STA 221: LECTURE 1

krishna balasubramanian

(University of California, Davis)

COURSE DETAILS

- ▷ Instructor: Krishna Balasubramanian
- ▷ Office: Zoom
- ▷ Office Hours: Monday 4:30-5:30PM
- ▷ Email: kbala@ucdavis.edu
- ▷ TA: Si Teng Hao

COURSE DETAILS

- Learning Objecives:
  - Learning to use Python for big-data analytics
  - Learning popular Machine Learning Techniques

PREREQUISITES

- ▷ Basics of python ?
- ▷ Basic statistics and machine learning ?
- ▷ Basics of Linear Algebra ?

Rate yourself out of 10 (10 being expert) in each of the above.

EVALUATION

- 3 homeworks: 60% of total points
  - Each homework is worth 20%.
- Final project: 40% of total points
  - May 4th and 6th proposal (individual meetings)
  - June 2nd and 3rd demonstration (individual meetings)
  - Final report due June 10th.

INSTRUCTIONS

- Follow instructions for each homework and project carefully!
- Late submission **WILL NOT** be accepted.
- Follow UC Davis Code of Academic Conduct carefully!
- Violations **WILL NOT** be tolerated!

COURSE CONTENT

- Two master algorithms:
    - Power method for computing eigenvectors
    - Gradient descent for optimization
    - Note: Both methods are kind of related.
- Bonus algorithm:
    - Randomized matrix multiplication - for learning/testing basics of Python

COURSE CONTENT

- Computing eigenvectors: Unsupervised learning. We will use scikit learn python package.
- Gradient descent: Supervised learning. We will use scikit learn and Pytorch.

UNSUPERVISED LEARNING

- Given large amounts of unlabelled data:
  - How to group/cluster them accordingly ?
  - How to visualize them for exploratory data analysis ?
- Some techniques (all based on eigenvalue computation):
  - k-means, mixture models, spectral clustering, page rank.
  - PCA, Kernel PCA, Manifold Learning

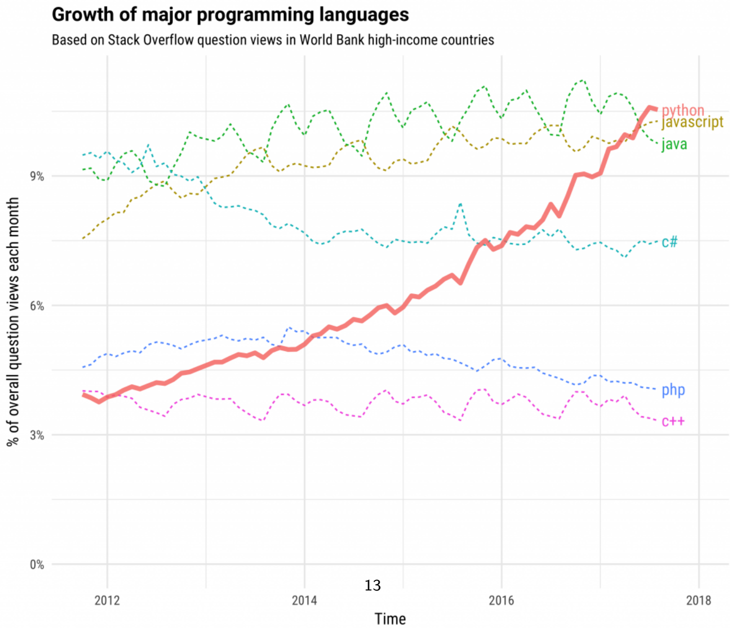
SUPERVISED LEARNING

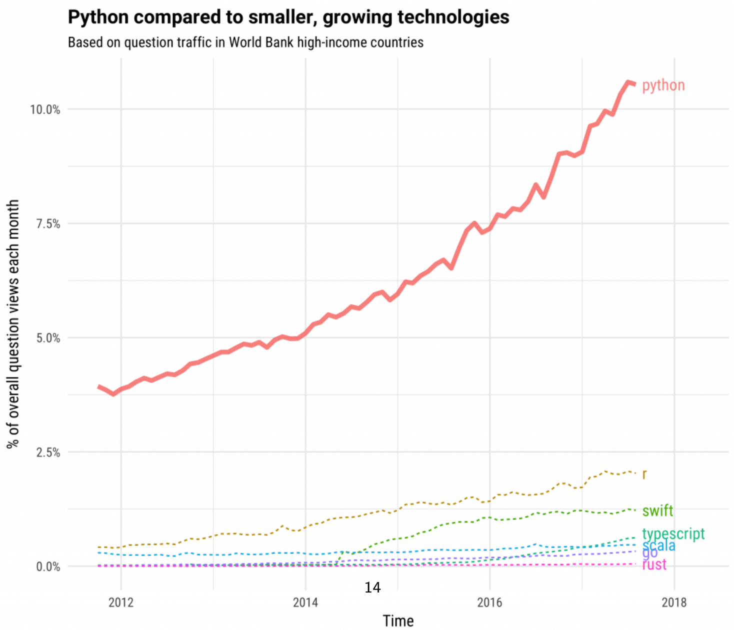
- Predict/Classify new data based on learning from labelled training data:
  - How to efficiently learn from labeled data ?
- Techniques (all based on gradient descent (or variations)):
  - Discriminant analysis, Logistic Regression
  - Support vector machine, Deep neural networks

MAIN THEOREM

NO FREE LUNCH !

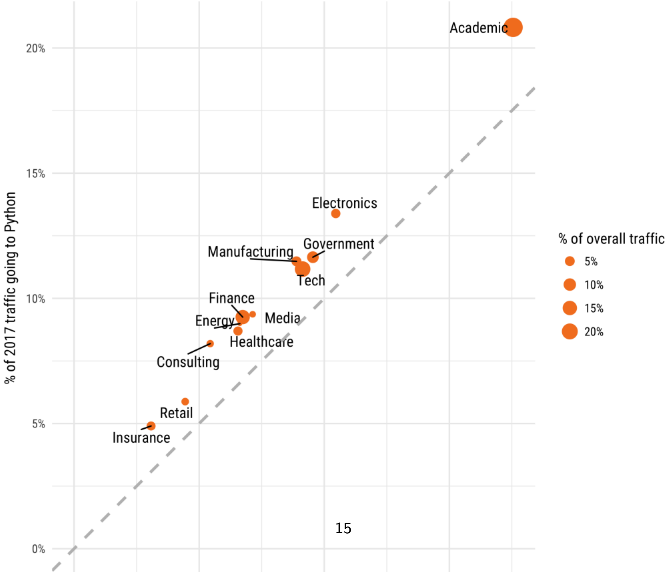
Lecture 1: Python Basics


Growth of major programming languages
Based on Stack Overflow question views in World Bank high-income countries
% of overall question views each month
9%
6%
3%
0%
2012
2014
2016
2018
Time
python
javascript
java
c#
php
c++

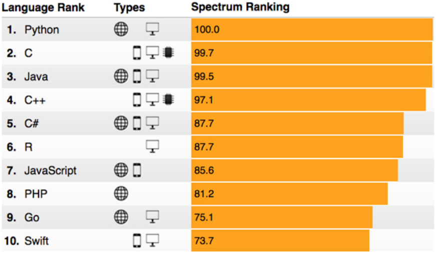

Python compared to smaller, growing technologies
Based on question traffic in World Bank high-income countries
% of overall question views each month
10.0%
7.5%
5.0%
2.5%
0.0%
2012
2014
2016
2018
Time
python
r
swift
typescript
scala
go
rust

Traffic by industry to Python

Comparing Jan-Aug of each year, in the United States and United Kingdom.

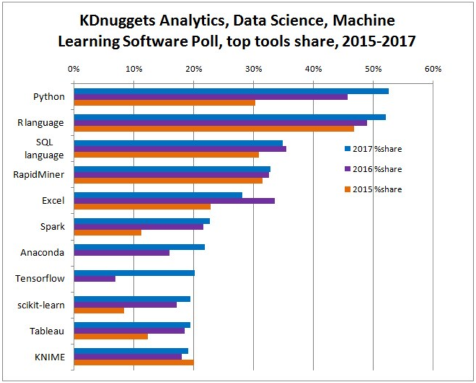


IEEE spectrum



Language Rank	Types	Spectrum Ranking
1. Python	🌐 🖥	100.0
2. C	📱 🖥 ▦	99.7
3. Java	🌐 📱 🖥	99.5
4. C++	📱 🖥 ▦	97.1
5. C#	🌐 📱 🖥	87.7
6. R	🖥	87.7
7. JavaScript	🌐 📱	85.6
8. PHP	🌐	81.2
9. Go	🌐 🖥	75.1
10. Swift	📱 🖥	73.7

WHY PYTHON


KDnuggets Analytics, Data Science, Machine Learning Software Poll, top tools share, 2015-2017
0%
10%
20%
30%
40%
50%
60%
Python
R language
SQL language
RapidMiner
Excel
Spark
Anaconda
Tensorflow
scikit-learn
Tableau
KNIME
🟦 2017 %share
🟪 2016 %share
🟧 2015 %share