LA ANALÍTICA AVANZADA EN LA EMPRESA

Ignasi Puig de Dou

ignasi.puig@datancia.com
https://www.datancia.com/
https://www.theloyaltytool.com/

Indice

1. ¿Quienes somos?
2. ¿Qué es la Analítica Avanzada?
3. Tecnologías
4. Los proyectos de Analítica en las empresas

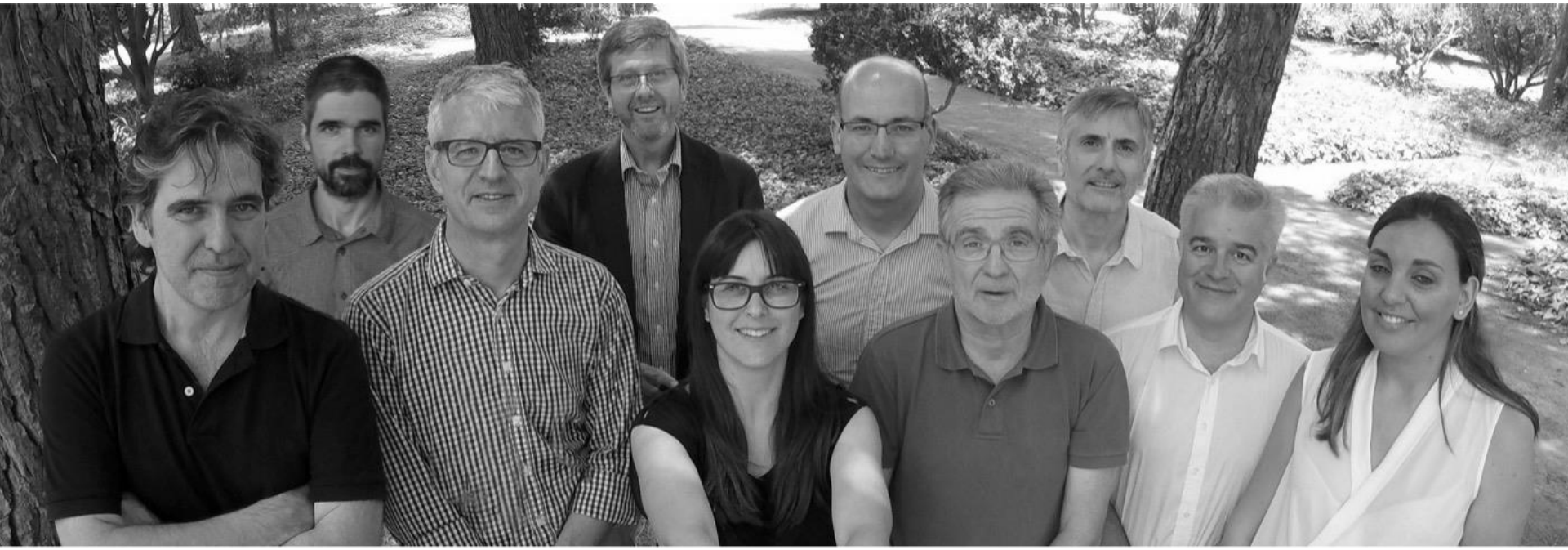
www.datancia.com

CUATRECASAS
ACELERA

Empresa seleccionada en la
4ª Edición de Cuatrecasas
Acelera

Somos una compañía de reciente creación, formada por profesionales de diferentes orígenes (Multinacionales, Universidad, Investigación y Consultoría).

Nos dedicamos a identificar cambios en el comportamiento de máquinas (y ahora también personas) monitorizando de manera regular las señales que producen.



2. Qué es la analítica avanzada?

Computadores comerciales

Bases de Datos

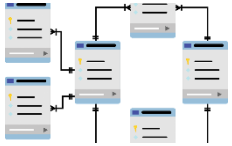
Bases de Datos relacionales

Ordenadores Personales

Redes

Internet WWW

50's 60's 70's 80's 90's 90's



50's 60's 70's 80's 90's 90's

PLC

Modbus
RFID

Carnegie Mellon
Coca cola

Machine Learning. Ethem Alpaydin. MIT Press

Laptops y móviles | Sensores | Smartphones | 4G | 5G

90's | 00's | 05's | 10's | 15's

- Benefon GPS (1999)
- Samsung accelerometer (2005)

- Smarter: more computer, less phone
- From talking to visualizing

LoRa sigfox

- Short range area network.
- 10-100m
- 250kbs

- Low power wide area network.
- 1-10km
- 50kbs

90's | 00's | 05's | 10's | 15's

Zigbee | LPWA | IPv6

DATANCIA EMPOWER YOUR DATA

CLUB EXCELENCIA EN GESTIÓN

II. Capacidad de cómputo

1994 amazon

2003

2006 hadoop

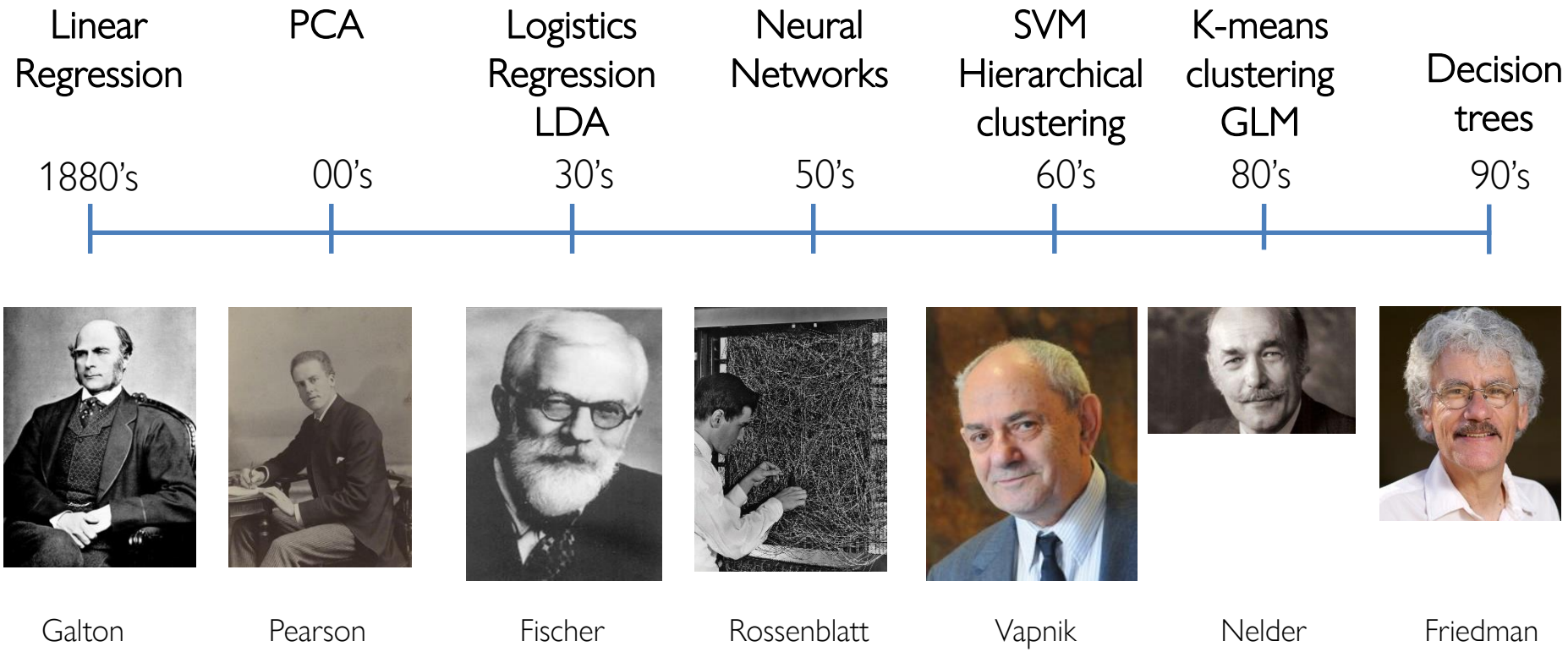
2008 cassandra

1995 Google

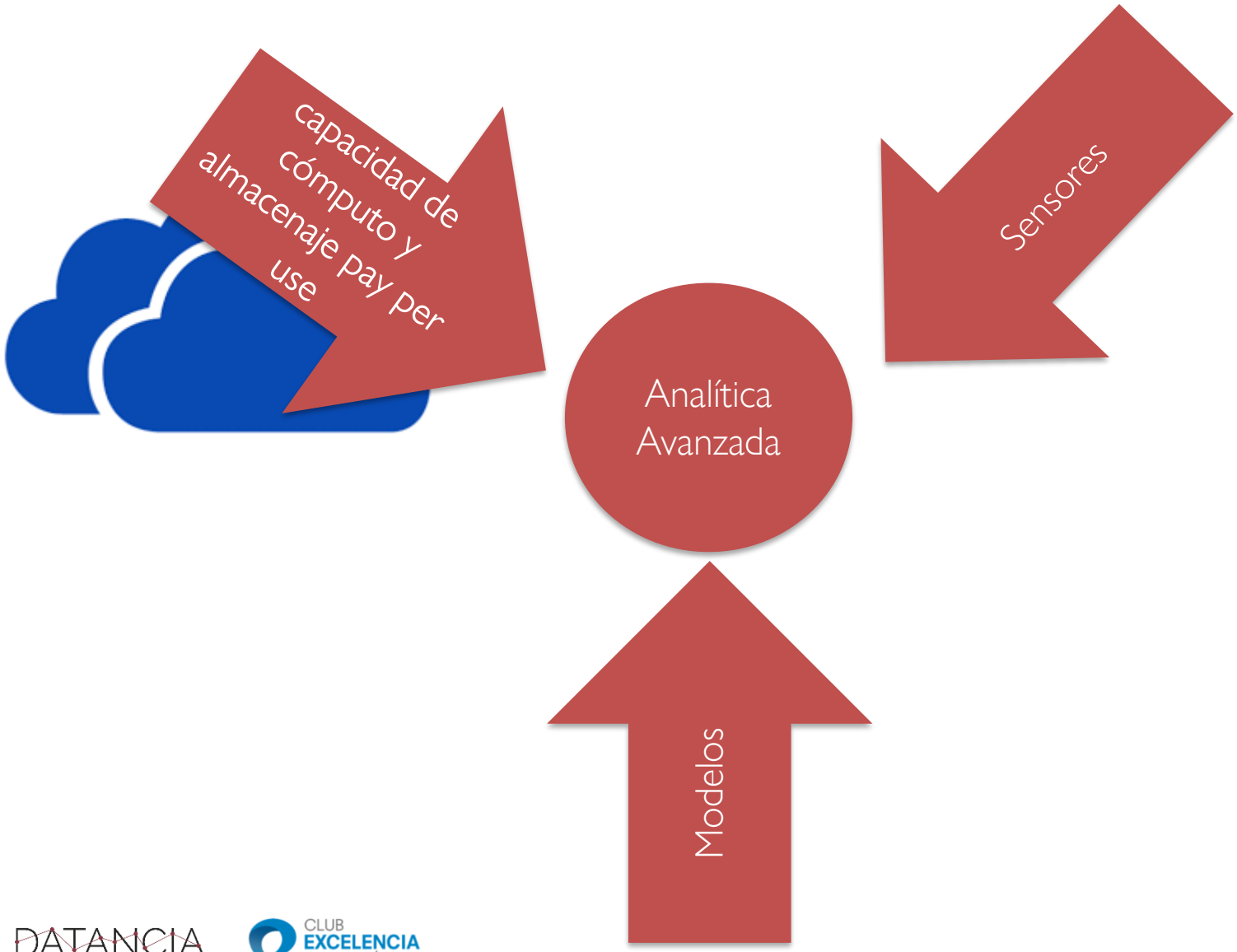
2003 Google File System

2007 amazon DynamoDB aws

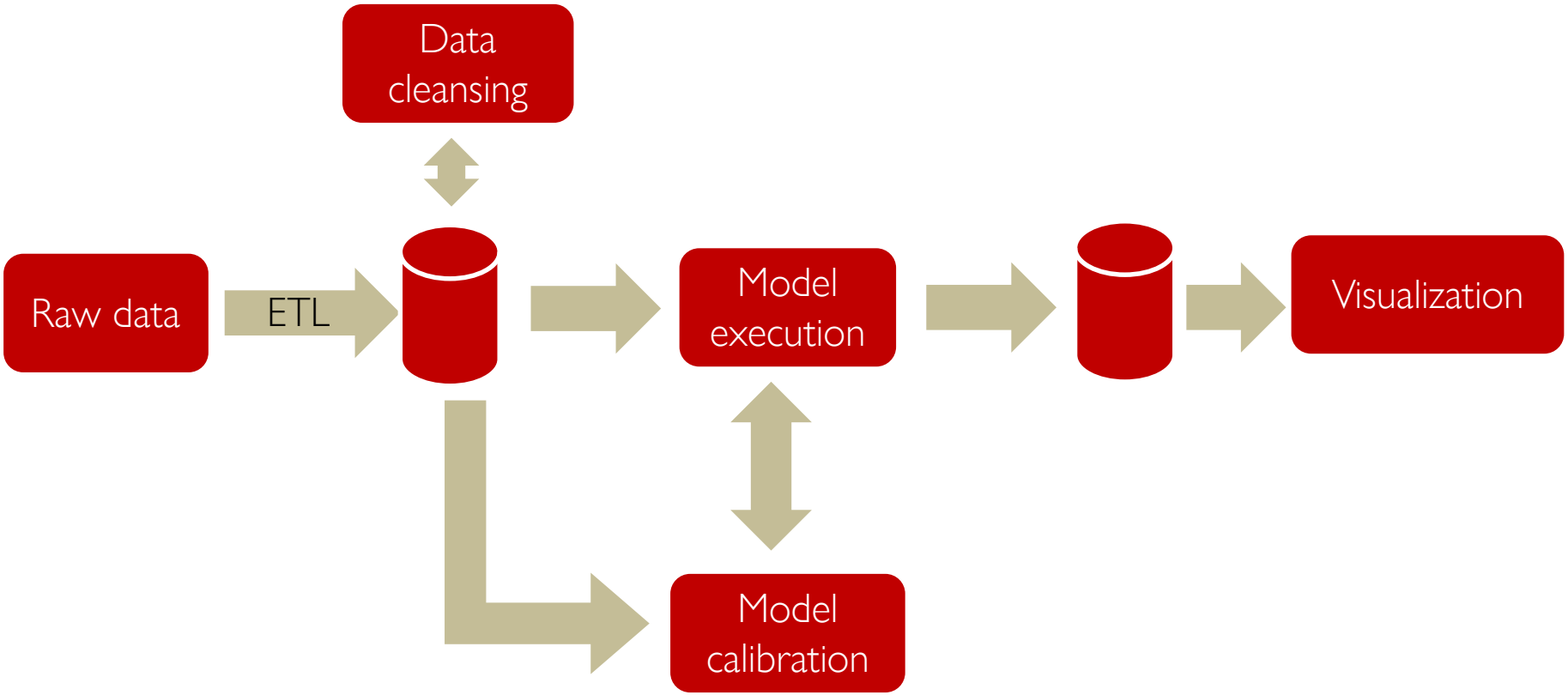
III. Modelos

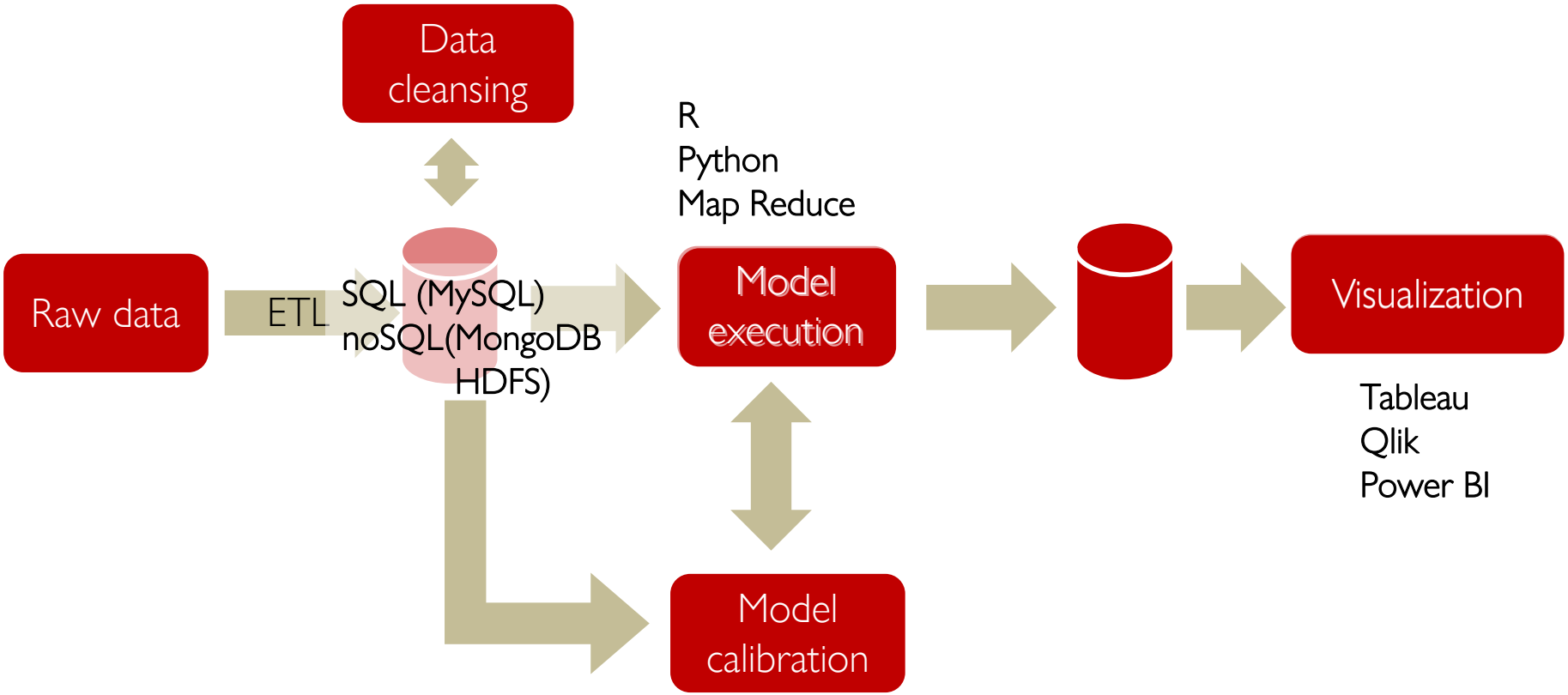


By Source (WP:NFCC#4), Fair use, https://en.wikipedia.org/w/index.php?curid=42979948

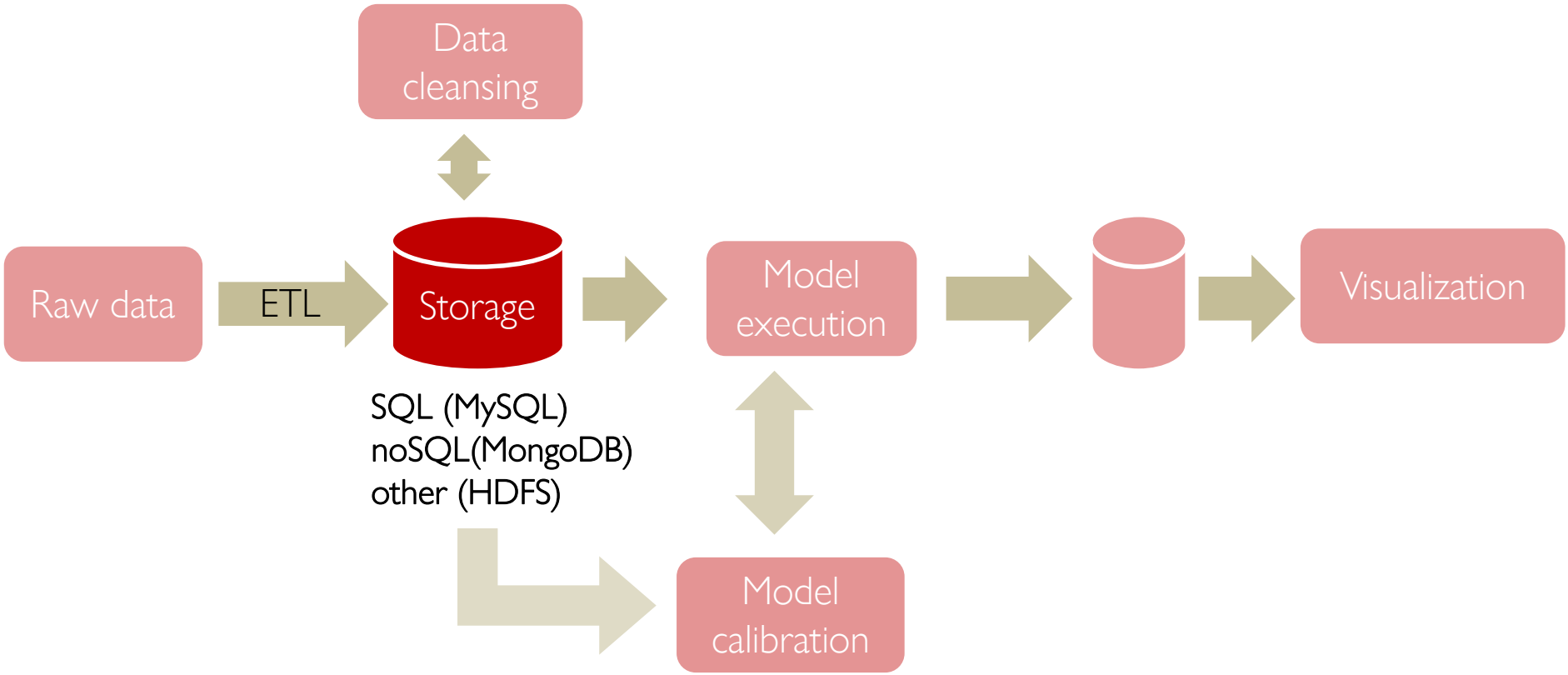


3. Tecnología

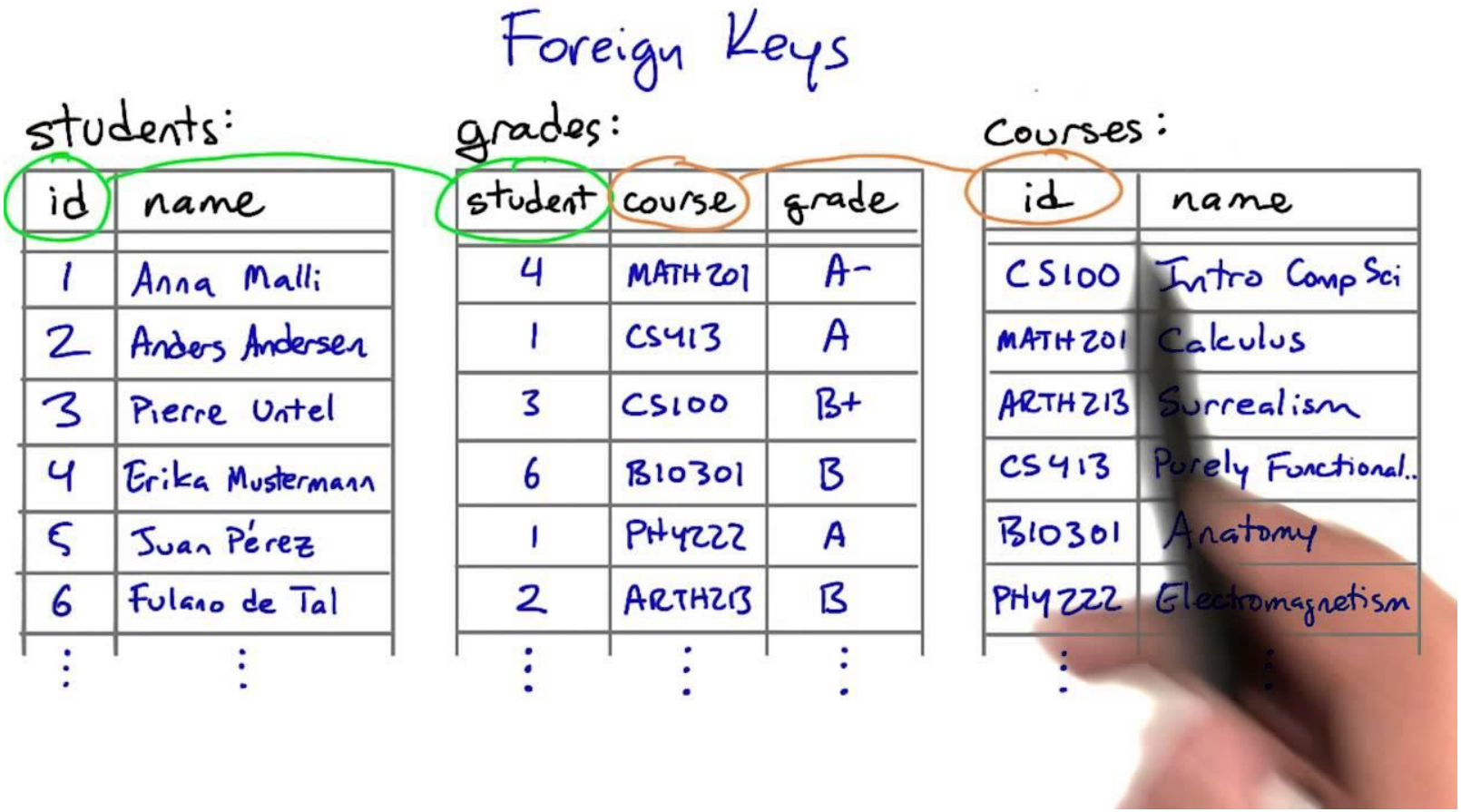




I. Almacenaje

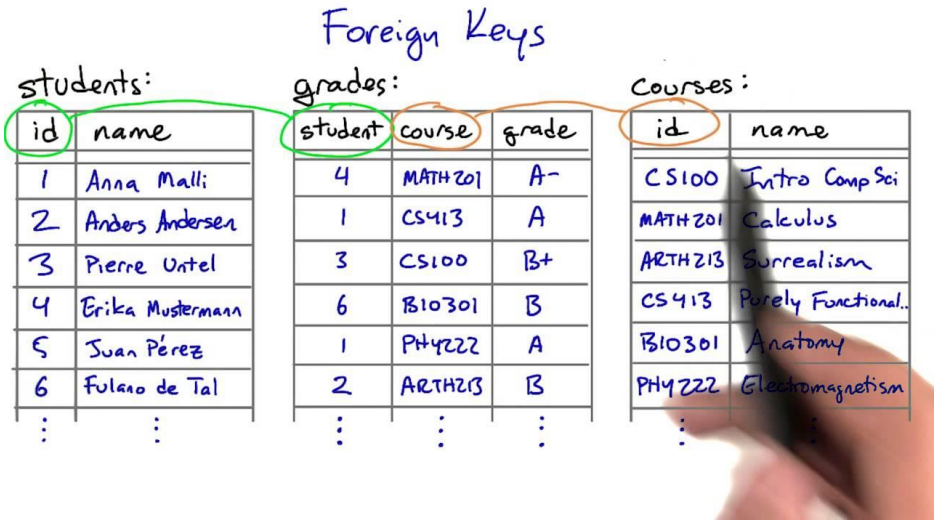


Bases de datos SQL o Relacionales



Bases de datos SQL o Relacionales

1. Datos agrupados en **tablas** y **relaciones**.
2. Las tablas contienen **registros** compuestos por **campos**, comunes y constantes dentro de la tabla.
3. Cada intersección de campo-registro solo puede contener un dato.



Las bases de datos relacionales son **ACID**

Atomicity

date	USD
10:00	2,000
11:00	1,300

700

date	USD
10:00	1,000
11:00	1,700

Isolation

200

date	USD
10:00	2,000
11:00	1,300
11:00	1,500

700

date	USD
10:00	1,000
11:00	1,700

Consistency

date	USD
10:00	2,000
11:00	1,300

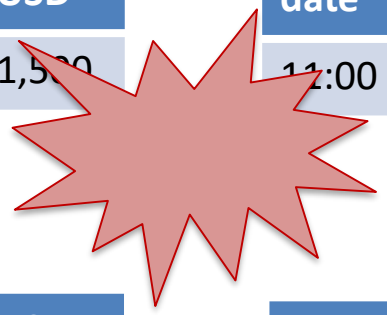
date	USD
10:00	1,000
11:00	1,700

cantidad total en el banco:
10:00 3,000USD
11:00 3,000USD

Durabilitiy

date	USD
11:00	1,5[illegible]

date	USD
11:00	1,700



date	USD
11:00	1,500

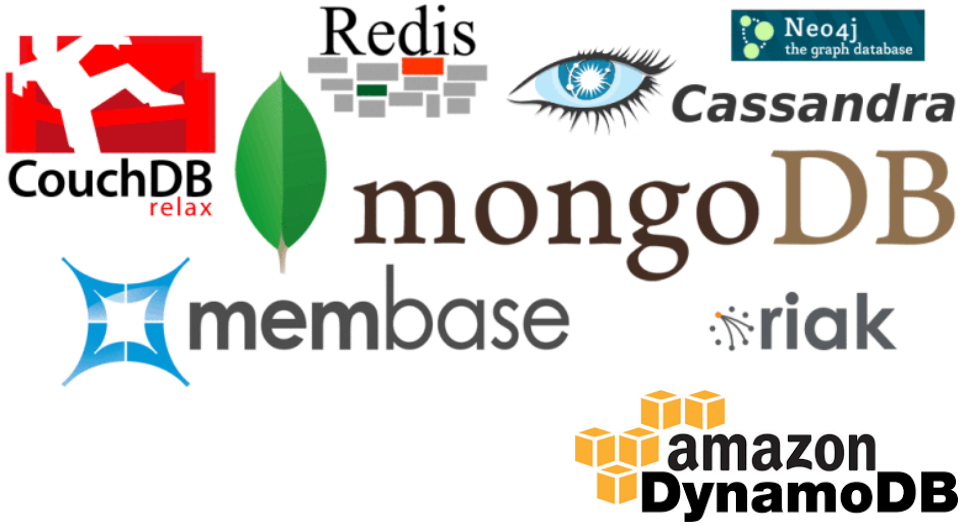
date	USD
11:00	1,700

Bases de datos SQL o Relacionales

1. **Atomicity:** cada transacción es todo o nada
2. **Consistency:** cada transacción lleva la base de datos de un estado válido a otro (los datos escritos son válidos de acuerdo con todas las reglas definidas: constraints, cascadas, triggers, etc.)
3. **Isolation:** transacciones concurrentes son administradas como procesamiento secuencial
4. **Durability:** una vez que se confirma una transacción, es ininmutable

La necesidad de mantener las propiedades ACID en las bases de datos compatibles con SQL hace que sean más difíciles de escalar (por ejemplo, consistencia basada en clúster)

Bases de datos no-SQL



Las bases de datos no-SQL simplifican el procesamiento al renunciar a algunas de las propiedades **ACID** para convertirlas en **BASIC**

Basically available: una query siempre recibirá una respuesta, aunque puede ser inconsistente o fallo
Soft state: las transacciones no son write-consistent ni las réplicas son mutuamente consistentes todo el tiempo
Eventually consistent: la base de datos será a largo plazo consistente una vez se dejen de recibir transacciones y los datos se propaguen a todas partes.

Bases de datos no-SQL

1. orientada a **clave-valor**
2. orientada a **documento**
3. orientada a **columna**
4. orientada a **gráfico**

Bases de datos no-SQL
bases de datos clave-valor (e.g. Redis, Dynamo DB)
Estructura de diccionario o table hash

- No hay información de los valores almacenaods.
- No se pueden relacionar la claves.
- Muy simples

Key	Value
K1	AAA,BBB,CCC
K2	AAA,BBB
K3	AAA,DDD
K4	AAA,2,01/01/2015
K5	3,ZZZ,5623

Bases de datos no-SQL

Bases de datos orientadas a **documento**(e.g. MongoDB)

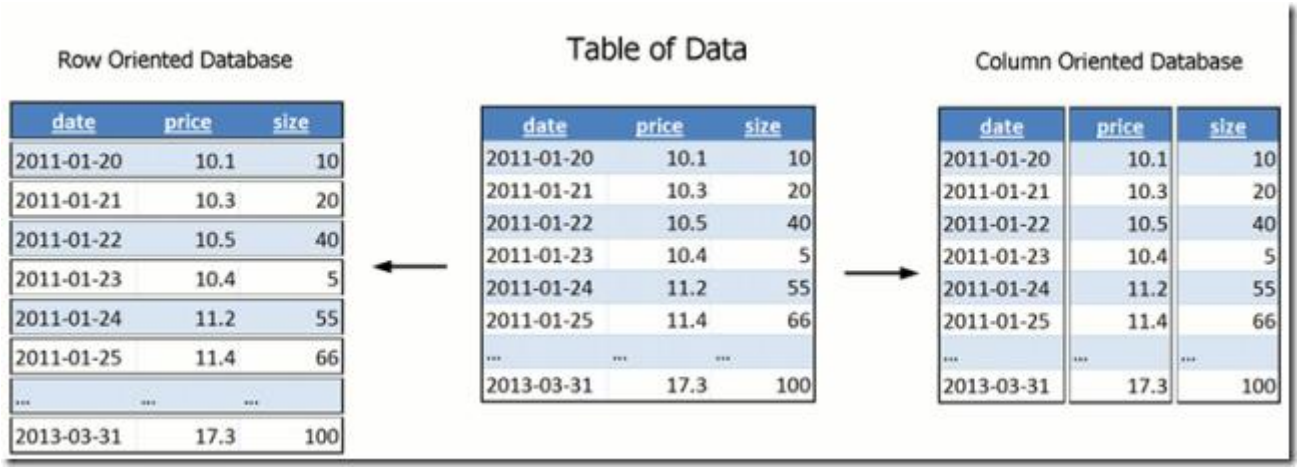
Parecidas a las BB.DD. clave-valor pero el valor almacenado no es totalmente "opaco". Existe un formato explícito del mismo (XML, JSON) que permite ser interpretado por el gestor de la base de datos.

<table>
<tr>
<td>DNI: 00A</td>
<td>name: Smith
age: 57
address:
<table>
<tr><td>add1</td><td>BCN</td></tr>
<tr><td>add2</td><td>MAD</td></tr>
</table>
</td>
</tr>
<tr>
<td>DNI: 00B</td>
<td>name: Codd
age: 57
languages:
<table>
<tr><td>English</td><td>certification:
first: A
advnaced: B
proficiency: F</td></tr>
<tr><td>Spanish</td><td>level:
Native</td></tr>
</table>
</td>
</tr>
</table>

Bases de datos no-SQL
Bases de datos columna (e.g. Vertica)

BB.DD. relacionales son eficientes obteniendo información de un registro completo (fila): datos del producto XYZ.

La ejecución de búsquedas por conjuntos son más eficientes en las BB.DD. columna: obtener el nombre del proveedor de los productos con precios comprendidos entre 10 y 15€.

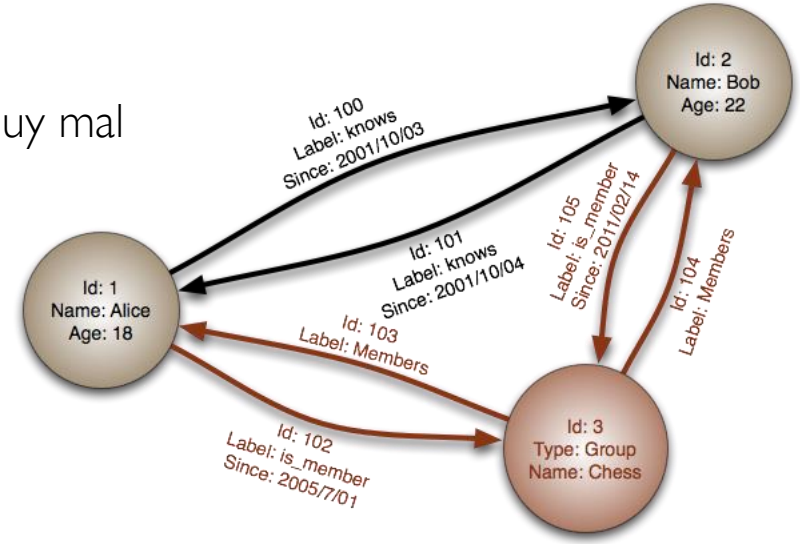


Bases de datos no-SQL

Bases de datos orientadas a **grafos**(e.g. Neo4j)

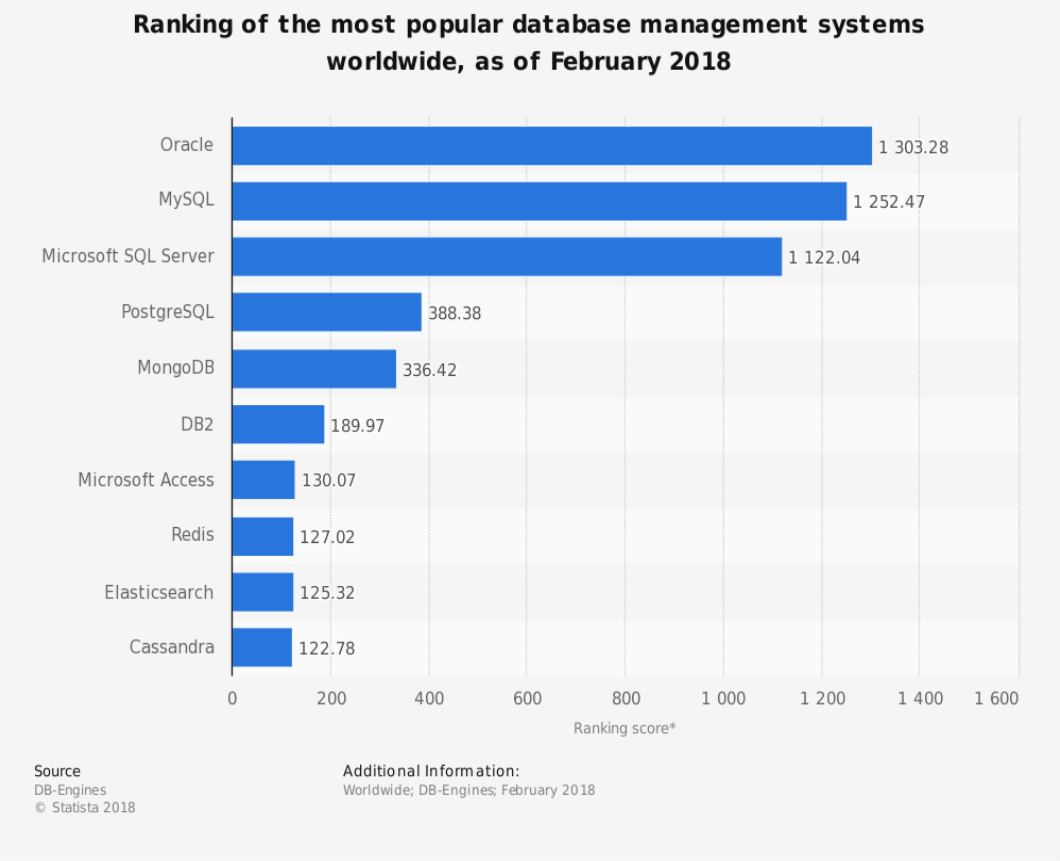
Nodos (e.g. personas) unidas por arcos (e.g. relaciones)

Las estructuras de nodo-arco se gestionan muy mal en bases de datos relacionales

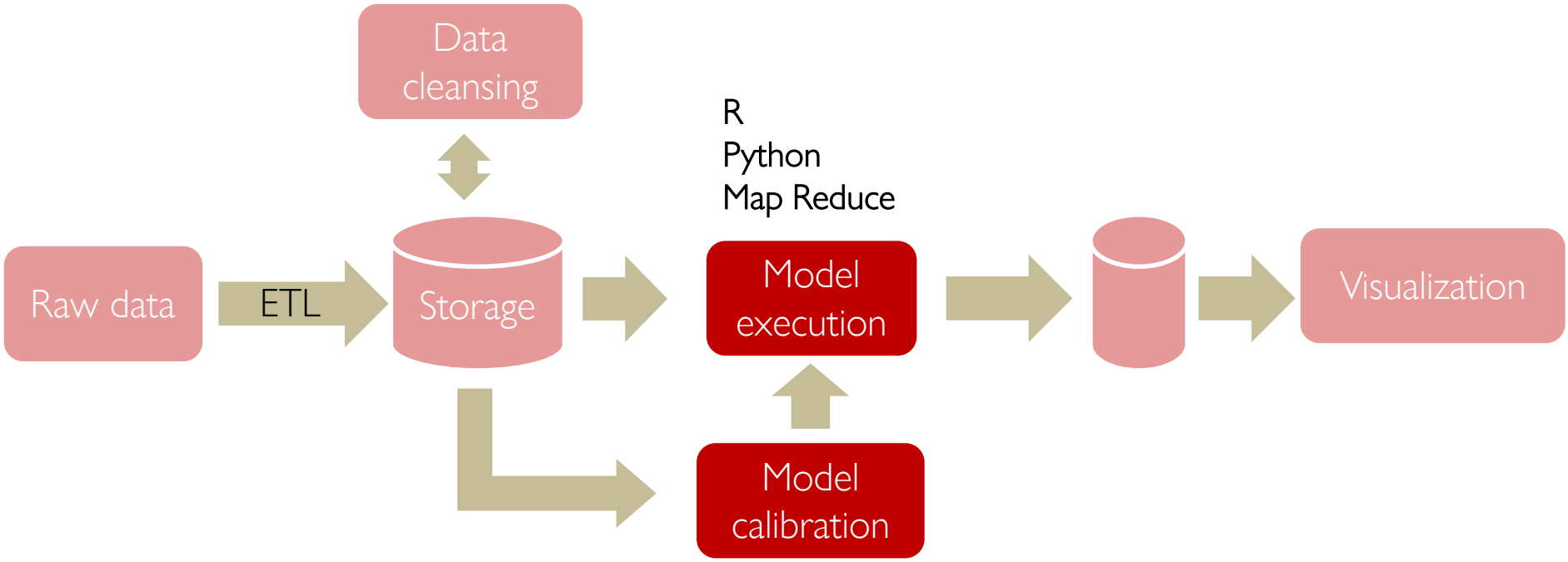


https://commons.wikimedia.org/w/index.php?curid=19279472

Uso de bases de datos



www.statista.com



Model execution software

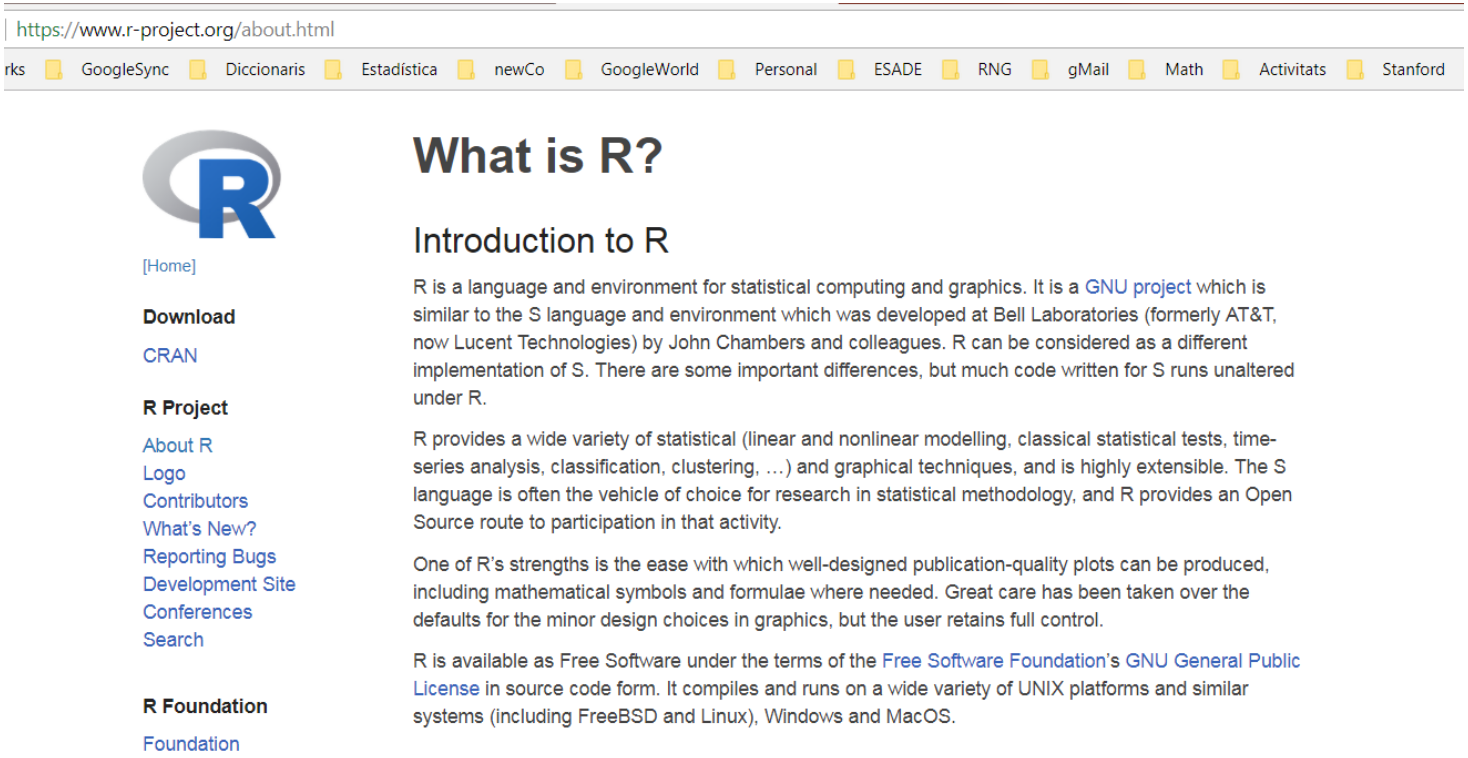
Do-it-yourself

We-do-it-for-you

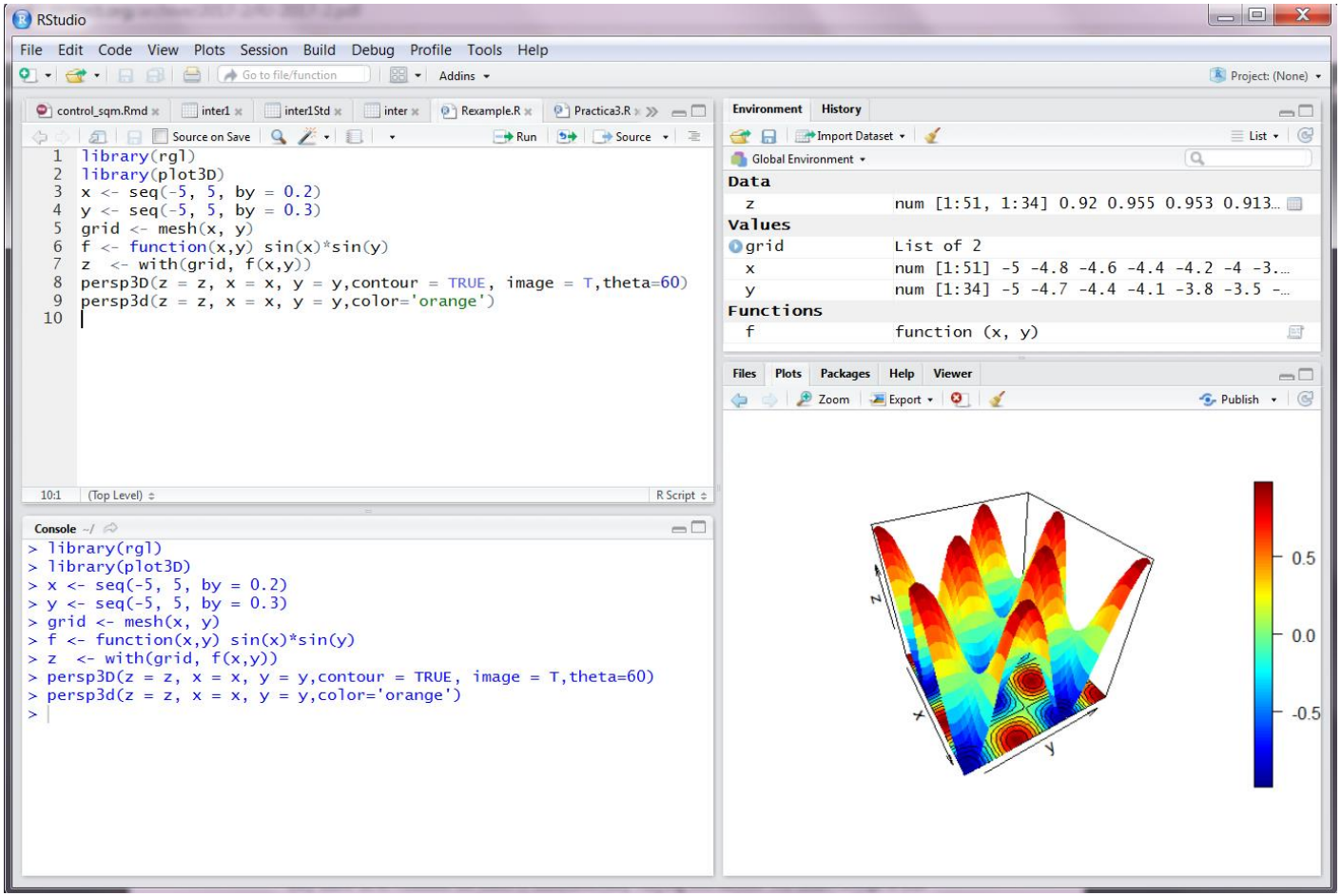
DataRobot

Azure Machine Learning

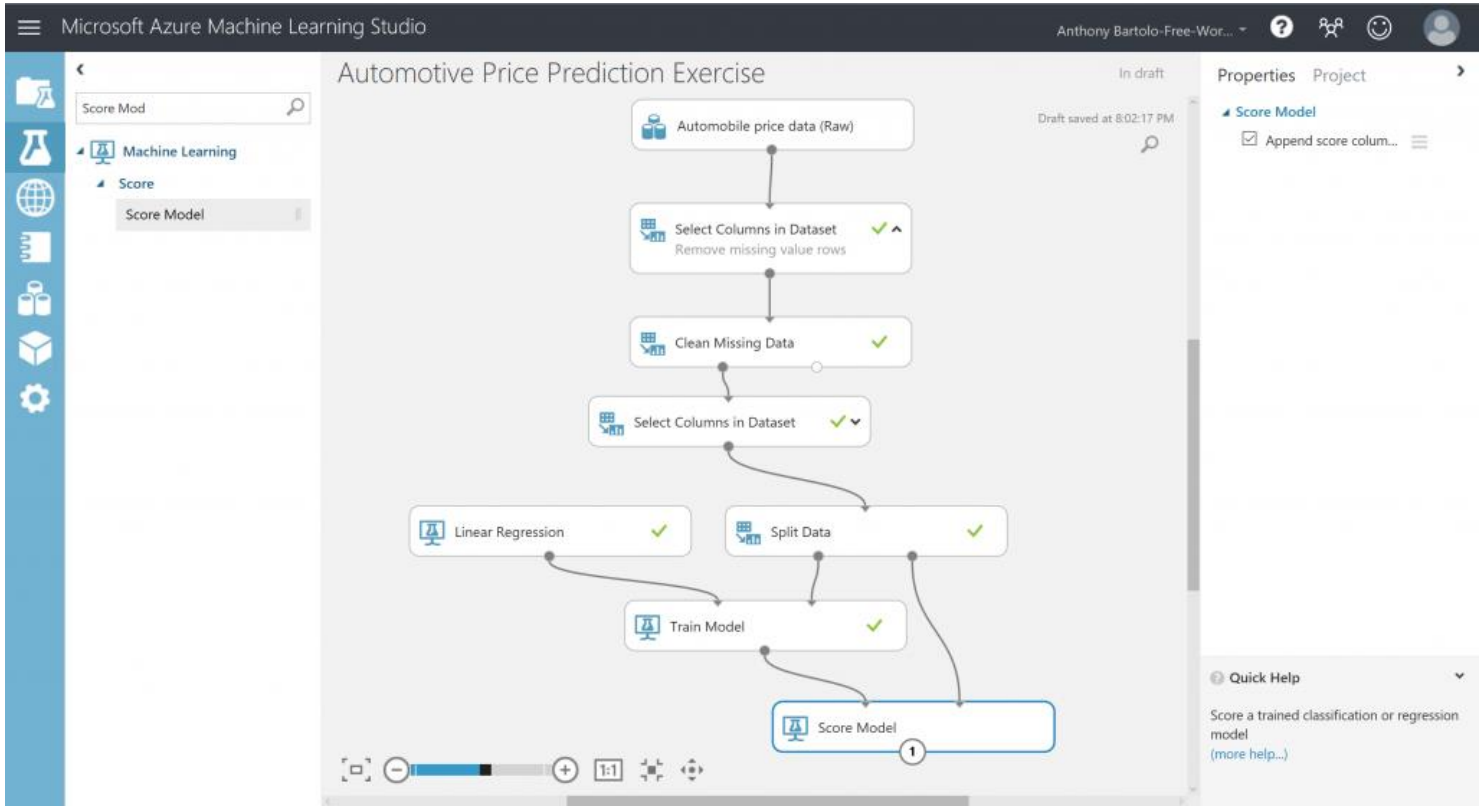
Do-it-yourself



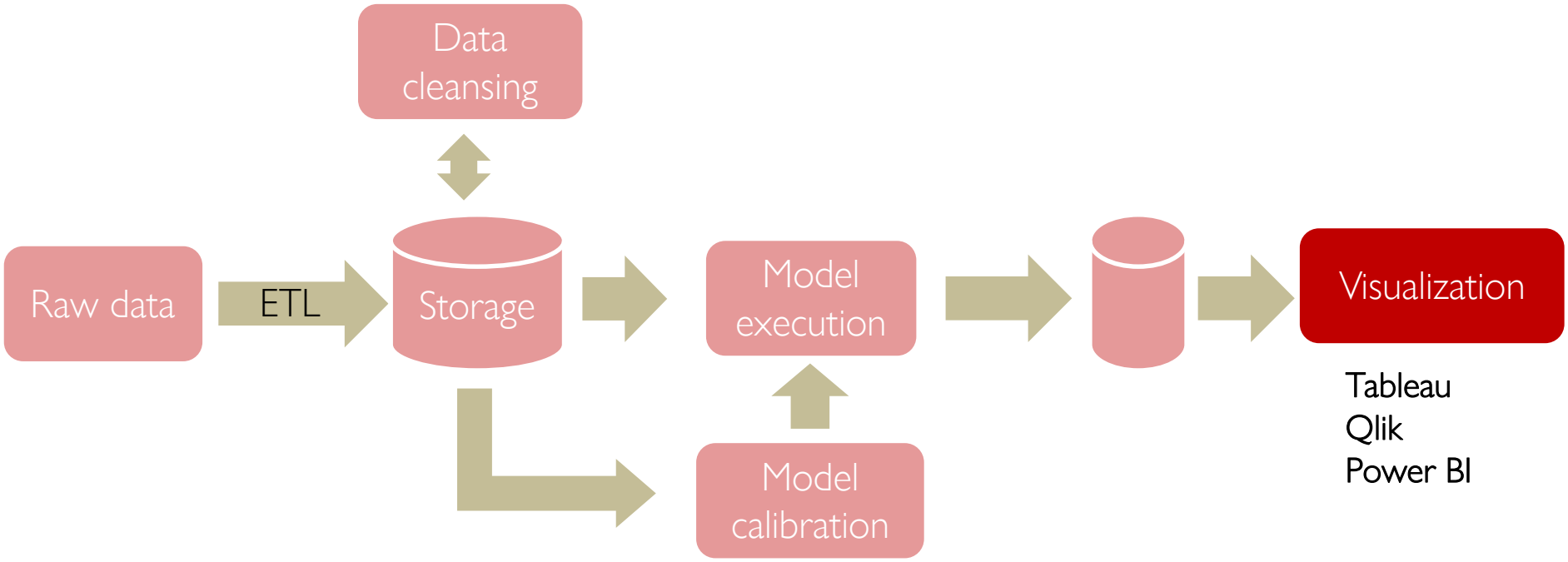
Do-it-yourself



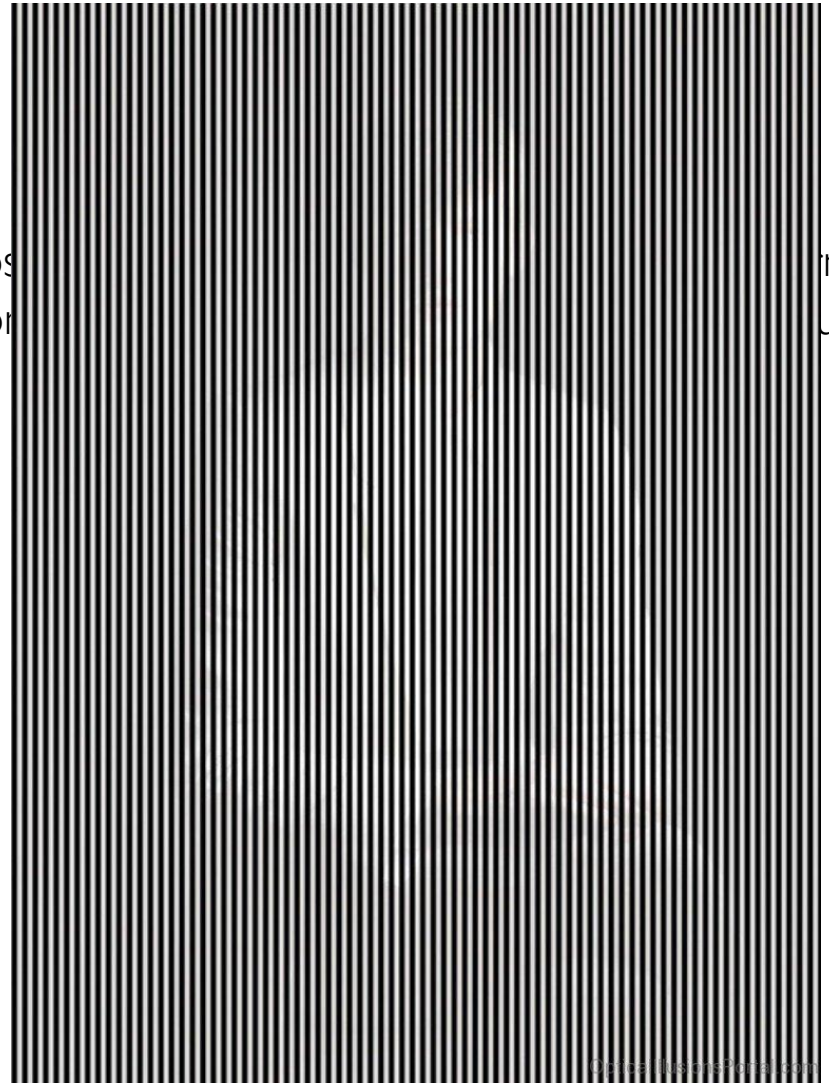
We-do-it-for-you



III. Visualización



Los humanos somos mación visual. La correcta visualizació ución del problema.



Gartner Magic Quadrant para software de Analítica

	Tableau	Qlik	Power BI (Microsoft)
location	Washington	Pennsylvania (Sweden)	Washington
employees	4,000 (2019)	2,500 (2015)	
revenue	1,2b$ (2018)	700M$(2017)	

Category	Criteria	Qlik	PowerBI	Tableau
Overall user experience	Ease to learn and help	4,0	4,0	4,5
	User base	4,0	4,5	4,5
	Ease to use	3,0	4,0	4,0
	Technical Support	2,5	2,0	2,0
Statistics	Data preparation	4,5	3,5	4,5
	Statistical analysis supported	3,0	2,5	3,5
Content	Statistical computations integration	2,5	2,0	3,5
	Look and feel	4,0	3,5	4,5
Share	Ease to publish	2,5	3,0	3,5
	Platform compatibility	4,0	3,5	3,0
Cost	License cost	3,0	4,0	2,0

Resumen

1. Las tecnologías de análisis avanzado se pueden dividir en tres grupos según la necesidad que aborden: 1) **data storage**, 2) **data modelling** y 3) **data visualization** .

2. El **data storage** incluye bases de datos **SQL** y **NoSQL**. Aunque Big Data está relacionado con las bases de datos NoSQL, las bases de datos SQL antiguas están aquí para quedarse. Ambos coexistirán en el futuro cubriendo diferentes requisitos en tamaño, disponibilidad y consistencia de datos.

3. La mayoría de los proyectos de Analítica Avanzada pueden gestionarse con bases de datos SQL para sus necesidades de almacenamiento de datos. La Analítica Avanzada es un proceso de aprendizaje. Uno empieza con lo que está acostumbrado y cuando llega el momento y la necesidad se está preparado para el cambio.

4. El software libre de creación de modelos es bueno, incluso mejor que el de pago. **R** y **Python** son el estándar para la construcción de modelos.

5. Las herramientas de análisis de datos **do-it-yourself** funcionan para casos sencillos. Han de usarse con precaución.

6. R y Python y la mayoría del software de construcción de modelos tienen buenas bibliotecas de visualización. Sin embargo, no están hechos pensado en la visualización de los datos a usuarios finales.

7. La visualización avanzada para usuarios finales debe dejarse en manos del software de visualización. Los líderes son **QlikView**, **Tableau** y **PowerBI**.

8. QlikView, Tableau y PowerBI **no son software de modelaje**. Aunque implementan modelos básicos su objetivo es la visualización.

4. Proyectos de Analítica Avanzada

"Analytics must be able to do three things: 1) solve a problem, 2) be predictive and 3) be implementable"

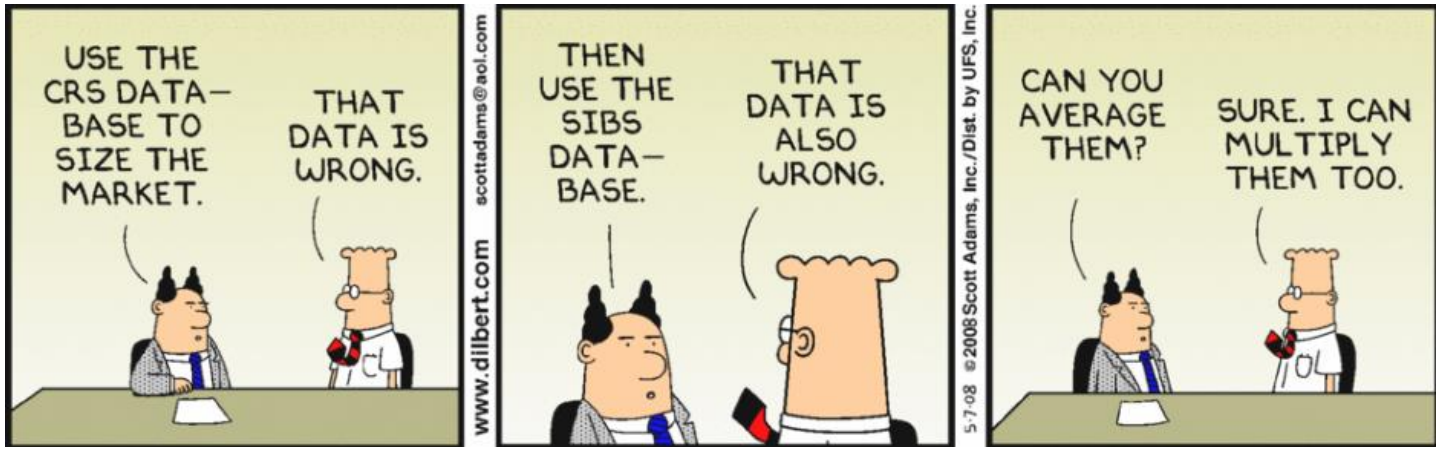
"Good data can yield to good decisions if captured, analyzed, communicated and acted upon on a timely and efficient fashion"

https://www2.deloitte.com

1. Objetivo

Tener clara la **necesidad de negocio** a cubrir con la solución y estimar su impacto.

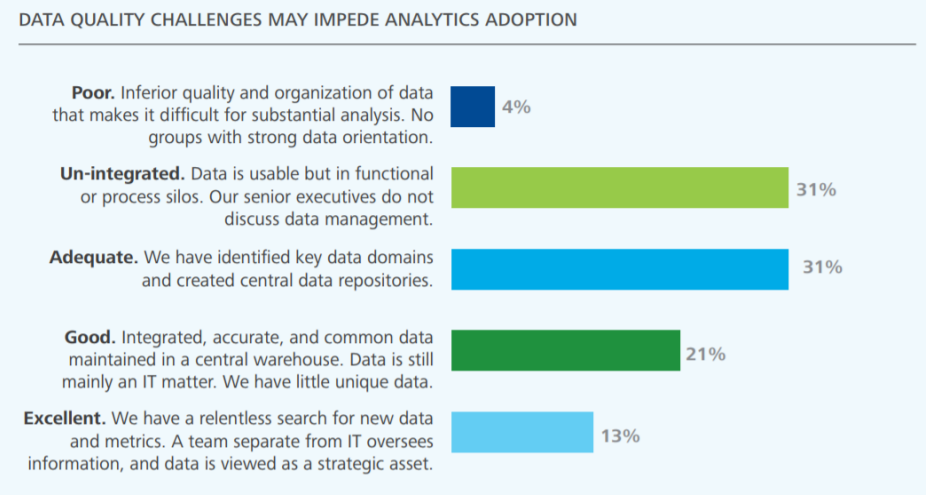
E.g. indicar como la empresa utilizará la solución una vez desarrollada.



2. Datos

Asegurar la **calidad** de los datos y su **relevancia**.

"Data management is becoming a bigger and bigger part of the puzzle, and a bigger and bigger challenge for us to overcome"
Vicepresident of marketing. Software Company



https://www2.deloitte.com

3. Recursos humanos

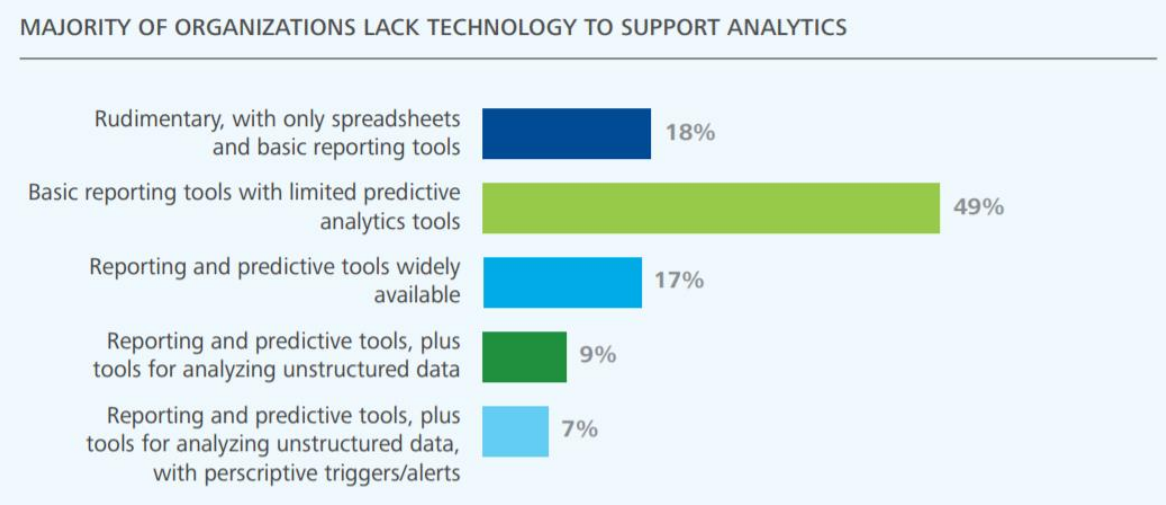
Muchas compañías no tienen los individuos o las habilidades necesarias para implementar Analítica Avanzada en la organización.

- Equipos internos vs. externos.
- Gap de skills.
- Calidad de los recursos.

4. Tecnología:

La infraestructura tecnología es “rudimentaria” o “básica” para el Desarrollo de Analítica Avanzada.

Tecnologías de visualización sin herramientas de modelado predictivo.



https://www2.deloitte.com

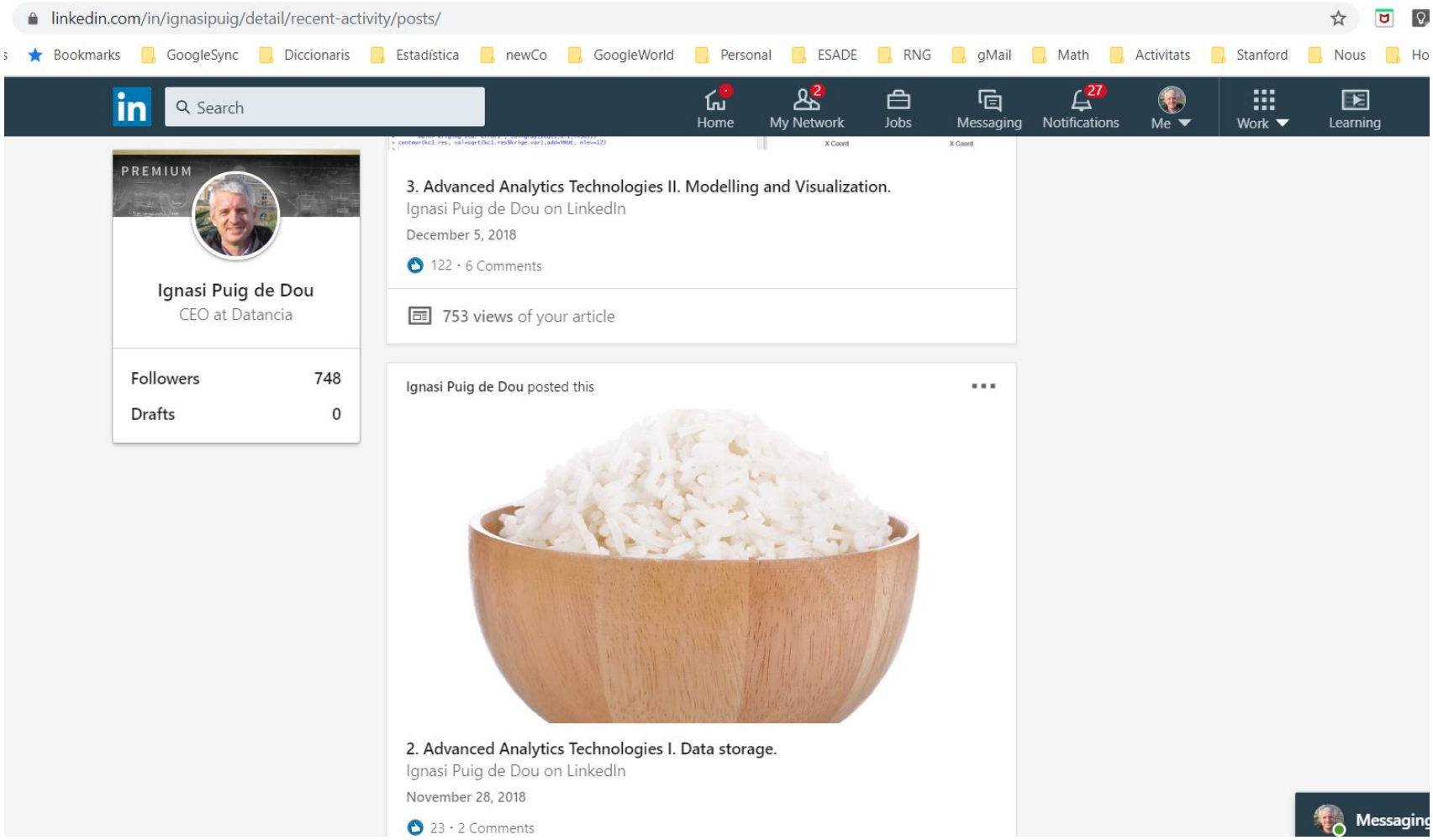
Requisitos para el éxito de un proyecto de Analítica Avanzada

I. Los resultados de los proyectos de Analítica Avanzada deben **incorporarse en los procesos de negocio** para ser útiles.

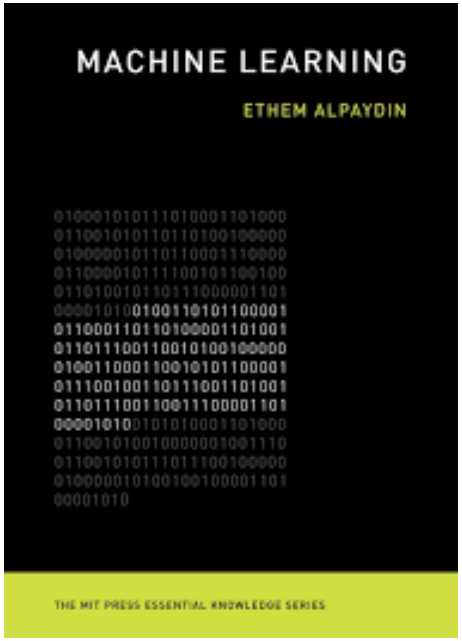
II. Aceptar que no siempre el resultado deseado está asegurado.

III. Pensar a **largo plazo** pero empezar **rápido** con proyectos a corto (**pilotos**) con el **soporte** del management de la empresa.

Further reading



Further reading



Este documento ha sido descargado de la web del Club Excelencia en Gestión: www.clubexcelencia.org