

# **Machine Learning aplicado a las Ciencias Sociales**

## **Clase 1. Introducción y análisis no supervisado 1 (PCA)**



# ¿Qué es Machine Learning?

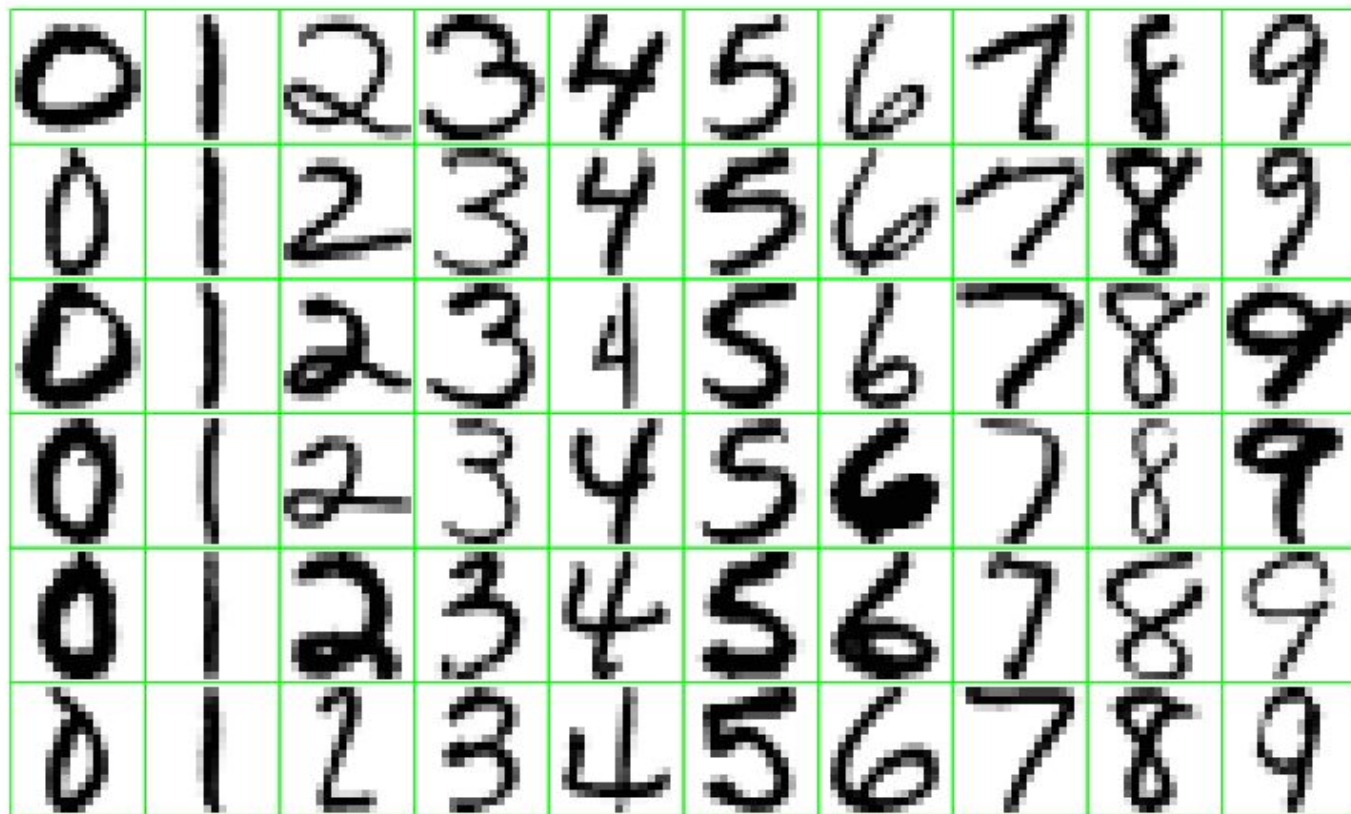


# Machine Learning

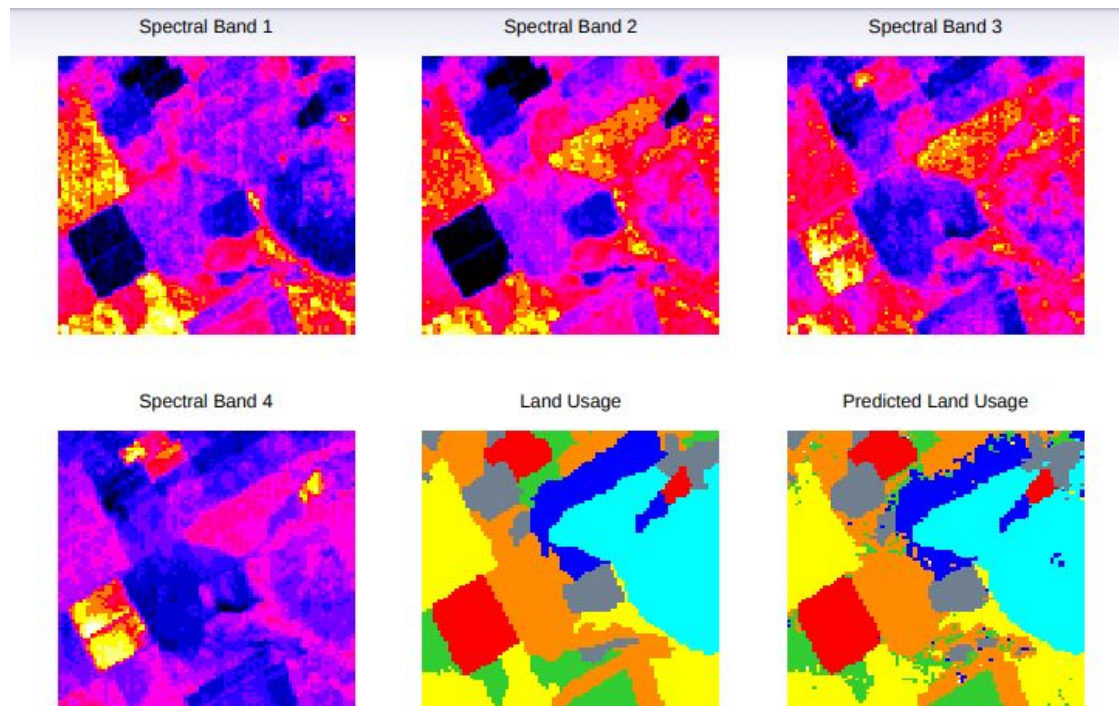
- ¿Podría una computadora ir más allá de “lo que sea que sepamos decirle que haga” y realmente “aprender” por su cuenta como realizar una determinada tarea?
- ¿Podría ser posible el aprendizaje automático de estas reglas a partir de los datos?



# Reconocimiento de dígitos



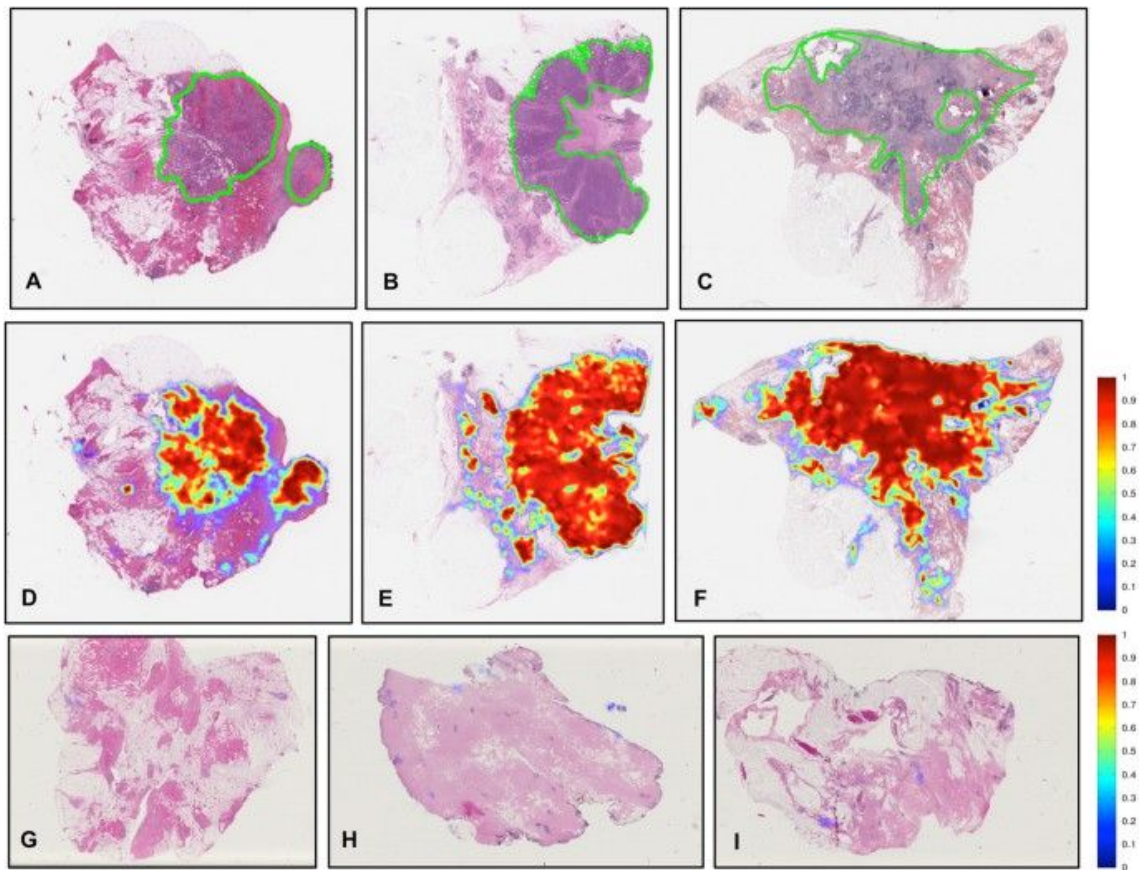
# Clasificación de imágenes satelitales



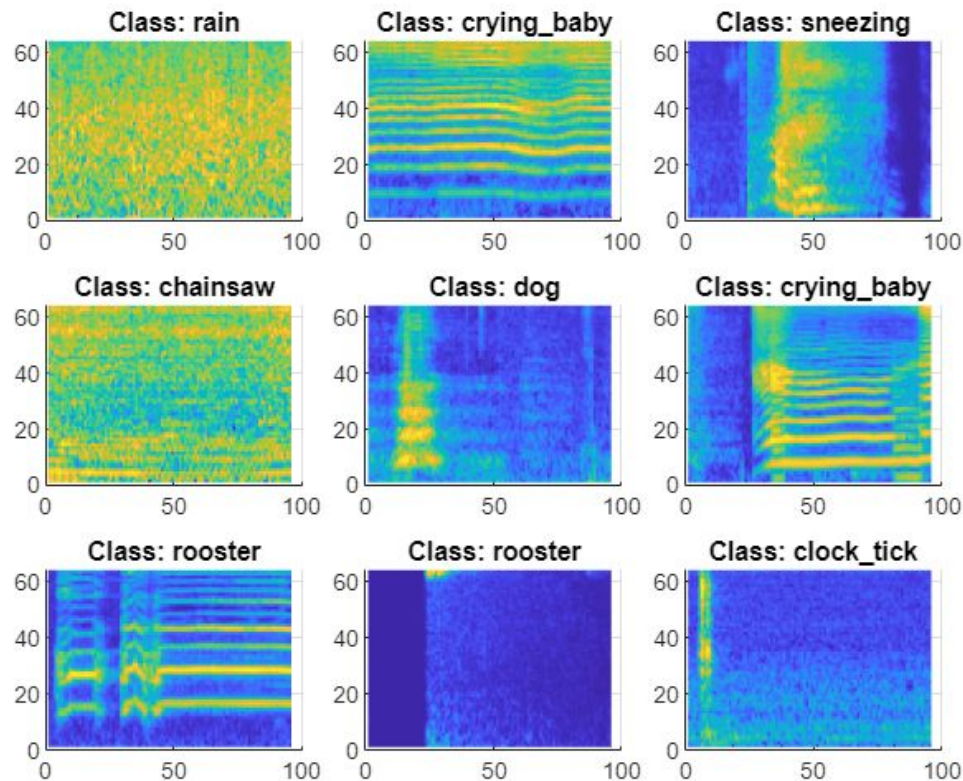
*Usage  $\in \{\text{red soil, cotton, vegetation stubble, mixture, gray soil, damp gray soil}\}$*



# Detección de células cancerígenas

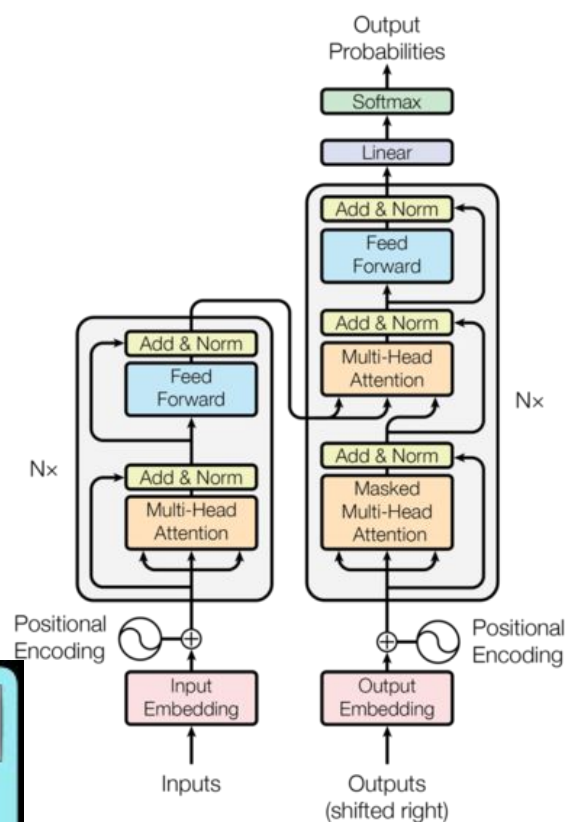
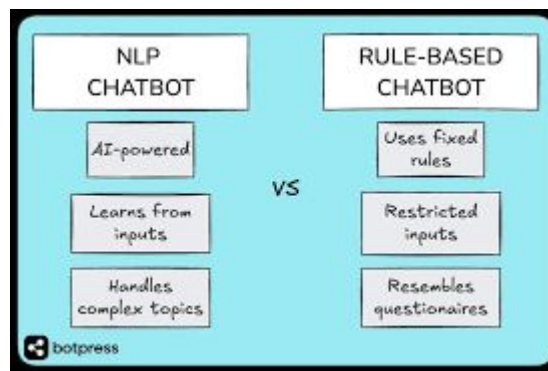
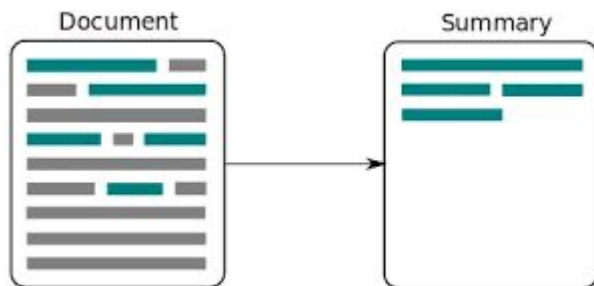
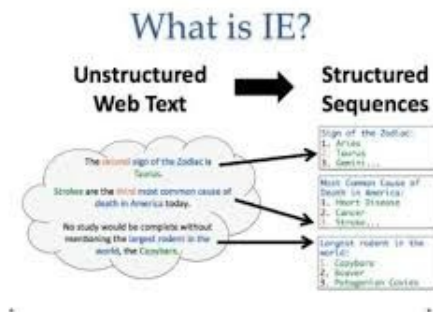
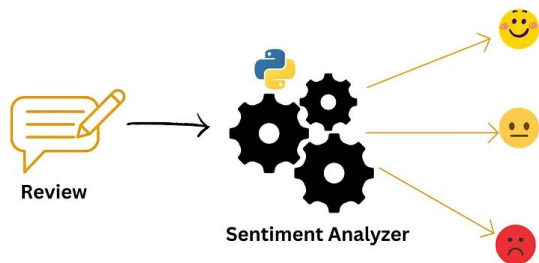


# Identificación de objetos / clasificación de audio





# Procesamiento de lenguaje natural

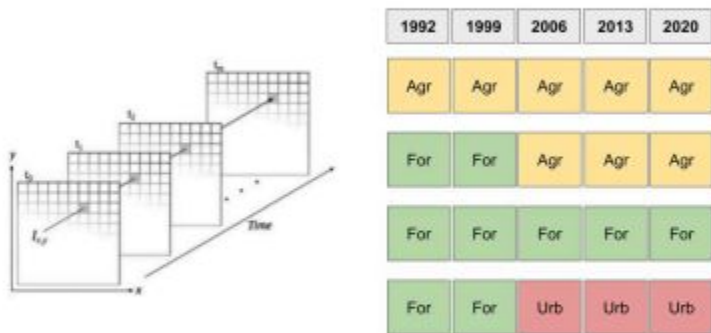




# Detección/predicción de trayectorias diversas

ESQUEMA 1

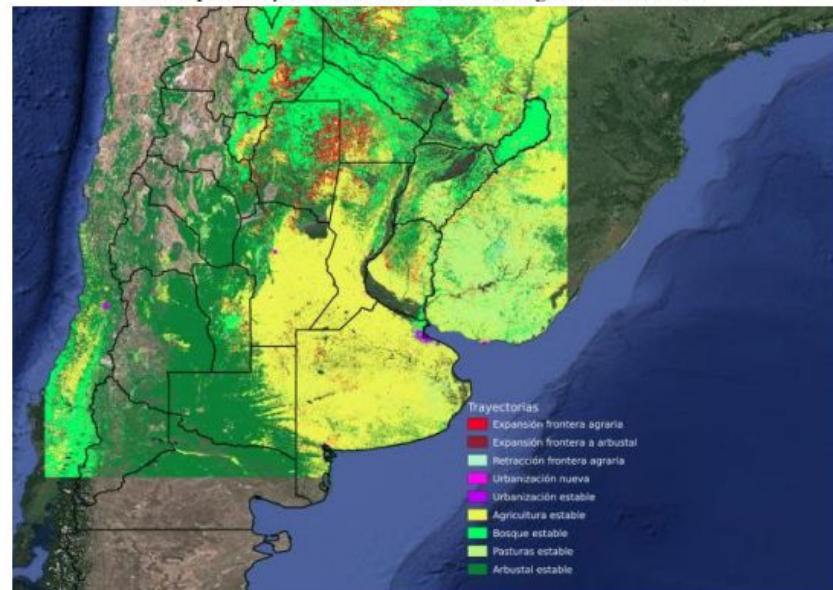
Ejemplos de trayectorias de uso del suelo a nivel pixel



Fuente: elaboración propia en base Maciel et al (2019) y a datos ESA-CCI-LC.

MAPA 1

Principales trayectorias de uso del suelo, Argentina 1992-2020



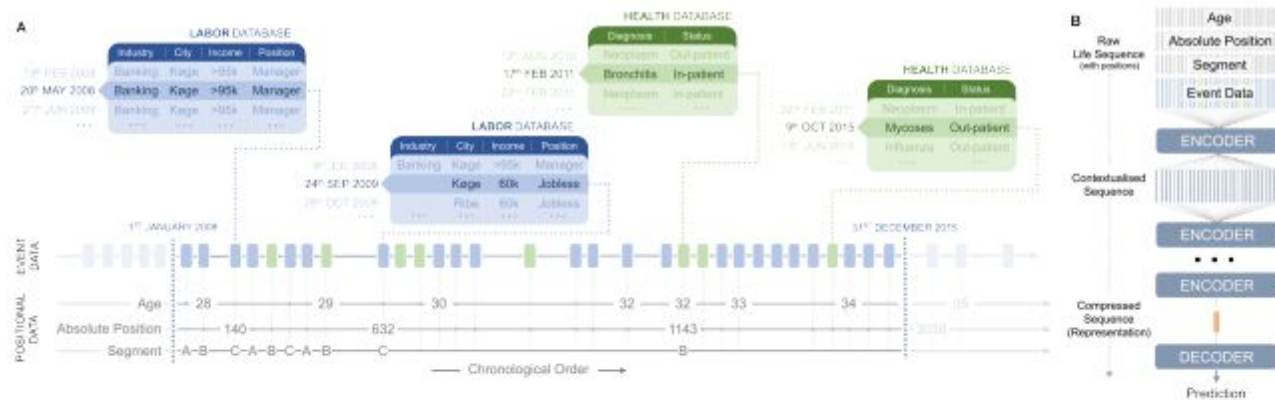
Fuente: elaboración propia en base a datos ESA-CCI-LC.



factor-data  
EIDAEs\_UNSAM

<https://www.geograficando.fahce.unlp.edu.ar/article/view/geoe130>

# Detección/predicción de trayectorias diversas



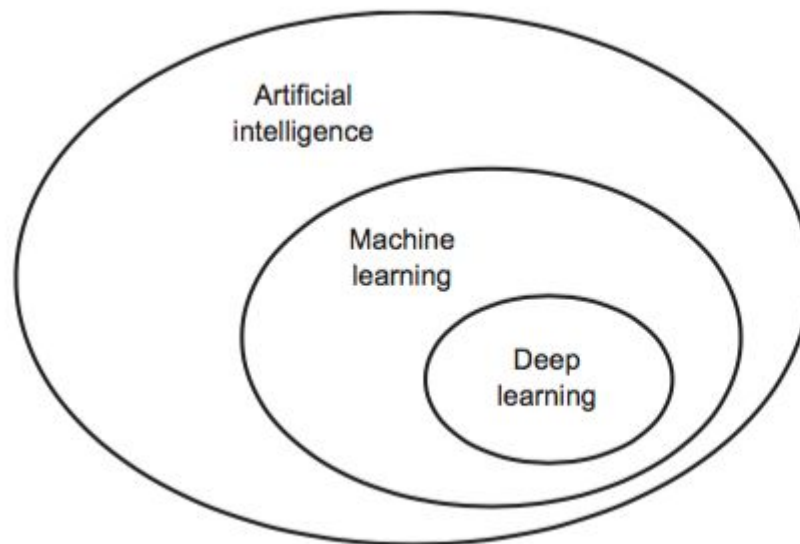
**Figure 1: A schematic individual-level data representation for the life2vec model.** (A) We organize socio-economic and health data from the Danish national registers from 1st January 2008 until 31st December 2015 into a single chronologically ordered *life-sequence*. Each database entry becomes an event in the sequence, where an event has associated positional and contextual data. The contextual data include variables associated with the entry (e.g., industry, city, income, job type). The positional data includes the person's age (expressed in full years), absolute position (number of days since 1st January 2008), and segment (alternating sequence of three elements). The raw life-sequence is then passed to the model described in panel (B). The model consists of multiple stacked encoders. The first encoder combines contextual and positional information to produce a contextual representation of each life event. The following encoders output deep contextual representations of each life event (considering the overall content of the life-sequence). The final encoder layer fuses the representations of life-events to produce the representation of a life-sequence. The decoder uses the latter to make predictions.

<https://arxiv.org/pdf/2306.03009>



factor-data  
EIDAES\_UNSAM

# Machine Learning



**Figure 1.1** Artificial intelligence, machine learning, and deep learning

[Chollet, 2017]



factor-data  
EIDAEs\_UNSAM

<https://www.manning.com/books/deep-learning-with-python>

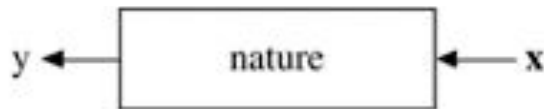
# Un poco de epistemología... ¿qué es un modelo?

- Básicamente: una manera de proponer hipótesis sobre la forma en que se combinan variables
- En general vamos a estar tratando de general modelos de la forma  $Y = f(x) + e$
- Todo el problema es estimar  $f(X)$ , es decir, de qué forma(s) se combinan las  $X$  para generar un output
- Una posibilidad es suponer que  $Y$  es una combinación lineal de las  $X$



# Las dos culturas... (Breiman, 2001)

“Todos los modelos son equivocados, algunos son útiles” George Box

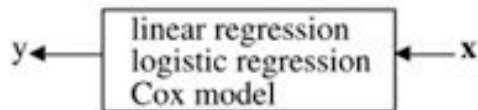


- El mundo como productor de outputs -y - en base a features -X -
- Problemas: ¿cuál es la manera en que el mundo produce resultados?
- Una forma común es asumir que los datos son generados por
- extracciones independientes de output = f (predictores, ruido, parámetros)



# Las dos culturas... (Breiman, 2001)

## Modelado estadístico

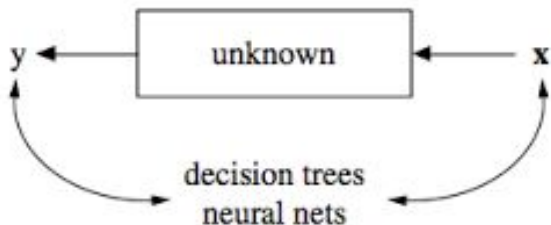


- Énfasis en  $f(x)$ . El modelo se postula en base a supuestos sobre  $f(x)$
- Conocimiento acumulado, teoría, diseño de experimentos
- Los parámetros son estimados con los datos y luego se realizan las predicciones.
- Evaluación del modelo: estimadores insesgados, robustos, mínima varianza



# Las dos culturas... (Breiman, 2001)

## Modelado algorítmico



- Énfasis en  $\hat{y}$
- El enfoque es encontrar una función  $f(x)$  -un algoritmo- que opera sobre las  $x$  para predecir las  $y$ .
- El modelo se “aprende” de los datos
- Evaluación del modelo: performance predictiva



# Los datos, los algoritmos y la ciencia social

CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

## THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE



*“Olvídense de la teoría del comportamiento humano, desde la lingüística hasta la sociología. Olvídense de las taxonomías, de la ontología y de la psicología. ¿Quién sabe por qué la gente hace lo que hace?. El punto es que lo hacen y que podemos trackear y medir eso que hacen con una precisión sin precedentes. Con suficientes datos, los números hablan por sí mismos.”*



factor-data  
EIDAEs\_UNSAM

<https://www.wired.com/2008/06/pb-theory/>

# Las dos culturas... (Breiman, 2001)

## Regresión lineal

$$y_i = x_i \beta + u_i, \quad i = 1, \dots, n,$$

$$\hat{y}_i = x_i \hat{\beta} + u_i$$



# Las dos culturas... (Breiman, 2001)

## Regresión lineal - Visión desde la estadística

$$y_i = x_i \beta + u_i, \quad i = 1, \dots, n,$$

$$\hat{y}_i = x_i \hat{\beta} + u_i$$

Ex-ante

Modelo ( $f(x)$ ) dado  $\Rightarrow$  estimarlo

- Problema: inferencia sobre  $\beta$
- ¿Cómo? Con un estimador
$$\hat{\beta} = (x_i^T x_i)^{-1} (x_i^T y_i)$$
- Insesgadez:  $E(\hat{\beta}) = \beta$
- Varianza mínima...
- Gauss-Markov, residuos, coeficientes, etc...



# Las dos culturas... (Breiman, 2001)

## Regresión lineal - Visión desde ML

$$y_i = x_i \beta + u_i, \quad i = 1, \dots, n,$$

$$\hat{y}_i = x_i \hat{\beta} + u_i$$

Ex-post

Modelo se “aprende” de los datos =>  
ECM “out of sample”

- Problema: obtener las mejores predicciones posibles  $\hat{y}_i \approx y_i$
- ¿Cómo? Minimizando una función de costo

$$ECM = \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

# Explicación (o comprensión) vs. predicción

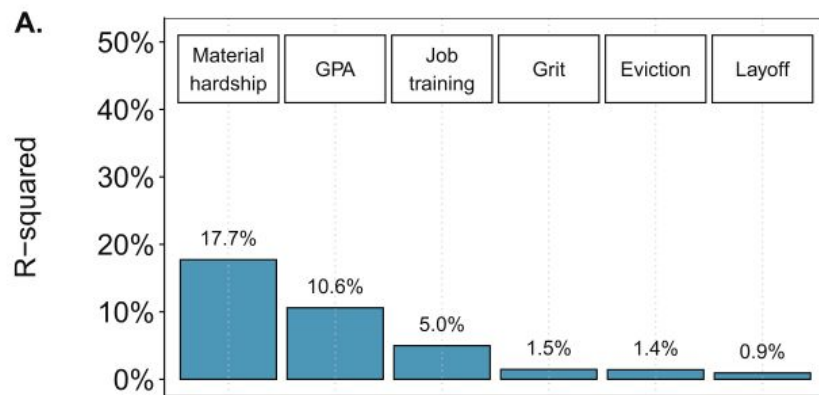
- Dicotomía presente desde la “disputa del método”
- Se asume la necesidad de elegir uno u otro
- Concepciones equivocadas al respecto
- Dos momentos complementarios de un proceso de investigación



# ¿Explicar, comprender, predecir o todo junto?

Tres virtudes del uso de la predicción

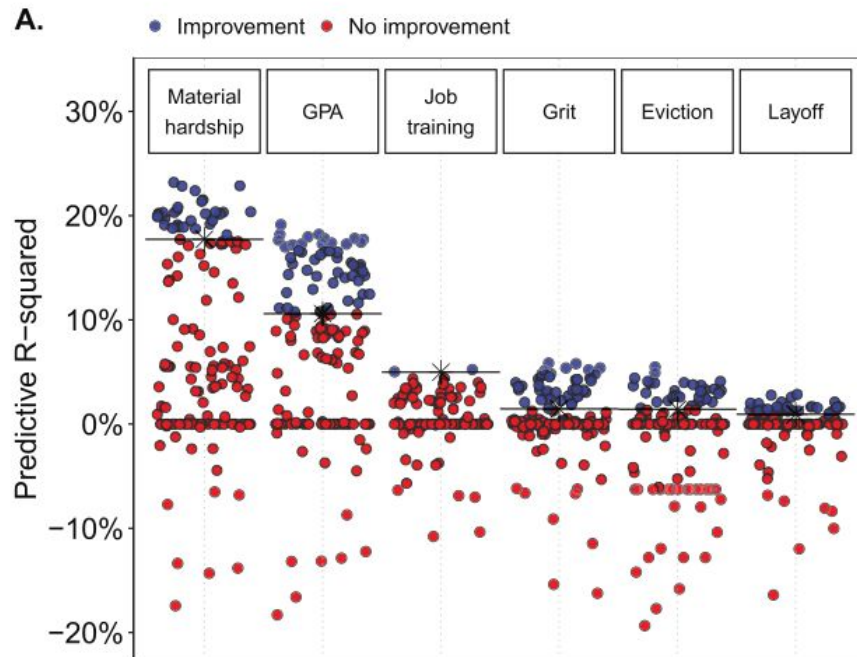
- Información sobre el ajuste
- Base de comparación a lo largo de diferentes modelos
- Información para interpretar modelos complejos



# ¿Explicar, comprender, predecir o todo junto?

Tres virtudes del uso de la predicción

- Información sobre el ajuste
- Base de comparación a lo largo de diferentes modelos
- Información para interpretar modelos complejos





# ¿Explicar, comprender, predecir o todo junto?

## Tres virtudes del uso de la predicción

- Información sobre el ajuste
- Base de comparación a lo largo de diferentes modelos
- Información para interpretar modelos complejos

Regresión logística para analizar la relación presencia de protestas (resolución 125) y entre los departamentos:

Donde

$Y_i$  es la presencia de conflictos en el departamento  $i$

$X_{1i}$  es la superficie con soja en el depto  $i$

$X_{2i}$  es el nivel de mecanización en el depto  $i$

$X_{3i}$  la proporción de pequeños productores capitalizados en el depto  $i$

$X_{4i}$  la proporción de explotaciones con trabajadores asalariados en el depto  $i$



# ¿Explicar, comprender, predecir o todo junto?

Tres virtudes del uso de la predicción

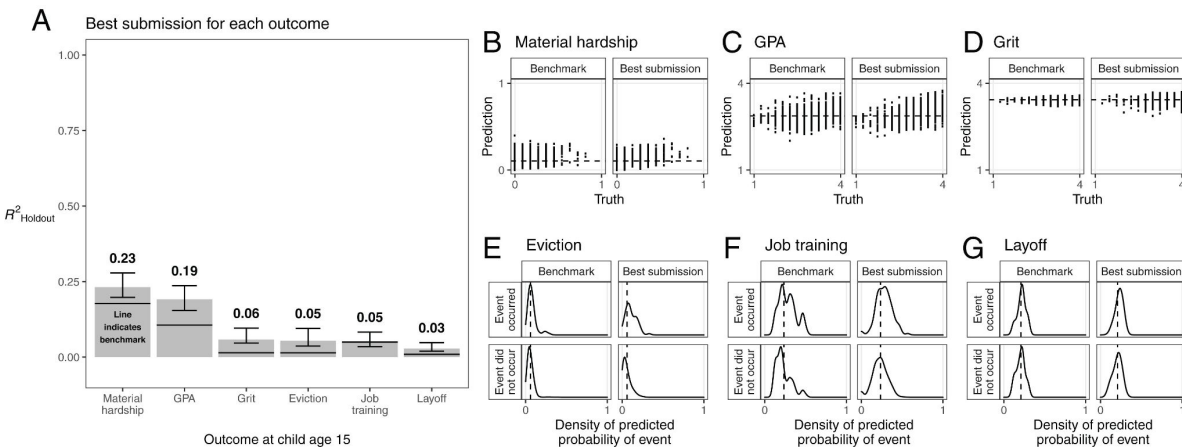
- Información sobre el ajuste
- Base de comparación a lo largo de diferentes modelos
- Información para interpretar modelos complejos

	<b>Peq. producción campesina</b>	<b>Gran producción extensiva</b>	<b>Peq. producción capitalista</b>	<b>Prod. capitalista desarrollada</b>
% Sup. con soja	Constante	Constante	Constante	Constante
Índice de <u>Mecaniz</u>	Baja	Baja	Alta	Alta
% <u>Peq. prod. T1</u>	Baja	Baja	Alta	Baja
% <u>EAP's con no fliares.</u>	Baja	Alta	Baja	Alta

**Nota:** constante (promedio); Baja (percentil 5); Alta (percentil 95)



# ¿Explicar, comprender, predecir o todo junto?



RESEARCH ARTICLE | SOCIAL SCIENCES | 8

f x in

## Measuring the predictability of life outcomes with a scientific mass collaboration

Matthew I. Salganik, Ian Lundberg, Alexander T. Kindel, and Sara McLanahan [Authors Info & Affiliations](#)

Contributed by Sara McLanahan, January 24, 2020 (sent for review October 1, 2019; reviewed by Senthil Mullainathan and Brian Uzzi)

March 30, 2020 117 (15) 8398-8403 <https://doi.org/10.1073/pnas.1915006117>

THIS ARTICLE HAS BEEN UPDATED

THIS ARTICLE HAS BEEN CORRECTED

VIEW RELATED CONTENT

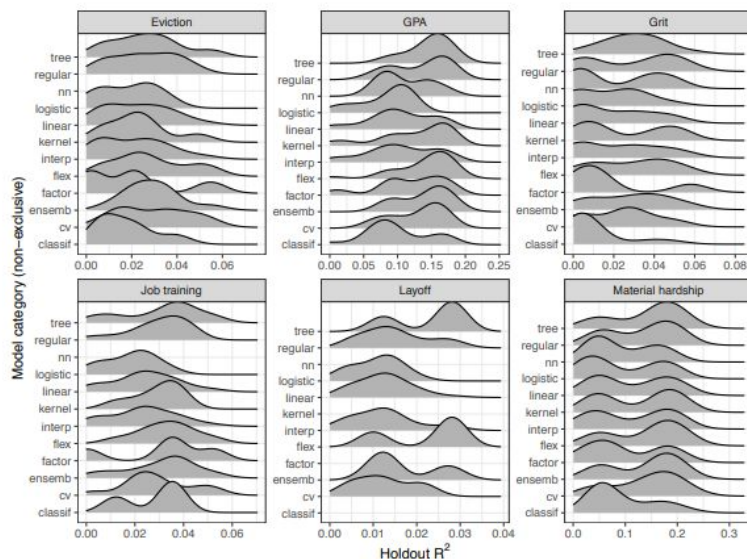
116,121 | 157

PDF/EPUB

### Significance

Hundreds of researchers attempted to predict six life outcomes, such as a child's grade point average and whether a family would be evicted from their home. These researchers used machine-learning methods optimized for prediction, and they drew on a vast

# ¿Explicar, comprender, predecir o todo junto?



**Fig. S18. Performance of submissions by type of statistical learning procedure.** Each ridge indicates the smoothed density of  $R^2_{\text{Holdout}}$  among qualifying submissions that used procedures of the corresponding type. A missing ridge indicates that a statistical learning procedure of that type was not used in three or more qualifying submissions for that outcome. X-axis scales vary by facet.

## Measuring the predictability of life outcomes with a scientific mass collaboration

Matthew I. Salganik , Jan Lundberg , Alexander T. Kindel , and Sara McLanahan  [Authors Info & Affiliations](#)

Contributed by Sara McLanahan, January 24, 2020 (sent for review October 1, 2019; reviewed by Senthil Mullainathan and Brian Uzzi)

March 30, 2020 | 117 (15) 8398-8403 | <https://doi.org/10.1073/pnas.1915006117>

THIS ARTICLE HAS BEEN UPDATED

THIS ARTICLE HAS BEEN CORRECTED +

VIEW RELATED CONTENT +

 116,121 | 157



PDF/Epub

### Significance

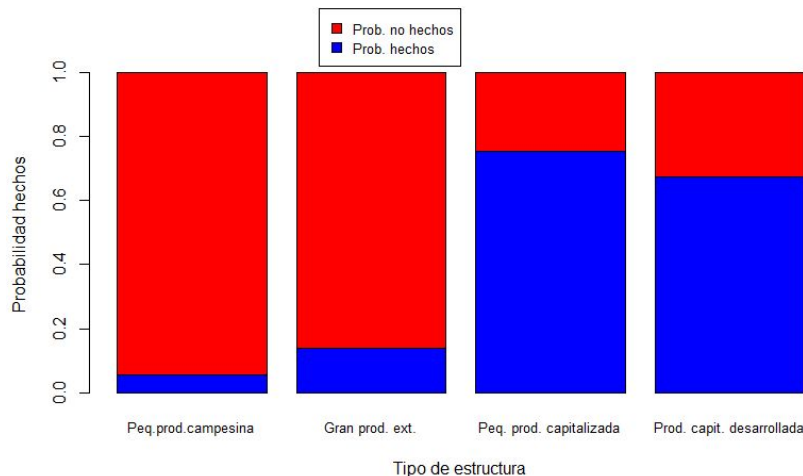
Hundreds of researchers attempted to predict six life outcomes, such as a child's grade point average and whether a family would be evicted from their home. These researchers used machine-learning methods optimized for prediction, and they drew on a vast

# ¿Explicar, comprender, predecir o todo junto?

Tres virtudes del uso de la predicción

- Información sobre el ajuste
- Base de comparación a lo largo de diferentes modelos
- Información para interpretar modelos complejos

Probabilidad de presentar hechos de protesta según tipo ideal de estructura agraria

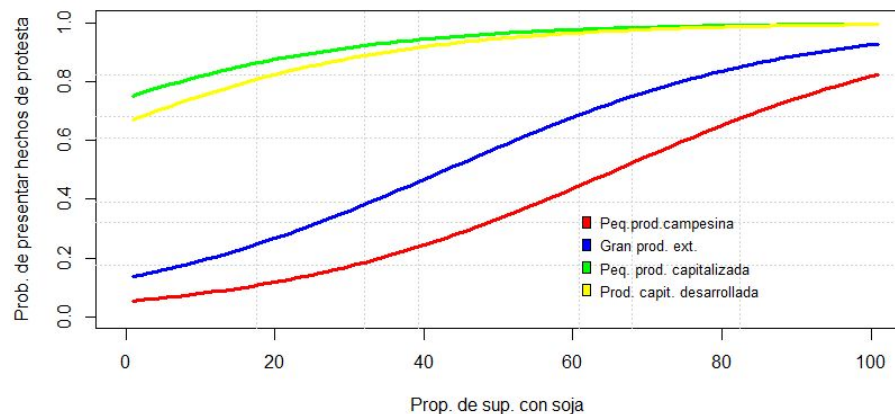


# ¿Explicar, comprender, predecir o todo junto?

Tres virtudes del uso de la predicción

- Información sobre el ajuste
- Base de comparación a lo largo de diferentes modelos
- Información para interpretar modelos complejos

Probabilidad de presentar hechos de protesta condicionada a la superficie sojera



# Bibliografía de consulta

- Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statistical Science, 16(3).  
<https://doi.org/10.1214/ss/1009213726>
- Chollet, F. (2018). Deep learning with Python. Manning Publications Co.
- Rosati, G. (2023). Analizando trayectorias de uso del suelo. Una propuesta de clusterización. Geograficando, 19(1), e130. <https://doi.org/10.24215/2346898Xe130>
- Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., ... McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. Proceedings of the National Academy of Sciences, 117(15), 8398–8403. <https://doi.org/10.1073/pnas.1915006117>





# Bibliografía de consulta

- Savcicens, G., Eliassi-Rad, T., Hansen, L. K., Mortensen, L., Lilleholt, L., Rogers, A., Zettler, I., & Lehmann, S. (2023). Using Sequences of Life-events to Predict Human Lives. *Nature Computational Science*, 4(1), 43–56.  
<https://doi.org/10.1038/s43588-023-00573-5>
- Verhagen, M. D. (2022). A Pragmatist's Guide to Using Prediction in the Social Sciences. *Socius: Sociological Research for a Dynamic World*, 8, 237802312210817.  
<https://doi.org/10.1177/23780231221081702>



# Aprendizaje Automático No Supervisado



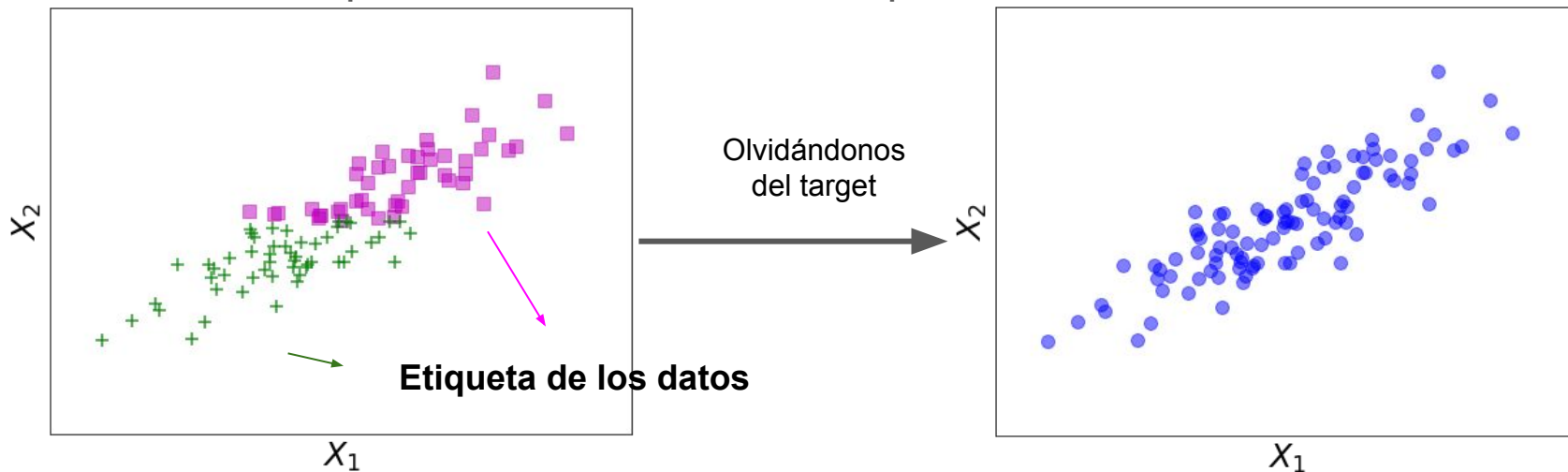
# Tipos de problemas

- **Aprendizaje Supervisado**
  - Variable dependiente => Y, resultado, target
  - Matriz de predictores (p), X, features, variables independientes, etc.
  - Problemas de regresión: Y es cuantitativa
  - Problemas de clasificación: Y es cualitativa
  - Tenemos datos de entrenamiento (conjuntos de  $X_i, Y_i$ ), observaciones
  - Podemos definir una (o varias) métricas para evaluar los modelos
- **Aprendizaje no Supervisado**
  - **No hay variable target (Y)**
  - **Solo hay X**
  - **Es más difícil evaluar qué tan bien funciona el modelo**



# ¿Qué es el aprendizaje no supervisado?

- El estudio o la exploración de cómo están representados datos y qué conclusiones podemos sacar de dicha representación.



**Aprendizaje no supervisado** es todo lo que podemos hacer con solo la representación de los datos en el espacio de features.



# ¿Qué podemos hacer solo con las features?

- Clustering (todavía no...): agrupar casos “parecidos”, que tengan una combinación de valores **similar** en las variables independientes (espacio de features) (próxima clase). Buscamos subconjuntos de datos muy parecidos entre sí.
- Reducción de la dimensionalidad (clase de hoy): encontrar combinaciones de features que reemplacen a los originales para reducir la dimensión del problema. “Reescribir” los datos en una menor cantidad de variables (o dimensiones)

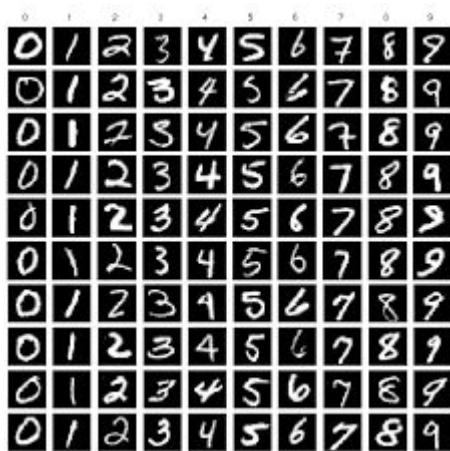


# ¿Por qué reducir la dimensionalidad del problema?

- ¿Todos los features aportan información relevante? ¿Es necesario trabajar con todos?
- Reduciendo la dimensión podemos:
  - Visualizar los datos en el espacio de dimensión reducida, más fácil de interpretar
  - Comprimir la información: nos permite separar la señal del ruido (abstracción)
  - Insumo para clustering: instancias parecidas en un espacio multidimensional son más parecidas en un espacio reducido (emergencia de estructuras).

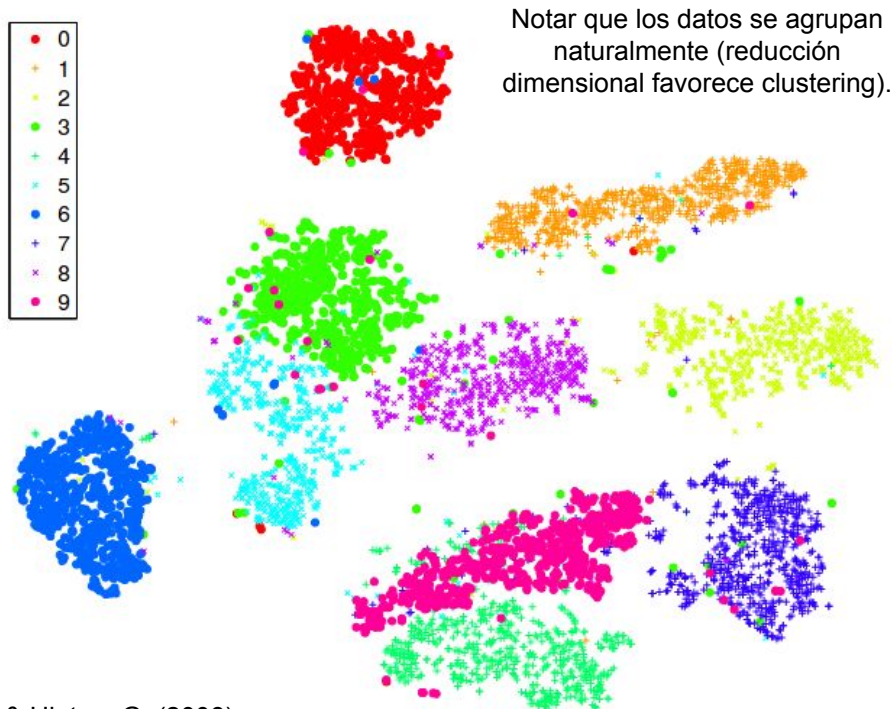


# ¿Por qué reducir la dimensionalidad del problema?



**MNIST dataset**  
Datos representados en  
un espacio de **748**  
**píxeles**.

Visualización de  
datos  
multidimensionales  
en 2D



Notar que los datos se agrupan  
naturalmente (reducción  
dimensional favorece clustering).



# ¿Por qué reducir la dimensionalidad del problema?

Compresión (con pérdida) de la información: separación de la señal del ruido.

Con 100 dimensiones ( $\leq 4096$ ) ya logramos una reproducción fiel de la imagen original.

+ dimensiones del espacio reducido



# componentes principales: 2



# componentes principales: 10



# componentes principales: 25



# componentes principales: 50



# componentes principales: 100



Datos representados en un espacio de **4096 píxeles**.



factor-data  
EIDAES\_UNSAM

Imagen reconstruida de un espacio de dimensión reducida

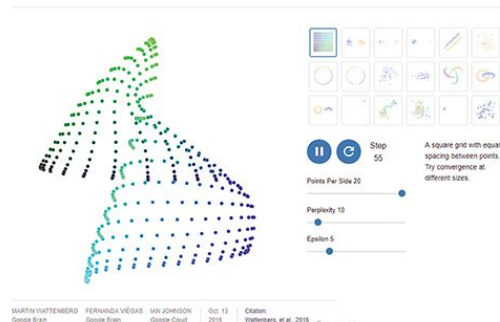
# Algunos métodos de reducción dimensional

Doit

ABOUT PRIZE SUBMIT

## How to Use t-SNE Effectively

Although extremely useful for visualizing high-dimensional data, t-SNE plots can sometimes be mysterious or misleading. By exploring how it behaves in simple cases, we can learn to use it more effectively.



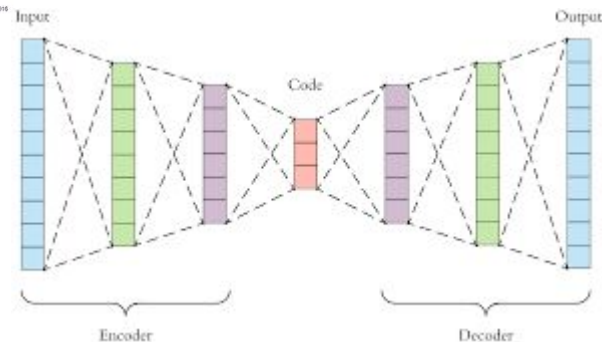
MARTIN WATTEBERG, FERNANDA VEDIS, VIN JONSSON, Oct. 13, 2016. Citation: Watteberg, et al., 2016



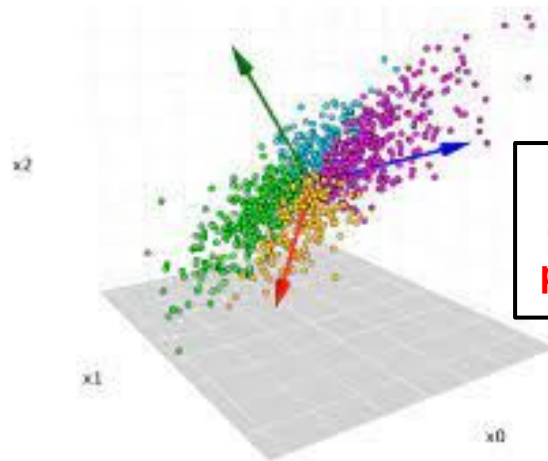
factor-data  
EIDAES\_UNSAM

$$\begin{matrix} W \\ \begin{bmatrix} \square & \square \\ \square & \square \\ \square & \square \\ \square & \square \end{bmatrix} \end{matrix} \times \begin{matrix} H \\ \begin{bmatrix} \square & \square & \square & \square & \square & \square & \square & \square \end{bmatrix} \end{matrix} \approx \begin{matrix} V \\ \begin{bmatrix} \square & \square & \square & \square & \square & \square & \square & \square \\ \square & \square & \square & \square & \square & \square & \square & \square \\ \square & \square & \square & \square & \square & \square & \square & \square \\ \square & \square & \square & \square & \square & \square & \square & \square \end{bmatrix} \end{matrix}$$

**Non-negative matrix factorization (NMF)**



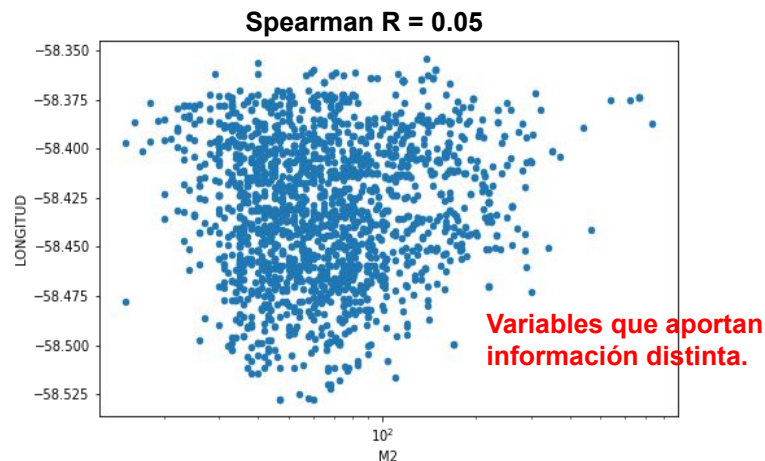
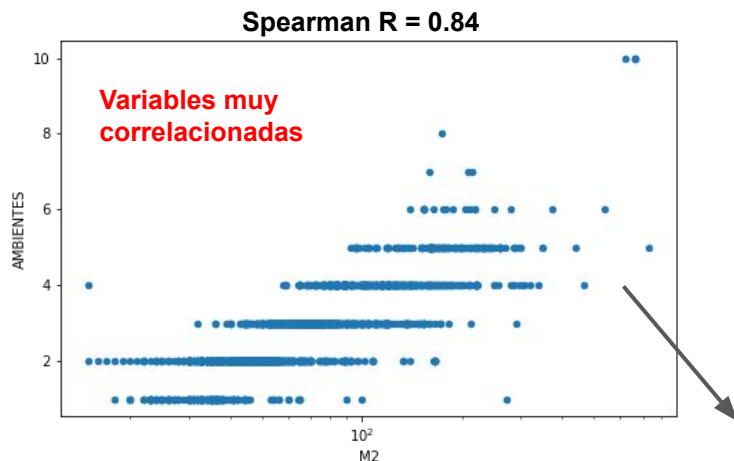
**Autoencoders**



**Descomposición  
en componentes  
principales (PCA)**

# Análisis de componentes principales (PCA)

Problema: muchas veces hay variables que contienen prácticamente la misma información que otras (están muy correlacionadas entre sí), por lo que agregan una dimensión más al problema sin aportar muchas más información.



Una opción para solucionar el problema sería tirar una de las dimensiones por redundante, sin embargo ¿estamos seguros que toda la información que tiramos no es necesaria? ¿Hay alguna otra forma más inteligente de tirar dimensiones?

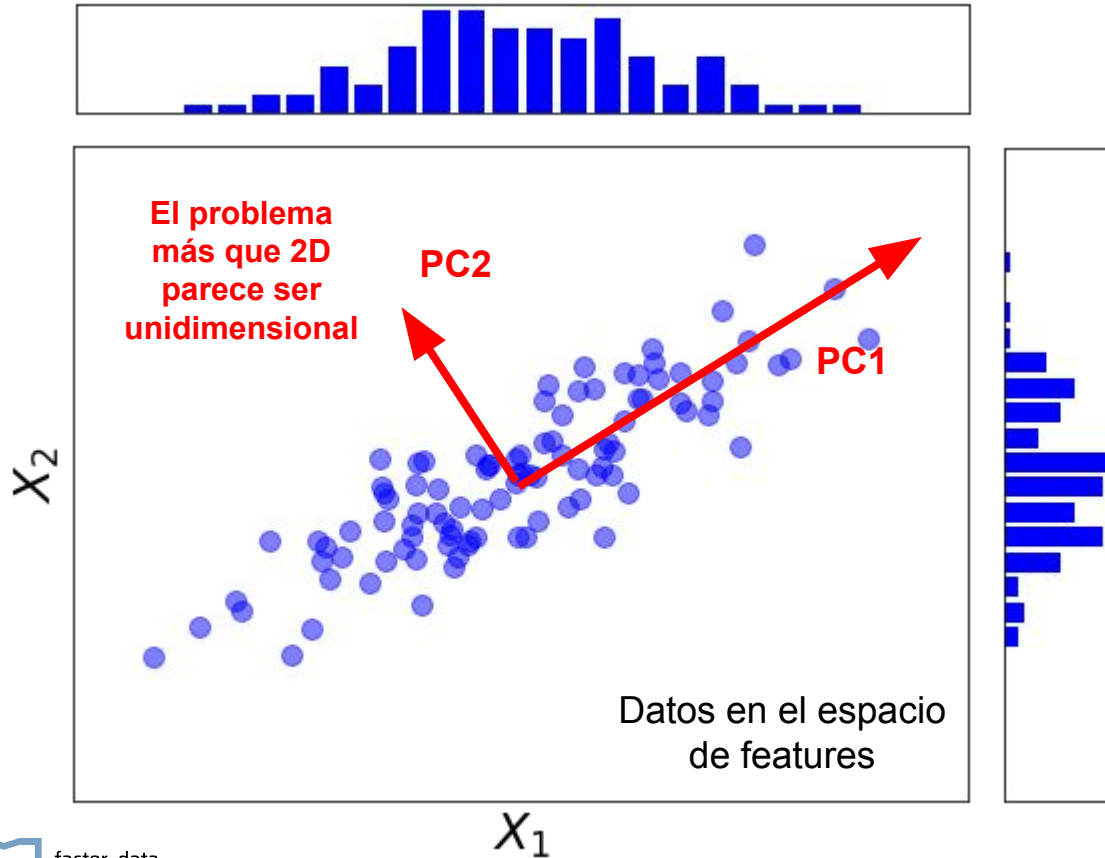


## Esquema general

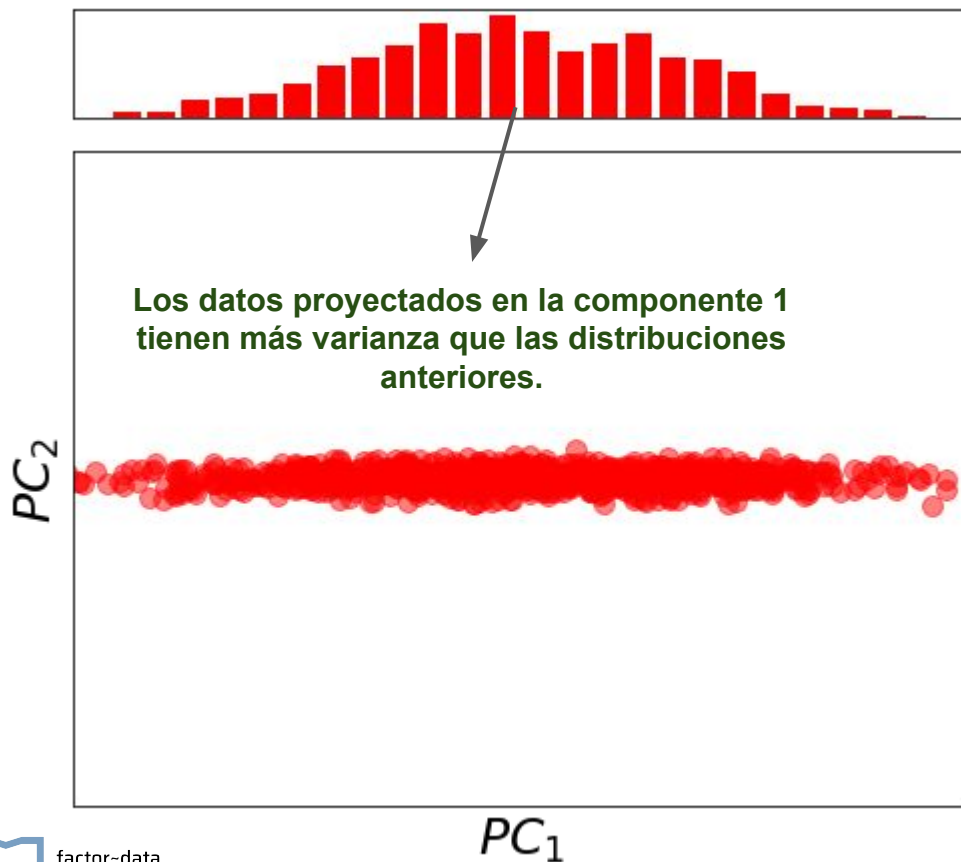
Podemos ver cómo se distribuyen los datos proyectando sobre cada uno de los features.

De las distribuciones podemos calcular su varianza, como una medida de qué tan dispersos están los datos en esa dirección.

Sin embargo, notamos direcciones (combinación lineal de features) donde los datos parecen variar más.



## Esquema general



Proyectando en esas nuevas direcciones (ortogonales) volvemos a ver la distribución de los datos proyectadas en las mismas.

La componente 2 tiene mucha menos variación (parece ser más ruido que información importante)

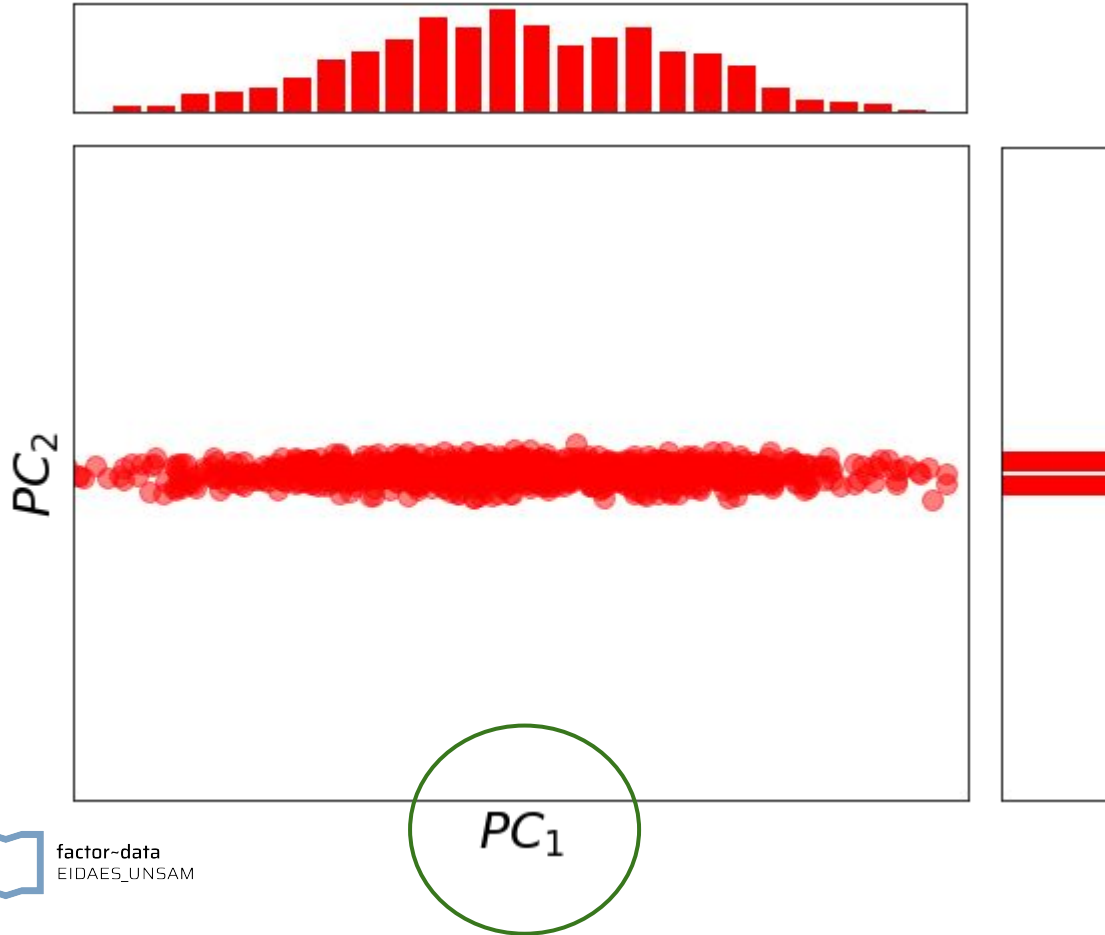


## Esquema general

Notar que si nos olvidamos de la componente 2, no perdemos tanta información como si hubiésemos tirado alguno de los features originales.

**Podemos reducir nuestro problema de 2 dimensiones a una sola.**

Las direcciones que se llevan la mayor cantidad de varianza de los datos se llaman **componentes principales**

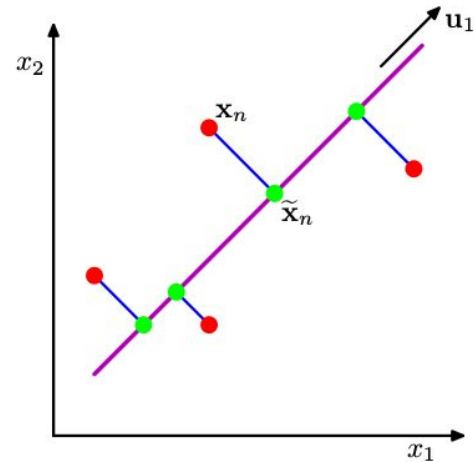


# PCA: descripción matemática

Dado un conjunto  $\{x_n\}$  de  $N$  datos en un espacio de dimensión  $D$  (cada  $x_n$  es un vector en el espacio de  $D$  dimensiones).

Las  $M < D$  componentes principales son:

- las  $M$  direcciones que maximizan la varianza de las proyecciones en ese subespacio,
- o, equivalentemente, las  $M$  direcciones que minimizan el error en la proyección (figura).



# PCA: descripción matemática

No nos vamos a detener en detalle en los diferentes algoritmos que se utilizan para estimar los componentes. Solo diremos que uno de ellos lleva a que podemos obtener los componentes principales a partir de la descomposición de la matriz de covarianza  $\mathbf{S}$ :

$$\bar{\bar{S}} \bar{u} = \lambda \bar{u}$$

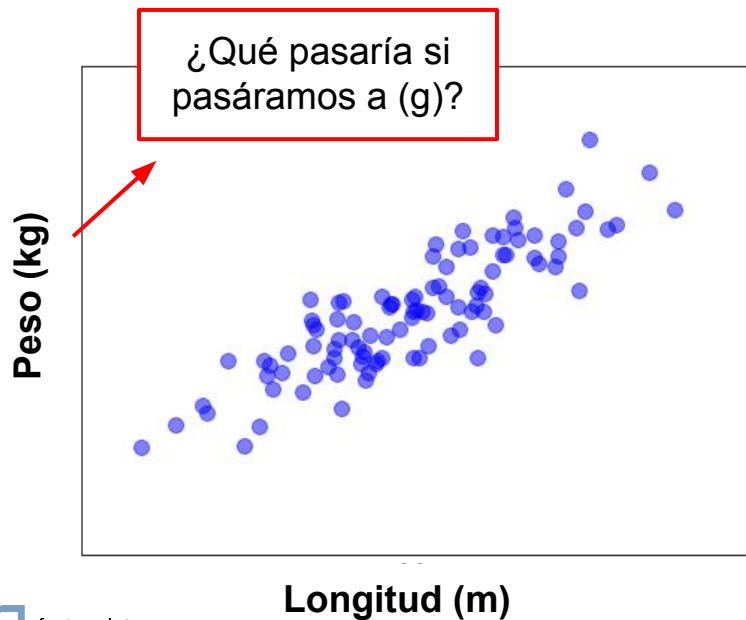
- Las componentes principales son ortogonales.
- Ordenando las componentes desde el autovalor más grande al más chico, las primeras componentes llevan la mayor varianza de los datos (es lo que efectivamente hace PCA).





# Dependencia de las unidades y escaleo

¿Qué pasa con la variabilidad si cambiamos las unidades de alguna de las variables?



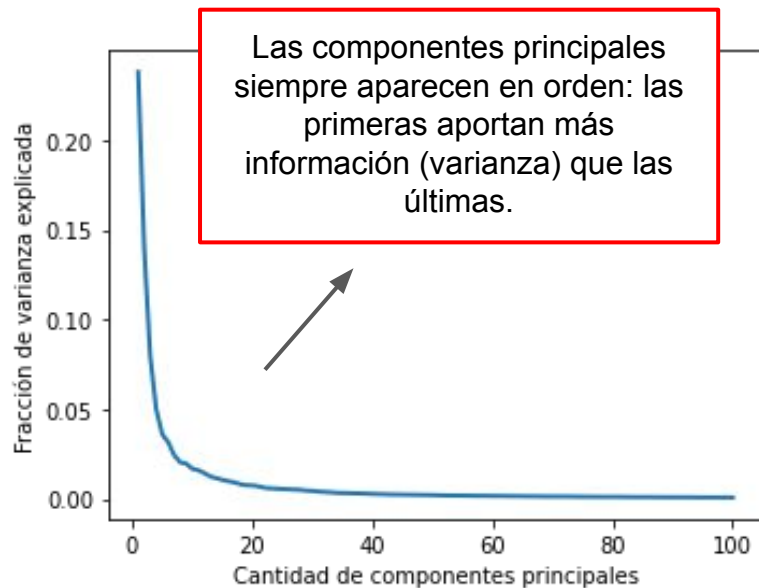
Para evitar introducir o sacar variabilidad por cambiar las unidades de alguna de las variables, una práctica habitual antes de hacer PCA es estandarizar las variables:

$$Z = \frac{X - \mu}{\sigma}$$

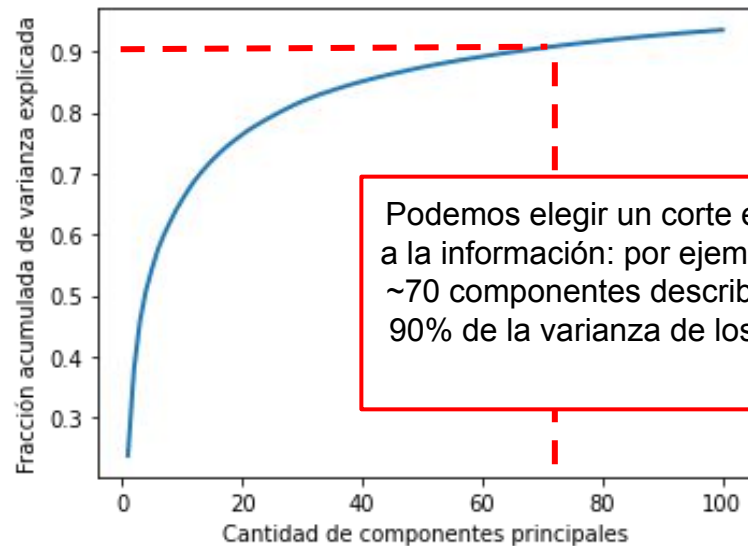


# ¿Cuántas componentes?

Al ser un problema de aprendizaje no-supervisado, no tenemos un conjunto de test para validar el número de componentes que elijamos. ¿Con cuántas nos quedamos?



**Fracción de varianza que aporta cada componente**



**Fracción de varianza acumulada**



# Reducción dimensional como input de otros modelos

Podemos usar las primeras componentes principales ( $Z_m$ ) como variables independientes en modelos de regresión o clasificación.

$$y \sim \beta_0 + \beta_1 Z_1 + \dots + \beta_M Z_M$$

Ventajas:

- Resolvemos el problema de colinealidad: En PCA las variables son por defecto ortogonales y por lo tanto buenas candidatas.
- Podemos validar nuestro modelo y por lo tanto tomar la cantidad de componentes principales como un hiperparámetro.



# Resumen de PCA

- Las componentes principales son una combinación lineal de los features originales y se corresponden con los autovectores de la matriz de covarianza de los datos.
- Las componentes están ordenadas de mayor a menor, en el sentido de la información (varianza) que se llevan. Si tiramos las últimas componentes, estamos reduciendo la dimensión de nuestro problema.
- Una práctica usual es estandarizar las variables antes de aplicar PCA (fundamental si los features corresponden a distintas magnitudes).
- Podemos usar las componentes principales para: comprimir la información, visualizar datos multidimensionales en bajas dimensiones, y como input para modelos de aprendizaje supervisado.



# Vamos al Notebook



factor-data  
EIDAES\_UNSAM