

FUNDAMENTOS DE ANALÍTICA



“You can't manage what you don't measure”,
Tom De Marco, 1982

“We're drowning in data but starving for knowledge”,
John Naisbitt, 1982

DESCRIPCIÓN

- En este curso se introducen los conceptos básicos de la analítica de datos, presentando las características de los modelos de **aprendizaje automático (*machine learning*)**, desde un enfoque teórico (introductorio) y práctico.
- Estudiamos **modelos supervisados** (permiten la predicción) y **no supervisados** (encuentran estructura en los datos)
- Usamos las **métricas de calidad** de los modelos y los **protocolos de evaluación** que permiten valorarlos y compararlos.

DESCRIPCIÓN

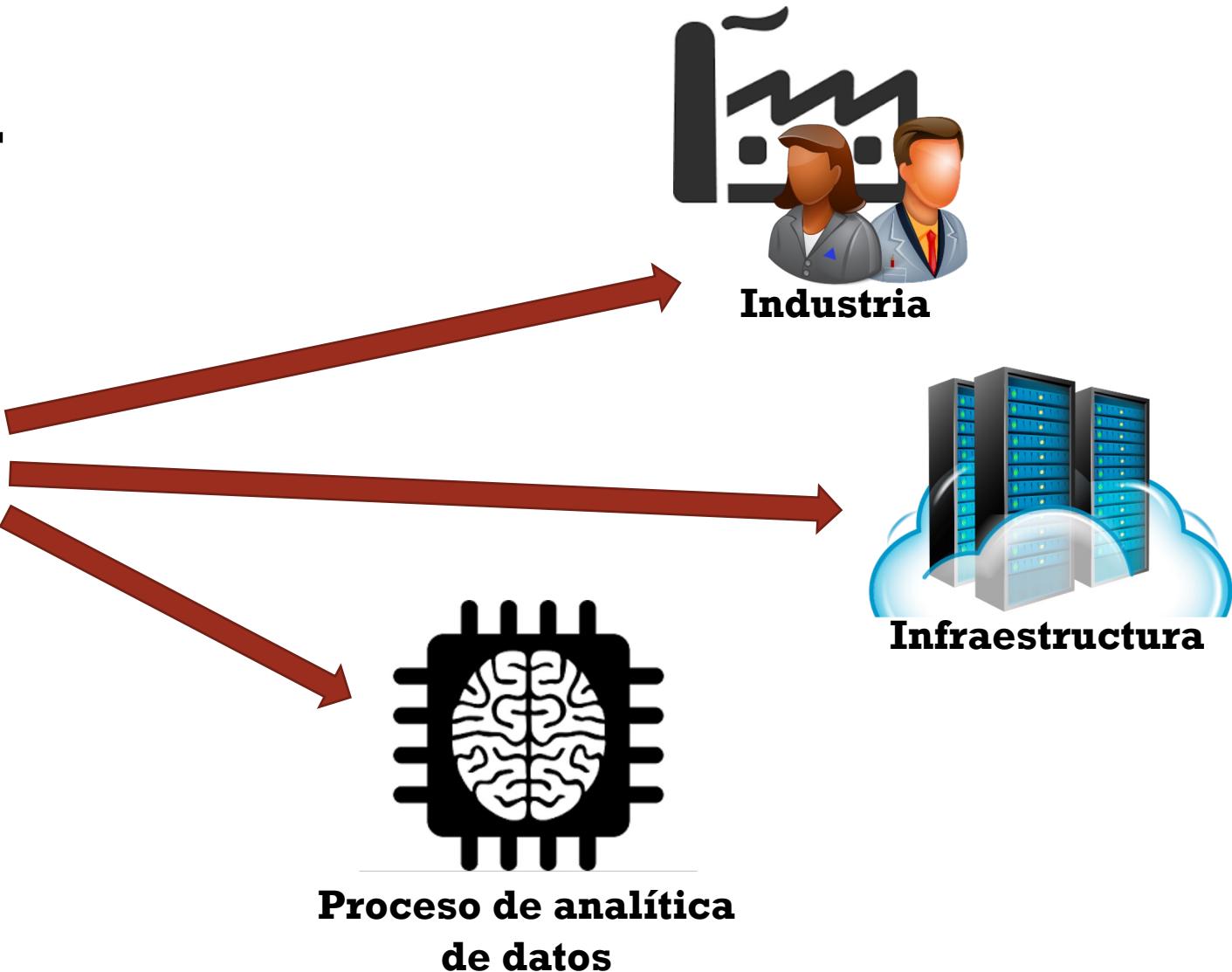
Aprendizaje Supervisado

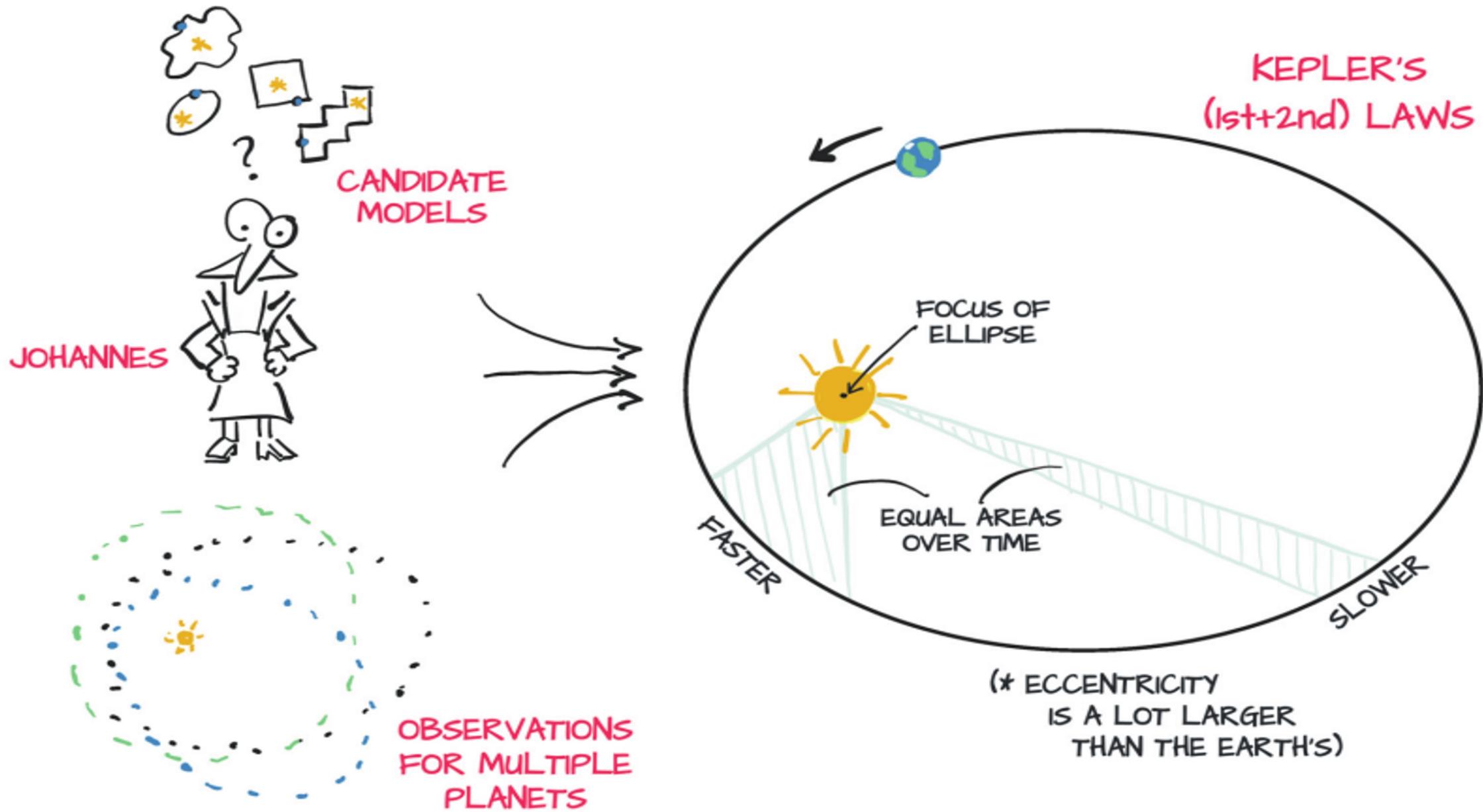
- Sobreaprendizaje
- Protocolos de evaluación
 - Holdout
 - Validación cruzada
- Modelos
 - K-NN
 - Árboles de decisión,
 - Regresión lineal y logística
 - Redes neuronales
 - Bayes Ingenuo

Aprendizaje No Supervisado

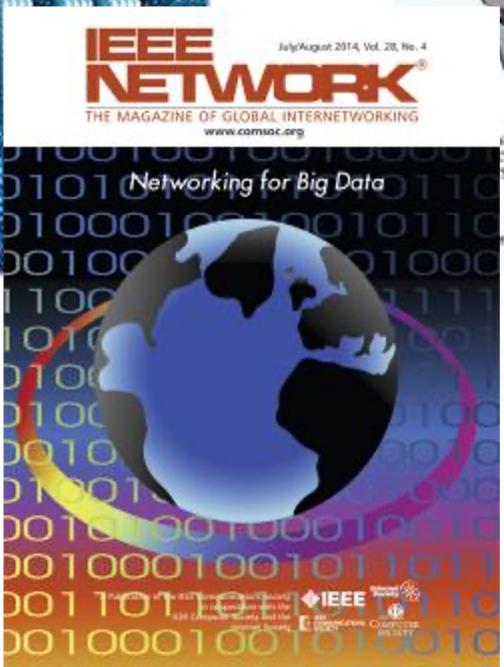
- Segmentación de datos
- Modelos
 - Clustering
 - K-Means
 - Jerárquico
 - Componentes Principales (PCA)

AGENDA





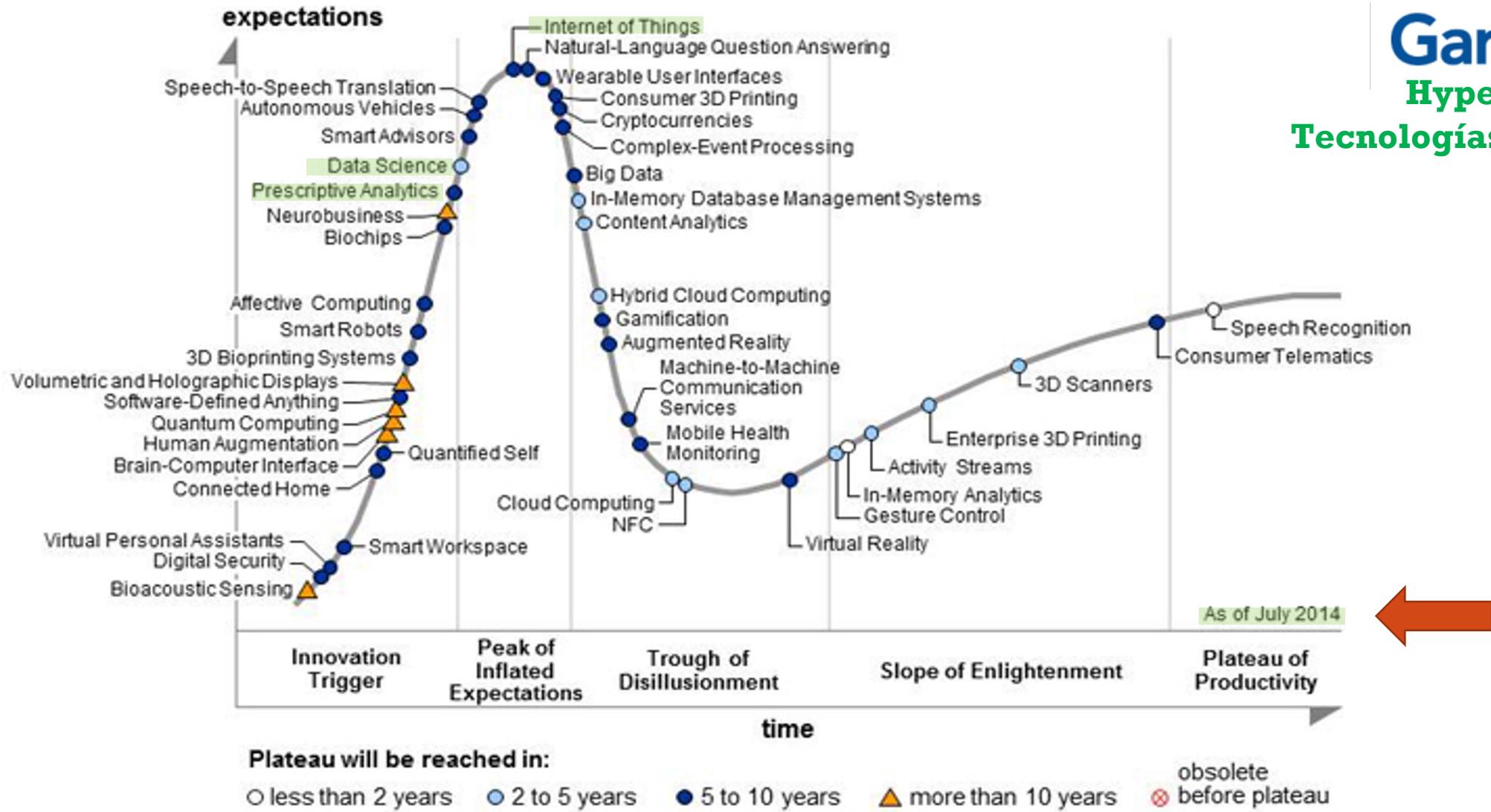
BUZZWORD



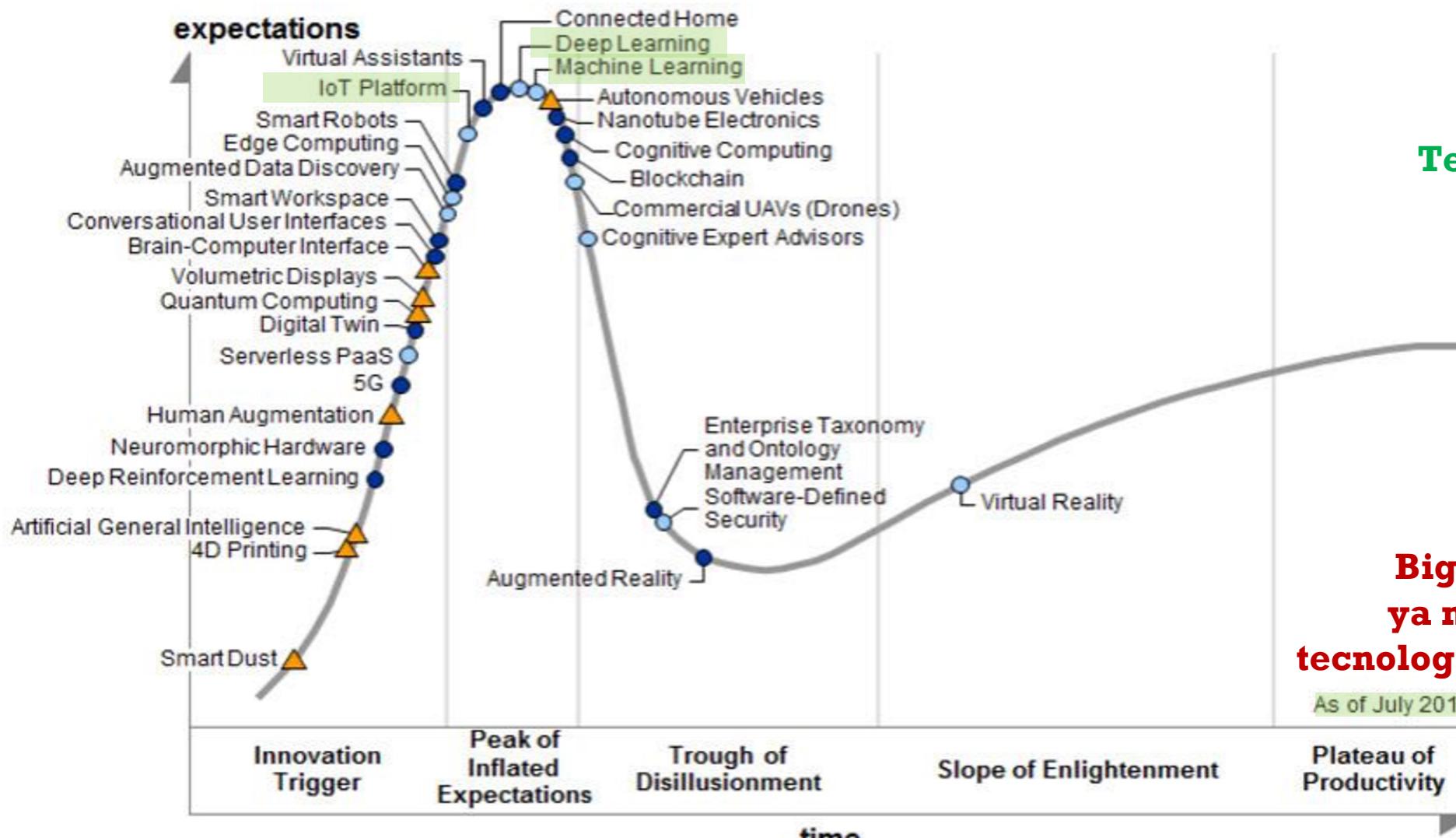
BUZZWORD



Gartner® Hype Cycle Tecnologías emergentes



Gartner® Hype Cycle Tecnologías emergentes



**Big data y analítica
ya no se consideran
tecnologías emergentes!**

As of July 2017

Years to mainstream adoption:

○ less than 2 years

● 2 to 5 years

● 5 to 10 years

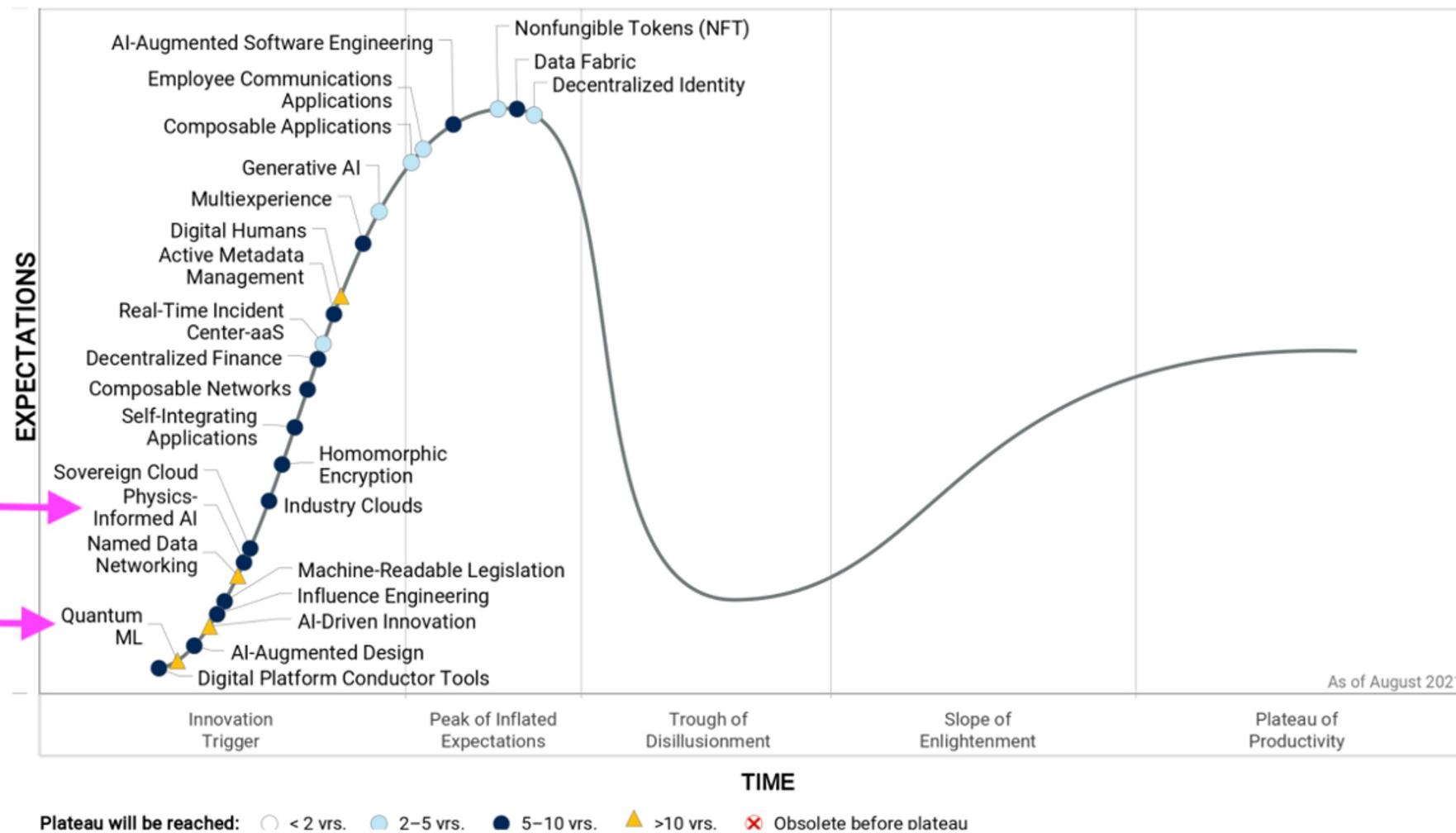
▲ more than 10 years

obsolete

✖ before plateau



Gartner® Hype Cycle Tecnologías emergentes



Source: Gartner (August 2021)

747576

UNIVERSIDAD
ICESI



BIG DATA ANALYTICS

World
Economic
Forum (1:49)

Forbes (2:45)

Harvard
Business
Review (2:44)



APLICACIONES EN LA INDUSTRIA



Inteligencia de clientes



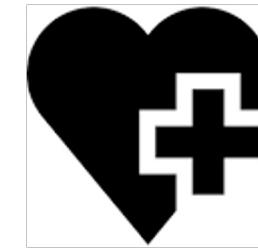
Tendencias



Optimización



Detección de Fraude



Salud



Materias primas



Riesgos



Infraestructura



Análisis de sentimientos



Capacidades



Gobierno



ANALÍTICA EN LA INDUSTRIA

- **Inteligencia de clientes:**

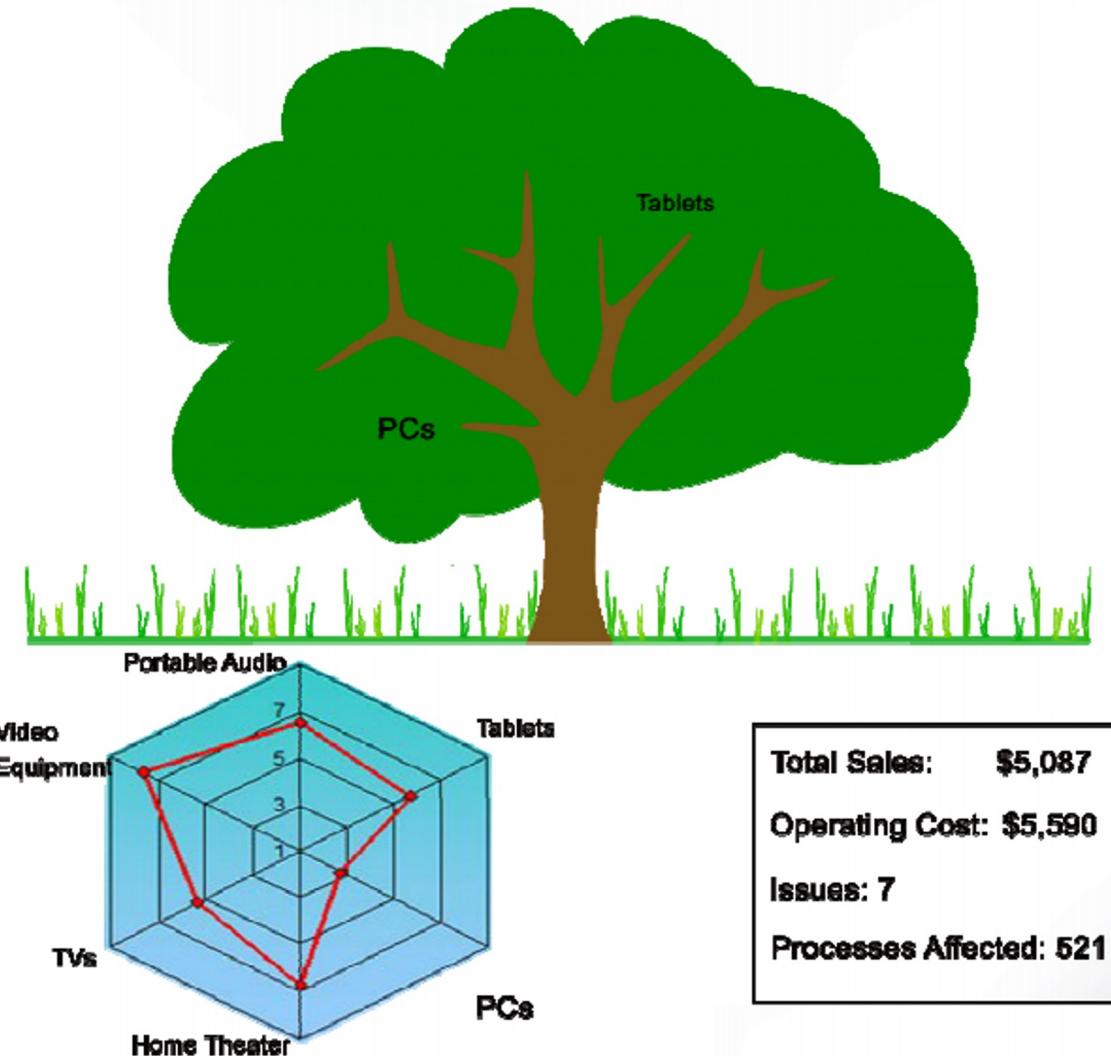
- ¿Cuáles de mis clientes me podrían abandonar para ir a la competencia?
- ¿Cuál es el valor que representan quienes aún no son clientes de mi compañía?
- ¿Qué clientes tienen un comportamiento o características similares?
- ¿Cuáles de mis clientes son propensos a la compra de un producto y cuáles no lo són?
- ¿Que productos o grupos de productos son los más idóneos o atractivos para cada cliente?
- ¿Cuál es nivel de riesgo de estos clientes potenciales aplicando a un crédito?
- ¿Qué percepción tienen los clientes respecto a mi negocio?



ANALÍTICA EN LA INDUSTRIA

Analítica descriptiva

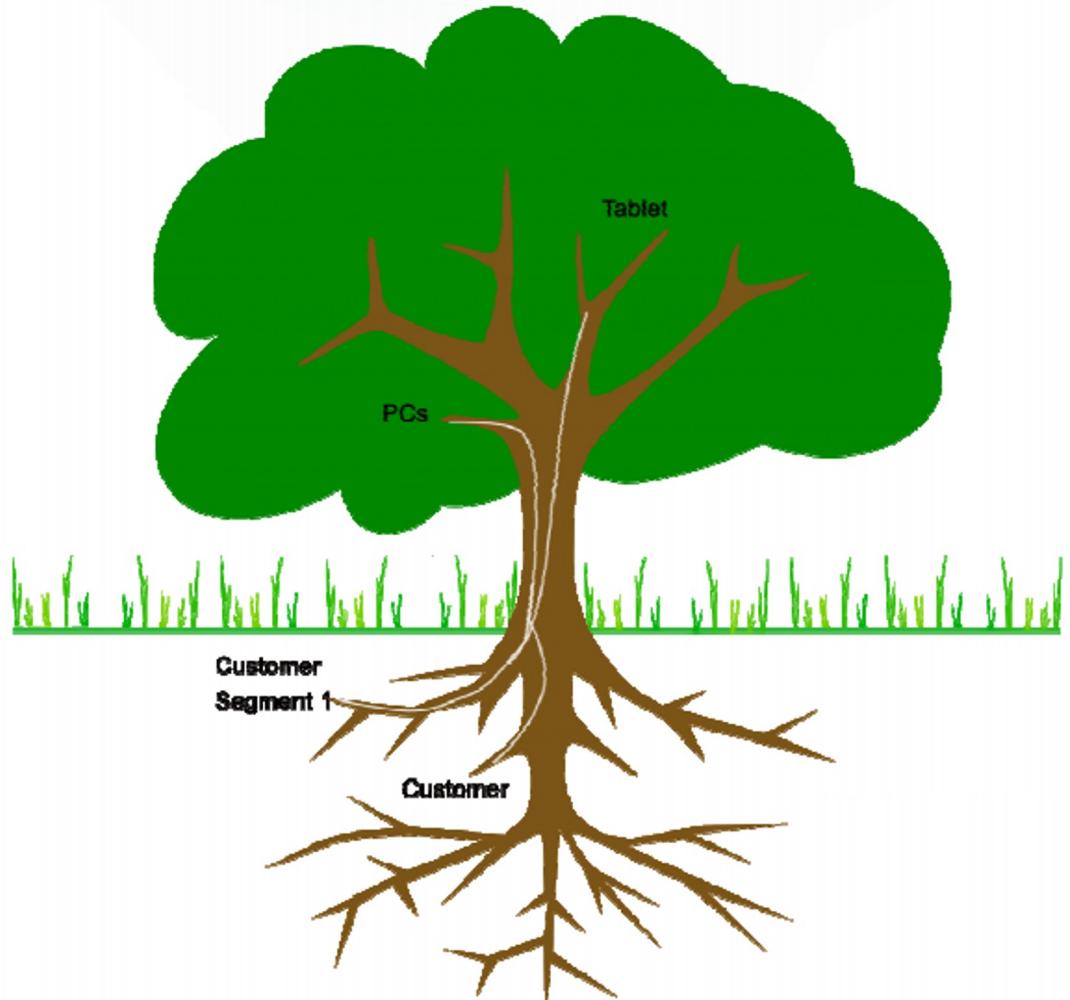
- ¿Qué está pasando? ¿Es bueno o malo?
- Biblioteca de reportes estáticos
- Peticiones de reportes desde todos los departamentos
- Intervención humana requerida
- Una vez se ve una representación, se enfrenta la necesidad de una explicación: ¿Por qué?



ANALÍTICA EN LA INDUSTRIA

Analítica diagnóstica

- ¿Por qué ocurrió esto?
- Explicar el por qué del momento
- Llegar a las causas de lo que se puede percibir
- Descubrir las relaciones entre los datos



ANALÍTICA EN LA INDUSTRIA

Analítica predictiva

- ¿Qué puede pasar?
- Analizar datos históricos para poder tener una idea más clara de lo que podría pasar en el futuro
 - Identificar un problema a resolver
 - Definir qué es lo que se quiere predecir
 - Indicar lo que logrará al hacerlo.



<http://www.modakanalytics.com/img/infographics/predictive-analytics.jpg>



ANALÍTICA EN LA INDUSTRIA

Analítica prescriptiva

- ¿Qué debemos hacer?

Estimar los posibles resultados según las decisiones

- Planificar
- Simular
- Optimizar



Optimization that helps achieve the best outcomes.



Used in producing the credit score which helps financial institutions decide the probability of a customer paying credit bills on time.



Stochastic optimization that helps understand how to achieve the best outcome and identify data uncertainties to make better decisions.

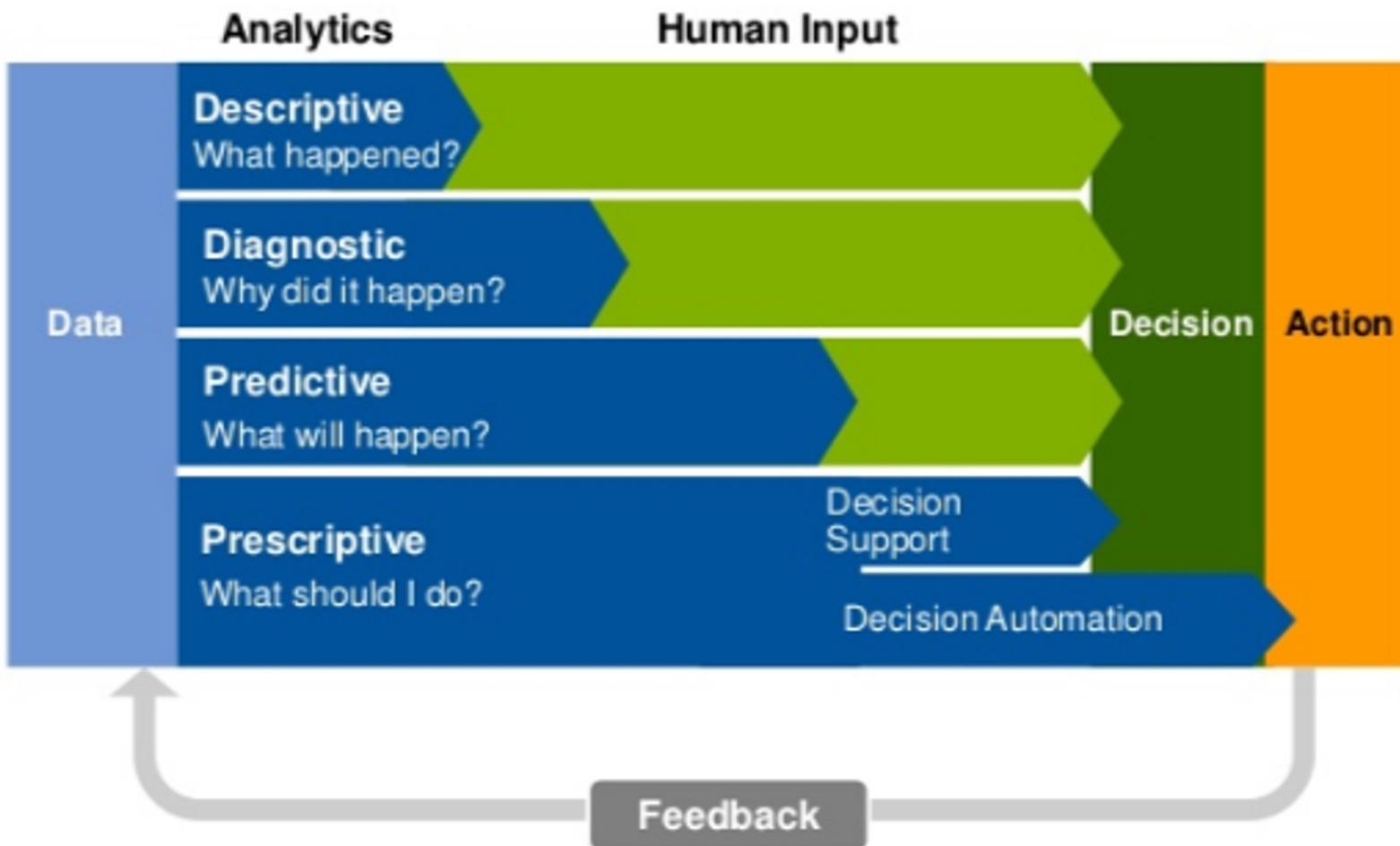


Aurora Health Care system saved \$6 million annually by using prescriptive analysis to reduce readmission rates by 10%.

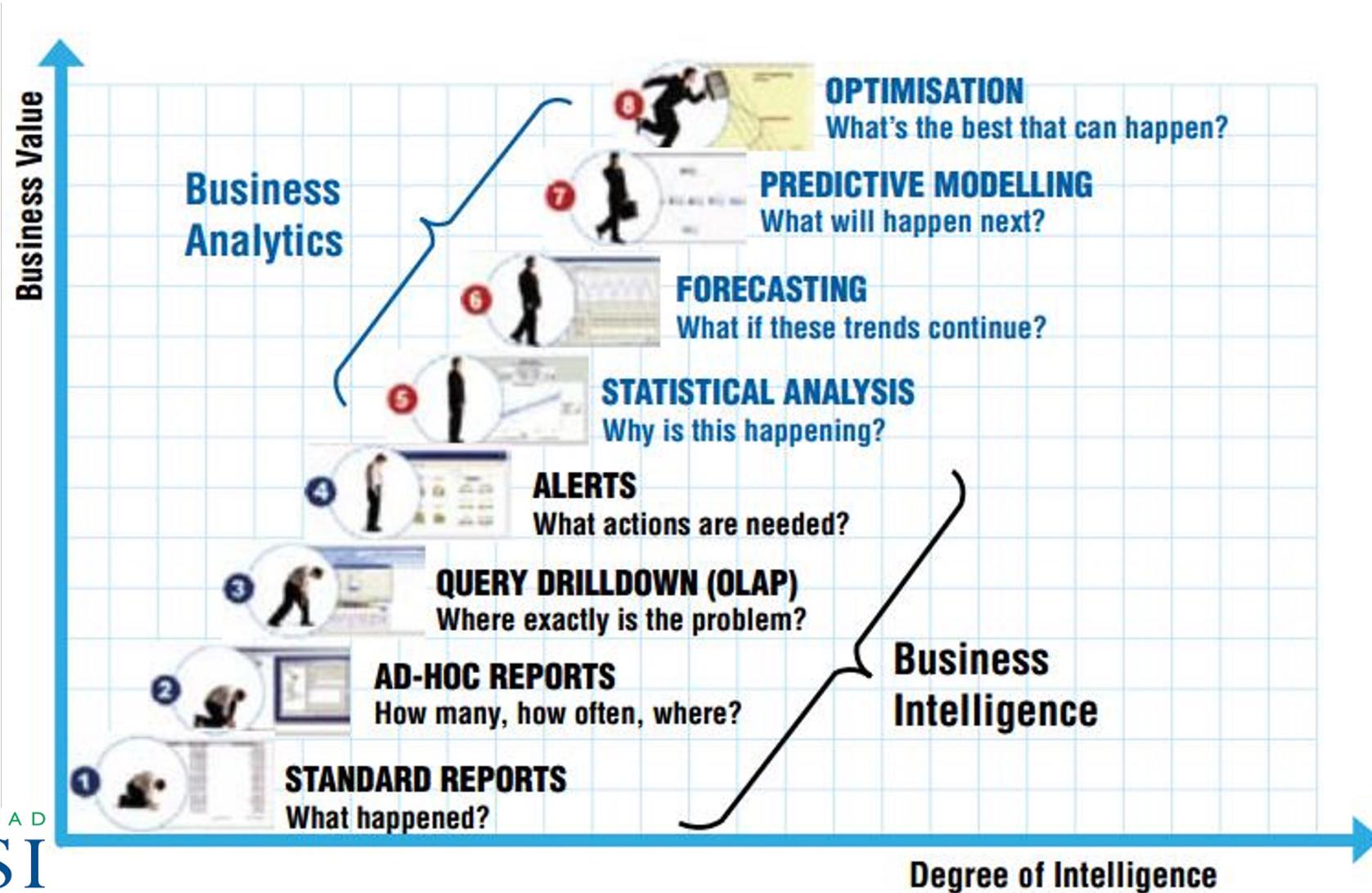


ANALÍTICA EN LA INDUSTRIA

Analytics Continuum



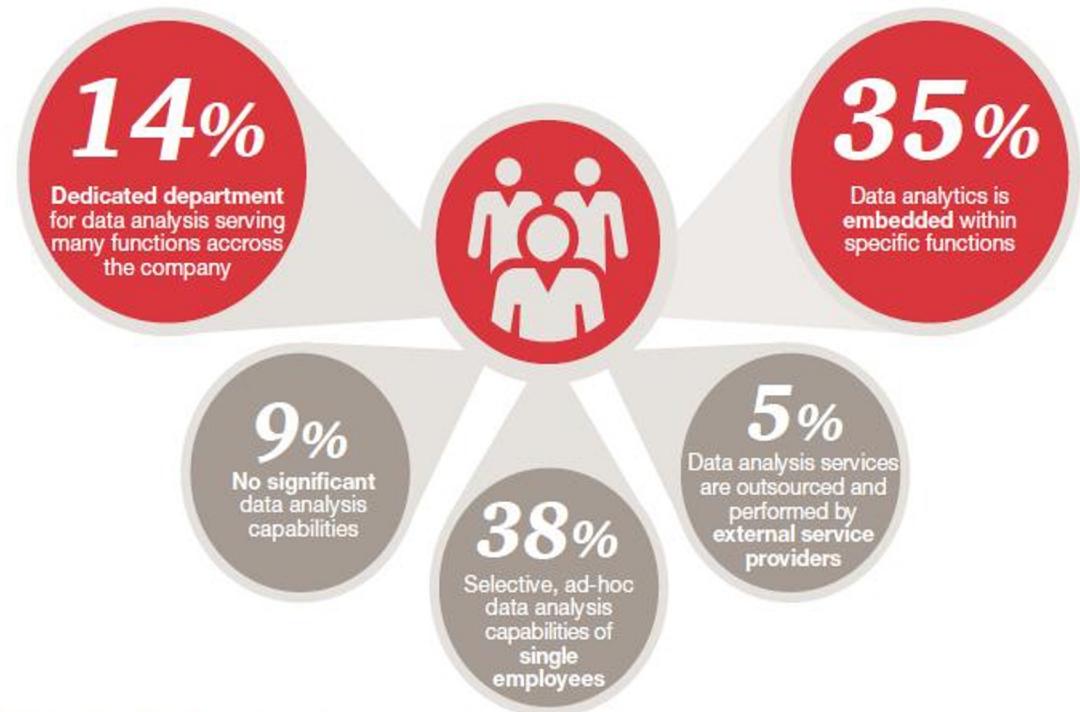
ANALÍTICA EN LA INDUSTRIA



ANALÍTICA EN LA INDUSTRIA

Figure 11: Nearly half of companies still need to develop a robust organisation that supports data analytics excellence

Muchas compañías todavía no están completamente preparadas para la toma de decisiones basadas en datos



Note: Answers shown are rounded

Q: How are data analysis capabilities organised in your company?



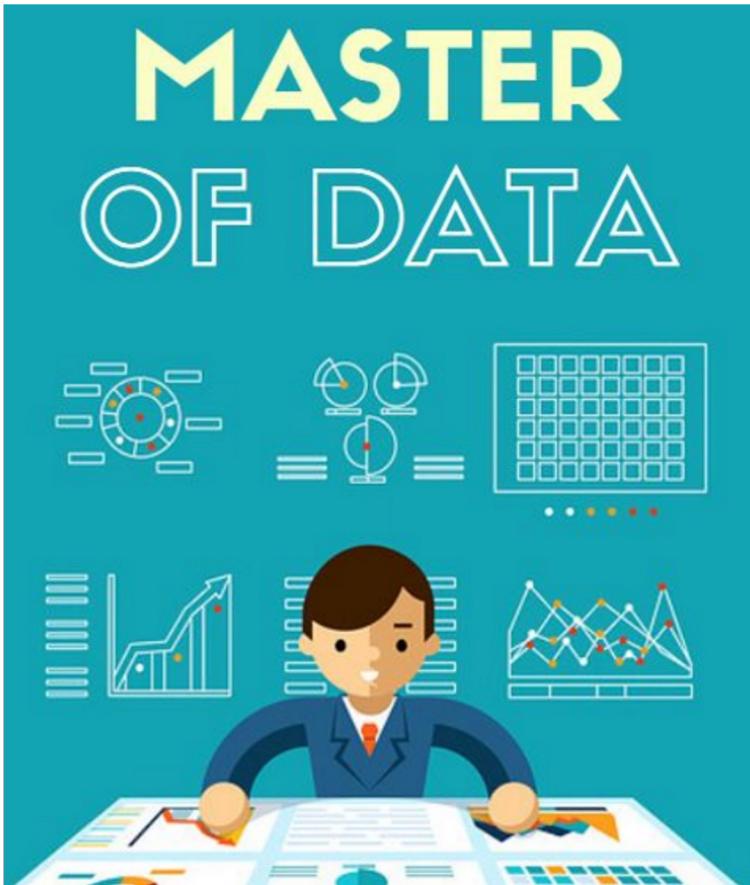
ANALÍTICA EN LA INDUSTRIA

Para tener mayor probabilidad de éxito, un proyecto de analítica debe conjugar:

- Los **datos**: calidad, cantidad, relevancia
- El **talento**: Diversidad de competencias necesaria: científicos de datos, estadísticos, ingenieros de desarrollo, gerentes de proyectos, **expertos de dominio**
- La **infraestructura**: procesamiento, software, unidades de almacenamiento
- Soporte **organizacional**: comenzando por el CEO y CIO



CHIEF DATA OFFICER



- Se enfoca en los datos como base para :
 - Desarrollar capacidades organizacionales
 - Tomar decisiones
 - Competir
 - Estructurar modelos de negocio novedosos
 - Evaluar modelos disruptivos
- Gestión de equipo multidisciplinario
- No se debe encargar de:
 - Manejar la infraestructura
 - Crear o mejorar los reportes existentes

57% de las grandes compañías (Fortune 1000) ya han reclutado un Chief Data Officer (CIO, Sep 2020)



CHIEF DATA OFFICER

Gobierno de datos

Discover

- Data discovery
- Data profiling
- Data inventories
- Process inventories
- CRUD analysis
- Capabilities assessment

Measure and Monitor

- Proactive monitoring
- Operational dashboards
- Reactive operational DQ audits
- Dashboard monitoring/audits
- Data lineage analysis
- Program performance
- Business value/ROI



Define

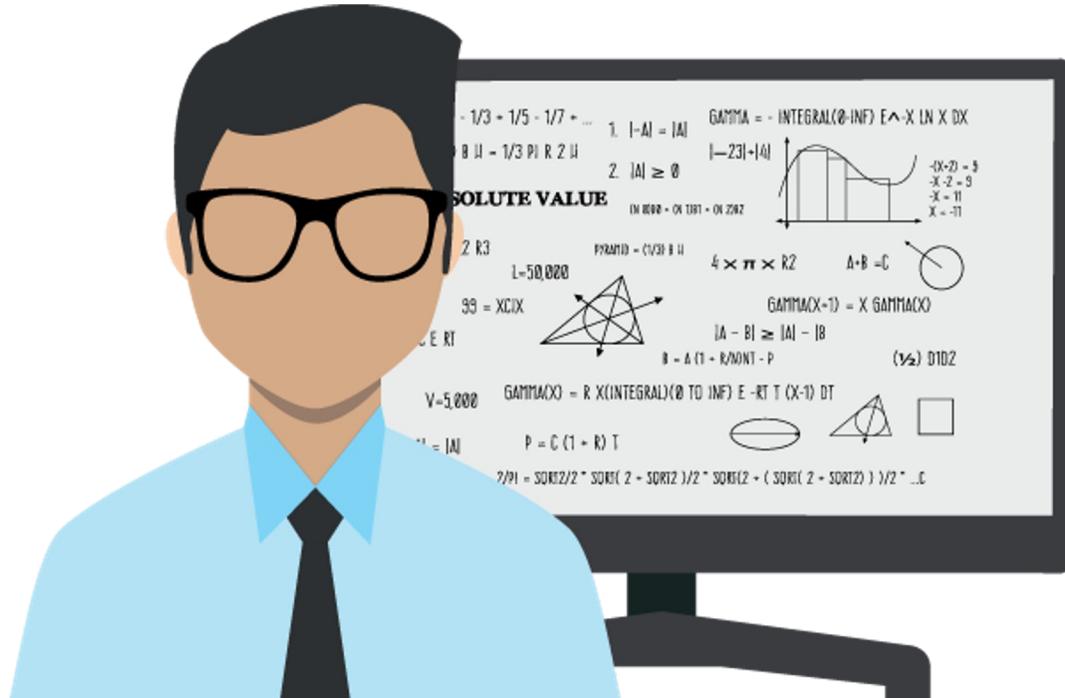
- Business glossary creation
- Data classifications
- Data relationships
- Reference data
- Business rules
- Data governance policies
- Other dependent policies
- Key Performance Indicators

Apply

- Automated rules
- Manual rules
- End to end workflows
- Business/IT collaboration



CIENTÍFICO DE DATOS



Senior data scientist salary

- Aplican sus capacidades analíticas para crear modelos y extraer conocimiento de los datos
- Multidisciplinariedad:
 - Matemáticas y probabilidad
 - Ciencias de la computación
 - Negocios y comunicaciones
 - Creatividad, recursividad, proactividad
- “El trabajo mas sexy del siglo 21”, DJ Patil
- Solo se logra cubrir el 20% de la demanda



INGENIERO DE DATOS



Roles de un ingeniero de datos

- Desarrolla, construye, evalúa y mantiene arquitecturas que aseguran el flujo de los datos entre servidores y aplicaciones:
 - Modelamiento de datos
 - Arquitecturas para big data
 - Sistemas operativos

TRADUCTOR ANALÍTICO



Analytic translator

- Identificar posibilidades de generación de valor a partir de los datos
- Puente entre los problemas de negocio, los modelos de analítica de datos, y las soluciones analíticas, conectando los científicos de datos con los gerentes
- Ayudan a comunicar los resultados de las soluciones analíticas desde el punto de vista del negocio
- Promueven la evangelización de la analítica en la cultura organizacional

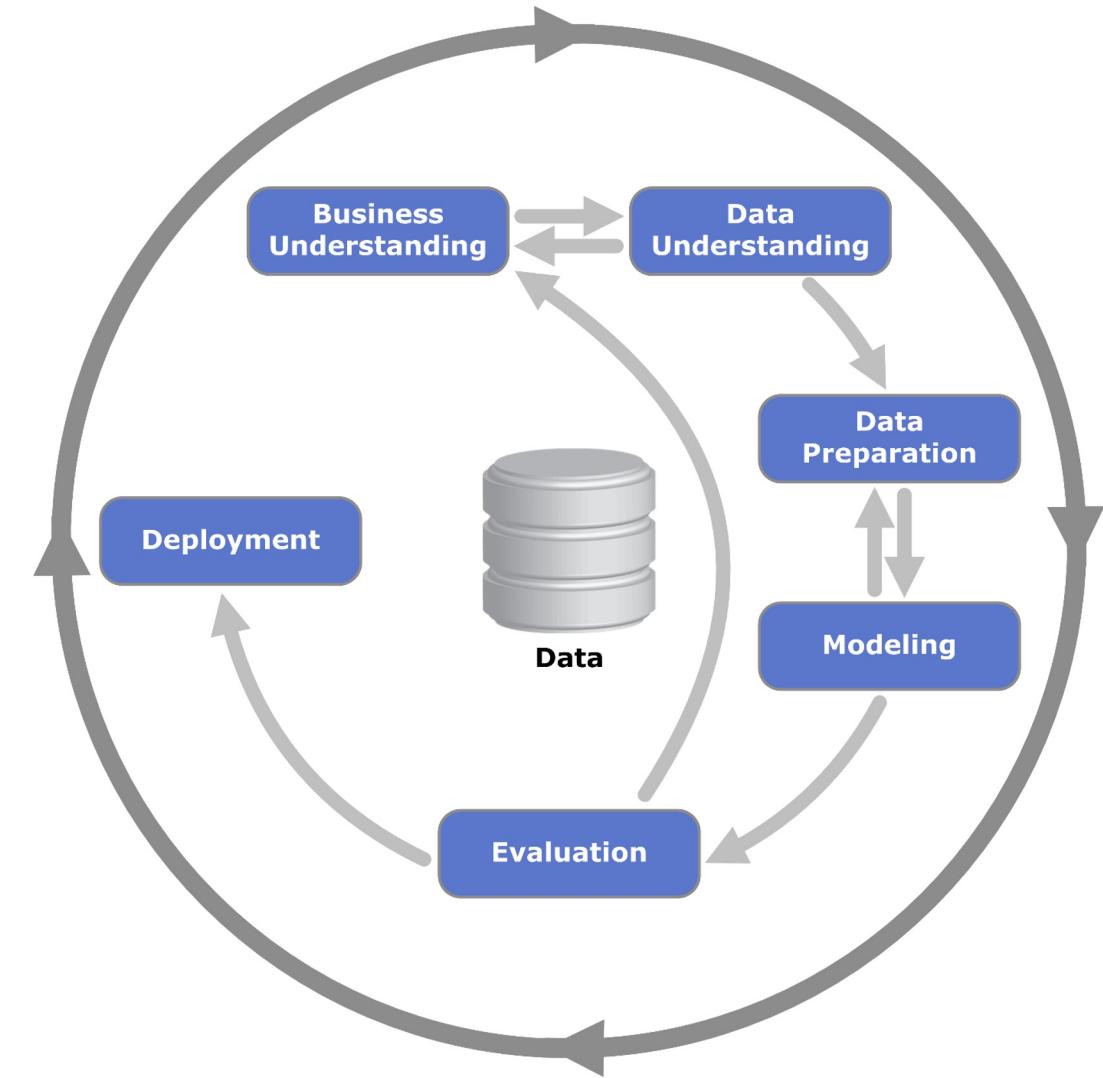


METODOLOGÍA

El proceso de la analítica no es un ciclo de desarrollo de SW!

CRISP-DM

- Cross-industry standard process for data mining:
 - Metodología: identifica 6 fases, sus tareas y dependencias
 - Modelo de proceso: definición iterativa de ciclo de vida
- Se puede llegar a una iteración sin necesariamente haber encontrado una respuesta a la pregunta inicial (sin resolver el problema)
- IBM SPSS incorpora una herramienta de gestión de proyectos que siguen CRISP-DM

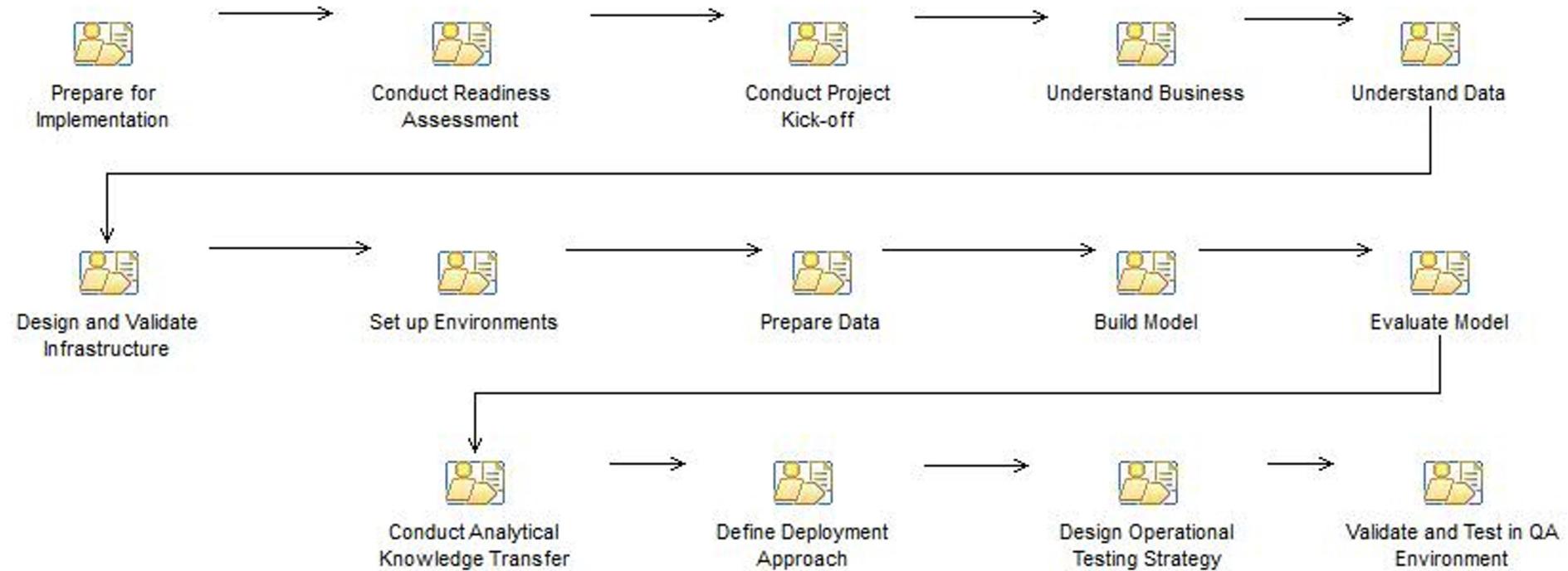


Pete Chapman et al., 2000

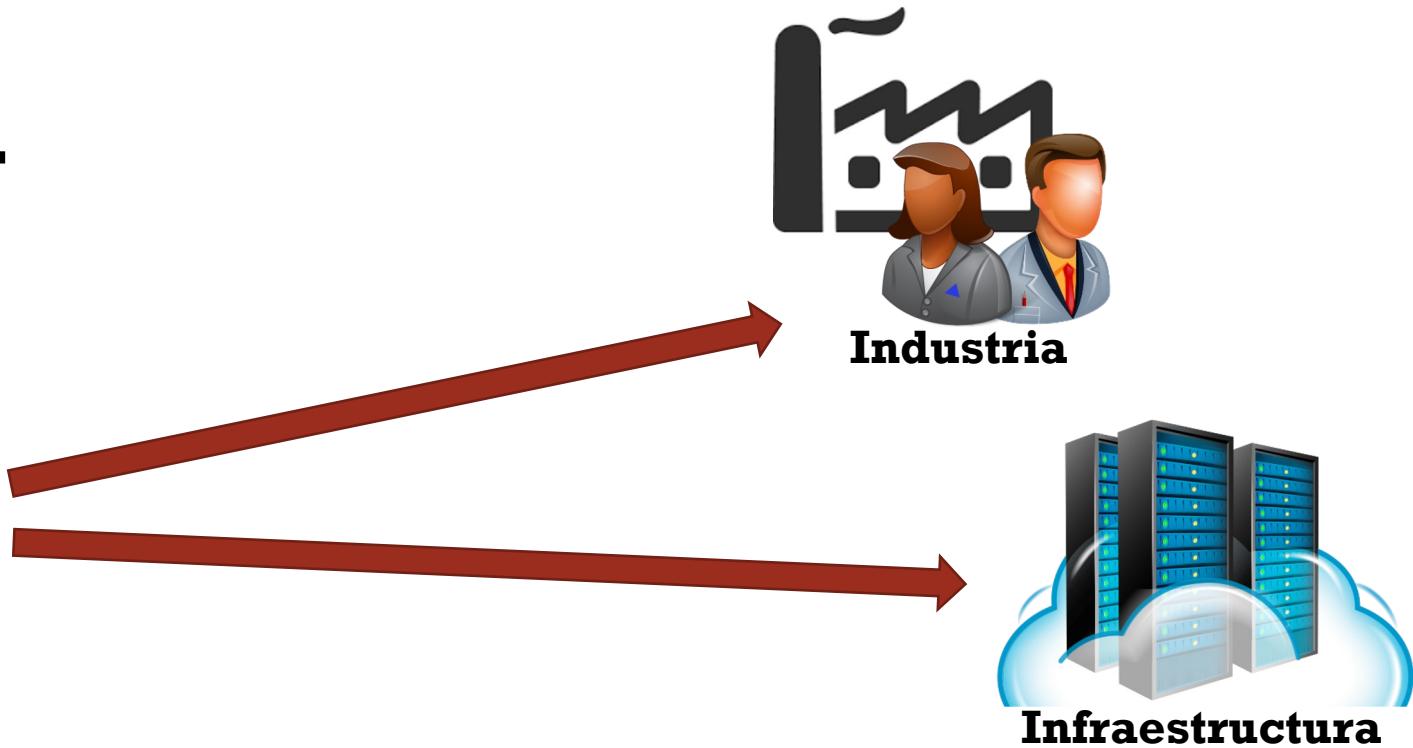


ASUM-DM

- **ASUM-DM**, parte de **CRISP-DM** y la complementa con procesos inspirados en **PMI**.



AGENDA



PROBLEMATICAS TÉCNICAS DE BIG DATA

- ¿Cómo almacenar gran cantidad de información?
- ¿Cómo garantizar la seguridad de la información?
- ¿Cómo garantizar la disponibilidad y la calidad de la información?
- ¿Cómo visualizar la información a tal escala?
- ¿Cómo analizar y convertir los datos en conocimiento?

APACHE HADOOP

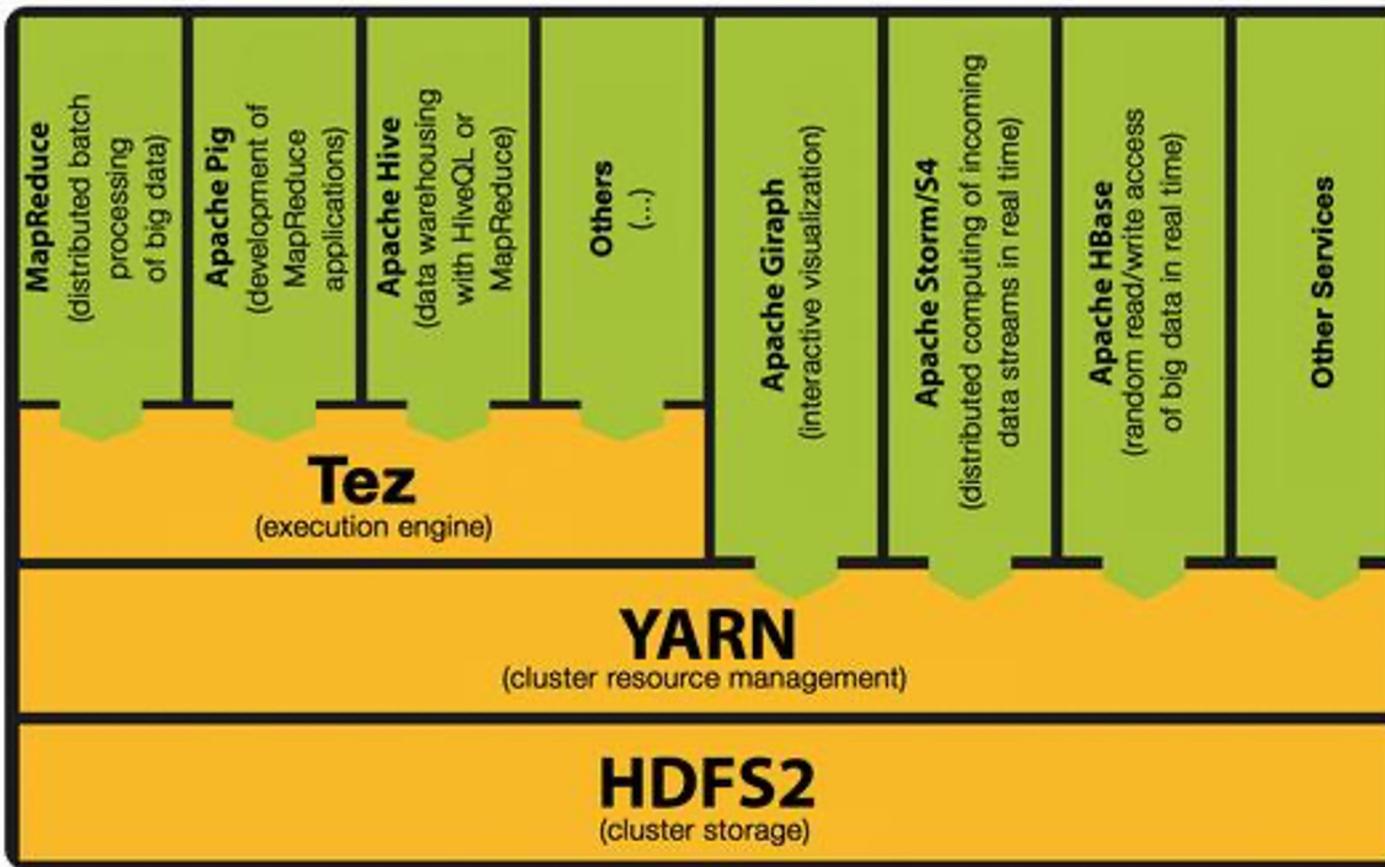


- Map Reduce + Google Filesystem (HDFS) = Nutch (2004) □ Yahoo: Hadoop (2006)
- Open source, licenciado y administrado por Apache Software Foundation

Framework que se encarga de:

- **Almacenar** datos de manera distribuida en los nodos del clúster (HDFS)
- Llevar el **procesamiento** donde se encuentren los datos
- Controlar el **balanceo de carga** de los nodos con respecto a los trabajos de map-reduce que se necesiten □ **escalabilidad**
- **Monitorear** su ejecución, **gestionar y corregir** errores en casos de fallos parciales
- Recolectar y asignar los **resultados intermedios** de la fase de map a nodos que ejecutarán la fase del reduce
- Disponibilizar el **resultado final** del tratamiento, o encadenarlo como datos de entrada de otra tarea map-reduce

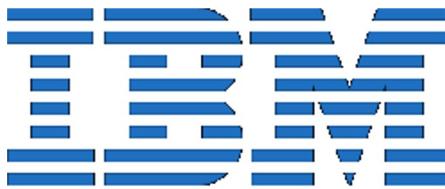
HADOOP ECOSYSTEM - STACK



<http://www.admin-magazine.com/layout/set/print/HPC/Articles/Hadoop-for-Small-to-Medium-Sized-Businesses>



APACHE HADOOP - U



Microsoft



Adobe



Apache



The New York Times

<http://wiki.apache.org/hadoop/PoweredBy>



APACHE HADOOP - DISTRIBUCIONES



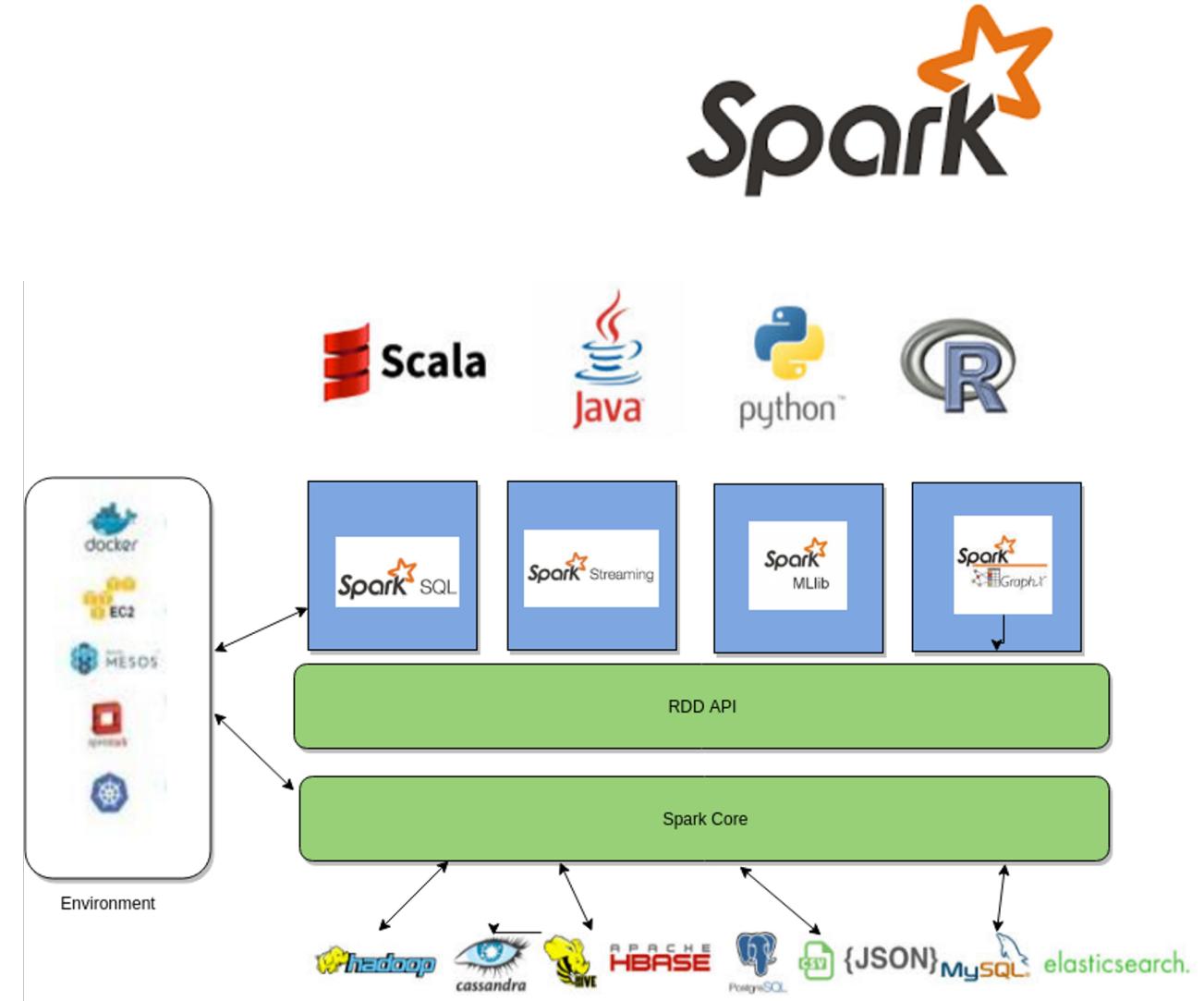
Microsoft Azure



Google Cloud Platform

SPARK

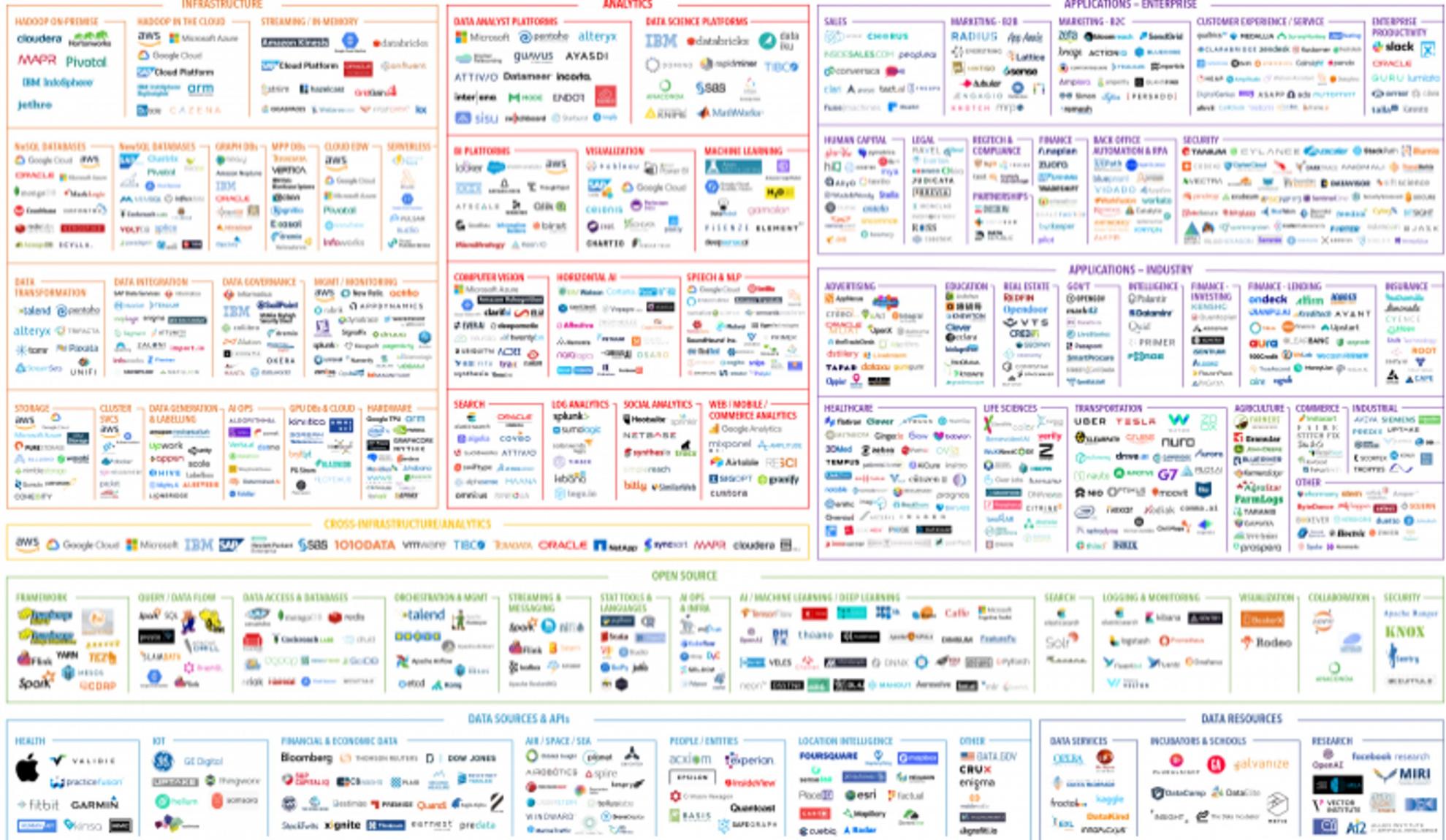
- Framework de procesamiento distribuido
- Alternativa a Map-Reduce, procesamiento en memoria, sin forzar el paso por el sistema de archivos □ Óptimo para MLearning
- Hasta 100 veces mas rápido que Map-Reduce
- APIs en Java, Scala, Python y R



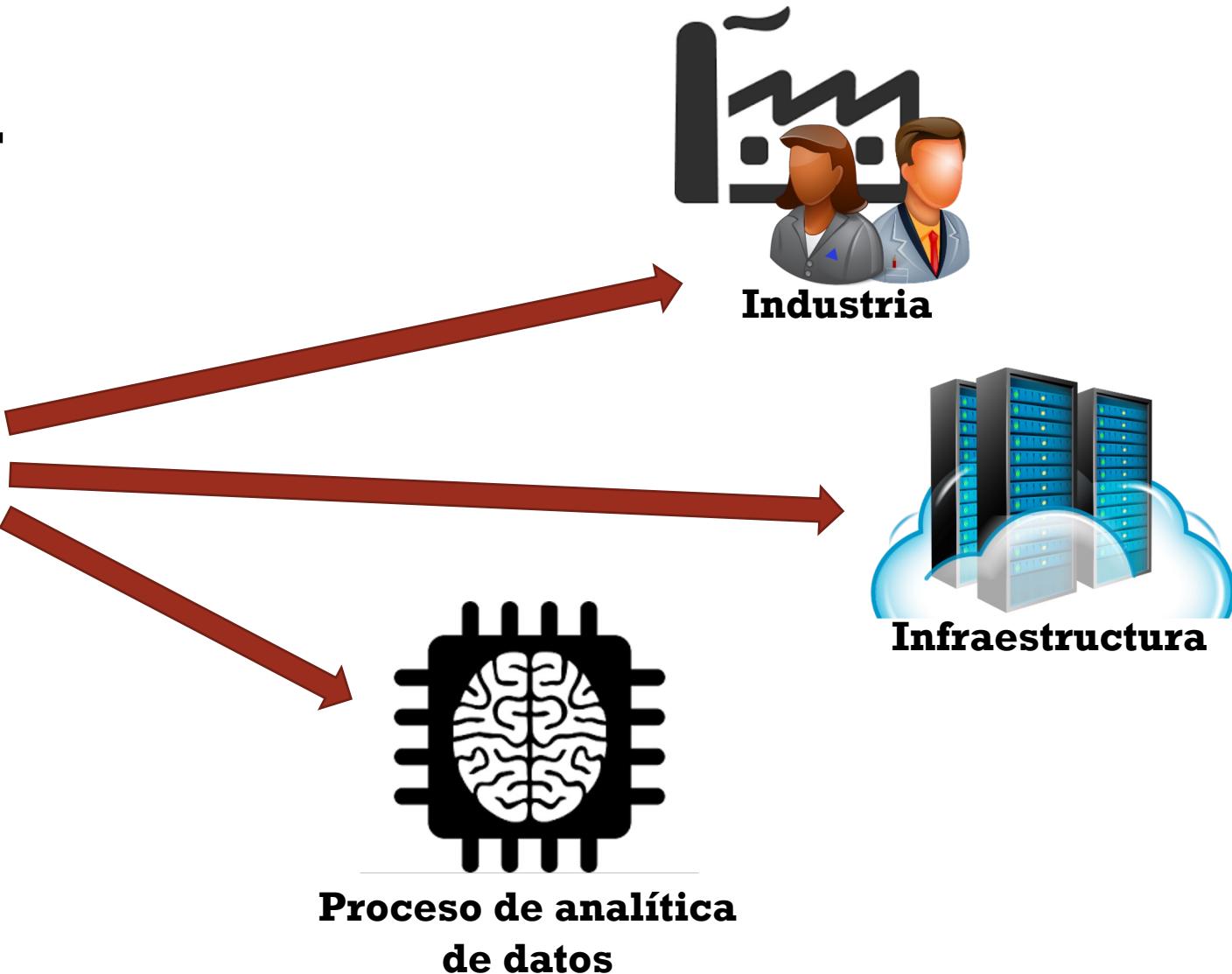
Singh,
2017



DATA & AI LANDSCAPE 2019



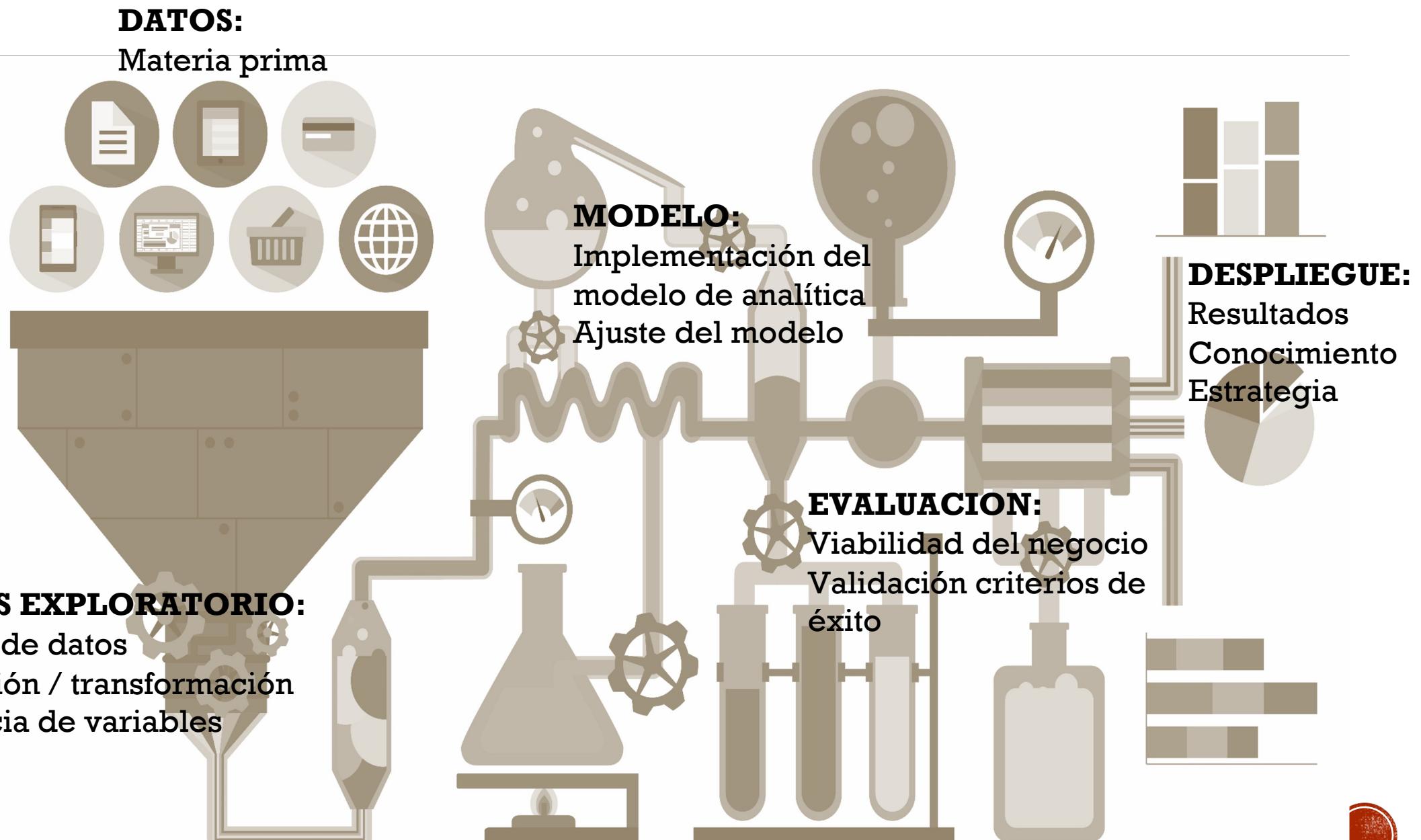
AGENDA



PREGUNTA:

¿Por qué?
¿Cómo?
¿Cuáles?
¿Cuándo?

ANALISIS EXPLORATORIO:
Limpieza de datos
Preparación / transformación
Escogencia de variables



LA PREGUNTA

- Responder a una **pregunta** específica
- Objetivo definido: **mejorar** la toma de decisiones teniendo los objetivos de negocio en mente



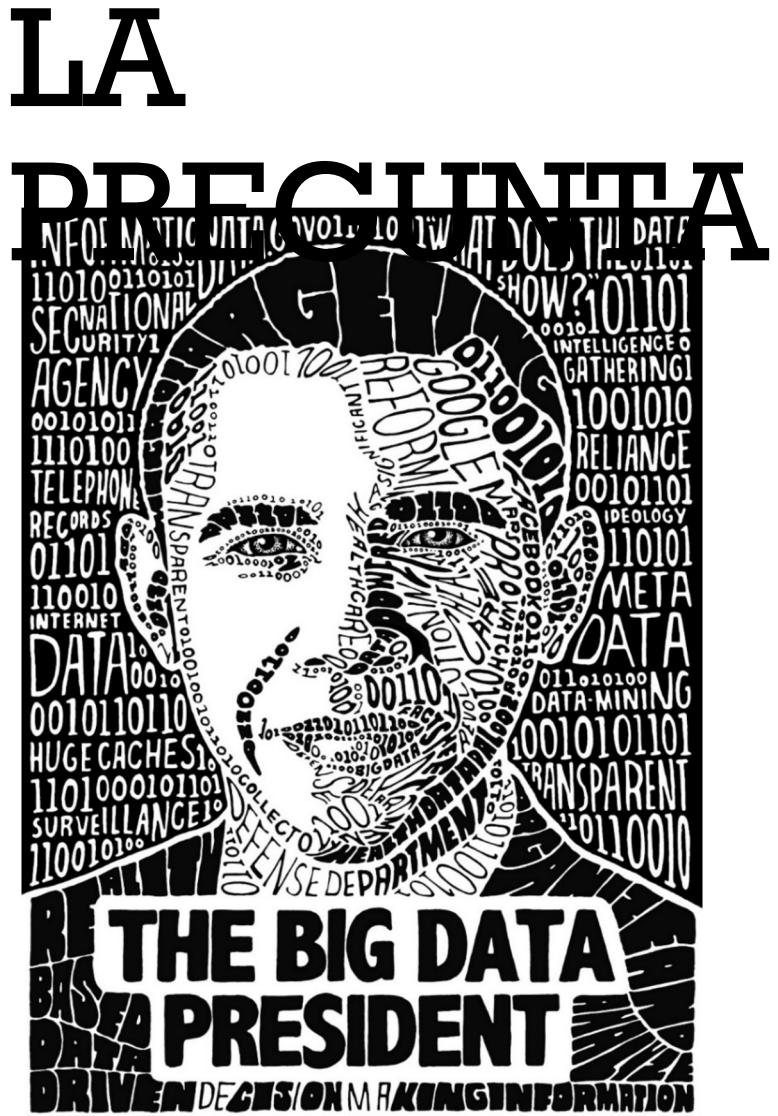
"My team has created a very innovative solution,
but we're still looking for a problem to go with it."

LA PREGUNTA

Clasificación de clientes

¿Cómo puedo identificar los clientes que están esperando un bebé?





¿Cómo maximizar las donaciones por internet de una campaña política?

Ejemplo de acción:

- Inicialmente, el 8,26% de las personas interesadas
- Donación de 21 US\$ en promedio
- A/B Test para mejorar el recaudo de fondos

4 mensajes:

- “Join us now”
- “Sign up now”
- “Sign up”
- “Learn more”

6 soportes visuales:

- 3 videos
- 3 imágenes

■ 24 permutaciones presentadas a 300 000 personas

■ Mejor combinación fue del 11.6%

■ 40,6% de mejora en el interés

■ 10 millones de personas accedieron

■ 60 000 000 US\$ adicionales de recaudo

Washington Post, 2013



LA PREGUNTA

Prevención del churn de clientes

¿Cuáles de mis clientes son los mas propensos a dejarme por mi competidor?



LA PREGUNTA

Sistemas de recomendación

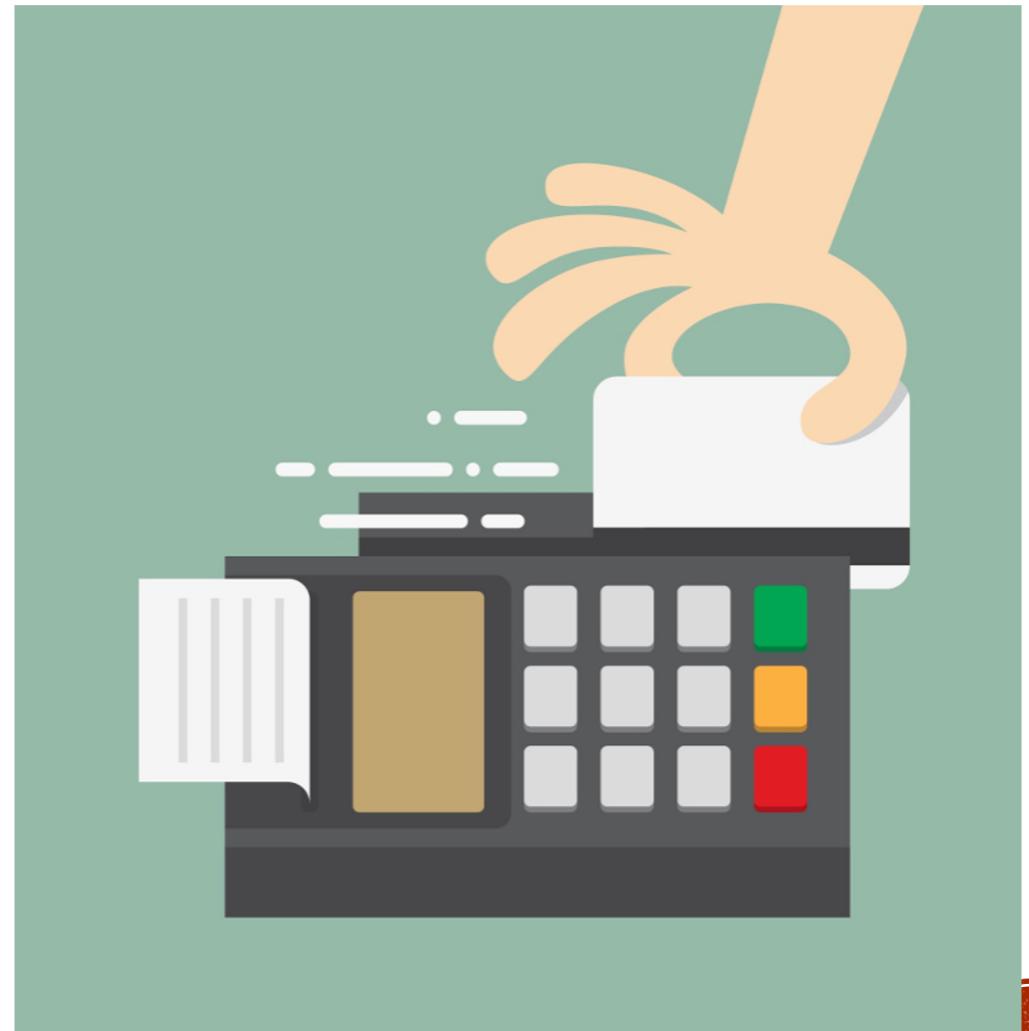
¿Qué películas son las mas apropiadas para este usuario?



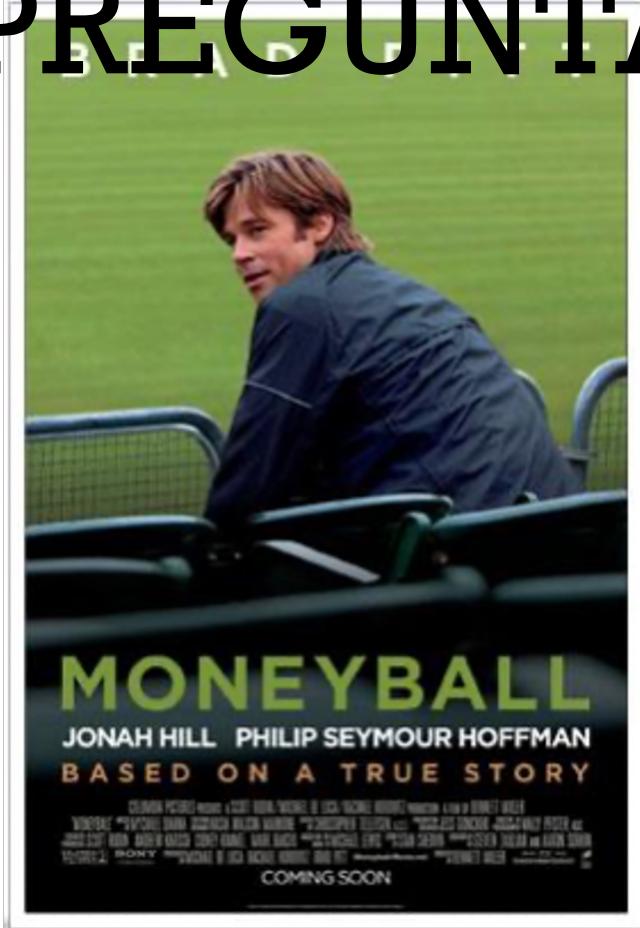
LA PREGUNTA

Detección de fraude de tarjeta de crédito

¿Es esta transacción de tarjeta de crédito
legítima o fraudulenta?

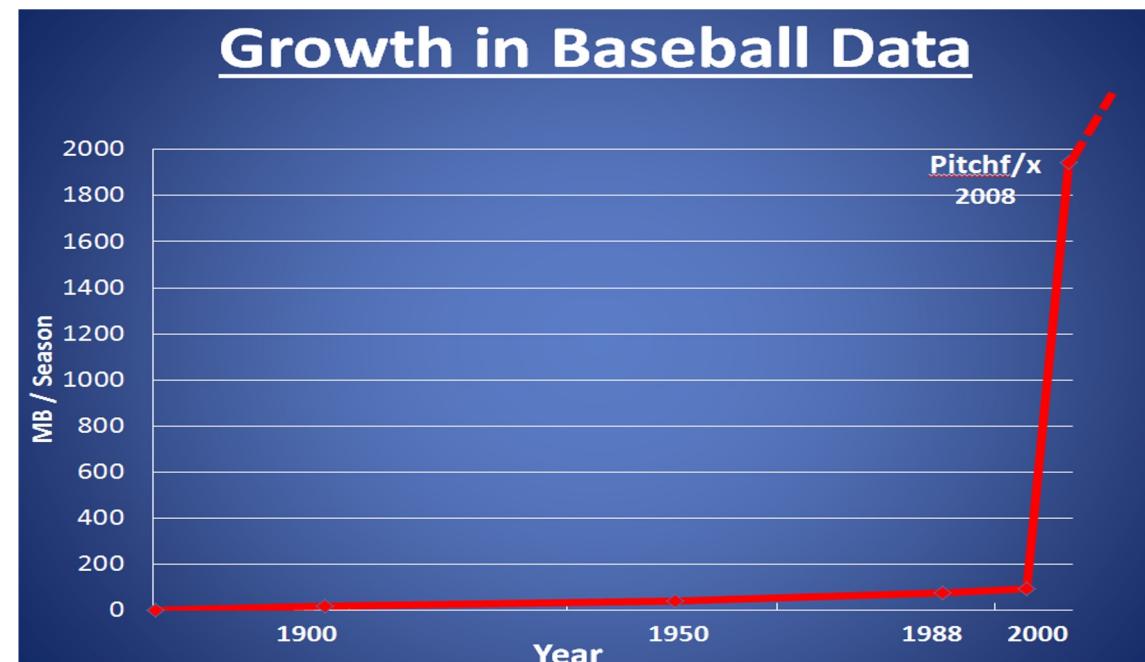


LA PREGUNTA



¿Cómo armar un equipo de beisbol ganador con un pequeño presupuesto?

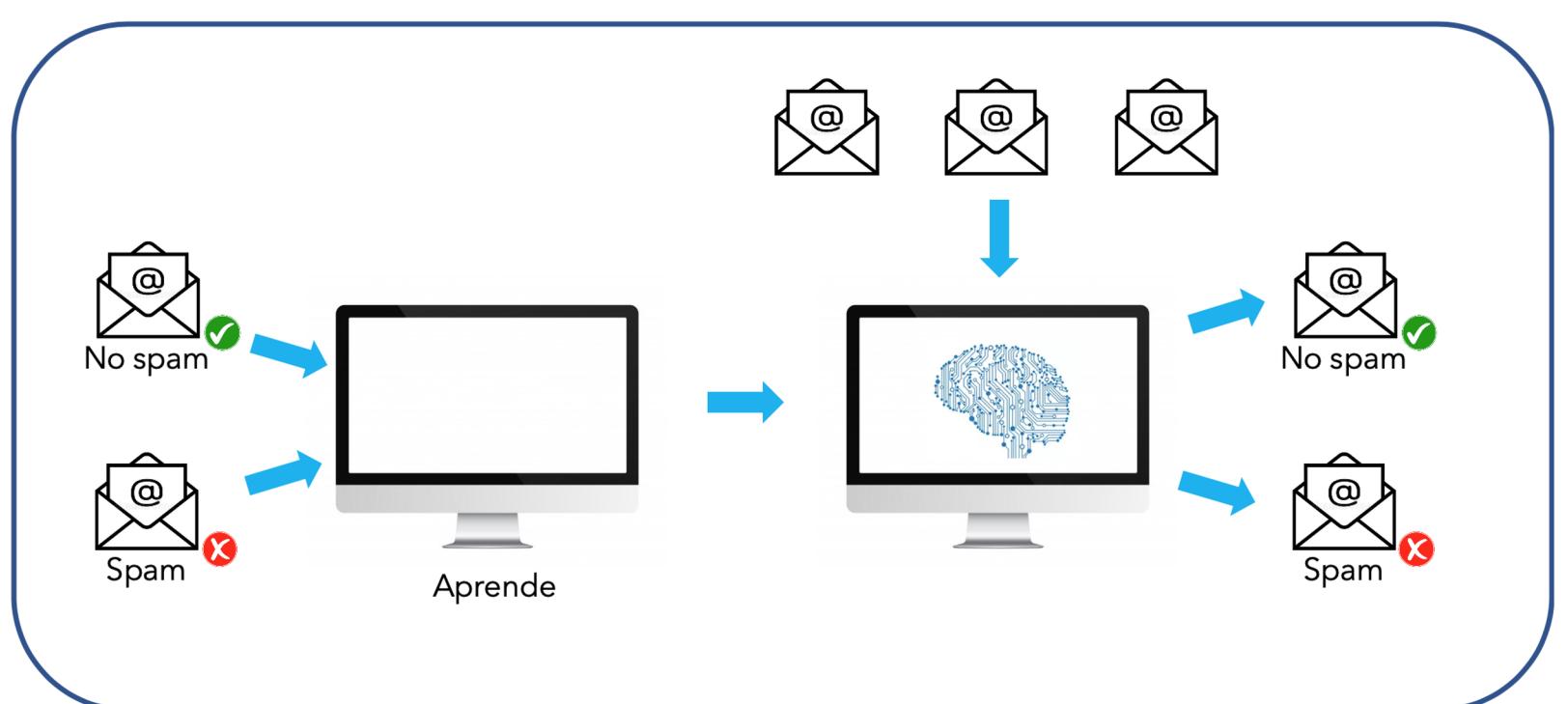
- Optimización de los recursos de un equipo de beisbol basada en datos
 - Pittsburg Pirates
 - Houston Astros
 - Milwaukee Brewers
 - Boston Red Sox



LA PREGUNTA

Reconocimiento de spam

¿Este correo es spam o ham?



LA PREGUNTA

Búsqueda de información

¿Cuáles son los mejores recursos de información para estos términos de búsqueda?



PREGUNTA:

¿Por qué?
¿Cómo?
¿Cuáles?
¿Cuándo?

DATOS:

Materia prima



ANALISIS EXPLORATORIO:

Limpieza de datos
Preparación / transformación
Escogencia de variables

MODELO:

Implementación del
modelo de analítica

EVALUACION:

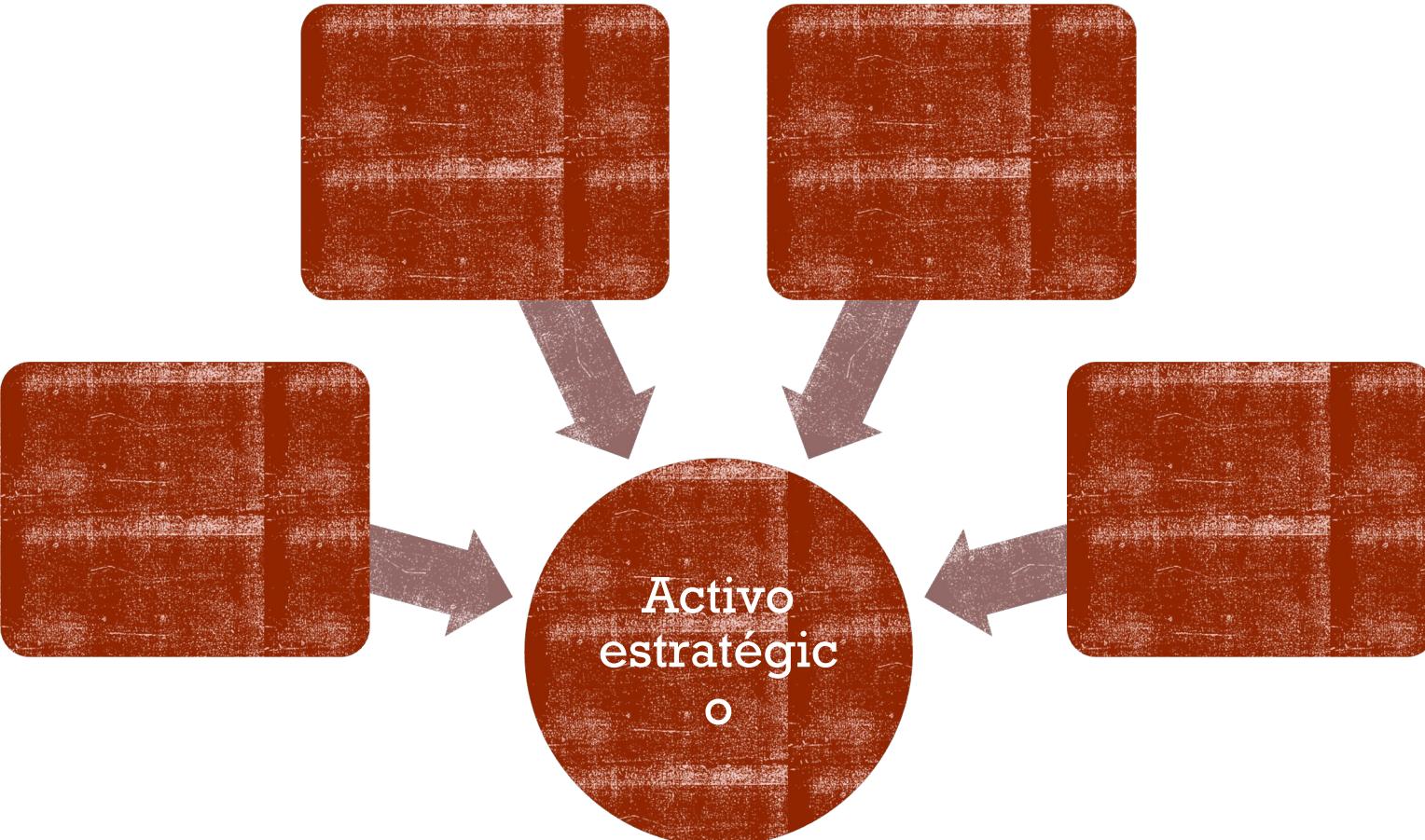
Calidad del modelo
Ajuste del modelo

DESPLIEGUE:

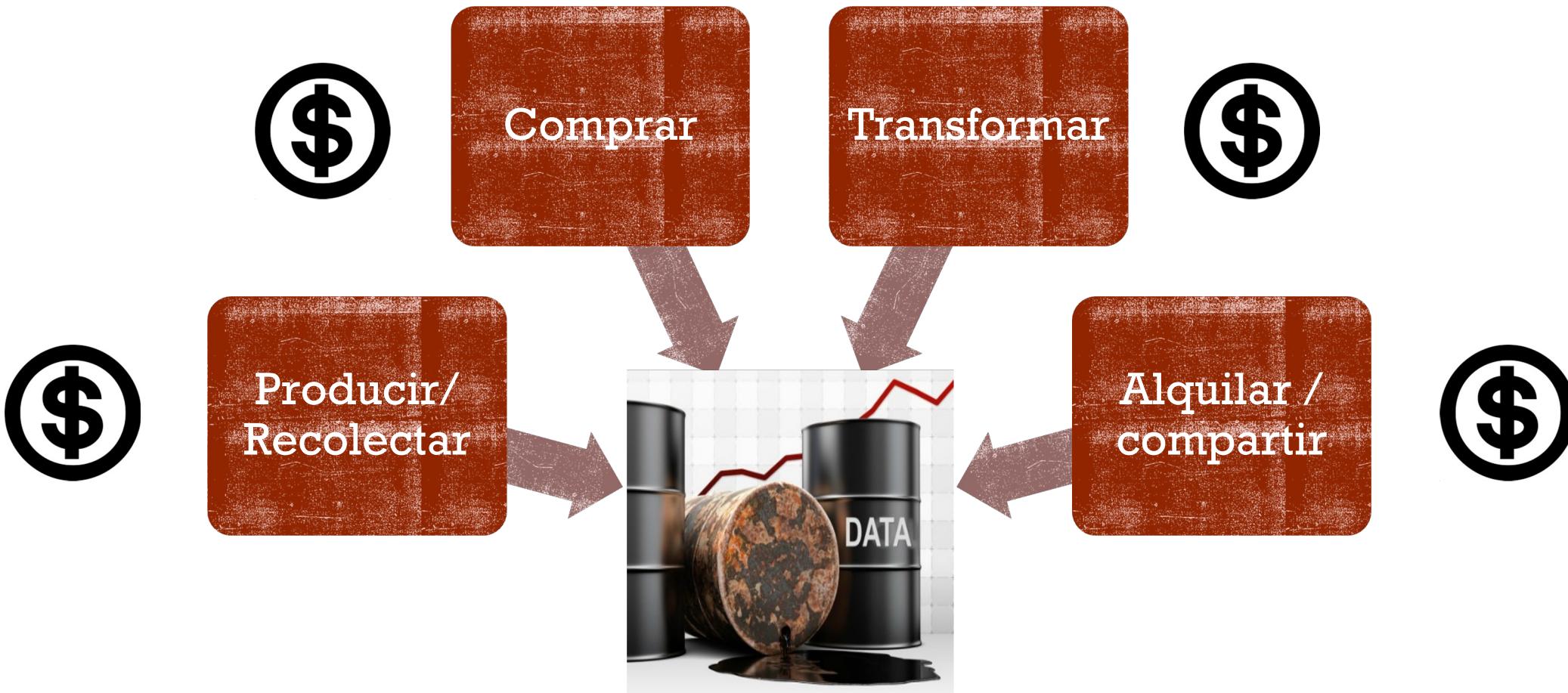
Resultados
Conocimiento
Estrategia



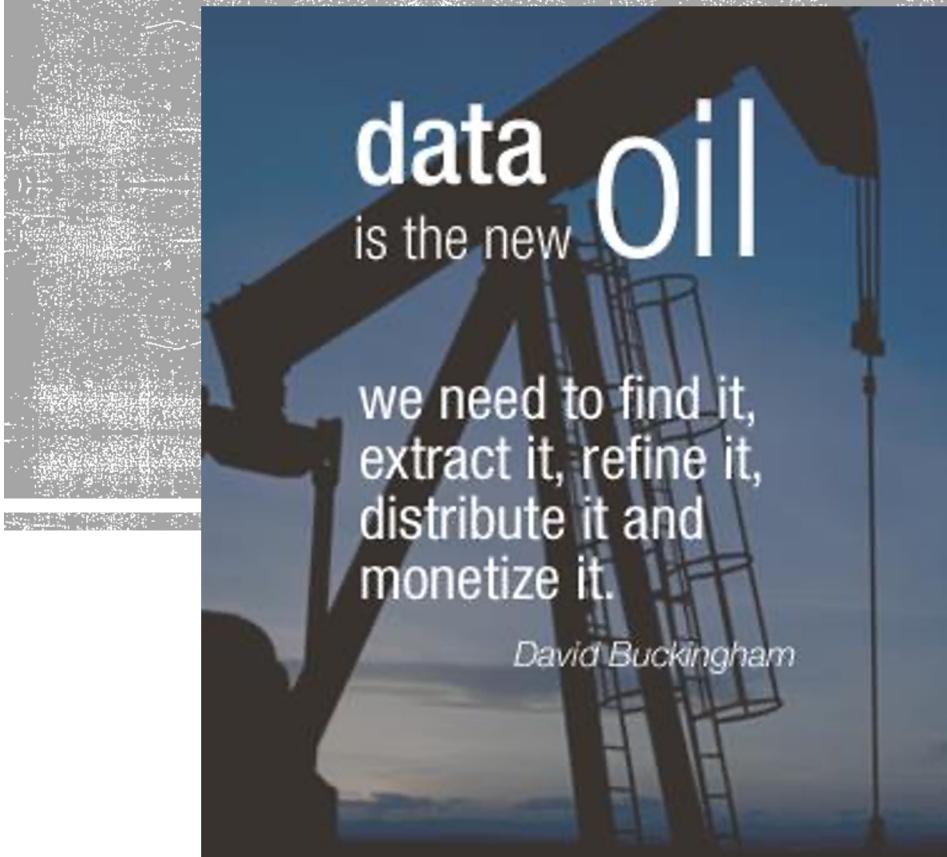
¿Cómo puede una empresa adquirir un activo estratégico?



¿Cómo puede una empresa adquirir un activo estratégico?



LOS DATOS



Materia prima

“Huella dactilar”

Necesitan tratamiento

Activo estratégico

Monetizable

Granularidad



How businesses use data to create value

LOS DATOS



Fuentes de datos

- **Fuentes internas:** ETLs
 - departamentos de producción, finanzas, riesgo, servicio al cliente, mercadeo, RH, ..
- Fuentes externas gratis
- Fuentes gubernamentales
- **Proveedores pagos:** Nielsen, Dunnhumby, Facebook, Twitter, ...
- **IoT:** Internet of Things (internet de las cosas)



LOS DATOS - MARKETING



- 50's: Reportes de ventas de tiendas
 - ¿Qué está pasando en cada tienda?
 - ¿Qué precios debo cobrar en cada tienda?
 - ¿Cuáles de mis productos debo ofrecer en cada tienda y con qué prioridad?
 - ¿Qué tan efectivas son las promociones? ¿los cupones?
 - ¿Qué efecto regional tiene la publicidad en las compras?

No se podía analizar a los clientes individualmente



LOS DATOS - MARKETING



- **60's-70's: compras por catálogo** dirigidas
 - ¿Qué compran en cada casa?
 - ¿Qué catálogo le envío a cada casa?
 - Experimentos con ofertas diferentes, analizando sus impactos
 - ¿Qué relación hay entre los precios y el nivel de compras por catálogo?
 - ¿Qué tipo de publicidad/mensajes son más efectivas?
 - ¿Cuál es el impacto de la frecuencia y el momento de envío de catálogos en las compras?
 - ¿Cuál es el impacto de la variedad de productos en las compras?
 - No se sabe qué compran los clientes por fuera de la publicidad por correo



LOS DATOS - MARKETING



- **80's: scanners en las tiendas + fidelización** marketing dirigido
 - ¿Qué le gusta a cada cliente?
 - ¿Quién compra estos productos?
 - ¿Con qué otros productos se compra este producto?
 - ¿Cuáles son los hábitos de compra de cada cliente? (Dónde, Cuándo)
 - ¿Qué tan susceptible es cada cliente a promociones y descuentos directos?

¿Quiénes son mis clientes?

Falta algo... Qué?



LOS DATOS - MARKETING



▪ 90's y 00's: Internet

- ¿Qué productos ha considerado cada cliente?
- ¿Cuántas veces ha considerado este producto este cliente sin haberlo aún comprado?
- ¿Qué sitios web le interesan a mis clientes?
- ¿Cuáles son los clientes para los cuales esta publicidad es la más indicada?
- ¿Qué sitios web son los más idóneos para esta publicidad?
- Offline + Online data
 - Social networks: ¿Quién conoce a quién?
 - Smartphones: ...
 - IOT (Internet Of Things)



PREGUNTA:

¿Por qué?
¿Cómo?
¿Cuáles?
¿Cuándo?

DATOS:
Materia prima



ANALISIS EXPLORATORIO:
Limpieza de datos
Preparación / transformación
Escogencia de variables

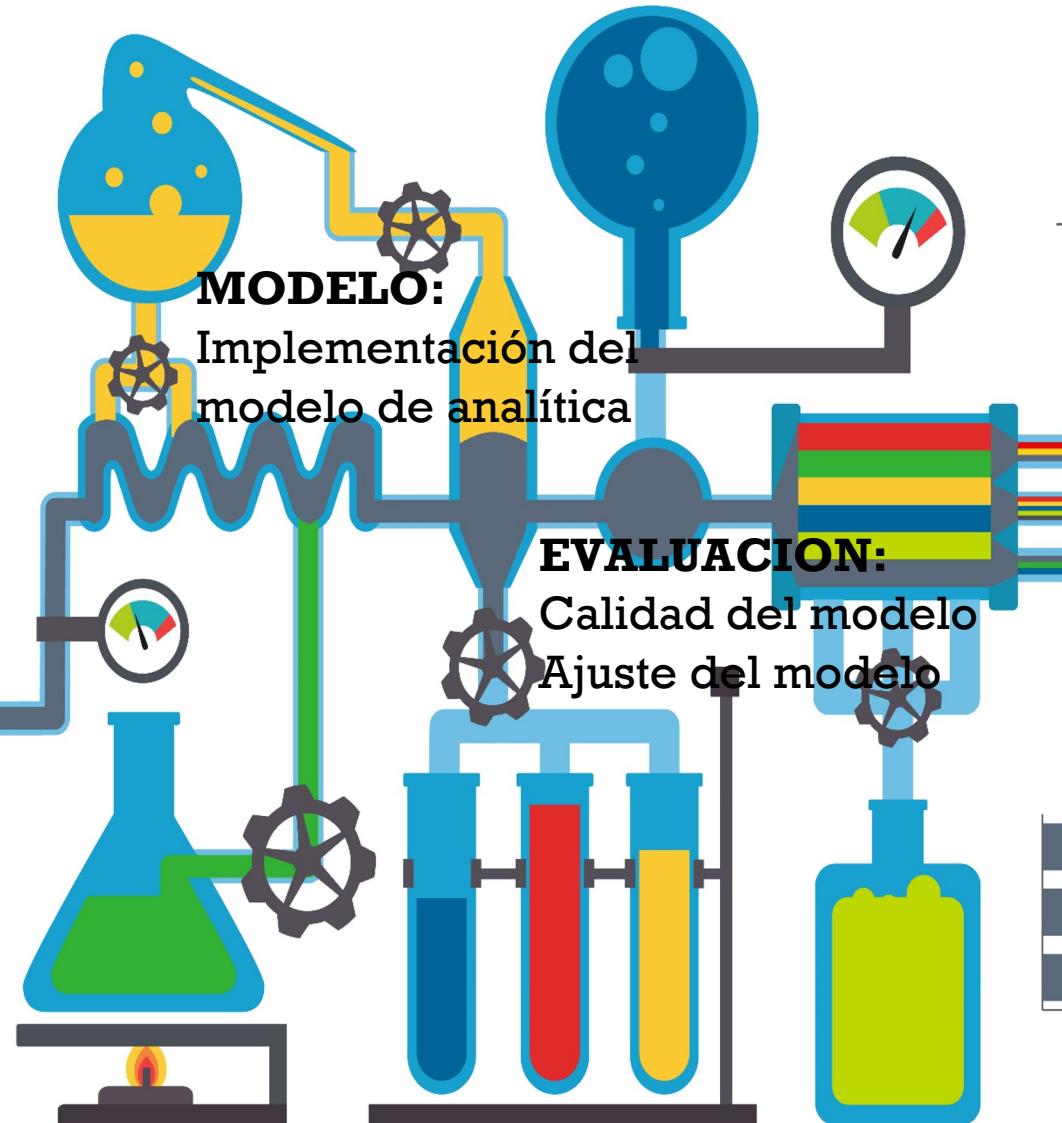
MODELO:

Implementación del
modelo de analítica

EVALUACION:

Calidad del modelo
Ajuste del modelo

DESPLIEGUE:
Resultados
Conocimiento
Estrategia



ANÁLISIS EXPLORATORIO



KEEP
CALM
AND
DO EXPLORATORY
DATA ANALYSIS



¿Ahora qué ya tengo los datos, qué
hago con ellos?

ANÁLISIS EXPLORATORIO

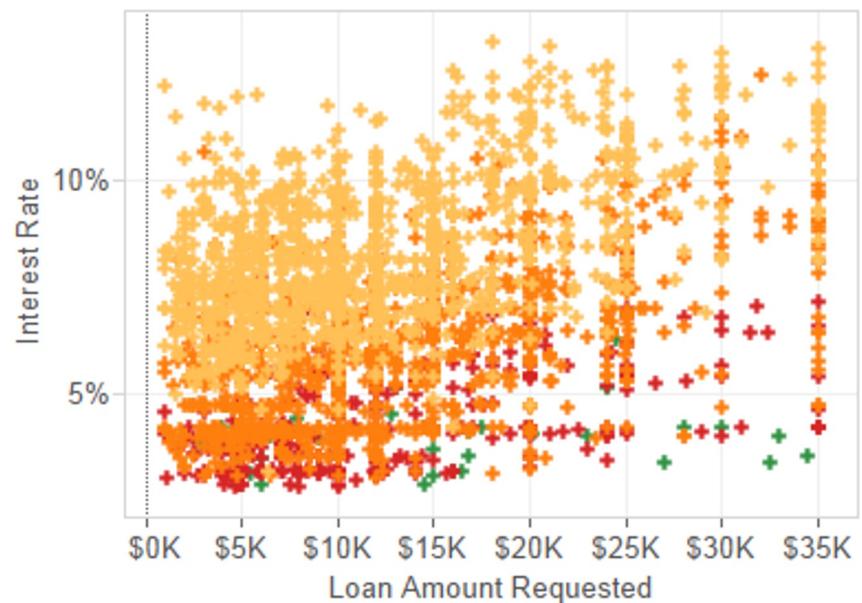


Entender los datos y prepararlos:

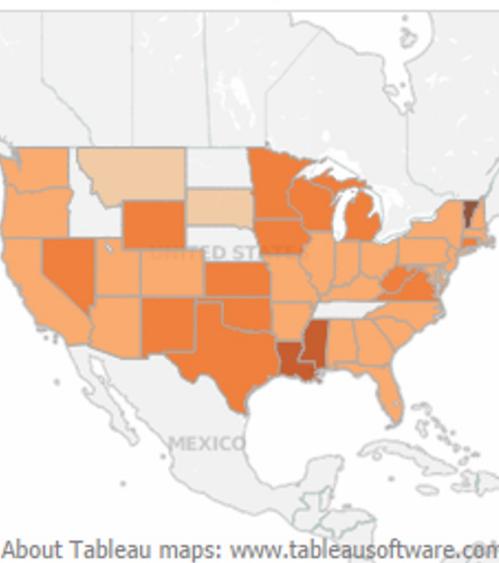
- Rangos, medidas de tendencia central, de dispersión
- Visualización
- Limpieza de datos
 - Analizar valores faltantes y tomar decisiones al respecto
 - Encontrar anomalías/excepciones
 - Transformaciones (standard, log)
- Decidir las variables que serán utilizadas
- Crear nuevas variables si se estima que es necesario



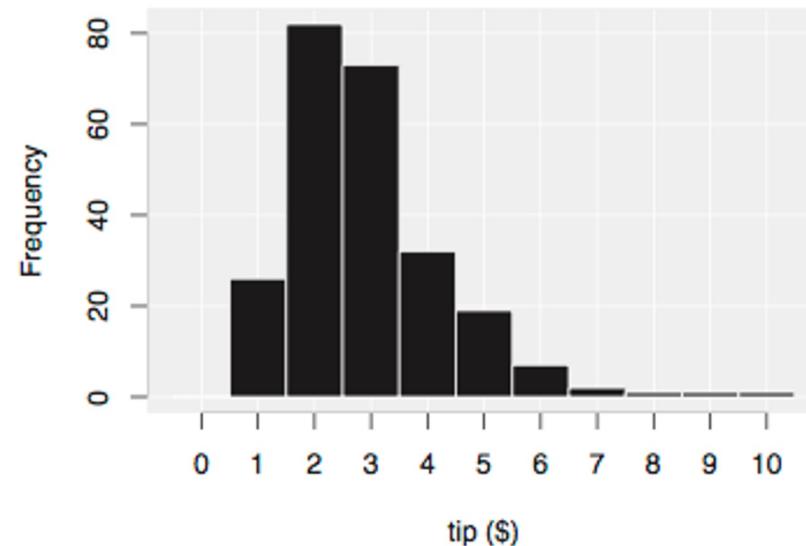
Scatter Plot



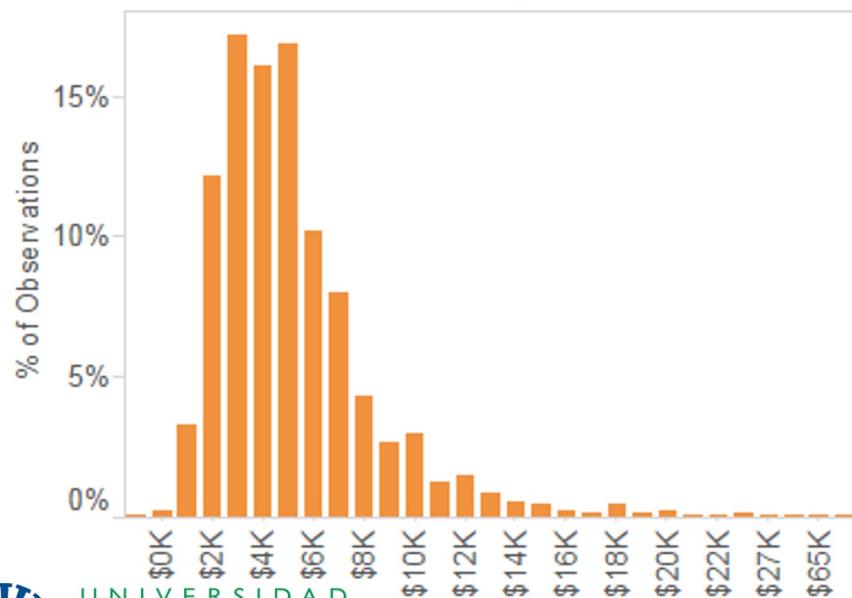
Map



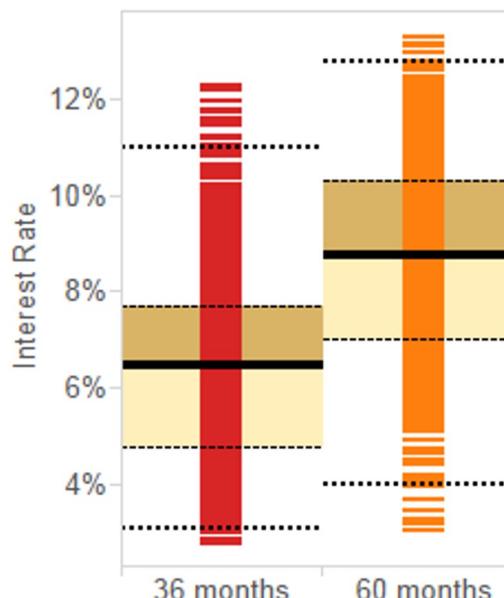
Bin width of \$1



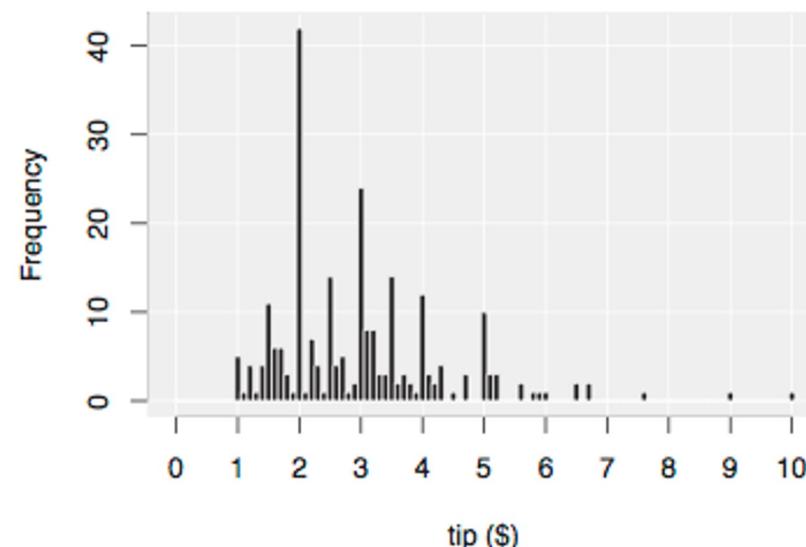
Histogram - Monthly Income

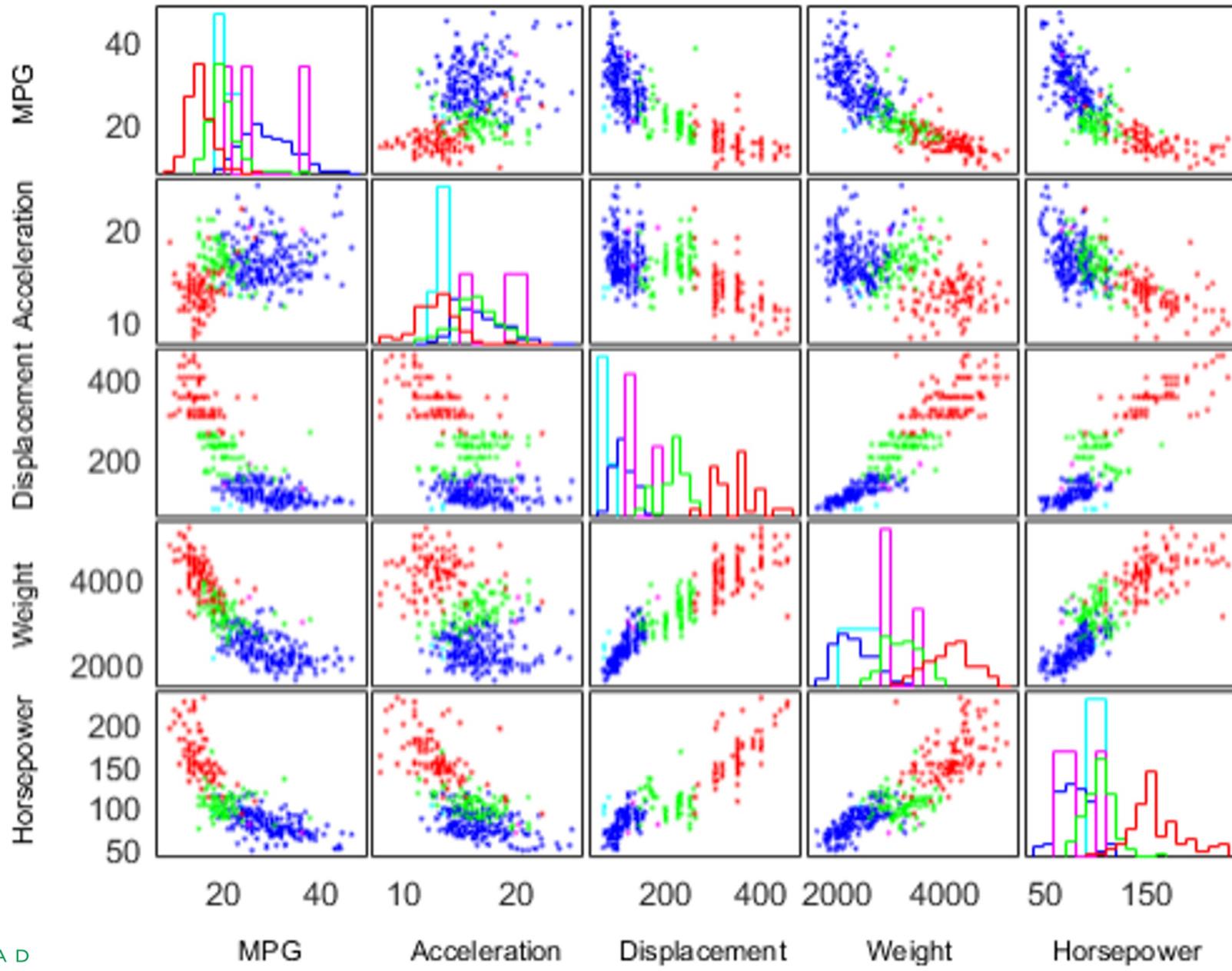


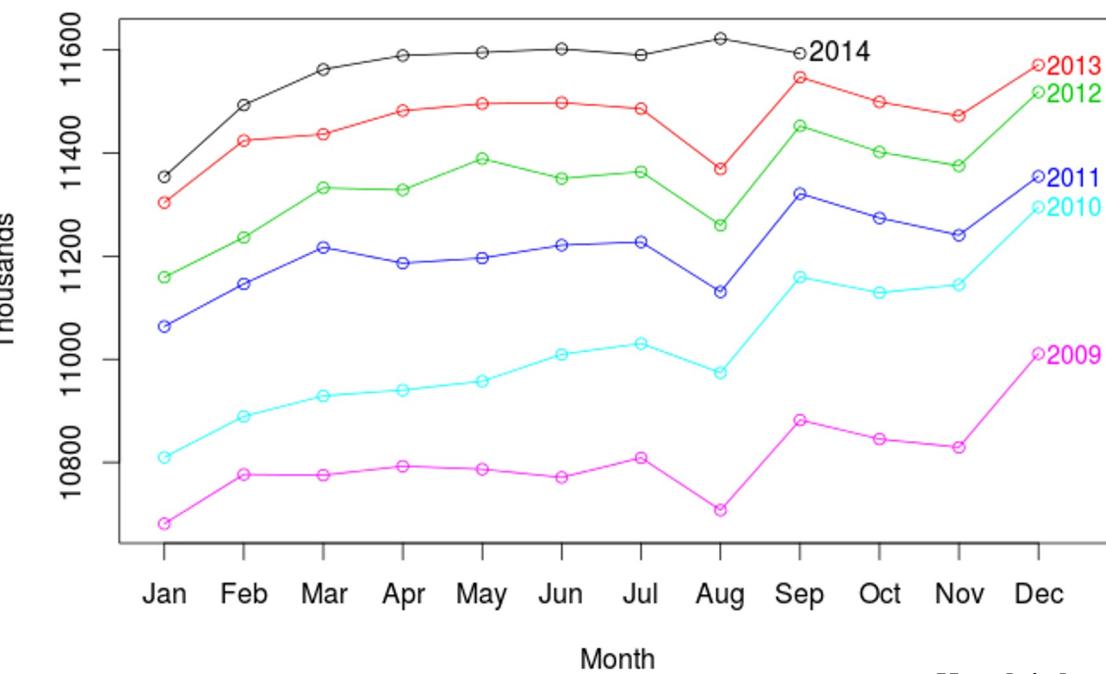
Box and Whisker Plot



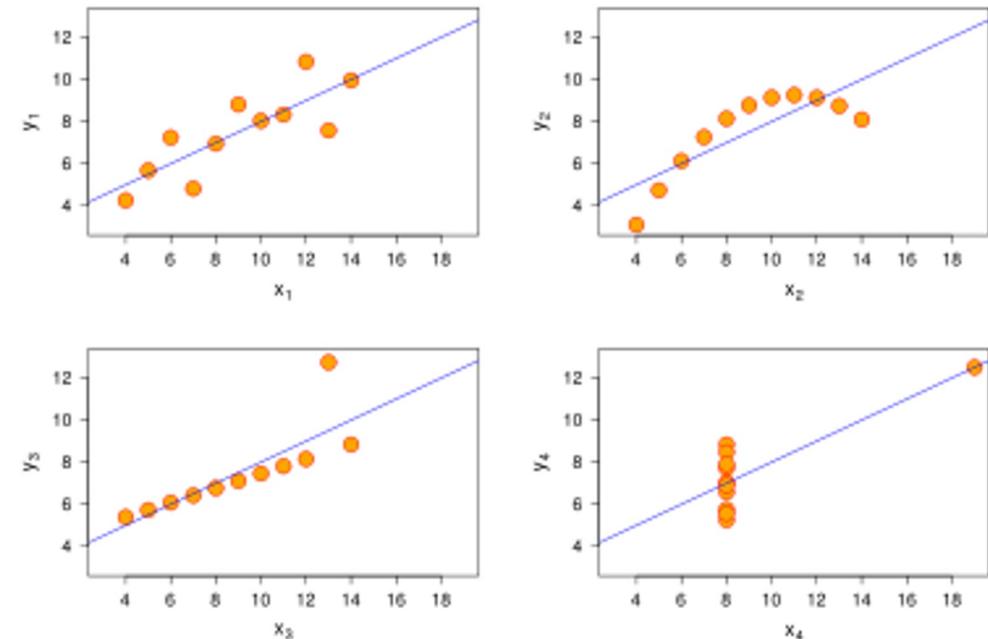
Bin width of 10c







Hyndsiht



I		II		III		IV		
x	y	x	y	x	y	x	y	
10	8,04	10	9,14	10	7,46	8	6,58	
8	6,95	8	8,14	8	6,77	8	5,76	
13	7,58	13	8,74	13	12,74	8	7,71	
9	8,81	9	8,77	9	7,11	8	8,84	
11	8,33	11	9,26	11	7,81	8	8,47	
14	9,96	14	8,1	14	8,84	8	7,04	
6	7,24	6	6,13	6	6,08	8	5,25	
4	4,26	4	3,1	4	5,39	19	12,5	
12	10,84	12	9,13	12	8,15	8	5,56	
7	4,82	7	7,26	7	6,42	8	7,91	
5	5,68	5	4,74	5	5,73	8	6,89	
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03



ANÁLISIS EXPLORATORIO



- ¿Son los datos disponibles suficientes para poder responder la pregunta de investigación?



EJEMPLO DE ANÁLISIS EXPLORATORIO DE DATOS

- 02-EDA-Ejemplo
 - **Desarrollo de un Análisis Exploratorio de Datos utilizando pandas, numpy, matplotlib y seaborn.**

PREGUNTA:

¿Por qué?
¿Cómo?
¿Cuáles?
¿Cuándo?

DATOS:

Materia prima



ANALISIS EXPLORATORIO:

Limpieza de datos
Preparación / transformación
Escogencia de variables

MODELO:

Implementación del
modelo de analítica

EVALUACION:

Calidad del modelo
Ajuste del modelo

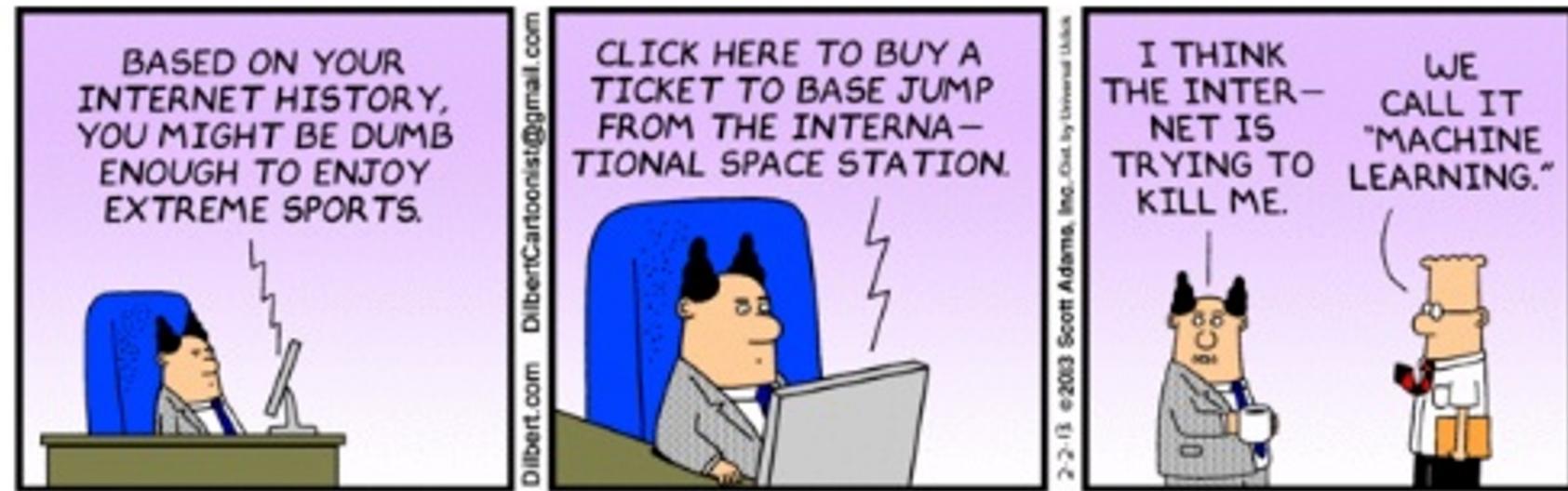
DESPLIEGUE:

Resultados
Conocimiento
Estrategia



ANALÍTICOS

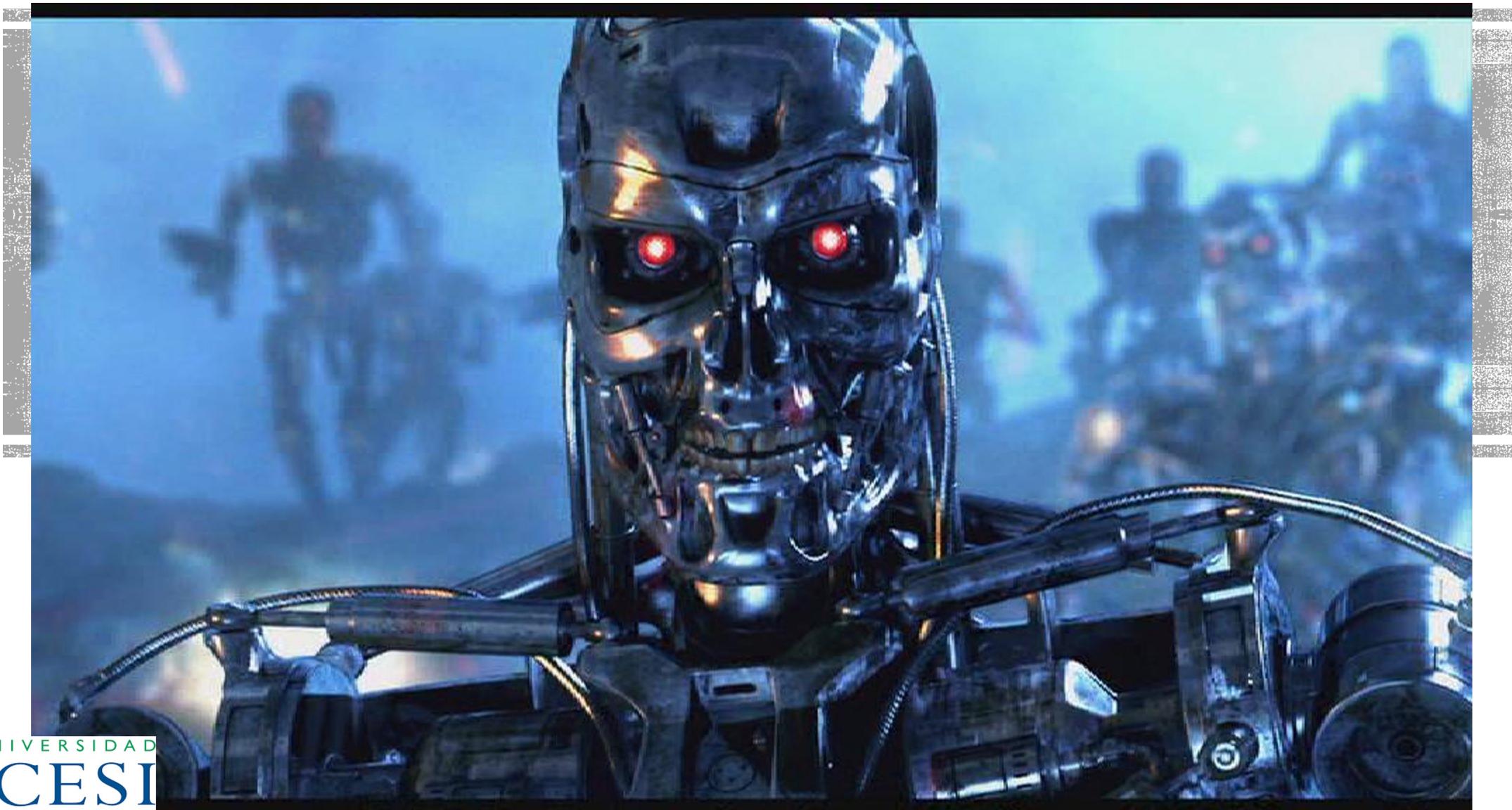
MACHINE



<http://dilbert.com/strips/comic/2013-02-02/>



MACHINE LEARNING



MACHINE LEARNING



Kasparov vencido por Deep Blue



Watson gana Jeopardy

APRENDIZAJE AUTOMÁTICO

- **¿Por qué es necesario?**

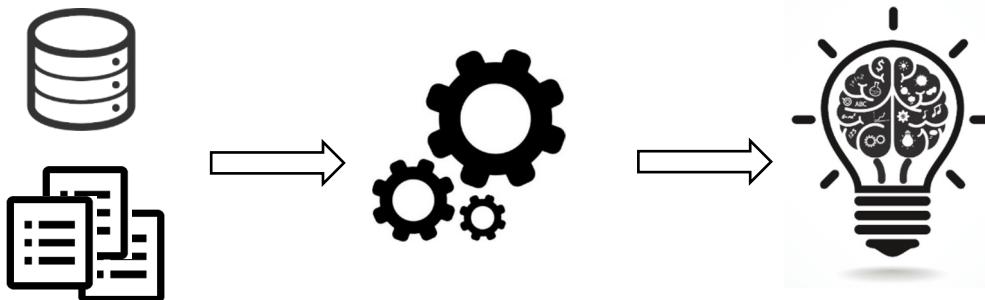
- Tareas complejas extremadamente difíciles de programar
- Poder computacional disponible para tratar grandes volúmenes de datos

Las máquinas tienen que aprender por sí solas

APRENDIZAJE AUTOMÁTICO

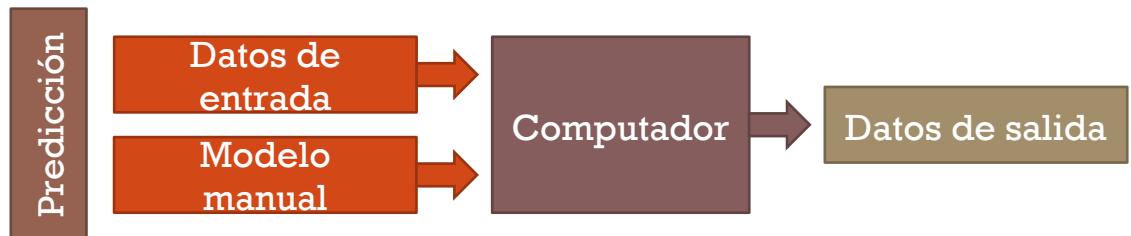
- Definición:

El aprendizaje automático es la ciencia que permite a los computadores aprender, sin ser explícitamente programados¹

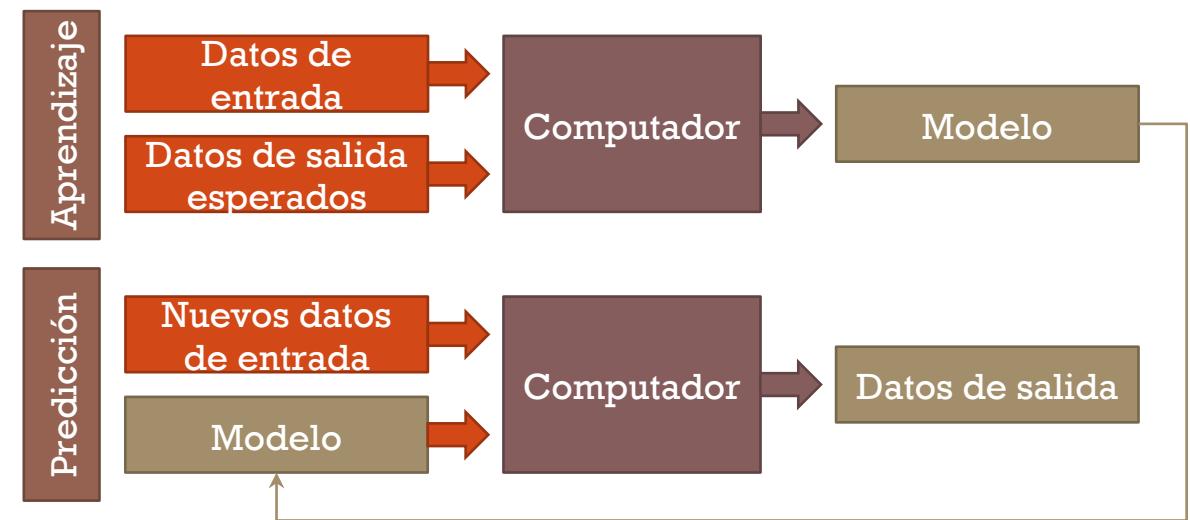


1. Andrew Ng, Stanford University, 2014

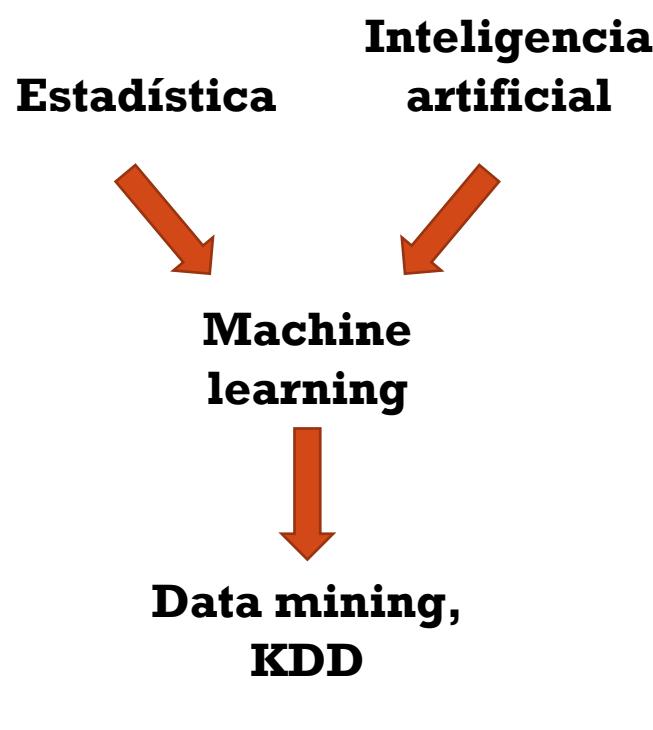
Modelo tradicional



Ciencia de datos



TERMINOLOGÍA



- **Inteligencia artificial:** la ciencia de automatizar comportamientos complejos como el aprendizaje, la resolución de problemas y la toma de decisiones.
- **Data mining:** El proceso de extraer información útil de grandes cantidades de datos complejos.
- **KDD:** Knowledge Discovery in Databases
- **Ciencia de datos:** El proceso de formular una pregunta que puede ser respondida después de haber recolectado, limpiado y analizado datos, y comunicar la respuesta a la pregunta a una audiencia relevante¹
- **Reconocimiento de patrones:** El descubrimiento automático de regularidades en los datos a partir de algoritmos computacionales²

1. Brian Caffo, 2015
2. Christopher Bishop, 2006

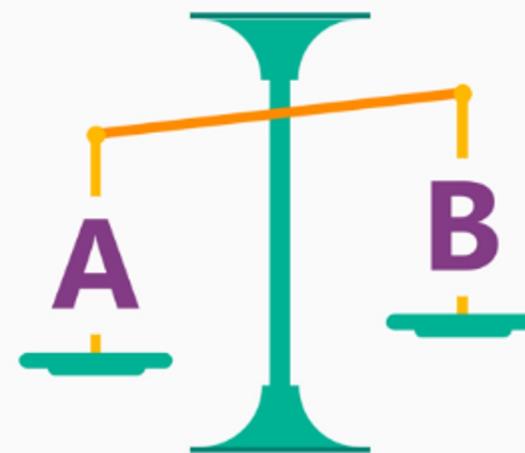


TAREAS DE APRENDIZAJE AUTOMÁTICO

- **¿Qué podemos hacer con los datos?:**
 - **Clasificación:** predecir la categoría de un ítem

Is this A or B?

Classification algorithms

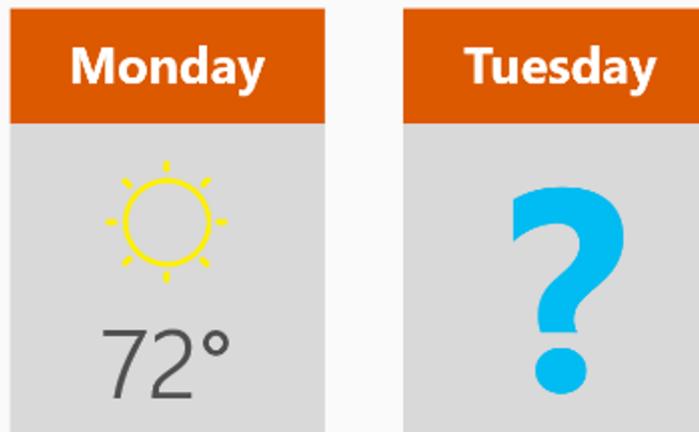


TAREAS DE APRENDIZAJE AUTOMÁTICO

- **¿Qué podemos hacer con los datos?:**
 - **Clasificación:** predecir la categoría de un ítem
 - **Regresión:** predecir un valor continuo

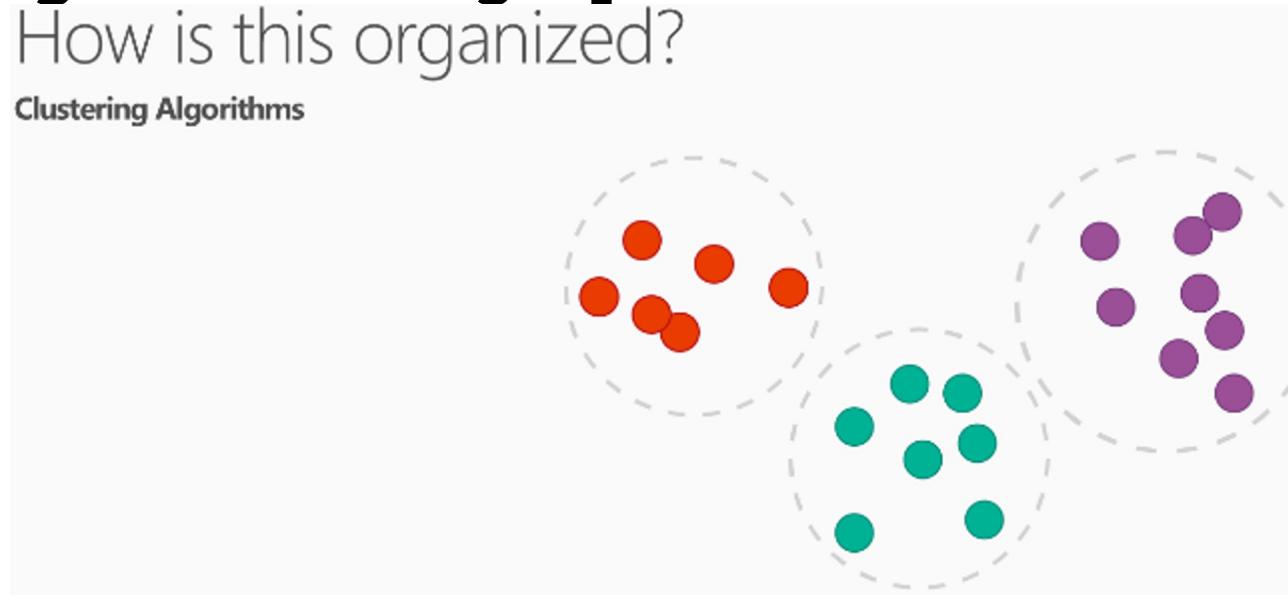
How much? How many?

Regression algorithms



TAREAS DE APRENDIZAJE AUTOMÁTICO

- **¿Qué podemos hacer con los datos?:**
 - **Clasificación:** predecir la categoría de un ítem
 - **Regresión:** predecir un valor continuo
 - **Detección de excepciones:** encontrar anomalías
 - **Clustering:** encontrar grupos de elementos similares



TAREAS DE APRENDIZAJE AUTOMÁTICO

- **¿Qué podemos hacer con los datos?:**
 - **Clasificación:** predecir la categoría de un ítem
 - **Regresión:** predecir un valor continuo
 - **Detección de excepciones:** encontrar anomalías
 - **Clustering:** encontrar grupos de elementos similares
 - **Refuerzo:** siguiente acción a tomar



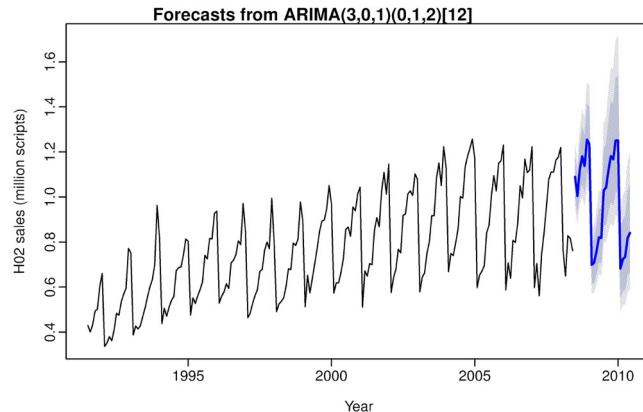
TAREAS DE APRENDIZAJE AUTOMÁTICO

- **¿Qué podemos hacer con los datos?:**
 - **Clasificación:** predecir la categoría de un ítem
 - **Regresión:** predecir un valor continuo
 - **Detección de excepciones:** encontrar anomalías
 - **Clustering:** encontrar grupos de elementos similares
 - **Refuerzo:** siguiente acción a tomar
 - **Asociaciones:** encontrar reglas de coocurrencia



TAREAS DE APRENDIZAJE AUTOMÁTICO

- **¿Qué podemos hacer con los datos?:**
 - **Clasificación:** predecir la categoría de un ítem
 - **Regresión:** predecir un valor continuo
 - **Detección de excepciones:** encontrar anomalías
 - **Clustering:** encontrar grupos de elementos similares
 - **Refuerzo:** siguiente acción a tomar
 - **Asociaciones:** encontrar reglas de coocurrencia
 - **Secuencias:** utilizar información temporal

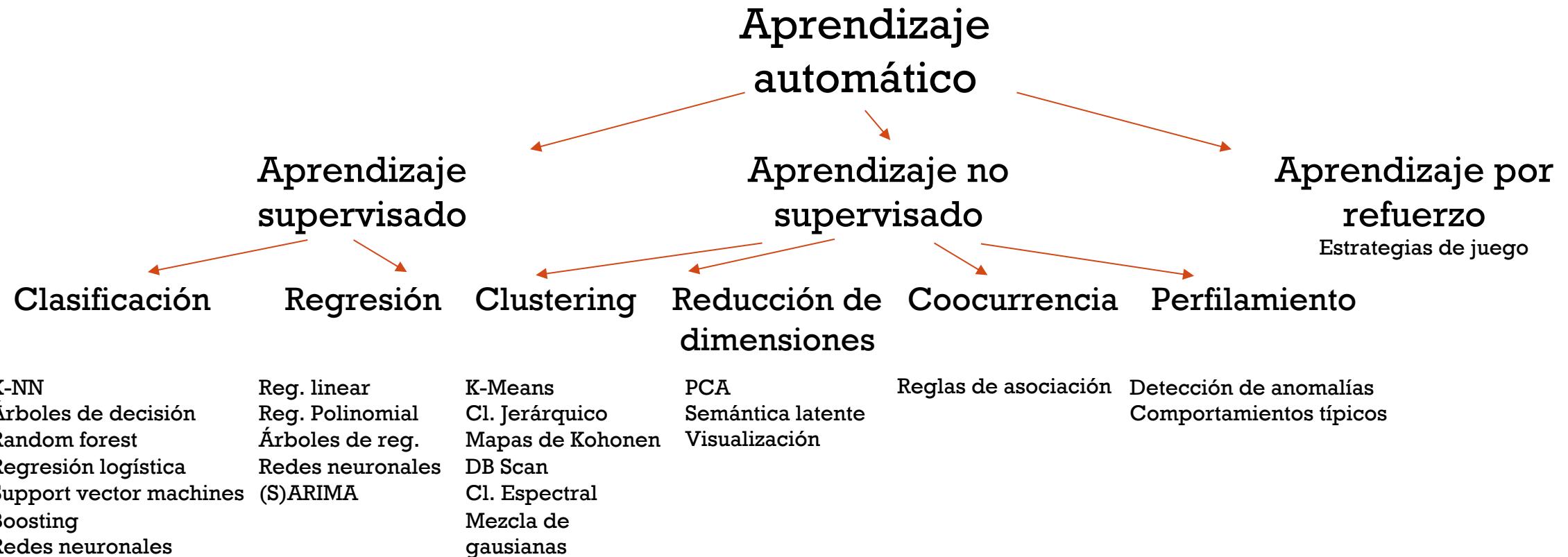


TAREAS DE APRENDIZAJE AUTOMÁTICO

- **¿Qué podemos hacer con los datos?:**
 - **Clasificación:** predecir la categoría de un ítem
 - **Regresión:** predecir un valor continuo
 - **Detección de excepciones:** encontrar anomalías
 - **Clustering:** encontrar grupos de elementos similares
 - **Refuerzo:** siguiente acción a tomar
 - **Asociaciones:** encontrar reglas de coocurrencia
 - **Secuencias:** utilizar información temporal
 - **Resumen:** simplificar la representación de información
 - **Visualización:** facilitar la comprensión y el descubrimiento



TAXONOMÍA



PREGUNTA:

¿Por qué?
¿Cómo?
¿Cuáles?
¿Cuándo?

DATOS:

Materia prima



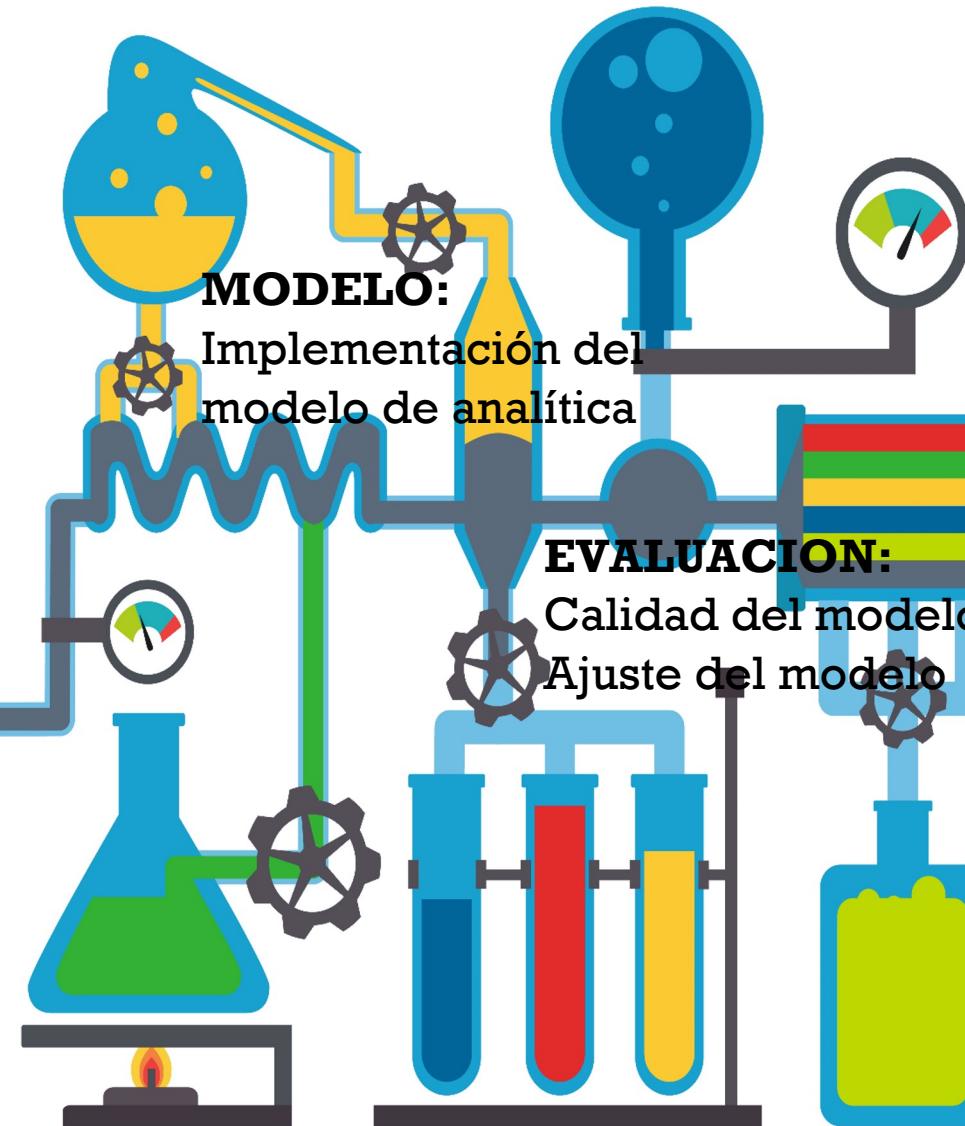
ANALISIS EXPLORATORIO:

Limpieza de datos
Preparación / transformación
Escogencia de variables



MODELO:

Implementación del
modelo de analítica



EVALUACION:

Calidad del modelo
Ajuste del modelo



DESPLIEGUE:
Resultados
Conocimiento
Estrategia



CUIDADO!

- Los resultados de la analítica deben ser evaluados e interpretados antes de ser explotados

EVALUACIÓN



- Criterios de éxito
- Evaluación de los resultados del modelo
- Revisión del proceso
(viabilidad, correctitud, etc.)



PREGUNTA:

¿Por qué?
¿Cómo?
¿Cuáles?
¿Cuándo?

DATOS:

Materia prima



ANALISIS EXPLORATORIO:

Limpieza de datos
Preparación / transformación
Escogencia de variables



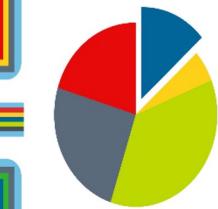
MODELO:

Implementación del
modelo de analítica



EVALUACION:

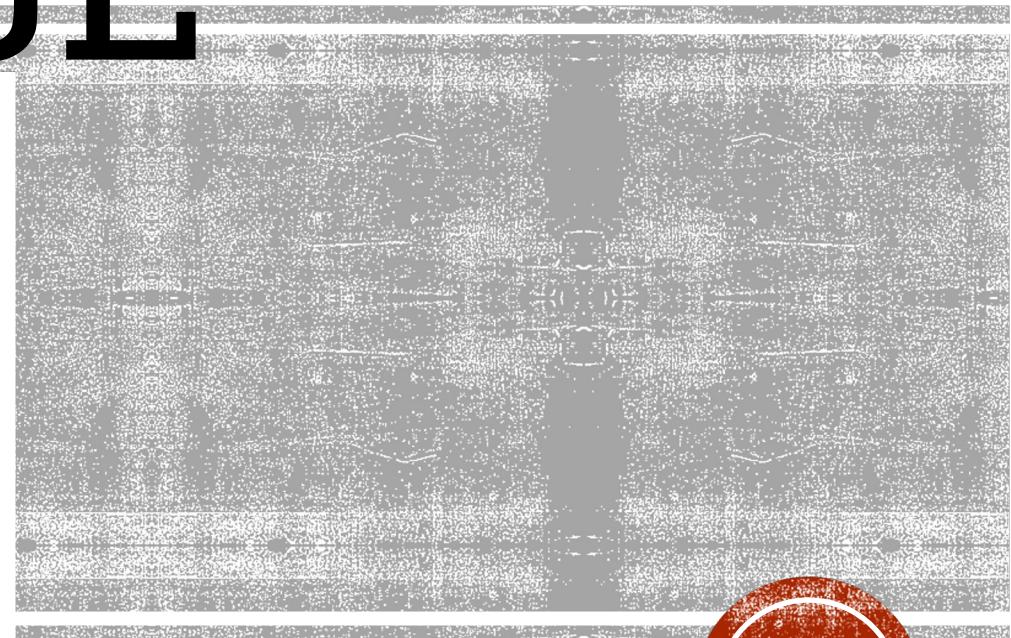
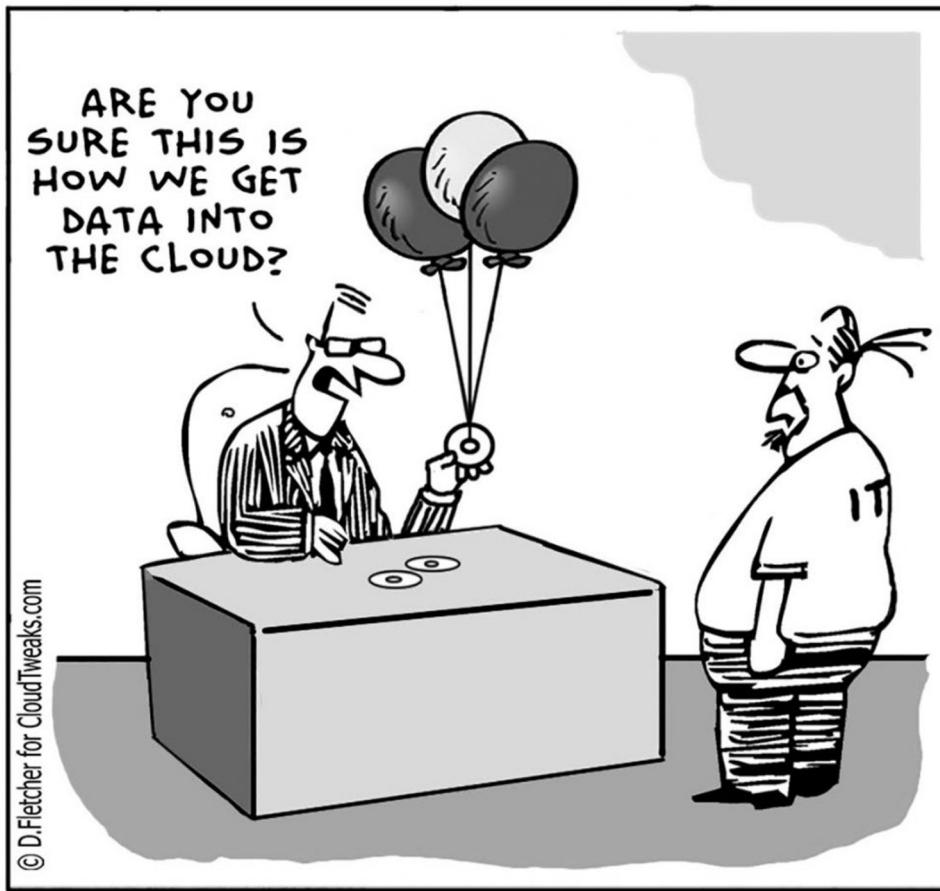
Calidad del modelo
Ajuste del modelo



DESPLIEGUE:
Resultados
Conocimiento
Estrategia



DESPLIEGUE



¿Cómo puedo explotar todo esto?

DESPLIEGUE

¿Cómo industrializo todo esto?

- ¿Aplicación stand-alone o integración con el sistema actual?
- ¿On premise o servicios cloud?
- ¿Reprogramo el modelo en otra plataforma?
- ¿Cómo automatizo la cadena de aprendizaje?
- ¿Cómo uso los resultados?
- ¿Cada cuánto actualizo los modelos?
- ¿Incluyo todos mis datos históricos en el aprendizaje?
- ¿Y las cuestiones de seguridad?

DESPLIEGUE

Buenas prácticas:

- Especificar requerimientos y características de ejecución
- Separar el modelo (software) de sus parámetros (configuración), control de versiones
- Implantar una infraestructura de back testing y de now testing
- Utilizar herramientas de prueba automática para monitorear la degradación de la calidad de los modelos
- Actualizar los modelos, puede que solo sea necesario actualizar sus parámetros

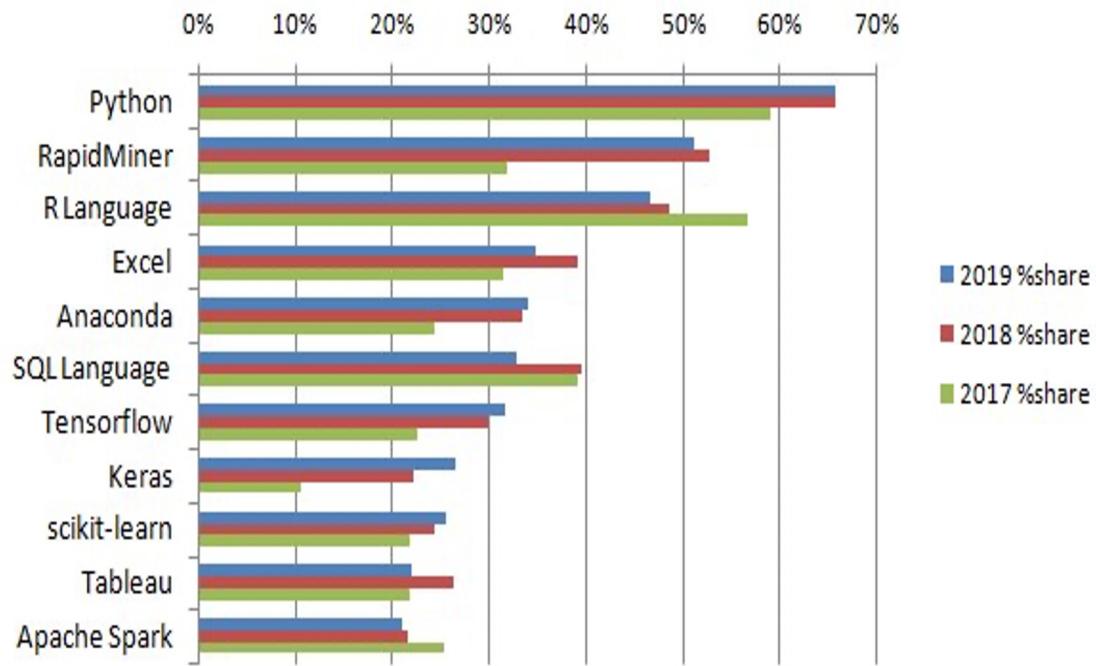
DESPLIEGUE

Herramientas

- Excel
- R
- Python
- Weka
- Knime
- SPSS
- Orange
- Matlab
- Rapidminer
- Anaconda
- Hadoop
- Cloudera
- AWS
- Azure
- Mahout
- Mllib
- ...

DESPLIEGUE

Top Analytics, Data Science, Machine Learning Software 2017-2019, KDnuggets Poll



Data Science and Machine Learning Platforms Magic Quadrant



Gartner - Magic Quadrant for Data Science and Machine Learning Platforms



VIDEOS

- Google analytics ad, shopping online

<https://www.youtube.com/watch?v=3Sk7cOqB9Dk>

<https://www.youtube.com/watch?v=N5WurXNec7E>

REFERENCIAS

- *How to win at digital transformation*, Forbes, 2016
- *Data Science for Business*, Foster Provost & Tom Fawcett, O'Reilly, 2013
- *CRISP-DM 1.0 Step-by-step data mining guides*, Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth (2000)
- *An external IBM ASUM Implementation Roadmap for data mining and predictive analytics projects*, IBM
- <https://www.gartner.com/webinar/2931518>
- <https://hbr.org/2013/10/are-you-ready-for-a-chief-data-officer>
- <https://developer.ibm.com/predictiveanalytics/2015/10/16/have-you-seen-asum-dm/>

REFERENCIAS

- *Predictive Analytics (2nd Edition)*, Eric Siegel, Wiley, 2016
- *Data Mining (4th Edition)*, Ian Witten, Eibe Frank, Mark A. Hall & Christopher J. Pal, Elsevier, 2016
- *Machine Learning (Coursera)*, Andrew Ng, Stanford University, 2016
- *The Golden Age of Marketing (Coursera)*, Eric Bradlow, Wharton School, University of Pennsylvania, 2016
- http://ana.blogs.com/maestros/2006/11/data_is_the_new.html
- <http://dismagazine.com/discussion/73298/sara-m-watson-metaphors-of-big-data/>