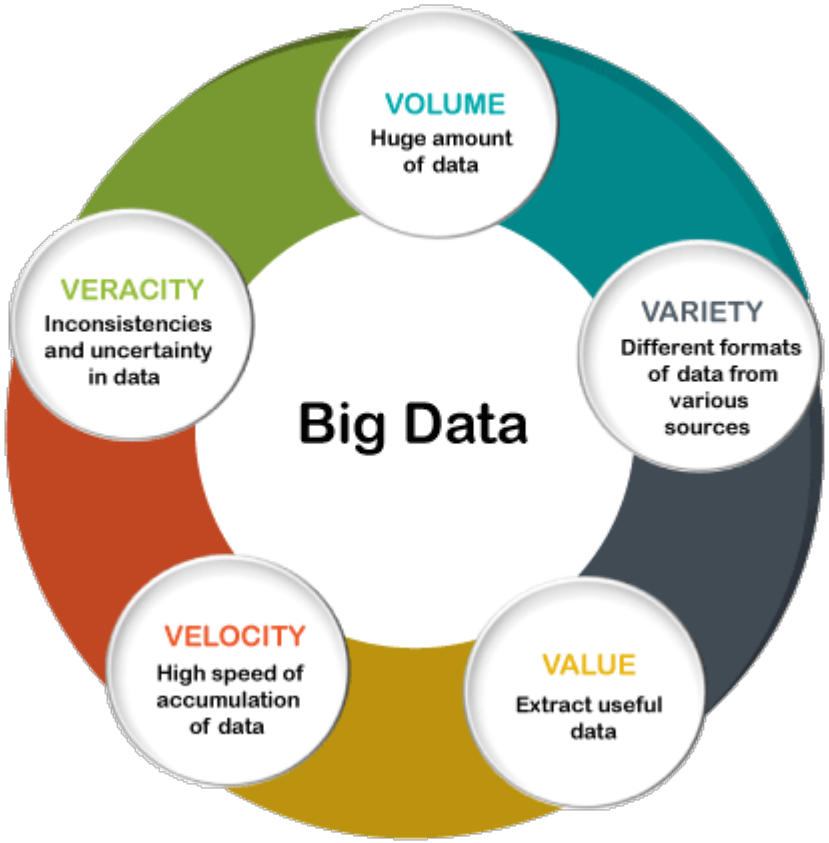
Big Recap!

Justin Post

Welcome to Big Data Analysis!

What is Big Data?

- 5 V's of Big Data
  - Volume
  - Variety
  - Velocity
  - Veracity (Variability)
  - Value



Course Plan

- Course split into four topics
    1. Programming in `python`
    2. Big Data Management
    3. Modeling Big Data (with `Spark` via `pyspark`)
    4. Streaming Data

Python Recap!

- Markdown capabilities of `JupyterLab`
- Modules
- Basic data types
  - Strings, Numeric Types, Booleans
  - Lists, Tuples, Dictionaries
- Advanced data types
  - Numpy arrays
  - Pandas series and data frames

Python Recap!

- Markdown capabilities of `JupyterLab`
- Modules
- Basic data types
  - Strings, Numeric Types, Booleans
  - Lists, Tuples, Dictionaries
- Advanced data types
  - Numpy arrays
  - Pandas series and data frames
- Writing Functions
- Control flow (if/then/else, Looping)
- Summarizing Data
  - via pandas
  - via matplotlib

Python Recap!

- Functions, Methods, and Attributes

Python Recap!

- Functions, Methods, and Attributes
- Sequence type objects act similarly

Python Recap!

- Functions, Methods, and Attributes
- Sequence type objects act similarly
- Iterable type objects

Python Recap!

- Functions, Methods, and Attributes
- Sequence type objects act similarly
- Iterable type objects
- Indenting for code formatting

```
if condition:
    code to execute

def my_fun(arg):
    body of function
```

Python Recap!

- Functions, Methods, and Attributes
- Sequence type objects act similarly
- Iterable type objects
- Indenting for code formatting

```
if condition:
    code to execute

def my_fun(arg):
    body of function
```

- List and dictionary comprehensions

Moving Towards Data Analysis

Four major goals with data:

1. Description
2. Prediction/Classification
3. Inference
4. Pattern Finding

EDA

- Essentially **Descriptive Statistics** with a bit more big picture stuff about your data
- EDA generally consists of a few steps:
    - Understand how your data is stored
    - Do basic data validation
    - Determine rate of missing values
    - Clean data up data as needed
    - Investigate distributions
        - Univariate measures/graphs
        - Multivariate measures/graphs
    - Apply transformations and repeat previous step

Statistical Learning

Statistical learning - Inference, prediction/classification, and pattern finding

- Supervised learning - a variable (or variables) represents an **output** or **response** of interest
    - May model response and
        - Make **inference** on the model parameters
        - **predict** a value or **classify** an observation

Statistical Learning

Statistical learning - Inference, prediction/classification, and pattern finding

- Supervised learning - a variable (or variables) represents an **output** or **response** of interest
    - May model response and
        - Make **inference** on the model parameters
        - **predict** a value or **classify** an observation

Goals:

- Understand basic modeling ideas
- Fitting a model
- Evaluating the model
- Testing/Training
- Cross Validation