Data Curation Techniques

Siddharth R

Syllabus

Unit 1: Introduction to Data Lifecycle
The data lifecycle (creation, storage, access, preservation) - Data objects and attribute types - Basic statistical descriptions of data - Measuring data similarity and dissimilarity - Database vs data warehouse - Data Quality: Why Preprocess the Data?

Unit 2: Data Preprocessing Techniques
Data cleaning workflow - Handling missing values, noisy data, outlier - Data integration : redundancy and correlation, duplication - Data transformation : normalization - data discretization using Binning, histogram, clusters - concept hierarchy generation for nominal data. Hands-on activity: Data preprocessing using Python libraries (Numpy, Pandas)

Unit 3: Data Reduction
Attribute subset selection method : forward selection, backward elimination, equal width and equal frequency histogram - sampling with and without replacement - data aggregation and summarization - overview of data cube - Basics of feature selection and feature extraction. Hands on: Data reduction using python (scikit-learn)

Unit 4: Data Management
ETL vs ELT - Data governance - data modeling and design - data integration and interoperability - Challenges of working with heterogeneous data sources - Common data formats (CSV, JSON, XML)- master data management - metadata - Discussion: Integration challenges in IoT data

Unit 5: Data Architecture
Principles of good data architecture - Architecture concepts : tight vs loose coupling - user access : single vs multi tenant - event driven architecture - data storage systems - storage abstraction - Hot, warm, and cold data - Discussion : Latest trends in storage using open source tools.

Text / Reference Books

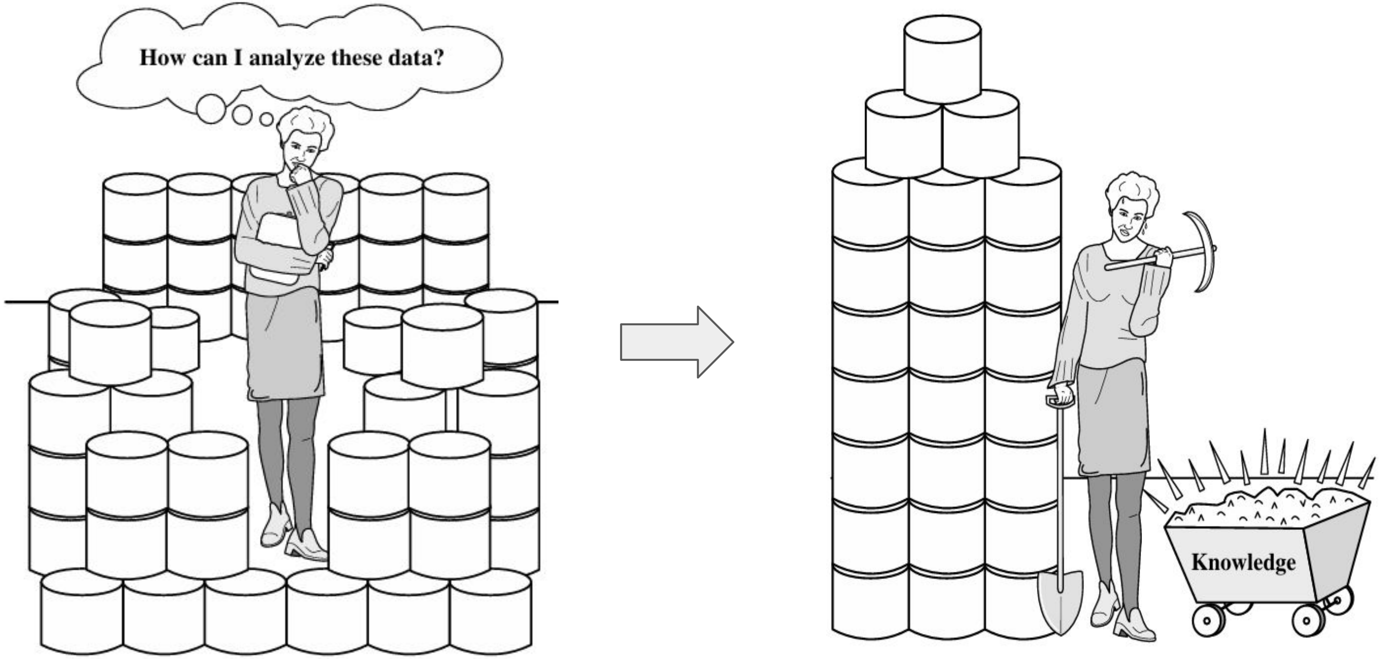
1. *"Data Mining: Concepts and Techniques"* by Jiawei Han, Jian Pei, Hanghang Tong , Morgan Kaufmann, 2022, ISBN: 9780128117613 (For Unit 1 to 3)
2. "Fundamentals of Data Engineering" by Joe Reis and Matt Housley, O'Reilly Media, 2022, ISBN: ISBN: 9781098108304 (for Unit 4 and 5)
3. Latest related research articles from reputed journal/conferences

Assessment components

Quiz - 1	: 5%
Mid - Term	: 25%
Quiz - 2	: 5 %
Lab Components	: 25%
End Sem	: 40%

Why Data Curation?

- The world is data rich but information poor.



Why Data Curation?

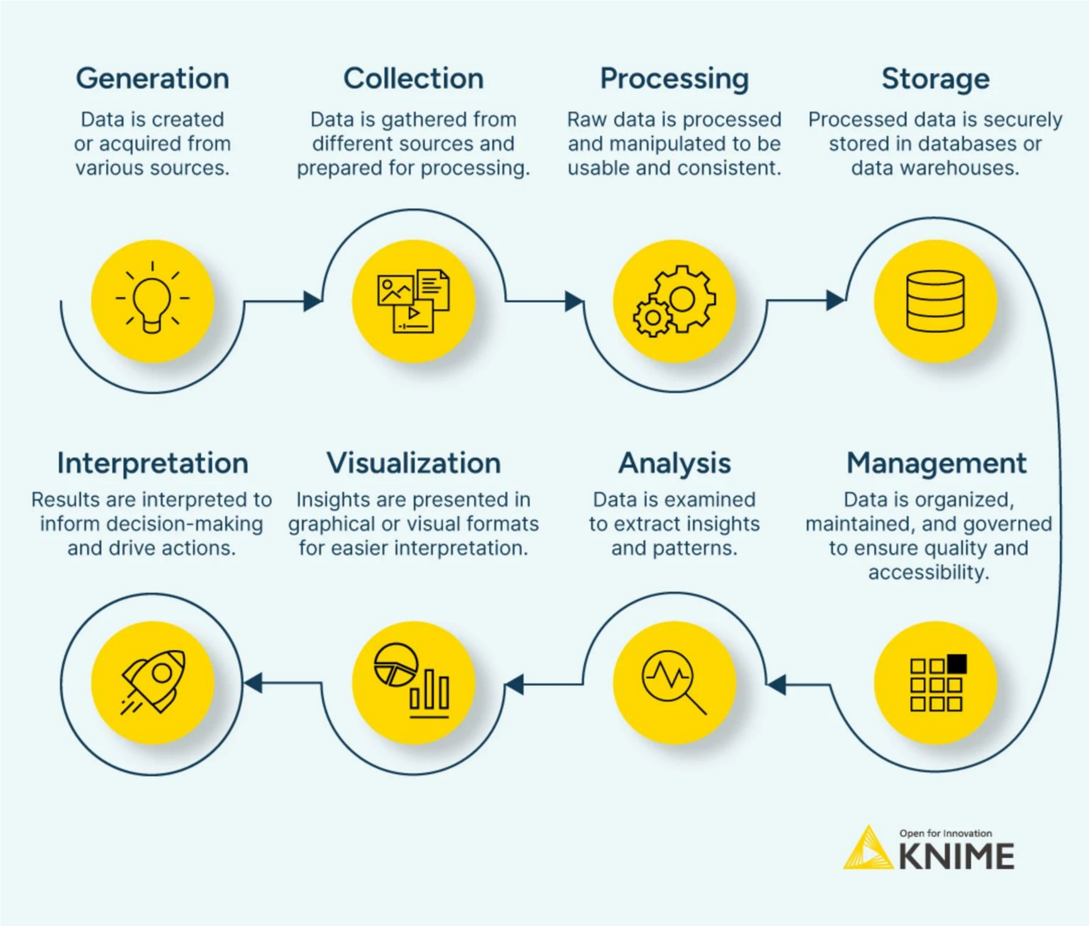
Key steps:

- Data Cleaning
- Data Integration
- Data Selection
- Data Transformation

↓

Data Analysis and Visualization

Data Life Cycle



Generation

Data is created or acquired from various sources.

Collection

Data is gathered from different sources and prepared for processing.

Processing

Raw data is processed and manipulated to be usable and consistent.

Storage

Processed data is securely stored in databases or data warehouses.

Management

Data is organized, maintained, and governed to ensure quality and accessibility.

Analysis

Data is examined to extract insights and patterns.

Visualization

Insights are presented in graphical or visual formats for easier interpretation.

Interpretation

Results are interpreted to inform decision-making and drive actions.

Source: https://www.knime.com/blog/the-data-lifecycle

Data Generation vs Data Collection

- Either active or passive
- Common sources:
  - Human-generated (e.g., surveys, forms)
  - Machine-generated (e.g., IoT sensors, logs)
  - Transactional systems (e.g., purchases, banking records)
  - Web scraping

What is Web Scraping (Common Data Collection)

- Extraction of data from website
- How the website content is presented?
- Is it structured ?
- How you are going to save the extracted content ?
- Can you name few common applications??
  - Sentiment Analysis?

Possible Ways to do Web Scraping

- Using Libraries in Programming Languages
- Browser Automation Tools
- APIs for Data Retrieval
- Headless Browsers
- No-Code or Low-Code Tools

Reading Assignment

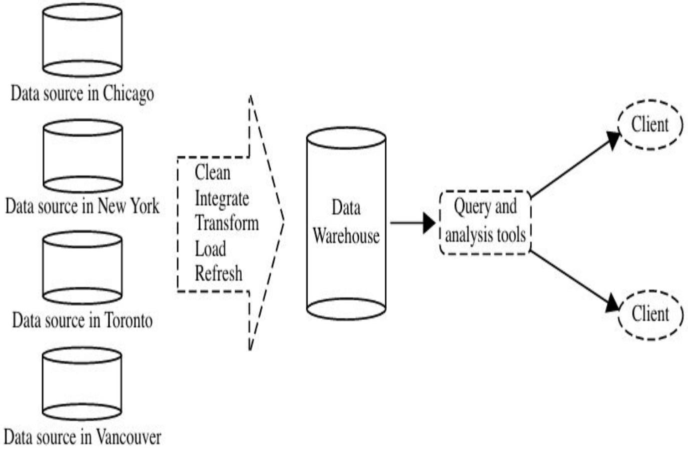
Headless Browser - Puppeteer and Playwright

Key Applications :

1. E-commerce Price tracking
2. Social media monitoring

Common Data Sources

- Database data
    - A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows).
    - Model : Entity - Relationship Model
- Data Warehouse
    - A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site.
    - Model : Data cube



Data Objects

- Data sets are made up of data objects.
- A data object represents an entity
    - in a sales data, the objects may be customers, store items, and sales
    - In a university data ???
- Data objects can also be referred to as samples, examples, instances, data points, or objects.
- If the data objects are stored in a database, they are data tuples.
- Data objects are typically described by attributes.
- That is, the rows of a database correspond to the data objects, and the columns correspond to the attributes.
- The attribute, dimension, feature, and variable are often used interchangeably

Types of data attributes

- Nominal attribute:
  - Relating to name
  - Categorical
  - Data without any inherent order Examples: Gender, color, city
  - Is your aadhar number nominal?
  - How do you measure the central tendency ?

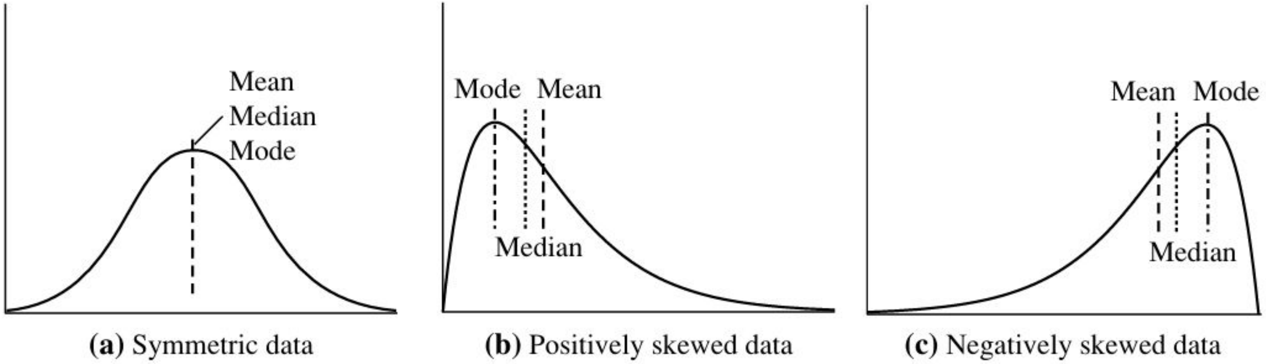
- Binary attribute:
  - Nominal attribute with only categories
  - Also called as Boolean
  - Give some examples ??
  - Symmetric vs Asymmetric binary attribute

Types of data attributes

- Ordinal attribute
    - Data with a specific order but without equal intervals between categories.
    - Examples: Education level, product rating (low, medium, high)
    - What is the preferred measure of central tendency?
- Whether the nominal, binary and ordinal data are quantitative or qualitative ?
- Numerical attribute:
    - Interval scaled attributes : No true zero point Example: Temperature in celsius / fahrenheit
    - Ratio scaled attributes : true zero point Example: Temperature in Kelvin
    - Continuous data: Data with infinite possible values within a given range. Examples: Height, weight, temperature
    - Discrete data: Data with a finite number of values. Examples: Number of children, number of products sold

Statistical Description of Data

- To have overall picture of your data
- Central tendency - measure the location of the middle or center of a data distribution.
- Dispersion of data - how are the data spread out
- Common approaches for Measuring the Central Tendency: Mean, Median, and Mode



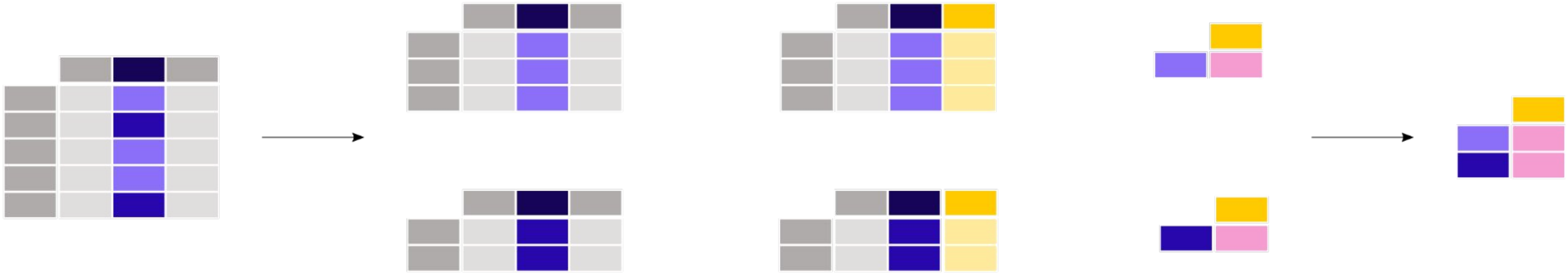
(a) Symmetric data (b) Positively skewed data (c) Negatively skewed data

Summary Statistics using Pandas [Refer Pandas code]

Key points to remember:

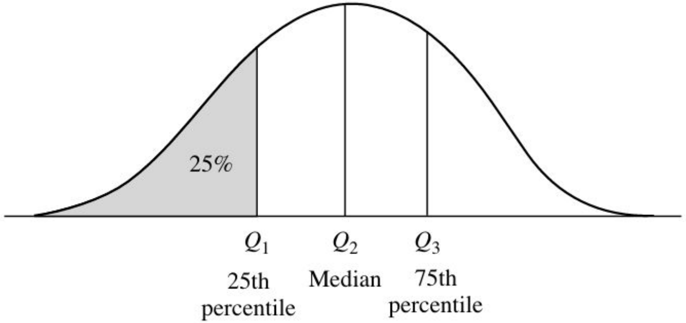
- Whereas size includes NaN values and count excludes the missing values.
- value_counts is a convenient shortcut to count the number of entries in each category of a variable.
- Split-Apply-Combine

```
titanic      .groupby("Sex")     ["Age"]      .mean()
```



Measuring the Dispersion of Data

- Range : (Maximum - minimum)
- Standard Deviation : It provides a measure of the average distance between each data point and the mean
- Variance : The average of the squared differences between each data point and the mean.
- Quartiles: Quartiles divide the dataset into four equal parts when it is sorted.
  - Q1 (1st quartile): Median of the lower half (25th percentile).
  - Q2 (2nd quartile): Median of the dataset (50th percentile).
  - Q3 (3rd quartile): Median of the upper half (75th percentile).
- Interquartile Range (IQR) : The distance between the first and third quartiles , IQR = Q3 - Q1



Measuring Data Similarity / Dissimilarity

- Why we need to measure the similarity / dissimilarity
- Also called as measure of proximity
- A similarity measure for two objects, i and j, will typically return the value 0 if the objects are unalike.
- The higher the similarity value, the greater the similarity between objects. (Typically,a value of 1 indicates complete similarity,that is,the objects are identical.)
- A Dissimilarity Measure works the opposite way.
- Data matrix vs Dissimilarity matrix

Data Matrix vs Dissimilarity Matrix

Data Matrix :This structure stores the n data objects in the form of a relational table, or n-by-p matrix (n objects X p attributes)

- Also called as “two-mode” matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}.$$

Dissimilarity Matrix: This structure stores a collection of proximities that are available for all pairs of n objects.

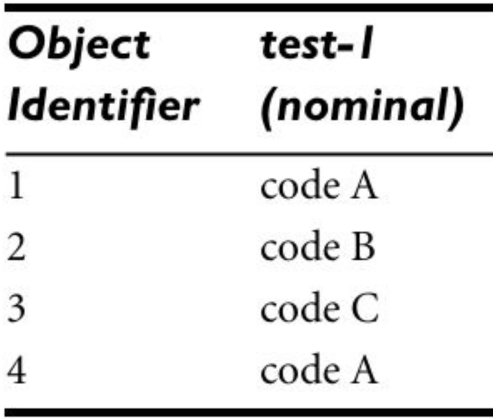
where d(i, j) is the measured dissimilarity or difference between obj and j.

sim(i, j) = 1 - d(i, j), where *sim* is the similarity

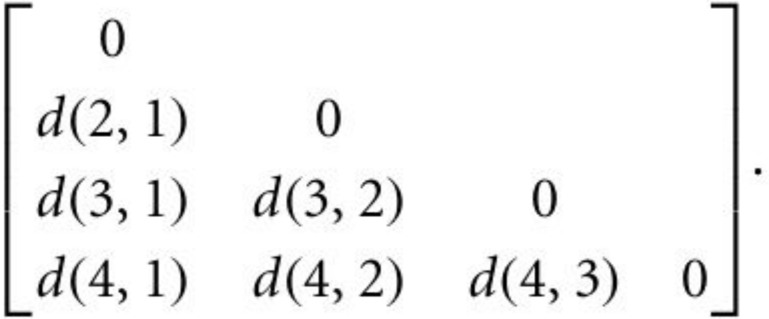
- Also called as one-mode matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix},$$

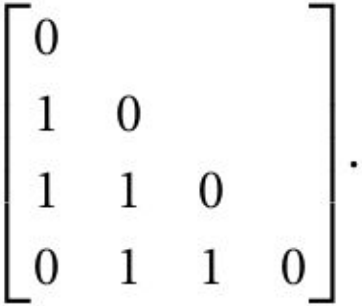
Measuring Dissimilarity - Nominal Attributes



Object Identifier	test-1 (nominal)
1	code A
2	code B
3	code C
4	code A



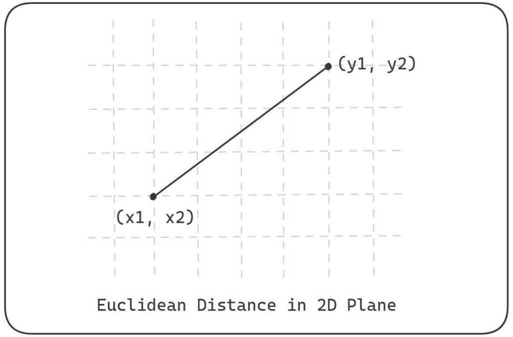
$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}.$$



$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

Euclidean Distance

- Measures the straight-line distance (as the crow flies).
- Diagonal movement allowed
- Euclidean distance is the shortest distance between any two points
- Mathematically, the Euclidean distance between the points x and y in two-dimensional plane is given by:



$$d(\mathbf{x},\mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

- Extending to n dimensions, the points x and y are of the form x = (x1, x2, …, xn) and y = (y1, y2, …, yn),

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

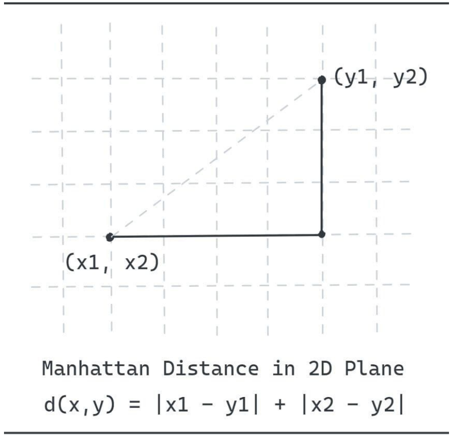
Manhattan Distance (city-block)

- Measures the sum of the absolute differences of the coordinates.
- In a 2-D plane, the Manhattan distance between the points x and y is given by:

$$d(\mathbf{x}, \mathbf{y}) = |x_1 - y_1| + |x_2 - y_2|$$

- In n-dimensional space, where each point has n coordinates, the Manhattan distance is given by:

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|$$



Manhattan Distance in 2D Plane

d(x,y) = |x1 − y1| + |x2 − y2|

Minkowski Distance

- Minkowski distance is a generalization of the Euclidean and Manhattan distances. It is defined as

$$d(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p} \quad for\ p \geq 1$$

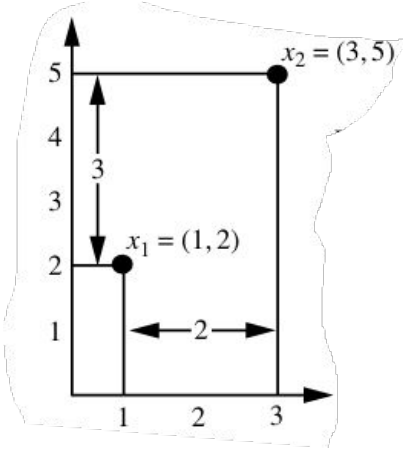
- If p=1, then Minkowski distance equation takes the same form as that of Manhattan distance (L1 - norm)
- Similarly, for p = 2, the Minkowski distance is equivalent to the Euclidean distance (L2-norm)

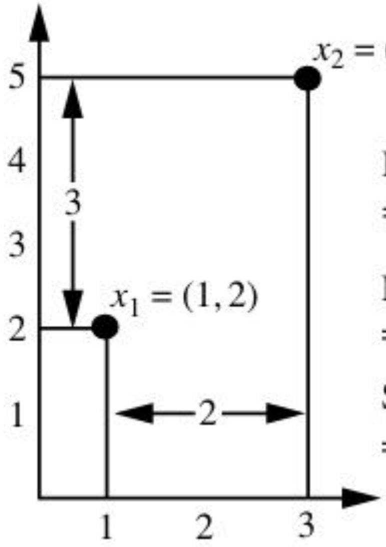
Chebyshev / Supremum distance

- It is a measure of distance that calculates the maximum difference along any coordinate dimension between two points in a multidimensional space.
- Also known as Lmax,L∞ norm or uniform norm

$$d_{\mathrm{Chebyshev}}(P, Q) = \max_i(|p_i - q_i|)$$

- Let x_1 = (1, 2) and x_2 = (3, 5) , find Euclidean, Manhattan and Supremum





Euclidean distance
$= (2^2 + 3^2)^{1/2} = 3.61$

Manhattan distance
$= 2 + 3 = 5$

Supremum distance
$= 5 - 2 = 3$

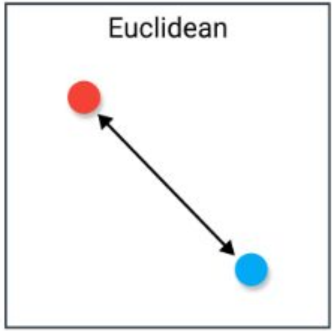
Mathematical Properties

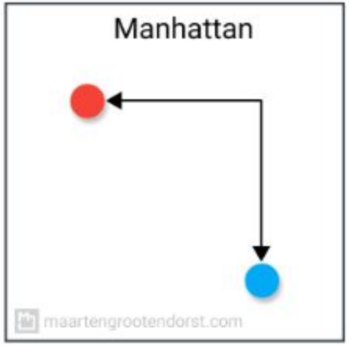
- Non-negativity : distance is always non-negative , $d(x, y) \geq 0$
- Identity of indiscernibles : distance between a point and itself is always 0. $d(x, y) = 0$ *if and only if* $x = y$
- Symmetry: The order of points doesn't matter in distance calculation. $d(x, y) = d(y, x)$
- Triangle inequality : The maximum difference between x and z along any dimension cannot be greater than the sum of the maximum differences from x to y and from y to z. $d(x, z) \leq d(x, y) + d(y, z)$

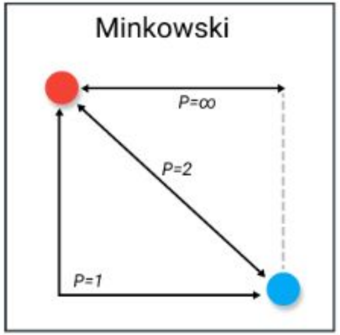
Jaccard Distance

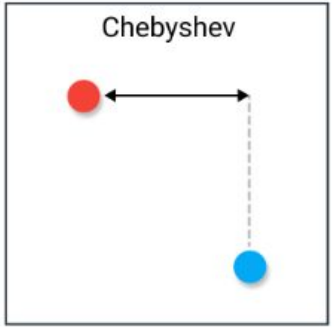
- ○ Quantify the dissimilarity between two sets of data.
- ○ Derived from the Jaccard Index (or Similarity Coefficient)
- ○ Jaccard Distance=1−Jaccard Index, where Jaccard Index=$|A\cap B|$ / $|A\cup B|$
- ○ $A\cap B$: The number of elements common to both sets A and B (intersection).
- ○ $A\cup B$: The number of unique elements in either set A or B (union).
- ○ The Jaccard Index ranges from 0 to 1, where 1 indicates identical sets and 0 indicated disjoint sets
- ○ Set A={1,2,3,4} and Set B={3,4,5,6} , $|A\cap B|$ =? , $|A\cup B|$ =? , Jaccard Index =?, Jaccard Distance =?
- ○ Use case: Compare documents or text similarity, sets of pixels in images for similarity

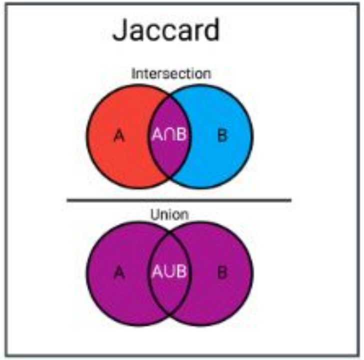
Overview


Euclidean


Manhattan
maartengrootendorst.com


Minkowski
P=∞
P=2
P=1


Chebyshev


Jaccard
Intersection
A
A∩B
B
Union
A
A∪B
B

Thank You !!!