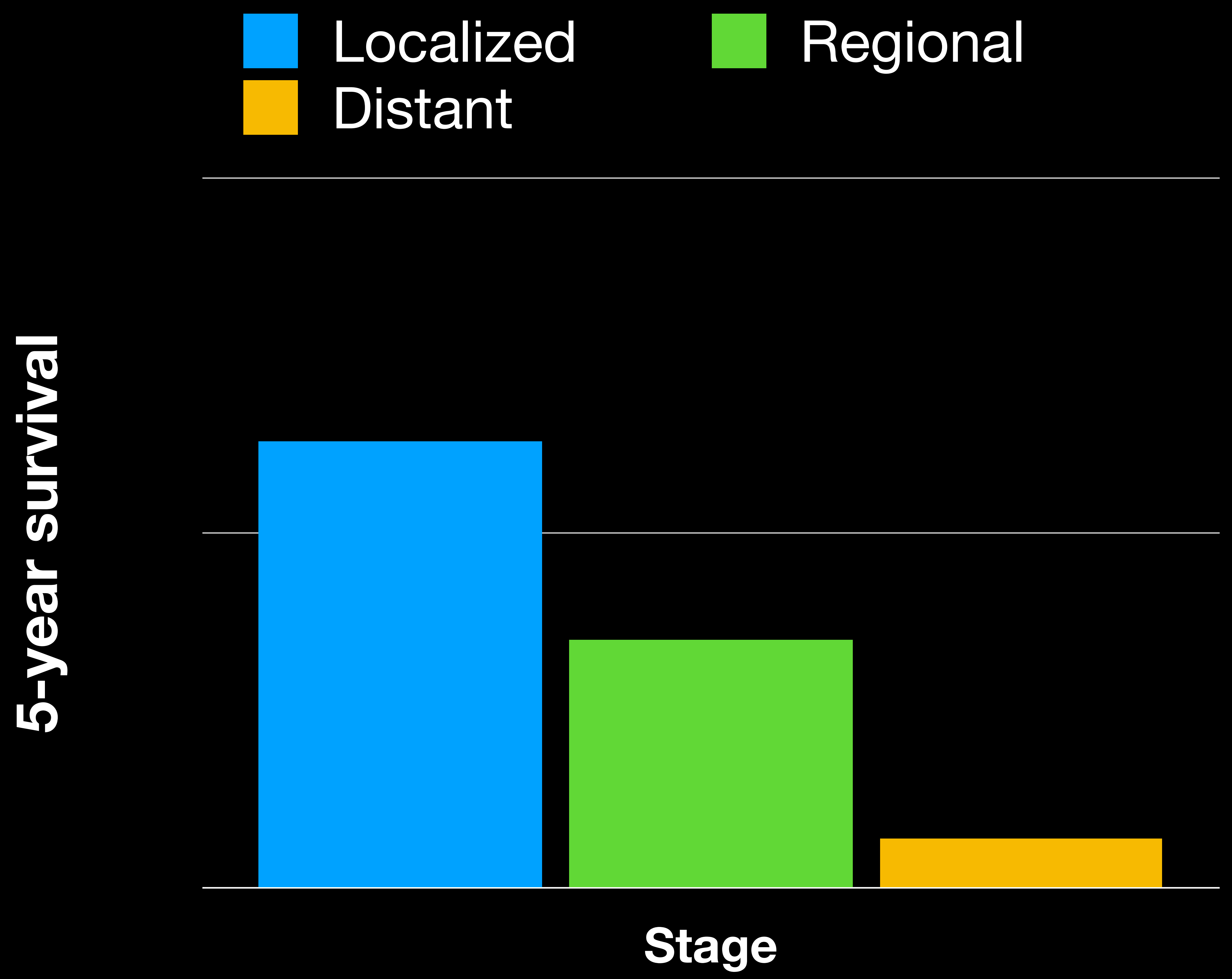
CPH 100A: Machine Learning Foundations I

Instructor: Adam Yala, PhD (yala@berkeley.edu)

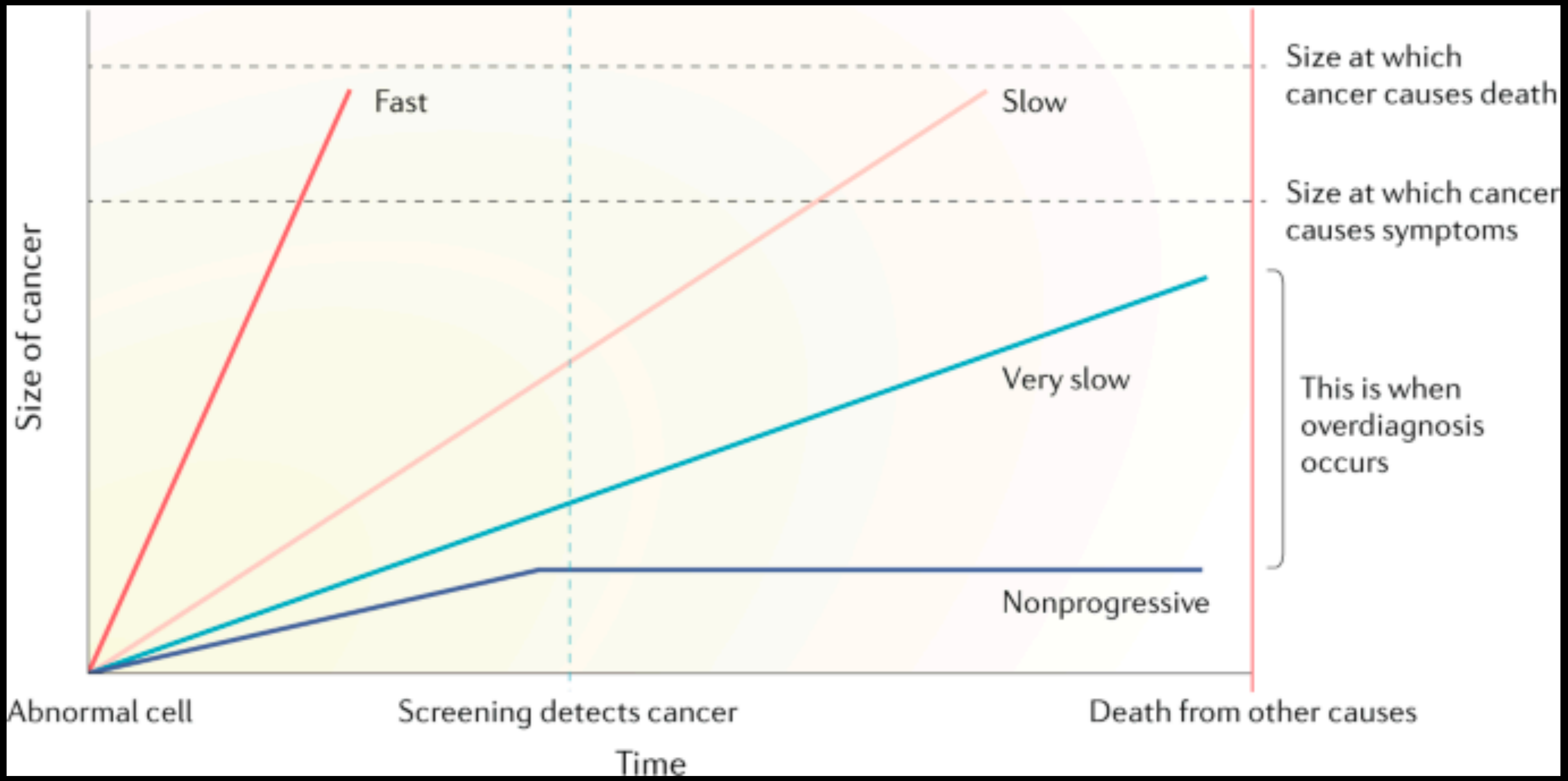
Computational
PRECISION HEALTH

Berkeley | UCSF

Problem Motivation: Early Detection is critical



Lung Cancer



RCTs reduce lung cancer mortality

ORIGINAL ARTICLE

Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening

The National Lung Screening Trial Research Team

NLST reduces lung cancer mortality by 20%

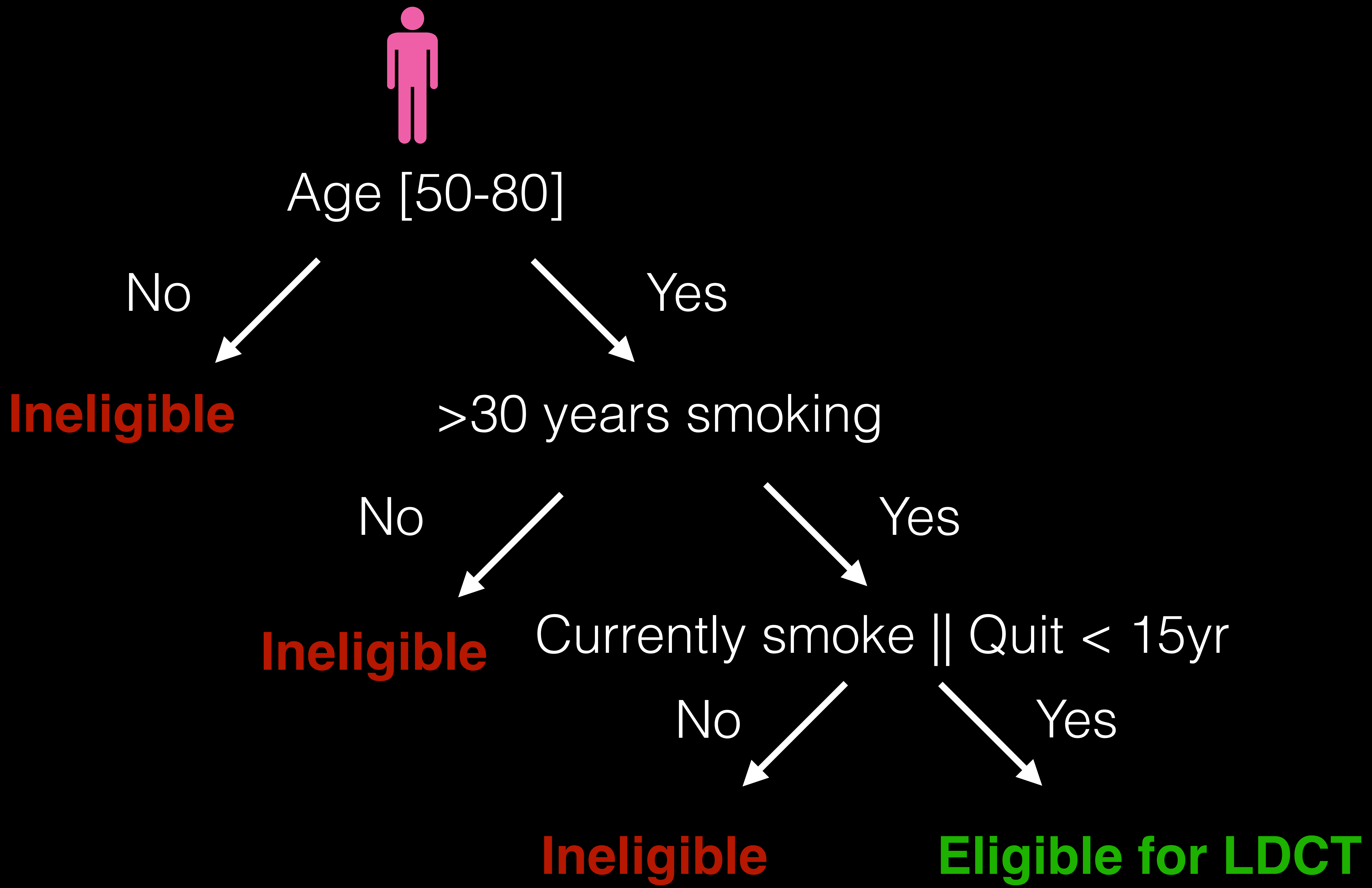
ORIGINAL ARTICLE

Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial

Harry J. de Koning, M.D., Ph.D., Carlijn M. van der Aalst, Ph.D., Pim A. de Jong, M.D., Ph.D., Ernst T. Scholten, M.D., Ph.D., Kristiaan Nackaerts, M.D., Ph.D., Marjolein A. Heuvelmans, M.D., Ph.D., Jan-Willem J. Lammers, M.D., Ph.D., Carla Weenink, M.D., Uraujh Yousaf-Khan, M.D., Ph.D., Nanda Horeweg, M.D., Ph.D., Susan van 't Westeind M.D., Ph.D., Mathias Prokop, M.D., Ph.D., et al.

NELSON reduces lung cancer mortality by 24%

NLST screening criteria

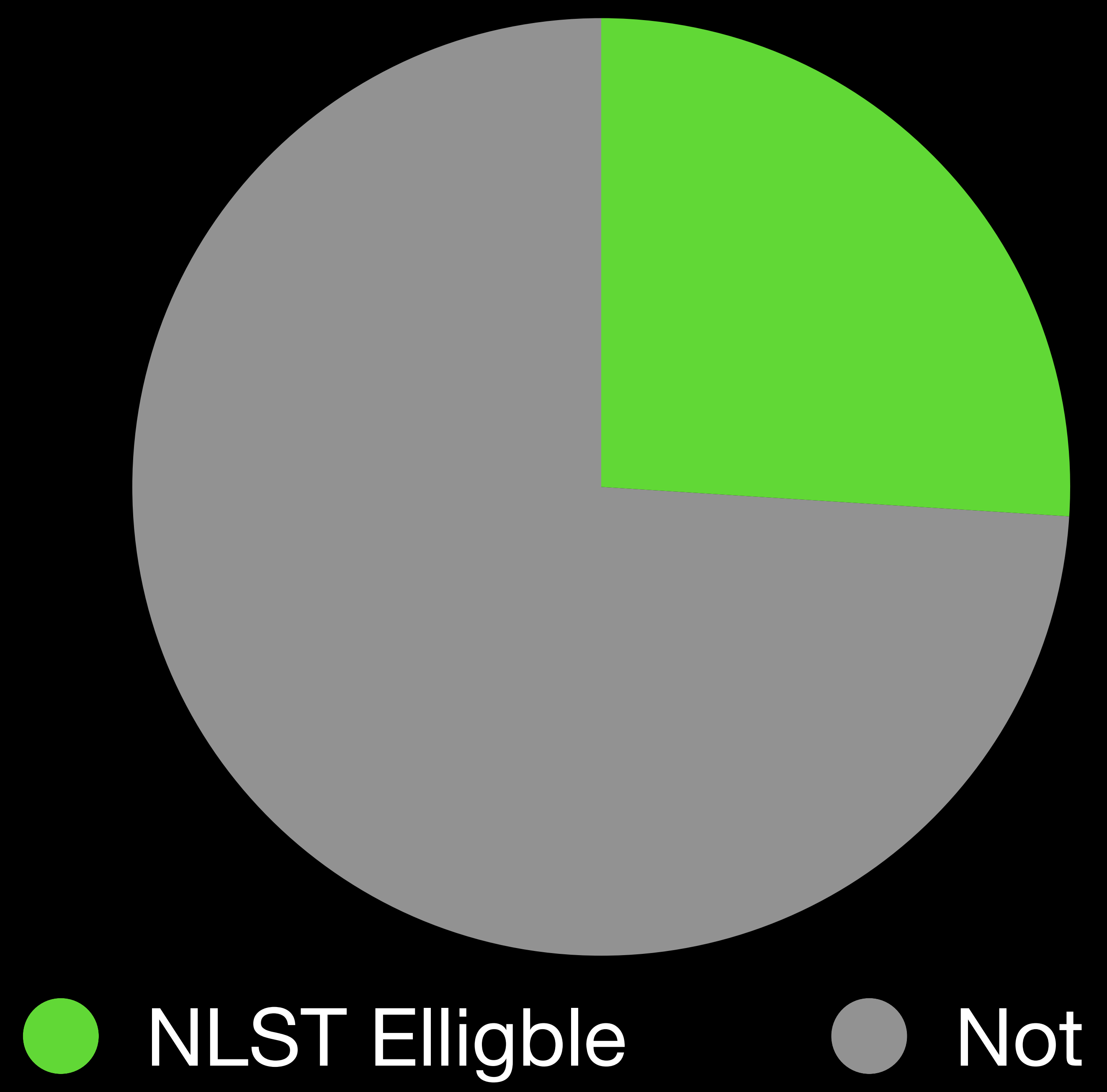


Efficacy of a screening program

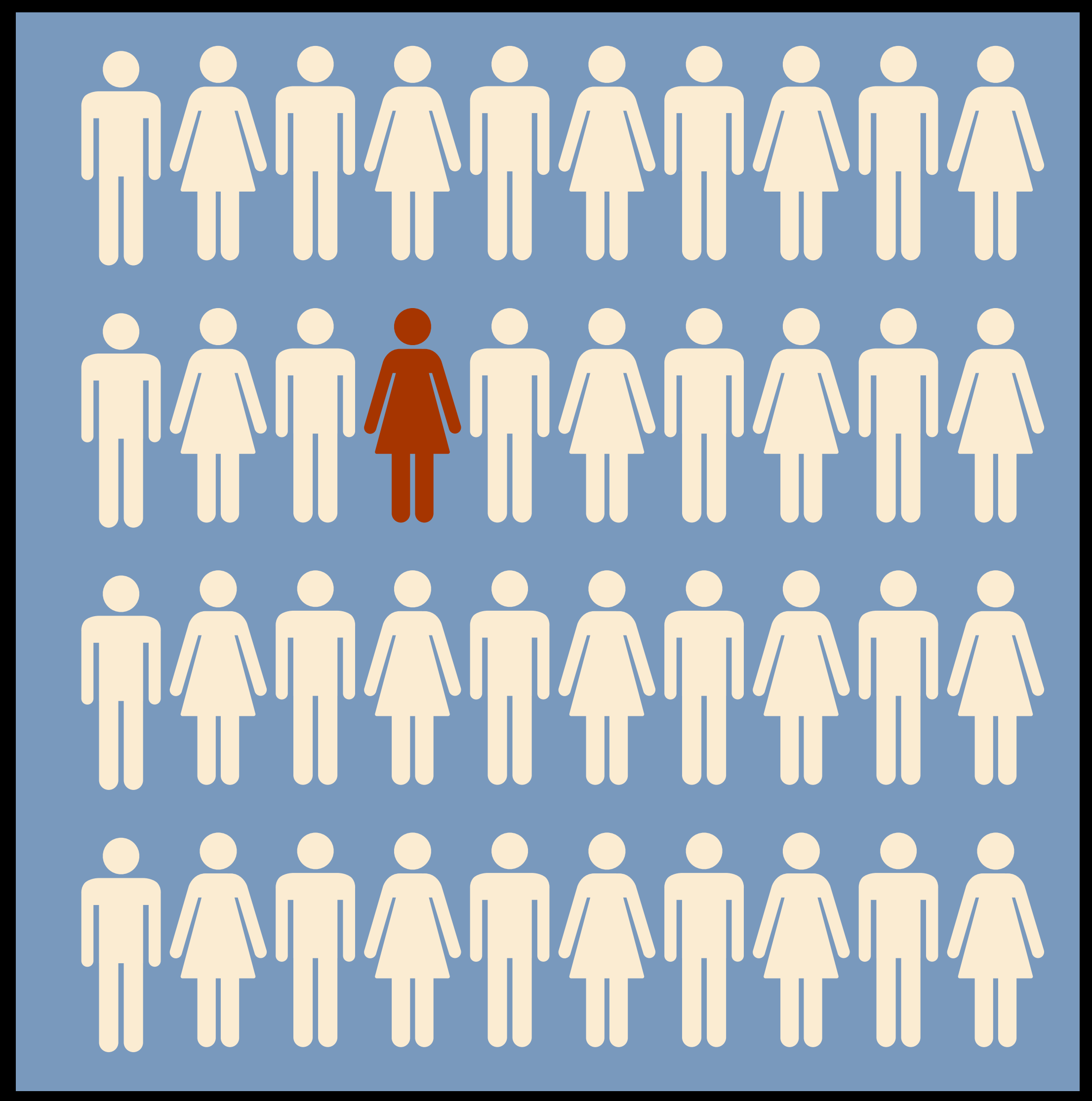
Fundamental challenge is **cost-effectiveness**

How much benefit does it achieve?

How much harm does the program do?



PMID: 23060474



1000 screens

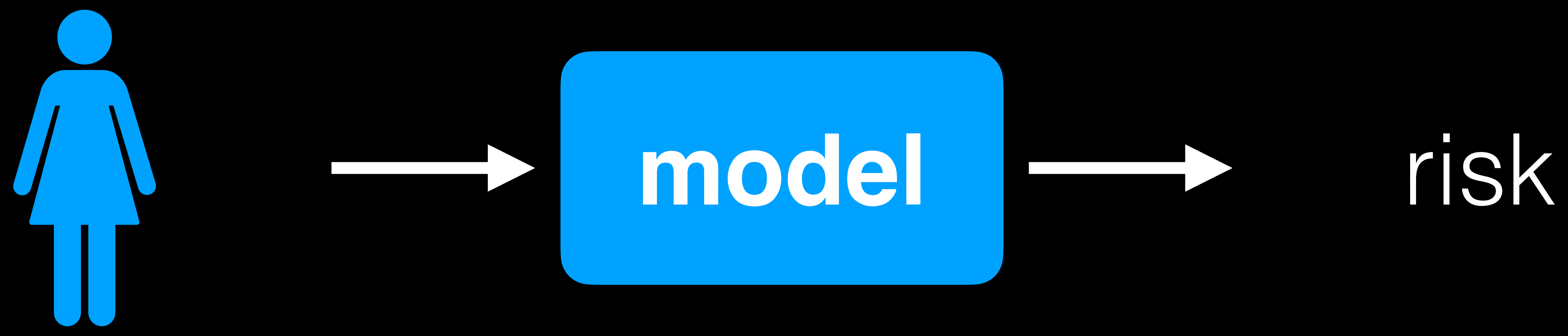
↓

240 positives

↓

6 cancers

Can we do better?



Predict probability of cancer *(proxy for prob screening benefit)*

Identify population with **higher specificity** and **higher sensitivity**

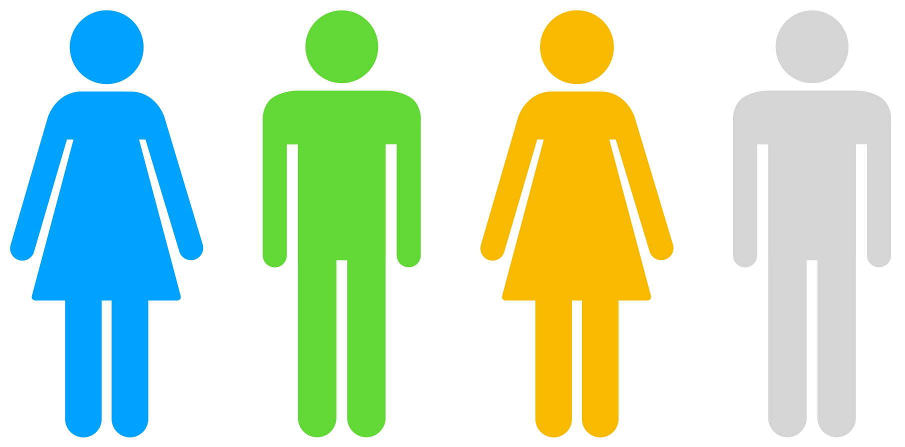
Key Question for today:

How do these models work?

How should we be evaluating them?

What we know

n historical patients

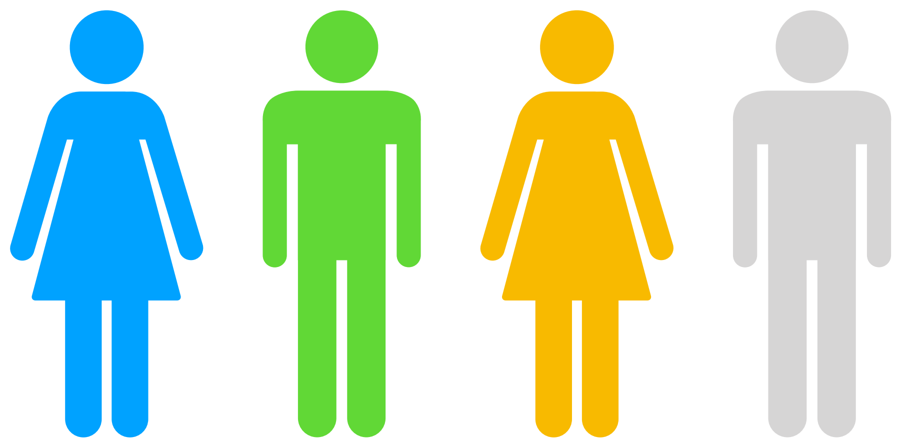


x $\qquad$ $x \in \mathscr{R}^D$

y $\qquad$ $y \in \{0,1\}$

What we know

n historical patients



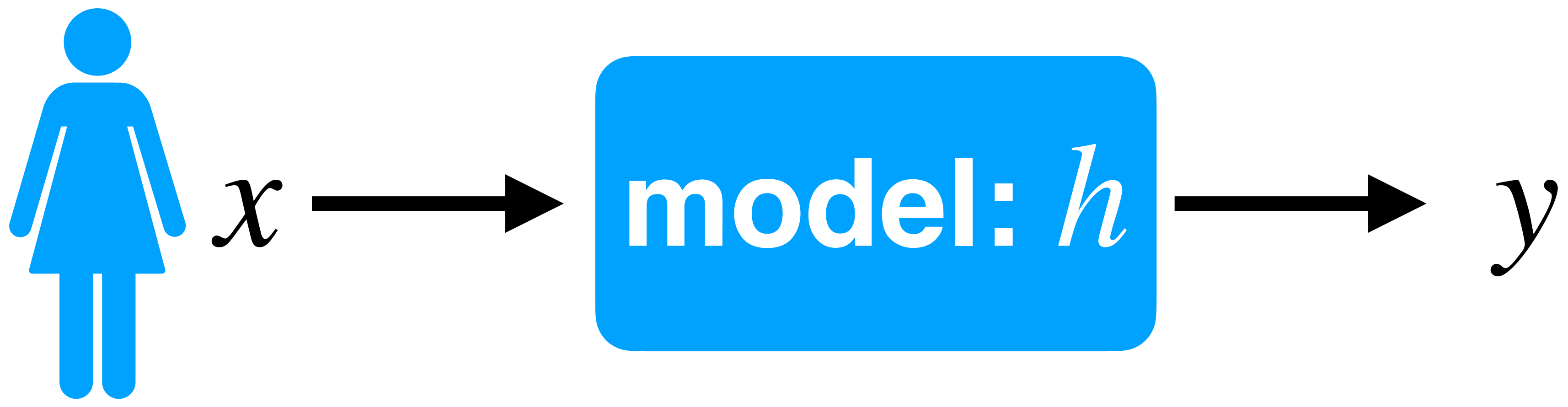
x — age, smoking status, weight, ⋮ — d

y → cancer

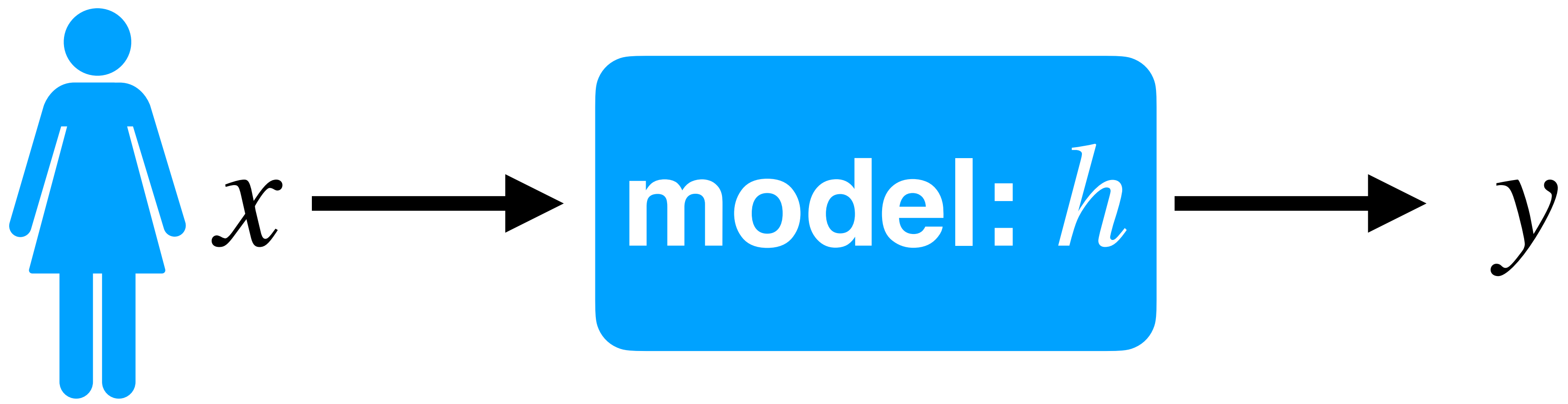
$x \in \mathscr{R}^D$

$y \in \{0,1\}$

What we want



What we want



Some func h $\quad x \to h \to y$

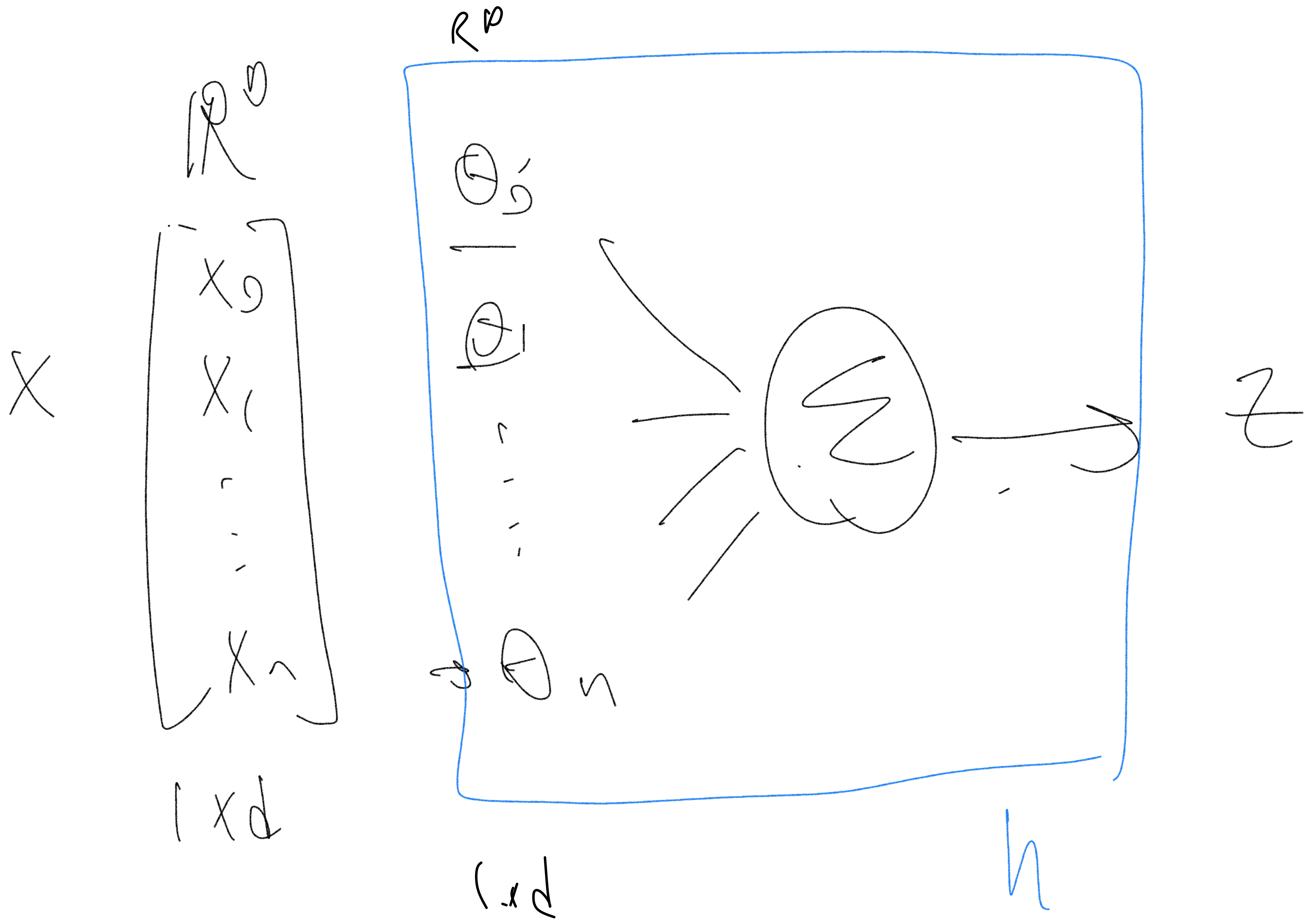
and h to be "good"

What is h?

$$h : \mathbb{R}^D \to \mathbb{R}$$

Todays Hypothesis class: linear models

Todays Hypothesis class: linear models

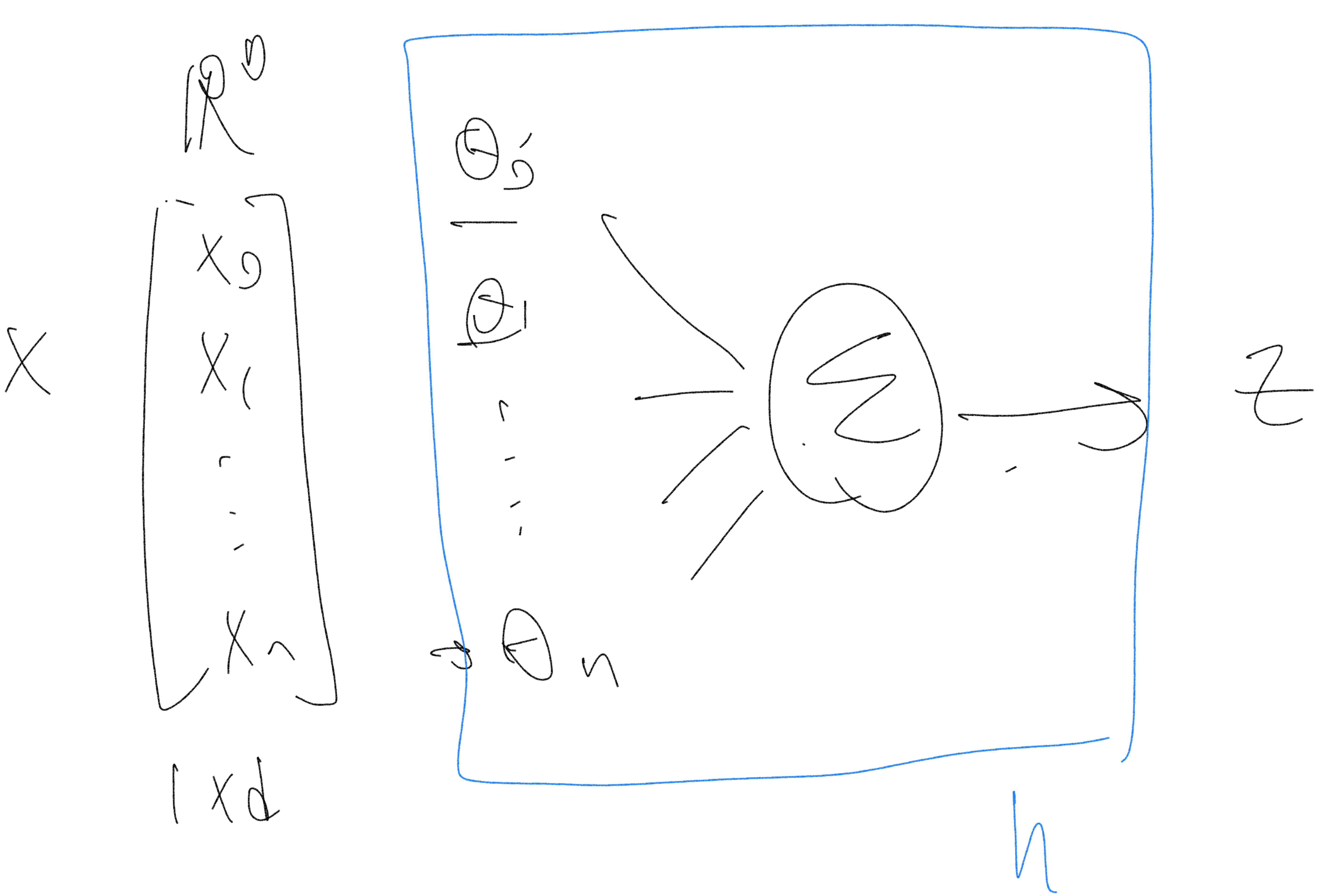


$$z = \sum_{i}^{N} \theta_i x_i + b$$

$$\boxed{z = \theta x^T}$$

simplified notation

Interpreting Linear Models



$x_0 \to$ age

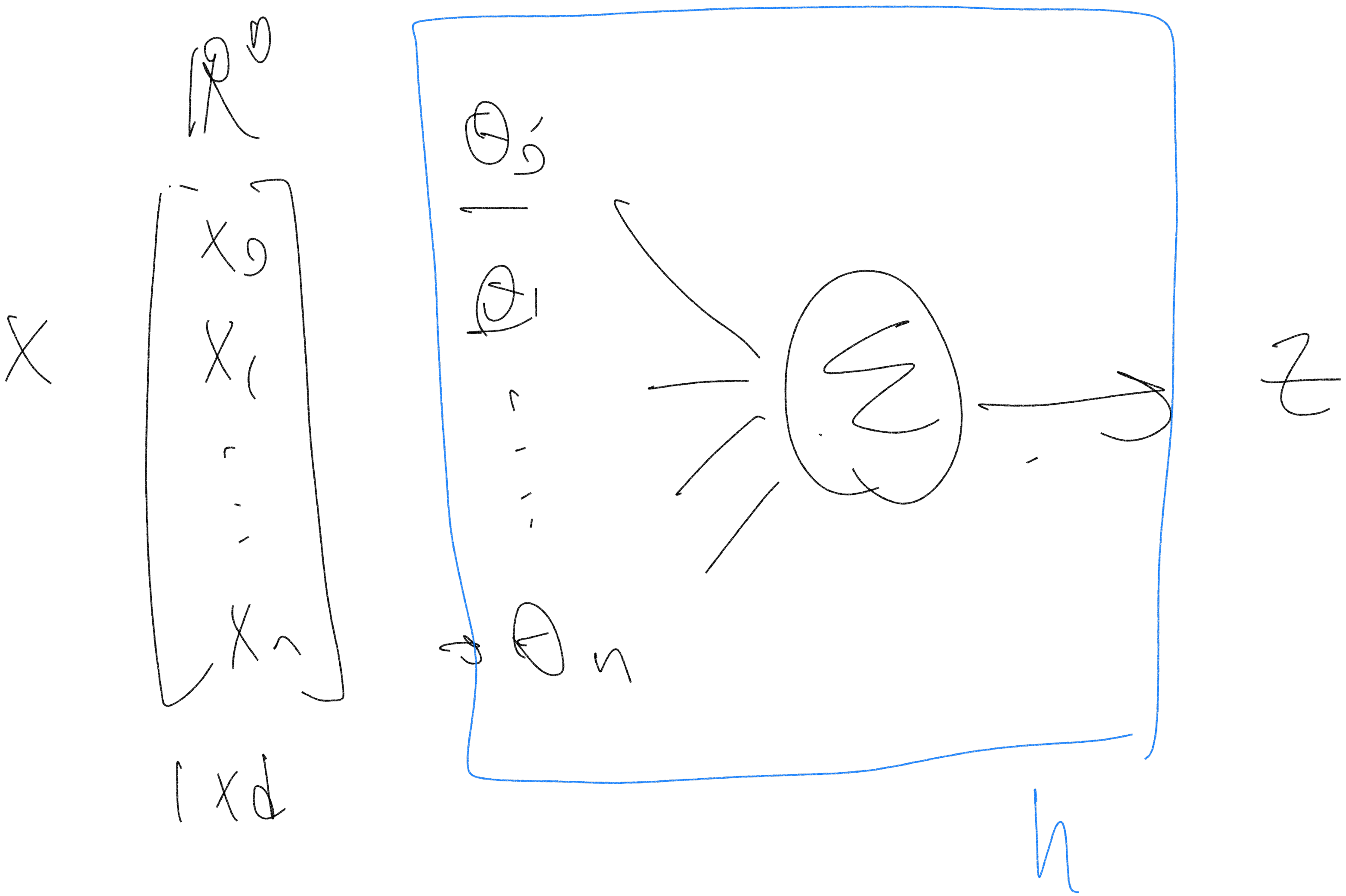
$x_1 \to$ smoking dur

$\uparrow z$:

$\uparrow \theta_0$:

$\uparrow \theta_1$:

Interpreting Linear Models



$x_0 \to$ age

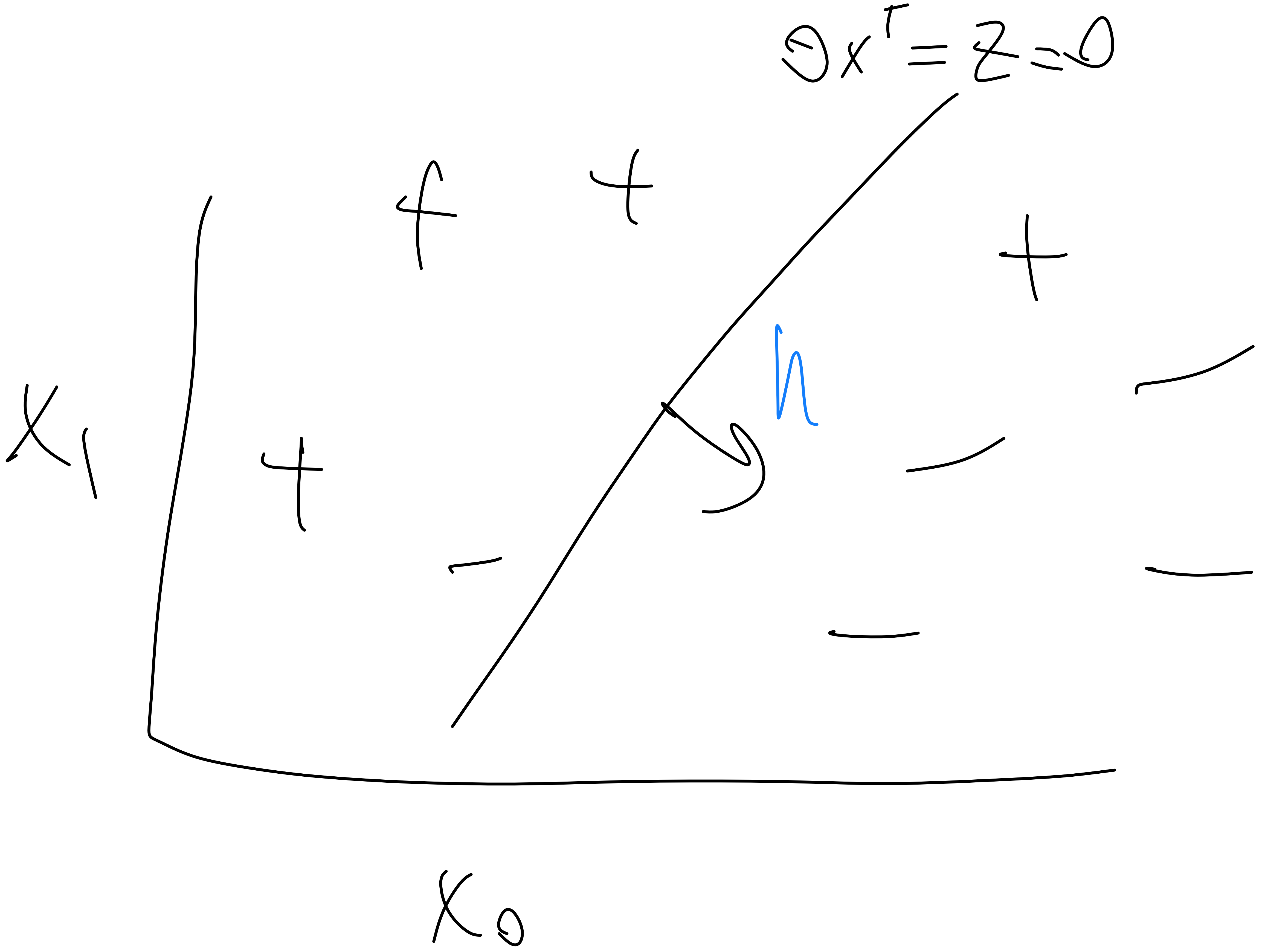
$x_1 \to$ smoking dur

$\uparrow z$: higher cancer risk

$\uparrow \theta_0$: age pos assoc

$\uparrow \theta_1$: smoking pos assoc

Geometric View



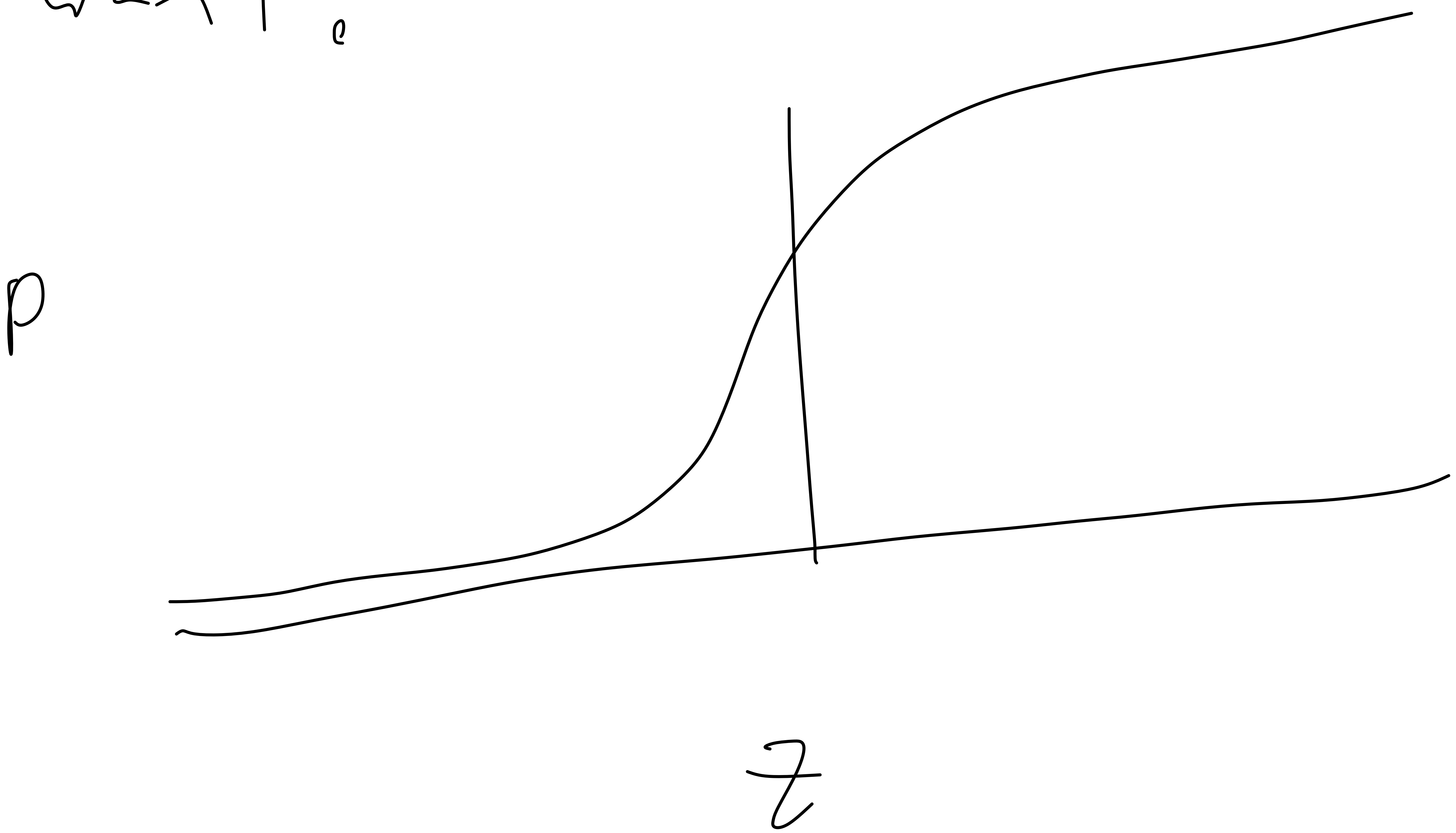
Capturing Uncertainty

Capturing Uncertainty

$\mathcal{H} : \theta x^T$ but we want $\mathbb{R}^D \to [0,1]$

$\mathbb{R}^D \to \mathbb{R}$

we want:

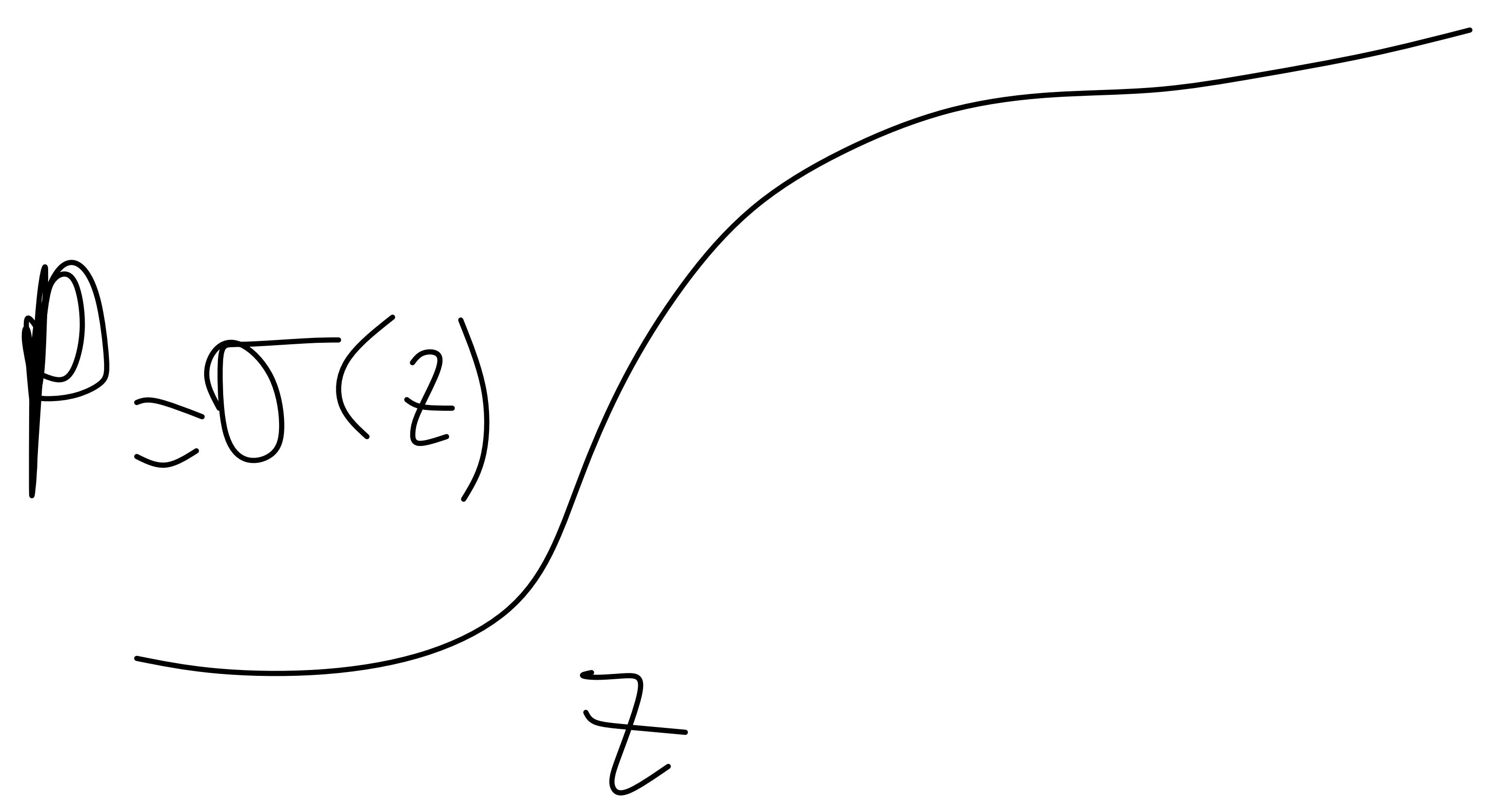


$\uparrow$ scores $\to p(y)=1$

$\downarrow$ score $\to p(y)=0$

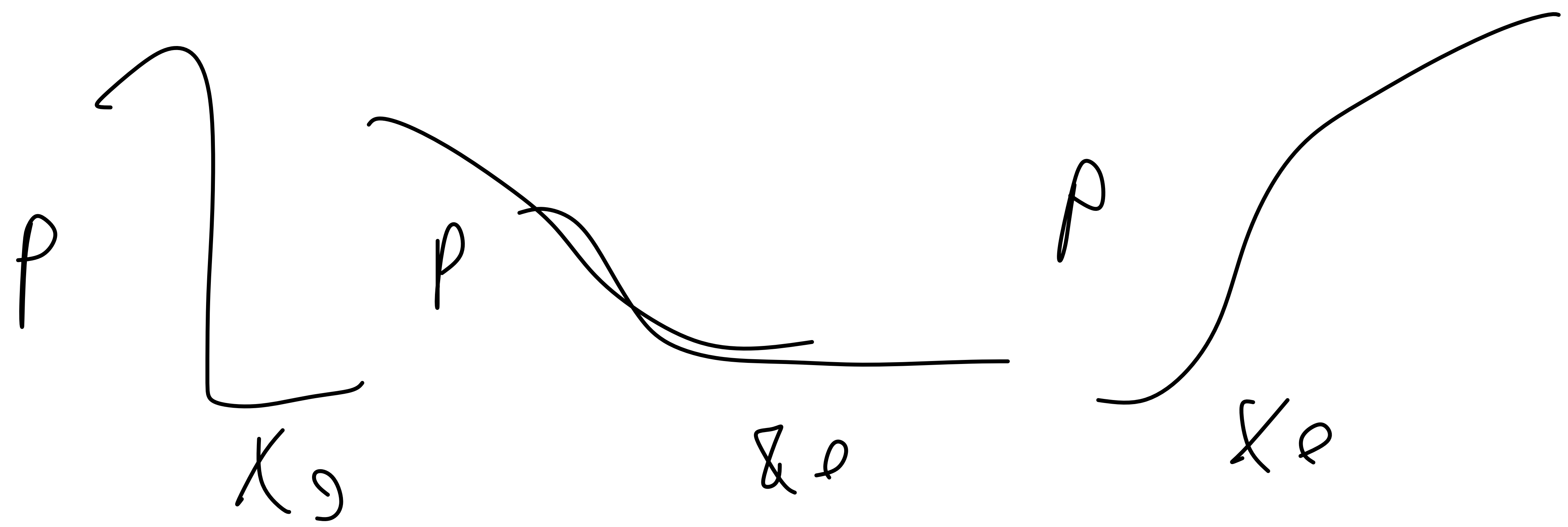
Loglinear Models

Loglinear Models



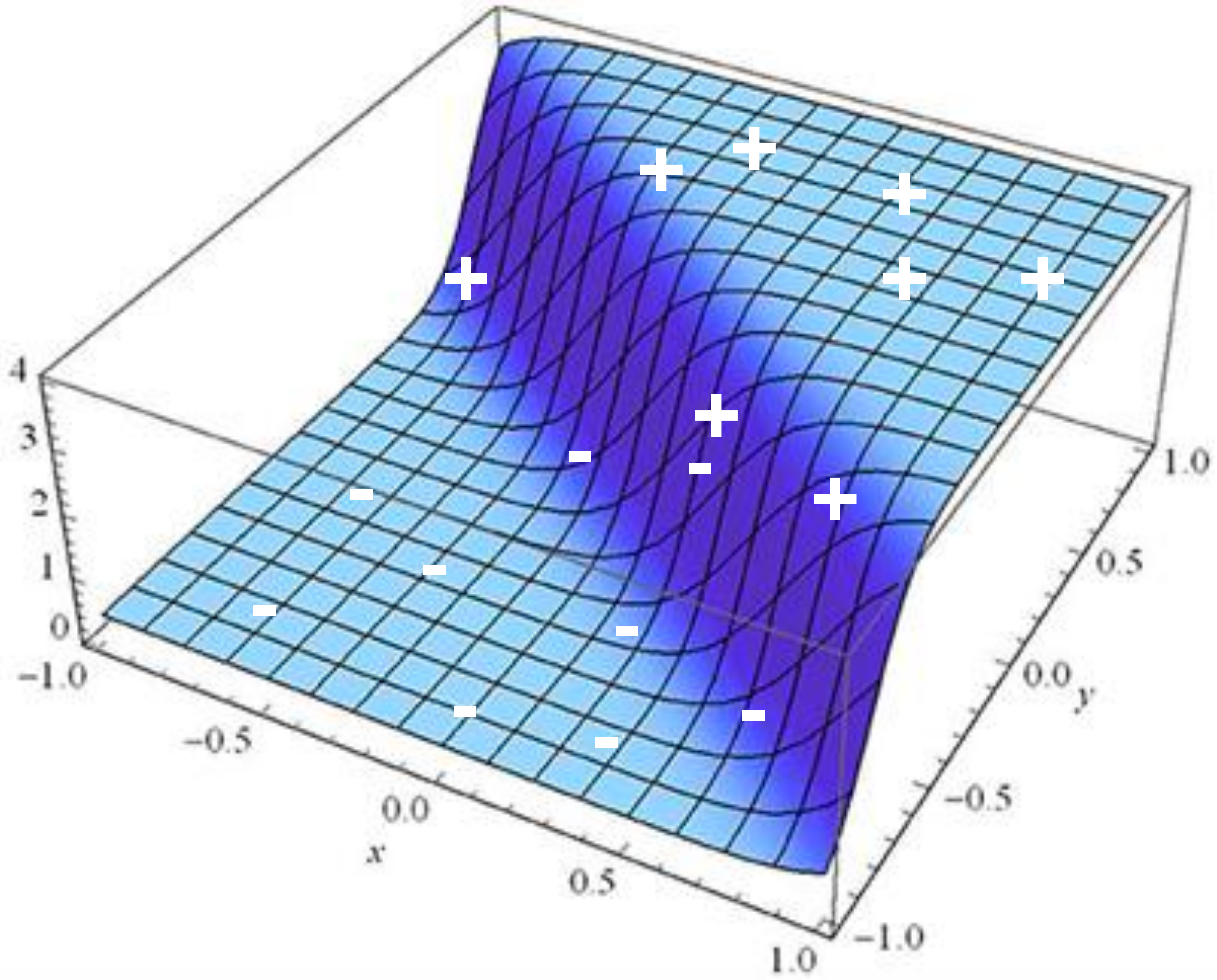
$$\sigma(z) = \frac{1}{1+e^{-z}}$$

$$h(x) = \sigma(\theta x^T)$$



1D case

Geometric View

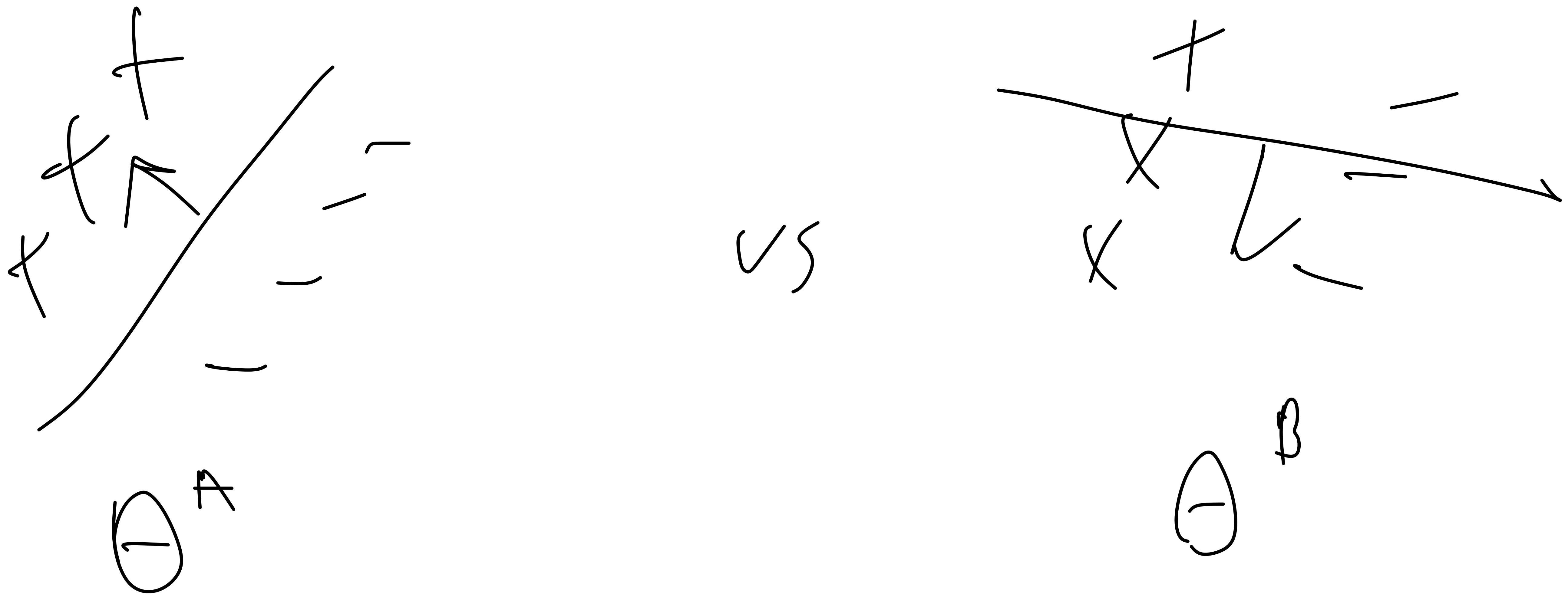


Empirical Risk Minimization

How do find a "good" h?

Empirical Risk Minimization

How do find a "good" h?



How good are the probs?

$$R(\theta) = \frac{1}{n} \sum_{i}^{N} L(y, p) = \frac{1}{n} \sum_{i}^{N} L(y, \sigma(\theta x^T))$$

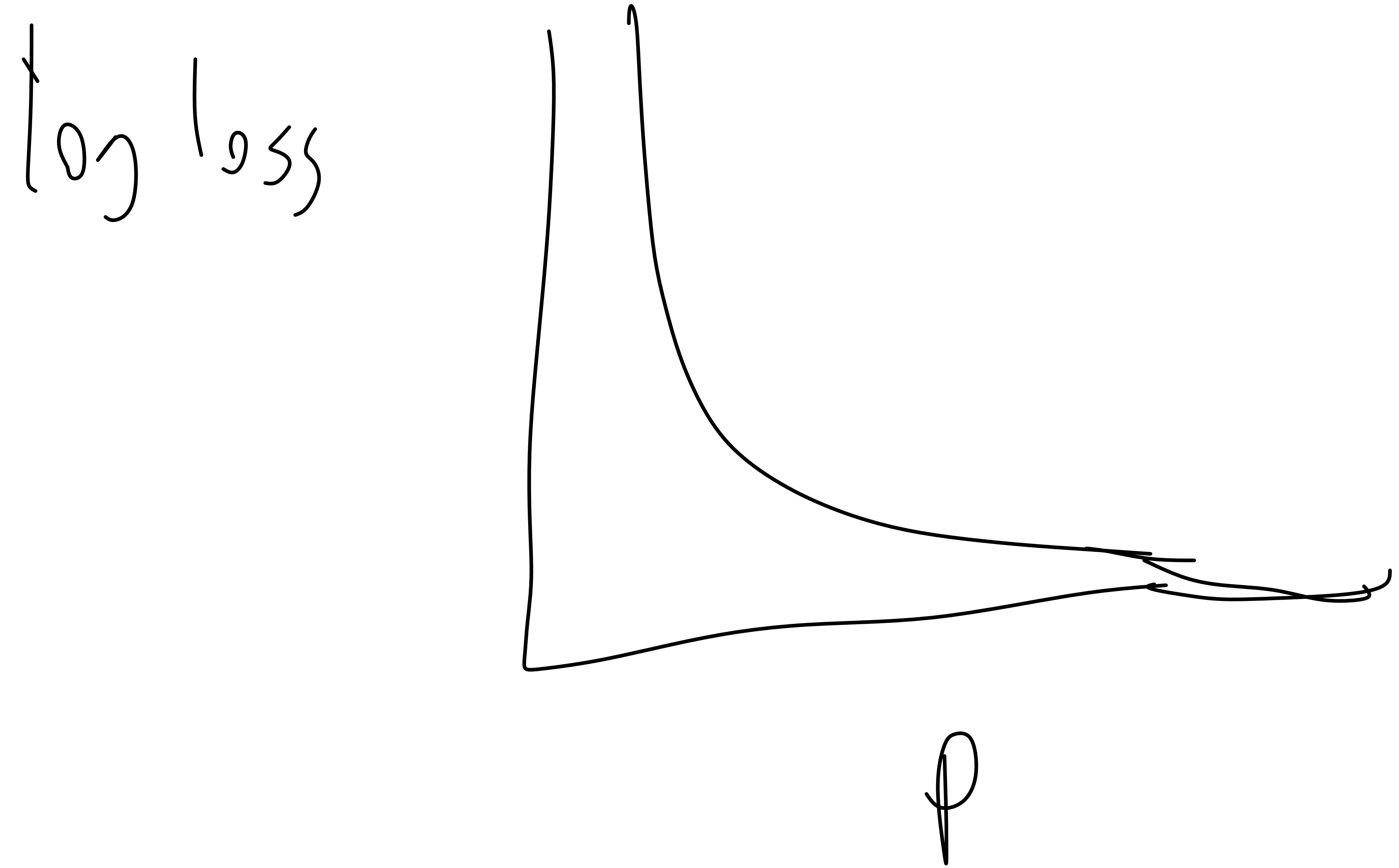
Loss Function: Cross Entropy

Loss Function: Cross Entropy

Likelihood of observed data

$$L(y, p) = -\left(y \log p + (1-y) \log(1-p) \right)$$

$y = 1$



penalize very wrong probs.

Maximum Likelihood Estimation!

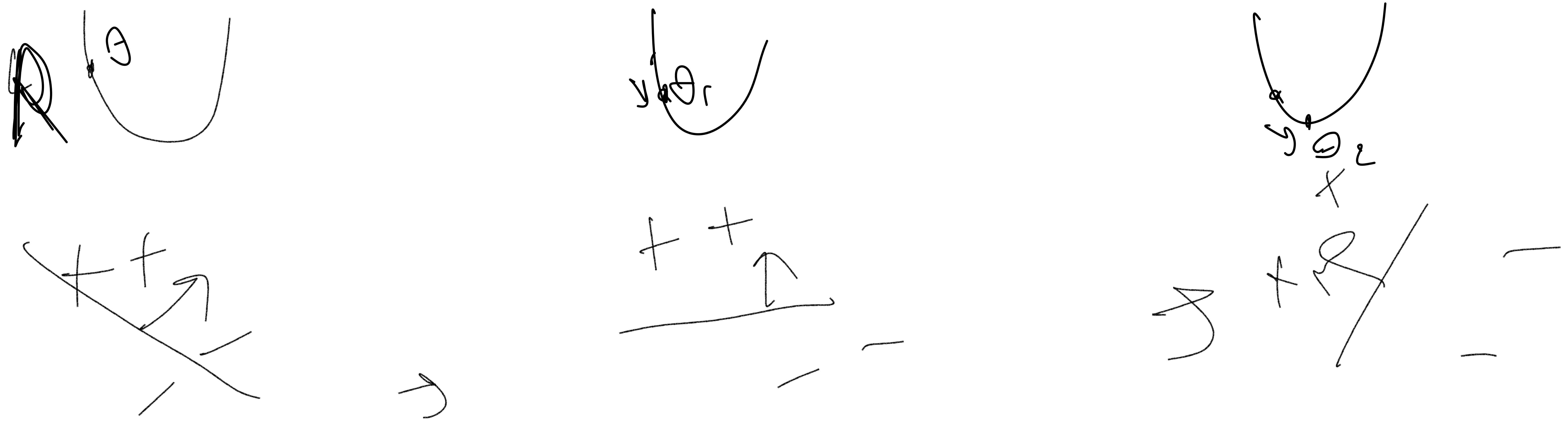
Optimization

How do we find a good h?

Optimization

How do we find a good h?

gradient descent!



θ_0 θ_1 θ_2

How? Gradient Descent

How? Gradient Descent

$$\theta_{new} = \theta_{old} - \eta \frac{\partial R(\theta_{old})}{\partial \theta_{old}}$$

η

learning rate

$$\frac{\partial R(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta}\left(\frac{1}{n}\sum_{i}^{N} L(y, p)\right) = \frac{1}{n}\sum^{N} \frac{\partial L}{\partial \theta}(y_i, p_i)$$

$$= \frac{1}{n}\sum^{n} (p_i - y_i)\, x_i$$

$$\frac{\partial L}{\partial \theta}(y, p) = -\frac{\partial}{\partial \theta}\left(\right)$$

$$\frac{\partial L}{\partial \theta}(y, p) = -\frac{\partial}{\partial \theta}\left(y \log \underbrace{\sigma(\theta x^T)}_{p} + (1-y)\log(1-\sigma(\theta x^T)) \right)$$

$$= -\left(\frac{y\,\cancel{(1-p)\,p\,x^T}}{\cancel{p}} + \frac{(1-y)\,\cancel{p\,(1-p)}(x)}{\cancel{1-p}} \right)$$

$$-\left(yx - \cancel{py}x - px + \cancel{py}x \right)$$

$$= -(y-p)x \qquad = (p-y)x$$

Putting it all together

$$\sigma(\theta x_i^T)$$

$$\theta_{new} = \theta_{old} - \eta \frac{\sum_i^N (p_i - y_i) x_i}{N}$$

Putting it all together

$$\theta_i x_i^T$$

$$\theta_{new} = \theta_{old} - \eta \frac{\sum_i^N (p_i - y_i) x_i}{N}$$

θ_0 init rand

while not converged:

$$\theta_{t+1} = \theta_t - \eta \frac{\partial L}{\partial \theta_t}$$

return θ_{end}

What if N is too large?

What if N is too large?

Stochastic Gradient Descent

Estimate emporical risk R
with B random samples

$\theta_i x_i^T$

$$\theta_{new} = \theta_{old} - \eta \frac{\sum_i^B (p_i - y_i) x_i}{B}$$

θ_0 init rand

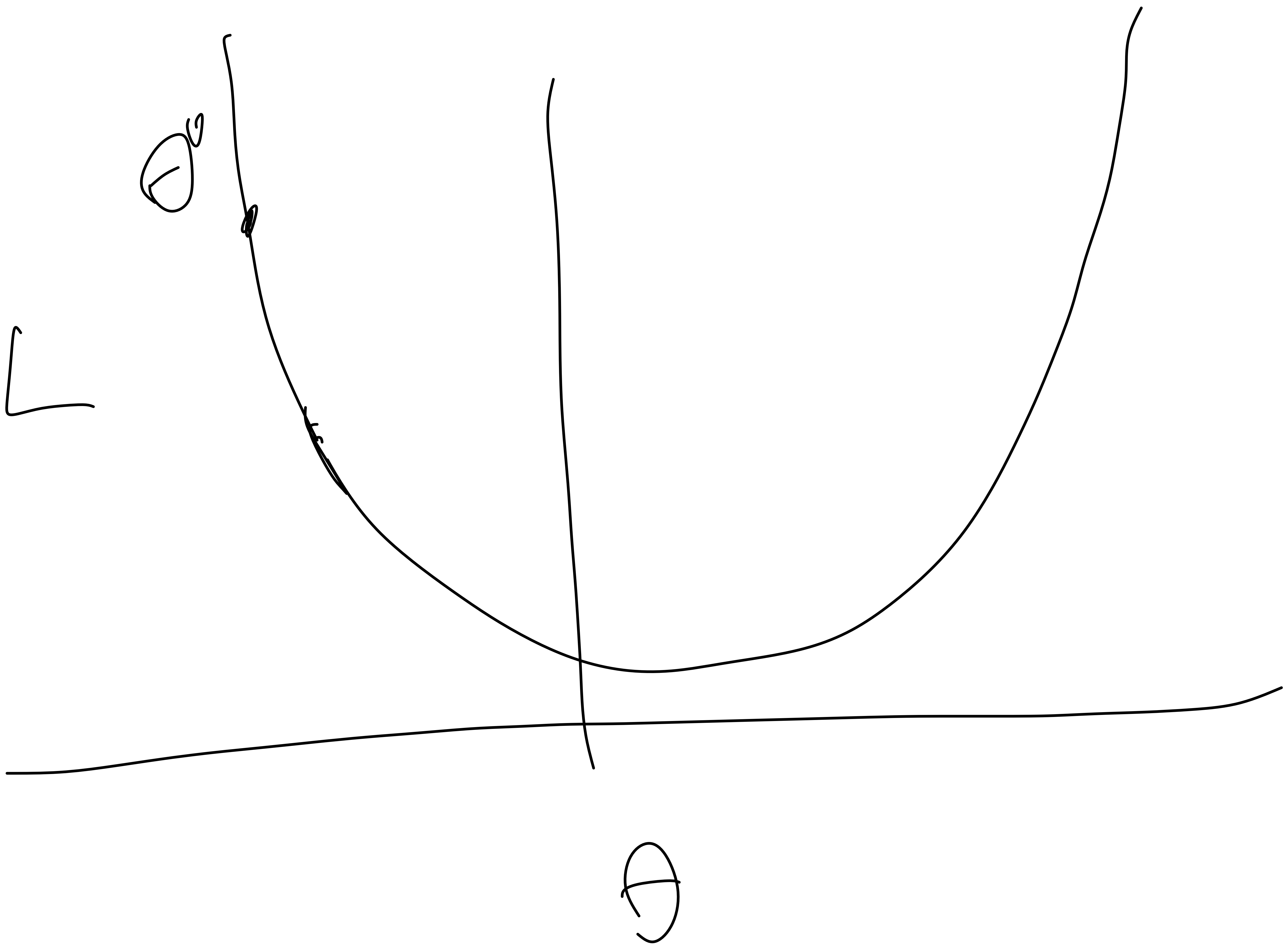
while not converged:

$$\theta_{t+1} = \theta_t - \eta \frac{\partial L}{\partial \theta_t}$$

return θ_{end}

Choosing your learning rate

What if LR is too small?



What if LR is too large?

How can we tell?

Choosing your learning rate

What if LR is too small?

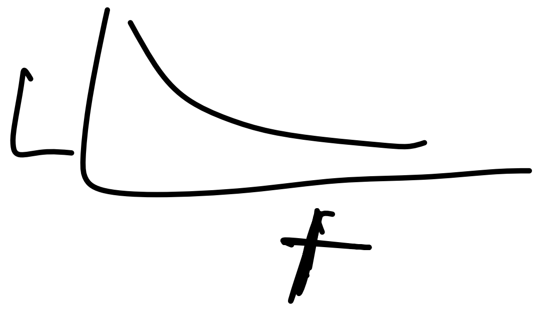
Slow optim

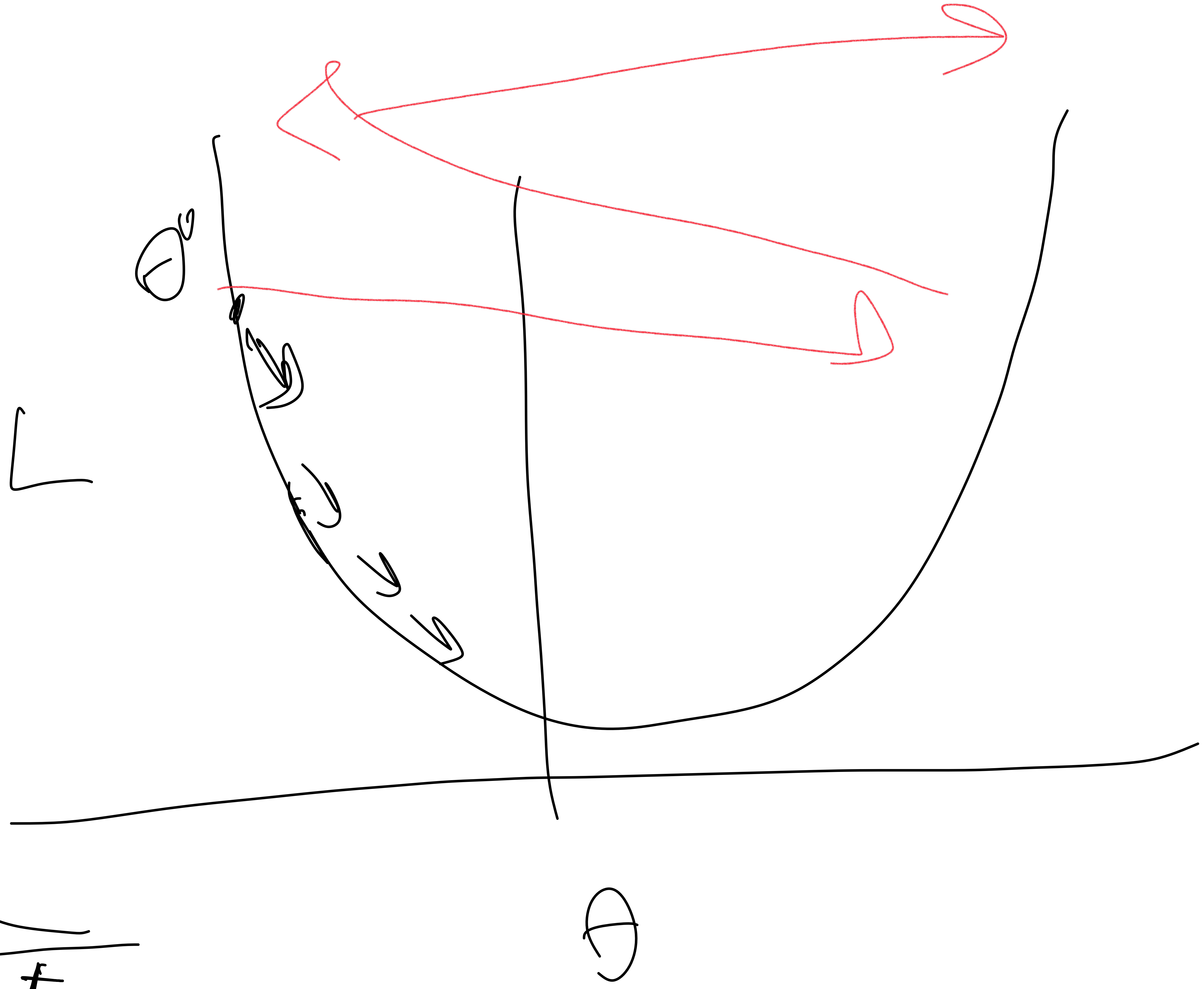
What if LR is too large?

Diverge

How can we tell?

Training Curve

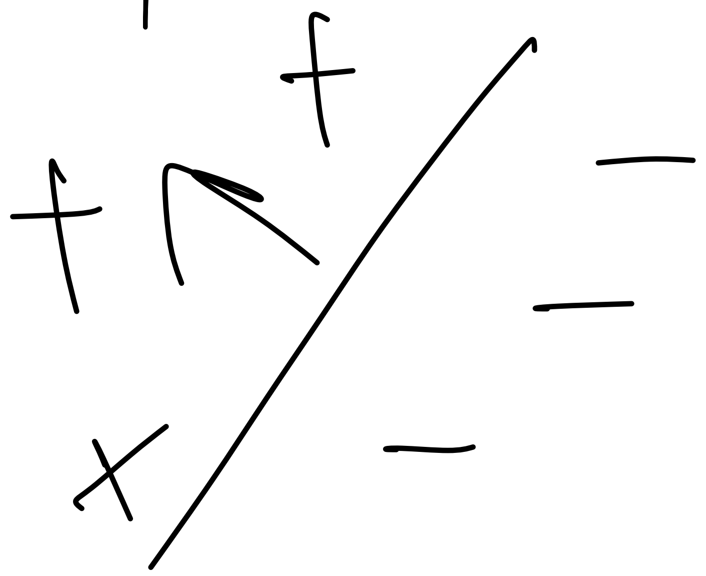




Recap

Recap

Now we have a model h w param θ

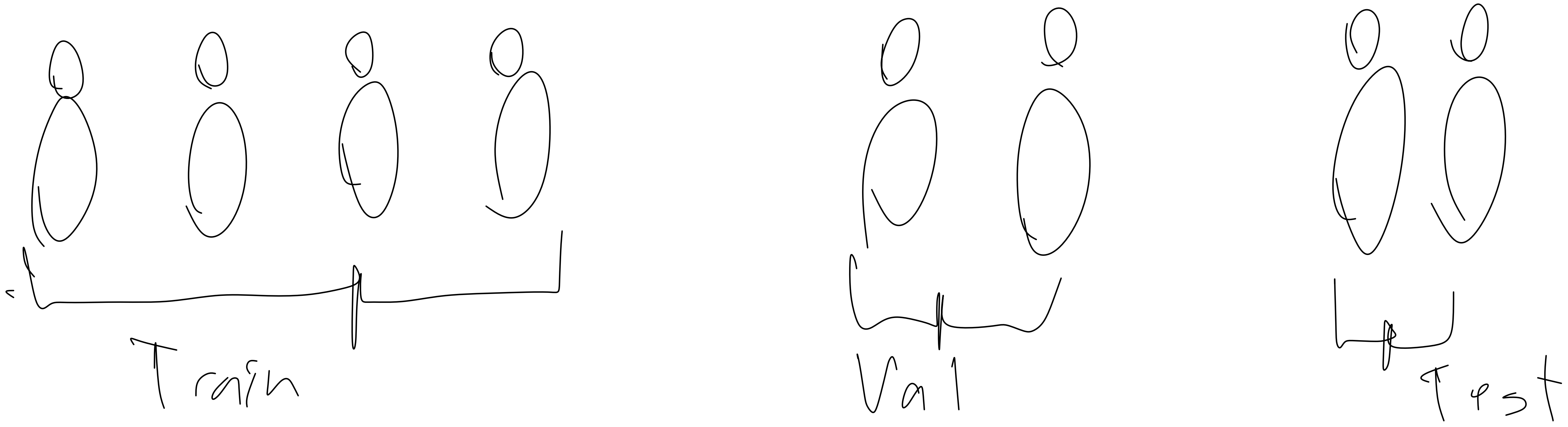


$$\theta = \arg\min_{\theta} R(\theta) = \frac{\sum_{i}^{n} L(p_i, y_i)}{n}$$

is this enough?

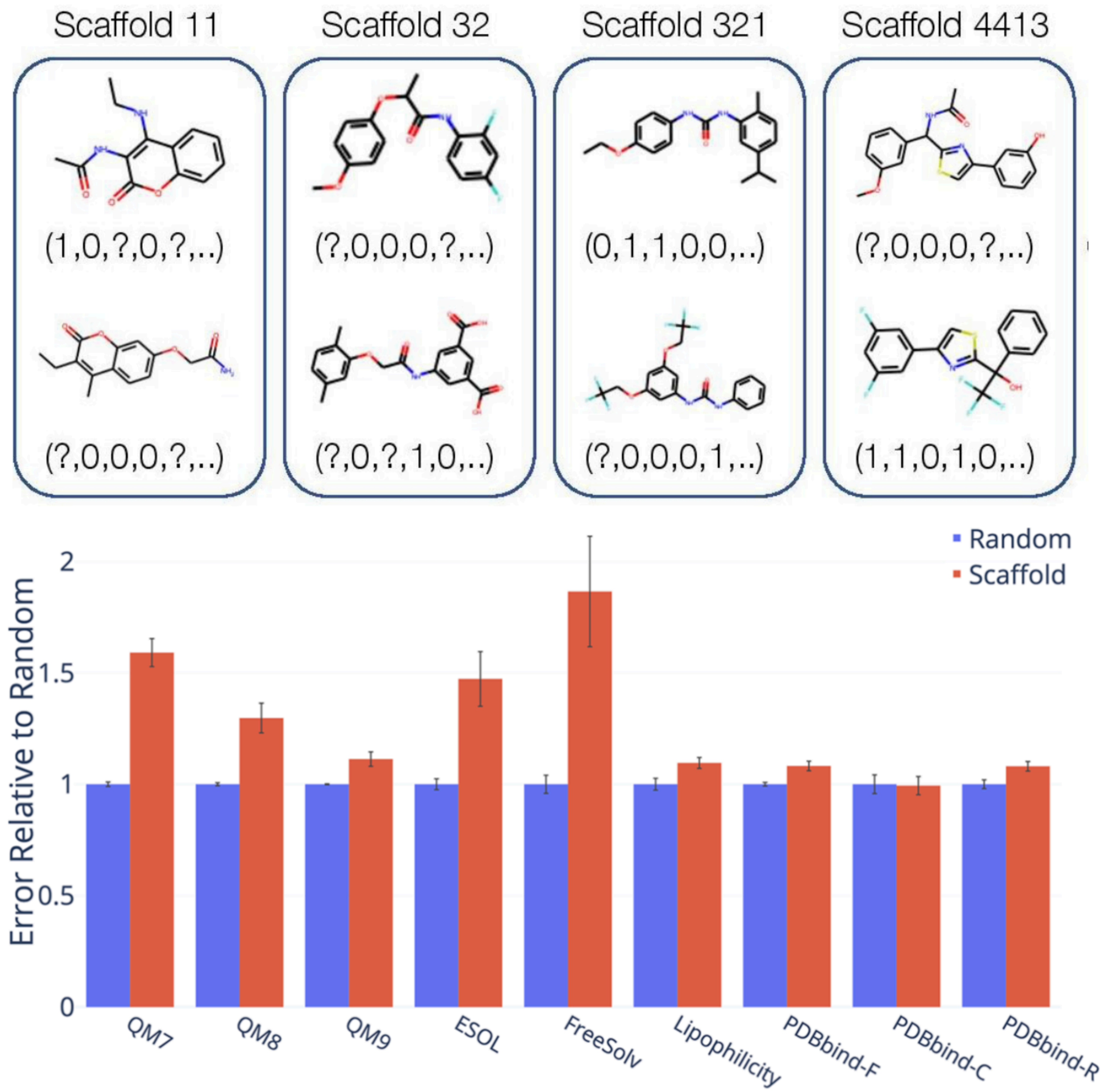
No

We want h to do well on NEW patients ...



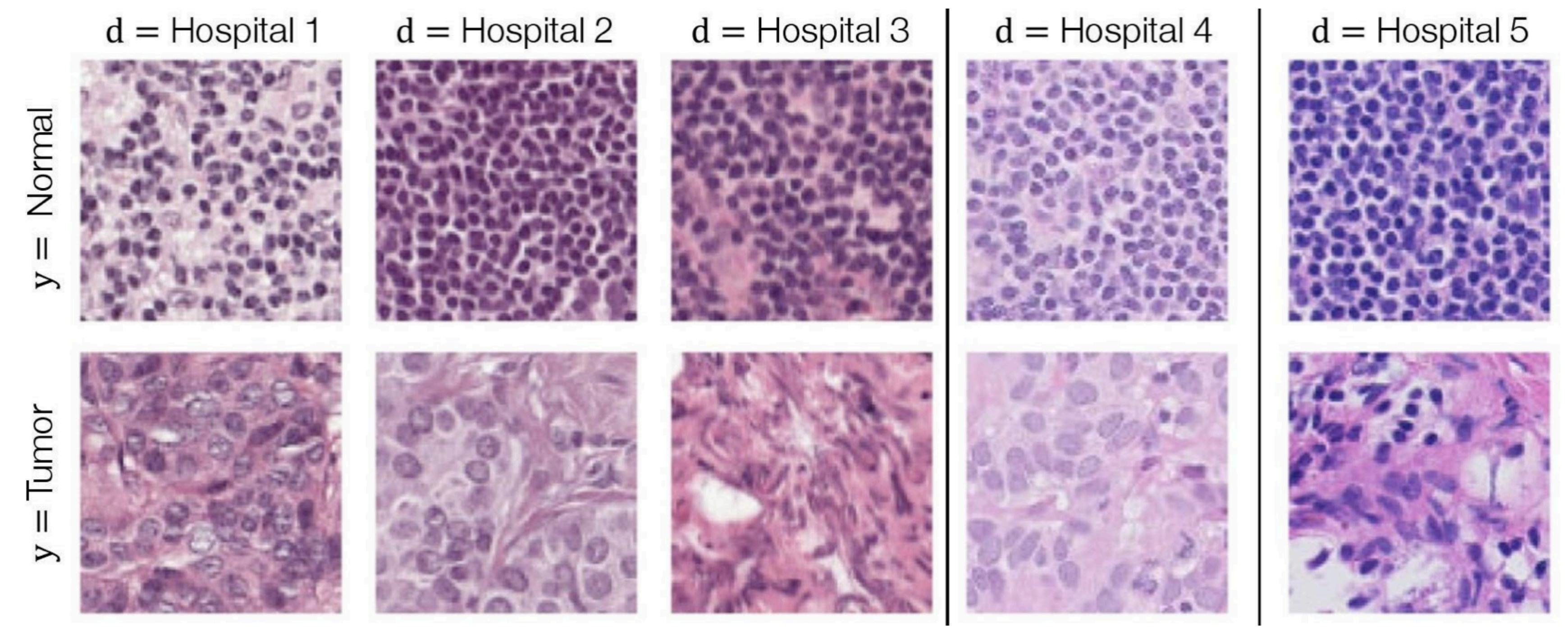
The importance of data splitting

Scaffold split in property prediction



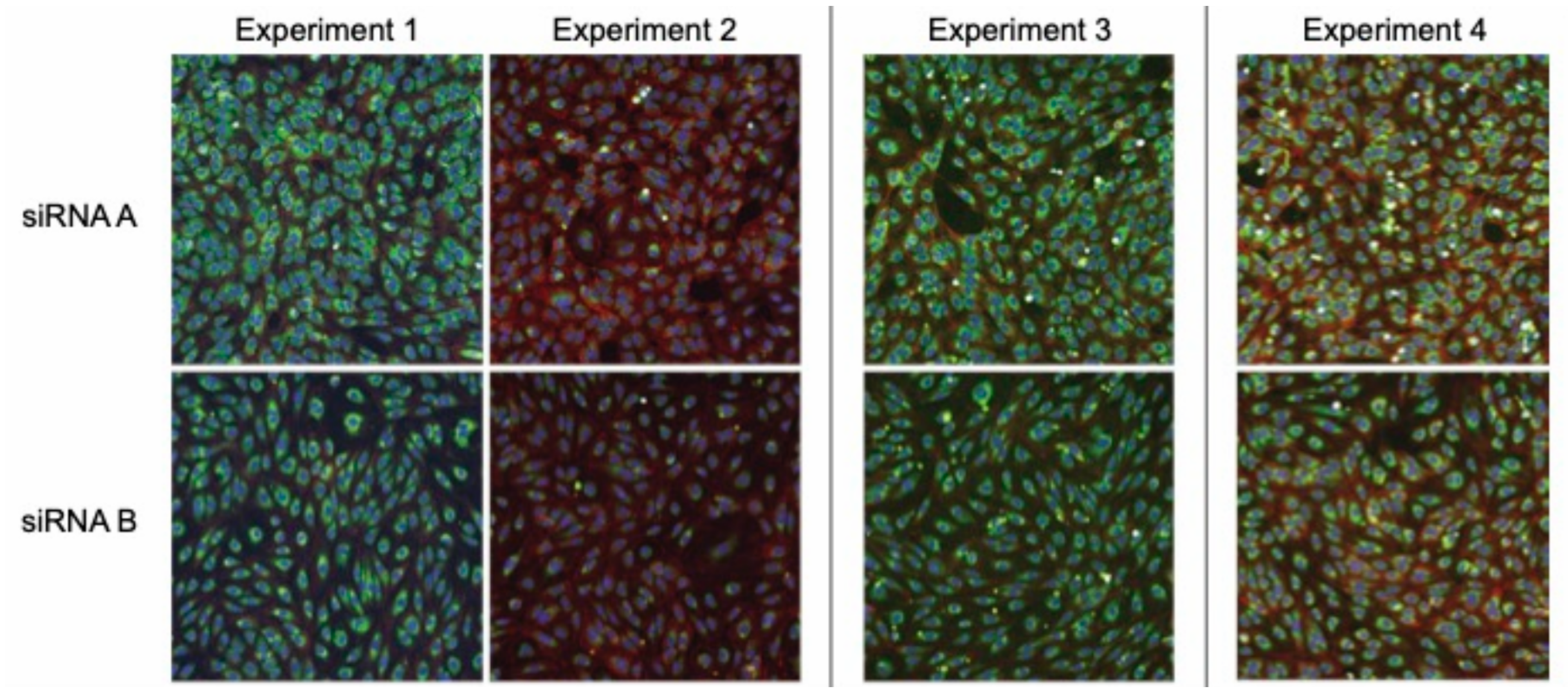
Yang, Kevin, et al. "Analyzing learned molecular representations for property prediction." *Journal of chemical information and modeling* 59.8 (2019): 3370-3388.

Hospital source in pathology



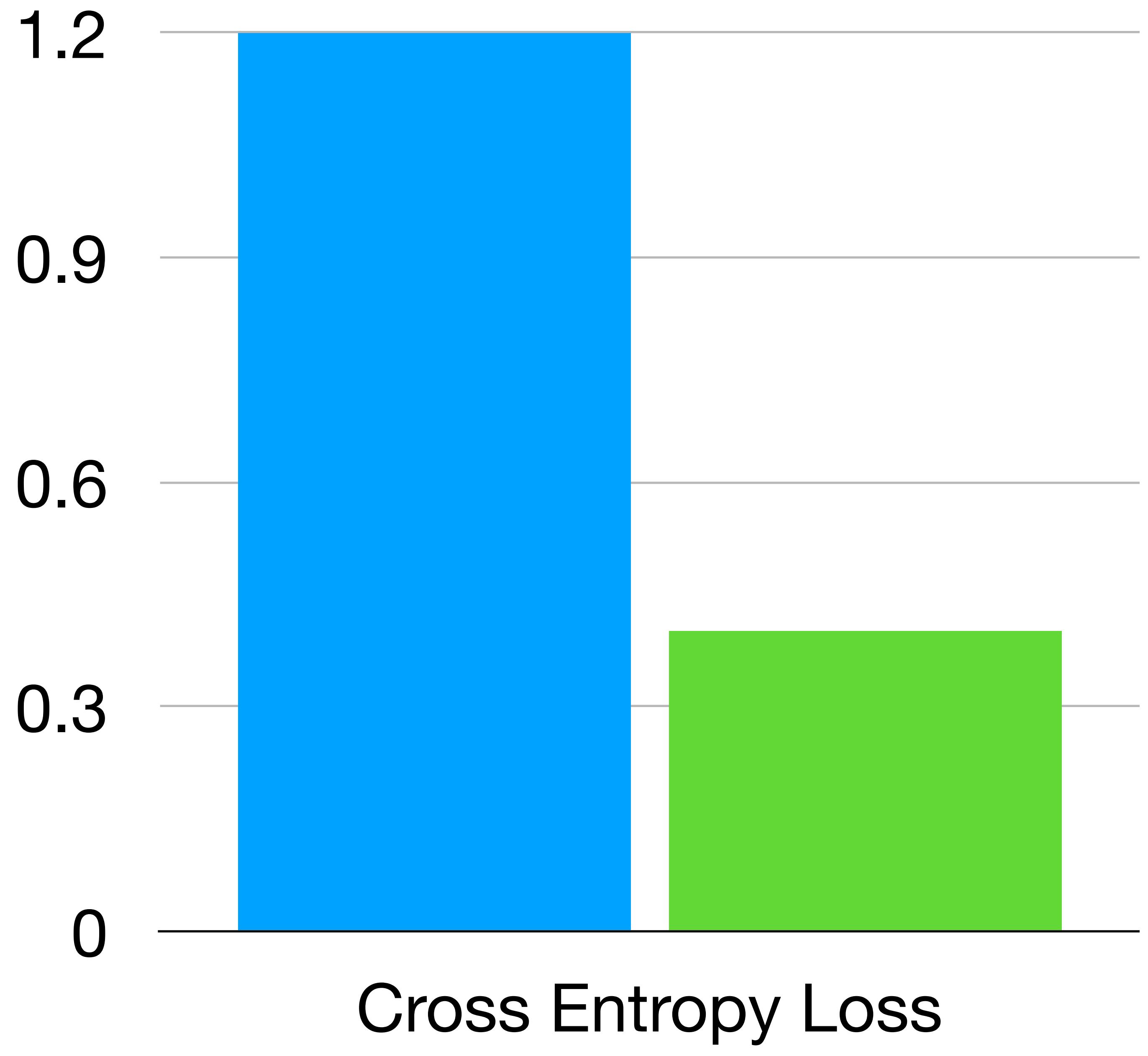
Bandi, Peter, et al. "From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge." *IEEE transactions on medical imaging* 38.2 (2018): 550-560.

Batch effects in high-throughput screening



Taylor, J., et al. "RxRx1: An Image Set for Cellular Morphological Variation Across Many Experimental Batches." *The 7th International Conference on Learning Representations*. 2019.

Model Evaluation



Modeling objective

Achievable performance

$1: h(x) \geq p$

$0: h(x) < p$

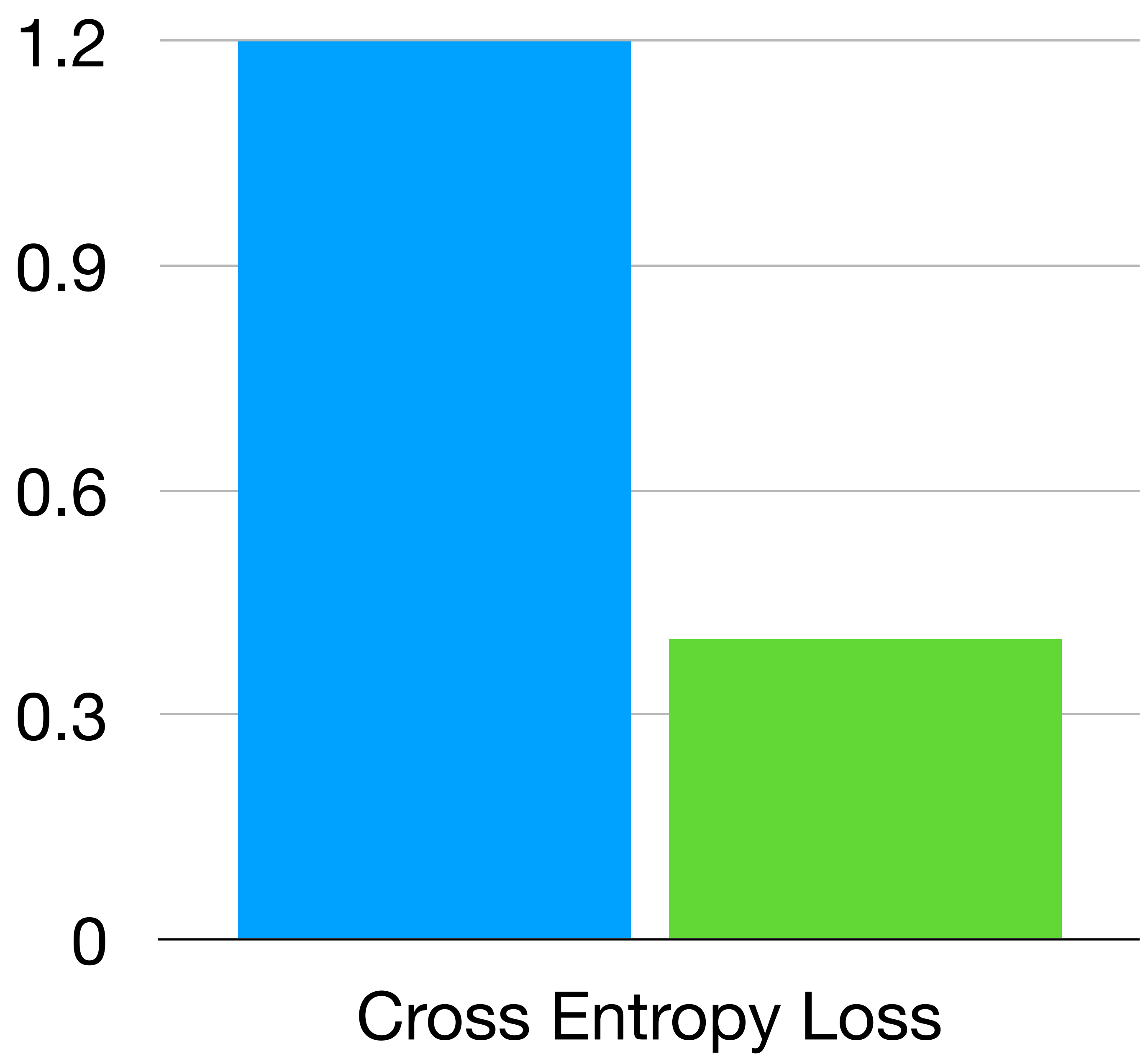
$TPR = \frac{TP}{\#+}$

$FPR = \frac{FP}{\#-}$

$AUC: P(p_i > p_j \mid y_i = 1, y_j = 0)$

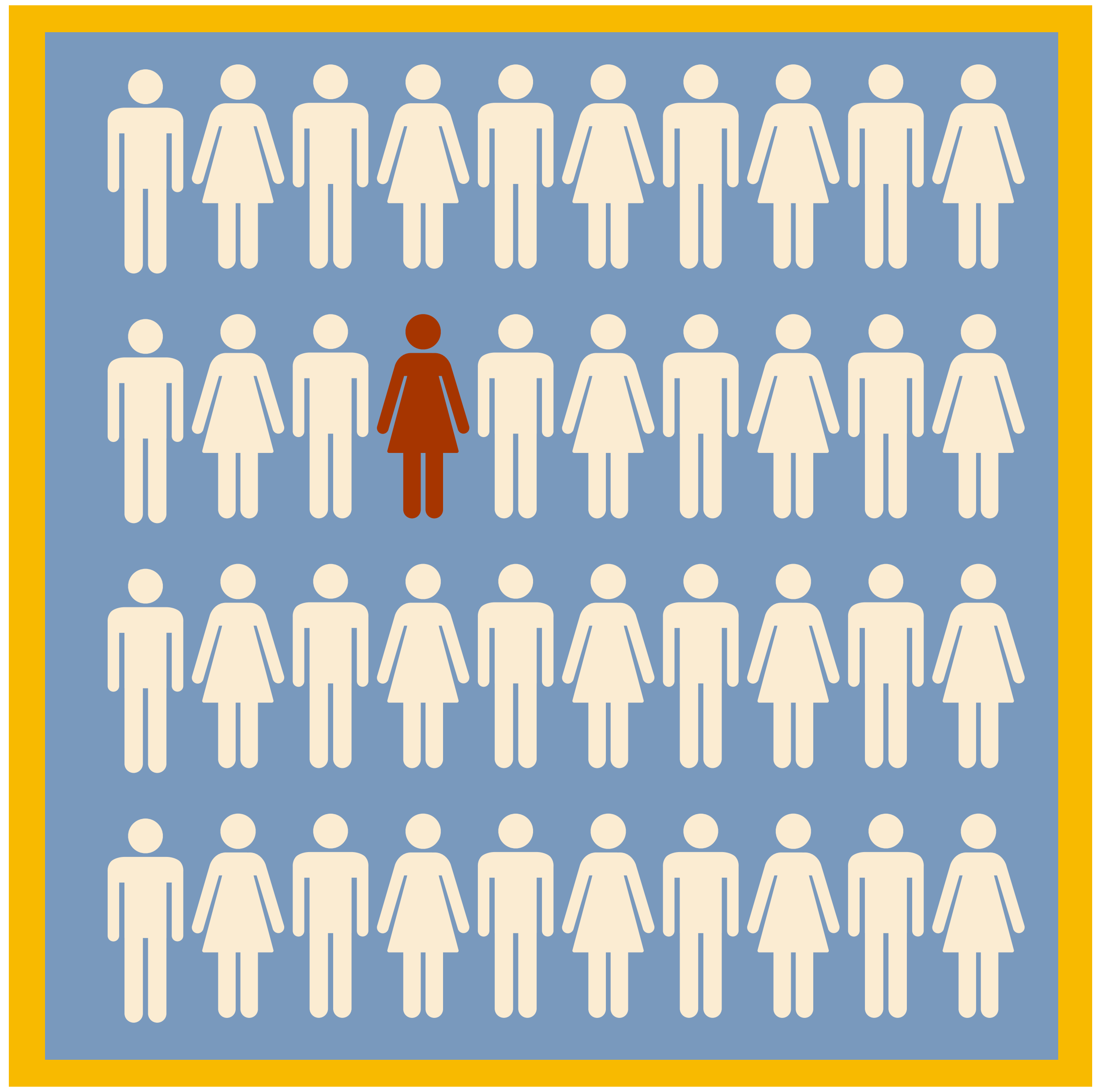
Computational
PRECISION HEALTH

Model Evaluation



Modeling objective

Achievable performance



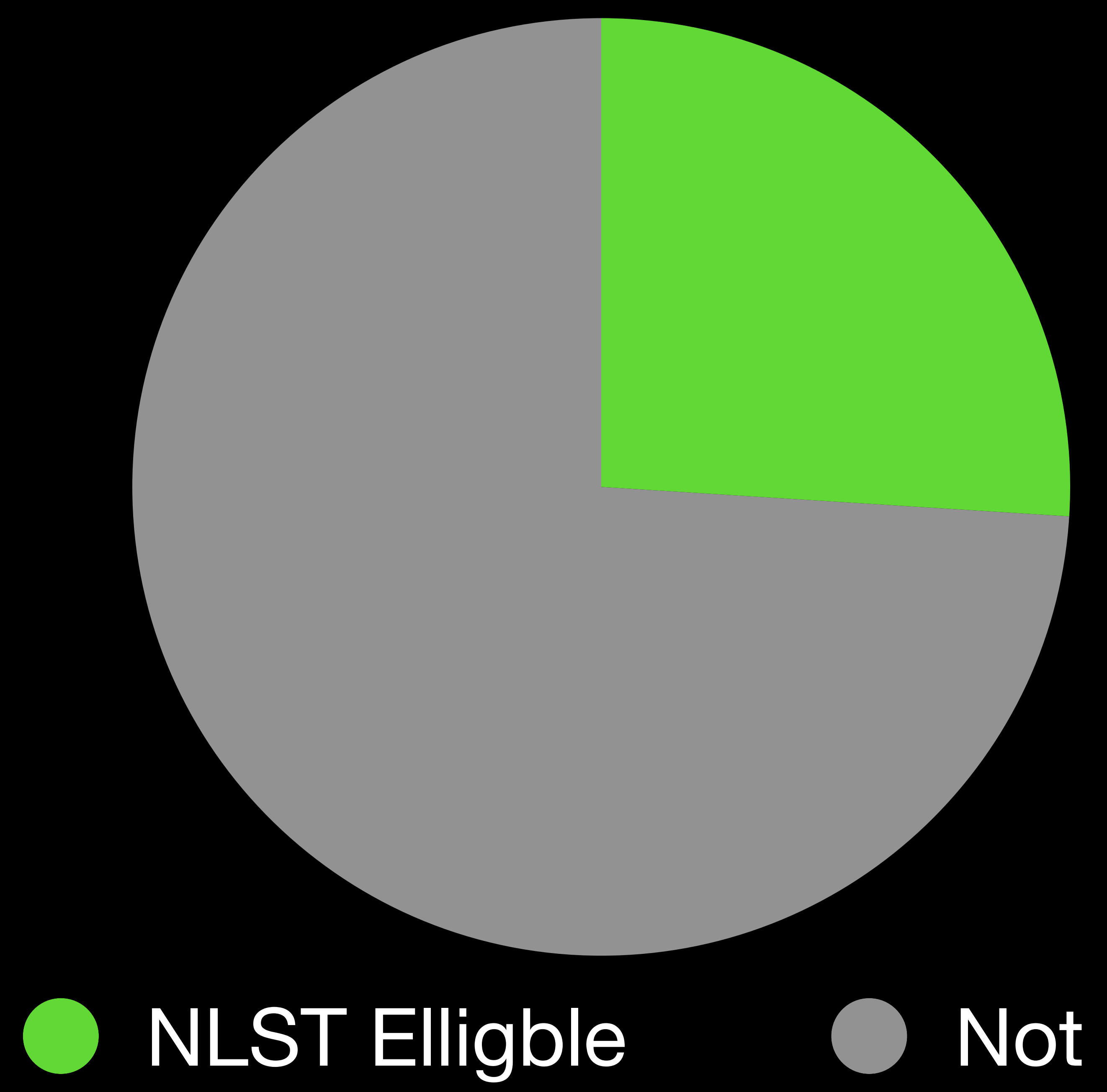
Simulated clinical utility

Efficacy of a screening program

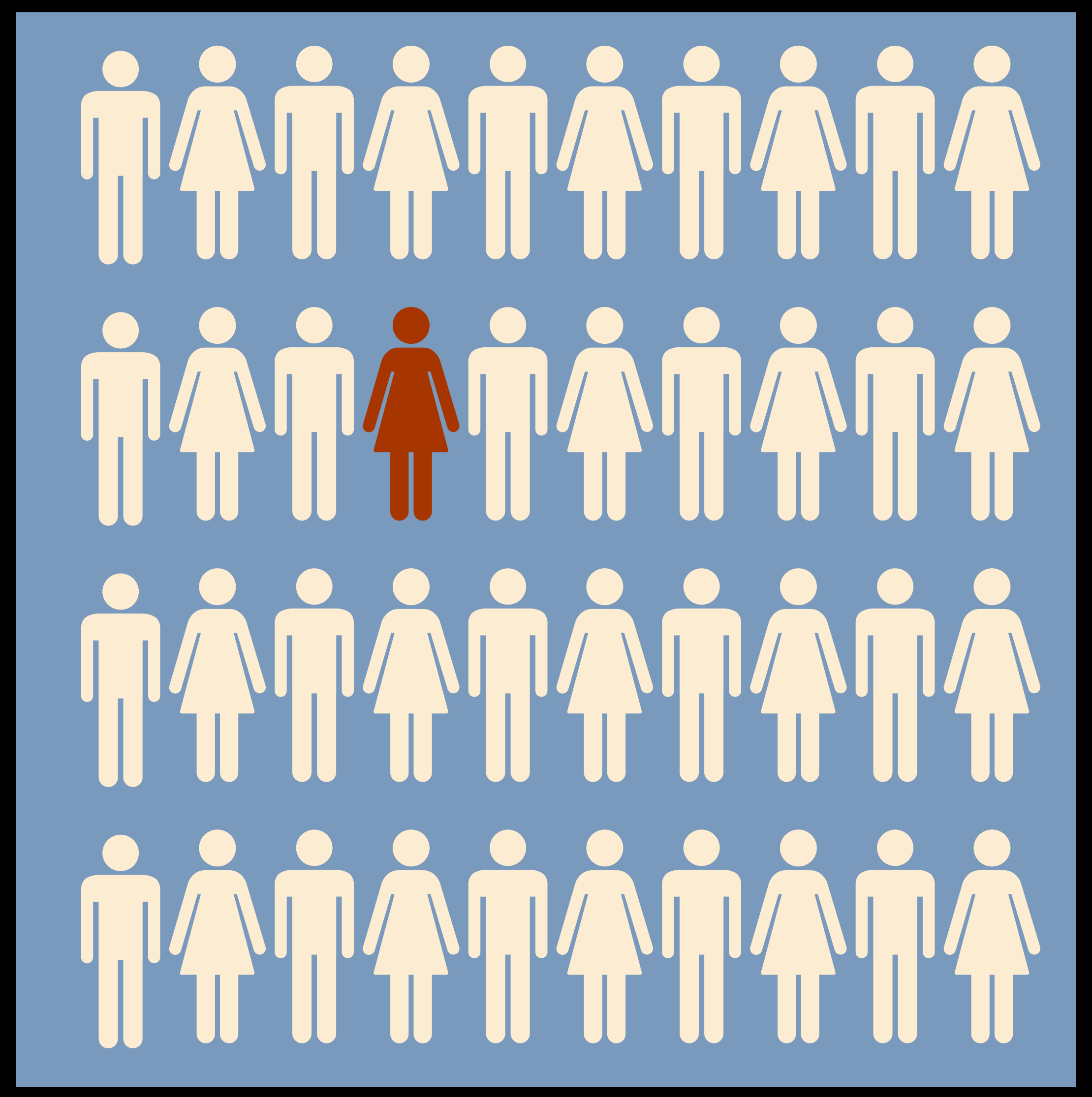
Fundamental challenge is **cost-effectiveness**

How much harm does the program do?

How much benefit does it achieve?



PMID: 23060474



1000 screens

↓

240 positives

↓

6 cancers

Summary

All screening programs are classifiers

Effective screening programs need risk models to allocate care

Logistic Regression: Log-linear hypothesis class

Optimization: (Stochastic) Gradient Descent

Model selection and evaluation

Questions?