CPH 100A: Machine Learning Foundations II

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

Computational
PRECISION HEALTH

Berkeley | UCSF

Agenda

Recap

Feature Engineering and Regularization

Normalization and Optimization

Beyond Classification tasks: Regression and Survival Modeling

Recap: Reflecting on Ida's lecture

How can we do better for these patients?

Recap: Screening reduces 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.

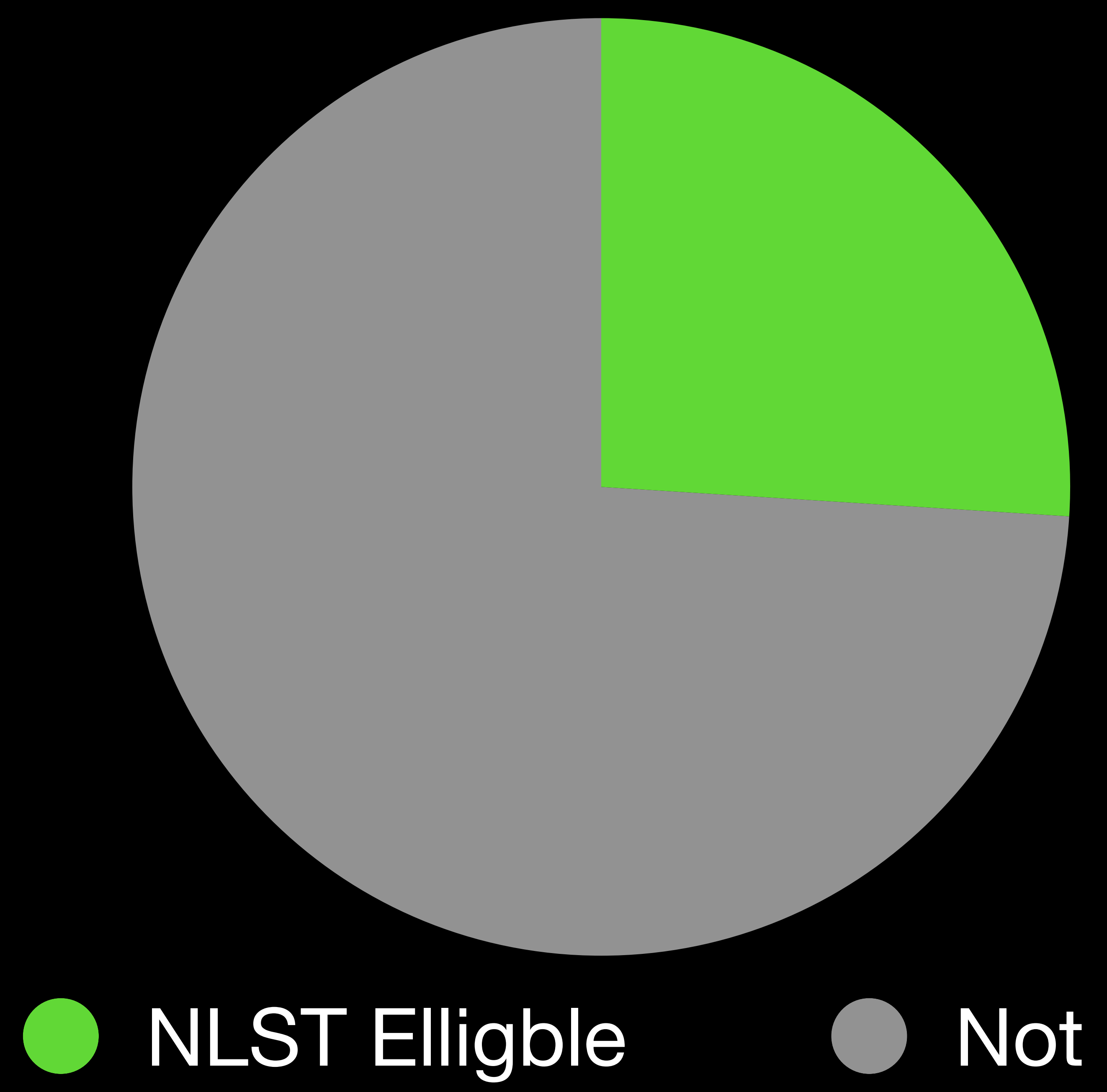
NELSON reduces lung cancer mortality by 24%

Recap: Efficacy of a screening program

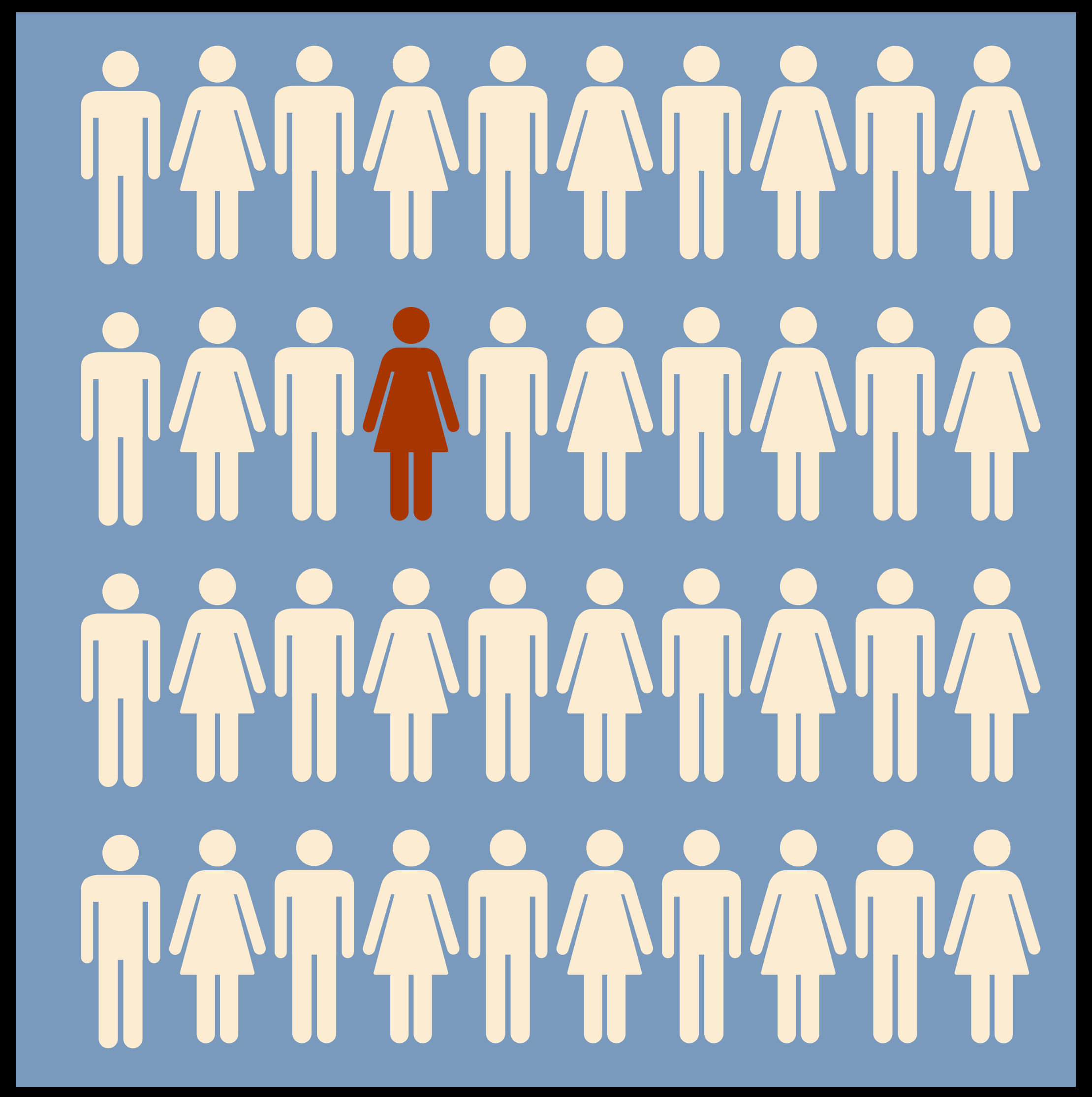
Fundamental challenge is **cost-effectiveness**

How much harm does the program do?

How much benefit does it achieve?



PMID: 23060474



1000 screens

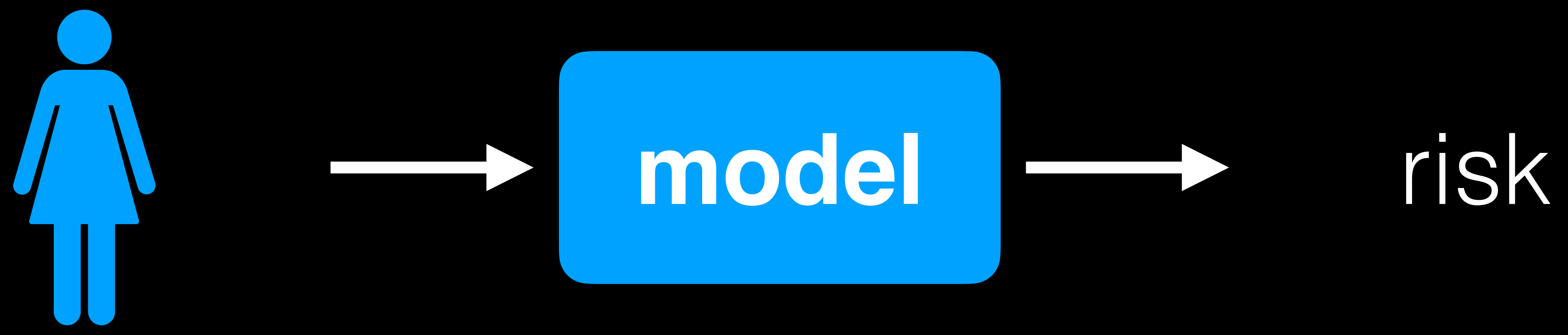
↓

240 positives

↓

6 cancers

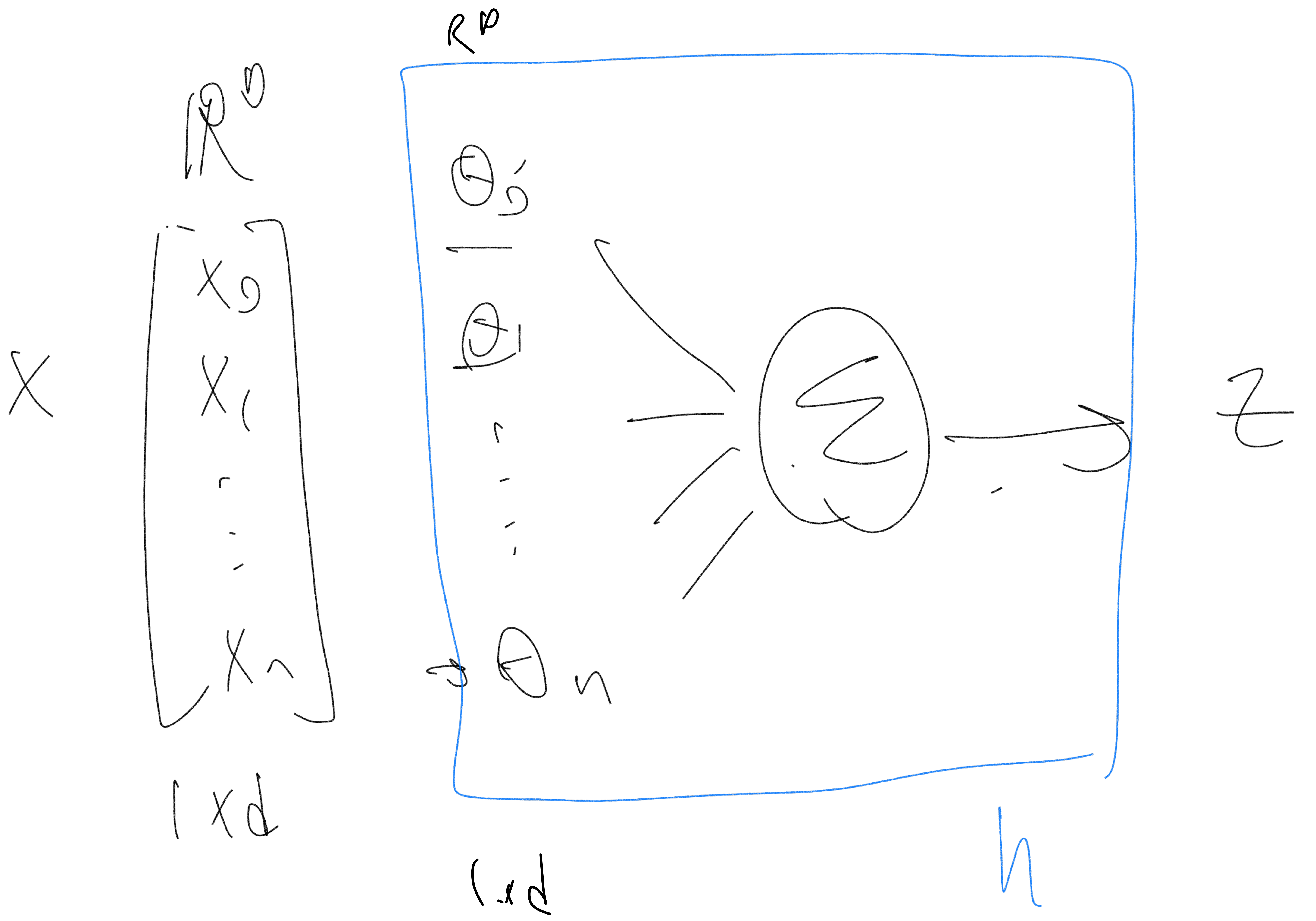
Recap: Can we do better?



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

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

Recap: Loglinear models



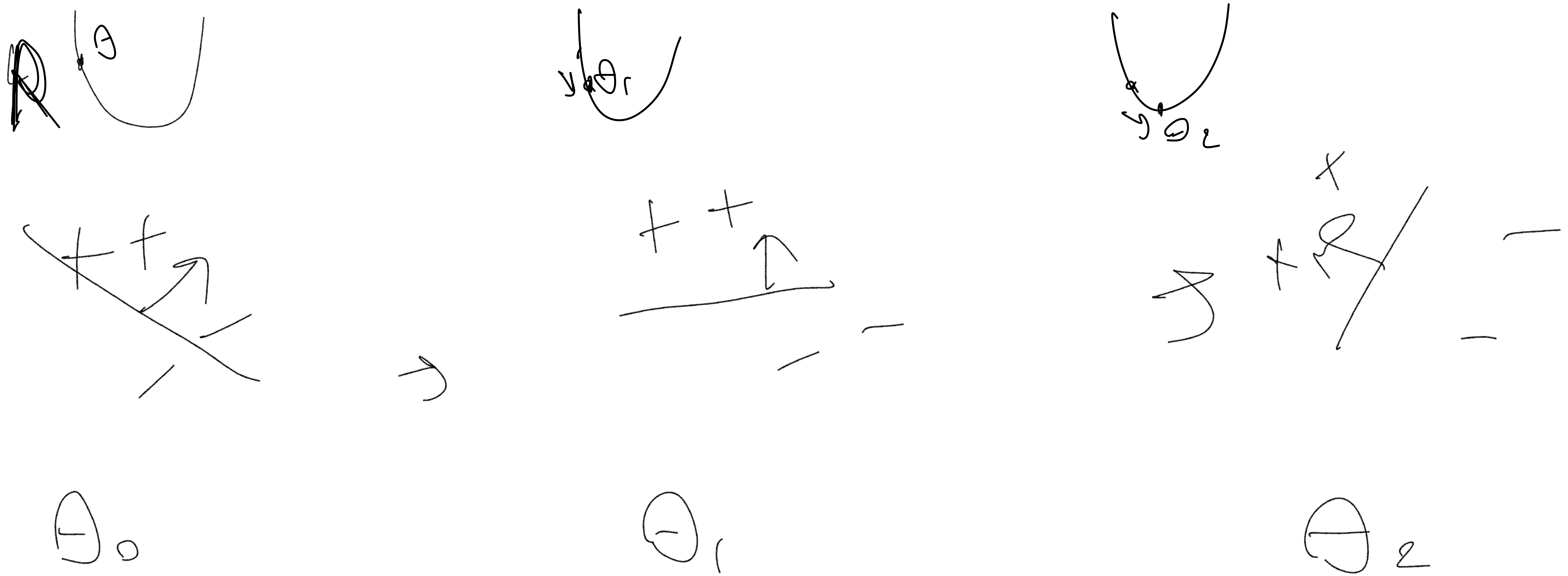
$$z = \sum_i^N \theta_i X_i$$

$$z = \theta X^T$$

quality of h: $R(\theta) = \sum \frac{L(y_i, p_i)}{n} = \sum \frac{\left(y_i \log p_i + (1-y_i) \log(1-p_i)\right)}{n}$

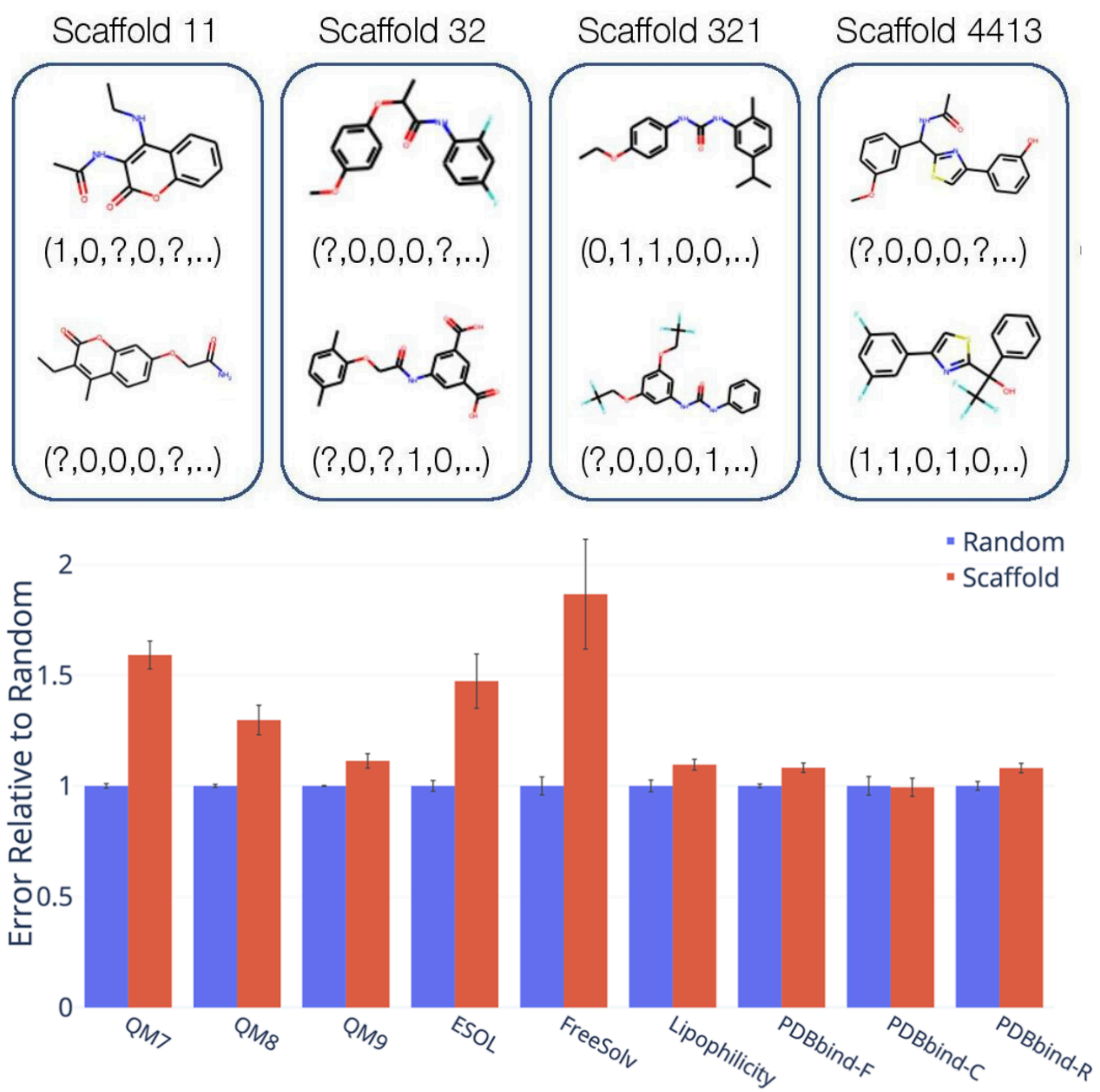
How do we find a good h?

gradient descent!



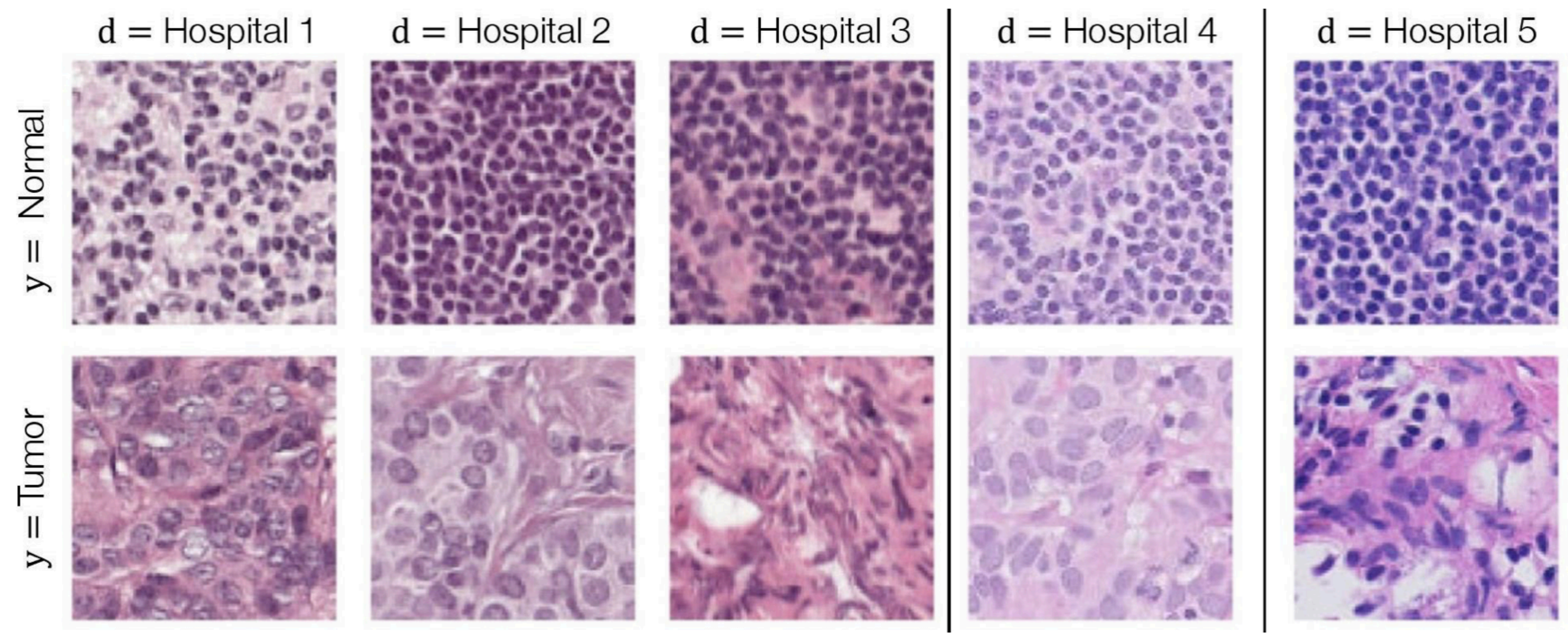
Generalization to what?

Scaffold split in property prediction



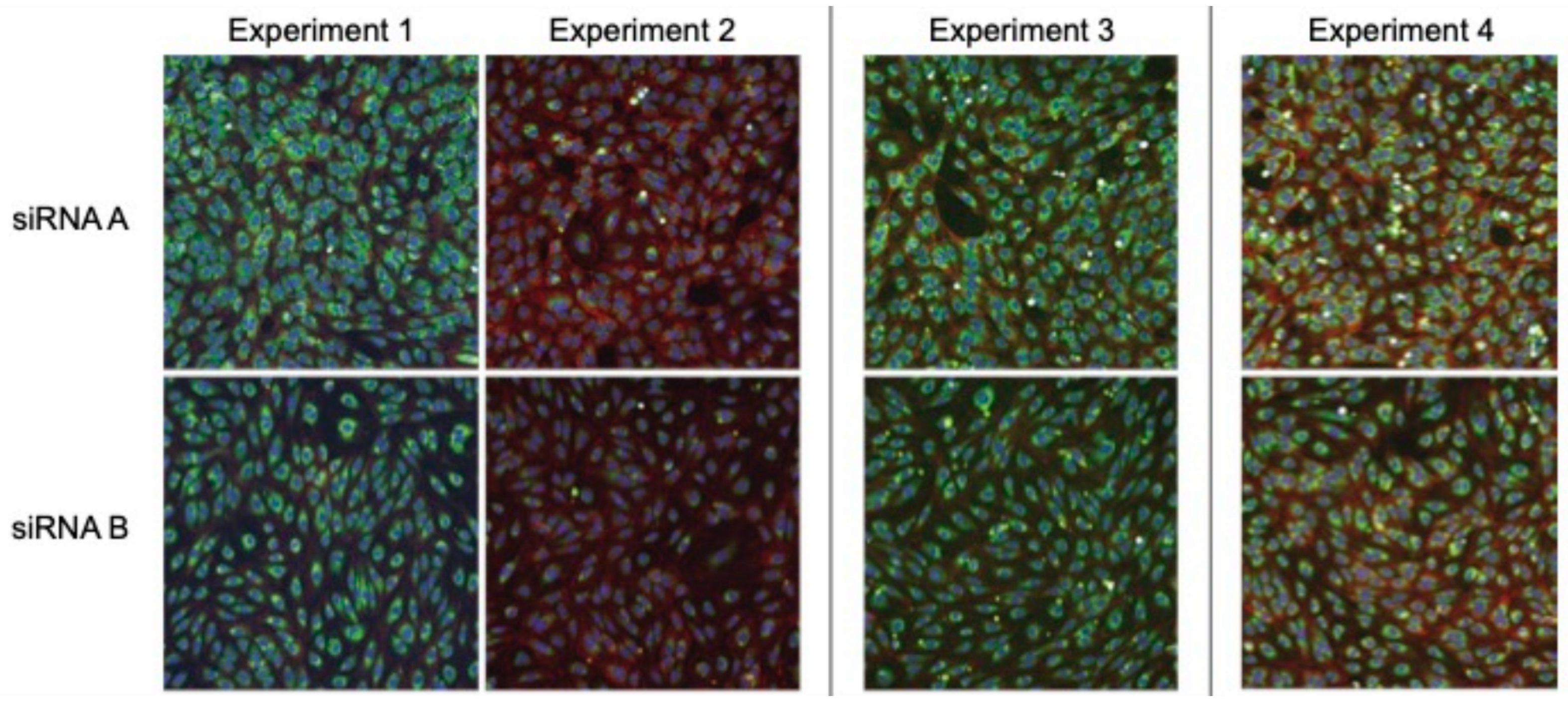
Yang, Kevin, et al. "Analyzing learned molecular representations for property prediction." *Journal of chemical information and modeling* 59.8 (2019): 3370-3388.

Hospital source in pathology



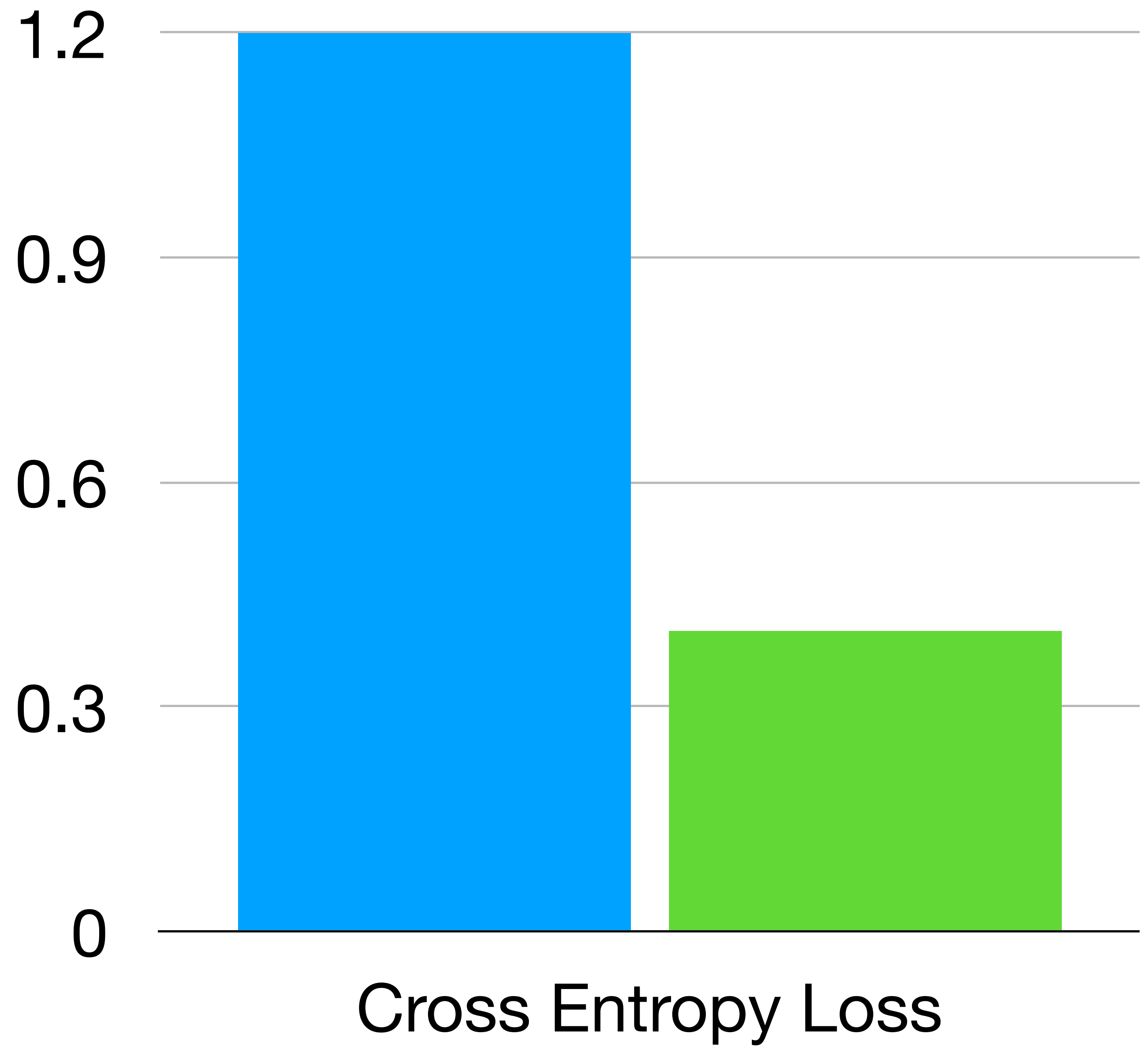
Bandi, Peter, et al. "From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge." *IEEE transactions on medical imaging* 38.2 (2018): 550-560.

Batch effects in high-throughput screening



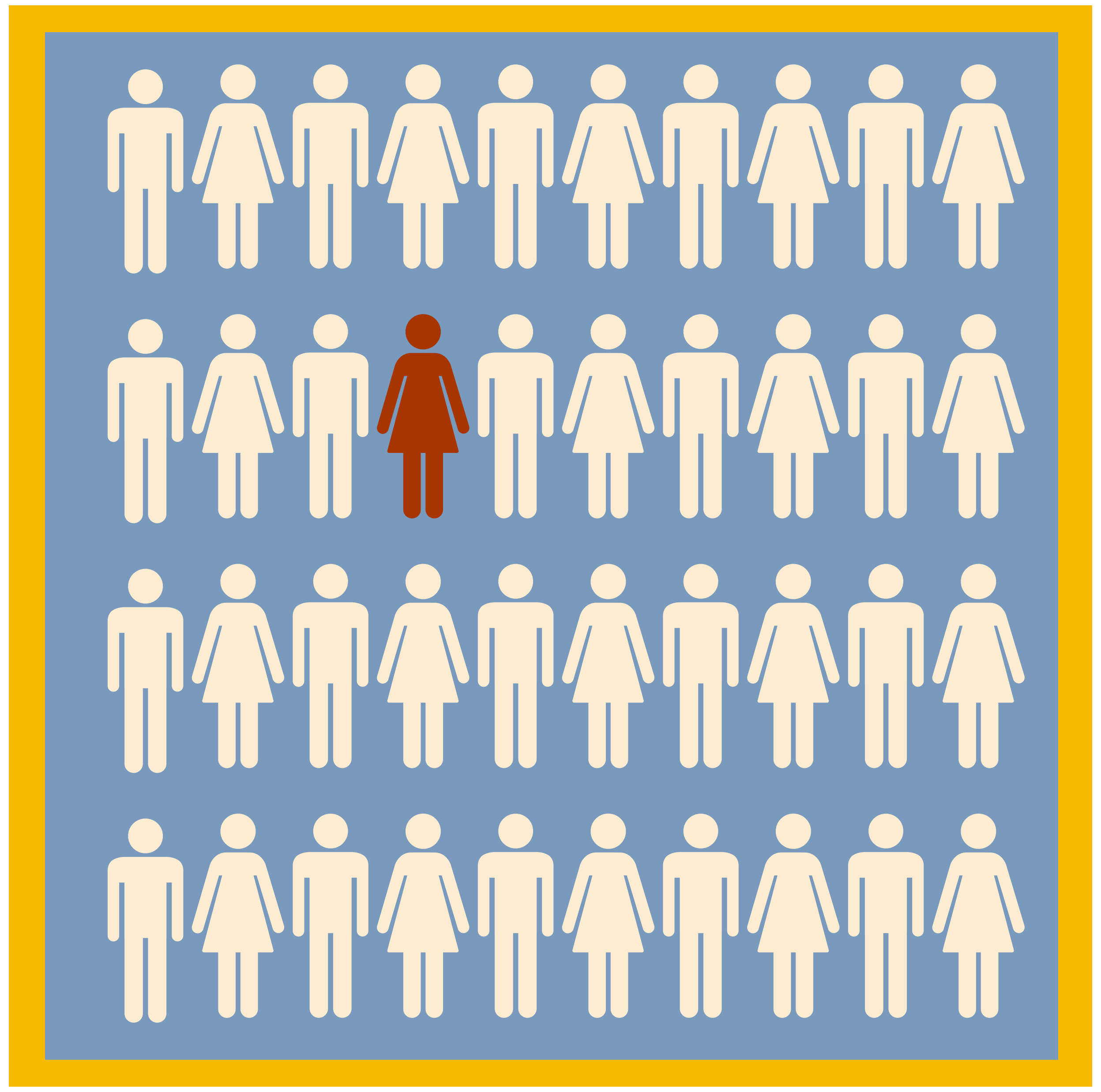
Taylor, J., et al. "RxRx1: An Image Set for Cellular Morphological Variation Across Many Experimental Batches." *The 7th International Conference on Learning Representations*. 2019.

Model Evaluation



Modeling objective

Achievable performance



Simulated clinical utility

Agenda

Recap

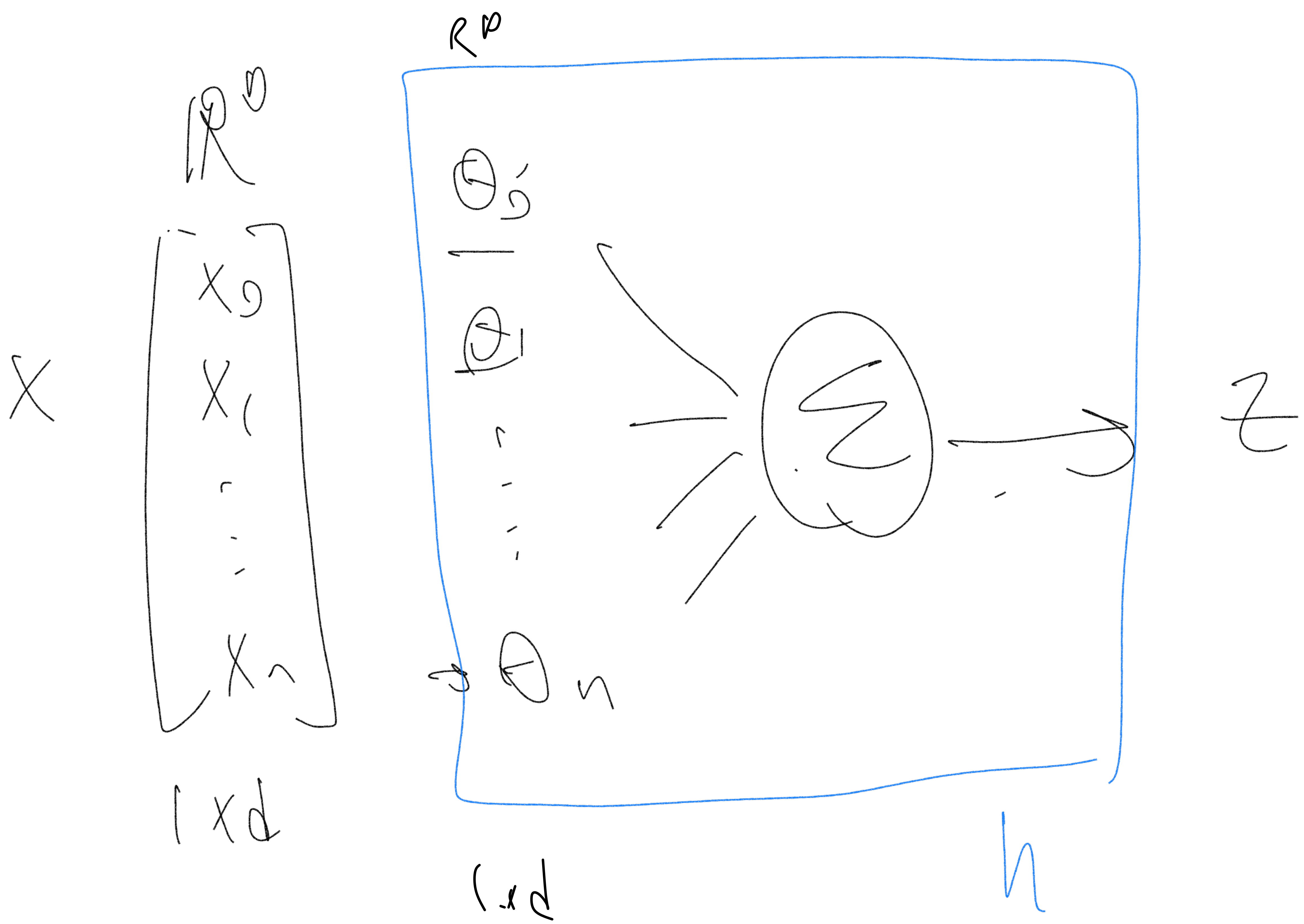
Feature Engineering and Regularization: Where the rubber hits the road

Normalization and Optimization

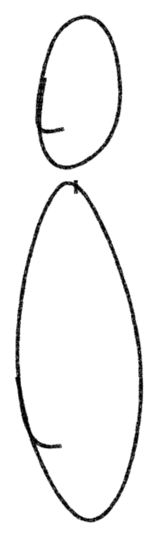
Beyond Classification tasks: Regression and Survival Modeling

What our data actually looks like

PID	Age	Smoking Status	Profession
1	55	Yes	Firefighter
2	65	Yes	Nurse
3	42	No	Chef
4	82	Yes	DJ



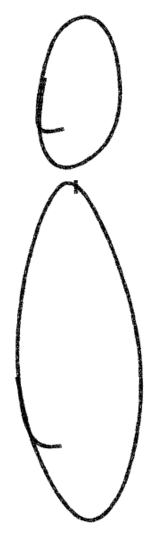
Feature Engineering: Categorical Data



Profession is

- firefighter
- nurse
- student

Feature Engineering: Categorical Data



Profession is

- firefighter → [1, 0, 0]
- nurse → [0, 1, 0]
- student → [0, 0, 1]

Assumption: categories are unrelated   firefighter · nurse = 0

Feature Engineering: Numerical Data

age X

$8 \to$ age $\to 73$

$73 \to$

Computational
PRECISION HEALTH

Berkeley | UCSF

Feature Engineering: Numerical Data

8 → age → 73

age X

73 → .73

age - category X

73 → 70+ → (0, 0, 1)

<50 / 50-70 / 70+

what do these assume when n is linear?

Computational
PRECISION HEALTH

Berkeley | UCSF

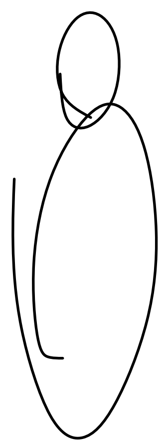
Feature Engineering: Ordinal Data

score x

Cough severity

None	0
light	1
mid	2
high	3

Feature Engineering: Ordinal Data

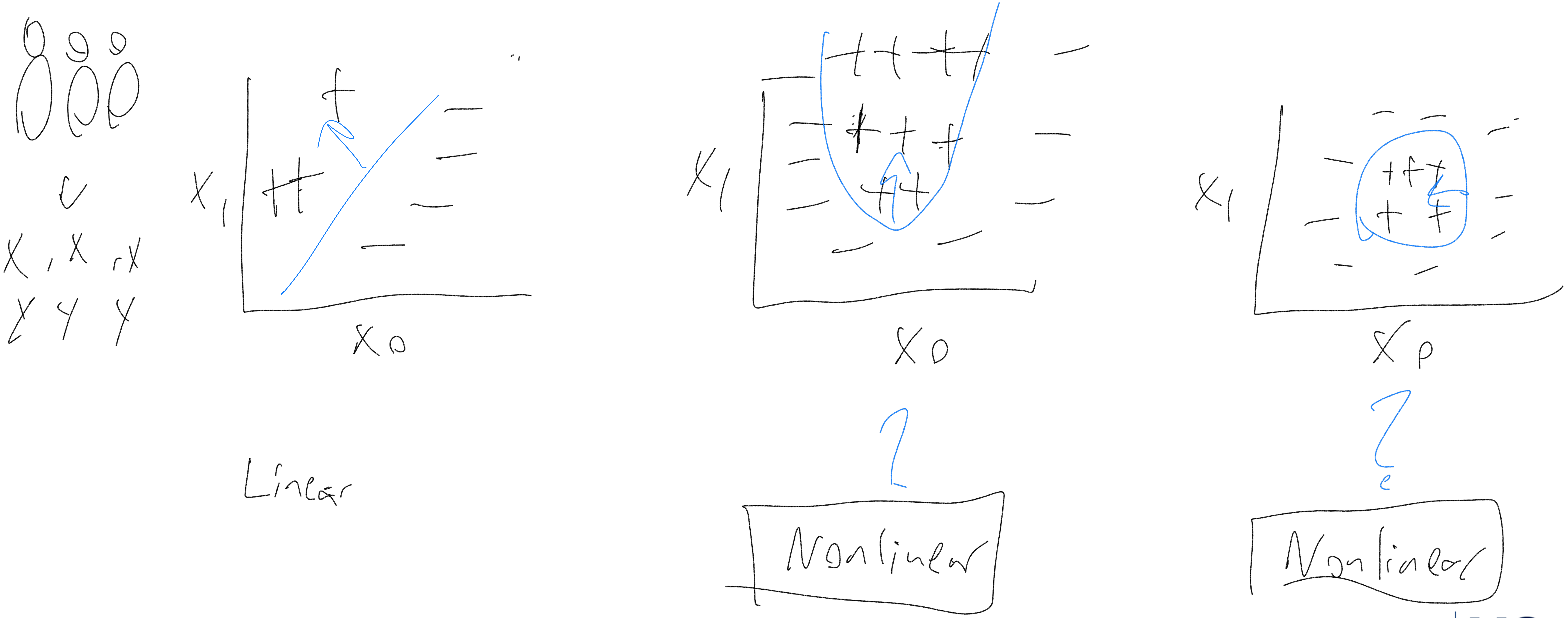


→ Cough severity

	score		x
None	0	→	[0, 0, 0]
light	1	→	[0, 0, 1]
mid	2	→	[0, 1, 1]
high	3	→	[1, 1, 1]

What does this assume? Why not view as numeric?

Feature Engineering: Dealing with non-linearity

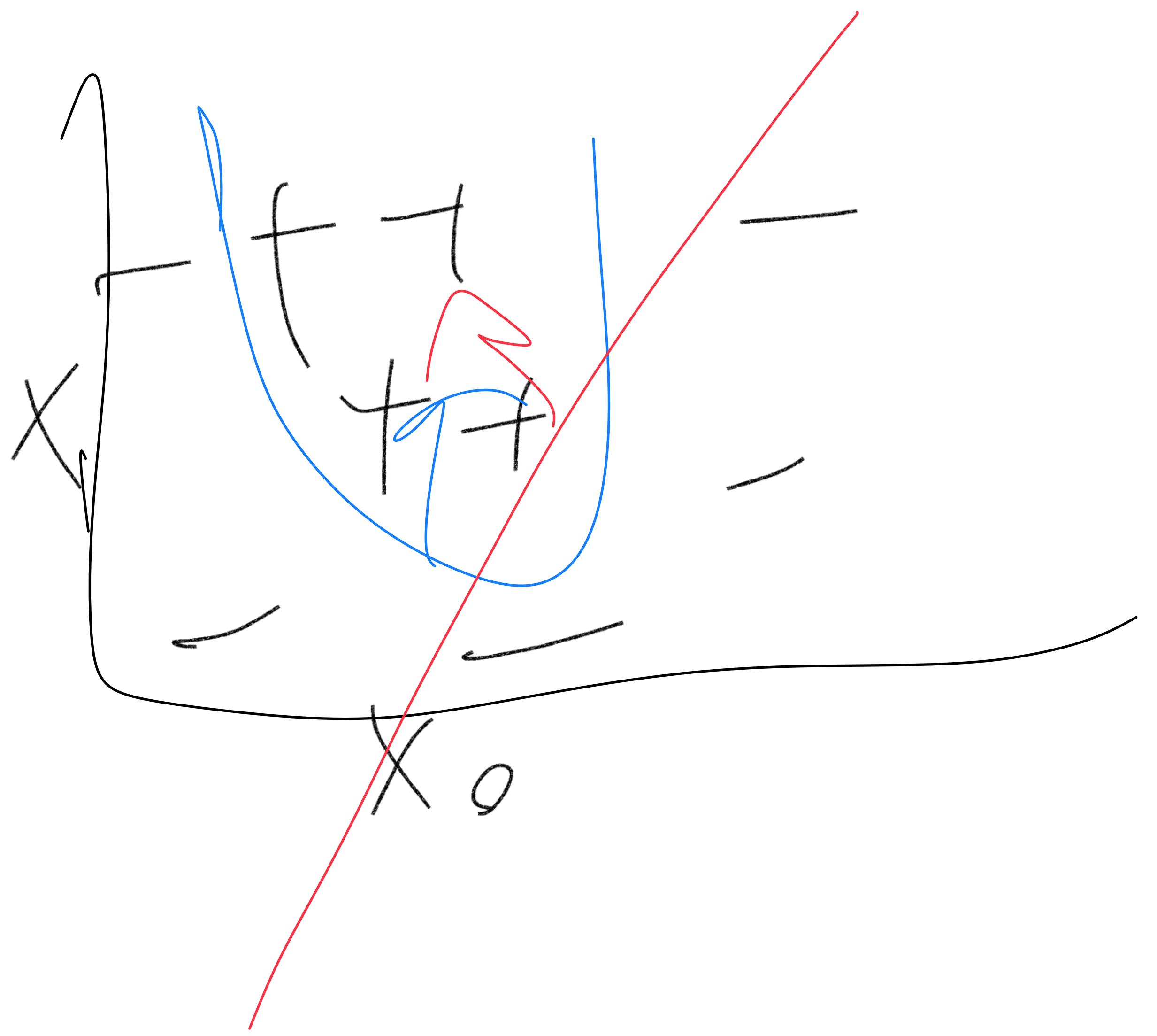


Computational
PRECISION HEALTH

Berkeley | UCSF

Feature Engineering: Dealing with non-linearity

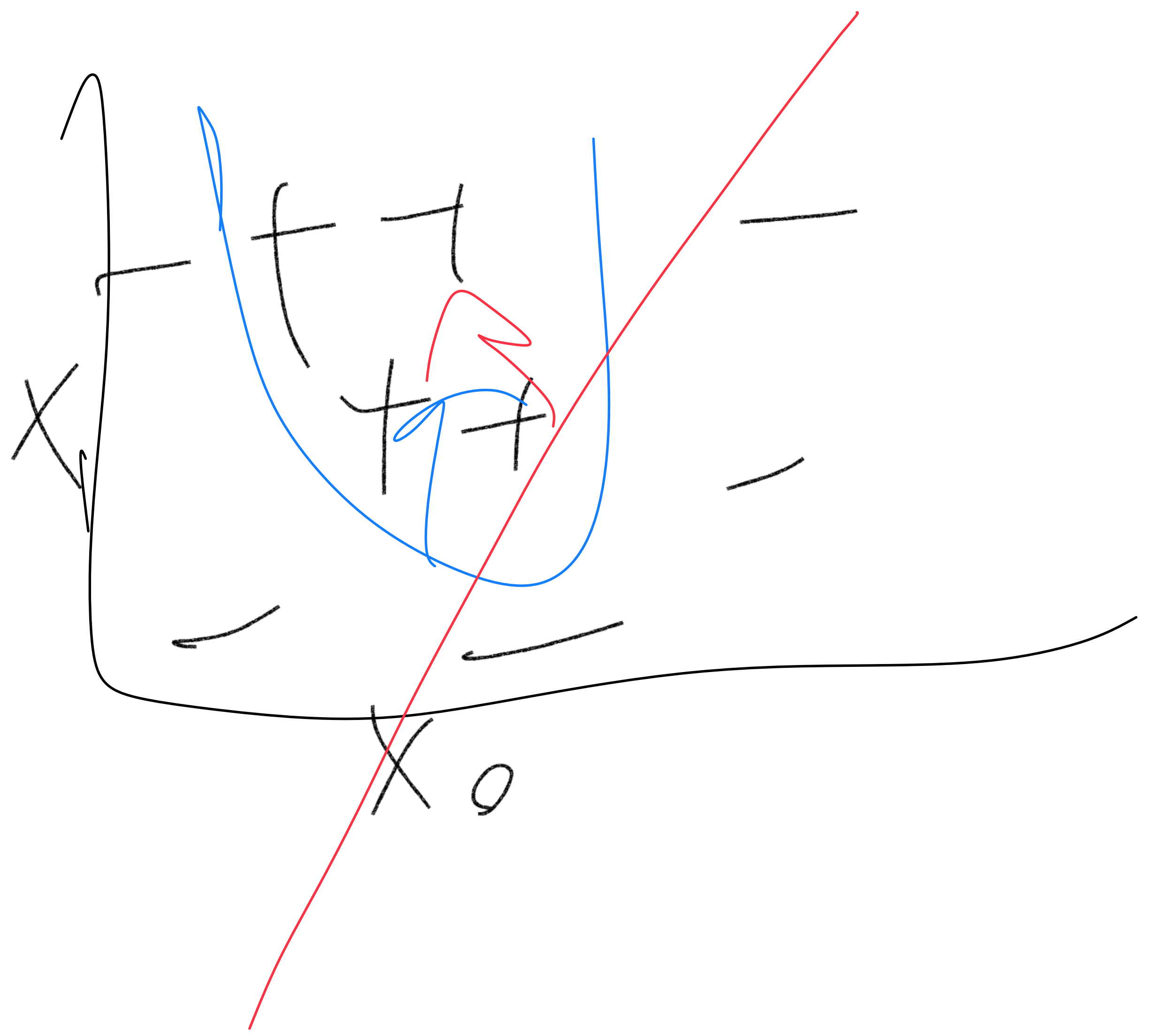
Convert non-linear to linear via features



orig features	expanded features
x_0
x_1

Feature Engineering: Dealing with non-linearity

Convert non-linear to linear via features



orig features	expanded features
x_0	x_0
x_1	x_1
	$x_0 x_1$
	x_0^2
	x_1^2

all linear!

Feature Engineering: Dealing with non-linearity

General: can expand capacity of linear model

$(a+b)^n \rightarrow$

order of features

Concretely:

[age, smoke] $\rightarrow$

Feature Engineering: Dealing with non-linearity

General: can expand capacity of linear model

$(a+b)^n \Rightarrow a^n + a^{n-1}b \ldots b^n$

order of features

Concretely:

[age, smoke] $\Rightarrow$ [age, age and smoke, smoke, ...]

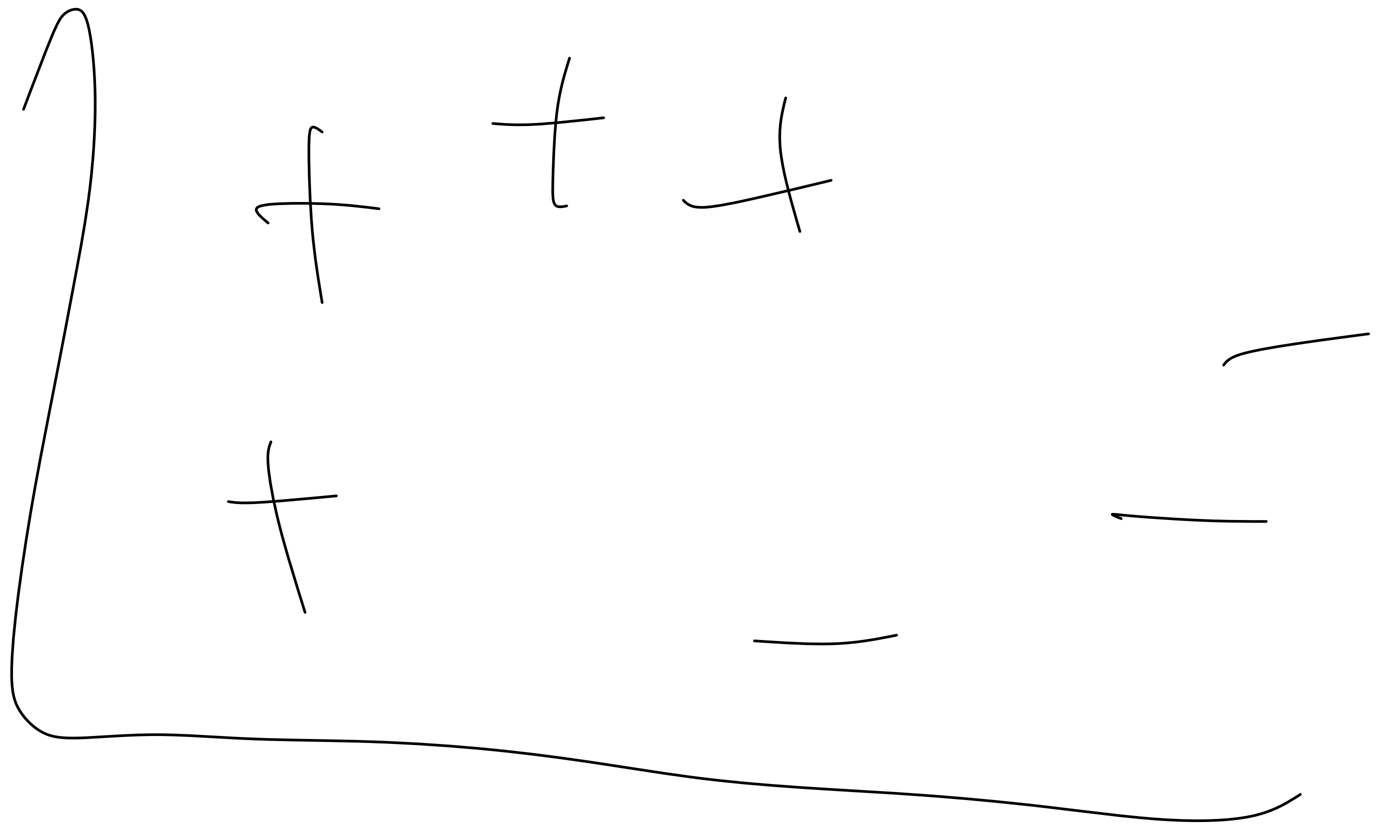
Computational
PRECISION HEALTH

Berkeley | UCSF

Feature Engineering: Dealing with non-linearity

Why not always set n high?

Supose orig $x \in R^D$ i.e $\dim(x) = D$

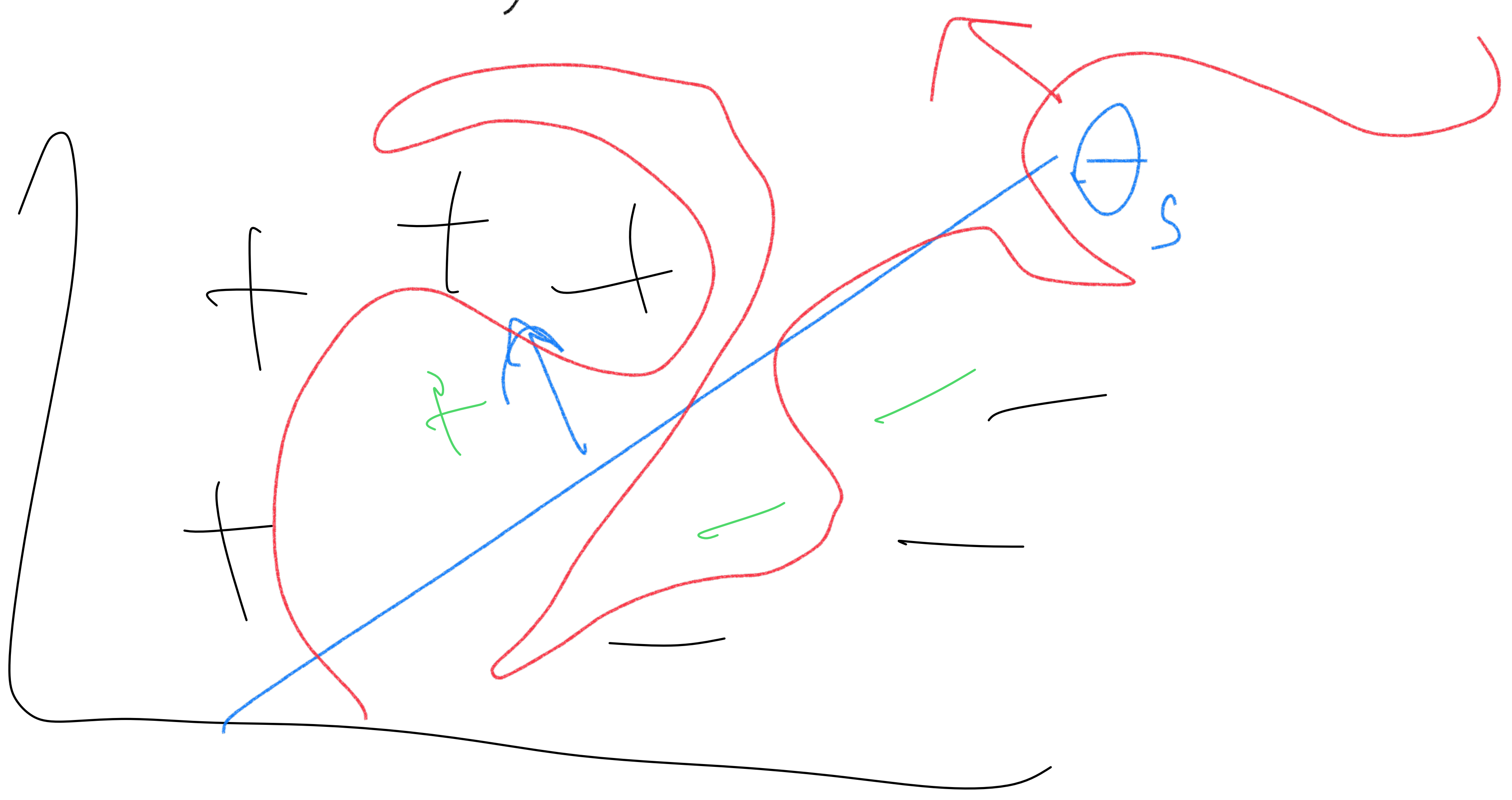


Feature Engineering: Dealing with non-linearity

Why not always set n high?

Suppose orig $x \in \mathbb{R}^D$ i.e. $\dim(x) = D$

$\dim(x^n) = O(D^n)$ ← doesn't scale



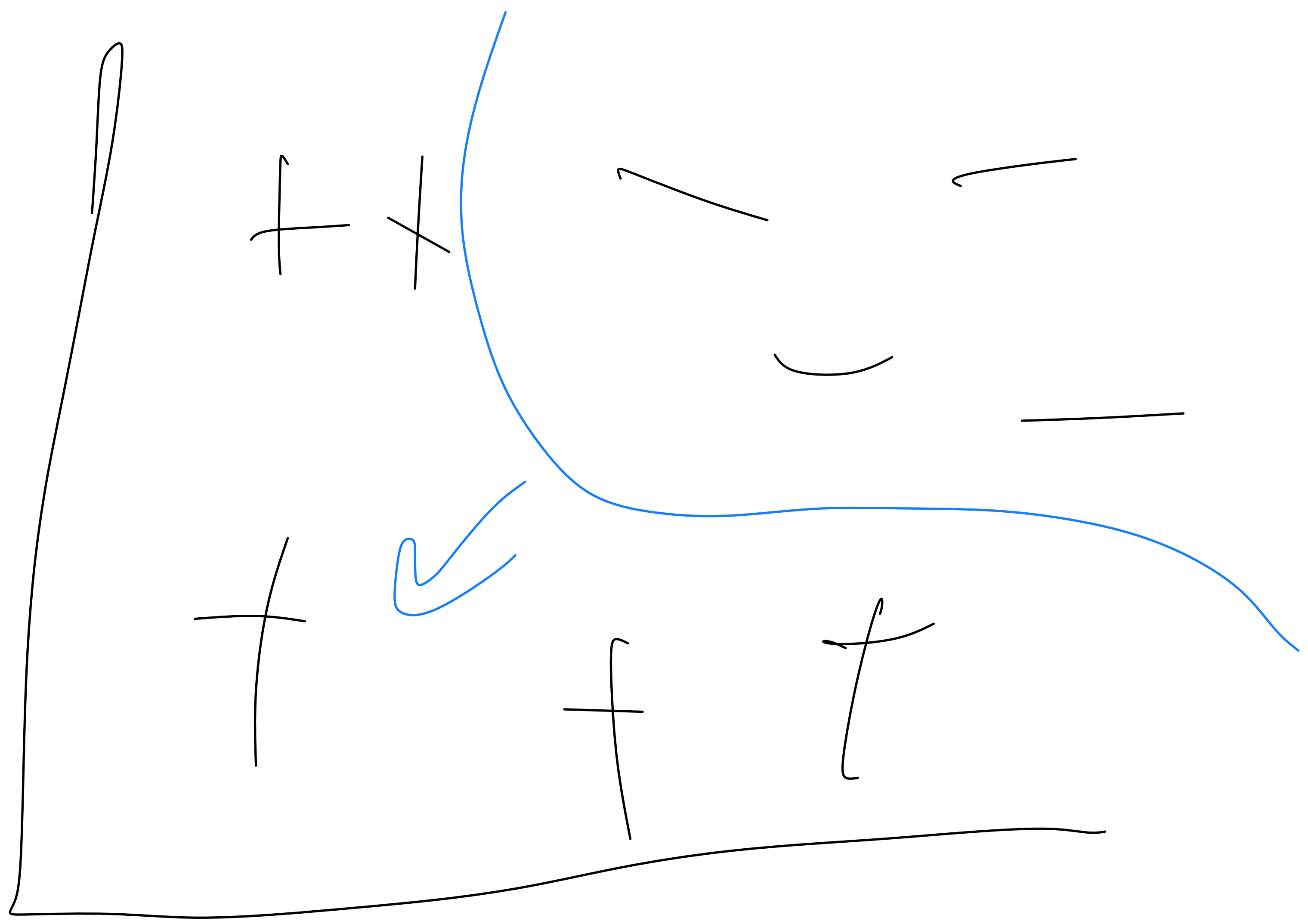
$\theta_s \to x^1$

$\theta_c \to x^{100}$

Regularization → Constrained Optimization

Bias model towards "good" behavior?

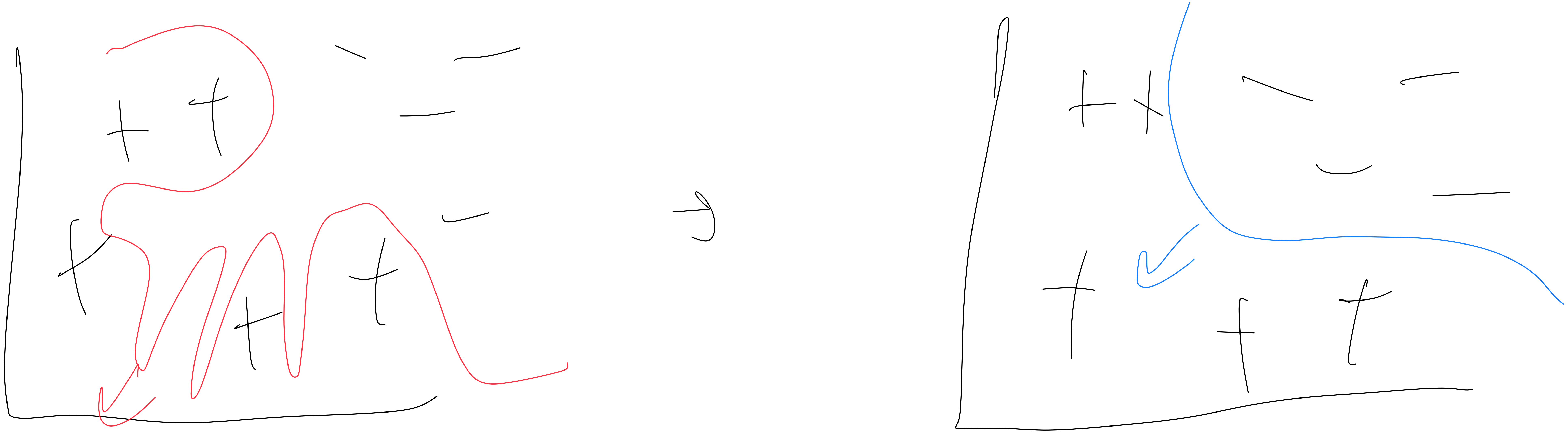
→



$L(y, p) = y \log p \ldots$

Regularization → Constrained Optimization

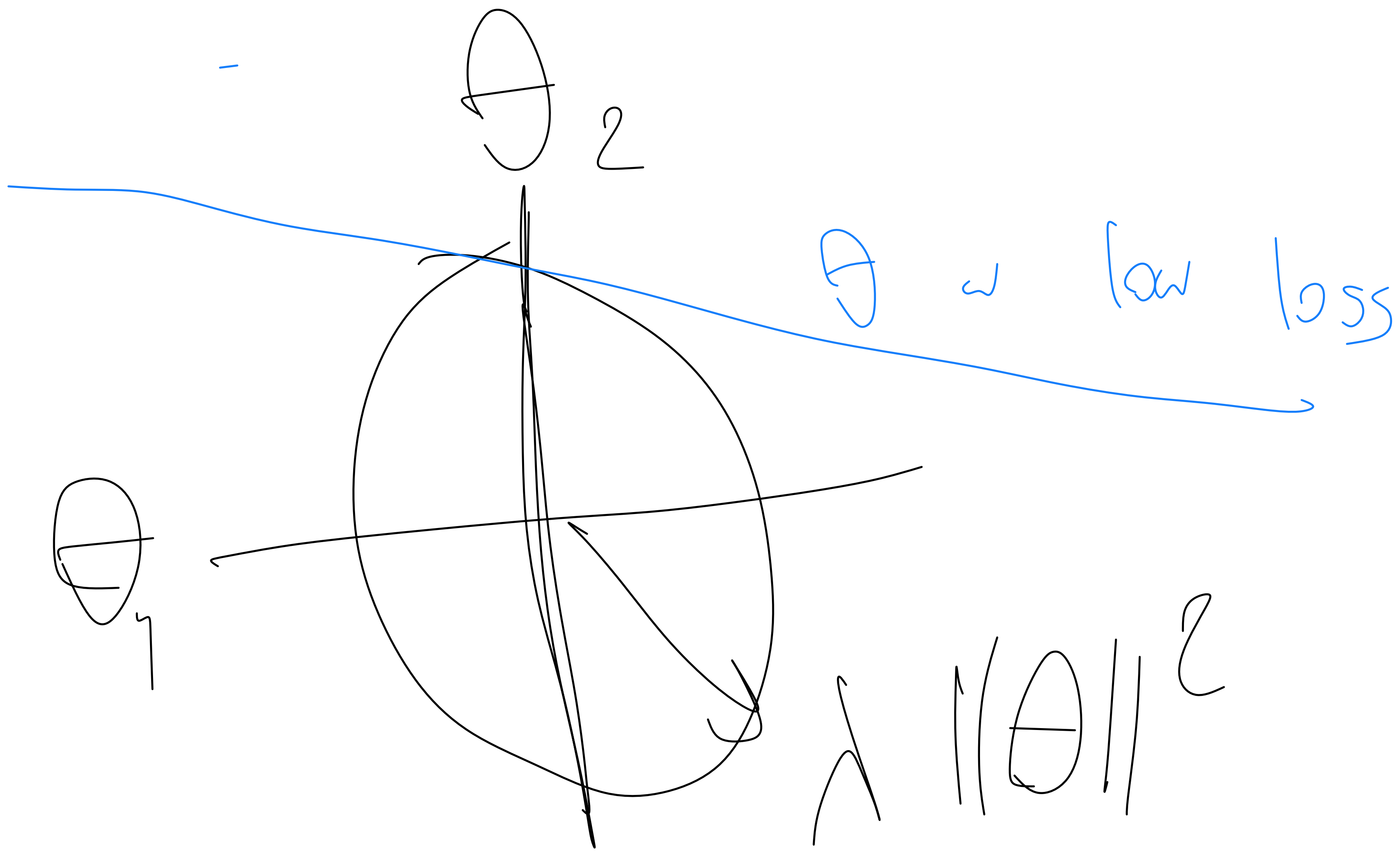
Bias model towards "good" behavior.



$L(y,p) = y \log p \ldots$

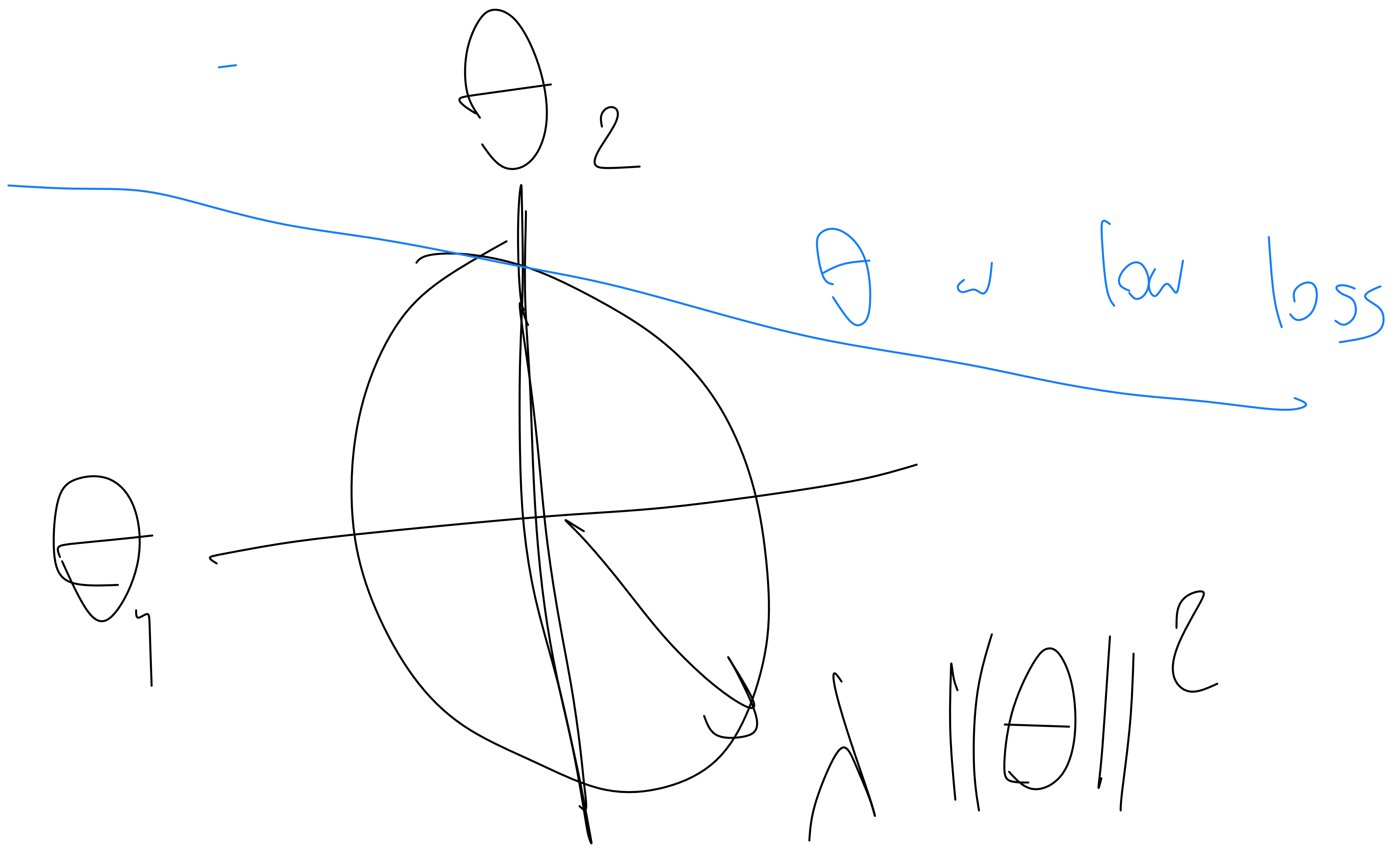
$+\lambda\|\theta\|^2$ ↑ small $\|\theta\|^2$, simpler

Regularization : Geometric perspective



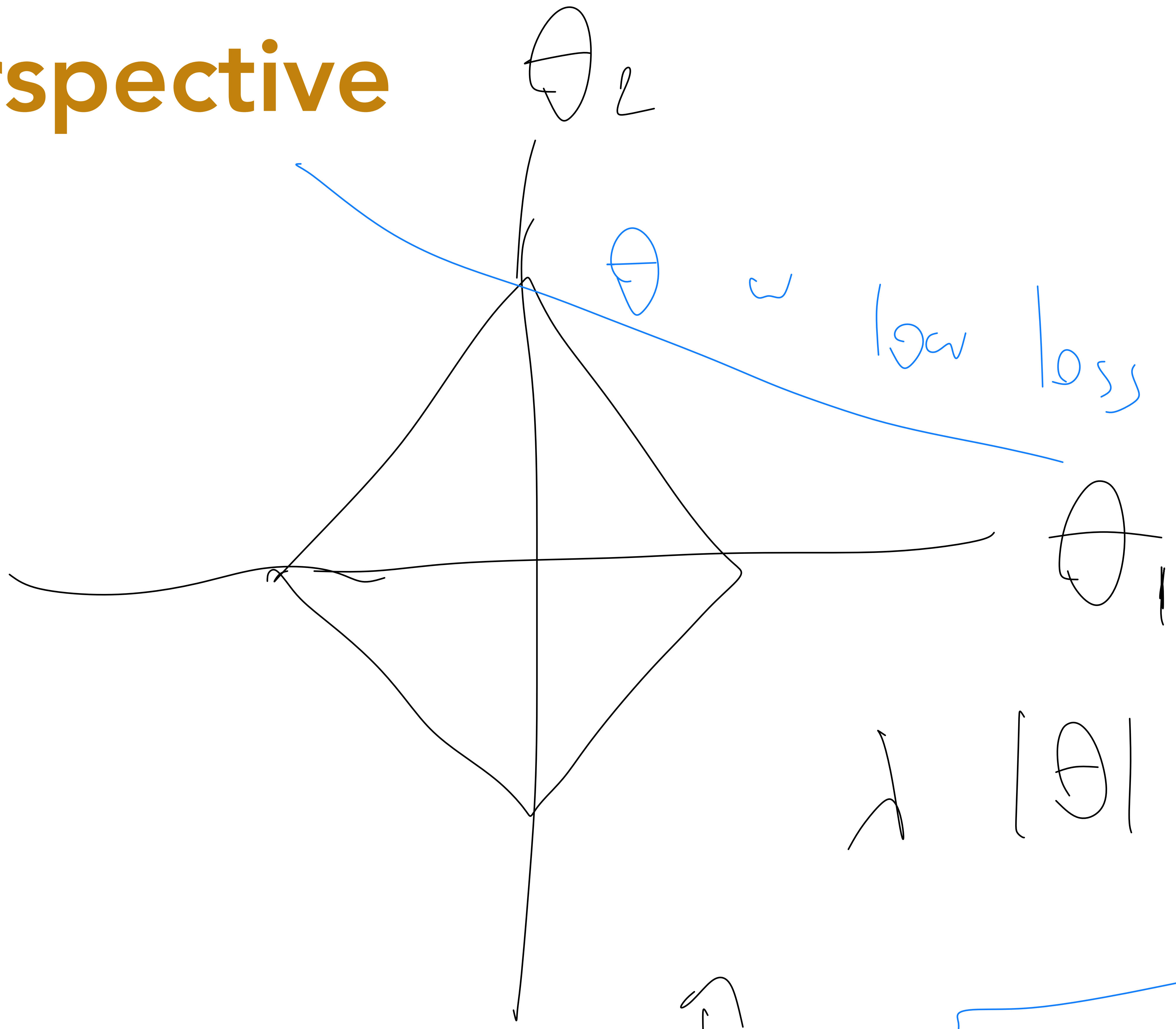
L2 or Weight Decay

Regularization : Geometric perspective



$\lambda \|\theta\|^2$

L2 or Weight Decay

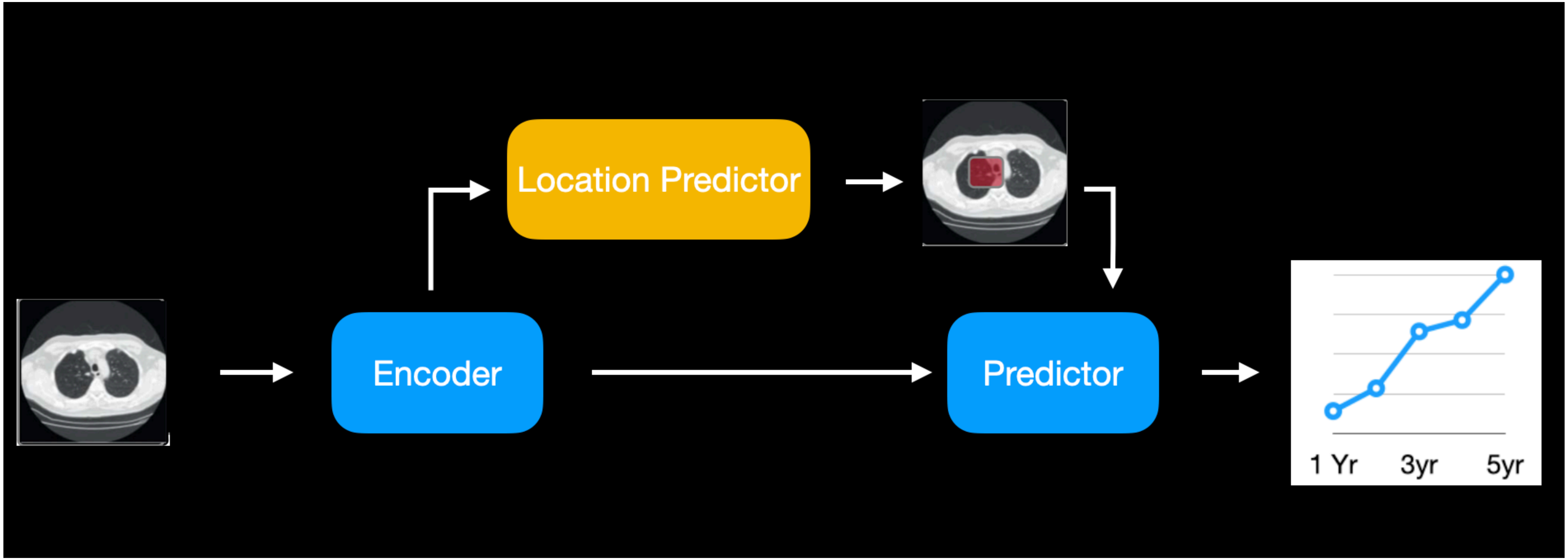


$\lambda |\theta|$

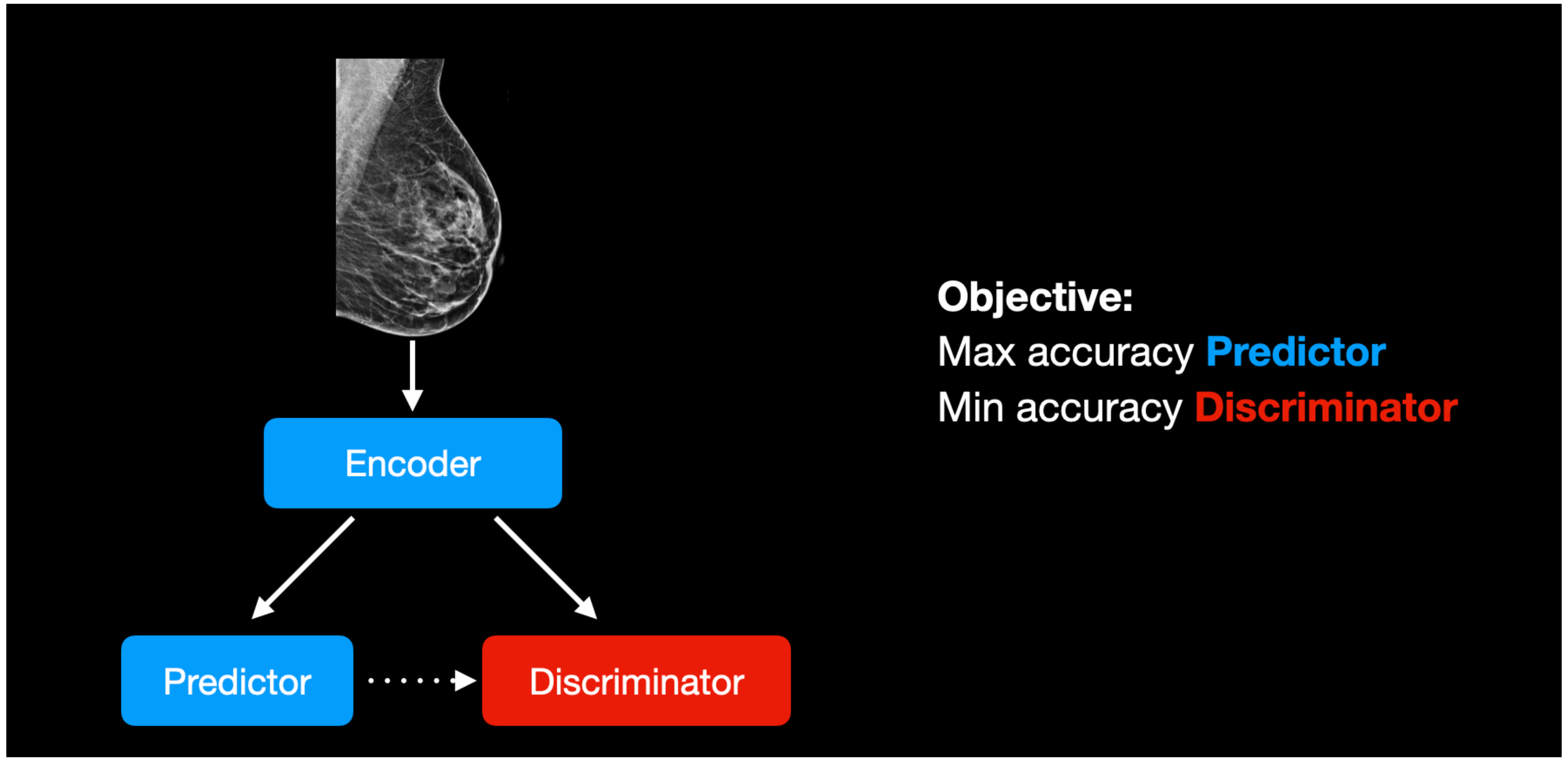
L1 (LASSO).

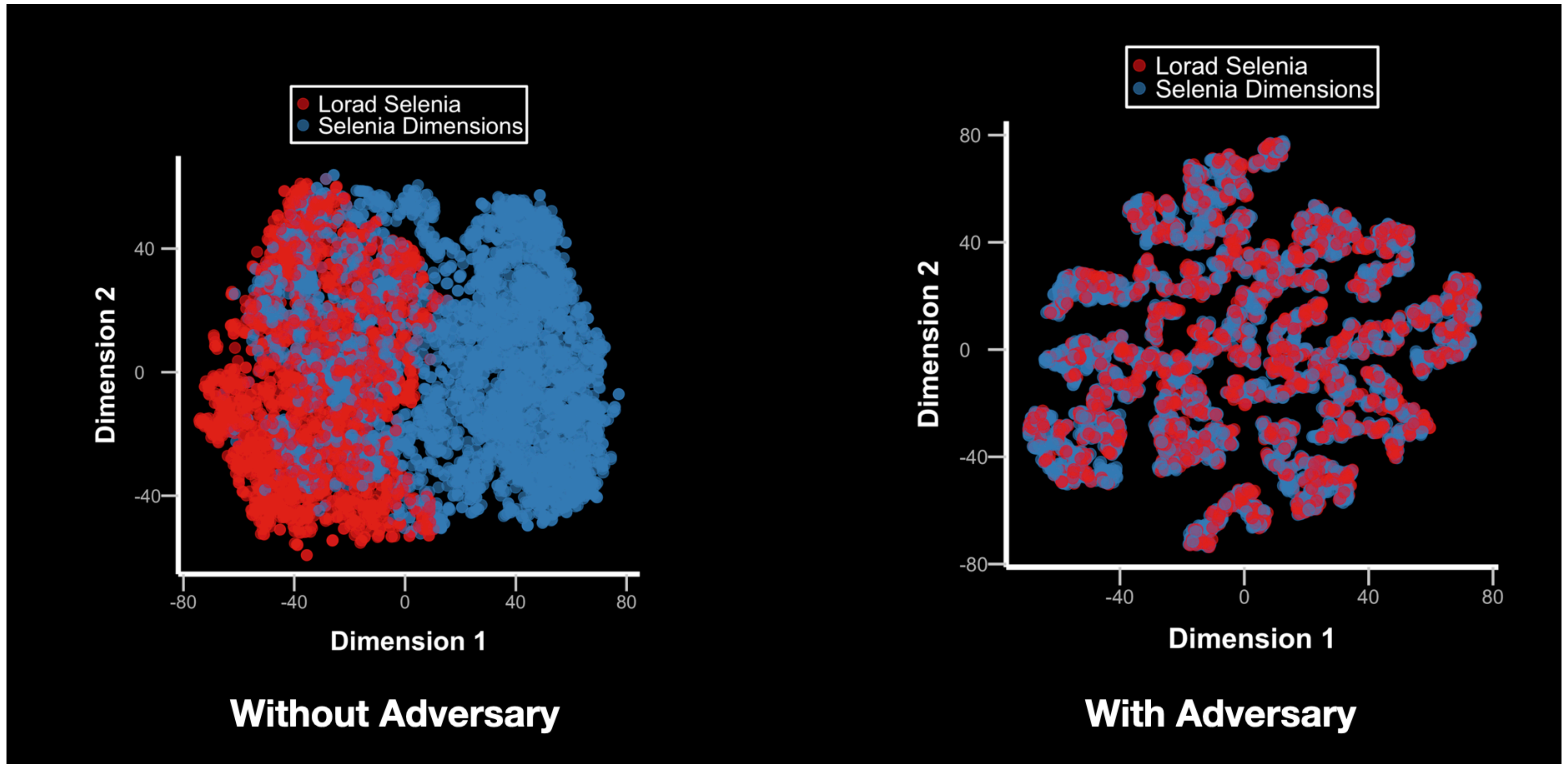
$\boxed{\text{Sparse } \theta!}$

Other examples of regularization



Other examples of regularization





Agenda

Recap

Feature Engineering

Normalization and Optimization

Beyond Classification tasks: Regression and Survival Modeling

Feature Normalization

Suppose 2 datasets D_1, D_2

$D_1 \begin{cases} x_0 & (1000, .1), \; y=1 \\ x_1 & (900, -.2), \; y=0 \end{cases}$

$D_2 - \begin{cases} x_0 & (.1 \;\; , 1), \; y=1 \\ x_1 & (.9, -.2), \; y=0 \end{cases}$

Feature Normalization

Suppose 2 datasets. D_1, D_2

D_1 $\begin{cases} x_0 & (1000, .1), \; y=1 \\ x_1 & (900, -.2), \; y=0 \end{cases}$

D_2 $\begin{cases} x_0 & (.1 \quad .1), \; y=1 \\ x_1 & (.9, -.2), \; y=0 \end{cases}$

Does it make a difference?

Yes! optimization matters

An optimization perspective

Let $\theta^0 = [.1, .9]$

$D: \begin{cases} x_0 & [1000, .1], \; y = 1 \\ x_1 & [900, -.2], \; y = 0 \end{cases}$

$\eta = 0.001$

$$\frac{\partial L}{\partial \theta} = (p - y)x$$

An optimization perspective

Let $\theta^0 = [.1, .9]$

$\eta = 0.001$

$D: \begin{cases} x_0 & [1000, .1], y=1 \\ x_1 & [900, -.2], y=0 \end{cases}$

$\frac{\partial L}{\partial \theta} = (p-y)x$

Assume x_0 1000x more important than x_1!

Can slow optimization! ↙ Poor rand init

↖ Poor η

What can we do?

An optimization perspective

Standard approach: X y

An optimization perspective

Standard approach: $X \to \dfrac{X - \mu_{\text{training set}}}{\sigma_{\text{training set}}}$

Set all x_i to 0 mean, std_dev $= \sigma = 1$

Other options?

Sep η and rand int per x_i

Computational
PRECISION HEALTH

Berkeley | UCSF

Recap: model selection

Many design decisions:

How do we choose hyperparams?

Recap: model selection

Many design decisions:

η, feature eng, normalization, initilization ...

How do we choose hyperparams?

Val set from same dist at test

Why not train?

Agenda

Recap

Feature Engineering

Normalization and Optimization

Beyond Classification tasks: Regression and Survival Modeling

High level Machine Learning Process

Find an important problem: e.g. Lung cancer screening

Find a good test-set: UCSF test set

Find training data: NLST dataset

Define your training objective and hypothesis class: cross entropy, LR

Optimize model and choose params on validation data: NLST holdout

Test generalization and study clinical impact: UCSF test set

Predicting numerical data: Regression

Motivating example: Predicting tumor (mm) size after treatment

Predicting numerical data: Regression

Motivating example: Predicting tumor (mm) size after treatment

$R(\theta) = \frac{1}{n}\sum_{n} L(y, p)$

$h_\theta : \theta X^T$

$L(y, p) = \frac{1}{2}\|y - p\|^2$

$\frac{\partial L}{\partial \theta} = (p - y)X$

Rest is the same!

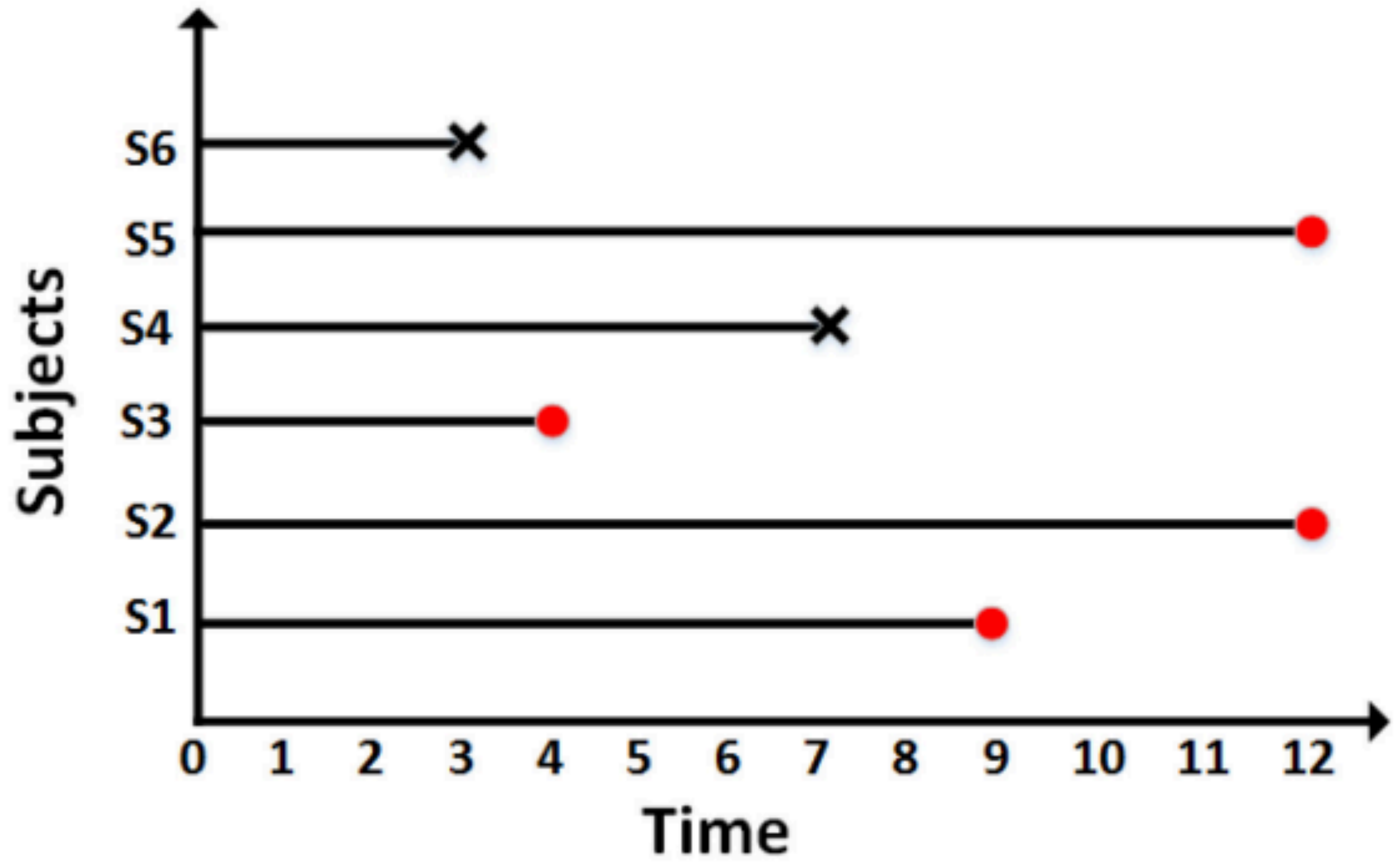
Computational
PRECISION HEALTH

Berkeley | UCSF

Predicting time-to-event data: Survival

In real world scenarios, we lose people to followup (right censoring)

Examples: Cancer screening, drug trials, etc.

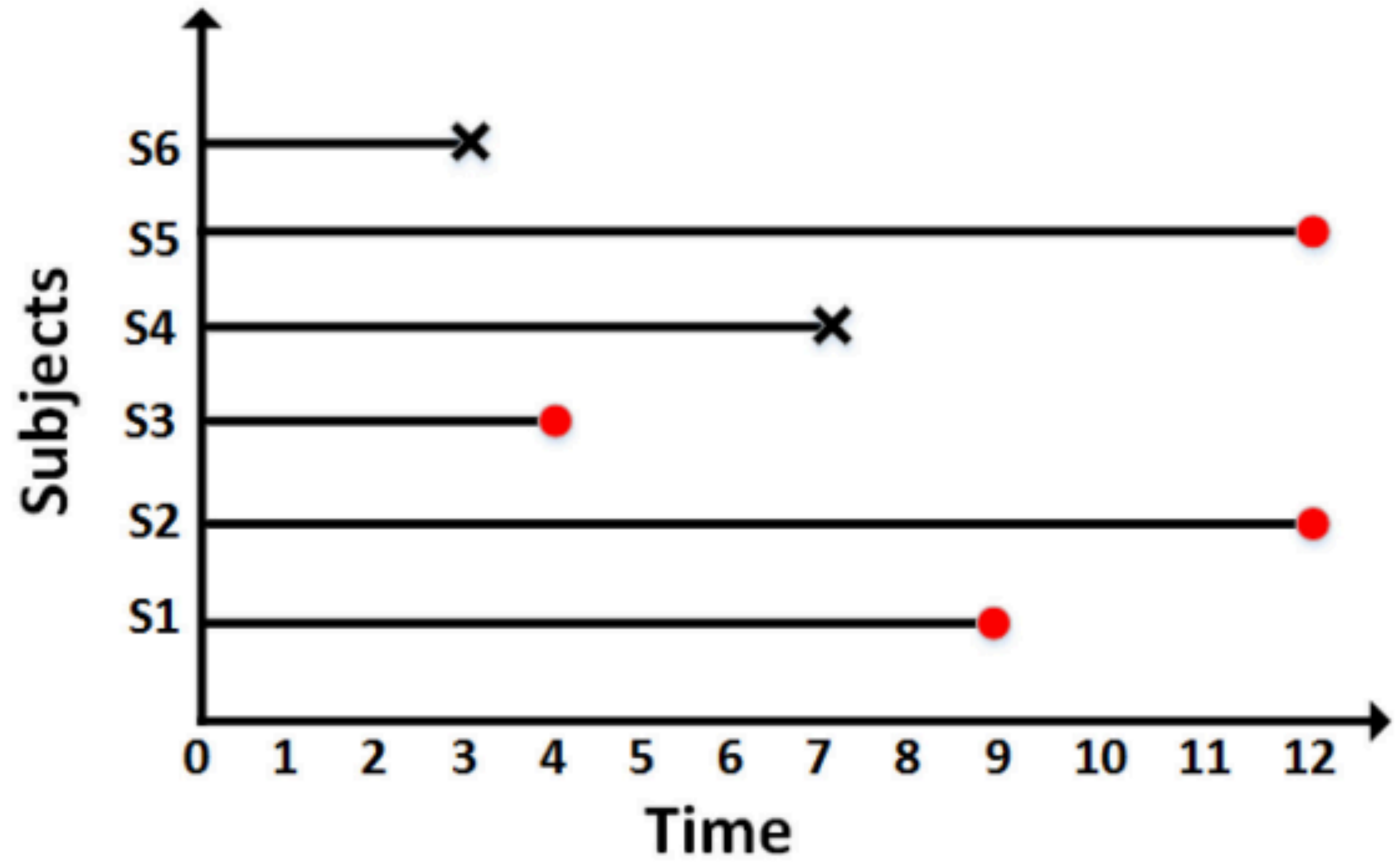


Wang, Ping, Yan Li, and Chandan K. Reddy. "Machine learning for survival analysis: A survey." *ACM Computing Surveys (CSUR)* 51.6 (2019): 1-36.

Survival Modeling

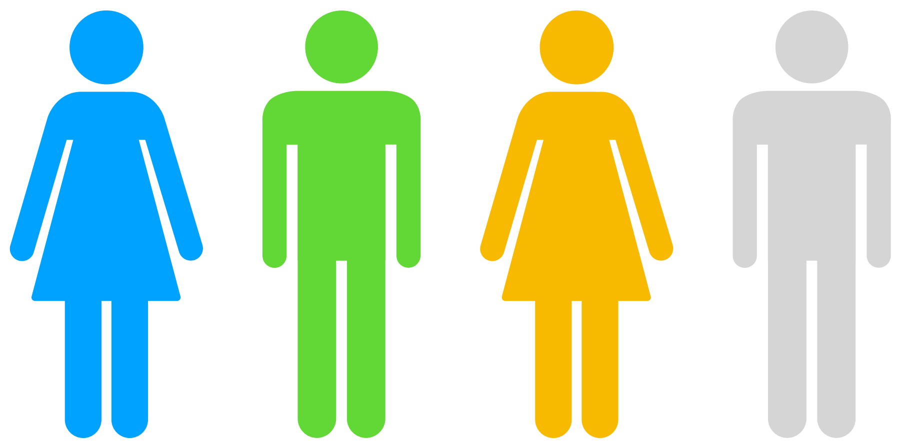
Why not view as classification task?

Why not view as regression task?



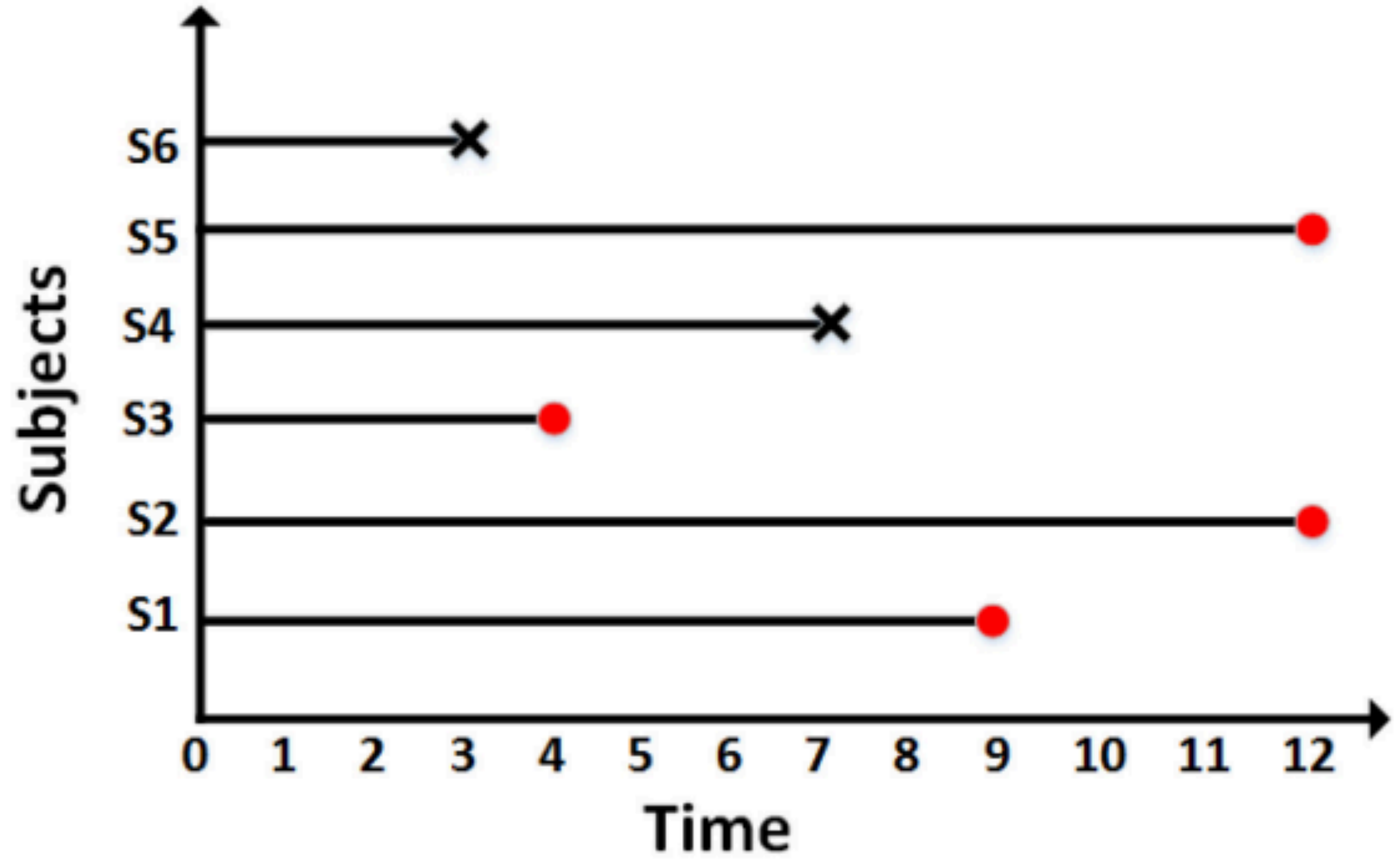
Survival Modeling: What we know

n historical patients



$(x, y, c) =$ (feature features, time, censoring)

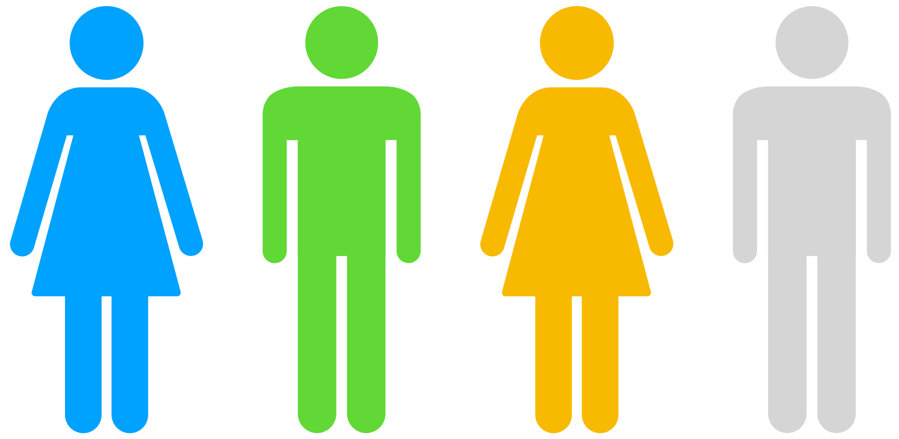
$c = 0$ -> event at time y
$c = 1$ -> censoring at time y



Wang, Ping, Yan Li, and Chandan K. Reddy. "Machine learning for survival analysis: A survey." *ACM Computing Surveys (CSUR)* 51.6 (2019): 1-36.

Survival Modeling: What we want

n historical patients



Estimate:

$$S(t) = P(T > t) = \int_t^{\infty} f(x)dx$$

$$F(t) = 1 - S(t)$$

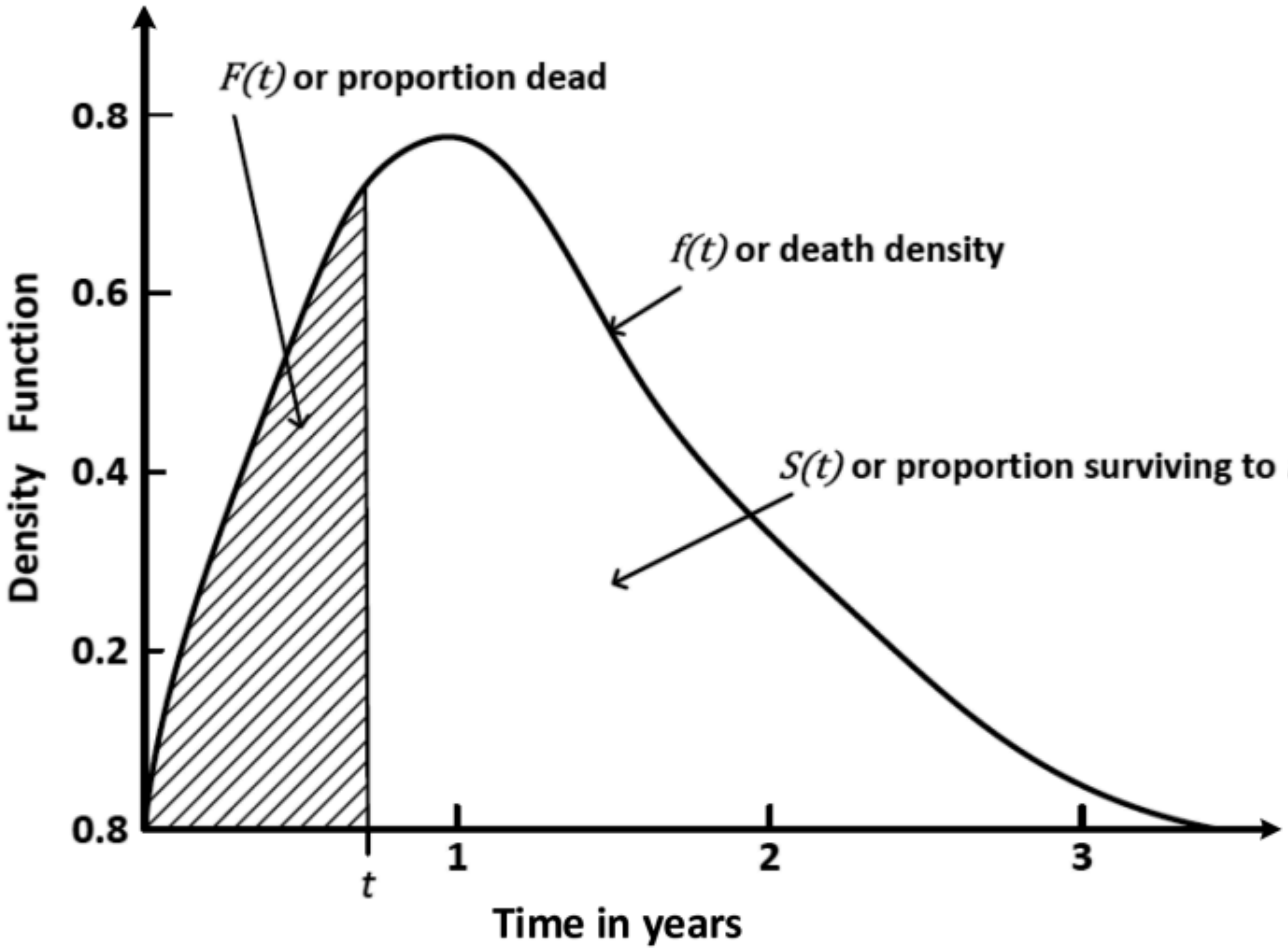


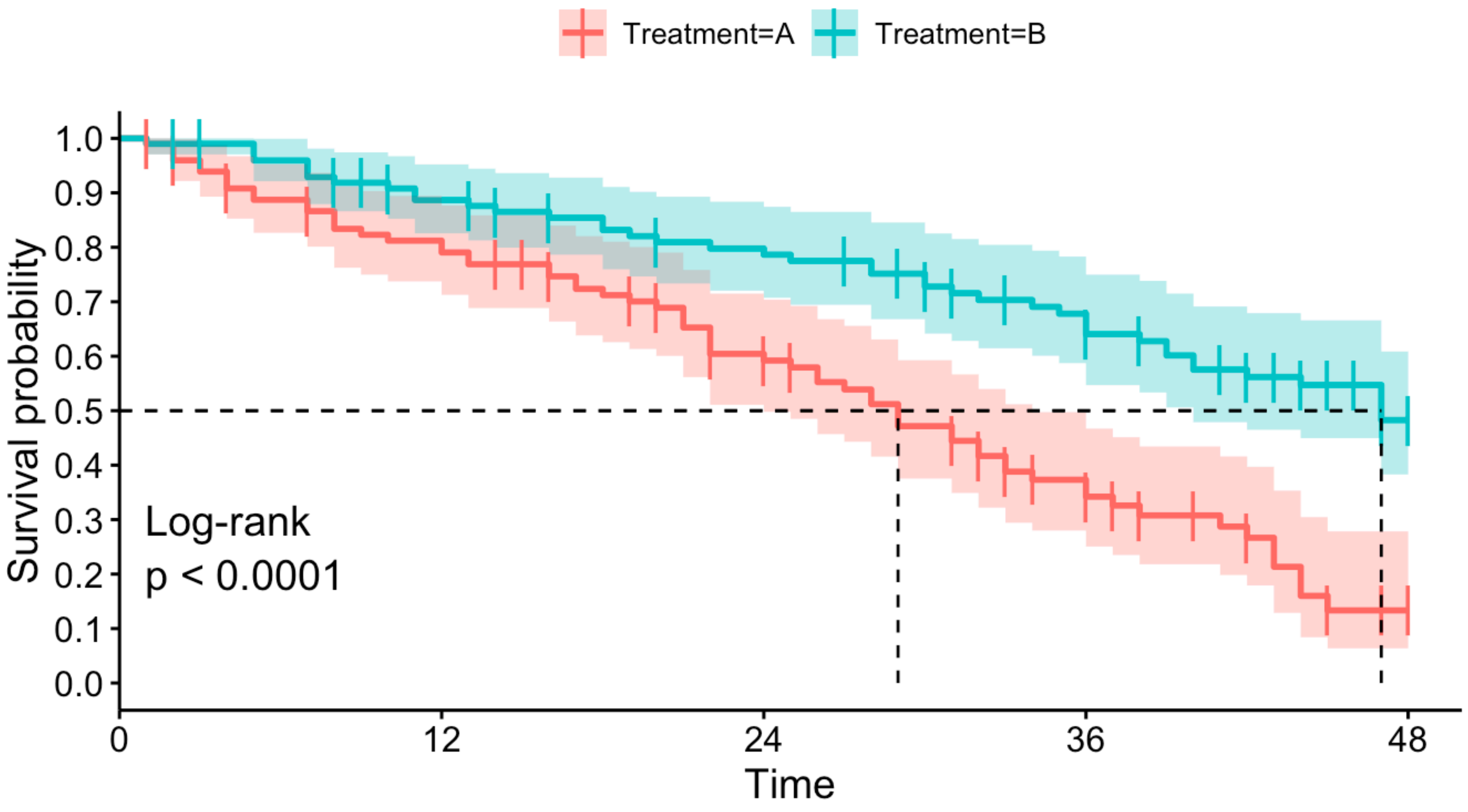
Fig. 2. Relationships between the different entities $f(t)$, $F(t)$, and $S(t)$.

Wang, Ping, Yan Li, and Chandan K. Reddy. "Machine learning for survival analysis: A survey." *ACM Computing Surveys (CSUR)* 51.6 (2019): 1-36.

Computational
PRECISION HEALTH

Berkeley | UCSF

A non-parametric approach: Kaplan-Meier estimators



Sort data by event times:

$$y_1 < y_2 \ldots < y_N$$

d_i = # events at time y_i

n_i = # uncensored subjects w.o event

$$S(t) = \prod_{i:y_i \leq t} (1 - \frac{d_i}{n_i})$$

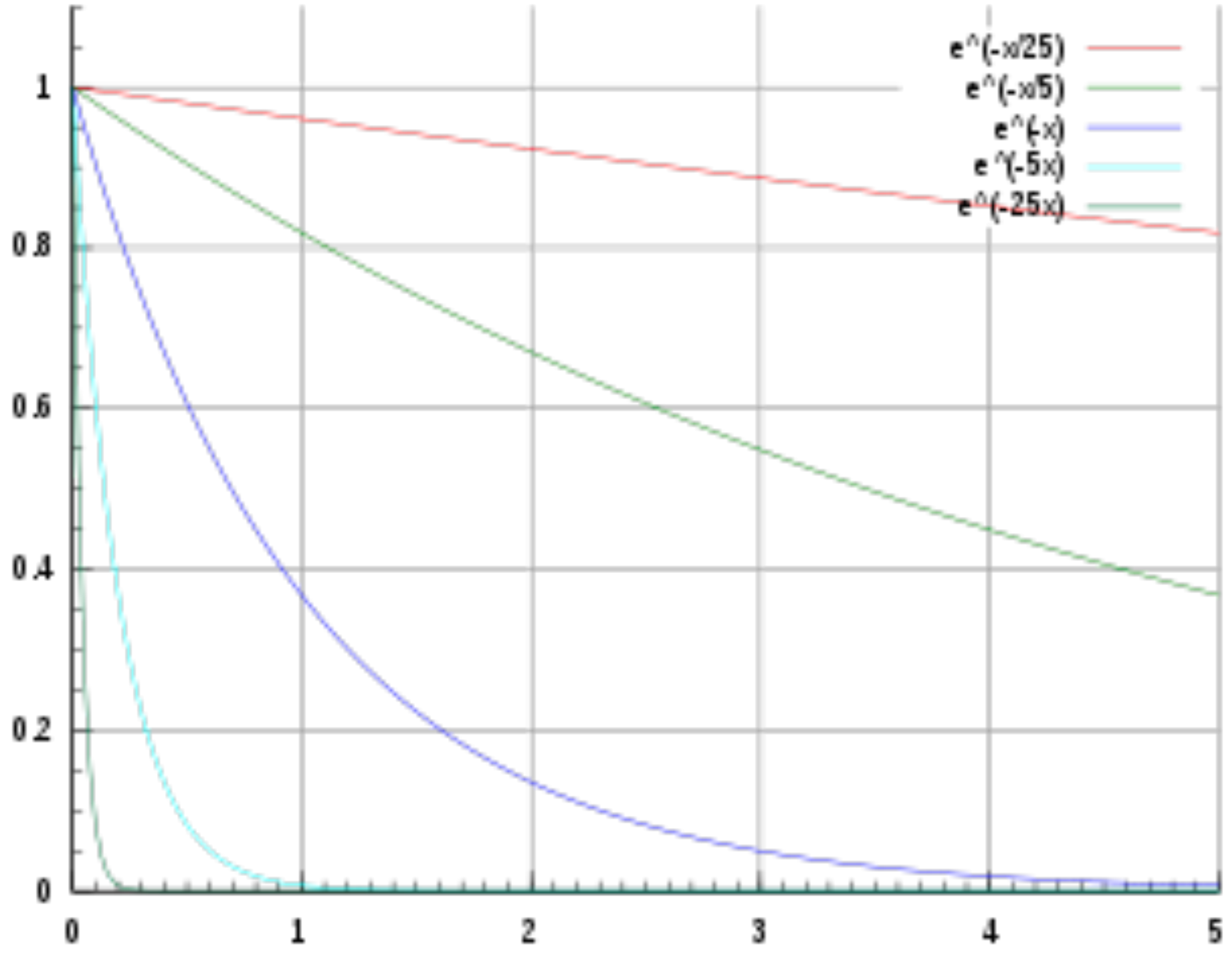
Parametric approaches

Choose parametric form of $f(t), S(t)$

Maximize likelihood of observations under parametric model

Examples: *Exponential Decay* $S(t) = e^{-\lambda t},\ f(t) = \lambda e^{-\lambda t}$

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



What does this assume?

Parametric approaches

Choose parametric form of $f(t), S(t)$

Maximize likelihood of observations under parametric model

Many common options reflecting different *hypothesis classes*

Distribution	**PDF** $f(t)$	**Survival** $S(t)$
Exponential	$\lambda exp(-\lambda t)$	$exp(-\lambda t)$
Weibull	$\lambda k t^{k-1} exp(-\lambda t^k)$	$exp(-\lambda t^k)$
Logistic	$\frac{e^{-(t-\mu)/\sigma}}{\sigma(1+e^{-(t-\mu)/\sigma})^2}$	$\frac{e^{-(t-\mu)/\sigma}}{1+e^{-(t-\mu)/\sigma}}$
Log-logistic	$\frac{\lambda k t^{k-1}}{(1+\lambda t^k)^2}$	$\frac{1}{1+\lambda t^k}$
Normal	$\frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(t-\mu)^2}{2\sigma^2})$	$1-\Phi(\frac{t-\mu)}{\sigma})$
Log-normal	$\frac{1}{\sqrt{2\pi}\sigma t} exp(-\frac{(log(t)-\mu)^2}{2\sigma^2})$	$1-\Phi(\frac{log(t)-\mu}{\sigma})$

Wang, Ping, Yan Li, and Chandan K. Reddy. "Machine learning for survival analysis: A survey." *ACM Computing Surveys (CSUR)* 51.6 (2019): 1-36.

Training Parametric Survival Models: MLE

Censored Observations: $(x_i, y_i, c = 1)$ -> $S(t = y_i, x = x_i) = 1$

Likelihood = $\prod_{c=1} S(t = y_i, x = x_i)$

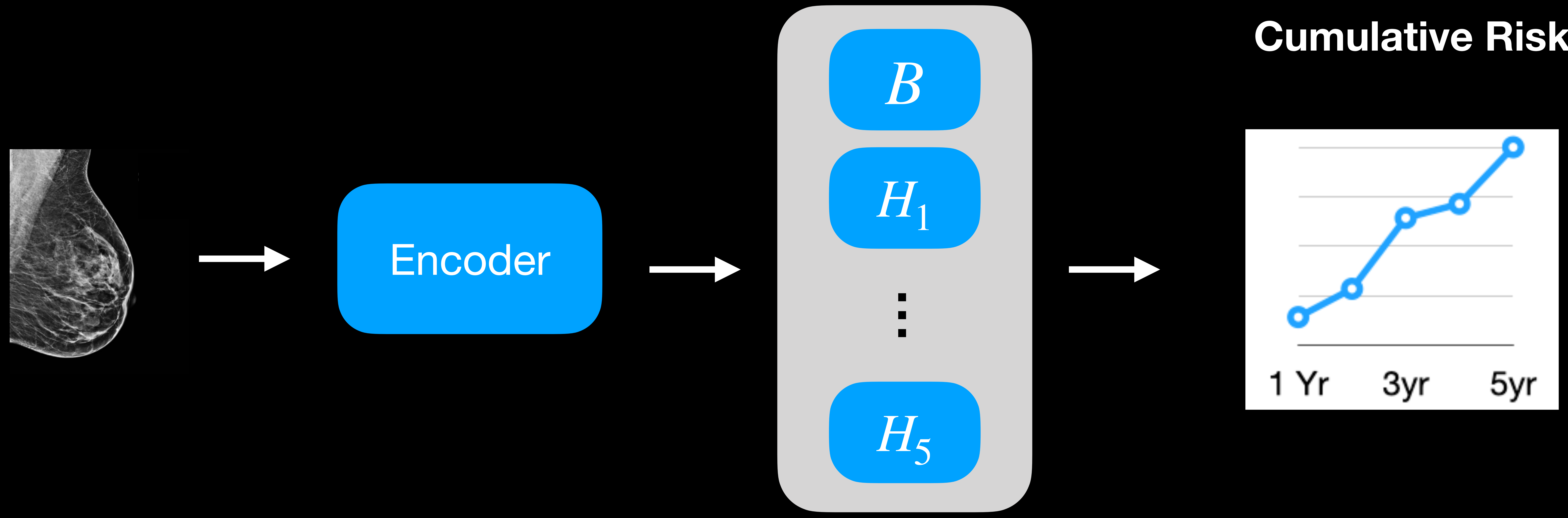
Uncensored Observations: $(x_i, y_i, c = 0)$ -> $f(t = y_i, x = x_i) = 1$

Likelihood = $\prod_{c=0} f(t = y_i, x = x_i)$

Overall Log-Likelihood = $\Sigma_{c=1} log(S(t = y_i, x = x_i)) + \Sigma_{c=0} log(f(t = y_i, x = x_i))$

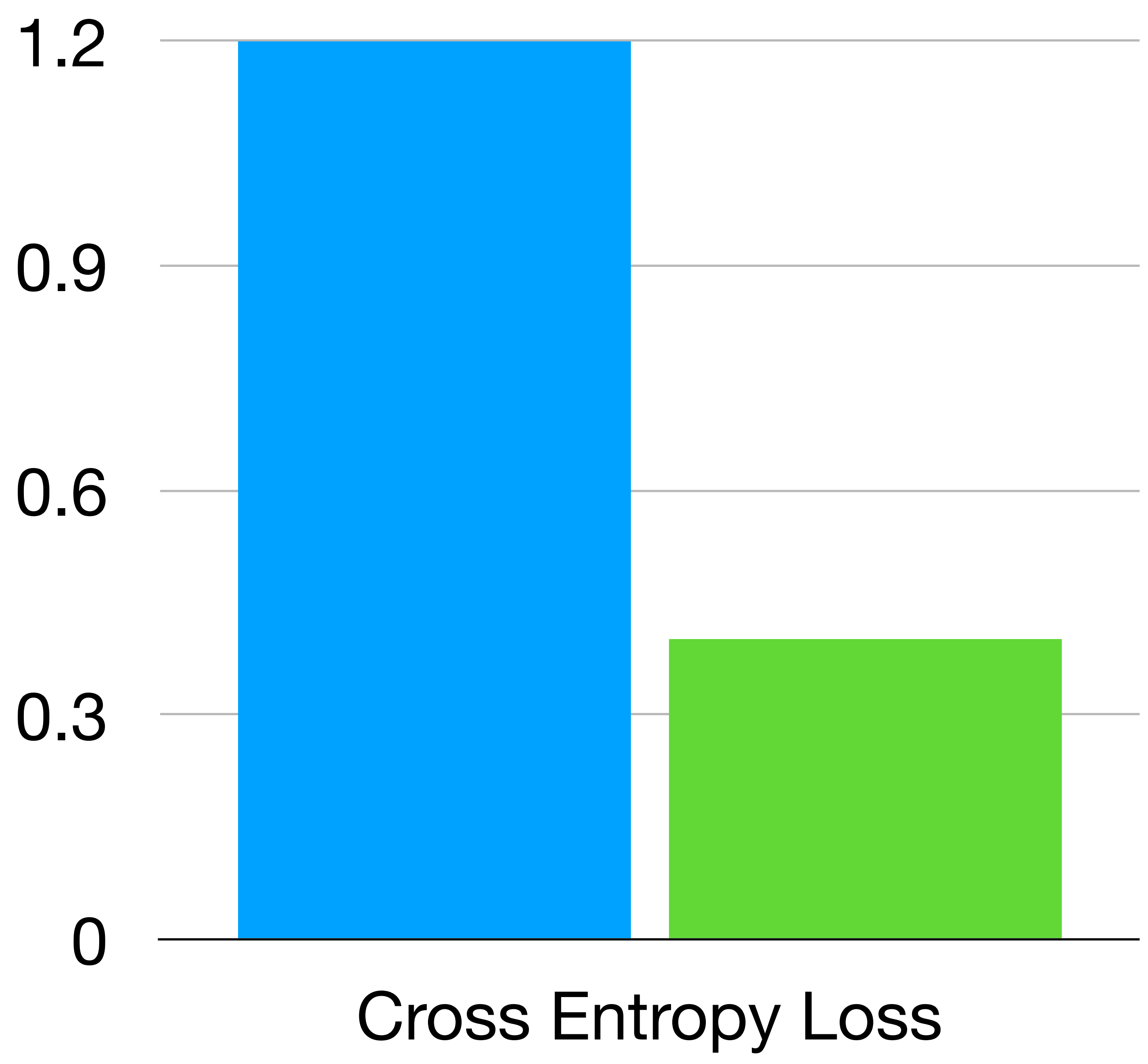
Optimization: Gradient Descent or Stochastic Gradient Descent!

Discrete time parametric approach



$$F(t) = P(t_{cancer} = k \mid x) = B(E(x)) + \sum H_i(E(x))$$

Model Evaluation



Modeling objective

Achievable performance

$1: h(x) \geq p$

$0: h(x) < p$

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

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

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

Common Survival Metrics

C-index: Generalized ROC AUC for survival modeling

$$c = P(\hat{y}_1 > \hat{y}_2 \,|\, y_1 > y_2) = \frac{1}{N}\Sigma_{i:c=1}\Sigma_{j:y_i<y_j} I[S(\hat{y}_j) > S(\hat{y}_i)]$$

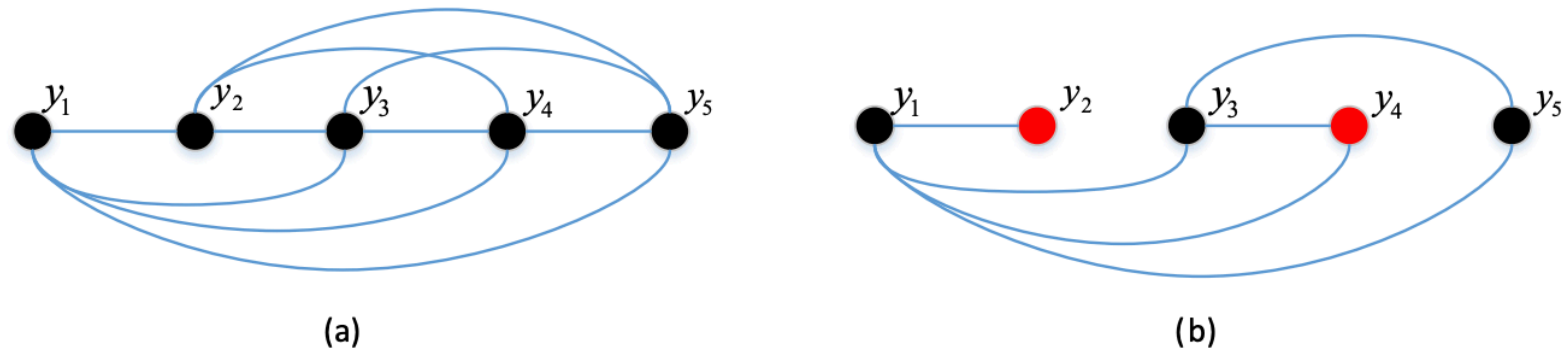
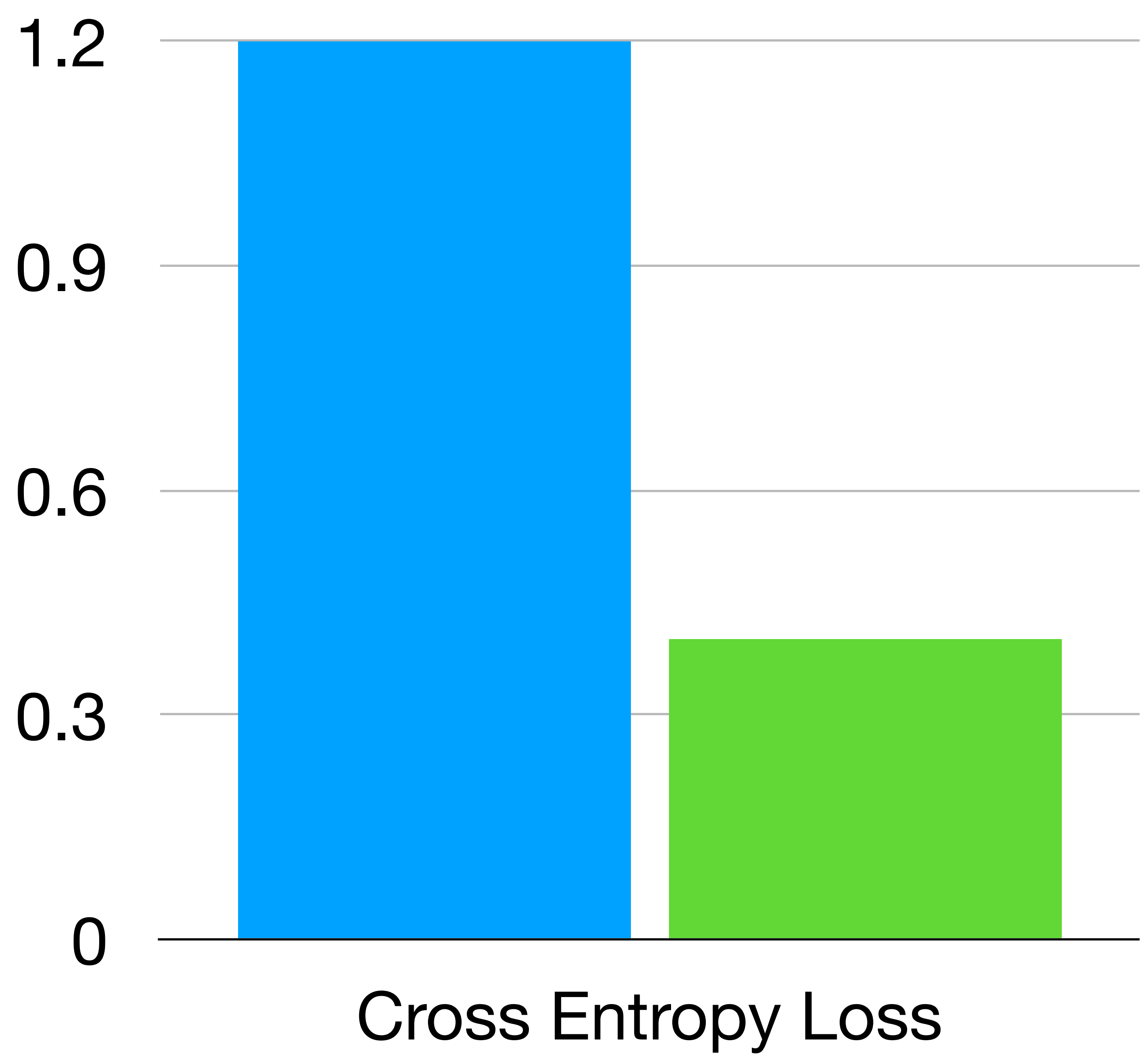


Fig. 4: Illustration of the ranking constraints in survival data for C-index calculations ($y_1 < y_2 < y_3 < y_4 < y_5$). Here, black circles indicate the observed events and red circles indicate the censored observations. (a) No censored data and (b) With censored data.

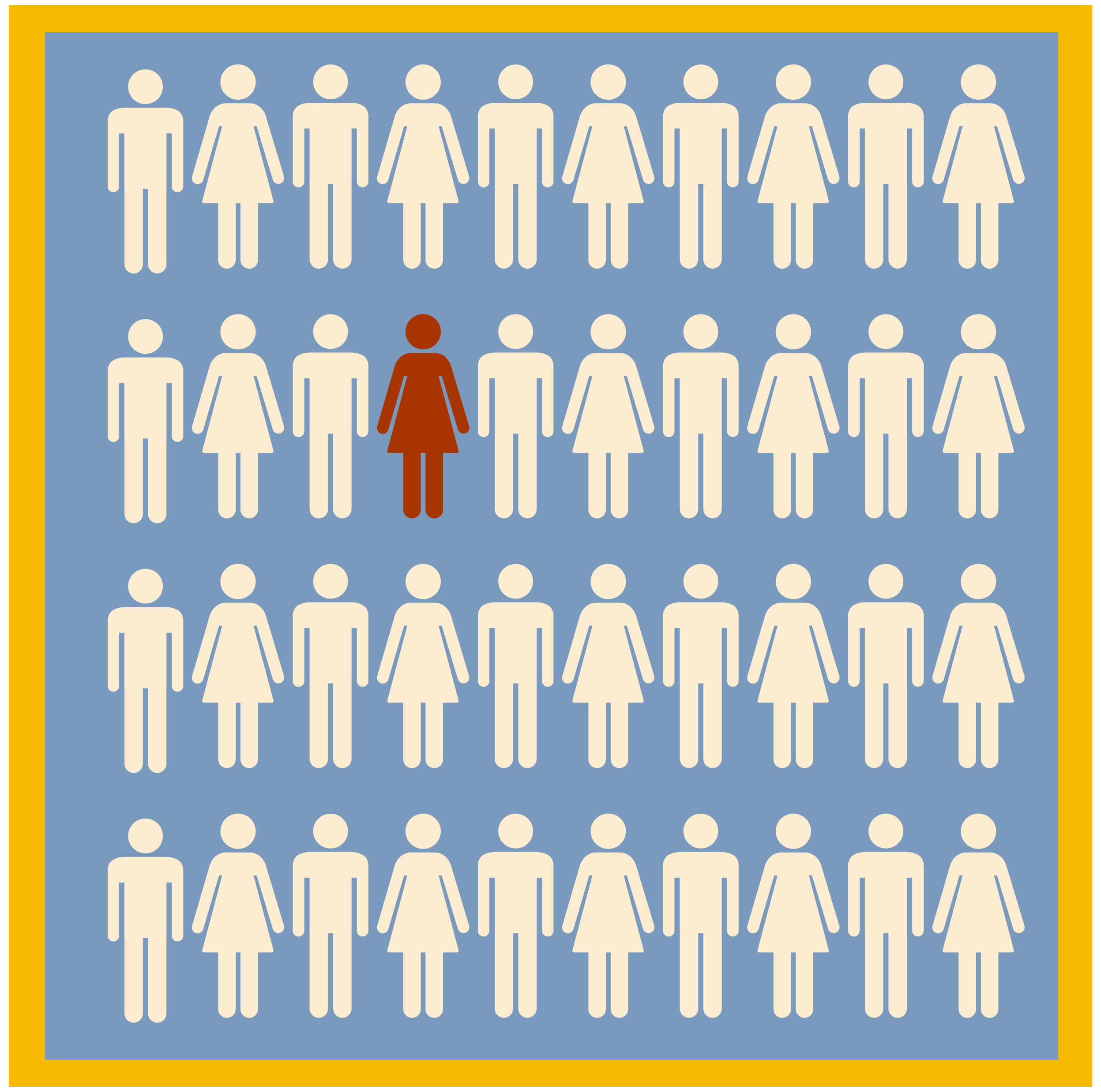
Wang, Ping, Yan Li, and Chandan K. Reddy. "Machine learning for survival analysis: A survey." *ACM Computing Surveys (CSUR)* 51.6 (2019): 1-36.

Model Evaluation



Modeling objective

Achievable performance



Simulated clinical utility

Summary

Representing categorical, numerical and ordinal data

Feature expansion and regularization

Feature Normalization

Regression

Survival Modeling: Non-parametric and-parametric estimators

Questions?

Computational
PRECISION HEALTH

Berkeley | UCSF