



Data Science

Ubai Sandouk, PhD

2021



Recap – DataOps



Data Storage + Management + Analysis

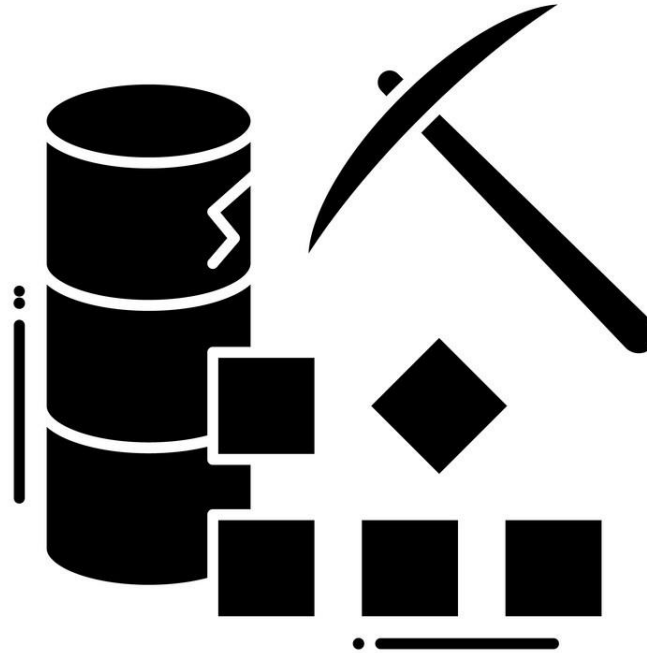
Big Data + Data Quality + **Data Mining**

Agenda



- Types of Data Analysis
- KDD Process
- Data Description
- Data Exploration
- Data Prediction
- Other Data Tasks
- References

Data Mining



DATA MINING

www.vectorstock.com

Types of Data Analysis

- Six types of Data Analysis:
 - Descriptive
 - Exploratory
 - Predictive
 - Inferential
 - Causal
 - Mechanistic

Data Mining

Data Mining

- Extracting non-trivial, implicit, previously unknown and potentially useful patterns of information from huge amounts of data.
- A.K.A.: Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, etc.
- Making sense of the data.
- Data could be structured or unstructured, with different complications.



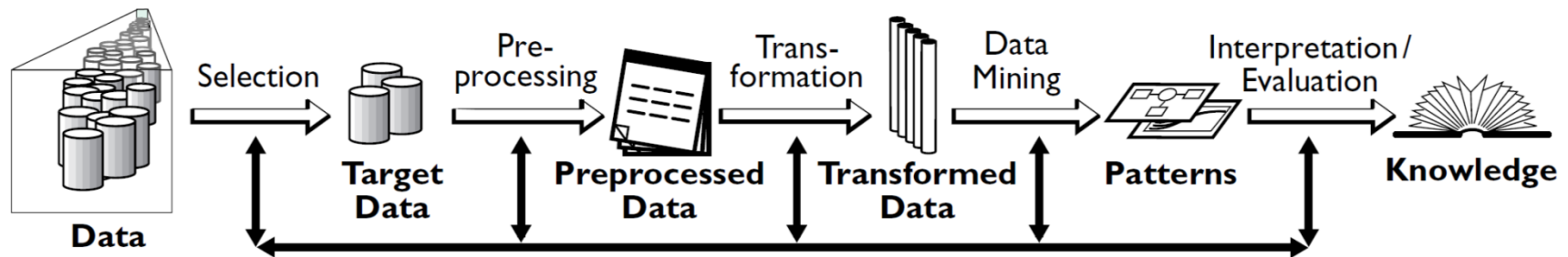
U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM. 1996.

Data Mining



- *KDD is defined as:*
“The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.”

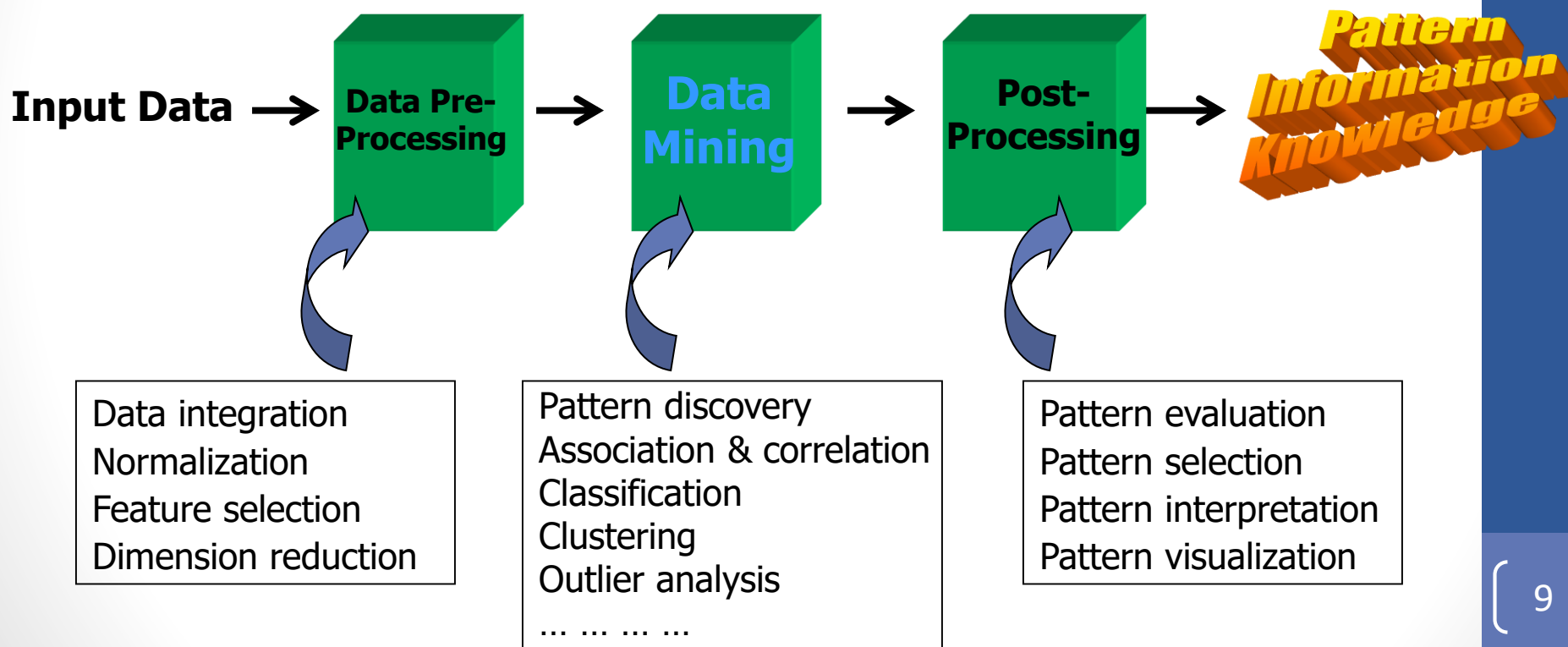
Figure 1. Overview of the steps constituting the KDD process



U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM. 1996.

KDD Process

- A Typical View from ML and Statistics



KDD Process



- Evaluation of Knowledge
 - Are all mined knowledge interesting?
 - One can mine tremendous amount of “patterns” and knowledge
 - Some may fit only certain dimension space (time, location, ...)
 - Some may not be representative, may be transient, ...
 - Evaluation of mined knowledge → directly mine only interesting knowledge?
 - Coverage
 - Typicality vs. novelty
 - Accuracy
 - Timeliness
 - Etc...

Data Mining - Bonferroni's



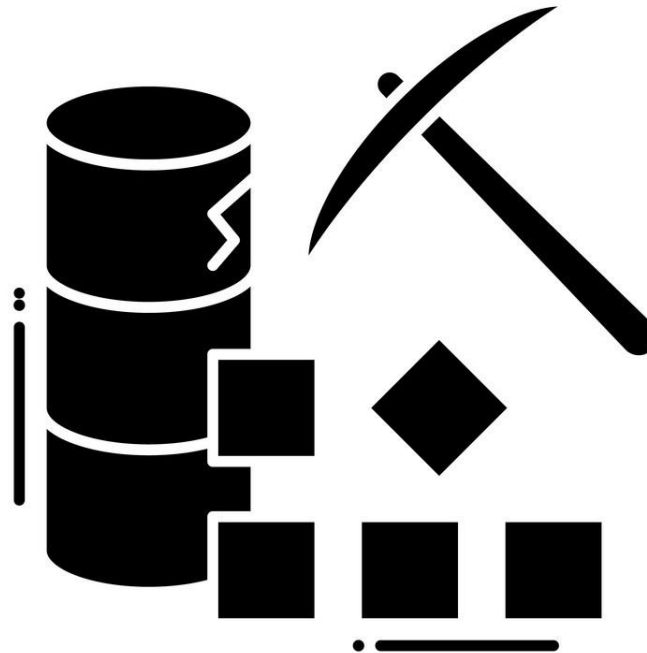
- An analyst can “discover” patterns that are meaningless!
- Bonferroni's principle:
 - if you look in more places for interesting patterns than your amount of data will support, you are bound to find nonsense!
- Find (unrelated) pairs of people who at least twice have stayed at the same hotel on the same day
 - 10^7 people being tracked, for a 1,000 days
 - Number of pairs of people available = $0.5 * 10^7 * 10^7 = 5 * 10^{13}$
 - Number of pairs of days available = $0.5 * 10^3 * 10^3 = 5 * 10^5$
 - Each person stays in a hotel 1% of time (1 day out of 100)
 - Hotels hold 100 people (so 10^3 hotels)
 - P of two people visiting any hotel on a day = 10^{-4}
 - P of two people visiting the same hotel on a day = 10^{-7}
 - P of two people visiting the same hotel on two days = 10^{-14}
 - Number of suspicious activities = $5 * 10^{13} * 5 * 10^5 * 10^{-14} = 250,000!!$



Data Mining: Multi-Dimensional View

- Data to be mined
 - Database data, data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, social and information networks
- Knowledge to be extracted
 - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
 - Descriptive vs. predictive data mining
- Techniques utilized
 - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, etc.
- Applications adapted
 - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Data Description



DATA MINING

www.vectorstock.com

Types of Data Sets

- Record
 - Relational records
 - Data matrix, e.g., numerical matrix, crosstabs
 - Document data: text documents: term-frequency vector
 - Transaction data
- Graph and network
 - World Wide Web
 - Social or information networks
 - Molecular Structures
- Ordered
 - Video data: sequence of images
 - Temporal data: time-series
 - Sequential Data: transaction sequences
 - Genetic sequence data
- Spatial, image and multimedia:
 - Spatial data: maps
 - Image data:
 - Video 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

Characteristics of Data

- Dimensionality
 - Curse of dimensionality
- Sparsity
 - Only presence counts
- Resolution
 - Patterns depend on the scale
- Distribution
 - Centrality and dispersion

Characteristics of Data

- **Data Objects:**
- Data sets are made up of data objects.
 - A data object represents an entity.
- Examples:
 - sales database: customers, store items, sales
 - medical database: patients, treatments
 - university database: students, professors, courses
- Also called samples, examples, instances, data points, objects, tuples.
 - Data objects are described by attributes.
 - Database rows -> data objects; columns -> attributes.

Characteristics of Data

- **Data Attributes:**
- Dimension, feature, variable, data field:
 - Representing a characteristic or feature of a data object.
- Types:
 - Nominal
 - Binary
 - Numeric: quantitative
 - Interval-scaled
 - Ratio-scaled

Attribute Types

- Nominal
 - Categories, States, or “names of things”
 - Hair_color = {auburn, black, blond, brown, grey, red, white}
 - marital status, occupation, ID numbers, zip codes
- Binary
 - Nominal attribute with only 2 states (0 and 1)
 - Symmetric binary: both outcomes equally important
 - e.g., gender
 - Asymmetric binary: outcomes not equally important.
 - e.g., medical test (positive vs. negative)
 - Convention: assign 1 to most important outcome (e.g., HIV positive)
- Ordinal
 - Values have a meaningful order (ranking) but magnitude between successive values is not known.
 - *Size = {small, medium, large}*, grades, army rankings

Attribute Types

- Interval
 - Measured on a scale of equal-sized units
 - Values have order
 - No true zero-point
- Ratio
 - Inherent zero-point
 - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K° is twice as high as 5 K°).



Attribute Types

- Discrete Attribute
 - Has only a finite or countably infinite set of values
 - Sometimes, represented as integer variables
 - Note: Binary attributes are a special case of discrete attributes
- Continuous Attribute
 - Has real numbers as attribute values
 - Practically, real values can only be measured and represented using a finite number of digits
 - Continuous attributes are typically represented as floating-point variables

Attribute Statistics

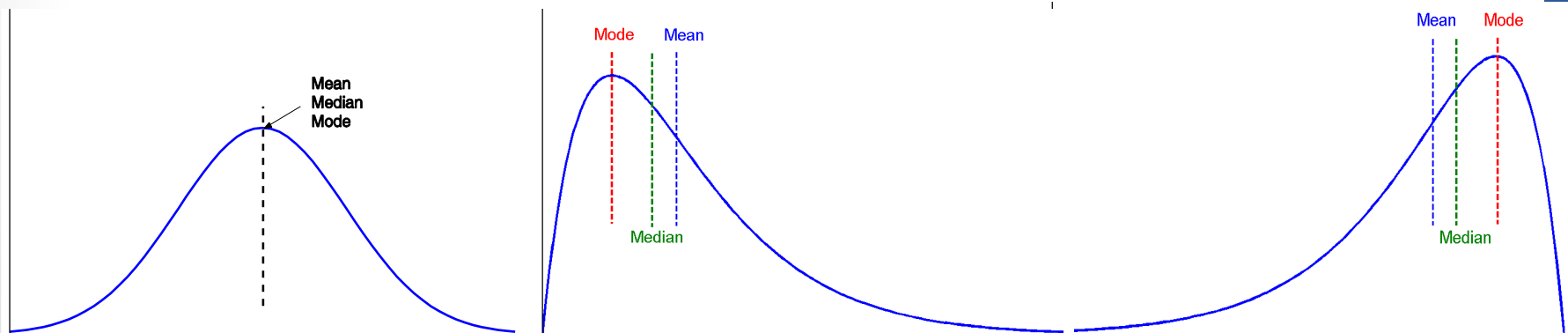
- Data dispersion characteristics
 - analyzed with multiple granularities of precision
 - median, max, min, quantiles, outliers, variance, etc.
 - Boxplot or quantile analysis on sorted intervals

Attribute Statistics

- Mean: n is sample size and N is population size.
 - Weighted arithmetic mean
- Median:
 - Middle value if odd number of values,
 - or average of the middle two values otherwise
- Mode
 - Value that occurs most frequently in the data
 - Unimodal, bimodal, trimodal

$$\mu = \frac{\sum x}{N}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$



Attribute Statistics

- Quartiles: Q1 (25th percentile), Q3 (75th percentile)
- Five number summary: min, Q1, median, Q3, max
- Variance and standard deviation (population: σ , sample: s)

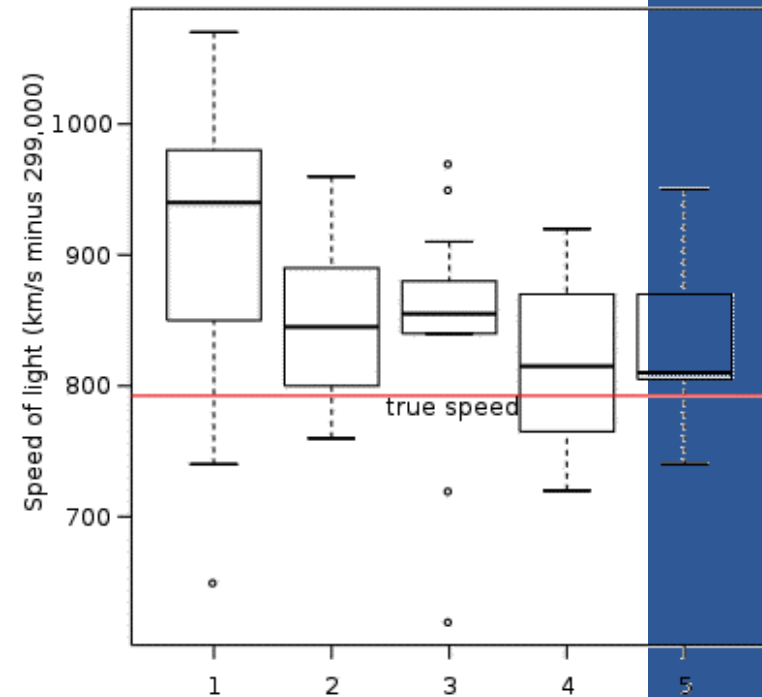
$$\sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right]$$

- Standard deviation s (or σ) is the square root of variance

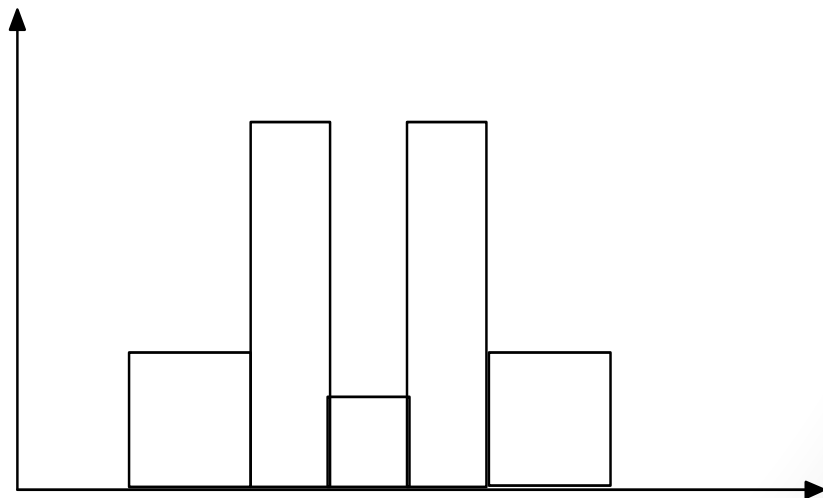
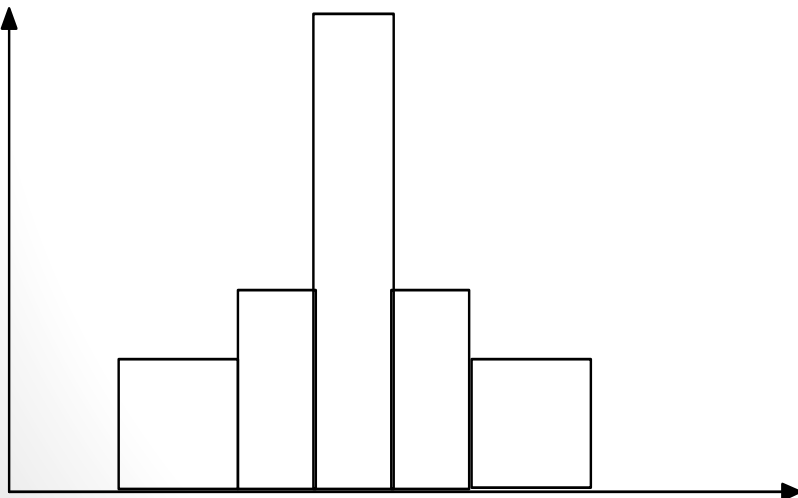
Attribute Statistics

- Five-number summary of a distribution
 - Minimum, Q1, Median, Q3, Maximum
- Boxplot
 - Data is represented with a box
 - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
 - The median is marked by a line within the box
 - Whiskers: two lines outside the box extended to Minimum and Maximum
 - Outliers: points beyond a specified outlier threshold, plotted individually



Attribute Statistics

- Two histograms may have the same boxplot representation
- The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions





Scale And Standardization

- Scale

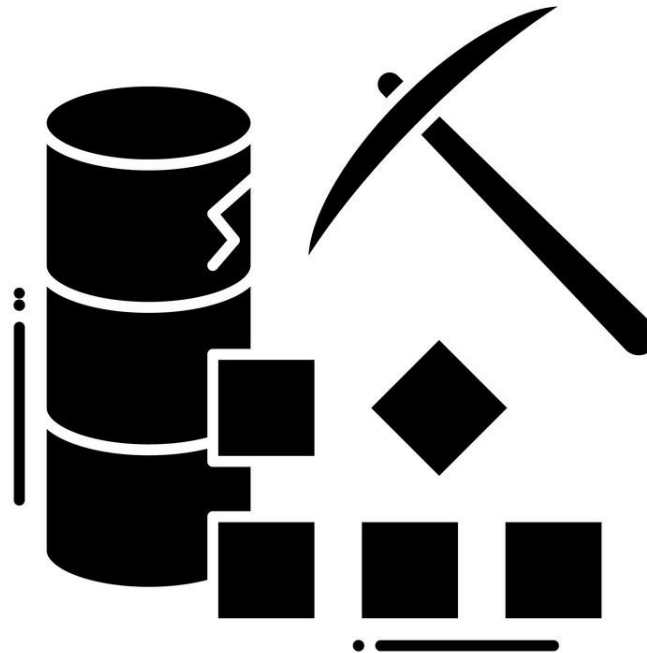
$$\bar{x} = \frac{x - \min(X)}{\max(X) - \min(X)}$$

- Z-score:

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

- X: raw score to be standardized, μ : mean of the population, σ : standard deviation
- the distance between the raw score and the population mean in units of the standard deviation
- negative when the raw score is below the mean, “+” when above

Data Exploration



DATA MINING

www.vectorstock.com

Similarity and Dissimilarity



- Similarity
 - Numerical measure of how alike two data objects are
 - Value is higher when objects are more alike
 - Often falls in the range $[0,1]$
- Dissimilarity (e.g., distance)
 - Numerical measure of how different two data objects are
 - Lower when objects are more alike
 - Minimum dissimilarity is often 0
 - Upper limit varies
- Proximity refers to a similarity or dissimilarity

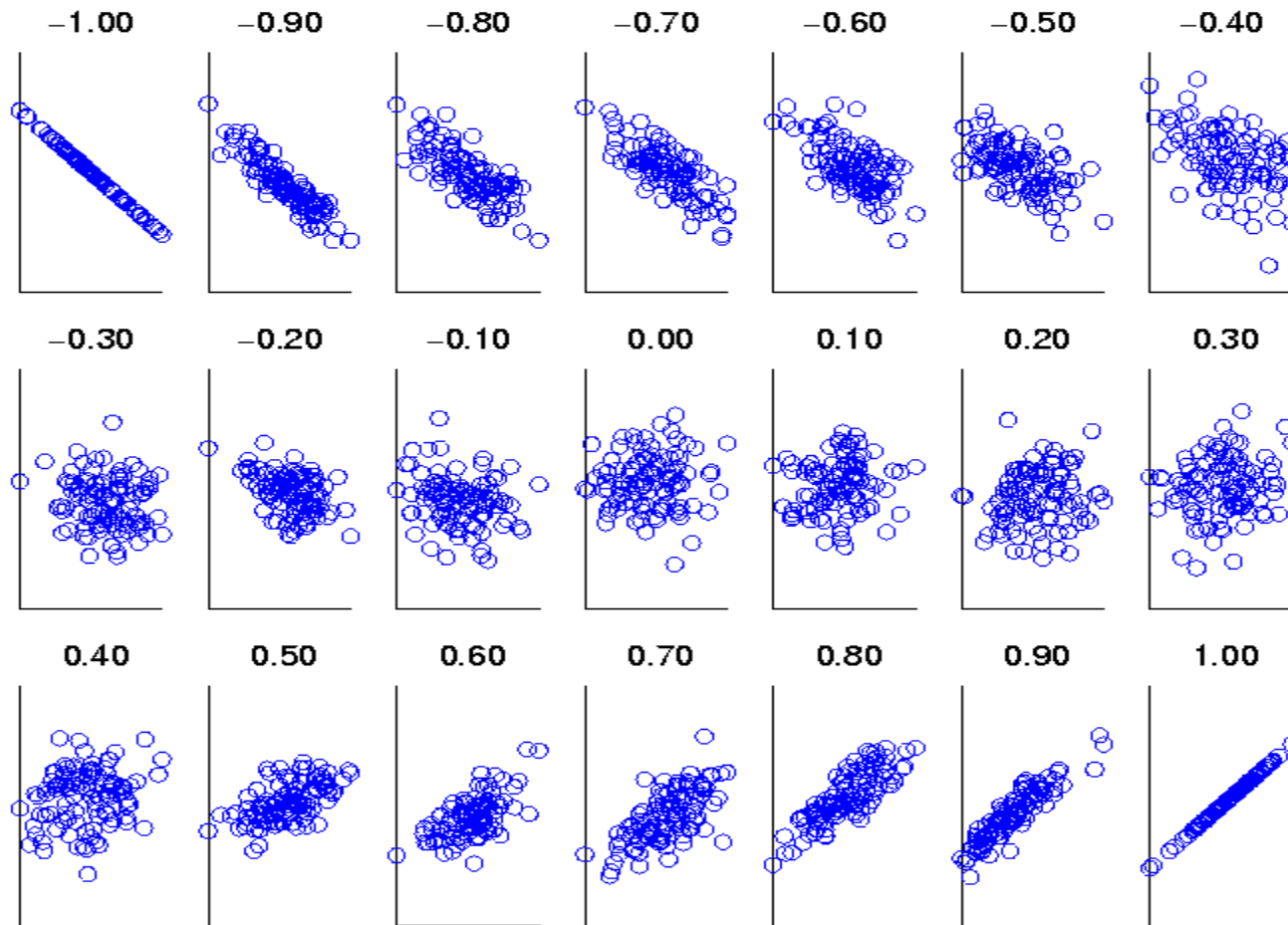
Correlation

- Correlation coefficient (also called Pearson's coefficient)

$$r_{A,B} = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum_{i=1}^n (a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

- where n is the number of tuples, \bar{A} and \bar{B} are the respective means of A and B , σ_A and σ_B are the respective standard deviation of A and B , and $\sum(a_i b_i)$ is the sum of the AB cross-product.
- If $r > 0$, A and B are positively correlated.
 - The higher, the stronger correlation.
- $r = 0$: independent
- $r < 0$: negatively correlated

Correlation

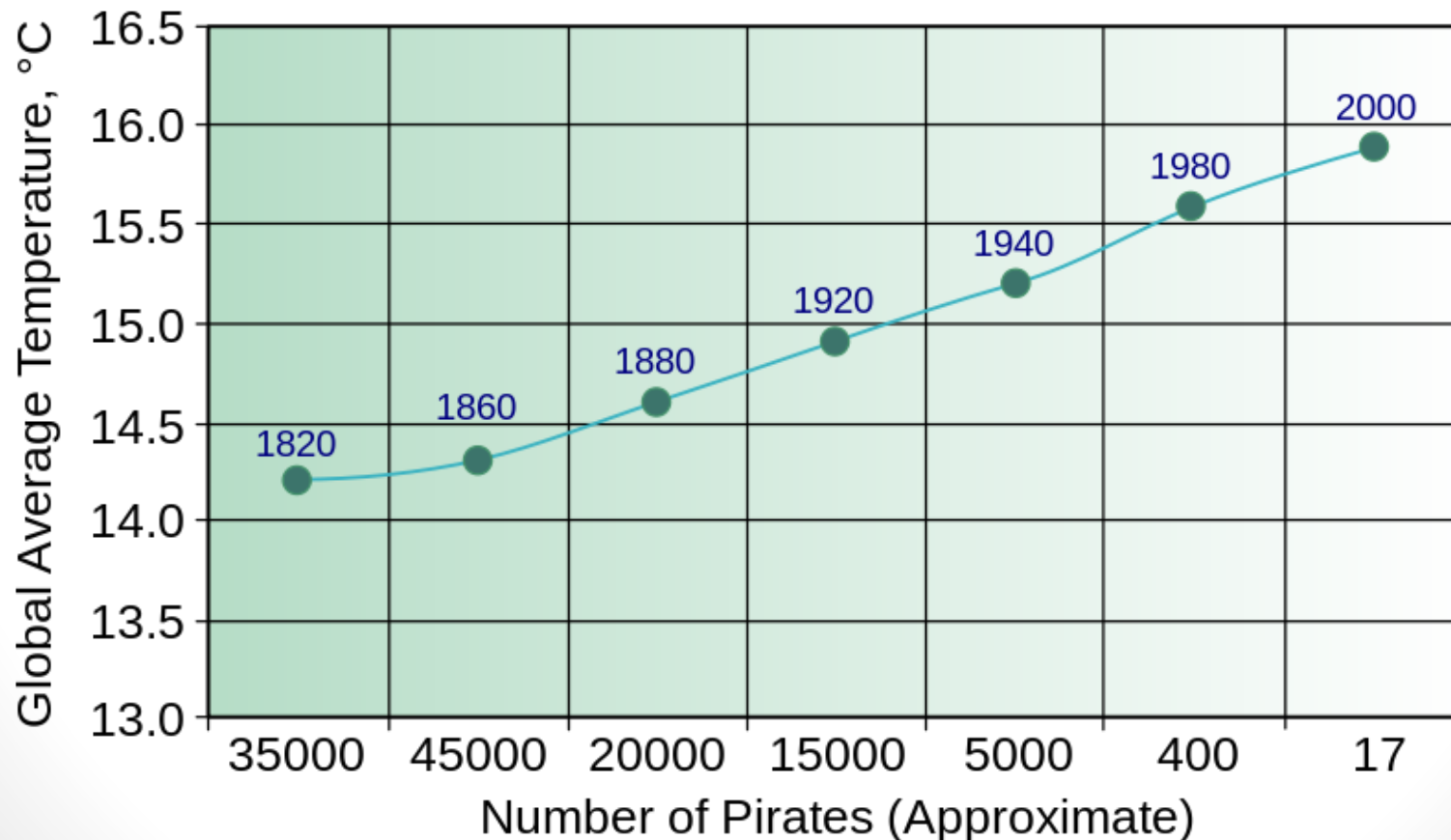


Scatter plots showing the similarity from -1 to 1 .

Correlation

- Correlation does not imply causality

Global Average Temperature vs. Number of Pirates



Correlation

- χ^2 (chi-square) test

$$\chi^2 = \sum \frac{(X_i - E[X_i])^2}{E[X_i]}$$

- The larger the χ^2 value, the more the variables are related
- The cells that contribute the most to the χ^2 value are those whose actual count is very different from the expected count

Correlation



	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- χ^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction and play_chess are correlated in the group

Association Analysis

- Generalised form of correlation.
- Finding association rules within the data
 - When features a , b , and c (*etc...*) have certain values or within certain ranges; then feature d will probably have a certain value.
- An example
 - $\{\text{Butter, Bread}\} \Rightarrow \{\text{Milk}\}$
- Accuracy measured by support, confidence and lift.

Association Analysis

- Rule: {Butter, Bread} => {Milk}

- $Support = \frac{1}{5} = 20\%$

- $Confidence = \frac{1}{1} = 100\%$

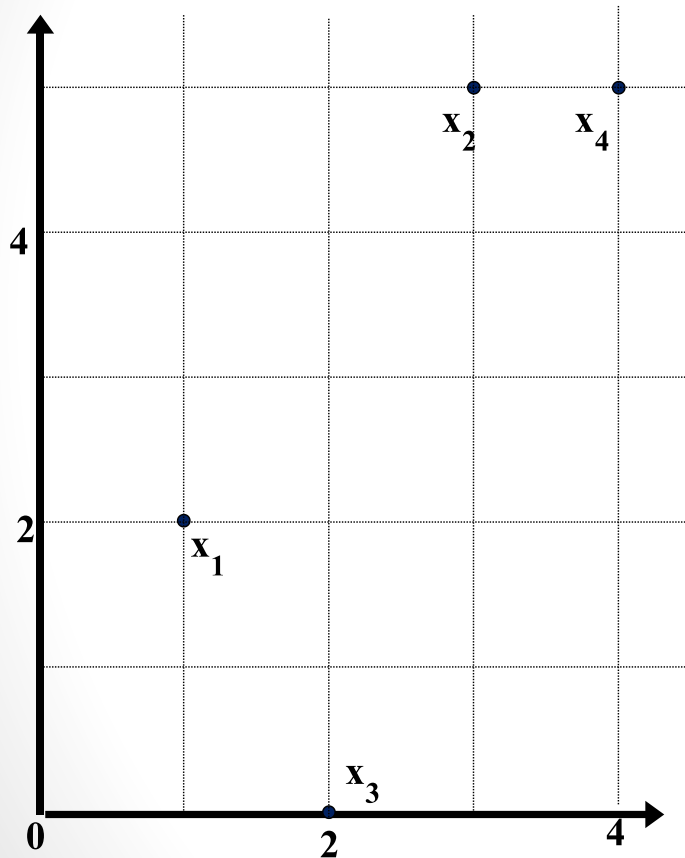
- $Lift = \frac{1}{2*1} = 50\%$

	A	B	C	D	E	F
1	ID	milk	bread	butter	beer	diapers
2	1	✓ 1	✓ 1	✗ 0	✗ 0	✗ 0
3	2	✗ 0	✗ 0	✓ 1	✗ 0	✗ 0
4	3	✗ 0	✗ 0	✗ 0	✓ 1	✓ 1
5	4	✓ 1	✓ 1	✓ 1	✗ 0	✗ 0
6	5	✗ 0	✓ 1	✗ 0	✗ 0	✗ 0
7						

https://en.wikipedia.org/wiki/Association_rule_learning

- Downward-closure* guarantees that for a frequent set, all its subsets are also frequent!
- Use **Apriori algorithm** to extract association rules.

Distance Measures



point	attribute1	attribute2
$x1$	1	2
$x2$	3	5
$x3$	2	0
$x4$	4	5

	$x1$	$x2$	$x3$	$x4$
$x1$	0			
$x2$	3.61	0		
$x3$	5.1	5.1	0	
$x4$	4.24	1	5.39	0

Distance Measures

- Minkowski distance: A popular distance measure

$$d(x, y) = \sqrt[h]{|x_1 - y_1|^h + |x_2 - y_2|^h + \dots + |x_p - y_p|^h}$$

- Properties
 - $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (Positive definiteness)
 - $d(i, j) = d(j, i)$ (Symmetry)
 - $d(i, j) \leq d(i, k) + d(k, j)$ (Triangle Inequality)
- A distance that satisfies these properties is a metric

Distance Measures

- $h = 1$: **Manhattan** distance (city block, L1 norm)
 - Hamming distance

$$d(i, j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \dots + |x_{i_p} - x_{j_p}|$$

- $h = 2$: **Euclidean** distance (L2 norm)

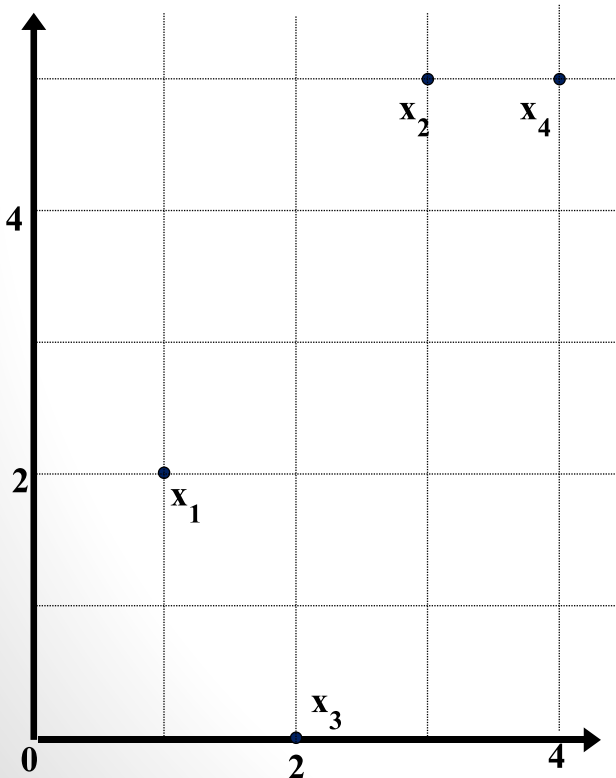
$$d(i, j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$$

- $h \rightarrow \infty$. **“supremum”** distance (L_{\max} norm).
 - This is the maximum difference between any component (attribute) of the vectors

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f |x_{if} - x_{jf}|$$

Distance Measures

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Distance Measures

- Cosine measure:

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| ||d_2|| ,$$

- \bullet indicates vector dot product
- $||d||$: the length of vector d
- One fewer dimension than Euclidean distance

Distance Measures

- Nominal Attributes: Generalization of a binary attribute
- Method 1: Simple matching
 - m: # of matches, p: total # of variables

$$d(i, j) = \frac{p - m}{p}$$

- Method 2: Use a large number of binary attributes
 - creating a new binary attribute for each of the M nominal states
- In asymmetric case: ignore dominant cases

Distance Measures

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank
- Can be treated like interval-scaled

- replace by their rank

$$r_{if} \in \{1, \dots, M_f\}$$

- map the range of each variable onto $[0, 1]$ by replacing i -th object in the f -th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- compute the dissimilarity using methods for interval-scaled variables

Distance Measures

- A database may contain all attribute types
 - Nominal, symmetric binary, asymmetric binary, numeric, ordinal
- One may use a weighted formula to combine their effects

$$d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

- f is binary or nominal:
 - $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise
- f is numeric: use the normalized distance
- f is ordinal
 - Compute ranks r_{if} and
 - Treat z_{if} as interval-scaled

Clustering



- Cluster: A collection of data objects
 - similar (or related) to one another within the same group
 - dissimilar (or unrelated) to the objects in other groups
- Categorising “similar” examples into groups, or clusters.
 - And separating “dissimilar” examples into differing groups.
- Similarity is introduced by a designer.
 - Most famously, the Euclidean distance.
 - Issues with Euclidean distance?

Clustering



- A good clustering method will produce high quality clusters
 - high **intra-class** similarity: cohesive within clusters
 - low **inter-class** similarity: distinctive between clusters
- The quality of a clustering method depends on
 - the similarity measure used by the method
 - its implementation, and
 - Its ability to discover some or all of the **hidden** patterns

Clustering

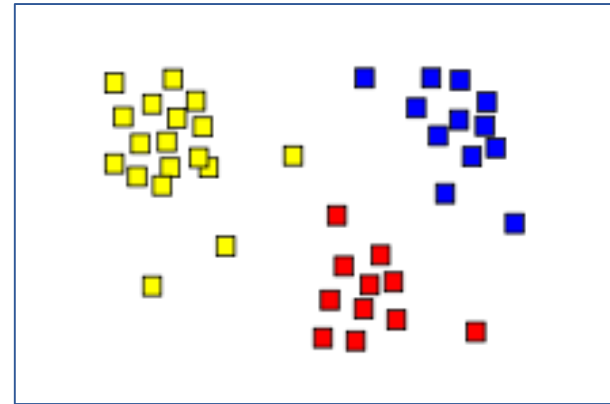


- Partitioning approach:
 - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
 - Typical methods: k-means
- Hierarchical approach:
 - Create a hierarchical decomposition of the set of data (or objects) using some criterion
 - Typical methods: BIRCH
- Density-based approach:
 - Based on connectivity and density functions
 - Typical methods: DBSCAN

Clustering



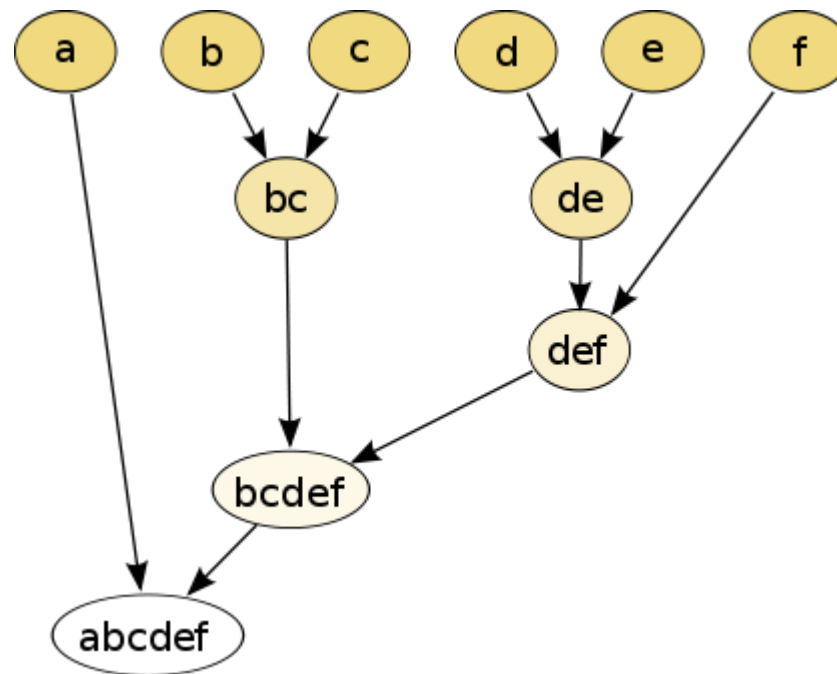
- Partition Clustering
- K-means clustering
 - Based on cluster centres
- Number of clusters is also introduced by a designer.



https://en.wikipedia.org/wiki/Cluster_analysis

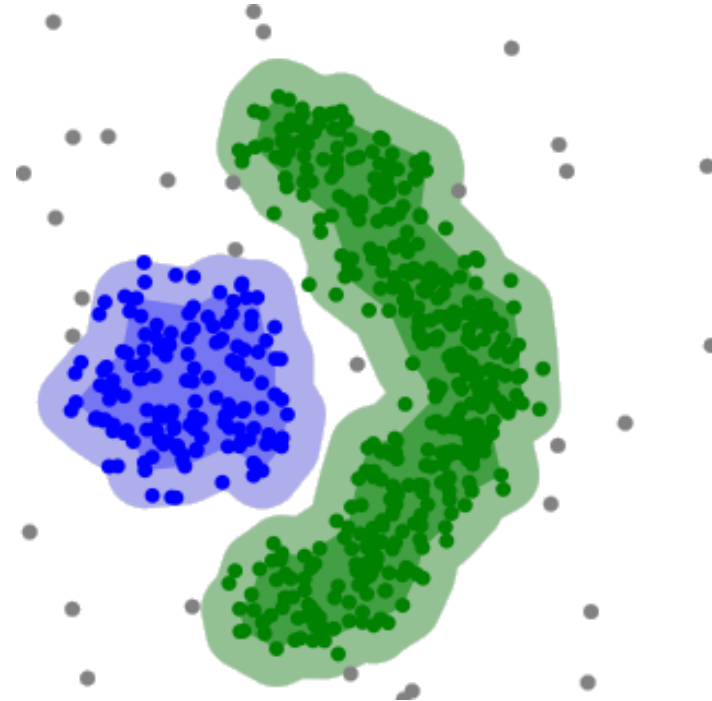
Clustering

- Hierarchical clustering
 - Results in dendrogram



Clustering

- Density-based clustering
 - clusters are defined as areas of higher density than the remainder of the data set
 - Possibly Soft clustering
 - Useful for feature extraction.

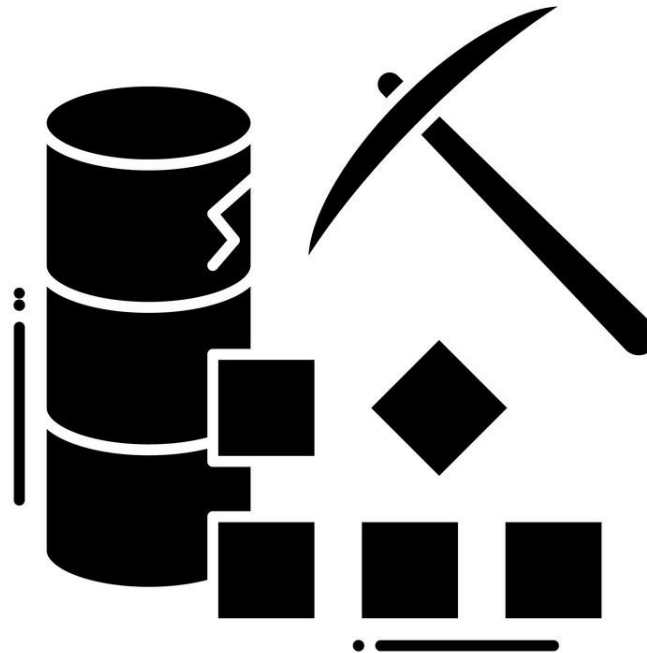


<https://en.wikipedia.org/wiki/DBSCAN>

Clustering Evaluation

- Two methods: extrinsic vs. intrinsic
 - Extrinsic: supervised, i.e., the ground truth is available
 - Compare a clustering against the ground truth using certain clustering quality measure
 - Ex. BCubed precision and recall metrics
 - Intrinsic: unsupervised, i.e., the ground truth is unavailable
 - Evaluate the goodness of a clustering by considering how well the clusters are separated, and how compact the clusters are
 - Ex. Silhouette coefficient

Data Prediction



DATA MINING

www.vectorstock.com



Classification

- Constructing a predictive model which can associate features of an example with its label, or class.
- A special case of Predictive Data Analysis.
- $f(\text{features}; \text{parameters}) \rightarrow \text{Label}$
- Care about the *model*, *learning algorithm* and the *cost/loss function*.

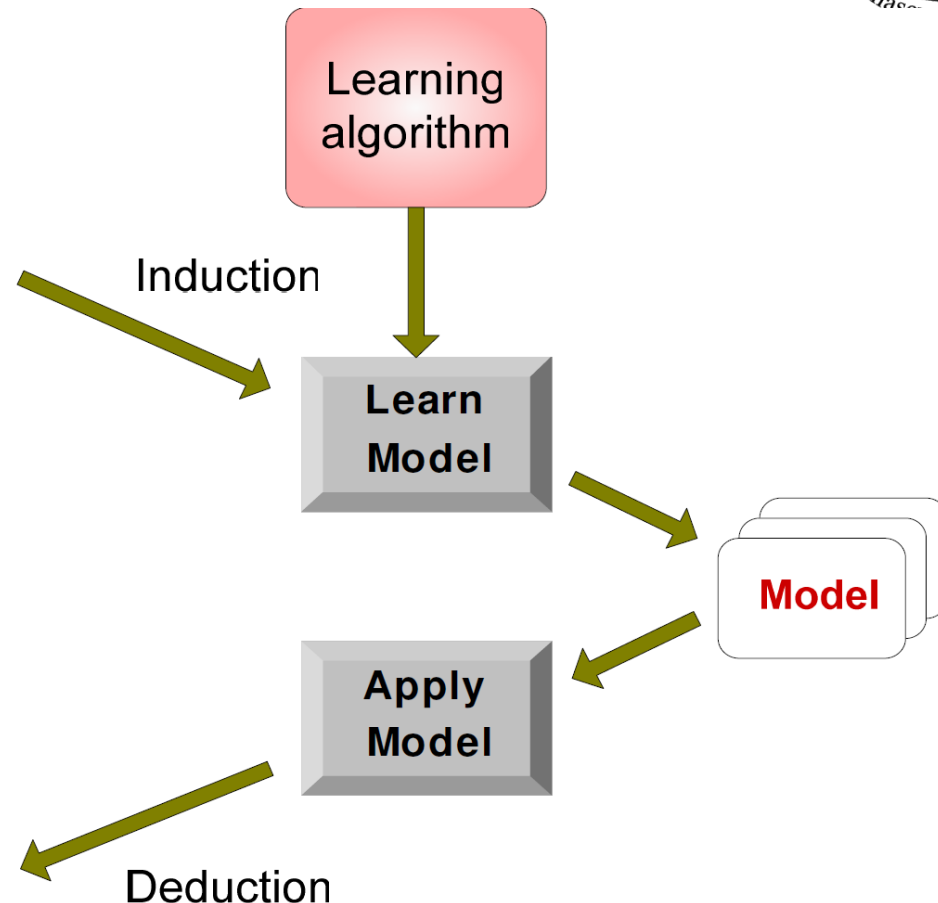
Classification

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

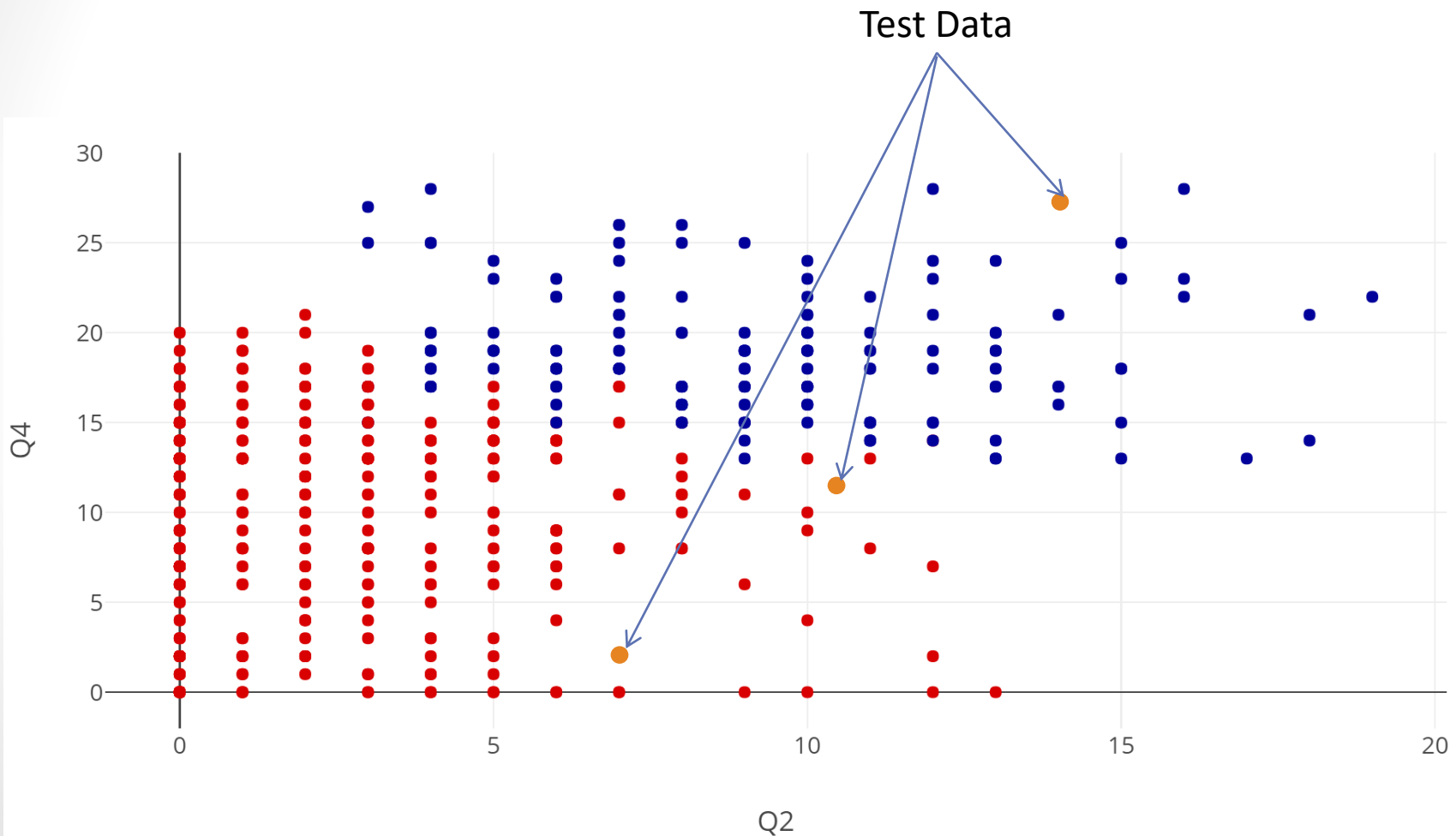
Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



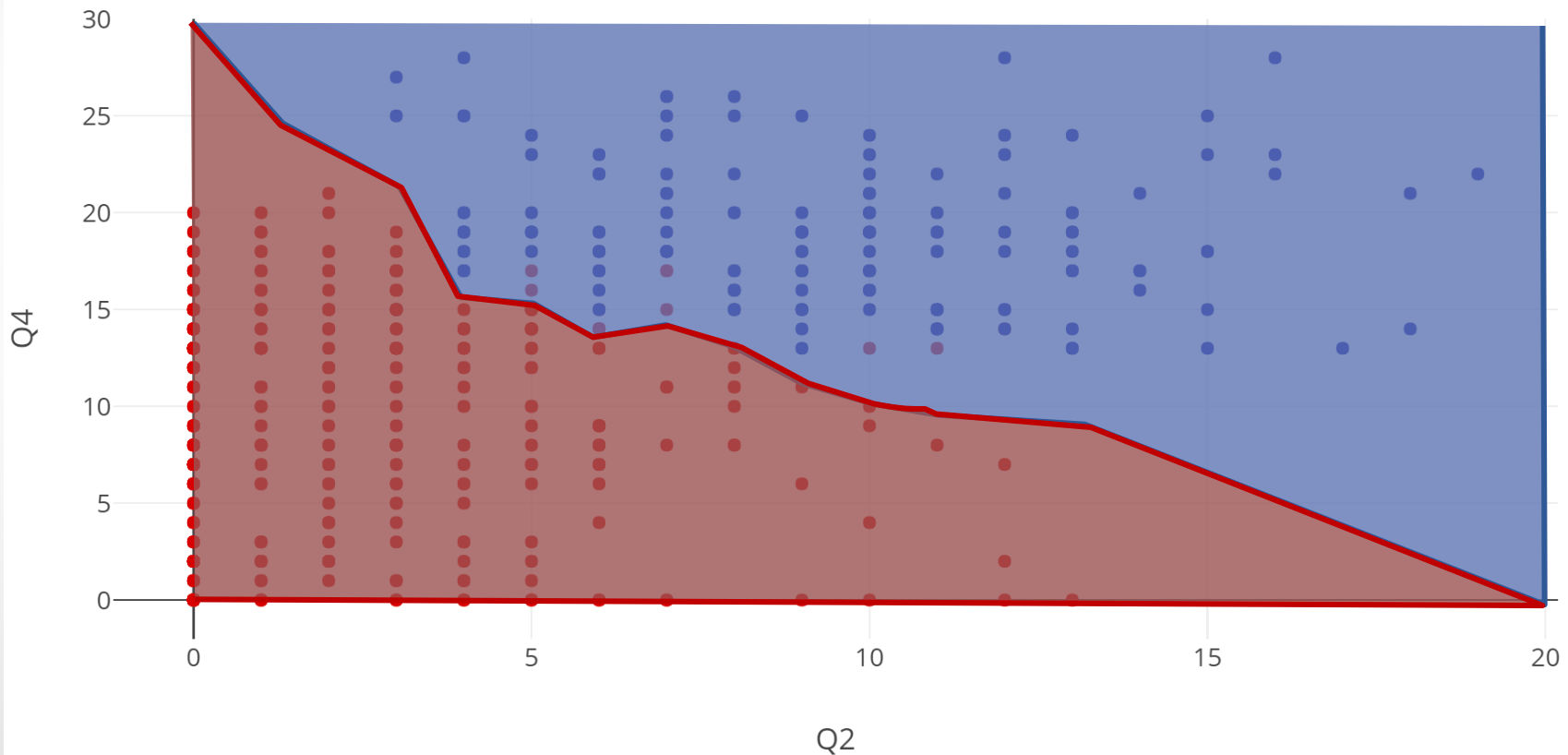
Classification



Classification



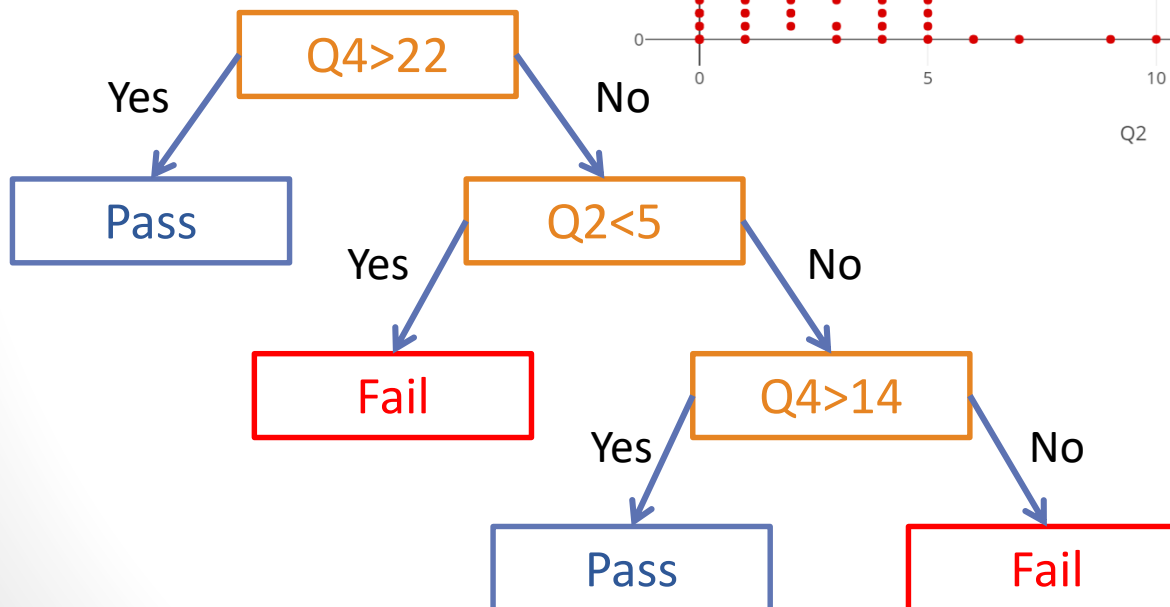
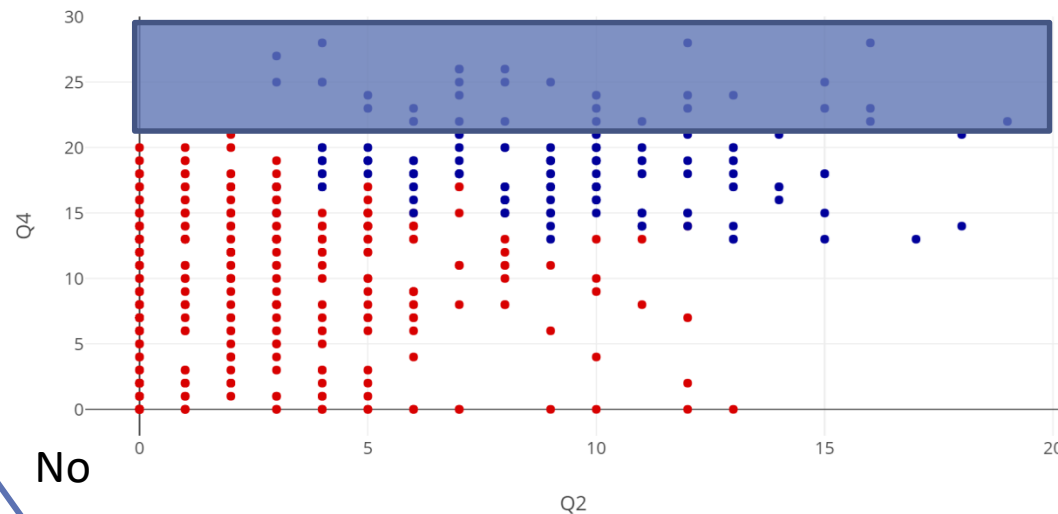
- K-NN



Classification



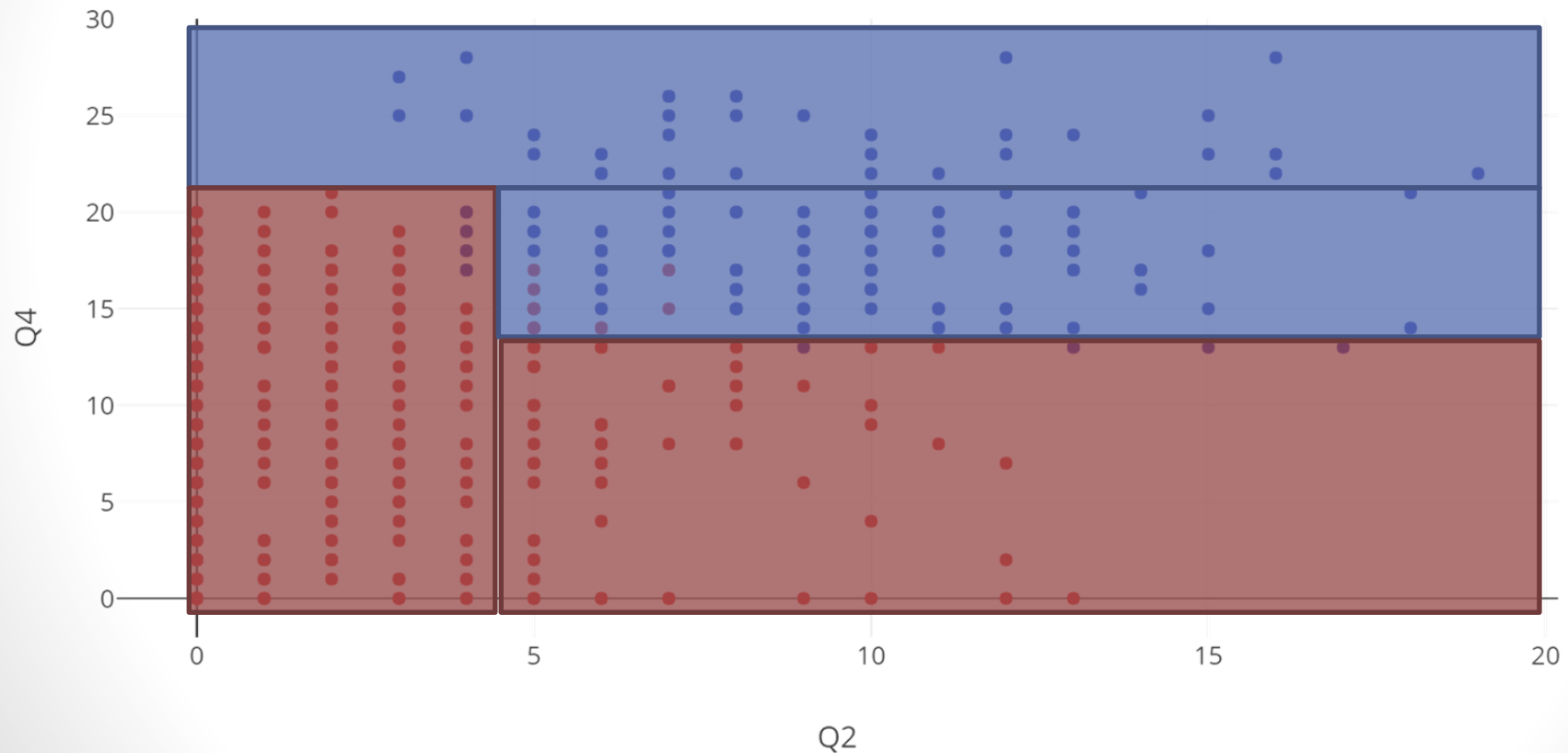
- Decision Tree



Classification



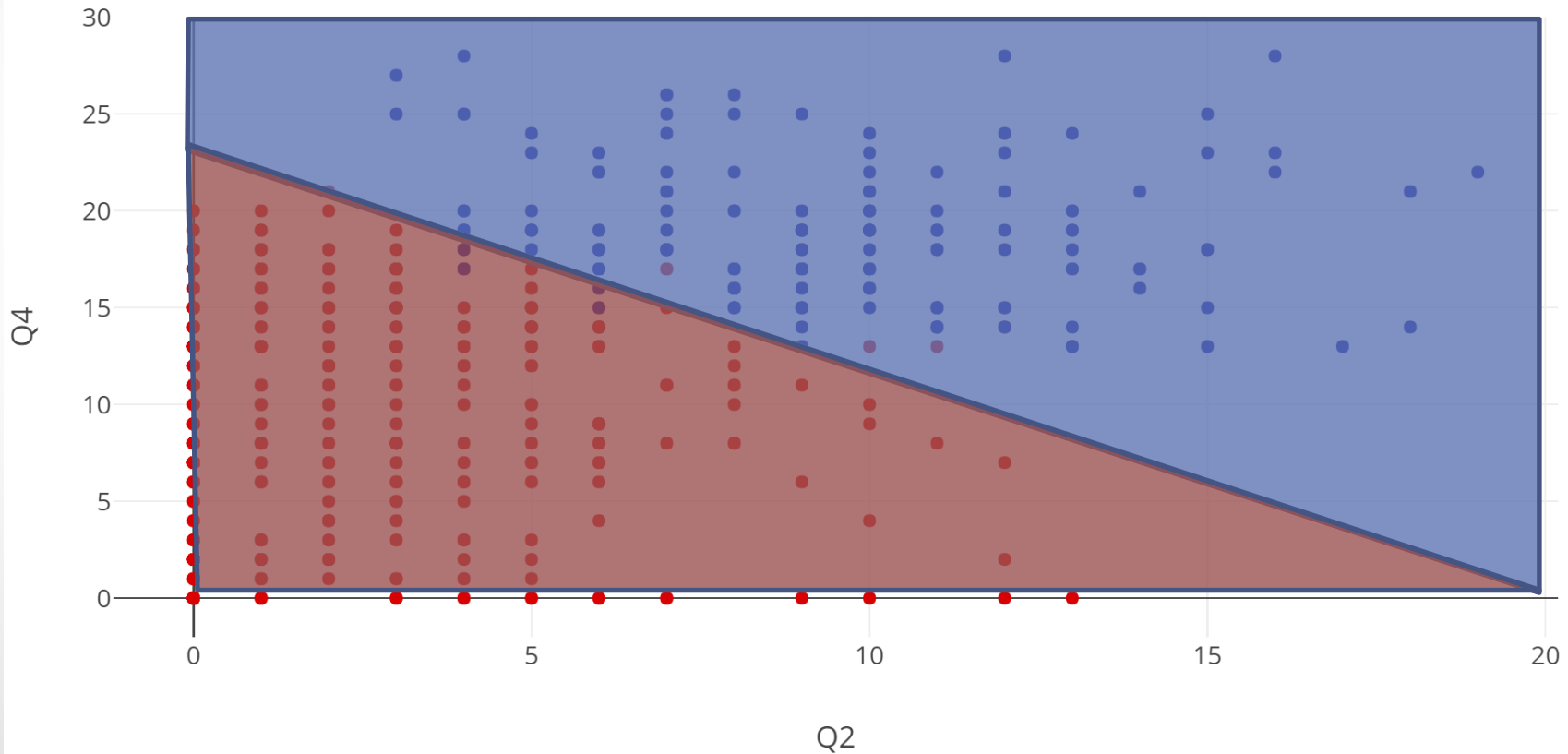
- Decision Tree



Classification



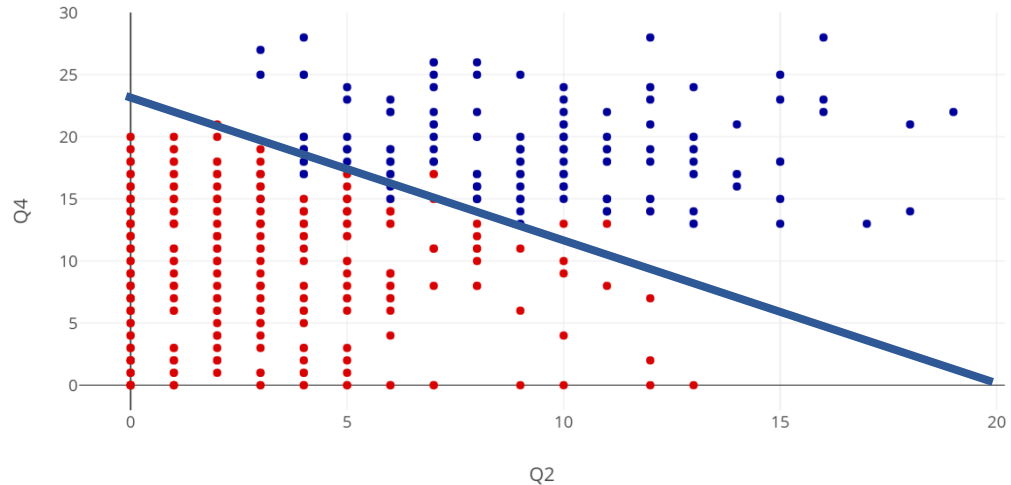
- Perceptron



Classification



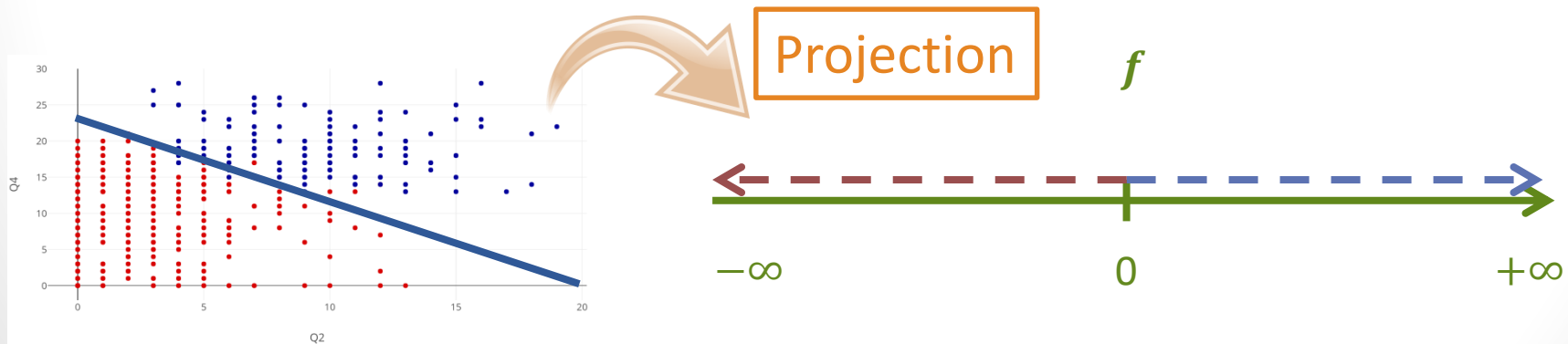
- Perceptron



- $f(Q1, Q2; w1, w2, c) = w1 * Q2 + w2 * Q4 + c \leq 0$
- $f(5, 10; w1, w2, c) = w1 * 5 + w2 * 10 + c < 0$
- $f(15, 20; w1, w2, c) = w1 * 15 + w2 * 20 + c > 0$

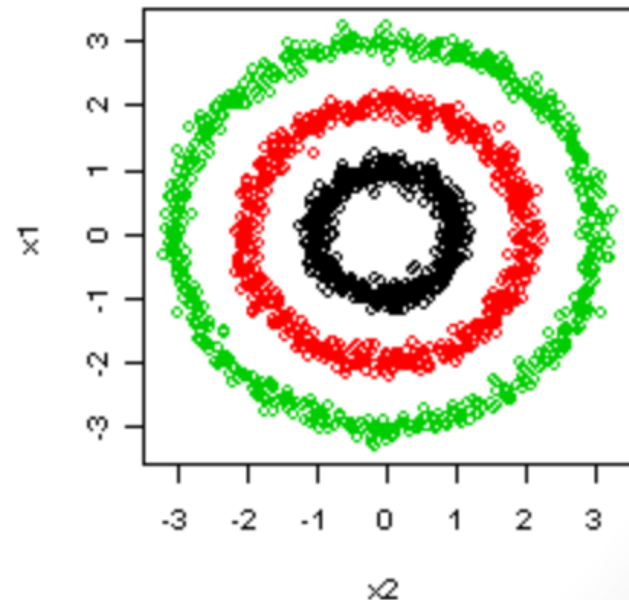
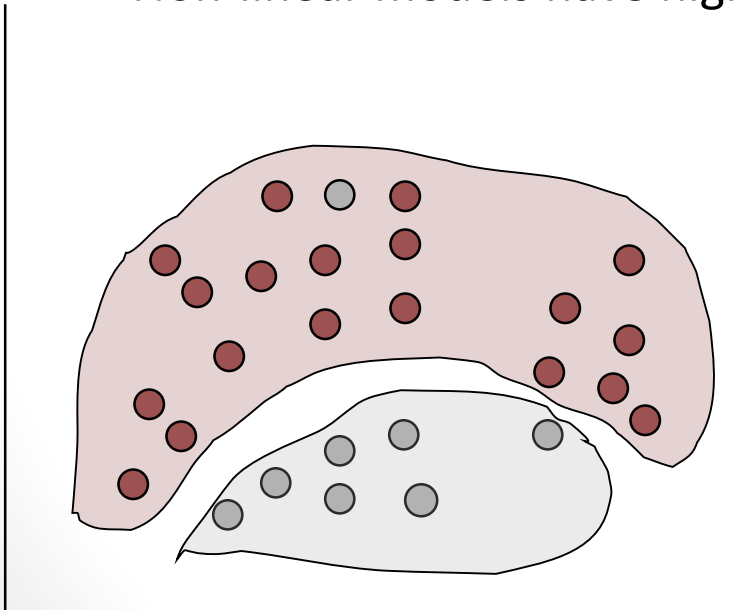
Classification

- $f(5,10; w1, w2, c) = w1 * 5 + w2 * 10 + c < 0$
- $f(15,20; w1, w2, c) = w1 * 15 + w2 * 20 + c > 0$
- ...
- $f(X; \theta) = X * \theta$
 - linear projection of data onto a **decision space**



Classification

- Projections may be non-linear
 - Kernel SVM, neural networks, etc...
 - The concept is the same
 - Require huge amounts of data to tune non-linear models
 - Non-linear models have higher capacity to model the real world.



Model Evaluation

Confusion Matrix:

Actual class \ Predicted class	C_1	$\neg C_1$
C_1	True Positives (TP)	False Negatives (FN)
$\neg C_1$	False Positives (FP)	True Negatives (TN)

Example of Confusion Matrix:

Actual class \ Predicted class	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	6954	46	7000
buy_computer = no	412	2588	3000
Total	7366	2634	10000

Model Evaluation

- Use **validation** set when selecting model
- Use **testing** set when comparing model
- Methods for estimating a classifier's accuracy:
 - Holdout method, random subsampling
 - Cross-validation
 - Bootstrap
- Comparing classifiers:
 - Confidence intervals
 - Cost-benefit analysis and ROC Curves



Model Evaluation

- Holdout method
 - Given data is randomly partitioned into two independent sets
 - Training set (e.g., $2/3$) for model construction
 - Test set (e.g., $1/3$) for accuracy estimation
 - Random sampling: a variation of holdout
 - Repeat holdout k times, accuracy = avg. of the accuracies obtained
- Cross-validation (k-fold, where $k = 10$ is most popular)
 - Randomly partition the data into k mutually exclusive subsets, each approximately equal size
 - At i -th iteration, use D_i as test set and others as training set
 - Leave-one-out: k folds where $k = \#$ of tuples, for small sized data
 - *Stratified cross-validation*: folds are stratified so that class dist. in each fold is approx. the same as that in the initial data

Model Evaluation

A\P	C	¬C	
C	TP	FN	P
¬C	FP	TN	N
	P'	N'	All

- **Classifier Accuracy**, or recognition rate: percentage of test set tuples that are correctly classified

$$\text{Accuracy} = (TP + TN) / \text{All}$$

- **Error rate**: $1 - \text{accuracy}$, or

$$\text{Error rate} = (FP + FN) / \text{All}$$

■ Class Imbalance Problem:

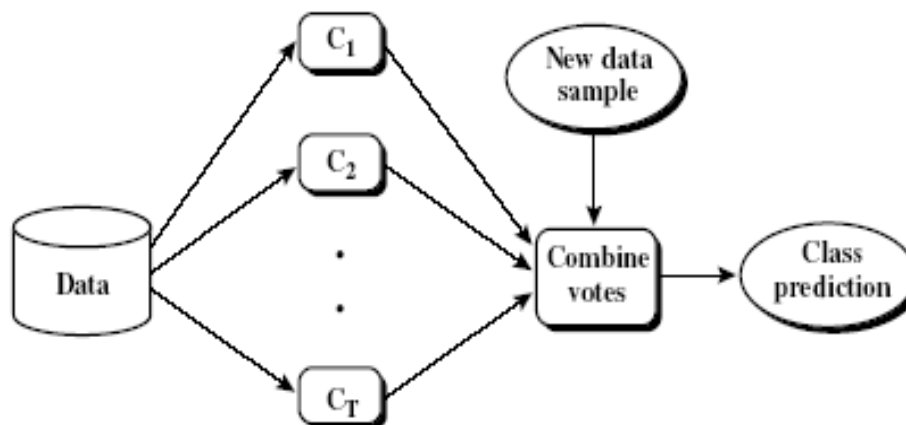
- One class may be *rare*, e.g. fraud, or HIV-positive
- Significant *majority of the negative class* and minority of the positive class
- **Sensitivity**: True Positive recognition rate
 - **Sensitivity** = TP / P
- **Specificity**: True Negative recognition rate
 - **Specificity** = TN / N

Model Evaluation

- Class-imbalance problem: Rare positive example but numerous negative ones, e.g., medical diagnosis, fraud, oil-spill, fault, etc.
- Traditional methods assume a balanced distribution of classes and equal error costs: not suitable for class-imbalanced data
- Typical methods for imbalance data in 2-class classification:
 - Oversampling: re-sampling of data from positive class
 - Under-sampling: randomly eliminate tuples from negative class
 - Threshold-moving: moves the decision threshold, t , so that the rare class tuples are easier to classify, and hence, less chance of costly false negative errors
- Ensemble techniques:
 - Ensemble multiple classifiers introduced above
- Still difficult for class imbalance problem on multiclass tasks

Ensamble

- Ensemble methods
 - Use a combination of models to increase accuracy
 - Combine a series of k learned models, M_1, M_2, \dots, M_k , with the aim of creating an improved model M^*
- Popular ensemble methods
 - Bagging: averaging the prediction over a collection of classifiers
 - Boosting: weighted vote with a collection of classifiers
 - Ensemble: combining a set of heterogeneous classifiers



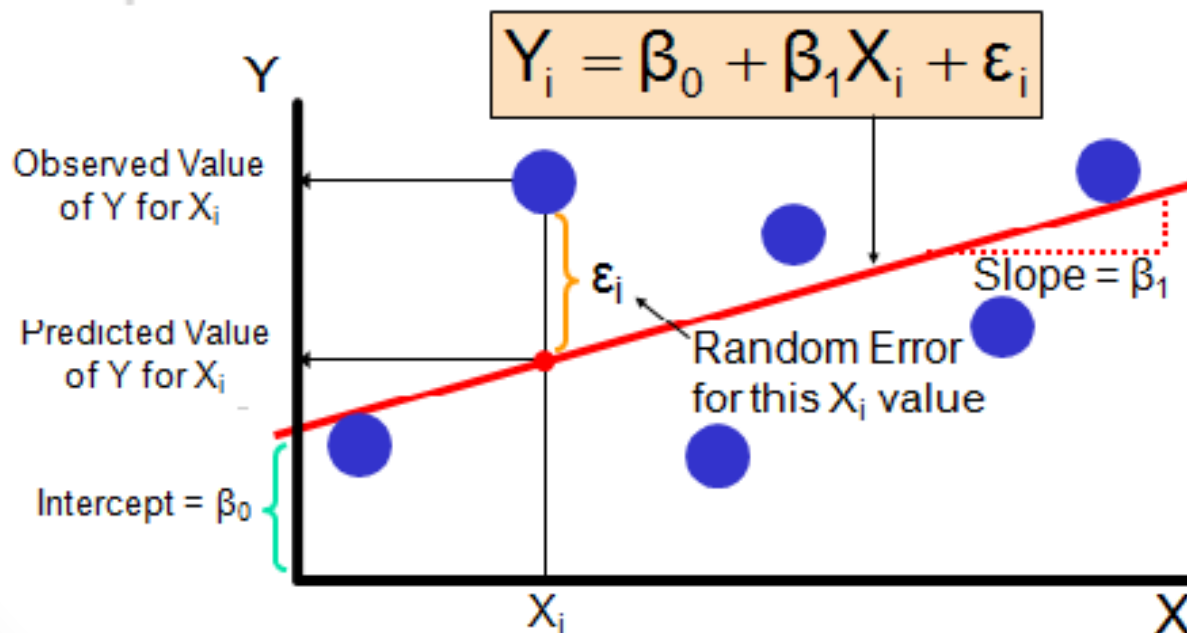
Regression



- Supervised learning where the label is a continuous numerical value.
- Often called *Curve Fitting*.
- Fundamentally, regression is the act of using a set of independent variables (features) to estimate the value of another variable (label value).
- An example is the use of marks in Pr1, Pr2 and Pr3 in order to predict the mark in SE.

Regression

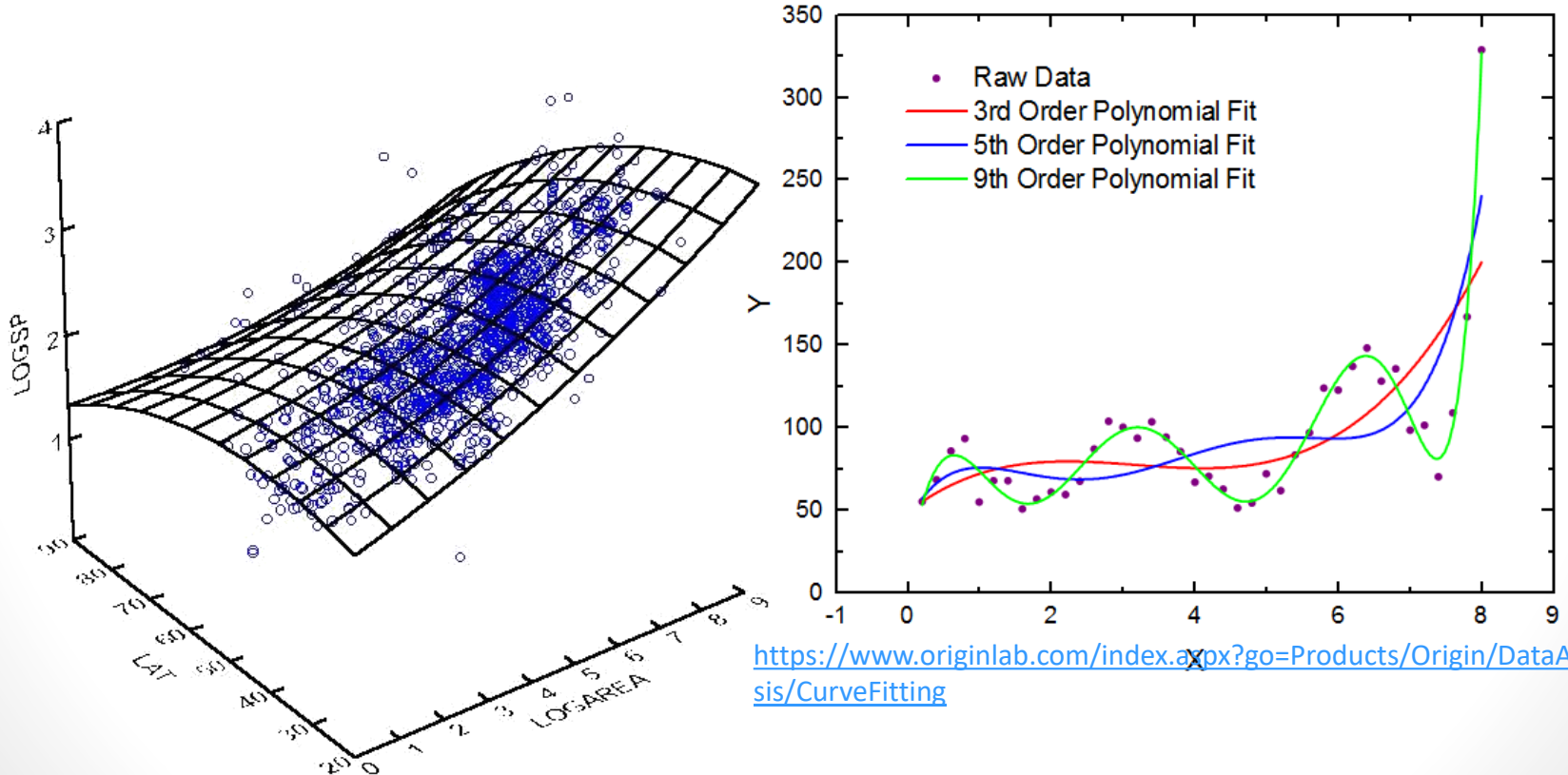
- Any deviation away from real target values is encoded as the error which must be minimised on training data during the training phase.
 - Ridge Regression, LASSO Regression, Polynomial Regression, etc...



Regression

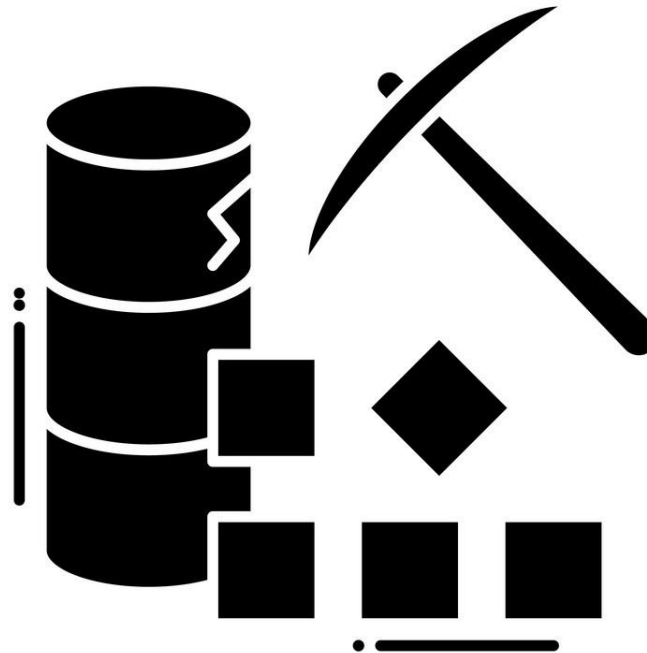


- $f(\text{features}; \text{parameters}) \rightarrow \text{Target} \pm \epsilon$



<https://www.originlab.com/index.aspx?go=Products/Origin/DataAnalysis/CurveFitting>

Other Data Tasks



DATA MINING

www.vectorstock.com



Point of Interest Detection

- Detection of examples which are out of the “norm”.
- Requires the definition of “norm”
- Examples:
 - Fraud Detection;
 - Scene Change;
 - Speaker Change;
 - etc...
- Closely related to decision theory and confidence levels.



Information Filtering


























- Using past behaviour in order to filter items which are of interest to the user.
- Analyse data as well as interaction with the data.
- May model items (examples) and their relatedness, users and their relatedness, and interaction between user/item.
- Can be seen as a special type of regression models predicting rating of an item that would be given by a user.

Information Filtering

- Recommender systems produce recommendations use:
 - *Collaborative filtering*: build a model from a user's past behaviour (items previously purchased or selected and/or numerical ratings given to those items) as well as similar decisions made by other users; or
 - *Content-based filtering*: utilize a series of discrete characteristics of an item in order to recommend additional items with similar properties; or
 - *A hybrid of the two*.

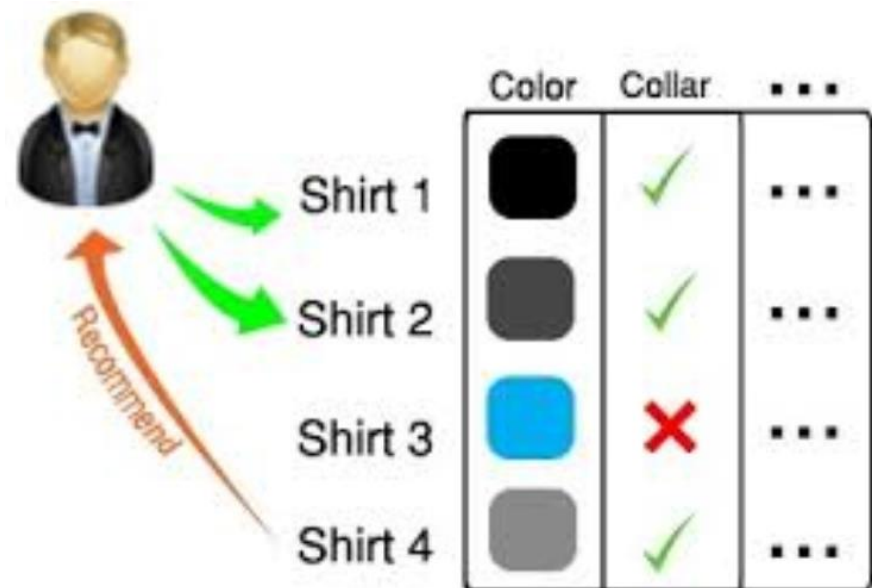
Information Filtering

Collaborative filtering

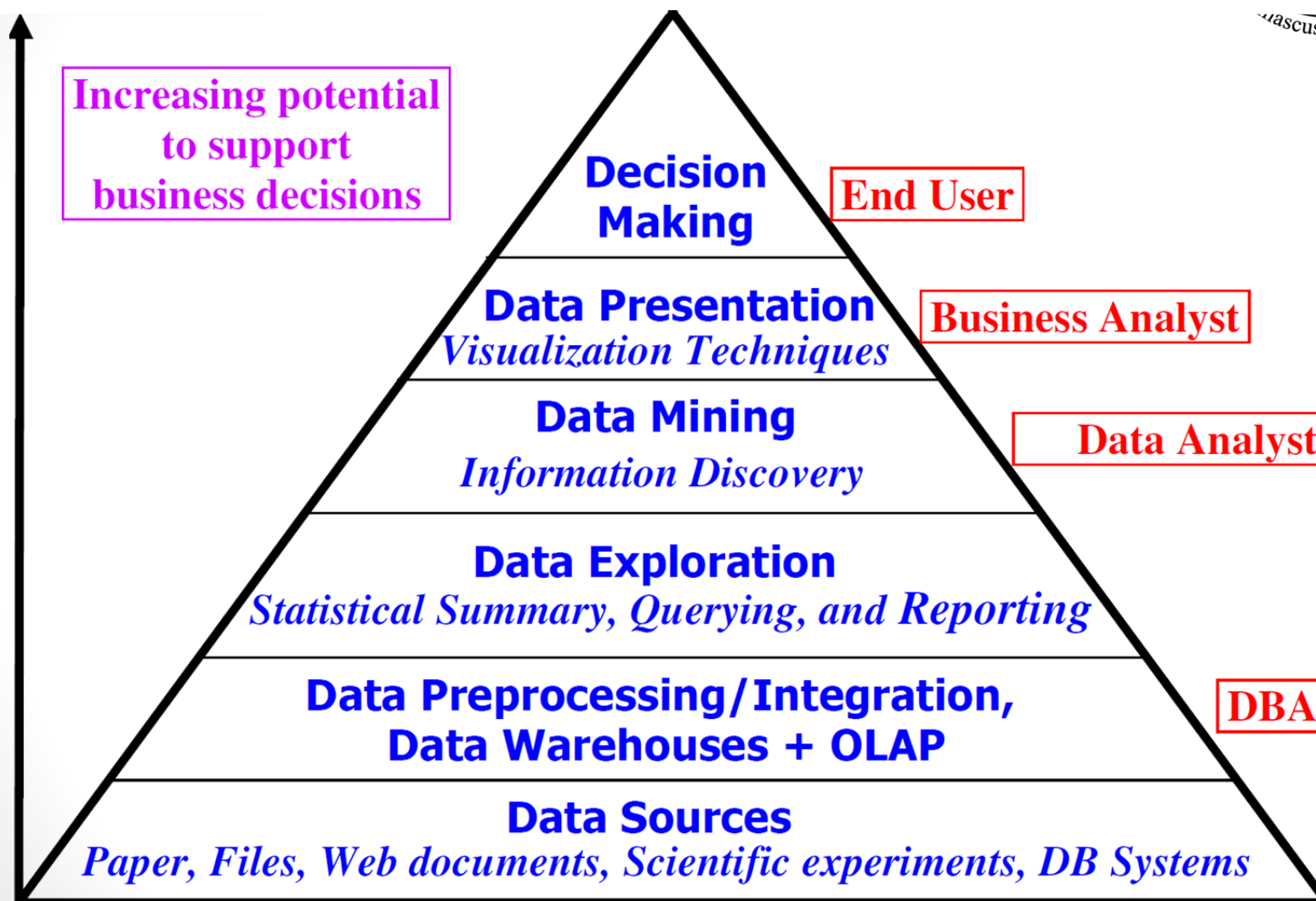
https://en.wikipedia.org/wiki/Collaborative_filtering

Content-based filtering

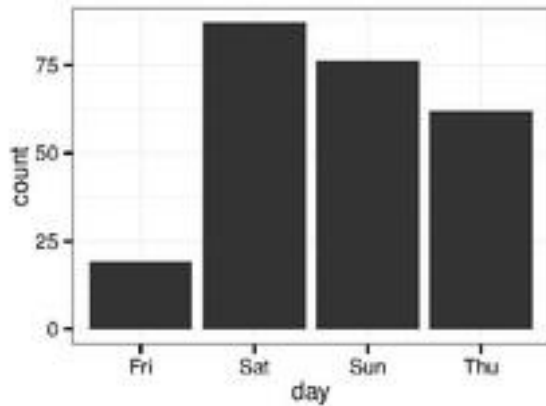


<http://ashishvs.in/2017-03-21-recommender-systems-and-collaborative-filtering-walkthrough/>

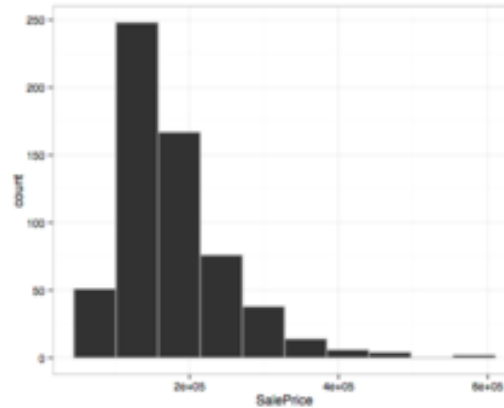
Visualization



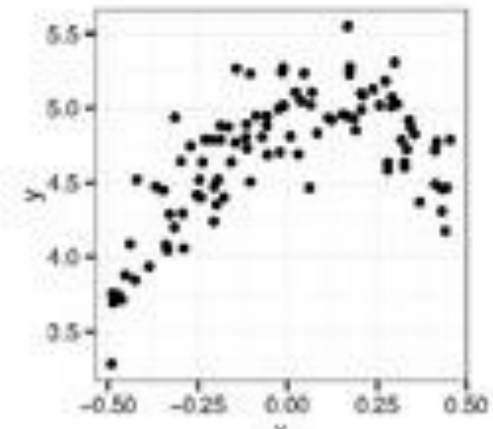
Visualization



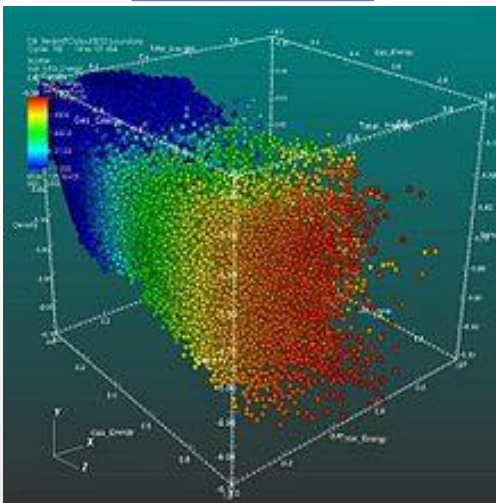
Bar chart



Histogram



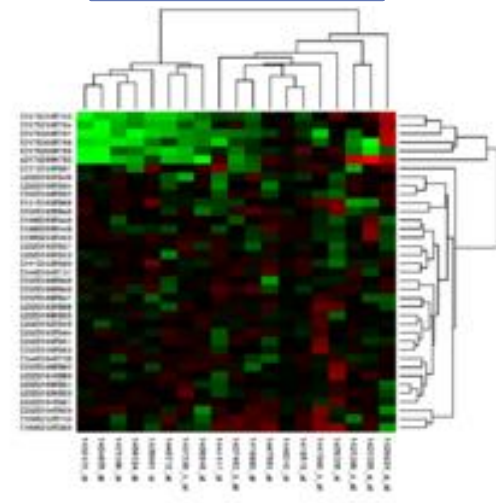
Scatter plot



Scatter plot (3D)

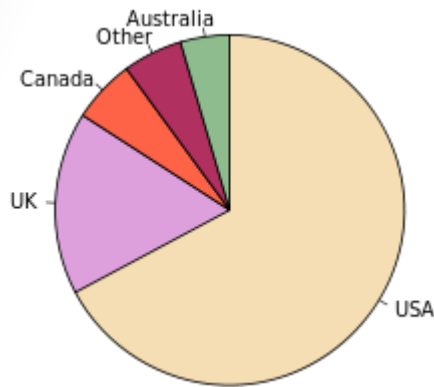


Network

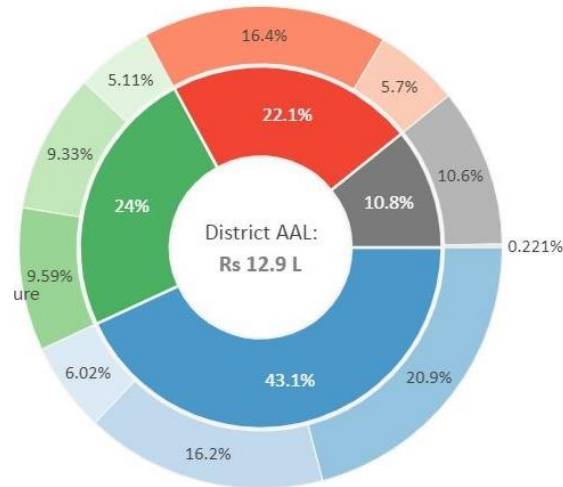


Heat map

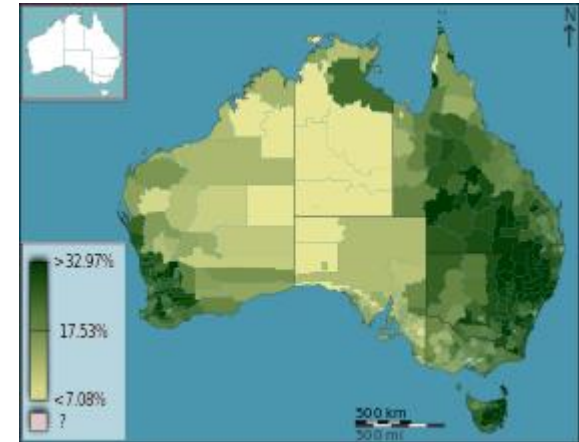
Visualization



Pie chart



Donut chart



Map chart

Summarization

- Summarization of data:
 - Compact description for a subset of data.
 - e.g. mean and standard deviations for all fields.
 - Summarization functions are often used in *interactive exploratory* data analysis and automated report generation; i.e., dashboards.
- *Automatic Summarization*: the process of shortening a text document or a video clip while maintaining the original points of interest and resulting in a coherent summary.
- Generally, techniques based on extraction and abstraction.
- Heavily used in Executive-Level systems.

Structuring Data

- Techniques for encoding single examples and improving their representation for data mining tools
 - Text (or Text Mining)
 - Images (or Computer Vision)
 - Audio
 - Video
 - Categorical data
 - Missing values
 - Feature normalisation
 - Representation Learning

Sources



- Data mining and KDD (SIGKDD: CDROM)
 - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
 - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
 - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
 - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
 - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
 - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- Web and IR
 - Conferences: SIGIR, WWW, CIKM, etc.
 - Journals: WWW: Internet and Web Information Systems,
- Statistics
 - Conferences: Joint Stat. Meeting, etc.
 - Journals: Annals of statistics, etc.
- Visualization
 - Conference proceedings: CHI, ACM-SIGGraph, etc.
 - Journals: IEEE Trans. visualization and computer graphics, etc.

References



Assessed Reading:

- U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM. 1996.
- Cluster analysis. Wikipedia. https://en.wikipedia.org/wiki/Cluster_analysis
- Data visualization. Wikipedia. https://en.wikipedia.org/wiki/Data_visualization

Bibliography:

- G. Piatetsky-Shapiro, G. Parker. Data Mining Course.
http://www.kdnuggets.com/data_mining_course/index.html
- G. Piatetsky-Shapiro, G. Parker. Lecture 2. Machine Learning: finding patterns.
<https://ameensheriffmca.files.wordpress.com/2014/08/dm2-intro-machine-learning-classification.ppt>
- S. Clayton. Building a Recommendation Engine in C#. 2018.
<https://www.codeproject.com/Articles/1232150/Building-a-Recommendation-Engine-in-Csharp>