Data Foundations: Course Introduction

Instructor: Anthony Rios

Outline

Course Introduction

- What is Data Science?
- The Data Science Pipeline: Prediction

Readings

Computational and Inferential Thinking: The Foundations of Data Science
By Ani Adhikari and John DeNero
Chapters 1.1 and 1.2

Data Mining: The Textbook
by Charu C. Aggarwal
Chapter 1

What is Data Science?

Drawing useful conclusions from data using computation

- **Exploration**
  - Identifying patterns in information
  - Uses visualizations
- **Inference**
  - Quantifying whether those patterns are reliable
  - Uses randomization
- **Prediction**
  - Making informed guesses
  - Uses Machine Learning

Exploration

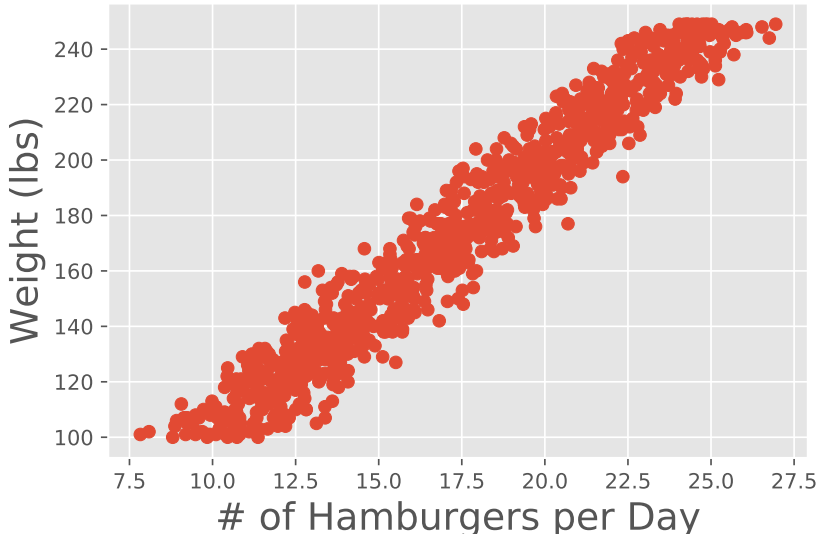
Frequent Itemset Mining

ID	Items
1	{Bread, Milk}
2	{Bread, **Diapers**, **Beer**, Eggs}
3	{Milk, **Diapers**, **Beer**, Cola}
4	{Bread, Milk, **Diapers**, **Beer**}
5	{Bread, Milk, Diapers, Cola}
...	...

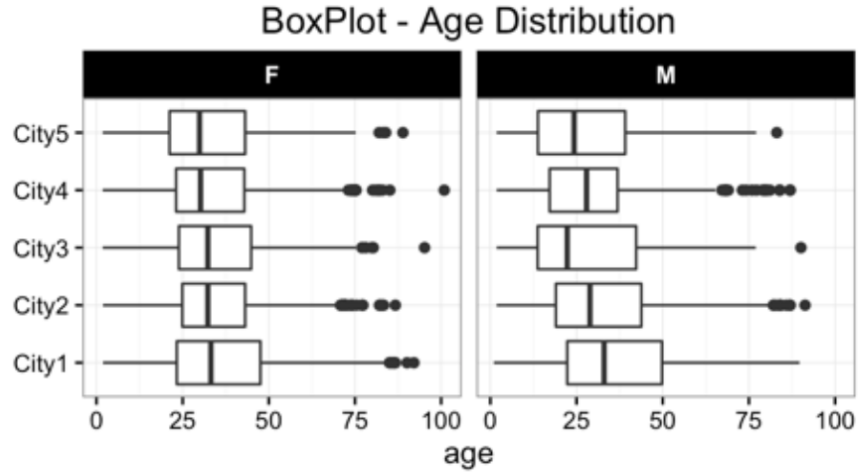
Example Freq. Itemset:
{Diapers, Beer}

Example Association Rule:
{Diapers} $\rightarrow$ {Beer}

Plotting Correlations



Other Plots



Inference

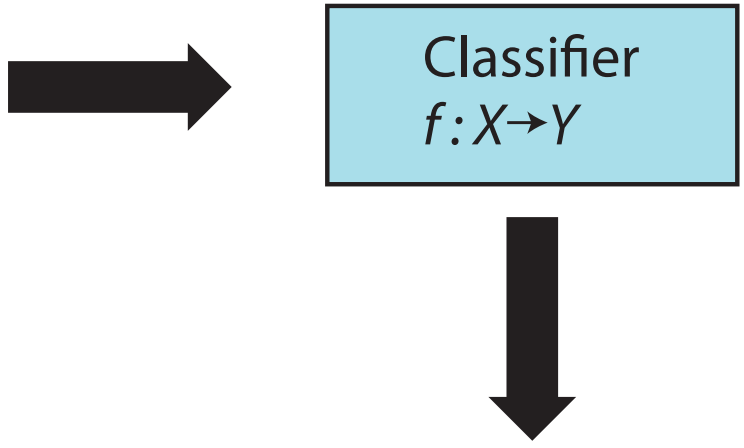
Hypothesis (A/B) Testing:

- Randomly Select 50% of users to see headline A
  - Trump Shared Classified Data With Russians, Officials Say.
- Randomly Select 50% of users to see headline B
  - My Psychic Dog has Healing Powers!
- Which headline do people click more?

Prediction

Machine Learning, Neural Networks, SVMs, Feature Engineering, ...

X ~ images



Y ~ Classes {"Cat", "Dog"}

Predict House Prices, Sentiment Classification, Scene Parsing, ...

`https://www.youtube.com/watch?v=VUrqddjkxok`

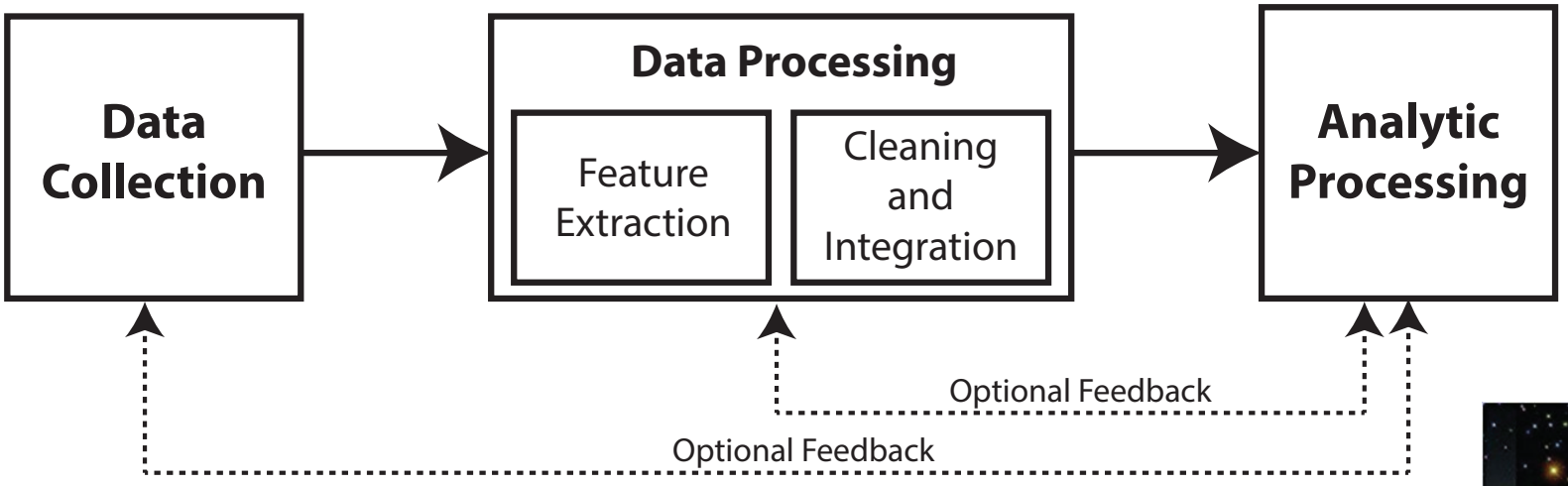
What is Data Science?

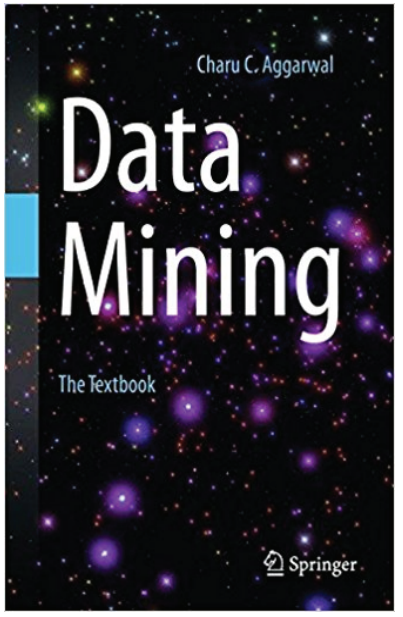
Drawing useful conclusions from data using computation

- **Exploration**
  - *DA 6223: Data Analytics Tools and Techniques
  - DA 6233: Data Analytics Visualization and Communication
- **Inference**
  - *DA 6213: Data Driven Decision Making and Design
  - STA 6443: Data Analytics Algorithms I
- **Prediction**
  - *DA 6213: Data Driven Decision Making and Design
  - *DA 6223: Data Analytics Tools and Techniques
  - STA 6543: Data Analytics Algorithms II
  - IS 6713: Data Foundations

* marks classes that cover multiple data science areas.

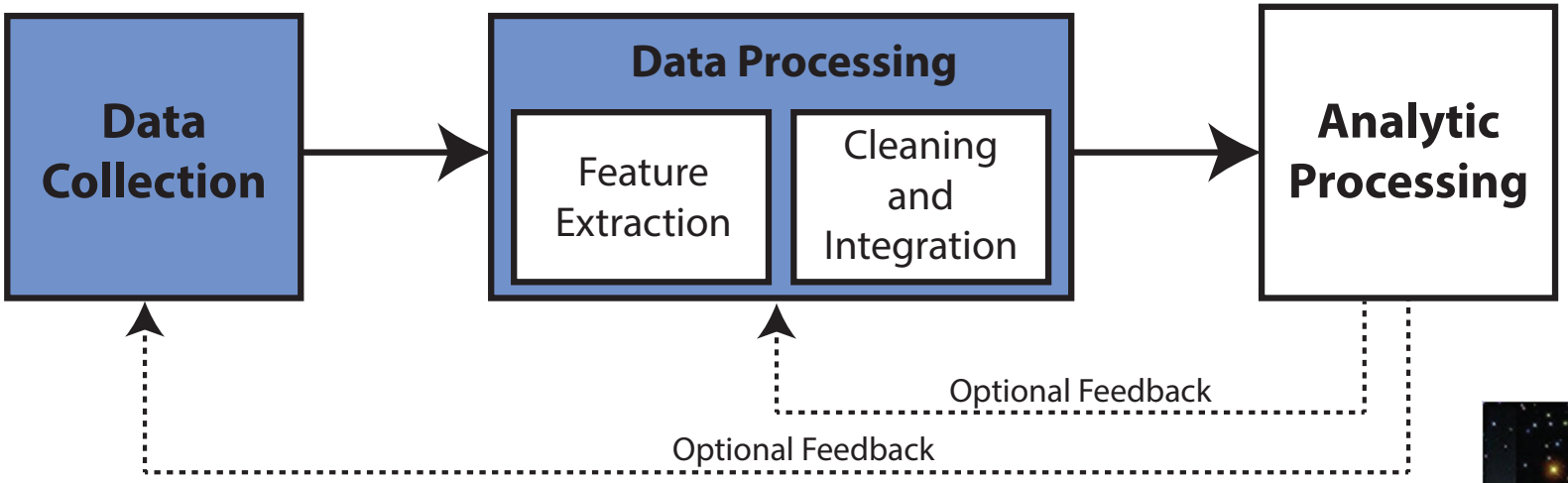
Data Processing Pipeline



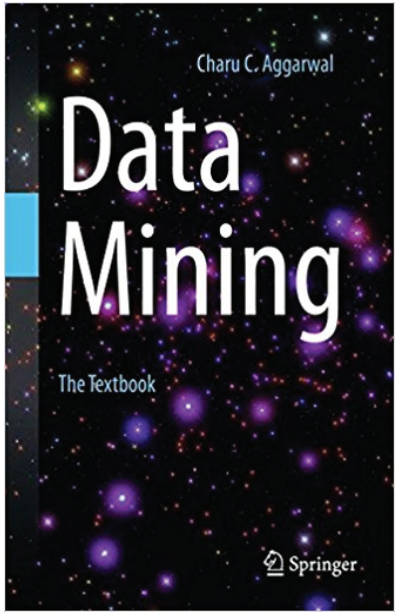


Data Processing Pipeline: Prediction

What we will cover in this course:



- This course focuses on data collection and data processing issues not covered in other courses.



Data Collection

This course will touch on the following subjects:

- Annotating Data
- Evaluating Annotations
- Pulling Data From The Web
- Loading different data formats
  - JSON(L)
  - XML
  - Text
  - CSV
- "Processing" data

Everything in this course is built on a strong coding background. So, the first part of the semester is focused on developing your coding ability.

Data Processing

This course will touch on the following subjects:

- Feature Engineering
- Combining different data modalities
- Feature Selection
- "Cleaning"
  - Handling Missing Features
  - Unbalanced Datasets

The End