Data Mining

Lecture 2

Data: Part 1

Mohammed Brahimi & Sami Belkacem

Outline

1. Data

2. Data preprocessing

3. Similarity measures

1- Data

What is Data ?

- **Data set**: collection of objects and their attributes
- **Attribute**: property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
- **Attribute** is also known as variable, field, characteristic, dimension, or feature
- **Object**: collection of attributes
- **Object** is also known as record, point, case, sample, entity, or instance

Attributes

Objects

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 Attributes

- **Nominal** (Categories)
  - Examples: ID numbers, eye color, zip codes
- **Ordinal** (Ordered Categories)
  - Examples: Rankings (e.g., taste of potato chips on a scale from 1-10), grades, height (tall, medium, short)
- **Interval** (Equal Intervals, No True Zero)
  - Examples: Calendar dates, temperatures in Celsius or Fahrenheit
- **Ratio** (Equal Intervals, True Zero)
  - Examples: Temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

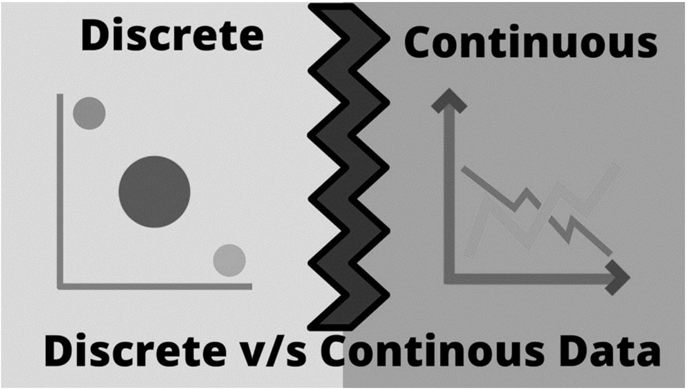
Properties of Attribute Values

- **Nominal**
  - Distinctness (=, ≠)
- **Ordinal**
  - Distinctness (=, ≠)
  - Order (< >)
- **Interval**
  - Distinctness (=, ≠)
  - Order (< >)
  - Meaningful Differences (+, -)
- **Ratio**
  - Distinctness (=, ≠)
  - Order (<, >)
  - Meaningful Differences (+, -)
  - Meaningful Ratios (* ,/)

Discrete vs. Continuous Attributes

- **Discrete Attribute**: Values from a finite or countably infinite set.
  - **Examples:** Zip codes, counts, or words in documents.
- Represented as integers.
- **Note:** Binary attributes are a special case of discrete attributes.

- **Continuous Attribute:** Values are real numbers.
  - **Examples:** Temperature, height, weight.
- Real values, practically measured with finite digits
- Represented as floating-point variables.



Asymmetric Attributes

- In **asymmetric attributes**, only the presence (non-zero value) matters.
  - **Examples:** Words present in documents: Focus on words that appear.
  - Items present in customer transactions: Emphasize purchased items.

- ***Real-Life Scenario***:

In a grocery store encounter, would we say:

"Our purchases are similar because we didn't buy most of the same products?"

Important Characteristics of Data

- **Dimensionality (Number of Attributes)**
  - High-dimensional data presents unique challenges.
- **Sparsity**
  - Emphasizes the importance of presence over absence.
- **Resolution**
  - Patterns can vary based on the scale of measurement.
- **Size**
  - The type of analysis often depends on the data's size.
- **Distribution**
  - Considers centrality and dispersion in the data.

Types of Data Sets

- **Record Data:** records with fixed attributes
  - Relational records
  - Data matrix …
  - Transaction Data
- **Graphs and Networks**
  - Transportation network
  - Social or information networks…
  - Molecular Structures
- **Ordered (Sequence) Data**
  - Video: sequence of image
  - Genetic Sequence Data
  - Temporal sequence …
- **Spatial Data**
  - RGB Images
  - Satellite images

Person:

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rom

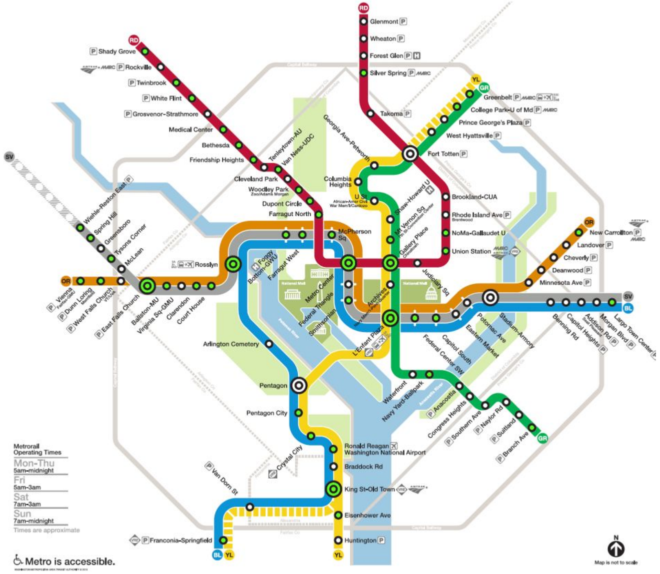
no relation

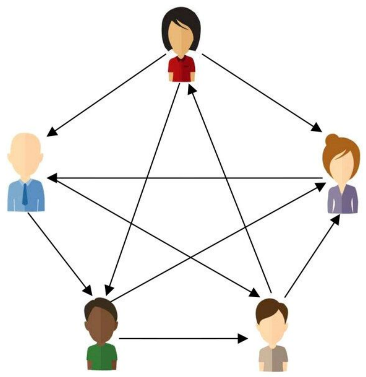
Car:

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

Types of Data Sets

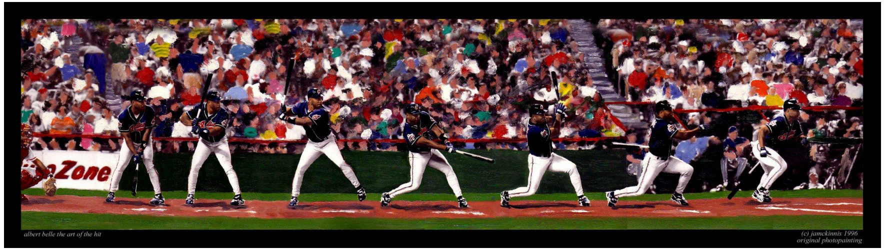
- **Record Data:** records with fixed attributes
  - Relational records
  - Data matrix …
  - Transaction Data
- **Graphs and Networks**
  - Transportation network
  - Social or information networks…
  - Molecular Structures
- **Ordered (Sequence) Data**
  - Video: sequence of image
  - Genetic Sequence Data
  - Temporal sequence …
- **Spatial Data**
  - RGB Images
  - Satellite images

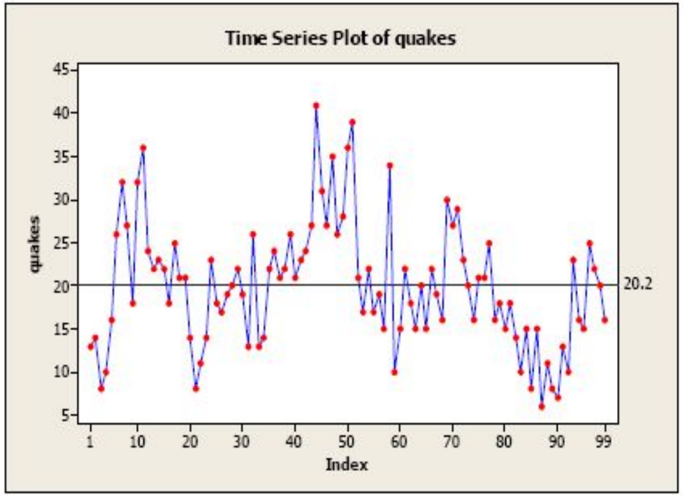




Types of Data Sets

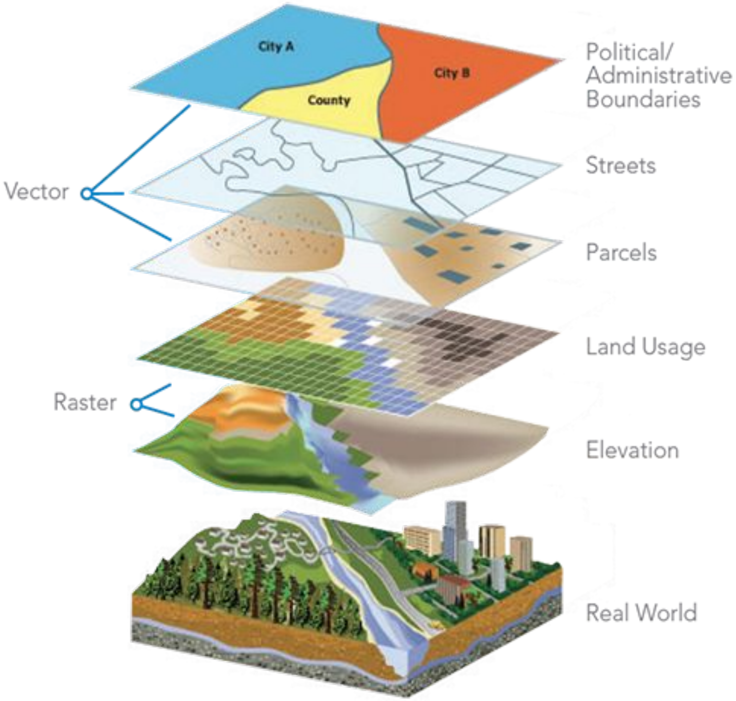
- **Record Data:** records with fixed attributes
  - Relational records
  - Data matrix …
  - Transaction Data
- **Graphs and Networks**
  - Transportation network
  - Social or information networks…
  - Molecular Structures
- **Ordered (Sequence) Data**
  - Video: sequence of image
  - Genetic Sequence Data
  - Temporal sequence …
- **Spatial Data**
  - RGB Images
  - Satellite images





Types of Data Sets

- **Record Data**
  - Relational records
  - Data matrix …
  - Transaction Data
- **Graphs and Networks**
  - Transportation network
  - Social or information networks…
  - Molecular Structures
- **Ordered (Sequence) Data**
  - Video: sequence of image
  - Genetic Sequence Data
  - Temporal sequence …
- **Spatial Data**
  - RGB Images
  - Satellite images



2- Data preprocessing

What is Data Preprocessing? — Major Tasks

Data cleaning

- Handle missing data, smooth noisy data, identify/remove outliers, and resolve inconsistencies

Data integration

- Integration of multiple databases, data cubes, or files

Data transformation and data discretization

- Normalization
- Discretization
- Sampling

Data reduction (covered in the next chapter)

- Dimensionality reduction
- Data compression

Data Quality

- Poor Data Quality adversely affects data processing efforts.

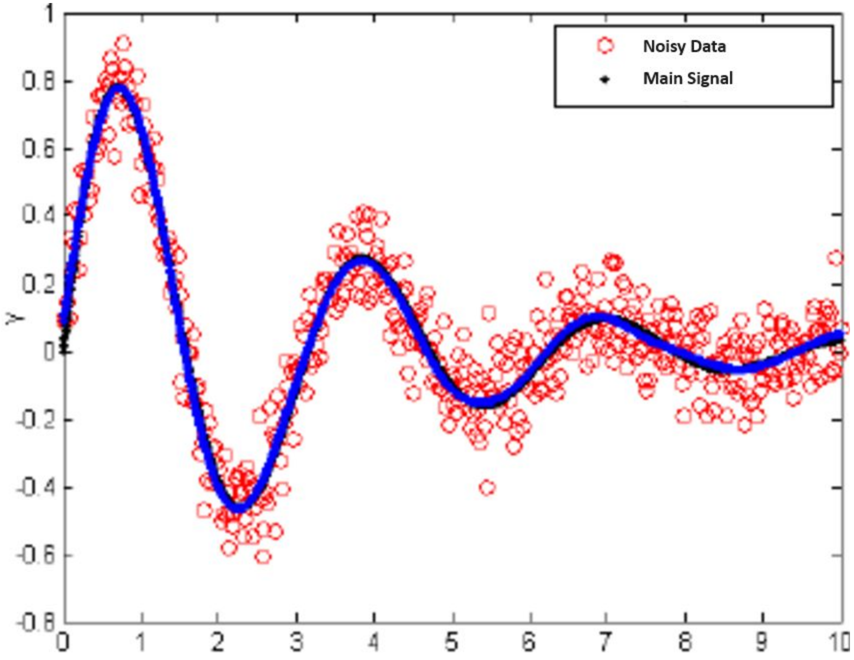
Example: Poor data can result in wrong loan decisions.
- *Some credit-worthy candidates are denied loans*
- *More loans are given to individuals that default*

- What types of data quality issues exist ? and how can we identify them?
- What can we do about these problems?

- Examples of data quality problems:
  - Noise and outliers
  - Wrong data
  - Fake data
  - Missing values
  - Duplicate data

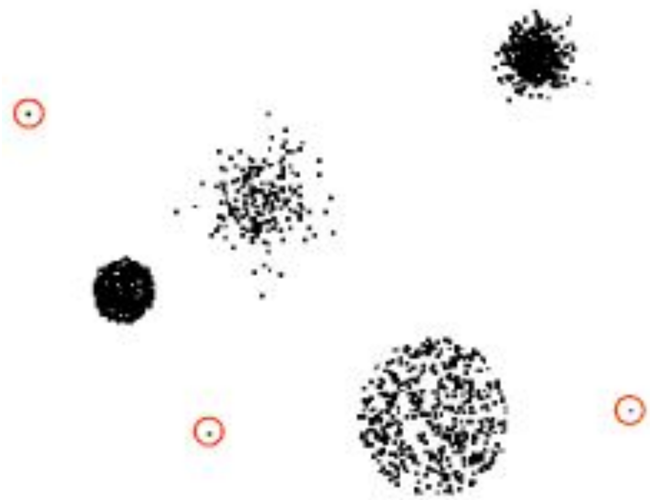
Noise

- **Noise in Objects:** Extraneous elements affecting data integrity.
- **Noise in Attributes:** Modification of original attribute values.
- **Examples:**
  - Distorted voice on a poor phone line.
  - "Snow" on a television screen.
  - Erroneous entries caused by data entry errors or system glitches.



Outliers

- Data objects with characteristics significantly different from the majority in the dataset.
- **Case 1:** Outliers as Noise:
    - Outliers can be noise that disrupts data analysis.
- **Case 2:** Outliers as the Focus:
    - In certain scenarios, outliers are the primary focus of analys
    - Credit card fraud detection
    - Intrusion detection.
- **Determining Causes:**
    - Explore the reasons behind the presence of outliers.



How to Handle Noisy Data?

- **Binning:** Sort data into bins, enabling smoothing using means, medians, or boundaries.
- **Regression:** Smooth data through regression functions.
    - Use other attributes to predict the noisy attributes
- **Clustering**: Identify and eliminate outliers.
- **Semi-supervised:** Combine automated and human inspection to identify noise and outliers.

Missing Values

Missing values

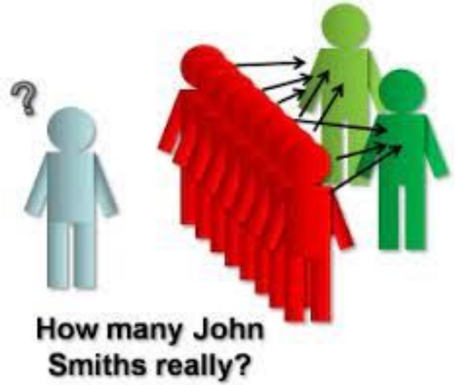
Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
22	1	0	A/5 21171	7.25		S
38	1	0	PC 17599	71.2833	C85	C
26	0	0	STON/O2. 3101282	7.925		S
35	1	0	113803	53.1	C123	S
35	0	0	373450	8.05		S
	0	0	330877	8.4583		Q

- **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 or variables
  - Estimate missing values
    - **Example**: time series of temperature
    - **Example**: census results
  - Ignore the missing value during analysis

Duplicate Data

- Occurrence of identical or nearly identical data objects.
- Common when merging data from diverse sources.
    - **Example:** Identical individuals with multiple email addresses.



How to handle duplicate data

- Remove duplicate data objects.
- Keep Duplicate Data: When and Why?
    - Customers with multiple accounts may unintentionally accumulate points separately.
    - Keeping duplicate data ensures they receive all earned benefits.

Data Transformation

- **Normalization:** Scaling data to a standard range (e.g., 0 to 1).
- **Discretization:** Converting continuous data into discrete categories.
- **Sampling:** Selecting a subset to represent a larger population.

Normalization

- Normalization ensures that variables are on a consistent scale.
- Normalization is crucial for many data mining algorithms.
- Improved Algorithm Convergence.

Min-max normalization: to [new_min$_A$, new_max$_A$]

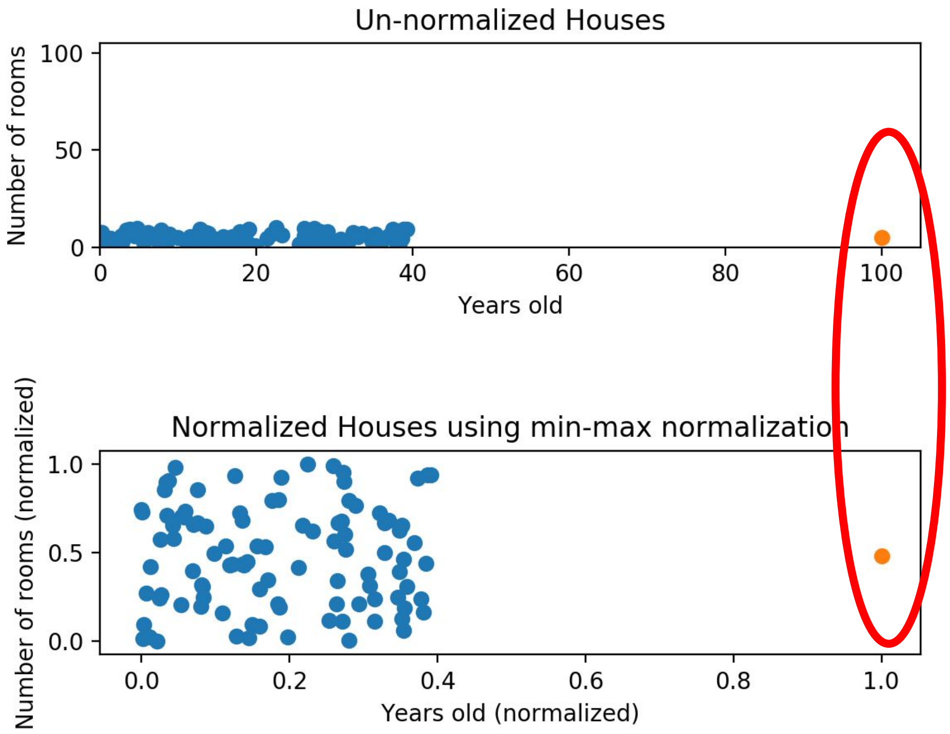
$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Z-score normalization (**μ**: mean, **σ**: standard deviation):

$$z = \frac{X - \mu}{\sigma}$$

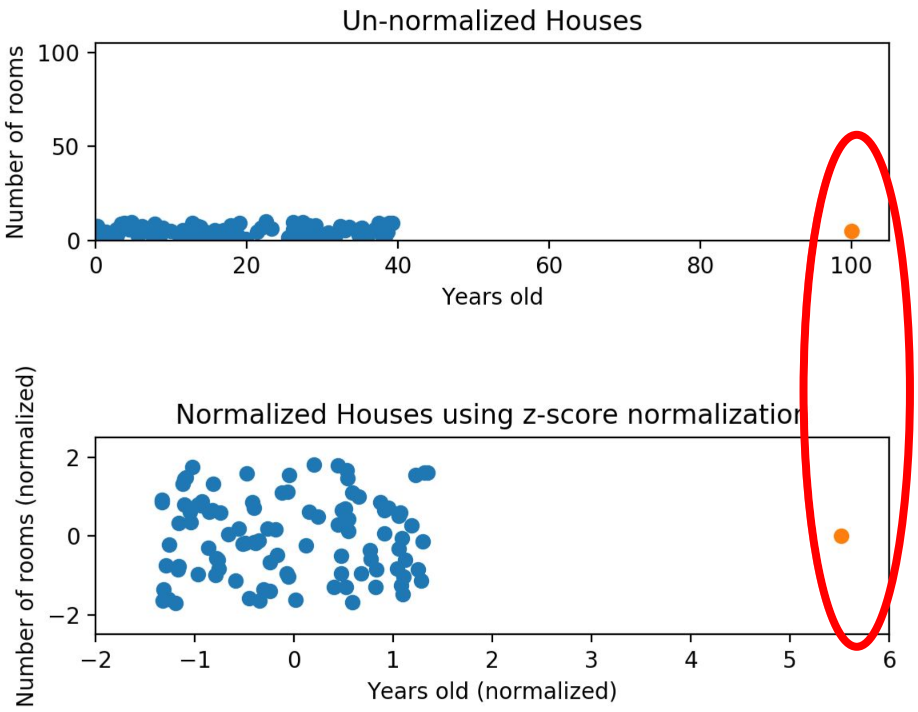
Min-max normalization **VS** Z-score normalization

- **Min-max normalization:**
  - Guarantees all attributes will have the exact same scale.
  - Does not handle outliers well.
- **Z-score normalization:**
  - Handles outliers.
  - Does not produce normalized data with the exact same scale.

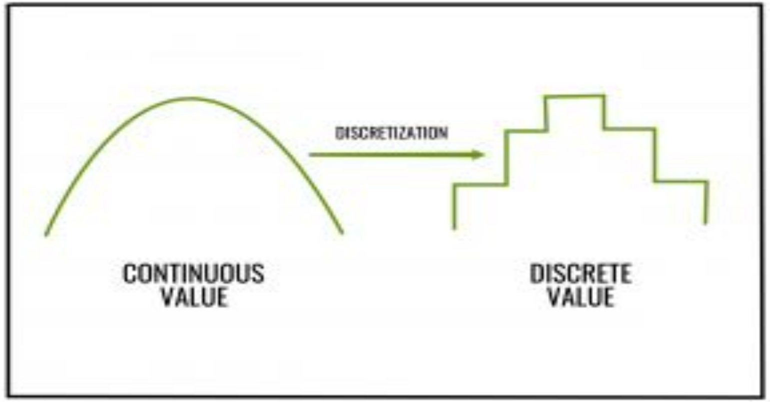


Min-max normalization **VS** Z-score normalization

- **Min-max normalization:**
  - Guarantees all attributes will have the exact same scale.
  - Does not handle outliers well.
- **Z-score normalization:**
  - Handles outliers.
  - Does not produce normalized data with the exact same scale.



Discretization



- Converting a continuous attribute into an ordinal attribute.
- A potentially infinite number of values are mapped to a small number of categories.
- Discretization is used in both unsupervised and supervised settings.

Discretization

- **Unsupervised**
  - **Binning:** Top-down split
  - **Histogram analysis:** Top-down split
  - **Clustering analysis:** top-down split or bottom-up merge
- **Supervised**
  - **Decision-tree analysis:** top-down split
  - **Correlation analysis:** bottom-up merge
- **Note:** All the methods can be applied recursively

Sampling

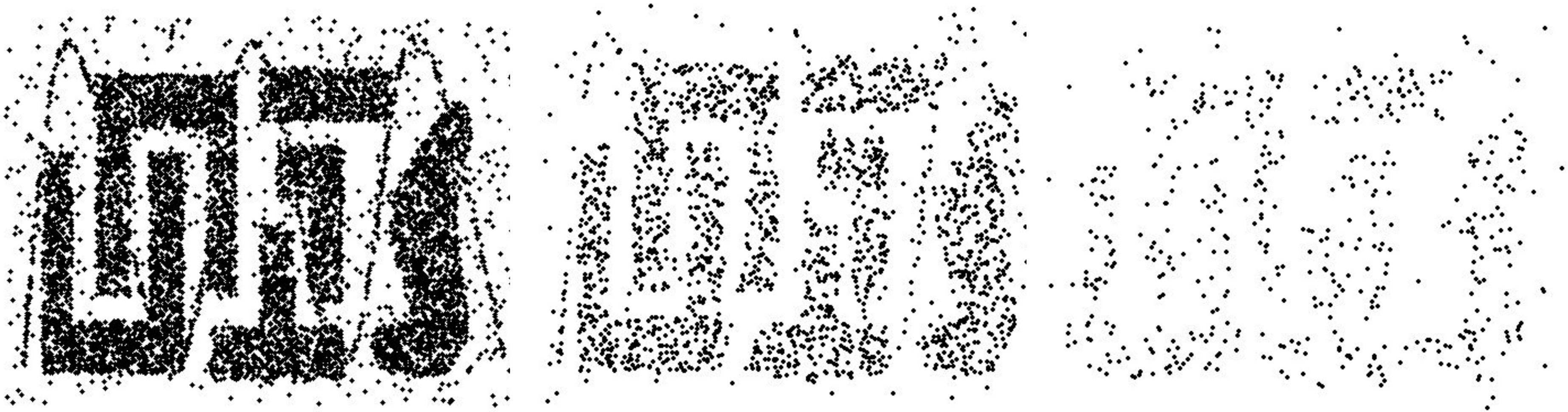
Sampling is selecting a subset of data from a larger dataset to make it more manageable for analysis while maintaining its representativeness.

- We use sampling because obtaining the entire dataset of interest is:
  - **Expensive:** Collecting, storing, and processing vast amounts of data can be cost-prohibitive.
  - **Time-consuming:** Analyzing the complete dataset can be impractical due to time constraints.
- Challenges:
  - Ensuring the sample is representative of the population.
  - Addressing potential bias in the sampling process.

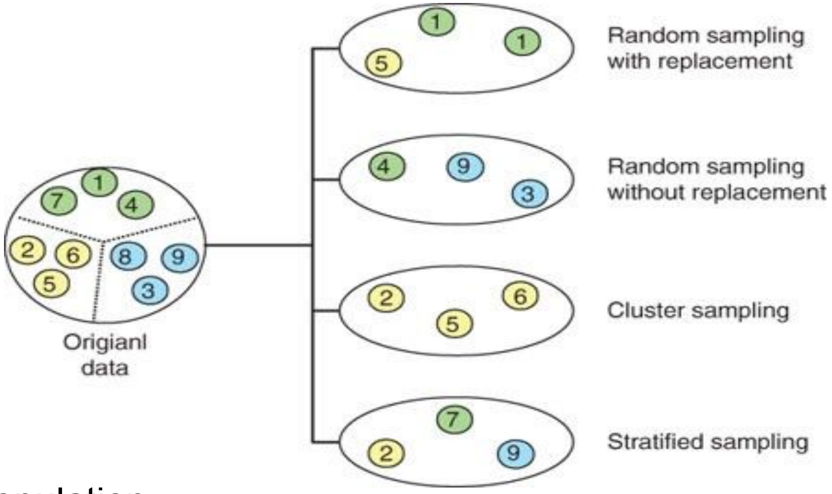
Sampling is an essential tool in data analysis, achieving a crucial equilibrium between **resource efficiency** and the **ability to derive meaningful insights**.

Sample size

Selecting an appropriate sample size is a critical decision in research and analysis.



Sampling methods



Simple random sampling

- Equal probability of selecting any particular item

Sampling without replacement

- Once an object is selected, it is removed from the population

Sampling with replacement

- A selected object is not removed from the population

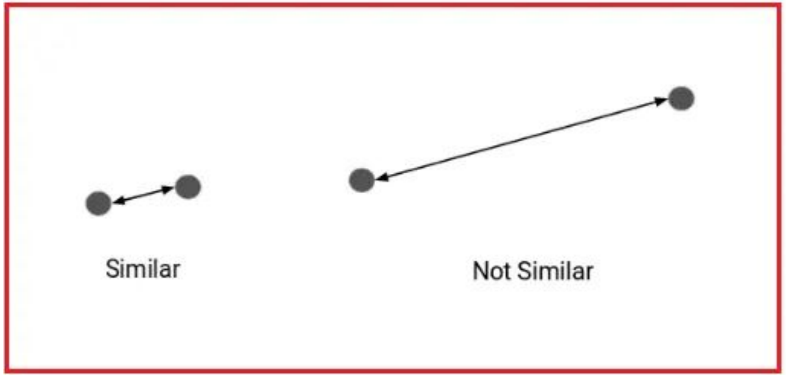
Stratified sampling

- Partition (or cluster) the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)

3- Similarity and Dissimilarity Measures

Similarity and Dissimilarity Measures

- **Similarity Measure:**
  - Quantifies data object likeness.
  - Higher values indicate greater similarity.
  - Typically within the range [0, 1].
- **Dissimilarity Measure:**
  - Quantifies data object differences.
  - Lower values indicate greater similarity.
  - Often starts at 0 and varies in the upper limit.
- **Proximity:**
  - Refers to either similarity or dissimilarity.



Similarity reveal valuable data relationships for pattern recognition, clustering, and classification.

Properties of a Distance

- Distance t is a **metric** if it satisfies these properties :
    - **Non-Negativity**:
        - $d(x, y) \geq 0$ for all x and y.
        - $d(x, y) = 0$ if and only if $x = y$.
    - **Symmetry**:
        - $d(x, y) = d(y, x)$ for all x and y.
    - **Triangle Inequality**:
        - $d(x, z) \leq d(x, y) + d(y, z)$ for all x, y, and z.
- Metrics ensure that distances align with real-world geometric properties

Metrics guarantee meaningful and reliable distance measurements in data analysis.

Properties of a Similarity

- **Identity:**
  - $s(x, y) = 1$ (or maximum similarity) only if $x = y$.
  - **Note:** This property may not always hold, e.g., cosine similarity.

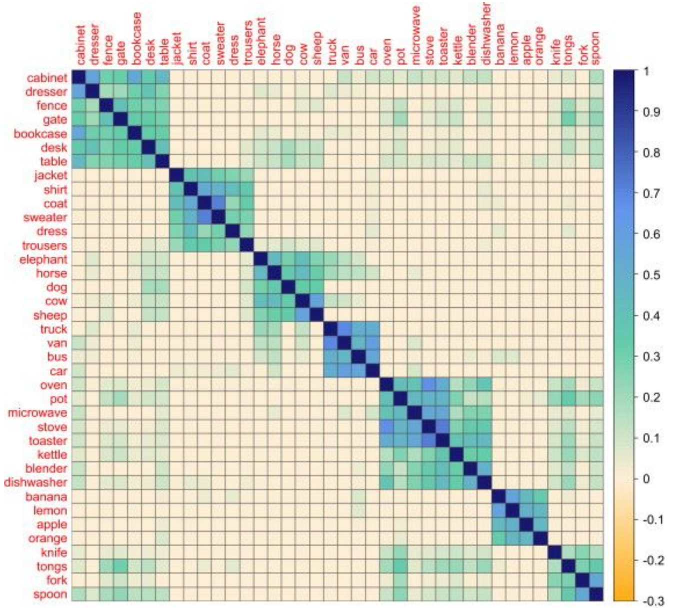
- **Symmetry:**
  - $s(x, y) = s(y, x)$ for all x and y.
  - Symmetry ensures that the order of comparison does not affect the similarity score.

Understanding these properties helps ensure the reliability and consistency of similarity measures in data analysis.

Similarity and dissimilarity matrix

- **Distance Matrix**
  - Distances between all data objects in a dataset.
  - Useful for clustering and nearest neighbor algorithms.
  - Symmetric, with values reflecting dissimilarities.

- **Similarity Matrix**
  - Similarities between data objects.
  - Valuable for clustering, recommendation systems, …
  - Often symmetric, with higher values indicating stronger similarities.



Distances and similarity examples

- **Proximity measures for numerical vectors**
  - Euclidean Distance
  - Minkowski Distance
  - Cosine Similarity
  - Linear correlation

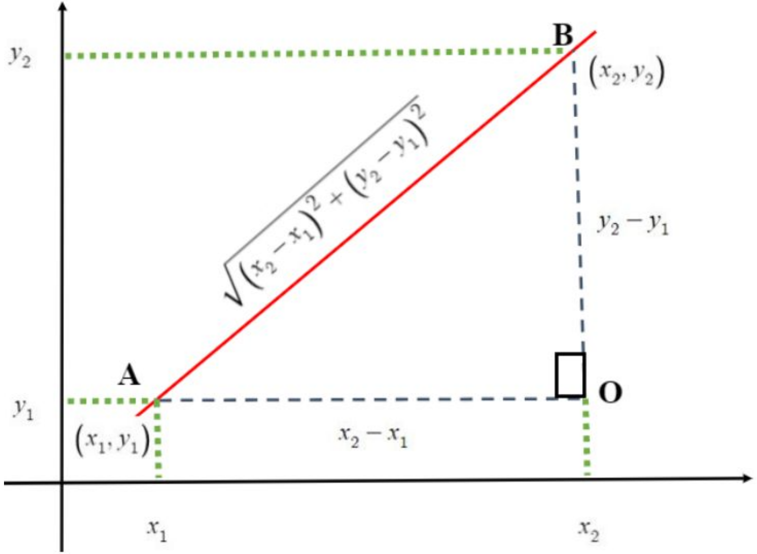
- Proximity measures for binary vectors
  - Simple Matching Coefficient (SMC)
  - Jaccard Coefficient

Euclidean Distance (applicable to numerical vectors)

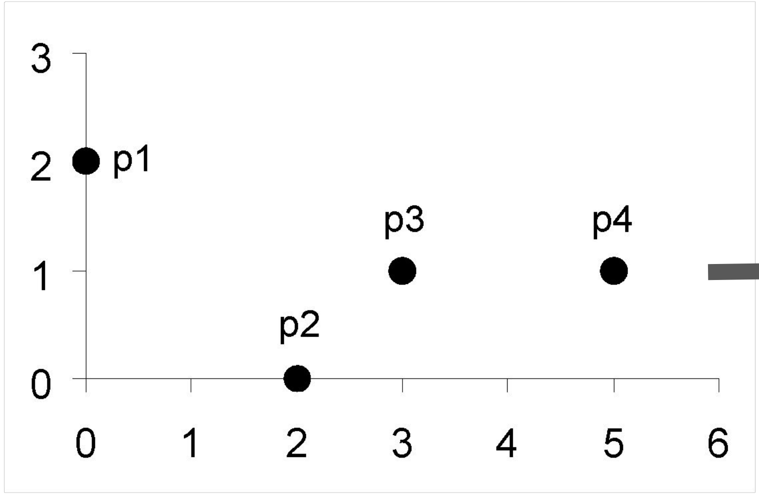
$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

- n : number of attributes.
- x_k, y_k : kth attributes for objects x and y, respectively.

Standardization is necessary, if scales differ.



Example: Euclidean Distance matrix



	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

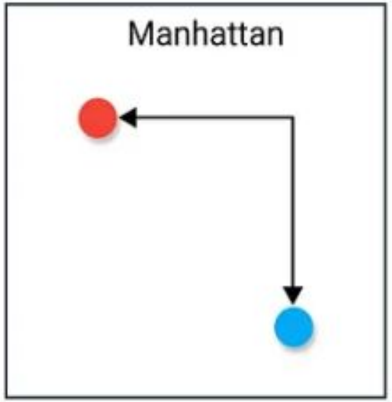
Minkowski Distance (applicable to numerical vectors)

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

- Generalization of Euclidean Distance.
- r : parameter
- n: number of attributes
- x_k and y_k are, respectively, the k^{th} attributes or objects x and y.
- The hyperparameters r Allows to adapt the distance to the characteristics of data.

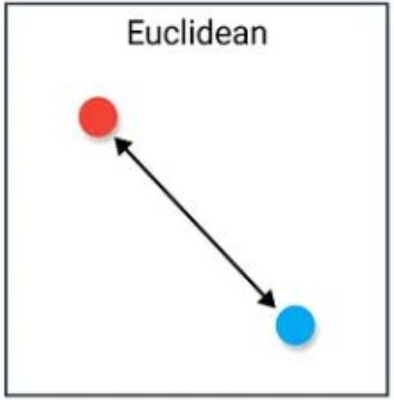
Special Cases of Minkowski Distance

- **City Block Distance (r = 1):**
    - Also known as Manhattan, taxicab, or L1 norm distance.
    - Ideal for measuring distances in grid-like paths.
    - Binary vector example: Hamming distance counts differing bits.
- **Euclidean Distance (r = 2):**
    - The most commonly used distance metric.
    - Measures the straight-line distance in Euclidean space.
- **Supremum Distance ($r \to \infty$):**
    - Also called **Lmax** norm or **L∞** norm distance.
    - Calculates the maximum difference between any component of vectors.
    - Appropriate when movement is unrestricted in any direction.



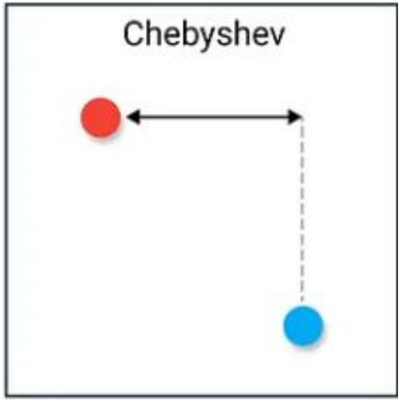
Special Cases of Minkowski Distance

- **City Block Distance (r = 1):**
  - Also known as Manhattan, taxicab, or L1 norm distance.
  - Ideal for measuring distances in grid-like paths.
  - Binary vector example: Hamming distance counts differing bits.
- **Euclidean Distance (r = 2):**
  - The most commonly used distance metric.
  - Measures the straight-line distance in Euclidean space.
- **Supremum Distance ($r \to \infty$):**
  - Also called **Lmax** norm or **L∞** norm distance.
  - Calculates the maximum difference between any component of vectors.
  - Appropriate when movement is unrestricted in any direction.



Special Cases of Minkowski Distance

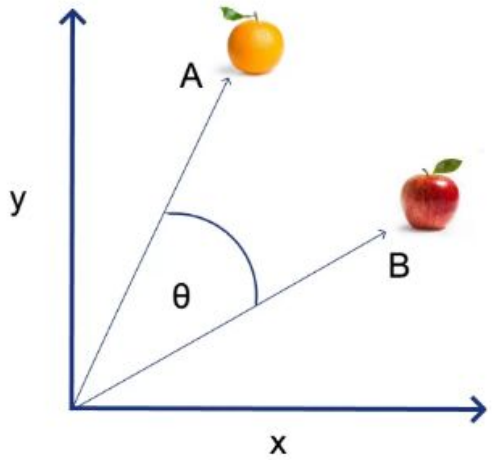
- **City Block Distance (r = 1):**
  - Also known as Manhattan, taxicab, or L1 norm distance.
  - Ideal for measuring distances in grid-like paths.
  - Binary vector example: Hamming distance counts differing bits.
- **Euclidean Distance (r = 2):**
  - The most commonly used distance metric.
  - Measures the straight-line distance in Euclidean space.
- **Supremum Distance ($r \to \infty$):**
  - Also called **Lmax norm**, **L∞ or chebyshev** distance.
  - Calculates the maximum difference between any component of vectors.
  - Appropriate when movement is unrestricted in any direction.



Cosine Similarity (applicable to numerical vectors)

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

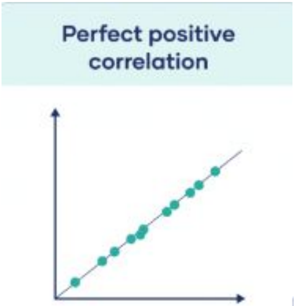
- **A . B** is dot product of the two vectors
- It is cosine of the angle between two vectors
- Non-sensitive to magnitudes, focusing on orientation.
- Values are between -1 and 1:
  - -1 (completely dissimilar)
  - 1 (perfect similarity).
  - 0 means orthogonal (no similarity).

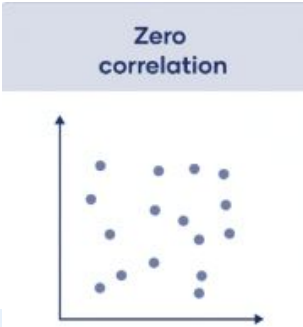


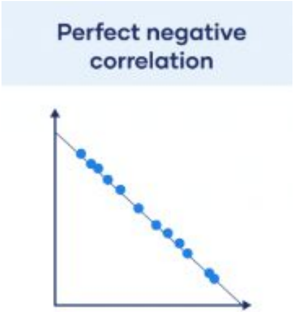
Linear correlation (applicable to numerical vectors)

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

- Measure the linear relationship between two variables.
- Evaluates how well one variable predicts another one.
- Values are between -1 and 1:
  - -1 (perfect inverse correlation)
  - 1 (perfect correlation).
  - 0 means orthogonal (no linear relationship).
- Commonly used in statistical analysis and data exploration.
- It unable to capture nonlinear associations.







Distances and similarity examples

- **Proximity measures for numerical vectors**
  - Euclidean Distance
  - Minkowski Distance
  - Cosine Similarity
  - Linear correlation

- **Proximity measures for binary vectors**
  - Simple Matching Coefficient (SMC)
  - Jaccard Coefficient

Similarity Between Binary Vectors

- **Simple Matching Coefficient (SMC)**: the number of matches divided by the total number of attributes.

$$\mathbf{SMC} = (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{00} + f_{11})$$

  - f_{01} = the number of attributes where $\mathbf{x}$ was 0 and $\mathbf{y}$ was 1
  - f_{10} = the number of attributes where $\mathbf{x}$ was 1 and $\mathbf{y}$ was 0
  - f_{00} = the number of attributes where $\mathbf{x}$ was 0 and $\mathbf{y}$ was 0
  - f_{11} = the number of attributes where $\mathbf{x}$ was 1 and $\mathbf{y}$ was 1

Similarity Between Binary Vectors

- **Jaccard Coefficient (J)**: the number of "11" matches relative to the total number of “00” non-zero attributes.
- It is designed for asymmetric binary attributes.

$$\mathbf{J} = f_{11} \,/\, (f_{01} + f_{10} + f_{11})$$

- f_{01} = the number of attributes where $\mathbf{x}$ was 0 and $\mathbf{y}$ was 1
- f_{10} = the number of attributes where $\mathbf{x}$ was 1 and $\mathbf{y}$ was 0
- f_{00} = the number of attributes where $\mathbf{x}$ was 0 and $\mathbf{y}$ was 0
- f_{11} = the number of attributes where $\mathbf{x}$ was 1 and $\mathbf{y}$ was 1

Example: SMC vs Jaccard Coefficient

x = 1 0 0 0 0 0 0 0 0 0

y = 0 0 0 0 0 0 1 0 0 1

- f_{01} = 2
- f_{10} = 1
- f_{00} = 7
- f_{11} = 0

SMC = 0.7

Jaccard = 0

How to Choose the Proximity Method ?

Choice of the right proximity measure depends on the domain

- Comparing Documents Using Word presence
  - Proximity Measure: Jaccard Coefficient
  - Similarity: Documents are considered similar if they use high number of common words.
- Comparing Temperature in Celsius of Two Locations
  - Proximity Measure: Euclidean Distance
  - Similarity: Two locations are considered similar if their temperatures are similar in magnitude.
- Comparing Two Time Series of Temperature (Celsius)
  - Proximity Measure: Cosine Similarity
  - Similarity: Two time series are considered similar if their "shape" is similar, i.e., they vary in the same way over time.
- Measuring Linear Relationship
  - Proximity Measure: Linear Correlation
  - Similarity: Measures the linear relationship between two variables. .

Similarity and Dissimilarity and attribute type

Similarity/dissimilarity between two objects, **x** and **y**, with only one attribute:

Attribute Type	**Dissimilarity**	**Similarity**
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$