Módulo 1. Introducción a la ciencia de datos

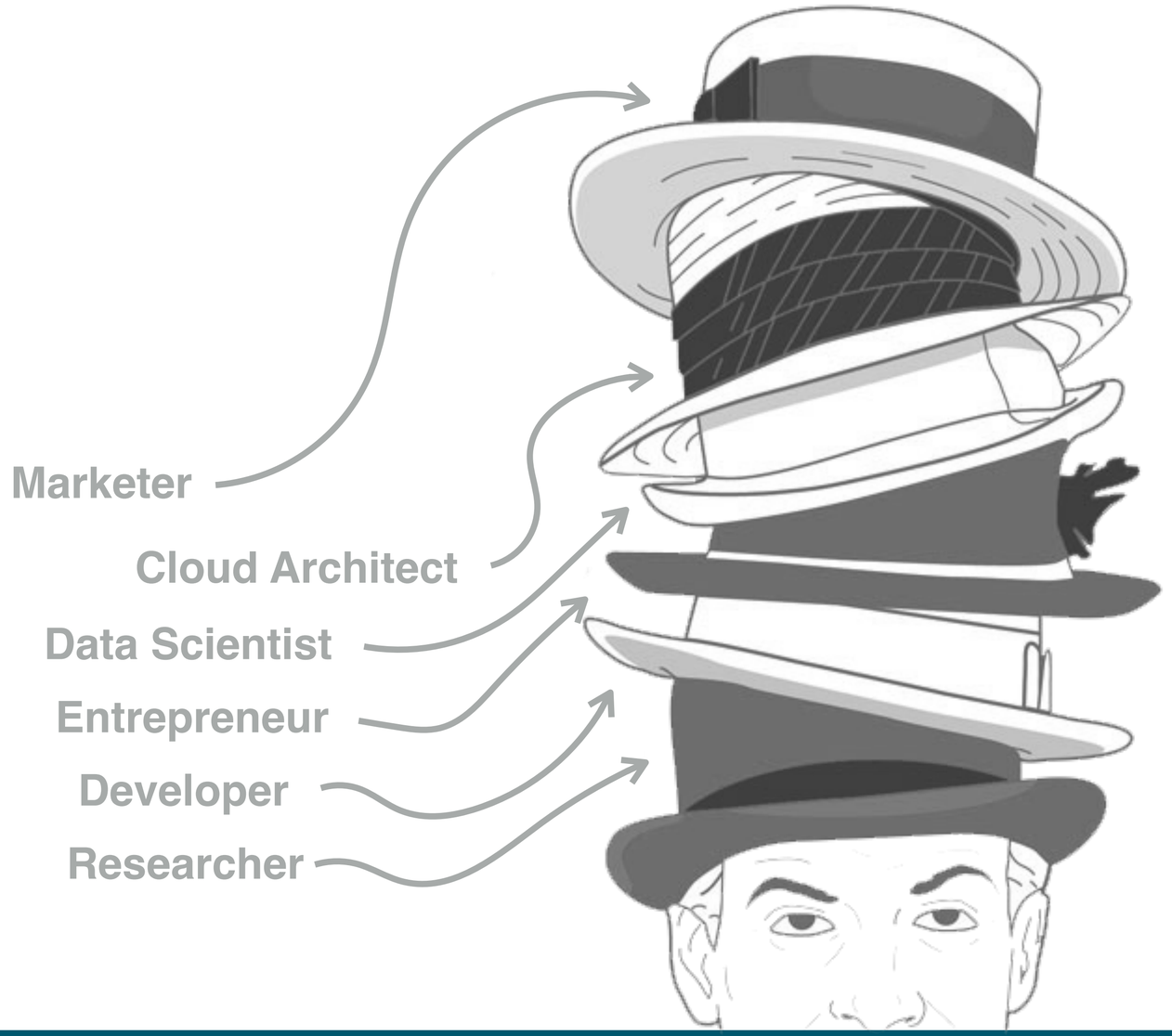
Introducción a la Ciencia de Datos e Ingeniería de Datos

Jacinto Arias

Máster en Ciencia de Datos e Ingeniería de Datos en la Nube

Sobre mi

- Informático
- Científico
- Desarrollador
- Emprendedor



Taidy

Visit taidy.cloud

in/jacintoarias


DATA SCIENCE
DETECTION
BIG DATA
COMPUTING
INFORMATION TECHNOLOGY
SOCIAL MEDIA
PROMOTION
WWW
E-MARKETING
COMMUNICATION
MODELS
SERVICES
PROCESSING
BRANDING
COMPUTER
MULTIMEDIA NETWORK
PROJECTS
CONTENT
CONSUMER
CONSUMER
DEMAND
MARKETS
PREDICTIVE
ORGANIZATION
PLANNING
CODING
SOFTWARE
WEB MARKETING
PROGRAM
ANALYTICS
EVENTS
PROGRAMMING
DATA MINING
MACHINE LEARNING
VISION
ENGINEERING
RESEARCH
PROBABILITY
INFORMATION
WEB SERVICES
BIG DATA
KDD
WEB DEV
COMPUTING
MATHS
PATTERN
PROJECTS
SERVICE
STRATEGY
WORLDWIDE
BIG DATA
ENGINEERING
VISUALIZATION
NETWORK
SOLUTIONS
PLANNING
MEDIA
MOBILE
PRO
CODING
PRICING
SOCIAL
SEGMENTATION
STATISTICS
INFORMATION
DIGITAL
TARGET
SOCIAL NETWORKS


BIG DATA
GOVERNMENT
TECHNOLOGIES
PROCESS
VOLUME
SOCIAL
SAN
NAS
SOFTWARE
MANAGEMENT
DATABASES
PROCESSING
SIZE
EXABYTES
PETABYTES
ANALYSIS
RESEARCH
LARGE
INFORMATION
STORAGE
LEARNING
INTERNET
MILLION
SETS
DEVELOPMENT
TECHNOLOGY
SEARCH
WORLD
BASED
SCIENCE
FUTURE
BILLION
NEW
PARALLEL
INITIATIVE
TERABYTES
PAST
TIME

“I keep saying that the sexy job in the next ten years will be statisticians. And I’m not kidding.”

Hal Varian (2009)
Chief Economist @ Google

Buzzwords


DATA SCIENCE
DATA SCIENTIST
DATA ENGINEER
MACHINE LEARNING
DEEP LEARNING
BIG DATA

¿Qué es un Científico de datos?

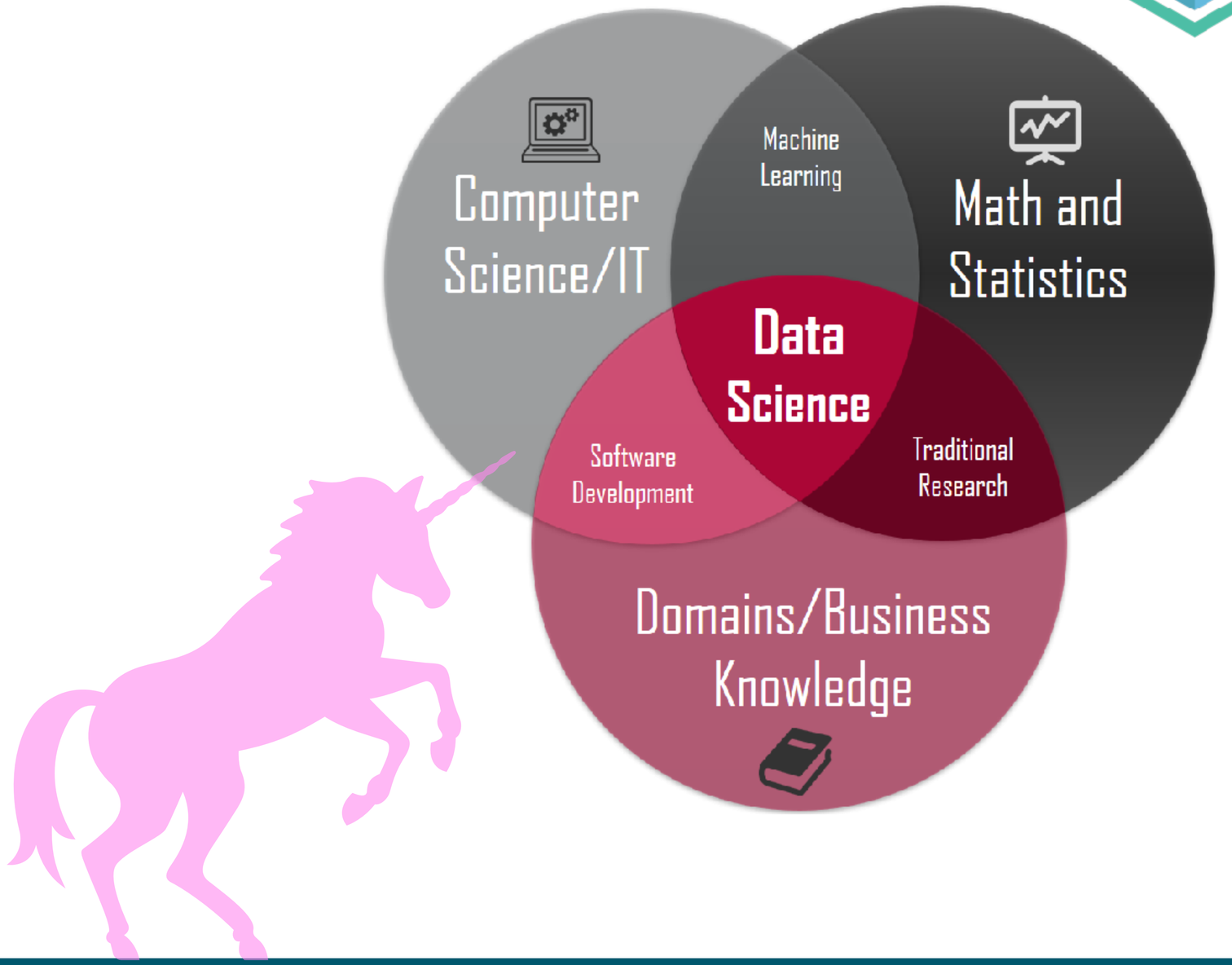
An Statistician who lives in San Francisco

Statistics on a Mac

Someone who is better at statistics than a software engineer, and better at coding than any statistician

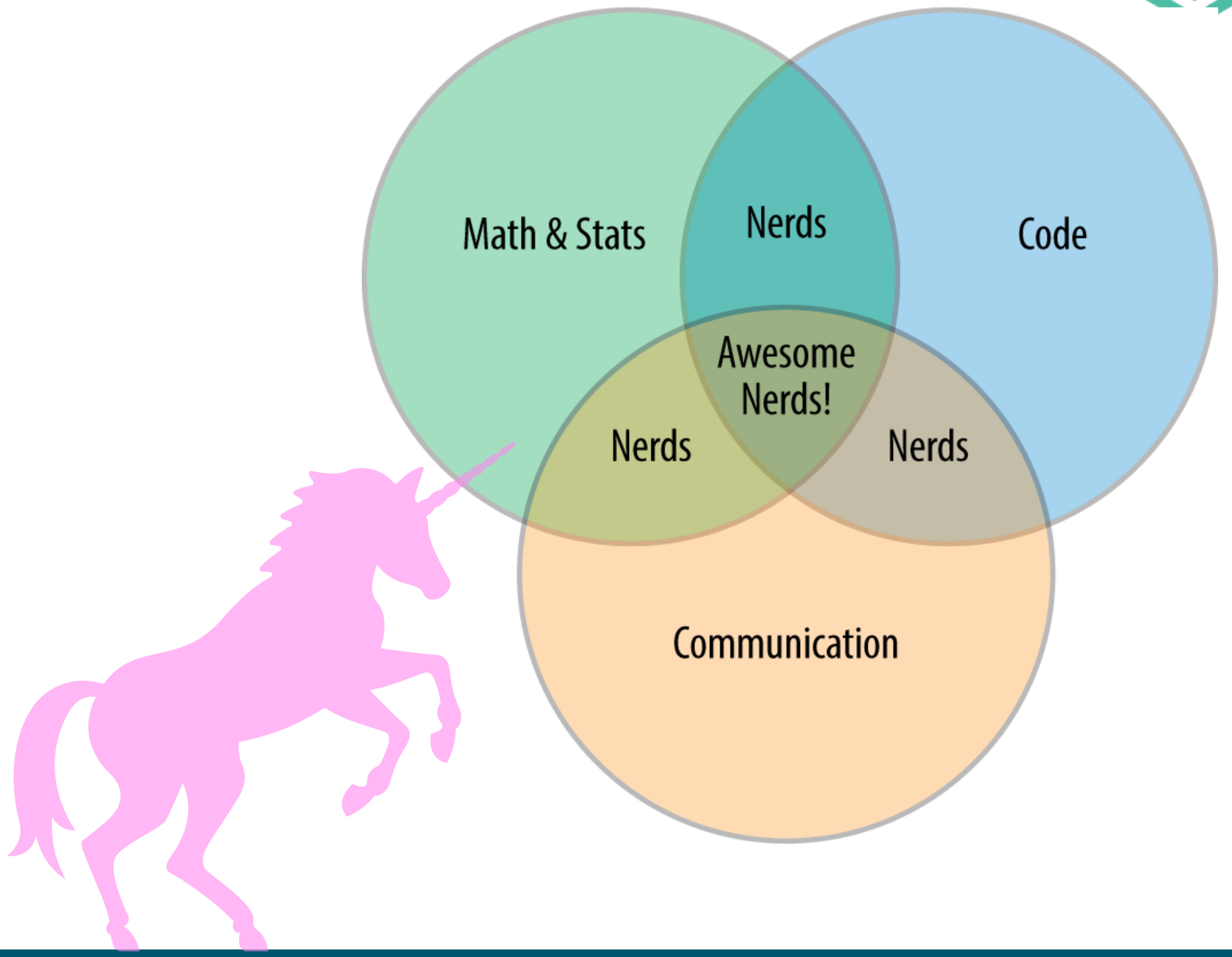
¿Qué es la Ciencia de Datos?

- Acuñado en 2008 (Patil@LinkedIn, Hammerbacher@Facebook)
- “Data Science is what a Data Scientist does”
- “Data science is the discipline of making data useful.”
-



¿Qué es la Ciencia de Datos?

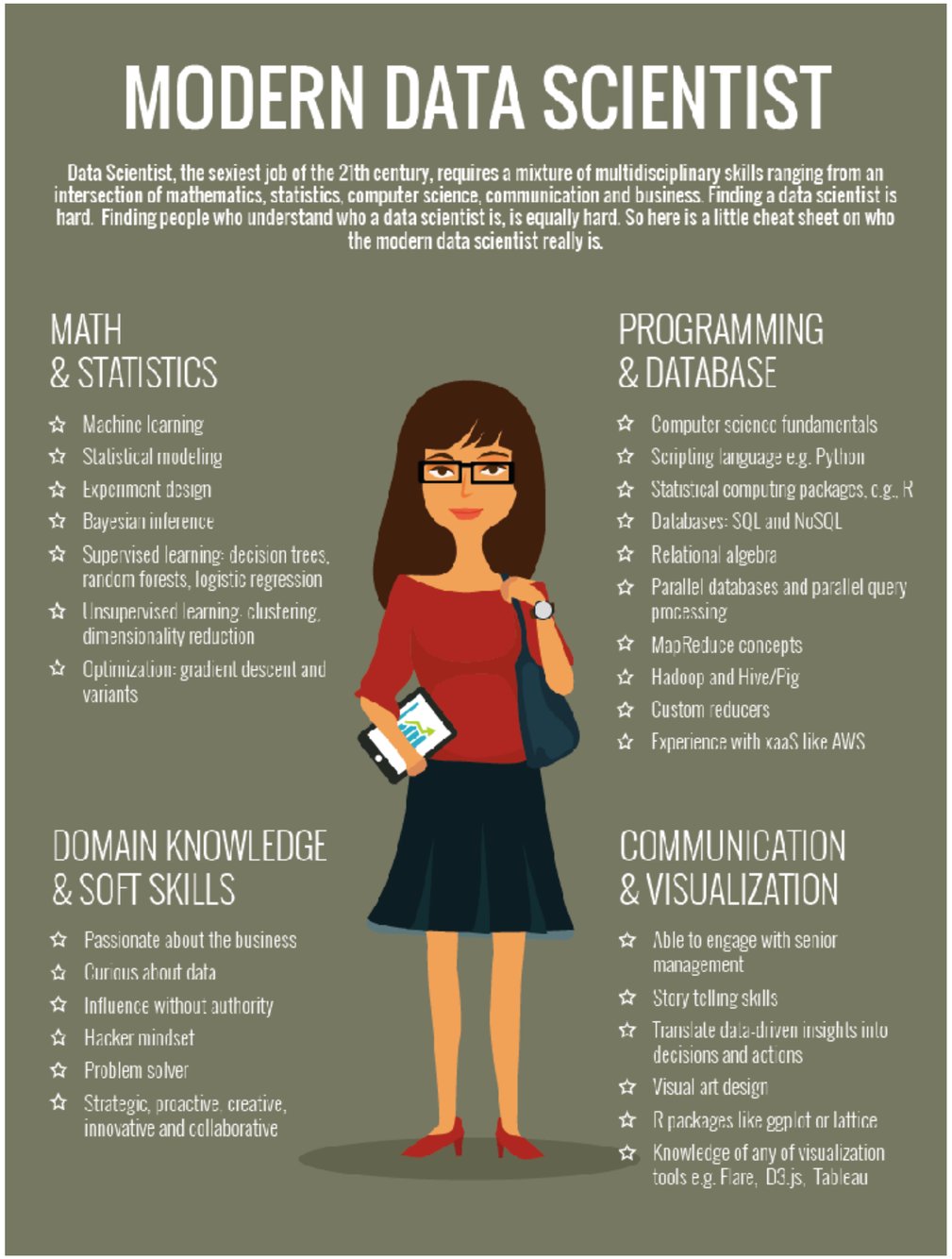
- Acuñado en 2008 (Patil@LinkedIn, Hammerbacher@Facebook)
- “Data Science is what a Data Scientist does”
- “Data science is the discipline of making data useful.”
-



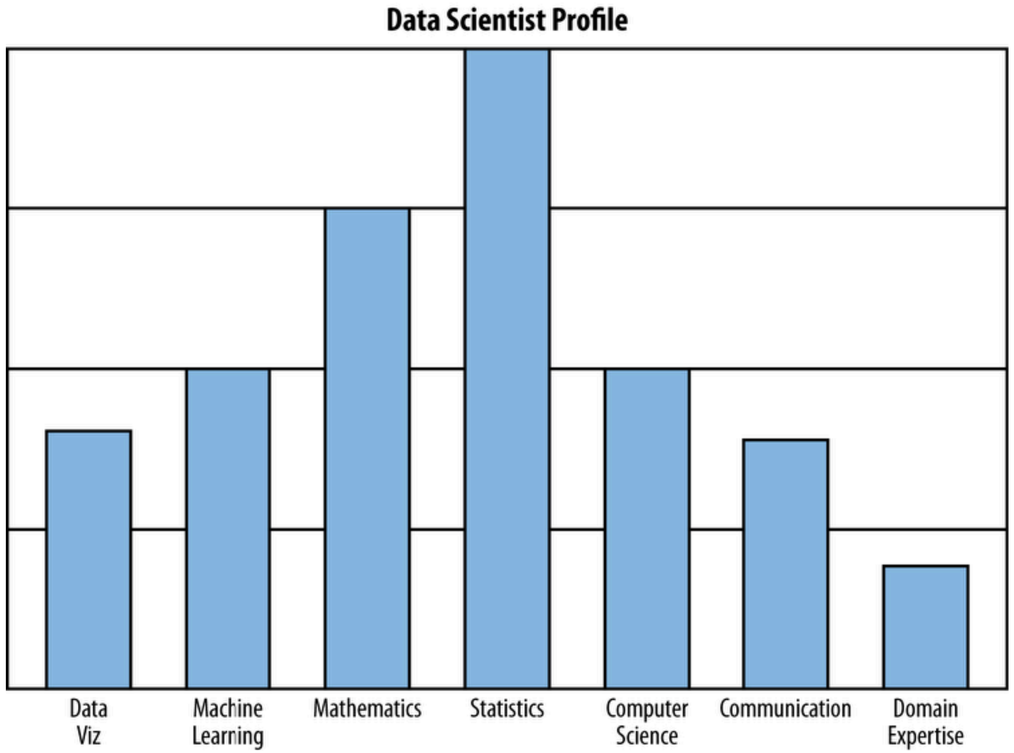
¿Quién es un Científico de Datos?

- Suele utilizarse como título de un puesto de trabajo
- Acuñado en 2008 (Patil@LinkedIn, Hammerbacher@Facebook)

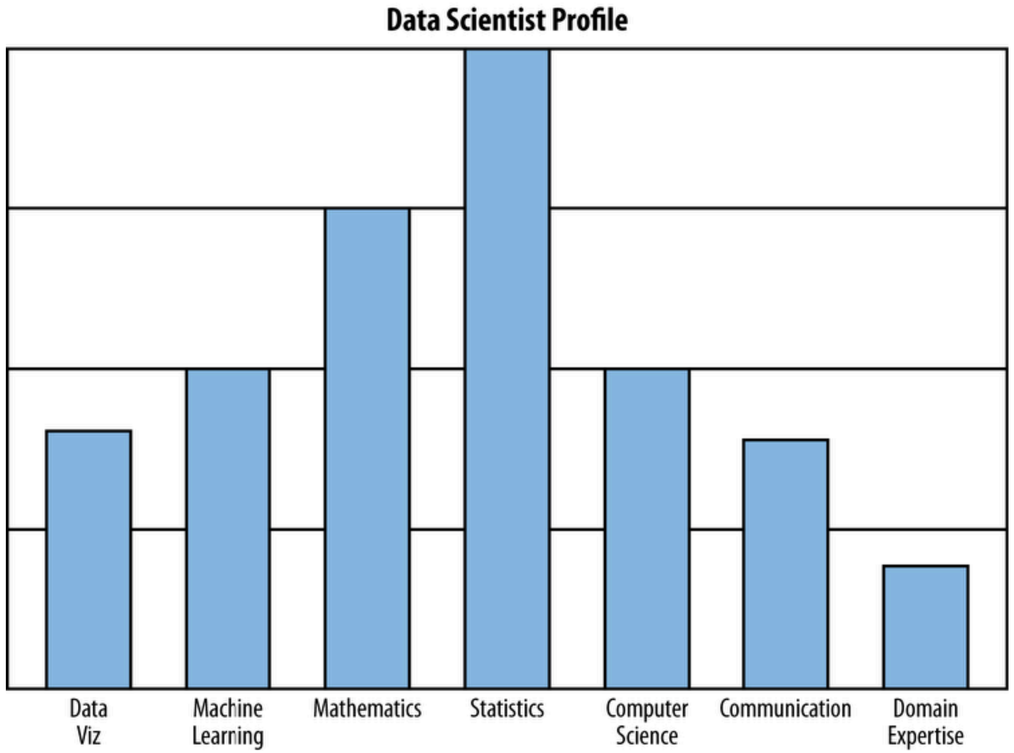
Es casi la descripción de un departamento entero

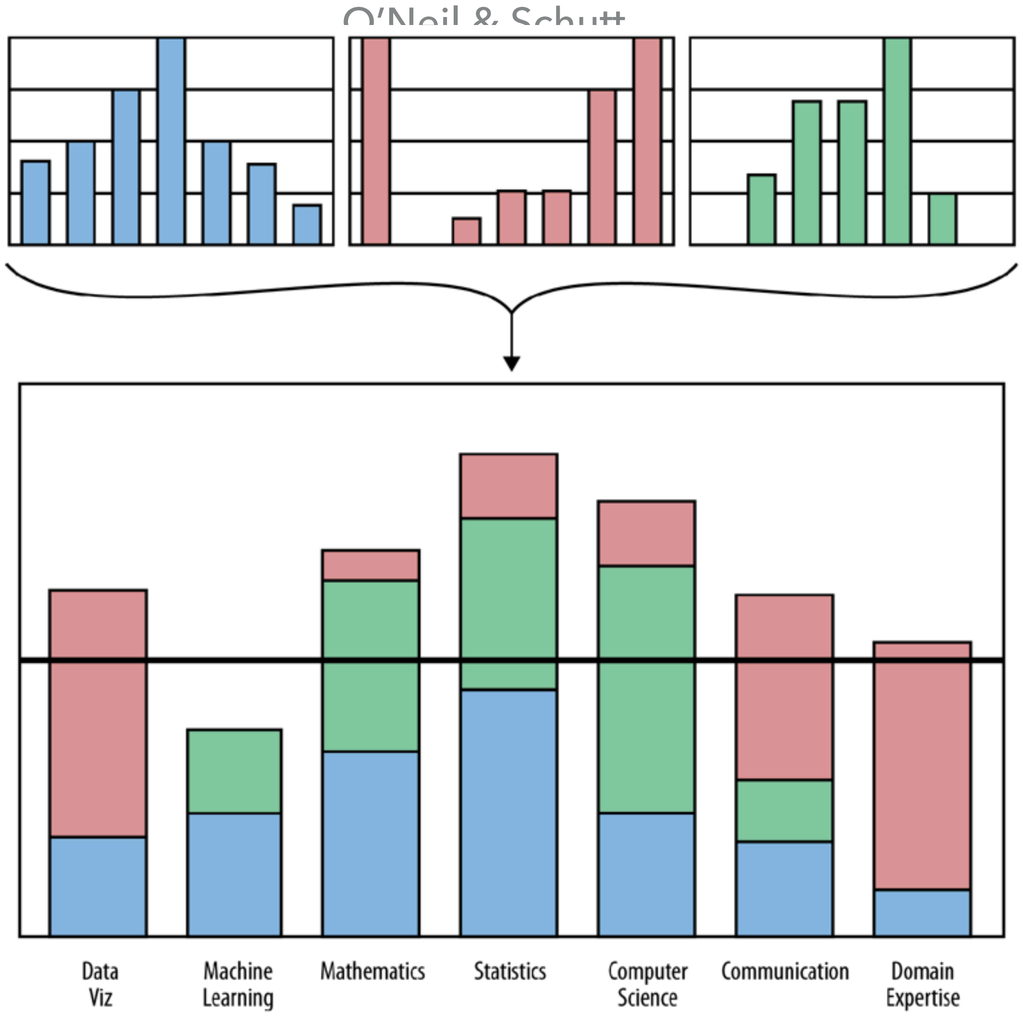


El Perfil de un Data Scientist



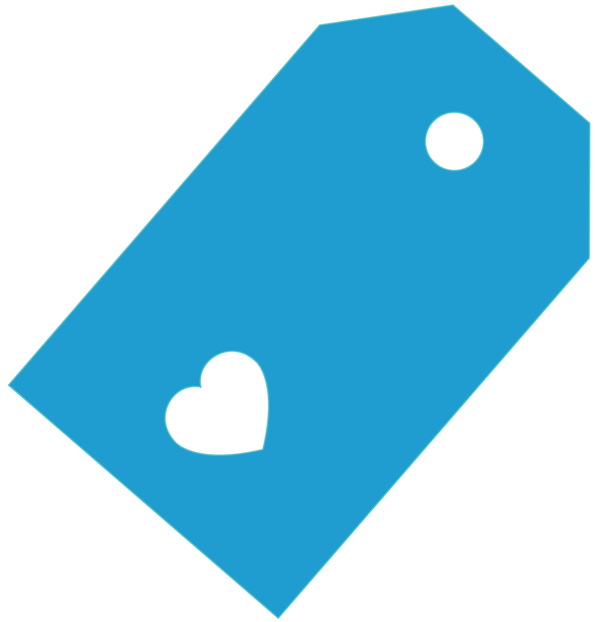
El Perfil de un Equipo de Data Science





Perdón, Data… ¿Qué?

- Data Analyst
- Data Scientist
- Data Engineer
- AI / ML Engineer
- Data Translator
- Citizen Data Scientist
- …



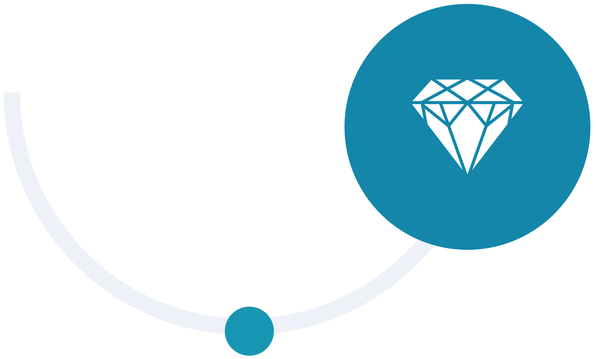
“Making Data Useful”

Datos

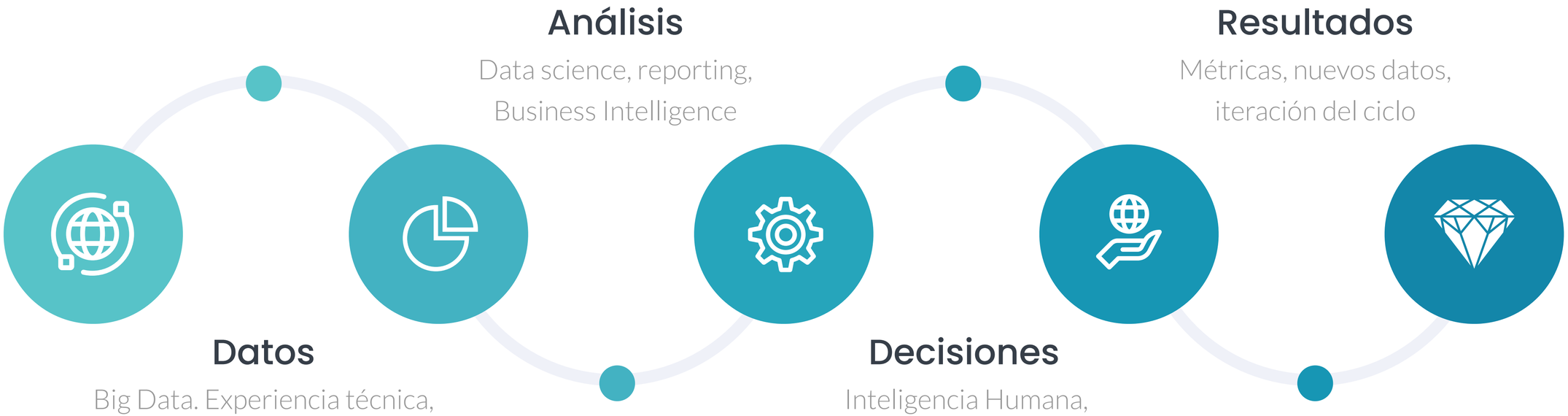
Big Data. Experiencia técnica, almacenamiento, modelado

Resultados

Métricas, nuevos datos, iteración del ciclo



“Making Data Useful”



¿Para qué?

Hacer que los datos tengan utilidad

¿Cómo?

¿Para qué?

Para tomar decisiones…

“It’s through our decisions that we affect the world around us.”

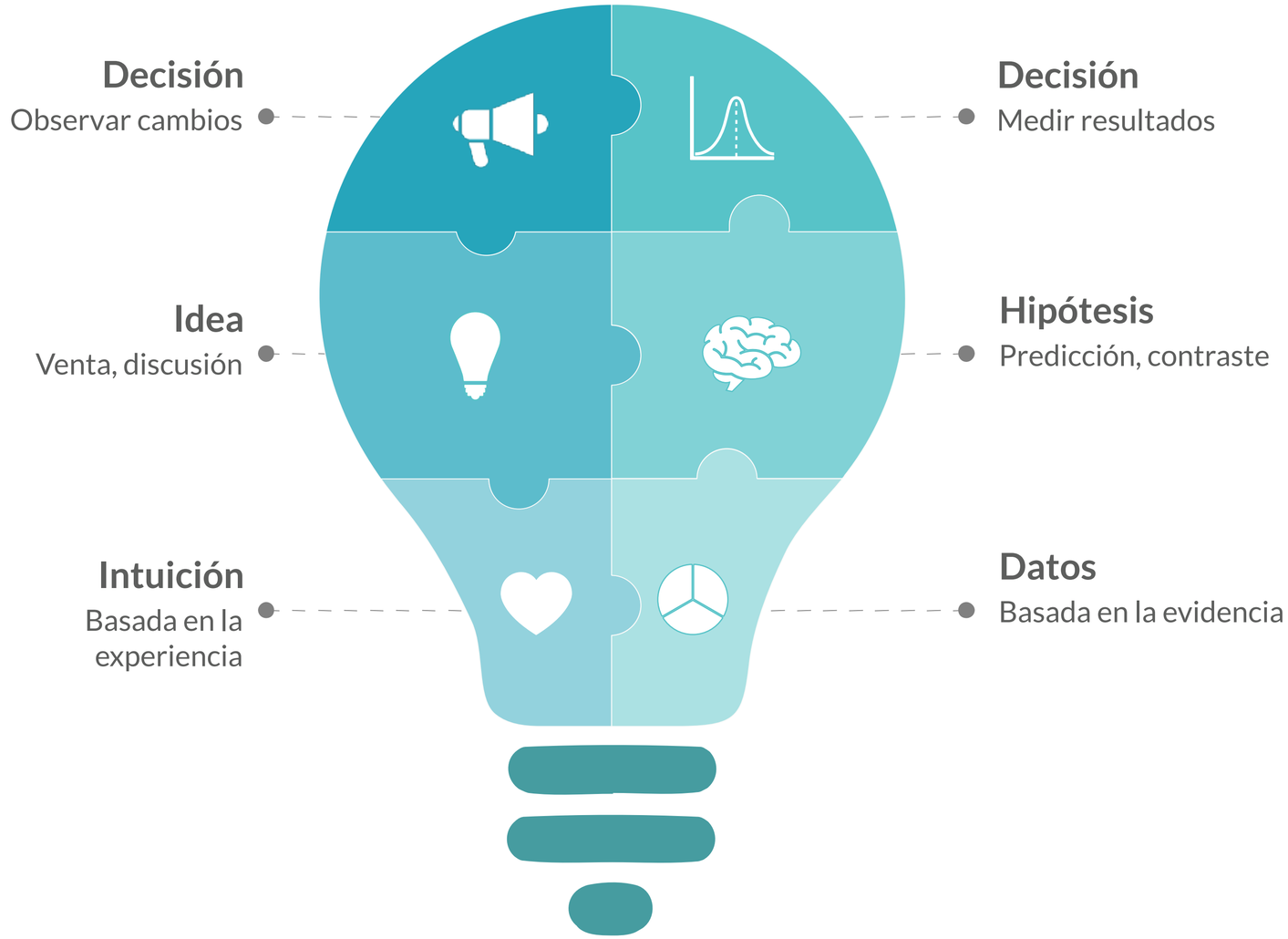
¿Para qué?

Para tomar decisiones…
…basadas en datos

¿Podemos hacerlo de otra manera?

Antipatrón: The Hippo

Un cambio de paradigma



“In God we trust the rest bring data.”

W. EDWARDS DEMING (~1970)
Engineer and Scientist

“If we have data, let's look at the data. If all we have are opinions, let's go with mine.”

JIM BARKSDALE (1998)
CEO @ Netscape

“Many people only use data to feel better about decisions they’ve already made.”

CASSIE KOZYRKOV (2018)
Chief Decision Maker @ Google

¿Para qué?

Hacer que los datos tengan utilidad

El Ecosistema

Personas

Procesos

Tecnología

El Ecosistema

Personas

Procesos

Tecnología

¿Cuál es la diferencia con…?

- Estadísticos
- Analístas de negocio
- Desarrolladores de software
- Administradores de Bases de Datos
- Administradores de sistemas
- …

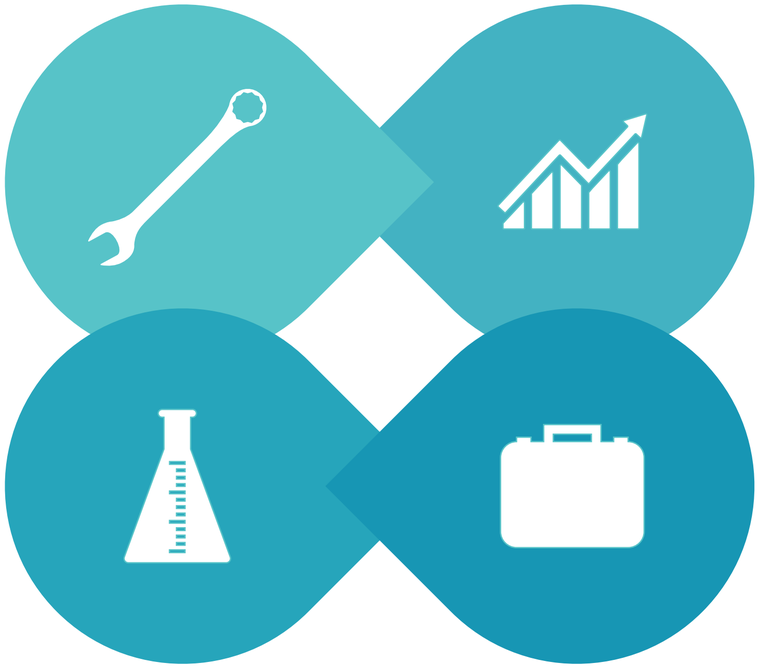
Combinar y adaptar

Data Engineers

Expertos en tecnología y en datos

Data Scientists

Machine Learning y curiosidad



Business Analyst

Traducir e interpretar los resultados

Decisión Makers

Liderazgo basado en datos

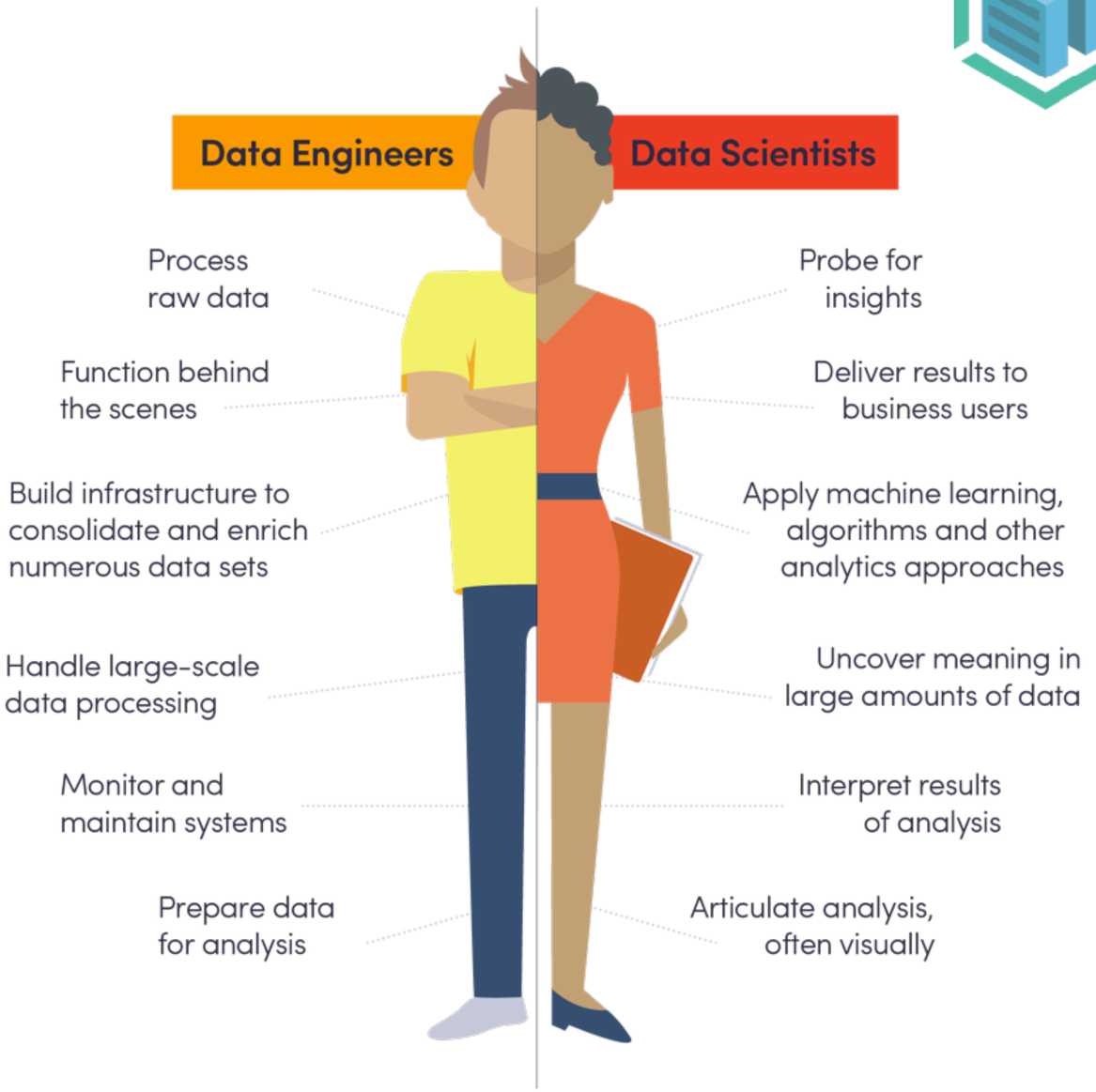
Data What?

Data Engineer

Al principio de la cadena de valor

Data Scientist

Más cerca del valor final



El Ecosistema

Personas

Procesos

Tecnología

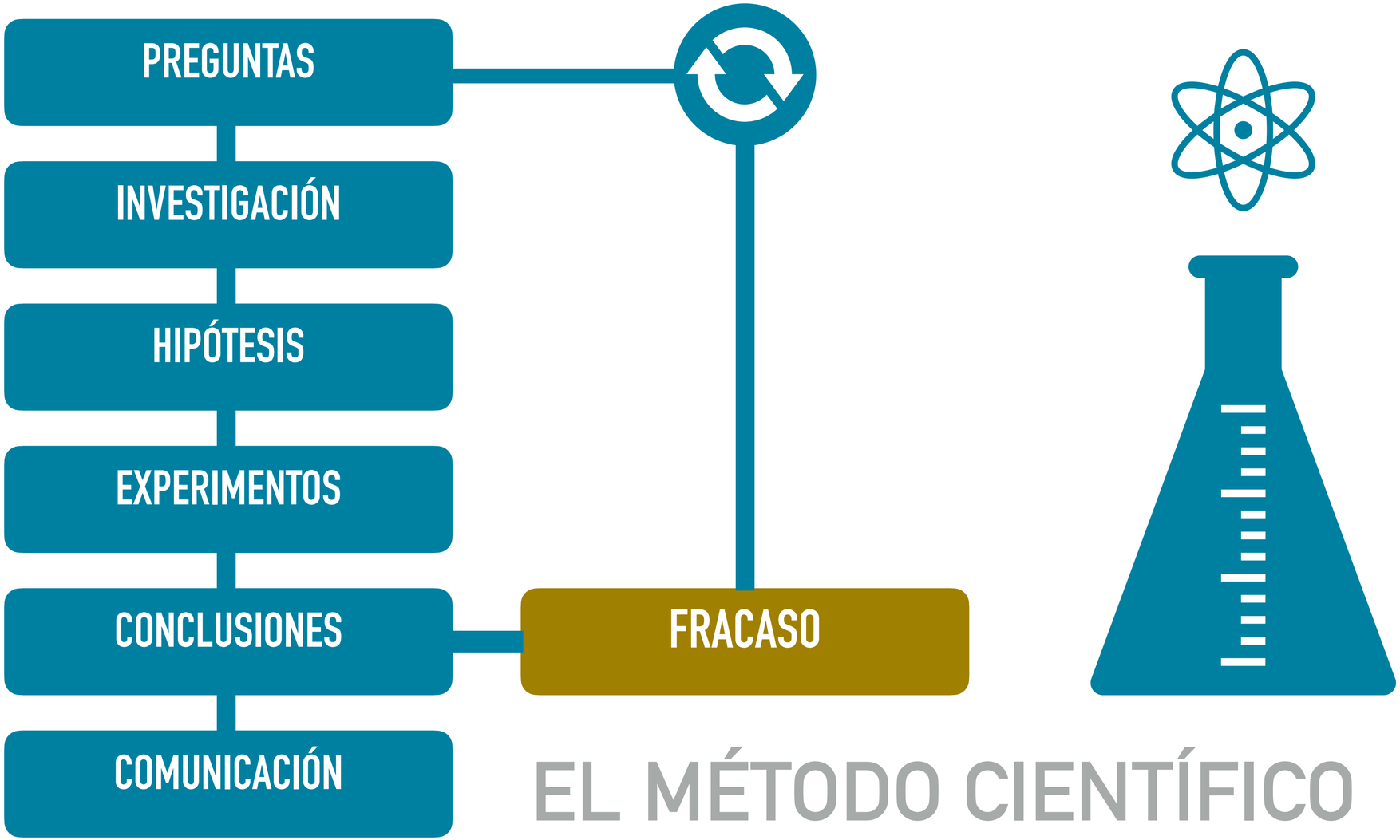
El ciclo de valor de los datos

La ciencia de datos es un proceso de principio a fin

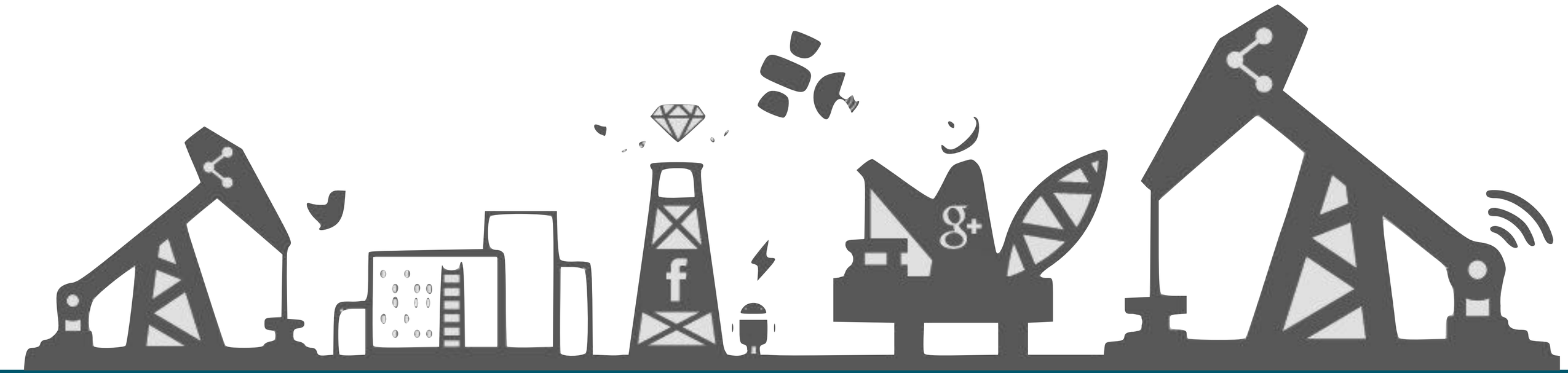
- Orientado a producto/cliente: Validado en escenarios reales
- Iterativo: Experimental y orgánico, con mejoras y adaptaciones continuas
- Data Driven: Basado en el estado de los datos

Metodologías Data Science: CRISP-DM

La “Ciencia” de la Ciencia de Datos



La Datificación (Transformación Digital) de las empresas

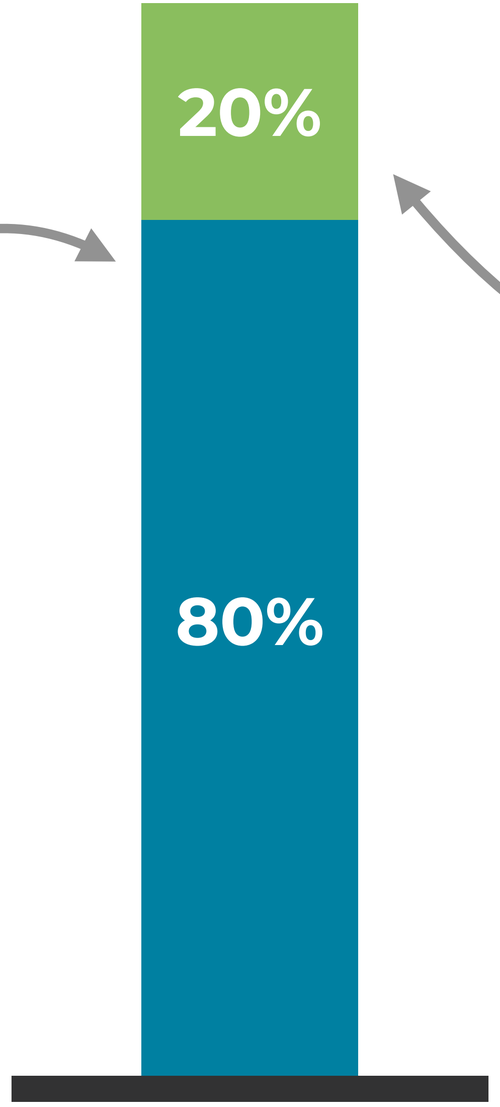


El ciclo de vida de los datos

- Extracción
- Preprocesamiento
- Almacenamiento
- Exploración
- Modelado
- Visualización

El “Trabajo Sucio”

Data Wrangling
Data Munging
Data Pre-processing
Data Tidying
Data Cleaning
Data Cleansing



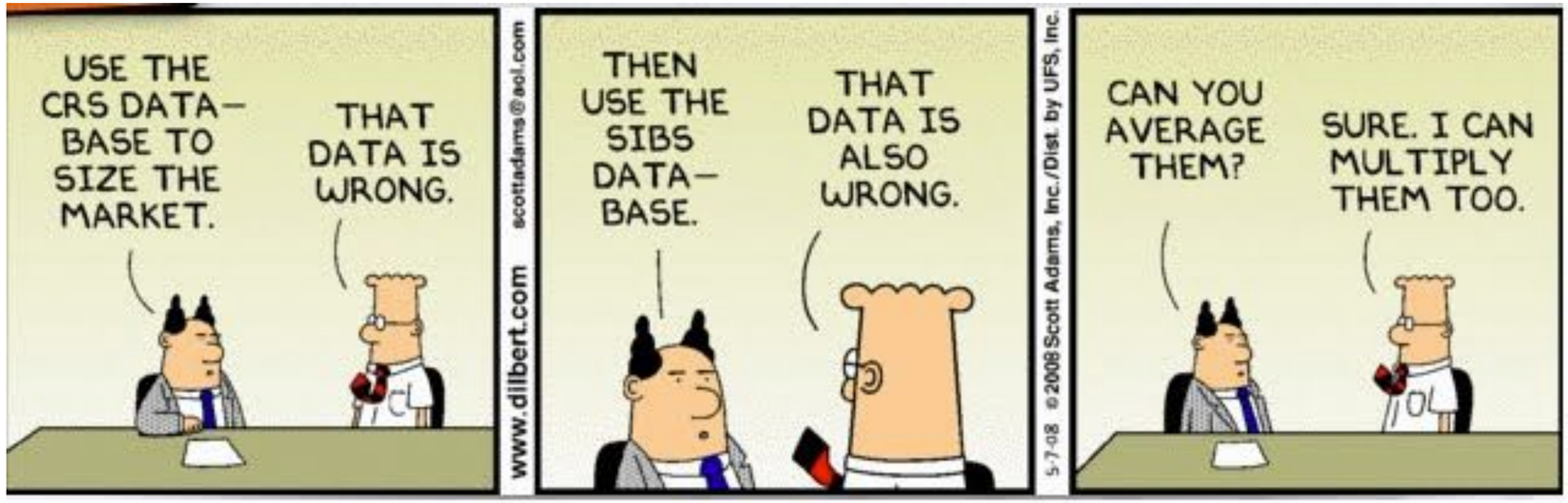
Modelling
Exploration
Learning
Inference
Prediction
Insights

Arquitecturas de datos

- El 80% de lo consideramos como análisis de datos no lo es realmente
- Pasamos la mayor parte del tiempo transformando y trasladando los datos
- Un formato o un sistema nunca es óptimo para todos los procesos
- Estos procesos requieren distintos perfiles en cada etapa

Calidad del dato

- Los fallos de calidad del datos cuestan aproximadamente $600 mil millones en EEUU anualmente
- “Better data will always beat better models”
- “Garbage in = Garbage Out”

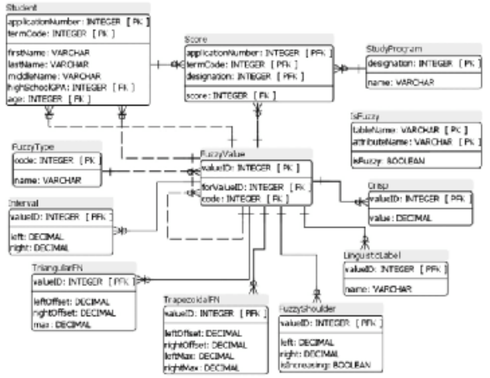


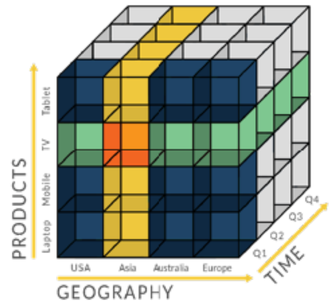
Limpieza/Preprocesado/Transformación

- Responsabilidad compartida en la organización:
  - Problemas de captura
  - Fallos de Hardware
  - Bugs en el Software
  - Errores humanos en la entrada
  - Errores experimentales

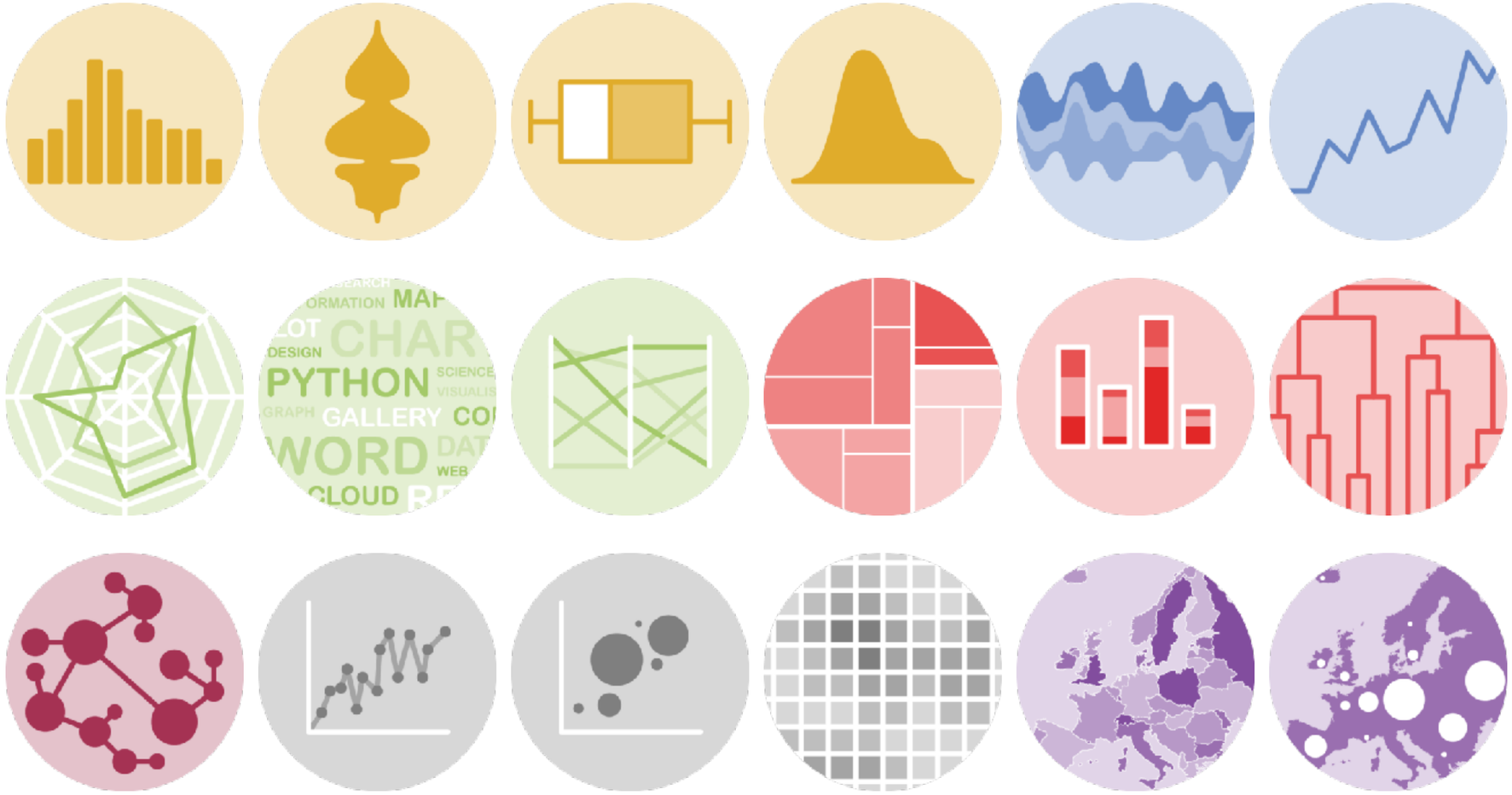
Almacenamiento

- Los datos se necesitan en distintos formatos
- La complejidad de las aplicaciones impone tecnologías concretas (SQL vs NoSQL)
- Distintas tareas requieren consultas y modelos de datos diferentes, ej: Machine Learning vs Business Intellgence





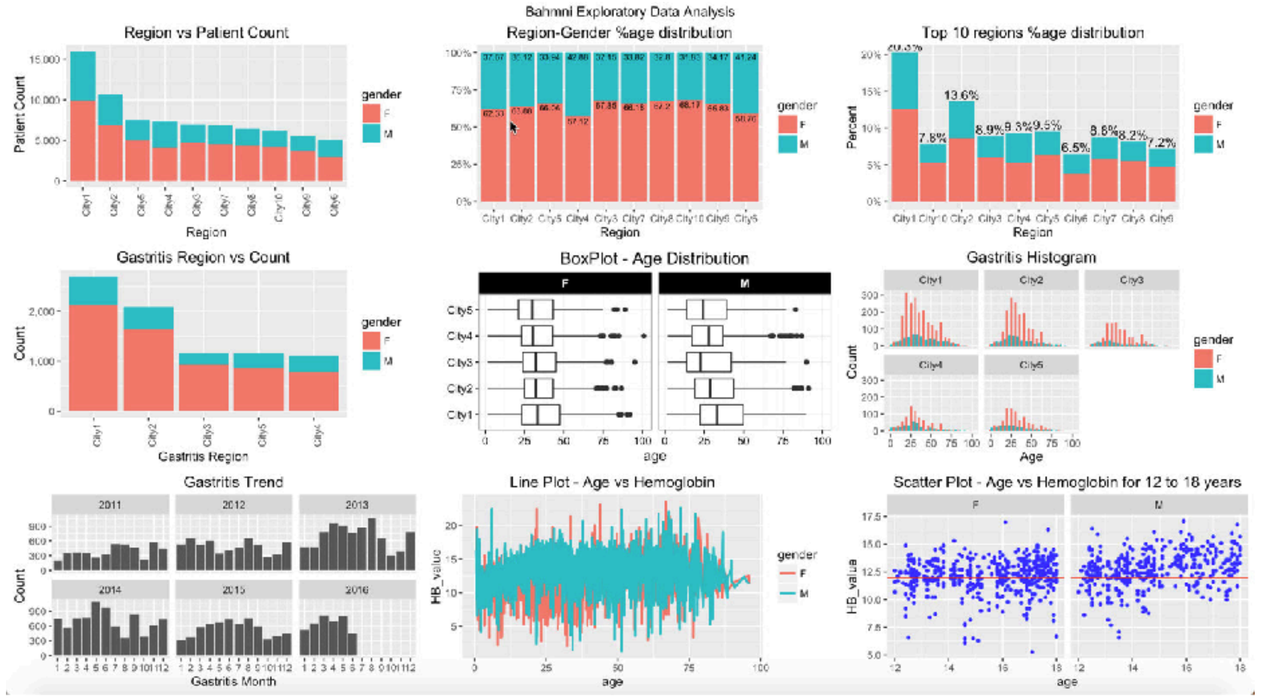
Visualización de datos



"Don't trust everything you see, even salt looks like sugar"

Análisis exploratorio (EDA)

- Realizado desde en etapas iniciales del proceso
- Aporta familiaridad y detecta problemas
- Centrado en conceptos estadísticos



Reporting y Business Intelligence

- Realizado en las etapas finales del proyecto
- Centrado en el valor final del datos en la monitorización y control de objetivos de negocio
- Realizado por expertos de negocio menos técnicos, utilizando herramientas especializadas

Modelado e Inteligencia

- El mundo real es complejo, incierto y genera un flujo continuo de datos
- Podemos capturar una muestra de los datos generados
- Pretendemos explicar el proceso original usando esa muestra
- Construiremos un modelo que abstraiga el proceso tras los datos

Proceso Generativo

Muestra de Datos

Algoritmo

Modelo

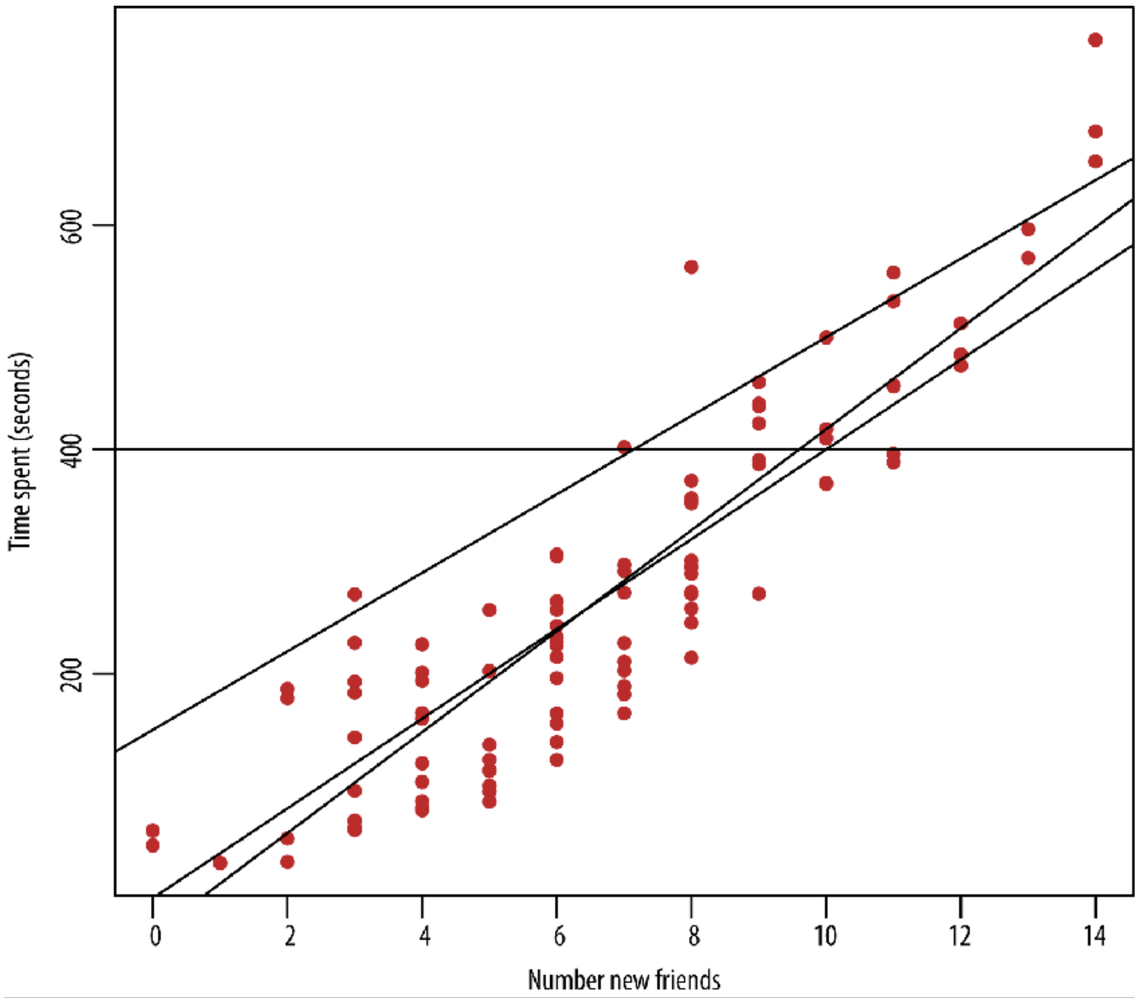
```
f(x)
```

Modelado Estadístico

- Ajustar un modelo dada una muestra de datos

$$y = \alpha + x\beta$$

- Aprender los parámetros que mejor se ajusten (que menos se equivoquen)

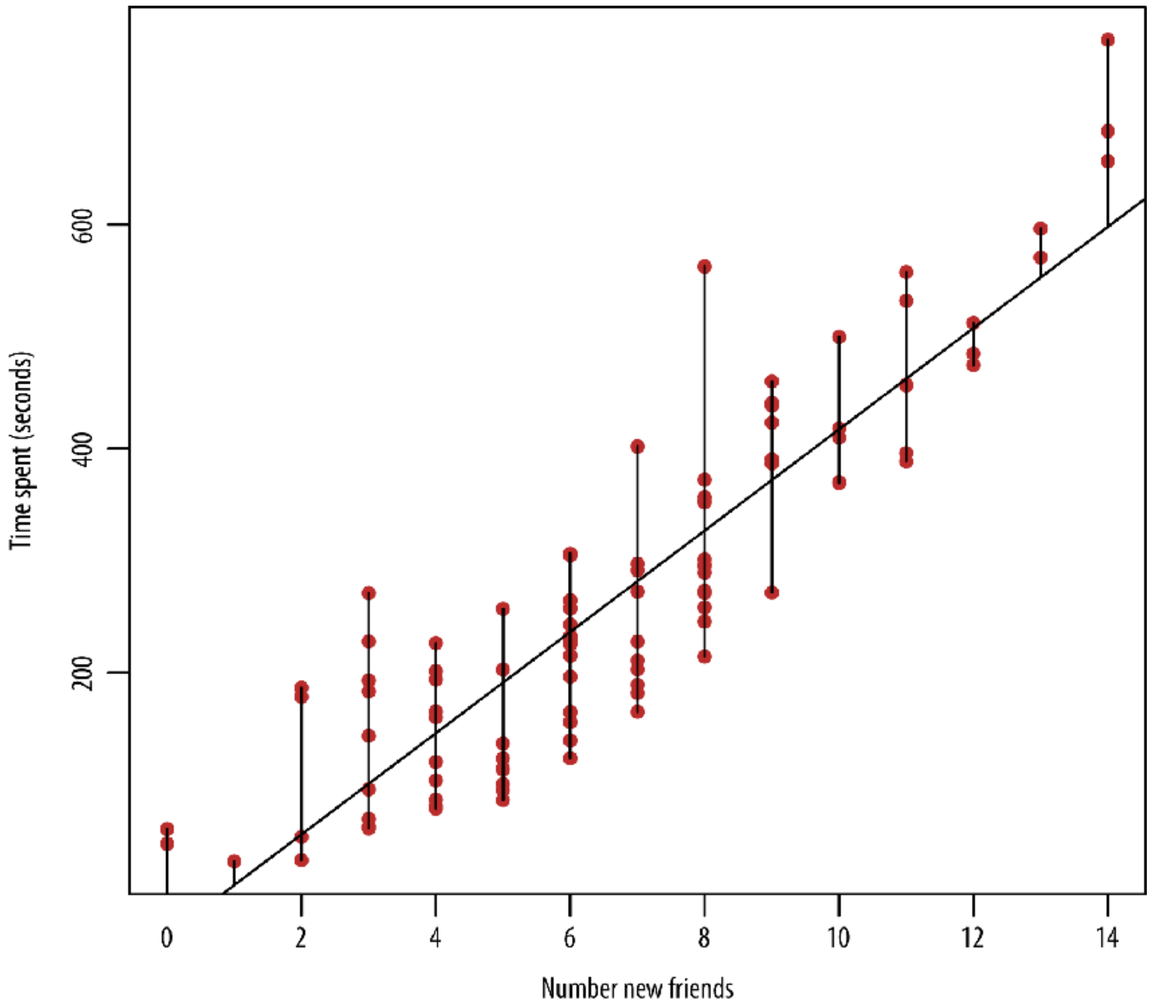


Modelado Estadístico

- Ajustar un modelo dada una muestra de datos

$$y = \alpha + x\beta$$

- Aprender los parámetros que mejor se ajusten (que menos se equivoquen)

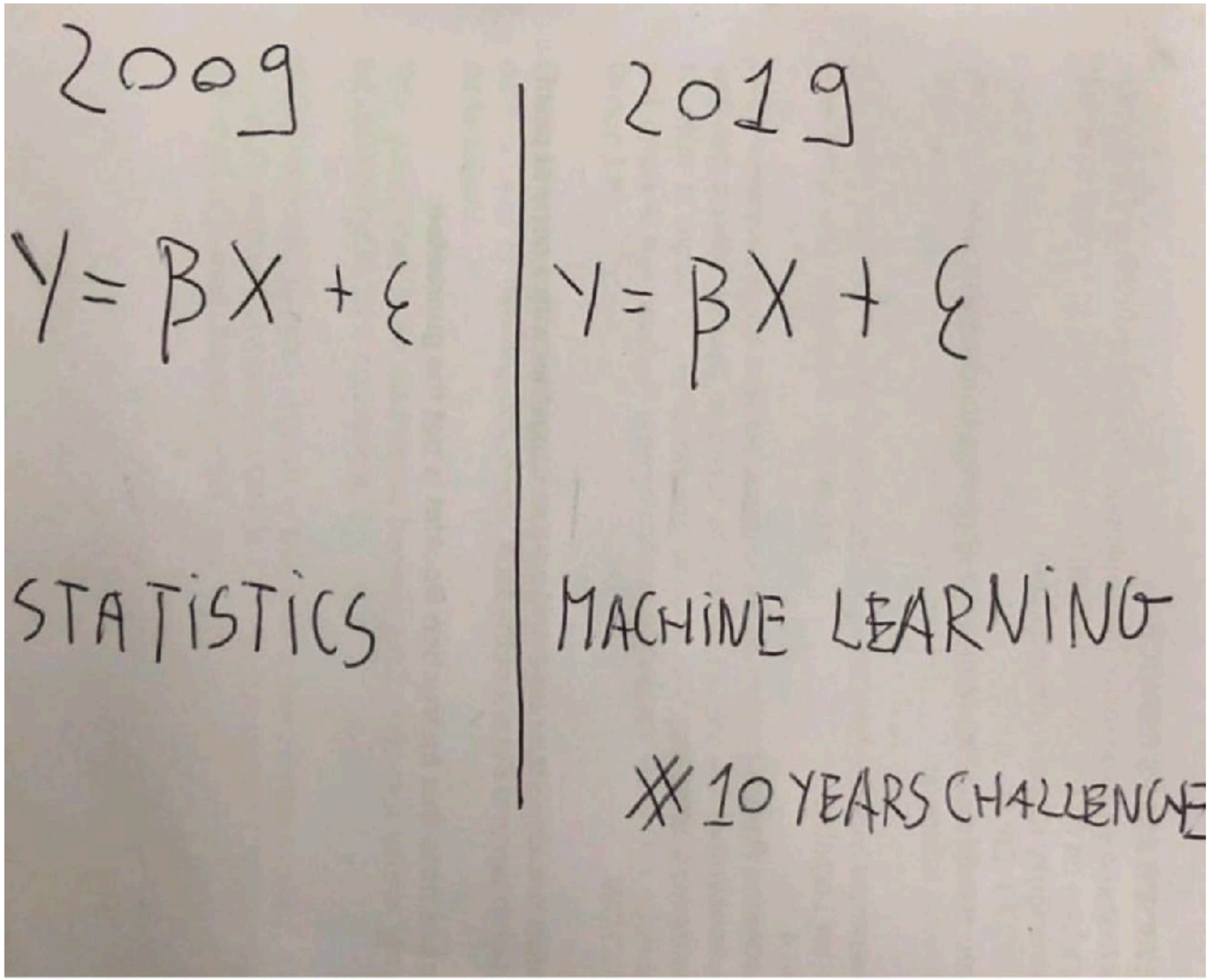


Machine Learning

- Aprendizaje de patrones a partir de los datos utilizando algoritmos estadísticos automáticos
- Usar los modelos para realizar procesos de inferencia tales como predicción, clasificación…

- ¿Cuál es la diferencia entre machine learning y modelado estadístico?

Isn’t machine learning just glorified statistics?



Machine Learnins VS Modelado Estadístico

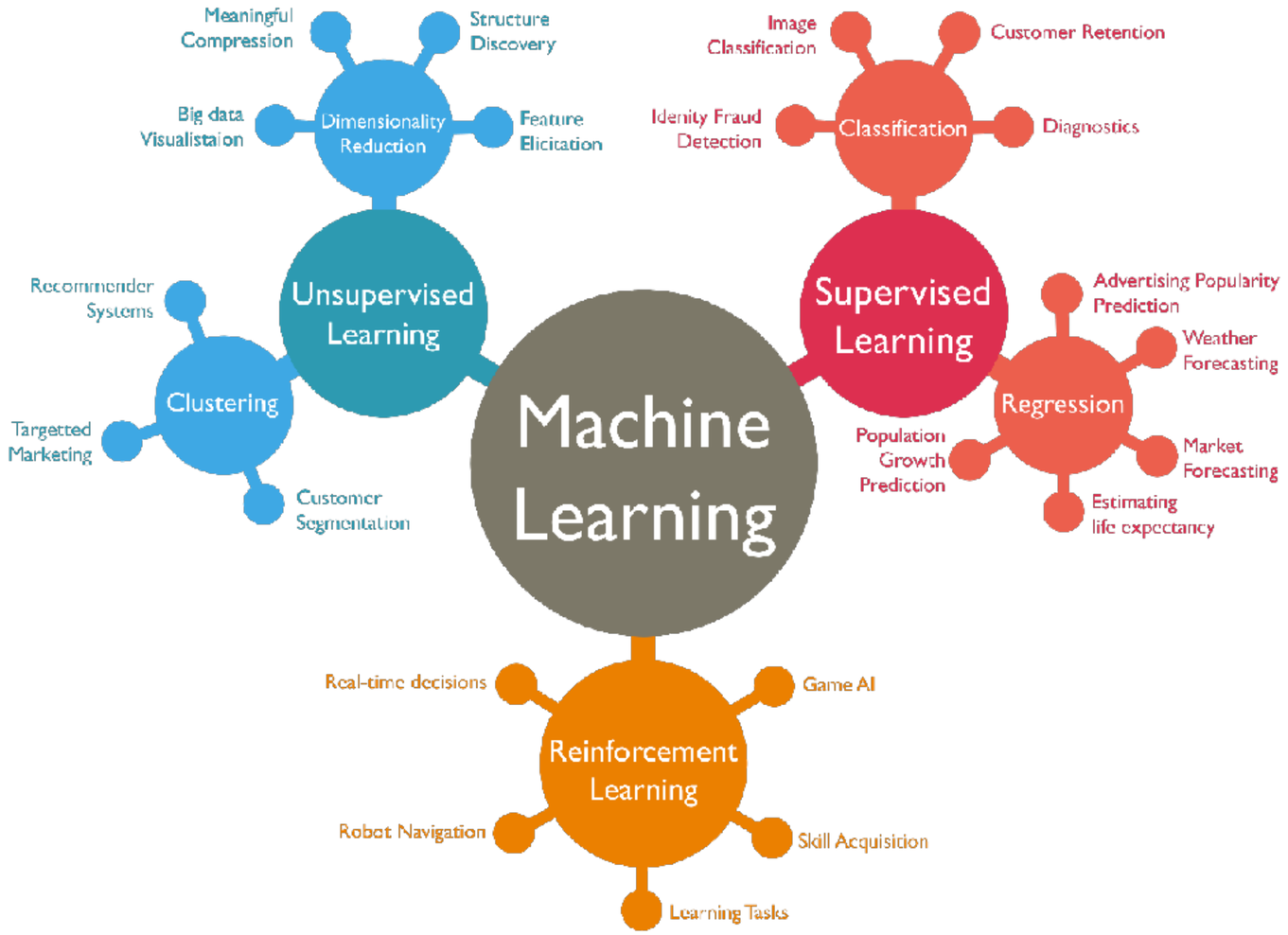
STATISTICAL MODELING

- Interpretabilidad de los parámetros
- Explicación de los resultados
- Obtener la mejor aproximación del modelo generativo

MACHINE LEARNING

- Centrado en los resultados
- Suele producir modelos “caja negra”
- Busca la automatización e integración

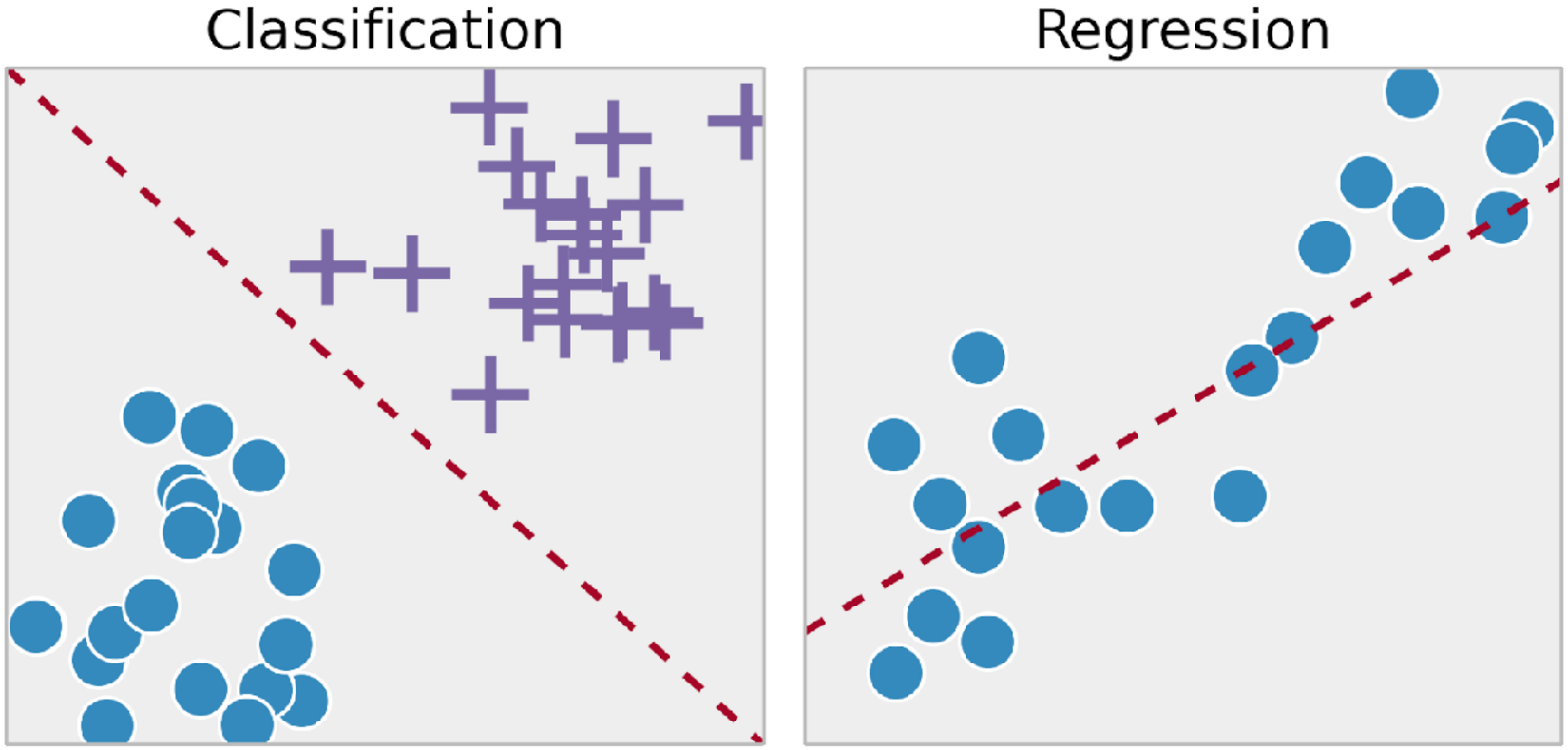
Tipos de Machine Learning



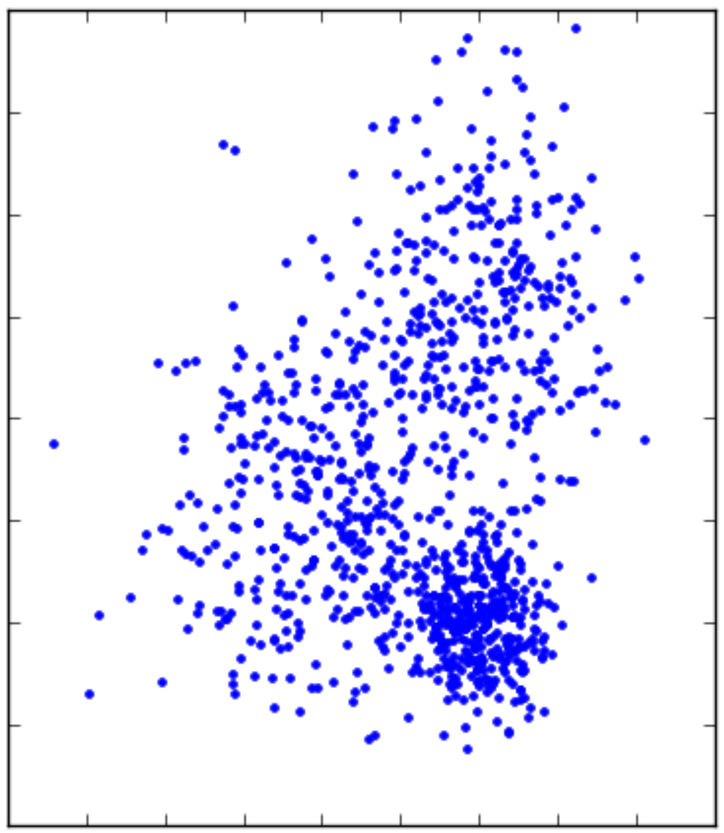
Hacer las preguntas correctas

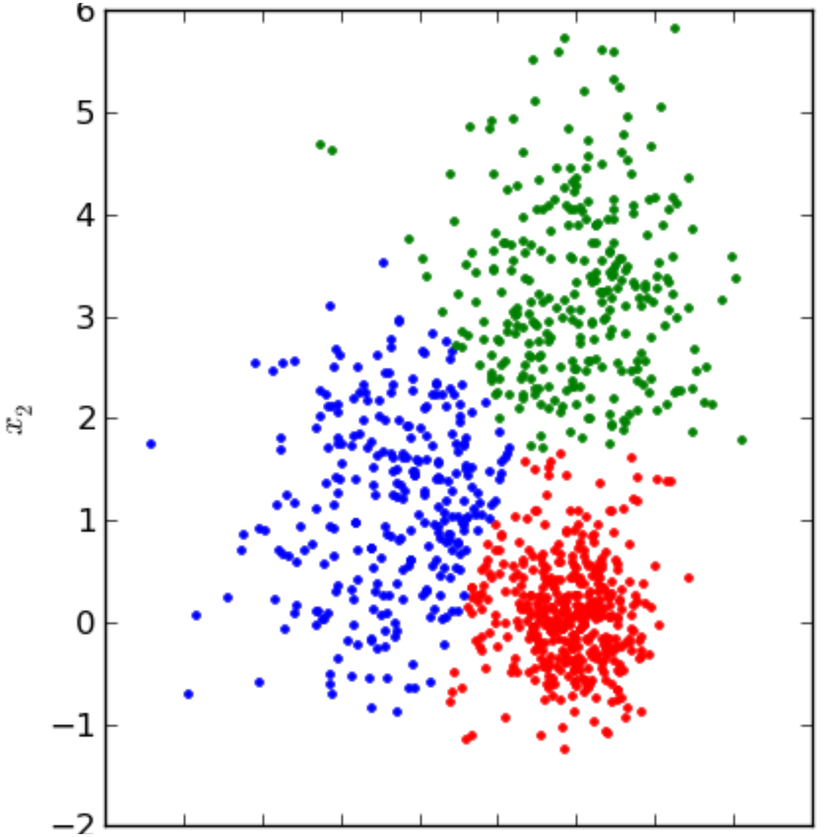
Clasificación	*"¿Esto es un perro o un gato?"*
Regresión	*"¿Cuánto valdrá esta vivienda?"*
Anomalías	*"¿Es esta la temperatura habitual?"*
Clustering	*"¿Cuantos tipos de bacteria hay?"*
Recomendación	*"¿Qué película preferirá este usuario?"*
Refuerzo	*"¿Tendré un accidente si cambio de carril?"*

Aprendizaje Supervisado

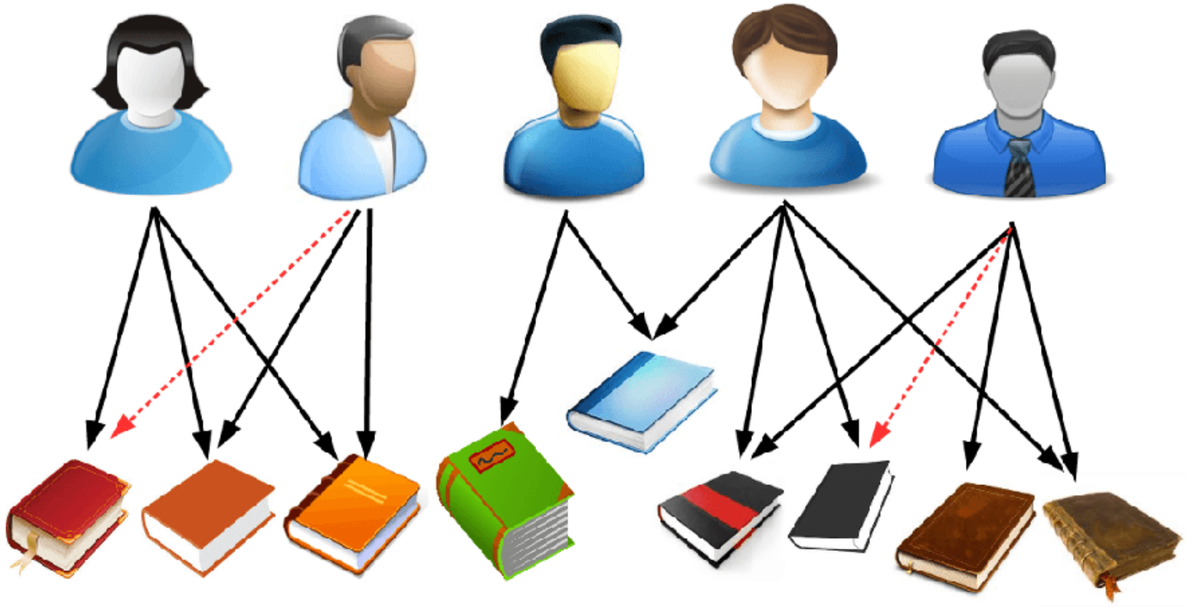


Aprendizaje no Supervisado



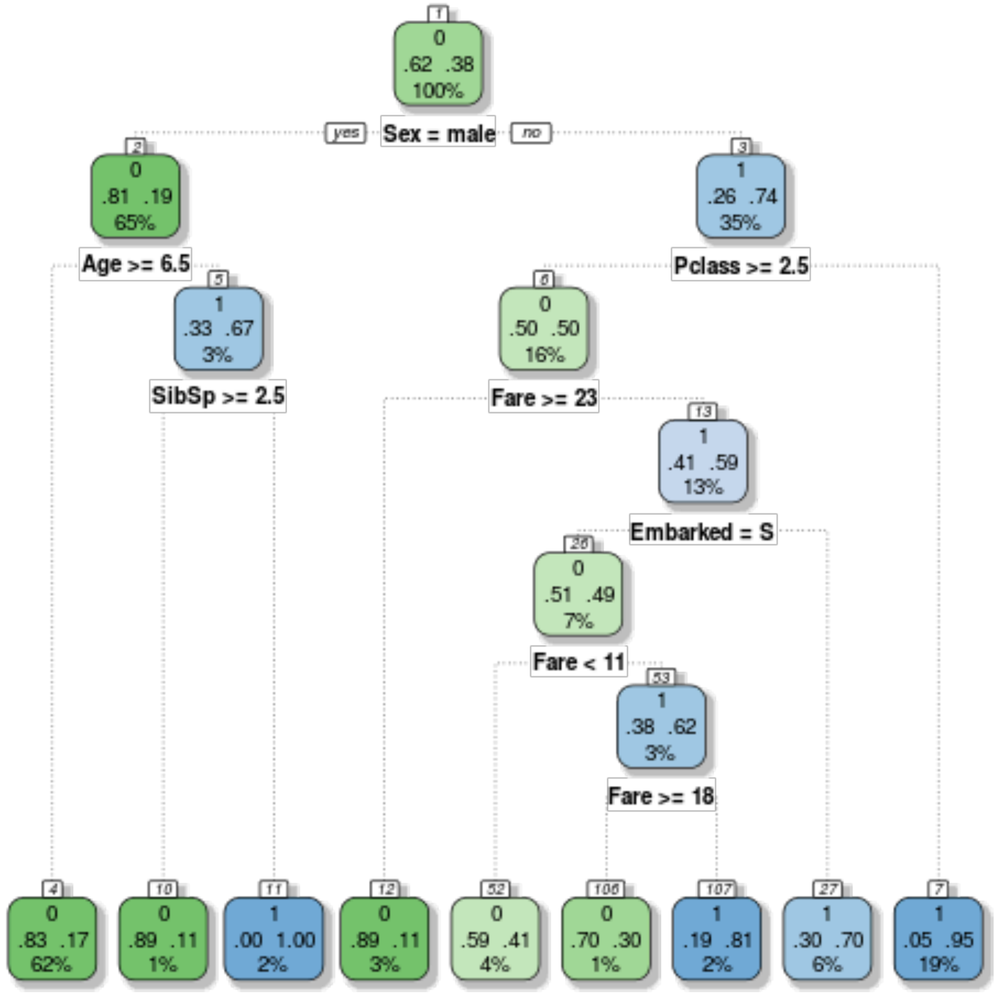


Sistemas de Recomendación



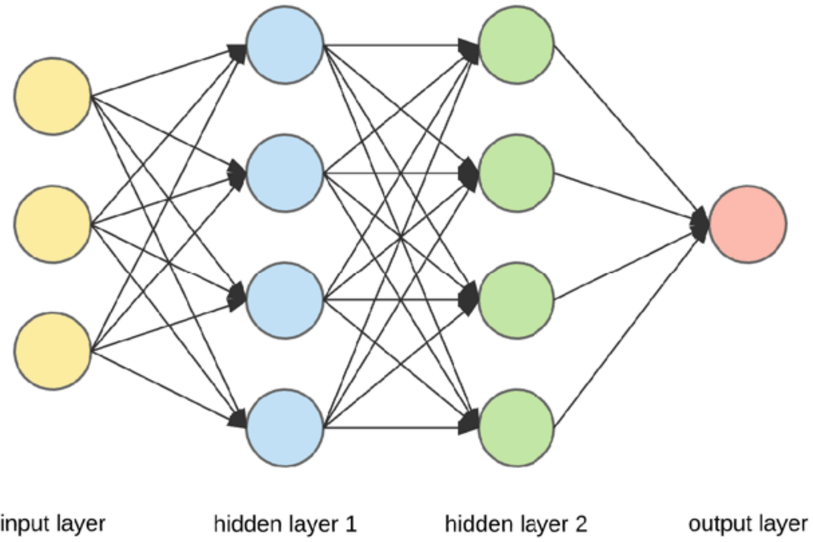
Tipos de Modelos

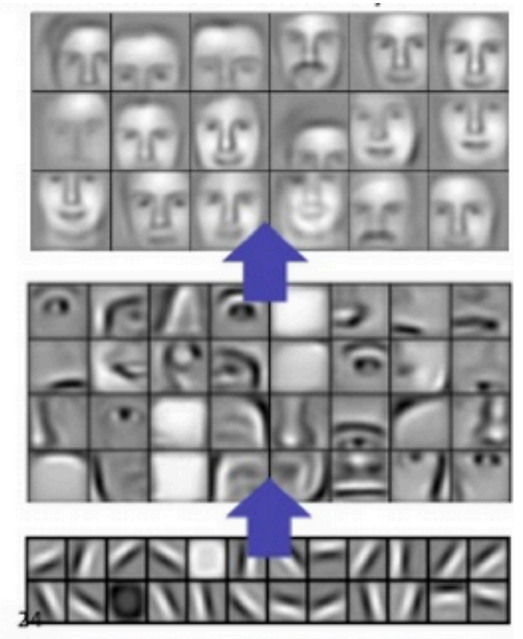
- Árbol de decisión
- Aprende un árbol binario que clasifica objetos
- Utiliza métricas de probabilidad y teoría de la información
- Puede ser fácilmente interpretado



Tipos de Modelos

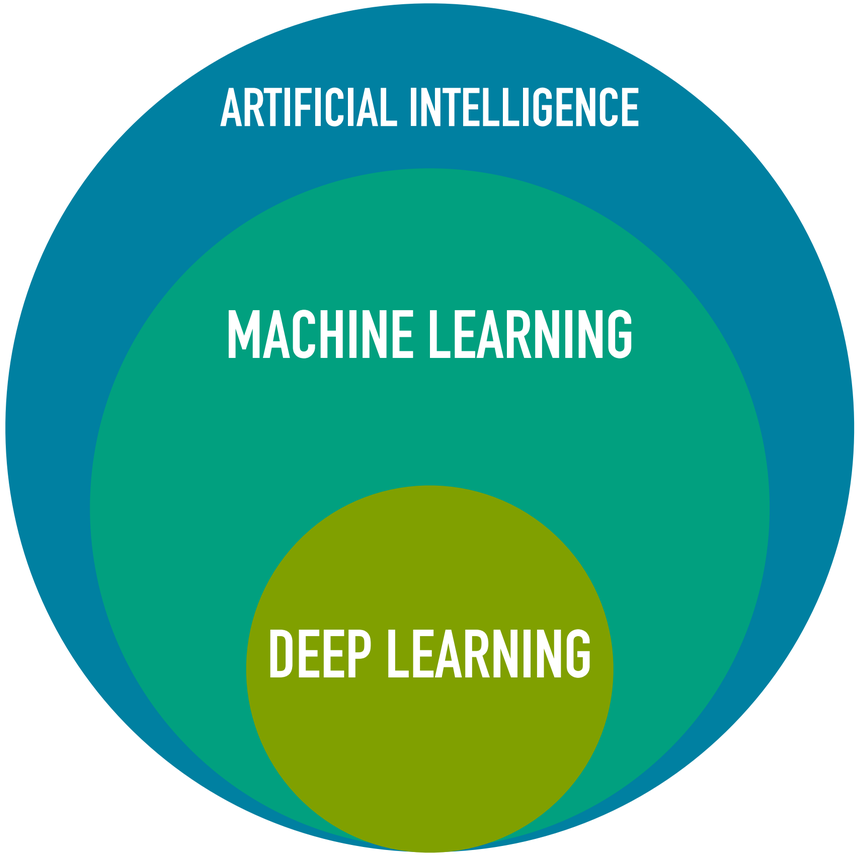
- Red Neuronal
- Modelo de caja negra
- Muy poca o nula interpretación
- Inspirado en como se conectan las neuronas entre si





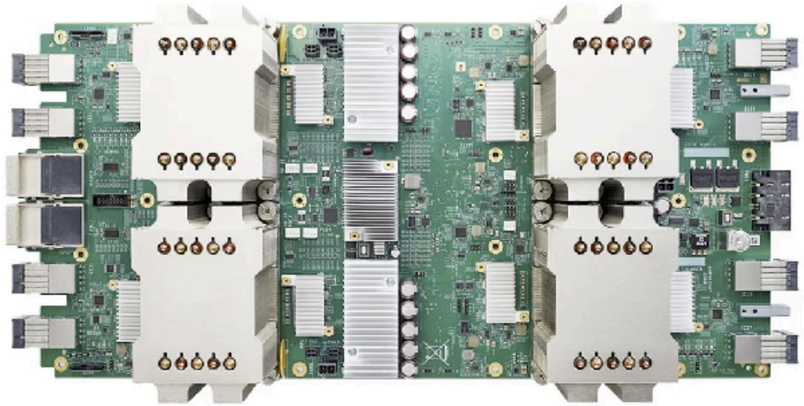
Deep Learning

- Técnicas avanzadas que aprenden redes inmensas
- Requieren grandes cantidades de datos y capacidad de cómputo
- Los modelos no pueden ser interpretados
- En ocasiones se pueden transferir a otros dominios



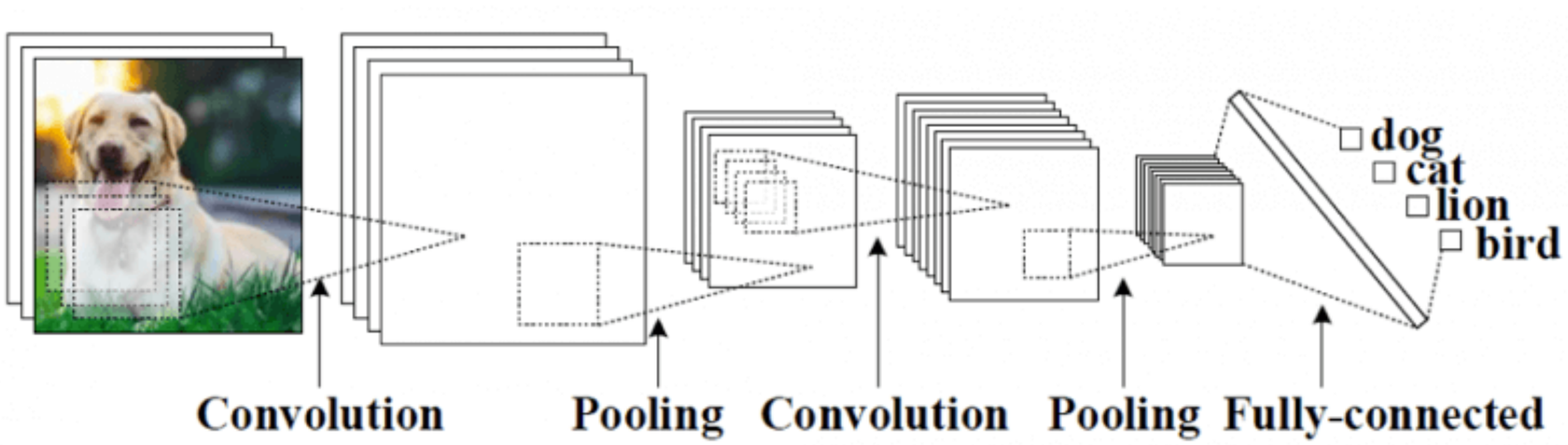
Deep Learning = Datos + Cómputo Intensivo

- La teoría básica data de los 50s y 80s
- Popularizado en los 2000s
- Reservado a grandes volúmenes de datos y hardware especializado (GPU y TPU)



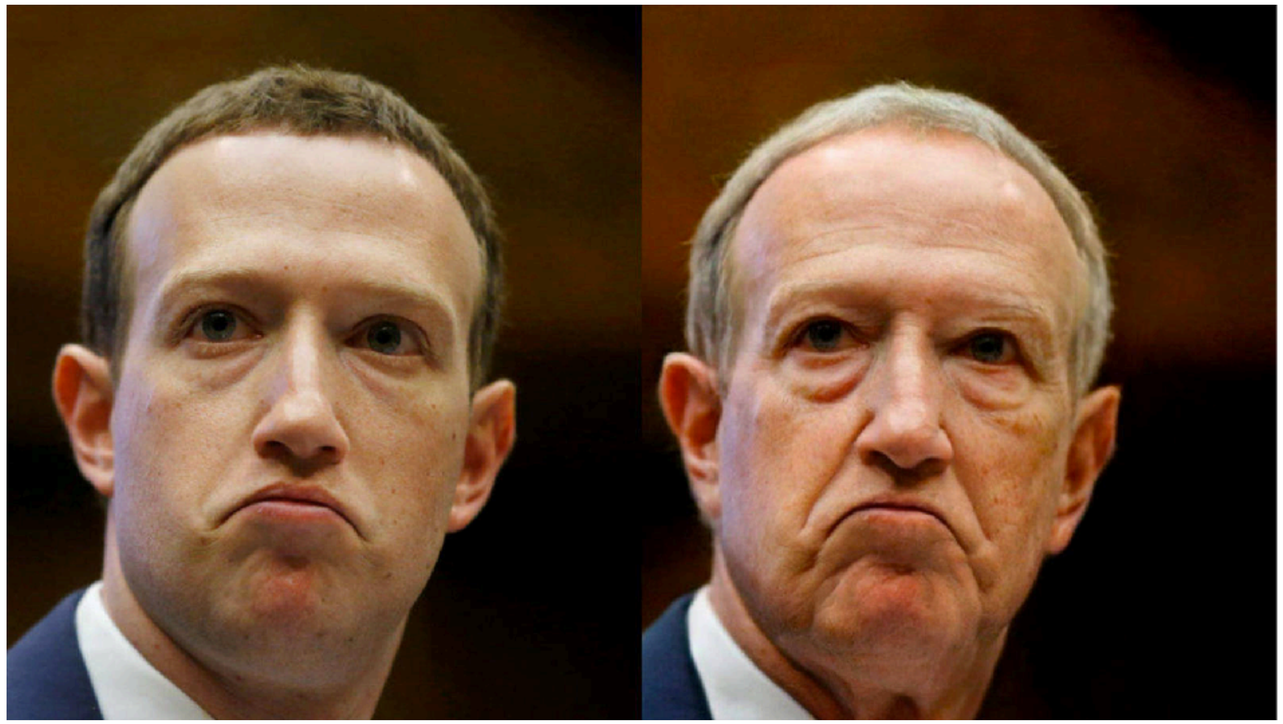
Ejemplos de Deep Learning

- Redes Convolucionales: Clasificación y regresión
- Reducción de la dimensionalidad sin pérdida
- Aprendizaje de muchas características intermedias



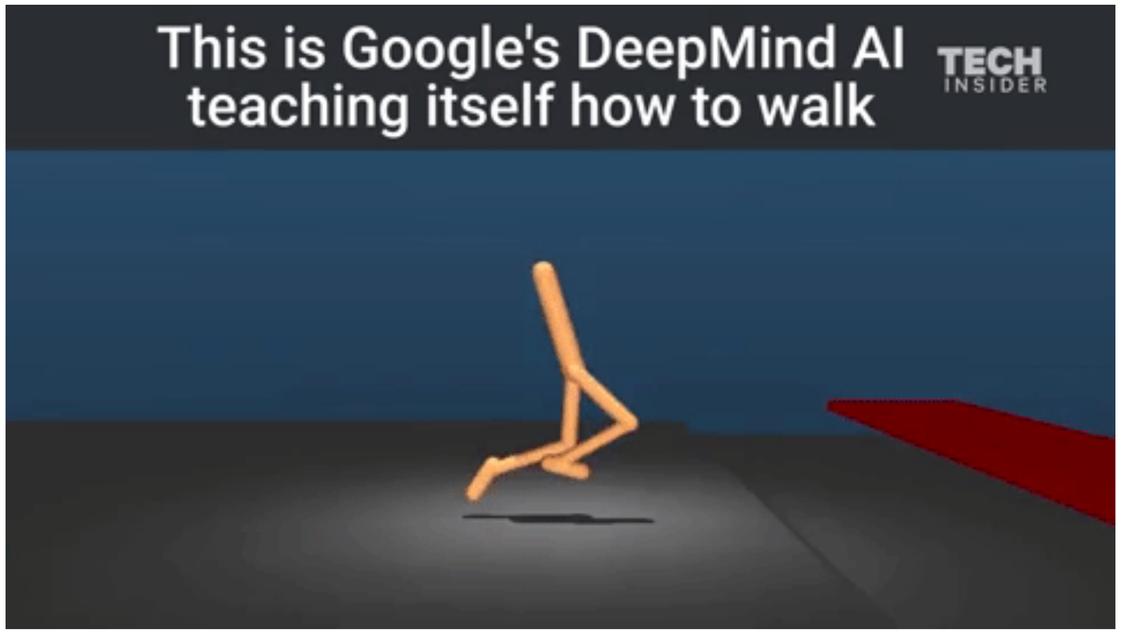
Ejemplos de Deep Learning

- Redes generativas adversariales (GANs): Generación de datos sintéticos



Ejemplos de Deep Learning

- Aprendizaje por refuerzo: Aprende de la experiencia, datos por simulación. Aplicaciones en robótica, videojuegos, navegación…



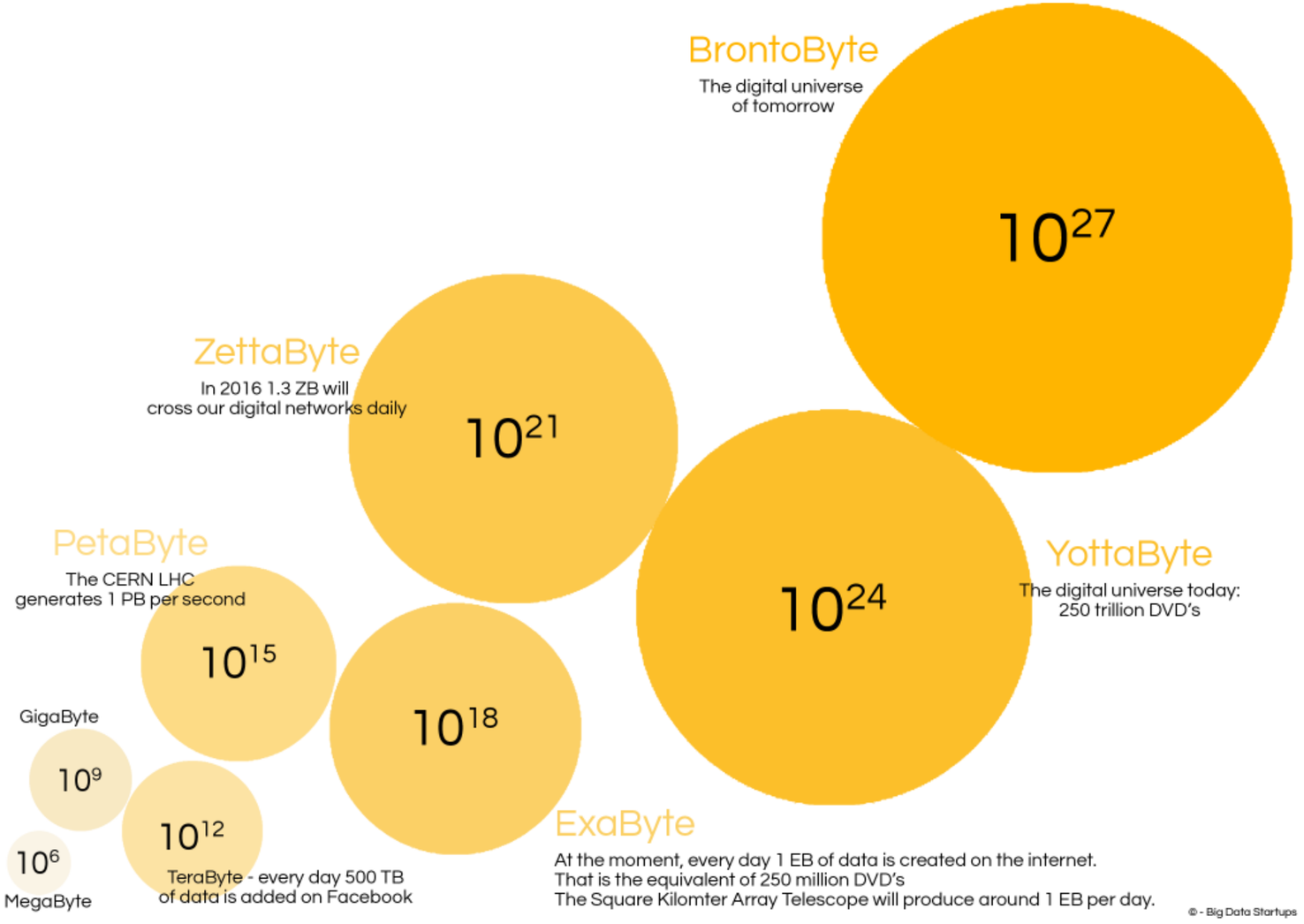
El Ecosistema

Personas

Procesos

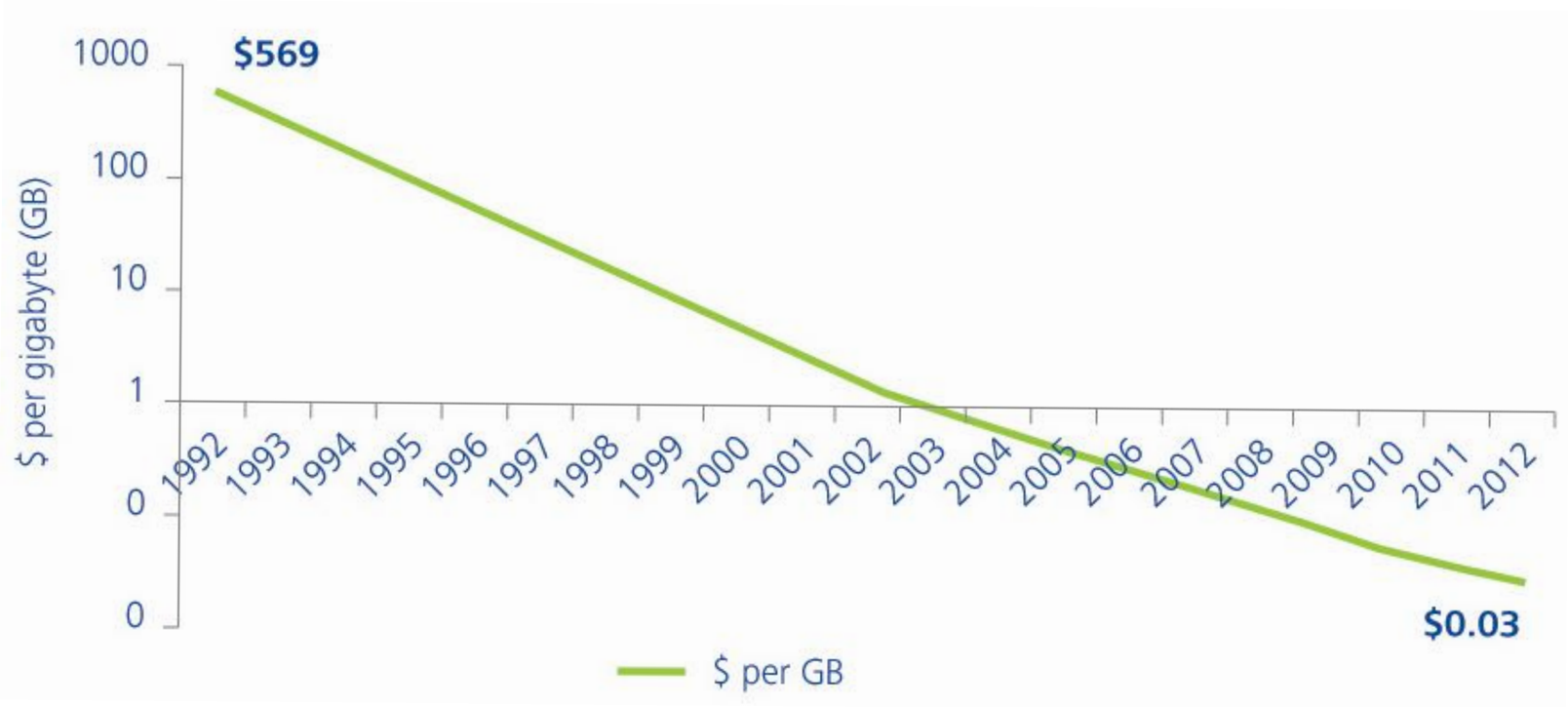
Tecnología

Las escalas son relativas a la tecnología



La tecnología impone limites

- “Big” es cuando no nos cabe en una sola máquina

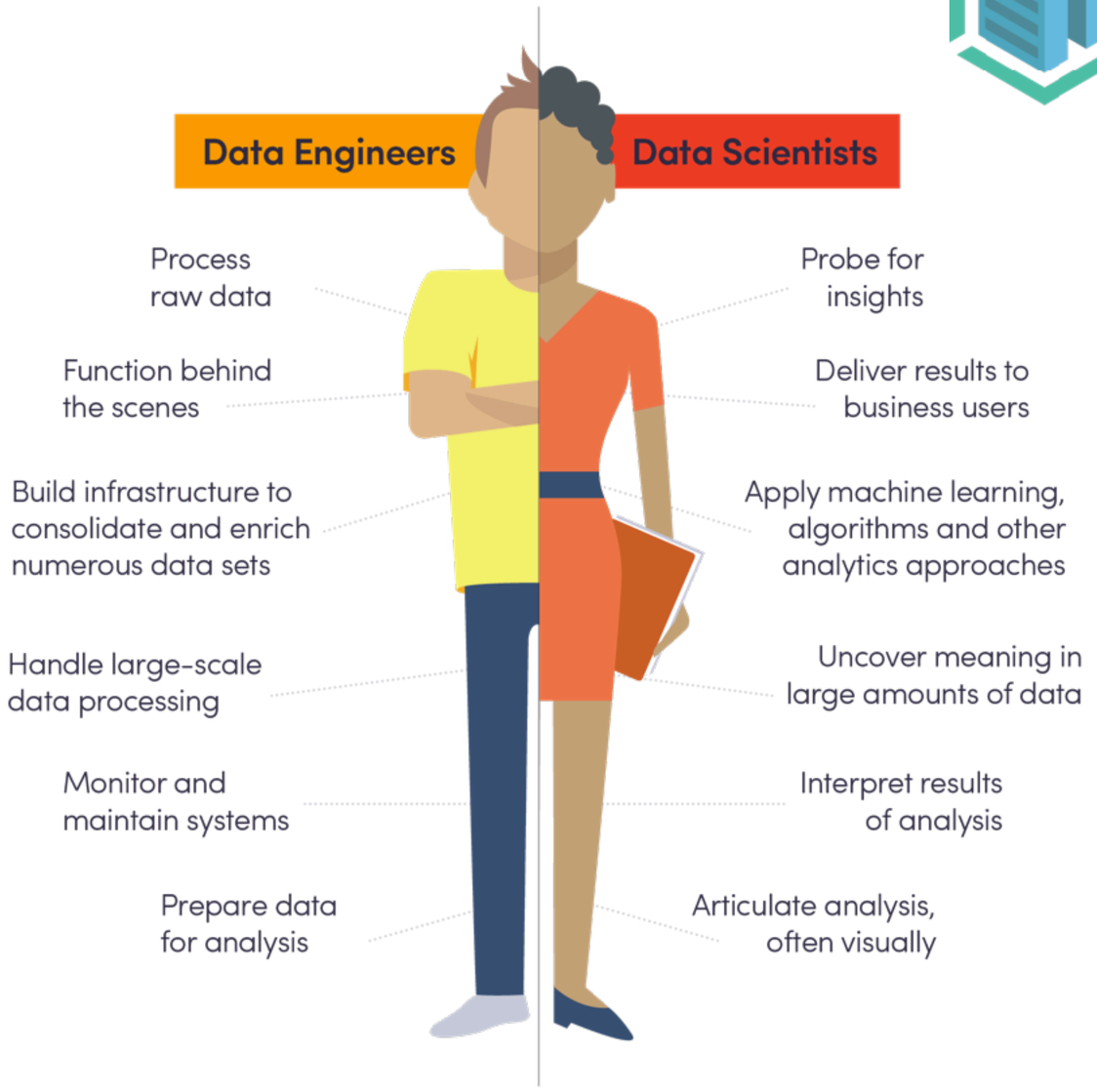


Big Data = Big Computing Power

- Un set específico de tecnologías que requieres habilidades concretas
- Gestionar múltiples máquinas simultáneamente es costoso
- Necesitamos abstraer la deslocalización de los datos y la ubicuidad del cómputo

El perfil del Ingeniero de Datos

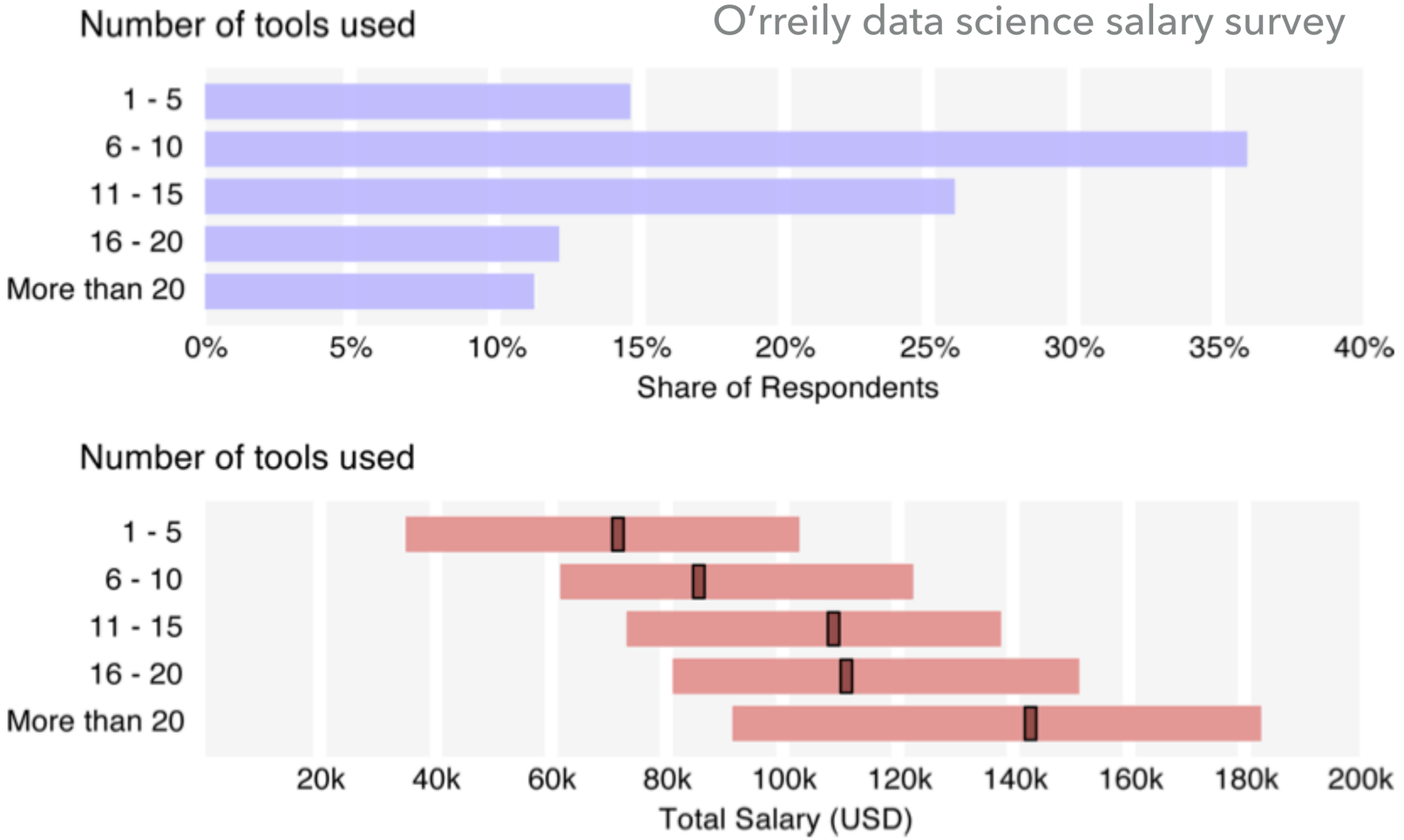
- Centrado en los datos, pero especializado en la tecnología
- Gestiona el complejo stack de big data y cloud
- Trabaja cerca del origen y la infraestructura que aloja los datos
- Clave en la captura y limpieza de los datos



YA TENGO SUFICIENTES HERRAMIENTAS

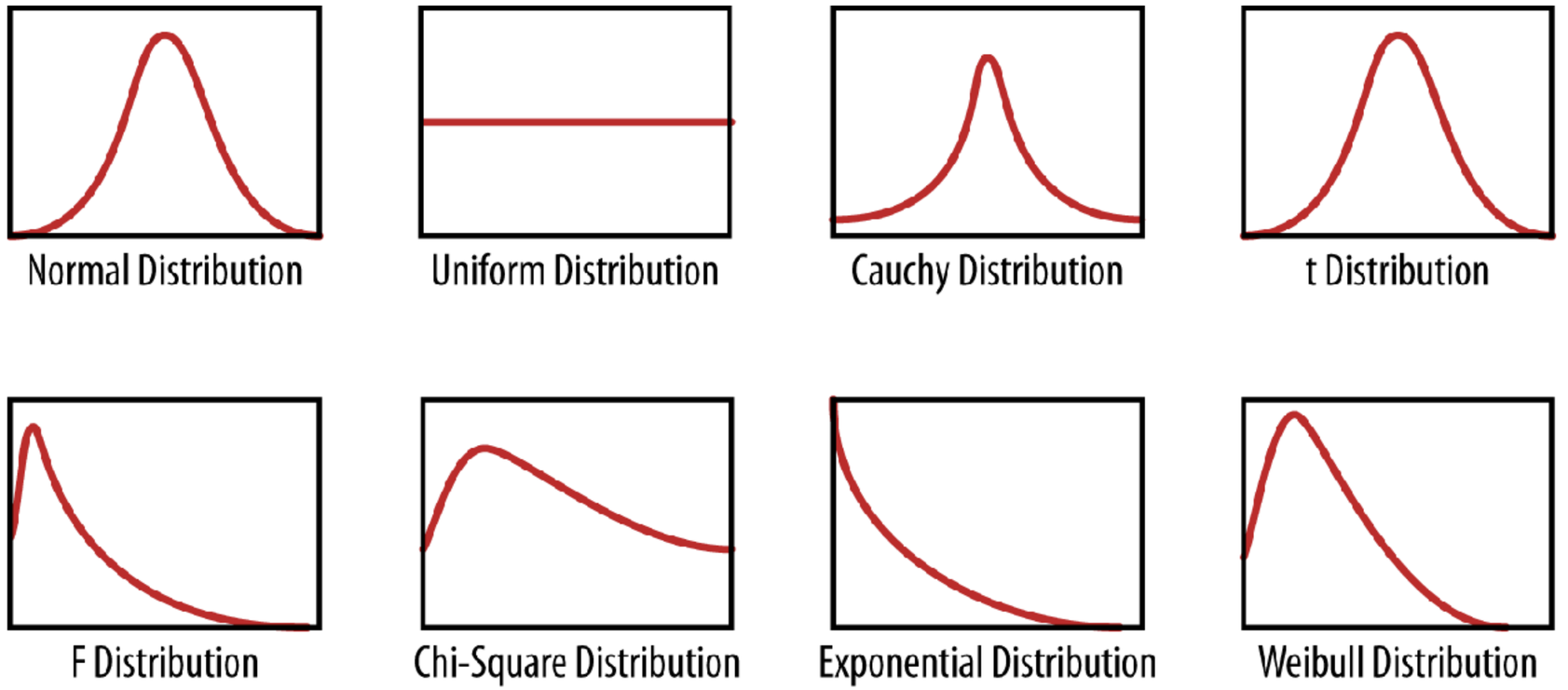
Esto no lo dijo un Data Scientist

Un ecosistema altamente especializado



¡Matemáticas y Estadística!

- Álgebra Lineal
- Probabilidad
- Estadística descriptiva e inferencial
- Teoría de la información
- Teoría de grafos



Lenguajes y habilidades de programación

- Hay que diferenciar entre las habilidades de programación y el lenguaje/librerías concretos que se utilizan
- Los fundamentos son abstractos y se pueden reutilizar
- Los lenguajes son herramientas y se solapan

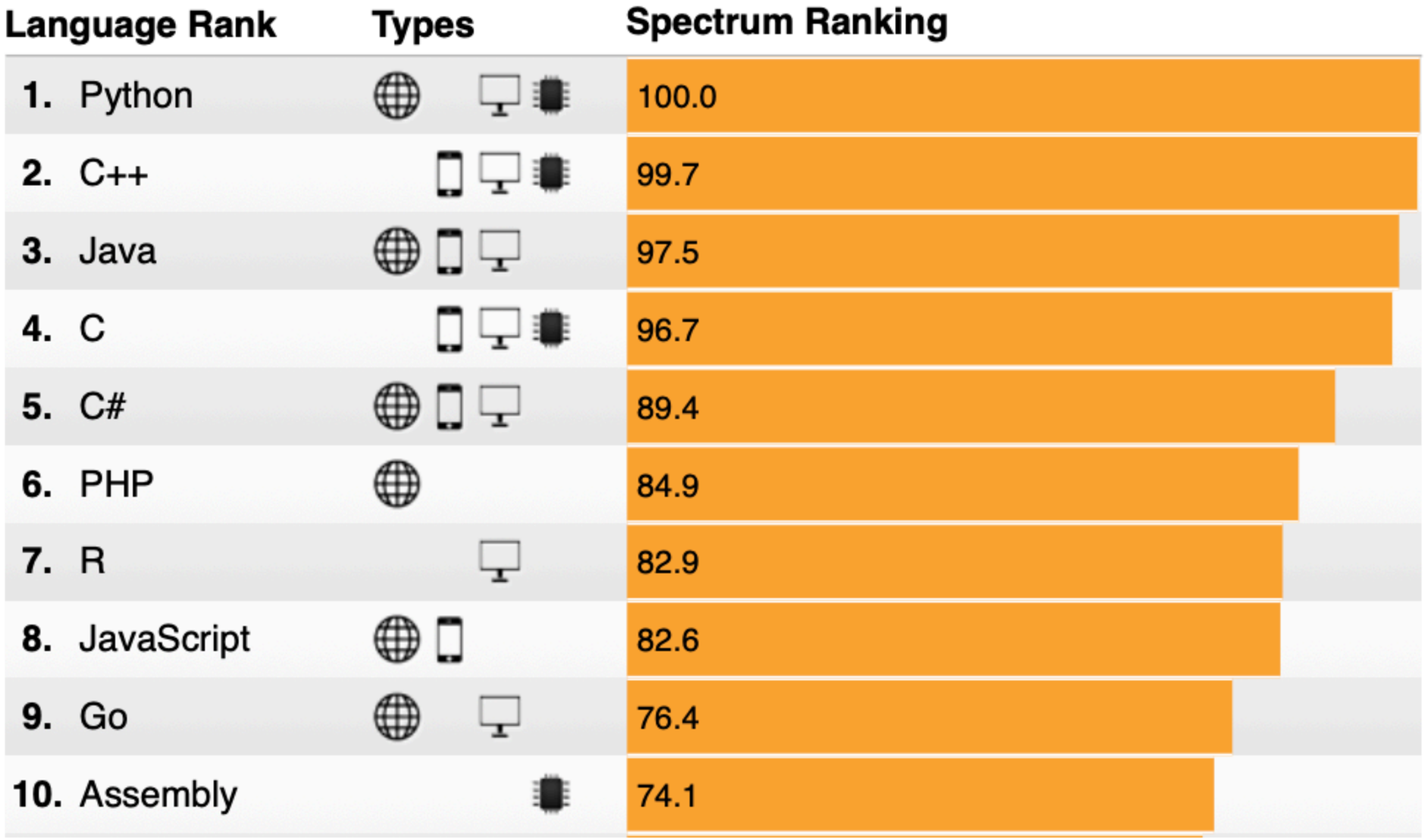
Lenguajes y habilidades de programación

- Hay que diferenciar entre las habilidades de programación y el lenguaje/librerías concretos que se utilizan
- Los fundamentos son abstractos y se pueden reutilizar
- Los lenguajes son herramientas y se solapan

Lenguajes y habilidades de programación

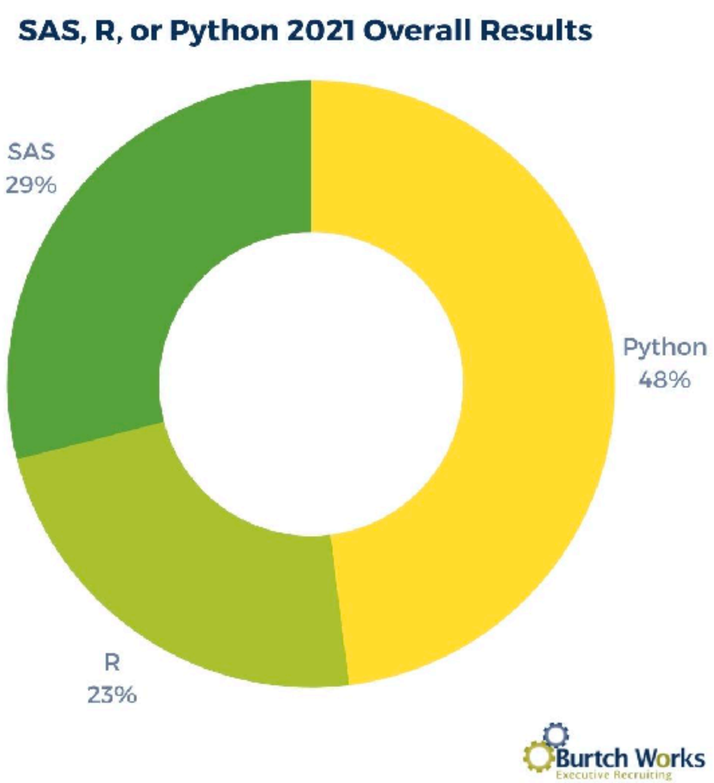
- Hay que diferenciar entre las habilidades de programación y el lenguaje/librerías concretos que se utilizan
- Los fundamentos son abstractos y se pueden reutilizar
- Los lenguajes son herramientas y se solapan

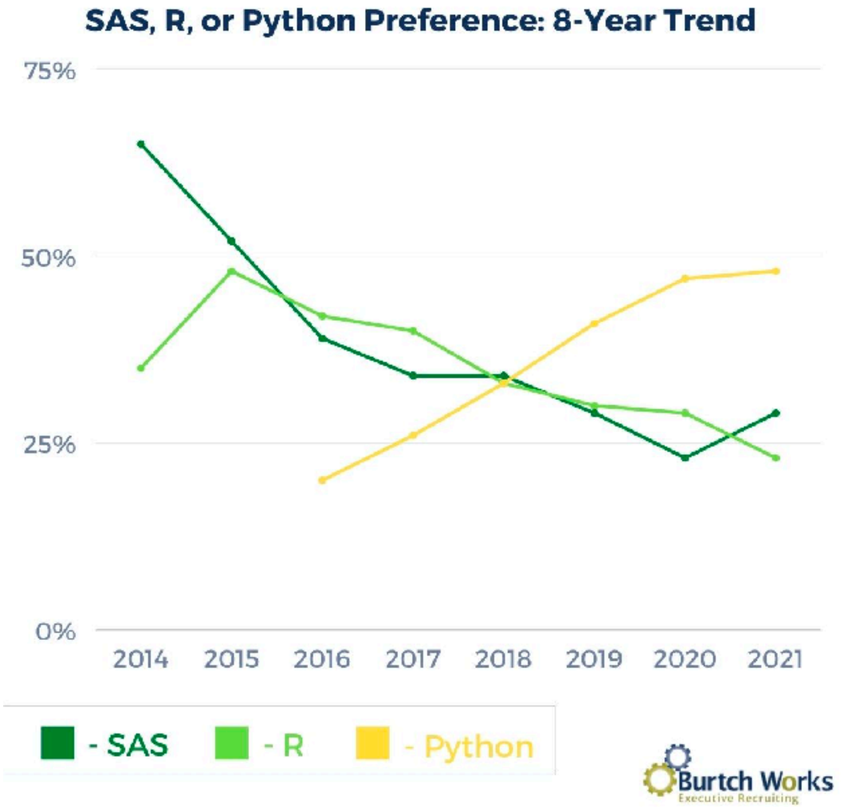
Lenguajes de programación



Language Rank	Types	Spectrum Ranking
1. Python		100.0
2. C++		99.7
3. Java		97.5
4. C		96.7
5. C#		89.4
6. PHP		84.9
7. R		82.9
8. JavaScript		82.6
9. Go		76.4
10. Assembly		74.1

Lenguajes de programación



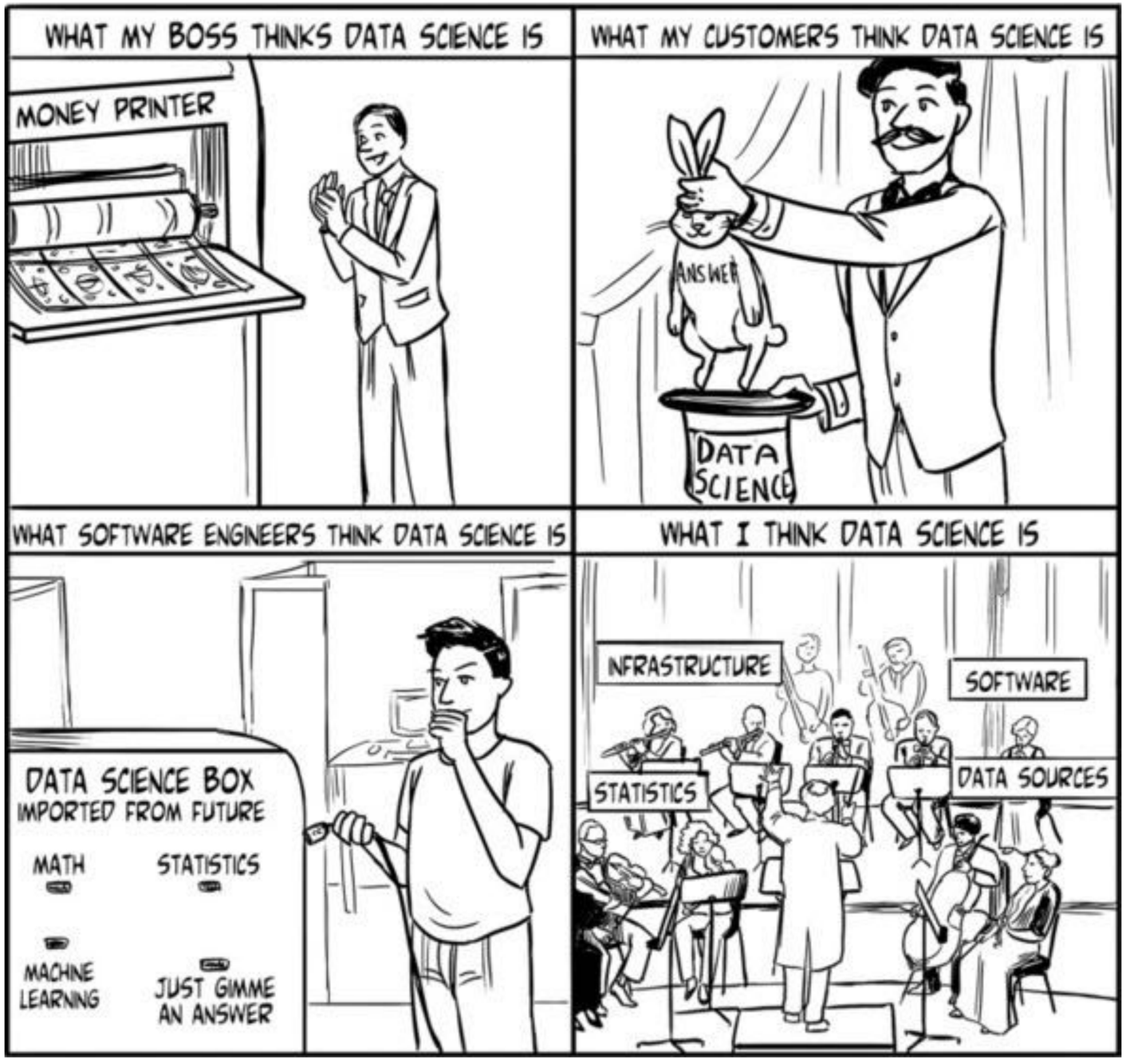


Ecosistema python

Ecosistema Big Data

A hombros de gigantes: Proveedores cloud público

A hombros de gigantes: Ecosistema AWS


WHAT MY BOSS THINKS DATA SCIENCE IS
MONEY PRINTER
WHAT MY CUSTOMERS THINK DATA SCIENCE IS
ANSWER
DATA SCIENCE
WHAT SOFTWARE ENGINEERS THINK DATA SCIENCE IS
DATA SCIENCE BOX
IMPORTED FROM FUTURE
MATH
STATISTICS
MACHINE LEARNING
JUST GIMME AN ANSWER
WHAT I THINK DATA SCIENCE IS
INFRASTRUCTURE
SOFTWARE
STATISTICS
DATA SOURCES

Módulo 1. Introducción a la ciencia de datos

Introducción a la Ciencia de Datos e Ingeniería de Datos

Jacinto Arias

Máster en Ciencia de Datos e Ingeniería de Datos en la Nube