Tampereen yliopisto

aistico

# DEEP DIVE INTO PRACTICAL MACHINE LEARNING

***WITH SOME EVANGELIZING ON ARTIFICIAL INTELLIGENCE ON TOP***

Lakeus Devaus Meetup

5.5.2020 / **Jussi Rasku**

About the affiliations: I'm currently postdoctoral research fellow at Tampere University, positioned at the University Consortium of Seinäjoki. However, some in the audience may know me as the founding member and previous chair of the Sepeli ry (Seinäjoen pelikehittäjät). Recently I've been working also for myself with a ltd company Aistico Oy me and few friends put together in 2020.

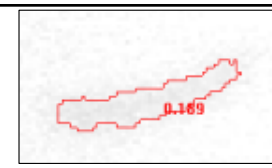
By the way, the images are generated from variations of the title prompt using an open-source AI called dalle-mini: https://huggingface.co/spaces/dalle-mini/dalle-mini

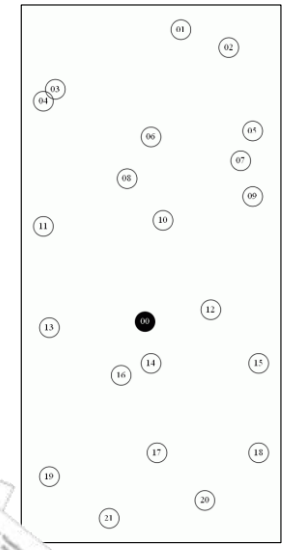
Tampereen yliopisto

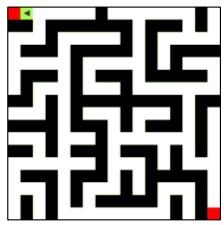
# AI AND ME

- My background is in SWE with an AI twist:
  - *"Statistical pattern classifier in machine vision quality control"* (2010), MSc. thesis, TUT.
  - *"Toward Automatic Customization of Vehicle Routing Systems"* (1.11.2019), **PhD** disseration, University of Jyväskylä.

  + many side/work/hobby projects.





AI and I have a long history together. Side projects have ranged from text, speech, machine vision analysis and game AI's and procedural content generation.
Taken together, have learned one or two things on the topic and even written a book on the topic: https://jyx.jyu.fi/handle/123456789/65790

Tampereen yliopisto

## CONTENTS

**Audience: SWEs, not data scientists**
**(no heavy math, and our goal is practical conceptual understanding)**

1) **AI recap (motivation)**
2) **Data is king (incl. preprocessing)**
   **or**
3) Machine learning methods
4) ~~Some practical tools and best practices~~

"Next time, I promise" 😂

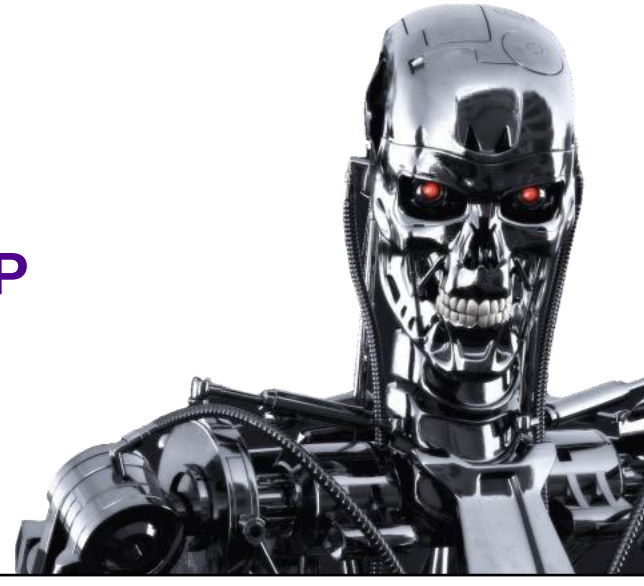
**Today, our goal is to form a mental image on what AI/ML is, what is needed, and what one can expect to do with it.**

We do not have the time to go through all of the material, but we can also pick and choose the content that you find most interesting.

Tampereen yliopisto

For those that missed
the 2020-10-23 meetup

# 1) AI RECAP



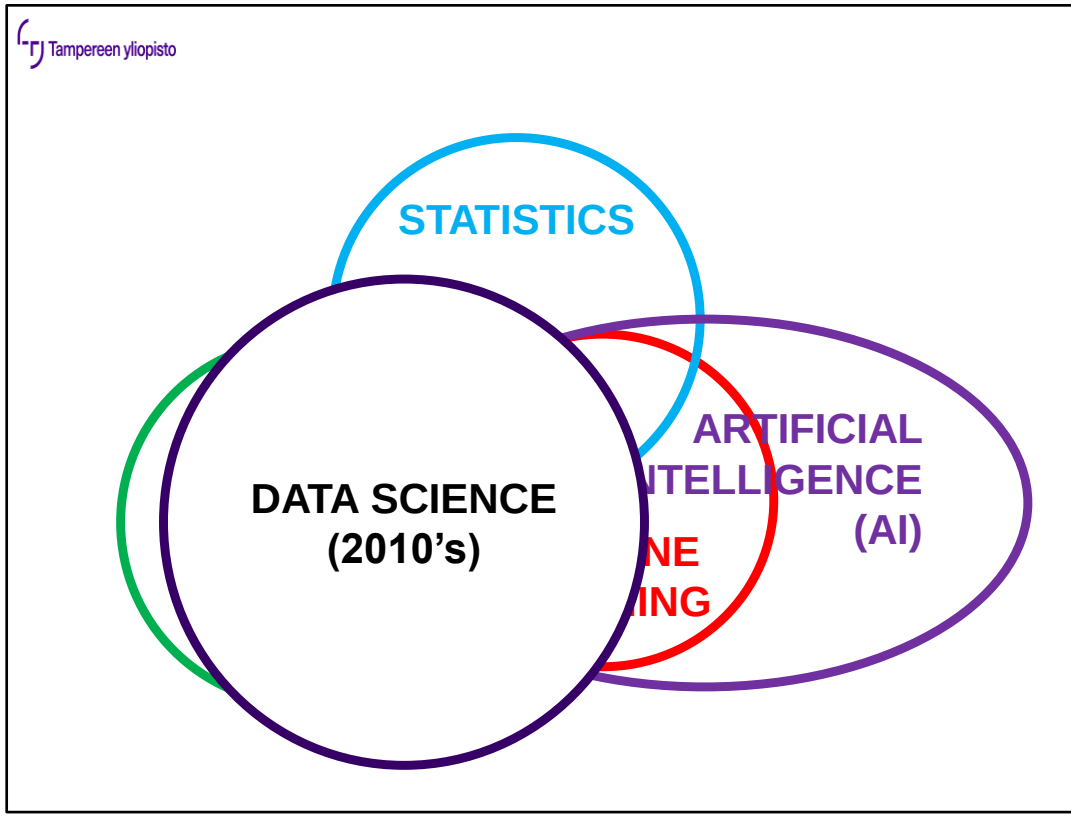
It is good to have a short recap for those that missed the previous meetup on 2020-10-23 meetup . This is important as there are lots of ***misconceptions*** regarding AI. Even the popular culture most often than not sees it as a threat, not as an opportunity.

Tampereen yliopisto

# WE STAND ON THE BRINK OF THE NEXT TECHNOLOGICAL REVOLUTION

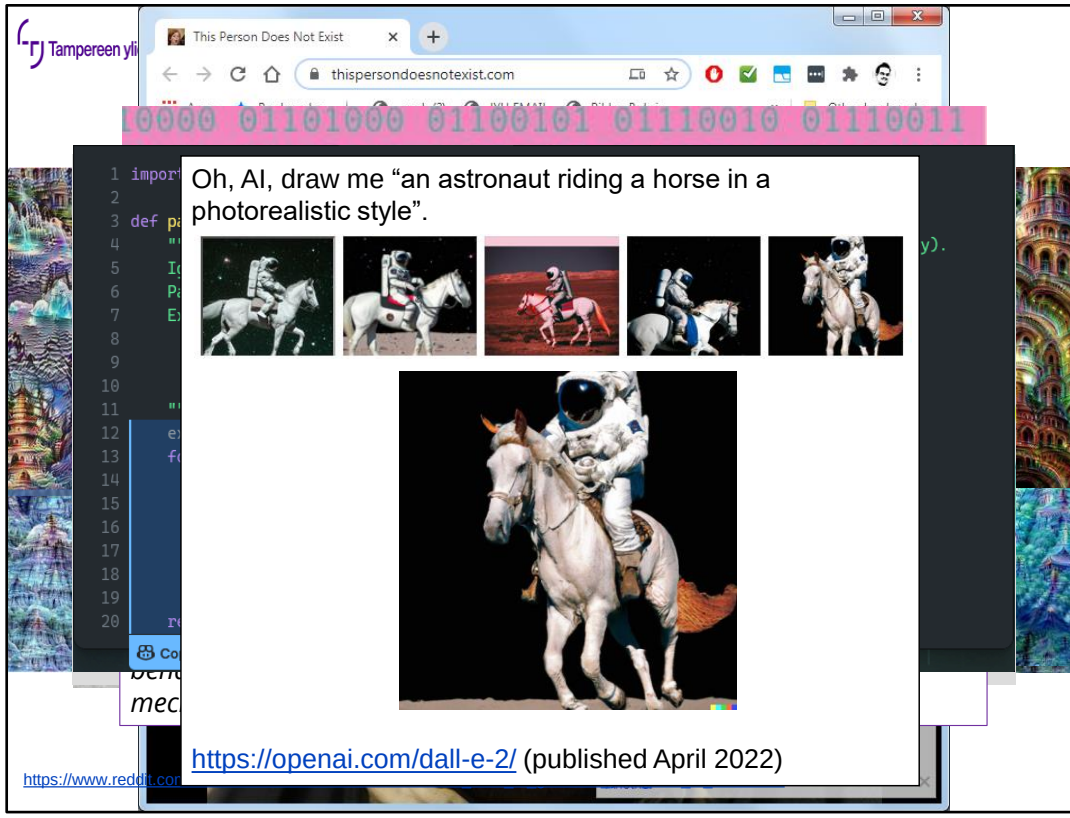
I strongly believe that we are on the brink of the next industrial revolution. This time the driver is not the steam engine, electricity or even the transistor, but **an intelligence that can be copied.**

Exponential change is hard to see up close as the progress seems linear. However, looking historical trends, the use of artificial thinking is rapidly accelerating.




The field is full of concepts that mean almost, but not quite, the same. This is because of the long history of the field and perverse academic incentives: one who comes up and publishes a new popular term will be cited a lot. This is why the same topic that combines data wrangling with applied statistics, and we have gone from Expert systems, Data Mining to Data Science.

What they all do is they manage data and try  increase its refinement level. Hence, a new field has been born that tries to capture these.




Arts, composer, journalist, secretary, taxi driver, doctor, even coder. It seems no occupation is safe. After GPT-3 I became more certain that human level AI is created within few decades. After Dall-e 2, I’m convinced.

Tampereen yliopisto

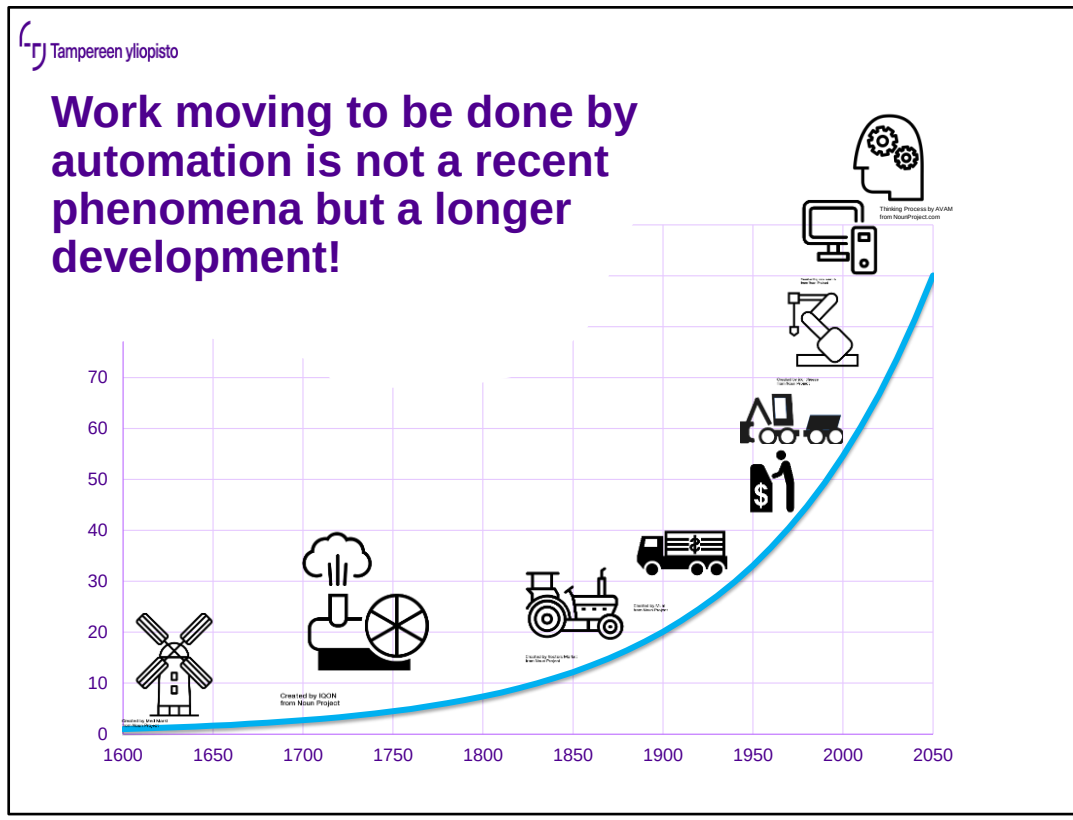
## I, FOR ONE, WELCOME OUR NEW COMPUTER OVERLORDS (OR AUGMENTED HUMANS)


From CGP Grey

*Brynjolfsson and A. McAfee, Race Against The Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy. Digital Frontier Press, 2011.*

However, now and in the near future we still can do better, not by competing with the machine, but augmenting ourselves with AI. Computer is a tool that allows us to think thoughts were not otherwise possible.

A quick introduction can be found from the CGP Grey video:
https://www.youtube.com/watch?v=7Pq-S557XQU
And a good relatively concise and optimistic book of the topic is
https://www.amazon.com/Race-Against-Machine-Accelerating-Productivity-ebook/dp/B005WTR4ZI



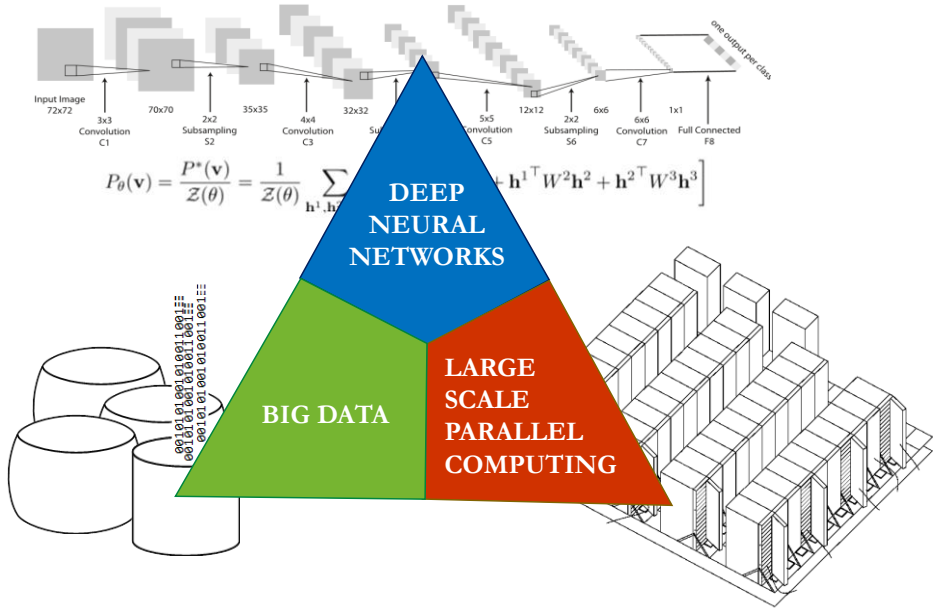

The first fully functional and practical steam-powered device was for a mining water pump. It was patented by Thomas Savery in 1698. It took a while, but Watt and Boulton made the first shaft-rotating machine in 1785. The first agricultural tractors were built in the mid-19th century and by the 20th century they had become mostly rely on internal combustion engine. Then, the proliferation of cars and especially trucks yielded a significant improvement in logistics.

Meanwhile in Finland … the first forest harvester PIKA 75 was introduced by system engineer Sakari Pinomäki in 1973. So what happened? By the 70's, there were an estimated 100k loggers in Finland. Today? -5k lumberjacks. The hard work shifted to the machines. And no wonder when I compare how long I have to felling and pruning a spruce compared to how Ponsse's machine does the job (20 min vs 20 seconds).

And it does not stop there: Automatic cash registers, eCommerce outcompeting brick-and-mortar, automated securities trading, self driving cars, robot chefs, power robots, etc. We have only seen the beginning of robotization of the society. This all means that more and more work is done with mechanical muscles. And now even thinking and creativity are being automated. The change will be comparable to the changes brought about by industrialization.

Tampereen yliopisto

## The new age of AI is driven by …

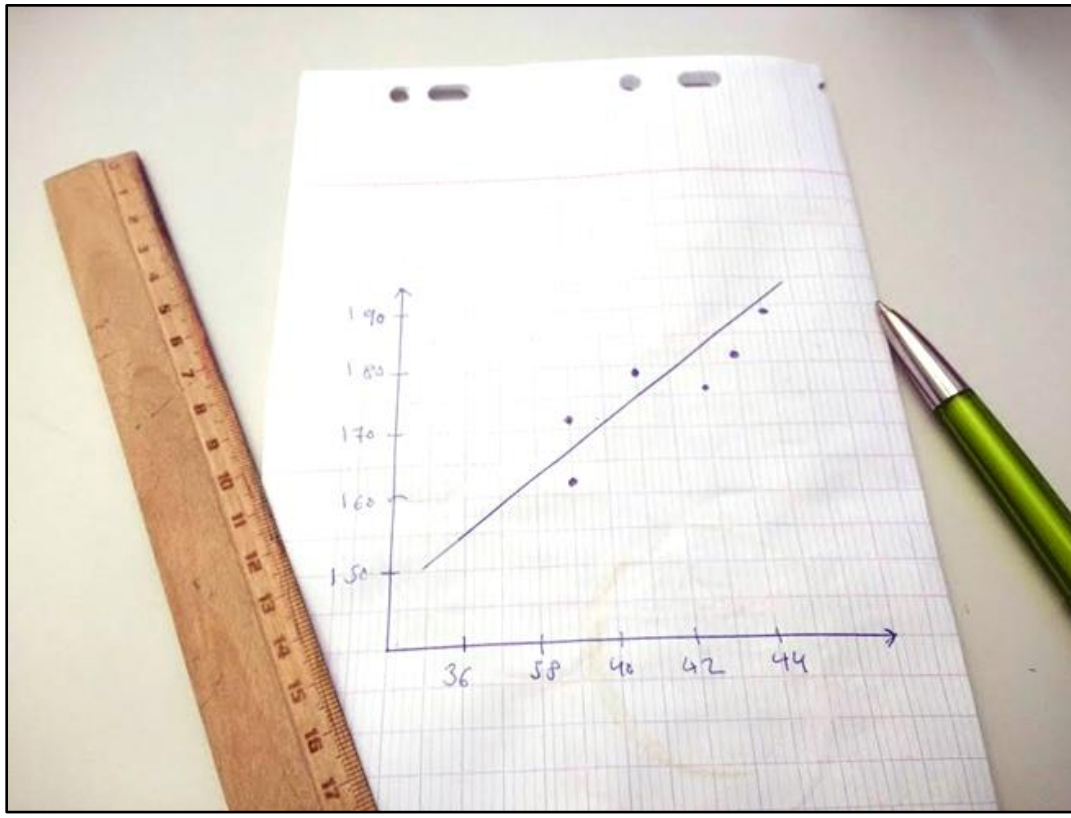



One might remind me that there has been two AI winters 1974-1980, 1987-1993. During these times there was less money and interest (the hype bubbles burst).

So, what is different **NOW**? Recent advances have been powered by:

1) Big data (cheap to store, cheap to record, cheap to move around)
2) Massive parallel computing capacity. Datacenters are buildings stacked full of computers.
3) New breakthroughs in the neural network architecture and training methods.

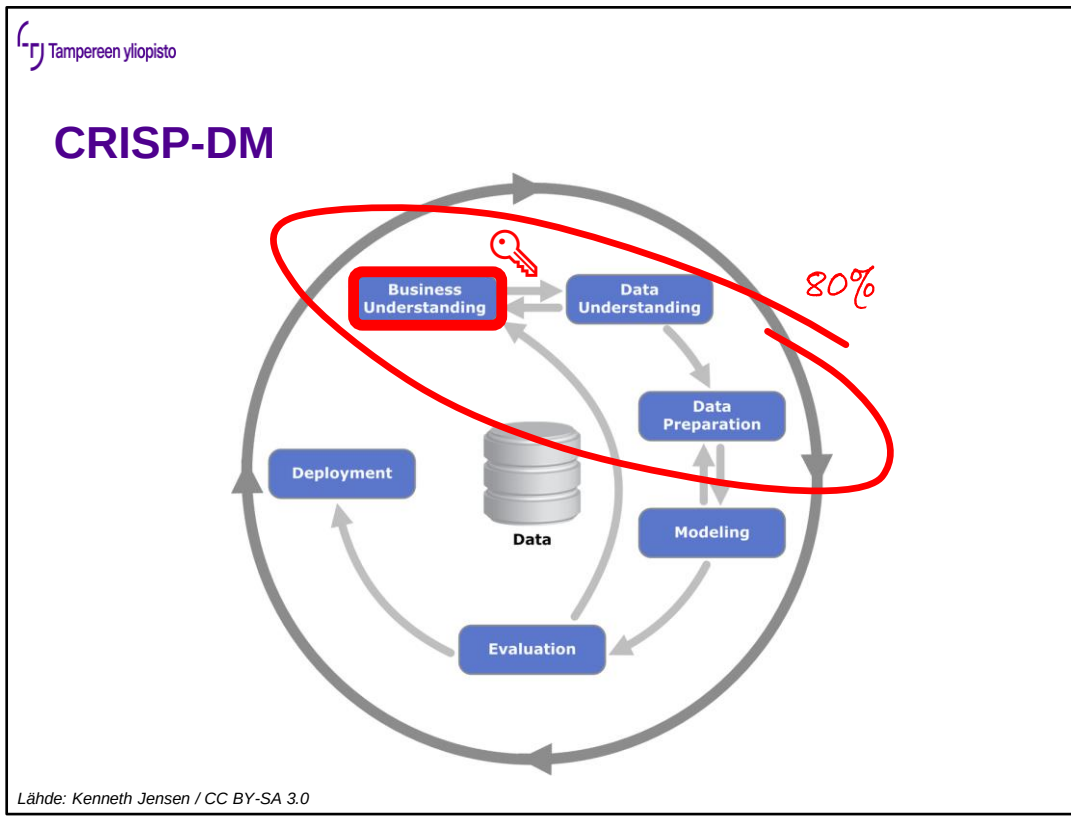
All this means we can do useful stuff. Truly useful stuff.




This does not seem too scary, right?

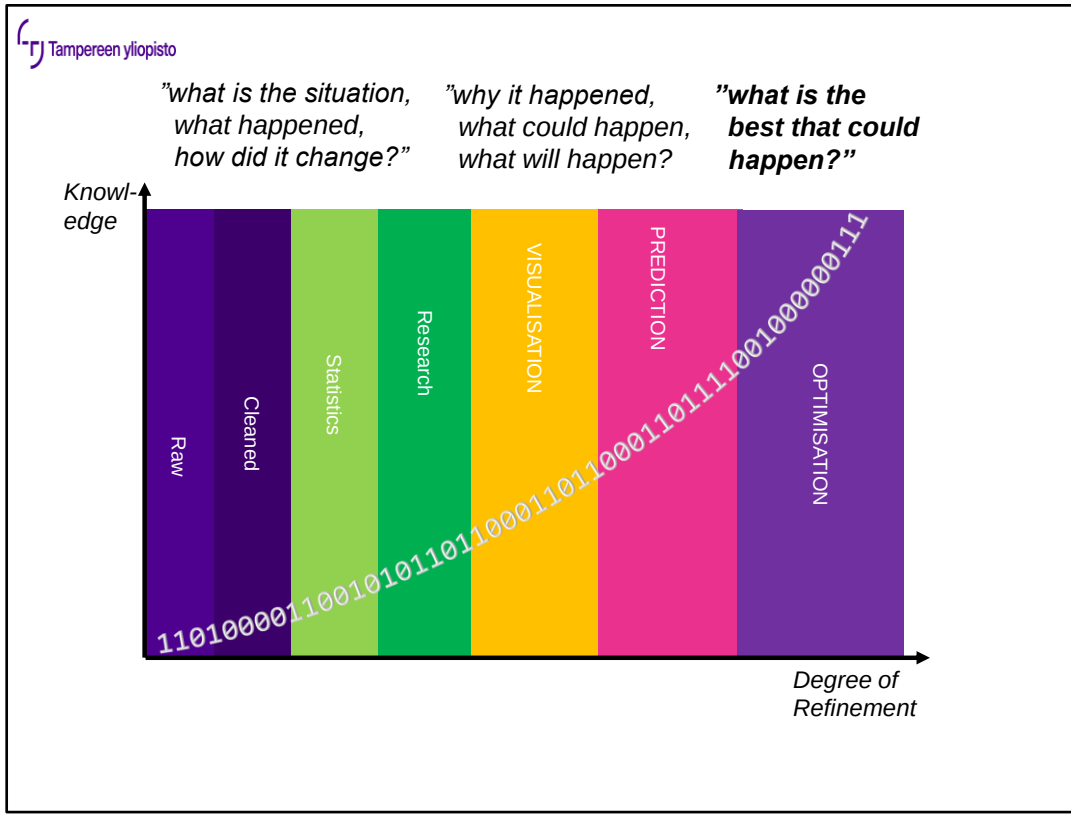
There is nothing mystic in machine learning. Ultimately it is nothing else than that what have been done in the elementary school physics classes for decades: **fitting a curve to the data.** Only the complexity of the task increases in typical machine learning tasks.

**Still: Is the fit good? optimal? Does it make any difference?**




Cross Industry Standard Process for Data Mining - a data mining process model.
SPSS Modeler product uses this!

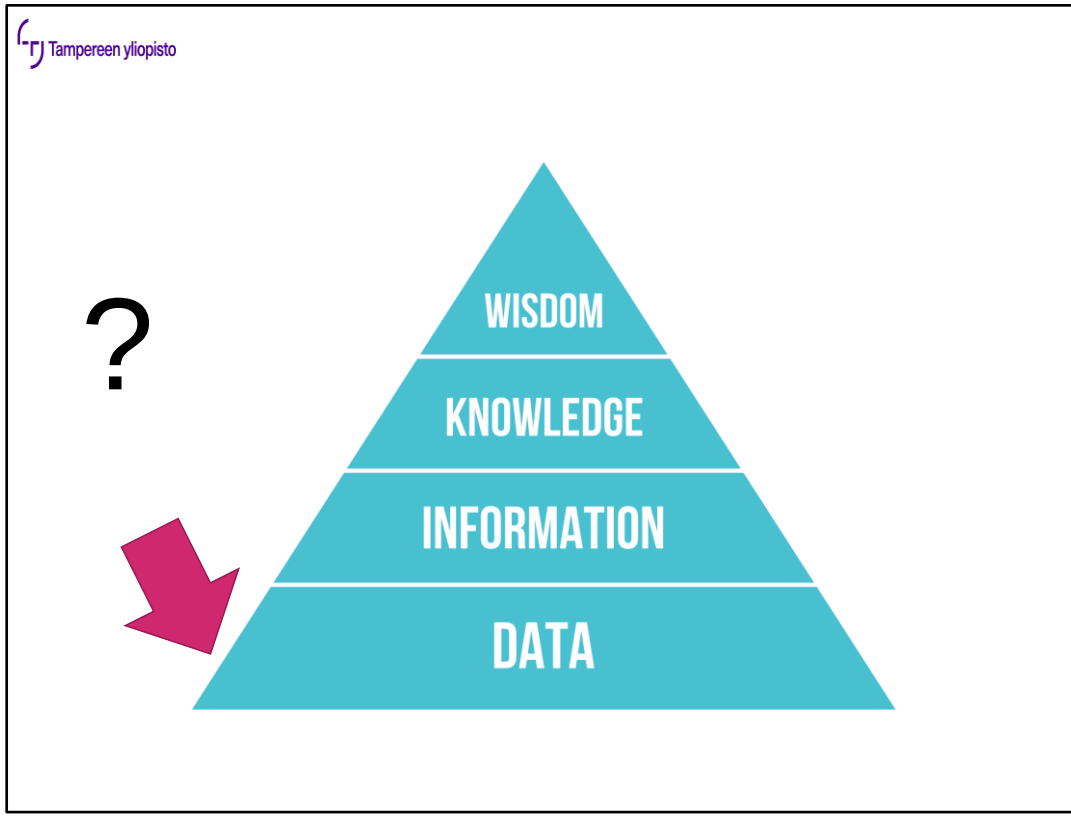
**Business Understanding!!:** Understanding the project objectives and requirements from a business perspective
Converting this knowledge into a data mining problem definition and plan
**Data Understanding:** Initial data collection and activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets.
**Data Preparation:** Activities to construct the dataset that will be fed into the modeling tool(s). Multiple iterations. Data and attribute selection, transformations, inputations, cleaning up.
**Modeling:** Various modeling techniques are selected and applied. Typically, there are several techniques for the same data mining problem type. The preparation may need to be revisited.
**Evaluation:** Evaluate the built model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached.
**Deployment:** The knowledge gained from the model will need to be organized and presented in a way that is useful to the customer/context. The actions which will need to be carried out in order to actually make use of the created models.




The data in itself is worthless. Only when we can use it to answer questions, some valuable knowledge, or even wisdom is created.

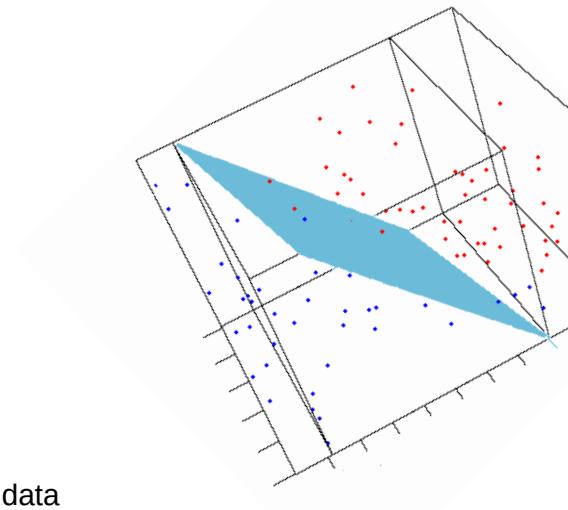
The more traditional methods only answer to questions related to current or past state of affairs.
Visualization and prediction are needed to gain understanding of the ongoing trends and future events.

Also, as an optimization person I'd like to think that optimization is the best we can do with the data as it
Allows answering the question ***"what is the best that could happen?".***




The data is the foundation, but distilling it into information or even knowledge with ML is perhaps more interesting? You decide!

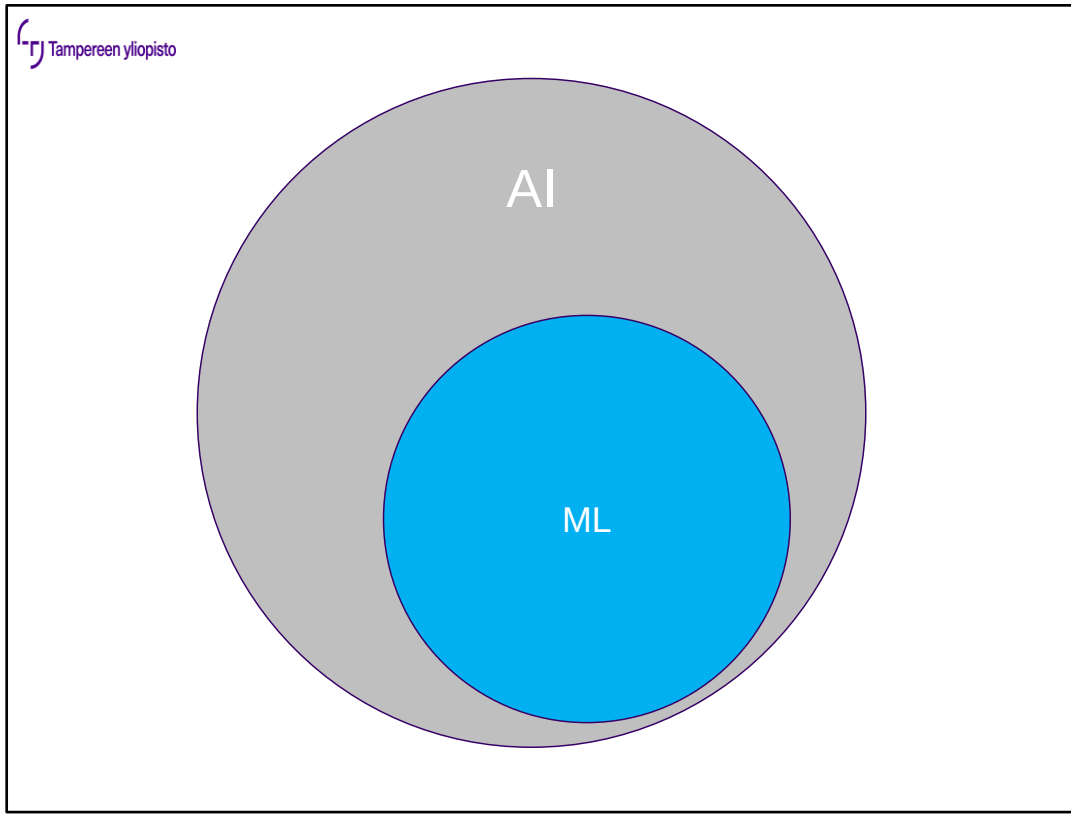
Tampereen yliopisto



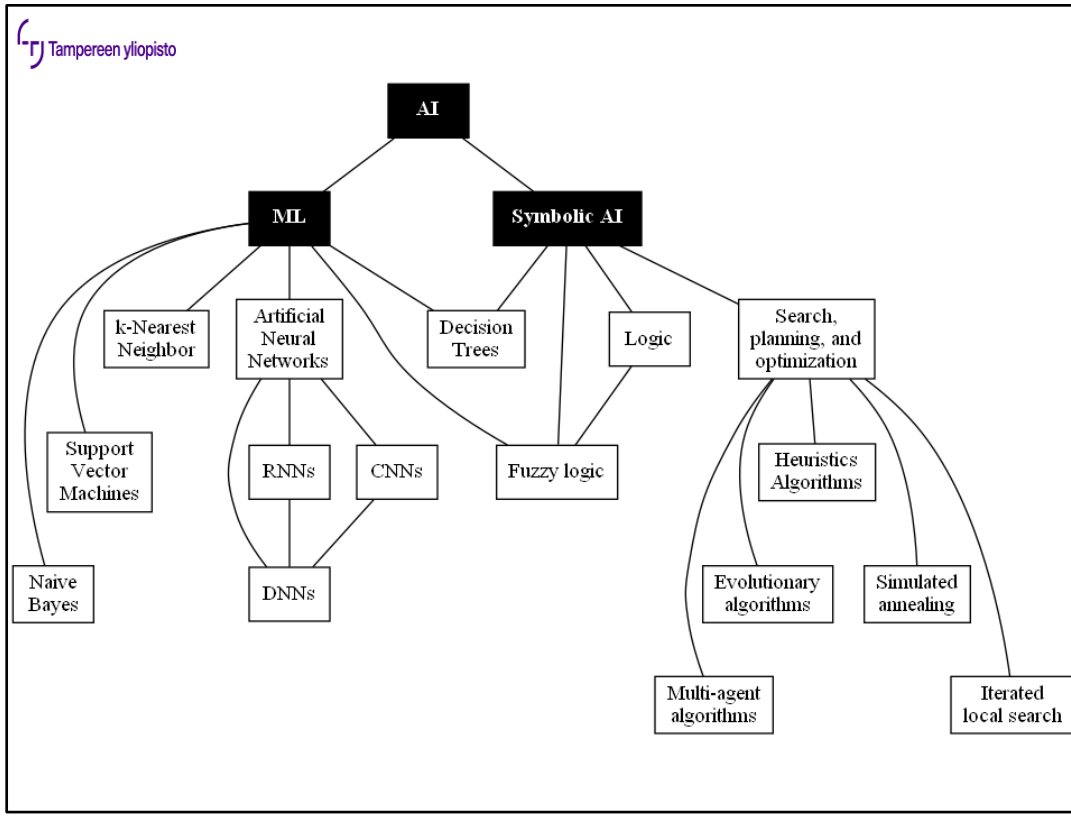
Finally, fitting the model to the data

# 2) MACHINE LEARNING

ML allows computer programs to adapt its behavior via observations. Can augment or replace humans in some tasks. Makes it possible to build self-improving tools.




By concentrating on ML, we rule out part of AI methods outside of this discussion (not all AI needs data). More of that on the next slide. But do not worry, ML by itself is a large enough topic to cover!




Symbolic AI is ”just” behaviour/intelligence a human has encoded into the machine. Even adaptivitiy is something that is more or less ”hard-coded”.

With ML the approach is different. Just throw (good quality) data and significant amount of compute to the problem until you start seeing results.

Highly educated human labor is very expensive, hence the “win” of ML over symbolic ML (and the saltyness of its proponents).

However, there are some problems that are difficult for the data+compute approach. Many search, planning and optimization problems rely on heuristics (i.e., manually crafted algorithms) and approaches to build machine “intuition” to solve these problems has met lackluster success. Still, ML and the “intuition” the model learns, can be used to steer those lower-level algorithms. This was the focus of my PhD.

Tampereen yliopisto

## Reminder : How to apply machine learning

1. **Find out what is the goal?**
2. Get the data
   collect, choose, and store
3. Do data preprocessing
   cleanup, merging, joining, transformations and pruning
4. **Choose the model, algorithm and do the training**
   **incl. regression, association rule mining, classification and clustering**
5. **Interpret, evaluate, and validate**
6. Leverage the increased understanding

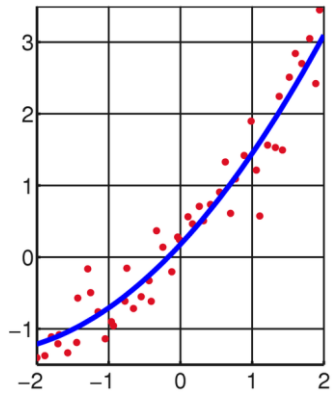
Ask the right questions, and choose the models based on that **when you know what you are trying to do.**

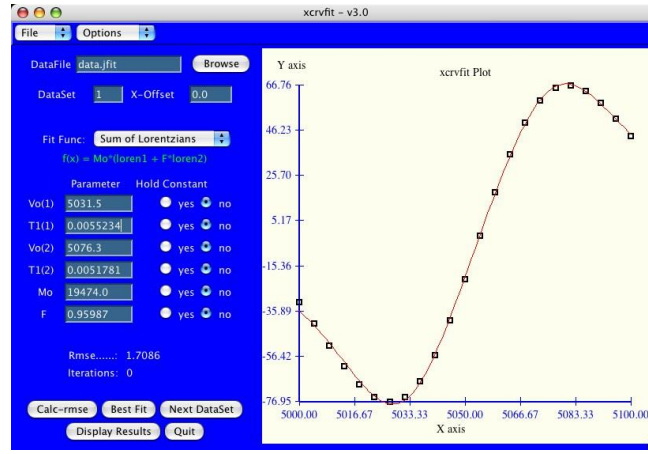
Tampereen yliopisto

# 3a) REGRESSION

**The task:**
For some numbers (feature values $x$), determine the output value $y$.

- Regression (linear, non-linear)
- I.e., Correlations
- I.e., ”curve fitting”



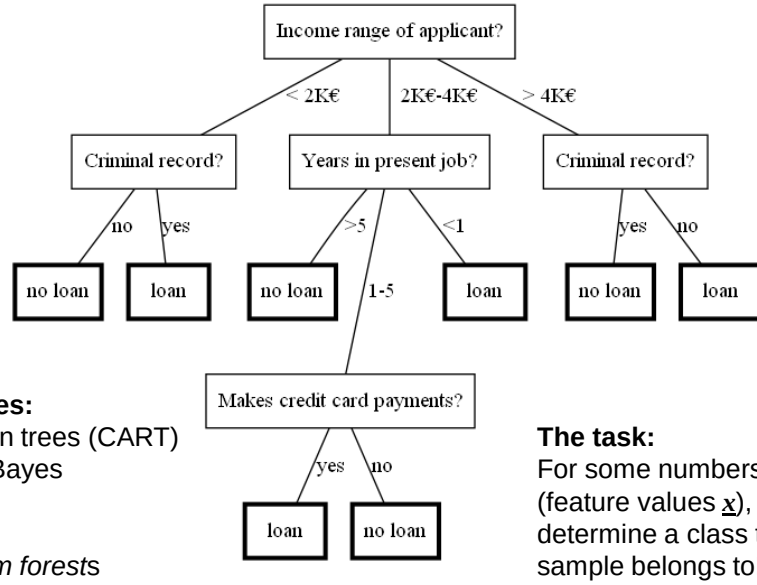




*For example, xcrvfit has some 30 curves to fit!*

To do the fitting the well known least squares method can be used or some more advanced optimization technique.

Tampereen yliopisto

# 3b) Classification

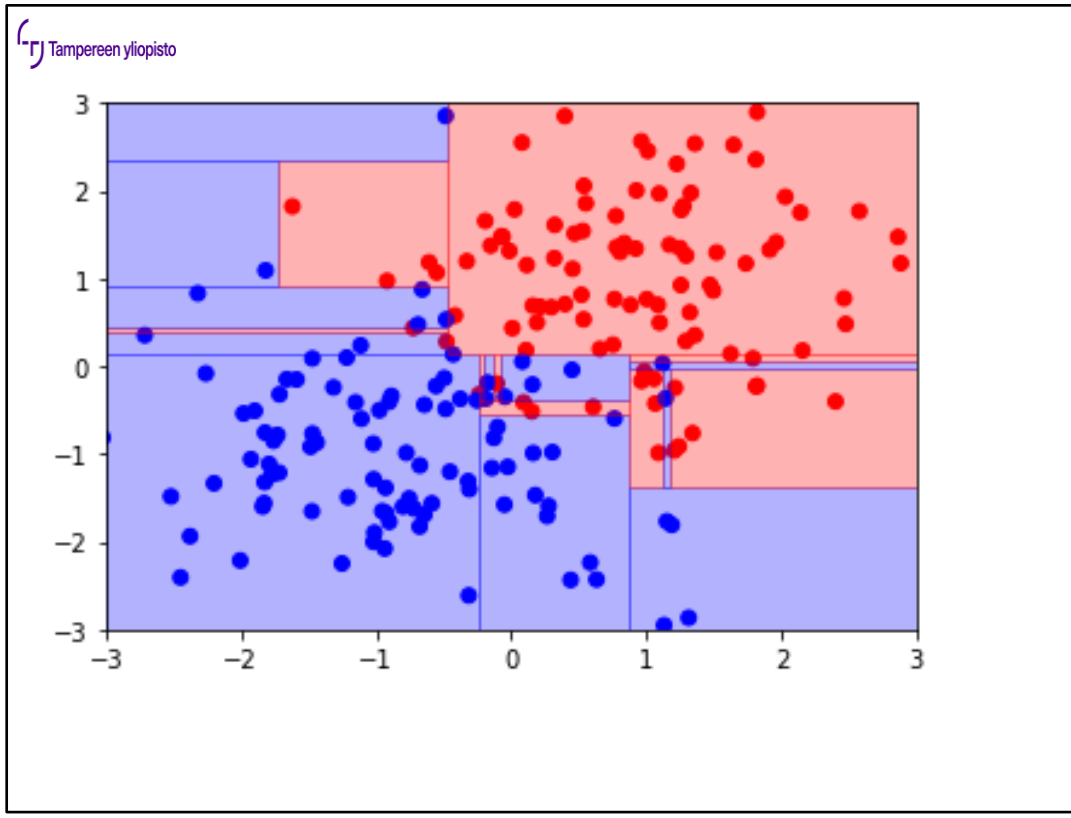



**Techniques:**

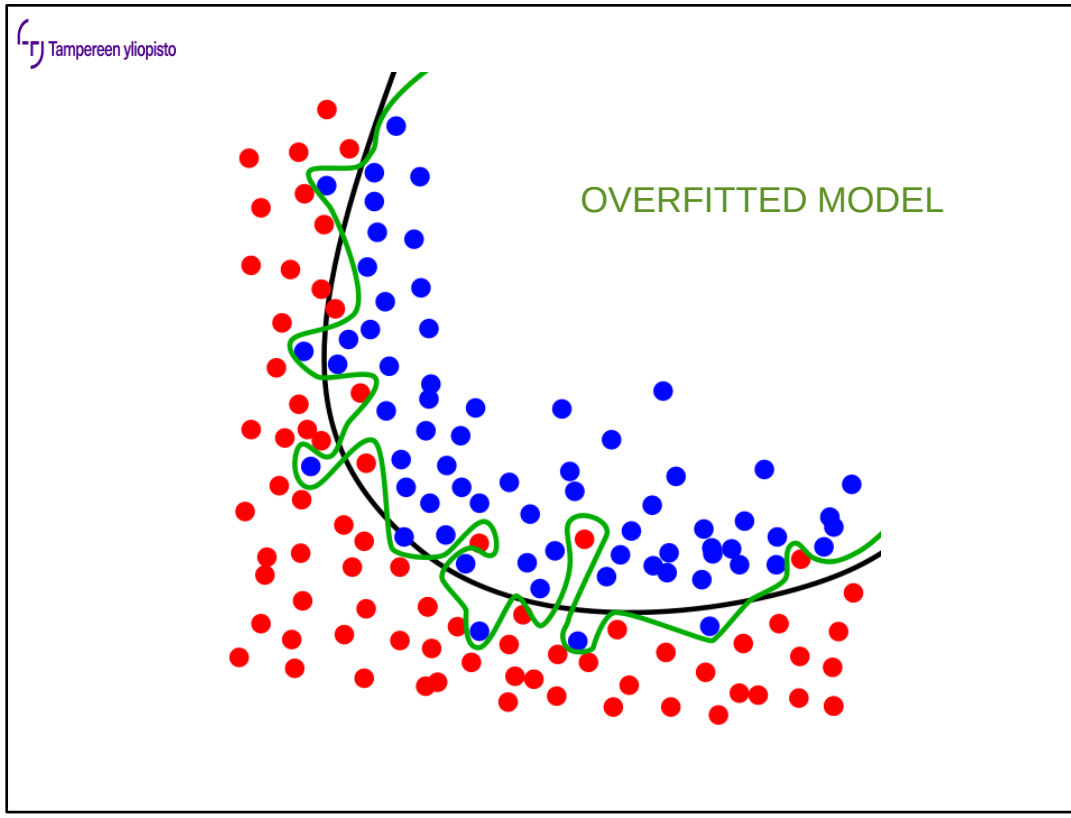
- Decision trees (CART)
- Naïve Bayes
- k-NN
- SVM
- *Random forest*s
- …

**The task:**
For some numbers (feature values ***x***), determine a class the sample belongs to!

If more performance is needed, bagging and boosting can increase robustness and accuracy. Basically, train and use many models in parallel and give 1 vote for each.




Decision trees split the space to rectangular “blocks”. One can see how this fails in some border cases. E.g., setting a point that should be blue at roughly (-2, 0.4) would get a red label.

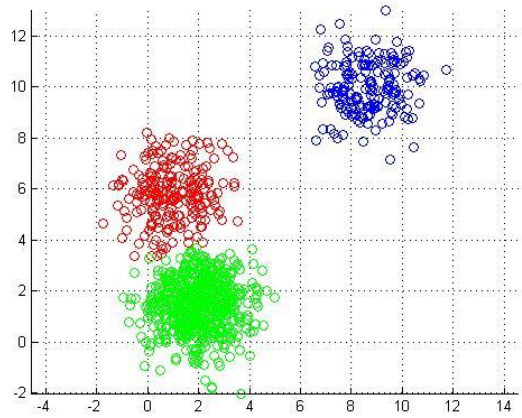



How a good model is, is outside of the scope of this talk. Basically, it depends on the exact methodology and oftentimes there are several alternatives for searching a suitable model parameters. That is, it can be seen as an optimization problem and tens, perhaps hundreds different optimization techniques can be applied.

However, it is good to be aware of the dangers of overfitting. More accurate and fine-grained model (green line) is not necessarily the best. One seeks a model that can capture the phenomena in a way that is generic (black line). The real world data is always messy and full of static. It is important to take this into account when training the model.

Tampereen yliopisto

## 3c) Clustering




**Methods:**

- k-Means
- DBSCAN / OPTICS
- BIRCH
- EM
- …

Note: The metric used to calculate distances is **very, very** important here. Not all dimensions are necessarily commeasurable and for clustering one may need to carefully consider each dimension/feature and scale them accordingly (so that no single feature/variable/dimension dominates when doing the clustering).

k-Means -> Centroid based clustering
DBSCAN -> Density based clustering
BIRCH -> Hierarchical clustering (connectivity-based clustering)
EM = Expectation–maximization -> Distribution based clustering

Tampereen yliopisto

# 3d) Association mining

| Trans-action | milk | bread | butter | beer |
|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 |
| 2 | 0 | 0 | 1 | 0 |
| 3 | 0 | 0 | 0 | 1 |
| 4 | 1 | 1 | 1 | 0 |
| 5 | 0 | 1 | 0 | 0 |

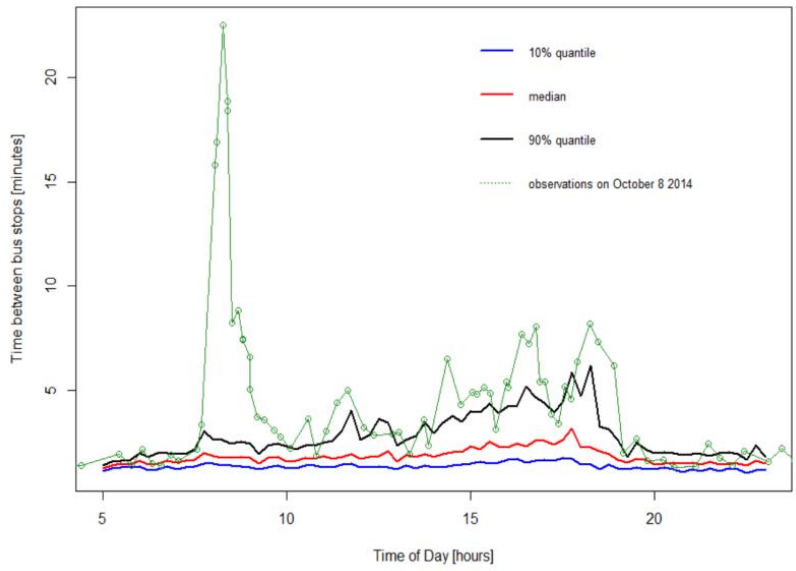
**Methods:**
- Apriori
- GUHA/ASSOC
- OPUS
- …


$$\{\text{milk}, \text{bread}\} \Rightarrow \{\text{butter}\}$$

Of the methods, Apriori is classical but there are others like GUHA.

Tampereen yliopisto

## 3e) *ANOMALY DETECTION*

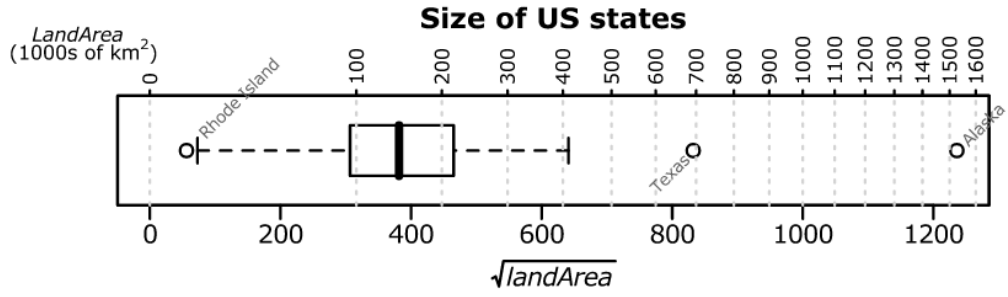



*Lähde: Nummenmaa, Metodifestivaali 2015: "Big data"*

One typical application is network security and detecting suspicious behavior from the network logs. Another could be, for example, early detection of machinery failure.

Tampereen yliopisto

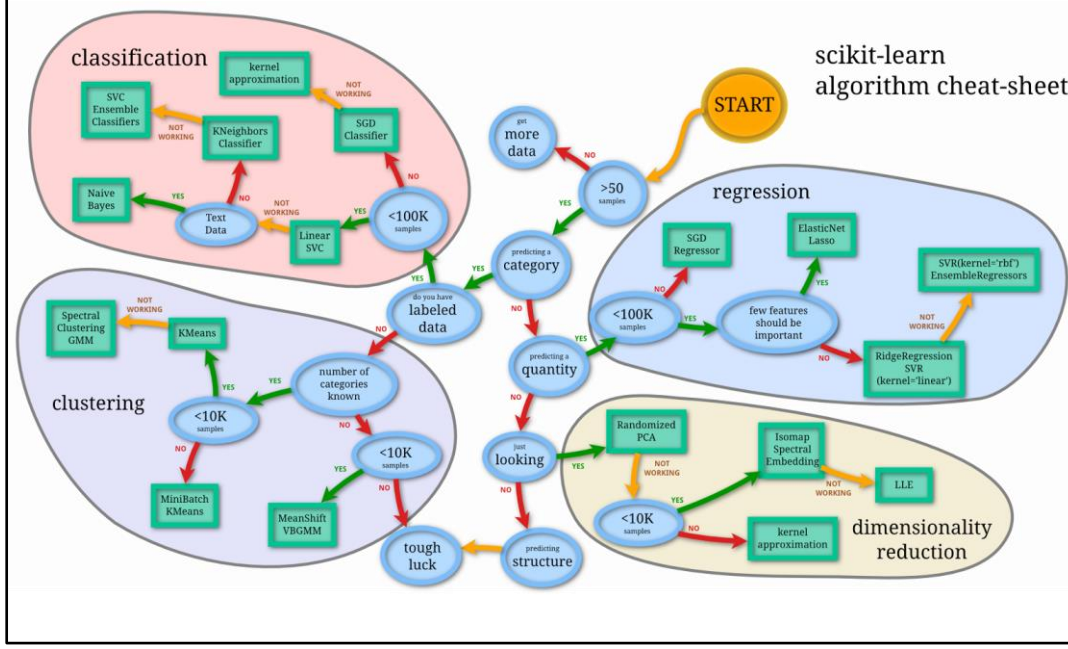
# Outlier detection / Anomaly detection (in 1D)




Outlier detection may also be needed when cleaning up the data. Rarely occurring event might be ones we do not need to consider when making predictions. There can be outliers, but there must be great many of them for the model to capture those special cases. However, the details of outlier detection is better to be discussed within the topic of data preprocessing.

Tampereen yliopisto

# Selecting a model




In popular ML libraries such as Python module scikit-learn, there are dozens and dozens of models to fit and algorithms to use. Here is an example of decision chart that may help you to select the right tool for the right job. You can also always ask me.

Tampereen yliopisto

## Or let computer do it?

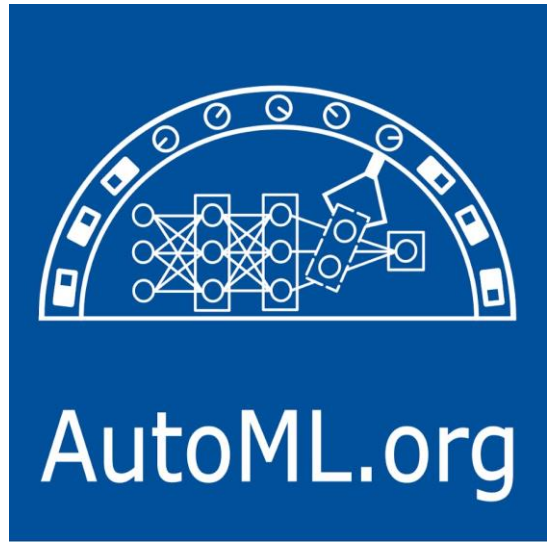
Parameter tuning

Automatic algorithm configuration

Metalearning

Algorithm selection

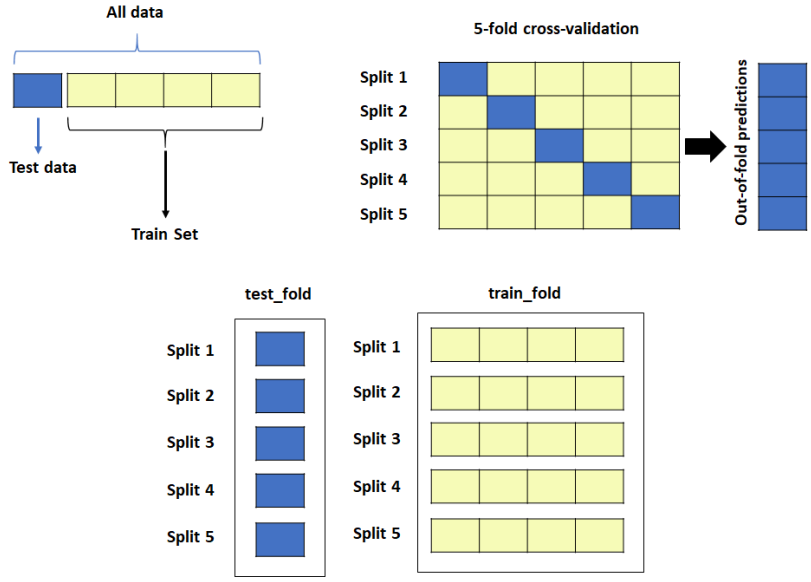
Hyperparameter search




There are also methods that can automate the search of a model, training algorithm and its parameters. In fact, the Internet giants (Microsoft, Amazon, Google) have their competing offering on AutoML. The goal is one-press solution for ML: “Just give us the data, we train the best model for it!”. Fortunately there are also many OSS solutions for those so inclined.

Tampereen yliopisto

# How to train and evaluate?




https://drzinph.com/nested-cross-validation-cross-validation-series-part-2a/

Finally, the biggest pitfall in applying ML is in using improper training and evaluation

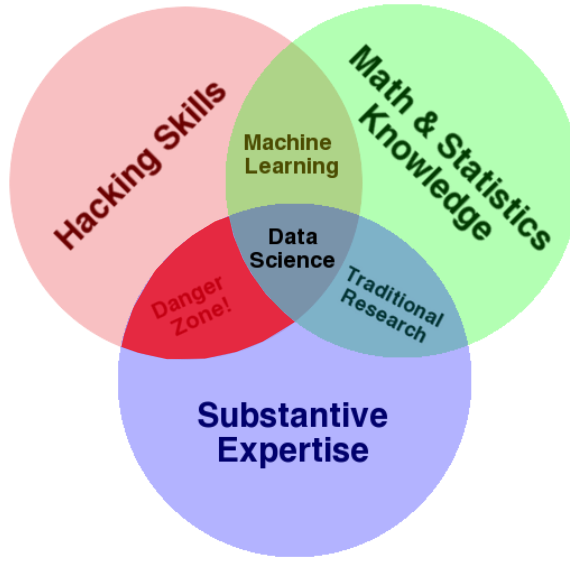
Tampereen yliopisto

## Baseline


**Sanity check:**
Ask yourself, is your model performing better than rolling the dice!

Tampereen yliopisto

## Warning for the laymen




*Source: http://drewconway.com*

Now, one should have some understanding what happens when machines learn. What are the possibilities, what are the pitfalls. How to build a model and how to test.

**AND MOST OF ALL: ensure you are solving the right problem.**

**Beware ways of misusing the machine:**

- **Interrogate data until it confesses**
- **Mixing correlation and causation**
- **GIGO :** garbage in, garbage out (including: biased data, duplicate data, irrelevant data, too little of it)

Tampereen yliopisto

# You are welcome for the information dump on hammers Now, YOU should SEE NAILS **EVERYWHERE**


## THANKS!

Jussi Rasku

etunimi . sukunimi (miukumauku)
tuni.fi
aistico.com

NEXT TIME (if you have me)
CONTEMPLATIONS ON DATA

I hope you feel that now you have many nails you will go and hit. Until next time (on data)