Ciencia de datos: El trabajo más sexy del siglo XXI

XIV Semana de la Ciencia y la Tecnología. IES Floridablanca, Murcia

Antonio Maurandi López

amaurandi@um.es

Universidad de Murcia

6 de febrero de 2018

Introducción

¿Quién soy?

cv informal

1999 Licenciatura en CC Matemáticas

Estudios de Ing. Informática

Programador

Administrador de BBDD

Programador *freelance*

2005 Servicio de Apoyo a la Investigación

Más y más formación (*cursos*, *másteres*, *doctorado*, *etc...*)

2017 Actualmente profesor en la Facultad de Educación

¿Por qué este título?

Ciencia de datos: El trabajo más sexy del siglo XXI

Data scientist: the sexiest job of the 21st century. Davenport y Patil - Harvard Business Review, 2012

El caso de Linkedin

https://es.linkedin.com/

"People You May Know"

2006. Jonathan Goldman PhD en Física de Stanford, comenzó a probar que pasaría si se presentara a los usuarios, personas con las que aún no se habían conectado, pero que parecía probable que se conocieran. Investigó patrones que le permitían predecir en qué redes caería un perfil dado.

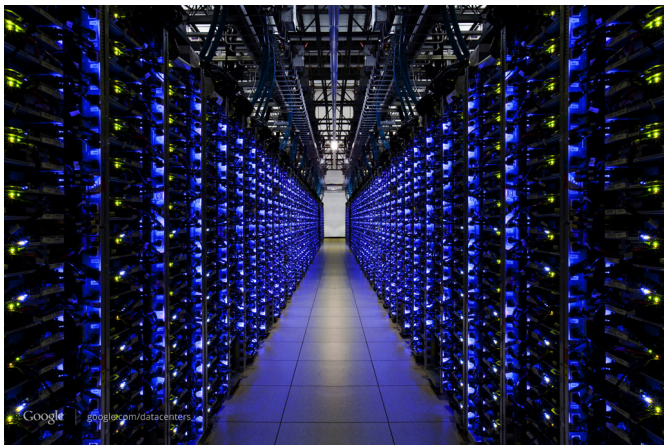
Científicos y científicas de datos

Un nuevo perfil profesional

Datos ubicuos

- Nunca antes ha habido tantos datos disponibles

Desarrollo tecnológico

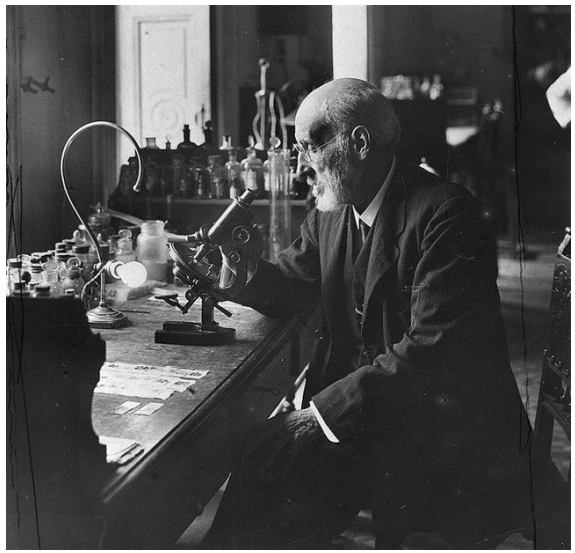


hay que sacarles partido!

Un nuevo perfil profesional

Científico o científica de datos Es alguien que saca significado de los datos, traduce grandes volumenes de datos en información comprensible que aporta valor.

1. Trabaja con hipóteis, que trata de probar
2. Hipóteis basadas en DATOS



¿Qué es un científico/a de datos?

Una persona formada en las ciencias **matemáticas** y las **estadísticas** que domina la **programación** y sus diferentes lenguajes, ciencias de la computación y analítica. Debe dominar la tecnología y las bases de datos para modificar y mejorar la orientación de los negocios de la empresa para la que trabaja.

- parte de hipótesis
- responde preguntas, usando datos
- prueba las hipótesis

- Escasez: Cuesta encontrar perfiles adecuados
- Ya existían de antes

Herramientas

Fundamentalmente un `Data Scientist` trabaja las *hipótesis* sobre datos de dos maneras:

1. Visualización
2. Algoritmos avanzados

Visualización

3 funciones

Exploración

- Problema: Poco conocimiento sobre el comportamiento de los datos
- Tarea: Generar hipótesis

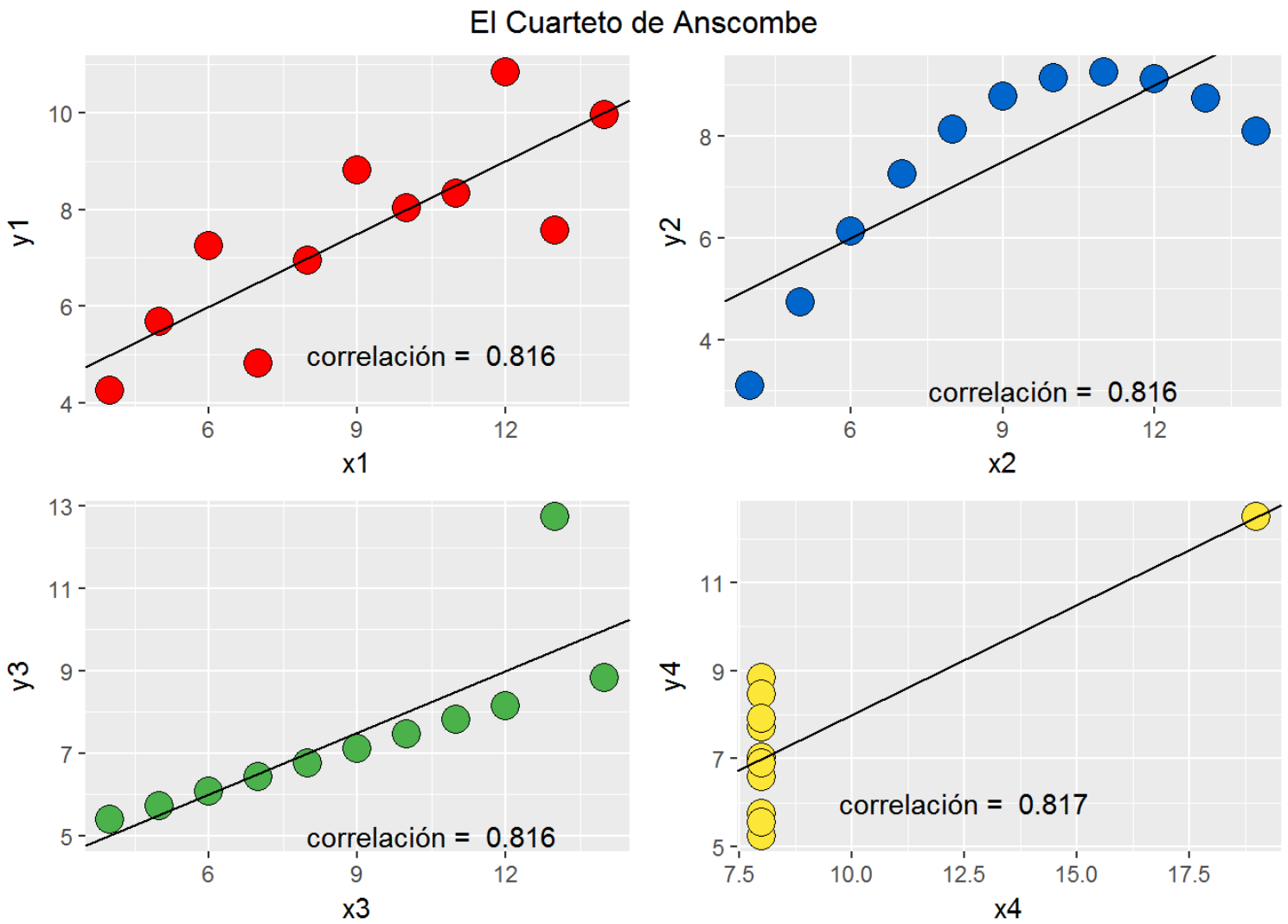
Comunicación

- Problema:cierto conocimiento sobre los datos
- Tarea: Presentarlos visualmente de una manera intuitiva/comprensiva

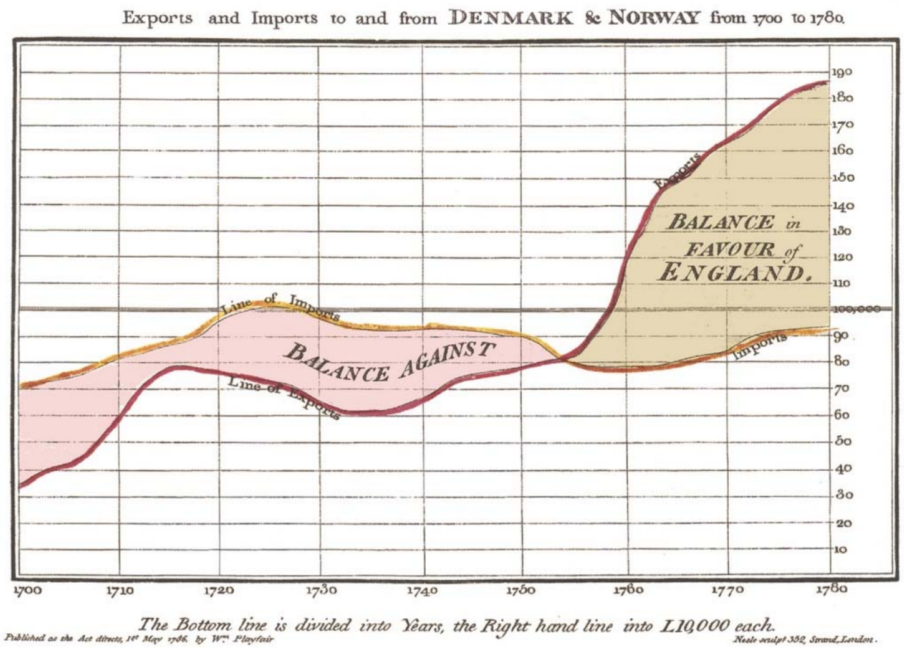
Confirmación

- Problema: Se tiene/formula hipótesis sobre el *comportamiento que subyace* a los datos
- Tarea: Confirmar o rechazar hipótesis

El cuarteto de Anscombe (explorar)

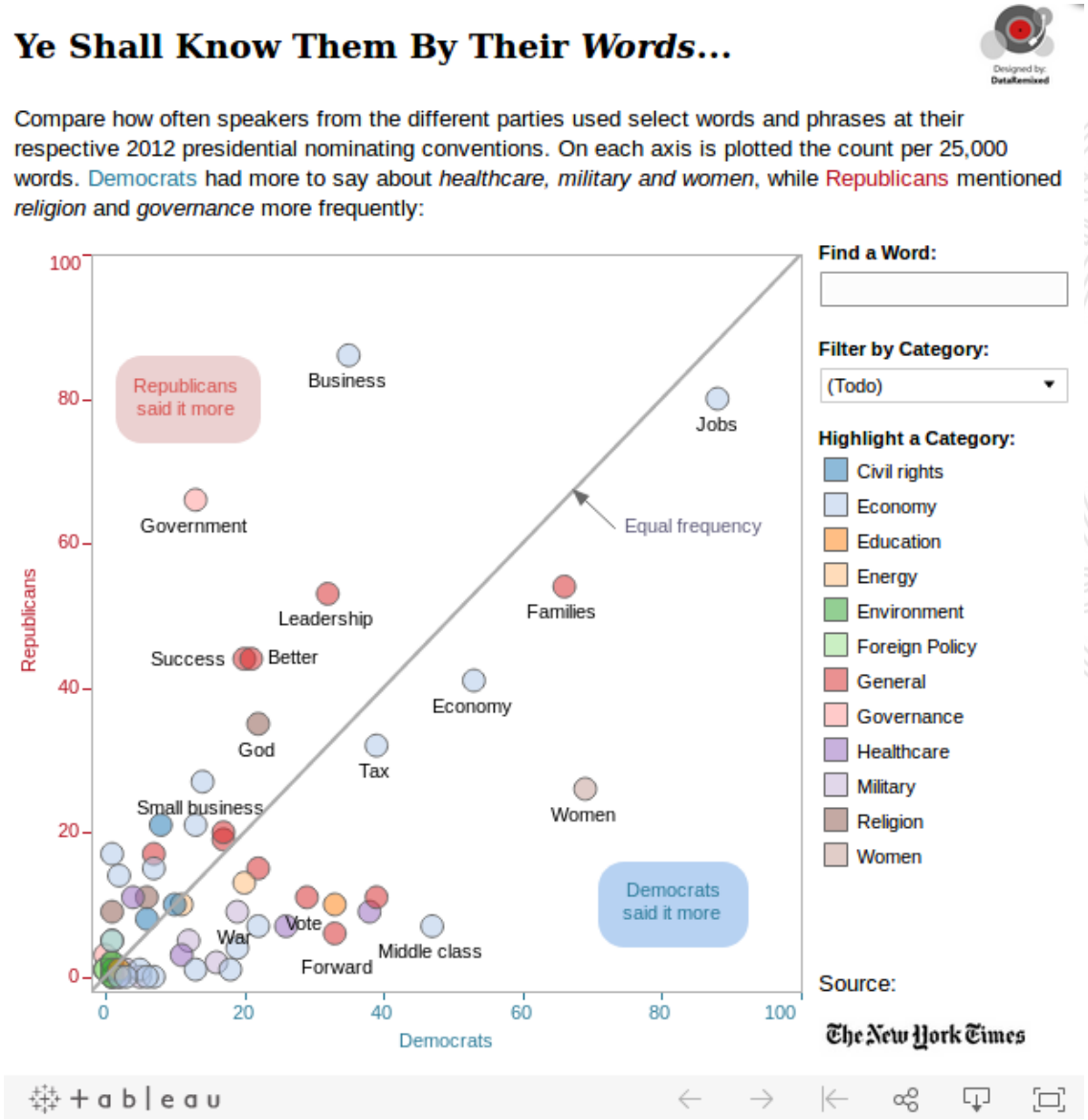


Una nueva forma de razonar



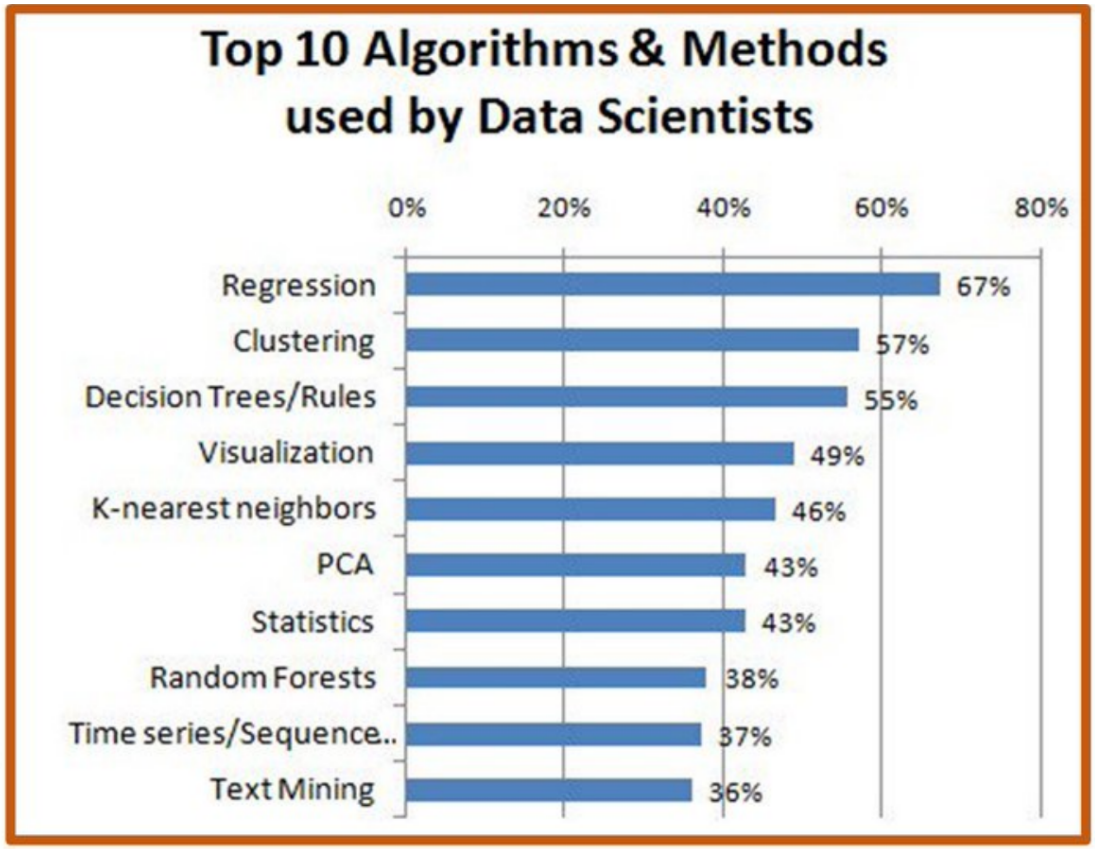
Izda: 1786. A William Playfair se atribuyen los primeros gráficos de carácter estadístico: gráficos de barras, gráficos de sectores y series temporales. *Decha*: Mapa realizado por el doctor John Snow durante la epidemia de cólera que sufrió Londres en 1854.

Ejemplo

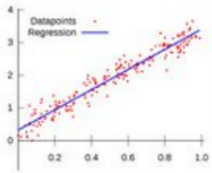

Ye Shall Know Them By Their Words...
Compare how often speakers from the different parties used select words and phrases at their respective 2012 presidential nominating conventions. On each axis is plotted the count per 25,000 words. Democrats had more to say about healthcare, military and women, while Republicans mentioned religion and governance more frequently:
Republicans
Democrats
Republicans said it more
Democrats said it more
Equal frequency
Business
Jobs
Government
Leadership
Families
Success
Better
Economy
God
Tax
Small business
Women
Vote
War
Middle class
Forward
Find a Word:
Filter by Category:
(Todo)
Highlight a Category:
Civil rights
Economy
Education
Energy
Environment
Foreign Policy
General
Governance
Healthcare
Military
Religion
Women
Source:
The New York Times
+ableau

Algoritmos avanzados

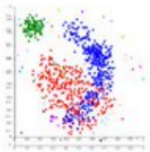
Los más usados



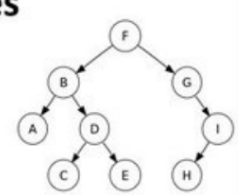
1. Regresion



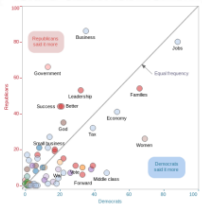
2. Clustering



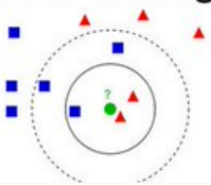
3. Decision Trees/ Rules



4. Visualization

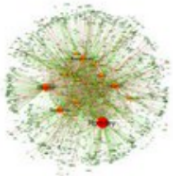


5. K-nearest neighbors

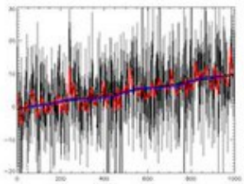


6. PCA

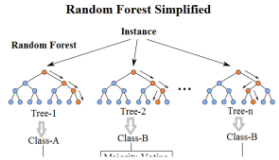
10. Text mining



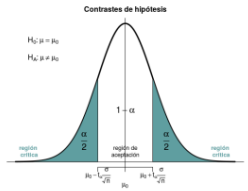
9. Time Series/sequence



8. Random forest



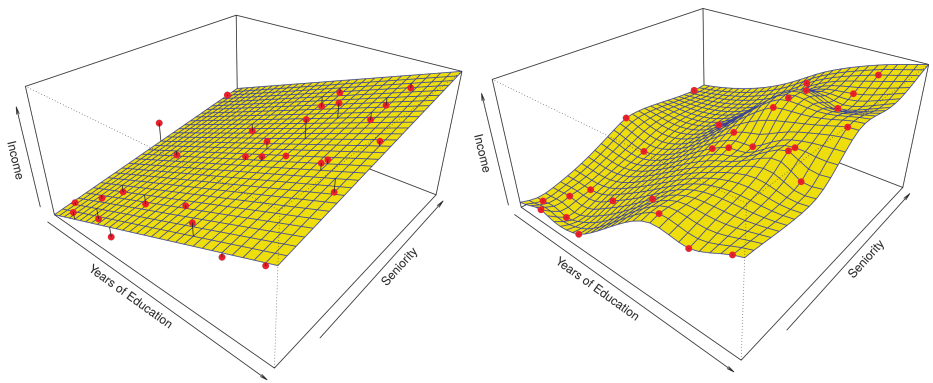
7.Statistics



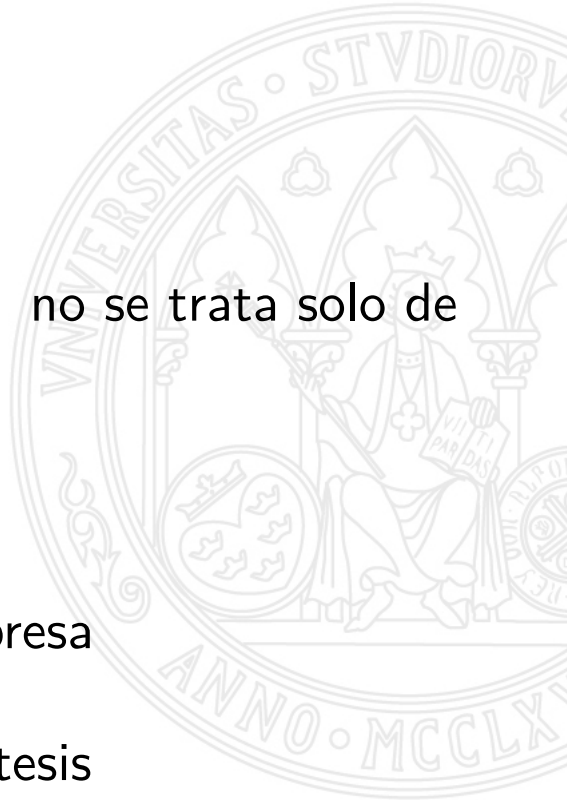
https://www.kdnuggets.com/2016/09/poll-algorithms-used-data-scientists.html

No free lunch in machine learning and statistics

1. Modelos paramétricosy no paramétricos
2. Interpretabilidad *versus* flexibilidad
3. Aprendizaje supervisado frente a no supervisado
    - predicción
    - inferencia
    - regresión y clasificación



Qué NO es un/a cientíifico/a de datos



- No es un programador/a
  - no es un prog. de `JAVA` que sabe usar `Hadoop`,... no se trata solo de habilidades técnicas
- No es un/a analista de *Bussines Intelligent*
  - Visualizan datos
  - Identificar errores, debilidades y fuertes de la empresa
  - Conocen bien el negocio (*domain knowledge*)
  - Pero no emplean algoritmos para probar sus hipótesis

No existe un titulo universitario concreto

Definición conjuntista de un Data Scientist

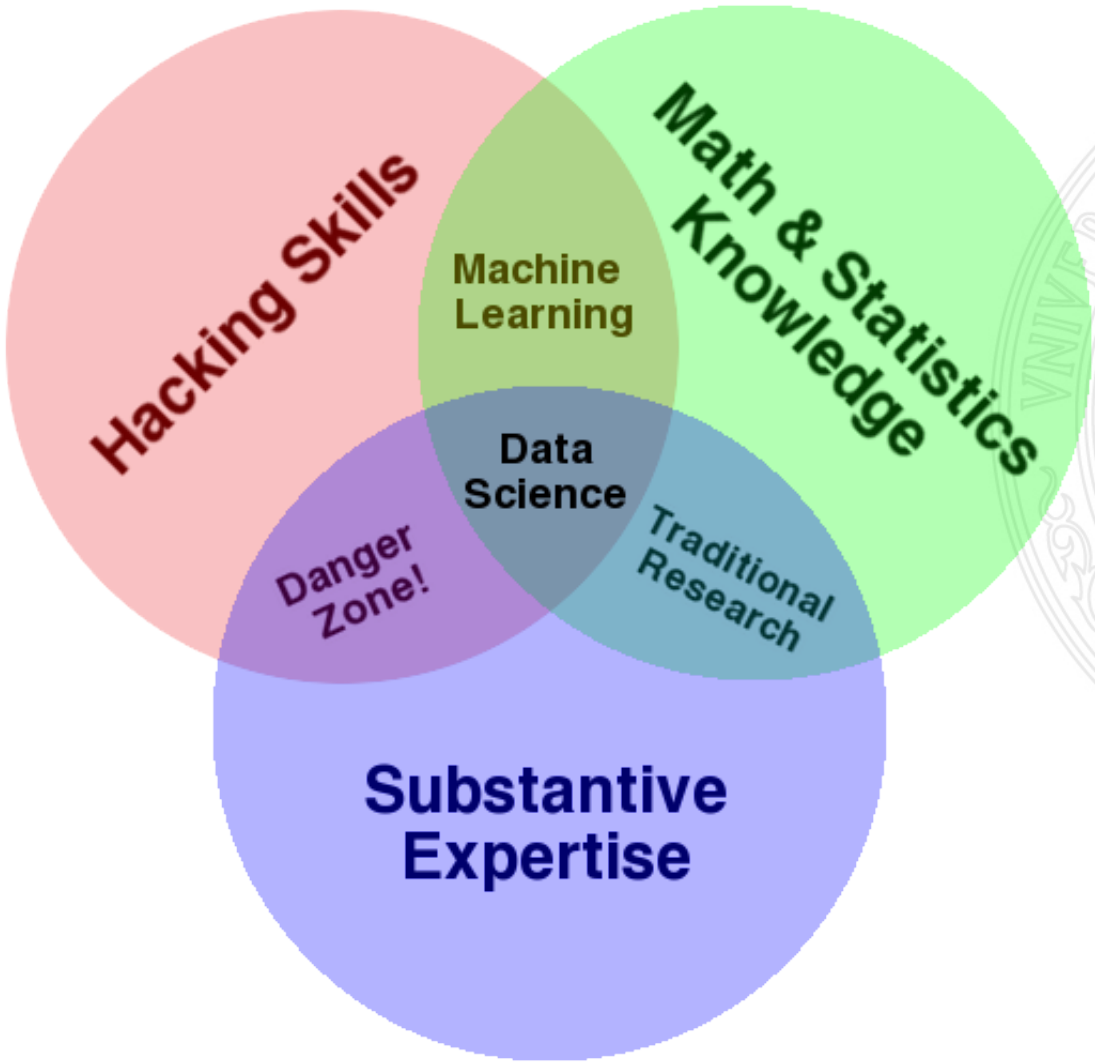


Figura 2: Drew Conway(2010). Diagrama de Venn del *"Científico de datos"*

Volvemos a la definición de Data Scientist

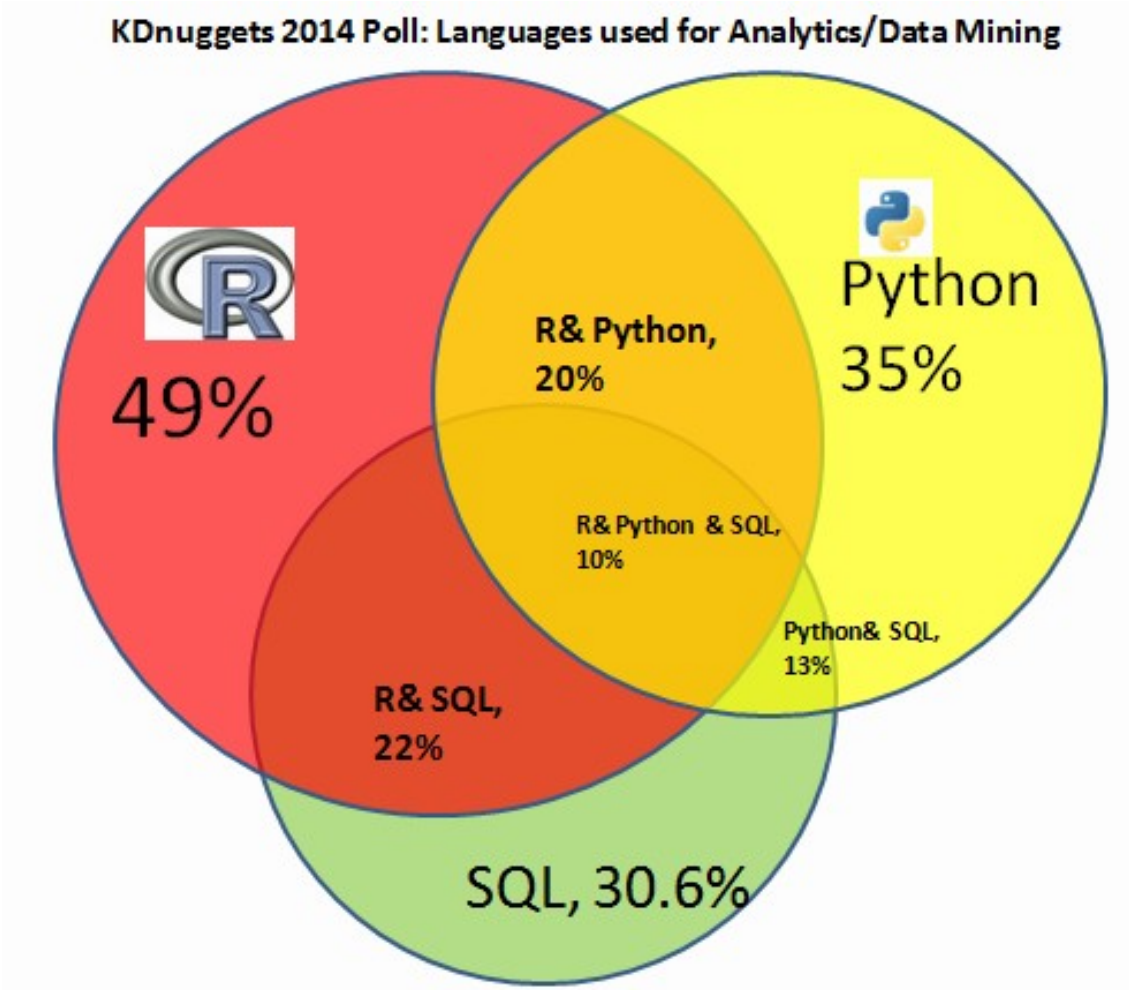
Profesional que *combinando*
conocimientos de **matemáticas**, **estadística** y **programación**, se encarga de analizar los grandes volúmenes de datos.
A diferencia de la estadística tradicional que utilizaba `muestras`, aplica sus conocimientos estadísticos para resolver **problemas de negocio** aplicando las nuevas tecnologías,
que permiten realizar cálculos que hasta ahora no se podían realizar y que aportan valor.

El día a día

Ciclo de trabajo

Recoger/reunir datos Limpiar datos Ampliar conjuntos de datos	80%
Plantear modelos Ejecutar modelos Evaluar modelos	19%
Informar	1%

Lenguajes de programación



¿Qué es R?

- Wikipedia: R (desambiguación) > En matemáticas:
  - En estadística, R es un lenguaje y entorno de programación para análisis estadístico y gráfico; (R en Wikipedia)

What is R

Figura 3: vídeo traducido, en inglés aquí

Un ejemplo de análisis

Varios pasos

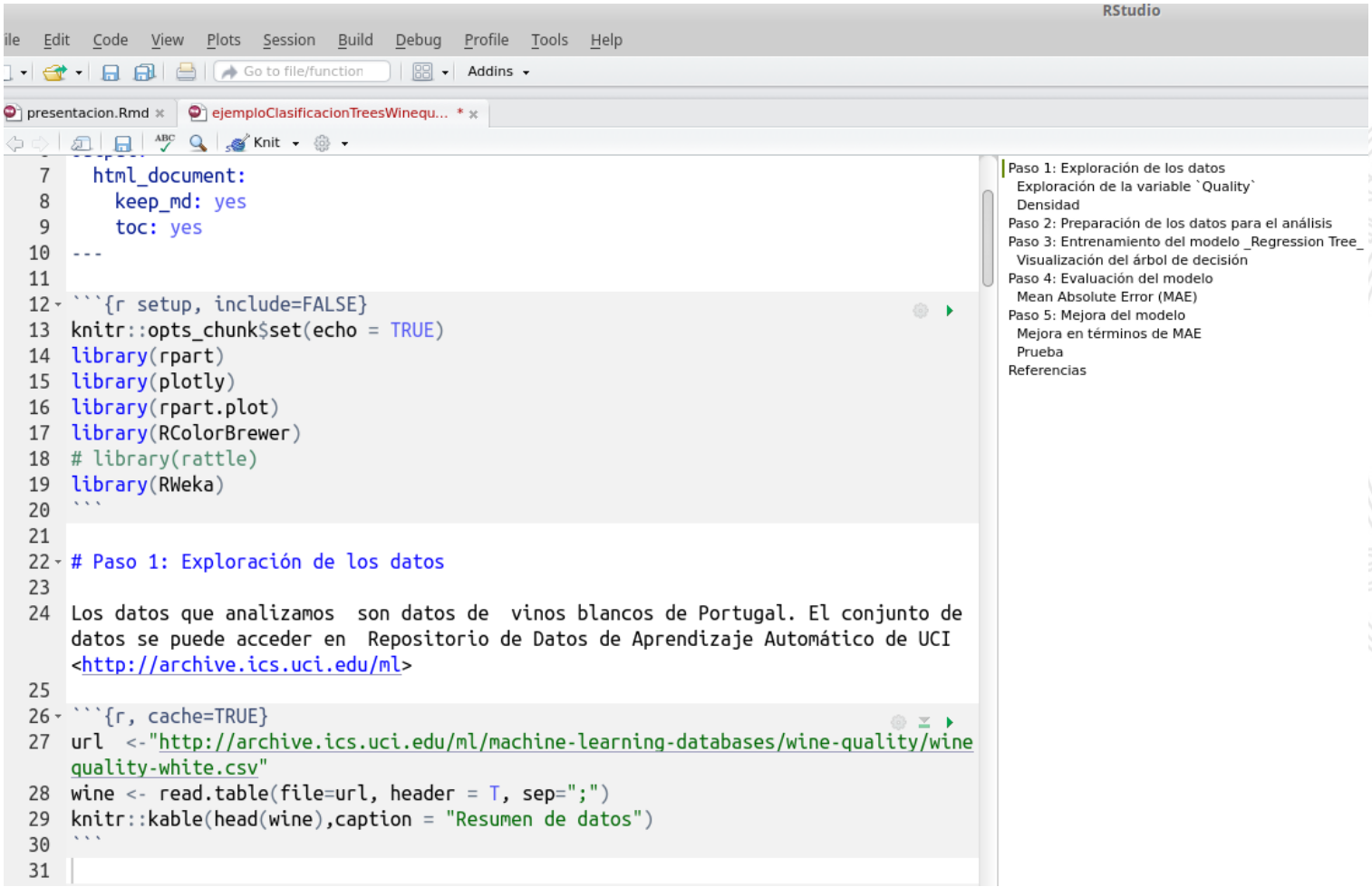
Ejemplo de clasificación por árboles de decisión

Estimación de la calidad del Vino (Wine Quality)

6 de febrero de 2018

`http://rpubs.com/amaurandi/ejemploTree`

Programación literaria



permite desarrollar programas en el **orden** fijado por la `lógica` y el `flujo de nuestros pensamientos`

Ejemplo: investigación reproducible

Los datos que analizamos son datos de vinos blancos de Portugal. El conjunto de datos se puede acceder en Repositorio de Datos de Aprendizaje Automático de UCI http://archive.ics.uci.edu/ml

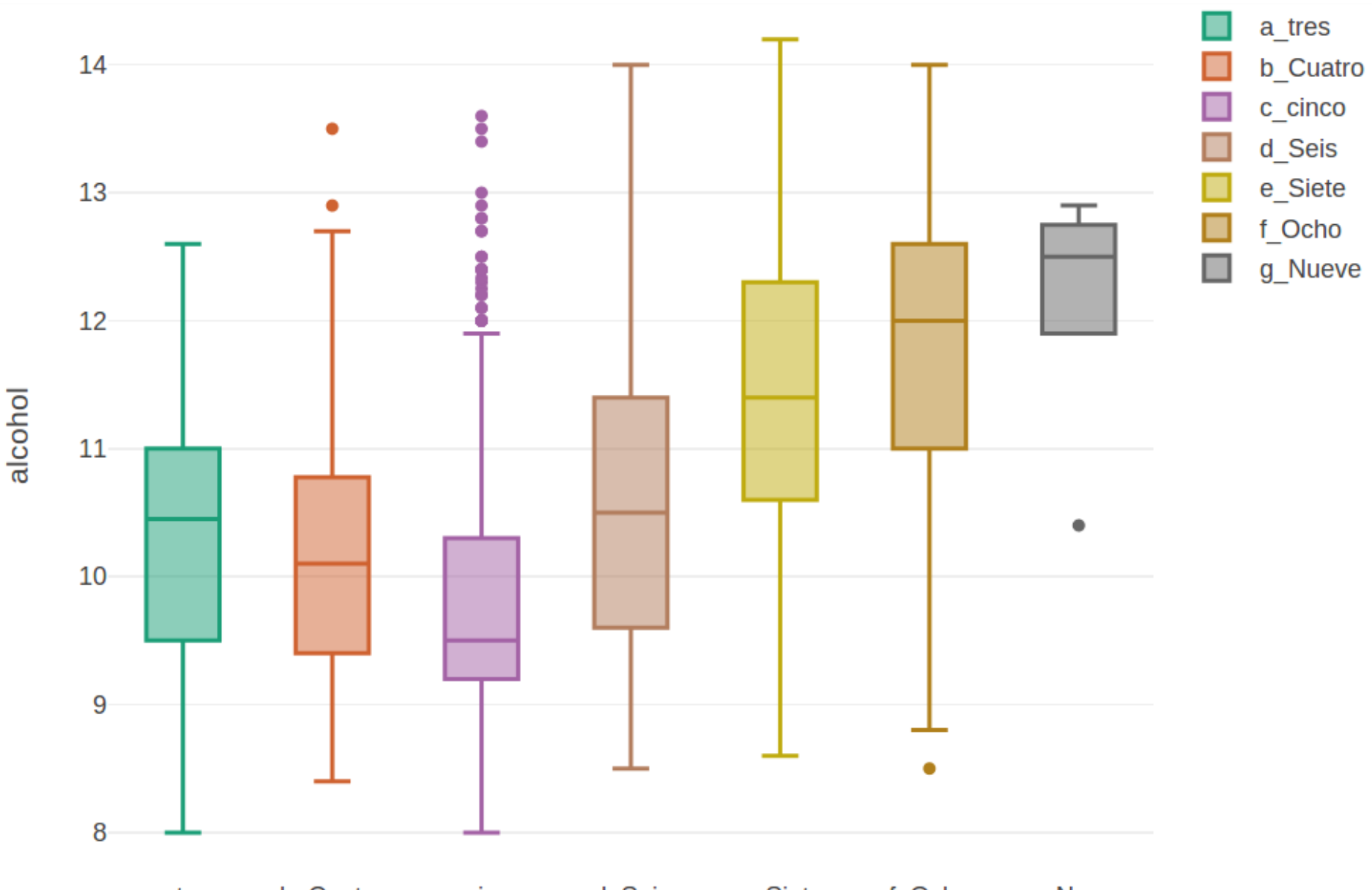
```
url  <-"http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv"
wine <- read.table(file=url, header = T, sep=";")
knitr::kable(head(wine),caption = "Resumen de datos")
```

De cada detalle queda constancia

```
wine2 <- wine
wine2$qualitychar <- ifelse( wine2$quality == 3, "a_tres"
                           , ifelse(wine2$quality == 4, "b_Cuatro"
                           , ifelse(wine2$quality == 5, "c_cinco"
                           , ifelse(wine2$quality == 6, "d_Seis"
                           , ifelse(wine2$quality == 7, "e_Siete"
                           , ifelse(wine2$quality == 8, "f_Ocho"
                           , "g_Nueve"))) )))

plot_ly(data = wine2, x = ~qualitychar, y = ~alcohol
        , color = ~qualitychar
        , type = "box"
        , colors = "Dark2"
        )
```

Ejemplo: gráfico



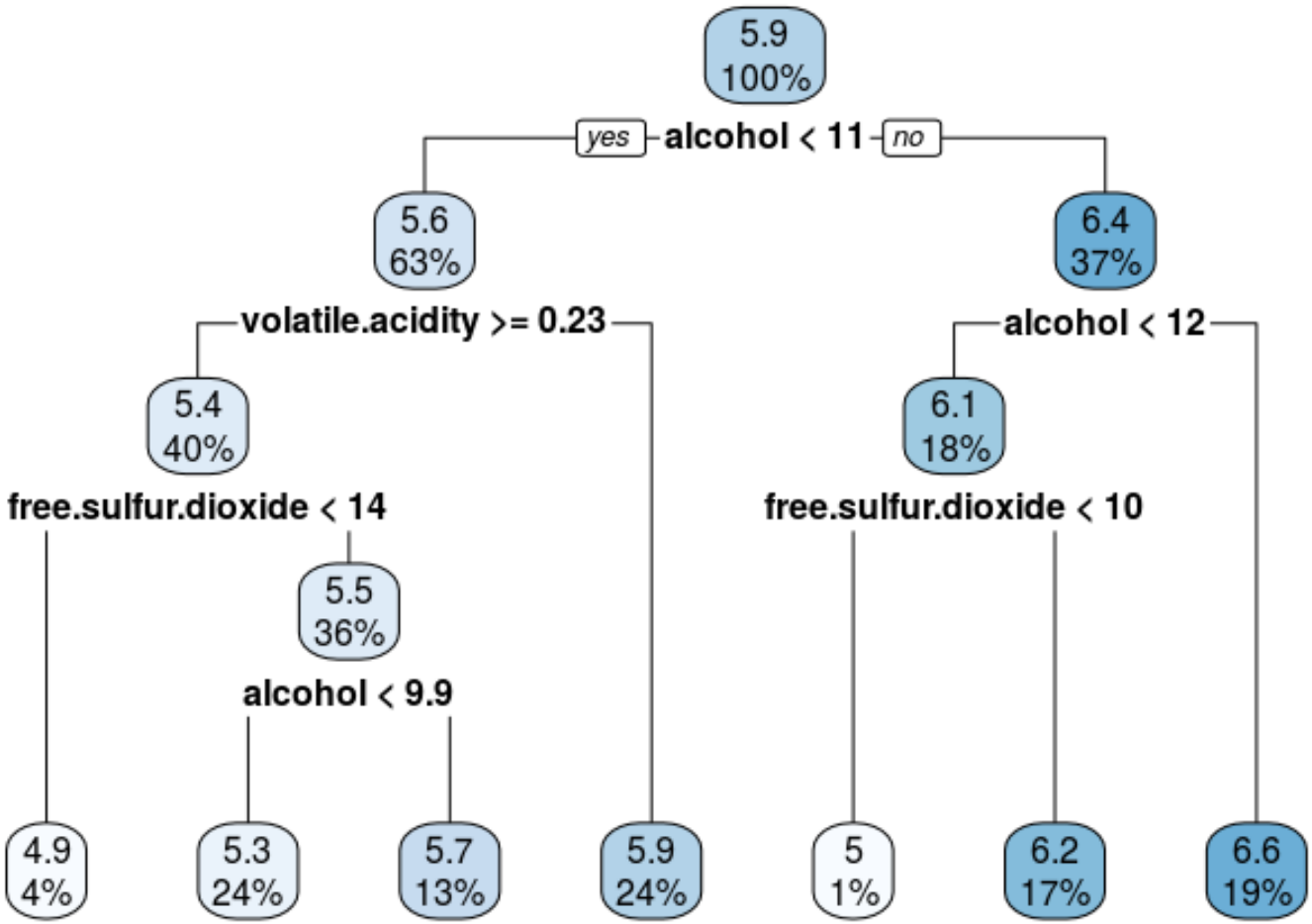
Ejemplo: asjuste de modelo

El paquete `rpart` (partición recursiva) ofrece una muy buena implementación árboles de regresión CART.

```
library(rpart)

m.rpart <- rpart(quality ~.
                 , data = wine_train)
m.rpart
```

Ejemplolo: Árbol de decisión



sessionInfo()

```
sessionInfo()
```

```
## R version 3.4.3 (2017-11-30)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 16.04.3 LTS
##
## Matrix products: default
## BLAS: /usr/lib/libblas/libblas.so.3.6.0
## LAPACK: /usr/lib/lapack/liblapack.so.3.6.0
##
## locale:
##  [1] LC_CTYPE=es_ES.UTF-8          LC_NUMERIC=C
##  [3] LC_TIME=es_ES.UTF-8           LC_COLLATE=es_ES.UTF-8
##  [5] LC_MONETARY=es_ES.UTF-8       LC_MESSAGES=es_ES.UTF-8
##  [7] LC_PAPER=es_ES.UTF-8          LC_NAME=es_ES.UTF-8
##  [9] LC_ADDRESS=es_ES.UTF-8        LC_TELEPHONE=es_ES.UTF-8
## [11] LC_MEASUREMENT=es_ES.UTF-8    LC_IDENTIFICATION=es_ES.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
```

Ofertas de trabajo

Mercado laboral

Una oferta de empleo: Amazon

¿Te apasiona aprovechar los datos para ofrecer una visión práctica que impacte en las decisiones empresariales diarias de Amazon? ¿Te entusiasma la perspectiva de tratar con un volumen masivo de datos?

Data Scientist

Amazon.com ★★★½☆ 20,864 valoraciones - Seattle, WA

Ver o postular al empleo

Basic Qualifications

- Bachelor's degree in Math, Finance, Statistics, Engineering or related discipline
- Experience with Python, R, or other statistics/machine learning packages
- 3+ years experience as a business analyst, data scientist or similar job function, including 1+ years of relevant experience with building statistical models and machine learning.
- Ability to develop experimental and analytic plans for data modeling processes, use of strong baselines, ability to accurately determine cause and effect relations

Preferred Qualifications

- Advanced degree in Math, Statistics, Engineering, Computer Science or related discipline
- Demonstrated ability to frame complex analytical problems and extract insights that

otra: Google

Data Scientist/Quantitative Analyst Intern, Summer 2018

Google
Software Engineering
Mountain View, CA, United States

Contratamos a personas con un amplio conjunto de habilidades técnicas que están dispuestas a asumir algunos de los mayores desafíos de la tecnología y a tener un impacto en millones, si no miles de millones de usuarios. . . .

Responsabilidades:

- Investigación sobre temas como el modelo de negocio de Google y técnicas de búsqueda novedosas.
- Elaborar modelos cuantitativos y cualitativos de las dinámicas de negocio, comportamiento de los usuarios, etc.
- Identificar áreas de investigación y crear métodos de análisis innovadores.

Qualifications

Minimum qualifications:

- Currently pursuing a PhD in statistics, biostatistics, computer science, mathematics, operations research, or another discipline involving experimental design and quantitative analysis of experimental data.
- Must be enrolled in a full-time degree program and returning to the program after the end of the internship.
- Experience using technology to work with datasets such as scripting, Python, statistical software packages (R, S-Plus, SAS or similar).

Preferred qualifications:

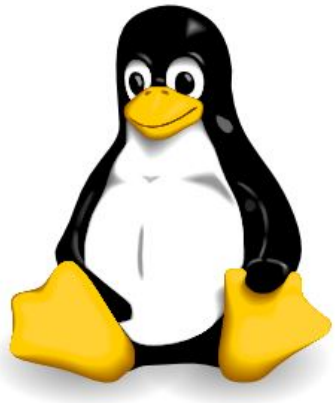
- Expected graduation date in late 2018 or Spring/Summer of 2019.
- Experience with statistical data analysis such as linear models, multivariate analysis, stochastic models, and sampling methods.
- Strong track record of developing intellectual capital such as published works.
- Authorization to work in the United States.

`https://careers.google.com/jobs#!t=jo&jid=/google/data-scientist-quantitative-analyst-google-building-41-1600-amphitheatre-2935340019&`

Para terminar

Como curiosidad: Software Libre

La mayor parte del software que se emplea en este ambiente Software libre

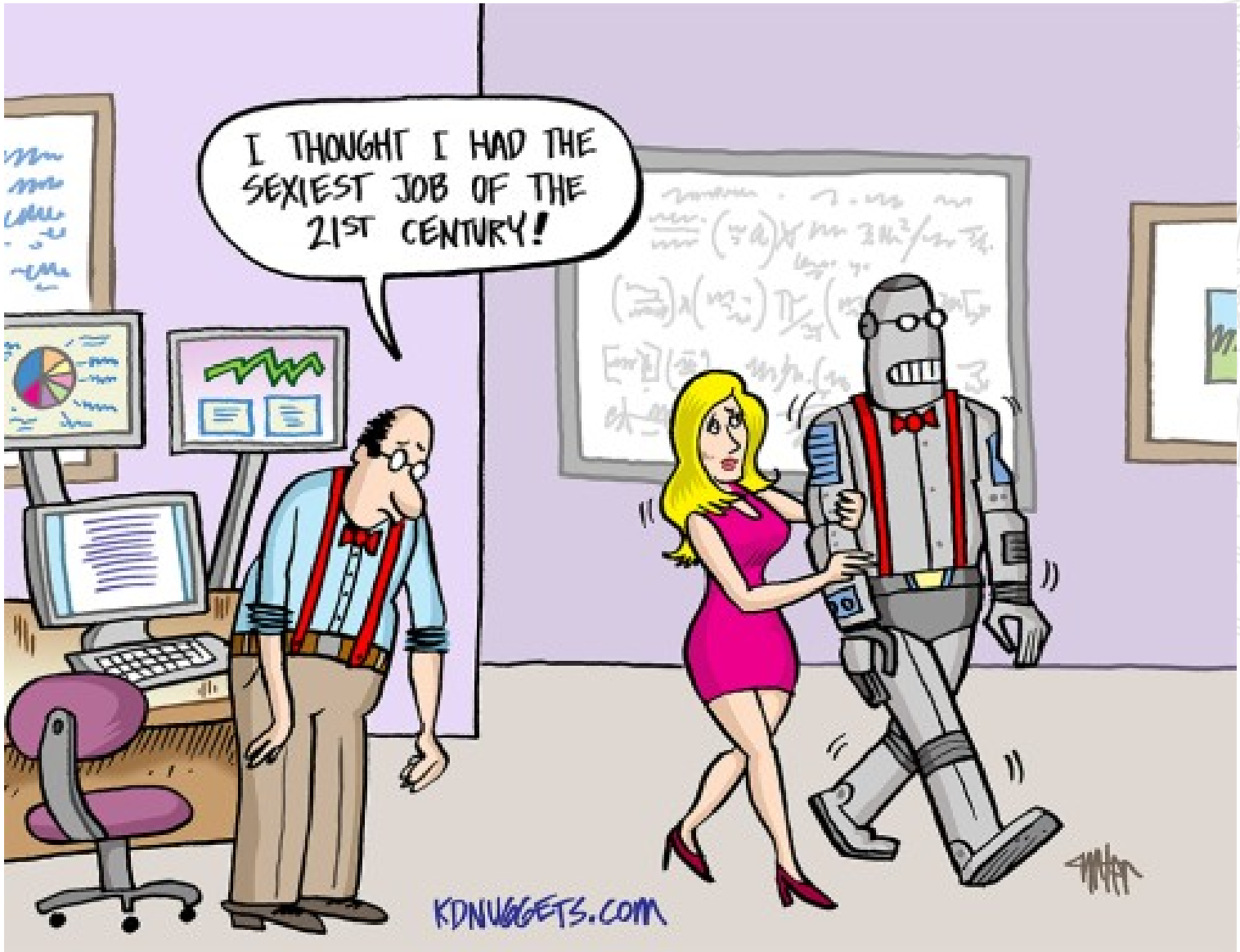


GNU/Linux

Dónde seguir

- Coursera: `https://www.coursera.org/specializations/jhu-data-science`
- The Home of Data Science & Machine Learning: `https://www.kaggle.com/`
- UMU: Máster Universitario en Tecnologías de Análisis de Datos Masivos: BIG DATA

El futuro de la profesión



Enlaces y referencias

Enlaces y referencias

- Thomas H. DavenportD.J. Patil (2012). *Data Scientist: The Sexiest Job of the 21st Century*. Harvard Business Review. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century
- Mike Gualtieri (2013). What Is A Data Scientist?. https://www.youtube.com/watch?v=iQBat7e0MQs.
- Gregory Piatetsky(2016). Top Algorithms and Methods Used by Data Scientists _ KDnuggets_ https://www.kdnuggets.com/2016/09/poll-algorithms-used-data-scientists.html

Muchas Gracias

Ciencia de datos: El trabajo más sexy del siglo XXI

XIV Semana de la Ciencia y la Tecnología. IES Floridablanca, Murcia

Antonio Maurandi López

amaurandi@um.es

Universidad de Murcia

6 de febrero de 2018