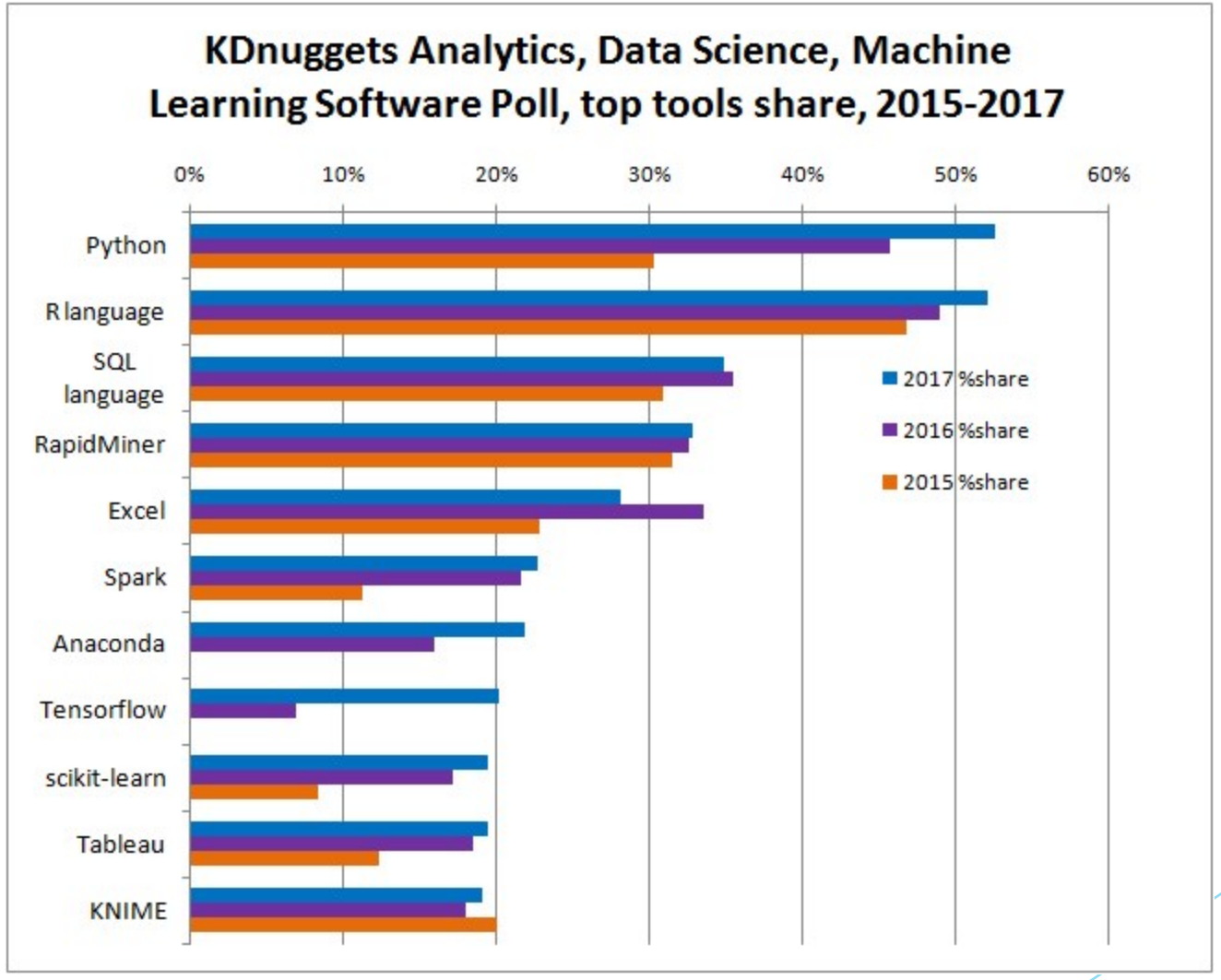
Brief Intro to Python for Statistics

By Ryan Cole

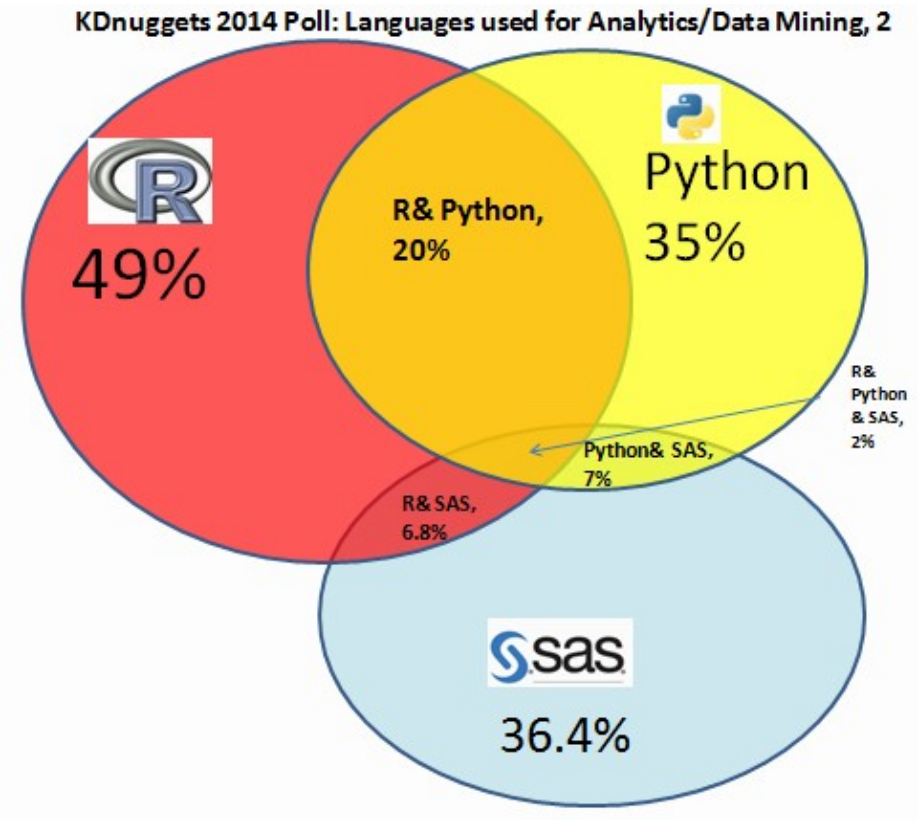
What is Python?

- Python is a popular object-oriented programming language
  - Scripting language, so is compiled in real time
  - Major features:
    - Ease of access
    - elegant syntax
    - large standard library, including interfacing with multiple software layers
  - Disadvantages:
    - Code base is fractured somewhat (two major versions are currently 3.6 and 2.7)
    - Not intended for high performance

How popular is ‘popular’?



Compare to 2014



Where do I get it?

- Recommended:
  - Install Python 3 using Anaconda
    - https://www.anaconda.com/download/
    - Platform for data science
    - Includes virtual environments
    - Includes Spyder IDE

Great, how do I learn?

- Lots of resources online
- General Python:
  - www.codeacademy.com
    - Free interactive python tutorial (also basic SQL)
    - Geared towards web development but great for basic structure
- Data Science with Python:
  - Python Data Science Handbook, by Jake VanderPlas

Notable Packages for Python

- SciPy
  - open-source software for mathematics, science, and engineering that includes:
    - Pandas uses data frames similar to R, particularly tidyverse package in R
    - NumPy array package useful for computation
    - Matplotlib for visualization – include Seaborn for quick boost to quality
- StatsModels
  - Classes and functions to estimate statistical models and conduct tests
  - Tested for correctness
  - Graphics output is less than optimal
- Scikit learn
  - Popular machine learning suite
  - with a number of procedures implemented