

LEARNING FROM IMPRECISE AND FUZZY DATA: ON THE NOTION OF DATA DISAMBIGUATION

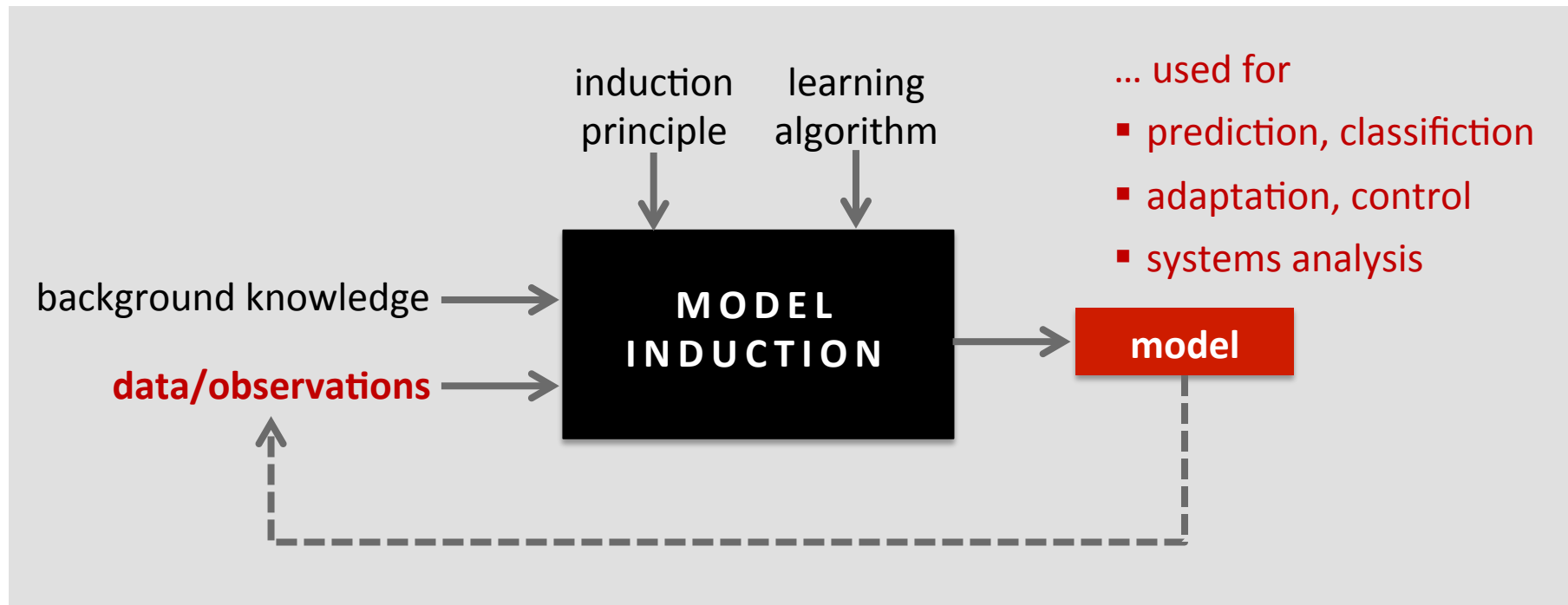
Eyke Hüllermeier

Computational Intelligence Group
Department of Mathematics and Computer Science
Marburg University, Germany



MACHINE LEARNING

SUPERVISED LEARNING: Algorithms and methods for discovering (alleged) dependencies and regularities in a domain of interest, expressed through appropriate models, from specific observations or examples.



MACHINE LEARNING

FUZZY MACHINE LEARNING: Learning **FUZZY MODELS** from **CRISP DATA**!

10	0.34	0
12	0.43	1
21	0.82	0
15	0.93	0
22	0.72	1
18	0.82	1
16	0.62	1

DATA















```
IF x1=high AND x2=low THEN Y=0
IF x1=low  AND x2=low  THEN Y=1
IF x1=high AND x2=high THEN Y=1
IF x1=low  AND x2=high THEN Y=0
```

FUZZY RULES

LEARNING FROM FUZZY DATA

X_1	X_2	X_3	X_4	Y
10	0.42	0	132	10.5
12	0.90	1	154	32.6
17	0.61	1	211	55.2
11	0.17	1	423	94.2
28	0.66	0	654	12.6
19	0.93	0	127	37.4
32	0.72	1	336	33.8
15	0.12	0	798	62.5
...

LEARNING FROM FUZZY DATA

X_1	X_2	X_3	X_4	Y
10	0.42	0	132	10.5
12	0.90	1	154	
	0.61	1		
11		1		94.2
	0.66	0	654	12.6
19		0	127	
32	0.72	1		
15	0.12	0		62.5
...

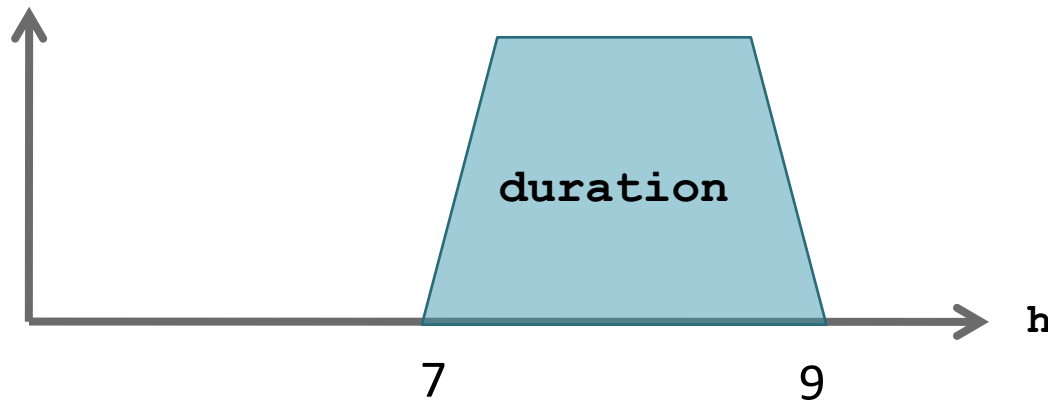
HOW TO ANALYZE AND LEARN FROM SUCH DATA?

TWO INTERPRETATIONS OF A FUZZY SET

The „ontic“ view (conjunctive interpretation):

- a fuzzy set is a real data entity;
- an attribute can assume a fuzzy set as a „value“, i.e.,
- we have a (fuzzy set)-valued attribute.

EXAMPLE: Duration of sunshine in Vilamoura today.



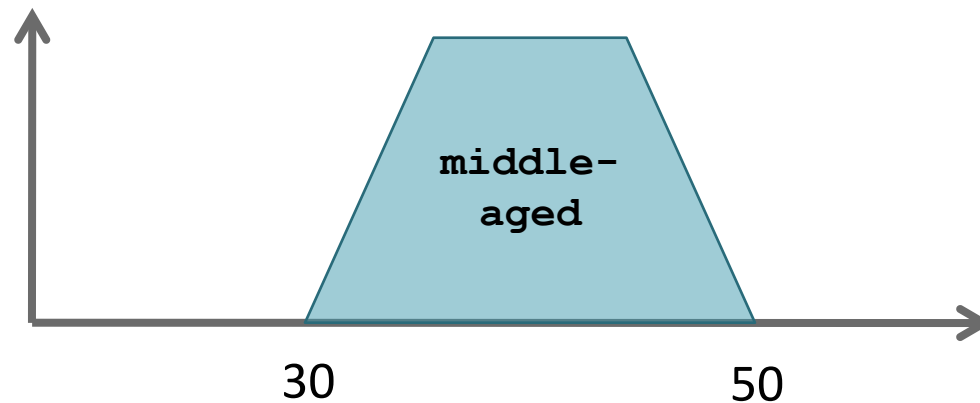
THE ONTIC VIEW: REMARKS

- In line with the general trend of **analyzing „complex“ data** (e.g., interval-valued, histogram-valued, functional, etc.)
- Questionable relevance for machine learning/data analysis:
 - A systematic collection of fuzzy data of that kind requires a suitable **„measurement device“** producing fuzzy sets (membership functions).
 - What is the **meaning of a membership degree**, if not related to frequency (and hence probability distributions)?

TWO INTERPRETATIONS OF A FUZZY SET

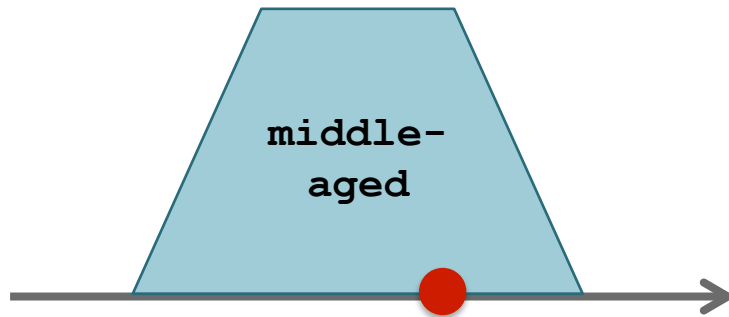
The „epistemic“ view (disjunctive interpretation):

- The true value of the attribute is precise, and a fuzzy set is used to express imprecise knowledge about this value (possibility distribution).

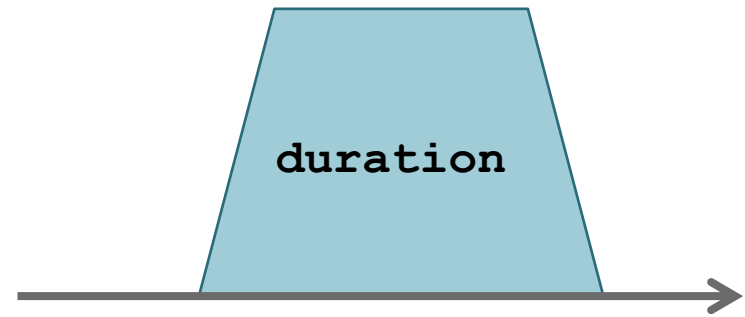


A FUZZY SET IS NOT THE **DATA OBJECT**, BUT REPRESENTS
KNOWLEDGE ABOUT THIS OBJECT!

TWO INTERPRETATIONS OF A FUZZY SET



Fuzzy set could be replaced by a precise value on the basis of additional knowledge.



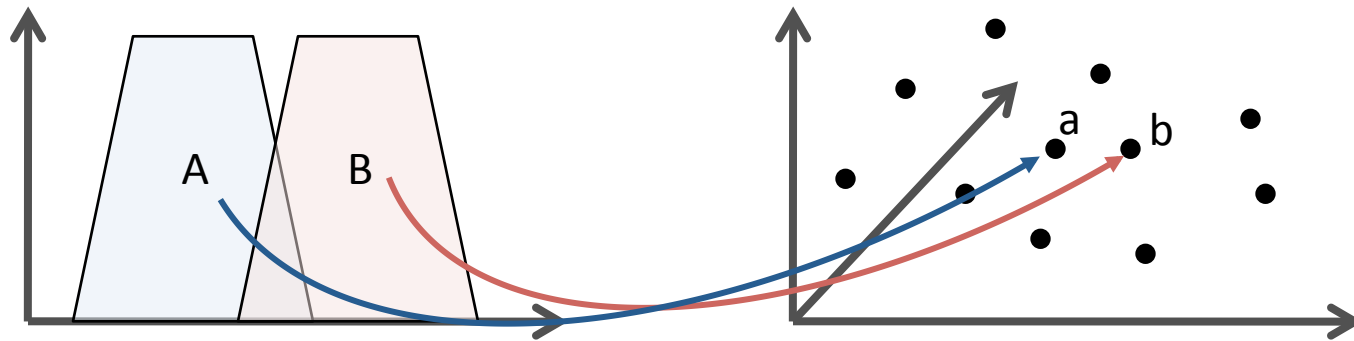
A further „precisiation“ of the data is not legitimate.

TWO INTERPRETATIONS OF FUZZY DATA

THE TWO INTERPRETATIONS, ONTIC AND EPISTEMIC, CALL FOR VERY
DIFFERENT EXTENSIONS OF METHODS FOR DATA ANALYSIS!

ANALYZING COMPLEX DATA

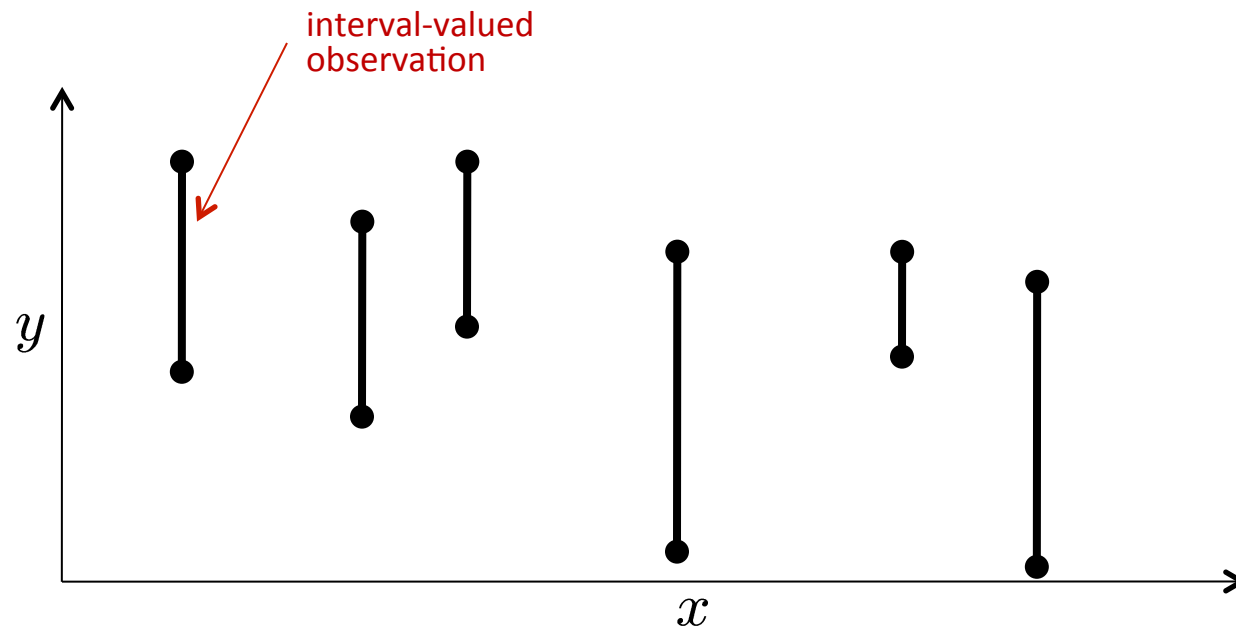
The ontic view essentially calls for „lifting“ a method to a **complex data space**, in which data entities are fuzzy sets, and to extend the underlying operations correspondingly.



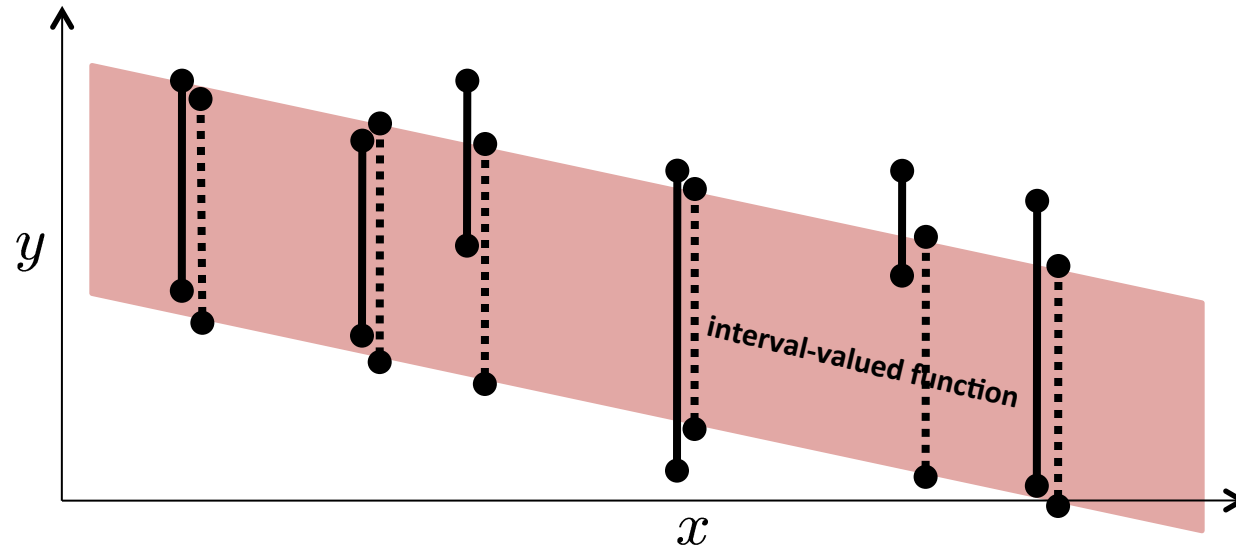
„Fuzzy observations“ are embedded as points in a (high-dimensional) space
(e.g., a fuzzy metric space).

Special case of **structured output prediction**, for which kernel-based learning methods are quite popular (kernels for sequences, graphs, etc.)

REGRESSION WITH INTERVAL OUTPUTS



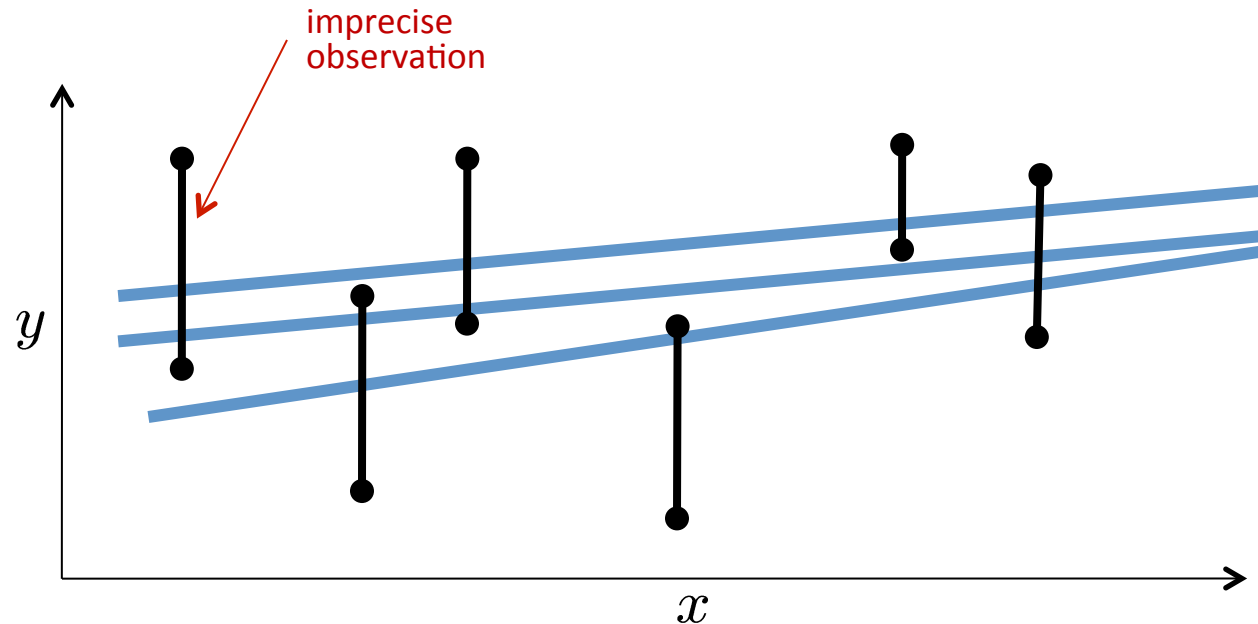
REGRESSION WITH INTERVAL OUTPUTS



Ontic view: Reproducing interval observations by means of an interval-valued function

$$F^* \in \arg \min_{F \in \mathcal{F}} \sum_i D(Y_i, F(x_i))$$

THE EPISTEMIC VIEW



Epistemic view: Solution is a

- fuzzy set of **REAL-VALUED** regression functions
- instead of a **single FUZZY SET-VALUED** regression function

THE EPISTEMIC VIEW

A model is deemed possible if there is a possible set of precise observations (a **SELECTION**) for which it is an optimal fit

→ **EXTENSION PRINCIPLE (applied to a data analysis method) ?**

THE EXTENSION PRINCIPLE

- The extension principle generalizes a function

$$f : X_1 \times X_2 \times \dots \times X_n \rightarrow Y$$

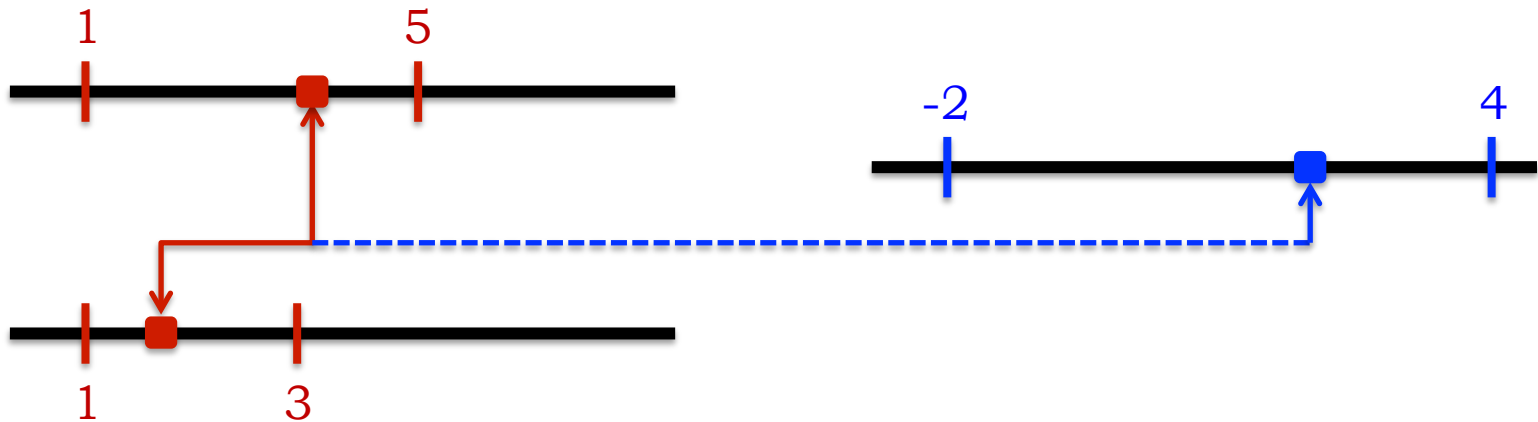
from „crisp“ to fuzzy inputs:

$$\begin{array}{ccccccc} f(x_1, x_2, \dots, x_n) = y & & & & & & \\ \downarrow & \downarrow & \downarrow & & \downarrow & & ? \\ F(A_1, A_2, \dots, A_n) = Y & & & & & & \end{array}$$

$$\mu_Y(y) = \sup_{\mathbf{x}=(x_1, \dots, x_n)} \left\{ \min(A_1(x_1), \dots, A_n(x_n)) \mid f(\mathbf{x}) = y \right\}$$

THE EXTENSION PRINCIPLE

- For example, interval arithmetics: $[1, 5] \ominus [1, 3] = [-2, 4]$



All selections of (single-valued) input values are treated the same and equally contribute to the output!

THE EXTENSION PRINCIPLE

- A learning algorithm is a **mapping from data to models**:

$$f : \mathbf{D}^n \rightarrow \mathbf{M}, \mathbf{d} = (d_1, \dots, d_n) \mapsto M$$

- So, the extension now reads as follows:

$$F(D)(M) = F(D_1, \dots, D_n)(M) = \sup_{\mathbf{d}=(d_1, \dots, d_n)} \left\{ \min_i D_i(d_i) \mid f(\mathbf{d}) = M \right\}$$

- Thus, a model is plausible insofar there is a plausible selection of precise data points supporting that model:

$$\pi(M) = \sup_{\mathbf{d}} \left\{ \mu_D(\mathbf{d}) \mid f(\mathbf{d}) = M \right\}$$

with $\mu_D(\mathbf{d}) = \min(D_1(d_1), \dots, D_n(d_n))$.

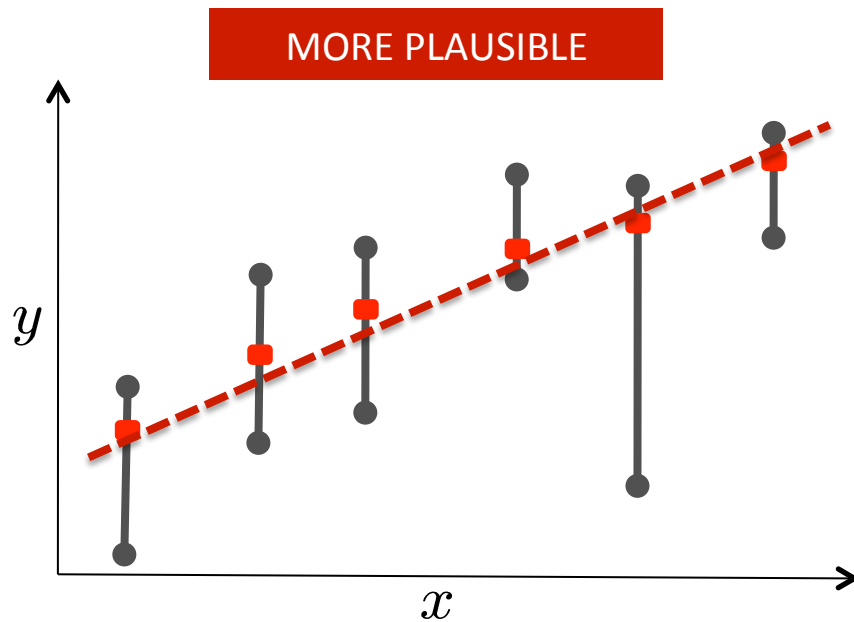
THE EXTENSION PRINCIPLE

Questioning the equal treatment of all selections ...

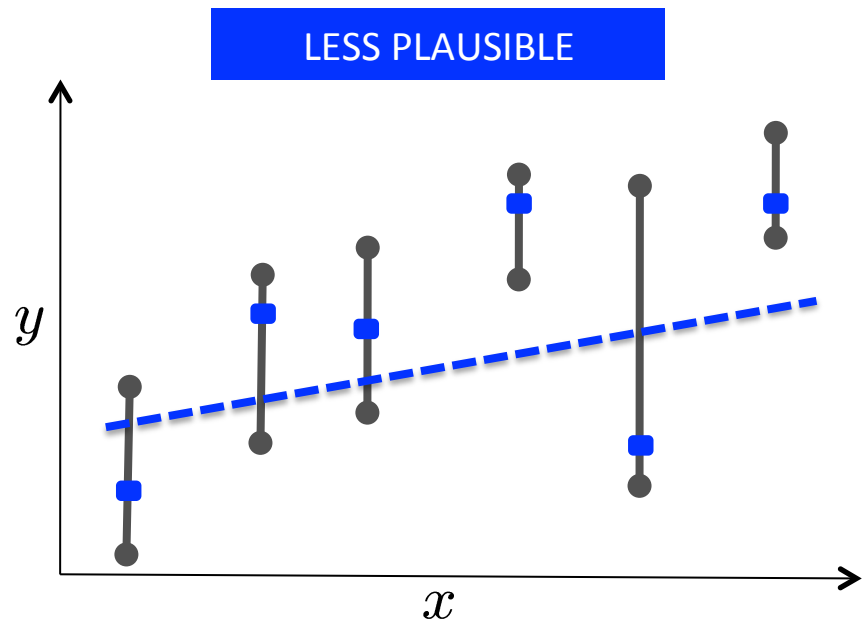
In data analysis, a method inducing a model from a set of data always comes with certain **MODEL ASSUMPTIONS**, and under these assumptions, specific selections may appear more plausible than others!

... to be explained through some simple examples.

DATA DISAMBIGUATION

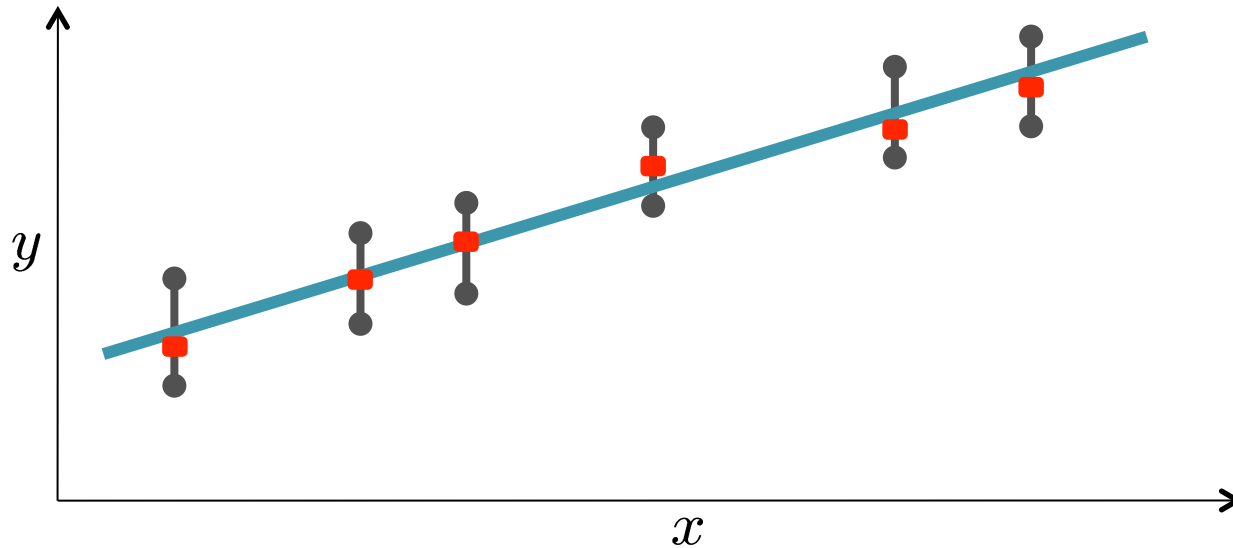


A plausible selection that can be fitted quite well with a **LINEAR** model!



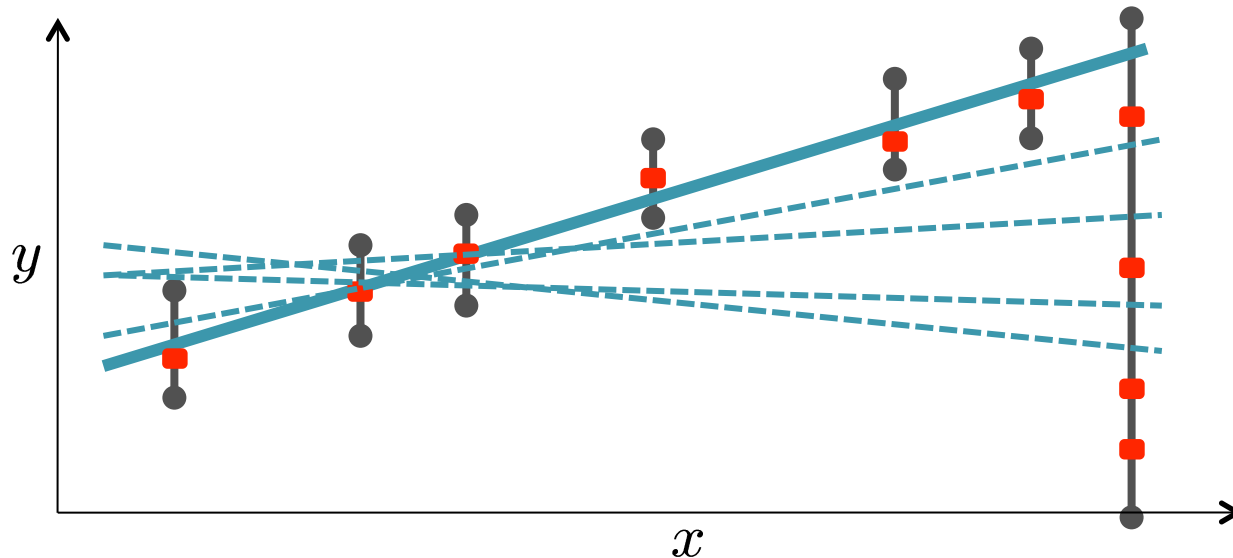
A less plausible selection, because there is no **LINEAR** model with a good fit!

DATA DISAMBIGUATION



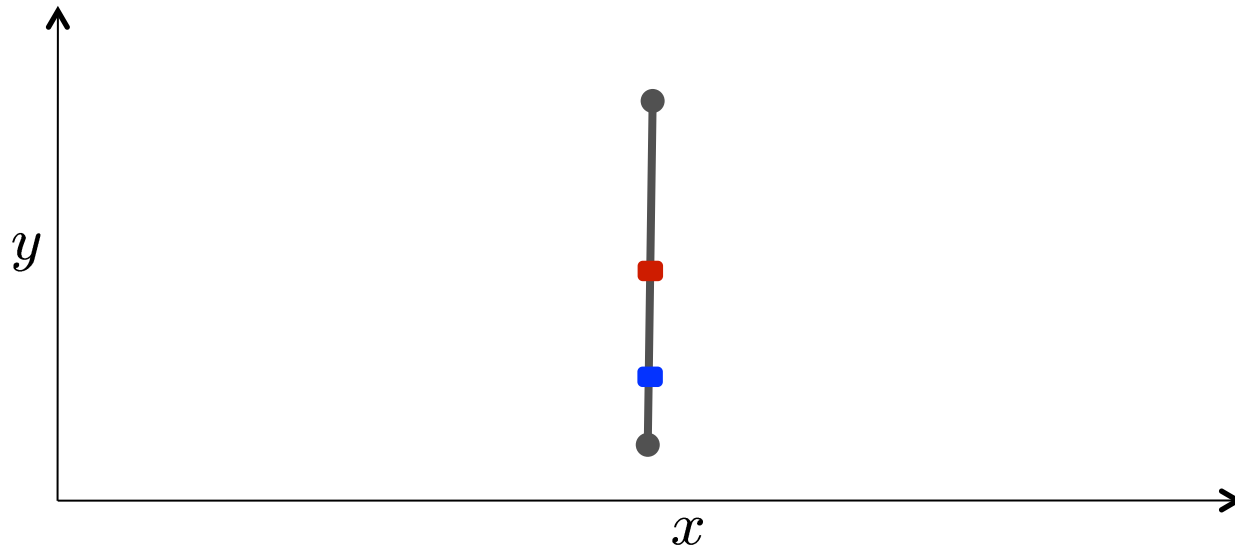
Adding **non-informative data** will have an influence on the plausibility of models!

DATA DISAMBIGUATION



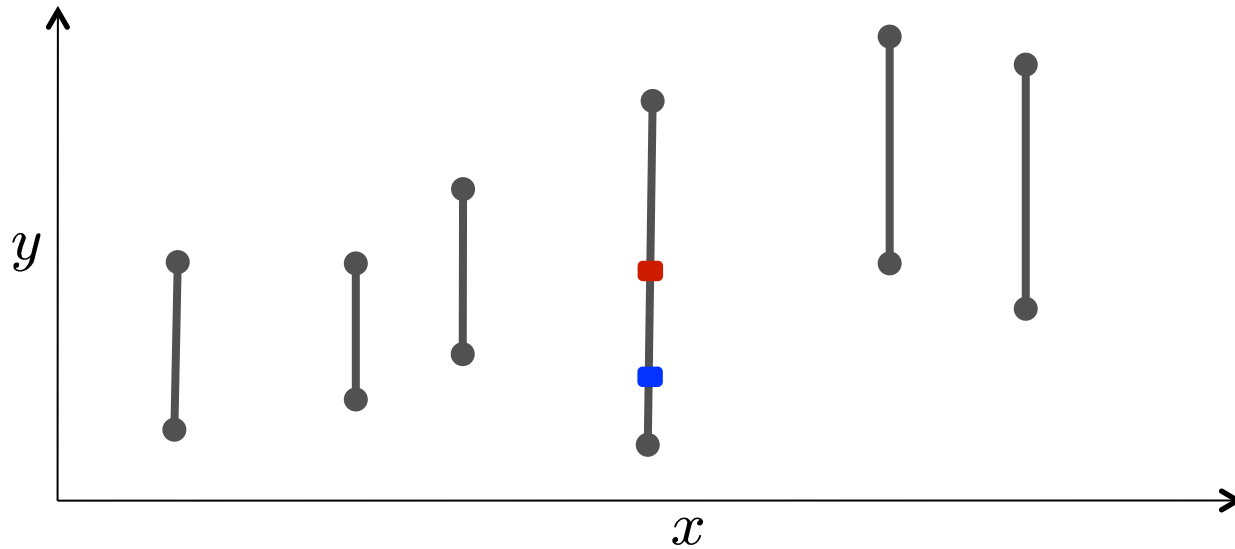
Adding **non-informative data** will have an influence on the plausibility of models!

DATA DISAMBIGUATION



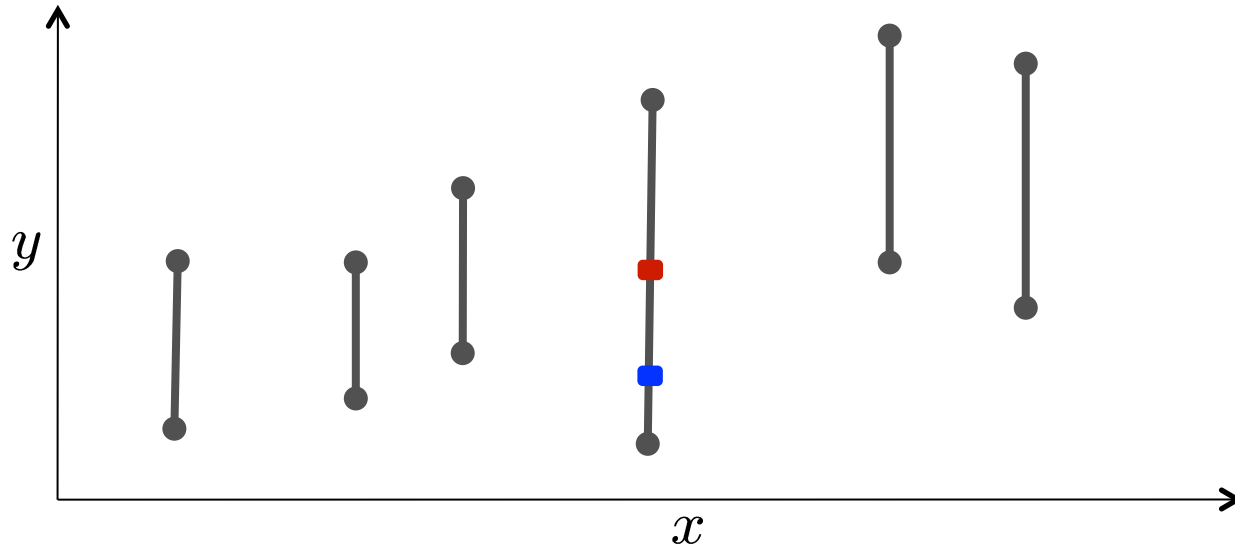
A single (imprecise) observation doesn't tell us very much ...

DATA DISAMBIGUATION



... and neither does a set of them.

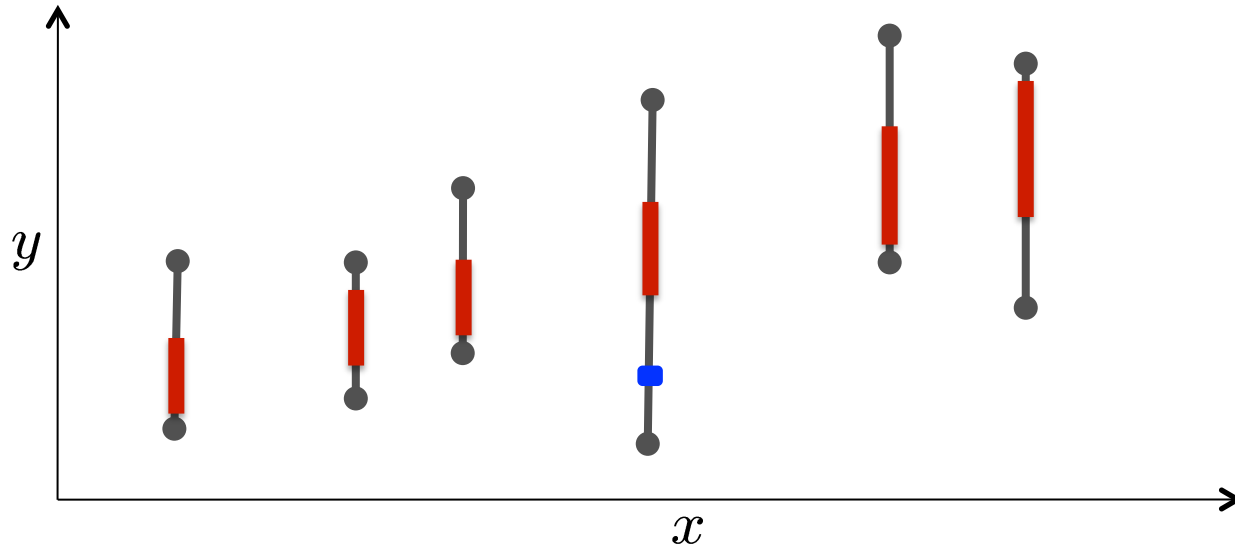
DATA DISAMBIGUATION



Yet, when looking at the data **AS A WHOLE**, and taking into account the **RELATION BETWEEN THEM**, some possible values become impossible!

→ constraint propagation

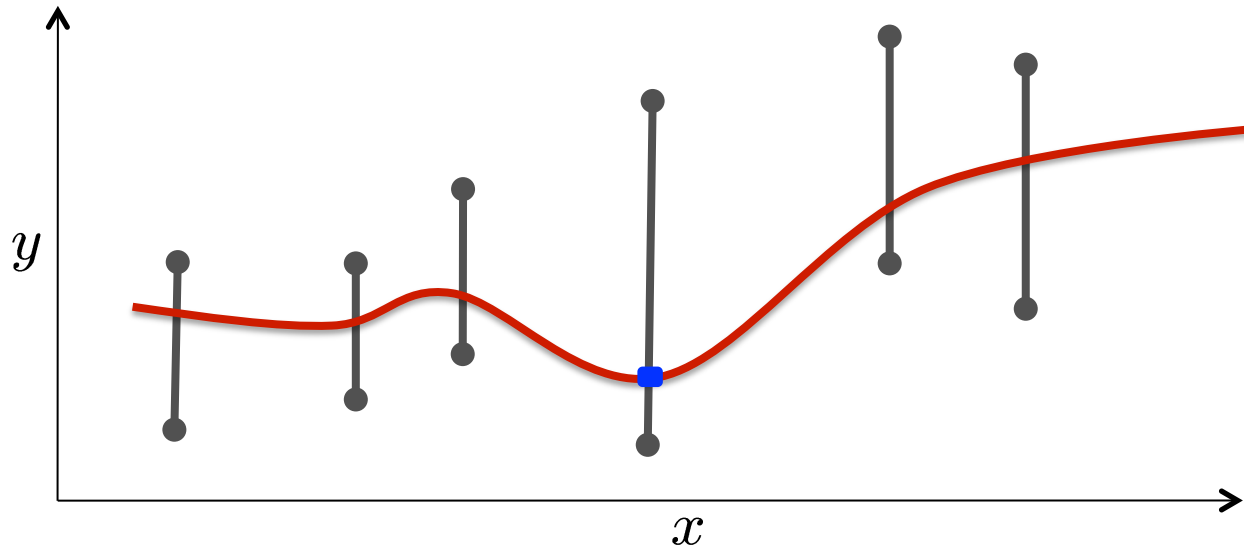
DATA DISAMBIGUATION



Yet, when looking at the data **AS A WHOLE**, and taking into account the **RELATION BETWEEN THEM**, some possible values become impossible!

→ constraint propagation

DATA DISAMBIGUATION

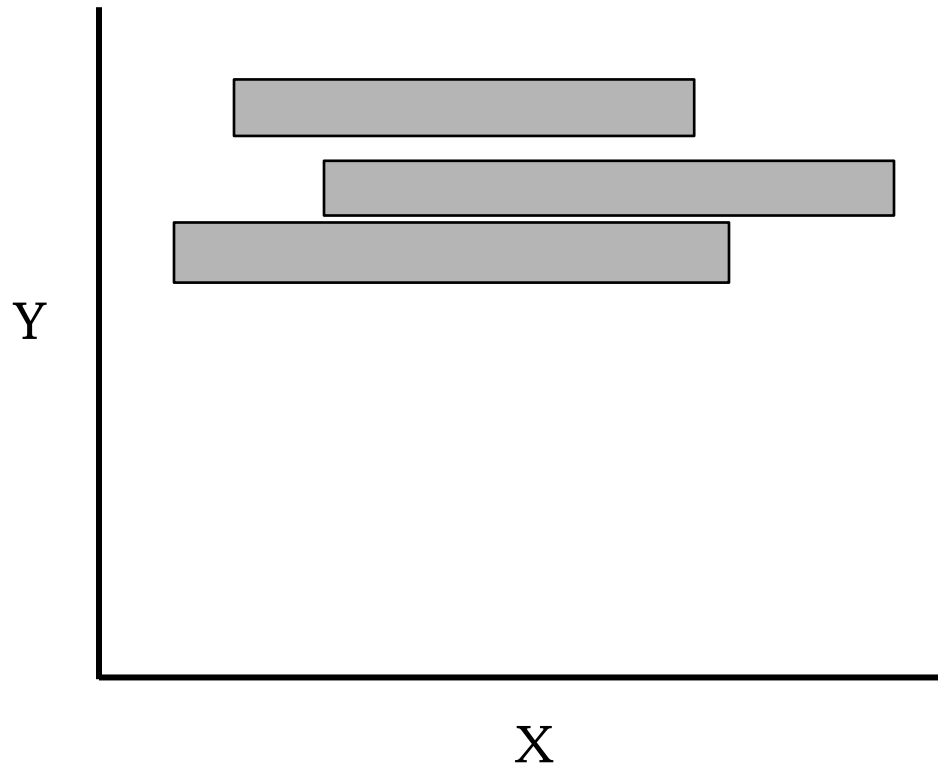


Yet, when looking at the data **AS A WHOLE**, and taking into account the **RELATION BETWEEN THEM**, some possible values become impossible!

→ constraint propagation

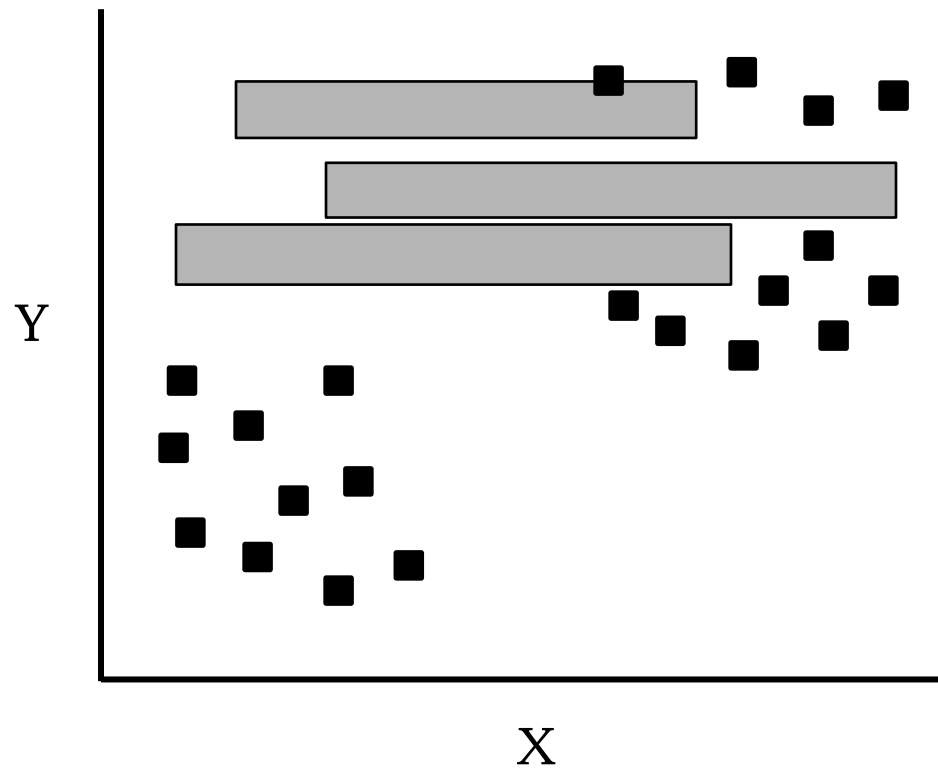
DATA DISAMBIGUATION IN CLUSTERING

Imprecise x-values modeled as intervals.



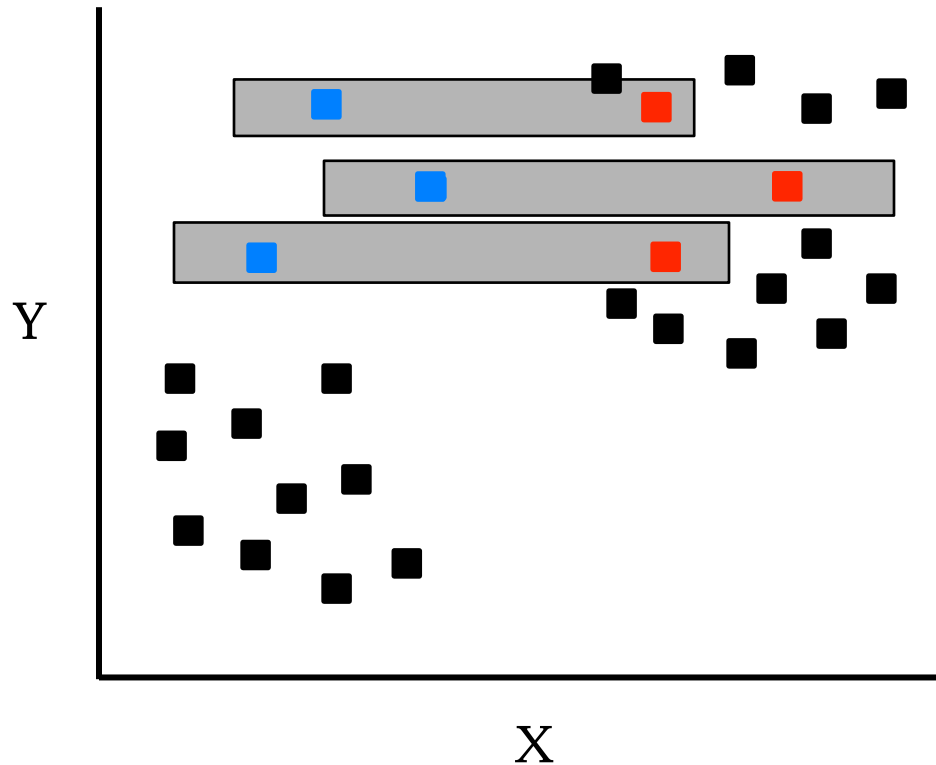
DATA DISAMBIGUATION IN CLUSTERING

Imprecise x-values modeled as intervals.



DATA DISAMBIGUATION IN CLUSTERING

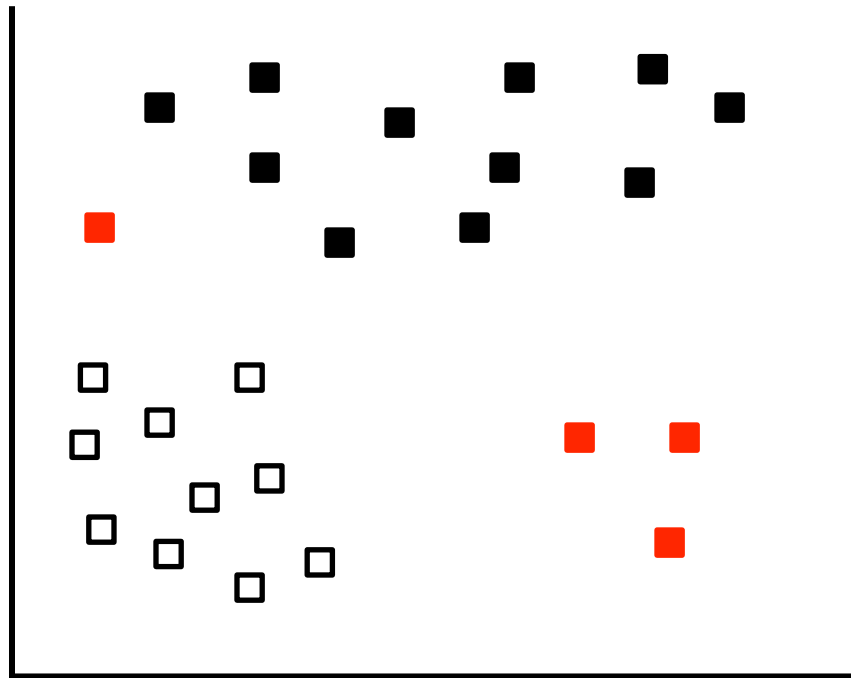
Scenario „red“ more likely than „blue“ !



The red scenario (two clusters) appears to be more plausible than the blue one (three clusters)!

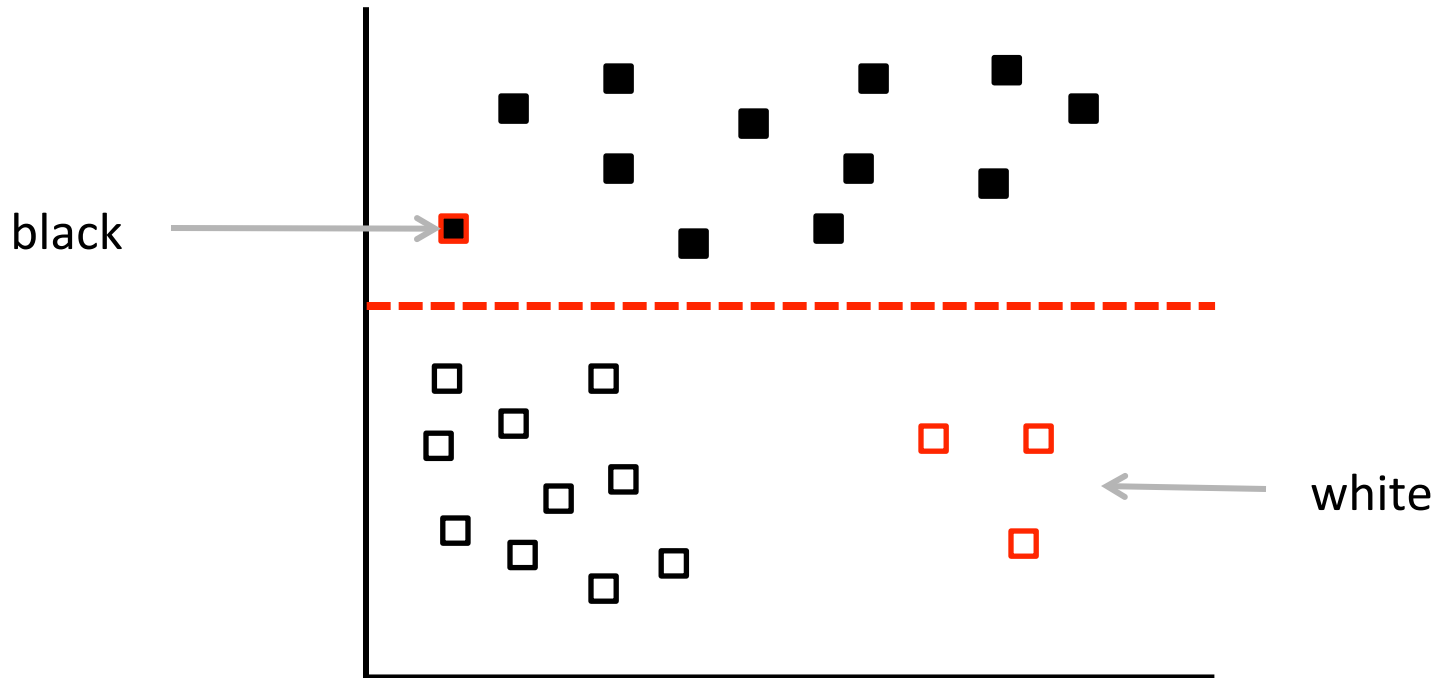
DATA DISAMBIGUATION IN CLASSIFICATION

The class of the red training points is not known.



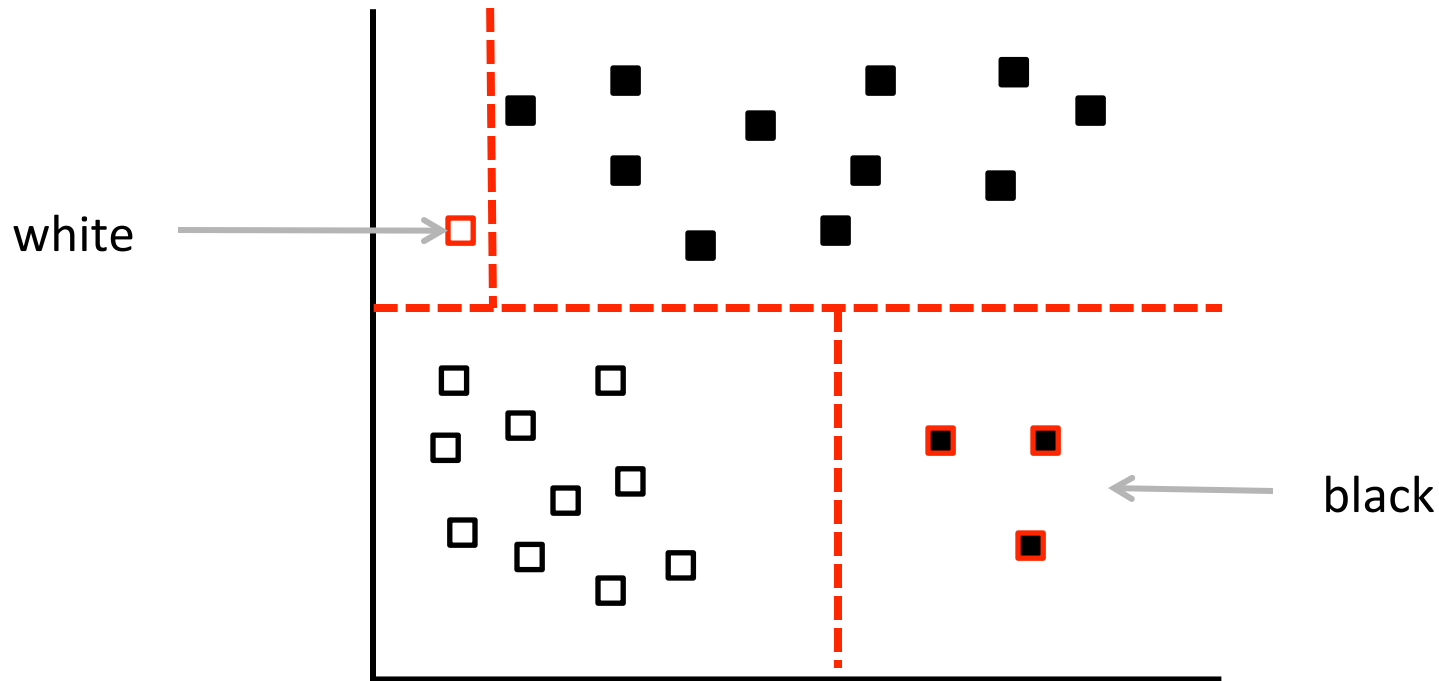
DATA DISAMBIGUATION IN CLASSIFICATION

This scenario allows one to fit a very simple decision tree, while other scenarios call for more complex models.



DATA DISAMBIGUATION IN CLASSIFICATION

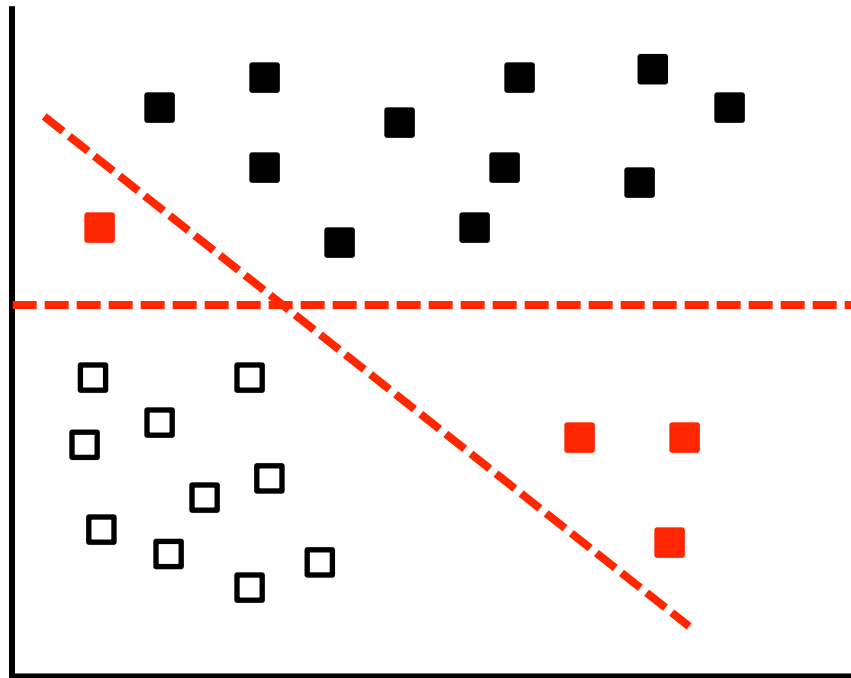
This scenario allows one to fit a very simple decision tree, while other scenarios call for more complex models.



Looking at the data from the point of view of a decision tree learner, the former scenario appears more likely than the latter.

DATA DISAMBIGUATION IN CLASSIFICATION

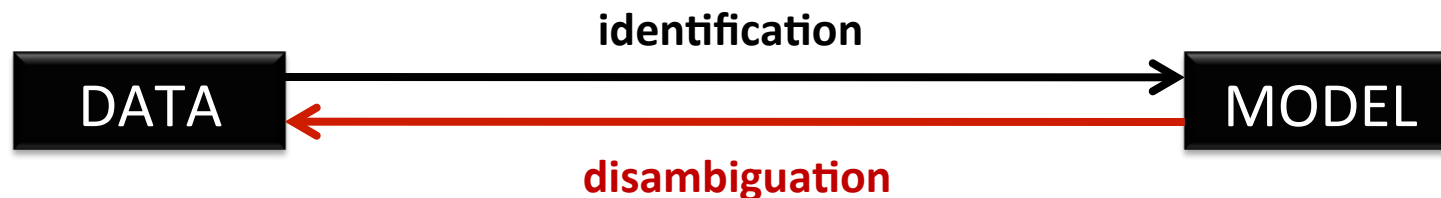
The same does not necessarily hold under different model assumptions !



It all depends on how you look at the data!

DATA DISAMBIGUATION

Under the epistemic view, **model identification** and **data disambiguation** should be performed simultaneously:



The loss minimization approach ...

THE EMPIRICAL RISK MINIMIZATION PRINCIPLE

Many (supervised) learning methods are based on minimization of the **empirical risk**

$$\mathcal{R}_{emp}(M) = \frac{1}{N} \sum_{i=1}^N L(y_i, M(\mathbf{x}_i))$$

or a regularized version thereof:

$$\mathcal{R}_{reg}(M) = \underbrace{\frac{1}{N} \sum_{i=1}^N L(y_i, M(\mathbf{x}_i))}_{\text{average loss on training data}} + \underbrace{\lambda C(M)}_{\text{complexity term to prevent overfitting}}$$

GENERALIZED EMPIRICAL RISK MINIMIZATION

Consider an imprecise observation (\boldsymbol{x}, Y) and let $\hat{y} = M(\boldsymbol{x})$.

How much should M be penalized for this prediction?

In agreement with the idea of data disambiguation, we look at the smallest possible loss, namely

$$L^*(Y, \hat{y}) = \min \{ L(y, \hat{y}) \mid y \in Y \} ,$$

and the value for which it is obtained:

$$y^* = \arg \min \{ L(y, \hat{y}) \mid y \in Y \} .$$

Given the model M , this value appears to be the most plausible in Y .

GENERALIZED EMPIRICAL RISK MINIMIZATION

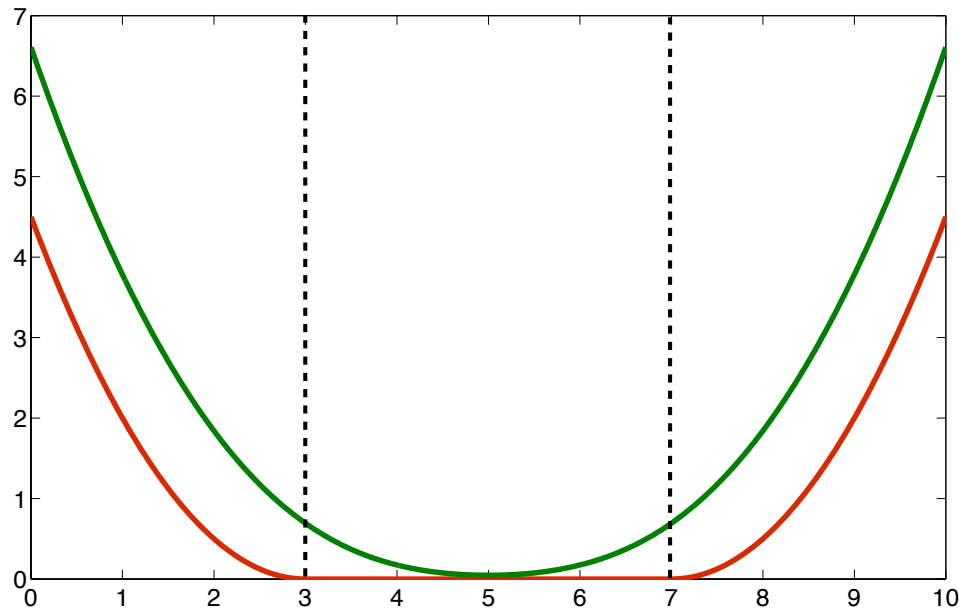
On the basis of the generalized loss function L^* , we define

$$\mathcal{R}_{emp}(M) = \frac{1}{N} \sum_{i=1}^N L^*(Y_i, M(\mathbf{x}_i)) .$$



how well the „crisp“ model
fits the imprecise data

GENERALIZED LOSS FUNCTION: THE INTERVAL CASE



Note similarity to
eps-insensitive
loss in support
vector regression

generalized

standard

GENERALIZATION TO THE CASE OF FUZZY DATA

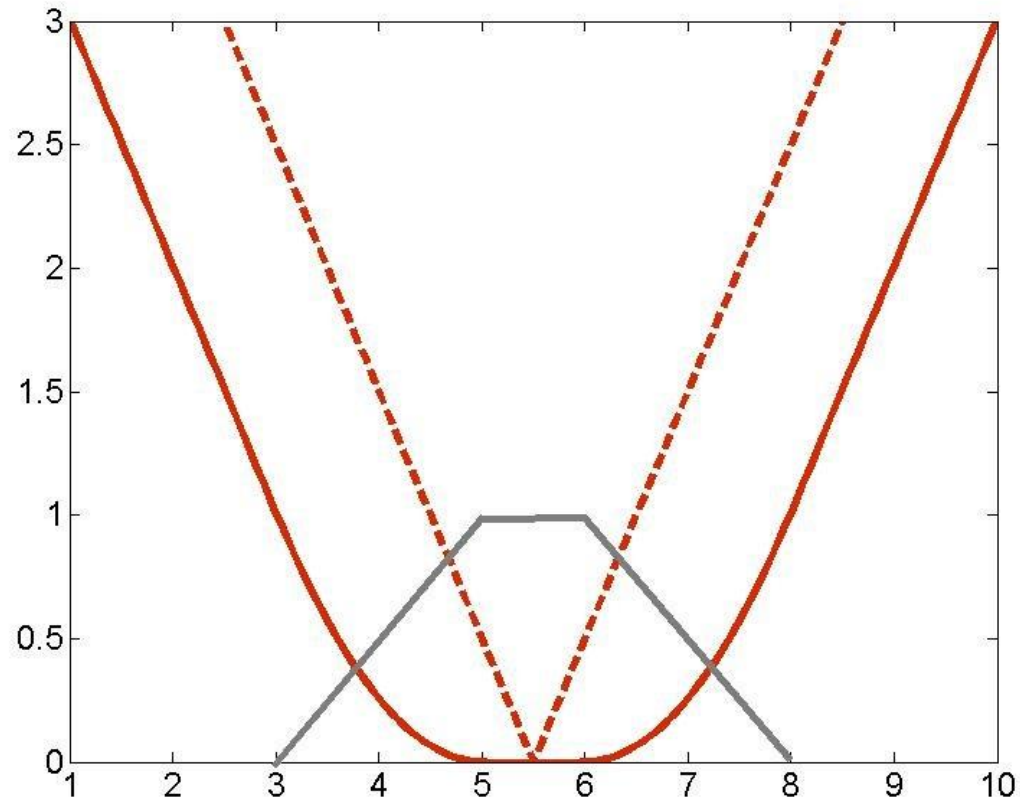
$$\mathcal{L}(Y, \hat{y}) = \int_0^1 L^*([Y]_\alpha, \hat{y}) d\alpha$$

LOSS

$$\overline{\mathcal{R}}_{emp}(M) = \frac{1}{N} \sum_{i=1}^N \mathcal{L}(Y_i, M(\mathbf{x}_i))$$

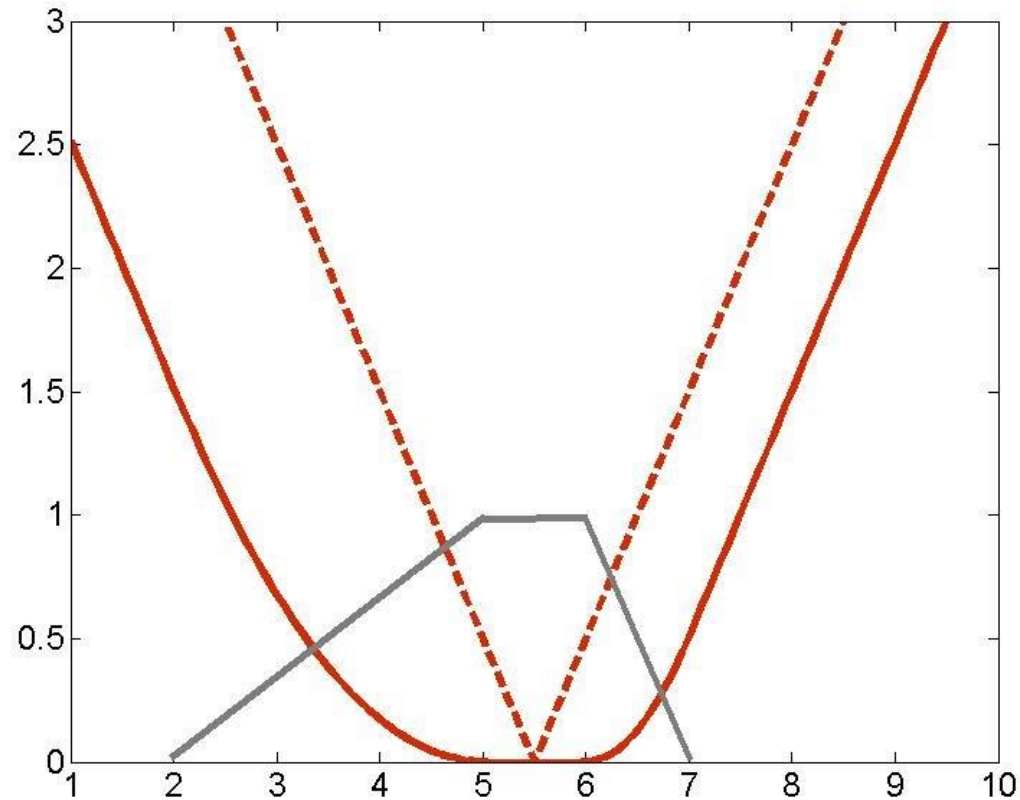
RISK

FUZZIFICATION OF L1 LOSS



→ close connection to Huber loss!

FUZZIFICATION OF L1 LOSS



→ overestimation is worse than underestimation!

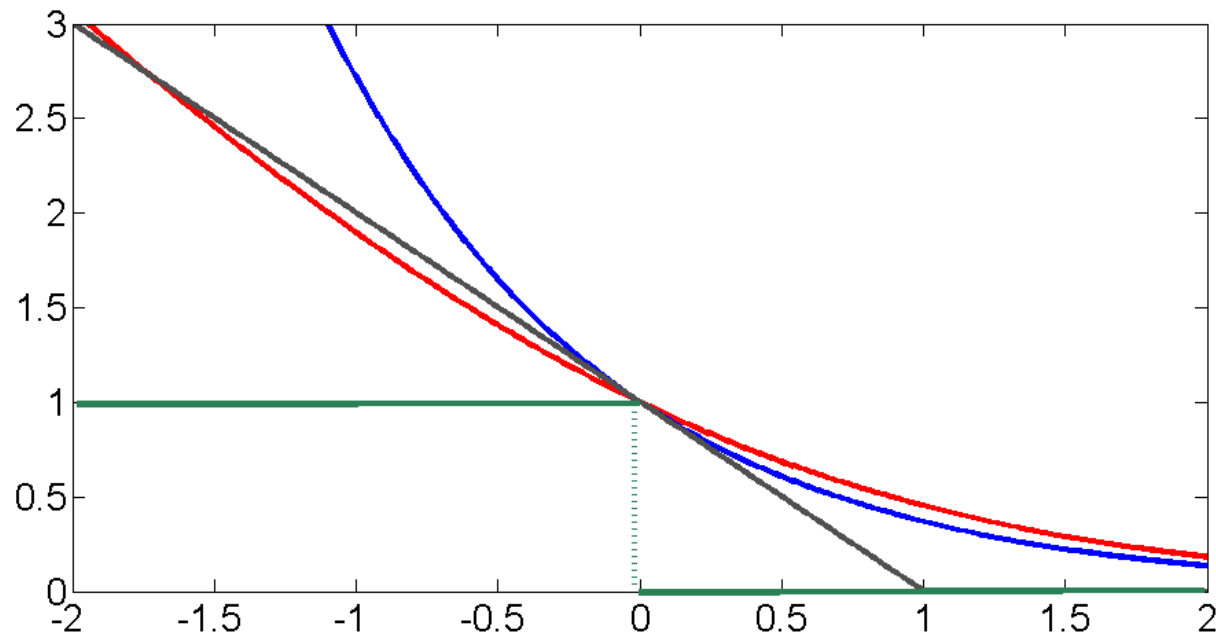
MARGIN LOSSES

Let $\mathcal{Y} = \{-1, +1\}$ and consider a class of scoring classifiers $M : \mathcal{X} \rightarrow \mathbb{R}$.

A margin loss is a function of the form

$$L(y, s) = f(ys) ,$$

where $f : \mathbb{R} \rightarrow \mathbb{R}$ is a non-increasing function.



0/1 loss

logistic

hinge

exponential

MARGIN LOSSES

Let $\mathcal{Y} = \{-1, +1\}$ and consider a class of scoring classifiers $M : \mathcal{X} \rightarrow \mathbb{R}$.

A margin loss is a function of the form

$$L(y, s) = f(ys) ,$$

where $f : \mathbb{R} \rightarrow \mathbb{R}$ is a non-increasing function.

Hinge loss:

$$L(y, s) = \max(1 - ys, 0)$$

Log-loss:

$$L(y, s) = \log(1 + \exp(-ys))$$

Exponential loss:

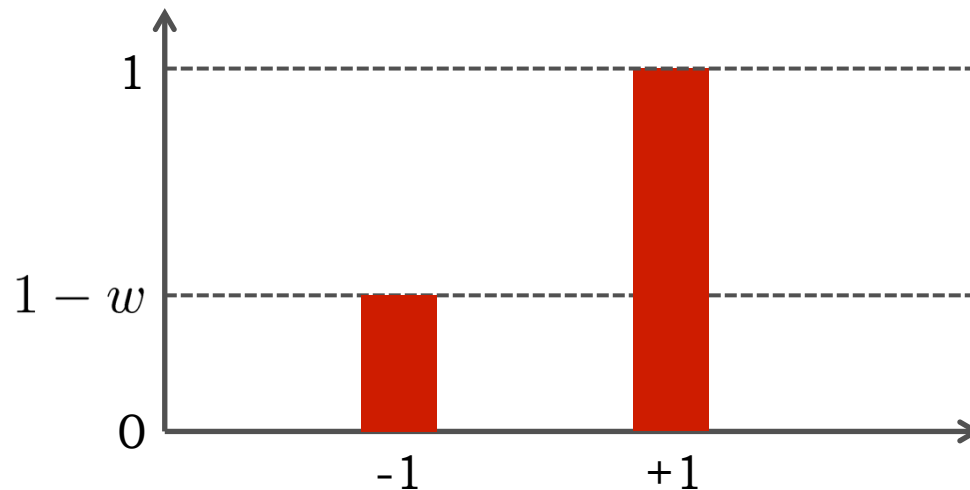
$$L(y, s) = \exp(-s)$$

FUZZY MARGIN LOSSES

Suppose the output is a fuzzy subset Y with membership degrees

$$\mu_Y(\lambda) = \begin{cases} 1 & \text{if } \lambda = y \\ 1 - w & \text{if } \lambda = \bar{y} \end{cases},$$

where $y, \bar{y} \in \{-1, +1\}$ such that $y\bar{y} = -1$, and w can be interpreted as a degree of confidence in y .



FUZZY MARGIN LOSSES

Suppose the output is a fuzzy subset Y with membership degrees

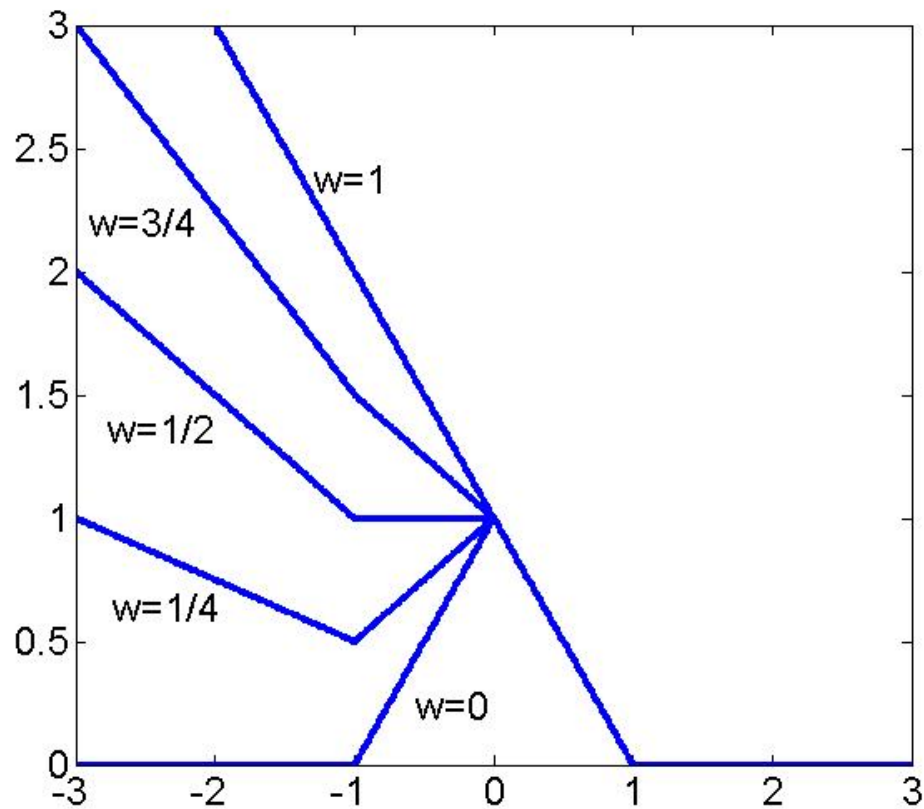
$$\mu_Y(\lambda) = \begin{cases} 1 & \text{if } \lambda = y \\ 1 - w & \text{if } \lambda = \bar{y} \end{cases} ,$$

where $y, \bar{y} \in \{-1, +1\}$ such that $y\bar{y} = -1$, and w can be interpreted as a degree of confidence in y .

Then, the fuzzy loss function is given by

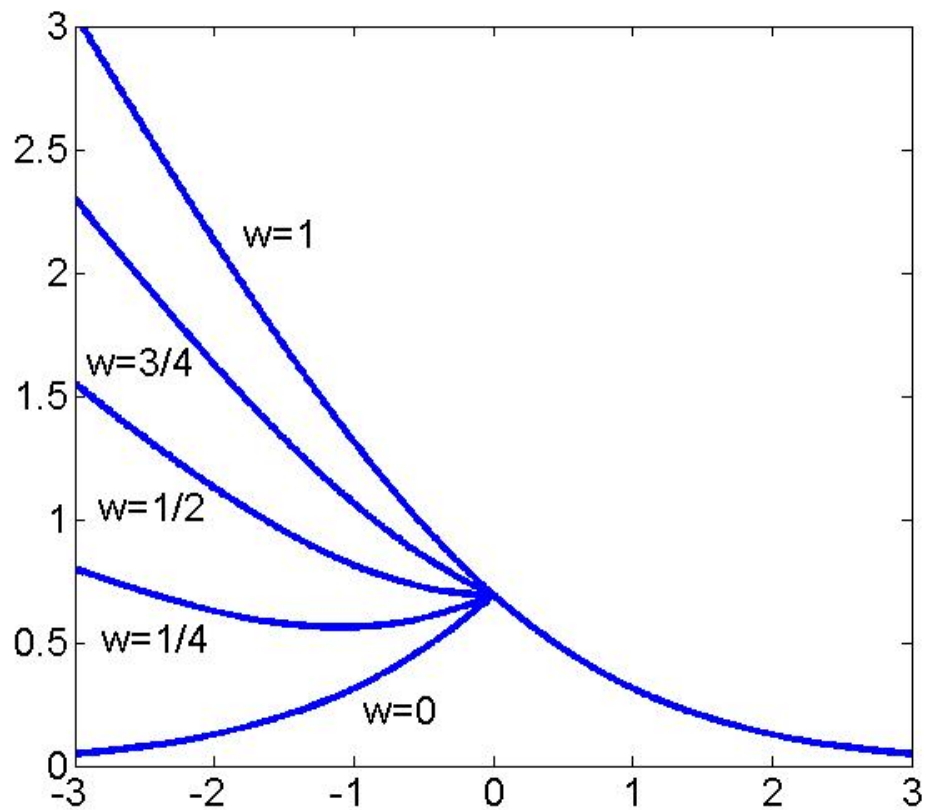
$$\mathcal{L}(Y, s) = f_w(ys) = w \cdot f(ys) + (1 - w) \cdot f(|ys|) .$$

FUZZY MARGIN LOSSES



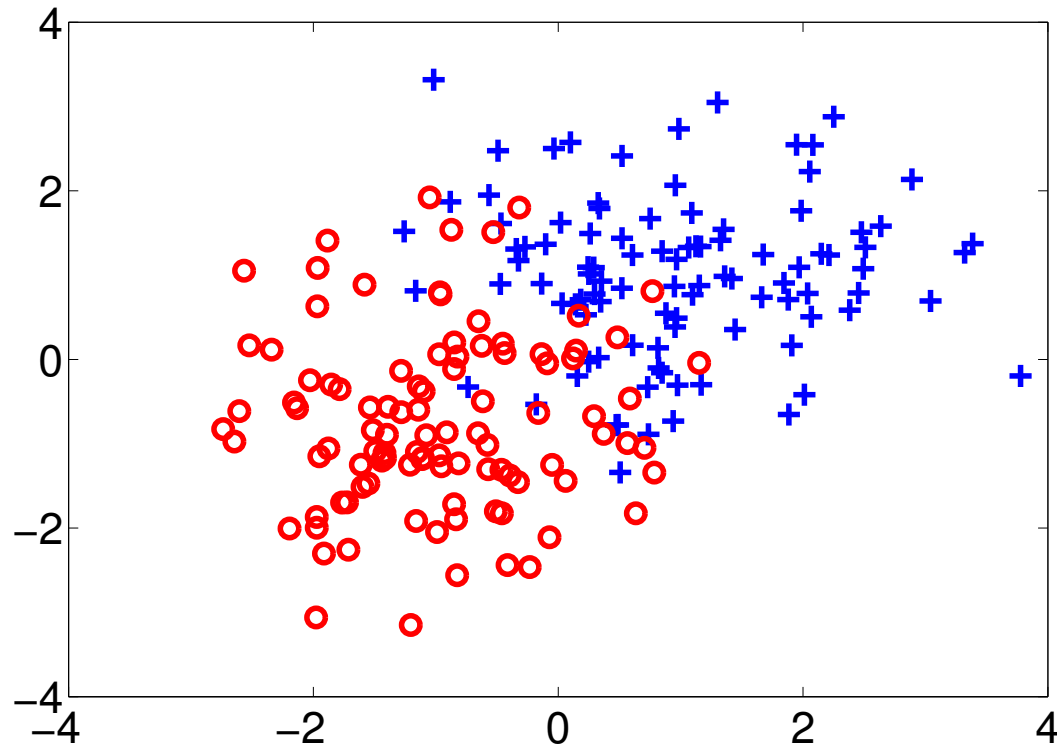
HINGE LOSS

FUZZY MARGIN LOSSES



LOG-LOSS

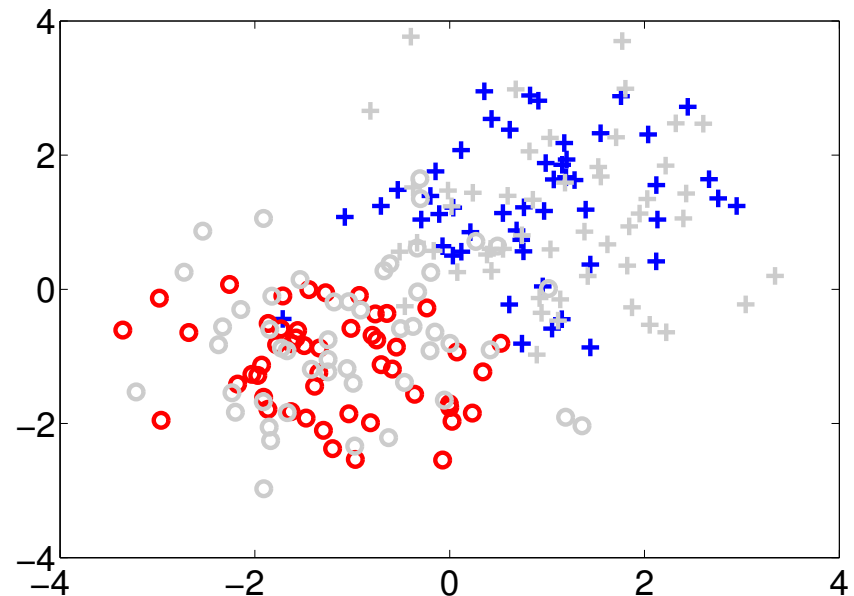
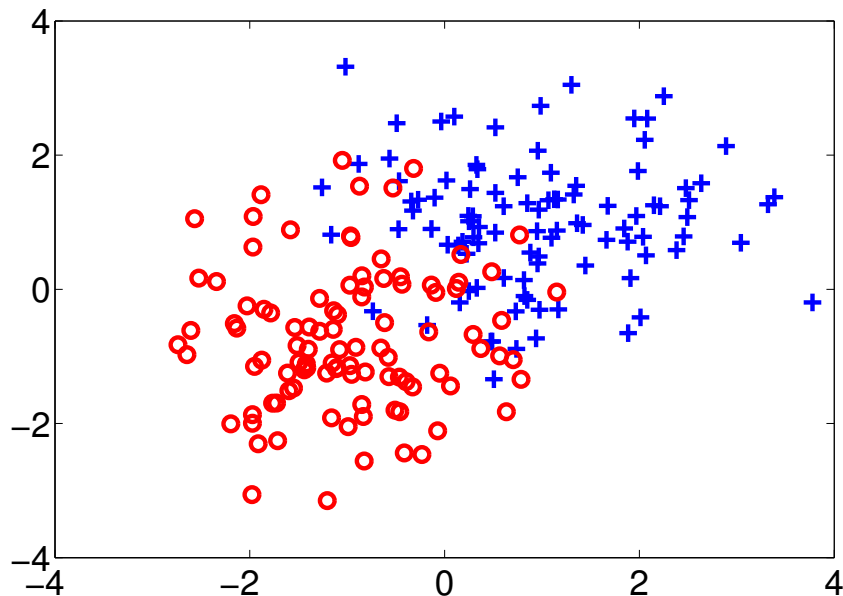
AN ILLUSTRATION



Two classes, both normally distributed, sample size 200.

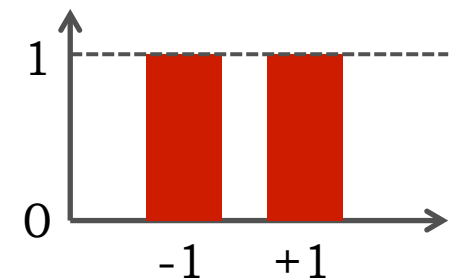
FIRST EXPERIMENT

- Class information was partly removed from the training instances.
- More specifically, each of the 200 instances was declared „unlabeled“ with a fixed probability γ .
- Thus, we are in a **semi-supervised setting**, in which approximately $200(1-\gamma)$ of the instances are labeled.

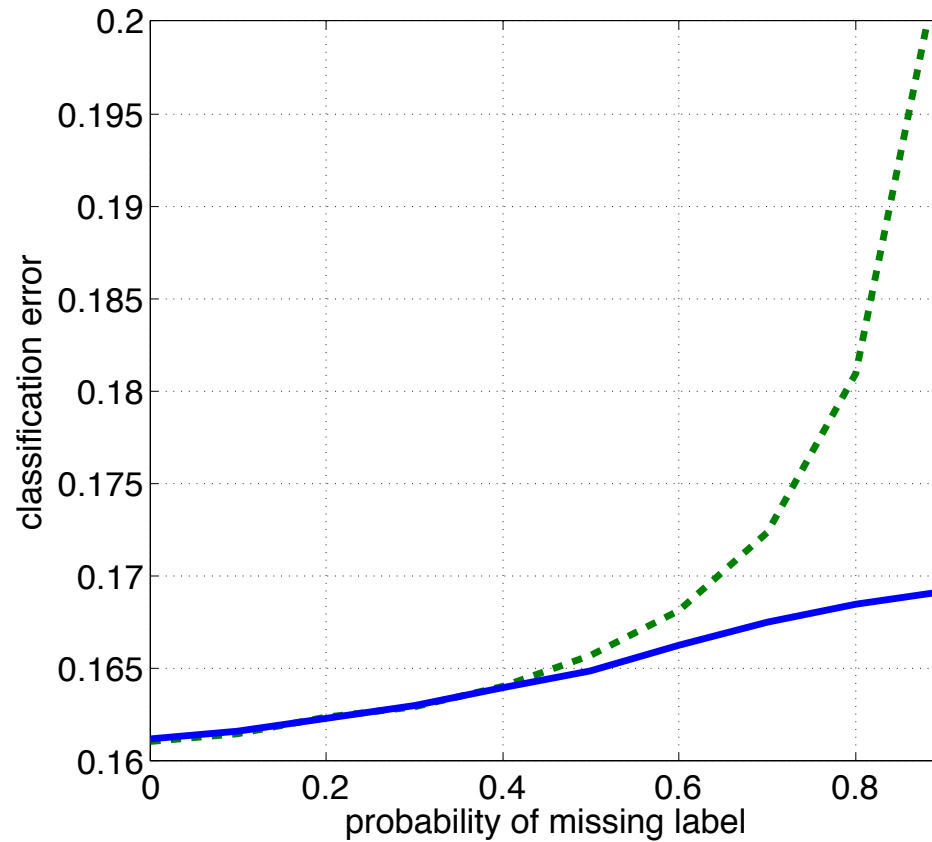


FIRST EXPERIMENT

- Class information was partly removed from the training instances.
- More specifically, each of the 200 instances was declared „unlabeled“ with a fixed probability γ .
- Thus, we are in a **semi-supervised setting**, in which approximately $200(1-\gamma)$ of the instances are labeled.
- In our approach, the unlabeled instances are considered as being labeled with the fuzzy set that assigns a membership degree of 1 to both the positive and the negative class.
- Then, a model is trained using the fuzzy log-loss.
- Standard logistic regression is used for comparison.

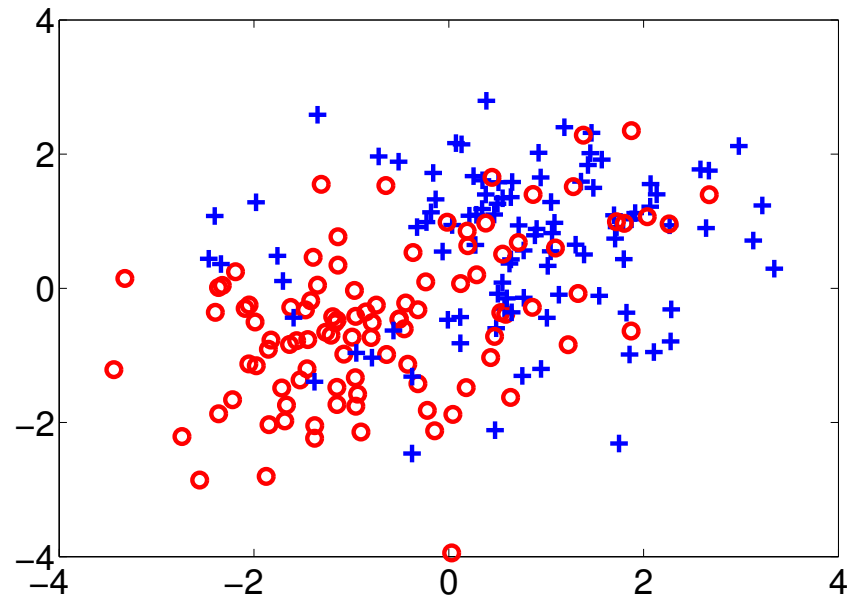
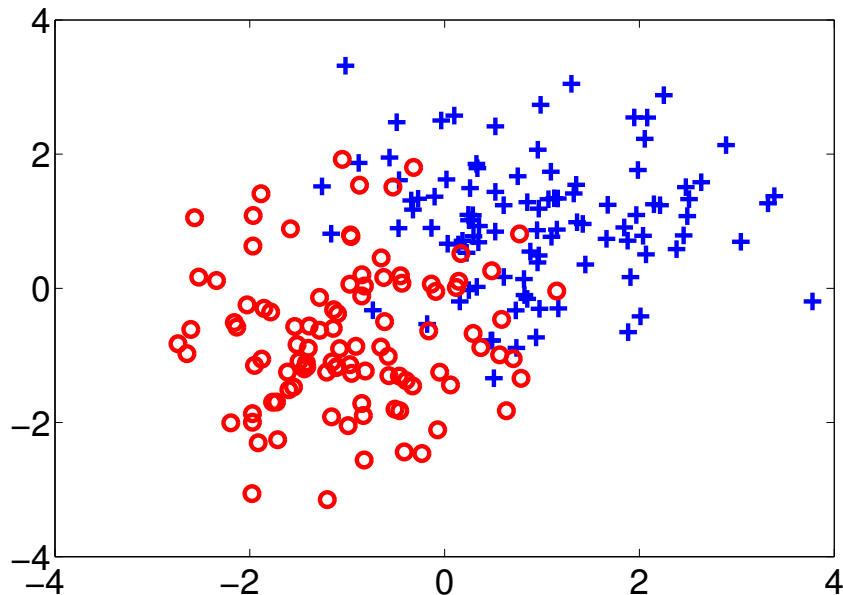


FIRST EXPERIMENT



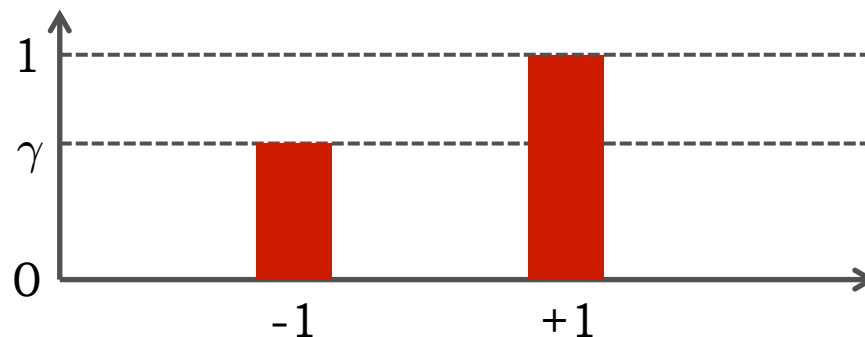
SECOND EXPERIMENT

- The label of each example is switched with a fixed probability γ .
- This noise level is supposed to be known, whereas for each individual training example, it is not known whether the observed label corresponds to the original one or has been switched.

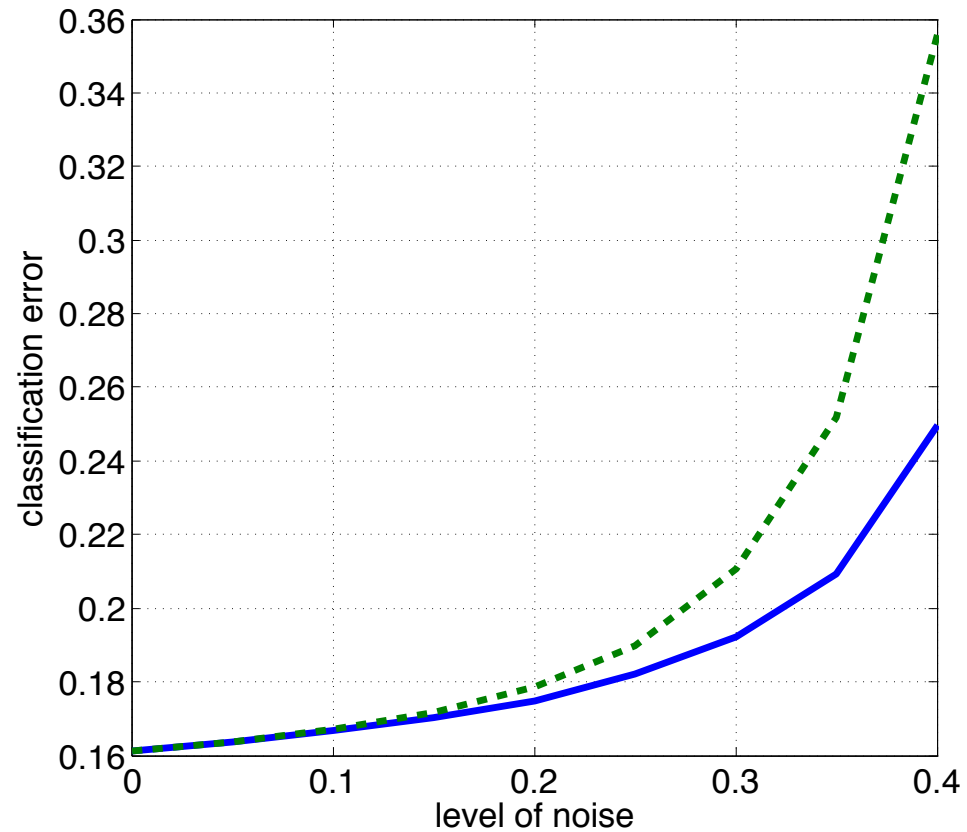


SECOND EXPERIMENT

- The label of each example is switched with a fixed probability γ .
- This noise level is supposed to be known, whereas for each individual training example, it is not known whether the observed label corresponds to the original one or has been switched.
- We model the label information in terms of a fuzzy set with a membership degree of 1 to the observed and of γ to the other label.
- Standard logistic regression simply uses the observed label information, which is the best it can do.

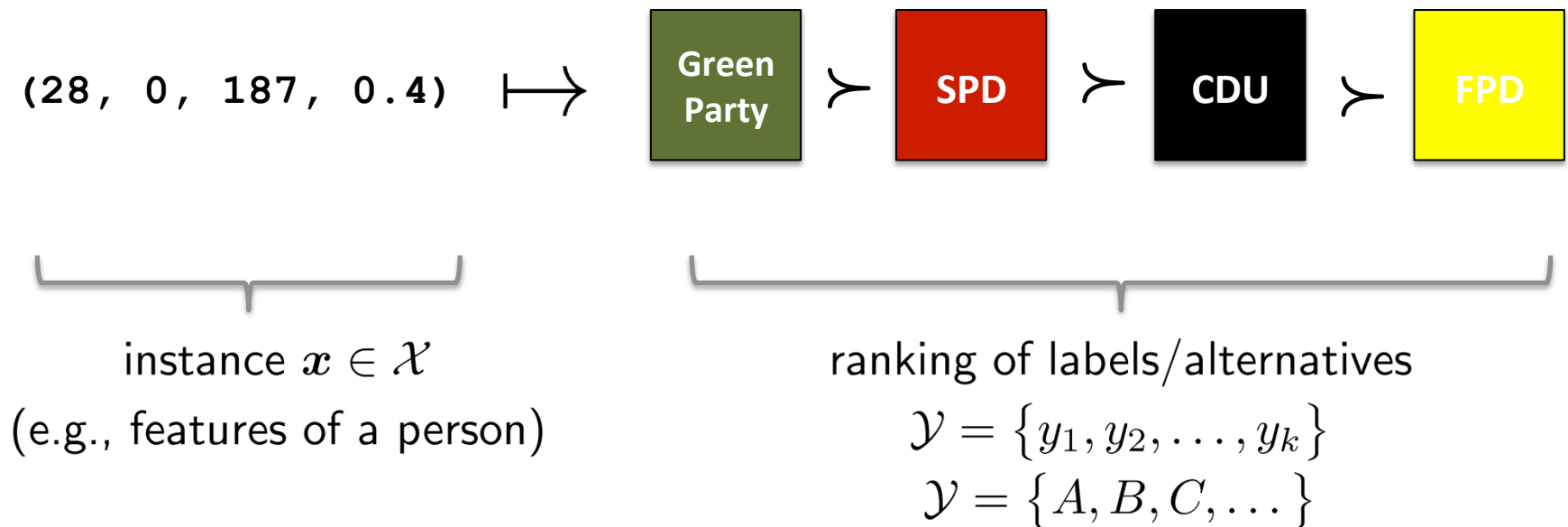


SECOND EXPERIMENT



LABEL RANKING

... mapping instances to **TOTAL ORDERS** over a fixed set of alternatives/labels:




LABEL RANKING: TRAINING DATA

TRAINING

X1	X2	X3	X4	preferences
0.34	0	10	174	$A \succ B, C \succ D$
1.45	0	32	277	$B \succ C$
1.22	1	46	421	$B \succ D, A \succ D, C \succ D, A \succ C$
0.74	1	25	165	$C \succ A, C \succ D, A \succ B$
0.95	1	72	273	$B \succ D, A \succ D$
1.04	0	33	158	$A \succ B, A \succ C$

Instances are associated with pairwise preferences between labels

- 
- rank of A between 1 and 2
 - rank of B between 2 and 4
 - rank of C between 2 and 4
 - rank of D between 1 and 4

LABEL RANKING

Performance in terms of Kendall's tau on synthetic data: missing-at-random (above) and top-rank setting (below).

	complete ranking		30% missing labels		60% missing labels	
	LWD	PL	LWD	PL	LWD	PL
authorship	.933±.016	.936±.015	.925±.018	.833±.030	.891±.021	.601±.054
glass	.840±.075	.841±.067	.819±.078	.669±.064	.721±.072	.395±.068
iris	.960±.036	.960±.036	.932±.051	.896±.069	.876±.068	.787±.111
pendigits	.940±.002	.939±.002	.924±.002	.770±.004	.709±.005	.434±.007
segment	.953±.006	.950±.005	.914±.009	.710±.013	.624±.020	.381±.020
vehicle	.853±.031	.859±.028	.836±.032	.753±.032	.767±.037	.520±.050
vowel	.876±.021	.851±.020	.821±.022	.612±.027	.536±.034	.327±.033
wine	.938±.050	.947±.047	.933±.054	.919±.059	.921±.062	.863±.094
authorship	.933±.016	.936±.015	.932±.017	.927±.017	.923±.015	.886±.022
glass	.840±.075	.841±.067	.838±.074	.809±.066	.815±.075	.675±.069
iris	.960±.036	.960±.036	.956±.036	.926±.051	.932±.048	.868±.070
pendigits	.940±.002	.939±.002	.933±.002	.918±.002	.837±.004	.794±.004
segment	.953±.006	.950±.005	.943±.005	.874±.008	.844±.010	.674±.015
vehicle	.853±.031	.859±.028	.851±.033	.838±.030	.818±.032	.765±.035
vowel	.876±.021	.851±.020	.867±.021	.785±.020	.800±.021	.588±.024
wine	.938±.050	.947±.047	.936±.049	.926±.061	.930±.059	.907±.066

SUMMARY AND CONCLUSION

- **Learning from fuzzy data** is gaining increasing attention.
- Different **interpretations** of fuzzy data exist and suggest different ways of extending machine learning and data analysis methods:

ontic interpretation → data reproduction

epistemic interpretation → **data disambiguation** (~~extension principle~~)

- We proposed a method based on **generalized (fuzzy) loss functions and risk minimization**: A fuzzy set properly „modulates“ the loss associated with an individual observation → **data modeling**
- Our framework covers several existing approaches as **special cases** (Huber loss, instance weighting, semi-supervised learning), but also supports the systematic development of **new methods**.

SUMMARY AND CONCLUSION

E. Hüllermeier. **Learning from Imprecise and Fuzzy Observations: Data Disambiguation through Generalized Loss Minimization.**
International Journal of Approximate Reasoning (to appear).

Preprint version: `arXiv:1305.0698`