COMP 499 Introduction to Data Analytics

Lecture 6 — Data Cleaning

Greg Butler

Data Science Research Centre

and

Centre for Structural and Functional Genomics

and

Computer Science and Software Engineering
Concordia University, Montreal, Canada

gregb@cs.concordia.ca

Overview of Lecture

1. Data Cleaning
2. Data Compatibility
3. Missing Data
4. Outliers
5. Remove Duplicates
6. Consistency of Data

Data Cleaning

Data Cleaning

detecting and correcting corrupt or inaccurate records

Error

is information lost in data acquisition

Artifact

systematic problem arising from data processing

Sniff Test

look closely to see if something may be wrong

Data Cleaning Process

Iterative

Keep raw data from data acquisition

You need a reference data set
you will need to re-process it many times

Script your cleaning steps

to refine and re-run your process

Data Compatibility

Comparing apples to apples

- unit conversions on numbers
- character code representations
- name unification
- time unification
- date unification
- financial unification

... And normalize to same scale

Z-scores

Missing Data

Delete observations or columns

that have missing values

Imputation — assign a value by inference

- fixed value, eg zero
- mean value (of column values)
- random value
  from randon observations
- by interpolation
  from similar observations
  by linear regression
  or machine learning

Outliers

Duplicates

Remove duplicates

Duplicates arise due to merging multiple data sets

Consistency of Data

Cluster/sort data values

To bring together
duplicate and similar data values
to make it easy to see differences/errors
(See OpenRefine video 1 of 3)

Cluster observations

To bring together
duplicate and similar observations
to make it easy to see differences/errors

Check for consistency

Differences need to be investigated