Introducción al Data Science

Alberto Torres Barrán

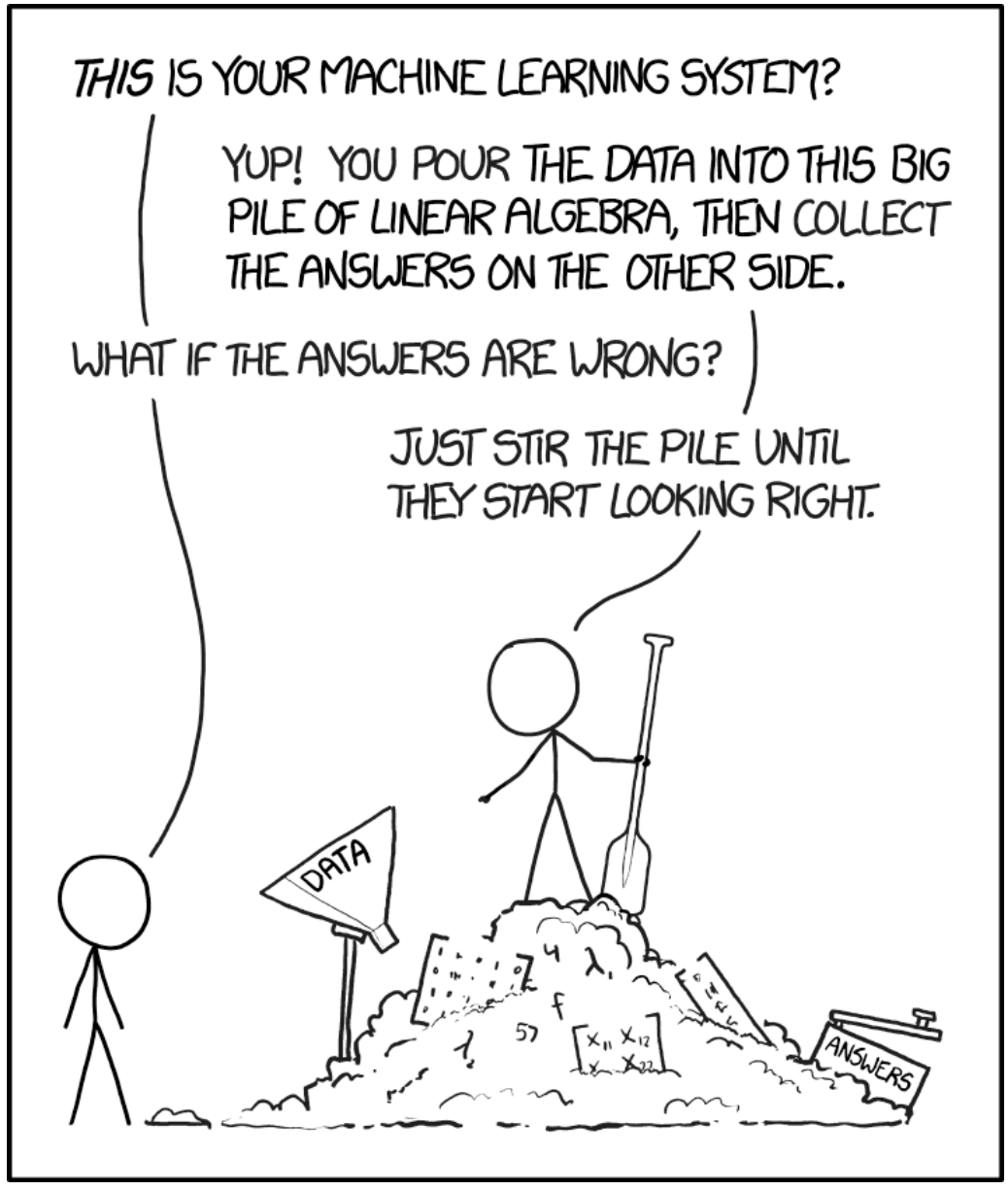
3 de Enero del 2020

¿Qué es el *Machine Learning*?

De la Wikipedia:

> *Machine learning is a subfield of* ***computer science*** *that evolved from the study of* ***pattern recognition*** *and computational learning theory in* ***artificial intelligence****. In 1959, Arthur Samuel defined machine learning as a "Field of study that gives* ***computers the ability to learn without being explicitly programmed****". Machine learning explores the study and construction of algorithms that can learn from and make predictions on data.*

Está íntimamente ligado con otras disciplinas.


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

Fuente


Data Science
Data Mining
Statistics
Artificial Intelligence
Machine Learning
Pattern Recognition

Statistics Más antigua (aprox. 1749), el resto de disciplinas utilizan algunas de sus técnicas: estadística descriptiva, análisis de regresión, inferencia.

Artificial Intelligence Más moderna, 1940. Algunos problemas que intenta resolver: procesamiento lenguaje natural, planificación, visión por computador, robótica.

Machine Learning Rama de la IA, 1946. Se utiliza para resolver algunos de los problemas que tiene la IA.

Pattern Recognition En general se usa como sinónimo de *Machine Learning*.

Data Mining Técnicas de modelado estadístico y *machine learning* aplicadas a un dominio en concreto.

Data Science Término más moderno, mezcla de todo lo anterior.

Data scientist

De la Wikipedia:

> *Data science employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, including signal processing, probability models, machine learning, statistical learning, data mining, database, data engineering, pattern recognition and learning, visualization, predictive analytics, uncertainty modeling, data warehousing, data compression, computer programming, artificial intelligence, and high performance computing.*

La conclusión es que se tratan de un perfil muy amplio, con un conjunto de habilidades poco definido.

Data science en la empresa

- En 2012, Harvard Business Review publicó el artículo:


DATA

Data Scientist: The Sexiest Job of the 21st Century

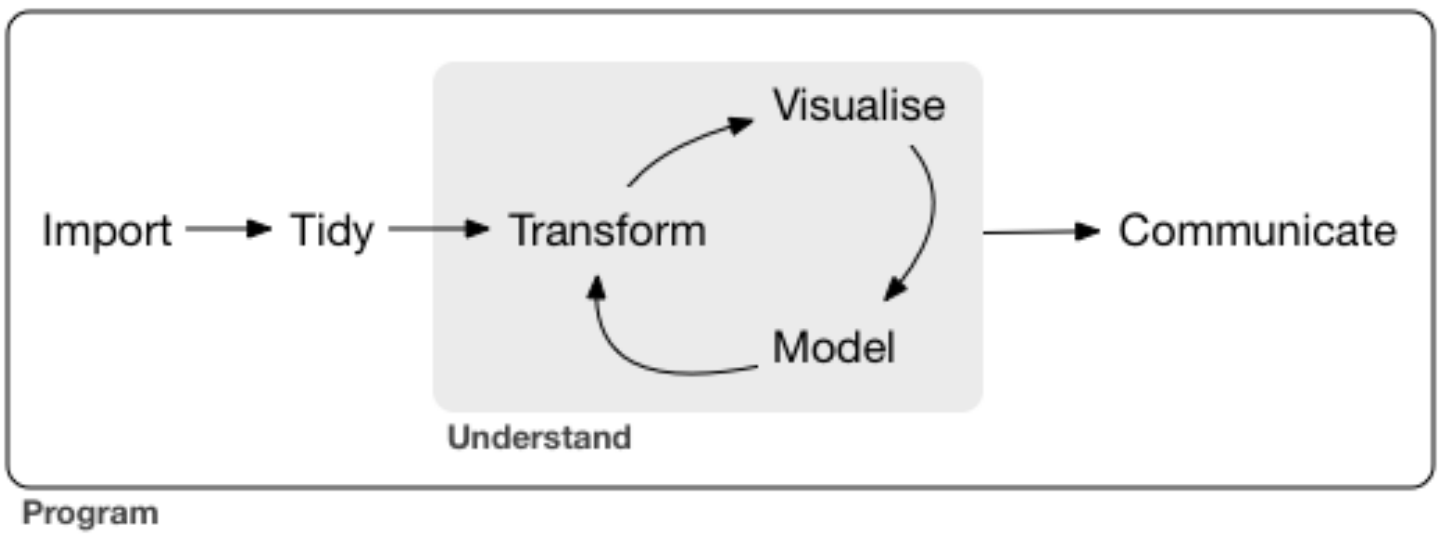
by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

- La consultora McKinsey calcula que para 2018 habrá una demanda de entre 140,000-190,000 puestos de *data science* sin cubrir.
- Nadie sabe muy bien lo que es pero todo el mundo quiere uno.

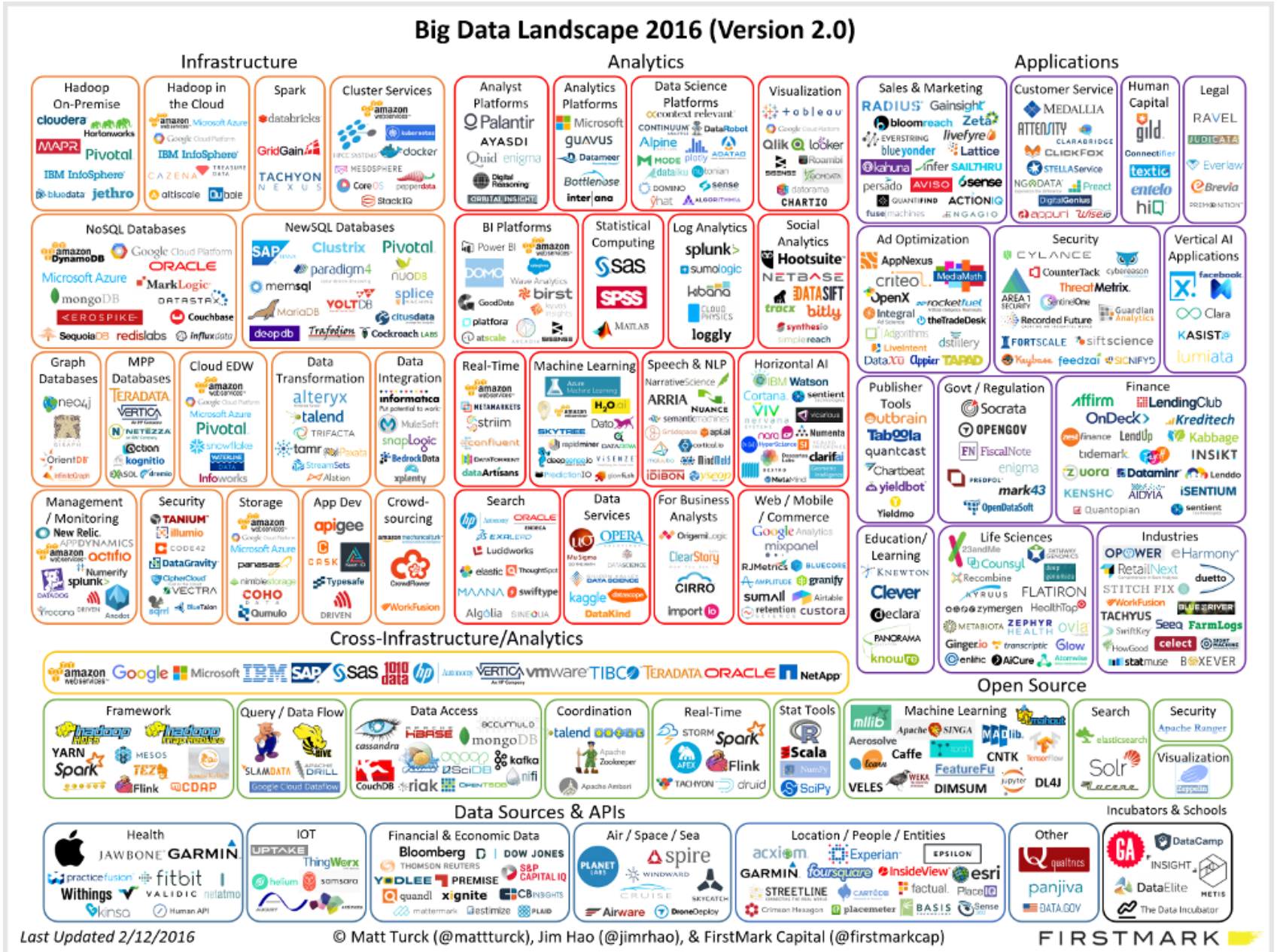
Flujo de trabajo de un equipo de *data science*

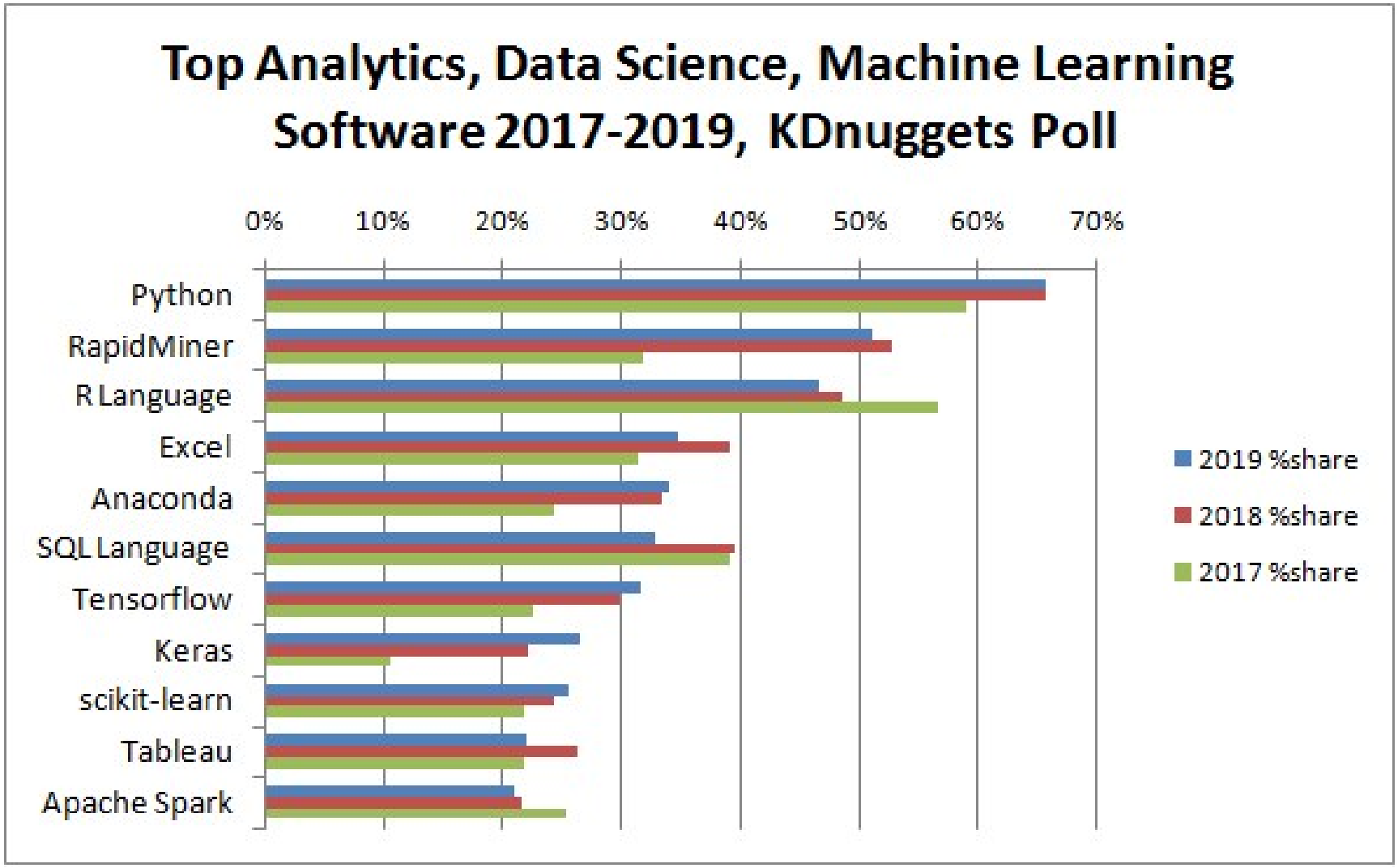
- En ocasiones un único perfil realiza todas las tareas.
- Sin embargo, cada vez es más habitual tener un equipo donde cada integrante esté especializado en distintas partes del proceso.



Fuente: Hadley Wickham, R for Data Science

Herramientas de análisis



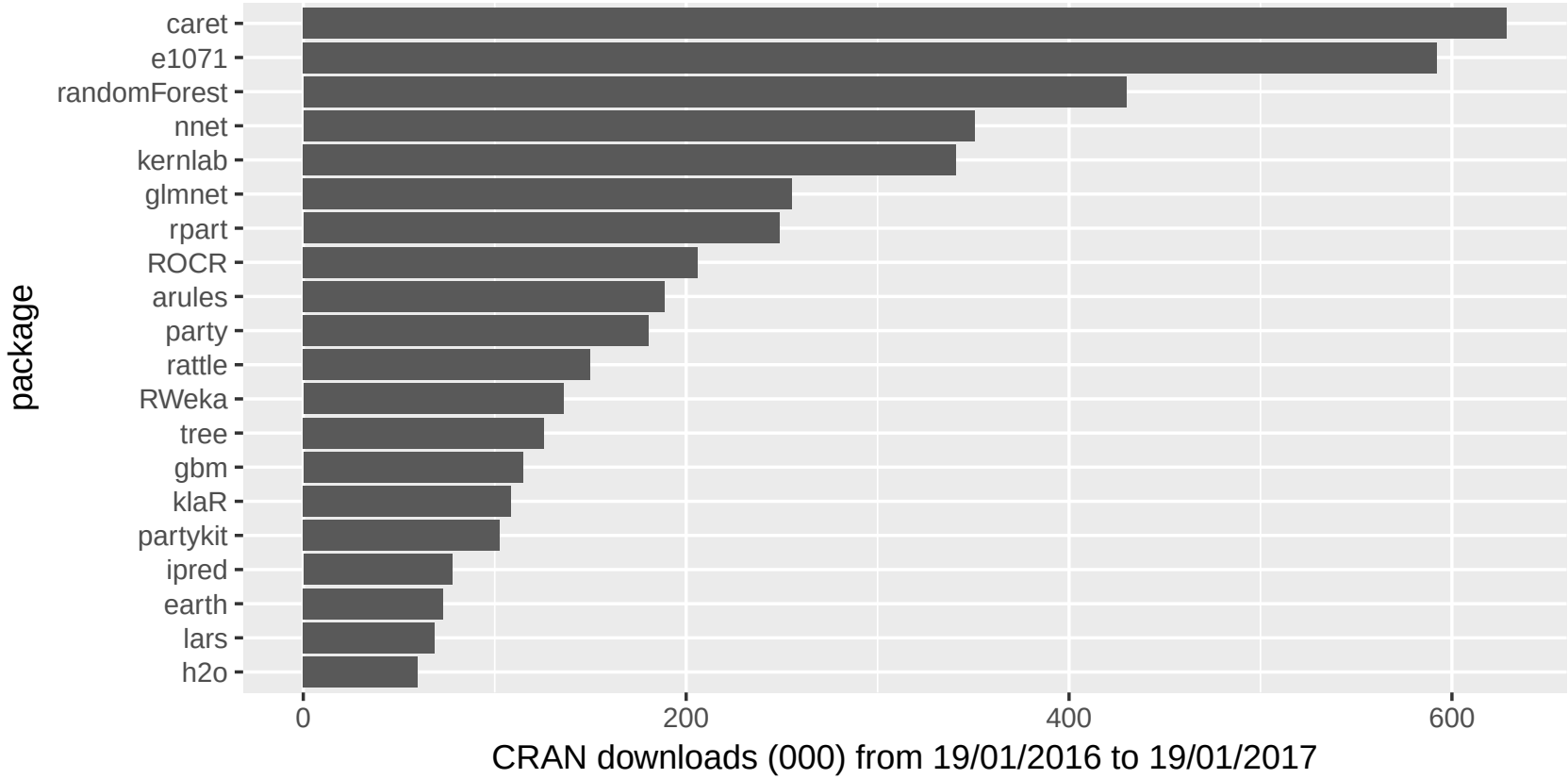


Software para proyectos de *analytics*, y *data science*. Encuesta de KDnuggets.

Colecciones de paquetes útiles

- En este enlace se puede ver una lista muy reciente de paquetes útiles.
- Machine Learning in R es una colección de paquetes de aprendizaje automático.
- High-Performance computing in R es una colección de paquetes de útiles para la computación de alto rendimiento.
- Recientemente aplicaciones web que permiten la visualización de resultados también gozan de gran popularidad, por ejemplo los notebooks de Jupyter.
- En R destaca Shiny, que permite convertir código R en aplicaciones web interactivas.

Paquetes más populares de ML



Descargas Enero 2016 - Enero 2017. Fuente

Resumen paquetes destacados

- Modelos lineales y utilidades:
  - `caret`, utilidades para clasificación y regresión
  - `MASS`, ridge regression
  - `ridge`, ridge regression con selección automática del hiper-parámetro
  - `glmnet`, GLMs con regularización Lasso o Elastic Net
  - `glmnetUtils`, utilidades para `glmnet`
  - `gam`, modelos aditivos generalizados
  - `mgcv`, modelos adictivos generalizados (recomendado)
- Algunos modelos más complejos:
  - `nnet`, Redes neuronales, también regresión logística multinomial
  - `e1071`, Support Vector Machines
  - `gbm`, Gradient Boosting
  - `randomForest`, Random Forest
  - `xgboost`, Extreme Gradient Boosting

Tidyverse

- Conjunto de paquetes creados por Hadley Wickham que comparten una misma API y contienen funciones para el análisis de datos:
  - `ggplot2`, para hacer gráficos avanzados.
  - `dplyr`, para manipular datos.
  - `tidyr`, para limpiar datos.
  - `readr`, para importar datos.
  - `purrr`, para programación funcional.
  - `tibble`, implementa *tibbles*, una versión moderna de los data.frames.
- El paquete *tidyverse* instala y carga los paquetes anteriores.
- También instala otros paquetes que pueden ser útiles aunque no los carga por defecto.
- Para más información y la lista completa de paquetes: https://github.com/tidyverse/tidyverse.

Libros y manuales

En general, se pueden encontrar muchos manuales en las secciones *Manuals* y *Contributed* de CRAN, así como ejemplos en la web RPubs. Algunos recursos más específicos:

Libros
- R for Data Science [url].
- An Introduction to Statistical Learning with Applications in R [url].

E-Books
- YaRrr! The Pirate's Guide to R [url].
- The R Inferno [url].
- R Programming [url].

Blogs
- RTutorial [url].
- Quick-R [url].
- RStudio [url].
- RBloggers [url].

FAQs y comunidades

- StackOverflow: las preguntas con el tag R contienen mucha información y problemas resueltos. Además, las nuevas preguntas se responden en cuestión de horas.
- CrossValidated: no es una comunidad específica de R (más bien de estadística), pero hay mucha información acerca de cómo realizar procedimientos concretos de análisis de datos y aprendizaje automático en R.
- @RLangTip: Twitter que publica consejos y trucos diarios.
- R Programming for Data Analysis: Comunidad de Google+.
- Statistics and R: Otra comunidad de Google+.
- The R Project for Statistical Computing: Grupo de LinkedIn.

Algunas referencias

1. Jerome H. Friedman. *Data Mining and Statistics: What's the Connection?* (1998). [url]
2. Leo Breiman. *Statistical Modeling: The Two Cultures* (2001). [url]
3. Cross Validated. *What is the difference between data mining, statistics, machine learning and AI?* (2010). [url]
4. Sakthi Dasan Sekar. *What is the difference between Artificial Intelligence, Machine Learning, Statistics, and Data Mining* (2014). [url].
5. Cross Validated. *What exactly is Big Data?* (2015). [url]
6. David Donoho. *50 years of Data Science* (2015). [url]