Programa de Doctorado en Investigación Biomédica

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

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

ANÁLISIS, PREDICCIÓN Y CLASIFICACIÓN DE DATOS BIOMÉDICOS

Jesús María Cortés

Profesor de Investigación Ikerbasque. Instituto de Investigación Sanitaria Biocruces

jesus.m.cortes@gmail.com

Javier Rasero

Instituto de Investigación Sanitaria Biocruces

jrasero.daparte@gmail.com

Osakidetza

biocruces
health research institute

Estructura del curso

Tema 1: Introducción al Machine Learning (ML)
Tema 2: Aprendizaje Supervisado I. Regresión
Tema 3: Aprendizaje Supervisado II Clasificación
Tema 4: Selección del modelo e hiperparametros y su evaluación
Tema 5: Aprendizaje no Supervisado

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

- **Sorfware**
  - SPSS (en sus distintas versiones)

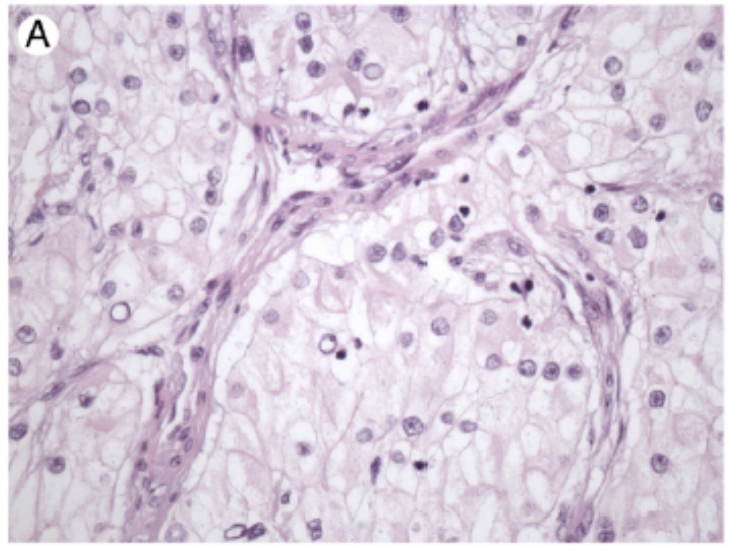
- **Análisis de datos que realiza habitualmente un patólogo**
  - t-test
  - Chi2
  - Correlaciones (Spearman, Pearson)
  - Regresión univariante
  - Regresión multivariante
  - Curvas de supervivencia
  - Curvas de mortalidad

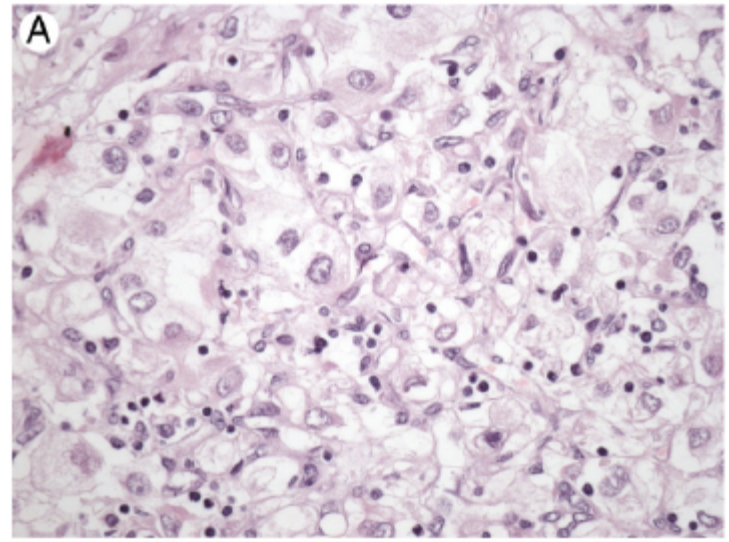
Los clínicos en general manejan de forma subóptima e imperfecta los datos

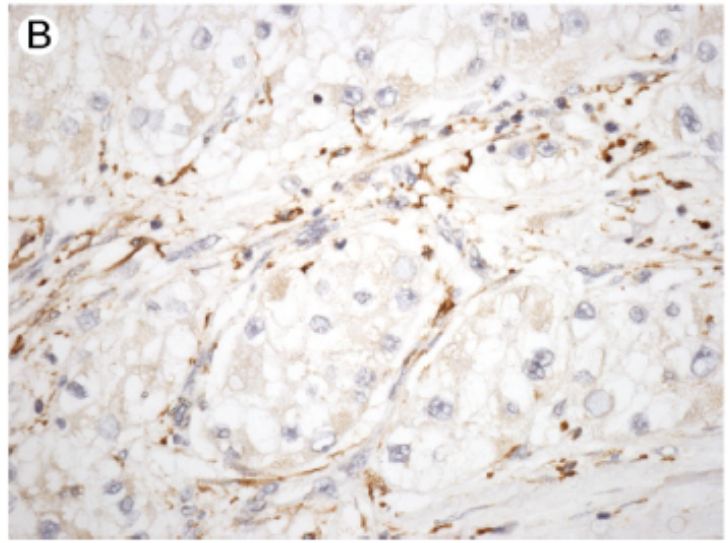
- **Sorfware**
  - SPSS (en sus distintas versiones)

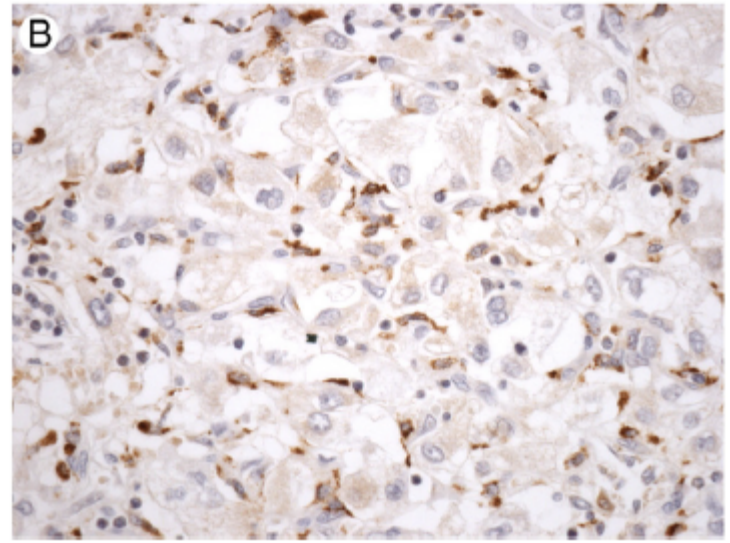
- **Análisis de datos que realiza habitualmente un patólogo**
  - t-test
  - Chi2
  - Correlaciones (Spearman, Pearson)
  - Regresión univariante
  - Regresión multivariante
  - Curvas de supervivencia
  - Curvas de mortalidad

… Y versiones no paramétricas


A


A


B


B

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 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.

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

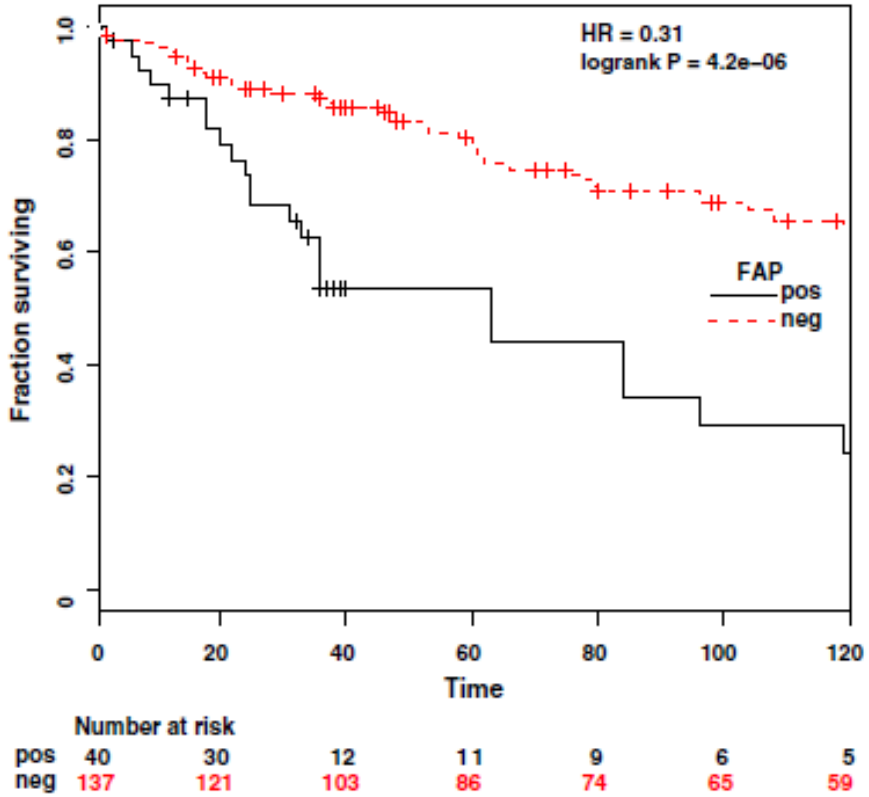
Table 2 Univariate regression analysis

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.



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				
58	M	CCRCC	3	5				
75	F	CCRCC	3	8				
44	M	CCRCC	2	2,8				

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.

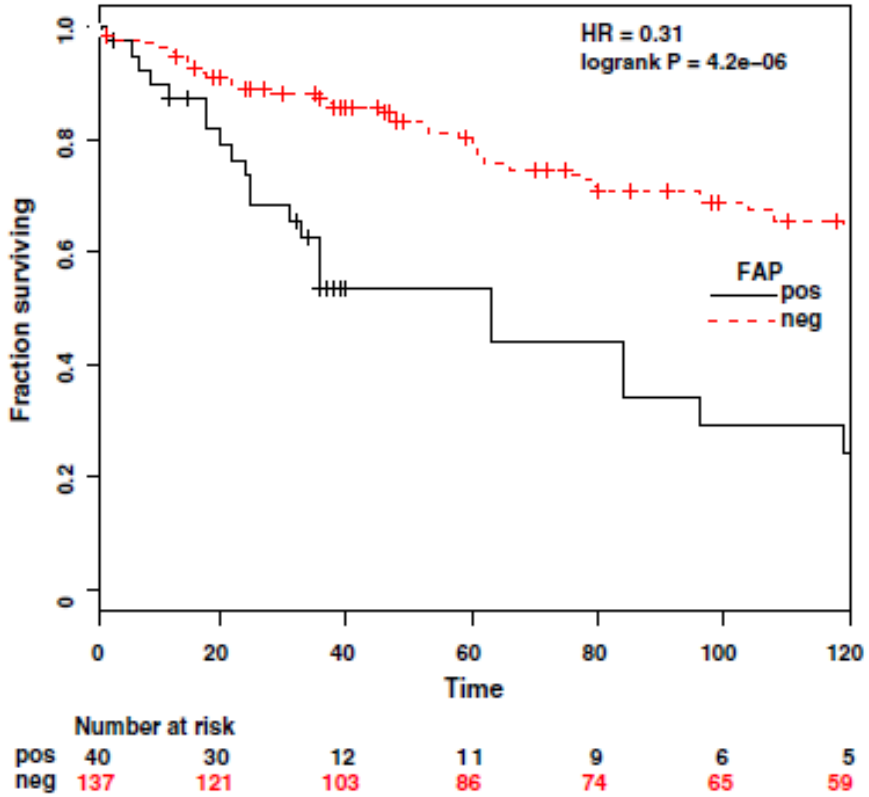


Table 1 Log-rank *P* for 5-, 10-, and 15-year survival

	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.

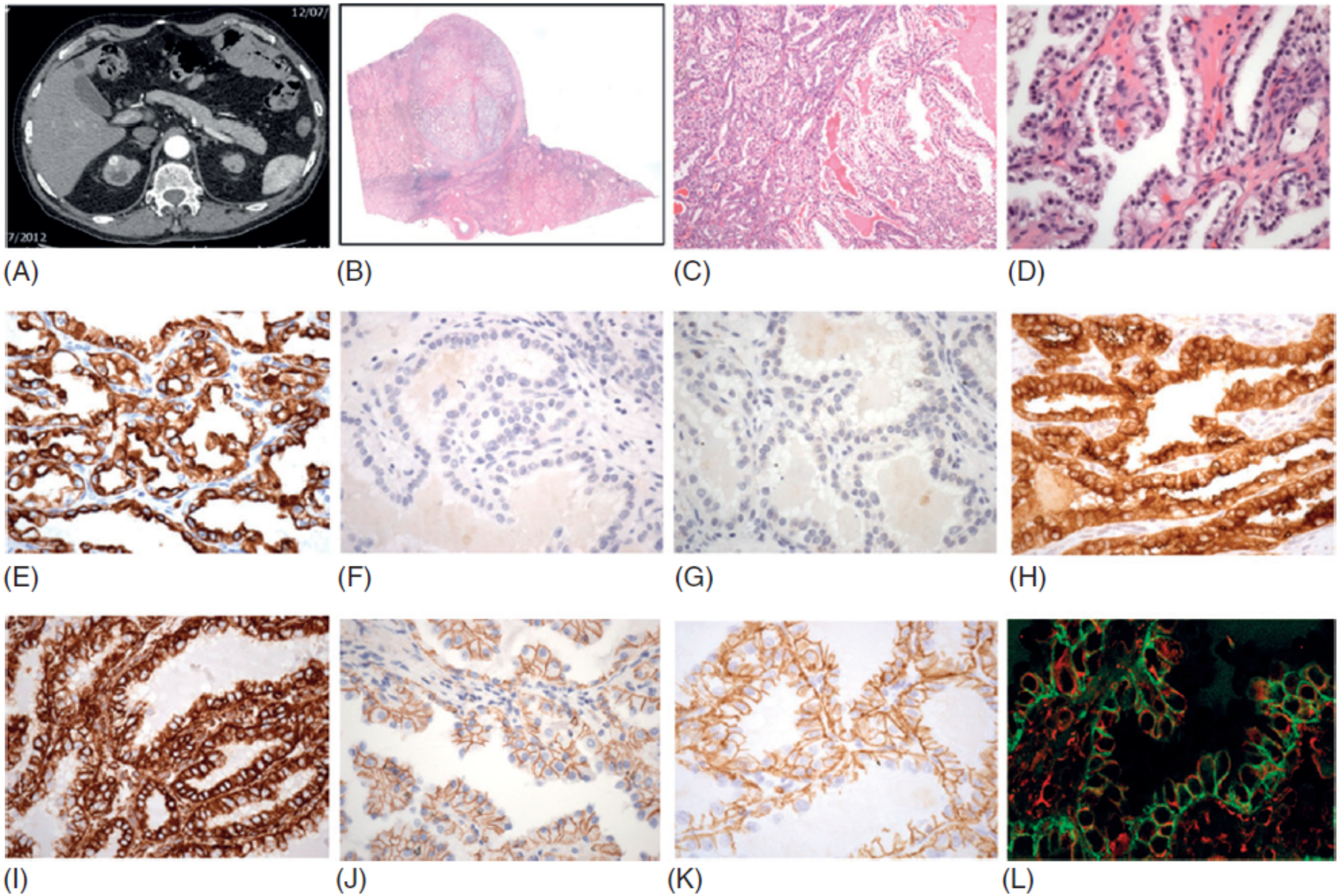
Human Pathology (2016) 54, 100–105

Human
PATHOLOGY
www.elsevier.com/locate/humpath

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]


12/07/
7/2012
(A)
(B)
(C)
(D)
(E)
(F)
(G)
(H)
(I)
(J)
(K)
(L)

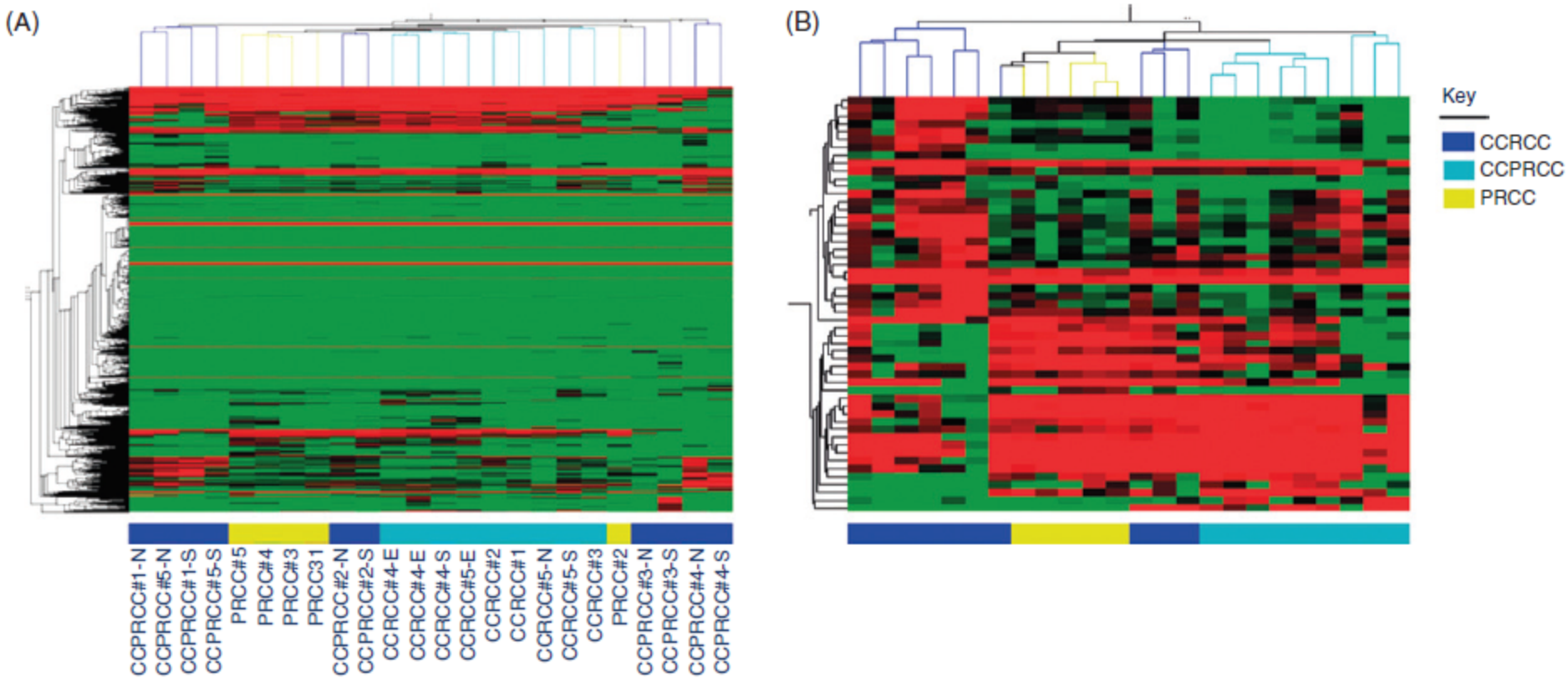


Figure 2. The small non-coding RNA expression profile of CCPRCC is distinct from either CCRCC or PRCC cases. Unsupervised cluster analysis of (A) mature miRNA expression data and (B) cluster analysis of samples according to expression levels of 53 differentially expressed miRNAs (> two-fold, $p < 0.05$). Expression levels of (C) *miR-155*, (D) *miR-210*, (E) *miR-34a*, (F) *miR-182*, (G) *miR-16*, (H) *miR-200a*, (I) *miR-200b*, (J) *miR-200c*, (K) *miR-141* and (L) *miR-429* in CCPRCC ($n = 22$), CCRCC ($n = 14$), PRCC ($n = 9$) cases and controls ($n = 5$) measured by qRT–PCR. *p* values were calculated by Mann–Whitney independent *t*-test.

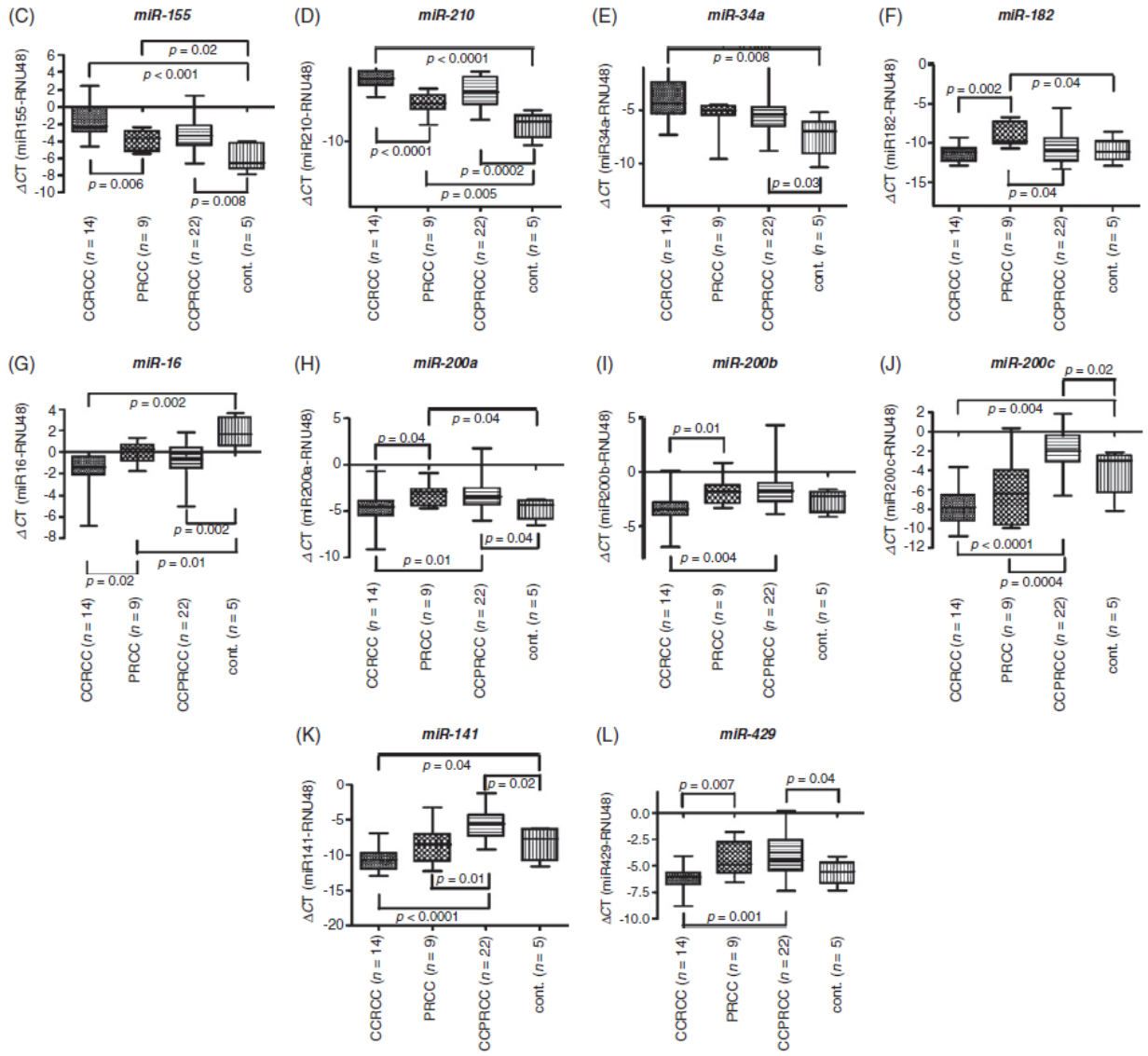


Figure 2. The small non-coding RNA expression profile of CCPRCC is distinct from either CCRCC or PRCC cases. Unsupervised cluster analysis of (A) mature miRNA expression data and (B) cluster analysis of samples according to expression levels of 53 differentially expressed miRNAs (> two-fold, $p < 0.05$). Expression levels of (C) *miR-155*, (D) *miR-210*, (E) *miR-34a*, (F) *miR-182*, (G) *miR-16*, (H) *miR-200a*, (I) *miR-200b*, (J) *miR-200c*, (K) *miR-141* and (L) *miR-429* in CCPRCC ($n = 22$), CCRCC ($n = 14$), PRCC ($n = 9$) cases and controls ($n = 5$) measured by qRT–PCR. *p* values were calculated by Mann–Whitney independent *t*-test.

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]

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]

Generalmente, datos más sofisticados permite publicar en revistas de mayor impacto

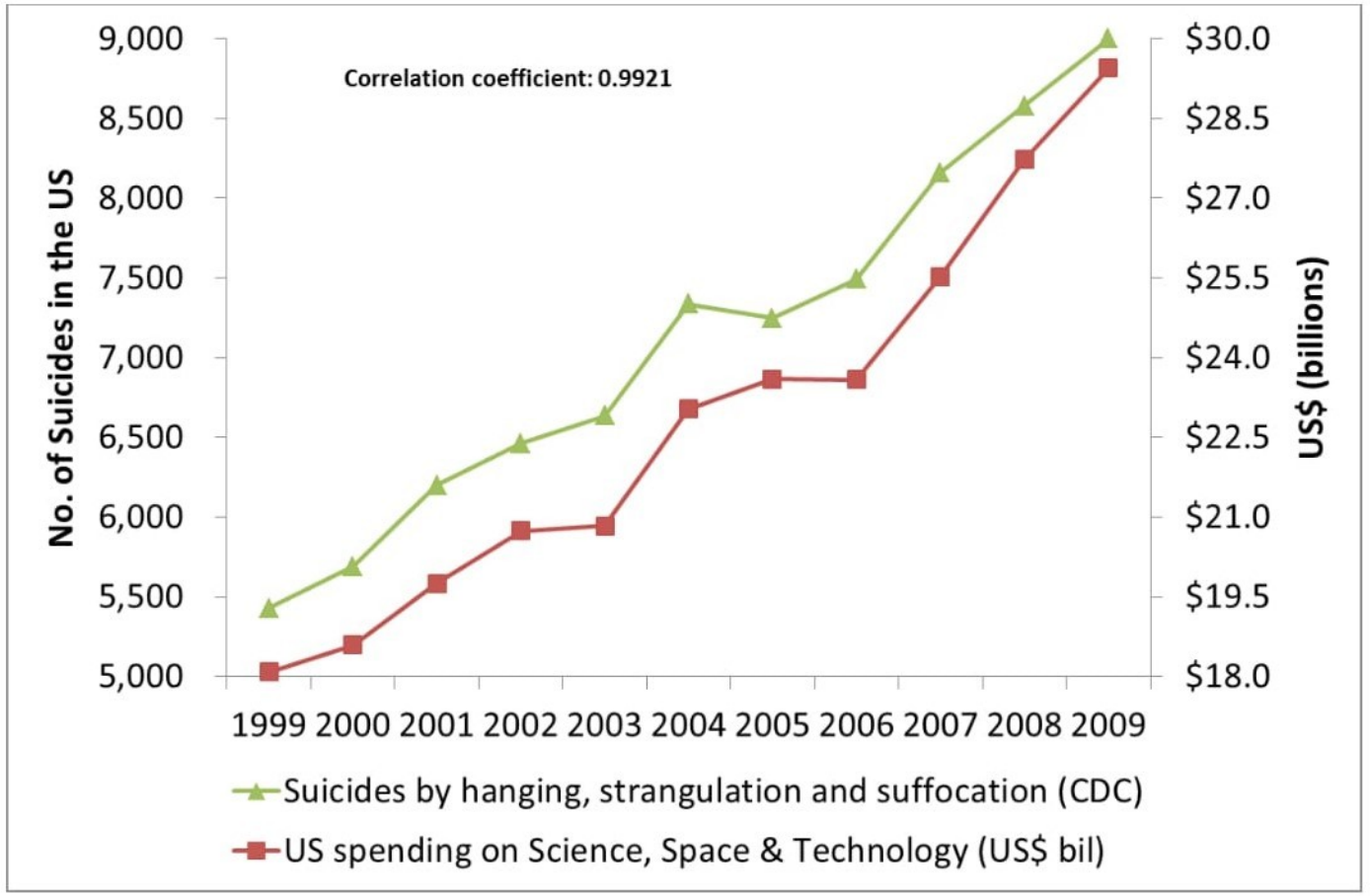
Rutina científica en anatomía patológica

age	sex	G	G agrupado	diam	T	histol	snail	slug	zeb1	twist	b-cat	e-cadh	viment	p110a	6H2	425	17A	akt	setD2
88	M	hom	hom	6,5	pT1b	hom	hom neg	hom neg	hom neg	hom neg	hom neg	hom neg	hom neg	het	hom neg	het	het	hom neg	hom neg
42	M	het	het	6,6	pT1b	het	het	het	het	hom neg	hom pos	het	het	hom pos	het	hom pos	hom pos	hom neg	hom pos
75	M	hom	hom	3,5	pT1a	hom	het	hom neg	het	het	hom pos	het	hom pos	hom pos	het	het	hom pos	hom neg	hom pos
74	M	het	hom	5,5	pT1b	hom	het	hom neg	het	het	hom pos	hom neg	het	hom pos	hom neg	hom pos	hom pos	hom neg	het
79	M	het	hom	3,5	pT1a	hom	het	hom neg	het	hom neg	hom pos	hom neg	hom pos	hom pos	hom neg	het	hom pos	hom neg	het
67	M	het	het	5	pT1b	het	het	hom neg	hom neg	het	het	hom neg	het	hom pos	hom neg	hom pos	hom pos	het	het
59	M	het	hom	4	pT1a	hom	het	hom neg	het	het	hom pos	hom neg	het	het	hom neg	het	hom pos	hom neg	het
55	F	het	het	2,5	pT1a	hom	het	het	het	hom neg	het	hom neg	het	hom pos	hom neg	het	hom pos	hom neg	het
59	M	het	hom	5	pT1b	hom	het	hom neg	het	het	hom pos	hom neg	het	hom pos	hom neg	hom pos	hom pos	het	hom pos
85	F	het	hom	4	pT1a	hom	het	hom neg	het	hom neg	hom pos	hom pos	het	hom pos	het	hom pos	hom pos	het	hom pos
70	M	hom	hom	6	pT1b	hom	het	hom neg	hom neg	het	het	hom neg	het	hom pos	hom neg	hom pos	hom pos	het	hom pos
68	F	het	hom	2,8	pT1a	hom	het	hom neg	het	hom neg	hom neg	hom neg	het	het	hom neg	het	hom pos	het	het

Las relaciones entre los datos pueden ser realmente sorprendentes…

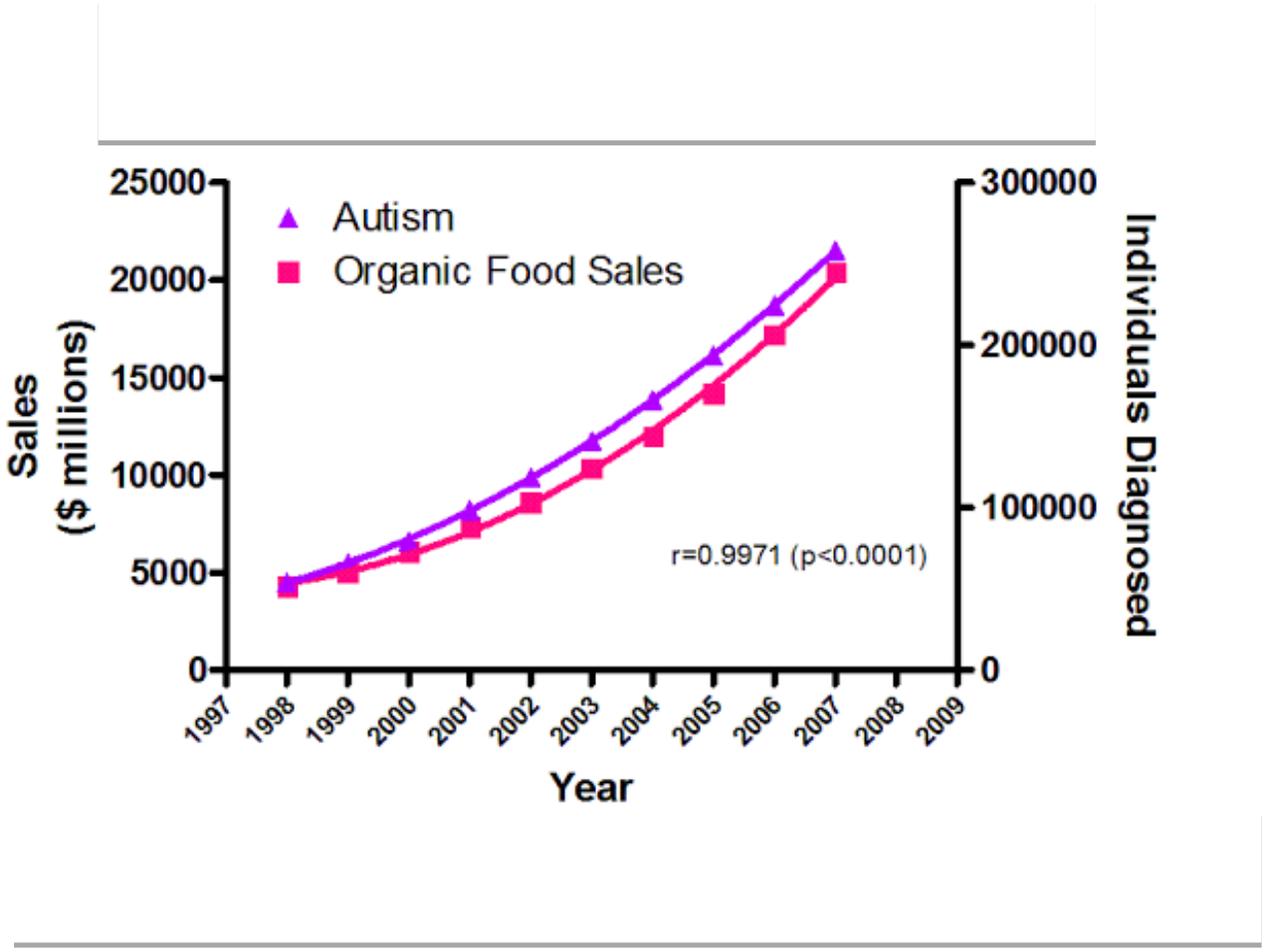
Suicidios vs Inversión en ciencia

Suicidios vs Inversión en ciencia



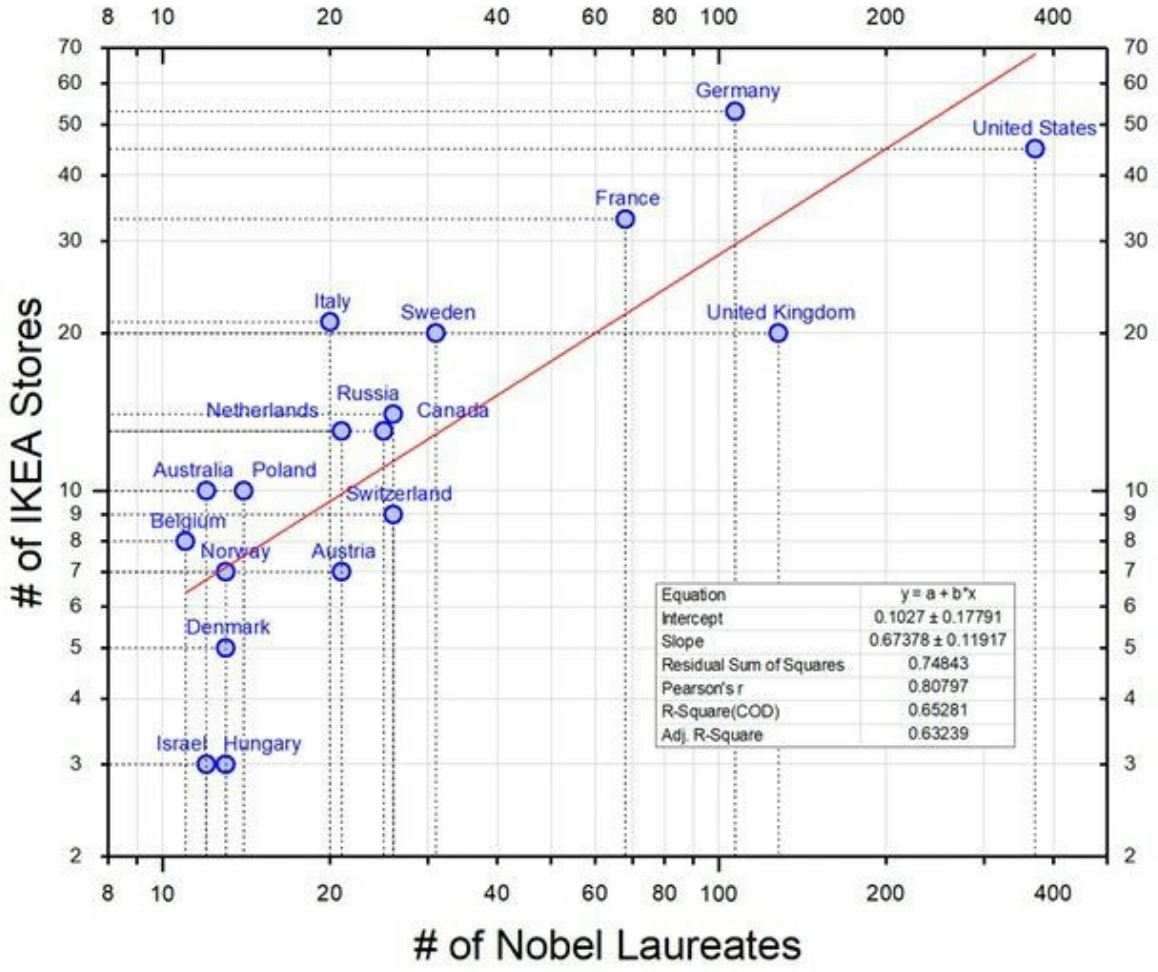
Autismo vs Productos de alimentación orgánicos

Autismo vs Productos de alimentación orgánicos



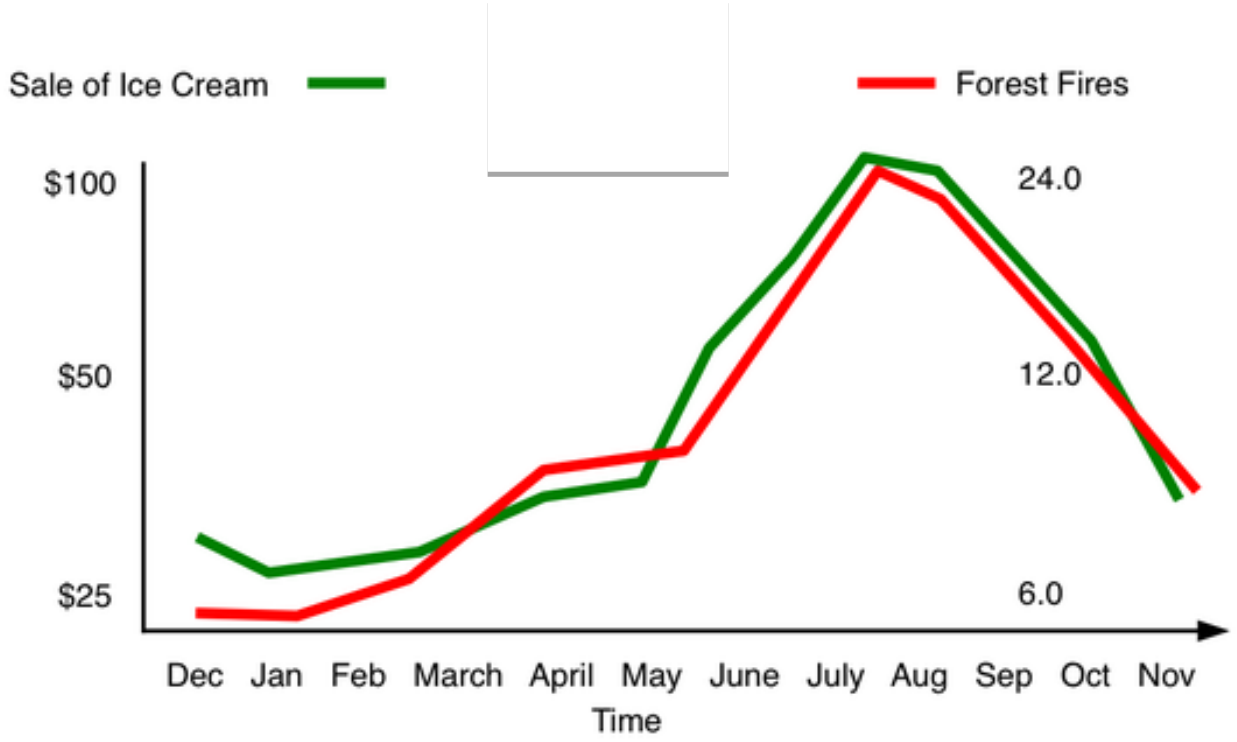
Tiendas de IKEA vs Premios Nobel

Tiendas de IKEA vs Premios Nobel



Venta de helados vs Incendios forestales

Venta de helados vs Incendios forestales



Importante recordar:

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

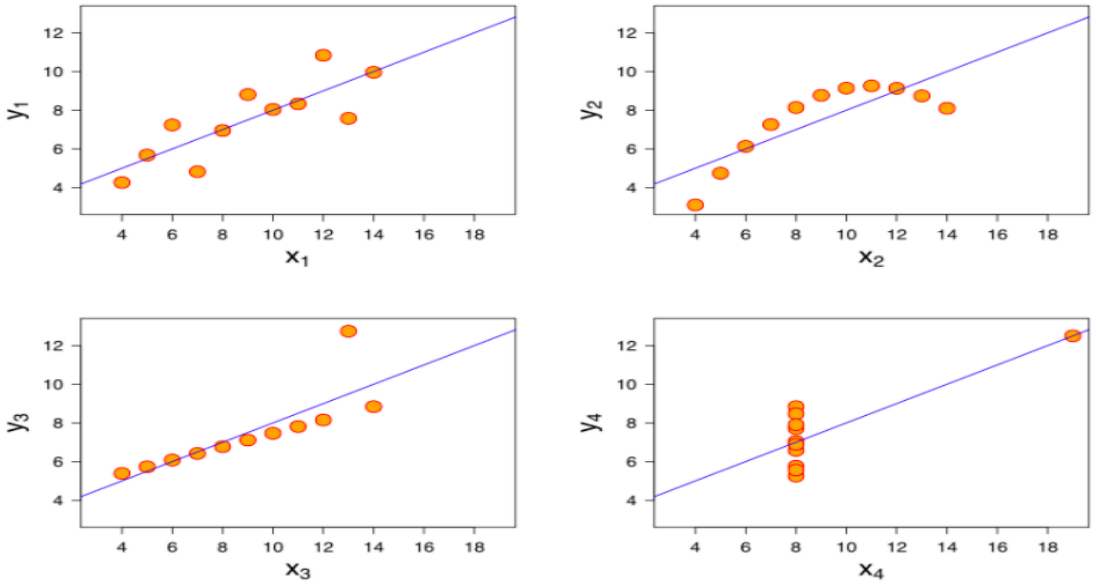
Importante recordar:

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

Más interesante que encontrar las relaciones entre las variables, es entender qué mecanismos las producen

Y muy importante es visualizar:

Y muy importante es visualizar:



¿Qué valores de x tienen mayor media?

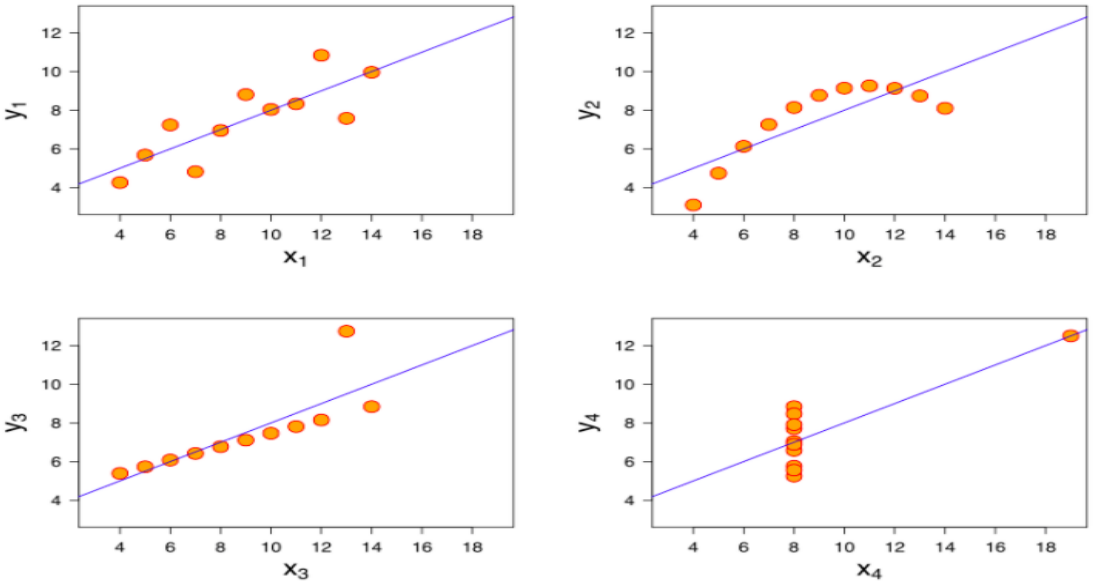
¿Qué valores de y tienen mayor media?

¿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?

Y muy importante es 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?

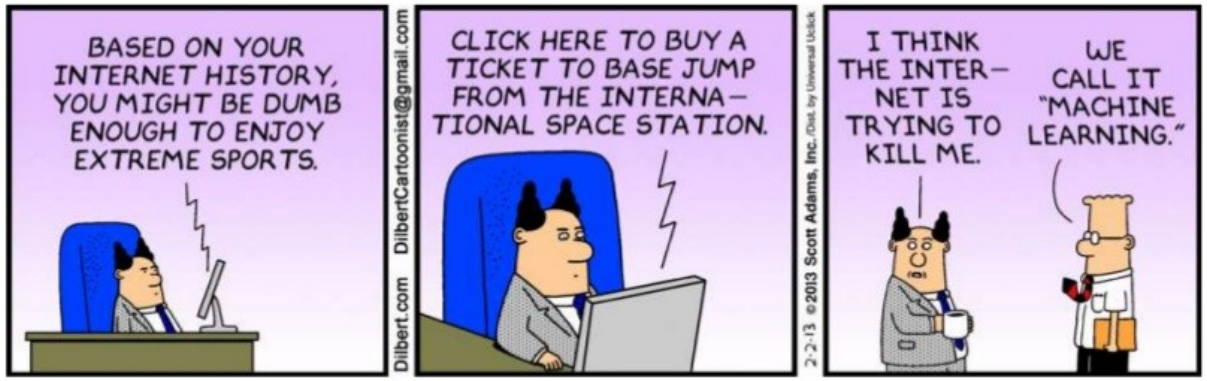
Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	4.125	plus/minus 0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of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

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

F. Anscombe 1973

Ciencias de los datos (Machine Learning)

WHAT IS MACHINE LEARNING?

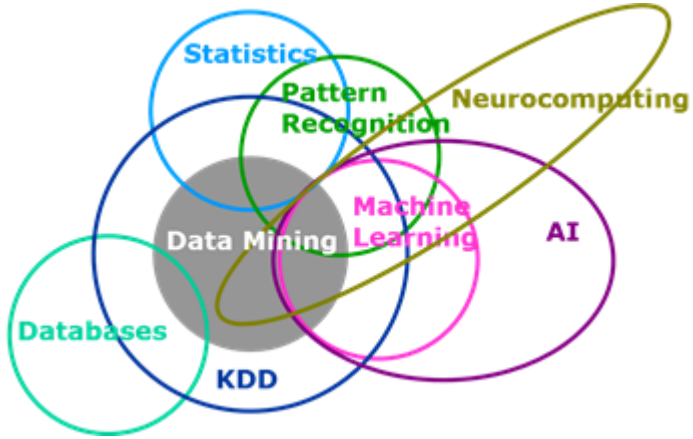


Data

Algorithm

Model

$$f(\mathbf{x})$$



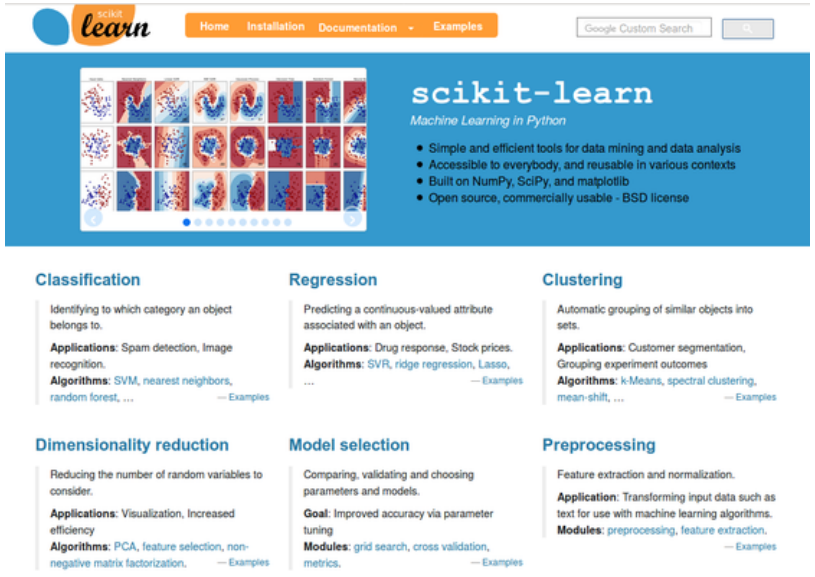
Ciencias de los datos (Machine Learning)

- 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)

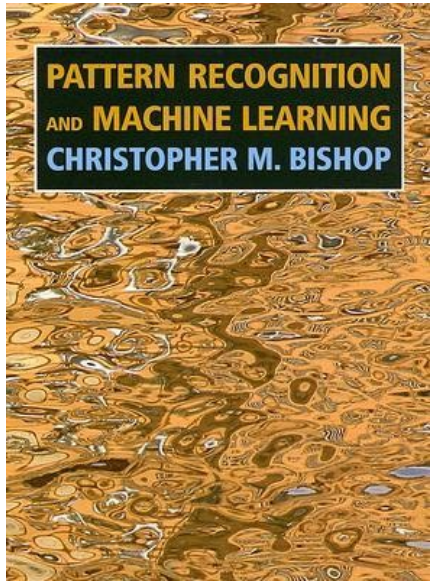
- En este curso, nosotros usaremos **scikit-learn,** que es una librería de machine learning escrita en python:
- Simple y efficiente.
- Accessible para todo el mundo y reutilizable.
- Open source.
- Tiene implementados una gran cantidad de algortimos y está en continuo desarrollo.
- Gran documentación.

Referencias generales 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:Data Mining, Inference, and Prediction.** 2º edition, Springer, 2009