Introduction to data science

Patrick Shafto

Department of Math and Computer Science

Plan for today

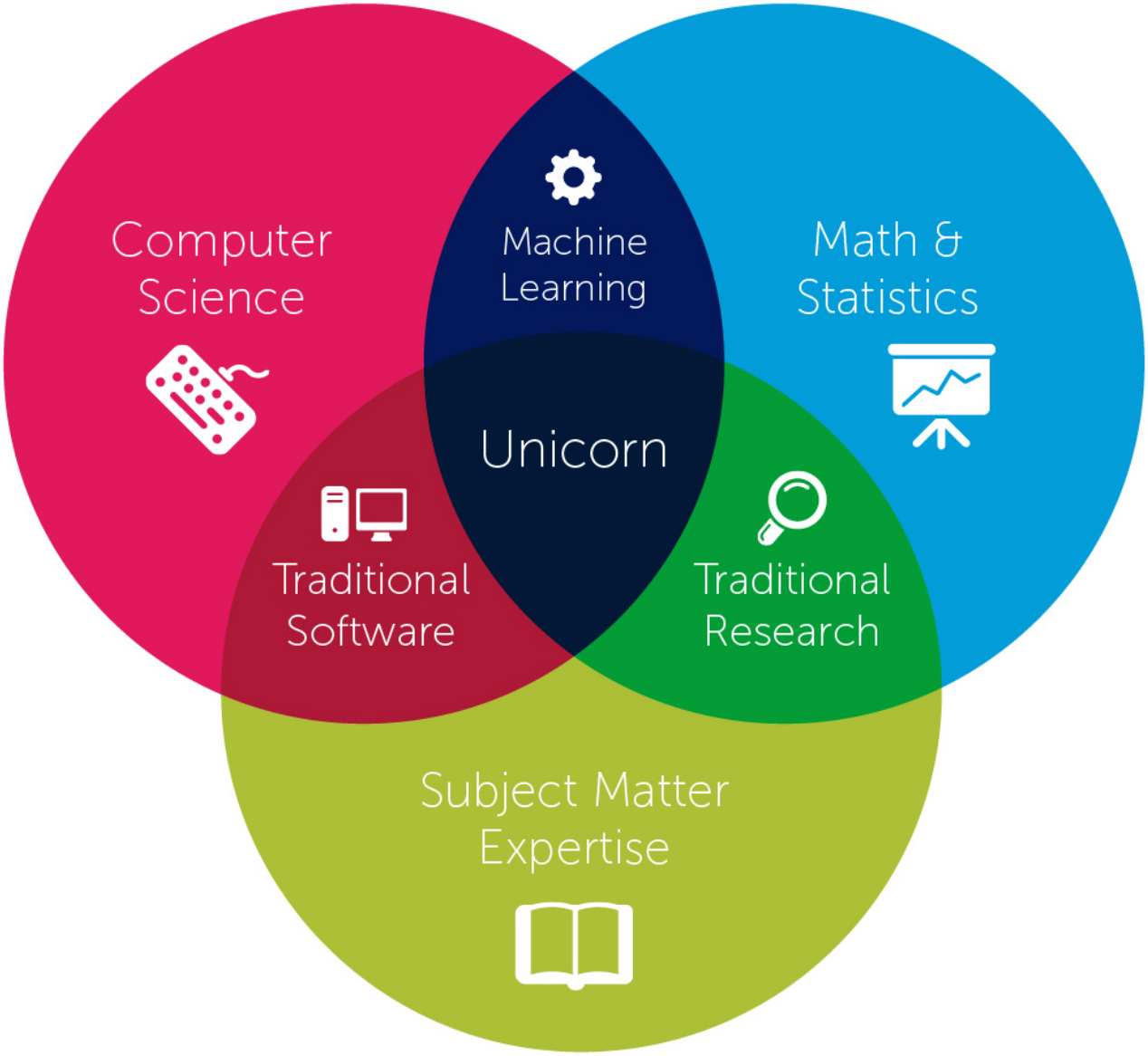
- Programming for data science
- Superfast intro to python

HW 2

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

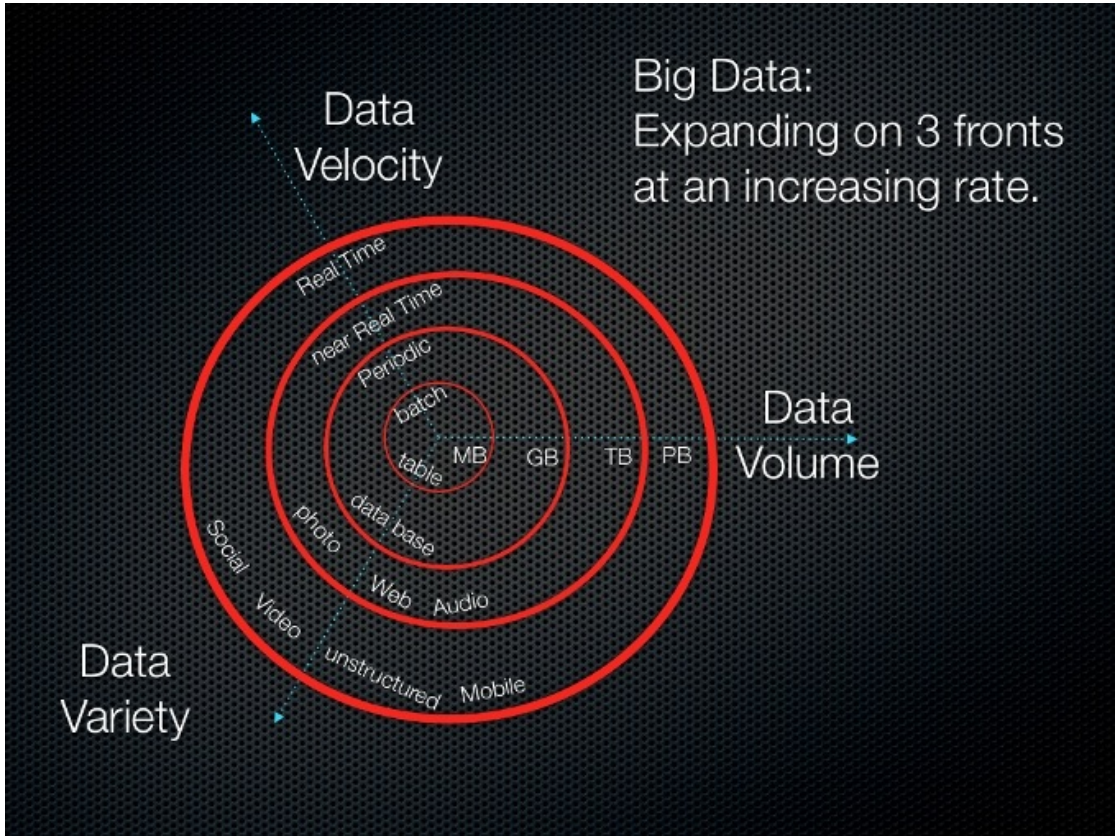
- Evaluations due Tuesday by 11:59pm
    - Use the rubric!
    - Justify your scores on each section of the rubric!

Data Science



Copyright © 2014 by Steven Geringer Raleigh, NC.
Permission is granted to use, distribute, or modify this image,
provided that this copyright notice remains intact.

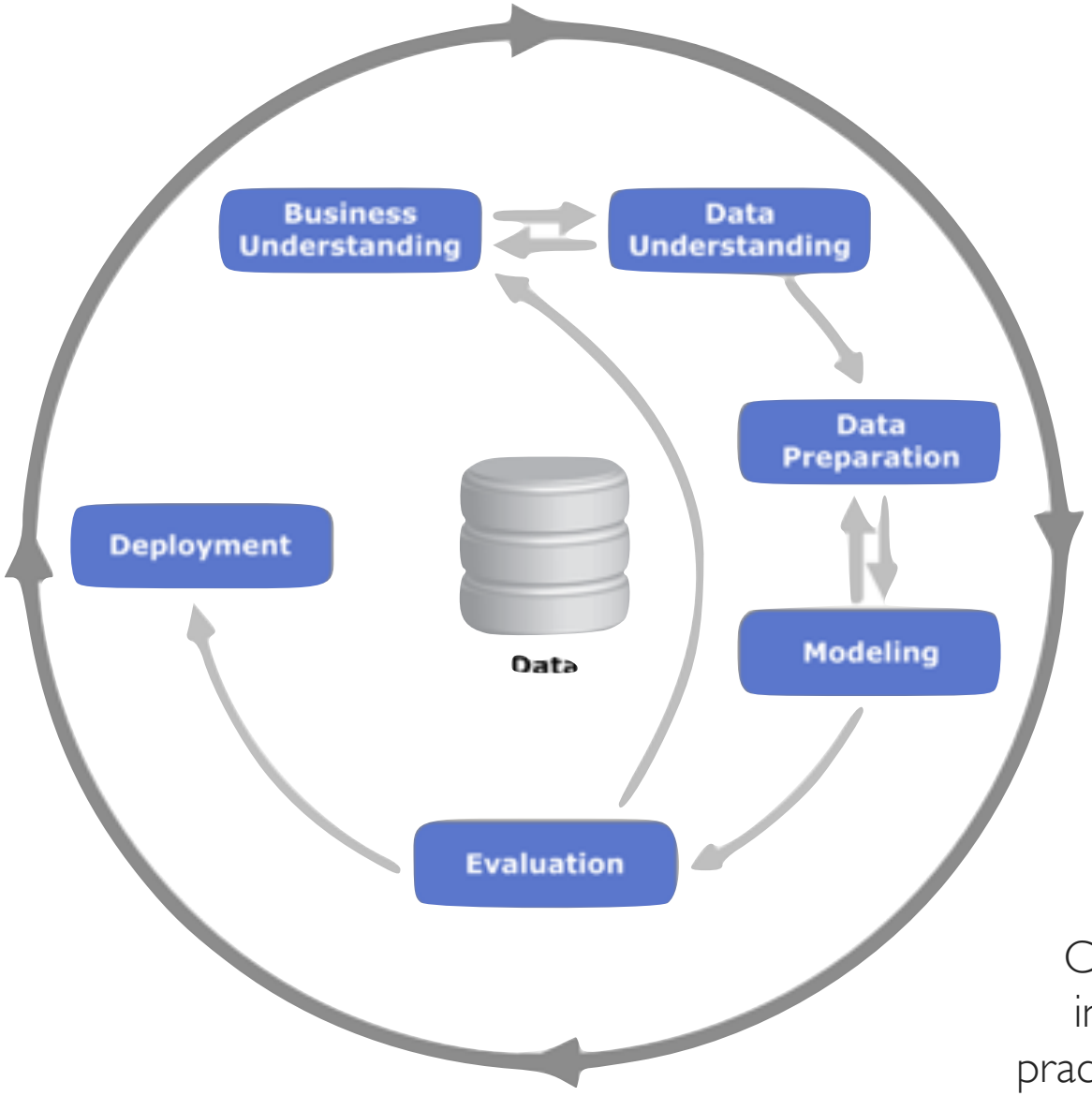
"Big data": Promise and problems



Promise: Personalization of prediction, service, education, healthcare, …

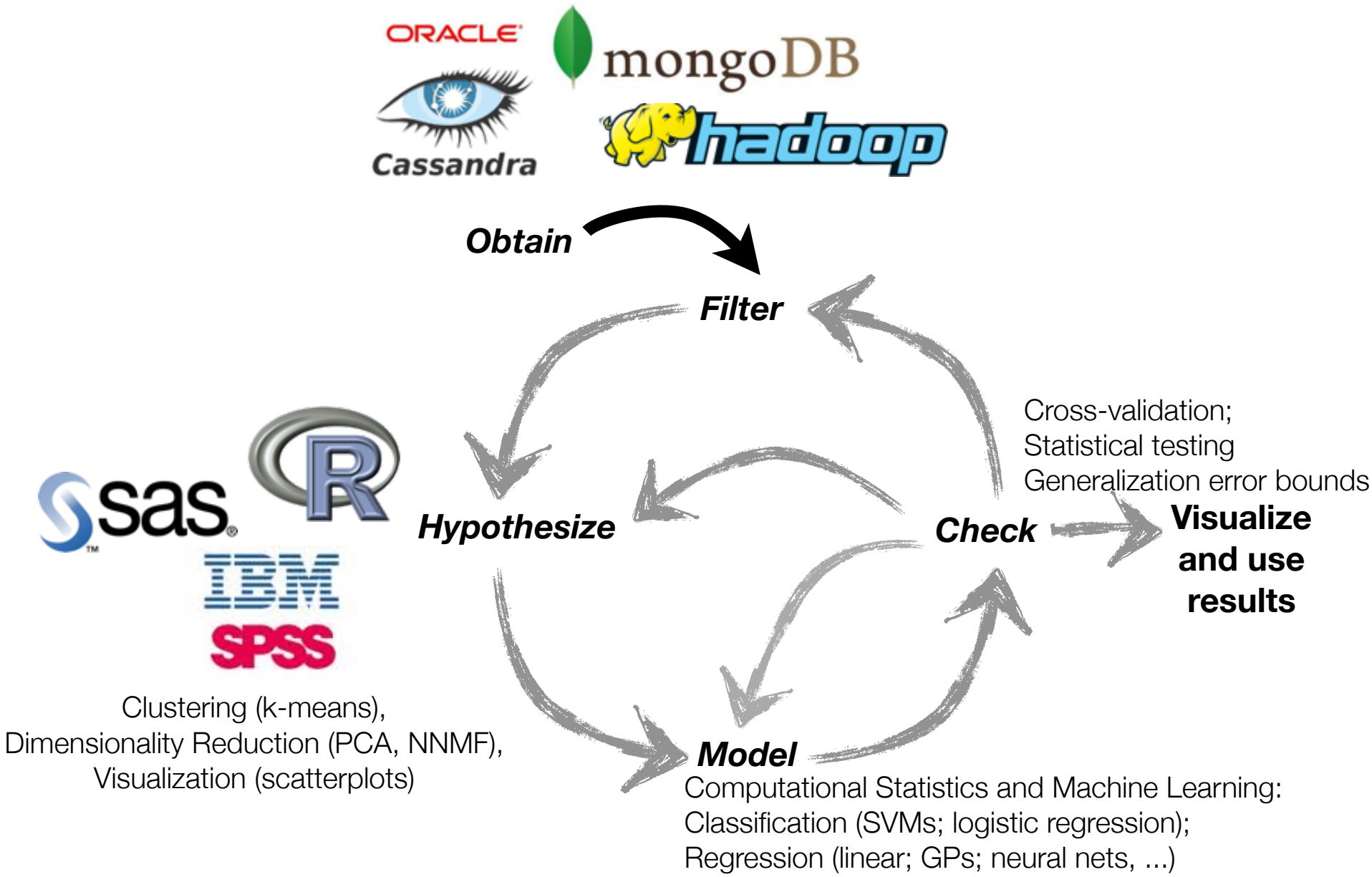
Problems: High dimensional, heterogeneously typed data; missing values; noisy variables

Big Data: Acquisition, processing, and analysis cycle

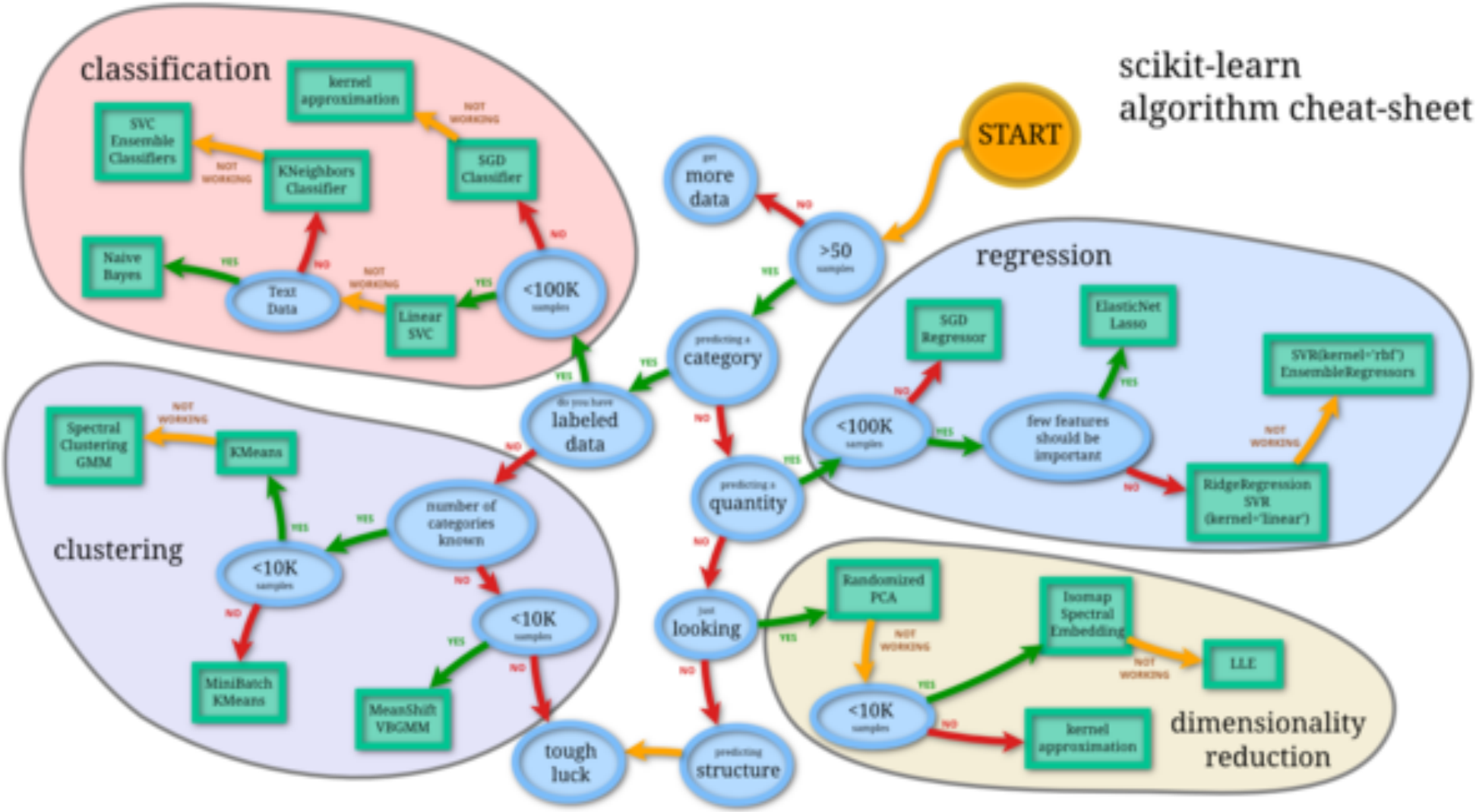


CRISP-DM: Cross-industry standard practice for data mining

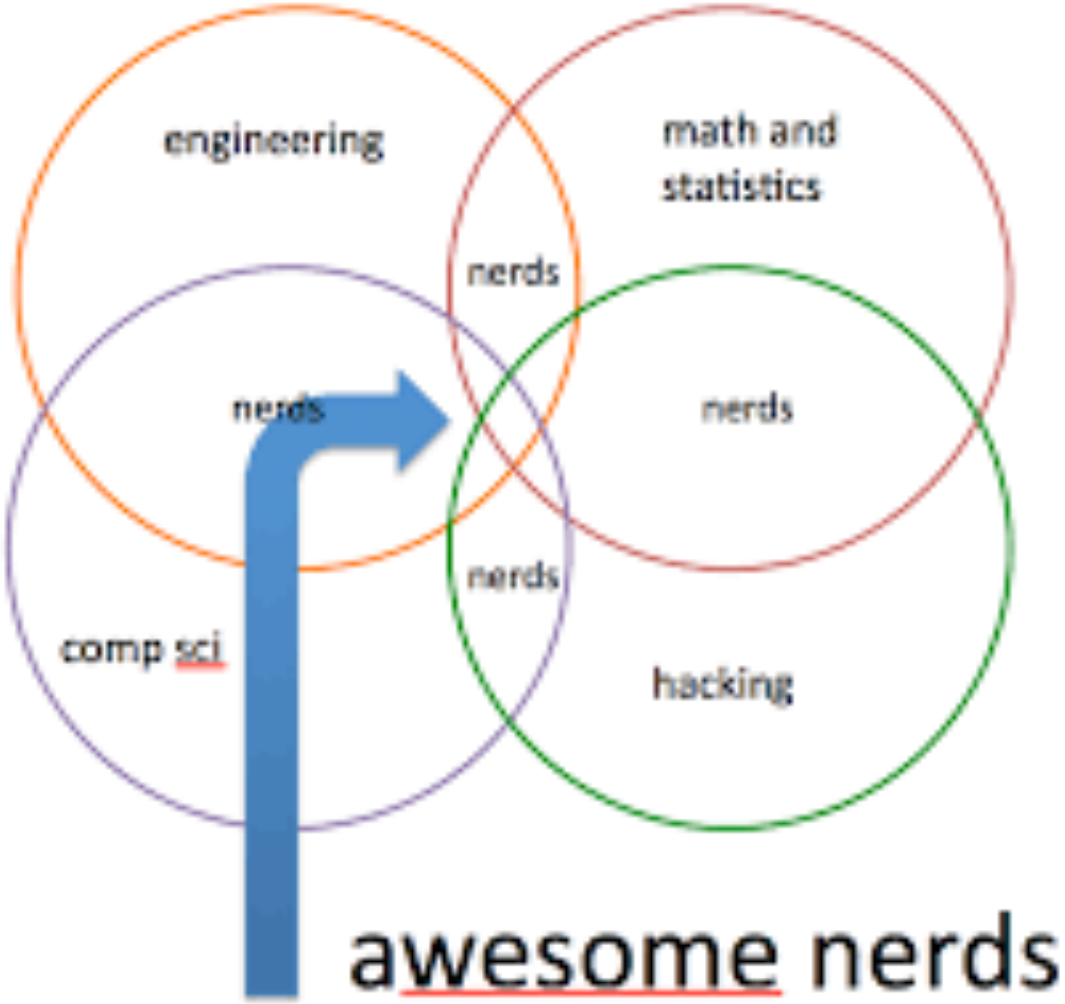
Big Data: Acquisition, processing, and analysis cycle



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.

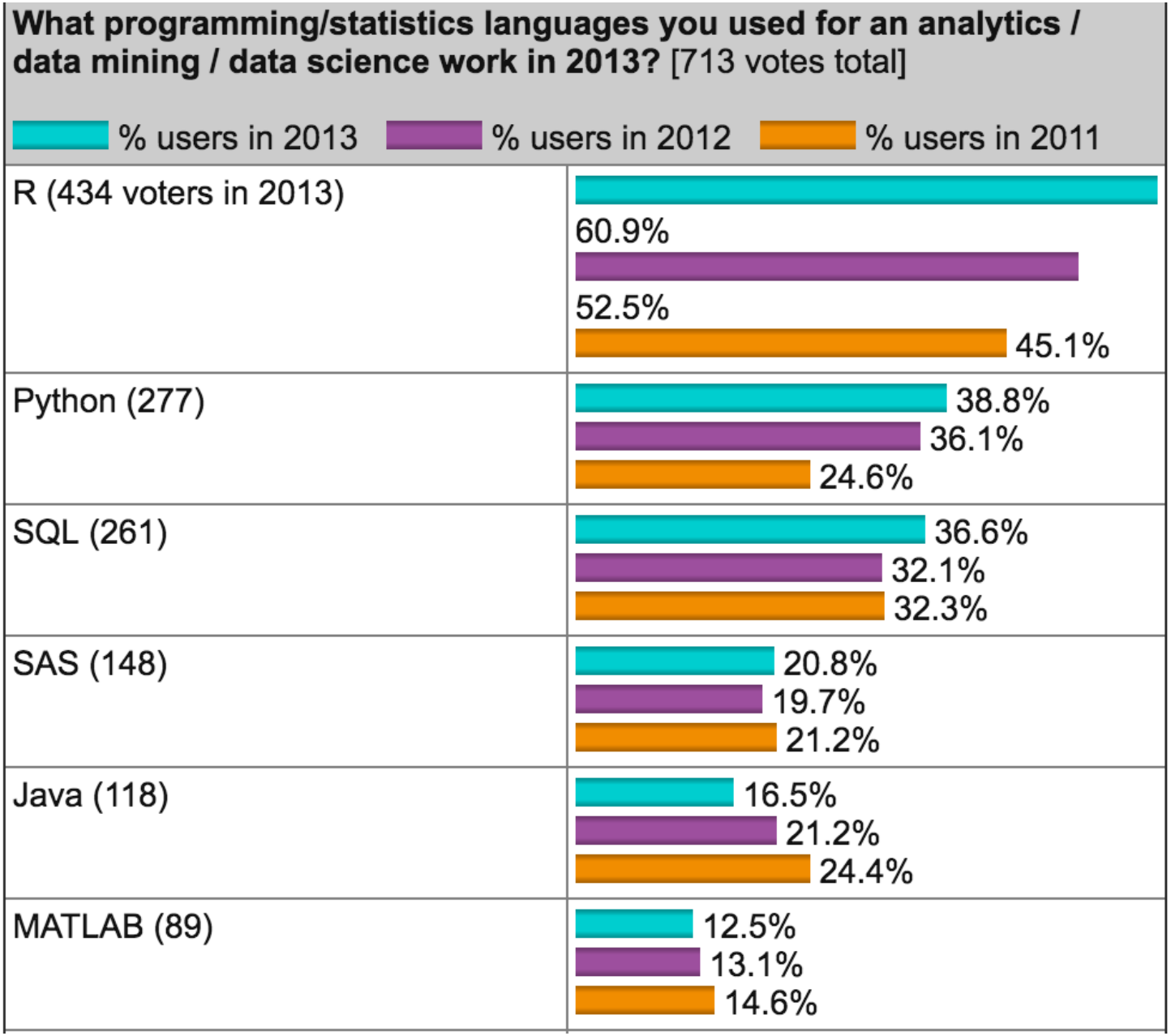
10. Is it better to have too many false positives, or too many false negatives? Explain.

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

Programming for data science

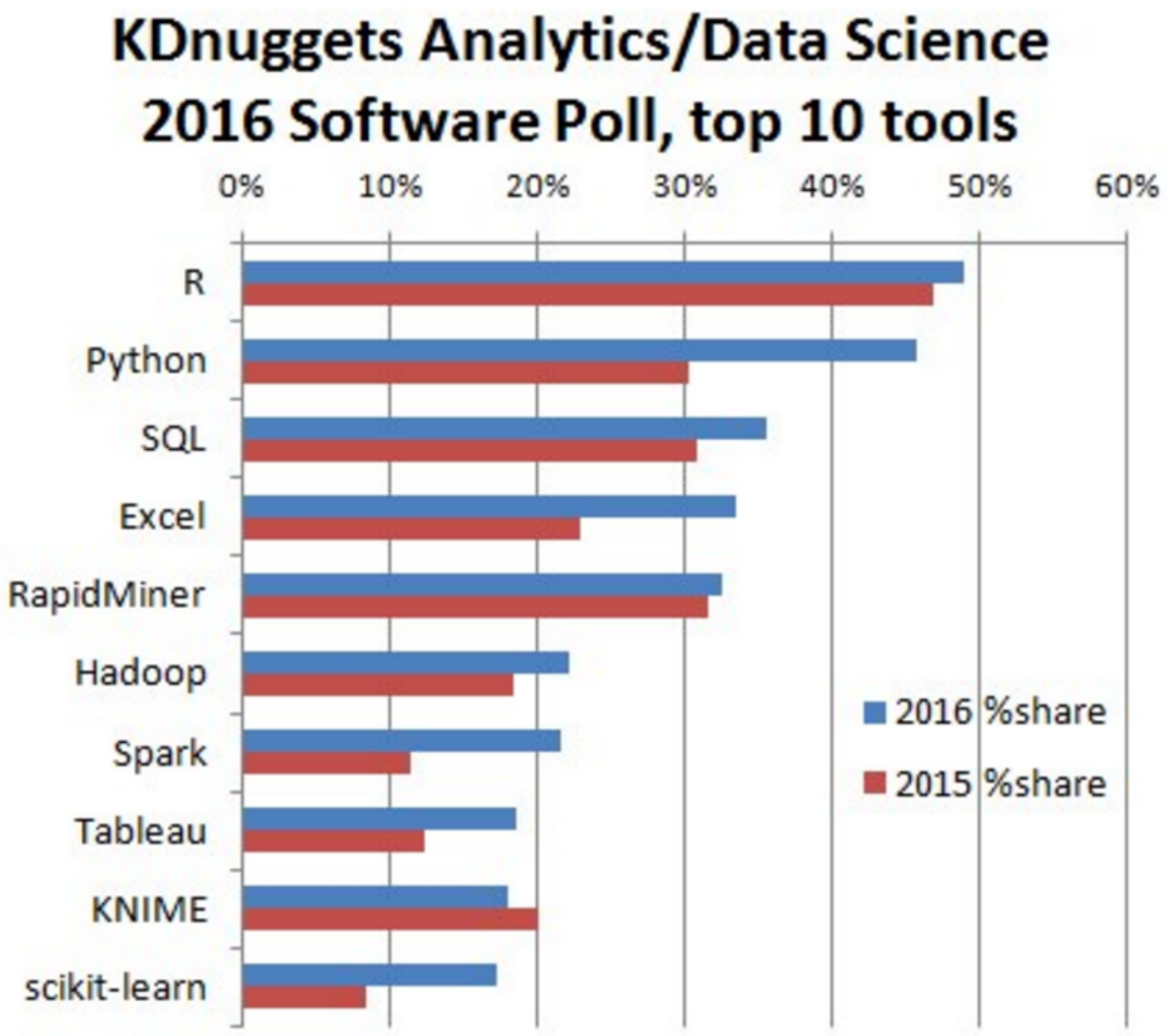
What languages and packages are being used?

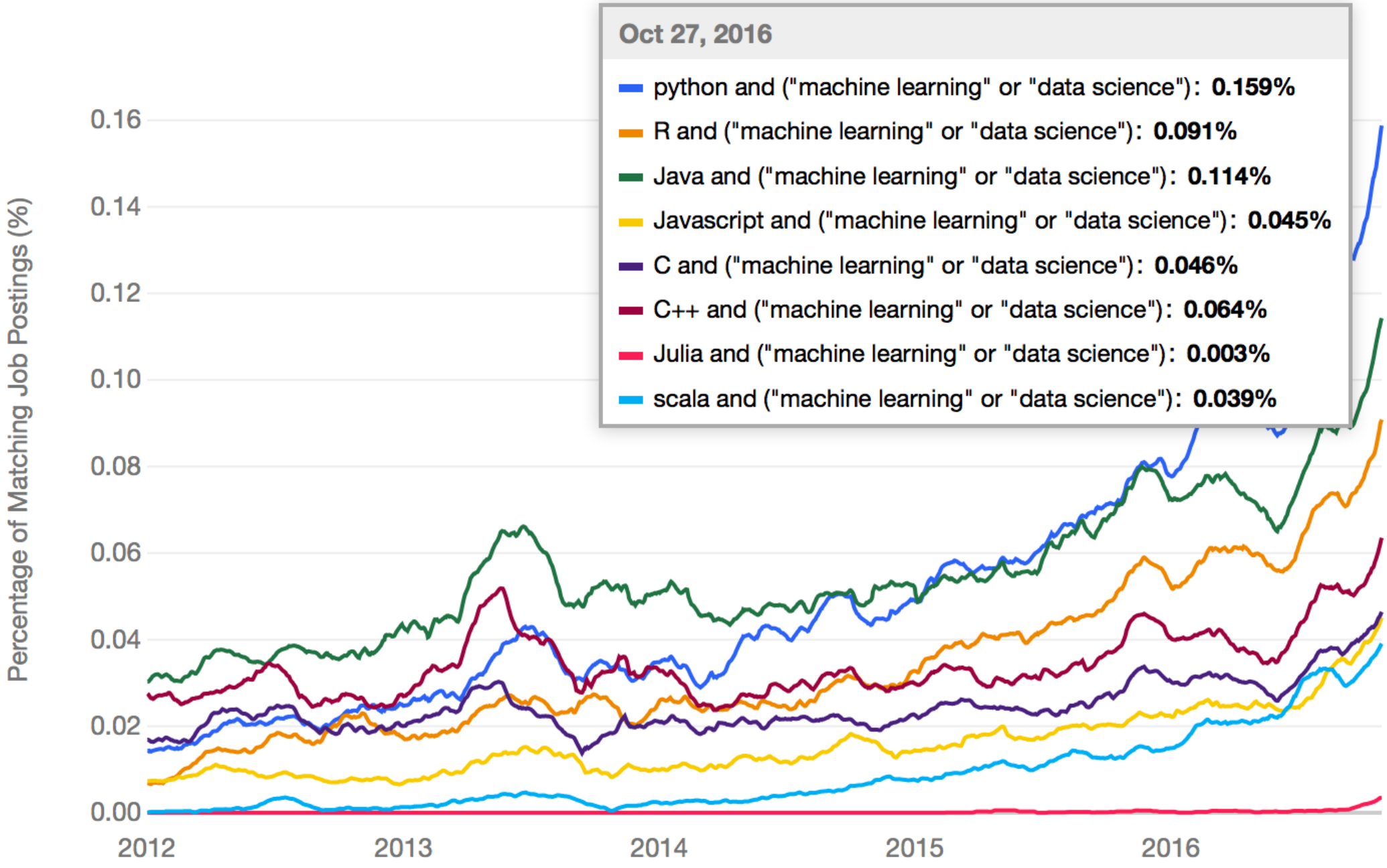
How is this changing over time?

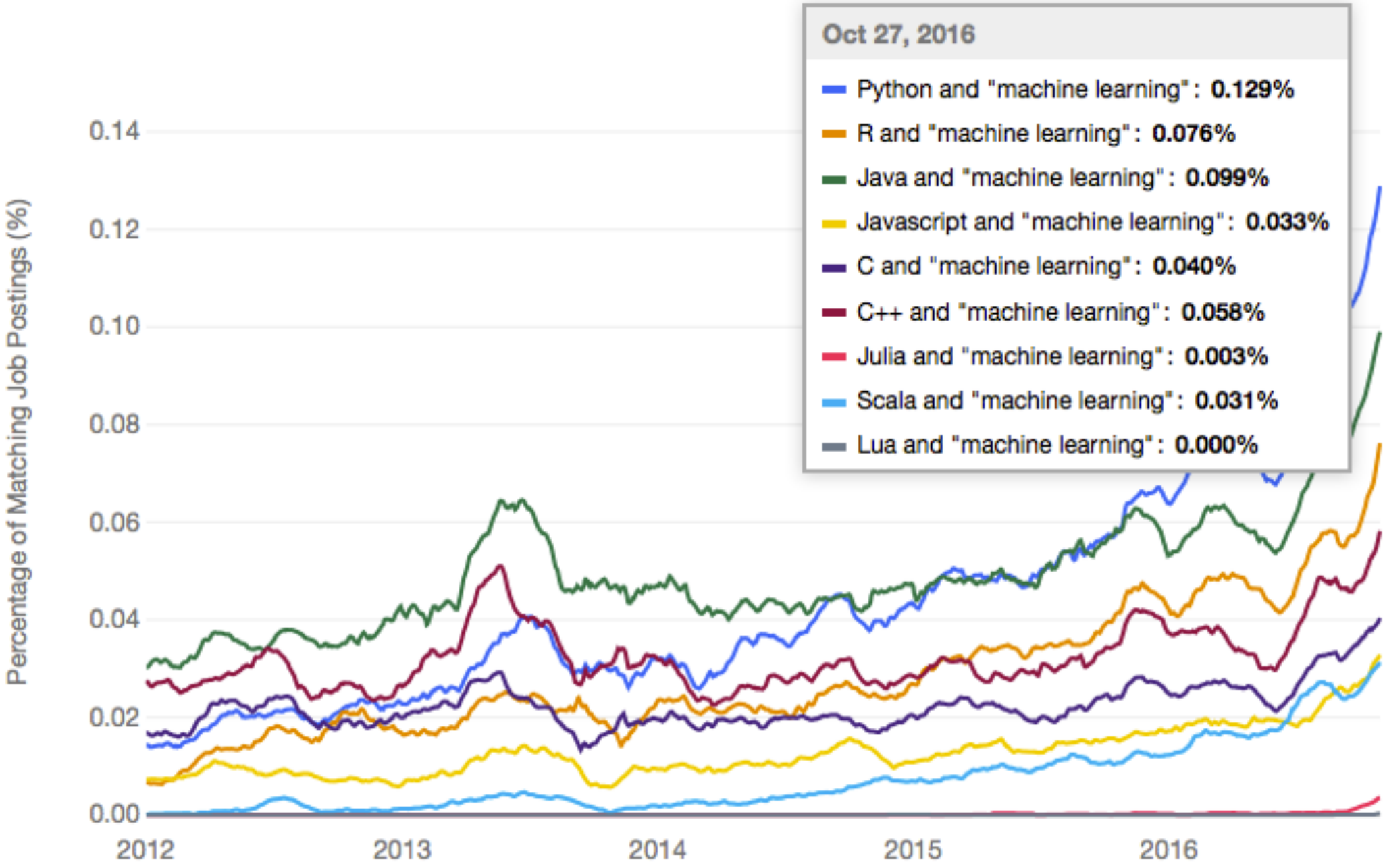

What programming/statistics languages you used for an analytics / data mining / data science work in 2013? [713 votes total]
% users in 2013
% users in 2012
% users in 2011
R (434 voters in 2013)
60.9%
52.5%
45.1%
Python (277)
38.8%
36.1%
24.6%
SQL (261)
36.6%
32.1%
32.3%
SAS (148)
20.8%
19.7%
21.2%
Java (118)
16.5%
21.2%
24.4%
MATLAB (89)
12.5%
13.1%
14.6%

What programming/statistics languages you used for an analytics / data mining / data science work in 2014?

Language used	% voters in 2014 (719 total)	% voters in 2013 (713 total)	% voters in 2012 (579 total)
R (352 voters in 2014)	49.0%	60.9%	52.5%
SAS (262)	36.4%	20.8%	19.7%
Python (252)	35.0%	38.8%	36.1%
SQL (220)	30.6%	36.6%	32.1%
Java (89)	12.4%	16.5%	21.2%


KDnuggets Analytics/Data Science
2016 Software Poll, top 10 tools
0%
10%
20%
30%
40%
50%
60%
R
Python
SQL
Excel
RapidMiner
Hadoop
Spark
Tableau
KNIME
scikit-learn
2016 %share
2015 %share


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%

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.

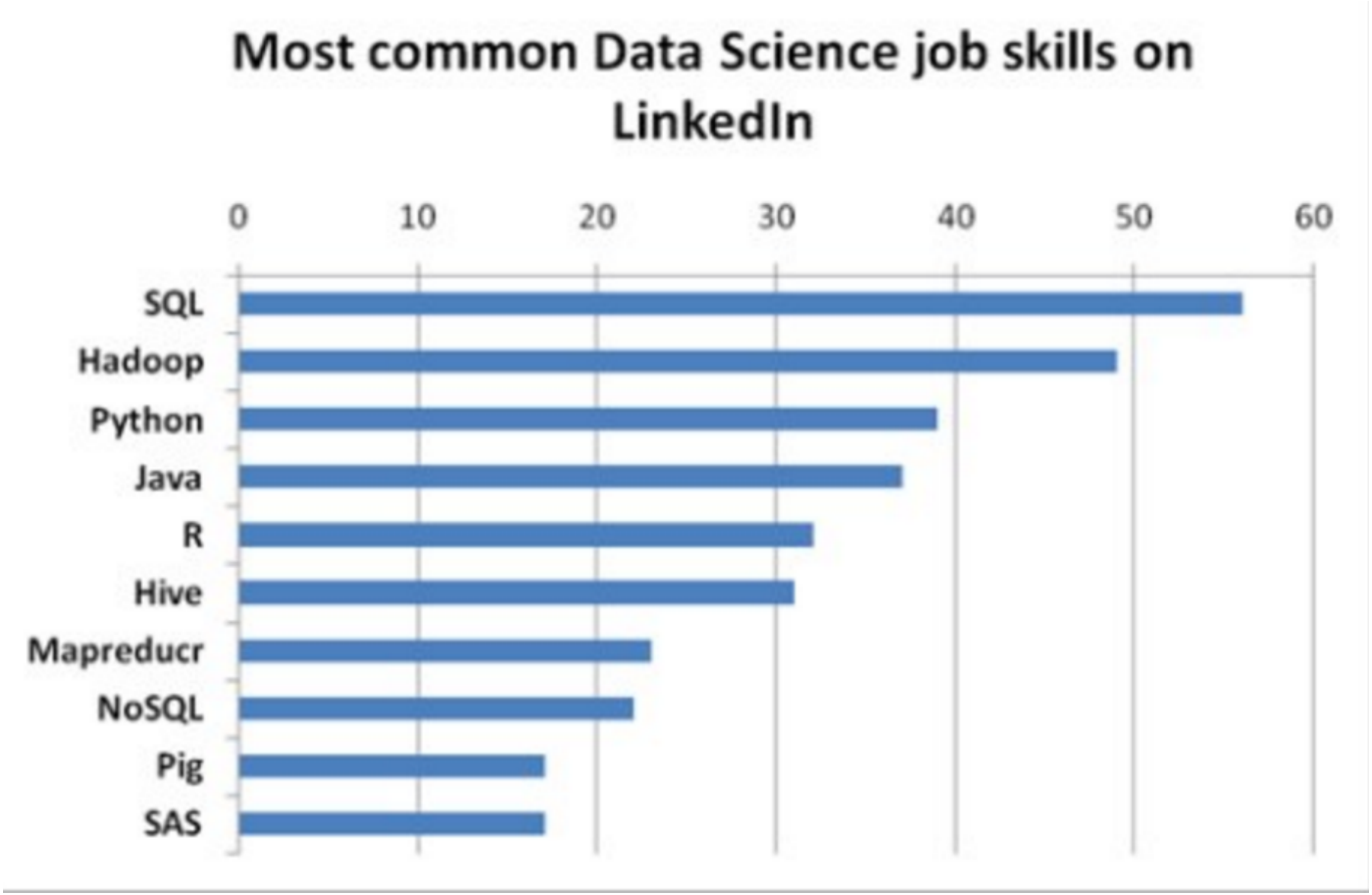
Google Tensorflow jumped to first place, displacing last year leader Theano/Pylearn2 ecosystem.

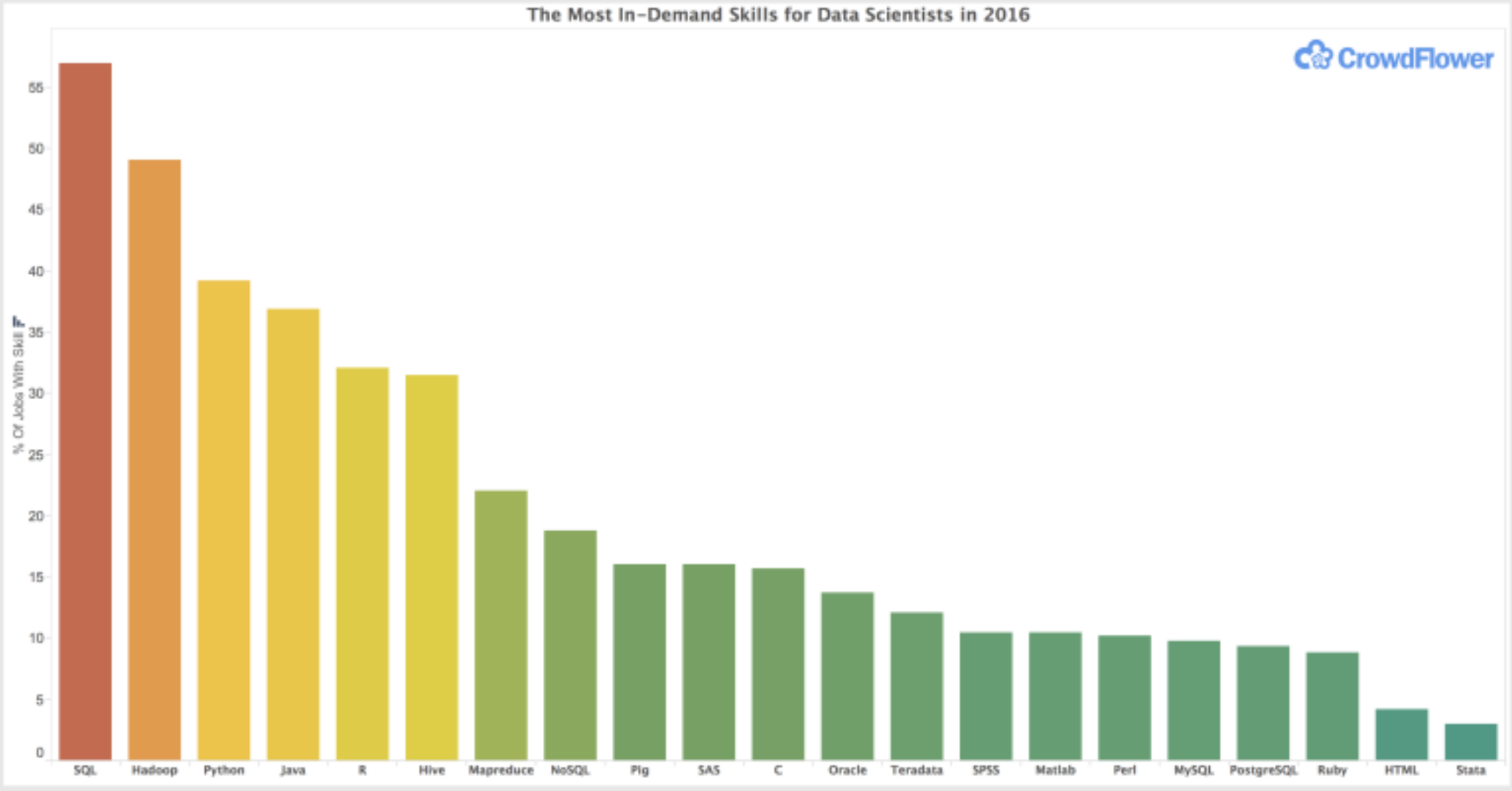
Top tools:

- 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%

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%

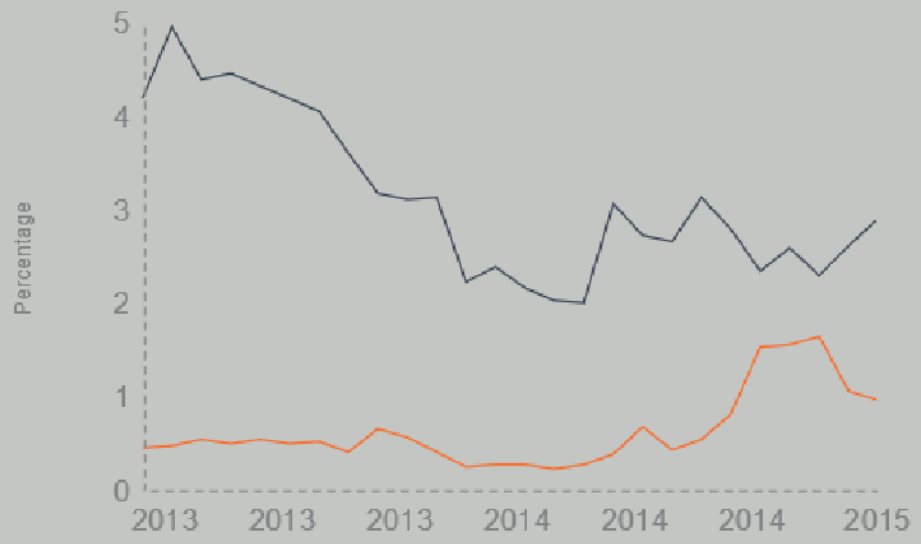

Most common Data Science job skills on LinkedIn
0
10
20
30
40
50
60
SQL
Hadoop
Python
Java
R
Hive
Mapreducr
NoSQL
Pig
SAS


The Most In-Demand Skills for Data Scientists in 2016
CrowdFlower
% Of Jobs With Skill
55
50
45
40
35
30
25
20
15
10
5
0
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)

	2012	2013	2014	2015
Python	4	4	5	4
R	17	17	15	13

Jobs And Salary?

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

$ 115,531

$94,139

Getting Started

IDE

RStudio

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.

Python for data science

Libraries for python:

- **NumPy** stands for Numerical Python. The most powerful feature of NumPy is n-dimensional array. This library also contains basic linear algebra functions, Fourier transforms, advanced random number capabilities and tools for integration with other low level languages like Fortran, C and C++
- **SciPy** stands for Scientific Python. SciPy is built on NumPy. It is one of the most useful library for variety of high level science and engineering modules like discrete Fourier transform, Linear Algebra, Optimization and Sparse matrices.
- **Matplotlib** for plotting vast variety of graphs, starting from histograms to line plots to heat plots.. You can use Pylab feature in ipython notebook (ipython notebook –pylab = inline) to use these plotting features inline. If you ignore the inline option, then pylab converts ipython environment to an environment, very similar to Matlab. You can also use Latex commands to add math to your plot.
- **Pandas** for structured data operations and manipulations. It is extensively used for data munging and preparation. Pandas were added relatively recently to Python and have been instrumental in boosting Python's usage in data scientist community.
- **Scikit Learn** for machine learning. Built on NumPy, SciPy and matplotlib, this library contains a lot of effiecient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

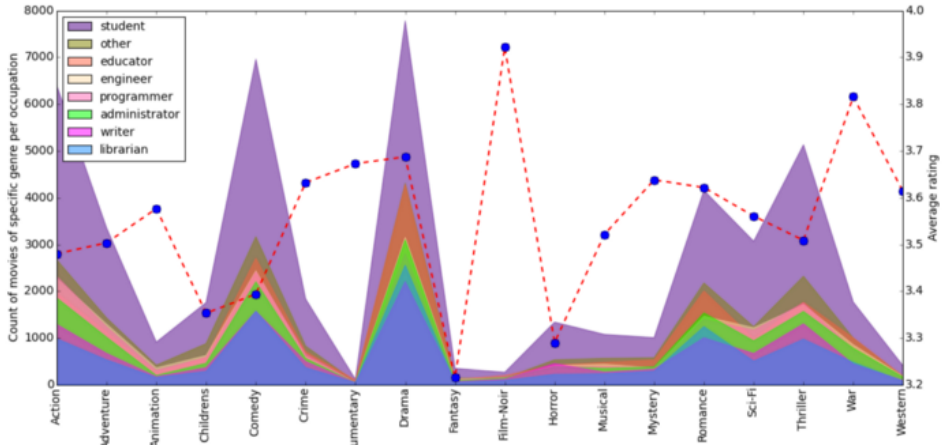
- **Statsmodels** for statistical modeling. Statsmodels is a Python module that allows users to explore data, estimate statistical models, and perform statistical tests. An extensive list of descriptive statistics, statistical tests, plotting functions, and result statistics are available for different types of data and each estimator.
- **Seaborn** for statistical data visualization. Seaborn is a library for making attractive and informative statistical graphics in Python. It is based on matplotlib. Seaborn aims to make visualization a central part of exploring and understanding data.
- **Bokeh** for creating interactive plots, dashboards and data applications on modern web-browsers. It empowers the user to generate elegant and concise graphics in the style of D3.js. Moreover, it has the capability of high-performance interactivity over very large or streaming datasets.
- **Blaze** for extending the capability of Numpy and Pandas to distributed and streaming datasets. It can be used to access data from a multitude of sources including Bcolz, MongoDB, SQLAlchemy, Apache Spark, PyTables, etc. Together with Bokeh, Blaze can act as a very powerful tool for creating effective visualizations and dashboards on huge chunks of data.
- **Scrapy** for web crawling. It is a very useful framework for getting specific patterns of data. It has the capability to start at a website home url and then dig through web-pages within the website to gather information.

Python libraries for data science 2

1. NumPy (Commits: 15980, Contributors: 522)
2. SciPy (Commits: 17213, Contributors: 489)
3. Pandas (Commits: 15089, Contributors: 762)

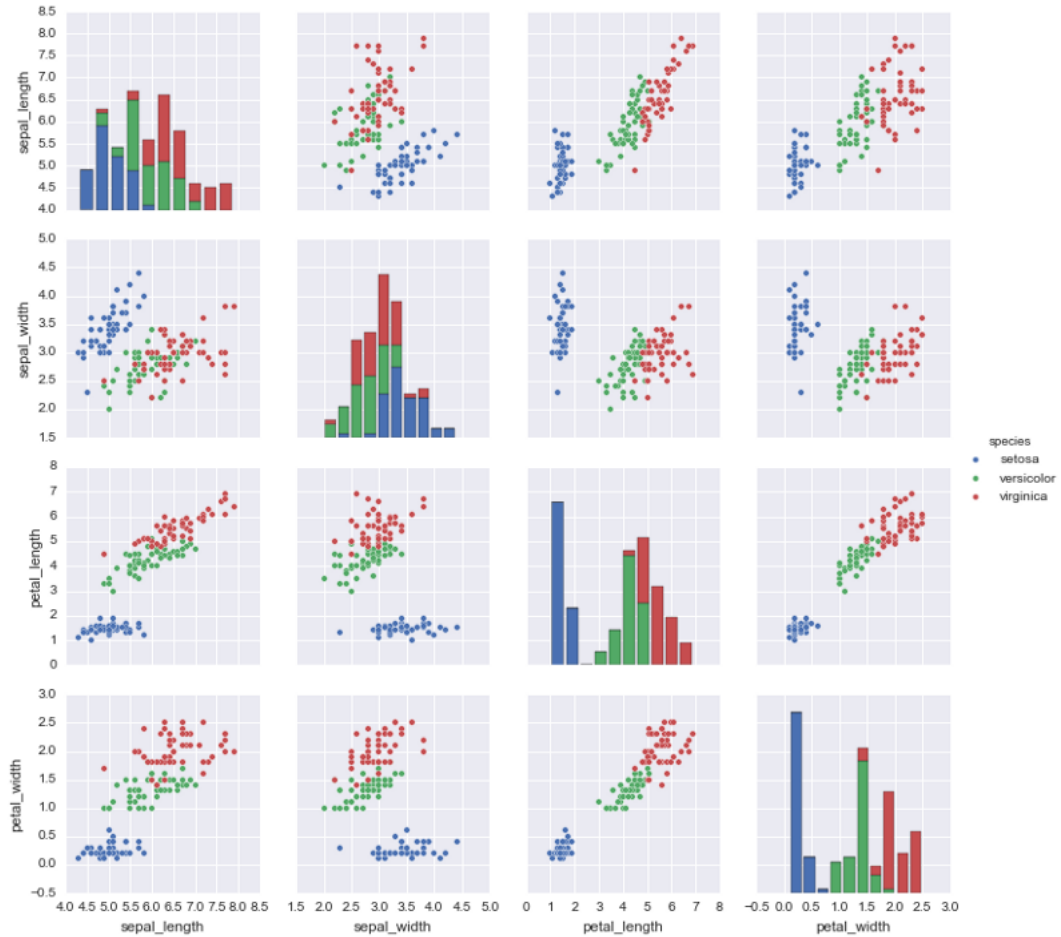
	A	B	C	D
0	A0	B0	C0	D0
1	A1	B1	C1	D1
2	A2	B2	C2	D2
3	A3	B3	C3	D3
4	X0	X1	X2	X3

4.Matplotlib (Commits: 21754, Contributors: 588)



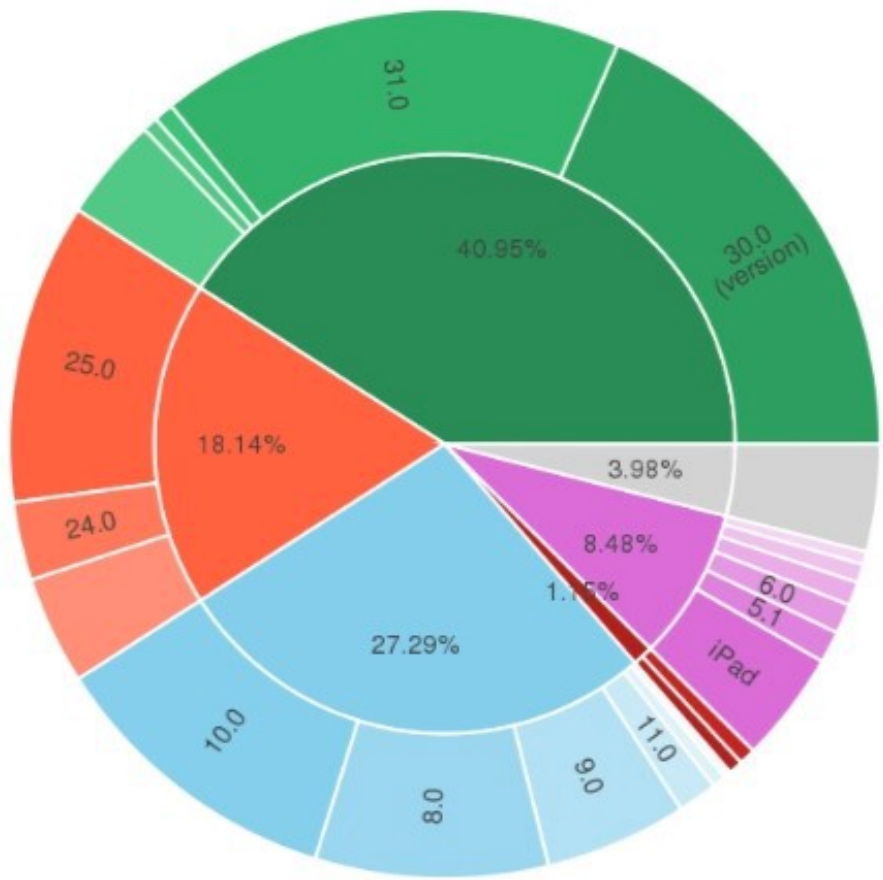
Python libraries for data science 2

5. Seaborn (Commits: 1699, Contributors: 71)



Python libraries for data science 2

6. Bokeh (Commits: 15724, Contributors: 223)



Python libraries for data science 2

7. Plotly (Commits: 2486, Contributors: 33)
8. SciKit-Learn (Commits: 21793, Contributors: 842)
9.Theano. (Commits: 25870, Contributors: 300)
10. TensorFlow. (Commits: 16785, Contributors: 795)
11. Keras. (Commits: 3519, Contributors: 428)
12. NLTK (Commits: 12449, Contributors: 196)
13. Gensim (Commits: 2878, Contributors: 179)
14. Scrapy (Commits: 6325, Contributors: 243)
15. Statsmodels (Commits: 8960, Contributors: 119)

HW 2

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

- Evaluations due Tuesday by 11:59pm
    - Use the rubric!
    - Justify your scores on each section of the rubric!