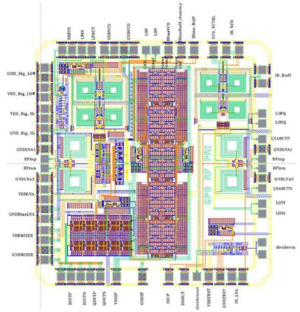
What is Data Science?

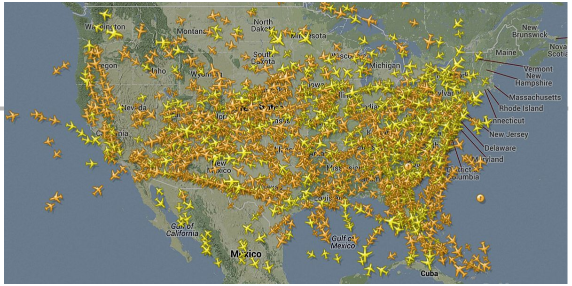
and how to learn it

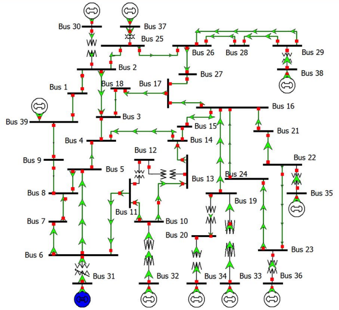
Template by: Laurent Lessard

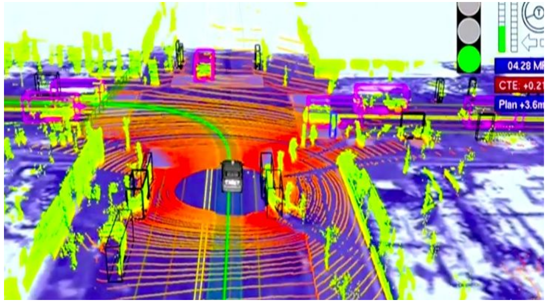
Examples

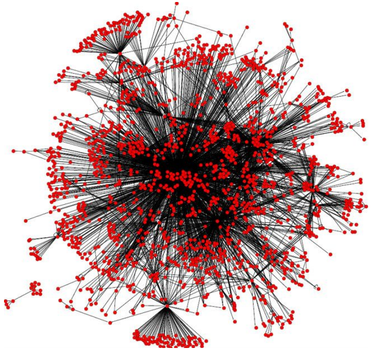
- Engineering
  - Infrastructure and logistics
  - Power systems
  - Wireless telecommunication
  - Robotics and manufacturing
  - Chip design and optimization
  - Autonomous vehicles
- Science
  - Particle physics
  - Biochemistry
  - Neuroscience

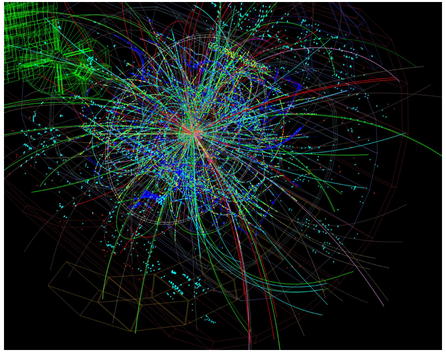






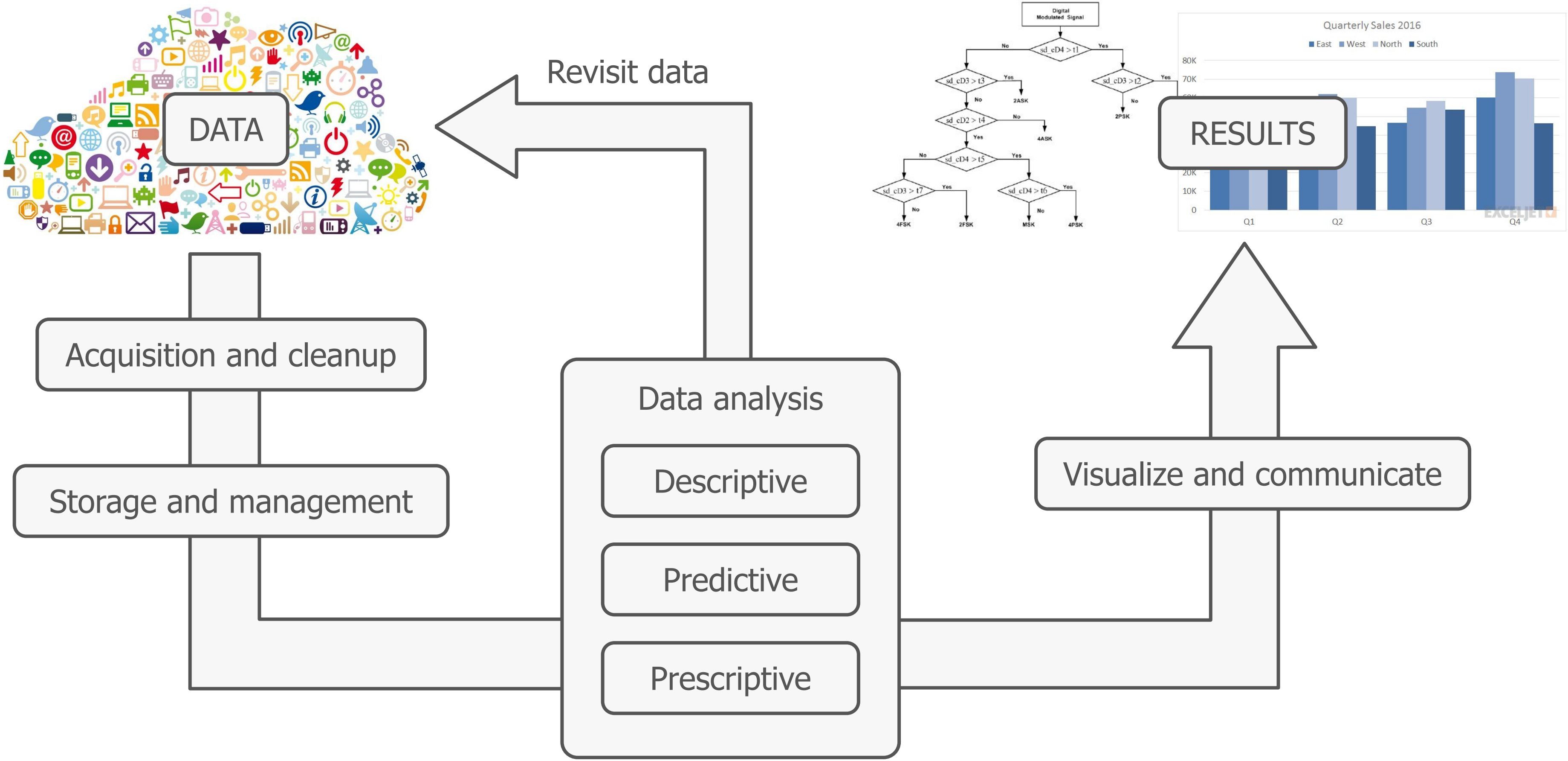






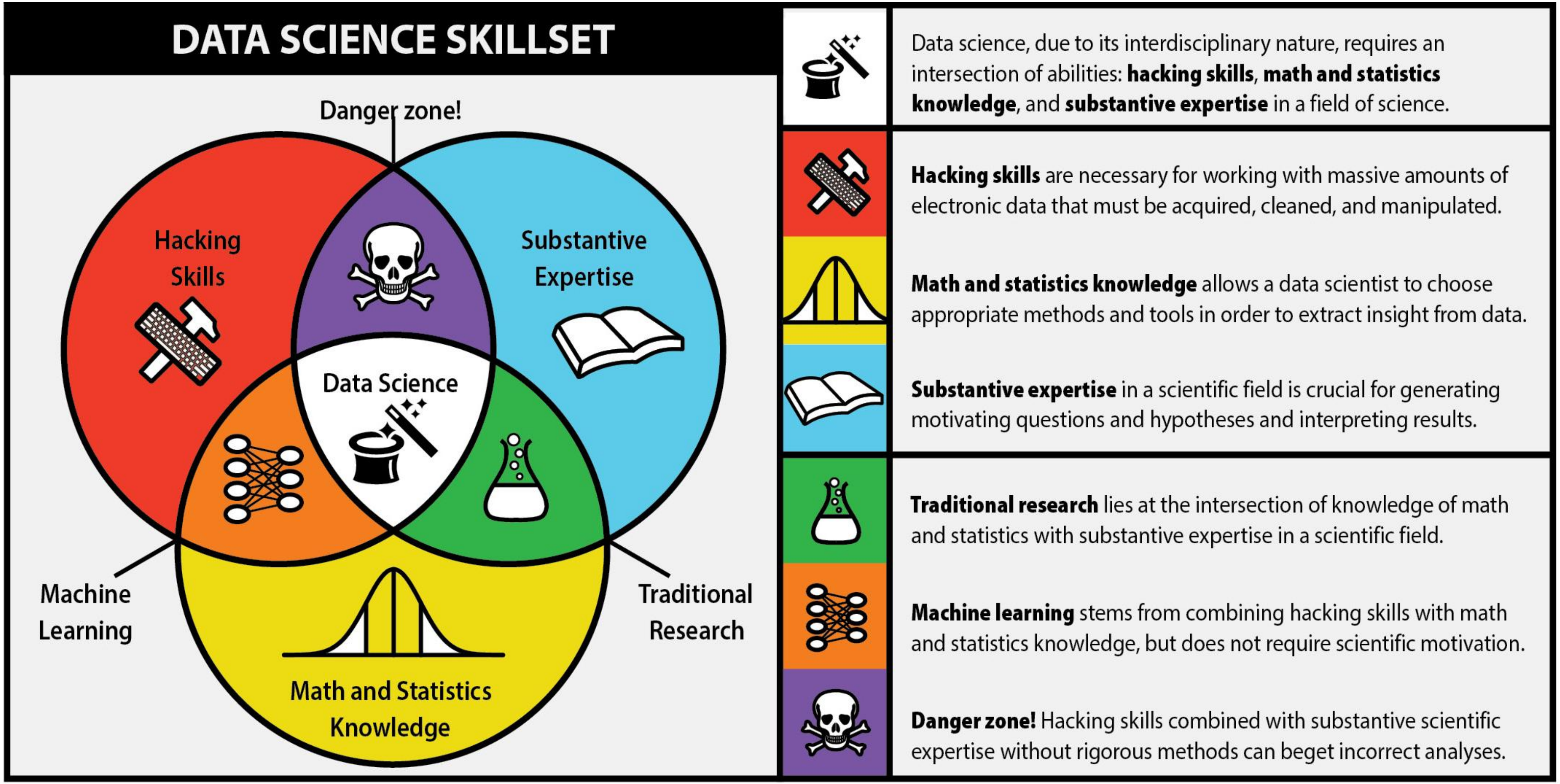
Sources: chip design, flight tracking, power network, self-driving car, E. Coli gene network, LHC data

What is data science?



What's in a name?

- **Data science (DS)**: very broad; everything on the prev. slide!
- **Machine learning (ML)**: everything in the "data analysis" box. Usually used to refer to specific mathematical techniques.
- **Deep learning (DL)**: specific kind of machine learning model used in applications such as image recognition, speech recognition, sentiment analysis, and more.
- **Artificial intelligence (AI)**: very broad; any software designed to make a computer think/behave/learn like a human.



Sources: http://berkeleysciencereview.com/article/first-rule-data-science/
http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Overview of class topics

- “Hacking” skills
  - Coding in Python
  - Using Jupyter Notebooks
  - Managing messy data
  - “Learn to learn”
- Avoiding the danger zone
  - Conceptual and practical pitfalls
  - Bias, privacy, ethical issues
  - Stuff not in the flowchart diagram!
- Descriptive analysis
  - Finding structure in data
  - Communicating/visualizing data
- Predictive analysis
  - Modeling techniques
  - Supervised learning
- Case studies
  - Work with real data
  - Domain-specific issues

Tools and technologies

python

jupyter

SciPy

NumPy

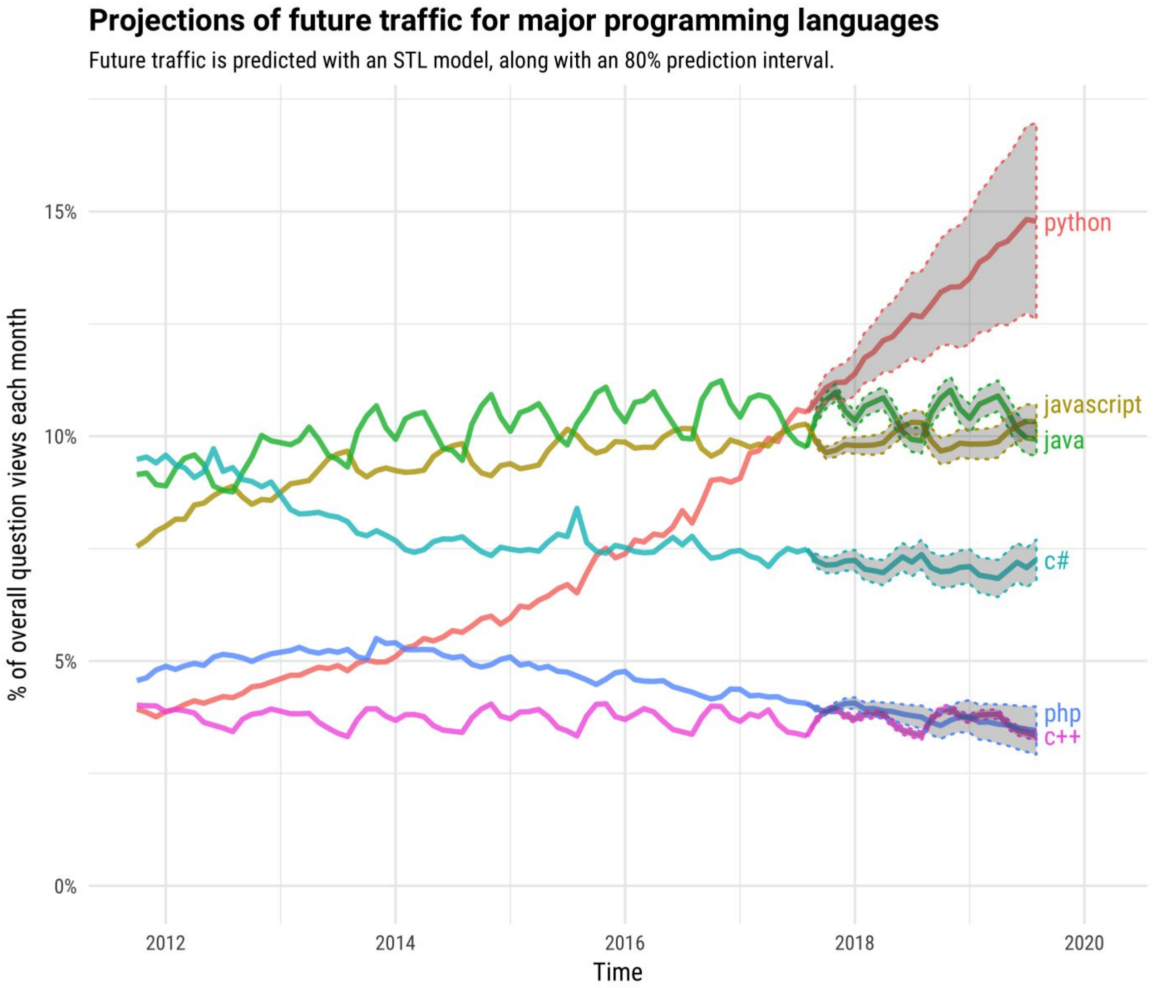
matplotlib

IP[y]: IPython
Interactive Computing

pandas $y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

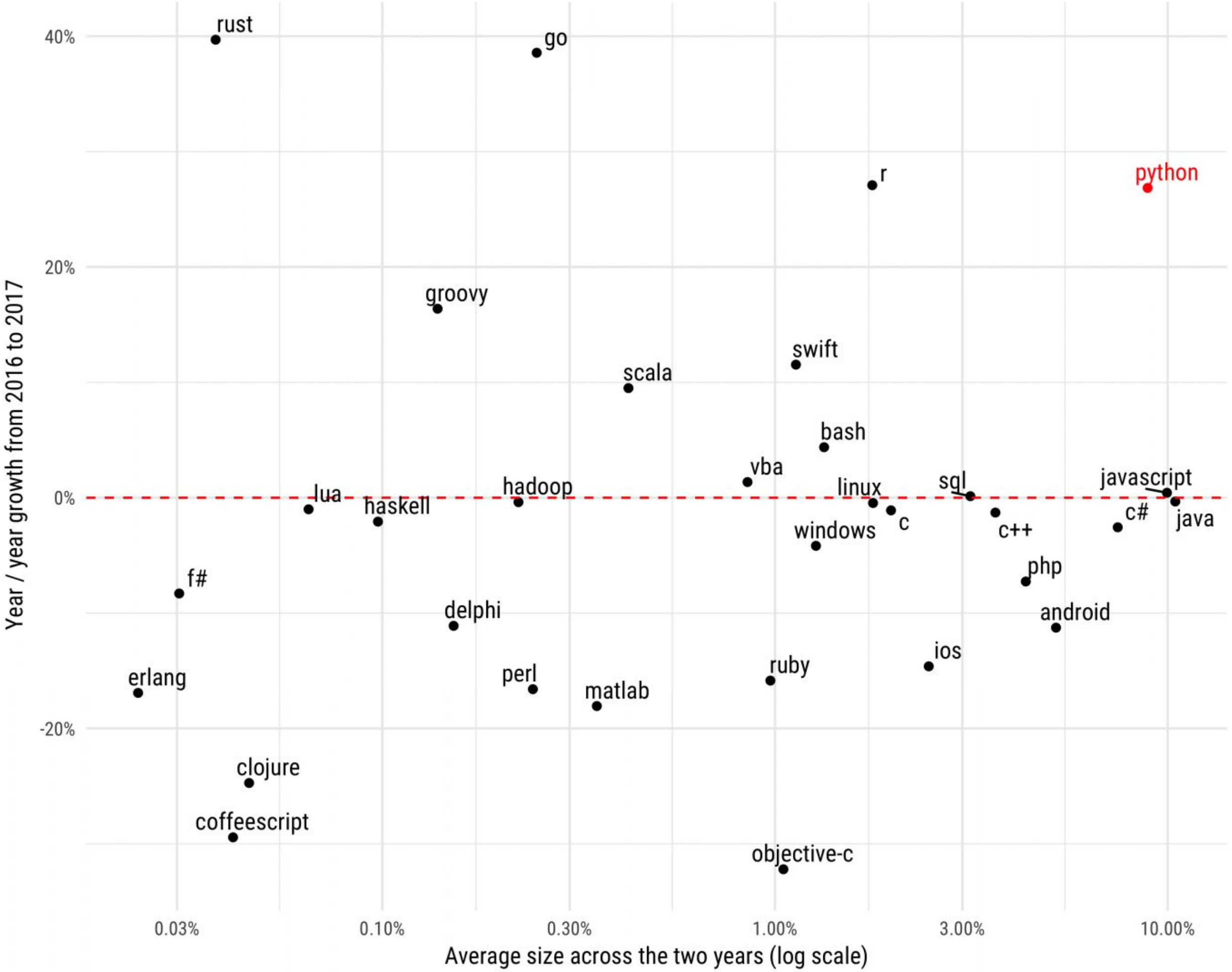
scikit learn

Why Python?



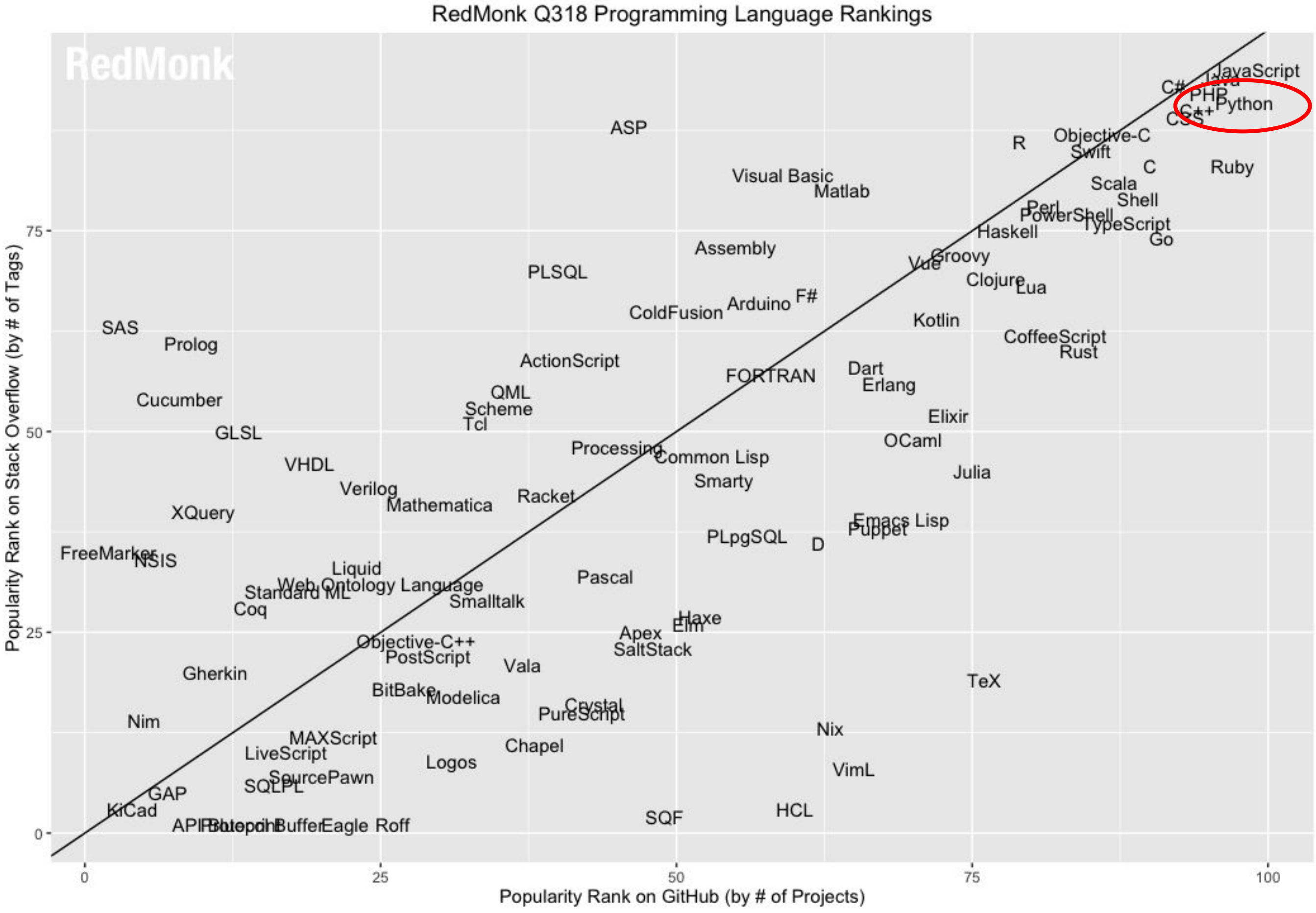
Source: https://stackoverflow.blog/2017/09/06/incredible-growth-python/

Why Python?



Source: https://stackoverflow.blog/2017/09/06/incredible-growth-python/

Why Python?



Source: https://redmonk.com/sogrady/2018/08/10/language-rankings-6-18/

Learn to Learn

What's popular today may not be popular tomorrow

- Strategy #1: learn *concepts* rather than recipes.

Classifiers

What is a classification problem?

What are the desirable properties for a classifier? How does this vary for different applications?

What are typical things to watch out (what can fail?) when solving a classification problem?

vs

Classifiers

linear least squares, naive Bayes, logistic regression, support vector machines, linear discriminant analysis, k-nearest neighbor, decision trees, boosted trees, random forests, perceptron, neural networks, convolutional nets, deep nets,...

Learn to Learn

What's popular today may not be popular tomorrow

- Strategy #2: learn how to figure things out!

Administrative stuff

- Syllabus (see Canvas page)