Intro to Data Science Workshop

Why are we here?

1. Intro to the Codeup experience
2. Overview of Data Science
3. Intro to visualizing, exploring, and modeling data. *We will do some machine learning today!*

Why Codeup?

- Focus on student outcomes
- Placement services and quality of network
- Immersion works. Full-time, live instruction for 5 months works.
- Projects simulate the work environment from real world data to presenting findings

What is Data Science?

- Interdisciplinary applied science intersecting programming, statistics, and domain expertise
- The application of the scientific method of hypothesis-experiment-analyze-repeat to analyze and infer outcomes from data.
- A broad description of approaches ranging from business analysis and visualizations to machine learning and deep neural network analysis.
- An increasingly accessible field

What *isn't* Data Science?

- Only statistics or only mathematics: let the computer compute and the people think
- Free from scrutiny. Methods and findings deserve serious technical and ethical scrutiny.

Isn't data science just statistics?

- *Future of Data Analysis*, Tukey 1962, `https://projecteuclid.org/euclid.aoms/1177704711`
- *50 Years of Data Science* by Donoho, `https://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf`

The five questions Data Science can answer

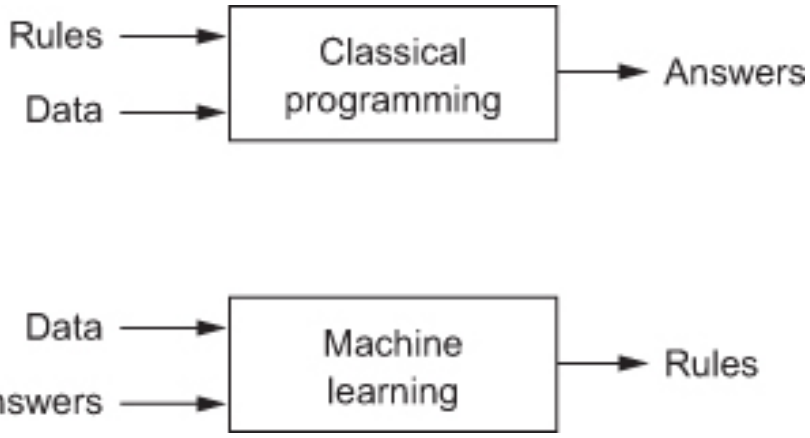
1. How many or how much of something? What will sales be next year?
2. Which category does this thing belong to: Is this A or B?
3. What groups? Is there a clustering of the data that tells a story? Who are our best customers?
4. Is this weird? Anomaly detection like looking for fraudulent transactions
5. What should we do next? Infer likely outcomes

Overview of Machine Learning

Broadly, machine learning is the process of using previous data as the fuel for determining rules for making predictions of outcomes from future data.

Classical programming takes business rules and data to produce answers. Ex. TurboTax software.

Machine learning takes in data (and sometimes answers/labels for some data) and produces rules or predictions for future data. The example here is text message autocomplete.



Types of Machine Learning

- Supervised = labeled data. Example is marking messages as spam to train a classifier
- Unsupervised = no labels or human provided answers
- Semi-supervised = mixture of supervised/unsupervised
- Reinforcement learning = use algorithm A to "grade" the results from algorithm B

Main Challenges of Machine Learning

- Garbage in, garbage out
- Insufficient quantity of data
- Nonrepresentative data
- Poor quality data
- Overfitting or underfitting
- Bias in, bias out

Homework and next steps

- `https://www.kaggle.com/ryanorsinger/101-exercises` for deep Python practice.
- *Data is the new oil*, so sharpen your data skills.
- Talk to admissions and apply today!