COMP 333 — Week 7 Welcome

Data Cleaning

In Week 6 and Week 7 (this lecture) we cover Data Wrangling
which is the most time-consuming phase of Data Analytics.

This week we cover Data Cleaning
an important step in Data Wrangling.

It is very important to clean and organize your data.

Remember GIGO (Garbage-In, Garbage-Out)

A concise discussion of using data cleaning
so GIGO (Garbage-In, Garbage-Out) does not effect machine learning
is `https://elitedatascience.com/data-cleaning`

This week we cover

- Data Cleaning
- Missing Values
  including the imputation (or inference) of missing values
- Unification
  including normalization and the role of Z-scores
- Entity Resolution
  which is also called Entity Recognition
- Examples using OpenRefine
  of data cleaning in OpenRefine
  which is open source, spreadsheet-like tool

READ the files marked READ.

For these topics, it is very worthwhile to also read/watch the supplementary material.

Go back to the video of the POI system development several times
to see each step of Data Wrangling in action.
This includes Data Cleaning.

Chapter 7 of the `pandas` book is on Data Wrangling

Chapter 5 of the `pandas` book has a section on *Handling Missing Values*

and Chapter 9 covers more advanced features for *Data Aggregation and Group Operations*

All the best, Greg.