Week 6

Foundations

For loops

The for-loop and sampling framework

(demo)

Sampling

- **Parameter**
  - A number associated with the population
- **Statistic**
  - A number calculated from the sample

Probability

Key words: AND, OR, At least

And - multiply ex. probability of a 6 and a 1 in two rolls

Or - addition ex probability of a 6 or a 1 in two rolls

At least - find the complement

(demo)

Moving beyond data 8

Trajectory

- Programming - CS 61A
- Data Structures - CS 61B
- Industry level Data Science - Data 100
- Probability - Stat 140, 134
- Linear Algebra - Math 54
- Machine Learning - CS 189
- Ethics - Info 188

Careers in Data Science

Four main choices:

- Data Engineer
- Data Scientist
- Data Analyst
- Using data science in a domain

Data science as a field is relatively new, roles are very ill-defined. Many of them overlap

Data Engineer

More behind the scenes, how do we prepare data to be analyzed?

Answers questions like:

How do we store data (size, format)?

Tools:

SQL, Hadoop, Spark, “Ground Level” programming languages (C, Java, + Python)

Sample Job

Data Scientist

Analyzing data, creating models, testing for significance.

Answers questions like:

Does this data tell us something significant? What models can we create off of it? What can we predict about the future?

Sample Job

Data Analyst

Forward facing,

Primarily trying to advise policy, strategy based upon data.

Answers questions like:

What industries should a company try to target? How do we increase profits?

Sample Job

Using data science in domains

- Data Science becoming more prevalent in all fields
- Emerging fields: Biostatistics, law, criminal justice, journalism, politics, econometrics, medicine

- https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2985861

Data Science Tools

Anaconda - https://www.anaconda.com/download/

Terminal

RStudio

Text editors - Atom, Sublime, TextWrangler

IDEs - Pycharm, IntelliJ

Languages/Packages

SQL

Pandas/Numpy

Matplotlib/Seaborn

Python

R

Excel/Tableau

Terminal

Mac OS X/Linux - preinstalled. Windows - Download Git Bash

Basic commands:

cd

mkdir

ls

mv

git

jupyter notebook

Version Control

Github- public place to host your code, projects, helps you collaborate with others

Using git, can sync between your computer and the web

Basic commands: git clone, git add, git commit, git push origin master, git pull origin master

https://sp18.datastructur.es/materials/guides/using-git.html