



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



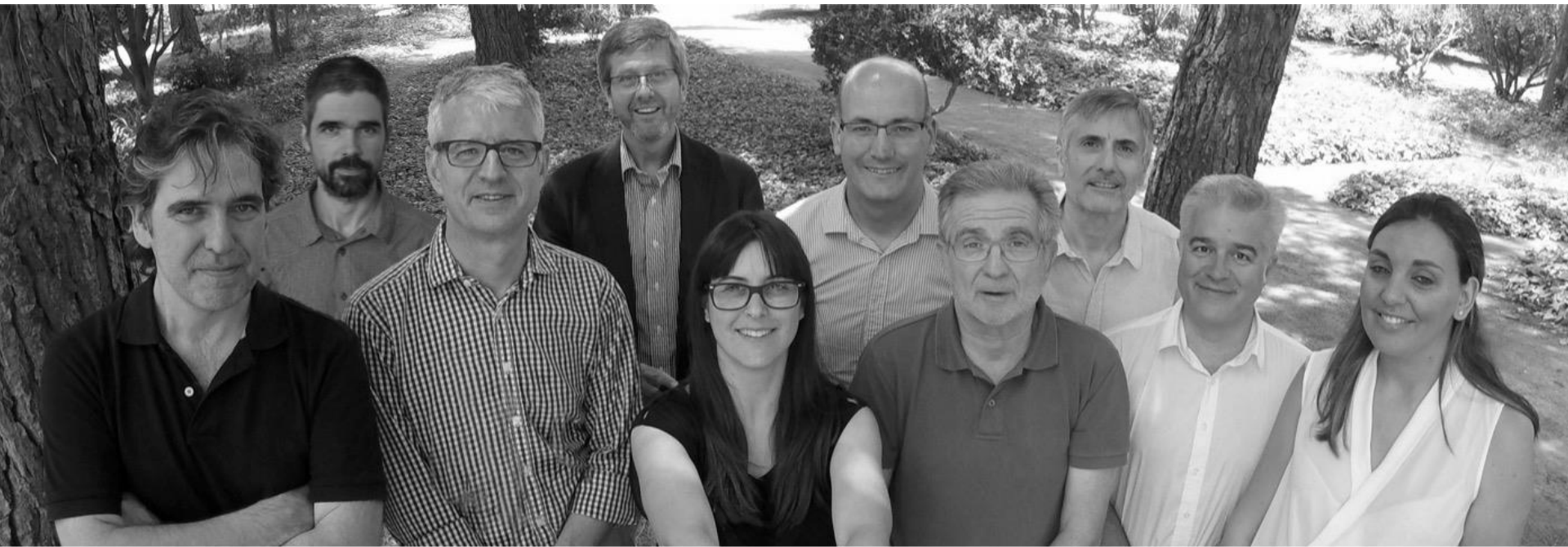
CUATRECASAS
ACELERA

Empresa seleccionada en la
4ª Edición de Cuatrecasas
Acelera



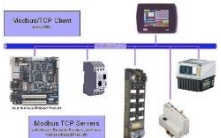
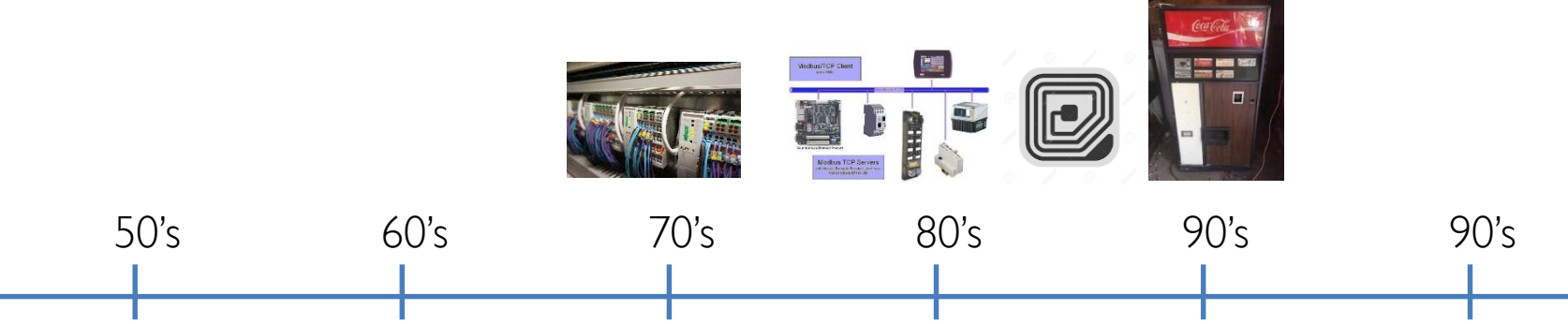
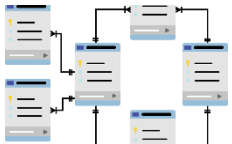
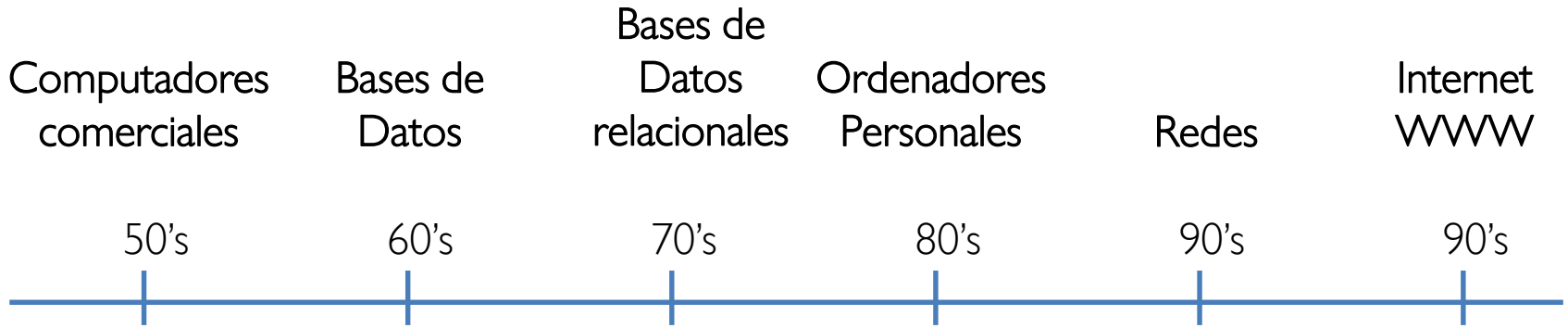
Somos una compañía de reciente creación, formada por profesionales de diferentes orígenes (Multinacionales, Universidad, Investigación y Consultoría).

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?



PLC

Modbus
RFID

Carnegie Mellon
Coca cola



I. Sensores

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



- 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



II. Capacidad de cómputo

amazon

1994



2003



2006



cassandra

2008

1995

Google

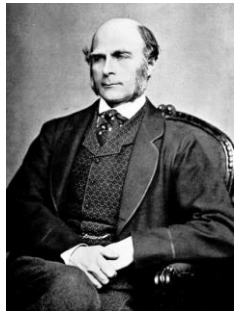
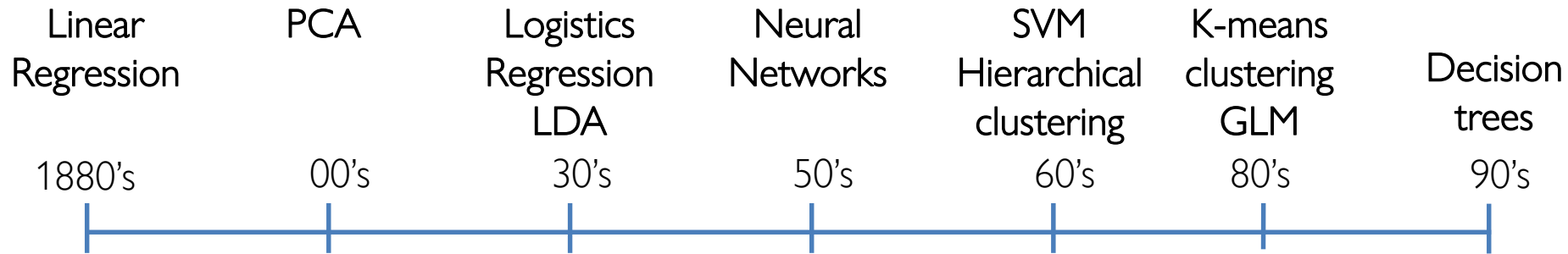
2003

Google File System



2007





Galton



Pearson



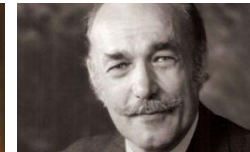
Fischer



Rossenblatt



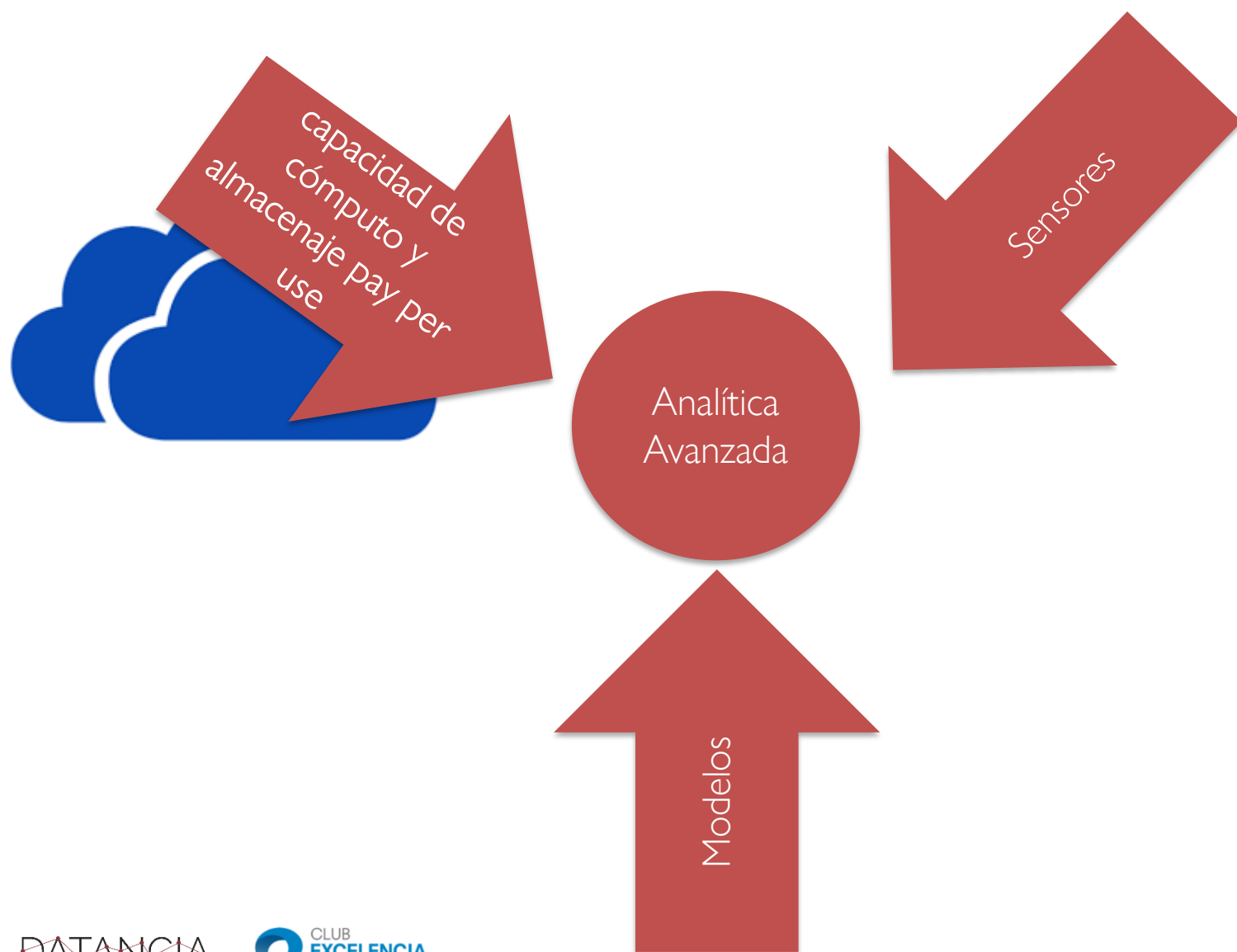
Vapnik



Nelder



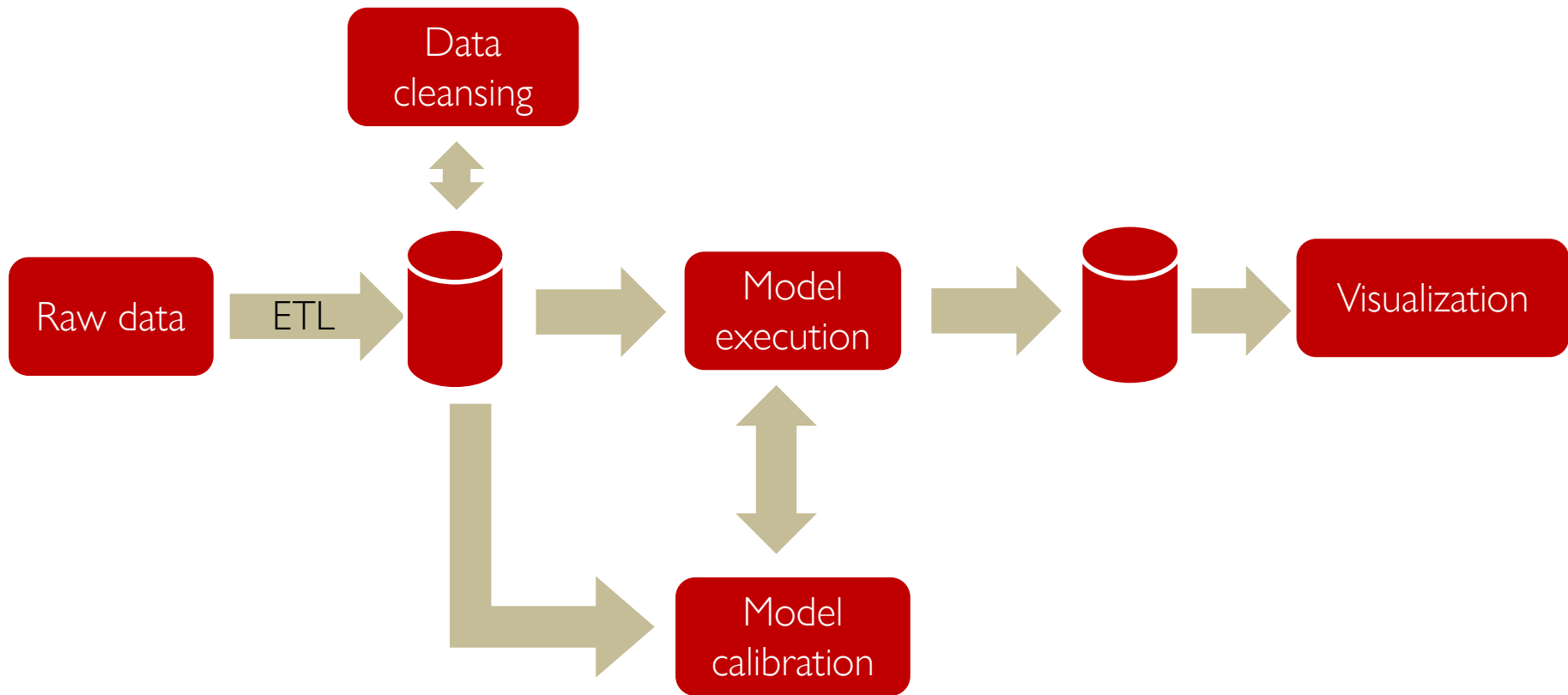
Friedman

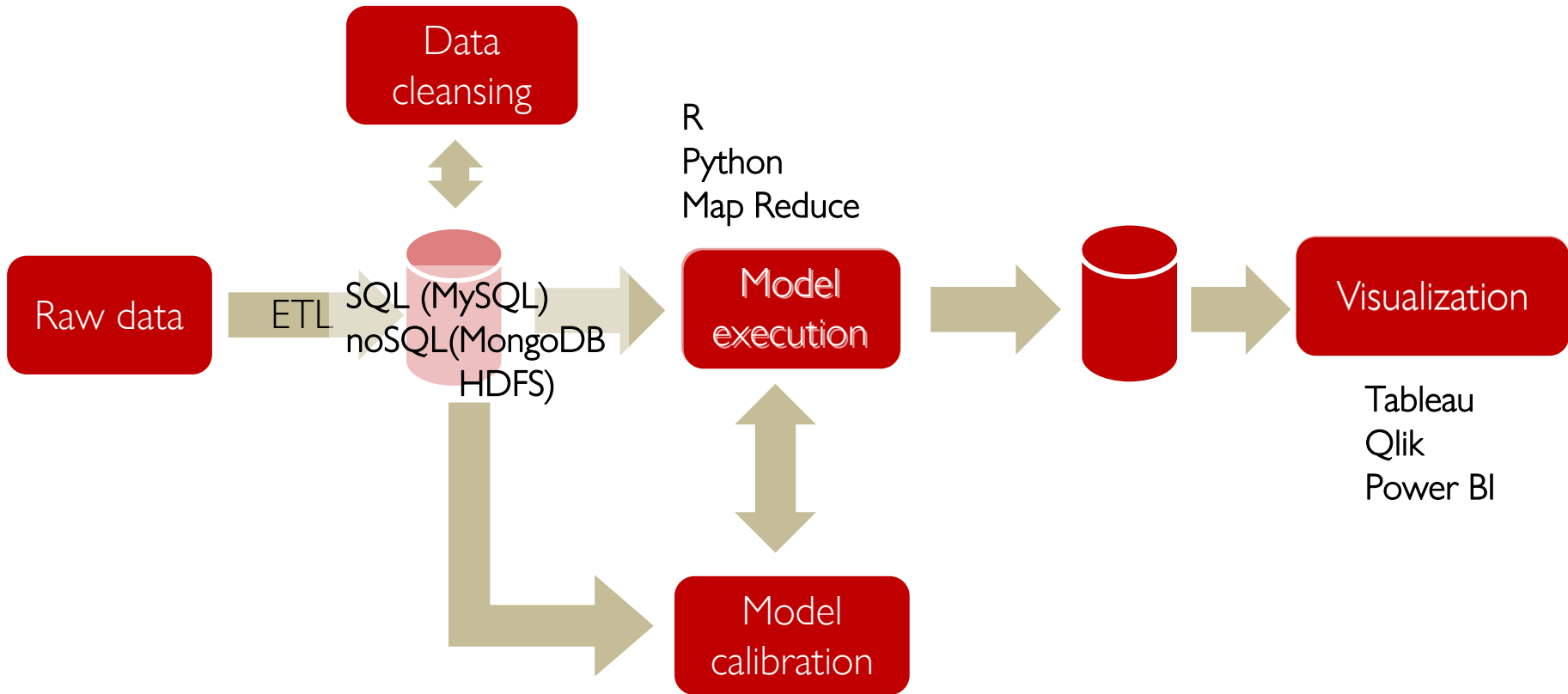


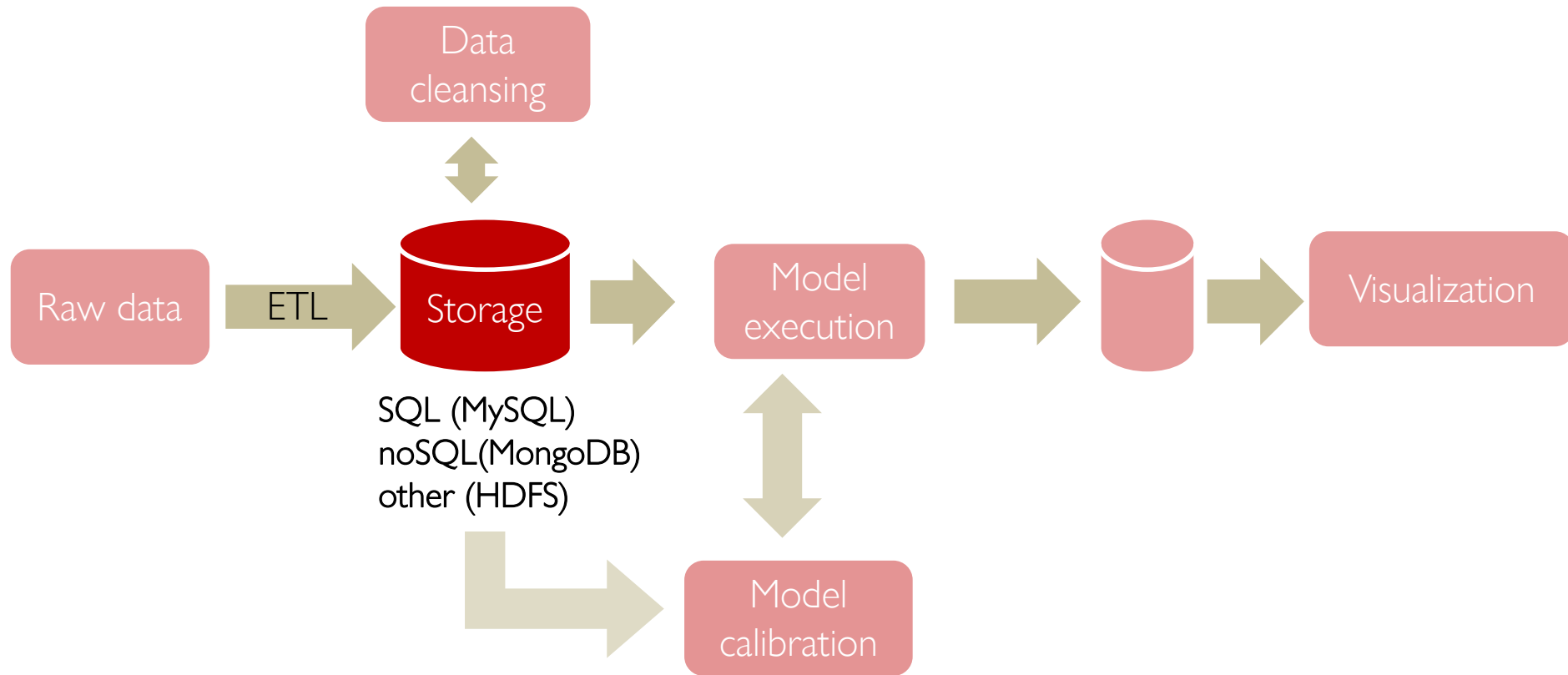


3. Tecnología











Bases de datos SQL o Relacionales

Foreign Keys

students:

id	name
1	Anna Malli
2	Anders Andersen
3	Pierre Untel
4	Erika Mustermann
5	Suan Pérez
6	Fulano de Tal
⋮	⋮

grades:

student	course	grade
4	MATH201	A-
1	CS413	A
3	CS100	B+
6	B10301	B
1	PHY222	A
2	ARTH213	B
⋮	⋮	⋮

Courses:

id	name
CS100	Intro Comp Sci
MATH201	Calculus
ARTH213	Surrealism
CS413	Purely Functional..
B10301	Anatomy
PHY222	Electromagnetism
⋮	⋮



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.

Foreign Keys

id	name
1	Anna Malli
2	Anders Andersen
3	Pierre Untel
4	Erika Mustermann
5	Juan Pérez
6	Fulano de Tal
⋮	⋮

student	course	grade
4	MATH201	A-
1	CS413	A
3	CS100	B+
6	B10301	B
1	PHY222	A
2	ARTH213	B
⋮	⋮	⋮

id	name
CS100	Intro Comp Sci
MATH201	Calculus
ARTH213	Surrealism
CS413	Purely Functional..
B10301	Anatomy
PHY222	Electromagnetism
⋮	⋮



Las bases de datos relacionales son **ACID**



I. Almacenaje

Atomicity

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

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

700

Isolation

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

200

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

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

Durability

date	USD
11:00	1,500

date	USD
11:00	1,700

date	USD
11:00	1,500

date	USD
11:00	1,700



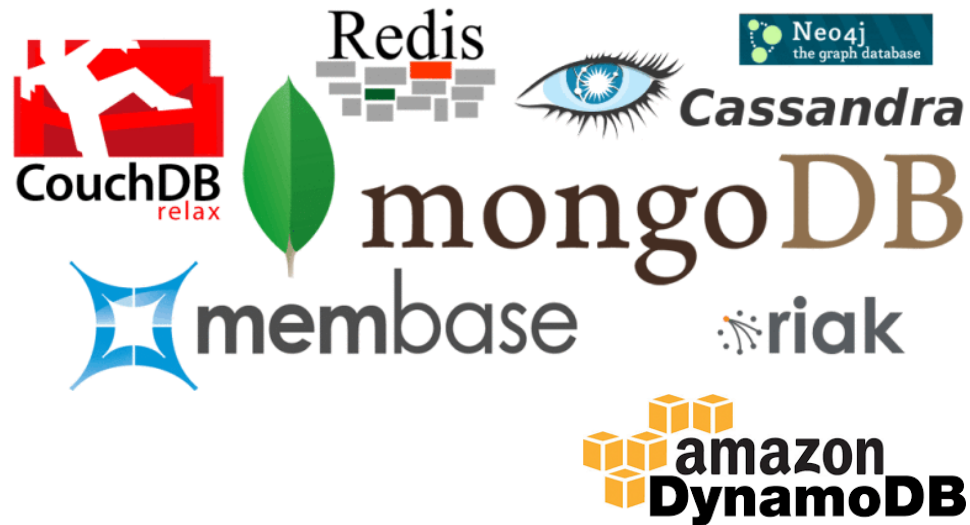
Bases de datos SQL o Relacionales

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 inmutable

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 almacenados.
- 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.

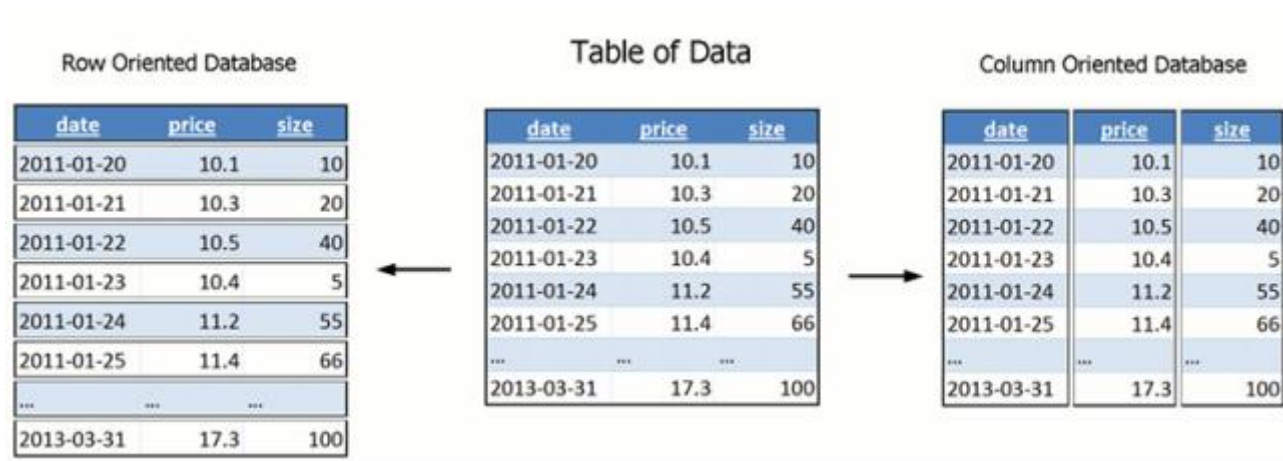
DNI: 00A	name: Smith age: 57 address: <table border="1"><tr><td>add1</td><td>BCN</td></tr><tr><td>add2</td><td>MAD</td></tr></table>	add1	BCN	add2	MAD
add1	BCN				
add2	MAD				
DNI: 00B	name: Codd age: 57 languages: <table border="1"><tr><td>English</td><td>certification: first: A advnaced: B proficiency: F</td></tr><tr><td>Spanish</td><td>level: Native</td></tr></table>	English	certification: first: A advnaced: B proficiency: F	Spanish	level: Native
English	certification: first: A advnaced: B proficiency: F				
Spanish	level: Native				

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€.

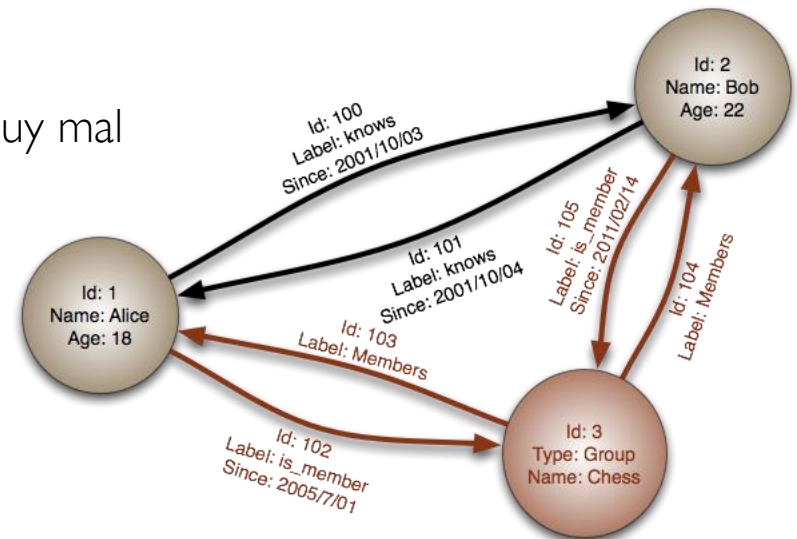


Bases de datos no-SQL

Bases de datos orientadas a grafos (e.g. [Neo4j](#))

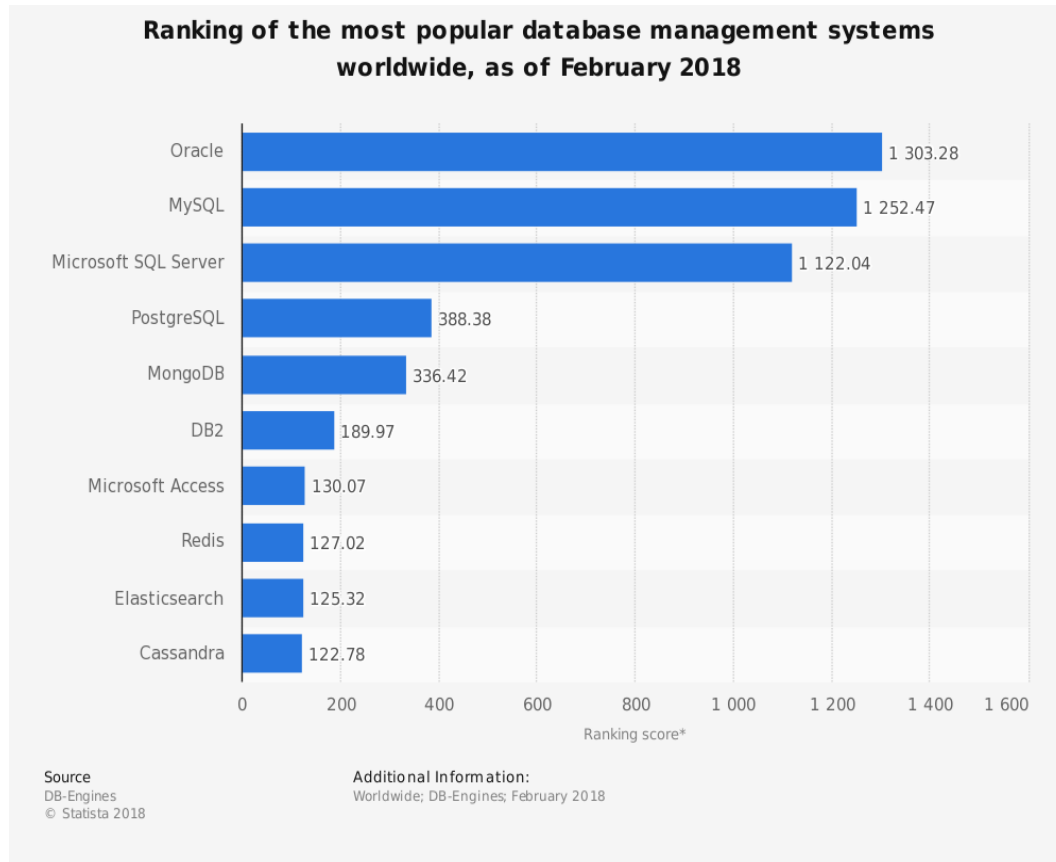
Nodos (e.g. personas) unidas por arcos (e.g. relaciones)

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



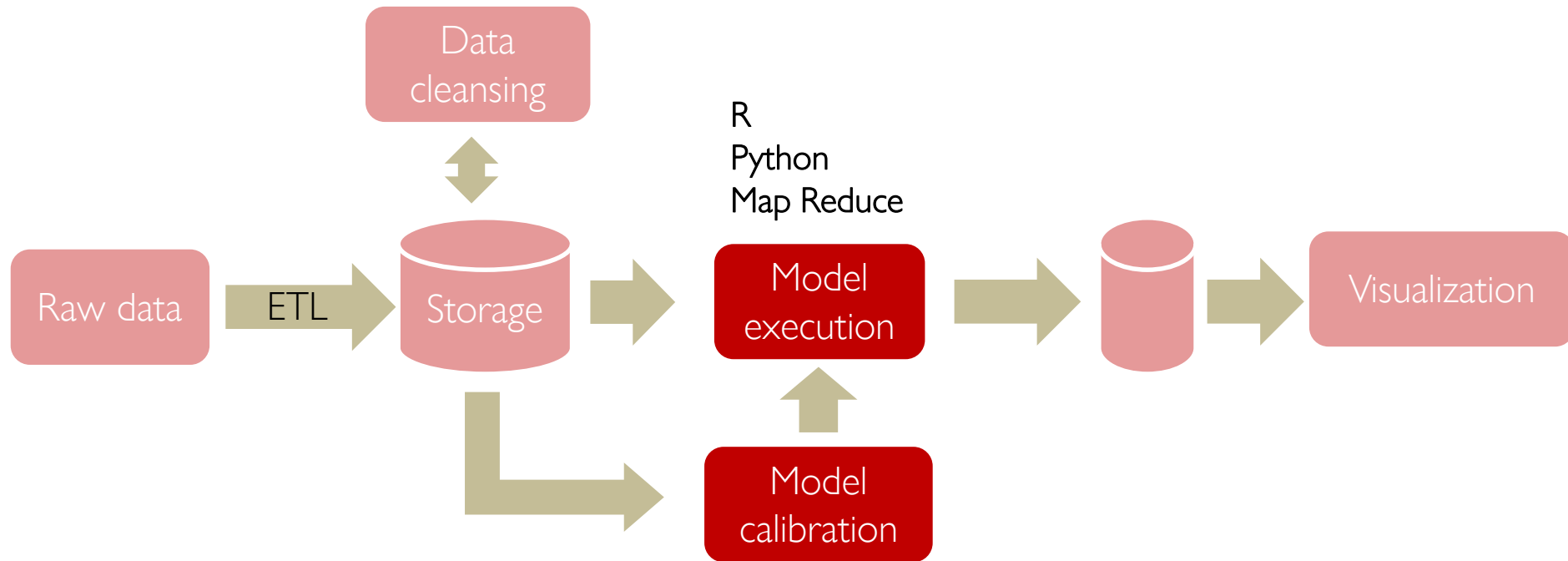


Uso de bases de datos





II. Desarrollo de modelos



Model execution software

Do-it-yourself



We-do-it-for-you



DataRobot




Azure Machine Learning



Do-it-yourself

https://www.r-project.org/about.html

rks GoogleSync Diccionario Estadística newCo GoogleWorld Personal ESADE RNG gMail Math Activitats Stanford



[Home]

Download
[CRAN](#)

R Project
[About R](#)
[Logo](#)
[Contributors](#)
[What's New?](#)
[Reporting Bugs](#)
[Development Site](#)
[Conferences](#)
[Search](#)

R Foundation
[Foundation](#)

What is R?

Introduction to R

R is a language and environment for statistical computing and graphics. It is a [GNU project](#) which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R is available as Free Software under the terms of the [Free Software Foundation's GNU General Public License](#) in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.



Do-it-yourself

The screenshot displays the RStudio interface. The source editor on the left contains the following R code:

```
1 library(rgl)
2 library(plot3D)
3 x <- seq(-5, 5, by = 0.2)
4 y <- seq(-5, 5, by = 0.3)
5 grid <- mesh(x, y)
6 f <- function(x,y) sin(x)*sin(y)
7 z <- with(grid, f(x,y))
8 persp3D(z = z, x = x, y = y,contour = TRUE, image = T,theta=60)
9 persp3D(z = z, x = x, y = y,color='orange')
10
```

The console at the bottom shows the execution of these commands:

```
> library(rgl)
> library(plot3D)
> x <- seq(-5, 5, by = 0.2)
> y <- seq(-5, 5, by = 0.3)
> grid <- mesh(x, y)
> f <- function(x,y) sin(x)*sin(y)
> z <- with(grid, f(x,y))
> persp3D(z = z, x = x, y = y,contour = TRUE, image = T,theta=60)
> persp3D(z = z, x = x, y = y,color='orange')
>
```

The Environment pane on the right shows the following data and functions:

Data	
z	num [1:51, 1:34] 0.92 0.955 0.953 0.913...

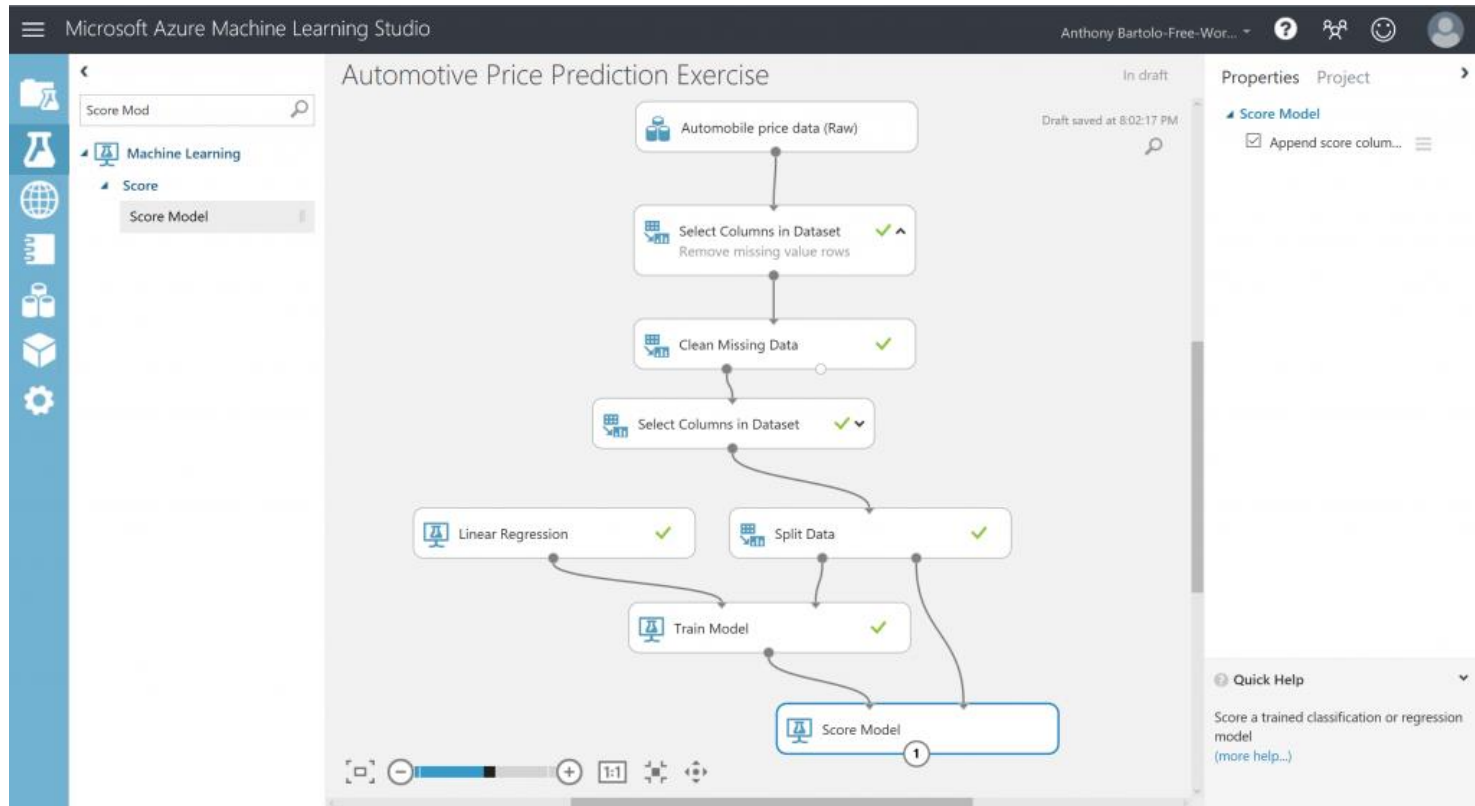
Values	
grid	List of 2
x	num [1:51] -5 -4.8 -4.6 -4.4 -4.2 -4 -3.8...
y	num [1:34] -5 -4.7 -4.4 -4.1 -3.8 -3.5 ...

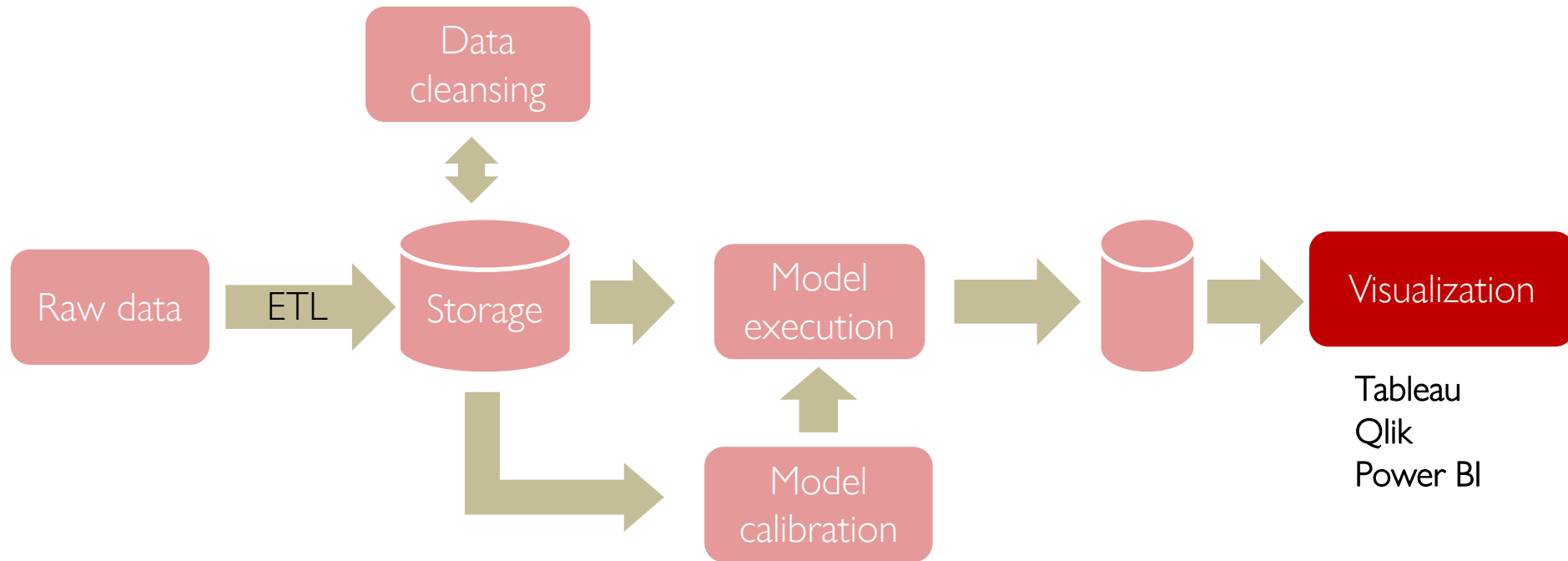
Functions	
f	function (x, y)

The Plots pane on the right shows a 3D surface plot of the function $z = \sin(x) \sin(y)$. The plot is rendered in orange and includes a contour overlay. A color scale on the right indicates values from -0.5 (blue) to 0.5 (red).



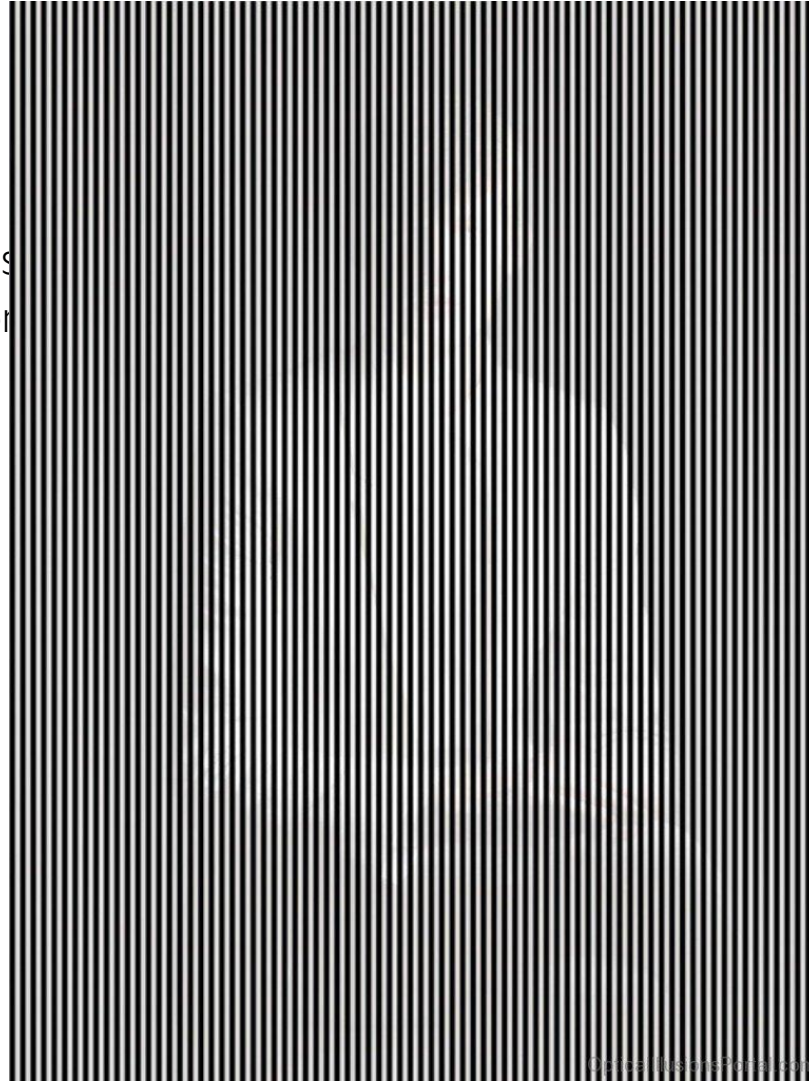
We-do-it-for-you







Los humanos somos
correcta visualización



Información visual. La
olución del problema.



Gartner Magic Quadrant para software de Analítica



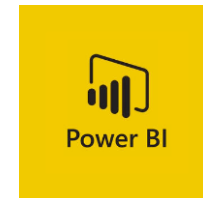


	Tableau	Qlik	Power BI (Microsoft)
location	Washington	Pennsylvania (Sweden)	Washington
employees	4,000 (2019)	2,500 (2015)	
revenue	1,2b\$ (2018)	700M\$(2017)	



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



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 Análítica Avanzada pueden gestionarse con bases de datos SQL para sus necesidades de almacenamiento de datos. La Análítica 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 pensando 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



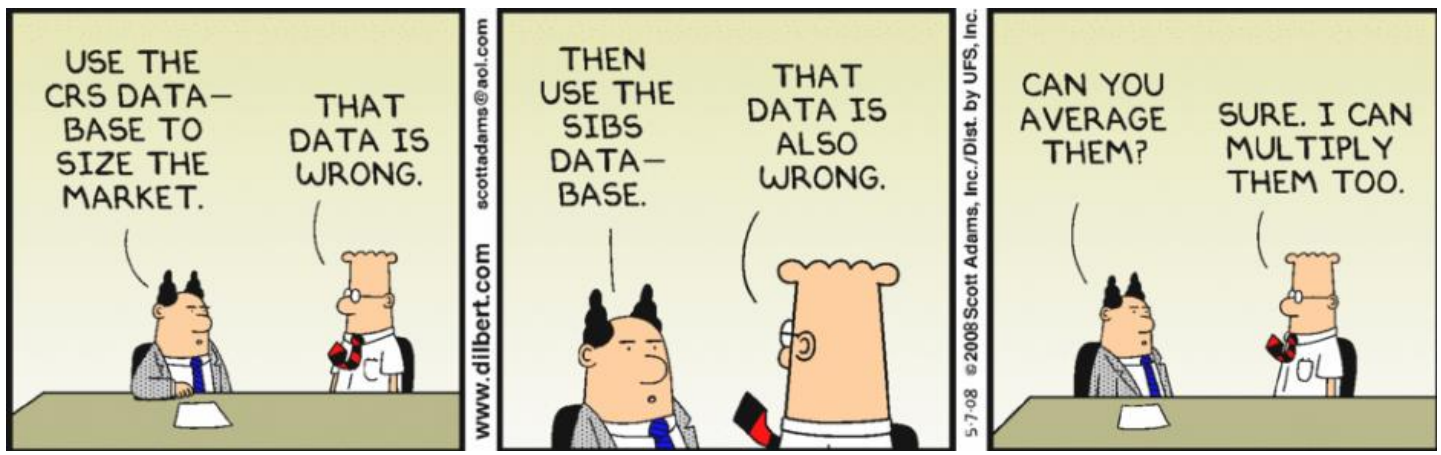
“Analytics must be able to do three things: 1) solve a problem, 2) be predictive and 3) be implementable”

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

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.



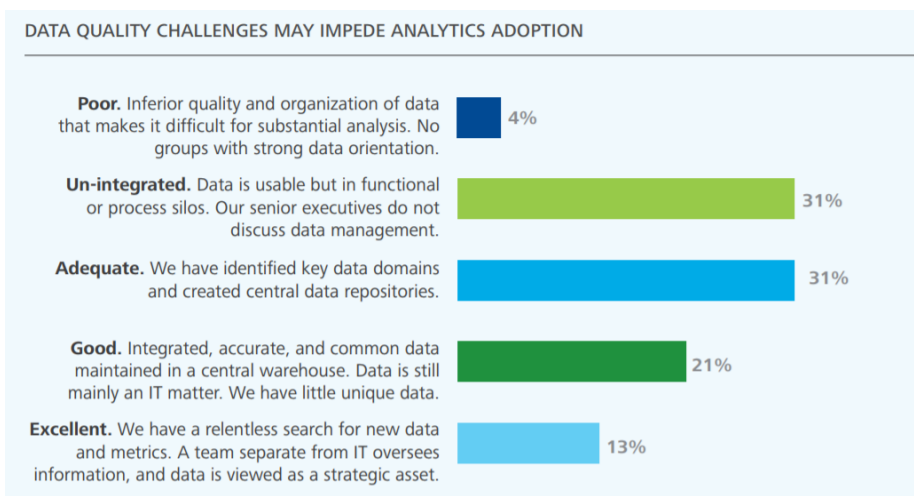


2. Datos

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



3. Recursos humanos

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

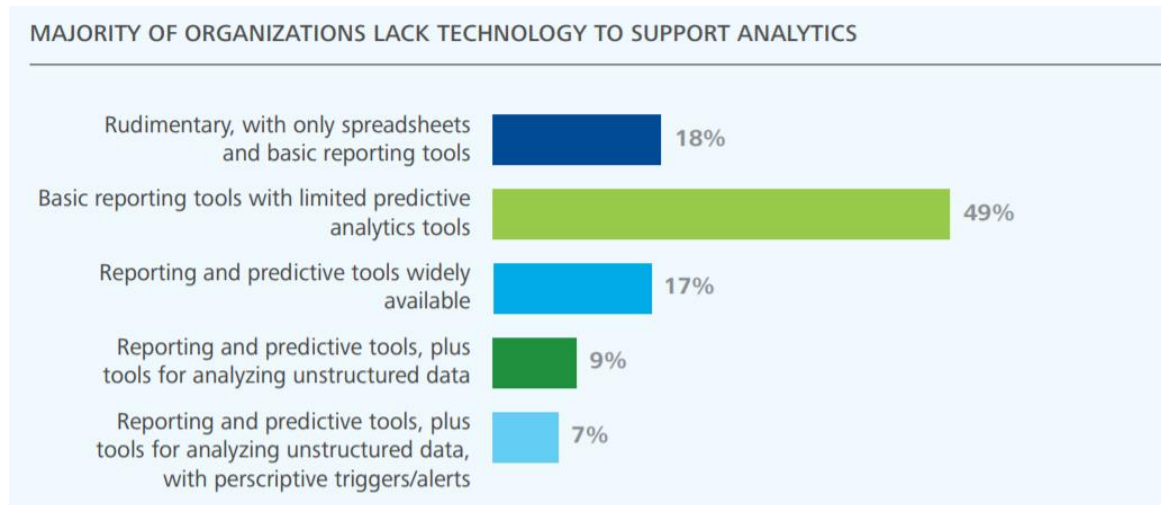
- Equipos internos vs. externos.
- Gap de skills.
- Calidad de los recursos.



4. Tecnología:

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

Tecnologías de visualización sin herramientas de modelado predictivo.



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

linkedin.com/in/ignasiPuig/detail/recent-activity/posts/

Bookmarks GoogleSync Diccionario Estadística newCo GoogleWorld Personal ESADE RNG gMail Math Activitats Stanford Nous Ho

in Search Home My Network Jobs Messaging Notifications Me Work Learning

PREMIUM




Ignasi Puig de Dou
CEO at Datancia

Followers 748
Drafts 0

3. Advanced Analytics Technologies II. Modelling and Visualization.
Ignasi Puig de Dou on LinkedIn
December 5, 2018
122 · 6 Comments

753 views of your article

Ignasi Puig de Dou posted this



2. Advanced Analytics Technologies I. Data storage.
Ignasi Puig de Dou on LinkedIn
November 28, 2018
23 · 2 Comments

Messaging

Further reading



