

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

ANIBAL SOSA

Formación

- Matemático de la Universidad del Valle, 2000
- Maestría en Ciencias Matemáticas, Universidad del Valle, 2004
- Maestría en Ciencias Computacionales, Universidad de Texas El Paso, USA, 2011
- Doctorado en Ciencias Computacionales, Universidad de Texas El Paso, USA, 2012

Experiencia académica

- Posdoctorado Cyber-ShARE Center of Excellence, Universidad Texas El Paso, USA, 2013
- Desde agosto 2004, profesor tiempo completo ICESI, Facultad de Ingeniería
- Ex-Becario Fulbright como profesor visitante en el Oak Ridge National Laboratory, Tennessee, USA, 2019



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, distinguiendo entre los **modelos supervisados** (permitiendo la predicción) y **no supervisados** (encontrando patrones estructurales en los datos), y estudiando las **métricas de calidad** de los mismos y los **protocolos de evaluación** que permiten valorarlos y compararlos.



DESCRIPCIÓN

Aprendizaje Supervisado

- Sobreaprendizaje
- Protocolos de evaluación
 - Holdout
 - Validación cruzada (Cross Validation – CV)
- Modelos
 - K-NN
 - Bayes Ingenuo
 - Árboles de decision - ensambles.
 - SVM (opcional)

Aprendizaje No Supervisado

- Modelos de clustering
 - K-Means
 - Jerárquico
 - DBScan (opcional)
- Análisis de Componentes Principales (PCA)



UNIDADES

Unidad 1
Introducción a
la Analítica

Unidad 2
Modelos de
Aprendizaje
Supervisado

Unidad 3
Modelos de
Aprendizaje No
supervisado



METODOLOGÍA

Desarrollo por unidades:

- Espacios de discusión
- Aplicación y análisis de los conceptos
- Participación activa de los estudiantes.

Sesión de clase

- Lectura previa basada en guías.
- Quices al inicio de algunas clases
- Resolución de dudas al inicio de la clase
- Presentación de tema
- Talleres de aplicación en grupos.
 - Excel y Python
- Talleres prácticos en sesiones extra de 2h con el monitor Daniel Osorio (daniel.osorio777@gmail.com)



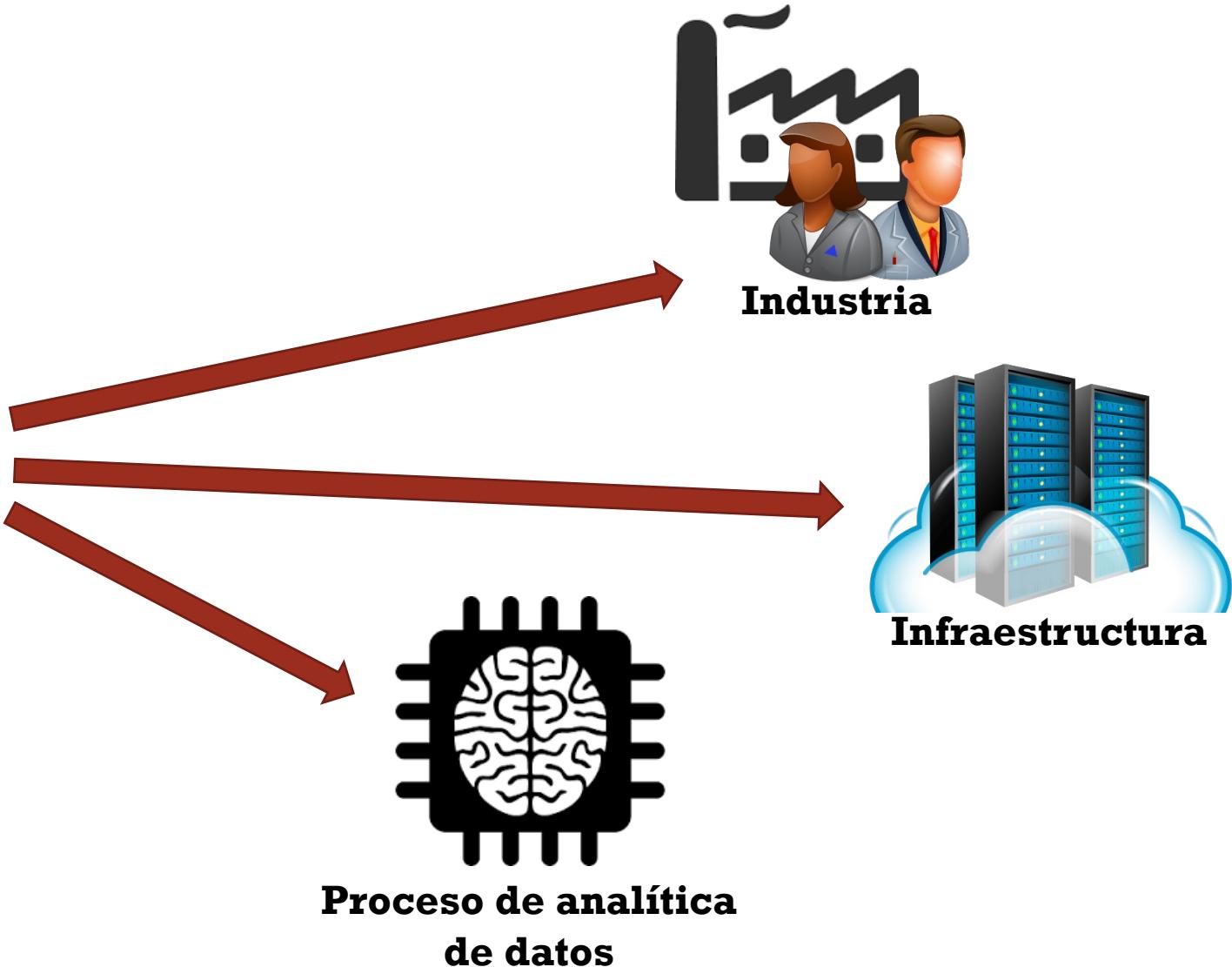
EVALUACIÓN

Forma de Evaluación	Porcentaje
Quices / Tareas	20%
Examen Parcial 1	25%
Examen Parcial 2	25%
Trabajo de aplicación	30%
Total	100%

- Los exámenes del curso comprenden los temas en sus aspectos teóricos (en forma de cuestionario de escogencia múltiple, incluyendo posiblemente ejercicios cortos), con algunos casos de aplicación.
- El trabajo final se hará sobre un dataset con datos reales de alguna entidad particular o de un repositorio en la web.

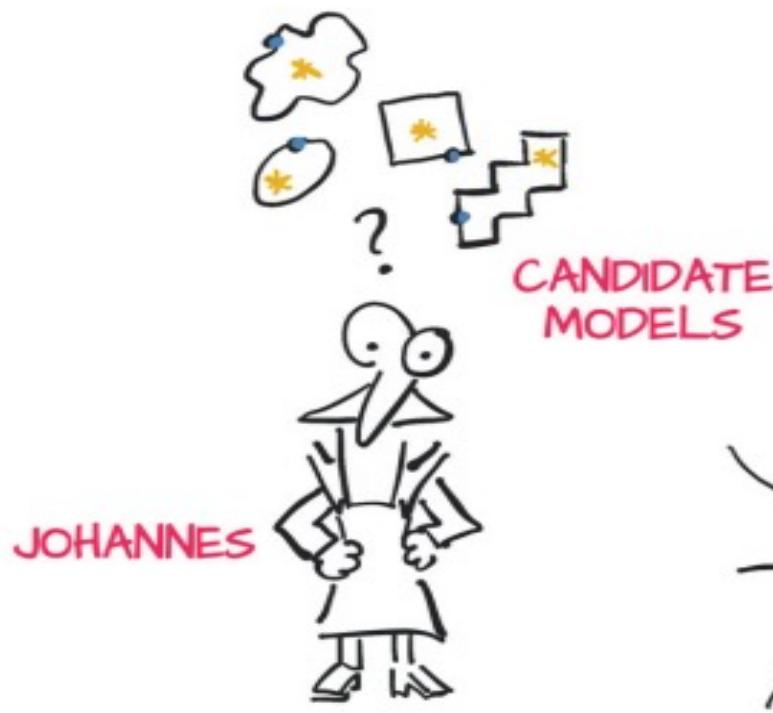


AGENDA

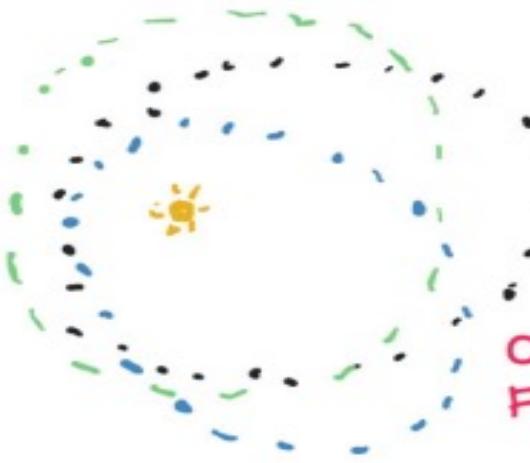


Proceso de analítica
de datos

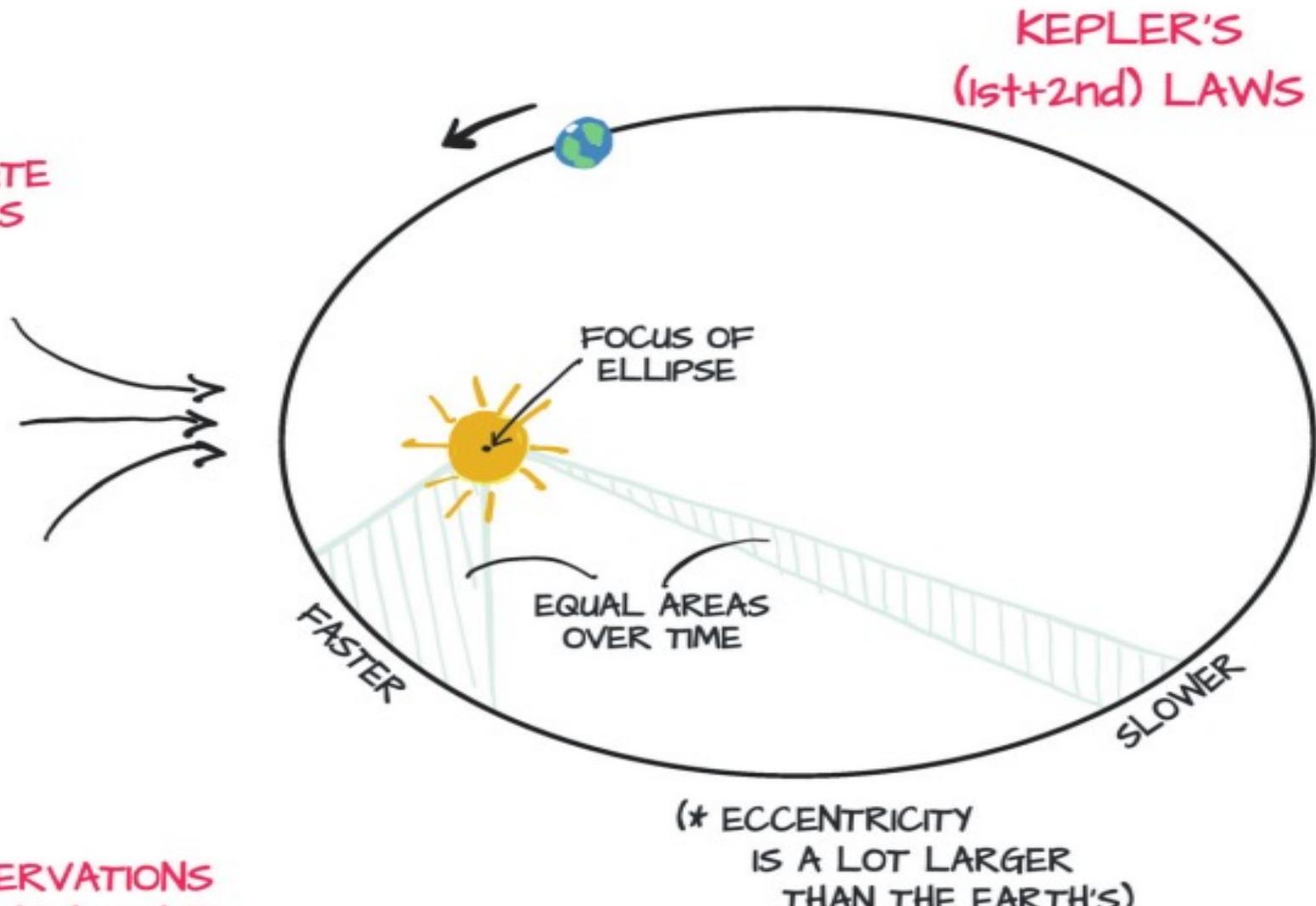




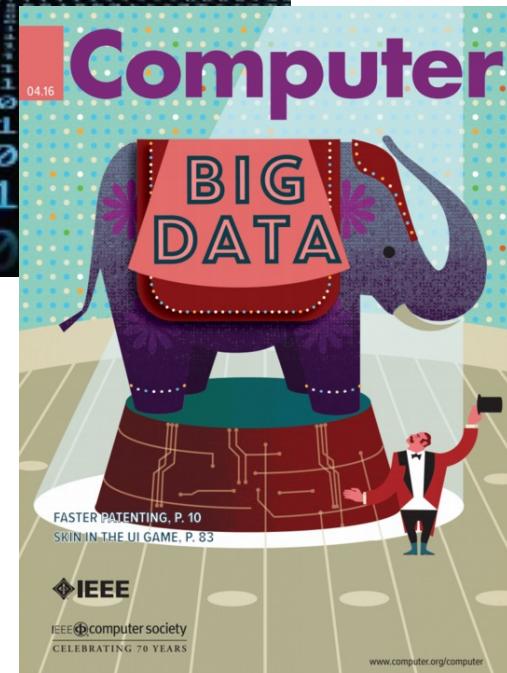
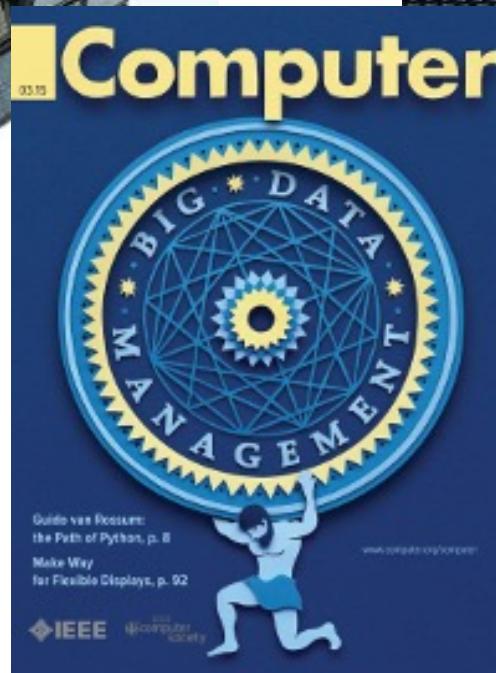
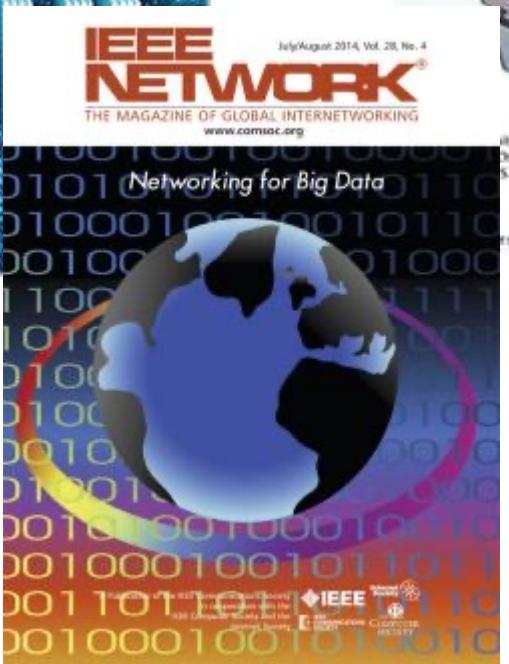
JOHANNES



OBSERVATIONS
FOR MULTIPLE
PLANETS



BUZZWORD



BUZZWORD





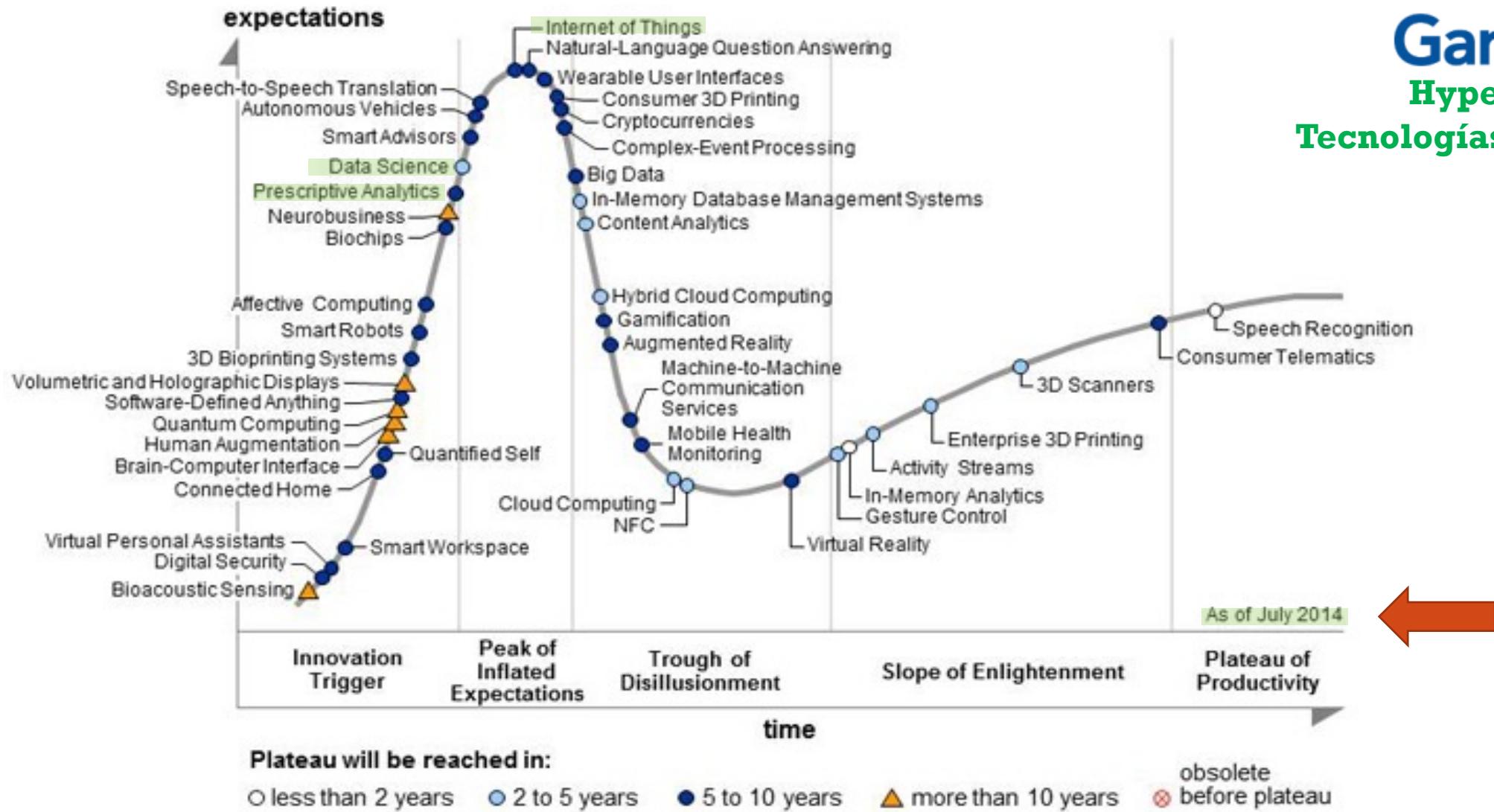
The state of AI in 2023: Generative AI's breakout year

August 1, 2023 | Survey

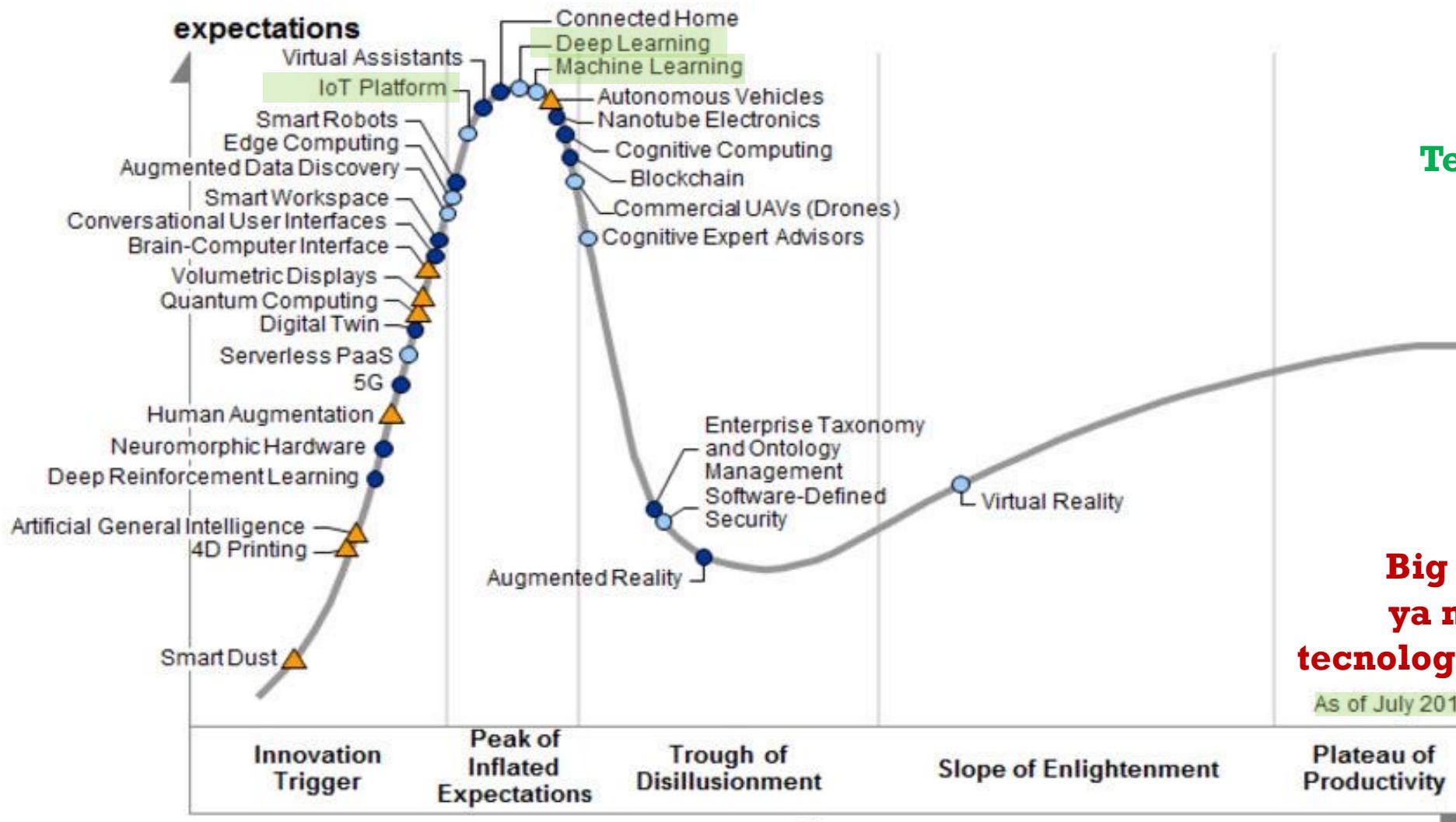


“Less than a year after many of these tools debuted, one-third of our survey respondents say their organizations are using gen AI regularly in at least one business function”

Gartner® Hype Cycle Tecnologías emergentes



Gartner® Hype Cycle Tecnologías emergentes



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

As of July 2017

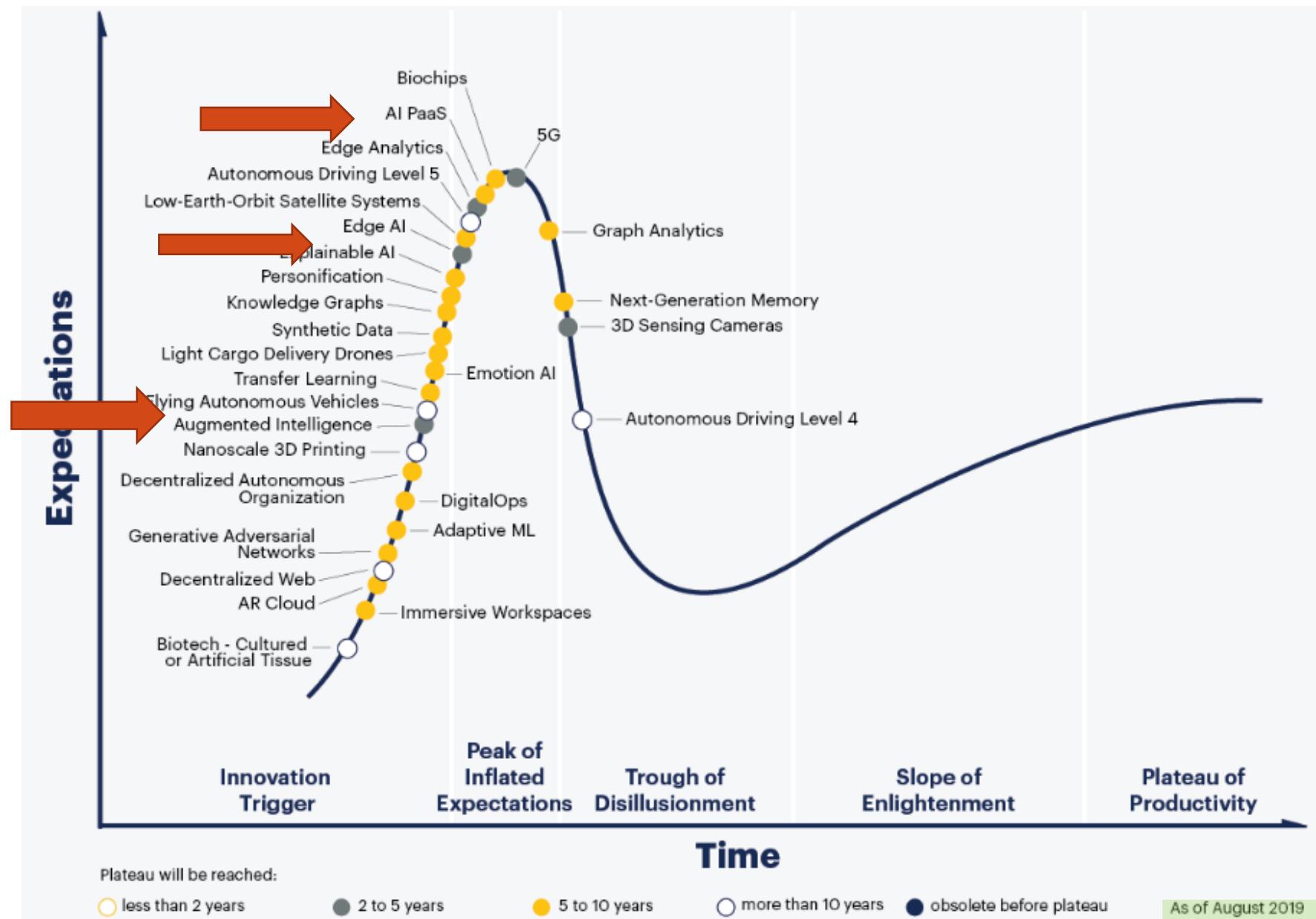
Years to mainstream adoption:

○ less than 2 years ● 2 to 5 years ● 5 to 10 years ▲ more than 10 years ○ obsolete

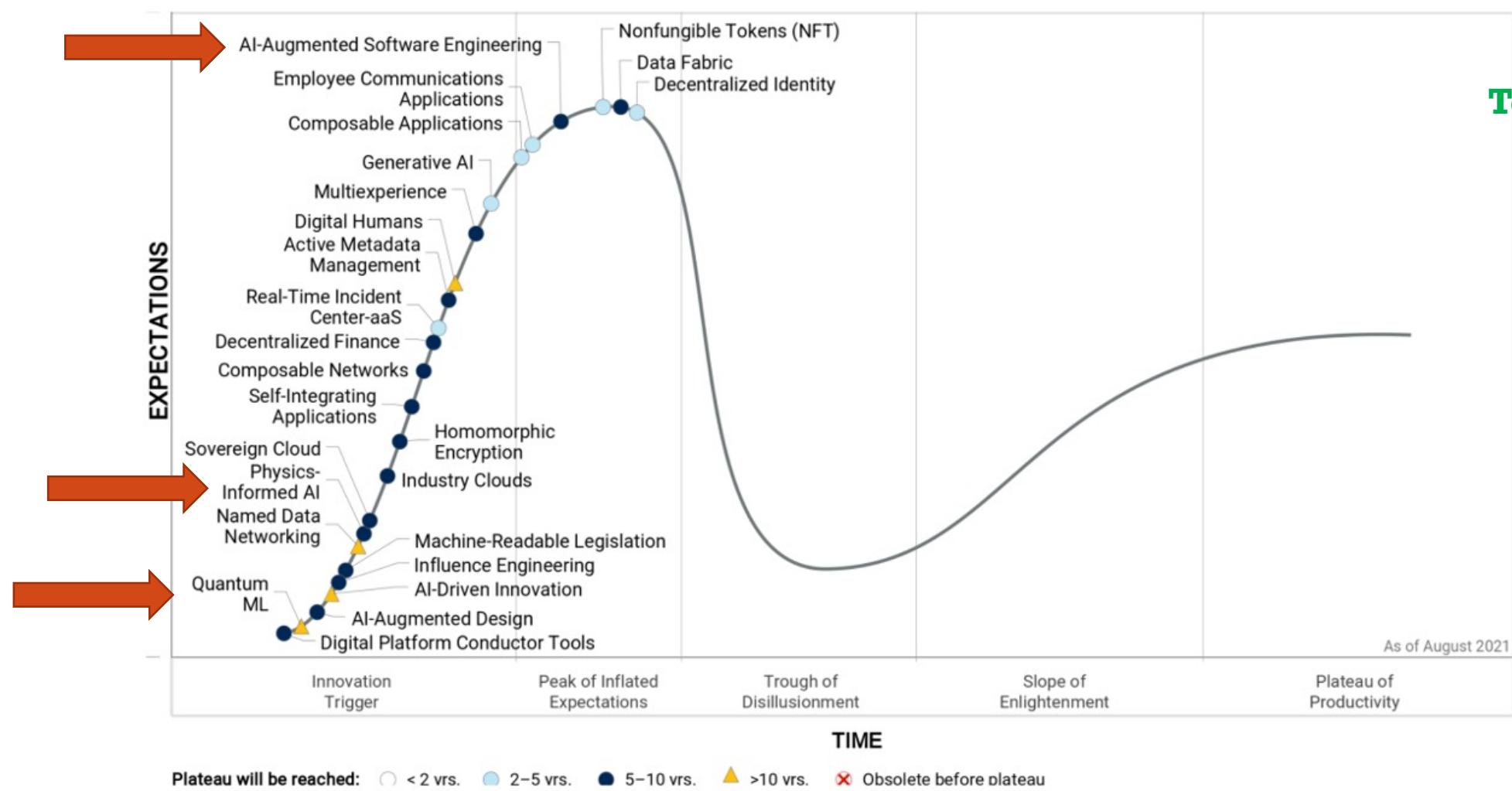
○ before plateau



Gartner® Hype Cycle Tecnologías emergentes



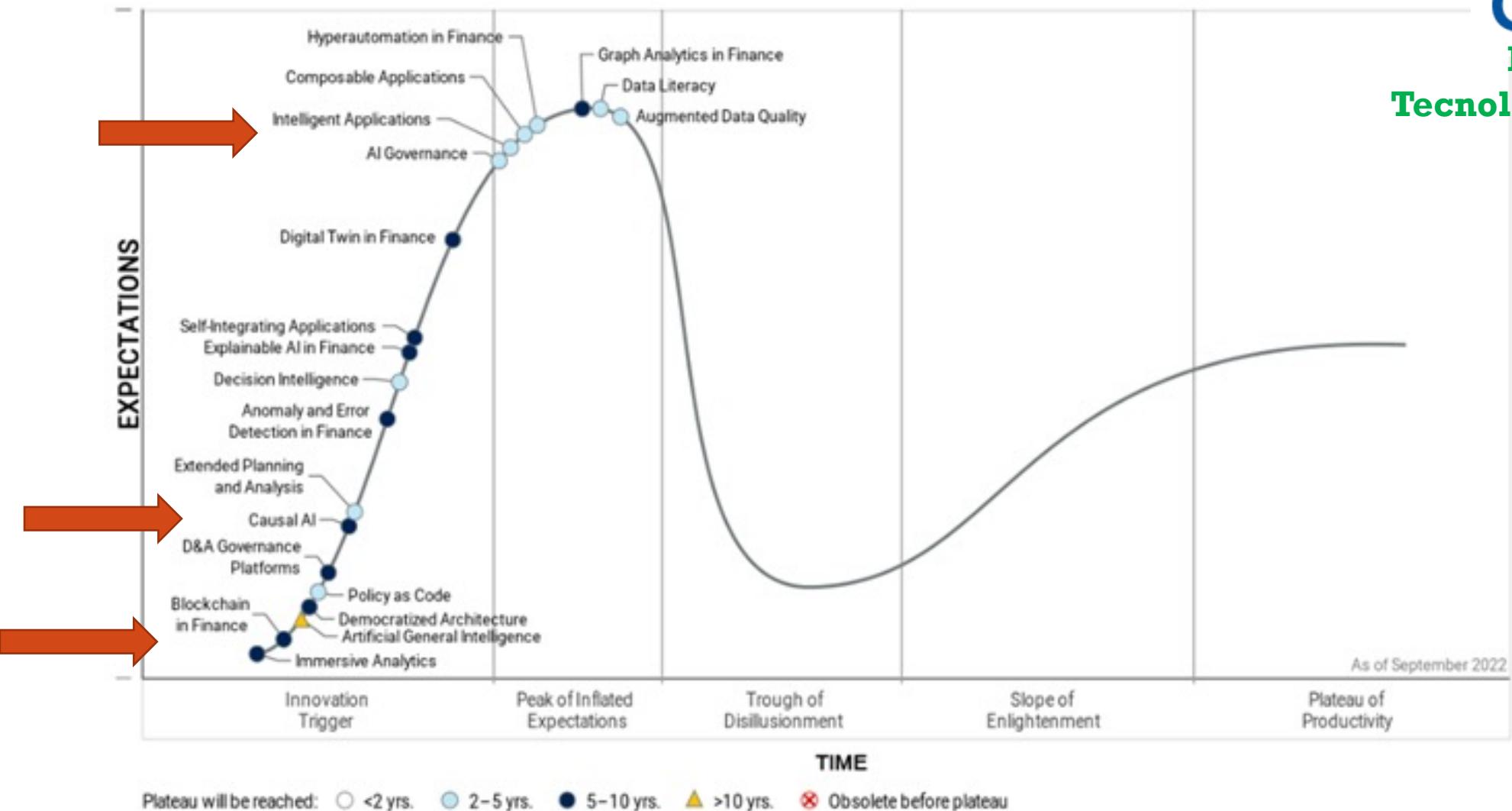
Gartner® Hype Cycle Tecnologías emergentes



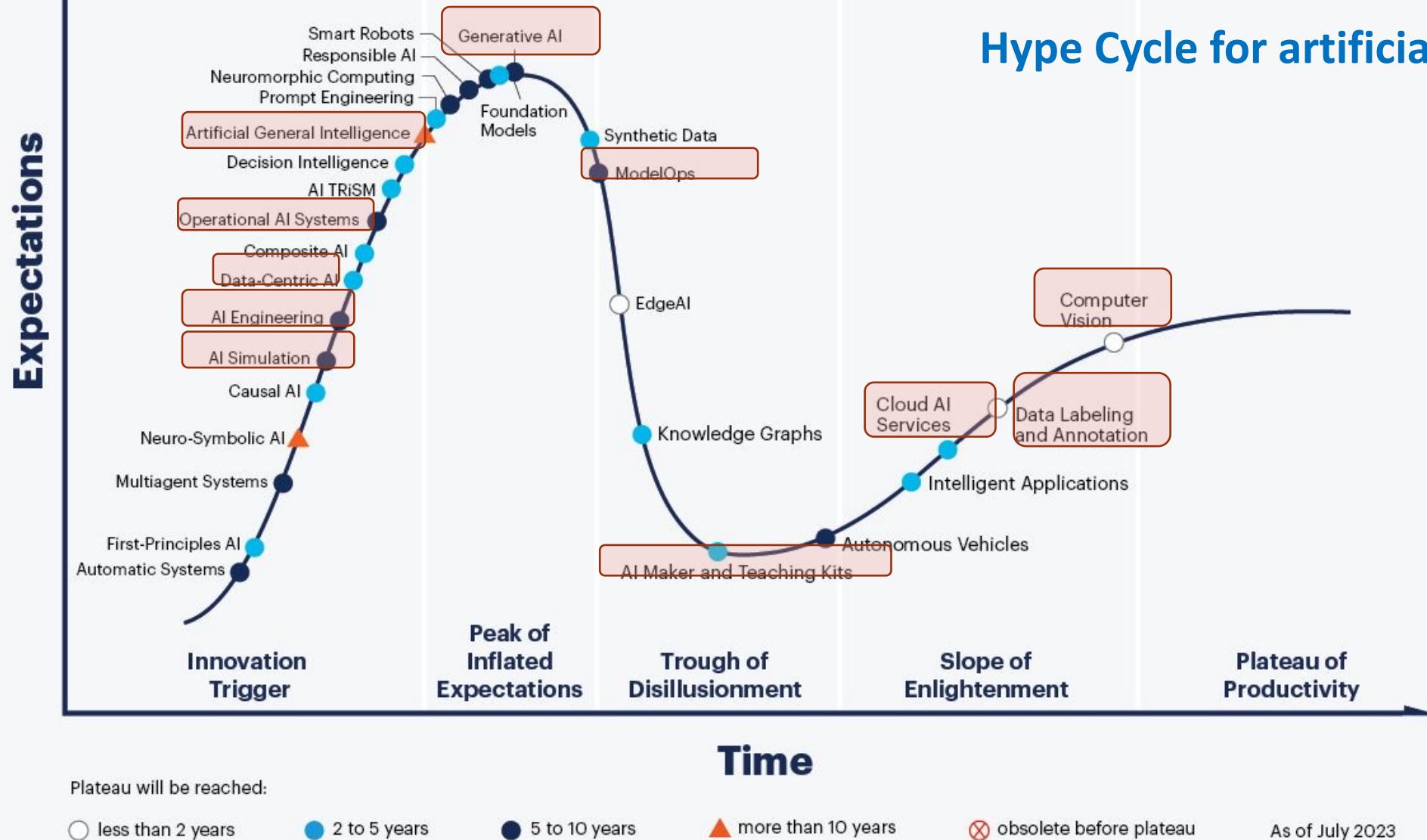
Source: Gartner (August 2021)



Gartner® Hype Cycle Tecnologías emergentes



Hype Cycle for artificial Intelligence



As of July 2023

TRANSFORMACIÓN DIGITAL



- Enfoque estratégico y de inversión:
 - Enfoque top-down o de una línea de negocio a la vez
 - Inversión en tecnología y en **capacidades de analítica de datos**
- Enfoque en el negocio y en la innovación:
 - Nuevos modelos de negocio, productos y servicios → **Basados en datos y analítica**
 - Nuevas tecnologías → Reducción de costos, ganancia en eficiencia
- **Potencial de datos y de analítica:**
 - Analítica de datos para generar nuevos ingresos (91% de las empresas, Forbes, 2019)
 - Enfoque en la experiencia de clientes (retención, segmentación, comunicación personalizada (mercadeo), riesgo, innovación, redes de suministro complejas)
- Adopción a través de toda la empresa:
 - Tecnología fusionada a todas las líneas y funciones de negocio → **Generación de datos**
- Balance entre gente y tecnología:
 - Cloud, mobile, **analytics**, IoT, social media como drivers de la transformación digital
 - Desafío: **talento** (recursos humanos) para aprovechar las tecnologías



BIG DATA (MCKINSEY, 2011)

Datos cuya escala, distribución, diversidad y/o validez en el tiempo requieren el uso de:

- Nuevas arquitecturas
- Modelos de analítica

para extraer conocimientos que revelen nuevas fuentes de valor para el negocio



BIG DATA ANALYTICS

World
Economic
Forum (3:13)

Forbes (2:45)

Harvard
Business
Review (2:44)



APLICACIONES EN LA INDUSTRIA



Inteligencia
de clientes



Tendencias



Optimización



Fraude



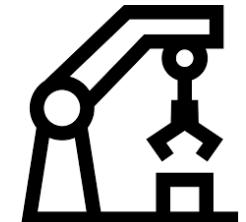
Salud



Materias
primas



Riesgos



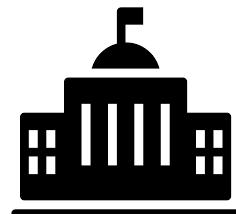
Infraestructura



Análisis de
sentimiento



Capacidades



Gobierno



ANALÍTICA EN LA INDUSTRIA

- Inteligencia de clientes

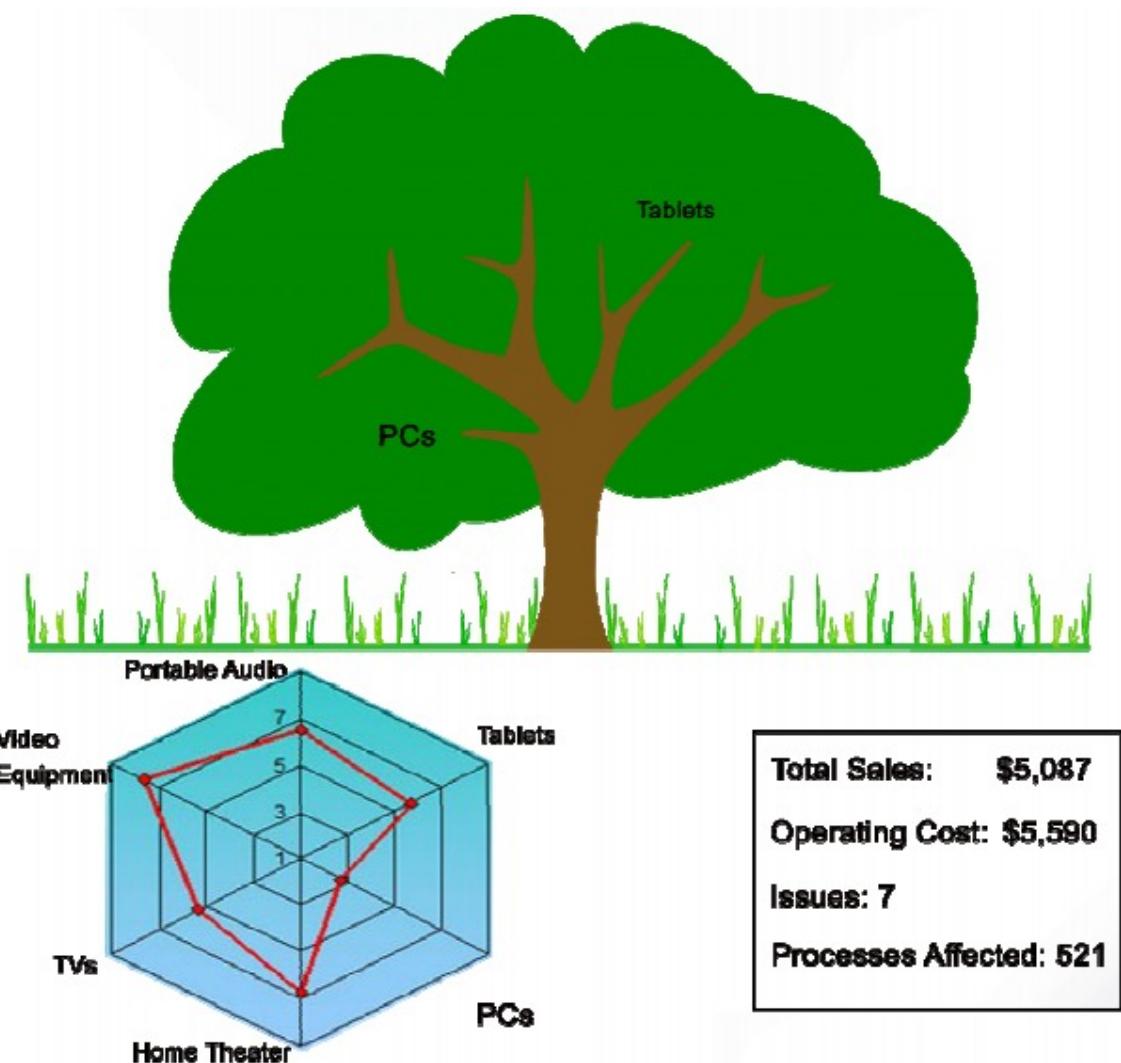
- ¿Cuáles de mis clientes me podrían abandonar para ir a la competencia?
- ¿Cuál es el valor potencial de los que aún no son clientes de mi compañía?
- ¿Qué clientes tienen un comportamiento o características similares?
- ¿Cuáles de mis clientes son propensos a la compra de un producto y cuáles no lo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



ANALÍTICA EN LA INDUSTRIA

Analítica descriptiva

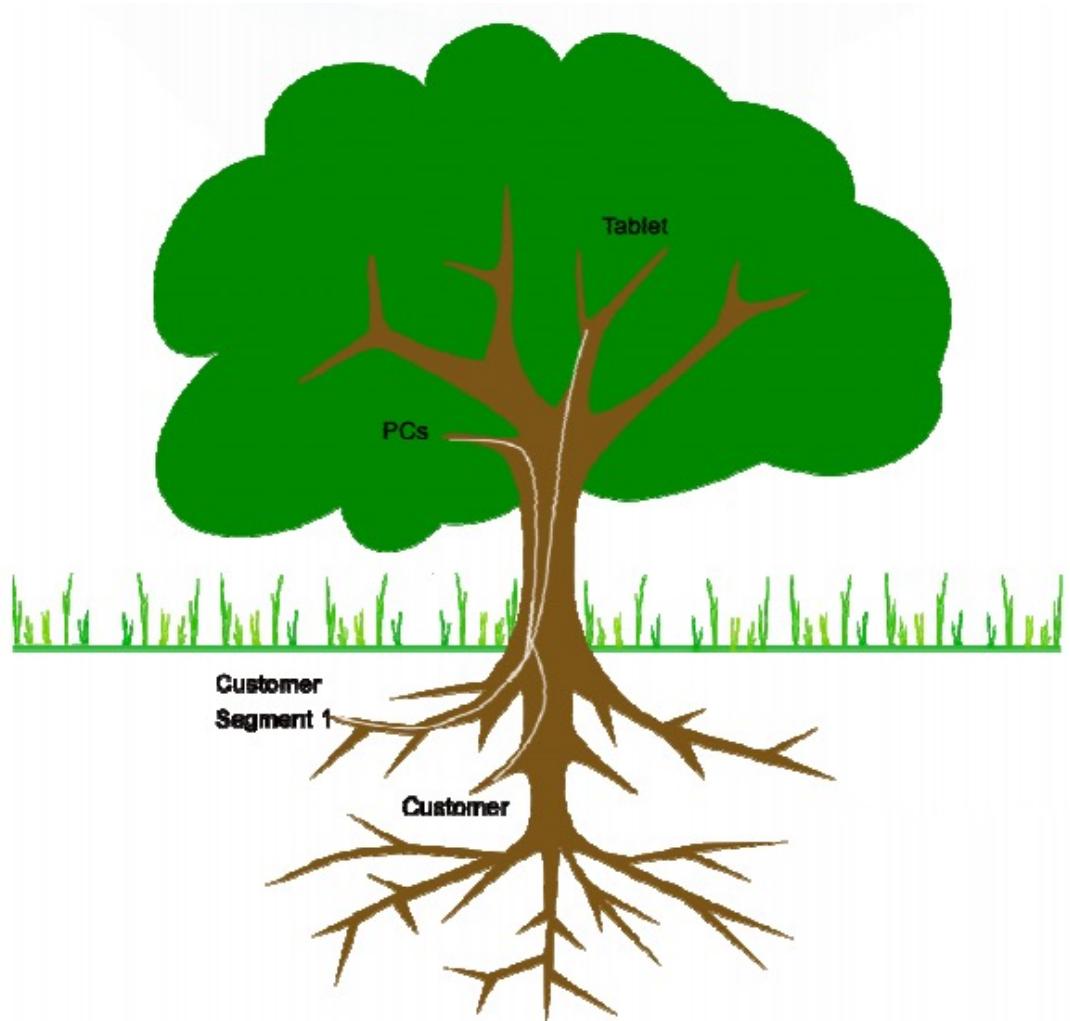
- ¿Qué está pasando? ¿Es bueno o malo?
- Biblioteca de reportes estáticos
- Peticiones de reportes desde todos los departamentos
- Intervención humana requerida
- Una vez se ve una representación, se enfrenta la necesidad de una explicación: ¿Por qué?



ANALÍTICA EN LA INDUSTRIA

Analítica diagnóstica

- ¿Por qué ocurrió esto?
- Explicar lo que está pasando
- 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



<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



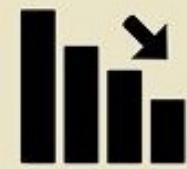
Optimization that helps achieve the best outcomes.



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



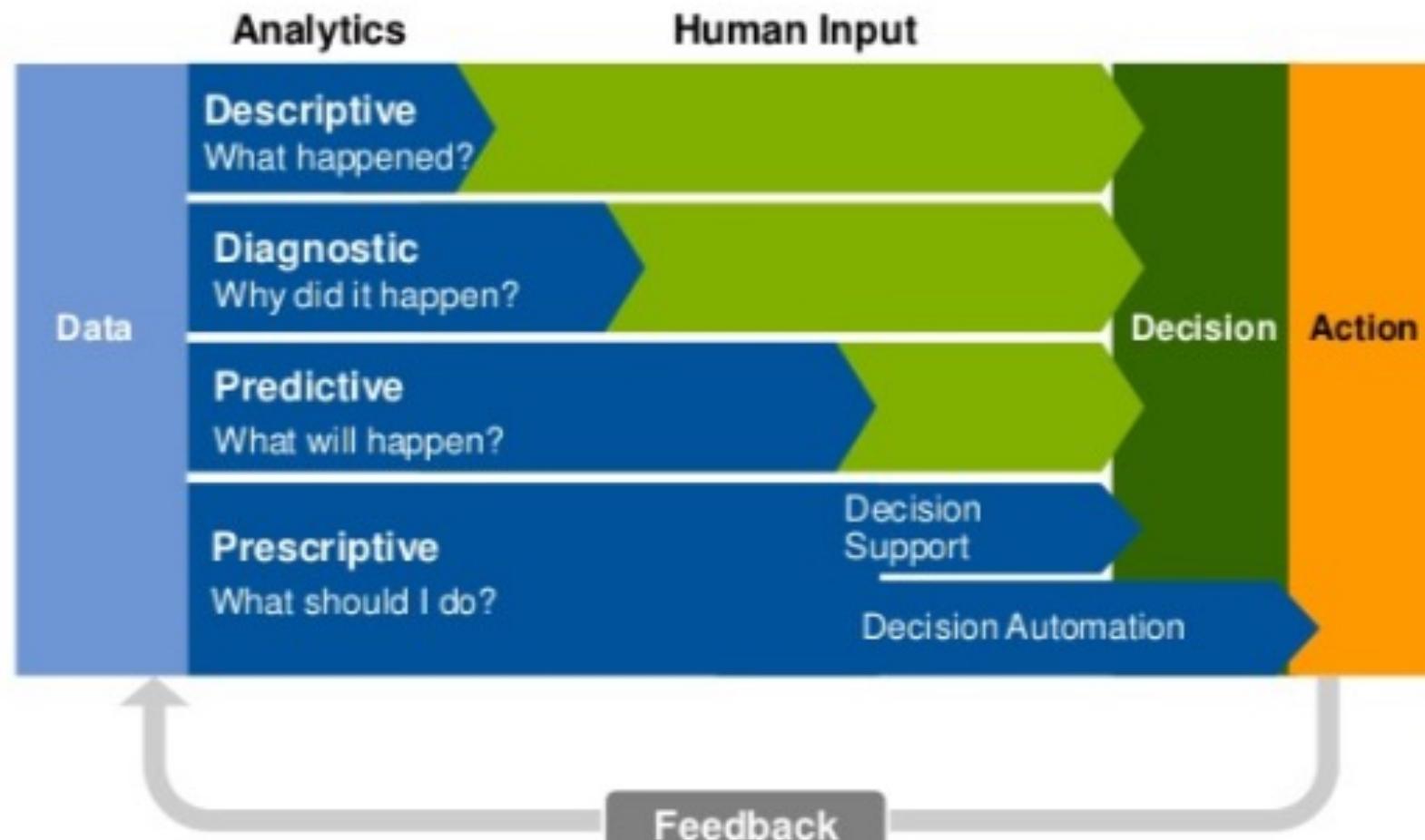
Stochastic optimization that helps understand how to achieve the best outcome and identify data uncertainties to make better decisions.



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



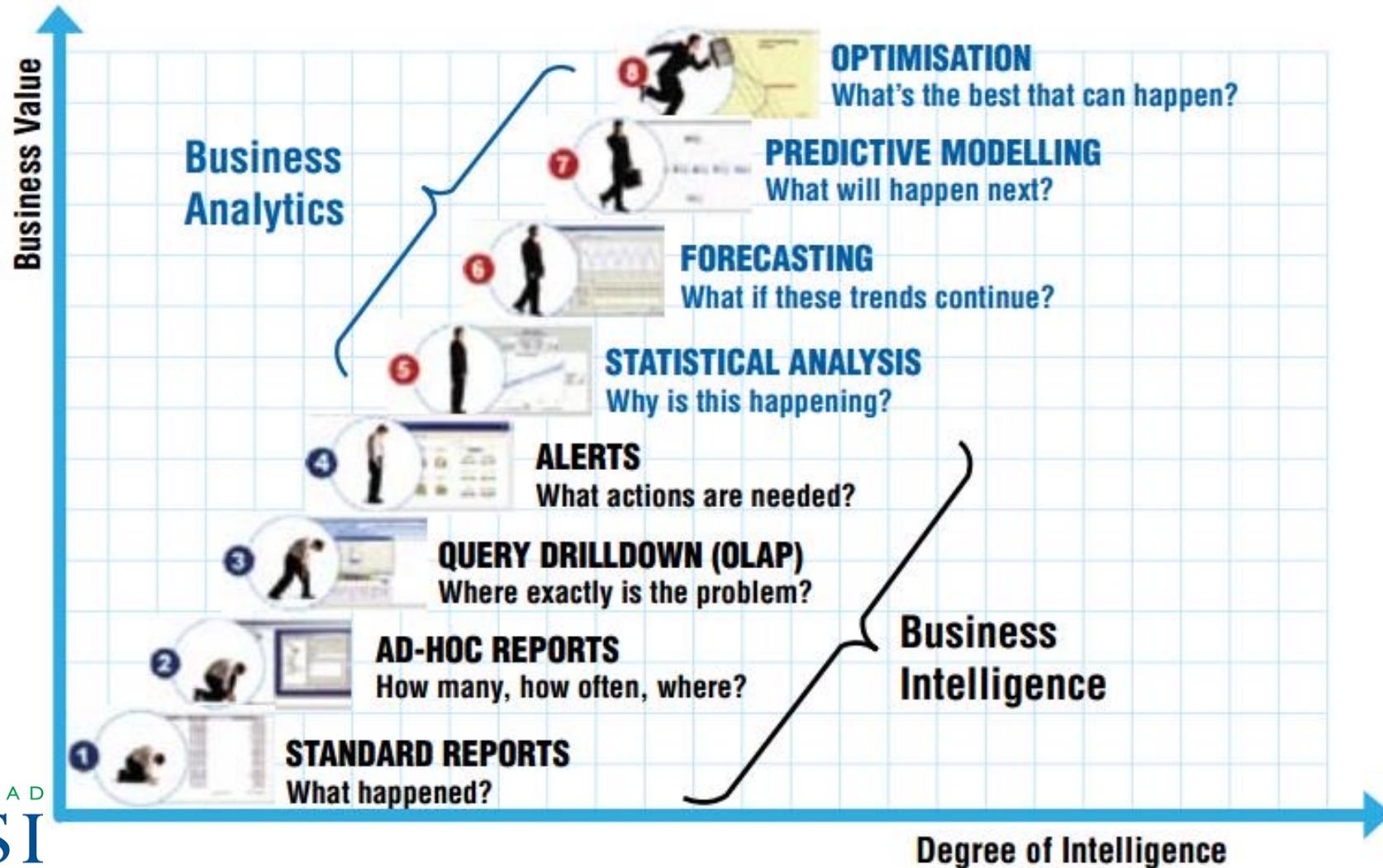
ANALÍTICA EN LA INDUSTRIA



Gartner, 2014



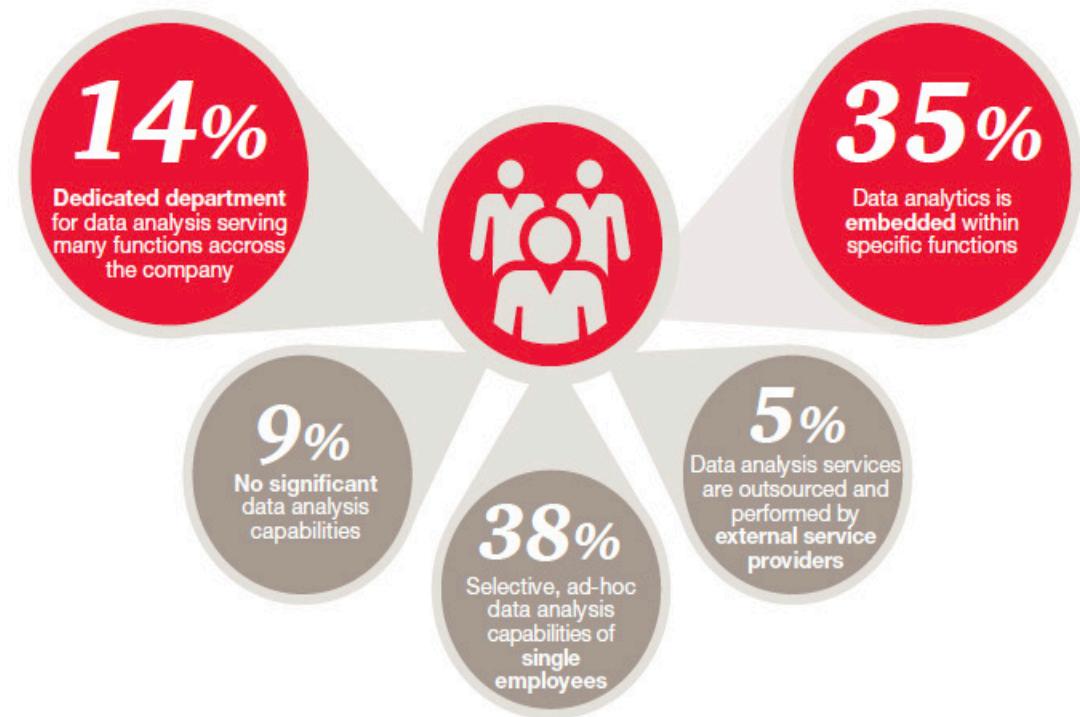
ANALÍTICA EN LA INDUSTRIA



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?

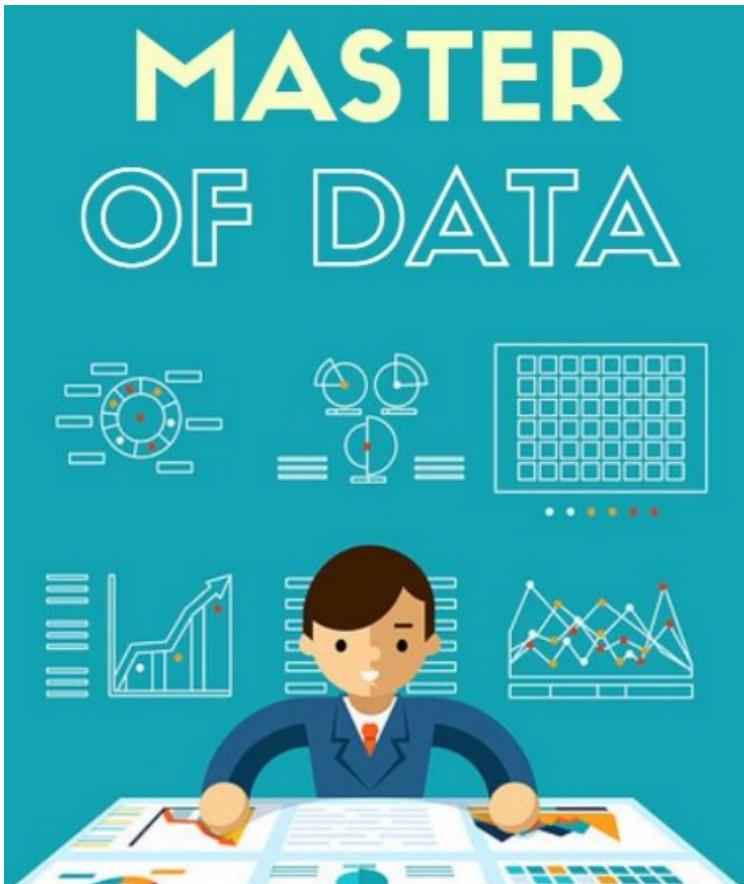
ANALÍTICA EN LA INDUSTRIA

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

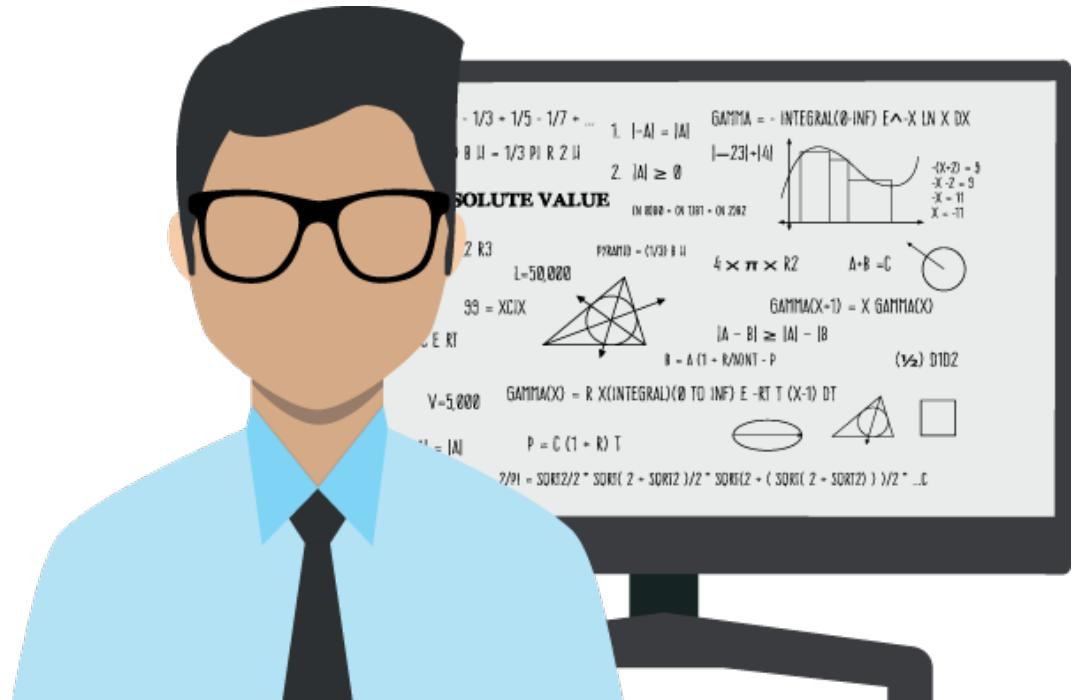


- Se enfoca en los datos como base para :
 - Desarrollar capacidades organizacionales
 - Tomar decisiones
 - Competir
 - Estructurar modelos de negocio novedosos
 - Evaluar modelos disruptivos
- Gestión de equipo multidisciplinario
- No se debe encargar de:
 - Manejar la infraestructura
 - Crear o mejorar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

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



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



- 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

Roles de un ingeniero de datos

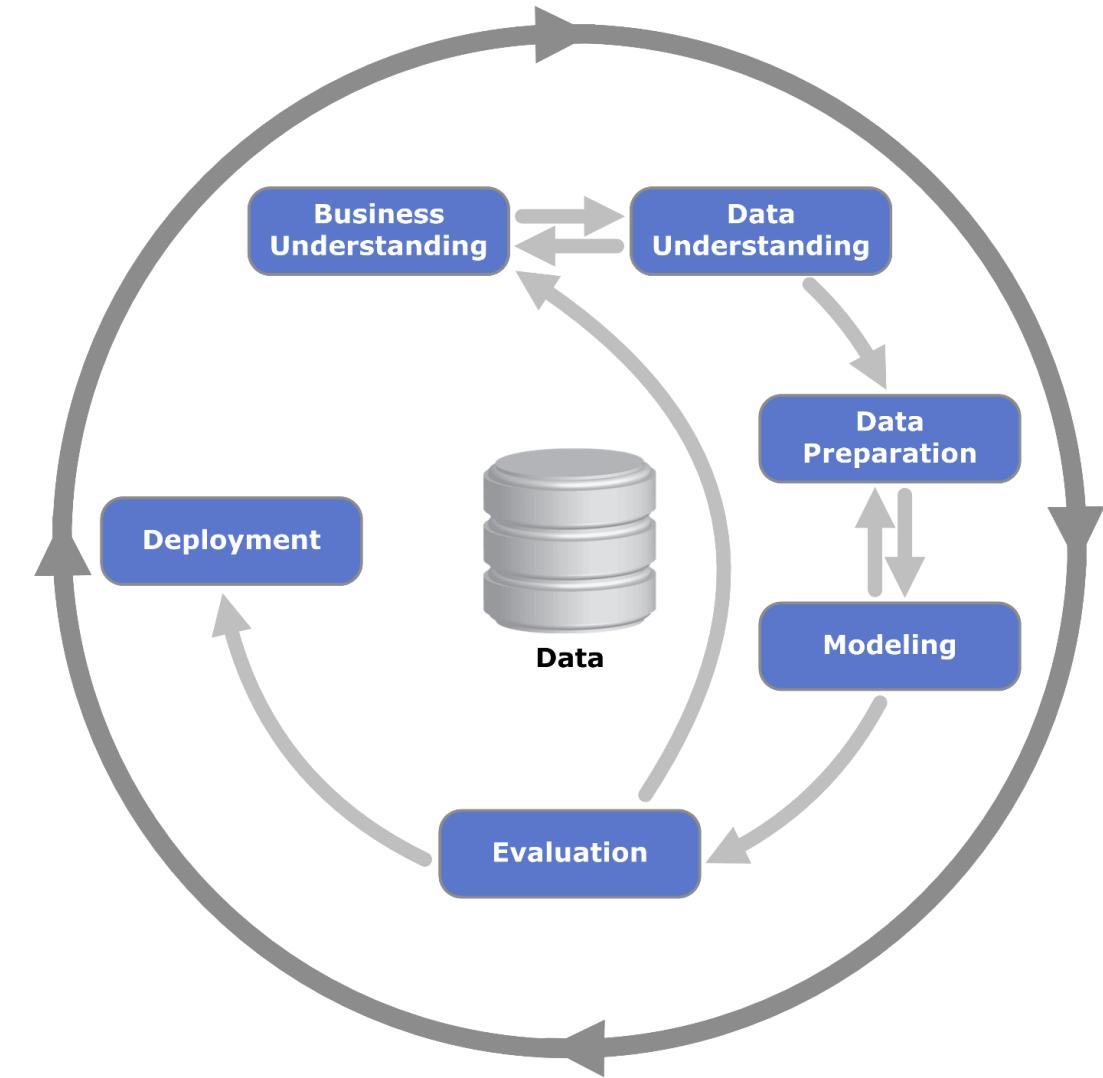


METODOLOGÍA

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

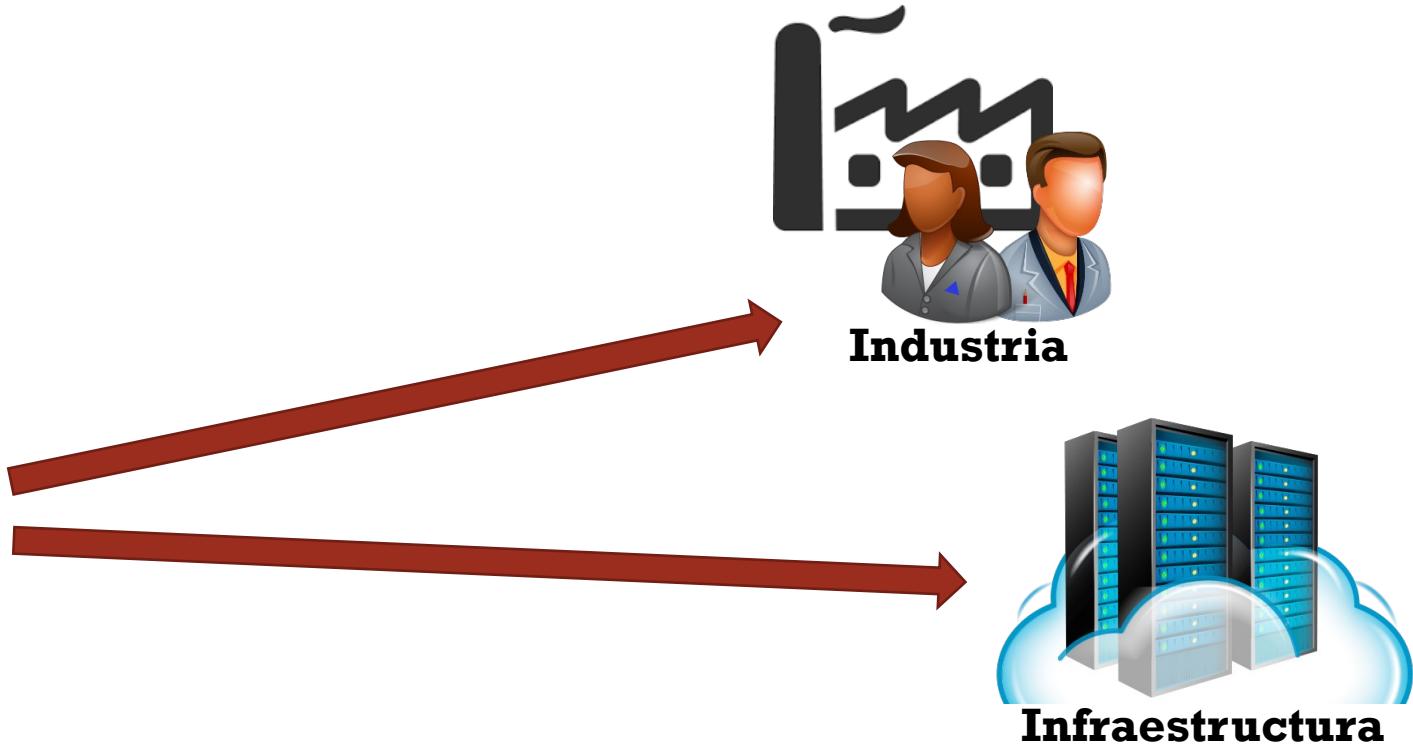
CRISP-DM

- 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

AGENDA



PROBLEMATICAS TÉCNICAS DE BIG DATA

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



APACHE HADOOP



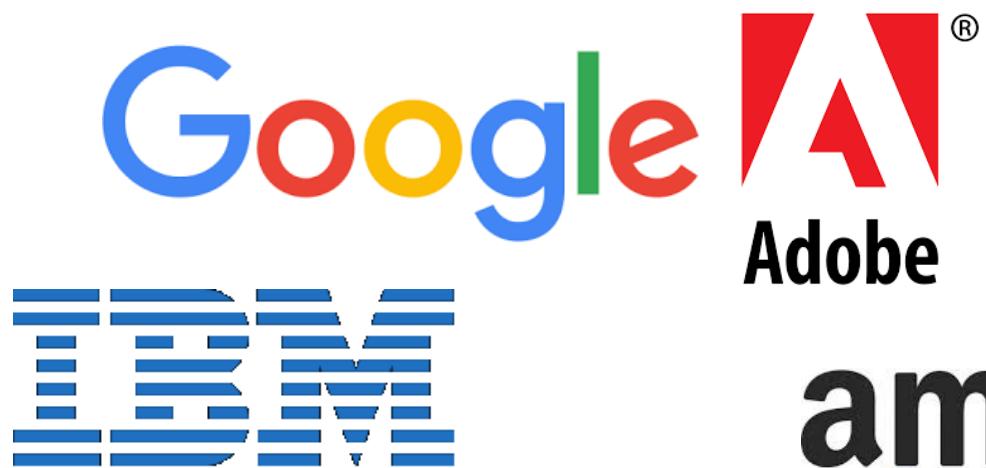
- Map Reduce + Google Filesystem (HDFS) = Nutch (2004) → Yahoo: Hadoop (2006)
- Open source, licenciado y administrado por Apache Software Foundation

Framework que se encarga de:

- **Almacenar** datos de manera distribuida en los nodos de un 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 ejecutaran la fase del reduce
- Disponibilizar el **resultado final** del tratamiento, o encadenarlo como datos de entrada de otra tarea map-reduce



APACHE HADOOP - USUARIOS



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



APACHE HADOOP - DISTRIBUCIONES



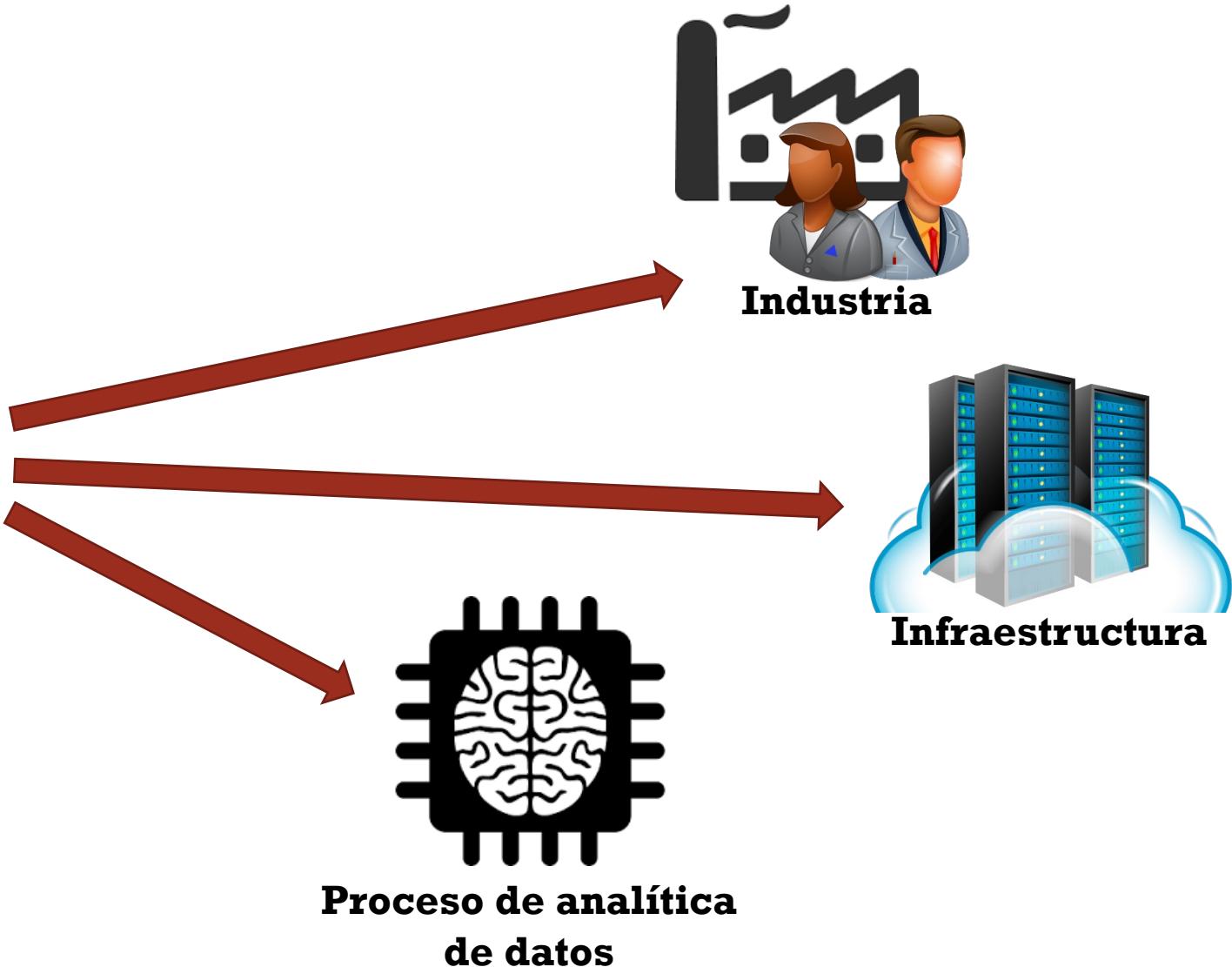
Microsoft Azure



Google Cloud Platform



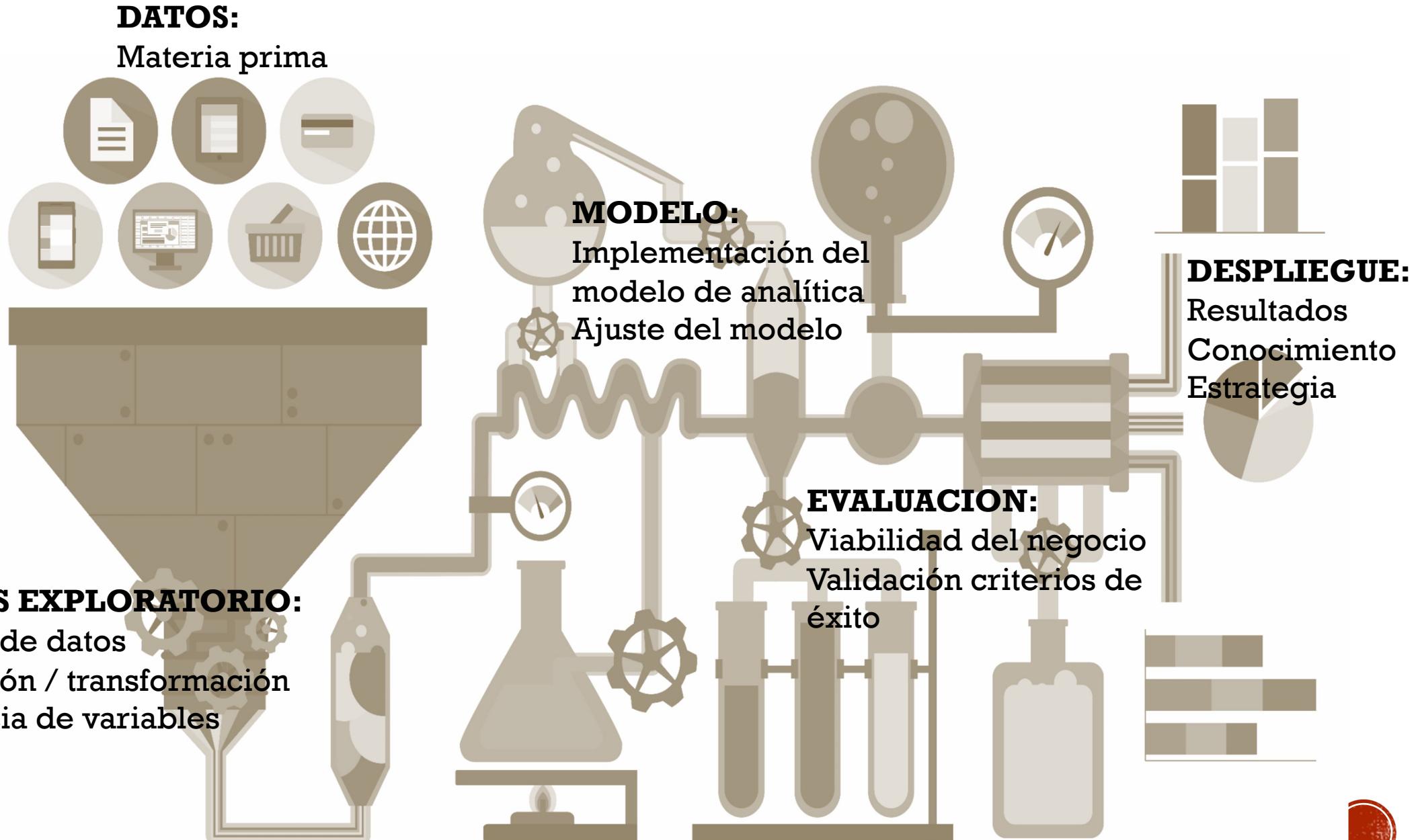
AGENDA



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."*

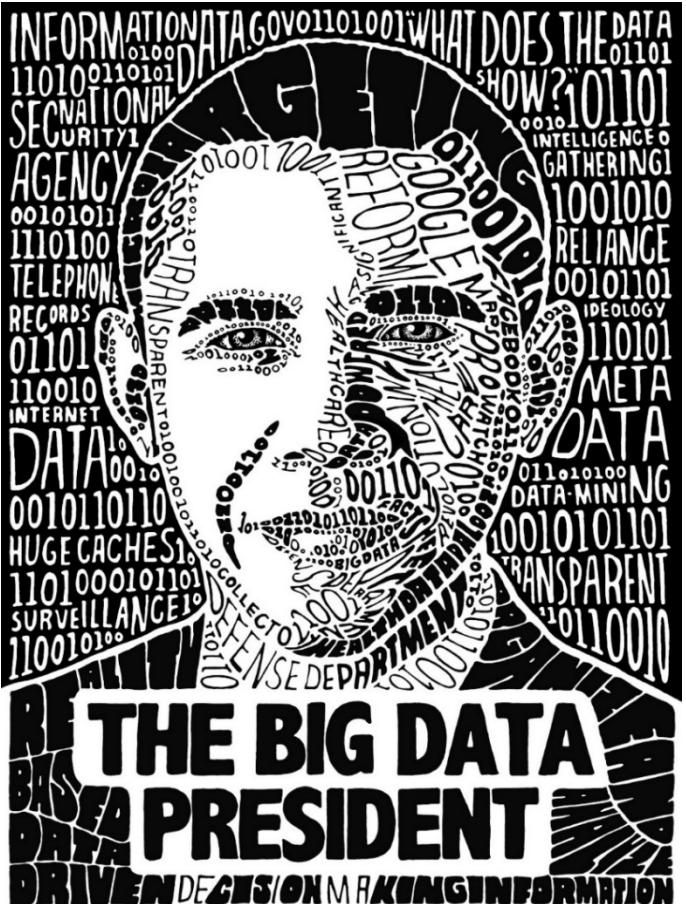
LA PREGUNTA

Clasificación de clientes

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



LA PREGUNTA



¿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”

→ 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

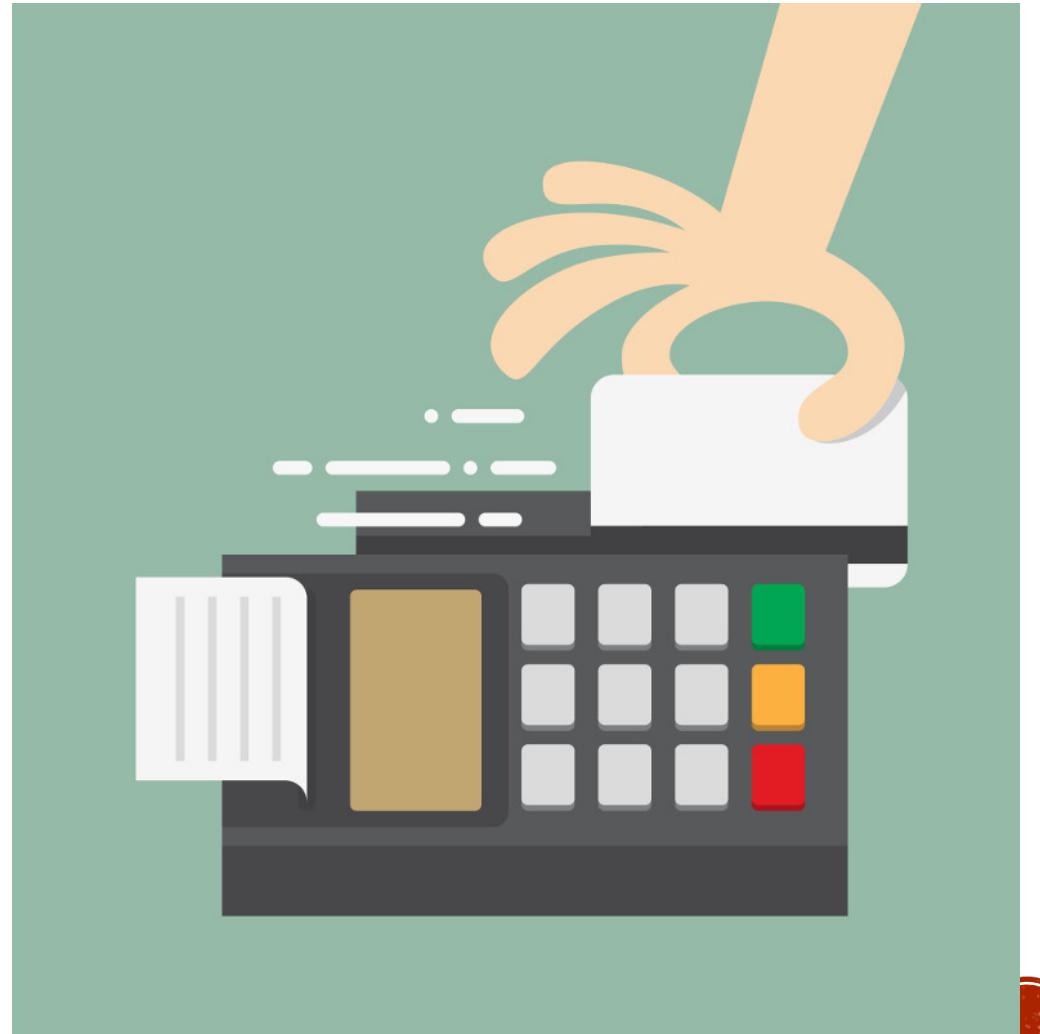
¿Cuáles de mis clientes son los mas propensos a dejarme por mi competidor?



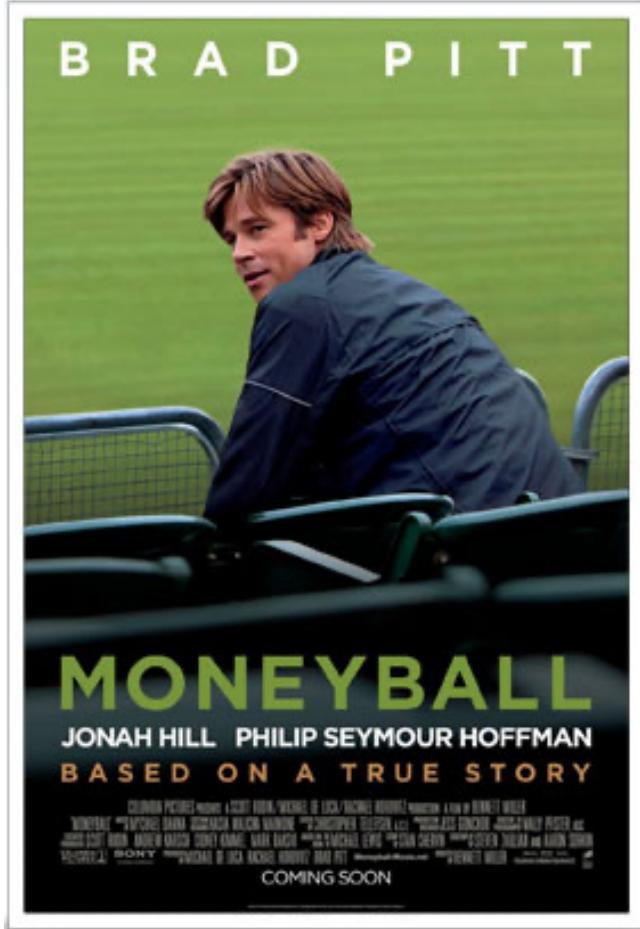
LA PREGUNTA

Detección de fraude de tarjeta de crédito

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

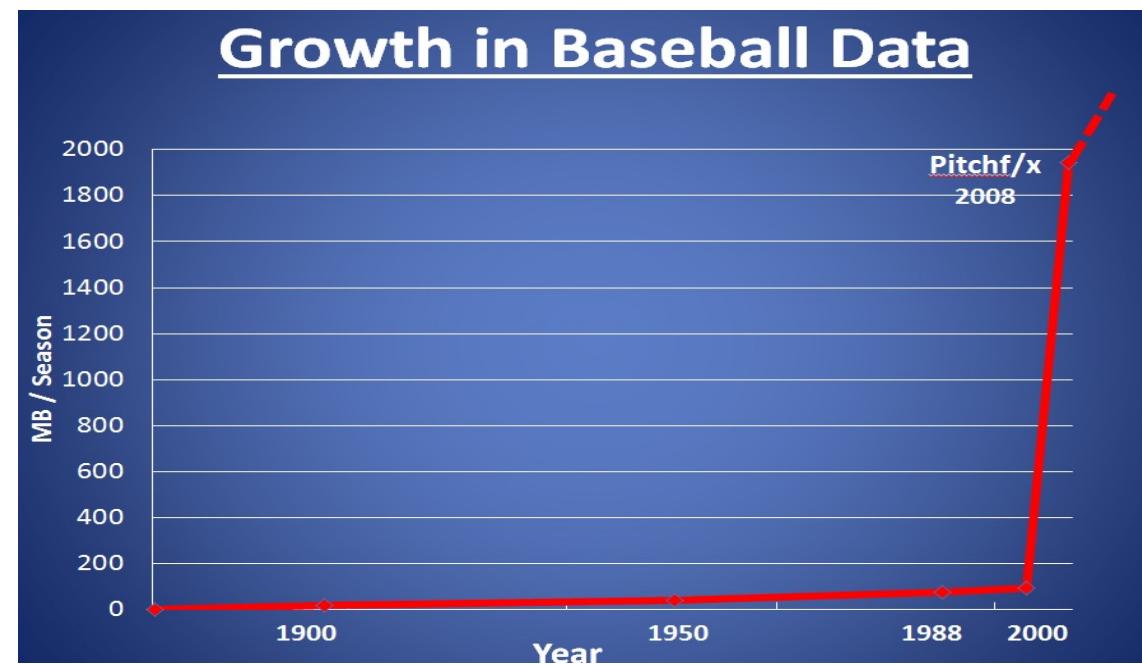


LA PREGUNTA



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

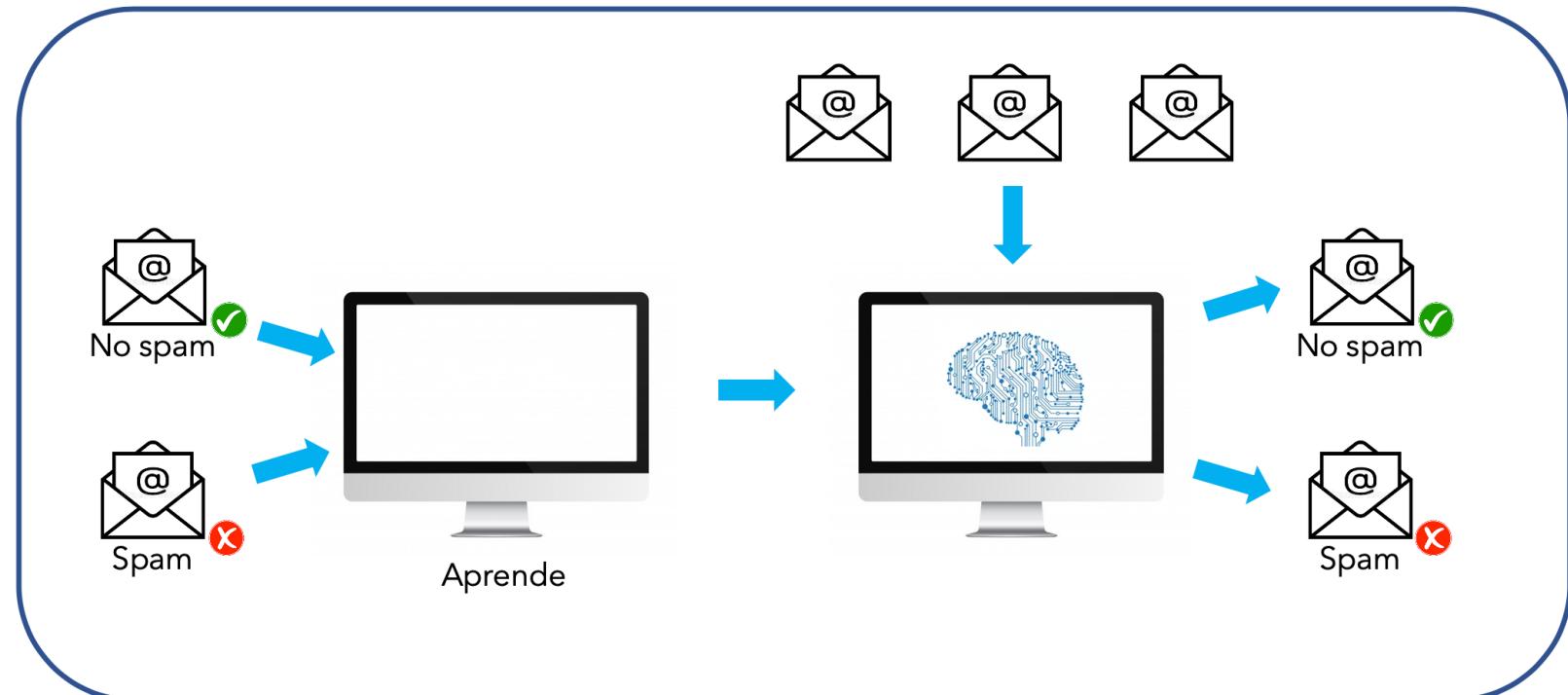
- Optimización de los recursos de un equipo de béisbol basada en datos
 - Pittsburgh Pirates
 - Houston Astros
 - Milwaukee Brewers
 - Boston Red Sox



LA PREGUNTA

Reconocimiento de spam

¿Este correo es spam o ham?



LA PREGUNTA

Sistemas de recomendación

¿Qué películas son las mas apropiadas para este usuario?



LA PREGUNTA

Búsqueda de información

¿Cuales 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:

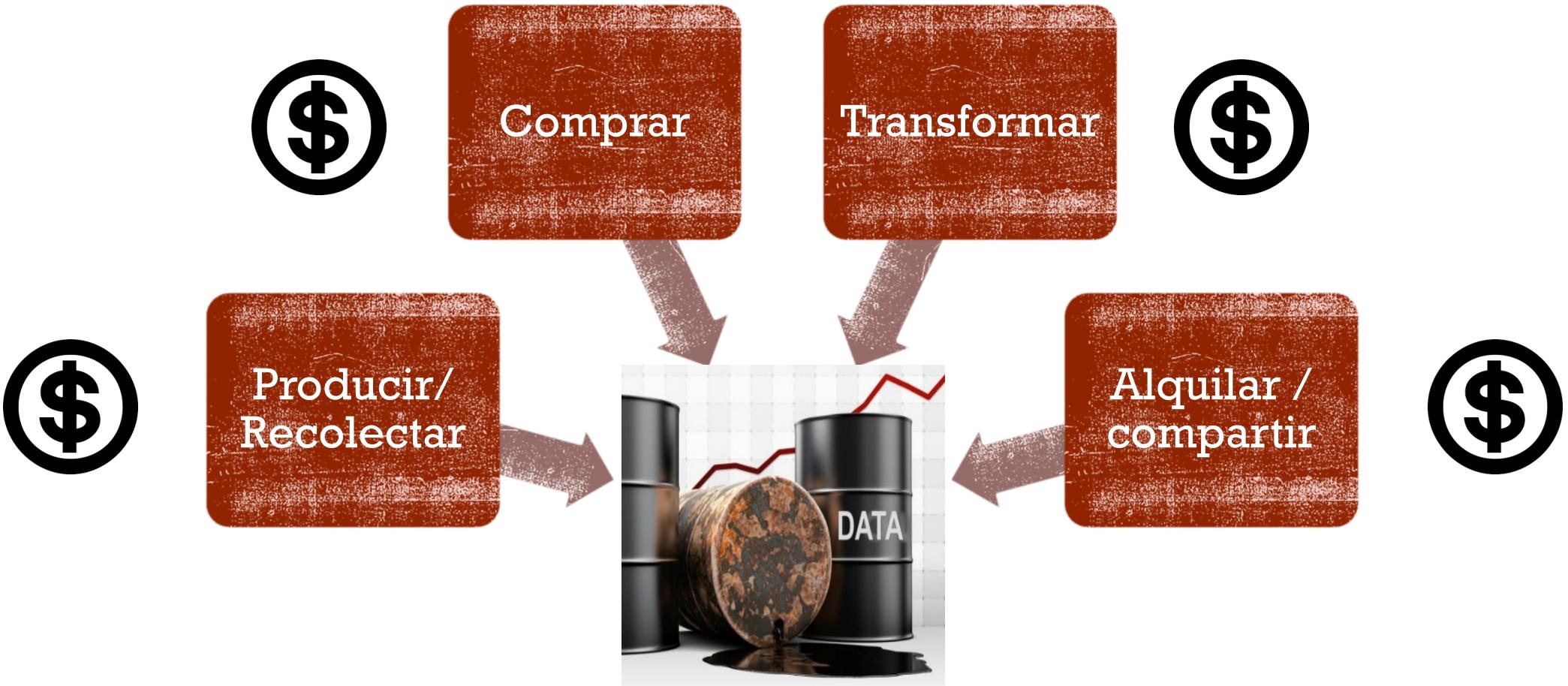
Resultados
Conocimiento
Estrategia



¿Cómo puede una empresa adquirir un activo estratégico?



¿Cómo puede una empresa adquirir un activo estratégico?



LOS DATOS



Materia prima

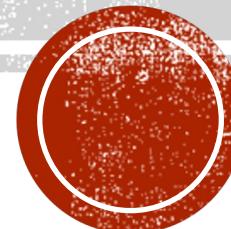
“Huella dactilar”

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)



TALLER: TELCO

En concordancia con la metodología CRISP-DM, esta tarea corresponde al entendimiento del negocio y la comprensión de los datos:

1. Defina un par de preguntas de negocio a responder en cada proyecto
2. Defina el tipo de información interna y externa que podría ayudar a responderlas.
3. Incluya en su descripción de los datos requeridos restricciones a tener en cuenta, como por ejemplo la antigüedad de los clientes o la temporalidad de los datos requeridos
4. Identifique los resultados esperados que podrían ayudar a tomar nuevas decisiones y cumplir con el objetivo establecido.



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 (normal, log)
- Decidir las variables a utilizar
- Crear nuevas variables si se estima necesario



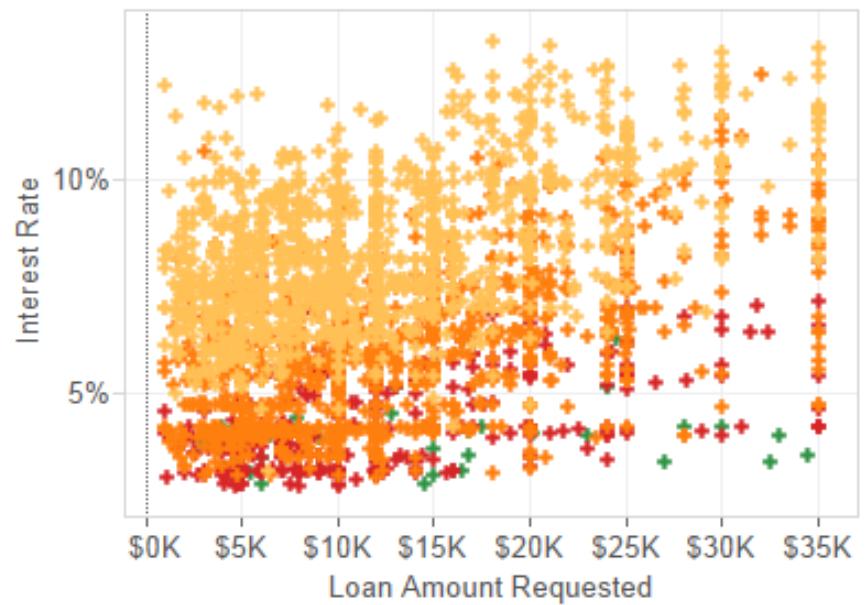
ANÁLISIS EXPLORATORIO



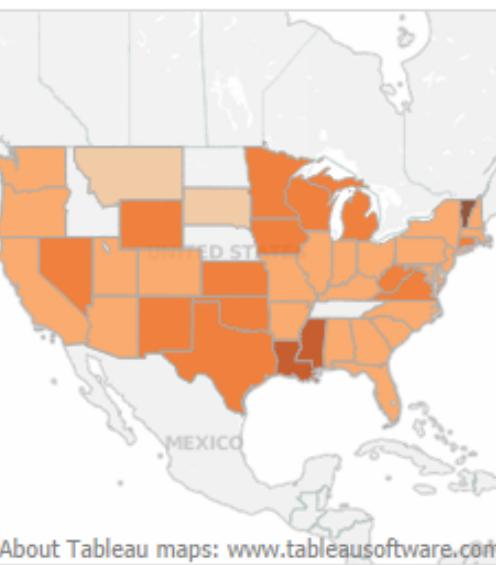
- ¿Son los datos disponibles suficientes para poder responder la pregunta de investigación?



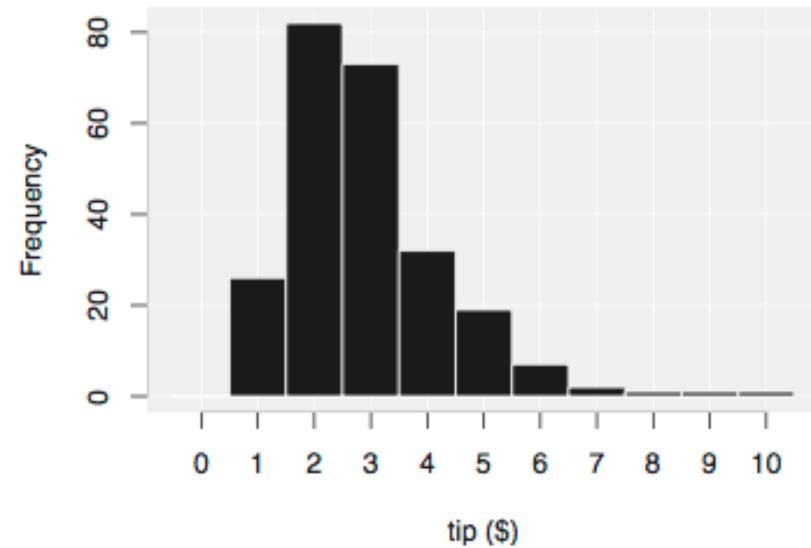
Scatter Plot



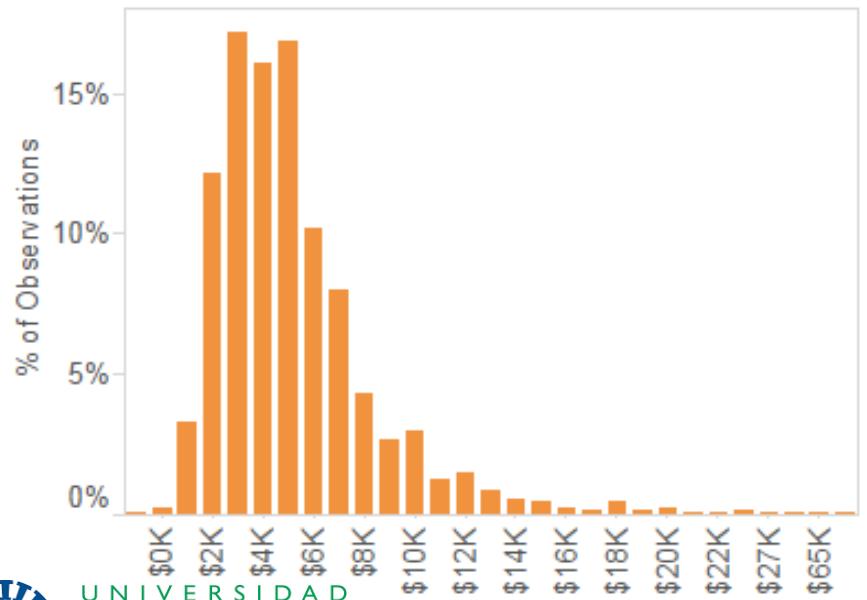
Map



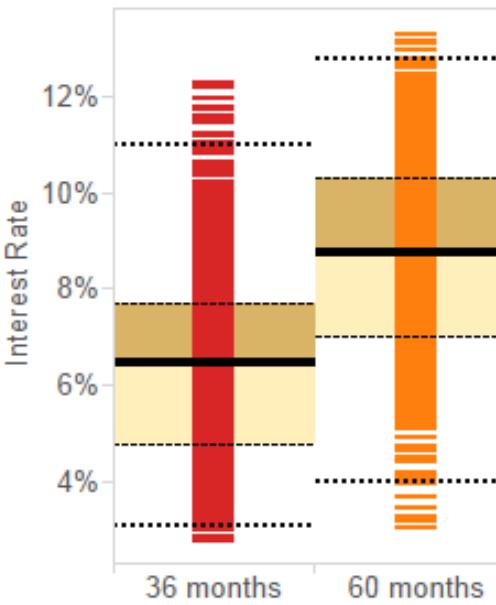
Bin width of \$1



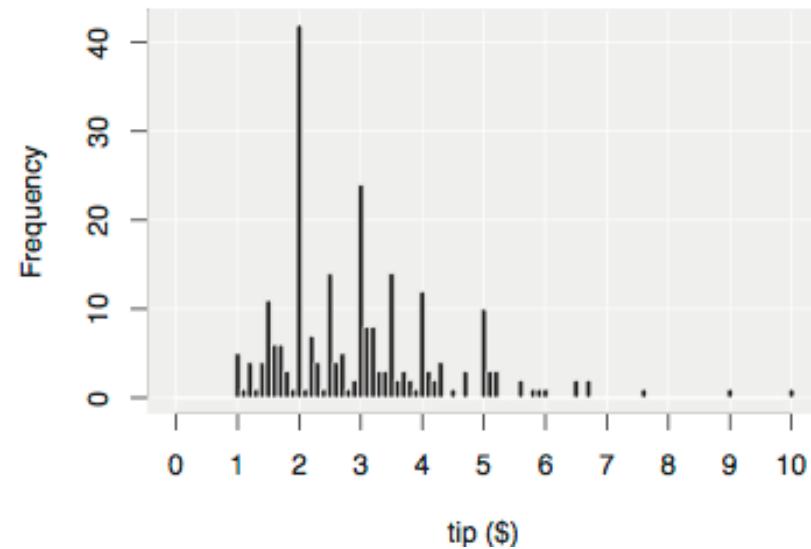
Histogram - Monthly Income

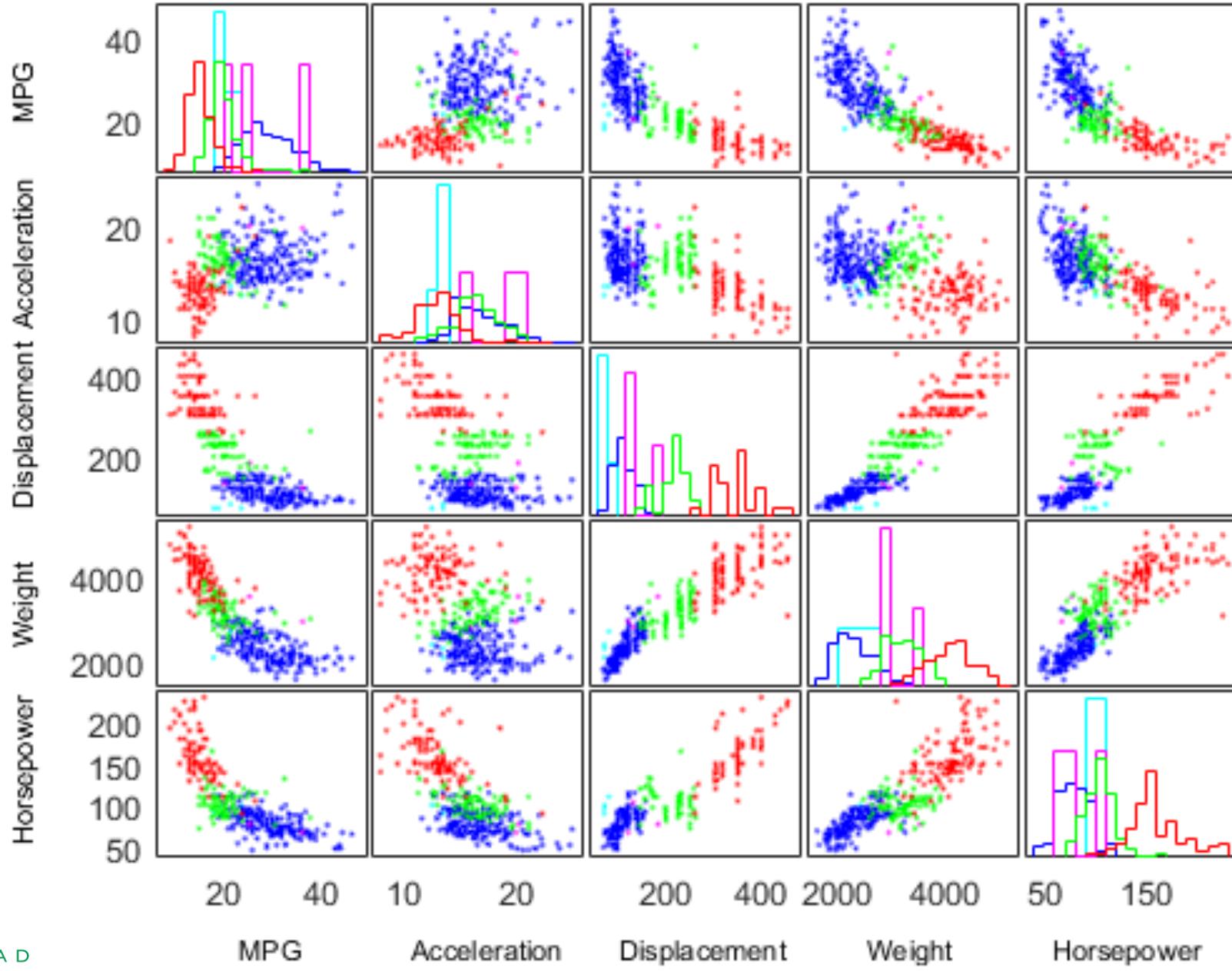


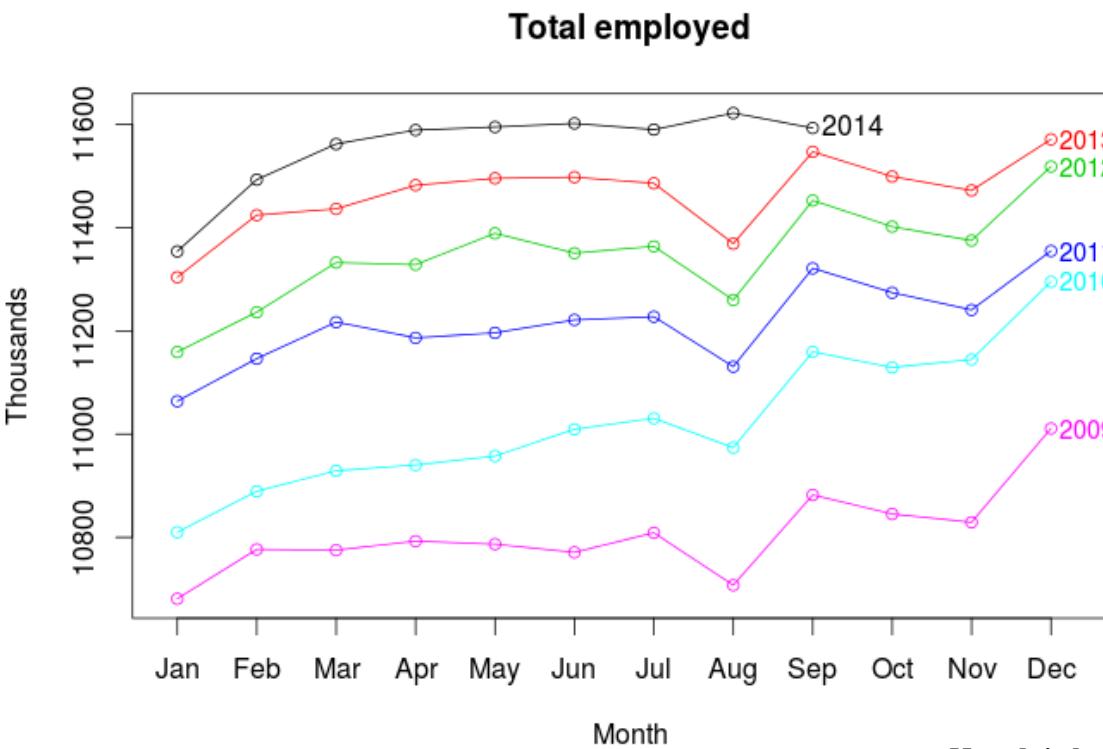
Box and Whisker Plot



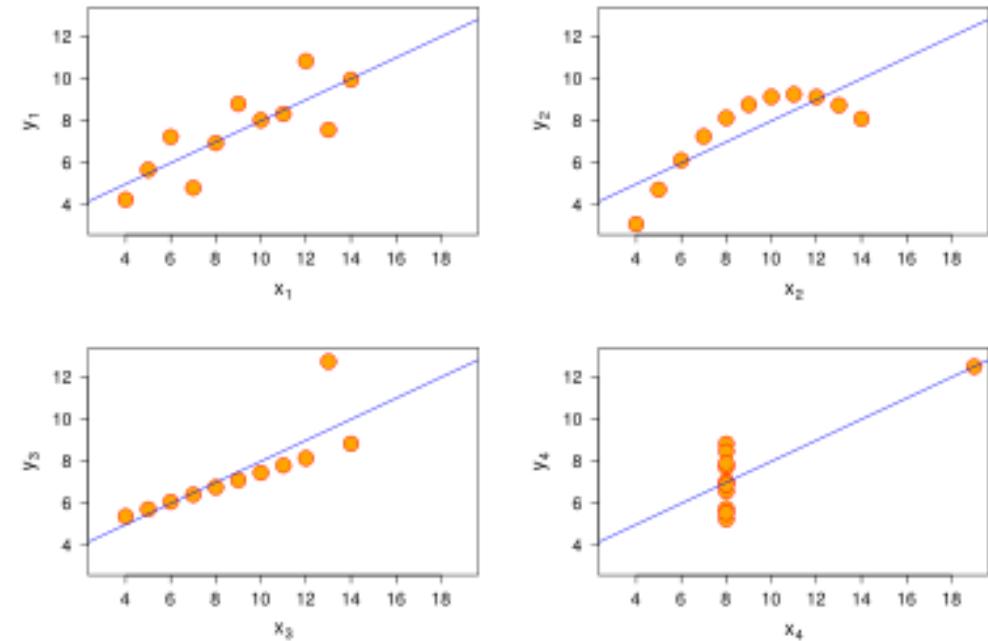
Bin width of 10c







Hyndsiht



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



EJEMPLO DE ANÁLISIS EXPLORATORIO DE DATOS

- 02-EDA-Ejemplo
 - Análisis Exploratorio del problema DatosCorazon 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

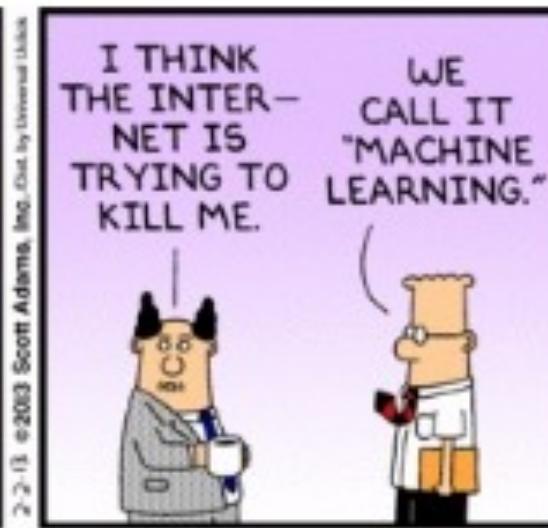
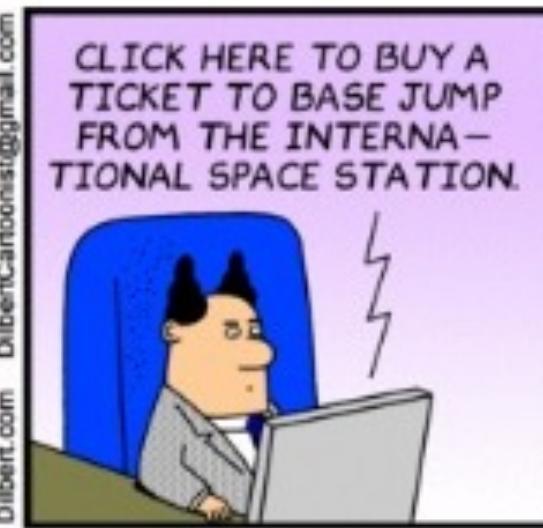
DESPLIEGUE:

Resultados
Conocimiento
Estrategia



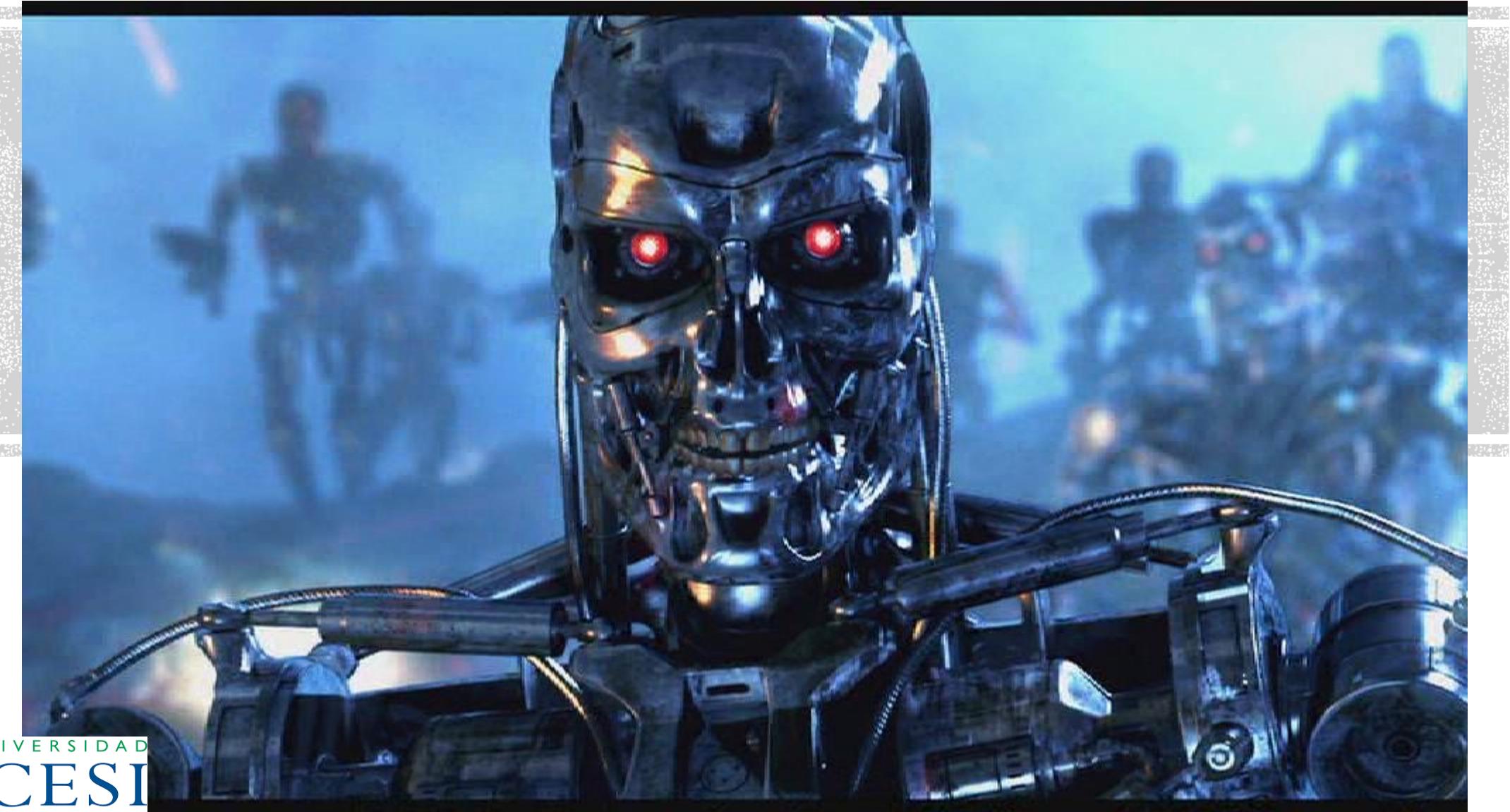
MODELOS DE ANALÍTICA

→ MACHINE LEARNING



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

MACHINE LEARNING



MACHINE LEARNING



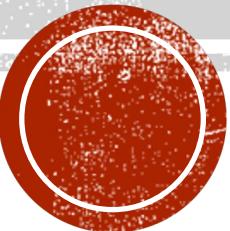
Kasparov vencido por Deep Blue



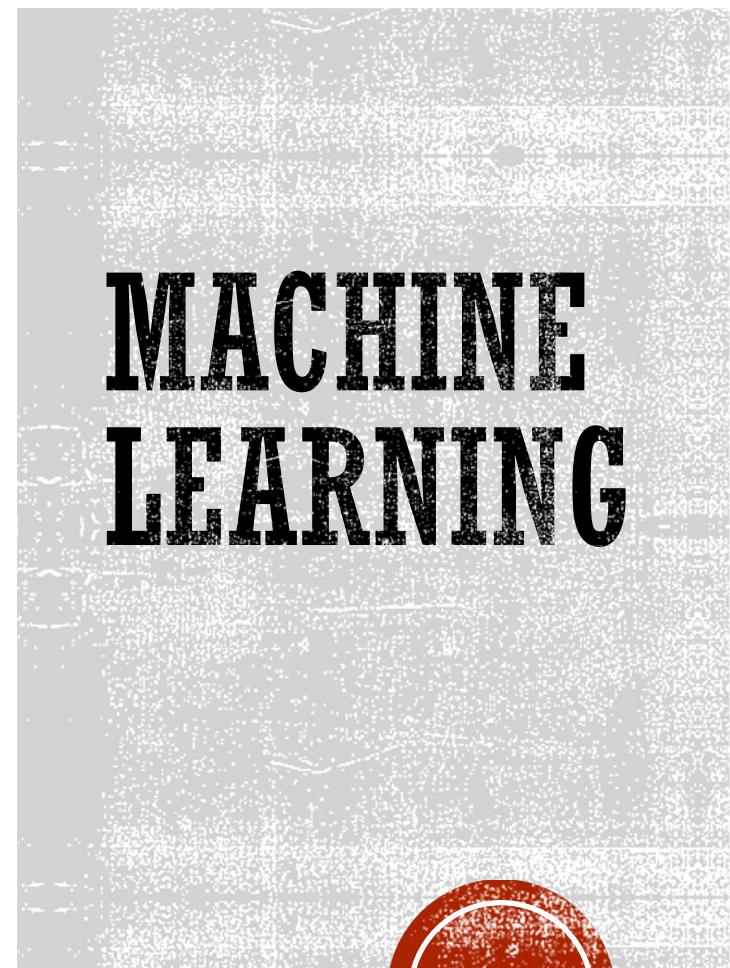
Watson gana Jeopardy

MACHINE LEARNING

The AI Forecaster: Machine Learning Takes on Weather Prediction



Improved weather forecasts from 2 weeks to 2 months could improve disaster preparedness for storms like Hurricane Irma, shown here battering Saint Martin in the Caribbean. A new machine learning weather prediction system shows promise on this subseasonal time frame. Credit: Netherlands Ministry of Defence, CC BY SA 4.0



APRENDIZAJE AUTOMÁTICO

- **¿Por qué es necesario?**

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

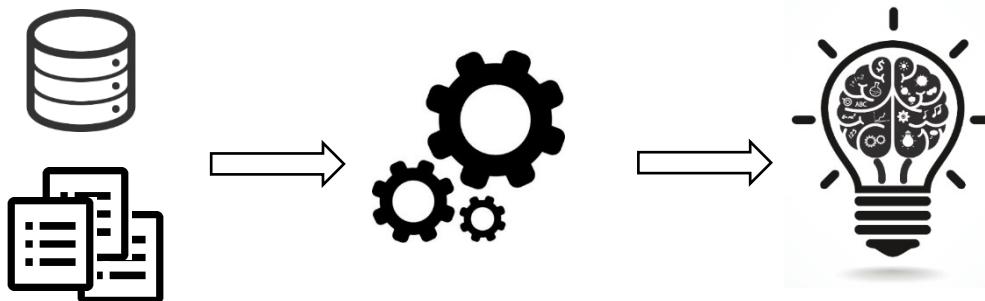
Las máquinas deben aprender por sí solas



APRENDIZAJE AUTOMÁTICO

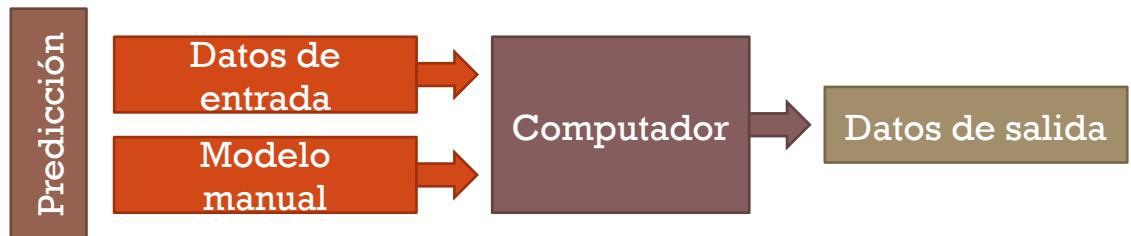
- Definición:

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

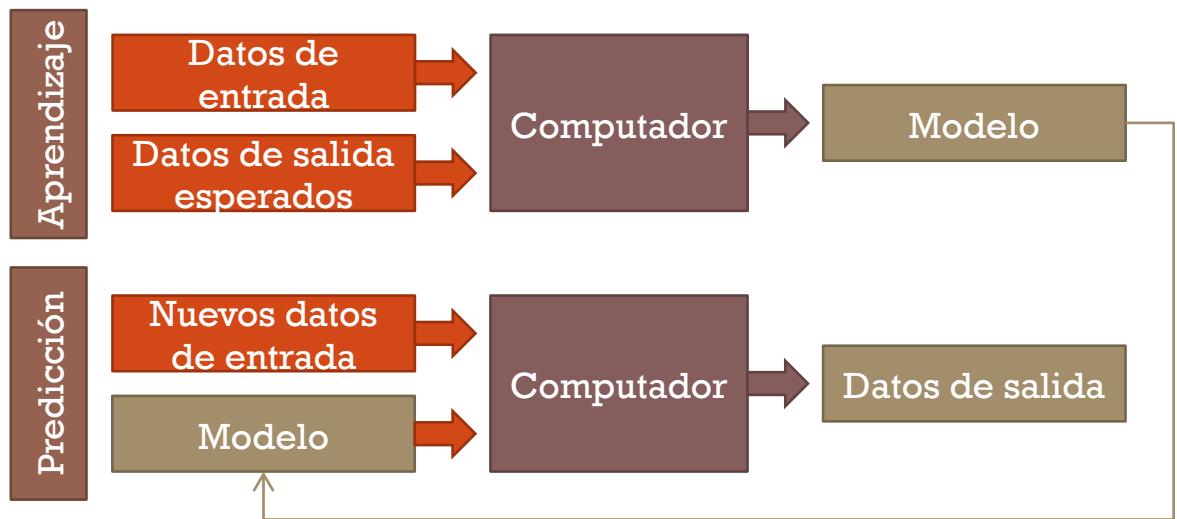


1. Andrew Ng, Stanford University, 2014

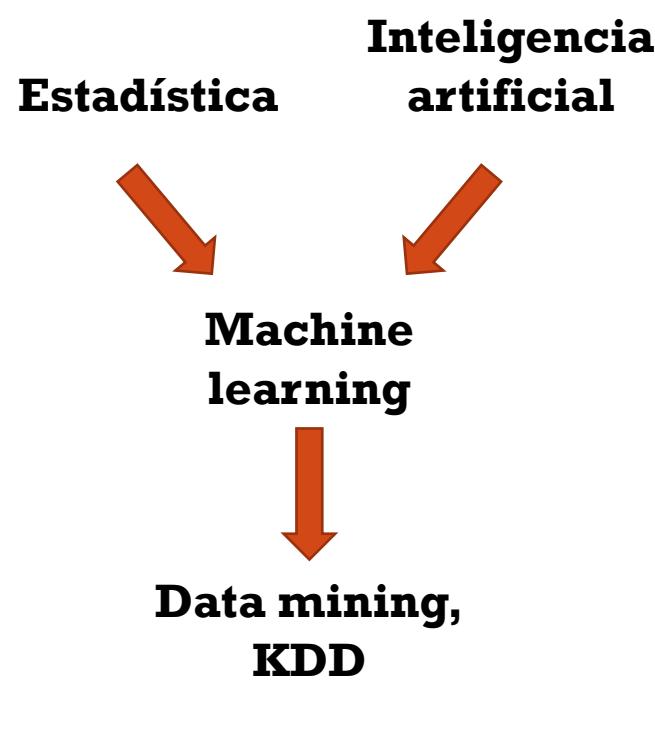
Modelo tradicional



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

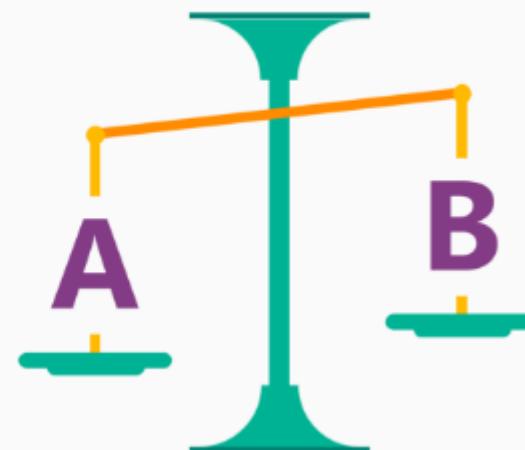


TAREAS DE APRENDIZAJE AUTOMÁTICO

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

Is this A or B?

Classification algorithms



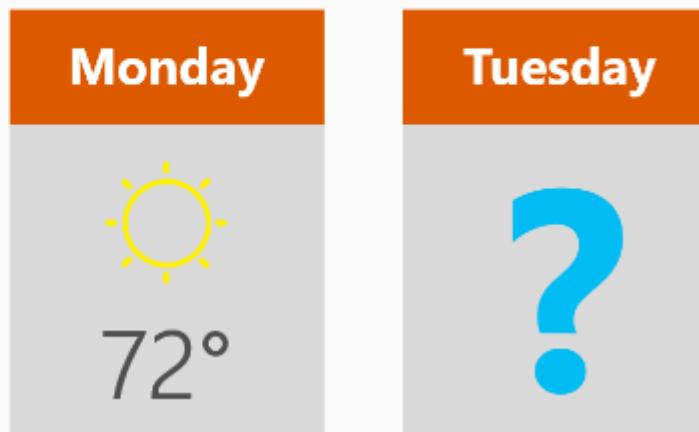
TAREAS DE APRENDIZAJE AUTOMÁTICO

- **¿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

How is this organized?

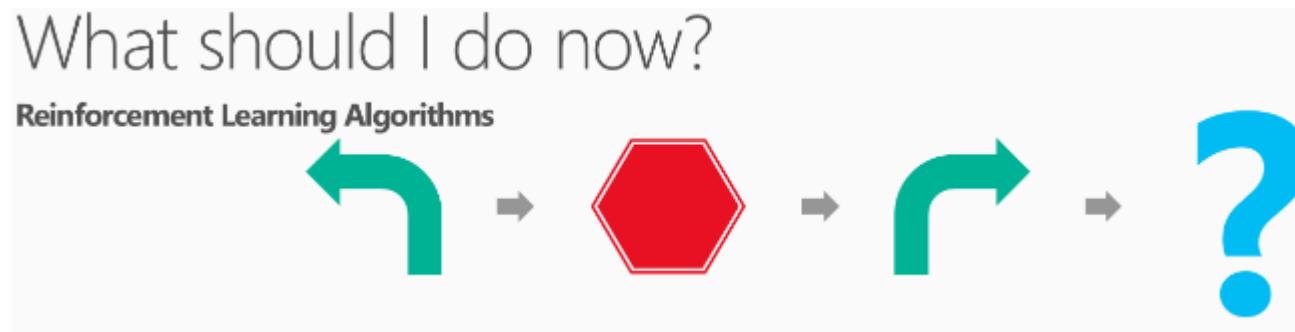
Clustering Algorithms



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
- **Refuerzo:** siguiente acción a tomar



TAREAS DE APRENDIZAJE AUTOMÁTICO

- **¿Qué podemos hacer con los datos?:**

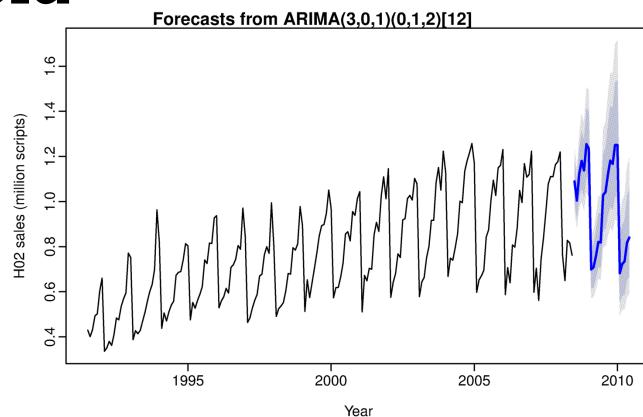
- **Clasificación:** predecir la categoría de un ítem
- **Regresión:** predecir un valor continuo
- **Detección de excepciones:** encontrar anomalías
- **Clustering:** encontrar grupos de elementos similares
- **Refuerzo:** siguiente acción a tomar
- **Asociaciones:** encontrar reglas de coocurrencia



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

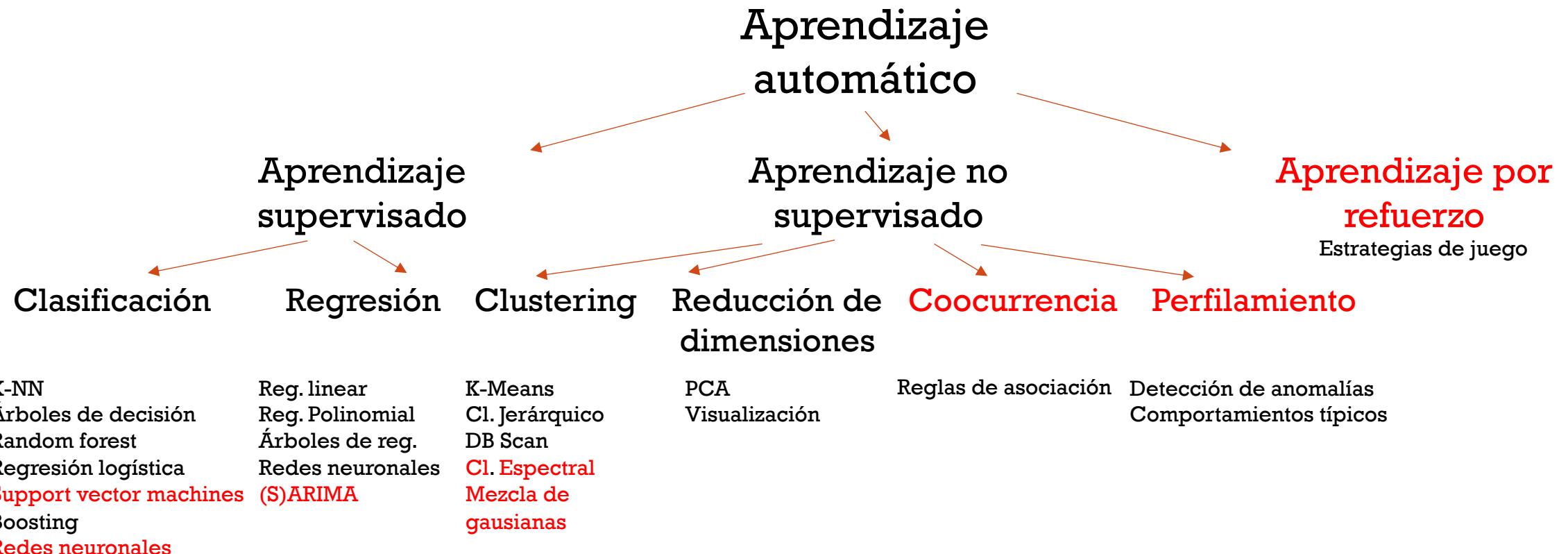


TALLER: TIPOS DE PROBLEMAS

1. ¿Cuáles son los prospectos más susceptibles de convertirse en clientes?
2. ¿Cuántas personas necesito en navidad para soportar la carga adicional?
3. Tengo dinero para hacer 4 campañas de mercadeo. ¿Cómo segmento los clientes?
4. Dado que este cliente tiene cuenta de ahorros, tarjeta de crédito y seguro de carro con mi entidad, ¿Cuál es el mejor producto para ofrecerle ahora?
5. ¿Cuáles son los clientes más parecidos a este cliente?
6. ¿Cuánto me va a comprar cada gran cliente el próximo mes?
7. ¿Cuál es la probabilidad de que este cliente me pague en los próximos 30 días?
8. ¿A qué máquinas debería hacerles mantenimiento?
9. ¿Cuánta plata debería invertir en publicidad en radio, televisión e internet?
10. ¿Cuál es el nivel de inventario óptimo para cada producto en cada semana del año?
11. ¿Le estoy cobrando adecuadamente a mis clientes, teniendo en cuenta sus consumos?



TAXONOMÍA



PREGUNTA:

¿Por qué?
¿Cómo?
¿Cuáles?
¿Cuándo?

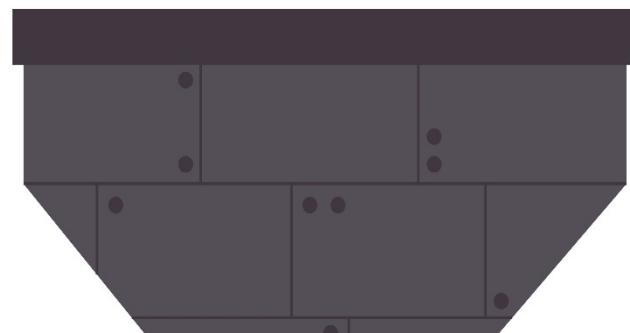
DATOS:

Materia prima



ANALISIS EXPLORATORIO:

Limpieza de datos
Preparación / transformación
Escogencia de variables



MODELO:

Implementación del
modelo de analítica

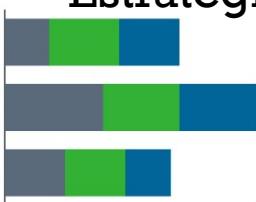


EVALUACION:

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

Marketing Analytics



EVALUACIÓN



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



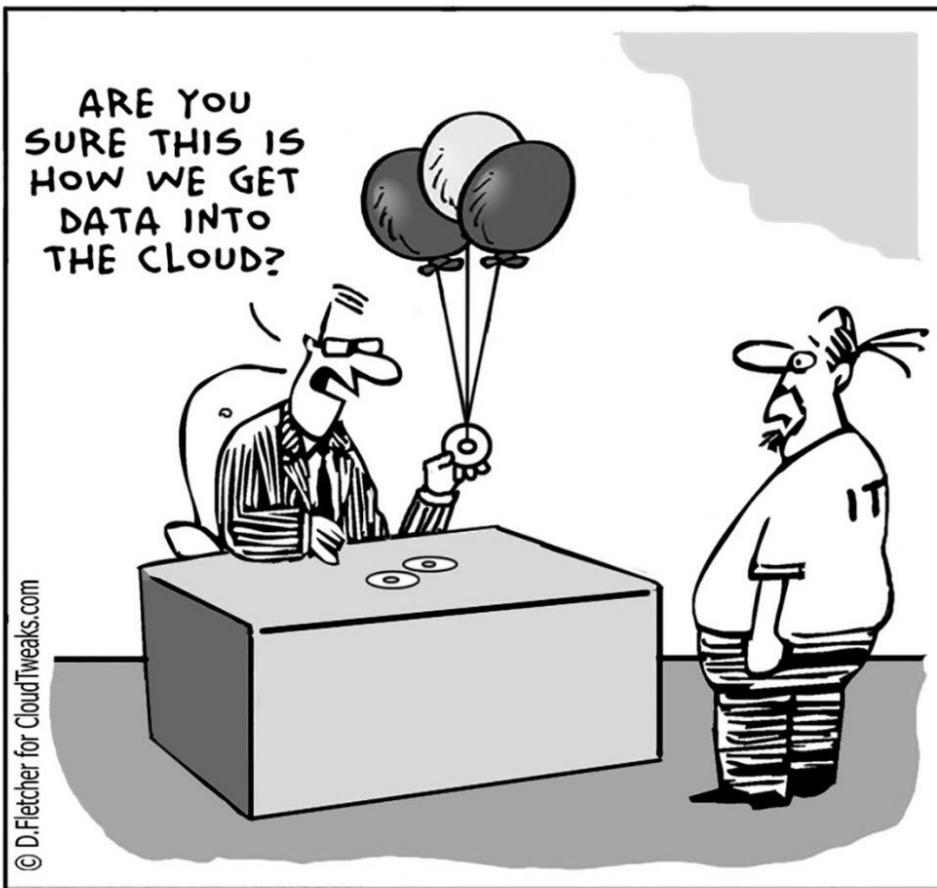
PREGUNTA:

¿Por qué?
¿Cómo?
¿Cuáles?
¿Cuándo?

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



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), versionar
- 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 coeficientes



DESPLIEGUE

Herramientas

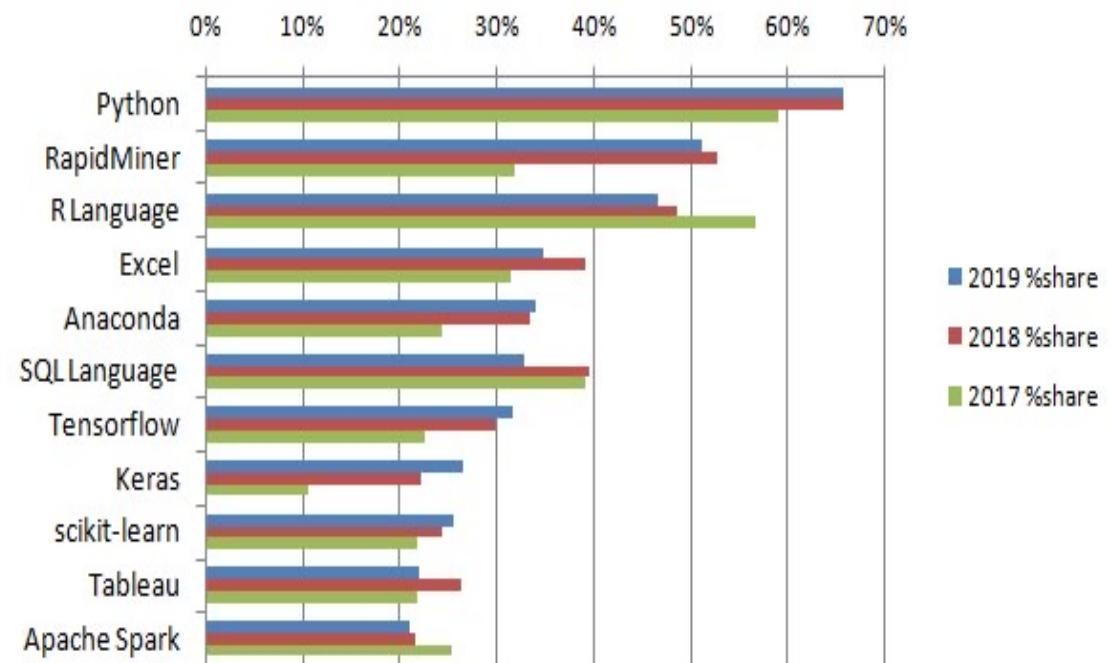
- Excel
- R
- Python
- Weka
- Knime
- SPSS
- Orange
- Matlab
- Rapidminer

- Anaconda
- Hadoop
- Cloudera
- AWS
- Azure
- Mahout
- Mllib
- ...



DESPLIEGUE

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



Top 10 Data Science Tools

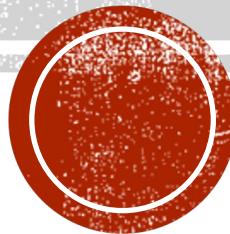


Gartner - Magic Quadrant for Data Science and Machine Learning Platforms





GRACIAS



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/>

