Data-driven: como las organizaciones utilizan nuestros datos para tomar decisiones

Pietro Marini

Distopías Cotiás
25 de Mayo de 2020

Introducción

Big Data: agrupaciones de datos cuyo almacenamiento, análisis y recopilación requieren el uso de nuevas técnicas

	Espacio ocupado	Ratio
Smartphone	10 GB	1
Disco Duro de PC Portátil	1 TB	100
Data Center del CERN [1]	300 PB	30 000 000
Datos almacenados en el mundo en 2025 [2]	160 ZB	10^{13}

- Ley de Moore: los procesadores ofrecen rendimientos siempre mejores.
- Grid Computing: distribución de la carga de computación en múltiples máquinas en paralelo.
- Software: desarrollo de estructuras de programación que facilitan la paralelización.
- Modelo de las “3V”:
  - **Volumen**
  - **Variedad:** textos, audio, imágenes, vídeos, datos non estructurados
  - **Velocidad:** aplicaciones en tiempo real

Ejemplo de Big Data

La empresa X fabrica, distribuye y vende smartphones en todo el mundo.

Esta empresa crea una herramienta de recepción y tratamiento de datos:

- Activación garantía: dirección IP, nombre, apellidos, email, número de teléfono, etc
- Transmisión posición de una aplicación de gestión del teléfono (contactos, batería, instalación otras aplicaciones, etc.). Cada minuto.
- Transmisión datos técnicos (procesador, memoria, numero de app instaladas/activas/desactivadas, etc.). Cada 10 minutos.

50 millones de smartphones vendidos al año, estimación de volumen anual 10 PB equivalente a ~ 100 000 PC portátiles

Un recurso fundamental: los datos

Aplicaciones:

- **Marketing, análisis de mercado y de clientes:** análisis de mercado, segmentación clientes, personalización canales/mensajes de marketing
- **Procesos de negocio:** gestión siniestros, gestión pagos, RRHH, finanzas, interacciones entre departamentos
- **Procesos industriales**: manutención, consumos energéticos, línea de producción, monitorización maquinarias
- **Servicios al ciudadano/cliente:** tramitaciones, gestión call center, quejas y feedback usuarios para mejora del producto/servicio

- ✔ elemento común: reconocimiento del rol fundamental de los datos

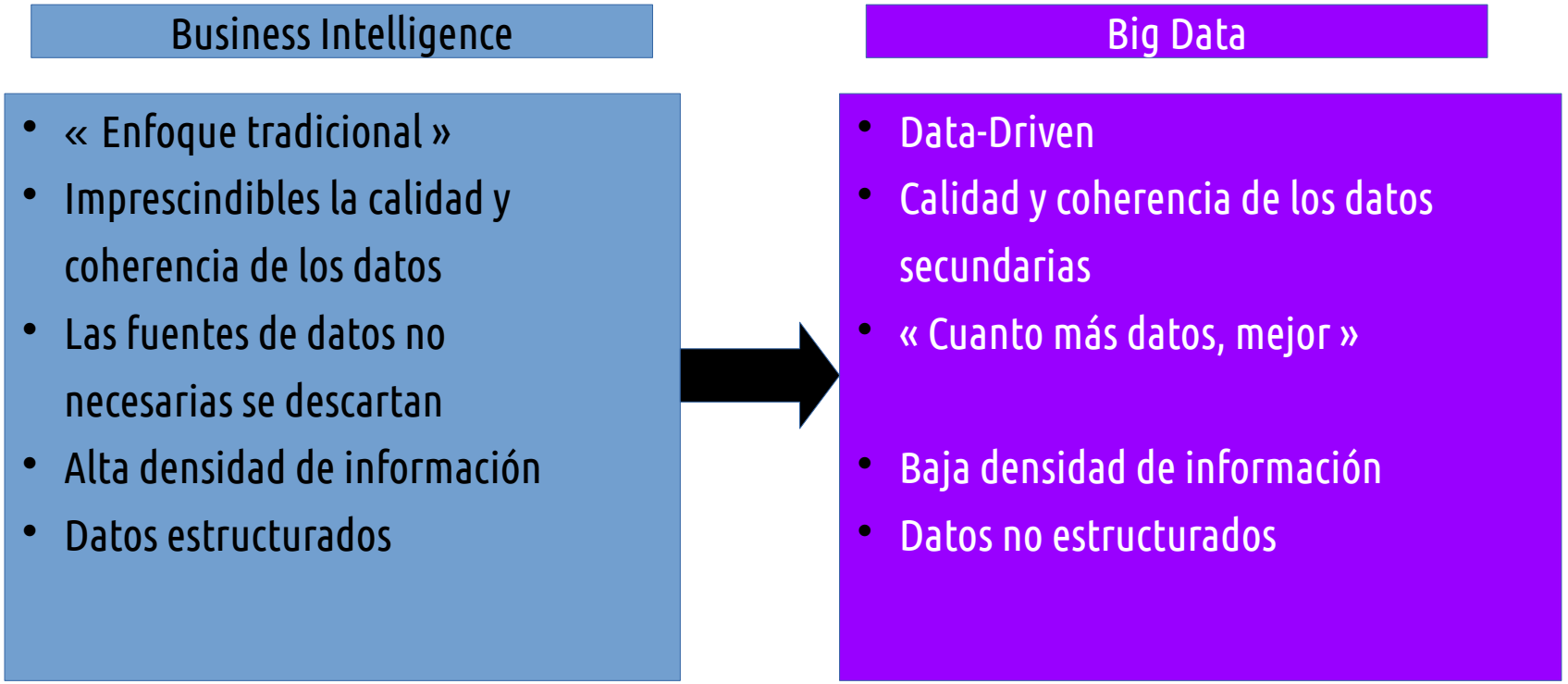
Business Intelligence

- BI (Inteligencia Empresarial): Conjunto de técnicas y procesos de soporte a la toma de decisiones
- Funcional (Business) y técnico (Tecnología de la Información)
- Problemáticas:
  - ¿Que fuentes de datos se necesitan para hacer un análisis?
  - ¿Como combinamos y ponemos en valor los datos a disposición para verificar una o más hipótesis de trabajo?
  - ¿Quien y como puede acceder a la información?
- Flujo de datos:

- Aplicaciones: Bases de datos relacionales, Excel, Aplicaciones Web, Paneles de Control
- Ejemplo: dirección comercial de una multinacional encarga un análisis de ventas de coches por diferentes variables: geográfica (España, Italia, Francia, EE. UU., etc.), temporal (último mes, último trimestre, último año), tipo de vehículo (turismo, familiar, furgonetas, …), canal de marketing (redes sociales, redes sociales, anuncios de televisión, cartel publicitario)

BI: ¿Predecesor del Big Data?

- Big Data como evolución de BI:
  - Aumento exponencial cantidad y variedad de datos => Enfoque diferente al consumo y al análisis
  - Bajo coste de almacenamiento y aumento rendimiento procesadores => Tendencia a guardar todas las fuentes e introducir nuevas => Enorme complejidad técnica y organizativa



Del Big Data a la Inteligencia Artificial

- Fase de realización de plataformas Big Data => Acumulación de datos
- Conexión con IA:

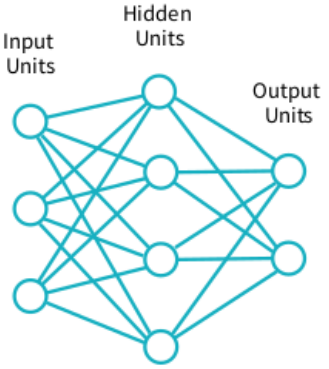
¿Podemos ***aprender*** de los datos?

↓

Sí, gracias al ***Aprendizaje Automático***

Confusión en los medios de comunicación: Inteligencia Artificial (IA) con Aprendizaje Automático (AA).

- AA:
  - [def. Arthur L. Samuel, US, 1901-1990] "rama de la IA que estudia como dar a los ordenadores la capacidad de aprender sin ser explícitamente programados"
  - A través de un proceso de *entrenamiento* y/u observación, buscar, para un *algoritmo*, los *parámetros* que mejor explican un fenómeno.
- IA: ¿podemos replicar la mente humana en un ordenador? Intuición, razonamiento, sentimientos, sentido común, imaginación ...



Aprendizaje Automático – Algunos Ejemplos

Familia de algoritmos	Aplicaciones	Tipo de datos
Sistemas de recomendación	Sugerencias en páginas web que ofrecen: • contenido (películas, noticias …) • servicios (vacaciones, trabajo …) • productos (libros, ropa …)	Datos de navegación web, hábitos de lectura, reseñas, comentarios ...
Visión Artificial	• Reconocimiento facial • Vehículo autónomo • Clasificación de galaxias • Diagnóstico de cáncer de pulmón	Imágenes fijas o en streaming
Análisis de sentimiento y text mining	• Predicción de cotizaciones de activos • Rendimiento del servicio de atención al cliente	Texto, libros, audio, comentarios y recensiones, prensa ...
Otros algoritmos de clasificación/regresión	• Identificación de la señal de colisiones entre partículas al CERN • Manutención predictiva • Probabilidad de clic en un web banner	Series temporales, datos técnicos, datos meteorológicos , texto, imágenes ...

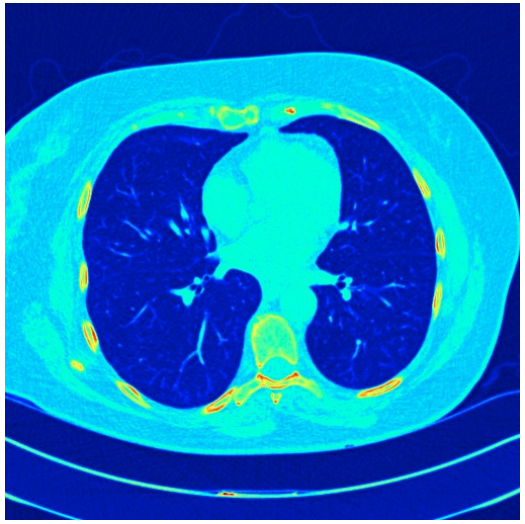
Aprendizaje Automático – ¿Como funciona?

¿Que banner voy a mostrar al Señor Alonso cuando se conecte a mí página web?

Definición del objetivo: el operador humano estudia e identifica el objetivo que el algoritmo tiene que lograr/ aproximar/modelar lo mejor posible	**Probabilidad de click en el banner**
Recogida de datos: generación, tratamiento y preparación de los datos en un formato adaptado para la aplicación del algoritmo	**Para cada visita, se recoge:** • **Entrada: datos de navegación y personales** • **Objetivo: si el usuario ha hecho click o no**
Entrenamiento: el algoritmo a través de un procedimiento iterativo busca los valores de los parámetros que dan la mejor función de aproximación al objetivo. Esta búsqueda se hace gracias a ejemplos donde conocemos la respuesta y juzgamos la calidad de sus respuestas sobre ejemplos que no le dejamos ver (generalización)	**La mejor publicidad para A. es la que obtuvo más éxito entre los usuarios más similares a él**
Predicción: una vez logrados resultados satisfactorios, el modelo restituye su predicción sobre datos que no ha visto en fase de entrenamiento	**Cuando A. se conecte, se recogen todos los valores de las variables utilizadas en entrenamiento; con ellos, el algoritmo restituye la mejor publicidad**

Aprendizaje Automático – ¿Como funciona? (2)

¿Que probabilidad tiene un paciente de desarrollar un cáncer de pulmón?



Definición del objetivo	Probabilidad de desarrollar un cáncer en el próximo año
Recogida de datos	Imágenes de tomografía, meta-datos (dispositivo utilizado, lugar, condiciones de la toma de imágenes …), historia clínica del paciente, otros datos (ej. incidencia de enfermedades similares en la población local)
Entrenamiento	Entrenamiento del algoritmo sobre un gran número de ejemplos positivos y negativos
Predicción	Para cada imagen nueva tomada en las mismas condiciones que las de entrenamiento, este nos restituye la probabilidad de desarrollo de una enfermedad

Aprendizaje Automático – Algunas Observaciones

- Teoría y aplicación práctica coinciden: más datos => mejor generalización => justificación de Big Data => “Data is the new oil”

- Complejidad e implicaciones en la definición del objetivo del algoritmo:
  - La toma de decisión depende de lo que “dice” el algoritmo
  - Pequeñas variaciones del objetivo pueden llevar a predicciones muy diferentes
    - Ej. cáncer de pulmón: ¿tres clases de riesgo en vez de dos? ¿Probabilidad a 6, 12 o 24 meses?
    - “Catastrophic Forgetting”

- Los algoritmos trabajan con variables sin ningún sentido ni conocimiento del contexto:
  - Encuentran correlación, no causalidad
  - Sus previsiones son el resultado de una optimización matemática, ¿lo podemos acercar al razonamiento/inteligencia?

- Problemática de la interpretación:
  - Definir y entender las predicciones de un algoritmo de AA
  - Capacidad de interpretación disminuye al crecer la capacidad de generalización

Muchas cuestiones abiertas

- Papel central de los datos y de las tecnologías de los datos en la sociedad moderna:
  - Reglamentación y privacidad: RGPD (Reglamento General de Protección de Datos)
    - ¿Modelo de negocio predominante de servicios en Internet tendría que replantearse?
  - Impacto en el ámbito científico: “muerte de la teoría”?
  - Digitalización:
    - Cada gesto en nuestro día a día está digitalizado => somos creadores de datos, pero consumidores a través del filtro de las empresas => derecho y deber sobre su uso

- Algoritmos de AA cada vez más autónomos:
  - ¿Aumento de nuestras facultades? ¿Peso que tenemos que dar a las decisiones algorítmicas?
  - ¿Simplificación de nuestras vidas vs vigilancia?
  - ¿Como definimos la responsabilidad para sistemas autónomos de decisión?
  - Problema de los prejuicios (ingles, *bias*)?

Graciñas!

Fuentes y Recursos

- [1]: Información sobre el centro de datos del CERN
- [2]: https://en.wikipedia.org/wiki/Big_data
- MIT Tech Review: AI could help with the next pandemic—but not with this one
- Nubes de palabras - Titulo
- The Institute for Ethical AI & Machine Learning
- MIT Tech Review: Nearly half of Twitter accounts pushing to reopen America may be bots