Jasper De Bock, Adrián Detavernier, Rodrigo Lassance

Robustness Quantification

using imprecise probabilities to assess the reliability of probabilistic classifiers

GHENT UNIVERSITY

WUML 2026
February 3




FLip

Foundations Lab for imprecise probabilities

GHENT UNIVERSITY



Adrián
Detavernier

Rodrigo
Lassance

MACHINE
LEARNING

IMPRECISE
PROBABILITIES

Adrián
Detavernier

Rodrigo
Lassance

MACHINE
LEARNING

IMPRECISE
PROBABILITIES

ROBUSTNESS
QUANTIFICATION

Adrián
Detavernier

Rodrigo
Lassance

MACHINE
LEARNING

IMPRECISE
PROBABILITIES

MACHINE LEARNING

… is unreliable


media saying AI will
take over the world


my neural network:
Dog

CLASSIFICATION

features x

FEATURES

photo

medical data

...

CLASSIFICATION

CLASSIFICATION

features x

predicted class $\hat{y} = h(x)$

try to predict

FEATURES → **CLASS**

photo

medical data

...

label

best treatment

...

PROBABILISTIC CLASSIFIER h

predict the class y that maximizes $P_{\text{classif}}(y \mid x)$

CLASSIFICATION

features x

predicted class $\hat{y} = h(x)$

try to predict

FEATURES → **CLASS**

photo

medical data

...

DISCRIMINATIVE PROBABILISTIC CLASSIFIER h

label

best treatment

...

learn $P_{\text{classif}}(Y \mid x)$

predict the class y that maximizes $P_{\text{classif}}(y \mid x)$

CLASSIFICATION

features x

predicted class $\hat{y} = h(x)$

try to predict

FEATURES → **CLASS**

photo

medical data

...

label

best treatment

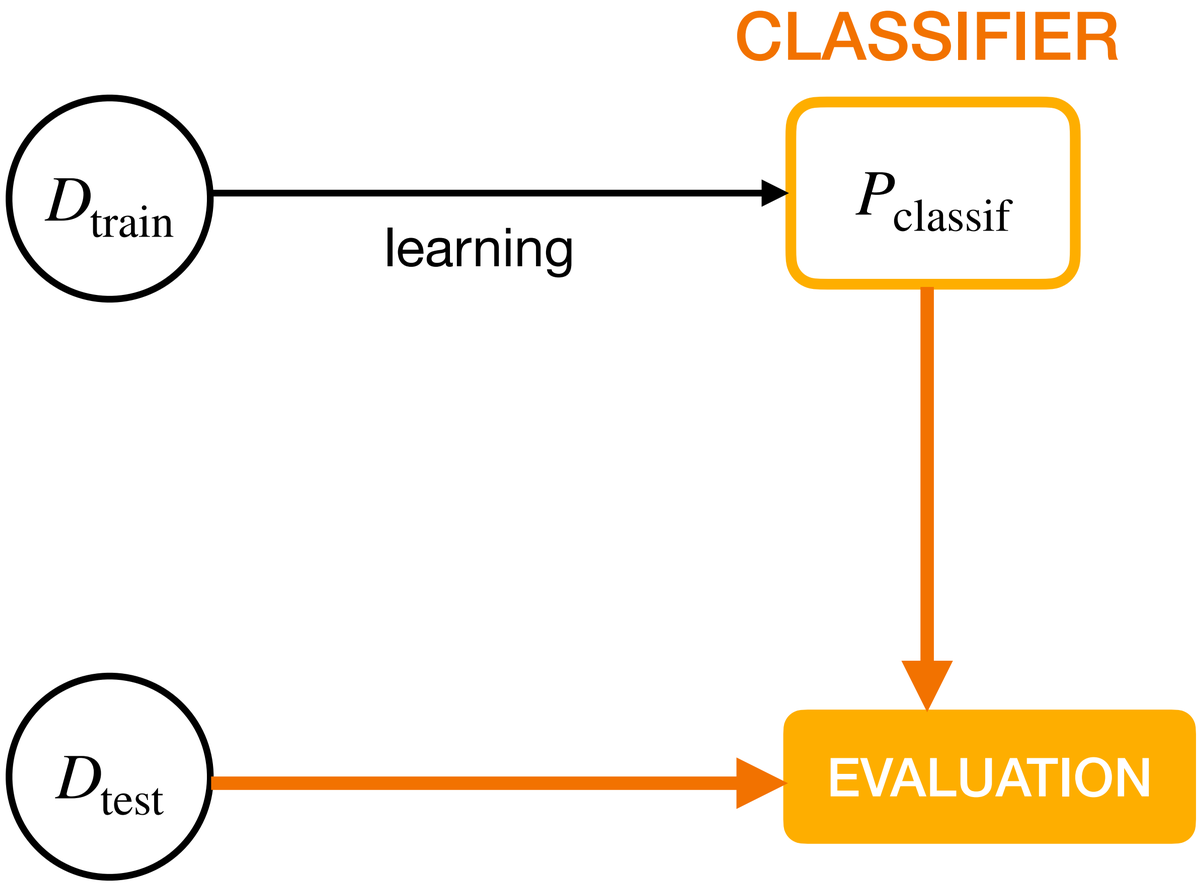
...

GENERATIVE PROBABILISTIC CLASSIFIER h

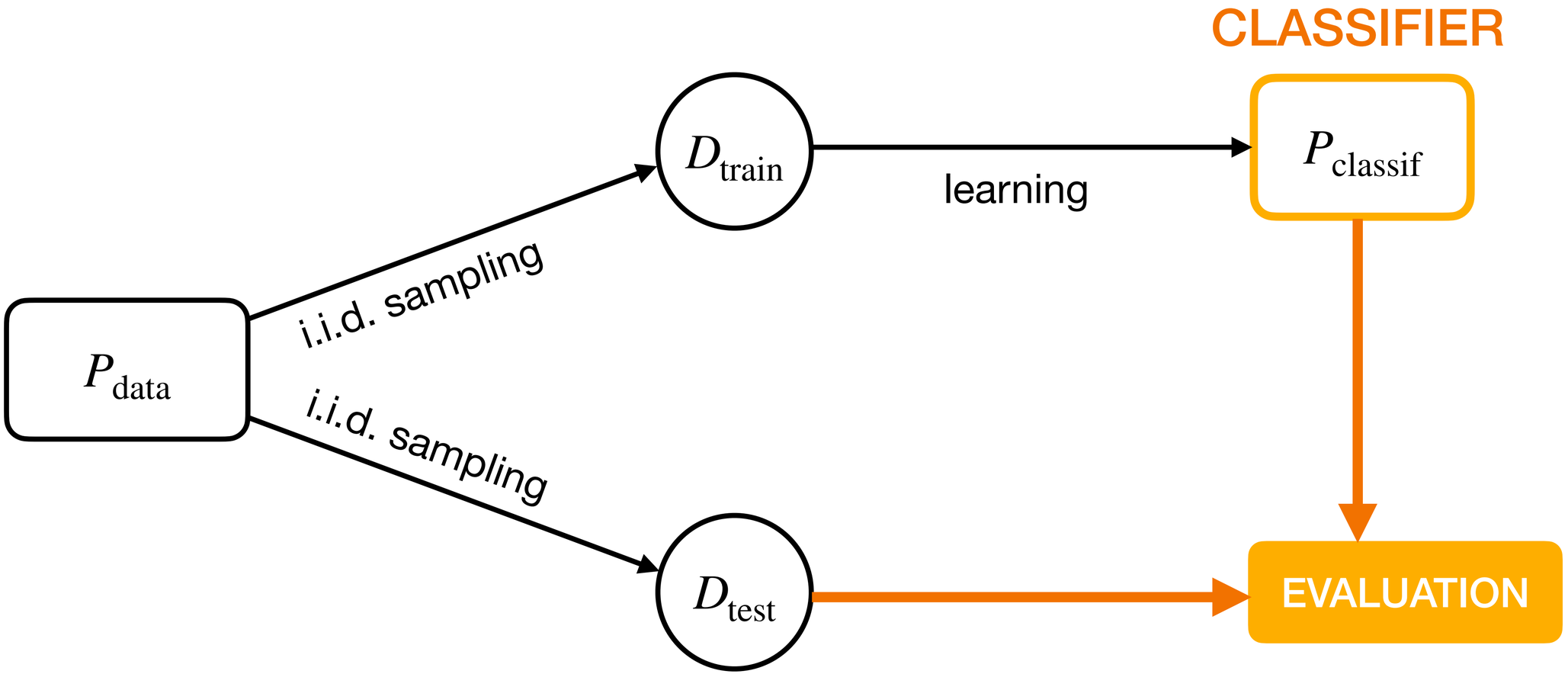
learn $P_{\text{classif}}(X, Y)$

predict the class y that maximizes $P_{\text{classif}}(y \mid x)$

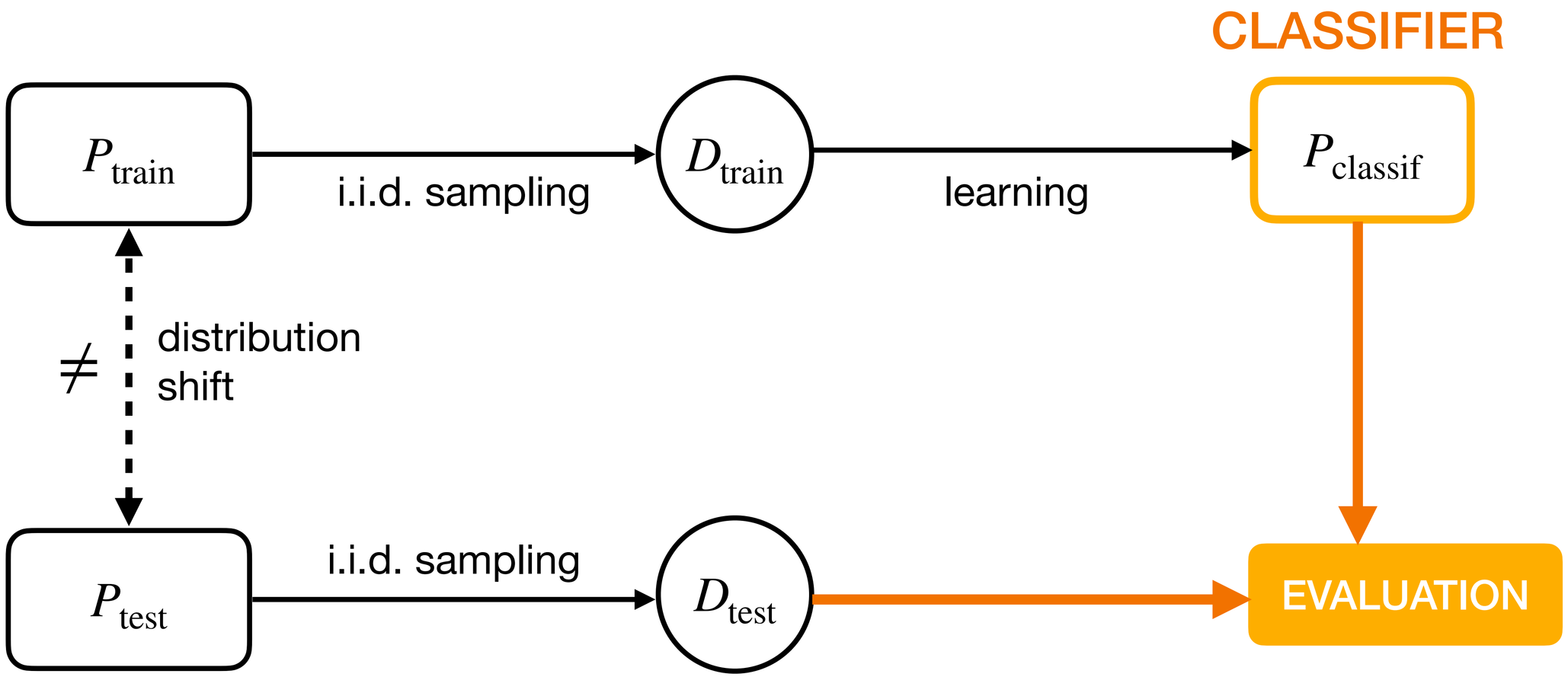
CLASSIFICATION



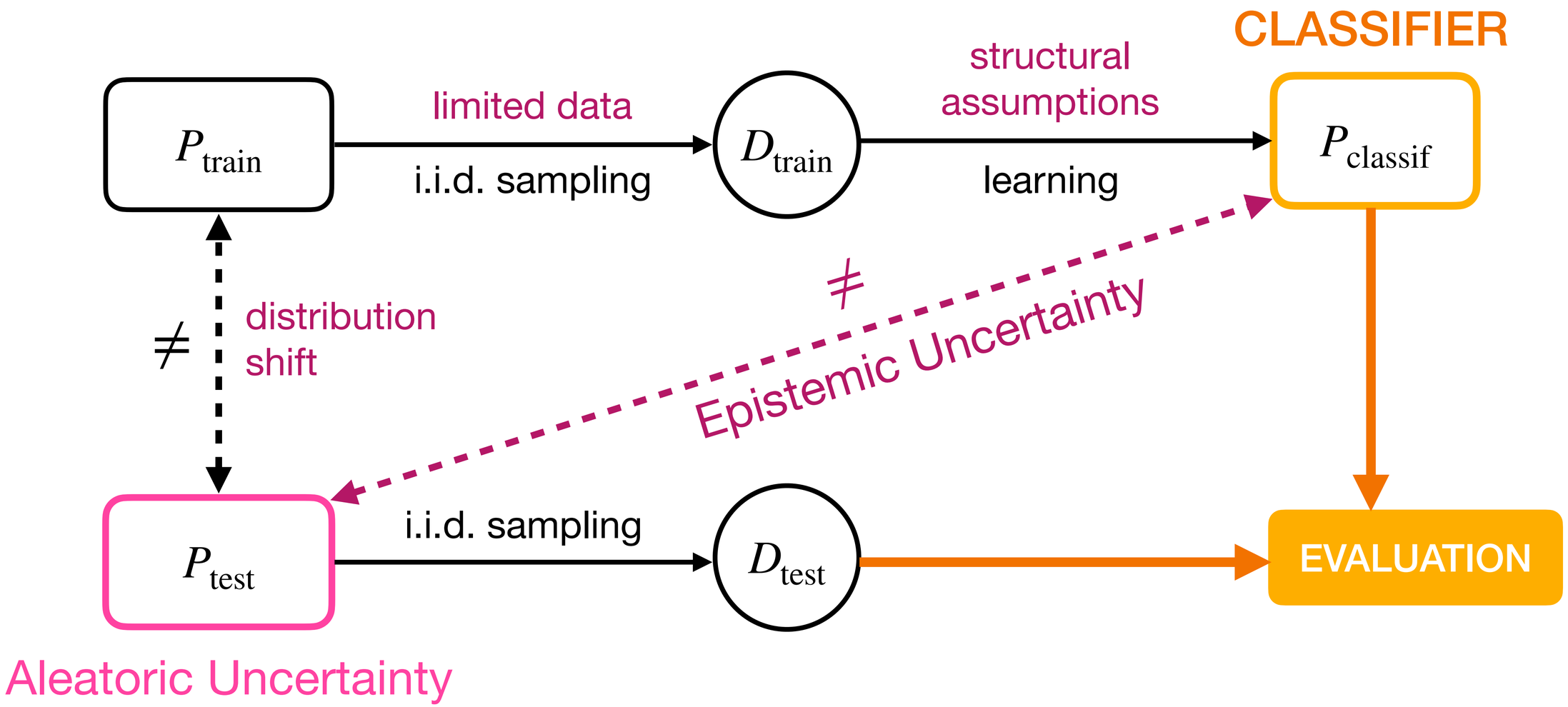
CLASSIFICATION



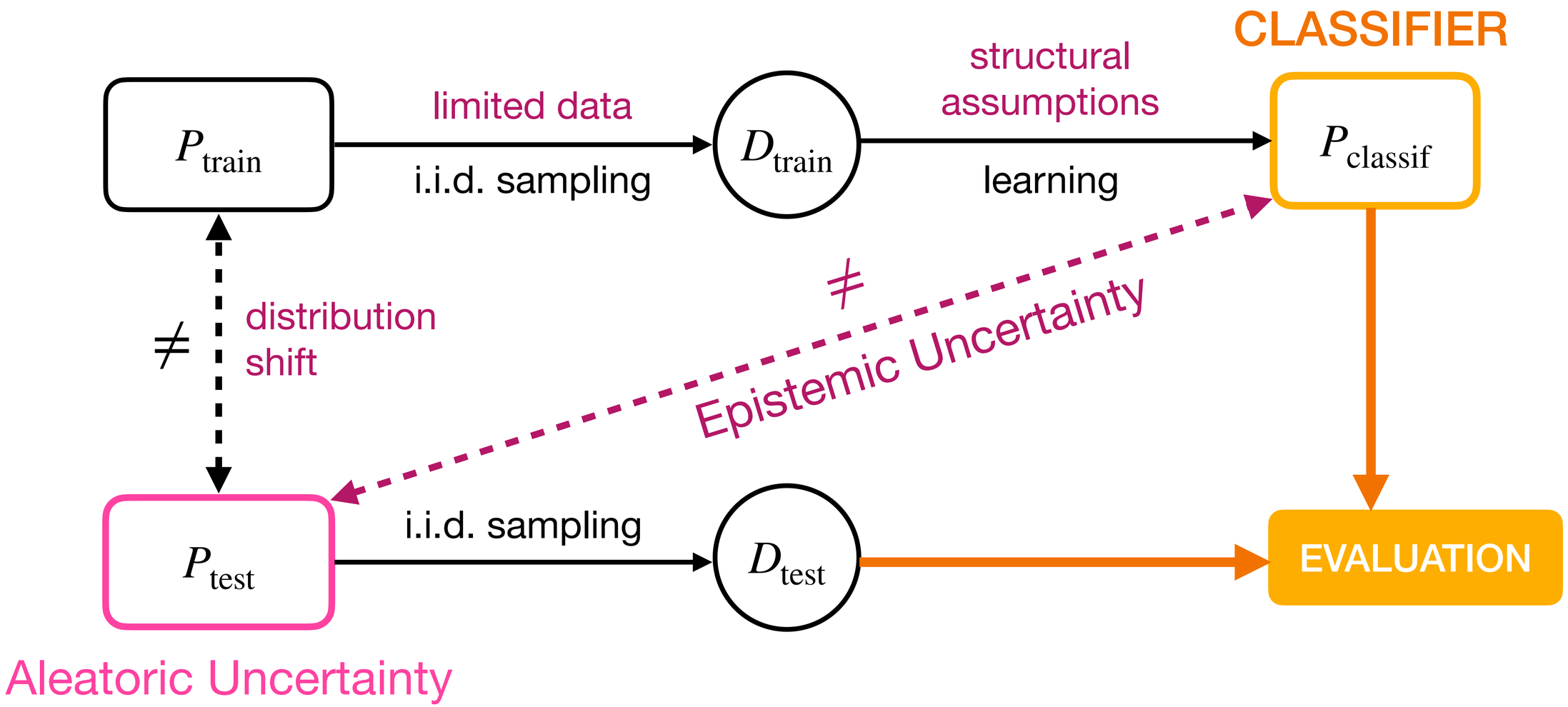
CLASSIFICATION



CLASSIFICATION … is unreliable



UNCERTAINTY QUANTIFICATION



MACHINE
LEARNING

IMPRECISE
PROBABILITIES

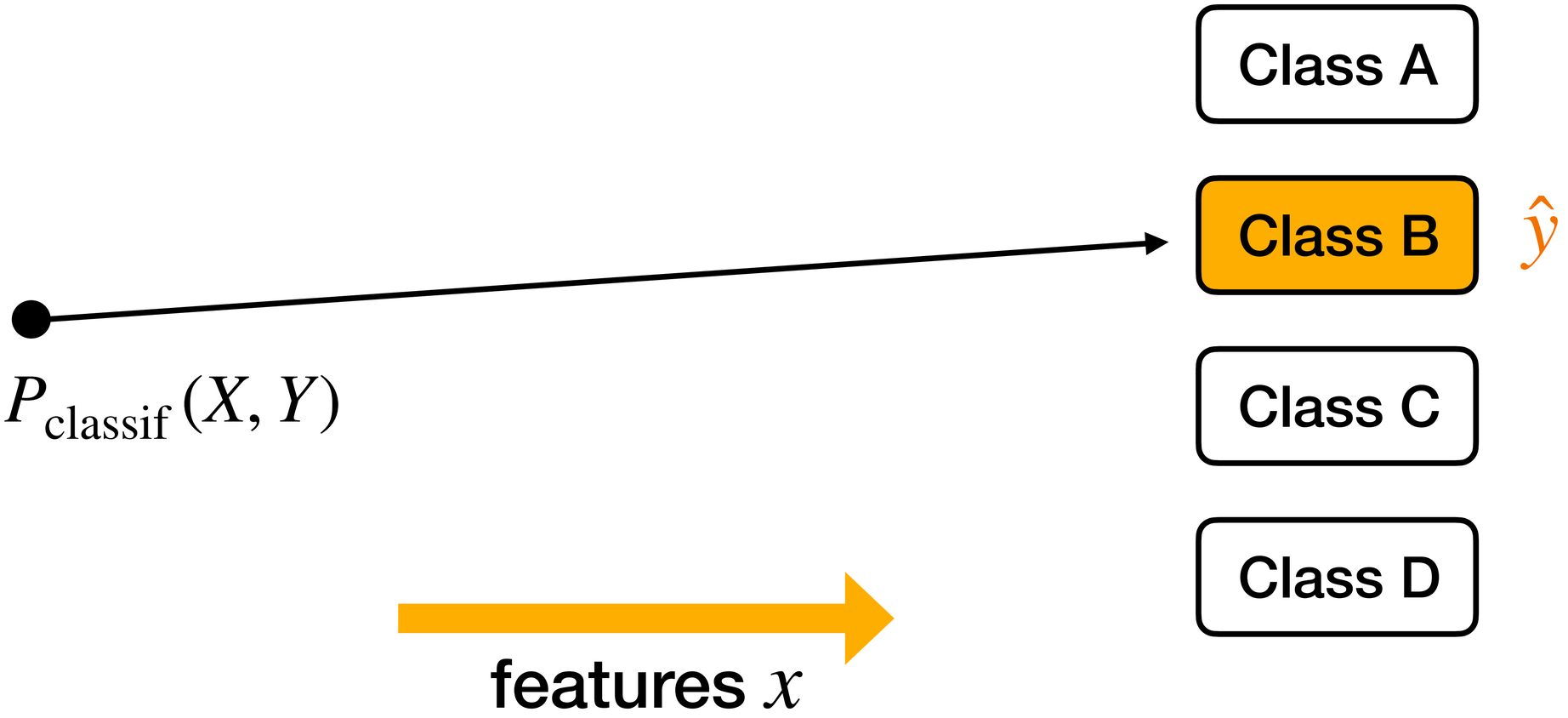
ROBUSTNESS
QUANTIFICATION

IMPRECISE PROBABILITIES

enable us to study robustness

PROBABILISTIC CLASSIFIER

predict the class y that
maximizes $P_{\text{classif}}(y \mid x)$



PROBABILISTIC CLASSIFIER

predict the class y that
maximizes $P_{\text{classif}}(y \mid x)$

ROBUST prediction

set of distributions
$\mathscr{P}$ that contains
$P_{\text{classif}}(X, Y)$

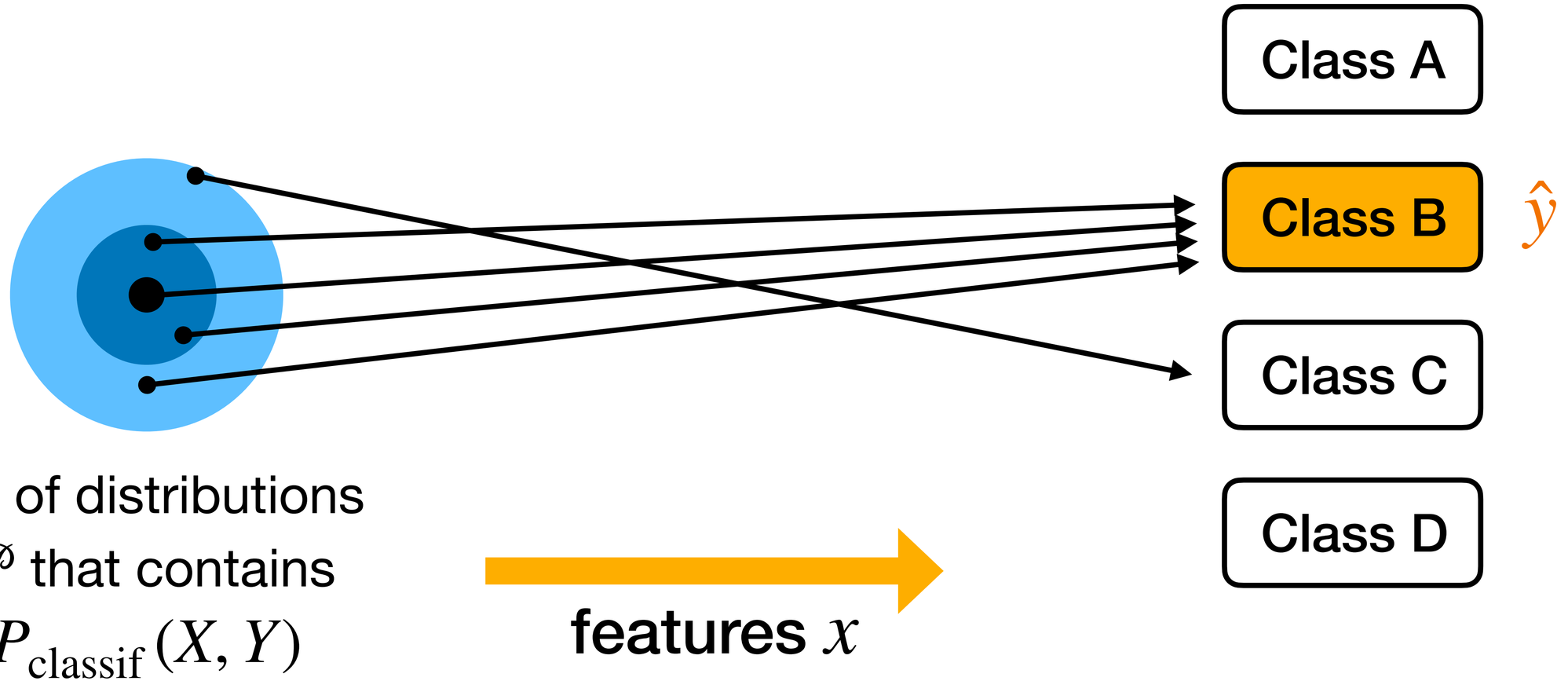
features x

- Class A
- Class B → $\hat{y}$
- Class C
- Class D

PROBABILISTIC CLASSIFIER

predict the class y that
maximizes $P_{\text{classif}}(y \mid x)$

~~ROBUST~~ prediction

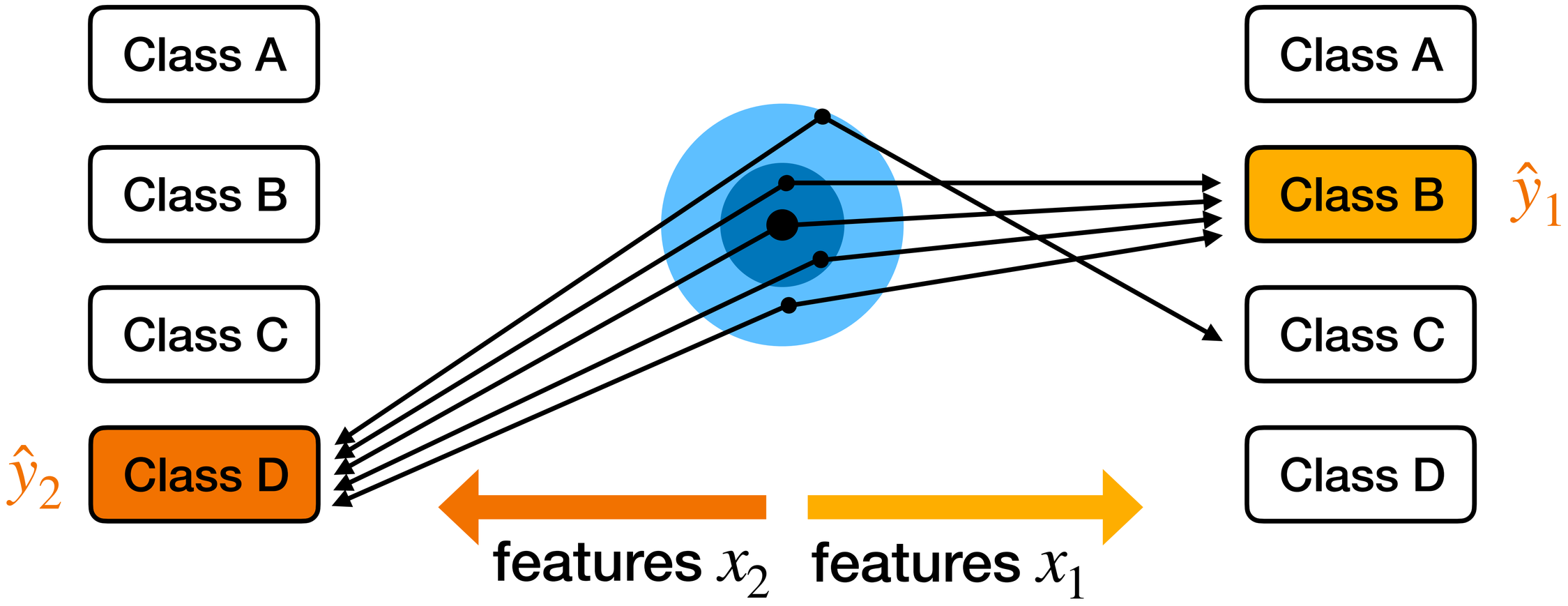


ROBUST prediction

PROBABILISTIC CLASSIFIER

predict the class y that maximizes $P_{\text{classif}}(y \mid x)$

~~ROBUST prediction~~

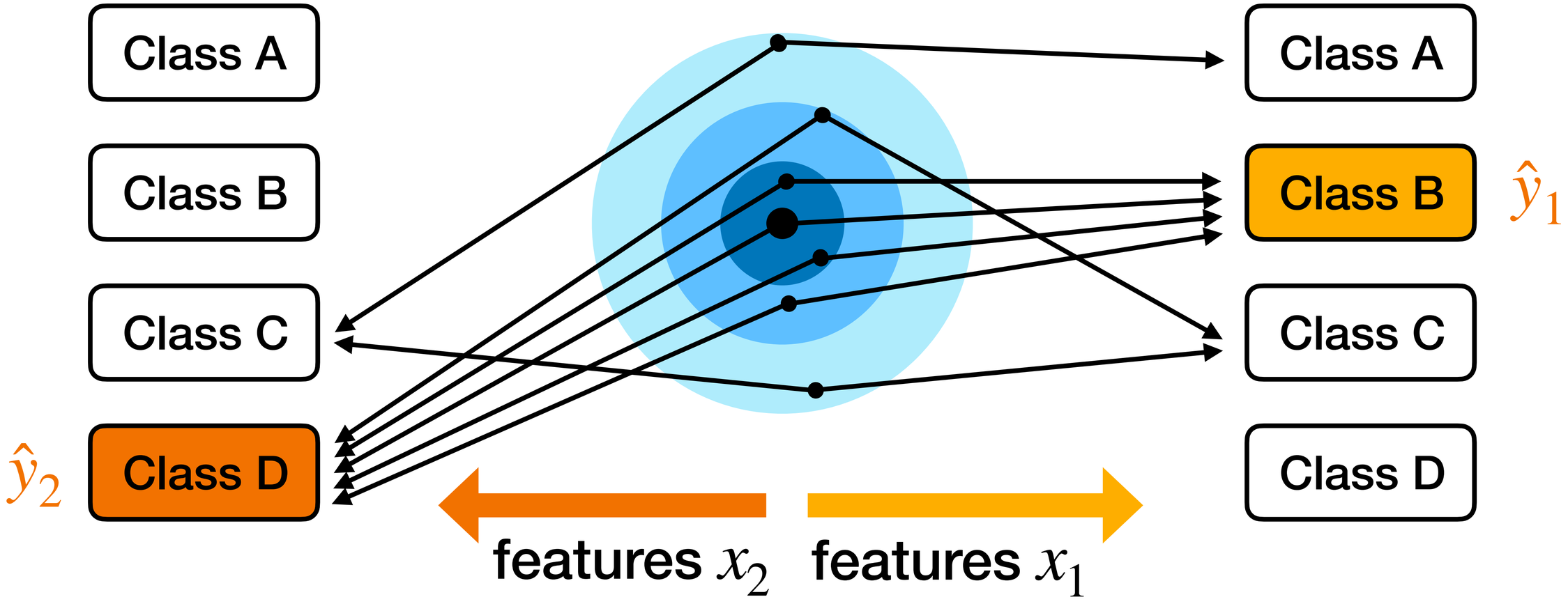


ROBUST prediction

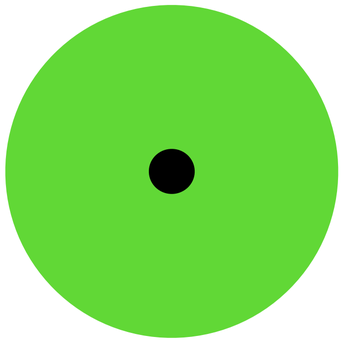
PROBABILISTIC CLASSIFIER

predict the class y that
maximizes $P_{\text{classif}}(y \mid x)$

ROBUST prediction



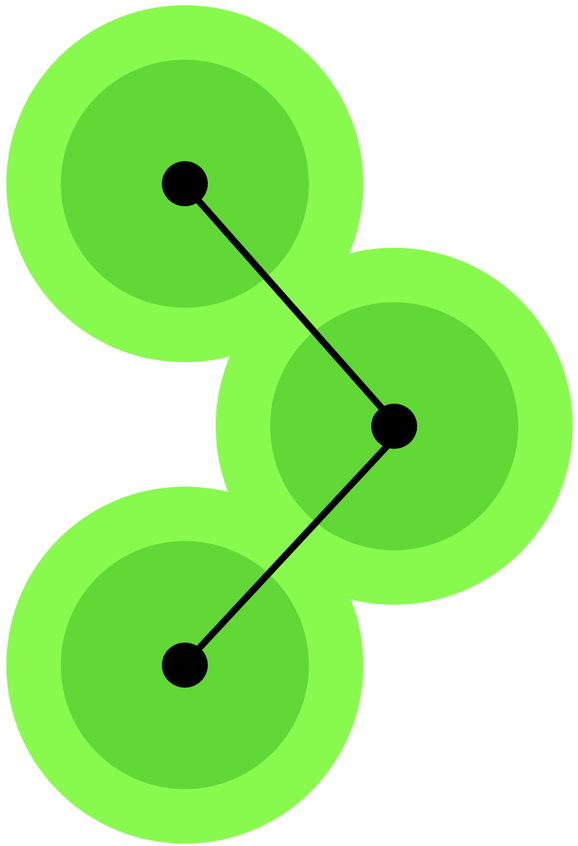
ROBUSTNESS QUANTIFICATION

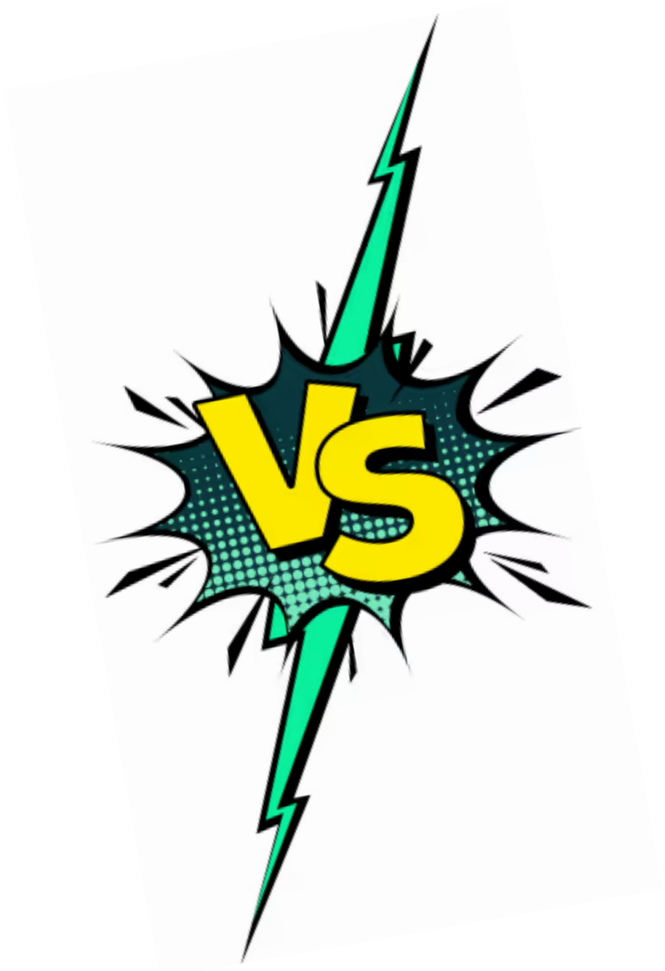


set of distributions
$\mathscr{P}$ that contains
$P_{\text{classif}}(X, Y)$

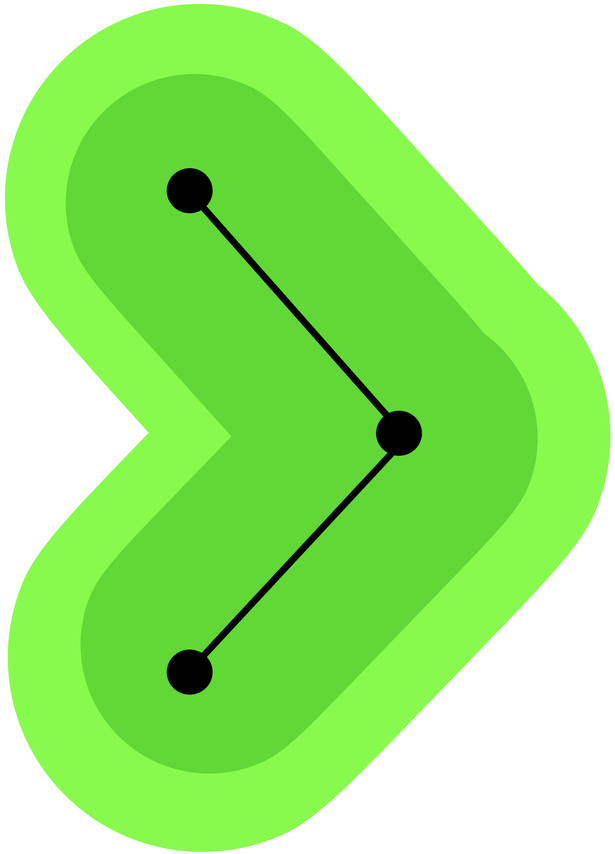
ROBUSTNESS:
"size" of largest $\mathscr{P}$ for which that particular prediction is robust

LOCAL

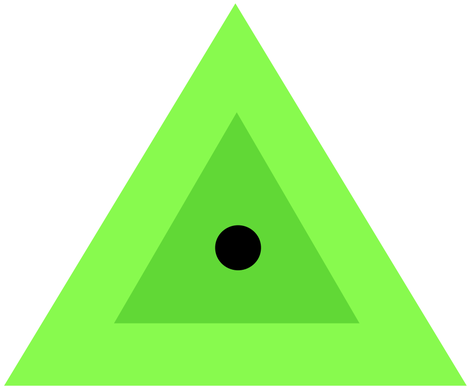




GLOBAL



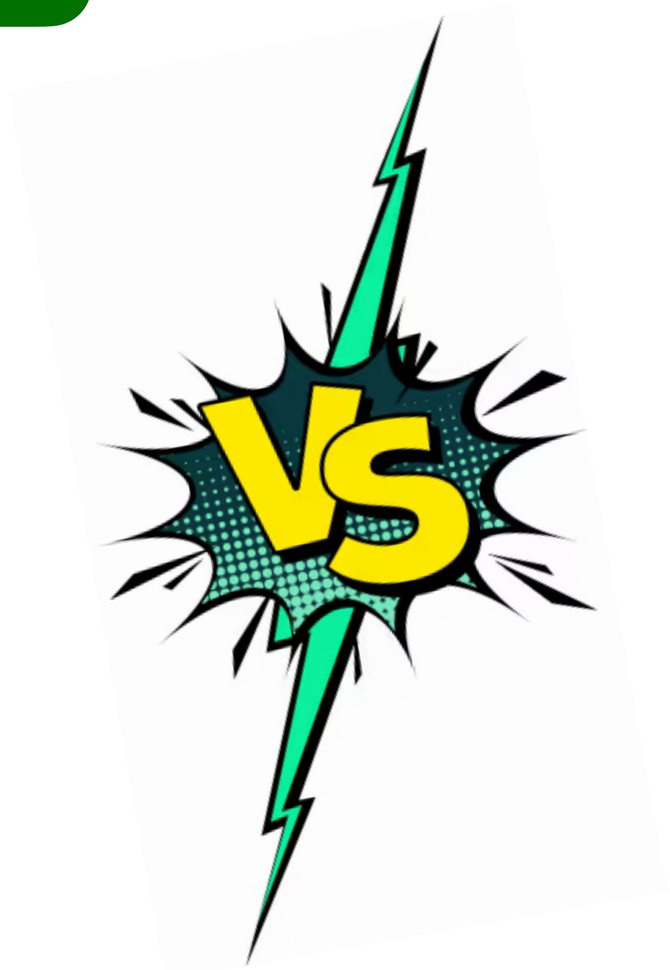
ϵ-CONTAMINATION



$\mathscr{P}_\epsilon$

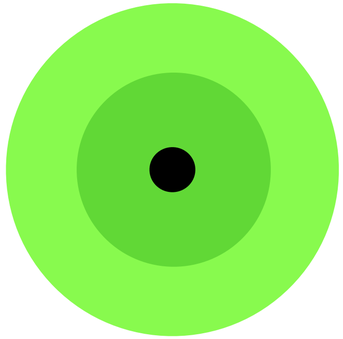
$\|$

$\{(1-\epsilon)P_{\text{classif}} + \epsilon P : P \in \mathbb{P}\}$



OTHER STUFF

distance-based, …

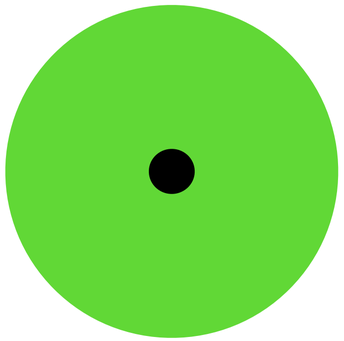


$\mathscr{P}_\delta$

$\|$

$\{P \in \mathbb{P} : \mathrm{d}(P_{\text{classif}}, P) < \delta\}$

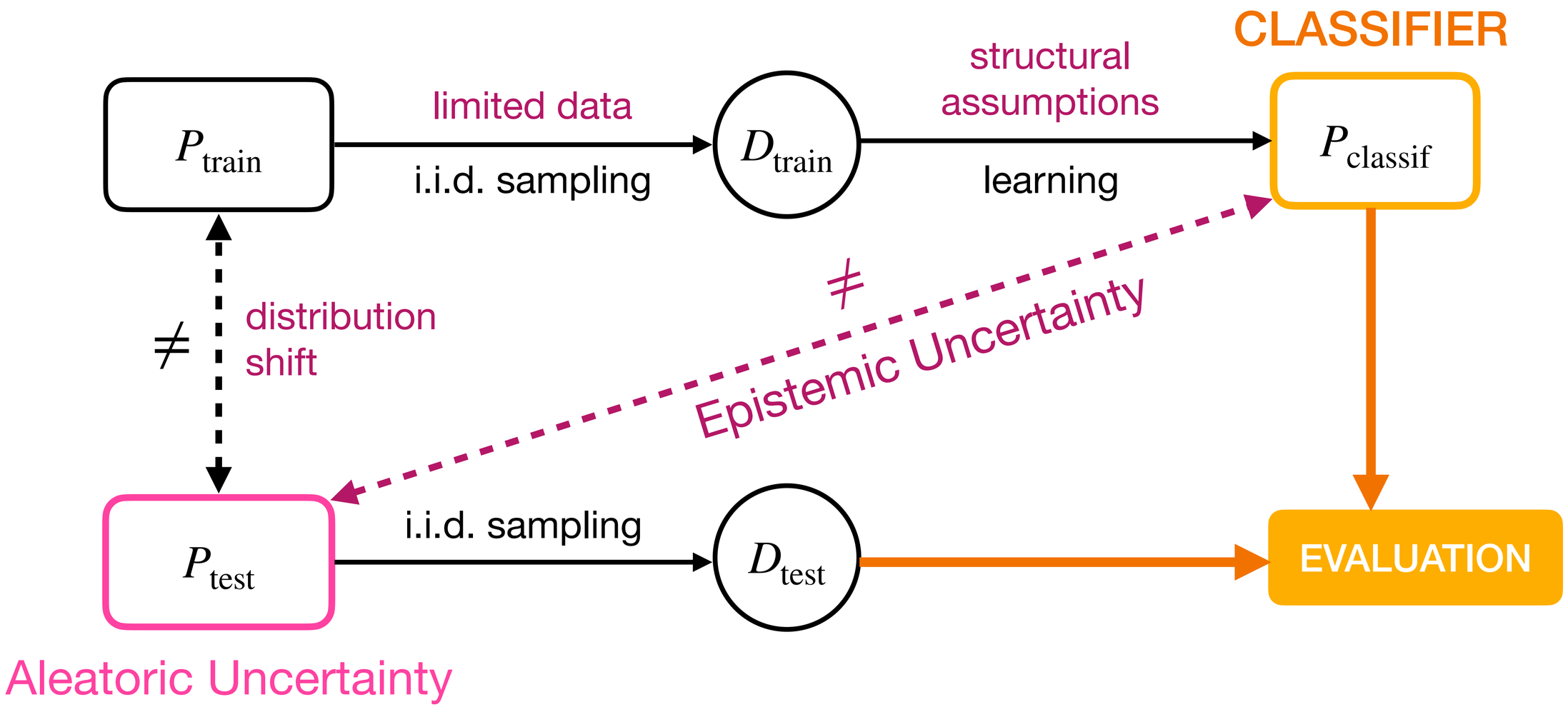
ROBUSTNESS QUANTIFICATION



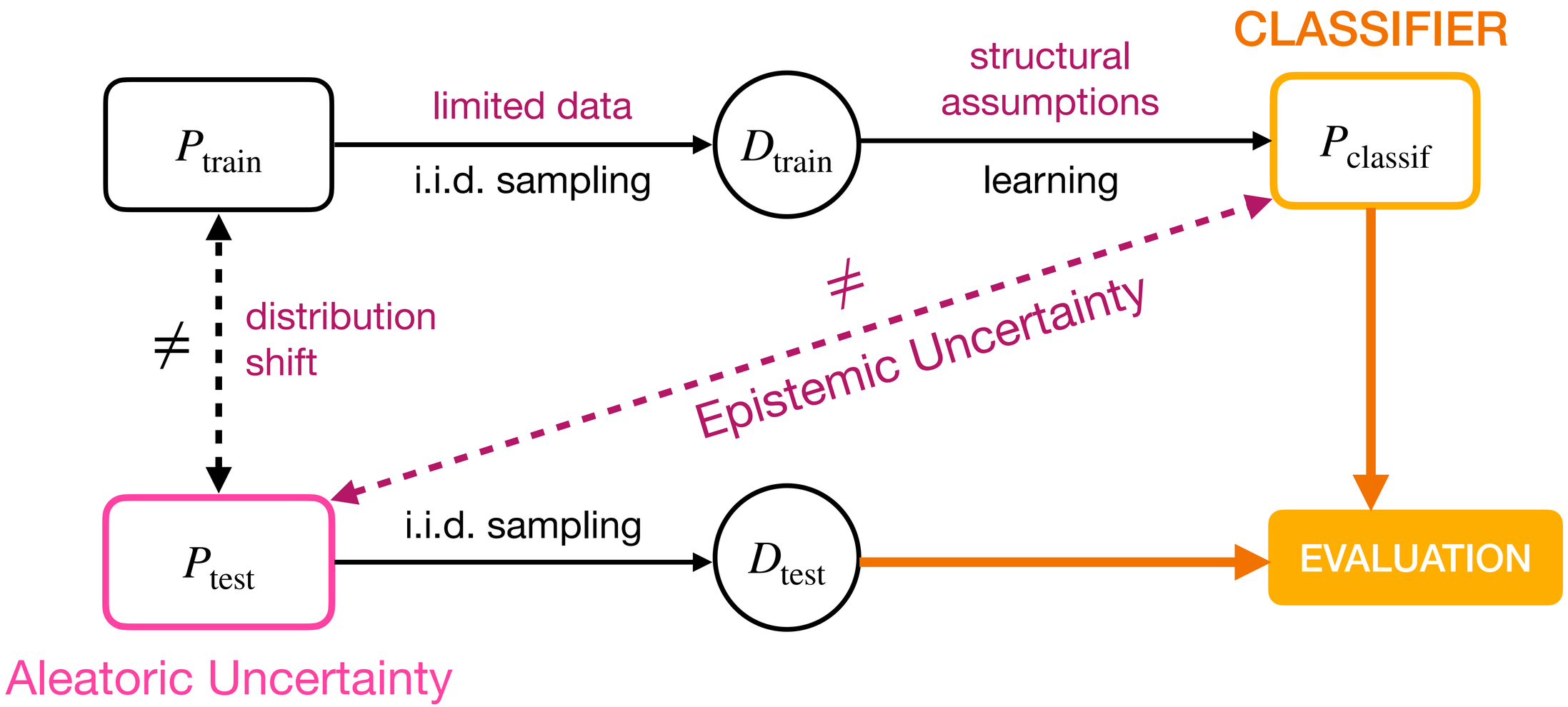
set of distributions
$\mathscr{P}$ that contains
$P_{\text{classif}}(X, Y)$

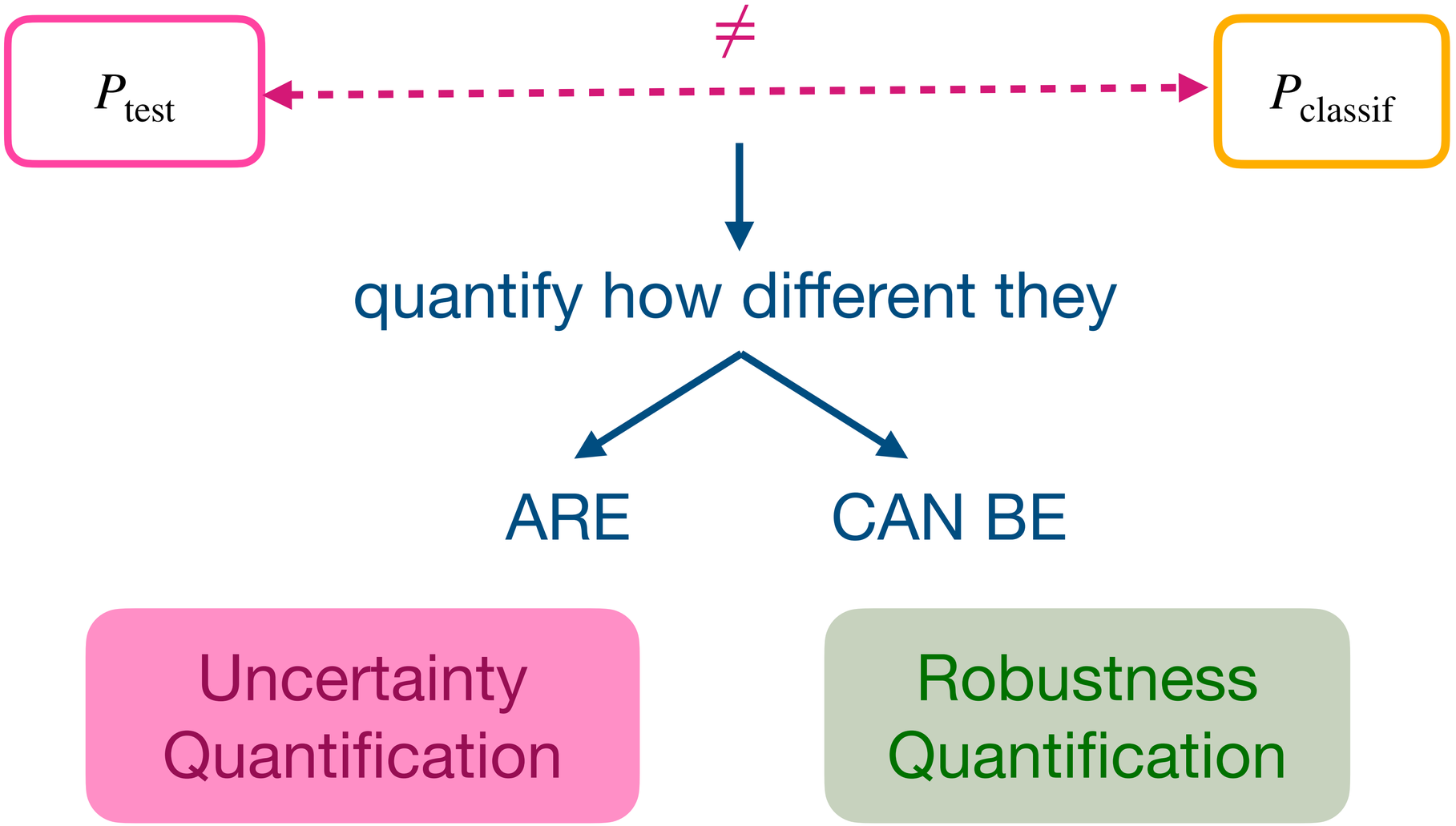
ROBUSTNESS:
"size" of largest $\mathscr{P}$ for which that particular prediction is robust

UNCERTAINTY QUANTIFICATION



ROBUSTNESS QUANTIFICATION




P_{test}
$\neq$
P_{classif}
quantify how different they
ARE
CAN BE
Uncertainty Quantification
Robustness Quantification

Adrián
Detavernier

Rodrigo
Lassance

MACHINE
LEARNING

IMPRECISE
PROBABILITIES

ROBUSTNESS
QUANTIFICATION

Adrián
Detavernier

MACHINE
LEARNING

IMPRECISE
PROBABILITIES

ROBUSTNESS
QUANTIFICATION

2014

MRF

BN

Cassio de Campos

Alessandro Antonucci

[1]

Global Sensitivity Analysis for MAP Inference in Graphical Models

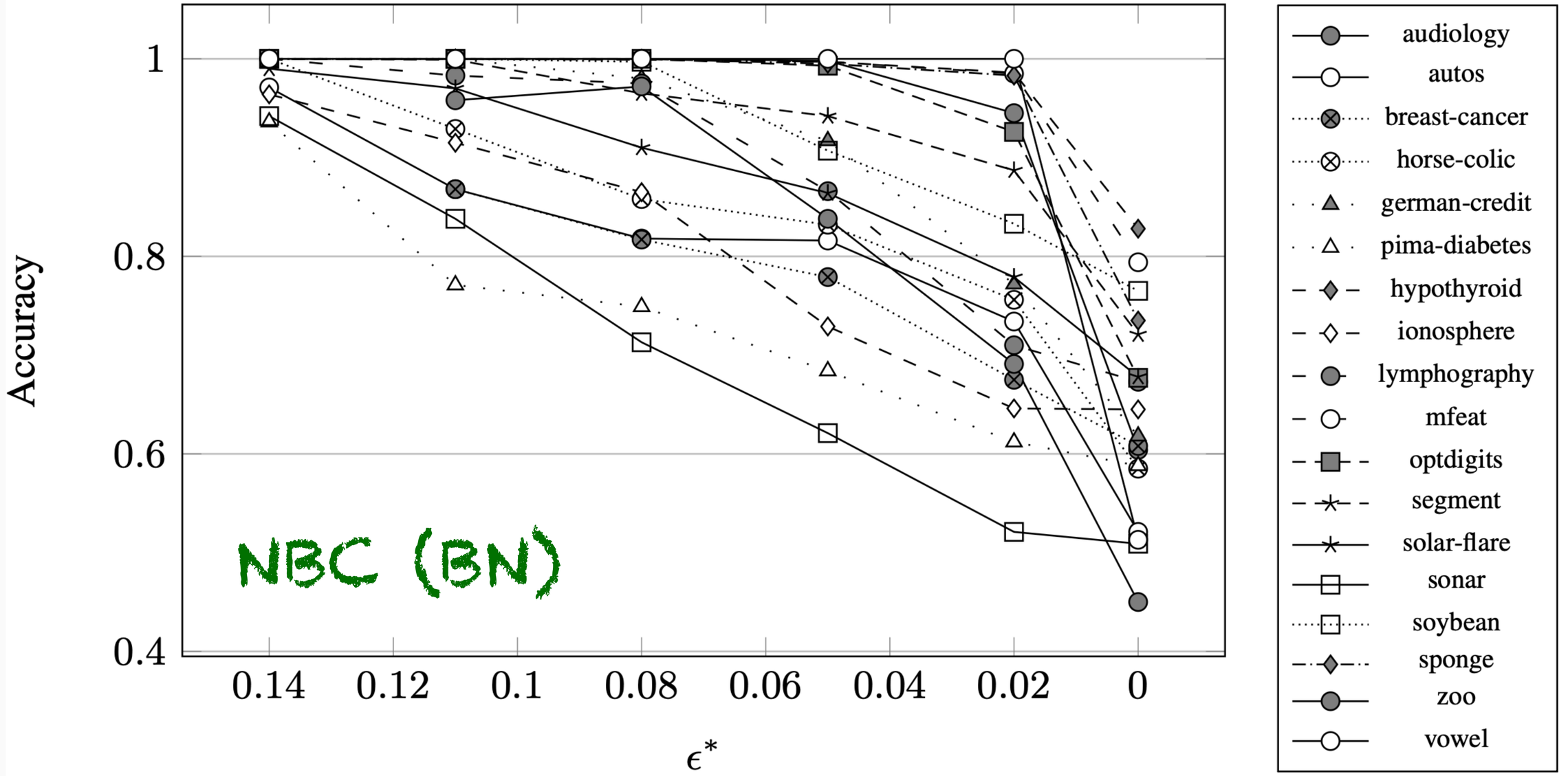
Jasper De Bock
Ghent University, SYSTeMS
Ghent (Belgium)
jasper.debock@ugent.be

Cassio P. de Campos
Queen's University
Belfast (UK)
c.decampos@qub.ac.uk

Alessandro Antonucci
IDSIA
Lugano (Switzerland)
alessandro@idsia.ch

Abstract

We study the sensitivity of a MAP configuration of a discrete probabilistic graphical model with respect to perturbations of its parameters. These perturbations are global, in the sense that simultaneous perturbations of all the parameters (or any chosen subset of them) are allowed. Our main contribution is an exact algorithm to check whether the MAP configuration is robust with respect to given perturbations. Its complexity is essentially the same as that of obtaining the MAP configuration once, and can be easily used with minimal effort. We use our


Accuracy
1
0.8
0.6
0.4
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
ϵ^*
NBC (BN)
audiology
autos
breast-cancer
horse-colic
german-credit
pima-diabetes
hypothyroid
ionosphere
lymphography
mfeat
optdigits
segment
solar-flare
sonar
soybean
sponge
zoo
vowel

[1]

2017-2020

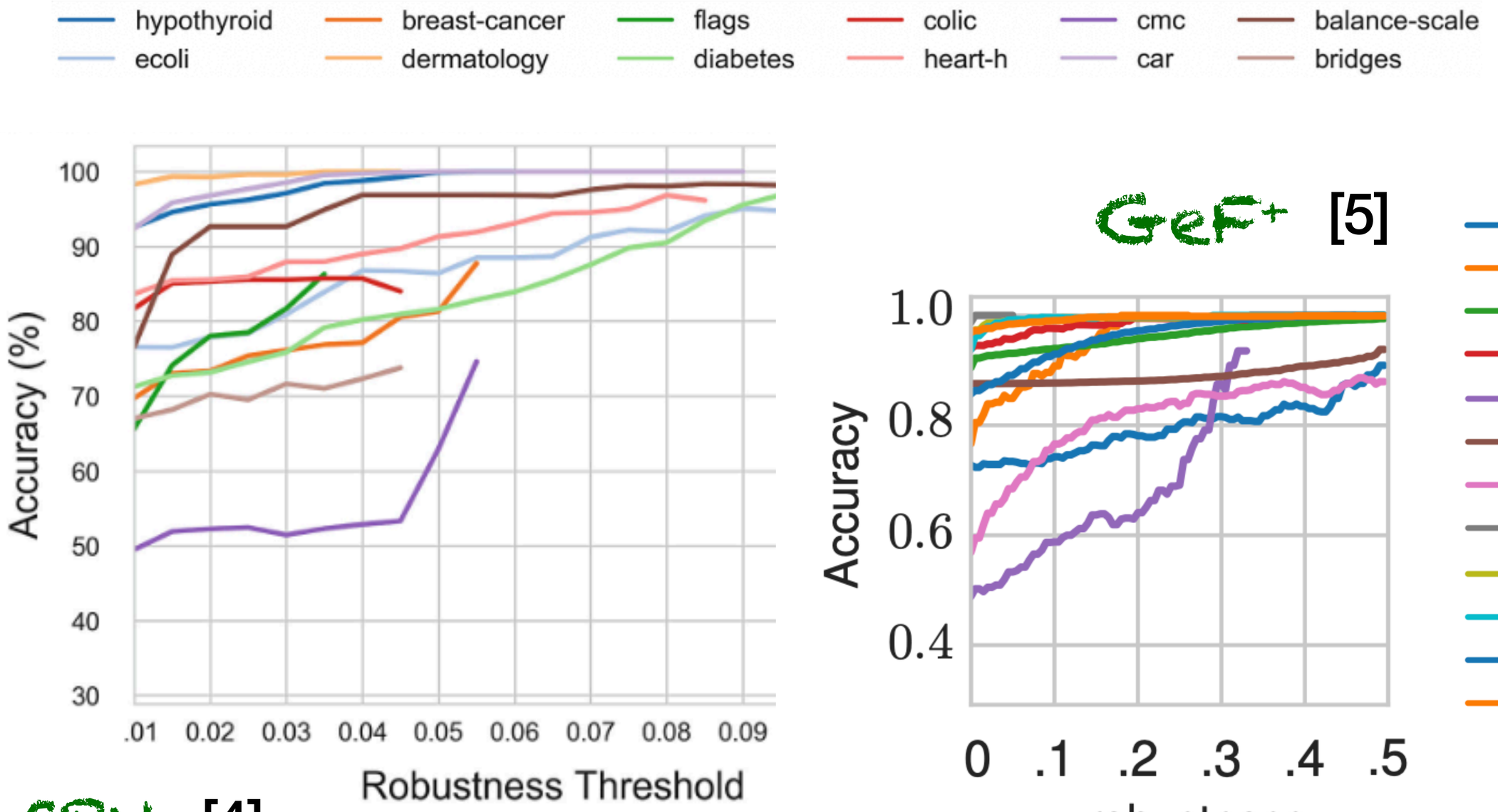
SPN

[2-4, ...]

[5]

GeF+

Cassio de Campos & various co-authors


hypothyroid
ecoli
breast-cancer
dermatology
flags
diabetes
colic
heart-h
cmc
car
balance-scale
bridges
100
90
80
70
60
50
40
30
Accuracy (%)
.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
Robustness Threshold

SPN [4]

GeF+ [5]

1.0
0.8
0.6
0.4
Accuracy
0
.1
.2
.3
.4
.5
ϵ-robustness
diabetes
german
bank
vowel
cmc
electricity
gesture
mice
texture
dna
jungle
phishing

ROBUSTNESS QUANTIFICATION

- correlates nicely with accuracy ✓
- works for different types of model architectures ✓

2025-...

[6, 7, ...]

Robustness quantification: a new method for assessing the reliability of the predictions of a classifier

Adrián Detavernier[1] **Jasper De Bock**[1]

[1]Foundations Lab for imprecise probabilities, Ghent University, Belgium

ness quantification compare in cases wh
data is limited or when there is a distrib
train and test data. Our motiv
by the fact that the
have a big

Robustness and uncertainty: two complementary aspects of the reliability of the predictions of a classifier

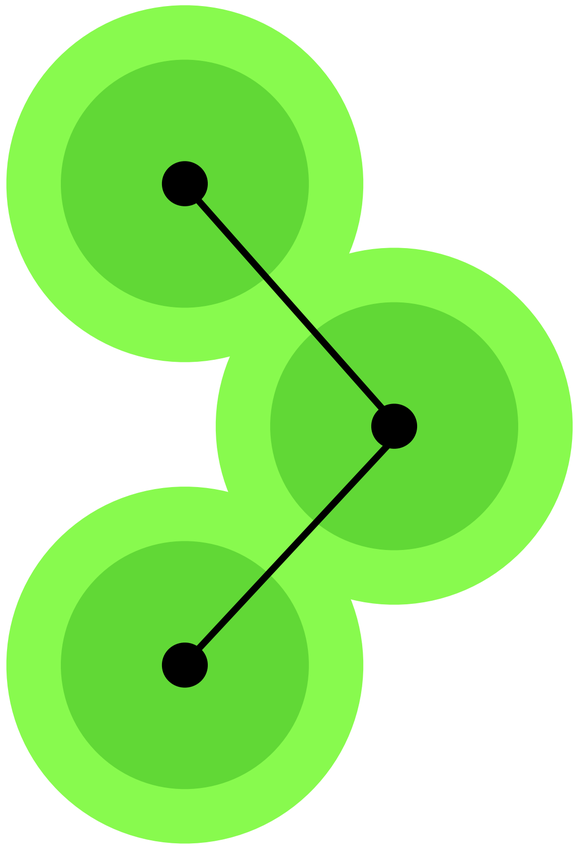
Adrián Detavernier
dations Lab for imprecise probabilities
Ghent University
Belgium

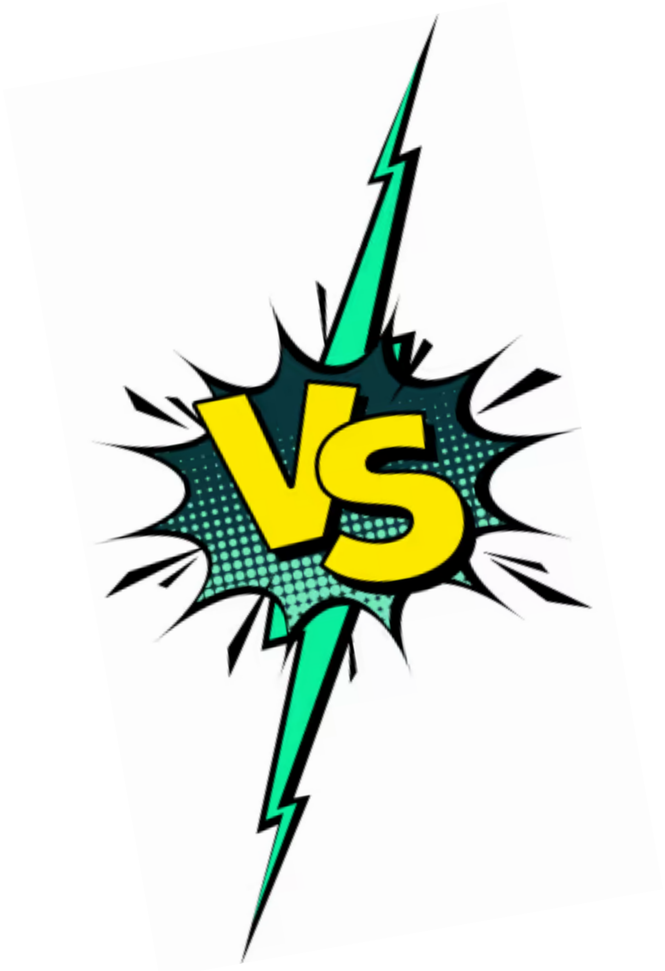
Jasper De Bock
Foundations Lab for imprecise probabilities
Ghent University
Belgium

Abstract

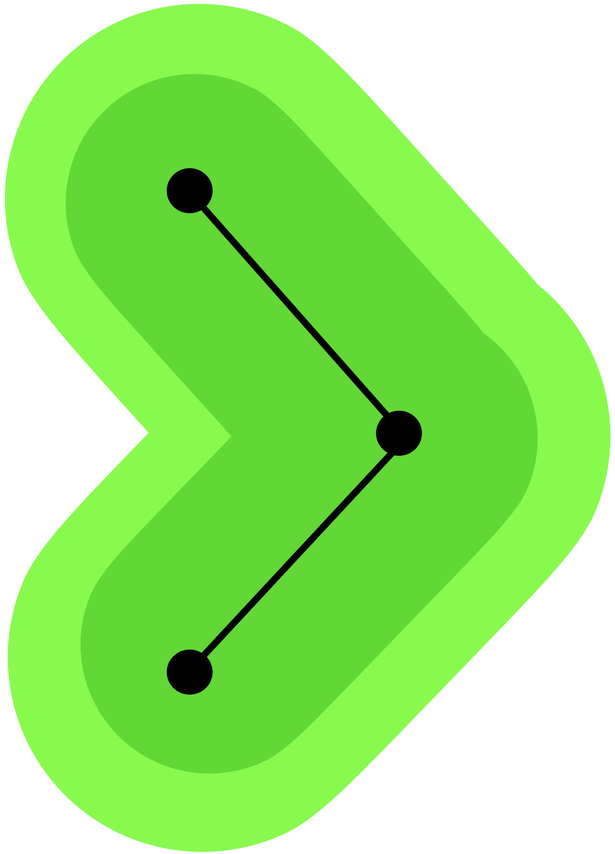
ally different approaches for assessing the reliability
of a classifier: Robustness Quantification (RQ) an
We compare both approaches on a number
ere is no clear winner between the two
ombined to obtain a hybrid approa
t of our approach, for each d
nce of uncertainty and

LOCAL

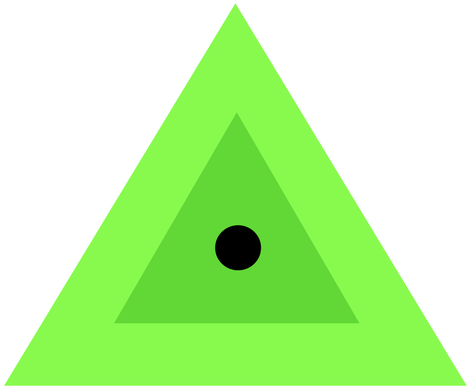




GLOBAL



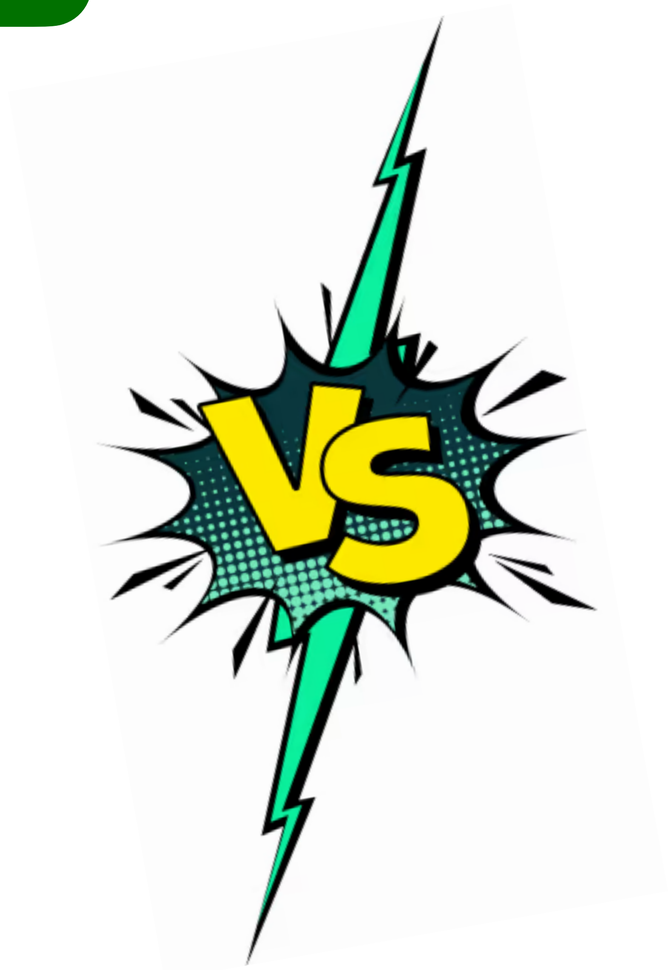
ϵ-CONTAMINATION



$\mathscr{P}_\epsilon$

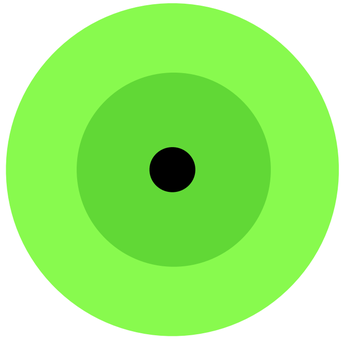
$\|$

$\{(1-\epsilon)P_{\text{classif}} + \epsilon P : P \in \mathbb{P}\}$



OTHER STUFF

distance-based, …

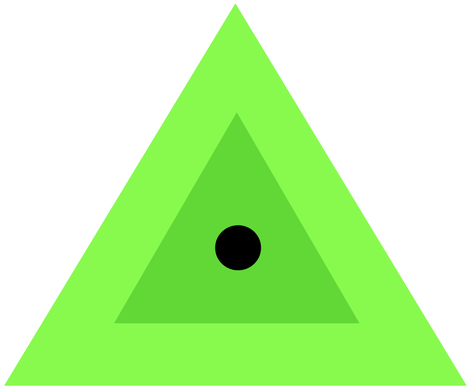


$\mathscr{P}_\delta$

$\|$

$\{P \in \mathbb{P} : \mathrm{d}(P_{\text{classif}}, P) < \delta\}$

ϵ-CONTAMINATION



$\mathscr{P}_\epsilon$

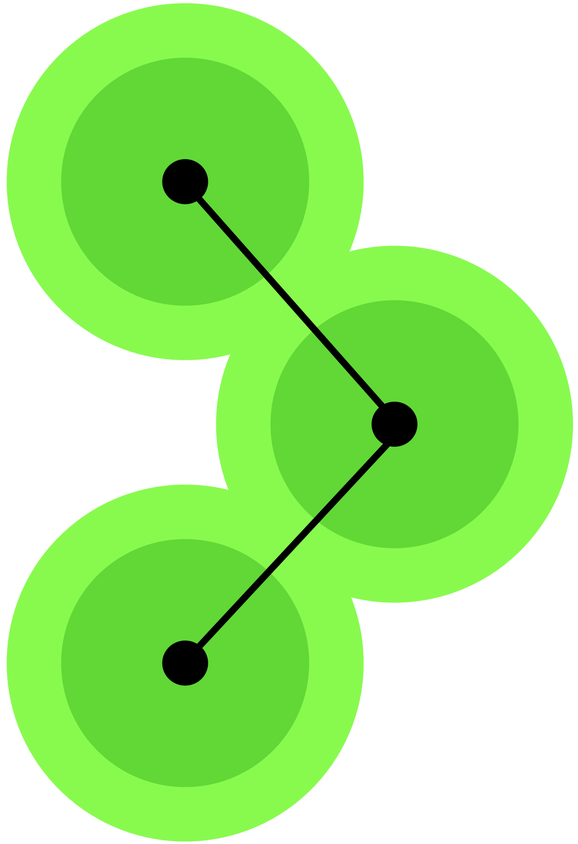
$\|$

$\{(1-\epsilon)P_{\text{classif}} + \epsilon P : P \in \mathbb{P}\}$

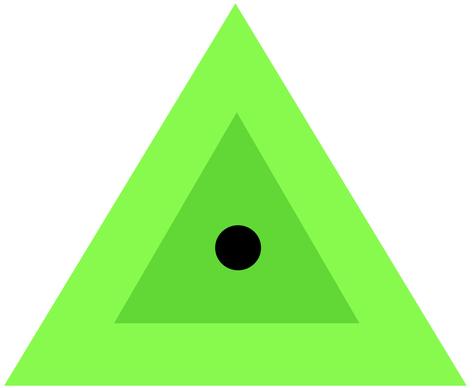
earlier work

+

LOCAL



ϵ-CONTAMINATION



$\mathscr{P}_\epsilon$

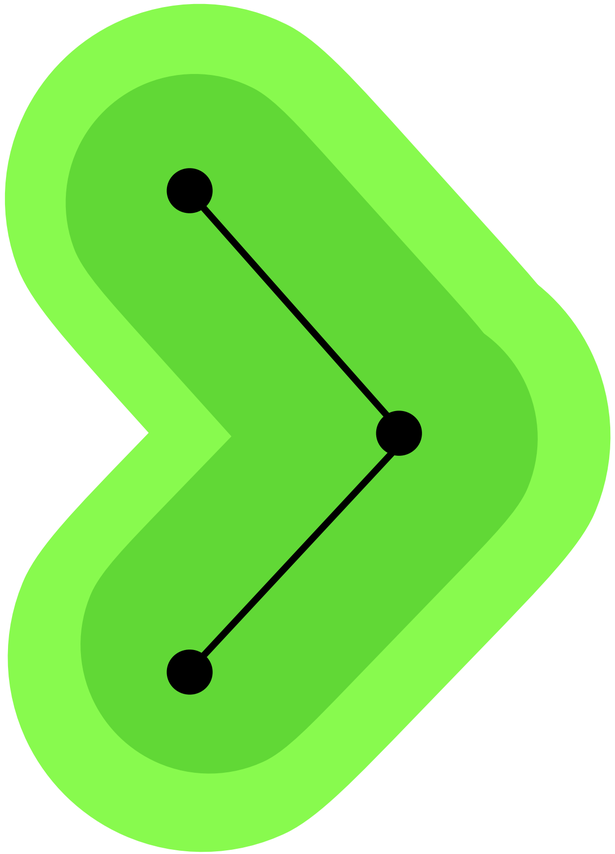
$\|$

$\{(1-\epsilon)P_{\text{classif}} + \epsilon P : P \in \mathbb{P}\}$

new work

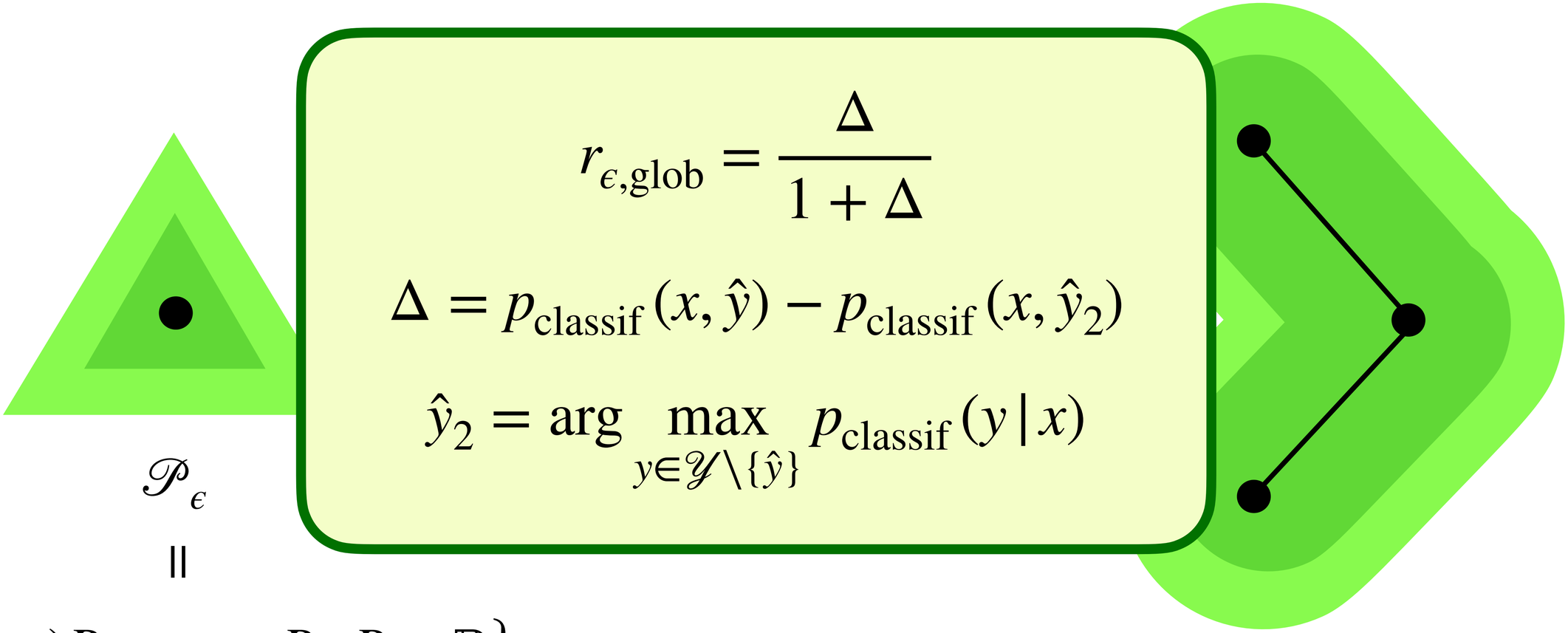
+

GLOBAL



ϵ-CONTAMINATION

GLOBAL



$$r_{\epsilon,\text{glob}} = \frac{\Delta}{1+\Delta}$$

$$\Delta = p_{\text{classif}}(x, \hat{y}) - p_{\text{classif}}(x, \hat{y}_2)$$

$$\hat{y}_2 = \arg \max_{y \in \mathcal{Y} \setminus \{\hat{y}\}} p_{\text{classif}}(y \mid x)$$

$$\mathscr{P}_\epsilon$$

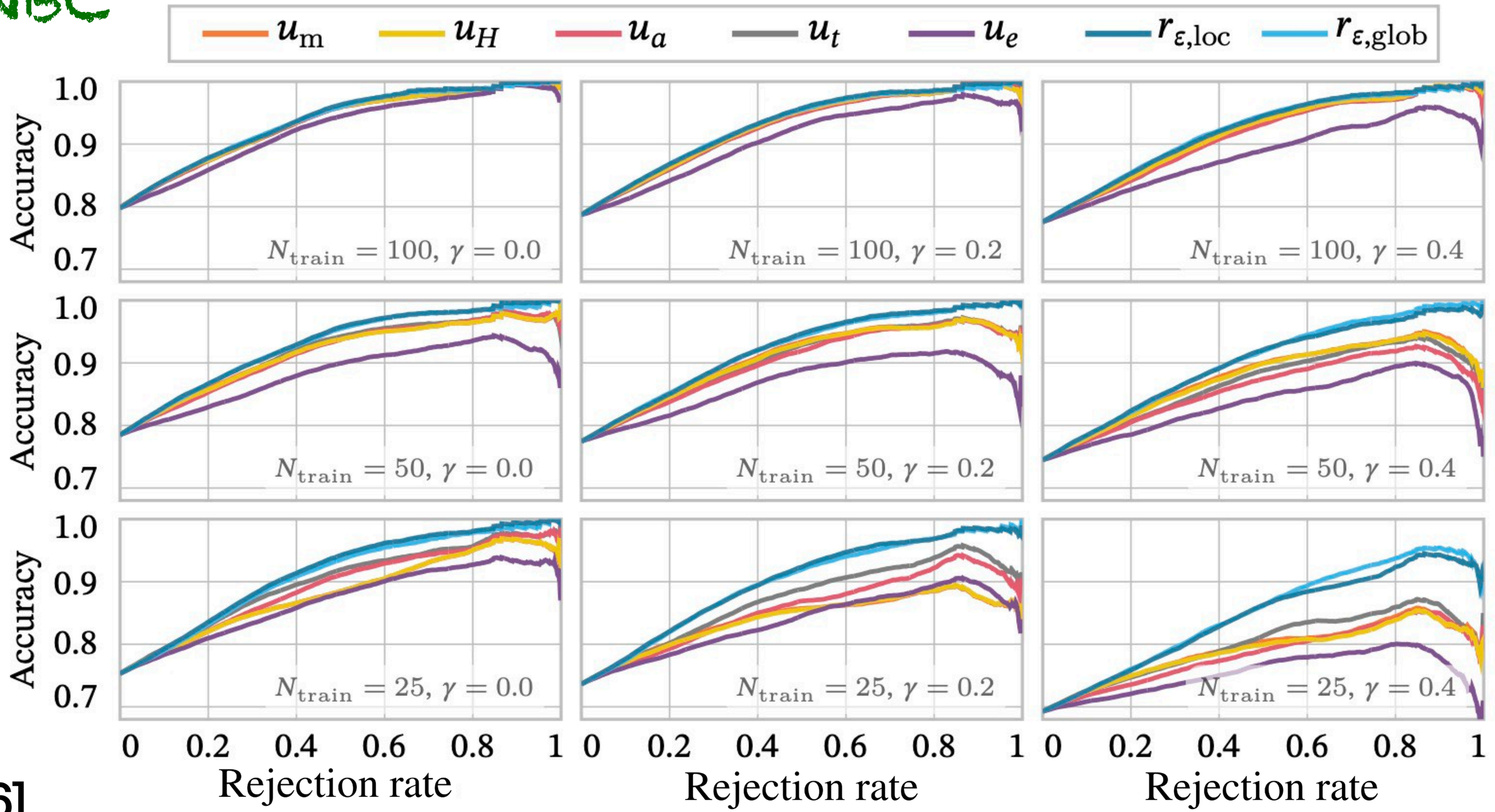
$$\|$$

$$\left\{(1-\epsilon)P_{\text{classif}} + \epsilon P \colon P \in \mathbb{P}\right\}$$

ROBUSTNESS QUANTIFICATION

- correlates nicely with accuracy
- works for different types of model architectures
- also works with global perturbations ✓

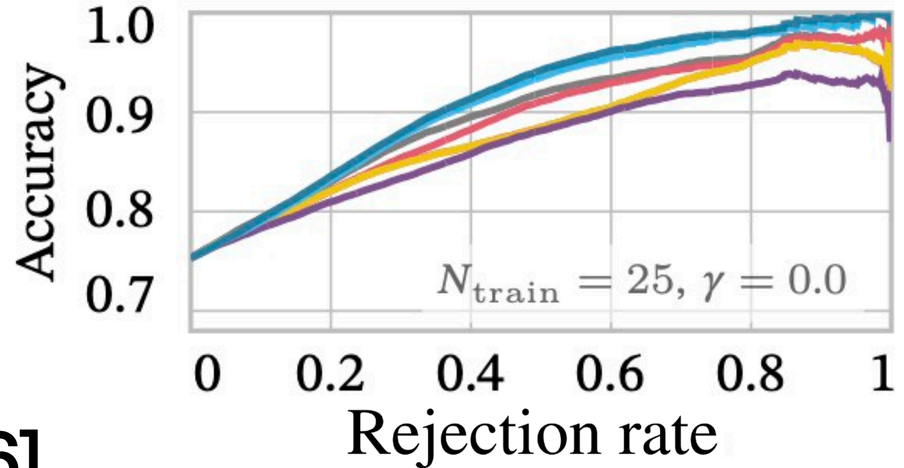
NBC


u_{m}
u_H
u_a
u_t
u_e
$r_{\varepsilon,\mathrm{loc}}$
$r_{\varepsilon,\mathrm{glob}}$
Accuracy
1.0
0.9
0.8
0.7
$N_{\mathrm{train}} = 100, \gamma = 0.0$
$N_{\mathrm{train}} = 100, \gamma = 0.2$
$N_{\mathrm{train}} = 100, \gamma = 0.4$
$N_{\mathrm{train}} = 50, \gamma = 0.0$
$N_{\mathrm{train}} = 50, \gamma = 0.2$
$N_{\mathrm{train}} = 50, \gamma = 0.4$
$N_{\mathrm{train}} = 25, \gamma = 0.0$
$N_{\mathrm{train}} = 25, \gamma = 0.2$
$N_{\mathrm{train}} = 25, \gamma = 0.4$
0
0.2
0.4
0.6
0.8
1
Rejection rate

[6]

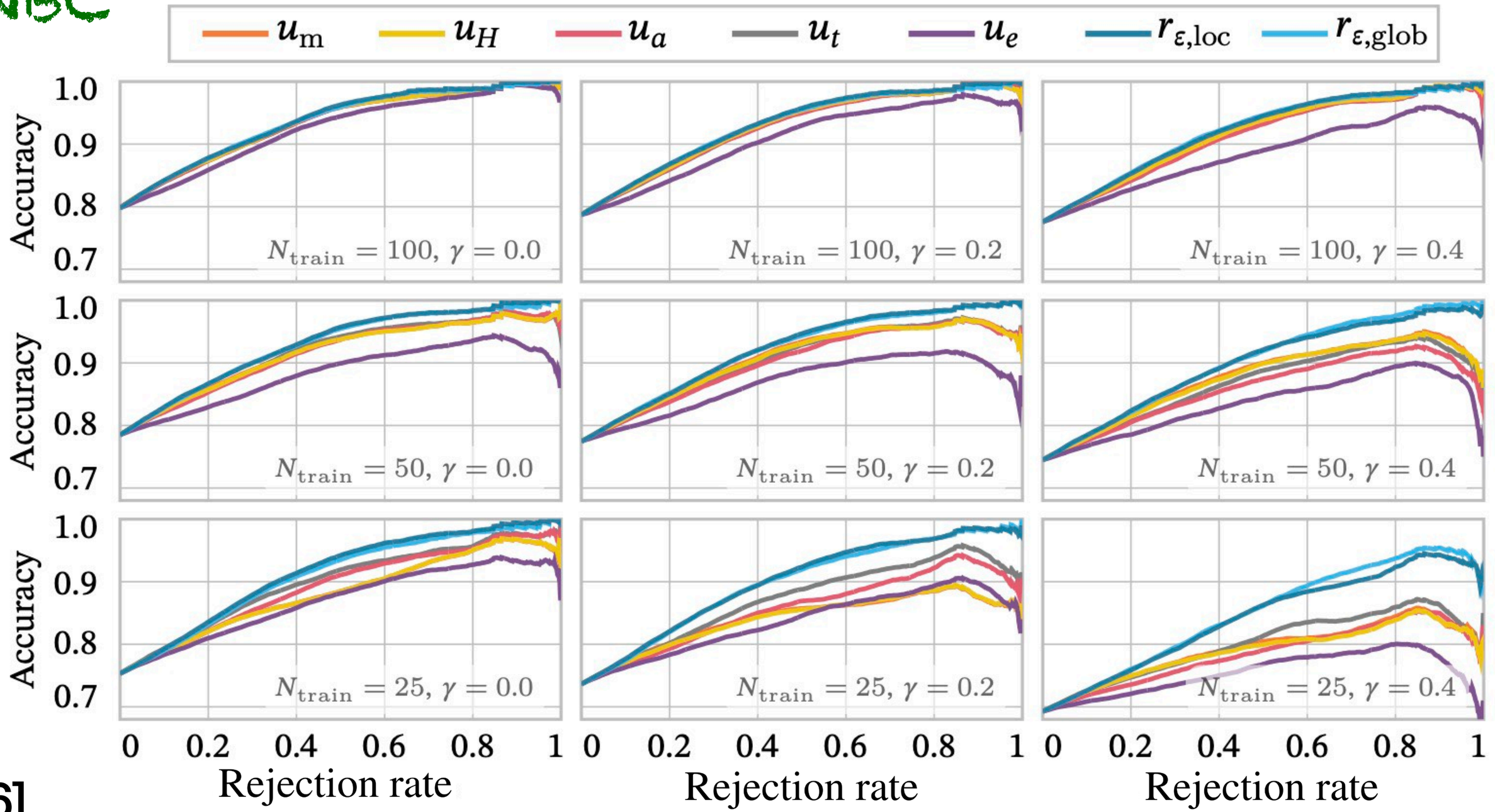
NBC


u_{m}
u_H
u_a
u_t
u_e
$r_{\varepsilon,\mathrm{loc}}$
$r_{\varepsilon,\mathrm{glob}}$


Accuracy
1.0
0.9
0.8
0.7
$N_{\mathrm{train}} = 25, \gamma = 0.0$
0
0.2
0.4
0.6
0.8
1
Rejection rate

[6]

NBC

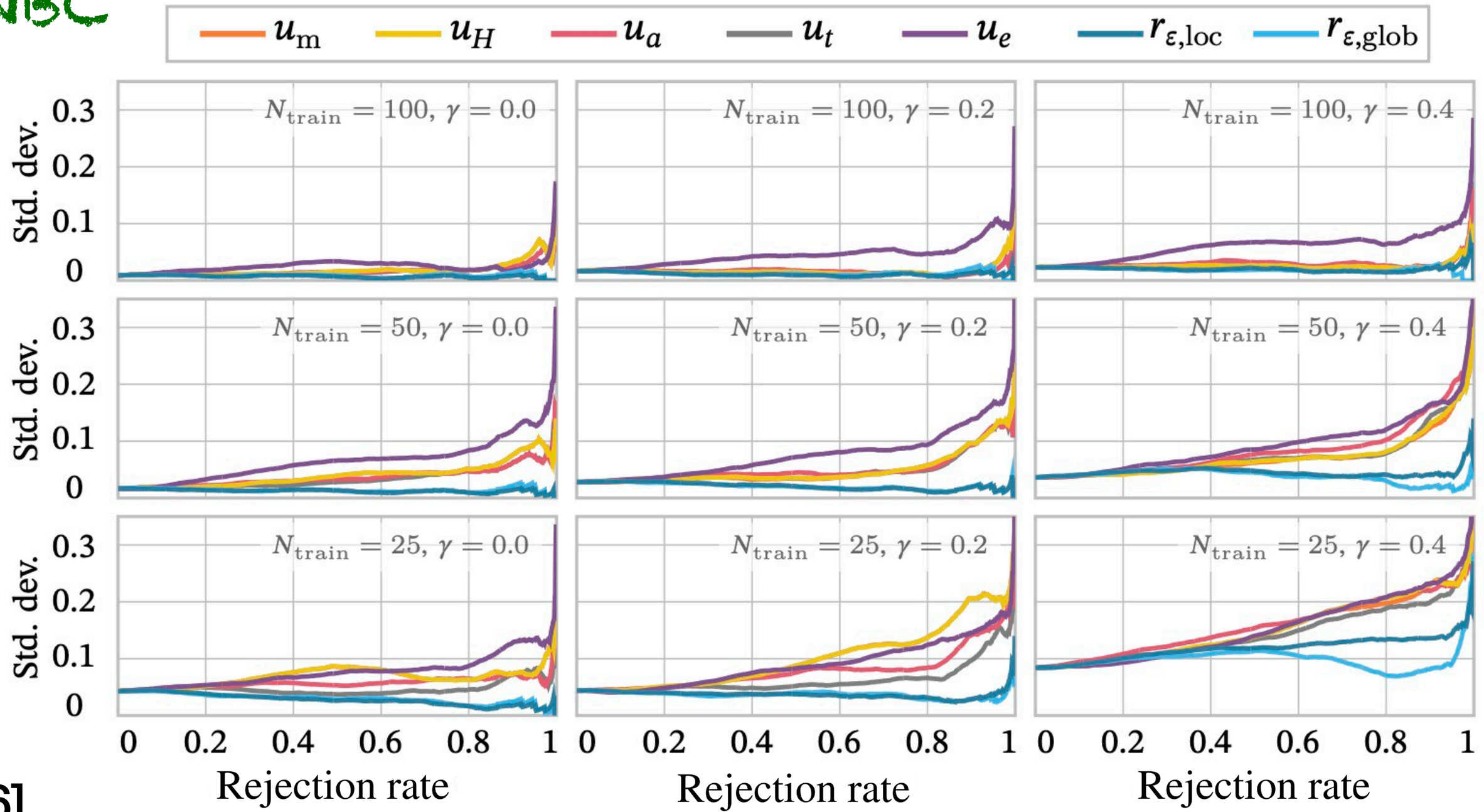

u_{m}
u_H
u_a
u_t
u_e
$r_{\varepsilon,\mathrm{loc}}$
$r_{\varepsilon,\mathrm{glob}}$
Accuracy
1.0
0.9
0.8
0.7
$N_{\mathrm{train}} = 100, \gamma = 0.0$
$N_{\mathrm{train}} = 100, \gamma = 0.2$
$N_{\mathrm{train}} = 100, \gamma = 0.4$
$N_{\mathrm{train}} = 50, \gamma = 0.0$
$N_{\mathrm{train}} = 50, \gamma = 0.2$
$N_{\mathrm{train}} = 50, \gamma = 0.4$
$N_{\mathrm{train}} = 25, \gamma = 0.0$
$N_{\mathrm{train}} = 25, \gamma = 0.2$
$N_{\mathrm{train}} = 25, \gamma = 0.4$
0
0.2
0.4
0.6
0.8
1
Rejection rate

[6]

ROBUSTNESS QUANTIFICATION

- correlates nicely with accuracy ✓
- works for different types of model architectures
- also works with global perturbations ✓
- is competitive with UQ ✓
- is good with distribution shift and small data sets ✓

NBC

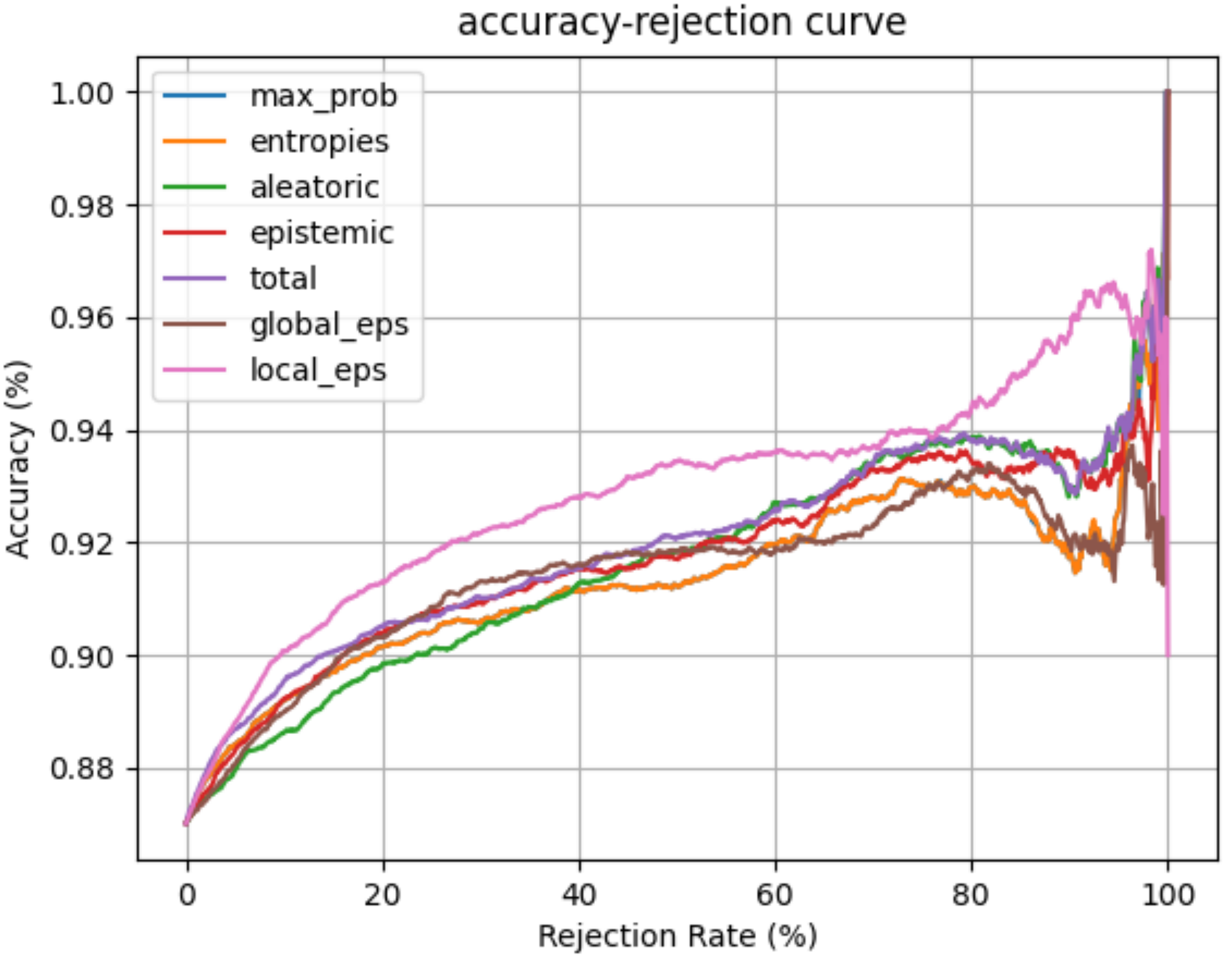

u_{m}
u_H
u_a
u_t
u_e
$r_{\varepsilon,\mathrm{loc}}$
$r_{\varepsilon,\mathrm{glob}}$
$N_{\mathrm{train}} = 100,\ \gamma = 0.0$
$N_{\mathrm{train}} = 100,\ \gamma = 0.2$
$N_{\mathrm{train}} = 100,\ \gamma = 0.4$
$N_{\mathrm{train}} = 50,\ \gamma = 0.0$
$N_{\mathrm{train}} = 50,\ \gamma = 0.2$
$N_{\mathrm{train}} = 50,\ \gamma = 0.4$
$N_{\mathrm{train}} = 25,\ \gamma = 0.0$
$N_{\mathrm{train}} = 25,\ \gamma = 0.2$
$N_{\mathrm{train}} = 25,\ \gamma = 0.4$
Std. dev.
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1
Rejection rate

[6]

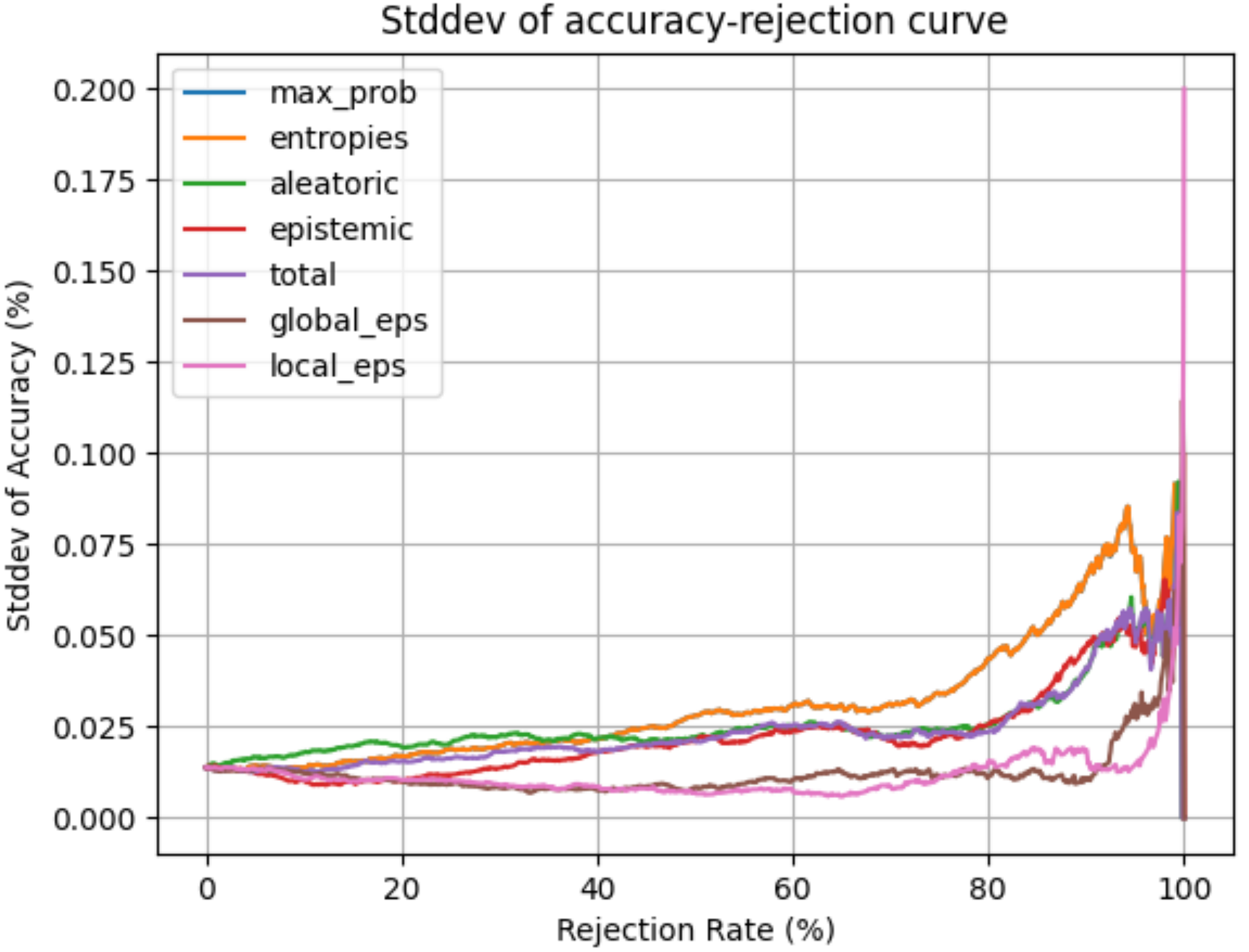
ROBUSTNESS QUANTIFICATION

- correlates nicely with accuracy
- works for different types of model architectures
- also works with global perturbations
- is competitive with UQ
- is good with distribution shift and small data sets
- is more stable than UQ ✓

NBC


accuracy-rejection curve
max_prob
entropies
aleatoric
epistemic
total
global_eps
local_eps
Accuracy (%)
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0
20
40
60
80
100
Rejection Rate (%)

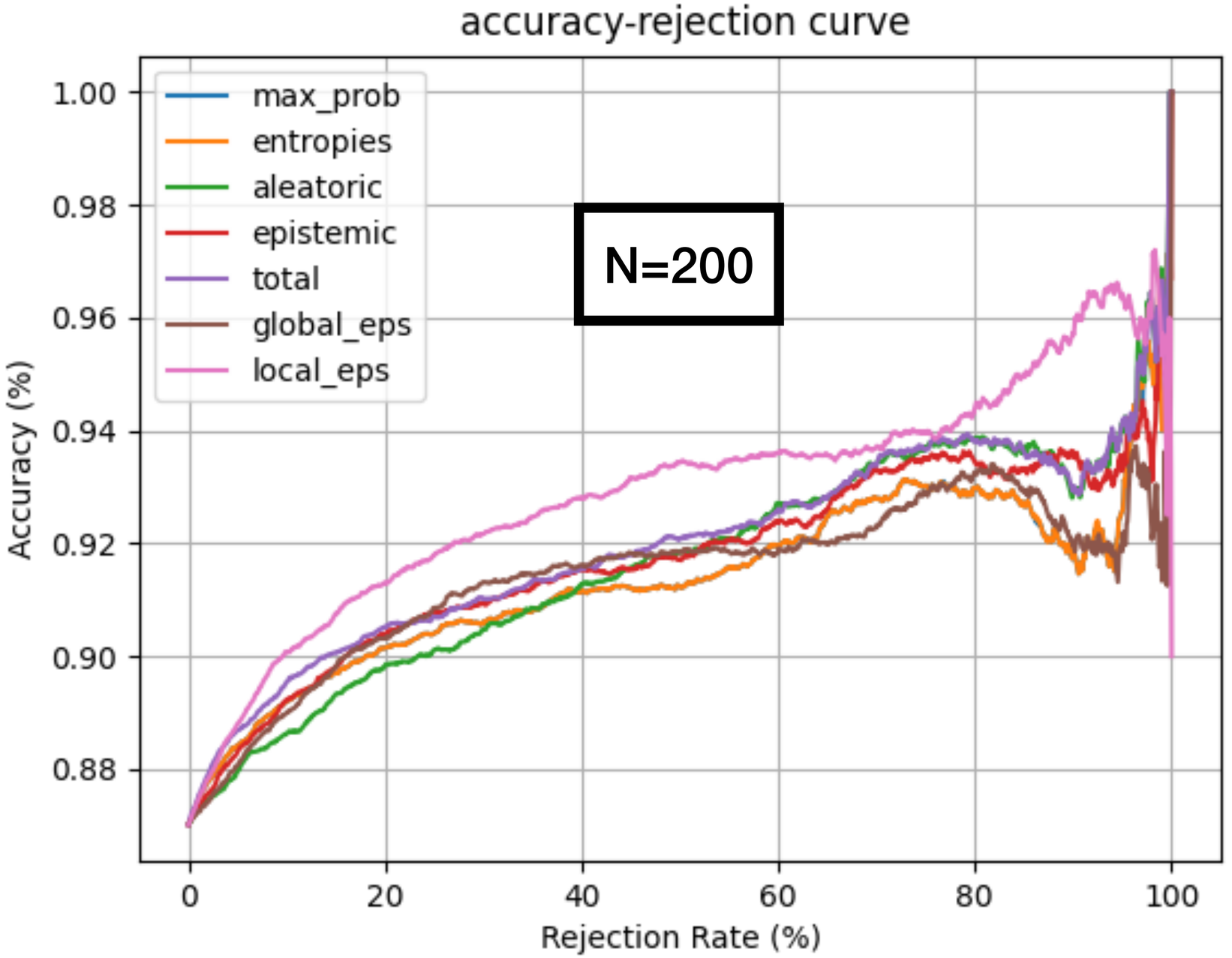
NBC


Stddev of accuracy-rejection curve
max_prob
entropies
aleatoric
epistemic
total
global_eps
local_eps
Stddev of Accuracy (%)
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
20
40
60
80
100
Rejection Rate (%)

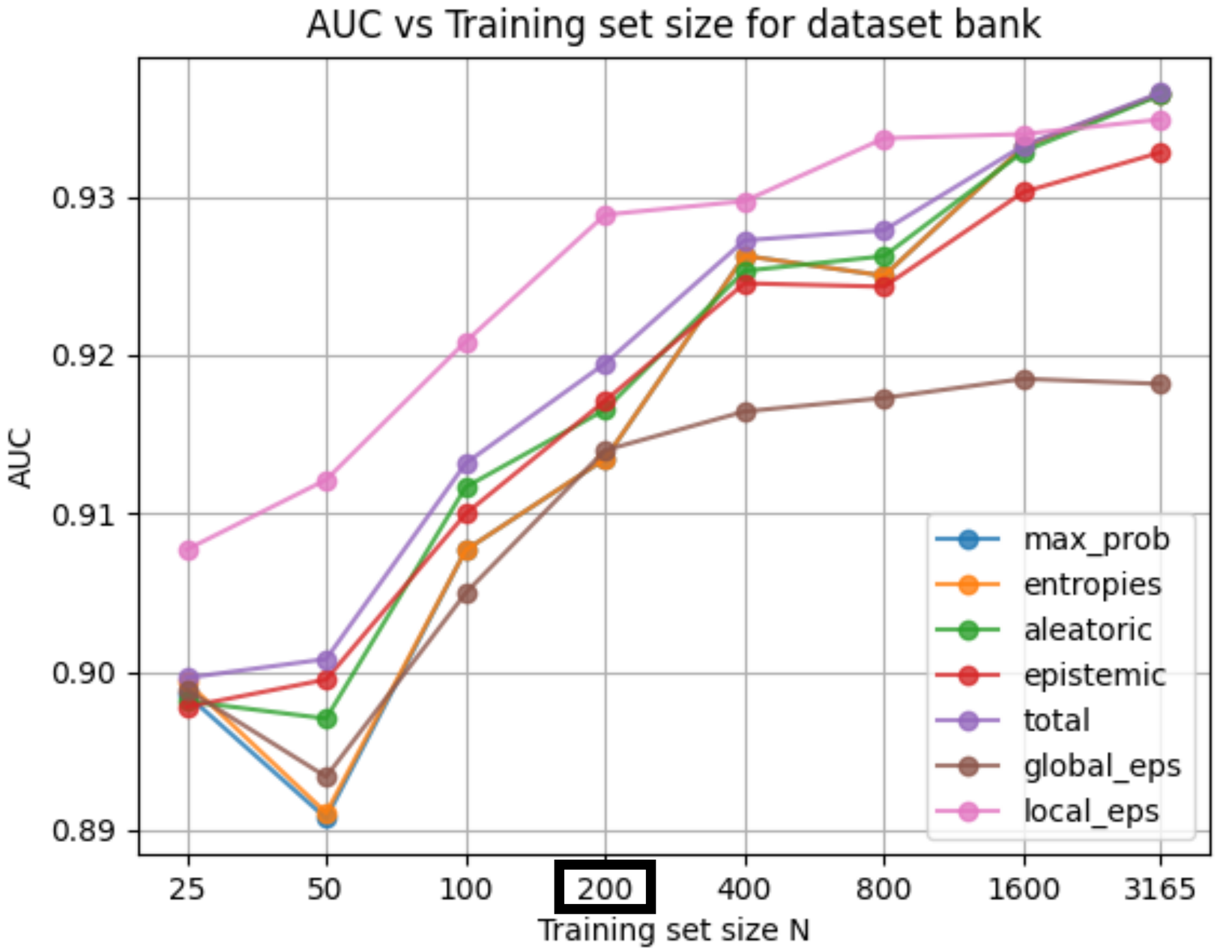
ROBUSTNESS QUANTIFICATION

- correlates nicely with accuracy ✓
- works for different types of model architectures
- also works with global perturbations ✓
- is competitive with UQ ✓
- is good with distribution shift and small data sets
- is more stable than UQ ✓

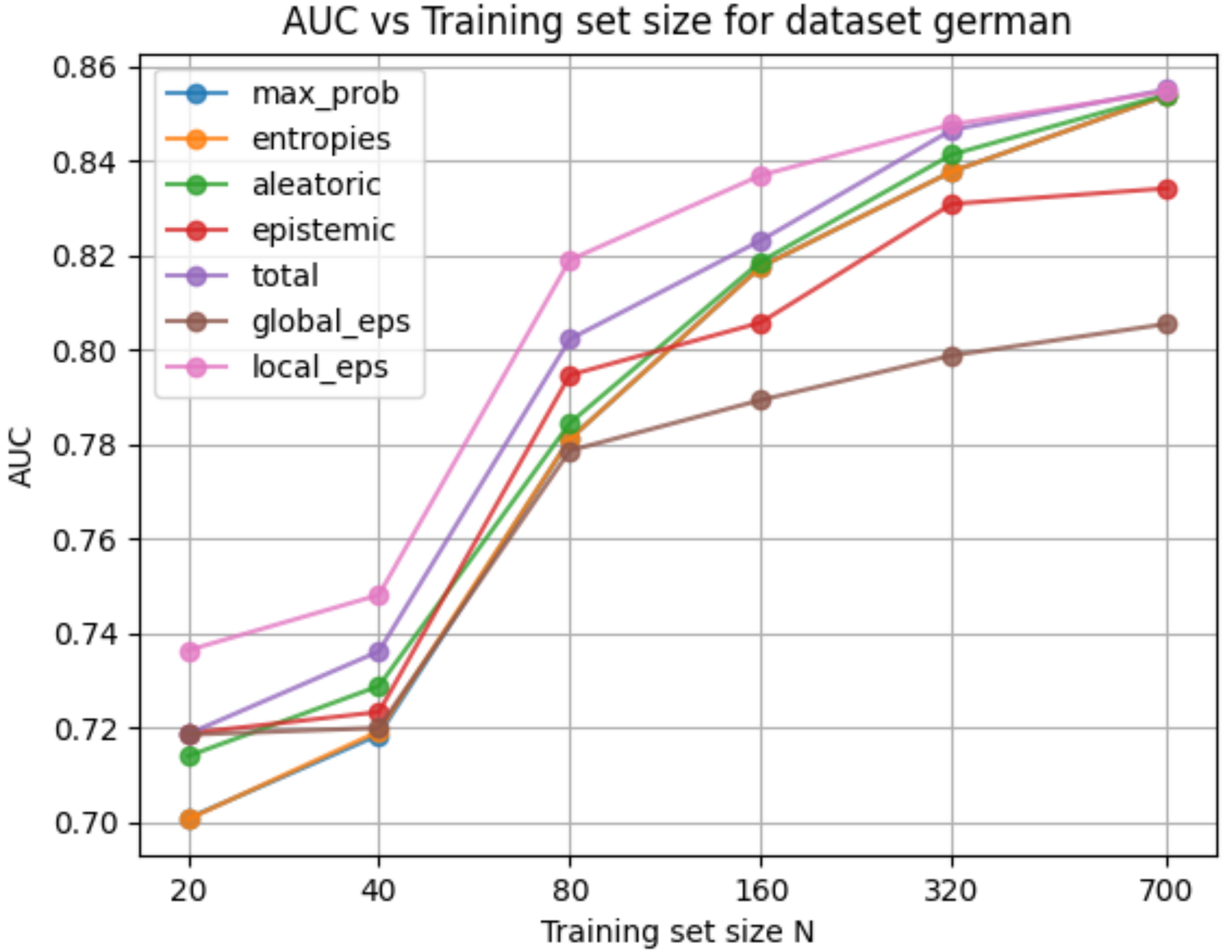
NBC


accuracy-rejection curve
max_prob
entropies
aleatoric
epistemic
total
global_eps
local_eps
N=200
Accuracy (%)
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0
20
40
60
80
100
Rejection Rate (%)

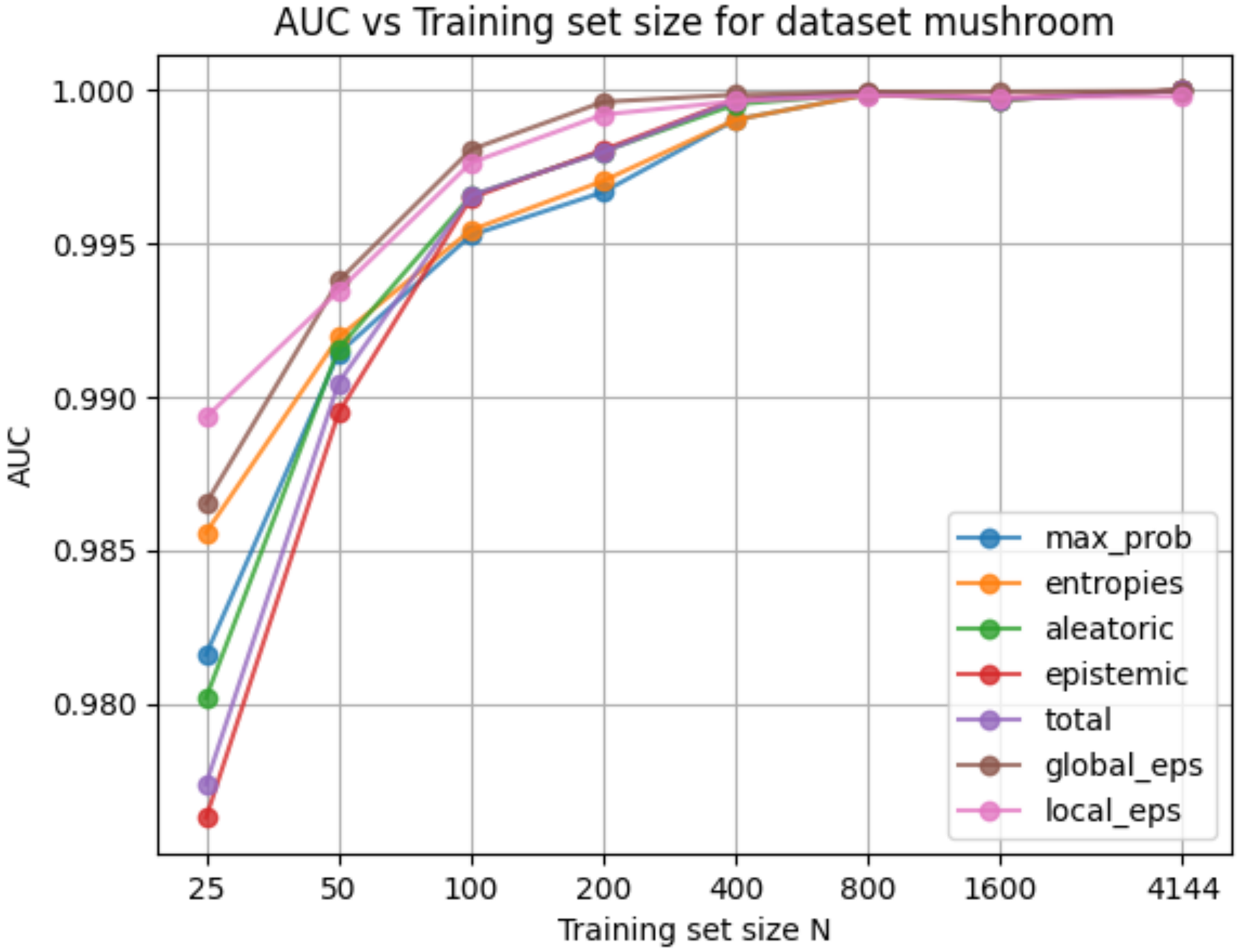
NBC


AUC vs Training set size for dataset bank
AUC
0.93
0.92
0.91
0.90
0.89
25
50
100
200
400
800
1600
3165
Training set size N
max_prob
entropies
aleatoric
epistemic
total
global_eps
local_eps

NBC


AUC vs Training set size for dataset german
max_prob
entropies
aleatoric
epistemic
total
global_eps
local_eps
AUC
0.86
0.84
0.82
0.80
0.78
0.76
0.74
0.72
0.70
20
40
80
160
320
700
Training set size N

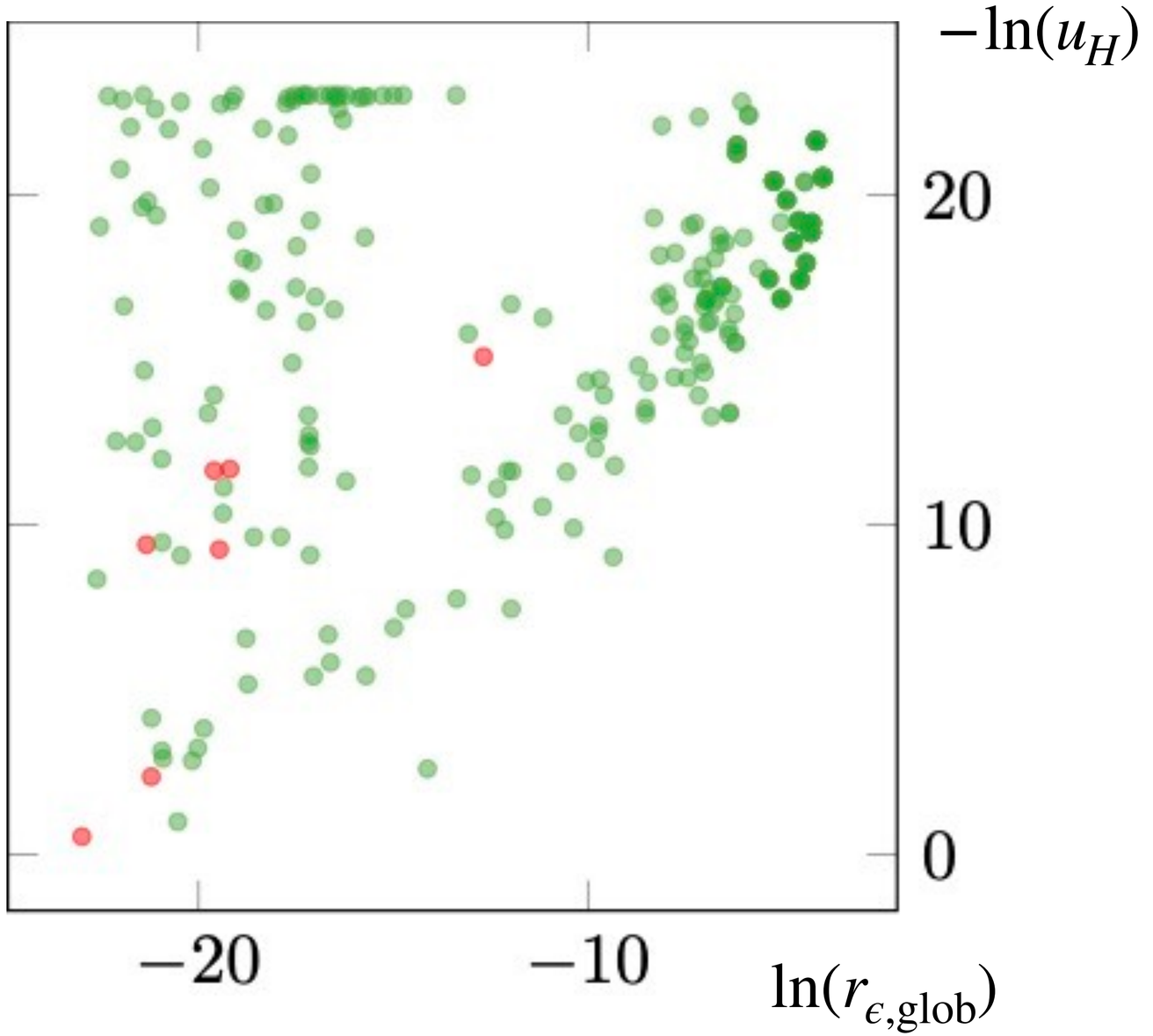
NBC


AUC vs Training set size for dataset mushroom
1.000
0.995
0.990
0.985
0.980
AUC
25
50
100
200
400
800
1600
4144
Training set size N
max_prob
entropies
aleatoric
epistemic
total
global_eps
local_eps

ROBUSTNESS QUANTIFICATION

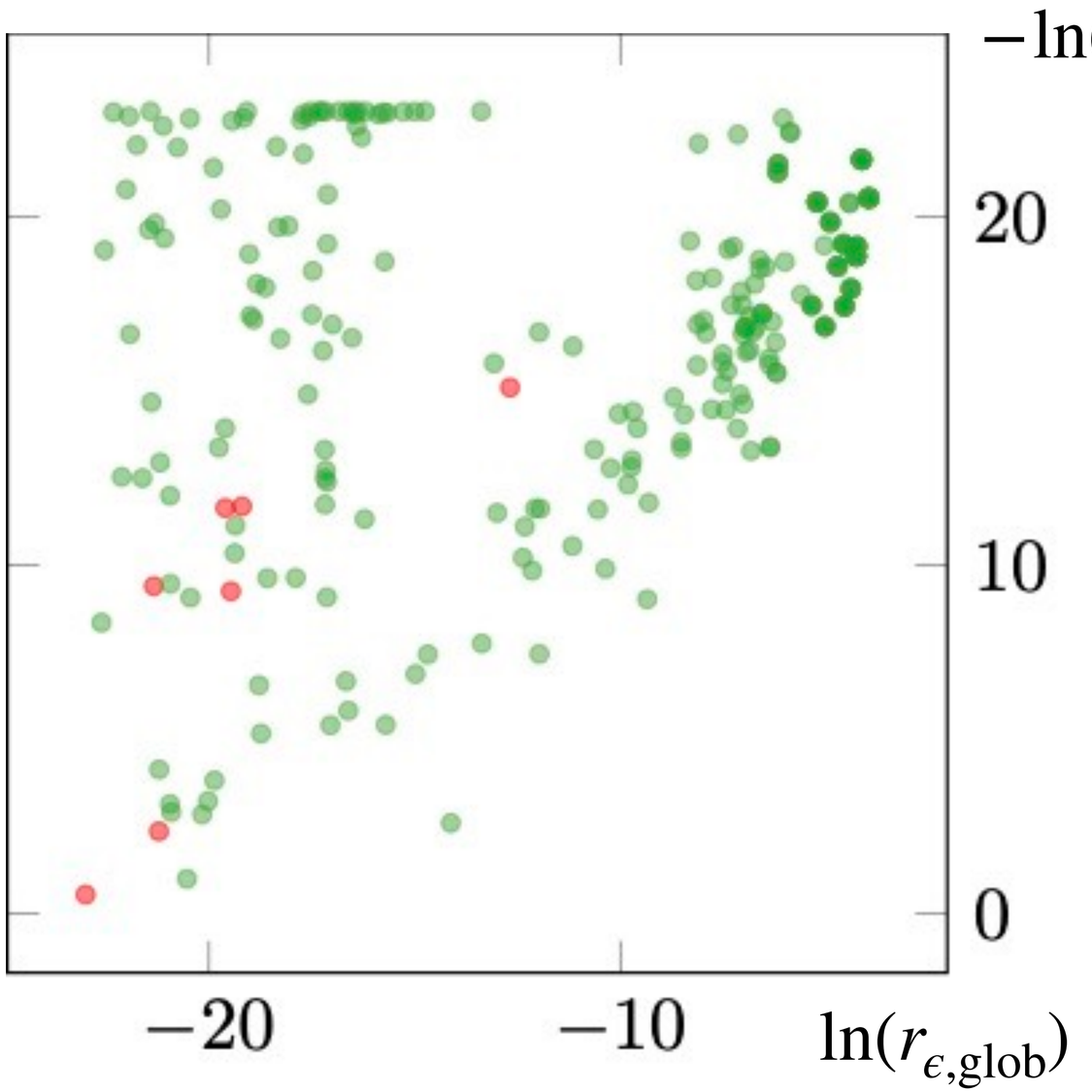
- correlates nicely with accuracy
- works for different types of model architectures
- also works with global perturbations ✓
- is competitive with UQ ✓
- is good with distribution shift and small data sets ✓
- is more stable than UQ ✓

NBC



[7]

NBC

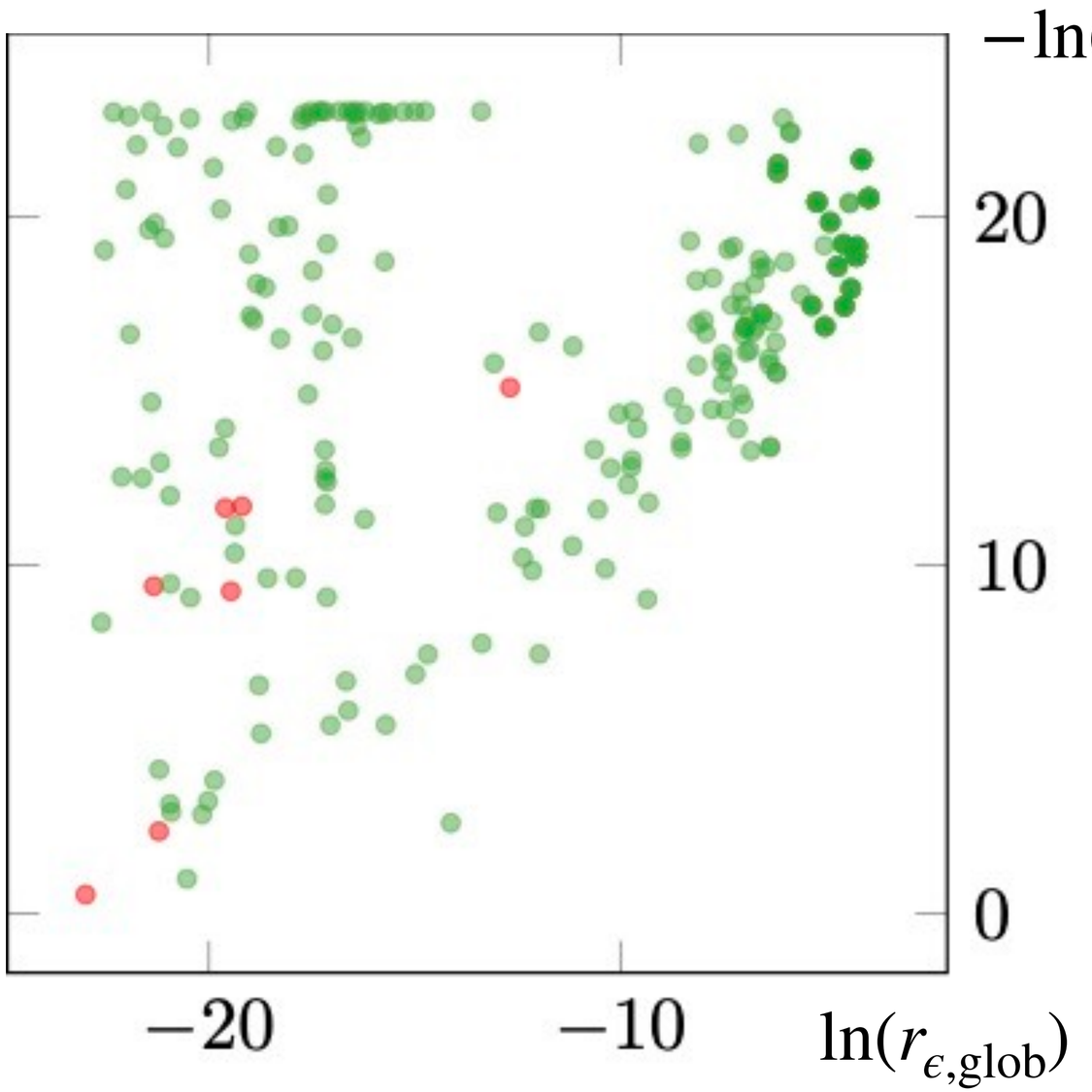


uncertainty ordering

$n_{u_H,i}$: order of i according to u_H

[7]

NBC



uncertainty ordering

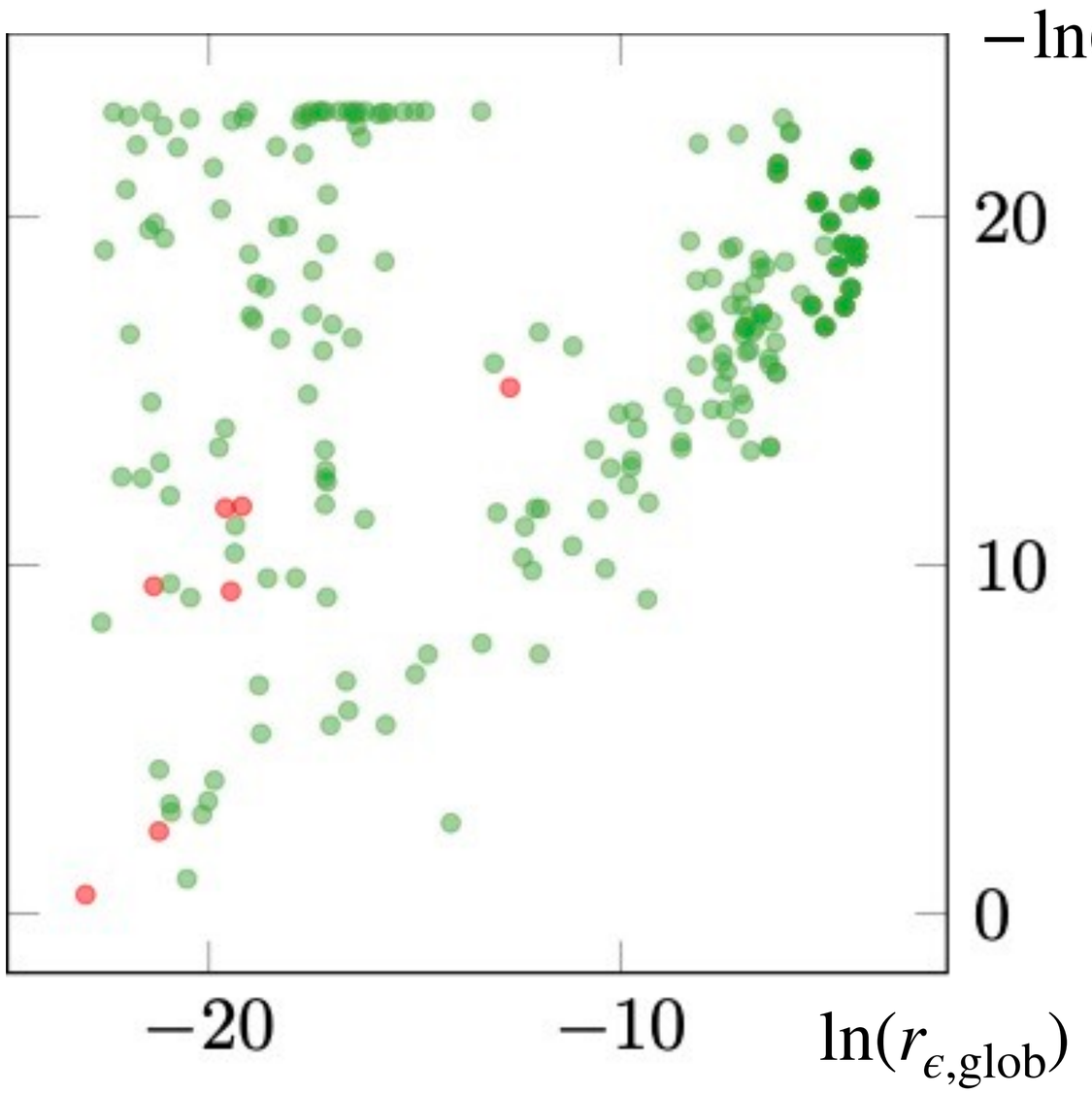
$n_{u_H,i}$: order of i according to u_H

robustness ordering

$n_{r_{\epsilon,\text{glob}},i}$: order of i according to $r_{\epsilon,\text{glob}}$

[7]

NBC



uncertainty ordering

$n_{u_H,i}$: order of i according to u_H

robustness ordering

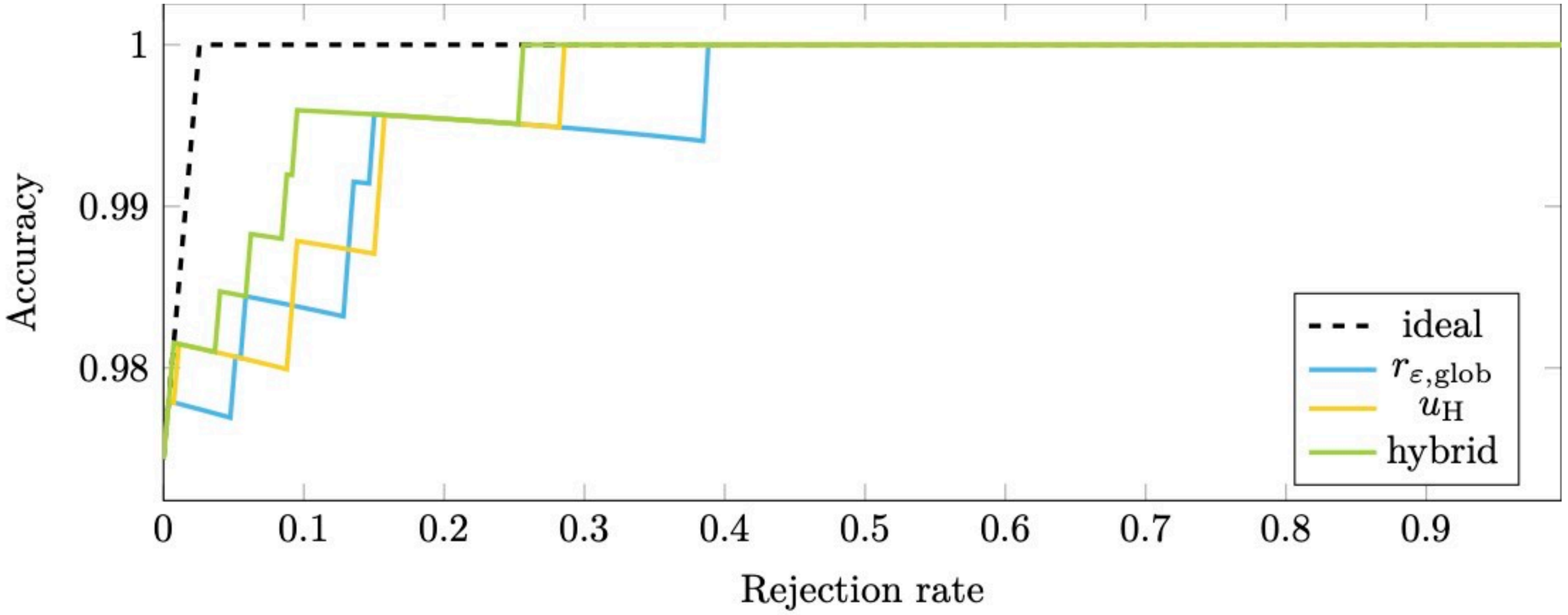
$n_{r_{\epsilon,\text{glob}},i}$: order of i according to $r_{\epsilon,\text{glob}}$

hybrid ordering

$$h_i = \gamma n_{u_H,i} + (1-\gamma) n_{r_{\epsilon,\text{glob}},i}$$

[7]

NBC


1
0.99
0.98
Accuracy
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Rejection rate
ideal
$r_{\varepsilon,\text{glob}}$
u_{H}
hybrid

[7]

NBC

Dataset	u_{H}	$r_{\varepsilon,\mathrm{loc}}$	hybrid	γ	$r_{\varepsilon,\mathrm{glob}}$	hybrid	γ
Adult	0.9295	0.9066	**0.9295**	1.00	0.7690	**0.9295**	1.00
Austr. Cr.	0.9236	0.9139	**0.9265**	0.75	0.8872	**0.9246**	0.86
Bank M.	0.9485	0.9452	**0.9485**	0.55	0.9299	0.9481	0.88
BCW	0.9968	0.9962	**0.9974**	0.52	0.9961	**0.9978**	0.53
German Cr.	0.8338	**0.8380**	0.8378	0.53	0.7972	**0.8376**	0.85
Heart dis.	0.7602	0.7540	**0.7602**	0.95	0.6761	0.7600	0.95
Lymphogr.	0.9440	0.9419	0.9428	0.77	0.8981	0.9425	0.88
NPHA	0.4962	**0.5021**	0.4917	0.77	**0.5159**	0.4913	0.96
Nursery	0.9813	0.9822	**0.9824**	0.28	0.9730	**0.9814**	0.91
Solar (big)	0.8603	**0.8926**	0.8874	0.23	0.8693	**0.8836**	0.71
Solar (small)	0.8709	0.8597	0.8666	0.19	0.7990	**0.8797**	0.78
SPECT	0.9458	0.8915	**0.9458**	0.99	0.5738	0.9457	0.99
Stud. Math	0.9434	0.9465	**0.9468**	0.31	0.9205	**0.9445**	0.60
Stud. Port	0.8898	**0.9276**	0.9093	0.77	0.8952	**0.9067**	0.79

[7]

NBC

Dataset	u_{H}	$r_{\varepsilon,\mathrm{loc}}$	hybrid	γ	γ^*	$r_{\varepsilon,\mathrm{glob}}$	hybrid	γ	γ^*
Adult	0.9322	0.9044	**0.9322**	1.00	1.00	0.7766	**0.9322**	1.00	1.00
Australian Credit	0.9268	0.9276	**0.9304**	0.83	0.56	0.9097	**0.9290**	0.88	0.65
Bank Marketing	0.9338	0.9343	**0.9345**	0.75	0.25	0.9241	0.9337	0.94	0.97
Breast Cancer	0.9945	0.9969	**0.9959**	0.62	0.15	0.9964	**0.9969**	0.59	0.26
German Credit	0.8670	**0.8644**	0.8680	0.67	0.53	0.8065	**0.8678**	0.86	0.92
Heart disease	0.7927	0.7852	**0.7927**	0.99	0.90	0.7466	0.7930	0.98	0.91
Lymphography	0.8804	0.8705	0.8867	0.39	0.60	0.8337	0.8880	0.66	0.79
NPHA	0.4947	**0.5350**	0.4887	0.85	0.28	**0.5220**	0.4856	0.91	0.26
Nursery	0.9814	0.9828	**0.9827**	0.33	0.17	0.9727	**0.9813**	0.93	0.96
Solar Flare (big)	0.8450	**0.8762**	0.8736	0.24	0.00	0.8638	**0.8736**	0.72	0.54
Solar Flare (small)	0.8635	0.8623	0.8560	0.64	0.43	0.8115	**0.8706**	0.77	0.84
SPECT Heart	0.9455	0.8915	**0.9455**	0.99	0.99	0.5738	0.9454	0.99	1.00
Student Math	0.9336	0.9381	**0.9376**	0.46	0.32	0.9289	**0.9396**	0.65	0.53
Student Port	0.8885	**0.9159**	0.9106	0.35	0.01	0.9032	**0.9073**	0.73	0.39

[7]

ROBUSTNESS QUANTIFICATION

- correlates nicely with accuracy ✓
- works for different types of model architectures
- also works with global perturbations ✓
- is competitive with and complementary to UQ ✓
- is good with distribution shift and small data sets
- is more stable than UQ

Rodrigo
Lassance

MACHINE
LEARNING

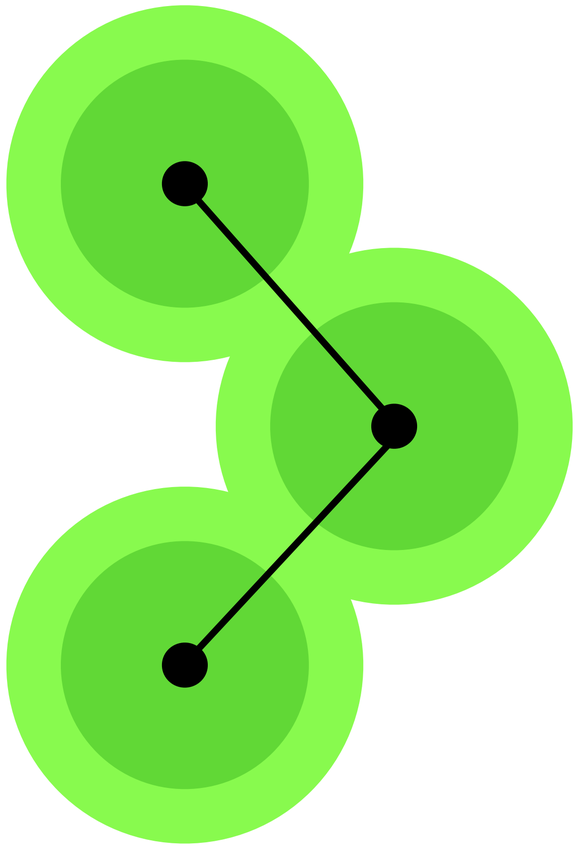
IMPRECISE
PROBABILITIES

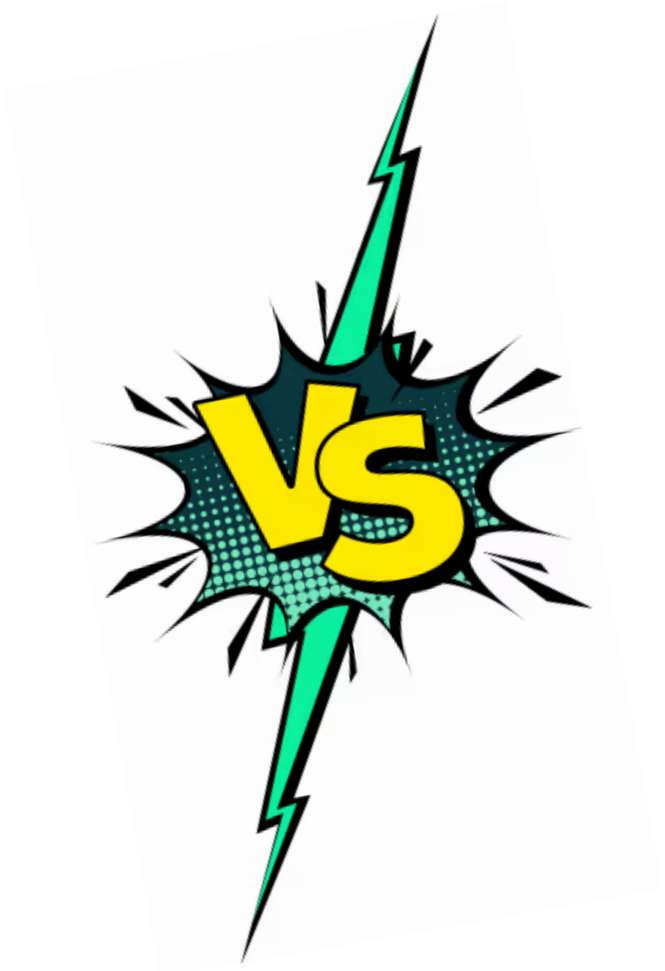
ROBUSTNESS
QUANTIFICATION

SMaLL
Statistical Machine Learning Lab

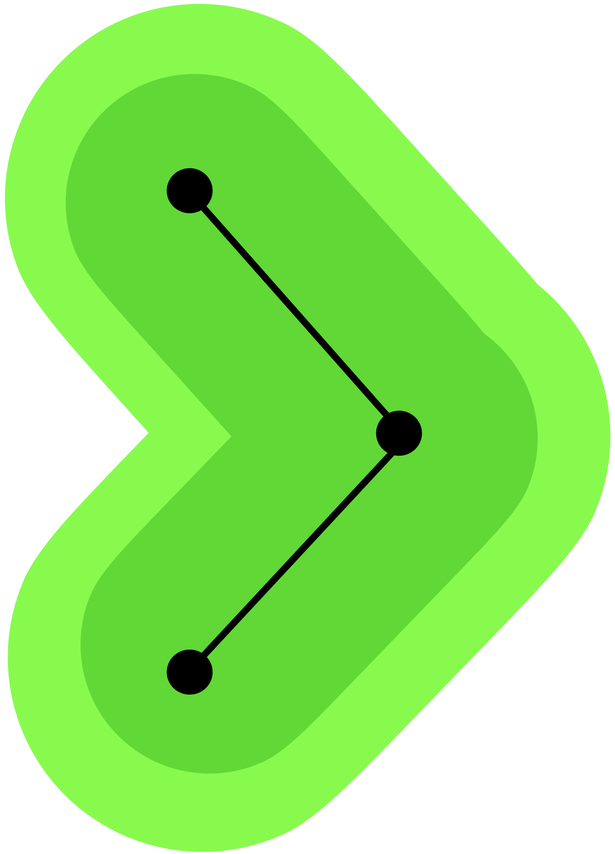
FLip
Foundations Lab for imprecise probabilities

LOCAL

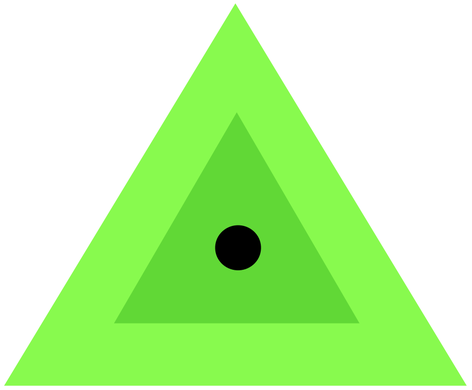




GLOBAL



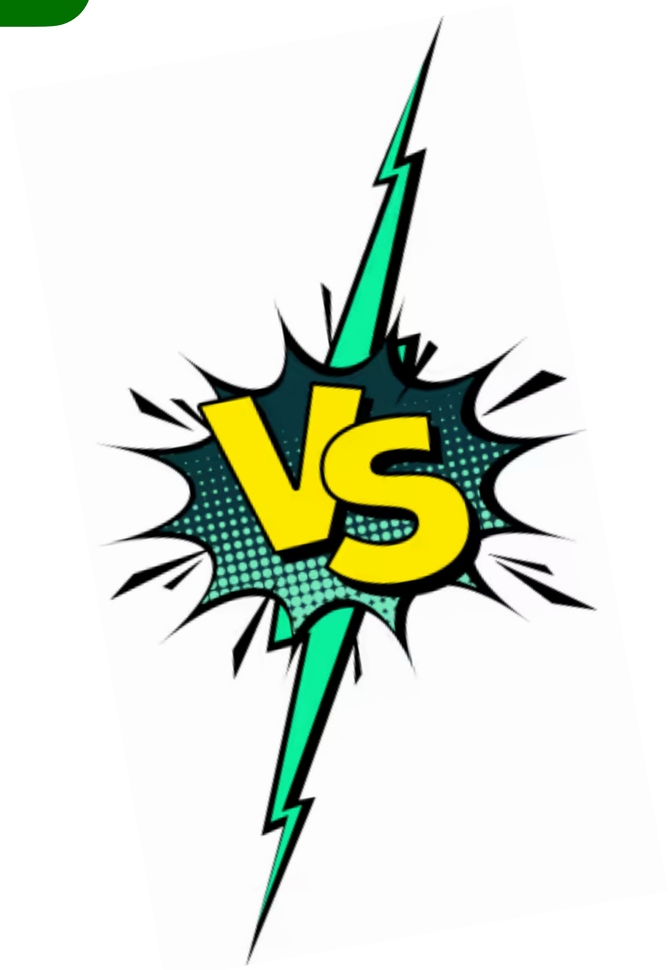
ϵ-CONTAMINATION



$\mathscr{P}_\epsilon$

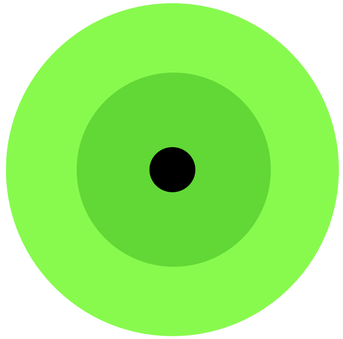
$\|$

$\{(1-\epsilon)P_{\text{classif}} + \epsilon P : P \in \mathbb{P}\}$



OTHER STUFF

distance-based, …

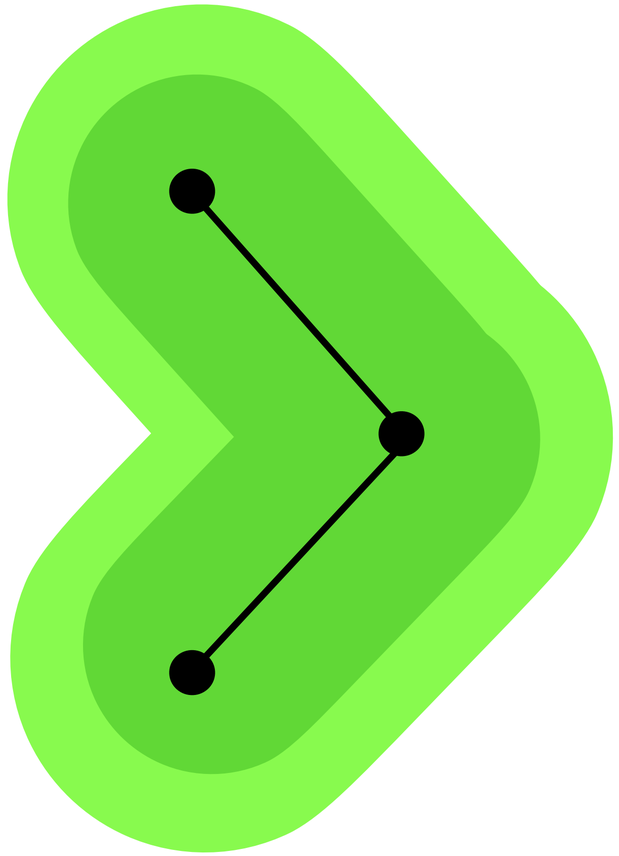


$\mathscr{P}_\delta$

$\|$

$\{P \in \mathbb{P} : \mathrm{d}(P_{\text{classif}}, P) < \delta\}$

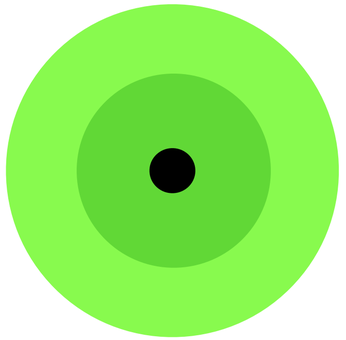
GLOBAL



+

OTHER STUFF

distance-based, …



$$\mathscr{P}_\delta$$

$$\|$$

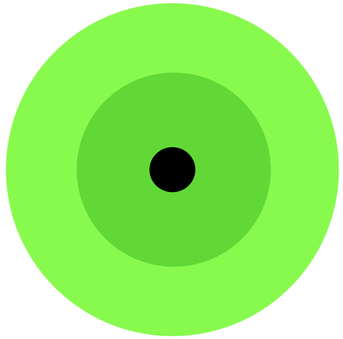
$$\{P \in \mathbb{P} : \mathrm{d}(P_{\text{classif}}, P) < \delta\}$$

GLOBAL

OTHER STUFF

distance-based, …

$$\mathrm{d}_{\mathrm{ratio}}(P_1, P_2) = \log \max \left\{ \sup_A \frac{P_1(A)}{P_2(A)}, \sup_A \frac{P_2(A)}{P_1(A)} \right\}$$



$$\mathscr{P}_\delta$$

$$\|$$

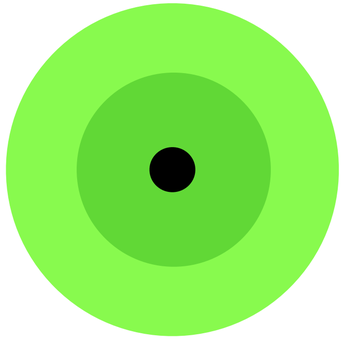
$$\left\{ P \in \mathbb{P} : \mathrm{d}(P_{\mathrm{classif}}, P) < \delta \right\}$$

GLOBAL

$$\mathrm{d}_{\mathrm{ratio}}(P_1, P_2) = \operatorname{ess\,sup}_{(x,y)\in\Omega} \left| \log\left(\frac{p_1(x,y)}{p_2(x,y)}\right) \right|$$

OTHER STUFF

distance-based, …



$$\mathscr{P}_\delta$$

$$\|$$

$$\left\{P \in \mathbb{P} : \mathrm{d}(P_{\mathrm{classif}}, P) < \delta\right\}$$

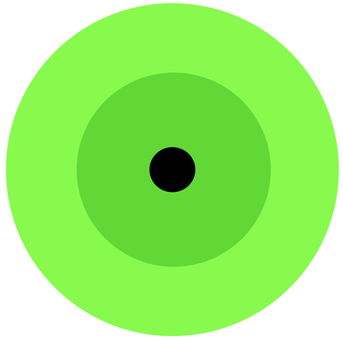
GLOBAL

OTHER STUFF

distance-based, …

$$\mathrm{d}_{\mathrm{ratio}}(P_1, P_2) = \operatorname{ess\,sup}_{(x,y)\in\Omega} \left| \log\left(\frac{p_1(x,y)}{p_2(x,y)}\right) \right|$$

$$\mathrm{d}_{\mathrm{diff}}(P_1, P_2) = \operatorname{ess\,sup}_{(x,y)\in\Omega} \left| p_1(x,y) - p_2(x,y) \right|$$



$$\mathscr{P}_\delta$$

$$\shortparallel$$

$$\left\{ P \in \mathbb{P} : \mathrm{d}(P_{\mathrm{classif}}, P) < \delta \right\}$$

GLOBAL

OTHER STUFF

distance-based, …

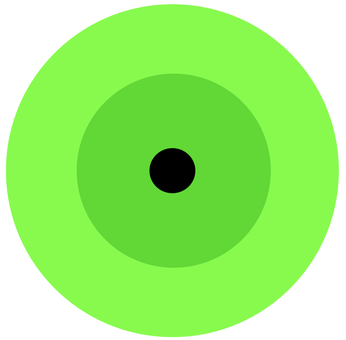
discrete features:

$$r_{\text{ratio}} = \log \min \left\{ \frac{1}{1-\Delta}, \sqrt{\frac{p_{\text{classif}}(x,\hat{y})}{p_{\text{classif}}(x,\hat{y}_2)}} \right\}$$

$$r_{\text{diff}} = \frac{1}{2}\left(p_{\text{classif}}(x,\hat{y}) - p_{\text{classif}}(x,\hat{y}_2)\right)$$

$$\hat{y}_2 = \arg \max_{y \in \mathcal{Y} \setminus \{\hat{y}\}} P_{\text{classif}}(y \mid x)$$

$$\Delta = p_{\text{classif}}(x,\hat{y}) - p_{\text{classif}}(x,\hat{y}_2)$$



$$\mathscr{P}_\delta$$

$$\shortparallel$$

$$\left\{P \in \mathbb{P} \colon \mathrm{d}(P_{\text{classif}}, P) < \delta\right\}$$

GLOBAL

OTHER STUFF

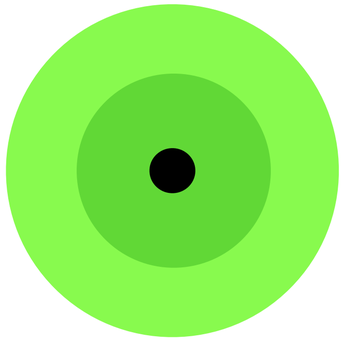
distance-based, …

at least one continuous feature:

$$r_{\text{ratio}} = \log \sqrt{\frac{p_{\text{classif}}(x, \hat{y})}{p_{\text{classif}}(x, \hat{y}_2)}}$$

$$r_{\text{diff}} = \frac{1}{2}\left(p_{\text{classif}}(x, \hat{y}) - p_{\text{classif}}(x, \hat{y}_2)\right)$$

$$\hat{y}_2 = \arg \max_{y \in \mathcal{Y} \setminus \{\hat{y}\}} P_{\text{classif}}(y \mid x)$$



$$\mathscr{P}_\delta$$

$$\|$$

$$\left\{P \in \mathbb{P} : \mathrm{d}(P_{\text{classif}}, P) < \delta\right\}$$

ROBUSTNESS QUANTIFICATION

- correlates nicely with accuracy
- works for different types of model architectures
- also works with global perturbations
- is competitive with and complementary to UQ
- is good with distribution shift and small data sets
- is more stable than UQ
- is not limited to epsilon-contamination ✓

2017-2020

SPN

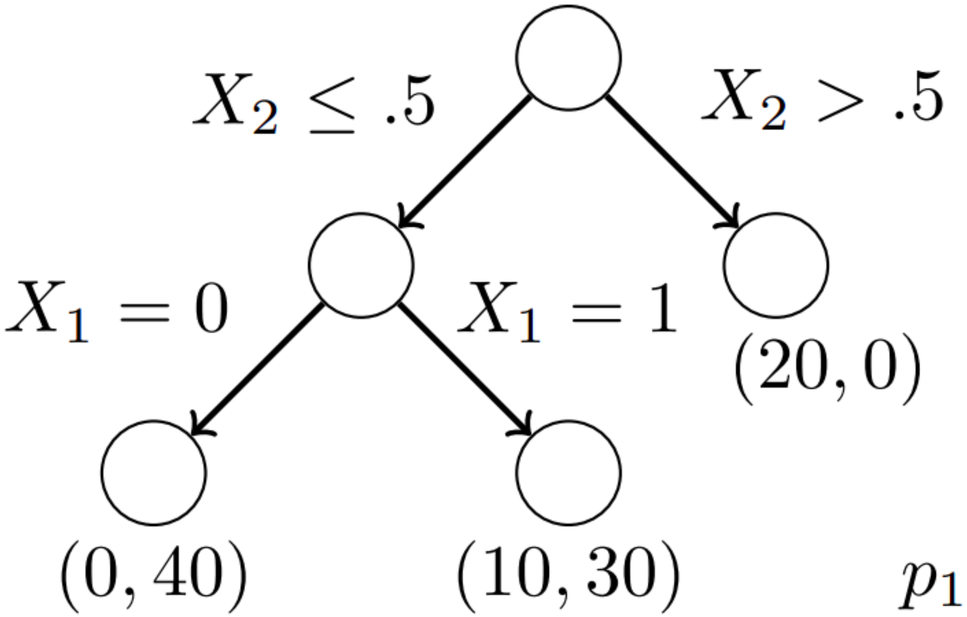
[2-4, ...]

[5]

GeF+

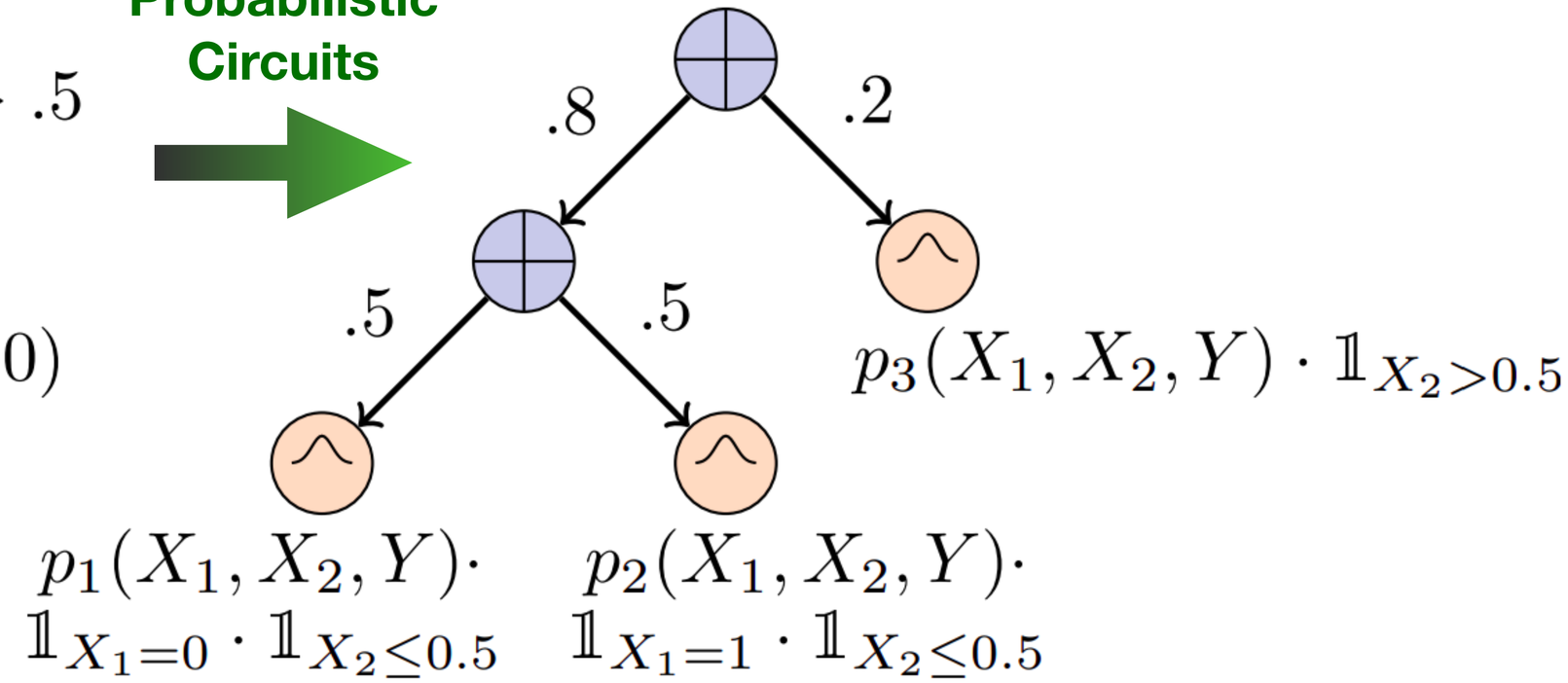
Cassio
de Campos
& various
co-authors

Decision Tree (DT)



Probabilistic Circuits

Generative DT



LOCAL

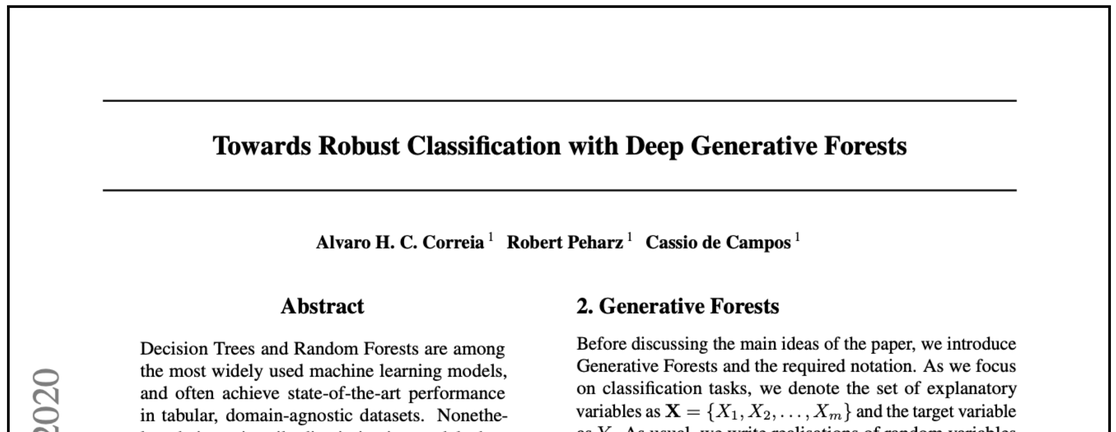
ϵ-CONTAMINATION

continuous features

$$r_{\epsilon,\mathrm{GeF}} = \text{complicated}$$

[5]

GeF+



Towards Robust Classification with Deep Generative Forests

Alvaro H. C. Correia [1] Robert Peharz [1] Cassio de Campos [1]

Abstract

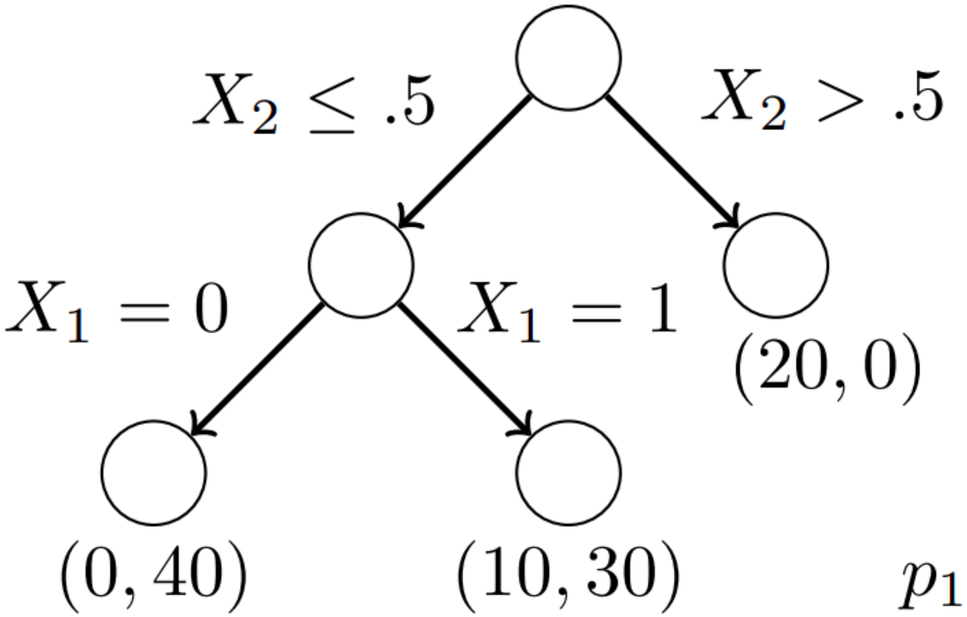
Decision Trees and Random Forests are among the most widely used machine learning models, and often achieve state-of-the-art performance in tabular, domain-agnostic datasets. Nonethe-

2. Generative Forests

Before discussing the main ideas of the paper, we introduce Generative Forests and the required notation. As we focus on classification tasks, we denote the set of explanatory variables as $\mathbf{X} = \{X_1, X_2, \ldots, X_m\}$ and the target variable

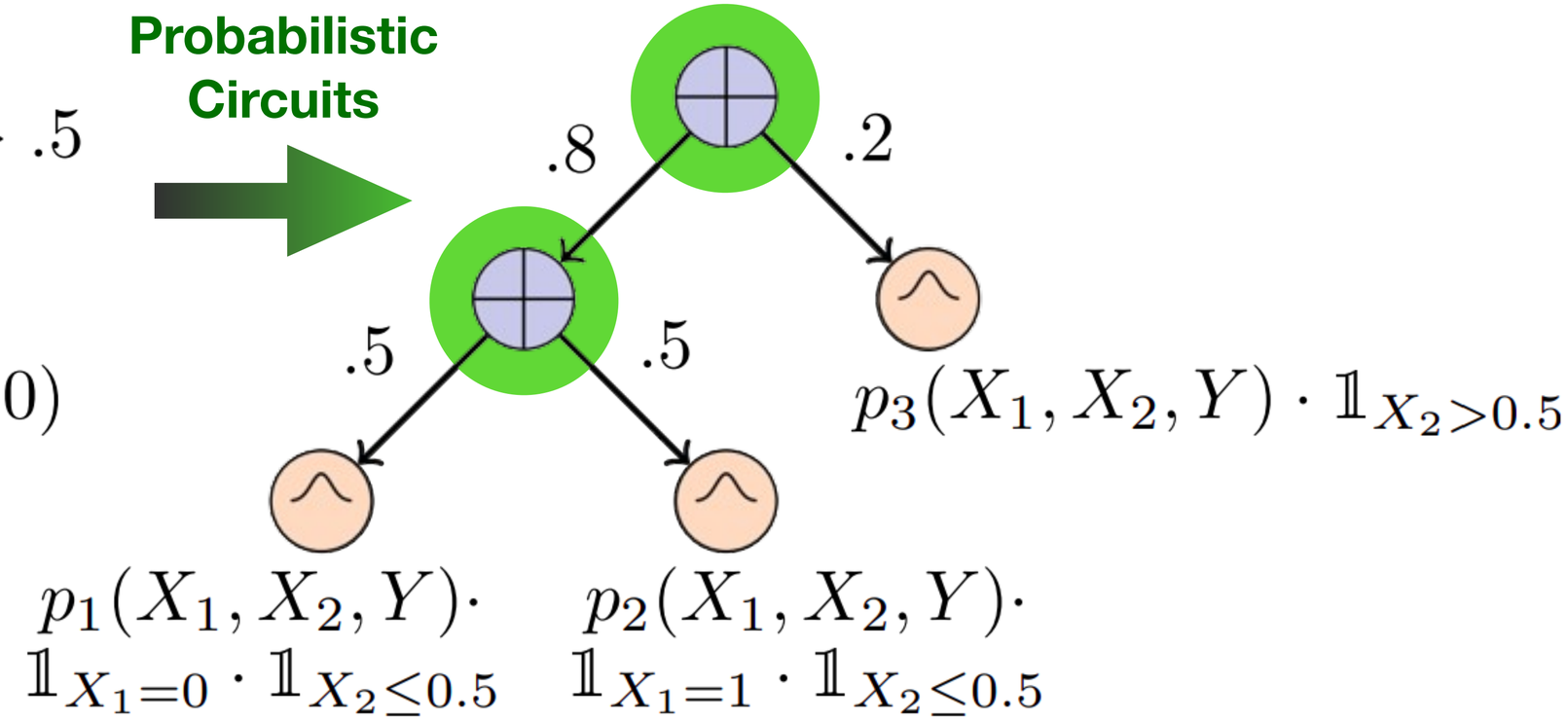
2020

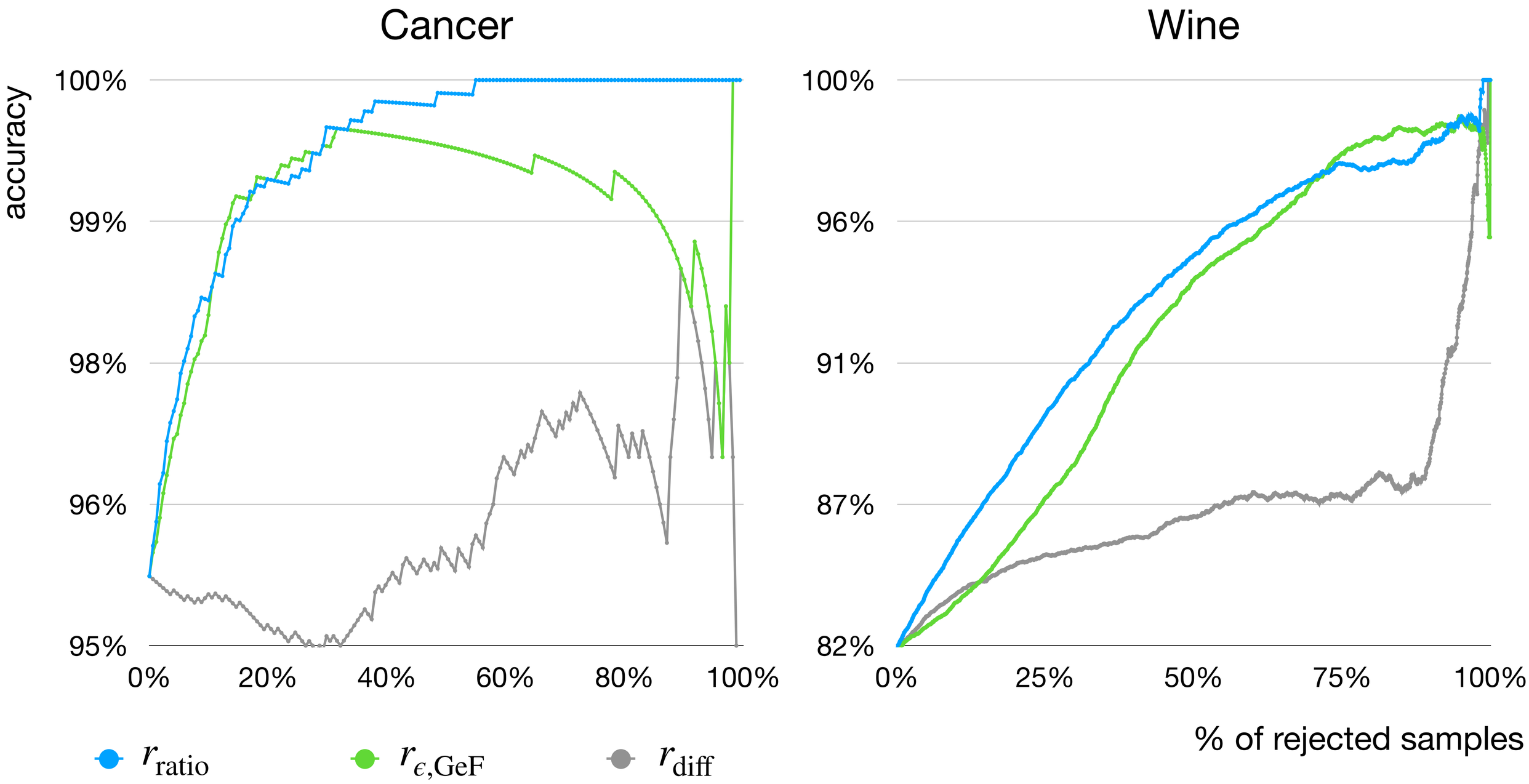
Decision Tree (DT)

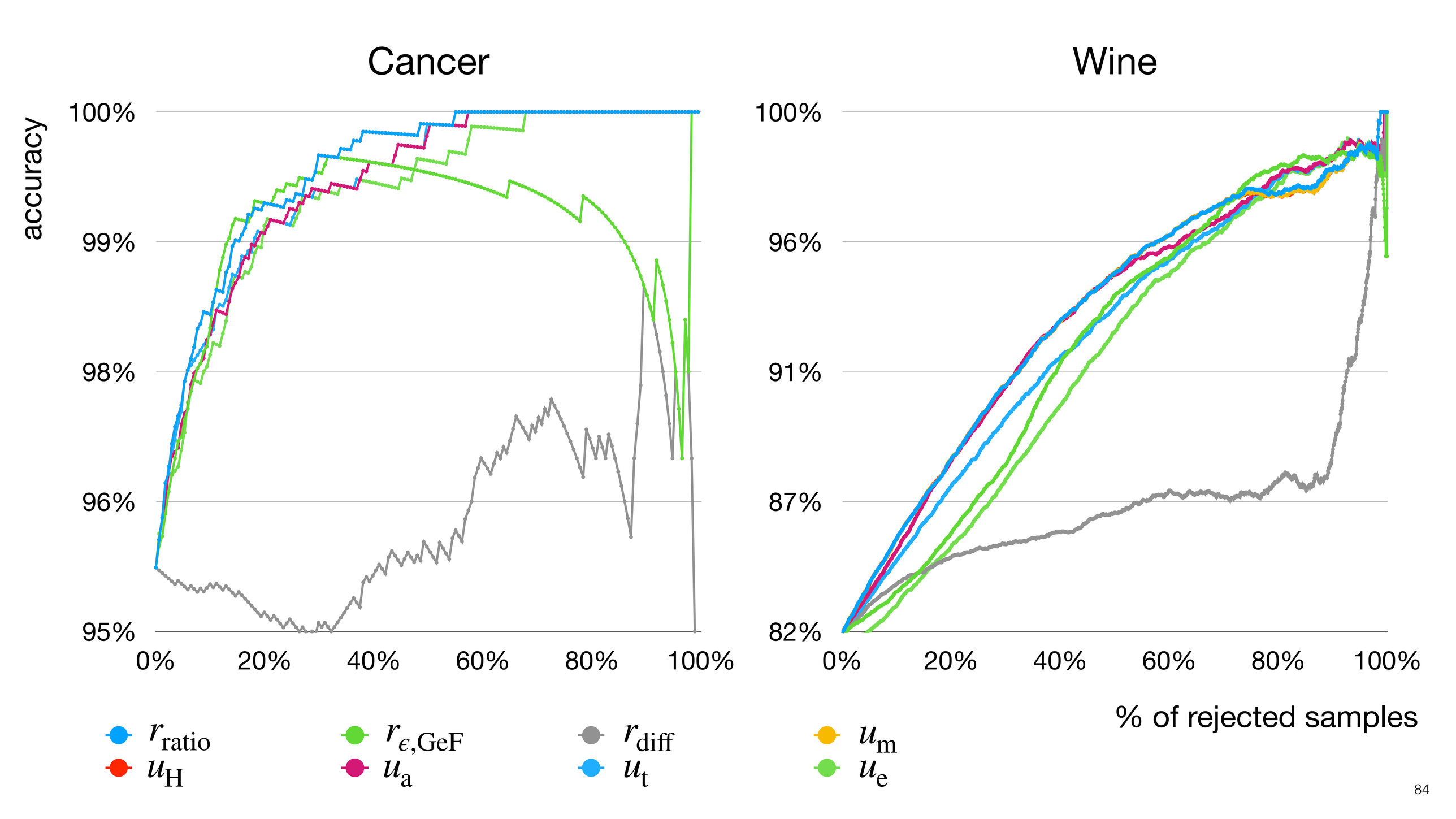


Generative DT

Probabilistic Circuits




Cancer
Wine
accuracy
100%
99%
98%
96%
95%
0%
20%
40%
60%
80%
100%
100%
96%
91%
87%
82%
0%
25%
50%
75%
100%
% of rejected samples
r_{ratio}
$r_{\epsilon,\text{GeF}}$
r_{diff}


Cancer
Wine
accuracy
100%
99%
98%
96%
95%
100%
96%
91%
87%
82%
0%
20%
40%
60%
80%
100%
% of rejected samples
r_{ratio}
$r_{\epsilon,\text{GeF}}$
r_{diff}
u_{m}
u_{H}
u_{a}
u_{t}
u_{e}

ROBUSTNESS QUANTIFICATION

- correlates nicely with accuracy ✓
- works for different types of model architectures ✓
- also works with global perturbations ✓
- is competitive with and complementary to UQ ✓
- is good with distribution shift and small data sets
- is more stable than UQ
- is not limited to epsilon-contamination ✓

CLASSIFICATION

features x

predicted class $\hat{y} = h(x)$

try to predict

FEATURES → **CLASS**

photo

medical data

...

label

best treatment

...

GENERATIVE PROBABILISTIC CLASSIFIER h

learn $P_{\text{classif}}(X, Y)$

predict the class y that maximizes $P_{\text{classif}}(y \mid x)$

GLOBAL

OTHER STUFF

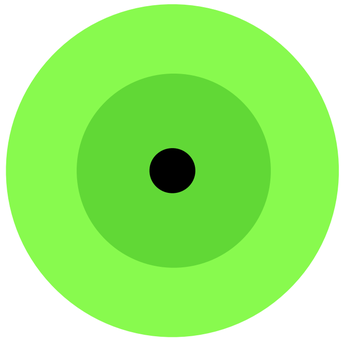
distance-based, …

at least one continuous feature:

$$r_{\text{ratio}} = \log \sqrt{\frac{p_{\text{classif}}(x, \hat{y})}{p_{\text{classif}}(x, \hat{y}_2)}}$$

$$= \frac{1}{2}\left(p_{\text{classif}}(x, \hat{y}) - p_{\text{classif}}(x, \hat{y}_2)\right)$$

$$\hat{y}_2 = \arg \max_{y \in \mathcal{Y} \setminus \{\hat{y}\}} P_{\text{classif}}(y \mid x)$$



$$\mathscr{P}_\delta$$

$$\|$$

$$\left\{P \in \mathbb{P} : \mathrm{d}(P_{\text{classif}}, P) < \delta\right\}$$

GLOBAL

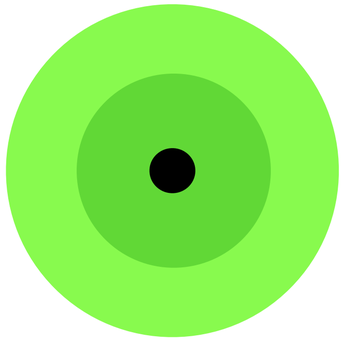
OTHER STUFF

distance-based, …

at least one continuous feature:

$$r_{\text{ratio}} = \log \sqrt{\frac{p_{\text{classif}}(x, \hat{y})}{p_{\text{classif}}(x, \hat{y}_2)}} = \log \sqrt{\frac{p_{\text{classif}}(\hat{y} \mid x)}{p_{\text{classif}}(\hat{y}_2 \mid x)}}$$

$$r_{\text{[illegible]}} = \frac{1}{2}\left(p_{\text{classif}}(x, \hat{y}) - p_{\text{classif}}(x, \hat{y}_2)\right)$$



$$\hat{y}_2 = \arg \max_{y \in \mathcal{Y} \setminus \{\hat{y}\}} P_{\text{classif}}(y \mid x)$$

$$\mathscr{P}_\delta$$

$$\|$$

$$\left\{P \in \mathbb{P} : \mathrm{d}(P_{\text{classif}}, P) < \delta\right\}$$

CLASSIFICATION

features x

predicted class $\hat{y} = h(x)$

try to predict

FEATURES → **CLASS**

photo

medical data

...

DISCRIMINATIVE PROBABILISTIC CLASSIFIER h

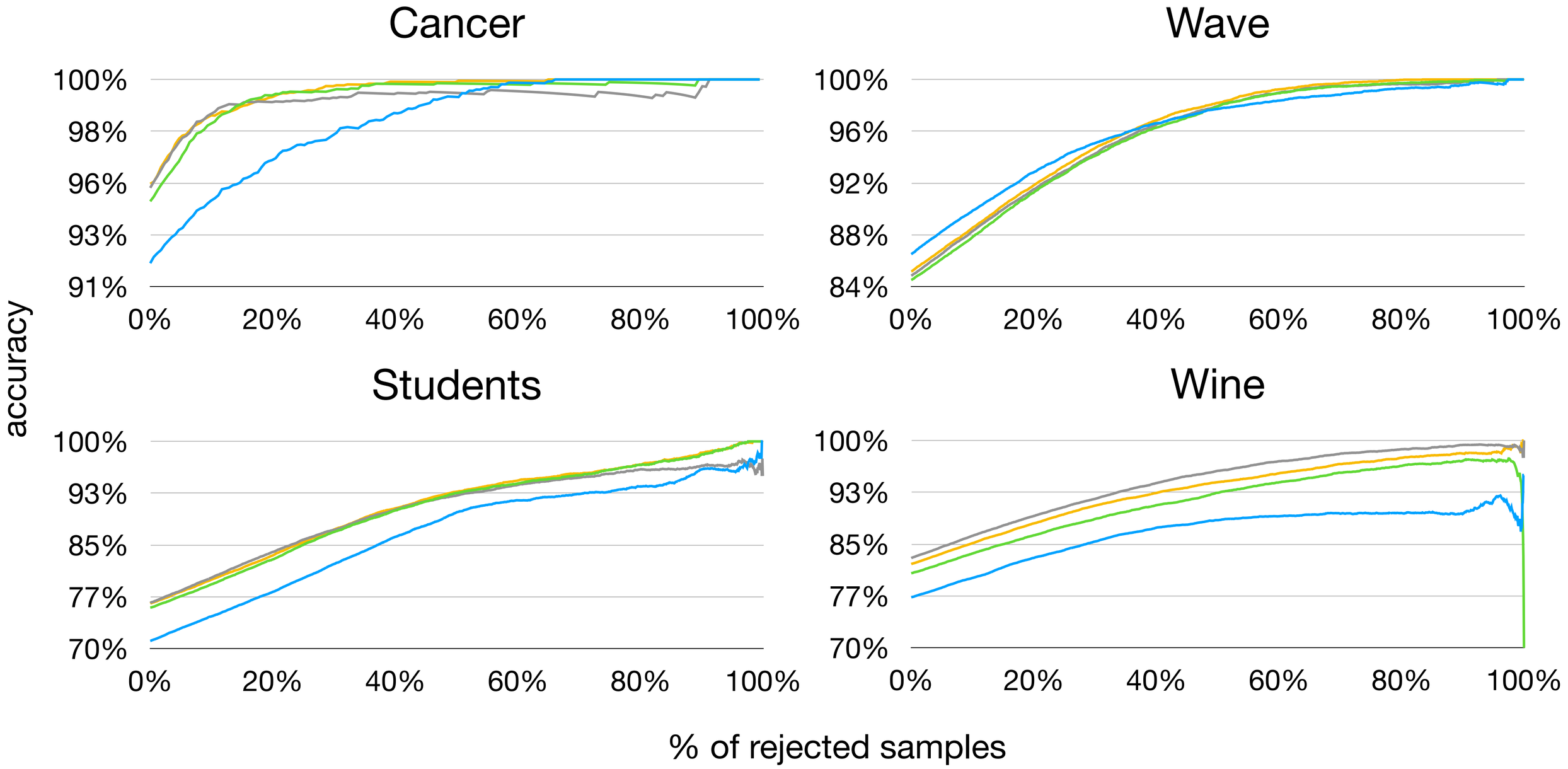
label

best treatment

...

learn $P_{\text{classif}}(Y \mid x)$

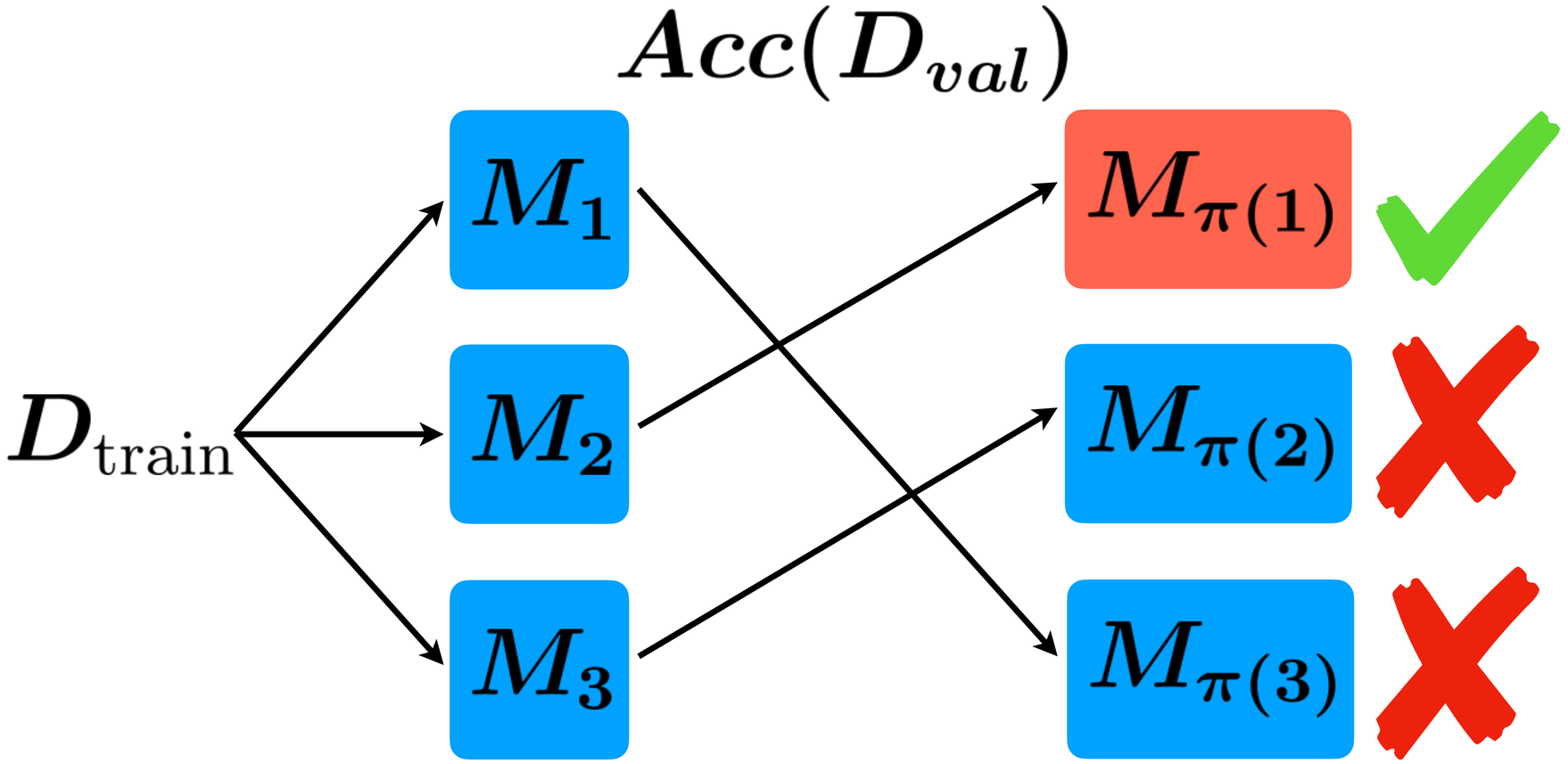
predict the class y that maximizes $P_{\text{classif}}(y \mid x)$


Cancer
Wave
Students
Wine
accuracy
100%
98%
96%
93%
91%
92%
88%
84%
85%
77%
70%
0%
20%
40%
60%
80%
100%
% of rejected samples
Logistic Regression
Gradient Boosting
Random Forest
XGBoost

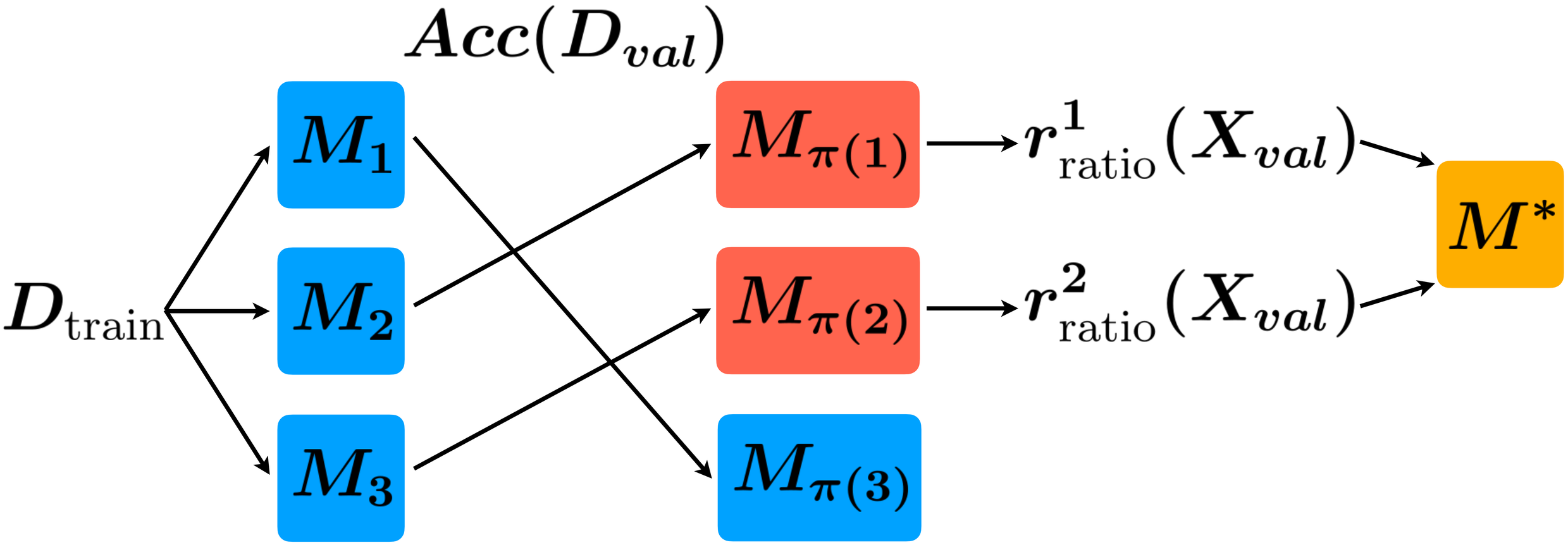
ROBUSTNESS QUANTIFICATION

- correlates nicely with accuracy
- works for different types of model architectures ✓
- also works with global perturbations
- is competitive with and complementary to UQ
- is good with distribution shift and small data sets
- is more stable than UQ
- is not limited to epsilon-contamination
- also works for discriminative classifiers ✓

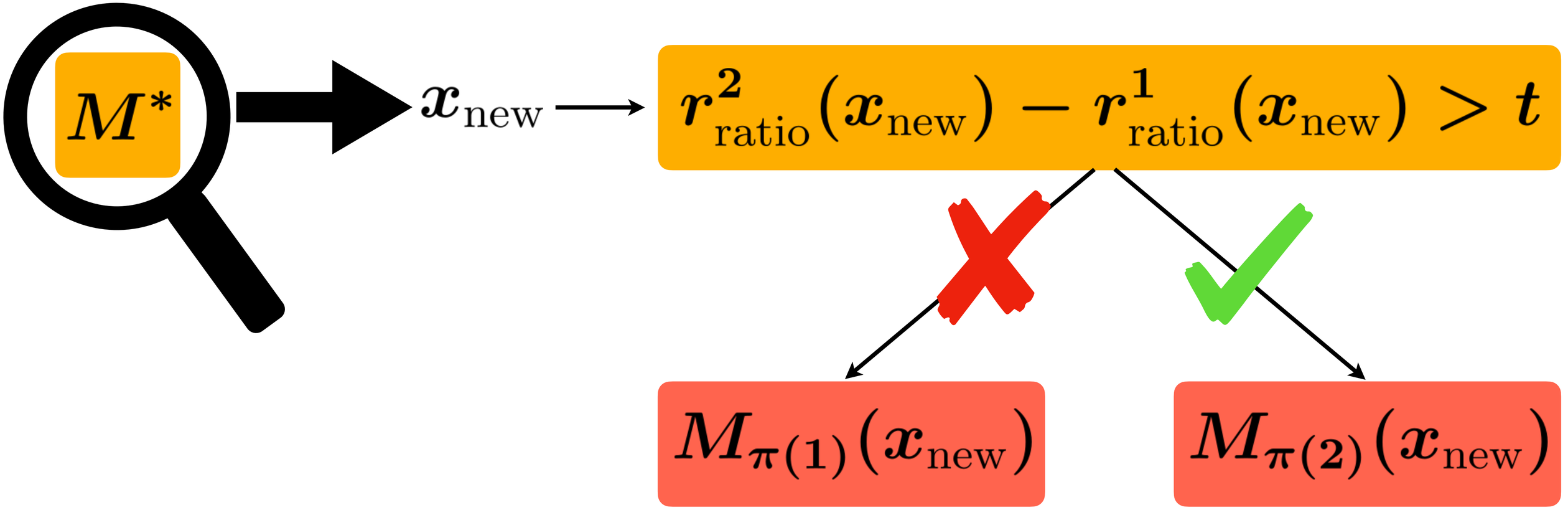
MODEL SELECTION



INSTANCE-WISE MODEL SELECTION

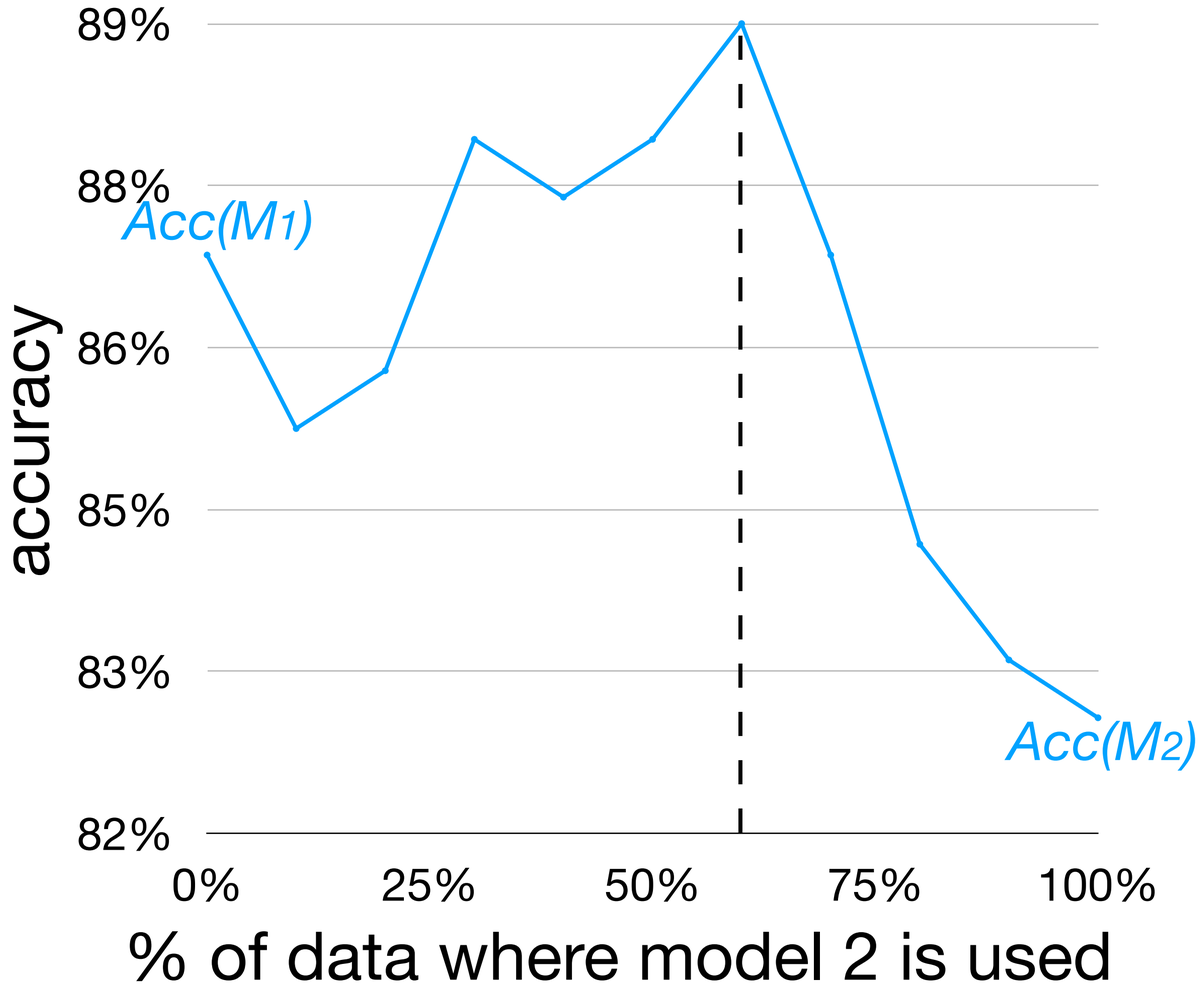


INSTANCE-WISE MODEL SELECTION

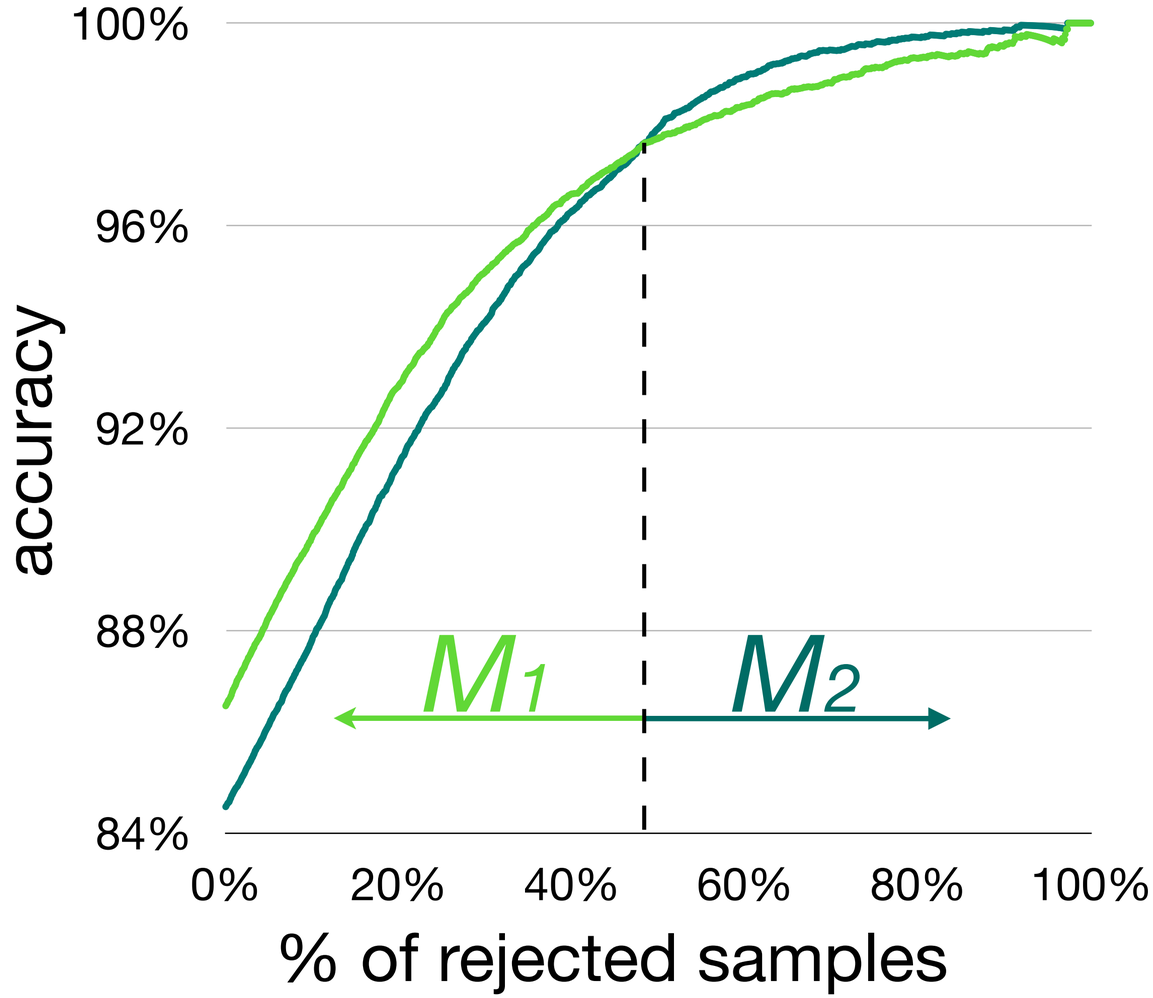


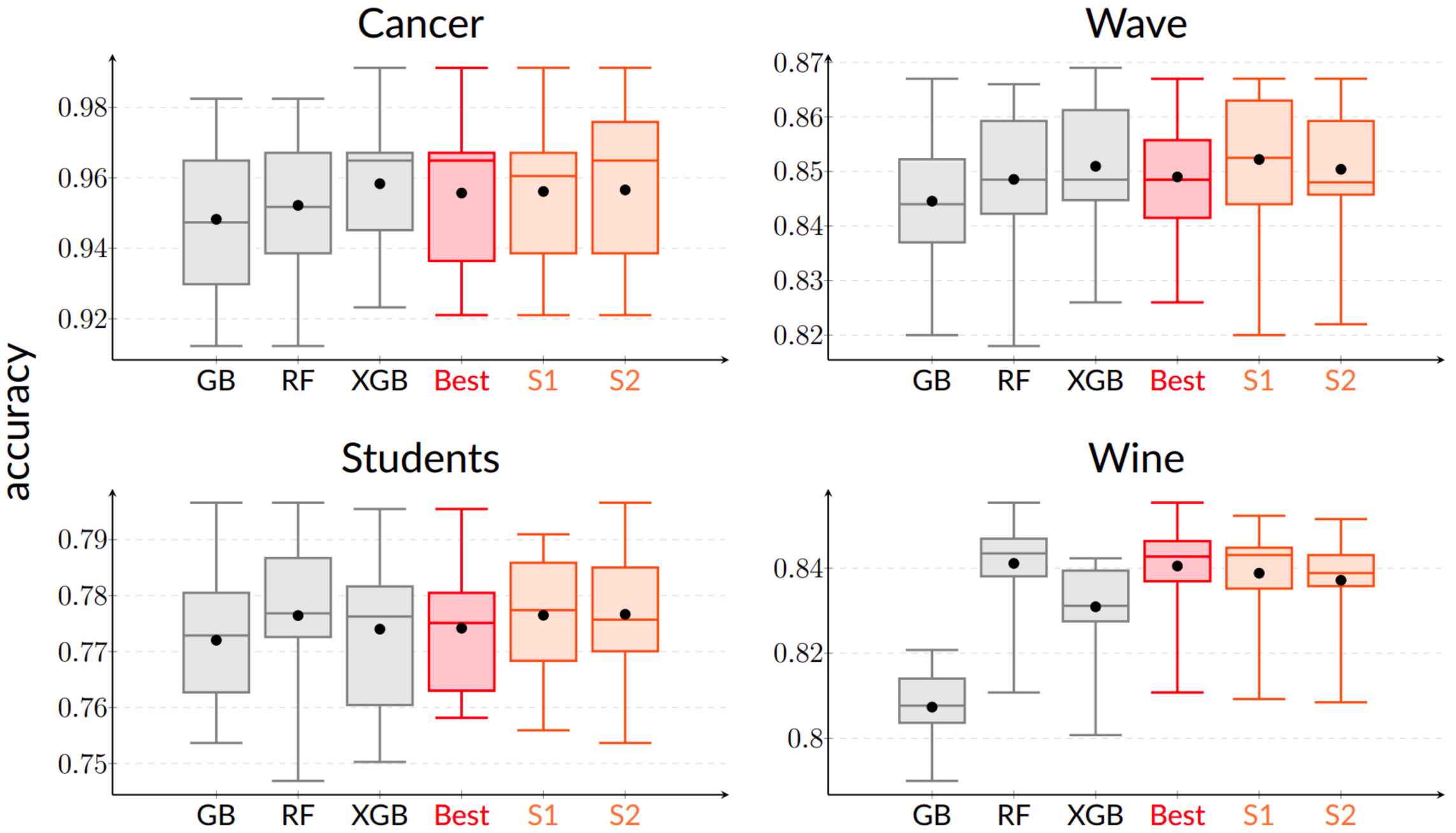
INSTANCE-WISE MODEL SELECTION

Strategy 1



Strategy 2




Cancer
Wave
Students
Wine
accuracy
0.98
0.96
0.94
0.92
0.87
0.86
0.85
0.84
0.83
0.82
0.79
0.78
0.77
0.76
0.75
0.84
0.82
0.8
GB
RF
XGB
Best
S1
S2

ROBUSTNESS QUANTIFICATION

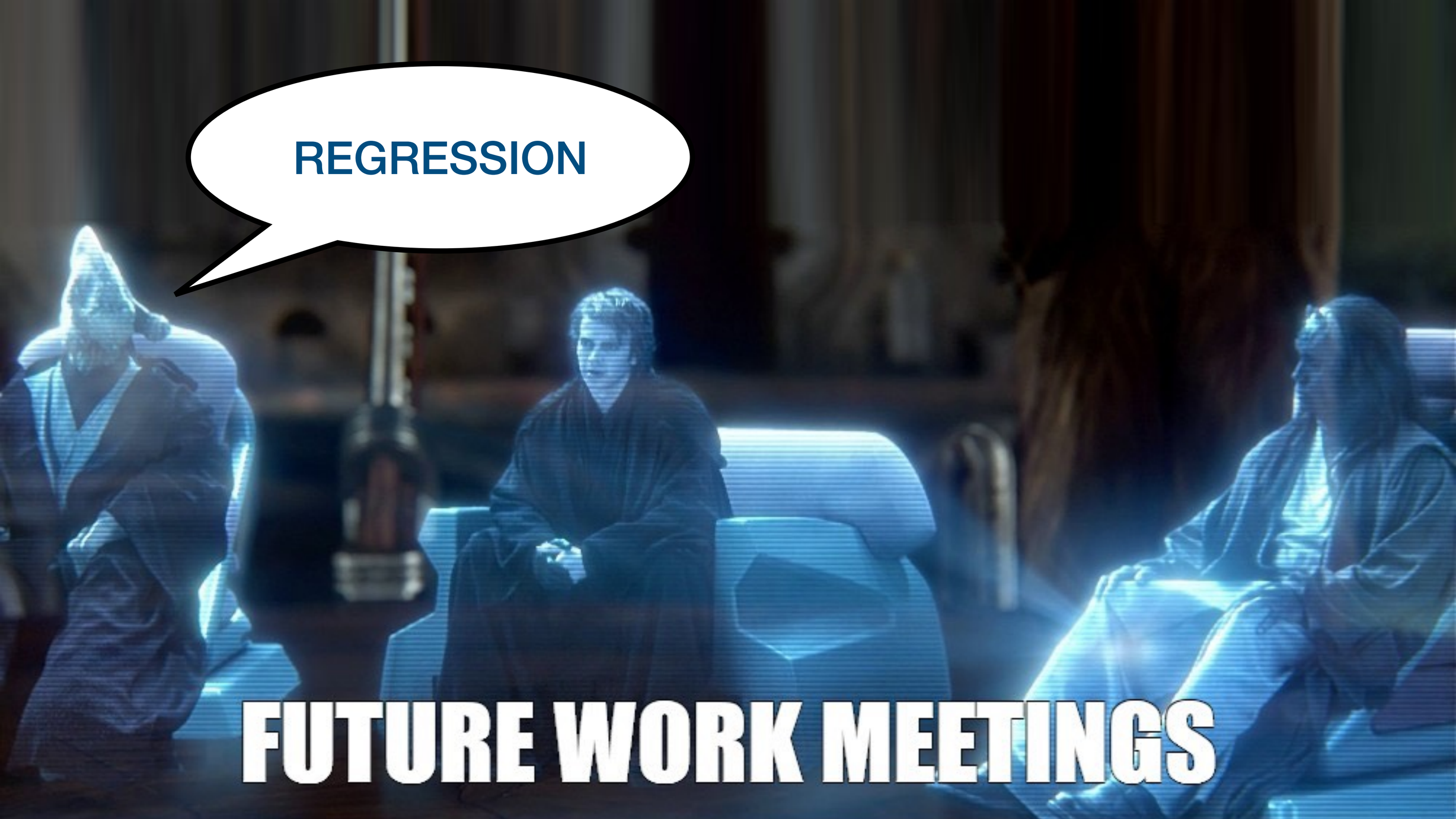
- correlates nicely with accuracy
- works for different types of model architectures ✓
- also works with global perturbations
- is competitive with and complementary to UQ
- is good with distribution shift and small data sets
- is more stable than UQ
- is not limited to epsilon-contamination
- also works for discriminative classifiers ✓
- can be used for instance-wise model selection ✓

ROBUSTNESS QUANTIFICATION

- correlates nicely with accuracy ✓
- works for different types of model architectures ✓
- also works with global perturbations ✓
- is competitive with and complementary to UQ ✓
- is good with distribution shift and small data sets ✓
- is more stable than UQ ✓
- is not limited to epsilon-contamination ✓
- also works for discriminative classifiers ✓
- can be used for instance-wise model selection ✓


FUTURE WORK MEETINGS


CALIBRATION
FUTURE WORK MEETINGS


REGRESSION
FUTURE WORK MEETINGS


CONFORMAL
PREDICTION
FUTURE WORK MEETINGS

jasper.debock@ugent.be
adrian.detavernier@ugent.be
rodrigo.lassance@ugent.be

[1] Global Sensitivity Analysis for MAP Inference in Graphical Models. De Bock, de Campos & Antonucci. 2014.

[2] Credal sum-product networks.
Mauá, Cozman, Conaty & de Campos. 2017.

[3] Robustifying sum-product networks.
Mauá, Conaty, Cozman, Poppenhaeger & de Campos. 2018.

[4] Towards Scalable and Robust Sum-Product Networks.
Correia & de Campos. 2019.

[5] Towards Robust Classification with Deep Generative Forests.
Correia, Peharz & de Campos. 2020

[6] Robustness quantification: a new method for assessing the reliability of the predictions of a classifier.
De Tavernier, De Bock. 2025.

[7] Robustness and uncertainty: two complementary aspects of the reliability of the predictions of a classifier.
De Tavernier, De Bock. 2025.

[8] A hierarchy of sum-product networks using robustness.
Conaty, Martínez del Rincon & de Campos. 2019.

[9] A Robust Dynamic Classifier Selection Approach for Hyperspectral Images with Imprecise Label Information.
Li, Huang, De Bock & Pižurica. 2020.