UNIVERSIDAD DE SALAMANCA

Departamento de Estadística

Máster en Análisis Avanzado de Datos Multivariantes

Trabajo Fin de Máster

LA GEOGRAFÍA Y LA ESTADÍSTICA.

DOS NECESIDADES PARA ENTENDER BIG DATA

PEDRO JUANES NOTARIO

ROSA AMANDA SEPÚLVEDA CORREA

2014

Dpto. de Estadística

Universidad de Salamanca

ROSA AMANDA SEPÚLVEDA CORREA

Profesora Contratada Doctor del Departamento de Estadística de la Universidad de Salamanca

---

CERTIFICA que **D. Pedro Juanes Notario** ha realizado en la Universidad de Salamanca, bajo su dirección, el trabajo que, para optar al título de Máster en Análisis Avanzado de Datos Multivariantes, presenta con el título ***La Geografía y la Estadística. Dos necesidades para entender Big Data***, autorizando expresamente su lectura y defensa.

Y para que así conste, firma el presente certificado en Salamanca a 15 de septiembre de 2014.

Rosa Amanda Sepúlveda Correa

LA GEOGRAFÍA Y LA ESTADÍSTICA.

DOS NECESIDADES PARA ENTENDER BIG DATA

Dpto. de Estadística

Universidad de Salamanca

Trabajo para optar al título de

'Máster en Análisis Avanzado de Datos Multivariantes'

Por la Universidad de Salamanca

Presenta:

PEDRO JUANES NOTARIO

SALAMANCA

2014

"*Haz siempre lo que temas hacer, pues haciéndolo matarás el temor*"

R.W. Emerson

"The true value of Big Data analytics will be realised only when it acts as the unifying force that will take geospatial de the present state of a collection of building blocks to a self-reliant discipline"

(Dasgupta, 2011)

DEDICATORIA

A MI MADRE. Siempre.

AGRADECIMIENTOS

A María Purificación Galindo Villardón, directora del Máster y madre académica, por aceptarme y aguantarme ya en dos ocasiones. El siguiente, juntos.

A Rosa Amanda Sepúlveda Correa quien, con su conocimiento y consejos, me ha ayudado a enfocar aquello que no sabía ver.

A los profesores del Departamento de Estadística: Puri, Inmaculada, María José, José Luis, Santiago, Carmelo, Antonio, Javier…

A mi médico y a mi acupuntor, por darme ese poquito de salud que lo ha hecho posible.

A mi familia y amigos, por su apoyo y por creer siempre en mí.

Pedro Juanes Notario

Septiembre 2014

1 ÍNDICE

2 RESUMEN

La integración de la Geografía como ciencia con carácter analítico holístico y la Estadística Multivariante como herramienta lógica de investigación, posibilitan muchos de los elementos necesarios para el desarrollo de una nueva disciplina naciente dentro del conjunto de las Ciencias Sociales que debe dar el contexto académico necesario para el estudio y el análisis del Big Data: la GISciencia la cual se engloba dentro de la corriente cuantitativista más reciente para este conjunto de disciplinas.

En el presente Trabajo de Fin de Máster, a través de una exhaustiva revisión bibliográfica realizada para todos los aspectos que inciden horizontalmente dentro de estas dos ciencias con Big Data, se crea la aproximación teórica necesaria que permite encuadrar este fenómeno y presentar su actual estado del arte.

Palabras Clave:

Big Data, Estadística Multivariante, Geografía Cuantitativa, GISciencia.

Key Words:

Big Data, Multivariate Statistic, Quantitative Geography, GIScience.

3 INTRODUCCIÓN

La Geografía es la rama del conocimiento que estudia y analiza la interacción entre el hombre y el espacio. La misma, da lugar a la creación de figuras paisajísticas singulares en relación a las posibilidades que ambos ofrecen y disponen. Esta interacción, desde el punto de vista humano, ha sido analizada históricamente desde muchas perspectivas, pero ahora surge la necesidad de abrirla a una nueva realidad provocada por la capacidad que posee la sociedad en relación a su desarrollo tecnológico para crear un nuevo espacio de tipo virtual, pero totalmente real, con consecuencias directas de tipo socioeconómico e incluso político sobre las propias personas (en su manifestación individual y colectiva). Un territorio, por tanto, de naturaleza totalmente antrópica donde la información existente conforma un dominio espacial masivo de datos principalmente desestructurados: el Big Data.

Los datos conforman una nueva realidad social, política y económica presente y futura al abrirnos a un mundo contrapuesto, donde el límite entre el conocimiento y la ignorancia definen tanto sus nuevas fronteras, como su cambiante orografía. Y es que adecuando lo señalado por Boisier (2001), la **información** requiere **datos**, el **conocimiento** presupone información y la **compresión** precisa del conocimiento. De esta manera, la necesidad permanentemente de adquirirlo se ha convertido en un imperativo tanto para quien gestiona una empresa o gobierno, como para cualquier individuo. En el Siglo XXI, el conocimiento y su propia evolución son las claves tanto del crecimiento económico, como de la definición del puesto que ocupa cada país en su ordenamiento presente y futuro como territorios dentro de un mundo absolutamente globalizado. La actual Sociedad del Conocimiento, se basa en el permanente avance de las Tecnologías de la Información y de las Comunicaciones (TIC), las cuales han consolidado el conocimiento como nuevo factor de producción básico y diferenciador, ya que determinan la capacidad para innovar en un entorno de creciente facilidad de acceso a la información (Sakaiya, 1995). Y es que los datos, desde finales del Siglo XX, han generado espacio y el mismo, por lógica, debe ser estudiado por su ciencia: la Geografía, utilizando para ello la Estadística como herramienta natural de análisis.

Este Trabajo de Fin de Máster (TFM, a partir de ahora) no pretende analizar en profundidad la incorporación de estas ideas al campo del desarrollo territorial adscrito a la naciente GISciencia (ya que formará parte de los objetivos de un futuro trabajo de Tesis Doctoral, del cual el presente TFM es su preámbulo), pero sí es necesario señalar que, para el mismo, tal y como señalaba Boisier (2001) ha comenzado a configurarse un nuevo paradigma cuyos elementos claves son la interconexión y la

interactividad, la actuación individual y colectiva, la escala geográfica y virtual, el conocimiento y las innovaciones. Ya Wong-González (1999) hizo enumeración de las nuevas estructuras surgidas a raíz de la expansión de este hecho: se crean realidades, productos, empresas, monedas, centros de enseñanza, comunidades e incluso regiones todas ellas virtuales. Estas últimas, señala además, se encuentran en el punto que define la nueva modalidad de configuración territorial y responde a la lógica geográfica del actual capitalismo tecnológico propio de la globalización y ya no directamente del voluntarismo político de los Gobiernos. La virtualidad es el producto de las TIC, cuyo uso requiere tanto de hardware como de software, en definitiva, de conocimiento.

Actualmente se estima que alrededor del 80% de la información que las empresas y los gobiernos gestionan está georreferenciada y este porcentaje aumentará ya que estas organizaciones son conscientes de la importancia que tiene este atributo. La mayoría de los datos creados nacen con naturaleza espacial y por lo tanto, es natural, que la GISciencia asuma el rol principal dentro de las Ciencias Sociales para el análisis y comprensión de los datos individuales, locales, regionales y globales. El presente TFM pretende dar sentido y explicación a este hecho y enlazarlo con una realidad cada vez más necesaria: disponer estos datos con el objetivo global de analizarlos de forma correcta para conseguir conocimiento y de las oportunidades y soluciones que éste ha de generar; en definitiva, cómo se han de estudiar los registros que provienen del Big Data. Para ello es necesario plantear una metodología lógica de análisis en relación a sus características y es, en este aspecto tan importante, donde se encuentra la razón de aplicación de la Estadística Multivariante. Y es que la descripción de cualquier situación real requiere tener en cuenta de manera simultánea muchas variables y precisamente el análisis de datos multivariante tiene como objeto el estudio estadístico de varias variables medidas en elementos de una población, pretendiendo los siguientes objetivos (Peña, 2002):

- Resumir el conjunto de variables observadas en unas pocas nuevas variables hipotéticas, construidas como transformaciones de las originales con la mínima pérdida de información.
- Encontrar agrupaciones de los datos.
- Clasificar nuevas observaciones en grupos definidos.
- Relacionar dos o más conjuntos de variables.

Los cuales, definen las principales necesidades de análisis para comenzar a entender cualquier fenómeno que queramos estudiar vinculado al Big Data para, posteriormente, trabajar dentro del ámbito de la GISciencia con los mismos.

Para el desarrollo conceptual de todo lo mencionado, hemos estructurado el TFM en distintos capítulos a lo largo de los cuales se han desarrollado las características fundamentales para cada uno de los temas tratados. De esta manera, los aspectos que han influido y deben de influir en la realización de nuestro propósito final, que es la generación de soluciones analíticas de carácter holístico para el fenómeno Big Data, y que dará pie a la segunda parte del proyecto en forma de Tesis Doctoral, tienen cabida en el mismo. Así, se ha desarrollado un capítulo para cada apartado con entidad e importancia dentro de este propósito. Así, Internet (capítulo 6), Big Data (Capítulo 7), la Estadística Multivariante (capítulo 8) y la Geografía Cuantitativa (capítulo 9), conforman la estructura principal del TFM, realizando (dentro de cada uno de los desarrollos al respecto) una profunda revisión bibliográfica a tal efecto, que define y precisa su actual estado del arte.

4 OBJETIVOS

Uno de los requisitos que se presuponen siempre a la hora de comenzar un nuevo trabajo de investigación es el de aportar a la comunidad científica una nueva visión, conocimiento o herramienta al conjunto de recursos de los cuales disponemos los investigadores para el desarrollo de la Ciencia y ésta fue precisamente nuestra idea desde la propia gestación del proyecto. Las sucesivas búsquedas bibliográficas realizadas con el objetivo de conocer qué técnicas estadísticas se utilizan en Big Data nos han permitido concluir que son muchas las aportaciones científicas que implantan soluciones multivariantes a este fenómeno que parecen las más lógicas para su análisis, pero para casi todas ellas parten de la premisa de la necesidad de adaptarlas a la nueva realidad y necesidades que impone Big Data.

Por lo tanto, además, ante nosotros se planteaba un doble reto, el primero, que dio origen y sentido a este trabajo, fue el de recopilar la información necesaria para llevar a cabo en un futuro la Tesis Doctoral y el segundo, aún más atrayente, continuar perfeccionando el conocimiento necesario para el desarrollo de la Ciencia Estadística y la GISciencia en el ámbito del Big Data, las cuales se convierten en herramientas imprescindibles para el profesional en la investigación social de esta realidad.

Además, como parte de mi formación y especialización en Estadística Multivariante, rama en la cual he realizado los cursos de Doctorado y Máster por la Universidad de Salamanca, se ha incluido una capítulo en el TFM que ayuda a complementar sobremanera la profunda revisión bibliográfica llevada a cabo en el mismo; realizando un análisis de cada una de las técnicas multivariantes clásicas que ayudará al especialista, en primer lugar, a entender la repercusión que tiene el estudio del Big Data en el contexto científico a nivel mundial; y en segundo lugar, permitirá comprender la naturaleza general, especialización y particularidades de la investigación de este hecho también a ese nivel.

De igual manera, para poder analizar, comprender y ‘hacer uso’ de un fenómeno en concreto, necesitamos establecer un objetivo relacionado y disponer de una serie de herramientas que estén a nuestro alcance. Actualmente, el desarrollo de la GISciencia aborda esta necesidad, tal y como se señala en el apartado que ocupa este aspecto dentro del TFM (9.5). Pero más allá de la aplicación de técnicas cuantitativas sobre espacios físicos, es necesario definir la importancia que tiene para la Geografía el conocimiento de este ‘espacio virtual’, el cual está compuesto exclusivamente por datos generados antrópicamente. Por lo tanto, y en relación a esta realidad, ‘emerge’

hace escasos 3 años un fenómeno denominado Big Data que define una nueva forma de pensar, de observar y entender el mundo, que se basa en la enorme capacidad que existe en la actualidad de obtener y analizar grandes volúmenes de datos para establecer conclusiones. *Comprender el mundo a través de los datos masivos antrópicamente generados: no hay nada más geográfico que esta cuestión.*

En definitiva, partimos de un **objetivo general** que es: "la definición del estado actual del Big Data en el ámbito científico, a través de una profunda revisión bibliográfica de las dos disciplinas que han de interaccionar para su correcto análisis en el ámbito de las Ciencias Sociales: la Geografía Cuantitativa y la Estadística Multivariante".

Y como consecuencia del desarrollo del TFM, nos encontramos con una serie de **objetivos específicos** que dan lugar y tienen su reflejo en cada uno de los capítulos relacionados:

- Definir de manera concreta y concisa Big Data como fenómeno.
- Justificar la Estadística Multivariante aplicada a Big Data. La clásica 'falla'.
- Clasificar y calificar las técnicas multivariantes clásicas que se pueden aplicar a Big Data, utilizando para ello un lenguaje científico, pero no específicamente matemático.
- Presentar la GISciencia como el vehículo curricular adecuado para el tratamiento académico, conceptual y de desarrollo analítico final del Big Data en el escenario de las Ciencias Sociales.
- Establecer este trabajo, realizado con el necesario rigor científico, como 'punto inicial' para el desarrollo de una investigación que maneje por primera vez y de manera conjunta, en nuestro país y en nuestro idioma (hasta donde conocemos), las necesidades del Big Data teniendo en cuenta la interacción detallada de la Estadística Multivariante y la GISciencia.

5 MATERIAL Y MÉTODOS

Debido al carácter eminentemente teórico de este TFM, cuya pretensión final es realizar la descripción más detallada posible sobre el estado del arte acerca de todos los elementos y disciplinas científicas que han de interaccionar sobre cada uno de los elementos propios del Big Data, para posibilitar el análisis de sus datos, proceder a la extracción de la información contenida y finalmente obtener conocimiento a partir del mismo, su desarrollo ha estado condicionado por este objetivo y por ello la etapa de recopilación de la información necesaria ha sido laboriosa y ha ocupado buena parte del tiempo y del esfuerzo invertido para su realización. El material consultado para el desarrollo de cada uno de los capítulos ha sido seleccionado rigurosamente entre todos los documentos referidos a cada aspecto estudiado. Hay que destacar que, debido a la 'novedad' del fenómeno Big Data, las publicaciones que se han manejado en los puntos relacionados con el mismo son sumamente recientes debido a la dimensión e importancia que ha tomado en muy poco tiempo este fenómeno.

Mención aparte requiere el capítulo 8 dedicado al análisis de las técnicas estadísticas multivariantes clásicas, donde para su realización se utilizó el material y la documentación complementaria proporcionada por el **Departamento de Estadística de la Universidad de Salamanca** durante el desarrollo lectivo del '*Máster en Análisis Avanzado de Datos Multivariantes*' que he cursado. Y a la hora de objetivar el uso de cada una de estas metodologías en un entorno Big Data, se realizó su búsqueda concreta en Google Académico (*Google Scholar: http://scholar.google.es/*) utilizando para ello unos criterios muy específicos: así, las consultas realizadas (siempre en inglés, sin filtro temporal y buscando por tipo de documento 'PDF' para que los resultados de la misma nos mostraran aquellos artículos científicos disponibles en ese formato) y su resultado, a 20/06/2014, para cada uno de los métodos fueron los siguientes:

NOMBRE DE LA TÉCNICA	CONSULTA REALIZADA	RESULTADOS
Análisis de Componentes Principales	"Principal Component Analysis" + "Big Data" filetype:pdf	378
Análisis Factorial	"Factorial Analysis" + "Big Data" filetype:pdf	10
Análisis de Coordenadas Principales	"Principal Coordinate Analysis" + "Big Data" filetype:pdf	1 (Ampliamos su búsqueda a **www.google.com**, donde obtuvimos 29 resultados)
Multidimensional Scaling	"Multidimensional Scaling" + "Big Data" filetype:pdf	106
Análisis de Correspondencias	"Correspondence Analysis" + "Big Data" filetype:pdf	40
Análisis de Cluster	"Cluster Analysis" + "Big Data" filetype:pdf	399
Análisis Discriminante	"Discriminant Analysis" + "Big Data" filetype:pdf	173
Análisis de Correlación Canónica	"Canonical Correlation Analysis" + "Big Data" filetype:pdf	43

TABLA 1. Descripción y resultados de las consultas realizadas para las técnicas multivariantes analizadas.

Este buscador dio lugar a una serie de resultados en forma de artículos y presentaciones científicas y la consulta realizada individualmente a cada uno de los textos seleccionados para ser incorporados en la revisión realizada nos condujo a utilizar las más importantes bases de datos electrónicas donde se encontraban las mismas, entre ellas:

- *CrossRef*: http://en.wikipedia.org/wiki/CrossRef
- *Dialnet*: http://es.wikipedia.org/wiki/Dialnet
- *Science Direct*: http://en.wikipedia.org/wiki/ScienceDirect
- *Google Books*: http://es.wikipedia.org/wiki/Google_Books
- *Jstor*: http://es.wikipedia.org/wiki/Jstor
- *Open WorldCat*: http://es.wikipedia.org/wiki/Online_Computer_Library_Center
- *Scopus*: http://es.wikipedia.org/wiki/Scopus

A muchas de las cuales hemos podido tener acceso gracias a los recursos proporcionados a los alumnos y personal docente que ofrece el '*Servicio de Bibliotecas de la Universidad de Salamanca'*.

Otro aspecto que ha requerido de una metodología concreta ha sido la elaboración del glosario para todos aquellos términos que deben ser conocidos para poder seguir correctamente, por parte del lector, el desarrollo del TFM pero que no son explicados directamente en los capítulos correspondientes. Para ello y por cada término se ha realizado una búsqueda concreta del mismo (normalmente en inglés) utilizando Wikipedia como recurso principal de consulta.

6 INTERNET

El ‘fenómeno Big Data’ no sería posible, ni se podría entender sin la existencia de Internet, el espacio virtual en el cual ‘reside’ y se ‘desarrolla’, constituyendo de manera conjunta un auténtico ecosistema. De igual manera, absolutamente todo lo que pretenda analizar esta realidad funcional debe adecuarse a las necesidades y exigencias propias de Big Data y su hábitat. Tradicionalmente los desarrollos informáticos relacionados con soluciones estadísticas y GISciencia han sido conceptualizados para un entorno cliente-servidor (en el mejor de los supuestos) que no es el adecuado para las necesidades actuales de procesamiento de datos vinculados a la velocidad, la variedad, la veracidad y el volumen de los mismos en Big Data (ver 7.1). Por eso, resulta imprescindible cambiar absolutamente de filosofía, plantear las particularidades de esa interacción y, ante las mismas, definir nuevos requerimientos para que las soluciones informáticas creadas ‘*ad hoc*’ permitan gestionar de manera adecuada todo lo anteriormente postulado. En la actualidad se estima que el 80% de los datos no estructurados del mundo están desaprovechados, y sin embargo adquieren una tremenda importancia en el proceso de búsqueda del conocimiento en pleno siglo XXI. Utilizando un esquema computacional lógico (por su sencillez) los datos podrían ser adquiridos mediante tecnologías Web asincrónicas de tipo ‘*Open Source*’, almacenados y procesados usando bases de datos distribuidas, extraídos mediante técnicas de ‘*Data Mining*’ y ‘*Machine Learning*’, y finalmente presentados y visualizados por el usuario final (¿?) en estaciones de trabajo compuestas por ordenadores multinúcleo (Ch’ng, 2014). Y es que las características propias del fenómeno Big Data necesitan la adecuada infraestructura tecnológica para la gestión total y efectiva de los datos contenidos.

Por lo tanto es evidente que resulta necesario, para desarrollar el cuerpo teórico de este TFM, realizar una revisión conceptual de Internet y de la revolución que se está produciendo en los 3 últimos años por la aparición de una serie de realidades en su seno que han configurado 3 nuevas dimensiones (la ‘*nube*’, el ‘*Internet de las Cosas*’ y el ‘*Internet de Todo*’) en algo tan novedoso como él mismo y que también detallaremos puntualmente a lo largo de este capítulo. Para ello, hemos utilizado las premisas definidas por la Fundación ‘*Internet Society*’ (principal fuente independiente mundial sobre su política, sus estándares tecnológicos y su desarrollo) identificando el qué, el cómo, el cuándo y el porqué de esta ‘*red de redes*’.

Es evidente que Internet ha revolucionado la informática y las comunicaciones, pero aún es más cierto que ha cambiado el mundo. Su acceso 'libre' y 'abierto' ha revolucionado la forma en la cual los individuos se comunican y colaboran, y en cómo las empresas, los gobiernos y los ciudadanos interactúan. Al mismo tiempo, ha establecido un modelo abierto totalmente revolucionario para su propio desarrollo y administración, que abarca a todas las partes interesadas. Las antiguas invenciones (entre otras) del telégrafo, el teléfono, la radio y el ordenador, sentaron las bases para la integración de unas funcionalidades antes inimaginables. Internet, es a la vez, una herramienta de emisión y distribución de información sin fronteras posibles y un medio para la colaboración y la interacción entre personas y ordenadores sin la necesidad de una ubicación geográfica determinada, representando el ejemplo de mayor éxito acerca de los beneficios que conllevan la inversión y el compromiso continuo en el campo de la investigación y el desarrollo de las TIC en un ámbito de colaboración 'en abierto'. Y es que el desarrollo de Internet está basado en el establecimiento original de procesos abiertos. Fundamentalmente, Internet es una 'red de redes' cuyos protocolos están diseñados para permitir a esas redes interoperar de manera continua y creciente. Al principio, estaban representadas por algunas instituciones académicas, un gobierno (el estadounidense) y unas pocas comunidades de investigación, cuyos miembros necesitaban cooperar, desarrollando estándares de comunicación, para gestionar y aprovechar aquellos recursos que les eran comunes. Más tarde, cuando Internet fue comercializado, sus operadores se sumaron al proceso de desarrollo bajo ese protocolo abierto, dando lugar al inicio de una época de crecimiento e innovación sin precedentes. De manera más concreta, podemos sugerir que Internet aparece a finales de los años 60 con el proyecto ARPANET, consistente en la elaboración de una red de ordenadores integrada por varias universidades y algunos centros de investigación de EEUU, e inicialmente fue financiada con fondos del Departamento de Defensa de este país con un objetivo puramente militar y estratégico: proteger su sistema de comunicaciones ante un potencial ataque terrorista. Así, la información se transmitía por cauces alternativos y si una línea era dañada se podía redirigir a otra; además un mensaje o archivo no era enviado por la misma línea, sino que se dividía en paquetes y se remitía por distintas vías aprovechando las menos usadas en un momento determinado. Los principales servicios que disponía ese primer sistema eran los de intercambio de mensajes (e-mail), el envío y la recepción de archivos (FTP) y el acceso remoto a otra máquina (TELNET). Estos servicios posibilitaron el impulso de distintos proyectos de colaboración que integraban a varios grupos de investigación en lugares distantes, con la consiguiente mejora en la transferencia de los resultados y la disponibilidad de los recursos; en definitiva, del conocimiento.

Durante los años 70, ARPANET siguió su desarrollo y crecimiento y, ante este hecho, la agencia DARPA planteó un proyecto, denominado **Internetting** (del cual deriva su actual nombre), para desarrollar un protocolo que permitiera que varias redes intercambiaran información de manera segura; así surgió TCP/IP, vigente hoy en día en la red. Mientras que los protocolos de comunicación anteriores se basaban en considerar que cada nodo de la red era seguro y que existía capacidad suficiente para asegurar que los mensajes llegaran correctamente hasta el siguiente nodo, TCP/IP considera la red como 'poco fiable' y se basa en la confirmación única del mensaje entre el punto de origen y el de destino. De esta manera, si la red es atacada y parte de sus nodos destruidos, se asegura la recepción del mensaje reenviándolo automáticamente por caminos alternativos. A mediados de los años 80 la fundación norteamericana NSF propuso ampliar ARPANET a otros centros y universidades no involucradas en el proyecto original y se produjo su primera ampliación, comenzando a conectarse unas redes con otras creando una interconexión en cadena. Un mensaje pasa de una computadora a otra, de una red a otra hasta llegar a su destino. Es la 'red de redes', que en 1988 ya se conocía con su nombre actual: **Internet**. A finales de esta década su ámbito se expande a otros países y otras redes de desarrollo paralelo como Bitnet y Usenet, fueron uniéndose de manera progresiva hasta conformar un todo único. En 1991 el Congreso Norteamericano permitió su uso a las empresas privadas y a partir de ese momento comenzó su comercialización. Sin embargo, su expansión a gran escala y la revolución social que ha implicado no se produce hasta mediados de los años 90, cuando se desarrollaron lenguajes que permitieron crear programas en entorno gráficos, que posibilitaron al usuario acceder a la información disponible en ella de una manera sencilla.

Actualmente, la información disponible en Internet (dejando a un lado el denominado 'Internet Profundo') es enorme y continúa aumentando exponencialmente año tras año (para más detalle, ver informe a fecha de 9 junio de 2014: http://bit.ly/ZqZpOn) y su impacto en la sociedad ha cambiado radicalmente incluso nuestra forma de vida a través de cada uno de los aspectos que a continuación desarrollaremos y a los cuales hacíamos una ligera referencia al principio del capítulo:

- La 'Nube'.
- El 'Internet de las Cosas' (IoT).
- El 'Internet de Todo' (IoE).

6.1 La 'Nube'

La creación en Internet de un espacio virtual denominado '*nube'* (término castellano proveniente del original '*cloud'* en inglés) es un hecho, y resulta imprescindible definir en qué consiste debido a la importancia que adquiere para cada uno de los conceptos que estamos definiendo e interpretando. Y es que, tal y como señalábamos al comienzo de este capítulo, cualquier desarrollo que pretenda tener carácter analítico sobre Big Data deberá realizarse conociendo las características actuales y evolutivas de este medio para poder actuar con eficacia.

El concepto '*Cloud Computing'* (o 'Computación en la Nube') se crea como consecuencia de los servicios ofrecidos por proveedores de Internet a gran escala como Google y Amazon que construyeron, en relación a sus necesidades, su propia infraestructura y que originó el desarrollo de una nueva arquitectura basada en un sistema de recursos distribuidos horizontalmente, escalados masivamente, manejados como recursos configurables y mantenidos automáticamente de manera continua, que configuran un nuevo sistema TIC de servicios virtuales y que sigue el modelo de arquitectura propuesto por Gilder (2006). El National Institute of Standards and Technology (NIST) define el término '*Cloud Computing*' como un modelo que permite el acceso a recursos de red de forma ubicua, conveniente y bajo demanda a través de un área compartida (o '*Shared Pool'*) compuesta por recursos computacionales configurables (redes, servidores, dispositivos de almacenamiento, aplicaciones y otros servicios) y que pueden ser rápidamente dimensionados con un mínimo esfuerzo de gestión o servicio por parte del proveedor (Mell & Grance, 2011). Este modelo se caracteriza por la existencia de:

5 CARACTERÍSTICAS ESENCIALES	3 MODELOS DE SERVICIO	4 MODELOS DE IMPLEMENTACIÓN
Autoservicio Bajo Demanda	Software as a Service (SaaS)	Nube Privada
Acceso Ubicuo a la Red	Platform as a Service (PaaS)	Nube Comunitaria
Agrupación de Recursos	Infrastructure as a Service	Nube Pública
Rápida Elasticidad	(IaaS)	Nube Híbrida
Servicio a Medida		

TABLA 2. Definición '*Cloud Computing*'. Fuente: NIST.

5 características esenciales: La computación en '*nube*' presenta las siguientes claves (Mell & Grance, 2011):

- *Autoservicio Bajo Demanda ('On-demand self-service').* Un consumidor de forma unilateral puede determinar, en relación a sus necesidades, la infraestructura informática que necesita en cada momento (el tipo de servidor, el ancho de banda de la red, el tiempo y el tamaño requerido para el almacenamiento de su información, etc.) y este cambio se realiza de manera automática y sin requerir la interacción humana por parte del proveedor de los servicios que contrata.
- *Acceso Ubicuo a la Red ('Broad Network Access')*. Los recursos de la red tienen una alta y sencilla disponibilidad, permitiendo su acceso a través de dispositivos estándar, promoviendo el uso de distintas plataformas (teléfonos móviles, tabletas, portátiles, estaciones de trabajo, etc.)
- *Agrupación de Recursos ('Resource Pooling').* Los recursos informáticos del proveedor se agrupan con el objeto de servir a múltiples y diferentes consumidores utilizando un modelo de multitenencia, que maneja diferentes recursos tanto físicos, como virtuales, reasignándolos dinámicamente según la demanda para un momento concreto. Generalmente el cliente tiene un sentimiento de deslocalización respecto a los mismos, ya que no tiene el control ni el conocimiento de la ubicación exacta de esos recursos, pero puede llegar a ser capaz de especificar esa localización a un nivel de abstracción superior (país, región o Datacenter).
- *Rápida Elasticidad ('Rapid Elasticity').* Sus recursos puede ser gestionados automáticamente de manera totalmente elástica de acuerdo a la demanda de cada momento. El consumidor puede llegar a percibir que estos recursos, que le proveen de servicios, parecen ilimitados y que son los apropiados para cualquiera de sus necesidades en cualquier momento y bajo cualquier circunstancia.
- *Servicio a Medida ('Measured Service').* Los recursos están, de forma automática, permanentemente controlados con el objetivo de optimizar su rendimiento mediante técnicas de medición en tiempo real. Este uso puede ser monitorizado, controlado e informado en el momento en el cual ocurre, ofreciendo una absoluta transparencia para el proveedor y para el consumidor.

- *3 modelos de servicio*: La '*nube'* es el intermediario virtual entre el proveedor del servicio y el usuario final, el cual puede acceder a ella desde cualquier dispositivo con acceso a Internet. De esta manera, el proveedor gestiona desde sus instalaciones los recursos que presta a sus clientes y estos recursos se virtualizan y pasan a formar parte de este espacio virtual. Cuando nos referimos a desarrollar o trabajar en la '*nube*', tenemos que puntualizar de qué manera lo vamos a hacer ya que existen distintas formas para ello. Así, se pueden identificar tres modelos de servicio (Fernández, Leyton & González, 2011; Mell & Grance, 2011), que detallamos a continuación y que también se definen gráficamente en la IMAGEN 1:
  - *'Software as a Service' (SaaS):* La capa '*Software como Servicio'*. El proveedor pone a disposición de los clientes su software, evitándoles su mantenimiento o la compra de licencias. Este concepto ha existido desde hace mucho tiempo y básicamente se trata de cualquier servicio basado en la web, donde el cliente normalmente accede a través de un navegador sin atender al software. Todo el desarrollo, mantenimiento, actualizaciones y copias de seguridad es responsabilidad del proveedor. En este caso el cliente tiene poco control, situándose en la parte superior de la capa del servicio. Este tipo de servicio es el adecuado para la fase de consumo de aplicaciones finales en la nube. Ejemplos: Google Docs, Dropbox y Gmail.
  - *'Platform as a Service' (PaaS):* La capa '*Plataforma como Servicio*'. Muy ligada a la capa SaaS, constituye el medio donde los desarrolladores programan las aplicaciones que se ejecutan en la '*nube*' preocupándose únicamente de este hecho, ya que la infraestructura y su mantenimiento lo proporciona exclusivamente la plataforma (compuesta por uno o varios servidores de aplicaciones y una base de datos), ofreciendo la posibilidad de ejecución de esas aplicaciones. El proveedor puede llegar incluso a proporcionar las herramientas para su desarrollo dentro de su plataforma. El inconveniente es que este servicio deja el control de los datos alojados enteramente al proveedor, con la problemática que puede conllevar. Este tipo de servicio es el adecuado para la fase desarrollo de aplicaciones en la nube (fase de 'construcción': entorno de desarrollo y prueba de software). Ejemplo: Google App Engine.
  - *'Infrastructure as a Service' (IaaS):* La capa '*Infraestructura como Servicio*'. Es la parte física de la nube, donde el cliente gestiona la infraestructura. A través del manejo de máquinas virtuales, elige el tipo de instancias que desea utilizar (Windows o Linux), así como la capacidad de memoria y el tipo de procesador de cada una de las máquinas, todo ello de manera totalmente transparente, ya

que el hardware que el cliente gestiona es virtual, pagando a un proveedor para que sea éste quien sea el propietario físico de esa infraestructura TIC y se encargue, también físicamente, de su mantenimiento y optimización. En ocasiones, este modelo también es denominado *HaaS* (‘*Hardware as a Service’*). Este tipo es el adecuado para el ‘*Hosting*’ (cómputo, conexión y almacenamiento). Ejemplos: Amazon Web Service y Microsoft Azure.

IMAGEN 1. Modelos de servicio ‘Cloud Computing’. Fuente: Adaptado de Gómez (2013).

- *4 modelos de implementación:* Existen cuatro grandes ‘tipos de nube’ dependiendo de su modelo, el cual está definido por su propósito y por su localización en relación las necesidades de cada usuario, el modelo de servicio ofrecido y la implementación de la misma (Furht & Escalante, 2010; Mell & Grance, 2011; Sosinsky, 2010). A continuación definimos textualmente cada una de ellas, pudiendo visualizar esquemáticamente las mismas en la IMAGEN 2:
    - *Nube Pública (‘Public Cloud’):* Su infraestructura se determina para el uso en abierto de los servicios para que puedan ser consumidos por cualquier persona que quiera disponer de ellos. Es el modelo estándar de servicio en la nube, donde los productos que se ofrecen se encuentran en servidores externos al usuario, el cual tiene acceso a los mismos de forma gratuita o pagando por ellos. Contiene aplicaciones de software, sitios de almacenamiento y otros recursos digitales.
    - *Nubes Privadas* (‘*Private Cloud*’). Su infraestructura se determina para el uso exclusivo de una sola organización que comprende varios consumidores (por ejemplo, sus unidades de negocio). Puede ser administrada por la misma organización, por una tercera parte o por alguna combinación de ellos y sus infraestructuras pueden existir dentro o fuera de su ubicación. Las nubes privadas son una buena opción para las

compañías que necesitan alta protección y edición de sus datos. En este tipo de nubes, el cliente controla quién utiliza sus aplicaciones y cómo lo hace.

- *Nubes Comunitarias* (‘*Community Cloud*’). Su infraestructura se determina para el uso exclusivo de una comunidad específica de consumidores pertenecientes a dos o más organizaciones que comparten objetivos (requisitos de seguridad, política, consideraciones de cumplimiento, etc.) Puede ser administrada por una o más de las organizaciones de la comunidad, por una tercera parte o por alguna combinación de ellos y sus infraestructuras pueden existir dentro o fuera de su ubicación. La nube comunitaria normalmente comparte el almacenamiento de los datos y los servicios electrónicos con el objetivo de minimizar los costes derivados de la gestión y administración de esta tecnología.
- *Nubes Híbridas (‘Hybrid Cloud’)*. Su infraestructura está compuesta por dos o más tipos de nubes (privadas, comunitarias o públicas) y éstas pertenecen a cada uno de los propietarios originales, uniéndose entre ellos para, a través del uso de tecnología estandarizada (o propietaria), posibilitar el intercambio de datos y aplicaciones. Las más comunes, son aquellas que combinan los recursos de una nube privada con una nube pública. De esta manera, la infraestructura privada ve mejorada su capacidad con los servicios de computación en nube de la infraestructura pública. Esto permite a una organización mantener el control de sus principales aplicaciones y aprovechar la computación de esa nube externa solamente cuando considere necesario.

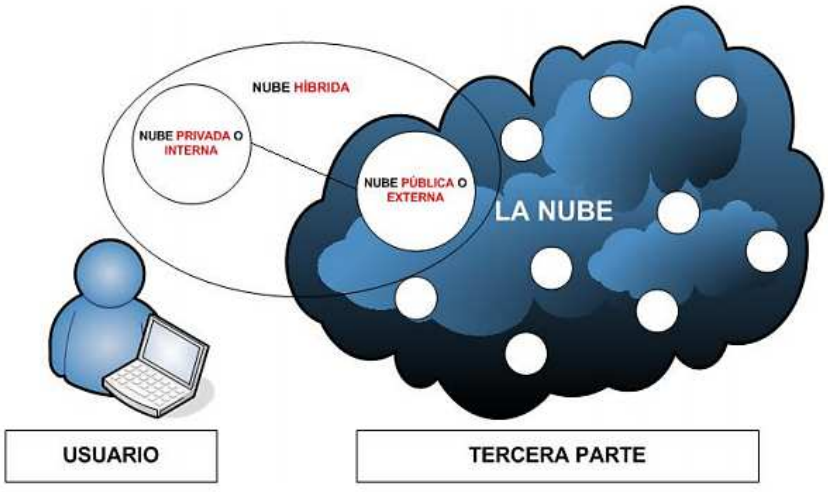

IMAGEN 2. Tipos de ‘*Nubes’*. Fuente: Gómez (2013).

6.2 El Internet de las Cosas (IoT)

El **Internet de las Cosas** (IoT a partir de ahora, por sus siglas en inglés '*Internet of Things*') no es una idea novedosa. Weiser (1993) director del '*Xerox Palo Alto Research Center'*, introdujo el concepto de '*Computación Ubicua*', que definía un futuro en el cual la informática (hardware y software), estaría tan integrada en el entorno que desaparecería de la vista y formaría parte integral de nuestra vida diaria resultando absolutamente 'transparente'. El término exacto fue postulado por el Auto-ID Center del MIT a finales de los años 90 pero, sin embargo, la idea del IoT no ha tomado relevancia práctica hasta la última década gracias a la rápida evolución de la electrónica y sobre todo por el impulso de multinacionales como CISCO Systems.

La progresiva implantación del IoT, significa que multitud de objetos cotidianos estarán dotados de sensores que harán las veces de nuestros sentidos y les permitirán generar continuamente información tanto del medio que les rodea '*Context Awareness'*, como de su posición geográfica '*Location Awareness*'. Este entorno ha sido ya denominado de diferentes formas: '*Pervasive Computing*' (Weiser, 1993), '*Web Squared*' (O'Reilly & Battelle, 2009) o '*Everyware'* (Greenfield, 2010) y supone que el futuro-presente de Internet estará definido por su inevitable encuentro con el mundo físico y real. Esta enorme cantidad de nueva información formará una piel digital que cubrirá el mundo físico y generará nuevas oportunidades para las organizaciones privadas y públicas (Chui, Löffler & Roberts, 2010).

Tradicionalmente en las organizaciones (gubernamentales y empresariales), tanto la información que generaban, como la que obtenían de manera externa (a través de fuentes públicas, directamente de Internet o comprada a los proveedores de información) era canalizada a través de sus redes, se alojaba en sus bases de datos y se analizaba en sus sistemas para generar informes que eran utilizados por sus órganos de gestión. Pero las vías de obtención de la información están cambiando; así, el mundo físico se está convirtiendo en un verdadero sistema de información gracias al desarrollo del IoT. El despliegue masivo que se está produciendo, y que verá exponencialmente incrementada la implantación en los próximos años de sensores y otros dispositivos con muy diversas funcionalidades, insertados en multitud de objetos físicos con acceso continuo a Internet, está creando volúmenes enormes de datos que deben llegar a las infraestructuras computacionales para proceder a su análisis. Cuando los objetos puedan 'sentir el ambiente' y 'comunicarse' con él (incluso trabajando sin la intervención humana, tal y como exponíamos en el párrafo anterior)

habremos dado un paso muy importante para comprender su complejidad y responder rápidamente a cualquier cambio que se produzca.

La adopción generalizada del IoT llevará tiempo, pero está avanzando rápidamente gracias a una serie de mejoras que se están produciendo en las tecnologías subyacentes:

- Los avances en la tecnología de red inalámbrica.
- La estandarización de los protocolos de comunicación (que permiten recoger datos de estos sensores en cualquier lugar y momento).
- La miniaturización de los chips de silicio construidos para este propósito y que están adquiriendo nuevas capacidades y posibilidades, como su incremento masivo en potencia de cálculo, siguiendo el patrón marcado por Gordon Moore cofundador de Intel que en 1965 postuló la '*Ley de Moore*', la cual señala que el número de transistores que contiene un chip se duplica cada dos años aproximadamente, lo cual se ha venido cumpliendo durante los últimos cuarenta años (Manyika et al., 2011).
- La posibilidad de almacenar esa información disponible a través del '*Cloud Computing*'.
- La caída de los costos de producción de esta tecnología, al haber entrado su fabricación en una economía de escala masiva.

Todos estos factores harán posible la implantación del IoT en todos los ámbitos imaginables (Chui, Löffler & Roberts, 2010), con una característica además esencial para nuestro objetivo que motiva uno de nuestros principales propósitos de estudio: la información que se obtenga de manera contínua a partir del IoT estará georreferenciada desde su origen (se estima que en más de un 80%) y será necesario desarrollar métodos específicos de análisis estadístico implementados en tecnología propia de la GISciencia para atender a las necesidades propias de este fenómeno y su evolución.

6.3 El Internet de Todo (IoE)

El verdadero origen y '*leitmotiv'* de la revolución que acompaña a Big Data es el **Internet de Todo** (IoE, a partir de ahora, por sus siglas en inglés '*Internet of Everything')*. Este fenómeno consiste en la conexión a la 'red de redes' de personas, procesos, datos y objetos en cualquier momento y ubicación. El beneficio del IoE deriva del impacto producido por la conexión de todos ellos y del valor que esta creciente interconexión crea como consecuencia de que 'todo' se transmite y de que 'todo' está en línea. Cada vez son más las personas conectadas a la red y cada vez son más los objetos implantados en nuestro entorno que tienen sensores integrados con capacidad de 'comunicación', lo que provoca que las actuales barreras que separan la realidad de lo virtual se estén difuminando paulatinamente. El mundo se está convirtiendo en un campo de información global y la cantidad de datos que circulan por las red está creciendo exponencialmente (Feller et al., 2011). Las operaciones que décadas atrás requerían de un ordenador del tamaño de una habitación son, hoy en día, realizadas por dispositivos electrónicos de una dimensión tan minúscula como un microchip. El tamaño, el coste y el consumo de energía del hardware se han reducido drásticamente, por lo que ahora es posible fabricar dispositivos electrónicos diminutos a un precio muy reducido. Estos pequeños ingenios, junto con la expansión de las redes inalámbricas de comunicación, permiten incorporar 'inteligencia' y conexión a los objetos del mundo real y están transformando lo que era una red global de personas en una red global de '*todas las cosas*' (de todos) y '*para todo*' (Wong, 2005). Esta es la base del Big Data, aquí aparece, aquí se desarrolla socialmente y aquí cobra completo sentido, de ahí su importancia.

Está claro que vivimos en un mundo interconectado. Las redes sociales a finales del año 2013 casi alcanzaban los 1.900 millones de usuarios activos y cada día esta cifra se incrementa enormemente. Actualmente 2.700 millones de personas se conectan a Internet, comparten información y se comunican a través de mails, blogs, redes sociales y muchos otros medios (Carnés, 2014; UIT News, 2013). Ahora, con el IoE, los objetos (IoT) se unen al Internet 2.0 y a la Web Semántica (que algunos ya denominan Internet 3.0) y conforman este fenómeno que cobra una tremenda importancia.

De acuerdo con el estudio '*The Digital Universe of Opportunities*' desarrollado por IDC (2014), los datos generados crecerán 10 veces para el año 2020, pasando de los 4.4 ZB existentes a los 44 ZB (ver Anexo1). Para ese año se estiman que 32.000 millones de dispositivos IoT estarán conectados a Internet y el 53% de los datos estarán en *'la*

nube'. Por su parte, la empresa CISCO Systems (una de las grandes impulsoras de este fenómeno, como señalábamos en el apartado anterior) calcula que en 2014, 100 nuevas 'cosas' (personas, lugares, procesos y objetos) por segundo se están conectando a Internet y en el año 2020 serán 250 nuevas 'cosas' las que lo harán a esa velocidad. También pronostica que de los 13.000 millones de 'cosas' conectadas en 2014, pasaremos a 50.000 millones (7 por cada persona existente en el mundo en ese momento) en 2020. Estas estimaciones se han realizado siguiendo una metodología estadística precisa que es explicada concisamente en la siguiente dirección: http://bit.ly/1tvjFVP donde además se puede visualizar, a través de un contador (ver IMAGEN 3), el incremento 'en tiempo real' de ese número de 'cosas' que se están conectando a Internet acumulativamente en ese momento preciso (siguiendo los patrones marcados por esa estimación).

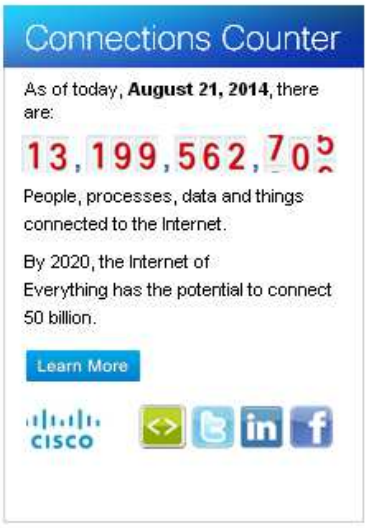

IMAGEN 3: Contador de conexiones a Internet ('Cosas'/Segundo) Fuente: http://bit.ly/1tvjfvp

Día tras día, como podemos comprobar, los ejemplos siguen aumentando, el IoE es real, ha llegado y sin embargo, no ha hecho más que empezar y, evidentemente, va a cambiar la forma en casi todas las personas viven, trabajan, juegan, aprenden… es evidente que esos miles de millones de 'cosas' conectadas generarán grandes cantidades de datos que pueden cambiar radicalmente incluso la forma en la cual concebimos la vida en nuestro mundo (ver IMAGEN 4. ILUSTRACIÓN). Pero, como estamos señalando a lo largo de todo el TFM, los datos en bruto no son suficientes para ello, todos esos bits deben ser 'tamizados' para encontrar información útil y luego ser transformados en conocimiento, para finalmente traducirlos a sabiduría. El potencial social, político y económico acumulativo de esta convergencia de fuerzas

tecnológicas será entonces casi inimaginable e inevitablemente podrá cambiar la forma en que las personas, los gobiernos y las empresas interactuarán entre ellos.

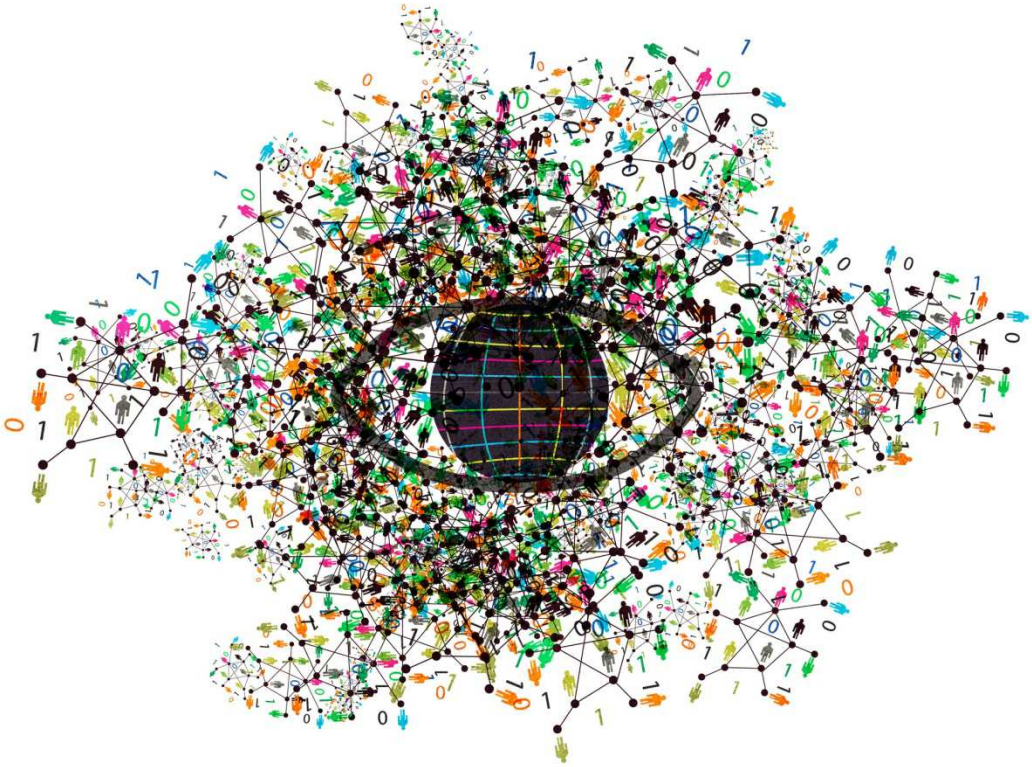

IMAGEN 4. Ilustración 'IOE'. Fuente: Lohr (2012).

7 BIG DATA

Paradójicamente, los datos siempre han sido grandes y ahora aparece Big Data como ¿un fenómeno novedoso? Aunque es cierto que la verdadera dimensión de los datos masivos se ha creado prácticamente en los 3 últimos años (el 90% de los registros existentes se ha generado en este periodo de tiempo), también es real que los investigadores, las empresas y los gobiernos desde mediados del siglo pasado han manejado enormes volúmenes de información, recabando, manejando y manteniendo grandes conjuntos de registros para, entre otros aspectos, analizar el clima, desarrollar fármacos, generar modelos físicos y económicos, etc. y todos estos objetivos implican 'mover' grandes cantidades de registros (Ohlhorst, 2012). Se habla de la Big Ciencia y con este concepto, se describen y engloban una serie de cambios en la investigación científica ocurridos en los países industrializados desde el inicio de la 2ª Guerra Mundial, donde el progreso científico se aceleró notablemente y comenzaron a gestarse proyectos a gran escala, por lo general financiados gubernamentalmente a escala nacional e internacional.

Además, a partir de la irrupción de esos megaproyectos, lo grande ha sido un adjetivo cambiante y continuamente modificado al alza por la disminución acelerada en los costos de procesamiento y almacenamiento en informática. A pesar de esta larga historia de uso en la administración y las empresas como elemento de ayuda en sus análisis y tomas de decisiones, Big Data ha surgido recientemente como un área de investigación por sí misma y el que lo haya hecho con tanta fuerza es producto, probablemente, de un conjunto de circunstancias, como es el hecho de que nuestra posibilidad para detectar, recopilar y procesar gran cantidad de datos heterogéneos provenientes de múltiples fuentes está superando nuestra capacidad para su correcto almacenamiento y gestión (Warden, 2011). Pero es tan rápida la evolución de las TIC, que a la par que se resuelven los problemas técnicos derivados de su manejo, ya se están generando soluciones que permiten extraer valor del mismo como nunca antes se había hecho desde un punto de vista del conocimiento (Buchholtz, Bukowski & Śniegocki, 2014). Al mismo tiempo, incluso estamos empezando a ver un cambio importante en la forma en que muchos científicos están procesando estos datos y procediendo a su análisis para generar hipótesis mediante procesos abductivos (desarrollando modelos directamente a partir del conjunto de datos observados), que parece que podría provocar un profundo cambio en el paradigma del pensamiento científico (García & de Prado, 2012), al afirmar que es suficiente con analizar la 'correlación' existente en un 'fenómeno Big Data' para poder extraer conocimiento del

mismo, olvidando la importancia de la 'causalidad' y abogando por que estamos entrando científicamente en la 'era de la correlación', lo cual puede llegar a ser desastroso y hace más necesario aún plantear la necesidad de desarrollar una ciencia de carácter social que maneje de manera conveniente y trabaje, con las herramientas adecuadas, estos datos generados de manera antrópica: la GISciencia (Farmer & Pozdnoukhov, 2012).

Cada día el término Big Data aparece en cualquier tipo de publicación. Ya no es un fenómeno puramente informático y se ha convertido, incluso, en una tendencia social. Aun así, su propio concepto sigue generando confusión; y es que la amplia cobertura mediática que está recibiendo no permite distinguir claramente cuál es su naturaleza y en qué consiste su realidad (Ohlhorst, 2012). Big Data se ha utilizado para trasladar a las personas todo tipo de conceptos de carácter incluso dispar, entre los que se incluyen '*enormes cantidades de datos'*, '*analítica de redes sociales*', '*herramientas de última generación para la gestión de datos*', '*información en tiempo real (streaming)*' y un largo etcétera. Independientemente de la etiqueta que le otorguemos, es un fenómeno tan novedoso, en realidad, que actualmente en su etapa inicial estamos comenzando a comprender y explorar nuevas formas de procesamiento y análisis que le den la cobertura adecuada (Schroeck, Shockley, Tufano, Smart, & Romero-Morales, 2013).

7.1 Definiendo Big Data

Siempre es difícil identificar cómo y cuándo surgen las ideas y los conceptos en un entorno tan rápidamente cambiante como el de las TIC. Big Data no supone una excepción y es difícil identificar si surge como consecuencia de otros conceptos como el Open Data, pero lo que está claro es que la popularización del término está ligada al documento publicado por el McKinsey Global Institute (MGI) en Junio de 2011, donde lo define como "*conjuntos de datos cuyo tamaño va más allá de la capacidad de captura, almacenamiento, gestión y análisis de las herramientas de base de datos*" (Manyika et al., 2011).

A primera vista, el término Big Data es bastante impreciso y parece hacer referencia 'a algo' que es grande y que está lleno de información. Esta descripción se ajusta, de hecho, de manera sucinta a la realidad pero, sin embargo, no proporciona ninguna declaración sobre lo que exactamente es. A menudo, este fenómeno es descrito (como veíamos antes) como un conjunto de datos extremadamente grande, que ha crecido más allá de lo que permiten la capacidad de gestión y análisis de las

herramientas de procesamiento de datos tradicionales (Manyika et al., 2011; Slocum, 2011), pero verdaderamente, las principales dificultades están relacionadas con la adquisición, el almacenamiento, la búsqueda, el intercambio, pero sobre todo con el análisis y la visualización de la información obtenida a partir de estos datos. De esta manera, el mismo significado del concepto ha ido evolucionado en los 2 últimos años para incluir, no sólo su tamaño, sino también los procesos que intervienen para su aprovechamiento. Big Data incluso ha llegado a ser sinónimo de ideas de negocio como el *'Business Intelligent'* (BI) o el '*Data Mining*' (Manyika et al., 2011). Actualmente, no se puede llegar a una definición casi universal compartida por la mayoría de los que promueven esta ideología, ya que se puede describir:

- *En forma de problema*: Big Data define una situación en la que el conjunto de datos existente ha llegado a tener una dimensión tan grande, una heterogeneidad tan diversa y un crecimiento tan exponencial que las TIC convencionales no pueden manejarlo de manera efectiva y aún resulta más difícil generar información objetiva a partir del mismo.
- *En forma de solución*: Big Data es el conjunto de herramientas, procesos y aptitudes que van a permitir la gestión de estas enormes cantidades de información para mejorar los resultados de las organizaciones.

Para Buchholtz et al. (2014), Big Data es producto de la última fase de desarrollo de las TIC y junto al Open Data conforman su revolución actual, consecuencia lógica de las mejoras exponenciales que ha sufrido el hardware y el software desde finales de la década de los 60 y complementada por un cambio de mentalidad respecto a la apertura en la tenencia de los datos.

Big Data también es conocido a través de otros términos como '*Big Data Analytics'*, '*Value Data'*, '*Smart Data', 'Fast Data'* y aunque algunas de las discusiones existentes están referidas a cómo definirlo correctamente, lo verdaderamente importante es atender un conjunto de necesidades que abarcan varios aspectos del mismo (Feller et al., 2011):

- Almacenamiento y procesamiento masivo.
- Heterogeneidad e integración.
- Facilidad de explotación.
- Análisis avanzado.
- 'Data Mining'.

Términos como Gigabyte (mil millones de bytes), o Terabyte (un billón de bytes) se están quedando desactualizados y dan paso a los Petabytes (mil billones de bytes) o Exabytes (un trillón de bytes) que reflejan mejor la cantidad real de la información global (Anexo1), y que pronto también quedarán desfasados. En muchos textos se preocupan más por la necesidad de almacenar esa gran cantidad de datos heterogéneos y, sin embargo, este hecho no responde esencialmente a la necesidad final del usuario (Feller et al., 2011). En cifras tomadas de UIT News (2013) y Carnés (2014), a diciembre de 2013 (los datos de esta realidad a septiembre de 2014 ya han sido ampliamente superados):

- La población Mundial se estimó en 7.095 millones de habitantes.
- El volumen de usuarios de telefonía móvil alcanzaba los 6.573 millones, de los cuales 2.100 millones eran usuarios de Smartphones u otro tipo de dispositivos móviles con capacidad de acceso a Internet.
- Según la UIT (Unión Internacional de Telecomunicaciones) 2.700 millones de personas (equivalente al 39% de la población mundial) utilizaban Internet, y de éstos casi 1.857 millones eran usuarios activos de alguna de las plataformas sociales existentes. Esto supone que haya 750 millones de hogares conectados a la red.
- En lo que respecta al acceso regional, Europa registra la tasa de utilización de Internet más elevada del mundo (75%), seguida del conjunto formado por los países del continente Americano (61%). La Comunidad de Estados Independientes (CEI) ocupa el tercer lugar (52%), seguida de los Estados Árabes (38%), la región de Asia-Pacífico (32%) y finalmente África (16%).

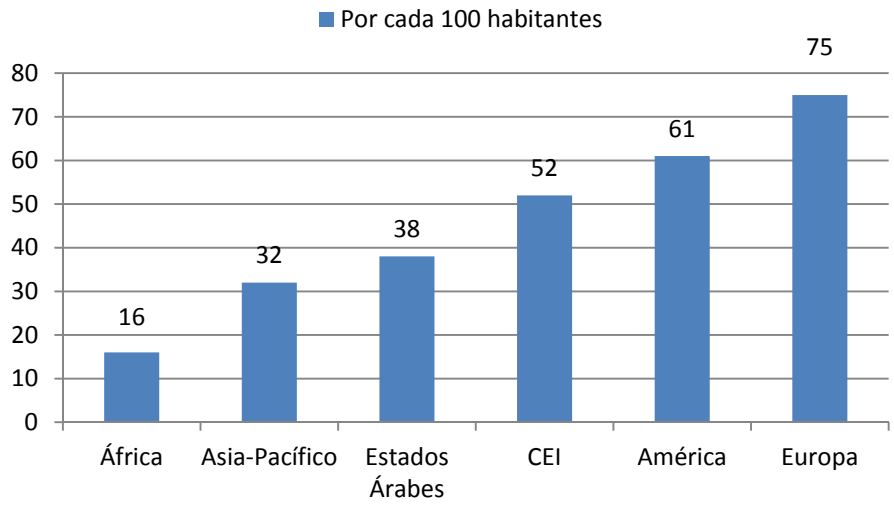

GRÁFICO 1. Usuarios de Internet por región. Fuente: http://bit.ly/1qf3eqn

- China es el país del Mundo con mayor volumen de Usuarios de Internet (alrededor de 604 millones), de Telefonía Móvil (más de 1.206 millones) y de Smartphones (246 millones).
- Actualmente, hay más de 650 millones de páginas web (más de un millón de ellas son falsas, se clonan para engañar al usuario y extraerle información).
- Existen unos 148,5 millones de dominios, de los cuales:
    - 112,5 millones son .COM.
    - 15,25 millones son .NET.
    - 10,4 millones son .ORG.
    - 5,85 millones son .INFO.
    - 2,65 millones son .BIZ.
- El número de cuentas de correo electrónico en el mundo asciende a 3.900 millones, de las cuales aproximadamente el 76% corresponden a usuarios particulares y el 24% a cuentas de empresa, y el número de usuarios de correo electrónico en el Mundo es de unos 2.420 millones.

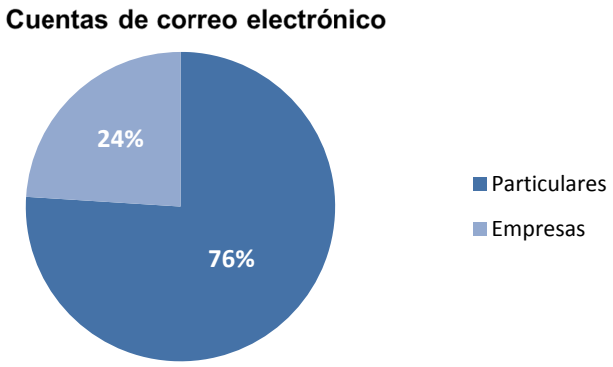

GRÁFICO 2. Cuentas de e-mail por tipo de usuario. Fuente: http://bit.ly/1qf3eqn

- El servicio Web Mail gratuito más utilizado es Gmail (solución de correo electrónico de Google), con unos 450 millones de Usuarios.
- Diariamente se envían unos 183.000 millones de correos electrónicos, de los cuales el 68% son Spam (casi el 51% del total emails enviados son Spam proveniente de la industria farmacéutica), y el 2,3% de éstos contienen algún tipo de Malware.
- La búsqueda continúa siendo la categoría online que concentra mayor volumen de audiencia, alrededor del 91,5% del total de Usuarios de Internet (2.640 millones) hacen 'consultas' a través de algunos de sus principales motores (Google, Yahoo!, Bind, etc.)

- En cuanto a las redes sociales, a continuación en la IMAGEN 5 se muestran los datos actualizados para las principales plataformas.

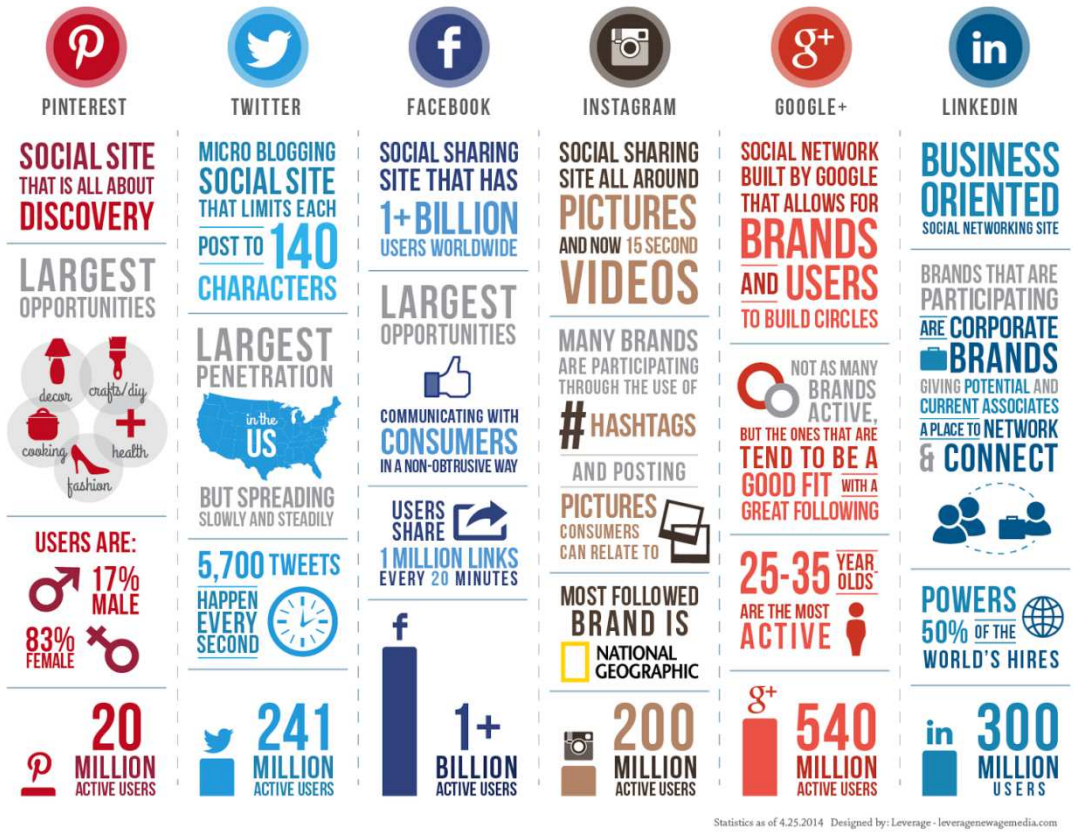

IMAGEN 5. Infografía de las características y grado de utilización de las 6 principales redes sociales, con datos de 24/04/2014. Fuente: http://bit.ly/1uw7uzn

- En 2013, el universo digital alcanzó los 900.000 Petabytes (1 Petabyte es un millón de Gigabytes). En 2014 llegará a los 2,1 millones de Petabytes o 2,1 Zetabytes. Si se mantiene este ritmo (advirtiendo que la estimación puede quedar muy desfasado a la baja), para el año 2020 la realidad digital será cuarenta y cuatro veces más grande que en el año 2013. Se señala además que el 90% de los datos presentes hoy en la red fueran creados en los últimos 3 años.
- El MGI estima que las empresas y los consumidores almacenaron un total de 13 Exabytes de datos nuevos en el año 2010 en todo el mundo (no hemos encontrado información más reciente al respecto). Además, calcula que el volumen de información creada, capturada y replicada en 2010 fue de 1,8 millones de Petabytes, lo que supone un crecimiento del 125% respecto al año 2009. Para el 2020, la consultora estima que se alcanzarán los 35 millones de Petabytes, lo que representaría un crecimiento del 1,845% (ver IMAGEN 6,

donde se define, para un momento preciso de nuestra vida, la cantidad de información que los individuos generamos en la actualidad).

IMAGEN 6. Fuente: Adaptado de http://humanfaceofbigdata.com/

- Google procesa 20 Petabytes de datos diariamente y el CERN en Ginebra genera 40 Terabytes por segundo. Pero no sólo estos gigantes manejan estas cantidades de información, según el MGI, en 15 de los 17 sectores económicos analizados, la empresa americana media de más de 1.000 empleados almacena más de 235 Terabytes diarios.

Además, Feller et al. (2011) señalan que el almacenamiento de toda la información que fluye en Internet es un negocio que está creciendo de manera exponencial, el cual está conformado por grandes centros de datos que actúan como 'hoteles' para los servidores que gestionan Internet. Google tiene más de 30 centros de datos, que contienen más de 1.000.000 de servidores de última generación. Para alcanzar el despliegue global de Google, Microsoft está invirtiendo miles de millones de dólares en añadir hasta 20.000 servidores al mes en sus soluciones tecnológicas. Diversos factores, como el inmenso calor que desprenden y el consumo de energía que requieren, han obligado a situar estos centros de datos en los lugares más remotos del mundo (se espera que en el año 2020, el consumo de estos centros equivalga al consumo actual de electricidad de Alemania, Canadá y Brasil juntos).

En definitiva, tal y como señala Manyika et al. (2011), el aumento del volumen y del detalle de la información gestionada por las organizaciones, el incremento en contenido multimedia, las redes sociales y el IoE impulsarán el crecimiento exponencial de los datos en un futuro inmediato.

Algunos de los autores consultados como Cantero y Alonso (2012), Schroeck et al. (2013), Singh (2013) y Tamaki (2014), llegan a caracterizar, en principio, tres dimensiones en Big Data (’las tres Vs’): ‘*Volumen*’, ‘*Variedad*’ y ‘*Velocidad*’, para en los últimos artículos y documentación consultada, Buchholtz et al. (2014) y Kaisler, Armour, Money y Espinosa (2015), añadan dos más: la ‘*Veracidad’* y el *‘Valor’* (ver IMAGEN 8 e IMAGEN 9). La convergencia de estas dimensiones ayuda tanto a definir, como a diferenciar, esta realidad terminológica:

1. **Volumen**: Hace referencia a la existencia de cantidades masivas de datos. Sea lo que se considere en este preciso momento como un volumen grande, mañana lo será más. Es la característica que se asocia con mayor frecuencia a Big Data y las organizaciones intentan aprovecharla para mejorar la toma de decisiones en su análisis.
2. **Variedad**: Se refiere a los distintos tipos de datos y sus diversos orígenes. La cuestión es: cómo se han de integrar, gestionar y analizar datos estructurados, semiestructurados y no estructurados. Y es que con la irrupción de sensores, dispositivos inteligentes y tecnologías de colaboración social, los registros que se generan presentan innumerables formas: archivos de texto, bases de datos, geolocalizaciones, URL´s, tuits, registros de sensores, audios, vídeos, secuencias de clic del ratón, archivos de registro y un largo etcétera cuya variedad sigue aumentando cada día (ver IMAGEN 7).

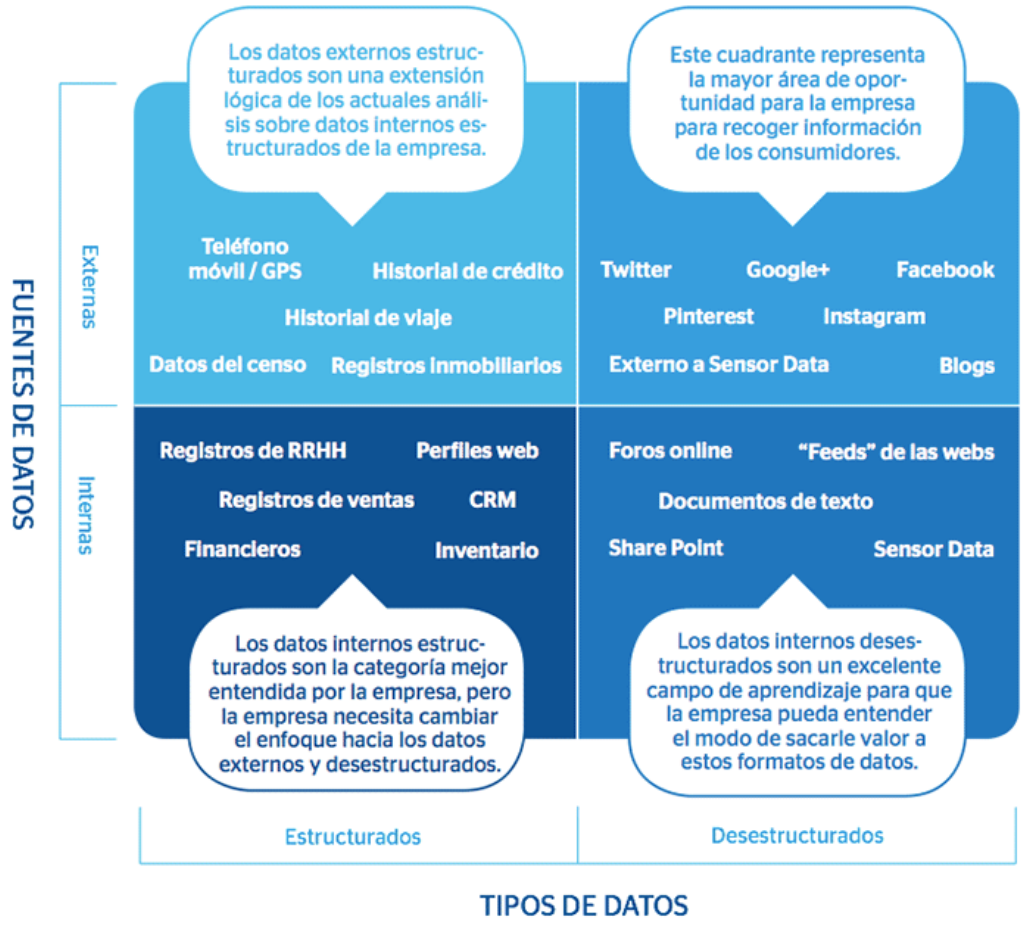

IMAGEN 7. Las fuentes de datos son internas y externas y los tipos de datos son estructurados y desestructurados. Fuente: https://www.centrodeinnovacionbbva.com

3. **Velocidad**: Hace referencia a cómo de rápido se crean y se procesan los datos. La velocidad está relacionada con la 'latencia', es decir: el tiempo de espera entre el momento en el que se crean estos registros, el momento en el cual se captan y el momento en el que son accesibles para su análisis. Esta velocidad está aumentando continuamente, lo que hace que los sistemas tradicionales no sean eficaces en su recogida, almacenamiento y análisis. Para los procesos en los que el tiempo resulta fundamental, ciertos tipos de datos deben analizarse en tiempo real (streaming) para que resulten útiles para el objetivo.
4. **Veracidad**: la incertidumbre de los datos. Si bien, desde un punto de vista tradicional (dentro de lo tradicional que puede llegar a ser algo tan novedoso) las tres dimensiones anteriores engloban los principales atributos de Big Data, es necesario tener en cuenta esta cuarta e importante realidad. La '*Veracidad*' ha sido junto al '*Valor*', los dos últimos atributos en incorporarse y lo han hecho cuando Big Data ha trascendido a otras áreas distintas de la ciencia

informática. En contraste con los 3 Vs originales, no está referida a sus características intrínsecas, sino más bien a la calidad que es necesaria para hacerla útil en su aplicación práctica (Buchholtz et al., 2014). Este término pone de relieve la importancia de abordar y gestionar la incertidumbre inherente a algunos tipos de datos y hace referencia al nivel de fiabilidad asociado a éstos (ya que ignorar esta incertidumbre puede generar incluso más problemas que su propio tratamiento). Esforzarse por conseguir inicialmente unos datos de alta calidad es un requisito importante y un reto fundamental, aunque partamos de la premisa de que los mejores métodos de 'limpieza' no pueden eliminar toda la imprevisibilidad inherente para algunos tipos de registros (como los relacionados con el clima, la economía o las futuras decisiones de compra de un cliente). No puede haber datos que cumplan los tres criterios originales, pero sean inaplicables en la práctica debido a su mala calidad o a una baja credibilidad de sus fuentes; pero, al mismo tiempo, los 3 primeras Vs hacen que esta cuarta (la '*Veracidad*') sea más fácil de lograr ya que cuando hay gran un número mediciones independientes los errores de medición ordinarios se convierten en un problema menor, ya que éstos tienden a estabilizarse y podemos gestionar su correcto tratamiento utilizando para ello técnicas estadísticas que aumenten la robustez de los datos. Por lo tanto, la clave para lograr la veracidad en Big Data no conlleva garantizar la medición perfecta, sino, más bien, requiere evitar los errores sistemáticos y controlar la confiabilidad de sus fuentes y aquí la Estadística Multivariante juega un papel fundamental.

5. **Valor**: Representa el resultado económico y social del desarrollo y la implantación del Big Data. Resultaría ya muy complicado en la actualidad realizar un análisis detallado de cómo el uso de este fenómeno aporta valor a aquellas organizaciones que lo manejan, requiriendo una monografía específica para definir cada aspecto en cada entorno implicado, aunque a continuación expondremos brevemente para los ámbitos empresarial y gubernamental algunos ejemplos representativos que hemos encontrado en la bibliografía consultada. De esta manera, desde un punto de vista empresarial Manyika et al. (2011), señalan que hay 5 grandes formas en las que el uso del Big Data puede crea valor:

- Recopilando los datos de una manera más exacta y detallada.
- Posibilitando realizar análisis de clientes, productos y servicios más detallados, precisos y a medida.
- Haciendo la información más transparente y utilizable.
- Mejorando sustancialmente el proceso de la toma de decisiones.
- Optimizando el desarrollo de la próxima generación de productos y servicios.

Desde el punto de vista de la Administración Pública, Valero (2013) señala que son numerosos los proyectos que ya se han puesto en marcha y que, por ejemplo, han permitido plantear políticas públicas mejor enfocadas en sectores claves como la Sanidad y las incipientes *'Smart Cities'* incrementando notablemente la eficacia y eficiencia de los sistemas de gestión de los recursos y los bienes de dominio público. Para este último aspecto, los tratamientos de la información propios del Big Data ofrecen posibilidades innovadoras respecto a las actuaciones de comprobación que pueden realizar las Administraciones Públicas en diversos sectores, en particular aquellas que se basan en el procesamiento de información en tareas de Inspección, permitiendo realizar comprobaciones masivas de forma automatizada no sólo con sus propias bases de datos o, en su caso, de otras Administraciones Públicas o entidades privadas a través de los distintos convenios que pueden ser establecidos sino, incluso, con la información desestructurada a la que pueda accederse libremente a través de Internet y que no se encuentre protegida con medidas de seguridad adecuadas. Por lo tanto, se convierte en una herramienta de tremendo potencial en este tipo de actividad en sectores como la recaudación de impuestos, el control de las bajas médicas y su incidencia sobre el sistema de prestaciones, la gestión de subsidios por desempleo y la supervisión del otorgamiento de subvenciones.

IMAGEN 8. Las 5Vs en Big Data. Fuente: Adaptado de Gómez (2013).

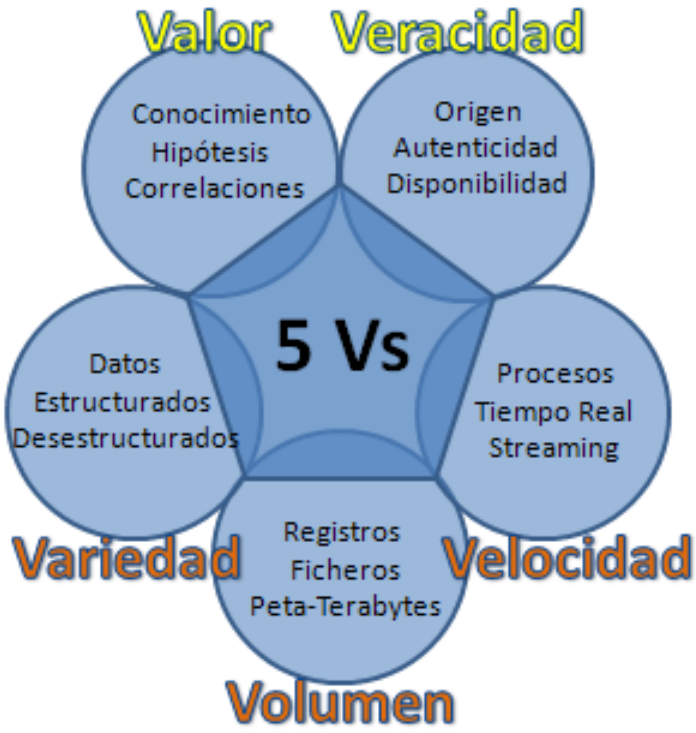

IMAGEN 9. Las 5Vs en Big Data. Fuente: Adaptado de http://bit.ly/1wfs7no

7.2 Localizando Big Data

Ya hemos comentado anteriormente que, hasta hace muy pocos años, las fuentes de datos disponibles solían ser escasas, poco diversificadas y conllevaban un coste económico muy alto. Además, era muy difícil recopilar y mantener los datos, requiriendo también un esfuerzo adicional su estructuración (aspecto imprescindible para poder aplicar sobre ellos técnicas tradicionales de análisis) y este esfuerzo se producía, incluso, aunque esos registros se hubieran generado como producto de la propia actividad de una organización. Pero con el desarrollo de las nuevas tecnologías TIC, esta realidad ha comenzado a cambiar al incorporarse nuevos conjuntos de técnicas desarrolladas ex profeso para recopilar, almacenar y analizar esta información y que comienzan a permitir a las organizaciones realizar todas estas tareas con un menor coste y esfuerzo que, además, junto a la creciente apertura de los datos (el Open Data), conforman un escenario inimaginable hace tan solo un par de años. En este sentido Buchholtz et al. (2014) postulan que las principales fuentes de Big Data son gratuitas y que, en la actualidad, el valor de los datos depende no sólo de su cantidad y calidad, sino también de su interconexión con otros conjuntos de registros, ya que la combinación de varias fuentes puede dar lugar a nueva información que aumenta el valor de cada origen en particular (un ejemplo de este proceso de superposición de tecnología que crea nuevos y valiosos recursos de este tipo es el IoT). A continuación, se detallan cuáles son esas fuentes de Big Data y, por lo tanto, dónde se puede focalizar el esfuerzo para obtener su contenido:

- *Internet*: Es la principal fuente y flujo de los datos creados y donde, además, es posible capturar las acciones que un usuario realiza para obtener información útil: desde crear estadísticas tan simples que permiten analizar qué páginas visita, hasta poder descubrir sofisticados patrones del uso que hace de los recursos disponibles en la red (páginas web, libros electrónicos, música, videos) a través de procesos robóticos de monitorización de la actividad individual, que permiten a las empresas ver qué buscan los usuarios (por ejemplo, cuánto tiempo pasan en diferentes secciones de una página web, en qué parte de un video abandonan su visualización, qué capítulos de un libro vuelven a leer, etc.) Pero ésta no es la única manera de recopilar datos ya que, en la era de Internet 2.0 y de las redes sociales, han surgido nuevos y grandes flujos de datos proporcionados por los propios usuarios, donde ya es posible obtener el registro de la interacción entre esos individuos y su comportamiento como grupo para ampliar aún más la información que puede obtenerse de manera directa e indirecta.

- *Las TIC*: Su implantación en el sector público y privado es otro elemento que posibilita la gran revolución de los datos. Un ejemplo claro de esta realidad es su uso en el *sector financiero*: la 'digitalización' del dinero ha permitido, por ejemplo, el registro de miles de millones de transacciones, el análisis en tiempo real de los mercados y la creación de líneas de crédito absolutamente personalizadas. Otro sector donde su integración ha provocado la generación de montañas de datos son los *servicios de salud* donde los flujos de registros generados son enormes y se producen en tiempo real y en todas las formas posibles (el texto de un diagnóstico, las recetas electrónicas, las imágenes y los videos de distintas técnicas de exploración, etc.) También es un fenómeno que implica a los *gobiernos*, ya que las autoridades públicas han manejado grandes cantidades de datos mucho antes de la era de la digitalización (los censos, los procedimientos de gestión y recaudación de impuestos…) y el uso de las TIC les ha ayudado a manejarlos y usarlos para mejorar y optimizar sus recursos. Y, finalmente, su implantación masiva en el *sector empresarial* donde ya es casi inconcebible una empresa no informatizada.
- *Los Sensores*: La tercera fuente de Big Data está relacionada con este tipo de dispositivos que conforman el ya analizado IoT. Y es que los avances tecnológicos ha vuelto ubicuos los registros, han facilitado su recogida a través de la implantación masiva de redes inalámbricas y han posibilitado su almacenamiento y gestión de una manera ágil y a bajo coste gracias al '*Cloud Computing*'. De esta manera, la medición detallada y continua de un producto es posible no sólo durante su proceso de producción, sino también durante toda su 'vida útil', y la cantidad de datos que se obtienen en tiempo real para ser analizados es enorme y continuará creciendo a medida que se vayan implantando en todos y cada uno de los elementos que un investigador o una organización deseen considerar.
- *La Investigación Científica*: Es la cuarta fuente del Big Data y hasta el momento la menos mencionada por los autores: la *'Big Ciencia'*. Los grandes proyectos científicos con presupuestos de billones de euros en áreas como la Medicina, la Genética, la Física de Partículas, la Astronomía o la Neurobiología generan volúmenes enormes de datos (ejemplos tan conocidos como el CERN que fue galardonado con el último premio Príncipe de Asturias (año 2013) o el '*Human Brain Project*', son algunos de los mejores ejemplos de lo señalado). A través de ellos, los científicos pretenden explotar todo su potencial con el objeto de que revierta directamente como un gran beneficio en la sociedad a través de un proceso de mejora contínua en la productividad de la propia investigación, ya que la digitalización de los resultados de las investigaciones y su publicación en Internet en tiempo real, posibilitan a otros investigadores hacer uso de ese enorme

conjunto de conocimientos y observaciones relevantes que permiten combinar ideas de diferentes ámbitos científicos, y esta convergencia es posible en gran medida por las tecnologías del Big Data de la cual, además, los gobiernos y las empresas podrán aprovechar toda la experiencia acumulada por la '*Big Ciencia*' para mejorar la gestión realizada sobre sus propios datos.

7.3 Las Consecuencias del Big Data

Big Data se ha convertido ya en un elemento transformador que afecta a todas las esferas de nuestra sociedad. Este fenómeno está cambiando la forma en cómo vivimos y trabajamos y la manera en la cual se relacionan los gobiernos, las empresas y las personas. Y éste es sólo el inicio de la revolución.

Según Buchholtz et al. (2014) existen tres maneras en que los registros provenientes de Big Data pueden convertirse en un valor añadido para cualquier organización y se convierten en el escenario idóneo de evolución en su implantación como parte fundamental de su 'modelo de negocio':

- *Del Dato a la Información*: Se convierte en la primera escala donde Big Data puede aportar valor. Los datos se extraen mediante distintas técnicas y son analizados 'ad hoc' para crear información. Es especialmente importante para aquellas organizaciones que requieren modelos de gestión donde la información es su principal fuente de negocio (aseguradoras, financieras, e-commerce, etc.)
- *Del Dato al Producto/Proceso*: El efecto se produce cuando el análisis de datos realizado necesita ser trasladado 'al mundo físico' para aportar valor a una organización (automoción, empresas de manufactura, etc.)
- *Del Dato a la Gestión*: trasladar el contenido útil de Big Data directamente al valor de su negocio constituye el mayor desafío para las organizaciones. Big Data debe aportar información basada en los datos de manera sistemática al proceso de toma de decisiones de una organización. La compañía no sólo necesita extraer con éxito esos datos y adaptar esos conocimientos para que tenga efectos directo en sus procesos de gestión, sino que también debe 'ajustar' su tradicional cultura corporativa y pasar de la toma de decisiones realizada de manera tradicional, muchas veces, por 'instinto' e intuición a la toma de decisiones basadas en datos reales y objetivos. Esta oportunidad está abierta a todo tipo de organizaciones, pero se beneficiarán primeramente y principalmente aquellas que presentan un uso intensivo y altas tasas de adopción de las TIC (Administración Pública, Sanidad y grandes multinacionales).

Zhou, Chawla, Jin y Williams (2014) señalan que la aplicación de '*técnicas Big Data*' proporcionará una serie de beneficios que se pueden resumir en tres aspectos:

- *Mejoras en la eficiencia de los recursos*: a través del máximo conocimiento de toda la información relativa a los recursos producto de la producción, la distribución y las actividades comerciales (marketing).
- *Mejoras en los productos y los procesos*: a través de una innovación contínua basada en actividades de I+D, en el monitoreo de procesos y en el feedback con los consumidores.
- *Mejoras en la gestión:* como resultado de la mejora que proporciona su implementación en el proceso de toma de decisiones.

En definitiva, la buena praxis, sobre cada una de las dimensiones de este fenómeno hace que los grandes problemas que se presuponen (y existen) cuando se inicia un proyecto que maneja Big Data, pueden hacer que se conviertan en grandes oportunidades, lo cual se expresa claramente en la IMAGEN 10. DEL '*GRAN PROBLEMA*' AL '*GRAN VALOR*'. FUENTE: ADAPTADO DE

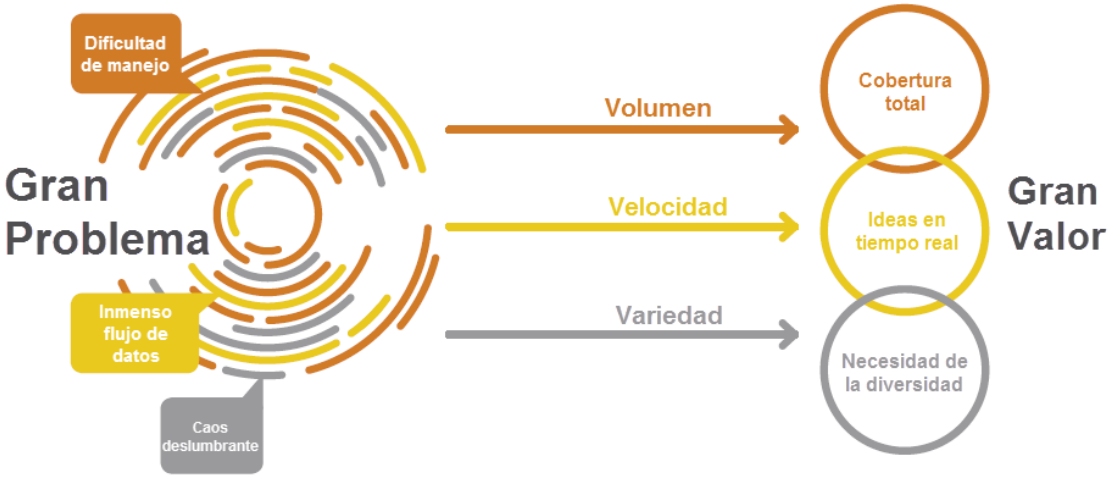

IMAGEN 10. Del '*Gran Problema*' al '*Gran Valor*'. Fuente: Adaptado de Buchholtz et al. (2014).

8 LA ESTADÍSTICA MULTIVARIANTE

Citando a Cuadras (2012) “el Análisis Multivariante es un conjunto de métodos estadísticos y matemáticos, destinados a describir e interpretar los datos que provienen de la observación de varias variables estadísticas, estudiadas conjuntamente”.

La información multivariante se recoge en matrices de datos $n \times p$ siendo, generalmente, n el número de individuos y p el número de variables (algunos métodos estadísticos trabajan con otro tipo de matrices: distancias, disimilaridades, etc.) De esta manera, la matriz de datos multivariantes X se puede definir de la siguiente manera:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Donde, de forma genérica, las filas de la matriz se identifican con los individuos y las columnas con las variables.

Como norma general es conveniente, antes de realizar cualquier tipo de análisis multivariante, examinar y revisar los datos. Para ello se utiliza la técnica conocida como **Análisis Exploratorio de Datos** (‘EDA’, acrónimo de la denominación original en inglés ‘*Exploratory Data Analysis*’) también conocida como **Estadística Descriptiva**. Desarrollada por Tukey (1977) padre de esta técnica y del software estadístico moderno, el autor la definió como *“el conjunto de herramientas gráficas y descriptivas utilizadas para el descubrimiento de patrones de comportamiento en los datos y el establecimiento de hipótesis con la menor estructura posible”*. El ‘EDA’ según Tukey (1993) se realiza sin ninguna hipótesis a priori, utilizando técnicas muy sencillas, donde abundan las representaciones gráficas, con el objetivo de:

- Familiarizarse con la naturaleza de los datos y analizar las principales características de la distribución de las variables.
- Sugerir y evaluar hipótesis sobre las causas de los fenómenos observados (en las cuales basaremos la inferencia estadística).
- Apoyo en la selección de las técnicas y herramientas estadísticas apropiadas.

- Extraer las variables más importantes.
- Detectar los valores atípicos (‘*outliers*’) y las anomalías.

Una vez realizado este análisis, el investigador tiene que plantearse qué metodología ha de utilizar en relación al tipo de datos que maneja. Existe, dentro de la Estadística Multivariante, un conjunto de técnicas que se pueden aplicar sobre los datos a analizar y éstas dependen, en gran parte, de la naturaleza de las observaciones que vayamos a realizar. Así, la decisión final depende básicamente de:

- El objetivo principal del problema a resolver.
- La estructura del conjunto de datos.

Estas herramientas de análisis que actualmente tienen su marco de aplicación en todos los campos científicos, comenzaron desarrollándose para resolver problemas de clasificación en Biología, se extendieron para encontrar variables indicadoras y factores en Psicometría, Marketing y otras Ciencias Sociales y han alcanzado un marco muy amplio de aplicación en Ingeniería y Ciencias de la Computación como herramientas para resumir la información y diseñar sistemas de clasificación automática y de reconocimiento de patrones (Peña, 2002).

Para el presente TFM nos centraremos en el estudio y revisión Bibliográfica de los denominados **Métodos Multivariantes Clásicos** para, en el futuro Trabajo de Tesis Doctoral, ampliar el mismo a los métodos **Biplot**, las denominadas **Técnicas de Análisis de Tablas de Tres Vías** y los **Modelos con Variable Respuesta** las cuales, a priori, parecen tener una capacidad mayor de implantación y desarrollo en la realidad del Big Data dadas sus características.

8.1 La Estadística Multivariante en Big Data

Hasta la irrupción del Big Data, el análisis multivariante estaba relacionado con estudios científicos en los cuales la excepcionalidad de los mismos les confería un halo de unicidad, pero este fenómeno plantea retos que requieren la aplicación continua de métodos estadísticos de este tipo junto a técnicas de ‘*Machine Learning*’ (*ML*) e Ingeniería Informática. De esta manera, la Estadística Multivariante ha de situarse en una posición dominante dentro del análisis para todas las áreas científicas y ámbitos socio-económicos. Y es que, por ejemplo, los sondeos estadísticos a gran escala generan mucha más información de la que puede ser asimilada por la persona que interpreta los resultados. Con las medidas de resumen clásicas como tablas de frecuencias, coeficientes de correlación, etc. es muy difícil encontrar patrones en las

interrelaciones entre las variables, especialmente si su número es muy elevado. La cantidad y complejidad de los datos generados en la investigación científica contemporánea continúa creciendo rápidamente siguiendo la Ley de Moore, tal y como hemos señalado anteriormente. Desde los dominios de la Genómica, pasando por la Climatología, el Marketing, la Ciencia Computacional y un largo etcétera, los estadísticos ya son participantes activos dentro de equipos de investigación multidisciplinares. Así, mientras que en algunas áreas los procesos automatizados recogen y procesan enormes cantidades de datos; en otras se diseñan simulaciones de sistemas complejos para generar información sobre comportamientos de modelos a gran escala y, en el ámbito general, los datos son el producto intrínseco de todo lo concerniente a la era de la información en la cual nos encontramos (March, 2008). En la actualidad en muchos de estos dominios, comienza a ser considerable la actividad científica que necesita implantar soluciones tradicionales o desarrollar nuevas ideas estadísticas, métodos y software que den respuestas a las preguntas que se plantean desde las mismas, como ya hemos comentado en el apartado introductorio de este capítulo.

La historia de la Estadística ha demostrado en multitud de ocasiones (tal y como hemos comprobado también a lo largo del Máster), que una técnica concreta y su teoría pueden desarrollarse en el ámbito de una disciplina específica y para unos datos muy determinados y posteriormente ser 'recogida' para su aplicación en otros ámbitos muy distintos del original. Ejemplos muy conocidos son el '*Análisis de Cluster*', el '*Multidimensional Scaling*' y el '*Análisis de Componentes Principales*' (por no citar cada una de las técnicas vistas durante el desarrollo del curso). Por lo tanto, existe también la gran oportunidad de avanzar en el estudio de las interrelaciones resultantes de la investigación estadística dentro del ámbito científico con el objeto de crear o reutilizar teorías, herramientas y métodos relevantes que sean útiles en múltiples dominios de la investigación científica siendo también necesario instituir, para ello, un cuerpo o núcleo académico que dé respuesta a esta realidad.

A continuación definimos a través de dos breves reseñas históricas dentro de la ciencia estadística dos hitos en su desarrollo que permiten su enlace lógico con las necesidades presentes (¿y futuras?) del Big Data.

8.1.1 ¿La necesidad de un nuevo enfoque dentro de la Estadística?

Ratner (2012) señala acertadamente que ya Tukey (1962) expresó su preocupación por que la ciencia estadística no avanzaba de manera conveniente. Advertía que los estadísticos mostraban demasiado énfasis en las matemáticas y no el suficiente en el análisis de los datos y señaló la necesidad de un profundo cambio que desbloqueara la rigidez en su desarrollo disciplinar. En el mismo artículo, Tukey ya se autodefinió como un '*analista de datos*' (y no como un Estadístico) que *"debía actuar como un detective*". Sin embargo, no fue hasta la publicación de su obra '*Exploratory Data Analysis*' (Tukey, 1977) donde el propio autor condujo esta disciplina de los rigores de la férrea inferencia estadística a un nuevo área conocido como 'EDA' (de la que ya hemos señalado en el punto introductorio de este capítulo algunas de sus características), la cual ofrecía un nuevo enfoque libre de hipótesis, contemplando pruebas no paramétricas que resolvieran problemas en los que el análisis es guiado por los propios datos de manera iterativa, probando y modificando el mismo como resultado de la evaluación por retroalimentación con el objeto de obtener los resultados finales más 'confiables'. Y es que 'EDA' incluye las siguientes características que han de extrapolarse al análisis multivariante en Big Data:

- 'Flexibilidad'.
- 'Practicidad'.
- 'Innovación'.
- 'Universalidad'.
- 'Simplicidad'.

Resulta necesario, e importante, mencionar que Tukey nunca habló específicamente de Big Data; sin embargo, él como impulsor de la necesidad de la informática en la Estadística ya predijo la necesidad creciente de la misma para esta ciencia y que sus costes económicos y operativos derivados irían decreciendo con el tiempo a medida que iban aumentando sus necesidades por la avalancha de información que una sociedad informatizada provocaría en un futuro próximo.

8.1.2 Integración de la Estadística en el ámbito computacional moderno

Samuel (1967) define el término '*Machine Learning*' ('ML') como el campo de estudio que pretende desarrollar en los ordenadores la habilidad de aprender sin ser explícitamente programados para ello (*aprendizaje automático*). En otras palabras, 'ML' investiga las formas en que las computadoras pueden adquirir conocimientos a

partir de los datos y así aprender a resolver problemas. Morgan y Sonquist (1963) dieron lugar a una auténtica revolución sobre los supuestos restrictivos de la estadística clásica al desarrollar el método 'AID' ('*Automatic Interaction Detection*') que es una técnica de segmentación que se sirve de la informática para encontrar y aprender de patrones multidimensionales y de relaciones entre los datos (detectando la presencia de interacción en un modelo de predicción), constituyendo una alternativa no paramétrica muy válida a los tradicionales análisis de regresión desarrollados para la predicción y la clasificación de los objetos. Muchos autores opinan que 'AID' marcó el comienzo del enfoque 'ML' para resolver problemas estadísticos y desde su desarrollo inicial se han realizado múltiples mejoras y ampliaciones del algoritmo inicial que incluso hemos analizado y estudiado detenidamente en el Máster (como el CHAID para el cual se han realizado importantes aportaciones desde el Departamento de Estadística de la Universidad de Salamanca), siendo actualmente una de las técnicas más utilizadas en Data Mining por su accesibilidad y eficacia. Por lo tanto, podemos considerar el conjunto de estas técnicas como el primer estadístico 'ML', ya que son técnicas informáticas intensivas que necesitan el 'auxilio' del ordenador, una condición necesaria para estos métodos; pero sin embargo, tal y como señala Ratner (2012) no llegan a ser verdaderos 'ML' porque usan criterios explícitamente estadísticos (test Chi-Cuadrado entre otros) para su aprendizaje.

8.1.3 El presente ahora: el futuro de la Estadística Multivariante

La creciente colaboración que se ha producido en los últimos años entre los estadísticos y los científicos de todas las disciplinas que requieren manejar estos enormes volúmenes de datos debe conducir a importantes desarrollos en todos los campos afectados. Pero junto a estas oportunidades, aparecen nuevos problemas motivados por las características del Big Data (ver capítulo 7) que, por su necesidad de solución y los múltiples ámbitos de aplicación, deben plantearse como retos multidisciplinares (Rajaraman & Ullman, 2011; Ratner, 2012; Warden, 2011; Witten & Frank, 2005):

- *Procesamiento de datos complejos en streaming*: Además de los (obvios) problemas de procesamiento y almacenamiento, deben resolverse los involucrados al manejo de cantidades masivas de datos en tiempo real (decidir qué calcular y qué almacenar en cada momento). Dos frentes abiertos:
    - La '*nube*' puede ser la respuesta ya que los servidores virtuales permiten escalar tanto los recursos como los costes disponibles, lo que resulta

esencial para poder procesar grandes volúmenes de datos con rapidez a través de distintas máquinas clusters.

- ‘ML’: Desarrollo de sistemas que permitan automatizar la toma de decisiones basadas directamente en los datos.

- *Desarrollo de métodos estadísticos más robustos.* Técnicas más tolerantes al estado y ‘confiabilidad’ de los datos y también a los *‘outliers’*. De esta manera:
  - *Análisis de las diversas fuentes de datos*: la mayoría de las Bases de Datos de acceso público más interesantes (internas, externas, estructuradas o no), se encuentran mal organizadas, ‘cargadas de ruido’ y son, normalmente, de difícil acceso a través de los estándares actuales de programación.
  - *La importancia de los ’outliers’*: en marketing (una de las disciplinas con más necesidad de implantar soluciones para Big Data) por ejemplo, no se deben descartar estos registros, ya que coinciden, a menudo, con las cuentas/clientes de mayor (o menor) valor de toda la población.
  - *El desarrollo del ‘Natural Language Processing’* (‘NLP’): un objetivo muy importante es transformar los contenidos textuales generados por los usuarios en información valiosa a través de procesos estadísticos computacionalmente sencillos.
- *Mejora en las técnicas de visualización*: La representación gráfica facilita la comunicación y el entendimiento de los datos, transformándolos directamente en información.

En definitiva, el objetivo y gran reto en Big Data para los sistemas de análisis estadístico y su capital humano asociado, es la transformación de todos esos datos en información relevante y de valor.

8.2 Métodos Estadísticos Multivariantes Clásicos en Big Data

A continuación se detallan las técnicas clásicas multivariantes más utilizadas como herramientas de análisis dentro de los diferentes ámbitos a los cuales se vincula Big Data y que hemos encontrado referidas en la bibliografía consultada. Para ello y contemplando las ideas propuestas por Witten y Frank (2005), Miller (2010), Ohlhorst (2012), Rajaraman y Ullman (2011) y Chen y Zhang (2014) en las cuales definen las necesidades analíticas que requieren de métodos estadísticos para extraer valor de Big Data, hemos realizado la clasificación de estas técnicas en relación a su finalidad,

proponiendo la existencia de dos grandes grupos que coinciden con los objetivos iniciales planteados, para este fenómeno, por estos autores:

- **Técnicas de Reducción de la Dimensión**: cuyo objetivo es simplificar los datos resumiendo la información de los mismos a través de un número pequeño de componentes que presenten la información más relevante. Dentro de ellas, la bibliografía consultada nos señala como técnicas más utilizadas:
    - 'Análisis de Componentes Principales'.
    - 'Análisis Factorial'.
    - 'Análisis de Coordenadas Principales'.
    - 'Multidimensional Scaling'.
    - 'Análisis de Correspondencias'.
- **Técnicas de Clasificación**: cuya finalidad es agrupar y clasificar los datos mediante la división adecuada de éstos y la aplicación de estas normas a nuevos conjuntos de registros. Las metodologías relacionadas con este grupo de técnicas son:
    - 'Análisis de Cluster'.
    - 'Análisis Discriminante'.
    - 'Análisis de Correlación Canónica'.

Comentar finalmente, que la fuente para la elaboración de cada una de las descripciones realizadas para las técnicas revisadas son las propias del '*Máster en Análisis Avanzado de Datos Multivariantes*' que he cursado y, si ha resultado necesaria la ampliación conceptual para alguna de ellas, se ha efectuado siguiendo la bibliografía de consulta proporcionada en el mismo.

8.2.1 Técnicas de Reducción de la Dimensión

Su objetivo general es identificar un conjunto de variables latentes o hipotéticas, resultado de la combinación de las variables iniciales, que recojan la mayor parte de la información original y que permitan, por lo tanto, reducir el número de dimensiones original para facilitar su análisis, interpretación y representación. Los métodos basados en variables latentes pretenden reducir la dimensionalidad del conjunto de datos a dos o tres dimensiones manteniendo la mayor parte de la información posible. Esto puede hacerse debido a que muchas de las cuestiones planteadas en los estudios o muchas de las variables medidas en una investigación son, en realidad, aspectos de la misma característica básica. El propósito general, por lo tanto, es explicar las relaciones entre las variables manifiestas a través de un conjunto de variables latentes de forma que,

dadas éstas las otras sean independientes (desde un punto de vista más intuitivo se trata de extraer qué es lo que tienen en común las variables manifiestas y resumirlo en un conjunto reducido de variables hipotéticas).

Esta reducción de las variables originales a unos pocos factores se utiliza, en muchos casos analizados en Big Data, como paso previo a la aplicación de otros análisis posteriores; por ejemplo, un diagrama de dispersión de las primeras componentes con el objeto de encontrar 'clusters' para contrastar similitudes o diferencias entre los individuos.

8.2.1.1 Análisis de Componentes Principales (PCA)

El **Análisis de Componentes Principales** ('PCA') es una técnica descriptiva (en principio no inferencial) que trabaja con variables cuantitativas continuas y no considera la diferencia, ni la existencia, de variables dependientes e independientes. Su objetivo es la reducción de la dimensión de los datos iniciales con el objeto de encontrar transformaciones ortogonales de las variables originales para conseguir un nuevo conjunto de variables incorreladas e independientes denominadas Componentes Principales ('CP'), que se obtienen en orden decreciente de importancia y que son aquellos ejes de los planos que absorben la máxima cantidad de información de los datos (variabilidad). La solución que ofrece este método es única.

La obtención de las 'CP' puede realizarse por varios métodos:

- Buscando aquella combinación lineal de las variables que maximiza la variabilidad (varianza máxima). Las componentes principales son los ejes debidos a la combinación lineal de las variables originales (Hotelling, 1933). Éste es el método más utilizado y que a continuación de esta enumeración continuamos con su planteamiento y discusión.
- Método de los Mínimos Cuadrados (Pearson, 1901): Buscando el subespacio de mejor ajuste al minimizar la suma de los cuadrados de las distancias de cada punto en el subespacio de baja dimensión.
- Método de las Coordenadas Principales (Gower, 1966): Minimizando la discrepancia de las distancias euclídeas entre los puntos calculadas en el espacio original y en el subespacio de baja dimensión.
- Mediante regresiones alternadas utilizando para ello métodos Biplot (Gower, Lubbe, & Roux, 2011). De manera genérica señalaremos que los Biplot son un conjunto de técnicas que permite la representación en un mismo gráfico de las filas (individuos) y las columnas (variables) de una matriz de datos con la máxima

calidad de representación. En el Trabajo de la Tesis Doctoral explicaremos de manera profusa las características de los Biplots vinculados a su utilización en Big Data.

Siguiendo a Hotelling (1933), una vez centrados los datos originales, las 'CP' se obtienen por la descomposición en Valores y Vectores Propios de la matriz de varianzas/covarianzas de los datos (matriz cuadrada y simétrica que contiene en la diagonal las varianzas y fuera de ésta las covarianzas entre las variables, la cual mide la dependencia lineal entre las variables, y que llamaremos a partir de ahora, para simplificar, matriz de covarianzas). Es una técnica que concede más importancia a los individuos que otras técnicas de reducción de la dimensión. Una vez obtenidas las 'CP' calculamos la proyección de las observaciones en ese espacio de dimensión reducida (con pérdida de información mínima) para observar su comportamiento y analizar posibles agrupaciones que los caractericen y categoricen.

Dentro de la bibliografía consultada hemos encontrado su aplicación para el Big Data en Microbiología (Ikeda et al., 2013) como método inicial de reducción dimensional implementado en un programa de mapeo de enzimas metabólicas (tomando como referencia una importante base de datos de uso científico ya estandarizado de secuencias de péptidos) y en análisis de series temporales (Wang, Han, & Liu, 2013). Combinado con el *Análisis Discriminante Lineal (LDA)* (que desarrollaremos en el punto 8.2.2.2 del TFM), en la clasificación de superficies de agua en relación a su composición geoquímica (Filella, Pomian-Srzednicki, & Nirel, 2014), en el desarrollo de sistemas automáticos de reconocimiento del habla (Viszlay, Pleva, & Juhàr, 2011), en la creación de algoritmos vinculados a la Genética (Sabatier & Reynès, 2008) y en Farmacogenómica (Fan & Liu, 2013). Finalmente combinado con el método de *Regresión de Mínimos Cuadrados Parciales ('PLS')* la encontramos desarrollada en varios artículos relacionados con la Espectroscopia (Gallello, Kuligowski, Pastor, Diez, & Bernabéu, 2013; Kemsley, 1996; Wold & Sjöström, 1998); con la Genética, buscando el tamaño ideal de una raza de bovino productora de carne (Arango & Van Vleck, 2002), en el análisis de las transacciones de pago mediante tarjetas de crédito (Laukaitis, 2008) y en la optimización de sistemas de "*Business Intelligent*" (Işık, Jones, & Sidorova, 2013).

Los principales puntos fuertes del 'PCA' en Big Data son:

- Técnica de cálculo computacional muy sencillo y, por lo tanto, muy popular e implementada en multitud de paquetes estadísticos, lo cual ha hecho que sea una de las más utilizadas en este ámbito multidisciplinar.
- Permite la reconstrucción de la información original a partir de la información reducida. Un ejemplo práctico de su aplicación es el restablecimiento de imágenes en espectrometría e imágenes ráster provenientes de cualquier fuente de generación de imágenes (cámaras, satélites, etc.), tal y como hemos visto en la bibliografía comentada anteriormente.

Las principales limitaciones que encuentran los investigadores con esta técnica son las siguientes:

- Cuando las escalas de las variables son muy diferentes (lo cual en Big Data es casi la norma) da lugar a que la variabilidad dominadora sea aquella relacionada con las variables de mayor magnitud. Para evitarlo se realiza su estandarización, pero el problema es que, con ésta, se pierde una información que puede llegar a ser necesaria por su relevancia.
- Problemas computacionales importantes para estimar la matriz de covarianzas cuando la dimensión de los datos es muy grande (propio también de Big Data). Por ello los investigadores proponen diferentes modificaciones para ajustarla a sus datos, la más utilizada es la técnica denominada '*Sparce Principal Component Analysis*' ('SPCA') que Zou, Hastie y Tibshirani (2006) desarrollaron en el ámbito de la Genómica para resolver el problema que se planteaba con el uso del 'PCA' en el caso de la existencia de un número de variables (genes) muchísimo mayor que el número de observaciones (individuos). Su aplicación actual abarca ámbitos muy diversos como el análisis de series temporales (Wang et al., 2013), la inteligencia artificial (d'Aspremont, Bach, & Ghaoui, 2008), la necesidad de mejoras en los tiempos computacionales (Shen & Huang, 2008), etc.
- Técnica no adecuada para datos mixtos (discretos, continuos…). Una de las características más importantes del Big Data es, tal y como hemos analizado, la variedad en los datos generados desde fuentes muy diversas.

8.2.1.2 Análisis Factorial (FA)

Es una técnica que trabaja con variables cuantitativas continuas y al igual que el 'PCA' no considera la diferencia, ni la existencia, de variables dependientes e independientes. El **Análisis Factorial** ('FA') pretende, de manera genérica, la reducción de la dimensión de los datos con el fin de explicar las relaciones entre las variables a partir de un número menor de factores comunes hipotéticos independientes: nuevas variables con *Comunalidad* grande y *Unicidad* pequeña (donde la *Comunalidad* es la parte de la variabilidad de cada variable que consigue explicar cada factor común y la *Unicidad* es la parte de la variabilidad de cada variable original que no se consigue explicar con esos factores comunes). Partimos de la estandarización y el centrado de los datos originales y según el propósito tenemos dos tipos de análisis:

- *Análisis Factorial Exploratorio ('EFA')*: Examina las dimensiones latentes de las variables. Su objetivo es explicar las relaciones entre las variables manifiestas a través de un conjunto de variables latentes de forma que, dadas éstas las otras sean independientes. Desde un punto de vista más intuitivo se trata de extraer qué es lo que tienen en común las variables manifiestas y explicarlo mediante un conjunto reducido de variables latentes. Dentro del 'EFA' existen distintos métodos de extracción de la matriz factorial, entre ellos:
    - *Método de las Componentes Principales*: La dependencia lineal entre pares de variables se mide a través de la matriz de correlación (matriz cuadrada y simétrica que tiene 1s (unos) en la diagonal principal y fuera de ella los coeficientes de correlación lineal entre pares de variables). Esta técnica calcula los Valores y Vectores Propios directamente de la matriz de correlaciones de las variables originales (despreciando la Unicidad). Los Factores Comunes son las 'q' primeras 'CP'. Este método fue estudiado como técnica descriptiva para la reducción de la dimensión y puede considerarse también como una solución para el problema factorial, de ahí que muchas veces exista confusión entre estas dos técnicas.
    - *Método del Factor Principal*: Es un método iterativo que comienza con la estimación previa de las comunalidades, a continuación calcula la matriz de correlaciones reducidas, se diagonaliza la misma y se estima la matriz factorial. Se calculan de nuevo las comunalidades y si son iguales a las del paso previo se para y se obtiene la solución; si no es así, se vuelve a calcular la matriz de correlaciones y continúa el proceso hasta encontrar la convergencia.

- *Análisis Factorial Confirmatorio ('CFA')*: El Análisis Factorial también se puede plantear en sentido confirmatorio; es decir, estableciendo una estructura factorial de acuerdo con el problema objeto de estudio y seguidamente aceptando o rechazando esta estructura mediante un test de hipótesis. Estos métodos, por lo tanto, permiten realizar inferencia. Su objetivo es sugerir y comprobar hipótesis científicas utilizando para ello la reducción de la dimensión de los datos con el fin de explicar las relaciones entre las variables a partir de un número menor de factores comunes hipotéticos independientes. Las dos metodologías más utilizadas para la extracción de la matriz factorial relacionadas con este objetivo son:
    - Métodos de los Mínimos Cuadrados:
        - *Método de los Mínimos Cuadrados No Ponderados:* Para un número fijo de factores, se genera una matriz de coeficientes que minimiza la suma de las diferencias al cuadrado entre las matrices de correlación observada y la reproducida, eliminando en las diferencias los elementos de la diagonal.
        - *Método de los Mínimos Cuadrados Generalizados*: Permite utilizar el contraste de hipótesis para determinar el número de factores. Al igual que el anterior, minimiza la suma de las diferencias al cuadrado entre las matrices de correlación observada y la reproducida, pero para este caso pondera las correlaciones inversamente por la varianza del factor específico.
    - *Método de Máxima Verosimilitud*: La estimación de la matriz factorial se plantea como un problema de estimación de la matriz de covarianzas. Asumiendo normalidad en los datos, se define una distancia entre la matriz de covarianzas observada y los valores predichos para esta matriz por el modelo del Análisis Factorial. La expresión de dicha distancia se precisa a través de una función y las estimaciones de los pesos factoriales se obtienen minimizando ésta, lo que es equivalente a maximizar la función de verosimilitud del modelo factorial. Su principal inconveniente radica en que al realizarse la optimización de la función de verosimilitud por métodos iterativos, si las variables originales no son normales puede haber problemas de convergencia sobre todo para muestras finitas.

La obtención de la matriz factorial, por aplicación de cualquiera de los métodos que hemos expuesto constituye el primer paso del 'FA'. Normalmente la matriz obtenida define una serie de factores comunes siempre ortogonales (incorrelados) que no son fácilmente identificables y para ello es necesario rotar los factores obtenidos hacia una

configuración interpretable (y es que mientras que en el 'PCA' obtenemos una solución factorial única, en el 'FA' las soluciones son infinitas). Existen diferentes técnicas de rotación para transformar la matriz factorial original con el objeto de obtener la denominada '*Estructura Simple'* para esos factores, donde buscamos conseguir que unas saturaciones sean muy altas a costa de otras, para así destacar la influencia de los factores comunes sobre las variables observables (Cuadras, 2012). Existen dos tipos de rotaciones:

- *Ortogonales*: Los nuevos factores siguen siendo incorrelados (rotaciones '*Quartimax'*, '*Varimax'*, etc.)
- *Oblicuas*: No se respeta la ortogonalidad inicial. Los nuevos factores están relacionados (rotaciones '*Quartimin'*, '*Covarimin'*, '*Equamax'*, etc.)

Finalmente, es necesario señalar que las puntuaciones de los individuos en un 'FA' pueden utilizarse para realizar su representación gráfica (de la misma manera que hacíamos en el 'PCA') y para ello existen dos metodologías:

- *Método basado en los autovectores*: la matriz factorial se calcula a partir de un conjunto de vectores propios (el cálculo de las puntuaciones es sencillo ya que se trata de una simple proyección, como en las 'CP').
- *Método basado en regresiones*: si los factores son incorrelados podemos calcular la puntuación de los individuos a través del cálculo de una función de regresión de las variables originales.

Dentro de la bibliografía analizada hemos encontrado su aplicación en pocos artículos de investigación vinculados al Big Data (son más los libros de consulta que la mencionan como técnica multivariante apropiada), entre ellos: en la definición de los factores de riesgo latentes a partir de las variables que provocan enfermedades coronarias en mujeres que sufren diabetes (Edwards et al., 1994), en la definición de modelos económicos (Bai & Ng, 2008), en el ámbito de la Psicología (Matsunaga, 2010), en el análisis de series temporales multidimensionales (Hallin & Lippi, 2013) y en estudios de ingeniería genética (Kurome et al., 2013).

Los principales puntos fuertes de esta técnica en Big Data son:

- Técnica clásica de reducción de la dimensión implantada en la mayoría de los paquetes estadísticos.
- Algunas soluciones son invariantes con respecto a los cambios de escala.

- Se presenta como una herramienta inicial que luego sirve para trabajar los factores extraídos con otras técnicas estadísticas.

Las principales limitaciones que encuentran los investigadores con esta técnica son:

- El cálculo de las matriz de correlación requiere mucho 'esfuerzo' a nivel de máquina cuando trabajamos con volúmenes grandes de datos
- En modelos complejos, los cálculos que estiman las comunalidades son problemáticos desde un punto de vista computacional.
- Técnica no adecuada para datos mixtos, discretos, continuos… propios de Big Data.

8.2.1.3 Análisis de Coordenadas Principales (PCoA)

Es una técnica descriptiva desarrollada por Gower (1966) que a diferencia del 'PCA' y del 'FA' no sólo trabaja con variables continuas. El objeto del **Análisis de Coordenadas Principales** ('PCoA') es representar un conjunto de individuos u objetos pertenecientes a una población, respecto a unas variables que pueden ser cuantitativas, cualitativas o una combinación de ambas y su objetivo es la reducción de la dimensión de los datos con el fin de interpretar las similaridades o disimilaridades (distancias) entre los individuos de una manera simple. Por lo tanto, trabaja con medidas que analizan esta característica calculadas a partir de la matriz inicial que puede estar formada por:

- *Datos brutos*: Se dispone de la medida de una serie de variables tomada de un conjunto de individuos. Pueden ser continuos, binarios, categóricos, ordinales, mixtos, etc.
- *Distancias o disimilaridades entre pares de objetos*: Se dispone de una matriz simétrica que contiene una medida de la disimilitud entre los pares de objetos. La diagonal principal encierra sólo ceros.
- *Similaridades entre pares de objetos*: Se dispone de una medida de la similitud entre pares de objetos. Las medidas se organizan en una matriz simétrica.
- *Productos escalares entre pares de objetos*: la matriz de covarianzas (con las columnas centradas), o la matriz de correlaciones (con las columnas estandarizadas) son ejemplos de matrices de productos escalares.

Una cuestión importante para esta técnica, es que es posible pasar de un tipo de matriz a otra. Por ejemplo, si tenemos una matriz de similaridades (acotadas entre 0 y 1), es fácil obtener una de disimilaridades y dada una matriz de disimilaridades

(distancias) es posible convertirla en una matriz de productos escalares. Dependiendo del tipo de datos se utilizarán unos coeficientes de cálculo específicos para ello:

- *Cuantitativos*: Medidas de Distancia (Euclídea, Minkowsky, Camberra, etc.)
- *Binarios* (presencia/ausencia): Coeficientes de Asociación o Similaridad (Jaccard, Dice y Sorensen, Tanimoto, etc.)
- *Mixtos*: Coeficiente de Similaridad de Gower ('SG').

Dentro de la bibliografía consultada hemos encontrado su aplicación vinculada al Big Data, sobre todo, como parte inicial de distintos algoritmos que pretenden como objetivo final la clasificación de los individuos observados. Así, Podani (1997) lo incorpora como método de reducción de la dimensión dentro de un algoritmo que tiene la finalidad de analizar la taxonomía de comunidades vegetales; Woodward, Gay y Baird (2013) lo utilizan para establecer similitudes entre las comunidades de macroinvertebrados para diferentes lugares de muestreo; Muller, Glaab, May, Vlassis y Wilmes (2013) como parte de una técnica que pretende la generación de modelos para comunidades microbiales naturales; y Vilhena et al. (2014) en un interesante artículo relacionado con el análisis de citas académicas. Finalmente es necesario señalar que el 'PCoA' forma parte de una interesante técnica desarrollada en el Departamento de Estadística de la Universidad de Salamanca por Vicente-Villardón, Galindo y Blázquez (2006) denominada 'Biplot Logístico' basada en un modelo de respuesta logística, en el cual las coordenadas de los individuos y las variables están calculadas para obtener respuestas de este tipo a lo largo de las dimensiones Biplot para grandes matrices de datos. Esta metodología combina, en un mismo algoritmo, el 'PCoA' y la Regresión Logística ('RL').

Los principales puntos fuertes de esta técnica en Big Data son:

- Técnica clásica muy conocida entre los investigadores que permite reducir la dimensión inicial y que está implantada en muchas soluciones informáticas estadísticas estándar.
- Es una herramienta de extracción factorial. Estos factores posteriormente son utilizados por otras técnicas estadísticas para, principalmente, proceder a la clasificación de los individuos.
- Trabaja con matrices brutas de los datos transformándolas en distancias o con matrices de distancias directamente observadas. Podemos utilizar, por lo tanto, datos de naturaleza diversa, tal y como se produce en Big Data.

Las principales limitaciones que encuentran los investigadores con esta técnica son:

- Esta técnica utiliza la Descomposición en Valores Singulares ('DVS') como método de factorización de la matriz y su cálculo presenta dificultades computacionales cuando el tamaño de los datos es verdaderamente grande.

8.2.1.4 Multidimensional Scaling (MDS)

El **Multidimensional Scaling** ('MDS') o Escalamiento Multidimensional (en castellano) es un conjunto de técnicas de análisis multivariante introducidas por Torgerson (1952) que permiten la representación espacial en un espacio de baja dimensión de un conjunto de elementos o estímulos a partir de medidas de proximidad entre pares de objetos, donde éstas pueden ser:

- *Similaridades* (a valores más altos, objetos más semejantes).
- *Disimilaridades* (a valores más altos, objetos menos semejantes).
- *Preferencias* (se utiliza la técnica '*Unfolding*': Caso especial del 'MDS No Métrico' apropiado para este tipo de datos).

La representación espacial de las proximidades se hace de modo que si dos estímulos son valorados como muy parecidos, se encontrarán a poca distancia uno del otro y viceversa. Su representación en un espacio de baja dimensión facilita la interpretación de las proximidades al mostrarlas de manera visual, en lugar de numéricamente, recogiendo lo esencial de la información original y reduciendo el error existente en los datos; de esta manera, es posible observar la 'estructura oculta' de los datos. Se puede decir, que estas técnicas son una generalización de la idea de las 'CP' cuando en lugar de disponer de una matriz de observaciones por variables se dispone de una matriz de distancias o disimilaridades entre los 'n' elementos de un conjunto. A través de ellas podemos trabajar con cualquier tipo de entidad que necesitemos analizar: Sujetos (normalmente) o Variables y podemos utilizarlas como:

- *Técnicas Exploratorias*: Para ver esa 'estructura oculta' de los datos ya comentada con anterioridad.
- *Técnicas Confirmatorias*: Para contrastar hipótesis estructurales basadas en teorías previas.
- *Técnicas para la Obtención de Evaluaciones Comparativas*: Cuando las bases específicas de comparación no se conocen, o no están definidas, a priori.

Trabaja con matrices de similaridades/disimilaridades y las transforma en distancias. Su objetivo es determinar el número de dimensiones adecuado a los datos, definir su función de representación y graficar la misma utilizando un conjunto de variables ortogonales que permiten localizar los objetos sobre cada dimensión, de manera que las distancias euclídeas entre las coordenadas de los elementos respecto a estas variables sean iguales (o lo más próximas posibles) a las distancias o disimilaridades de la matriz original (Micó, 2012; Peña, 2002). En definitiva, buscamos una configuración de puntos (coordenadas) en un espacio euclídeo k-dimensional adecuado y reducido (normalmente 2 o 3 dimensiones) de manera que las distancias en ese espacio sean lo más parecido a las disimilaridades iniciales.

Podemos definir la existencia de dos categorías dentro de las técnicas del 'MDS':

- *'MDS MÉTRICO'*: Partimos de una matriz inicial de distancias y su función de representación es lineal. Tipos:
    - Modelo de Torgerson (1952). El algoritmo utilizado está basado en el teorema propuesto por Young y Householder (1938) según el cual "una matriz de productos escalares derivada a partir de una matriz de distancias euclidianas (de 'n' puntos en un espacio de 'r' dimensiones) puede descomponerse en el producto de una matriz de coordenadas (de los 'n' puntos en las 'r' dimensiones) y su transpuesta" y elimina la restricción incluso, para el caso de que las distancias no cumplan totalmente las condiciones euclídeas.
    - *Modelo Métrico de Razón*. Desarrollado por Ramsay (1977), es un modelo de tipo confirmatorio basado en la distribución lognormal biparamétrica que permite el análisis de datos de disimilaridad obtenidos en escalas de clasificación dentro de un contexto de máxima verosimilitud. Se desarrolla bajo las hipótesis de que los datos son independientes, positivos y poseen un origen determinado en cero (Vera & González, 1996). Dado que está basado en un comportamiento muy concreto de la distribución de los datos, el modelo está casi exclusivamente orientado al análisis de tipo confirmatorio, utilizándose para ello un gran número de estadísticos de contraste (por ejemplo para confirmar hipótesis sobre el número de dimensiones a retener o para determinar la bondad del tipo de modelo aplicado).
    - *Modelo Métrico de Intervalos*: Carroll y Chang (1970) desarrollaron el algoritmo INDSCAL utilizado para el análisis de las diferencias individuales. Obtienen una representación a partir de varias matrices de distancias

asumiendo que éstas difieren entre sí de forma sistemática y no aleatoria (actualmente las modificaciones realizadas a partir del algoritmo INDSCAL original permiten realizar también 'MDS No Métrico').

- *'MDS NO MÉTRICO'*: Resuelve los mismos problemas que el 'MDS Métrico', pero utilizando sólo la información ordinal contenida en los datos, con lo que se evitan ciertos supuestos sobre la distribución y la variabilidad de los mismos. Micó (2012) señala que el fundamento teórico del MDS No Métrico se basa (igual que en el caso métrico), en ajustar las distancias a las disimilaridades, pero mientras que en éste último estaban relacionadas linealmente con las proximidades, en el No Métrico están relacionadas monotónicamente (esta condición incluye no sólo la función lineal, sino también la exponencial, la potencial, la logarítmica y en general, aquellas monótonamente crecientes o decrecientes). Tipos:
    - Modelo de Shepard (1962): Trabaja con distancias en un espacio euclídeo utilizando sólo la información ordinal contenida en los datos, evitando las restricciones propias del modelo métrico sobre la distribución y la variabilidad de los mismos. Los datos de partida no son magnitudes de razón, sino de rango. Este método trabaja obteniendo una representación de 'n' puntos en n-1 dimensiones, y reduciendo progresivamente esta dimensionalidad. Supuso la generalización del 'MDS' a un amplio número de disciplinas científicas.
    - Modelo de Kruskal (1964): A partir de las investigaciones de Shepard este autor desarrolló un nuevo modelo no métrico que a diferencia del anterior comienza obteniendo la mejor representación posible de los 'n' puntos en el espacio k-dimensional, siendo 'k' mucho más pequeño que 'n' y estando especificado antes del análisis en vez de después. Kruskal, además, introdujo el estadístico 'STRESS' como medida de discrepancia entre los datos de entrada (proximidades) y los datos de salida (distancias de la configuración) para valorar el grado de ajuste del modelo. Actualmente este es el modelo No Métrico más extendido, siendo más conocido como el de Shepard-Kruskal (Micó, 2012).

Otro criterio para la clasificación de los modelos 'MDS', es el número de matrices de entrada que soportan. A los modelos que permiten una sola matriz de entrada se les denomina *modelos de dos vías*, donde la primera vía son las filas y la segunda las columnas (los modelos de Shepard, Kruskal y Torgerson son de este tipo). A los modelos que permiten más una matriz de entrada se les denomina *modelos de tres*

vías, las dos primeras coinciden con lo expuesto anteriormente y la tercera vía es 'el eje de profundidad' (el modelo INDSCAL y el propuesto por Ramsay son de este tipo).

Además, cuando se define un 'MDS' es necesario especificar también el número de modos (que son las distintas entidades que se incluyen en la matriz de entrada). De esta manera, las matrices de entrada de '*1 Modo'* únicamente incluyen una entidad: los *estímulos*, las de '*2 Modos'* incluyen dos entidades: *estímulos* y *sujetos* o *estímulos* y *atributos*. Finalmente las de '*3 Modos'* incluyen tres entidades: *estímulos*, *sujetos* y *atributos* (ver IMAGEN 11).

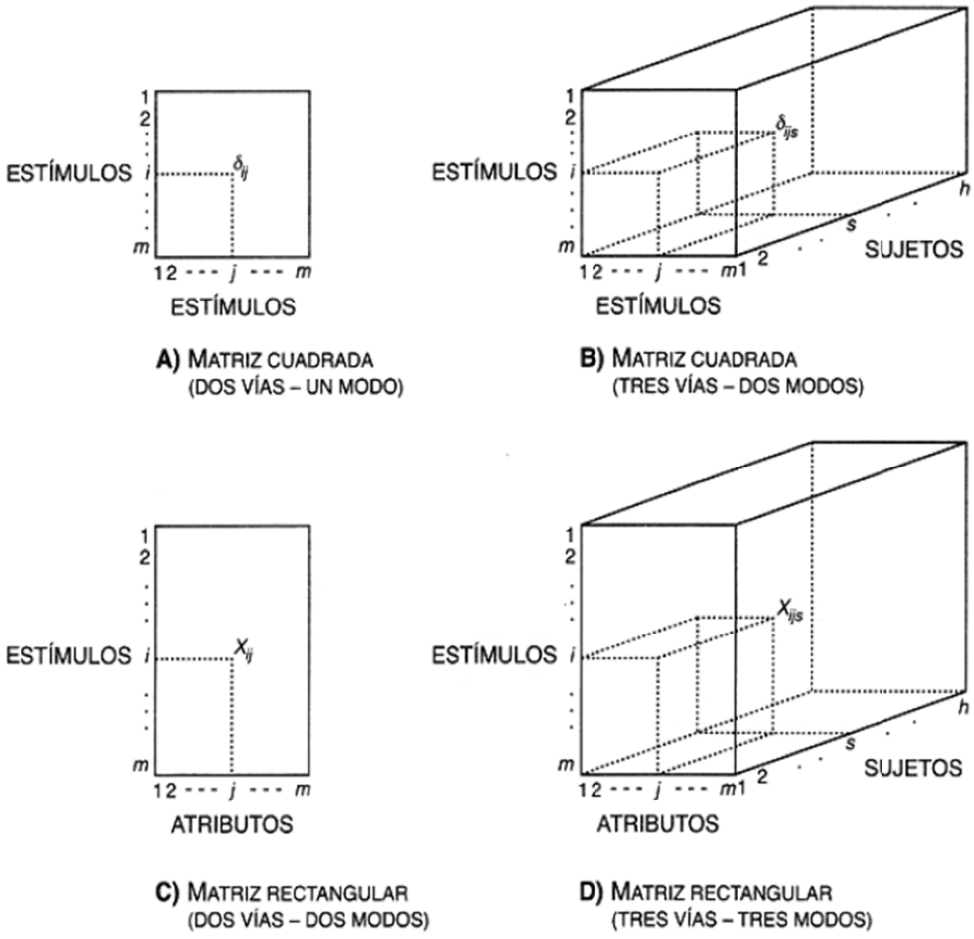

IMAGEN 11. Tipo de matrices de entrada en 'MDS'. Fuente: Martín (2014).

Vinculado a esta técnica, tanto para los Modelos Métricos como para los No Métricos, se han desarrollado distintos algoritmos cuyo objetivo es reproducir los datos en dimensión reducida. Los dos más utilizados en 'MDS' (al estar implementados en programas como SPSS y SAS) dentro del ámbito de la investigación clásica son:

- *El algoritmo ALSCAL* (‘*Alternating Least Squared sCALing*’): desarrollado por Takane, Young y de Leeuw (1977), puede considerarse como el más completo y es, tradicionalmente, el más empleado ya que permite realizar análisis Métricos y No Métricos, siendo sus resultados métricos similares a los obtenidos con INDSCAL en el caso de diferencias individuales.
- *El algoritmo PROXSCAL (‘PROXimity SCALing’):* desarrollado por Commandeur y Heiser (1993), se basa en la utilización simultánea de técnicas de mínimos cuadrados alternados y de mayorización iterativa. Formalmente, el problema que resuelve este algoritmo respecto al resto, es la minimización de la función de pérdidas de mínimos cuadrados o ‘STRESS’ de Kruskal (Montero, Vera, Fernando, Muñoz, & Andrés, 2003).

Dentro de la bibliografía consultada hemos encontrado su aplicación vinculada al Big Data en múltiples disciplinas, ya que su ámbito de empleo es muy amplio dadas sus características (no requiere supuestos de linealidad, ni que las variables sean métricas, ni un tamaño mínimo de muestra, etc.) Entre todos estos campos queremos destacar, por su interés de desarrollo metodológico dentro de la ciencia estadística, los artículos que definen mejoras sobre los algoritmos de la metodología clásica. Así, de Silva y Tenenbaum (2004) desarrollan un modelo computacional eficiente a partir del algoritmo del ‘*Multidimensional Scaling Clásico*’ (‘CMDS’) denominado ‘LANDMARK MDS’ (‘LMDS’), para su uso cuando el conjunto de datos es muy grande y demuestran su eficacia con una serie de ejemplos aplicados. Tzeng, Lu y Li (2008), dentro del ámbito del estudio de microarrays en Genómica, desarrollan otro método denominado ‘SC-MDS’ (‘*Split-and-Combine-MDS*’) que reduce considerablemente (según los autores, 3 veces más rápido) el tiempo de procesamiento en máquina para obtener los mismos resultados que el ‘CMDS’ y plantean su uso para analizar el genoma humano completo. Por su parte Agarwal, Phillips y Venkatasubramanian (2010), dentro del ámbito del desarrollo computacional, proponen el uso de un algoritmo denominado ‘PLACECENTER’ como marco unificado para diferentes tipos de ‘MDS’ (‘CMDS’, ‘MDS Robusto’ y ‘MDS Esférico’) que mejora la calidad de la solución de convergencia respecto a los algoritmos originales que utilizan cada una de estas técnicas por separado. También es necesario mencionar el trabajo de Bae, Qiu y Fox (2012) donde describen un algoritmo de reducción dimensional conocido como ‘SMACOF’ cuya ventaja es minorar las necesidades de memoria de los ordenadores que operan con ‘MDS’ para grandes cantidades de datos. Finalmente, en un muy buen artículo que recoge la evolución de los distintos algoritmos creados, Aflalo y Kimmel (2013), también dentro del ámbito del desarrollo computacional, presentan también un nuevo

método denominado ‘SPECTRAL MDS’ (‘SMDS’) en la misma línea que los anteriores artículos, reduciendo los tiempos en máquina y mejorando los resultados finales para la agrupación obtenida de los individuos.

Podemos señalar las siguientes ventajas de esta técnica:

- Técnica de reducción de dimensiones muy conocida por los investigadores.
- Gran capacidad para adaptarse a las necesidades crecientes que se producen al manejar grandes volúmenes de información y que se procesan a grandes velocidades.
- Permite trabajar con modelos de dos y tres vías y con uno, dos y tres modos.
- Permite crear representaciones espaciales individuales o mapas combinando los grupos de sujetos (dualidad muy interesante en Big Data).
- Admite la posibilidad de combinarse con el Análisis de Cluster. En la investigación empírica se ha demostrado que utilizando estas dos técnicas conjuntamente posibilita descubrir aspectos complementarios muy relevantes de los datos (Gras, 1996).
- Es una técnica que por su naturaleza y bondad matemática puede ser utilizada en muchos campos y disciplinas científicas.

Las desventajas del MDS:

- Los primeros algoritmos se formularon en los años sesenta del siglo XX, pero el ‘MDS’ no se hizo popular hasta su implantación en algunos de los paquetes estadísticos más populares. El problema es que éstos muestran claras limitaciones cuando son utilizados en matrices enormes de datos tal y como hemos señalado en la revisión bibliográfica, lo que hace necesaria su revisión y actualización para adaptarse a las características del Big Data. Aun así, son todavía escasos los trabajos orientados a optimizar su correcta utilización cuando existen observaciones empatadas, o para definir correctamente el número de dimensiones necesarias o para tratar la existencia de condicionalidad en los datos originales (Micó, 2012).
- El investigador tiene poca ayuda para determinar científicamente la dimensionalidad de la representación espacial y la representatividad final de la solución en los desarrollos informáticos realizados en la actualidad.
- Cuando el conjunto de datos analizado es muy grande, los algoritmos originalmente desarrollados para los métodos clásicos en ‘MDS’ son muy lentos en máquina. Esta limitación se ha vuelto más urgente recientemente, con la creciente

disponibilidad de conjuntos de datos muy grandes y la correspondiente necesidad creciente por parte de los investigadores de algoritmos de reducción de dimensionalidad escalables. El cuello de botella en los 'MDS' clásicos se produce en el cálculo de los 'K' primeros valores propios y de los vectores propios de la matriz obtenida a partir de la matriz de distancias original (de Silva & Tenenbaum, 2004).

8.2.1.5 Análisis de Correspondencias (CA)

El **Análisis de Correspondencias** ('CA'), también llamado 'Análisis Factorial de Correspondencias' es una técnica que trabaja con tablas de contingencia (donde recogemos las frecuencias de aparición de dos o más variables cualitativas en un conjunto de elementos) y con la distancia Chi-Cuadrado (distancia Euclídea ponderada que mide distancias entre perfiles, calculada de forma que, las categorías más frecuentes se ponderan menos y las menos frecuentes se ponderan más). Representa por igual filas y columnas y trata de encontrar planos en los cuales la inercia sea máxima. Su objetivo es transformar una tabla con información numérica en una representación gráfica, donde se visualizan las categorías de las variables como puntos en un plano buscando unos ejes (variables hipotéticas) que explican distribuciones de frecuencia. Existen dos escuelas que han estudiado este análisis:

- Para Benzecri (1973) y su escuela francesa, el 'CA' es una técnica que se utiliza para representar las filas y las columnas de una tabla de contingencia como puntos en un espacio vectorial de baja dimensión, de forma que los correspondientes espacios se pueden superponer para obtener una representación conjunta. Tiene un enfoque marcadamente exploratorio. Da importancia a la representación conjunta de las variables.
- Según la escuela Holandesa (Gifi, 1990) es una técnica multivariante que estudia las relaciones de dependencia entre variables categóricas a partir de una tabla de contingencia para extraer variables ficticias cuantitativas a partir de las variables cualitativas originales. Da importancia a la descripción de la relación entre las variables y está enfocado en la construcción de modelos restrictivos.

En el 'CA' podemos diferenciar entre el **Análisis de Correspondencias Simple** ('SCA') y el **Análisis de Correspondencias Múltiple** ('MCA'). Se diferencian por el número de variables implicadas: mientras que el primero trabaja con dos variables, el múltiple puede analizar más de dos variables por lo que, en este caso, las tablas de contingencia son mucho más complejas. Para el caso que nos ocupa dada la

dimensión de los datos, el 'MCA' es la técnica utilizada por los investigadores en el ámbito del Big Data, y por lo tanto será la que analicemos a continuación.

Greenacre (2008) señala que el 'MCA' se ocupa de analizar las relaciones entre un conjunto de variables, generalmente homogéneas al hacer referencia a un mismo tema, siendo también las escalas de respuestas iguales. Además, podemos trabajar sobre una matriz que contenga los datos codificados en 0 o 1 (matriz binaria), o bien sobre una matriz formada por todos los cruzamientos posibles entre variables (matriz de Burt). La aplicación de uno u otro método es prácticamente equivalente y la diferencia entre ellas se halla en su resultado para las inercias principales: las de la matriz de Burt son los cuadrados de las de la matriz binaria (por ello los porcentajes de inercia para esta técnica son siempre más elevados que para el binario). Pero para el autor la geometría resultante para las aproximaciones obtenidas a través de estas dos metodologías en MCA no está clara y propone el '**Análisis de Correspondencias Conjunto**' ('ACCo') para superar este inconveniente.

Dentro de la bibliografía consultada hemos encontrado su aplicación vinculada al Big Data, en múltiples disciplinas. De esta manera, Fellenberg et al. (2001) en el campo de la Genómica (una vez más) demuestran la aplicabilidad del 'ACM' en el análisis de los datos de microarrays, mostrando la asociación existente entre la hibridación genética como resultado de la aplicación de distintas condiciones experimentales. Por su parte, Otsuki y Kawamura (2013) a través de una API de Twitter extraen información de las compras realizadas en el periodo de Navidad del 2013 con el objeto de establecer cluster de comportamiento en relación a la localización geográfica de los consumidores, para ello utilizan el 'MCA' como técnica intermedia para definir las medida de similitud entre las observaciones, para después realizar las distintas agrupaciones, objeto final del estudio. Fleites, Ha, Yang y Chen (2014) en un buen libro dedicado al Cloud Computing definen una nueva técnica para la clasificación semántica dentro de Map Reduce basada en el 'MCA'. Blasius y Greenacre (2014) en su libro '*Visualization and Verbalization of Data*', en el cual revisan la metodología para adaptarlas a las necesidades (o limitaciones) actuales, muestran la importancia que tiene esta técnica dentro del ámbito del Big Data. Chakradeo, Reaves, Traynor y Enck, (2013) desarrollan dentro del ámbito de la informática un software denominado 'MAST' ('*Mobile Application Security Triage*') cuyo objetivo es gestionar dentro de los e-markets (tiendas online de aplicaciones para dispositivos móviles) los escasos recursos existentes para la detección de Malware dentro de la propia plataforma de venta, dirigiendo éstos a las apps más utilizadas y mejor calificadas por los usuarios, para ello implantan dentro del algoritmo el 'MCA' como método para determinar cuáles

son sus objetivos de análisis, demostrando finalmente con un caso práctico (en el cual analizan la interacción entre más de 15.000 aplicaciones de Google Play y más de 700 aplicaciones Malware) la importante reducción de costos que supone su aplicación (¡esto sí es Big Data!). Finalmente, Mercy y Padmavathi (2014) desarrollan un método para cruzar una base de datos que contiene la información de un grupo de proteínas a analizar, contra bases de datos ya existentes que contienen información acerca de proteínas "con carga informativa errónea", encontrando la relación y pudiendo extraer automáticamente las proteínas que coinciden (y que presentan un "comportamiento erróneo"), utilizando para ello esta metodología estadística multivariante.

Las ventajas de este método:

- Su principio geométrico permite representar simultáneamente individuos y variables en un mismo espacio, construyendo una 'cartografía' y caracterizando, mediante los factores resultantes, la existencia de distintos perfiles.
- Para enriquecer los resultados e interpretaciones, esta técnica permite incluir variables y/o individuos 'suplementarios' que no participan en la construcción del espacio geométrico de las variables.
- Sus resultados gráficos facilitan la comprensión e interpretación de los datos, de una manera directa al investigador.

Las desventajas:

- Esta técnica, al igual que el 'PCA' y el 'MDS', utiliza la 'DVS' como método de factorización de la matriz y su cálculo, como ya hemos señalado para estas técnicas, presenta dificultades cuando el tamaño de los datos es muy grande.
- La existencia de variables mixtas en una investigación hace que deban de realizarse una serie de tareas previas para adecuar los datos a uno de los principales requisitos de esta técnica: trabajar con variables de tipo cualitativo. Para ello existen diferentes propuestas (como la codificación de Escofier) que permiten, después de su aplicación, realizar el 'CA' a toda la matriz de datos.
- Muchas veces la interpretación de los resultados requiere un trabajo extra para los investigadores.

8.2.2 Técnicas de Clasificación

Una de las tareas fundamentales de cualquier disciplina vinculada a la Ciencias Experimentales es la clasificación ordenada de los objetos y los fenómenos naturales observados. Sólo mediante una buena clasificación es posible establecer relaciones entre la gran variedad de resultados propios de la observación científica (Cuadras, 2012). El objetivo de todo este conjunto de técnicas es identificar las características de los elementos analizados que son más útiles para diferenciarlos en categorías que pueden haberse especificado a priori (o no) por parte del analista y obtener diferentes grupos de individuos y poder clasificarlos. Permite combinar observaciones en agrupaciones (cluster) de forma que obtengamos:

- Grupos homogéneos de individuos (las observaciones dentro de cada grupo han de ser similares entre sí).
- Grupos bien diferenciados (las observaciones de un grupo deben diferenciarse de las observaciones de los otros grupos).

Las relaciones entre los objetos se establecen calculando una matriz de similaridades o disimilaridades, que informe sobre la analogía o diferencias entre unos y otros individuos sobre la base de las características cualitativas elegidas. Cuando se utilizan variables cuantitativas se trabaja con una matriz de correlaciones o de distancias.

8.2.2.1 Análisis de Cluster

Se conoce como **Análisis de Cluster** a un conjunto de técnicas de carácter descriptivo (no inferencial) que trabaja con variables tanto cualitativas, como cuantitativas y su objetivo es obtener agrupaciones de los individuos y proceder, de esta manera, a su clasificación. Se trata de elaborar sucesivas particiones (‘*clusterings*’), organizadas en diferentes niveles jerárquicos, estando cada partición formada por clases disjuntas (‘*clusters*’).

Se considera que la obra que más ha influido en el enfoque numérico de clasificación, frente a los criterios tradicionales, es el libro escrito por Sokal y Sneath (1963) en el que se expone “*el estudio teórico de la clasificación incluyendo sus bases, principios, procedimientos y reglas*” (Cuadras, 2012).

El dendograma es la representación gráfica que mejor ayuda a interpretar el resultado de un Análisis de Cluster. Los objetos se representan como nodos y las ramas del árbol indican los sujetos que se han fusionado en un cluster y su longitud indica la distancia de la fusión.

Existen diferentes técnicas dentro de este tipo de análisis:

- *CLUSTER JERÁRQUICOS*: No se conoce el número de cluster a priori. Los grupos se van fusionando progresivamente mientras decrece la homogeneidad entre las agrupaciones cada vez más amplias que se van formando. Se utiliza tanto para variables cuantitativas, como cualitativas. Puede ser de tipo:
    - *Aglomerativo*: Se parte inicialmente de todos los objetos que se van progresivamente fusionando para formar particiones sucesivas. Partimos de 'n' cluster para 'n' individuos.
    - *Divisivo*: Se parte del conjunto total que se subdivide progresivamente en grupos más finos. Partimos de 1 cluster para 'n' individuos.
- *CLUSTER NO JERÁRQUICOS*: Requiere conocer el número de cluster a priori. Se forman grupos homogéneos excluyentes, pero con máxima divergencia entre las clases y sin establecer relaciones, u ordenación jerárquica, entre los mismos. Sólo puede ser aplicado a variables cuantitativas. Puede realizarse para un número de objetos relativamente grande pues no requiere el cálculo de todas las posibles distancias.
- *CLUSTER EN DOS ETAPAS* (*bietápicos*): Pensado para realizar Minería de Datos ('*Data Mining*'), es decir para estudios con un número muy grande de observaciones que pueden tener problemas de clasificación si se aplican los procedimientos anteriores. Permite trabajar conjuntamente con variables de tipo mixto (cualitativo y cuantitativo) y puede realizarse cuando el número de cluster es conocido a priori y también cuando no se conoce. Es un método de agrupación en dos pasos (IBM Knowledge Center, 2013):
    - El primero consiste en hacer una única 'pasada' por los datos, para comprimir los registros de entrada de la fila en un conjunto de subclusters administrable.
    - El segundo utiliza un método de agrupación en clusteres jerárquico para fundir progresivamente los subclusters obtenidos en el primer paso en clusteres cada vez más grandes.

Dentro de la bibliografía consultada hemos encontrado su aplicación vinculada al Big Data en múltiples disciplinas ya que uno de los principales objetivos en todos los análisis en Big Data es conseguir la clasificación de los individuos para definir comportamientos y estructuras adaptando los datos, para posteriormente hacer uso de otra técnica estadística que, normalmente permita hacer inferencia sobre los mismos. Entre los artículos de investigación más interesantes que hemos encontrado en la búsqueda realizada tenemos que citar el realizado por Dopazo, Zanders, Dragoni,

Amphlett y Falciani (2001) que analizan el uso de los Cluster Jerárquicos (aglomerativos y divisivos) como técnicas exploratorias muy válidas dentro del campo de la genómica (una vez más) para proponer la necesidad de combinar estos resultados con técnicas de análisis estadístico robustas que permitan la predicción, por ejemplo, de la pertenencia de una enfermedad o la modelización de las variables biológicas a través de la propia expresión génica, detallando la conveniencia de aplicar para ello el 'Análisis Discriminante Lineal' ('LDA') y otras técnicas de inferencia estadística, como el 'PCA'. Este grupo de científicos dirigidos por el español Joaquín Dopazo, que pertenece al Centro Nacional de Investigaciones Oncológicas (CNIO), ya en 2001 señalaron la importancia que tiene la estadística en este campo en el momento en el que la explosión de los datos que manejan es una realidad, señalando que este hecho permitirá la evolución de la Biología a una ciencia más madura al formar parte activa, dentro de su investigación experimental base, los enfoques matemáticos. Siguiendo esta línea de investigación, en un trabajo en el cual también participa el autor antes mencionado, Herrero, Valencia, y Dopazo (2001) presentan la aplicación del algoritmo jerárquico divisivo y configurable que denominan 'SOTA', y que crearon Dopazo y Carazo (1997), en el ámbito de la genómica y que mejora notablemente las prestaciones respecto a los algoritmos utilizados hasta la fecha en su ámbito disciplinar. Valkonen, Kolehmainen, Lakka y Salonen (2002), en un artículo posterior, aplican el algoritmo de clusterización 'SOM' (ya mejorado por los anteriores investigadores) en el ámbito de las enfermedades cardiovasculares provocadas por el 'síndrome insulínico', señalando las limitaciones del método que ya habían superado los investigadores anteriormente citados a través del desarrollo del algoritmo 'SOTA' lo que demuestra la necesidad que señalábamos en la presentación de este capítulo, al postular la necesidad del desarrollo de un, cada vez más requerido, cuerpo curricular estadístico como herramienta central para el conjunto de disciplinas científicas. Banas et al. (2013) desarrollan mediante el lenguaje de programación R un método de agrupación jerárquica de tipo automático para el análisis de los tipos de radiación producida por las emisiones de Rayos X, iniciando el estudio con la aplicación de un 'PCA' sobre los datos para, posteriormente, con los resultados obtenidos aplicar el método de clustering antes comentado. También es necesario volver a comentar el artículo de Otsuki y Kawamura (2013) que a través de una API de Twitter extraen información sobre las compras realizadas en la Navidad de 2013 con el objeto de establecer cluster del comportamiento de compra en relación a la localización geográfica de los consumidores, para ello utilizan un 'Análisis de Correspondencias Múltiple' ('MCA') como técnica intermedia para definir las medida de similitud entre las observaciones, para después realizar las distintas agrupaciones (objeto final del

estudio). Buza, Nagy y Nanopoulos (2014) desarrollan un nuevo algoritmo denominado 'SOHAC' que mejora notablemente el rendimiento de las computadoras en su gestión de almacenamiento y análisis de las transacciones que se producen en un mercado financiero, donde sus extremas velocidades y volúmenes crean verdaderos problemas si se adoptan soluciones informáticas de software 'clásicas'. Finalmente, Tsumoto S., Iwata, Hirano y Tsumoto Y. (2014) en un interesante artículo presentan los resultados de aplicar técnicas cluster en '*Data Mining*' para analizar y visualizar el comportamiento temporal de las actividades globales en un hospital japonés y sus resultados demuestran que la reutilización de los datos se define como una herramienta de gran valor dentro del ámbito de la gestión hospitalaria ya que conlleva una mejora de suss servicios (de todos los artículos revisados, tenemos que señalar que éste se asemeja a lo que verdaderamente significa gestionar y analizar Big Data).

Los puntos fuertes de esta técnica en Big Data son los siguientes:

- Es un método universal dentro del análisis estadístico, cuyo objetivo coincide exactamente con una de las principales necesidades del Big Data: la agrupación de las observaciones que permita 'hacer manejable' el volumen, la velocidad y la variedad de los registros.
- Se puede considerar como un método 'puente', ya que los investigadores lo utilizan como una técnica inicial o final, dependiendo de las necesidades planteadas por un proyecto. Por ejemplo, se puede combinar con un 'PCA' (ya que, primeramente, mediante esta técnica se 'homogeneizan' los datos) lo que permite después aplicar un Análisis de Cluster sobre los componentes obtenidos. O se puede utilizar con otras técnicas que posibilitan, una vez obtenidas las agrupaciones de las observaciones, inferir sobre los datos.

Entre sus limitaciones podemos destacar los siguientes hechos:

- Es una técnica únicamente descriptiva. Tal y como señalábamos antes, el investigador deberá hacer uso de otras metodologías estadísticas después de haber conseguido las agrupaciones de los elementos para inferir sobre los mismos.
- A veces los Cluster obtenidos son difícilmente interpretables, sobre todo cuando se utilizan métodos no jerárquicos.
- Por lo general los algoritmos de clusterización desarrollados hasta ahora, trabajan con sólo un conjunto de datos y en Big Data es muy interesante analizar dos o más conjuntos multivariados para buscar relaciones entre ellos. Por ejemplo: el primer

conjunto de datos puede representar las funciones de consulta disponibles antes del tiempo de ejecución, y el segundo conjunto de datos representa las características de rendimiento medido disponibles después de ejecutar cada consulta y pretendemos descubrir las relaciones estadísticas entre esos dos conjuntos de datos multivariantes para poder predecir el rendimiento basado en las características de esa consulta.

8.2.2.2 Análisis Discriminante (DA)

Englobadas en el **Análisis Discriminante** ('DA') existe un conjunto de técnicas que estudian las diferencias entre grupos mediante métodos multivariantes con el objeto de clasificar a un nuevo individuo en una de varias poblaciones. Tienen su origen en el trabajo de Fisher (1936) y su objetivo es obtener una función capaz de clasificar a un nuevo individuo a partir del conocimiento de los valores de ciertas variables, permitiendo describir (si existen) las diferencias entre grupos de objetos sobre los que se observan 'n' variables (variables discriminantes). De manera genérica, se comparan y describen las medias de las 'n' variables clasificadoras a través de los grupos y en el caso de que estas diferencias existan, el análisis intentará explicar en qué sentido se dan, permitiendo proporcionar procedimientos de asignación sistemática para las nuevas observaciones a uno de los grupos analizados, utilizando para ello sus valores en las variables clasificadoras. Podemos ver este procedimiento como un modelo de predicción de una variable respuesta (variable grupo), a partir un conjunto de variables explicativas (variables clasificatorias). A diferencia del Análisis de Cluster, se deben conocer previamente tanto los grupos, como a qué agrupación pertenecen ciertos individuos (de los que también se conoce sus valores en las variables discriminantes). Por lo tanto, podemos señalar que el 'DA' es una técnica estadística multivariante que trata de analizar si existen diferencias entre una serie de grupos en los que se divide una población, con respecto a un conjunto de variables y, en caso afirmativo, intenta averiguar a qué se deben, proporcionando procedimientos sistemáticos de clasificación de nuevas observaciones en alguno de los grupos considerados.

Dependiendo del modelo que siguen los datos, existen diferentes tipos de 'DA', siendo el nexo común para todos ellos, tratar de estimar la probabilidad que un individuo pertenezca a uno de los grupos considerados. De esta manera:

- *Análisis Discriminante Lineal ('LDA')*. Es un método muy conocido y por ello muy utilizado en la bibliografía consultada, donde la función discriminante es una combinación lineal de las variables originales. Parte de los supuestos de que todas las poblaciones siguen una distribución normal y además, requiere que la matriz de

covarianzas sea también igual para todas las poblaciones. La proyección óptima (o transformación) en el 'LDA' clásico se obtiene minimizando la varianza dentro de cada clase y maximizando la varianza entre esas clases simultáneamente, logrando así la máxima discriminación posible.

- *Análisis Discriminante Cuadrático ('CDA')*. Al igual que en el método anterior, se parte del supuesto de normalidad de los datos, pero en este caso la matriz de covarianzas de cada población es distinta. Si se ha comprobado este hecho, es necesario usar estas matrices para calcular la probabilidad de pertenencia al grupo. Los inconvenientes que plantea respecto al 'LDA' es que las regiones resultantes con estas funciones de segundo grado son típicamente disjuntas y, a veces, difíciles de interpretar en varias dimensiones. Además, el número de parámetros a estimar en el caso cuadrático es mucho mayor que en el caso lineal (por ejemplo, con 10 variables y 4 grupos pasamos de estimar 95 parámetros en el caso lineal, a 260 en el cuadrático). Este gran número de parámetros hace que, salvo que tengamos muestras muy grandes, la discriminación cuadrática sea bastante inestable y, aunque las matrices de covarianzas sean muy diferentes, se obtengan con frecuencia mejores resultados con la función lineal. Otro problema adicional es que es muy sensible a las desviaciones de la normalidad de los datos, por lo cual, en líneas generales, podemos afirmar que la clasificación lineal es, en estos casos, más robusta (Peña, 2002).
- *Análisis Discriminante Logístico*. Cuando no se verifican las condiciones de aplicación de los análisis discriminantes anteriores (distribuciones normales y varianzas iguales), ya que es frecuente que los datos disponibles para la clasificación no asuman estas restricciones (por ejemplo, en muchos problemas de clasificación se utilizan variables binarias) se ha desarrollado el Análisis Discriminante Logístico. Con esta técnica, tratamos de estimar la probabilidad de que un individuo pertenezca a uno de los 2 grupos cuando tiene una combinación concreta de variables explicativas, mediante un modelo de respuesta logística. Una vez que se han estimado los parámetros y se han calculado las probabilidades de pertenencia a cada una de las poblaciones, el individuo será asignado a aquella población para la cual su probabilidad es mayor.
- Existen otros métodos de 'Análisis Discriminante', algunos no-paramétricos, otros para variables mixtas: como el método del núcleo, del vecino más próximo, el basado en el "location model" de Krzanowski, etc. (Cuadras, 2012).

Dentro de estos análisis, existen varios métodos de clasificación dependiendo del número de grupos a clasificar (dos o más grupos), de las hipótesis hechas acerca del

comportamiento de las variables en cada grupo (Normalidad, Homocedasticidad…), así como del criterio utilizado para llevar a cabo dicha clasificación. Entre ellas, las más utilizadas son las siguientes:

- *Discriminadores Lineales*: Su objetivo es, en lugar de trabajar con las 'p' variables originales, definir un vector de 'r' variables canónicas, que se obtengan como combinación lineal de las originales y que permitan resolver el problema de clasificación de la siguiente manera: proyectando las medias de las variables en los grupos sobre el espacio determinado por las 'r' variables canónicas, proyectamos un nuevo elemento y realizamos su clasificación en aquella población a cuya media se encuentre más próximo. Las distancias se miden con la distancia euclídea en el espacio de las variables canónicas (Peña, 2002).
- *Regla de Máxima-Verosimilitud*: Se puede obtener una regla de clasificación asignando un elemento a la población donde su razón de verosimilitud sea mayor. Este criterio es más genérico que el geométrico y está asociado a funciones discriminantes logísticas. En el caso de cumplirse la normalidad multivariante y la existencia de una matriz de covarianzas común, los discriminadores máximo-verosímiles coinciden con los lineales; pero si las matrices de covarianzas son diferentes, entonces este criterio da lugar a discriminadores cuadráticos (Cuadras, 2012).
- *Teorema de Bayes*: Utiliza este teorema y se aplica cuando se conoce a priori la probabilidad de que una observación pertenezca a una población y calcula la función de densidad de probabilidad. Este enfoque permite dar una solución general al problema de clasificación tanto cuando los parámetros son conocidos, como cuando los parámetros deben estimarse a partir de los datos, aportando una solución directa del problema teniendo en cuenta la incertidumbre en la estimación de los mismos que, a diferencia del enfoque clásico, ignora este hecho. La solución es válida sean o no iguales las matrices de covarianzas. El procedimiento para clasificar una observación dada la muestra, es asignarla a la población con mayor probabilidad, para ello se obtiene el máximo de las probabilidades a posteriori de que la observación a clasificar, dada una muestra, provenga de una de las poblaciones (Peña, 2002).
- *Análisis Discriminante basado en Distancias*. Permite mejorar la ordenación y la formación de clusters cuando las variables son binarias, categóricas o mixtas (los 3 métodos anteriores funcionan bien con variables cuantitativas). Aceptando y aplicando el principio de que siempre es posible definir una distancia entre observaciones, es posible dar una versión del 'DA' utilizando solamente este tipo

de medida. Geométricamente el criterio consiste en asignar el individuo a la población más cercana, midiendo esa cercanía a partir de la distancia de Mahalanobis (Cuadras, Fortiana & Oliva, 1997; Cuadras, 2012).

Conviene finalmente hacer notar, que el DA no es la única técnica estadística implicada en el proceso de clasificación de observaciones en grupos previamente fijados por el analista. Otra alternativa interesante viene dada por los modelos de regresión con variable dependiente cualitativa (de los que el 'Análisis Discriminante' podría considerarse un caso particular) como son, por ejemplo, los modelos de Regresión Logit.

Las técnicas de 'DA' tienen numerosas aplicaciones en Big Data y se utilizan para abordar problemas muy interesantes y a la vez complejos tanto en la biología (estudios genéticos), como en otras disciplinas como la medicina (detección precoz del cáncer), la ingeniería (reconocimiento de voz y caras humanas), la informática (clasificación de correo spam), la economía (adjudicación de créditos bancarios por previsiones de insolvencia), el arte (autoría de manuscritos y pinturas de autoría desconocida) y un largo etcétera. Entre los artículos de investigación que hemos encontrado en la búsqueda realizada podemos realizar una clasificación de los mismos en relación a su objetivo. De esta manera:

- *Artículos que proponen modificaciones sobre los Algoritmos inicialmente propuestos para el 'DA'*: Du y Nekovei (2005) desarrollan una variación del algoritmo 'LDA' inicialmente propuesto por Fisher (el 'FLDA'), denominado 'CLDA' (*'Constrained Linear Discriminant Analysis'*) para su uso en el análisis, en tiempo real, de imágenes multiespectrales provenientes de satélites, demostrando su mayor efectividad y sencillez computacional respecto a otros métodos de uso muy extendido en ese área. Por su parte Ye (2007), propone otra modificación del 'FLDA' utilizando para ello la técnica de mínimos cuadrados y lo denomina 'LS-LDA' (*'Least Squares-Linear Discriminant Analysis'*) y demuestra su eficacia aplicándolo a tres ámbitos muy distintos: el análisis de documentos de texto, el de imágenes de rostros humanos y el de datos genéticos. Otra mejora propuesta sobre el 'FLDA' la realizan Sabatier y Reynès (2008) y la denominan 'GA-SLDA' (*'Genetic Algorithm-Simple Linear Discriminant Analysis'*) incorporando '*Algoritmos Genéticos*' ('GA') al mismo y señalando su correcta aplicación en el campo del reconocimiento de la escritura y los estudios de diagnósis médica, desarrollando un ejemplo para la diagnosis de enfermedades coronarias. Cai, He y Han (2008) proponen otro algoritmo, que lo denominan '*Spectral Regression Discriminant*

Analysis' ('SRDA') el cual, utilizando el análisis gráfico espectral, encaja el 'DA' en el ámbito de la regresión lo que facilita tanto su cómputo, como la necesidad de utilizar otras técnicas auxiliares; específicamente, el 'SRDA' sólo tiene que resolver una serie de problemas de mínimos cuadrados regularizados, sin realizar el cálculo de los vectores propios en el cómputo, lo que supone un enorme ahorro de tiempo y memoria; y demuestran su eficacia en el ámbito del reconocimiento facial. De nuevo, con el mismo objetivo, Jin, Cao, Ruan y Wang (2014) desarrollan el algoritmo denominado '*Multiview Smooth Discriminant Analysis*' ('MSDA') basado en una nueva técnica dentro del 'ML' denominada '*Extreme Learning Machines*' ('ELM').

- *Artículos que destacan la importancia del 'DA' como técnica de análisis en Big Data* y proponen su aplicación a diversos ámbitos de investigación y desarrollo dentro del mismo (Chen, Chiang & Storey, 2012; March, 2008; Chen & Zhang, 2014; Witten & Frank, 2005).
- *Artículos relacionados con el teorema de Bayes* en el campo de la inferencia y que haremos referencia a los mismos en el futuro trabajo de Tesis.

Las ventajas de este conjunto de técnicas:

- El 'DA' por su capacidad para la clasificación, identificación, asignación, reconocimiento de patrones y selección es uno de los métodos de análisis multivariante que ha tenido mayor desarrollo en Big Data.
- Es una técnica que permite su combinación con otras metodologías estadísticas para mejorar los resultados de los análisis (ya hemos visto su aplicación combinado con 'PCA', Análisis de Cluster, etc.)

Las desventajas:

- Las restricciones que impone el modelo 'LDA' (ampliamente utilizado) al asumir la normalidad multivariante y la igualdad de las matrices de covarianzas y que, como ya hemos visto, resuelven los investigadores a través de distintas propuestas de mejora respecto al algoritmo original de discriminación.
- Si todas las variables son continuas, es frecuente que aunque los datos originales no sean normales sea posible transformar las variables y, de esta manera, los métodos discriminantes clásicos pueden aplicarse a las variables transformadas. Sin embargo, cuando tengamos variables binarias o categóricas para clasificar, la hipótesis de normalidad multivariante resulta poco realista, y es necesario trabajar

con otros enfoques que pueden dar solución a estos casos, como el modelo discriminante logístico para varias poblaciones '*MultiLogit*' (Peña, 2002).

- Esta técnica, al igual que hemos visto para otras técnicas ya analizadas ('PCA', 'MDS', etc.) utiliza la 'DVS' como método de factorización de la matriz y su cálculo presenta dificultades cuando el investigador trabaja con conjuntos de datos muy grandes.
- Limitación del análisis de la función discriminante cuando existen correlaciones entre las variables predictivas.
- Problemas ante la existencia de muestras desequilibradas.
- Complicaciones computacionales por la presencia de *'outliers'*.

8.2.2.3 Análisis de Correlación Canónica (CCA)

El **Análisis de Correlación Canónica** ('CCA') inicialmente desarrollado por Hotelling (1936), es una técnica confirmatoria que analiza varias variables dependientes y varias variables independientes de naturaleza tanto cuantitativa, como cualitativa. Según Peña (2002) esta técnica puede ser utilizada cuando se desea estudiar la relación entre un conjunto de variables que puede dividirse en dos grupos homogéneos (por distintos criterios) y, también, cuando los dos grupos se corresponden a las mismas variables pero medidas en dos momentos distintos en el tiempo, en el espacio, etc.

Su principio básico es desarrollar una combinación lineal para cada conjunto de variables, de tal manera que la correlación entre esos grupos sea máxima. La primera etapa, por lo tanto, consiste en obtener una o varias funciones canónicas, donde cada función consiste en un par de variables aleatorias, una en representación de las variables independientes y otra para las variables dependientes, que consisten en la suma ponderada de cada variable en cada conjunto de datos, maximizando a través de los pesos de cada variable la correlación entre estos dos conjuntos tal y como señalábamos anteriormente. Su cálculo es similar al procedimiento utilizado con el 'Análisis Factorial' sin rotar; es decir: la primera función extraída representa la cantidad máxima de la varianza para el conjunto de variables originales, la segunda se calcula a continuación de manera que represente la máxima varianza posible no contabilizada por la primera y así sucesivamente, hasta que se hayan extraído todas las funciones necesarias (siendo su número máximo igual al número de variables que presenta el conjunto de datos más pequeño).

Cada par de variables canónicas obtenidas es ortogonal a todas las demás y es necesario medir la potencia de relación entre ellas mediante el 'Coeficiente de Correlación Canónica', que representa la cantidad de varianza de una variable

canónica explicada por la otra variable canónica (cantidad de varianza compartida). De todas las funciones canónicas extraídas las que deben ser interpretadas son aquellas cuyos 'Coeficientes de Correlación Canónica' son estadísticamente significativos y, para ello, la prueba más utilizada es el '*Estadístico F*' basada en la Lambda de Wilks, aunque también en la mayoría de los paquetes comerciales están disponibles otras medidas como la Traza de Hotelling, la Raíz Mayor de Roi y la Traza de Pillai. Además, se tiene que medir la magnitud de la Correlación Canónica que se basa (de nuevo, como en el 'Análisis Factorial') en la varianza explicada para las variables canónicas obtenidas (y no sobre las originales); para, finalmente, evaluar la medida de redundancia para el porcentaje de varianza explicada por los dos conjuntos de variables y con ello, podemos conocer qué funciones canónicas son las adecuadas para definir y representar con precisión la correlación existente para el conjunto de variables.

Existen dos tipos de 'Análisis de Correlación Canónica':

- El **Lineal**, cuya correcta aplicación requiere la evaluación de una serie de supuestos como:
    - La existencia de *Normalidad* para las variables, ya que este hecho estandariza su distribución permitiendo una mayor correlación entre ellas.
    - Este tipo de análisis se basa en la existencia de la *relación lineal* entre variables, por lo tanto si no la cumplen hay que proceder a su transformación.
    - La presencia de *Homocedasticidad*, que disminuye la correlación entre las variables. Por ello, si se observa también hay que proceder a su tratamiento.
    - La existencia de *Multicolinealidad*, que hace que la representación obtenida sea poco fiable.
- El **No Lineal** se corresponde con el 'Análisis de Correlación Canónica' para datos categóricos (método OVERALS). Posibilita la descripción y el análisis de estructuras, modelos y tipologías relacionales muy complejas. Sus características principales (a diferencia del lineal) y que tienen mucha importancia para aplicarlo en Big Data son las siguientes:
    - Permite analizar más de dos conjuntos de variables.
    - Las variables pueden ser escaladas como ordinales, nominales o numéricas y, de esta manera, las relaciones no lineales entre las variables pueden ser analizadas.

- No maximiza las correlaciones entre los conjuntos de las variables (como en el caso lineal), sino que éstos se comparan con un 'conjunto compromiso' desconocido y definido por las puntuaciones de los objetos.

El uso de esta técnica en Big Data está muy extendido (al igual que otros métodos de regresión) porque, en general, es adecuada cuando existe un conjunto de variables que pueden dividirse en dos o más grupos homogéneos (creados por diferentes criterios que se pueden corresponder, por ejemplo, con las mismas variables medidas en momentos distintos en el tiempo, en el espacio, etc.) y se desea estudiar la relación entre esos conjuntos de variables. Y es que mediante otras técnicas multivariantes, como el 'PCA' o el 'Análisis de Cluster', no es posible realizar este análisis conjunto ya que su aplicación tradicional se limita a un solo grupo de datos (podríamos aplicar estas técnicas por separado para cada uno de ellos, pero no lograríamos identificar las correlaciones existentes entre esos grupos). Pero para su correcta implantación en Big Data los investigadores, conocedores de sus limitaciones iniciales, pero también sabedores de su potencialidad, parten de su planteamiento base y definen las modificaciones necesarias para eliminar las restricciones propias de la métrica y las necesidades de aplicarlo sobre grandes y variados conjuntos de datos. De esta manera, la bibliografía consultada determina de manera consecuente lo ya mencionado; así los investigadores en cada artículo analizado proponen variaciones que mejoran el rendimiento de esta técnica en el ámbito computacional del Big Data. De esta manera, Ganapathi et al. (2009) aplican dentro del entorno del 'ML', para predecir el comportamiento de las consultas que se realizan sobre bases de datos que crecen continuamente, una modificación de este análisis denominado '*Kernel Canonical Correlation Analysis*' ('KCCA') que reemplaza el cálculo de los Productos Escalares Euclídeos del 'CCA' (para dos vectores en un espacio de esta naturaleza, se define como el producto de sus longitudes por el coseno del ángulo que forman) por lo que denominan '*Kernel Functions*', las cuales (siguen señalando) se encuentran en el centro de muchos de los desarrollos más importantes realizados dentro del 'ML' para el reconocimiento automático de similitudes en los datos. De nuevo, Ganapathi, Chen, Fox, Katz y Patterson (2010) en otro interesante artículo utilizan esta técnica para dar solución a problemas de predicciones inadecuadas de dimensionalidad para los recursos computacionales en la '*nube*', introduciendo el algoritmo directamente en MapReduce (solución estandarizada para la gestión del Big Data). Por otra parte Lin et al. (2013), señalan la necesidad imperativa en el ámbito de la genómica, ante la aparición de grandes conjuntos de datos provenientes de muy diferentes fuentes y plataformas (lo cual ha provocado una mejora considerable en la compresión de la

influencia que tienen los factores genómicos en el desarrollo de las denominadas 'enfermedades raras') de analizar pormenorizadamente la relación entre esos conjuntos de datos. Para ello, aplican una variación del 'CCA' denominado '*Sparse Canonical Correlation Analysis*' ('SCCA') para evitar los problemas que la técnica original tiene cuando trabaja con matrices donde los datos de las muestras son significativamente menores que la de los biomarcadores (lo cual es habitual para este tipo de datos tal y como hemos podido comprobar también con la técnica denominada '*Sparce Principal Component Analysis*' ('SPCA') que Zou et al. (2006) desarrollaron en este mismo ámbito para resolver el problema que se planteaba con el uso del 'PCA' en el caso de la existencia de un número de variables (genes) muchísimo mayor que el número de observaciones). Lu (2013) a su vez desarrolla la implantación del '*Multiple Principal Component Analysis*' ('MPCA') en el ámbito de reconocimiento de imágenes en 2D y 3D y comprueba su fiabilidad y mejores resultados frente a las técnicas tradicionales que se utilizan habitualmente ('CCA' y '2D-CCA'). Por su parte, Shen, Sun y Yuan (2013) proponen un nuevo método denominado '*Orthogonal Canonical Correlation Analysis'* ('OCCA') donde reemplazan las limitaciones de ortogonalidad propias en el 'CCA' clásico mediante la introducción de restricciones a sus particularidades, creando un nuevo algoritmo basado en la doble descomposición en valores singulares ('*Twin Eigen Decomposition'*) para obtener de manera gradual los vectores canónicos ortogonales y, finalmente, demostrar su eficacia frente a otras técnicas de correlación canónica en el ámbito del análisis y reconocimiento facial. Por último, Shen, Sun, Tang y Priebe (2014), en un artículo aún en fase de revisión definen la conveniencia de utilizar el método que han desarrollado denominado *'Generalized Canonical Correlation Analysis*' ('GCCA') para múltiples conjuntos de datos multivariantes ya que mejora el rendimiento de clasificación (en comparación con el 'CCA'), incluso utilizando sólo dos conjuntos de datos. Para ello, ilustran sus resultados teóricos con un experimento en el cual realizan la clasificación de documentos de texto provenientes de Wikipedia.

Las ventajas de esta técnica aplicada a Big Data.:

- Es una técnica multivariante perteneciente a los métodos de dependencia y es muy útil en investigaciones donde su objetivo está asociado, simultáneamente, a las relaciones existentes entre múltiples variables dependientes y múltiples variables independientes, es decir dos o más conjuntos de datos para los cuales queremos conocer su correlación (muy propio del fenómeno Big Data).
- Permite trabajar con datos de naturaleza muy diversa (registros ordinales, nominales, cuantitativos…)

- Se revela como una técnica muy valiosa, por su fiabilidad y resultados en la predicción a priori, de comportamientos no deseados para entornos de desarrollo y explotación complejos caracterizados por grandes y desestructuradas bases de datos. Permitiendo planificar adecuadamente, junto a técnicas de 'ML', la dimensionalidad y los tiempos precisos en máquina para dar una respuesta satisfactoria a las necesidades cambiantes propias del desarrollo de un sistema en contínua evolución.

Las limitaciones propias de este método se extrapolan a Big Data:

- Debemos asumir la existencia de una serie de presupuestos que condicionan enormemente la solución (normalidad y relación lineal entre las variables). Además, se debe eliminar la Homocedasticidad y la existencia de Multicolinealidad (que se produce cuando las variables independientes, al igual que las dependientes, están correlacionadas entre sí, lo que imposibilita aislar este efecto de cada variable) da lugar a interpretaciones poco fiables. Por ello es preferible utilizar otras técnicas si nos encontramos con estos supuestos.
- Esta técnica trabaja con matrices de covarianzas y correlaciones muestrales; y utiliza la Descomposición en Valores Singulares ('DVS') como método de factorización de esa matriz y, como ya hemos analizado, su cálculo presenta dificultades cuando el tamaño de los datos es muy grande.

9 LA GEOGRAFÍA CUANTITATIVA

Según los estudios historiográficos consultados de Capel y Sáez (1985) y Montello y Sutton (2006), durante los años comprendidos entre las décadas de los 40 y 60 del Siglo XX se generalizaron en todas las Ciencias Sociales profundos cambios metodológicos relacionados con el triunfo de la corriente filosófica positivista que propugnaba que las Matemáticas, así como la 'racionalidad cognitiva', eran las únicas fuentes seguras para llegar al verdadero conocimiento, oponiéndose al resto de fenómenos no verificables. En el arranque de esta reformulación teorética para la Geografía, tuvo un papel muy destacado el trabajo de Schaefer (1953) donde se expusieron con claridad la mayor parte de los problemas epistemológicos que, posteriormente, los autores más destacados de la **Nueva Geografía** (como también se la denominó a la **Geografía Cuantitativa**) desarrollarían. Schaefer expresó, por primera vez, la idea del 'excepcionalismo' para designar la atribución a la Geografía de un carácter singular respecto al resto de ciencias y señaló la necesidad de incorporar a la misma, cuanto antes, métodos verdaderamente científicos. Para este autor, la Geografía era una ciencia que se encontraba, todavía, en la 'fase juvenil' de las clasificaciones y no había pasado a la 'fase adulta' de la formulación de leyes propias de otras disciplinas. Su actitud crítica ante la pura descripción geográfica tradicional y la necesidad del establecimiento de esas leyes científicas generales fueron ideas que, unos años después, Bunge (1960) utilizó y amplió y puede decirse que es en estos trabajos donde las distintas escuelas cuantitativas (sobre todo la anglosajona y la escandinava) encontraron los principios básicos para establecer su discurso teórico. A partir de ese momento, en buena parte del pensamiento geográfico (no sin una fuerte oposición por parte de los sectores más conservadores), se adopta el método científico que se rige por leyes que explican el funcionamiento de un fenómeno que la ciencia ha de descubrir y enunciar. Así, se adoptó esta metodología basada en la observación de los hechos, el planteamiento de una hipótesis para explicarlos, su comprobación experimental y, si ésta es plausible, su aceptación de manera provisional como ley o teoría con validez general, representándola de forma simplificada a través de modelos.

Johnston (1981) definió la Geografía Cuantitativa como "aquella (ciencia) que utiliza modelos matemáticos y análisis estadísticos y no únicamente la manipulación aritmética de los datos"; y su idea central según Capel y Urteaga (1982) "es que por debajo de la diversidad y de la compleja madeja que forman los fenómenos espaciales existe un orden que permite explicarlos... para encontrar ese orden que rige la

organización espacial debe relegarse a un segundo plano el estudio de los fenómenos singulares o accidentales, y detenerse en las regularidades, en los procesos de tipo general que afectan a la superficie terrestre".

Los Geógrafos, a partir de entonces, comenzaron a intentar definir leyes que rigen la localización y la distribución de determinadas manifestaciones espaciales (asentamientos, usos del suelo, etc.), utilizando tanto el método científico, como el lenguaje matemático y cartográfico. Así, se lanzaron a elaborar hipótesis, teorías y modelos para una representación simplificada del espacio geográfico que se concibe de manera objetiva y abstracta, adquiriendo formas geométricas, representando los fenómenos mediante figuras concretas (puntos, líneas, polígonos, etc.) y definiendo las relaciones entre ellos mediante trazas (Anselin & Rey, 2009).

Dentro de este gran desarrollo científico y procedimental, podemos señalar como pilares básicos sobre los cuales se ha asentado y consolidado la rama cuantitativa de la Geografía, las siguientes disciplinas o campos del conocimiento científico (y que a continuación desarrollaremos):

- El Análisis Exploratorio de Datos Espaciales.
- La Aplicación de Técnicas Multivariantes Clásicas en Geografía.
- La Estadística Espacial.
- La Geoestadística.
- Los Sistemas de Información Geográfica.
- La GISciencia.

9.1 Análisis Exploratorio de Datos Espaciales

Para el estudio de los datos georreferenciados (al igual que para otros procedimientos estadísticos), la primera etapa que se debe acometer es el 'EDA' (Tukey, 1977), tal y como señalábamos en el punto introductorio relacionado con la Estadística Multivariante (capítulo 8). El **Análisis Exploratorio de Datos Espaciales** ('ESDA' acrónimo de la denominación original en inglés '*Exploratory Spatial Data Analysis*') surge como una disciplina dentro del 'EDA', orientada al tratamiento de datos con impronta geográfica. Sus objetivos son descriptivos (más que confirmatorios) y trata de detectar patrones de comportamiento en los datos espaciales para la formulación de hipótesis basadas en la dimensión geográfica de los registros y la definición de modelos espaciales (Haining, Wise & Signoretta, 2000). En los textos de Chasco

(2010), Góngora (2007) y Moreno y Cañada (2006) añaden, a los objetivos particulares del EDA, los siguientes aspectos:

- La descripción y visualización de la distribución espacial.
- La identificación de valores espaciales extremos (‘*outliers*’) y su localización.
- El análisis de las formas de autocorrelación espacial (de carácter global o local).
- La definición de la covariación espacial entre variables, en concreto, la variación direccional de los datos.
- La definición de estructuras dentro del espacio geográfico.

Otra de las características del ‘ESDA’ es que es una disciplina que combina al análisis estadístico de los datos con la utilización de gráficos y cartografía, lo que permite la visualización espacial de éstos. Así, a los contrastes estadísticos (de dependencia y heterogeneidad espacial), se unen representaciones dinámicas que posibilitan reproducir la información georreferenciada y que hacen de estos métodos algo más que simples mapas o gráficos de visualización estática (Anselin & Rey, 2009). Además para Madrid y Ortiz (2005) las técnicas que se utilizan son robustas desde un punto de vista estadístico (al no afectar a sus resultados la existencia de ‘*outliers*’). De acuerdo con Cressie (1993), el ESDA puede ser abordado desde dos puntos de vista, según se trate de un análisis desarrollado por:

- La ‘*Geoestadística’*: El objeto de este tipo de estudios se encuentra, por lo general, en el entorno de las ciencias medioambientales (física, geología, hidrología, etc.) y se centra en una muestra de datos puntuales procedentes de distribuciones geográficas continuas (precipitación atmosférica, humedad de la tierra, altura del océano, etc.)
- La ‘*Econometría Espacial*’: que analiza localizaciones geográficas discretas de puntos o polígonos (provincias, municipios…) Es lo que se denomina ‘perspectiva de retícula’ o ‘Lattice’ y se encuentra mucho más centrado en el estudio de fenómenos socioeconómicos (distribución de la renta, clientes, votantes, etc.)

Finalmente, para Rocco (2012) es importante comprender que cuando se trabaja con datos espaciales se debe tener en cuenta los siguientes aspectos:

- Es imprescindible conocer la posición relativa o absoluta donde se producen los fenómenos que se están analizando: la *‘Georreferenciación’* de todos los elementos.
- Las relaciones que se establecen entre las observaciones espaciales son de carácter *‘multidireccional*’ y ‘*multidimensional*’. Debido a estas dos propiedades de

los datos geográficos, la tendencia espacial sólo puede ser representada correctamente utilizando mapas y gráficos, los cuales son capaces de expresar la evolución de las variables para las distintas unidades geográficas a estudio. Por eso, las técnicas del 'ESDA' son muy parecidas a las herramientas desarrolladas para la representación cartográfica. La diferencia principal entre estas dos técnicas es que mientras que para la Cartografía el mapa es el elemento central, las técnicas del 'ESDA' se apoyan en ésta para trabajar con elementos gráficos propios de la estadística utilizando herramientas de análisis como:

- El Histograma.
- Los Mapas de Polígonos de Voronoi.
- Los Gráficos de Probabilidad Normal Q-Q.
- El Análisis de Tendencia.
- El Gráfico del semivariograma/covarianza.
- El Gráfico de Covarianza Cruzada.
- Los Gráficos de Coordenadas Paralelas.
- Y un largo etcétera.

9.2 La Aplicación de Técnicas Multivariantes Clásicas en Geografía

En los primeros años de desarrollo de la Geografía Cuantitativa (década de los 50), el interés por realizar estudios científicos que se apartaran del tradicional ensayo descriptivo paisajístico realizado hasta entonces, produjo que muchos autores introdujeran las primeras aplicaciones de índices, de coeficientes de correlación y de regresión a sus análisis. Pero, tal y como señala Bradshaw (1983), los primeros trabajos que verdaderamente incorporaron técnicas estadísticas multivariantes a la Geografía estuvieron posibilitados por el desarrollo de la informática y la consecuente difusión de los datos censales informatizados y fueron definidos a partir del empleo de técnicas propias de la **Ecología Factorial**, disciplina que fue iniciada a principios de los años 60 a raíz de los estudios realizados por John Sweeter para las ciudades de Helsinki y Boston, donde postuló lo adecuado de la utilización del 'Análisis Factorial' para extraer y analizar los aspectos socioeconómicos que tenían importancia en la ordenación del medio urbano. Berry (1968) amplió esta idea al proponer la utilización de otros métodos estadísticos como el 'Análisis de Componentes Principales' y el 'Análisis de Cluster' con el mismo objetivo: comprender la 'estructura interna' de las ciudades a partir de sus datos sociales, económicos y demográficos. Pero a finales de esa década, comenzó a producirse un rechazo a estos planteamientos, tanto desde las posturas más humanistas de esta ciencia (como la Geografía Radical) al señalar

que se había perdido la verdadera naturaleza de querer comprender el origen del fenómeno estudiado frente a la obtención de resultados simplemente cuantitativos, como también desde la propia Geografía Cuantitativa al poner en duda lo adecuado de su implantación al utilizar estas técnicas de una manera 'simplista' y sin dar solución a circunstancias geográficas tan importantes como el '*Problema de la Unidad Espacial Modificable*' (PUEM), la influencia de la autocorrelación espacial en los datos y la necesaria adecuación de esas metodologías a la realidad de los registros geográficos medibles. Por eso, en los años 70, comenzaron a ser 'sustituidas' por otras metodologías más adecuadas a sus características y necesidades; de esta manera, aparecen los primeros trabajos que incluyen análisis de series espacio-temporales, el uso de modelos probit y logit para el análisis de datos categóricos, etc.

A partir de entonces, la Geografía ha ido incorporando 'tímidamente' otras técnicas estadísticas como el 'EDA', el 'MDS', el 'DA', el 'AID', el 'PCoA', etc. sin que su aportación al desarrollo como ciencia de la Geografía haya trascendido más allá de la singularidad que adquieren estos trabajos en los cuales son utilizadas, de manera circunstancial, como herramientas de análisis y en la mayoría de las ocasiones como producto de la colaboración con especialistas en Estadística dentro de proyectos que requieren un marcado carácter multidisciplinar (como en los análisis de ordenación territorial). Miller (2010) señala que en la actualidad la mayoría de los avances y la literatura técnica que se están realizando en esta dirección se encuentran ubicados en el ámbito de los GIS, la GISciencia y el Análisis Espacial; y que para el resto de disciplinas dentro de la Geografía, como la denominada 'Ciencia Regional' (que está cobrando una gran importancia en los últimos años), se encuentran con la circunstancia, sin precedentes, de la posibilidad de acceder a una enorme cantidad de datos urbanos, económicos, políticos, sociales, etc. muy detallados y la mayoría de ellos con referenciación espacial y temporal (propios del Big Data), que una vez analizados permitirían adquirir nuevas ideas y conocimientos lo que provocaría un rápido avance en su modelado teórico como disciplina científica. Sin embargo, reflejando lo disertado por Lazer et al. (2009), en la literatura geográfica científica existe poca conciencia sobre los importantes beneficios que ofrece la práctica del enfoque estadístico-computacional para el análisis exploratorio de las grandes bases de datos que conlleve al descubrimiento del conocimiento. La principal razón esgrimida, puede achacarse a la falta de formación y, por lo tanto, de conocimiento de los Geógrafos sobre estas técnicas, cuestión que comparto absolutamente como licenciado en esta ciencia, lo cual constituye un importante reto a resolver en los próximos años. Además, este autor también señala el hecho de que está surgiendo

una nueva ciencia social de carácter computacional (la GISciencia, a la cual hacíamos referencia en capítulos anteriores) basada en la necesidad de recopilar y analizar cantidades masivas de datos provenientes tanto de las acciones provocadas por el comportamiento individual, como de los grupos sociales. El problema es que está desarrollándose casi exclusivamente en el ámbito del sector privado (ligado a empresas como Yahoo o Google) y de organizaciones gubernamentales como la Agencia de Seguridad Nacional de los Estados Unidos ('NSA'), teniendo poca repercusión en las principales revistas científicas, hecho (entre otros) que no está facilitando su adecuado tratamiento académico y, por lo tanto, tampoco la acumulación y difusión de este conocimiento como servicio público y social. Y ésta es una cuestión que adquiere una tremenda importancia, por lo que resulta preciso crear e implantar una doctrina académica adecuada impartida desde los centros universitarios que permita extender esta disciplina al ámbito social, al igual que históricamente se ha realizado con el resto de ciencias (ver capítulo 9.4 e IMAGEN 12. Desarrollo CURRICULAR ACADÉMICO PROPUESTO PARA LA FORMACIÓN EN GISCIENCIA.

Fuente: DIBIASE (2006).

9.3 Estadística Espacial y Geoestadística

Es necesario comenzar este apartado señalando, tal y como concreta Cressie (1993), y recoge de este texto original Giraldo (2002), que la Estadística Clásica, en sus modelos más simples, tiende a considerar muestras aleatorias en las cuales para los fenómenos analizados se presupone independencia y equidistribución. De esta manera, se 'suaviza' la teoría que se desarrolla para establecer los modelos de comportamiento pertinentes. Pero en el mundo real, en muchas ocasiones, esta circunstancia no es válida ya que debemos analizar fenómenos que suceden existiendo una dependencia clara entre ellos. Así, nos encontramos con dos casos específicos dentro de la Estadística que necesitan investigar esta relación entre las observaciones:

- La Estadística que analiza series temporales.
- La Estadística que estudia datos espaciales.

La diferencia entre ambas es que, para esta última, la dependencia entre las observaciones está presente en todas las direcciones y, además, se vuelve más débil según se alejan unas de otras.

Cressie (1993) definió la **Estadística Espacial** como "la rama de la Estadística que abarca teorías y aplicaciones para procesos aleatorios con índices espaciales continuos". Para Giraldo (2002) "es la reunión de un conjunto de metodologías apropiadas para el análisis de datos que corresponden a la medición de variables aleatorias en diversos sitios (puntos del espacio o agregaciones espaciales) de una región". Guyón (2010) señala que esta disciplina estudia fenómenos aleatorios indexados por un conjunto espacial dónde podemos encontrarnos con las siguientes realidades:

- Fenómenos de naturaleza unidimensional.
- Fenómenos de naturaleza bidimensional.
- Fenómenos de naturaleza tridimensional.

Además, siguiendo con este autor, la existencia de diferentes tipos de datos espaciales ha dado lugar a la necesidad de desarrollar técnicas de análisis concretas relacionadas con su naturaleza:

- *'Datos Puntuales':* También denominados *'Procesos Puntuales' ('PP'*, a partir de ahora). Los registros se sitúan en ubicaciones que pueden ser discretas o continuas y su selección no depende del investigador (localización aleatoria). El propósito de las técnicas que analizan este tipo de datos es determinar si su distribución geográfica es debida al azar ('PP' de Poisson). Y si no lo es, definir el modelo de dependencia espacial, la existencia de clusters, etc.
- *'Datos de Red (Lattices)'*: Los datos están localizados en ubicaciones discretas seleccionadas a priori por el investigador (localización determinista). Haining et al. (2000) definen los 'Lattices' como "*aquéllos datos referidos a regiones con formas no regulares en el espacio*". Dominan dos familias de modelos: las Auto-Regresiones Espaciales y los Campos de Markov. Algunos de los objetivos de estudio de esta área son la modelización, la estimación, el control de la correlación espacial y la restauración de imágenes digitales.
- *'Datos Geoestadísticos'*: Los datos están localizados en ubicaciones continuas y son seleccionadas a juicio del investigador (localización determinista). La Geoestadística (que es la rama dentro de la Estadística Espacial que analiza estos datos) se preocupa por la identificación y estimación del modelo para, después, definir la predicción de valores en los lugares no observados. Desde un punto de vista histórico, Matheron a principios de los años 60, desarrolló la metodología 'Kriging', que extendía la teoría de predicción de procesos estocásticos definida por Kolmogorov (1933) a los procesos espaciales ocurridos en un espacio

geográfico continuo (Mandrekar & Masani, 1997). Su desarrollo se llevó a cabo para hacer frente a la necesidad práctica de predecir los rendimientos en las explotaciones mineras, basándose en la interpolación de datos dispersos. Las investigaciones, desde este punto de vista, han ido buscado los métodos más eficientes que proporcionen la mayor información posible para los datos disponibles, utilizando para ello el mejor estadístico que minimice la varianza del error cuadrático medio de estimación (Mateu, 2003).

Incidiendo en este último punto, tenemos que señalar que Matheron (1963) definió la **Geoestadística** como *"la aplicación del formalismo de las variables aleatorias al reconocimiento y estimación de las variables naturales"* y junto a su equipo de la Escuela de Minas de Fontainebleau (Francia), desarrolló dentro de la Estadística Espacial esa disciplina científica, que parte del supuesto de que la continuidad espacial de las variables distribuidas en el espacio tiene una estructura particular que se analiza a través de las dependencias existentes entre ellas. Recogiendo lo señalado por García (2009), dichos fenómenos se caracterizan por la distribución geográfica de una o más variables denominadas '*Variables Regionalizadas'* las cuales son definidas por Giraldo (2002) como "*aquellas variables medidas en el espacio de forma que presentan una estructura de correlación"*. Los objetivos de la Geoestadística son la estimación, la predicción y la simulación (modelización) de estas variables (Warrick & Myers, 1987) y de ello se sirven ciencias como la Geografía, la Geología y cualquier otra disciplina que trabaja con datos recogidos en diferentes locaciones espaciales que necesitan desarrollar modelos que indiquen si existe dependencia espacial entre las mediciones realizadas para las variables analizadas.

Cuando el objetivo es realizar predicciones se trabaja, básicamente, en dos etapas:

- Realizar un análisis estructural, en el cual se describa la correlación entre los puntos ubicados en el espacio.
- Realizar la predicción para los lugares de la región no muestreados por medio de la técnica **Kriging**. Este proceso calcula un promedio ponderado para las observaciones muestrales, donde los pesos asignados a los valores son determinados por la estructura espacial de correlación establecida en la primera etapa y por la propia configuración del muestreo (Petitgas, 1996).

Al igual que ocurre con el resto disciplinas que se sirven de modelos matemáticos que intentan reflejar algún aspecto de la realidad, la Geoestadística en su pretensión de análisis y predicción de la distribución espacial para fenómenos georreferenciados, establece y desarrolla un conjunto de hipótesis hasta obtener unas conclusiones. Y,

finalmente, si éstas se ajustan a lo que observamos en la realidad estimamos que es un modelo válido.

La columna vertebral del análisis geoestadístico, tal y como señala Giraldo (2002), es la determinación de la estructura de autocorrelación entre los datos, la modelización/simulación y su uso en la predicción a través de las técnicas conocidas como el Kriging y el **Cokriging**. La diferencia entre ambas (para las cuales existen distintos métodos), es que mientras que el Kriging utiliza únicamente la correlación espacial para determinar los coeficientes del estimador lineal de predicción, el Cokriging utiliza tanto la correlación espacial, como la correlación existente entre las funciones aleatorias (método que consiste en hacer predicciones espaciales de una variable partiendo de la información que disponemos de ella y de la de algunas variables auxiliares con las cuales está correlacionada espacialmente).

Es aquí donde 'entramos' en el análisis multivariante dentro de esta disciplina ya que los métodos tradicionales de la **Geoestadística Básica Lineal** son convenientes para analizar la estructura espacial de una sola variable a partir de un conjunto de datos, pero con frecuencia es necesario caracterizar las correlaciones espaciales para un número mayor de variables simultáneamente (Einax & Soldt, 1998) y es donde la **Geoestadística Multivariante** proporciona herramientas complementarias para el estudio de varias variables simultáneamente (Tolosana, 2011). Los métodos de análisis de datos multivariantes pueden utilizarse para descubrir relaciones multivariantes entre muestras o elementos y esta información 'latente' puede ser extraída de la matriz de covarianzas. De esta manera, el uso de métodos geoestadísticos multivariantes pretende constituir una herramienta eficaz para superar el problema de los métodos de la estadística multivariante tradicional que ignoran las correlaciones espaciales entre puntos de muestreo y que incluyen una parte importante de la información.

Pardo-Iguzquiza, Chica-Olmo, Rigol-Sánchez, Luque-Espinar y Rodríguez-Galiano (2011) señalan que "es bien conocido que el cokrigeaje es la extensión multivariante del método de interpolación espacial geoestadística conocido como krigeaje". Así, continúan señalando, que para el caso más sencillo de Cokriging, la estimación para una variable regionalizada de interés (variable primaria) se realiza a partir de los valores experimentales de ésta y los de una variable secundaria que está correlacionada con la primera. De esta forma, su valor (en una localización donde no se ha muestreado) se estima mediante la combinación lineal de las dos variables aplicando para ello una ponderación (Oliver, Bosch, Slocum, Kleingold, & Krige, 2002).

En la bibliografía consultada, hemos comprobado que existe confusión entre el Cokriging y la Geoestadística Multivariante ya que ambas técnicas tratan con un conjunto de variables, pero esta última a diferencia de la anterior utiliza métodos clásicos de la Estadística Multivariante como el 'Análisis Factorial' y el 'Análisis de Componentes Principales'. Estos métodos geoestadísticos multivariantes se basan en la teoría de la corregionalización lineal que fue desarrollada por Matheron (1982) y que plasmó en la técnica denominada **Factorial Kriging Analysis** ('FKA') donde la matriz de correlación resultado del análisis geoestadístico realizado sobre los datos se utiliza para detectar los factores correlacionados, que pueden ser atribuidos a fuentes comunes no observables directamente. Yuan y Myers (1994) señalan que el 'FKA' puede ser visto como un caso particular del Cokriging y así lo denominan investigadores como Jiménez-Espinosa y Chica-Olmo (1992) y Castrignanò et al. (2005) denominándolo **Cokrigeaje Factorial**. Esta técnica según Wackernagel (1995), Pardo-Iguzquiza y Dowd (2002) y Demšar, Harris, Brunsdon, Fotheringham y McLoone (2012) consiste en realizar una serie de procesos siguiendo una serie de etapas:

- Análisis de Semivariogramas y Semivariogramas Cruzados estimados a partir de los datos experimentales.
- Ajuste de un modelo para el conjunto de Semivariogramas.
- Generación del Cokriging Factorial para estimar un Factor Individual (una componente individual o una combinación de componentes de cualquiera de las variables corregionalizadas).
- Interpretación del Factor Estimado (Componente).

En relación a otros desarrollos realizados dentro de la Geoestadística Multivariante, tal y como señala Wackernagel (1995) "el Análisis de Componentes Principales es el método de análisis de datos multivariante más ampliamente utilizado debido a la simplicidad de su álgebra y su interpretación directa" y lo traslada al ámbito Geoestadístico denominándolo **Análisis de Componentes Principales Regionalizado** ('ACPR'). Este método se fundamenta en la realización del 'Análisis de Componentes Principales' ('PCA') sobre matrices de corregionalización y la técnica consiste en generar los ejes principales de la forma tradicional, para luego realizar su interpretación en términos de la variabilidad explicada de cada componente obtenida respecto a cada variable original, para finalmente llevar a cabo un análisis geoestadístico a través de la estimación de la función de semivarianza y la aplicación de algún procedimiento Kriging de los datos originales sobre los ejes generados. De esta manera, la interpretación del mapa de predicciones obtenido sobre las

componentes resultantes permite obtener una visión integral del comportamiento conjunto de las variables consideradas dentro del sistema de estudio (Manly, 1994).

En la actualidad la práctica de esta disciplina abarca ámbitos muy diversos como la Minería, la Geología, la Industria Petrolífera, la Meteorología, la Climatología, la Edafología, la Hidrología, la Ecología, la Epidemiología, etc.

Retomando la definición de Cressie (1993) que ofrecíamos al principio de este punto, es necesario señalar que la misma deja abierta la posibilidad de que el fenómeno a analizar provenga o no de la naturaleza, lo que amplía su posible respuesta a los avances de la tecnología; aspecto que podemos extrapolar a este estudio, ya que por su propósito inicial y las técnicas desarrolladas al efecto y fundamentadas en la estimación o predicción (siempre a partir de información escasa) del valor de una variable en una localización determinada, queda vinculada a nuestro objeto de análisis. Y es que aunque partimos de la premisa contraria, es decir tenemos grandes volúmenes de datos, también necesitamos determinar la correlación espacial existente para establecer conclusiones precisas propias del objeto de estudio.

9.4 Sistemas de Información Geográfica y GISciencia

Auspiciado por los avances de la tecnología informática y por los desarrollos sobre los datos con carácter espacial dentro de las Matemáticas y la Estadística, durante los primeros años de la década de los 60 se diseñaron las primeras formas automatizadas de cartografía que, posteriormente, condujeron al desarrollo de los **Sistemas de Información Geográfica** (GIS, de su acrónimo original '*Geographic Information System*'). Recogiendo lo señalado por Bosque (2000) podemos considerar, a modo de resumen, que estos Sistemas son el producto de la unión de varias líneas de trabajo:

- Los Sistemas de Cartografía Asistida por Ordenador.
- El planteamiento teórico-práctico propuesto por McHarg (1995) para la superposición de mapas.
- El desarrollo de la Teledetección y la Fotogrametría.
- Los avances técnicos y metodológicos de la Geografía Cuantitativa.

Aunque para la mayoría de los autores los GIS son un programa de ordenador, en realidad son el resultado de la combinación de tres componentes (Huxhold & Levinsohn, 1995):

- El Hardware.
- El Software.
- Un contexto organizativo apropiado compuesto por un equipo multidisciplinar.

Un GIS, por lo tanto, es un producto informático resultado de la combinación de los tres elementos antes definidos y desarrollado para manejar registros de cualquier naturaleza, con el objeto de facilitar su entendimiento como fenómenos con naturaleza espacial y basados en la gestión de los datos como antecedente necesario para la obtención de la información necesaria para el conocimiento preciso del objeto a analizar (Burrough, McDonnell, & Lloyd, 2013). En este punto es necesario remarcar la diferencia existente entre '*Dato*' e '*Información*' que ya definían Puebla y Gould (1994), y que trataremos detalladamente en el capítulo 9.5, así denominan '*Dato'* a "*la observación que se almacena*" e '*Información'* a "*la respuesta a una pregunta*" y precisamente estos programas son capaces de convertir y transformar esos datos en información generando por ello respuestas al investigador, justo el objetivo que pretendemos con el Big Data y para el cual deberemos utilizar, de manera conveniente, la Estadística Multivariante.

Hay que continuar señalando que los datos geográficos coinciden con los manejados por otras especialidades al poder ser descompuestos en dos elementos que los conforman:

- *La Observación* (o individuo): que constituye una entidad dentro de la realidad sobre la cual se realiza la investigación.
- *La Variable* (o atributo temático): que nos muestra las diferentes respuestas que puede adoptar ese individuo dentro del conjunto.

Pero hay un elemento que diferencia este tipo de datos del resto haciéndolos únicos: *su soporte está localizado en el espacio*, constituyendo este hecho una cuestión esencial dentro del propio enfoque estadístico analítico de la Geografía y de las mismas propiedades de un GIS (Ebdon, 1985). Es por tanto, en la manipulación y en el análisis de datos donde se produce una marcada diferencia con otro tipo de programas que reciben acepciones genéricas tales como Programas 'CAD' ('*Computer Aided Design*') referidos a sistemas de dibujo y diseño asistido, o Programas 'CAE' ('*Computer Assisted Engineering*') especializados en el diseño y cálculo de elementos propios de la Ingeniería Civil. Así, los GIS, a diferencia del resto, nos permiten, mediante la correcta gestión de los registros insertados en las bases de datos, obtener

información. Todo ello, les confiere un objetivo puntual: el ayudar y asistir a la toma de decisiones humanas en cualquiera de los campos ante los cuales se enfrenta un investigador trabajando con esta tecnología. Justo el objetivo que perseguimos, es decir: *cómo hemos de transformar los datos provenientes de una realidad, para obtener información que nos permita entenderla*.

Un programa puede denominarse GIS cuando ofrece, en mayor o menor medida, las siguientes funciones básicas (Juanes P., Juanes J.A., Bajo & Diéguez, 2001):

- Un sistema de representación gráfica, que permita simular la mayoría de las entidades gráficas típicas de la cartografía automática, tales como las líneas, los puntos, los símbolos, las tramas, las imágenes, etc., referenciadas mediante coordenadas geográficas o cartesianas.
- Una base de datos que permita gestionar de forma sencilla tanto conjunta, como separadamente los datos alfanuméricos y gráficos referentes a un espacio territorial.
- Una organización de su base de datos que posibilite relaciones espaciales entre los registros, conocidas como relaciones topológicas.
- Un sistema de acceso selectivo a los datos (por ejemplo, el lenguaje SQL) que permita realizar búsquedas y simulaciones, tanto con los datos gráficos como con los alfanuméricos.
- Un sistema de generación de cartografía automática a partir de consultas y simulaciones.
- Un sistema de generación de documentación alfanumérica formada por listados e informes realizados a partir de las consultas y las simulaciones.
- Un sistema de importación y exportación de los datos y la información generada que permita una adecuada interacción con otras aplicaciones.

El GIS supone la aplicación práctica sobre casos reales de las técnicas matemáticas y estadísticas fundamentadas en la revolución cuantitativa en la Geografía ya explicada en la introducción de este capítulo. Y es que, si en la actualidad existe un área de conocimiento e investigación donde los datos y el tratamiento de éstos tienen importancia dentro de la Geografía y donde las aportaciones de esta disciplina al ámbito científico general están teniendo una trascendencia significativa es en este campo; pero a la vez al manejar esa concepción *'formalista'* del espacio geográfico, donde se priorizan las formas, las relaciones geométricas y topológicas entre los elementos, existe un vacío importante dentro de ellos: *que es el análisis tanto de los procesos causales generadores de estas formas, como de los estudios temporales*.

Actualmente la iniciativa más interesante que intenta dar respuesta a esta problemática es la Ciencia de la Información Geográfica (o GISciencia a partir de ahora) que Bosque (2005) define como el "*cuerpo de conocimiento que pretende el estudio, la investigación y el desarrollo de los conceptos teóricos, los algoritmos matemáticos, los programas informáticos, los instrumentos físicos, las bases de datos, las nuevas formas de uso y la búsqueda de nuevos campos de aplicación, en relación a las tecnologías de la información geográfica*".

Así, la GISciencia constituye un fundamento conceptual y teórico más profundo que el uso que venía haciéndose de los GIS, resultado de su propia evolución como campo unificado que estudia estos Sistemas de Información y los métodos de la Estadística Espacial considerándolos como un apoyo teórico holístico esencial para el análisis y la comprensión del comportamiento de los datos espaciales (Goodchild & Haining, 2005). Por lo tanto, la evolución de la GISciencia debe mucho a los desarrollos experimentados en los GIS y al dominio del análisis de los datos espaciales, cuestiones que han mejorado su capacidad técnica para manejar datos de naturaleza geográfica acercándolos a las necesidades diarias de la sociedad (Duckham, Goodchild, & Worboys, 2003).

Y es que bajo el acrónimo GIS se esconden dos aspectos distintos pero íntimamente relacionados: por una lado, el '*Sistema de Información Geográfica*' (que podría definirse en castellano como '*GIS-Sistema*') y por otro lado la '*Ciencia de la Información Geográfica*' (que denominaríamos '*GISciencia'*, tal y como hemos señalado anteriormente). Sus traducciones, con sus acrónimos, en castellano podrían ser las siguientes:

- 'Sistema de Información Geográfica' ('GIS').
- '*Ciencia de Información Geográfica*' ('*CIG*'): Que en el ámbito académico francés se denominada, a menudo, '*Geomática'*.

Sin embargo, es frecuente que el término GIS se utilice en la literatura de forma indiscriminada y sea el contexto el que tenga que orientar al propio investigador sobre la correcta interpretación de la palabra. Schuurman (2004) señala que el '*GIS-Sistema es el medio por el que se expresan y materializan las ideas de la GISciencia*'. Para aclararlo propone un ejemplo: El GIS-Sistema es un programa que tiene implementado un algoritmo de cálculo sobre un modelo que el usuario da por válido (el GIS-Sistema actúa como una caja negra). La GISciencia es la encargada de construir y validar ese modelo (construye esa caja negra) y evalúa su campo de aplicación, su utilidad y

repercusión. Para una completa revisión del concepto teórico-práctico ver DiBiase (2006).

9.5 GISciencia en Big Data

A mayores de lo señalado en el capítulo anterior acerca de los GIS y la GISciencia y enlazándolo con lo expuesto en el capítulo 7 en el cual hemos definido el Big Data, es necesario, por su importancia conceptual, iniciar este apartado volviendo a remarcar que, tal y como definen Moreno y Cañada (2006), un GIS no es un programa cartográfico, ni un software de tipo CAD; aunque genere mapas y tenga funcionalidades que le permiten 'dibujar', lo específico de estos sistemas reside en su capacidad para almacenar y gestionar grandes masas de información georreferenciada y su potencia para analizar la misma, lo cual los hace idóneos como herramientas vinculadas a la toma de decisiones. Un GIS implica un modelo del territorio, mediante el cual se hace una representación virtual de la realidad geográfica lo cual requiere que los datos seleccionados sean registros parciales cuya 'criba' se realiza teniendo en cuenta el objetivo que se persigue (lo cual enlaza con la primera de las Vs ya expuestas para el Big Data: la necesidad de 'Veracidad' en los registros, cuestión imprescindible de solventar utilizando para ello las técnicas estadísticas adecuadas que permitan manejar de manera coherente los datos contenidos). Así, este modelo descompone en partes el todo, y cada una de esas partes es un fenómeno con naturaleza identitaria que ha de ser sumado al conjunto de datos para generar la información objetivo. La manera de organizar y almacenar en un GIS cada una de estas partes es a través de capas (o '*layers*') que contienen los datos referidos a un aspecto de la realidad, lo que posteriormente permite relacionar y combinar de nuevo varias capas entre sí para mostrar aspectos fundamentales para el estudio que la complejidad de la realidad impide percibir directamente. Esta idea es básica para enfrentar la solución del complejo armazón del Big Data.

Goodchild y Haining (2005) definieron una serie de paradigmas a los cuales se tuvo que 'enfrentar' la ciencia para el desarrollo de los GIS. Haciendo analogía (al igual que la hicimos para la Estadística Multivariante en el capítulo 8), estas mismas circunstancias nos las encontramos en la actualidad para el tratamiento del Big Data:

1. Las dificultades para obtener mediciones precisas de los mapas, así como la simplicidad para obtener dichas medidas por medio de una representación digital: Complejidad de los datos y necesidad de simplificación y modelización.

2. La *necesidad de analizar datos con múltiples tipos de características* (análisis de densidades, de rutas, de lugares, censos poblacionales, etc.) y la relación entre ellos en proyectos de grandes dimensiones: *Variedad de los datos.*
3. La *necesidad de integrar múltiples niveles de información*. Por ejemplo, al valorar los posibles impactos territoriales en proyectos de desarrollo regional donde confluyen enormes conjuntos de información distribuidos en múltiples capas: *Dificultad en la integración de los datos disponibles.*
4. Los problemas para gestionar gran cantidad de registros. Por ejemplo, en el análisis de los datos censales: Manejo de grandes volúmenes de datos.

En el pasado, el factor limitante para el desarrollo e implantación como solución masiva de los GIS (primero) y la GISciencia (en el presente) fueron los datos y es que esta disciplina requiere de información detallada, tanto de los atributos, como de la ubicación del fenómeno a analizar (Farmer & Pozdnoukhov, 2012). Moldes (1995), Bosque (2000) y otros autores, señalan que *"el primer paso imprescindible para la generación de un modelo GIS es la disponibilidad de datos geográficos, que serán introducidos y procesados en el ordenador, con el objetivo de obtener información"*. Y coinciden en que muchas veces este aspecto se convertía en la tarea más difícil para el analista debido a su carestía. La escasa disponibilidad de información digital (la única admitida de forma directa por los GIS) fue, sin lugar a dudas hasta los inicios del siglo XXI, el principal obstáculo para el investigador; pero también lo fueron la escasa posibilidad de obtener datos analógicos. Por todo ello, la cuestión previa era dónde buscar y qué procedimientos utilizar para disponer de los registros que tenían que ser introducidos en un GIS. Pero ésta ya es una realidad pasada, y es que en contraposición con lo ocurrrido en los años ochenta y noventa donde, a pesar del rápido desarrollo producido dentro del campo teórico del GIS y de la Geografía Cuantitativa, quedaron muchas teorías e ideas sin poder ser comprobadas de manera fehaciente debido a la falta la cantidad de datos necesaria, en la actualidad (para la GISciencia al igual que para otras disciplinas científicas) se está produciendo el problema opuesto: el rápido ritmo de captación de un conjunto extraordinariamente grande de datos, excede el alcance de sus métodos y de su actual marco teórico (Farmer & Pozdnoukhov, 2012). En un mundo donde los registros y, por supuesto, la información obtenida de ellos eran escasos nos habíamos acostumbrado a razonar utilizando hipótesis: se proponía una idea, se confirmaba con unas pocas referencias recogidas al efecto y, gracias a este proceso, determinábamos la causalidad, el porqué de los fenómenos.

Con el Big Data la causalidad parece perder terreno a favor de la correlación, basada en cuantificar la relación estadística entre dos conjuntos de datos (nos damos cuenta de que cuando algo sucede, muy probablemente sucede otra cosa). Esta nueva manera general de explicar y comprender el mundo supone poner en crisis los fundamentos de toda la cultura científica tradicional. Históricamente nos indujeron la necesidad de comprender y conocer los motivos que provocaba la existencia de un fenómeno. Frente a este enfoque causístico tradicional, la nueva propuesta que parecen abrazar muchos autores pasa por aceptar el desorden y la imprecisión del mundo para ser capaces de hacer predicciones. Señalan que, mientras la causalidad nos ayudaba a entender lo que sucedió en el pasado, la correlación nos lleva a poder, en gran medida, predecir el futuro. El cambio cuantitativo que nos trae Big Data, por lo tanto, se convierte en una transformación cualitativa y radical (Mayer-Schönberger & Cukier, 2013). Por lo tanto, en la actualidad, la necesaria adopción y adecuación a la realidad que origina este fenómeno dentro del estudio geográfico y del análisis científico en general, presupone para muchos autores cambiar radicalmente la manera de hacer ciencia que ya mencionábamos al principio del capítulo 7, asumiendo su desarrollo dentro de un proceso abductivo donde la correlación entre los datos señala la causa y el efecto. Así, manejando de manera correcta esta gran cantidad de datos, con un objetivo concreto, permitirá concluir a los analistas y científicos que cuando un fenómeno se produce se da otro hecho o conjunto de hechos, pero el riesgo aparece porque podemos comenzar a ignorar la causa (el cómo se produce). Saber tanto, puede tener su precio: ignorar cómo lo sabemos o cómo se produce exactamente un fenómeno, sencillamente lo sabemos, diluyéndose el vínculo causal en pro del correlacional (pautas y correlaciones por encima de causalidades). Y es que hasta ahora, las investigaciones científicas que perseguían esta causalidad eran complejas y costosas y por lo tanto se comienza a plantear el Big Data (y la búsqueda de la correlación en sus datos) como la gran alternativa a esta metodología. Tal y como señalan Zhou et al. (2014) una de las ideas erróneas más importantes que ha aparecido vinculada a esta realidad es que la 'correlación es suficiente' (*'correlation is enough*') y esta afirmación está presente incluso en algunas de los monografías más populares sobre este fenómeno, incluyendo el mediático libro de Mayer-Schönberger y Cukier (2013) citado anteriormente, donde llegan a postular esa idea incluso como una necesidad. Y es que descubrir la causalidad representa analizar y entender de manera minuciosa y pormenorizada los datos, lo cual generalmente es muy complicado para determinados contextos, pero creemos necesario enfatizar que la correlación se encuentra lejos de ser suficiente y que el papel de la causalidad no puede ser sustituido por el de la correlación. La razón radica en el hecho de que se invierte en el

análisis de datos porque se quiere obtener información útil para tomar las decisiones más sabias y adecuadas, mientras que el abuso de la correlación puede llegar a ser engañosa o incluso desastrosa. Por lo tanto y por todo expuesto anteriormente, el científico y el analista no deben nunca olvidar el análisis del porqué, de ahí lo necesario de la GISciencia que ha de modelizar adecuadamente la realidad para dar lugar al conocimiento científico adecuado de la misma (Farmer & Pozdnoukhov, 2012). Un ejemplo de esta propuesta académica la podemos analizar en la IMAGEN 12 (DiBiase, 2006).

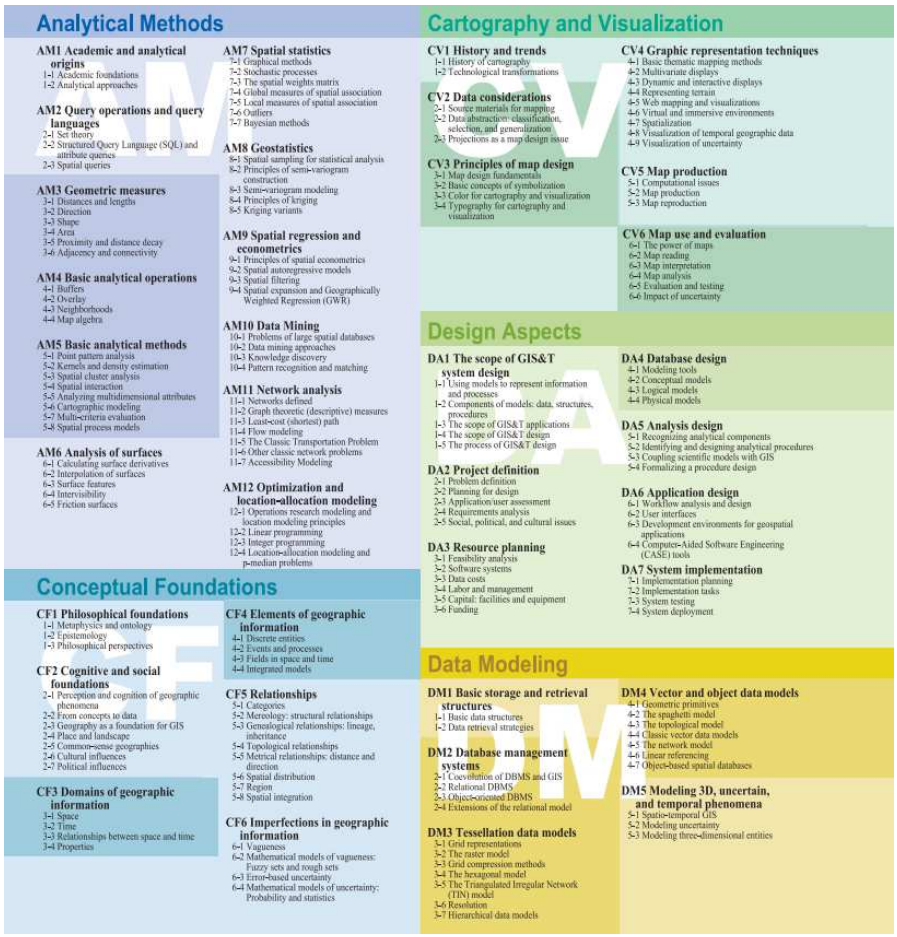

IMAGEN 12. Desarrollo curricular académico propuesto para la formación en GISciencia.

Fuente: DiBiase (2006).

Además, más allá de la aplicación de técnicas cuantitativas sobre espacios físicos, es necesario definir la importancia que tiene para la Geografía el conocimiento de este 'espacio virtual' el cual está compuesto exclusivamente por datos generados antrópicamente. Por lo tanto, y en relación a esta realidad, Big Data define una nueva forma de pensar, de observar el mundo, que se basa en la enorme capacidad que existe en la actualidad de obtener y analizar datos para establecer conclusiones. Pero para llegar a este conocimiento, en estos momentos, según Feller et al. (2011) dada

la importancia de procesar toda la información, el almacenamiento y la velocidad de búsqueda no constituyen los únicos retos para la Estadística y la GISciencia (así como para el resto de disciplinas científicas). Así, la capacidad para analizar mucha información en tiempo real es fundamental y se convierte en la verdadera ventaja competitiva (los datos pueden llegar a constituir sabiduría). Se habla de la 'Pirámide del Conocimiento' o la '*Jerarquía DIKW*' (acrónimo en inglés de '*Data'* + '*Information'* + '*Knowledge'* + '*Wisdom'*), donde en su nivel más básico se encuentran los '*Datos'*. Al dotar de contexto a un conjunto de datos, se obtiene '*Información'*. Y esta información será '*Conocimiento'*, sólo si se sabe cómo utilizarla. Por último, la '*Sabiduría'* (también señalada como la '*Inteligencia'* en algunos textos) responde a porqué se está utilizando (Allee, 2002).

Recogiendo lo señalado por Davenport y Prusak (1998), a continuación exponemos las definiciones para estos términos:

- '**Dato'** ('*Data'*): es el componente básico a partir del cual los procesos y los elementos se relacionan. Pueden hacer referencia a un hecho tangible o intangible, subjetivo u objetivo. Los 'datos' aparecen una vez que se ha superado el proceso de estar atentos y conscientes de que 'algo' está pasando.
- '**Información'** ('*Information*'): es 'algo' que se genera a partir de los 'datos' y que les otorga una 'forma distinta'. Proporciona a esos 'datos' un sentido de utilidad y para ello, previamente, es necesario encontrar sentido, coherencia, contexto, significado y forma a los mismos.
- '**Conocimiento'** ('*Knowledge'*): Cuando hacemos uso de la 'información', adquirimos 'conocimiento' y éste nos permite construir teorías e hipótesis que luego podrán ser verificadas. Es una mezcla de 'información', 'experiencia', 'valores' y 'saber hacer' ('Kwon-How') que sirve como marco para la incorporación de nuevas experiencias y es útil para la acción.
- '**Sabiduría'** ('*Wisdom'*): Se entiende como un 'don' para descubrir y discernir si los 'conocimientos' que se han adquirido, las 'experiencias' que se tienen, así como el 'entendimiento' que se ha logrado, son benignos o no, o en el caso práctico, si hay algo de valor o incluso de novedad o innovación (de aporte), después de todo lo realizado (cognitivamente hablando).

En la IMAGEN 13 tenemos una representación esquemática en la cual se describe, de manera muy sencilla, el proceso de transformación del '*Dato'* en '*Conocimiento*' (obviamos el concepto 'sabiduría' en este análisis, por su absoluta impronta humana).

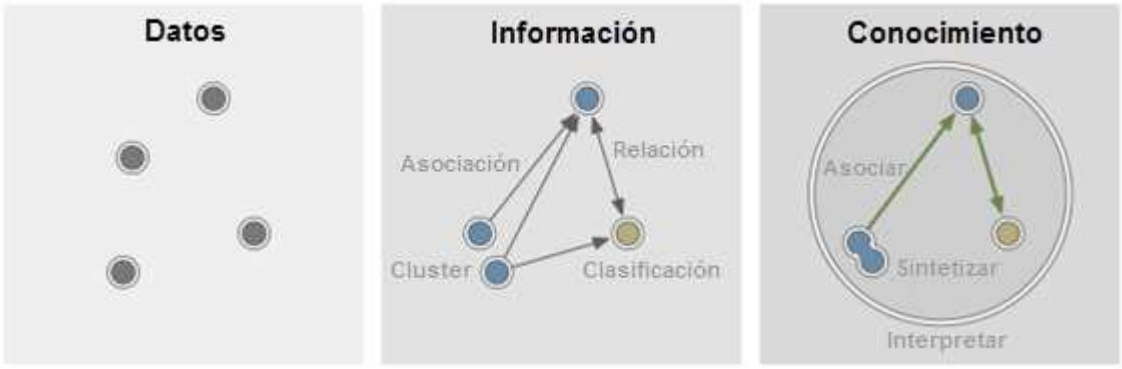

IMAGEN 13. Definición gráfica de los conceptos vinculados a la 'pirámide del conocimiento' Fuente: Modificado de Diakopoulus (2011).

Y precisamente es en este aspecto donde la GISciencia deberá aportar la metodología adecuada en Big Data para, una vez tratados los 'Datos' desde un punto de vista estadístico y ser transformados en 'Información', obtener el 'Conocimiento' necesario para poder actuar en consecuencia sobre un fenómeno observado utilizando para ello el máximo rigor científico (ver **IMAGEN 14**).

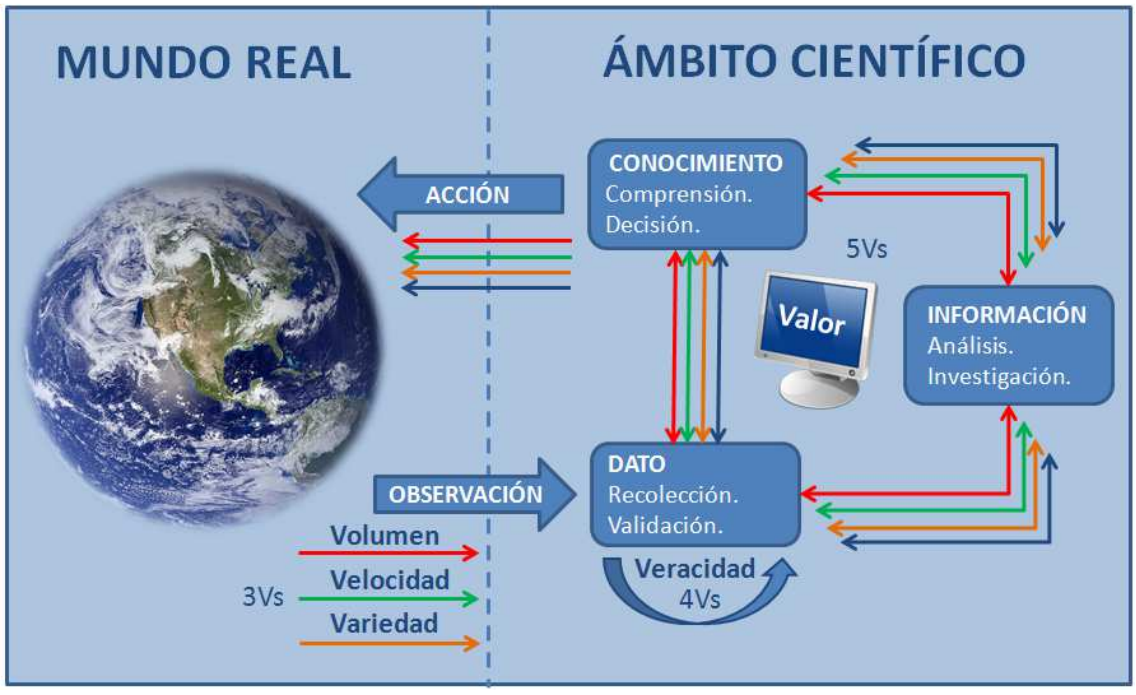

IMAGEN 14. Del '*Dato*' al '*Conocimiento*' en Big Data. Fuente: Elaboración propia.

10 CONCLUSIONES

1. **Big Data** es un fenómeno que surge en 2011 como producto de la última fase de desarrollo de las '**TIC**' (consecuencia de las mejoras exponenciales que ha sufrido el hardware y el software desde finales de la década de los 60) y del impulso al '***Open Data***' (cambio de mentalidad respecto a la apertura en la tenencia de los datos), conformando una realidad y una revolución que en la actualidad afecta a todos los ámbitos de nuestra sociedad.
2. **Cinco características** lo definen y la ciencia Estadística ha de dar respuesta a cada una estas dimensiones:
    - '***Veracidad'***: implantando soluciones que extraigan del conjunto de datos aquellos registros que verdaderamente aporten valor. Importancia de los '*outliers*'.
    - '***Velocidad'***: desarrollando algoritmos que permitan el procesamiento en '*streaming*'.
    - '***Volumen'***: mediante técnicas que permitan reducir la dimensión original de los datos y conseguir su clasificación.
    - '***Variedad'***: adoptando métodos que sean capaces de tratar con registros de diversa naturaleza.
    - '***Valor'***: utilizando desarrollos que posibiliten predecir comportamientos y tomar decisiones sobre los datos (obtención de conocimiento).
3. Es evidente que la estadística univariante no puede hacer frente a las necesidades del Big Data y que además, tal y como hemos comprobado en la bibliografía consultada, las **técnicas multivariantes clásicas** son utilizadas pero con importantes limitaciones.
4. Existe **confusión en la bibliografía científica** entre Big Data y grandes matrices de datos ('Large Data Set'). Como ya hemos visto a lo largo del capítulo dedicado a la Estadística Multivariante, tratar enormes conjuntos registros donde se recoge información de millones de observaciones es sinónimo de Big Data en este ámbito. Los científicos siguen confundiendo este hecho con el verdadero fenómeno que estamos analizando (cuestión que parece estar mucho más clara dentro del ámbito empresarial y gubernamental).
5. Hemos dejado en mano de las **grandes empresas** los grandes fenómenos informáticos y Big Data es otro hecho, pero al trascender de este ámbito tiene importantes repercusiones en la propia sociedad. *Las soluciones se están aportando mayormente desde el ámbito empresarial donde ya existen desarrollos que se encargan de gestionar, almacenar y procesar los registros*

de Big Data. La clave es 'cómo analizamos la misma' (¿dicotomía correlación-causalidad?) y la necesidad de una profunda alianza entre el ámbito académico y el empresarial (que provocará un incremento exponencial en I+D+i).

6. Big Data está dando lugar a una **revolución en el ámbito científico** ('*Big Ciencia*'). Las distintas disciplinas, para dar respuesta a los fenómenos que analizan (y que ya están inmersos en este fenómeno), están adoptando irremediablemente herramientas, antes auxiliares (informática y estadística), que comienzan a integrarse como parte fundamental de su núcleo teorético. La explosión de los datos que manejan es una realidad y puede postularse como su principal factor de evolución a disciplinas científicas 'más maduras' (al formar parte activa dentro de su investigación experimental base los enfoques matemáticos).
7. Un ejemplo claro de lo anterior es la Geografía. La '*localización'* como atributo imprescindible para todos los registros en Big Data (IoT), provoca la necesidad de determinar nuevos objetivos y la aparición de nuevas oportunidades para esta ciencia, que requieren el desarrollo de un cuerpo disciplinar capaz de amoldarse a las características del Big Data: la **GISciencia**, la cual debe postularse como el eje vertebrador del resto de Ciencias Sociales por el rol central que tiene en la actualidad el espacio geográfico para las mismas; asumiendo para ello, el diseño de un plan curricular que abarque aspectos verticales tan necesarios donde se integren metodologías multidisciplinares como la Informática y la Estadística.
8. La dispersión que supondría acometer el estudio de Big Data desde cada uno de estos puntos de vista puede suponer un gran error y puede convertirse en la última oportunidad para avanzar en el estudio de las interrelaciones resultantes de la investigación estadística dentro del ámbito científico general con el objeto de crear teorías, herramientas y métodos relevantes que sean útiles en múltiples dominios de la investigación científica siendo necesario crear, para ello, un cuerpo o núcleo académico que dé respuesta a esta realidad. La historia de la Estadística ha demostrado en multitud de ocasiones (tal y como hemos comprobado a lo largo del Máster), que una técnica concreta y su teoría pueden desarrollarse en el ámbito de una disciplina específica para unos datos muy determinados y, posteriormente, ser 'recogida' para su aplicación en otros ámbitos muy distintos del original. Postulamos, por lo tanto, la necesidad del desarrollo del cada vez más requerido espacio para **la Estadística como ciencia de análisis central** dentro del resto de disciplinas científicas.

9. Big Data requiere **desarrollar herramientas y habilidades analíticas** para convertir sus datos en conocimiento. Las organizaciones han de hacer frente a una creciente variedad de herramientas, al mismo tiempo que han de hacer frente a la escasez de habilidades analíticas. La eficacia de Big Data depende de abordar esta importante laguna, centrando la atención en el desarrollo profesional de los analistas internos familiarizados con los retos y los procesos de una organización y, al mismo tiempo, dentro del mundo académico, las universidades y los propios individuos (independientemente de su formación) están obligados a desarrollar esta habilidad.

11 BIBLIOGRAFÍA

Aflalo, Y., & Kimmel, R. (2013). Spectral Multidimensional Scaling. *Proceedings of the National Academy of Sciences*, 110(45), 18052–18057. doi:10.1073/pnas.1308708110

Agarwal, A., Phillips, J. M., & Venkatasubramanian, S. (2010). Universal Multidimensional Scaling. En ACM (Ed.) Procedente del 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1149-1158. Washington DC, USA.

Allee, V. (2002). The Future of Knowledge. Increasing Prosperity Through Value Networks. Butterworth-Heinemann. Business Books.

Anselin, L., & Rey, S. J. (2009). *Perspectives on Spatial Data Analysis*. Berlin Heidelberg. Springer.

Arango, J., & Van Vleck, L. D. (2002). Size of beef cows: early ideas, new developments. *Faculty Papers and Publications in Animal Science*, 237.

Bae, S. Qiu, J., & Fox, G. (2012). High performance multidimensional scaling for large high-dimensional data visualization. *IEEE Transaction of Parallel and Distributed System*. 0(0), 1-14.

Bai, J., & Ng, S. (2008). Large Dimensional Factor Analysis. *Foundations and Trends in Econometrics*, *3*(2), 89–163. doi:10.1561/0800000002

Banas, K., Banas, A. M., Gajda, M., Kwiatek, W. M., Pawlicki, B., & Breese, M. B. H. (2013). Analysis of synchrotron radiation induced X-ray emission spectra with R environment. Procedente del 11th International School and Symposium on Synchrotron Radiation in Natural Science (ISSRNS), *93*(0), 82-86. doi:10.1016/j.radphyschem.2013.04.026

Benzecri, J.P. (1973). Analyse des Donnés. Tôme 1: La Classification. Tôme 2: L'Analyse des Correspondances. Paris: Dunod.

Berry, B. (1968). *Spatial Analysis. A reader in Statistical Geography*. Englewood Cliffs: Prentice-Hall.

Blasius, J., & Greenacre, M. (2014). *Visualization and Verbalization of Data*. Boca Raton: Chapman and Hall/CRC

Boisier, S. (2001). Sociedad del conocimiento, conocimiento social y gestión territorial. *Interações*, *2*(3), 9–28.

Bosque Sendra, J. (2000). *Sistemas de Información Geográfica*. Madrid: Rialp.

Bosque Sendra, J. (2001). Planificación y Gestión del Territorio. De los SIG a los Sistemas de ayuda a la decisión espacial. *El Campo de las Ciencias y las Artes*, (138), 137–174.

Bosque Sendra, J. (2005). Espacio Geográfico y Ciencias sociales: nuevas propuestas para el estudio del territorio. *Investigaciones Regionales*, (6), 203–224. Recuperado de http://dialnet.unirioja.es/descarga/articulo/2124771.pdf

Bradshaw, R. (1983). El futuro de la Geografía Cuantitativa. Ponencia presentada en el curso: Geografía Teórica y Cuantitativa. Concepto y Métodos. Universidad de Oviedo, Oviedo, España.

Buchholtz, S., Bukowski, M., & Śniegocki, A. (2014). *Big and Open Data in Europe. A growth engine or a missed opportunity?* Varsovia. Warsaw Institute for Economic Studies (WISE Institute).

Bunge, W. (1960). *Theoretical Geography*. Lund: University Microfilms. Recuperado de http://books.google.es/books?id=3jJ_AAAAMAAJ

Burrough, P. A., McDonnell, R. A., & Lloyd, C. D. (2013). *Principles of Geographical Information Systems*. Oxford: OUP.

Buza, K., Nagy, G. I., & Nanopoulos, A. (2014). Storage-optimizing clustering algorithms for high-dimensional tick data. *Expert Systems with Applications*, *41*(9), 4148-4157. doi:10.1016/j.eswa.2013.12.046

Cai, D., He, X., & Han, J. (2008). SRDA: An Efficient Algorithm for Large-Scale Discriminant Analysis. *IEEE Transactions on Knowledge and Data Engineering*, *20*(1), 1-12. doi:10.1109/TKDE.2007.190669

Capel, H., & Sáez, H. C. (1985). *Geografía humana y Ciencias Sociales*. Barcelona: Montesinos Editor, S.A.

Capel, H., & Urteaga, L. (1982). *Las nuevas geografías*. Madrid: Salvat.

Carnés, N. (2014). Cifras de Internet en 2013. [Mensaje en un blog]. Recuperado de http://nachocarnes.wordpress.com/2014/01/28/cifras-de-internet-en-2013/

Carroll, J. D., & Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of “Eckart-Young” decomposition. *Psychometrika*, *35*(3), 283–319. doi:10.1007/BF02310791

Castrignanò, A., Cherubini, C., Giasi, C. I., Castore, M., Di Mucci, G., & Molinari, M. (2005). Using multivariate geostatistics for describing spatial relationships among some soil properties. Procedente de *ISTRO Conference,* Brno, Czech Republic.

Chakradeo, S., Reaves, B., Traynor, P., & Enck, W. (2013). Mast: triage for market-scale mobile malware analysis. Procedente de *Sixth ACM conference on Security and privacy in wireless and mobile networks* (pp. 13–24). Washington DC, USA.

Chasco Yrigoyen, C. (2010). Métodos gráficos del análisis exploratorio de datos espaciales. Instituto LR Klein, Departamento de Economía Aplicada, Universidad Autónoma de Madrid. Disponible en: http://www.Asepelt.org/ficheros/File/Anales/2003%2020Almeria/asepeltPDF/93.PDF

Chen, C. L., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*, 314–347. doi:10.1016/j.ins.2014.01.015

Ch’ng, E. (2014). The Value of Using Big Data Technologies in Computational Social Science. *ArXiv* reimpreso *arXiv:1408.3170*. Recuperado de http://arxiv.org/abs/1408.3170

Chui M., Löffler, M., & Roberts, R. (2010). The Internet of Things. Recuperado 20 Agosto de 2014 http://www.mckinsey.com/insights/high_tech_telecoms_internet/the_internet_of_things

Commandeur, J. J. F., & Heiser, W. J. (1993). *Mathematical derivations in the proximity scaling (PROXSCAL) of symmetric data matrices*. Leiden: Department of Data Theory, University of Leiden.

Cressie, N. A. C. (1993). *Statistics for spatial data*. New York: John Wiley & Sons, Ltd.

Cuadras, C. M. (2012). *Nuevos Métodos de Análisis Multivariante*. Barcelona: CMC Editions.

Cuadras, C. M., Fortiana, J., & Oliva, F. (1997). The Proximity of an Individual to a Population with Applications in Discriminant Analysis. *Journal of Classification*, *14*(1), 117–136. doi:10.1007/s003579900006

D' Aspremont, A., Bach, F., & Ghaoui, L. E. (2008). Optimal solutions for Sparse Principal Component Analysis. *The Journal of Machine Learning Research*, *9*, 1269-1294.

Davenport, T. H., & Prusak, L. (1998). *Working Knowledge: How Organizations Manage what They Know*. New York: Harvard Business School Press.

Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A. S., & McLoone, S. (2012). Principal Component Analysis on Spatial Data: An Overview. *Annals of the Association of American Geographers*, *103*(1), 106-128. doi:10.1080/00045608.2012.689236

De Silva, V., & Tenenbaum, J. B. (2004). *Sparse Multidimensional Scaling using landmark points*. Stanford: Stanford University.

Diakopoulus N. (2012). Data, Information, Knowledge Visualization. Recuperado de http://www.nickdiakopoulos.com/2011/12/16/data-information-knowledge-visualization/

DiBiase, D. (2006). *Geographic information science and technology body of knowledge*. Washington, D.C.: Association of American Geographers.

Dopazo, J., & Carazo, J. M. (1997). Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *Journal of Molecular Evolution*, *44*(2), 226-233. Recuperado de http://link.springer.com/article/10.1007/PL00006139

Dopazo, J., Zanders, E., Dragoni, I., Amphlett, G., & Falciani, F. (2001). Methods and approaches in the analysis of gene expression data. *Gene Expression Technologies*, *250*(1–2), 93–112. doi:10.1016/S0022-1759(01)00307-6

Duckham, M., Goodchild, M. F., & Worboys, M. (2003). *Foundations of geographic information science*. New York: Taylor & Francis.

Du, Q., & Nekovei, R. (2005). Implementation of real-time constrained linear discriminant analysis to remote sensing image classification. *Pattern Recognition*, *38*(4), 459–471. doi:10.1016/j.patcog.2004.09.008

Ebdon, D. (1985). Statistics in Geography: A Practical Approach - Revised with 17 Programs. John Wiley & Sons, Ltd.

Edwards, K. L., Austin, M. A., Newman, B., Mayer, E., Krauss, R. M., & Selby, J. V. (1994). Multivariate analysis of the insulin resistance syndrome in women. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *14*(12), 1940–1945. doi:10.1161/01.ATV.14.12.1940

Einax, J. W., & Soldt, U. (1998). Multivariate Geostatistical Analysis of soil contaminations. *Fresenius' Journal of Analytical Chemistry*, *361*(1), 10-14.

Fan, J., & Liu, H. (2013). Statistical analysis of big data on pharmacogenomics. *Advanced Drug Delivery Reviews*, *65*(7), 987-1000.

Farmer, C. J., & Pozdnoukhov, A. (2012). Building streaming GIScience de context, theory, and intelligence. *Proceedings of the Workshop on GIScience in the Big Data Age. Columbus, Ohio*.

Fellenberg, K., Hauser, N. C., Brors, B., Neutzner, A., Hoheisel, J. D., & Vingron, M. (2001). Correspondence analysis applied to microarray data. *Proceedings of the National Academy of Sciences*, *98*(19), 10781-10786.

Feller, G., Gaudiano, P., González, J., Hirshberg, P., Green, E., Horn, P.,… Lok Yoon, J. (2011). *El Internet de las Cosas. En un mundo conectado de objetos inteligentes*. Fundación de la Innovación Bankinter.

Fernández, V., Leyton, J., & González, A. *Cloudcomputing*. Valparaíso: Universidad Técnica Federico Santa María.

Filella, M., Pomian-Srzednicki, I., & Nirel, P. M. (2014). Development of a powerful approach for classification of surface waters by geochemical signature. *Water Research*, *50*, 221-228. doi:10.1016/j.watres.2013.11.046

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, *7*, 179-188.

Fleites, F. C., Ha, H.-Y., Yang, Y., & Chen, S.-C. (2014). Large-Scale Correlation-Based Semantic Classification Using MapReduce. *Cloud Computing and Digital Media: Fundamentals, Techniques, and Applications*, 169.

Furht, B., & Escalante, A. (2010). *Handbook of Cloud Computing*. New York. Springer. Recuperado de http://books.google.es/books?id=jLNGCPs6rr4C

Ganapathi, A., Chen, Y., Fox, A., Katz, R., & Patterson, D. (2010). Statistics-driven workload modeling for the cloud. En *Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference* (87-92). Fukuojama, Japón. Recuperado de http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5452742

Ganapathi, A., Kuno, H., Dayal, U., Wiener, J. L., Fox, A., Jordan, M. I., & Patterson, D. (2009). Predicting multiple metrics for queries: Better decisions enabled by machine learning. En *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference* (592-603). Los Ángeles. EEUU. Recuperado de http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4812438

García Bastante, F. (2009). *Fundamentos de Geoestadística*. E.T.S. Enxeñeria de Minas. Universidad de Vigo.

García Cantero, J., & De Prado Alonso, I. (2012). Big Data: Cómo la avalancha de datos se ha convertido en un importante beneficio. TIC Beat.

Gifi, A. (1990). Nonlinear Multivariate Analysis. West Sussex: John Wiley & Sons, Ltd.

Gilder, G. (2006). The Information Factories. Obtenida en Octubre de 2006, de http://archive.wired.com/wired/archive/14.10/cloudware.html

Giraldo, H. R. (2002). *Introducción a la Geoestadística: Teoría y aplicación*. Disertación doctoral no publicada. Bogotá DC: Universidad Nacional de Colombia.

Gómez Martínez, C. (2013). Procesamiento de grandes volúmenes de datos en entornos Cloud Computing utilizando Hadoop MapReduce. Disertación doctoral no publicada. Almería: Universidad de Almería.

Góngora Gómez, J.L. (2007). Dimensión espacial de las remesas de migrantes internacionales en México. México D.C. CRIM-UNAM / UAEM.

Goodchild, M. F., & Haining, R. P. (2005). SIG y análisis espacial de datos: perspectivas convergentes. *Investigaciones Regionales*, (6), 175-202.

Gower, J. C. (1966). Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. *Biometrika*, *53*(3/4), 325. doi:10.2307/2333639

Gower, J. C., Lubbe, S. G., & Roux, N. J. L. (2011). *Understanding Biplots*. West Sussex: John Wiley & Sons, Ltd. Recuperado de http://books.google.es/books?id=66gQCi5JOKYC

Gras, J. A. (1996). Métodos y técnicas avanzadas de análisis de datos en ciencias del comportamiento. Barcelona: Universitat de Barcelona.

Greenacre, M. (2008). La práctica del análisis de correspondencias. Bilbao: BBVA Foundation.

Greenfield, A. (2010). *Everyware: The Dawning Age of Ubiquitous Computing.* Berkeley: Pearson Education.

Guyón, X. (2010). Modelación para la estadística espacial. *Revista de Investigación Operacional*, *31*(1), 1-33.

Haining, R., Wise, S., & Signoretta, P. (2000). Providing scientific visualization for spatial data analysis: Criteria and an assessment of SAGE. *Journal of Geographical Systems*, *2*(2), 121-140.

Hallin, M., & Lippi, M. (2013). Factor models in high-dimensional time series. A time-domain approach. *Stochastic Processes and Their Applications*, *123*(7), 2678-2695. doi:10.1016/j.spa.2013.04.001

Herrero, J., Valencia, A., & Dopazo, J. (2001). An hierarchical unsupervised growing neural network form clustering gene expression patterns. *Bioinformatics*, *17*(2), 126-136.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417.

Hotelling, H. (1936). Relations between two sets of variants. *Biometrika*, *28*, 321-377.

Huxhold, W. E., & Levinsohn, A. G. (1995). *Managing Geographic Information System Projects*. New York: Oxford University Press.

IBM Knowledge Center. (2013, January 1). [CT701]. Recuperado 13 de Agosto de 2014, de http://www-01.ibm.com/support/knowledgecenter/SS3RA7_16.0.0/com.ibm.spss.modeler.help/clementine/nodes_clusteringmodels.htm?lang=es

IDC. (2014). The Digital Universe of Opportunities. Recuperado 13 de Abril de 2014, de http://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf

Ikeda, S., Abe, T., Nakamura, Y., Kibinge, N., Hirai Morita, A., Nakatani, A.,… Kanaya, S. (2013). Systematization of the Protein Sequence Diversity in Enzymes Related to Secondary Metabolic Pathways in Plants, in the Context of Big Data Biology Inspired by the KNApSAcK Motorcycle Database. *Plant and Cell Physiology*, *54*(5), 711–727. doi:10.1093/pcp/pct041

Işık, Ö., Jones, M. C., & Sidorova, A. (2013). Business intelligence success: The roles of BI capabilities and decision environments. *Information & Management*, *50*(1), 13-23. doi:10.1016/j.im.2012.12.001

Jiménez- Espinosa, R., & Chica-Olmo, M. (1992). Aplicación del Krigeaje Factorial al estudio de anomalías geoquímicas. *Boletín Geológico Y Minero*, *103*(4), 723–729.

Jin, Y., Cao, J., Ruan, Q., & Wang, X. (2014). Cross-Modality 2D-3D Face Recognition via Multiview Smooth Discriminant Analysis Based on ELM. *Journal of Electrical and Computer Engineering*, *2014*, 1–9. doi:10.1155/2014/584241

Juanes, P., Juanes, J. A., Villarroel, R., & Cabero, V. (2001). *Sistema de procesamiento digital de información georreferenciada de áreas periventriculares cerebrales*. Universidad de Salamanca, Facultad de Geografía e Historia, Departamento de Geografía.

Kaisler, S., Armour, F., Money, W., & Espinosa, J.A. (2015). Big Data Issues and Challenges. En Mehdi Khosrow-Pour (Ed.), *Encyclopedia of Information Science and Technology, Third Edition* (pp. 363–370). Hershey, PA, USA: IGI Global. Recuperado de http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-4666-5888-2.ch035

Kemsley, K. (1996). Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. *Chemometrics and Intelligent Laboratory Systems*, *33*, 47–61.

Kolmogorov, A. N. (1933). (En alemán) Grundbegriffe der Wahrscheinlichkeitsrechnung. Berlín: Springer. O traducción: Kolmogorov, Andrey (1956) Foundations of the Theory of Probability (2ª Edición.). Chelsea Publishing Company. New York.

Kruskal, J. B. (1964). Multidimensional Scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1-27. doi:10.1007/BF02289565

Kruskal, J. B., & Wish, M. (1978). *Multidimensional Scaling* (Vol. 11). London: Sage.

Kurome, M., Geistlinger, L., Kessler, B., Zakhartchenko, V., Klymiuk, N., Wuensch, A. … others. (2013). Factors influencing the efficiency of generating genetically engineered pigs by nuclear transfer: multi-factorial analysis of a large data set. *BMC Biotechnology*, *13*(1), 43.

Laukaitis, A. (2008). Functional data analysis for cash flow and transactions intensity continuous-time prediction using Hilbert-valued autoregressive processes. *European Journal of Operational Research*, *185*(3), 1607-1614. doi:10.1016/j.ejor.2006.08.030

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D. … Van Alstyne, M. (2009). Computational Social Science. *Science*, *323*(5915), 721-723. Recuperado de http://www.sciencemag.org/content/323/5915/721.short

Lin, D., Zhang, J.-G., Li, J., Calhoun, V. D., Deng, H.-W., & Wang, Y.-P. (2013). Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinformatics*, *14*(1), 245. Recuperado de http://www.biomedcentral.com/content/pdf/1471-2105-14-245.pdf

Lohr, S. (2012). Data-Driven Discovery Is Tech's New Wave - Unboxed. *The New York Times*. Recuperado 8 de septiembre de 2012 de http://www.nytimes.com/2012/09/09/technology/data-driven-discovery-is-techs-new-wave-unboxed.html

Lu, H. (2013). Learning canonical correlations of paired tensor sets via tensor-to-vector projection. Procedente del *Twenty-Third international joint conference on Artificial Intelligence* (pp. 1516-1522). AAAI Press. Beijing, China. Recuperado de http://dl.acm.org/citation.cfm?id=2540346

Madrid Soto, A., & Ortiz López, L. M. (2005). *Análisis y síntesis en cartografía: algunos procedimientos*. Bogotá: Universidad Nacional de Colombia. Recuperado de http://www.bdigital.unal.edu.co/1239/

Mandrekar, V., & Masani, P. R. (1997). *Proceedings of the Norbert Wiener Centenary Congress, 1994*. Michigan State: American Mathematical Soc.

Manyika J., Chui M., Brown D., Bughin J., Dobbs R., Roxburgh C., & Hung Byers A. (2011). *Big Data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

March, P. (2008). Discovery in Complex or Massive Datasets: Common Statistical Themes. National Science Foundation.

Martín Casado, A.M. (2014). *Escalamiento Multidimensional (Multidimensional Scaling)*. Apuntes del 'Máster en Análisis Avanzado de Datos Multivariantes', Universidad de Salamanca.

Mateu, J. (2003). *Geoestadística y modelos matemáticos en hidrogeología*. Castellón de la Plana: Universitat Jaume I.

Matheron, G. (1963). *Traité de Géostatistique Appliquée*. París: Éditions Technip.

Matheron, G. (1982). Pour une analyse krigeante des données regionalisées. Fontainebleau: Centre de Géostatistique.

Matsunaga, M. (2010). How to Facto-Analyze your data right: Do´s, Don´ts, and How-To´s. *International Journal of Psychological Research*, *3*(1), 97–110.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: La revolución de los datos masivos*. Madrid: Turner.

McHarg, I. L. (1995). *Design with Nature*. San Val Incorporated.

Mell, P., & Grance, T. (2011). *The NIST definition of cloud computing*. Recuperado de http://faculty.winthrop.edu/domanm/csci411/Handouts/NIST.pdf

Mercy, A.I., & Padmavathi, S. (2014). Discovering and Mining Links for Protein Databases. *Compusoft*, *3*(1).

Micó, G.A. (2012). Escalamiento multidimensional Métrico vs. No-Métrico: Intervalos de error en la interpretación de los resultados. Departamento de Psicología (Área de Metodología). Universidad de las Illes Balears.

Miller, H. J. (2010). The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science*, *50*(1), 181–201. doi:10.1111/j.1467-9787.2009.00641.x

Moldes Teo, F. J. (1995). Tecnología de los Sistemas de Información Geográfica. Madrid: Ra-Ma.

Montello, D., & Sutton, P. C. (2006). *An Introduction to Scientific Research Methods in Geography*. Thousand Oaks: SAGE Publications.

Montero Alonso, M. A., Vera, V., Fernando, J., Muñoz Orellana, J. A., & Andrés, G. C. (2003). Caracterización de indicadores y movimientos migratorios mediante MDS. Presentado en el 27 Congreso Nacional de Estadística e Investigación Operativa, Lleida. Recuperado de http://web.udl.cat/usuaris/esi2009/treballs/P2_51.pdf

Moreno Jiménez, A., & Cañada Torrecillas, R. (2006). Sistemas y análisis de la información geográfica: manual de autoaprendizaje con ArcGIS. Paracuellos de Jarama: Ra-Ma.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*, *58*(302), 415. doi:10.2307/2283276

Muller, E. E. L., Glaab, E., May, P., Vlassis, N., & Wilmes, P. (2013). Condensing the omics fog of microbial communities. *Trends in Microbiology*, *21*(7), 325–333. doi:10.1016/j.tim.2013.04.009

Ohlhorst, F. J. (2012). *Big Data Analytics*. New York. John Wiley & Sons, Ltd.

Oliver, M. A., Bosch, E., Slocum, K., Kleingold, W. J., & Krige, D. G. (2002). *Wavelets and Kriging for filtering and data reconstruction*. Cape Town: Geostatical Association of Southern Africa.

O'Reilly, T., & Battelle, J. (2009). *Web squared: Web 2.0 five years on* (Vol. 20). O'Reilly Media.

Otsuki, A., & Kawamura, M. (2013). The Study about the Analysis of Responsiveness Pair Clustering To social Network Bipartite Graph. *Advanced Computing: An International Journal*, *4*(6), 1–14. doi:10.5121/acij.2013.4601

Pardo-Iguzquiza, E., Chica-Olmo, M., Rigol-Sánchez, J. P., Luque-Espinar, J. A., & Rodríguez-Galiano, V. (2011). Una revisión de las nuevas aplicaciones metodológicas del cokrigeaje en Ciencias de la Tierra. *Boletín Geológico Y Minero*, *122*(4), 497-516.

Pardo-Iguzquiza, E., & Dowd, P. A. (2002). FACTOR2D: a computer program for Factorial Cokriging. *Computers & Geosciences*, *28*(8), 857-875.

Pearson, K. (1901). On lines and planes of closets fit to systems of points in the space. *Philosophical Magazine*, *2*, 559-572.

Peña, D. (2002). *Análisis de datos multivariantes* (Vol. 24). Madrid: McGraw-Hill.

Petitgas, P. (1996). Geostatistics and their applications to fisheries survey data. En B. Megrey & E. Moksness (Eds.), *Computers in Fisheries Research* (pp. 113-142). Springer Netherlands.

Podani, J. (1997). On the sensitivity of ordination and classification methods to variation in the input order of data. *Journal of Vegetation Science*, *8*(1), 153–156.

Puebla, J. G., & Gould, M. (1994). *SIG: Sistemas de información geográfica*. Madrid: Síntesis.

Rajaraman, A., & Ullman, J. D. (2012). *Mining of massive datasets*. New York: Cambridge University Press.

Ramsay, J. O. (1977). Maximum likelihood estimation in multidimensional scaling. *Psychometrika*, *42*(2), 241–266. doi:10.1007/BF02294052

Ratner, B. (2012). Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data, Second Edition. Taylor & Francis.

Rocco, C. M. (2012). Análisis exploratorio de datos espaciales en estudios de demanda eléctrica. En III Congreso Venezolano de redes y energía eléctrica. Universidad Central de Venezuela. Bogotá. Venezuela.

Sabatier, R., & Reynès, C. (2008). Extensions of simple component analysis and simple linear discriminant analysis using genetic algorithms. *Computational Statistics & Data Analysis*, *52*(10), 4779–4789. doi:10.1016/j.csda.2008.03.021

Sakaiya, T. (1995). *Historia del futuro: la sociedad del conocimiento*. Santiago de Chile: Editorial Andrés Bello.

Samuel, A. L. (1967). Some studies in machine learning using the game of checkers. II Recent Progress. *IBM Journal of Research and Development*, *11*(6), 601–617.

Schaefer, F. K. (1953). Exceptionalism in Geography: A Methodological Examination. *Annals of the Association of American Geographers*, *43*(3), 226. doi:10.2307/2560876

Schroeck M., Shockley R., Tufano P., Smart J., & Romero-Morales D. (2013). *Analytics: el uso de Big Data en el mundo real*. IBM Global Business Services.

Schuurman, N. (2004). *GIS: A Short Introduction*. John Wiley & Sons, Ltd.

Shen, C., Sun, M., Tang, M., & Priebe, C.E. (2014). Generalized Canonical Correlation Analysis for Classification. *Journal of Multivariate Analysis*, *130*, 310–322.

Shen, H., & Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, *99*(6), 1015–1034. doi:10.1016/j.jmva.2007.06.007

Shen, X.-B., Sun, Q.-S., & Yuan, Y.-H. (2013). Orthogonal canonical correlation analysis and its application in feature fusion. En *Information Fusion (FUSION), 2013 16th International Conference on* (pp. 151–157). IEEE. Recuperado de http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6641189

Shepard, R. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, *27*(3), 219–246. doi:10.1007/BF02289621

Singh, J. (2013). Big Data Analytic and Mining with Machine Learning Algorithm. Oracle Data Paper Junio 2013.

Slocum M. (2011). *Big data now current perspectives de O'Reilly radar*. Sebastopol, CA: O'Reilly Media.

Sokal, R., & Sneath, P. (1963). Principles of Numerical Taxonomy. *Freeman*, *6*(2), 139–140. doi:10.1002/jobm.19660060216

Sosinsky, B. (2010). *Cloud Computing Bible*. John Wiley & Sons, Ltd. Recuperado de http://books.google.es/books?id=hvv2pDEAbOEC

Srivastava, M. S. (2007). Multivariate theory for analyzing high dimensional data. *Journal of the Japan Statistical Society*, *37*(1), 53–86. Recuperado de http://jlc.jst.go.jp/JST.JSTAGE/jjss/37.53?de=Google

Takane, Y., Young, F. W., & De Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. *Psychometrika*, *42*(1), 7–67. Recuperado de http://link.springer.com/article/10.1007/BF02293745

Tamaki G. (2014). La hora del Big Data. www.techroi.com.pe

Tolosana Delgado, R. (2011). Guía para el análisis espacial de datos composicionales. *Boletín Geológico y Minero*, 122 (4): 469-482.

Torgerson, W. (1952). Multidimensional Scaling: I. Theory and method. *Psychometrika*, *17*(4), 401–419. doi:10.1007/BF02288916

Tsumoto, S., Iwata, H., Hirano, S., & Tsumoto, Y. (2014). Similarity-based behavior and process mining of medical practices. *Future Generation Computer Systems*, *33*, 21–31. doi:10.1016/j.future.2013.10.014

Tukey, J.W. (1962). The Future of Data Analysis. *Annals of Mathematical Statistics*, *3*, 1–67.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison: Wesley Publishing Company. Recuperado de http://books.google.es/books?id=UT9dAAAAIAAJ

Tukey, J. W. (1993). Exploratory Data Analysis: Past, Present and Future. DTIC Document.

Tzeng, J., Lu, H., & Li, W.-H. (2008). Multidimensional scaling for large genomic data sets. *BMC Bioinformatics*, *9*(1), 179. doi:10.1186/1471-2105-9-179

UIT News. (2013). Lo más destacado del mundo en 2013: datos y cifras relativos a las TIC - El uso de Internet. Recuperado de https://itunews.itu.int/es/3781-Lo-mas-destacado-de-El-mundo-en-2013-datos-y-cifras-relativos-a-las-TIC.note.aspx

Valero Torrijos, J. (2013). El Big Data en las Administraciones Públicas: El difícil equilibrio entre eficacia de la actividad administrativa y garantía de los derechos de los ciudadanos. Presentado en Big Data: Retos y Oportunidades. Actas del IX Congreso Internacional Internet, Derecho y Política. Universitat Oberta de Catalunya, Barcelona: UOC-Huygens Editorial.

Valkonen, V.-P., Kolehmainen, M., Lakka, H.-M., & Salonen, J. T. (2002). Insulin resistance syndrome revisited: application of self-organizing maps. *International Journal of Epidemiology*, *31*(4), 864–871.

Vera, J., & González, A. (1996). Un procedimiento de MDS a dos vías para el análisis mediante máxima verosimilitud de datos de disimilaridad con origen indeterminado. *Qüestíio*, *20*(1), 29–43.

Vicente-Villardón, J., Galindo, M., & Blazquez-Zaballos, A. (2006). Logistic Biplots. In *Multiple correspondence analysis and related methods*. En Greenacre M. y Blasius J. (Ed.), *Multiple Correspondence Analysis* (pp. 491–509). Londres. CRC Press.

Vilhena, D., Foster, J., Rosvall, M., West, J., Evans, J., & Bergstrom, C. (2014). Finding Cultural Holes: How Structure and Culture Diverge in Networks of Scholarly Communication. *Sociological Science*, *1*, 221–238. doi:10.15195/v1.a15

Viszlay, P., Pleva, M., & Juhàr, J. (2011). Dimension Reduction with Principal Component Analysis Applied to Speech Supervectors. *Journal of Electrical and Electronics Engineering*, *4*(1).

Wackernagel, H. (1995). Multivariate Geostatistics: An Introduction with Applications. Heidelberg: Springer.

Wang, Z., Han, F., & Liu, H. (2013). Sparse principal component analysis for high dimensional multivariate time series. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics* (pp. 48–56).

Warden, P. (2011). *Big Data Glossary*. O'Reilly Media.

Warrick, A. W., & Myers, D. E. (1987). Optimization of sampling locations for variogram calculations. *Water Resources Research*, *23*(3), 496–500.

Weiser, M. (1993). Some computer science issues in ubiquitous computing. *Commun. ACM*, *36*(7), 75–84.

Witten, I. H., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques*. Amsterdam: Elsevier Morgan Kaufmann.

Wold, S., & Sjöström, M. (1998). Chemometrics, present and future success. *Chemometrics and Intelligent Laboratory Systems*, *44*(1), 3–14.

Wong-González, P. (1999). Globalización y virtualización de la economía: impactos territoriales. En V Seminario de la Red Iberoamericana de Investigadores sobre Globalización y Territorio. Universidad Autónoma de México.

Woodward, G., Gay, C., & Baird, D. J. (2013). Biomonitoring for the 21st Century: new perspectives in an age of globalization and emerging environmental threats. *Limnetica*, *32*(2), 159-174.

Ye, J. (2007). Least squares linear discriminant analysis. En *Proceedings of the 24th international conference on Machine learning* (pp. 1087-1093). ACM.

Young, G., & Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, *3*(1), 19-22. Recuperado de http://link.springer.com/article/10.1007/BF02287916

Yuan, Z., & Myers, D. (1994). Simply and Ordinary Factorial Cokriging. *Science Terre*, *32*, 49-62.

Zhou, Z.-H., Chawla, N. V., Jin, Y., & Williams, G. J. (2014). Big Data Opportunities and Challenges: Discussions de Data Analytics Perspectives. *IEEE Computational Intelligence Magazine*. Recuperado de http://www.soft-computing.de/CIM_BD.pdf

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, *15*(2), 265–286. doi: 10.1198/106186006X113430.

12 ANEXO

Unidad	Tamaño	Lo que significa
Bit (b)	1 o 0	Diminutivo de 'dígito binario' (binary digit) por el código binario (1 o 0) que los ordenadores utilizan para almacenar y procesar datos
Byte (B)	8 bits	Información suficiente como para crear un carácter. Es la unidad básica de la informática
Kilobyte (KB)	1.000 o 2^{30} bytes	Kilo en griego significa 1.000. Una página de texto son 2KB
Megabyte (MB)	1.000KB; 2^{20}bytes	Mega en griego significa grande. Las obras completas de Shakespeare son 5MB. Una canción suele tener alrededor de 4MB.
Gigabyte (GB)	1.000MB; 2^{30}bytes	Giga en griego significa gigante. Una película de dos horas puede comprimirse en entre 1 y 2GB.
Terabyte (TB)	1.000GB; 2^{40}bytes	Tera en griego significa monstruo. Todos los libros de la biblioteca del congreso estadounidense suman un total 15TB
Petabyte (PB)	1.000TB; 2^{50}bytes	Todas las cartas entregadas por el servicio postal estadounidense sumarán alrededor de 5PB. Google procesa aproximadamente 1PB cada hora.
Exabyte (EB)	1.000PB; 2^{60}bytes	El equivalente a 10.000 millones de copias de The Economist
Zettabyte (ZB)	1.000EB; 2^{70}bytes	Se calcula que al final del año habrá un total de 1,2ZB de información en total
Yottabyte (YB)	1.000ZB; 2^{80}bytes	Aún es imposible imaginarlo

ANEXO 1: Definición de las unidades de almacenamiento digital. Fuente: García y De Prado (2012).

13 GLOSARIO

A continuación, se muestran e incluyen todos aquellos términos poco conocidos, de difícil interpretación, o que no hayan sido comúnmente utilizados en el contexto en que aparecen a lo largo del TFM. Cada uno de estos vocablos viene acompañado de su respectiva definición que ha sido recogida de Wikipedia (procediendo a corregir convenientemente las erratas encontradas). Si no se encontró un término en este '*repositorio enciclopédico informatizado*' se buscó su descripción directamente de la fuente original.

A

'ALGORITMO'

http://es.wikipedia.org/wiki/Algoritmo

Es un conjunto prescrito de instrucciones o reglas bien definidas, ordenadas y finitas que permite realizar una actividad mediante pasos sucesivos que no generen dudas a quien deba realizar dicha actividad.

'ALGORITMO GENÉTICO'

http://es.wikipedia.org/wiki/Algoritmo_gen%C3%A9tico

En los años 70, de la mano de John Henry Holland, surgió una de las líneas más prometedoras de la inteligencia artificial, la de los algoritmos genéticos. Son llamados así porque se inspiran en la evolución biológica y su base genético-molecular. Estos algoritmos hacen evolucionar una población de individuos sometiéndola a acciones aleatorias semejantes a las que actúan en la evolución biológica (mutaciones y recombinaciones genéticas), así como también a una selección de acuerdo con algún criterio, en función del cual se decide cuáles son los individuos más adaptados (que sobreviven), y cuáles los menos aptos (que son descartados). Los algoritmos genéticos se enmarcan dentro de los algoritmos evolutivos, que incluyen también las estrategias evolutivas, la programación evolutiva y la programación genética.

'AMAZON WEB SERVICES'

http://es.wikipedia.org/wiki/Amazon_Web_Services

Amazon Web Services (AWS abreviado) es una colección de servicios de computación en la nube (también llamados servicios web) que en conjunto forman una plataforma de servicios ofrecidas a través de Internet por Amazon.com. Es usado en aplicaciones

tan populares como Dropbox. Es una de las ofertas internacionales más importantes de la computación en la nube y compite directamente contra servicios como Windows Azure y Google Cloud Platform. Es considerado como un pionero en este campo.

'API'

http://es.wikipedia.org/wiki/API

API (del inglés '*Application Programming Interface*') o Interfaz de Programación de Aplicaciones (IPA) es el conjunto de funciones y procedimientos (o métodos, en la programación orientada a objetos) que ofrece ciertas bibliotecas para ser utilizadas por otro software como una capa de abstracción.

'ARPANET'

http://es.wikipedia.org/wiki/ARPANET

La red de computadoras '*Advanced Research Projects Agency Network*' (ARPANET) fue creada por encargo del Departamento de Defensa de Estados Unidos ('DOD' por sus siglas en inglés) como medio de comunicación para los diferentes organismos del país. El primer nodo se creó en la Universidad de California (Los Ángeles) y fue la espina dorsal de Internet hasta 1990, tras finalizar la transición al protocolo TCP/IP iniciada en 1983. El concepto de una red de computadoras capaz de comunicar usuarios en distintas computadoras fue formulado por J.C.R. Licklider de Bolt, Beranek and Newman (BBN) en agosto de 1962, en una serie de notas que discutían la idea de 'red galáctica'.

'AUTO-ID CENTER'

http://www.3c3cc.com/web/ES/assets/executives/pdf/Internet_of_Things_IoT_IBSG_0411FINAL.pdf

El 'Auto-ID Center' es el precursor del moderno '*Auto-ID Labs*' que conforma un grupo de investigación en el ámbito de la red de identificación por radiofrecuencia (RFID) y las tecnologías de sensores (IoT). Los laboratorios están vinculados a siete universidades de investigación ubicadas en cuatro continentes diferentes.

B

‘BIG CIENCIA’

http://es.wikipedia.org/wiki/Megaciencia

La Big Ciencia (‘*Big Science*’ en inglés), es un término usado por los científicos, y particularmente usado en la Historia de la ciencia y de la tecnología. Con este concepto, se describen y engloban una serie de cambios en la investigación científica ocurridos en los países industrializados durante y con posterioridad a la Segunda Guerra Mundial. Hacia el fin de la primera mitad del siglo XX, el progreso científico se aceleró notoriamente y comenzaron a realizarse proyectos a gran escala, por lo general financiados por gobiernos nacionales o grupos de gobiernos (el proyecto Manhattan, el programa atómico de Japón, la carrera espacial, el telescopio espacial Hubble, la exploración de Marte, el colisionador de partículas’, el proyecto Genoma Humano, etc.)

‘BITNET’

http://es.wikipedia.org/wiki/Bitnet

Bitnet era una antigua red internacional de computadoras de centros docentes y de investigación que ofrecía servicios interactivos de correo electrónico y de transferencia de ficheros utilizando un protocolo de almacenaje y envío basado en los protocolos ‘*Network Job Entry*’ de IBM. Se conectaba a Internet a través de una pasarela de correo electrónico.

‘BUSINESS INTELLIGENCE’

http://es.wikipedia.org/wiki/Inteligencia_empresarial

Se denomina inteligencia empresarial, inteligencia de negocios o BI (del inglés ‘*Business Intelligence*’) al conjunto de estrategias y aspectos relevantes enfocados a la administración y creación de conocimiento sobre el medio, a través del análisis de los datos existentes en una organización o empresa.

C

‘CAD‘

http://es.wikipedia.org/wiki/Dise%C3%B1o_asistido_por_computadora

Del inglés ‘*Computer Aid Design*’. El diseño asistido por computadora, es el uso de un amplio rango de herramientas computacionales que asisten a ingenieros, arquitectos y diseñadores.

‘CAE’

http://es.wikipedia.org/wiki/Ingenier%C3%ADa_asistida_por_computadora

Del inglés *‘Computer Aid Engineering*’. Ingeniería asistida por ordenador, es la disciplina que se encarga del conjunto de programas informáticos que permiten analizar y simular los diseños de ingeniería realizados con el ordenador, o creados de otro modo e introducidos en el ordenador, para valorar sus características, propiedades, viabilidad, y rentabilidad. Su finalidad es optimizar su desarrollo y consecuentes costos de fabricación, y reducir al máximo las pruebas para la obtención del producto deseado.

‘CERN’

http://es.wikipedia.org/wiki/Organizaci%C3%B3n_Europea_para_la_Investigaci%C3%B3n_Nuclear

La Organización Europea para la Investigación Nuclear (nombre oficial), comúnmente conocida por la sigla ‘CERN’ (sigla provisional utilizada en 1952, que respondía al nombre en francés ‘*Conseil Européen pour la Recherche Nucléaire*’ y en castellano ‘Consejo Europeo para la Investigación Nuclear’), es el mayor laboratorio de investigación en física de partículas en el ámbito mundial.

‘CIENCIA REGIONAL’

http://en.wikipedia.org/wiki/Regional_science

La Ciencia Regional es una disciplina de la Ciencia Social que se ocupa de los problemas analíticos que son específicamente urbanos, rurales o de enfoque regional. Los temas de la ciencia regional incluyen (sin limitarse a la teoría de la localización y la economía del espacio): el modelado de localización y el transporte, del análisis de la migración, del uso del suelo y la planificación urbana, del análisis industrial, ambiental y ecológico, de la gestión de los recursos urbanos y el análisis de la política regional.

En el sentido más amplio, cualquier análisis de la Ciencia Social que tiene una dimensión espacial es estudiado por científicos de la región.

'CISCO SYSTEMS'

http://es.wikipedia.org/wiki/Cisco_Systems

Cisco Systems es una empresa global con sede en San José (California, Estados Unidos), principalmente dedicada a la fabricación, venta, mantenimiento y consultoría de equipos de telecomunicaciones tales como:

- Dispositivos de conexión para redes informáticas: routers, switches (conmutadores) y hubs (concentradores).
- Dispositivos de seguridad como Cortafuegos y Concentradores para VPN.
- Productos de telefonía IP como teléfonos y CallManagers.
- Software de gestión de red como CiscoWorks.
- Equipos para redes de área de almacenamiento.

'CLOUD COMPUTING'

http://es.wikipedia.org/wiki/Computaci%C3%B3n_en_la_nube

Nuevo paradigma de la computación en '*nube'*, según el cual cualquier cosa que pueda hacerse en informática puede trasladarse a esa '*nube*' o, lo que es lo mismo, a la Red. Este modelo implica el uso de recursos informáticos como un suministro más, igual que si se tratara de la electricidad o el teléfono. Estos recursos son ofrecidos por proveedores de '*nube'*, que los gestionan en grandes centros de datos remotos y prestan servicio a múltiples clientes que acceden a ellos a través de cualquier dispositivo conectado a Internet.

D

'DATACENTER'

http://es.wikipedia.org/wiki/Centro_de_procesamiento_de_datos

Se denomina Centro de Procesamiento de Datos (CPD) a aquella ubicación donde se concentran los recursos necesarios para el procesamiento de la información de una organización. También se conoce como centro de cómputo en Hispanoamérica, o centro de cálculo en España o centro de datos por su equivalente en inglés Datacenter. Dichos recursos consisten esencialmente en unas dependencias debidamente acondicionadas, computadoras y redes de comunicaciones.

‘DATA MINING’

http://es.wikipedia.org/wiki/Miner%C3%ADa_de_datos

(Minería de Datos). Es un campo de las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. El objetivo general del proceso de minería de datos consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. Además de la etapa de análisis en bruto, que involucra aspectos de bases de datos y de gestión de datos, de procesamiento de datos, del modelo y de las consideraciones de inferencia, de métricas de Intereses, de consideraciones de la Teoría de la complejidad computacional, de post-procesamiento de las estructuras descubiertas, de la visualización y de la actualización en línea.

‘DARPA’

http://es.wikipedia.org/wiki/Defense_Advanced_Research_Projects_Agency

DARPA acrónimo de la expresión en inglés ‘*Defense Advanced Research Projects Agency*’ (Agencia de Proyectos de Investigación Avanzados de Defensa) es una agencia del Departamento de Defensa de Estados Unidos responsable del desarrollo de nuevas tecnologías para uso militar. Fue creada en 1958 como consecuencia tecnológica de la llamada Guerra Fría, y del que surgieron, década después, los fundamentos de ARPANET, red que dio origen a Internet. La agencia, denominada en su origen simplemente como ARPA, cambió su denominación en 1972, conociéndose en lo sucesivo como DARPA. Fue responsable de dar fondos para desarrollar muchas tecnologías que han tenido un gran impacto en el mundo: satélites, robots, incluyendo redes de ordenadores, (empezando con ARPANET, que después se desarrolló como Internet), así como NLS, el cual fue tanto un sistema de hipertexto como un precursor de la interfaz gráfica de usuario contemporánea.

‘DROPBOX’

http://es.wikipedia.org/wiki/Dropbox

Dropbox es un servicio de alojamiento de archivos multiplataforma en la nube, operado por la compañía Dropbox. El servicio permite a los usuarios almacenar y sincronizar archivos en línea y entre ordenadores, para así compartir archivos y carpetas con otros. Existen versiones gratuitas y de pago, cada una de las cuales tiene opciones variadas.

E

‘E-COMMERCE’

http://es.wikipedia.org/wiki/Comercio_electr%C3%B3nico

El comercio electrónico, también conocido como e-commerce (‘*Electronic Commerce*’ en inglés), consiste en la compra y venta de productos o de servicios a través de medios electrónicos, tales como Internet y otras redes informáticas. Originalmente el término se aplicaba a la realización de transacciones mediante medios electrónicos tales como el intercambio electrónico de datos, sin embargo con el desarrollo de Internet y la World Wide Web a mediados de los años 90 comenzó a referirse principalmente a la venta de bienes y servicios a través de Internet, usando como forma de pago medios electrónicos, tales como las tarjetas de crédito. La cantidad de comercio llevada a cabo electrónicamente ha crecido de manera extraordinaria debido a Internet. Una gran variedad de comercio se realiza de esta manera, estimulando la creación y utilización de innovaciones como la transferencia de fondos electrónica, la administración de cadenas de suministro, el marketing en Internet, el procesamiento de transacciones en línea (OLTP), el intercambio electrónico de datos (EDI), los sistemas de administración del inventario y los sistemas automatizados de recolección de datos. La mayor parte del comercio electrónico consiste en la compra y venta de productos o servicios entre personas y empresas, sin embargo un porcentaje considerable del comercio electrónico consiste en la adquisición de artículos virtuales (software y derivados en su mayoría), tales como el acceso a contenido "Premium" de un sitio web.

‘E-MAIL’

http://es.wikipedia.org/wiki/Correo_electr%C3%B3nico

Correo electrónico (en inglés: e-mail), es un servicio de red que permite a los usuarios enviar y recibir mensajes (también denominados mensajes electrónicos o cartas electrónicas) mediante sistemas de comunicación electrónica. Principalmente se usa este nombre para denominar al sistema que provee este servicio en Internet, mediante el protocolo SMTP, aunque por extensión también puede verse aplicado a sistemas análogos que usen otras tecnologías. Por medio de mensajes de correo electrónico se puede enviar, no solamente texto, sino todo tipo de documentos digitales dependiendo del sistema que se use. Su eficiencia, conveniencia y bajo coste están logrando que el correo electrónico desplace al correo ordinario para muchos usos habituales.

‘EDA’

http://es.wikipedia.org/wiki/An%C3%A1lisis_exploratorio_de_datos

El análisis exploratorio de datos definido por John W. Tukey (del inglés ‘*Exploratory Data Analysis*’) es, básicamente, el tratamiento estadístico al que se someten las muestras recogidas durante un proceso de investigación en cualquier campo científico. Para mayor rapidez y precisión, todo el proceso suele realizarse por medios informáticos, con aplicaciones específicas para el tratamiento estadístico. Los EDA, no necesariamente, se llevan a cabo con una base de datos al uso, ni con una hoja de cálculo convencional; no obstante el programa SPSS y R (lenguaje de programación) son las aplicaciones más utilizadas, aunque no las únicas.

‘EXPERIENCIA’

http://es.wikipedia.org/wiki/Experiencia

Experiencia (del latín *experiri*, ‘comprobar’) es una forma de conocimiento o habilidad derivados de la observación, de la participación y de la vivencia de un evento o proveniente de las cosas que suceden en la vida, es un conocimiento que se elabora colectivamente.

F

‘FTP’

http://es.wikipedia.org/wiki/File_Transfer_Protocol

FTP (del inglés ‘*File Transfer Protocol*’) Protocolo de Transferencia de Archivos, en informática, es un protocolo de red para la transferencia de archivos entre sistemas conectados a una red TCP (‘*Transmission Control Protocol*’), basado en la arquitectura cliente-servidor. Desde un equipo cliente se puede conectar a un servidor para descargar archivos desde él o para enviarle archivos, independientemente del sistema operativo utilizado en cada equipo. El servicio FTP es ofrecido por la capa de aplicación del modelo de capas de red TCP/IP al usuario, utilizando normalmente el puerto de red 20 y el 21. Un problema básico de FTP es que está pensado para ofrecer la máxima velocidad en la conexión, pero no la máxima seguridad, ya que todo el intercambio de información, desde el login y password del usuario en el servidor hasta la transferencia de cualquier archivo, se realiza en texto plano sin ningún tipo de cifrado, con lo que un posible atacante puede capturar este tráfico, acceder al servidor y/o apropiarse de los archivos transferidos.

‘FEEDBACK’

http://es.wikipedia.org/wiki/Realimentaci%C3%B3n

El *‘feedback’* (o retroalimentación) es un mecanismo por el cual una cierta proporción de la salida de un sistema se redirige a la entrada, con objeto de controlar su comportamiento. La realimentación se produce cuando las salidas del sistema o la influencia de las salidas del sistema en el contexto, vuelven a ingresar al sistema como recursos o información. La realimentación permite el control de un sistema y que el mismo tome medidas de corrección con base en la información realimentada.

G

‘GIS’

http://es.wikipedia.org/wiki/Sistema_de_informaci%C3%B3n_geogr%C3%A1fica

Del inglés ‘*Geographic Information System*’ (Sistemas de Información Geográfica). Integración organizada de hardware, software y datos geográficos diseñada para capturar, almacenar, manipular, analizar y desplegar en todas sus formas la información georreferenciada con el fin de resolver problemas complejos de planificación y gestión geográfica. También puede definirse como un modelo de una parte de la realidad referido a un sistema de coordenadas terrestres, construido para satisfacer unas necesidades concretas de información.

‘GMAIL’

http://es.wikipedia.org/wiki/Gmail

Gmail, llamado en otros lugares Google Mail (Austria y antes en Alemania -hasta 2012- y Reino Unido -hasta 2009-) por problemas legales, es un servicio de correo electrónico con posibilidades POP3 e IMAP gratuito proporcionado por la empresa estadounidense Google, Inc. a partir del 15 de abril de 2004 y que ha captado la atención de los medios de información por sus innovaciones tecnológicas, su capacidad, y por algunas noticias que alertaban sobre la violación de la privacidad de los usuarios.

‘GOOGLE ACADÉMICO’

http://es.wikipedia.org/wiki/Google_Acad%C3%A9mico

‘Google Scholar’ o Google Académico es un buscador de Google especializado en artículos de revistas científicas, enfocado en el mundo académico, y soportado por una

base de datos disponible libremente en Internet que almacena un amplio conjunto de trabajos de investigación científica de distintas disciplinas y en distintos formatos de publicación.

'GOOGLE APP ENGINE'

http://es.wikipedia.org/wiki/Google_App_Engine

'Google App Engine' es un servicio de alojamiento web que presta Google de forma gratuita hasta determinadas cuotas, este servicio permite ejecutar aplicaciones sobre la infraestructura de Google. Si no se cuenta con un dominio propio, Google proporciona uno con la siguiente estructura: *midominio.appspot.com*. También permite implementar un dominio propio a través de Google Apps. Por el momento las cuentas gratuitas tienen un límite de 500 megabyte de almacenamiento permanente y la suficiente cantidad de ancho de banda y CPU para cinco millones de visitas mensuales, y si la aplicación supera estas cuotas, se pueden comprar cuotas adicionales.

'GOOGLE DOCS'

http://es.wikipedia.org/wiki/Google_Docs

Google Documentos y Hojas de cálculo, oficialmente '*Google Docs & Spreadsheets*' es un programa gratuito basado en Web para crear documentos en línea con la posibilidad de colaborar en grupo. Incluye un Procesador de textos, una Hoja de cálculo, Programa de presentación básico, un creador de dibujos y un editor de formularios destinados a encuestas. '*Google Docs*' junto con Gmail, '*Google Calendar*' y '*Google Talk*'; el 7 de julio de 2009, dejaron su calidad de Beta y pasaron a ser productos terminados.

'GOOGLE DRIVE'

http://es.wikipedia.org/wiki/Google_Drive

'Google Drive' es un servicio de alojamiento de archivos. Fue introducido por Google el 24 de abril de 2012. '*Google Drive*' es un reemplazo de '*Google Docs*' que ha cambiado su dirección de enlace de docs.google.com por drive.google.com entre otras cualidades. Cada usuario cuenta con 15 gigabytes de espacio gratuito para almacenar sus archivos, ampliables mediante pago. Es accesible por su página web desde ordenadores y dispone de aplicaciones para iOS y Android que permiten editar documentos y hojas de cálculo.

H

‘HOSTING’

http://es.wikipedia.org/wiki/Alojamiento_web

El alojamiento web (en inglés ‘*Web Hosting*’) es el servicio que provee a los usuarios de Internet un sistema para poder almacenar información, imágenes, vídeo, o cualquier contenido accesible vía web. Es una analogía de ‘hospedaje o alojamiento en hoteles o habitaciones’ donde uno ocupa un lugar específico, en este caso la analogía alojamiento web o alojamiento de páginas web, se refiere al lugar que ocupa una página web, sitio web, sistema, correo electrónico, archivos etc. en Internet o más específicamente en un servidor que por lo general hospeda varias aplicaciones o páginas web. Las compañías que proporcionan espacio de un servidor a sus clientes se suelen denominar con el término en inglés ‘*Web Host*’. El hospedaje web aunque no es necesariamente un servicio, se ha convertido en un lucrativo negocio para las compañías de Internet alrededor del mundo. Se puede definir como ‘un lugar para tu página web o correos electrónicos’, aunque esta definición simplifica de manera conceptual el hecho de que el alojamiento web es en realidad espacio en Internet para prácticamente cualquier tipo de información, sea archivos, sistemas, correos electrónicos, videos etc.

‘HUMAN BRAIN PROJECT’

http://es.wikipedia.org/wiki/Proyecto_cerebro_humano

El Proyecto Cerebro Humano (HBP por sus siglas en inglés) es un proyecto médico-científico y tecnológico financiado por la Unión Europea y dirigido por Henry Makram, que tiene como fin reproducir tecnológicamente las características del cerebro humano, y de esta forma conseguir avances en el campo de la medicina y la neurociencia. Para que éste proyecto pueda desarrollarse es necesaria la investigación en nuevas TIC, o tecnologías de supercomputación avanzadas que permitan asociar y utilizar la información integrada en modelos informáticos y simulaciones del cerebro que identifiquen patrones, principios organizativos y posibles carencias que puedan ser subsanadas con nuevos experimentos.

I

‘I+D’

http://es.wikipedia.org/wiki/Investigaci%C3%B3n_y_desarrollo

El término investigación y desarrollo, abreviado I+D, (del inglés ‘*Research and Development*’, abreviado R&D), puede hacer referencia, según el contexto, a la investigación en ciencias aplicadas o bien ciencia básica utilizada en el desarrollo de ingeniería, que persigue con la unión de ambas áreas un incremento de la innovación que conlleve un aumento en las ventas de las empresas. Un fuerte vínculo entre la investigación y desarrollo para la investigación de ciencias aplicadas es, por un lado, una nueva fuente de ingresos para los institutos de las universidades gracias a la cooperación con las empresas, y, por otro, las empresas ven un futuro más prometedor si se implican en la investigación de forma continua.

‘I+D+I’

http://es.wikipedia.org/wiki/Investigaci%C3%B3n,_desarrollo_e_innovaci%C3%B3n

Investigación, desarrollo e innovación (habitualmente indicado por la expresión I+D+i o I+D+I) es un concepto de reciente aparición, en el contexto de los estudios de ciencia, tecnología y sociedad; como superación del anterior concepto de investigación y desarrollo (I+D). Es el corazón de las tecnologías de la información y comunicación. Mientras que el de desarrollo es un término proveniente del mundo de la economía, los de investigación e innovación provienen respectivamente del mundo de la ciencia y la tecnología, y su dinámica relación se encuentra en el contexto de la diferenciación entre ciencia pura y ciencia aplicada; cualquiera de ellos es de compleja definición. Esko Aho define provocativamente investigación como’ invertir dinero para obtener conocimiento’, mientras que innovación sería ‘invertir conocimiento para obtener dinero’, lo que expresa muy bien el fenómeno de retroalimentación que se produce con una estrategia exitosa de I+D+i (Citado por Alejandro Jadad y Julio Lorca Innovación no es lo mismo que novedad, en Andalucía Investiga, nº 38, febrero de 2007, pg. 44. También aportan su propia definición: investigar es ‘invertir recursos para obtener conocimiento’, en tanto que innovar es ‘invertir conocimiento para obtener valor’).

‘IAAS’

http://es.wikipedia.org/wiki/Computaci%C3%B3n_en_la_nube

Del inglés ‘*Infraestructure as a Service*’ (Infraestructura Como Servicio). También llamada ‘*Hardware as a Service’* (HaaS) se encuentra en la capa inferior de los

servicios disponibles en la '*nube'* y es un medio que ofrece almacenamiento básico y capacidades de cómputo como servicios estandarizados en la red. Servidores, sistemas de almacenamiento, conexiones, enrutadores y otros sistemas se concentran (por ejemplo a través de la tecnología de virtualización) para manejar tipos específicos de cargas de trabajo: desde el procesamiento en lotes, hasta el aumento de la capacidad del servidor y del almacenamiento durante los picos de las cargas de trabajo. El ejemplo comercial más conocido es '*Amazon Web Services*'.

'INFOMEDIARIO'

http://www.zorraquino.com/diccionario/marketing-online/infomediario.html

Modelo de negocio online dedicado a administrar el exceso de información propio de Internet; recopila grandes cantidades de datos provenientes de diversas fuentes, que analiza, criba y organiza de manera relevante para brindarlos finalmente en calidad de proveedor neutral a los usuarios que los requieran. Habitualmente tienden a la especialización, ofreciendo datos acerca de un sector exclusivo del mercado. Se pueden distinguir dos tipos de infomediario, según en qué extremo de la transacción se encuentren sus clientes: en el caso de ser consumidores, les facilita un determinado proceso comercial proveyendo para su consulta contenidos minuciosos acerca de los productos o marcas involucradas; si se trata de negocios, recopila para ellos información acerca de los públicos y sus hábitos de consumo que les ayudará en el desarrollo de productos y su comercialización. Los ingresos de un infomediario provienen de las inserciones realizadas en sus espacios publicitarios y de la comisión que le corresponda por intermediar en cada transacción conseguida.

'INTERNET 2.0'

http://es.wikipedia.org/wiki/Web_2.0

El término Web 2.0 comprende aquellos sitios web que facilitan el compartir información, la interoperabilidad, el diseño centrado en el usuario y la colaboración en la '*World Wide Web*'. Un sitio Web 2.0 permite a los usuarios interactuar y colaborar entre sí como creadores de contenido generado por usuarios en una comunidad virtual, a diferencia de sitios web estáticos donde los usuarios se limitan a la observación pasiva de los contenidos que se han creado para ellos. Ejemplos de la Web 2.0 son las comunidades web, los servicios web, las aplicaciones Web, los servicios de red social, los servicios de alojamiento de videos, las wikis, blogs, etc. Es la evolución de las aplicaciones estáticas a las dinámicas donde la colaboración del usuario es necesaria. En conclusión, la Web 2.0 nos permite realizar trabajo

colaborativo entre varios usuarios o colaboradores. Además, las herramientas que ofrece la Web 2.0 no sólo permitirán mejorar los temas en el aula de clase, sino también pueden utilizarse para trabajo en empresa. La Web 2.0 permite a estudiantes y docentes mejorar las herramientas utilizadas en el aula de clase. El trabajo colaborativo está tomando mucha importancia en las actividades que realicemos en Internet.

'INTERNET 3.0'

http://es.wikipedia.org/wiki/Web_3.0

Movimiento social orientado a crear contenidos accesibles por múltiples aplicaciones '*non-browser*' (sin navegador), promovido por el empuje de las tecnologías de Inteligencia Artificial, la Web Semántica, la Web Geoespacial o la Web 3D. La expresión es utilizada por los mercados para promocionar las mejoras respecto a la Web 2.0. Esta expresión Web 3.0 apareció por primera vez en 2006 en un artículo de Jeffrey Zeldman, crítico de la Web 2.0 y asociado a tecnologías como AJAX. Actualmente existe un debate considerable en torno a lo que significa Web 3.0, y cuál sea la definición más adecuada.

'INTERNET PROFUNDO'

http://es.wikipedia.org/wiki/Internet_profunda

Se conoce como 'Internet Profundo' o 'Internet Invisible' (en inglés: '*Deep Web*', '*Invisible Web*', '*Dark Web*' o '*Hidden Web*') a todo el contenido de Internet que no forma parte del 'Internet Superficial', es decir, de las páginas indexadas por las redes de los motores de búsqueda de la red. Esto se debe a las limitaciones que tienen las redes para acceder a todos los sitios web por distintos motivos. La principal causa de la existencia de la Internet profunda es la imposibilidad de los motores de búsqueda de encontrar o indexar gran parte de la información existente en Internet. Se estima que la Internet Profunda es 500 veces mayor que el 'Internet Superficial', siendo el 95% de esta información públicamente accesible. Si los buscadores tuvieran la capacidad para acceder a toda la información entonces la magnitud del 'Internet Profundo' se reduciría casi en su totalidad; sin embargo, no desaparecería totalmente porque siempre existirán páginas privadas. Los siguientes son algunos de los motivos por los que los buscadores son incapaces de indexar la Internet profunda:

- Páginas y sitios web protegidos con contraseña.
- Documentos en formatos no indexables.

- Enciclopedias, diccionarios, revistas en las que para acceder a la información hay que interrogar a la base de datos, como por ejemplo la base de datos de la RAE.

'INTERNET SOCIETY'

http://es.wikipedia.org/wiki/Internet_Society

'Internet Society' (ISOC) es una organización no gubernamental y sin ánimo de lucro, constituida como la única organización dedicada exclusivamente al desarrollo mundial de Internet y con la tarea específica de concentrar sus esfuerzos y acciones en asuntos particulares sobre Internet. Fundada en 1991 por una gran parte de los arquitectos pioneros encargados de su diseño, la ISOC tiene como objetivo principal ser un centro de cooperación y coordinación global para el desarrollo de protocolos y estándares compatibles. Tanto las cuotas de sus socios como las contribuciones económicas de particulares, organizaciones y empresas son completamente deducibles de impuestos en Estados Unidos, según la norma del IRC 501 así como en algunos otros países.

'IOE'

http://www.cisco.com/web/about/ac79/innov/IoE.html

Cisco define el 'Internet de Todo' (IoE) como el conjunto de conexiones a Internet de personas, procesos, datos y cosas que hacen este hecho más relevante y valioso que nunca, al convertir la información en acciones que crean nuevas capacidades, experiencias más valiosas y oportunidades económicas sin precedentes para las empresas, las personas y los países.

'IOT'

http://es.wikipedia.org/wiki/Internet_de_las_cosas

Del inglés '*Internet of the Things*' (Internet de las Cosas). Nuevo concepto que completa la evolución de las comunicaciones y la informática, aplicándola a los objetos, que facilita una mejor interacción con ellos. Se refiere a una red de objetos cotidianos interconectados a través de Internet.

J

‘JERARQUÍA DIKW’

http://es.wikipedia.org/wiki/Jerarqu%C3%ADa_del_conocimiento

La *Jerarquía del Conocimiento*, también conocida como ‘*Jerarquía DIKW’*, o ‘*Pirámide del Conocimiento’*, podría ser definida como un conjunto de modelos para representar las relaciones aparentemente estructurales entre los ‘Datos’, la ‘Información’, el ‘Conocimiento’, y en algunos casos la ‘Sabiduría’. Por lo general:

- La ‘Información’ se define en términos de ‘Datos’.
- El ‘Conocimiento’ se define en términos de ‘Información’.
- La ‘Sabiduría’ se define en términos de ‘Conocimiento’.

Entonces la secuencia de la jerarquía, de lo más básico a lo más complejo, es:

- ‘Datos’.
- ‘Información’.
- ‘Conocimiento’.
- ‘Sabiduría’.

No todas las versiones del conjunto de modelos comprenden estos cuatro componentes. Además de ser considerados una jerarquía y una pirámide, también se han considerado una cadena, una plataforma, y un continuo.

K

‘KNOW-HOW’

http://es.wikipedia.org/wiki/Know_how

El ‘*Know-How*’ (‘*saber hacer’*) o conocimiento fundamental es una forma de transferencia de tecnología. Es una expresión anglosajona utilizada en los últimos tiempos en el comercio internacional para denominar los conocimientos preexistentes no siempre académicos, que incluyen: técnicas, información secreta, teorías e incluso datos privados (como clientes o proveedores). Un uso muy difundido del término suele utilizarse en la venta de franquicias, ya que lo que se vende es el ‘saber cómo’. Las franquicias generalmente son vendidas por países o empresas ‘avanzadas’ que ‘ya lo han hecho’, casi siempre en el campo de los negocios, el saber cómo hacerlo a personas que saben poco del tema se convierte en un patrimonio de muchos años de madurez y una ventaja comparativa muy valiosa frente a la competencia.

‘KRIGING’

http://es.wikipedia.org/wiki/Krigeaje

Método de interpolación de Geoestadística basado en modelos estadísticos que incluyen la autocorrelación (relación estadística entre los puntos medidos). El Kriging pondera los valores medidos circundantes para derivar una predicción para las localizaciones sin medir. Los pesos están basados en la distancia entre los puntos medidos, las situaciones de la predicción y la composición global entro todos los puntos medidos.

L

‘LEY DE MOORE’

http://es.wikipedia.org/wiki/Ley_de_Moore

Expresa que aproximadamente cada dieciocho meses se duplica el número de transistores en un circuito integrado. Se trata de una ley empírica, formulada por el cofundador de Intel, Gordon E. Moore el 19 de abril de 1965, cuyo cumplimiento se ha podido constatar hasta hoy.

M

‘MALWARE’

http://es.wikipedia.org/wiki/Malware

El malware (del inglés ‘*Malicious Software’*), también llamado badware, código maligno, software malicioso o software malintencionado, es un tipo de software que tiene como objetivo infiltrarse o dañar una computadora o sistema de información sin el consentimiento de su propietario. El término malware es muy utilizado por profesionales de la informática para referirse a una variedad de software hostil, intrusivo o molesto. El término virus informático suele aplicarse de forma incorrecta para referirse a todos los tipos de malware, incluidos los virus verdaderos.

‘MAPREDUCE’

http://es.wikipedia.org/wiki/MapReduce

MapReduce es un modelo de programación utilizado por Google para dar soporte a la computación paralela sobre grandes colecciones de datos en grupos de computadoras y al ‘*Commodity Computing*’. El nombre está inspirado en los nombres de dos

importantes métodos, macros o funciones en programación funcional: Map y Reduce. MapReduce ha sido adoptado mundialmente, ya que existe una implementación OpenSource denominada Hadoop. Su desarrollo fue liderado inicialmente por Yahoo y actualmente lo realiza el proyecto Apache. En esta década de los años 2010 existen diversas iniciativas similares a Hadoop tanto en la industria como en la academia. Se han escrito implementaciones de bibliotecas de MapReduce en diversos lenguajes de programación como C++, Java y Python.

'MGI'

http://www.mckinsey.com/insights/mgi

Del inglés *'McKinsey Global Institute'*. El Instituto Global McKinsey es el departamento de investigación para las empresas y la economía de la fundación McKinsey. Fundado en 1990 para desarrollar el conocimiento y la comprensión de la economía mundial en evolución. La misión del MGI es proporcionar a los líderes de los sectores comerciales, públicos y sociales los hechos y puntos de vista que han de conocer para ayudar a definir sus decisiones de gestión y políticas de base.

'MICROSOFT AZURE'

http://es.wikipedia.org/wiki/Microsoft_Azure

'Microsoft Azure' (anteriormente *'Windows Azure*' y '*Azure Services Platform*') es una plataforma ofrecida como servicio y alojada en los '*Data Centers*' de Microsoft. Anunciada en el '*Professional Developers Conference*' de Microsoft (PDC) del 2008 en su versión beta, pasó a ser un producto comercial el 1 de enero del 2010. Windows Azure es una plataforma general que tiene diferentes servicios para aplicaciones, desde servicios que alojan aplicaciones en alguno de los centros de procesamiento de datos de Microsoft para que se ejecute sobre su infraestructura ('*Cloud Computing*') hasta servicios de comunicación segura entre aplicaciones.

'MIT'

http://es.wikipedia.org/wiki/Instituto_Tecnol%C3%B3gico_de_Massachusetts

El Instituto Tecnológico de Massachusetts (MIT por las iniciales de su nombre en inglés, '*Massachusetts Institute of Technology*') es una universidad privada localizada en Cambridge, Massachusetts (Estados Unidos). El MIT consta de cinco escuelas y una facultad:

- Escuela de Ciencia del MIT (MIT School of Science)
- Escuela de Ingeniería del MIT (MIT School of Engineering)
- Escuela de Arquitectura y Planeamiento del MIT (MIT School of Architecture and Planning)
- Escuela de Administración Sloan del MIT (MIT Sloan School of Management)
- Escuela de Humanidades, Artes y Ciencias Sociales del MIT (MIT School of Humanities, Arts, and Social Sciences)
- Facultad de Ciencias de la Salud y Tecnología Whitaker (Whitaker College of Health Sciences and Technology)

Incluyen un total de 32 departamentos académicos con un fuerte énfasis en la investigación, la ingeniería, y la educación tecnológica.

'ML'

http://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico

Del inglés *'Machine Learning'*. En ciencias de la computación el aprendizaje automático o aprendizaje de máquinas es una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. De forma más concreta, se trata de crear programas capaces de generalizar comportamientos a partir de una información no estructurada suministrada en forma de ejemplos. Es, por lo tanto, un proceso de inducción del conocimiento. En muchas ocasiones el campo de actuación del aprendizaje automático se solapa con el de la Estadística, ya que las dos disciplinas se basan en el análisis de datos. Sin embargo, el aprendizaje automático se centra más en el estudio de la complejidad computacional de los problemas. El aprendizaje automático puede ser visto como un intento de automatizar algunas partes del método científico mediante métodos matemáticos.

'MULTITENENCIA'

http://es.wikipedia.org/wiki/Tenencia_M%C3%BAltiple

En informática la multitenencia (o tenencia múltiple) se corresponde con un principio de arquitectura de software en la cual una sola instancia de la aplicación se ejecuta en el servidor, pero sirviendo a múltiples clientes u organizaciones (tenedor o instancia). Este modelo se diferencia de las arquitecturas con múltiples instancias donde cada organización o cliente tiene su propia instancia instalada de la aplicación. Con una arquitectura de tenencia múltiple, la aplicación puede dividir virtualmente sus datos y su configuración para que cada cliente tenga una instancia virtual adaptada a sus

requerimientos. Algunos expertos consideran la tenencia múltiple como un factor decisivo del paradigma del '*Cloud Computing*'.

N

'NATURAL LANGUAGE PROCESSING'

http://es.wikipedia.org/wiki/Procesamiento_de_lenguajes_naturales

El Procesamiento de Lenguaje Natural (abreviado PLN, o NLP del '*Natural Language Processing*') es un campo de las ciencias de la computación, inteligencia artificial y lingüística, que estudia las interacciones entre las computadoras y el lenguaje humano. El PLN se ocupa de la formulación e investigación de mecanismos eficaces computacionalmente para la comunicación entre personas y máquinas por medio de lenguajes naturales. El PLN no trata de la comunicación por medio de lenguajes naturales de una forma abstracta, sino de diseñar mecanismos para comunicarse que sean eficaces computacionalmente (que se puedan realizar por medio de programas que ejecuten o simulen la comunicación). Los modelos aplicados se enfocan no sólo a la comprensión del lenguaje en sí, sino a aspectos generales cognitivos humanos y a la organización de la memoria. El lenguaje natural sirve sólo de medio para estudiar estos fenómenos. Hasta la década de 1980, la mayoría de los sistemas de PNL se basaban en un complejo conjunto de reglas diseñadas a mano. A partir de finales de 1980, sin embargo, hubo una revolución en PNL con la introducción de algoritmos de aprendizaje automático para el procesamiento del lenguaje.

'NIST'

http://es.wikipedia.org/wiki/Instituto_Nacional_de_Est%C3%A1ndares_y_Tecnolog%C3%ADa

El Instituto Nacional de Normas y Tecnología (NIST por sus siglas del inglés, '*National Institute of Standards and Technology*'), llamada entre 1901 y 1988 Oficina Nacional de Normas (NBS por sus siglas del inglés '*National Bureau of Standards*'), es una agencia de la Administración de Tecnología del Departamento de Comercio de los Estados Unidos. La misión de este instituto es promover la innovación y la competencia industrial en Estados Unidos mediante avances en metrología, normas y tecnología de forma que mejoren la estabilidad económica y la calidad de vida.

‘NSA’

http://es.wikipedia.org/wiki/Agencia_de_Seguridad_Nacional

La Agencia de Seguridad Nacional (en inglés: ‘*National Security Agency*’, NSA), es una agencia de inteligencia del Gobierno de los Estados Unidos que se encarga de todo lo relacionado con la seguridad de la información. Con este propósito en ella trabajan distintos tipos de especialistas como matemáticos, criptógrafos, lingüistas, operadores de polígrafos, expertos en radiofrecuencias, programadores y hackers, operadores de puestos de escucha para espionaje, etc. Fue creada en secreto el 4 de noviembre de 1952 por el presidente Harry S. Truman como sucesor de la ‘*Armed Forces Security Agency*’ (AFSA).

‘NSF’

http://es.wikipedia.org/wiki/Fundaci%C3%B3n_Nacional_para_la_Ciencia

La Fundación Nacional para la Ciencia (del inglés ‘*National Science Foundation*’) es la agencia gubernamental de los Estados Unidos, que impulsa investigación y educación fundamental en todos los campos no médicos de la Ciencia y la Ingeniería. Es un líder mundial en desarrollo de normatividad y certificación de productos, educación y gestión de riesgos para la salud pública. Con un presupuesto anual de unos $6.000 millones (del año fiscal 2008), NSF financia aproximadamente el 20 por ciento de toda la investigación básica impulsada federalmente en los institutos y universidades de los Estados Unidos. En algunos campos, tales como Matemáticas, Informática, Económicas y las Ciencias Sociales, NSF es la principal fuente federal. También ha sido designado como Centro colaborador para la Organización mundial de la Salud para la alimentación y seguridad del agua y medio ambiente.

O

‘OPEN ACCESS’

http://es.wikipedia.org/wiki/Acceso_abierto

El concepto ‘Acceso Abierto’ (del inglés ‘Open Access’), es el acceso inmediato, sin requerimientos de registro, suscripción o pago (es decir sin restricciones) a material digital educativo, académico, científico o de cualquier otro tipo, principalmente artículos de investigación científica de revistas especializadas y arbitradas mediante el sistema de revisión por pares o ‘*peer review*’.

‘OPEN DATA’

http://es.wikipedia.org/wiki/Datos_abiertos

El concepto ‘Datos Abiertos’ (del inglés ‘*Open Data’*), es una filosofía y práctica que persigue que determinados tipos de datos estén disponibles de forma libre para todo el mundo, sin restricciones de derechos de autor, de patentes o de otros mecanismos de control. Tiene una ética similar a otros movimientos y comunidades abiertos, como el software libre, el código abierto (del inglés ‘*Open Source’*) y el acceso libre (del inglés ‘*Open Access’*).

‘OPEN SOURCE’

http://es.wikipedia.org/wiki/C%C3%B3digo_abierto

El concepto ‘Código Abierto’ (del inglés ‘*Open Source’*), es la expresión con la que se conoce al software distribuido y desarrollado libremente. Se focaliza más en los beneficios prácticos (acceso al código fuente) que en cuestiones éticas o de libertad que tanto se destacan en el software libre.

P

‘PAAS’

http://es.wikipedia.org/wiki/Computaci%C3%B3n_en_la_nube#Plataforma_como_servicio

Del inglés *‘Platform as a Service’* (Plataforma Como Servicio). La capa del medio es la encapsulación de una abstracción de un ambiente de desarrollo y el empaquetamiento de una serie de módulos o complementos que proporcionan, normalmente, una funcionalidad horizontal (persistencia de datos, autenticación, mensajería, etc.). De esta forma, un arquetipo de plataforma como servicio podría consistir en un entorno conteniendo una pila básica de sistemas, componentes o APIs preconfiguradas y listas para integrarse sobre una tecnología concreta de desarrollo (por ejemplo, un sistema Linux, un servidor web, y un ambiente de programación como Perl o Ruby). Las ofertas de PaaS pueden dar servicio a todas las fases del ciclo de desarrollo y pruebas del software, o pueden estar especializadas en cualquier área en particular, tal como la administración del contenido. Los ejemplos comerciales incluyen ‘*Google App Engine*’, que sirve aplicaciones de la infraestructura Google y ‘*Windows Azure*’ de Microsoft. Es una plataforma en la nube que permite el desarrollo y ejecución de aplicaciones codificadas en varios lenguajes y tecnologías como .NET, Java y PHP; y también la

Plataforma G, desarrollada en Perl. Servicios PaaS como éstos permiten gran flexibilidad, pero puede ser restringida por las capacidades que están disponibles a través del proveedor. En este modelo de servicio al usuario se le ofrece la plataforma de desarrollo y las herramientas de programación por lo que puede desarrollar aplicaciones propias y controlar la aplicación, pero no controla la infraestructura.

'PDF'

http://es.wikipedia.org/wiki/PDF

Siglas del inglés '*Portable Document Format*', formato de documento portátil. Es un formato de almacenamiento de documentos digitales independiente de plataformas de software o hardware. Este formato es de tipo compuesto (imagen vectorial, mapa de bits y texto). Fue inicialmente desarrollado por la empresa Adobe Systems, oficialmente lanzado como un estándar abierto el 1 de julio de 2008 y publicado por la Organización Internacional de Estandarización como ISO 32000-1.

'PUEM'

Bosque Sendra (2001) señala que el empleo de entidades geográficas, de carácter artificial, constituye un importante problema dentro del análisis de los datos geográficos, y está directamente relacionado con la permutabilidad de éstos. Este hecho es debido a su constitución artificial, lo que les otorga una gran arbitrariedad al no ser sus fronteras naturales ni fijas. Es lo que se conoce con el nombre de '*Problema de la Unidad Espacial Modificable'* (PUEM), y aparece debido a que los cambios en el trazado de las fronteras de entidades geográficas tiene grandes repercusiones sobre los valores alcanzados en ellas por una variable, sin que realmente haya cambiado el valor subyacente del hecho temático. Así, las autoridades pueden enmascarar datos estadísticos al enmascararlos en diferentes unidades administrativas. El análisis de las variables geográficas se ve fuertemente afectado por este efecto PUEM e incluso se ha cifrado la variación de los coeficientes de correlación lineal entre dos variables de un -0,90 a un +0,90, solo con modificar el número y el tamaño de los casos además del trazado de sus fronteras para el mismo área de estudio. En conjunto, se puede considerar que esta problemática pone en cuestión la validez de muchos análisis estadísticos realizados a partir de estas unidades de observación artificiales.

R

‘R’

http://es.wikipedia.org/wiki/R_(lenguaje_de_programaci%C3%B3n)

R es un lenguaje y entorno de programación para análisis estadístico y gráfico. Se trata de un proyecto de software libre, resultado de la implementación GNU del premiado lenguaje S. R y S-Plus (versión comercial de S) son, probablemente, los dos lenguajes más utilizados en investigación por la comunidad estadística, siendo además muy populares en el campo de la investigación biomédica, la bioinformática y las matemáticas financieras. A esto contribuye la posibilidad de cargar diferentes bibliotecas o paquetes con finalidades específicas de cálculo o gráfico. R se distribuye bajo la licencia GNU GPL y está disponible para los sistemas operativos Windows, Macintosh, Unix y GNU/Linux.

‘RFID’

http://es.wikipedia.org/wiki/RFID

Dispositivos pequeños, similares a una pegatina, que pueden ser adheridos o incorporados a un producto, un animal o una persona, y contienen antenas para permitirles recibir y responder a peticiones por radiofrecuencia desde un emisor-receptor RFID.

S

‘SAAS’

http://es.wikipedia.org/wiki/Software_como_servicio

Del inglés ‘*Software as a Service’.* Método que proporciona acceso a software y sus funciones como un servicio basado en la Web, pagando en base al consumo. SaaS permite a las organizaciones acceder a funcionalidades de negocio a un menor coste, sin los requisitos de inversión en grandes desarrollos tecnológicos. Dado que el software está alojado en remoto, los usuarios no tienen que invertir en hardware adicional para la nueva aplicación. SaaS elimina las necesidades de instalación, puesta en marcha, conservación y mantenimiento.

‘SHARED POOL’

http://es.wikipedia.org/wiki/%C3%81rea_Global_del_Sistema

En esta zona se encuentran las sentencias SQL que han sido analizadas. El análisis sintáctico de las sentencias SQL lleva su tiempo y Oracle mantiene las estructuras asociadas a cada sentencia SQL analizada durante el tiempo que pueda para ver si puede reutilizarlas. Antes de analizar una sentencia SQL, Oracle mira a ver si encuentra otra sentencia exactamente igual en la zona de SQL compartido. Si es así, no la analiza y pasa directamente a ejecutar la que mantiene en memoria. De esta manera se premia la uniformidad en la programación de las aplicaciones. La igualdad se entiende que es lexicográfica, espacios en blanco y variables incluidas. La base de datos Oracle asigna memoria a la Shared Pool cuando una nueva instrucción SQL se analiza. El tamaño de esta memoria depende de la complejidad de la instrucción. Si toda la Shared Pool ya ha sido asignada la base de datos Oracle puede liberar elementos de la Shared Pool hasta que haya suficiente espacio libre para nuevas sentencias. Al liberar un elemento de la Shared Pool el SQL asociado debe ser recompilado y reasignado a otra área de SQL compartida la próxima vez que se ejecute.

‘SMART CITY’

http://es.wikipedia.org/wiki/Ciudad_inteligente

La expresión ‘*Ciudad Inteligente’* es la traducción y adaptación del término en idioma inglés ‘*Smart City*’. Es un concepto emergente, y por tanto sus acepciones en español y en otros idiomas, incluso en el propio inglés, están sujetas a constante revisión. Es también un término actual, que se está utilizando como un concepto de marketing (mercadotecnia) en el ámbito empresarial, en relación a políticas de desarrollo, y en lo concerniente a diversas especialidades y temáticas. La ‘*Ciudad Inteligente’* a veces también llamada ‘*Ciudad Eficiente’* o ‘*Ciudad Súper-Eficiente’*, se refiere a un tipo de desarrollo urbano basado en la sostenibilidad que es capaz de responder adecuadamente a las necesidades básicas de instituciones, empresas, y de los propios habitantes, tanto en el plano económico, como en los aspectos operativos, sociales y ambientales. Una ciudad o complejo urbano podrá ser calificado de inteligente en la medida que las inversiones que se realicen en capital humano (educación permanente, enseñanza inicial, enseñanza media y superior, educación de adultos…), en aspectos sociales, en infraestructuras de energía (electricidad, gas), tecnologías de comunicación (electrónica, Internet) e infraestructuras de transporte, contemplen y promuevan una calidad de vida elevada, un desarrollo económico-

ambiental durable y sostenible, una gobernanza participativa, una gestión prudente y reflexiva de los recursos naturales, y un buen aprovechamiento del tiempo de los ciudadanos. Las ciudades modernas, basadas en infraestructuras eficientes y durables de agua, electricidad, telecomunicaciones, gas, transportes, servicios de urgencia y seguridad, equipamientos públicos, edificaciones inteligentes de oficinas y de residencias, etc., deben orientarse a mejorar el confort de los ciudadanos, siendo cada vez más eficaces y brindando nuevos servicios de calidad, mientras que se respetan al máximo los aspectos ambientales y el uso prudente de los recursos naturales no renovables.

'SMARTPHONE'

http://es.wikipedia.org/wiki/Tel%C3%A9fono_inteligente

Un teléfono inteligente (en inglés: Smartphone) es un teléfono móvil construido sobre una plataforma informática móvil, con una mayor capacidad de almacenar datos y realizar actividades semejantes a una minicomputadora, y con una mayor conectividad que un teléfono móvil convencional. El término 'inteligente', que se utiliza con fines comerciales, hace referencia a la capacidad de usarse como un computador de bolsillo, y llega incluso a reemplazar a un computador personal en algunos casos.

'SMTP'

http://es.wikipedia.org/wiki/Simple_Mail_Transfer_Protocol

El '*Simple Mail Transfer Protocol*' (SMTP) (Protocolo para la Transferencia Simple de Correo Electrónico), es un protocolo de red utilizado para el intercambio de mensajes de correo electrónico entre computadoras u otros dispositivos (PDA, teléfonos móviles, etc.)

'SPAM'

http://es.wikipedia.org/wiki/Spam

Se llama spam, correo basura o mensaje basura a los mensajes no solicitados, no deseados o de remitente no conocido (correo anónimo), habitualmente de tipo publicitario, generalmente enviados en grandes cantidades (incluso masivas) que perjudican de alguna o varias maneras al receptor. La acción de enviar dichos mensajes se denomina '*Spamming'*. La palabra spam proviene de la segunda guerra mundial, cuando los familiares de los soldados en guerra les enviaban comida enlatada; entre estas comidas enlatadas estaba una carne enlatada llamada spam, que en los Estados Unidos era y sigue siendo muy común. No obstante, este término

comenzó a usarse en la informática décadas más tarde al popularizarse por un sketch de los Monty Python en el que se incluía spam en todos los platos. Aunque se puede hacer spam por distintas vías, la más utilizada entre el público en general es la basada en el correo electrónico. Otras tecnologías de Internet que han sido objeto de correo basura incluyen grupos de noticias, Usenet, motores de búsqueda, redes sociales, páginas web, wiki, foros, blogs, a través de ventanas emergentes y todo tipo de imágenes y textos en la web.

'SQL'

http://es.wikipedia.org/wiki/SQL

El lenguaje de consulta estructurado o SQL (por sus siglas en inglés '*Structured Query Language*') es un lenguaje declarativo de acceso a bases de datos relacionales que permite especificar diversos tipos de operaciones en ellas. Una de sus características es el manejo del álgebra y el cálculo relacional que permiten efectuar consultas con el fin de recuperar de forma sencilla información de interés de bases de datos, así como hacer cambios en ellas.

'STREAMING'

http://es.wikipedia.org/wiki/Streaming

El streaming (también denominado lectura en continuo, difusión en flujo, lectura en tránsito, difusión en continuo, descarga continua o media flujo) es la distribución de multimedia a través de una red de computadoras de manera que el usuario consume el producto, generalmente archivo de video o audio, en paralelo mientras se descarga. La palabra streaming se refiere a una corriente continua (que fluye sin interrupción). Este tipo de tecnología funciona mediante un búfer de datos que va almacenando lo que se va descargando en la estación del usuario para luego mostrarle el material descargado. Esto se contrapone al mecanismo de descarga de archivos, que requiere que el usuario descargue por completo los archivos para poder acceder a su contenido.

T

'TCP/IP'

http://es.wikipedia.org/wiki/Familia_de_protocolos_de_Internet

La familia de protocolos de Internet es un conjunto de protocolos de red en los que se basa Internet y que permiten la transmisión de datos entre computadoras. En

ocasiones se le denomina conjunto de protocolos TCP/IP, en referencia a los dos protocolos más importantes que la componen: Protocolo de Control de Transmisión (TCP) y Protocolo de Internet (IP), que fueron dos de los primeros en definirse, y que son los más utilizados de la familia. Existen tantos protocolos en este conjunto que llegan a ser más de 100 diferentes, entre ellos se encuentra el popular HTTP ('*HyperText Transfer Protocol*'), que es el que se utiliza para acceder a las páginas web, además de otros como el ARP ('*Address Resolution Protocol*') para la resolución de direcciones, el FTP ('*File Transfer Protocol*') para transferencia de archivos, y el SMTP ('*Simple Mail Transfer Protocol')* y el POP ('*Post Office Protocol*') para correo electrónico, TELNET para acceder a equipos remotos, entre otros. TCP/IP fue desarrollado y demostrado por primera vez en 1972 por el Departamento de Defensa de los Estados Unidos, ejecutándolo en ARPANET.

'TELNET'

http://es.wikipedia.org/wiki/Telnet

Telnet ('*TELecommunication NETwork*') es el nombre de un protocolo de red que nos permite viajar a otra máquina para manejarla remotamente como si estuviéramos sentados delante de ella. También es el nombre del programa informático que implementa el cliente. Para que la conexión funcione, como en todos los servicios de Internet, la máquina a la que se acceda debe tener un programa especial que reciba y gestione las conexiones.

'TIC'

http://es.wikipedia.org/wiki/Tecnolog%C3%ADas_de_la_informaci%C3%B3n_y_la_comunicaci%C3%B3n

'*Tecnologías de la Información y la Comunicación'*. A veces denominadas nuevas tecnologías de la información y la comunicación (NTIC) son un concepto muy asociado a la Informática. Si se entiende esta última como el conjunto de recursos, procedimientos y técnicas usadas en el procesamiento, almacenamiento y transmisión de información, esta definición se ha matizado de la mano de las TIC, pues en la actualidad no basta con hablar de una computadora cuando se hace referencia al procesamiento de la información. Internet puede formar parte de ese procesamiento que, quizás, se realice de manera distribuida y remota. Y al hablar de procesamiento remoto, además de incorporar el concepto de telecomunicación, se puede estar haciendo referencia a un dispositivo muy distinto a lo que tradicionalmente se entiende por computadora pues podría llevarse a cabo, por ejemplo, con un teléfono móvil o

una computadora ultra-portátil, con capacidad de operar en red mediante comunicación inalámbrica y con cada vez más prestaciones, facilidades y rendimiento.

U

'UIT'

http://es.wikipedia.org/wiki/Uni%C3%B3n_Internacional_de_Telecomunicaciones

La '*Unión Internacional de Telecomunicaciones'* es el organismo especializado de Telecomunicaciones de la Organización de las Naciones Unidas encargado de regular las telecomunicaciones a nivel internacional entre las distintas administraciones y empresas operadoras.

'USENET'

http://es.wikipedia.org/wiki/Usenet

Usenet es el acrónimo de '*Users Network*' (Red de Usuarios), consistente en un sistema global de discusión en Internet, que evoluciona de las redes UUCP (acrónimo del inglés '*Unix to Unix CoPy*', 'Copiador de Unix a Unix'). Fue creado por Tom Truscott y Jim Ellis, estudiantes de la Universidad de Duke, en 1979. Los usuarios pueden leer o enviar mensajes (denominados artículos) a distintos grupos de noticias ordenados de forma jerárquica. El medio se sostiene gracias a un gran número de servidores distribuidos y actualizados mundialmente, que guardan y transmiten los mensajes.

W

'WEB SEMÁNTICA'

http://es.wikipedia.org/wiki/Web_sem%C3%A1ntica

La web semántica (del inglés '*Semantic Web'*) es un conjunto de actividades desarrolladas en el seno de '*World Wide Web Consortium*' tendente a la creación de tecnologías para publicar datos legibles por aplicaciones informáticas (máquinas en la terminología de la web semántica). Se basa en la idea de añadir metadatos semánticos y ontológicos a la '*World Wide Web*'. Esas informaciones adicionales (que describen el contenido, el significado y la relación de los datos) se deben proporcionar de manera formal, para que así sea posible evaluarlas automáticamente por máquinas de procesamiento. El objetivo es mejorar Internet ampliando la interoperabilidad entre

los sistemas informáticos usando '*agentes inteligentes*' (programas en las computadoras que buscan información sin operadores humanos).

'WIKIPEDIA'

http://es.wikipedia.org/wiki/Wikipedia

Wikipedia es una enciclopedia libre, políglota y editada colaborativamente. Es administrada por la Fundación Wikimedia, una organización sin ánimo de lucro. Sus más de 37 millones de artículos en 284 idiomas (cantidad que incluye idiomas artificiales como el esperanto, lenguas indígenas o aborígenes como el náhuatl, el maya y las lenguas de las islas Andamán, o lenguas muertas, como el latín, el chino clásico o el anglosajón) han sido redactados conjuntamente por voluntarios de todo el mundo y prácticamente cualquier persona con acceso al proyecto puede editarlos. Iniciada en enero de 2001 por Jimmy Wales y Larry Sanger, es la mayor y más popular obra de consulta en Internet.

X

'XEROX PALO ALTO RESEARCH CENTER'

http://es.wikipedia.org/wiki/Xerox_PARC

Xerox PARC ('*Palo Alto Research Center*' o su equivalente en castellano: 'Centro de Investigación de Palo Alto') es una empresa de investigación y desarrollo, propiedad de Xerox Corporation. Con sede en Palo Alto (California, EE.UU.), fue fundado en 1970 inicialmente como una división de investigación. Desde entonces ha sido reconocida mundialmente por sus contribuciones e importantes desarrollos en la industria del hardware y software, siendo el creador de algunos de los estándares actuales más comúnmente usados. El Xerox PARC ha sido responsable por desarrollos bien conocidos e importantes tales como la impresión por láser, el estándar Ethernet, el moderno computador personal, la interfaz gráfica de usuario (GUI), la metáfora de escritorio, la programación orientada a objetos, la computación ubicua, aplicaciones de silicio amorfo (a-Si), y avances en el desarrollo del dispositivo apuntador ratón o mouse y los semiconductores de muy alta escala de integración (VLSI). En su momento, Xerox llegó a invertir más de 100 millones de dólares en diversos proyectos que aunque no llegaran a rentabilizar sus esfuerzos en algunos de sus productos, se convirtieron en la principal fuente de inspiración de las nuevas tecnologías de los años setenta y buena parte de los ochenta.