Big Data and Artificial Intelligence

M. Granitzer, Lehrstuhl für Data Science, Universität Passau

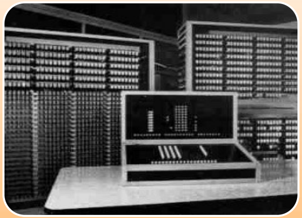
FIT Europe Seminar, 16.02.2021

Co-funded by the
Erasmus+ Programme
of the European Union

How to “teach” machines to do what we want?

Artificial Intelligence

- Information / Numbers as more abstract objects
- Mathematical rules (discrete)
- We describe how machines can change
- Environment and data as language



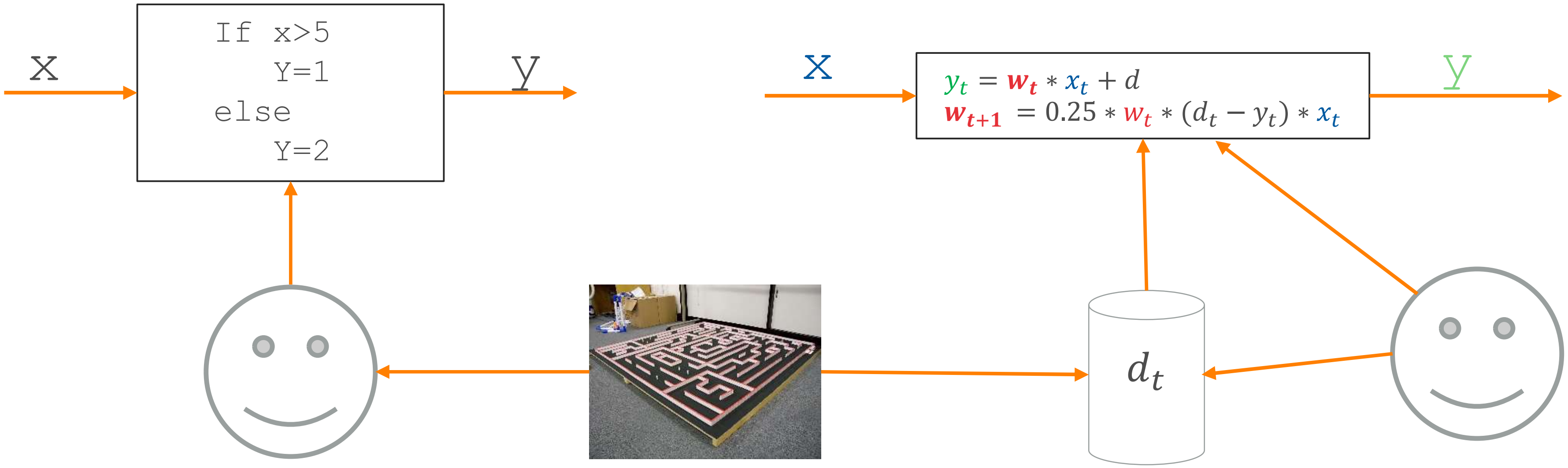
Digital Machines

- Information / Numbers as more abstract objects
- Mathematical rules (discrete)
- Precise description of the behaviour
- Language the same for all solvealbe problems

Physical Machines

- Concrete Objects
- Physical Law as Language
- Precise description of the behaviour.
- Language differs per Machine

Programming vs. AI



```
If x>5
   Y=1
else
   Y=2
```

$$y_t = w_t * x_t + d$$
$$w_{t+1} = 0.25 * w_t * (d_t - y_t) * x_t$$

d_t

This Photo by Unknown Author is licensed under CC BY-SA

Questions:

- Can we learn every possible programme?
- Are there different “AI languages”?
- Properties of data /environments to be used?
- How much data / how many steps do I need?
- How to understand what the machine is doing?
- How to transfer expert knowledge to the machine?

Knowledge Representation and Reasoning

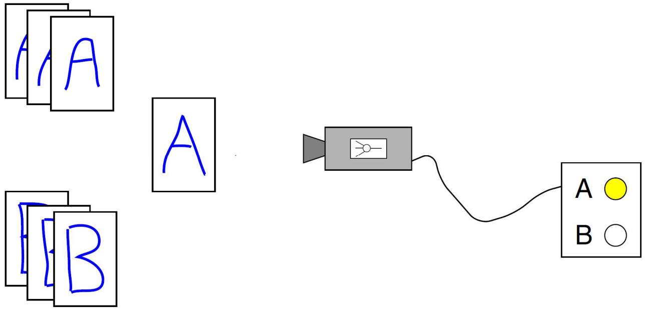
Two ingredients necessary

1. Formal representation of knowledge (of our problem)
2. Reasoning to infer new knowledge (regarding our problem)

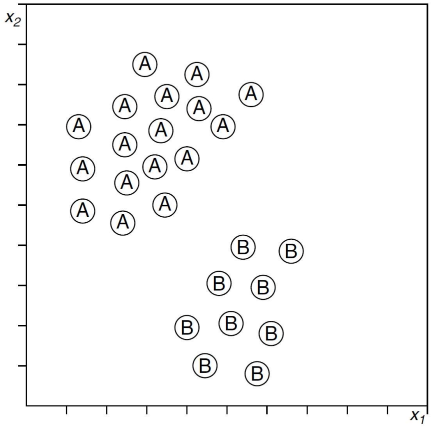
Different scopes / approaches

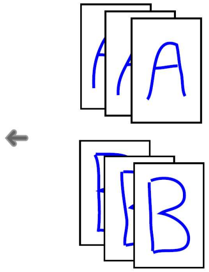
- **Formal Logics (Deductive Reasoning):** set of statements and rules (axioms) how to derive new statements
- **Machine Learning (Inductive Reasoning):** mathematical models which improve through experience on a certain task given a certain performance measure
  - Experience: Usually given as data set
  - Task: a very specific, mostly narrow task
  - Performance measure: a goal on what to improve

Example of a Machine Learning System



Represent the data: gathering, aggregation, cleaning, preparation, transformation

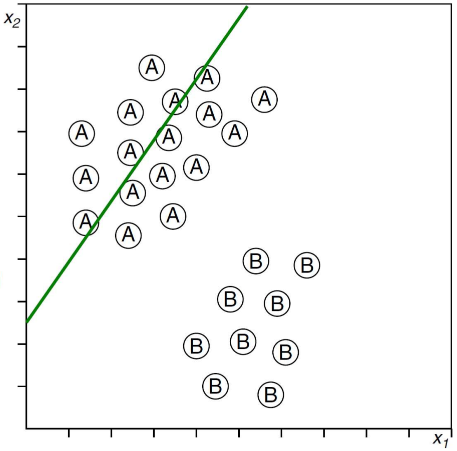




Formulate the Model

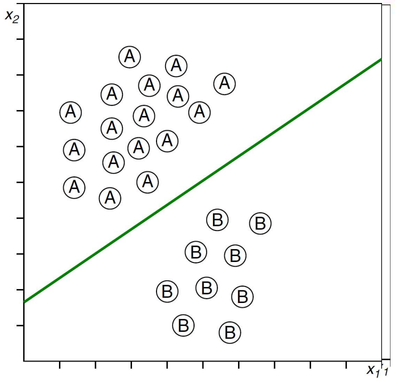
If $\sum_{j=1}^{p} w_j x_j \geq \theta$ then $h(\mathbf{x}) = 1$, and

if $\sum_{j=1}^{p} w_j x_j < \theta$ then $h(\mathbf{x}) = 0$.



Describe the learning algorithm and fit to the data

```
initialize_random_weights($\mathbf{w}$), $t = 0$
REPEAT
  $t = t + 1$
  $(\mathbf{x}, y(\mathbf{x})) = random\_select(D)$
  $error = y(\mathbf{x}) - heaviside(\mathbf{w}^T\mathbf{x})$
  FOR $j = 0$ TO $p$ DO
    $\Delta w_j = \eta \cdot error \cdot x_j$
    $w_j = w_j + \Delta w_j$
  ENDDO
UNTIL(convergence($D, h(D)$)) OR $t > t_{\text{max}}$
```



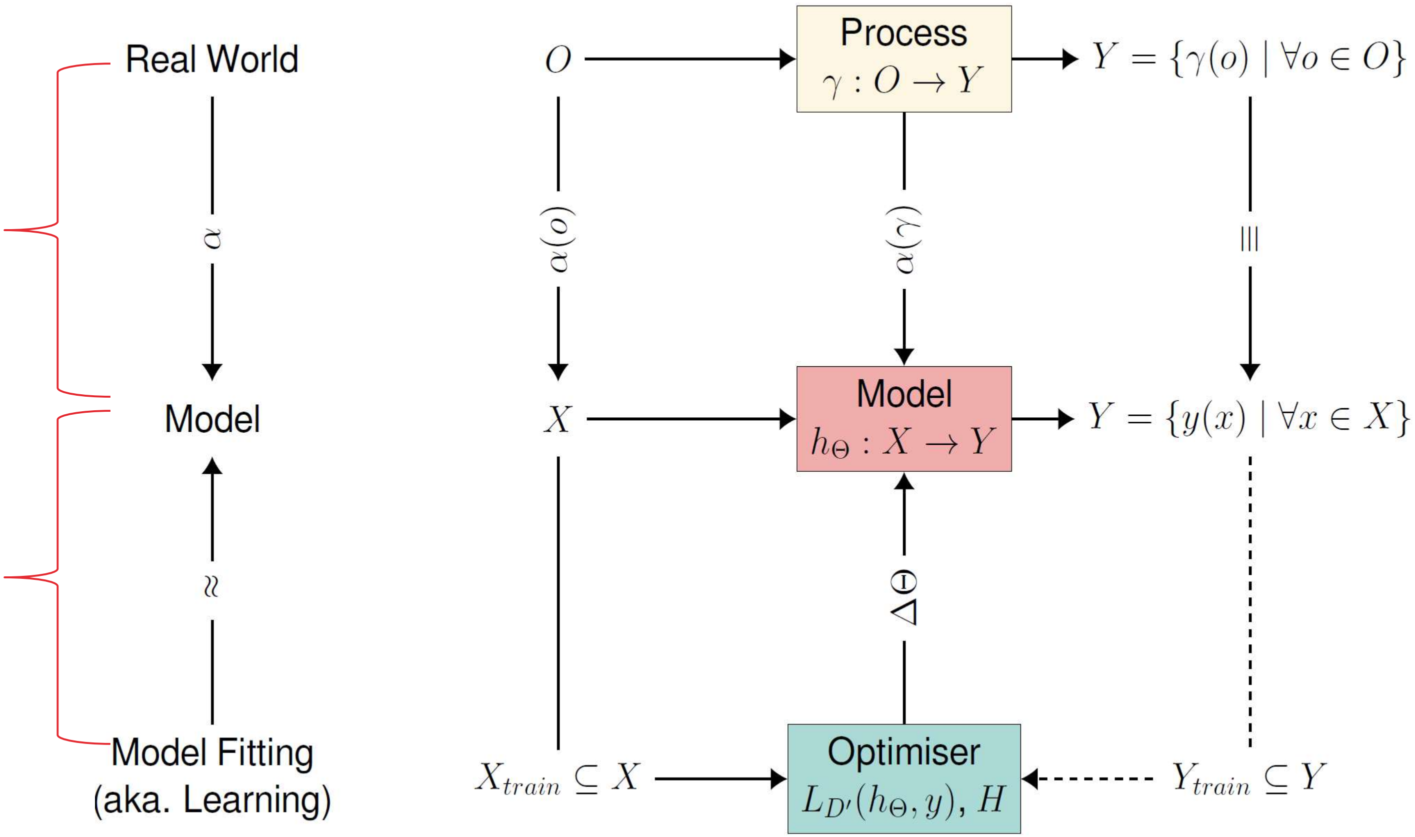
Components of a Machine Learning System (and what can go wrong)

Modelling Concerns

- O vs. X
- $\gamma\ vs.\ h$
- Does the process change over time?
- Does the objects change over time?
- Properties of X?
- Properties of Y?

Fitting and Data Concerns

- How well does our Dataset D represent the true set of objects O?
- Dataset size, data quality (missing value, noise)
- Bias in the data?
- Suitability of the Loss function
- Complexity for the optimiser to fit the loss function to the data (approximation error)



Most errors are hard / impossible to estimate a-priori

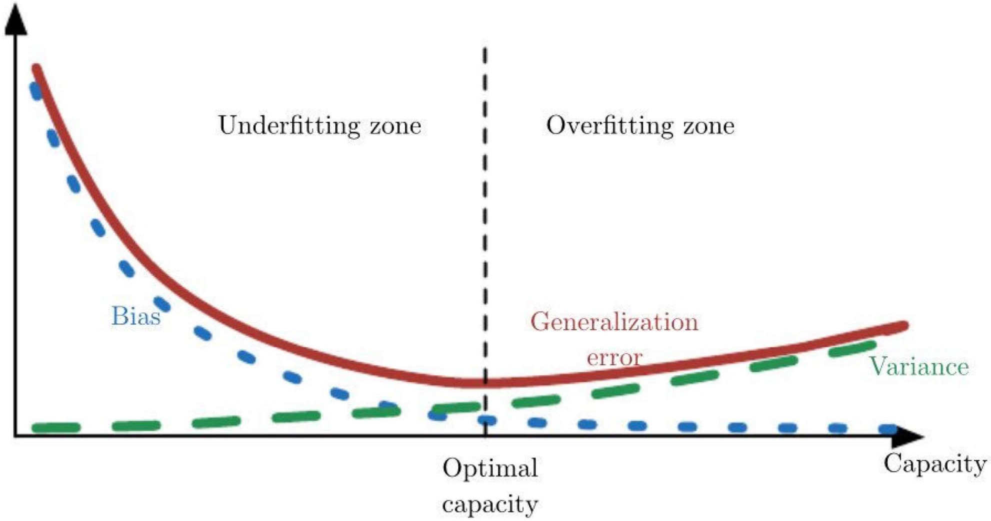
Some general remarks

Computationally Expensive: Exact solutions to complex problems are computationally expensive (most often NP complete). E.g. Probabilistic Reasoning, Logical reasoning. So most algorithms work on the basis of heuristics which are strongly domain dependent.

No-free-lunch: There is no single best algorithm.

Inductive Bias: Assumptions, by which the algorithm generalises. No bias, no learning.

Bias-Variance Trade-off: trade-off in fitting the data perfectly vs. sticking to assumptions in the model



Data?

Attributes

Objects

ID	Check	Status	Income	Risk
1	+	single	125 000	No
2	-	married	100 000	No
3	-	single	70 000	No
4	+	married	120 000	No
5	-	divorced	95 000	Yes
6	-	married	60 000	No
7	+	divorced	220 000	No
8	-	single	85 000	Yes
9	-	married	75 000	No
10	-	single	90 000	Yes

Attribute values may vary from one object to another or one time to another.

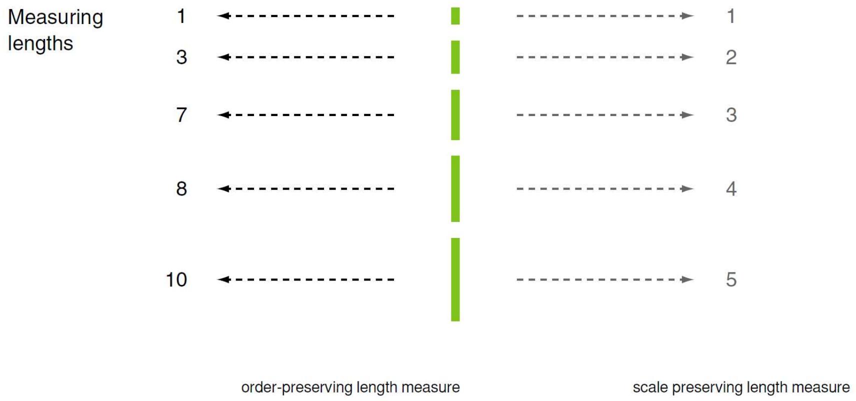
The same attribute can be mapped to different attribute values.

Example: height can be measured in feet or meters.

Different attributes can be mapped to the same set of values.

Example: attribute values for ID and age are integers.

The way an attribute is measured may not match the attribute’s properties:


Measuring lengths
1
3
7
8
10
1
2
3
4
5
order-preserving length measure
scale preserving length measure

Attribute Types

Type		Comparison	Statistics	Examples
categorical (qualitative)	nominal	values are names, only information to distinguish objects $= \quad \neq$	mode, entropy, contingency, correlation, χ^2 test	zip codes, employee IDs, eye color, gender: {male, female}
	ordinal	enough information to order objects $< \quad > \quad \leq \quad \geq$	median, percentiles, rank correlation, run tests, sign tests	hardness of minerals, grades, street numbers, quality: {good, better, best}
numeric (quantitative)	interval	differences are meaningful, a unit of measurement exists $+ \quad -$	mean, standard deviation, Pearson's correlation, t-test, F-test	calendar dates, temperature in Celsius, temperature in Fahrenheit
	ratio	differences and ratios are meaningful $* \quad /$	geometric mean, harmonic mean, percent variation	temperature in Kelvin, monetary quantities, counts, age, length, electrical current

Attribute Types

Type		Permissible transformation	Comment
categorical (qualitative)	nominal	any one-to-one mapping, permutation of values	A reassignment of employee ID numbers will not make any difference.
	ordinal	any order-preserving change of values: $x \to f(x)$, where f is a monotonic	An attribute encompassing the notion of "{good, better, best}" can be represented equally well by the values $\{1, 2, 3\}$.
numeric (quantitative)	interval	$x \to a \cdot x + b$, where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
	ratio	$x \to a \cdot x$, where a is a constant	Length can be measured in meters or feet.

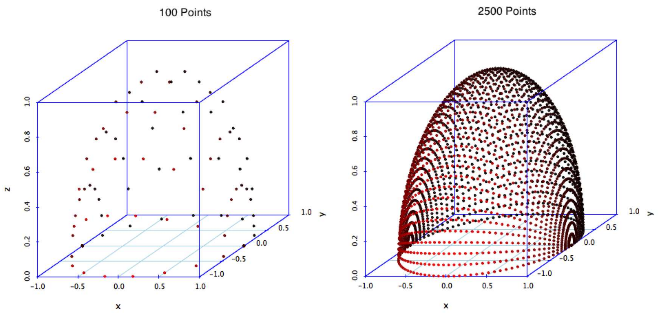
(Big) Data and Machine Learning

- Big Data characterised by 4 Vs:
  - Volume
  - Velocity
  - Variety
  - Veracity
  - One of them is sufficient

- Varity of different kinds of data sets
  - Unstructured Data: Text, image, sound, video
  - Graph Data: Social Networks, Information Networks
  - Record Data: Databases, Lists.
  - Geo-spatial-temporal data: Satellite data
  - Time-series data: Sensors, Transactions

- Large data sets and big data have certainly contributed to a large degree today's success of machine learning and AI.

(Big) Data Pitfalls



Curse of Dimensionality: object density decreases exponentially with the number of attributes

- Given d binary attributes, you can have 2^d objects

Corelation does not imply Causation: Only because to events A and B corelate, it does not mean that one event causes the other

- Giving high-dimensional data, it is more likely that several attribute encode the same information and are thus correlated (e.g. age and birthdate)

Heterogeneous Attribute Types can be challenging

- Nominal features with large number of values (e.g. like ID) can be challenging
- Combining nominal features (Gender) with ratio-scaled features

Solutions

Know your data set, know your algorithm.

- Visualise / explore your data to understand what you are doing

- Know the prerequisites of your algorithms
  - Supported attribute types?
  - Value ranges?
  - Correlated attributes?
  - Size?
  - Missing Values?
  - Noise?

- Preprocess the data accordingly.

Summary

- AI / Machine Learning as a new way to teach machines
- There are well-known theoretical limits to learning algorithms
  - Computationally Expensive, Heuristics, Bias
- Big Data can help, but needs to be handled with care
  - GIGO – Grabage In Grabage Out
  - Understand the data and the data generation process
  - Transform the data appropriately
  - Less is more most often, especially in terms of attributes
- Our Job: properly integrate domain knowledge via
  - Algorithmic choices
  - Data selection and preprocessing
  - (plus Monitoring the system)