# Python for Data Analysis and Visualization

Professor Widom's Instructional Odyssey

www.professorwidom.org

Stanford University

Google

Google Cloud Platform

Association for Computing Machinery

Very Large Data Bases Endowment Inc.

Instabase

amazon web services™

# Python

- Very popular general-purpose programming language
- Used from introductory programming courses to production systems

# Python Features

- Dynamically typed
  (rather than statically typed like Java or C/C++)
- Interpreted
  (rather than compiled like Java or C/C++)

Python programs are comparatively…

+ Quicker to write
+ Shorter
– More error-prone
– Slower to run

# Python for Data

- Fairly easy to read/write/process data using standard features

- Plus special packages for…
  - Numerical and statistical manipulations - numpy
  - Visualization (“plotting”) - matplotlib
  - Relational database like capabilities – pandas
  - Machine learning - sklearn
  - Network analysis - networkx
  - Unstructured data – re, nltk, PIL

# Python Versus R

## Python

- Good for beginners or experienced programmers
- Used by software engineers of all types
- Well integrated with general-purpose coding
- Not especially fast

## R

- Easier for experienced programmers
- Used by academics, researchers, hard-core data scientists
- Specialized code for complex analyses, statistics, graphics
- Extremely slow!

# Data Sets

## Europe Temperatures

**Cities:** city, country, latitude, longitude, temperature

**Countries:** country, population, EU, coastline

## 2010 World Cup

**Teams:** team, ranking, games, wins, draws, losses, goalsFor, goalsAgainst, yellowCards, redCards

**Players:** surname, team, position, minutes, shots, passes, tackles, saves

## Titanic

**Titanic:** last, first, gender, age, class, fare, embarked, survived

# Jupyter Notebooks

(formerly iPython notebooks)

- Modeled after “laboratory notebooks”
- In one notebook can combine text boxes with boxes containing executable code in a wide variety of languages
- Can run/re-run boxes (cells) individually, or run/re-run entire notebook

Rapid adoption in many sectors

# Jupyter Notebooks

- Can download to your computer (recommend *Anaconda*) but no one-button download yet
- We will use notebooks in the cloud, via *Google Colab*
- Either way, notebooks run in a web browser

To execute a code cell, click inside the box then click ▶.

Or use *shift*, *control*, or *command* with *enter* or *return*

# Agenda

1. Python basics
2. Data manipulation
3. Pandas
4. Plotting

(more in modules on Machine Learning, Data Mining, Network Analysis, Unstructured Data)

Plenty of your turn!

For help while working with Python:

Tutorials and help pages
(website Course Materials)

- ➢ Web search