eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Programa de doctorado en investigación biomédica

MDe
Master eta Doktorego Eskola
Escuela de Máster y Doctorado
Master and Doctoral School

ANÁLISIS, PREDICCIÓN Y CLASIFICACIÓN DE DATOS BIOMÉDICOS

Javier Rasero (jrasero.daparte@gmail.com)
Instituto Biocruces Bizkaia

biocruces bizkaia
osasun ikerketa institutua
instituto de investigación sanitaria

Estructura del curso

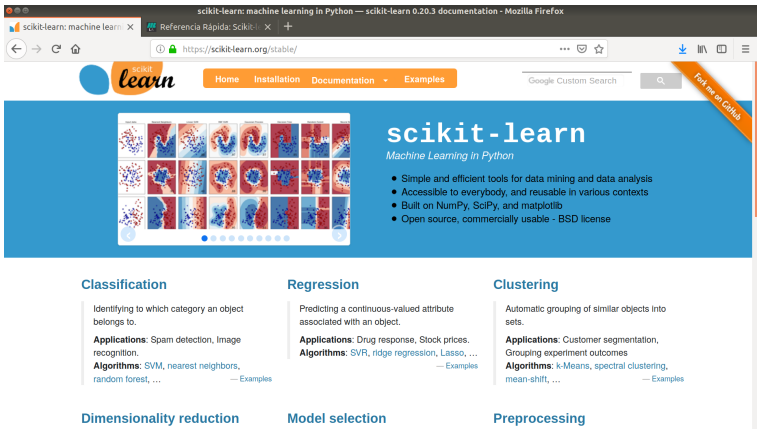
- **Semana 1:** Introducción al Machine Learning.
- **Semana 2:** Aprendizaje Supervisado I. Regresión.
- **Semana 3:** Aprendizaje Supervisado II. Clasificación.
- **Semana 4:** Selección del modelo e hiperparámetros y su evaluación.
- **Semana 5:** Aprendizaje no Supervisado: clustering y reducción de la dimensionalidad de los datos.

El curso

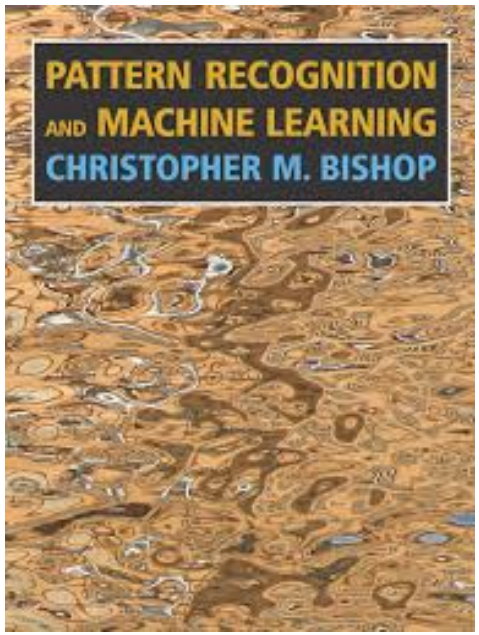
- El curso introduce al machine learning mediante la librería **scikit-learn**, escrita en Python.
- Consistirá en diapositivas con las nociones teóricas (miércoles), tutoriales (miércoles-jueves) y ejercicios (jueves-casa).
- Tanto los tutoriales como los ejercicios se realizarán a través de notebooks de **Jupyter** y siempre en **Python**.
- Todas las datasets usadas en este curso serán de dominio público, bien porque están incluidas en las diferentes librerías y plataformas de estadística y machine learning o porque se encuentran en repositorios de machine learning open-access.

Todo el material del curso lo podréis encontrar en
`jrasero.github.io/curso-scikit-ehu-2019`

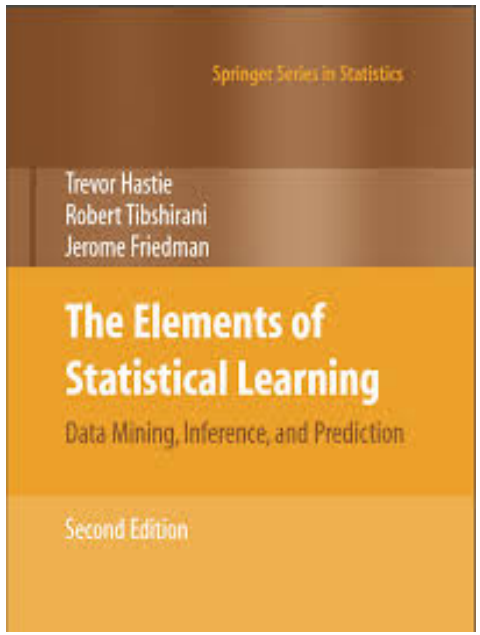
Referencias del curso



http://scikit-learn.org/

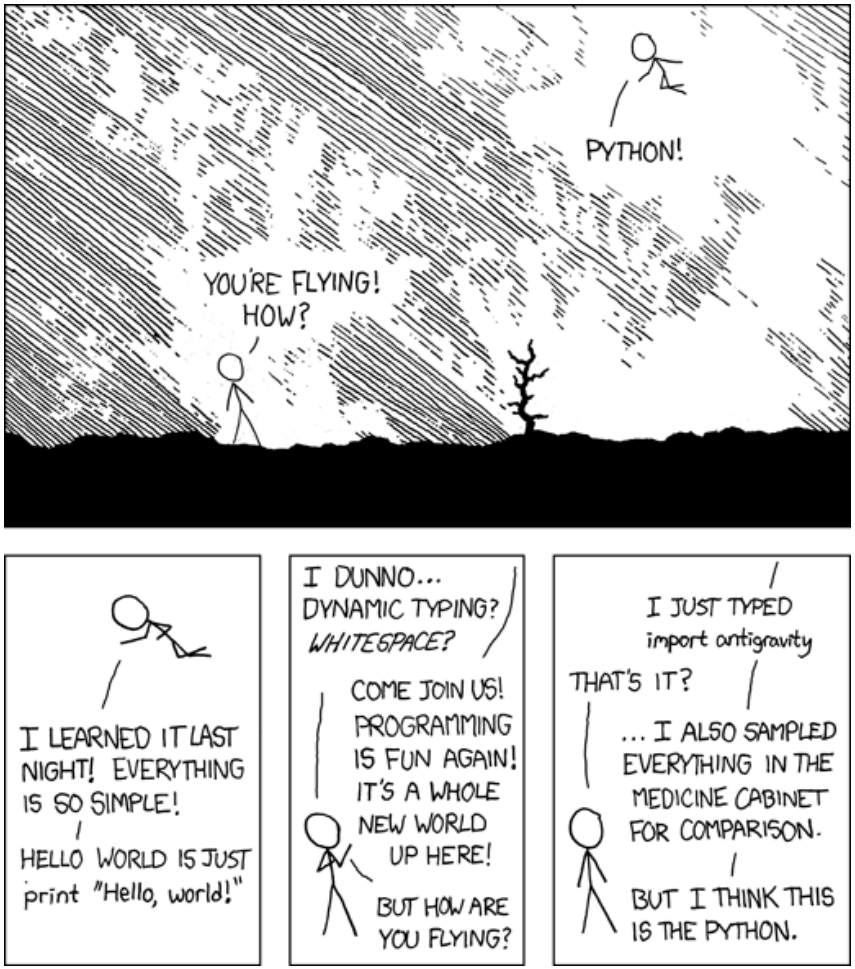


Christopher M. Bishop. Pattern Recognition and Machine Learning. Springer 2006



T. Hastie, R. Tibshirani, J.Friedman. The Elements of Statistical Learning. Springer, 2009

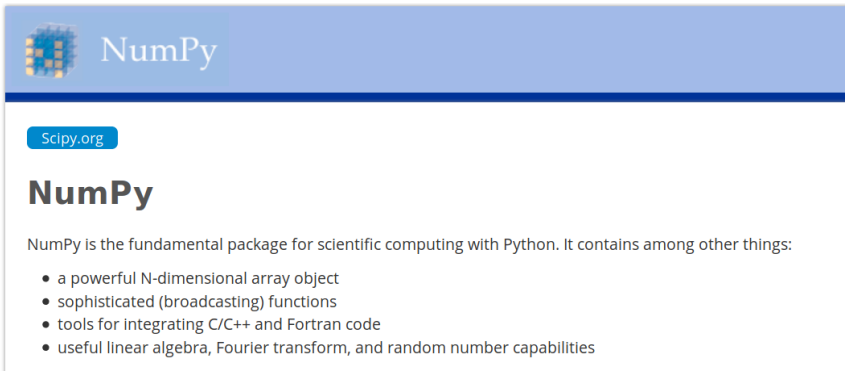
Código: Python



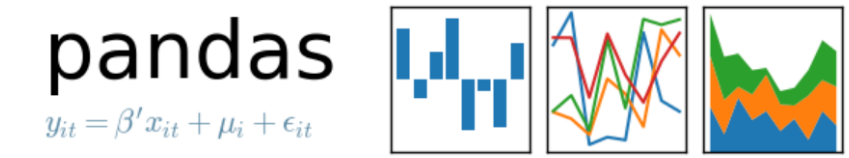
from xkcd.com

Librerías esenciales para el análisis de datos en python

Datos

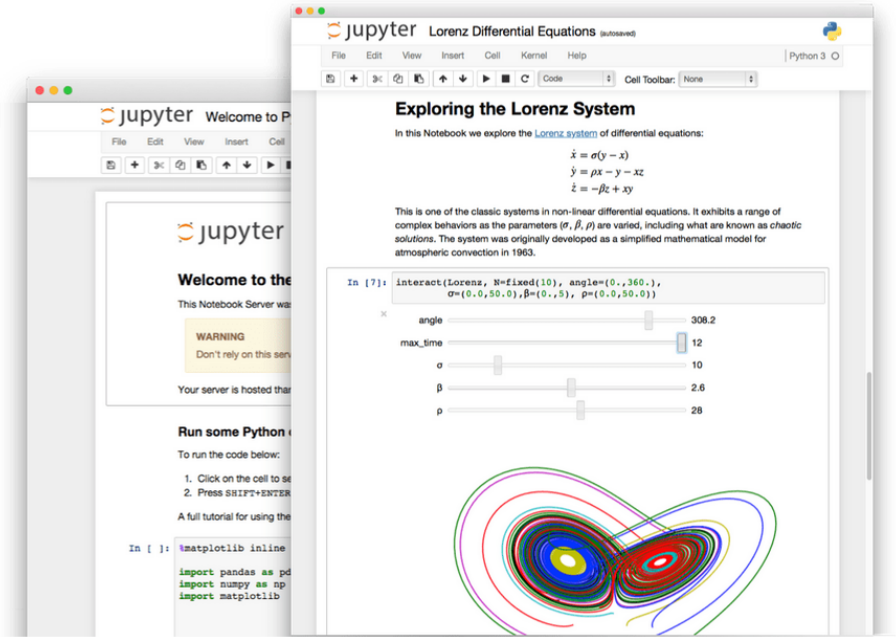


Manipulación y exploración



Visualización

Jupyter project



The Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Try it in your browser | Install the Notebook

Language of choice

Share notebooks

Interactive output

Big data integration

scikit-learn

En este curso, nosotros usaremos scikit-learn, que es una librería de machine learning escrita en python:

- Simple y efficiente.
- Open source.
- Gran cantidad de algoritmos
- En continuo desarrollo
- Perfectamente documentado

Introducción al Machine Learning (ML)

Los clínicos en general manejan de forma subóptima e imperfecta los datos.

- **Software.** Suelen usar software estadístico en un entorno gráfico, pero sin tener el 100% de control de lo que están haciendo
  - SPSS
  - STATA
  - SAS
- **Análisis.** Suele ser sencillo
  - t-test
  - χ^2
  - Correlaciones (Pearson, Spearman)
  - Regresión Univariada
  - Regresión multivariada
  - Curvas de supervivencia

EJEMPLO 1: Human Pathology (3.125)

Original contribution

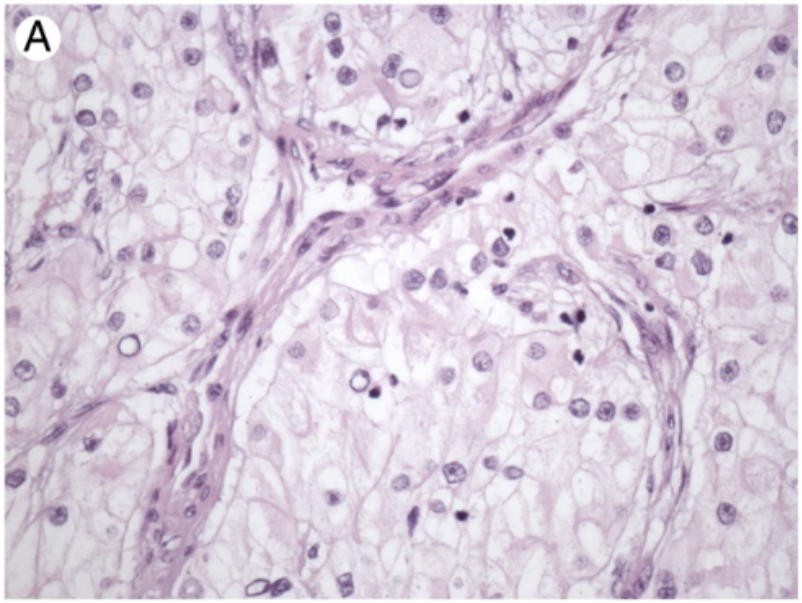
Fibroblast activation protein predicts prognosis in clear cell renal cell carcinoma ☆

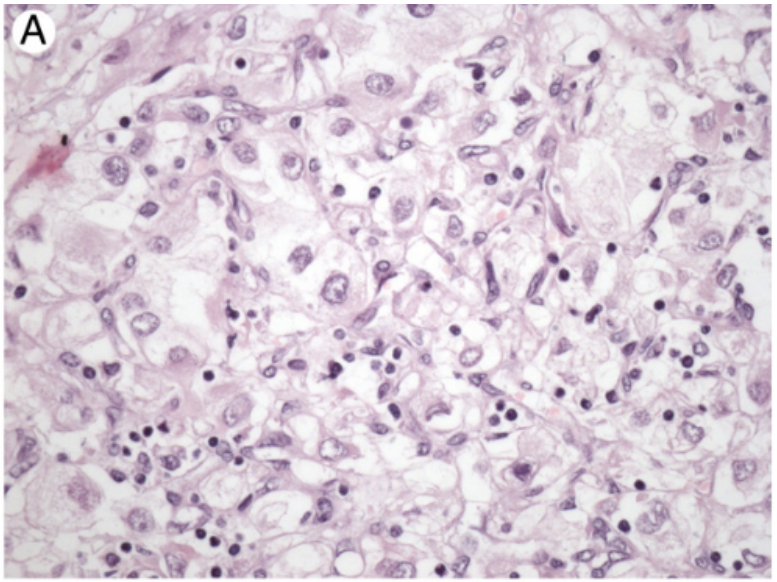
José I. López MD, PhD [a,b,*], Peio Errarte BSc [b,c], Asier Erramuzpe MSc [d], Rosa Guarch MD, PhD [e], Jesús M. Cortés PhD [d,f,j], Javier C. Angulo MD, PhD [g], Rafael Pulido PhD [b,f], Jon Irazusta PhD [b,c], Roberto Llarena MD [h], Gorka Larrinaga MD, PhD [b,c,i]

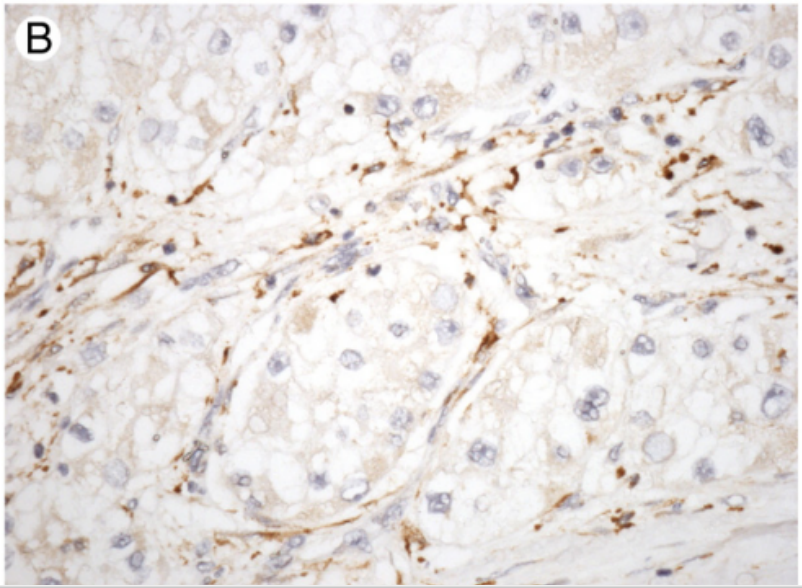
[a] *Department of Pathology, Cruces University Hospital, University of the Basque Country (UPV/EHU), Barakaldo 48903, Bizkaia, Spain*
[b] *Cancer Biomarkers Group, BioCruces Health Research Institute, Barakaldo 48903, Bizkaia, Spain*
[c] *Department of Physiology, University of the Basque Country (UPV/EHU), Leioa 48940, Bizkaia, Spain*
[d] *Quantitative Biomedicine Unit, BioCruces Health Research Institute, Barakaldo 48903, Bizkaia, Spain*
[e] *Department of Pathology, Complejo Hospitalario B de Navarra, Pamplona 31008, Navarra, Spain*
[f] *Ikerbasque, Basque Foundation for Science, Bilbao 48011, Bizkaia, Spain*
[g] *Department of Urology, Hospital de Getafe, Universidad Europea de Madrid, Getafe, 28907, Madrid, Spain*
[h] *Department of Urology, Cruces University Hospital, University of the Basque Country (UPV/EHU), Barakaldo 48903, Bizkaia, Spain*
[i] *Department of Nursing I, School of Nursing, University of the Basque Country (UPV/EHU), Leioa 48940, Bizkaia, Spain*
[j] *Department of Cell Biology and Histology, University of the Basque Country (UPV/EHU), Leioa, Spain*

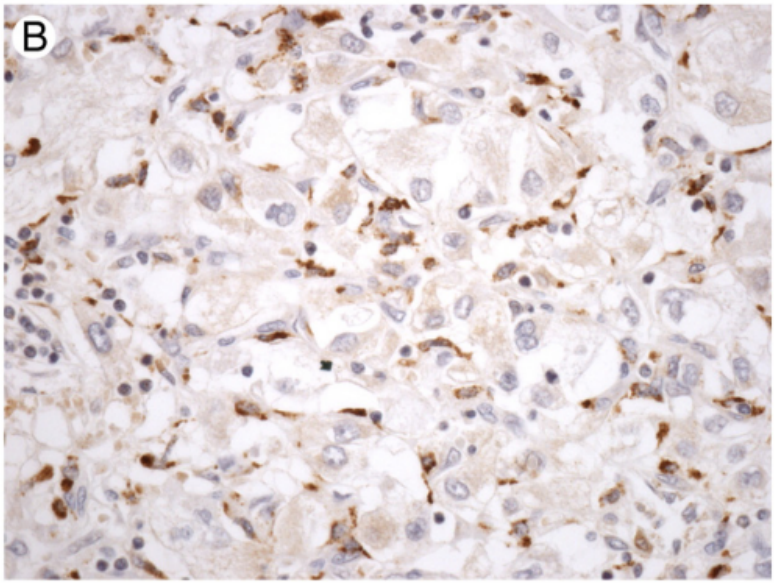
Received 16 January 2016; revised 22 February 2016; accepted 1 March 2016

EJEMPLO 1: Human Pathology (3.125)









EJEMPLO 1: Human Pathology (3.125)

age	sex	histologic type	grade	diam	stage	FAP	followup	situation
80	M	CCRCC	2	5,5	1b	NEG	180	alive
68	F	CCRCC	2	2,5	1a	NEG	183	alive
84	F	CCRCC	2	19	2	NEG	132	dead
39	M	CCRCC	3	10	3a	NEG	60	dead
54	M	CCRCC	2	4	1a	NEG	174	alive
84	F	CCRCC	3	13	3b	NEG	53	dead
73	M	CCRCC	4	8	2	POS	31	dead
89	F	CCRCC	4	9	3a	POS	37	alive
87	M	CCRCC	2	5	1b	NEG	48	dead
78	M	CCRCC	2	5,3	1b	NEG	13	dead
75	M	CCRCC	2	5	1b	POS	210	dead
87	M	CCRCC	4	7,5	3a	NEG	121	dead
77	F	CCRCC	4	10	3a	NEG	15	dead
60	M	CCRCC	3	6,5	1b	NEG	192	alive
59	M	CCRCC	1	7	3a	NEG	204	alive
91	F	CCRCC	3	6,5	3b	NEG	38	dead
53	M	CCRCC	3	12,5	2	POS	25	dead
69	M	CCRCC	4	5,5	3a	NEG	189	alive
51	M	CCRCC	2	9	2	NEG	180	alive
83	M	CCRCC	2	4	3a	NEG	156	dead
72	M	CCRCC	2	5	1b	NEG	201	alive
58	M	CCRCC	3	5	3b	NEG	144	dead
75	F	CCRCC	3	8	2	NEG	23	dead
44	M	CCRCC	2	2,8	1a	NEG	192	alive

Table 1 Log-rank *P* for 5-, 10-, and 15-year survival

	Grade [a]	Stage [b]	Diameter [c]	FAP [d]
5 y	.0000087	.00000000085	.0000001	.00015
10 y	.00000025	.000000031	.000013	.0000042
15 y	.000000083	.000000001	.0000082	.000043

[a] Fuhrman grade, low (G1/2) versus high (G3/4).
[b] American Joint Committee on Cancer 2010 stage, low (pT1/2) versus high (≥ pT3).
[c] Tumor diameter, small (≤4 cm) versus large (>4 cm).
[d] FAP expression, positive versus negative.

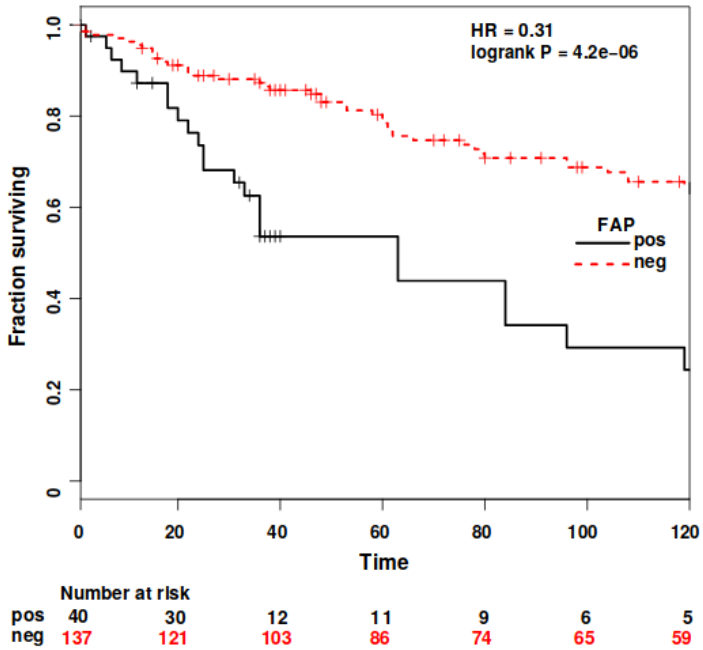
Table 2 Univariate regression analysis

Variable	*P*
Grade	.00000124
Stage	.000000000666
Tumor diameter	.000028
FAP expression	.000000764

Table 3 Multivariate regression analysis

Variable	*P*
Grade	.04162
Stage	.02106
Tumor diameter	.64408 [a]
FAP expression	.00117

[a] Not statistically significant.



EJEMPLO 2: Journal of Pathology (6.253)

Journal of Pathology
J Pathol 2014; **232:** 32–42
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/path.4296

ORIGINAL PAPER

Targeted next-generation sequencing and non-coding RNA expression analysis of clear cell papillary renal cell carcinoma suggests distinct pathological mechanisms from other renal tumour subtypes

Charles H Lawrie,[1,2,3]* Erika Larrea,[1] Gorka Larrinaga,[4] Ibai Goicoechea,[1] María Arestin,[1] Marta Fernandez-Mercado,[1] Ondrej Hes,[5] Francisco Cáceres,[6] Lorea Manterola[1] and José I López[7]

[1] *Oncology Area, Biodonostia Research Institute, San Sebastian, Spain*
[2] *Nuffield Department of Clinical Laboratory Sciences, University of Oxford, UK*
[3] *IKERBASQUE, Basque Foundation for Science, Bilbao, Spain*
[4] *Nursing School, University of the Basque Country (UPV/EHU), Leioa, Bizkaia, Spain*
[5] *Department of Pathology, Charles University Hospital, Plzen, Czech Republic*
[6] *Department of Urology, Cruces University Hospital, Barakaldo, Bizkaia, Spain*
[7] *Department of Pathology, Cruces University Hospital, BioCruces Research Institute, University of the Basque Country (UPV/EHU), Barakaldo, Bizkaia, Spain*

EJEMPLO 2: Journal of Pathology (6.253)

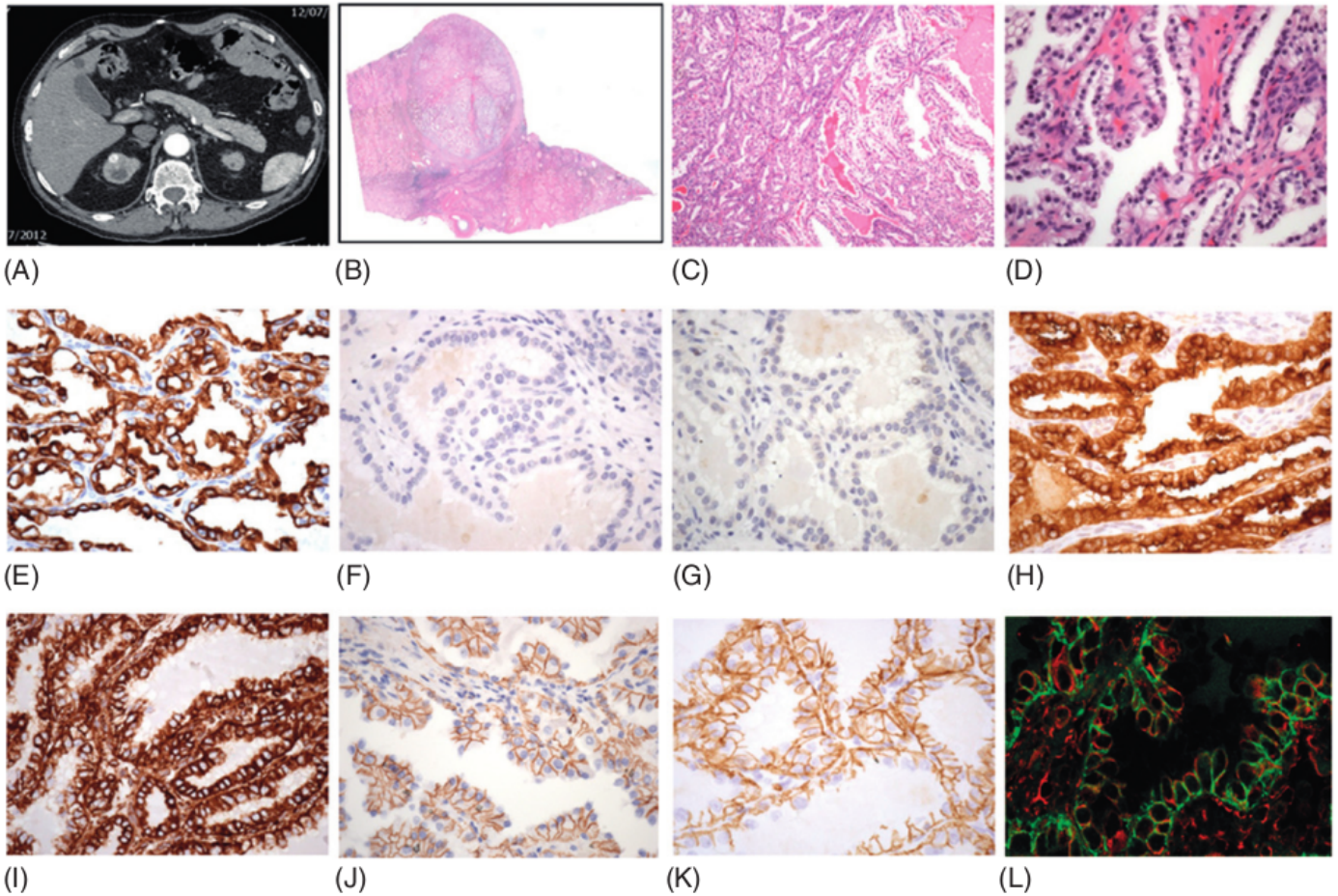
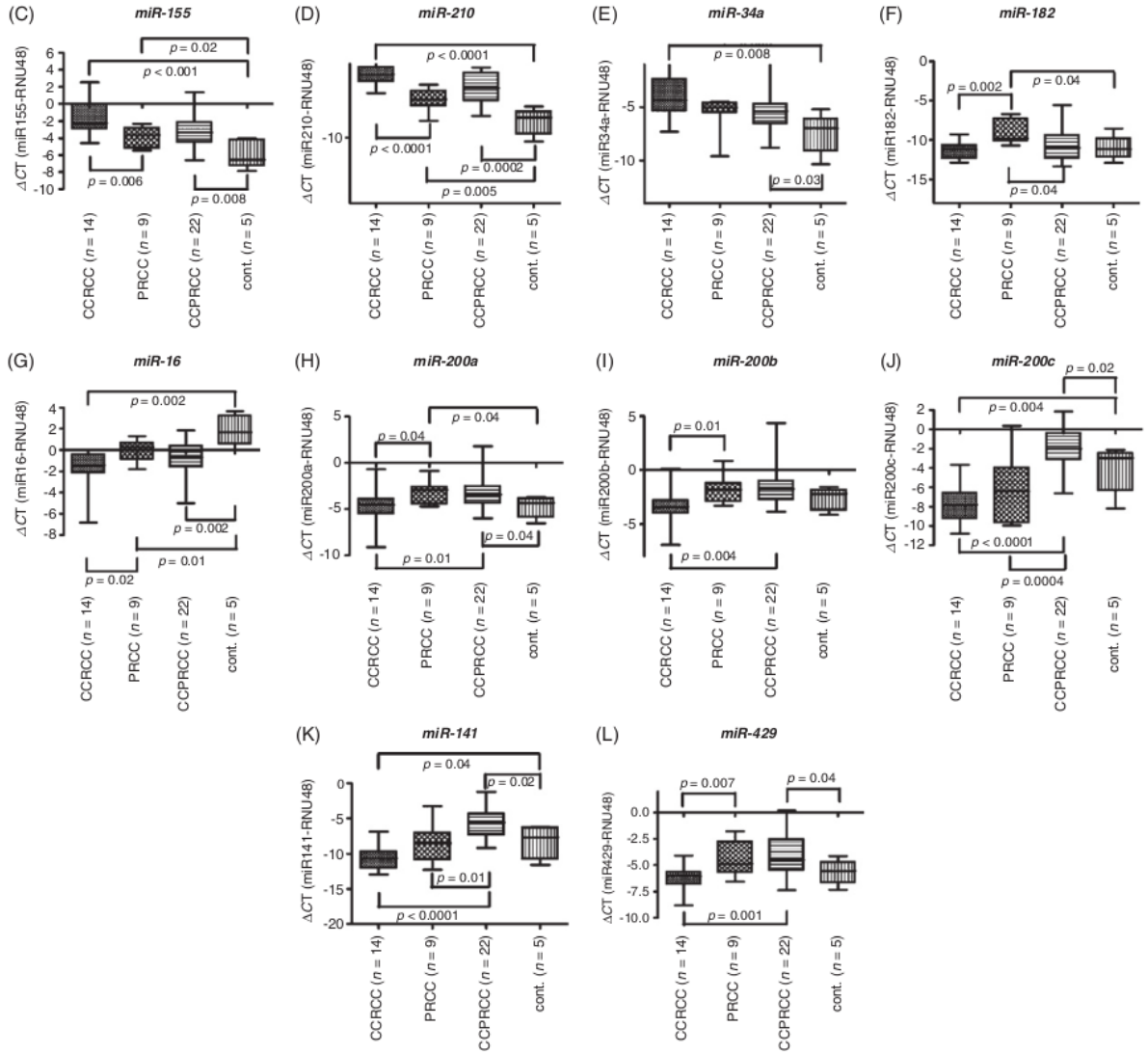
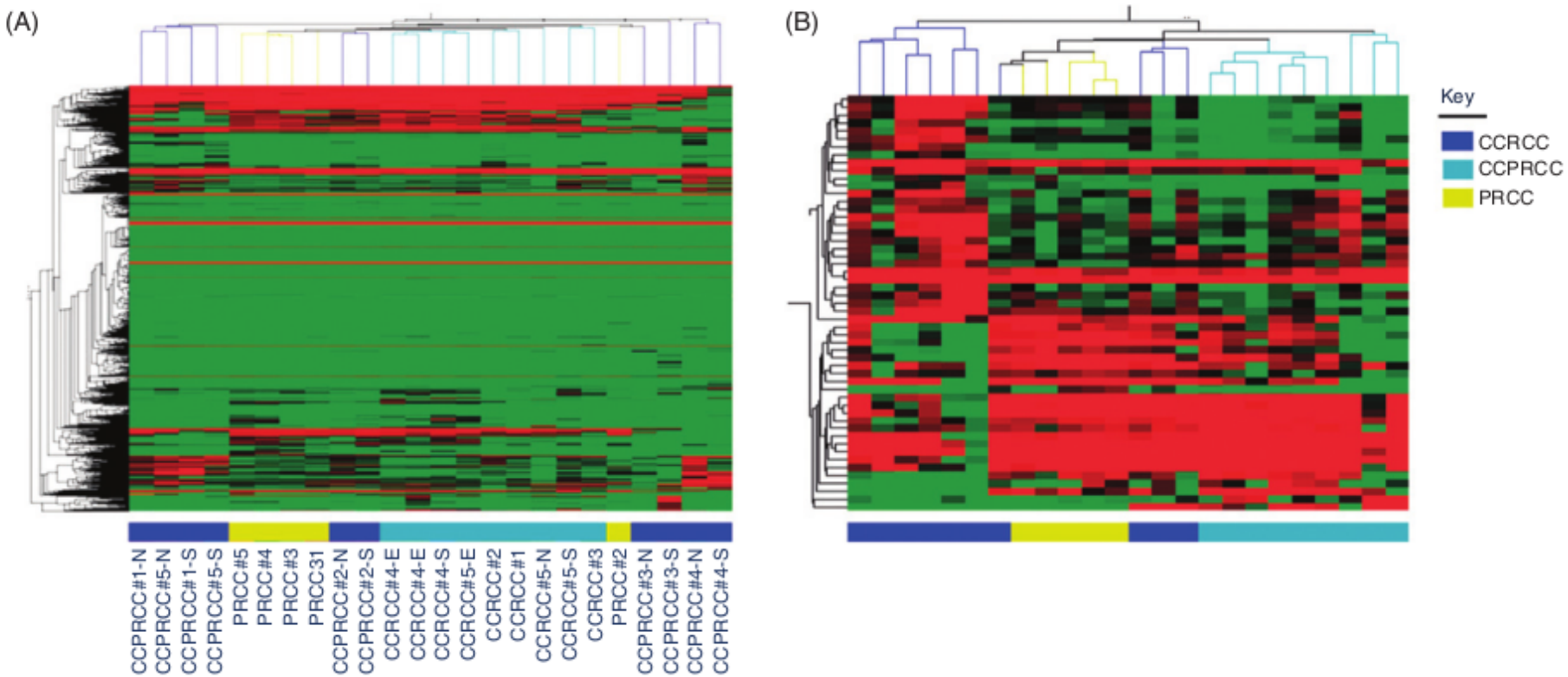


Figure 1. Histological and immunohistochemical features of CCPRCC. (A) CT scan showing small (10 mm diameter) peripheral tumour in the anterior side of the right kidney. (B) Macro/micro histological view of the same tumour showing complete capsulation. (C) Medium-power field showing typical CCPRCC tubulopapillary architecture (magnification, ×100). (D) High-power magnification showing papillae covered by clear cells with hyperchromatic low-grade nuclei located in the luminal side of the cells (magnification, ×400). (E) Diffuse immunostaining for CK7; (F) negative immunostaining for CD10; (G) negative immunostaining for AMACR; (H) diffuse immunostaining for EMA; (I) diffuse immunostaining for Vimentin; (J) diffuse membranous immunostaining for β-catenin; and (K) diffuse membranous immunostaining for E-cadherin. (L) Double imunofluorescent staining with E-cadherin (red) and vimentin (green).

EJEMPLO 2: Journal of Pathology (6.253)



EJEMPLO 2: Journal of Pathology (6.253)



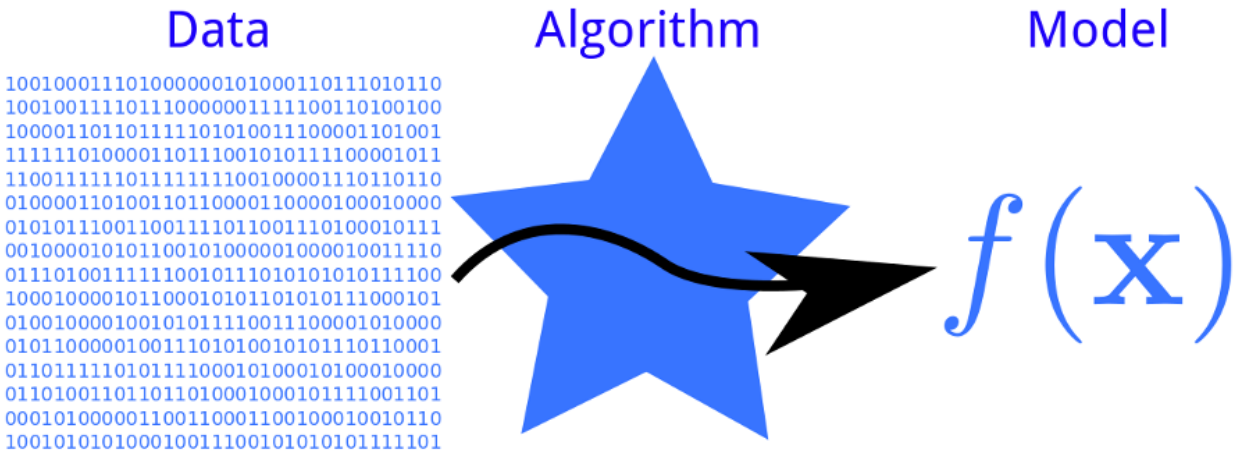
- El tipo de dato suele determinar el impacto de nuestro trabajos.
- De manera general, metodos más avanzados nos permiten publicar en revistas de mayor impacto.
- En biomedícina, es de vital importancia la correcta predicción (o diagnóstico) de los sujetos según su patología o desarrollo de la enfermedad.

- El tipo de dato suele determinar el impacto de nuestro trabajos.
- De manera general, metodos más avanzados nos permiten publicar en revistas de mayor impacto.
- En biomedícina, es de vital importancia la correcta predicción (o diagnóstico) de los sujetos según su patología o desarrollo de la enfermedad.

MACHINE LEARNING

Qué es machine Learning

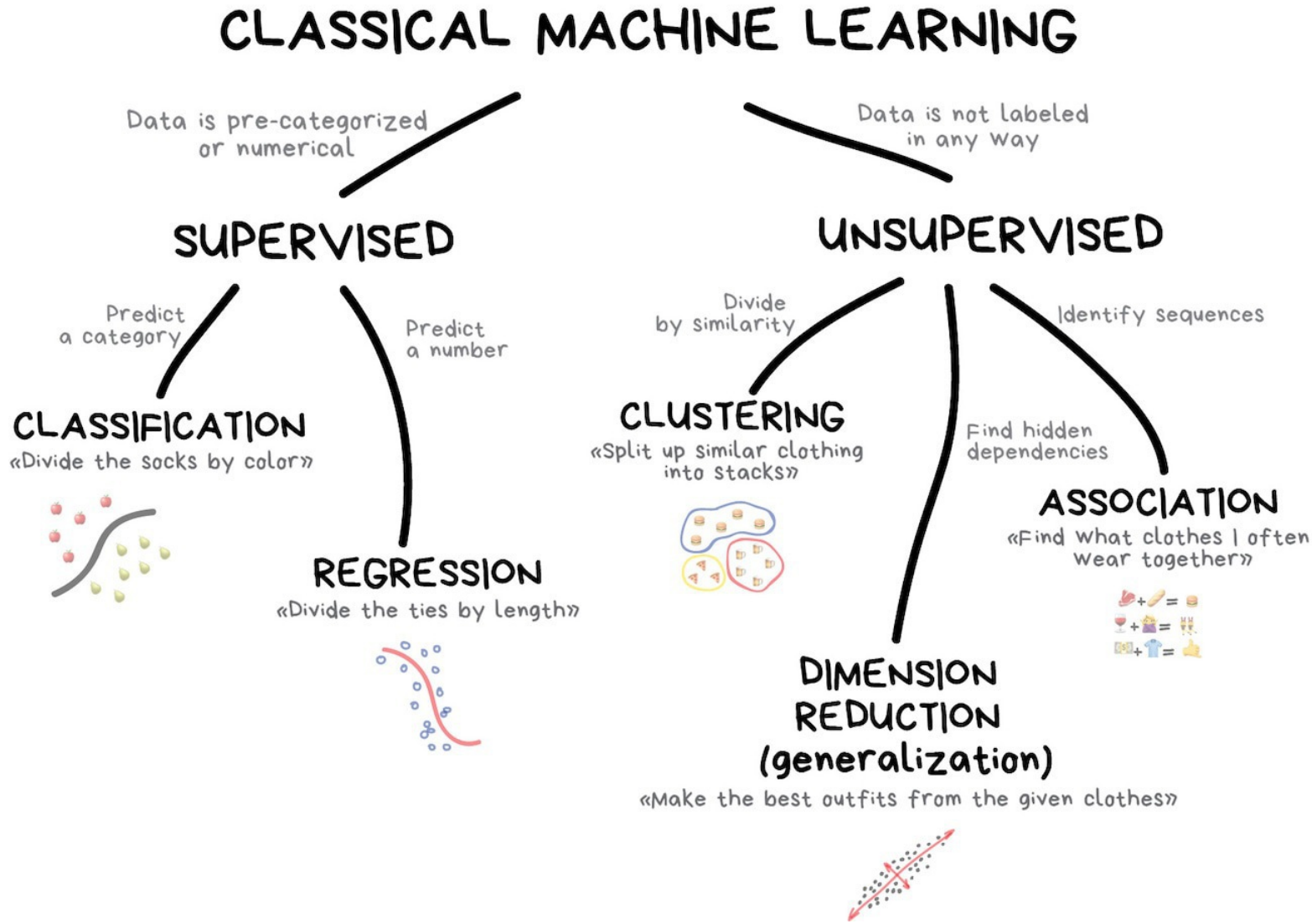
Dados unos datos, machine learning permite aprender la función que mejor los aproxima



from https://blogs.bmj.com/

Si la función tiene algún constraint, entonces hacemos aprendizaje supervisado. En caso contrario, es no supervisado.

Tipos de Machine Learning



from https://vas3k.com/

Machine Learning

Típicas tareas:

- Estructura de los datos
- Preprocesar los datos, entender sus rango de valores, las relaciones entre ellos
- Visualizar los datos
- Clasificarlos (de forma supervisada o no)
- Feature selection
- Cross-validation
- Métricas de calidad en la clasificación (matriz de confusión, precisión, especificidad)

ML en biomedicina

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed | machine learning[Title] | Search

Create RSS | Create alert | Advanced | Help

Article types
Clinical Trial
Review
Customize ...

Text availability
Abstract
Free full text
Full text

Publication dates
5 years
10 years
Custom range...

Species
Humans
Other Animals

Clear all

Show additional filters

Format: Summary **Sort by**: Most Recent **Per page**: 20 **Send to**

Search results

Items: 1 to 20 of 4522

<< First < Prev Page 1 of 227 Next > Last >>

1. **Machine learning** approaches to studying the role of cognitive reserve in conversion from mild cognitive impairment to dementia.
Facal D, Valladares-Rodriguez S, Lojo-Seoane C, Pereiro AX, Anido-Rifon L, Juncos-Rabadán O.
Int J Geriatr Psychiatry. 2019 Mar 10. doi: 10.1002/gps.5090. [Epub ahead of print]
PMID: 30854737
Similar articles

2. Live-cell phenotypic-biomarker microfluidic assay for the risk stratification of cancer patients via **machine learning**.
Manak MS, Varsanik JS, Hogan BJ, Whitfield MJ, Su WR, Joshi N, Steinke N, Min A, Berger D, Saphirstein RJ, Dixit G, Meyyappan T, Chu HM, Knopf KB, Albala DM, Sant GR, Chander AC.
Nat Biomed Eng. 2018 Oct;2(10):761-772. doi: 10.1038/s41551-018-0285-z. Epub 2018 Sep 17.
PMID: 30854249
Similar articles

Filters: Manage Filters

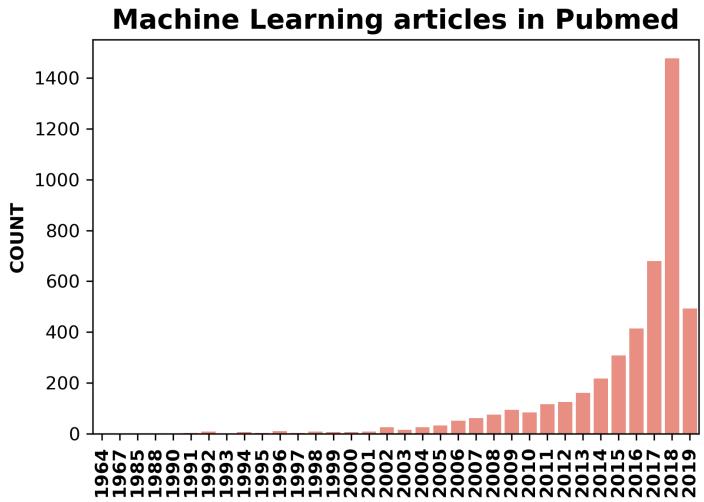
Sort by:
Best match | Most recent

Results by year

Download CSV

Titles with your search terms

Find related data
Database: Select



Ejemplos de ML en biomedicina

ARTICLE

doi:10.1038/nature26000

DNA methylation-based classification of central nervous system tumours

A list of authors and their affiliations appears in the online version of the paper.

Accurate pathological diagnosis is crucial for optimal management of patients with cancer. For the approximately 100 known tumour types of the central nervous system, standardization of the diagnostic process has been shown to be particularly challenging—with substantial inter-observer variability in the histopathological diagnosis of many tumour types. Here we present a comprehensive approach for the DNA methylation-based classification of central nervous system tumours across all entities and age groups, and demonstrate its application in a routine diagnostic setting. We show that the availability of this method may have a substantial impact on diagnostic precision compared to standard methods, resulting in a change of diagnosis in up to 12% of prospective cases. For broader accessibility, we have designed a free online classifier tool, the use of which does not require any additional onsite data processing. Our results provide a blueprint for the generation of machine-learning-based tumour classifiers across other cancer entities, with the potential to fundamentally transform tumour pathology.

Ejemplos de ML en biomedicina

ARTICLE RESEARCH

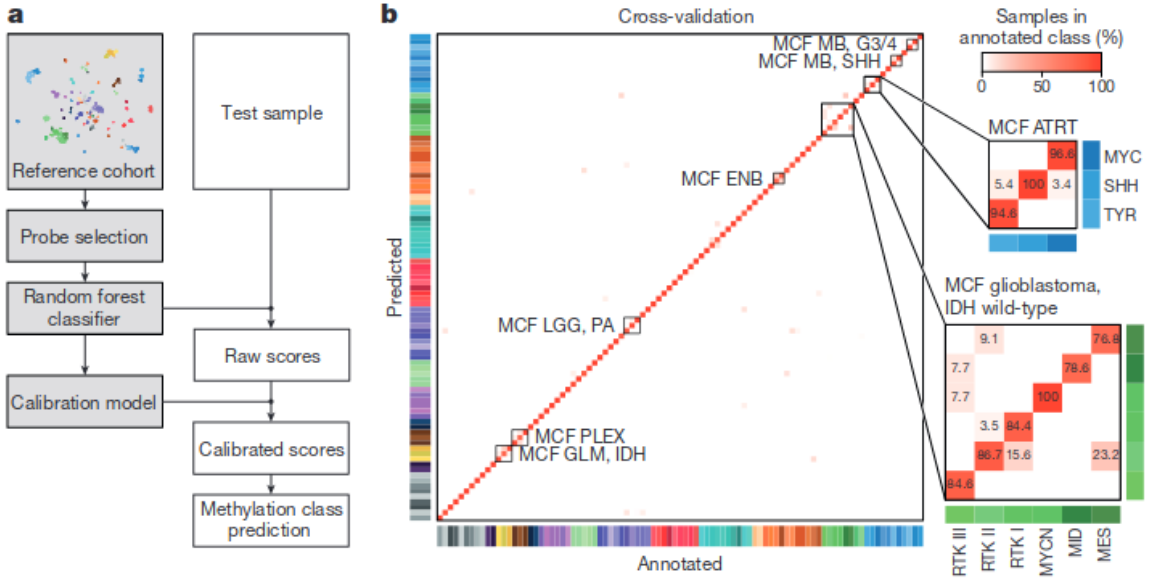


Figure 2 | Development and cross-validation of the DNA methylation-based CNS tumour classifier. a, Schematic of principal classifier components (grey) and processing steps for individual test samples (white). The most informative probes are selected for training of the random forest classifier. The classifier produces raw scores that represent the number of decision trees that assign a test sample to a specific methylation class. To enable inter-class comparability, a calibration model is used, which transforms raw scores into calibrated scores. Calibrated scores represent an estimated probability measure of methylation class assignment. **b**, Heat map showing results of a threefold cross-validation of the random forest classifier incorporating information of $n = 2{,}801$ biologically independent samples allocated to 91 methylation classes. Deviations from the bisecting line represent misclassification errors (using the maximum calibrated score for class prediction). Methylation class families (MCF) are indicated by black squares. The colour code and abbreviations are identical to Fig. 1a.

Ejemplos de ML en biomedicina

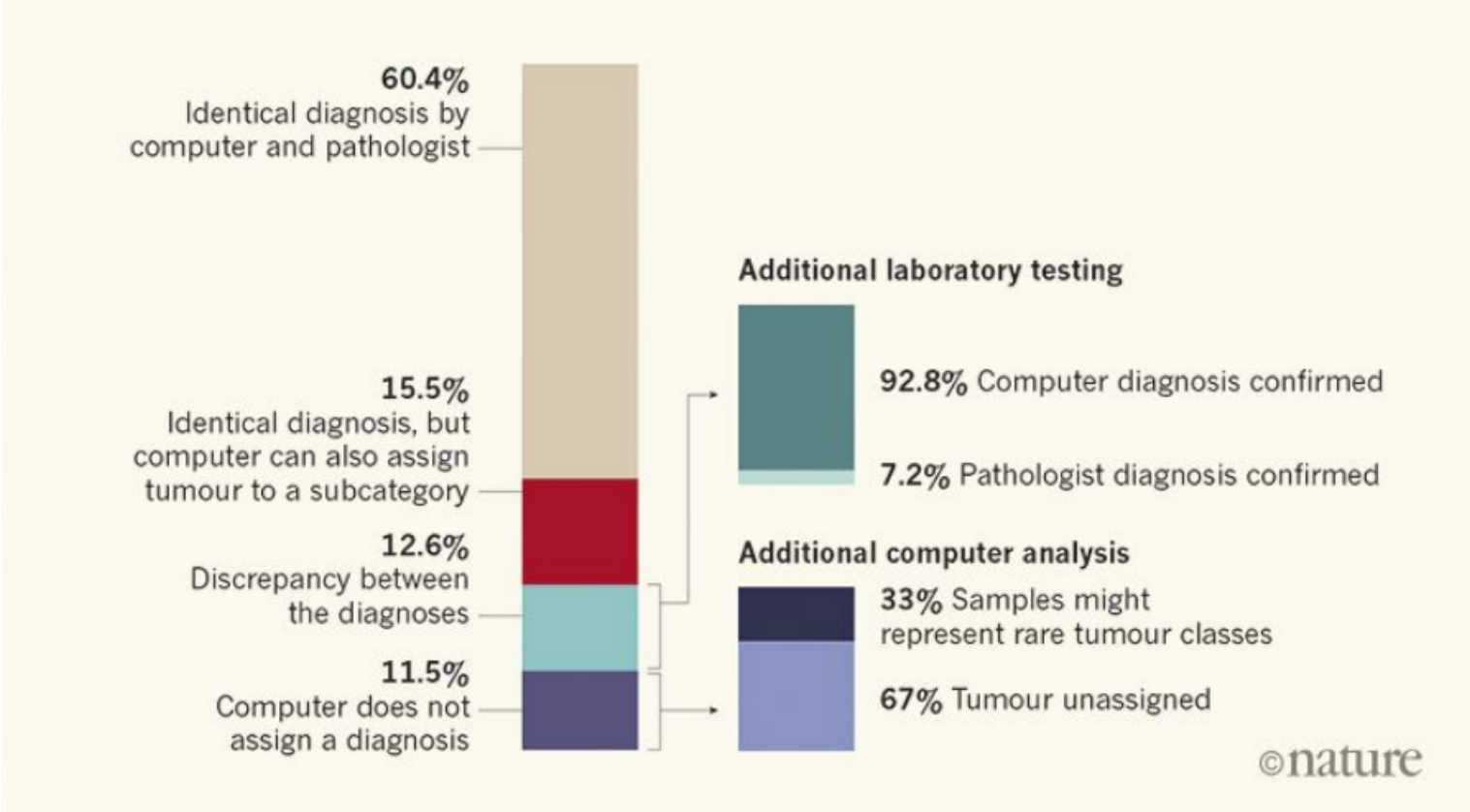


Figure 1 | Tumour classification using a machine-learning approach. Capper *et al.*[1] used a machine-learning approach to classify brain tumours on the basis of genome-wide patterns of a type of DNA alteration called methylation. The computer was trained using methylation data for tumour samples that had been diagnosed by pathologists using standard microscopy-based analysis or analysis of selected genes. After training, the computer was given 1,104 test cases. The authors compared the diagnoses made by the computer and by the pathologists. Although the machine was unable to diagnose all specimens, of the specimens that it classified, the machine-based diagnosis was more accurate or could assign tumours to more-specific subcategories than the classifications made by the pathologists.

Ejemplos de ML en biomedicina

nature biomedical engineering

ARTICLES

https://doi.org/10.1038/s41551-018-0285-z

Live-cell phenotypic-biomarker microfluidic assay for the risk stratification of cancer patients via machine learning

Michael S. Manak[1,6], Jonathan S. Varsanik[1,6], Brad J. Hogan[1], Matt J. Whitfield[1], Wendell R. Su[1], Nikhil Joshi[1], Nicolai Steinke[1], Andrew Min[1], Delaney Berger[1], Robert J. Saphirstein[1], Gauri Dixit[1], Thiagarajan Meyyappan[1], Hui-May Chu[2], Kevin B. Knopf[3], David M. Albala[4], Grannum R. Sant[5] and Ashok C. Chander[1]*

The risk stratification of prostate cancer and breast cancer tumours from patients relies on histopathology, selective genomic testing, or on other methods employing fixed formalin tissue samples. However, static biomarker measurements from bulk fixed-tissue samples provide limited accuracy and actionability. Here, we report the development of a live-primary-cell phenotypic-biomarker assay with single-cell resolution, and its validation with prostate cancer and breast cancer tissue samples for the prediction of post-surgical adverse pathology. The assay includes a collagen-I/fibronectin extracellular-matrix formulation, dynamic live-cell biomarkers, a microfluidic device, machine-vision analysis and machine-learning algorithms, and generates predictive scores of adverse pathology at the time of surgery. Predictive scores for the risk stratification of 59 prostate cancer patients and 47 breast cancer patients, with values for area under the curve in receiver-operating-characteristic curves surpassing 80%, support the validation of the assay and its potential clinical applicability for the risk stratification of cancer patients.

Ejemplos de ML en biomedicina

Table 1 | Machine-learning-derived GAPP, LAPP and MAPP clinical scores

	GAPP	LAPP	MAPP
Prostate	Any of the six adverse pathologies (SVI, PSM, EPE, PNI, LNP, LVI)	Seminal vesicle invasion (SVI) Positive surgical margins (PSM) Extra prostatic extension (EPE)	Perineural invasion (PNI) Lymph node positive (LNP) Lympho-vascular invasion (LVI)
Breast	Any of the four adverse pathologies (ENE, PSM, LVI, LI)	Extranodal extension (ENE) Positive surgical margins (PSM)	Lympho-vascular invasion (LVI) Lymph invasion (LI)

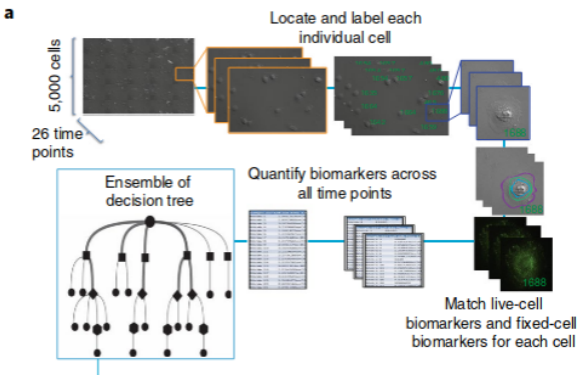


Table 2 | Predictive performance results for adverse pathologies from prostate tissue and breast tissue samples

Predicted adverse pathology	Sensitivity	Specificity	AUC	*N*	True positive	True negative	Predicted positive	Predicted negative
Prostate tissue								
Seminal vesicle invasion	0.89	0.96	0.93	57	9	48	8	46
Positive surgical margin	0.99	0.93	0.94	59	18	41	18	38
Extra-prostatic extension	0.95	0.97	0.96	53	21	32	20	31
Perineural invasion	0.99	0.99	0.99	50	37	13	37	13
Lymph node positive	0.95	0.96	0.81	47	4	43	4	41
Lymph vascular invasion	0.99	0.98	0.98	54	6	48	6	47
GAPP	0.91	0.93	0.88	59	45	14	41	13
LAPP	0.93	0.90	0.93	59	28	31	26	28
MAPP	0.95	0.84	0.89	59	40	19	38	16
Breast tissue								
Extra-nodal extension	0.99	0.73	0.84	37	14	23	13	19
Positive surgical margin	0.99	0.95	0.98	45	3	42	3	39
Lympho-vascular invasion	0.90	0.87	0.87	44	21	23	19	19
Lymph invasion	0.96	0.79	0.91	46	27	19	20	18
GAPP[a]	0.81	0.93	0.85	47	32	15	26	14
LAPP[a]	0.99	0.72	0.81	47	15	32	15	23
MAPP[a]	0.84	0.88	0.85	47	31	16	26	14
$MAPP_{LI}$[b]	0.90	0.85	0.83	32	19	13	15	12
$MAPP_{LVI}$[b]	0.83	0.87	0.86	32	18	14	15	12

Sensitivity, specificity, AUC, total number of samples, true-positive and true-negative numbers, and number of samples predicted positive or negative were obtained from machine-learning-derived statistical algorithms. [a]For breast samples, GAPP, LAPP and MAPP scores are derived from algorithms trained on all breast samples. [b]For breast samples, $MAPP_{LI}$ and $MAPP_{LVI}$ scores are derived from algorithms trained on DCIS positive samples only.

Estadística versus machine learning

Estadística versus machine learning

Statistics draws population inferences from a sample, and machine learning finds generalizable predictive patterns.

Nature Methods volume 15, pages 233–234 (2018)

Estadística versus machine learning

Statistical methods have a long-standing focus on inference, which creates a mathematical model of the data-generation process to formalize understanding or test a hypothesis about how the system behaves.

- it allows us compute a quantitative measure of confidence that a discovered relationship describes a 'true' effect that is unlikely to result from noise.
- if enough data are available, we can explicitly verify assumptions (e.g., equal variance) and refine the specified model, if needed.

Nature Methods volume 15, pages 233–234 (2018)

Estadística versus machine learning

Machine learning concentrates on prediction aims at forecasting unobserved outcomes or future behavior, such as whether a mouse with a given gene expression pattern has a disease.

- Helpful when one is dealing with 'wide data', where the number of input variables exceeds the number of subjects, in contrast to 'long data', where the number of subjects is greater than that of input variables.
- Minimal assumptions about the data-generating systems; they can be effective even when the data are gathered without a carefully controlled experimental design and in the presence of complicated nonlinear interactions.
- The lack of an explicit model can make Machine Learning solutions difficult to directly relate to existing biological knowledge.

Nature Methods volume 15, pages 233–234 (2018)

Estadística versus machine learning

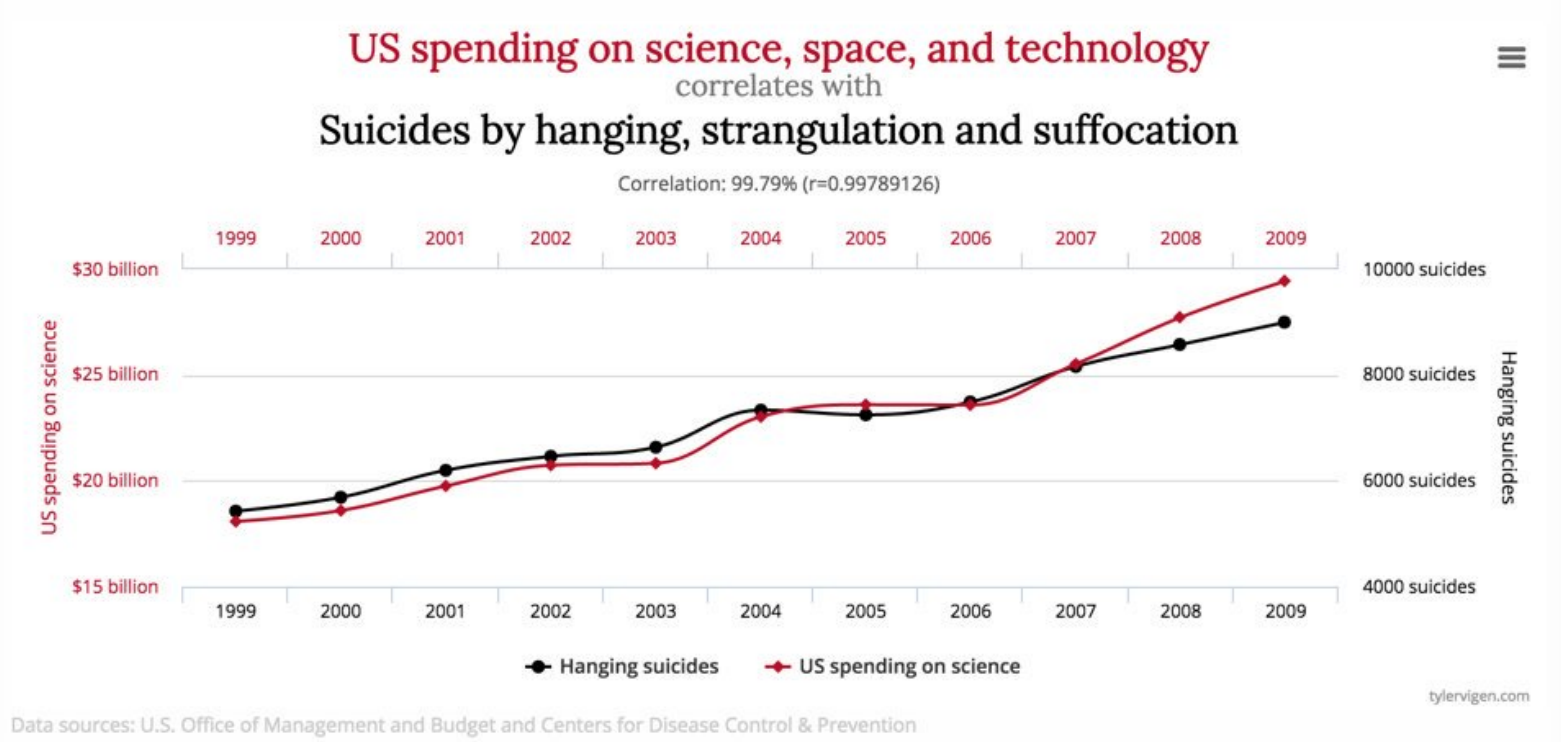
- Classical statistical modeling was designed for data with a few dozen input variables and sample sizes that would be considered small to moderate today. In this scenario, the model fills in the unobserved aspects of the system. However, as the numbers of input variables and possible associations among them increase, the model that captures these relationships becomes more complex. Consequently, statistical inferences become less precise and the boundary between statistical and ML approaches becomes hazier.

Nature Methods volume 15, pages 233–234 (2018)

AVISO 1

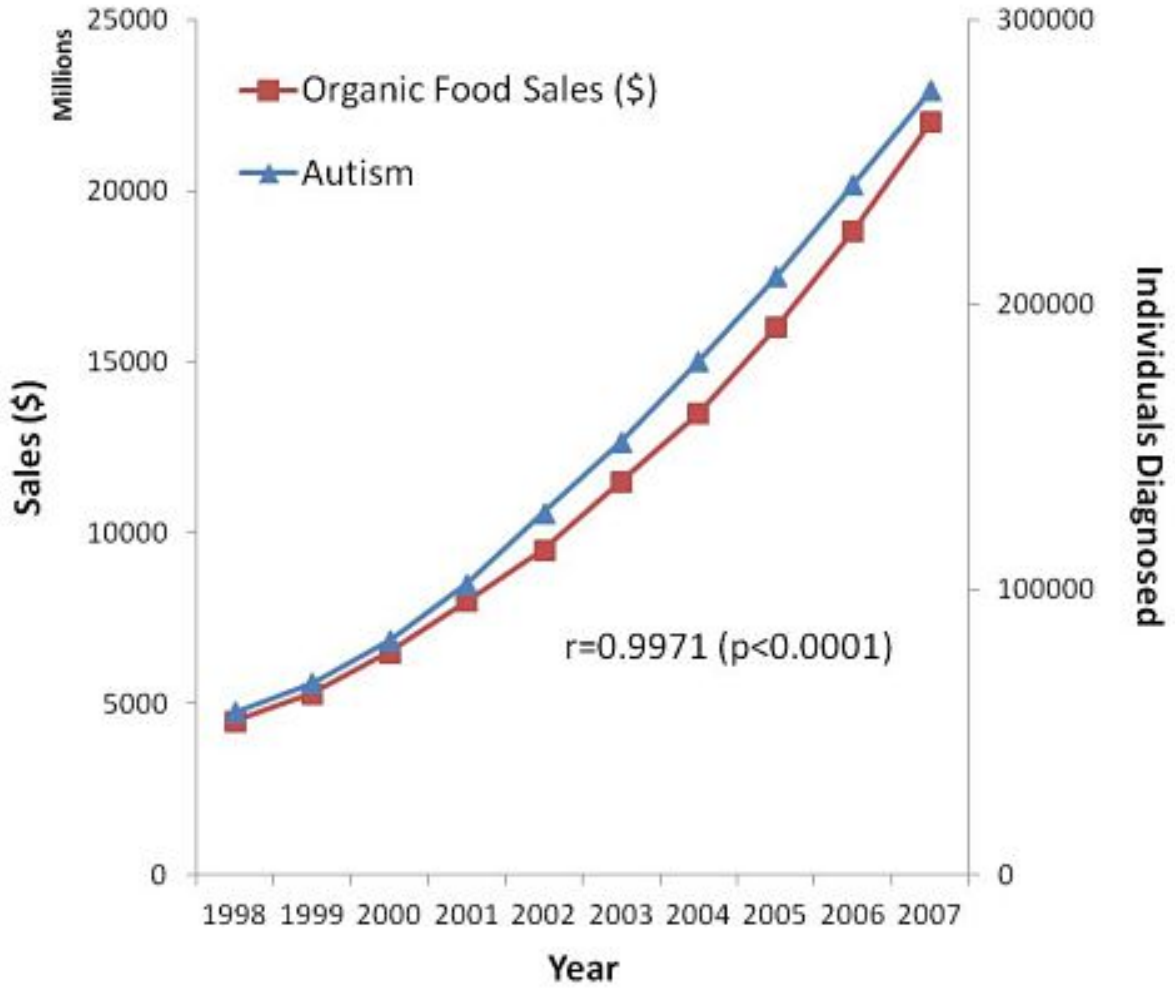
Cuidado con las conclusiones que extraemos de los resultados que podemos encontrar

Las relaciones entre los datos pueden ser sorprendente...



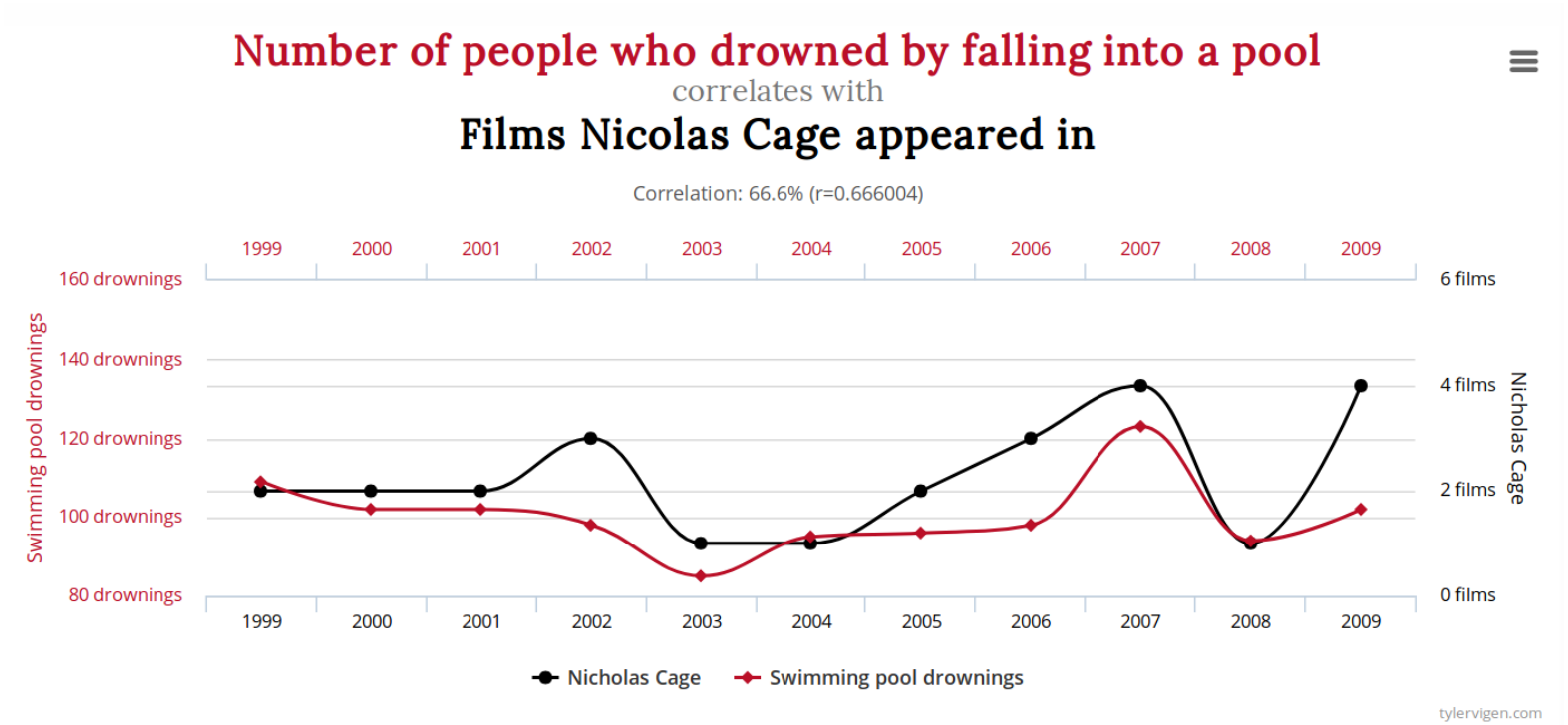
from http://www.tylervigen.com/spurious-correlations

Las relaciones entre los datos pueden ser sorprendente...



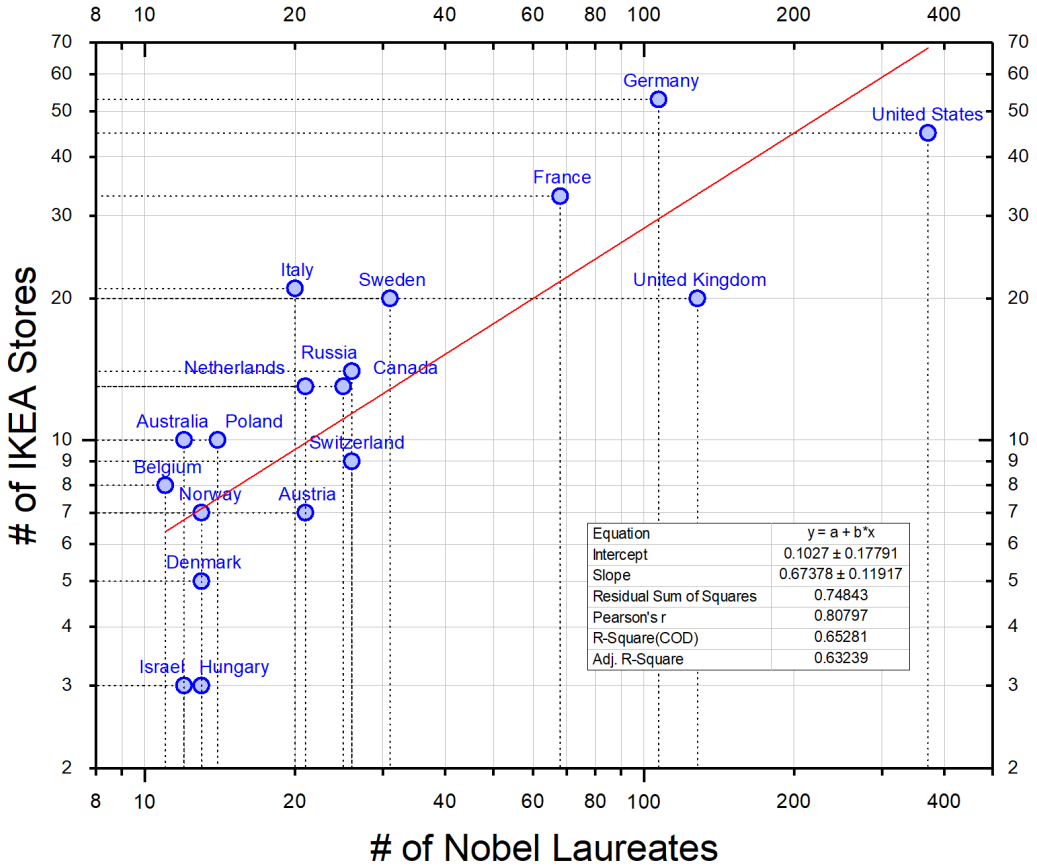
from reddit.com

Las relaciones entre los datos pueden ser sorprendente...



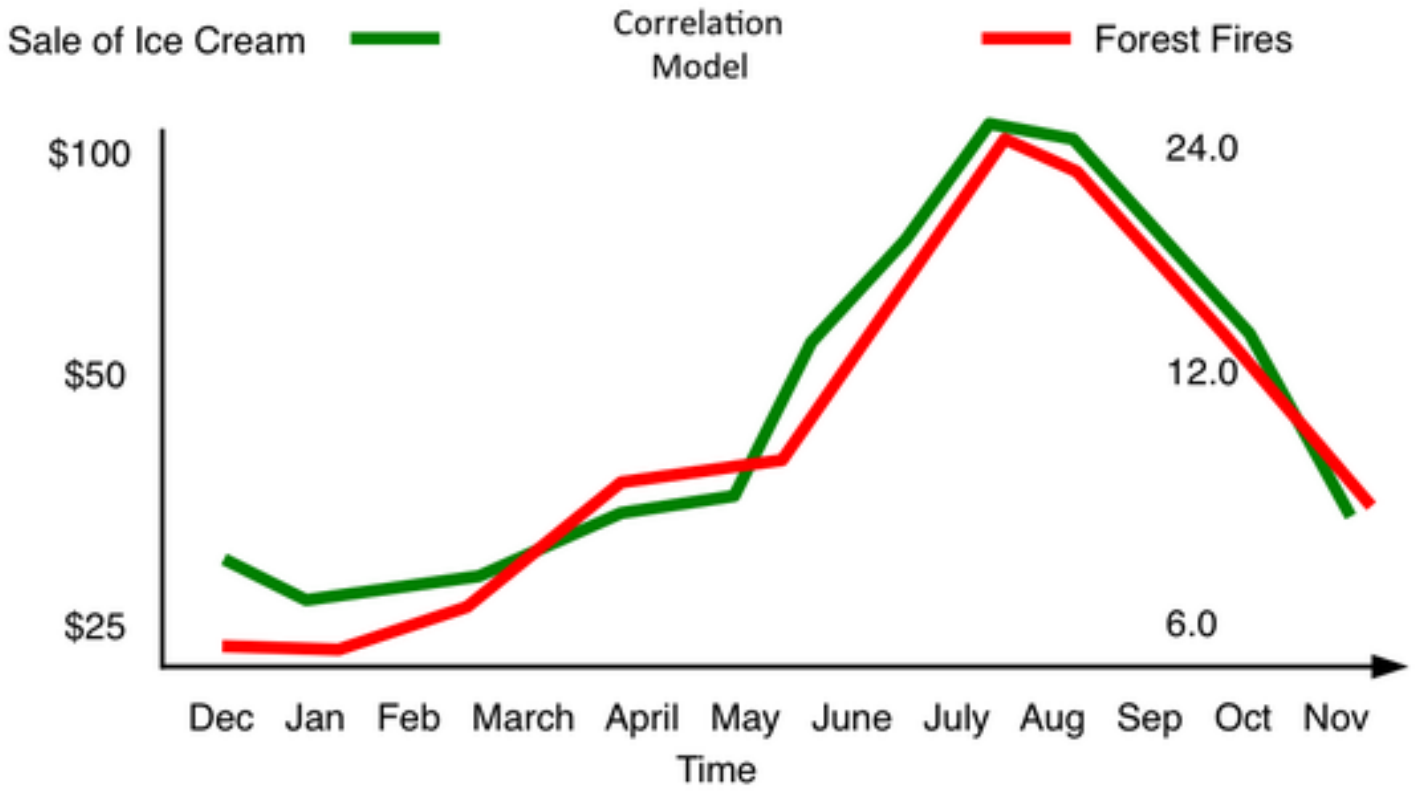
from http://www.tylervigen.com/spurious-correlations

Las relaciones entre los datos pueden ser sorprendente...



from reddit.com

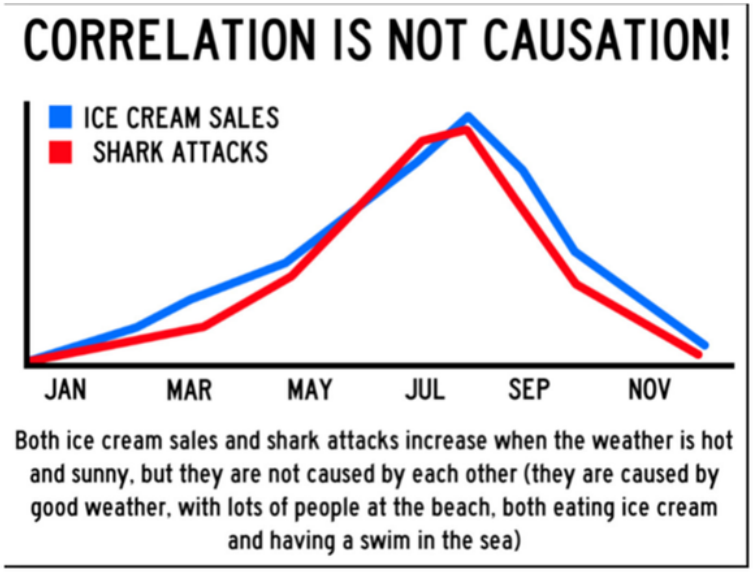
Las relaciones entre los datos pueden ser sorprendente...



from www.decisionskills.com

Importante recordar

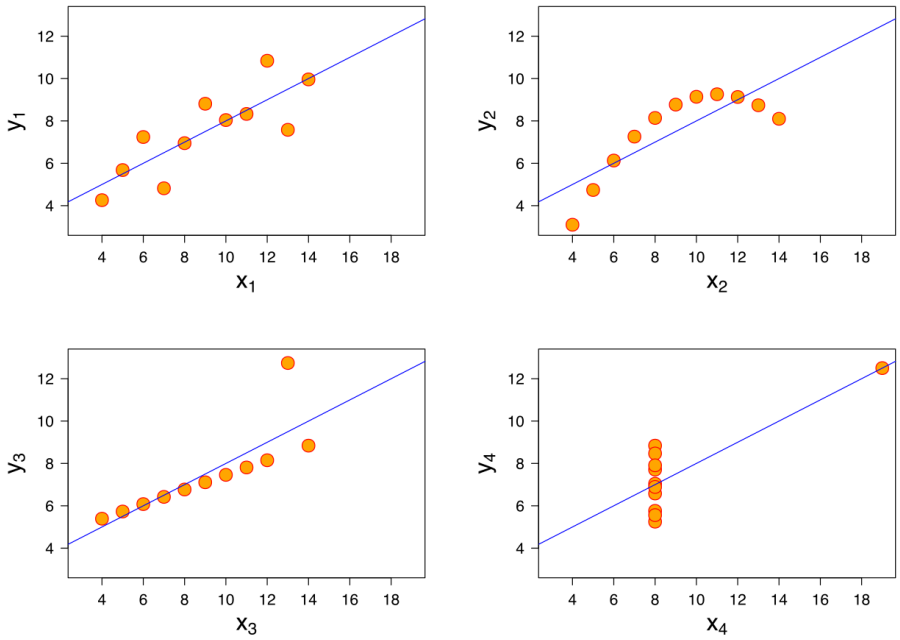
- Asociación (o correlación) entre variables no significa que haya una relación causa-efecto.



from quora.com

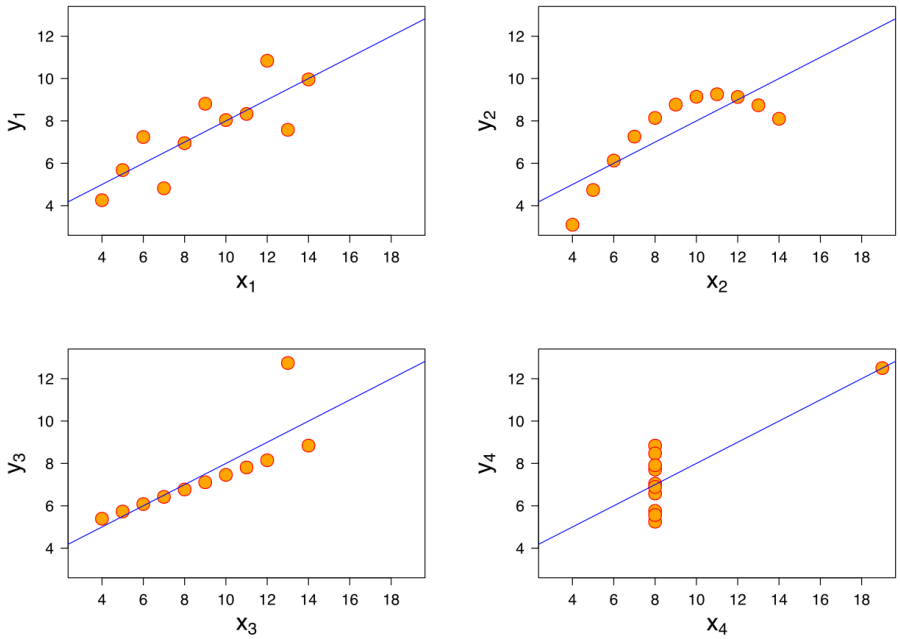
- Más interesantes que encontrar relaciones entre las variables, es entender qué mecanismos las producen.
- **Predecir $\neq$ Explicar!!**

Muy importante visualizar



- ¿Qué valores de x tienen mayor media?
- ¿Qué valores de y tienen mayor media?
- ¿Qué valores de x tienen mayor desviación estándar?
- ¿Qué valores de y tienen mayor desviación estándar?
- ¿Qué gráfica muestra más correlación entre x e y?

Muy importante visualizar



- ¿Qué valores de x tienen mayor media?
- ¿Qué valores de y tienen mayor media?
- ¿Qué valores de x tienen mayor desviación estándar?
- ¿Qué valores de y tienen mayor desviación estándar?
- ¿Qué gráfica muestra más correlación entre x e y?

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places

https://en.wikipedia.org/wiki/Anscombe's_quartet

Francis Anscombe 1973

AVISO 2

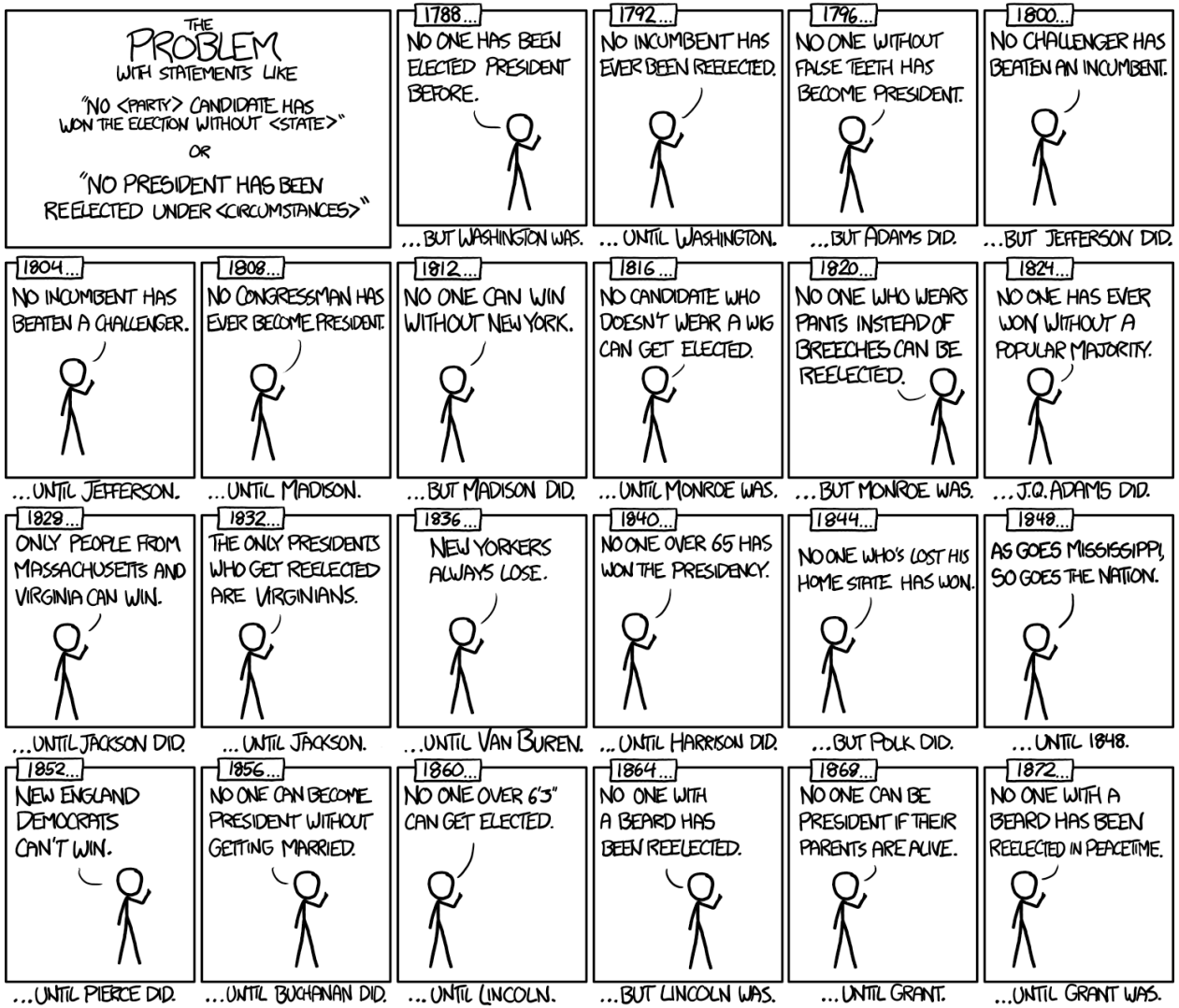
Como siempre, cuidado con querer publicar artículos a toda costa usando la metodología de machine learning.

Típica enfermedad en estadística...

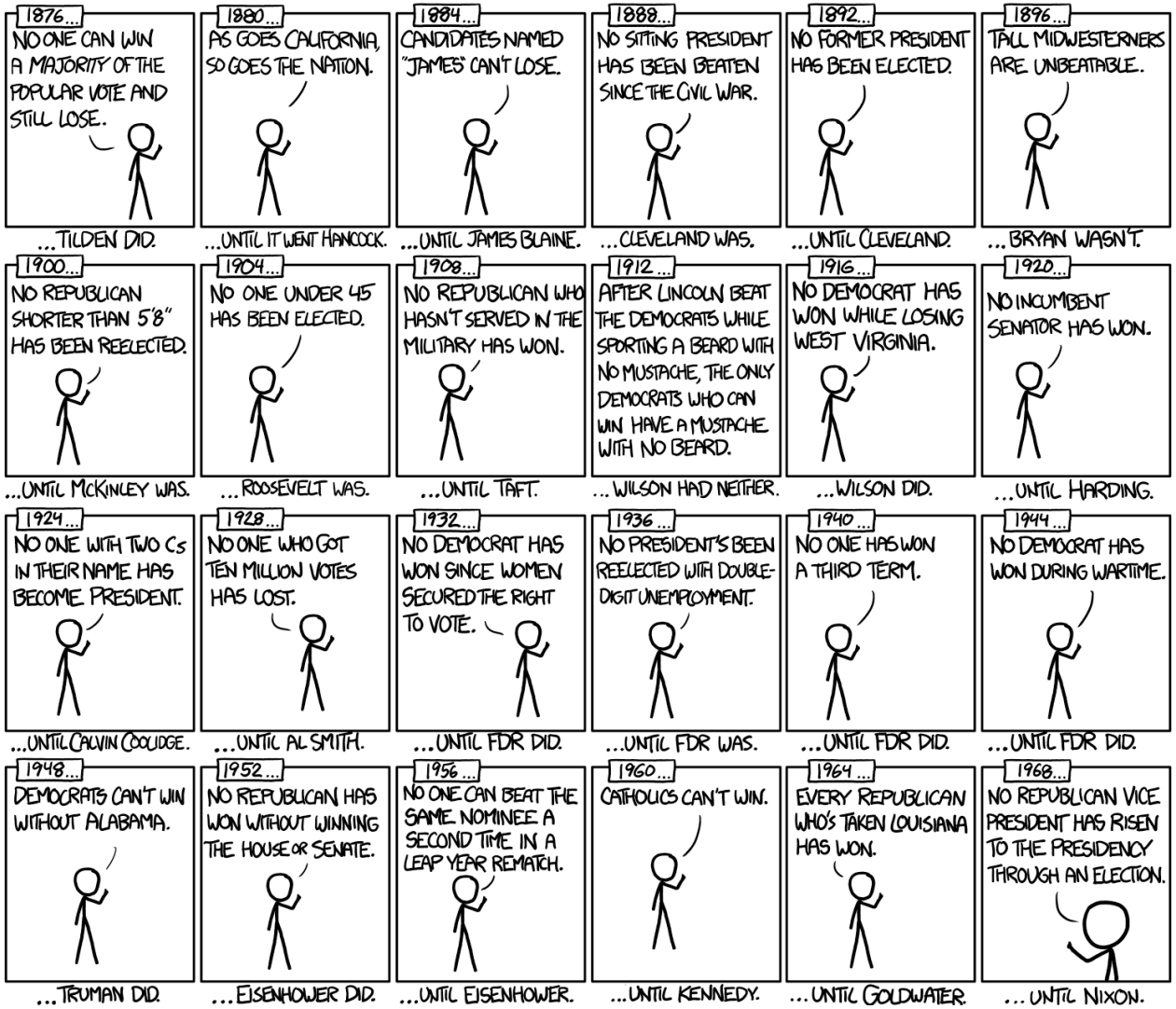
Falsos positivos y tamaño de los efectos pequeños

from garstats.wordpress.com

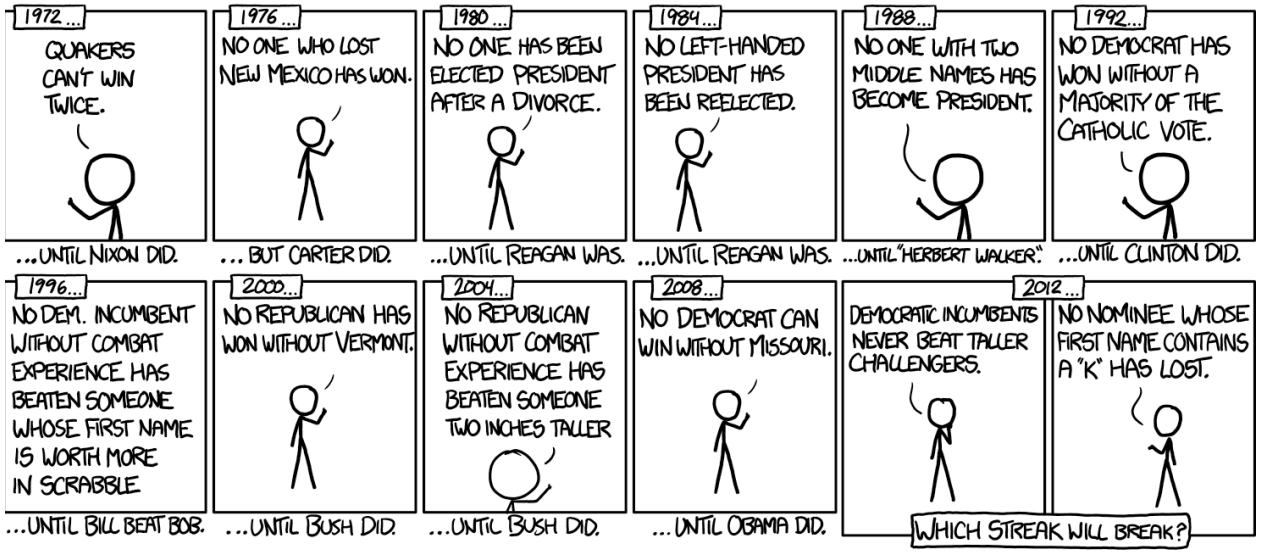
. . . que continua en ML



. . . que continua en ML



. . . que continua en ML

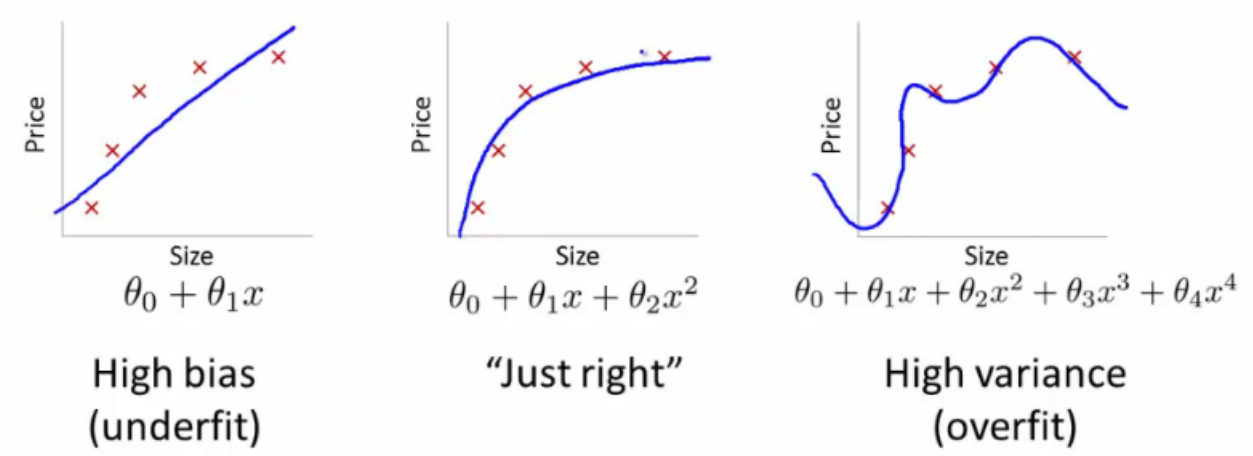


from xkcd.com

OVERFITTING

AVISO 2

- Es muy fácil reportar resultados no generalizables impresionados por el buen rendimiento del algoritmo.
- Scikit ofrece una gran variedad de herramientas para el diagnóstico y prevención de problemas de este tipo y que aprenderemos en este curso.
- Como antes, visualizar los resultados suele ser una de las mejores herramientas para diagnosticar nuestros problemas.



from http://blog.grio.com/2017/03/an-introduction-to-machine-learning.html