Data Science (not covered in text)

FROM READING ARTICLE (not to be covered in class):

What is Data Science?

What is Big Data?

Why is Big Data so buzzworthy today?

What are the three V's of Big Data?
  1. 
  2. 
  3. 

What is the difference between structured and unstructured data? What is semi-structured data?

What are three popular choices for statistical and algorithmic data processing/visualization?
  1. 
  2. 
  3. 

TO BE COVERED IN CLASS:

Are Data Science and Big Data “new fields”?

Chapter Data Science
Expertise: Data Scientists should be good at:
1.
2.
3.
4.

The Data Science Process:
1.
2.
3.
4.

**STEP #1:**

Where can you find data?
- Kaggle.com
- Data.gov
- Github - awesome-lists

Common Data Formats:
1.
2.
3.
4.

**STEP #2:**

How to clean data?
**STEP #3:**

Exploratory Analytics: determine key metrics about your data to understand it:
1. 
2. 
3. 
4. 

What is Anscombe’s Quartet?

Common Types of Data Visualizations:
1. 
2. 
3. 
4. 
5. 
6. 

**STEP #4:**

Predictive Analytics: Goal is to train a model based on the available data in order to predict unknown values.

How do we determine if the model is good?

Training/Selecting a Model
1. Split all data into TWO subsets: …
2. Split training dataset into TWO subsets: …
3. Train multiple models on the model dataset
4. To select best model, …
5. Confirm results are reasonable on the …

What is overfitting?
The Data Science Process (more formal names):
1. 
2. 
3. 
4. 

True/False: Big Data is a solution to a problem?

The three V’s of Big Data:
1. 
2. 
3. 

What does IBM/Wikipedia think s/b added?

Big Data Architecture

Databases are used to ....

Databases optimize for the ....

Data Warehouses are used to ...

Data Warehouses optimize for the ...

MapReduce: a framework for parallelizing calculations:
1. Map:
2. Reduce: