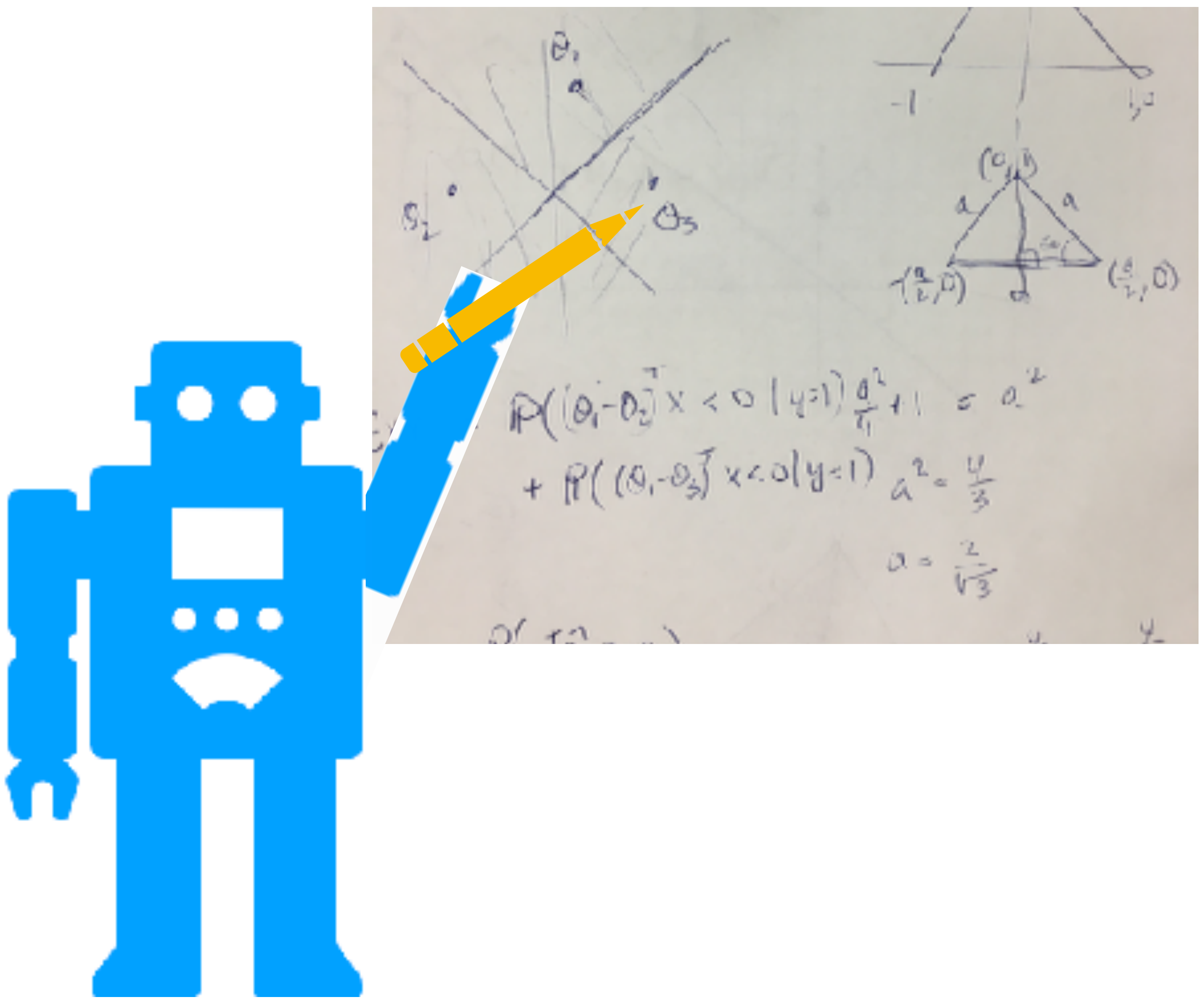
Active Learning from Theory to Practice



Steve Hanneke
Toyota Technological Institute at Chicago
steve.hanneke@gmail.com

Robert Nowak
UW-Madison
rdnowak@wisc.edu

ICML | 2019

Thirty-sixth International Conference on Machine Learning

Tutorial Outline

Part 1: Introduction to Active Learning (Rob)

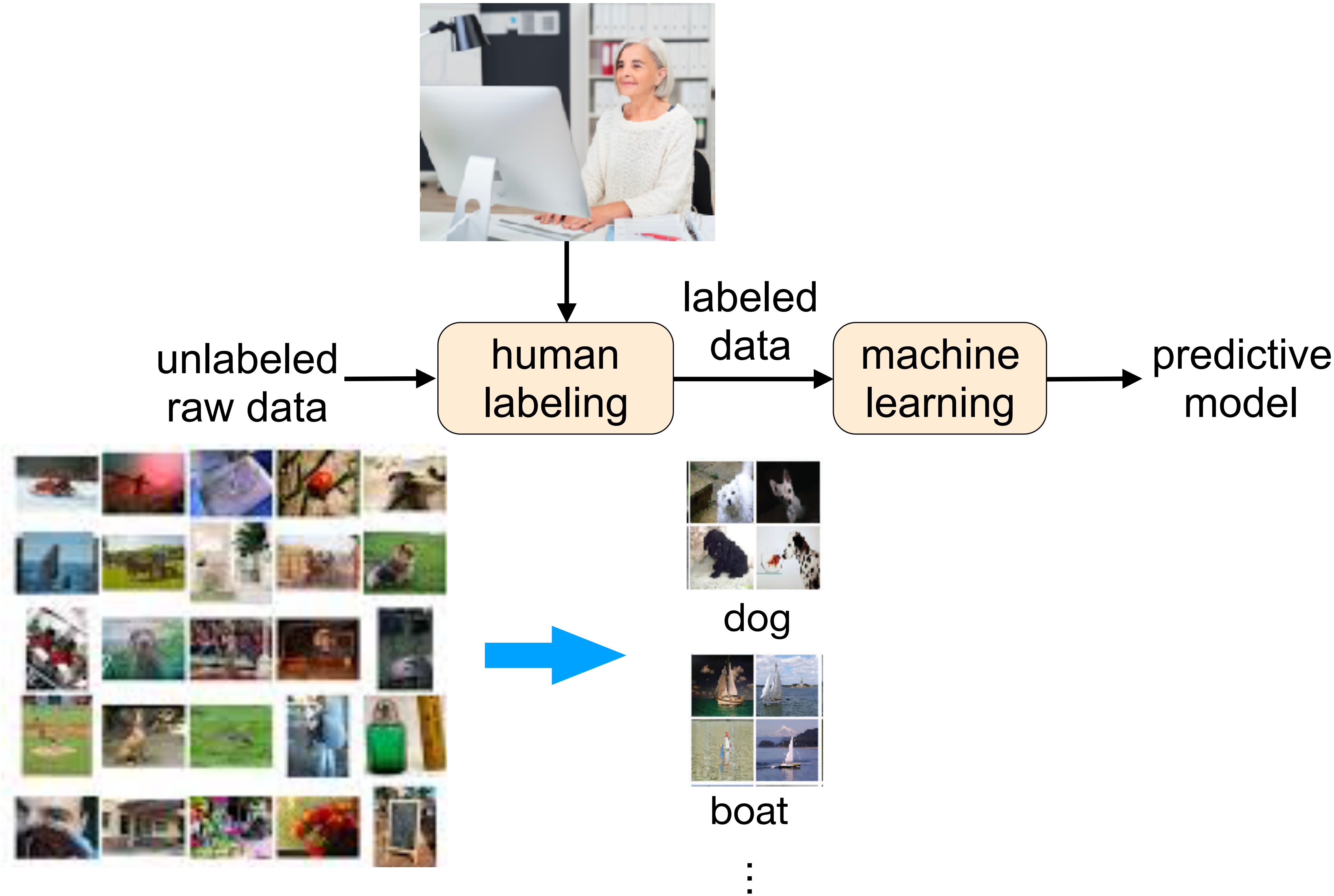
Part 2: Theory of Active Learning (Steve)

Part 3: Advanced Topics and Open Problems (Steve)

Part 4: Nonparametric Active Learning (Rob)

slides: http://nowak.ece.wisc.edu/ActiveML.html

Conventional (Passive) Machine Learning



ALL SYSTEMS GO ?

theguardian

Computers now better than humans at recognising and sorting images

millions of labeled images
1000's of human hours

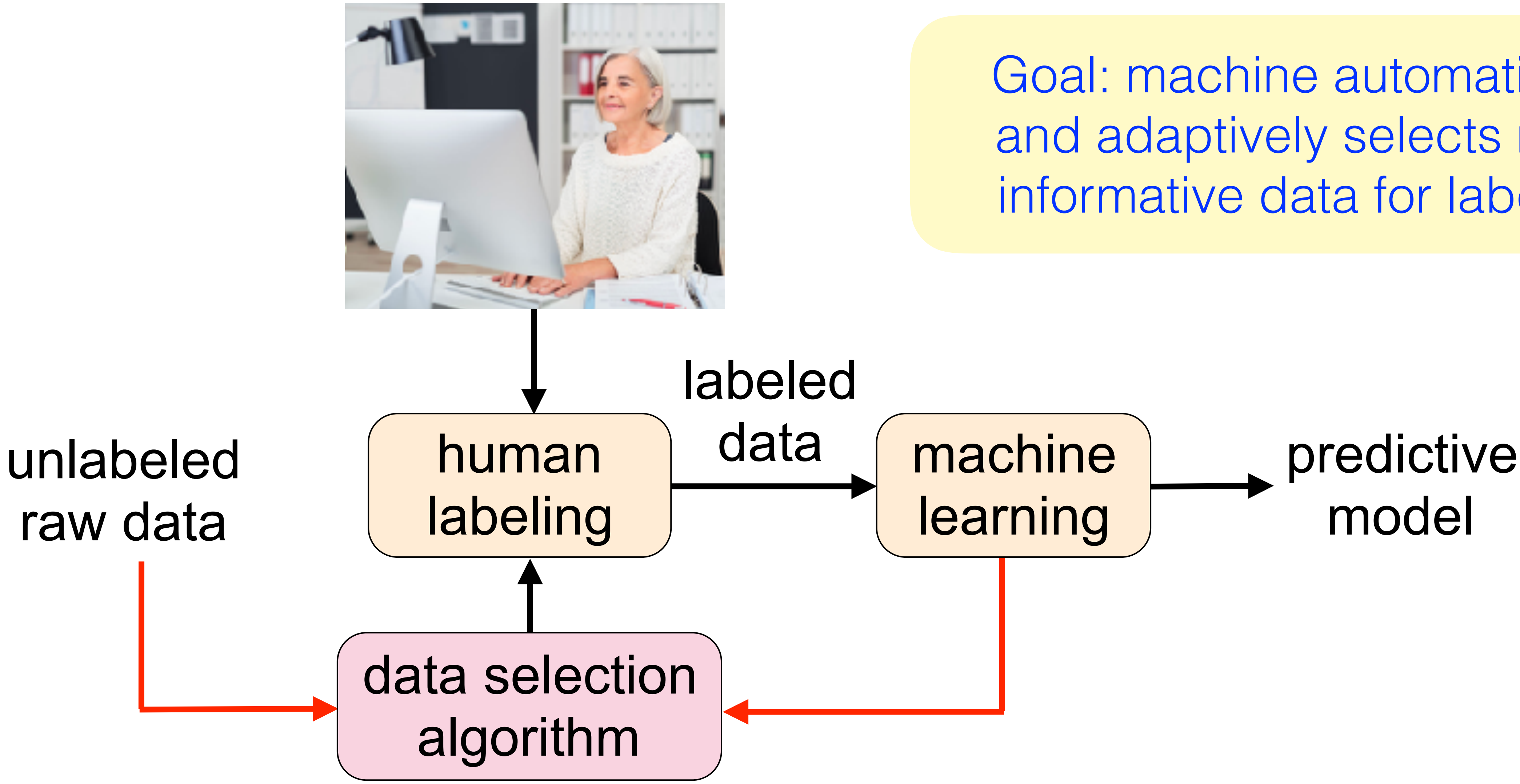
QUARTZ

Google says its new AI-powered translation tool scores nearly identically to human translators

trained on more texts than a human could read in a lifetime

Can we train machines with less labeled data and less human supervision?

Active Machine Learning



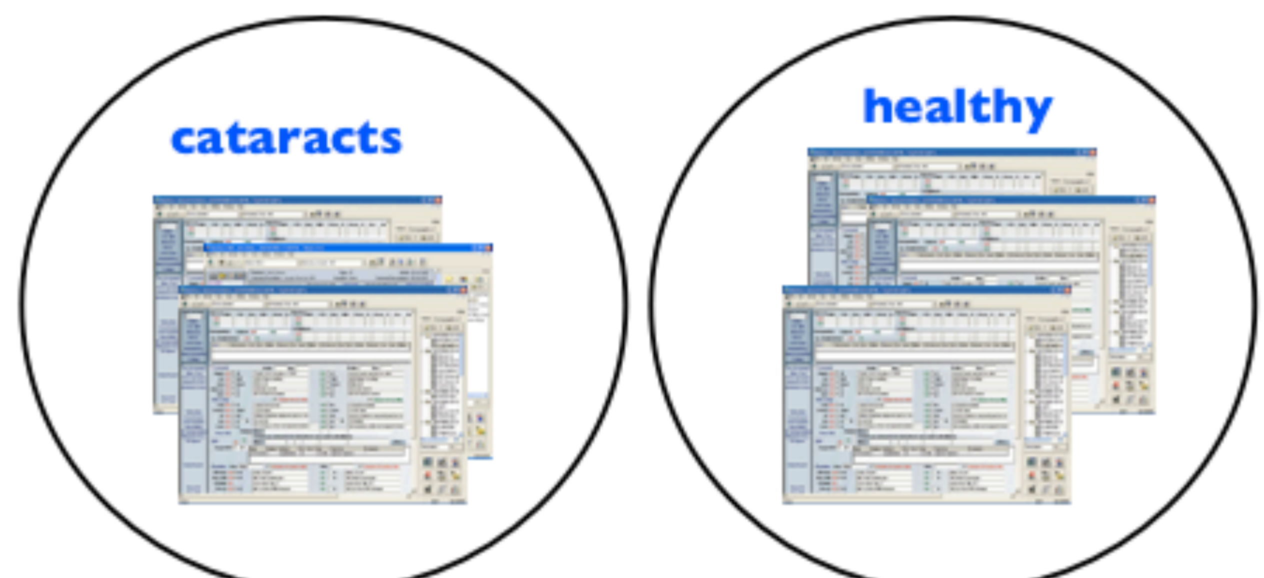
Goal: machine automatically and adaptively selects most informative data for labeling

Motivating Application

unlabeled electronic
health records (EHRs)

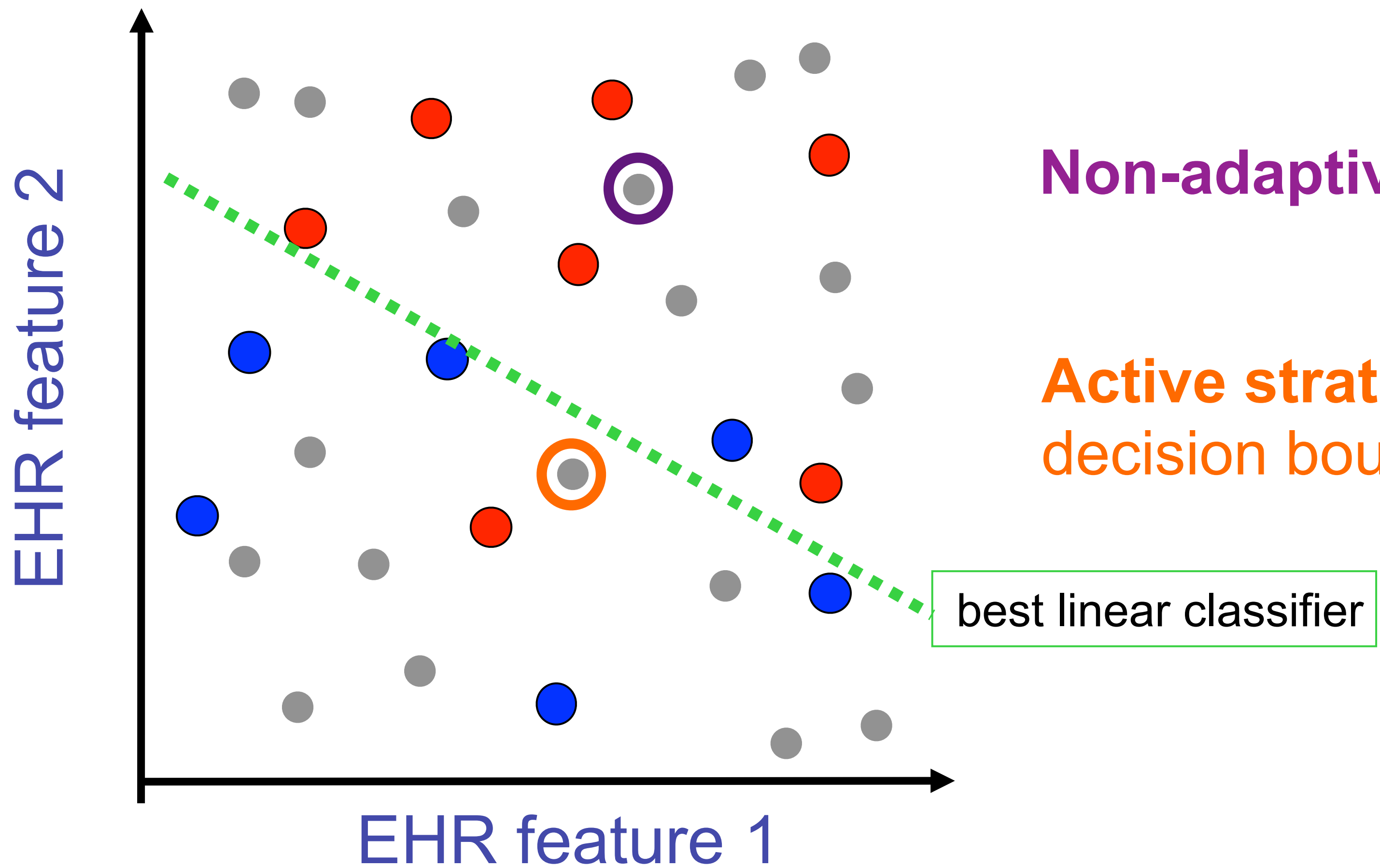
prediction rule
that can be applied
to unlabeled EHRs

human experts



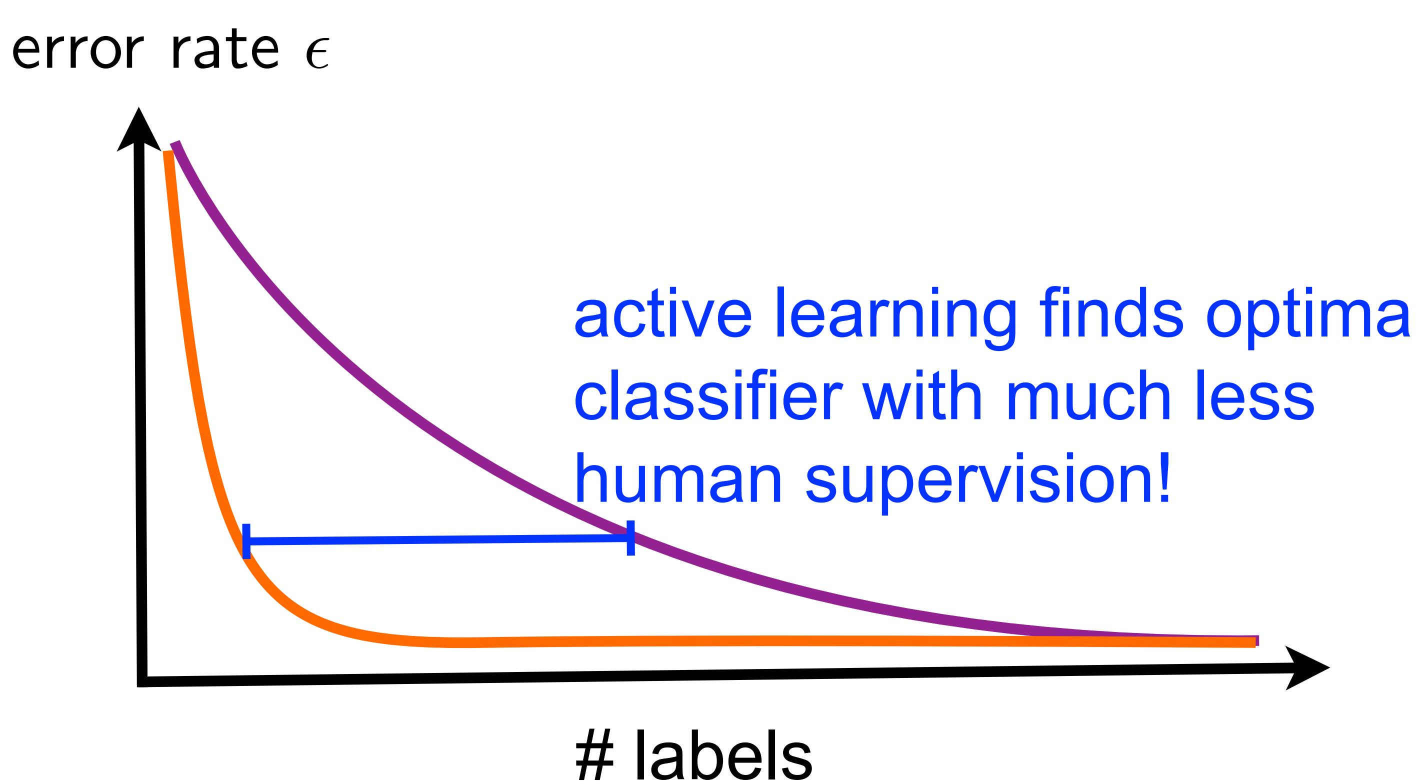
provides labels to machine learner
(several minutes / EHR)

Active Learning

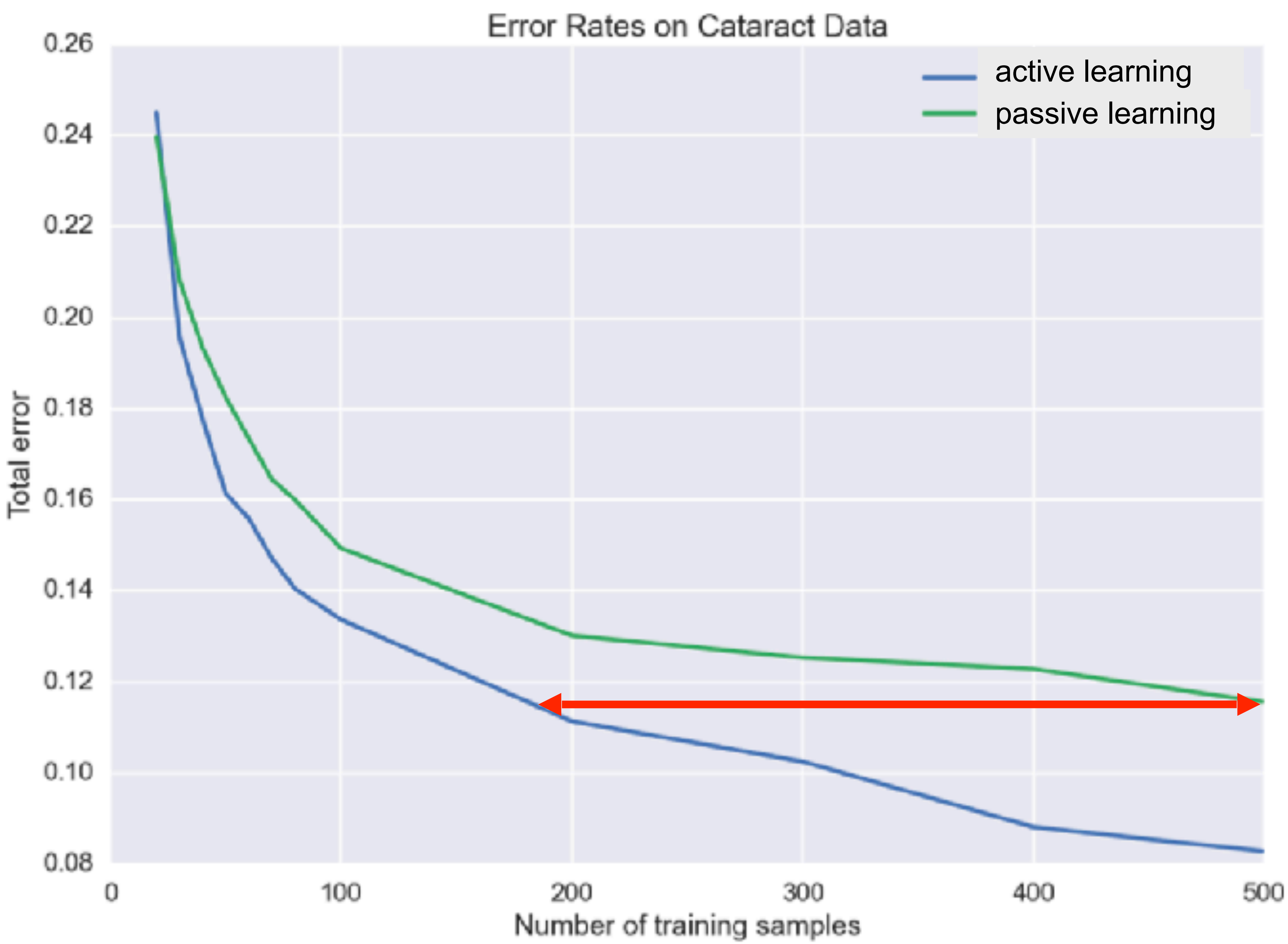


Non-adaptive strategy: Label a random sample

Active strategy: Label a sample near best decision boundary based on labels seen so far



Active Logistic Regression



11000 patient records

8000 positive

3000 negative

6182 Numerical Features

icd9 codes

lab tests

patient data

Classification task:
cataracts or healthy

less than half as many labeled examples needed by active learning

nextml.org

NEXT

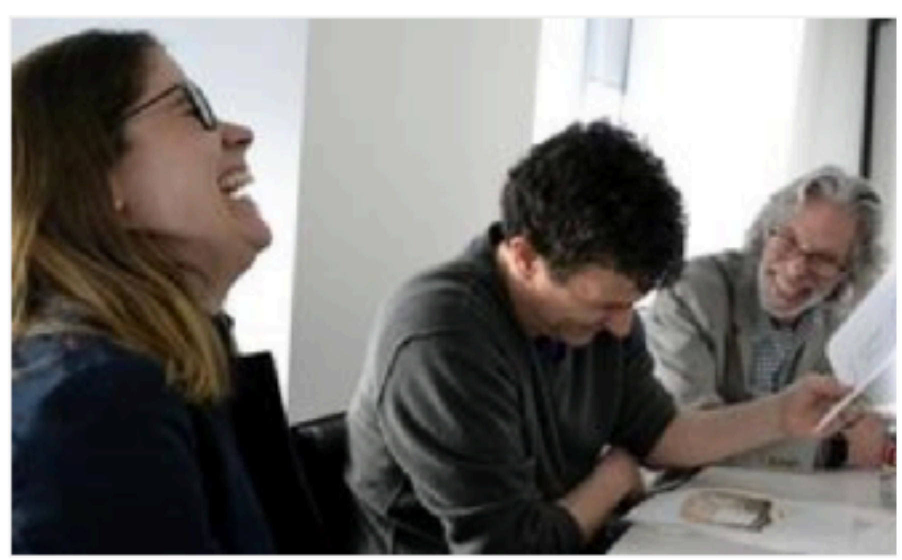
ASK BETTER QUESTIONS.
GET BETTER RESULTS.
FASTER. AUTOMATED.

GitHub | Paper | Docs

Blog | Team | Data

Active learning to optimize crowdsourcing and rating in New Yorker Cartoon Caption Contest

digg



BY DOING THE EXACT OPPOSITE

How New Yorker Cartoons Could Teach Computers To Be Funny

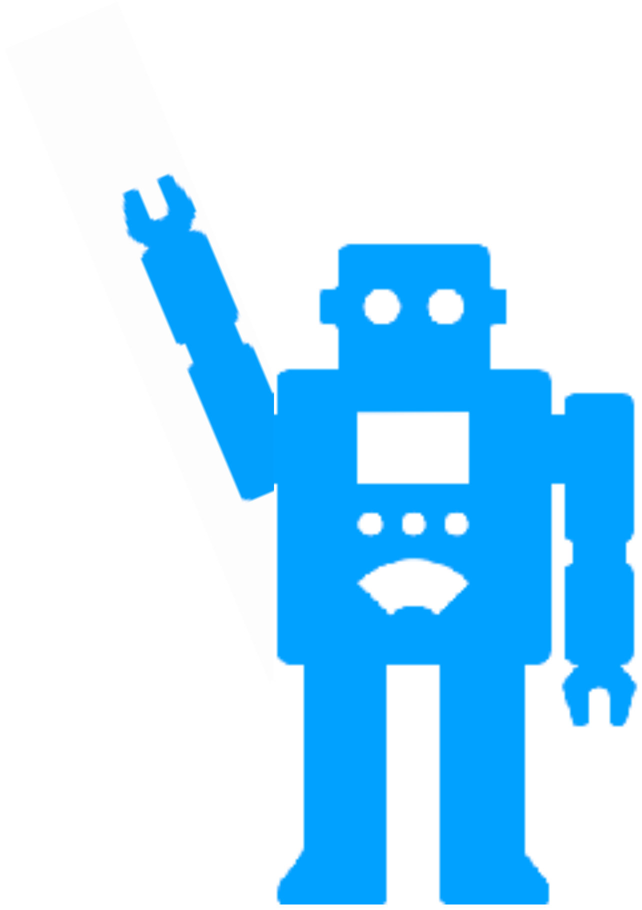
3 diggs CNET Technology

With the help of computer scientists from the University of Wisconsin at Madison, The New Yorker for the first time is using crowdsourcing algorithms to uncover the best captions.

3

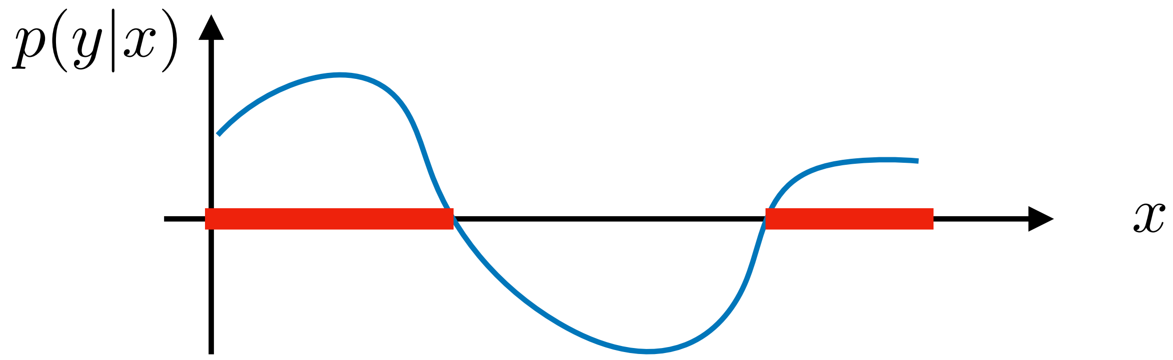
Actively learning user's beer preferences

Principles of Active Learning

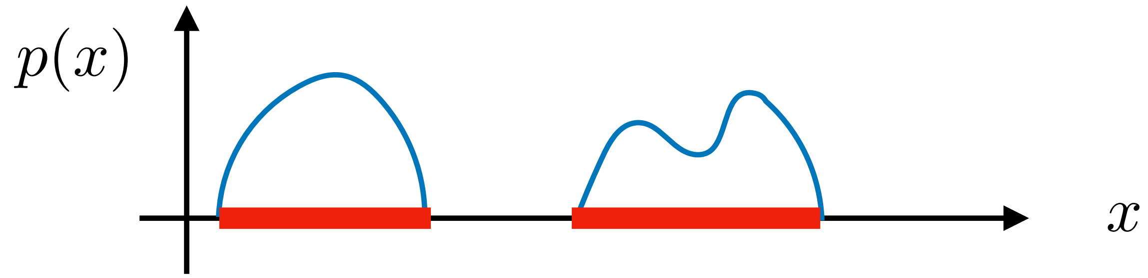


What and Where Information

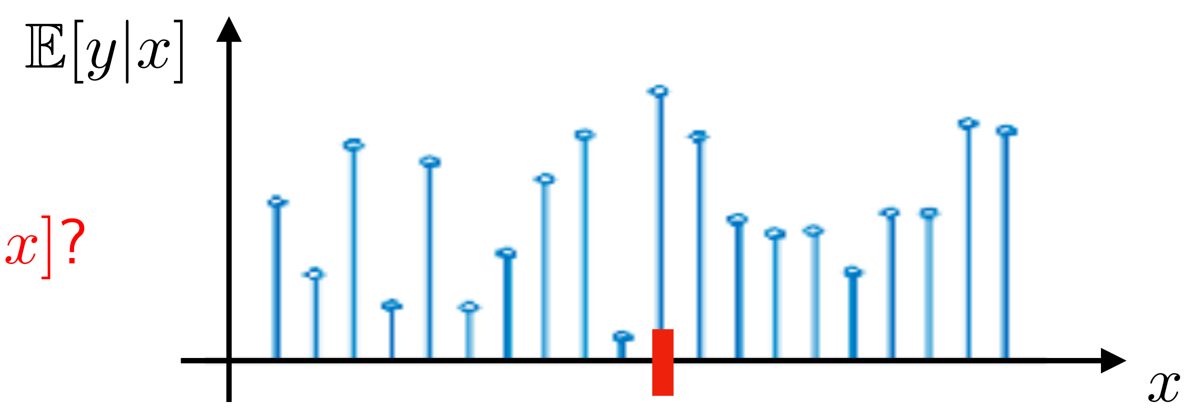
Density estimation: What is $p(y|x)$?
Classification: Where is $p(y|x) > 0$?



Density estimation: What is $p(x)$?
Clustering: Where is $p(x) > \epsilon$?



Function estimation: What is $\mathbb{E}[y|x]$?
Bandit optimization: Where is $\max_x \mathbb{E}[y|x]$?



Active learning is more efficient than passive learning for localized "where" information

Meta-Algorithm for Active Learning

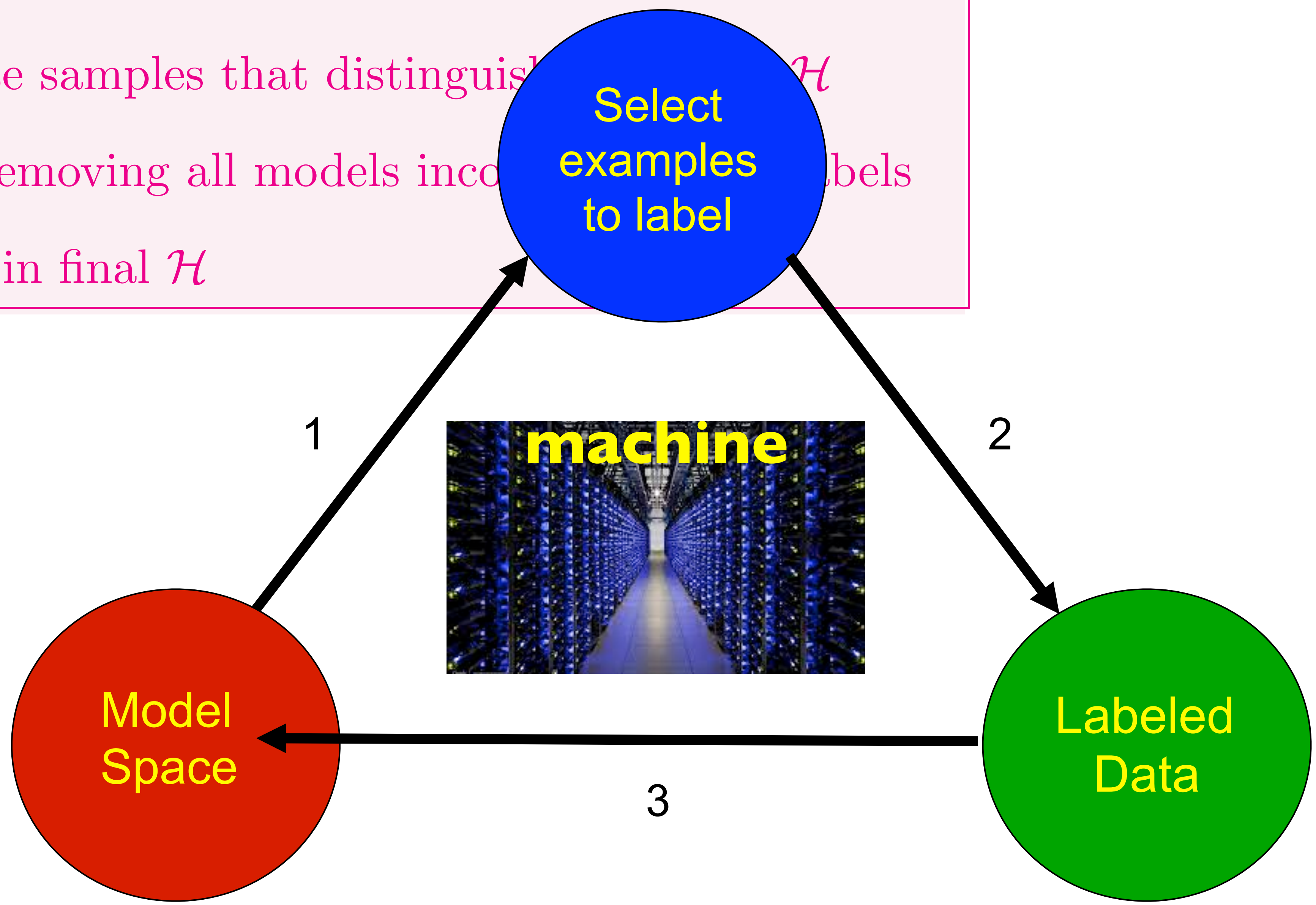
Version-Space (VS) Active Learning

initialize VS: $\mathcal{H}$ = all models/hypotheses

while (*stopping-criterion*) not met

1. **sample** at random from available dataset
2. **label** only those samples that distinguis … $\mathcal{H}$
3. **reduce** $\mathcal{H}$ by removing all models inco … bels

output: best model in final $\mathcal{H}$



Learning a 1-D Classifier

binary search quickly finds **decision boundary**

passive : err $\sim \quad n^{-1}$

active : err $\sim \quad 2^{-n}$

Vapnik-Chervonenkis (VC) Theory

Given training data $\{(x_j, y_j)\}_{j=1}^n$, learn a function f to predict y from x

Consider a possibly infinite set of hypotheses $\mathcal{F}$ with *finite VC dimension* d and for each $f \in \mathcal{F}$ define the risk (error rate):

$$R(f) \ := \ \mathbb{P}(f(x) \neq y)$$

error rate on training data: $$\widehat{R}(f) \ = \ \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\Big(f(x_i) \neq y_i\Big)$$ "empirical risk"

VC bound: $$\sup_{f \in \mathcal{F}} |R(f) - \widehat{R}(f)| \ \leq \ 6\sqrt{\frac{d\log(n/\delta)}{n}}$$

w.p. $\geq \ 1 - \delta$

Empirical Risk Minimization (ERM)

Goal: select hypothesis with true error rate within $\epsilon > 0$ of $\min_{f \in \mathcal{F}} R(f)$

$$f^* = \arg\min_{f \in \mathcal{F}} R(f)$$ true risk minimizer

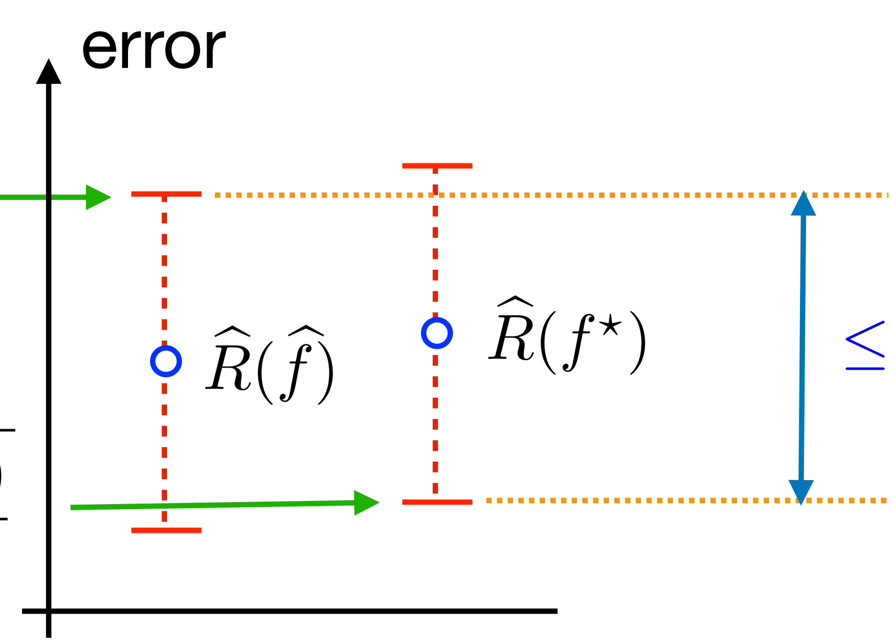
$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \widehat{R}(f)$$ empirical risk minimizer

$\widehat{f}$ minimizes empirical risk:

$$\widehat{R}(\widehat{f}) \leq \widehat{R}(f^*)$$

$$R(\widehat{f}) \leq \widehat{R}(\widehat{f}) + 6\sqrt{\frac{d\log(n/\delta)}{n}}$$

$$R(f^*) \geq \widehat{R}(f^*) - 6\sqrt{\frac{d\log(n/\delta)}{n}}$$



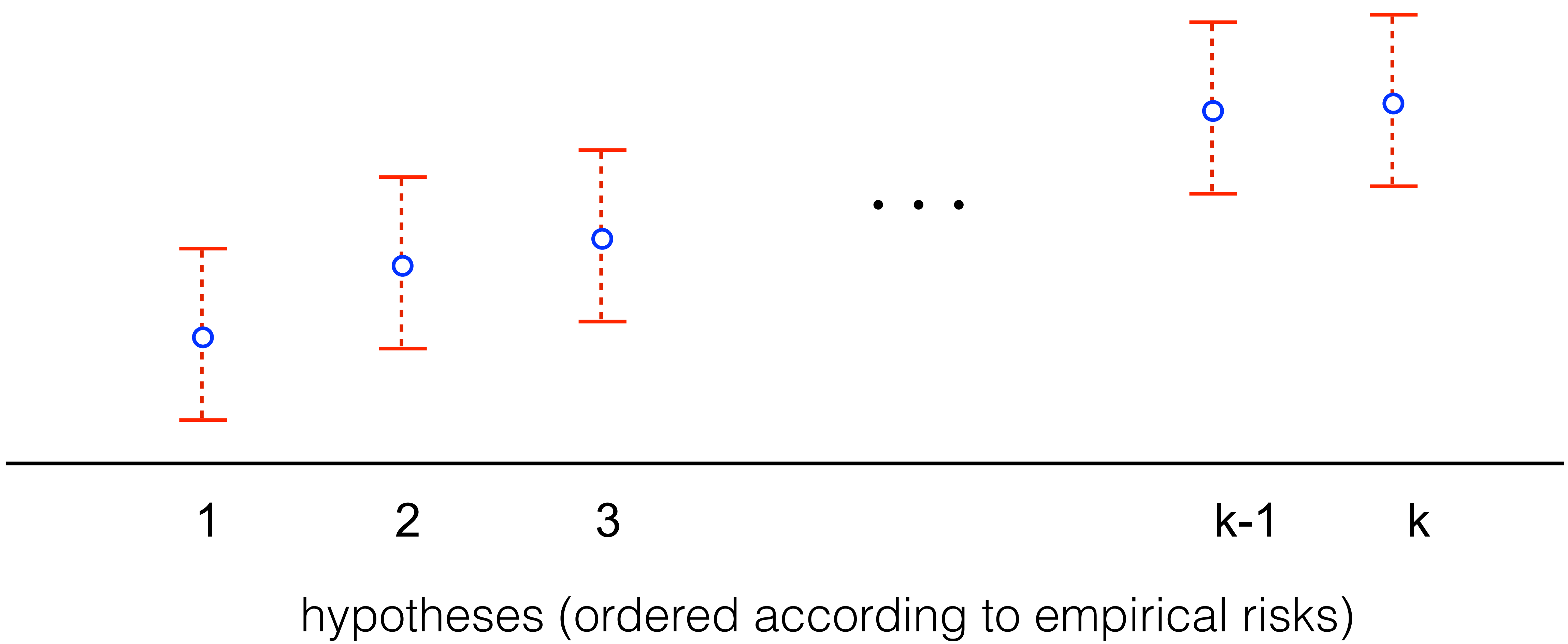
$$\leq \quad 12\sqrt{\frac{d\log(n/\delta)}{n}}$$

sufficient number
of training examples:

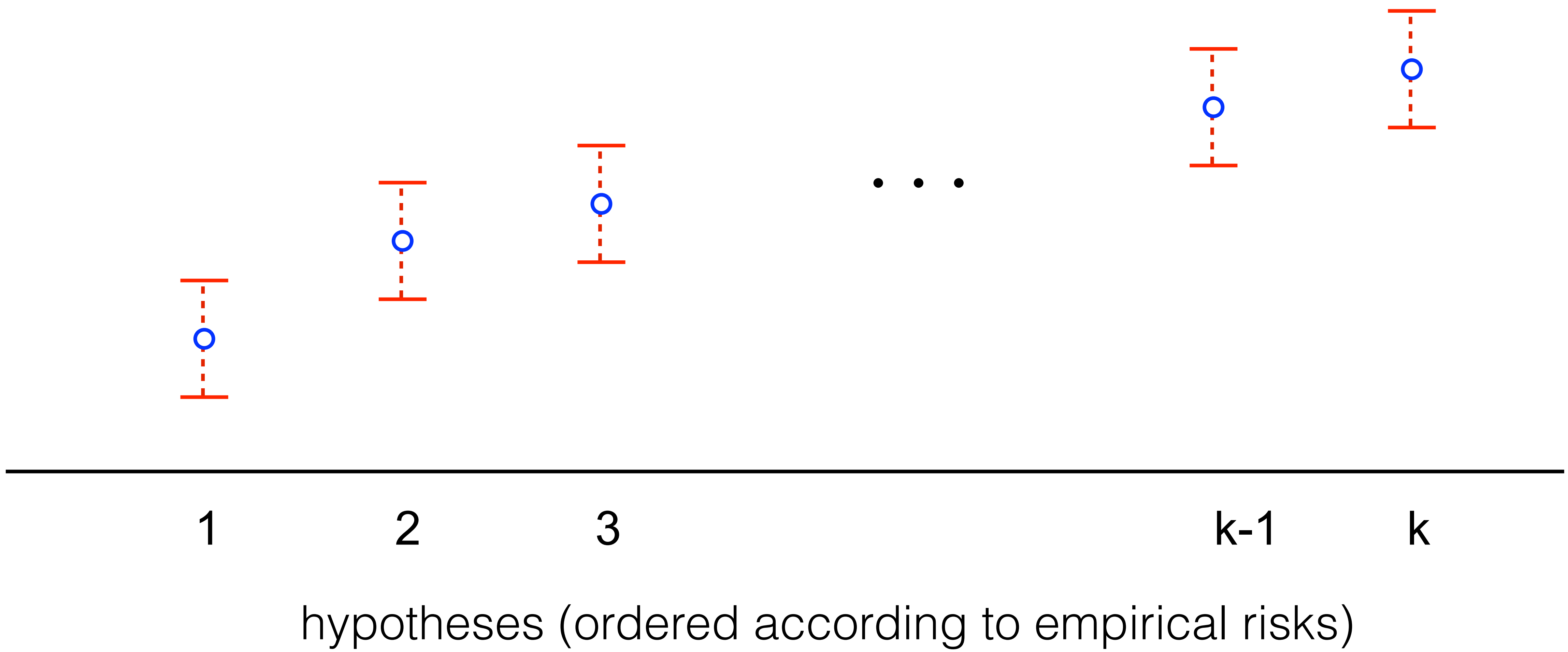
$$12\sqrt{\frac{d\log(n/\delta)}{n}} \leq \epsilon$$

$$n = \widetilde{O}\left(\frac{d\log(1/\delta)}{\epsilon^2}\right)$$

Empirical Risks and Confidence Intervals

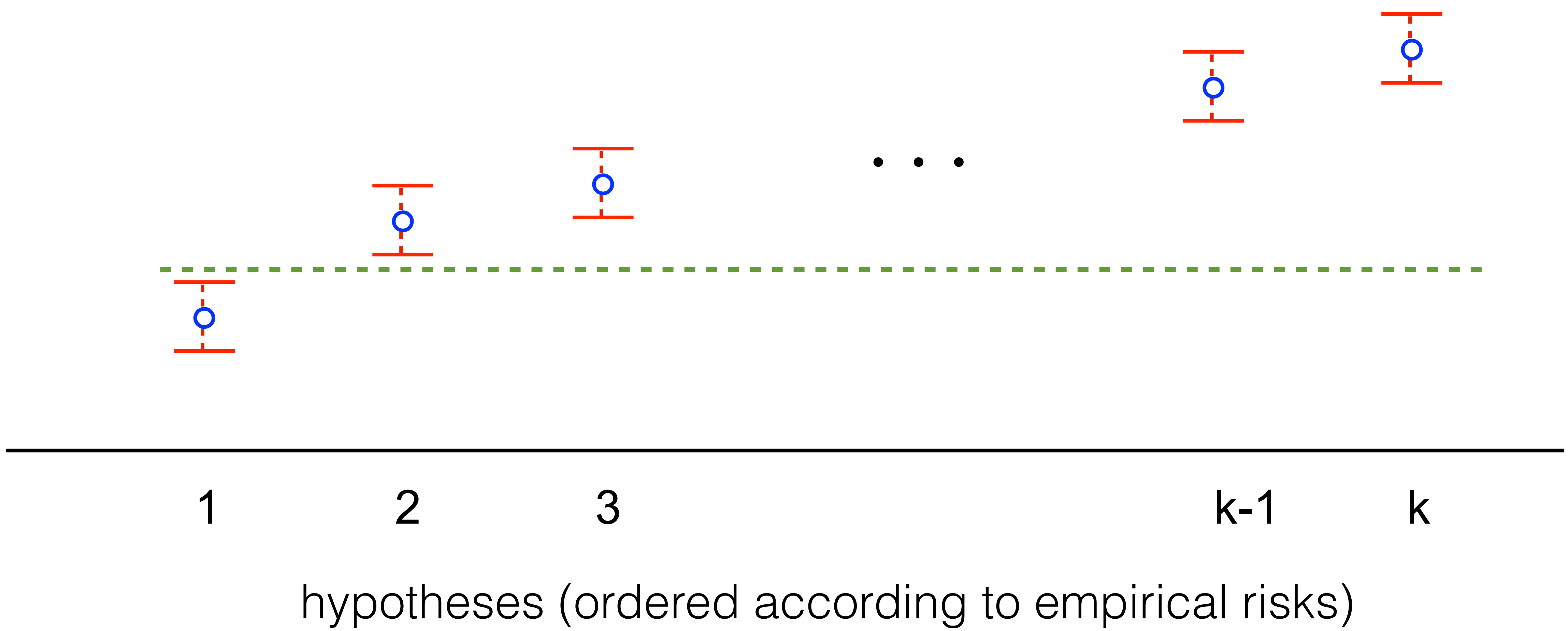


Empirical Risks and Confidence Intervals



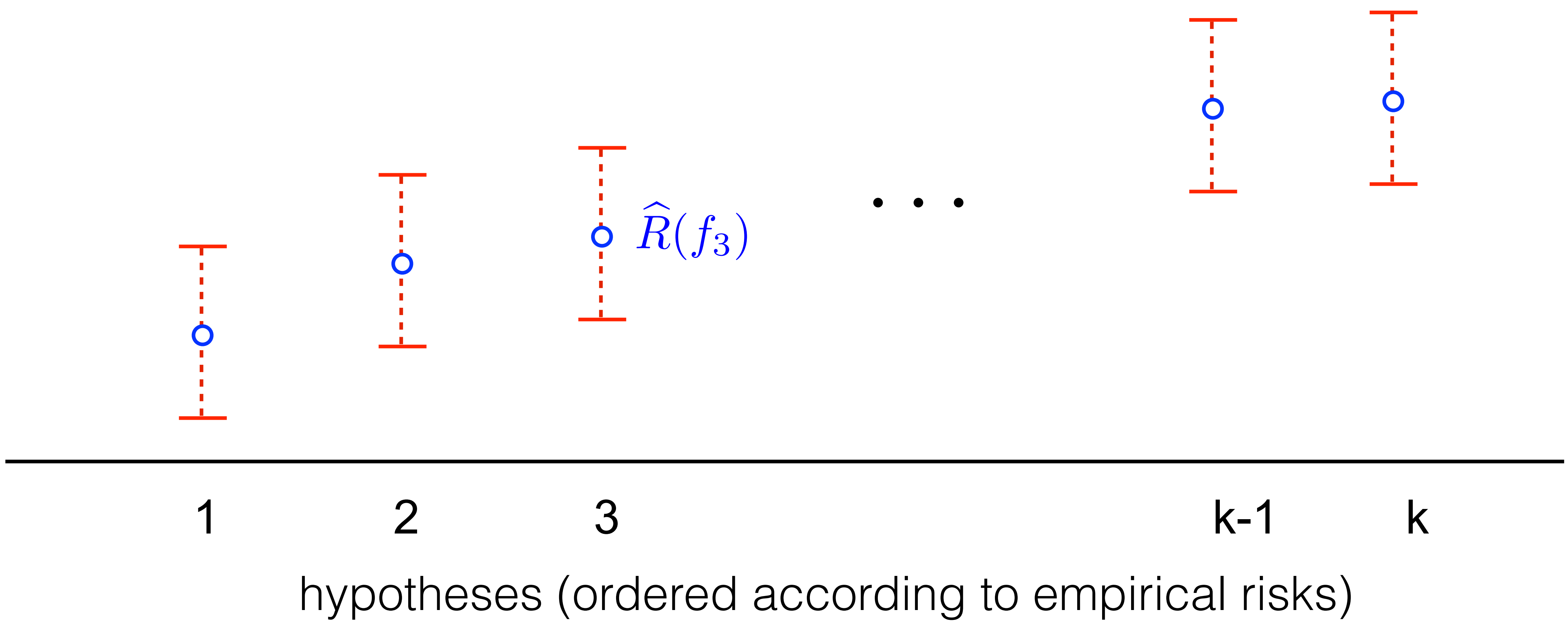
more training data $\Rightarrow$ smaller confidence intervals

Empirical Risks and Confidence Intervals

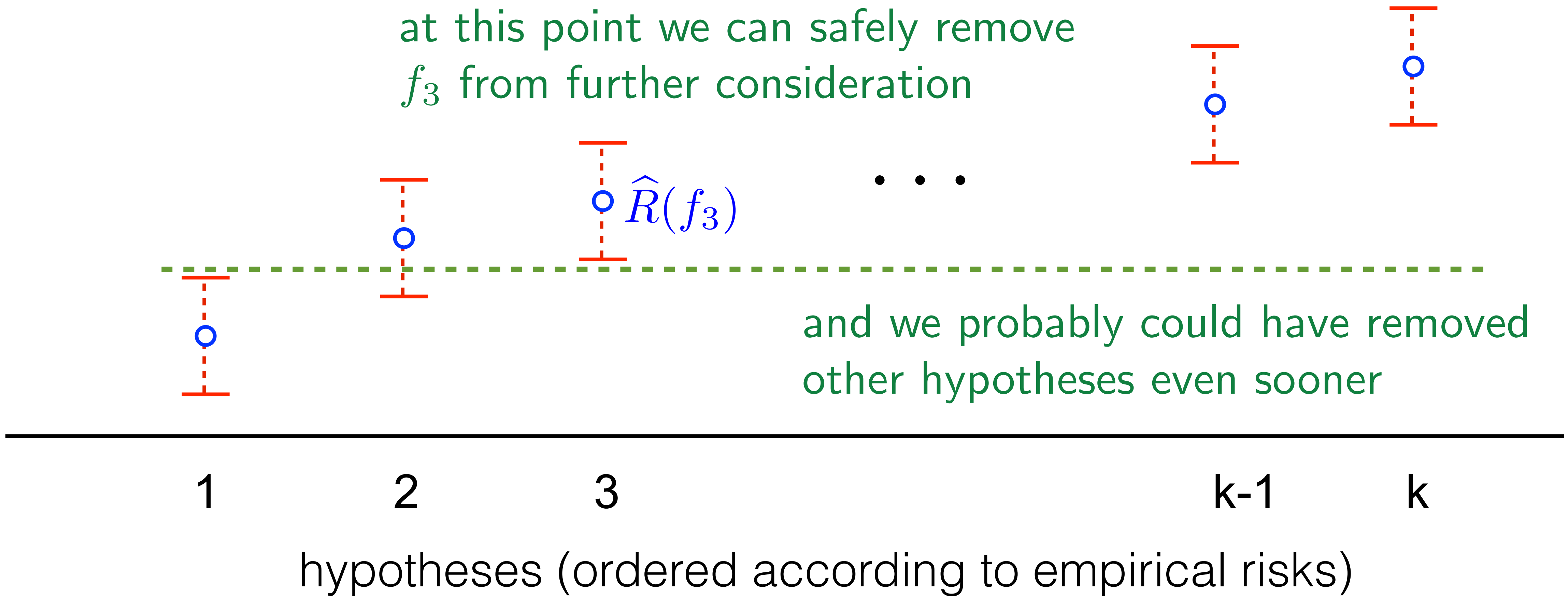


more training data $\Rightarrow$ smaller confidence intervals

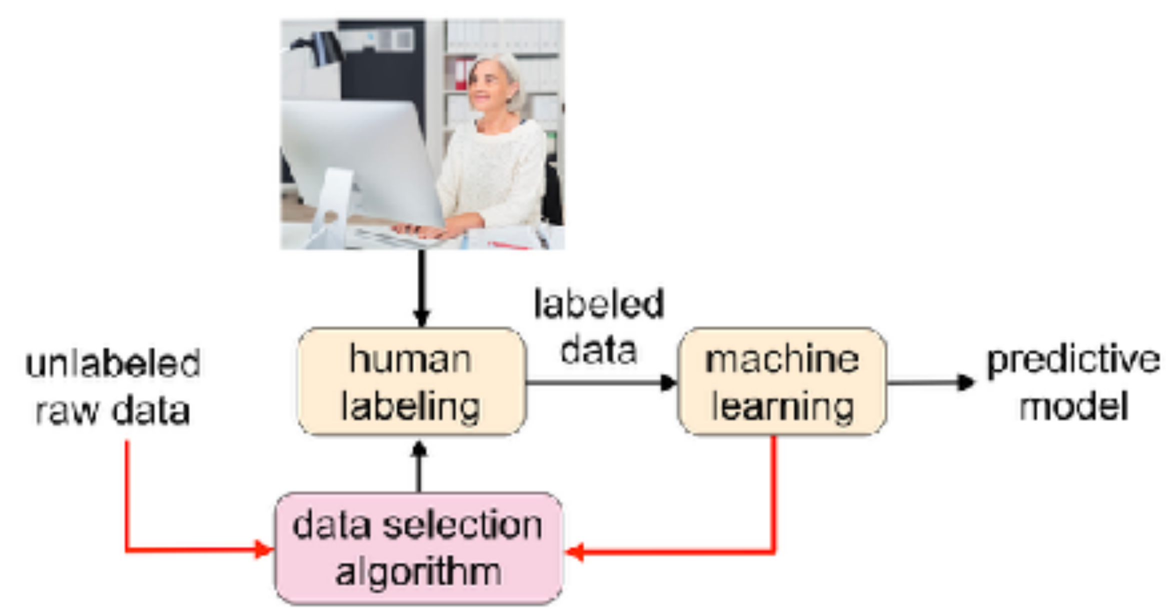
ERM is Wasting Labeled Examples



ERM is Wasting Labeled Examples

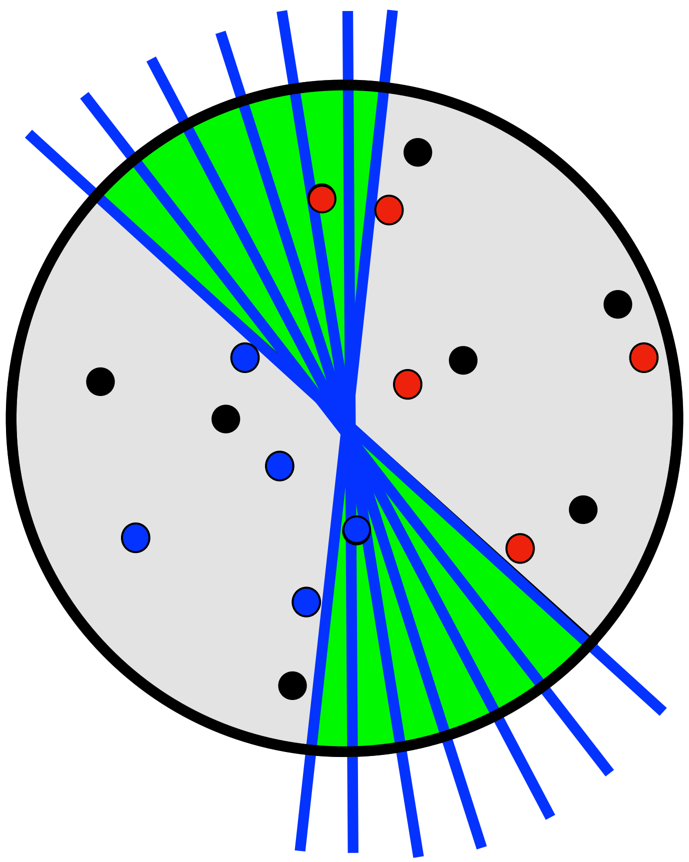


only require labels for examples that hypotheses 1 and 2 label differently (i.e., examples where they *disagree*)



Disagreement-Based Active Learning

consider points uniform on unit ball and
linear classifiers passing through origin



only label points in the
region of disagreement $\mathfrak{D}$

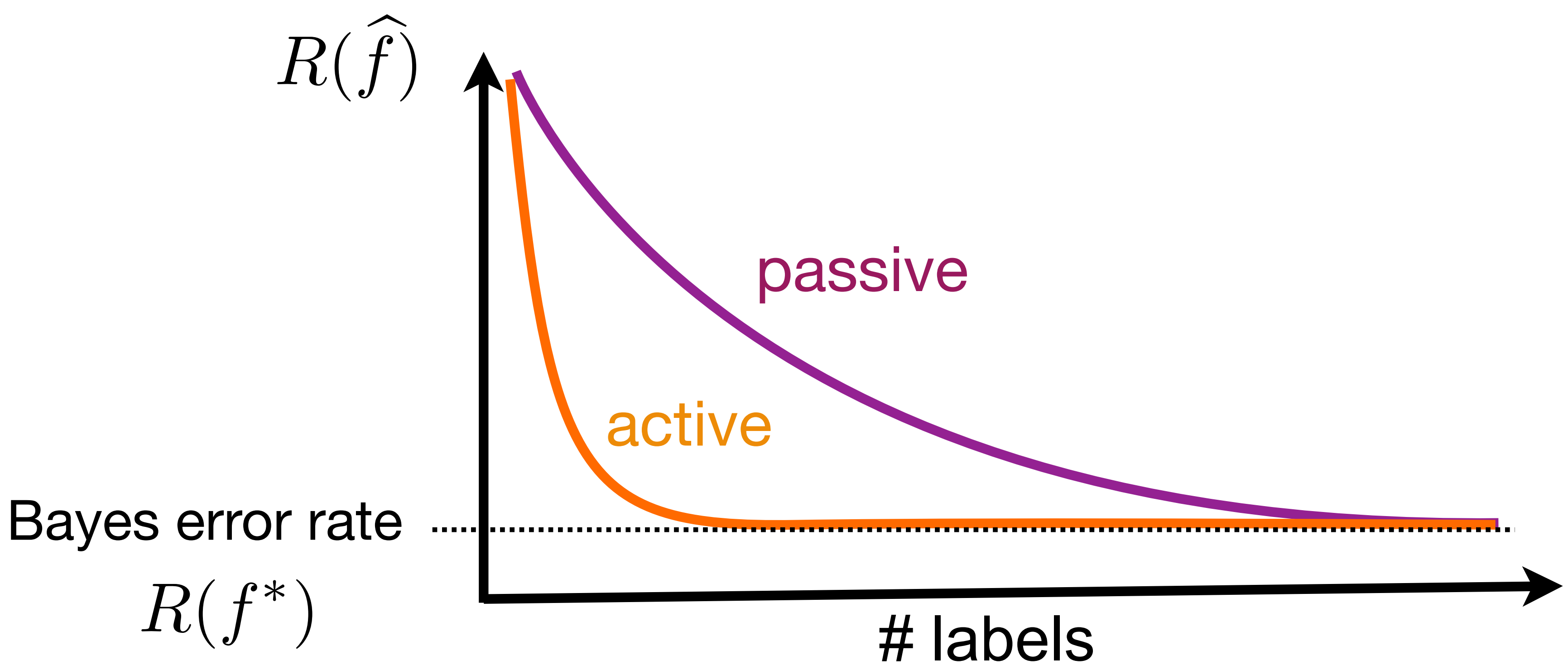
Active Binary Classification

Assuming optimal Bayes classifer f^* in VC class with dimension d and "nice" distributions (e.g., bounded label noise)

$$\epsilon = R(\widehat{f}) - R(f^*)$$

passive $\quad \epsilon \sim \dfrac{d}{n} \quad$ parametric rate

active $\quad \epsilon \sim \exp\left(-c\,\dfrac{n}{d}\right) \quad$ exponential speed-up



Tutorial Outline

Part 1: Introduction to Active Learning (Rob)

Part 2: Theory of Active Learning (Steve)

Part 3: Advanced Topics and Open Problems (Steve)

Part 4: Nonparametric Active Learning (Rob)

slides: http://nowak.ece.wisc.edu/ActiveML.html

Recommended Reading (Foundations of Active Learning)

Settles, Burr. "Active learning." *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6.1 (2012): 1-114.

Dasgupta, Sanjoy. "Two faces of active learning." *Theoretical computer science* 412.19 (2011): 1767-1781.

Cohn, David, Les Atlas, and Richard Ladner. "Improving generalization with active learning." *Machine learning* 15.2 (1994): 201-221.

Castro, Rui M., and Robert D. Nowak. "Minimax bounds for active learning." *IEEE Transactions on Information Theory* 54, no. 5 (2008): 2339-2353.

Zhu, Xiaojin, John Lafferty, and Zoubin Ghahramani. "Combining active learning and semi-supervised learning using gaussian fields and harmonic functions." *ICML 2003 workshop*. Vol. 3. 2003.

Dasgupta, Sanjoy, Daniel J. Hsu, and Claire Monteleoni. "A general agnostic active learning algorithm." *Advances in neural information processing systems*. 2008.

Balcan, Maria-Florina, Alina Beygelzimer, and John Langford. "Agnostic active learning." *Journal of Computer and System Sciences* 75.1 (2009): 78-89.

Nowak, Robert D. "The geometry of generalized binary search." *IEEE Transactions on Information Theory* 57, no. 12 (2011): 7893-7906.

Hanneke, Steve. "Theory of active learning." *Foundations and Trends in Machine Learning* 7, no. 2-3 (2014).