Introduction

Introduction to Data Science with Python

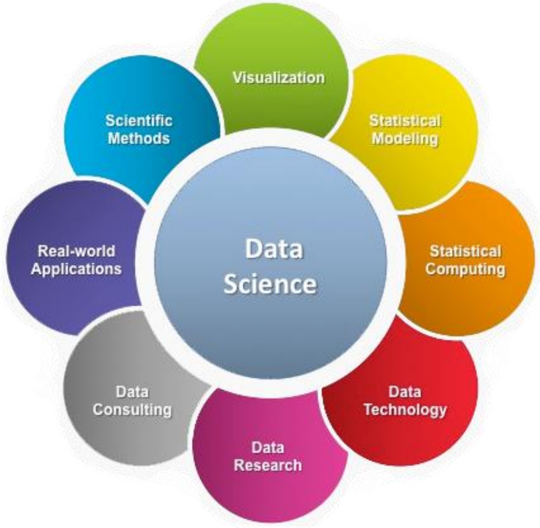
Let me introduce myself

- Kevin McCarty - https://www.linkedin.com/in/kevin-mccarty-96a4487/
- kevin.mccarty@my.gcu.edu
  - Adjunct Professor of Business Analytics at Grand Canyon University
  - Data Scientist and Facilitator for data science and analytics courses
  - BS in Theoretical Mathematics, MS, PhD in Computer Science from the University of Idaho
  - 30+ years in the industry
  - Former Army officer and Eagle Scout
  - Fun facts: Was on Family Feud in the 80s (won the big money once), started elementary school classified as "retarded/slow"

Let’s get acquainted!

- Name
- Job title
- Experience/background
- What you hope to get out of this course
- Any fun facts?

What is Data Science?



- Humans employ many techniques to understand their surroundings
  - Inference
  - Pattern recognition
  - Memory

- Unfortunately, when it comes to data, there too much of it to process for humans

- Data Science is the use of multiple disciplines in combination to make sense and gain insights of otherwise nonsensical data (kind of a hard definition to nail down)

What do Data Scientists do?

- Data Scientists bring together aspects of business, programming, statistics and communication to turn raw data into actionable insights
  - If this sounds “hard” it is because it is

- Some of the things data scientists “do”
  - Answer important business questions
    - Spending more here will increase sales of this by how much?
    - What can we do to be more efficient in our manufacturing process?
    - Which customers should we focus on for best results?
  - Explain the underlying data
    - Dashboards and charts

What do Data Scientists do?

- Develop programs using sophisticated algorithms
  - Prediction models
  - Statistical inference applications to data
  - Pattern recognition
  - Clustering
  - Recommenders
  - Anomaly (e.g. fraud) detection

- Serve as a liaison between subject matter experts and laypersons

The Data Science Life Cycle

- There are different takes on this such as CRISP-DM, TDSP and others but they all follow pretty much the same pattern
  - Business Understanding – the Question
    - What are we trying to accomplish?
  - Data Gathering
    - You should expect to find sources in many different databases, file repositories, the web and elsewhere
  - Data Analysis
    - This is where you discover whether you have the right data to answer the question, if it is in the right format, timely, available, etc.
  - Data Cleaning
    - Fix bad data, missing data, change names, datatypes, etc.

The Data Science Life Cycle

- Data Exploration/Analysis
  - Get statistics
  - Create visuals
- Feature Engineering
  - Dimension Reduction, PCA, Isomap
  - Creative recombination of variables
- Modeling
  - Build and train one or more models for prediction, classification, etc.
- Model Evaluation
  - See how the model does on real/unknown data
  - Interpret results

The Data Science Life Cycle

- Model Deployment
  - Put the thing in production
- Results Visualization
  - Provide data to customers
- Retraining/redeployment
  - Essentially we start the whole process over
  - As data grows stale, our models are likely to loss their effectiveness
- Keep these phases in mind as we move through this course

What will we learn in this course

- Since this is Data Science with Python we will focus on those libraries and methods which will often be used in pursuit of data science
- Setup with Python/Anaconda
- Python Crash Course
  - Variable
  - Data Structures
  - Control Flow
  - Functions and Modules
  - Object-oriented programming

What will we learn in this course

- Python Modules for Data Science
  - NumPy
  - Pandas
  - Matplotlib
  - Seaborn

- Data preparation
  - File IO
  - Database IO
  - Data cleaning
  - Dimension Reduction
  - Feature Engineering

What will we learn in this course

- We will look at some techniques data scientists employ
  - Inferential Statistics
  - Machine Learning
  - Model Evaluation
- Along with way we will do a series of exercises and labs to try to reinforce the material

Miscellaneous

- Each 4-hour session will include a 5-15 minute break at the end of each hour
  - I try to combine them with labs/exercises so they won’t be at a regular time
- Please be prompt, class will start right on time
- Feel free to ask questions via chat or mic (but mute mic when not speaking)
- Feel free to email me
- Let’s have a great time!!

Resources I Have Found Helpful

- Stack Overflow: https://stackoverflow.com
- Coursera: https://www.coursera.org/
- Udemy: https://www.udemy.com
- Kaggle: https://www.kaggle.com/
- DataCamp: https://www.kaggle.com/
- EdX: https://www.edx.org/
- Analytics Vidhya: https://www.analyticsvidhya.com
- Real Python: https://realpython.com

Lab 1 – The Data Science Lifecycle (20 min)

- Open up the Lab 1.docx file
  - List the individual(s) in your own organization who might fill in each role
  - Fill in the information for each step