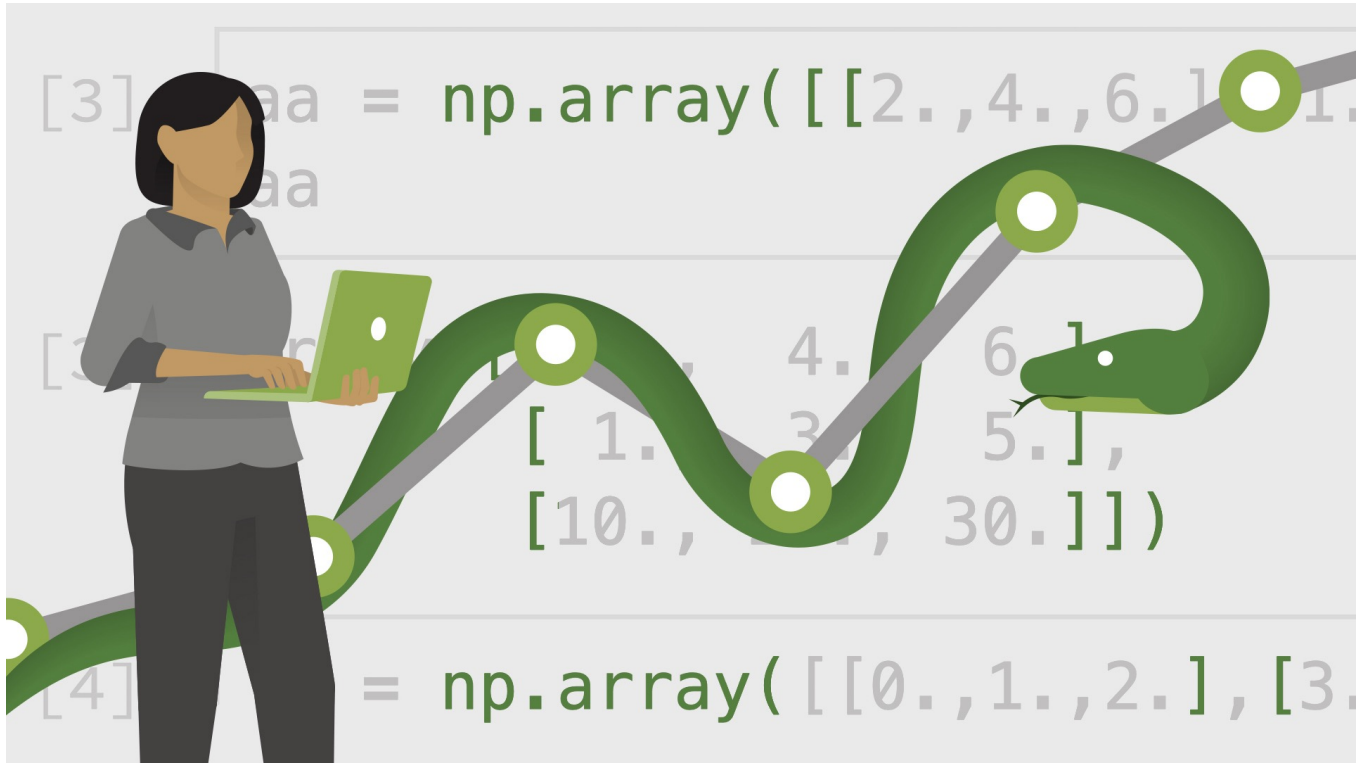
Teaching Data Science using Python



Overview

R vs. Python: which language to choose?

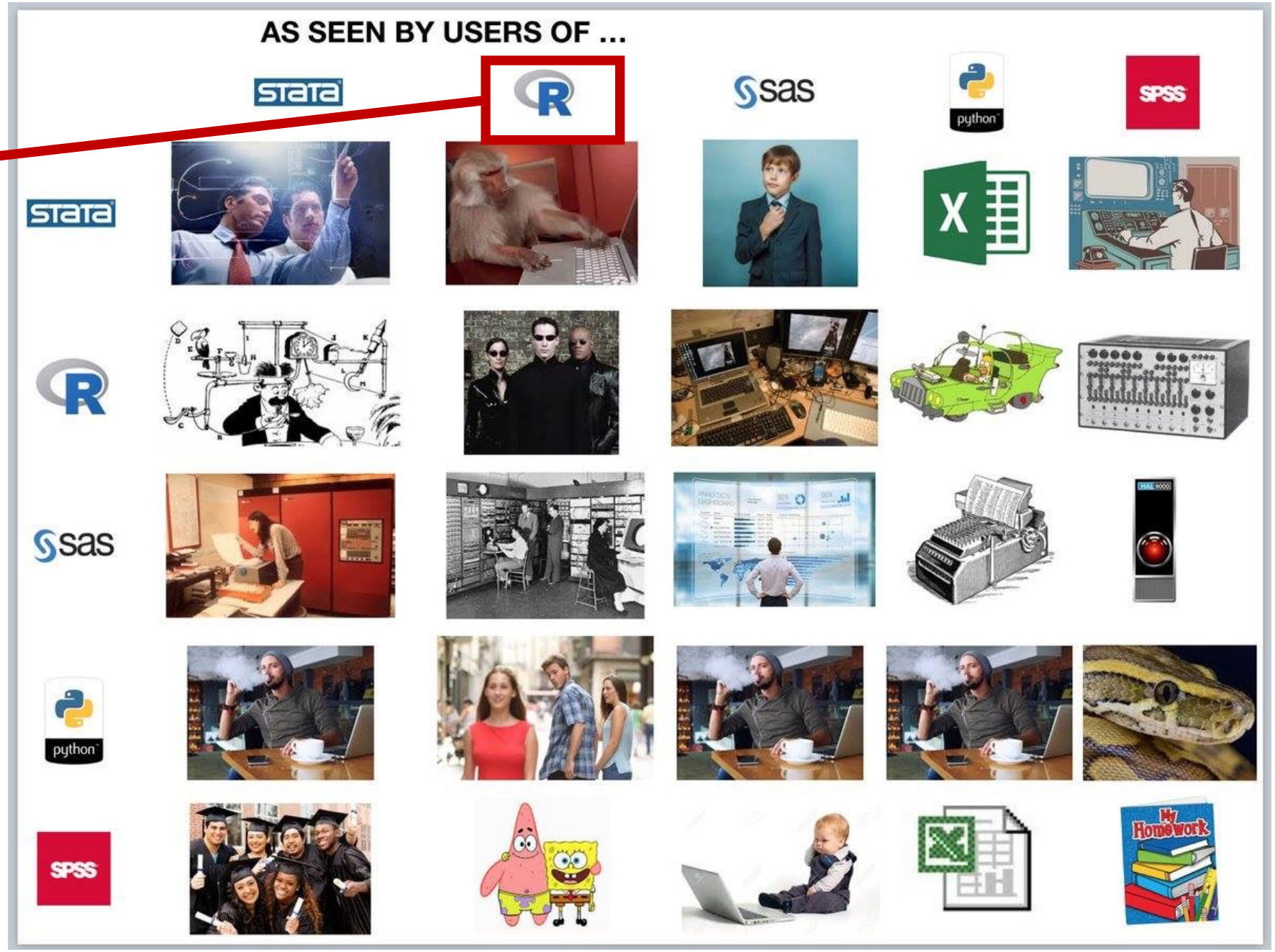
Teaching packages vs. industry packages

The Berkeley Data 8 course

Python Data Science packages

Does this seem accurate Nick and Ben?

Full disclosure



R vs. Python for teaching Data Science

General purpose programming language

Key Data Science packages

- Numpy, Pandas, Matplotlib, Seaborn

Strengths

- Better for building software
- Machine Learning

Language for data analysis

Key Data Science packages

- dplyr, ggplot, tidyr (tidyverse)

Strengths

- Better for publications
- Data manipulation

R vs. Python for teaching Data Science

Better choice if students already know Python

- Part of the CS curriculum

To learn Data Science useful to know the basics of Python

- Exception: Berkeley's Data 8 class

Better choice for students new to programming

- Levels the playing field

Can start right away on Data Science packages with little discussion of base R

Sandboxes vs. the real world

Packages exist that are designed for teaching

- Example: mosaic for teaching Intro Stats in R

Advantages of using teaching packages:

- Easier for students to learn the syntax
  - Syntax is more consistent/simpler

Disadvantages of using teaching packages:

- Does not translate as well after the class is over

Teaching Data Science using Python

Sandbox option: Berkeley's Data 8 course

- Uses the datascience package

Real world option: Python packages:

- Standard Python libraries
- Numpy
- Pandas
- Matplotlib
- Also: seaborn, statsmodels, scikitlearn

Berkeley's Data 8: The Foundations of Data Science

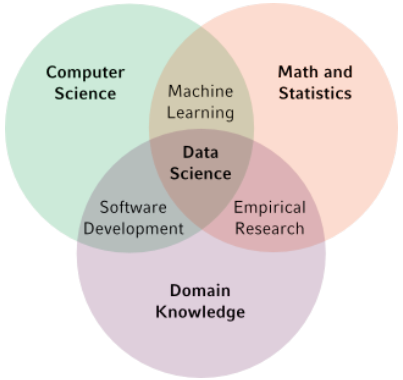
Description:

- The UC Berkeley Foundations of Data Science course combines three perspectives: inferential thinking, computational thinking, and real-world relevance. Given data arising from some real-world phenomenon, how does one analyze that data so as to understand that phenomenon? The course teaches critical concepts and skills in computer programming and statistical inference, in conjunction with hands-on analysis of real-world datasets, including economic data, document collections, geographical data, and social networks. It delves into social issues surrounding data analysis such as privacy and design.

Very broad class: lots of topics with less depth

- 1200 Berkeley students take it each semester

Berkeley's Data 8: Topics



- Cause and Effect
- Data Types
- Building Tables
- Census
- Charts
- Distributions
- Histograms
- Functions
- Groups
- Pivots and Joins
- Iteration
- Chance
- Sampling
- Models
- Comparing Distributions
- Decisions and Uncertainty
- A/B Testing
- Causality
- Confidence Intervals
- Center and Spread
- The Normal Distribution
- Sample means
- Design Experiments
- Correlation
- Linear Regression
- Least Squares
- Residuals
- Regression Inference
- Privacy
- Classification
- Classifiers

Also see YData123

Berkeley's Data 8: Resources

There are a number of resources that are available for the class:

- Online textbook
- Lecture material, assignments, etc.
  - Zero to Data 8 describes how to get started teaching the Data 8 class
- The datascience module
  - Easier syntax for table manipulation than using Pandas
- Connector classes are courses on a particular topic that reinforce Data 8 content
  - Example: YData Baseball

Berkeley's Data 8: the datascience package

The ***Table object*** is the main additional of the datascience package

- Easier to do data manipulation on tables compared to using pandas

Main methods of Table objects:

- Selecting columns: tb.select("column_name ")
- Filtering a subset of rows: tb.where("column_name", value)
- Aggregation: tb.group("column_name", aggregation_function)
- Visualization: tb.plot("column_x", "column_y")

Let's explore the datascience package!

We will use a Jupyter notebook to explore data manipulation and basic visualizations using the datascience package