

Machine Learning

Lecture 4-5: DATA

COURSE CODE: CSE451

2021



Course Teacher

Dr. Mrinal Kanti Baowaly

Assistant Professor

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

Email: baowaly@gmail.com



DATA

- Data can be any unprocessed fact, value, text, sound, picture or video that is not being interpreted and analyzed
- Data is the most important part of all Data Analytics, Machine Learning, Artificial Intelligence
- Without data, we can't train any model and all modern research and automation will go vain
- Big Enterprises are spending loads of money just to gather as much certain data as possible
- Example: Facebook acquires WhatsApp by paying a huge price of \$19 billion

Information and Knowledge

- Information: Data that has been interpreted and manipulated and has now some meaningful inference for the users
- Knowledge: Combination of inferred information, experiences, learning and insights
- Machine learning is a tool for turning information into knowledge



Data Types From A Machine Learning Perspective

- Numerical data: any data where data points are exact numbers
 - Continuous data
 - Discrete data
- Categorical data: two or more groups, don't have mathematical meaning
 - Ordinal data: ordered or ranked categories
 - Nominal data: categories with no rank or order between them
 - Binary data: nominal data with exactly two categories
- Time series data: a sequence of numbers collected at regular intervals over some period of time
- Text data: words

Source: [MachineLearning-blog](#)

Discrete and Continuous Attributes

Discrete Attribute

- Can be counted
- Has only a finite or countably infinite set of values
- Examples: the number of students in a class, the number of words in a document, the number of heads in 100 coin flips
- Often represented as integer variables.

Continuous Attribute

- Can only be measured
- Has any value (real number) within a range
- Examples: temperature, height, or weight.
- represented as real or floating-point variables.

Quantitative data vs Qualitative data

- Quantitative data: can be measured, e.g. distance, area, time, speed, volume, weight, temperature, cost, etc.
- Qualitative data: described in linguistic terms
 - Data can be observed but not measured
 - Description typically includes a clear subjective and/or contextual aspect
 - Long texts can also be considered to be qualitative data

More from [here](#).

What is Data set?

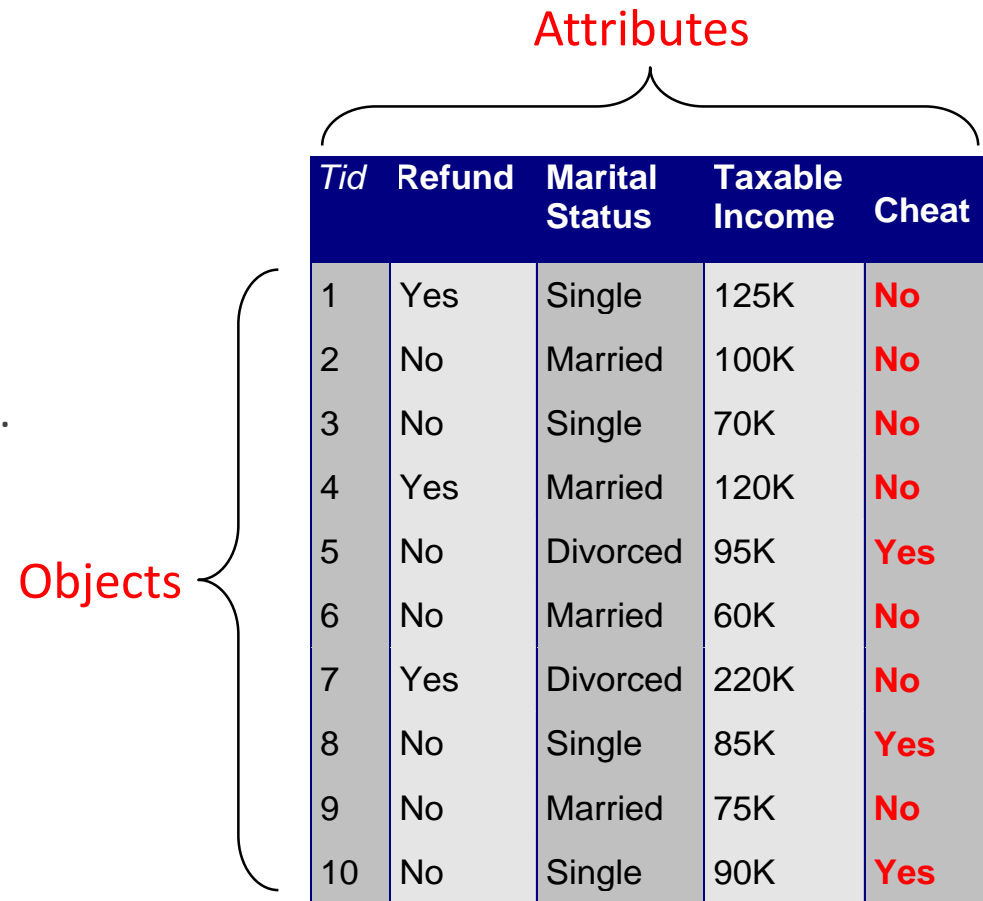
Collection of data objects and their attributes

An attribute is a property or characteristic of an object

- Examples: eye color of a person, temperature, etc.
- Attribute is also known as variable, field, characteristic, or feature

A collection of attributes describe an object

- Object is also known as record, point, case, sample, entity, or instance



Attributes				
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Types of Data sets

1. Record

- Data Matrix
- Document Data
- Transaction Data

2. Graph

- World Wide Web
- Molecular Structures

3. Ordered

- Spatial Data
- Temporal Data
- Sequential Data
- Genetic Sequence Data

1. Record Data

Data that consists of a collection of records, each of which consists of a fixed set of attributes

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data Matrix

If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data

Each document becomes a 'term' vector,

- each term is a component (attribute) of the vector,
- the value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Transaction Data

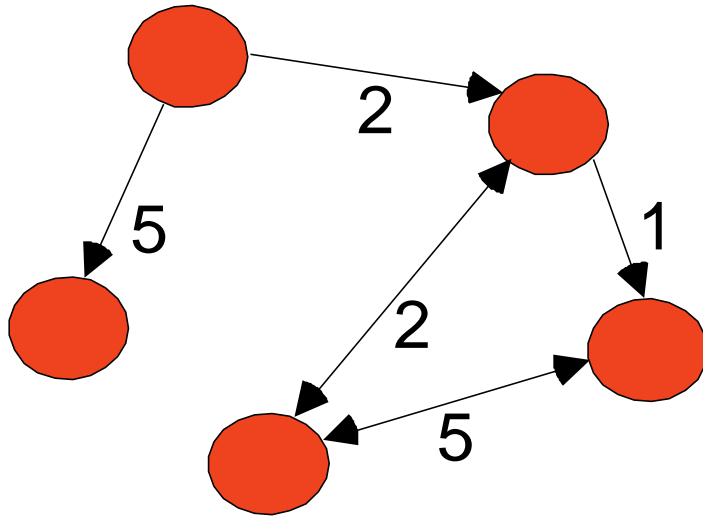
A special type of record data, where

- each record (transaction) involves a set of items.
- For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

2. Graph Data

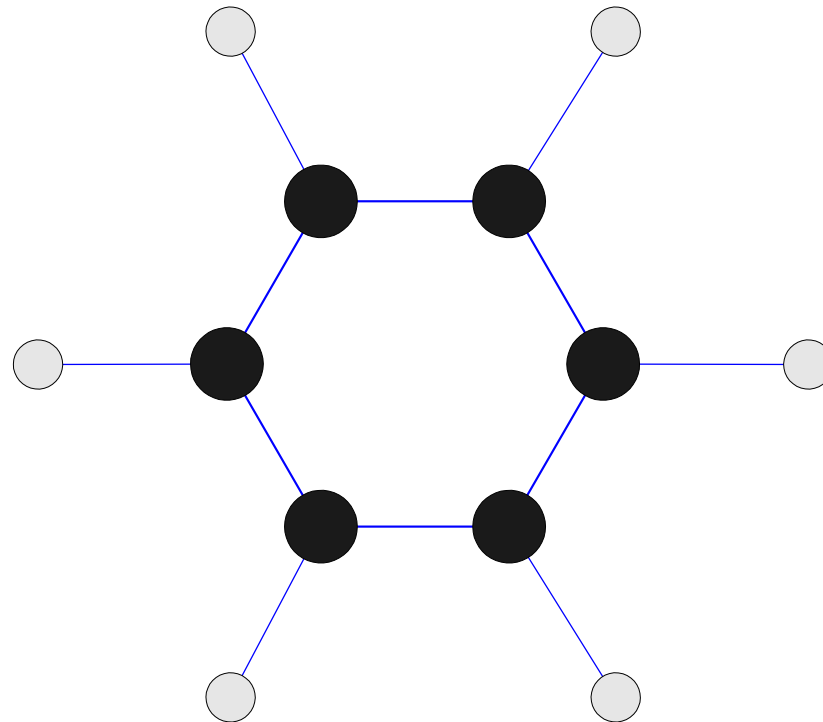
Examples: Generic graph and HTML Links



```
<a href="papers/papers.html#bbbb">  
Data Mining </a>  
<li>  
<a href="papers/papers.html#aaaa">  
Graph Partitioning </a>  
<li>  
<a href="papers/papers.html#aaaa">  
Parallel Solution of Sparse Linear System of Equations </a>  
<li>  
<a href="papers/papers.html#ffff">  
N-Body Computation and Dense Linear System Solvers
```

Chemical Data

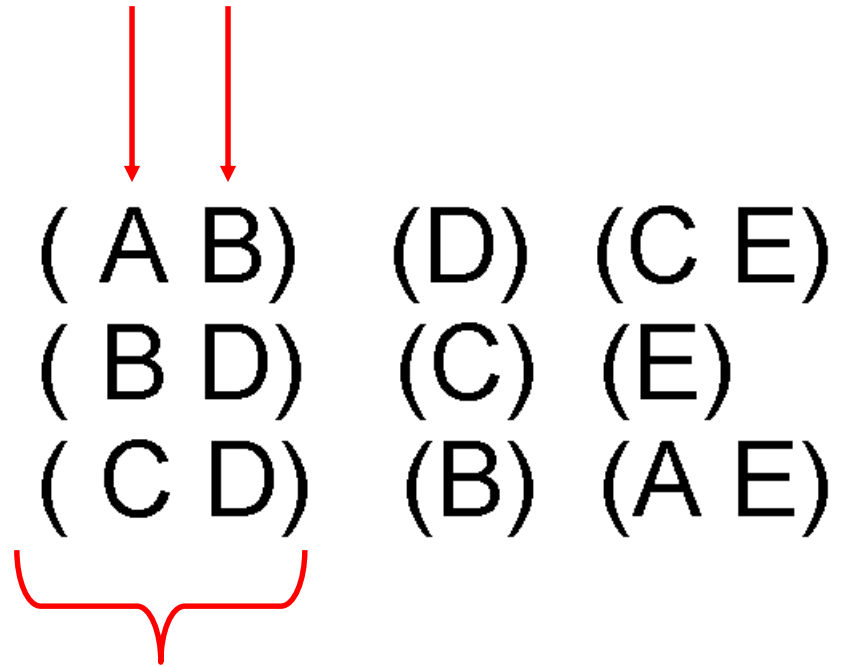
Benzene Molecule: C₆H₆



3. Ordered Data

Sequences of transactions

Items/Events



An element of the
sequence

3. Ordered Data (Cont.)

Genomic sequence data

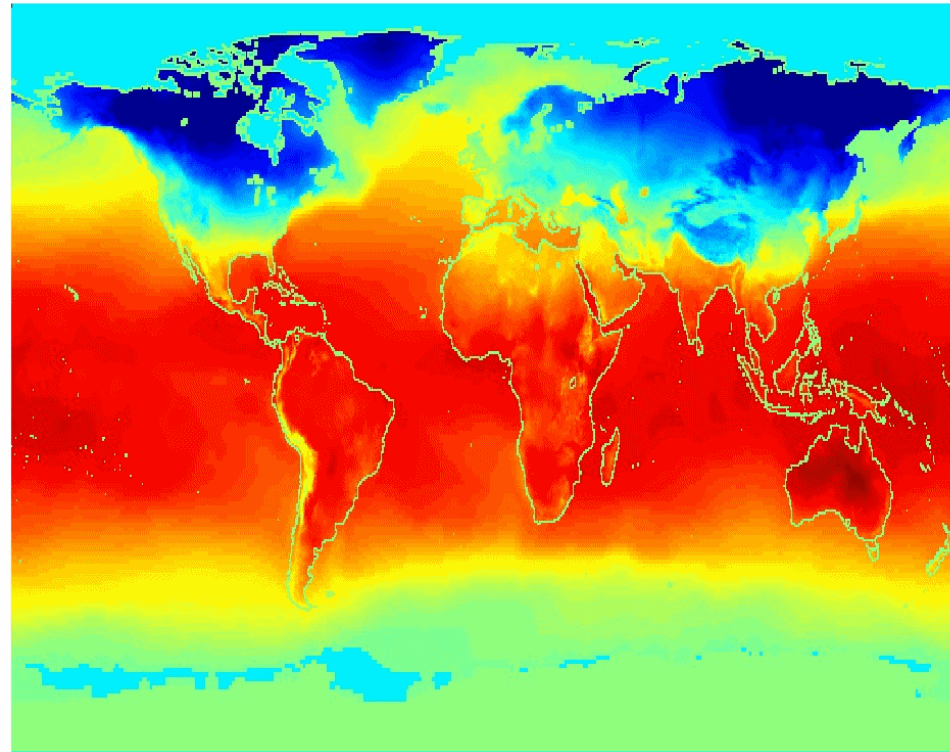
```
GGTTCCGCCTTCAGCCCCGCGCC  
CGCAGGGCCCGCCCCGCGCCGTC  
GAGAAGGGCCCGCCTGGCGGGCG  
GGGGGAGGCGGGGCCGCCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG
```

3. Ordered Data (Cont.)

Spatio-Temporal Data

Average Monthly
Temperature of
land and ocean

Jan



Test Your Understanding

- Take part in the following Quiz Test on Types of Data
- Click [here](#)



How to get datasets for Machine Learning

- Popular sources for Machine Learning datasets
 - [Kaggle Datasets](#)
 - [UCI Machine Learning Repository](#)
 - [Datasets via AWS](#)
 - [Google's Dataset Search Engine](#)
 - [Microsoft Datasets](#)
 - Government Datasets
 - [Computer Vision Datasets](#)
 - [Scikit-learn dataset](#)

Source: [JavaTPoint](#)

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?

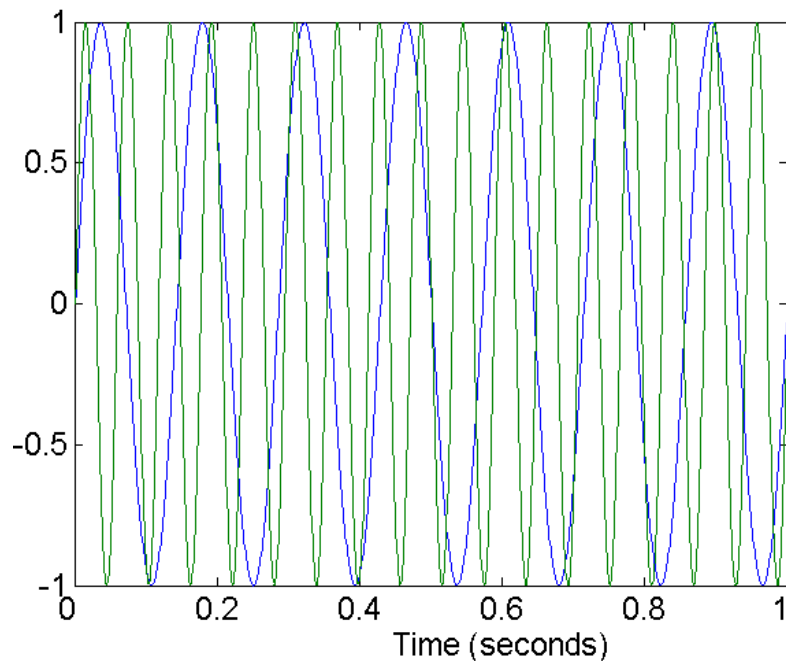
Examples of data quality problems:

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

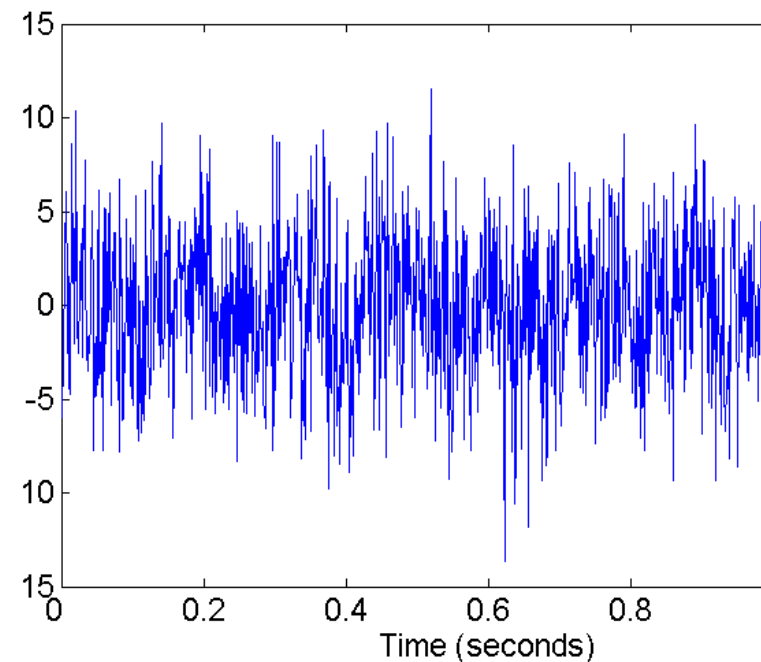
Noise

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

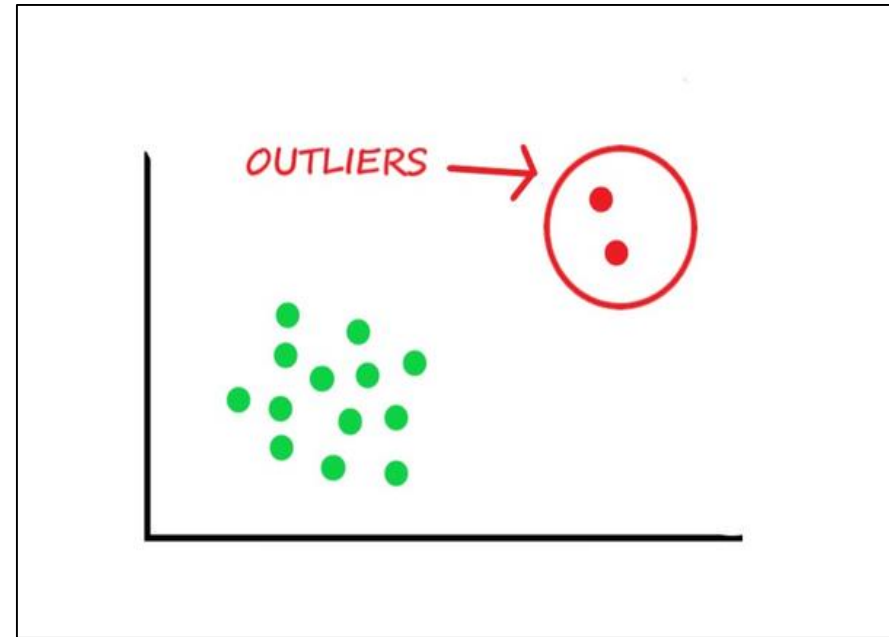
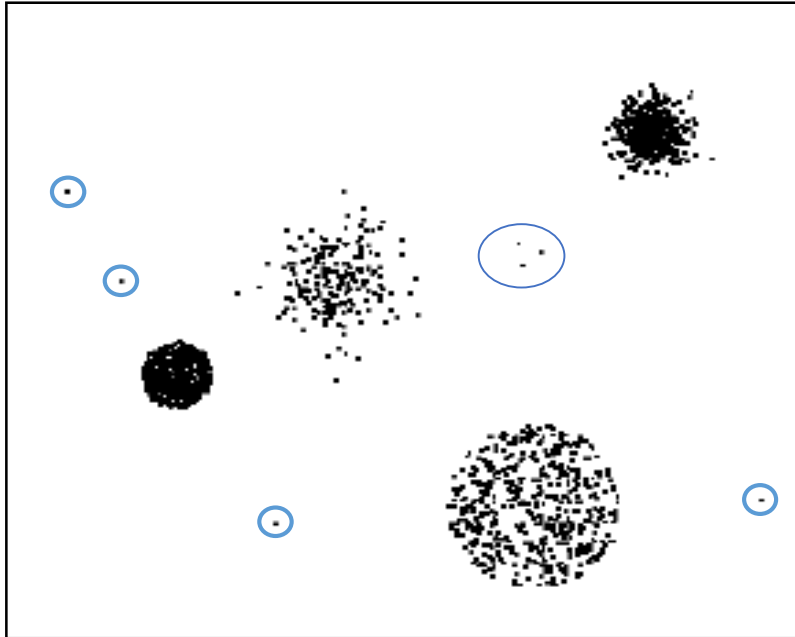
How to Handle Noisy Data

- 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.

Source: [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.

Source: [DataScience Foundation](#)

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

Lab Work: 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 Challenge: Handling missing values](#)

[Data Cleaning In Python Basics Using Pandas](#)

[Data Cleaning with Python and Pandas: Detecting Missing Values](#)