Machine learning

Bibliography

1. D.Larose, Discovering Knowledge in Data: An Introduction to Data Mining, Willey, 2014.
2. H.Jiawei, M.Kamber, Data Mining: Concepts and Techniques. Elsevier, 2006.
3. T.Pang-Ning, M.Steinbach, V.Kumar, Introduction to Data Mining, Pearson, 2014
4. G.James, D.Witten, T.Hastie, R.Tibshirani. An Introduction to Statistical Learning. New York: Springer, 2013.
5. L.Brett Machine learning with R, Packt PublishingPackt, 2015.
6. T.Hastie, R.Tibshirani, J.Friedman, The Elements of Statistical Learning, Springer, 2009
7. A.Navlani, A.Fandango, I.Idris, Python Data Analysis, Packt PublishingPackt, 2021.
8. S.Raschka, V.Mirjalili, Python Machine Learning, Packt PublishingPackt, 2019.

Bibliography

1. Saed Sayad. **Data Mining Map**.
   http://www.saedsayad.com/data_mining_map.htm.
2. Analytics, Data Mining, and Data Science.
   http://www.kdnuggets.com/.
3. Kaggle https://www.kaggle.com/datasets?fileType=csv.

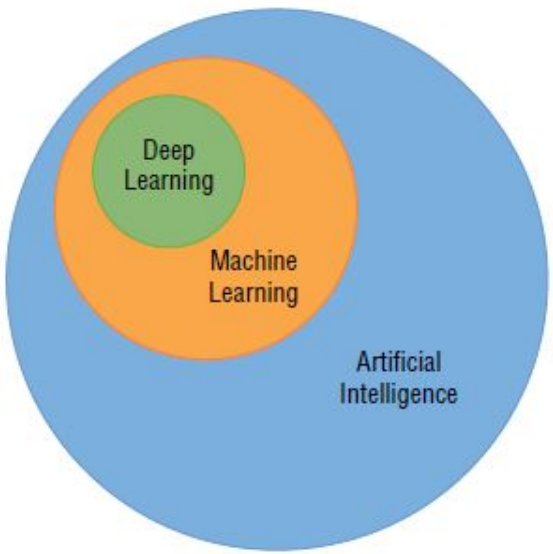
Data

The collection of the University of Lodz Library contains approximately 2.8 millions volumes. If the average size of a document was 1MB (although it usually has more), the library would take up 30 terabytes. Meanwhile, the database of courier shipments in a logistics company, is about 20 terabytes.

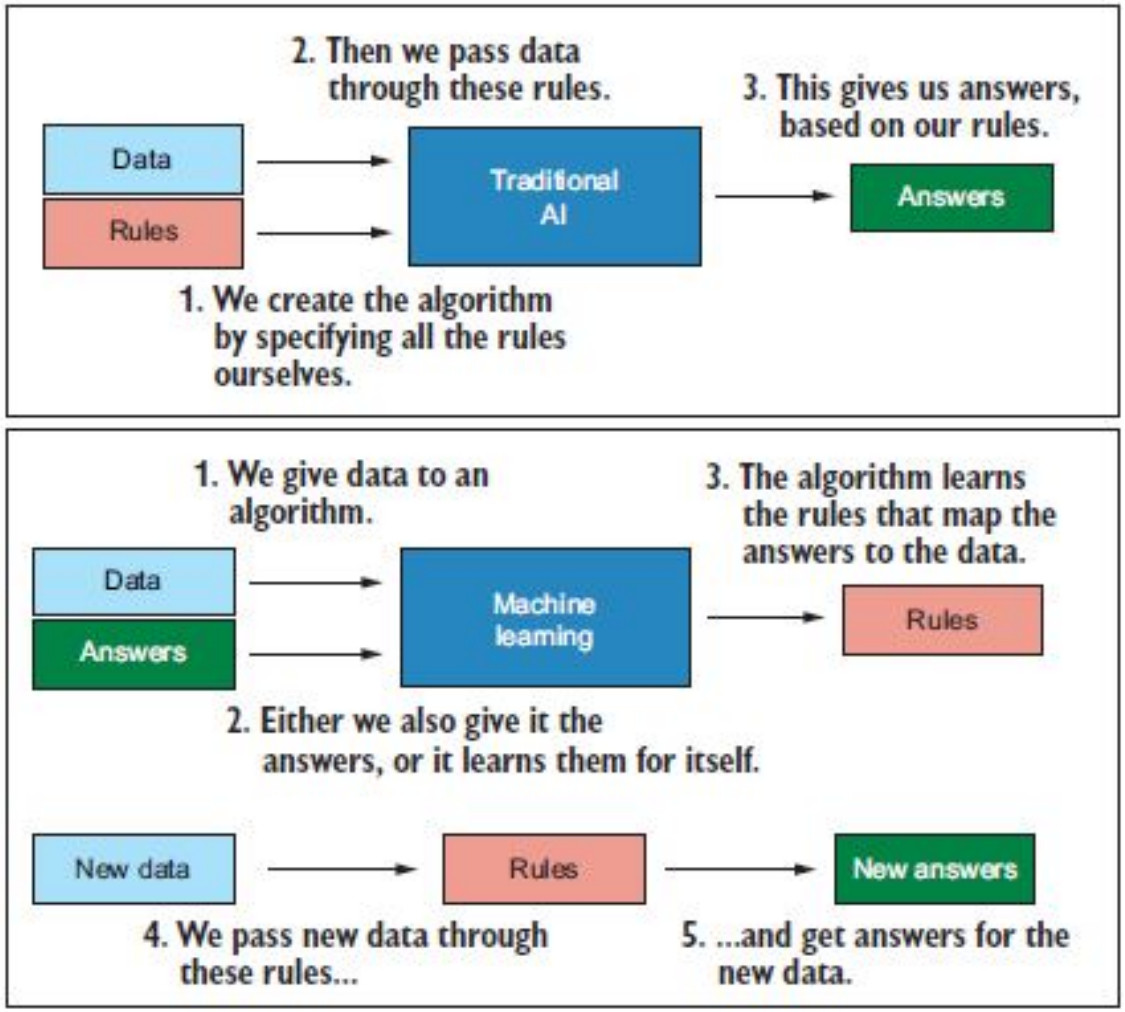
Wielokrotności bajtów					
Przedrostki dziesiętne (SI)			Przedrostki binarne (IEC 60027-2)		
Nazwa	Symbol	Mnożnik	Nazwa	Symbol	Mnożnik
kilobajt	kB	$10^3 = 1000^1$	kilobajt	KiB	$2^{10} = 1024^1$
megabajt	MB	$10^6 = 1000^2$	megabajt	MiB	$2^{20} = 1024^2$
gigabajt	GB	$10^9 = 1000^3$	gigabajt	GiB	$2^{30} = 1024^3$
terabajt	TB	$10^{12} = 1000^4$	**terabajt**	TiB	$2^{40} = 1024^4$
petabajt	PB	$10^{15} = 1000^5$	petabajt	PiB	$2^{50} = 1024^5$
eksabajt	EB	$10^{18} = 1000^6$	eksabajt	EiB	$2^{60} = 1024^6$
zettabajt	ZB	$10^{21} = 1000^7$	zettabajt	ZiB	$2^{70} = 1024^7$
jottabajt	YB	$10^{24} = 1000^8$	jottabajt	YiB	$2^{80} = 1024^8$

The relationship between Artificial Intelligence and Machine learning

- **Artificial intelligence (AI)** includes any type of technique where we are attempting to get a computer system to imitate human behavior. As the name implies, we are trying to ask computer systems to artificially behave as if they were intelligent.
- **Machine learning (ML)** is a subset of artificial intelligence techniques that attempt to apply statistics to data problems in an effort to discover new knowledge by generalizing from examples.

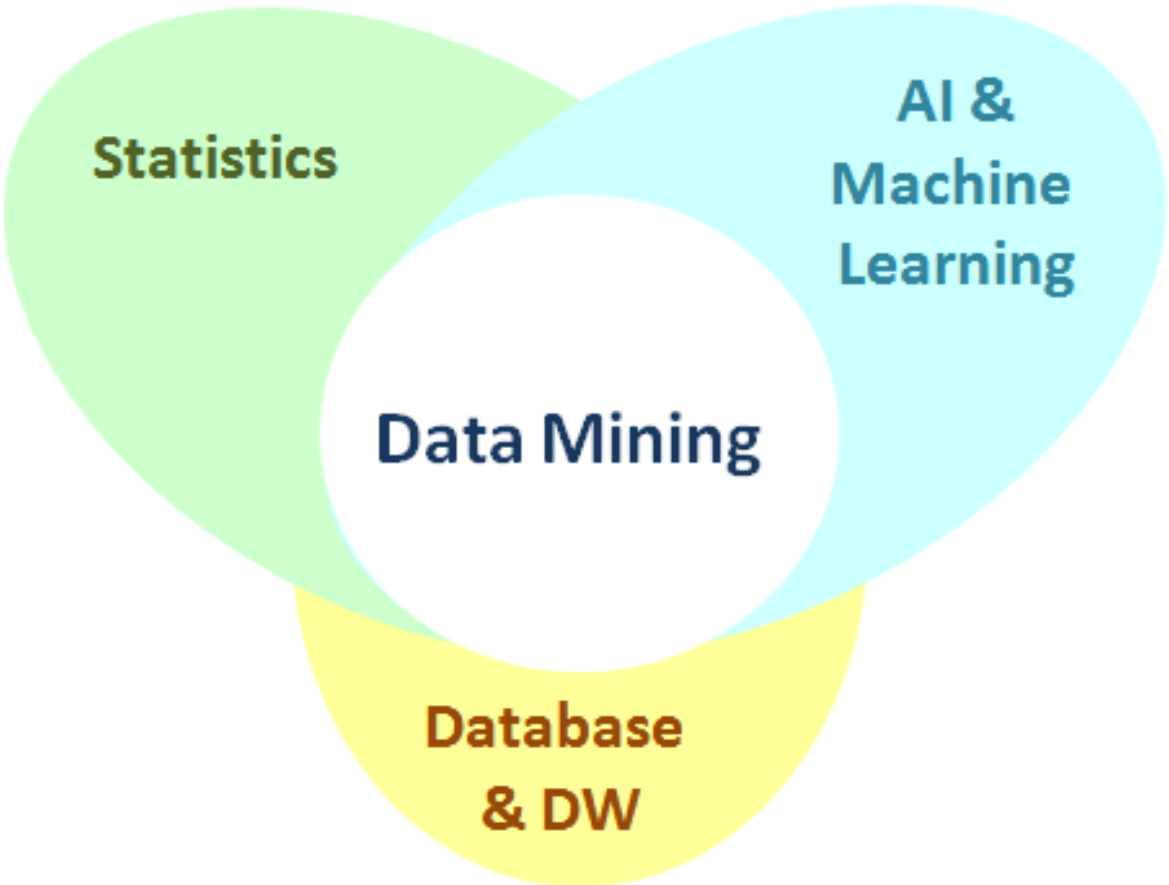


The relationship between Artificial Intelligence and Machine learning



The relationship between Data mining and Machine learning

Machine learning (ML) is a field of computer science that studies algorithms and techniques for automating solutions to complex problems.



Definition of datamining

Graham Williams

Data mining is the art and science of intelligent data analysis. The aim is to discover meaningful insights and knowledge from data. Discoveries are often expressed as models, and we often describe data mining as the process of building models. A model captures, in some formulation, the essence of the discovered knowledge. A model can be used to assist in our understanding of the world. Models can also be used to make predictions.

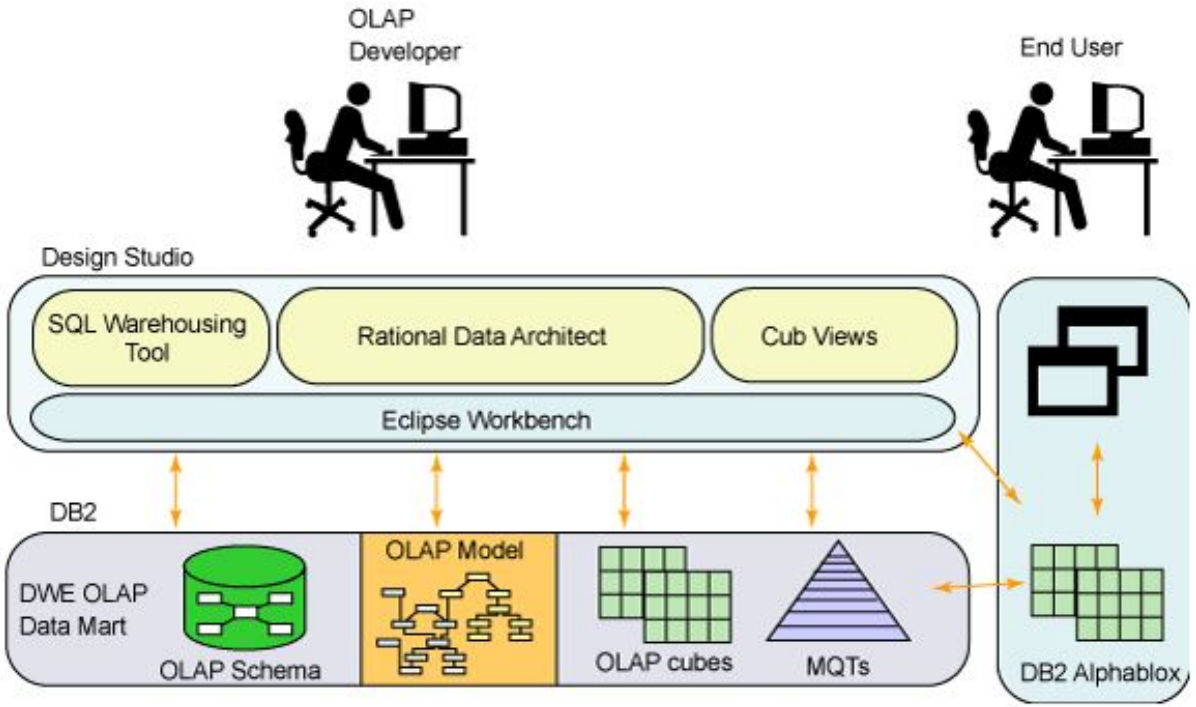
Data mining vs Machine learning

1. **Data mining** is a technique of discovering different kinds of patterns that are inherited in the data set and which are precise, new, and useful data. Data Mining is working as a subset of business analytics and similar to experimental studies. Data Mining's origins are databases, statistics.
2. **Machine learning** includes an algorithm that automatically improves through data-based experience. Machine learning is **a way to find a new algorithm from experience**. Machine learning includes the study of an algorithm that can automatically extract the data. Machine learning utilizes data mining techniques and another learning algorithm to construct models of what is happening behind certain information so that it can predict future results.

Data mining and **Machine learning** are areas that influence each other and have many things in common.

What is not dm and ml

Dm and ml is not **OLAP**.



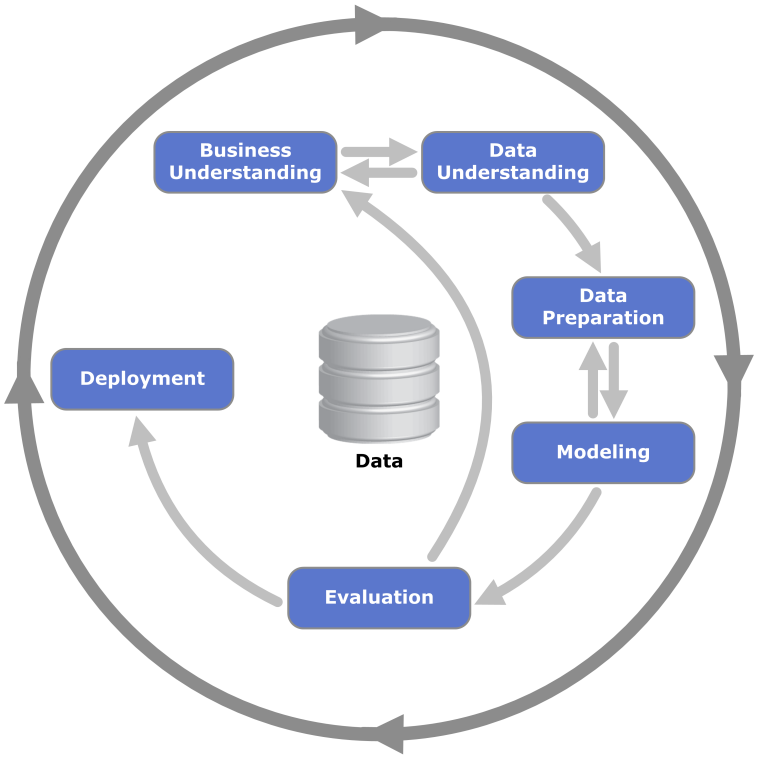
What is not dm and ml

1. How many customers who bought a suit bought a shirt?
2. Which customers are not paying back the loan?
3. Which customers have not renewed their insurance policies?

1. What product did the customers who bought the suit, buy?
2. What credit risk does the customer pose?
3. Which customers may leave for another company?

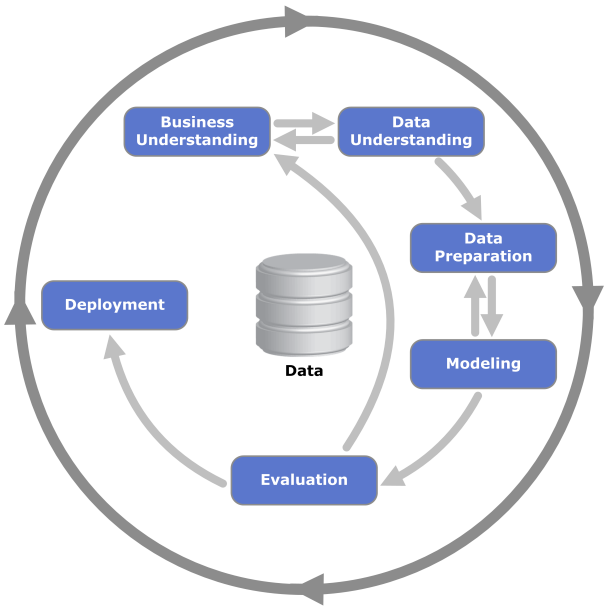
Data analysis process

The most commonly used approach is **Cross Industry Process for Data Mining CRISP-DM**, 1996).



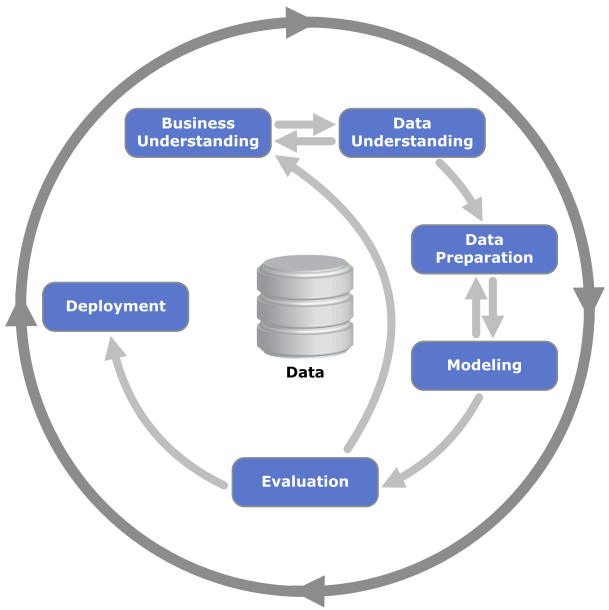
1. **Problem Understanding lub Business Understanding**
2. **Data Understanding**
3. **Data Preparation**
4. **Modeling**
5. **Evaluation**
6. **Deployment**

Data analysis process - Problem Understanding lub Business Understanding



This initial phase focuses on understanding the project aims and requirements from a business perspective, then converting this knowledge into a data analysis problem definition and a preliminary plan designed to achieve the aims.

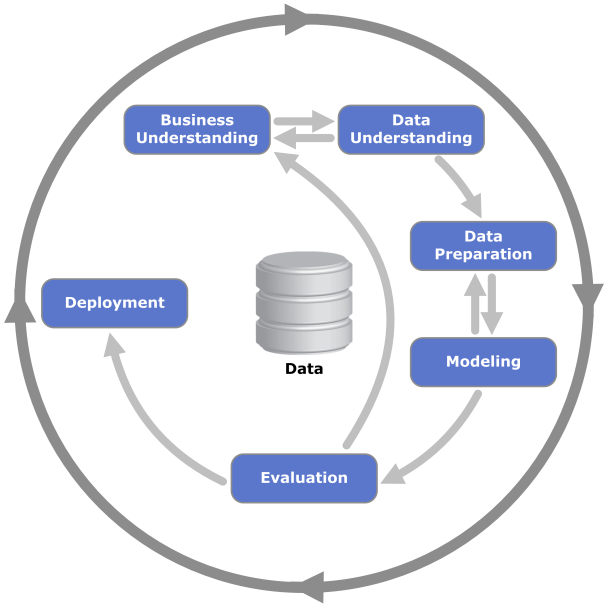
Data analysis process - Data Understanding



- The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems.

After this step you should know the answer to the following questions:

1. where did the data come from?
2. who collected them and what methods did they use to collect them?
3. what do the rows and columns in the data mean?
4. are there any obscure symbols or abbreviations in the data?

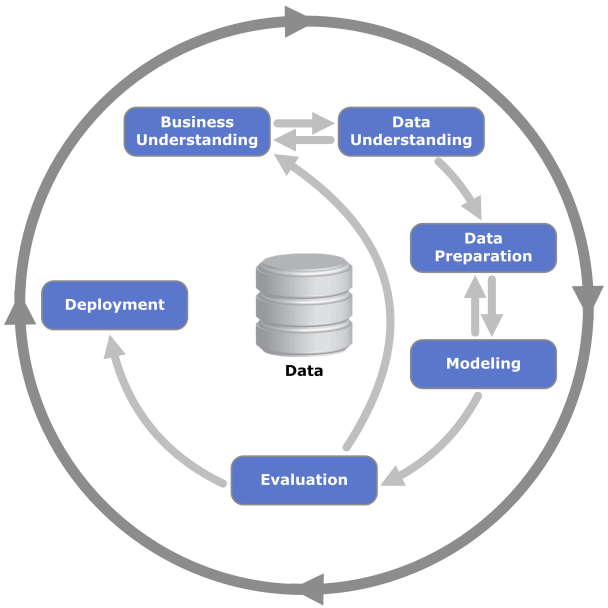


- The data preparation phase covers all activities to construct the final dataset from the initial raw data.

Data preparation most often requires:

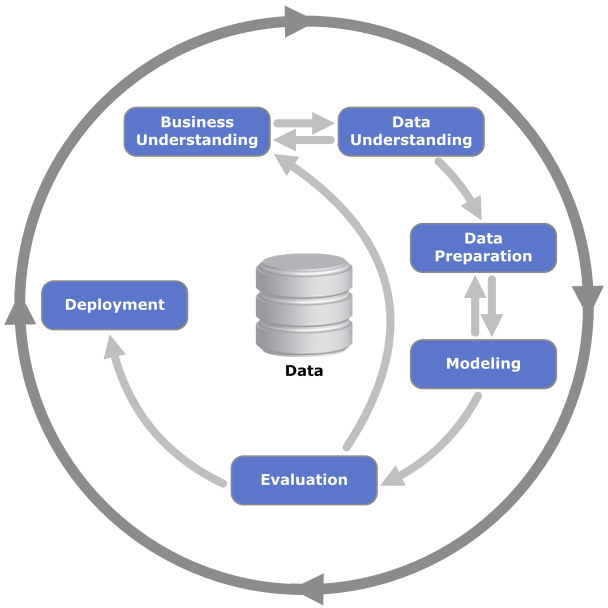
1. the joining of several data sets,
2. reducing the number of variables to only those which will be relevant for the process,
3. data cleaning (removal of anomalies, reformatting, normalisation, missing data).

Data analysis process - Modeling



At this stage, various modelling techniques are selected and applied, and their parameters are calibrated to optimal values.

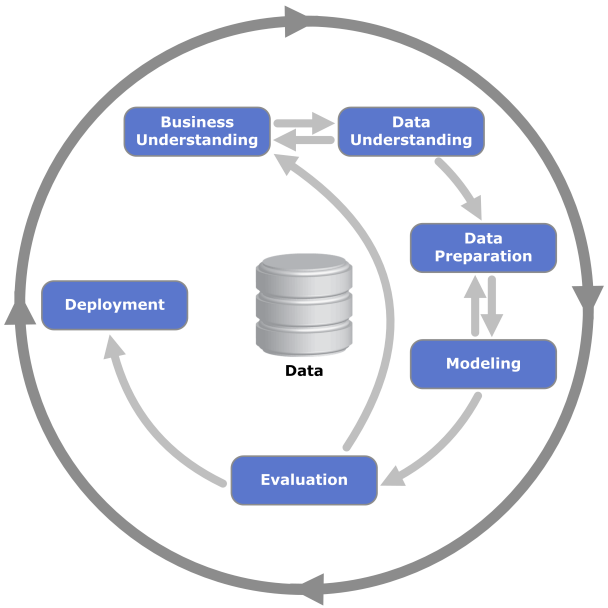
Data analysis process - Evaluation



Process evaluation consists of

- determining whether the model or models meet the assumptions established in the first stage (quality and efficiency)
- verifying whether there are any important business or research objectives that have not been taken into account deciding on the further use of the results

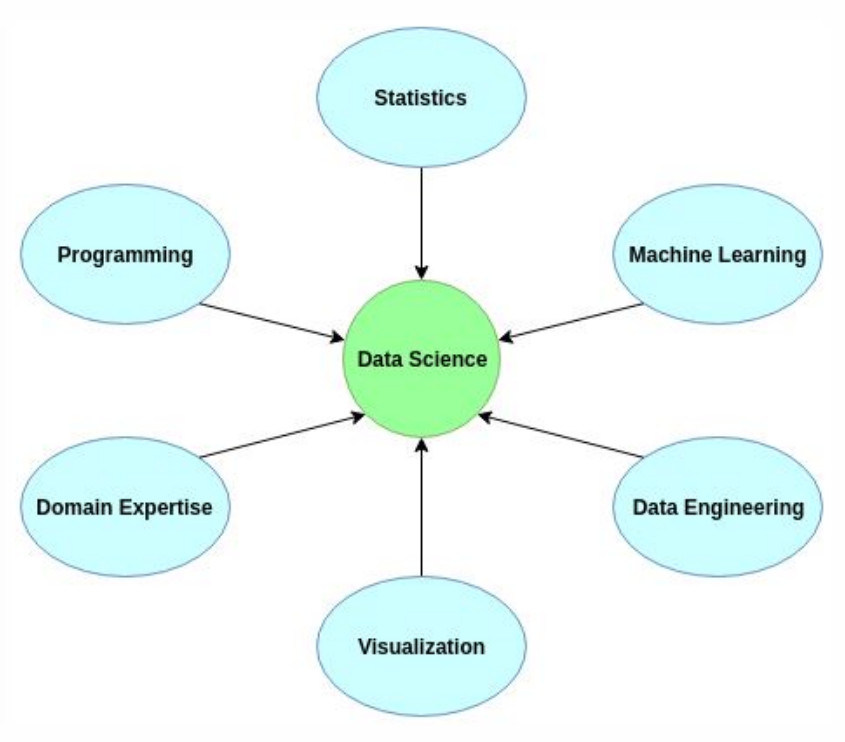
Data analysis process - Deployment



- Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it.

Data analysis engineering - summary

- http://www.kdnuggets.com/2017/02/analytics-grease-monkeys.html



https://www.purpleslate.com/what-is-data-mining/

The word data, is the plural of the word **datum**

datum (język łaciński) [edytuj]

znaczenia:

rzeczownik, rodzaj nijaki

(1.1) dar, podarek, datek[1]

(1.2) data[1]

odmiana:

(1.1) datum, datī (deklinacja II) [ukryj▲]

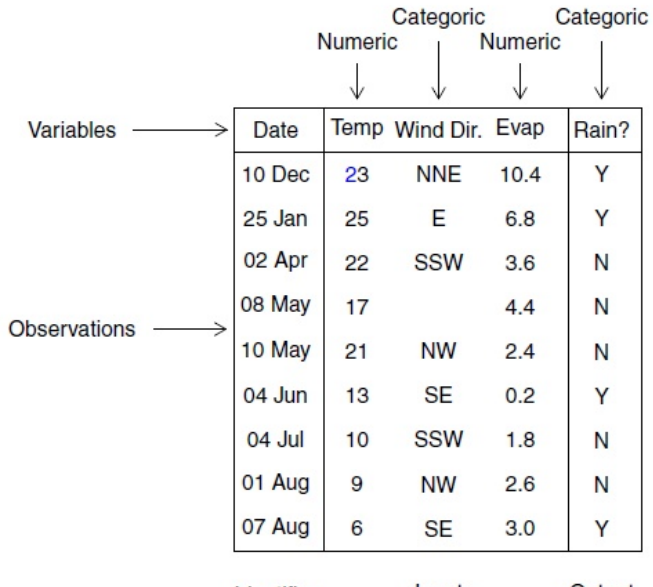
przypadek	liczba pojedyncza	liczba mnoga
mianownik	**datum**	**data**
dopełniacz	datī	datōrum
celownik	datō	datīs
biernik	datum	data
ablatyw	datō	datīs
wołacz	datum	data

Data

- The term was the first to be used by **Euclid**, in the work **Dedomena**.
- **data** is called, in Euclid's work, a quantity resulting directly from the terms of a given problem.
- **Euclid**

Data in the form of records

- each **record (object, sample, observation)** is described by a set of attributes (variables).
- each observation (record) has a fixed number of attributes (i.e. a fixed tuple length), so that it can be considered as a **vector in a multidimensional space** whose dimension is equal to the number of attributes.
- the data set can be represented as a matrix of type $m \times n$, where each of the m rows corresponds to an observation and each of the n columns corresponds to an attribute ($D = \{x_{ij}\}_{i=1}^{m}, j = 1, ..., n$.)

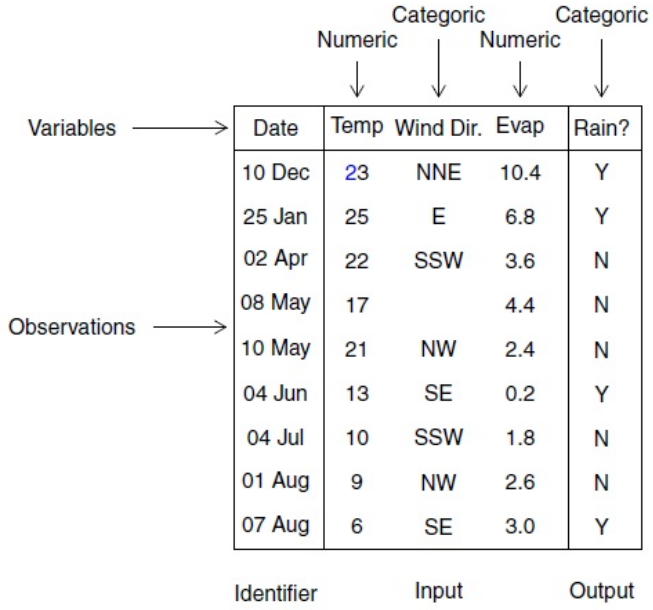


Date	Temp	Wind Dir.	Evap	Rain?
10 Dec	23	NNE	10.4	Y
25 Jan	25	E	6.8	Y
02 Apr	22	SSW	3.6	N
08 May	17		4.4	N
10 May	21	NW	2.4	N
04 Jun	13	SE	0.2	Y
04 Jul	10	SSW	1.8	N
01 Aug	9	NW	2.6	N
07 Aug	6	SE	3.0	Y

Data in the form of records

Diagnosis	GGT(u/l)	Cholesterol (mg/dL)	Albumin (g/dL)	Age (year)	Glycemia (mmol/L)	Sex
Cirrhosis	289	148	3.12	57	0.9	M
Hepatitis	255	258	3.5	65	1.1	M
Hepatitis	32	240	4.83	60	1.14	F
Hepatitis	104	230	4.06	36	0.9	F
Cirrhosis	585	220	2.7	55	1.5	F
Cirrhosis	100	161	3.8	57	1.02	M
Hepatitis	85	188	3.1	48	1.09	M
Cirrhosis	220	138	3.84	58	0.9	M
Cancer	1117	200	2.3	57	2.2	F
Cancer	421	309	3.9	44	1.1	M

Data in the form of records



- **Input variables** are also referred to as independent variables, observed variables or descriptive variables.
- **Output variables** are dependent on the input variables. They are referred to as target, response or dependent variables.

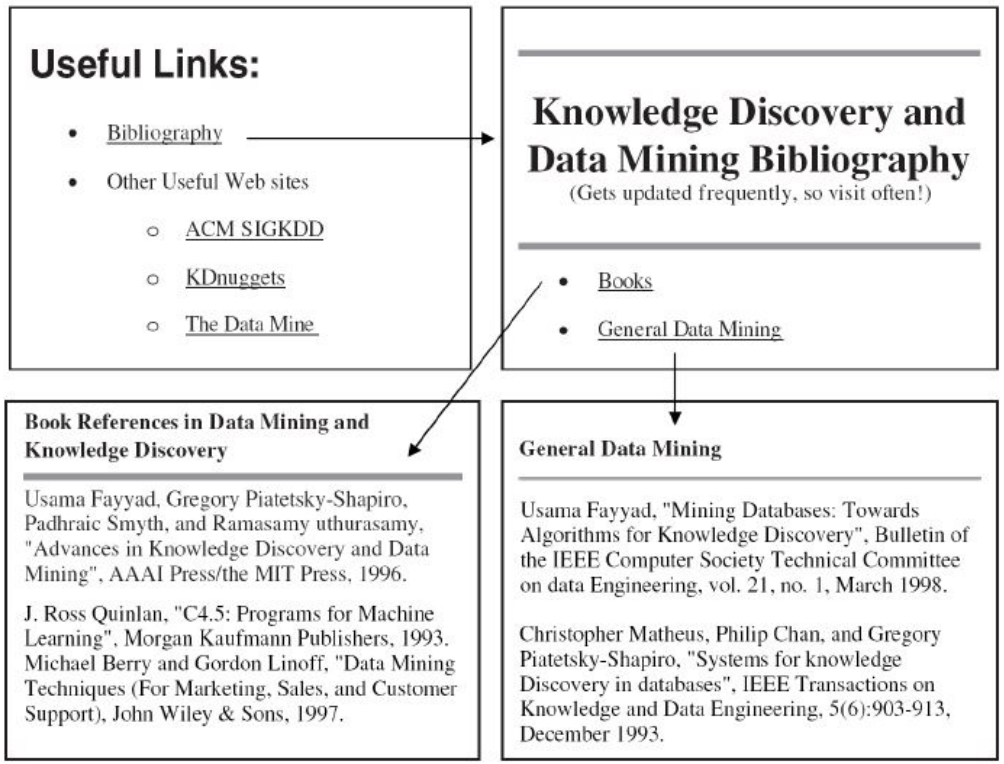
Transaction-based sets

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

- each transaction (purchase, observation) is assigned a vector.
- transaction components denote goods, objects, etc.

Data in graph form

The vertices of the graph are used to store the data, while the edges indicate the relationships between the data.



Data quality

- **Machine learning algorithms** are very sensitive to the quality of the source data;
- **GIGO (Garbage In, Garbage Out)** - results of processing incorrect data will be wrong regardless of the correctness of the processing procedure.

Data properties:

- completeness,
- correctness,
- actuality.

Noise in the data

- **label noise**
  - **inconsistent observations**
  - **classification errors** - observations that are labelled as a class other than the actual class.

Att 1	Att 2	Class
0.25	red	positive
0.25	red	negative
0.99	green	negative
1.02	green	positive
2.05	?	negative
=	green	positive

Often, the term **noised data** is used as a synonym for corrupted data.

Noise in the data

- **attribute noise** this refers to incorrect values of one or more attributes
  - **wrong attribute values** (1.02, green, class= positive) when we assume that an attribute has a bad value.
  - **missing or unknown attribute values**(2.05,?, class = negative) - we do not know the value of the second attribute.
  - **incomplete attributes or values** (=, green, class = positive) - the system cannot understand and correctly interpret values.
  - **outliers.**

Att 1	Att 2	Class
0.25	red	positive
0.25	red	negative
0.99	green	negative
1.02	green	positive
2.05	?	negative
=	green	positive

Types of Variable

We can divide attributes into just two types:

- **qualitative, (non-measurable)** (categorical) are data that cannot be uniquely characterised by numbers.
- **quantitative (numerical)**, e.g. the height of a person, the number of items sold articles;

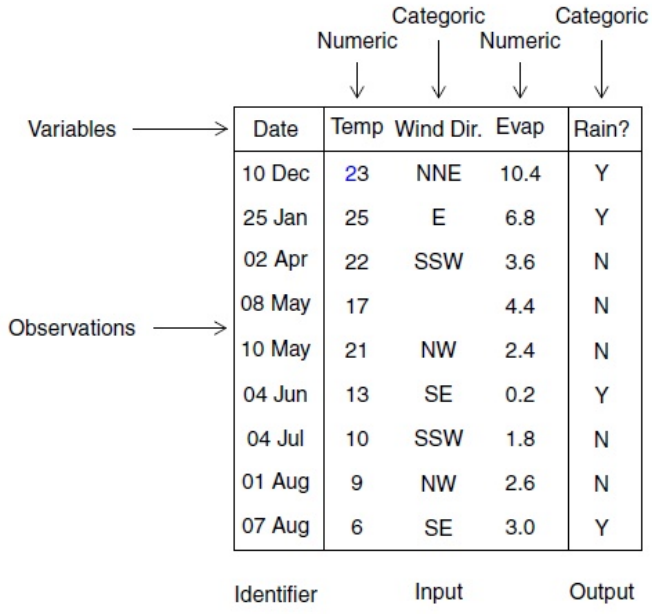
Qualitative data:

- the set of available values is always limited (e.g. in the case of months to 12); for this reason the values of categorical variables are called states;
- if the values can be compared with each other in a reasonable way (e.g. level of education), the categorical variable is called ordinal, otherwise we are dealing with a regular categorical variable (as for example in the case of gender);
- determining the distance between values is only possible within the framework of the adopted model;
- it is impossible to perform arithmetic operations on them.

Quantitative data:

- the values can be compared with each other;
- it is possible to determine the distance between values;
- arithmetic operations can be performed on them;
- the values can be
  - discrete (integers) - a finite or countable set of values;
  - continuous (real numbers).

Types of Variable dataset



Date	Temp	Wind Dir.	Evap	Rain?
10 Dec	23	NNE	10.4	Y
25 Jan	25	E	6.8	Y
02 Apr	22	SSW	3.6	N
08 May	17		4.4	N
10 May	21	NW	2.4	N
04 Jun	13	SE	0.2	Y
04 Jul	10	SSW	1.8	N
01 Aug	9	NW	2.6	N
07 Aug	6	SE	3.0	Y

Types of Variable

- Variables can take one or multiple values.
- **Single-valued variables are otherwise known as constants**.

Single-valued variables should not be used in the data analysis process as they carry no information.

- remark - it should be checked whether a given variable is single-valued in the selected sample or in the entire source data set - if some values of a variable occur very rarely (e.g. once in 500 000 cases), we will probably not find them in a sample of 10 000 rows. By removing such a variable from the training dataset of a data mining model, we may not only degrade the accuracy of its results, but also prevent it from recognising unusual cases, possibly the most interesting ones.

Types of Variable

- **Some variables are used to uniquely identify the observation**, this could be, for example, some official identification number. The identifier can also be, for example, the date of the observation.
- **Identifiers** are not used in the modelling.

Types of Variable

Another type of variables not useful for predictive models are **monotonic variables.**

- The values of such variables constantly increase or decrease.
  - this type of variable is very common, for example: the values of all time-related variables (such as invoice date or date of birth) are increasing;
  - wartości wielu niezwiązanych bezpośrednio z czasem zmiennych również są rosnące lub malejące – należą do nich między innymi numery faktur numery samochodów.

Knowledge and information

- Let us assume that we run a local shop and that we register all the details in the shop's database. We know our customers' details and what they buy each day.
- E.g. Alex , Jessica and Paul visit the shop every Sunday and buy candles. What we store in the database is just the **data**.
- Every time we want to know who the visitors are who buy the candles, we can search the database and get the answer. This is **information**. If we want to know how many candles were sold on each day of the week, then we can again direct a query to the database database and get the answer - this is also **information**.

- But suppose we have many other customers who also buy candles from us every Sunday (mostly with some level of freedom), and all of them are Christians (going to church). So so we can conclude that Alex, Jessica and Paul must also be Christians. The religion of Alex, Jessica and Paul was not recorded in the database, so we could not retrieve it as information from it. We learned this information indirectly. It is a **knowledge** that we discovered.
- Of course, it is likely that our findings as to Alex, Jessica and Paul may be wrong. Therefore it is important that our knowledge and findings are evaluated correctly.

Data pre-processing

Data pre-processing involves cleaning and transformation of data to prepare it for mining. It is estimated that data preprocessing is 70-80% of the process of of knowledge discovery.

For example, a database may contain

- fields that are out of date or redundant,
- records with missing values,
- outliers,
- data in a format unsuitable for machine learning models,
- values incompatible with principles or common sense.

To illustrate the necessity of data cleaning, we will examine the following example (D.Larose):

We will analyse the personal data from the following table

Customer ID	Zip	Gender	Income	Age	Marital Status	Transaction Amount
1001	10048	M	78,000	C	M	5000
1002	J2S7K7	F	−40,000	40	W	4000
1003	90210		10,000,000	45	S	7000
1004	6269	M	50,000	0	S	1000
1005	55101	F	99,999	30	D	3000

Example

Outliers

↳ Hay que analizarlos. En el caso de 10000000 no se trata de un error ya que el ZIP code 90210 es de Beverly Hills, un barrio rico

Thank you for your attention!!!