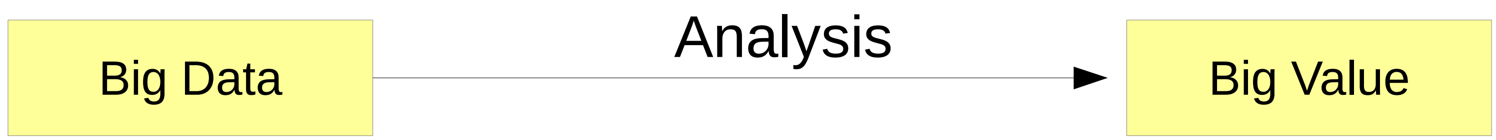
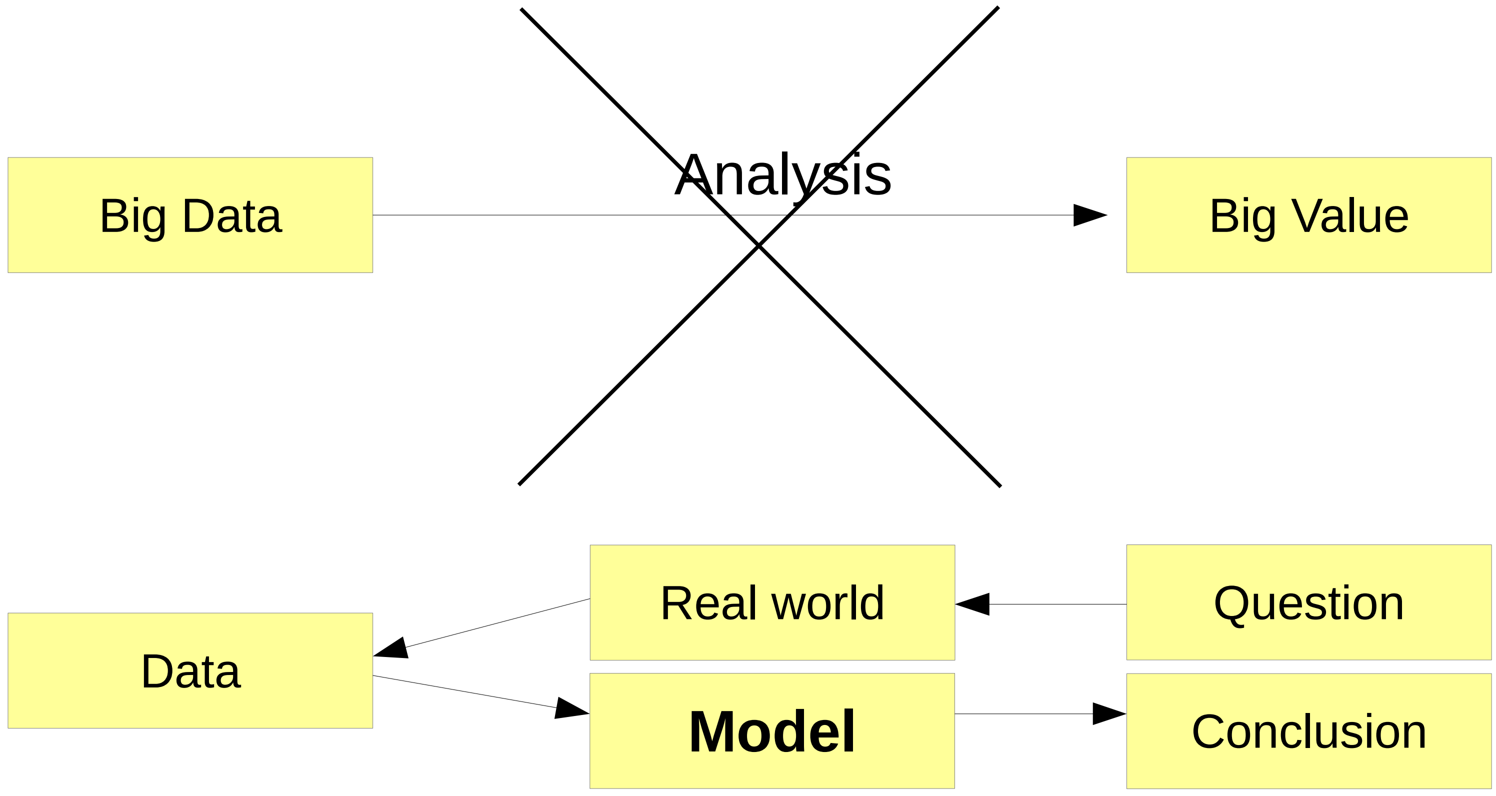
Machine Learning

Anders Holst
SICS

Big Data Analytics



Big Data Analytics



Machine Learning

Use real data to train a model, which can then be used to solve various tasks.

Machine Learning

Use real data to train a model, which can then be used to solve various tasks.

Tasks:

- Classification
- Clustering
- Prediction
- Anomaly detection

Machine Learning

Use real data to train a model, which can then be used to solve various tasks.

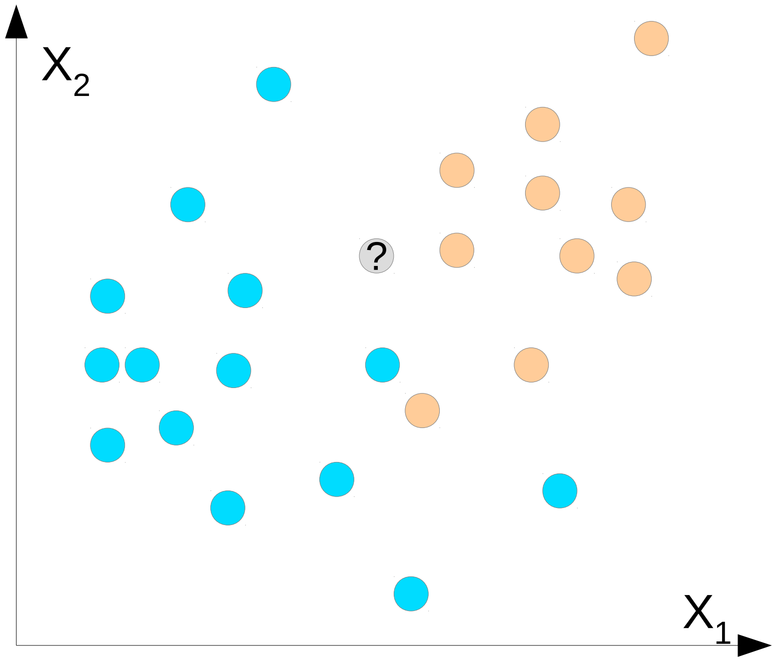
Tasks:

- Classification
- Clustering
- Prediction
- Anomaly detection

Applications:

- Medical diagnosis
- Computer vision
- Speech recognition
- Fraud detection
- Recommender systems
- Sales prediction

Machine Learning



Machine Learning

Data types:

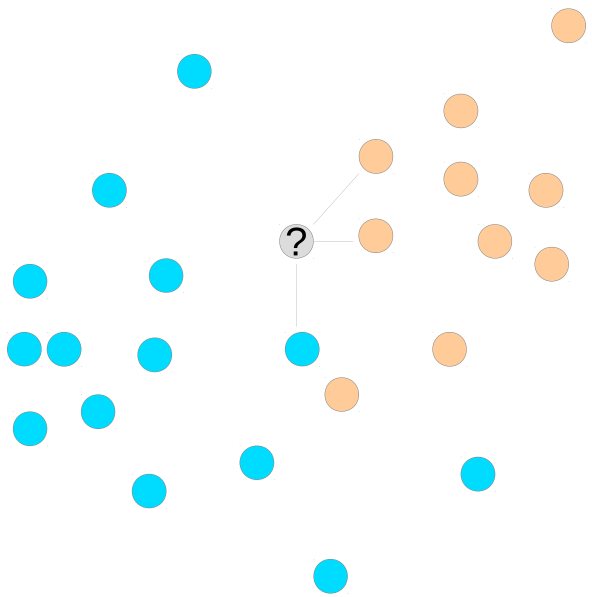
- Binary or discrete
- Continuous values
- Time series
- Natural language text
- Images
- Sound

Machine Learning Methods

- Case based methods
  Table lookup, Nearest neighbour, k-Nearest neighbour
- Logical Inference
  Inductive logic, Decision trees, Rule based systems
- Artificial Neural Networks
  Multilayer perceptrons, Self Organizing Maps, Bolzmann machines,
  Deep neural networks
- Statistical methods
  Naive Bayes, Mixture models, Hidden Markov models, Bayesian networks,
  MCMC, Kernel density estimators, Particle filters
- Heuristic search
  Genetic algorithms, Reinforcement learning, Simulated annealing,
  Minimum Description Length

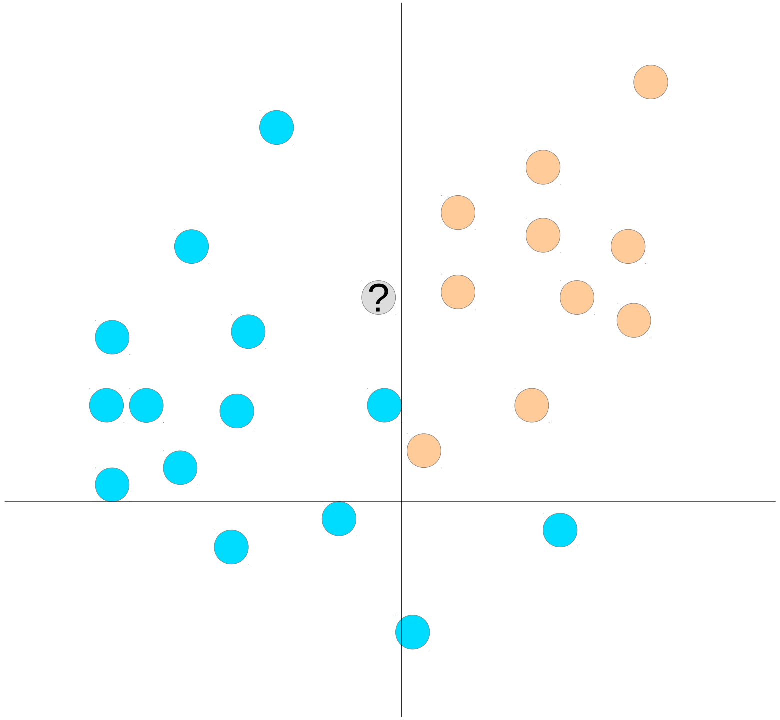
Case based methods

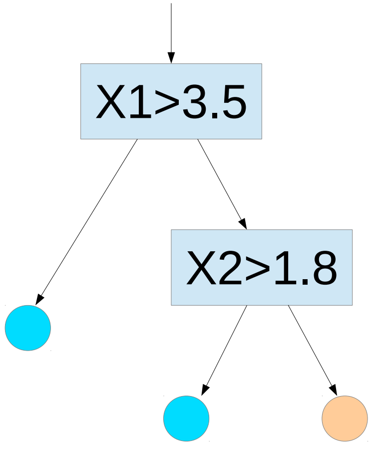
- ”Similar patterns belong to the same class”
- Easy to train (just save every pattern), but takes longer during recall, to find the similar patterns
- Model size increases with the number of seen examples
- Requires specification of a distance measure



Logical Inference

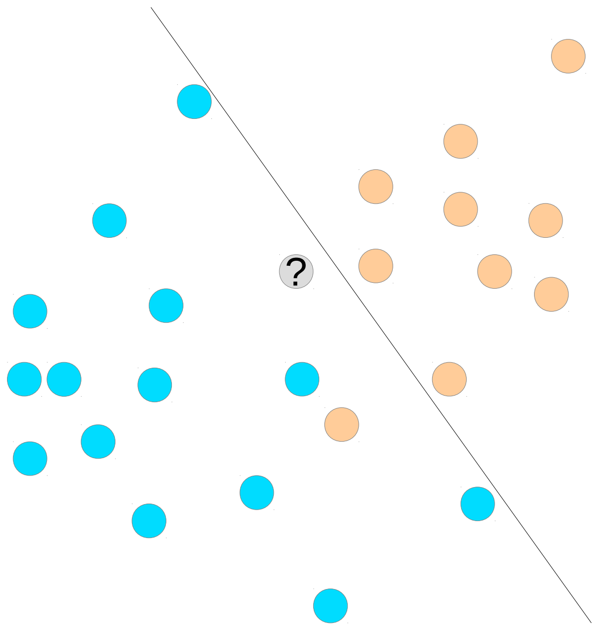
- Construct logical expressions that characterizes the classes
- Typically considers one feature at a time – axis parallell decision regions
- A decision tree be constructed using e.g. information theory

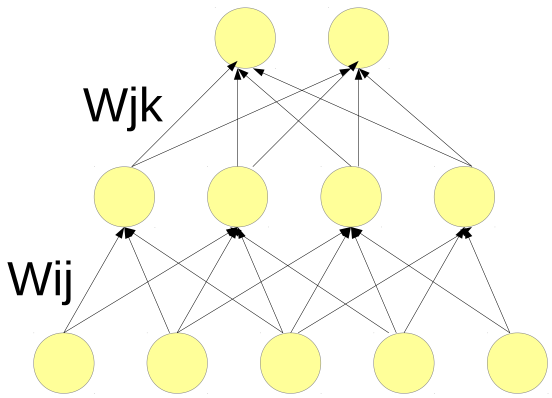




Artificial Neural Networks

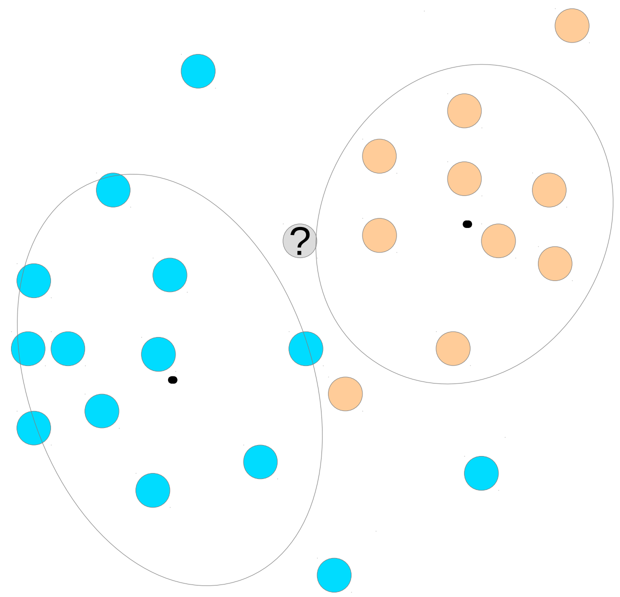
- Inspired by the neural structure of the brain
- Neural units connected by weights. Weights are adjusted to produce the best mapping.
- ”Deep” architectures has gained popularity – requires much data to train

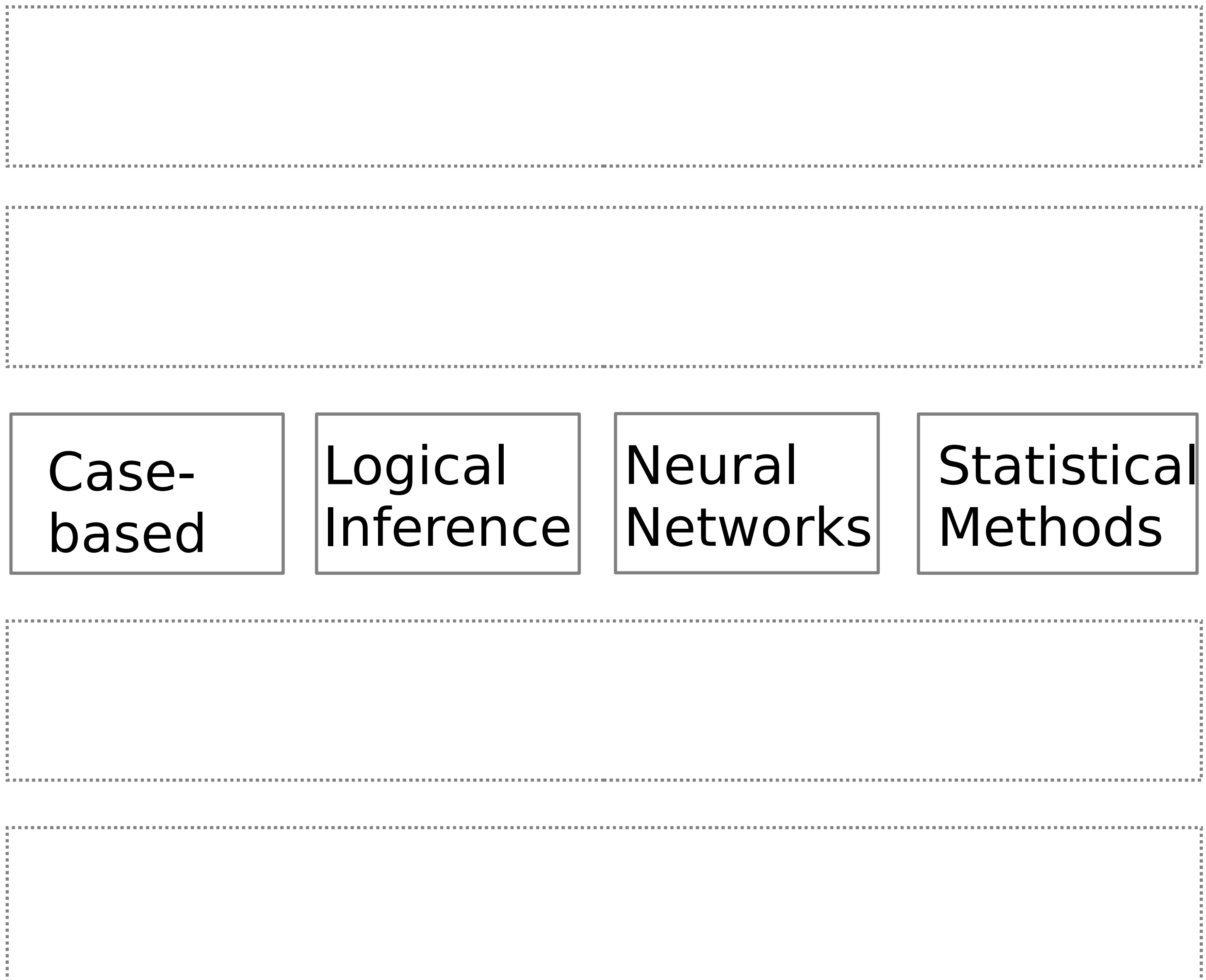




Statistical methods

- Large number of methods, from simple to complex
- The common idea is to calculate the probability of each class given a feature vector, P(c|**x**)
- Parametric versus nonparametric methods – depending on whether the forms of the class distributions are known or not




Case-based
Logical Inference
Neural Networks
Statistical Methods

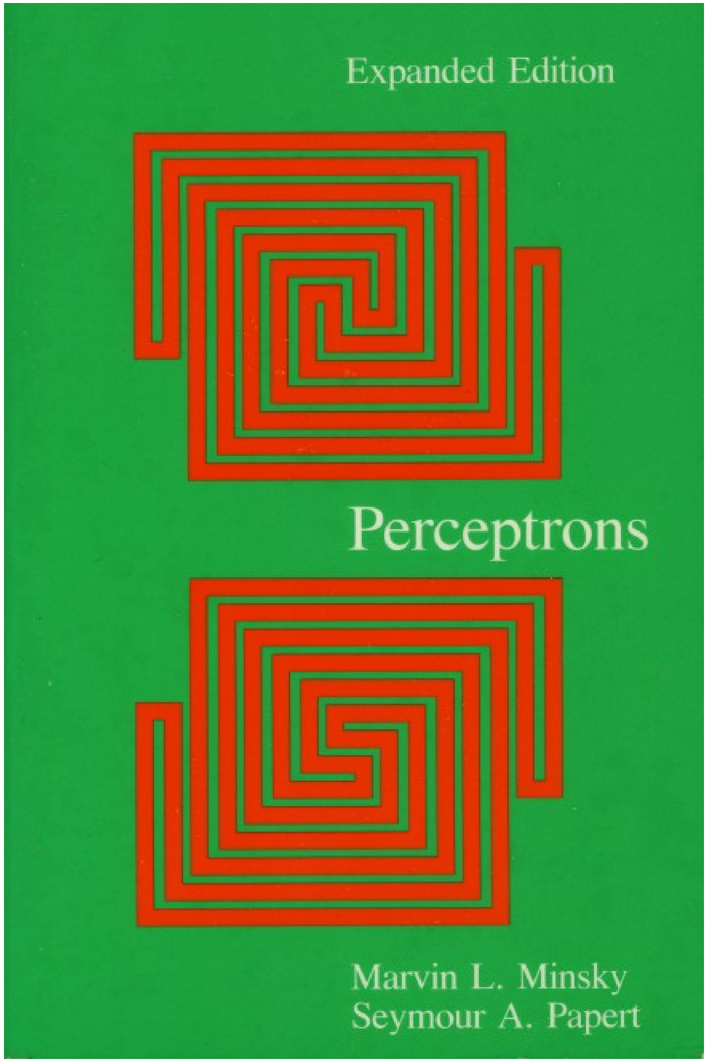
Representation

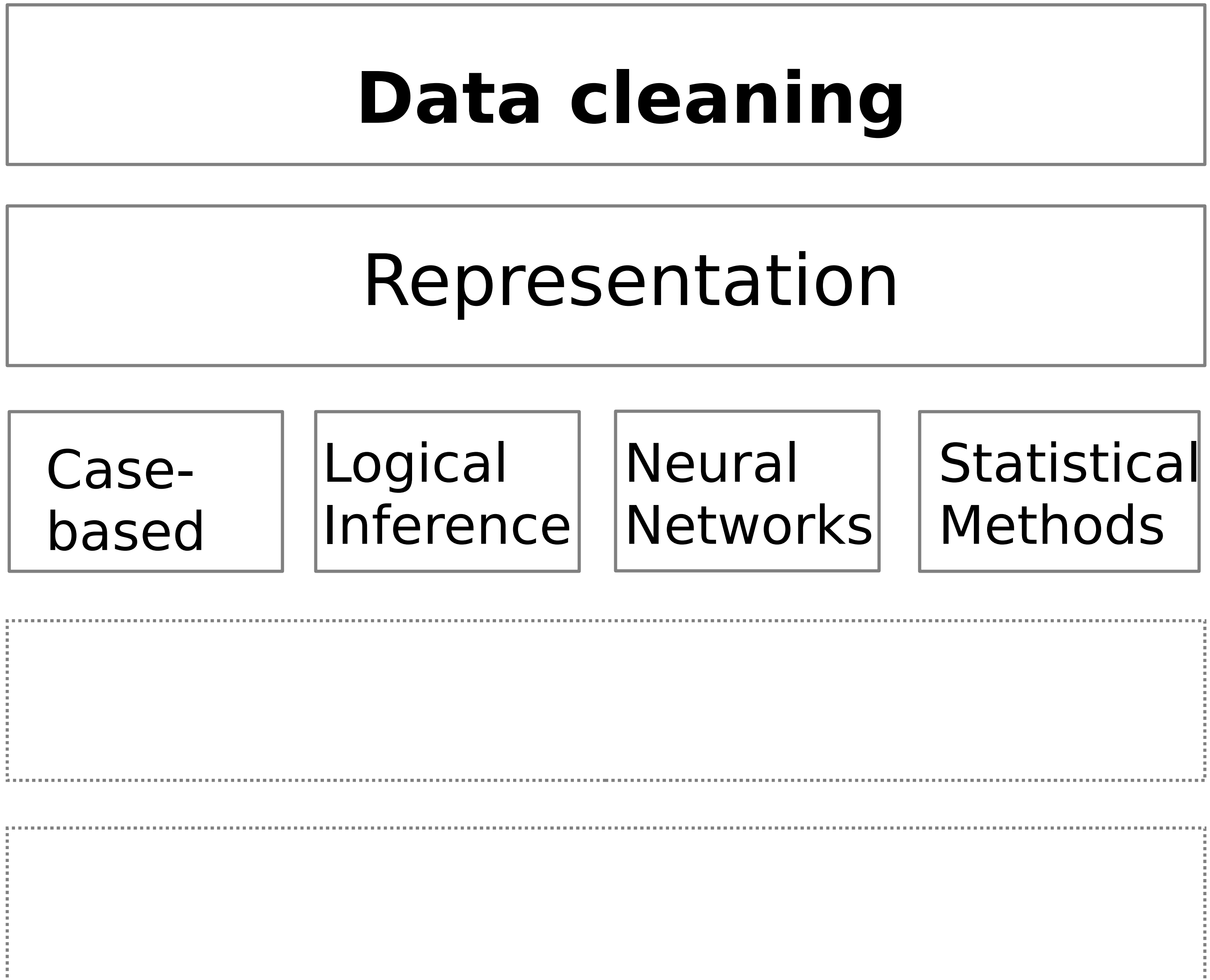
Case-based | Logical Inference | Neural Networks | Statistical Methods

Representation

- The exact choice of method is often not critical, but the choise of representation of features is:
  - With the wrong representation *no* method will succeed
  - Once you have found a good representation, almost *any* method will do
- Once preprocessing has turned data into something reasonable, a simple model may be sufficient
  - With limited amount of independent data, the number of parameters must be kept low, so keep it as simple as possible
- Finding a suitable representation requires much domain knowledge and problem understanding
  - No black box solution in general

Neural Network book, 1969




Data cleaning
Representation
Case-based
Logical Inference
Neural Networks
Statistical Methods

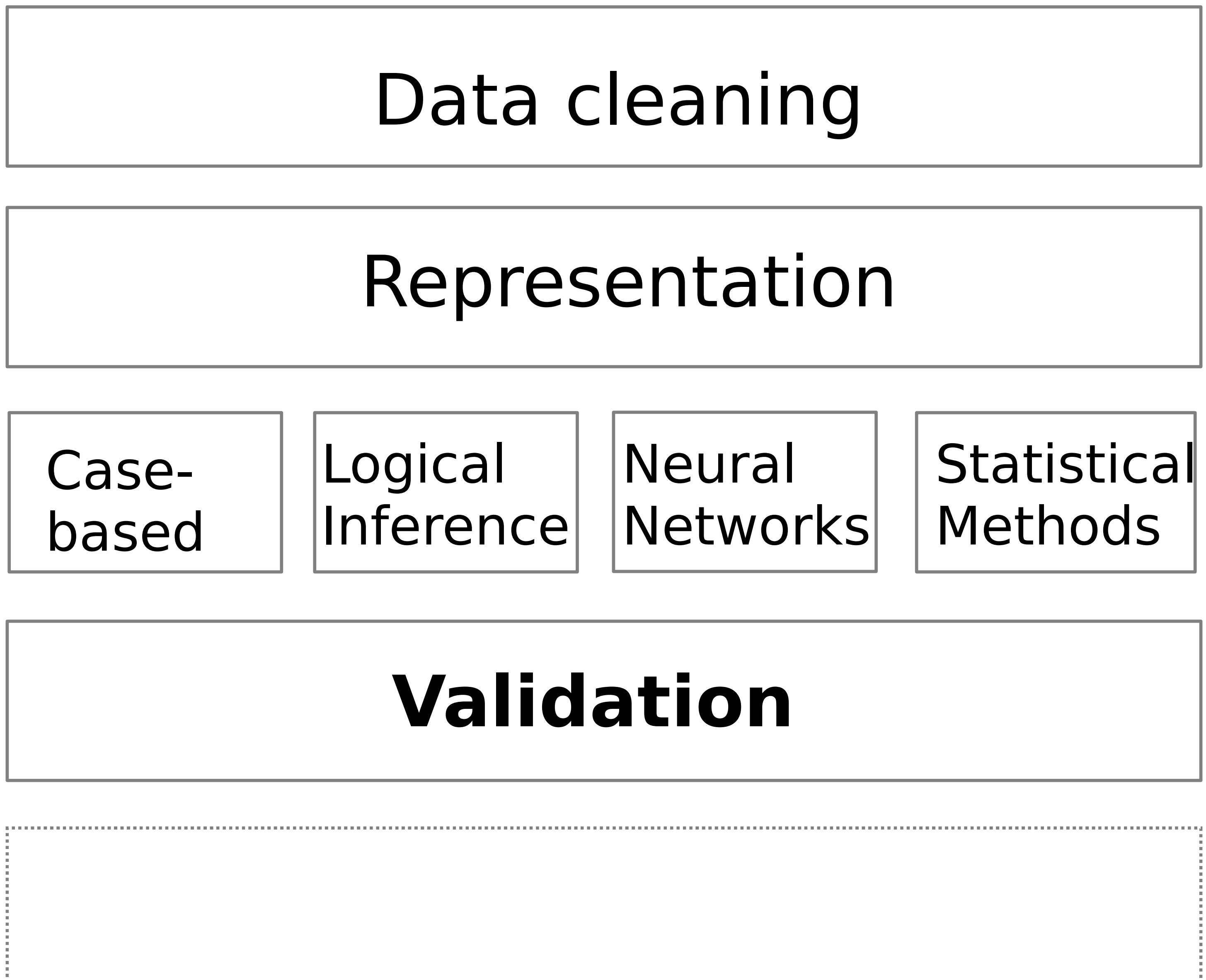
Data cleaning

Real data is not clean:

- Missing data
- Out of sync fields
- Misspellings
- Special values (temperature -9999)
- Spikes (10e+14)
- Dirty or drifting sensors (0.3 – 100.3 %)
- Data from different sources (old / new), with slightly different meaning
- Inconsistent data
- Irrelevant data

Data cleaning

Attr 1	Attr 2	Attr3	Attr 4	Attr 5
12.2827	2002080612220500	10.47	5.2	Cool. on
12.2826	2002080612220622	15.39	4.7	Switch
12.2825	2002080612220743	12.66	5.9	hasp temp 680
12.2824	2002080612220886	-999.0	22.8	Hasp-temp
1.22823	2002080612221012	-999.0	Overflow	cool
12.2819	2002080612221136	-999.0	Overflow	Cooling
12.2815	1858111700000000	13.49	Error	cooling on
122821	1858111700000000	25.85	Error	sw.
12.2823	2002080612221631	22.98	0.6	not in phase
...	...	...	...	...


Data cleaning
Representation
Case-based
Logical Inference
Neural Networks
Statistical Methods
Validation

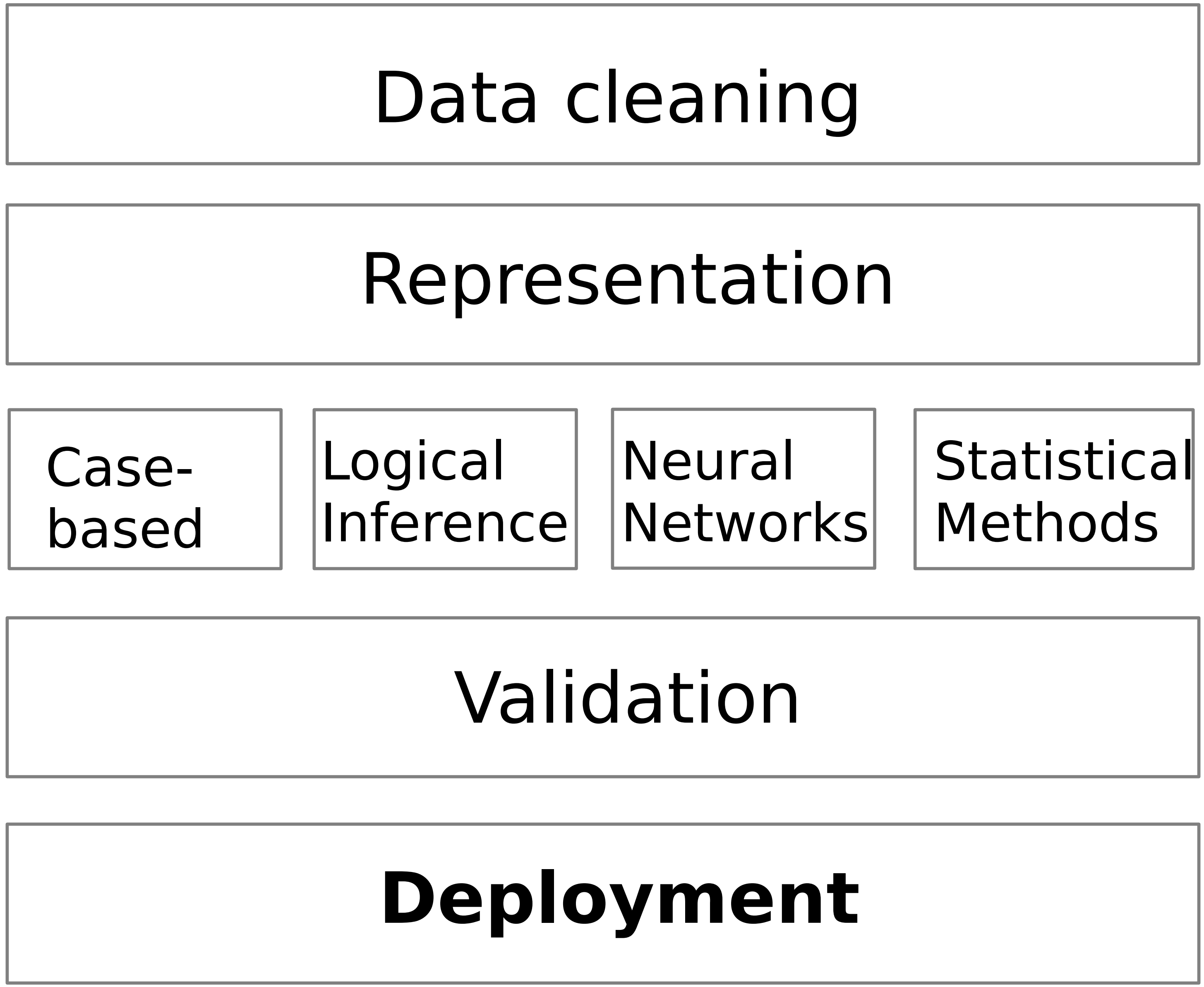
Validation

- “Validation” is used to estimate the performance on new data, i.e. how the model would perform when actually used
- To get good generalization you must avoid overtraining the machine learning model
- There are unimaginably many ways that makes the result look better in the laboratory than in the real life
- However hard you try to avoid it, you will *always* get too optimistic validation results!

Validation

Some ways to guarantee overtraining:

- Too few data samples
- Too complicated model
- Too similar training, test and validation samples
- Fine-tuning your parameters
- Evaluating several models with the same validation set


Data cleaning
Representation
Case-based
Logical Inference
Neural Networks
Statistical Methods
Validation
Deployment

Deployment

- The method is on its own
- Keep it simple and robust
- Must the network be regularly retrained?
  Can the “ground truth” be trusted?
  Can stability and performance be guaranteed?
- Did your pre-study test the right thing?
  Distinction between prediction and control
  Distinction between prediction and causation
- Be prepared to go all over the process again

Data cleaning

Representation

Case-based

Logical Inference

Neural Networks

Statistical Methods

Validation

Deployment

Conclusions

- Thoroughly understand the problem you are working on and try to understand the process that generated the data
- Select a suitable representation, of the relevant features
- Take extreme care with validation, and test the application on as much real-world data as you can
- Keep it as simple as possible (but still powerful enough to solve the problem at hand).