Introduction to data science

Patrick Shafto

Department of Math and Computer Science

Plan for today

- Discuss homework
- What is data science?
- Programming for data science

Homework

- Hand in your homework on time!

Homework

- Hand in your homework on time!
- Hand in your homework on time!

Homework

- Hand in your homework on time!
- Hand in your homework on time!
- Hand in your homework on time!

Homework

- Hand in your homework on time!
- Hand in your homework on time!
- Hand in your homework on time!
- Hand in your homework on time!

Homework

- Hand in your homework on time!
- Hand in your homework on time!
- Hand in your homework on time!
- Hand in your homework on time!
- Sunday by midnight for Monday's class. Tuesday by midnight for Wednesday's class.

HW1 - Rubric

- Excellent
- Good
- Fair
- Poor

Explain your rating!

Feedback to Learner

Solution to all the subsections looks good. It would have been better if you would have named each subsection. It becomes more readable. Rest everything is fine.

Feedback to Learner

Please provide the comments where ever required, there is no segregation between the different examples, if you could start by saying "Sorting"/"Dictionary", that would be of help.

Feedback to Learner

1.No markdowns 2.Slicing of strings in not present 3.Dictionary demo could have been more detailed (using for loop and built in functions such as del)

Feedback to Learner

You have really worked hard for this assignment. You can improve it a little better by first defining terms like what are strings and why do we need to use them and where? in your examples, you could have shown how to concatenate strings and functions like how do we take a particular string statement onto the next line which is a \n. In the lists, you could have given few more examples like min, max and reverse operations so that a user with coding experience could even learn from it. Overall it was a good tutorial just a little extra effort and it will be excellent.

Feedback to Learner

Good Work!!

Feedback to Learner

A good touch to the basic Python language. Kudos!

Feedback to Learner

Incomplete Assignment

Feedback to Learner

the assignment is good but it should have been more precised. It should have been more distinct . Use markdown to give a title to each of the following.

HW2

- Excellent
- Good
- Fair
- Poor

Explain your rating!

What is data science?

Data science

From Wikipedia, the free encyclopedia

Not to be confused with information science.

Data science, also known as **data-driven science**, is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured,[1][2] similar to data mining.

Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data.[3] It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, data mining, databases, and visualization.

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.[4][5]

When Harvard Business Review called it "The Sexiest Job of the 21st Century" [6] the term became a buzzword, and is now often applied to business analytics,[7] or even arbitrary use of data, or used as a sexed-up term for statistics.[8] While many university programs now offer a data science degree, there exists no consensus on a definition or curriculum contents.[7] Because of the current popularity of this term, there are many "advocacy efforts" surrounding it.[9]

What is data science?

- Method: Check out universities
  - Berkeley

What is Data Science?

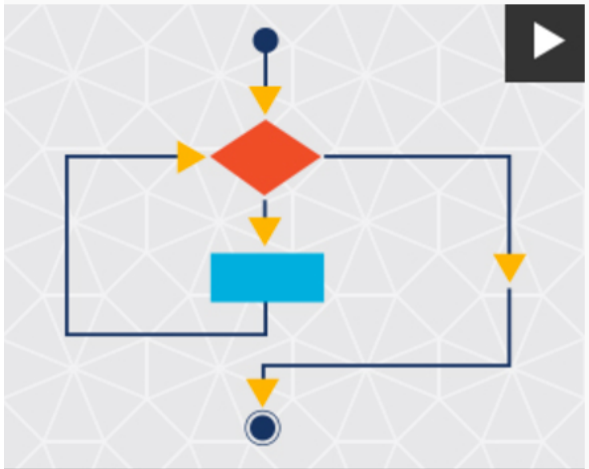
A New Field Emerges

There is significant and growing demand for data-savvy professionals in businesses, public agencies, and nonprofits. The supply of professionals who can work effectively with data at scale is limited, and is reflected by rapidly rising salaries for data engineers, data scientists, statisticians, and data analysts.

A recent study by the McKinsey Global Institute concludes, "a shortage of the analytical and managerial talent necessary to make the most of Big Data is a significant and pressing challenge (for the U.S.)." The report estimates that there will be four to five million jobs in the U.S. requiring data analysis skills by 2018, and that large numbers of positions will only be filled through training or retraining. The authors also project a need for 1.5 million more managers and analysts with deep analytical and technical skills "who can ask the right questions and consume the results of analysis of big data effectively."

The statistics listed below represent this significant and growing demand for data scientists.

#16	3,433	$105,395	#1

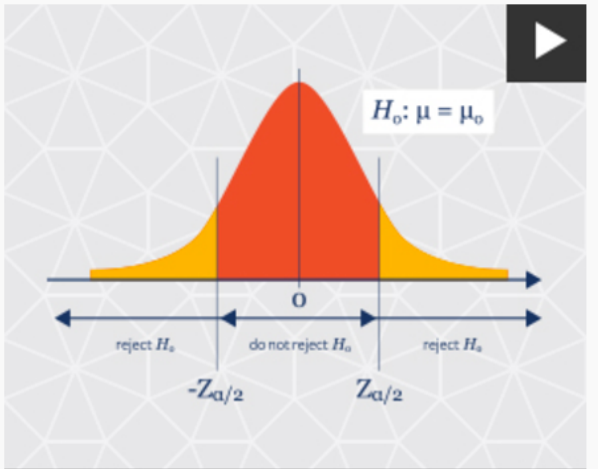


Python for Data Science

3 UNITS

Research Design and Application for Data and Analysis

3 UNITS



Statistics for Data Science

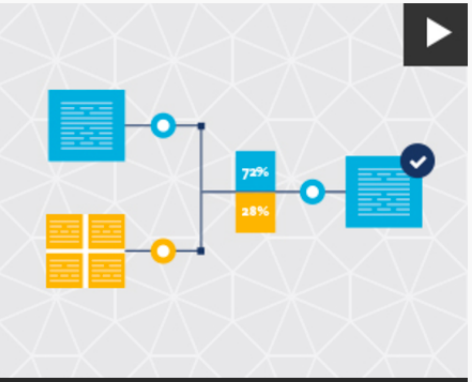
3 UNITS

Storing and Retrieving Data

3 UNITS

Applied Machine Learning

3 UNITS

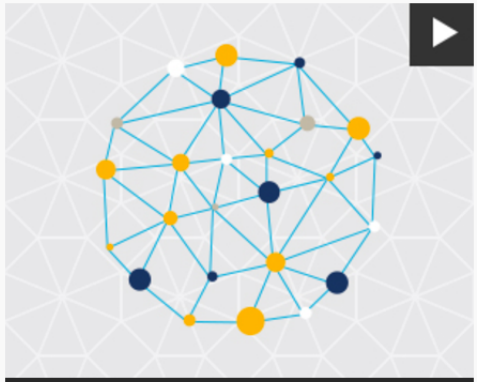


Experiments and Causality

3 UNITS

Behind the Data: Humans and Values

3 UNITS

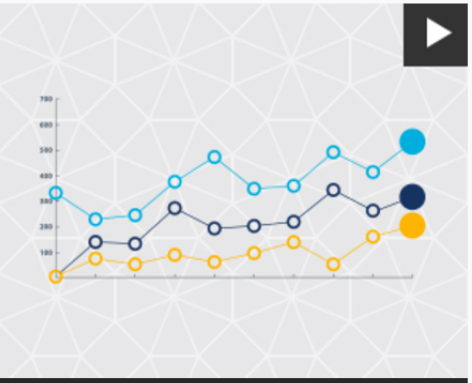


Scaling Up! Really Big Data

3 UNITS

Data Visualization

3 UNITS



Statistical Methods for Discrete Response, Time Series, and Panel Data

3 UNITS

Machine Learning at Scale

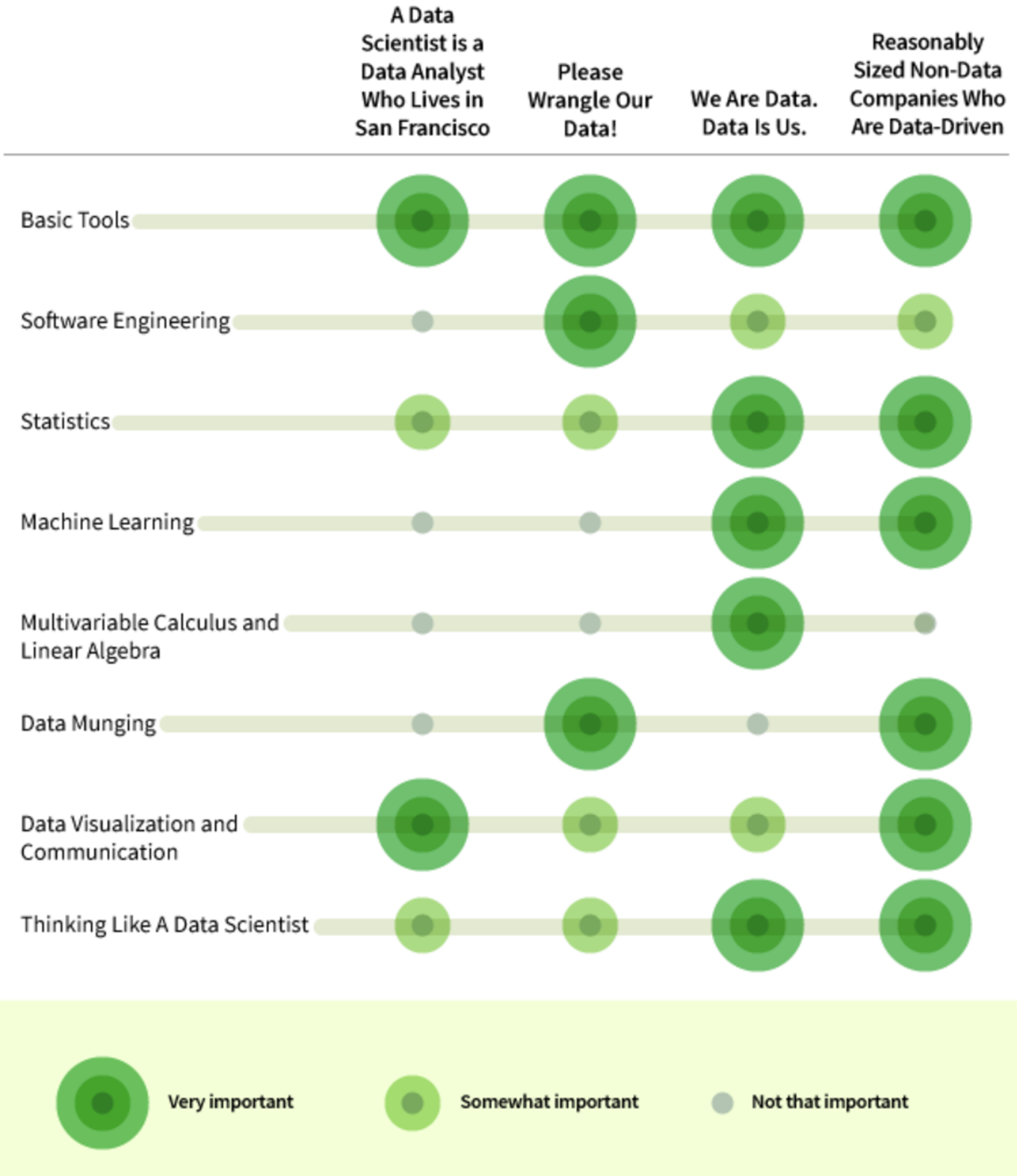
3 UNITS

Natural Language Processing with Deep Learning

3 UNITS

What is data science?

- Method: Check online courses
    - Udacity


A Data Scientist is a Data Analyst Who Lives in San Francisco
Please Wrangle Our Data!
We Are Data. Data Is Us.
Reasonably Sized Non-Data Companies Who Are Data-Driven
Basic Tools
Software Engineering
Statistics
Machine Learning
Multivariable Calculus and Linear Algebra
Data Munging
Data Visualization and Communication
Thinking Like A Data Scientist
Very important
Somewhat important
Not that important

What is data science?

- Method: Check online courses
    - Coursera (JHU)

List of modules

- The Data Scientist's Toolbox
- R Programming
- Getting and Cleaning Data
- Exploratory Data Analysis
- Reproducible Research
- Statistical Inference
- Regression Models
- Practical Machine Learning
- Developing Data Products
- Data Science Capstone

What is data science?

- Method: Ask a statistician
    - Larry Wasserman

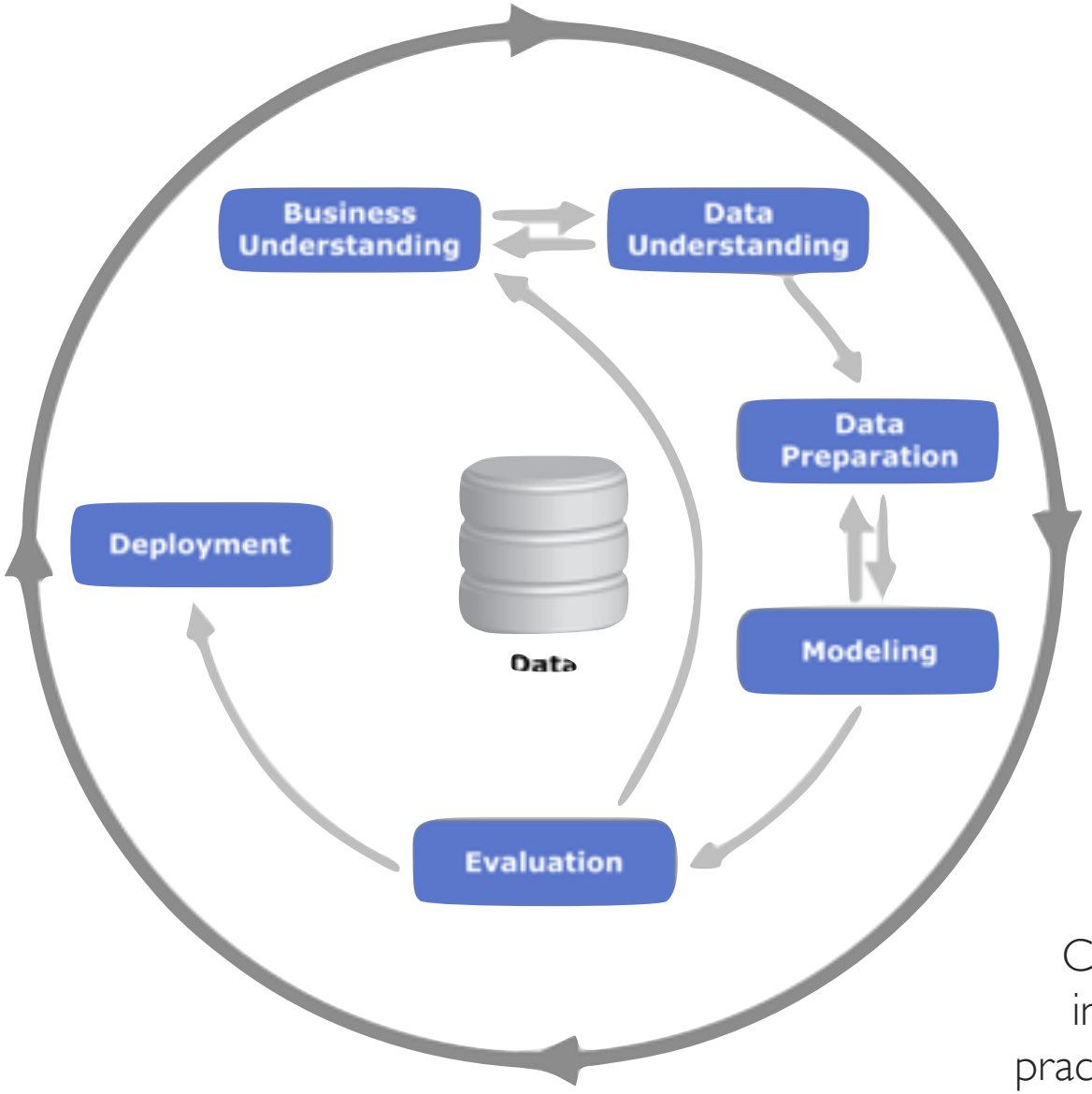
What is data science?

- Method: Check industry sources
- KDNuggets

What is data science?

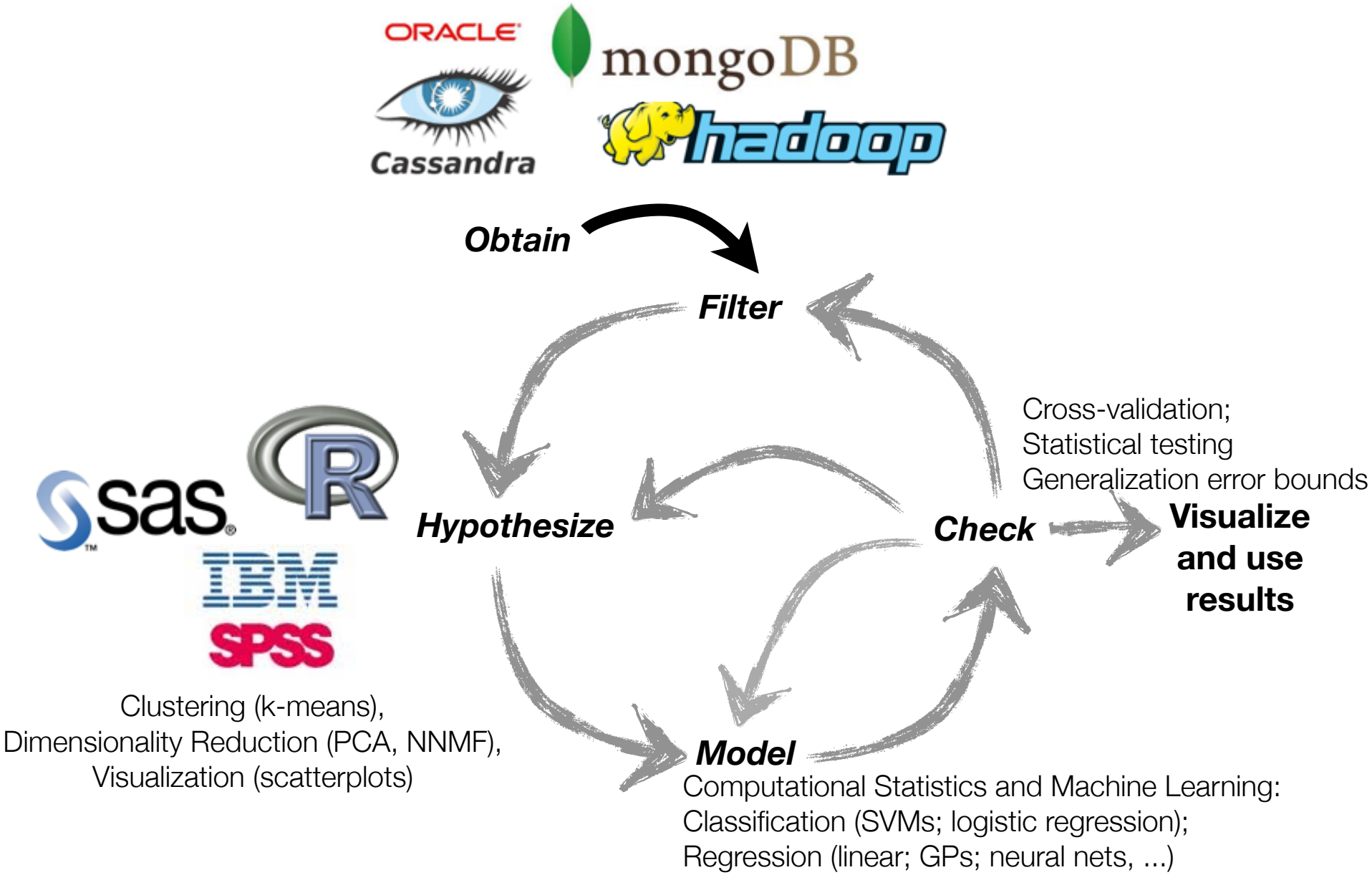
- Method: Google images!

Big Data: Acquisition, processing, and analysis cycle

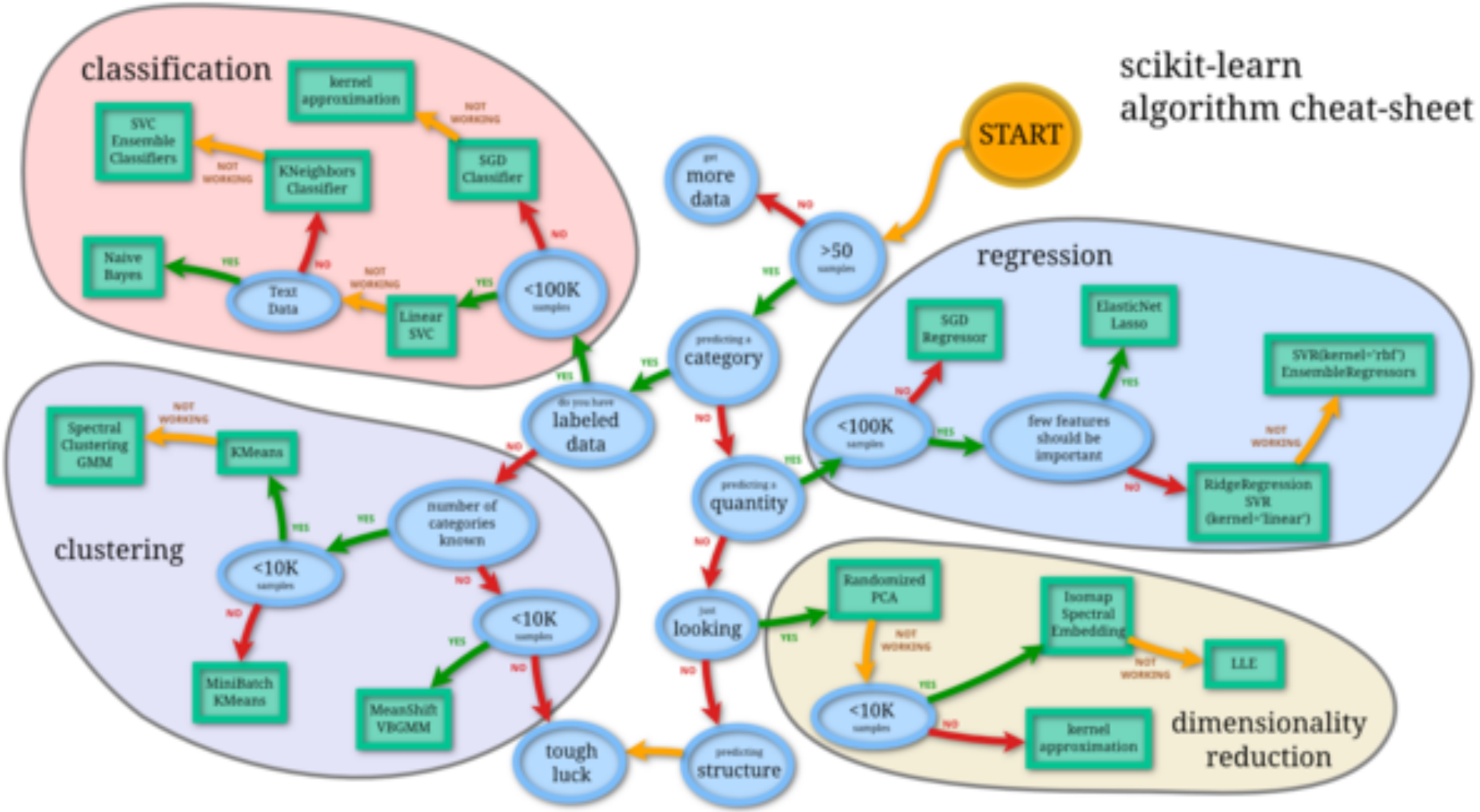


CRISP-DM: Cross-industry standard practice for data mining

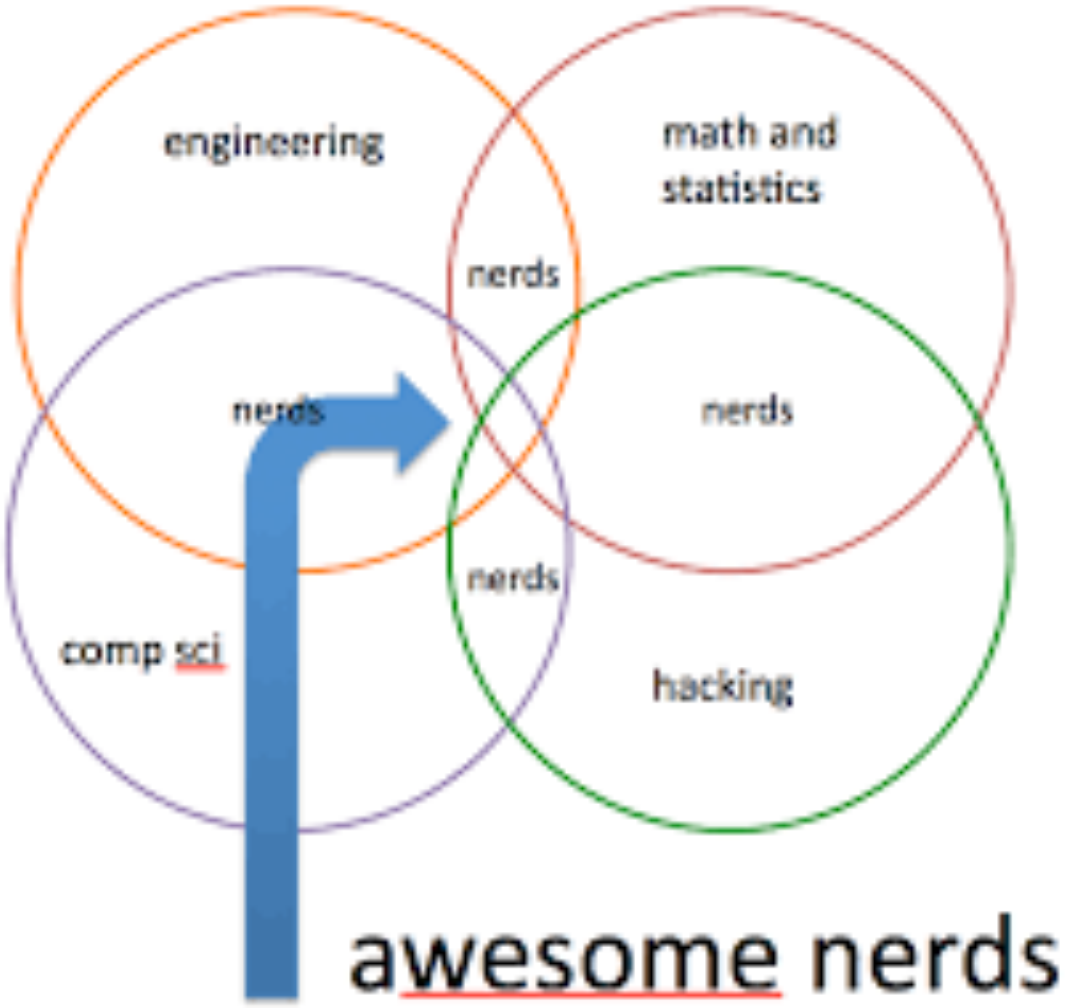
Big Data: Acquisition, processing, and analysis cycle



A "simple" graphic to help with modeling…



Data scientists?



Hilary Mason, of Fast forward labs
(formerly of bitly)

One last view on data science: Interview questions!

One last view on data science: Interview questions!

Q1. Explain what regularization is and why it is useful.

Q2. Which data scientists do you admire most? which startups?

Q3. How would you validate a model you created to generate a predictive model of a quantitative outcome variable using multiple regression.

Q4. Explain what precision and recall are. How do they relate to the ROC curve?

Q5. How can you prove that one improvement you've brought to an algorithm is really an improvement over not doing anything?

One last view on data science: Interview questions!

8. What is statistical power?

9. Explain what resampling methods are and why they are useful. Also explain their limitations.

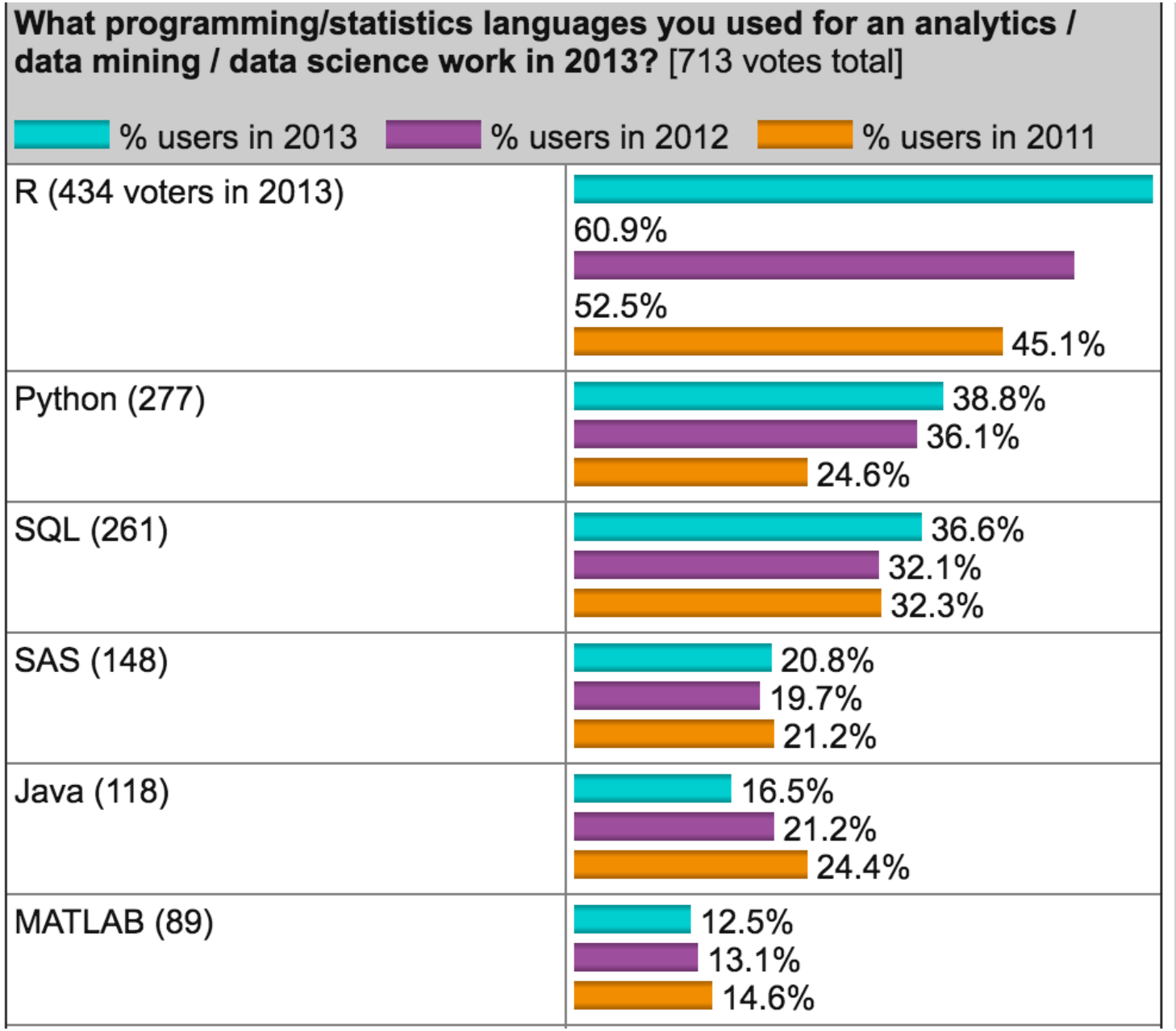
10. Is it better to have too many false positives, or too many false negatives? Explain.

11. What is selection bias, why is it important and how can you avoid it?

Programming for data science

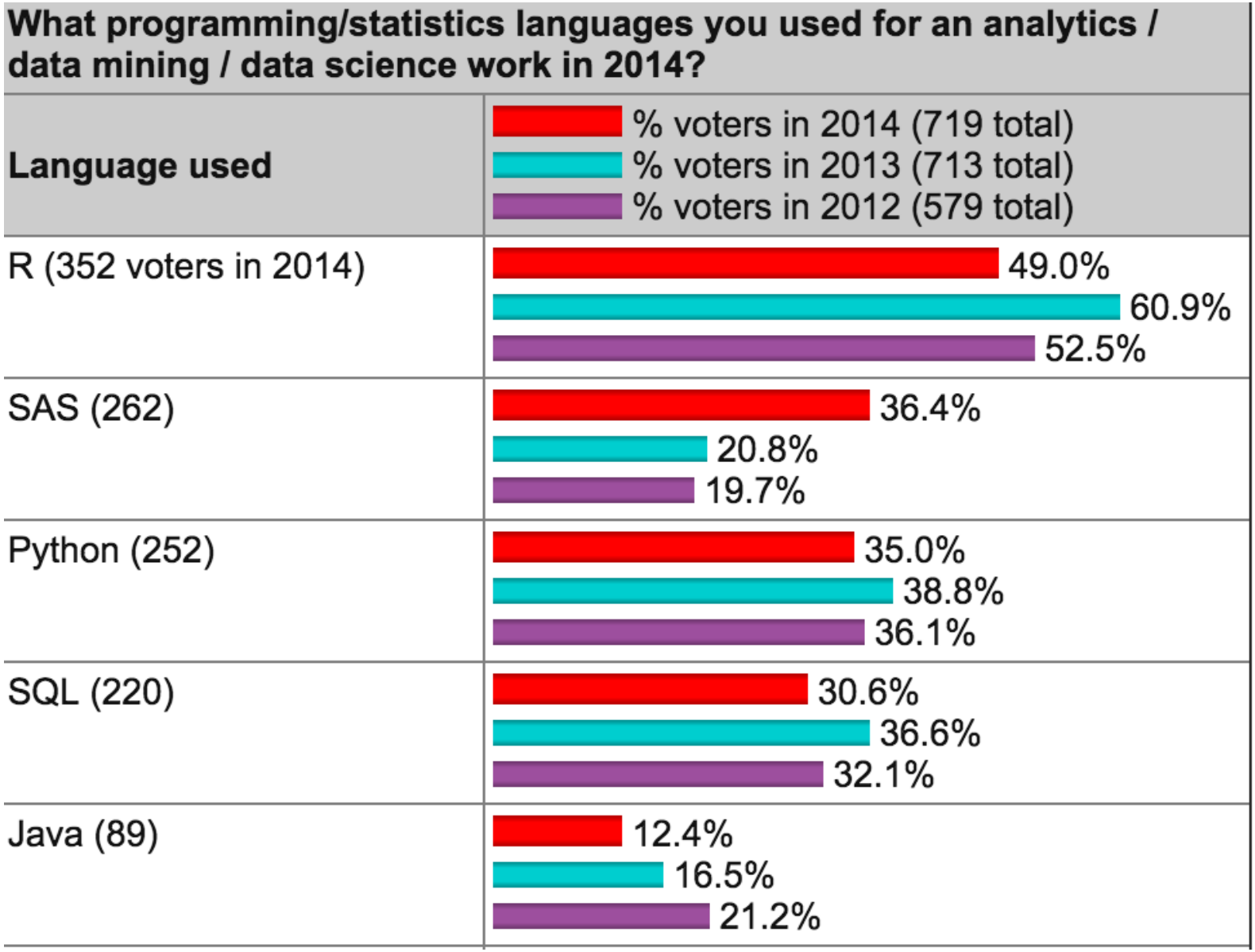
What languages and packages are being used?

How is this changing over time?

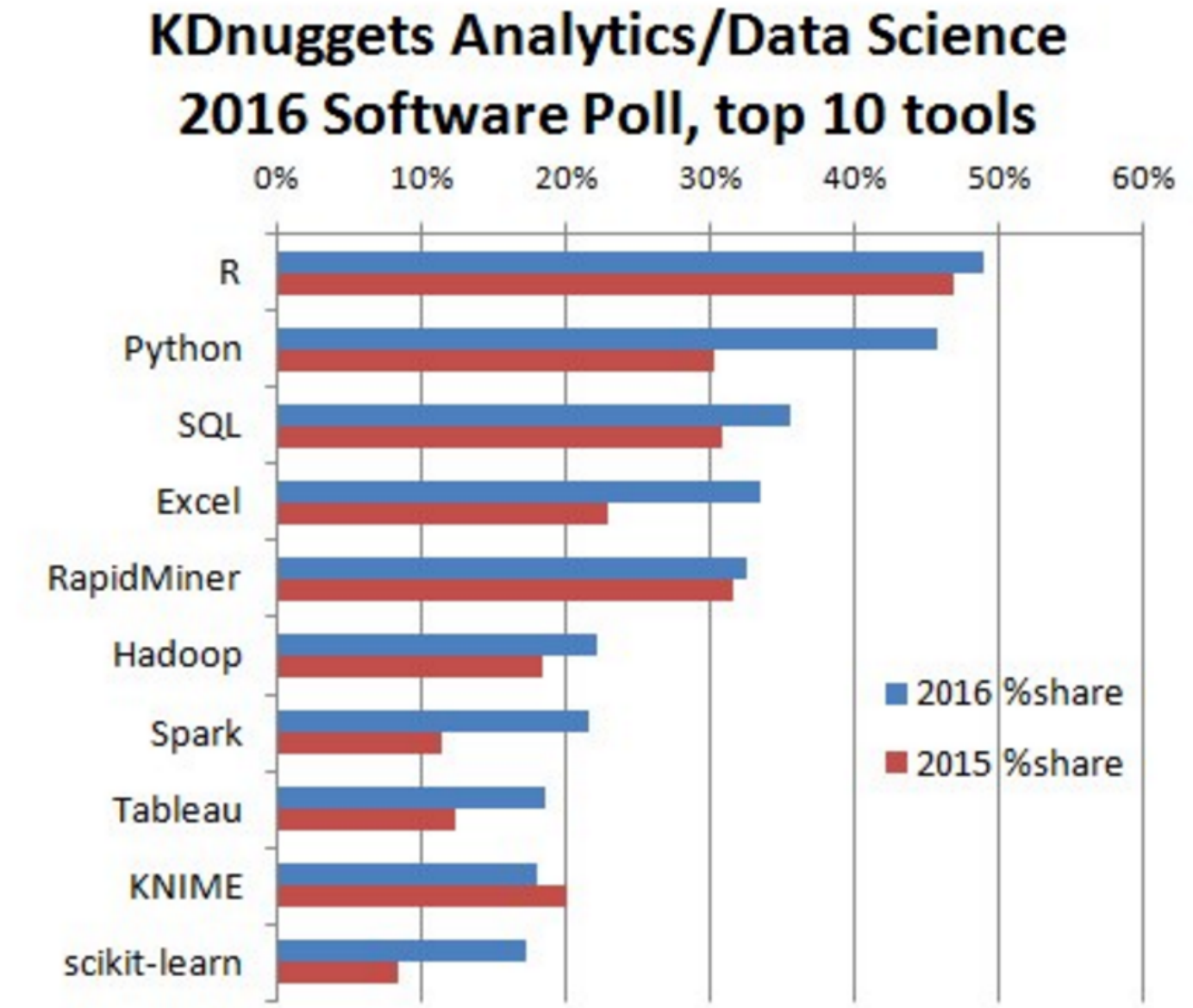


http://www.kdnuggets.com/2013/08/languages-for-analytics-data-mining-data-science.html

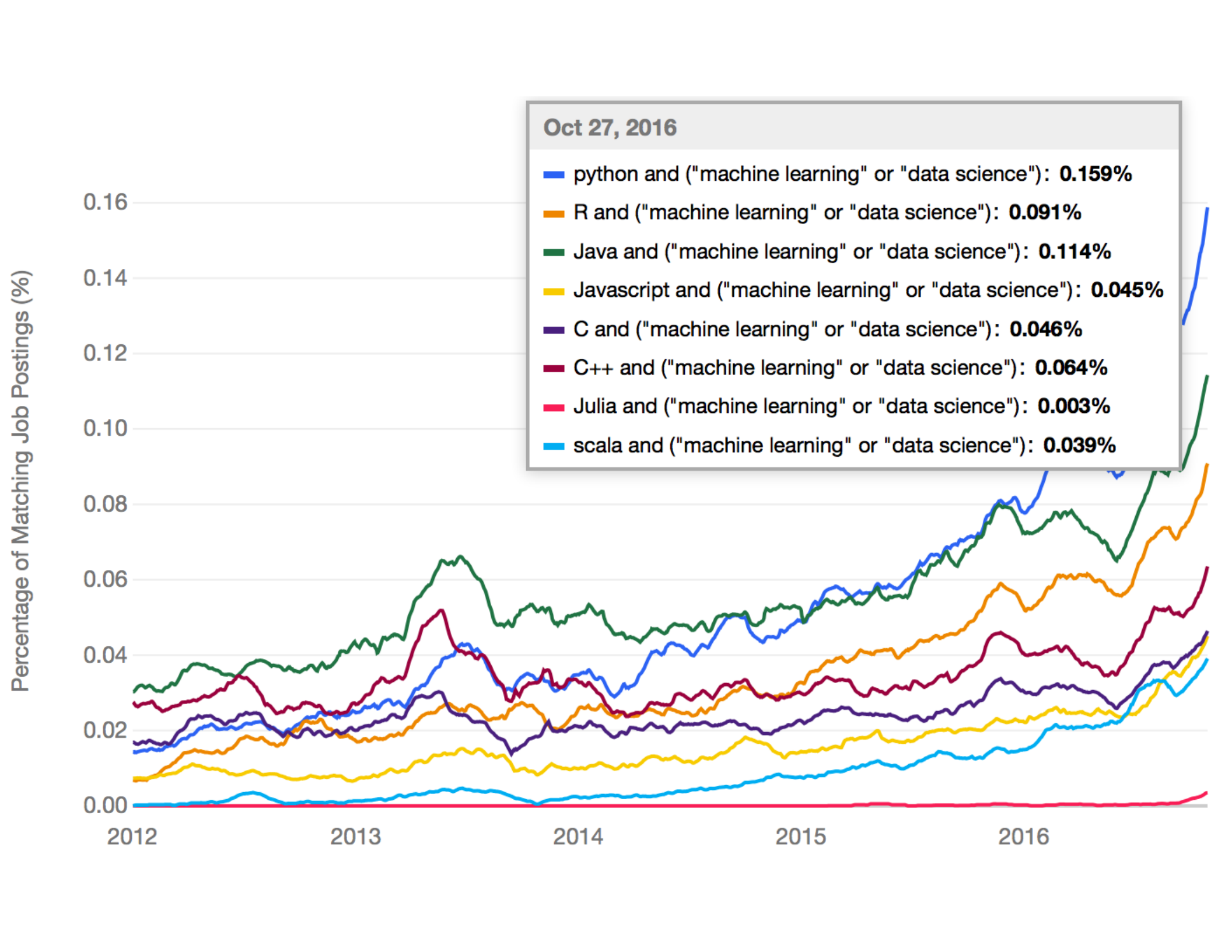
Aug 27, 2013

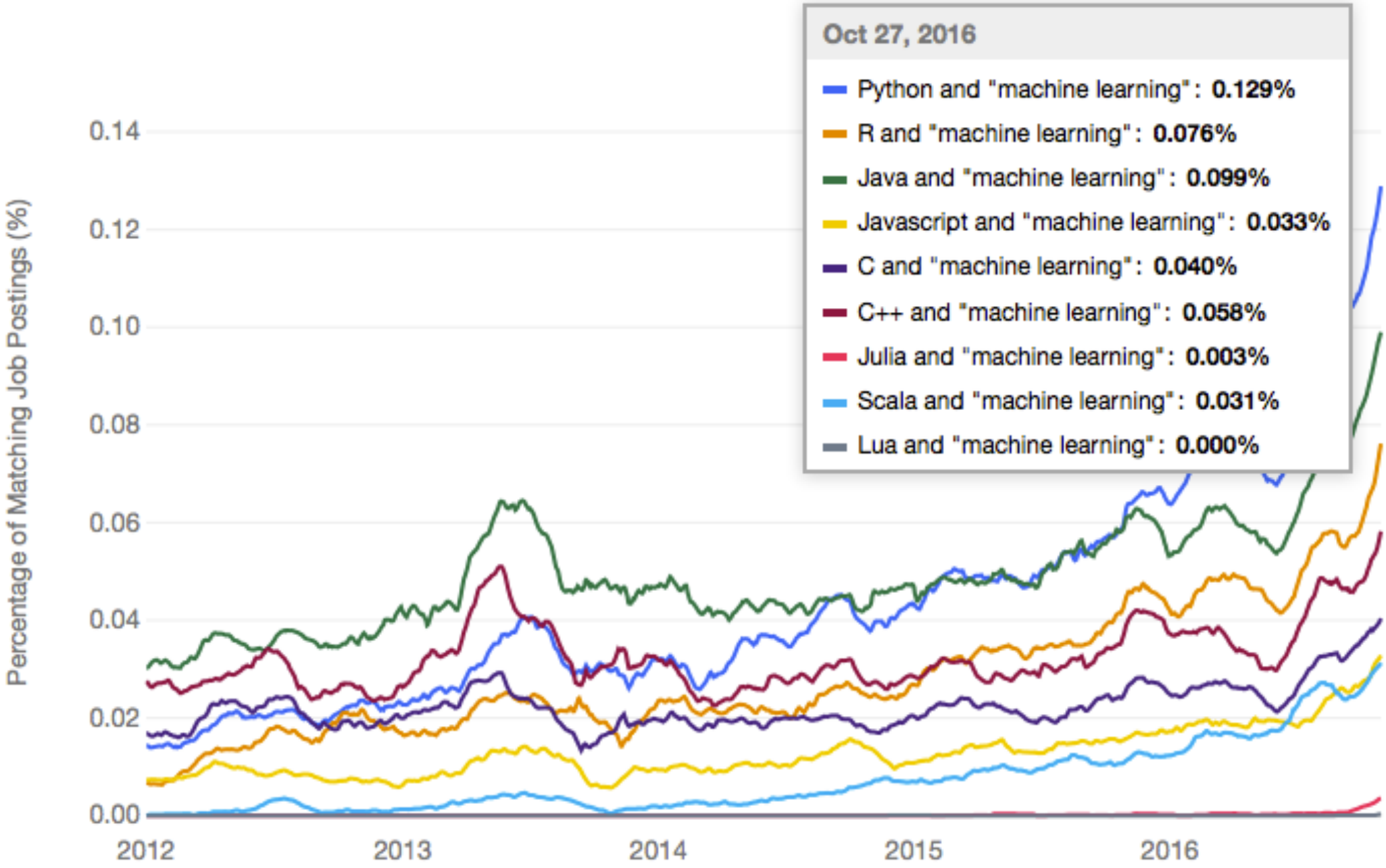


http://www.kdnuggets.com/2014/08/four-main-languages-analytics-data-mining-data-science.html
Aug, 2014



http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html
Jun, 2016


Oct 27, 2016
python and ("machine learning" or "data science"): 0.159%
R and ("machine learning" or "data science"): 0.091%
Java and ("machine learning" or "data science"): 0.114%
Javascript and ("machine learning" or "data science"): 0.045%
C and ("machine learning" or "data science"): 0.046%
C++ and ("machine learning" or "data science"): 0.064%
Julia and ("machine learning" or "data science"): 0.003%
scala and ("machine learning" or "data science"): 0.039%
Percentage of Matching Job Postings (%)
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
2012
2013
2014
2015
2016


Oct 27, 2016
Python and "machine learning": 0.129%
R and "machine learning": 0.076%
Java and "machine learning": 0.099%
Javascript and "machine learning": 0.033%
C and "machine learning": 0.040%
C++ and "machine learning": 0.058%
Julia and "machine learning": 0.003%
Scala and "machine learning": 0.031%
Lua and "machine learning": 0.000%
Percentage of Matching Job Postings (%)
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
2012
2013
2014
2015
2016

Programming Languages

Python, Java, Unix tools, Scala grew in popularity, while C/C++, Perl, Julia, F#, Clojure, and Lisp declined.

Here are the programming languages sorted by popularity.

- Python, 45.8% share (was 30.3%), 51% increase
- Java, 16.8% share (was 14.1%), 19% increase
- Unix shell/awk/gawk 10.4% share (was 8.0%), 30% increase
- C/C++, 7.3% share (was 9.4%), 23% decrease
- Other programming/data languages, 6.8% share (was 5.1%), 34.1% increase
- Scala, 6.2% share (was 3.5%), 79% increase
- Perl, 2.3% share (was 2.9%), 19% decrease
- Julia, 1.1% share (was 1.1%), 1.6% decrease
- F#, 0.4% share (was 0.7%), 41.8% decrease
- Clojure, 0.4% share (was 0.5%), 19.4% decrease
- Lisp, 0.2% share (was 0.4%), 33.3% decrease

Hadoop/Big Data Tools

The usage of Hadoop/Big Data tools grew to 39%, up from 29% in 2015 and 17% in 2014), driven mainly by big growth in Apache Spark, MLlib (Spark Machine Learning Library) and H2O, which we included among Big Data tools.

Here are the Big Data tools and their share in 2016, 2015, and %change.

Tool	2016 %Share	2015 %share	% change
Hadoop	22.1%	18.4%	+20.5%
Spark	21.6%	11.3%	+91%
Hive	12.4%	10.2%	+21.3%
MLlib	11.6%	3.3%	+253%
SQL on Hadoop tools	7.3%	7.2%	+1.6%
H2O	6.7%	2.0%	+234%
HBase	5.5%	4.6%	+18.6%
Apache Pig	4.6%	5.4%	-16.1%
Apache Mahout	2.6%	2.8%	-7.2%
Dato	2.4%	0.5%	+338%
Datameer	0.4%	0.9%	-52.3%
Other Hadoop/HDFS-based tools	4.9%	4.5%	+7.5%

http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html
Jun, 2016

Deep Learning Tools

For the second year KDnuggets poll include Deep Learning Tools. This year, 18% of voters used Deep Learning tools, doubling the 9% in 2015.

Google Tensorflow jumped to first place, displacing last year leader Theano/Pylearn2 ecosystem.

Top tools:

- Tensorflow, 6.8%
- Theano ecosystem (including Pylearn2), 5.1%
- Caffe, 2.3%
- MATLAB Deep Learning Toolbox, 2.0%
- Deeplearning4j, 1.7%
- Torch, 1.0%
- Microsoft CNTK, 0.9%
- Cuda-convnet, 0.8%
- mxnet, 0.6%
- Convnet.js, 0.3%
- darch, 0.1%
- Nervana, 0.1%
- Veles, 0.1%
- Other Deep Learning Tools, 3.7%

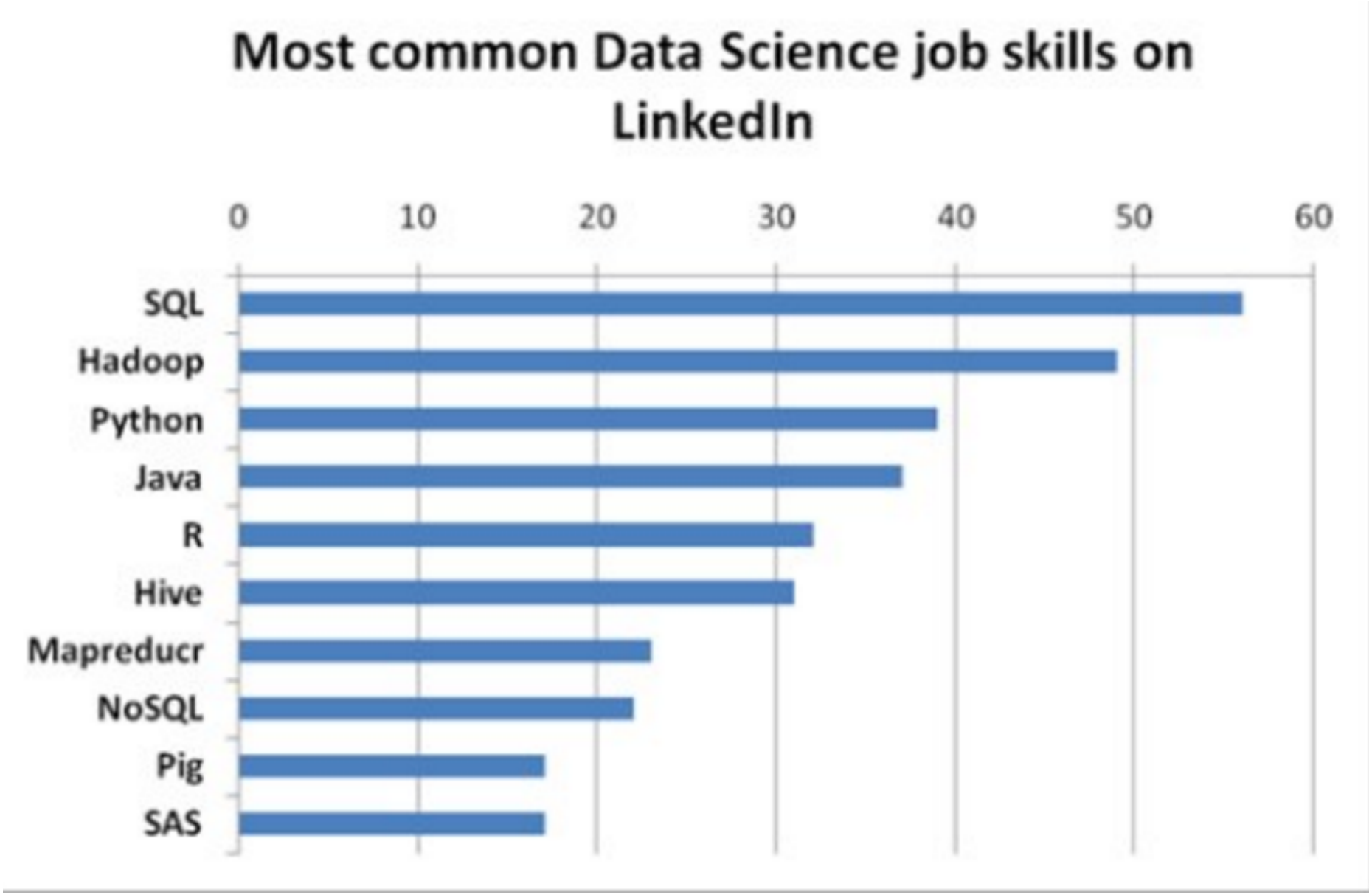
http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html

Jun, 2016

Next table has the top 10 most popular tools in 2016 poll

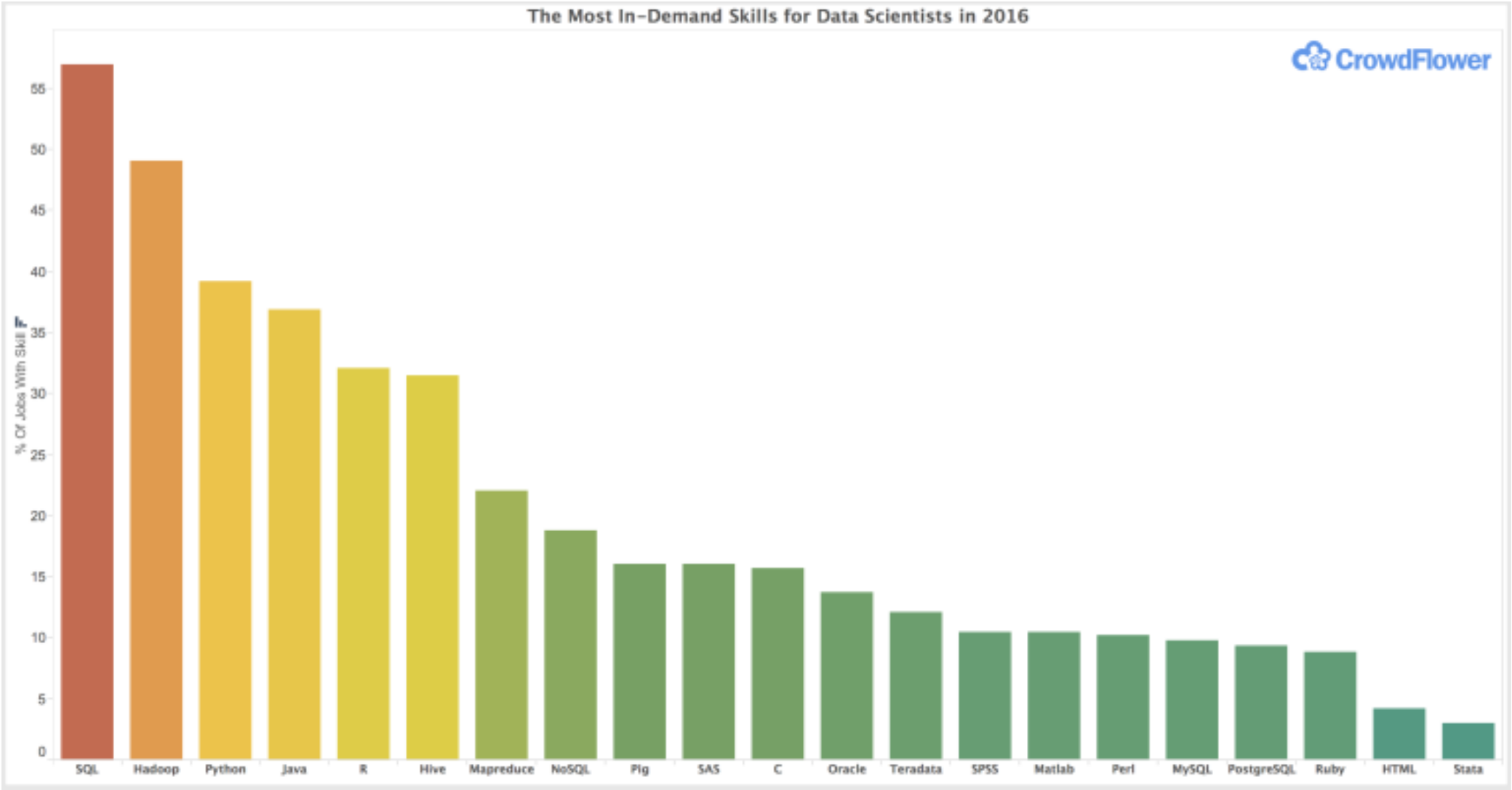
Tool	2016 % share	% change	% alone
R	49%	+4.5%	1.4%
Python	45.8%	+51%	0.1%
SQL	35.5%	+15%	0%
Excel	33.6%	+47%	0.2%
RapidMiner	32.6%	+3.5%	11.7%
Hadoop	22.1%	+20%	0%
Spark	21.6%	+91%	0.2%
Tableau	18.5%	+49%	0.2%
KNIME	18.0%	-10%	4.4%
scikit-learn	17.2%	+107%	0%

http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html
Jun, 2016



http://www.kdnuggets.com/2016/02/data-science-skills-2016.html

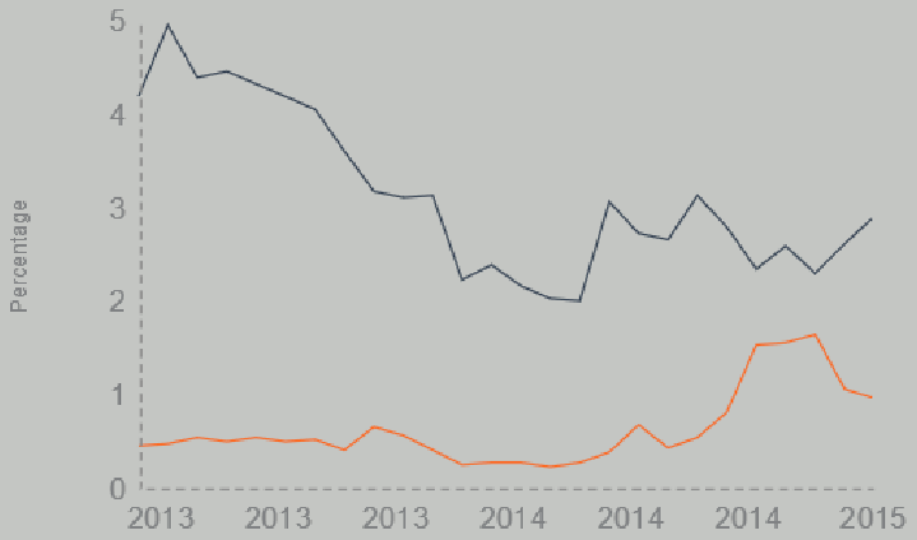
Feb, 2016


The Most In-Demand Skills for Data Scientists in 2016
CrowdFlower
% Of Jobs With Skill
0
5
10
15
20
25
30
35
40
45
50
55
SQL
Hadoop
Python
Java
R
Hive
Mapreduce
NoSQL
Pig
SAS
C
Oracle
Teradata
SPSS
Matlab
Perl
MySQL
PostgreSQL
Ruby
HTML
Stata

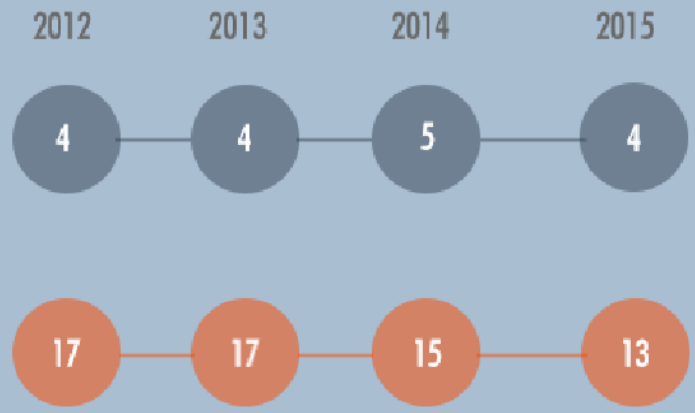

R
VS.
python

Popularity Rankings

R and Pythons popularity between 2013 and February 2015 (Tiobe Index)



Redmonk ranking, comparing the relative performance of programming languages on GitHub and Stack Overflow (September 2012 and January 2013, 2014, 2015)



Jobs And Salary?

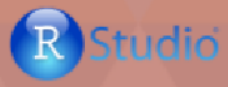
2014 Dice Tech Salary Survey:
Average Salary For High Paying Skills and Experience

$ 115,531

$94,139

Getting Started

IDE



Popular Packages

- ✓ dplyr, plyr and data.table to easily manipulate data.
- ✓ stringr to manipulate strings.
- ✓ zoo to work with regular and irregular time series
- ✓ ggvis, lattice and ggplot2 to visualize data.
- ✓ caret for machine learning.

Tip: check out DataCamp's online interactive courses and tutorials!

IDE

There are many Python IDEs to chose from. However, Spyder and IPython Notebook are most popular.

Tip: also look up Rodeo, the "data science IDE for Python"

Popular Libraries

- ✓ pandas to easily manipulate data.
- ✓ SciPy /NumPy for scientific computing.
- ✓ sckikit-learn to use machine learning methods.
- ✓ matplotlib to make graphics.
- ✓ statsmodels to explore data, estimate statistical models, and perform statistical tests and unit tests.

HW 2

- Write a tutorial on dicts (yes, again), functions, and classes
- Cover basic functionality with worked examples
- Due Sunday by 11:59 pm
- Evaluations due Tuesday by 11:59