KAZAKH-BRITISH TECHNICAL UNIVERSITY

School of IT and Engineering

MACHINE LEARNING

LECTURE #1

Number of credits: 3 (2/0/1)
Course code – CSCI3234
MS Teams team code – **y4ntwlz**

Adilet Yerkin, MS in Engineering, Senior Lecturer
a.yerkin@kbtu.kz

KBTU KAZAKH-BRITISH TECHNICAL UNIVERSITY

School of IT and Engineering

MS Teams team code – **y4ntwlz**

Adilet Yerkin, a.yerkin@kbtu.kz

NOT!
a_yerkin@kbtu.kz
ad_yerkin@kbtu.kz

- MS in Engineering, Senior Lecturer KBTU
- PhD Student 2nd year
- Head of Data Science Division at SCB
- Senior Scientific Researcher at Laboratory of Fuzzy Logic and Intelligent Systems

Number of credits: 3 (2/0/1)
Course code – CSCI3234
MS Teams team code – y4ntwlz

COURSE AIMS:

The course aims to equip students with fundamental knowledge and practical skills in machine learning for building, evaluating, and deploying predictive models. It focuses on the principles of learning from data, model generalization, and performance evaluation, while also addressing modern challenges related to scalability, reliability, and production deployment.

The course prepares students to apply machine learning methods effectively in both research and real-world systems.

Learning how to see the invisible:

- **Everyone else sees chaos, you'll see structure.**
- **Everyone else sees numbers, you'll see meaning.**
- **Everyone else sees data, you'll see knowledge and opportunities.**

SYLLABUS

This course provides a systematic introduction to machine learning, covering the complete lifecycle of machine learning solutions—from data preprocessing and feature engineering to model training, validation, and deployment.

Students study supervised and unsupervised learning methods, including regression, classification, clustering, and ensemble techniques, as well as similarity measures, dimensionality reduction, and outlier analysis. Special attention is given to model evaluation, validation strategies, handling imbalanced data, and preventing overfitting.

The course also introduces machine learning system design, data engineering fundamentals, model interpretability, automated machine learning, and deployment paradigms.

Practical work in Python complements theoretical lectures, enabling students to implement algorithms, analyze results, and understand the differences between experimental and production-level machine learning systems.

The course integrates theoretical lectures, independent practice works, a midterm exam, and a final course project, encouraging hands-on problem-solving and research-oriented thinking.

SYLLABUS

COURSE PREREQUISITES:

- Introduction to Machine Learning
- Basic Python programming
- Basic mathematics (algebra, simple calculus)
- Understanding of descriptive statistics (mean, median, variance)

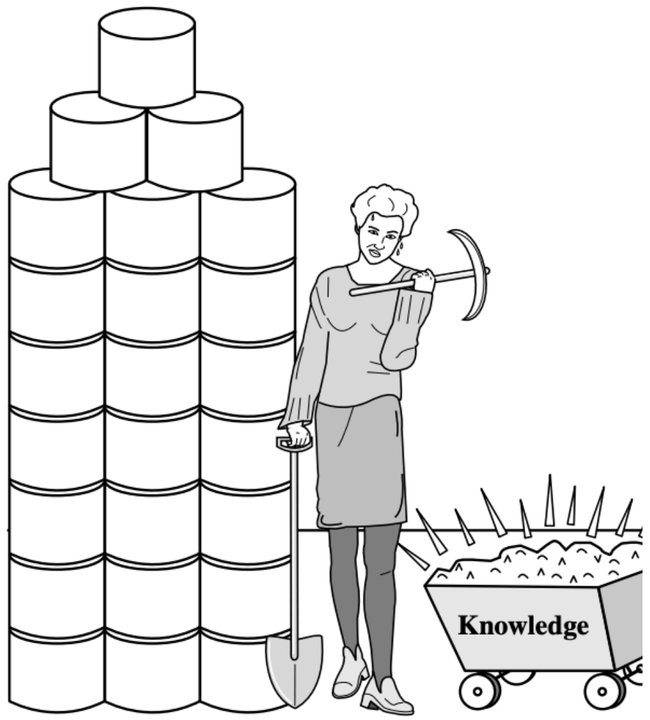
Type of activity	Final scores
Attendance /participation	**6**
Midterm/Endterm	**30**
SIS	**24**
Final exam - Project defense	**40**
Total	**100**

Week	Practice works	Cost (in points)
3	**SIS 1**	6
6	**SIS 2**	6
12	**SIS 3**	12
	Total	**24**

Introduction

With the ever increasing amounts of data in electronic form, the need for automated methods for data analysis continues to grow. The goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest.

- Machine learning is the most significant shift in computing since programmable computers
- System behavior emerges from data, not explicit instructions
- Traditional software engineering is insufficient for learning and adaptive systems
- ML systems address problems involving uncertainty and massive data
- Key applications: medicine, climate modeling, finance, autonomous systems
- Understanding ML engineering is essential to solve complex real-world problems



Searching for knowledge (interesting patterns) in data

ML Applications in Heavy Production Automation

Urban Infrastructure (Sewerage & Water)

- Predictive blockage and corrosion formation
- Autonomous flow, pump, and valve control
- Micro-forecasting of floods and contamination

Construction & Civil Engineering

- AI-driven construction sequencing and rescheduling
- Structural health monitoring of buildings and bridges
- Autonomous earth-moving and safety analytics

Mining & Ore Extraction

- Learning ore–waste boundaries in real time
- Autonomous drilling and blasting optimization
- Slope stability and collapse risk forecasting

Cross-Industry Impact

- Remaining-life prediction of critical equipment
- Subsurface modeling from sparse sensors
- ML-based emergency response and environmental protection

Sensors & IoT → ML Models → Autonomous Control → Safer, Cheaper, Sustainable Production

Data analytics = Turning Raw Data into Gold

Teaches you how to transform millions of rows of “boring” numbers into insights that solve real-world problems: predicting diseases, detecting fraud, recommending movies, or optimizing business decisions.
In other words: you learn how to find the story hidden in data.

The Core of AI and Machine Learning

ChatGPT, self-driving cars, all these AI breakthroughs rely on principles: classification, clustering, prediction, anomaly detection.

Data

A data set can often be viewed as a collection of data objects.

Other names for a data object are record, point, vector, pattern, event, case, sample, instance, observation, or entity

An attribute = a data field that describes a characteristic of an object.

Different names in literature:

- Feature → machine learning
- Attribute → data mining, databases
- Variable → statistics
- Dimension → data warehousing

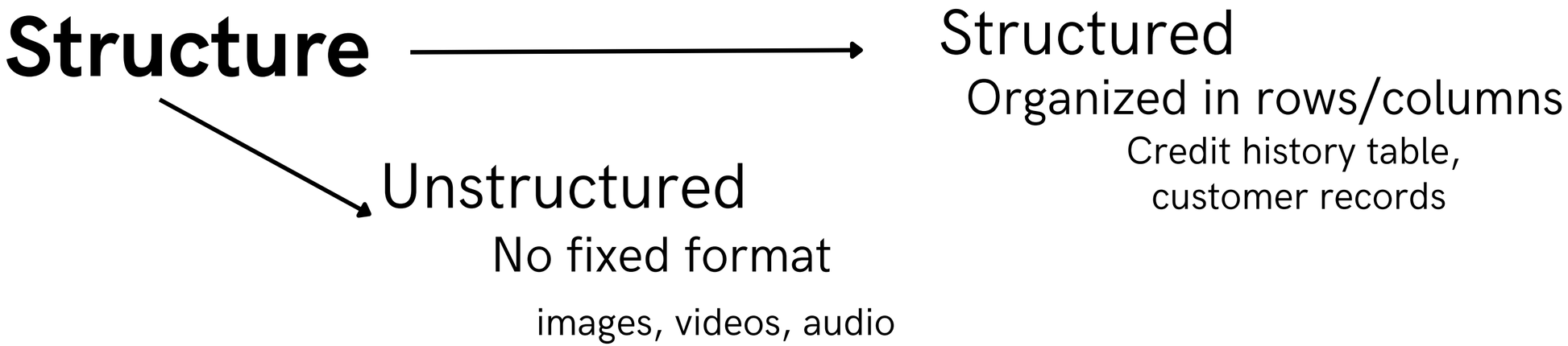
Golden Rules of Data (for future CDOs)

1. The data should be accurate.
2. The data should be stored according to data type.
3. The data should have integrity.
4. The data should be consistent.
5. The data should not be redundant.
6. The data should be timely.
7. The data should be well understood.
8. The data set should be complete.

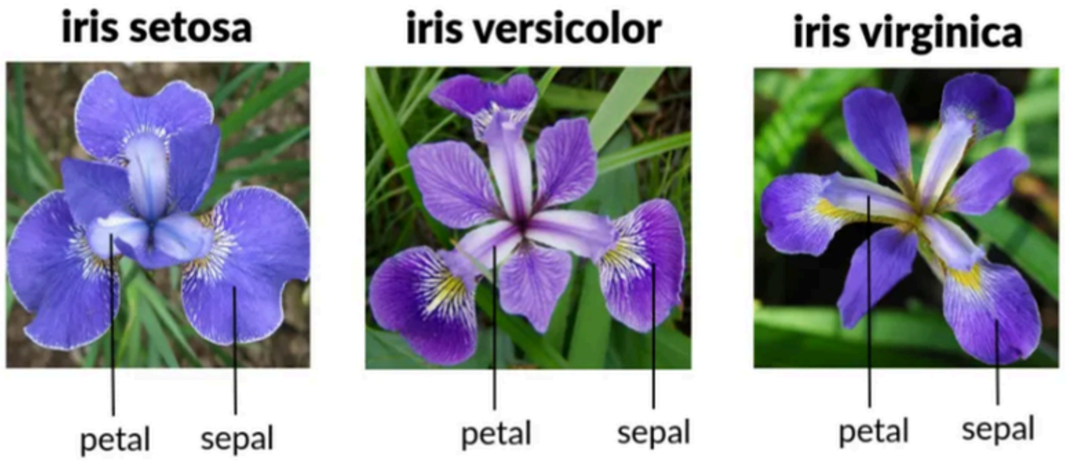
Data is often far from perfect

Data: Types of Data

Type	Subtype	Key Idea	Examples (Kazakhstan)
Categorical (qualitative)	Nominal	Labels, no order	City {Almaty, Astana}
	Ordinal	Ordered	Education level {Bach., Master}
Numeric (quantitative)	Discrete	Countable	Number of students
	Continuous	Any value in range	Temperature in Astana



Attributes



Iris flower

# sepal_length	# sepal_width	# petal_length	# petal_width	A species
4.3 – 7.9	2 – 4.4	1 – 6.9	0.1 – 2.5	3 unique values
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa

The iris dataset contains three classes of flowers, Versicolor, Setosa, Virginica.

Each class contains 4 features: ‘Sepal length’, ‘Sepal width’, ‘Petal length’, ‘Petal width’.

DATA PREPROCESSING

Data preparation (or preprocessing) is the set of techniques applied to raw data in order to transform it into a clean, consistent, and suitable form for analysis and modeling

In real-world applications, raw data is often:

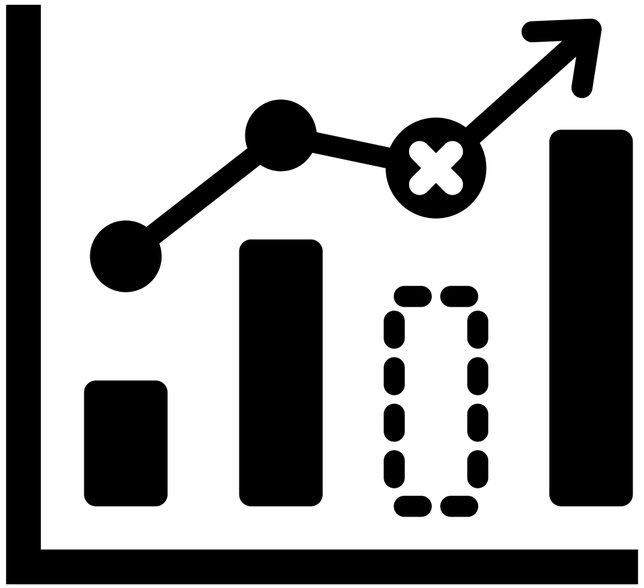
- **Incomplete (missing values, unknown attributes).**
- **Noisy (errors, outliers, random variations).**
- **Inconsistent (different coding systems, typos, contradictory records).**

Why it matters?

- **“Garbage in → Garbage out.” Quality of models depends on quality of input data.**
- **Up to 60–80% of the total effort in data mining projects is spent on preprocessing**

HANDLING MISSING DATA

Missing data occurs when some values are not recorded in the dataset.



Techniques:

Ignore (works only if dataset is large)

Fill with constants

Fill with mean/median/mode:

- Mean
- Median
- Mode

Predictive filling using regression, kNN, or decision trees.

Measuring the Central Tendency: Mean, Median, and Mode

The (arithmetic) mean

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

The **weighted arithmetic mean**
or the weighted average

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}$$

The weights reflect the significance, importance, or occurrence frequency attached to their respective values

Median: middle value of sorted data.

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width$$

- Data sets with one, two, or three modes are respectively called unimodal, bimodal, and trimodal.
- In general, a data set with two or more modes is multimodal/

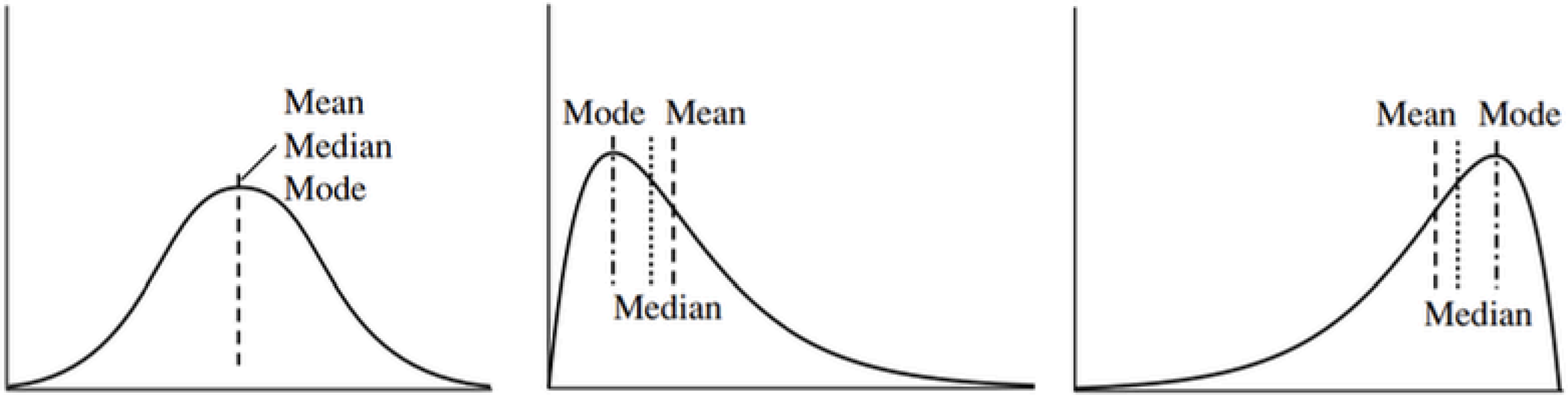
Mode: most frequent value

The mean is the singlemost useful quantity for describing a data set, it is not always the best way of measuring the center of the data. A major problem with the mean is its sensitivity to extreme (e.g., outlier) values. Even a small number of extreme values can corrupt the mean.

For skewed (asymmetric) data, a better measure of the center of data is the median, which is the middle value in a set of ordered data values. It is the value that separates the higher half of a data set from the lower half

Mode can be determined for qualitative and quantitative attributes. It is possible for the greatest frequency to correspond to several different values, which results in more than one mode

Measuring the Central Tendency: Mean, Median, and Mode

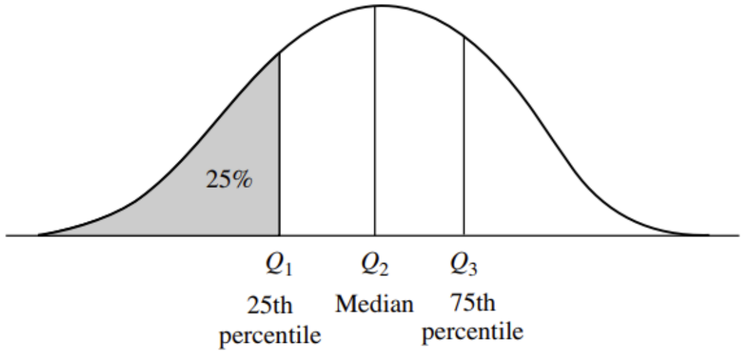


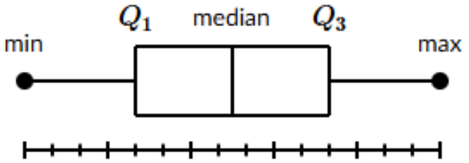
(a) Symmetric data **(b)** Positively skewed data **(c)** Negatively skewed data

Mean, median, and mode of symmetric versus positively and negatively skewed data.

Measuring the Dispersion of Data: Range, Quartiles, Variance, Standard Deviation, and Interquartile Range

The range of the set is the difference between the largest (max()) and smallest (min()) values



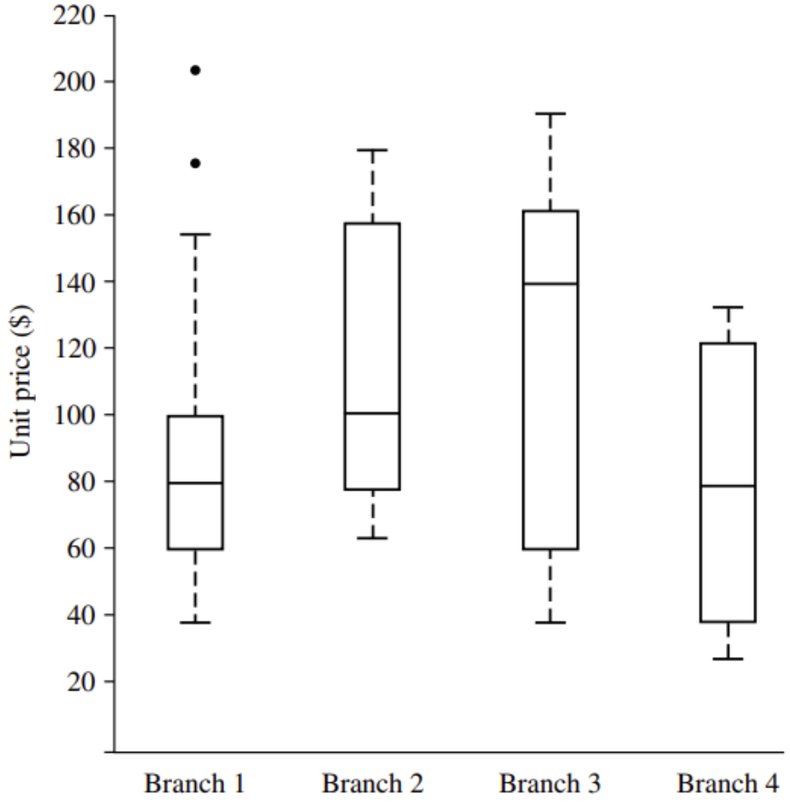


Interquartile range (IQR)

$$IQR = Q_3 - Q_1$$

Boxplots are a popular way of visualizing a distribution.
A boxplot incorporates the five-number summary as follows:

- The ends of the box are at the quartiles so that the box length is the interquartile range.
- The median is marked by a line within the box.
- Two lines (called whiskers) outside the box extend to the smallest (Minimum) and largest (Maximum) observations.



Measuring the Dispersion of Data: Range, Quartiles, Variance, Standard Deviation, and Interquartile Range

Variance and standard deviation are measures of data dispersion. They indicate how spread out a data distribution is.
A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.

Variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

The standard deviation, σ, of the observations is the square root of the variance

Chebyshev's inequality is a probabilistic inequality that provides a lower bound on the proportion of data values that lie within a certain number of standard deviations from the mean.

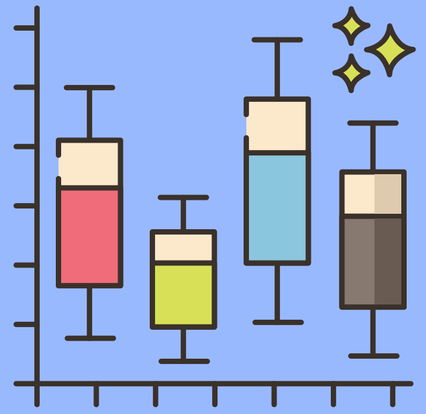
$$P\left(|X - \mu| < k\sigma\right) \geq 1 - \frac{1}{k^2}$$

Suppose the average monthly income of employees is: μ = 3000, with standard deviation σ = 500 . What fraction of employees have incomes between 2000 and 4000?
Here, 2000 = $\mu - 2\sigma$, 4000 = $\mu + 2\sigma$
So k = 2.
By Chebyshev's inequality:

$$P(|X - \mu| < 2\sigma) \geq 1 - \frac{1}{2^2} = 0.75$$

At least 75% of employees earn between 2000 and 4000

OUTLIER ANALYSIS



A data object that deviates significantly from the rest of the dataset and may indicate noise, rare events, or novel patterns.

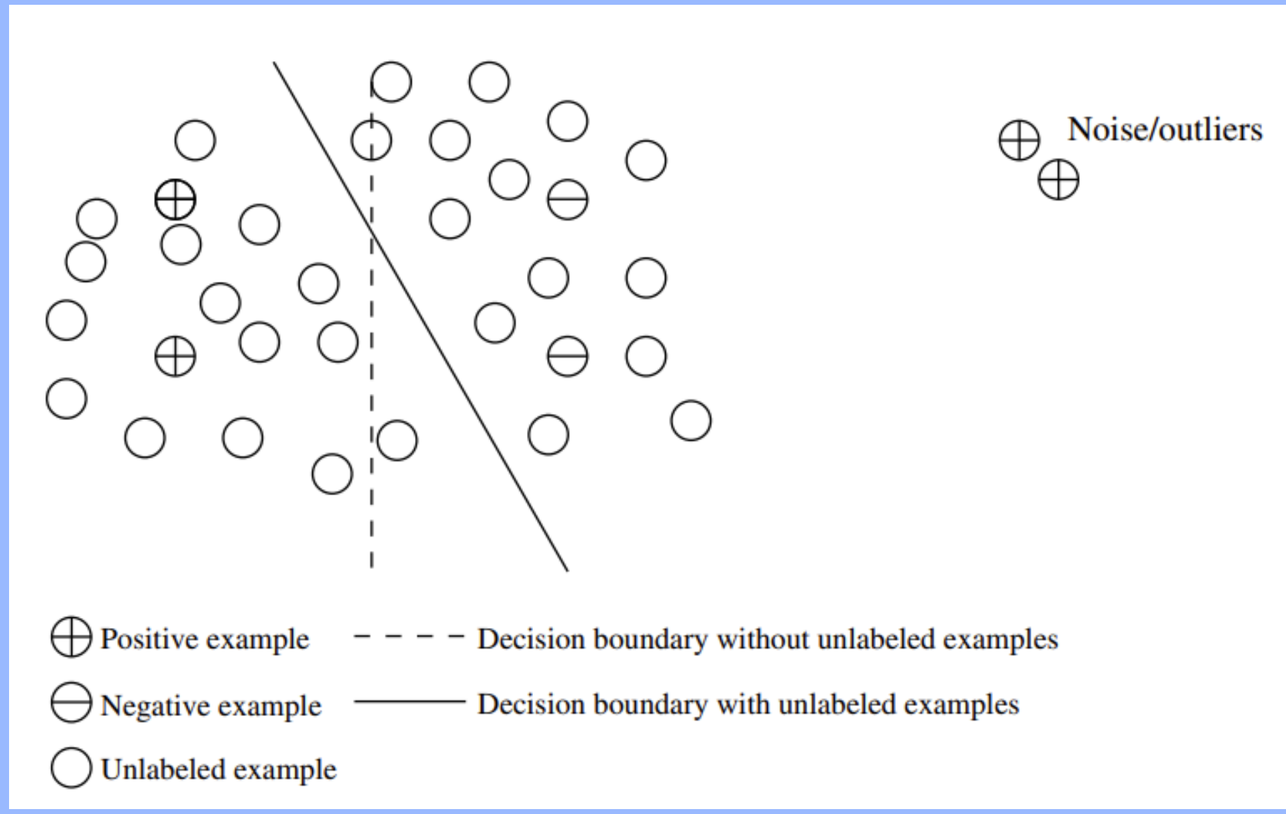
In large datasets, we often find observations that do not follow the general pattern of the data. Such observations are called outliers. An outlier is a data point that is significantly different from the majority of the data.
Examples: If the age of a person is recorded as –1, the value is clearly incorrect, probably caused by a default system setting for “unrecorded age.” If the number of children for one person is recorded as 25, the value is unusual. It may be an error (typo), but it could also be true — representing rare, valid variability. Thus, outliers may be caused by errors in data collection or by rare but real events.

Outliers can:

- Distort models(misrepresent): They may introduce skewness and make it hard to fit accurate models.
- Mislead analysis: Algorithms such as clustering, regression, or PCA are sensitive to unusual values.
- Contain valuable information: In fraud detection, rare outliers (e.g., suspicious credit card transactions) are the main focus of the analysis.

Applications

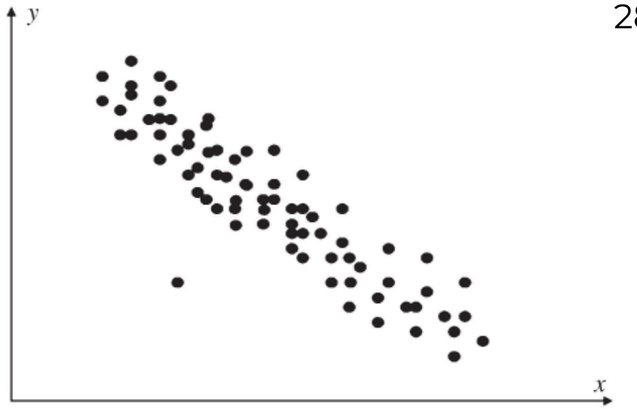
- Fraud detection: Unusual banking or credit card transactions may indicate fraud.
- Fault detection: Abnormal sensor readings can indicate machine breakdown.
- Cybersecurity: Outliers in network traffic may show intrusion attempts.
- Healthcare: Rare medical results may point to new diseases or errors in measuremen



OUTLIER DETECTION METHODS

Graphical

histograms, boxplots, scatterplots



Statistical

Z-score: $z = \frac{x-\mu}{\sigma}$.

Outlier if $|z| > 3$.

Age= {3,56,23,39, 156,52,41,22,9, 28, 139,31,55,20, -67,37,11,55,45,37}

then the corresponding statistical parameters are
Mean = 39.9
Standard deviation = 45.65

Threshold= Mean ± 2 x Standard deviation

Distance-based

k-nearest neighbor distance

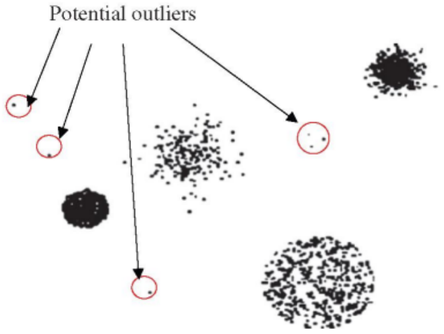
S= {(2,4), (3,2), (1,1), (4,3),(1,6), (5,3), (4,2)}

	s_1	s_2	s_3	s_4	s_5	s_6	s_7
s_1		2.236	3.162	2.236	2.236	3.162	2.828
s_2			2.236	1.414	4.472	2.236	1.000
s_3				3.605	5.000	4.472	3.162
s_4					4.242	1.000	1.000
s_5						5.000	5.000
s_6							1.414

Euclidean distances,

Model-based

clustering (DBSCAN), isolation forest



Determining outliers through clustering

DATA REDUCTION AND TRANSFORMATION

NORMALIZATION

Normalization is the process of rescaling values of a feature into a specific range, usually $[0, 1]$ or $[-1, 1]$

Decimal scaling

$$x' = \frac{x}{10^j} \quad \text{where } j = \min\{j : \max(|x'|) < 1\}$$

Min–Max Normalization

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Standard deviation normalization (Z-score Normalization)

$$x' = \frac{x - \mu}{\sigma}$$

Standardization

DISCRETIZATION AND BINARIZATION

Discretization

Transforming continuous attributes into categorica

Methods:

- Equal-width: split range into k intervals of equal size.
- Equal-frequency: each bin has same number of objects.
- Entropy-based: supervised, based on information gain.

Suppose we have ages of people:
22, 25, 29, 35, 42, 55, 63, 70
After discretization into 3 bins:

- Young (0–30): 22, 25, 29
- Middle-aged (31–50): 35, 42
- Old (51+): 55, 63, 70

Binarization

Converting categorical/numeric values into binary

Example: “Age > 30” → {0,1}

One-hot encoding for categorical variable
Colors: red, blue, green
After binarization (one-hot):

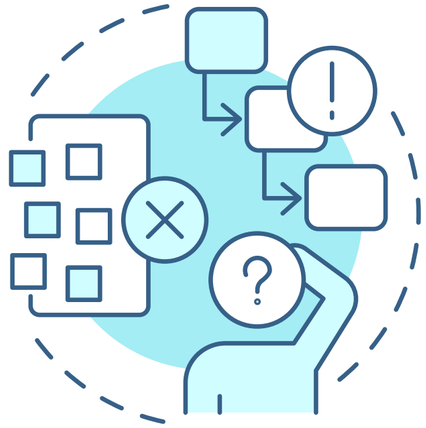
- Red → (1,0,0)
- Blue → (0,1,0)
- Green → (0,0,1)

HANDLING INCORRECT OR INCONSISTENT DATA

In real datasets, we often find incorrect or inconsistent entries.

- These can occur when the same information comes from different sources or when values are entered outside realistic ranges.
- Example: a person's height recorded as 6 meters is clearly wrong.

- Inconsistencies can also appear across sources (e.g., a full name in one source vs. initials in another).



Main approaches to handling incorrect entries:

- **Inconsistency detection:** Compare data across multiple sources to find duplicates or conflicts.

- **Domain knowledge:** Use known rules or ranges. For example, if Country = "Kazakhstan", then City cannot be "Paris." Data auditing tools often apply such rules.

Data reduction

Data reduction techniques obtain a reduced representation of the data while minimizing the loss of information content.

These include methods of dimensionality reduction, numerosity reduction, and data compression.

- **Dimensionality reduction** reduces the number of random variables or attributes under consideration. Methods include wavelet transforms, principal components analysis, attribute subset selection, and attribute creation. (+Feature selection - Redundant features)

- **Numerosity reduction** methods use parametric or nonparatmetric models to obtain smaller representations of the original data.

1. Parametric models store only the model parameters instead of the actual data. Examples include regression and log-linear models.
2. Nonparamteric methods include histograms, clustering, sampling, and data cube aggregation.

- **Data compression** methods apply transformations to obtain a reduced or “compressed” representation of the original data. The data reduction is lossless if the original data can be reconstructed from the compressed data without any loss of information; otherwise, it is lossy.

SAMPLING

Data sampling: The records from the underlying data are sampled to create a much smaller database.
Sampling is generally much harder in the streaming scenario where the sample needs to be dynamically maintained
Sampling can be used as a data reduction technique because it allows a large data set to be represented by a much smaller random data sample (or subset). Suppose that a large data set, D, contains N tuples.

Simple random sample without replacement (SRSWOR) of size - s

This is created by drawing s of the N tuples from D (s < N), where the probability of drawing any tuple in D is 1/N, that is, all tuples are equally likely to be sampled.

Simple random sample with replacement (SRSWR) of size s

This is similar to SRSWOR, except that each time a tuple is drawn from D, it is recorded and then replaced. That is, after a tuple is drawn, it is placed back in D so that it may be drawn again

Cluster sample

The population is divided into strata (subgroups) that are homogeneous within (similar inside) but different across

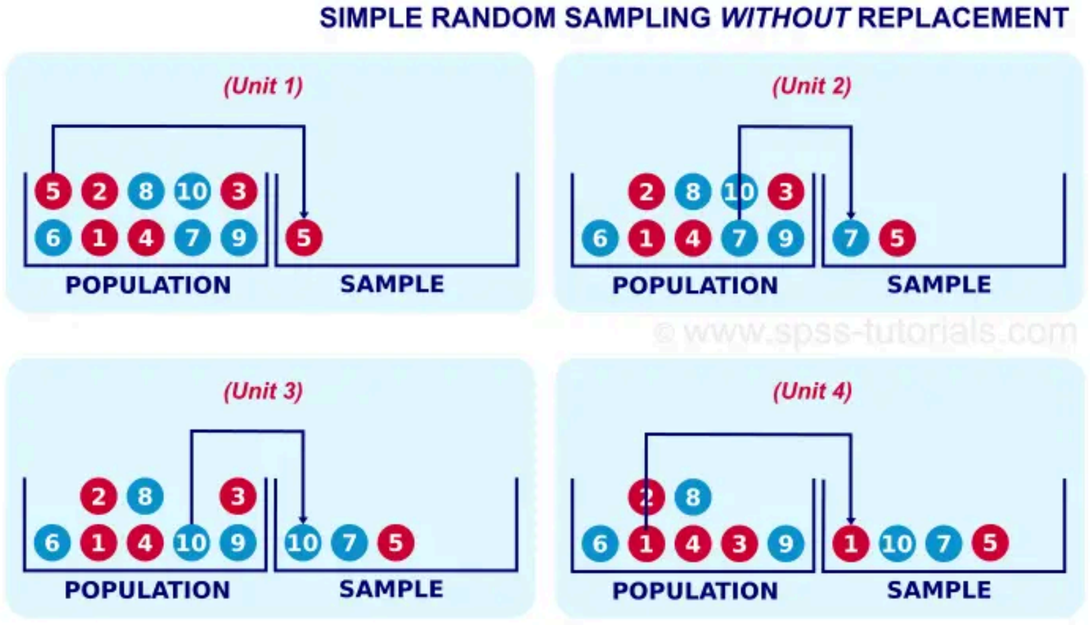
If the tuples in D are grouped into M mutually disjoint “clusters,” then an SRS of s clusters can be obtained, where s < M

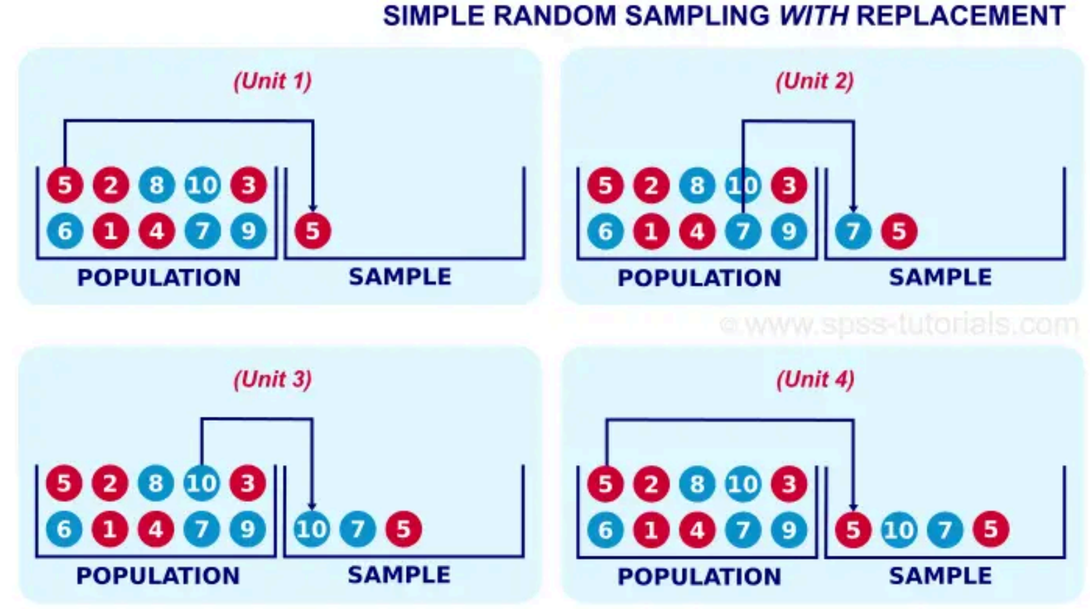
Stratified sample

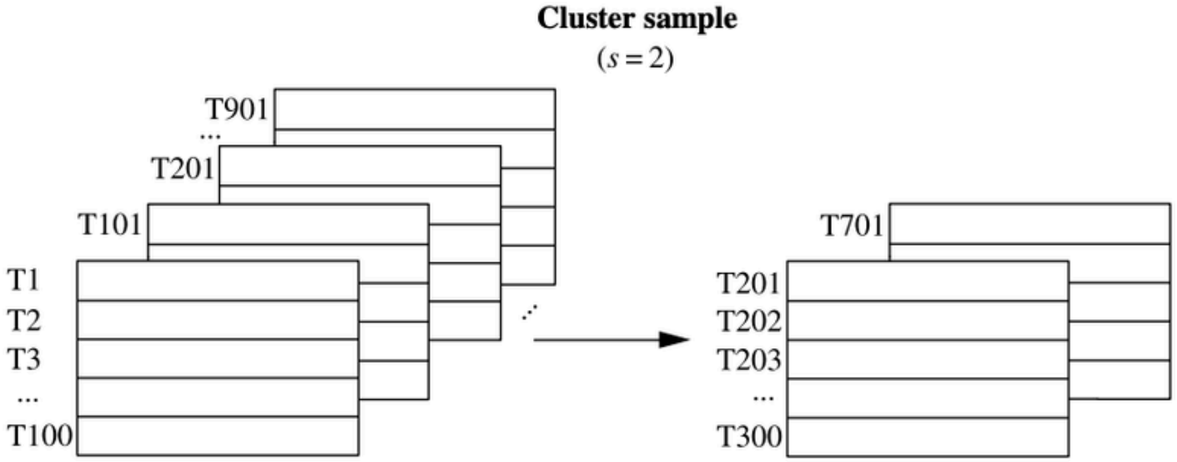
The population is divided into clusters (natural groups, often heterogeneous within)

If D is divided into mutually disjoint parts called strata, a stratified sample of D is generated by obtaining an SRS at each stratum.

Data Reduction







Startified sample
(according to *age*)

T38	youth
T256	youth
T307	youth
T391	youth
T96	middle_aged
T117	middle_aged
T138	middle_aged
T263	middle_aged
T290	middle_aged
T308	middle_aged
T326	middle_aged
T387	middle_aged
T69	senior
T284	senior

T38	youth
T391	youth
T117	middle_aged
T138	middle_aged
T290	middle_aged
T326	middle_aged
T69	senior

AGGREGATION

Sometimes “less is more,” and this is the case with aggregation , the combining of two or more objects into a single object

Transaction ID	Item	Store Location	Date	Price	…
⋮	⋮	⋮	⋮	⋮	
101123	Watch	Chicago	09/06/04	$25.99	…
101123	Battery	Chicago	09/06/04	$5.99	…
101124	Shoes	Minneapolis	09/06/04	$75.00	…

Task to complete before next class

1) Install and configure jupyter(anaconda) or google colab.
   https://www.anaconda.com/download
   https://colab.research.google.com
   https://www.kaggle.com

2) Connect to the Teams group, code - y4ntwlz

KBTU KAZAKH-BRITISH TECHNICAL UNIVERSITY

School of IT and Engineering

Thank you for your attention!