How to Become a Data Scientist

Ilya Ezepov

AlphaGo (by Google)

AlphaGo

?

Lee Sedol,
Last decade top player

AlphaGo (by Google)

AlphaGo 4:1

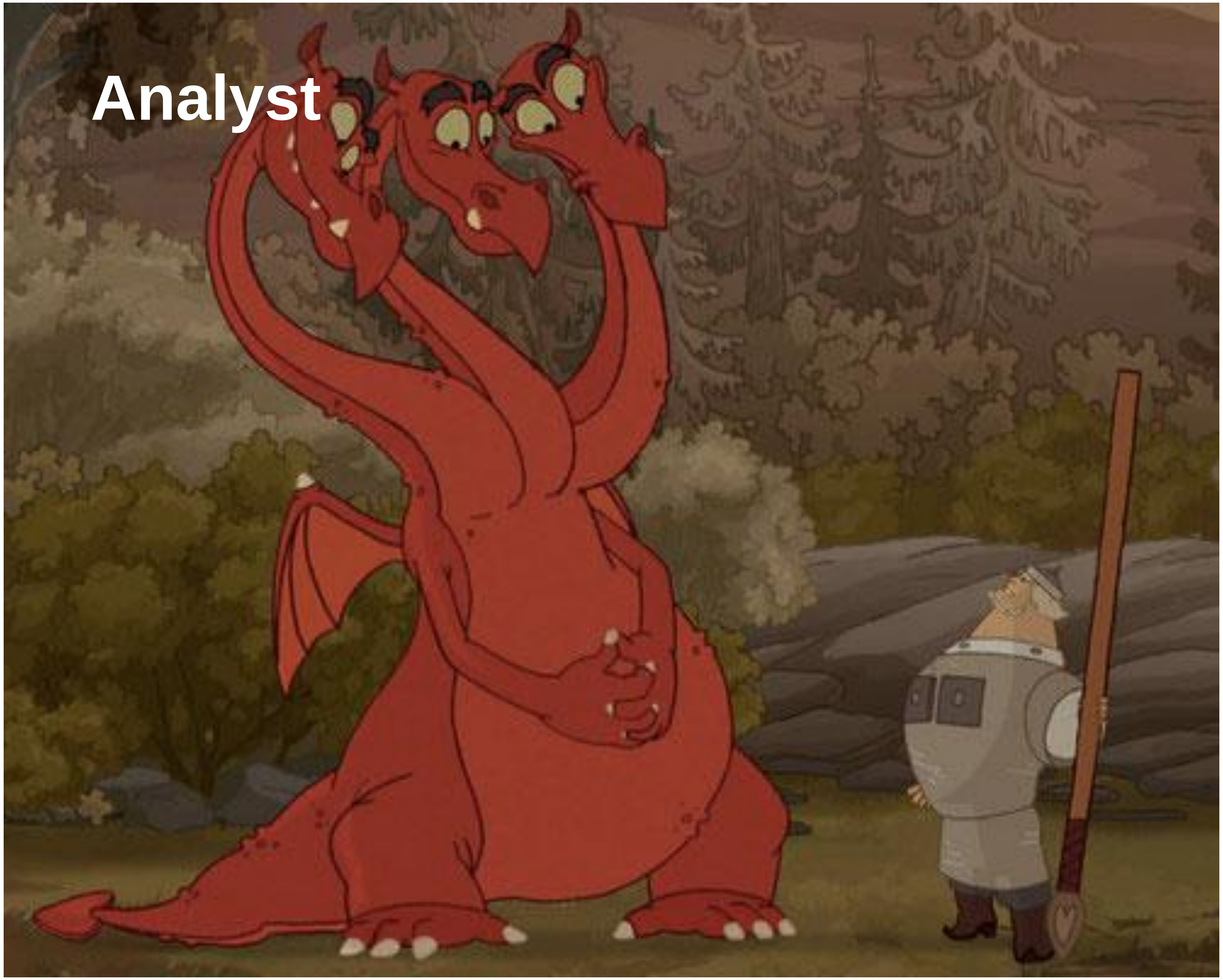
Lee Sedol,
Last decade top player

Agenda

- Math skills
- Programming skills
- Tech skills
- Soft skills
- Experience

Types of data scientist

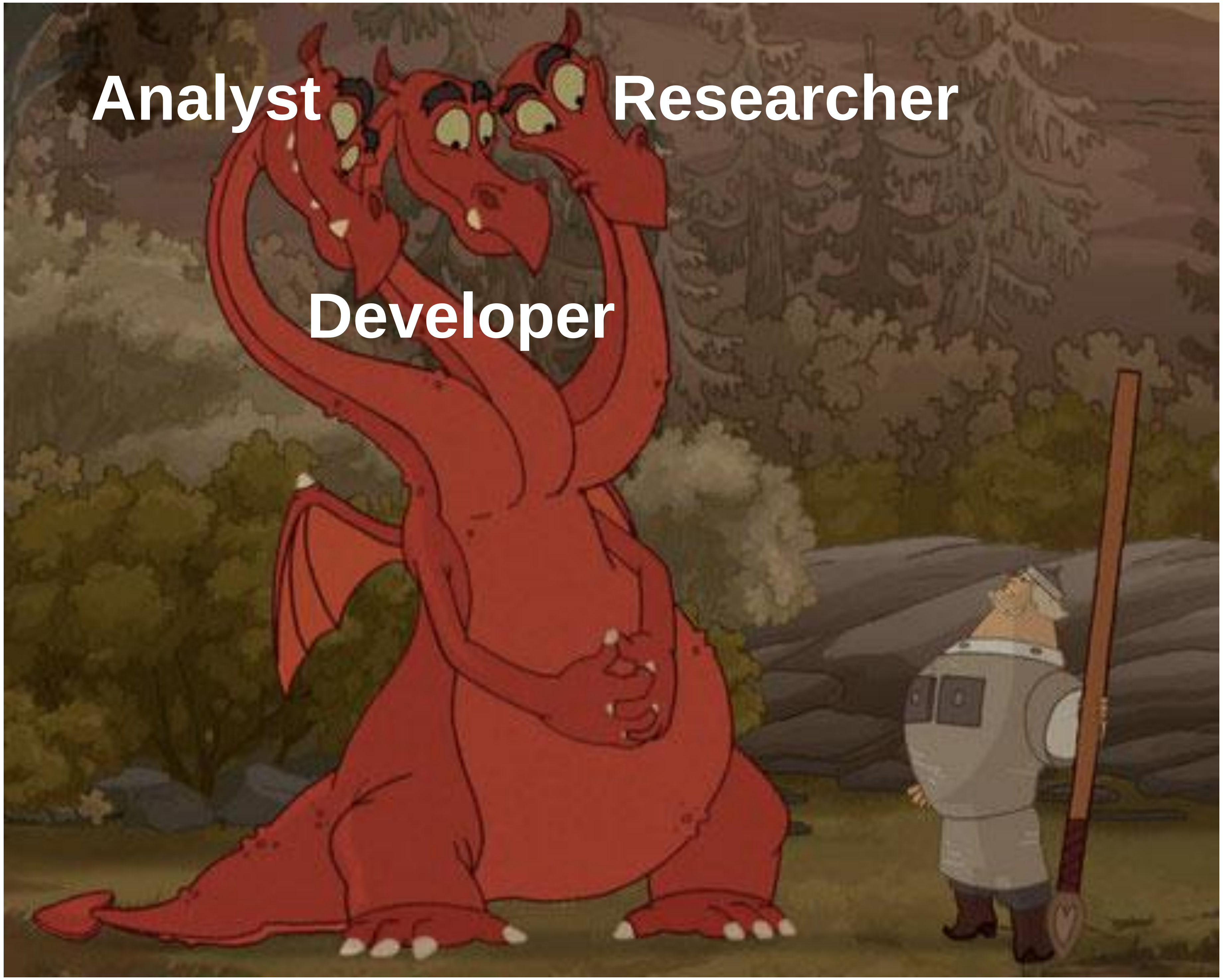
Types of data scientist



Types of data scientist

Analyst
Developer

Types of data scientist



Math

Math skills

- Theory of probability and mathematical statistics
- Calculus and algebra
- Calculus of variations and optimal control

- Not that much for analyst and developer ways

- No upper limit for researcher way

Math skills

- **Moscow is** (one of) **the best place to study!**
- A lot of great universities: MSU, HSE, MIPT, etc.
- Yandex School of Data Analysis
- IUM
- Big and active community: meet-ups, hackathons, lectures, courses, challenges, etc.

Machine learning resources

- **Andrew Ng course @ Coursera**
- https://www.coursera.org/learn/machine-learning
- 11 weeks x (~40 min videos, quiz, assignment)
- Assignments on Octave/MATLAB :(
- Video & quiz + assignment on python/R just for yourself is absolutely ok
- The classic machine learning course: the best place to start (with low math level)

Machine learning resources

- **Yandex video lectures (russian)**
- https://yandexdataschool.ru/edu-process/courses
- 24 x 90 min
- High math level

Machine learning resources

- **Yandex & HSE @ Coursera (russian)**
- https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie
- 7 weeks x 60 min
- Simplified version of Yandex Data School

Machine learning resources

- **Mining Massive Datasets @ Coursera**

- https://www.coursera.org/course/mmds
- http://mmds.org

- 7 weeks x (~200 min videos)

- A lot of material

- Not only about machine learning, but also about general data analysis

Machine learning resources

- **MIT AI course**
- http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-034-artificial-intelligence-fall-2010/
- 22 x 45 min
- ML course from Electrical Engineering and Computer Science department

Machine learning resources

- **Deep Learning @ Udacity**
- https://www.udacity.com/course/deep-learning--ud730
- Brief and clean introduction
- Assignments on TensorFlow (new python library)

Machine learning resources

- **Deep Learning**

- http://deeplearning.net/tutorial/index.html
- http://deeplearning.net/reading-list/

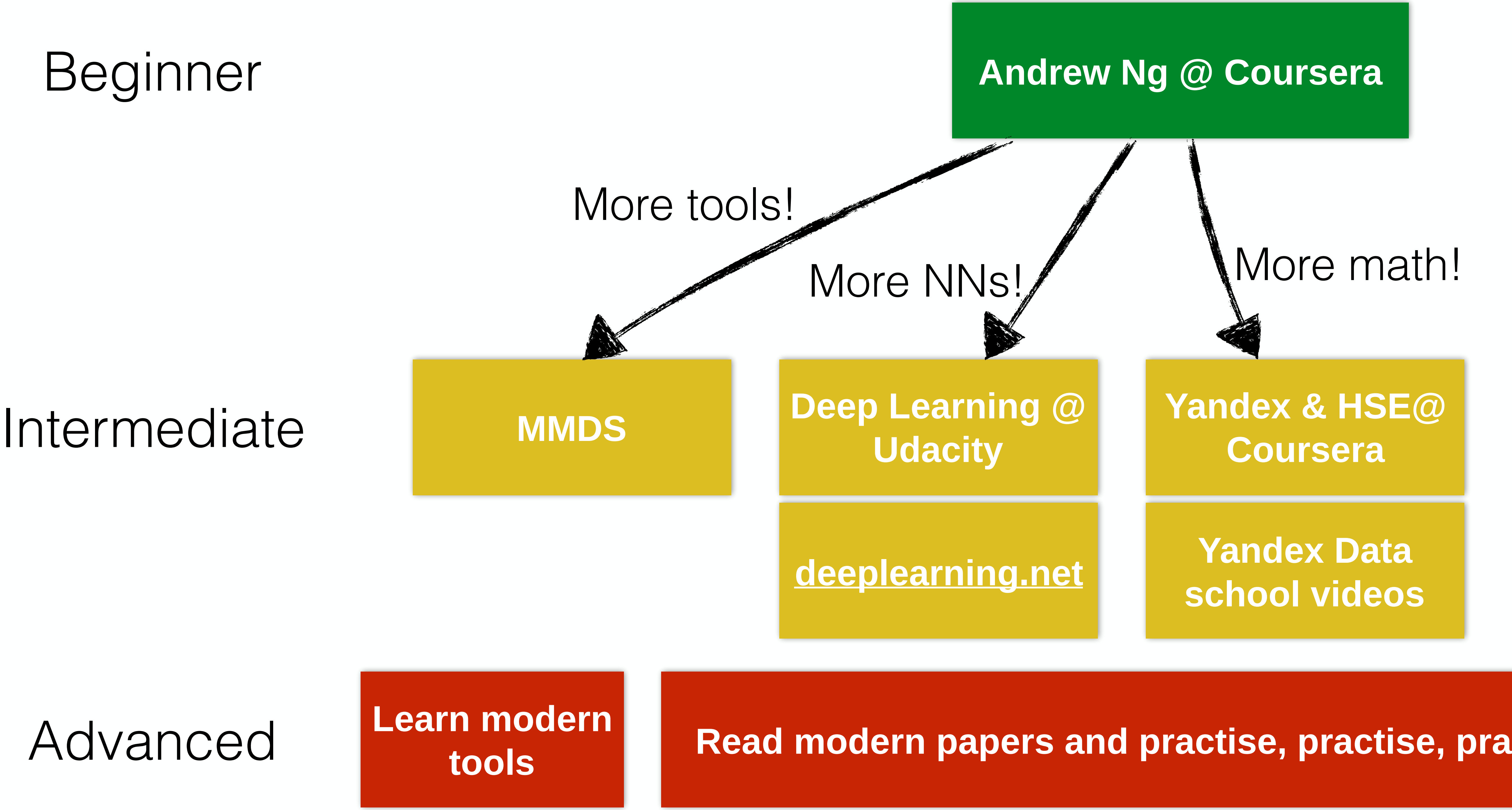
- Python tutorials (with Theano, python lib for tensor calculations)

- Deep understanding of deep learning

Machine learning resources

- **Reinforcement learning (David Silver)**
- http://www.youtube.com/watch?v=2pWv7GOvuf0&list=PL5X3mDkKaJrL42i_jhE4N-p6E2Ol62Ofa
- 10 x 90 min
- Great place to learn one the most hot field of machine learning

Machine learning resources



Beginner

Andrew Ng @ Coursera

More tools!

More NNs!

More math!

Intermediate

MMDS

Deep Learning @ Udacity

Yandex & HSE@ Coursera

deeplearning.net

Yandex Data school videos

Advanced

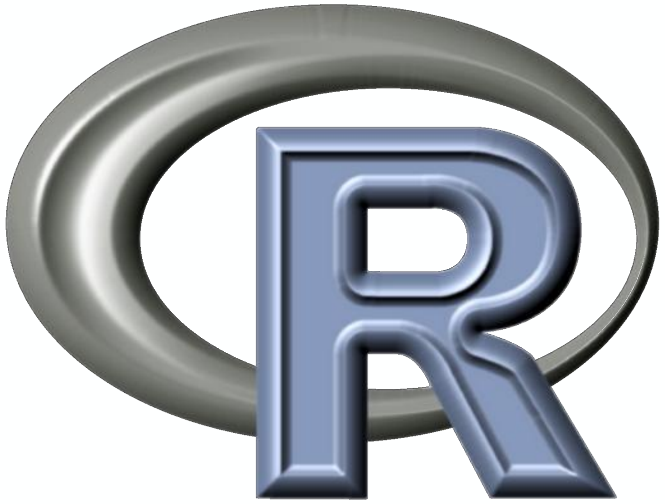
Learn modern tools

Read modern papers and practise, practise, pra

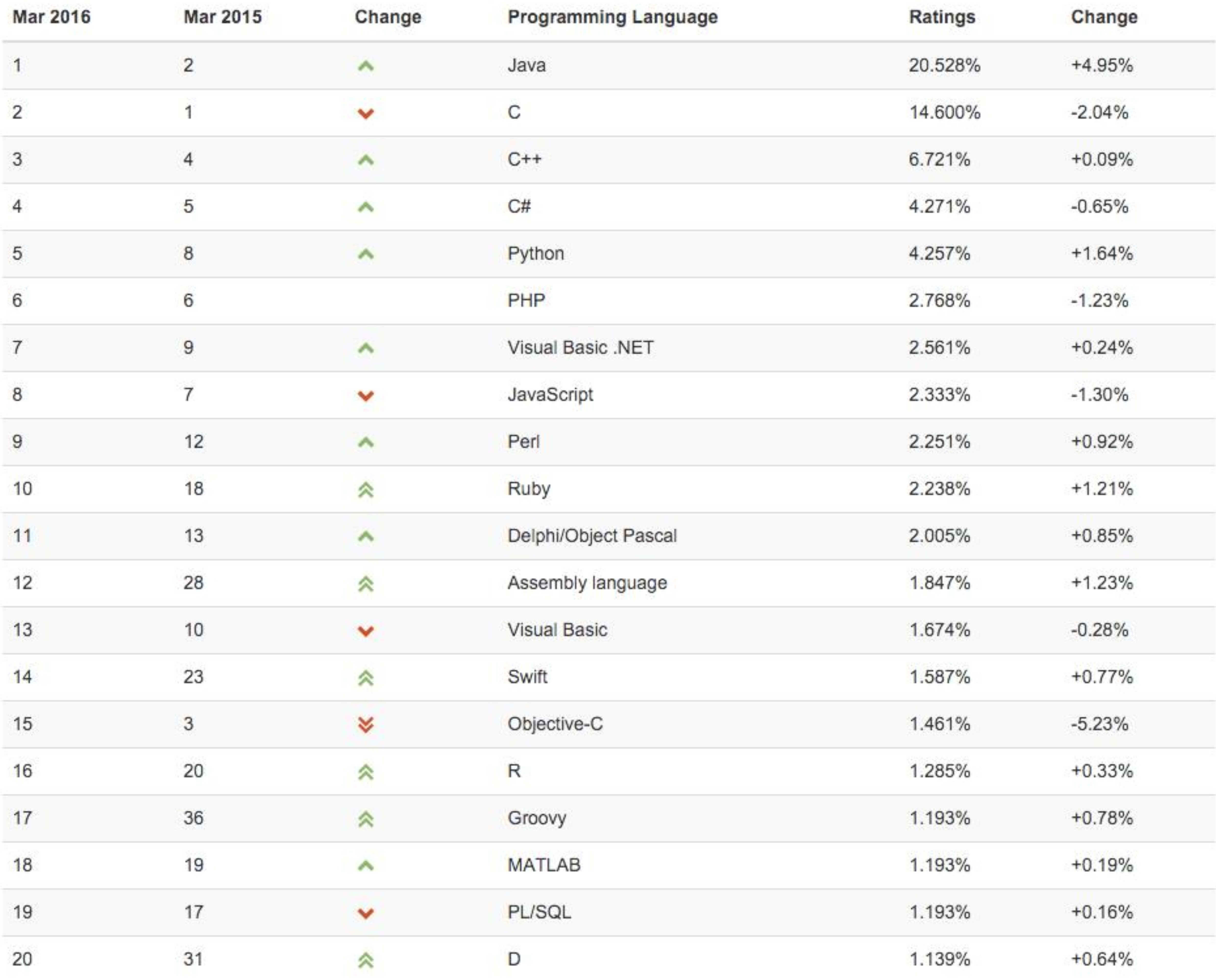
Programming

The best language?

Useful languages:



Fast languages:



Mar 2016	Mar 2015	Change	Programming Language	Ratings	Change
1	2	▲	Java	20.528%	+4.95%
2	1	▼	C	14.600%	-2.04%
3	4	▲	C++	6.721%	+0.09%
4	5	▲	C#	4.271%	-0.65%
5	8	▲	Python	4.257%	+1.64%
6	6		PHP	2.768%	-1.23%
7	9	▲	Visual Basic .NET	2.561%	+0.24%
8	7	▼	JavaScript	2.333%	-1.30%
9	12	▲	Perl	2.251%	+0.92%
10	18	▲▲	Ruby	2.238%	+1.21%
11	13	▲	Delphi/Object Pascal	2.005%	+0.85%
12	28	▲▲	Assembly language	1.847%	+1.23%
13	10	▼	Visual Basic	1.674%	-0.28%
14	23	▲▲	Swift	1.587%	+0.77%
15	3	▼▼	Objective-C	1.461%	-5.23%
16	20	▲▲	R	1.285%	+0.33%
17	36	▲▲	Groovy	1.193%	+0.78%
18	19	▲	MATLAB	1.193%	+0.19%
19	17	▼	PL/SQL	1.193%	+0.16%
20	31	▲▲	D	1.139%	+0.64%

The most important language

Fast languages

- C++ or Java is must-know for developer …
- And absolutely ok not to know for analyst
- So, experience with one is a plus, but not worth it to start learning

Useful languages

- Python or R: let the holy war begin!

Useful languages

- Python or R: let the holy war begin!
- My experience: ~2 years of R programming with complete switching to python
- Python is much wider:
  - A lot of machine learning libraries
  - Fast calculations (via numpy)
  - Web development
  - Game development
  - Enterprise development

Where to learn python?

- Great intro:
  - https://www.codecademy.com/learn/python
- Another one:
  - http://learnpythonthehardway.org/book/
- Next try to solve real problems (math, financial modelling, web-parsing, Kaggle, etc.)

Strange languages

“PHP is a minor evil perpetrated and created by incompetent amateurs, whereas Perl is a great and insidious evil perpetrated by skilled but perverted professionals.”

Jon Ribbens

Tech

Linux/UNIX

- Linux is great. But how to start?

- Virtual Machine: resources needed; pointless
- Setup Jupyter notebook on Amazon AWS t2.micro
- Run a web server on Digital Ocean
- For mac users: just open Terminal app

- Intro to bash: https://www.codecademy.com/learn/learn-the-command-line

Version control systems

- Learn Git, because:
  - Data scientist often work in teams
  - There is a lot of great stuff on github
  - It’s must-known to work in big companies
  - Sometimes it’s good even for personal long-term big projects
  - And it’s cool to have popular github account

- Use on of these:
  - https://www.codecademy.com/learn/learn-git
  - https://www.codeschool.com/learn/git
  - https://try.github.io

MapReduce

- Know MapReduce paradigm
- Know how to solve basic problems: word count, tf-idf, histograms, etc.
- Setup your single node Hadoop cluster
- Buy 2+ Raspberry Pi and setup real Hadoop cluster

Soft

Soft skills

- Communication
- Team work
- Imagination

Experience

Do Kaggle!

- Register on Kaggle
- Do Titanic challenge
- Do Digit Recognizer/Facial Keypoint challenges if you liked deep neural networks
- Do some real challenges. Participate in everything

Do Kaggle!

- Register on Kaggle
- Do Titanic challenge
- Do Digit Recognizer/Facial Keypoint challenges if you liked deep neural networks
- Do some real challenges. Participate in everything

Read no free hunch blog!

http://blog.kaggle.com/

Numerai

- Hedge fund with open data
- Like Kaggle, but only one on-going challenge
- Predict stock prices

https://numer.ai/

Kaggle-like

- https://www.innocentive.com/ar/challenge/browse
- http://tunedit.org/challenges
- https://www.crowdanalytix.com/community
- https://www.hackerrank.com/
- There always are a lot of challenges from companies

BlackBox

- Russian competition on reinforcement learning
- You need to play game with un-known rules

blackboxchallenge.com

“Data Scientist:
The Sexiest Job of the 21st Century.”

–Harvard Business Review