UNIVERSIDAD COMPLUTENSE MADRID

Introducción a Big Data

Enrique Martín (emartinm@ucm.es)
Sistemas de Gestión de Datos y de la Información
Master Ing. Informática
Fac. Informática

Contenidos

Explicando Big Data

¿Qué es Big Data?

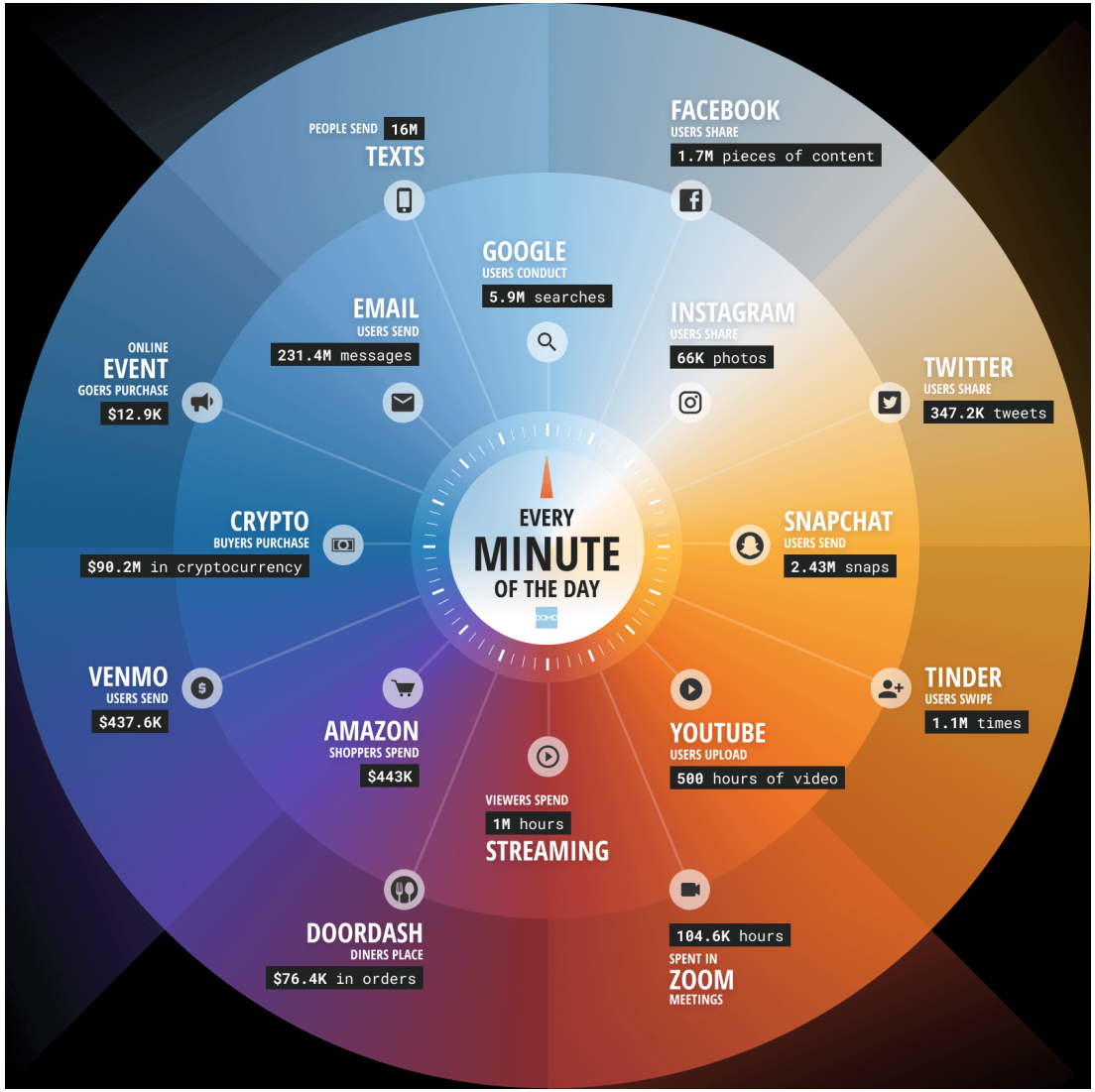
- **Big data** (grandes datos, grandes volúmenes de datos, o macrodatos) es la confluencia de distintas tendencias tecnológicas que venían madurando durante la década del 2000.
- Esto incluye tecnologías tanto para procesar (MapReduce, Hadoop, Storm, Spark...) como para almacenar (MongoDB, HBase, CouchDB, Cassandra...).

¿Cuánto es Big Data?

Depende del tamaño de la organización. En cualquier caso:

- Más cantidad de la que se puede almacenar y procesar con un **ordenador o servidor convencional**.
- **Crece a grandísima velocidad**. Hoy es una gran cantidad, pero cada día será más.

¿Cuánto es Big Data?



Fuente: `https://www.domo.com/data-never-sleeps`

¿Cómo es Big Data?

La información en Big Data proviene de fuentes diversas:

- Datos **estructurados**
- Datos **semiestructurados**
- Datos **no estructurados**

Datos estructurados

- Tienen una estructura fija y bien definida
- Suelen provenir de las bases de datos relacionales tradicionales que la organización utiliza en su negocio o sensores
- El **esquema** de la base de datos fuerza que todo lo que contenga tenga una determinada estructura fija, que no varía entre ejemplares

Datos semiestructurados

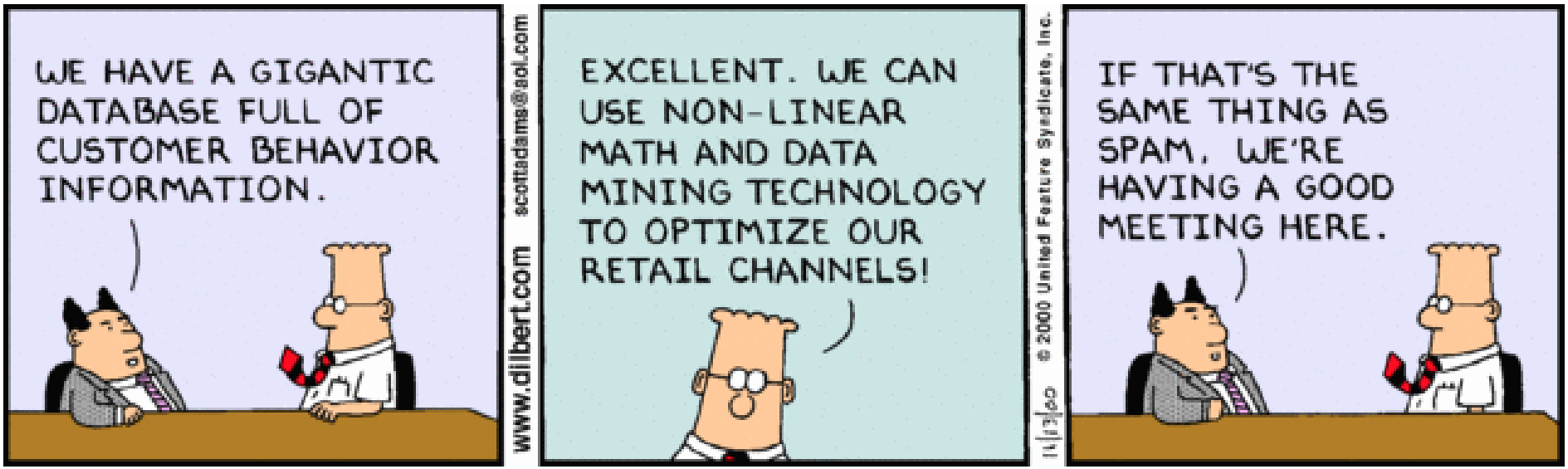
- Datos que no tienen un formato fijo, pero que tienen **etiquetas y otros marcadores** que permiten extraer la información con facilidad.
- Ejemplos: documentos XML, páginas web (HTML), ficheros de registro (logs)...

Datos no estructurados

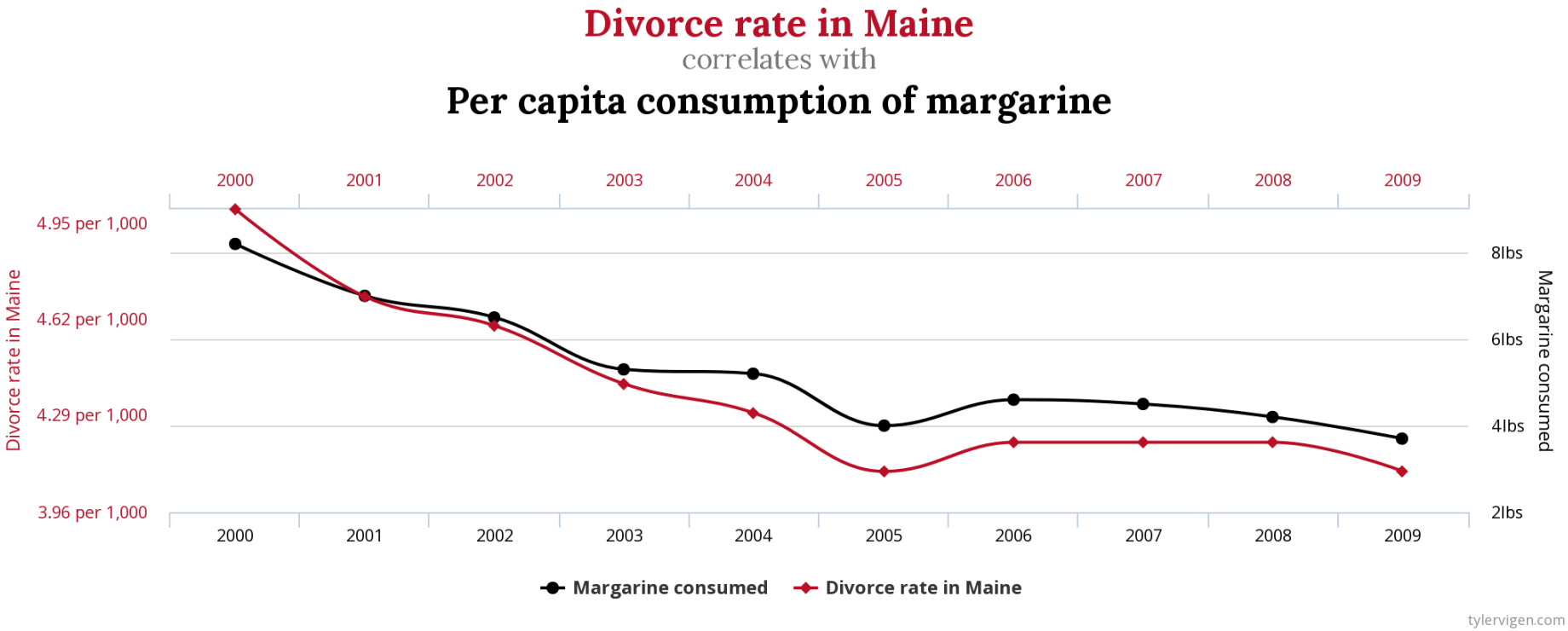
- No tienen tipo predefinido, ni estructura, ni etiquetas **que permitan extraer su información fácilmente**
- Ejemplos: ficheros multimedia (sonidos, vídeos o fotografías), mensajes de e-mail, SMS, mensajería instantánea (Whatsapp, Telegram)
- En algunos casos hay información que se puede obtener de manera sencilla a partir de sus **metadatos**: resolución de una imagen, fecha y lugar donde fue tomada, longitud de un vídeo
- En general, la información interesante es difícil de extraer: *¿Qué o quién aparece? ¿Qué hacen? ¿Qué ocurre?* Suele necesitar el uso de **inteligencia artificial**

¿Para qué Big Data?

- Pará qué puede ser útil el Big Data?
- ¿Qué provecho creéis que se puede sacar del análisis de mucha información?



¿Para qué Big Data?



Fuente: `http://www.tylervigen.com/spurious-correlations`

¿Para qué Big Data?

Tenemos acceso a una cantidad ingente de datos. ¿Cómo aprovecharlos?

- Principalmente para **entender mejor la realidad**:
  - La manera en la que una enfermedad infecciosa se transmite en la población
  - Las distintas maneras en las que se puede organizar espacialmente una proteína y sus implicaciones
  - en una tienda, los hombres de más de 50 años compran más un determinado producto cuando llueve
- Ese conocimiento de la realidad nos permite **ser más eficaces**:
  - Predecir contagios
  - Diseñar nuevos fármacos
  - Proporcionar ofertas y descuentos a determinados clientes para maximizar su consumo

Características de Big Data: Las 5 V

Características de Big Data

En Big Data se identifican principalmente 3 características, referidas como las 3 Vs:

1. Volumen
2. Velocidad
3. Variedad

Volumen

- Es la **principal característica de Big Data**, pero no la única
- Las organizaciones almacenan una cantidad muy grande de datos (TB, PB, etc.), pero este volumen sigue creciendo
- Es necesario tener tecnologías para **almacenar** y **analizar** estas cantidades de datos

Velocidad

- Esta característica se refiere a dos aspectos:
  - La velocidad con la que los datos son generados es muy alta. Puede llegar a varios PB diarios.
  - La gran cantidad de información nueva y acumulada hace que las organizaciones deban analizarla muy rápido, incluso llegando al tiempo real.
- Ejemplos de fuentes de datos con una gran velocidad de generación son los sensores (tráfico, clima, polución), los chips RFID, las cámaras, etc.

Variedad

Las fuentes de datos son de tipos muy diferentes:

- Texto: logs, e-mails, SMS
- Datos de sensores: clima, tráfico, polución, chips RFID
- Multimedia: webcams, cámaras de seguridad y vigilancia, fotos
- Flujos de clics en una sesión de navegación
- ***y muchos otros***

Más V de Big Data

Algunos autores añaden dos Vs más como características de Big Data:

- **Veracidad**: tener tanta cantidad de información de muchas fuentes tan diversas no inspira confianza. Habría que establecer la fiabilidad de los datos obtenidos.
- **Valor**: hay que poder obtener información de los datos de manera rentable y eficiente.

Resumen

- Big Data sirve para referirse a tecnologías utilizadas para obtener, almacenar y analizar cantidades inmensas de datos
- **Características principales**:
  1. Gran volumen de datos
  2. Hay que procesar rápidamente, ya que la cantidad de datos crece rápidamente
  3. Los datos se reúnen de distintas fuentes y con distintos formatos (con estructura o sin ella)