

DIPLOMADO EN CIENCIA DE DATOS

Módulo: Minería de Datos

Universidad Nacional de Colombia

Expositora

Juliana Guerrero Velásquez

- Pregrado en Estadística. Universidad Nacional de Colombia
- MsC. Statisticis. KU Leuven (Belgium)
- Intereses: Análisis social, de comportamiento y educación. Modelos mixtos-jerárquicos, variable latente, variable discreta
- Contacto: ljuliana.guerrero@gmail.com



Objetivos

Al final de este módulo los estudiantes:

Conocerán las principales herramientas para analizar grandes volúmenes de datos

Podrán utilizar herramientas para hallar patrones y valor en los datos, para optimizar las decisiones de negocio

Cómo aplicar algunas de las herramientas más importantes en R y Python

Básicamente: Fundamentos y bases de conocimiento esenciales para convertirse en un (mejor) científico de datos!

Logística de clase

Clases:

- Martes y Jueves (Octubre 20, 22, 27, 29): 6:00 – 9:00 pm
- Sábados (Octubre 24 y 31): 9:00 – 12:00 am

Práctico:

- Martes Noviembre 3 y Jueves Noviembre 5

Clases teórico- prácticas

Preguntas? Durante, antes o después de clase o por e-mail

Los slides estarán disponibles antes de la clase

El material principal para los temas tratados en clase serán las diapositivas del curso

Cronograma

Fecha	Hora inicio	Hora final	Tema
2020-10-20	18:00	21:00	Introducción a ciencia y minería de datos, etapas del proceso de minería de datos
2020-10-22	18:00	21:00	Vectores, matrices, distancias y reducción de dimensionalidad
2020-10-24	9:00	12:00	Reglas de asociación
2020-10-27	18:00	21:00	Algoritmos de agrupación (Clustering)
2020-10-29	18:00	21:00	Validación y caracterización
2020-10-31	9:00	12:00	Análisis de texto*
2020-11-03	18:00	21:00	Proyecto práctico
2020-11-05	18:00	21:00	Proyecto práctico

Prerrequisitos

Conocimientos básicos de estadística y analítica

Conocimientos básicos de R y Python

Motivación y ganas de trabajar

Evaluación

Quices cortos al inicio de cada clase

Proyecto:

- Grupos de 5-6 estudiantes
- Aplicación de las herramientas dadas en clase en un dataset 'real' / dataset propio
- Los entregables del proyecto serán el código del análisis y una presentación de 10 minutos con los resultados y hallazgos

¿PREGUNTAS?

DIPLOMADO EN CIENCIA DE DATOS

Módulo: Introducción a ciencia y minería de datos

Universidad Nacional de Colombia

Contenido

- ¿Qué es ciencia de datos?
- El equipo de ciencia de datos
- Motivación de la minería de datos
- Grandes enfoques de análisis de datos
- Proceso y etapas de un proyecto de ciencia de datos
- Conclusiones

¿QUÉ ES CIENCIA DE DATOS?

¿Qué es ciencia de datos?

Los datos tienen valor y conocimiento, la ciencia de datos los extrae

Pero... para extraer el conocimiento :

- Recolectarlos
- Almacenarlos
- Administrarlos
- Analizarlos
- Interpretarlos

Términos usados:

Data Mining ≈ Big Data ≈ Data Analytics ≈ Data Science ≈ Knowledge Discovery ≈
Artificial Intelligence ≈ Deep Learning ≈ Machine Learning

Un científico de datos debe tener habilidades cuantitativas, de programación, comunicación y visualización, entender el negocio y ser creativo.

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21st 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

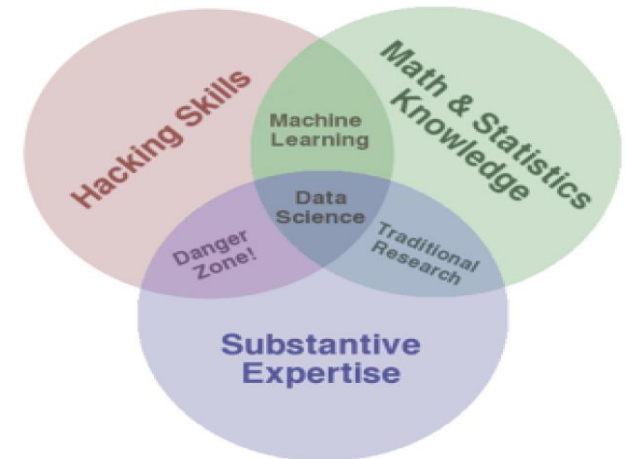
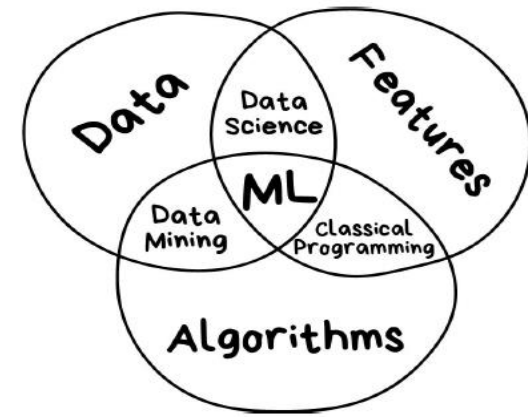

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, e.g., R
- ☆ Databases: SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with 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 tools e.g. Flare, D3.js, Tableau



No hay que dejarse llevar por los diagramas de venn



Josh Wills

@josh_wills



 Follow

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.



RETWEETS

1,046

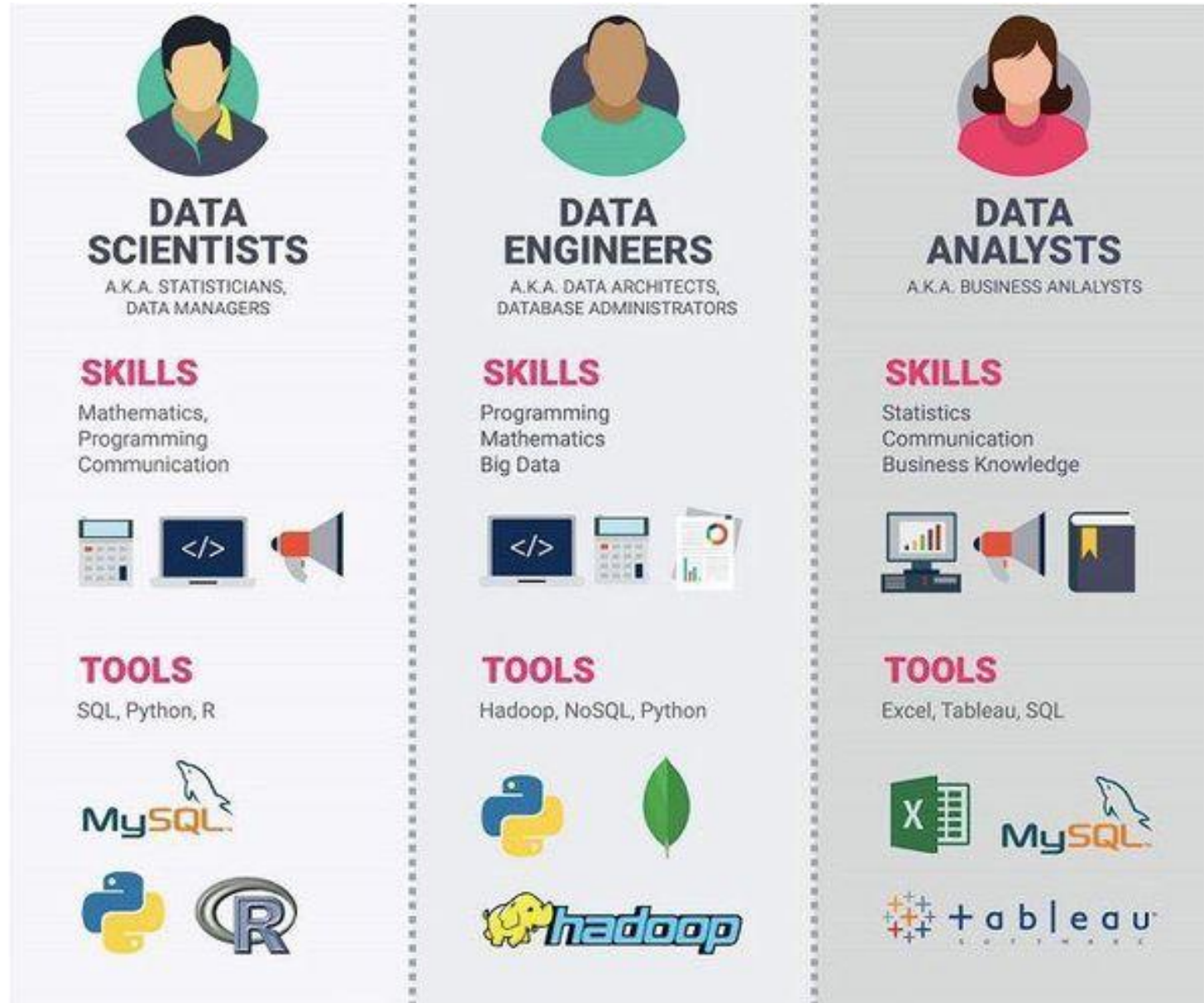
FAVORITES

532



6:55 PM - 3 May 2012

El data science team



¿De qué se trata?

Descubrir patrones y modelos que sean:

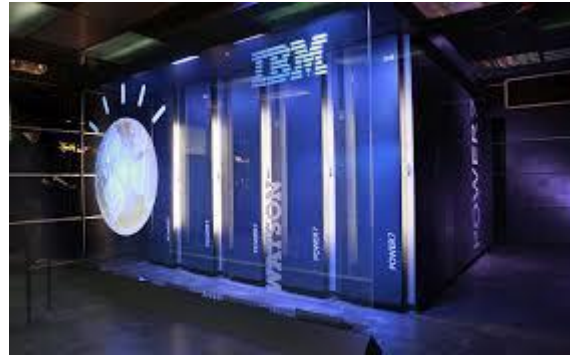
- **Válidos:** nuevos datos, generalizable, a través del tiempo, overfitting
- THE NEURAL NET TANK URBAN LEGEND** (<https://www.gwern.net/Tanks>)



¿De qué se trata?

Descubrir patrones y modelos que sean:

- **Útiles:** Accionable, pregunta del negocio, implementación, mantenimiento, fácil de usar
Podemos predecir quien va a dejar de utilizar nuestro negocio pero y luego?



<https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care>

¿De qué se trata?

Descubrir patrones y modelos que sean:

- **Inesperados:** no obvio, interesante, balance entre algo confiable y descubrimiento
Usuarios de Chrome y Firefox resultan ser mejores empleados (<https://community.spiceworks.com/topic/844160-study-finds-chrome-and-firefox-users-are-better-employees>).



Pero no todo lo inesperado es interesante o válido.

¿De qué se trata?

Descubrir patrones y modelos que sean:

- **Comprensibles:** Lo puede entender un humano “normal”, Black box Vs White box
Qué atributos en mi modelo son importantes? Cómo interpreto las interacciones?

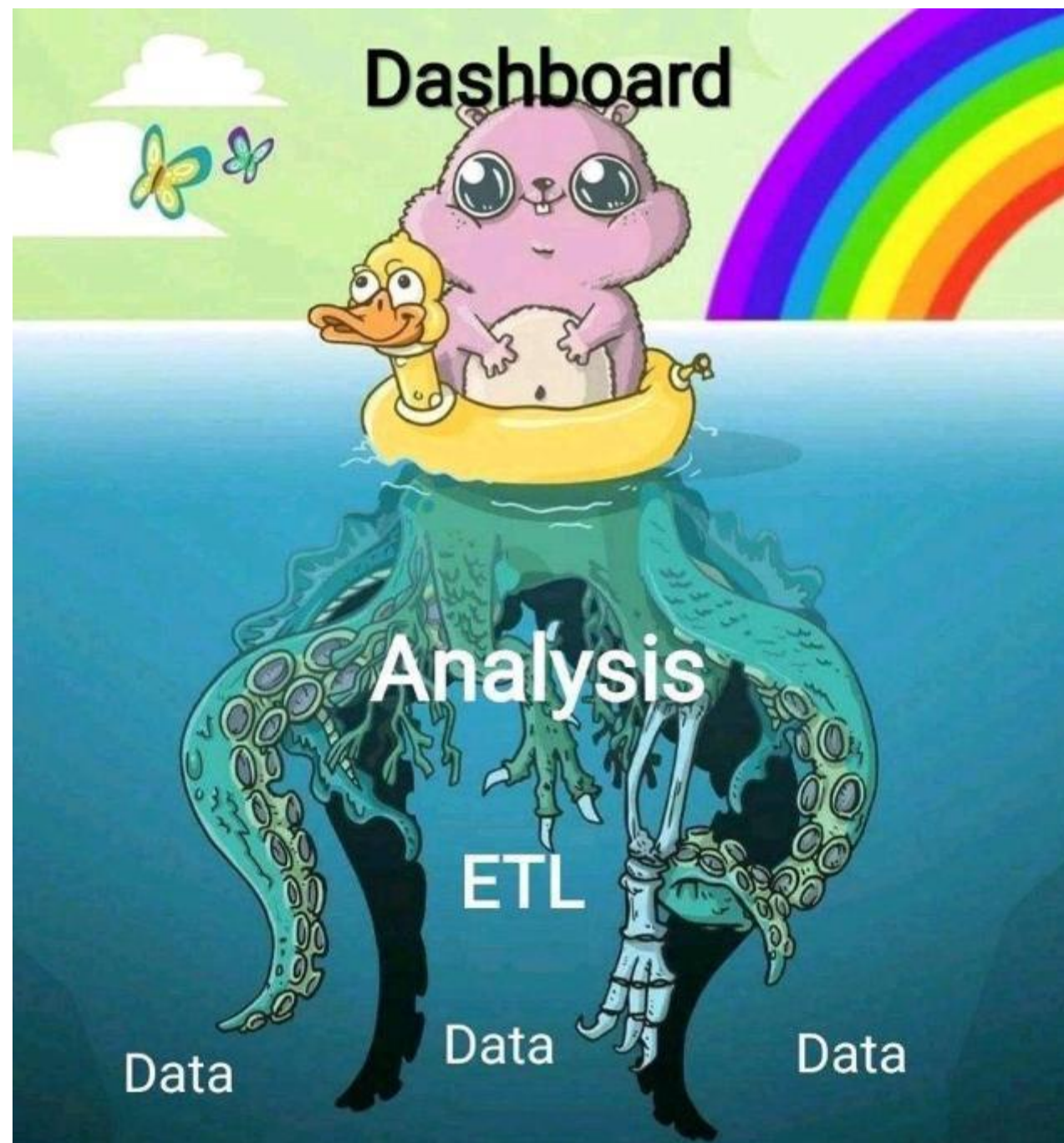


<https://bdtechtalks.com/2020/07/27/black-box-ai-models/>

Es diferente que el algoritmo se equivoque en Alpha Go a que se equivoque un carro que se maneja solo

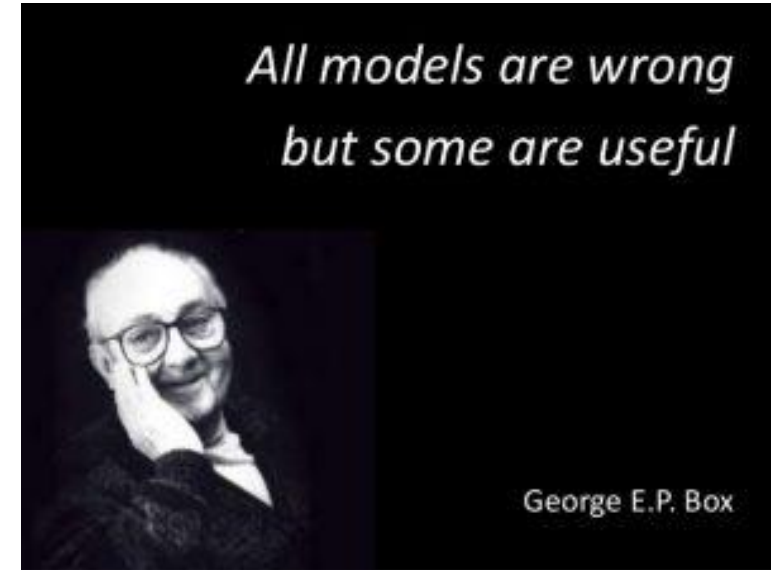
Retos...

- Llevar la idea de negocio (problema) a una técnica(s) de análisis. En DS no hay **UNA** respuesta correcta
- (No) saber todo el esfuerzo que puede llevar el pre-procesamiento de los datos
- Muy pocos datos, observaciones o variables. O demasiadas variables
- Imbalance**
- Calidad de datos, ruido
- Predecir el futuro es duro, difícil de extrapolar (los modelos son ingenuos y perezosos)
- Incluir conocimiento de experto, explicar los modelos
- Una solución válida toma tiempo y datos
- Organización, equipos



La Data

- Precisa (outliers, edad de 300 años vs salario 30M)
Atípicos vs Extremos
- Completa (los missing values son importantes? Falta información también es información)
- Sesgo (minimizarlo) y muestreo
- Actualizada/Relevante



Garbage in, Garbage out; messy data gives messy models

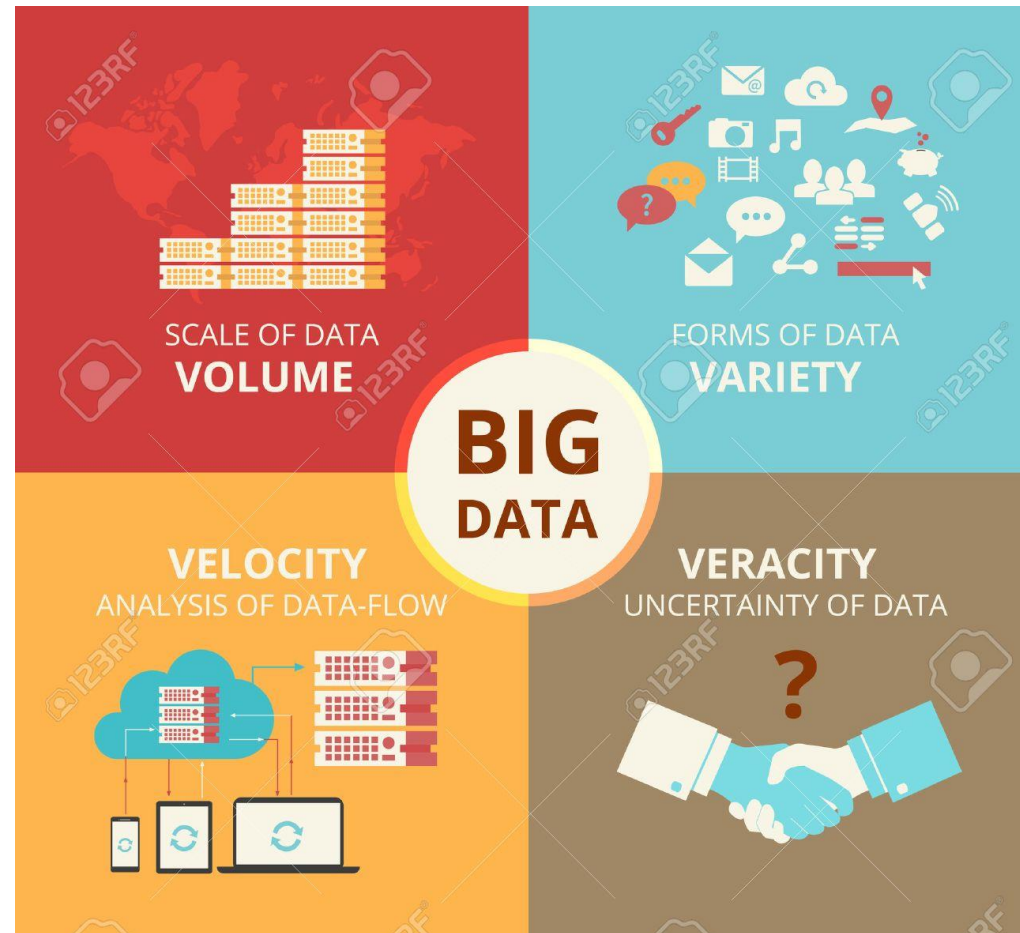
La mejor manera de mejorar el desempeño de un modelo no es buscar herramientas o técnicas sofisticadas, sino mejor la CALIDAD DE LOS DATOS PRIMERO!

Motivación de la minería de datos

- Grandes cantidades de datos recolectados y almacenados

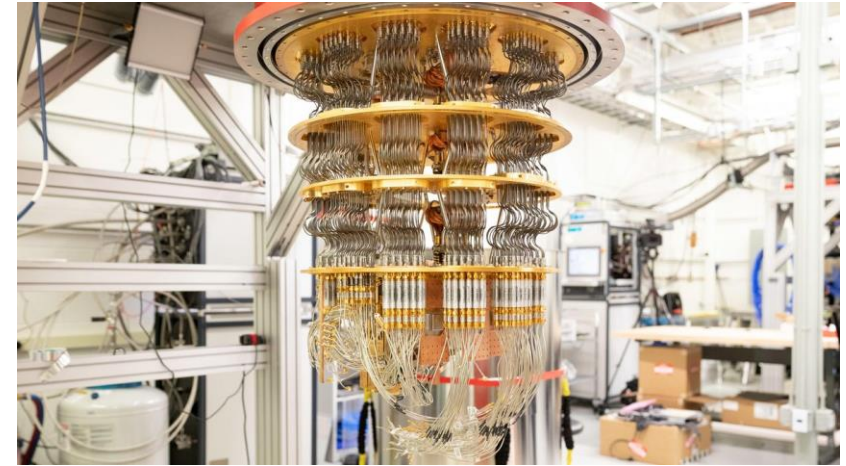
Comprar un café, pedir un Uber, enviar un email, domicilio de Rappi, ver una serie de Netflix, hacer un tweet, cambiar de canción en Spotify

Las 4 Vs del Big Data: Volume, velocity, variety, veracity



Motivación de la minería de datos

- El gasto computacional es cada vez menor, tenemos mejores máquinas para procesar



- La presión de competencia aumenta. Toma de decisiones basadas e informadas en la data. (mejorar servicio/producto, ser más rápido/eficiente, disminuir costos)

Motivación de la minería de datos

- La presión de competencia aumenta. Toma de decisiones basadas e informadas en la data. (mejorar servicio/producto, ser más rápido/eficiente, disminuir costos)

COMMUNITY

5 Stats That Show How Data-Driven Organizations Outperform Their Competition

<https://www.keboola.com/blog/5-stats-that-show-how-data-driven-organizations-outperform-their-competition>



| 5 Companies Using Big Data and AI to Improve Performance

1

<https://www.kolabtree.com/blog/5-companies-using-big-data-and-ai-to-improve-performance/>

PERO... la data no tiene que ser “big” para tener un reto

You don't need big data to do analytics
... but you can

Big data doesn't necessarily mean doing analytics
... but it can

Minería de datos - Nuestro enfoque

Extraer patrones de negocio y/o modelos de decisión matemáticos del procesamiento de un conjunto de datos que sean válidos, útiles, interesantes y comprensibles.

- Analizar eficientemente grandes volúmenes de datos para entenderlos y llegar a hallazgos
- Generar hipótesis a partir de lo observado, descubrir información útil antes desconocida
- Aprendizaje supervisado
- **Aprendizaje no supervisado**
- Deep Learning
- Text mining

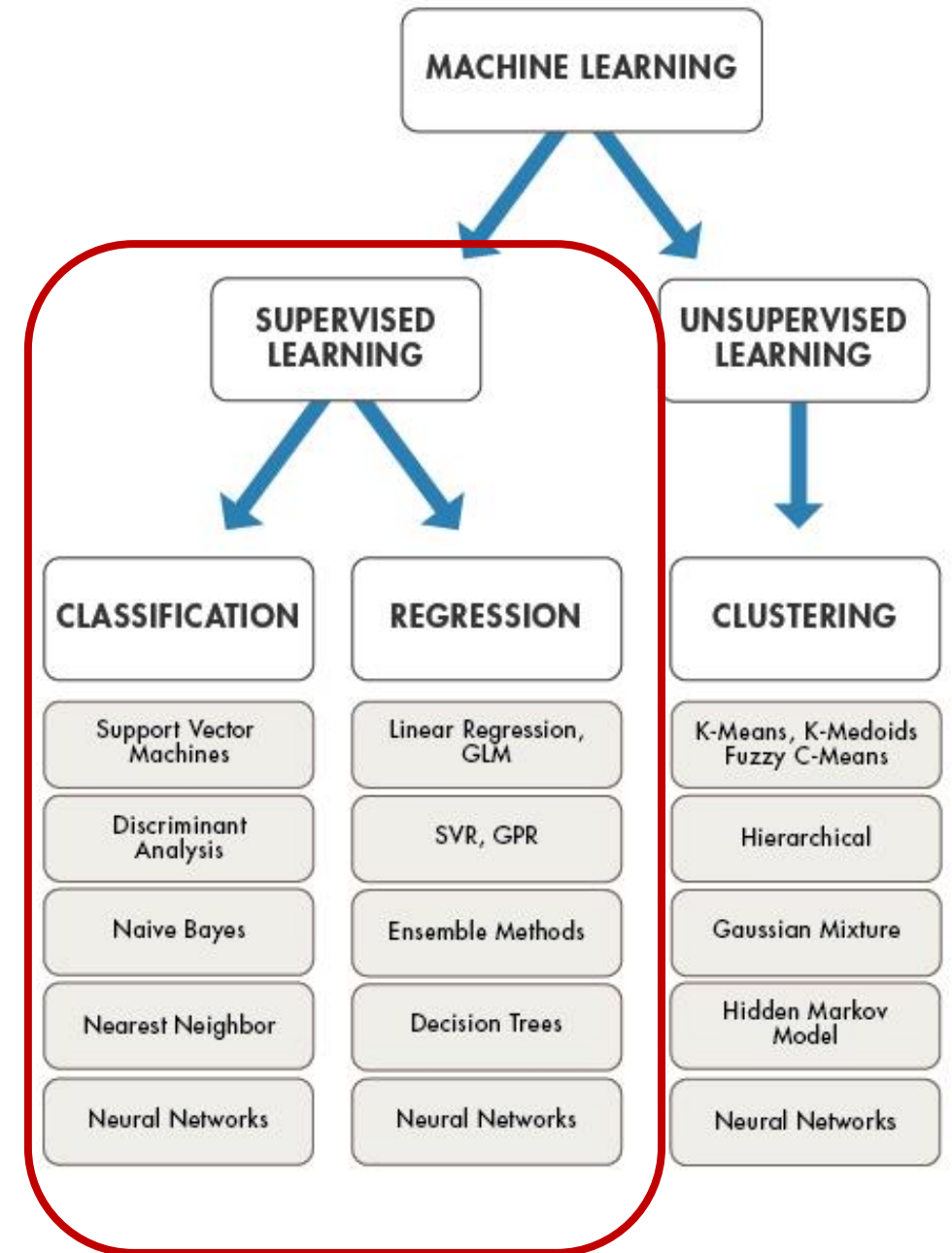
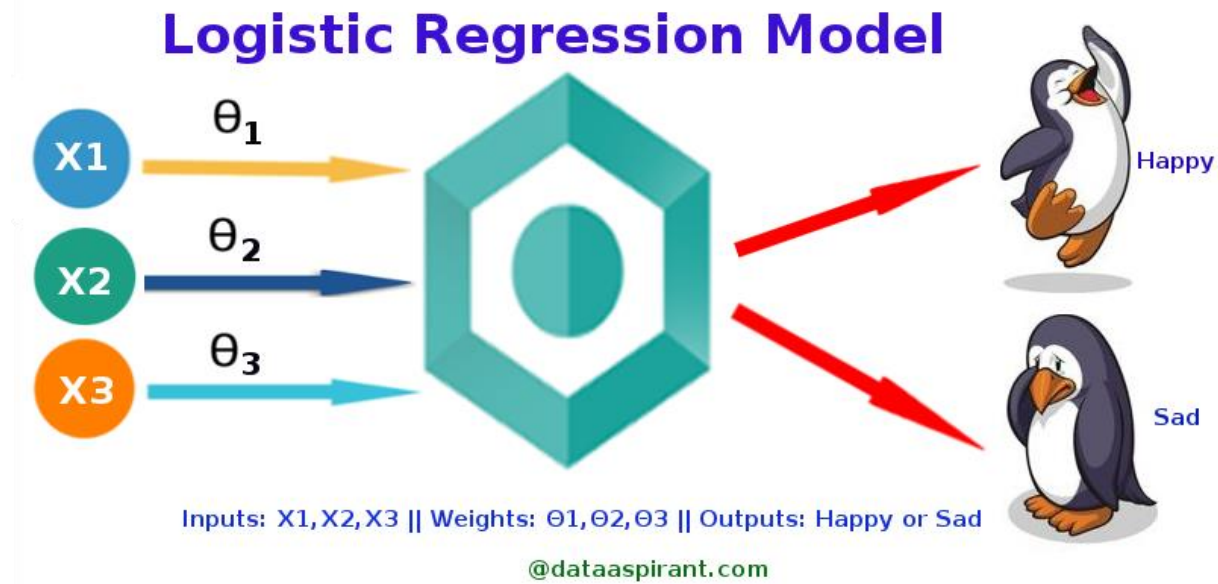
Cómo hacerlo?

Aprendizaje Supervisado

Predecir el futuro basado en patrones aprendidos con data del pasado

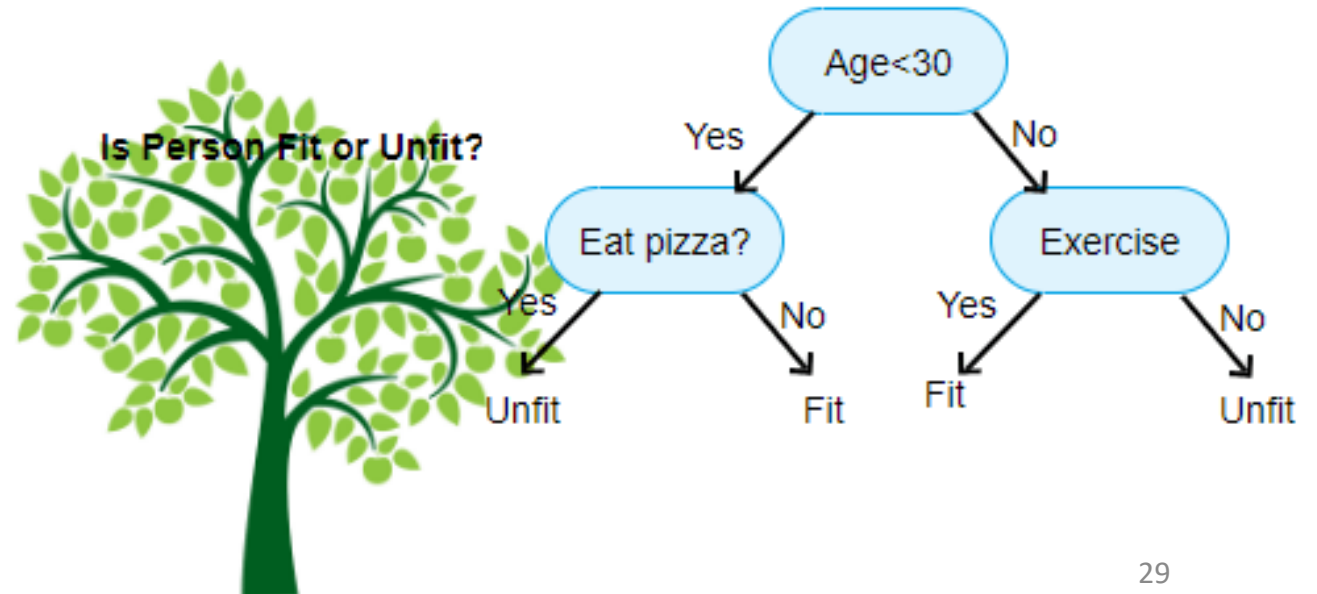
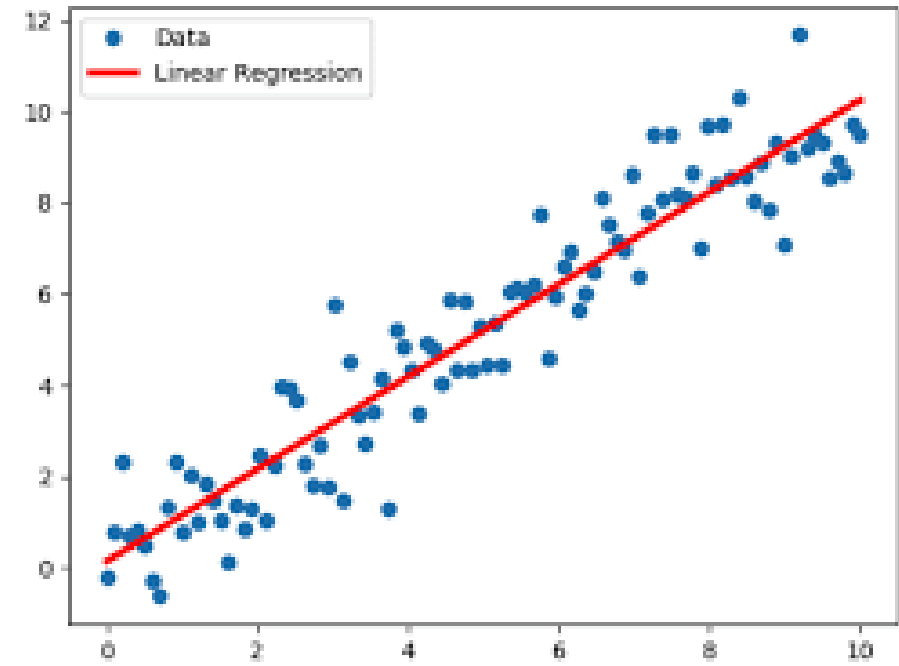
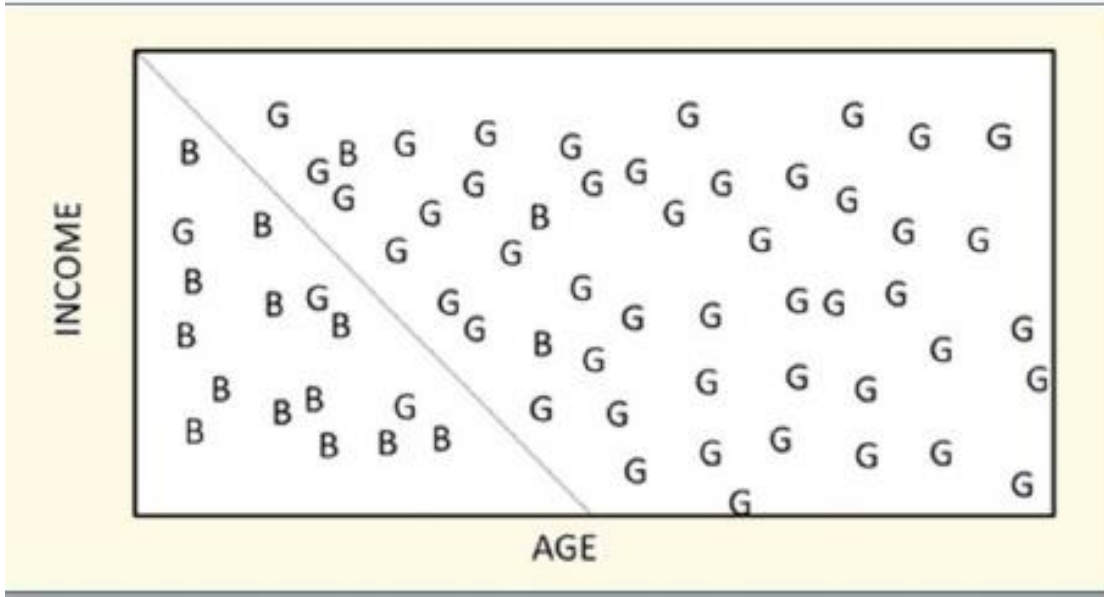
Clasificación (categórico) vs Regresión (continuo)

Se tiene un dataset con labels



Cómo hacerlo?

Aprendizaje Supervisado



Cómo hacerlo?

Aprendizaje Supervisado



Cómo hacerlo? – Nuestro enfoque

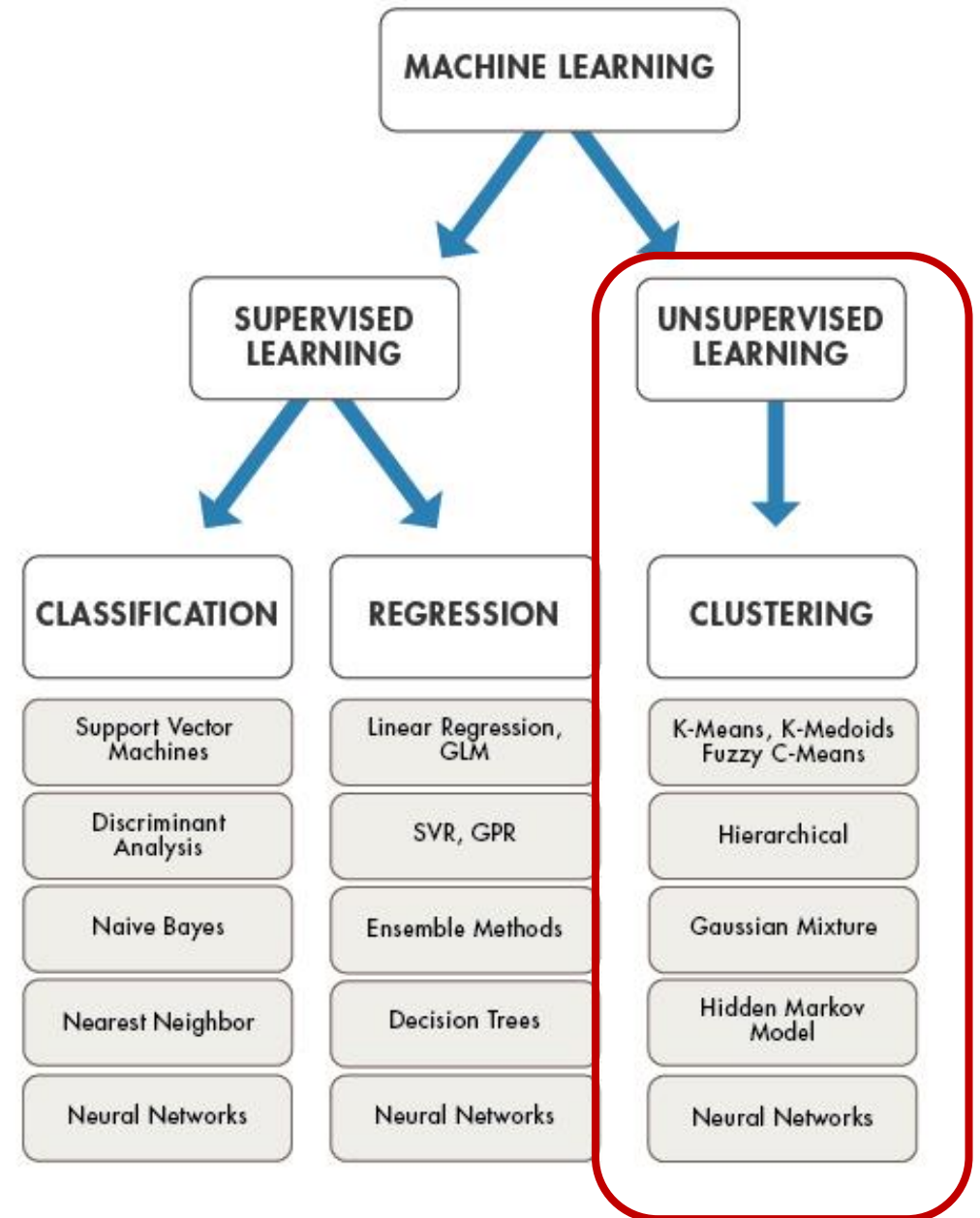
Aprendizaje no supervisado

Describir patrones en la data

Clustering, Reglas de asociación

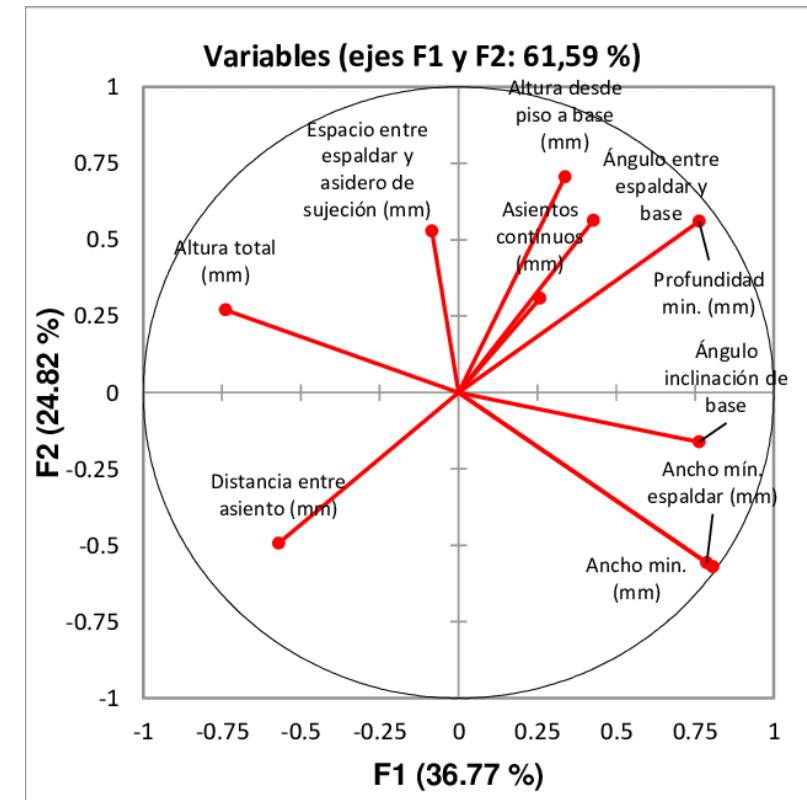
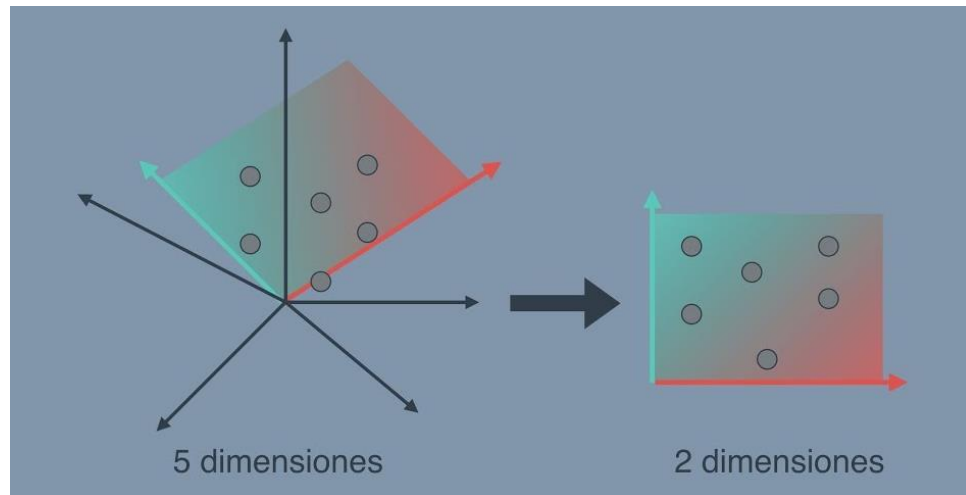
No se necesitan labels

(Hay mas que estas dos..)



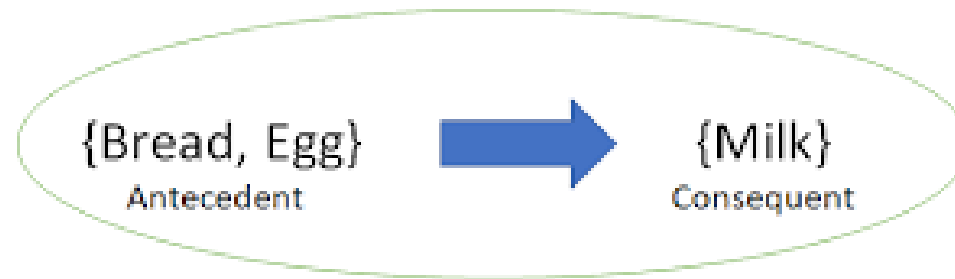
Aprendizaje no supervisado

Análisis de Componentes Principales



Aprendizaje no supervisado

Reglas de asociación

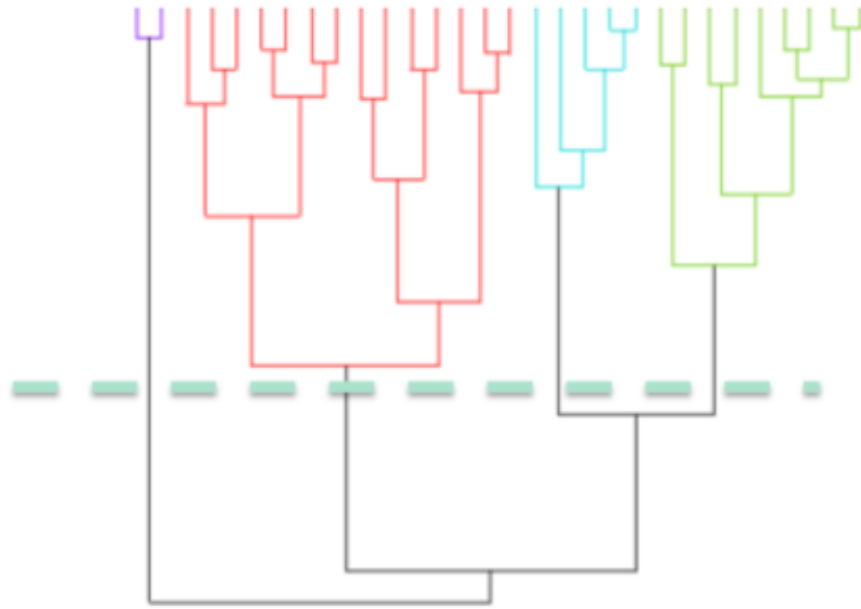


Itemset = {Bread, Egg, Milk}

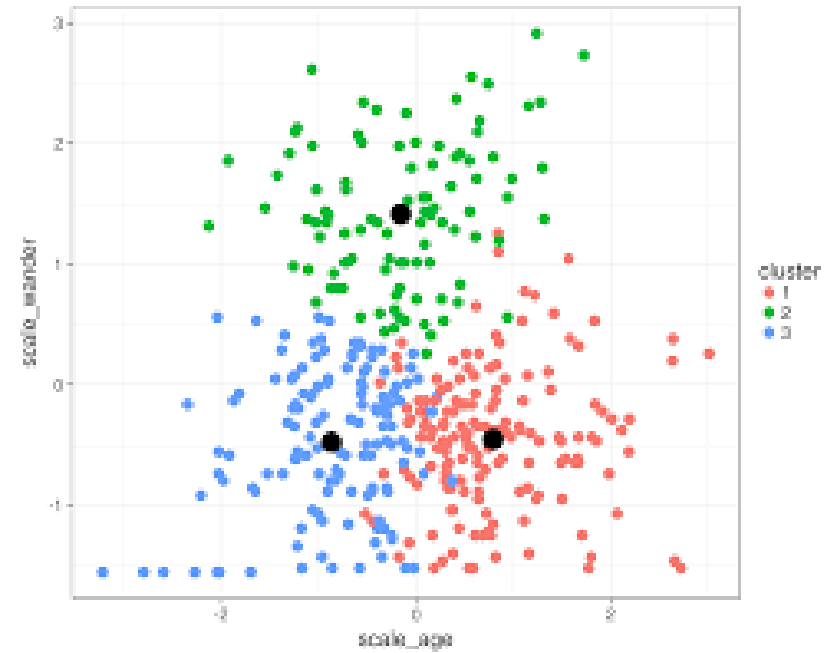
Aprendizaje no supervisado

Agrupación (Clustering)

Jerárquico

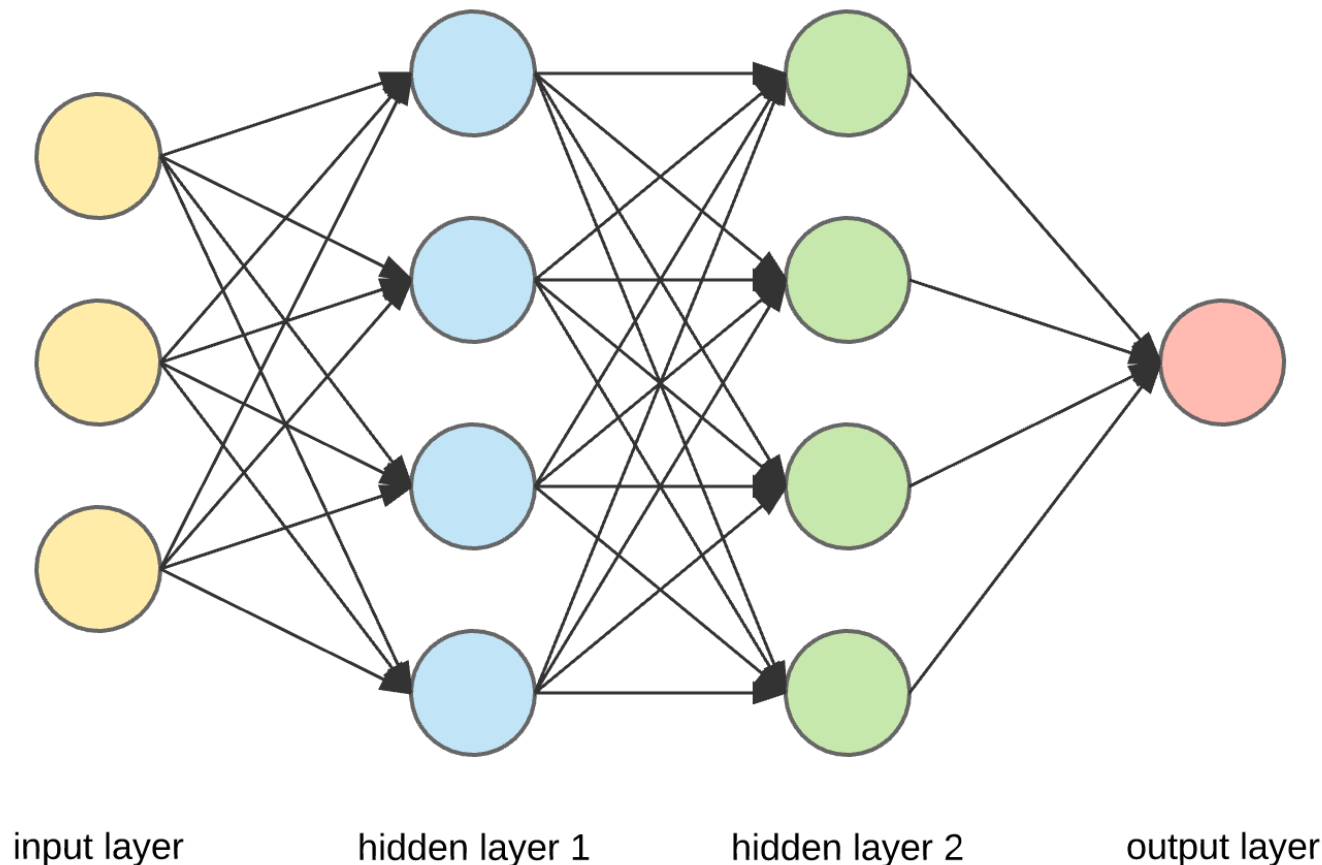


Particional



Aprendizaje profundo – Deep learning

“The goal is to create algorithms that can take in very unstructured data, like images, audio waves or text blocks (things traditionally very hard for computers to process) and predict the properties of those inputs” – Andrew Ng

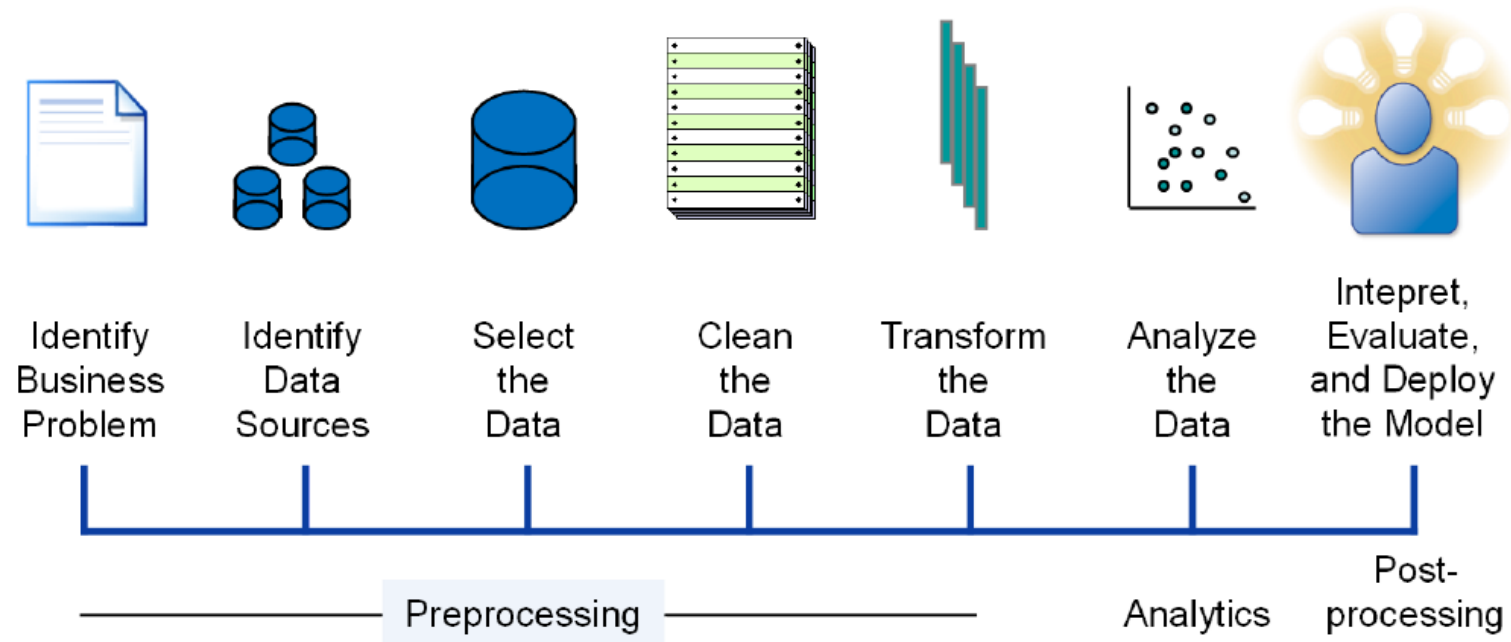


- Convolutional neural network
- Recurrent neural network
- Reinforcement learning
- Generative adversarial networks

Análisis de textos

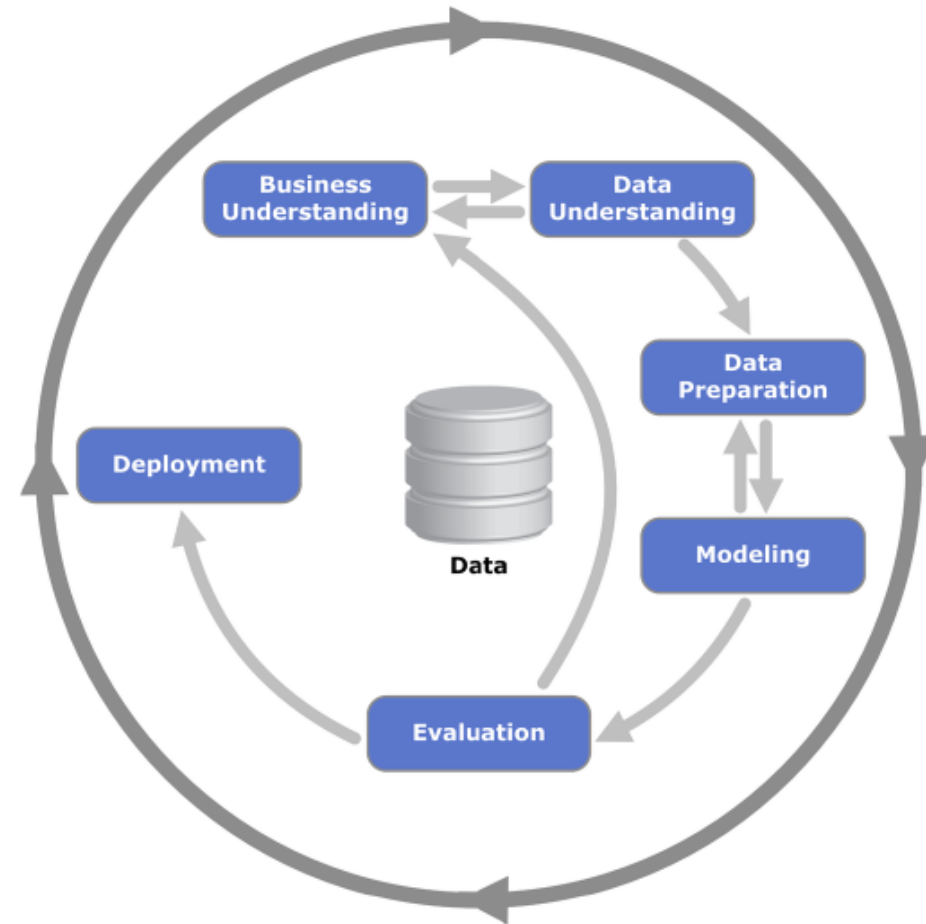


Metodología del proceso analítico



CRISP-DM

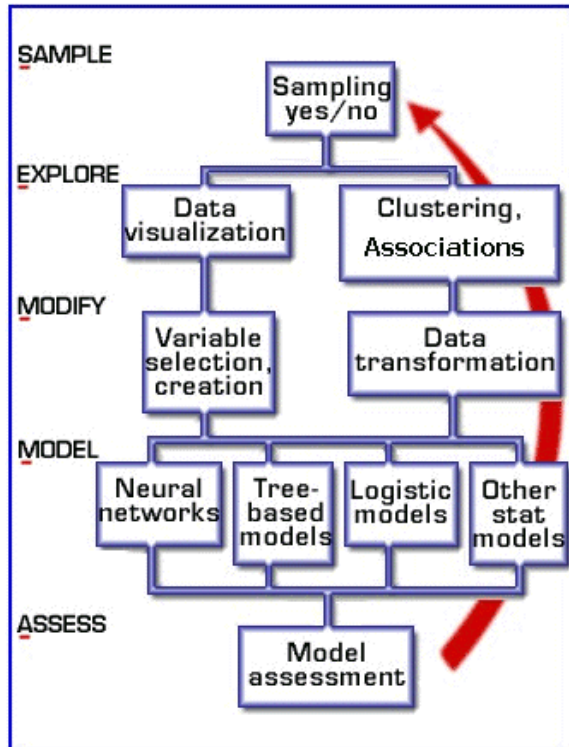
Cross Industry Standard Process for Data Mining



Otros

SEMMA

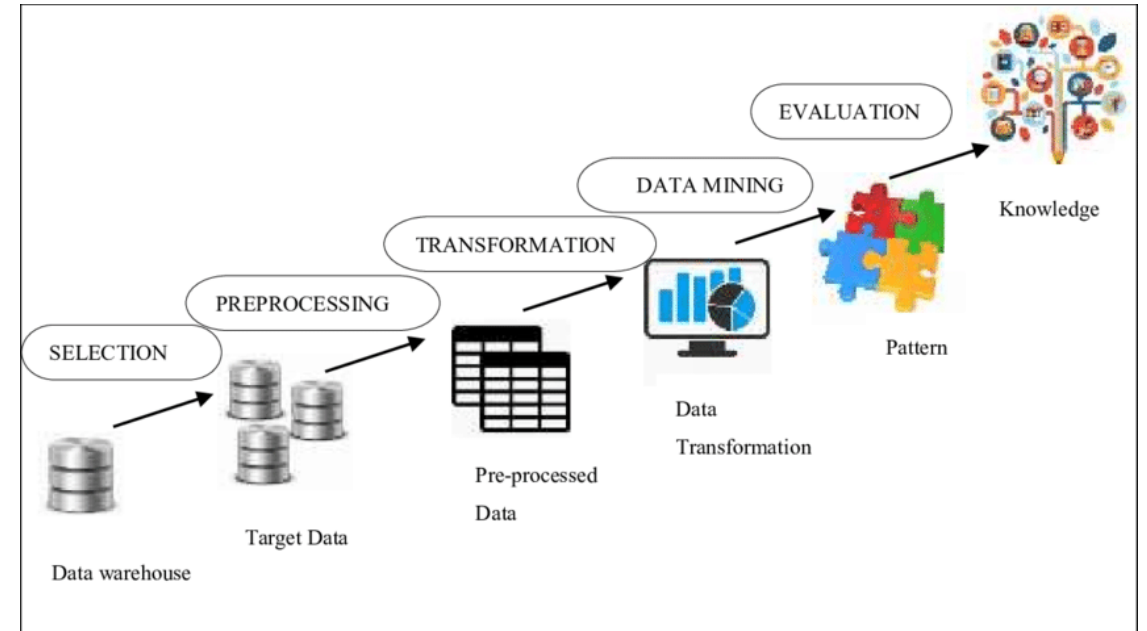
Sample, Explore, Modify, Model, Assess



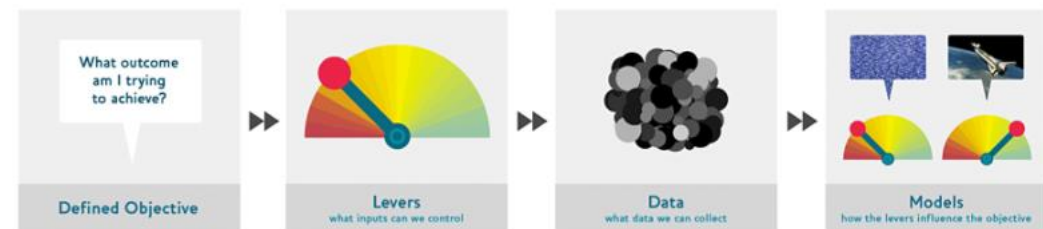
<https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jn8bbj1m1a2.htm&docsetVersion=14.3&locale=en>

KDD

Knowledge Discovery in Databases



The drivetrain approach



The four steps in the Drivetrain Approach.

<https://www.oreilly.com/radar/drivetrain-approach-data-products/>

Conocimiento/Pregunta de negocio

El análisis de los datos puede llegar a ser tan bueno como las preguntas que se hagan

Acotar. Definir objetivos y resultados claves

Redefinir el problema y/o su impacto en lenguaje de analítica

Conocimiento/Pregunta de negocio

Preguntas a hacerse al definirla:

Qué es lo que realmente quiero encontrar?

Cómo puedo medirlo? Qué métricas/variables?

Qué datos tengo?

Qué tan buenos son estos datos? (calidad, completitud)

Qué técnicas de análisis podrían llegar a solucionarlo?

Qué tanta limpieza y transformación?

Quién es mi usuario final?

Qué producto usaría este usuario final?

Qué preocupaciones hay antes de comenzar el proyecto?

Qué posibles versiones del proyecto? (MVP)

Qué tiempo tengo?

Conocimiento de los datos/ Selección de fuentes de datos

Búsqueda de fuentes de datos que podrían ayudar a ejecutar mi proyecto

Qué datos tengo disponibles?

Cuáles podría conseguir?

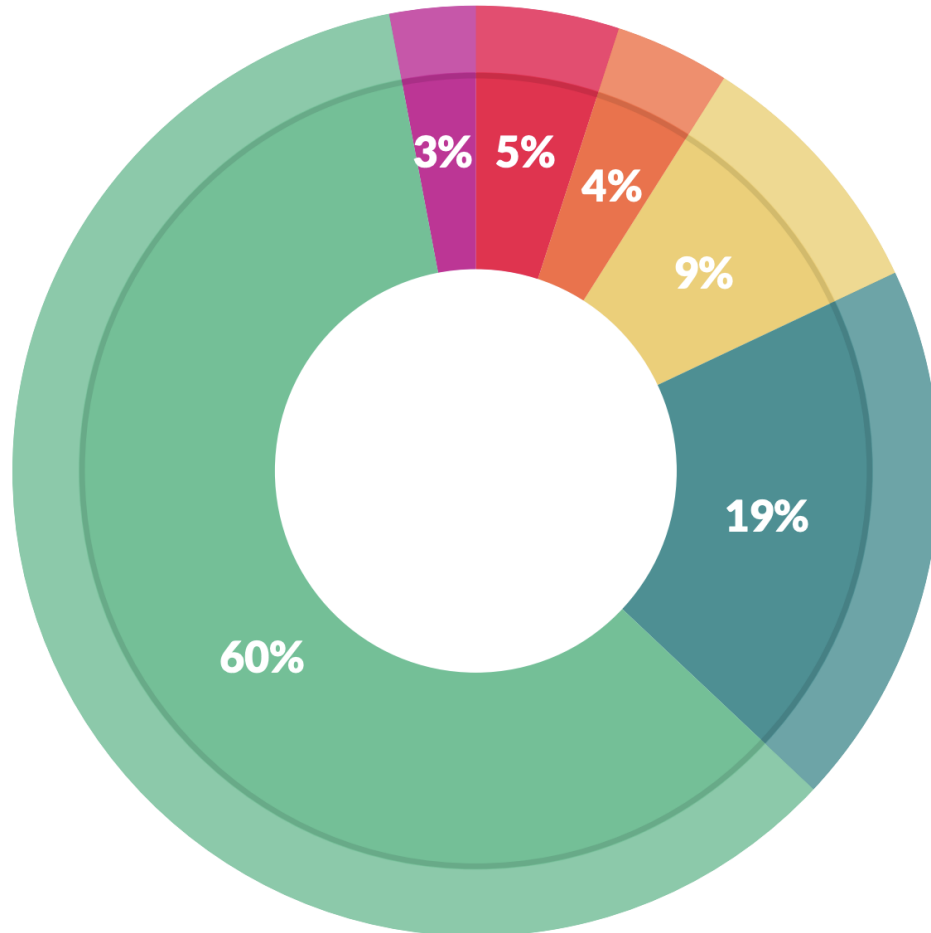
Puedo recolectar más?

Hacer una exploración inicial

Verificar y filtrar aquellas fuentes que realmente me sirven

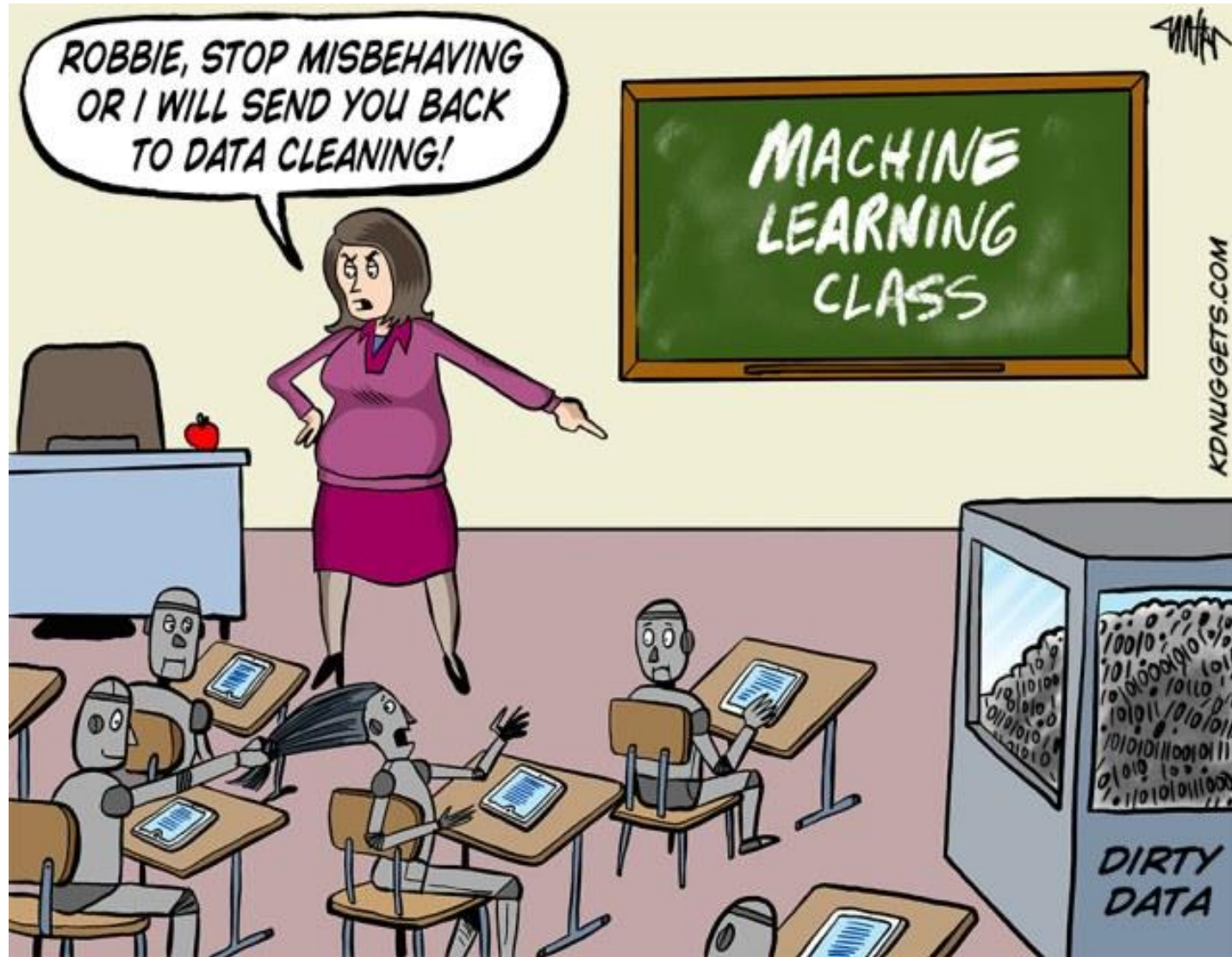
Preparación de la data

Limpieza y transformación



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%



Preparación de la data

Limpieza y transformación

- Datos actualizados
- Valores Missing
- Duplicados
- Valores atípicos
- Imputación
- Formato adecuado
- Construcción de variables
- (Re)categorización de variables

Modelamiento/Análisis

Llevar la pregunta/problema de negocio a la técnica/herramienta de análisis adecuada

No existe una solución única

Es relativo medir cuál solución es mejor que otra

No todos los problemas de ciencia de datos se resuelven con modelos

Evaluación

Comparación de soluciones, modelos, análisis

Es útil?

Es válido?

Es inesperado?

Es comprensible?

Cuál es mejor? Depende de la métrica y el contexto

Se puede mejorar?

Implementación

Definir la manera como nuestro cliente final puede acceder a nuestra solución

Una tabla

Informe de resultados

Dashboard

Debo alimentar mi solución con datos nuevos

Monitoreo del desempeño de modelos. Re-entrenamiento

10 formas en que un Proyecto de CD puede fallar

<http://www.martingoodson.com/ten-ways-your-data-project-is-going-to-fail/>

1. La data no está lista. Si los datos no han sido utilizados antes, añadir a los tiempos algo extra para la limpieza. Hacer una exploración previa antes de definir cronograma
2. Alguien escuchó que 'Data is the new oil'. Los datos por sí solos no suelen ser una mercancía, necesitan ser transformados y analizados para sacar su valor
3. El científico de datos está por renunciar...
Todas las partes de la cadena del Proyecto deben apropiarse en CD

Phd after years of data science research: I still have a a lot to learn



Data science enthusiast after finishing 3 certificates online



10 formas en que un Proyecto de CD puede fallar

<http://www.martingoodson.com/ten-ways-your-data-project-is-going-to-fail/>

4. No tiene un líder de ciencia de datos (no todos los científicos de datos saben lo que hacen)

Saber funciones de Python no te hace un Data Scientist

5. Tampoco se deben contratar científicos.

Si es limpieza y transformación, un data engineer
Si es hacer reportes, un BI analyst

6. Tu jefe leyó un blog sobre machine learning

Phd after years of data science research: I still have a a lot to learn



Data science enthusiast after finishing 3 certificates online



If God exists it's fucking me

10 formas en que un Proyecto de CD puede fallar

<http://www.martingoodson.com/ten-ways-your-data-project-is-going-to-fail/>

7. El modelo es demasiado complejo

Empezar con un modelo simple que pueda entender. Luego intentar algo más complejo, solo si es necesario

8. Tus resultados no son reproducibles

9. R&D (research and development) – Innovación no hace parte de la cultura de la empresa

10. Diseñar soluciones sin ver datos en la vida real

Las preocupaciones más grandes siempre estarán alrededor de la data!

Requisitos del modelo analítico

Relevancia para el negocio. Resolver un problema en particular

Desempeño estadístico. Significancia estadística, precisión de modelos, calidad en las predicciones del modelo

Interpretabilidad y justificación. Es subjetivo pero crucial! (depende mucho del tomador de decisiones). A veces necesario de balancear con el desempeño estadístico

Eficiencia operacional. Cómo los resultados de los modelos pueden ser integrados a las soluciones del negocio?

Costos económicos. Cuál es el costo de recolectar la data para ajustar y evaluar los modelos. Vale la pena comprar data extra y/o modelos

Cumplimiento de la normativa. De acuerdo con regulaciones y normas

Recuerden: queremos modelos que sean válidos (generalizables), útiles (accionables), inesperados (interesantes) y comprensibles

Conclusiones

- Entender los conceptos principales relacionados con ciencia de datos
- Introducción a las principales técnicas de análisis de datos
- Procesos y etapas para el desarrollo de proyectos de ciencia de datos
- Aspectos importantes para tener un proyecto de CD exitoso