

Practical Tools and Best Practices of Machine Learning

(mostly ramblings about data)

9.2.2023 / Jussi Rasku

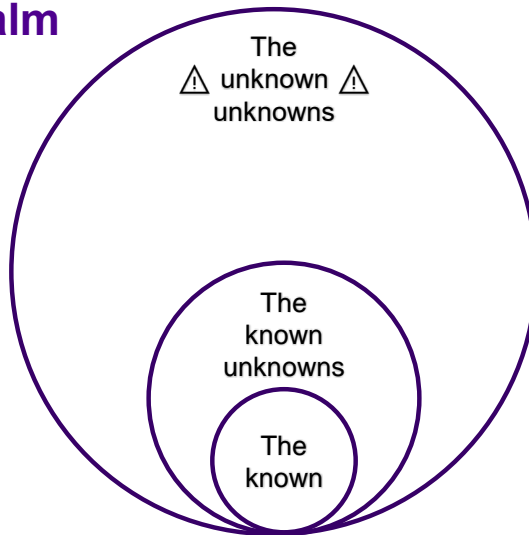


As previously, the image is generated by an AI. My last talk was little under a year ago and since then there has been further process in the field of AI, not limited to the AIs that can do illustrations.



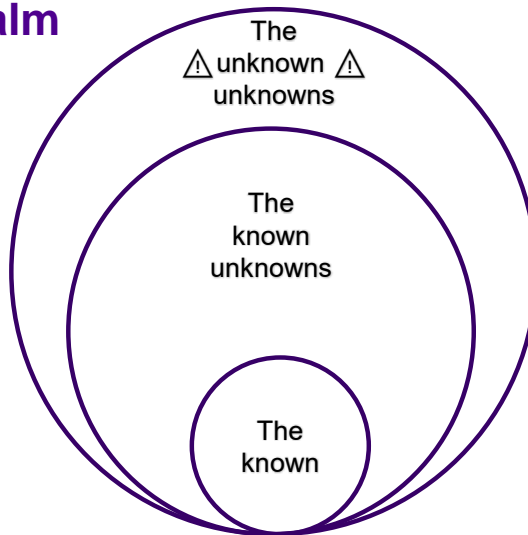
The data is the foundation the AI is built upon. Therefore, this talk will be mostly about data and how to form it into the shape it can be fed to the AI algorithms.

OUR AIM (in the realm of ML)









Unknown unknowns are the things that you do not know that you do not know. It is ignorance of the most dangerous kind.

OUR AIM (in the realm of ML)



Our aim is to push it back. I do not seek to grow "the known" much. If it happens, good, but the more important thing is to know what is out there to get you in ML.

		
<p>Software engineer jussi.rasku@aistico.com</p>	<p>Inventor & enthusiast jussi.rasku@sepeliry.fi</p>	<p>Operations researcher jussi.rasku@jyu.fi</p>
<p>Founder @ Aistico Oy</p>	<p>e.g., Secretary @ Game dev club of Seinäjoki</p>	<p>Postdoc @ Tampere University</p>
		

As one can see I wear many hats. I've been jokingly calling these:
Workwork,
Workhobby, and
Hobbywork

AI and I have a long history together and span all of these roles. Data analysis consultation and side projects have ranged from text, speech, machine vision analysis and game AI's and procedural content generation. AI also plays a major part in my research. Taken together, have learned one or two things on the topic.

Read more:
Statistical pattern classifier in machine vision quality control" (2010), MSc. thesis, TUT.
"Toward Automatic Customization of Vehicle Routing Systems" (2019) PhD disseration, University of Jyväskylä.


For those that missed
[the 2022-05-05 meetup](#)

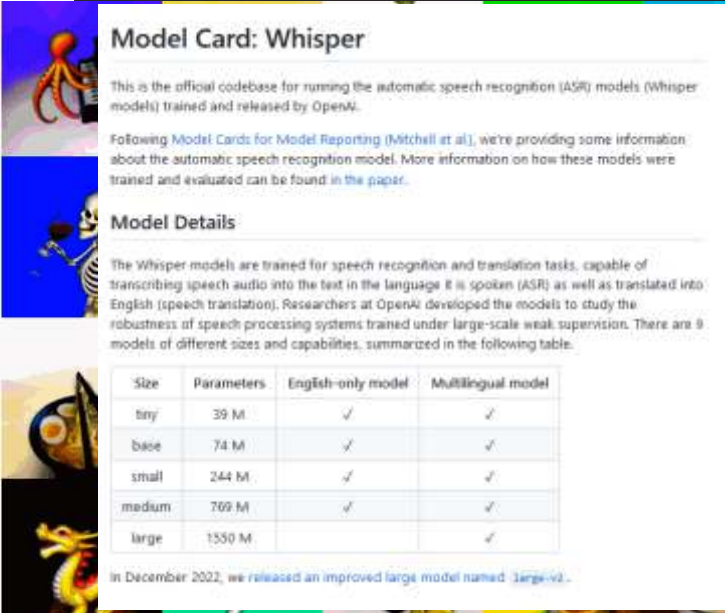
0) AI/ML RECAP




CC4 BY-NC

I promised there would be no evangelizing this time. However, I could not resist buy to give a brief update on what amazing things have happened since my previous presentation in May 22.







<https://lexica.art/prompt/2bf0be61-8301-4f2c-aa93-8346e9f05ec3>

The generative models have been becoming better. The state-of-the-art in image generation is pushed further and the frontier is now 3D and in video generation. Meanwhile, there has been major progress with other AI applications and models. For example, OpenAI has published MIT-licensed Whisper model that is almost as good as Google in speech-to-text. And one can run that model locally!

John Carmack's New Holy Grail: Artificial General Intelligence

John Carmack, the iconic Dallas game developer, rocket engineer, and VR pioneer, is setting his sights on AGI. If successful, his moonshot effort would be a 'change-the-world-level' event.

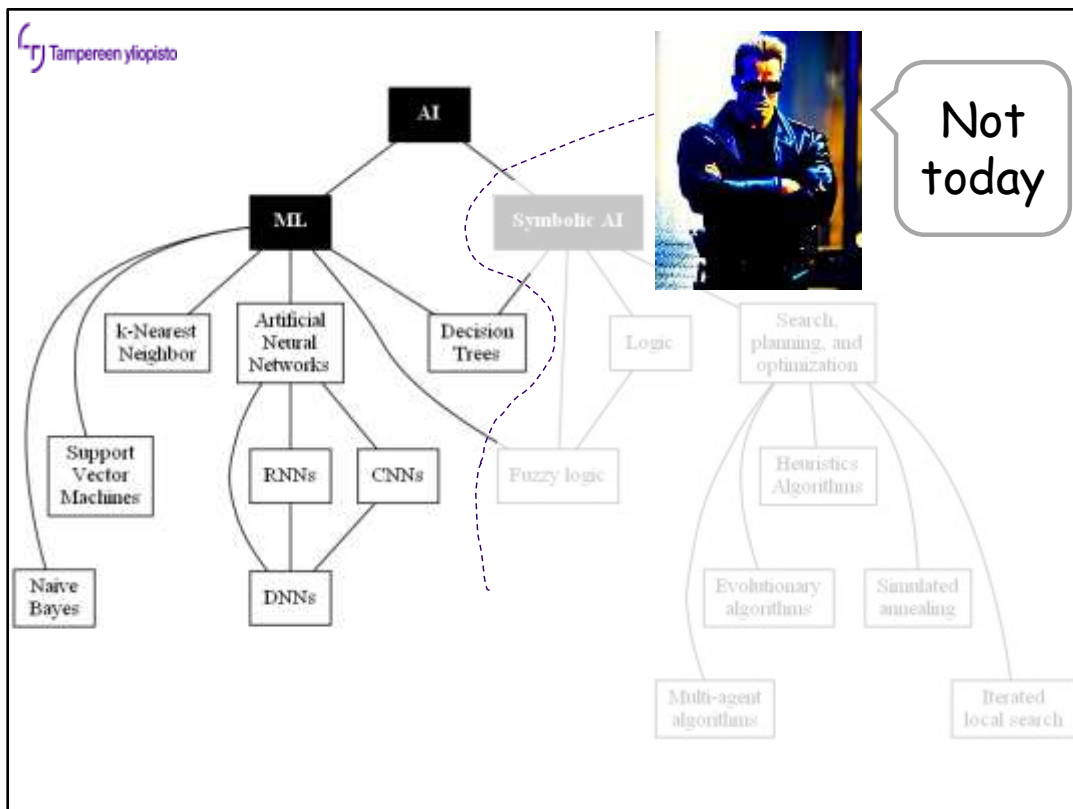


<https://dallasinnovates.com/carmacks-quest-artificial-general-intelligence/> 2.2.2023

And if you are not yet convinced, the messiah... I mean 100x engineer, I mean THE JOHN CARMACK has decided that after Oculus artificial intelligence is the topic he will spend the next 10 years with. John thinks this is THE NEXT THING, and I agree.



But, I diverge even if I promised to be concrete. So, back to the topic.



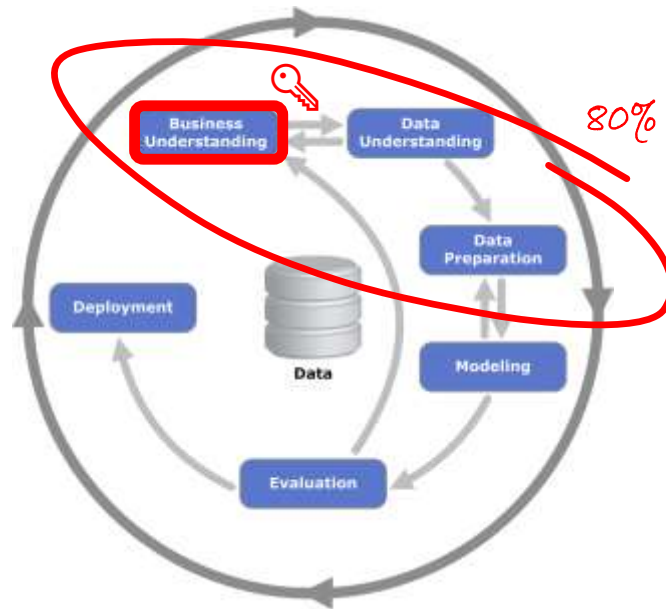
You might remember from the last talk that AI can be split in two: to Symbolic AI and Machine learning.

- Symbolic AI is “just” behavior/intelligence a human has encoded into the machine. Even adaptivity 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 recent “win” and success of ML over symbolic ML (and the [saltiness of symbolic AI proponents](#)).

However, there are some problems that are still difficult if using the “throw massive amounts of data+compute at it” approach. Many search, planning and optimization problems rely on heuristics, i.e., manually crafted algorithms. 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 – an approach I proposed in my PhD dissertation.

CRISP-DM



Lähde: Kenneth Jensen / CC BY-SA 3.0

This is a data mining process, but it still applies in modern ML. (Cross Industry Standard Process for Data Mining).

The steps can be recognized in most projects. Most of the time is spent ensuring we are building the right thing and getting the data right.

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, imputations, cleaning up.

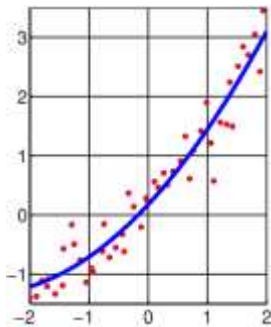
Modeling: Various modeling techniques are selected and applied. Typically, there are several techniques for the same data mining problem type. The preparation may need to be revisited.

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.

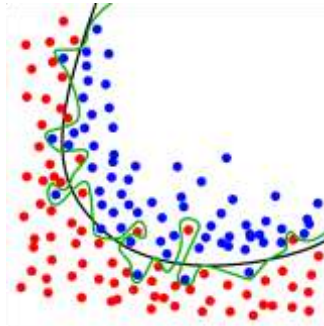
Deployment: The knowledge gained from the model will need to be organized and presented in a way that is useful to the customer/context. The actions which will need to be carried out in order to actually make use of the created models.

Reminder: ML approaches

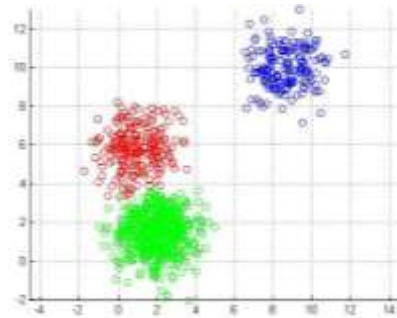
Regression



Classification



Clustering

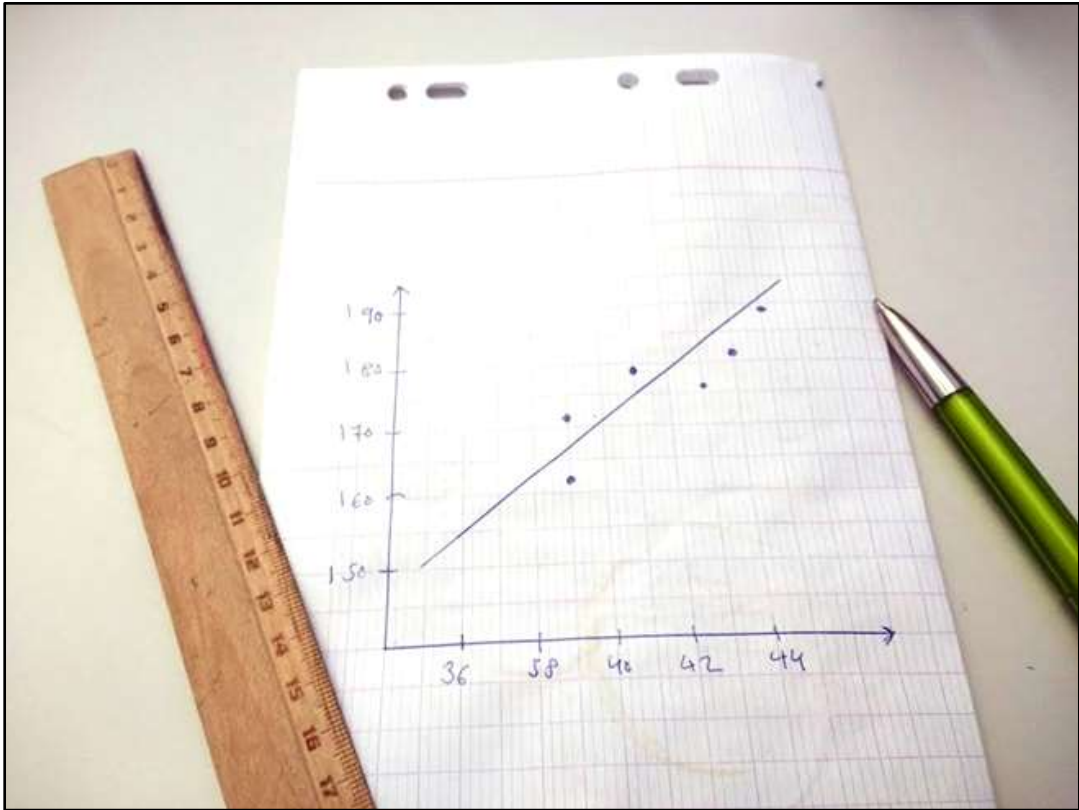


+ Association rule mining, anomaly detection, multidimensional scaling

Also, a reminder. There are three major categories in ML.

1. Regression, fitting a curve (i.e. plane in $>2D$) to predict an output for a new input sample (or to extrapolate)
2. Classification, to fit a curve (i.e. plane in $>2D$) "decision boundary" to divide the space into categories. Hence, we can assign a label/category for a new input sample.
3. Clustering, where using only the data we can automatically identify the different "kinds" of samples in our data. In the example these are different flower plant species. The algorithm has no understanding of this but was still able to recognize them.

There are also some other ML tasks (such as rule mining, MDS etc., but let's not discuss those today).



Remember: 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.

Can I has ur data, plz?

1) CONTEMPLATIONS ON THE DATA

High quality data is the most important thing to consider when planning to build and apply ML models.

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

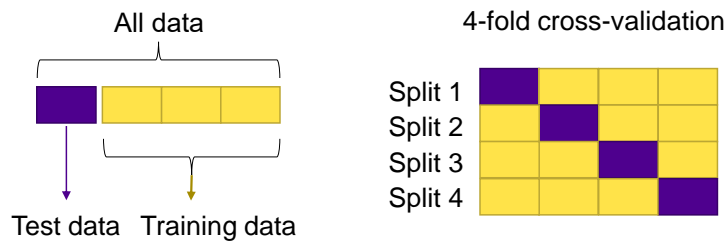
Presenting the data in a format that it reveals its secrets. In a format one can understand and extract KNOWLEDGE from.

As much art as a science.

Ask the right questions, find the right goal and make a plan how to get the data and from where. Then we must ensure it is of high enough quality and fit for purpose.

In the previous talks of the series, we have already discussed choosing the model and little bit of the training.

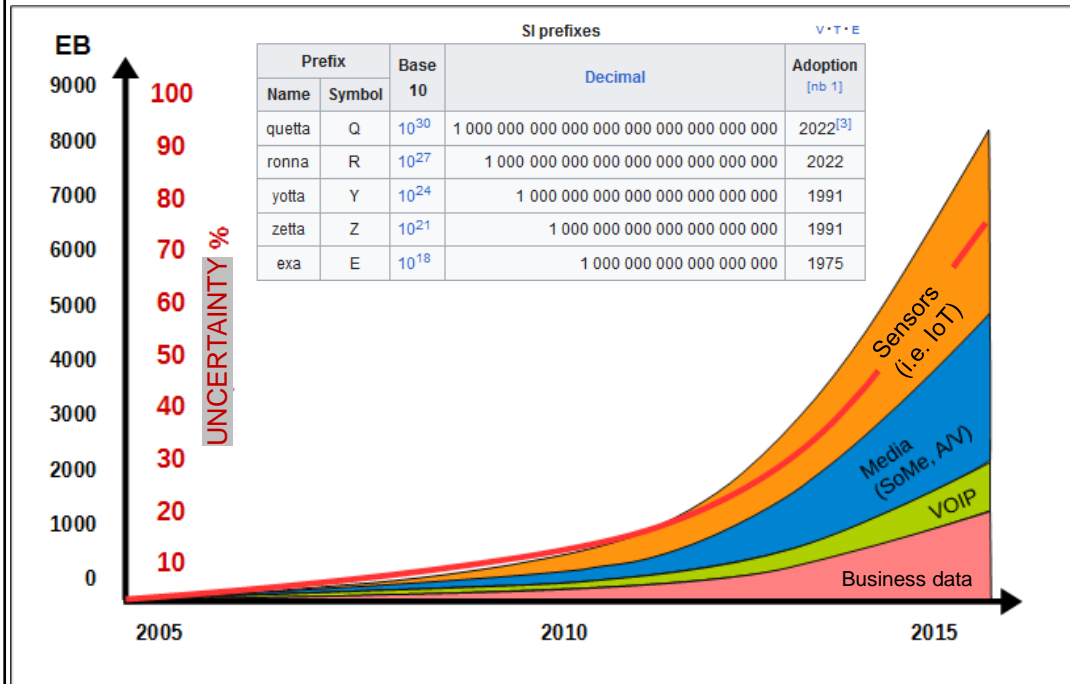
Eyes on the ball: we will use the data to train and evaluate?



Largest pitfall in applying ML is in using improper training and evaluation. If you have plenty of data, do a test / training / validation split. If you have less, do cross validation.

For smallest datasets one can do leave-one-out cross-validation. This means you will train as many models as you have datapoints. Each of the models is then evaluated on that withhold datapoint to calculate the expected performance of the model.

What is this data?



IoT and logs can grow quite big, quite fast.

1 EB = 1 Exabyte, the storage capacity of around 10 million workstations.

We are already living the zettabyte (1000 EBs) era since mid-2010s and the humanity is closing fast to the era of yottabyte. In fact, we needed new SI prefixes to anticipate this. ronna- and quetta- have been proposed last year in 2022.

Where to get some data?

Readily available

- Academic datasets
- Open data
- Data from stakeholders
 - Data lakes
 - Databases
 - Logs
 - Excel sheets
 - Files and ... gulp... papers

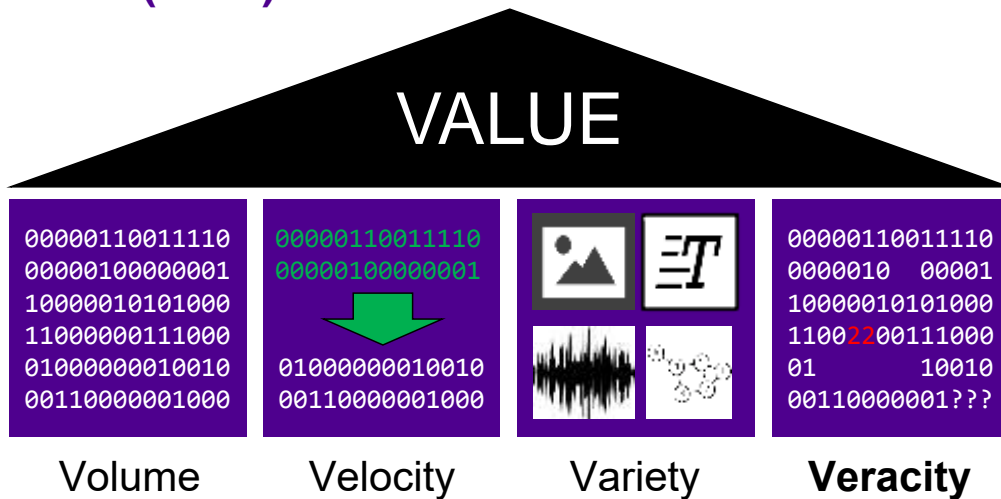


Collected

- Directed data collection
- Mturk etc.
- Questionnaires
- Automated extraction (e.g. NLP)

A lot of the data is in a format that cannot readily be processed. **Most of the data is unstructured. Utilizing it without machine learning or human effort is difficult.** Examples are text, images, video etc. A major investment must be made to make it usable.

The four v's 4 x V (+ 1 V)

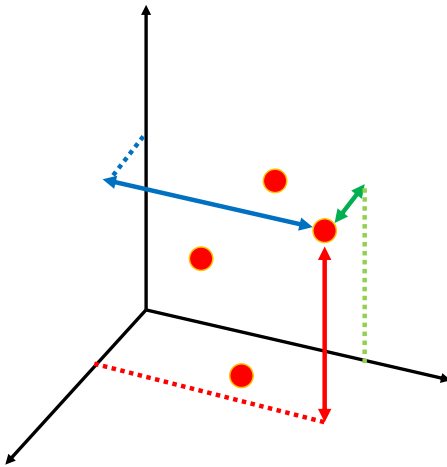


Velocity = pace at which data flows in

Veracity = biases, noise and abnormality in data (todenmukaisuus)

Let's proceed to discuss some tools to manage these four V's

DATA VARIABLE TYPES



1.2	M	9+
1.3	F	7-
+	+	+
Numeric	Categorical	Ordinal

The data is composed of a set of measurements. The variables are the dimensions of these measurements.

Variables can be

- **Categorical**, allowing only certain values.
- **Ordinal**, where variable values can be ordered.
- **Numerical**, where we can do some arithmetic and calculate distances (e.g. $6/2 = 3$)
- (sometimes non-structured, pictures, text, sound etc.)

Measurement gives the variable its value. Values together create a **data point**

*The **curse** of dimensionality*

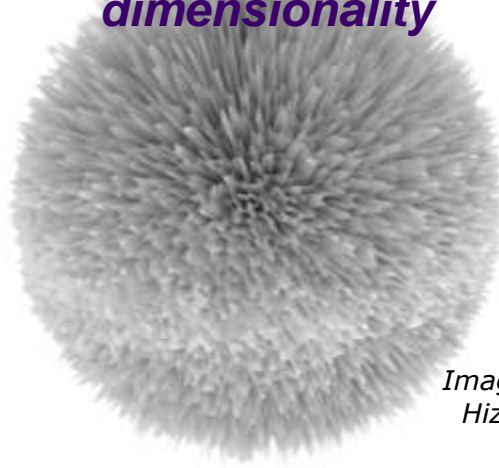


Image by
HizkiFW
(CC)

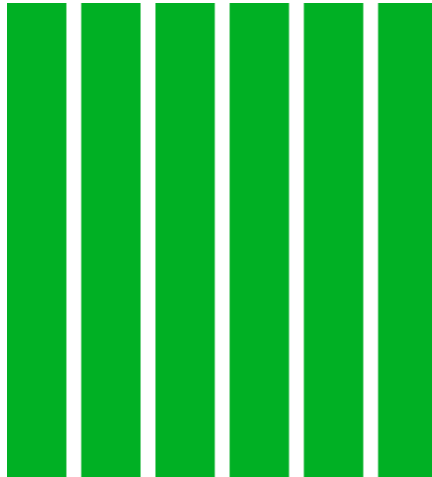
Too many variables?

No worries!

In a high dimensional space, every data point is close to the surface.

Here, the curve is not a line and instead of 2-dimensions the space can have, for example, 100 dimensions. Intuition breaks down. What is the model that you are even fitting? We humans are incapable of imaging such spaces. Still, one should remember that in the end this is still the same task you were doing curve fitting in the physics class. Only several magnitudes harder. Hence, better leave it to the computers.

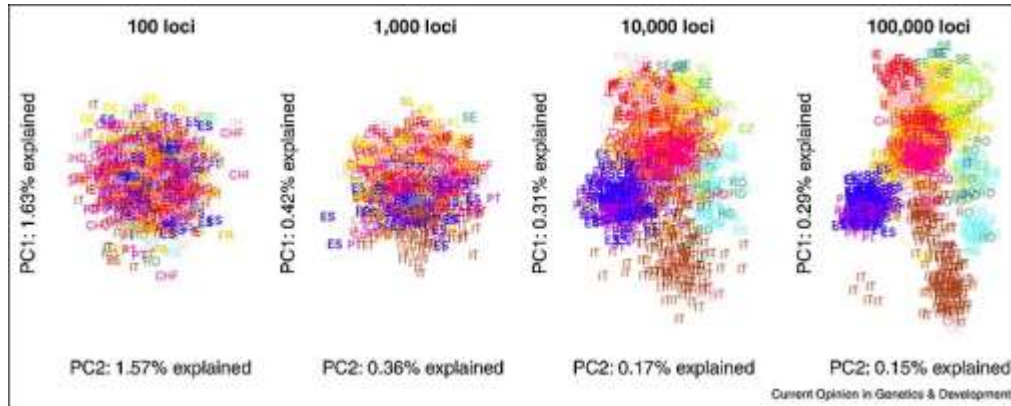
Too many variables? USE DIMENSION REDUCTION



→ Reduce the number of DATA COLUMNS, easy

The problem is that we try to keep **the right** columns that are relevant to our business problem.

PCA



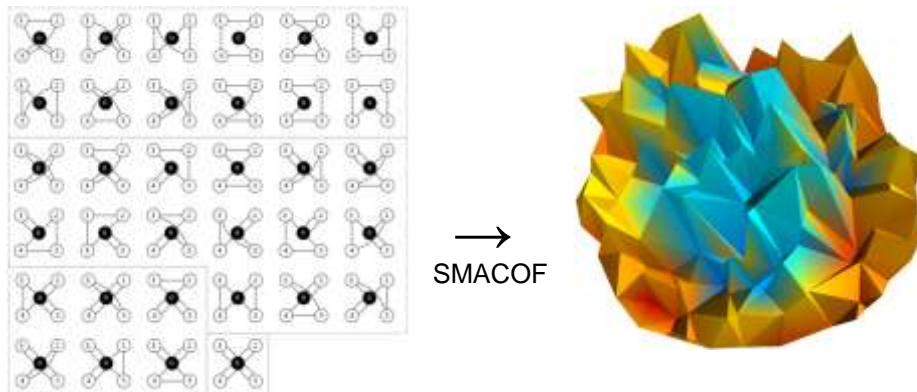
Unfortunately, the new variables PC1...PCN might not carry any sensible meaning anymore.

John Novembre, Benjamin M Peter (2016) Recent advances in the study of fine-scale population structure in humans. Current Opinion in Genetics & Development. Volume 41, 2016.

Genetic data is multidimensional. However, if we consider 100k gene locations (100k dimensions!) a pattern starts to merge. This pattern can be revealed with PCA and it matches the geography. Still, the coordinates PC1 & PC2 are **not** longitude and latitude. They just represent something similar about the dissimilarity of the loci.

Example of dimension reduction and multidimensional scaling (MDS)

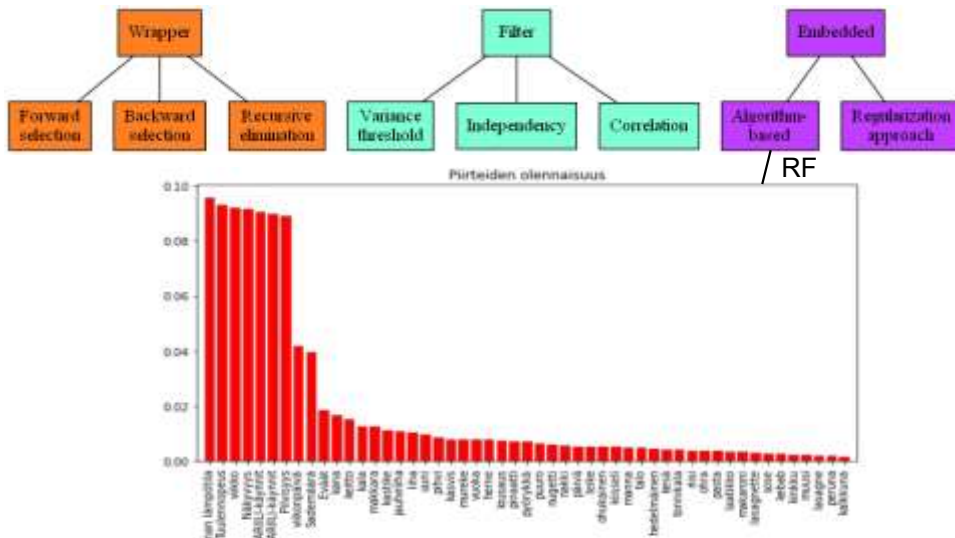
$$\{0,1\}^{206} \rightarrow \mathbb{R}^2$$



As discussed earlier, sometimes the dimensionality is too high and needs to be reduced.

As an example, here is a visualization technique from my earlier research representing 206-dimensional vehicle routing problems as a 3D surface. The algorithm is SMACOF stress majorization algorithm.

Still too many variables? USE FEATURE SELECTION



Feature selection is another way to get rid of useless, irrelevant measurements. It, e.g., helps in cases where two features have almost the same information content. Feature selection helps in the fight against the curse of dimensionality.

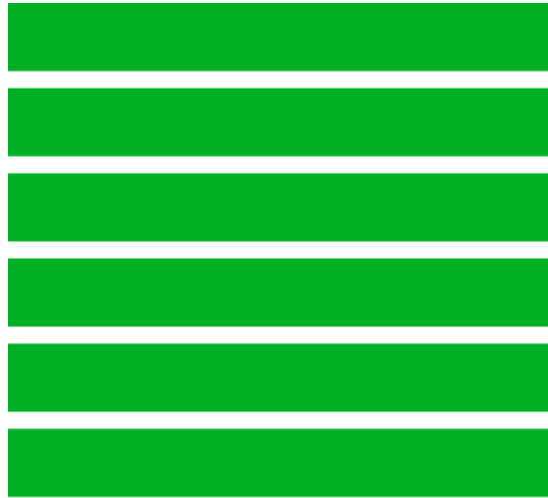
Consequently, with less features the training of models can be faster, models become simpler and easier to interpret and fine-tune.

Too much data?

No worries!

We discussed the issues data with too high dimensionality. What if you just have too much datapoints (i.e, rows)? This is so easy that it is almost non-issue.

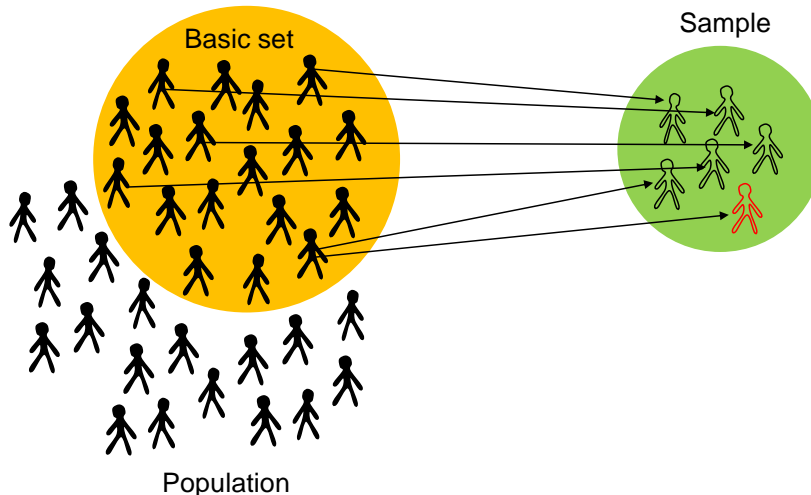
Sampling



→ Reduce the number of DATA ROWS

We try to select the rows in a way that the set remains representative (no cherry picking!)

Sampling



Randomly select samples from the base population. Most often it is done without returning the datapoint to the population, so each element can only be selected once.

If the sample is representative (*“edustava”* in Finnish), the conclusions drawn are valid for the entire population. It is usual to do stratified sampling (fin *“ositettu otanta”*) or clustered sampling (fin *“ryväsotanta”*) to ensure this.

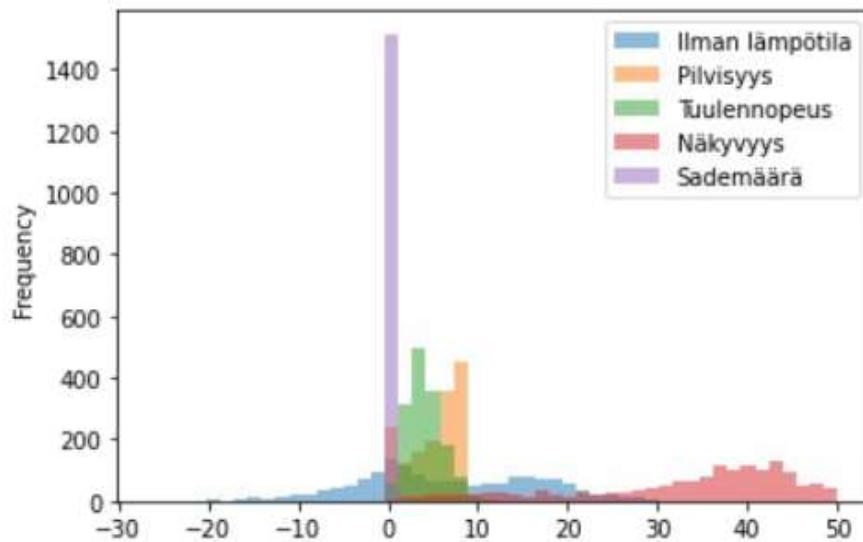
Be wary of accidental sampling (the basic set we sample is not representative of the entire population) and snowball sampling (the basic set is built by branching out using connections of members of the basic set).

My data is broken?

Hmm..
start to worry

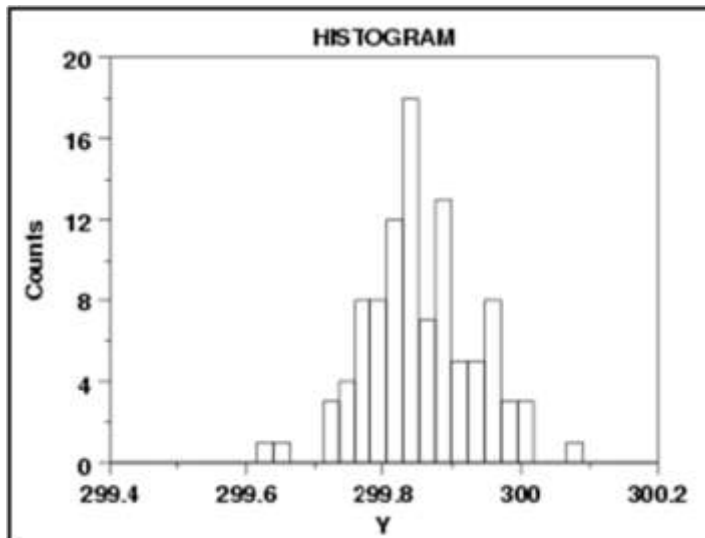
GiGo : Garbage in, garbage out can kill otherwise promising ML project.

How would I know?



Answer: look really hard. However, sometimes suspicious distributions can have a natural explanation. When you **do not** have an explanation, then you might have bad data.

Get a "FEEL" for the variable



Source: <https://www.itl.nist.gov>

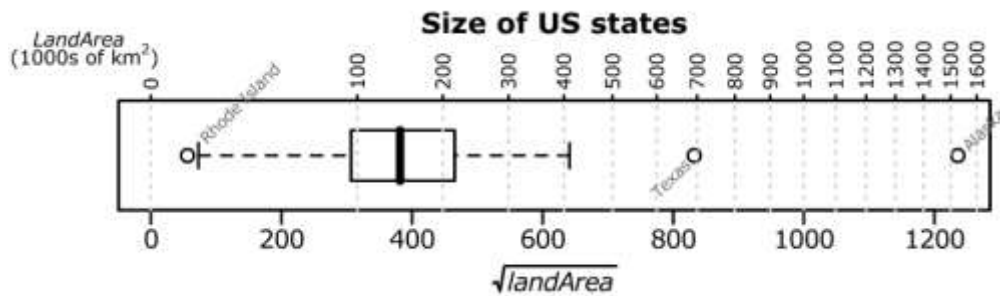
One can learn a great lot from histograms.

Q: How would two persons doing the testing (or different measurement methods) appear in the histogram? How about systematic or drifting error? Saturation of the measuring instrument?

A: Statistical moments, mean, stdev, skew, kurtosis.

Noise detection and considerations related to that are outside of scope of this talk.

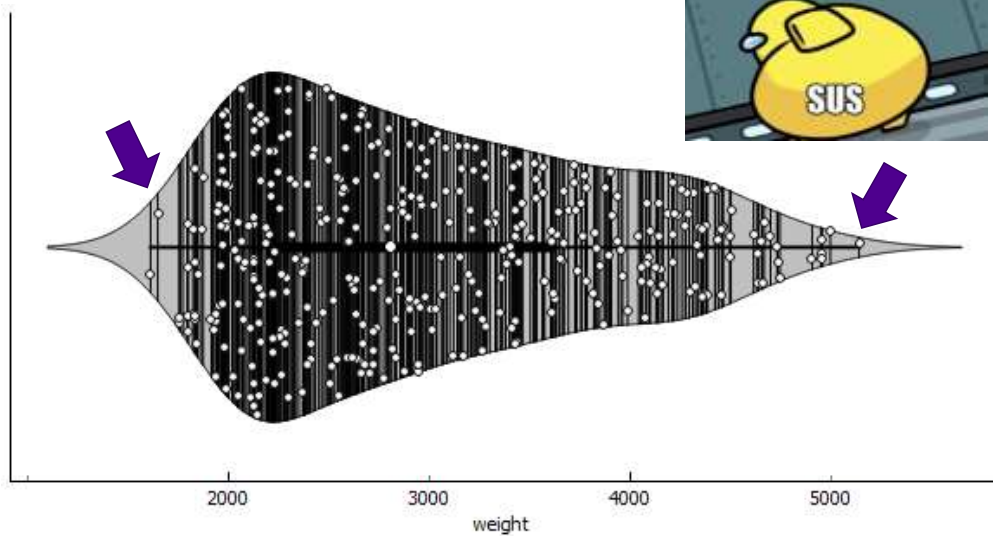
Another way to know: outlier detection



Boxplot can show the outliers and removing those may improve the **Veracity of the data**.

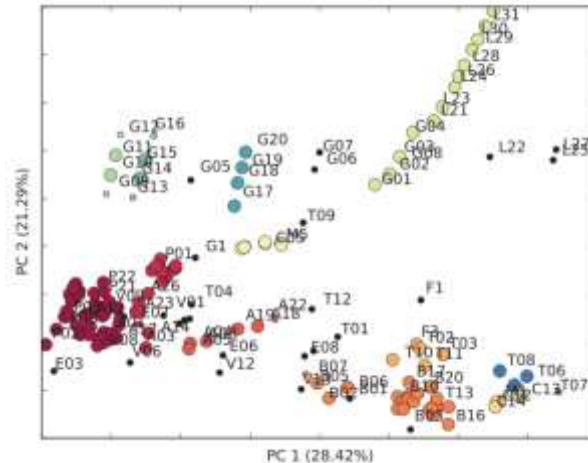
Outlier detection may also be needed when cleaning up the data. Rarely occurring event might be ones we do not need to consider when making predictions. There can be outliers, but there must be great many of them for the model to capture those special cases. However, the details of outlier detection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

**Sometimes you just know:
Just look at it!**



There is also violin plot. Use your predator visual cortex and visualizations to make the "sus" individuals stand out.

Outlier detection / Anomaly detection in higher dimensions



For an interactive version, see
<http://users.jyu.fi/~juherask/features/>

Some clustering methods such as DBSCAN can reveal outliers in high-dimensional data. This is another example from my research.

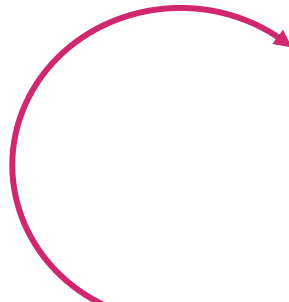
MISSING VALUES & IMPUTATION

Plan A) Through removal

1.2	M	9+
-	F	7-
8.4	F	8
...

If you have plenty of data, just remove the rows that do not have all the measurements.

MISSING VALUES & IMPUTATION Plan B) With average



1.2	M	9+
-	F	7-
8.4	F	8
...

=AVERAGE(A1:A100) = 1.4

However, you can substitute the missing value with the column average. Now you are doing imputation.

IMPUTATION, Plan C) Nearest neighbor

Most
similar

1.2	M	9+
-	F	7-
8.4	F	8
...
1.3	F	7
...

Find the most similar and use the value of that row when imputing.

The metric, i.e., how the nearest neighbor is recognized is very important here. For example, how relevant is gender in the above example? If interpreted as M=1, F=0, it affects as much as 100 cm in the height of the student! This is plainly incorrect. Devise a good metric with the domain expert when defining similarity between rows.

IMPUTATION, Plan D) Machine Learning

Predicted

1.2	M	9+
-	F	7-
8.4	F	8
...
1.3	F	7
...

One can also train another machine learning model to do imputation. However, beware, there may be madness here due to runaway bias.



Communicate with clarity,
precision and efficiency

2) VISUALIZATION OF THE DATA

Data visualization is the graphical representation of information and data. It is a powerful tool to communicate trends, outliers, and patterns present in the data using charts, graphs, and maps.

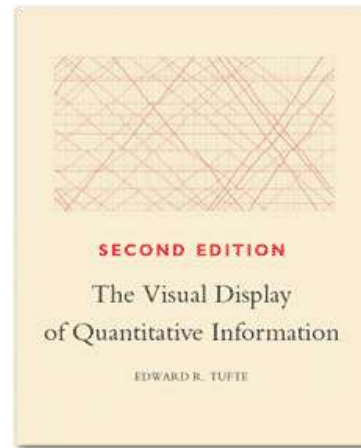
Humans are very visual and can take in a large amount of data with a single glance. Also, for any data set with more than 100 or so measurements we must resort to a) statistics b) visualization to gain any understanding of it. Visualization is a key tool to make sense of the big data we generate every day.

It is a very important topic, but in this talk we move on quite quickly. After all, good books have been written of the topic... (next slide)

Charts

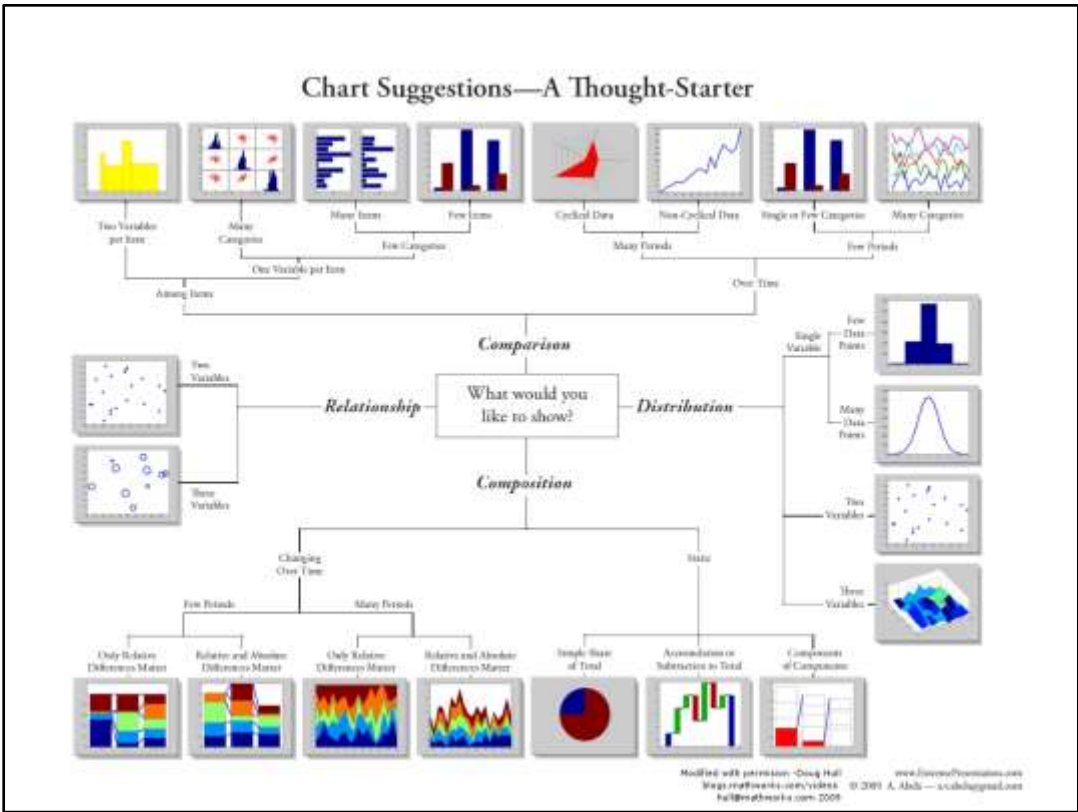
- Statistical moments (mean, median, MAD, SD, var, etc.)
- Histogram (distribution), boxplot
- Relations (graph repr. etc.)
- Point distribution (scatterplot)
- Radial projections
- Dimension reduction
- *Etc.*

SeAMK
KIRJASTO | LIBRARY



Tufte: *"maximize the data/ink ratio"*
"clarify by adding data"

Tufte has written classical on visual display of data. My MSc. supervisor required that everyone of his students reads this book before submitting their thesis. Do not tell Robert that I did not (however, I **did** read it during my PhD, perhaps I'm safe?).



There are also decision trees that may help you to choose the correct visualization. Study these and try to understand their rationale.



seaborn



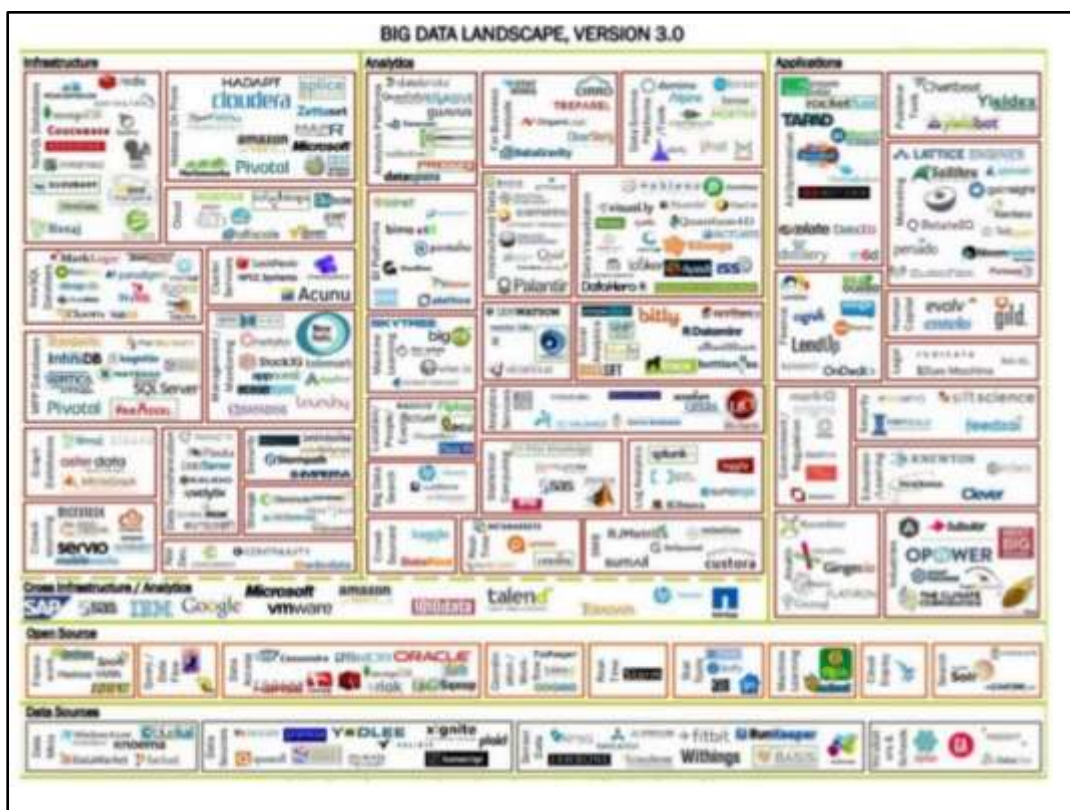
I've used all of these save D3.js. Mastering the use of a good plotting tool is highly recommend.

Some tools and examples

3) TOOLS

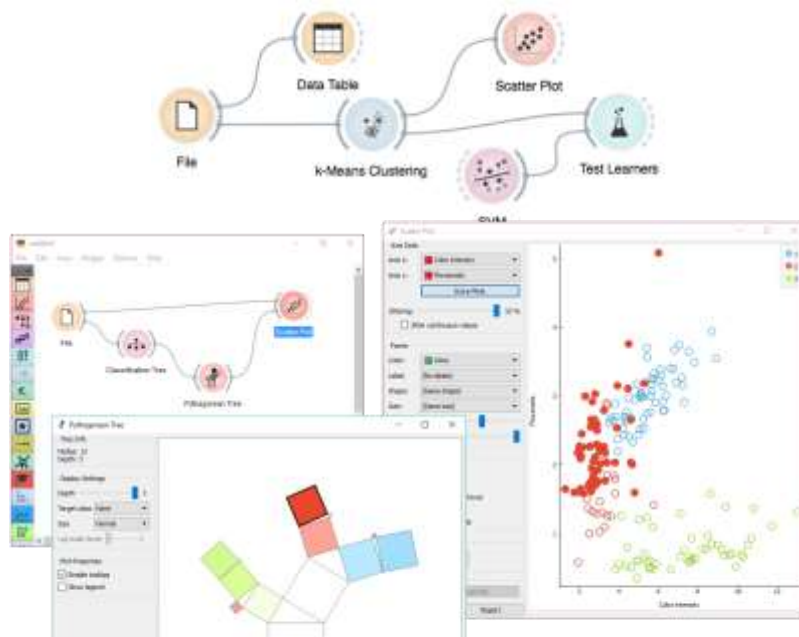


Finally, some recommendations of tools and techniques to crunch and preprocess that data.



The space of tooling for big data and ML is growing ridiculously large. I'd recommend choosing a boring and tried and true solution.

TOOL : ORANGE3

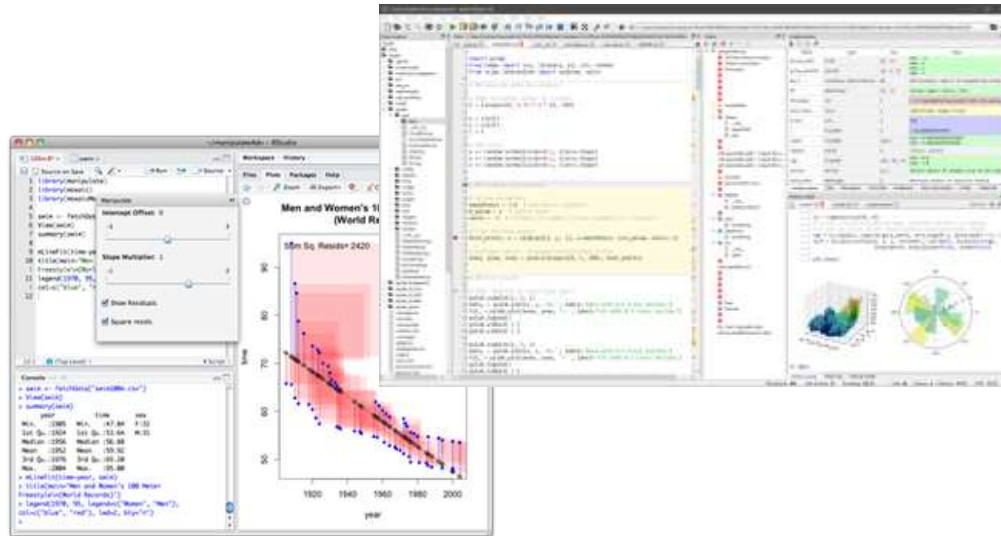


PROBLEM: You want to mine material that can fit on a CD.

SOLUTION: Graphical and interactive open-source no-code datamining environment.

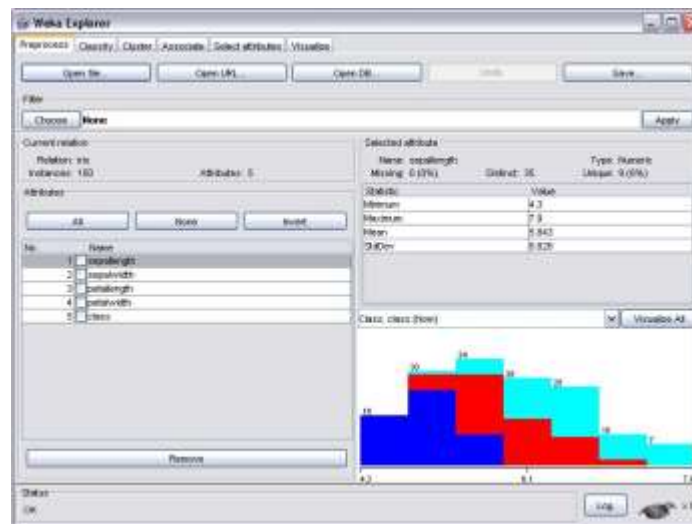
TOOL: <http://orange.biolab.si/>

IDEs : R studio & Spyder



Proper IDEs but tailored for data engineering and data science.

TOOL : WEKA



<http://www.cs.waikato.ac.nz/ml/>

PROBLEM: You need more methods to make method comparisons and for a larger amount of data.

SOLUTION: A versatile toolkit for machine learning and data mining.

TOOL: Weka has implementations of almost all classical ML algorithms.

TOOL: UNIX COMMAND LINE

```
tomi@data36-learn-server: ~/practice (ssh)
tomi@data36-learn-server:~/practice$ head flightdelays.csv | sed 's/SMF/DATA36_TUTORIALS/g'
Year,Month,DayofMonth,DayOfWeek,DepTime,CRSDepTime,ArrTime,CRSArrTime,UniqueCarrier,FlightNum,TailNum,ActualElapsedTime,CRSElapsedTime,AirTime,ArrDelay,DepDelay,Origin,Dest,Distance,TaxiIn,TaxiOut,Canceled,CancellationCode,Diverted,CarrierDelay,WeatherDelay,NASDelay,SecurityDelay,LateAircraftDelay
2007,1,1,1,1232,1225,1341,1340,WN,2891,N351,69,75,54,1,7,DATA36_TUTORIALS,ONT,389,4,11,0,,0,0,0,0,0,0
2007,1,1,1,1918,1905,2043,2035,WN,462,N370,85,90,74,8,13,DATA36_TUTORIALS,PDX,479,5,6,0,,0,0,0,0,0,0
2007,1,1,1,2206,2130,2334,2300,WN,1229,N685,88,90,73,34,36,DATA36_TUTORIALS,PDX,479,6,0,0,,0,3,0,0,0,31
2007,1,1,1,1230,1200,1356,1330,WN,1355,N364,86,90,75,26,30,DATA36_TUTORIALS,PDX,479,3,8,0,,0,25,0,0,0,3
2007,1,1,1,831,830,957,1000,WN,2278,N480,86,90,74,-3,1,DATA36_TUTORIALS,PDX,479,3,9,0,,0,0,0,0,0,0
2007,1,1,1,1430,1420,1553,1550,WN,2386,N611SW,83,90,74,3,10,DATA36_TUTORIALS,PDX,479,2,7,0,,0,0,0,0,0,0
2007,1,1,1,1936,1840,2217,2130,WN,409,N482,101,110,89,47,56,DATA36_TUTORIALS,PHX,647,5,7,0,,0,46,0,0,0,1
2007,1,1,1,944,935,1223,1225,WN,1131,N749SW,99,110,86,-2,9,DATA36_TUTORIALS,PHX,647,4,0,0,,0,0,0,0,0,0
2007,1,1,1,1537,1450,1819,1735,WN,1212,N451,102,105,90,44,47,DATA36_TUTORIALS,PHX,647,5,7,0,,0,20,0,0,0,24
tomi@data36-learn-server:~/practice$
```

PROBLEM: But I really have a lot of data. Excel won't open it anymore.

SOLUTION: If you know what information you are looking for, a Unix-style command line is excellent fit for ad-hoc style data processing.

TOOLS: Cat, grep, head, tail, uniq, awk, sed, wc, gnuplot, etc.

<http://www.gregreda.com/2013/07/15/unix-commands-for-data-science/>

<https://comsysto.wordpress.com/2013/04/25/data-analysis-with-the-unix-shell/>

<http://datavuu.blogspot.fi/2014/08/useful-unix-commands-for-exploring-data.html>

TOOL: (Deep) neural networks

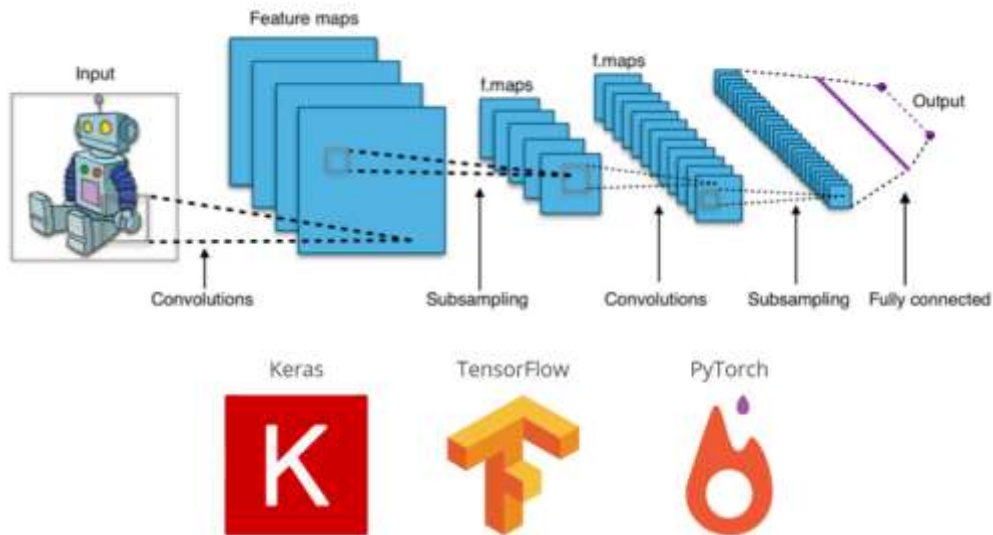
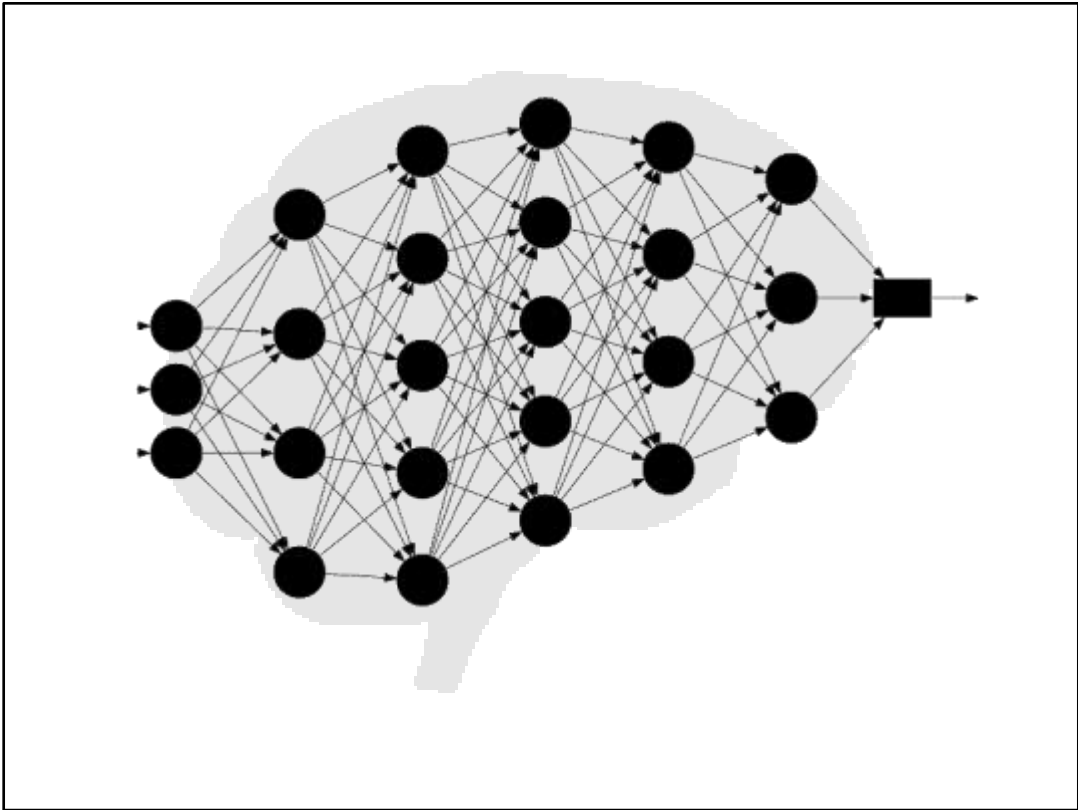


Image: Aphex34, CC BY-SA 4.0 via Wikimedia Commons

ANNs allow you to use non-structured data. However, they are quite difficult to use as one needs to come up with a suitable neural network architecture. Finding suitable activation functions, layers, normalization, regularization is difficult – even for someone with significant expertise. In my experience, this is no silver bullet.



What makes neural networks interesting, is that they draw inspiration from the actual neurons inside our heads. Also, the research done during the past decade or so has shown that when combined with large amount of data they are really powerful and can solve really difficult perception and generation problems.

Cloud ML, i.e., ”data science on someone else’s computer”

- Microsoft Azure Analytics
<https://azure.microsoft.com/en-us/product-categories/analytics/>



- Amazon AWS Analytics
<https://aws.amazon.com/machine-learning/>



- Google GCP Cloud Analytics
<https://cloud.google.com/solutions/smart-analytics>



IS your data Big data?

- Data volume (MB, GB, TB?)
- Query volume (1k, 1M, ?)
- Query response speed?
- Velocity
- Which data processing methods?
- Structured / unstructured data?
- Computational infrastructure?
- Fault tolerance?



Read more: <https://motherduck.com/blog/big-data-is-dead/>
e.g, *"Most people don't have that much data"*

-> IS your data Big data? Often, a smaller, simpler tool will do. Do not emulate the big (Google, Amazon, MS) if your data is small. That only accrues complexity to your technological foundations.

It is not just the size. Also high velocity may mean your data is big. However (next slide)

BIG GUNS of the "Big Data Ecosystem"

- Computation
 - Map & Reduce
 - Hadoop
 - Spark ym.
- Big storage
 - Hbase (noSQL)
 - HDFS / Hive
 - Cloudstore
 - Titan ym.
- Collection
 - Kafka, Stormy.
- Orchestration
 - Kubernetes
 - Airflow

These are only needed
if the scale is "Google".
Needs Deep Expertise.
YAGNI?

Only go there if a) traditional methods can no longer be used b) there is REALLY a lot of data c) it doesn't fit in the database, disk, or program d) the data is really diverse - including unstructured data.

However, if you need data duplication, you have to use calculation-intensive methods, you might need to borrow something from the

"Simple algorithms and lots of data trump complex models" - Halevy, Norvig and Pereira (Google)

Some Assembly Expertise required



Sadly, building AI based businesses is not easy.

Some Assembly Expertise required

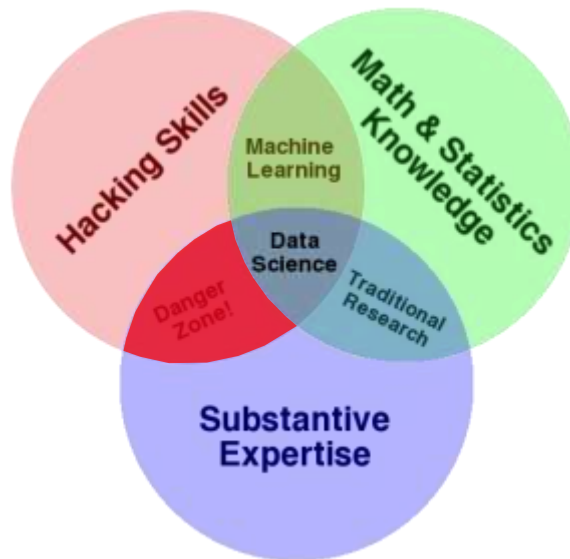
AI business	VS.	Software business
<ul style="list-style-type: none"> • Choose the right tools (models, preprocessing, ...) 		<ul style="list-style-type: none"> • Choose the right stack (Web, Desktop, Cloud ...)
<ul style="list-style-type: none"> • Define the problem and its constraints. • Gather a huge amount of data • Use data to understand the problem better 		<ul style="list-style-type: none"> • Specify the business problem and its constraints. • Build a prototype or mockups. • Use the prototype and requirements engineering to understand the problem.
<ul style="list-style-type: none"> • Capture the rules by iteratively training the ML model 		<ul style="list-style-type: none"> • Capture the rules by iteratively developing a software artifact
<ul style="list-style-type: none"> • Measure, cross-validate, deploy 		<ul style="list-style-type: none"> • Measure, test, deploy

Hence, Steve was right.

(Steve who? Although it is hard to believe, that sweaty guy is the previous CEO of MS)

The lesson here is that if you are dreaming of an AI powered business. Find a good developer and Hold on for dear life. The success of your business depends on your ability to execute the vision.

To conclude...



Source: <http://drewconway.com>

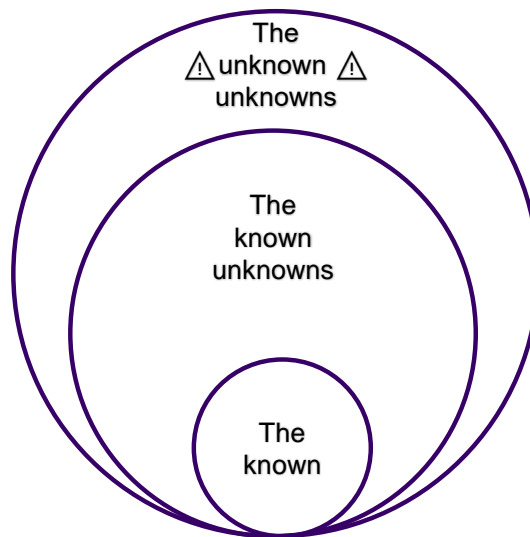
One should understand what happens when machines learn. What are the possibilities, what are the pitfalls. How to build a model and how to test.

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

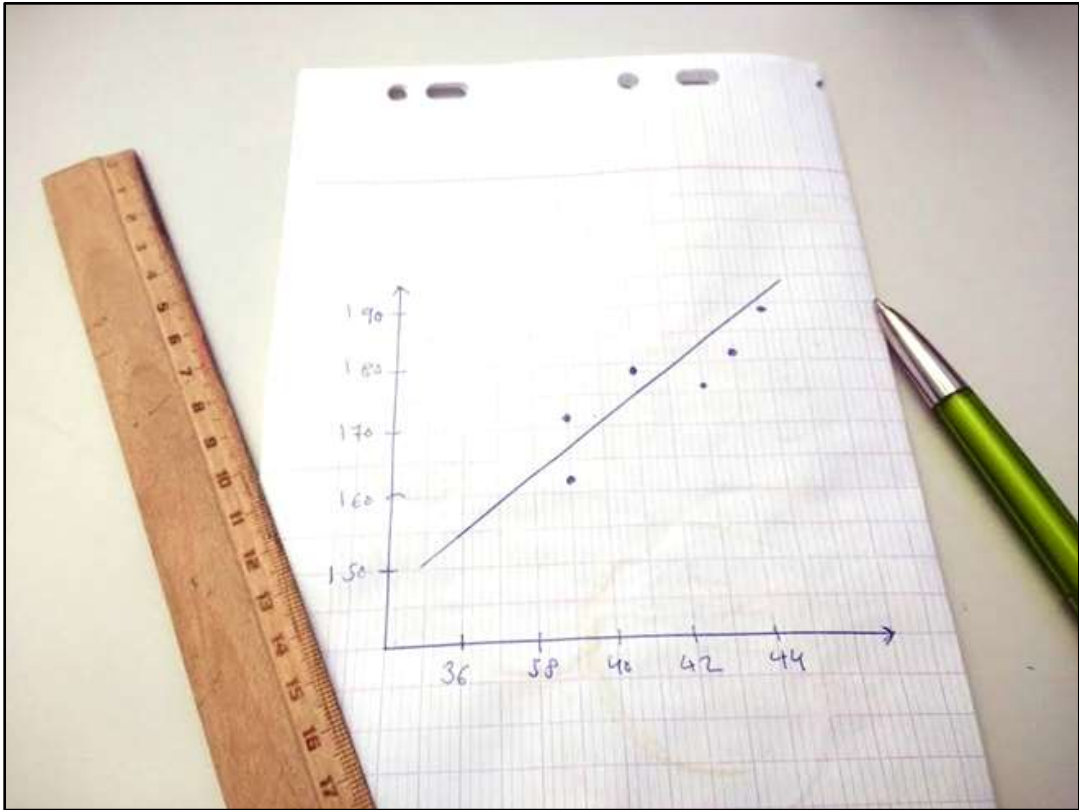
Some examples of misusing the machine

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

OUR AIM



I hope I was able to shrink the space of your unknown unknowns.

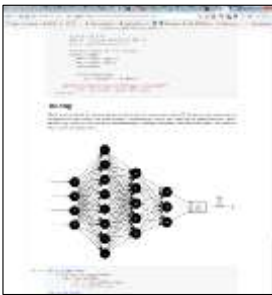


And, finally, if there would be just one thing you bring with you from this talk, it is this:
At its core, AI is nothing more than curve fitting. Thank you.

DEMO(S)

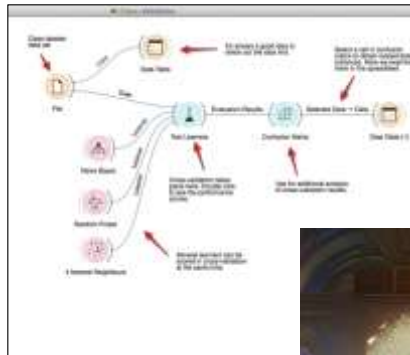
IPython Notebook

Artificial neural networks and differential evolution in Python programming language



Orange Data Miner

Data manipulation, visualization and ML model building.



Stable Diffusion

Local image generation with free and open generative AI model.



(if we have the time)