

Seminario

Introducción a Big Data, Data Science e Inteligencia Artificial

Octubre de 2020

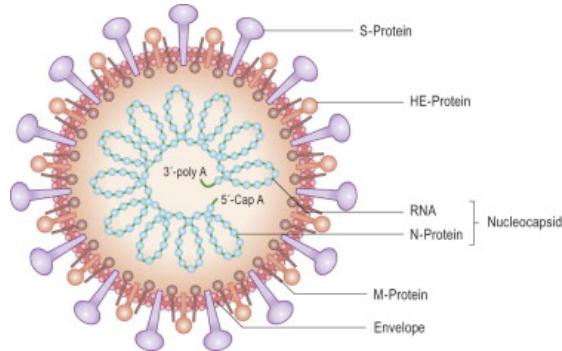
Dr. Jaime Navón
Departamento de Ciencia de la Computación
Escuela de Ingeniería de la UC

La Data es el nuevo petróleo

- ▶ Cambios profundos en la forma en que se produce, usa y gestionan los datos
- ▶ Se producen a través de cada interacción, conversación o proceso dentro o fuera de la empresa
- ▶ Redes sociales + *smartphones* + *IoT* se tiene acceso a un río de información permanente
- ▶ Este río de información incluye información no estructurada
- ▶ Se requieren técnicas para gestionarla: Big Data (volumen, velocidad, variedad)
- ▶ Este río de información sumado a las tecnologías de minería de datos y *machine learning* permite ser analizado y utilizado en la toma de decisiones
- ▶ Analítica compleja requiere integrar e interconectar los silos de datos

COVID-19: Estudio de big data permitiría pronosticar el peak del coronavirus en Chile

Cuando de pandemia se trata, la tecnología es capaz de anticiparnos que lo peor aún no ha llegado. La buena noticia es que con el manejo de datos es posible predecirlo y prepararse. Así lo señala este análisis hecho en nuestro país.



The power of data in a pandemic

The NHS is facing an unprecedented challenge. Responding to the Covid-19 crisis will require everything we have and more. In the fight against this pandemic, decision-makers will need accurate real-time information. To understand and anticipate demand on health and care services, we need a robust operating picture of the virus, how it's spreading, where it might spread next and how that will affect the NHS and social care services. On the supply side, we need to know where the system is likely to face strain first, be that on ventilators, beds or staff sickness.

CULTURA | CIENCIA

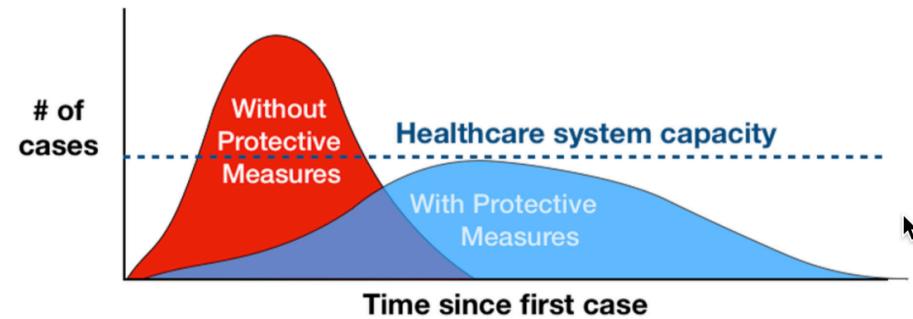
Cultura

Investigadores participan en mesa de datos sobre coronavirus en Chile liderada por el Ministerio de Ciencia

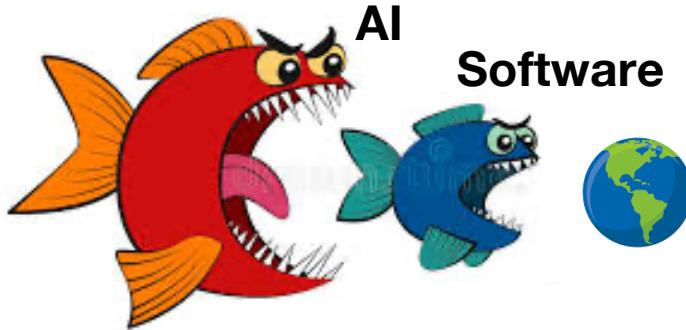
Académicos trabajan en "mesa de datos COVID-19" liderada por el Ministerio de Ciencia

Understanding the COVID-19 Pandemic as a Big Data Analytics Issue

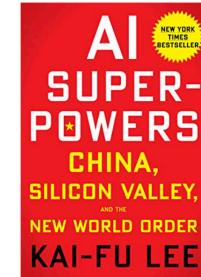
Big data analytics techniques are well-suited for tracking and controlling the spread of COVID-19 around the world.



"AI is eating the world"



- ▶ Lucha por la supremacía mundial entre USA y China
 - ▶ *If data is the new oil, then China is the new Saudi Arabia* Kai-Fu Lee
-
- ▶ Training deep learning models requires more of a brute force method than innovation; well suited to China's supposedly higher quantity but lower quality of software engineer compared to the US
 - ▶ China has fewer data protection regulations than other countries, so Chinese software collect more data on users.
 - ▶ Chinese tech startup culture is more "aggressive" than that of other countries', with fewer intellectual property restrictions and fewer barriers to vertical integration
 - ▶ The participation of China's central government in funding and raising the status of the AI industry





- ▶ Efectos enormes de la electricidad en la era de la revolución industrial
 - ▶ eliminó restricciones de ubicación de las máquinas
 - ▶ eliminó la necesidad de escala grande para las fábricas
 - ▶ permitió uso de motores eléctricos livianos y baratos
 - ▶ se creó un nuevo ecosistema de industrias, ingenieros, trabajadores, productos y negocios
- ▶ La IA es la nueva electricidad y Amazon, Google, Microsoft son las nuevas empresas eléctricas
 - ▶ Muchas de las restricciones de la era pre-digital desaparecieron
 - ▶ Es una tecnología de tipo general
 - ▶ Hasta ahora pocos han sabido sacar partido de ello
 - ▶ Quienes lo han hecho adquieren una enorme ventaja competitiva

Existe mucha confusión

- ▶ Informe Big Data ?
- ▶ Es AI ?
- ▶ Es Data Science ?
- ▶ Es Data Mining ?
- ▶ Es Business Intelligence ?
- ▶ Es Advanced Analytics ?

NACIONAL Crisis social

Informe "Big Data" fue elaborado por la empresa española Alto Data Analytics

Leslie Ayala y Juan Manuel Ojeda
30 DIC 2019 08:21 PM

A photograph showing a person standing in a debris-strewn area, looking at a large building engulfed in intense orange and yellow flames. Thick black smoke billows from the burning structure, partially obscuring the upper floors. In the background, other buildings and a white van are visible under a clear sky.

NACIONAL

Crisis social

Informe "Big Data" fue elaborado por la empresa española Alto Data Analytics

Leslie Ayala y Juan Manuel Ojeda

30 DIC 2019 08:21 PM



Trataremos de Precisar los términos mas usados

- ▶ Big Data
- ▶ Inteligencia Artificial
- ▶ Ciencia de Datos
- ▶ Machine Learning
- ▶ Deep Learning
- ▶ Analytics
- ▶ Business Intelligence
- ▶ Data Mining
- ▶ Analytics

Tecnologías claves relacionadas con Datos

- ▶ Databases, Data Warehouses y Datalakes
Repositorios donde los datos son almacenados y gestionados
- ▶ Big Data
Técnicas para manejar volúmenes gigantescos de datos estructurados y no estructurados que cambian rápidamente
- ▶ Minería de Datos
Encontrar patrones interesantes o de gran valor para el negocio a partir del análisis de cerros de información
- ▶ Machine Learning (Aprendizaje de Máquina)
Usar los datos (big data) para entrenar a un algoritmo a realizar una tarea que podría requerir habilidades cognitivas (detección de un cancer, chatbot inteligente, etc)
- ▶ Deep Learning
Una forma específica de lo anterior que usa varias capas de redes neuronales y que ha demostrado ser tremadamente poderosa
- ▶ Inteligencia de Negocios o Analítica de Negocios
Utilización de técnicas como las anteriores para el apoyo de las decisiones de negocios de alto nivel

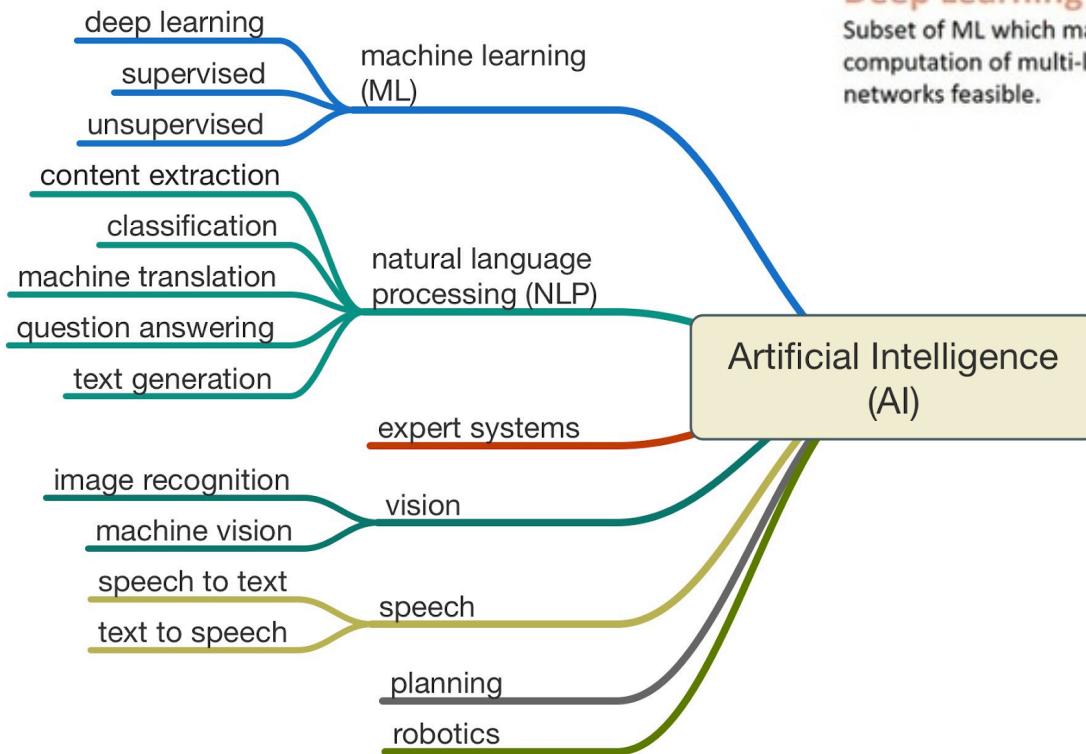
Big Data vs Data Science

- ▶ El foco de Data Science está en el análisis de la data
- ▶ Un *data scientist* debe conocer las técnicas para aprender desde los datos (data mining, data visualization, programming)
- ▶ Se hace mucho mas interesante con la llegada de big data
- ▶ El ser capaz de obtener, limpiar, almacenar y procesar la big data es el dominio de los *data engineers*
- ▶ Muchas veces ambas funciones son realizadas por las mismas personas

AI, ML, DL

- ▶ AI - Artificial Intelligence
 - ▶ lograr que una máquina sea capaz de realizar funciones cognitivas típicas de los humanos
- ▶ ML - Machine Learning
 - ▶ subconjunto de la AI que logra hacer que una máquina sea capaz de "aprender" a partir de un conjunto de datos (entrenamiento)
 - ▶ uno de los modelos para lograr esto son las redes neuronales
- ▶ DL - Deep Learning
 - ▶ subconjunto de ML basada en redes neuronales
 - ▶ técnica ha logrado resultados espectaculares los últimos años
 - ▶ requiere de big data

Gráficamente



Artificial Intelligence

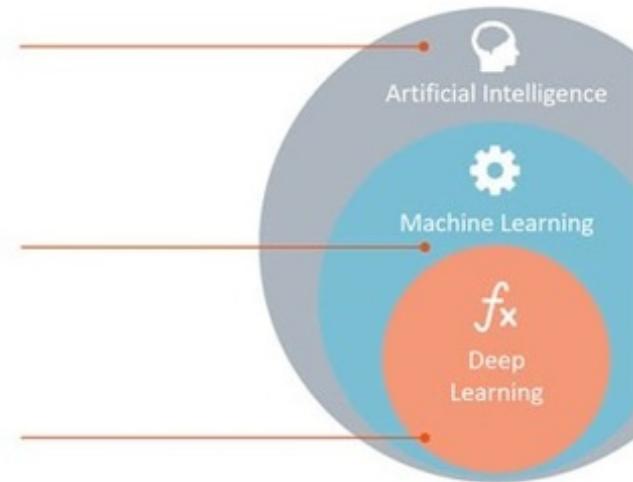
Any technique which enables computers to mimic human behavior.

Machine Learning

Subset of AI techniques which use statistical methods to enable machines to improve with experiences.

Deep Learning

Subset of ML which make the computation of multi-layer neural networks feasible.





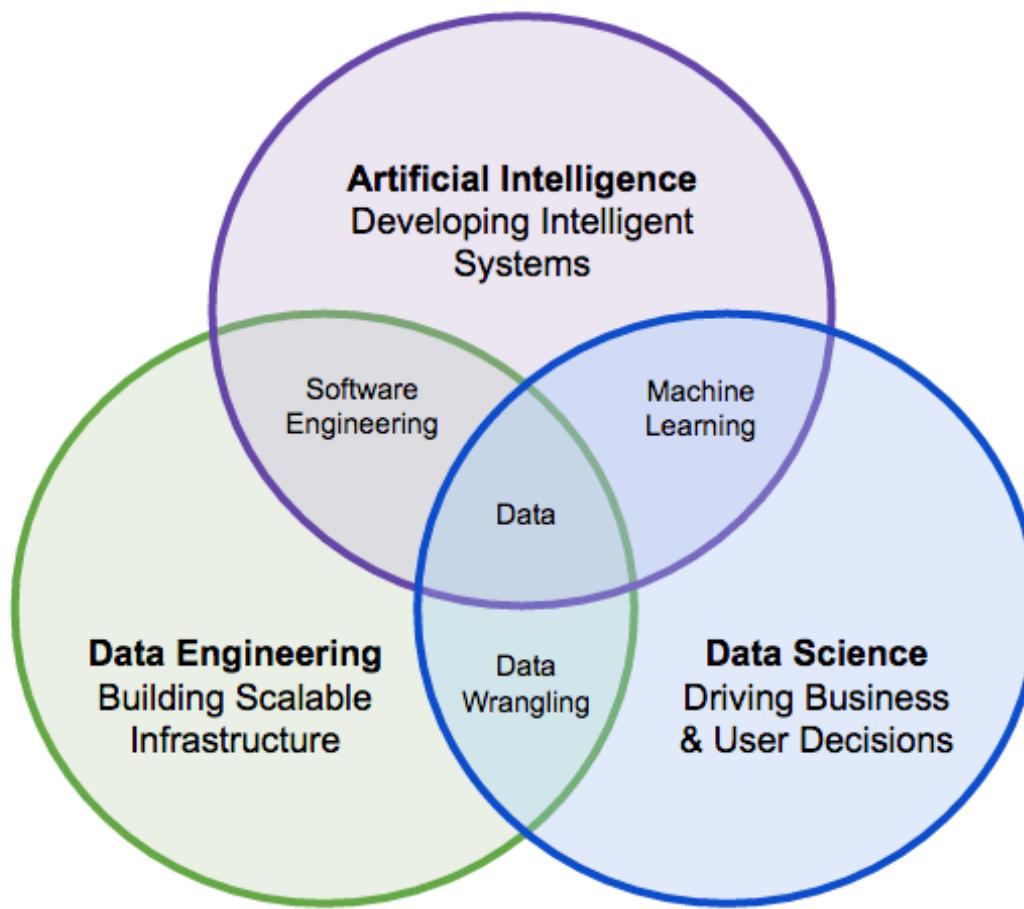
Ejemplos de AI

- ▶ AI: Un auto que se conduce en forma autónoma emulando una tarea que pareciera requerir habilidades cognitivas elevadas
- ▶ AI: Un robot capaz de interactuar con humanos incluso con expresiones faciales
- ▶ AI: Un chatbot inteligente que pueda actuar en forma eficiente en una mesa de ayuda
- ▶ AI: Asistentes capaces de entender lenguaje natural (Alexa)

Ejemplos de ML

- ▶ Alphago
- ▶ Sistemas recomendadores
- ▶ identificación de imágenes
- ▶ Prevención de fraudes
- ▶ Prevención de ataques cibernéticos
- ▶ Detectores de plagio en ensayos
- ▶ Predicción de ETA





Big Data

- ▶ Data que excede la capacidad de proceso de sistemas de BD convencionales
- ▶ Las 3 V's: volume, velocity, variability
 - ▶ volumen gigante
 - ▶ se mueve muy rápido
 - ▶ no calza en estructuras usuales
- ▶ Solo recientemente hay técnicas y software capaces de sacar partido de esto

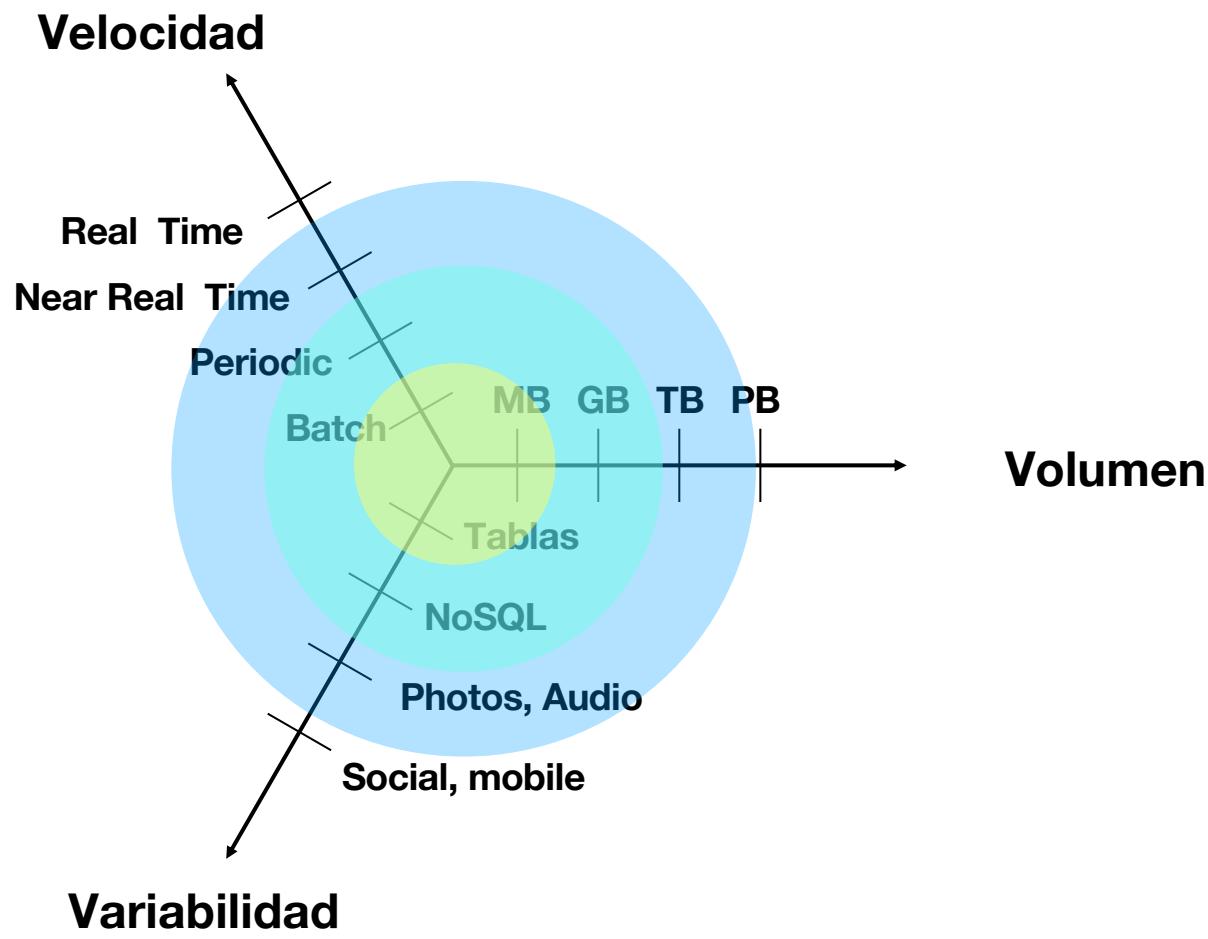


Volumen



Volumen + Velocidad

El mundo de Big Data



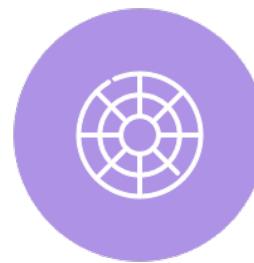
Algunos agregan otras V's



Volume



Velocity



Variety



Veracity



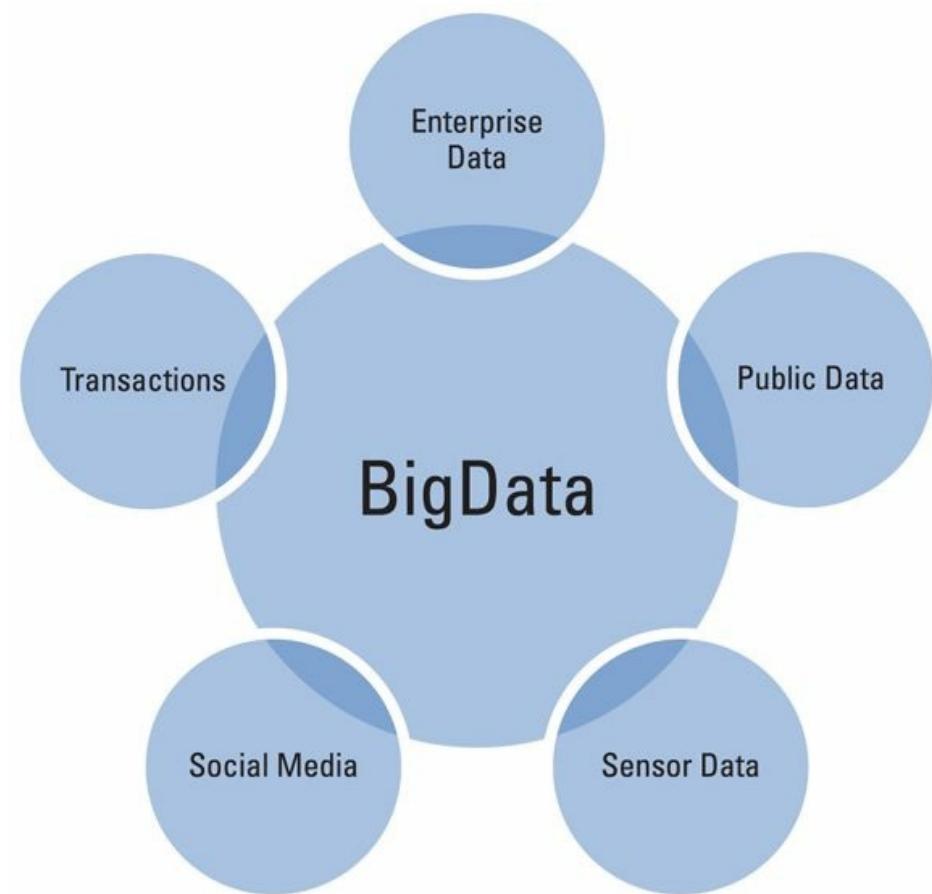
Value



Variability

Big Data: Fuentes de Datos

- ▶ conversaciones de redes sociales
- ▶ registros de acceso a servidores Web
- ▶ sensores de flujo de tránsito
- ▶ IoT
- ▶ imágenes satelitales
- ▶ broadcast de streams de audio
- ▶ scans de documentos de gobierno
- ▶ huellas de GPS
- ▶ resultados bursátiles mundiales



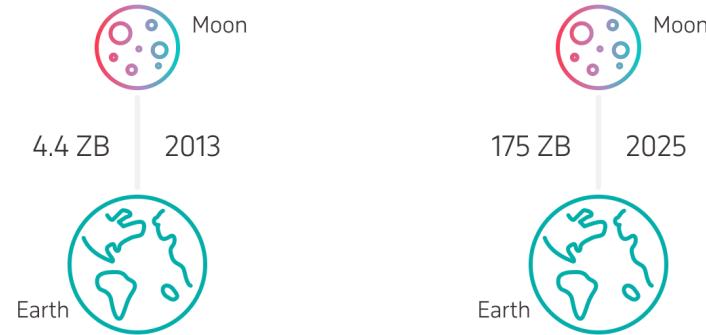
Volumen

- ▶ Capacidad de procesar volúmenes gigantescos de datos es el principal atractivo de Big Data
- ▶ Típicamente 1 TB hacia arriba
- ▶ Mas datos es a menudo mejor que tener mejores modelos
- ▶ Obliga a arquitectura escalable y esquema de consulta distribuido
- ▶ Data Warehousing comparte este requisito pero no se requiere la variabilidad
- ▶ Uso de Plataforma Hadoop o similar (distribución de la pega)

Creciendo rápidamente

The exponential growth of data

iPad Air—0.29" thick, 128GB



If the Digital Universe were represented by the memory in a stack of tablets, in 2013 it would have stretched two-thirds the way to the Moon.

By 2025, there would be 26.25 stacks from the Earth to the Moon.

Data source: seagate.com—The digitization of the world from edge to core, 2018

- ▶ 2013 – 4.4 zettabytes (1 zettabyte = 10^{21} bytes)
- ▶ 2025 - 175 zettabytes Equivale a 1.367.720.000.000 iPads de 128 GB
- ▶ Un iPad Air de 128 GB tiene un espesor de 0.29"
- ▶ Para completar los 175 zettabytes necesitaríamos 1.367.720.000.000 iPads
- ▶ Colocados uno sobre otro sería 26 veces la distancia de la tierra a la luna



Más números

- ▶ Este año habrá alrededor de 40 trillones de gigabytes de data (40 zettabytes o 40×10^{21} bytes)
- ▶ El 90% se creó en los últimos dos años
- ▶ Cada persona genera del orden de 1.7MB por segundo !
- ▶ Se generan aprox 2.5 exabytes de datos cada día (1 millón de terabytes o 1000 millones de gigas) en emails, websites y redes sociales
- ▶ La mayor parte es información no estructurada (texto, imágenes, videos)
- ▶ Usuarios de twitter generan más de medio millón de tweets cada minuto

Velocidad

- ▶ Cada vez los datos fluyen más rápido al interior de las organizaciones
 - ▶ actividad de los clientes minuto a minuto es canalizada directamente
 - ▶ aumento del uso de smartphones contribuye a esta tendencia (fuente permanente de audio e imágenes georeferenciadas)
 - ▶ se requiere reaccionar rápido (ejemplo de cruzar la calle de IBM)
- ▶ Hasta 30 GB/sec, o miles de mensajes por segundo
- ▶ Considerar procesamiento directo del stream (IBM InfoSphere Streams, Twitter Storm)

Variabilidad

- ▶ Fuente de datos muy diversa
- ▶ No calza en estructuras clásicas (esquemas relacionales)
- ▶ texto, imágenes, feed de sensores, etc
- ▶ BD orientadas a documentos (JSON, XML) o grafos (Neo4J) suelen funcionar mejor
- ▶ Principal problema de BD relacionales es la naturaleza estática de los esquemas
- ▶ Uso de Data Lakes en lugar de Data Warehouses

Herramientas

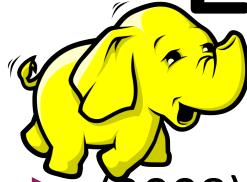


- ▶ Grandes ecosistemas o plataformas
 - ▶ Apache Hadoop
 - ▶ Apache Spark

Procesamiento Distribuido

- ▶ Aumento enorme de capacidades de almacenamiento no ha sido acompañado por aumento similar en velocidades de lectura/escritura
 - ▶ Leer un drive completo en 1990 (1Gb) – 5 minutos
 - ▶ Leer un drive completo hoy (1Tb = 1000 Gb) – 2.5 horas
- ▶ Desafíos de cómputo distribuido
 - ▶ tolerancia a fallas - HDFS
 - ▶ combinar resultados parciales - MapReduce

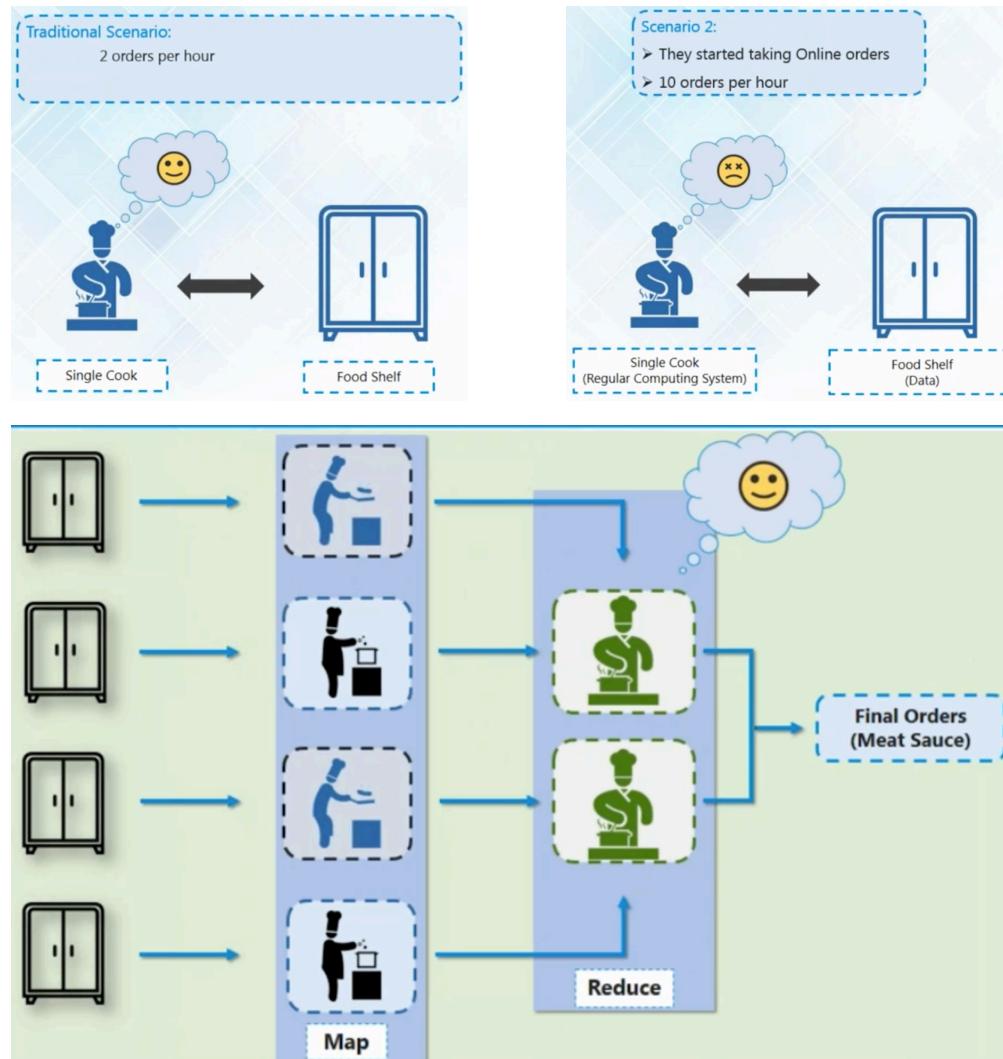
El Ecosistema Hadoop



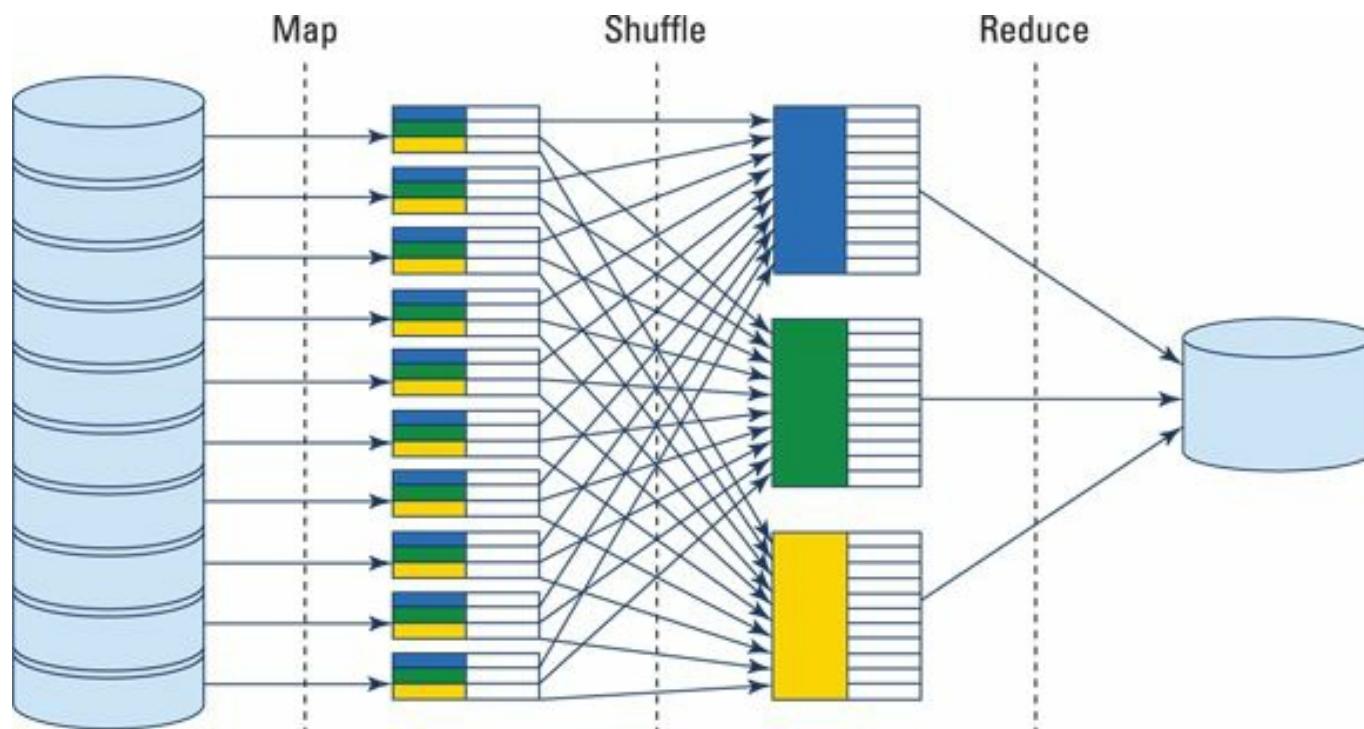
► (2008) Asociado en forma casi simbiótica con Big Data

- ▶ Responde adecuadamente a los dos problemas anteriores
 - ▶ aumento de la probabilidad de falla – HDFS un sistema de archivos distribuido
 - ▶ debemos combinar los resultados individuales – algoritmo map-reduce
- ▶ MapReduce permite tomar una consulta sobre un dataset, dividirla y correrla sobre múltiples nodos en forma simultánea para luego combinar los resultados en una solución final
- ▶ Hoy en día Hadoop representa todo un ecosistema de herramientas que permiten acceder a esta base con mayor facilidad

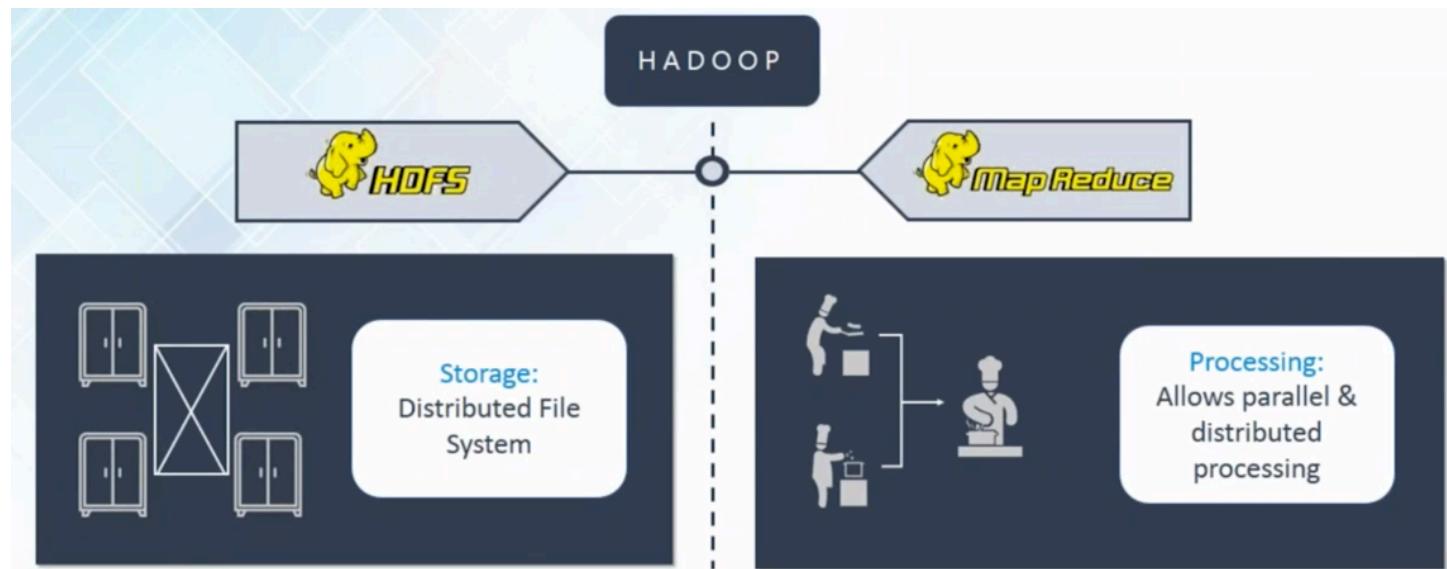
Escalabilidad gracias al procesamiento distribuido



Arquitectura MapReduce

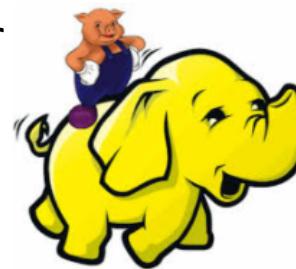


Las dos grandes componentes de Hadoop



Herramientas de mas alto nivel sobre Hadoop

- ▶ Aunque Hadoop puede hacer la pega, preparar las tareas puede no ser demasiado sencillo
- ▶ Se han construido herramientas que permiten no tener que pensar en términos de map-reduce aunque por debajo funcione de esa forma
- ▶ Dos de estas herramientas son Pig y Hive
 - ▶ Pig tiene un enfoque de scripting (similar a escribir un shell script)
 - ▶ Hive tiene un enfoque SQL
 - ▶ Ambos corren sobre Hadoop y por lo tanto requieren la JVM





- ▶ Proyecto Apache originado en UC Berkeley
- ▶ Ecosistema open source que disputa supremacía de Hadoop para procesamiento de Big Data
- ▶ Problemas cuando hay dos tareas de map reduce encadenadas (lectura-escritura innecesaria)
- ▶ Necesidad de procesar en tiempo real
- ▶ Puede correr de 10 a 100 veces mas rápido que Hadoop map reduce
- ▶ Mucho trabajo en memoria

Big Data por sí sola no sirve de nada

- ▶ estos ríos de datos gigantescos poco estructurados resultan demasiado complejos imposibles de entender
- ▶ al procesar estos datos con la ayuda de plataformas analíticas las organizaciones puede producir información precisa y "accionable"
- ▶ actionable significa el poder tomar una acción concreta a partir de esa información
- ▶ se pueden tomar mejores decisiones de negocio y mejorar las operaciones
- ▶ se puede obtener una ventaja competitiva importante

Data Science

- ▶ Aún siendo capaces de manejar la Big Data incluyendo el manejo de herramientas asociadas ello no garantiza el poder sacar partido de ella
- ▶ La palabra **Science** en Data Science es importante !
- ▶ En Ciencia la clave es ser capaces de formular buenas preguntas
- ▶ La data servirá para responder a esas preguntas y así aprender algo nuevo que puede ser usado en la organización

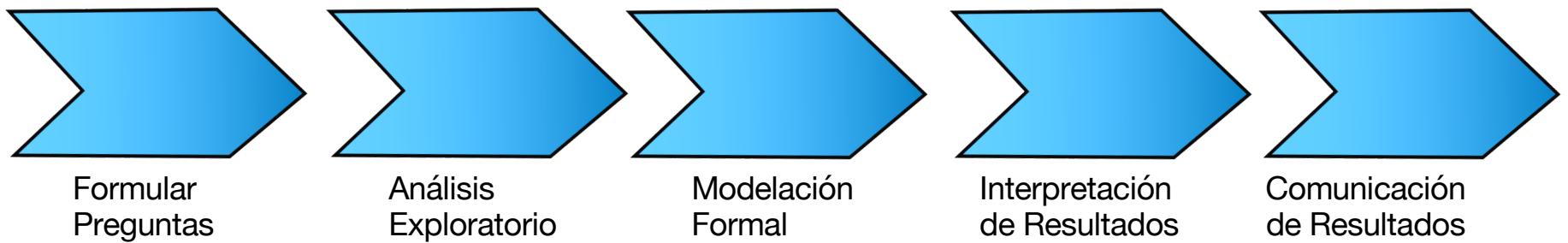
La importancia de partir con las preguntas y no con los datos

- ▶ Podemos tener Gigas de data pero solo una fracción es relevante para responder
- ▶ Puede detectarse la necesidad de recolectar otros datos o más datos
- ▶ Puede evitarse el perder tiempo con correlaciones extraídas ciegamente de los datos

El proceso de Data Science

- ▶ formular preguntas cuantitativas
- ▶ identificar la data que puede ser usada para responder
- ▶ limpiar y organizar la data
- ▶ analizar la data
 - ▶ visualización
 - ▶ análisis estadístico
 - ▶ machine learning
- ▶ comunicar los resultados a otras personas

Proyecto de Data Science



La importancia de la visualización

- ▶ Importante tanto en el proceso de un proyecto para llegar a los "insights" como también para comunicar a las personas
- ▶ A medida que hay más y mas datos disponibles se hace más difícil entenderlos y comunicarlos
- ▶ Los seres humanos somos muy buenos en entender información gráfica
- ▶ A veces incluso estamos bastante "ciegos" ante relativamente poca información

Cuatro sets de datos aparentemente parecidos

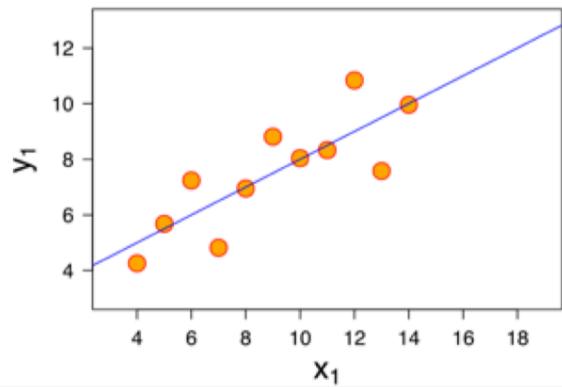
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Los cuatro sets tienen
todos estos parámetros iguales

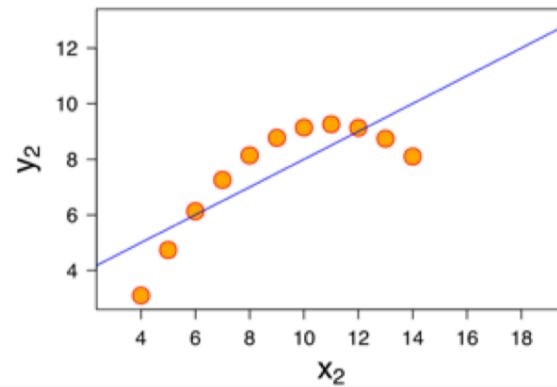
Property	Value
Mean of x	9
Sample variance of x	11
Mean of y	7.50
Sample variance of y	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination of the linear regression	0.67

La visualización deja en evidencia las diferencias

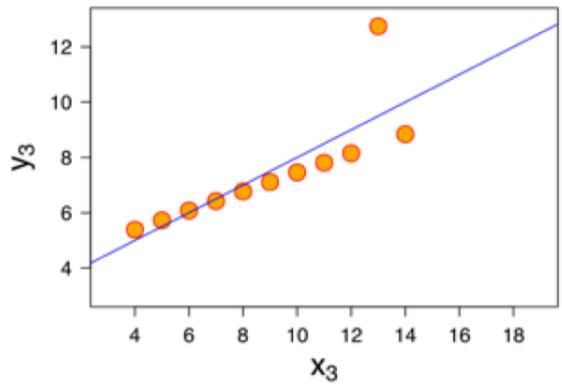
I



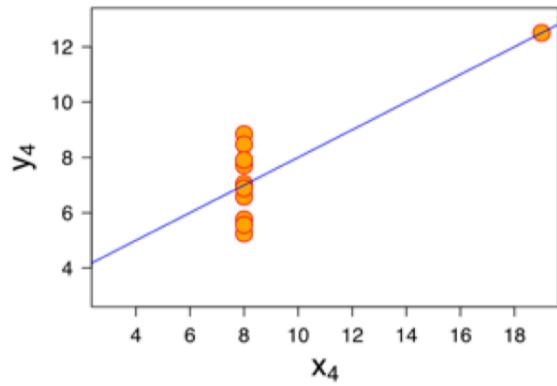
II



III



IV



Surgimiento del Data Scientist

- ▶ Combinación de competencias
 - ▶ científicas (capaz de formular preguntas interesantes)
 - ▶ matemáticas (estadísticas, álgebra lineal)
 - ▶ computacionales (programación, bases de datos, herramientas)
 - ▶ comunicacionales (capaces de crear narrativas de su trabajo)

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.



MATH & STATISTICS

- ★ Machine learning
- ★ Statistical modeling
- ★ Experiment design
- ★ Bayesian inference
- ★ Supervised learning: decision trees, random forests, logistic regression
- ★ Unsupervised learning: clustering, dimensionality reduction
- ★ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ★ Computer science fundamentals
- ★ Scripting language e.g. Python
- ★ Statistical computing package e.g. R
- ★ Databases SQL and NoSQL
- ★ Relational algebra
- ★ Parallel databases and parallel query processing
- ★ MapReduce concepts
- ★ Hadoop and Hive/Pig
- ★ Custom reducers
- ★ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ★ Passionate about the business
- ★ Curious about data
- ★ Influence without authority
- ★ Hacker mindset
- ★ Problem solver
- ★ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data-driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any of visualization

Herramientas de un Data Scientist

- ▶ Herramientas para almacenar los datos (bases de datos de distintos tipos, hojas de cálculo, etc)
- ▶ Herramientas para "masajear" los datos (limpiar, ajustar, normalizar, etc)
- ▶ Herramientas para analizar los datos incluyendo soporte para la visualización

Y qué es "analytics"

- ▶ alcance algo mas limitado que el de data science
- ▶ foco en encontrar respuestas a preguntas que conocemos
- ▶ uso de datasets disponibles
- ▶ resultados deben conducir a mejoras inmediatas

	Data Science	Analytics
Scope	Macro	Micro
Objetivo	Buenas Preguntas	Datos Valiosos
Requiere Big Data	Si	Si
Plazo	Mediano y Largo	Corto

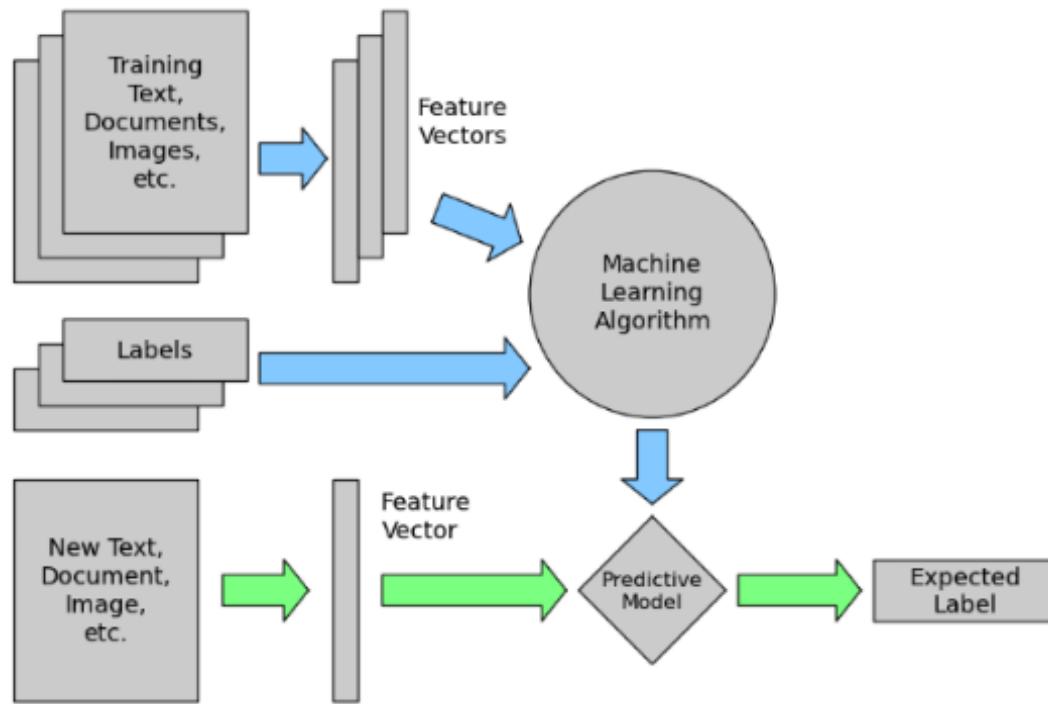
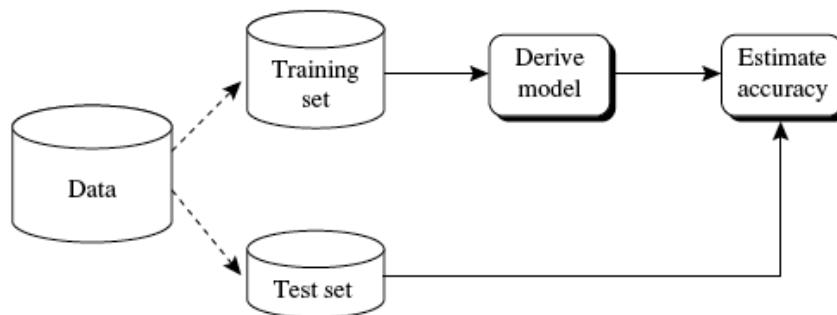
y Business Intelligence (BI) ?

- ▶ Hay mucho en común con Data Science
 - ▶ Usar los datos para lograr objetivos de negocio
- ▶ Principales diferencias
 - ▶ principalmente datasets internos (CRMs, Ventas, Desempeño)
 - ▶ tamaño mediano a grande, datos estructurados
 - ▶ modelos de datos multidimensionales (cubos)
 - ▶ uso de data warehouses
- ▶ Ambas cosas se han ido acercando

Machine Learning (ML)

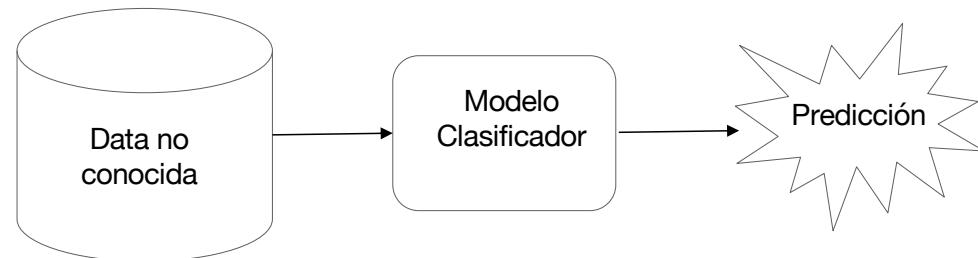
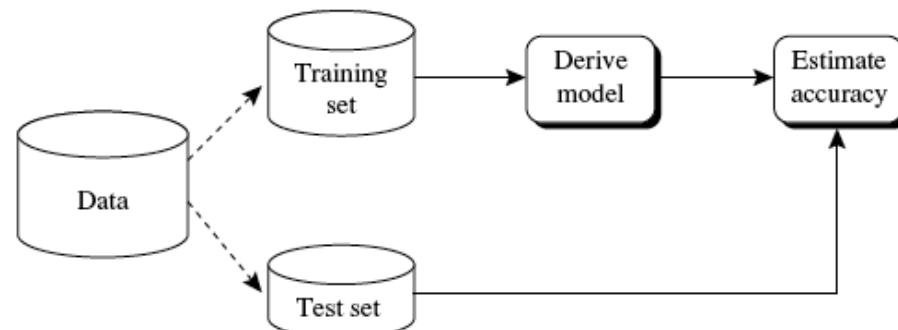
- ▶ Aprendizaje algorítmico a través de los datos
- ▶ El objetivo principal es poder hacer predicciones
- ▶ Dos grandes tipos de aprendizaje
 - ▶ aprendizaje no supervisado - algoritmos intentan descubrir patrones en los datos (por ejemplo clusters)
 - ▶ aprendizaje supervisado - colección de elementos predictores y resultados observados son usados para entrenar a un sistema de modo que pueda predecir un resultado que no ha sido observado

Aprendizaje Supervisado



Clasificación mediante Aprendizaje Supervisado

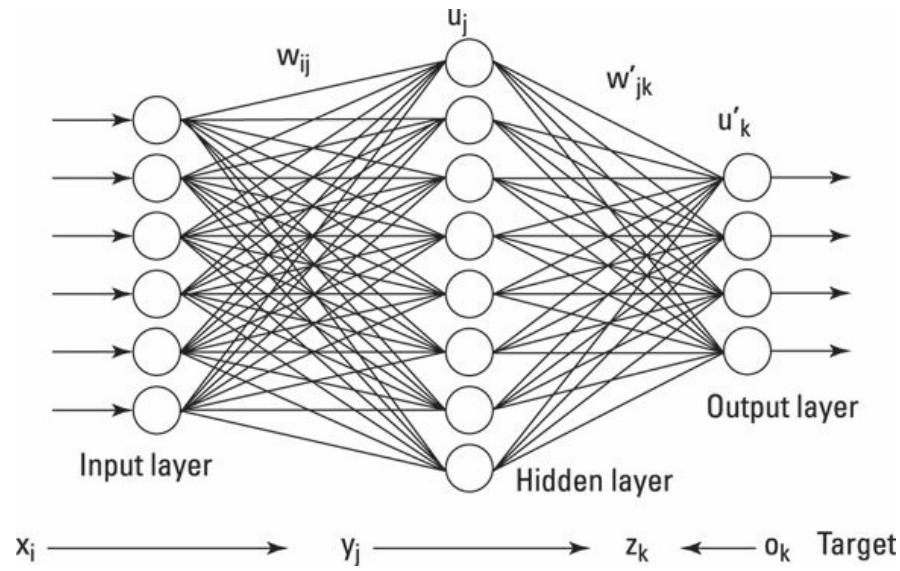
- ▶ Tres etapas
 - ▶ Entrenamiento
 - ▶ Test
 - ▶ Clasificación



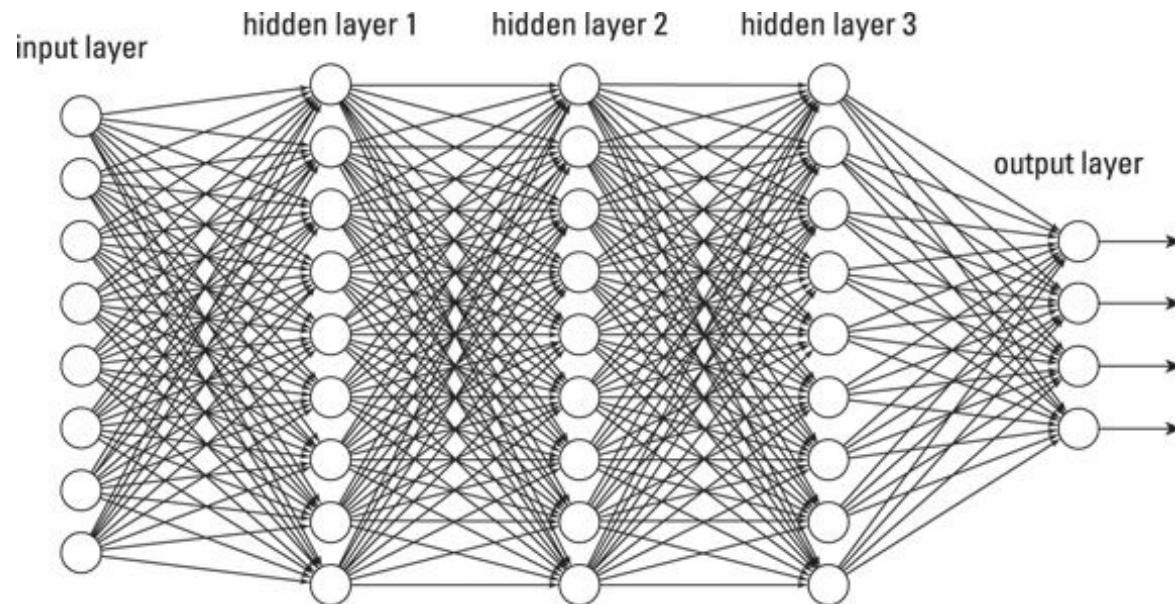
y Deep Learning ?

- ▶ Dentro de los algoritmos usados en ML para técnica de aprendizaje supervisado hay uno que ha cobrado especial importancia: las redes neuronales
- ▶ Una red neuronal permite generar un modelo ajustando permanentemente el peso de las conexiones entre las capas de células durante el proceso de entrenamiento
- ▶ Los algoritmos de deep learning llevan esta idea mas allá al incorporar numerosas capas de celdas entre la entrada y la salida
- ▶ Ello permite mejorar dramáticamente el desempeño de estos modelos

Deep Learning



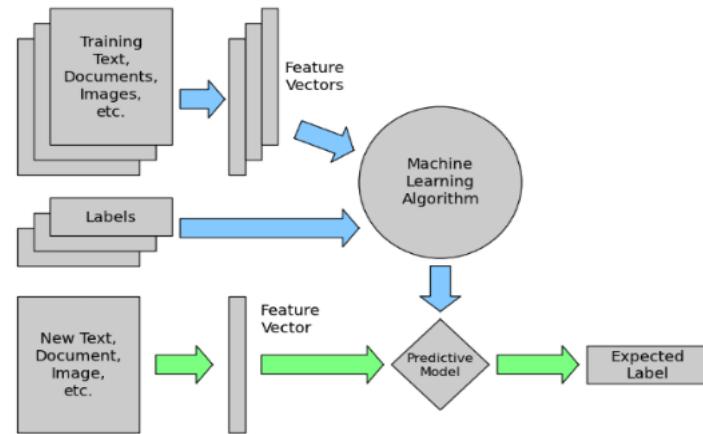
Una red neuronal simple



Una red neuronal para
Deep Learning

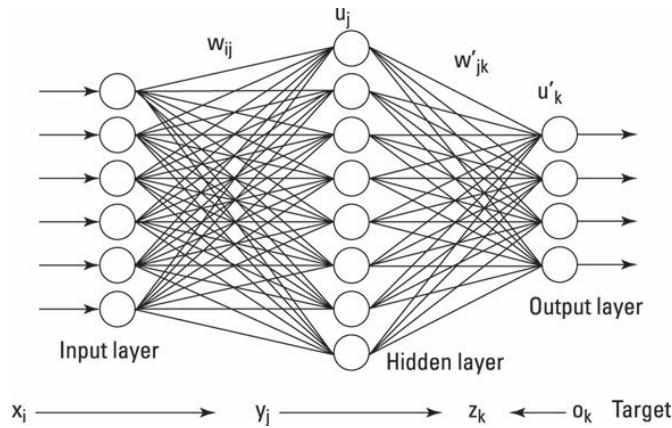
La relación de Big Data con ML

- ▶ Las técnicas de ML se centran en producir modelos
- ▶ El modelo se basa en aprender del pasado para predecir el futuro
- ▶ Mas datos de entrenamiento producen mejores modelos

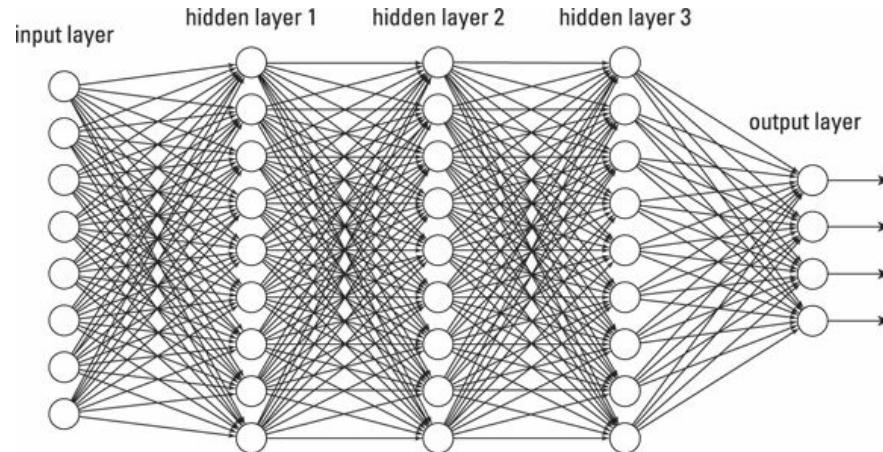


Deep learning requiere de muchos datos

Una red neuronal simple



Una red neuronal para Deep Learning



Grandes Momentos de Deep Learning

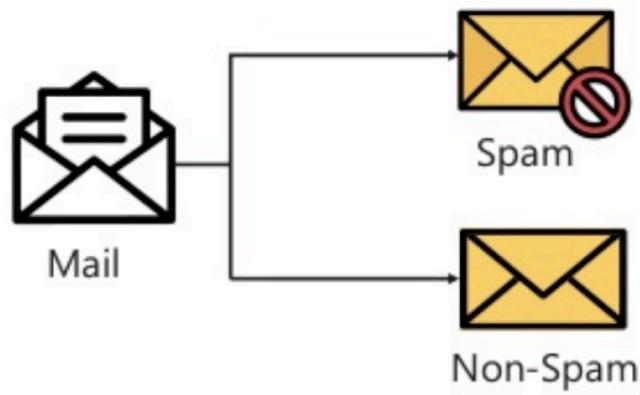
- ▶ 2011 - IBM Watson logra vencer a los dos campeones humanos de Jeopardy
- ▶ 2012 Large Scale Visual Recognition Challenge - Modelo basado en DL logra tasa de errores (16%) muy por debajo de los records (28%)
- ▶ 2016 - Google AphaGo logra vencer al campeón mundial de Go



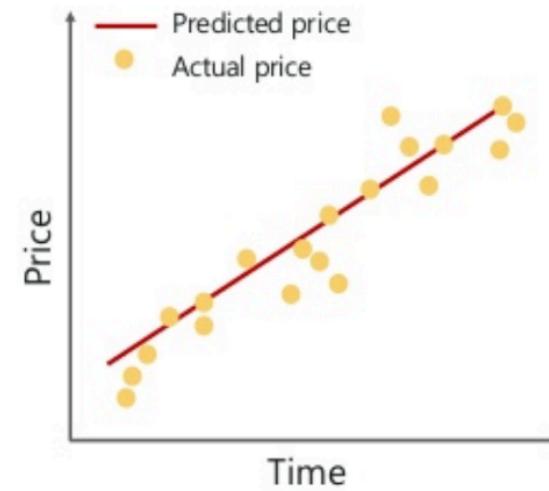
Modelos: Clasificación vs Regresión

- ▶ Clasificación
 - ▶ Modelo maneja valores discretos (pocos valores)
 - ▶ Modelo puede ser presentado como
 - ▶ Reglas de clasificación
 - ▶ Arboles de decisión
 - ▶ Redes neuronales (Deep learning)
- ▶ Regresión
 - ▶ Similar pero modelo maneja valores continuos en lugar de clases
 - ▶ Permite predecir valores inexistentes o no disponibles

Clasificación

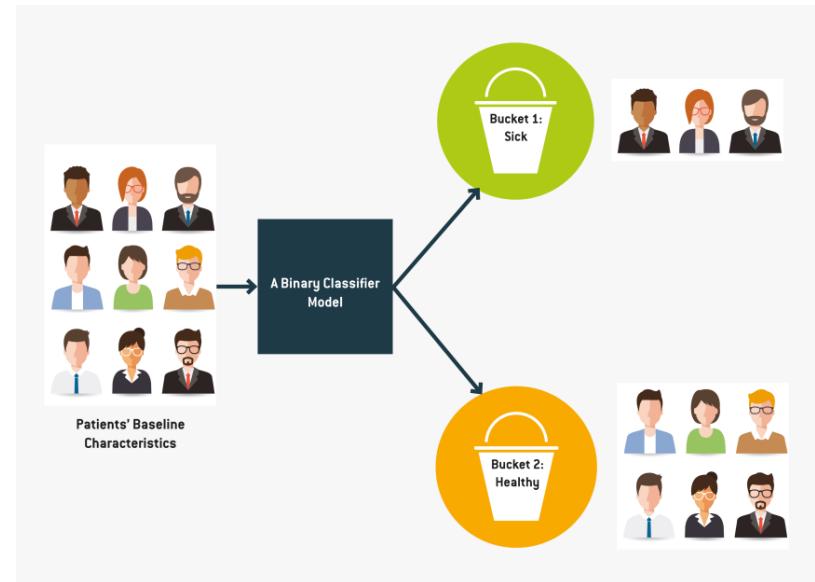


Regresión



Clasificación Binaria

- La mas común de los modelos predictivos
- se requiere una decisión si/no
- ejemplos
 - enfermo o saludable ?
 - se atrasará un vuelo más de 15 minutos ?
 - esta persona va a comprar o no una bicicleta ?
 - podemos prestarle dinero ?
 - tendremos un número de nuevos casos menor a 500 para el 25 de Octubre ?
- puede asociarse un grado de confianza (lloverá con un 80%)



Clasificación Multinivel

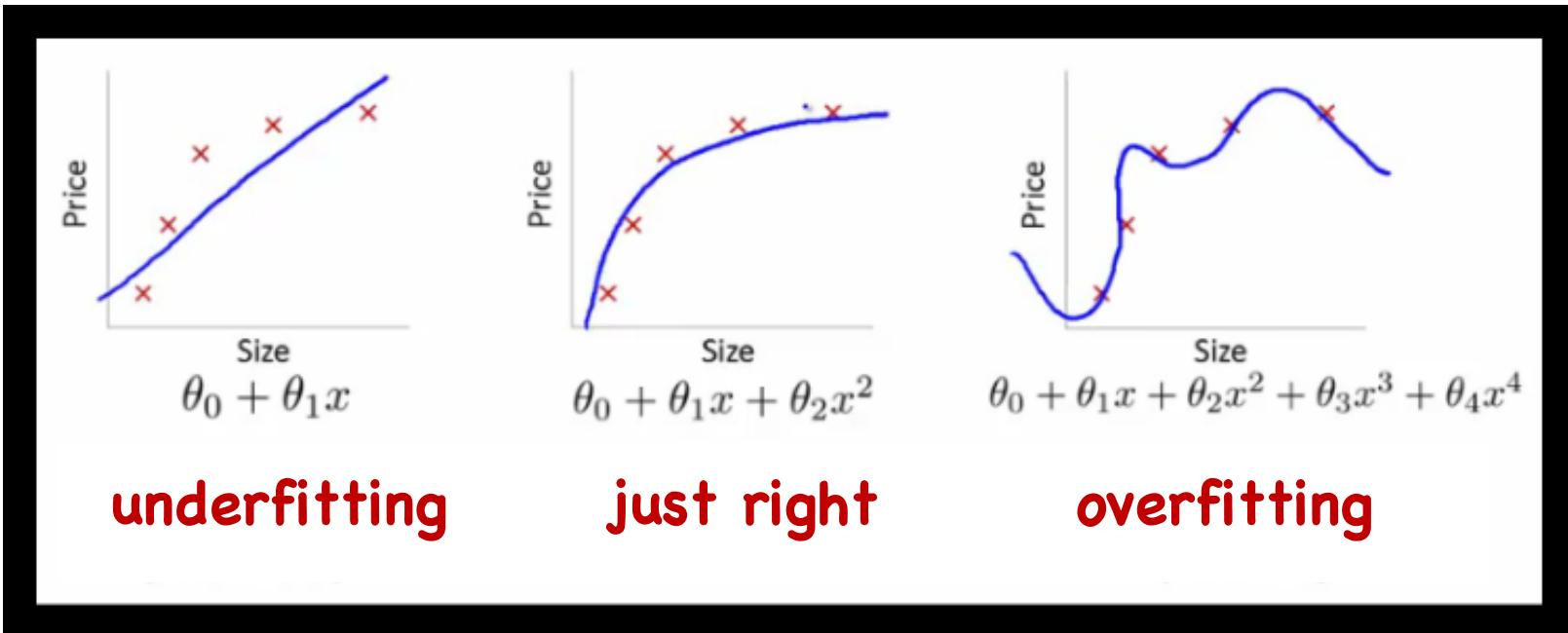
- similar a la anterior pero hay mas de dos intervalos
 - el árbol es un roble/fresno/abedul/castaño/nogal/otro
 - se da el crédito/se rechaza/se estudia
- probabilidades son mas complejas (no basta p y $1-p$)

Spruce/Fir
Lodgepole Pine
Ponderosa Pine
Cottonwood/Willow
Aspen
Douglas-fir
Krummholz

7 Cover Types



Modelos: *Underfitting* y *Overfitting*



Evaluación de un Modelo

- ▶ Para un clasificador binario podemos tener 4 situaciones distintas
 - ▶ true negative (dijo que no llovía y no llovió)
 - ▶ true positive (dijo que llovía y llovió)
 - ▶ false positive (dijo que llovía y no llovió)
 - ▶ false negative (dijo que no llovía y llovió)

$$\text{accuracy} = \frac{\text{decisiones correctas}}{\text{total de decisiones}}$$

- ▶ Si nos dicen que el clasificador tiene una precisión del 80% no estamos diferenciando falsos positivos de falsos negativos
- ▶ ¿Da lo mismo que se equivoque en predecir lluvia que en predecir no lluvia ?
- ▶ ¿Y si fuese un clasificador para decidir si una lesión es cancerosa o no ?

Mas información

Matriz de confusión

	Positive	Negative
Predicted Yes	true positives	false positives
Predicted No	false negatives	true negatives

Measure	Formula
Classification rate (accuracy)	$\frac{\text{true negative} + \text{true positive}}{\text{total observations}} \times 100$
Missclassification rate	$(1 - \frac{\text{true negative} + \text{true positive}}{\text{total observations}}) \times 100$
Sensitivity (true positive rate)	$\frac{\text{true positive}}{\text{true positive} + \text{false negative}} \times 100$
Specificity (true negative rate)	$\frac{\text{true negative}}{\text{false positive} + \text{true negative}} \times 100$
1-Specificity (false positive rate)	$\frac{\text{false positive}}{\text{false positive} + \text{true negative}} \times 100$

Reconocimiento de Imágenes

- ▶ Aplicación para reconocer alimentos desde una foto
- ▶ Casos fáciles



- ▶ Casos Difíciles



Data Sets de Entrenamiento

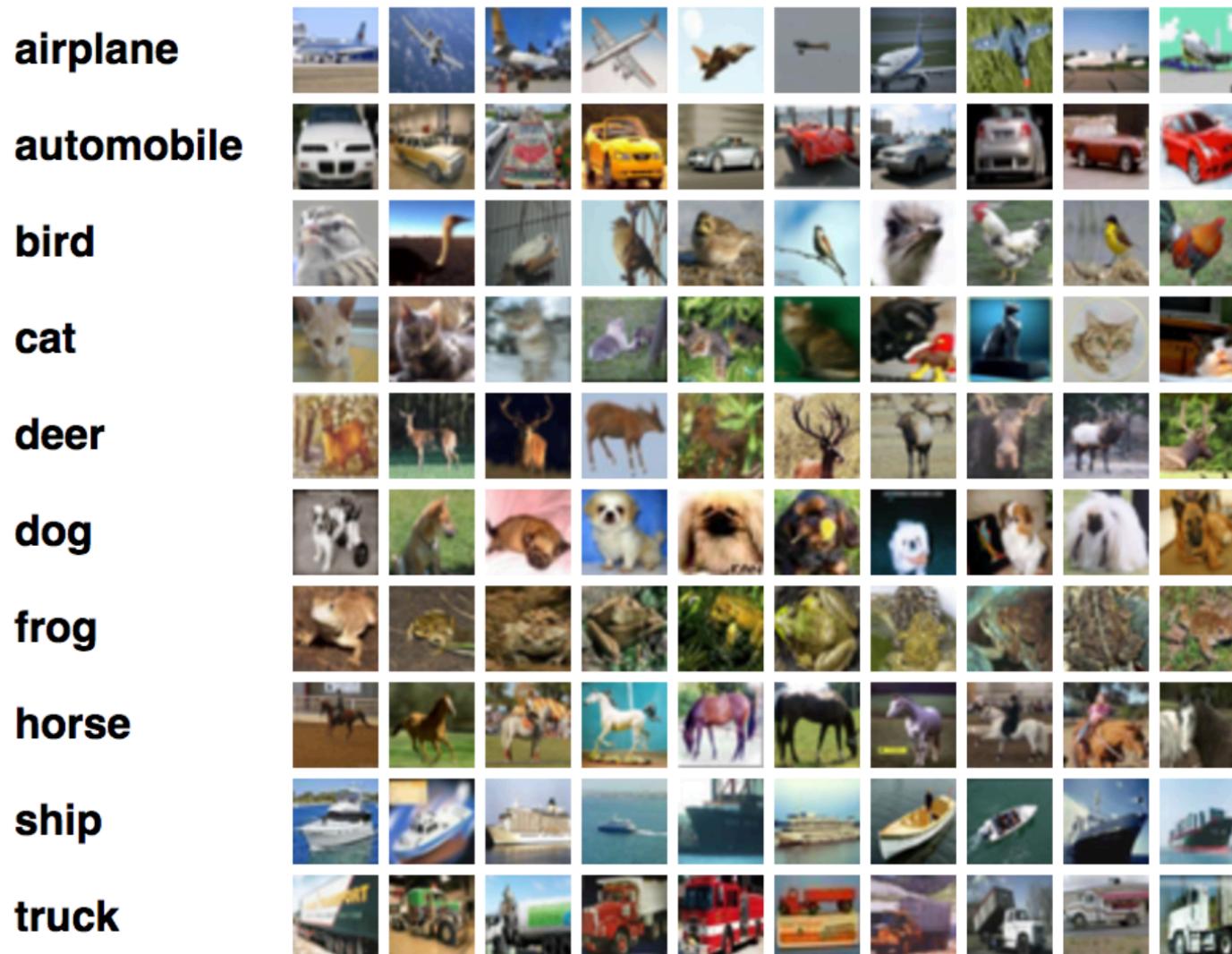


Illustration of images of CIFAR-10 data. Source [here](#)

Es una aplicación muy importante !

- ▶ Reconocimiento de rostros (en China usado intensamente)
- ▶ Identificación de patologías (cancer, ceguera)
- ▶ DART - Sistema para prevenir la ceguera



23 de mayo de 2018

Ministro de Salud presenta software que permitirá triplicar la cantidad de exámenes para prevenir la ceguera diabética

ML vs Estadística

- ▶ Desde hace muchos años los modelos estadísticos son usados para hacer predicciones
- ▶ En eso hay cierta superposición con ML (supervisada especialmente)
- ▶ Más énfasis en inferencia (desde muestra a la población)
- ▶ Características distintivas de ML
 - ▶ énfasis en predecir
 - ▶ evaluar desempeño de predicciones
 - ▶ preocupación por "overfitting"
 - ▶ preocupación por el desempeño

Y que es Minería de Datos ?

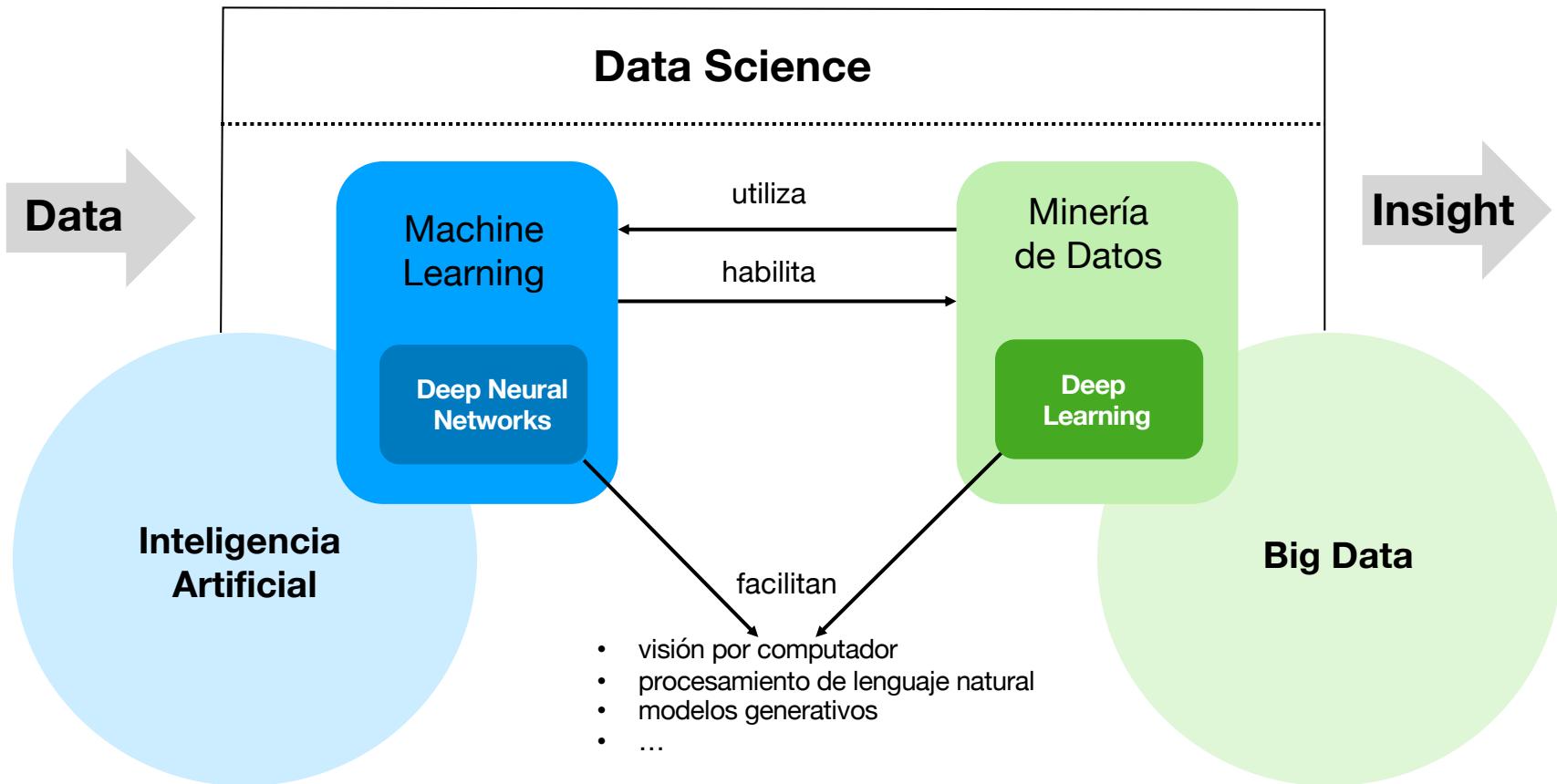


- ▶ Exploración y análisis de grandes sets de datos para descubrir patrones o reglas significativas
- ▶ Se considera una parte de lo que cae bajo Data Science
- ▶ Distinto a analytics porque el foco es predecir el futuro y no entender mejor la data
- ▶ Técnicas usadas en Minería de Datos son las que se usan en ML
- ▶ En el corazón de la minería de datos se encuentran técnicas de ML
- ▶ Minería de datos se asocia mas cercana a las bases de datos, ML mas cerca de la IA
- ▶ Aplicaciones en detección de fraudes, gestión de riesgo en créditos, filtrado de spam, etc.

Tipos de minería de datos

- ▶ Testeo de hipótesis (Estadística clásica)
 - ▶ Data es usada para responder una pregunta o ganar en comprensión de algo
- ▶ Minería de Datos dirigida
 - ▶ Construir un modelo que explica o predice el comportamiento de una o mas variables seleccionadas (target)
- ▶ Minería de Datos no dirigida
 - ▶ Encontrar patrones que no necesariamente están asociados a variables seleccionadas

Parece complicado ...



... pero en el fondo es muy simple

- ▶ Data está siendo producida en forma masiva en todo momento en todos lados
- ▶ Ríos de datos estructurados y no estructurados fluyen en los mundos físico y digital, es el mundo de la *big data*
- ▶ Ingenieros de datos para capturar y manejar estos inmensos flujos
- ▶ Científicos de datos para derivar información valiosa (*insights*) desde esta data: conclusiones y predicciones que permiten mejorar el negocio, la salud o la vida social

Proceso para hacer una predicción usando ML

1. Framing – qué es lo que estamos observando y qué exactamente es lo que queremos predecir
2. Data Collection – recopilar, limpiar y preparar los datos
3. Data Analysis – visualizar y analizar los datos
4. Feature Processing – seleccionar y preparar las variables que intervienen en la predicción
5. Model Building – diseñar y construir el algoritmo de aprendizaje
6. Training – Alimentar con datos el modelo y evaluar su calidad
7. Prediction – Usar el modelo para generar predicciones para nuevas instancias

Tipos de Tareas en ML

Pregunta	Tarea	Salud	Retail	Finanzas
Si o No	Detección	Detección de Cancer	Poner publicidad	Ciberseguridad
Qué tipo	Clasificación	Clasificación de imágenes	Análisis de carrito	Score Crédito
Qué tamaño	Segmentación	Tamaño del Tumor	Tipos de cliente	Análisis de Riesgo
Qué resultado	Predicción	Pronóstico	Sentimiento Comportamiento	Detección de Fraude
Que acción tomar	Recomendación	Terapia	Recomendación	Fast Trading

Cuando usar estas técnicas

- ▶ el problema no puede ser resuelto en base a soluciones basadas en reglas
- ▶ el modelo es demasiado complejo o hay demasiados factores en juego
- ▶ el número de inputs o factores puede escalar rápidamente

Por qué aprender un lenguaje de programación

- ▶ herramienta fundamental del data scientist (inevitable)
- ▶ puede apoyar en cada una de las categorías anteriores
- ▶ no es necesario ser un programador profesional o un ingeniero de software
- ▶ no siempre es posible obtener lo que uno quiere con un software comprado
- ▶ a veces es necesario un poco de código para mejorar o procesar algo más a lo obtenido con otras herramientas
- ▶ permite extraer y procesar datos de fuentes muy diversas
- ▶ permite explorar nuevas ideas o ideas "locas"

Por qué R

- ▶ fácil de aprender (incluso para gente que no programa)
- ▶ flexible
- ▶ muy enfocado a la computación estadística (excelente para sumarizar, calcular probabilidades, etc)
- ▶ fácil crear visualizaciones de los datos
- ▶ usado por mucha gente
- ▶ disponible en todas las plataformas
- ▶ muchos recursos de aprendizaje disponibles

Obtención de la Data

- ▶ Fuente principal son Bases de Datos convencionales (relacionales) - Uso de SQL
- ▶ Otros tipos de bases de datos - usar motor correspondiente o usar APIs de acceso a los datos
- ▶ Redes sociales - acceso a través de las APIs respectivas
- ▶ Información de la Web no disponible a través de APIs - usar web scraping

Bases de Datos Convencionales

- ▶ Set de información estructurada
- ▶ Acorde a un modelo de datos
- ▶ Se puede consultar información con facilidad (lenguaje de consultas)
- ▶ Se puede agregar/eliminar/actualizar la información en ella
- ▶ Se puede acceder en forma concurrente
- ▶ La información se mantiene consistente

Motores de Base de Datos

- ▶ Un producto de Software que permite
 - ▶ crear una BD
 - ▶ ingresar, actualizar y eliminar información de la BD
 - ▶ consultar la información contenida en la BD
 - ▶ controlar el acceso a usuarios autorizados
 - ▶ mantener la consistencia aún con actualizaciones concurrentes y posibles fallas
- ▶ Implementa uno o más modelos de datos
 - ▶ Los mas populares: Relacionales (PostgreSQL, MySQL, Oracle, SQL Server)
 - ▶ Documentos: MongoDB, CouchDB

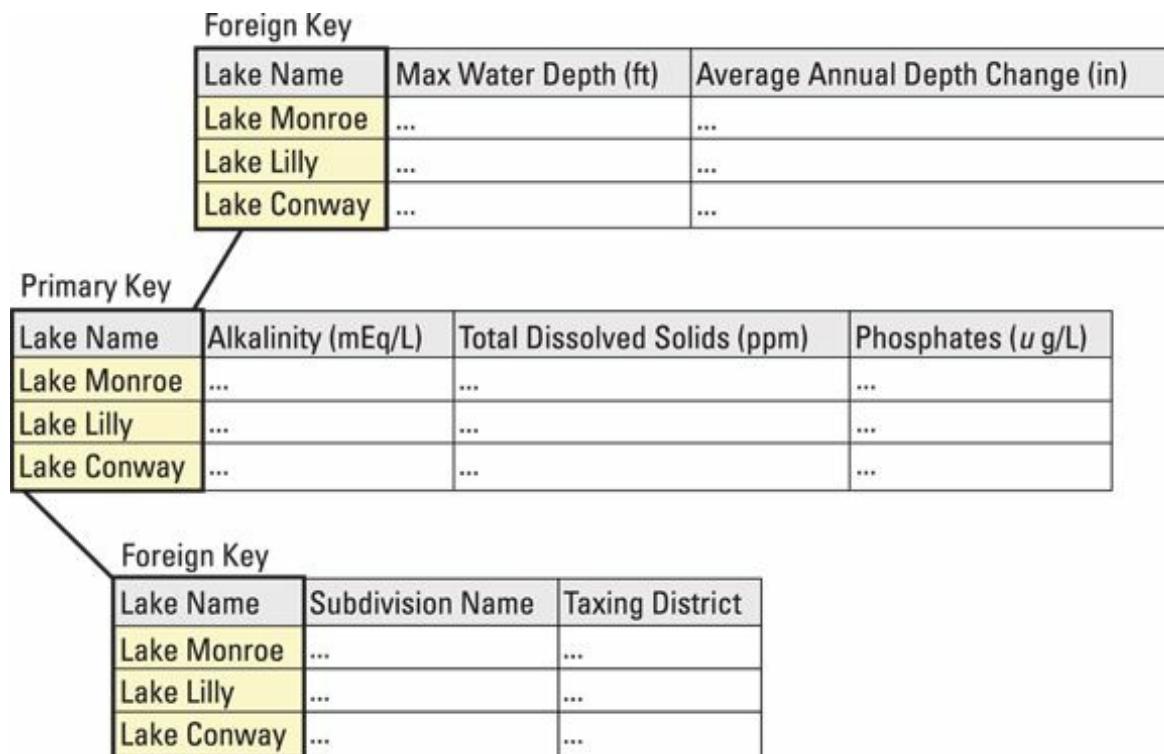
Modelos de Datos

- ▶ Diferencia entre Modelo genérico y Modelo de Datos
- ▶ Un Motor de BD implementa algún modelo genérico (Relacional, Documentos)
- ▶ La BD a almacenar debe ser modelada
 - ▶ cuales son las entidades a manejar y sus atributos
 - ▶ como se relacionan las entidades
- ▶ El modelo de datos puede ser llevado a uno de los modelos genéricos (relacional o documentos)

El Modelo Relacional

- ▶ Solo hay tablas
- ▶ Lenguaje de consulta estandard SQL (no procedural) fácil de aprender
- ▶ Tablas tienen un atributo que identifica la fila (tupla) que se denomina clave primaria
- ▶ Pueden tener otros atributos sobre los cuales se quiere facilitar el acceso
- ▶ Se relacionan entre si conectando claves primarias con claves foráneas

Primary Keys y Foreign Keys



Ejemplo Modelo Relacional

Publishers

PubID	Publisher	PubAddress
03-4472822	Random House	123 4th Street, New York
04-7733903	Wiley and Sons	45 Lincoln Blvd, Chicago
03-4859223	O'Reilly Press	77 Boston Ave, Cambridge
03-3920886	City Lights Books	99 Market, San Francisco

Titles

ISBN	AuthorID	PubID	Date	Title
1-34532-482-1	345-28-2938	03-4472822	1990	Cold Fusion for Dummies
1-38482-995-1	392-48-9965	04-7733903	1985	Macrame and Straw Tying
2-35921-499-4	454-22-4012	03-4859223	1852	Fluid Dynamics of Aqueducts
1-38278-293-4	663-59-1254	03-3920886	1967	Beads, Baskets & Revolution

Authors

AuthorID	AuthorName	AuthorBDay
345-28-2938	Haile Selassie	14-Aug-92
392-48-9965	Joe Blow	14-Mar-15
454-22-4012	Sally Hemmings	12-Sep-70
663-59-1254	Hannah Arendt	12-Mar-06

- ▶ Ejemplos de Consultas en SQL

¿Cuál es la dirección de Wiley ?

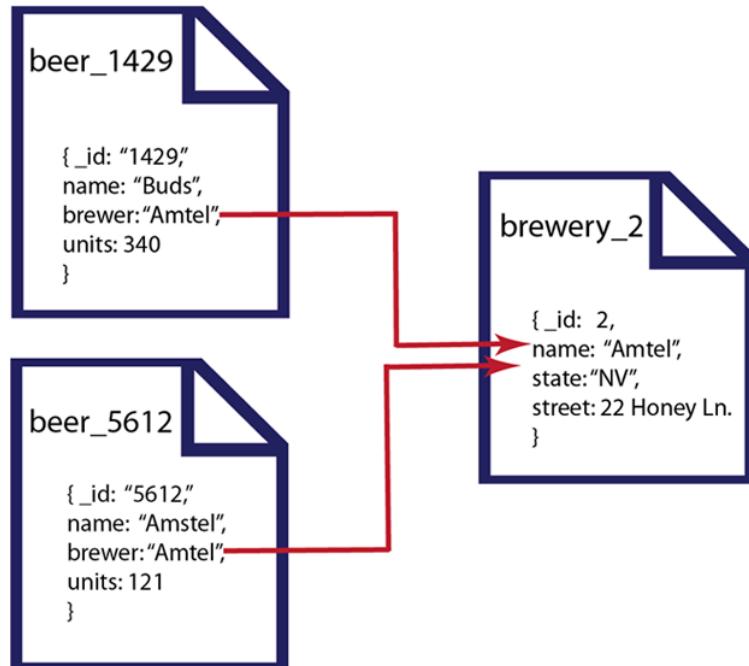
```
Select PubAddress  
From Publishers  
Where Publisher like '%Wiley%'
```

¿Qué títulos de Random House hay ?

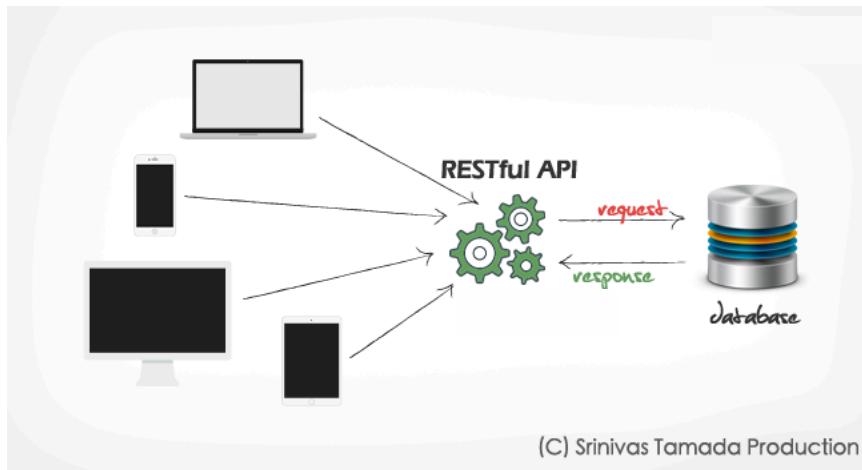
```
Select Title  
From Publishers Join Titles On Publishers.PubID = Titles.PubID  
Where Publisher = 'Random House'
```

El Modelo de Documentos

- ▶ Permite manejar información semiestructurada (JSON)



Acceso a través de una API



- ▶ Cada vez hay mas data pública disponible para obtener desde organizaciones gubernamentales, ONGs, etc (open data, world bank, etc)
- ▶ Se proveen mecanismos para visualizar la data pero también para obtenerla en forma cruda desde un programa

data.gov.uk | Find open data

Publish your data Support

We've been improving data.gov.uk to help you find and use open government data.
[Discover what's changed](#) and [get in touch](#) to give us your feedback.

[Don't show this message again](#)

Find open data

Find data published by central government, local authorities and public bodies to help you build products and services

Business and economy

Small businesses, industry, imports, exports and trade

Environment

Weather, flooding, rivers, air quality, geology and agriculture

Mapping

Addresses, boundaries, land ownership, aerial photographs, seabed and land terrain

Crime and justice

Courts, police, prison, offenders, borders and immigration

Government

Staff numbers and pay, local councillors and department business plans

Society

Employment, benefits, household finances, poverty and population

Defence

Armed forces, health and safety, search and rescue

Government spending

Includes all payments by government departments over £25,000

Towns and cities

Includes housing, urban planning, leisure, waste and energy consumption

Education

Students, training, qualifications and the National Curriculum

Health

Includes smoking, drugs, alcohol, medicine performance and hospitals

Transport

Airports, roads, freight, electric vehicles, parking, buses and footpaths

Otros Ejemplos

World Bank Open Data

Free and open access to global development data

Search data e.g. GDP, population, Indonesia

Browse by [Country](#) or [Indicator](#)



NASA's Open Data Portal

Search

Categories

Aerospace

Applied Science

Apps

Earth Science

Management/Operations

Show All...

24217 Results

Sort by [Recent](#)

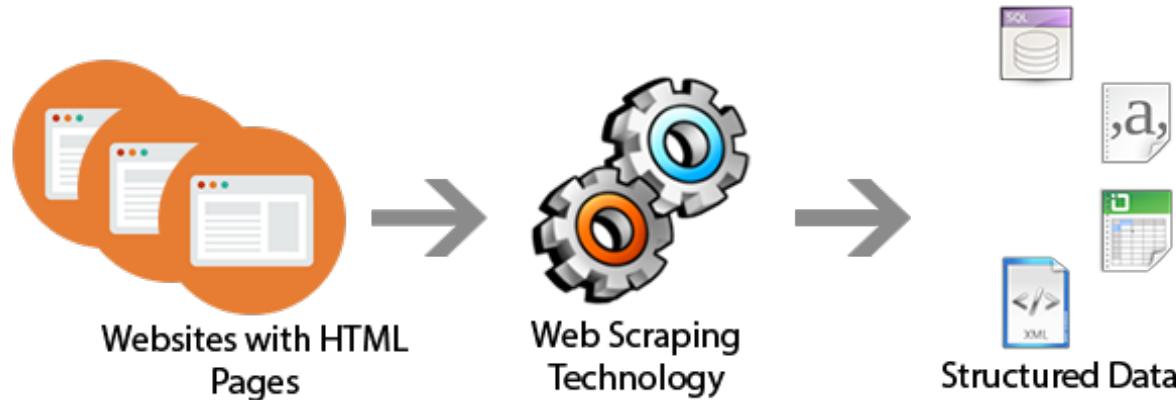
[Thematic Mapper \(TM\) Mosaics \(1984-1997\)](#)

Mosaic data products, which are also available for Tri-Decadal Global Landsat Orthorectified TM and ETM+ Pan-sharpened data, and may be...
[More](#)

Tags [land surface](#), [national geospatial data asset](#), [topography](#), [ngda](#), [surface radiative properties](#), and 1 more

[TES/Aura L2 Formic Acid Lite Nadir V007](#)

Y si no hay API disponible ?



- Web Scraping
 - Se utiliza un programa (R, Python u otros) para extraer los datos de interés directamente desde las páginas servidas a un navegador
 - Es necesario procesar el fuente (HTML, CSS, JavaScript) para extraer lo que es de interés para dejarlo en un formato estructurado o semiestructurado (JSON, CSV, XML, etc)

Este Diplomado ...

- ▶ **1. Seminario**
 - ▶ Vista desde 10.000 m de altura de toda el área de big data, machine learning y data science
- ▶ **2. Programación en R para Data Science**
 - ▶ Habilidades básicas de programación usando el lenguaje R
- ▶ **3. Arquitectura e Infraestructura para Big Data y Data Science**
 - ▶ Técnicas, arquitecturas, plataformas
- ▶ **4. Aplicaciones de Data Science**
 - ▶ Ejemplos del uso de técnicas de Data Science en el mundo real
- ▶ **5. Minería de Datos**
 - ▶ Fuentes de datos, preparación de los datos, extracción de patrones, herramientas, algoritmos
- ▶ **6. Visualización**
 - ▶ Fundamentos y aplicaciones con casos de uso reales
- ▶ **7. Machine Learning**
 - ▶ Introducción y aplicaciones del aprendizaje de máquina con énfasis en aprendizaje supervisado