Machine learning

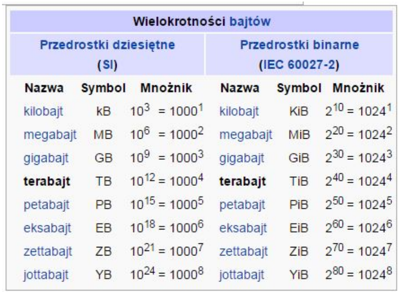
Bibliography

1. Larose, Daniel T. Discovering Knowledge in Data : An Introduction to Data Mining, Wiley, 2014.
2. Han Jiawei, Kamber Micheline, Data Mining: Concepts and Techniques. Elsevier, 2006.
3. Pang-Ning Tan, Steinbach Michael, Vipin 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. Raschka Sebastian, Yuxi Liu, Vahid Mirjalili Machine Learning with PyTorch and Scikit-Learn. Packt, 2022.

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.

What big data looks like?

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	$2^{80} = 1024^{8}$

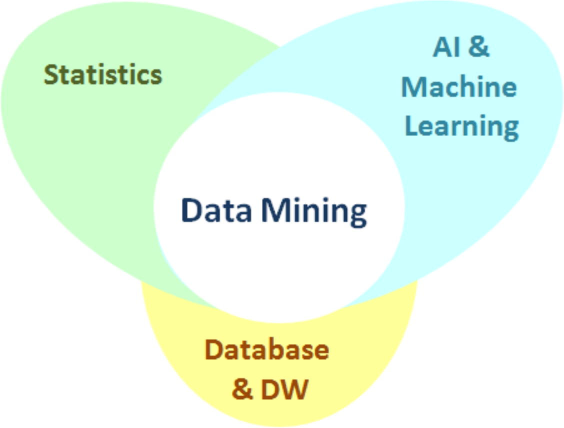
Machine learning

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

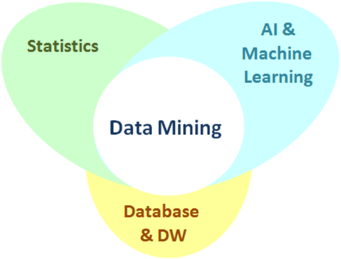
Machine learning (ML) is a term that was coined around 1960, consisting of two words— **machine**, which corresponds to a computer, robot, or other device, and **learning**, which refers to an activity intended to acquire or discover event patterns, which we humans are good at.

Machine learning

Machine learning (ML) machine learning is often also referred to as data mining or predictive analysis.



Relation with Artificial Intelligence (AI)



- Artificial intelligence (AI) is a much broader field of study than **machine learning (ML)** (ML is a subfield of AI).
- AI is all about making machines intelligent using multiple approaches, whereas ML is essentially about one approach – making machines that can learn to perform tasks. An example of an AI approach that's not based on learning is developing **expert systems**.

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.

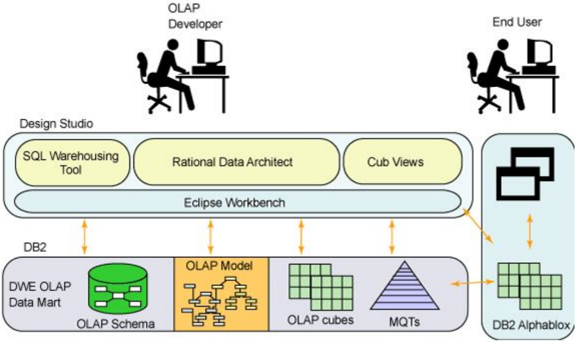
Data mining vs Machine learning

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**.



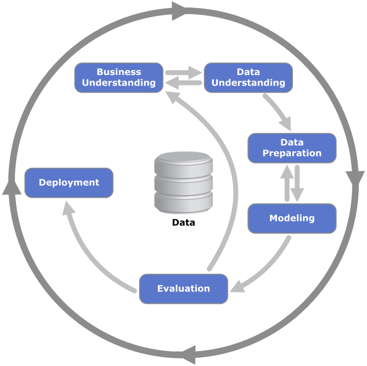
What is not dm and ml

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?

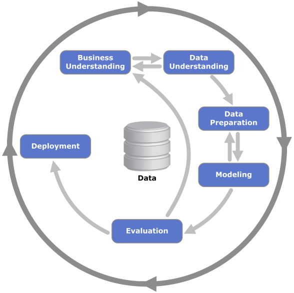
Data analysis process

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



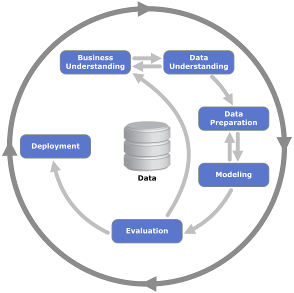
1. Problem Understanding lub Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

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

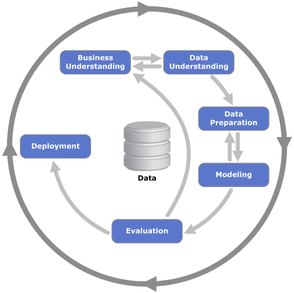


- The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems.

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?

Data analysis process - Data Preparation

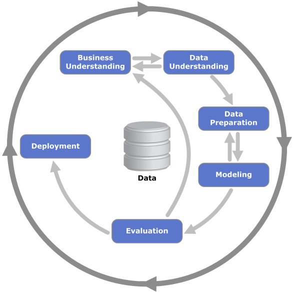


- The data preparation phase covers all activities to construct the final dataset from the initial raw data.

Data preparation most often requires:

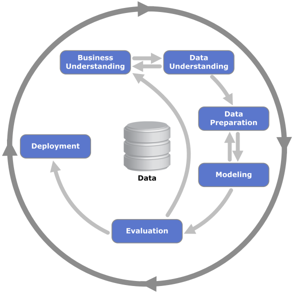
1. the joining of several data sets,
2. reducing the number of variables to only those which will be relevant for the process,
3. data cleaning (removal of anomalies, reformatting, normalisation, missing data).

Data analysis process - Modeling



In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

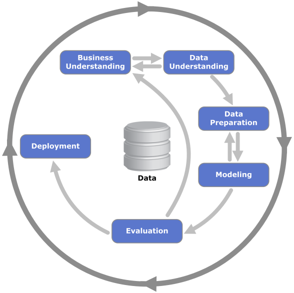
Data analysis process - Evaluation



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

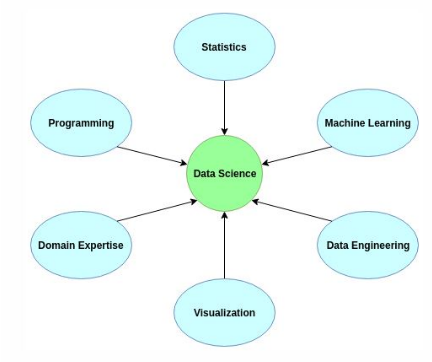
Data analysis process - Deployment



- 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/

Python 4 ML

Why Python is Preferred for Machine Learning?

- Python is known for its readability and simplicity , making it easy for beginners to grasp and valuable for experts due to its clear and intuitive syntax.
- Python offers many libraries and frameworks for machine learning and data analysis, such as **Scikit-learn, TensorFlow, PyTorch, Keras, and Pandas.**

These libraries provide prebuilt functions and utilities for mathematical operations, data manipulation, and machine learning tasks, reducing the need to write code from scratch.

Python 4 ML

Why Python is Preferred for Machine Learning?

- Python has a large and active community, providing ample tutorials, forums, and documentation for support, troubleshooting, and collaboration.
- The community ensures regular updates and optimization of libraries, keeping them up-to-date with the latest features and performance improvements.

Python's flexibility makes it suitable for projects of any scale, from small experiments to large, complex systems, and across various stages of software development and machine learning workflows.

- **NumPy** This library is fundamental for scientific computing with Python. It provides support for large, multi-dimensional arrays and matrices, along with a collection of high-level mathematical functions to operate on these arrays.
- **Pandas** Essential for data manipulation and analysis, Pandas provides data structures and operations for manipulating numerical tables and time series. It is ideal for data cleaning, transformation, and analysis.
- **Matplotlib** It is great for creating static, interactive, and animated visualizations in Python. Matplotlib is highly customizable and can produce graphs and charts that are publication quality.

Python 4 ML

- **Scikit-learn** Provides a range of supervised and unsupervised learning algorithms via a consistent interface. It includes methods for classification, regression, clustering, and dimensionality reduction, as well as tools for model selection and evaluation.
- **SciPy** Built on NumPy, SciPy extends its capabilities by adding more sophisticated routines for optimization, regression, interpolation, and eigenvector decomposition, making it useful for scientific and technical computing.

Thank you for your attention!!!