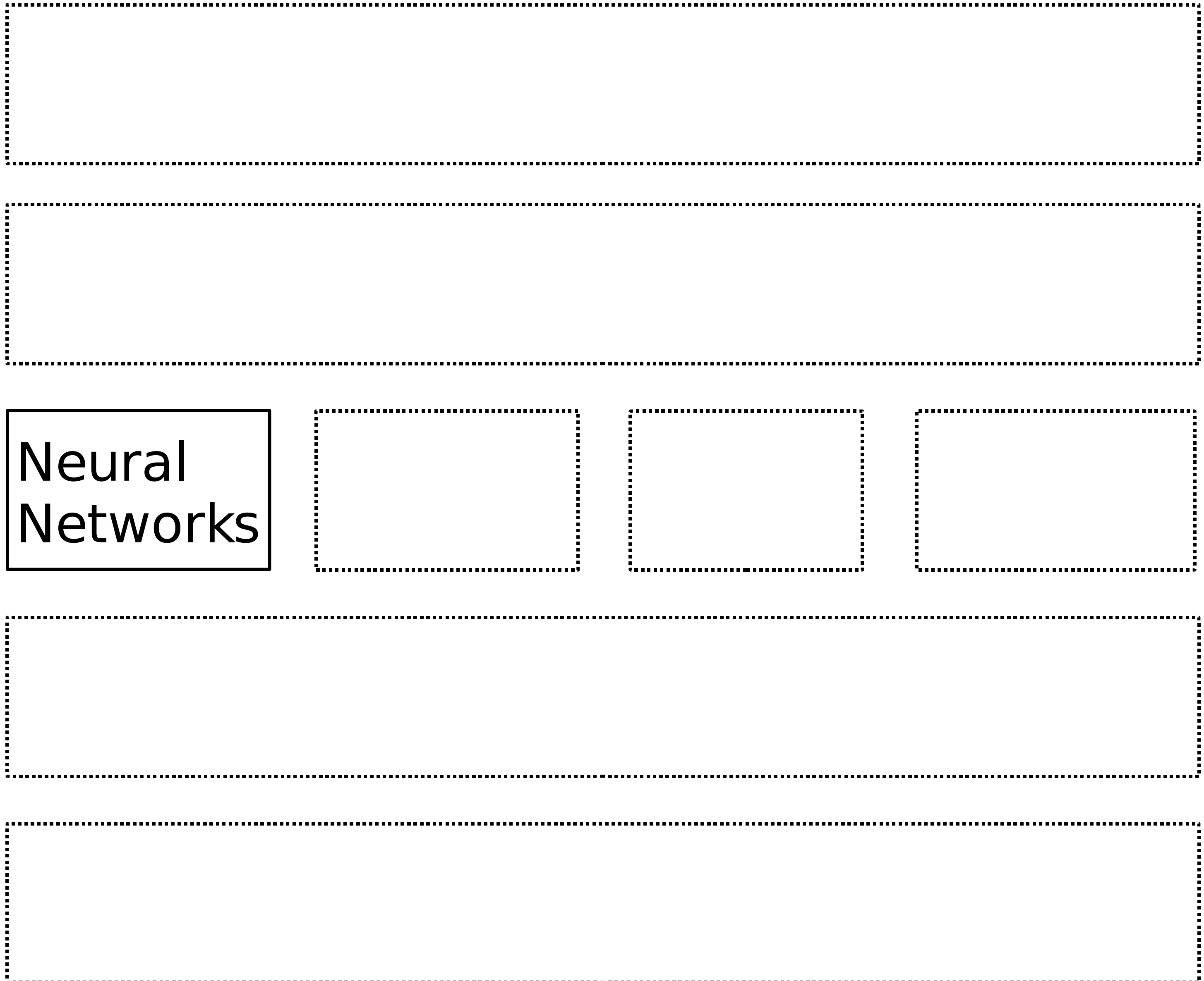
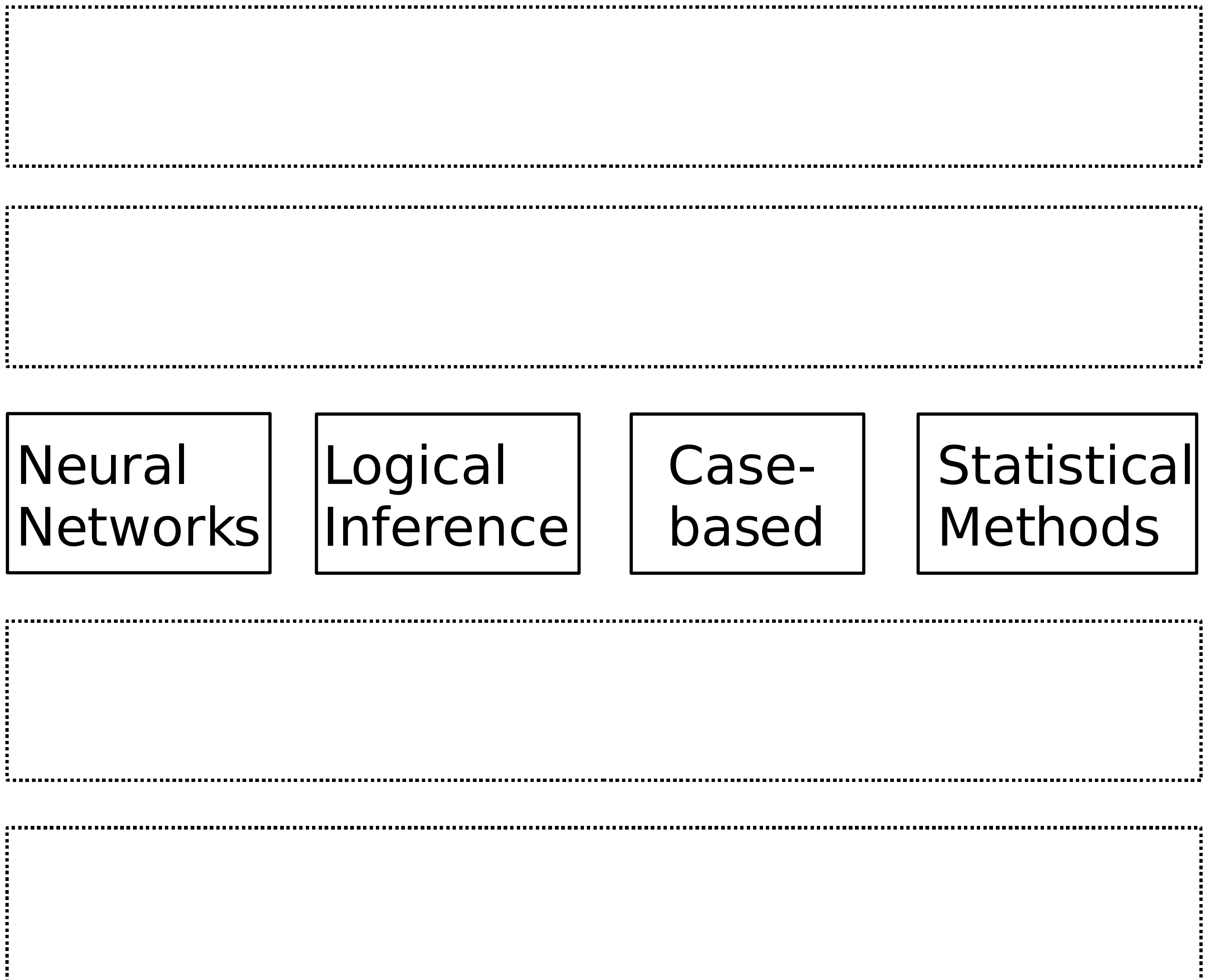
Aspects of practical application of ANN

Anders Holst
SICS

Neural Networks


Neural
Networks


Neural Networks
Logical Inference
Case-based
Statistical Methods

Machine Learning methods

Typical machine learning tasks:

- Classification / Diagnosis
- Prediction / Prognosis
- Clustering / Categorization

Different domain characteristics:

- Logical / Discrete / Continuous
- Few or many attributes
- Few or many classes
- Deterministic / Noisy

Machine Learning methods

Neural Networks
- Multi-layer Perceptrons
- Self-Organizing Maps
- Hopfield Networks
- Bolzmann machines

Case-based methods
- Table-lookup
- Nearest Neighbour
- k-Nearest Neighbour

Logical Inference
- Decision Rules
- Decision Trees
- Rule based systems

Statistical methods
- Regression
- Naïve Bayes
- Mixture Models
- Graph Models
- MCMC

Machine Learning methods

- The exact choice of method is usually not critical
- The choice of problem representation *is* critical!
- This requires domain understanding

Representation

Neural Networks

Logical Inference

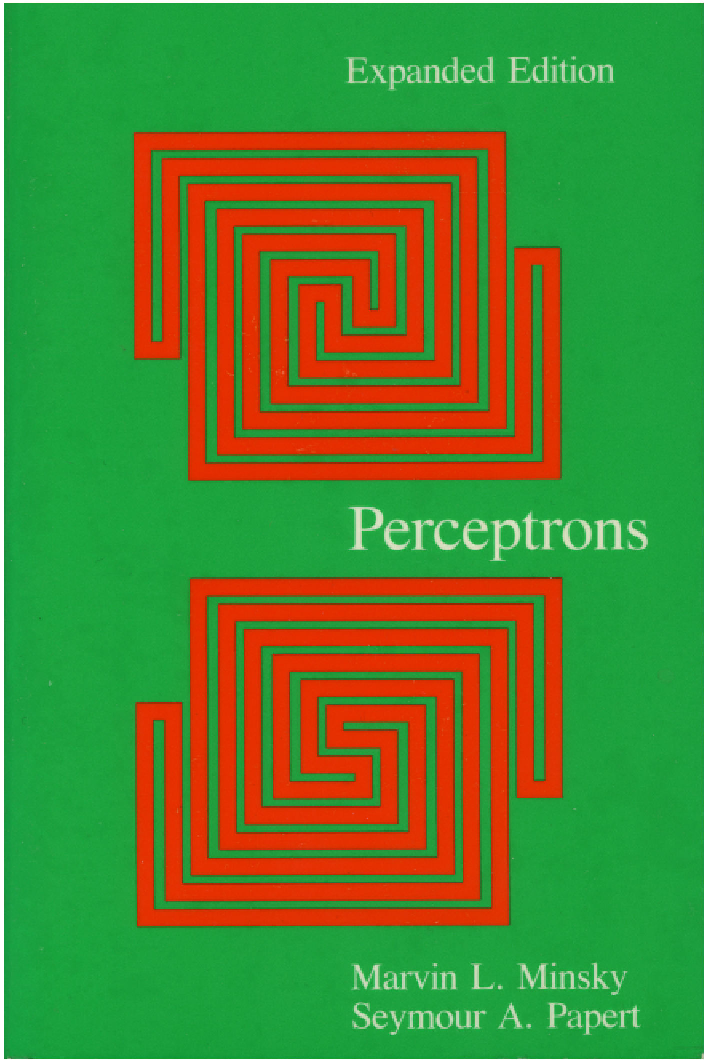
Case-based

Statistical Methods

Representation

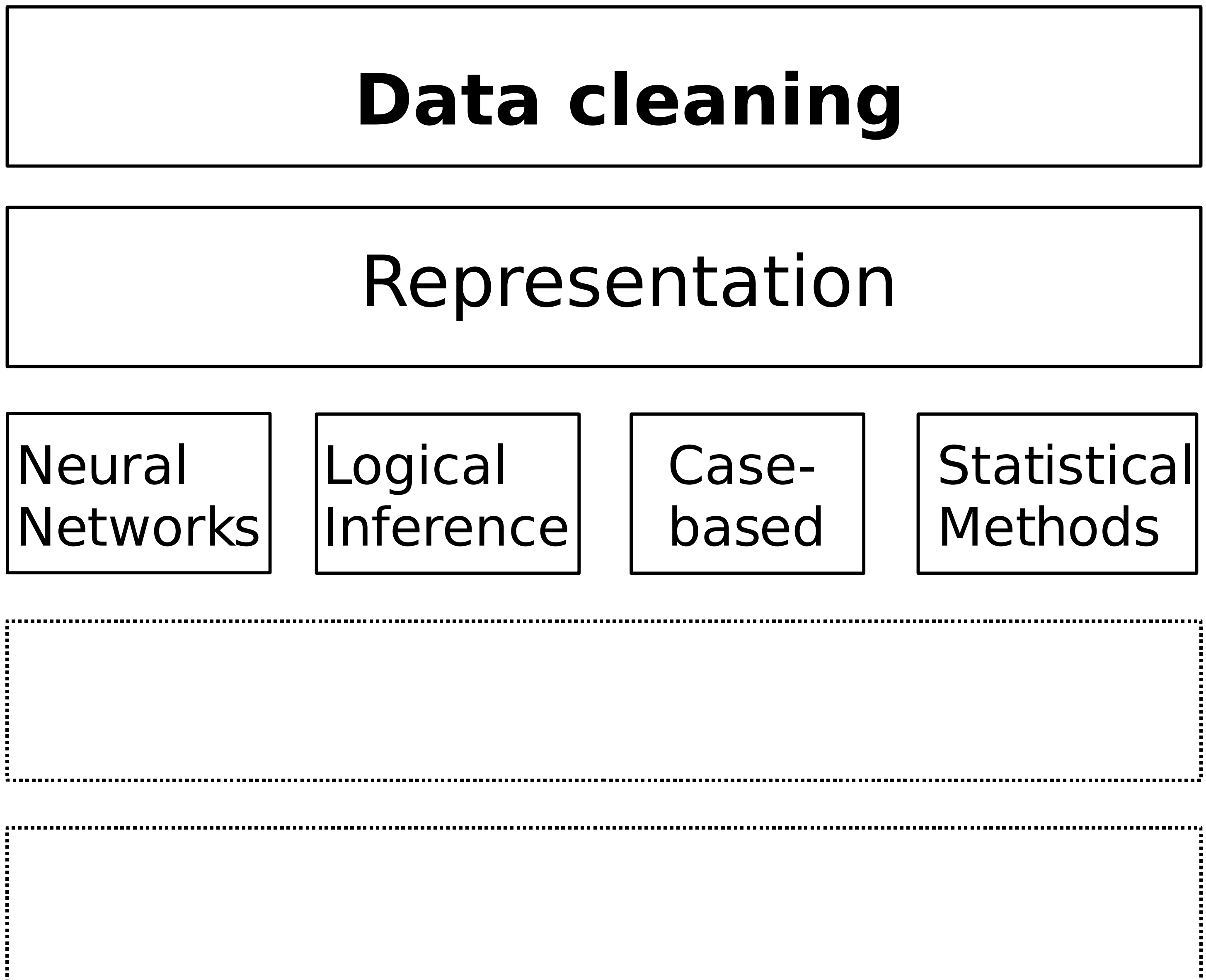
- With the wrong representation *no* method will succeed
- Once you have found a good representation, almost *any* method will do
- Once preprocessing has turned data into something reasonable, a simple model may be sufficient
- With limited amount of independent data, the number of parameters must be kept low

Neural Network book, 1969



Representation

- Many types of real data, e.g. images, sounds, time series, free text, binary register dumps, etc, require special pre-processing to pick out whats relevant
- Look out for invariances


Data cleaning
Representation
Neural Networks
Logical Inference
Case-based
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

Data cleaning

Attr 1	Attr 2	Attr3	Attr 4	Attr 5
12.2827	2002080612220500	10.47	5.2	Kyln. på
12.2826	2002080612220622	15.39	4.7	Valsbyte
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	kyln
12.2819	2002080612221136	-999.0	Overflow	Kylning
12.2815	1858111700000000	13.49	Error	karkylning på
122821	1858111700000000	25.85	Error	valsb.
12.2823	2002080612221631	22.98	0.6	ej i dragläge
...	...	...	...	...

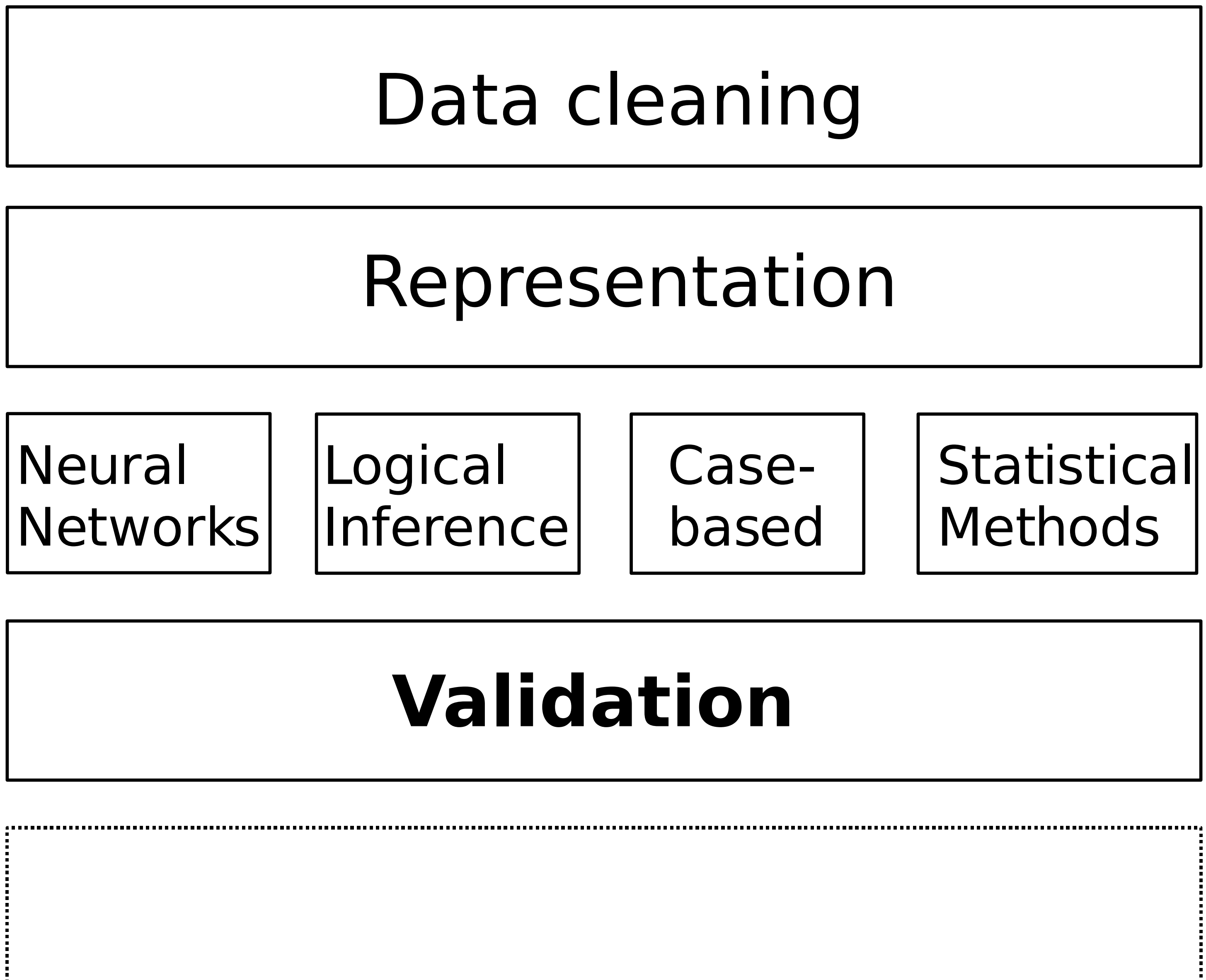
Data cleaning

One even earlier step: Getting data

- Not so easy, often takes time
- Data samples are always too few!
- Even when they claim there are huge amounts of data, the relevant part may be very small
- A large amount of input attributes is no guarantee for successful modelling. The relevant attributes may still not be included.
- An iterative process to get the right data

Mid-time conclusions

- Pre-processing (understanding the domain and the problem, getting relevant data, cleaning the data, and finding a good representation) takes more than 90% of the time of a project
- No black box method - domain knowledge is critical for success


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

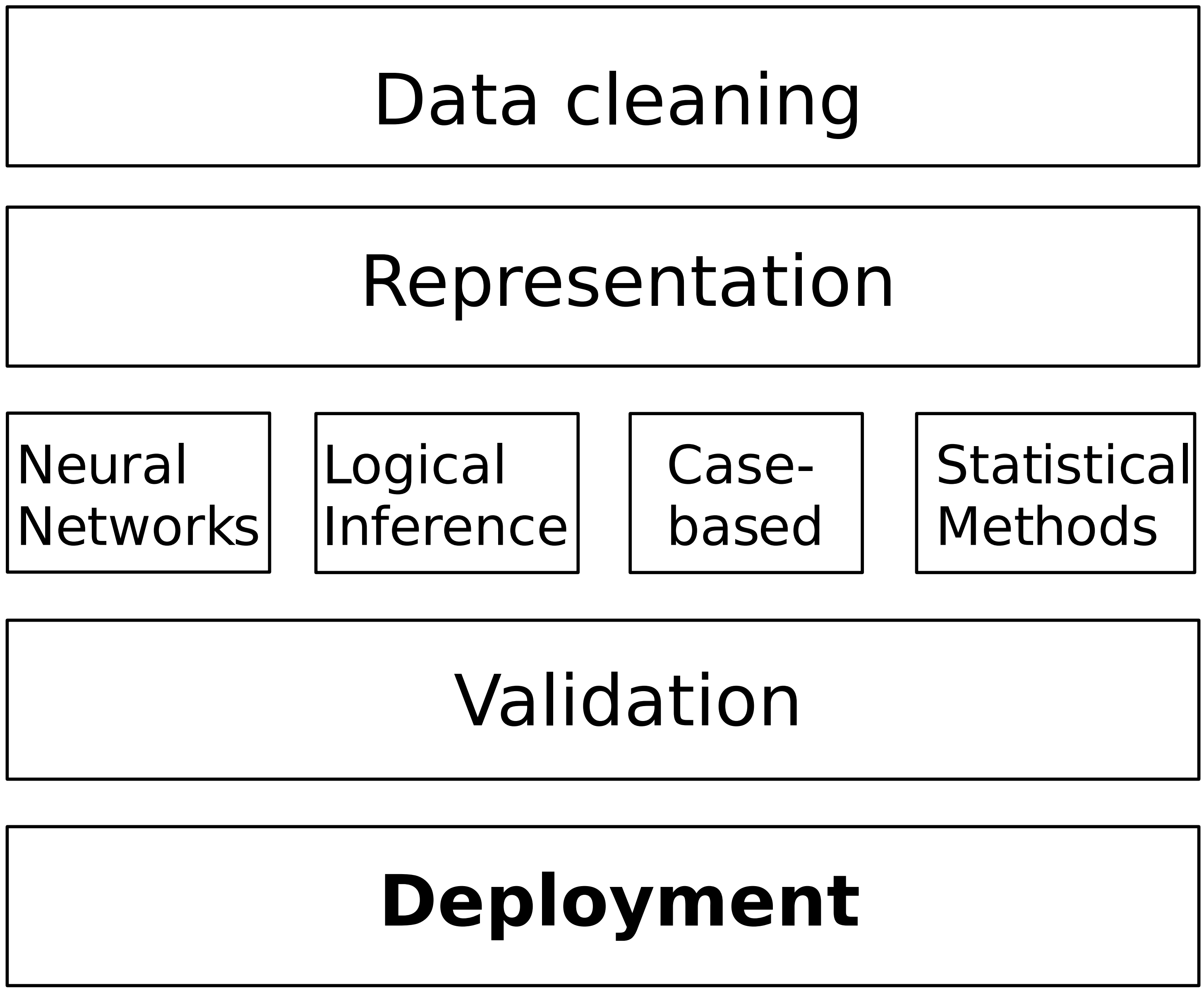
Validation

- “Validation” is used to estimate the performance on new data, i.e. how it would perform when actually used
- To get good generalization you must avoid overtraining the network
- 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

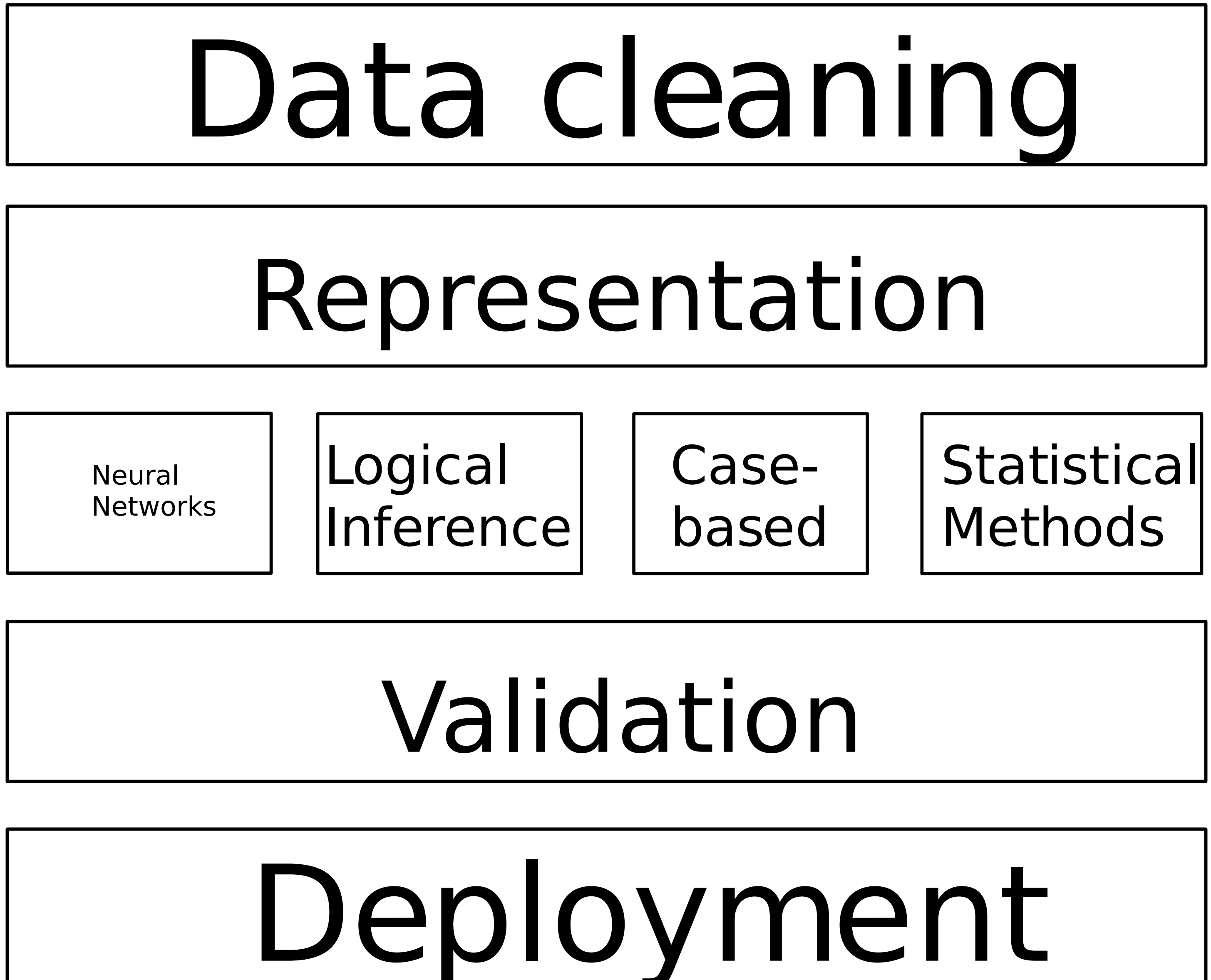

Data cleaning
Representation
Neural Networks
Logical Inference
Case-based
Statistical Methods
Validation
Deployment

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

Conclusions

- Thoroughly understand the problem you are working on and try to understand the process that generated the data
- Take extreme care with validation
- Test the application on as much real-world data as you can
- Keep it as simple as possible


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

Commercial applications of ANN

FossTecator:
ANN modelling of NIR data - a tool for managing local and global commercial transactions in grain trading.

CellaVision:
Provide systems for the healt-care sector for automatic digital cell morphology. DiffMaster - Blood cell classification using neural networks.

Pharma Vision Systems AB:
Automatic microscopy of particle shape and size.

Malcolm - ANN for graphic art quality control. (Halmstad University)