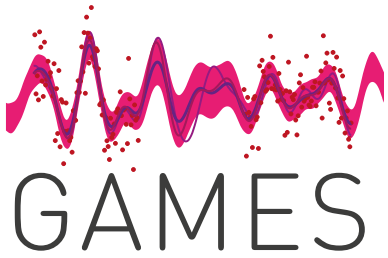
Data Science

Herramientas, lenguajes y Python

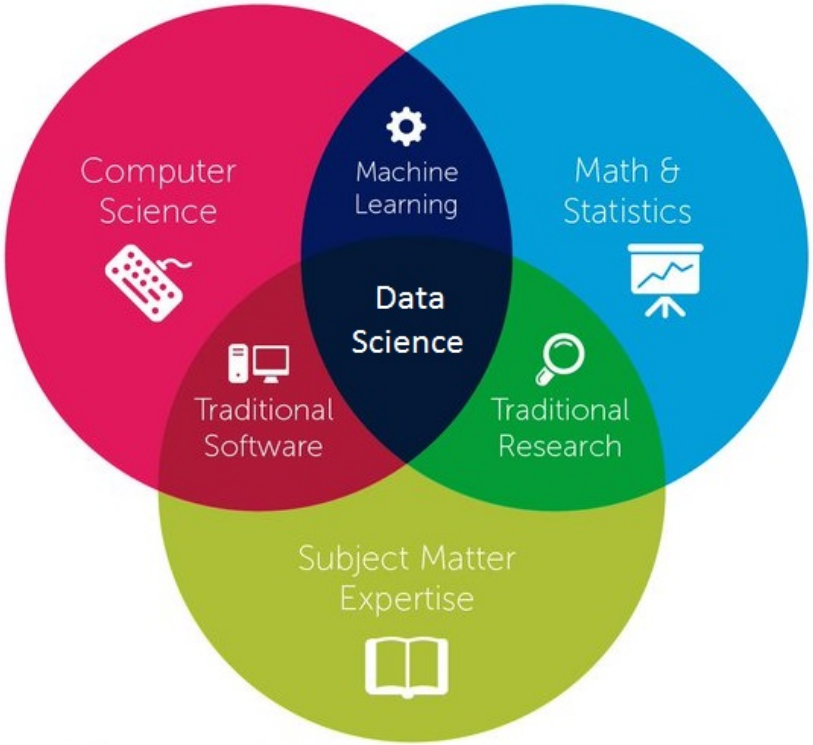
Gonzalo Rios



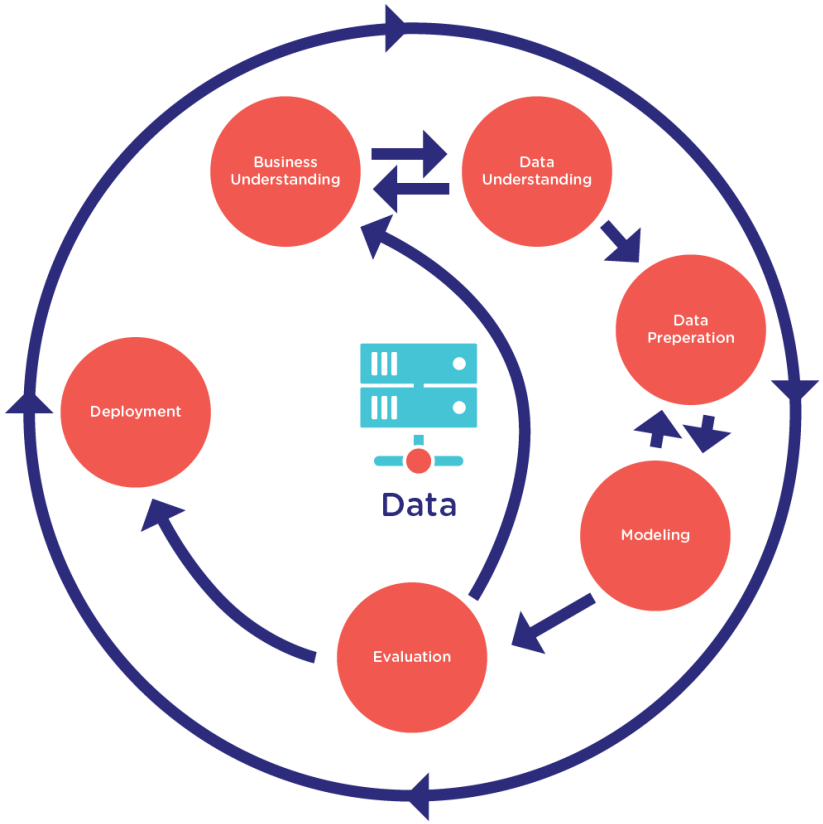
MA6201 - Computación Científica
Grupo de Aprendizaje de Máquinas, infErencia y Señales
Centro de Modelamiento Matemático
Universidad de Chile

Agosto 2017

Data Science es un campo interdisciplinario que aplica técnicas **matemáticas**, **estadísticas** y **computacionales** a diversas áreas: **biología**, **física**, **economía**, **sicología**, **sociología**, entre otras.



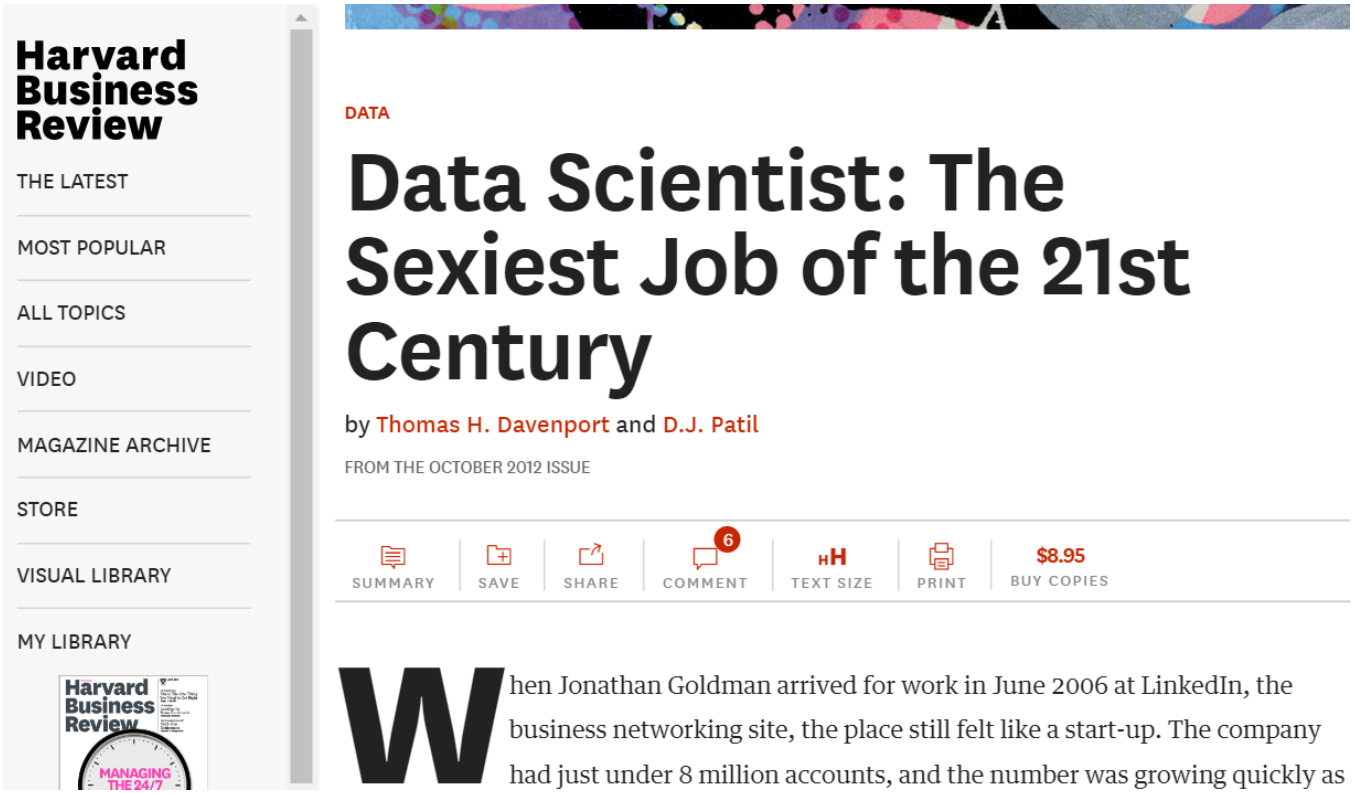
Data Science tiene la misión de **modelar**, **analizar**, **entender**, **visualizar** y **extraer** conocimiento a partir de **datos**



Data Science

¿Quién hace Data Science?

Data Scientists son los profesionales de **Data Science**, que necesitan conocer más **estadística** que un ingeniero de software, y saber más **ingeniería de software** que un estadístico



Harvard Business Review

THE LATEST

MOST POPULAR

ALL TOPICS

VIDEO

MAGAZINE ARCHIVE

STORE

VISUAL LIBRARY

MY LIBRARY

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

SUMMARY SAVE SHARE COMMENT 6 TEXT SIZE PRINT $8.95 BUY COPIES

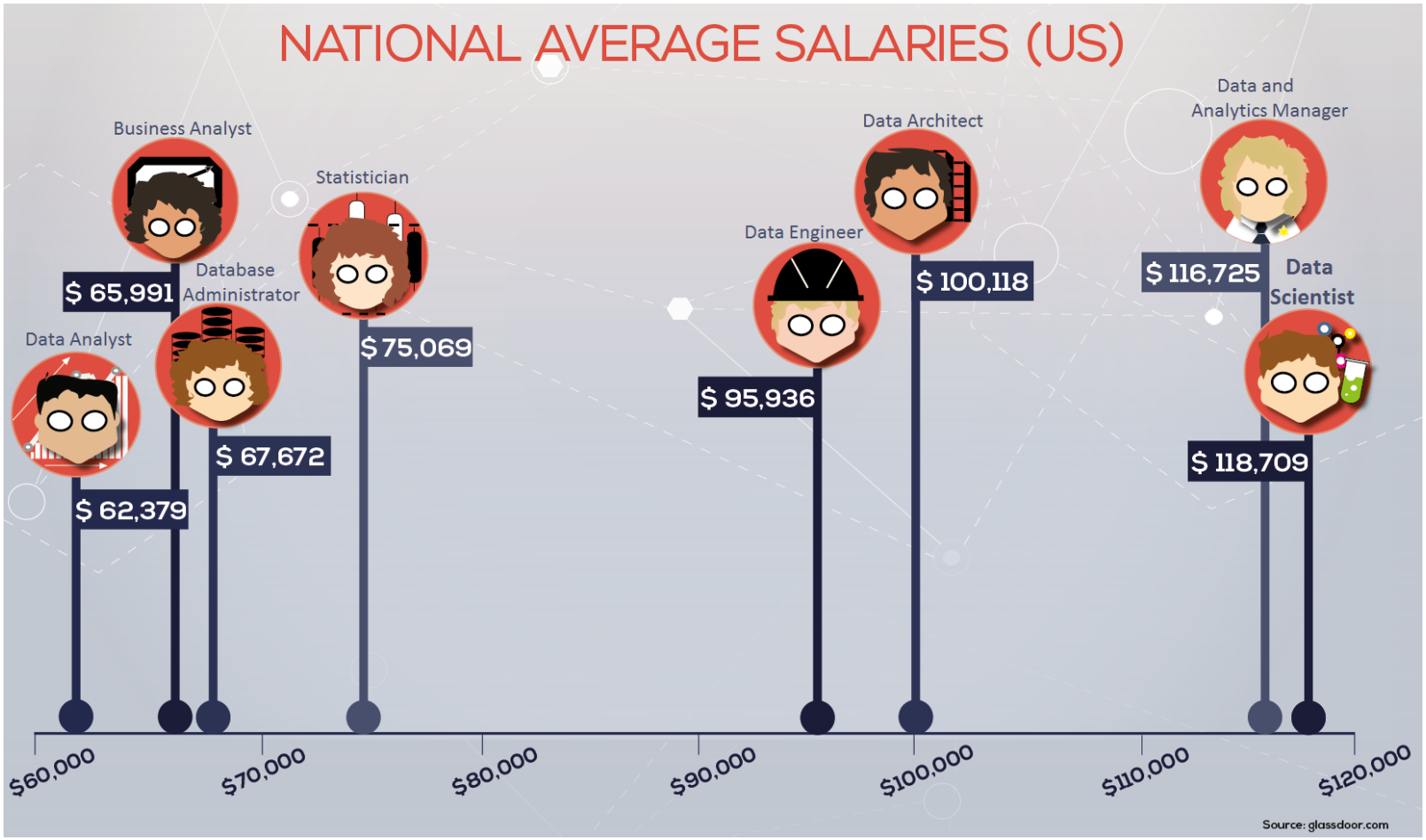
When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as

Ref: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Data Scientist

Valorados en la Industria Privada

Google, **Microsoft** y **Facebook** contratan **Data Scientists**

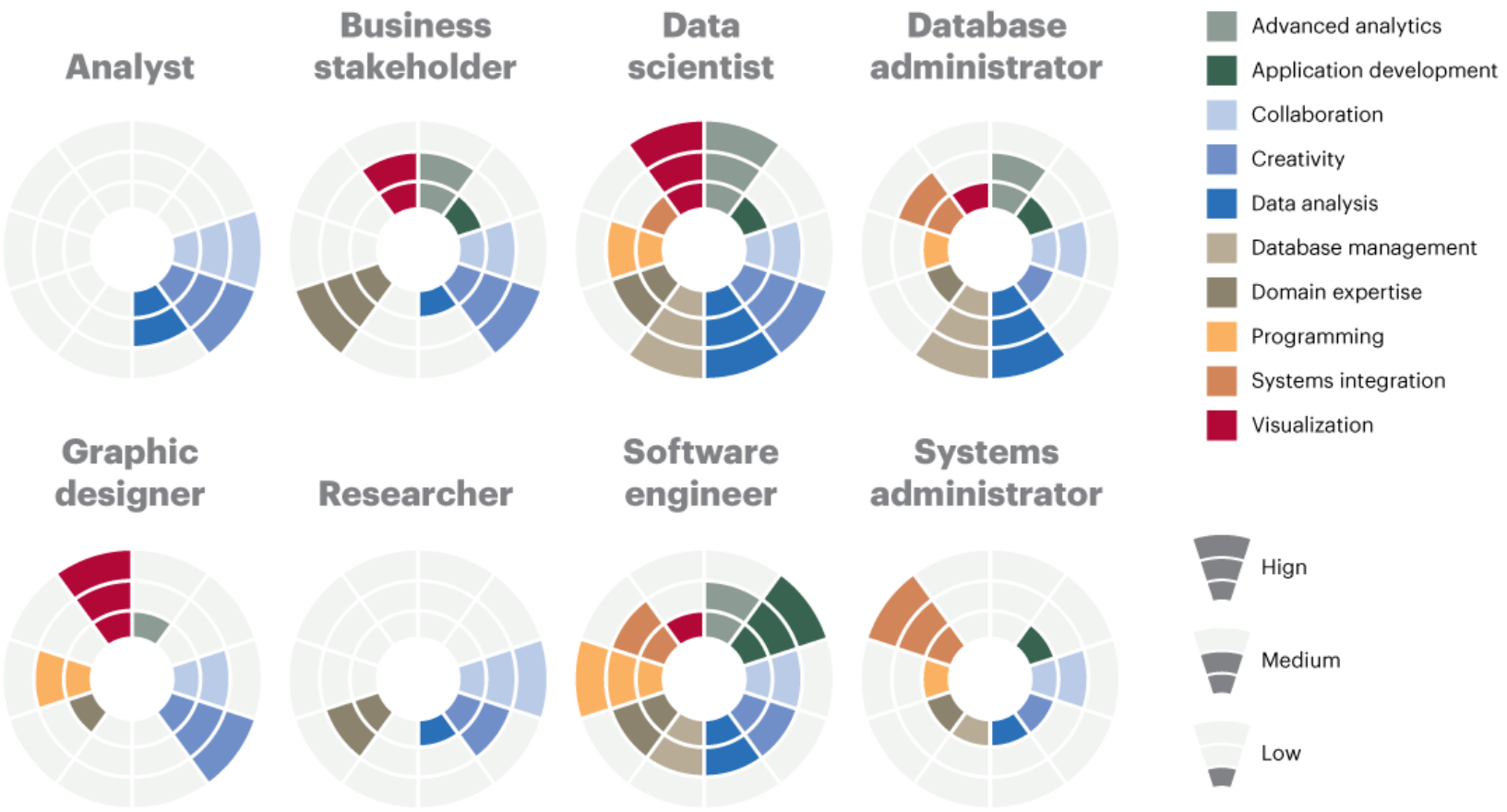


Ref: https://www.glassdoor.com

Data Scientists

Habilidades por Roles

Existen diferentes roles relacionados con **Data Science**

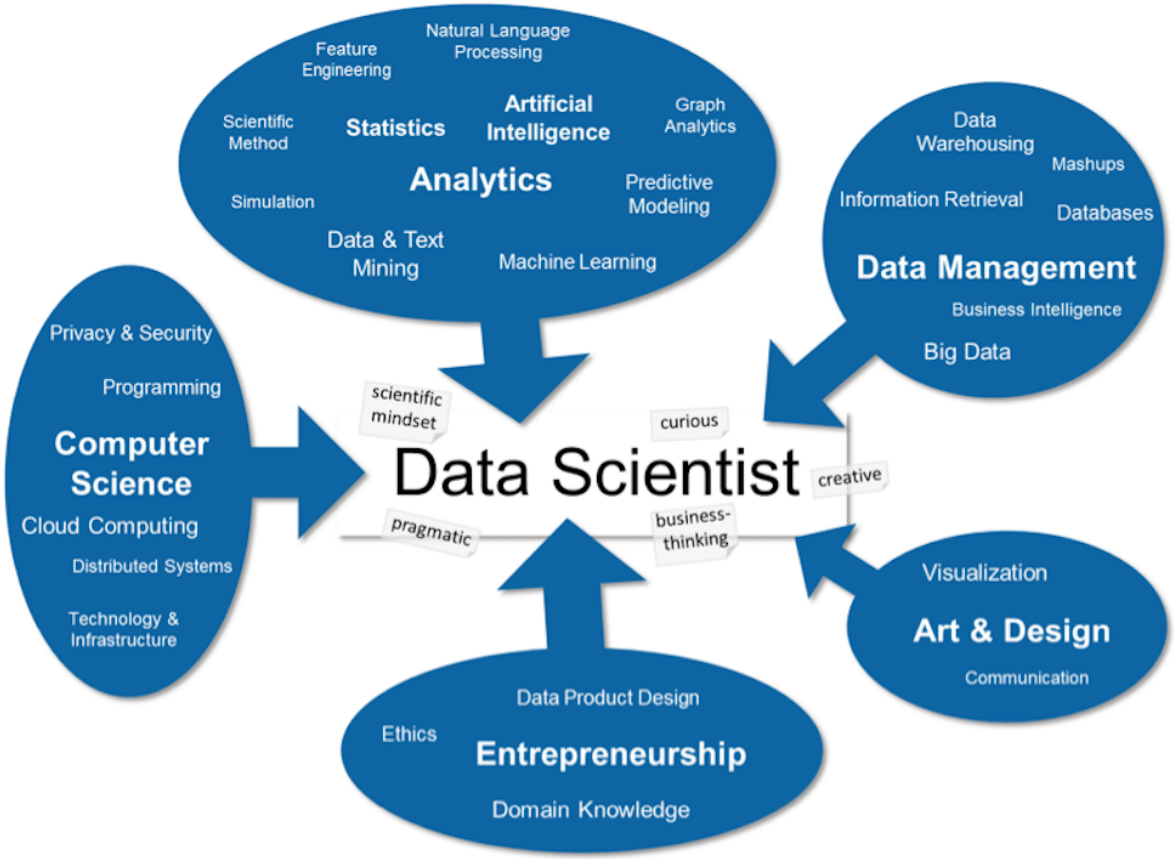


Source: A.T. Kearney analysis

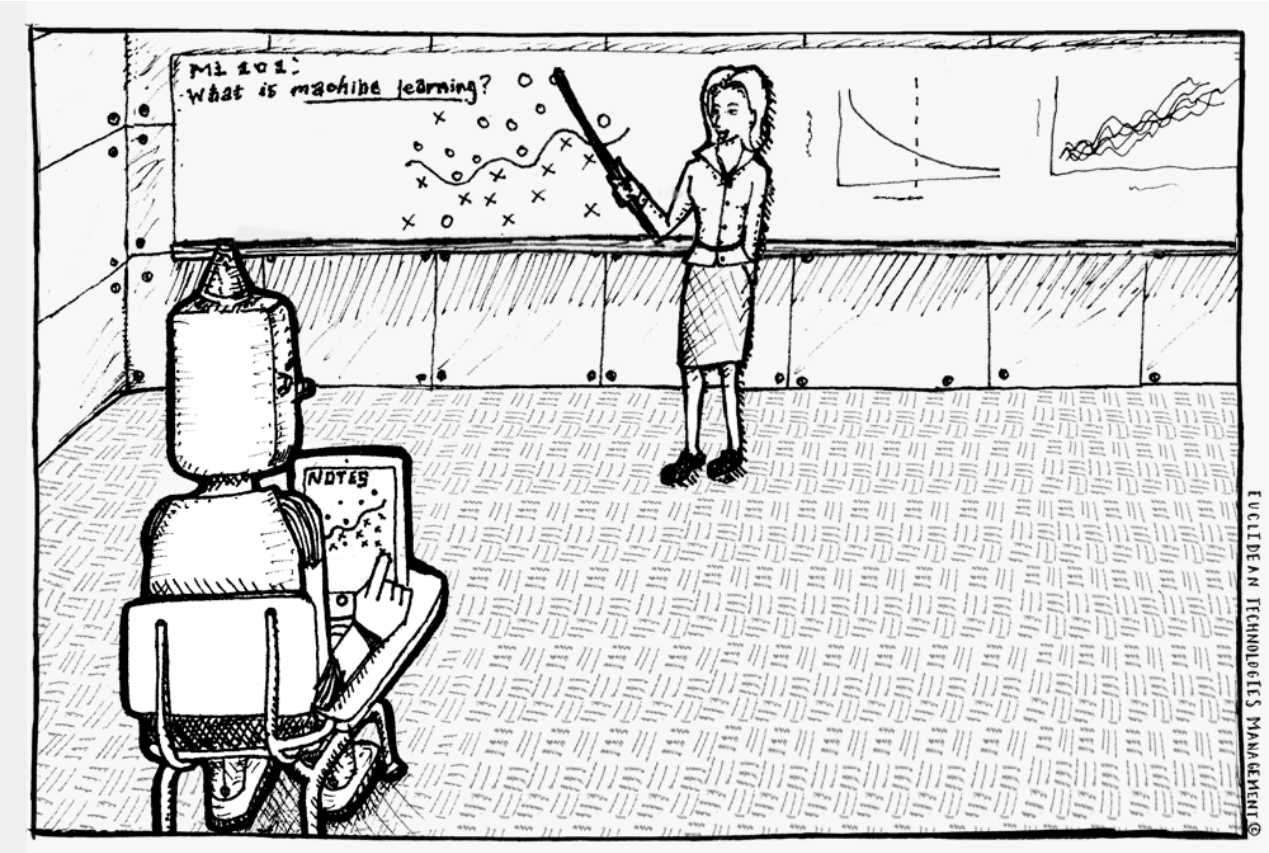
Data Scientists

Habilidades Interdisciplinarias

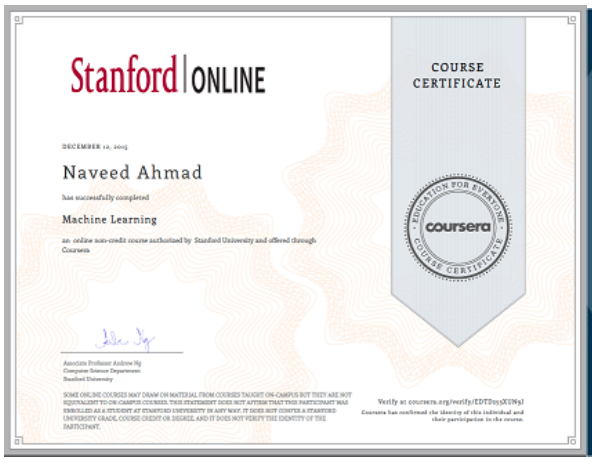
Un **Data Scientist** posee conocimientos (y **creatividad!**) en **modelación**, **visualización**, **bases de datos** y **programación**

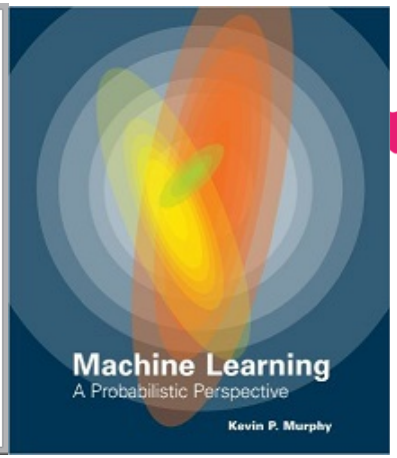


Machine Learning es un área cuyo objetivo es desarrollar **algoritmos** que permitan a las **computadoras aprender**.



Una buena forma de aprender **Machine Learning** es haciendo un **curso online**, leer un **libro** y asistir a **charlas**





Pronto empezarán los **Martes de Máquina**!

http://www.coursera.org/
http://mitpress.mit.edu/books/
http://games.cmm.uchile.cl/

Programming Languages

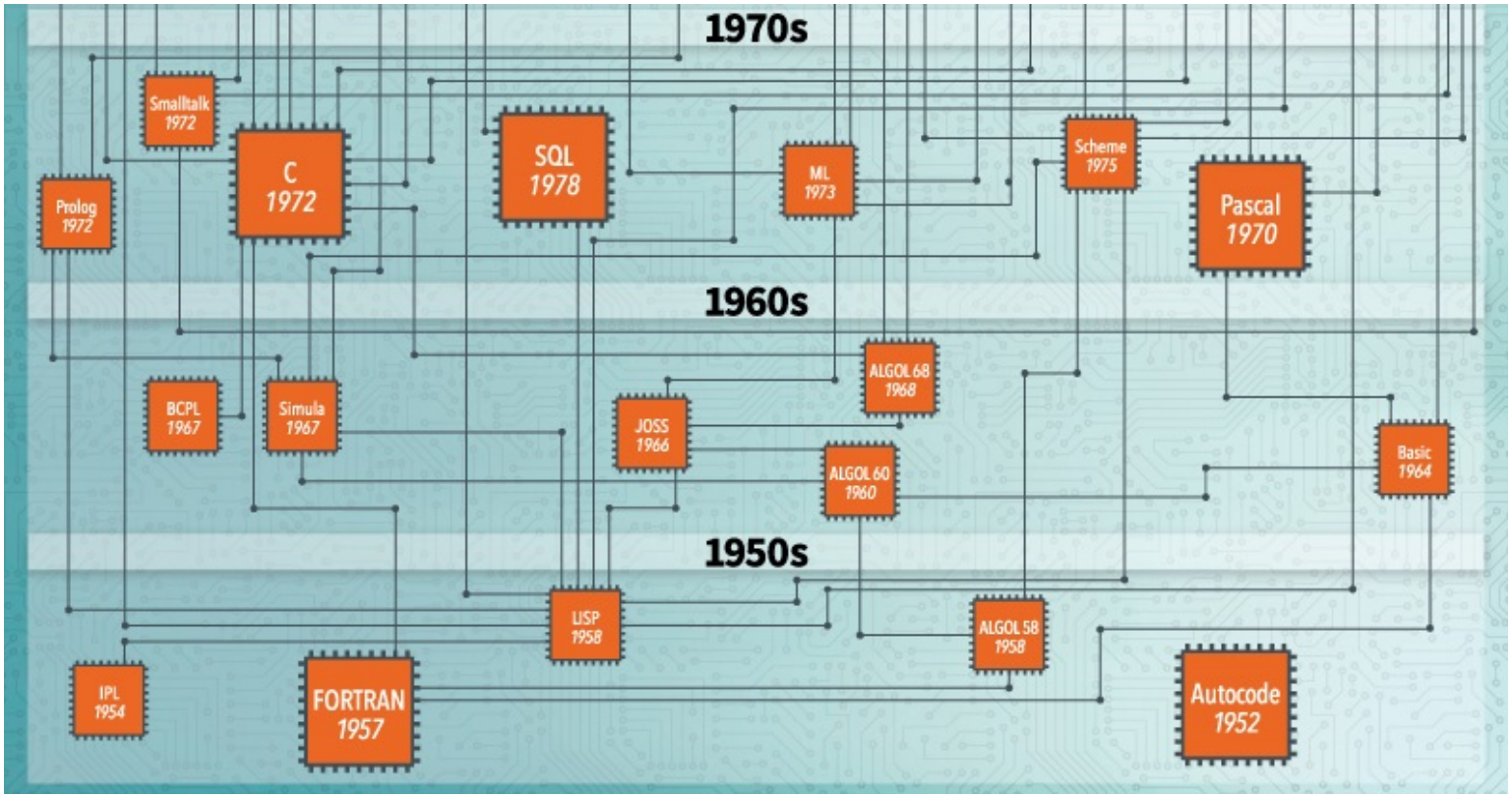
¿Cuál lenguaje de programación escoger?

Data Science necesita el uso de un **lenguaje de programación**, el problema es como **escoger** uno entre cientos (o miles)!

Programming Languages

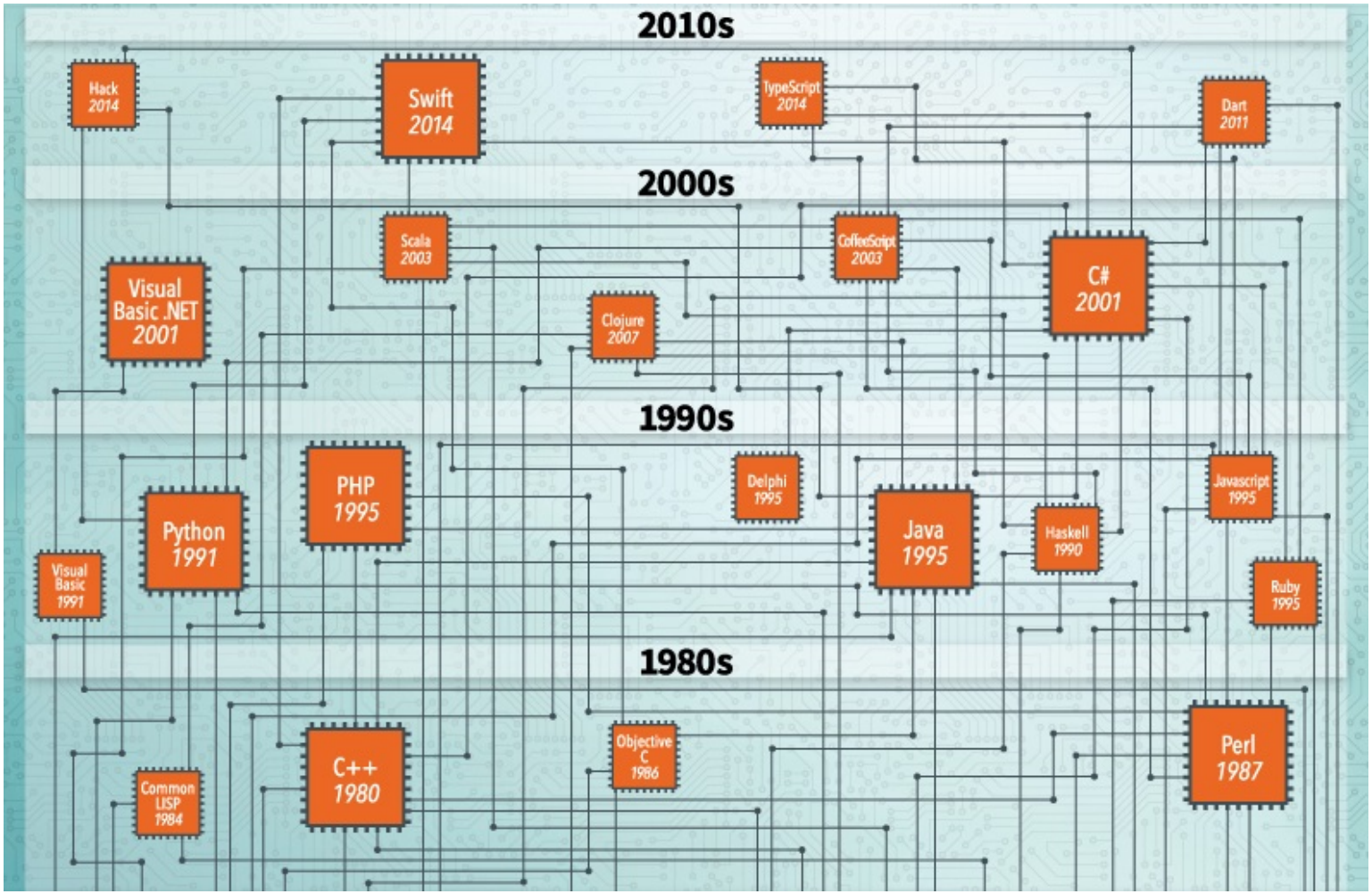
Historia: 1837 - 1979

En 1837, **Charles Babbage** diseñó la primera **máquina programable**, y fue **Ada Lovelace** quien publicó en 1843 el primer **programa**. En 1946 se construyó **ENIAC**, el primer computador Turing-completo programable de propósitos generales.

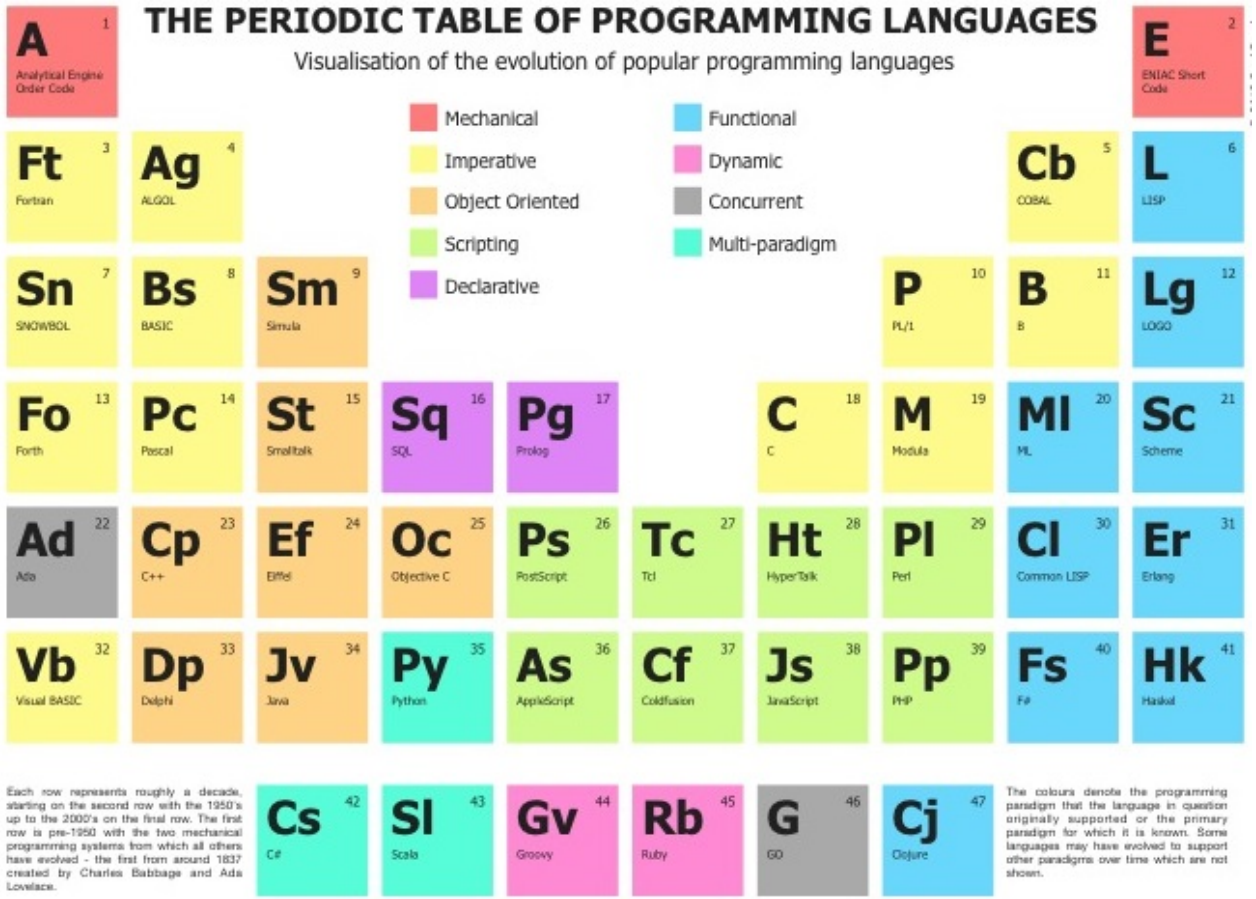


Programming Languages

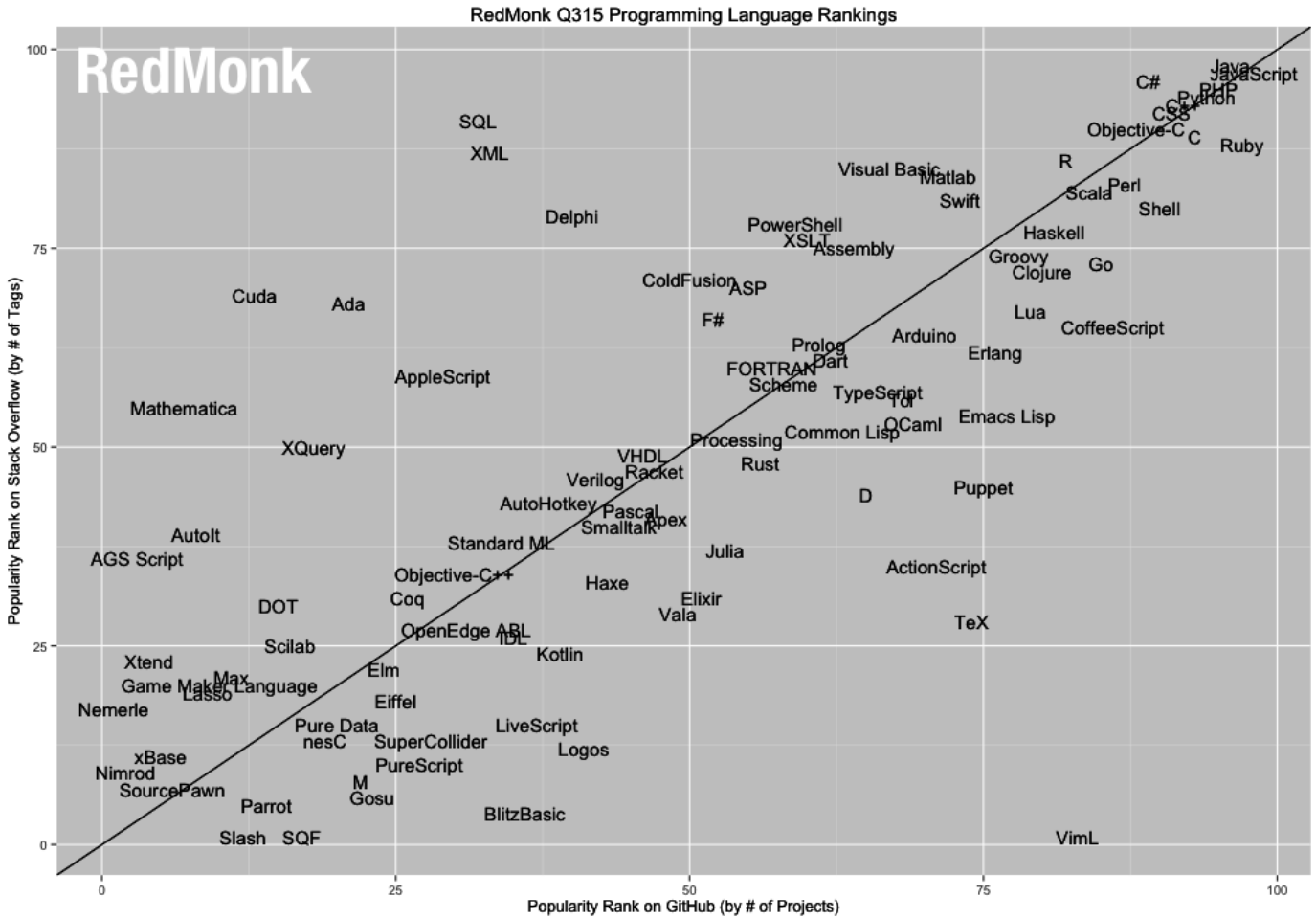
Historia: 1980 - Hoy



Los **paradigmas de programación** son las diferentes corrientes filosóficas de programación según la **abstracción de elementos**.



Documentación y **librerías** acelera los tiempos de desarrollo

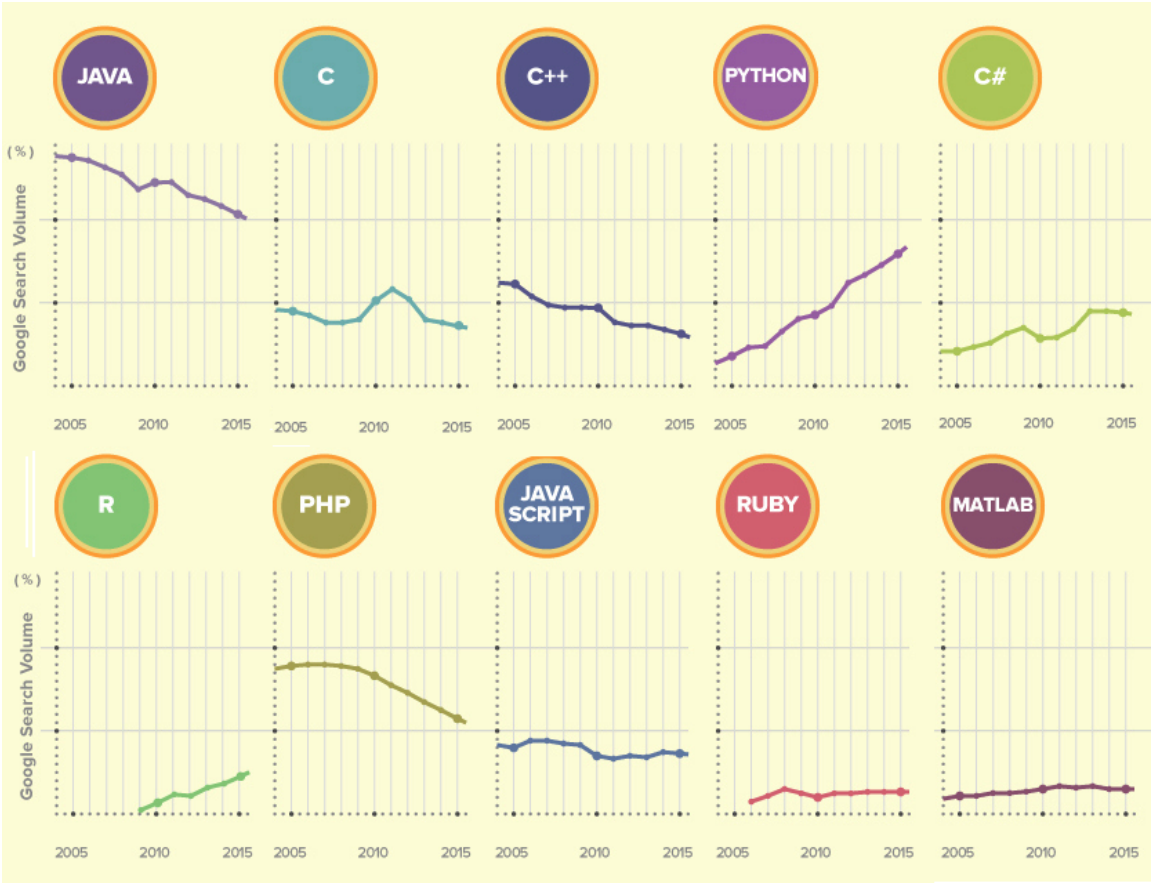


Ref: http://redmonk.com/

Programming Languages

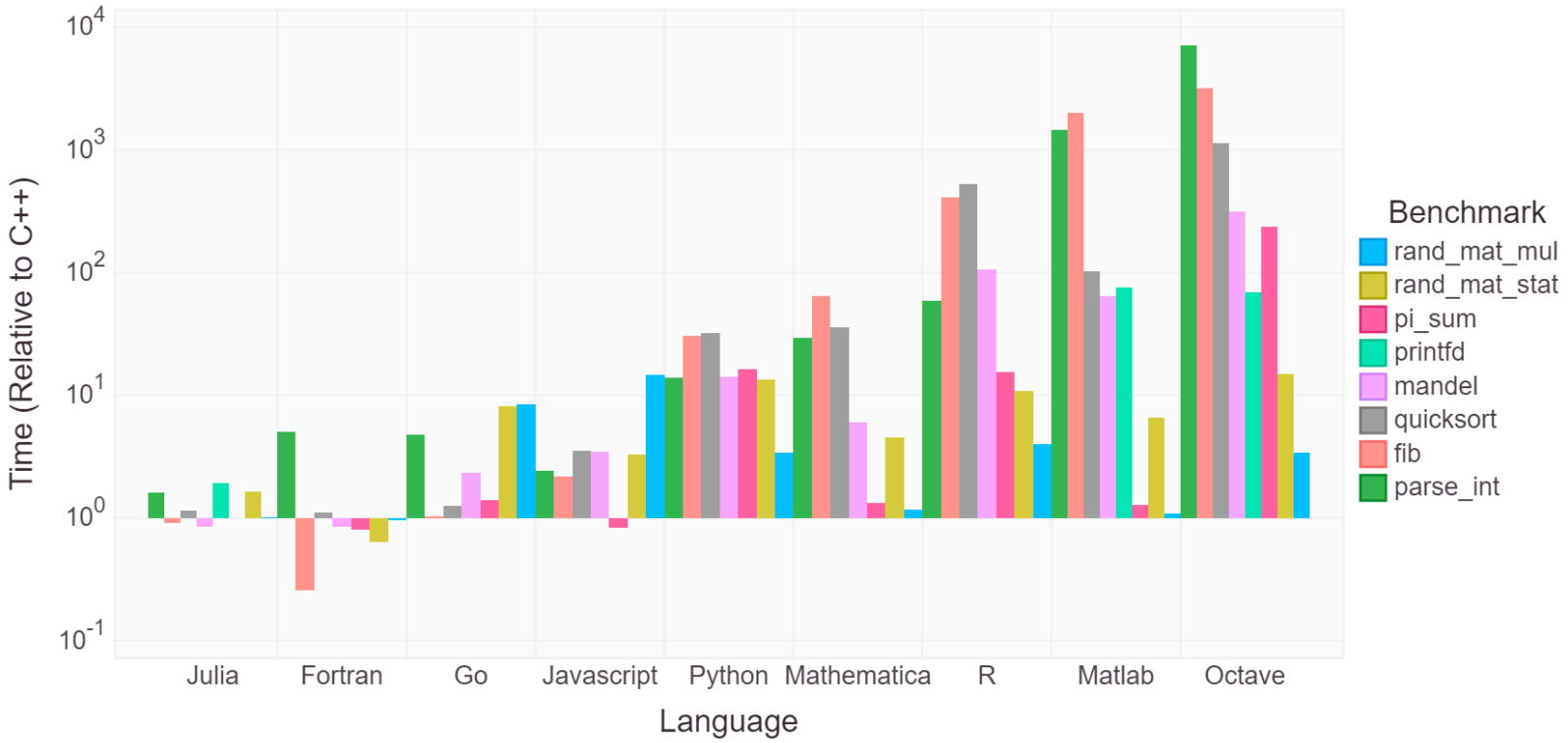
Top 10 IEEE Spectrum Ranking - Google Search Evolution

Los lenguajes **populares** tienden a ser más **robustos**, pero las **tendencias** van cambiando a lo largo del tiempo.



Ref: http://spectrum.ieee.org 2016 - http://www.udacity.com/

Un lenguaje más **eficiente** computacionalmente permite resolver problemas de mayor **complejidad**, tema fundamental en **big data**

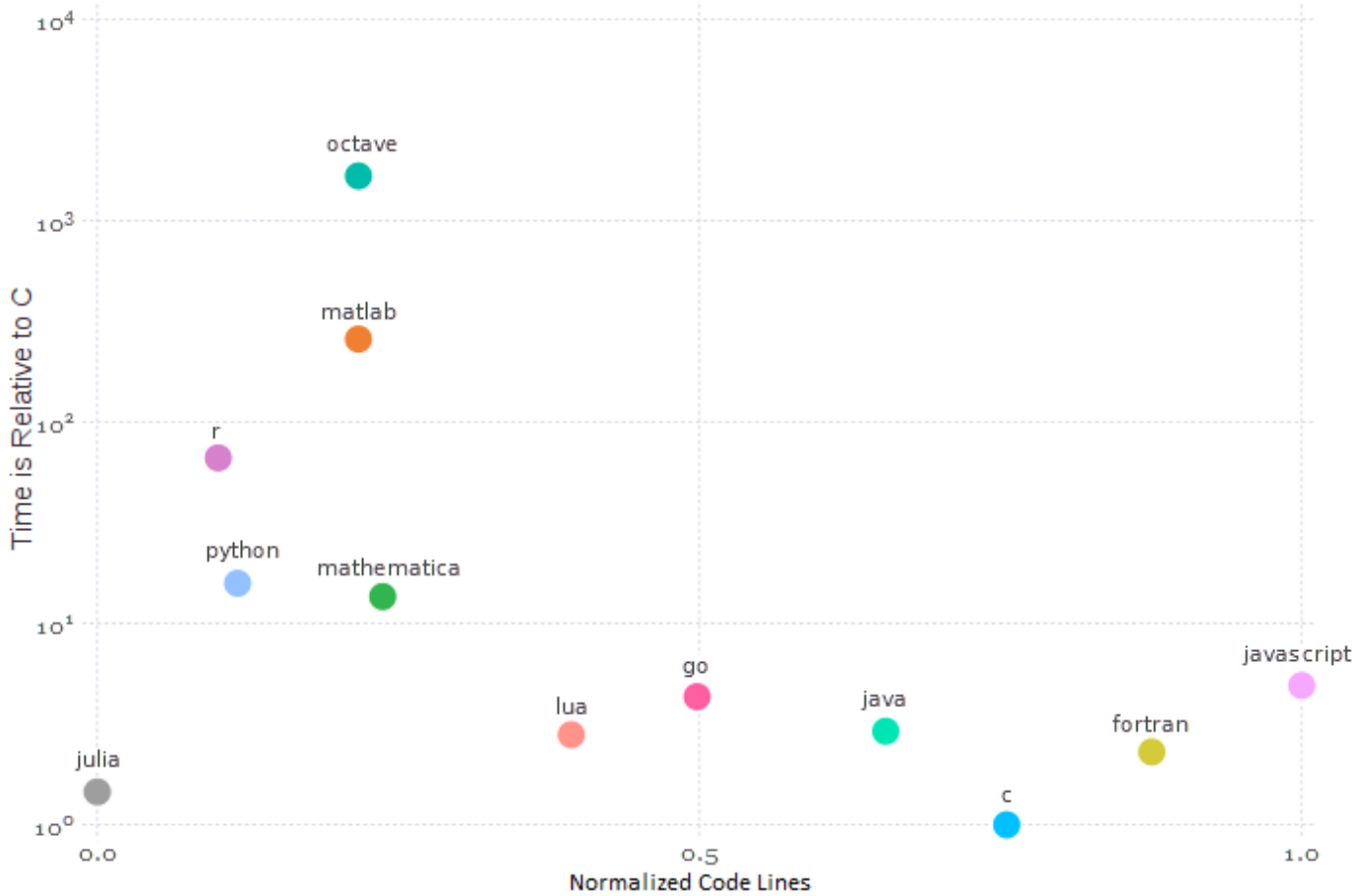


Ref: http://julialang.org/

Programming Languages

Velocidad de Ejecución vs Velocidad de Desarrollo

Al momento de programar, es importante **equilibrar** la velocidad de **ejecución** con la velocidad de **programación.**

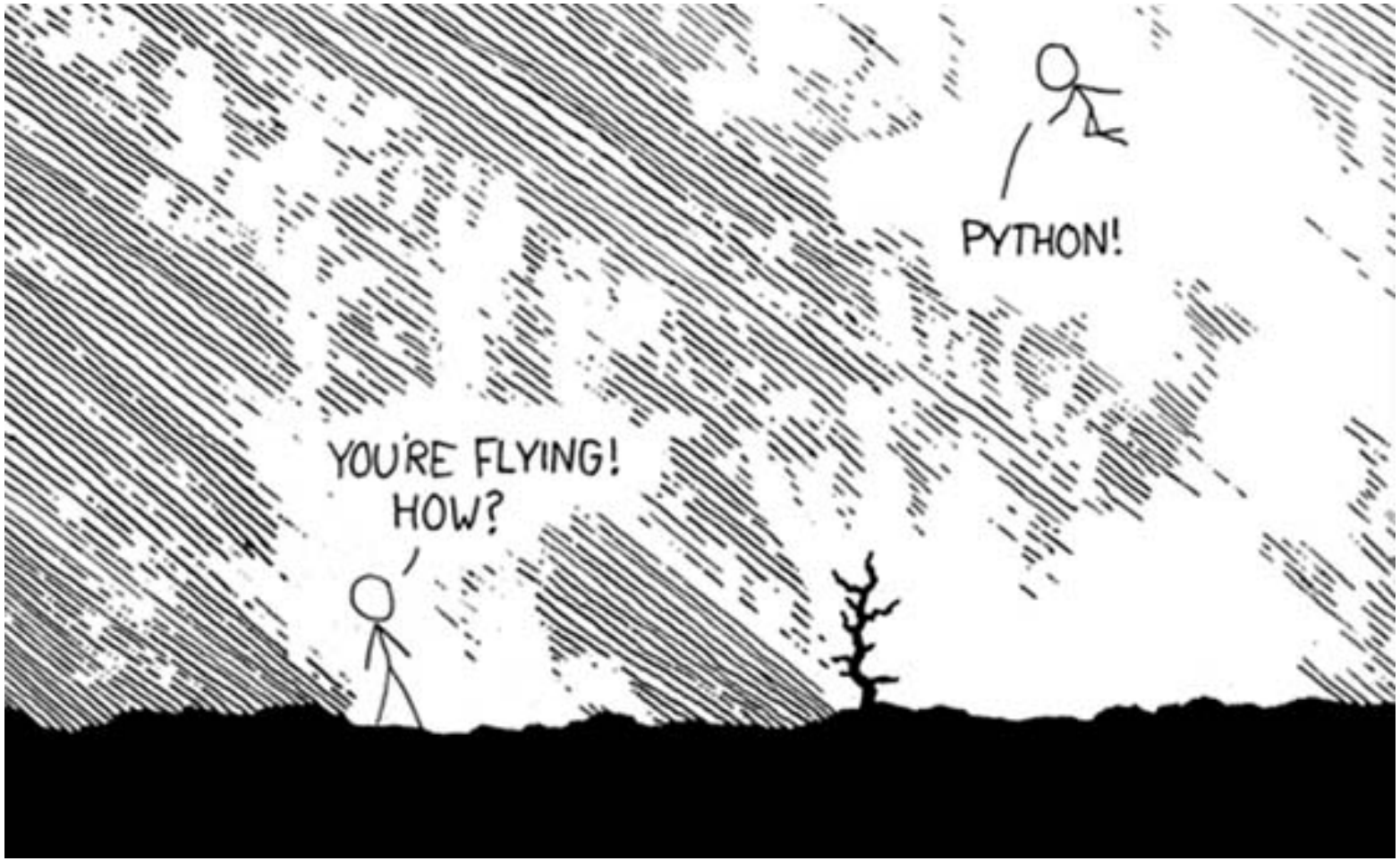


Para poder abarcar problemas de **Big Data**, es necesario poder contar con **librería**s especialmente diseñas para gestionar grandes **volúmenes** de datos y obtener una gran **velocidad** de cálculos en **complejas** estructuras de datos.

- **Apache Spark**: procesamiento de datos a gran escala
  Python, Java, R
- **Apache Mesos**: manejo de recursos computacionales
  Python, Java, C/C++
- **Open MPI**: computación paralela de alto desempeño
  Python, Java, C/C++, R, Fortran, Matlab
- **nVidia CUDA**: programación en GPUs
  Python, Java, C/C++, R Fortran
- **TensorFlow**:modelos de machine deep learning
  Python, C/C++

Programming Languages

Python



Python

Historia

- Monty Python (1969)
- Guido van Rossum (1991)
- Python Software Foundation (2001)
- Python 2.7.13 (2016) / 3.6.2 (2017)

Ref: http://www.python.org

Python reune las características necesarias para **Data Science**, además de ser un buen lenguaje de programación para **uso general**

Ref: https://www.python.org/dev/peps/pep-0020/

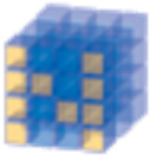
Python cuenta con diversas implementaciones, distribuciones, herramientas y gestores de paquetes.

- **Pip**: Python Package Index.
- **Anaconda**: Python distribution/installer.
- **Conda**: Binary Package Manager.
- **Jupyter**: Web-Based Interactive Notebook.
- **Django**: Framework Web.
- **CPython**: C bytecode interpreter.
- **Jython**: Java bytecode interpreter.
- **IronPython**: .NET bytecode interpreter.
- **PyPy**: Just-in-time compiler (RPython to machine code).
- **MicroPython**: Microcontroller compiler.

Python es capaz de ejecutar código de otros lenguajes de uso general como **C**, **Fortran** y **Java**, además de ser una alternativa completa a **lenguajes científicos** de uso específico

- **Cython**: ejecutar código C
- **F2py**: ejecutar código Fortran
- **Jep**: integrar con Java
- **Rpy2**: integrar con R
- **Numpy/Scipy**: alternativa Matlab
- **Sympy/SageMath**: alternativa a Mathematica

Python dispone de un rico ecosistema compuesto de **librerías open-source** para **matemáticas**, **estadística**, **machine learning** y ciencia en general.



NumPy
Base N-dimensional array package

SciPy library
Fundamental library for scientific computing

Matplotlib
Comprehensive 2D Plotting

IP[y]:
IPython

IPython
Enhanced Interactive Console

Sympy
Symbolic mathematics

pandas
Data structures & analysis

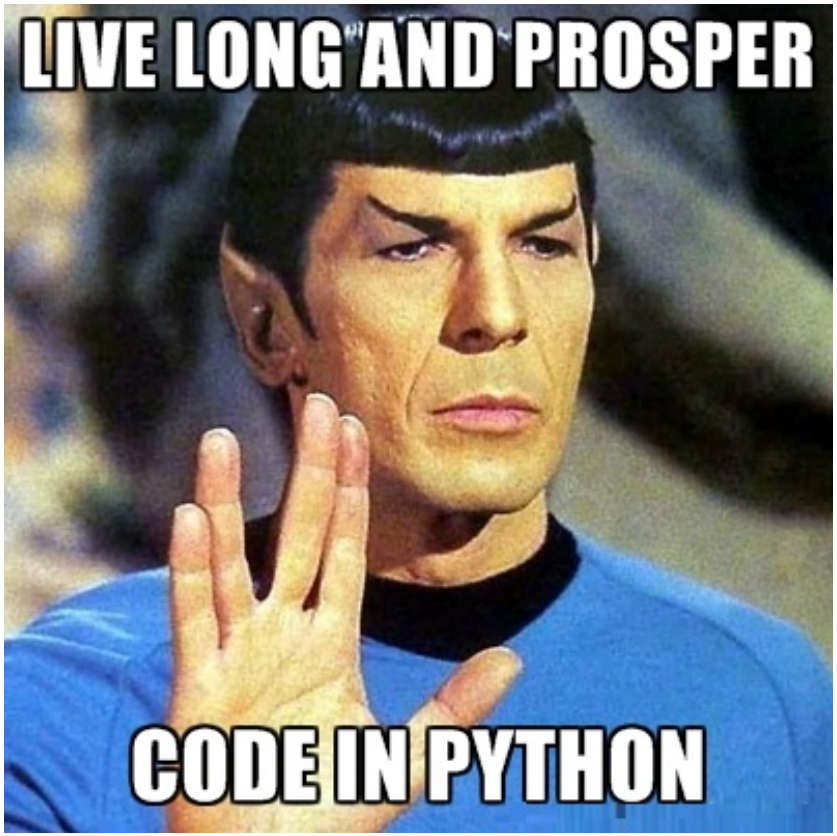
Ref: http://www.scipy.org

Una forma muy simple de aprender **Python** es hacer un curso online, recomiendo revisar **www.sololearn.com**. En las próximas sesiones del ramo aprenderemos:

- **Programación**: Ecosistema, sintaxis, librerias, paradigmas.
- **Manipulación**: Numpy, Pandas.
- **Visualización** : Matplotlib, Seaborn, Bokeh.
- **Modelación**: Scipy, Scikit-Learn, PyMC3.
- **Eficiencia**: Profiling, Compilation, Multiprocessing.

Gracias!

Los esperamos en las próximas sesiones!



PD: Bienvenido a la Iglesia Pythoneana de los Programadores de los Últimos Códigos