# Machine Learning

## Lecture 5: Data Quality

COURSE CODE: CSE451

2023

# Course Teacher

**Dr. Mrinal Kanti Baowaly**

Associate Professor

Department of Computer Science and Engineering, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Bangladesh.

Email: mkbaowaly@gmail.com

# Data Quality

- Data quality is a perception or an assessment of data's fitness to serve its purpose in a given context
- Components of data quality


Source: Link1

# #1: Completeness

- Completeness is defined as expected comprehensiveness.
- Data can be complete even if optional data is missing. As long as the data meets the expectations then the data is considered complete.
- For example, a customer's first name and last name are mandatory but middle name is optional; so a record can be considered complete even if a middle name is not available.

# #2: Consistency

- Consistency means data across all systems reflects the same information and are in synchronized with each other across the enterprise.

- Examples of some inconsistencies:
  - A business unit status is closed but there are sales for that business unit.
  - Employee status is terminated but pay status is active.

# #3: Conformity

- Conformity means the data is following the set of standard data definitions like data type, size and format.

- For example, date of birth of customer is in the format “mm/dd/yyyy”

# #4: Accuracy

- Accuracy is the degree to which data correctly reflects the real world object or an event being described.

- Examples:
  - Sales of the business unit are the real value.
  - Address of an employee in the employee database is the real address.

# #5: Integrity

- Integrity means validity of data across the relationships and ensures that all data in a database can be traced and connected to other data.

- For example, in a customer database, there should be a valid customer, address and relationship between them. If there is an address relationship data without a customer then that data is not valid and is considered an orphaned record.

# #6: Timeliness

- Timeliness references whether information is available when it is expected and needed.
- The data should be recorded as soon as possible after the real-world event because, with the passage of time, statistics become less useful and less accurate.
- Examples:
  - Companies that are required to publish their quarterly results within a given frame of time
  - Customer service providing up-to date information to the customers
  - Credit system checking in real-time on the credit card account activity

# Data Quality Problems

What kinds of data quality problems?

How can we detect problems with the data?

What can we do about these problems?

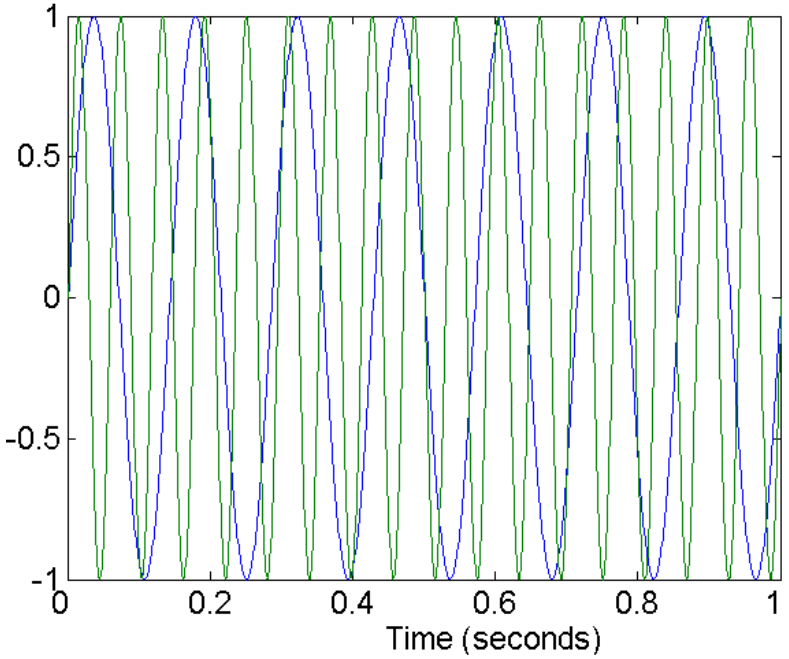
**Examples of data quality problems:**

- Noise
- Outliers
- Missing values
- Duplicate or Redundant data

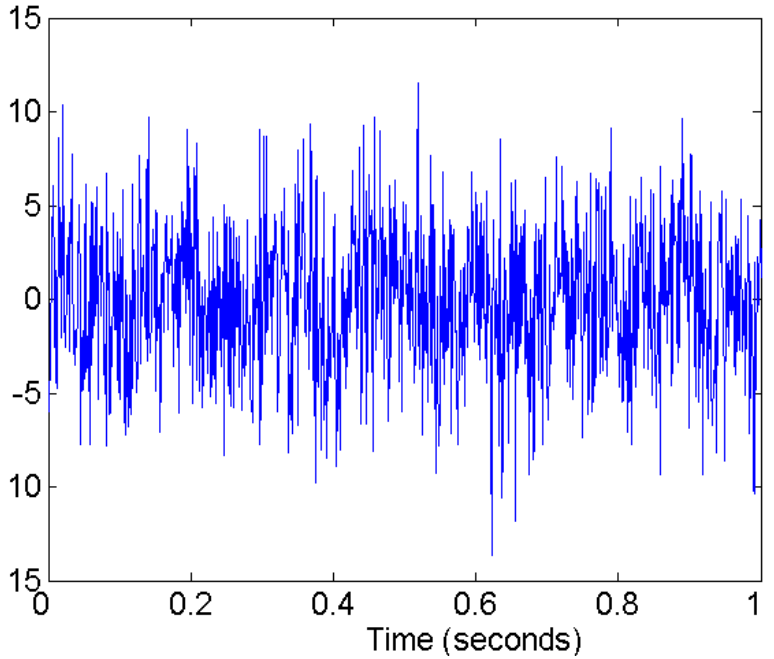
# Noise

## Noise refers to modification of original values.

- Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen




Two Sine Waves




Two Sine Waves + Noise

# Noisy Data

- Noisy data (or corrupt data) are meaningless information
- It cannot be understood and interpreted correctly by machines
- It unnecessarily increases the amount of storage space required and can adversely affect any data mining analysis results.
- Noisy data can be caused by faulty data collection instruments, human or computer errors occurring at data entry, data transmission errors, limited buffer size for coordinating synchronized data transfer, inconsistencies in naming conventions or data codes used and inconsistent formats for input fields( e.g. date).

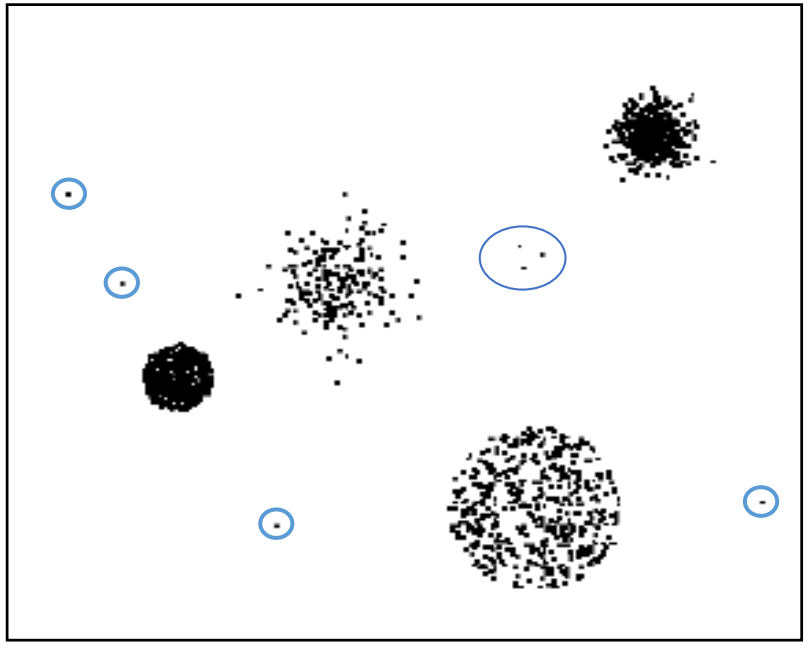
# How to Handle Noisy Data

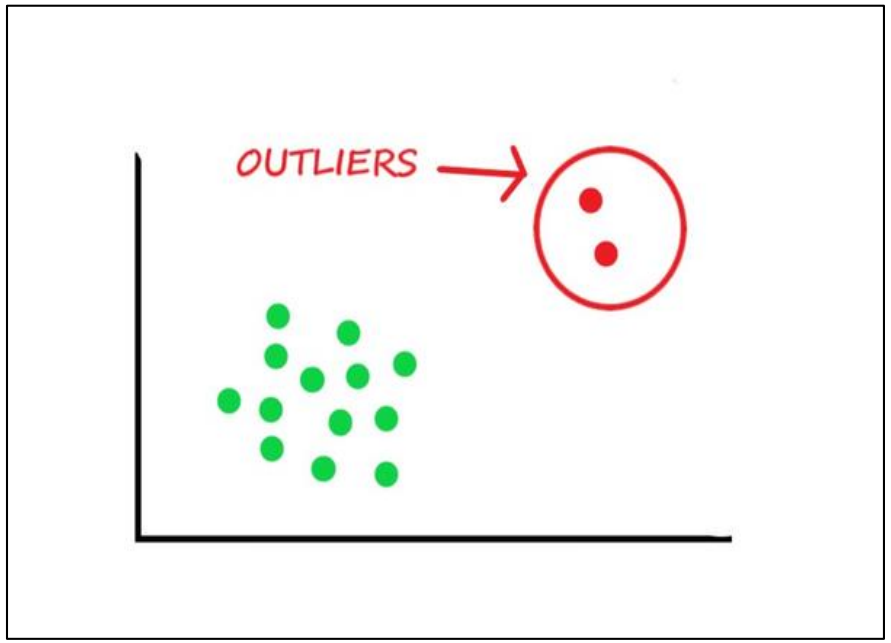
- Remove noise from data (called data smoothing) using binning method, regression, clustering
- Collect more data, it's the best way to cut the noise out but data is expensive
- Use Principal Component Analysis (PCA) for dimensionality reduction
- Use regularization and cross validation (CV) to prevent overfitting

Detail: Link

# Outliers

Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set






**How to detect outliers:** use various visualization methods, like Box-plot, Histogram, Scatter Plot. Link1, Link2

# How to Handle Outliers

- **Drop the outlier records:** Sometimes it's best to completely remove those records from your dataset to stop them from skewing your analysis.
- **Cap your outliers' data:** Another way to handle true outliers is to cap them. For example, if you're using income, you might find that people above a certain income level behave in the same way as those with a lower income. In this case, you can cap the income value at a level that keeps that intact.
- **Assign a new value:** If an outlier seems to be due to a mistake in your data, try imputing a new value. Common imputation methods include using the mean of a variable or utilizing a regression model to predict the missing value.
- **Try a transformation:** A different approach to true outliers could be to try creating a transformation of the data rather than using the data itself.

# Missing Values

## Reasons for missing values

- Information is not collected
  (e.g., people decline to give their age and weight)
- Attributes may not be applicable to all cases
  (e.g., annual income is not applicable to children)

## Handling missing values

- Eliminate Data Objects
- Estimate Missing Values (Mean/ Mode/ Median /Prediction etc.)
- Ignore the Missing Value During Analysis
- Replace with all possible values (weighted by their probabilities)

# Dealing with duplicate data

- You should probably remove duplicate data.
- Duplicate data will essentially lead to bias your fitted model or do the model overfitting.
- **But you should**
  1) be sure they are not real data that coincidentally have values that are identical
  2) try to figure why you have duplicates in your data. For example, sometimes people intentionally ‘oversample’ rare categories in training data

# HW: Data Cleaning with Python and Pandas and NumPy

*According to IBM Data Analytics you can expect to spend up to 80% of your time cleaning data.*

**Practice:**

Data Preprocessing | Data Cleaning Python

Data Cleaning In Python Basics Using Pandas

Pythonic Data Cleaning With Pandas and NumPy

# End of Lecture-5