https://github.com/fmohr/lcdb

```
pip install lcdb
```

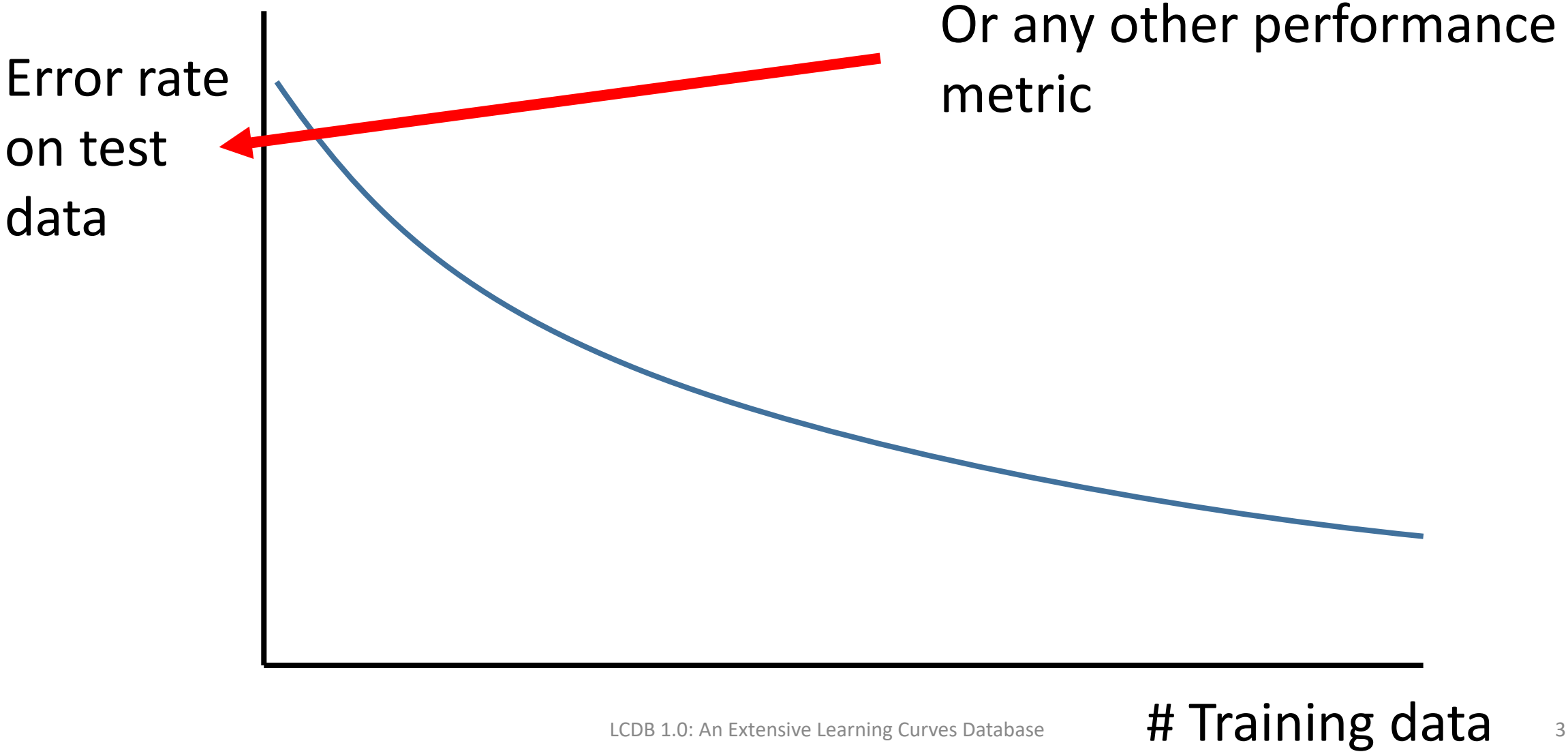
LCDB 1.0: An Extensive Learning Curves Database for Classification Tasks

Felix Mohr, Tom J Viering, Marco Loog, Jan N van Rijn

Overview

- Introduction: learning curves
  - What is a learning curve?
  - What can you do with them?
  - Why should we study them more?
- Our database
- Some preliminary results
- Discussion
- Version 2.0 of the database

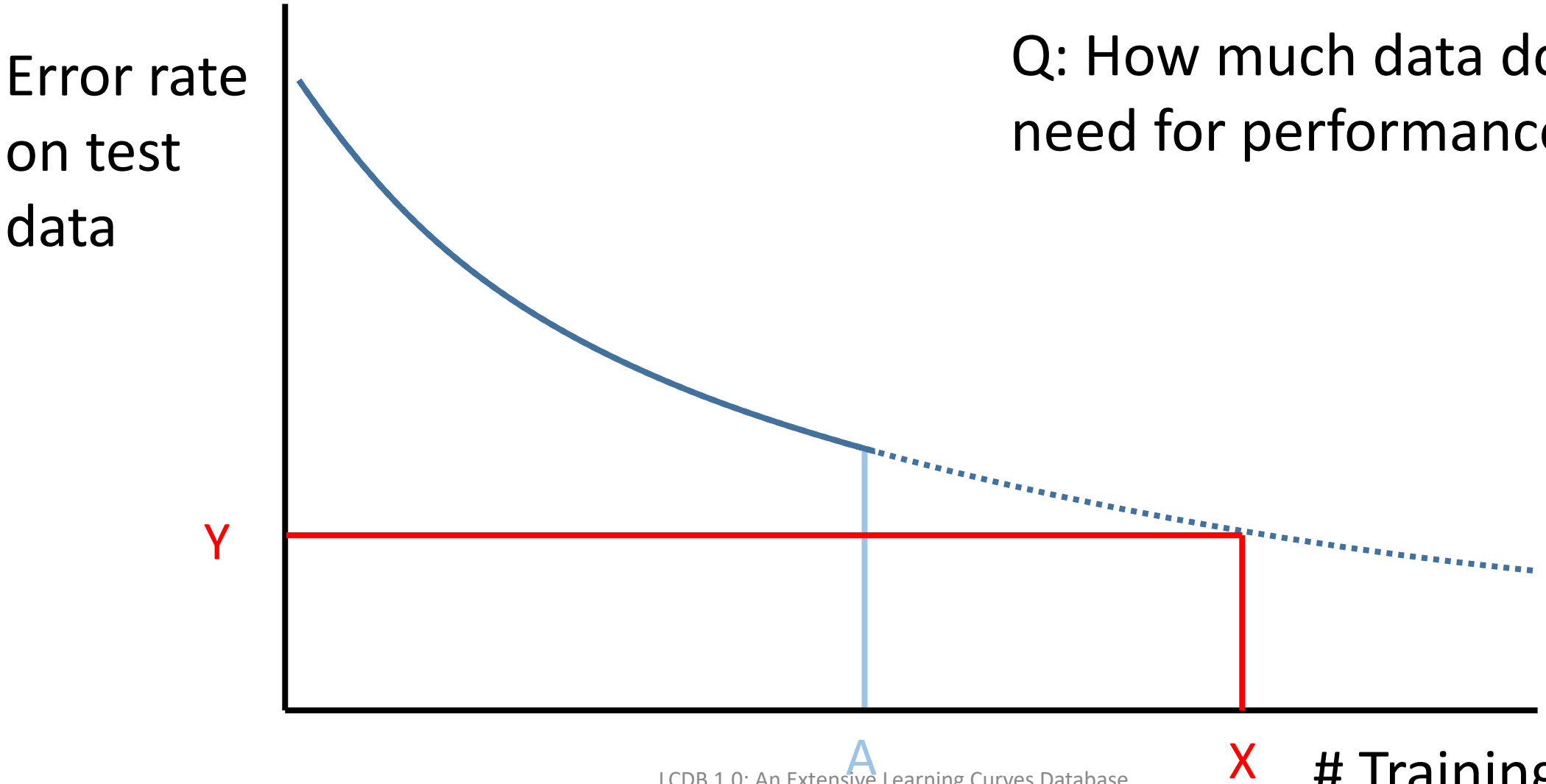
Learning Curve



What can we do with learning curves?

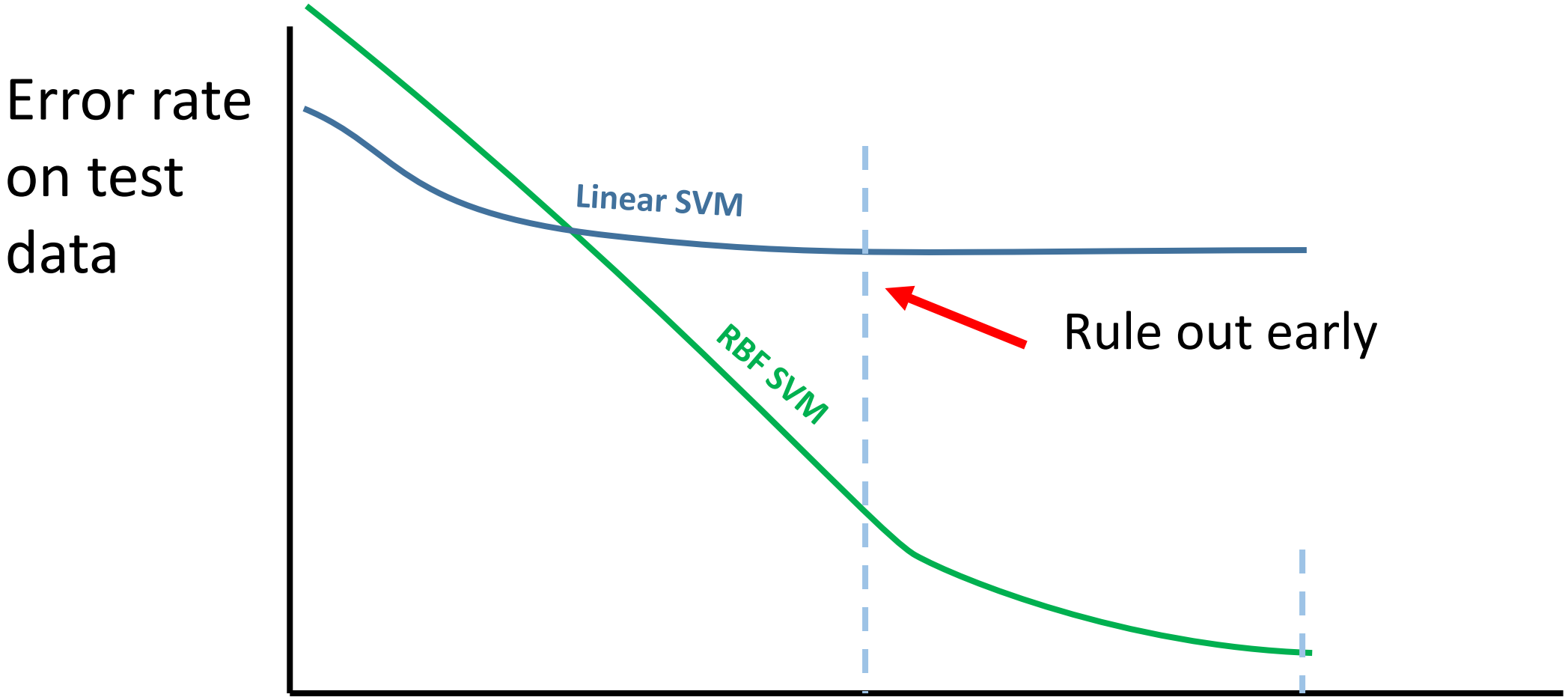
- Estimate the value of additional data
- Faster model selection
- Faster training

Data Value



Q: How much data do we need for performance Y?

Faster Model Selection

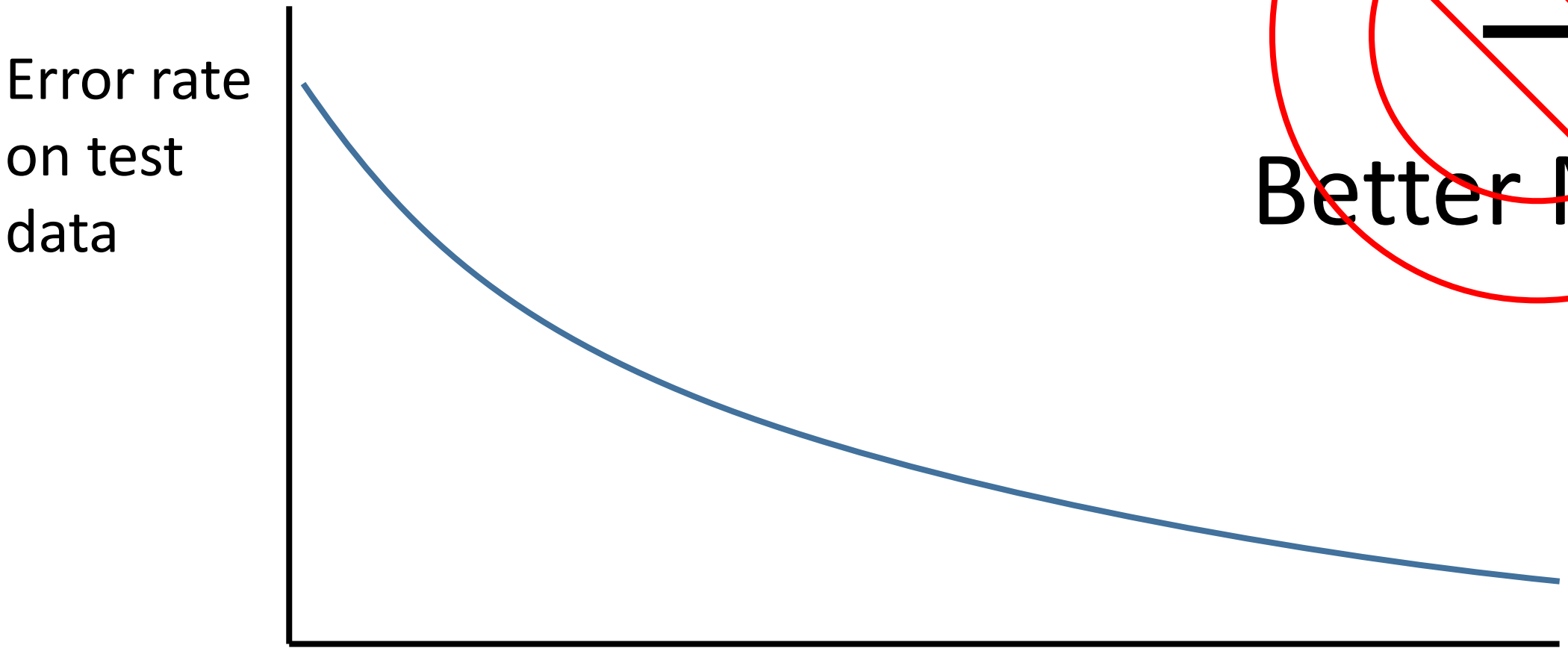


Mohr, Felix, and Jan N. van Rijn. "Towards model selection using learning curve cross-validation." *AutoML workshop at ICML* 2021.

What can we do with learning curves?

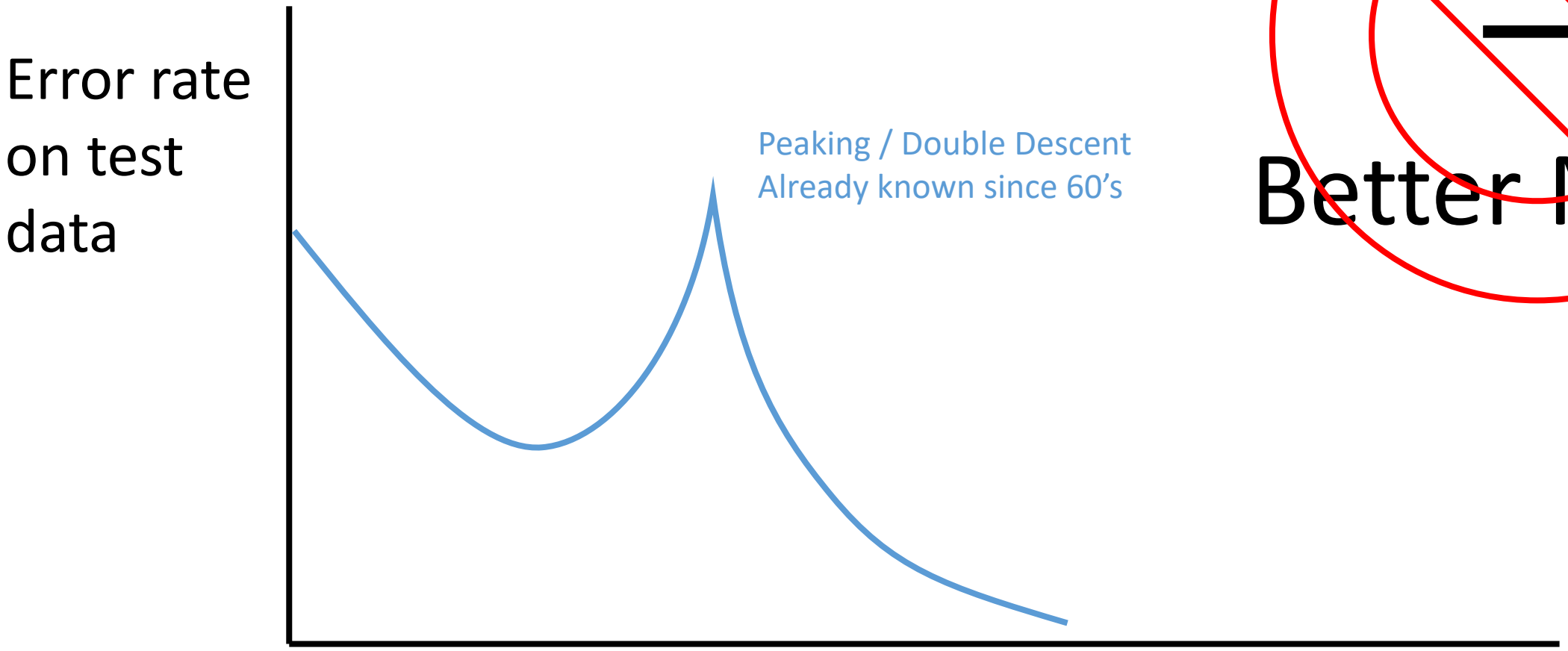
- Estimate the value of additional data
- Faster model selection
- Faster training

Why study Learning Curves?



[1] Loog, Marco, et al. "A brief prehistory of double descent." *PNAS* 2020
[2] Loog, Marco, Tom Viering, and Alexander Mey. "Minimizers of the empirical risk and risk monotonicity." *NeuRIPS* 2019
[3] Loog, Marco, and Robert PW Duin. "The dipping phenomenon."*IAPR SPR SSPR* 2012.

Why study Learning Curves?

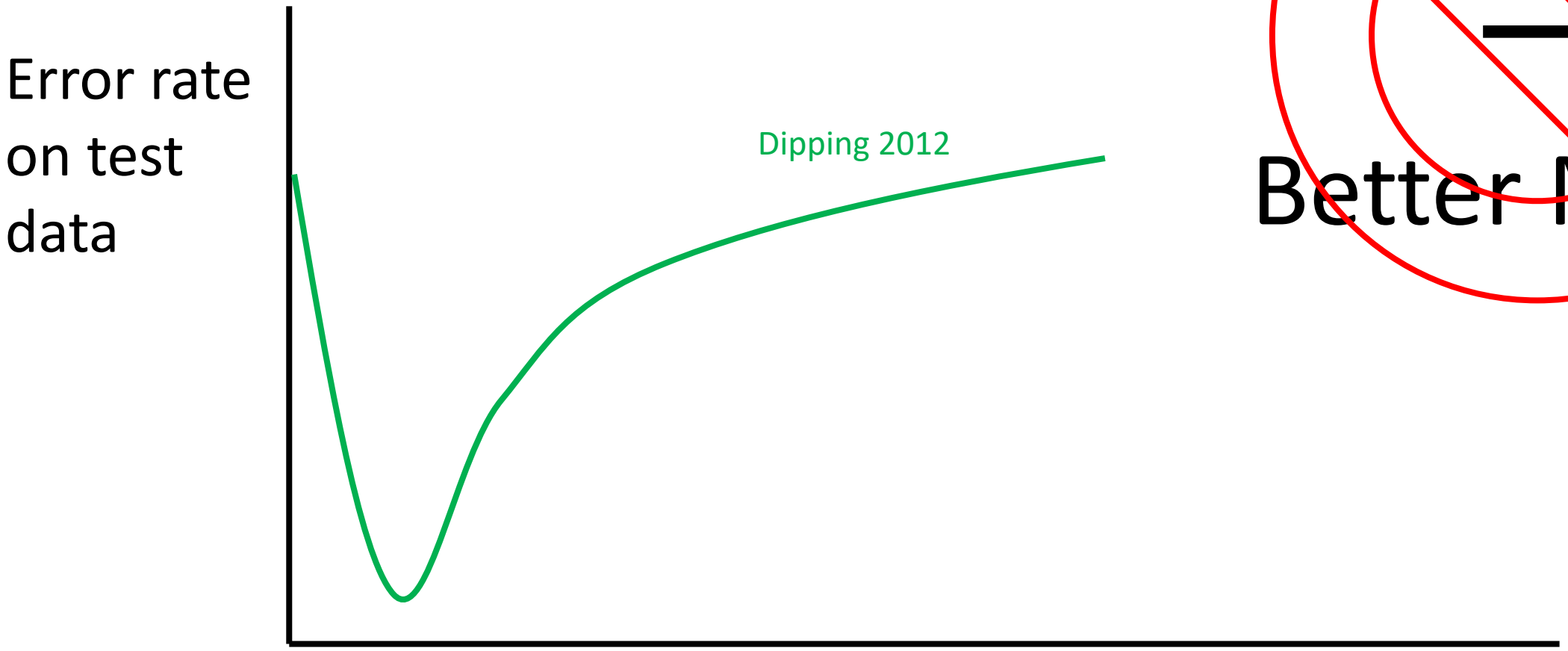


[1] Loog, Marco, et al. "A brief prehistory of double descent." *PNAS* 2020

[2] Loog, Marco, Tom Viering, and Alexander Mey. "Minimizers of the empirical risk and risk monotonicity." *NeuRIPS* 2019

[3] Loog, Marco, and Robert PW Duin. "The dipping phenomenon."*IAPR SPR SSPR* 2012.

Why study Learning Curves?



[1] Loog, Marco, et al. "A brief prehistory of double descent." *PNAS* 2020
[2] Loog, Marco, Tom Viering, and Alexander Mey. "Minimizers of the empirical risk and risk monotonicity." *NeuRIPS* 2019
[3] Loog, Marco, and Robert PW Duin. "The dipping phenomenon."*IAPR SPR SSPR* 2012.

Why study Learning Curves?

[1] Loog, Marco, et al. "A brief prehistory of double descent." *PNAS* 2020
[2] Loog, Marco, Tom Viering, and Alexander Mey. "Minimizers of the empirical risk and risk monotonicity." *NeuRIPS* 2019
[3] Loog, Marco, and Robert PW Duin. "The dipping phenomenon."*IAPR SPR SSPR* 2012.

Why study Learning Curves?

- Ill-behaving curves
- Do these behaviours also occur in “real-world” settings?
  - Do you have strange learning curves? Please let us know!

- We have written 2 surveys covering learning curve literature [1,2]
- **No consensus** on learning curve shape: exponential, power law, …???
  - PAC learning doesn’t say much about the shape [2,3]!
  - Scaling laws for CNN’s: power law seems to work relatively well [2]
  - What about tabular data / typical learners such as SVM’s, KNN, etc.?

What is the shape….?!
Our database (this talk)

[1] Mohr, F., van Rijn, J.N.: Learning curves for decision making in supervised machine learning - A survey. Arxiv 2022
[2] Viering, T.J., Loog, M.: The shape of learning curves: a review. Arxiv 2021
[3] Bousquet, Olivier, et al. A theory of universal learning. ACM SIGACT STOC. 2021.

Our database

- 20 learners on 246 datasets (openML)
- Training set sizes $s_i = \left\lceil 2^{(7+i)/2} \right\rceil$, $s = [16,23,32, \ldots]$
- Bootstrapping: 25 train / validation / test splits
- Precomputed error rate, F1, AUC ROC, log loss
- Provide all predictions (can compute any metric)
- Getting our data: `pip install lcdb`

Help us analyse all these curves!

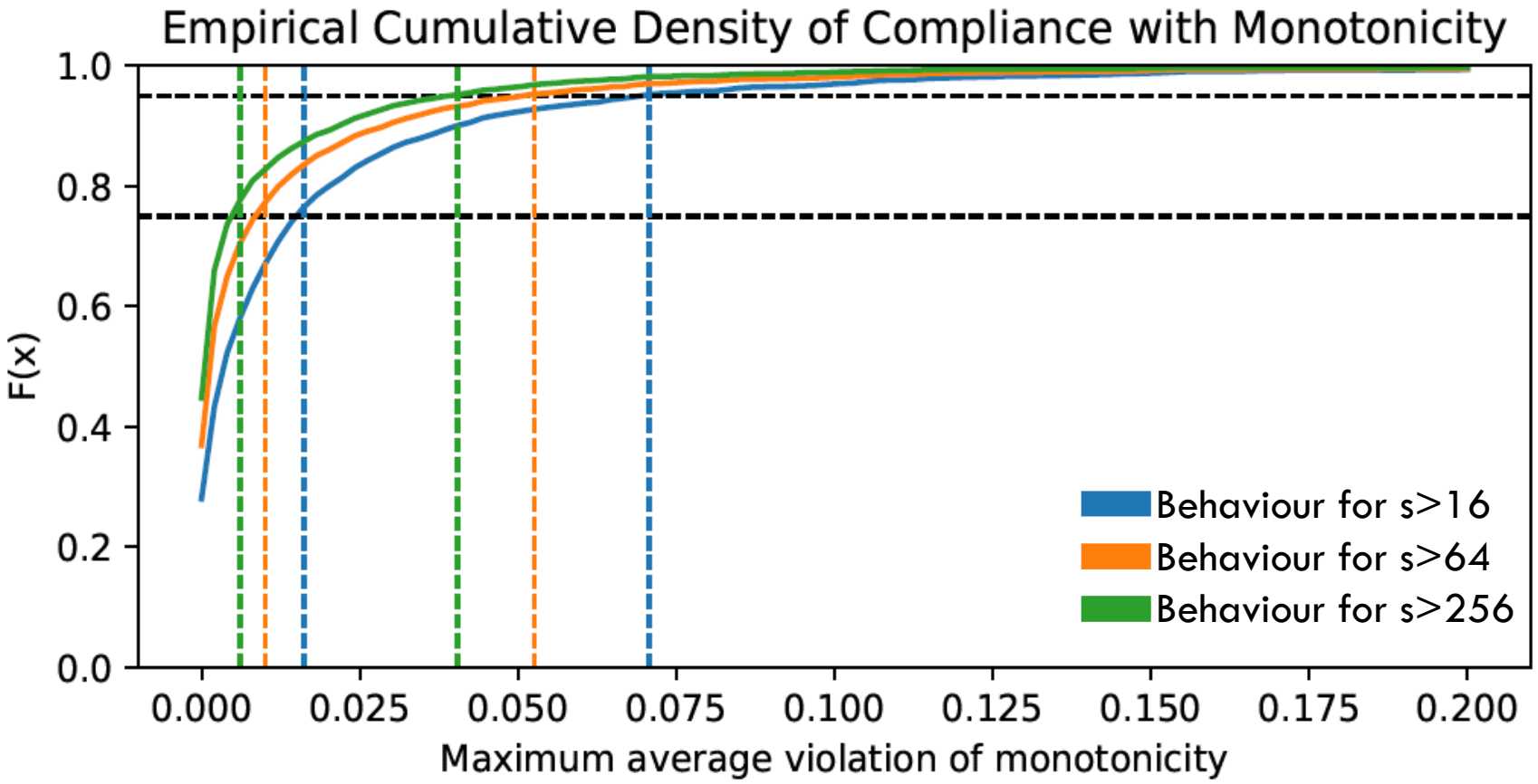
$20 \times 246 \times 25 \times 4 \approx$ half a million learning curves

Some preliminary results

- Monotonicity
- Peaks
- Crossings
- Shape: power law, exponential, …?

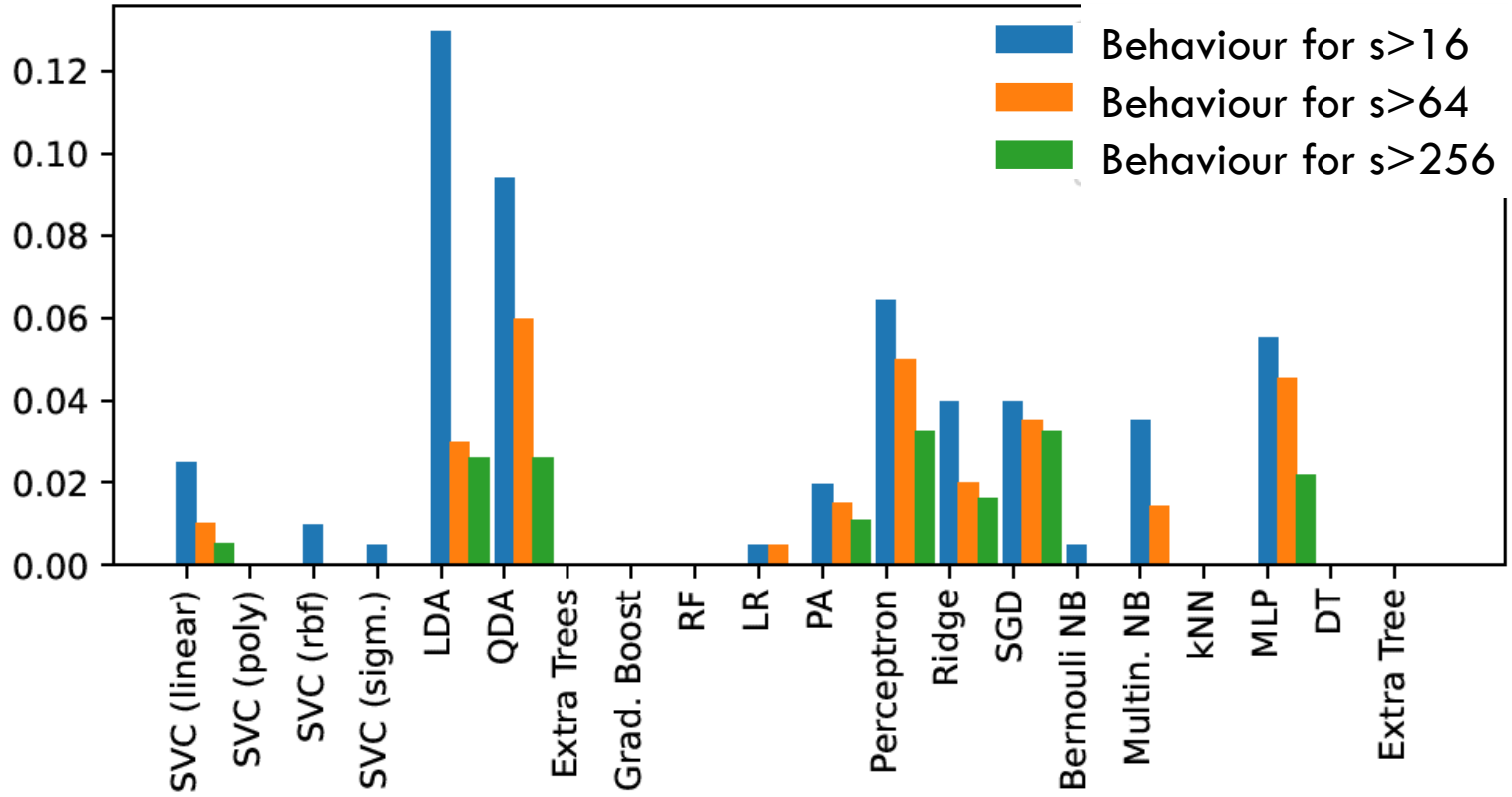
Monotone?

- Define x measuring the violation of monotonicity

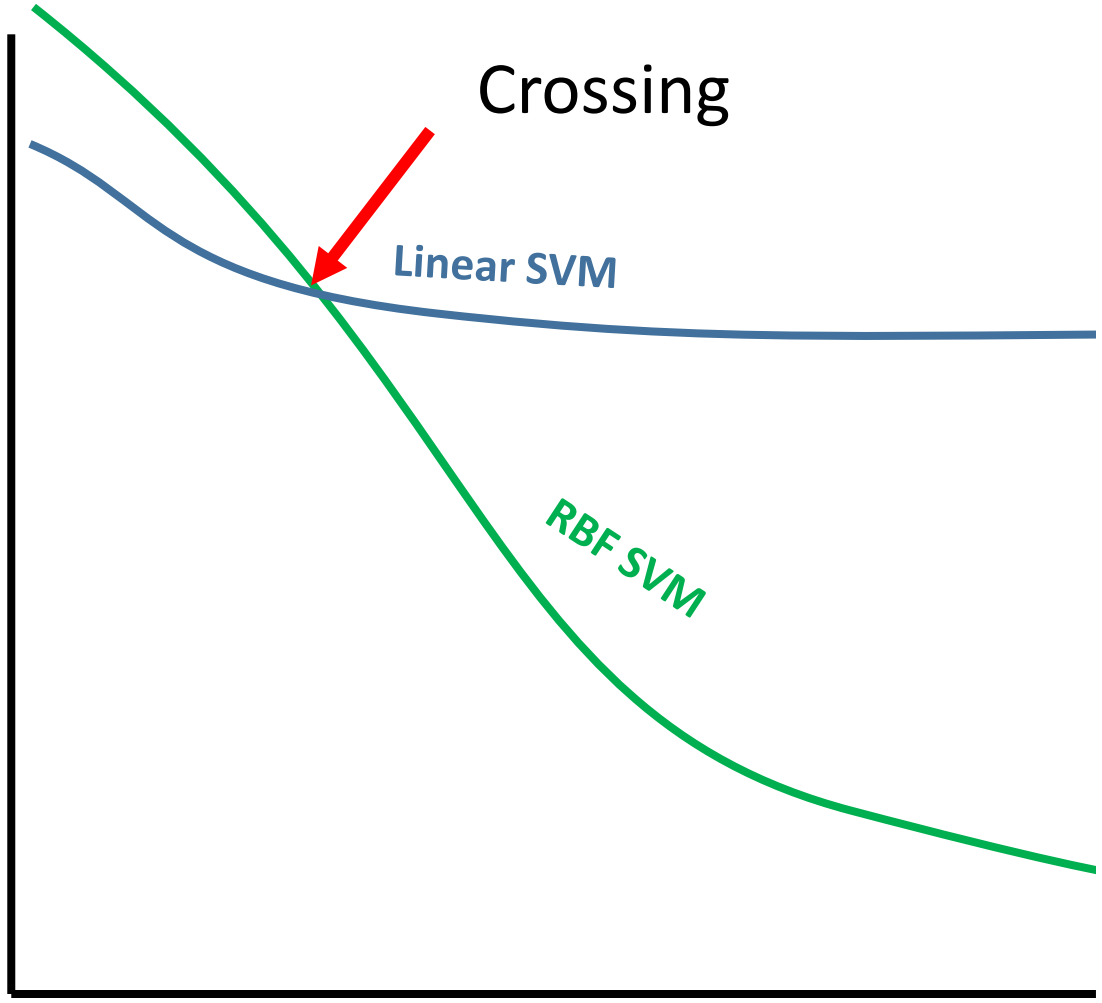


Peaks?

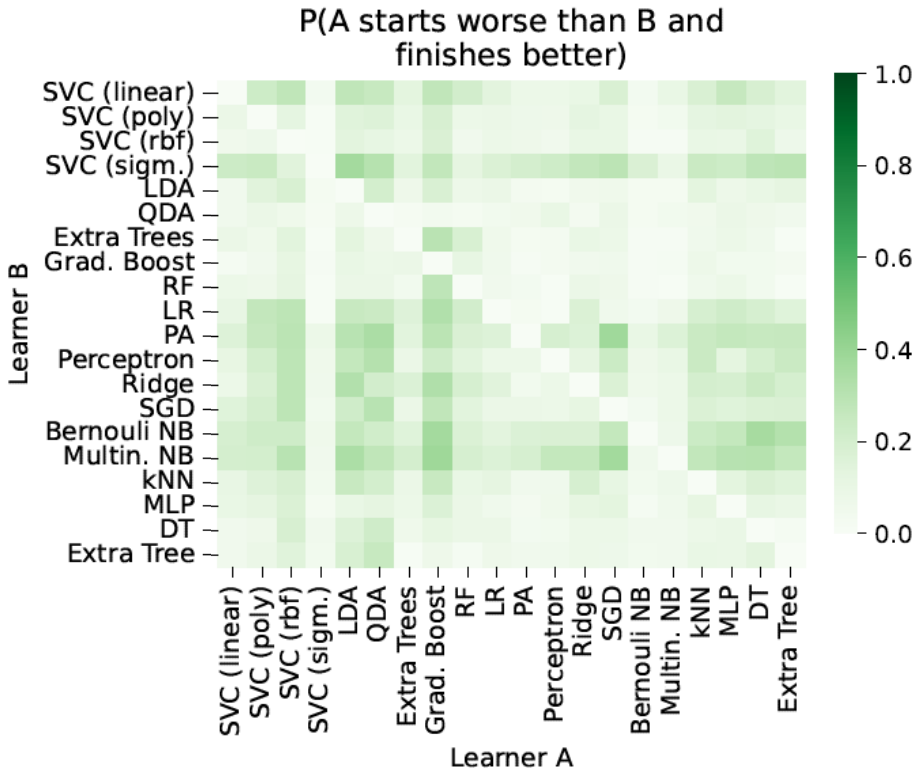
- Define “peak”: percentage of datasets with a peak in the curve



Do curves cross?



- Yes curves cross! 13%



Shape?

- We fit learning curves and extrapolate to the last point
- Mean Squared Error (MSE)
- Wbl4 and mmf4 perform best on average

Model	Formula	Model	Formula
last1	a	vap3	$\exp\left(a+\frac{b}{x}+c\log(x)\right)$
pow2	$-ax^{-b}$	expp3	$c-\exp\left((-b+x)^{a}\right)$
log2	$-a\log(x)+b$	expd3	$c-(-a+c)\,exp(-bx)$
exp2	$a\exp(-bx)$	logpow3	$a/\left((x\exp(-b))^{c}+1\right)$
lin2	$ax+b$	pow4	$a-b\,(d+x)^{-c}$
ilog2	$-a/\log(x)+b$	mmf4	$(ab+cx^{d})/(b+x^{d})$
pow3	$a-bx^{-c}$	wbl4	$-b\exp(-ax^{d})+c$
exp3	$a\exp(-bx)+c$	exp4	$c-\exp(-ax^{d}+b)$

Discussion

- **Surprisingly** 4 parameter models perform best for extrapolation
  - **Not** exponential or power law! (as suggested by most prior literature)
  - Avoided in prior studies due to overfitting concerns

- Curve fitting is *notoriously* difficult!
  - We average 25 learning curves to make curves smoother
  - **Still** 2% of fits **fail** (we use Levenberg-Marquadt, standard approach from literature)
  - Any curve fitting experts? Please come to us!

- Analysis is preliminary

Version 2.0 of the database

- Support for scikit pipelines
  - So we can do feature scaling, hyperparameter tuning, …
- We used default hyperparameters for all learners
  - Ideally tuned in 2.0
- We sample all training sets independently (bootstrap)
  - For 2.0 we want a variant where training sets s_i, s_{i+1} satisfy $s_{i+1} \subset s_i$

LCDB 1.0: An Extensive Learning Curves Database for Classification Tasks

Felix Mohr, Tom J Viering, Marco Loog, Jan N van Rijn





Help us analyse/fit curves!

https://github.com/fmohr/lcdb

```
pip install lcdb
```

Paper 1317

Poster tomorrow **Zone A**

Universidad de La Sabana

TU Delft

Universiteit Leiden