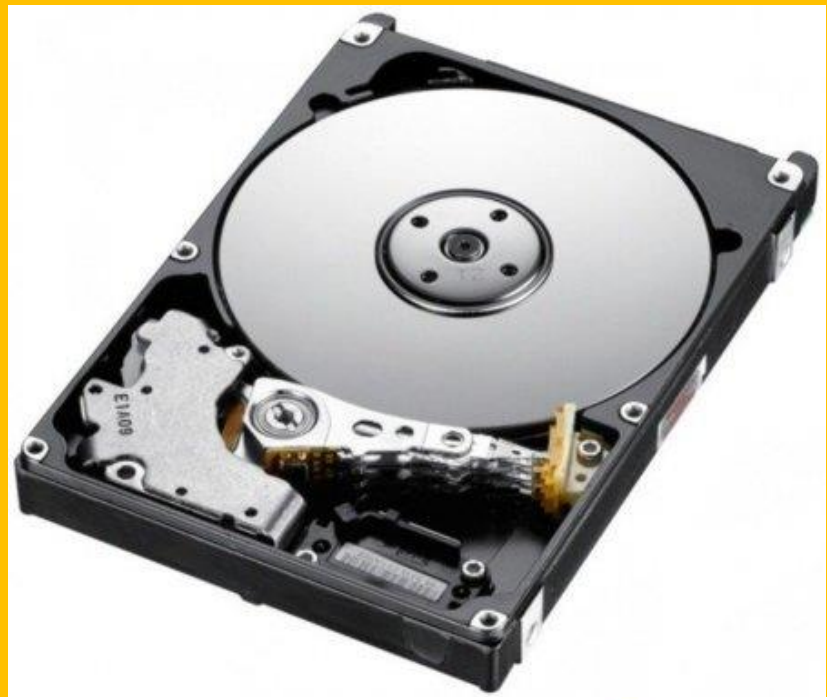
Big Data & Analytics

Aprendizaje Estadístico

Fernando Kornblit

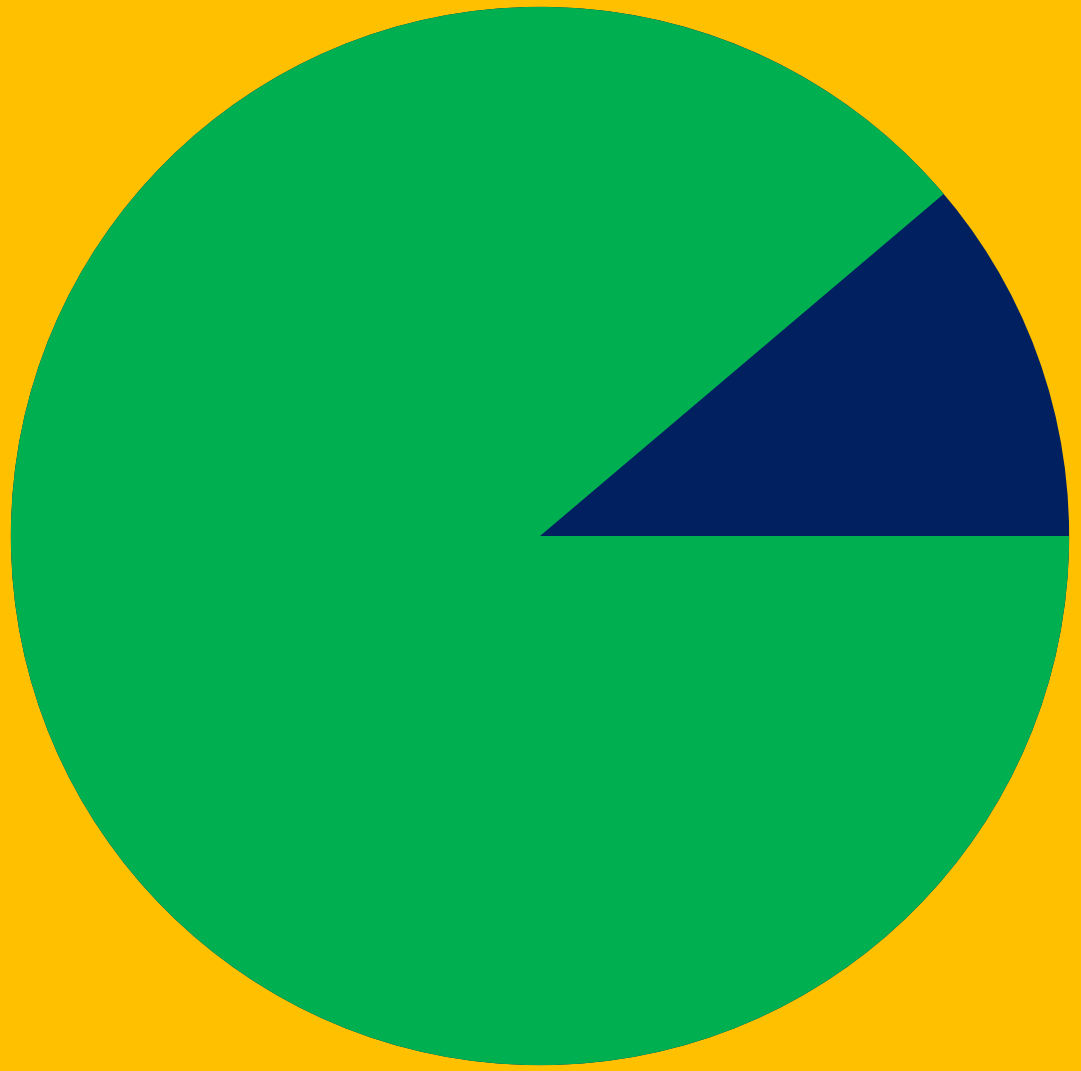
fkornblit@inti.gob.ar

Disco de Festos, Creta.
Edad de bronce, 2000 A.C.
(encontrado en 1908)



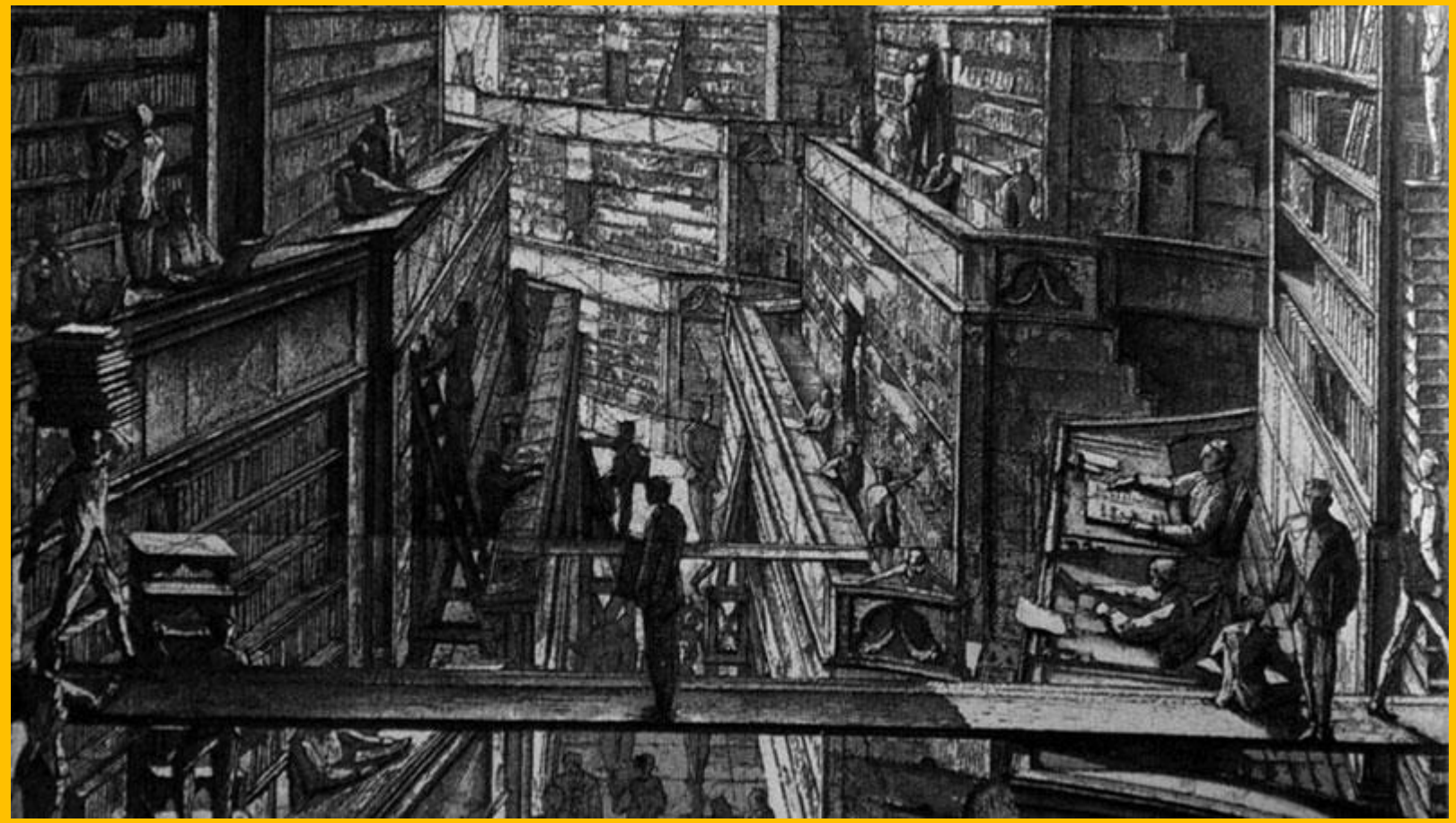
kB → MB → GB → TB →PB →EB → ZB →…

10^3 B 10^6 B 10^9 B 10^{12} B 10^{15} B 10^{18} B 10^{21} B



- Crecimiento exponencial
- En 2025 se generarán 163 ZB ($1{,}63 \times 10^{23}$ B)
- En 2003, se descifró por primera vez el genoma humano. Secuenciar tres mil millones de pares de bases de ADN requirió una década de trabajo. Diez años después, un solo laboratorio fue capaz de hacerlo en un día
- Si los autos hubieran avanzado igual que las TICs, podríamos dar la vuelta al mundo en 5h, y sin llenar el tanque


LA BIBLIOTECA
DE BABEL
DE
JORGE
LUIS BORGES



- Cambio no sólo cuantitativo, sino cualitativo
- No sólo tenemos más respuestas a las preguntas, sino más preguntas
- Los datos son tan buenos como las preguntas que les hacemos
- Nuevas demandas, nuevos requerimientos, nuevas aplicaciones (sensores inteligentes, IoT, …)
- Se requieren mejores algoritmos y tecnologías
- Tecnologías de almacenamiento, búsqueda, transmisión, procesamiento

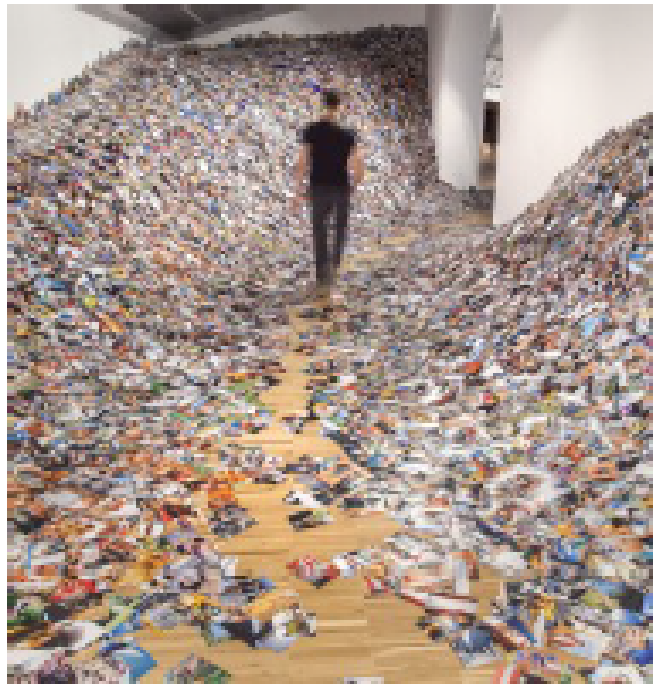
Datos “líquidos”

Downloading → Visiting

¿Datos = Información?

Oportunidades en el uso de datos masivos en la industria

- Detectar y aprovechar nuevos clientes y nuevos negocios
- Mejorar la calidad y competitividad
- Prevenir riesgos



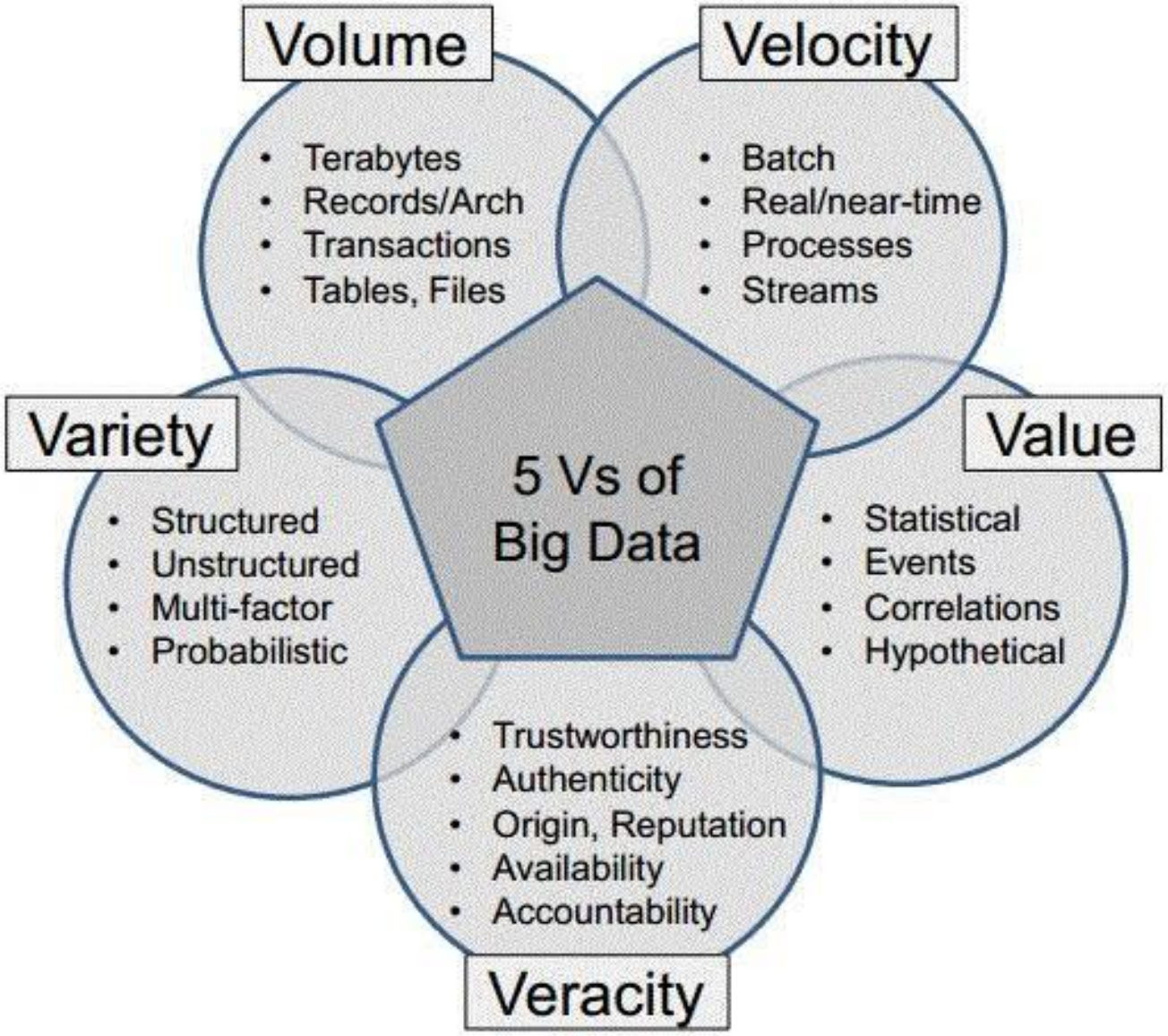
Debilidades para el uso inteligente y efectivo de datos

- Insuficiente nivel de desarrollo en la generación y utilización de datos
- Limitación de acceso a fuentes de datos
- Diversidad de formatos y estructuras
- Falta de normas o criterios de evaluación de su calidad
- Falta de regulación en el uso de datos

Desafíos

- Balancear el incremento de valor generado por la interconexión de sistemas y bases de datos, con la necesidad de proteger la privacidad y propiedad intelectual de sus generadores

- Políticas sobre ética de recolección y uso de datos a gran escala


Volume
Terabytes
Records/Arch
Transactions
Tables, Files
Velocity
Batch
Real/near-time
Processes
Streams
Variety
Structured
Unstructured
Multi-factor
Probabilistic
5 Vs of
Big Data
Value
Statistical
Events
Correlations
Hypothetical
Trustworthiness
Authenticity
Origin, Reputation
Availability
Accountability
Veracity

Metadatos: Datos que describen cómo son los datos

En el contexto 4.0, los datos son:

- Adquiridos automáticamente
- Transmitidos para ser procesados en otro lugar
- Integrados de múltiples orígenes, con calidades diferentes
- Operados y reciclados por otros usuarios para ganar información y tomar decisiones.

Calidad de datos → Calidad de información → Decisiones confiables

El problema ya no es cómo obtener datos sino saber cuán confiables son

¿Cómo aseguramos la calidad de esos datos, y la confiabilidad de la información?

Big data: (datos masivos, inteligencia de datos, datos a gran escala) conjuntos de datos tan grandes y complejos que requieren, para su adecuado tratamiento, aplicaciones no tradicionales

Curado de datos:
Organización e integración de datos colectados de diversas fuentes. Presentación de los datos de forma tal que su valor sea mantenido en el tiempo, y que resulte disponible para el reuso y preservación

Integración:
Agregado de data sets a partir de fuentes heterogéneas, por medio de enlaces, combinación o fusión

Interoperabilidad:
Habilidad de un sistema de crear, intercambiar, y consumir datos con un significado claro y no ambiguo

Procedencia:
Descripción de la historia de un dataset, conteniendo su origen, propósito para el que fue creado, y el registro de todas las modificaciones posteriores

Datos accionables automáticamente *(Machine actionable data)*:
Datos y metadatos que permiten a una computadora procesar, interpretar, relacionar, inferir a partir de ellos, y tomar decisiones sin intervención humana. **The machine knows what I mean**

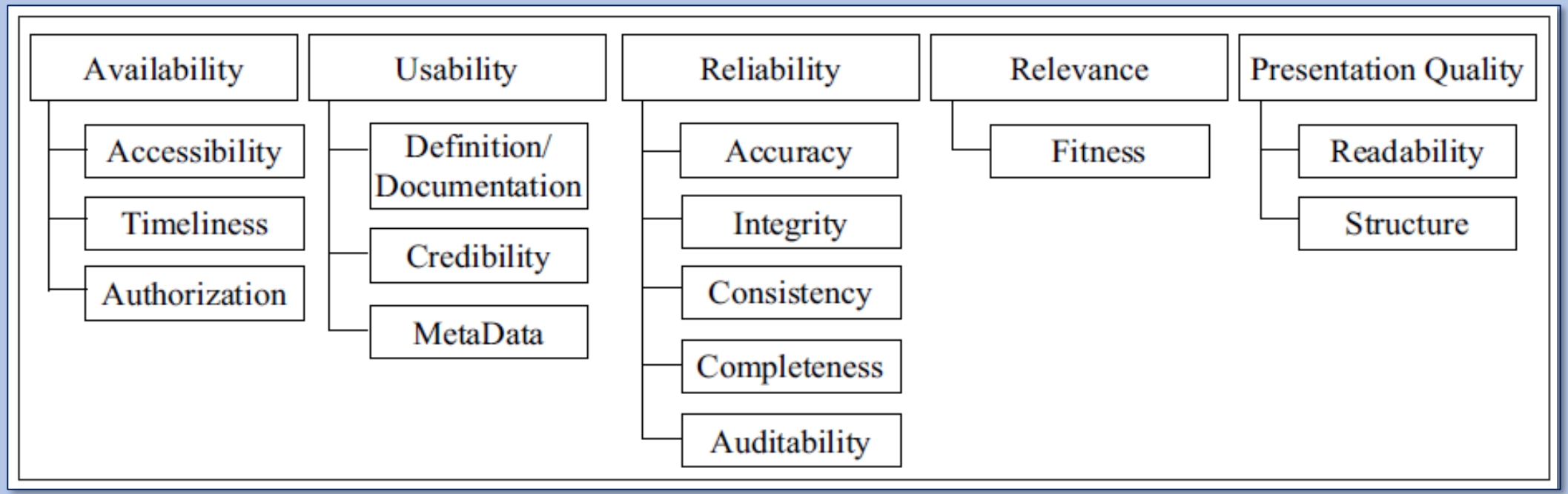
DATA SCIENCE JOURNAL

Cai, L and Zhu, Y 2015 The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, 14: 2, pp. 1-10, DOI: http://dx.doi.org/10.5334/dsj-2015-002

PROCEEDINGS PAPER

The Challenges of Data Quality and Data Quality Assessment in the Big Data Era

Li Cai[1,3] and Yangyong Zhu[2]



En 2016, se publicaron los '**FAIR Guiding Principles for scientific data management and stewardship**'

Findable
F1. (Meta)data are assigned a globally unique and persistent identifier
F2. Data are described with rich metadata (defined by R1 below)
F3. Metadata clearly and explicitly include the identifier of the data they describe
F4. (Meta)data are registered or indexed in a searchable resource

FAIR: Fully AI Ready

Accessible
A1. (Meta)data are retrievable by their identifier using a standardized communications protocol
A2. The protocol is open, free, and universally implementable by an authentication and authorization procedure
A3. Metadata are accessible, even when the data are no longer available

Interoperable
I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (Meta)data use vocabularies that follow FAIR principles
I3. (Meta)data include qualified references to other (meta)data

Reusable
R1. Meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (Meta)data are released with a clear and accessible data usage license
R1.2. (Meta)data are associated with detailed provenance

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON APPLICATIONS OF BIG DATA IN SOCIAL SCIENCES

Received February 10, 2019, accepted March 1, 2019, date of publication March 14, 2019, date of current version April 2, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2904286

Big Data Quality Assurance Through Data Traceability: A Case Study of the National Standard Reference Data Program of Korea

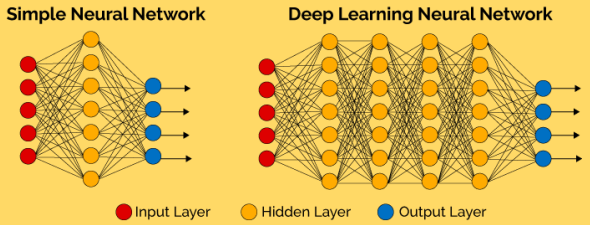
DOYOUNG LEE

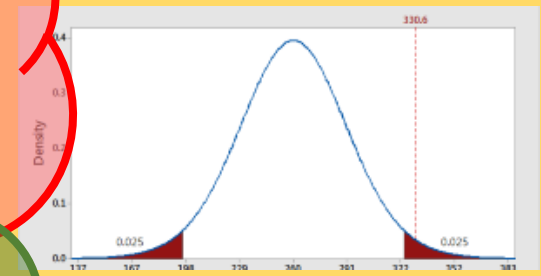
National Center for Standard Reference Data, Korea Research Institute of Standards and Science, Daejeon 34113, South Korea
Graduate School of Science and Technology Policy, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea

e-mail: dy.lee@kaist.ac.kr

This work was supported by the Korea Research Institute of Standards and Science.

¿Dónde vive la Ciencia de Datos?





Inteligencia Artificial

Estadística

Informática

Ciencia de Datos

Estadística Aplicada

Procesamiento de datos

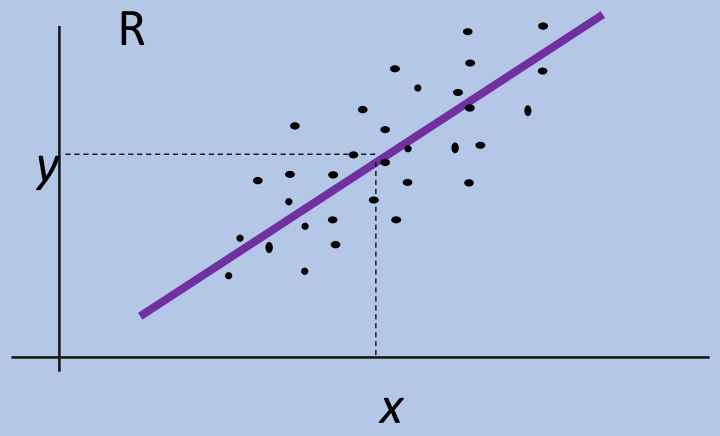
Experticia específica

De la Estadística a la Ciencia de Datos

- Más datos → se requieren algoritmos más rápidos y eficientes
- Más datos, pero quizás de menor calidad
- Relativización del concepto de muestreo representativo (¿n = todo?)
- Menos fundamentos probabilísticos y demostraciones matemáticas. Los mismos datos nos dicen si un algoritmo funciona
- Poder predictivo vs. Interpretabilidad

Covarianza y correlación

Supongamos por ejemplo que, en la fabricación de piezas de plásticos por extrusión, X es la presión aplicada, e Y la resistencia a la tracción de las piezas producidas. Se realizan n corridas de prueba, a diferentes presiones, y se ensaya la resistencia de las piezas producidas. Se representan los resultados obtenidos en el gráfico siguiente:



Coeficiente de correlación $$R = \frac{cov(X,Y)}{s_X \cdot s_y}$$

Coeficiente de determinación $$R^2$$

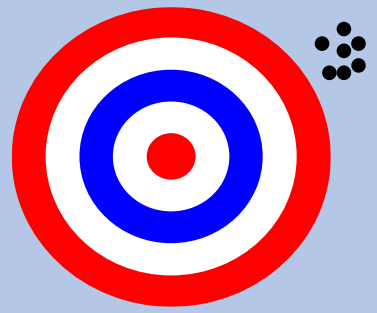
Si se obtiene un valor alto valor de R, vale la pena considerar un modelo lineal que relacione las variables.

El proceso de determinar la recta o que mejor ajusta a los datos medidos (x,y) se llama **regresión**

Una vez hallada la recta, es posible **predecir** la resistencia (Y) de una nueva pieza, en función de la presión (X) aplicada en la fabricación

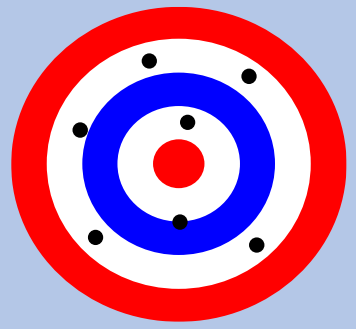
Propiedades de estimadores

Sesgo:



Alto sesgo,
baja varianza

Varianza:



Alta varianza,
bajo sesgo

Error cuadrático medio: suma del cuadrado de su sesgo mas su varianza

$$ECM(\hat{\theta}) = E[\hat{\theta} - \theta]^2 = b(\hat{\theta})^2 + V(\hat{\theta})$$

¿Cuándo un efecto es estadísticamente significativo?

- **Tests de hipótesis**: herramientas para poner una hipótesis a prueba

- **P-valor**: medida de la significancia de una hipótesis

Cuanto más cercano a 0 es el p-valor, más seguros estamos de la significatividad de un efecto o de una hipótesis

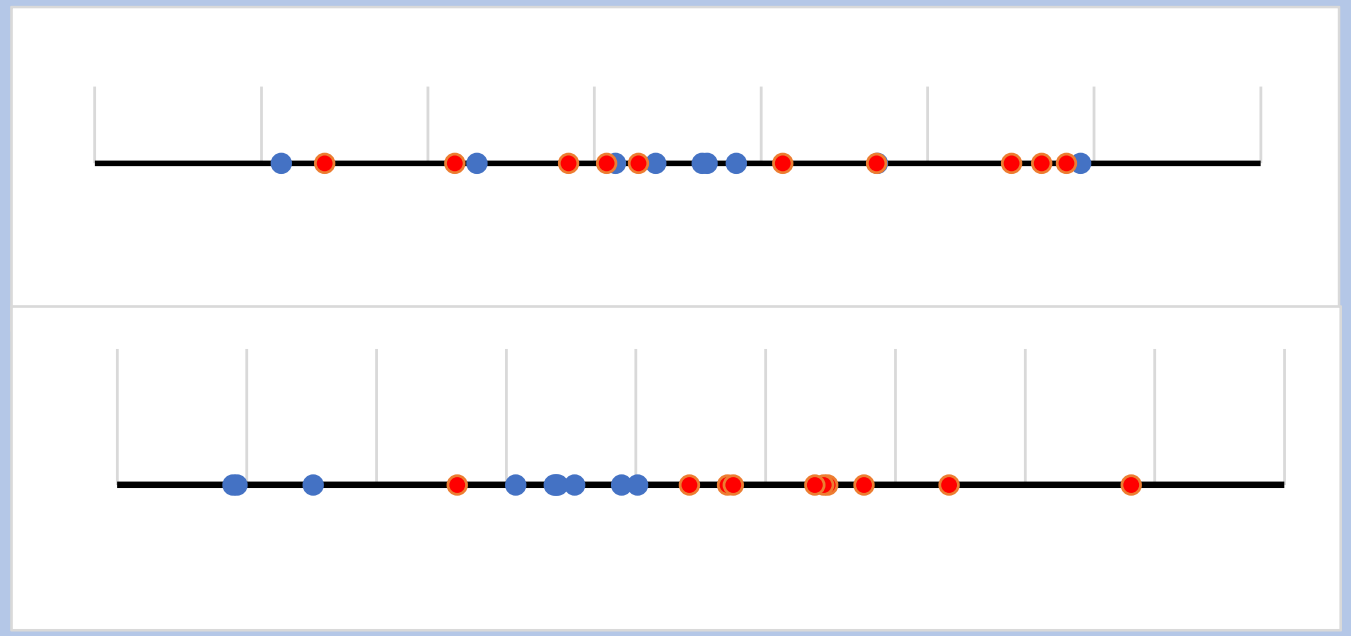
¿Cuándo un efecto es estadísticamente significativo?

- Tests de hipótesis: herramientas para poner una hipótesis a prueba
- P-valor: medida de la significancia de una hipótesis

Cuanto más cercano a 0 es el p-valor, más seguros estamos de la significatividad de un efecto o de una hipótesis

Ambas muestras no difieren significativamente (p=0.35)

Ambas muestras difieren significativamente (p=0.02)

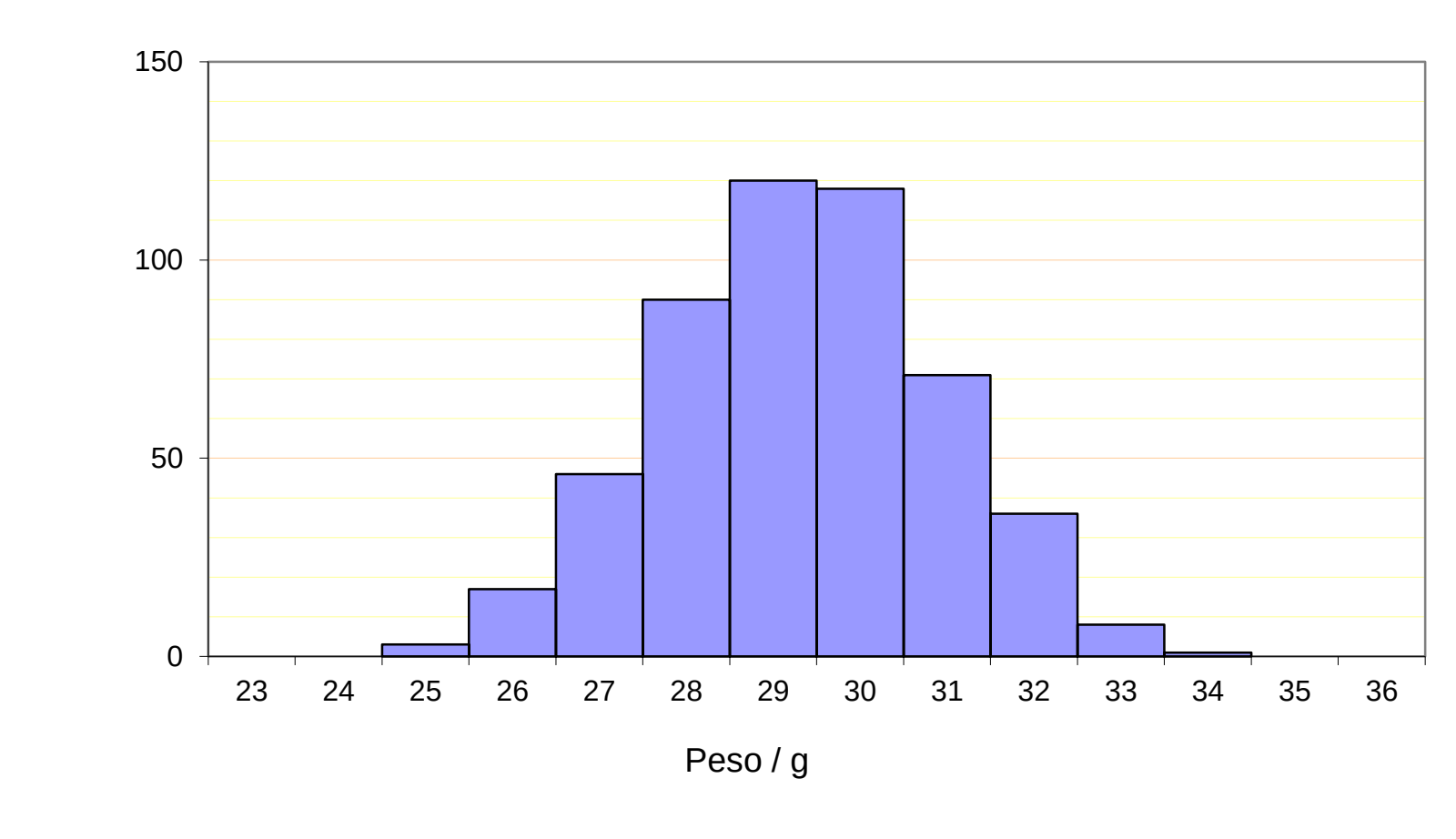


Algunos gráficos importantes

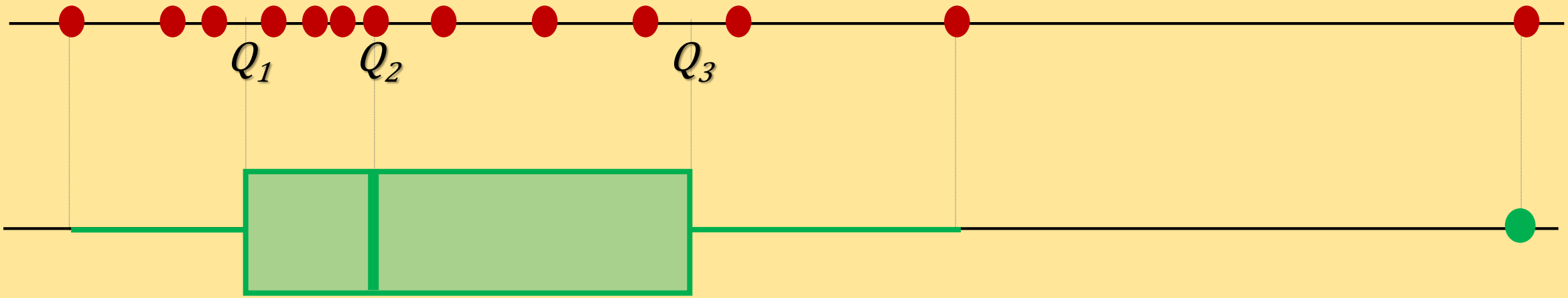
- Histogramas
- Gráficos de caja y bigote (Box-plots)
- Gráficos de correlación

Histograma:

Nos muestra cómo se distribuyen las distintas observaciones de una variable numérica

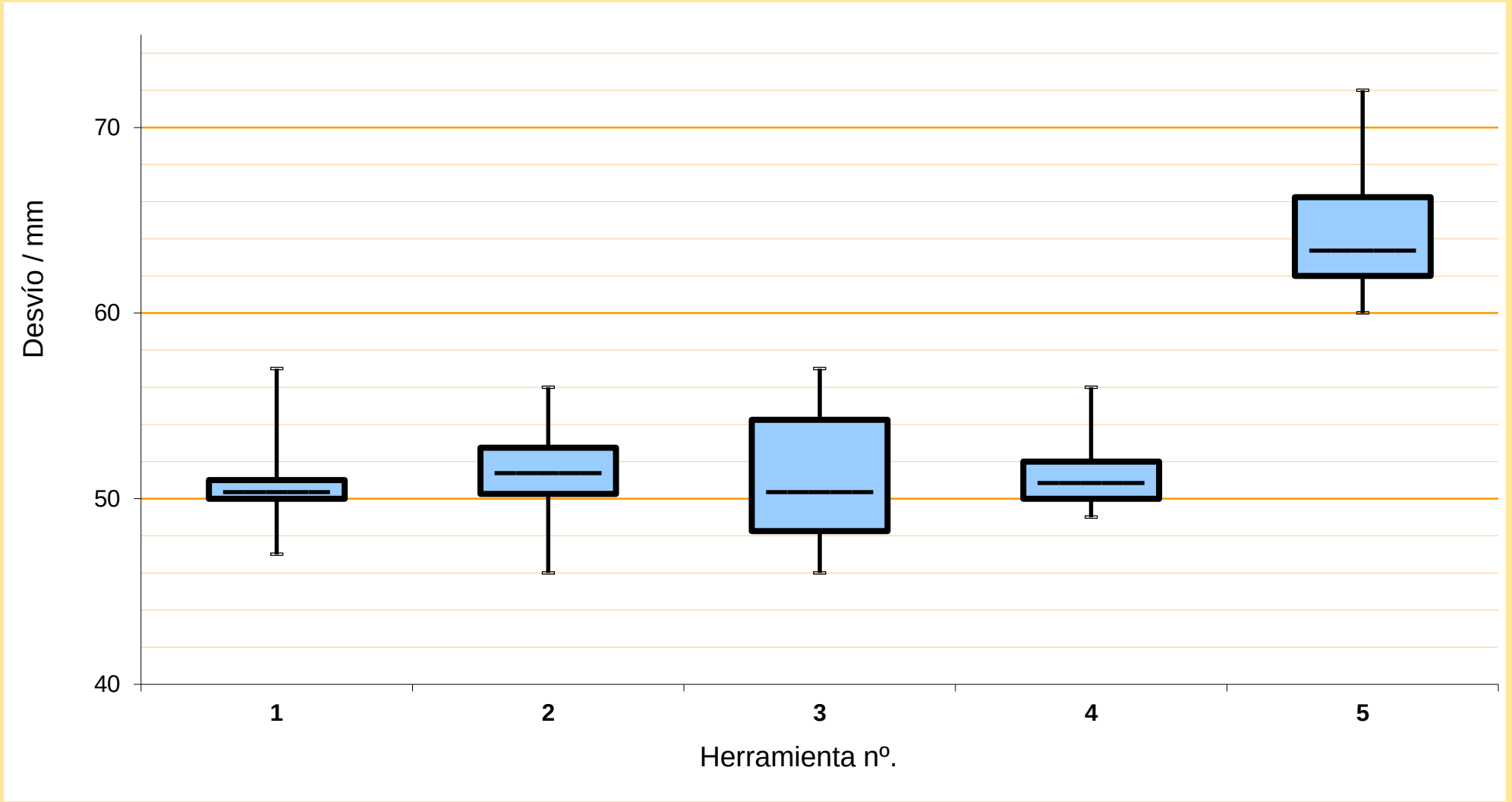


Boxplot (Gráfico de caja y bigotes)



$Q_1\,Q_2\,Q_3$: Cuartiles

1. Línea gruesa: mediana:
2. Caja: desde Q1 hasta Q3
3. Bigotes: Desde la caja hasta la última observación anterior a 1,5 RIQ
4. Outliers: los que quedan afuera de los bigotes

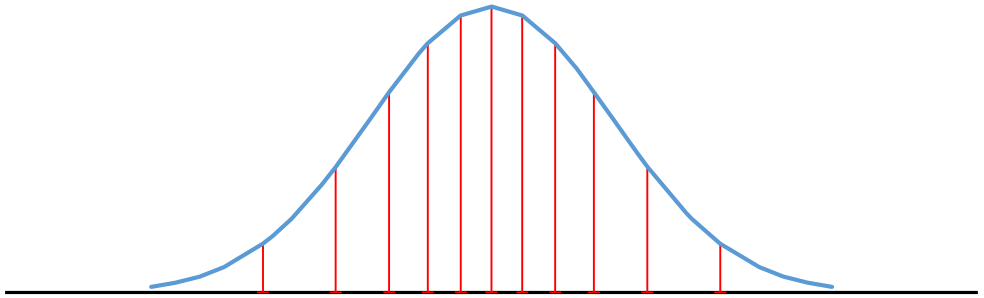

Desvío / mm
70
60
50
40
1
2
3
4
5
Herramienta nº.

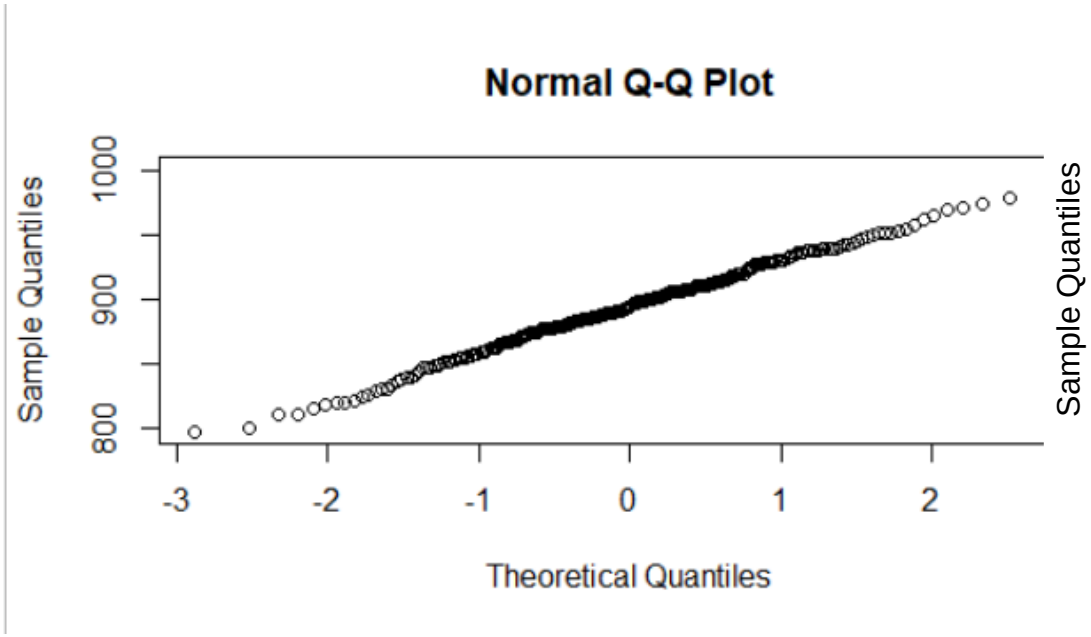
Q- Q Plots

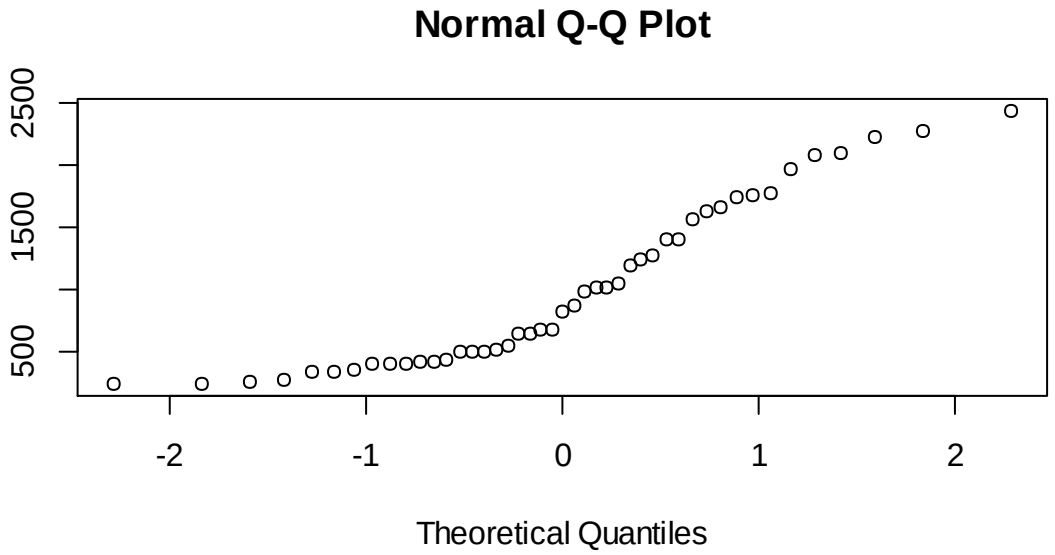
Una forma gráfica y muy práctica de chequear gráficamente si un serie de datos pueden modelarse según una determinada distribución (por ej. Normal)

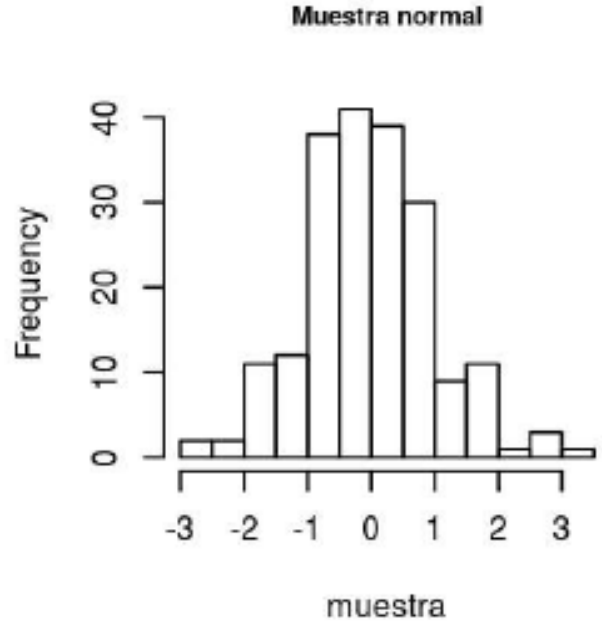
Se trata de comparar los datos (cuantiles muestrales) con cuantiles teóricos de la distribución en estudio

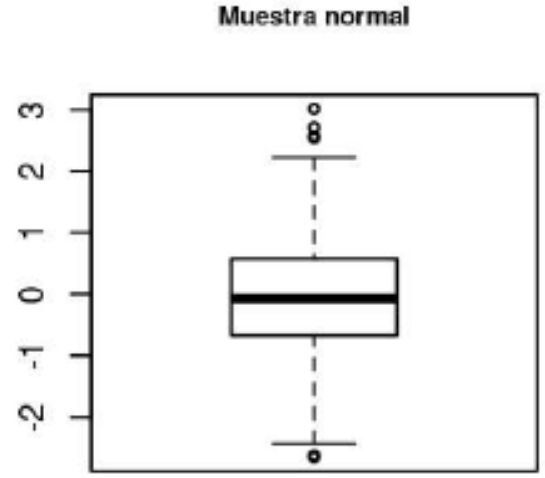
En R: *qqplot, qqnorm*

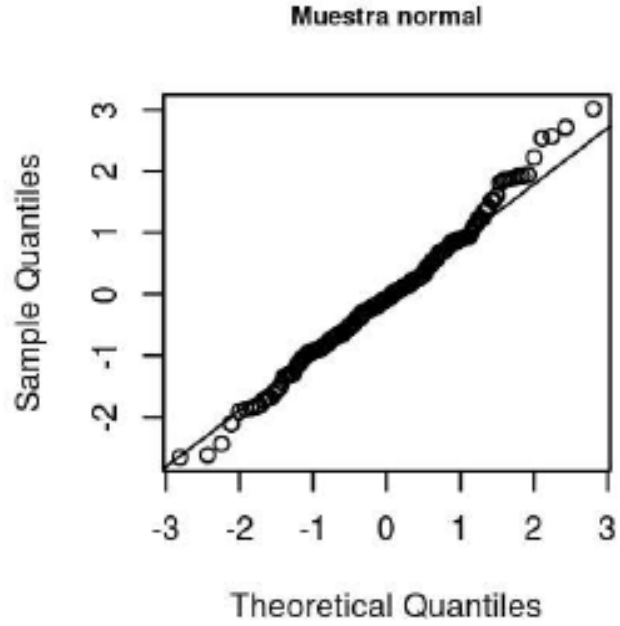


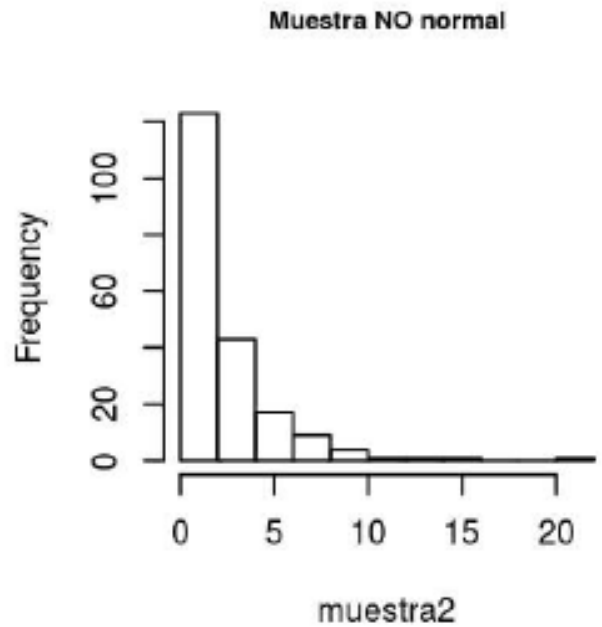


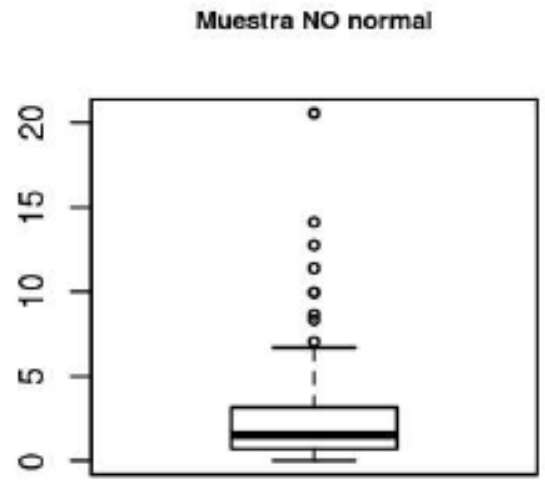


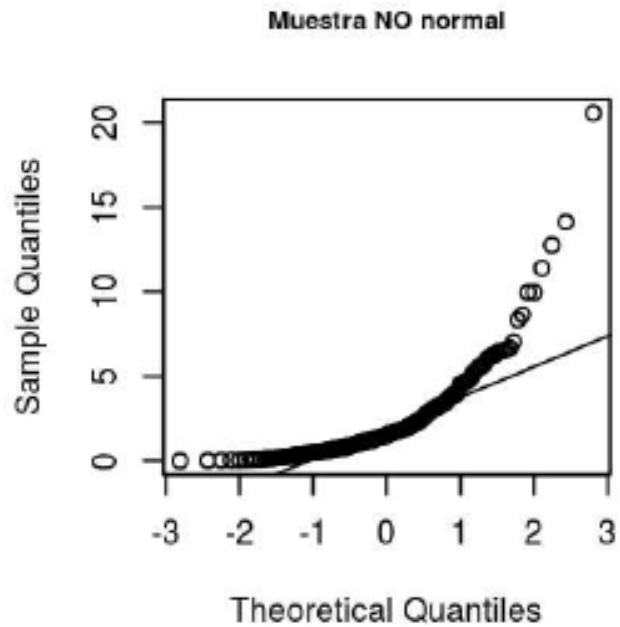

Muestra normal
Frequency
0
10
20
30
40
-3
-2
-1
0
1
2
3
muestra


Muestra normal
-2
-1
0
1
2
3


Muestra normal
Sample Quantiles
-2
-1
0
1
2
3
Theoretical Quantiles


Muestra NO normal
Frequency
0
20
60
100
0
5
10
15
20
muestra2


Muestra NO normal
0
5
10
15
20


Muestra NO normal
Sample Quantiles
0
5
10
15
20
-3
-2
-1
0
1
2
3
Theoretical Quantiles

Matrices de covarianza, matrices de correlación

$$var = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

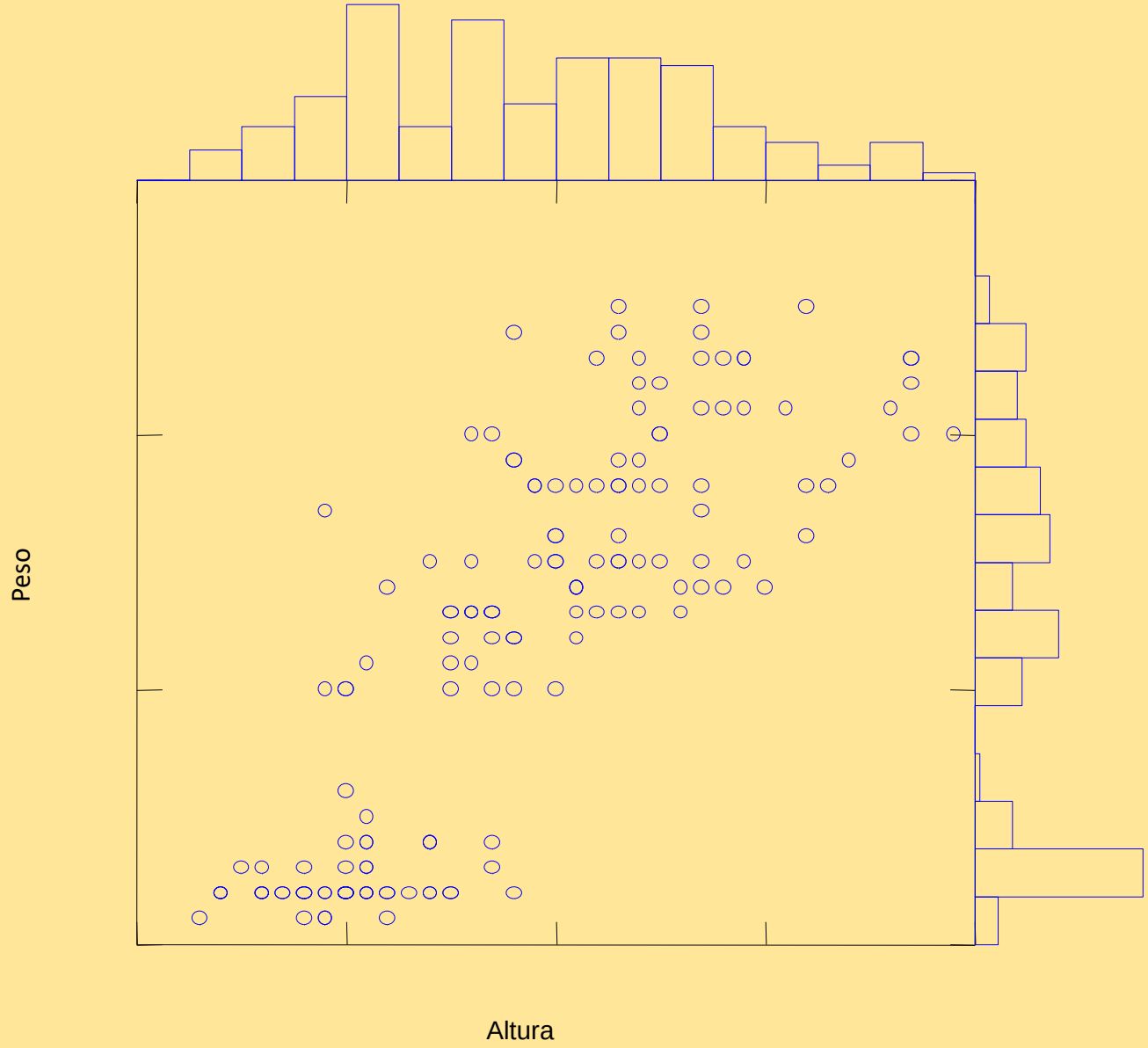
Covarianza: $cov(x,y) = \overline{x \cdot y} - \bar{x} \cdot \bar{y}$

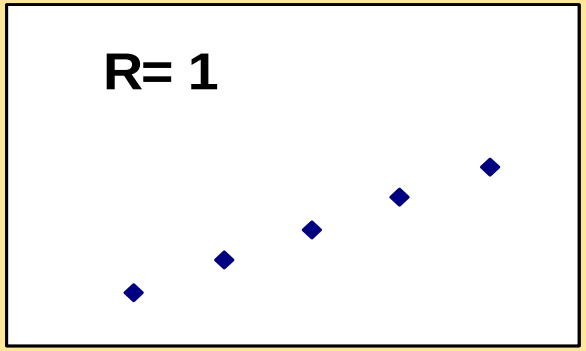
Correlación: $R = \frac{cov(X,Y)}{s_X \cdot s_y} \qquad -1 \leq R \leq 1$

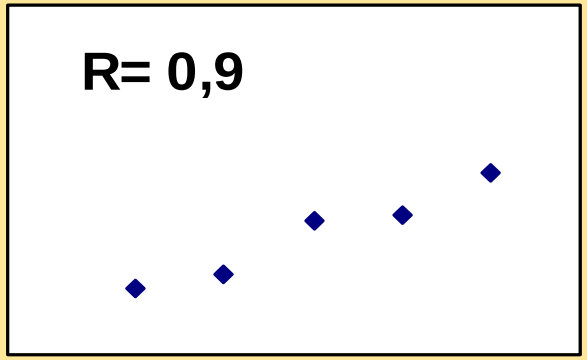
$$\text{cov}(\boldsymbol{X}) = \begin{bmatrix} V(x_1) & \cdots & cov(x_1,x_j) & \cdots & cov(x_1,x_p) \\ \vdots & \ddots & \vdots & \iddots & \cdots \\ cov(x_i,x_1) & \cdots & V(x_i) & \cdots & cov(x_i,x_p) \\ \vdots & \iddots & \vdots & \ddots & \cdots \\ cov(x_p,x_1) & \cdots & cov(x_p,x_i) & \cdots & V(x_p) \end{bmatrix}$$

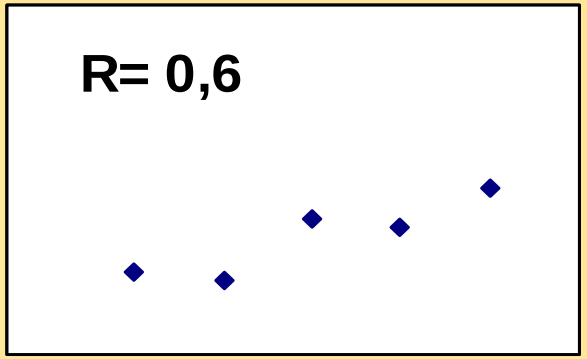
$$\text{cor}(\boldsymbol{X}) = \begin{bmatrix} 1 & \cdots & cor(x_1,x_j) & \cdots & cor(x_1,x_p) \\ \vdots & \ddots & \vdots & \iddots & \cdots \\ cor(x_i,x_1) & \cdots & 1 & \cdots & cor(x_i,x_p) \\ \vdots & \iddots & \vdots & \ddots & \cdots \\ cor(x_p,x_1) & \cdots & cor(x_p,x_i) & \cdots & 1 \end{bmatrix}$$

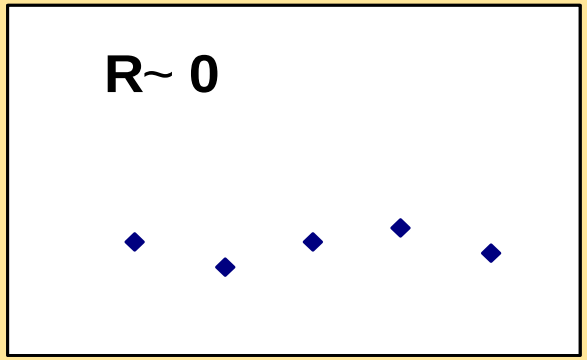
Gráficos de correlación

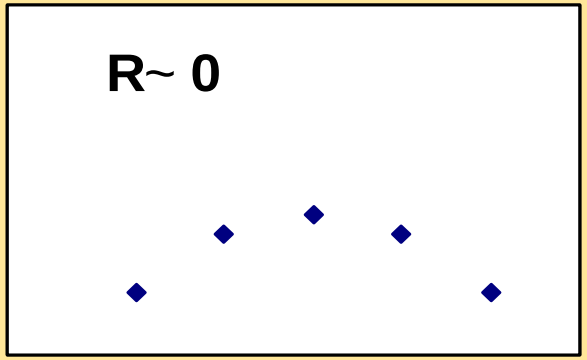


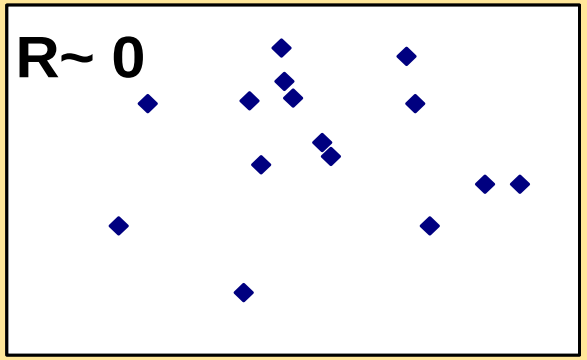

R= 1

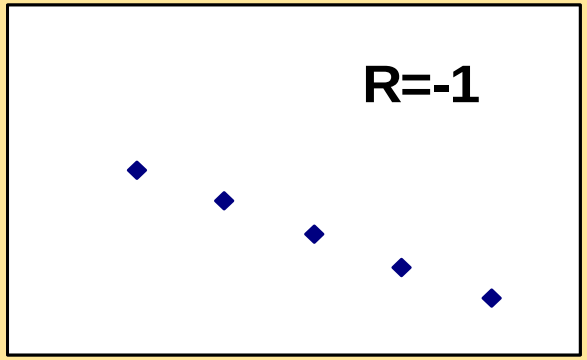

R= 0,9

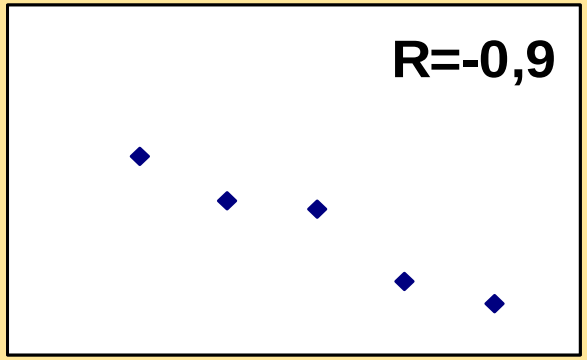

R= 0,6

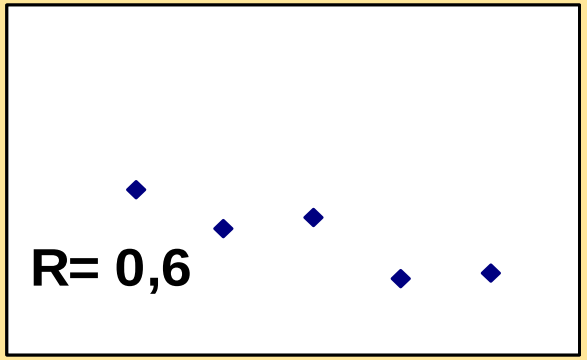

R~ 0

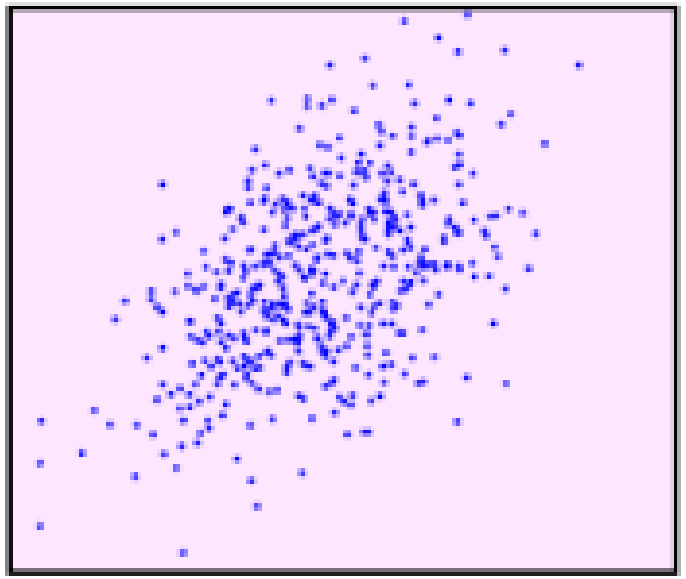

R~ 0

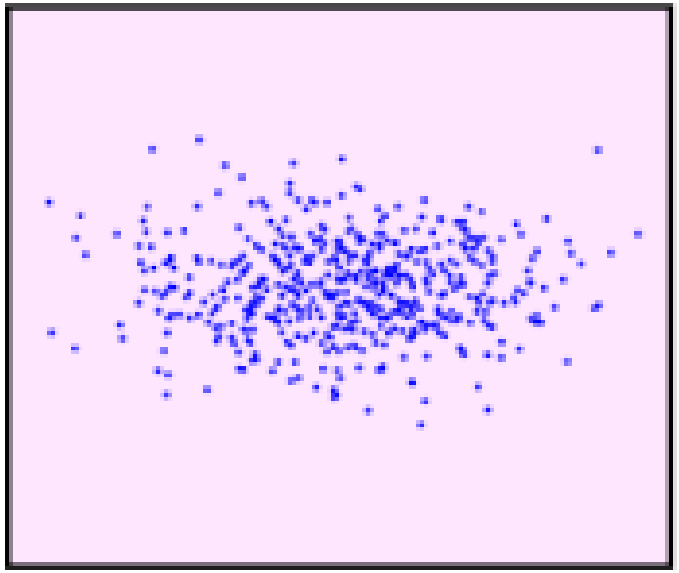

R~ 0

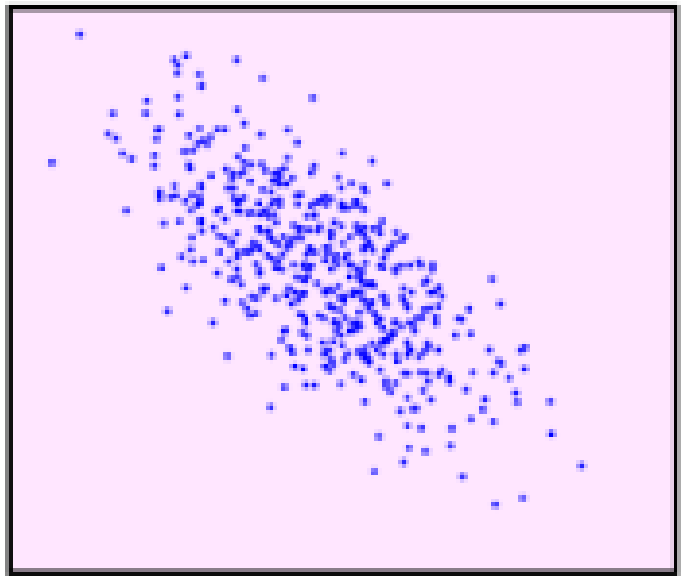

R=-1

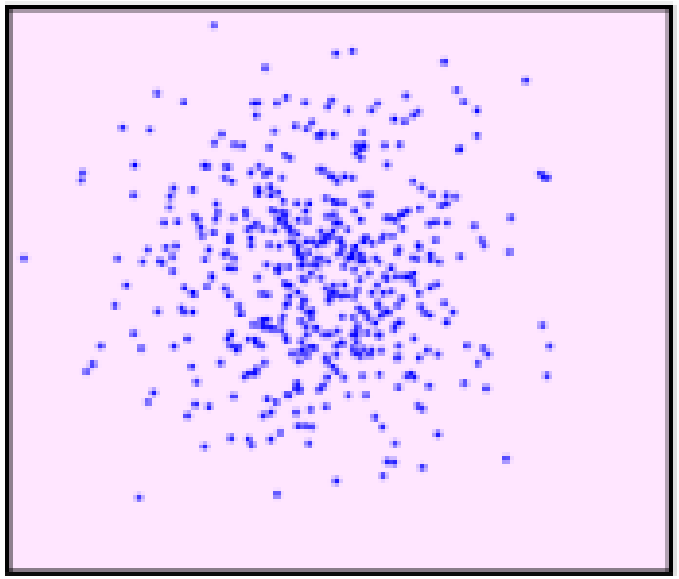

R=-0,9


R= 0,6









$$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\Sigma_2 = \begin{pmatrix} 1 & -0.7 \\ -0.7 & 1 \end{pmatrix}$$

$$\Sigma_3 = \begin{pmatrix} 1 & 0,5 \\ 0,5 & 1 \end{pmatrix}$$

$$\Sigma_4 = \begin{pmatrix} 5 & 0 \\ 0 & 1 \end{pmatrix}$$

MÉTODOS NO SUPERVISADOS

- Los métodos de **Estadística No Supervisada** se aplican cuando disponemos una seria de variables observadas $x_1, \ldots x_p$ No tenemos una variable respuesta (y)

- No estamos interesados en predecir observaciones futuras, sino sólo en describir los datos que ya tenemos o en encontrar aspectos interesantes en ellos

- Muchas veces es necesario encontrar descripciones simples de datasets con muchas variables y observaciones

- En general, los métodos no supervisados se utilizan como análisis exploratorios de datos, que anteceden a la aplicación de un método supervisado

- A veces los resultados pueden ser algo más subjetivos. No siempre es posible validar los métodos: no siempre hay una respuesta correcta

¿Qué podemos hacer cuando tenemos data-sets muy grandes (muchas observaciones y/o muchas variables)?

Reducir el número de variables (dimensión)

Agrupar o segmentar observaciones en grupos más pequeños, que puedan ser tratados por separado

Variables 1,…,p

Observaciones 1,…,n

id	Cement (comp 1)	Blast Furnace Slag	Fly Ash	Water	Superplasticizer	Coarse Aggregate	Fine Aggregate	Age (day)	Concrete compressive strength /Mpa
1	540,0	0,0	0,0	162,0	2,5	1040,0	676,0	28	79,99
2	540,0	0,0	0,0	162,0	2,5	1055,0	676,0	28	61,89
3	332,5	142,5	0,0	228,0	0,0	932,0	594,0	270	40,27
4	332,5	142,5	0,0	228,0	0,0	932,0	594,0	365	41,05
5	198,6	132,4	0,0	192,0	0,0	978,4	825,5	360	44,30
6	266,0	114,0	0,0	228,0	0,0	932,0	670,0	90	47,03
7	380,0	95,0	0,0	228,0	0,0	932,0	594,0	365	43,70

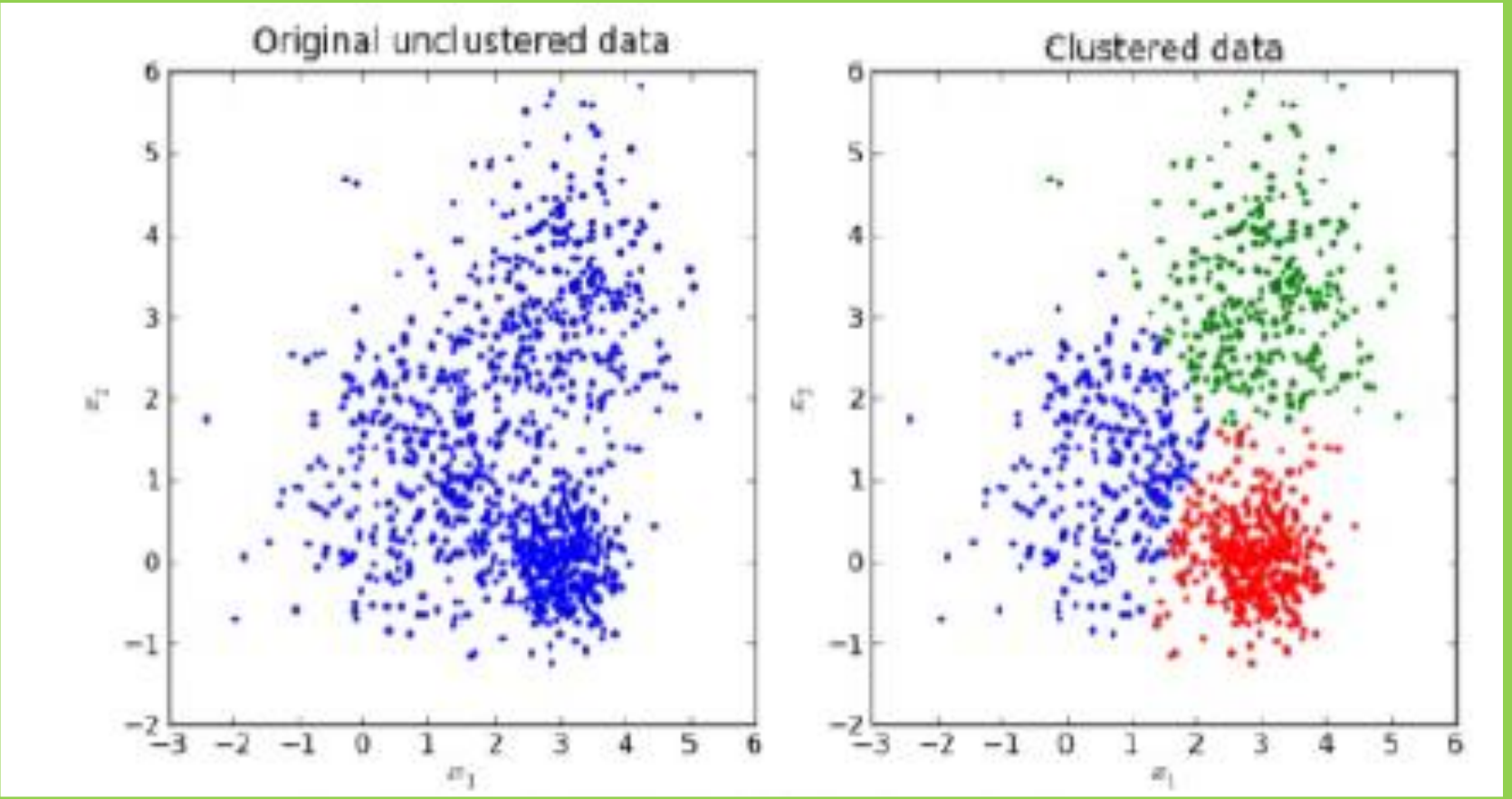
¿Qué podemos hacer cuando tenemos data-sets muy grandes muchas observaciones y/o muchas variables?

- Métodos de agrupamiento (clustering): Buscar grupos internamente homogéneos y con diferencias entre ellos

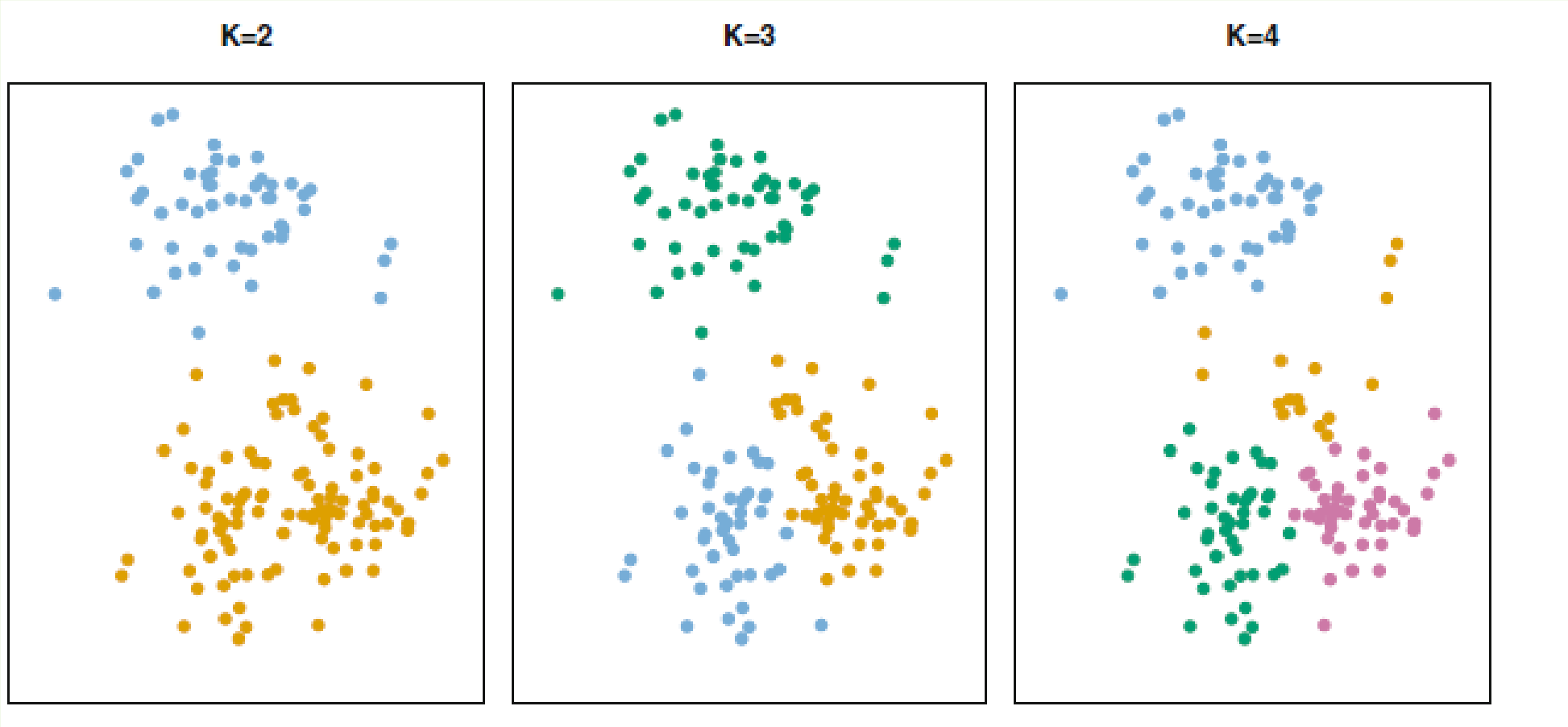
- Métodos de reducción de dimensión: Análisis de componentes principales, explicando una gran fracción de la varianza

¿Cómo agrupar “naturalmente” un conjunto de datos?

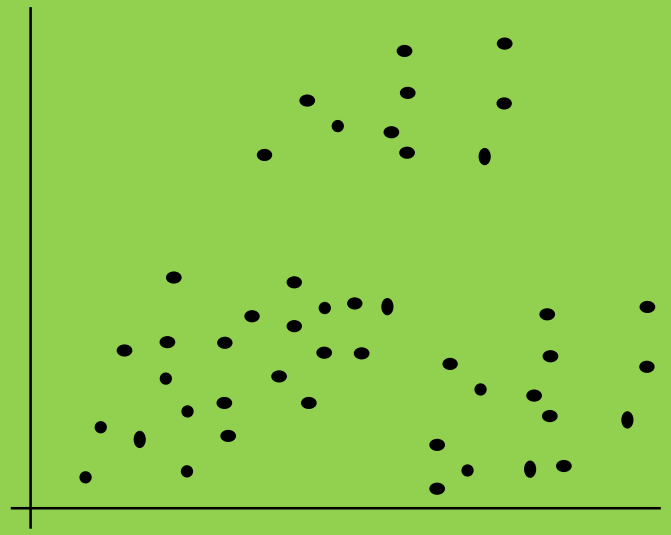
Sin criterio preestablecido



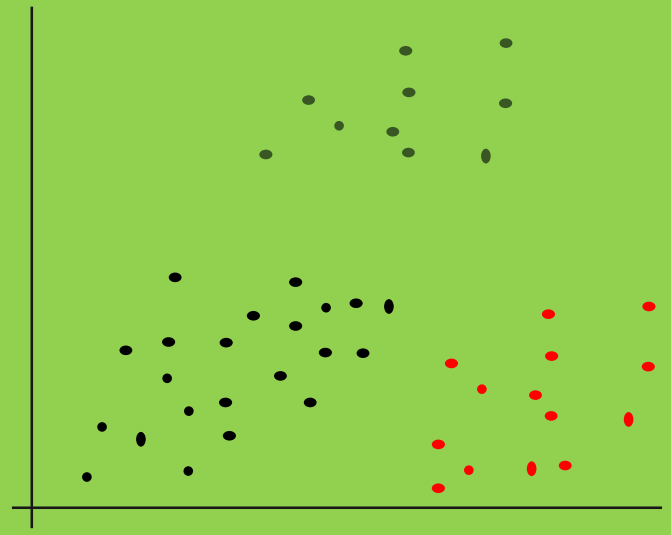
¿Cuántos grupos?



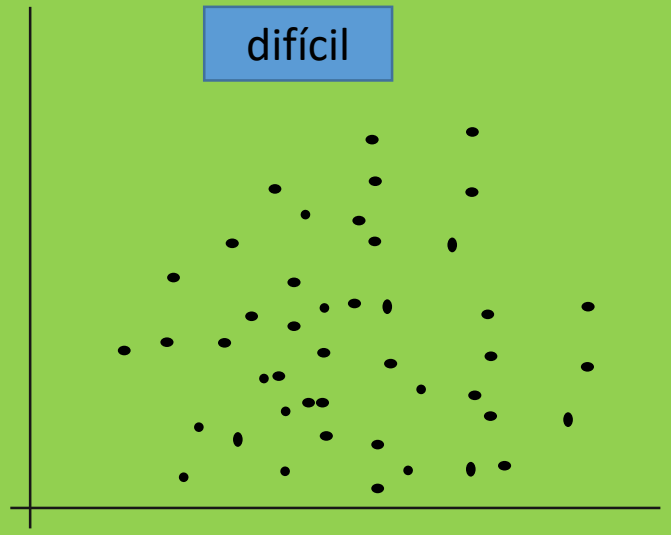
A veces resulta fácil agrupar, a veces no



fácil



difícil



Es esencial definir qué noción de distancia vamos a emplear

Algunos métodos de clustering

- K-medias, K-medianas
- Jerarquizado Bottom-up
- Jerarquizado Top-down

} dendogramas

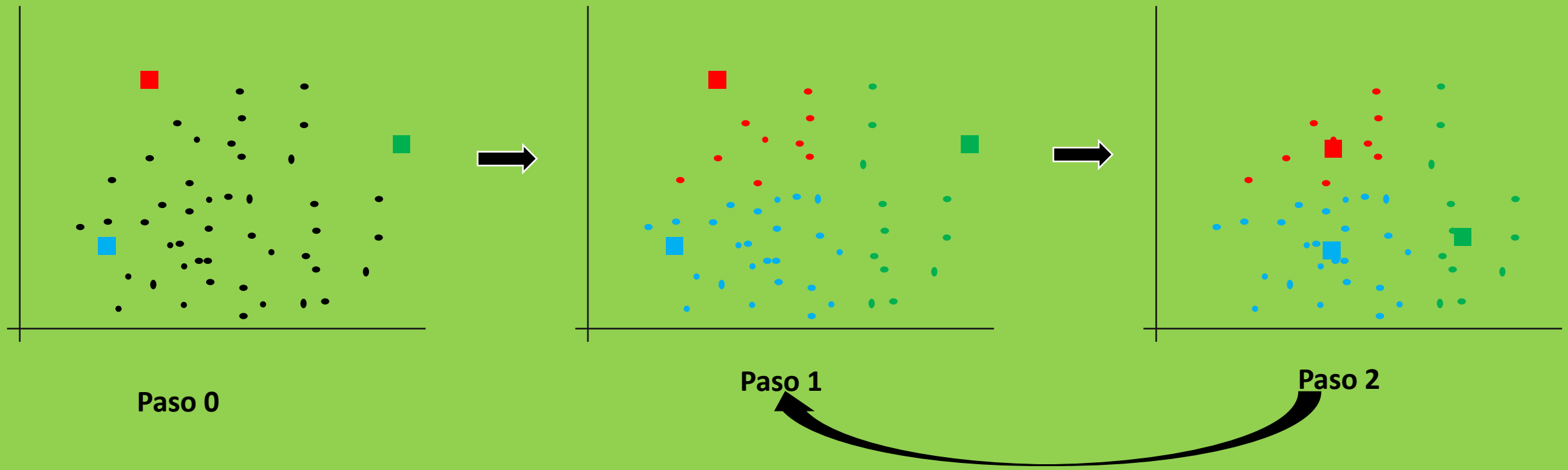
Distancias:

- Euclideana
- Minkovsky
- otras

K-medias:

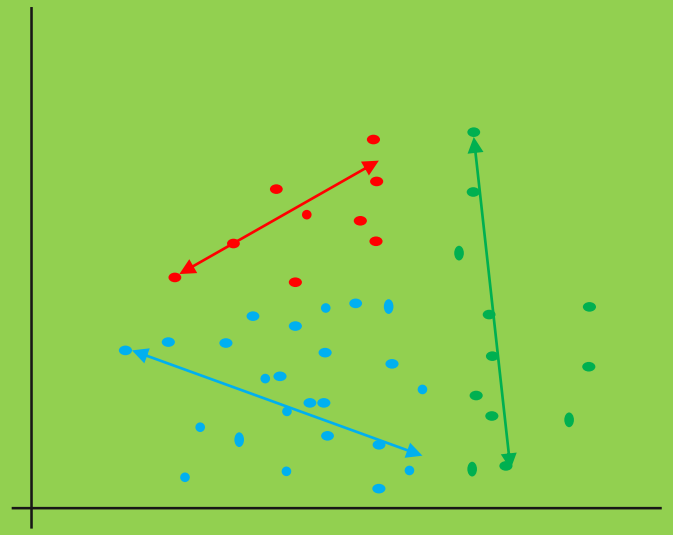
0. Elijo el número de grupos (k) y elijo arbitrariamente k centroides, uno por cada grupo
1. Asigno grupo a cada elemento en función del centroide más cercano
2. Recalculo los centroides como el promedio (mediana) de su grupo

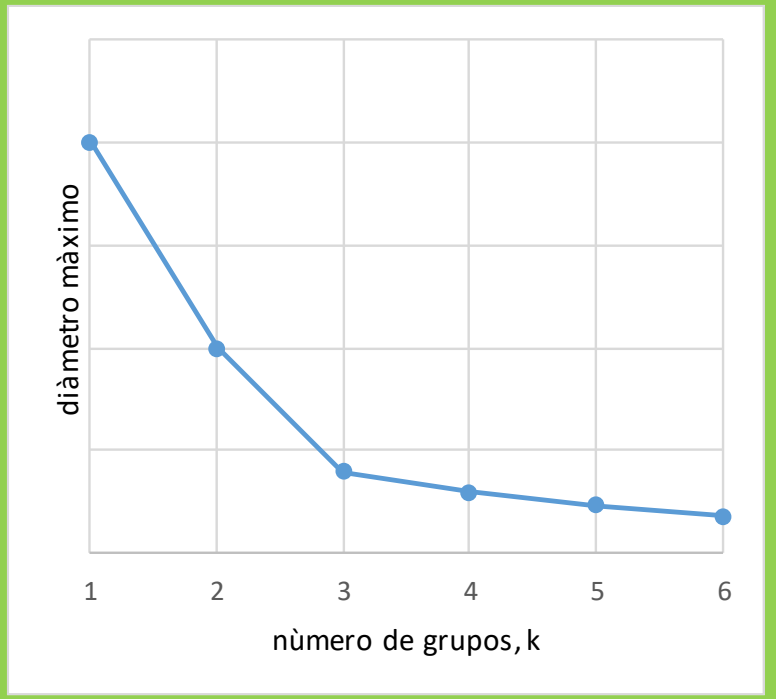
Repito 1 y 2 hasta converger

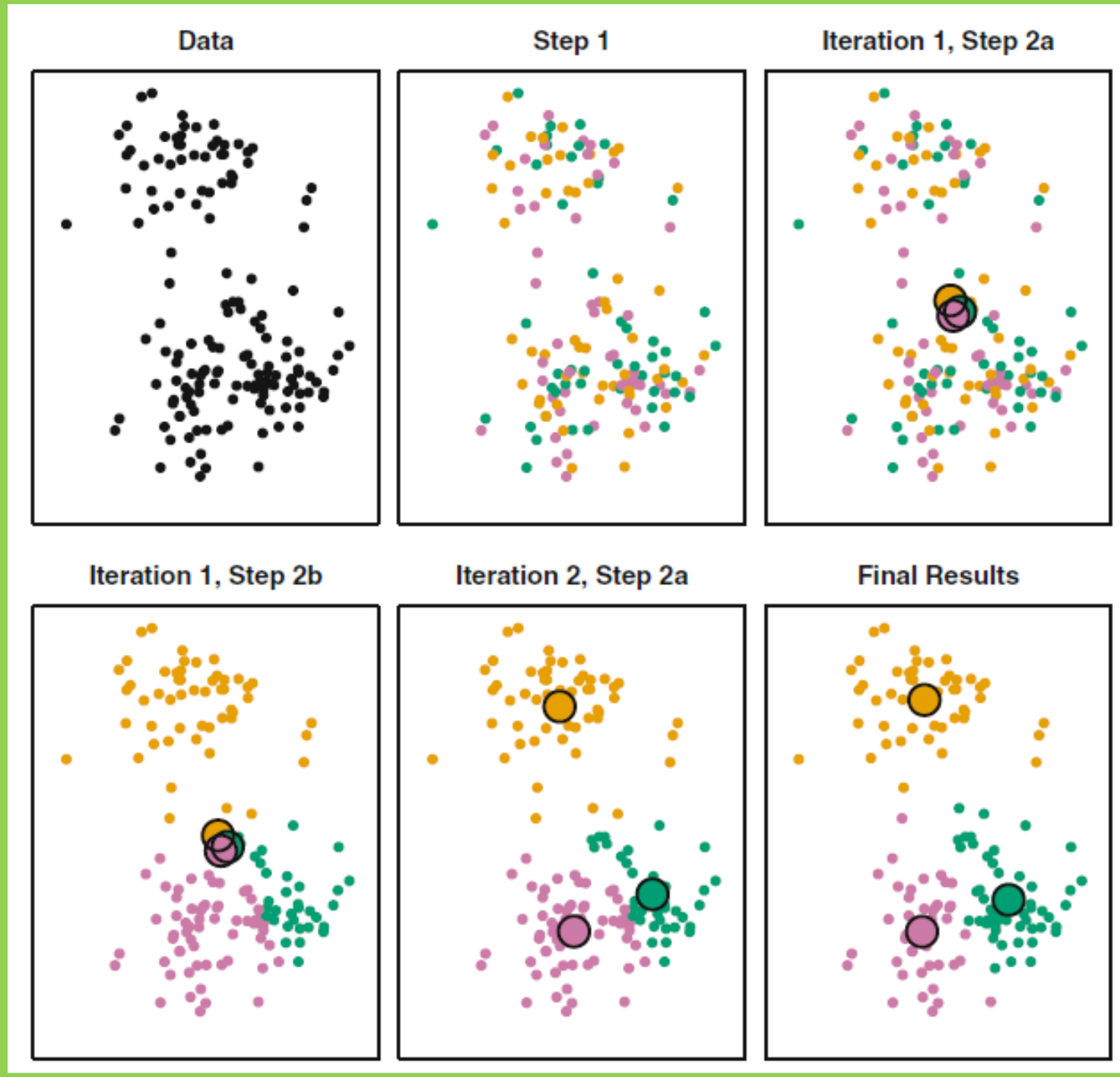


¿Cómo sé si el número de grupos es adecuado?

- Criterios de Análisis de la Varianza (ANOVA): minimizar la distancia intra-grupo, maximizar la distancia inter-grupos
- Graficar el diámetro máximo en función de k






Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a
Final Results

Medidas de homogeneidad de clusters

Dada una partición del dataset completo en *K* clusters $C_1 \cdots C_K$, de tamaño $n_1 \ldots n_K$ ($n_1 + \cdots + n_K = n$), una medida de homogeneidad del cluster *k* es

$$W(C_k) = \frac{\sum_{i,j} \|x_i - x_j\|^2}{n_k}$$

La mejor partición es aquella que minimiza

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}.$$

Algoritmo K-medias

1. Definir qué distancia se usará (por ejemplo, euclídea)

2. Definir el criterio de finalización del algoritmo

3. Elegir en número de grupos, k

4. Asignar aleatoriamente cada observación al grupo 1, 2, …., k: agrupamiento inicial

5. Iterar a-b hasta que se cumpla el criterio de finalización

    a. Recalcular el centroide, de cada uno de los k clusters, como la media coordenada a coordenada de sus observaciones

    b. Reasignar cada observación al cluster de centroide más cercano (según la distancia elegida)

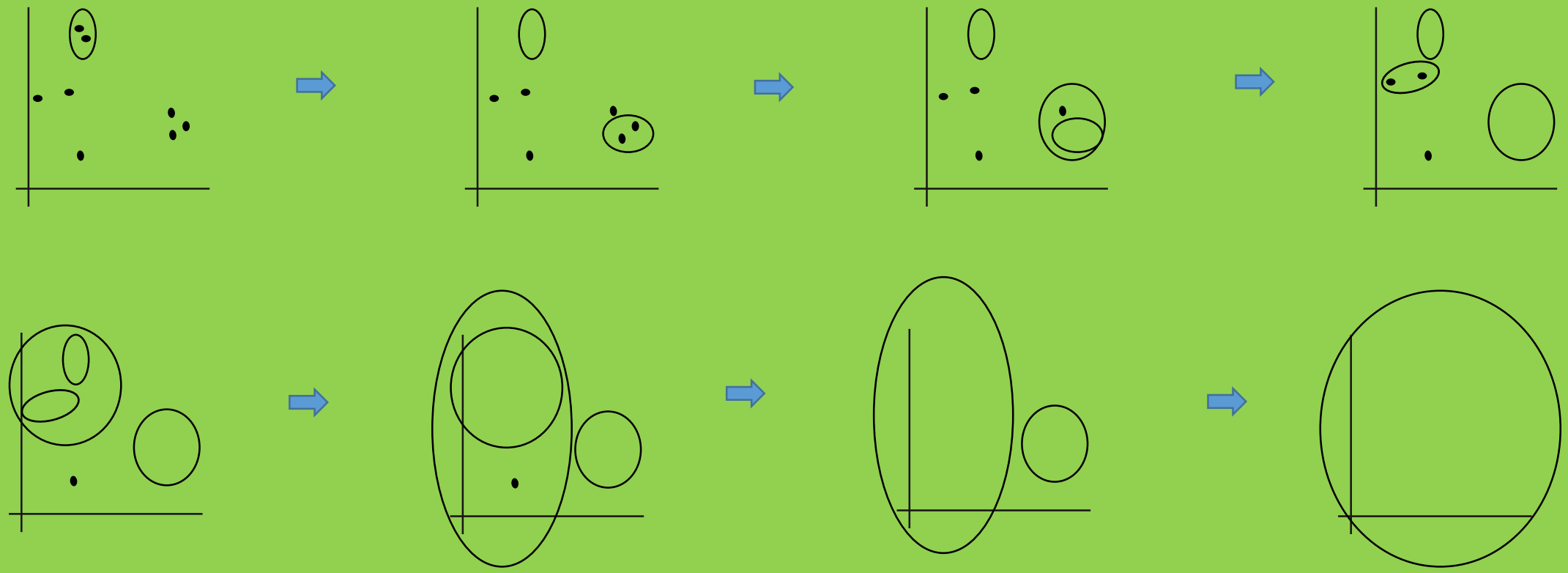
6. Repetir 4-5 n veces, y elegir el agrupamiento mejor

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}.$$

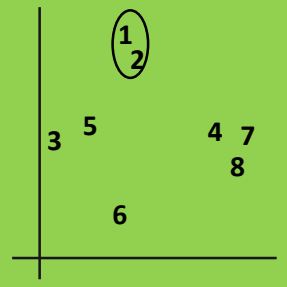
Algoritmo k-medias corriendo 6 veces las etapas 4-5

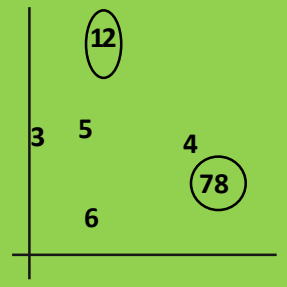
Métodos jerárquicos y dendogramas

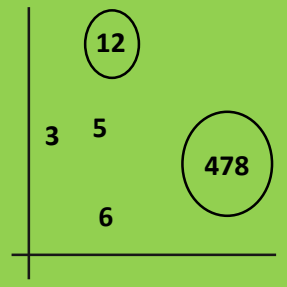
1. Identifico los dos elementos más cercanos y los uno en un grupo común
2. Repito 1 hasta que me quede un único grupo

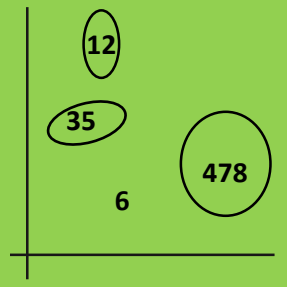


Bottom-up



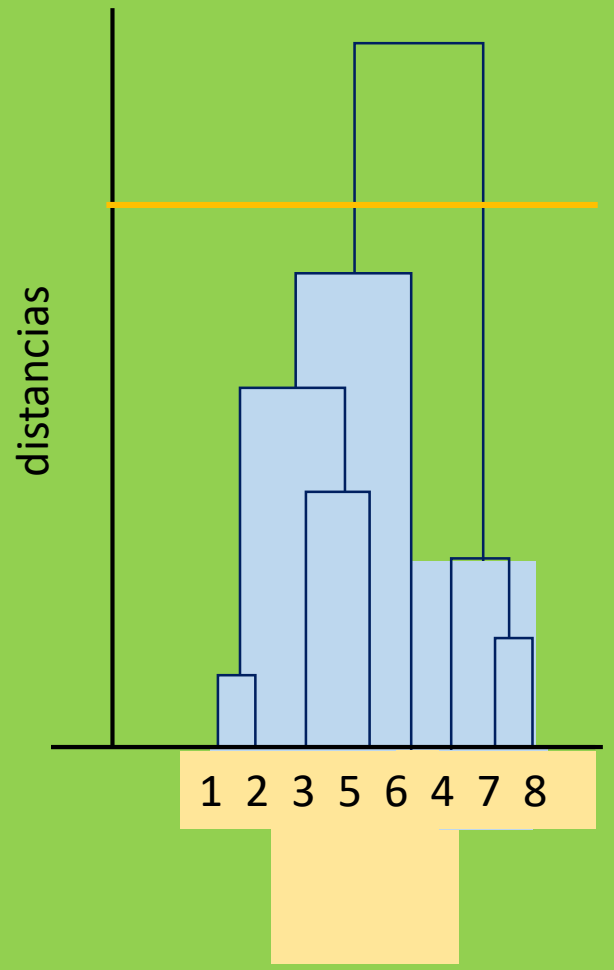






Formación del dendograma:

¿dónde quiero cortar?



Medidas de disimilaridad entre dos clusters A y B

Completa: Máxima disimilaridad entre los clusters A y B. Se obtiene computando el máximo ente todas las distancias $\|x_i - x_j\|$ para todos los pares de observaciones (i, j) con i en el cluster A y i en el cluster B

Simple: Mínima disimilaridad entre los clusters A y B. Se obtiene computando el mínimo ente todas las distancias $\|x_i - x_j\|$ para todos los pares de observaciones (i, j) con i en el cluster A y i en el cluster B

Promedio: Disimilaridad media entre los clusters A y B. Se obtiene computando el promedio ente todas las distancias $\|x_i - x_j\|$ para todos los pares de observaciones (i, j) con i en el cluster A y i en el cluster B

Centroide: Disimilaridad entre los centroides de los clusters A y B.

Algoritmo de agrupamiento jerárquico *(bottom up)*

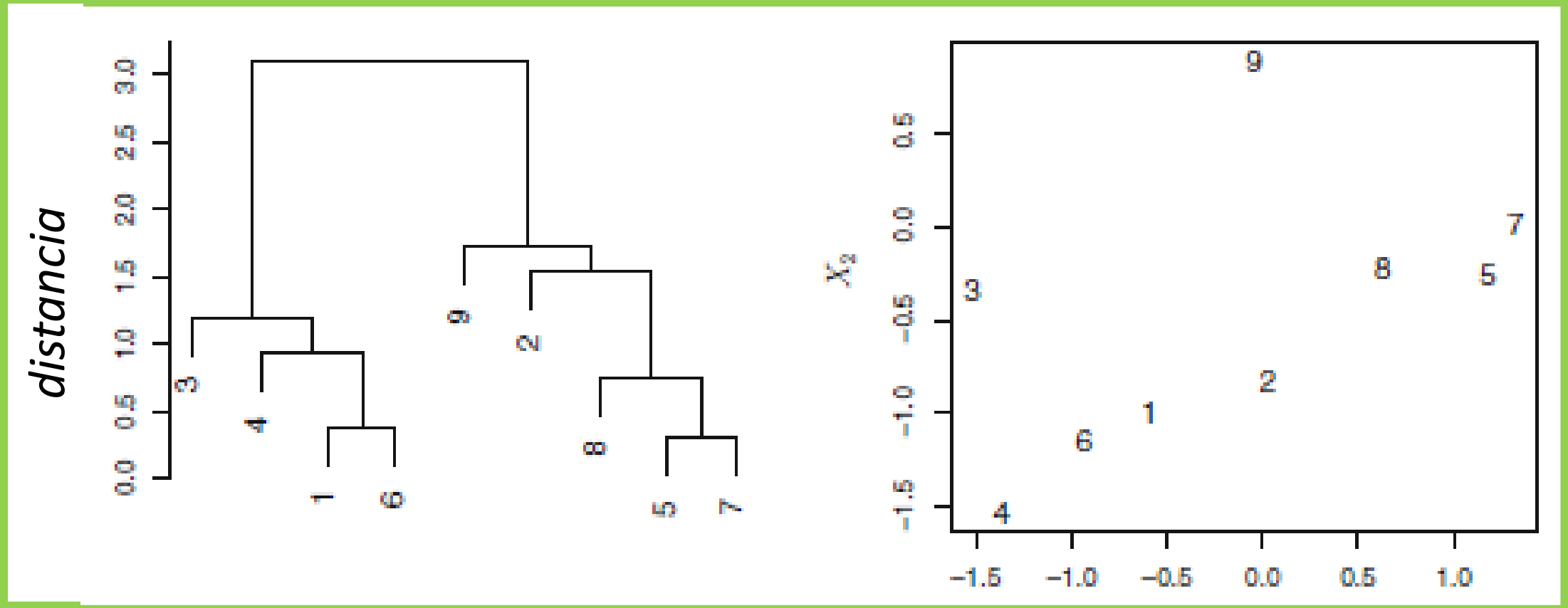
1. Definir qué distancia (por ejemplo, euclídea) y qué medida de disimilaridad entre clusters se usará. Inicialmente considerar cada observación como un cluster y Calcular la distancia entre cada uno de los $\binom{n}{2}$ pares de datos como medida inicial de disimilaridad.

2. Para i = n, n-1, ……, 2, hacer

   a. Encontrar el par de clusters más cercano, y fusionar esos dos clusters en uno. La distancia entre ambos indica a altura del dendograma correspondiente a la fusión

   b. Computar las nuevas distancias entre los pares de los i-1 clusters restantes

Interpretación del dendograma



ANÁLISIS DE COMPONENTES PRINCIPALES

¿Cómo reducir el número de variables (dimensión) perdiendo la menor información posible?

Variables 1,…,p

Observaciones 1,…,n

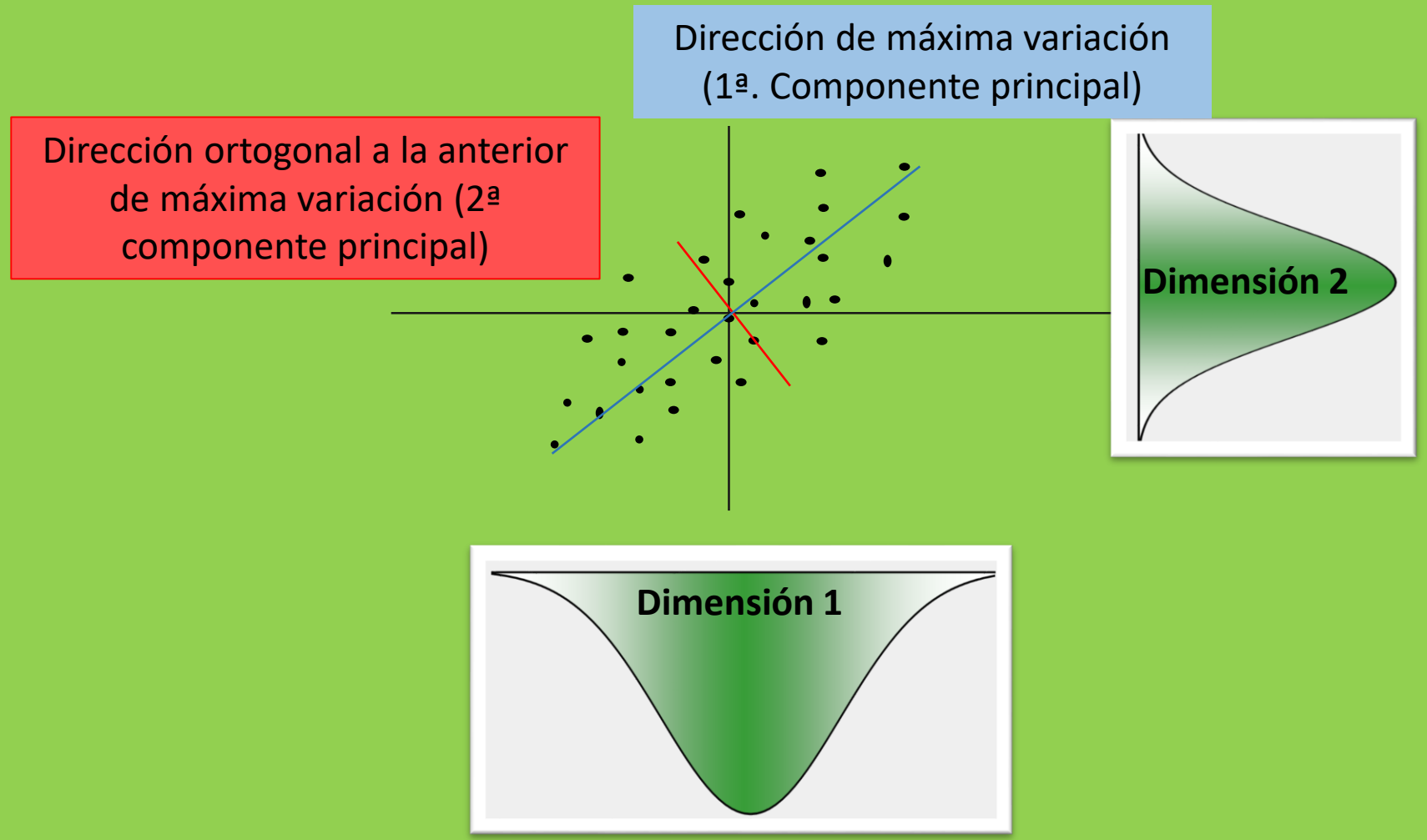
id	Cement (comp 1)	Blast Furnace Slag	Fly Ash	Water	Superplasticizer	Coarse Aggregate	Fine Aggregate	Age (day)	Concrete compressive strength /Mpa
1	540,0	0,0	0,0	162,0	2,5	1040,0	676,0	28	79,99
2	540,0	0,0	0,0	162,0	2,5	1055,0	676,0	28	61,89
3	332,5	142,5	0,0	228,0	0,0	932,0	594,0	270	40,27
4	332,5	142,5	0,0	228,0	0,0	932,0	594,0	365	41,05
5	198,6	132,4	0,0	192,0	0,0	978,4	825,5	360	44,30
6	266,0	114,0	0,0	228,0	0,0	932,0	670,0	90	47,03
7	380,0	95,0	0,0	228,0	0,0	932,0	594,0	365	43,70

En lugar de **eliminar** variables (como hacíamos por ejemplo en regresión)

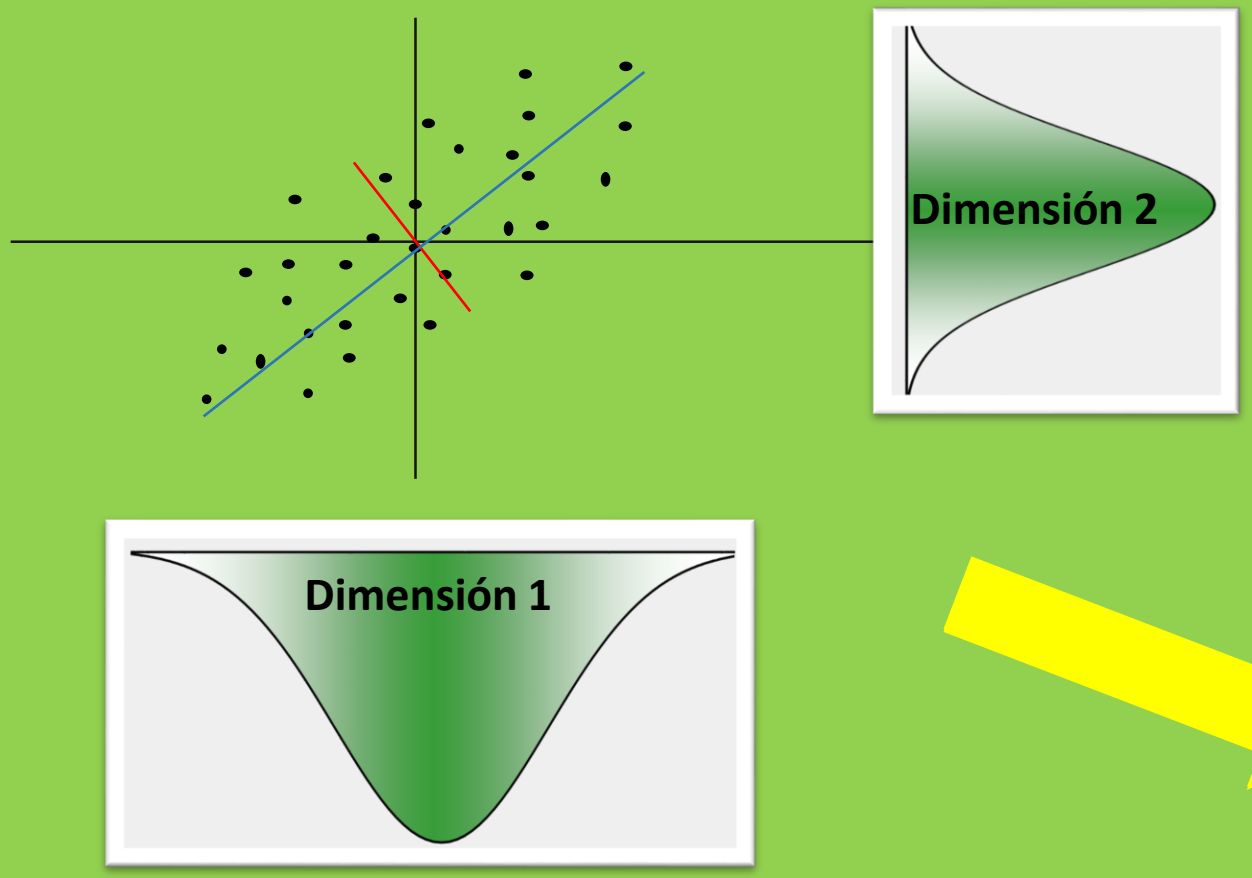
Intentaremos buscar aquellas combinaciones lineales de las variables originales que expliquen la mayor información posible

Ejemplo de reducción de dimensión (de dos dimensiones a una)

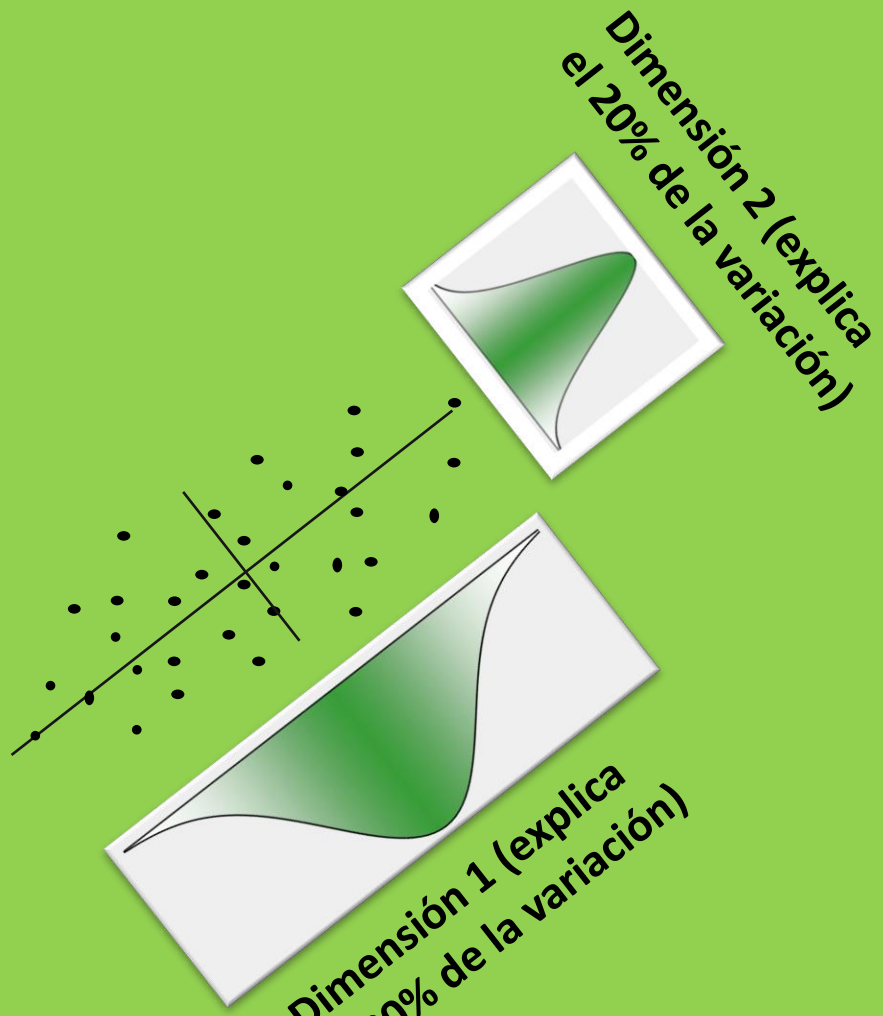
Si tenemos estos datos

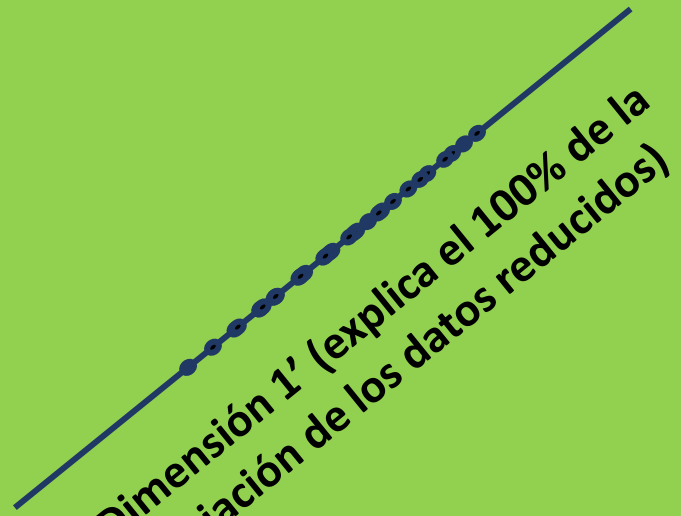


Componentes principales y reducción de dimensión



Componentes principales y reducción de dimensión



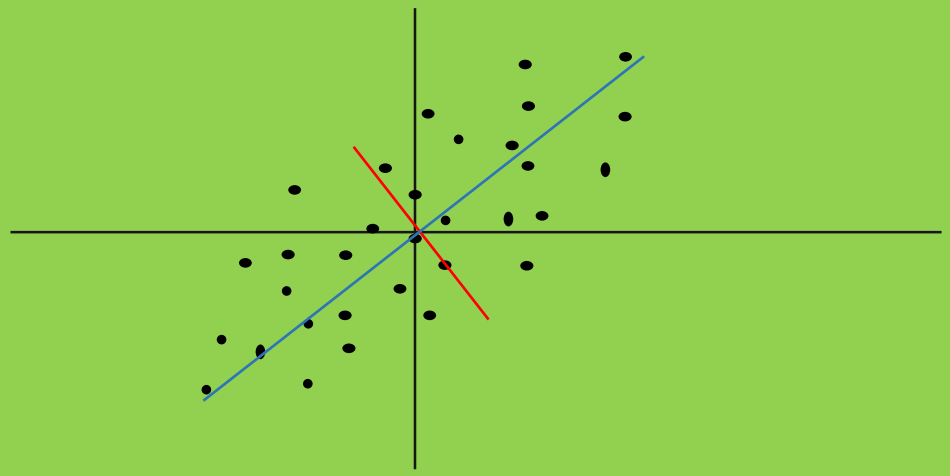


Pierdo el 20% de la información, gano en simplicidad y facilidad de interpretación

Reocordar: Matriz de varianzas y covarianzas

En dos variables (x,y)

$$\sum_X = \begin{pmatrix} var(x) & cov(x,y) \\ cov(x,y) & var(y) \end{pmatrix}$$



En general

$$\sum_X = \begin{bmatrix} V(X_1\ ) & cov(X_1,X_j) & cov(X_1,X_p) \\ cov(X_i,X_1) & V(X_i\ ) & cov(X_i,X_p) \\ cov(X_p,X_1) & cov(X_p,X_j) & V(X_p\ ) \end{bmatrix}$$

¿Qué hacemos con los datos?

Primero centramos las variables

$$
\begin{matrix}
x_{11}, x_{12}, \dots, x_{1p} \\
x_{21}, x_{22}, \dots, x_{2p} \\
. \\
. \\
. \\
. \\
. \\
x_{n1}, x_{n2}, \dots, x_{np}
\end{matrix}
$$

$$
\begin{matrix}
x_{11} - \bar{x}_1, & x_{12} - \bar{x}_2, \dots, & x_{1p} - \bar{x}_p \\
x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2, \dots, & x_{1p} - \bar{x}_p \\
 & . & \\
 & . & \\
 & . & \\
 & . & \\
 & . & \\
x_{n1} - \bar{x}_1, & x_{n2} - \bar{x}_2, \dots, x_{np} - \bar{x}_p &
\end{matrix}
$$

Después las escalamos

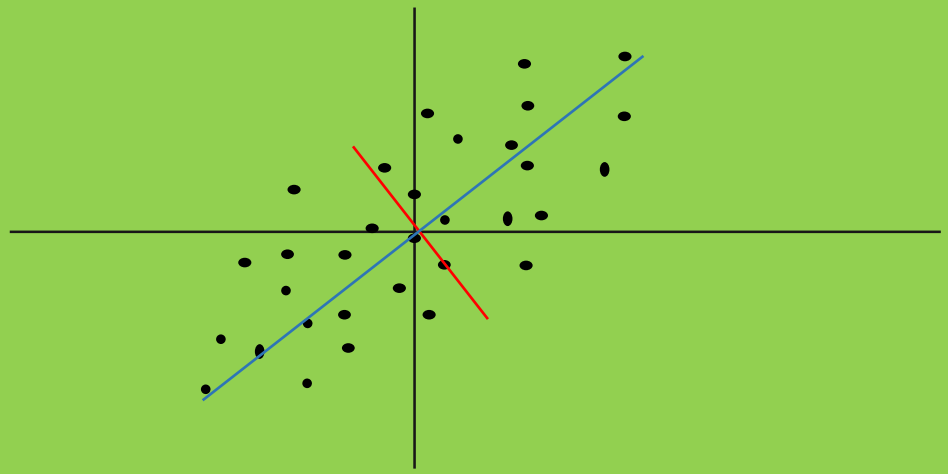
$$
\begin{matrix}
\frac{x_{11} - \bar{x}_1}{s_1}, \dots\dots\dots, \frac{x_{1p} - \bar{x}_p}{s_p} \\
\frac{x_{21} - \bar{x}_1}{s_1}, \dots\dots\dots, \frac{x_{2p} - \bar{x}_p}{s_p} \\
. \\
. \\
. \\
\frac{x_{n1} - \bar{x}_1}{s_1}, \dots\dots\dots, \frac{x_{np} - \bar{x}_p}{s_p}
\end{matrix}
$$

Primera componente principal

Combinación lineal (cambio de variables) que maximice la dispersión:

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \ldots + \phi_{p1} X_p$$

Lo mismo con las sucesivas componentes principales



5. Componentes principales

- La dirección de máxima variación es co-lineal al autovector de autovalor máximo α_1 de $\sum_X$

- El porcentaje de variación explicado por la primer componente principal es igual a

$$\frac{\alpha_1}{\alpha_1 + \cdots + \alpha_p}$$

- La segunda dirección de máxima variación es co-lineal al autovector del autovalor siguiente α_2 de $\sum_X$

- El porcentaje de variación de las dos primeras componentes principales es

$$\frac{\alpha_1 + \alpha_2}{\alpha_1 + \cdots + \alpha_p}$$

- Y así sucesivamente

Normalmente nos quedamos con las dos primeras componentes principales para representar los datos (biplot)

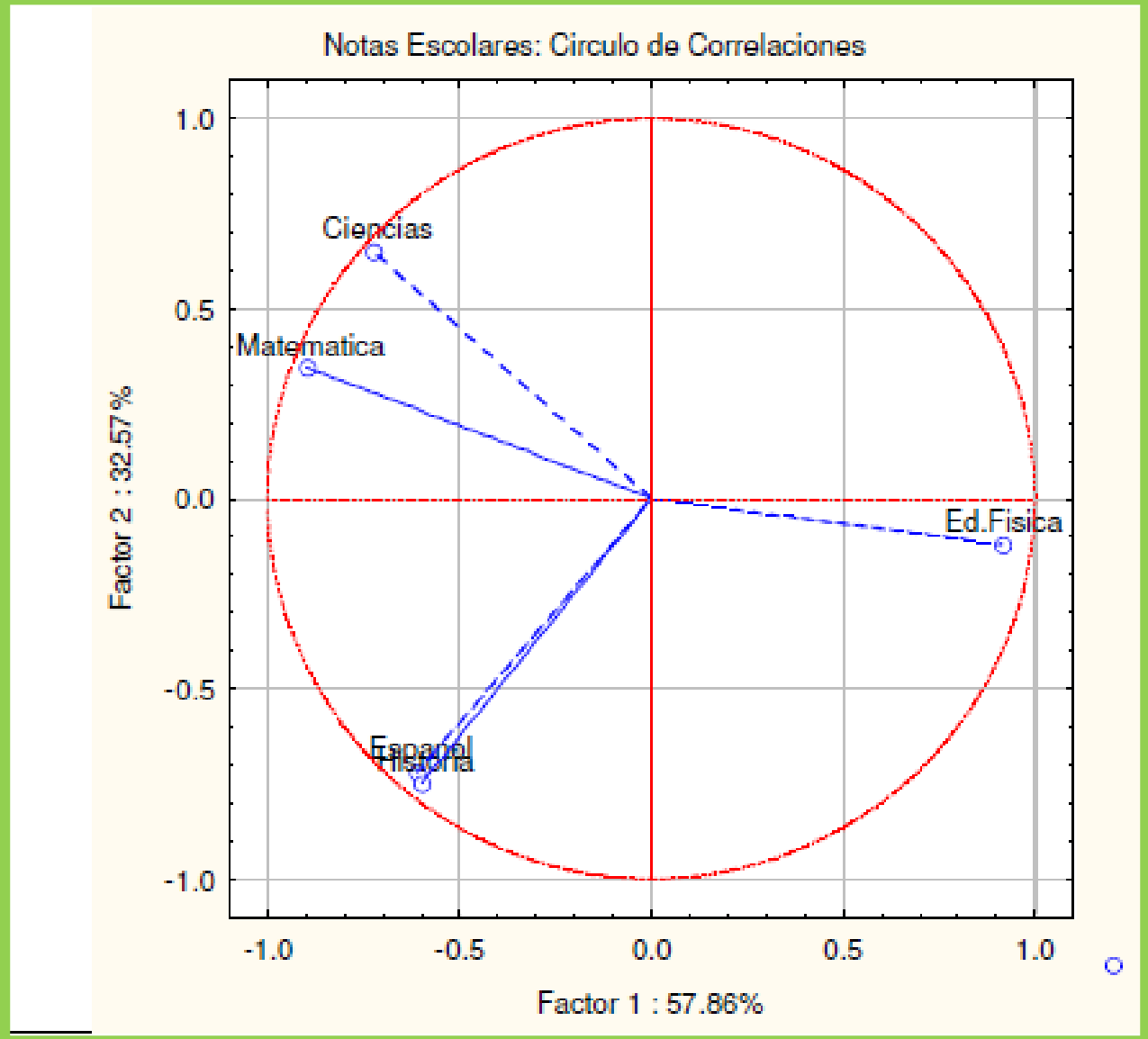
Ejemplo: notas de varias materias de la escuela para un grupo de alumnos

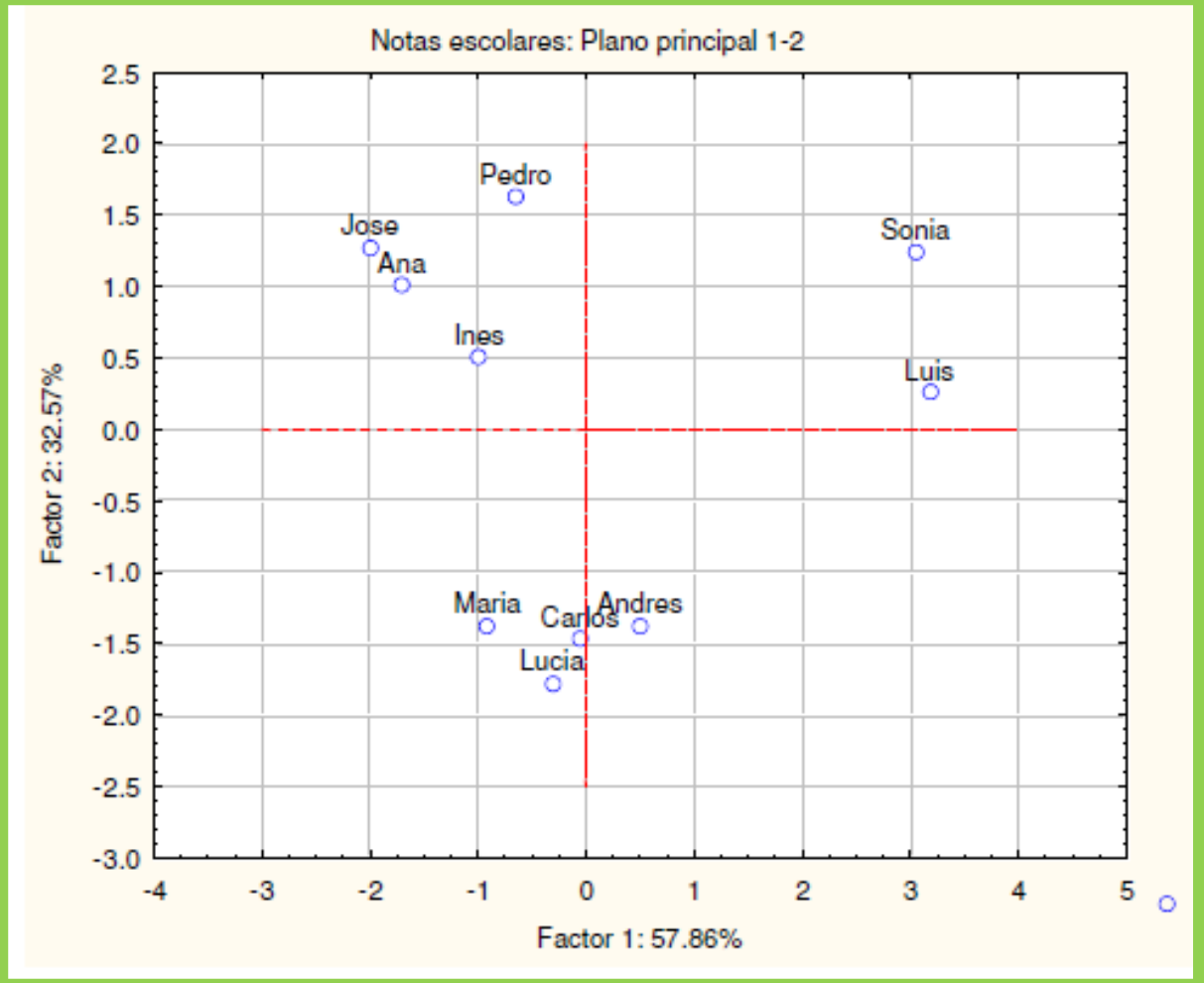
	MATE	CIEN.	ESPA	HIS.	GIM
LUCIA	7.0	6.5	9.2	8.6	8.0
PEDRO	7.5	9.4	7.3	7.0	7.0
INES	7.6	9.2	8.0	8.0	7.5
LUIS	5.0	6.5	6.5	7.0	9.0
ANDRES	6.0	6.0	7.8	8.9	7.3
ANA	7.8	9.6	7.7	8.0	6.5
CARLOS	6.3	6.4	8.2	9.0	7.2
JOSE	7.9	9.7	7.5	8.0	6.0
SONIA	6.0	6.0	6.5	5.5	8.7
MARÍA	6.8	7.2	8.7	9.0	7.0
PROM	6.79	7.65	7.74	7.9	7.42

Matriz de correlación

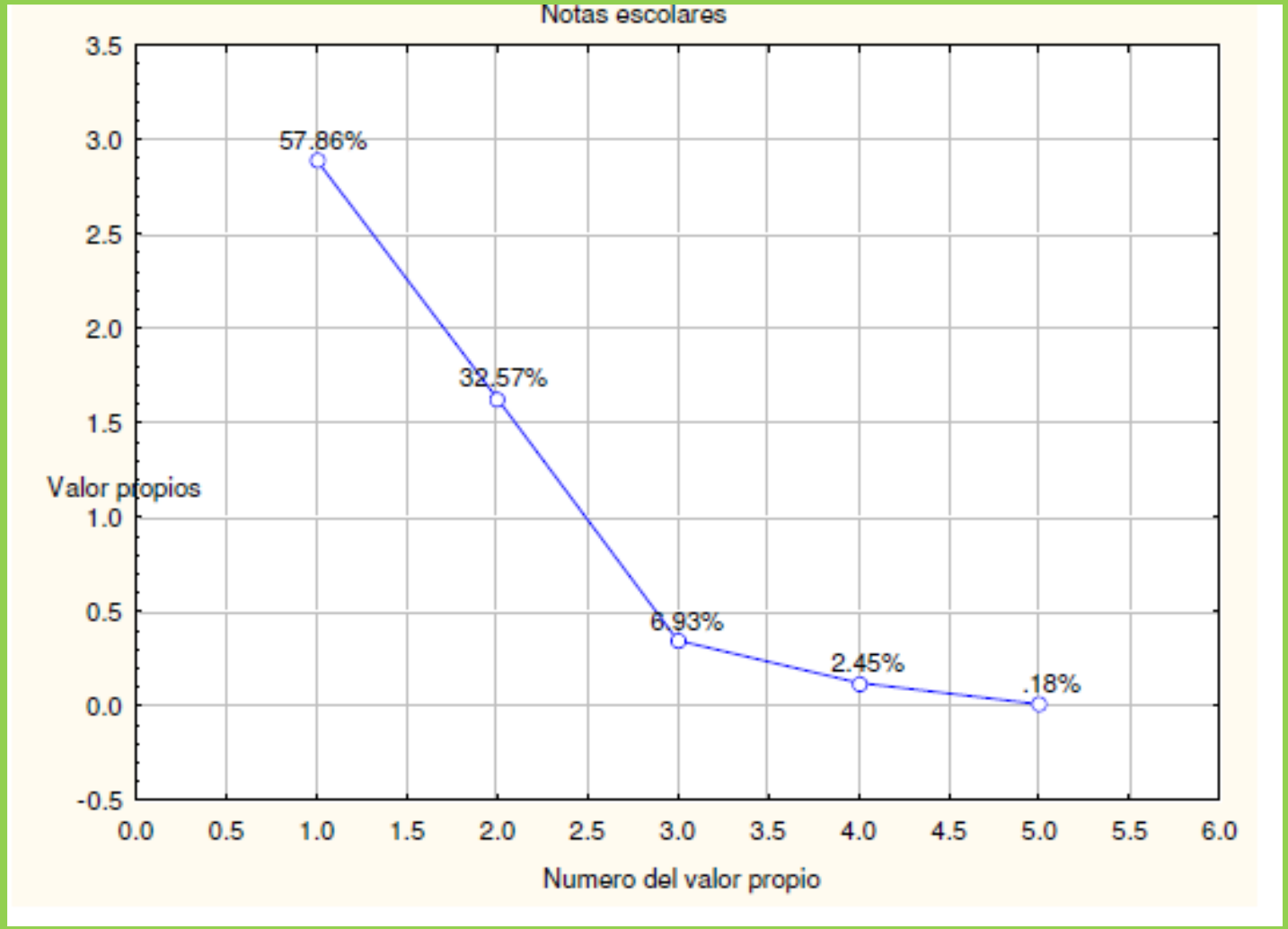
	MATE	CIEN	ESPA	HISTO	GIM
MATE	1	0.85	0.38	0.21	-0.79
CIEN	0.85	1	-0.02	-0.02	-0.69
ESPA	0.38	-0.02	1	0.82	-0.37
HISTO	0.21	-0.02	0.82	1	-0.51
GIM	-0.79	-069	-0.37	-0.51	1

Bi-plot:

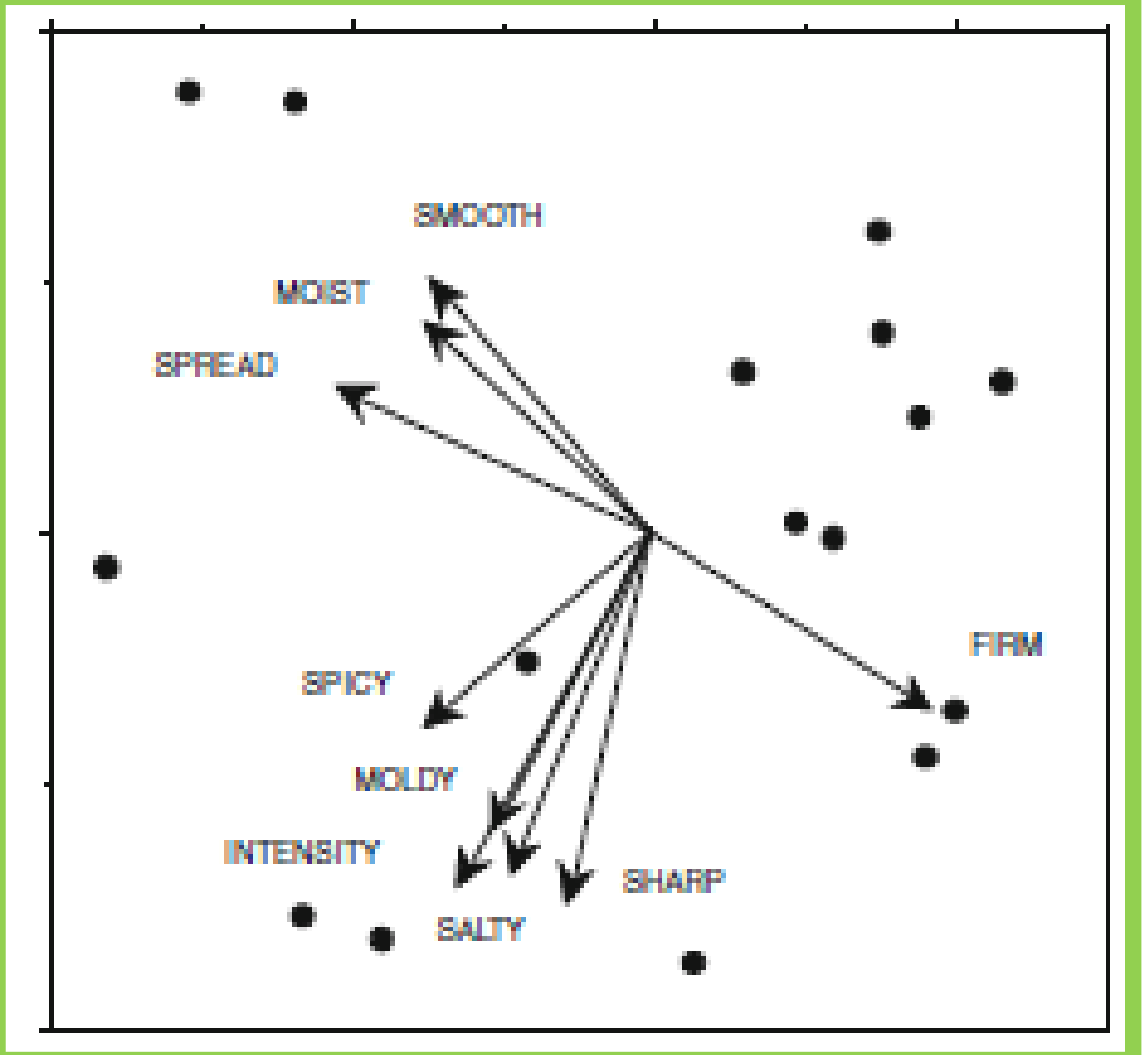

Notas Escolares: Circulo de Correlaciones
Ciencias
Matematica
Ed.Fisica
Espanol
Historia
Factor 2 : 32.57 %
Factor 1 : 57.86%
1.0
0.5
0.0
-0.5
-1.0

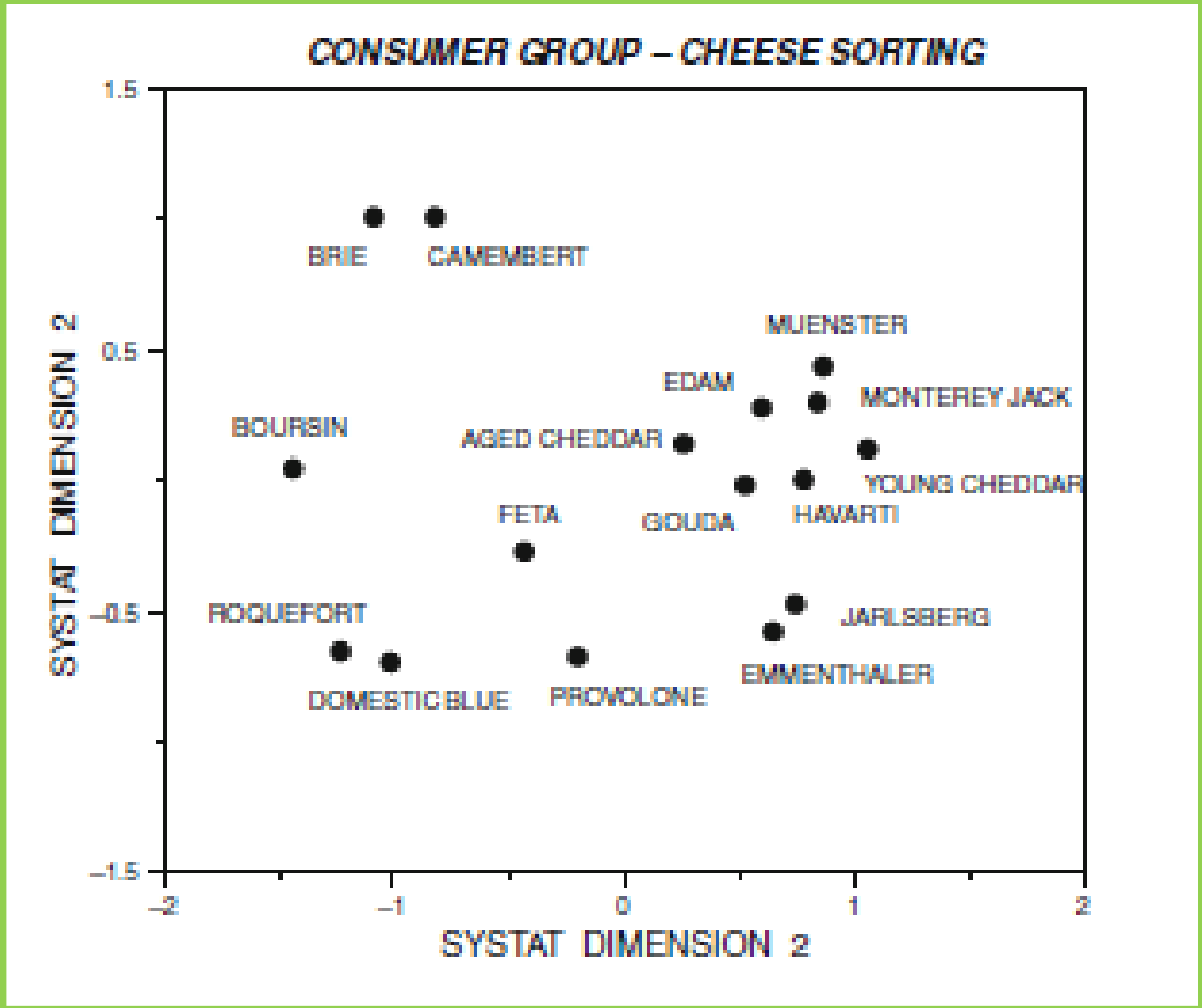

Notas escolares: Plano principal 1-2
Pedro
Jose
Ana
Sonia
Ines
Luis
Maria
Carlos
Andres
Lucia
Factor 2: 32.57%
Factor 1: 57.86%
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-4
-3
-2
-1
0
1
2
3
4
5

¿Cuánto explica cada componente?



Otro ejemplo: análisis sensorial de quesos





Special Issue Article

Advances in Mechanical Engineering

Advances in Mechanical Engineering
2018, Vol. 10(2) 1–16
© The Author(s) 2018
DOI: 10.1177/1687814018755519
journals.sagepub.com/home/ade
SAGE

Machine learning techniques for quality control in high conformance manufacturing environment

Carlos A Escobar[1,2] and Ruben Morales-Menendez[2]

Explainable Artificial Intelligence for Predictive Maintenance Applications

Stephan Matzka
School of Engineering - Technology and Life
Hochschule für Technik und Wirtschaft Berlin
12459 Berlin, Germany
stephan.matzka@htw-berlin.de

Since real predictive maintenance datasets are generally difficult to obtain and in particular difficult to publish, we present and provide a synthetic dataset [5] that reflects real predictive maintenance data encountered in industry to the best of our knowledge and experience. The dataset consists of 10,000 data points stored as rows with 6 features in columns

FACULTY & RESEARCH

FACULTY RESEARCH FEATURED TOPICS

... → Faculty & Research

Publications

OCTOBER 2012 ARTICLE HARVARD BUSINESS REVIEW

Data Scientist: The Sexiest Job of the 21st Century

Format: Print

Email Print Share Recommend 114 Sha

ABSTRACT

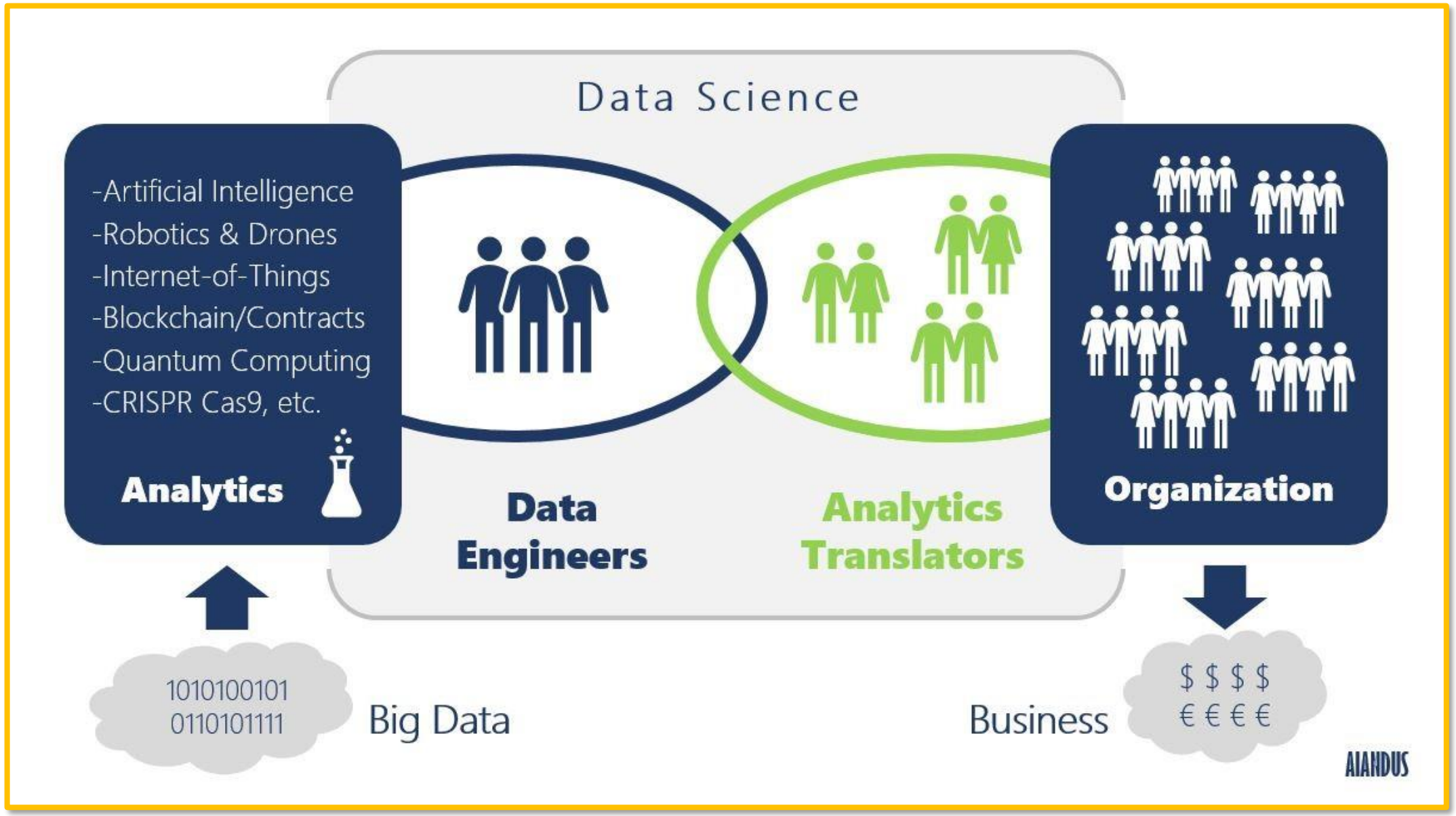
Key to the effective use of big data are the analytical professionals known as "data scientists," who can both manipulate large and unstructured data sources and create insights from them. Data scientists are difficult to hire and retain, but their skills will be necessary to any organization wishing to profit from bi data.

+

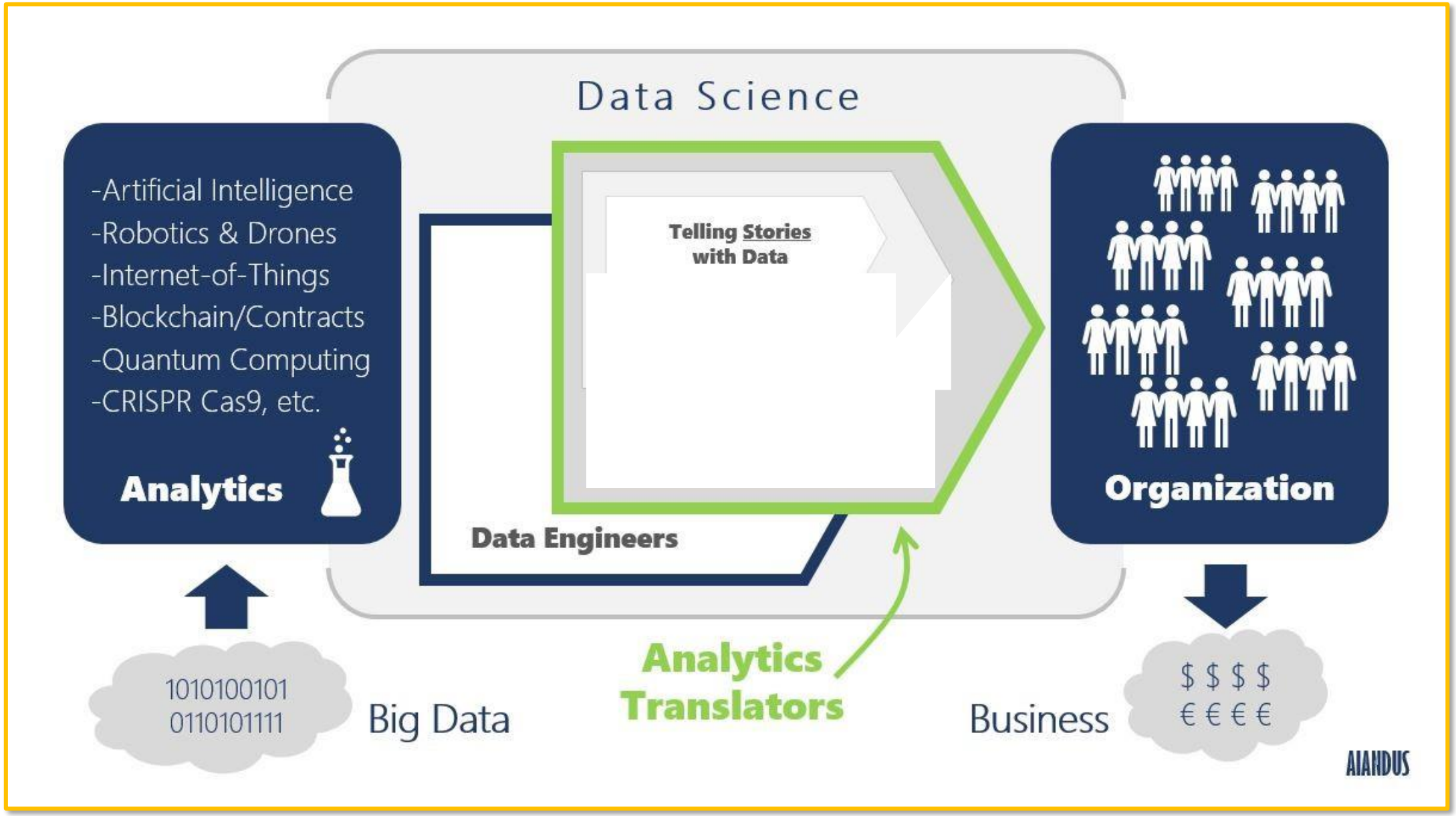
nerd + sexy

+

nerd + sexy



The New Sexiest Job of the 21st Century
Kees Groeneveld


Data Science
-Artificial Intelligence
-Robotics & Drones
-Internet-of-Things
-Blockchain/Contracts
-Quantum Computing
-CRISPR Cas9, etc.
Analytics
Telling Stories with Data
Data Engineers
Organization
1010100101
0110101111
Big Data
Analytics Translators
Business
$ $ $ $
€ € € €
AIANDUS