CS506 Lecture

Override filters	☐
Room	KCB 104

Notes Intro To Data Science:

We are mostly in the space of capturing relationships and not find equations for exact values when doing Data Science

Hypothesis: Remember a Good example must always be falsifiable

Confirmation Bias

- A set of examples may not always tell you what rule the examples are made of
- There may be infinitely many rules that may match an example
- Rules and examples can CHANGE over time
- There may be bias in your data collection process “99 percent of survey takers say they love taking surveys’

Positive Example:

- Shows your hypothesis is correct
- If you only look at positive examples then we are only going to get positive results

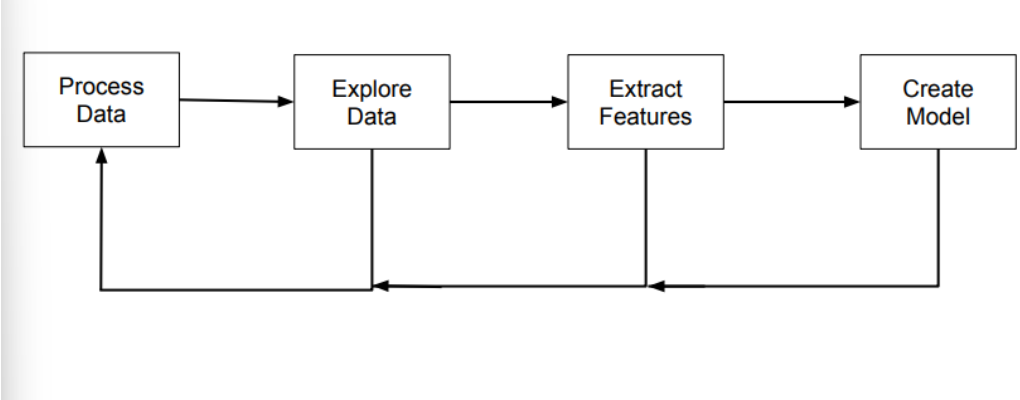
- If you only do positive examples you're only going to reaffirm your bias to your hypothesis

Negative Example:

- This shows your hypothesis is incorrect
- If you look at negative examples, you will see examples that fit the rule but not your hypothesis

Building Models

- When building models Its really all about the data
- Data Science Workflow simplified



What Data-Science we will be doing:

- m-dimensional points/vector
- Graphs like (330) using an adjacency matrix or Adjacency list
- Images using a Matrix of pixel
- Texts and Corpuses of texts: We represent a list of words or matrix of values of how many times the word appears in the corpus

Types of Learning:

- Supervised Learning:
  - Can you change some factors to test your hypothesis about some correlation?
  - For example Tumor Patients can you get younger patients to test your hypothesis to see if there is a relationship between age and tumor malignancy?
  - Tools
    -
- Unsupervised Learning
  - Goals
    - Better understand/describe the data
      - Data exploration/visualization
      - Find anomalies
  - Extract features
  - Fill in the Gaps
  - Make learning algorithms faster by getting rid of noise (Assuming that is unnecessary data)
  - Tools:
    - Clustering
    - Matrices (from CS-132)