Introduction to Data Science
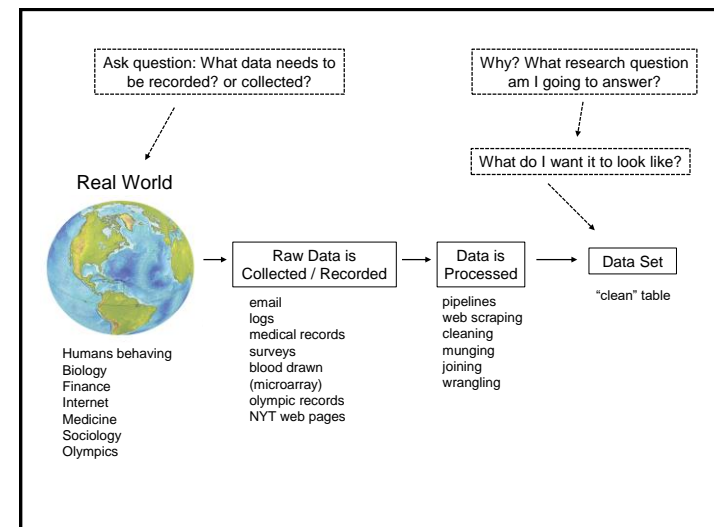
CS 5665
Utah State University
Department of Computer Science
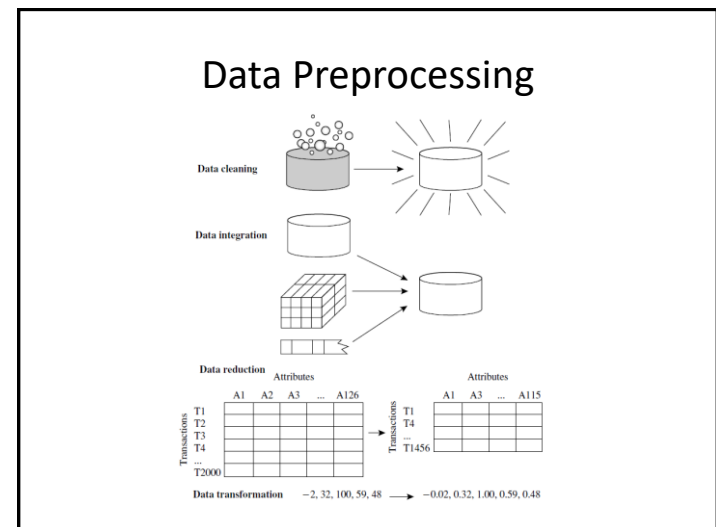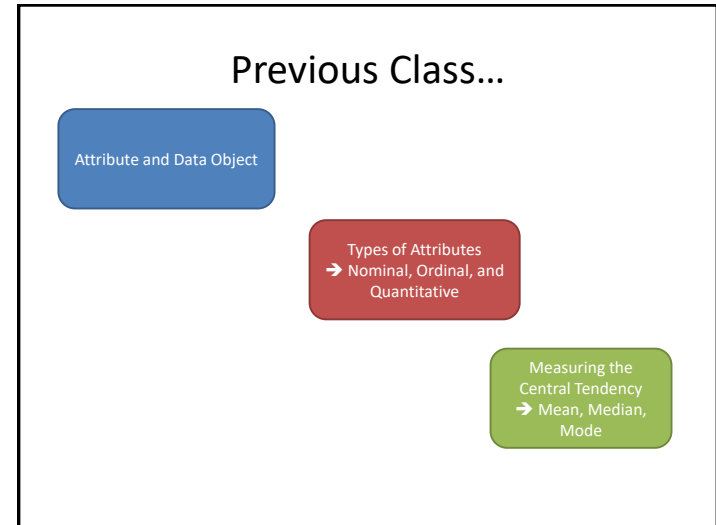Instructor: Prof. Kyumin Lee

---

Midterm (Oct. 27)

- The midterm exam is closed book.

- You may bring one standard 8.5" by 11" piece of paper with any notes you think appropriate or significant (front and back).

- No electronic devices allowed.

Previous Class…

Data Science Process

Slide 1

Possibly have to go
back and recollect

Can this be generalized
across domains?

Exploratory data
analysis

Clean
data set

Build data product

Machine learning
Statistical models

Visualize, communicate, report
findings

regression
classification
prediction
kNN
description
naive bayes
k-means

Make decisions

Slide 2

Previous Class…

Attribute and Data Object

Types of Attributes
➔ Nominal, Ordinal, and
Quantitative

Measuring the
Central Tendency
➔ Mean, Median,
Mode

Slide 3

Previous Class…

Measuring Data Similarity
and Dissimilarity
➔ Nominal (Binary),
Ordinal and Numeric
(Quantitative)

Measuring Distance of
Numeric Data
➔ Manhattan Distance
and Euclidean Distance

Measuring Similarity
of documents (texts)
➔Cosine Similarity

Slide 4

Data Preprocessing

Data cleaning

Data integration

Data reduction

Attributes
A1 A2 A3 ... A126

Transactions
T1
T2
T3
T4
...
T2000

Attributes
A1 A3 ... A115

Transactions
T1
T4
...
T1456

Data transformation   −2, 32, 100, 59, 48 ⟶ −0.02, 0.32, 1.00, 0.59, 0.48

Data Preprocessing

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases/data sources, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation

Previous Class…

Correlation Analysis
➔Correlation Coefficient
➔Covariance

Classification: Decision Tree

GINI Index

Previous Class…

Underfitting and Overfitting in Classificaiton

Evaluation
➔Accuracy
➔Cost Matrix

Previous Class…

MapReduce

Two Main Functions
➔ Map and Reduce

Google Glass

- Read https://en.wikipedia.org/wiki/Google_Glass
- Understand What Google Glass is