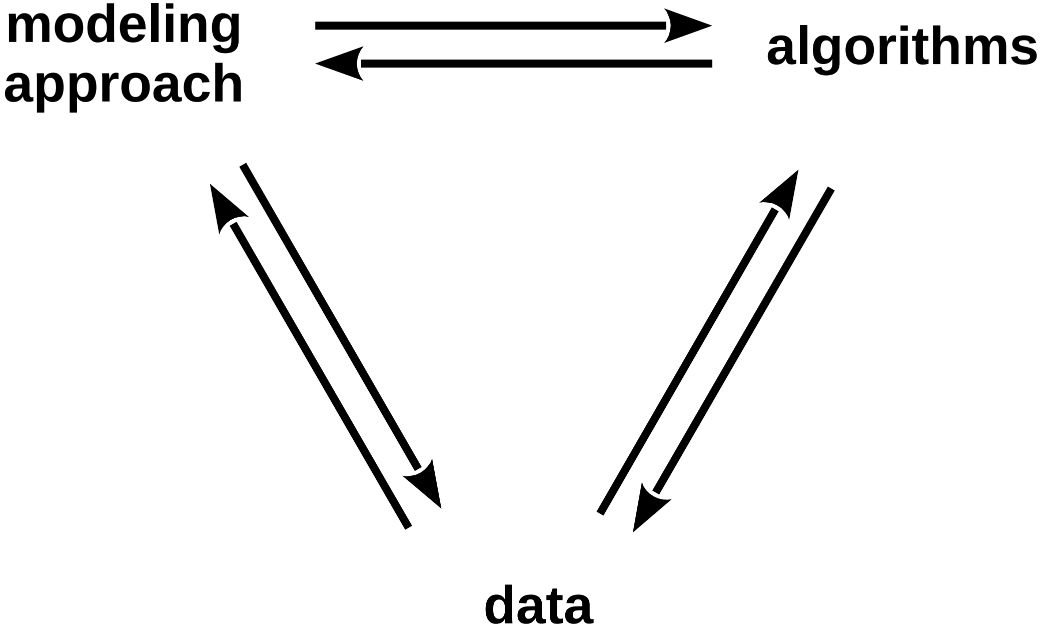
The potential of Machine Learning in biology

Robert Różański



modeling approach

algorithms

data

level of detail

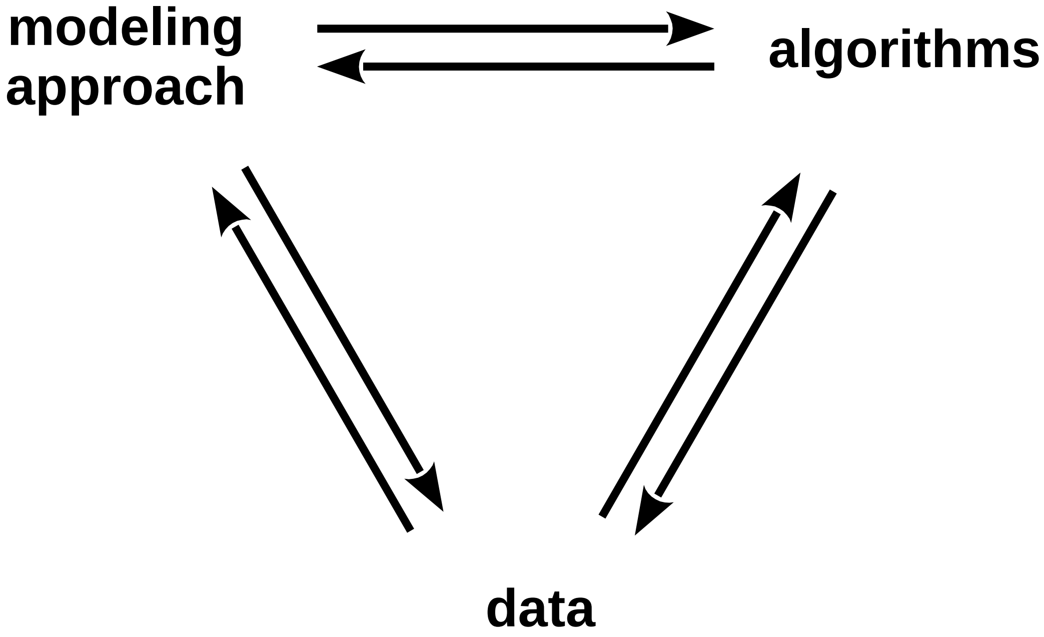
time

type of abstraction

space

randomness

quantitative aspects



modeling approach

algorithms

data

level of detail

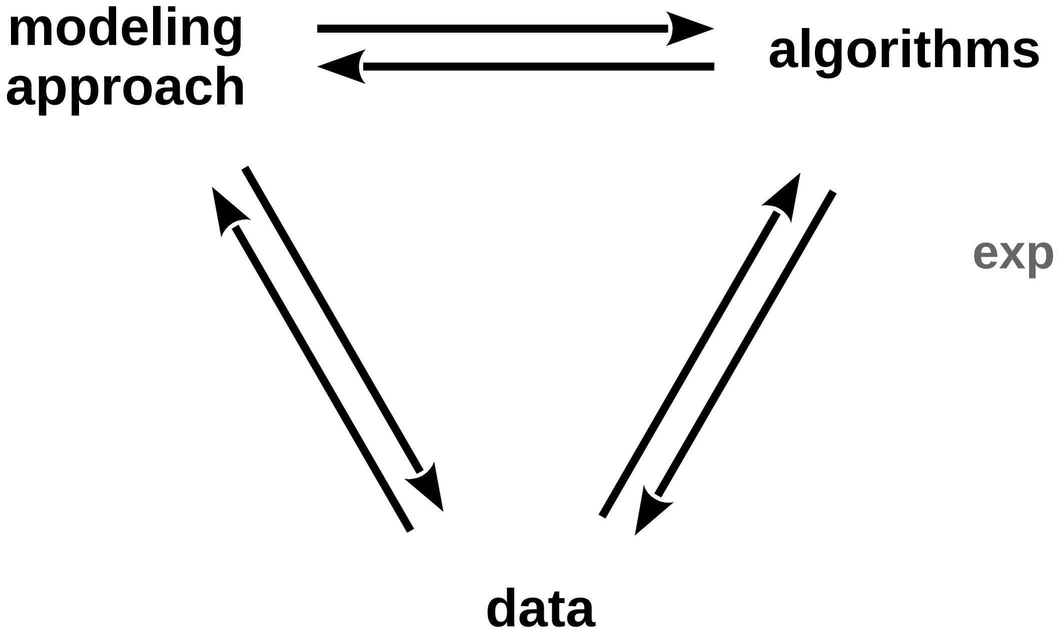
time

type of abstraction

space

randomness

quantitative aspects



modeling approach

algorithms

data

model checking

prediction

forward simulation

(in)consistency check

revision

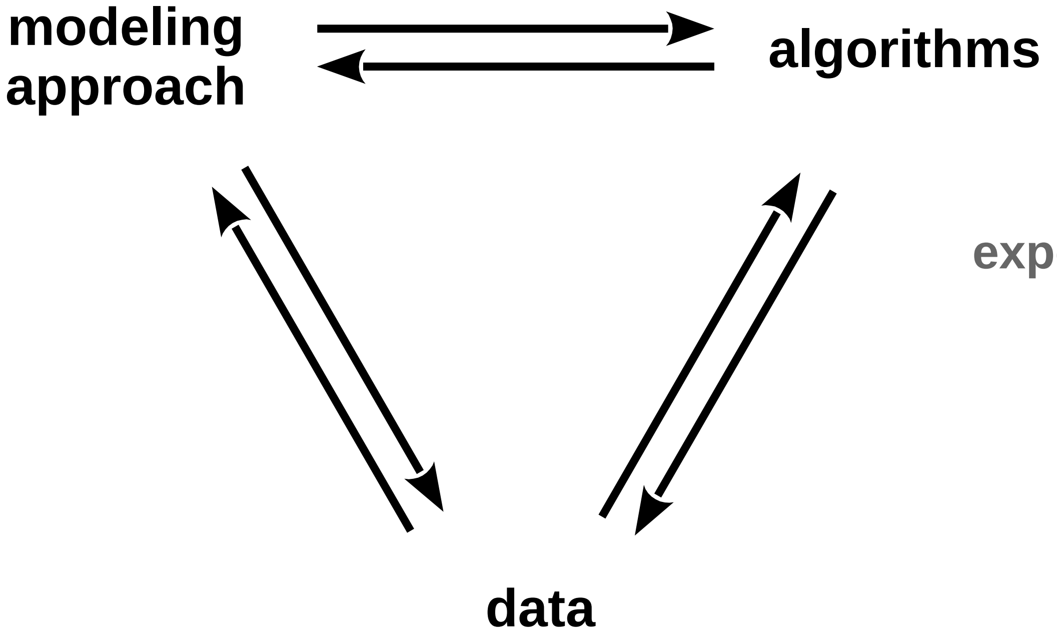
backward search

parameter fitting

model construction

experiment design

classification



modeling approach

algorithms

data

level of detail

time

type of abstraction

space

randomness

quantitative aspects

model checking

prediction

forward simulation

(in)consistency check

revision

backward search

parameter fitting

model construction

experiment design

classification

qualitative

quantitative

errors

availability

noise

representation of experiments

So what can be done?

- symbolic ML
- standard supervised ML
- standard unsupervised ML
- Deep Learning

Symbolic ML

Symbolic ML

- Modeling approach: symbolic models (Boolean Networks, Bayesian Networks, Petri Nets, Pathway Logic, ... (many more))

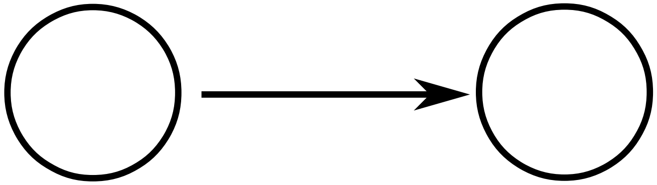
Symbolic ML

- Modeling approach: symbolic models (Boolean Networks, Bayesian Networks, Petri Nets, Pathway Logic, ... (many more))
- quantitative/spatial/temporal aspects abstracted away to some degree (models can have some discrete states and transition between them)

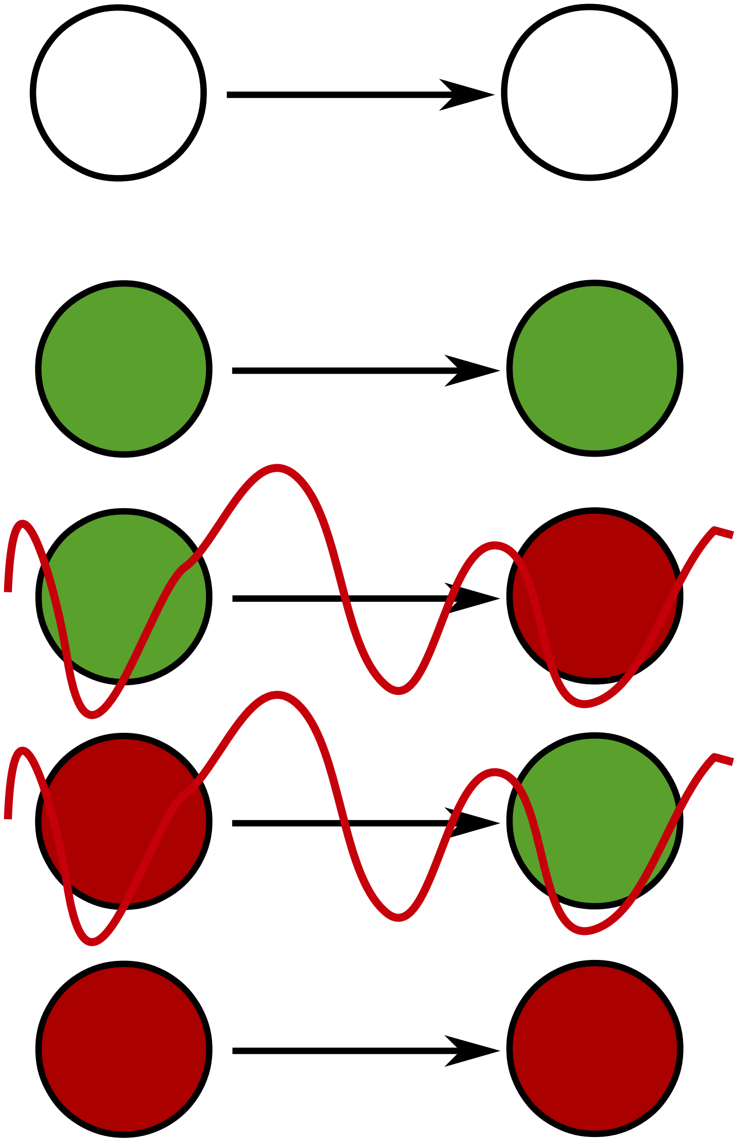
Symbolic ML

- Modeling approach: symbolic models (Boolean Networks, Bayesian Networks, Petri Nets, Pathway Logic, ... (many more))
- quantitative/spatial/temporal aspects abstracted away to some degree (models can have some discrete states and transition between them)
- many algorithms exist - the simpler the formalism, the better (simulation, model checking, construction, revision, exp. design, ...)

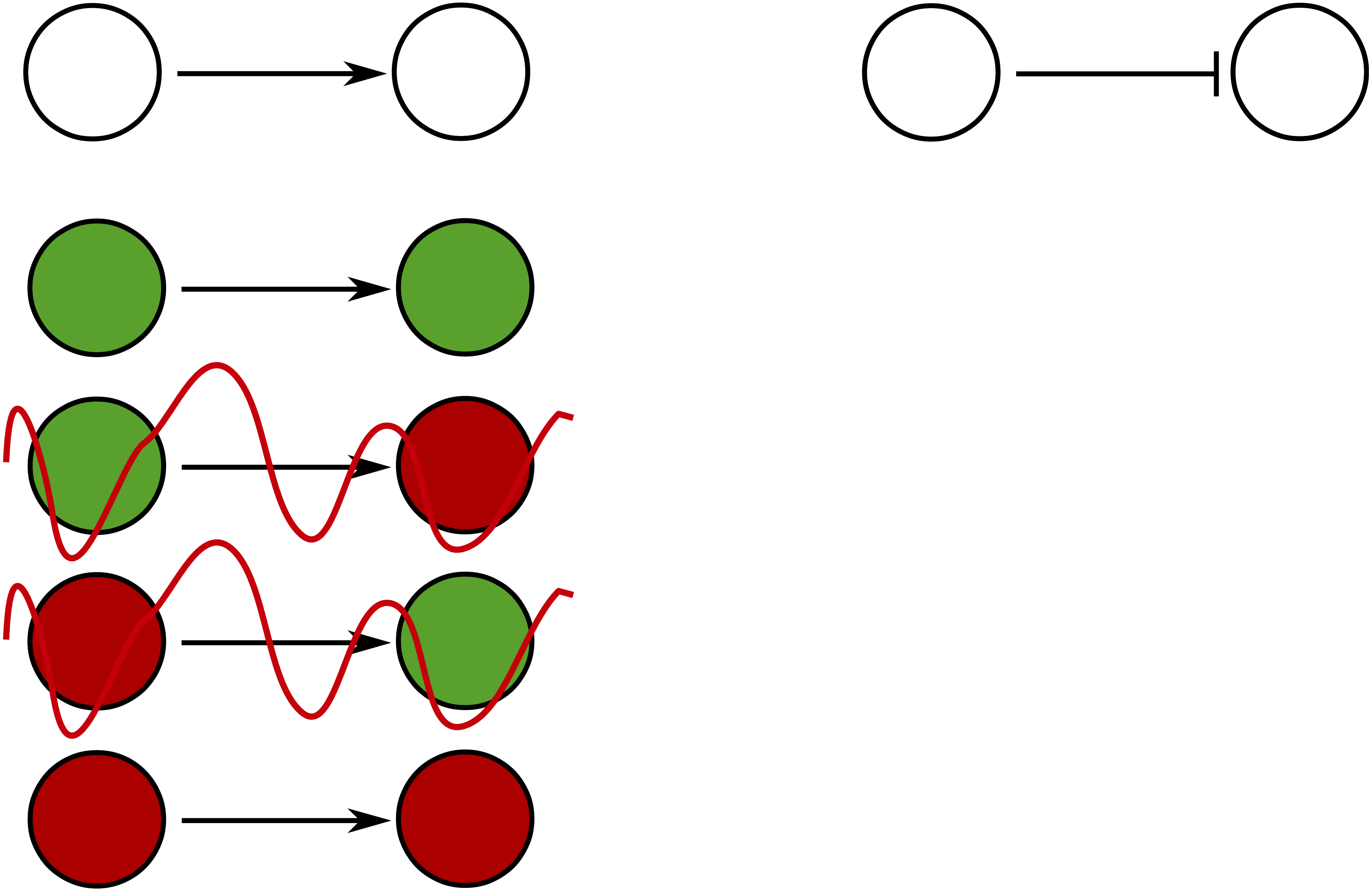
Symbolic ML - example 1



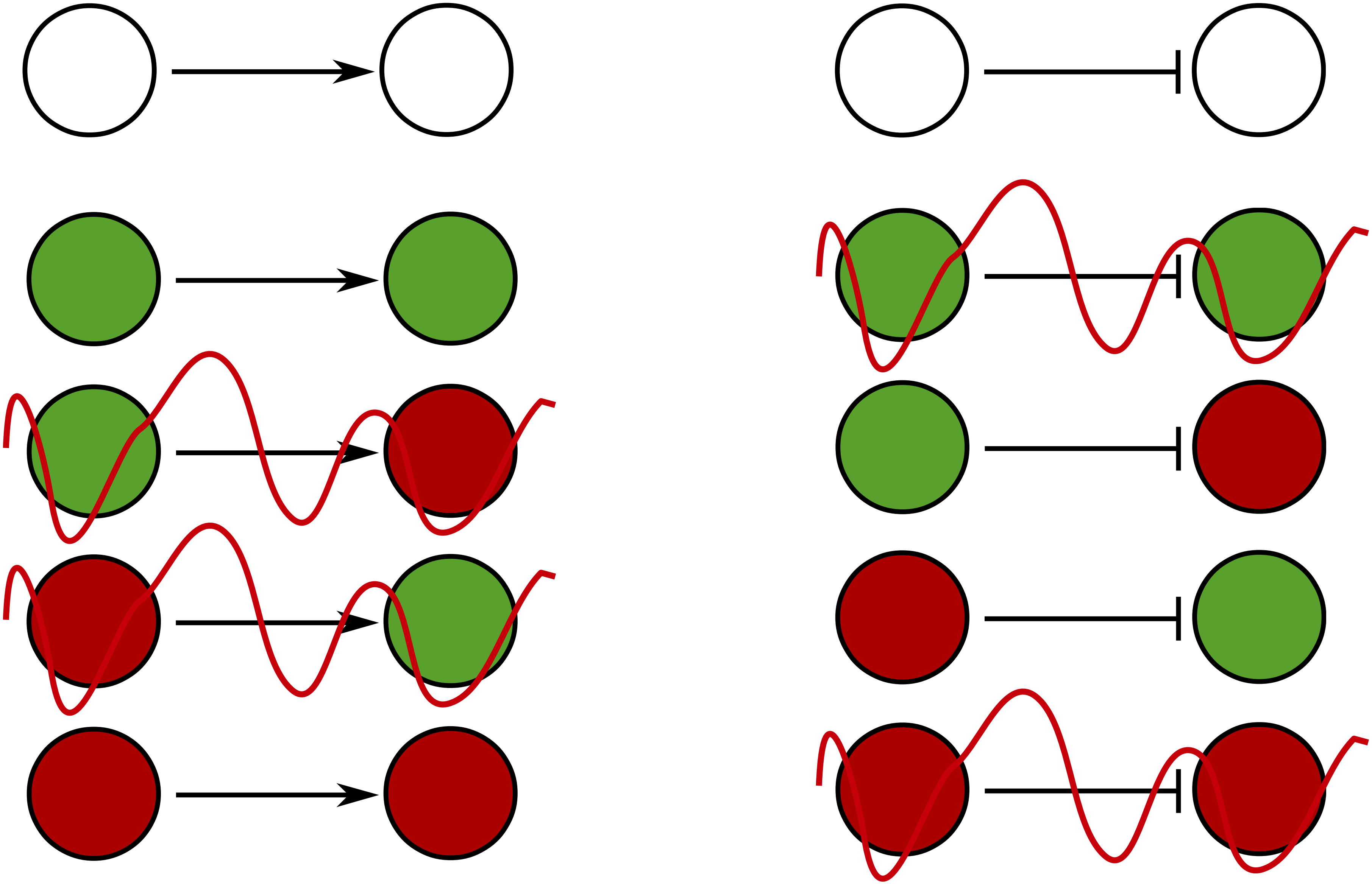
Symbolic ML - example 1



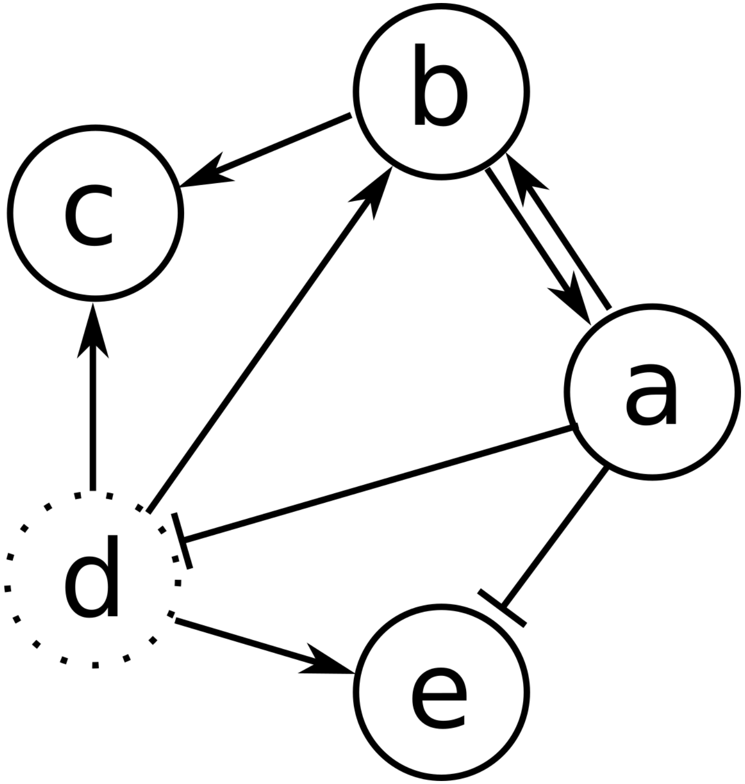
Symbolic ML - example 1



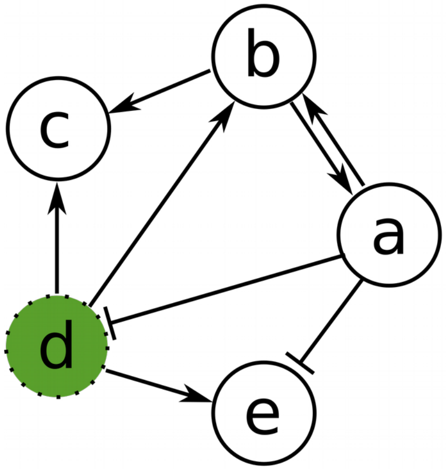
Symbolic ML - example 1



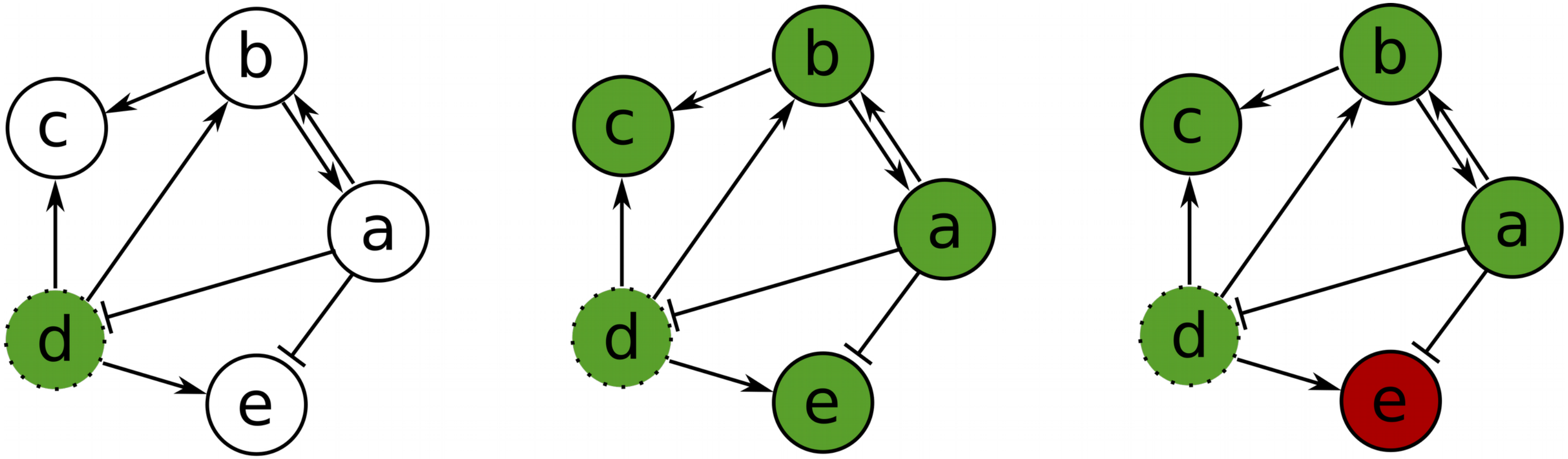
Symbolic ML - example 1



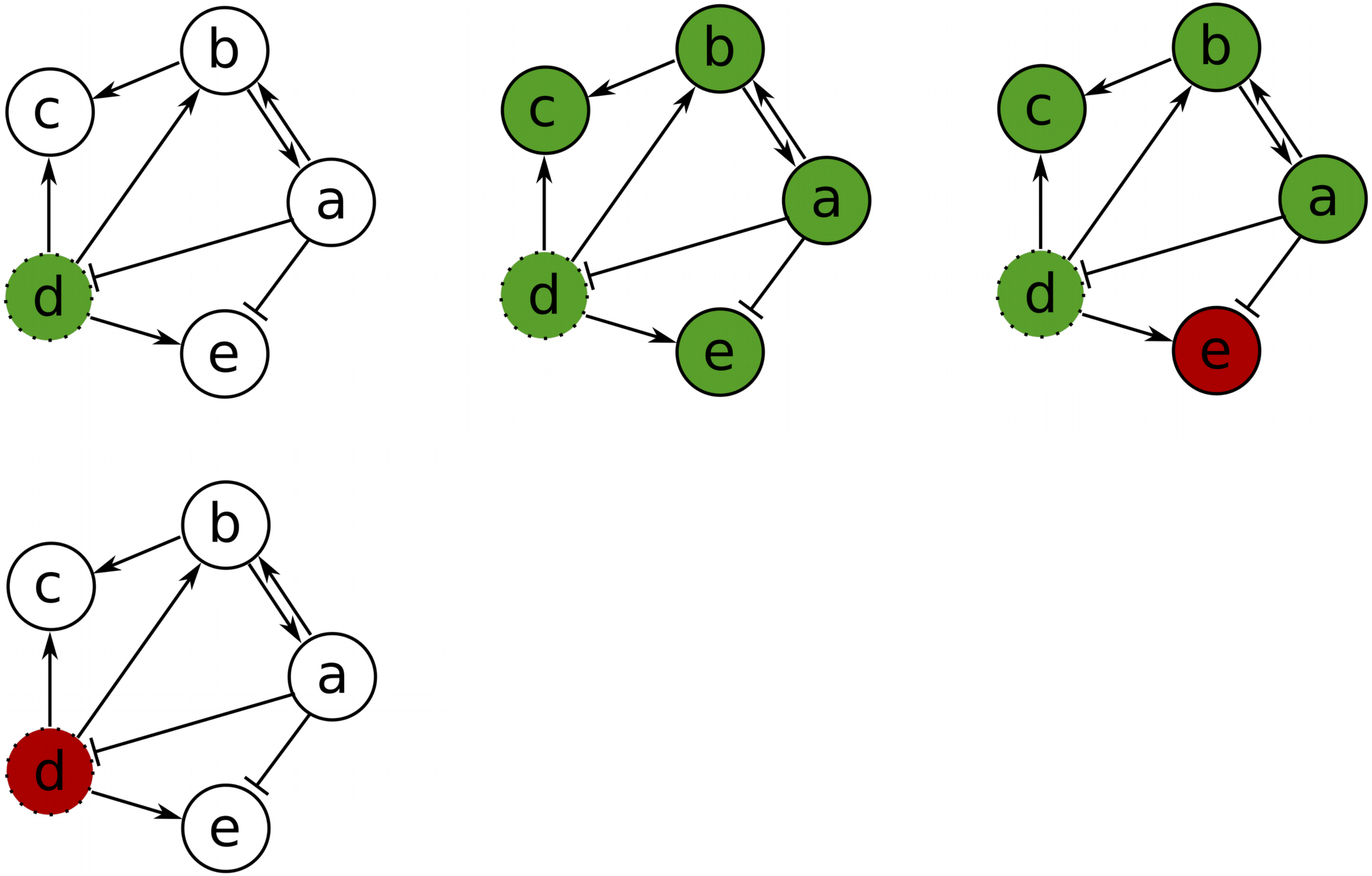
Symbolic ML - example 1



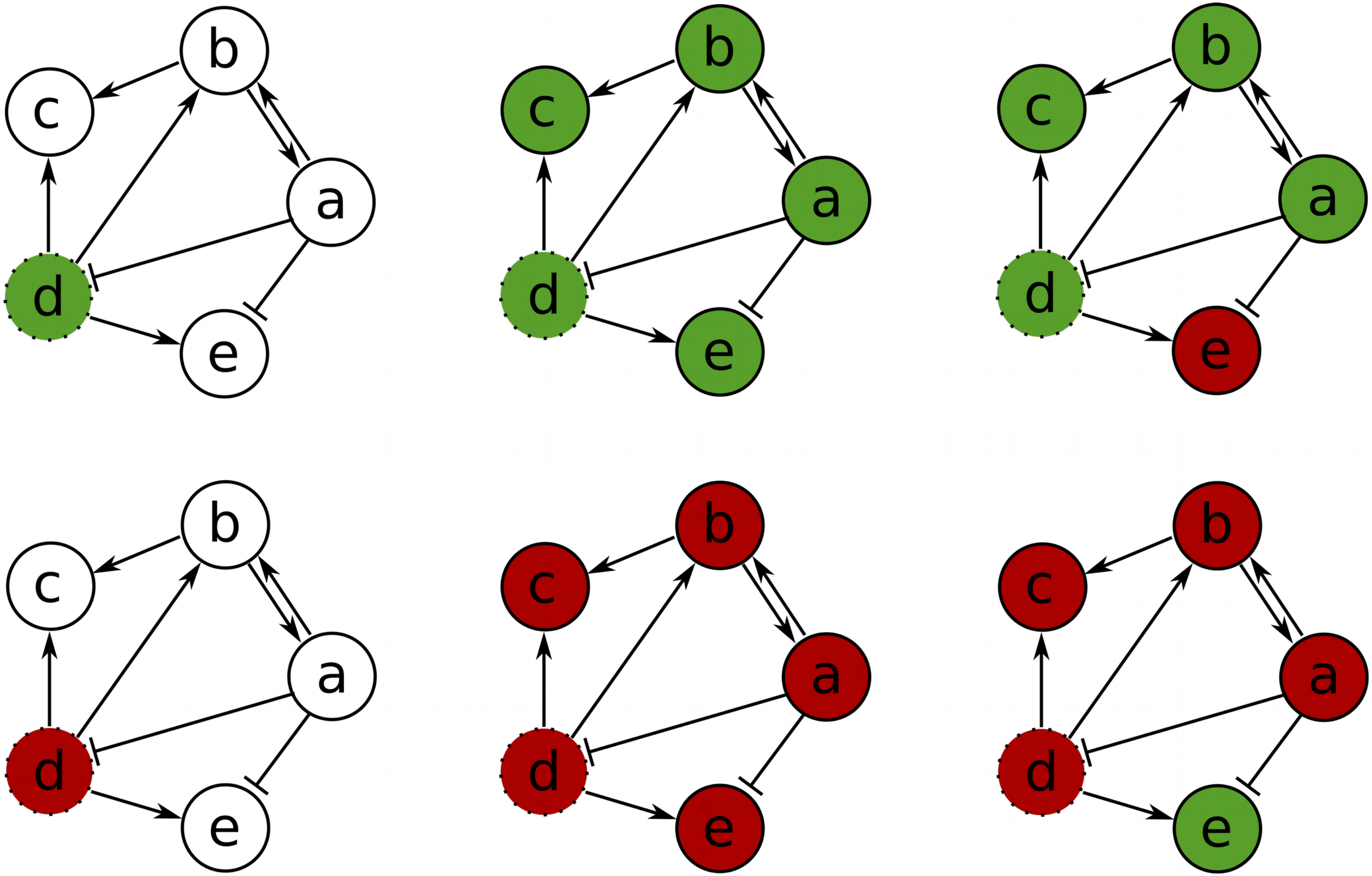
Symbolic ML - example 1



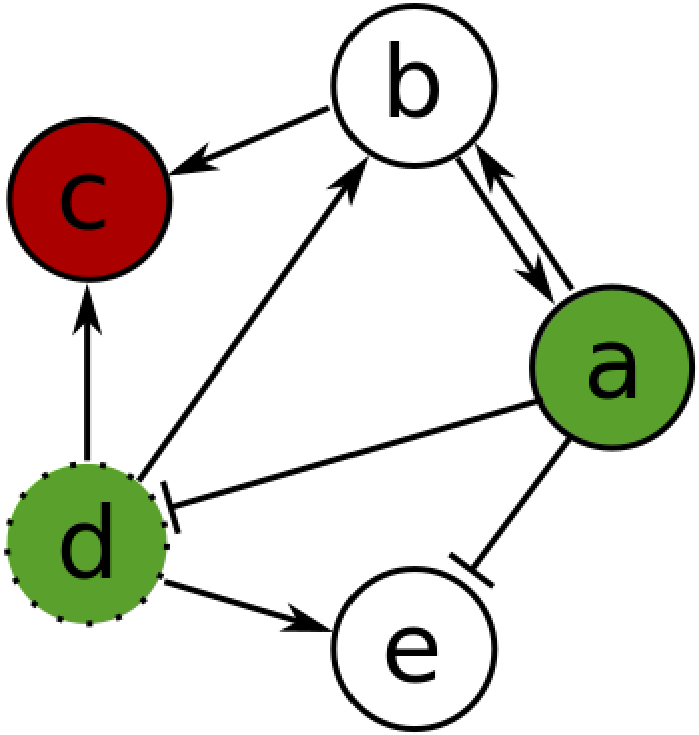
Symbolic ML - example 1



Symbolic ML - example 1



Symbolic ML - example 1



Gebser et al. (2010), Repair and Prediction (Under Inconsistency) in Large Biological Networks with Answer Set Programming

Symbolic ML - example 1

- Gebser et al. “Repair and Prediction (Under Inconsistency) in Large Biological Networks with Answer Set Programming”

Symbolic ML - example 1

- Gebser et al. “Repair and Prediction (Under Inconsistency) in Large Biological Networks with Answer Set Programming”
- Modeling approach: sign consistency graph (Boolean graph + annotations + consistency criteria)

Symbolic ML - example 1

- Gebser et al. “Repair and Prediction (Under Inconsistency) in Large Biological Networks with Answer Set Programming”
- Modeling approach: sign consistency graph (Boolean graph + annotations + consistency criteria)
- algorithms: Answer Set Programming (Clingo) to automate **consistency checks** and **model revision**

Symbolic ML - example 1

- Gebser et al. “Repair and Prediction (Under Inconsistency) in Large Biological Networks with Answer Set Programming”
- Modeling approach: sign consistency graph (Boolean graph + annotations + consistency criteria)
- algorithms: Answer Set Programming (Clingo) to automate **consistency checks** and **model revision**
- initial model: *E. coli* (RegulonDB, 5150 interactions)

Symbolic ML - example 1

- Gebser et al. “Repair and Prediction (Under Inconsistency) in Large Biological Networks with Answer Set Programming”
- Modeling approach: sign consistency graph (Boolean graph + annotations + consistency criteria)
- algorithms: Answer Set Programming (Clingo) to automate **consistency checks** and **model revision**
- initial model: *E. coli* (RegulonDB, 5150 interactions)
- data: Exponential-Stationary growth shift (Bradley et al. 2007) & Heatshock experiment (Allen et al. 2003)

Symbolic ML - example 1

	data repair (sign)	model repair (interaction sign or input node)
Exp.-Stationary growth	40	42
Heatshock	34	94

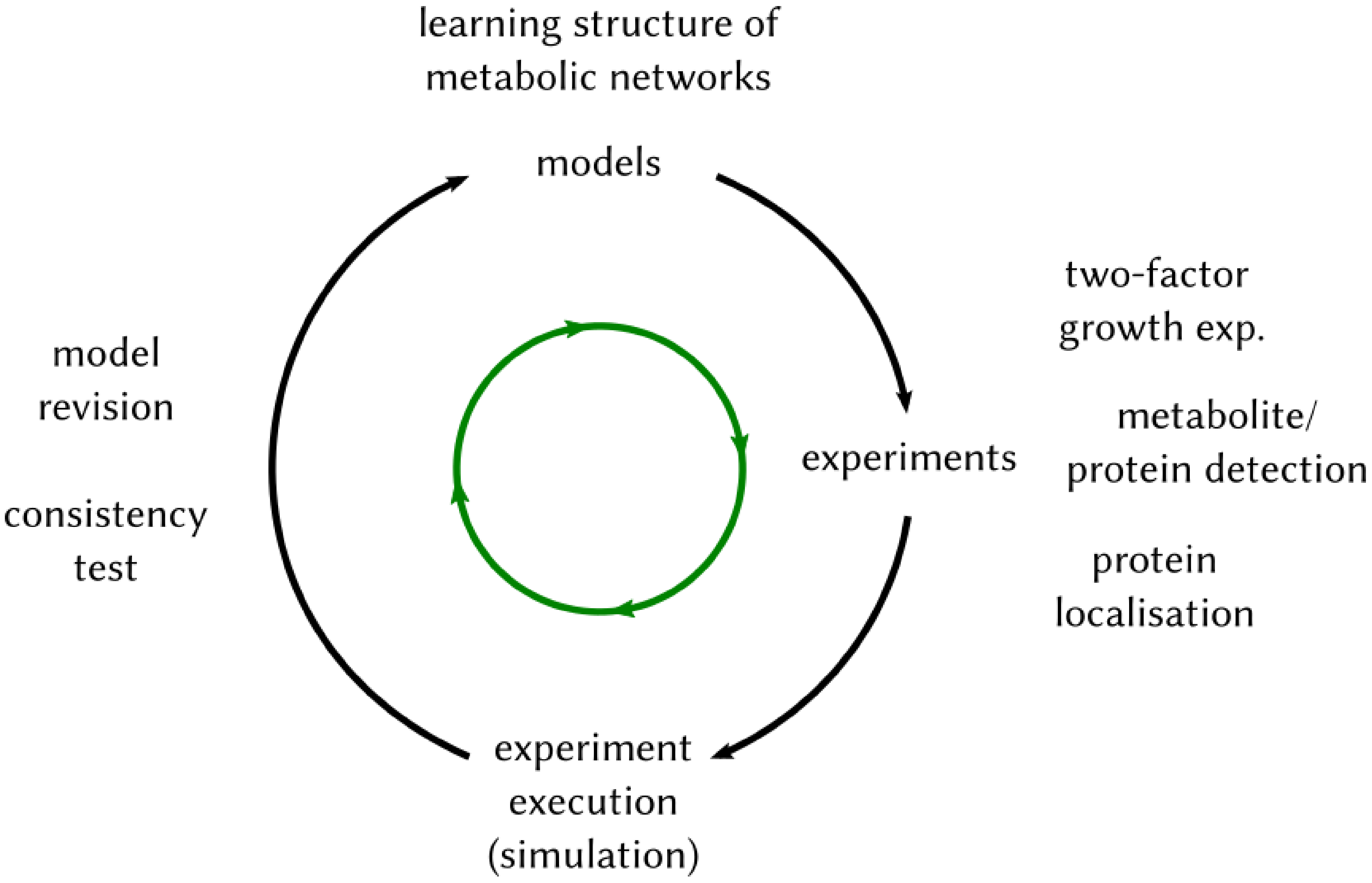
Symbolic ML - example 1

	data repair (sign)	model repair (interaction sign or input node)
Exp.-Stationary growth	40	42
Heatshock	34	94

the method produces repairs of high quality: predictions from minimal repairs for unobserved nodes conform with test data (>90% accuracy rate)

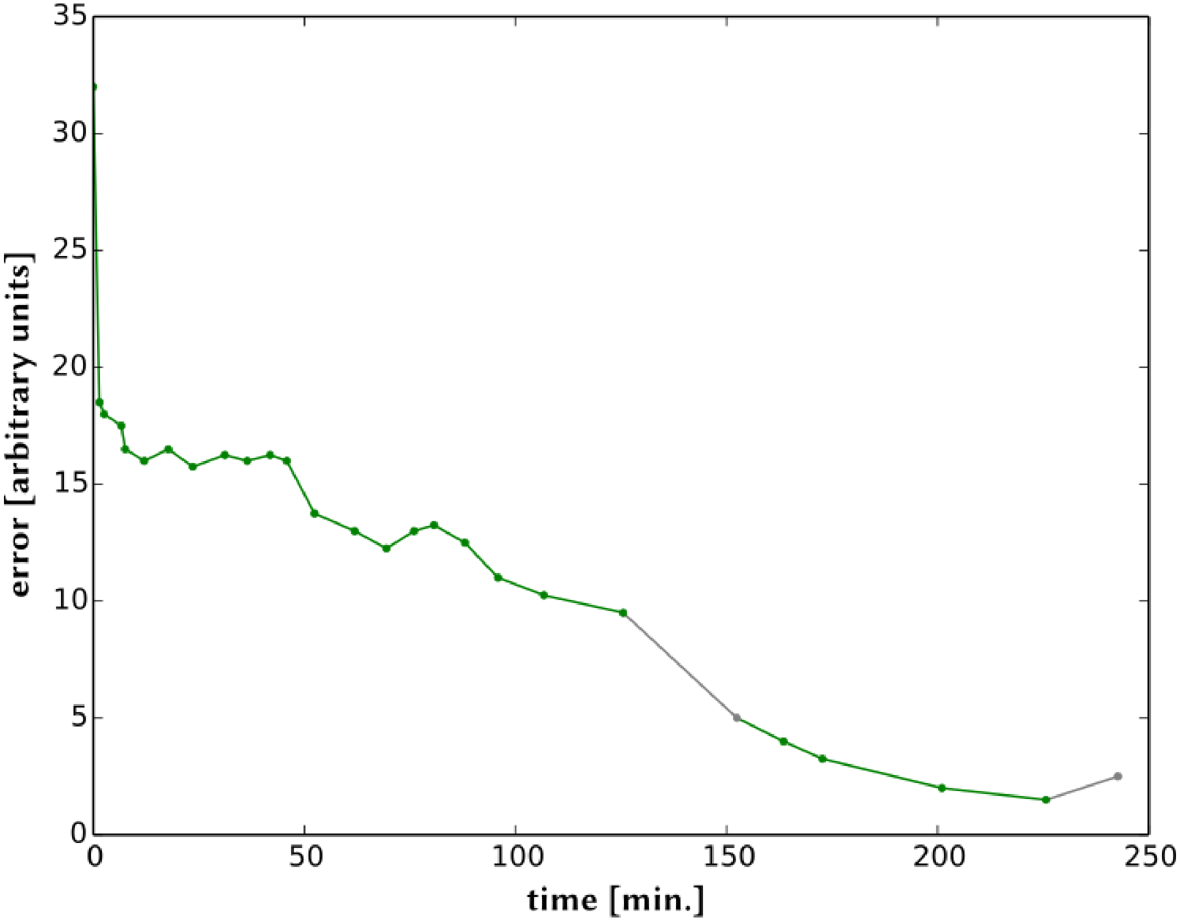
Symbolic ML - example 2

Symbolic ML - example 2



https://www.research.manchester.ac.uk/portal/files/60986384/FULL_TEXT.PDF

Symbolic ML - example 2



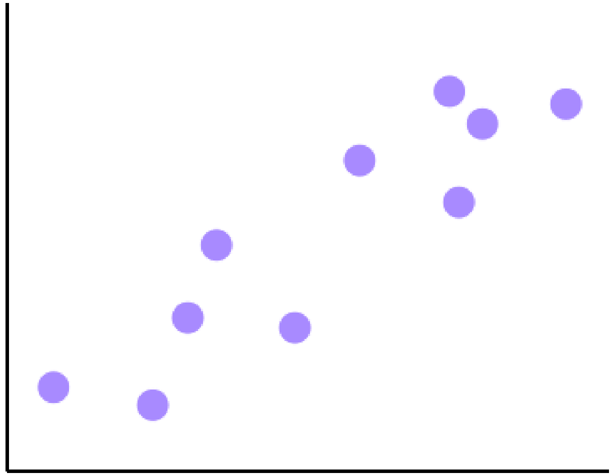
avg. error reduction: 76%

25-86 reactions

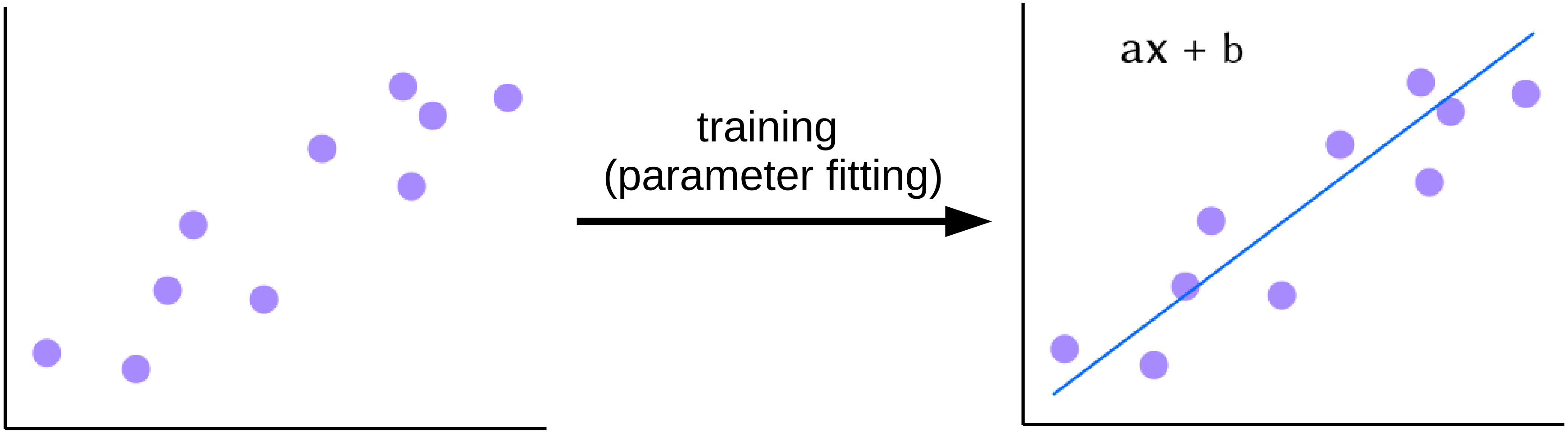
https://www.research.manchester.ac.uk/portal/files/60986384/FULL_TEXT.PDF

Standard Supervised ML

Standard Supervised ML

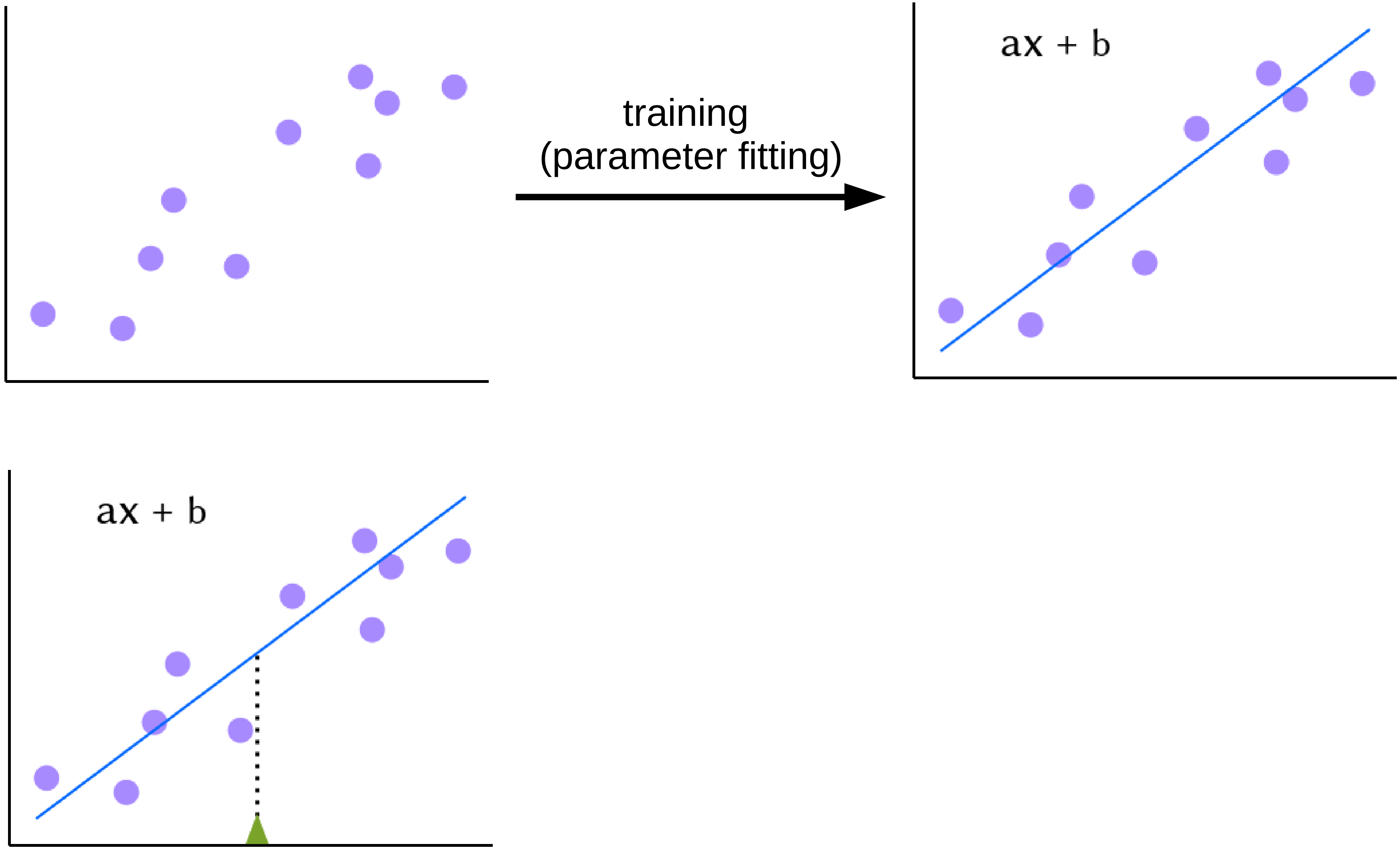


Standard Supervised ML



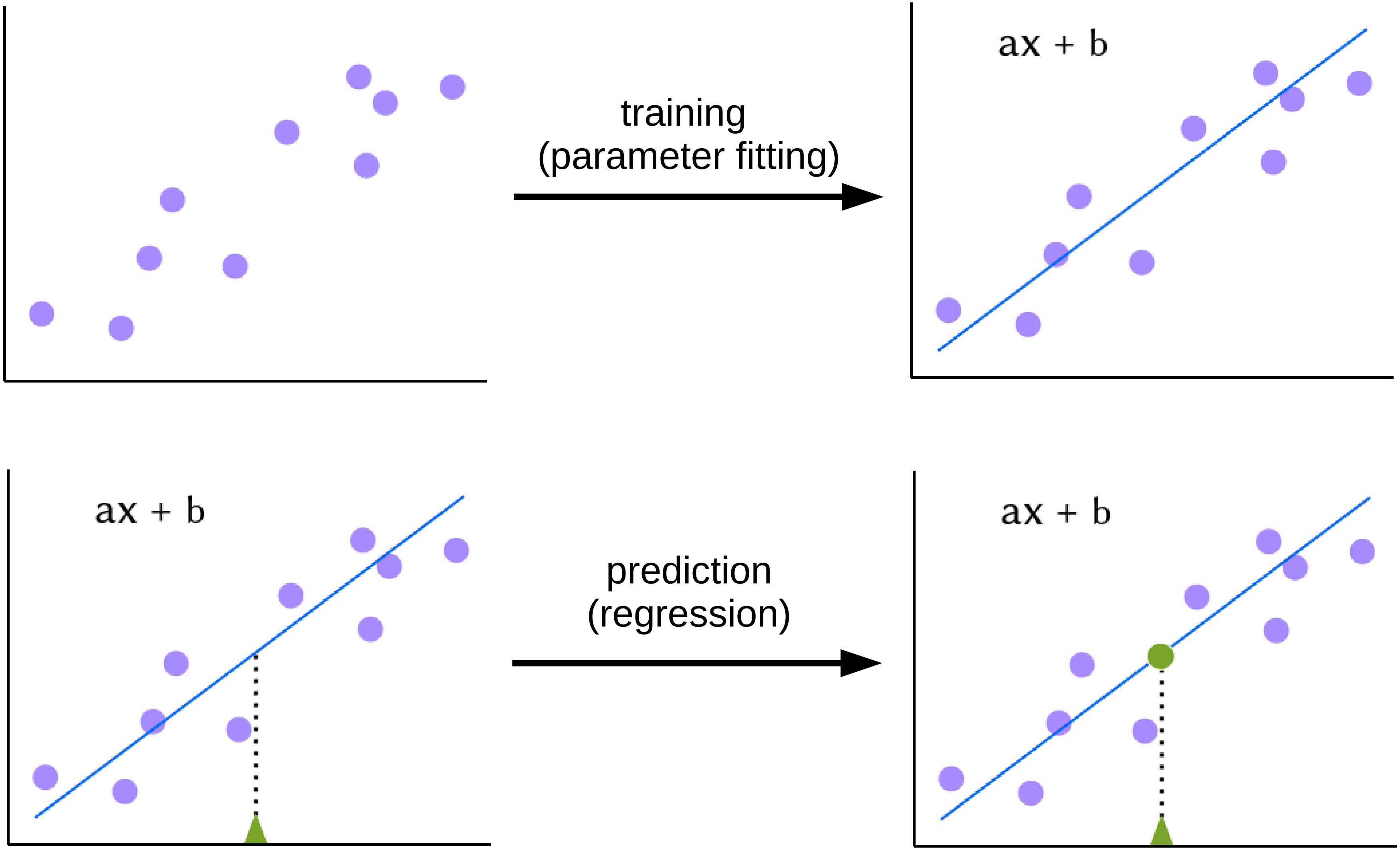
training
(parameter fitting)

Standard Supervised ML



training
(parameter fitting)

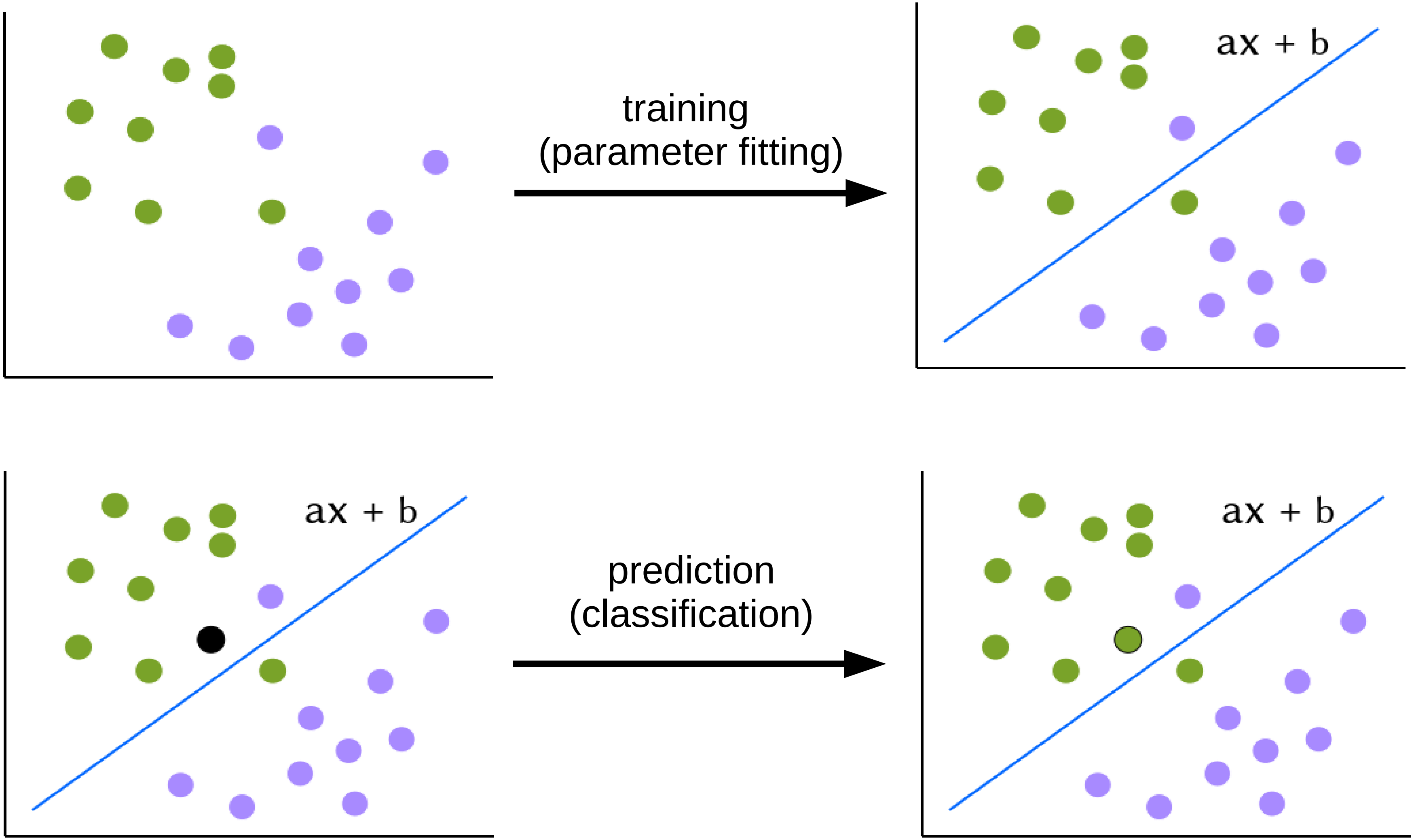
Standard Supervised ML



training
(parameter fitting)

prediction
(regression)

Standard Supervised ML



training
(parameter fitting)

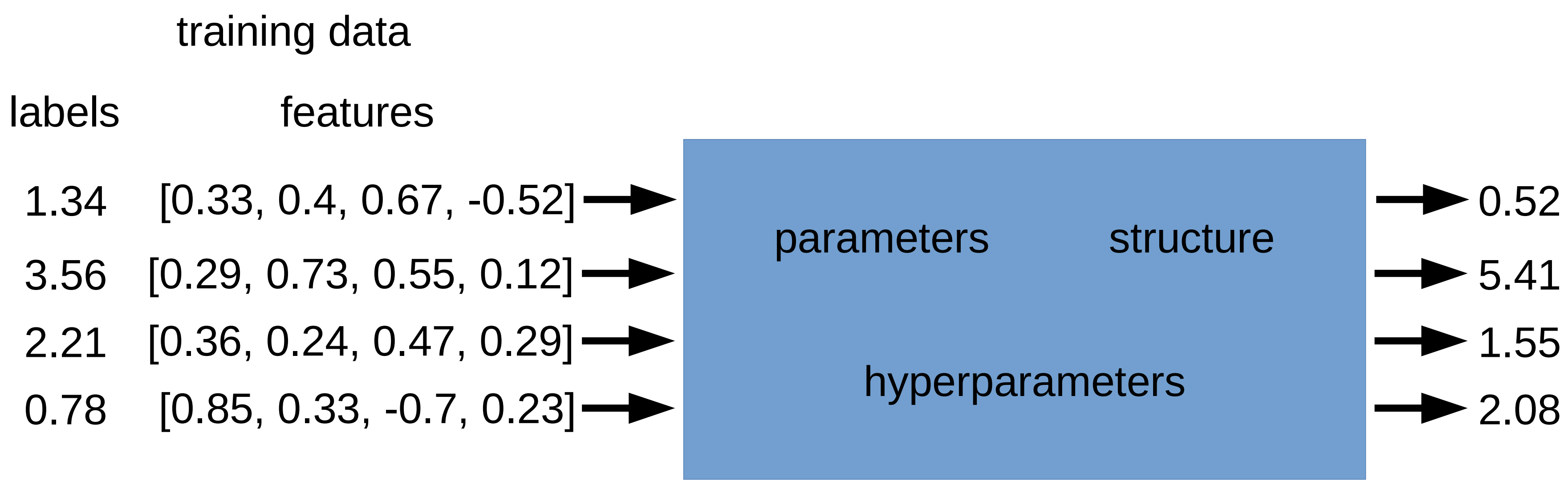
prediction
(classification)

Standard Supervised ML

training data

labels	features
1.34	[0.33, 0.4, 0.67, -0.52]
3.56	[0.29, 0.73, 0.55, 0.12]
2.21	[0.36, 0.24, 0.47, 0.29]
0.78	[0.85, 0.33, -0.7, 0.23]

Standard Supervised ML



training data

labels	features
1.34	[0.33, 0.4, 0.67, -0.52]
3.56	[0.29, 0.73, 0.55, 0.12]
2.21	[0.36, 0.24, 0.47, 0.29]
0.78	[0.85, 0.33, -0.7, 0.23]

parameters

structure

hyperparameters

0.52

5.41

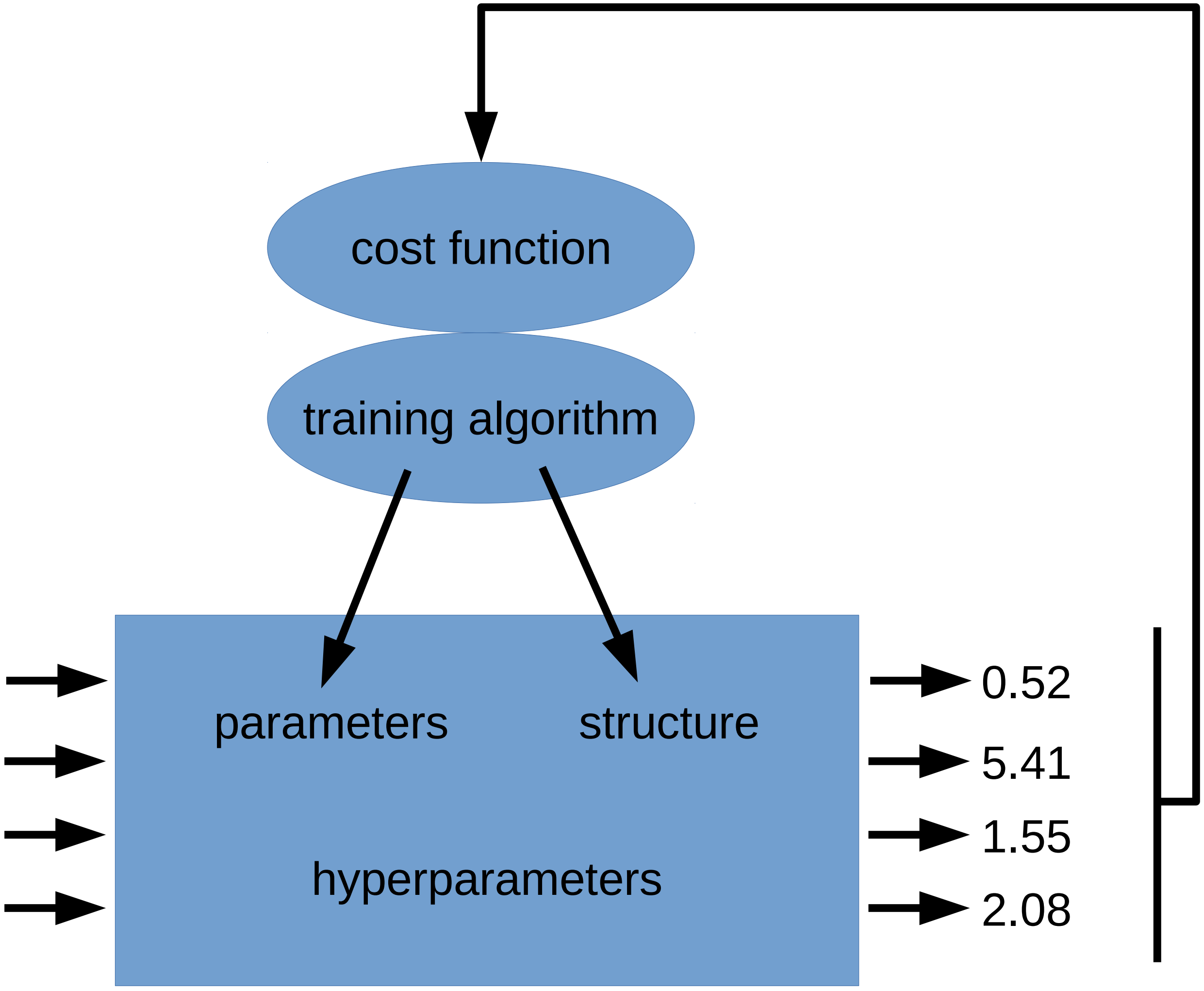
1.55

2.08

Standard Supervised ML

training data

labels	features
1.34	[0.33, 0.4, 0.67, -0.52]
3.56	[0.29, 0.73, 0.55, 0.12]
2.21	[0.36, 0.24, 0.47, 0.29]
0.78	[0.85, 0.33, -0.7, 0.23]



cost function

training algorithm

parameters

structure

hyperparameters

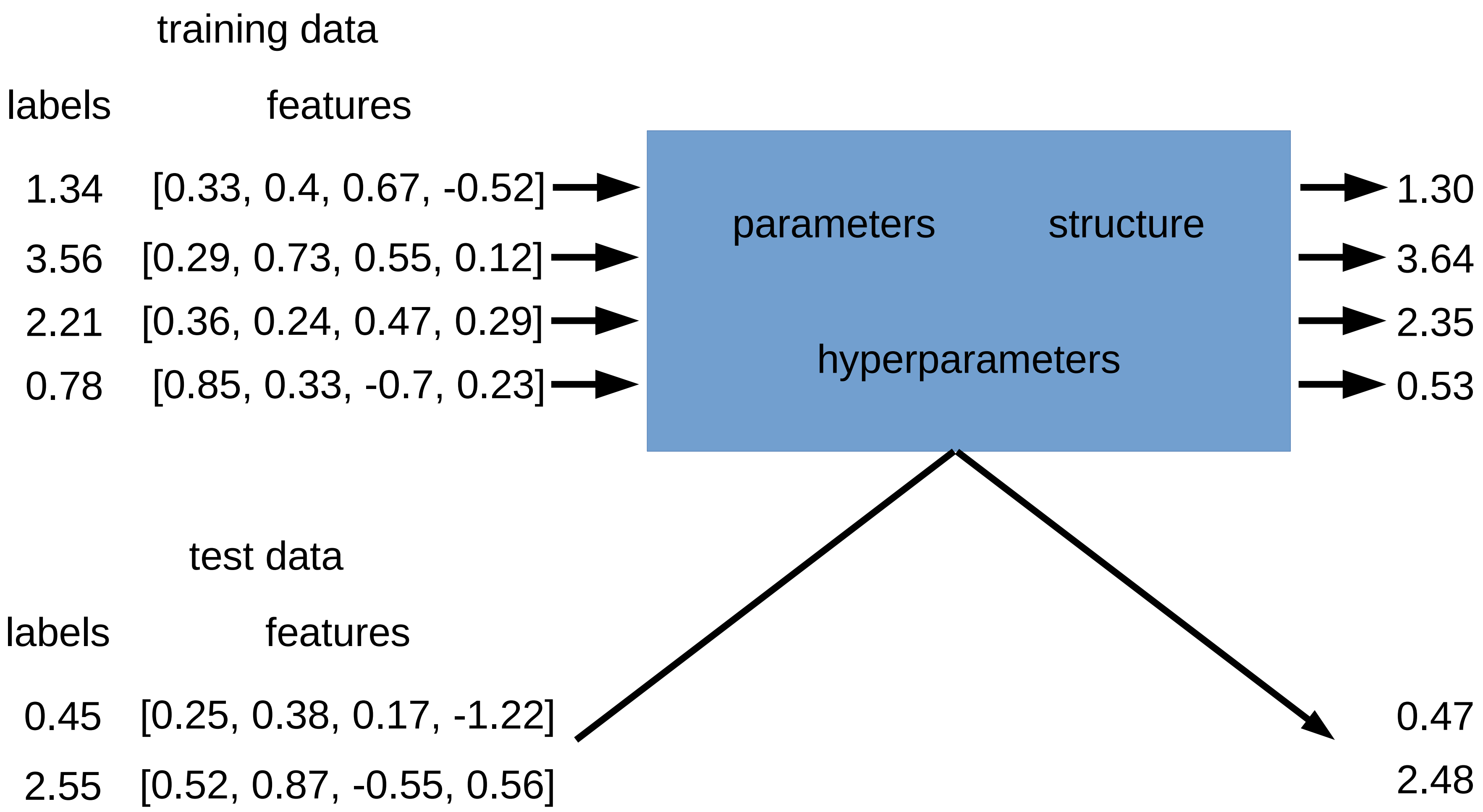
0.52

5.41

1.55

2.08

Standard Supervised ML



training data

labels	features
1.34	[0.33, 0.4, 0.67, -0.52]
3.56	[0.29, 0.73, 0.55, 0.12]
2.21	[0.36, 0.24, 0.47, 0.29]
0.78	[0.85, 0.33, -0.7, 0.23]

parameters

structure

hyperparameters

1.30

3.64

2.35

0.53

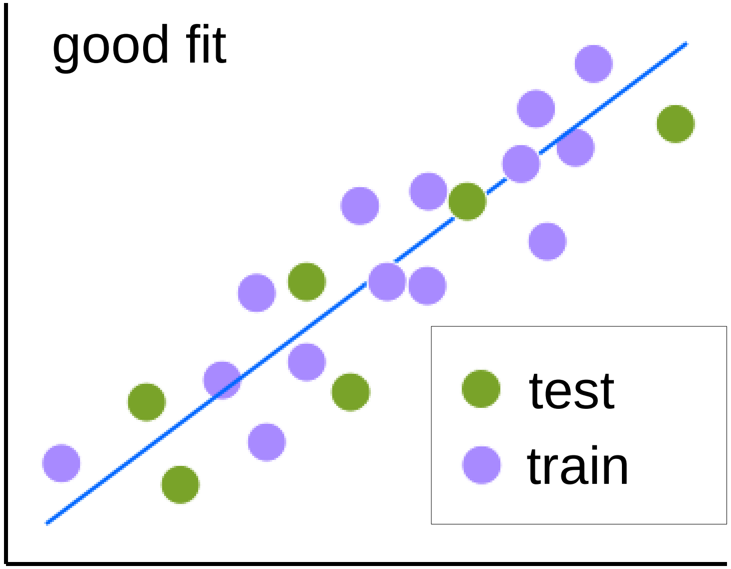
test data

labels	features
0.45	[0.25, 0.38, 0.17, -1.22]
2.55	[0.52, 0.87, -0.55, 0.56]

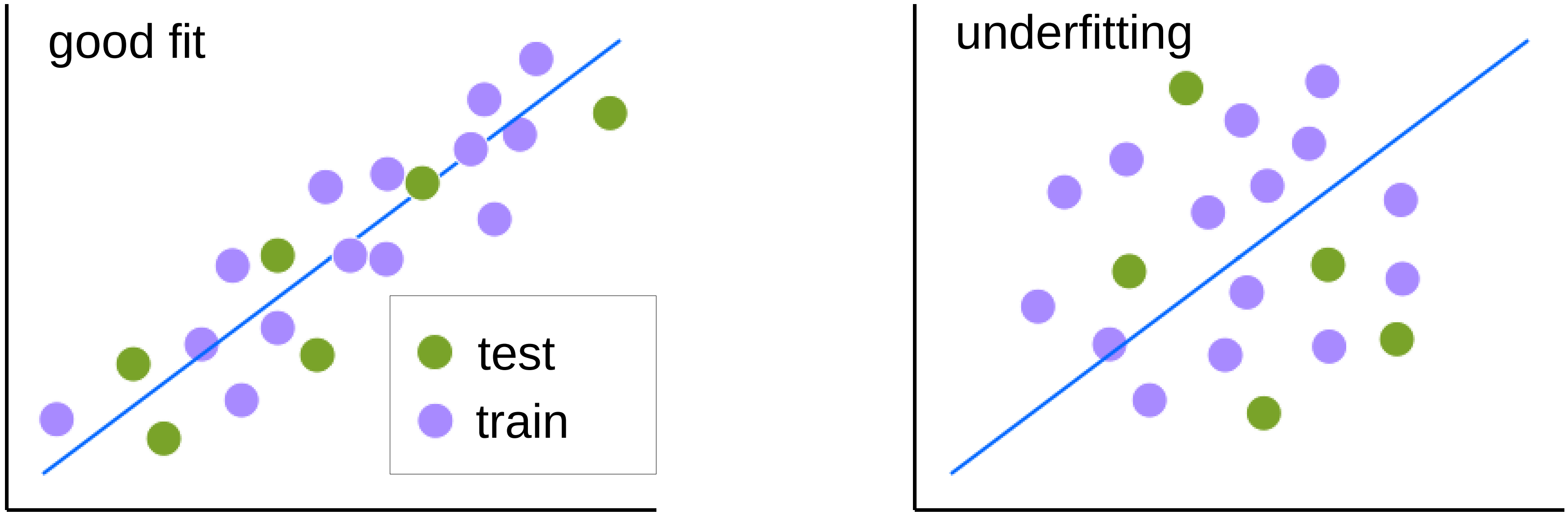
0.47

2.48

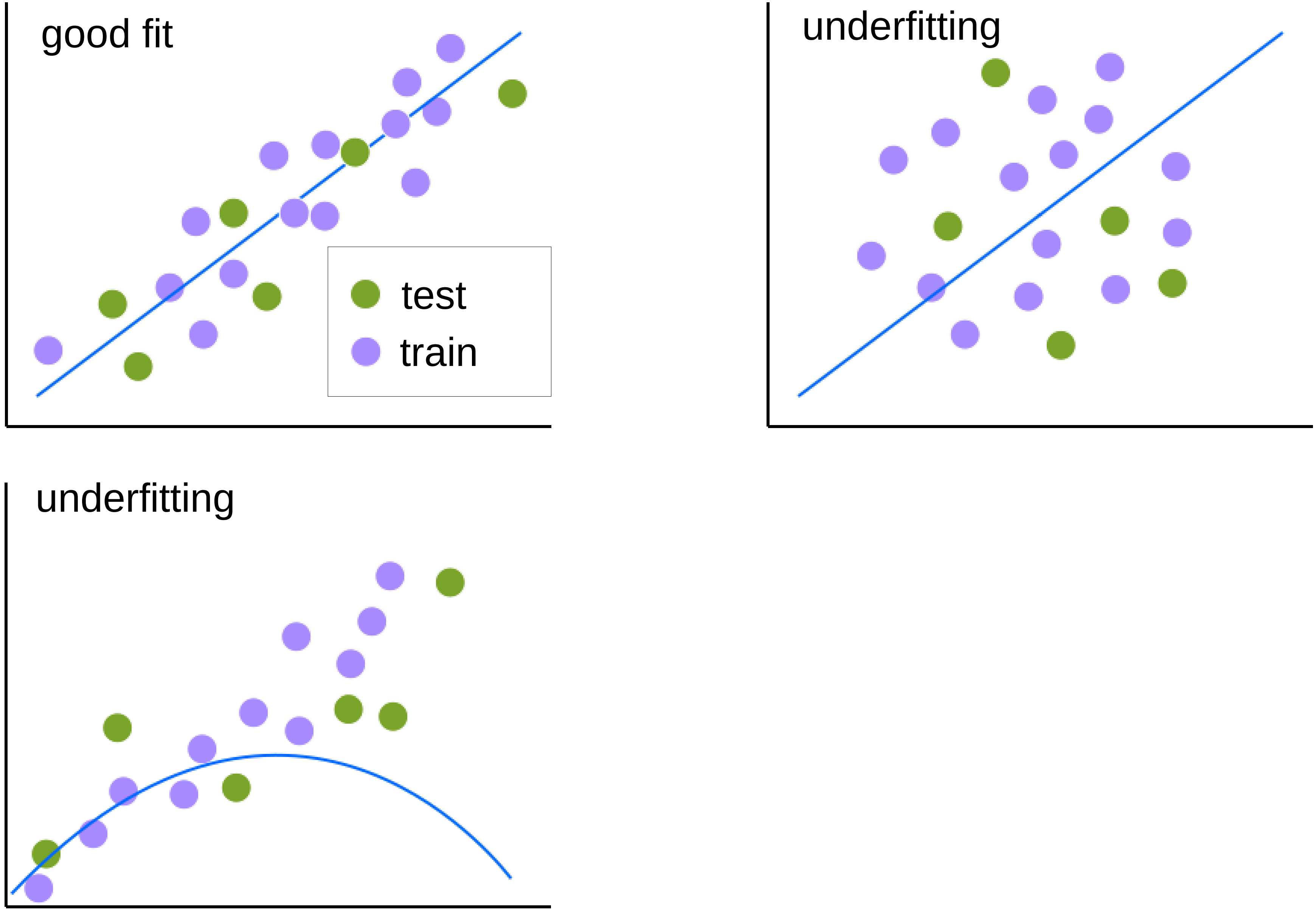
Standard Supervised ML



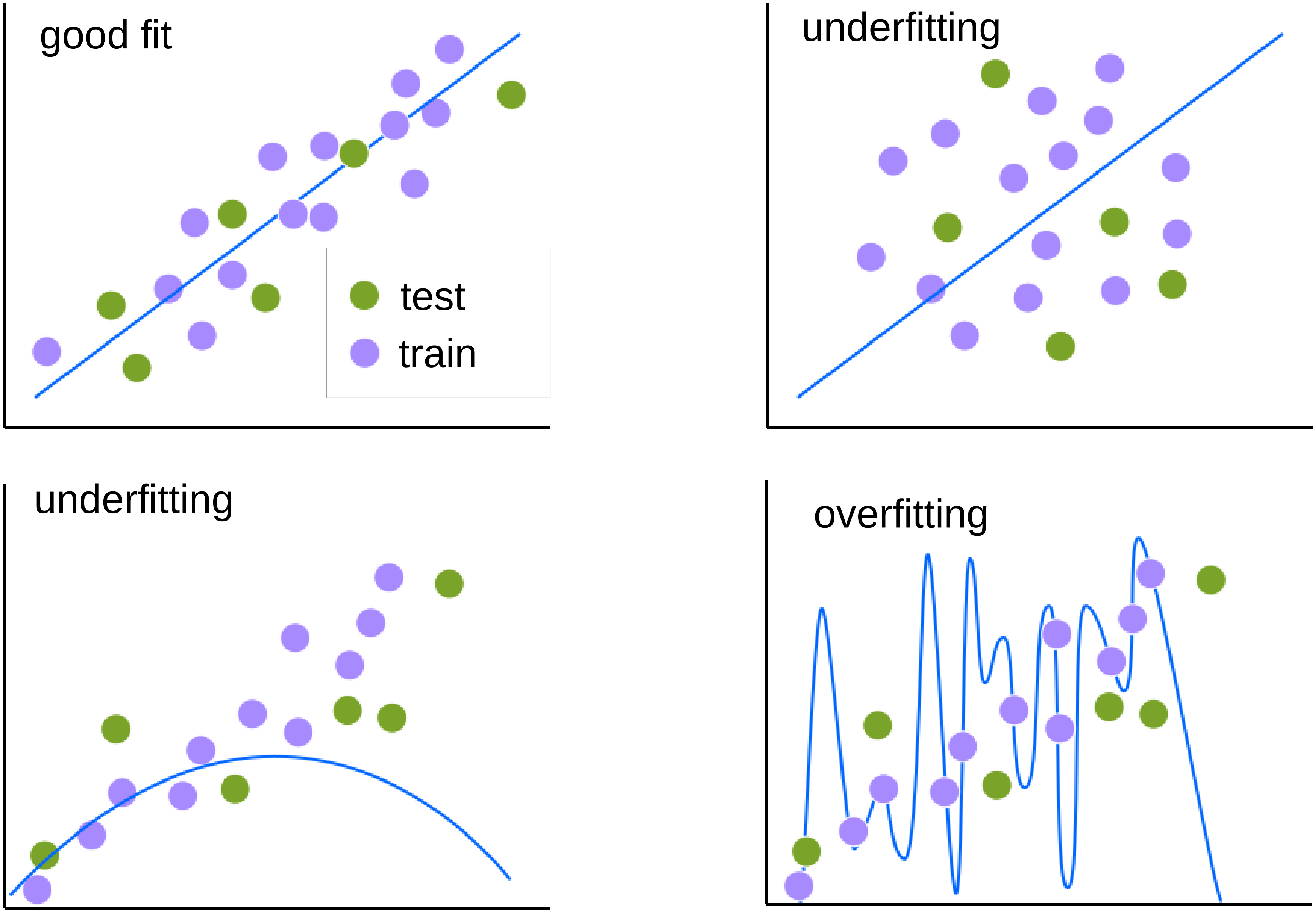
Standard Supervised ML



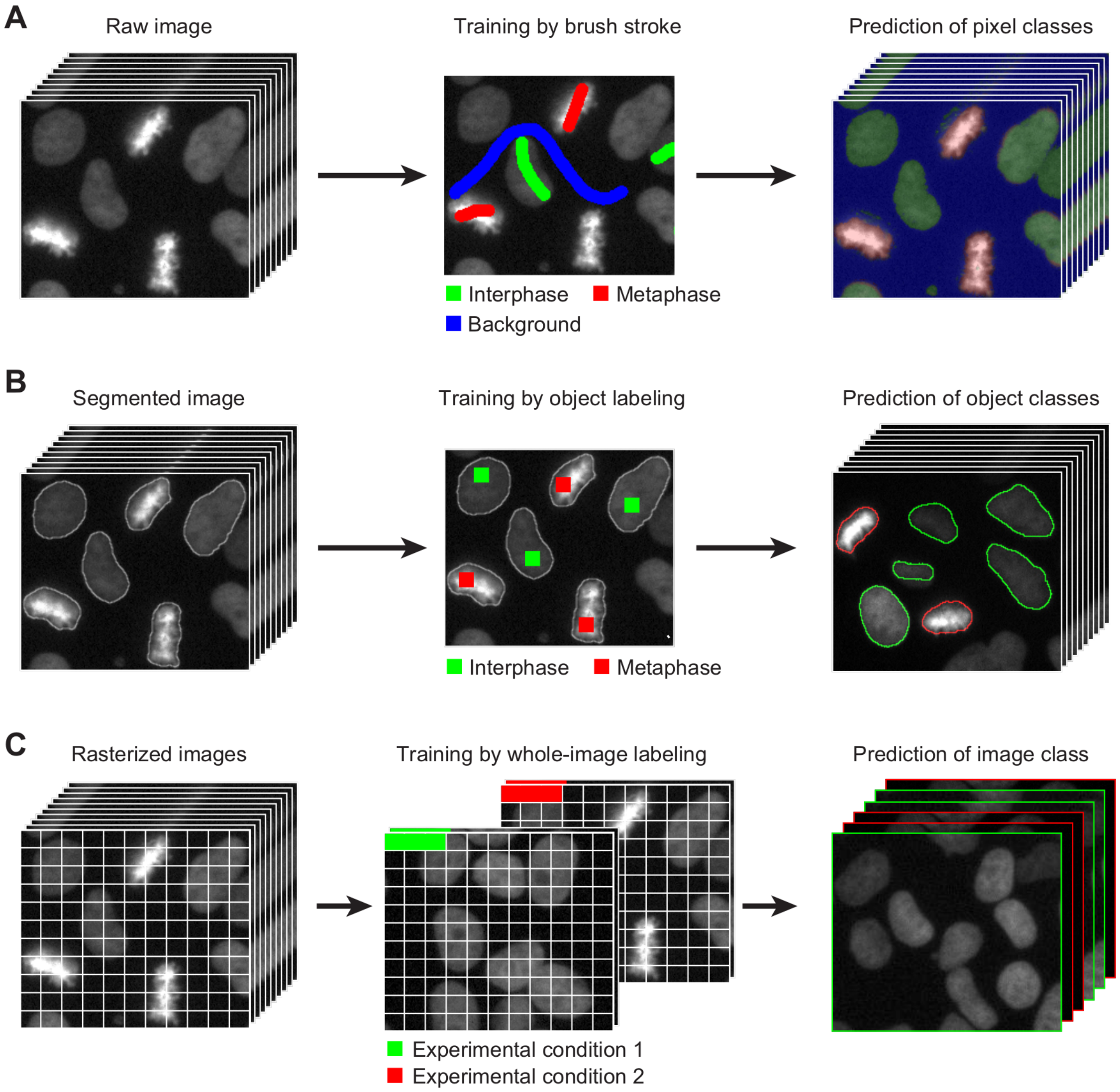
Standard Supervised ML



Standard Supervised ML



Standard Supervised ML: example 1


A
Raw image
Training by brush stroke
Prediction of pixel classes
🟩 Interphase
🟥 Metaphase
🟦 Background
B
Segmented image
Training by object labeling
Prediction of object classes
🟩 Interphase
🟥 Metaphase
C
Rasterized images
Training by whole-image labeling
Prediction of image class
🟩 Experimental condition 1
🟥 Experimental condition 2

Sommer et al. Machine learning in cell biology – teaching computers to recognize phenotypes (2013)

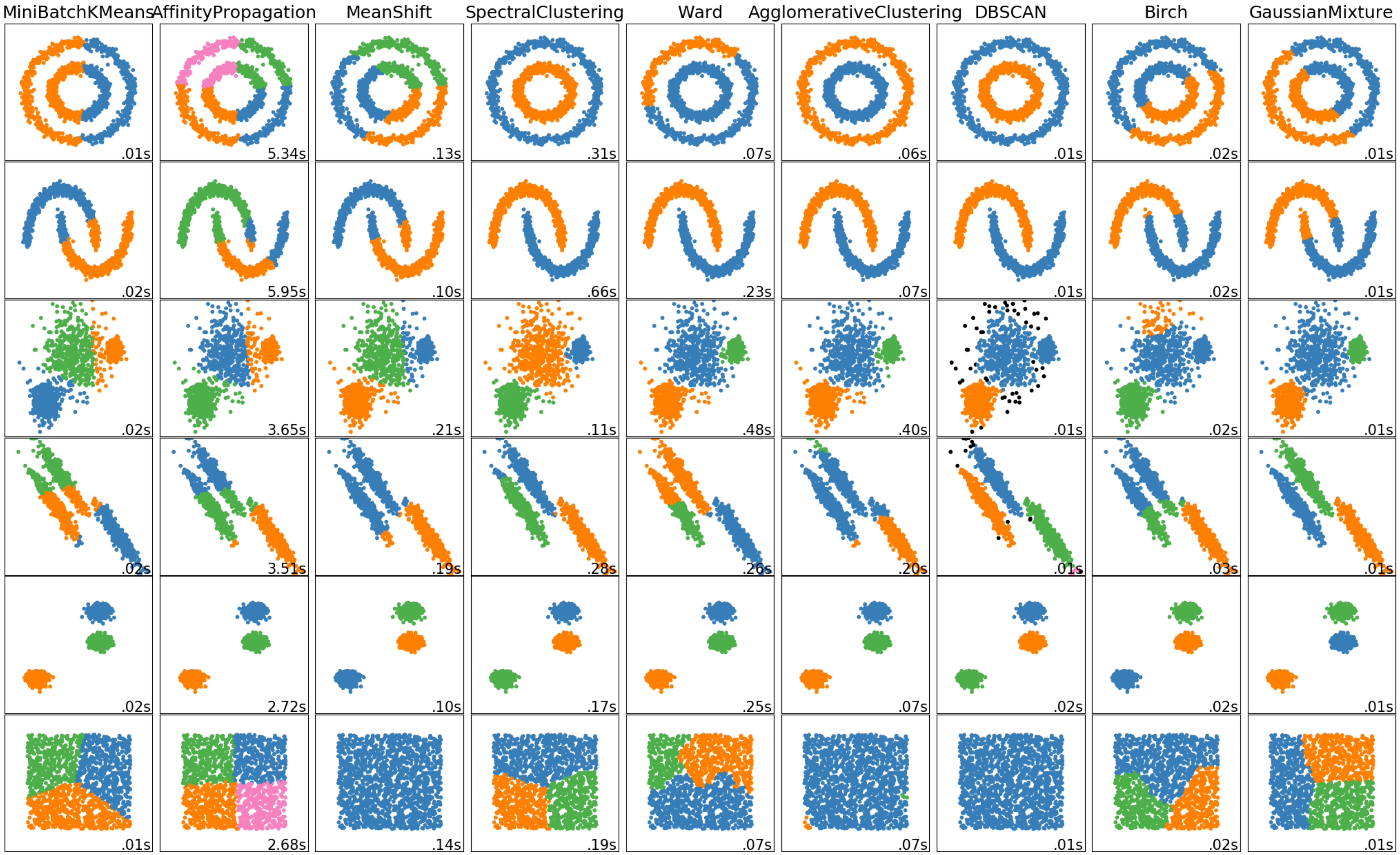
Standard Supervised ML: example 2

- Xu et al. A Gene Signature for Breast Cancer Prognosis Using Support Vector Machine (2012)
- 50 gene signature (microarray gene expression) used to predict metastasis using SVM (accuracy 0.97, sensitivity 0.99, specificity 0.93)
- improvement over 70 gene signature (Recursive Feature Elimination)

Standard Unsupervised ML

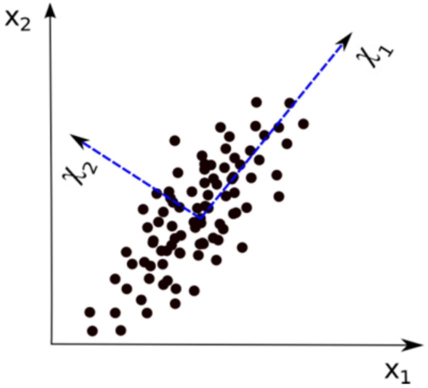
https://scikit-learn.org/stable/modules/clustering.html

Standard Unsupervised ML: clustering



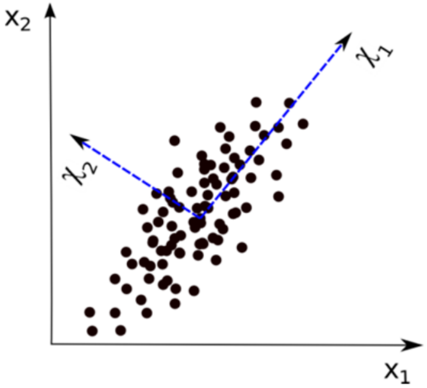
https://scikit-learn.org/stable/modules/clustering.html

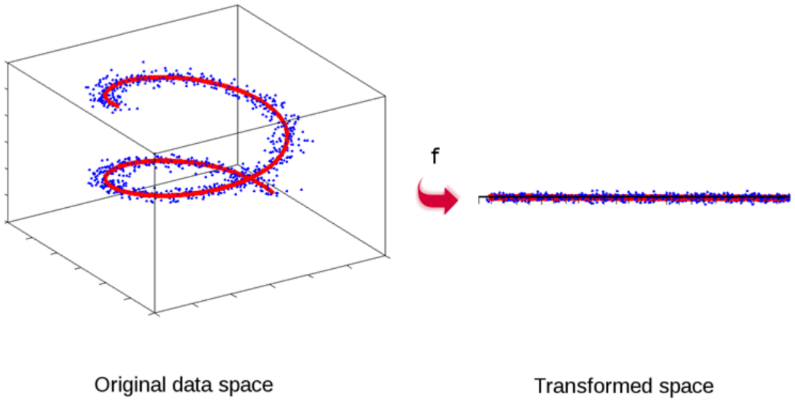
Standard Unsupervised ML: dimensionality reduction



Sorzano et al. “A survey of dimensionality reduction techniques”

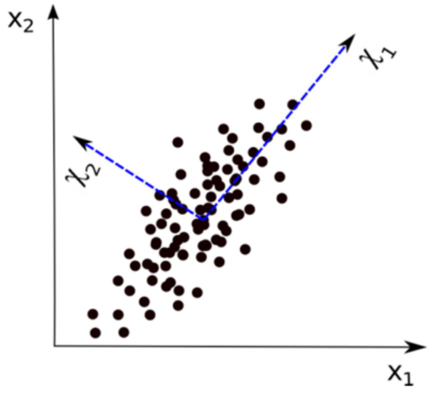
Standard Unsupervised ML: dimensionality reduction

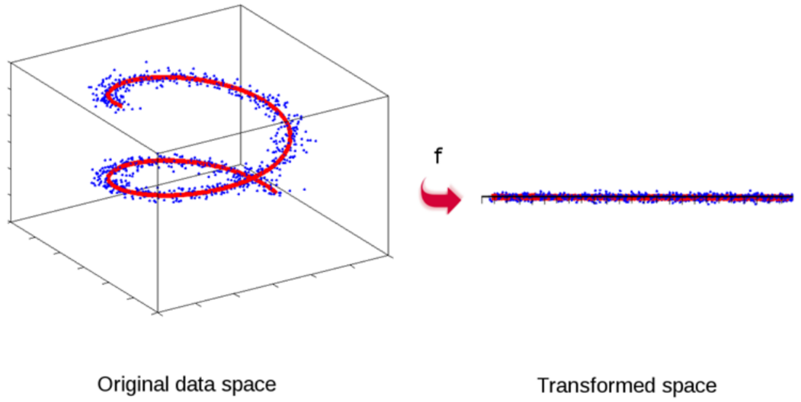




Sorzano et al. “A survey of dimensionality reduction techniques”

Standard Unsupervised ML: dimensionality reduction



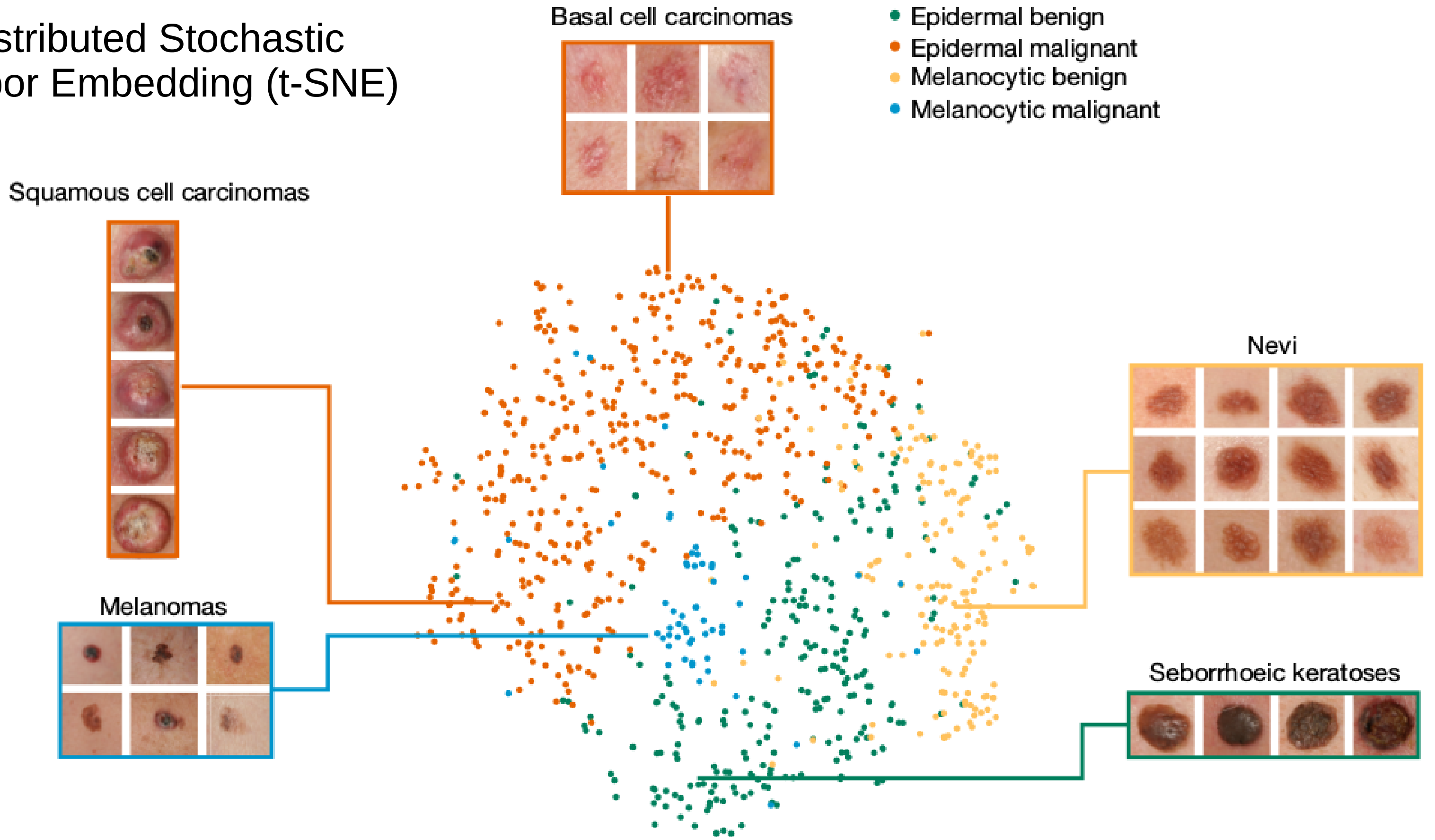


- preparation of data for further classification, regression, etc.
- visualization and analysis
- generative models

Sorzano et al. “A survey of dimensionality reduction techniques”

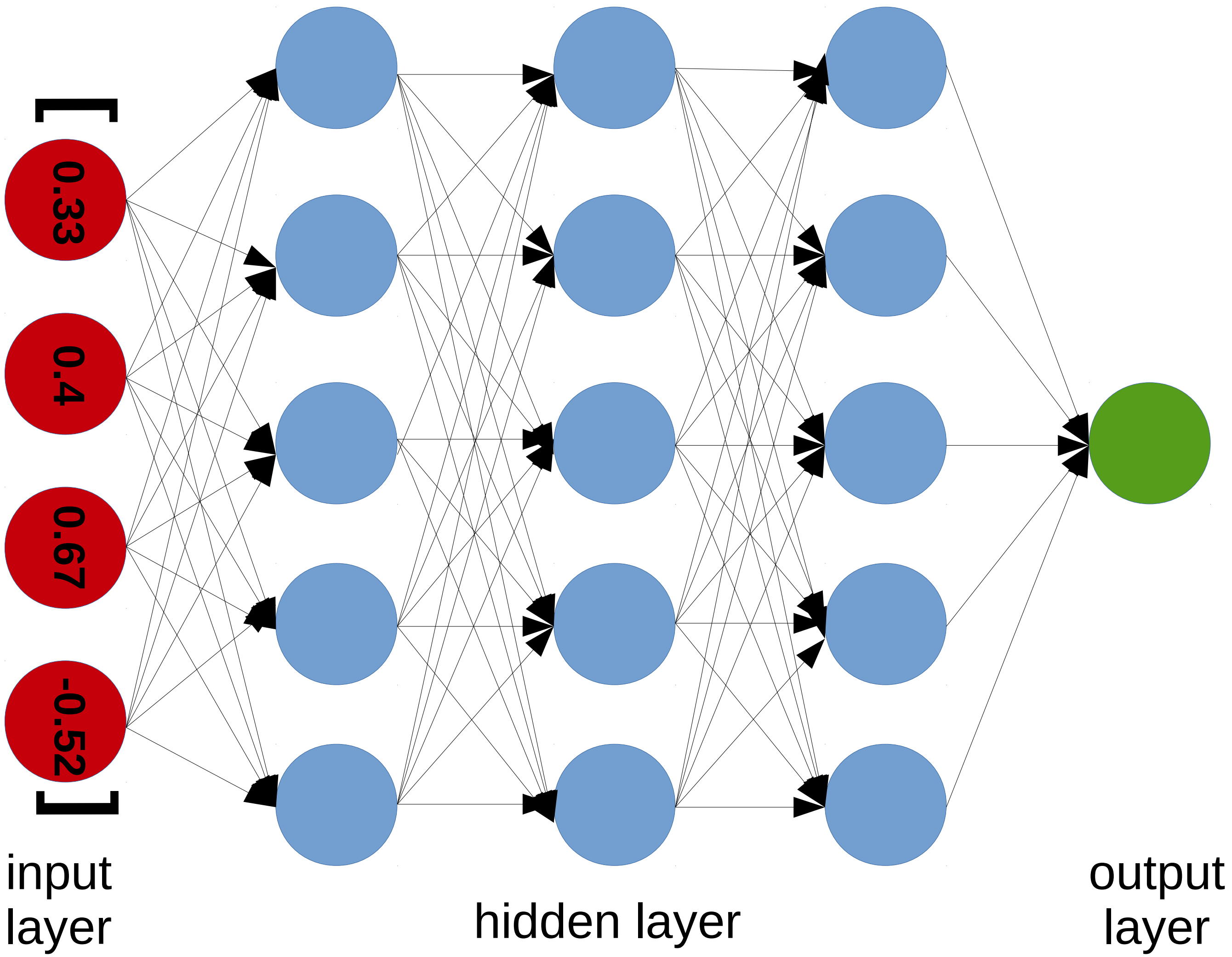
Standard Unsupervised ML: dimensionality reduction

t-Distributed Stochastic Neighbor Embedding (t-SNE)

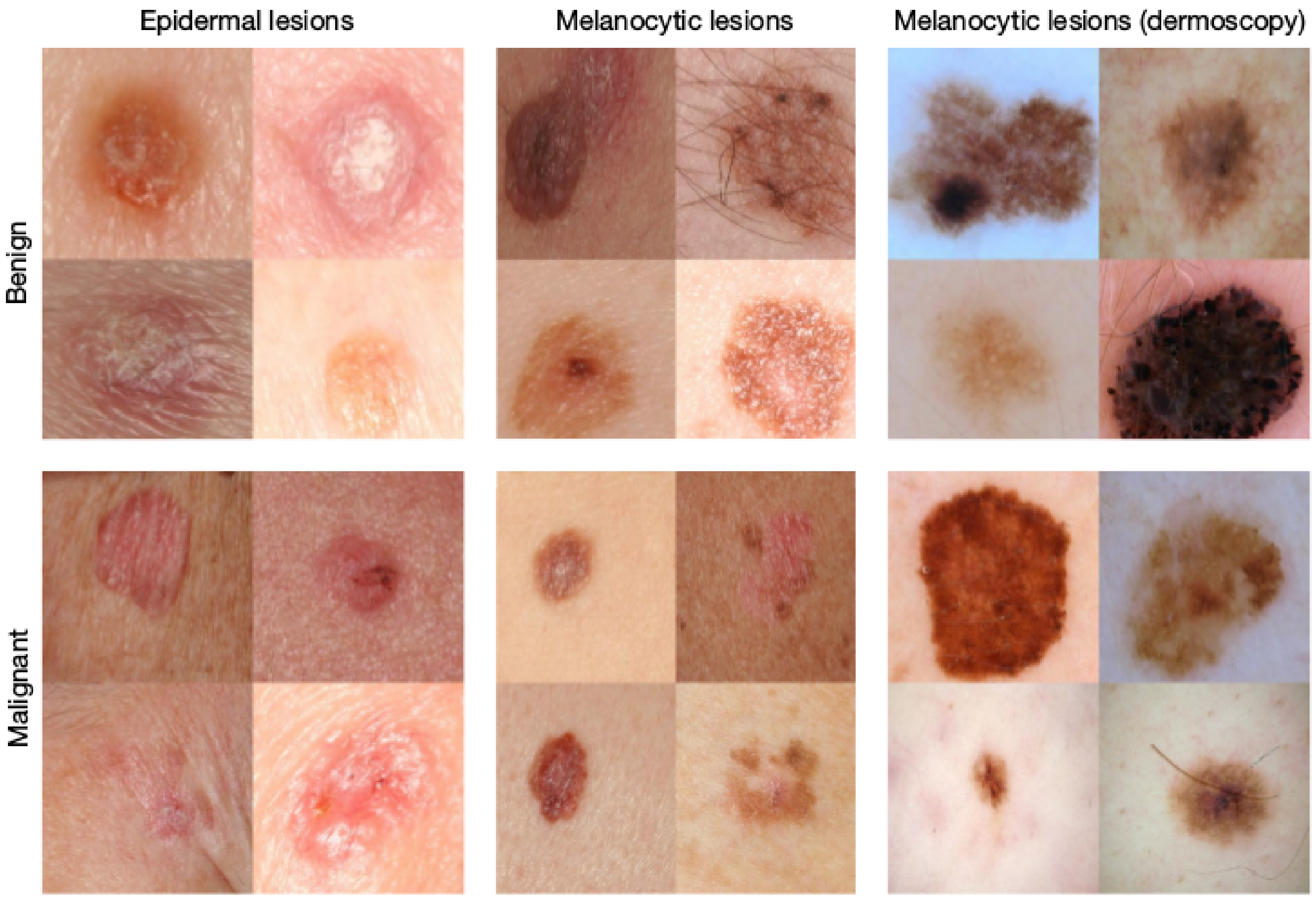


Esteva et al. Dermatologist-level classification of skin cancer with deep neural networks (2017)

Deep Learning (Neural Networks)

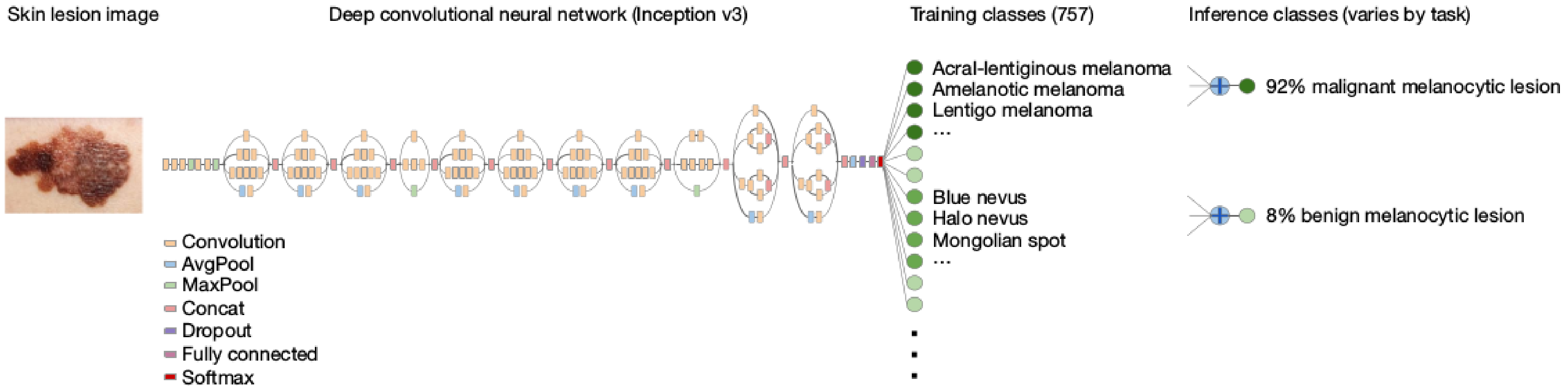


Deep Learning (Neural Networks): example



Esteva et al. Dermatologist-level classification of skin cancer with deep neural networks (2017)

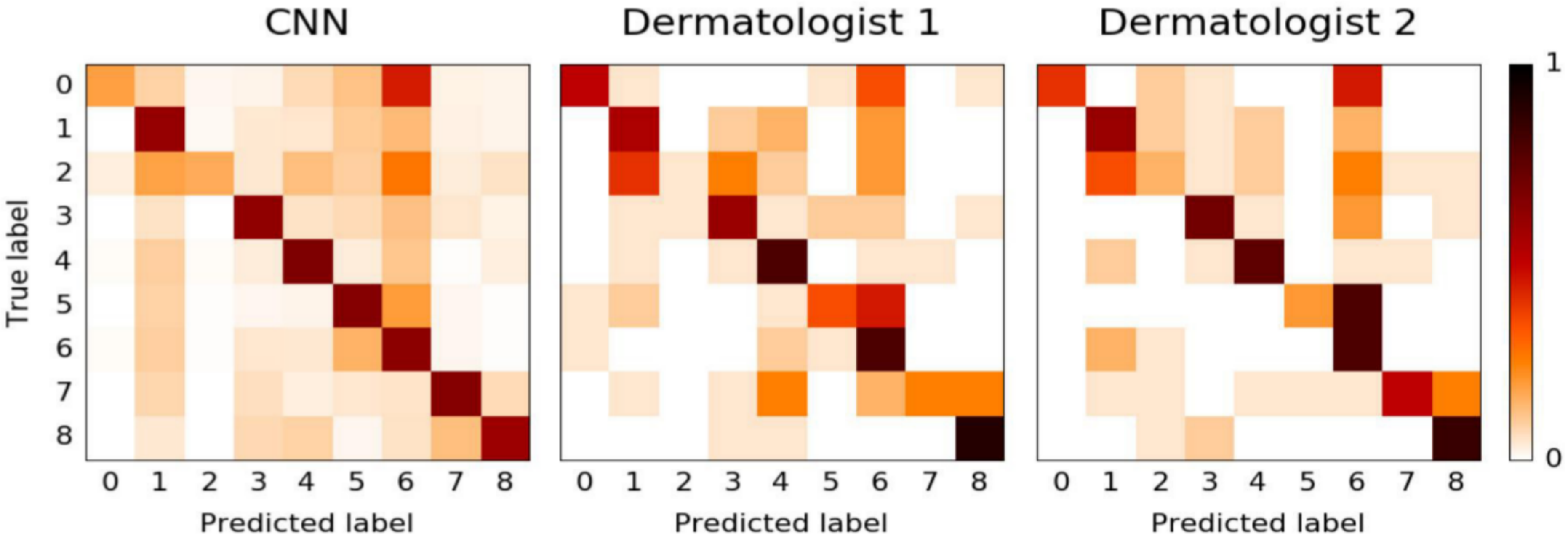
Deep Learning (Neural Networks): example



- base model: Google Inception v3 CNN (pretrained 1.28 million images / 1,000 classes)
- transfer learning: 129,450 skin lesions / 757 classes (2,032 different diseases)

Esteva et al. Dermatologist-level classification of skin cancer with deep neural networks (2017)

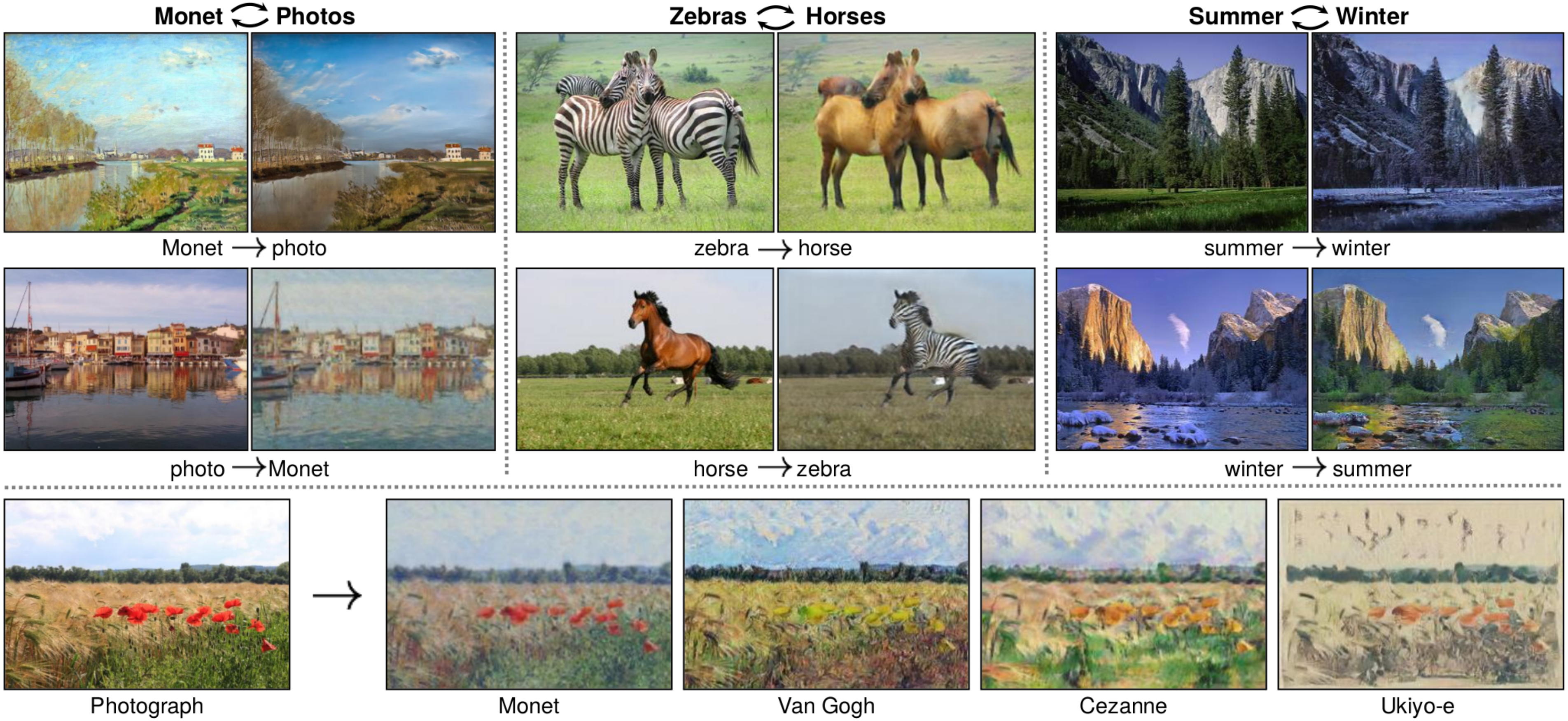
Deep Learning (Neural Networks): example



Extended Data Figure 2 | Confusion matrix comparison between CNN and dermatologists. Confusion matrices for the CNN and both dermatologists for the nine-way classification task of the second validation strategy reveal similarities in misclassification between human experts and the CNN. Element (*i*, *j*) of each confusion matrix represents the empirical probability of predicting class *j* given that the ground truth was class *i*, with *i* and *j* referencing classes from Extended Data Table 2d. Note that both the CNN and the dermatologists noticeably confuse benign and malignant melanocytic lesions—classes 7 and 8—with each other, with dermatologists erring on the side of predicting malignant. The distribution across column 6—inflammatory conditions—is pronounced in all three plots, demonstrating that many lesions are easily confused with this class. The distribution across row 2 in all three plots shows the difficulty of classifying malignant dermal tumours, which appear as little more than cutaneous nodules under the skin. The dermatologist matrices are each computed using the 180 images from the nine-way validation set. The CNN matrix is computed using a random sample of 684 images (equally distributed across the nine classes) from the validation set.

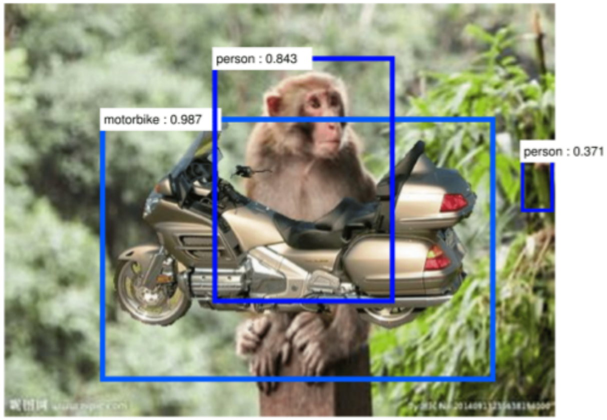
Esteva et al. Dermatologist-level classification of skin cancer with deep neural networks (2017)

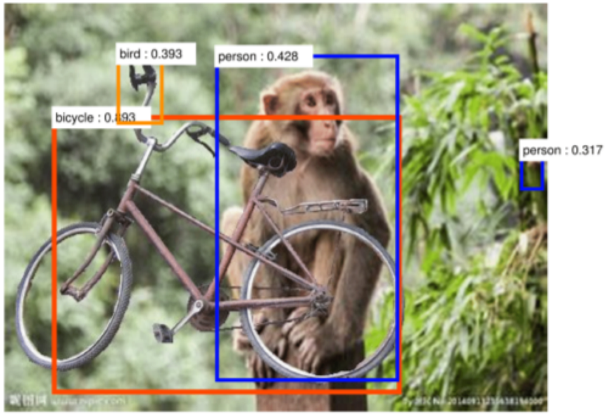
Deep Learning (Neural Networks)

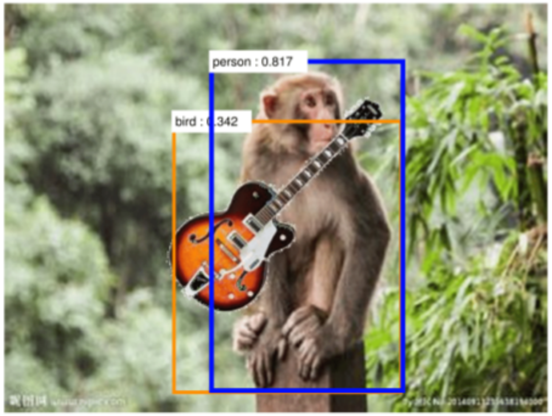


Zhu et al., Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

Deep Learning (Neural Networks)

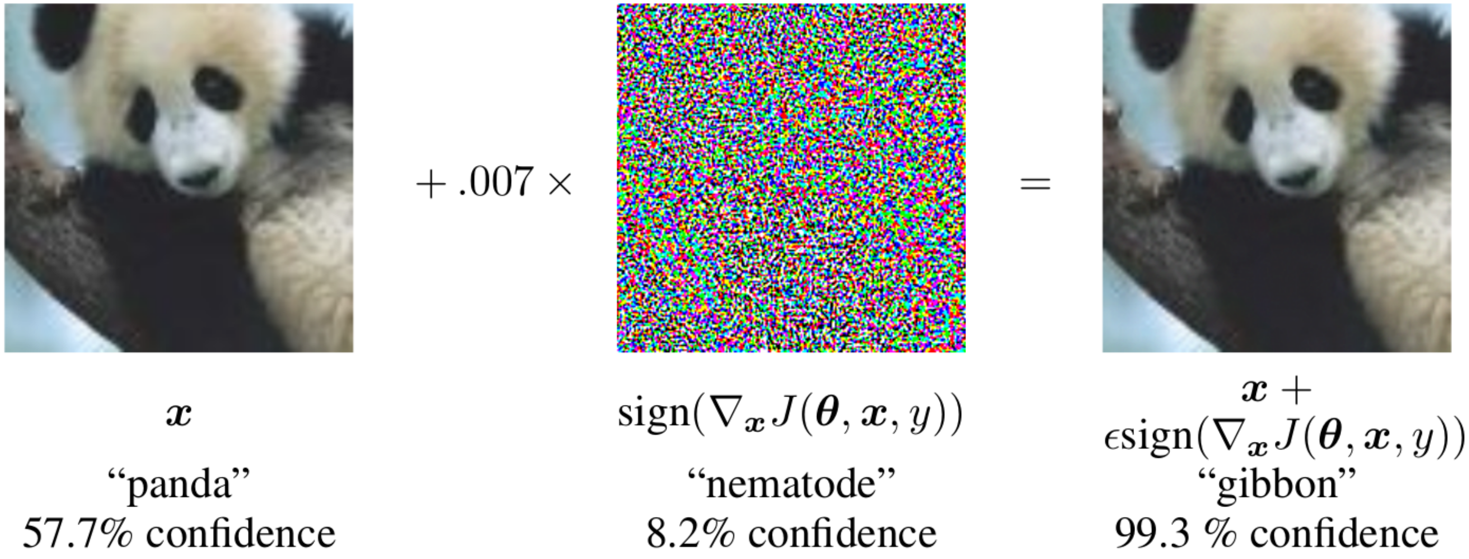






Wang et al. “Visual Concepts and Compositional Voting”

Deep Learning (Neural Networks)


$+ .007 \times$
$=$
$\boldsymbol{x}$
“panda”
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
“nematode”
8.2% confidence
$\boldsymbol{x} +$
$\epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
“gibbon”
99.3 % confidence

Goodfellow et al. “Explaining and Harnessing Adversarial Examples”

Symbolic ML:

- intelligible models
- can use established types of models (if suitable formalisms and algorithms exit)
- can automate various analysis, repair and design tasks
- easy to justify output
- can take advantage of small data
- poor handling of numerical parameters

Std. Supervised ML:

- models capture numerical patterns from labeled data
- main tasks are classification and regression
- human necessary in model selection and feature engineering
- need (often a lot of) labeled data

Std. Unsupervised ML:

- no labels needed
- can be used in concert with supervised methods (dimensionality red.)
- or to find patterns in data (semisupervised classification)

Deep Learning:

- hidden layers allow for feature learning
- state of the art performance on very complex tasks (Moravec’s paradox)
- hype (+/-)
- requires huge amount of data
- complex black box (generalisation?)

How hard is it going to be?

1) Are data available?

- quantity/cost: gathering data can be >80% of the work; also slows everything down
- quality: missing data? noise?
- relevance: spurious features? How easy to extract relevant features? (domain knowledge, another 80%)

How hard is it going to be?

1) Are data available?

- quantity/cost: gathering data can be >80% of the work; also slows everything down
- quality: missing data? noise?
- relevance: spurious features? How easy to extract relevant features? (domain knowledge, another 80%)

2) Does the problem match?

- classification/regression: should be straightforward
- deep learning: hard, unless transfer learning used
- symbolic:straightforward, if suitable formalism exists