Data Mining and Data Warehousing

1. Introduction in Data Mining

Drd. Horia Modran

Contact: horia.modran@unitbv.ro / modranhoria@gmail.com

Tel: 0770171577

Universitatea Transilvania din Brașov
FACULTATEA DE INGINERIE ELECTRICĂ ȘI ȘTIINȚA CALCULATOARELOR

2022 - 2023

Number of hours and credits

Number of credits	5
Total hours per semester	125
Hours in the curriculum	56
Study / individual work	69

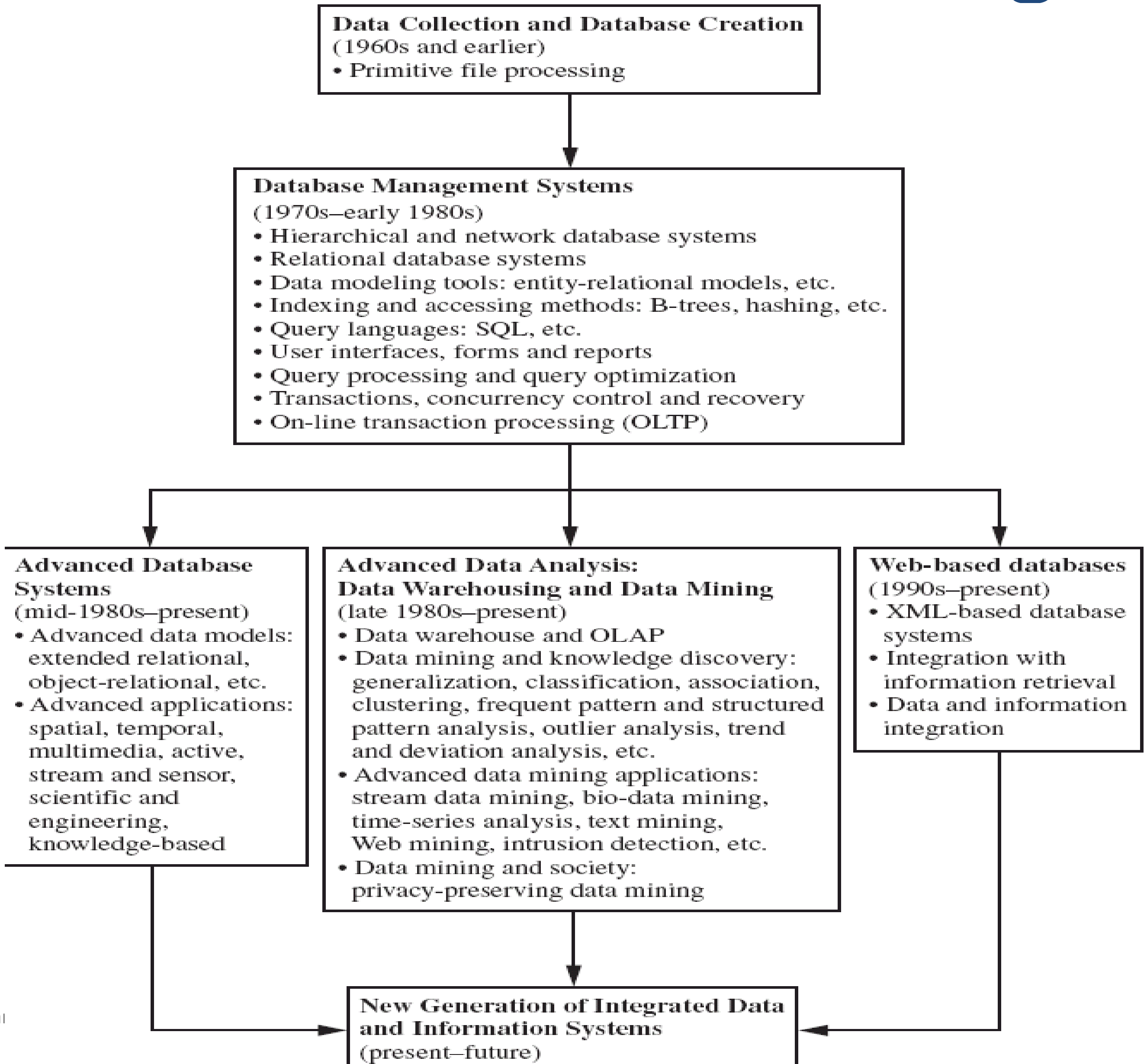
Sylabus

Content	# of hours
Introduction in Data Mining Applications and examples (Python)	3 1
Machine Learning – Supervised Learning Applications and examples (Python)	2 2
Machine Learning – Unsupervised Learning Applications and examples (Python)	2 2
Anomaly Detection Examples + Application in Cybersecurity	2 2
Neural Networks Design and Implementation	4 4

Evaluation

Criterion	Method	%
Exam – multiple choice questions	Written Exam	50%
Small Project Data Mining & Machine Learning	Presentation	50%
Attendance and activity at the course & laboratories	1 bonus point (max.)	-
TOTAL		100%

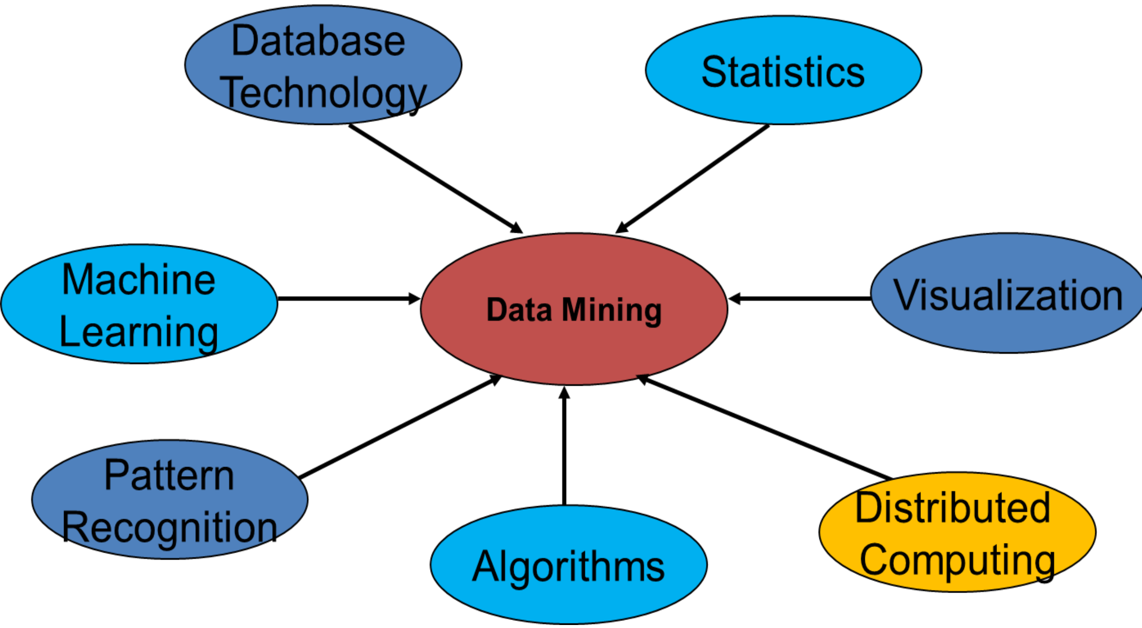
Evolution of DB Technologies



What is Data Mining?

- after years of data mining there is still no unique answer to this question
- the term "data mining" wasn’t coined until the 1990s
- possible definition: “Data mining is the use of efficient techniques for the analysis of very large collections of data and the extraction of useful and possibly unexpected patterns in data.”

Data Mining



- **Data Science:** Data is useful to understand a process and improve it
  - focuses on more immediate applications
- **Big Data:** Data appear everywhere. We should process it collectively and interconnect them. We need cloud infrastructure for this
  - more systems oriented
- **AI/Machine Learning/Deep Learning**: now we have the data to learn more complex models that are significantly more powerful
  - emphasis on scientific breakthroughs

Data Mining - Motivation

- the Explosive Growth of Data: from terabytes (1000^4) to yottabytes(1000^8) -> really huge amount of raw data !!
  - data collection and data availability
    - automated data collection tools, DB systems, web
  - major sources of abundant data
    - business: Web, e-commerce, transactions, stocks, etc.
    - science: bioinformatics, medical research
    - mobile devices, digital cameras, etc.
- How to analyze data?
- Data mining — automated analysis of massive data sets

Data Mining - Motivation

- large amounts of data can be more powerful than complex algorithms and models
  - Google has solved Natural Language Processing problems simply by looking at the data: misspelling, synonyms
- data is power
  - biggest assets of companies
- we need a way to harness the collective intelligence
- data is very complex: tables, time series, images, graphs

What is Data?

- 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.
- a collection of attributes describe an object
- object is also known as record, point, case, sample, entity, or instance

Attributes = Table columns

Objects = Table rows

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	**NULL**
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	**NULL**	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Size: Number of objects
Dimensionality: Number of attributes
Sparsity: Number of populated object-attribute pairs

Types of attributes

- There are different types of attributes
  - **Categorical**
    - examples: eye color, zip codes, words, rankings (e.g, good, fair, bad), height in {tall, medium, short}
    - nominal (no order or comparison) vs Ordinal (order but not comparable)
  - **Numeric**
    - Examples: dates, temperature, time, length, value, etc.
    - discrete (counts) vs Continuous (temperature)
    - special case: Binary attributes (yes/no, exists/not exists)

Numeric record data

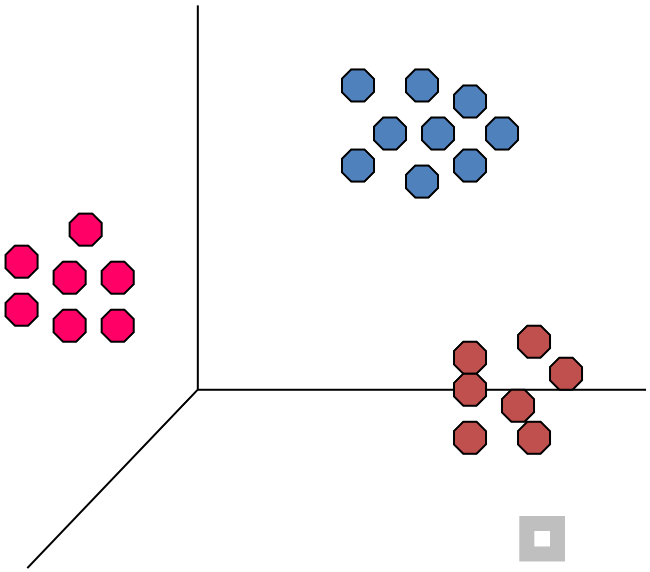
- if data objects have the same **fixed set** of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- such data set can be represented by an n-by-d **data matrix**, where there are n rows, one for each object, and d columns, one for each attribute

	Temperature	Humidity	Pressure
O1	30	0.8	90
O2	32	0.5	80
O3	24	0.3	95

30	0.8	90
32	0.5	80
24	0.3	95

Numeric data

- thinking of numeric data as points or vectors is very convenient
- for small dimensions we can plot the data
- we can use geometric analogues to define concepts like distance or similarity
- we can use linear algebra to process the data matrix
- we will often talk about points or vectors



Mixed relational data

- Data that consists of a collection of records, each of which consists of a fixed set of both numeric and categorical attributes

Takes numerical values but it is actually categorical

ID Number	Zip Code	Age	Marital Status	Income	Income Bracket	Refund
1129842	45221	55	Single	250000	High	0
2342345	45223	25	Married	30000	Low	1
1234542	45221	45	Divorced	200000	High	0
1243535	45224	43	Single	150000	Medium	0

Boolean attributes can be thought as both numeric and categorical
When appearing together with other attributes they make more sense as categorical
They are often represented as numeric though

Universitatea Transilvania din Brașov
FACULTATEA DE INGINERIE ELECTRICĂ ȘI ȘTIINȚA CALCULATOARELOR

Mixed relational data

- sometimes it is convenient to represent categorical attributes as boolean
  - add a Boolean attribute for each possible value of the

ID	Zip 45221	Zip 45223	Zip 45224	Age	Single	Married	Divorced	Income	Refund
1129842	1	0	0	55	0	0	0	250000	0
2342345	0	1	0	25	0	1	0	30000	1
1234542	1	0	0	45	0	0	1	200000	0
1243535	0	0	1	43	0	0	0	150000	0

We can now view the whole vector as numeric

Mixed relational data

- sometimes it is convenient to represent numerical attributes as categorical
  - group the values of the numerical attributes into bins

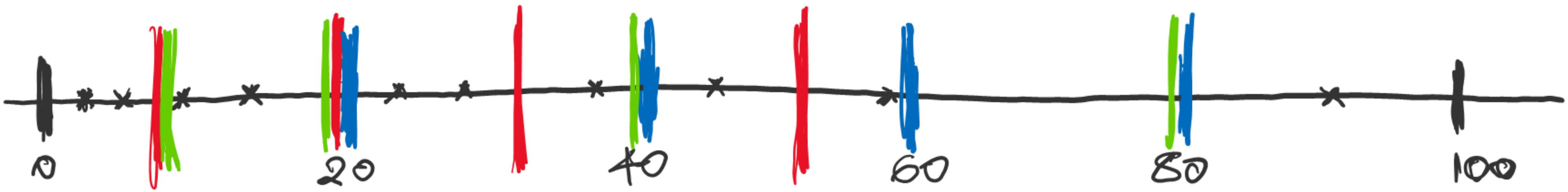
ID Number	Zip Code	Age	Marital Status	Income	Income Bracket	Refund
1129842	45221	50s	Single	High	High	0
2342345	45223	20s	Married	Low	Low	1
1234542	45221	40s	Divorced	High	High	0
1243535	45224	40s	Single	Medium	Medium	0

- Idea: split the range of the domain of the numerical attribute into bins (intervals)

Bucketization

- **Equi-width bins**: all bins have the same size
  - example: split time into decades
  - problem: some bins may be very sparse or empty
- **Equi-size (depth)** bins: select the bins so that they all contain the same number of elements
  - this splits data into quantiles: top-10%, second 10% etc
  - problem: some bins may be very small
- **Equi-log** bins: log end–log $start$ is constant
  - the size of the previous bin is a fraction of the current one
- Optimized bins: use a 1-dimensional clustering algorithm to create the bins

Bucketization - example



Blue: Equi-width [20,40,60,80]

Red: Equi-depth (2 points per bin)

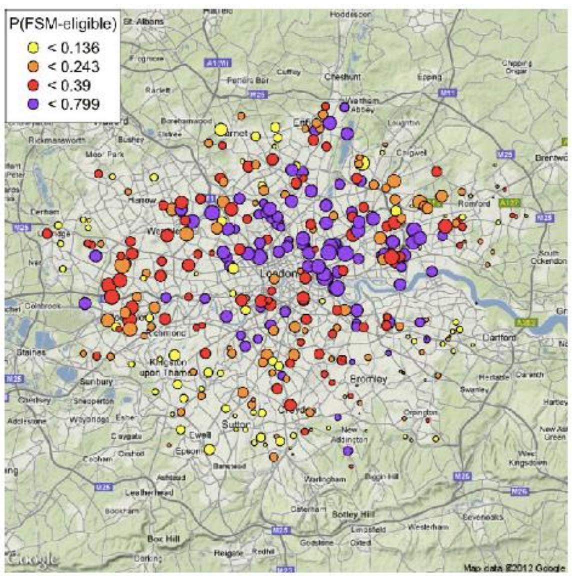
Green: Equi-log ($\frac{end}{start} = 2$)

Data Type - example

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Transaction data

Ordered data



Spatial data

	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

Document data

Types of data

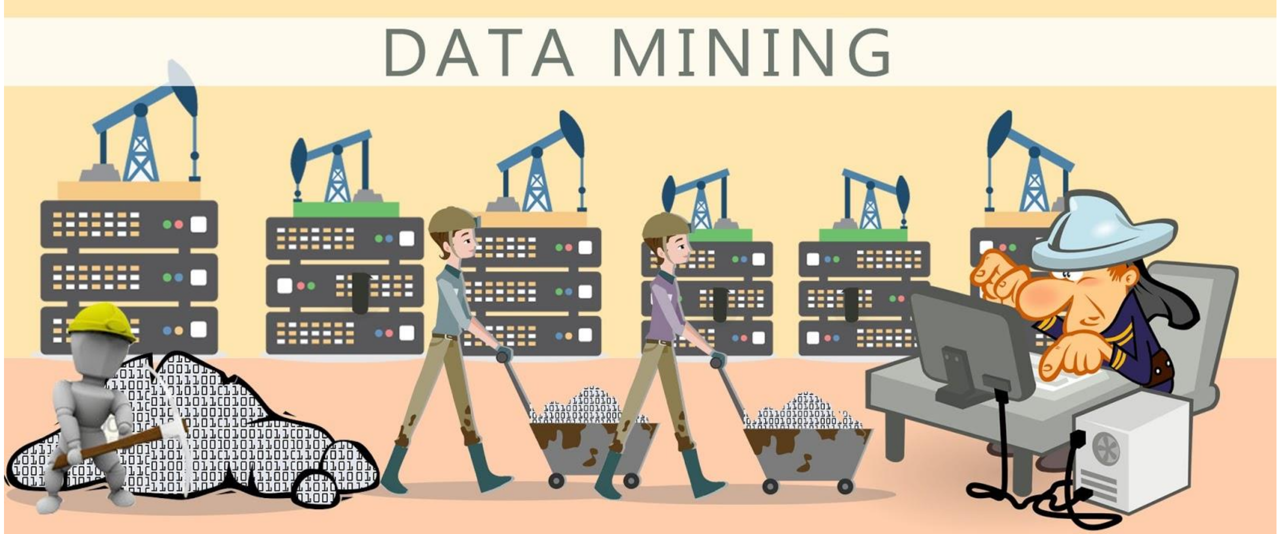
- Numeric data: each object is a point in a multidimensional space
- Categorical data: each object is a vector of categorical values
- Set data: each object is a set of values (with or without counts)
  - sets can also be represented as binary vectors, or vectors of counts
- Ordered sequences: each object is an ordered sequence of values.
- Graph data

Why data mining?

- **Commercial point of view** (companies – Facebook, Google, etc.)
  - data has become the key advantage of companies
  - being able to extract useful information out of the data is key for exploiting them commercially
- **Scientific point of view**
  - unprecedented position - collect TB of information (Sensor data, astronomy data, social network data, gene data)
  - we need the tools to analyze such data to get a better understanding of the world and advance science

Data mining usage

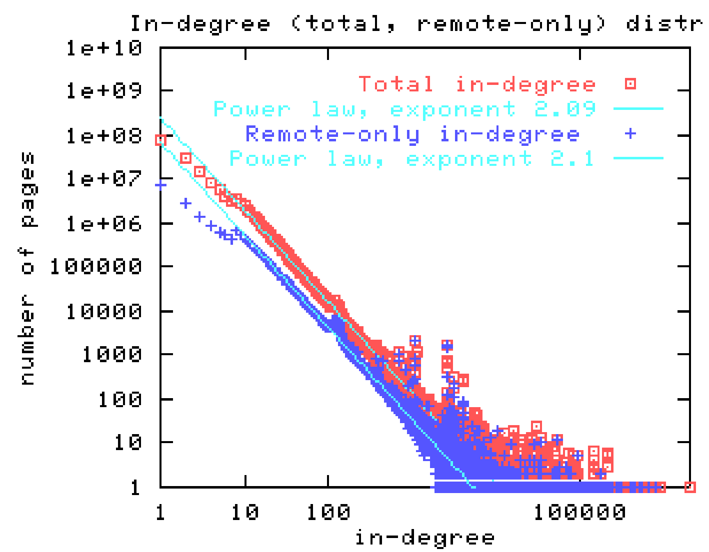
- Some usage of data mining:
  - frequent item sets (text mining, recommendations)
  - association Rules extraction
  - exploratory analysis
  - similarities
  - clustering
  - classification
  - ranking



Exploratory analysis

- Make measurements to understand what the data looks like
  - example: Posts
    - How often do users posts, how many posts per user, when do they post, is there a correlation between number of posts and number of friends, etc.
- This is one of the first steps when collecting data
  - metrics: **deciding what to measure is important**

The example of the Web graph



Exploiding similarities

- Consider the following data for six users:
  - number of times they have clicked on posts from these pages

What conclusion can we draw?

	NBA	ESPN	Sports.com	MSNBC	NY Times	Wall Street	Politico
A	100	50	73	10	1	1	4
B	500	200	400	20	10	4	1
C	80	100	60	1	3	1	1
D	4	2	1	12	90	100	80
E	9	3	4	9	100	80	70
F	3	4	5	30	300	200	500

How do we compute similarity?
How do we group similar users? Clustering

Making predictions

- filling the missing value can also be viewed as a prediction task
- types of prediction tasks:
  - predicting a real value: Regression
  - predicting a YES/NO value: Binary classification
  - predicting over multiple classes: Classification
- Can you think of prediction/classification tasks for your social network?

Ad click prediction

Like prediction

Predict if a post is offensive

Predict if a photo contains nudity

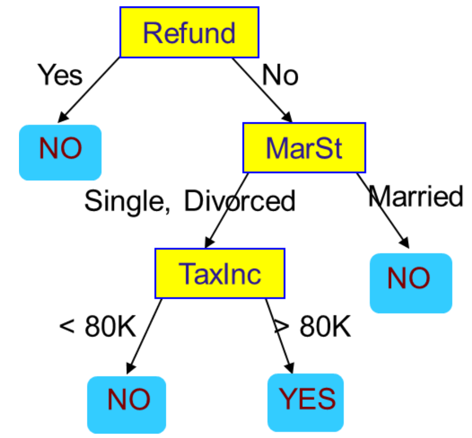
Ad clickthrough prediction

Predict if a user will like a post over another: Learning to rank

Classification

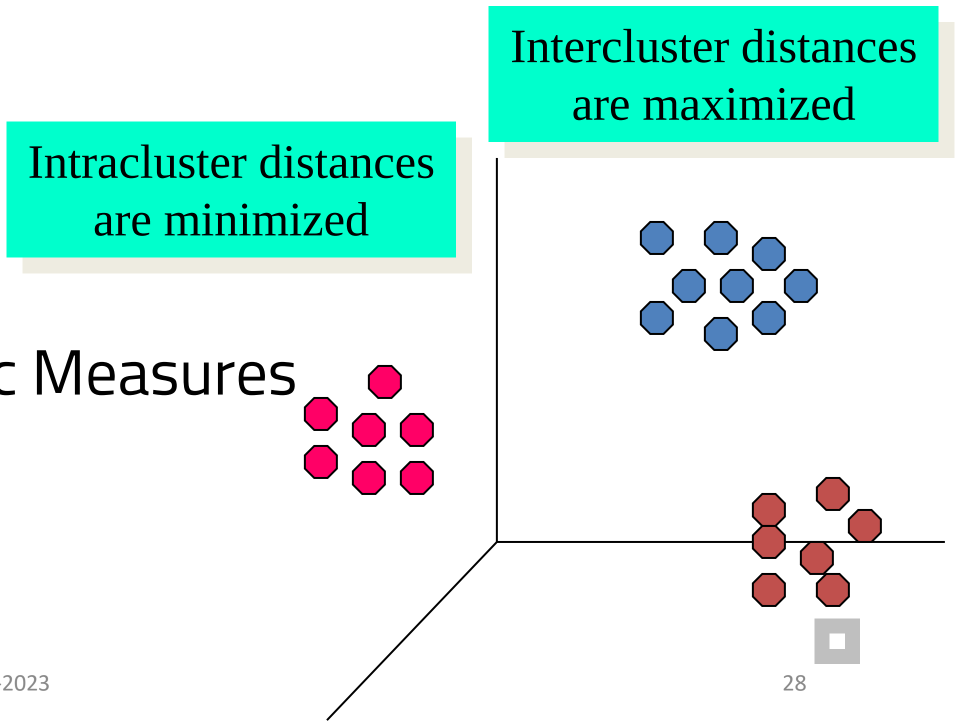
- Classification process:
  - find features that describe an entity
  - use examples of the classes to predict
  - learn a model (function) that predicts
- **Classification is the engine behind the AI revolution**
  - used in all systems that make decisions
  - became very powerful with Deep Learning
  - huge applications in vision

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



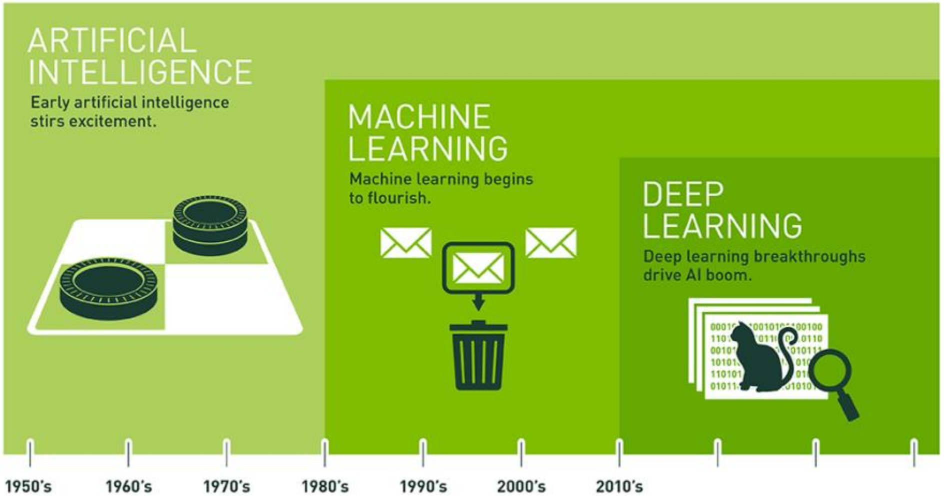
Clustering

- given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that:
  - Data points in one cluster are more similar to one another
  - Data points in separate clusters are less similar to one another
  - Similarity Measures?
    - Euclidean Distance
    - Other Problem-specific Measures



Deep Learning

- Machine learning systems that use neural networks with multiple layers and are trained on very large quantities of data
  - able to learn complex representations and powerful models
  - applications in recommendations, network analysis, text analysis, image recognition, car driving, playing games, etc.

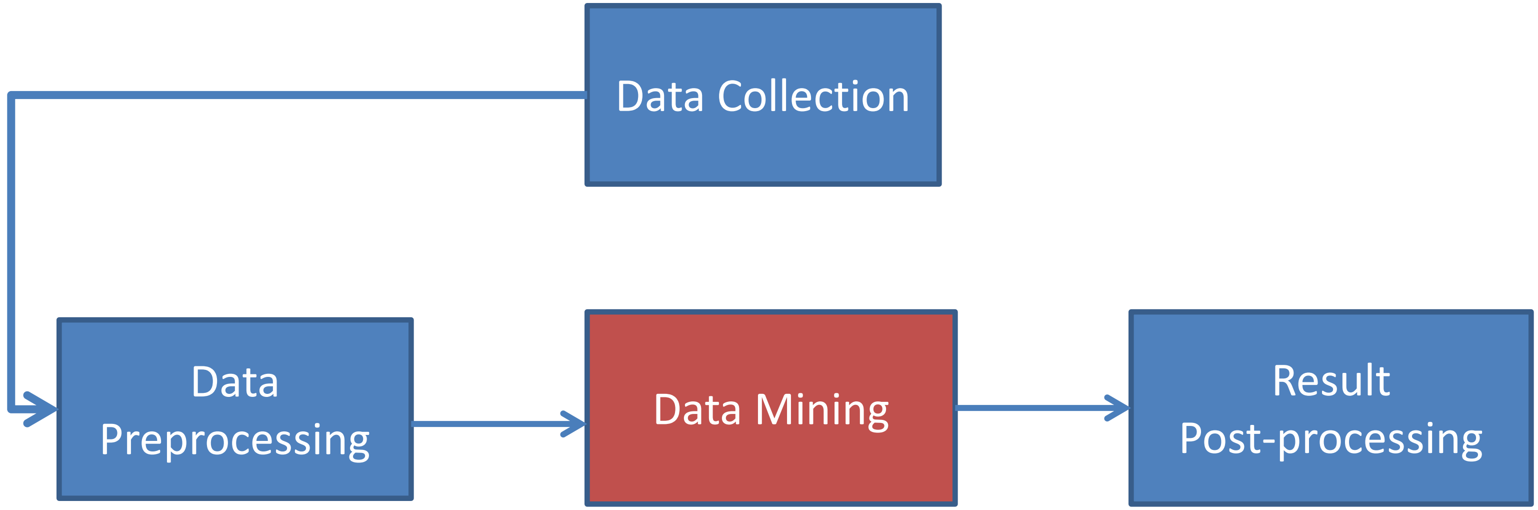


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

- require less feature engineering

Data Mining pipeline

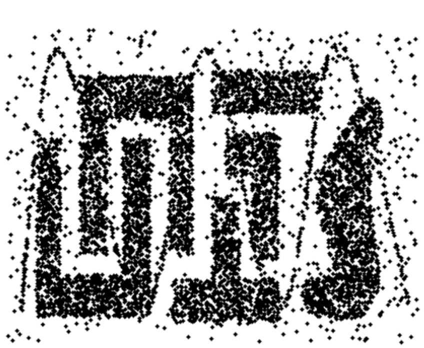
- mining is not the only step in the analysis process
- the data mining part is about the analytical methods and algorithms for extracting useful knowledge from the data
- Pre- and Post-processing are often data mining tasks as well



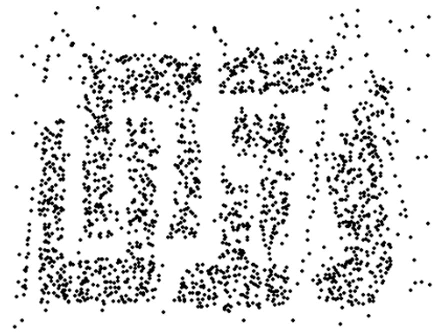
Data collection

- sampling is the main technique employed for data selection
  - it is often used for both the preliminary investigation of the data and the final data analysis
- statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming
  - example: what is the average height of a person in Romania?
- sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming
  - example: We have 1M documents. What fraction of pairs has at least 100 words in common?

Sampling size

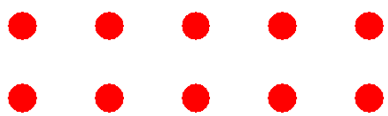


8000 points

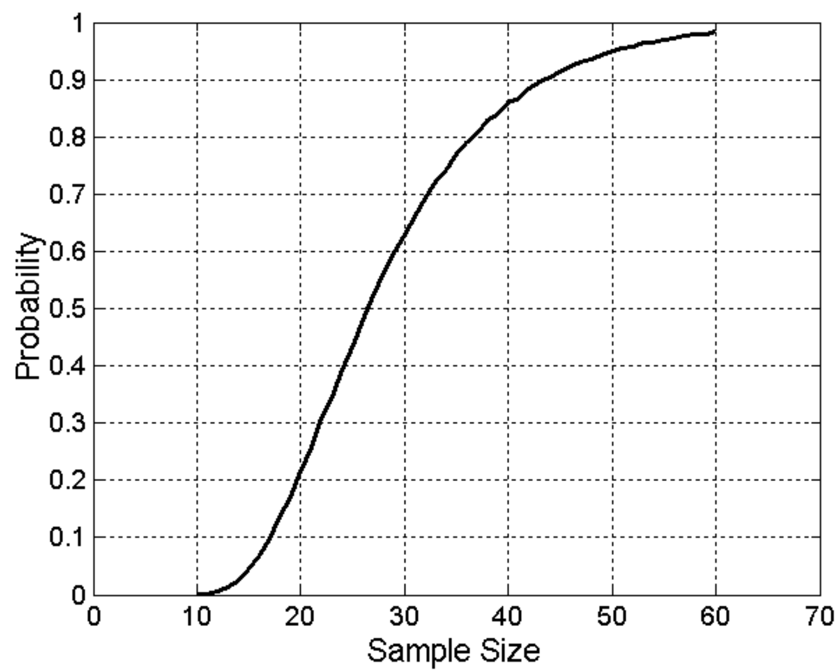


2000 Points

500 Points



What sample size is necessary to get at least one object from each of 10 groups



Feature extraction: Data cleaning

- we need to do **some cleaning**
- we need to extract some **features** to represent our data

Examples of data quality problems:

Noise and outliers

Missing values

Duplicate data

A mistake or a millionaire?

Missing values

Inconsistent duplicate entries

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	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
9	No	Single	90K	No

Feature Extraction

- the data we obtain are not necessarily as a relational table
- data may be in a very raw format
  - examples: text, speech, mouse movements, etc.
- we need to extract the features from the data
- feature extraction:
  - selecting the characteristics by which we want to represent our data
  - it requires some domain knowledge about the data
  - it depends on the application
- Deep learning: eliminates this step

Data normalization

- in many cases it is important to normalize the data
- the kind of normalization that we use depends on what we want to achieve
- Column normalization
  - subtract the minimum value and divide by the difference of the maximum value and minimum value for each attribute
  - brings everything in the [0,1] range, maximum is one, minimum is zero

Temperature	Humidity	Pressure
30	0.8	90
32	0.5	80
24	0.3	95

Temperature	Humidity	Pressure
0.75	1	0.33
1	0.6	0
0	0	1

new value = (old value – min column value) / (max col. value – min col. value)

Row normalization

- divide by the sum of values for each document (row in the matrix)
- transform a vector into a **distribution** *

	Word 1	Word 2	Word 3
Doc 1	28	50	22
Doc 2	12	25	13

Are these documents similar?

	Word 1	Word 2	Word 3
Doc 1	0.28	0.5	0.22
Doc 2	0.24	0.5	0.26

new value = old value / Σ old values in the row

* for example, the value of cell (Doc1, Word2) is the probability that a randomly chosen word of Doc1 is Word2

Row normalization

- Do these two users rate movies in a similar way?
- subtract the mean value for each user (row) – centering of data
- captures the deviation from the average behavior

	Movie 1	Movie 2	Movie 3
User 1	1	2	3
User 2	2	3	4

	Movie 1	Movie 2	Movie 3
User 1	-1	0	+1
User 2	-1	0	+1

new value = (old value – mean row value) [/ (max row value –min row value)]

Post processing

- Visualization
  - the human eye is a powerful analytical tool !!
- if we visualize the data properly, we can discover patterns and demonstrate trends
- visualization - present the data so that patterns can be seen
  - e.g., histograms and plots are a form of visualization
  - there are multiple techniques (a field on its own)

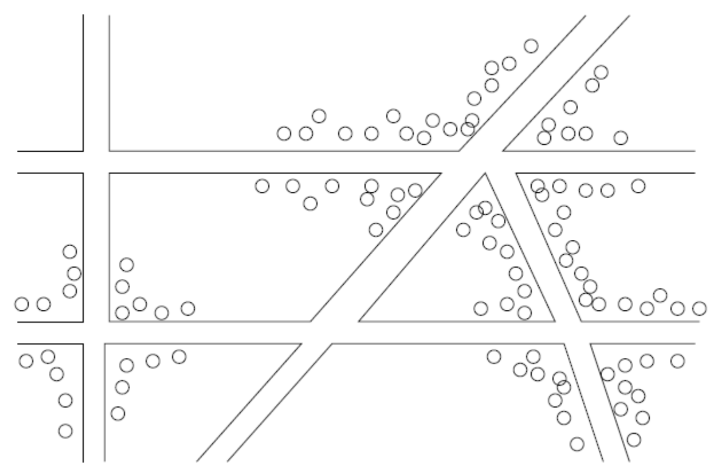


Figure 1.1: Plotting cholera cases on a map of London

Visualization on a map

Dimensionality reduction

- the human eye is limited to processing visualizations in two (at most three) dimensions
- one of the great challenges in visualization is to visualize high-dimensional data into a two-dimensional space
  - dimensionality reduction
  - distance preserving embeddings
- Dimensionality reduction is also a **preprocessing** technique:
  - reduce the amount of data
  - extract the useful information

Dimensionality reduction

- Consider the following 6-dimensional dataset

Each row is a multiple of two **vectors**

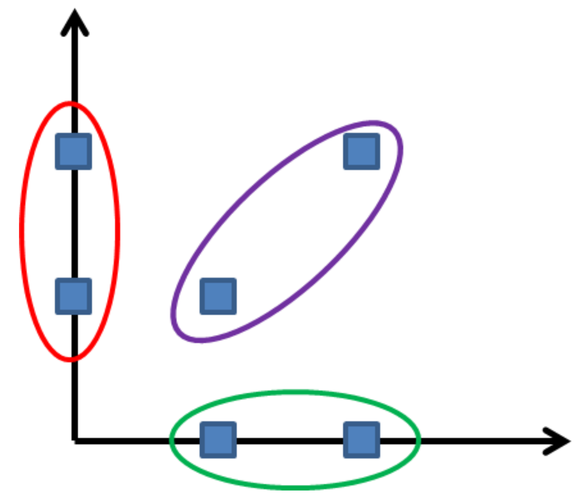
$$x = [1, 2, 3, 0, 0, 0]$$
$$y = [0, 0, 0, 1, 2, 3]$$

- $$D = \begin{bmatrix} 1 & 2 & 3 & 0 & 0 & 0 \\ 2 & 4 & 6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 2 & 4 & 6 \\ 1 & 2 & 3 & 1 & 2 & 3 \\ 2 & 4 & 6 & 2 & 4 & 6 \end{bmatrix}$$

What do you observe? Can we reduce the dimension of the data?

We can rewrite D as:

$$D = \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 0 & 1 \\ 0 & 2 \\ 1 & 1 \\ 2 & 2 \end{bmatrix}$$



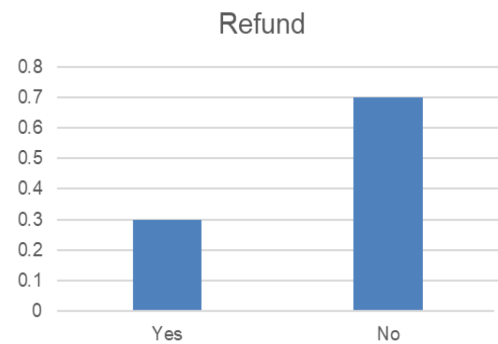
Exploratory Data Analysis

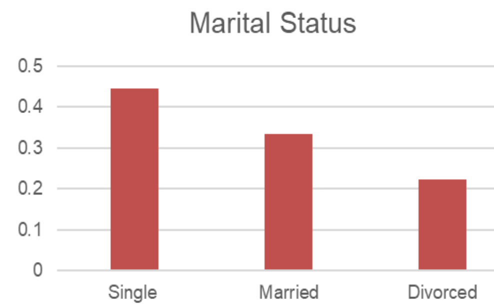
- Summary statistics: numbers that summarize data properties
- Summarized properties include frequency, location and spread
  - examples: location - mean

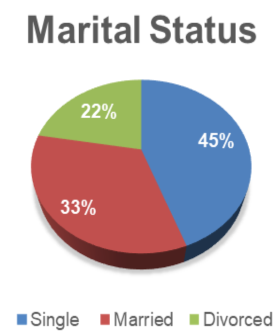
    spread - standard deviation
- the frequency of an attribute value is the percentage of time the value occurs in the data set
  - for example, given the attribute 'gender' and a representative population of people, the gender 'female' occurs about 50% of the time.
- the mode of an attribute is the most frequent attribute value
- we can visualize the data frequencies using a value histogram

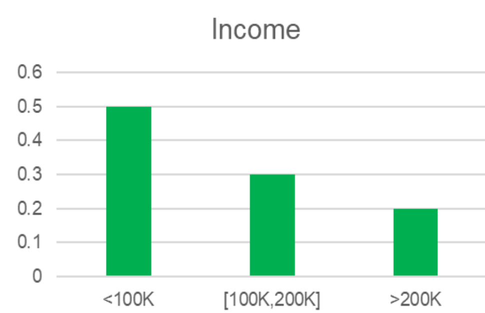
Example

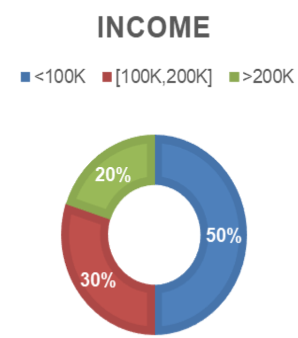
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
10	No	Single	90K	No











Mode: Single

Single	Married	Divorced	NULL
4	3	2	1

Single	Married	Divorced
44%	33%	22%

We can choose to ignore NULL values

Universitatea Transilvania din Brașov
FACULTATEA DE INGINERIE ELECTRICĂ ȘI ȘTIINȚA CALCULATOARELOR

Percentiles

- for continuous data, the notion of a percentile is more useful
- given an ordinal or continuous attribute *x* and a number *p* between 0 and 100, the *p*th percentile is a value x_p of x such that *p*% of the observed values of x are less or equal than x_p
- for instance, the 80th percentile is the value $x_{80\%}$ that is greater or equal than 80% of all the values of x we have in our data.

Taxable Income
10000K
220K
125K
120K
100K
90K
90K
85K
70K
60K

$x_{80\%}$ = 125K

Universitatea Transilvania din Brașov
FACULTATEA DE INGINERIE ELECTRICĂ ȘI ȘTIINȚA CALCULATOARELOR

Mean and median

- the **mean** is the most common measure of the location of a set of points. $\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$
- the **median** is also commonly used

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

- **Trimmed mean**: the mean after removing min and max values

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	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
10	No	Single	90K	No

Mean: 1090K

Trimmed mean (remove min, max): 105K

Median: (90+100)/2 = 95K

Attribute relations

- in many cases it is interesting to look at two attributes together to understand if they are correlated
  - e.g., How does your marital status relate with tax cheating?
  - e.g., Does refund correlate with average income?
  - Is there a relationship between years of study and income?

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	10000K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	90K	No
10	No	Single	90K	No

- How do we visualize these relationships?

Confusion Matrix

	No	Yes
Single	2	1
Married	4	0
Divorced	1	1

Joint Distribution Matrix

	No	Yes
Single	0.2	0.1
Married	0.4	0.0
Divorced	0.1	0.1

Confidence and error

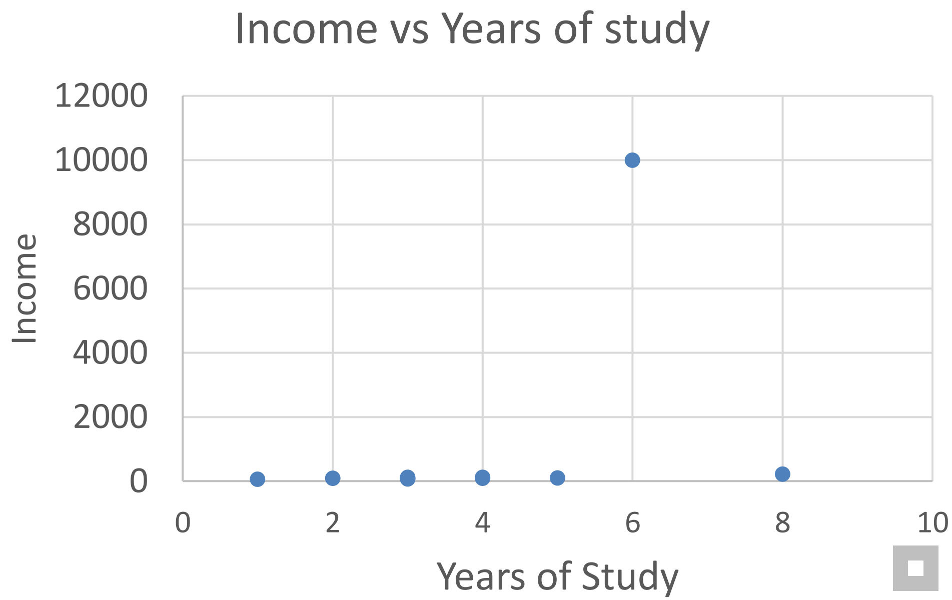
- we have a set of measurements X_i of incomes and we estimate the average income as:
  - $\hat{\mu} = \frac{1}{n}\sum_i X_i$
- the p-confidence interval of the value μ is an interval of values C_n such that: $P(\mu \in C_n) \geq p$
  - we usually ask for the 95% confidence interval
- if we have a measurement $\hat{\theta}$ that we estimate from the data, the standard error is defined as $se = \sqrt{Var(\hat{\theta})}$
- in our case our measurement is the average income which we estimate as:
  - $\hat{\mu} = \frac{1}{n}\sum_i X_i$

Correlating numerical attributes

- Scatter plot:
  - X axis is one attribute, Y axis is the other

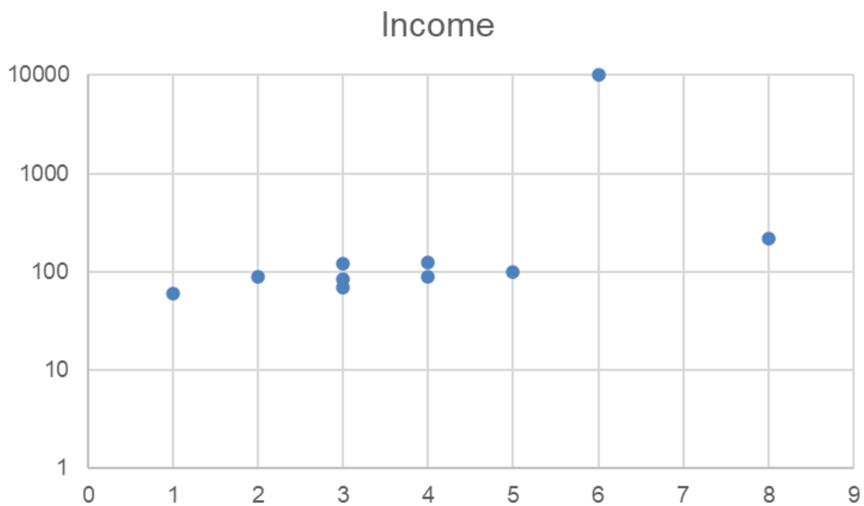
Tid	Refund	Marital Status	Taxable Income	Years of Study
1	Yes	Single	125K	4
2	No	Married	100K	5
3	No	Single	70K	3
4	Yes	Married	120K	3
5	No	Divorced	10000K	6
6	No	NULL	60K	1
7	Yes	Divorced	220K	8
8	No	Single	85K	3
9	No	Married	90K	2
10	No	Single	90K	4

- for each entry we have two values
- plot the entries as two-dimensional points

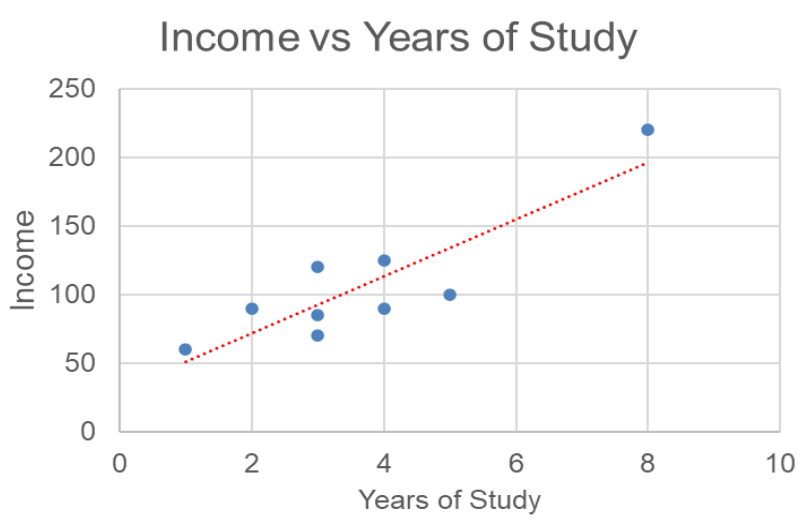


Correlating numerical attributes

- Log-scale in y-axis makes the plot look a little better



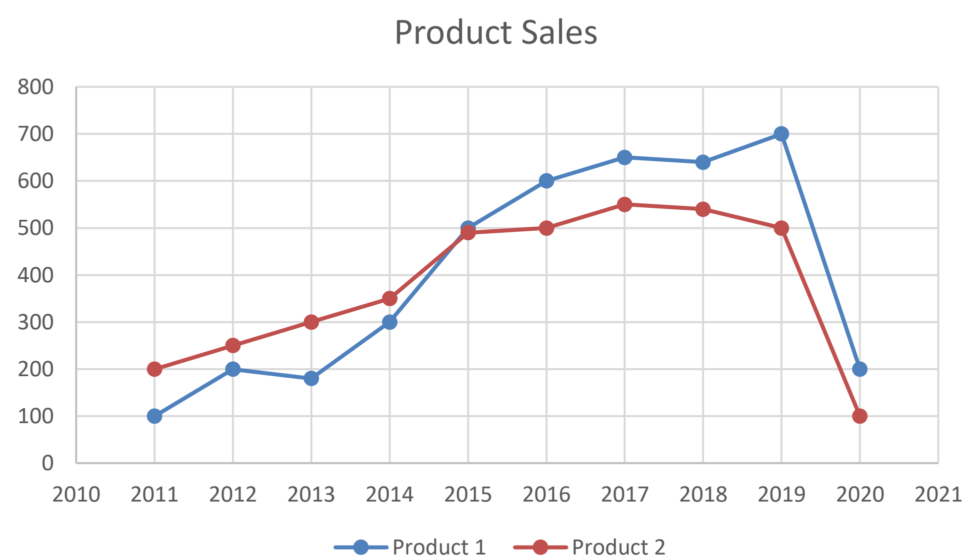
- **After removing the outlier value** there is a clear correlation



Plotting attributes

Year	Product 1	Product 2
2011	100	200
2012	200	250
2013	180	300
2014	300	350
2015	500	490
2016	600	500
2017	650	550
2018	640	540
2019	700	500
2020	200	100

- How would you visualize the differences between the product sales over time?



QUESTIONS ?