01MBID Fundamentos de la tecnología Big Data

Universidad Internacional de Valencia

Sesión 2
Tema 1 + Foro + Tema 2

De:
Planeta Formación y Universidades

Agenda

- **Dudas**
- **Tema 1 2de2: Introducción a Big Data**
- **Foro**
- **Tema 2: Fuentes de datos en entornos Big data**

Agenda

- **Dudas**
- **Tema 1 2de2: Introducción a Big Data**
- **Foro**
- **Tema 2: Fuentes de datos en entornos Big data**

Agenda

- Dudas
- **Tema 1 2de2: Introducción a Big Data**
- **Foro**
- **Tema 2: Fuentes de datos en entornos Big data**

Seguimos con el tema 1: Introducción a Big Data

Introducción a Big Data

1) ¿Qué es Big Data?

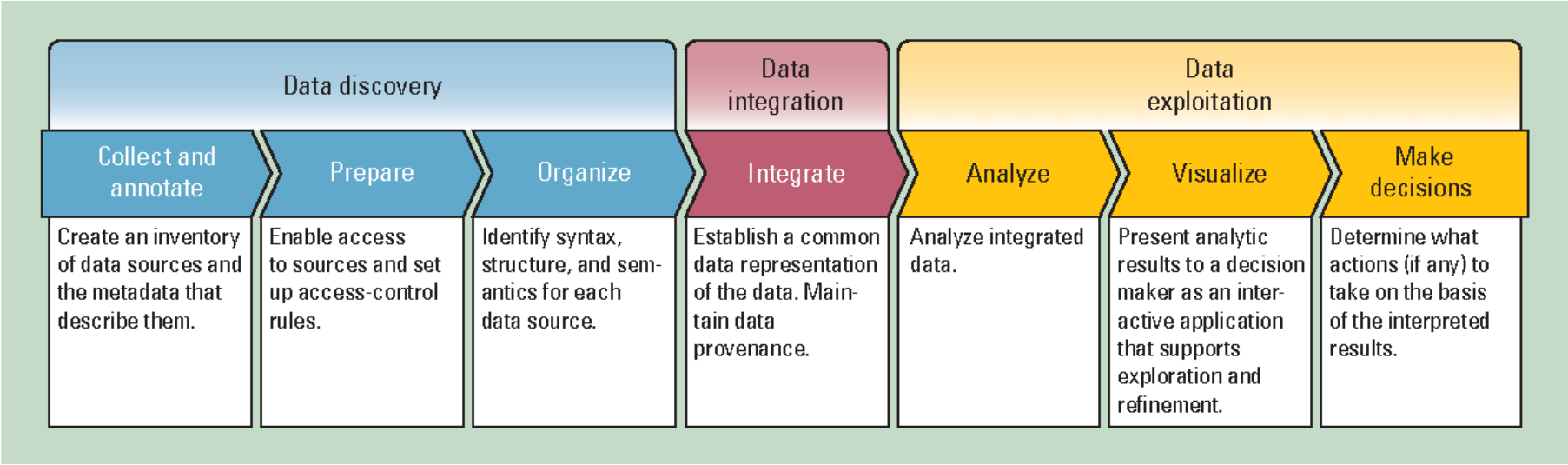
2) Cadena de valor y áreas del Big Data

3) Definiciones relacionadas con Big Data

4) Perfiles profesionales Big Data

Cadena de valor y áreas del Big Data

Cadena de valor y áreas del Big Data



Si hay precedencia, pero no es necesario hacer todo el proceso

* Miller, H.G., & Mork, P. (2013). From Data to Decisions: A Value Chain for Big Data. *IT Professional, 15*, 57-59.

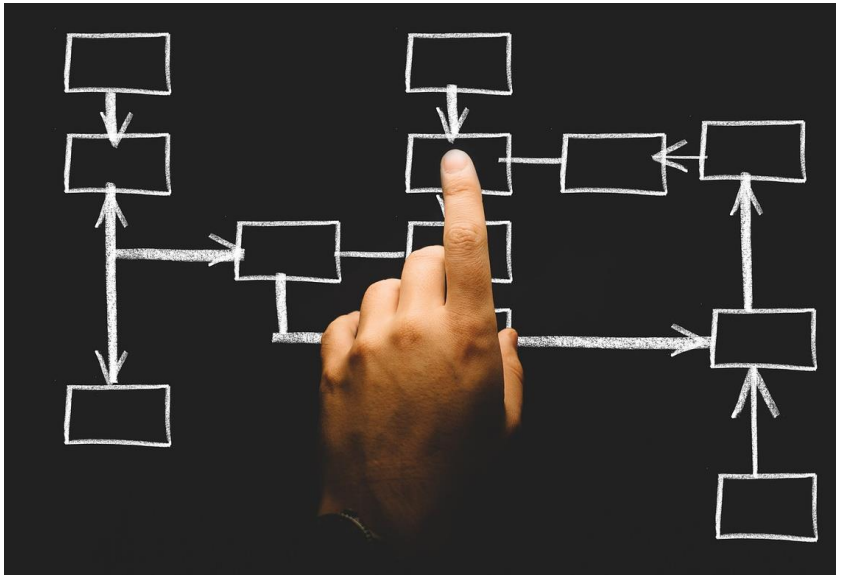
Cadena de valor y áreas del Big Data

Áreas

- Integración
- Infraestructura
- Preservación
- Análisis
- Explotación
- Visualización

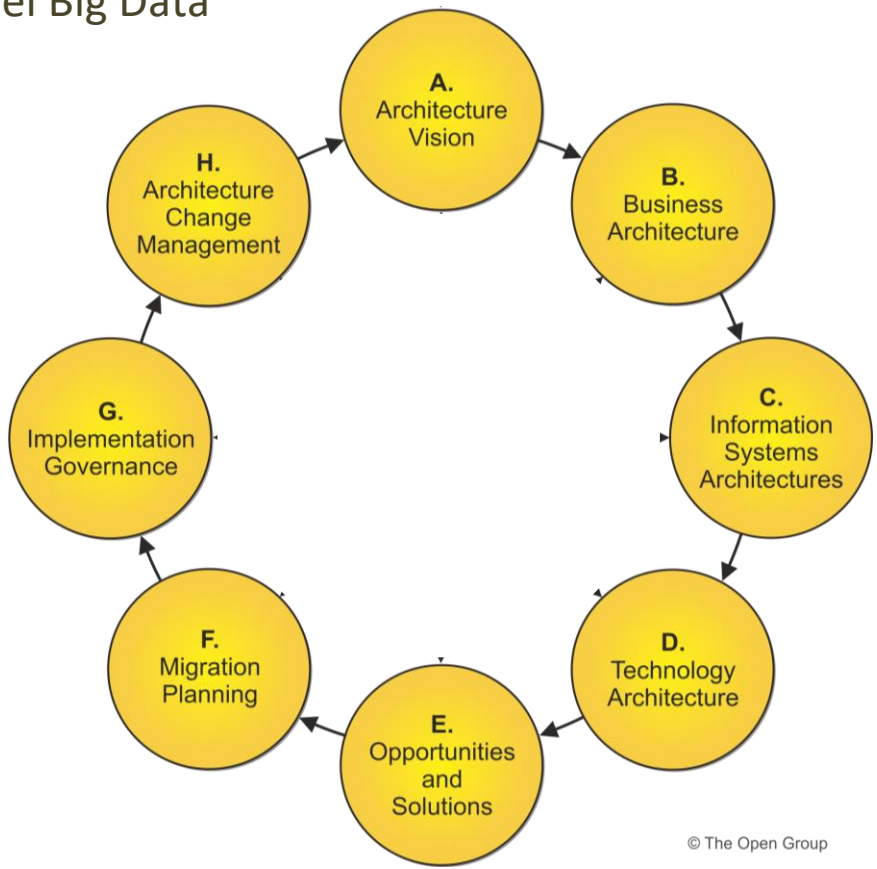
Integración

- Diferentes fuentes de datos
- Interoperabilidad
  - Interna
  - Externa
  - Legal
  - ...



Cadena de valor y áreas del Big Data

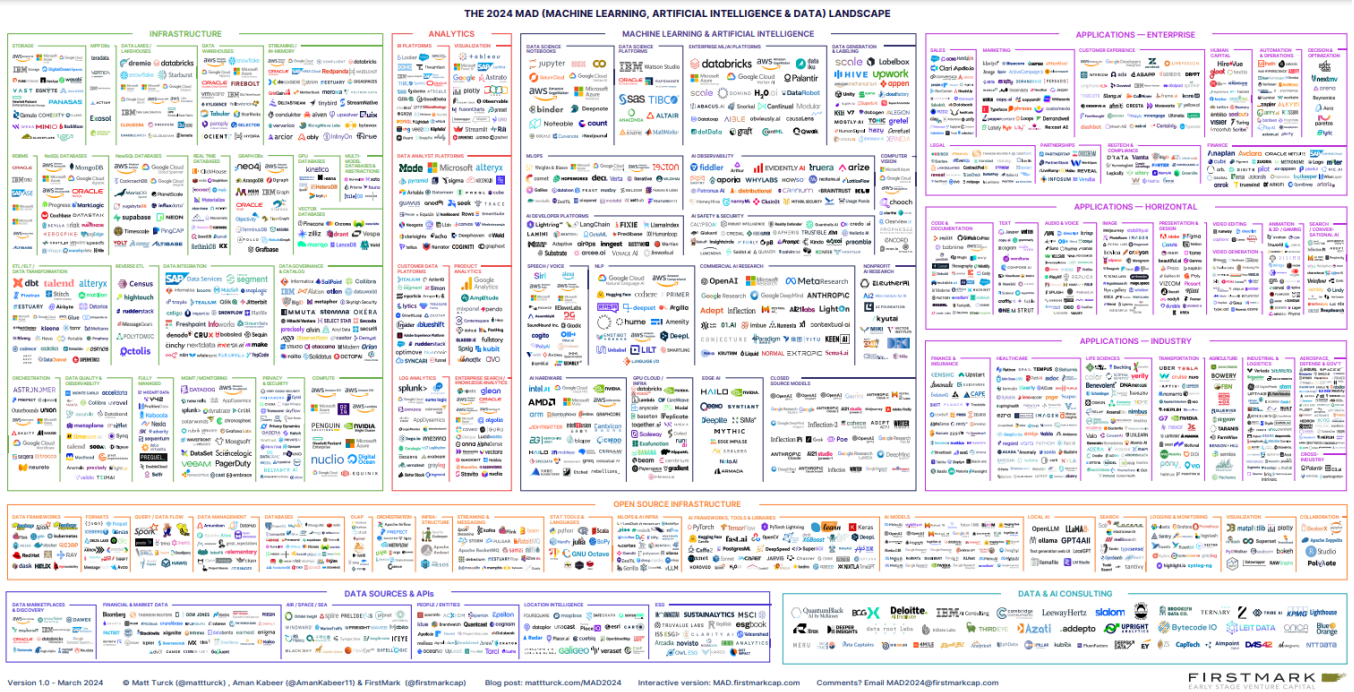
Integración



Universidad Internacional de Valencia

Cadena de valor y áreas del Big Data

Infraestructura

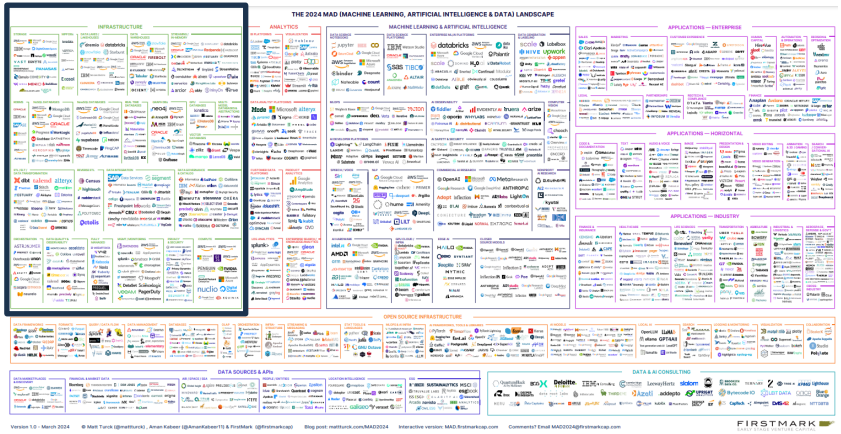


https://mattturck.com/mad2024/

Universidad Internacional de Valencia

Cadena de valor y áreas del Big Data

Infraestructura



https://mattturck.com/mad2024/

Preservación

RAE: Preservar es "*Proteger, resguardar anticipadamente alguien o algo, de algún daño o peligro.*"

- Preservación digital, de Miquel Térmens (2014): *"la preservación digital se refiere a una serie de actividades necesarias y muy bien administradas para asegurar el acceso continuo a los materiales digitales, por el periodo que sea necesario".*

- Problemas: Obsolescencia, Riesgos, Migraciones, Streaming, …

Preservación

Big Data

Hay que garantizar que la información almacenada, gestionada y procesada, perdure en el tiempo:

- Integridad
- Autenticidad
- Fiabilidad
- Legibilidad
- Funcionalidad

Cadena de valor y áreas del Big Data

Análisis

Pronóstico

- *Predicción de ventas*
- *Predicción de la carga de un servidor*
- *Evolución de una pandemia*

Riesgo y probabilidad

- *Elección de los mejores clientes para una campaña publicitaria*
- *Evaluar la conveniencia o no de aplicar una vacuna de forma masiva a la población*
- *Diagnósticos de enfermedades*

Cadena de valor y áreas del Big Data

Análisis

Recomendaciones

- *Determinación de los productos que se pueden vender juntos*
- *Recomendación de políticas públicas de salud*

Búsqueda de secuencias

- *Análisis de los artículos que los clientes han introducido en el carrito de la compra y predicción de posibles eventos*
- *En función de los síntomas detectados, sugerir pruebas para detectar los síntomas más probables y así determinar la enfermedad*

Cadena de valor y áreas del Big Data

Análisis

Agrupación

- *Distribución de clientes en grupos relacionados, y análisis y predicción de afinidades*
- *Determinación de grupos de riesgos para determinadas enfermedades*

Descriptivo

- *Para saber qué hacer para que suceda un determinado efecto*

Prescriptivo

- *Cómo actuar*

Cadena de valor y áreas del Big Data

Explotación

¿Quién puede acceder a los datos?

¿Para qué?

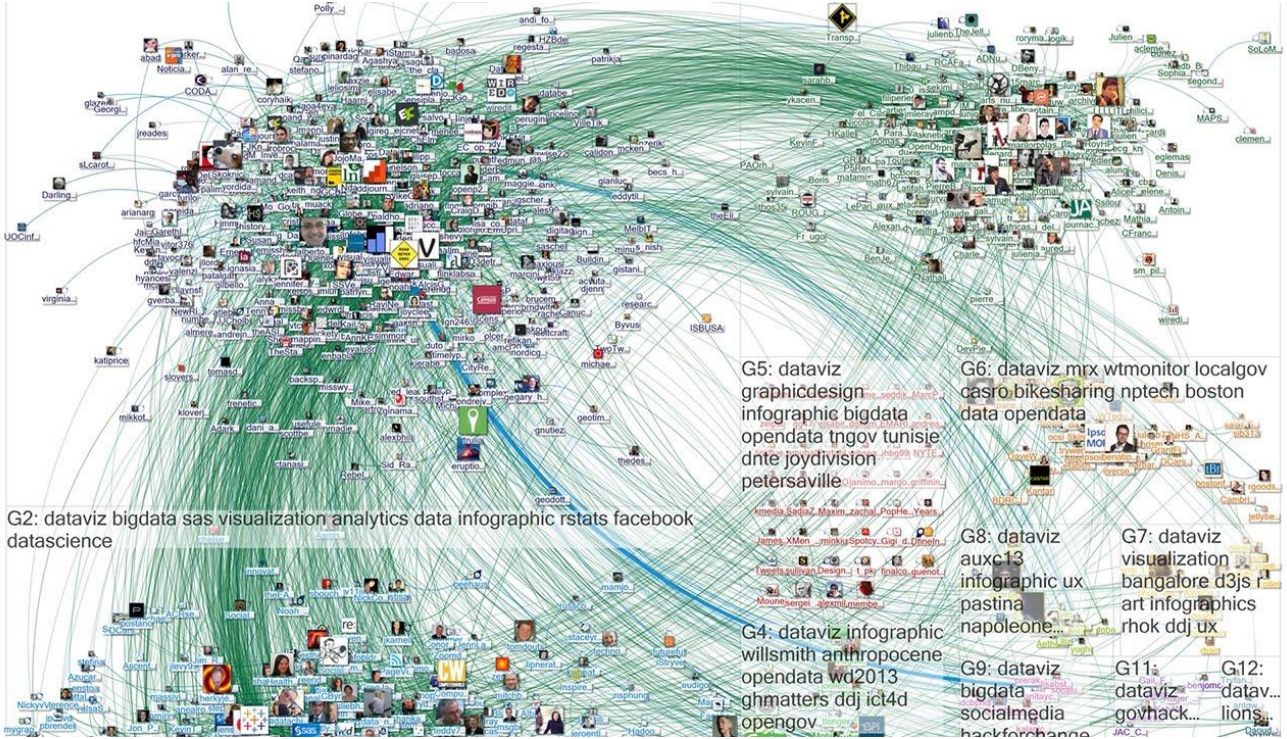
¿Qué uso se les da a los datos?

*(Gomez Garcia & Conesa i Caralt, 2015)

Universidad Internacional de Valencia

Cadena de valor y áreas del Big Data

Visualización

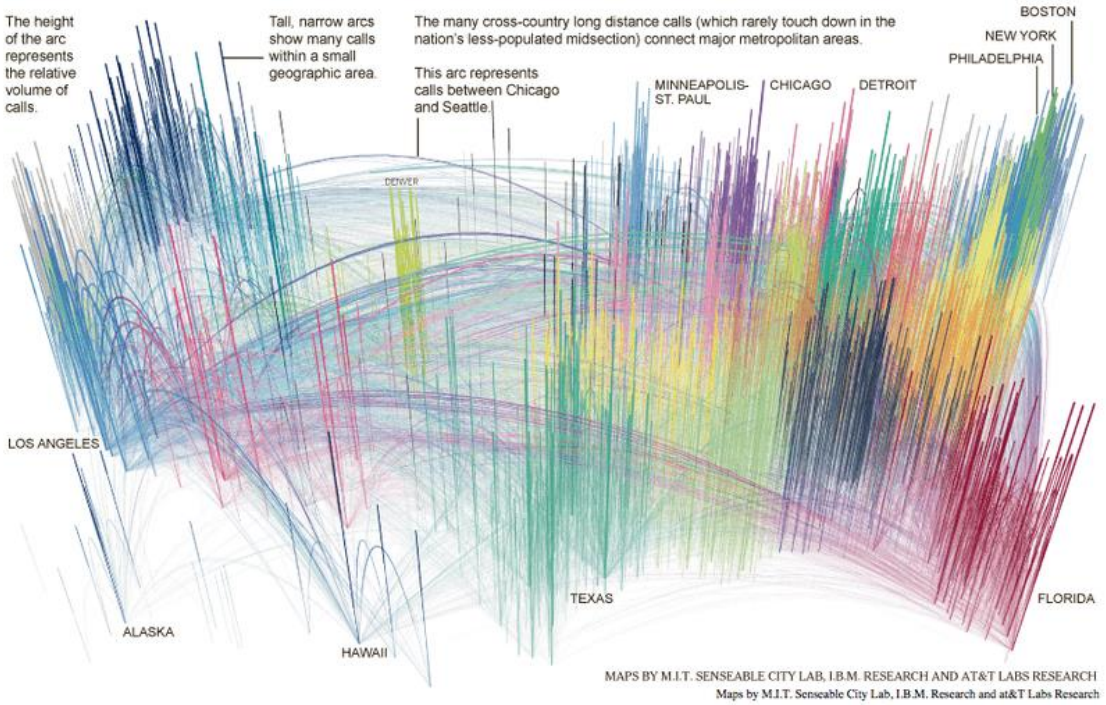


Interrelaciones

Universidad Internacional de Valencia

Cadena de valor y áreas del Big Data

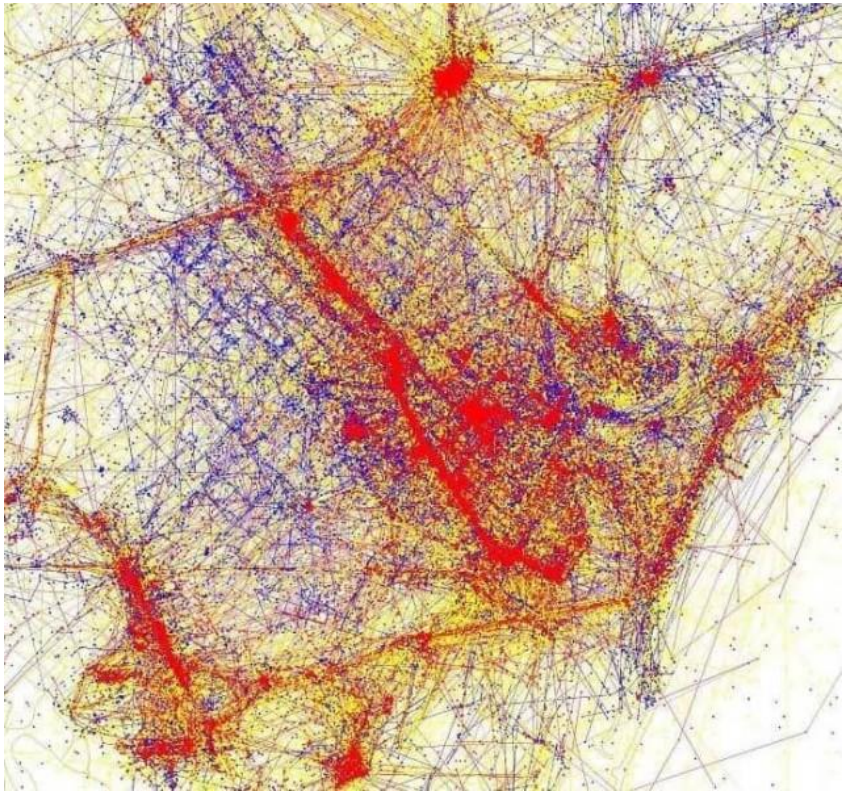
Visualización



Llamadas USA

Cadena de valor y áreas del Big Data

Visualización



Title: Locals vs. Tourists Visualized Through Geotagged Images Credits: Courtesy of Eric Fischer, under a Creative Commons Attribution-ShareAlike license. Base map data © OpenStreetMap contributors.

Introducción a Big Data

1) ¿Qué es Big Data?

2) Cadena de valor y áreas del Big Data

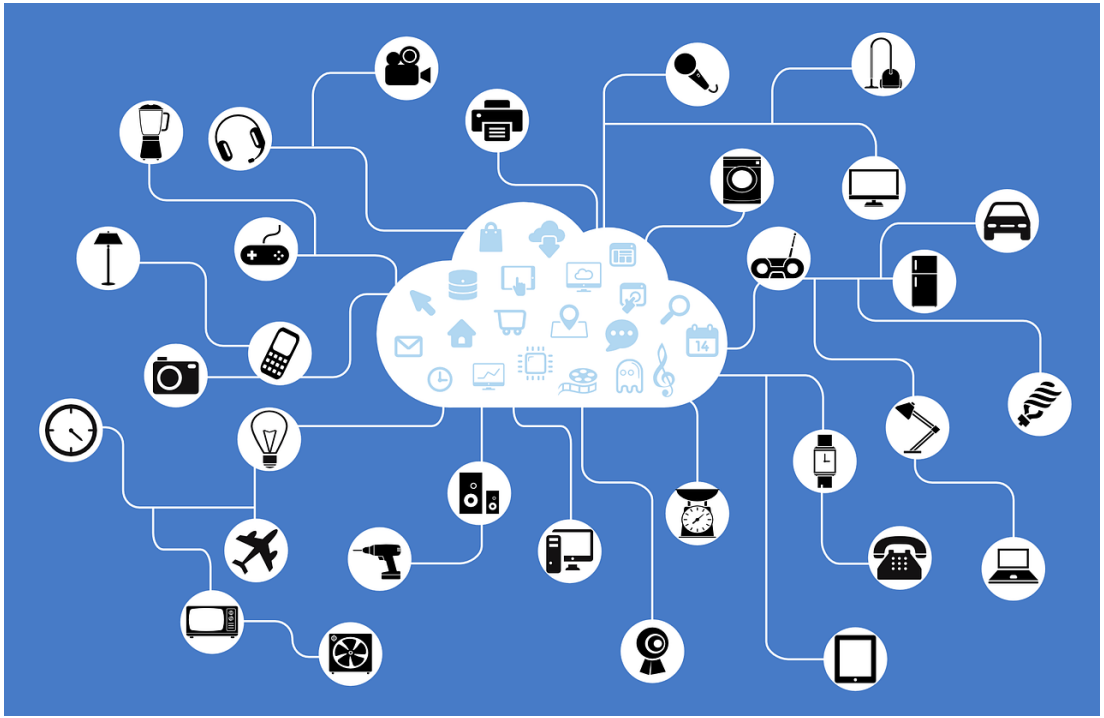
3) Definiciones relacionadas con Big Data

4) Perfiles profesionales Big Data

Definiciones relacionadas con Big Data

Definiciones relacionadas con Big Data

Internet of Things (IoT)



Internet of Things (IoT)

Big Data:

- Gran cantidad de datos
- Procesamiento de los datos
- Tiempo real

Data Analytics

Es el proceso de **examinar conjuntos de datos para extraer conclusiones** sobre la información que contienen, cada vez más con la ayuda de sistemas y software especializados.

Fases:

- Requisitos de datos
- Adquisición de datos
- Procesamiento de datos
- Limpieza de datos
- Modelado y algoritmos
- Informar/Visualizar/Comunica datos

≈ Big Data

* “Data Science and Data Scientist” de Alex Liu (2015)

Data Science

Campo **multidisciplinar** sobre procesos y sistemas para **extraer información** de **grandes volúmenes de datos** en diversas formas, ya sea **estructuradas** o **no estructuradas.**

Se considerar la continuació de ***Knowledge Discovery*** o ***Data Mining***.

** "Doing Data Science. Straight Talk from the Frontline" de Cathy O'Neil,Rachel Schutt. O'Really Media (2013).*

Data Science vs Data Analytics

En general, ***Data Science*** es un termino mucho **más amplio** que abarca a *Data Analytics* y también a *Big Data*.

Data Analytics se centra en el análisis de datos para la **extracción de conclusiones útiles** para el negocio.

Data Science engloba una serie de fundamentos y técnicas que tienen **muchos campos de aplicación**, como son: búsquedas en internet, sistemas de recomendación, publicidad digital, etc.

Data Mining

Objetivo es **descubrir patrones** y **extraer información** para transformarla.

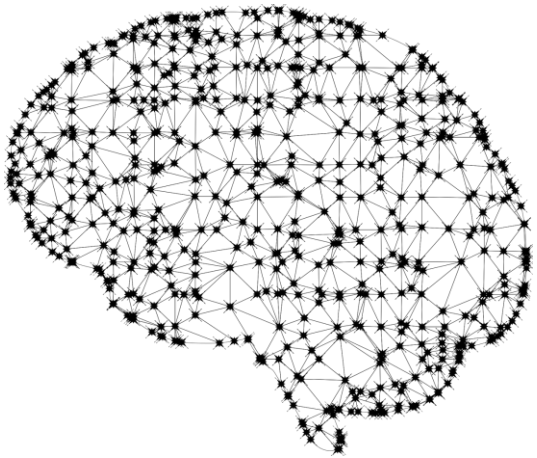
Data Mining = KDD

- *"Knowledge Discovery from Data"*
- *"Knowledge Discovery in Databases"*

Machine Learning

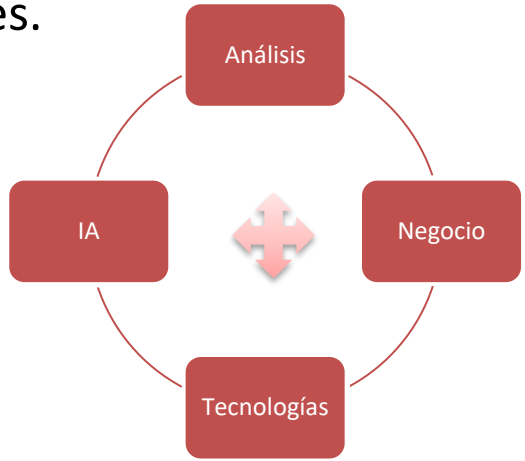
Una **especialidad** del área de **Inteligencia Artificial** para el aprendizaje automático dirigido al desarrollo de **técnicas para** que las **máquinas puedan aprender y tomar decisiones de forma autónoma**.

- AlgoritmosGenéticos
- Deep Learning
- Redes Bayesianas
- Simulación
- ...



Business Intelligence

Proceso dirigido por tecnologías para analizar los datos y transformarlos en información y conocimiento que den valor al negocio. Optimizando el proceso de toma de decisiones.



* *"BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT" de H.Chen, R.H.L.Chiang, y V.C. Storey. MIS Quarterly, 2012.*

Introducción a Big Data

1) ¿Qué es Big Data?

2) Cadena de valor y áreas del Big Data

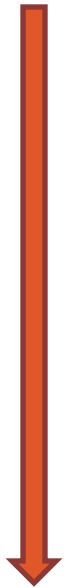
3) Definiciones relacionadas con Big Data

4) **Perfiles profesionales Big Data**

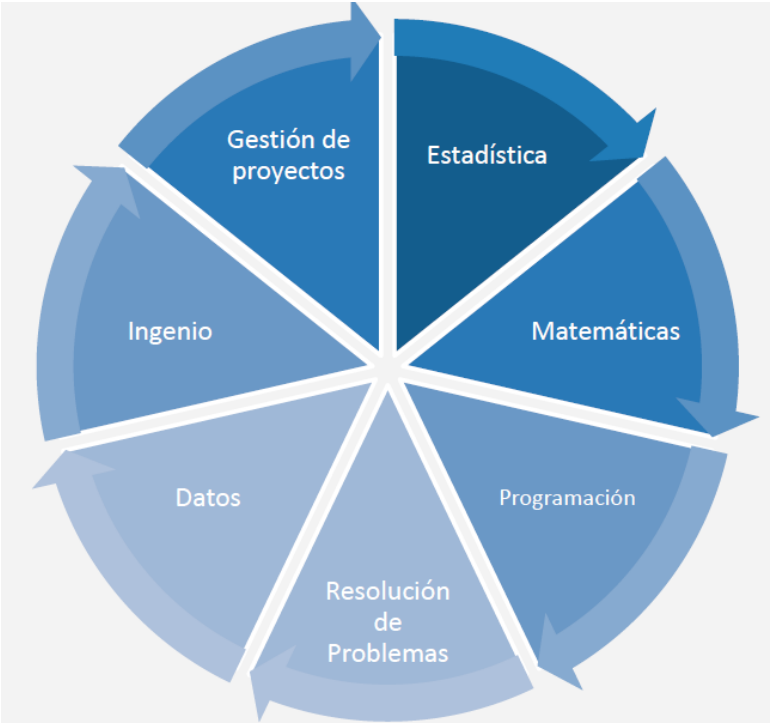
Perfiles profesionales Big Data

Perfiles profesionales Big Data

- Jefe Ejecutivo de los Datos - Chief Data Officer (CDO)
- Científico de datos
- Administrador de datos
- Ingeniero de Datos
- Otros



Perfiles profesionales Big Data



Jefe Ejecutivo de los Datos

Chief Data Officer (CDO)  CIO (Chief Information Officer), Jefe de Informática

- Responsable de los datos de la organización y de definir la estrategia de datos.

- Responsable de diseñar, implementar y supervisar el gobierno de datos de la organización en la que es responsable.

- Conocimientos de gestión de proyectos Big Data y de metodologías de desarrollo.

Científico de datos

- Procesa los datos para descubrir conocimiento, alcanzando toda la cadena de valor de Big Data.

- Recopila datos y los analiza para crear modelos de predicción.

- Competencias:
  - Estadística
  - Matemática
  - Programación
  - Resolución de Problemas
  - Gestión de los datos
  - Ingenio / Curiosidad

Administrador de datos

Data curator, data manager, o gestor de datos

- Entiende, gestiona y custodia los datos.

- Específicamente: preservar, mantener, archivar y depositar los datos para mantenerlos seguros, intactos y accesibles para su reutilización.

- Otros perfiles profesionales más concretos:
  - Higienista de datos
  - Explorador de datos.
  - Arquitecto de soluciones de negocio.
  - Experto en campañas.

Ingeniero de Datos

- Proporciona los datos de una manera accesible y apropiada a los usuarios y a los *Data Scientists.*

- Especializado en infraestructura tecnológica Big Data.

- Desarrolla y explota técnicas, procesos, herramientas y métodos que deben servir para el desarrollo de aplicaciones Big Data.

- Gran conocimiento en gestión de bases de datos, arquitecturas de clústeres, así como habilidades de programación y sistemas de procesamiento de datos.

Otros

- ***Business Data Analyst*** (Analista de Datos) participa en las iniciativas y proyectos de análisis de datos. Es la persona que recoge las necesidades de los usuarios de negocio para los *Data Scientisty* presenta resultados obtenidos.

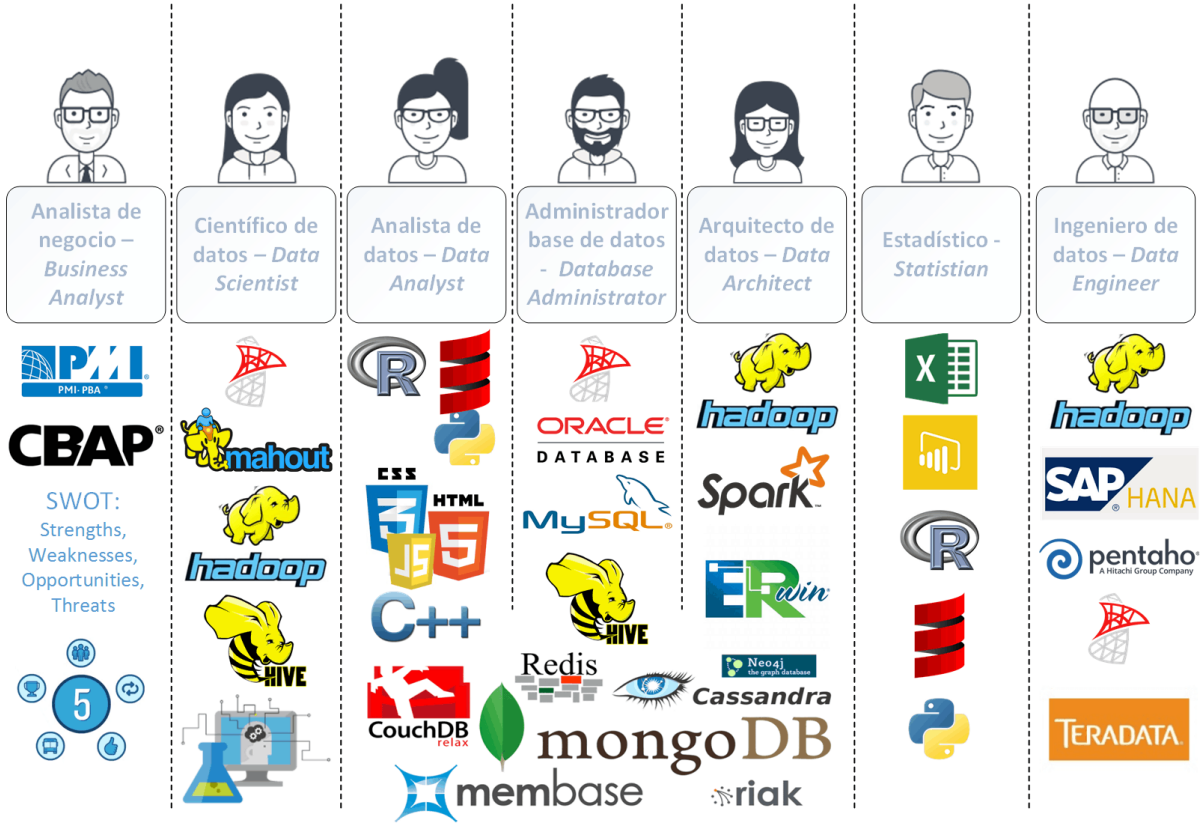
- ***Data Artist*** (*Business Analytics)* y son los responsables de crear los gráficos, infografías y otras herramientas visuales para ayudar a las diferentes personas de la organización a comprender datos complejos y el resultado de los análisis de dichos datos.

Otros

- **Estadístico**, dar soporte a actividades concretas de los otros perfiles profesionales.

- **Matemático**, dar soporte a actividades concretas de los otros perfiles profesionales.

Perfiles



** Relación de perfiles profesionales y tecnologías Big Data. Fuente: Rayo (2016).*

Agenda

- Dudas
- Tema 1 2de2: Introducción a Big Data
- **Foro**
- **Tema 2: Fuentes de datos en entornos Big data**

Foro Evaluado

Foro Evaluado

El foro no está disponible. Disponible a partir de: miércoles 23 de octubre de 2024 18H00' CEST.

Realiza dos aportaciones al foro, comentando al menos dos artículos. 11

Cada artículo tiene su propio hilo en el foro, por lo que debes hacer tus comentarios en el hilo correspondiente.

Si deseas hacer más de una contribución sobre algún artículo, ¡adelante! El requisito es hacer al menos una aportación por cada artículo leído, pero no hay un máximo de intervenciones.

Las contribuciones deben ser razonables, evitando que sean ni demasiado breves ni excesivamente largas.

CADENA

A qualitative and quantitative comparison between Web scraping and API methods for Twitter credibility analysis

Big Data A Review

Big data adoption-State of the art and research challenges

Big Data in Agriculture- A Challenge for the Future

Big data in healthcare management, analysis and future prospects

Big IoT Data Analytics- Architecture, Opportunities, and Open Research Challenges

Credibility Analysis for Available Information Sources on the Web- A Review and a Contribution

Credibility Analysis on Twitter Considering Topic Detection

Debating big data-A literature review on realizing value from big data

T-CREo_A_Twitter_Credibility_Analysis_Fra...

Web Scraping versus Twitter API- A Comparison for a Credibility Analysis

Agenda

- Dudas
- Tema 1 2de2: Introducción a Big Data
- Foro
- **Tema 2: Fuentes de datos en entornos Big data**

Tema 2: Fuentes de datos en entornos Big data

Fuentes de datos en entornos Big data

1) ¿Qué es una fuentes de datos?
2) Diferencias respecto a las tecnologías de datos tradicionales
3) Tipos de datos y flujo de datos
4) Calidad de Datos
5) Las V's del Big Data

Fuentes de datos en entornos Big data

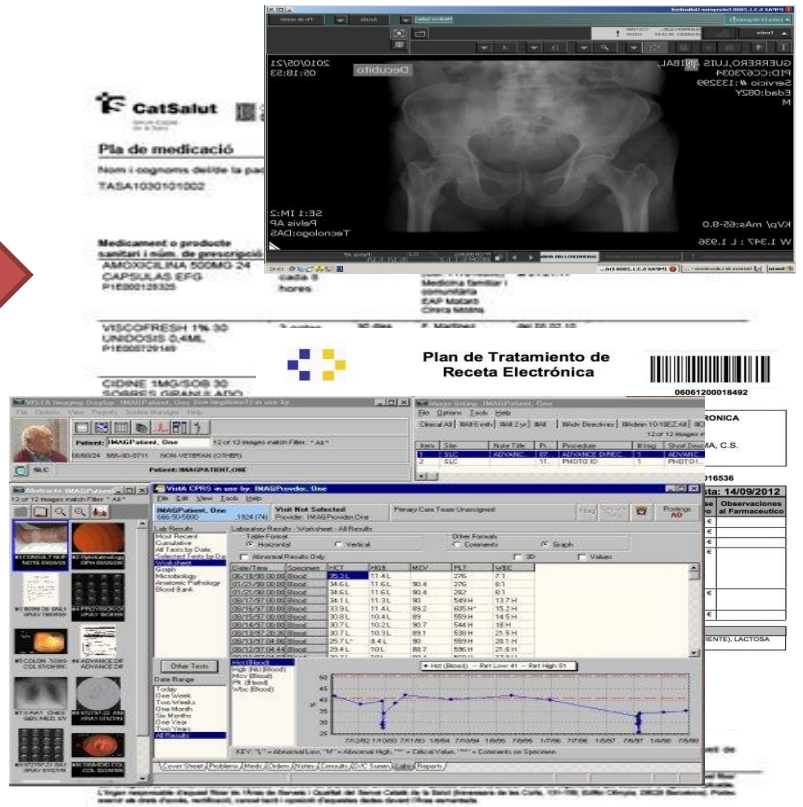
1) **¿Qué es una fuentes de datos?**
2) **Diferencias respecto a las tecnologías de datos tradicionales**
3) **Tipos de datos y flujo de datos**
4) **Calidad de Datos**
5) **Las V's del Big Data**

¿Qué es una fuentes de datos?

Cuando hablamos de fuentes de datos nos referimos a **información digital o** que es **digitalizada** para su procesamiento.

Los datos pueden proporcionarse de distintas formas y en una gran **variedad de formatos**.

¿Qué es una fuentes de datos?



¿Qué es una fuentes de datos?

- **Biométricos:** ADN, reconocimiento facial, …
- **IoT:** RFID, *smart devices*, …
- **Datos de transacciones:** Facturación, Transferencias, TC, …
- **Datos de geolocalización:** + información espacial
- **Generados por los humanos:** Voz, emails, RME, documentos, …
- **Web y medios sociales:** Clicks, búsquedas, publicaciones, Web, …

Fuentes de datos en entornos Big data

1) ¿Qué es una fuentes de datos?
2) **Diferencias respecto a las tecnologías de datos tradicionales**
3) **Tipos de datos y flujo de datos**
4) **Calidad de Datos**
5) **Las V's del Big Data**

Diferencias respecto a las tecnologías de datos tradicionales

Big Data es **útil** para muchas empresas por **proporciona respuestas** a **preguntas** que las organizaciones **ni siquiera sabían que tenían**.

- Reducción de coste
- Más rápida y mejor toma de decisiones
- Nuevos productos y servicios

Diferencias respecto a las tecnologías de datos tradicionales

Tecnologías Tradicionales	Tecnologías Big Data
BD Relacionales	BD Relacionales + NoSQL
Consultas	Consultas, Captura, Procesamientos
Datos Internos	Datos Heterogéneos
Ámbito de la Informática	Todos los ámbitos

Diferencias respecto a las tecnologías de datos tradicionales

Fuentes de datos en entornos Big data

1) ¿Qué es una fuentes de datos?
2) Diferencias respecto a las tecnologías de datos tradicionales
3) **Tipos de datos y flujo de datos**
4) **Calidad de Datos**
5) **Las V's del Big Data**

Tipos de datos y flujos de datos

- Estructurados
- No Estructurados
- Semiestructurados

Estructurados

Son aquellos datos que tienen **tamaño finito** y **formato**.

Ejemplos: Las fechas, los números o las cadenas de caracteres. En esta categoría entran los que se compilan en los censos de población, los diferentes tipos de encuestas, los datos de transacciones bancarias, las compras en tiendas online, etc.

Estructurados

- **Creados:** datos generados por nuestros sistemas de una manera predefinida (registros en tablas, ficheros XML asociados a un esquema).

- **Provocados:** datos creados de manera indirecta a partir de una acción previa (valoraciones de restaurantes, películas, empresas, TripAdvisor, …).

- **Dirigido por transacciones:** datos que resultan al finalizar una acción previa de manera correcta (facturas autogeneradas al realizar una compra, recibo de un cajero automático, etc.).

Estructurados

• **Compilados:** resúmenes de datos de empresa, servicios públicos de interés grupal. Entre ellos nos encontramos con el censo electoral, vehículos matriculados, viviendas públicas, entre otros).

• **Experimentales:** datos generados como parte de pruebas o simulaciones que permitirán validar si existe una oportunidad de negocio.

Estructurados

Ejemplos:

- Declaración Renta, SRI, DIAN, RUC ...
- Expediente académico
- ...

No Estructurados

Aquellos datos que **carecen de un formato determinado**. No pueden ser almacenados en una tabla. Pueden ser capturados o generados.

Conviene saber que este tipo de datos **no tiene campos fijos** y normalmente se tiene poco control sobre ellos. Su manipulación requiere **herramientas especializadas**, como Hadoop (la más popular) y/o bases de datos NoSQL, entre otras.

No Estructurados

• **Capturados:** datos creados a partir del **comportamiento de un usuario**. Por ejemplo, información biométrica de pulseras de movimiento, aplicaciones de seguimiento de actividades (carrera, ciclismo, natación, etc.), posición GPS.

• **Generados por usuarios:** datos que especifica un usuario (publicaciones en redes sociales, videos reproducidos en *YouTube,* búsquedas en Google, entre otros).

No Estructurados

Ejemplos:

- Archivos de audio
- Vídeo
- Fotografías
- Formatos de texto
- SMS, Artículos
- Correos electrónicos
- Etc.

Semi estructurados

Poseen **organización interna o marcadores que facilita el tratamiento** de sus elementos. No pertenecen a bases de datos relacionales.

Ejemplos: XML, HTML o los datos almacenados en bases de datos NoSQL.

Tienen una cierta estructura, aunque sin llegar a estar totalmente estructurados.

Semi estructurados

Se puede considerar también: **multi-estructurados o híbridos**:

- datos de mercados emergentes
- e-commerce
- datos meteorológicos

Hasta acá la sesión 2, en la sesión 3 seguiremos

Gracias

Universidad Internacional de Valencia