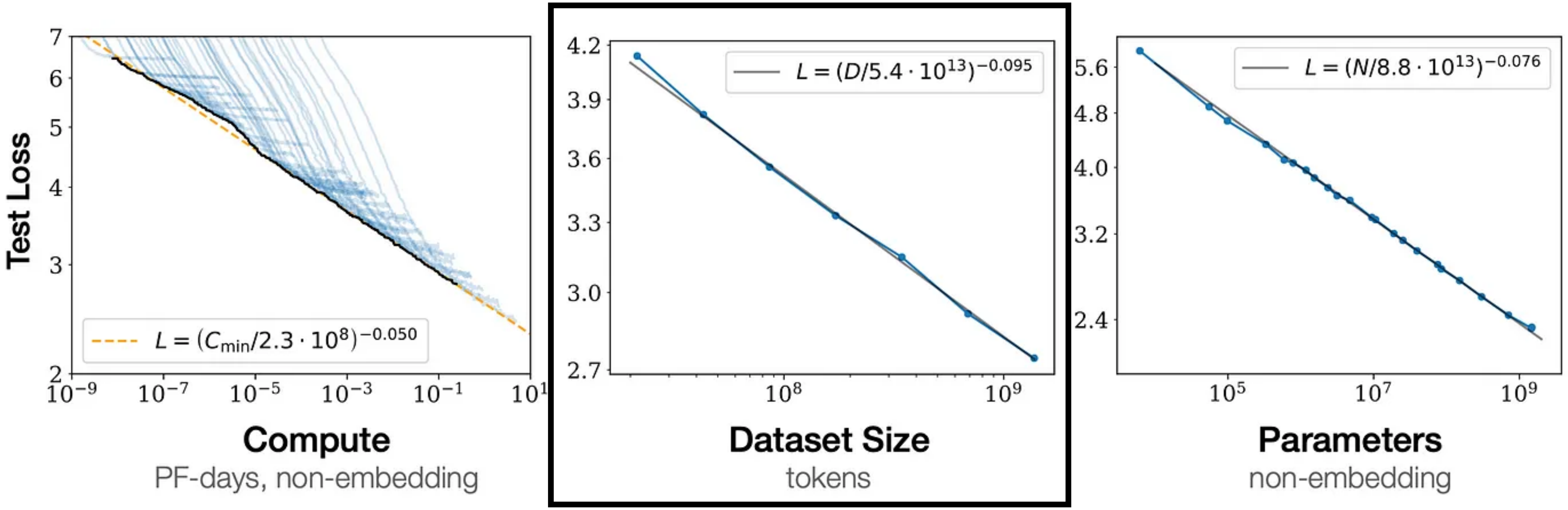
Towards a statistical theory of data selection under weak supervision

Germain Kolossov*, Andrea Montanari*, Pulkit Tandon*
Granica.ai

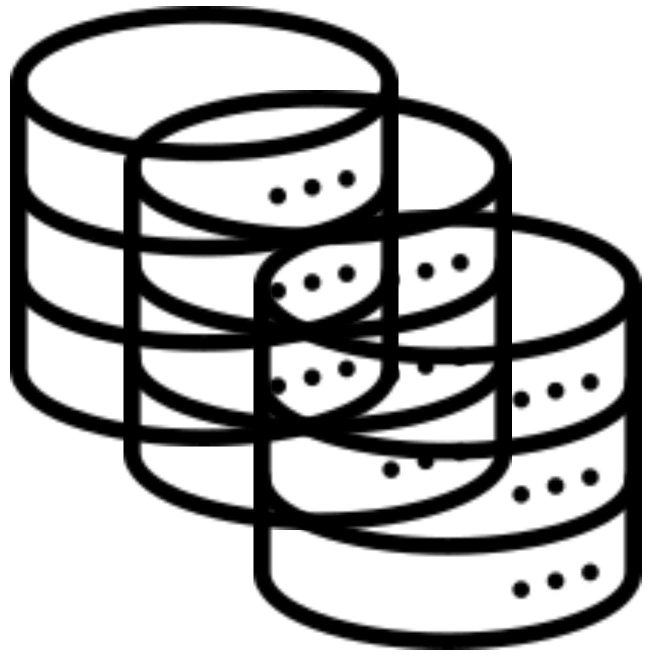

Modern AI is hungry for data!

Image credits: ChatGPT

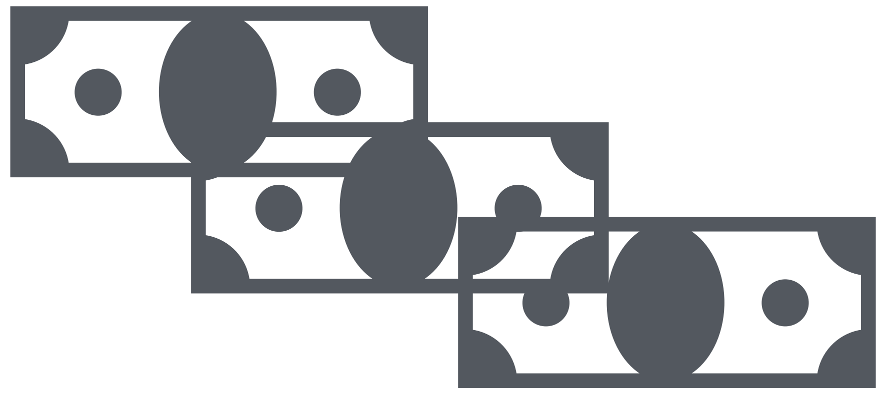
Modern AI is hungry for data!



Kaplan et al. (2020)



More data implies

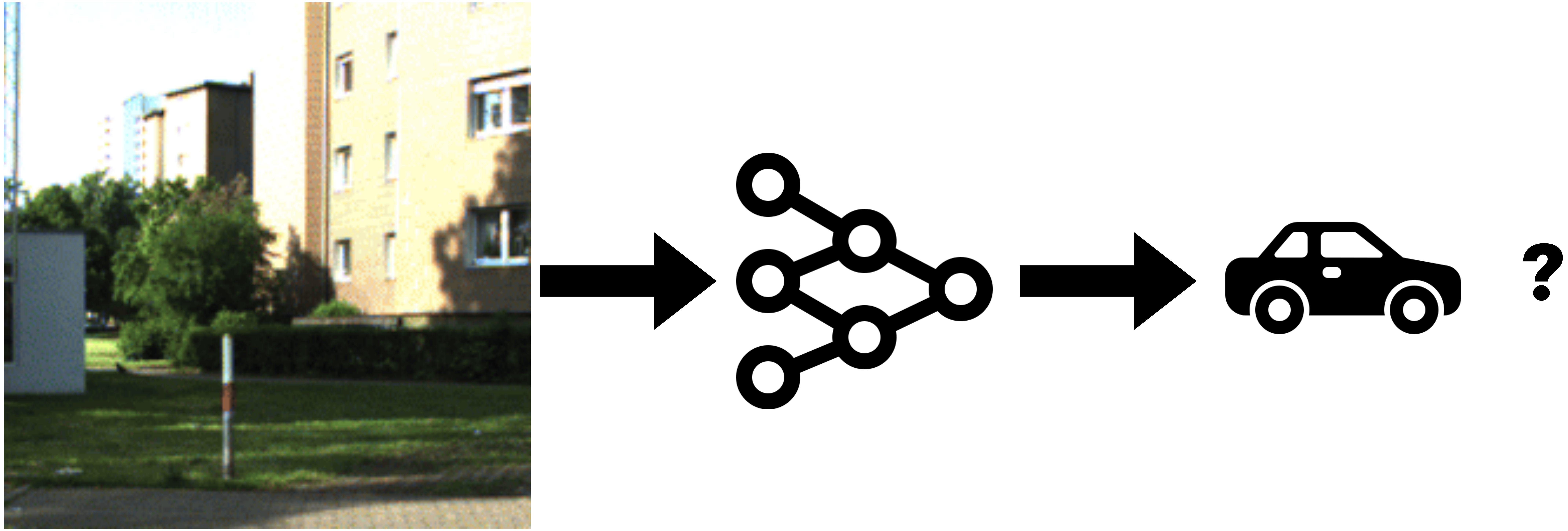


labeling, storage
costs

training
time

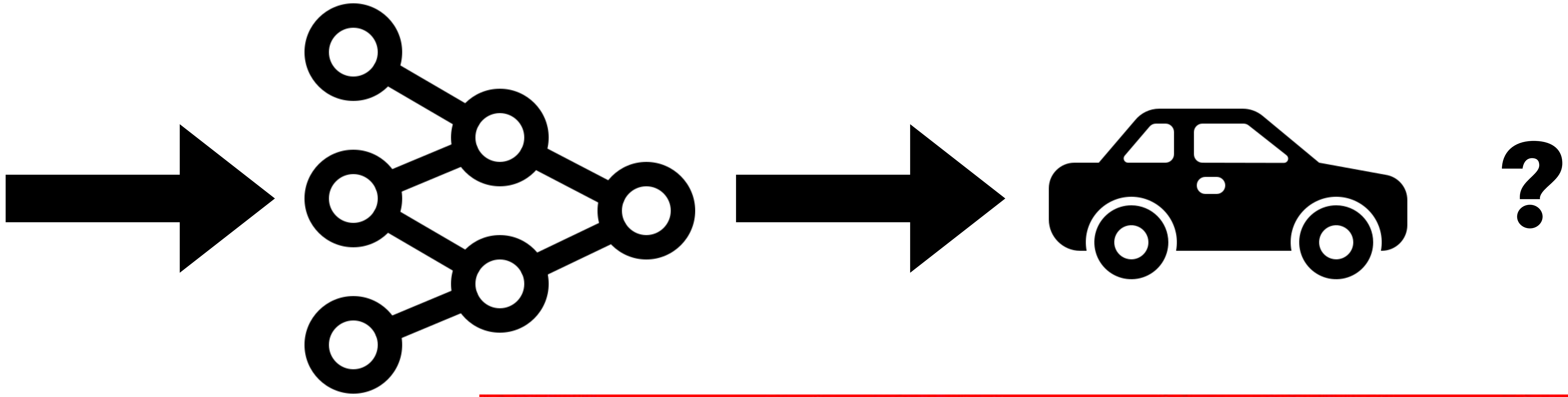
poorer
data quality

However, each datapoint does not contribute equally

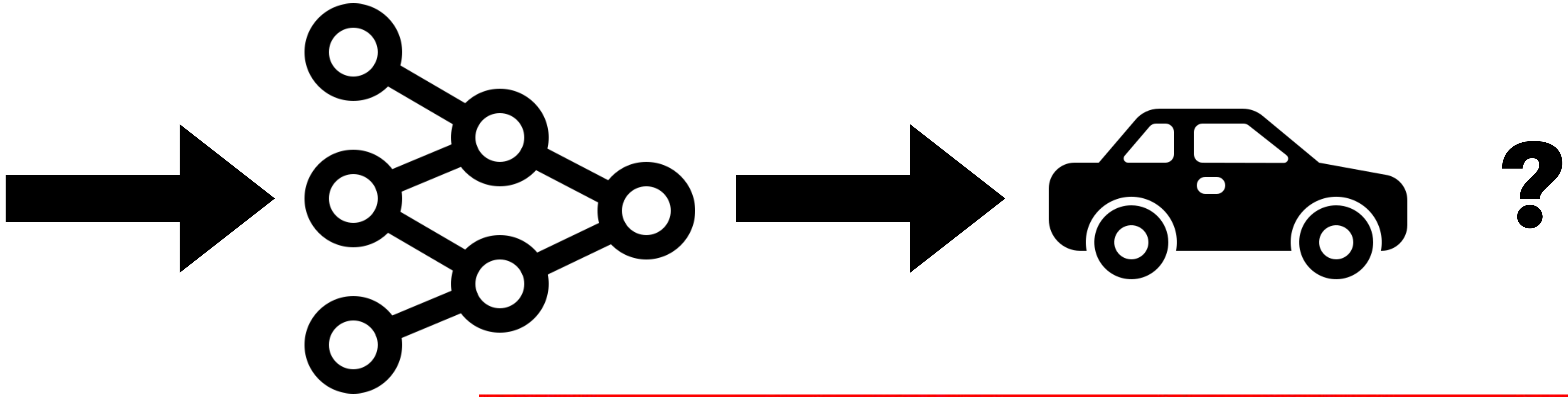


GIF obtained from KITTI-360 dataset (Liao et al., 2022)

However, each datapoint does not contribute equally



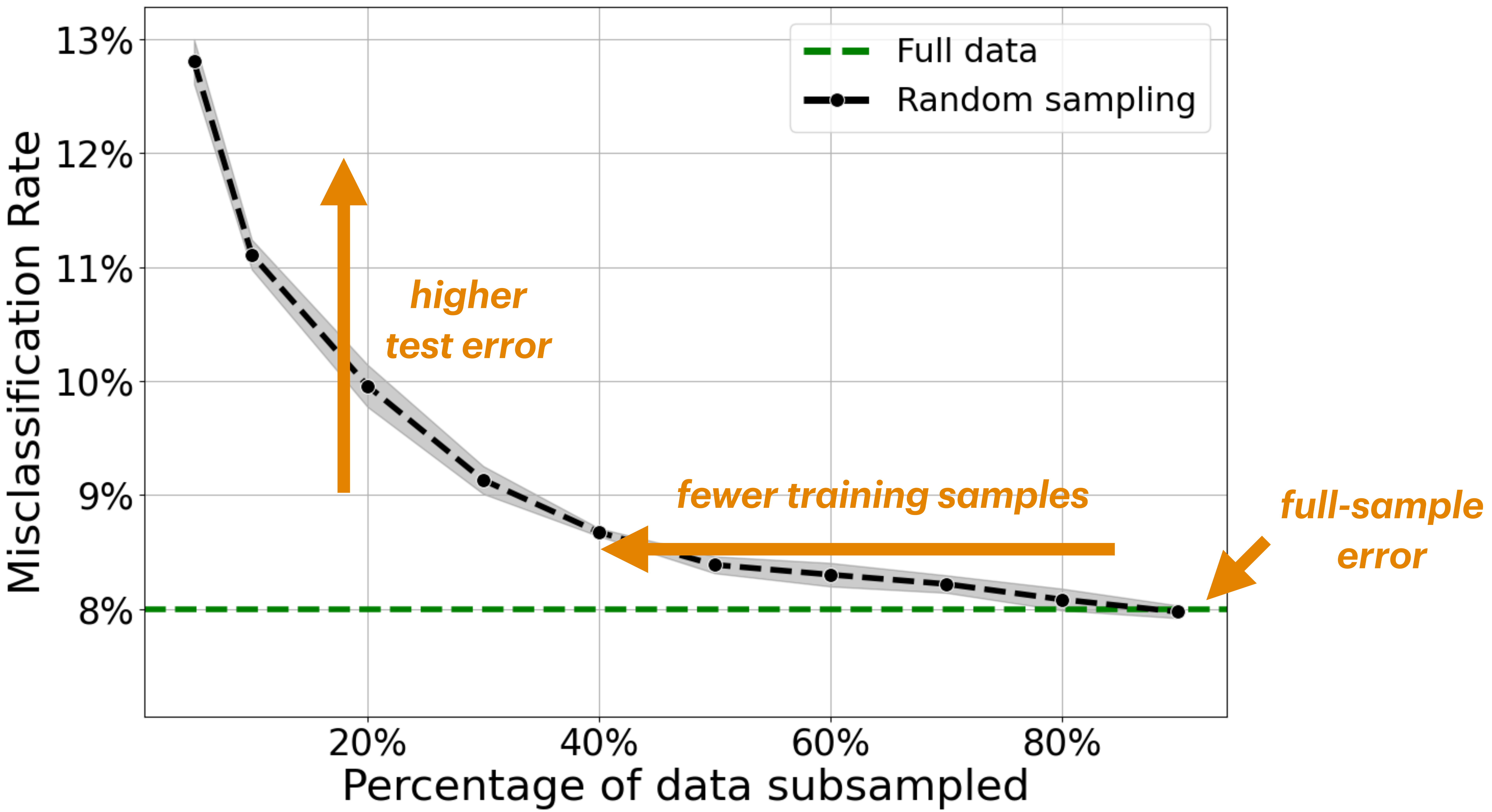
However, each datapoint does not contribute equally



Concordant with previous empirical results —

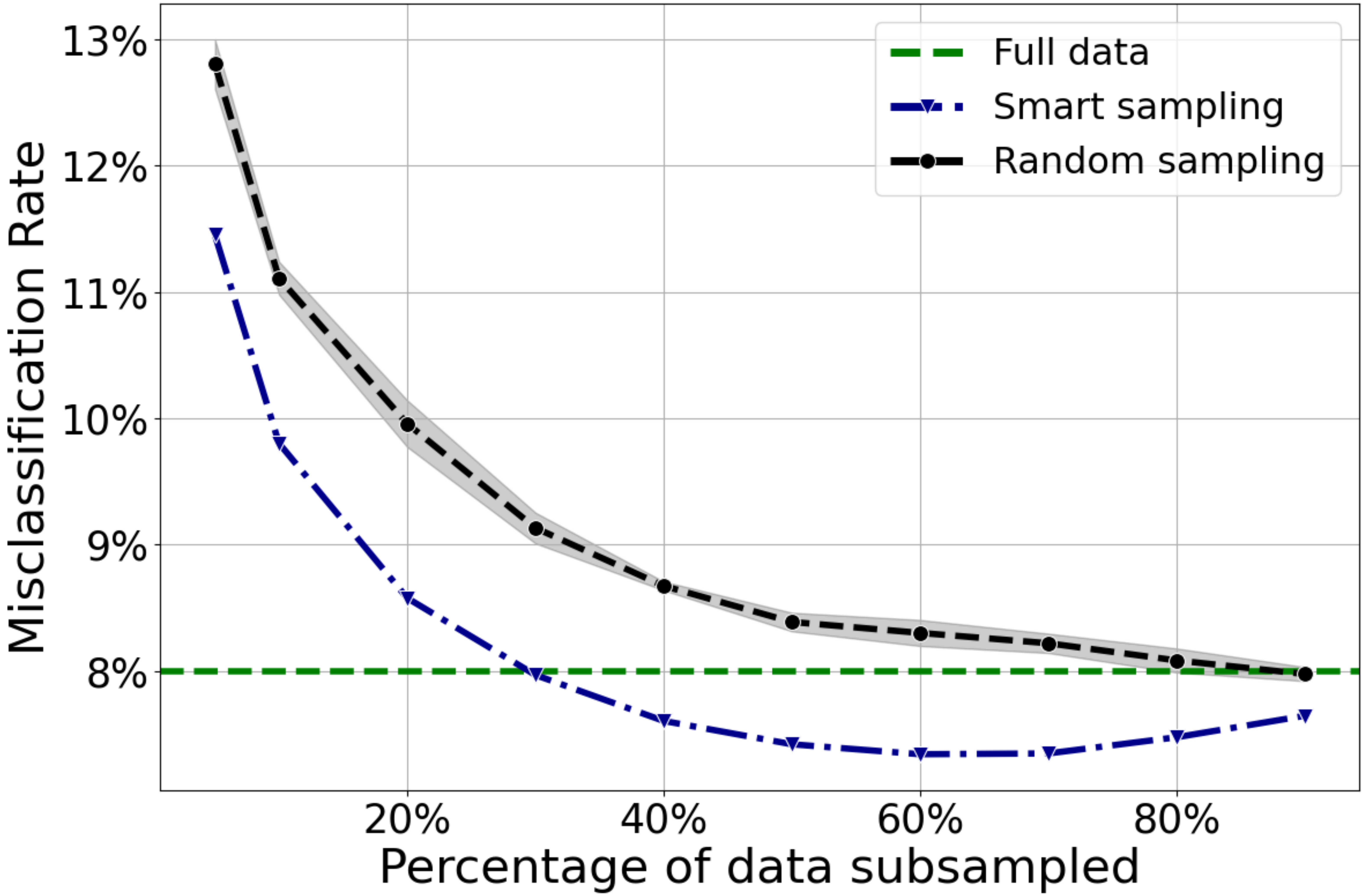
Nakkiran et al., 2021; Guo et al., 2022; Yang et al., 2022; Sorscher et al., 2022; Gadre et al., 2024, ...

However, each datapoint does not contribute equally *even in* **simple cases**



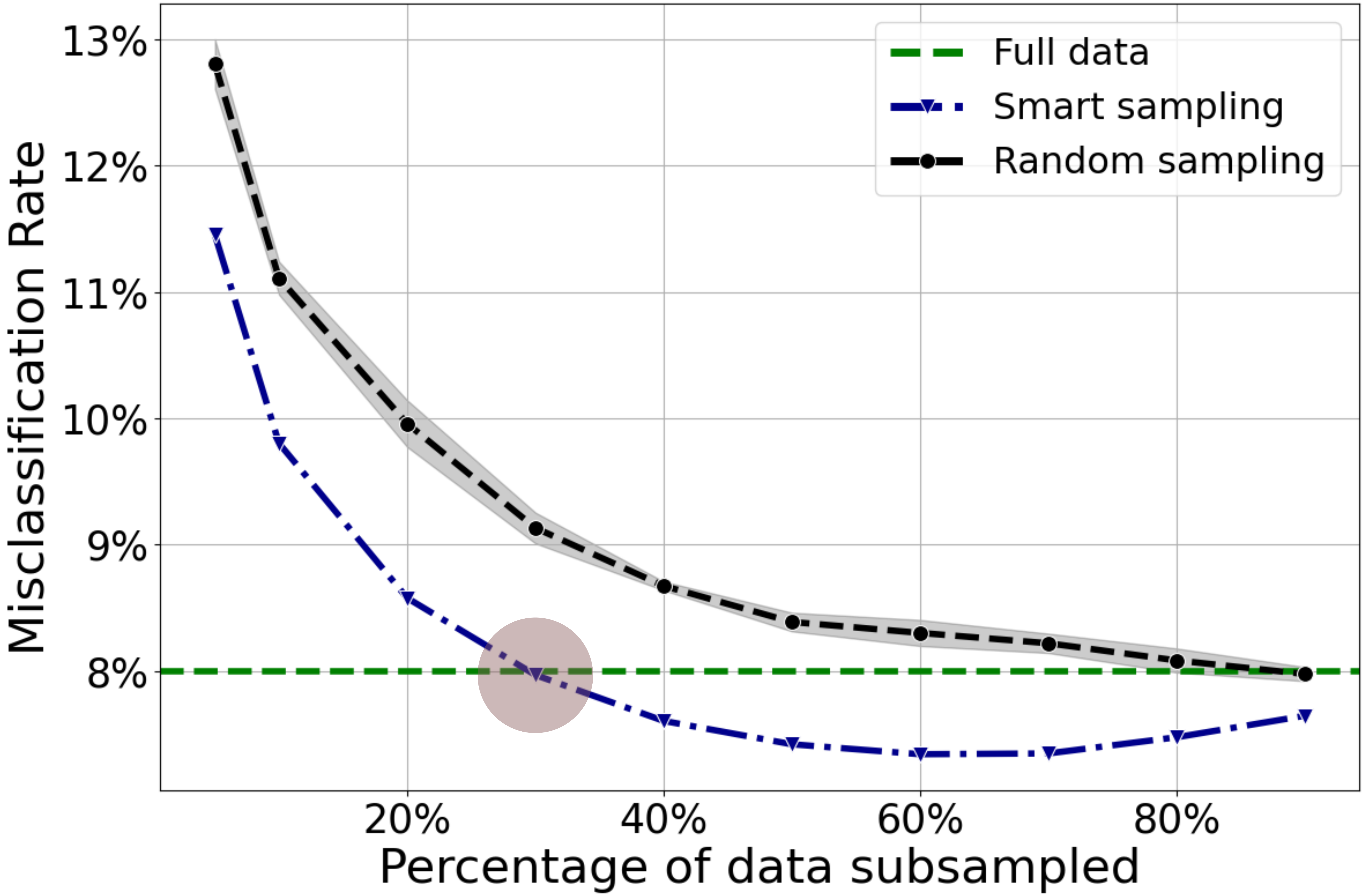
However, each datapoint does not contribute equally *even in* **simple cases**

“Smart” subsampling beats random.



However, each datapoint does not contribute equally *even in* **simple cases**

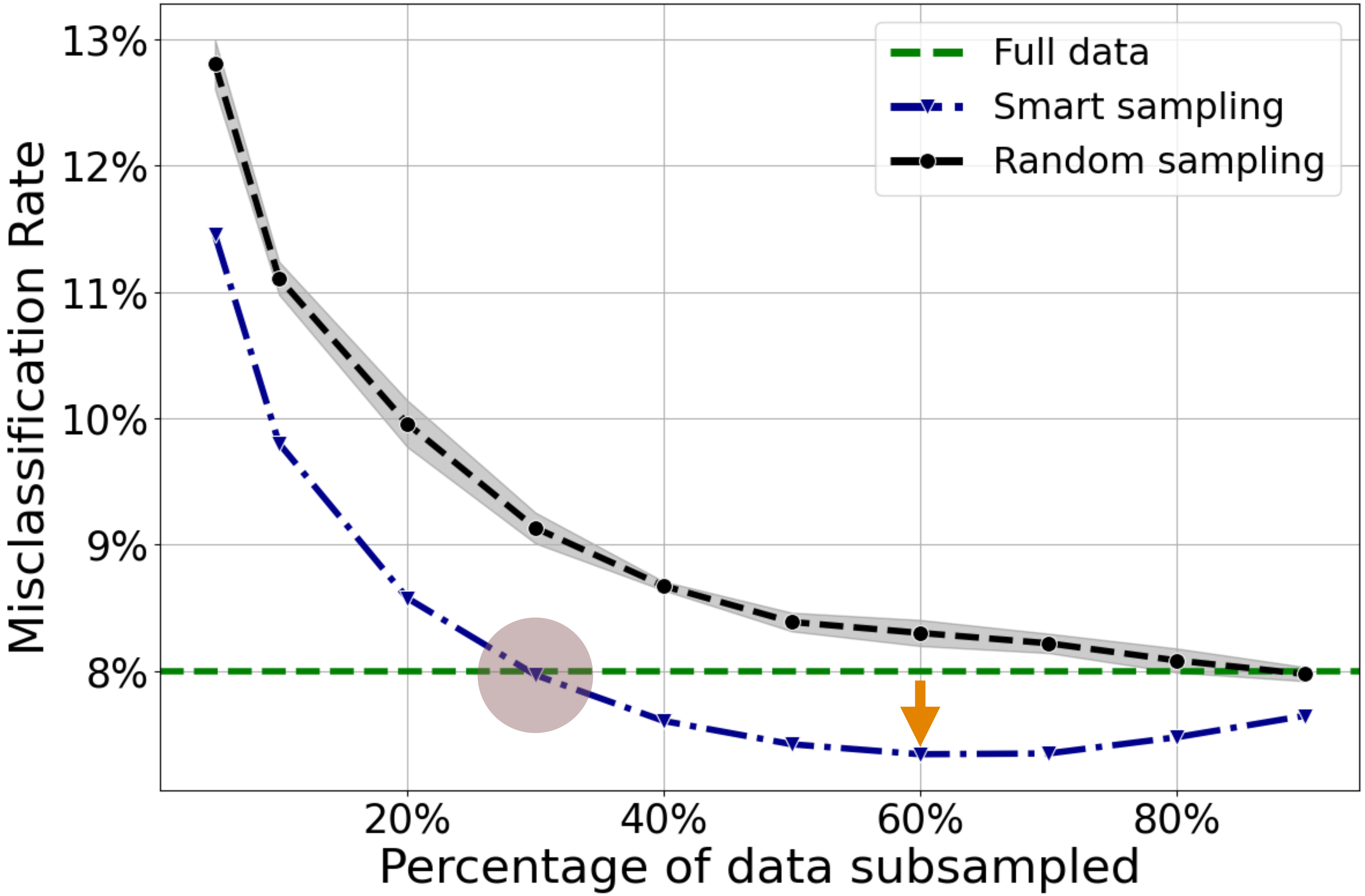
Full performance after throwing away 65% of the dataset



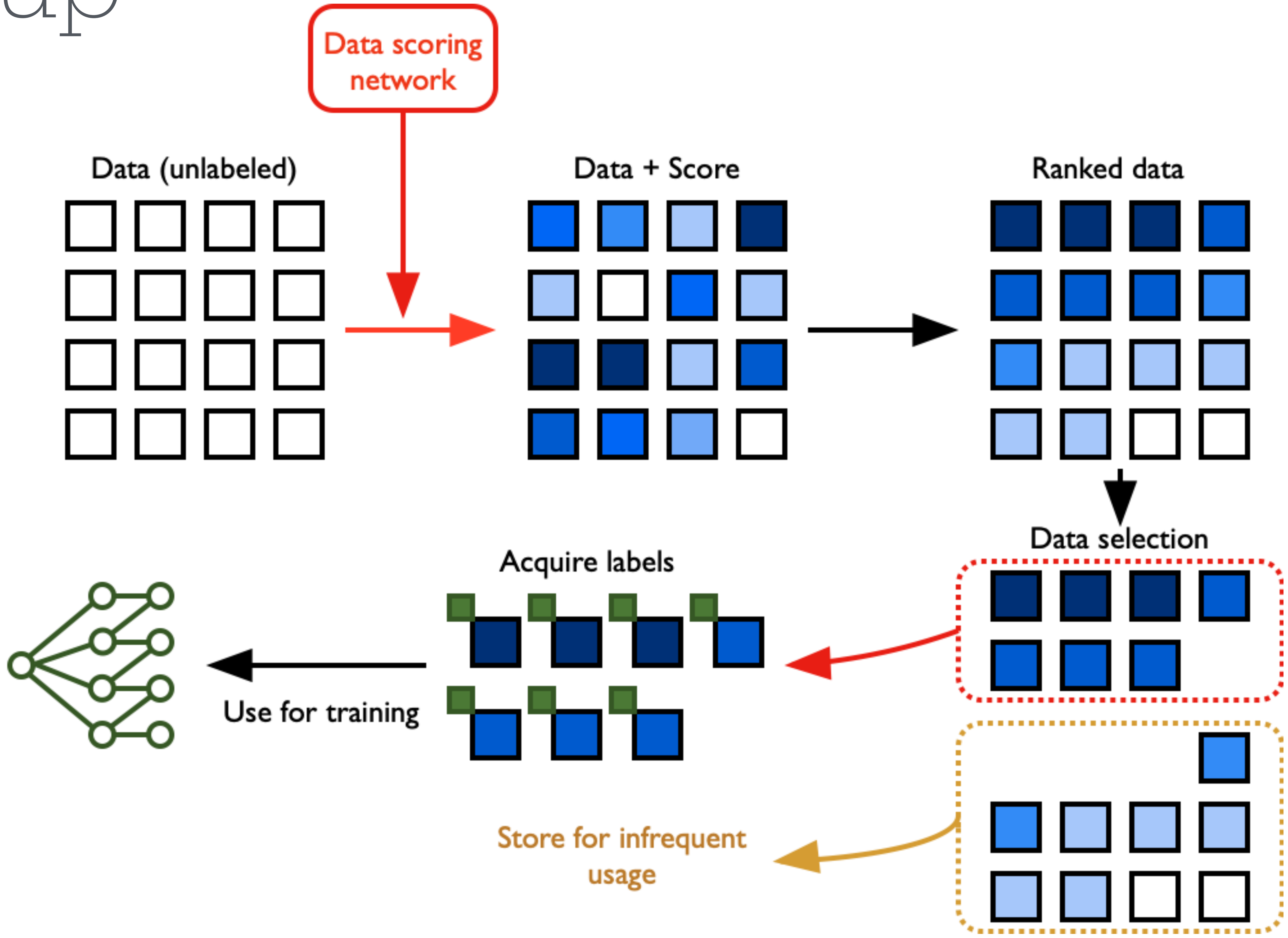
However, each datapoint does not contribute equally *even in* **simple cases**

Full performance after throwing away 65% of the dataset

Better performance with 60% of the data compared to full-sample



(informal) Setup



(informal) Setup

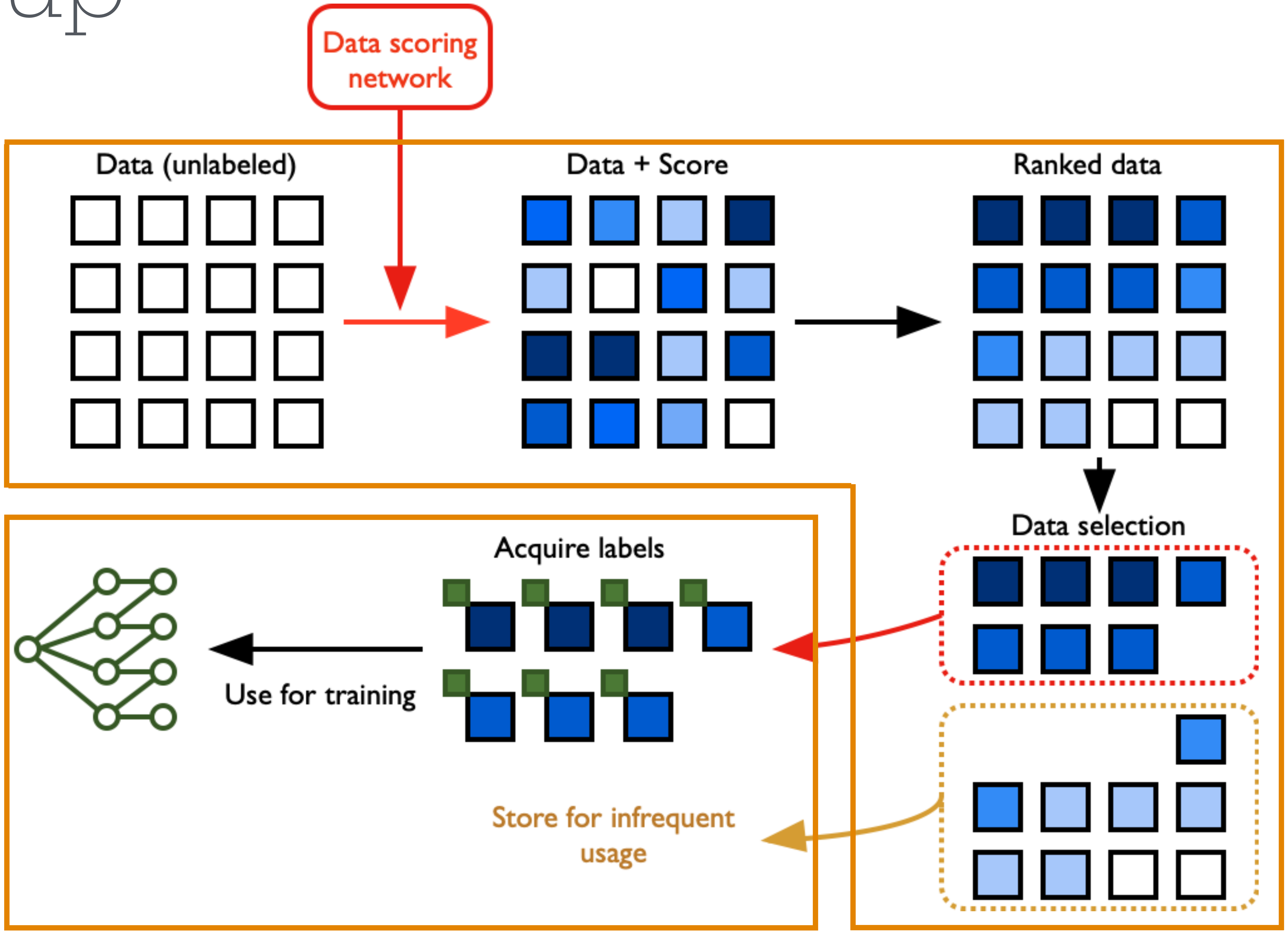
Main features

Two-step procedure: selection followed by training

Weakly Supervised — no access to data labels during selection but access to a "surrogate model"

Score-based subselection: "easy" or "hard" to classify

Sample or select points based on scores



(informal) Setup

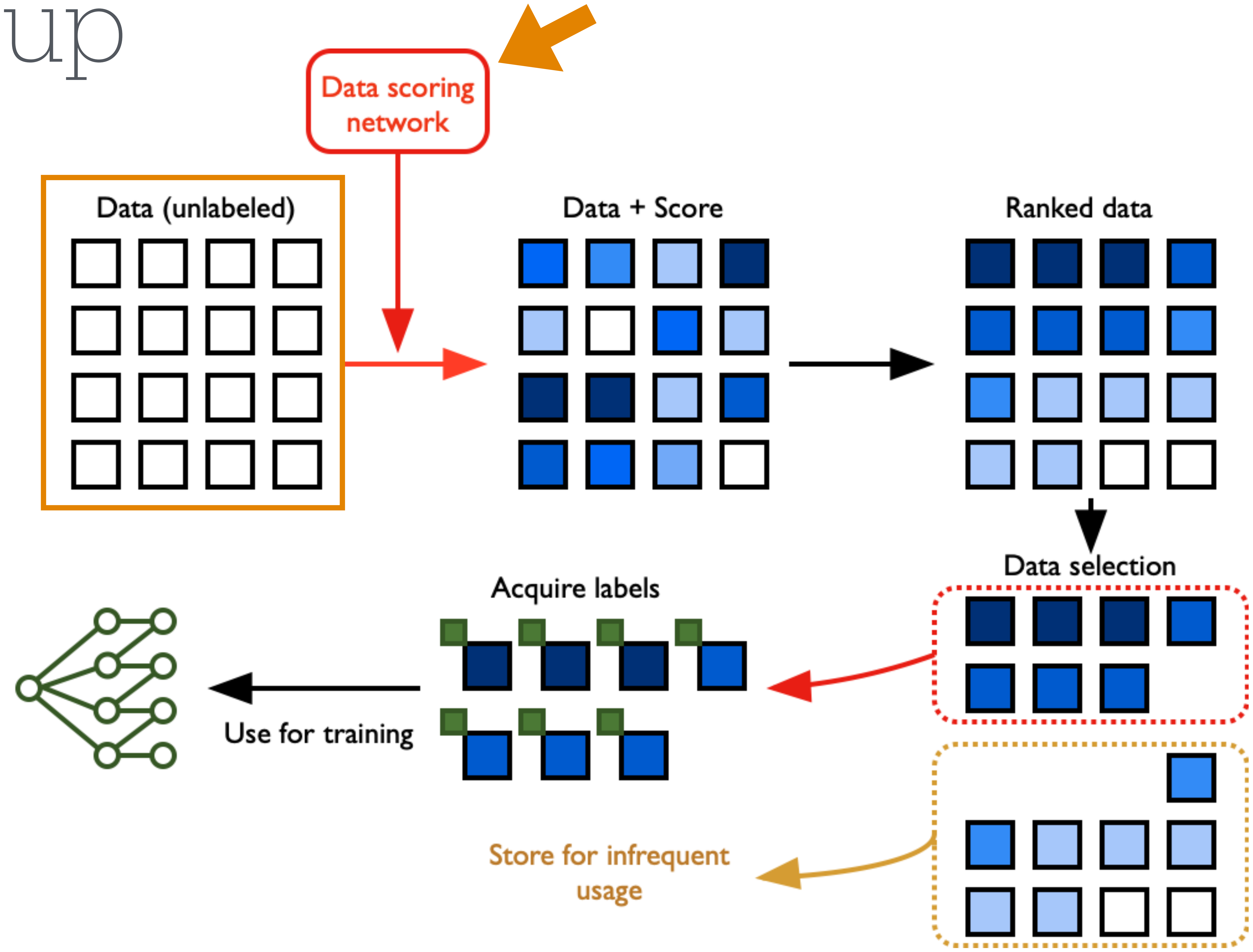
Main features

Two-step procedure:
selection followed by training

Weakly Supervised — no access to data labels during selection but access to a "surrogate model"

Score-based subselection:
"easy" or "hard" to classify

Sample or select points based on scores



(informal) Setup

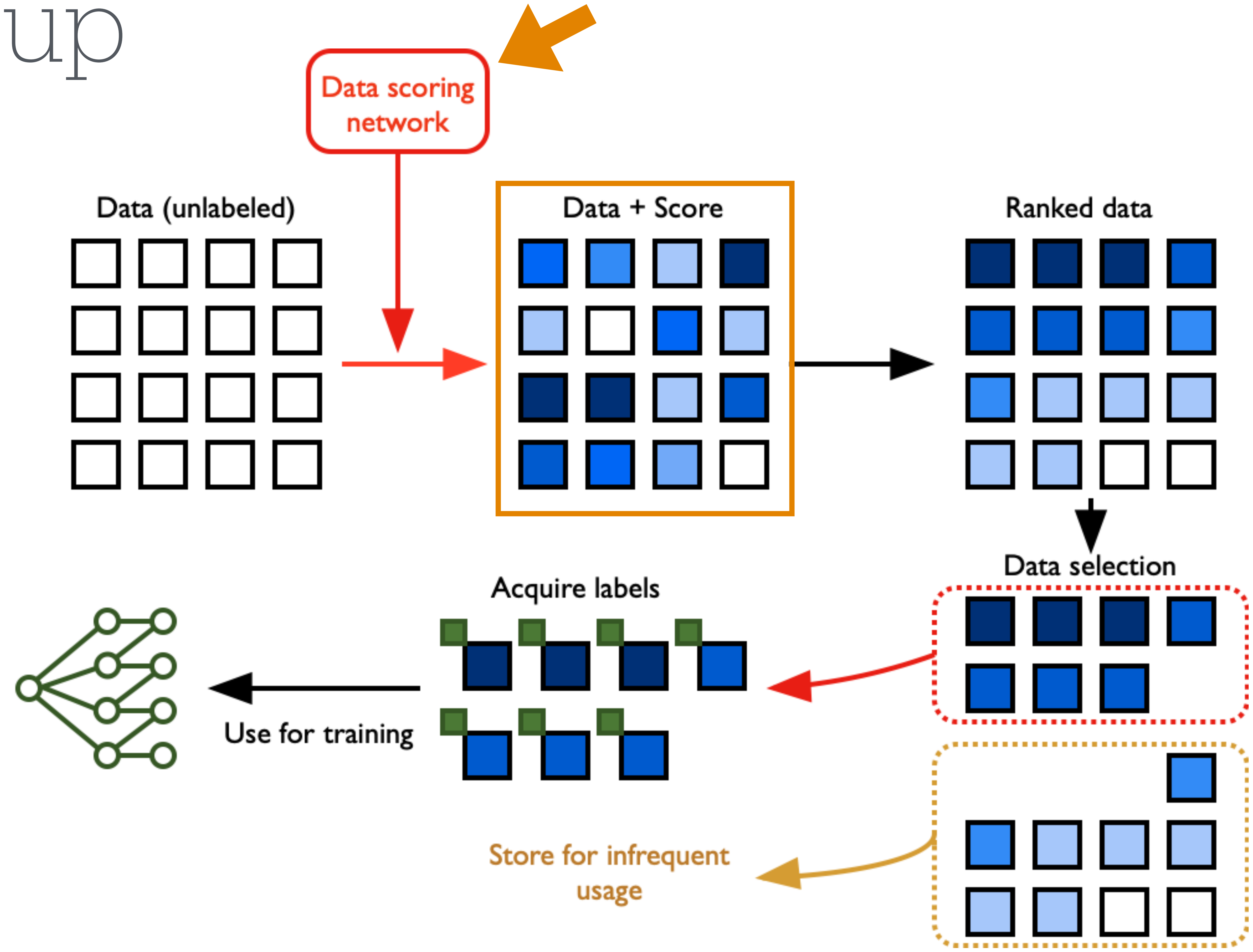
Main features

Two-step procedure:
selection followed by training

Weakly Supervised — no access to data labels during selection but access to a "surrogate model"

Score-based subselection: "easy" or "hard" to classify

Sample or select points based on scores



(informal) Setup

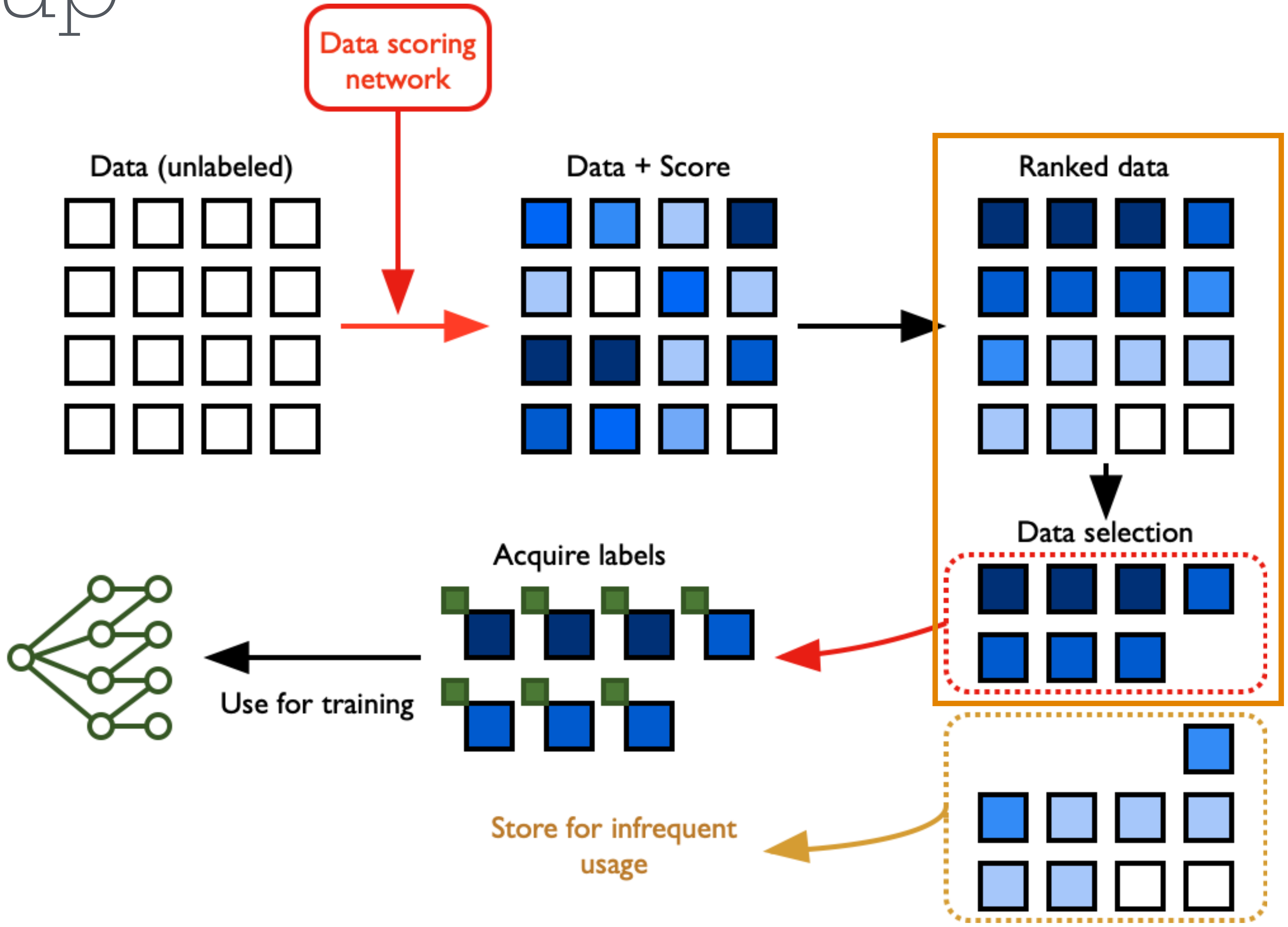
Main features

Two-step procedure:
selection followed by training

Weakly Supervised — no access to data labels during selection but access to a "surrogate model"

Score-based subselection: "easy" or "hard" to classify

Sample or select points based on scores



Formally

Weighted empirical risk minimization (ERM)

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \hat{R}_N(\boldsymbol{\theta})$$

$$\hat{R}_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} S_i(\boldsymbol{x}_i)\, \ell\left(y_i,\ f\left(\boldsymbol{x}_i\,; \boldsymbol{\theta}\right)\right) + \lambda\,\Omega(\boldsymbol{\theta})$$

Subsection scheme $S_i(\boldsymbol{x}_i)$ is defined by tuple (π_i, w_i)

$$\mathbb{P}(i \in G\,|\,\boldsymbol{X}, \boldsymbol{y}) = \pi(\boldsymbol{x}_i)\,, \qquad S_i(\boldsymbol{x}_i) = w(\boldsymbol{x}_i)\,\mathbf{1}_{i \in G}$$

(π_i, w_i) can depend on

(i) features $\boldsymbol{x}_i$

(ii) surrogate model $\mathsf{P}_{\mathsf{su}}(\,\cdot\,|\boldsymbol{x}_i)$

(iii) additional independent randomness.

1. Biased vs Unbiased subsampling

Unbiased loss function post subsampling:

$$w_i \propto 1/\pi_i$$

2. High vs Low-dim asymptotic

Proportional high-dimension asymptotics:

$$n, N, p \to \infty$$

$$n/N \to \gamma,\ \ N/p \to \delta_0$$

3. Imperfect vs Perfect Surrogates

Perfect Surrogate:

$$\mathsf{P}_{\mathsf{su}}(\,\cdot\,|\boldsymbol{x}_i) = \mathbb{P}(\,\cdot\,|\boldsymbol{x}_i)$$

Setup: numerical results

Binary logistic regression

Subselection Scheme
$\pi(\boldsymbol{x}_i) \propto \left(p_{su} \times (1 - p_{su})\right)^{\alpha}$

Setup: numerical results

Binary logistic regression

Subselection Scheme

$$\pi(\boldsymbol{x}_i) \propto \left(p_{su} \times (1 - p_{su})\right)^{\alpha}$$

Probability under surrogate model

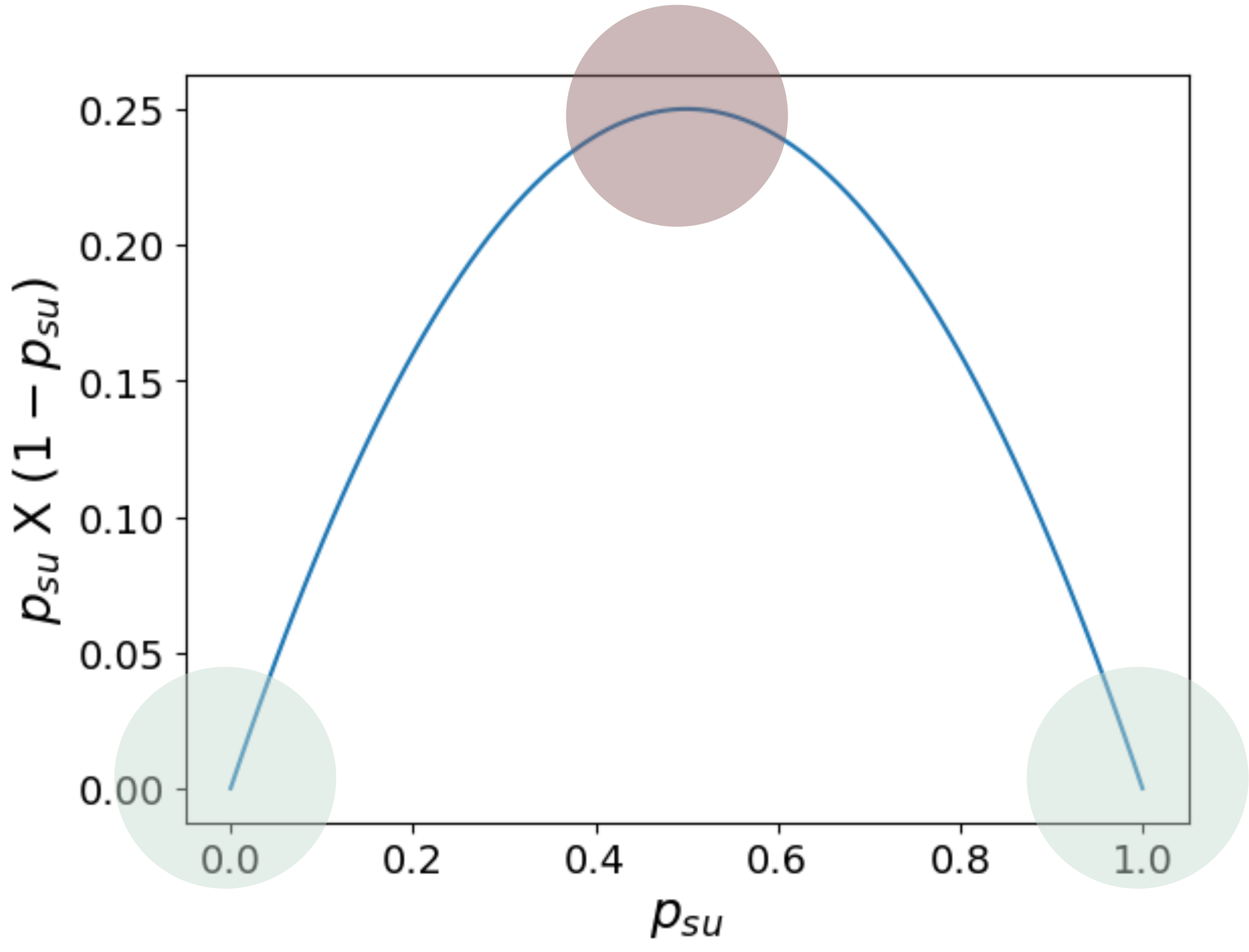
Setup: numerical results

Binary logistic regression

Subselection Scheme

$$\pi(\boldsymbol{x}_i) \propto \left(p_{su} \times (1 - p_{su})\right)^{\alpha}$$

Hardness score



"hard" examples under surrogate model

"***easy***" examples under surrogate model

Setup: numerical results

Binary logistic regression

Subselection Scheme

$$\pi(\boldsymbol{x}_i) \propto \left(p_{su} \times (1 - p_{su})\right)^{\alpha}$$

α determines hardness:

$\alpha > 0$ upsample hard points

Setup: numerical results

Binary logistic regression

Subselection Scheme

$$\pi(\boldsymbol{x}_i) \propto \left(p_{su} \times (1 - p_{su})\right)^{\alpha}$$

Synthetic Data

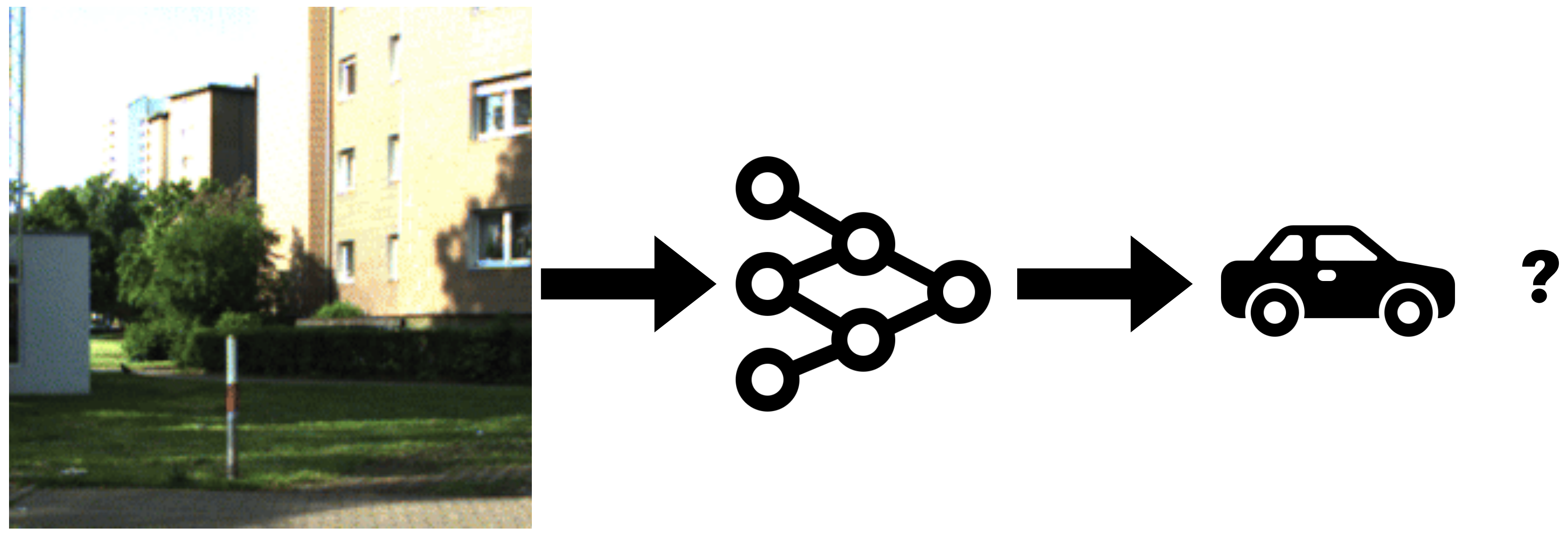
Isotropic Gaussian Covariates:

$\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{I}_p)$

GLM (well- or mis-specified):

$\mathbb{P}(y_i = +1 \,|\, \boldsymbol{x}_i) = f\left(\langle \boldsymbol{\theta}_0 , \boldsymbol{x}_i \rangle\right)$

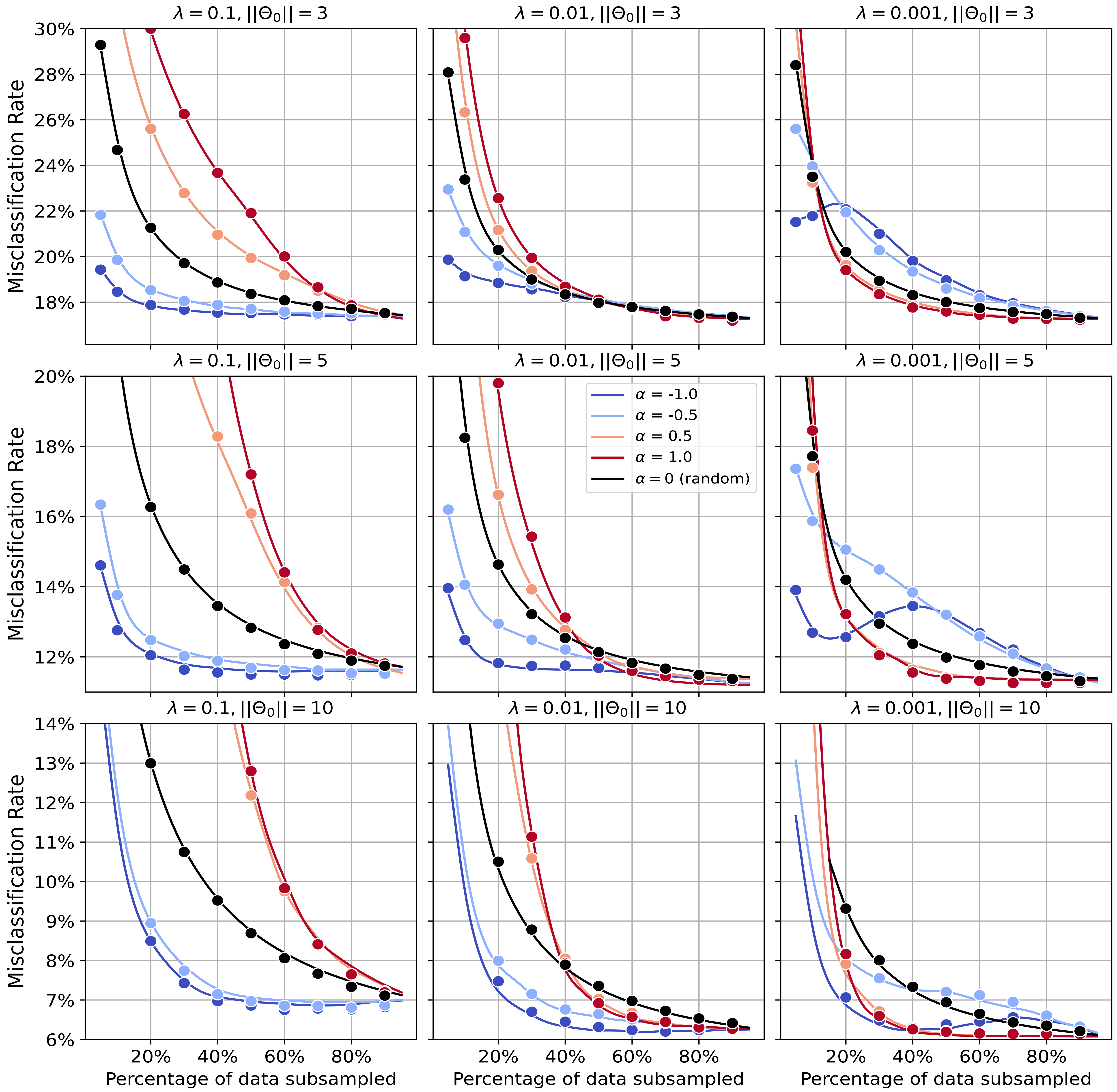
Real Data: AV dataset



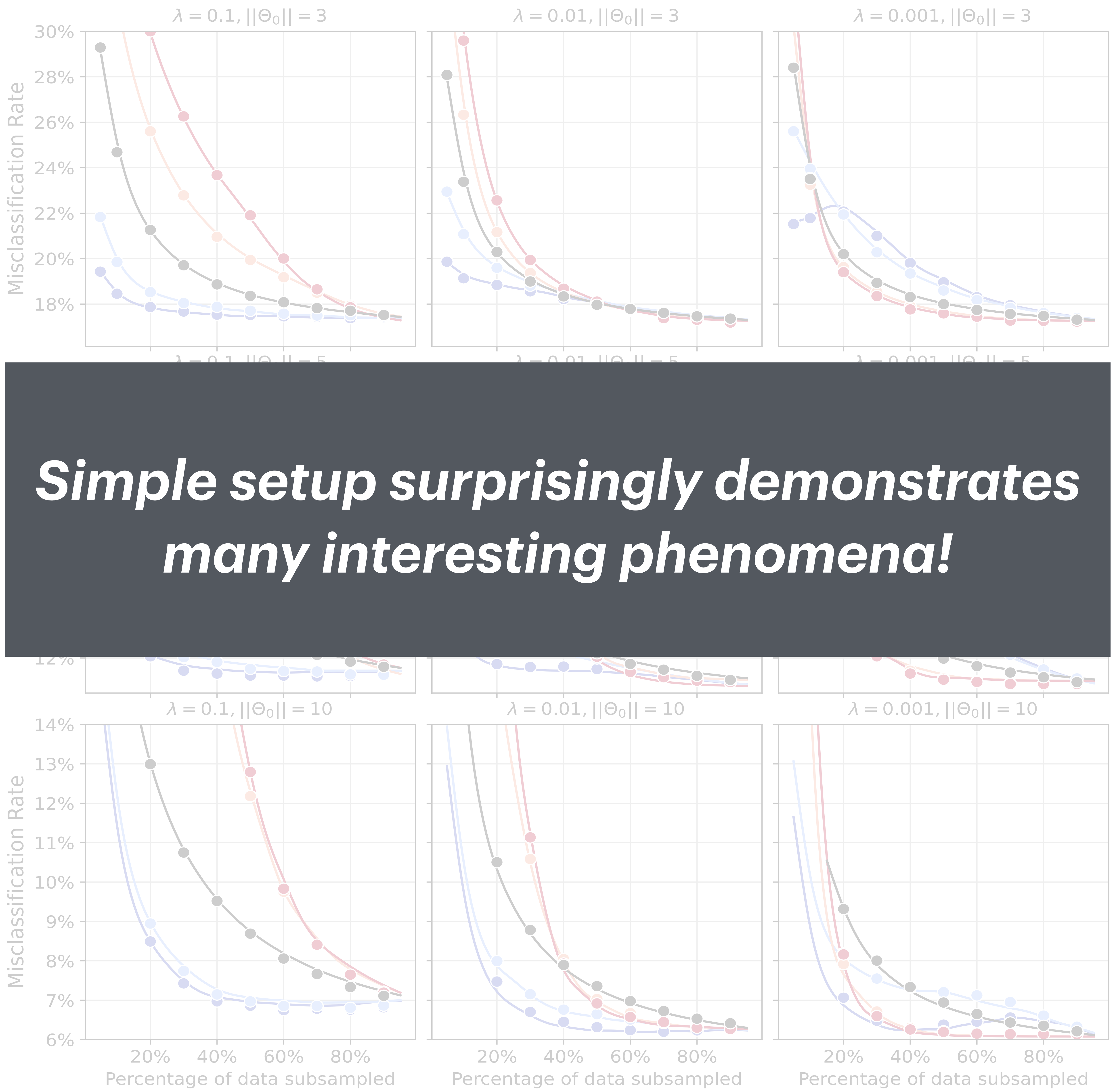
Theory predicts “exact” high-dim asymptotic test-error

Synthetic data

Circles: Simulations
Continuous lines: Theory



Theory predicts “exact” high-dim asymptotic test-error



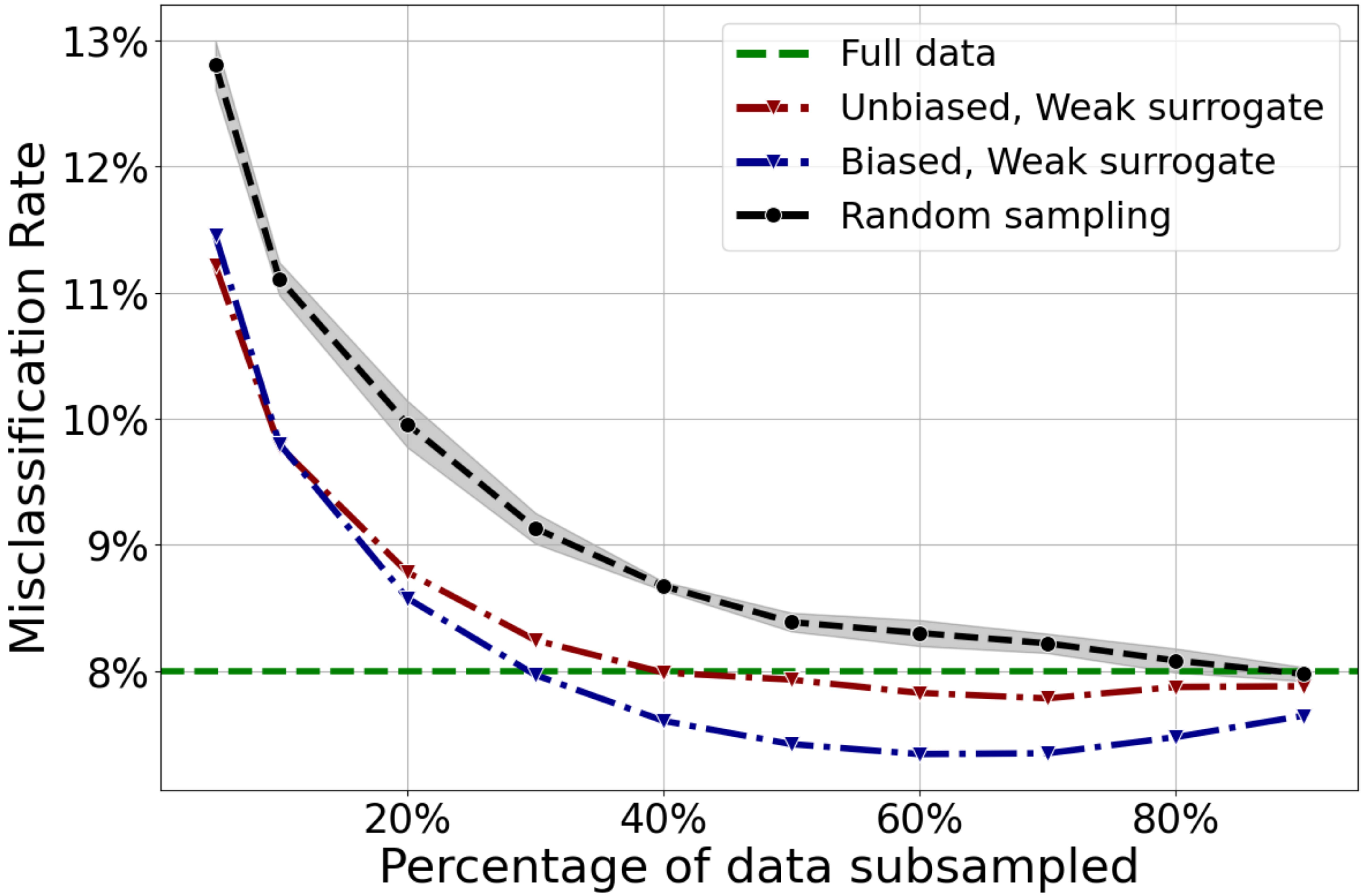
Simple setup surprisingly demonstrates many interesting phenomena!

1. Unbiased subsampling can be suboptimal

Real data:
AV dataset

Proposition

Under certain natural settings we have multiple theorems and specific constructions showing unbiased subsampling can be arbitrary worse.



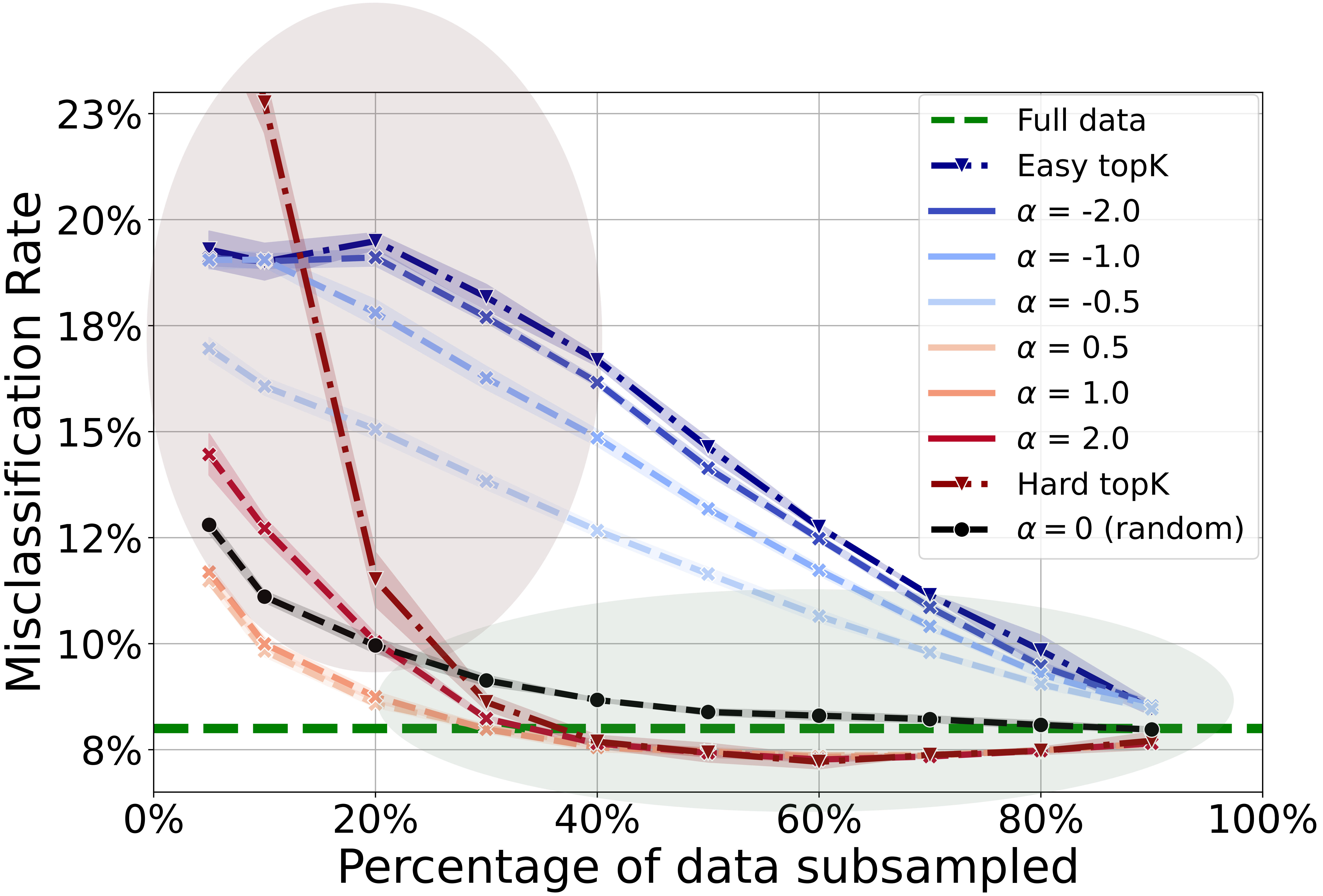
2. Choose "*hard*" but not the "*hardest*"

Real data:
AV dataset

Observation

Choosing ***"hard"*** examples work for this setup however,

picking ***"hardest"*** examples can lead to catastrophic failures!



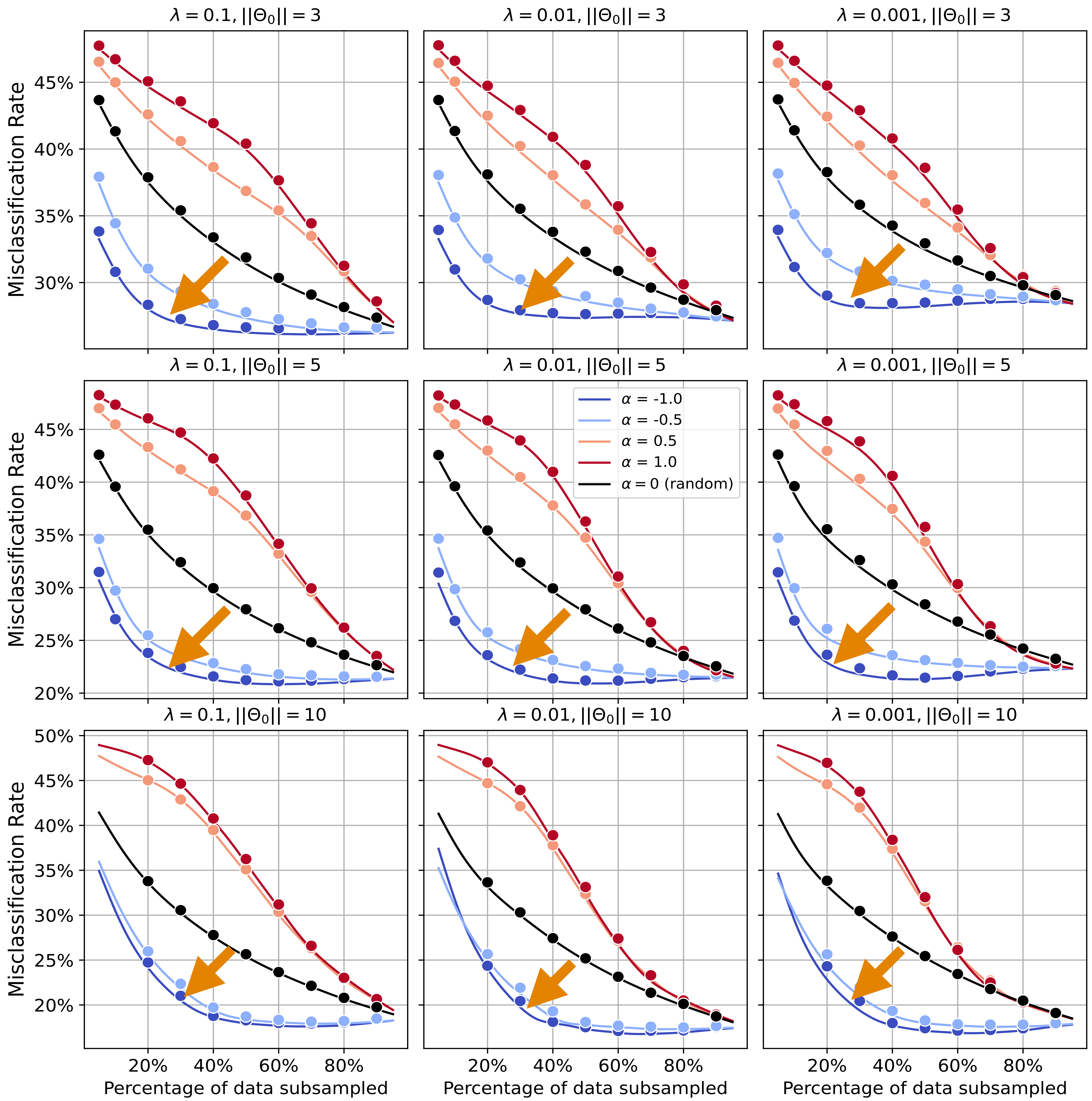
3. In high-dim settings choosing "*easy*" is better

Synthetic data

Observation

Blue curve (negative alpha), i.e. ***upsampling easy examples***, performs best for all settings (across regularizations and SNRs) in over-parameterization regime*

*corroborates Sorscher et al., 2022



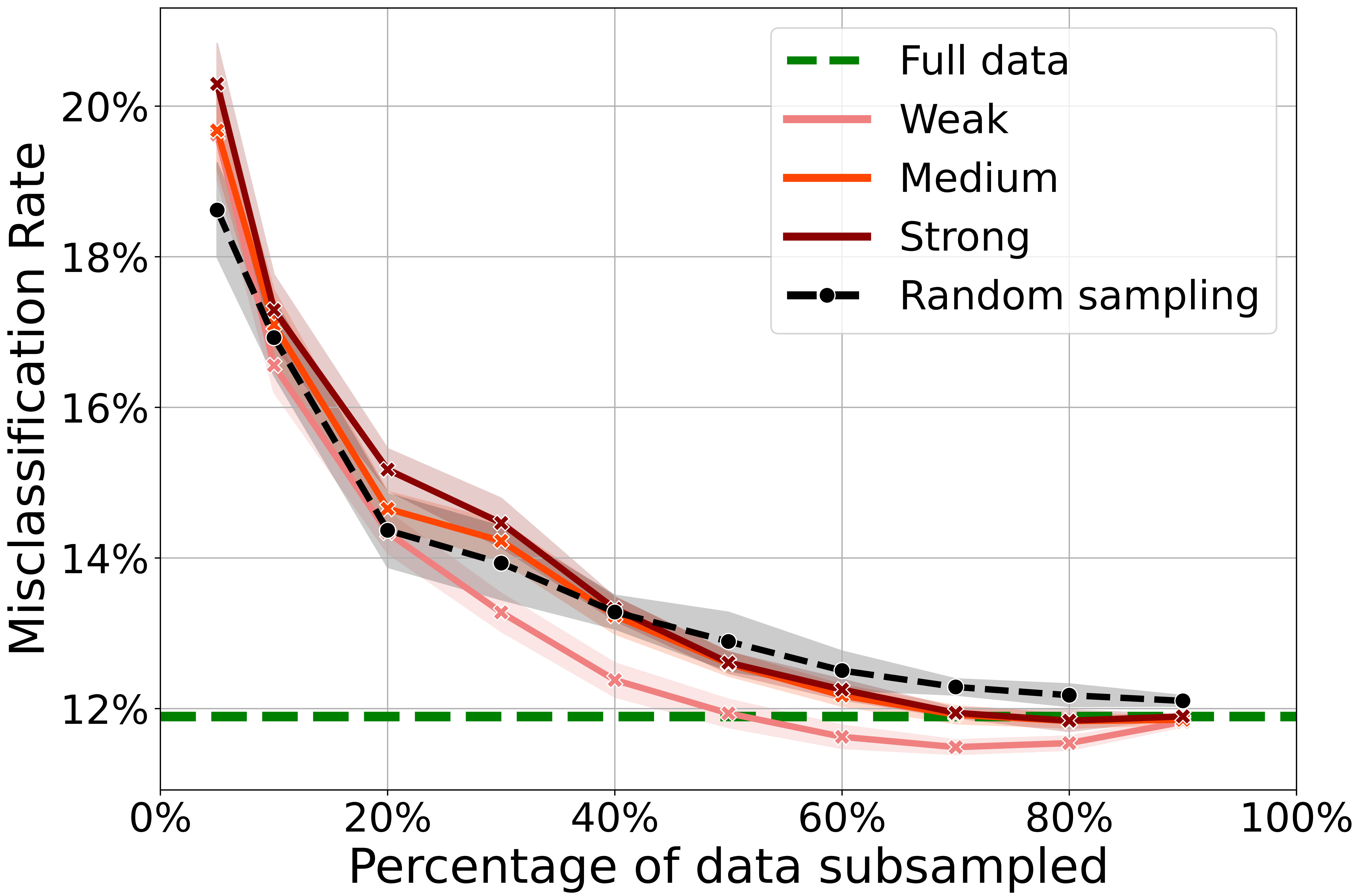
4. Better surrogate models != better selection

Real data:
AV dataset

Observation

"Weak" supervision, i.e. surrogate models trained on far-fewer independent samples, is sufficient for effective data selection.

In-fact, "stronger" surrogate models can hurt!



$N_{su}/N = 4\,\%\,,\,21\,\%\,,\,43\,\%$

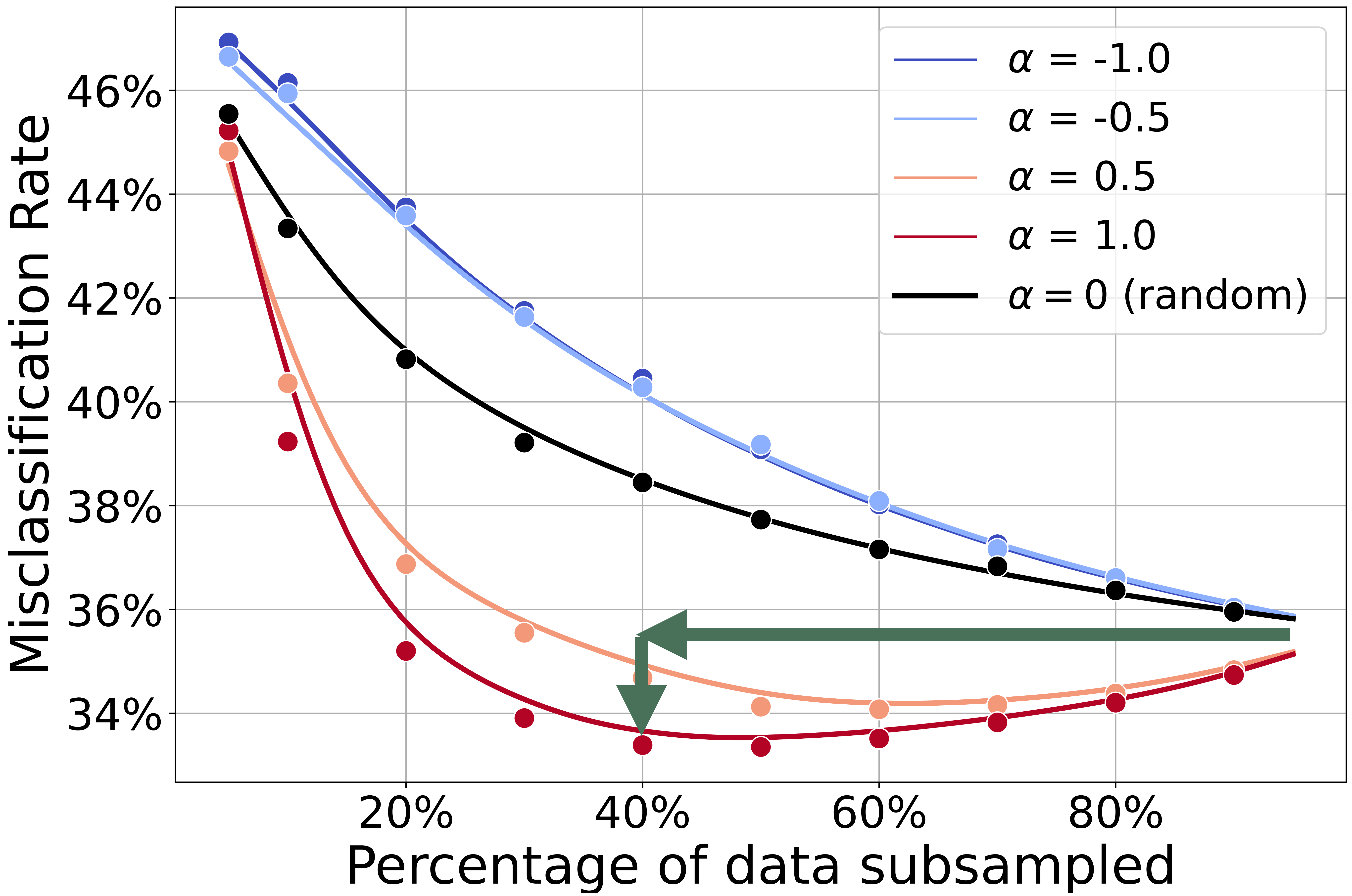
5. Subsampling can beat full-sample training

Synthetic data

Intuition

Observed in case of mis-specified models (true data does not follow logistic distribution).

Not all data samples provide new information when machine learning models and losses are mismatched!



Conclusions

Surprises

Popular techniques using “unbiased” subsampling can be suboptimal

Use of “weaker” surrogate models can outperform stronger surrogate models

Main Insight

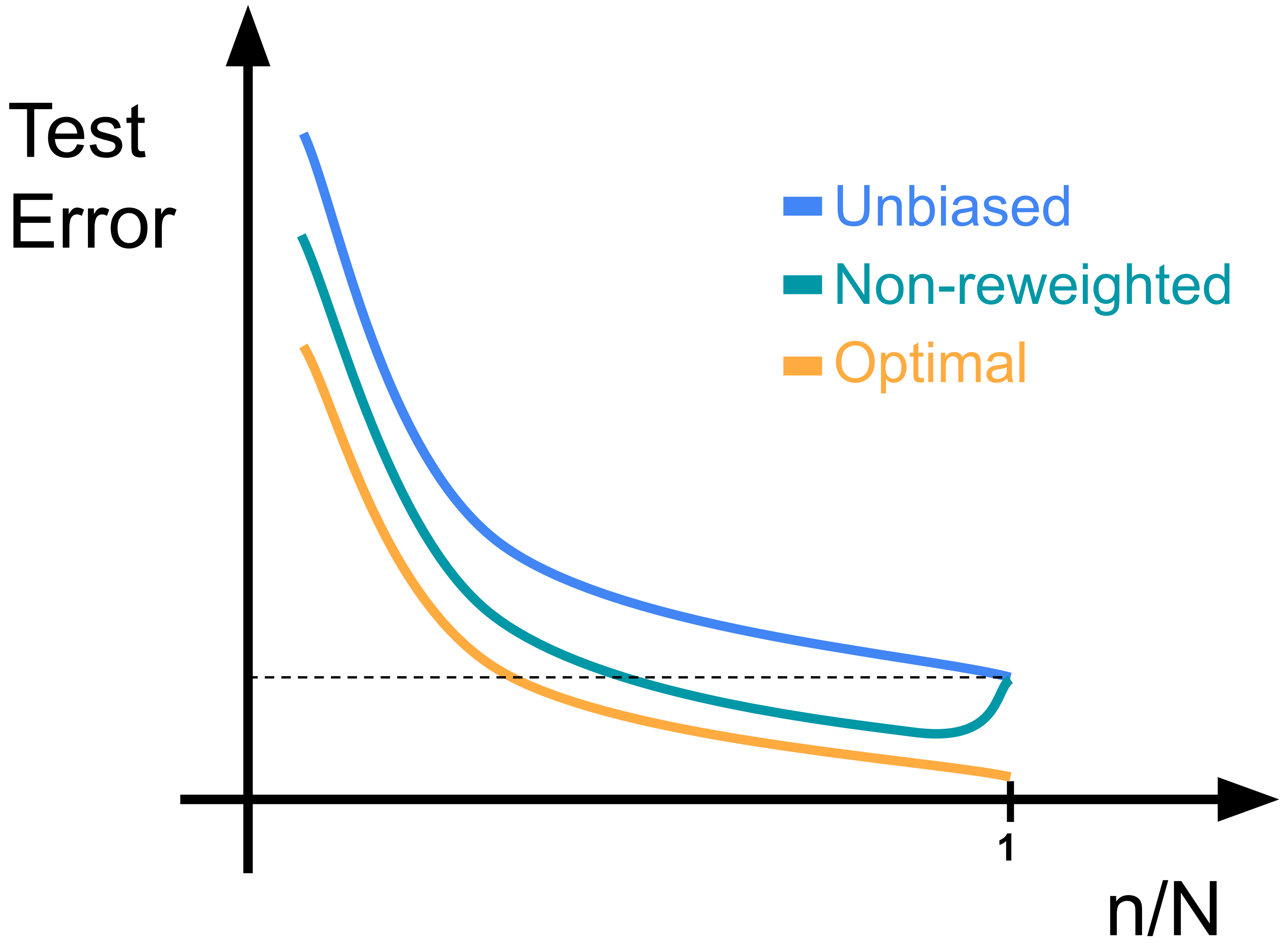
Uncertainty based subsampling can be effective though

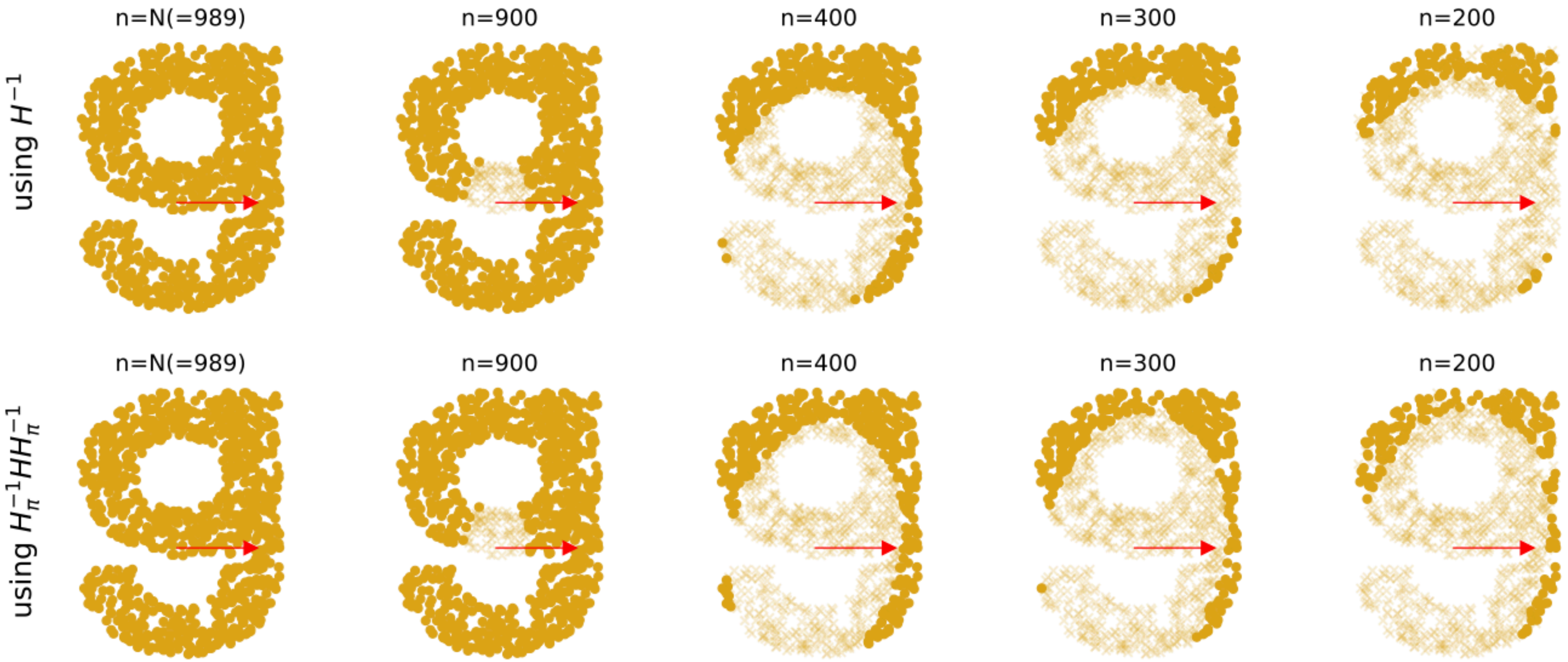
choosing “hardest” examples can be catastrophic

depending on setting such as parameterization ratio, regularization, mis-specification;
“easy” examples can be more beneficial than hard examples*

Don’t stir the pile,
be selective about it!

Questions?!


Test Error
Unbiased
Non-reweighted
Optimal
1
n/N


using H^{-1}
using $H_\pi^{-1} H H_\pi^{-1}$
n=N(=989)
n=900
n=400
n=300
n=200