CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

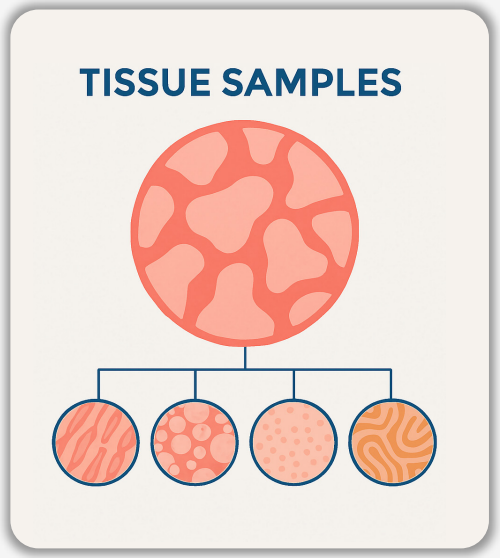
Learning Actionable Insights

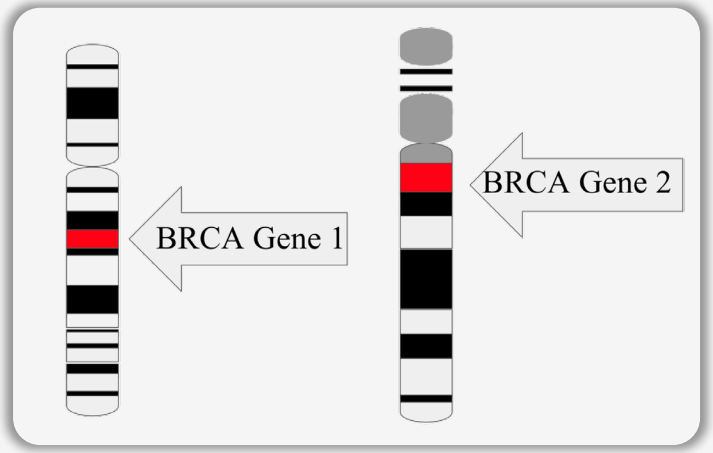
from scientific data and models

Jilles Vreeken | 19 May 2025

Common questions in biology

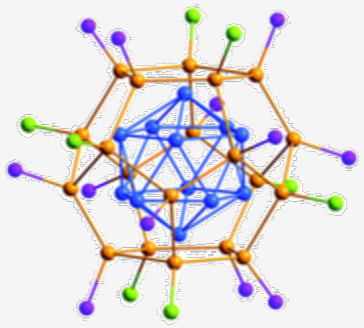
What is the difference between cancerous and benign tissue?

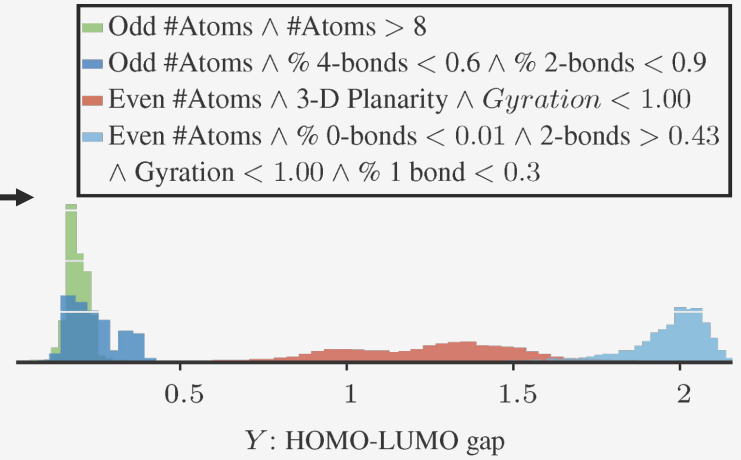




Common questions in physics

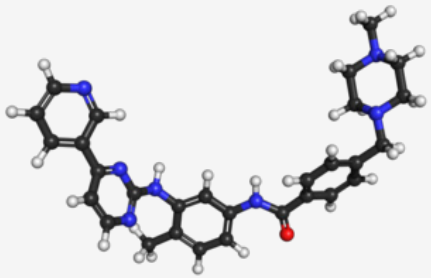
What materials have exceptional conductivity?

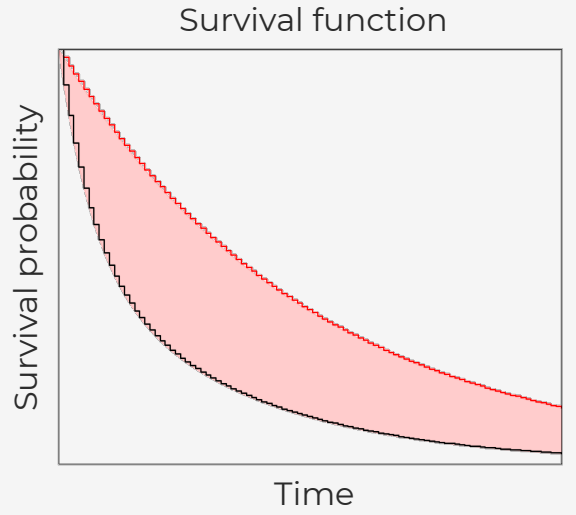




Common questions in medicine

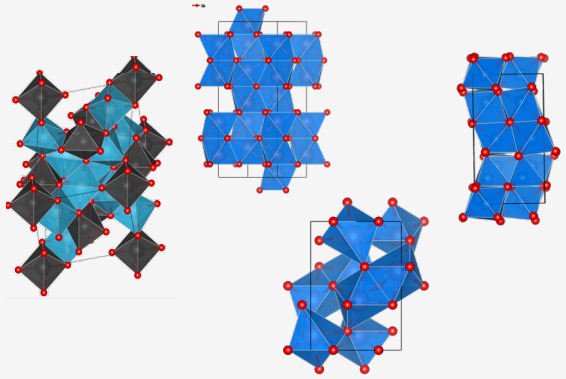
Which people do (not) benefit from a treatment?



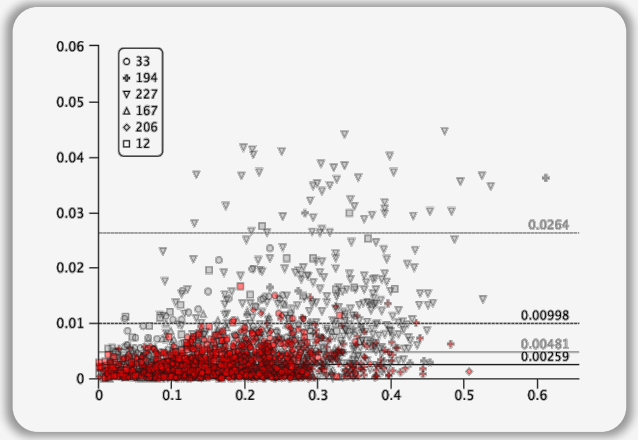


Common questions in reality

What is the domain of applicability of my ML model?

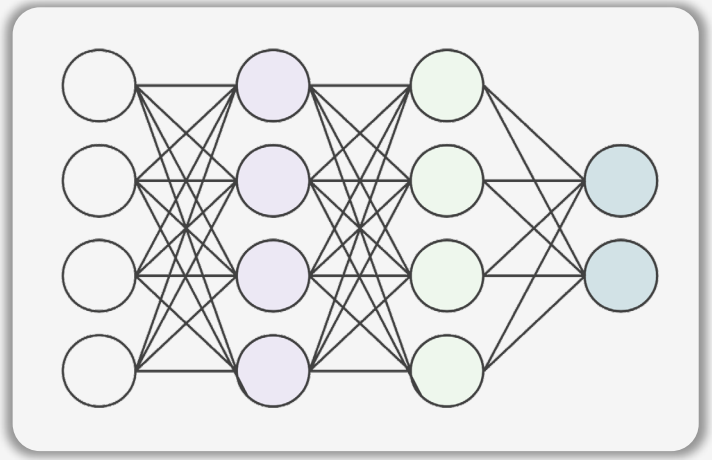


Determine conditions on x under which $\hat{f}(x)$ is reliable

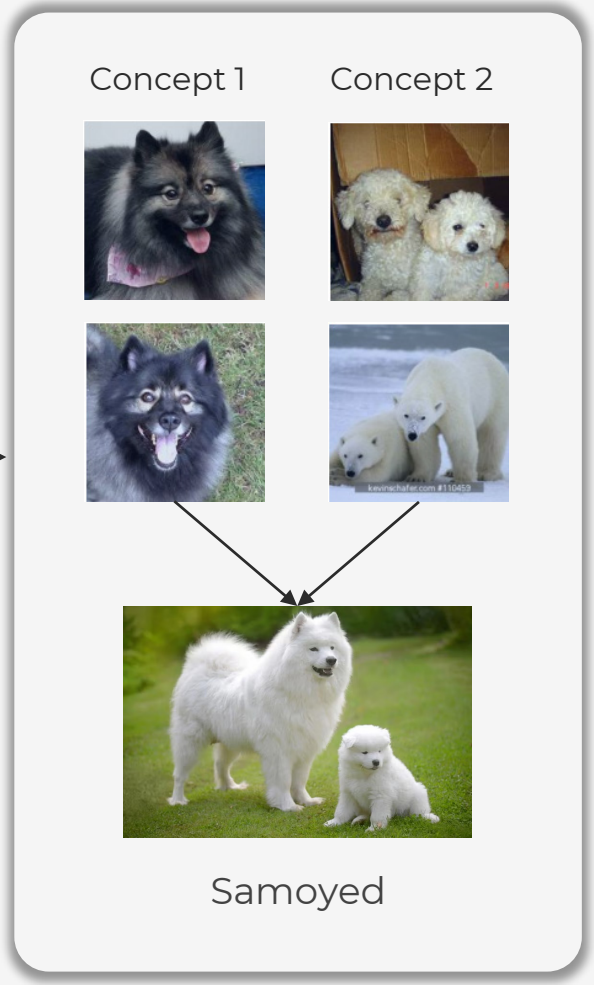


Common questions in interpretability

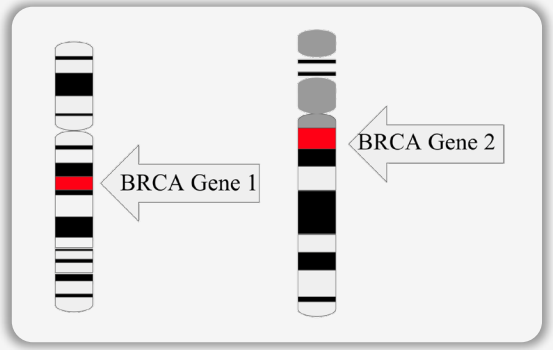
What internal reasoning influences predictions?

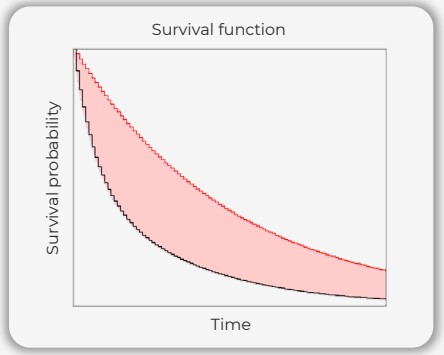


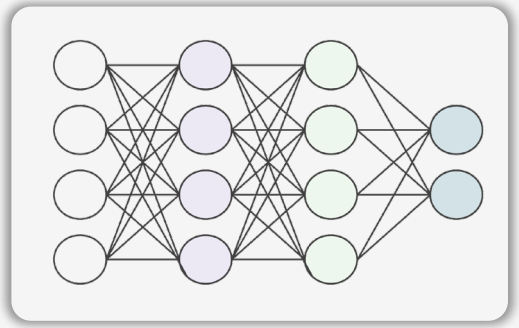
Extract interpretable concepts



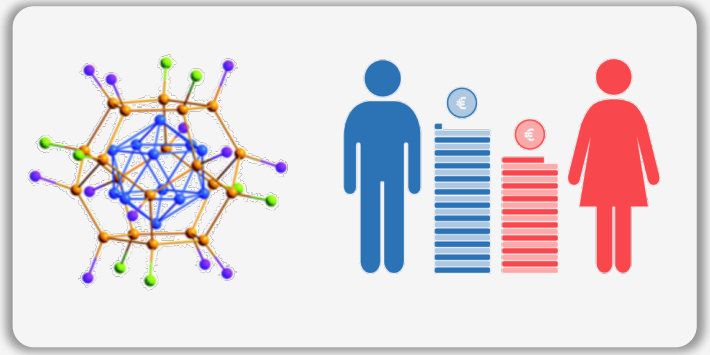
Common questions in ML

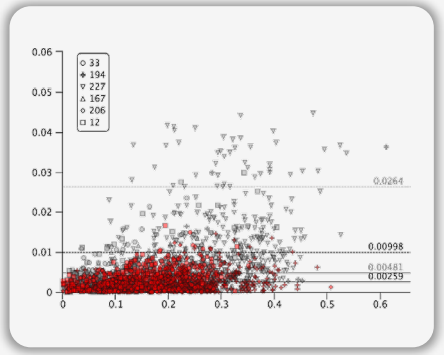




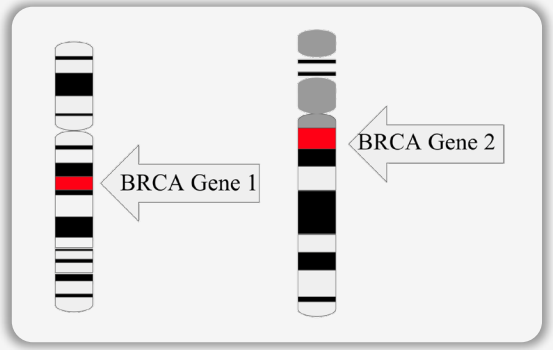


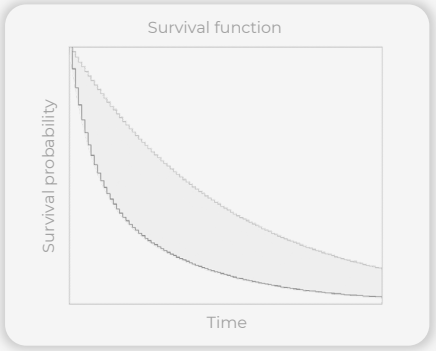
Data ← → Models

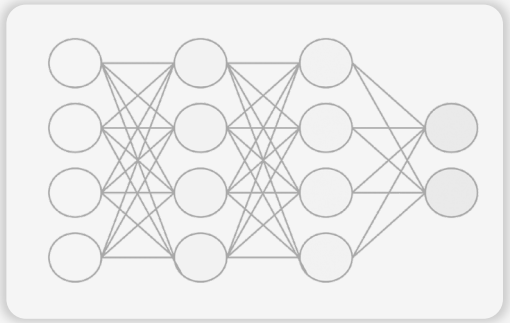




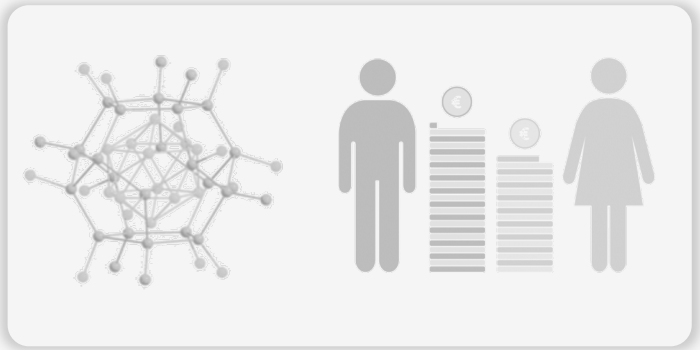
Common questions in ML

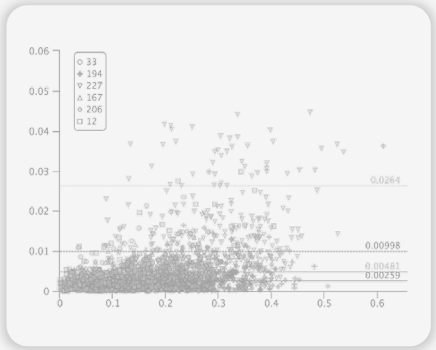






Data ← → Models

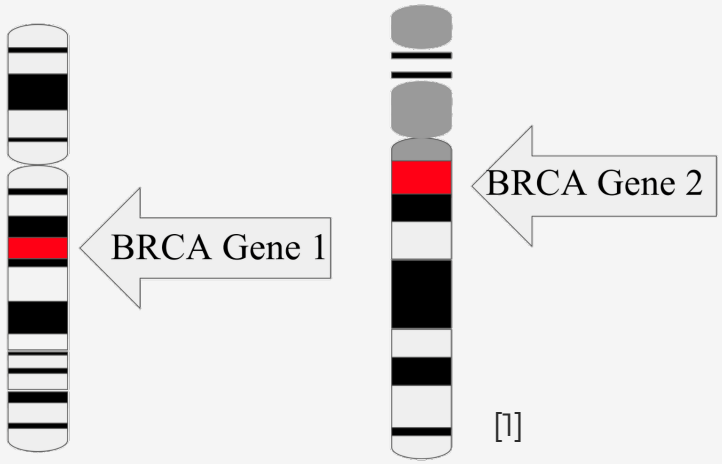




Understanding genetic data


Breast Cancer Data (BRCA)

- (Binarized) sequences of tissue samples
- Labeled with types of BRCA
- Very high-dimensional, low #samples

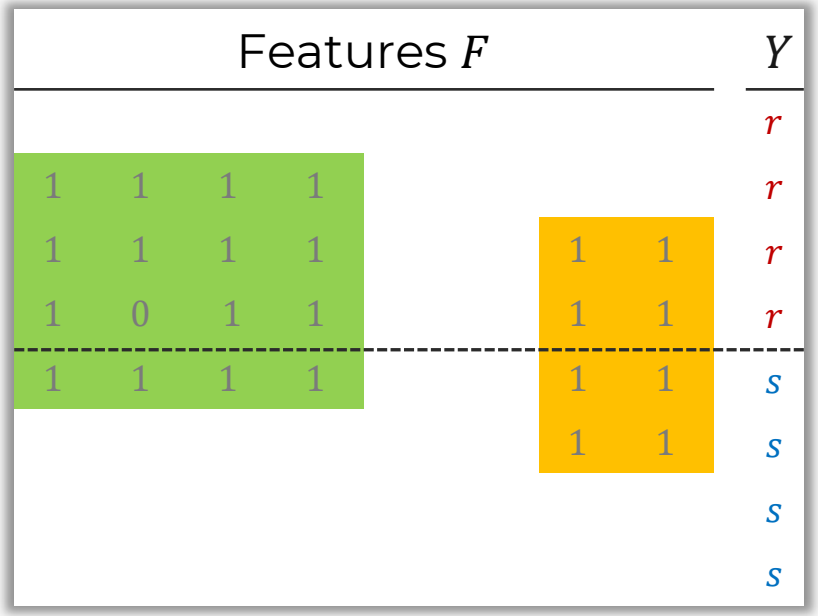


[1]

Goal: Find class-specific interactions of genes

Class-specific pattern:

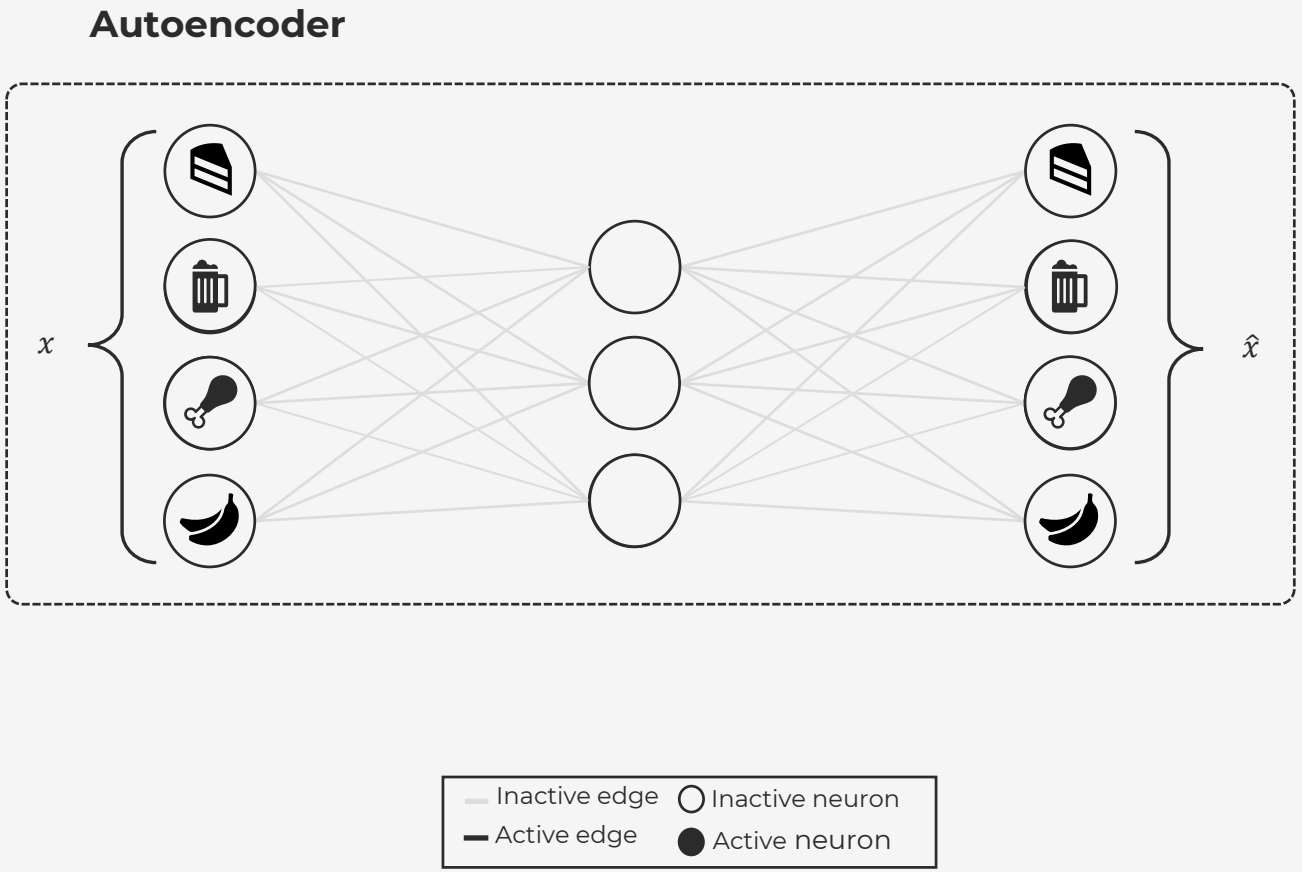
1. Occurs more often in class k
2. Is most predictive for class k



(Walter et al. 2024)

BiNaps – In a nutshell

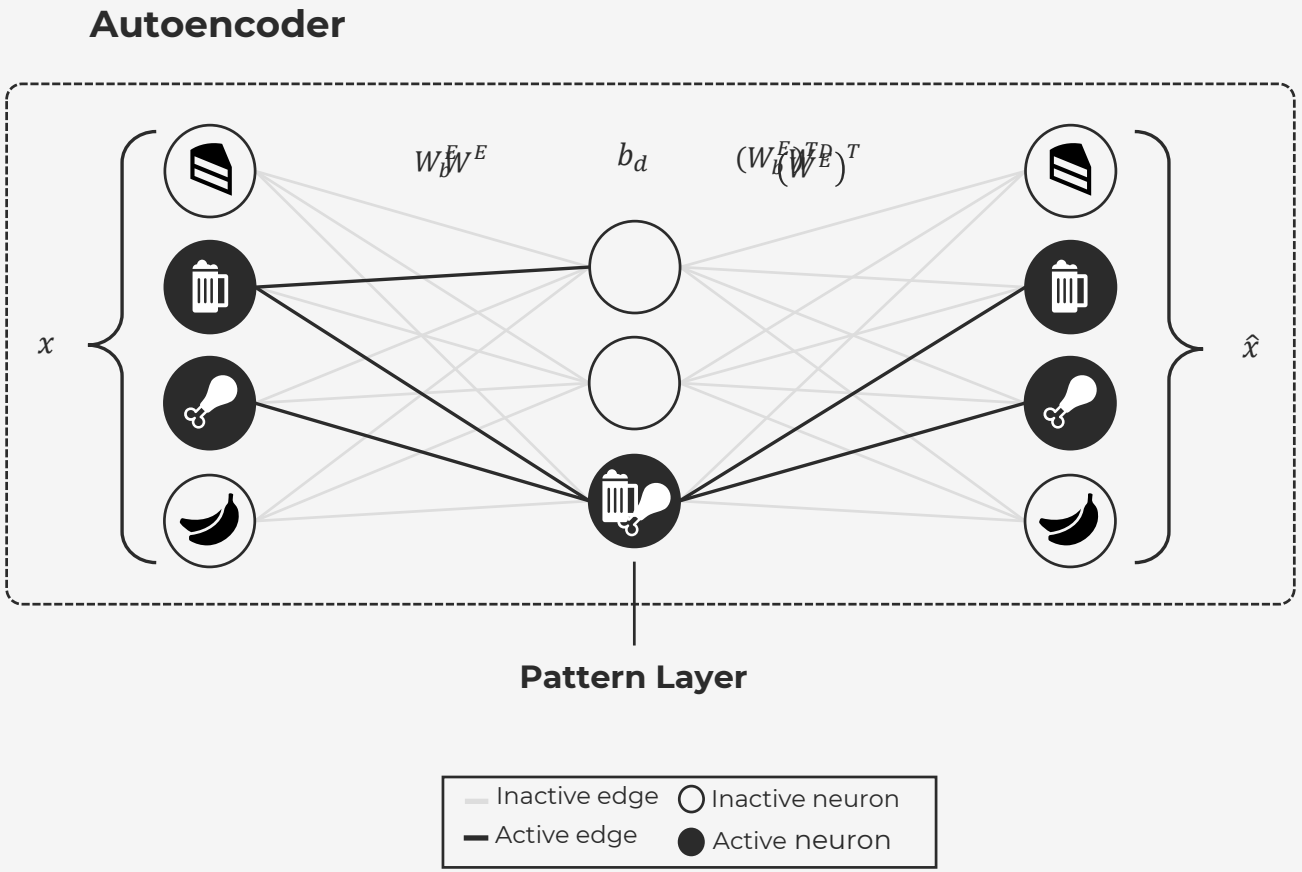
Main Idea: Compression → Patterns



(Fischer et al. 2021)

BiNaps – In a nutshell

Main Idea: Compression → Patterns

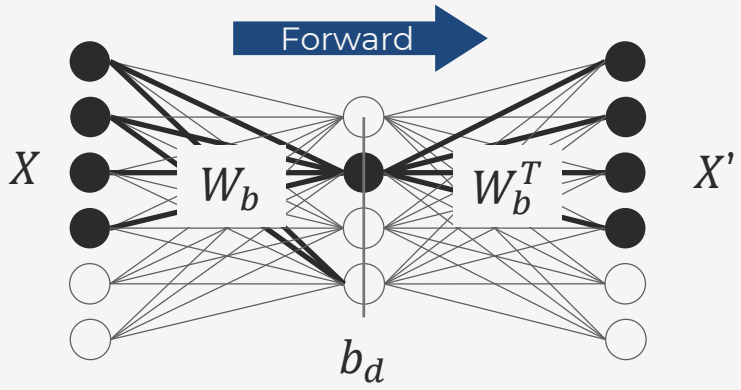


(Fischer et al. 2021)

BiNaps – In a nutshell

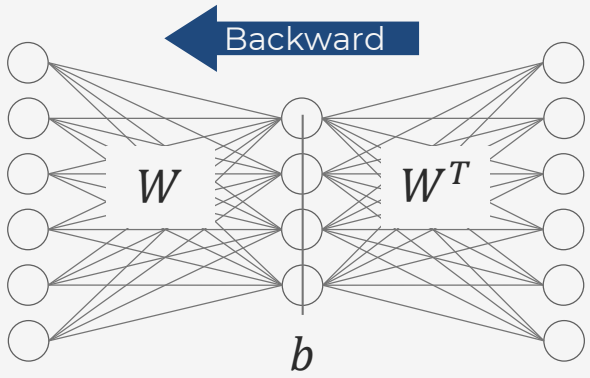
For $X \in D$

1. reconstruct



2. compute loss $L_{W,b}(X)$

3. backpropagate $\frac{dL_{W,b}(X)}{dW}$ $\frac{dL_{W,b}(X)}{db}$



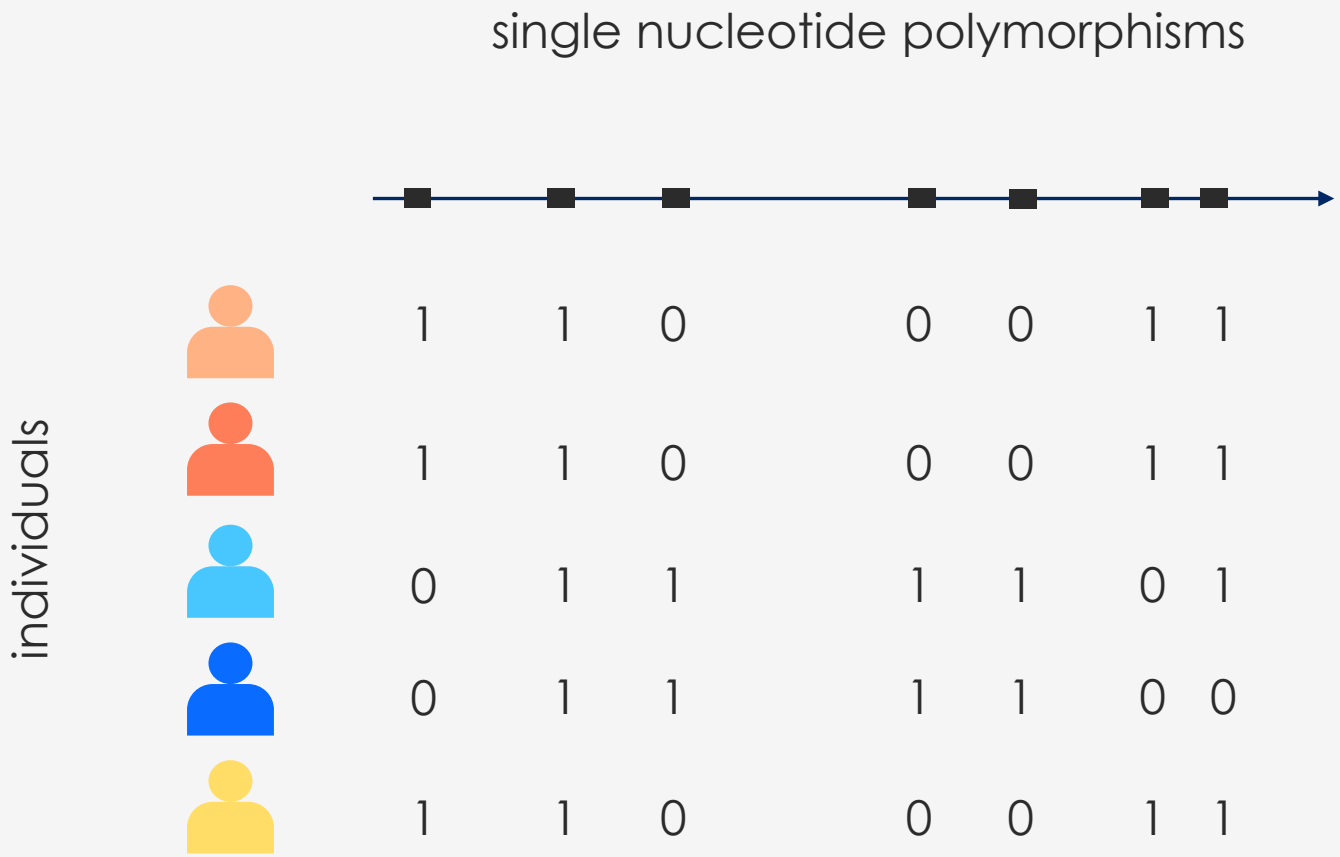
4. Clamp W, b and get binarized W_b, discretized b_d

PyTorch

Use ADAM for optimization

(Fischer et al. 2021)

Exploring human variation.



1000 Genomes project

Measuring variations across human populations (~2500 individuals)

Here we look at single nucleotide variations in genes of autosomes (~228k variants)

Exploring human variation.

What we find:

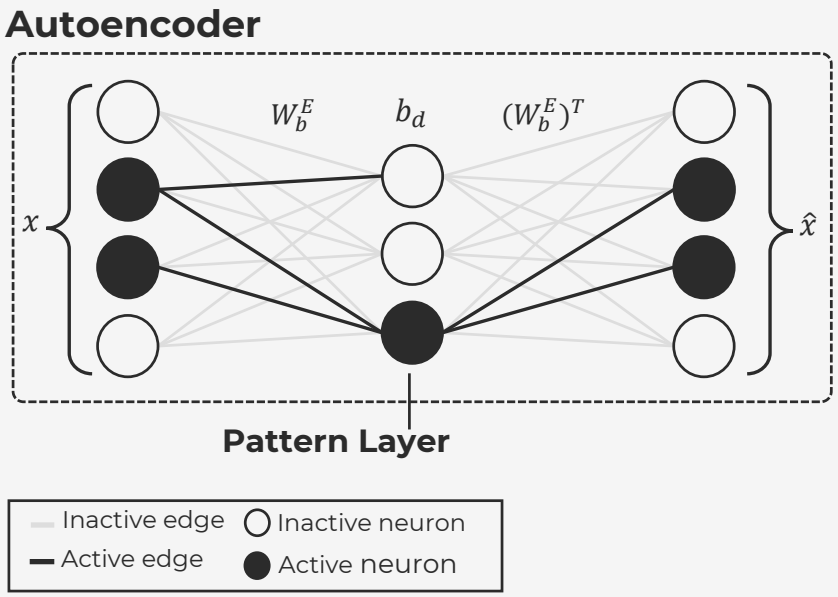
- Blocks of consecutive variation

- Patterns of alternate phasing across variants
  (in some individuals pattern 0|1, 0|1, 0|1 in others 1|0, 1|0, 1|0)

- Linking to genes:
  - Pattern of developmental genes together with genes with unknown role
  - Pattern of genes crucial for the ribosomal complex
  - Pattern of variants associated with type 2 diabetes, and unknown variant

BiNaps – In a nutshell

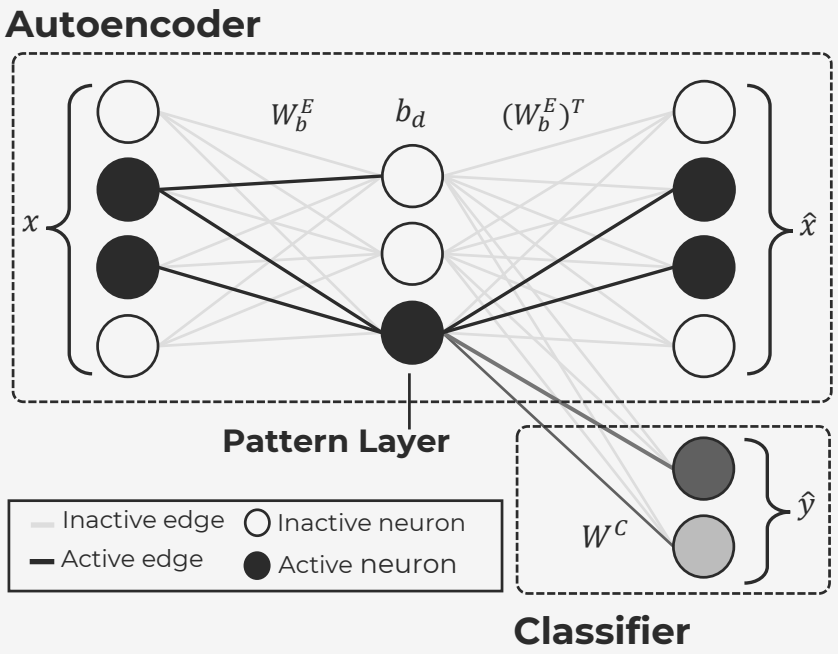
Main Idea: Compression → Patterns



(Walter et al. 2024)

DiffNaps – In a nutshell


Main Idea: Compression + Classification → Class-specific pattern



Forward Pass

1. Binary input x
2. Binarize weight matrix W_b^E
3. Compute hidden activation z

$$z = \lambda_E(W_b^E\, x)\,,$$

where λ_E is a binary activation function.

4. Compute reconstruction $\hat{x}$

$$\hat{x} = \lambda_D((W_b^E)^T z)$$

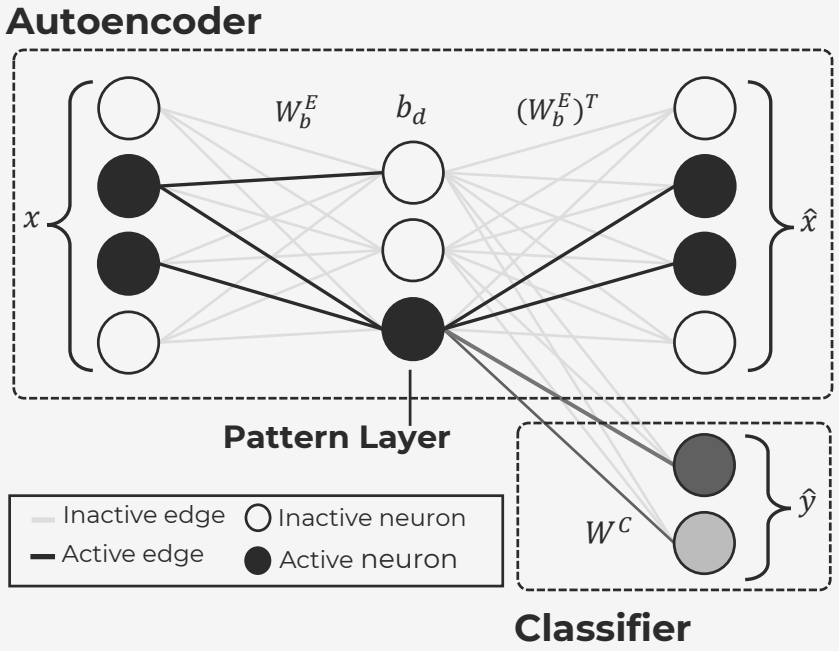
5. Compute classification $\hat{y}$

$$\hat{y} = \text{softmax}(W^C z)$$

Jointly optimize reconstruction and classification accuracy

(Walter et al. 2024)

DiffNaps – Extract patterns

Continuous **Discrete**

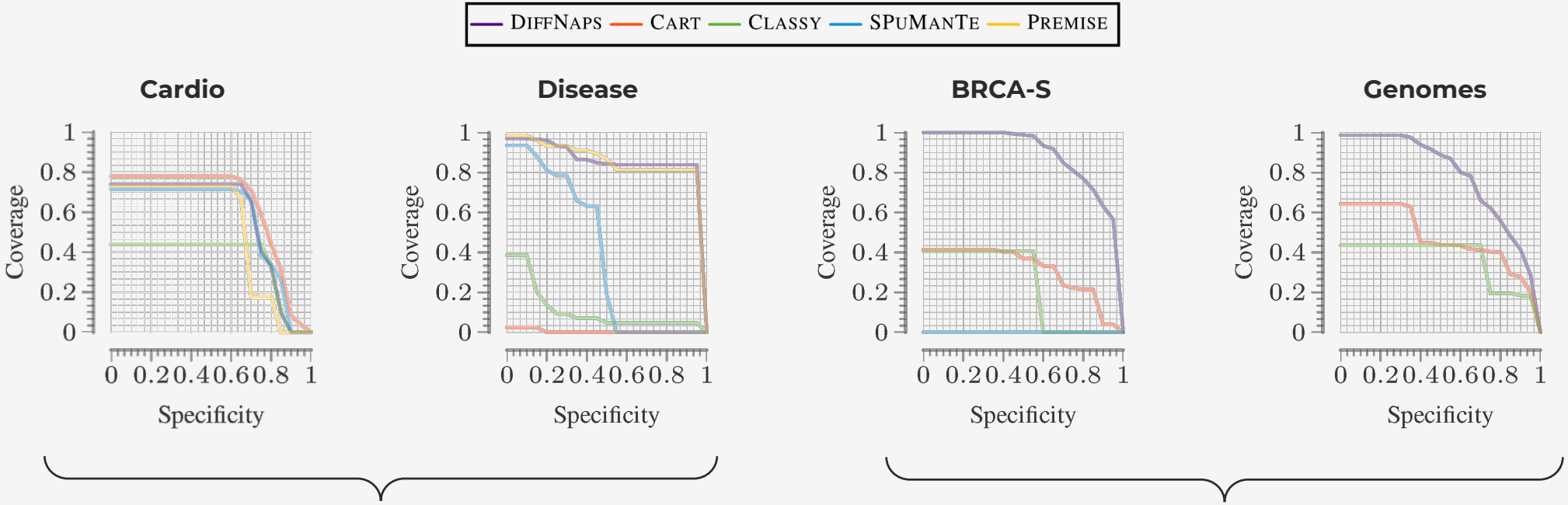
$$W^E = \begin{bmatrix} 0.02 & 0.87 & 0.02 & 0.87 \\ 0.8 & 0.01 & 0.04 & 0.9 \\ 0.0 & 0.99 & 0.8 & 0.09 \end{bmatrix} \xrightarrow{\mathcal{B}\tau_E(\cdot)} W_b^E = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

$$W^C = \begin{bmatrix} 0.07 & 0.01 & 0.9 \\ 0.02 & 0.9 & 0.2 \end{bmatrix} \xrightarrow{\mathcal{B}_{\tau_C}(\cdot)} W_b^C = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

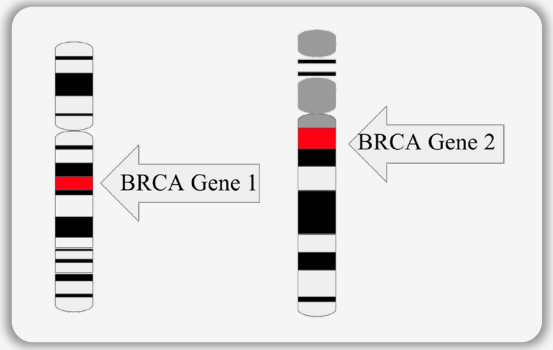
Class-specific patterns: $P^1 = \{\{2,3\}\}$ resp. $P^2 = \{\{1,4\}\}$

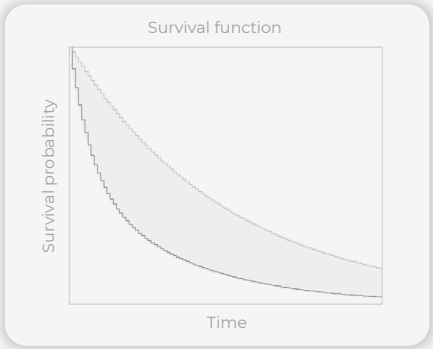
DiffNaps – Real World Data

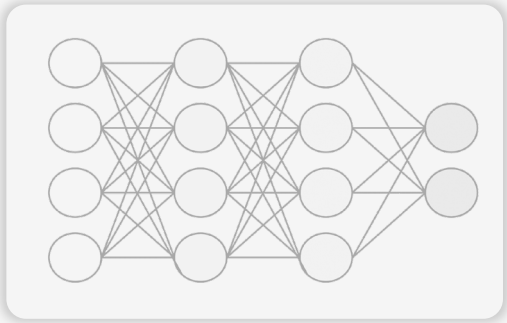

- **Cardio** ($m = 45$): Patient data annotated with heart disease
- **Disease** ($m = 131$) : Various symptoms annotated with disease
- **BRCA-S** ($m = 20k$): ncRNA annotated with cancer type → Found biologically **relevant** patterns!
- **Genomes** ($m = 225k$): DNA annotated with ethnicity



Common questions in ML

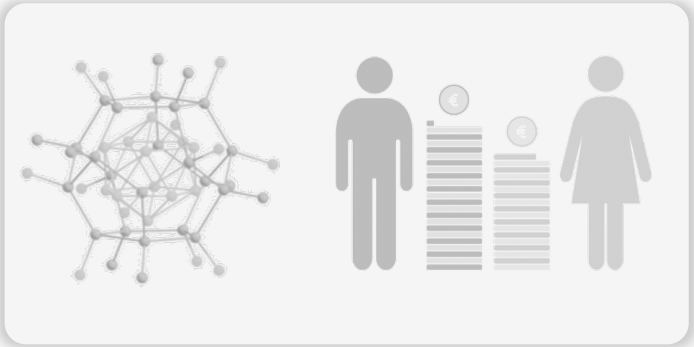


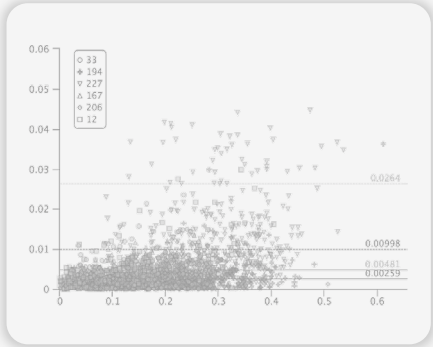




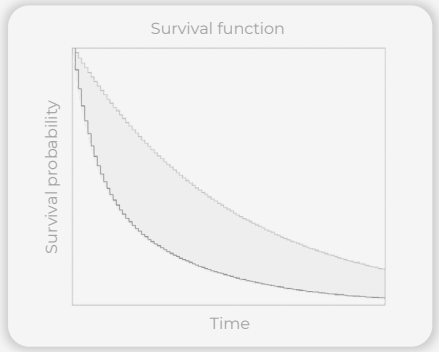
Data

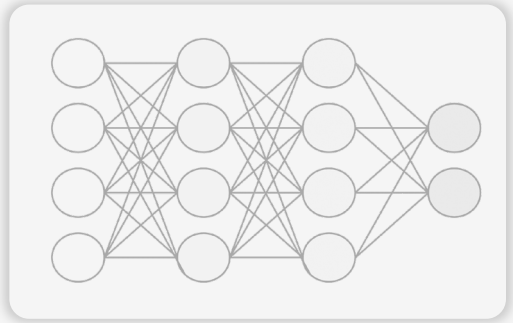
Models



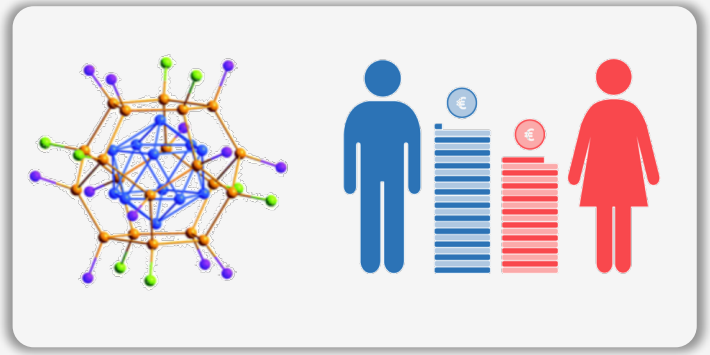


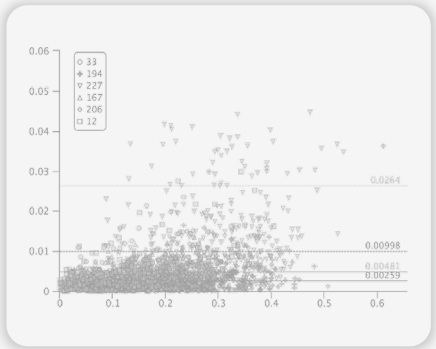
Common questions in ML





Data ← → Models

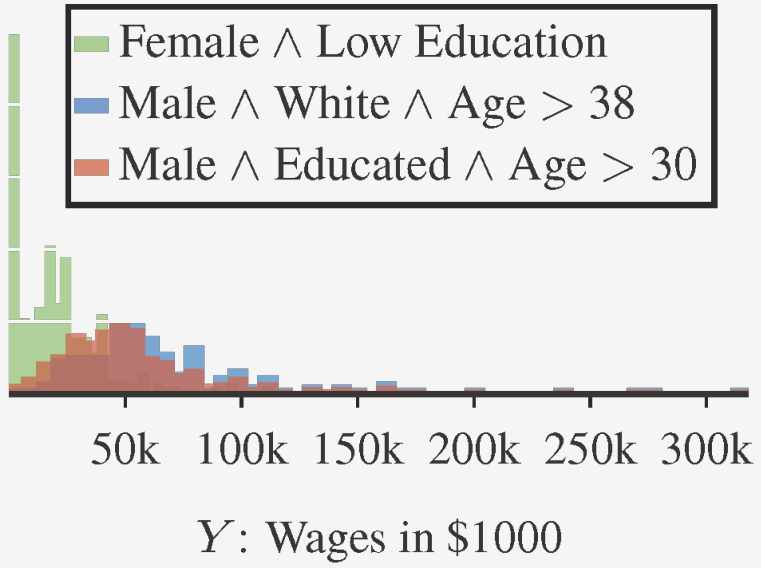




Discovering exceptional subgroups

Census Data

Sex	Height	Ethn	Education	Age	Income
♀	168	White	12	72	17k
♂	163	White	11	55	23k
♀	160	White	**5**	62	**1k**
♂	188	White	16	38	63k
♀	165	White	**9**	45	**4k**
♂	172	White	12	78	71k
♀	180	White	**8**	74	**1k**



Task – Subgroup Discovery:

1. Find **exceptional** subgroups
2. With an **interpretable** description

Traditional Subgroup Discovery

Subgroup Discovery

1. Pre-discretize features
   → **Likely to miss subgroups**

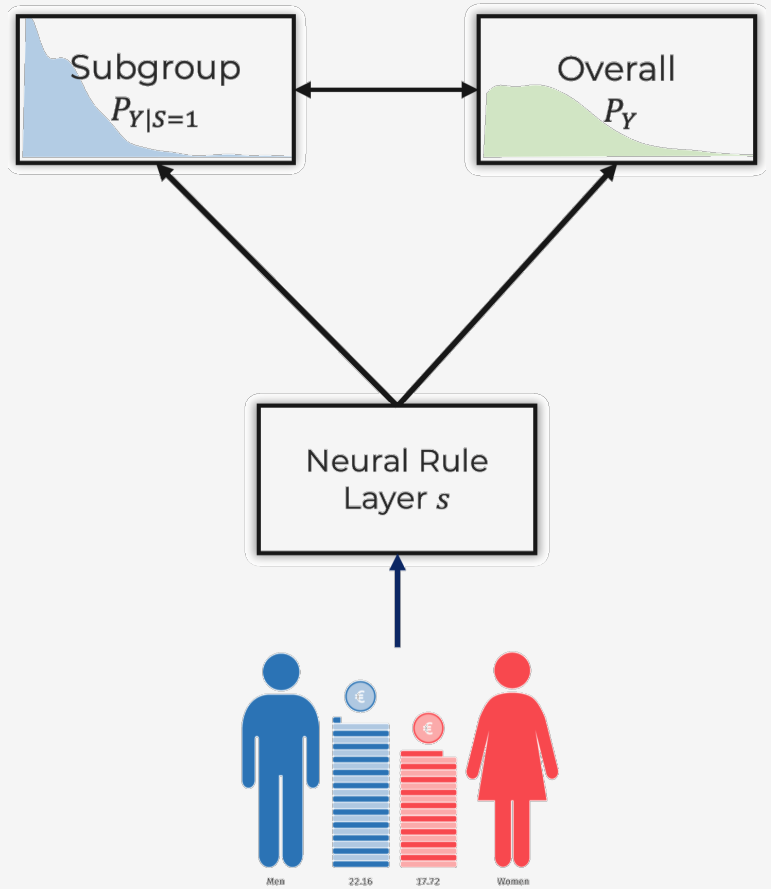
2. Strong assumptions on target
   → **Specialized methods needed**

3. Combinatorial optimization
   → **Does not scale**

Modern Subgroup Discovery

Subgroup Discovery	SyFlow
1. Pre-discretize features → **Likely to miss subgroups**	1. Learn predicates end-to-end → **Accurate Discretization**
2. Strong assumptions on target → **Specialized methods needed**	2. Use Normalizing Flows (NFs) → **No assumptions**
3. Combinatorial optimization → **Does not scale**	3. Maximize KL-divergence → **Highly general**
	4. Continous optimization → **Highly scalable**

$$\max KL(P_{Y|S=1} \mid P_Y)$$



(Xu et al. 2024)

SyFlow – Neural Rule Layer I

Goal: Find an crisp interpretable description

$$\sigma(x) = \neg Smoker \wedge 44 < Age < 64$$

Ingredients:

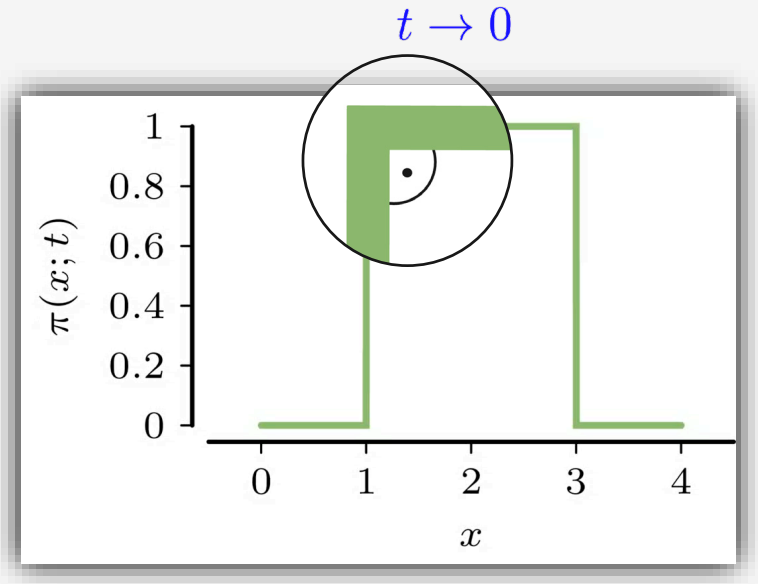
1. Differentiable binning predicate

$$\hat{\pi}(x_i; \alpha_i, \beta_i, t) = \frac{e^{\frac{1}{t}(2x_i - \alpha_i)}}{e^{\frac{1}{t}x_i} + e^{\frac{1}{t}(2x_i - \alpha_i)} + e^{\frac{1}{t}(3x_i - \alpha_i - \beta_i)}}$$

- Differentiable analog of:

$$\pi(x_i; \alpha_i, \beta_i) = \begin{cases} 1 & \text{if } \alpha_i < x_i < \beta_i \\ 0 & \text{otherwise} \end{cases}$$

- Temperature t controls crispness



Theorem 1 *Given its lower and upper bounds $\alpha_i, \beta_i \in \mathbb{R}$, the soft predicate of Eq. (1) applied on $x \in R$ converges to the crisp predicate that decides whether $x \in (\alpha, \beta)$,*

$$\lim_{t \to 0} \hat{\pi}(x_i; \alpha_i, \beta_i, t) = \begin{cases} 1 & \text{if } \alpha_i < x_i < \beta_i \\ 0.5 & \text{if } x_i = \alpha_i \vee x_i = \beta_i \\ 0 & \text{otherwise} \end{cases} .$$

(Yang et. al. 2018)

SyFlow – Neural Rule Layer II

Ingredients:

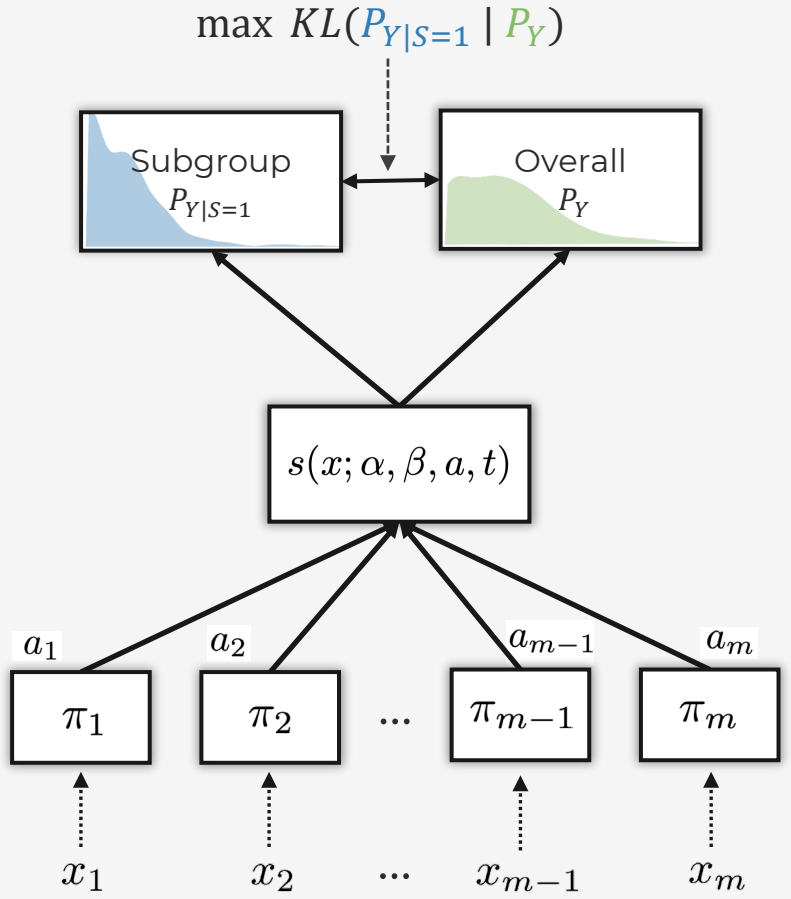
1. Differentiable binning predicate
2. Differentiable logical *AND*
   - Harmonic means behaves like an *AND*
     1. If one π_i is zero then evaluate to false
     2. If all π_i are one then evaluate to true
   - Implicit feature selection with a_i

Optimization

1. Learn the overall distribution P_Y
2. Learn the subgroup distribution $P_{Y|S=1}$
3. Optimize classifier weights and bins

} Repeat for N steps

4. Output: Subgroup

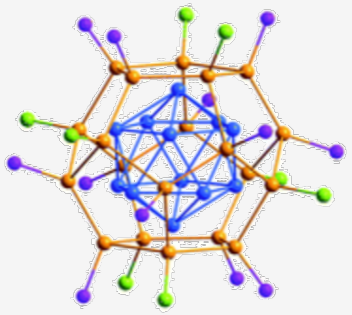


Fully differentiable!

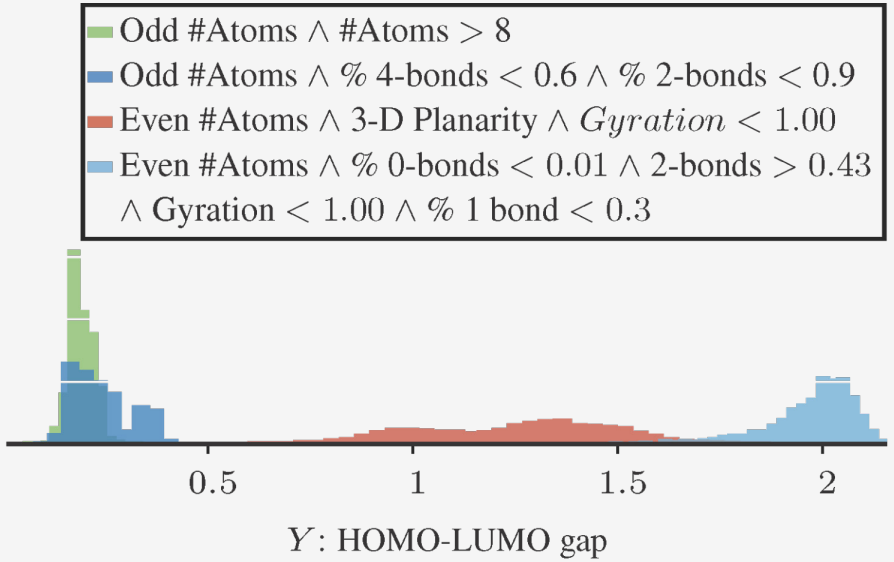
(Xu et al. 2024)

Experiments – Materials Science

Gold Nanoclusters



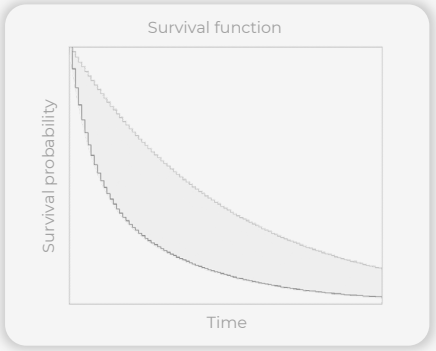
- Atom-based features
- Bond-based features
- Shape-based features

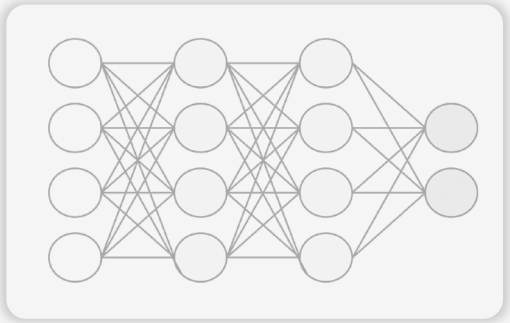


Target: HOMO-LUMO gap
~ stability and conductivity

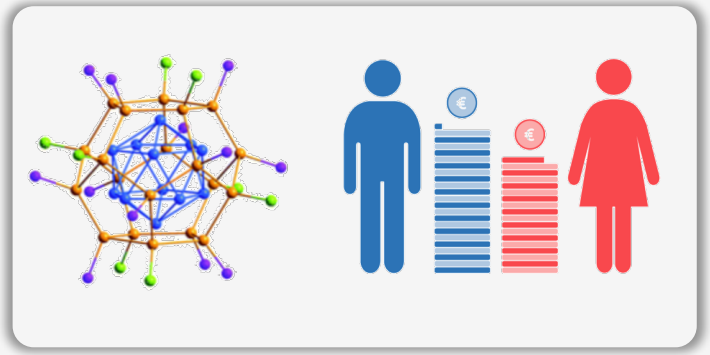
(Xu et al. 2024)

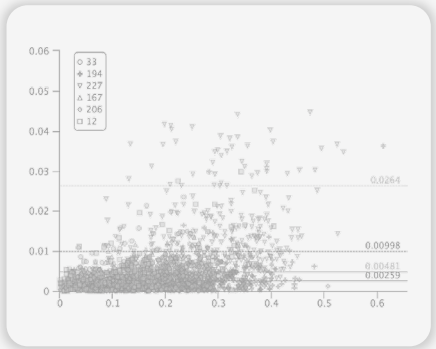
Common questions in ML



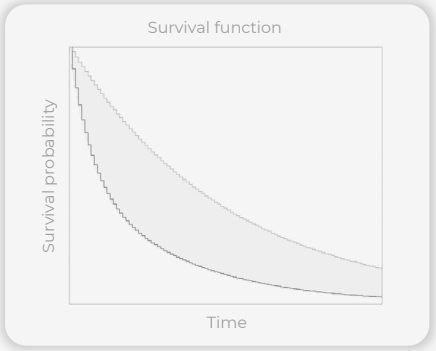


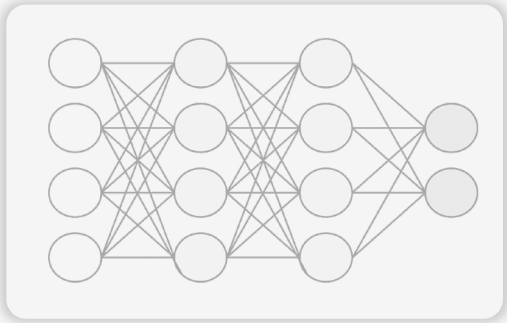
Data ← → Models



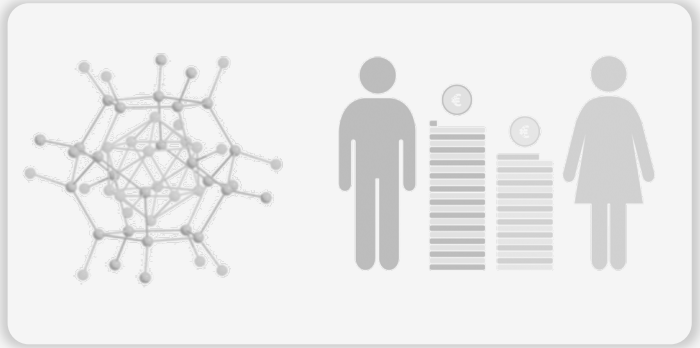


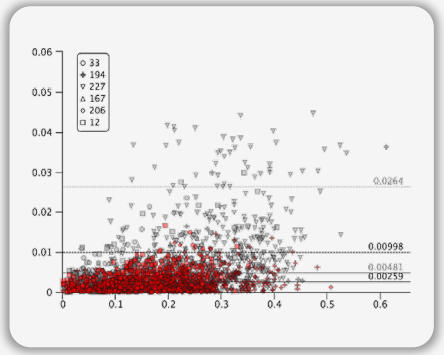
Common questions in ML





Data ⟵ ⟶ Models



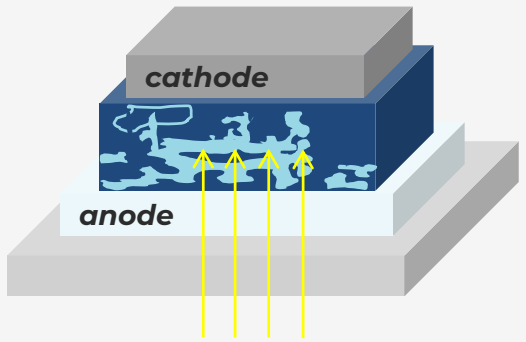


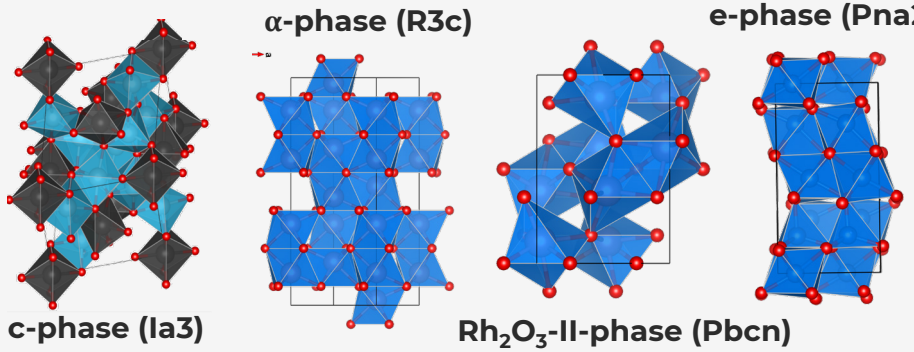
Model diagnostics

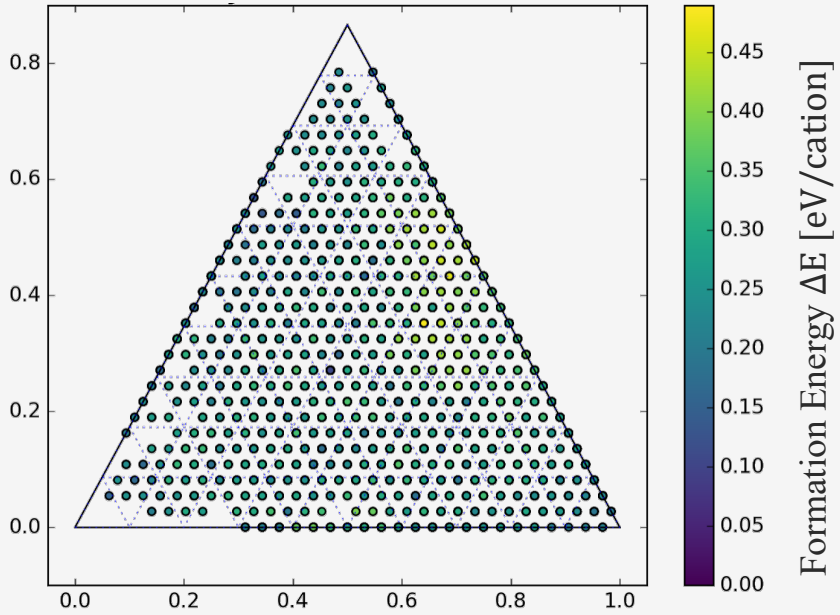
Population

$$P = \left\{c : c \in \left(\mathrm{In}_x \mathrm{Ga}_y \mathrm{Al}_y\right)_2 \mathrm{O}_3\right\}$$

Photovoltaics







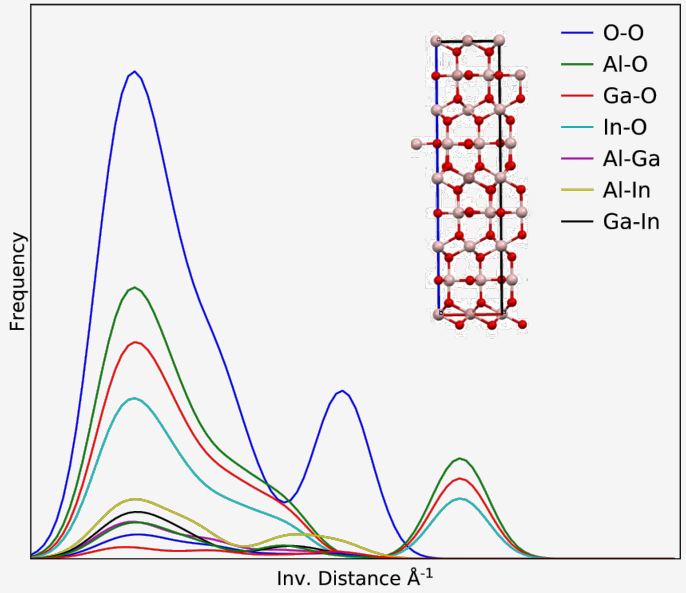
Model diagnostics

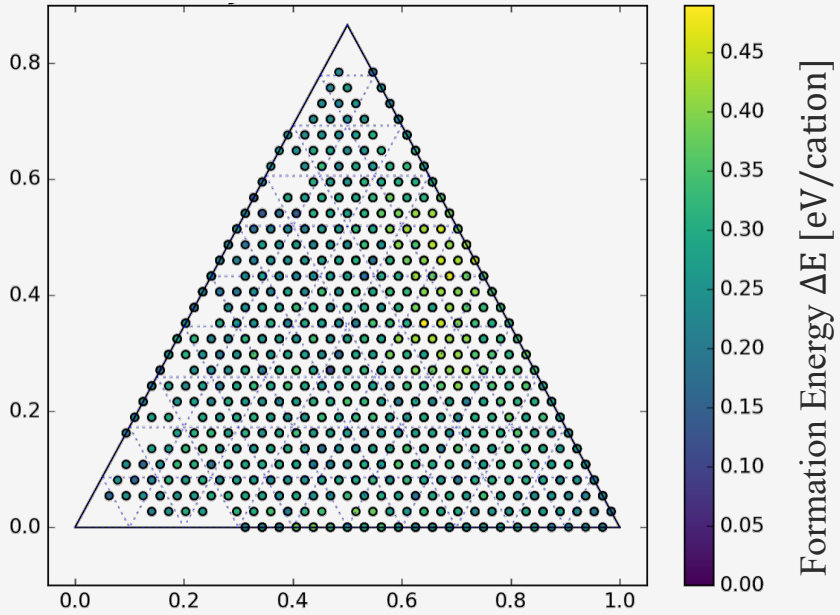
Population

$$P = \left\{c : c \in \left(\mathrm{In}_x \mathrm{Ga}_y \mathrm{Al}_y\right)_2 \mathrm{O}_3\right\}$$

Target

$y = |\Delta E - \tilde{f}_{\Delta E}(c)|$ of regression model $\tilde{f}_{\Delta E}$





Arbitrary model you want to use for predicting

(Hua and Rupp, 2018)

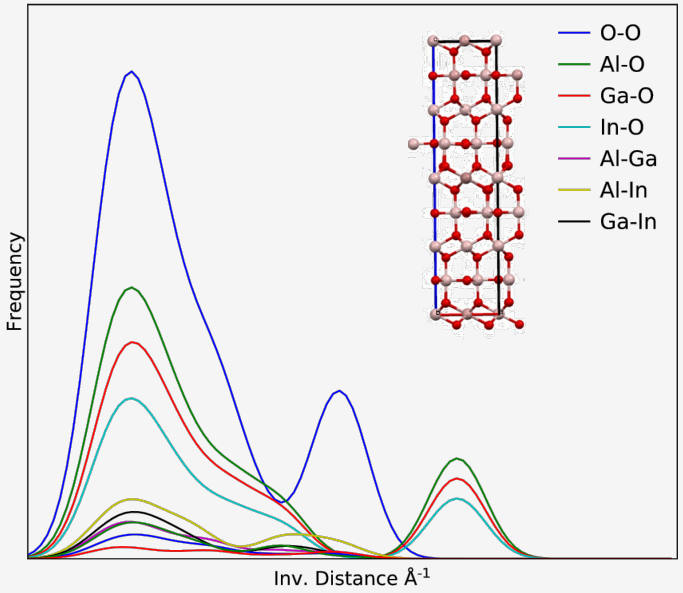
Model diagnostics

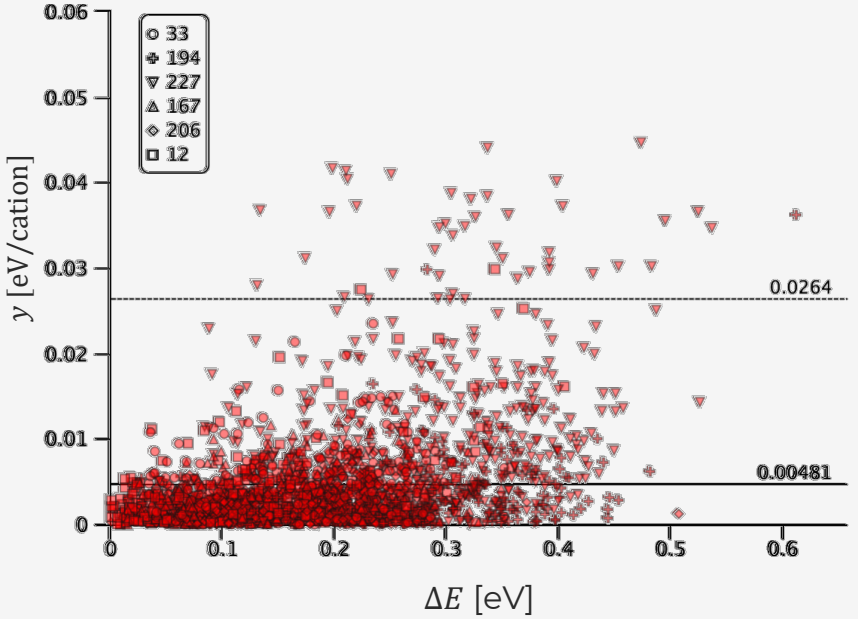
Population

$$P = \left\{c : c \in \left(\mathrm{In}_x \mathrm{Ga}_y \mathrm{Al}_y\right)_2 \mathrm{O}_3\right\}$$

Target

$y = |\Delta E - \tilde{f}_{\Delta E}(c)|$ of regression model $\tilde{f}_{\Delta E}$





Model has **low average error**
but its predictions are
generally unreliable

(Hua and Rupp, 2018)

Model diagnostics

Population

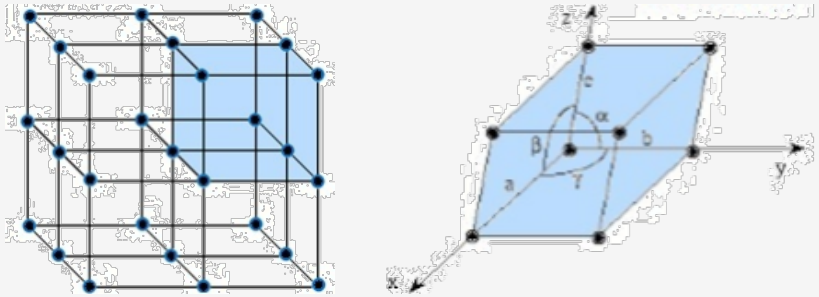
$$P = \left\{c : c \in \left(\mathrm{In}_x \mathrm{Ga}_y \mathrm{Al}_y\right)_2 \mathrm{O}_3\right\}$$

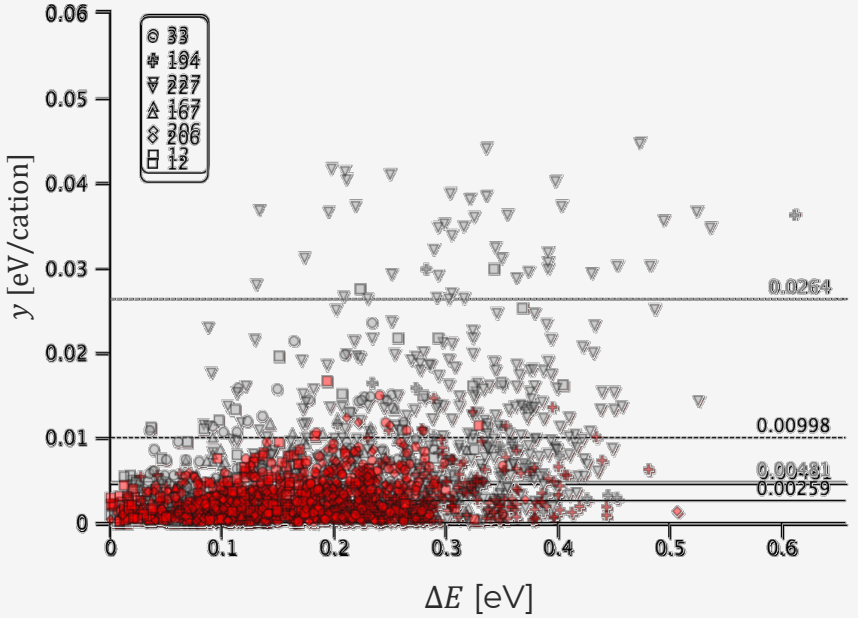
Target

$$y = |\Delta E - \tilde{f}_{\Delta E}(c)| \text{ of regression model } \tilde{f}_{\Delta E}$$

Features

$$x \in \{a, b, c, \alpha, \beta, \gamma, n, \dots\}$$





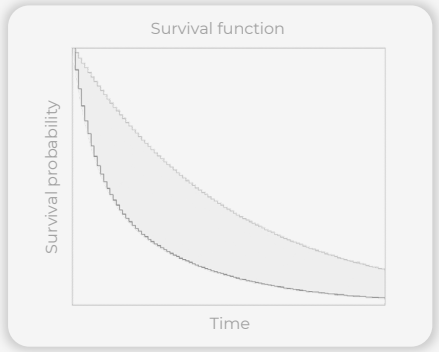
Model is **very reliable** under these conditions

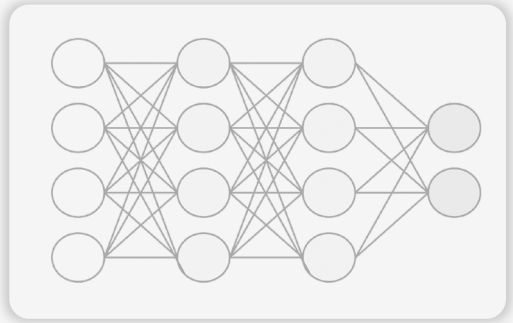
Selector $\sigma(c) \equiv n(c) \geq 60 \wedge \gamma(c) \leq 95 \wedge \alpha(c) \leq 121$

Parameters $\mathrm{cvr}(\sigma) = 0.6$ $\mathrm{eff}(\sigma) = 0.3$ $[\overline{y}(Q) = 0.003 \pm 0.002,\ \overline{y}(S) = 0.005 \pm 0.007]$

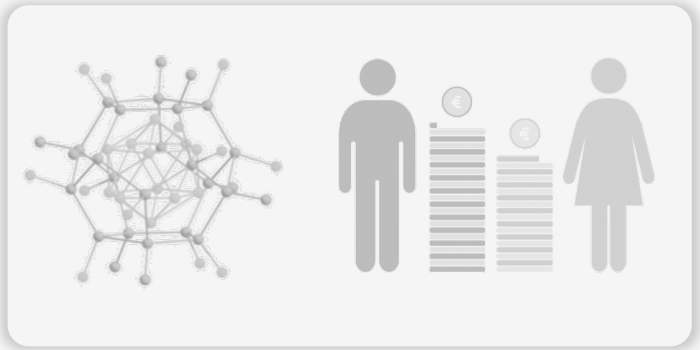
(Sutton et al, 2020)

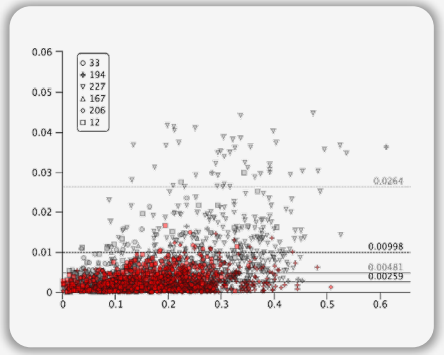
Common questions in ML



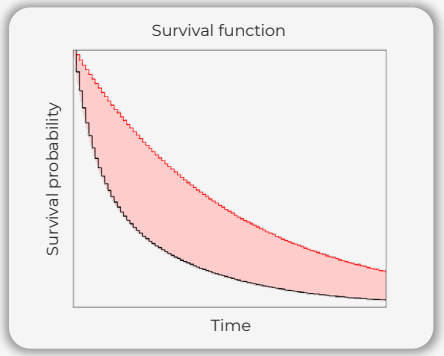


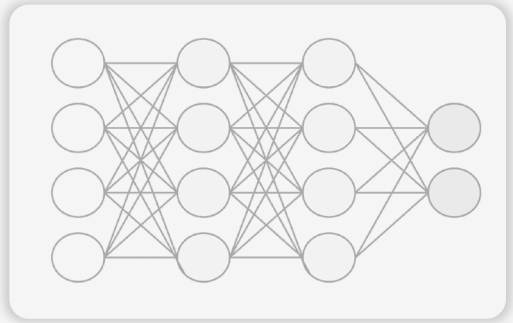
Data ⟵ ⟶ Models



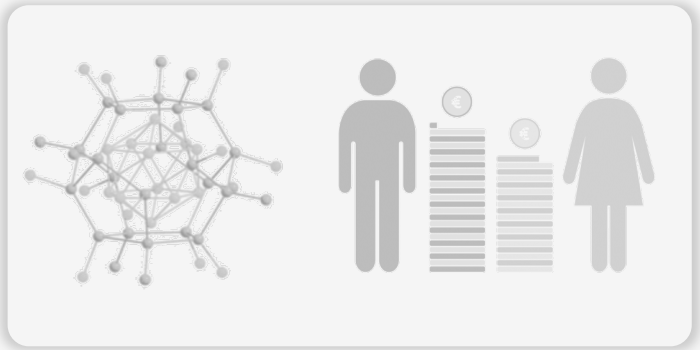


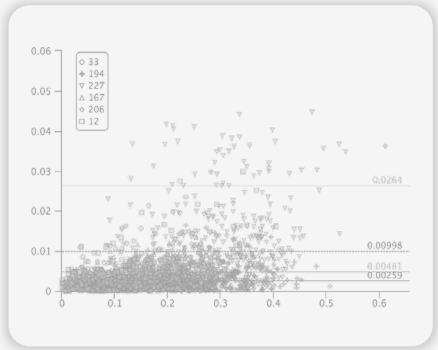
Common questions in ML





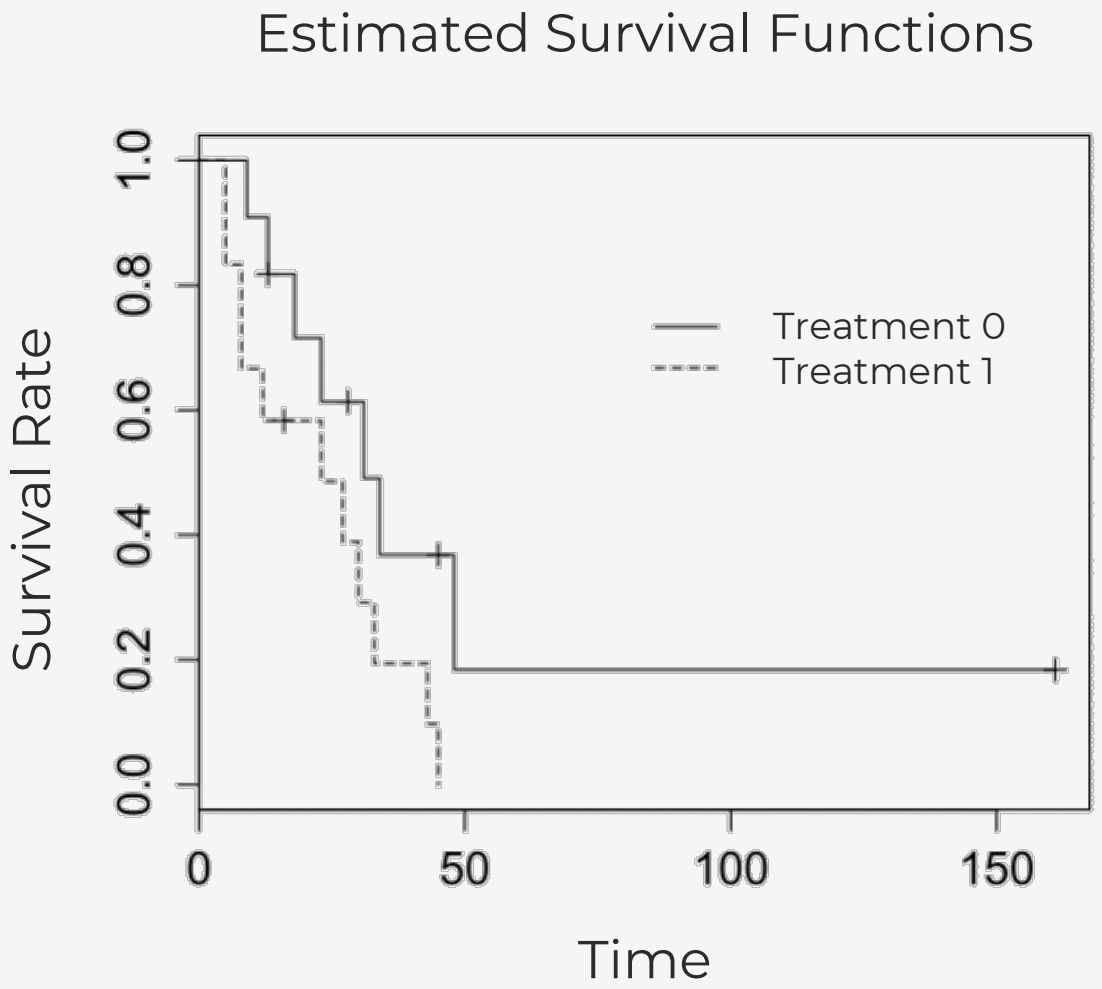
Data ← → Models





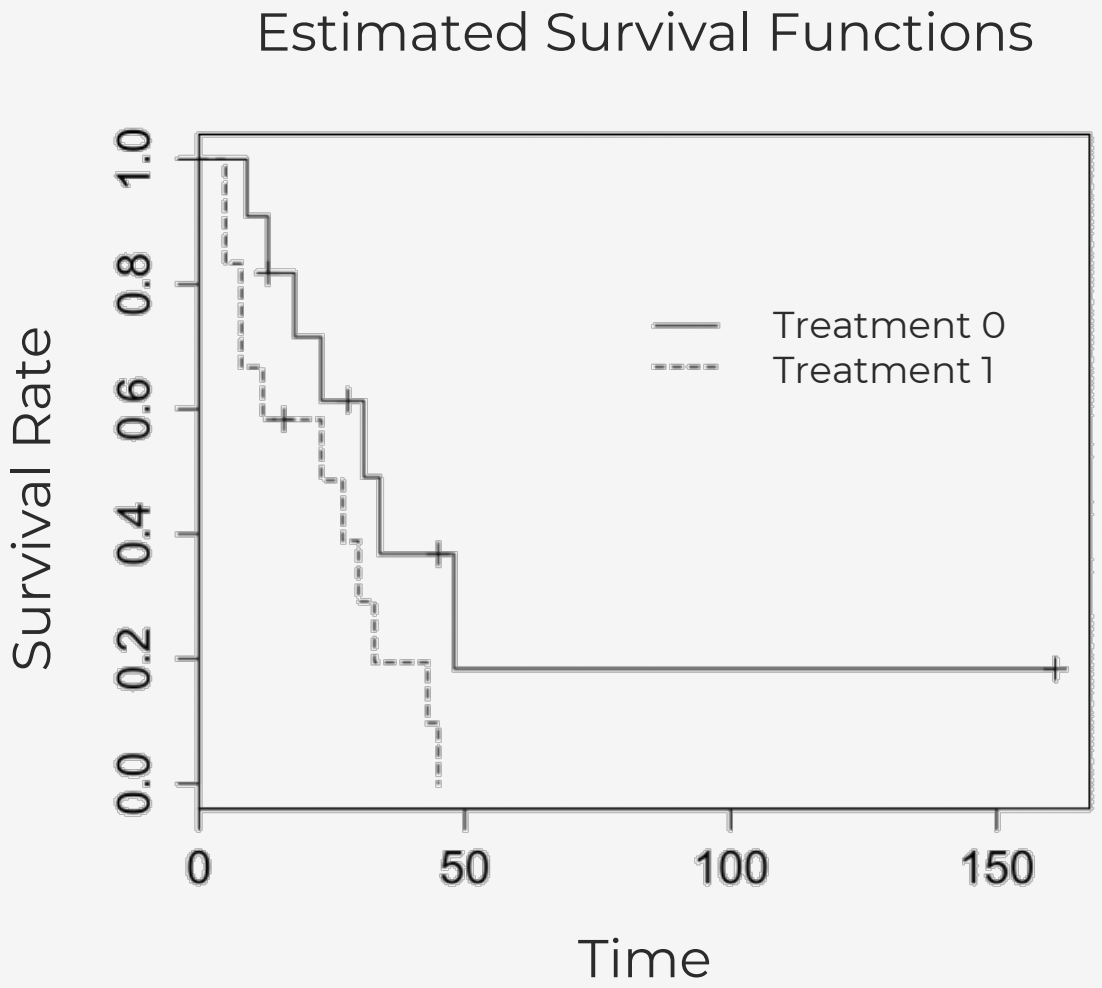
Time-to-event data

Participant	Treatment	Time of death
12	0	5
4	1	18
13	0	5
17	0	16
14	0	8
3	1	13
15	0	8
2	1	13
1	1	9
16	0	12



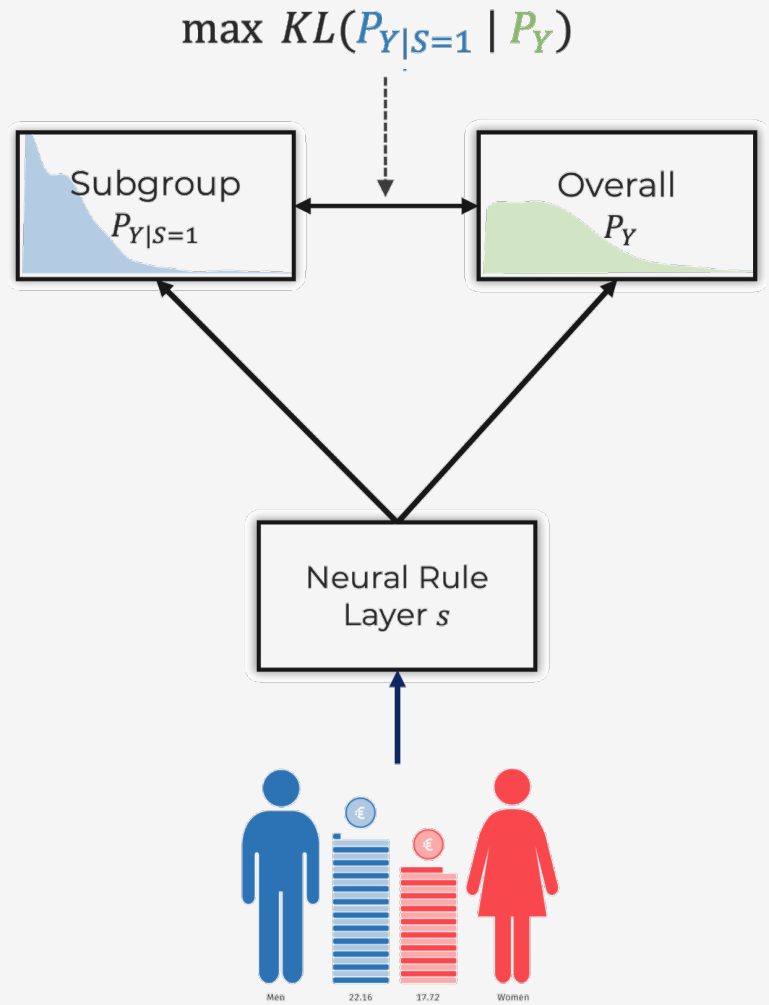
Time-to-event data

Participant	Treatment	Time of death
12	0	5
4	1	18
13	0	survived
17	0	left study
14	0	8
3	1	13
15	0	8
2	1	13
1	1	9
16	0	12



Survival Subgroups

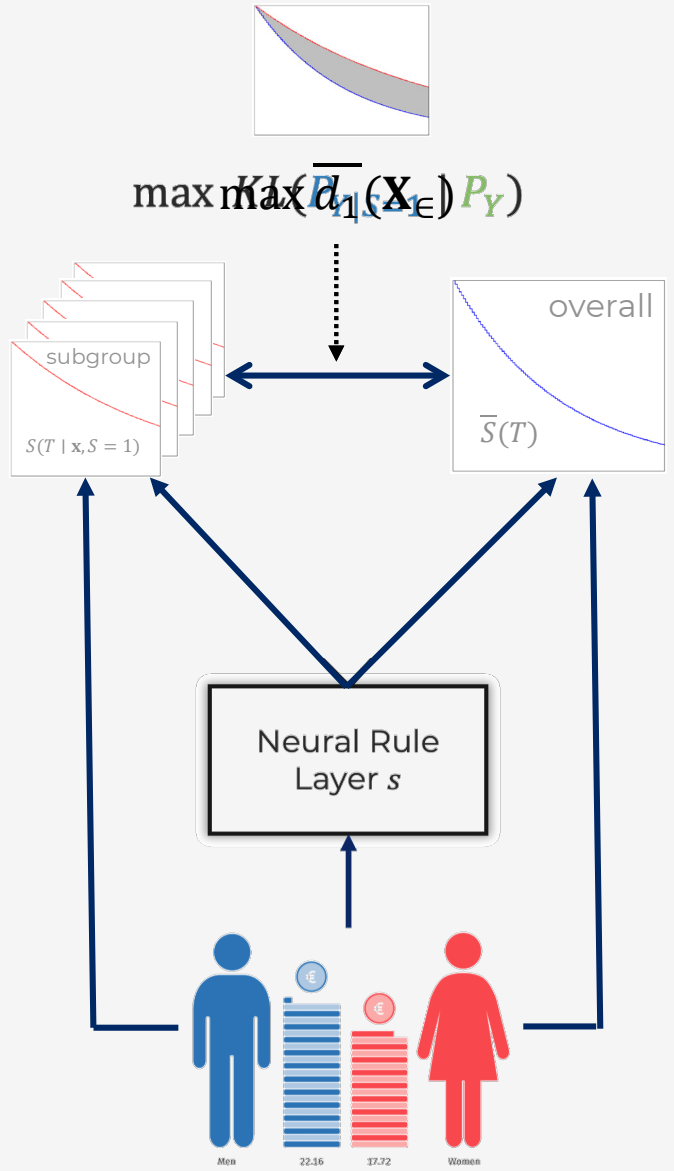
SyFlow	SySurv
1. Learn predicates end-to-end → **Accurate Discretization**	1. Learn predicates end-to-end → **Accurate Discretization**
2. Continous optimization → **Highly scalable**	2. Continous optimization → **Highly scalable**
3. Use Normalizing Flows (NFs) → **Ignores censoring**	



(Al Rahwanji, work in progress)

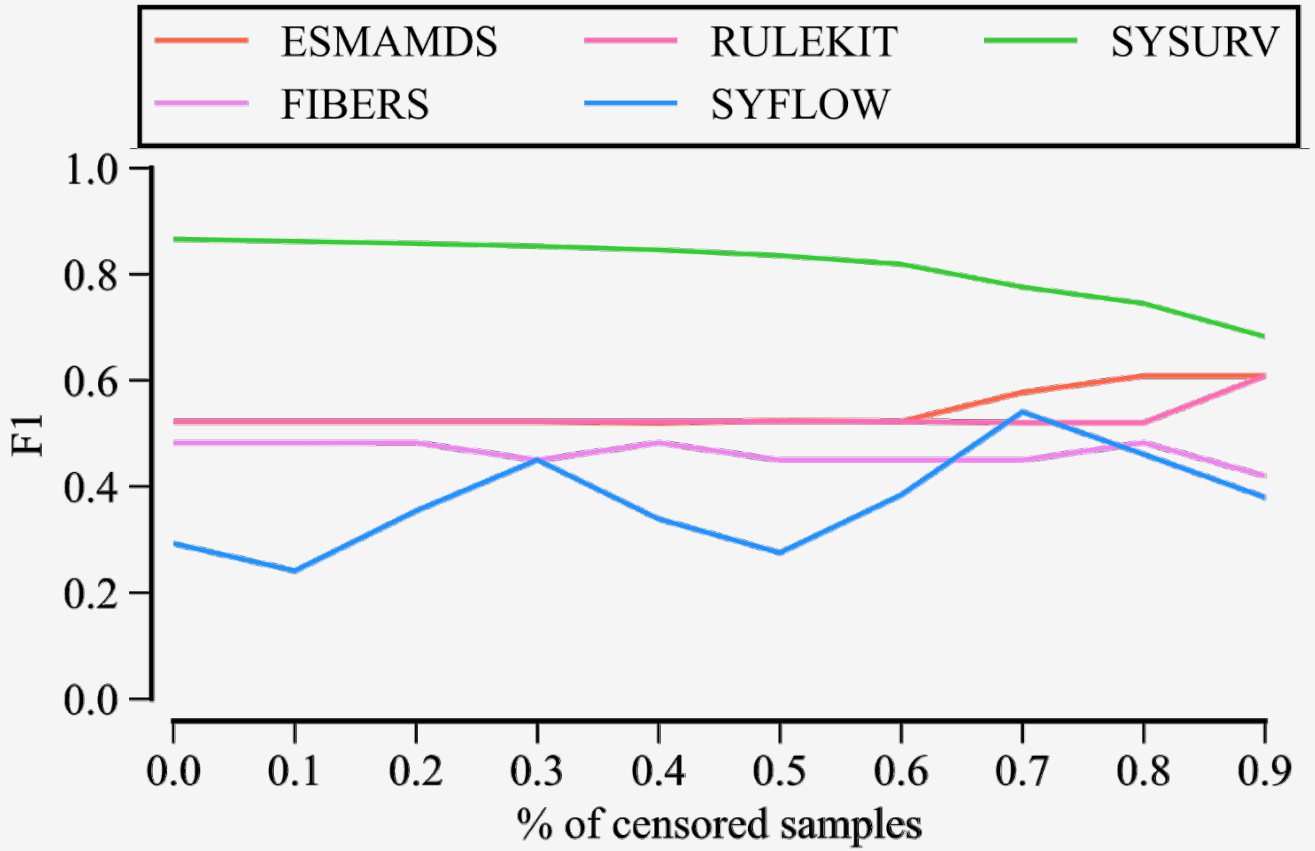
Survival Subgroups

SyFlow	SySurv
1. Learn predicates end-to-end → **Accurate Discretization**	1. Learn predicates end-to-end → **Accurate Discretization**
2. Continous optimization → **Highly scalable**	2. Continous optimization → **Highly scalable**
3. Use Normalizing Flows (NFs) → **Ignores censoring**	3. Use Random Survival Forests → **No assumptions**
4. Uses KL-Divergence → **Ignores survival**	4. Maximize curve distance → **No assumptions**



(Al Rahwanji, work in progress)

Experiments — Synthetic data



(Al Rahwanji, work in progress)

Experiments — Real world

	Our objective					KL-Divergence					Logrank statistic					Mean-shift					Adjusted mean-shift				
	SYSURV	SYFLOW	ESMAMDS	FIBERS	RULEKIT	SYSURV	SYFLOW	ESMAMDS	FIBERS	RULEKIT	SYSURV	SYFLOW	ESMAMDS	FIBERS	RULEKIT	SYSURV	SYFLOW	ESMAMDS	FIBERS	RULEKIT	SYSURV	SYFLOW	ESMAMDS	FIBERS	RULEKIT
UnempDur	**6.37**	3.20	3.79	3.79	3.79	0.08	**0.15**	0.13	0.13	0.13	19.26	125.99	125.00	125.00	**226.52**	-2.01	-1.74	**2.14**	**2.14**	**2.14**	-0.02	-0.00	0.00	0.00	0.00
nwtco	**1011.64**	205.02	680.98	208.59	680.98	**0.14**	0.02	0.06	0.00	0.06	**374.22**	0.21	31.36	0.00	281.27	**1161.97**	872.08	688.24	-0.56	688.24	6.92	0.40	1.50	-0.00	1.50
rott2	**559.22**	419.19	487.58	487.58	326.16	-	-	-	-	-	**266.95**	3.11	71.65	169.35	93.04	31.47	**-92.00**	25.23	25.23	10.03	0.10	-11.50	0.04	0.04	0.01
rdata	**259.24**	153.69	213.83	158.52	213.83	**0.08**	0.04	0.03	0.01	0.03	**113.38**	56.39	14.64	43.32	107.75	**1116.58**	-700.91	933.49	705.50	933.49	7.98	-1.72	3.60	1.62	3.60
Aids2	**66.64**	31.99	54.01	34.69	60.52	-	-	-	-	-	0.21	0.00	**4.27**	0.01	1.72	22.68	-1.45	**61.36**	-0.70	61.21	0.07	-0.00	0.10	-0.00	0.17
Dialysis	4.52	4.69	4.71	4.59	**4.74**	0.02	**0.24**	0.01	0.02	0.04	6.36	33.76	70.86	**305.94**	68.38	0.64	**-19.60**	3.05	4.47	5.74	0.00	-0.19	0.00	0.00	0.02
TRACE	261.76	**486.65**	332.10	261.58	272.52	0.00	-	0.02	**4.98**	0.01	0.00	17.74	93.24	0.00	**94.62**	0.00	**4.53**	1.73	-0.00	0.61	0.00	4.53	0.00	-0.00	0.00
support2	**132.83**	68.72	104.90	54.79	81.38	**2.71**	0.57	1.36	0.22	0.21	**251.00**	39.00	46.83	11.93	59.87	**405.72**	-193.91	405.52	-71.03	227.60	12.29	-1.41	5.13	-0.16	2.35
dataDIVAT2	**286.30**	99.92	172.47	141.30	141.30	0.16	**3.91**	1.41	2.19	2.19	**33.55**	15.90	10.94	31.60	31.60	**3.18**	-1.06	2.06	1.56	1.56	0.05	-0.00	0.01	0.00	0.00
prostateSurvival	**20.11**	9.20	14.13	5.38	7.51	-	-	-	-	-	**425.72**	103.66	82.50	82.50	64.74	**6.11**	5.11	3.37	-1.02	3.67	0.00	0.00	0.00	-0.00	0.00
actg	**34.58**	13.62	18.82	11.33	24.22	**0.09**	0.03	0.03	0.00	0.06	**59.97**	2.40	12.26	0.02	1.05	**54.13**	18.50	5.98	-0.37	27.30	1.42	0.07	0.03	-0.00	0.22
scania	**172.47**	101.59	147.42	84.74	147.42	**0.03**	0.00	**0.03**	0.01	**0.03**	4.65	1.91	13.66	9.93	**14.66**	**6.51**	-2.86	-2.44	-0.07	-2.44	0.05	-0.00	-0.01	-0.00	-0.01
grace	**38.56**	32.04	21.01	13.08	16.51	0.14	0.06	**7.53**	0.02	0.00	59.30	**70.71**	28.68	13.81	33.33	**64.03**	51.79	21.77	-11.14	16.00	1.94	0.54	0.08	-0.02	0.04
Avg. rank	**1.54**	3.85	2.46	4.23	2.92	**2.50**	2.67	2.80	3.75	3.05	**2.38**	3.46	3.08	3.65	2.42	**2.08**	2.85	2.85	4.23	3.00					

(Datasets were chosen to have 1000 samples or more and around 10 features.)

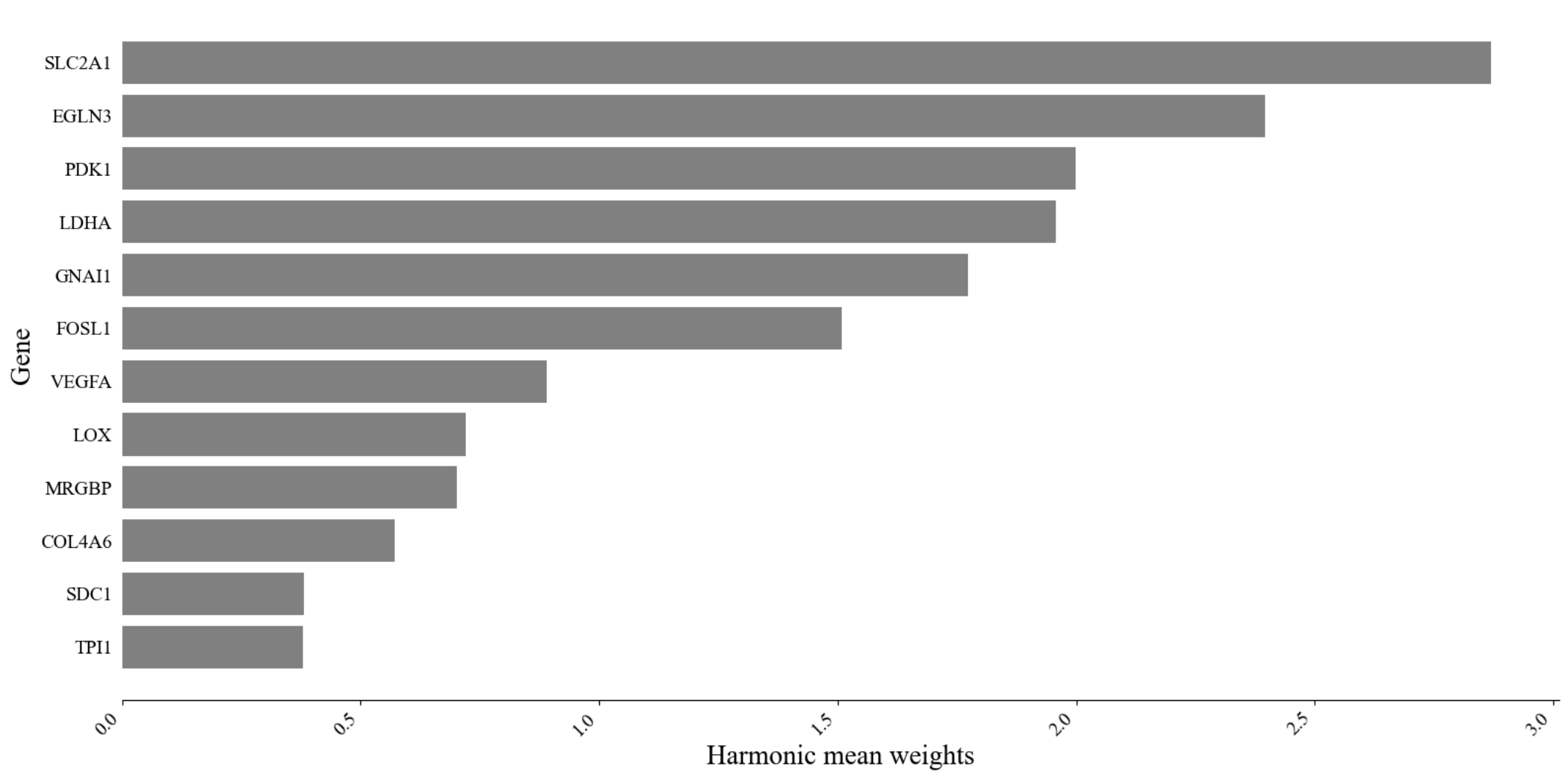
Experiments — Real world

Avg. Rank	SYSURV	SYFLOW	ESMAMDS	FIBERS	RULEKIT
Our objective	**1.54**	3.85	2.46	4.23	2.92
KL-Divergence	**2.50**	2.67	2.80	3.75	3.05
Logrank statistic	**2.38**	3.46	3.08	3.65	2.42
Mean-shift	**2.08**	2.85	2.85	4.23	3.00

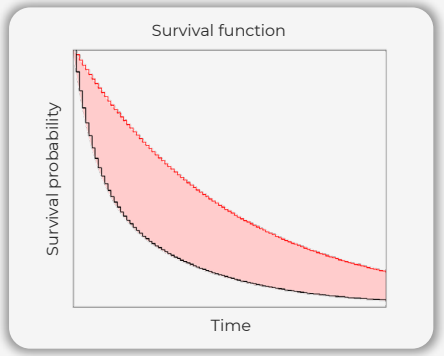
(Al Rahwanji, work in progress)

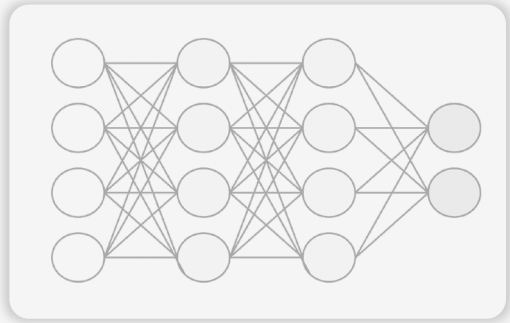
Case Study: Neck Carcioma

PCR - first rule (pruned) -> All associated with Hypoxia

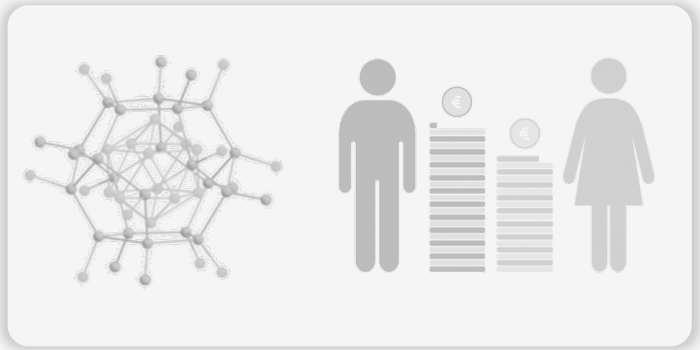


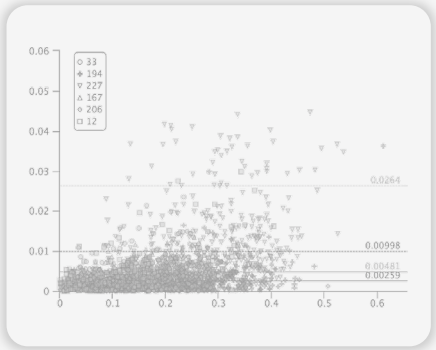
Common questions in ML



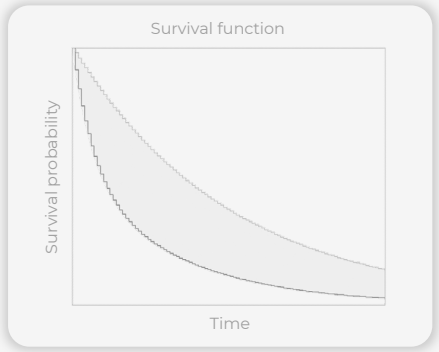


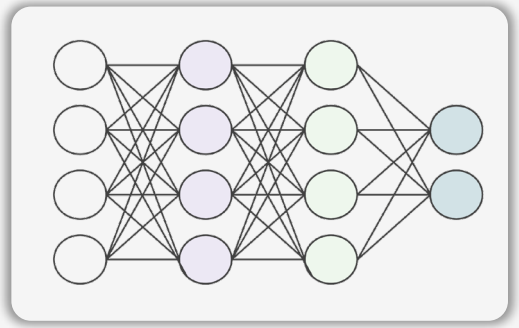
Data ← → Models



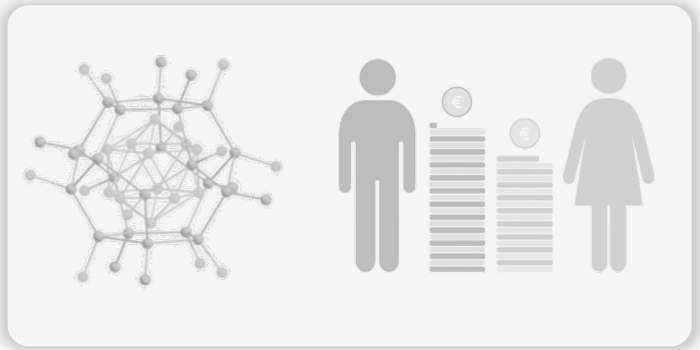


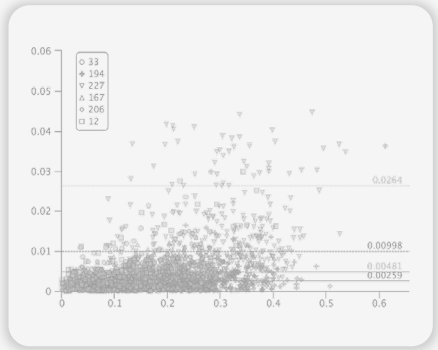
Common questions in ML



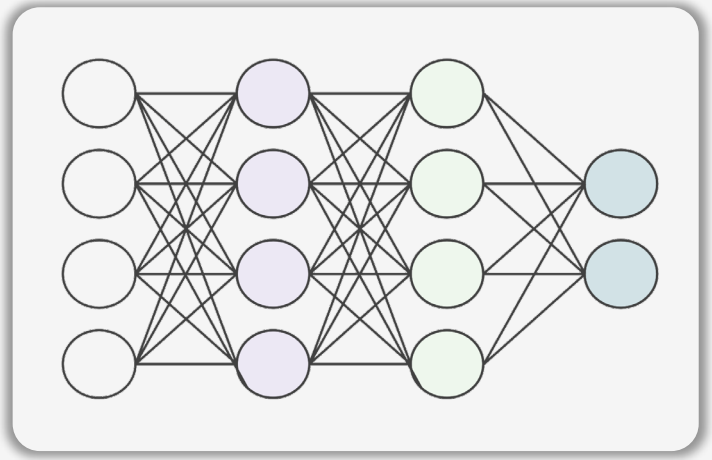


Data ← → Models





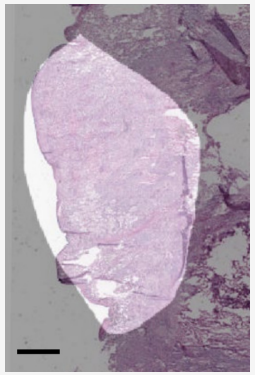
Understanding reasoning



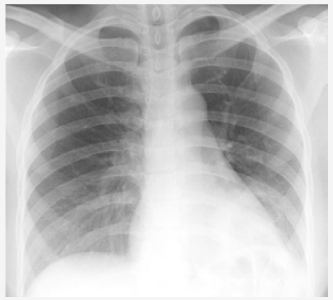
Extract interpretable concepts

Why is it important?

Gaining new scientific insights



"... 45 out of 54 of the TCGA images **misclassified by at least one of the pathologists** were **assigned** to the **correct** cancer type **by the algorithm**".



COVID-Net

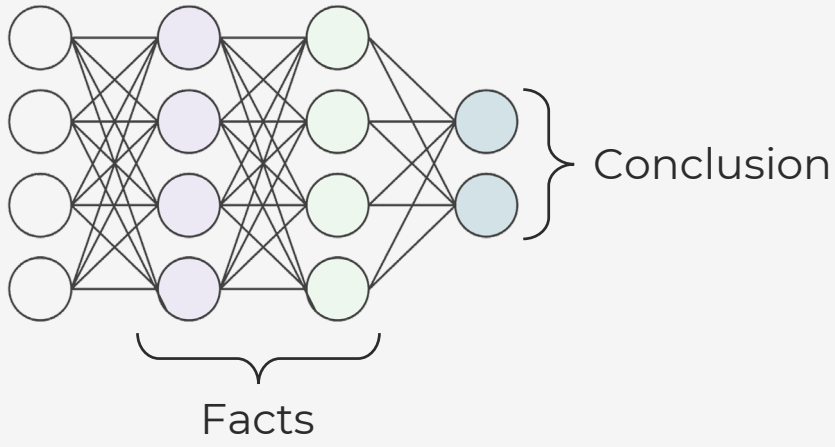
Gain insights into **critical factors** associated with **COVID** cases

What is **reasoning**?

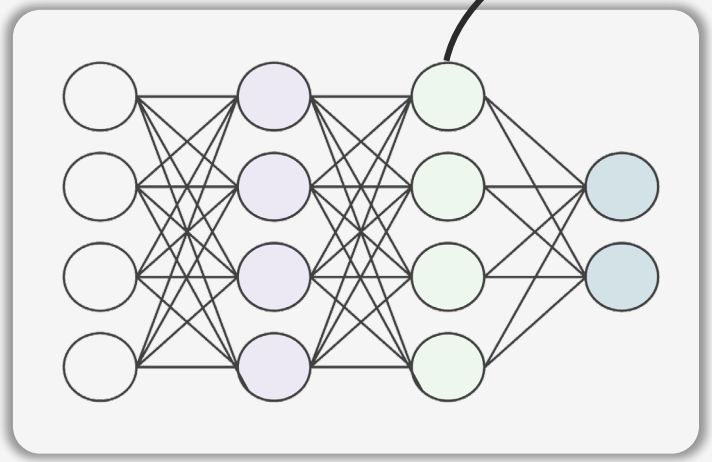
*„Reasoning is the process by which you reach a **conclusion** after thinking about all the **facts**."*

— Collins dictionary

i.e. a relation between **facts** and a **conclusion**

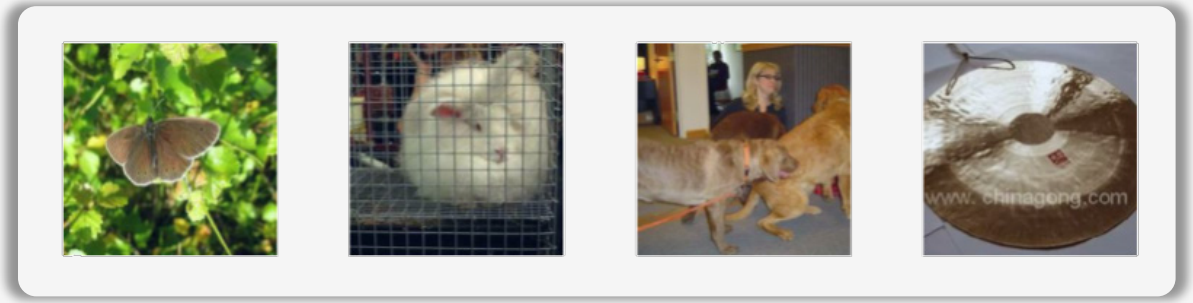


Extracting Features



Neuron activations								Y
0	1	1	1	1	1	1	1	c_1
1	1	0	0	0	0	1	1	c_1
1	1	1	1	0	1	0	1	c_1
0	1	1	1	0	1	0	0	c_2
0	1	1	1	0	1	0	0	c_2

Neurons are polysemantic



(Fischer et al. 2021)

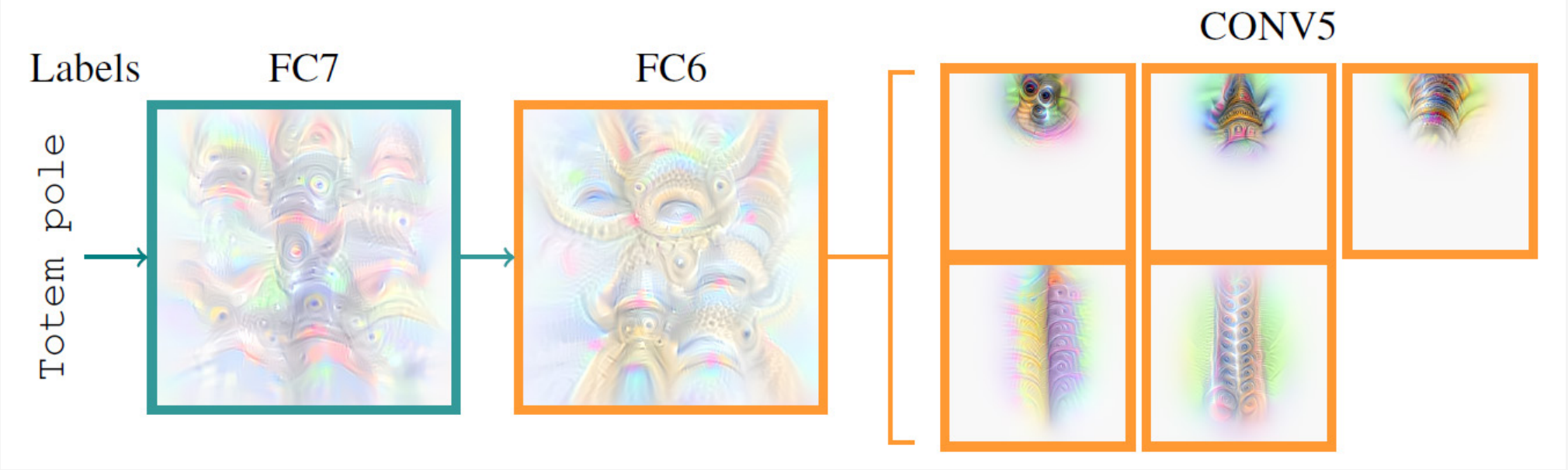
Sneak peek – Information Flow

How does information flow and how is it combined?

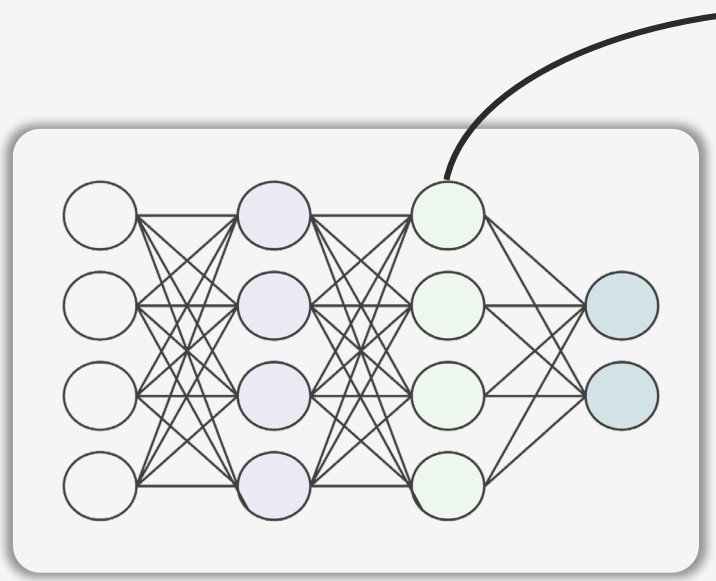
Rules: $X \rightarrow Y$, $X \in Labels,\ Y \in FC7$

$Y \rightarrow Z$, $Z \in FC6$

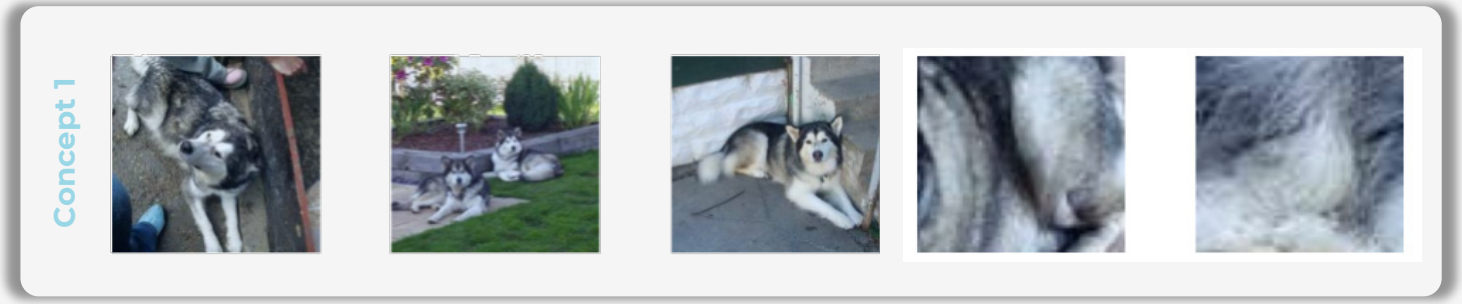
$Z \rightarrow Q$, $Q \in CONV5$



Extracting ~~Features~~ Concepts

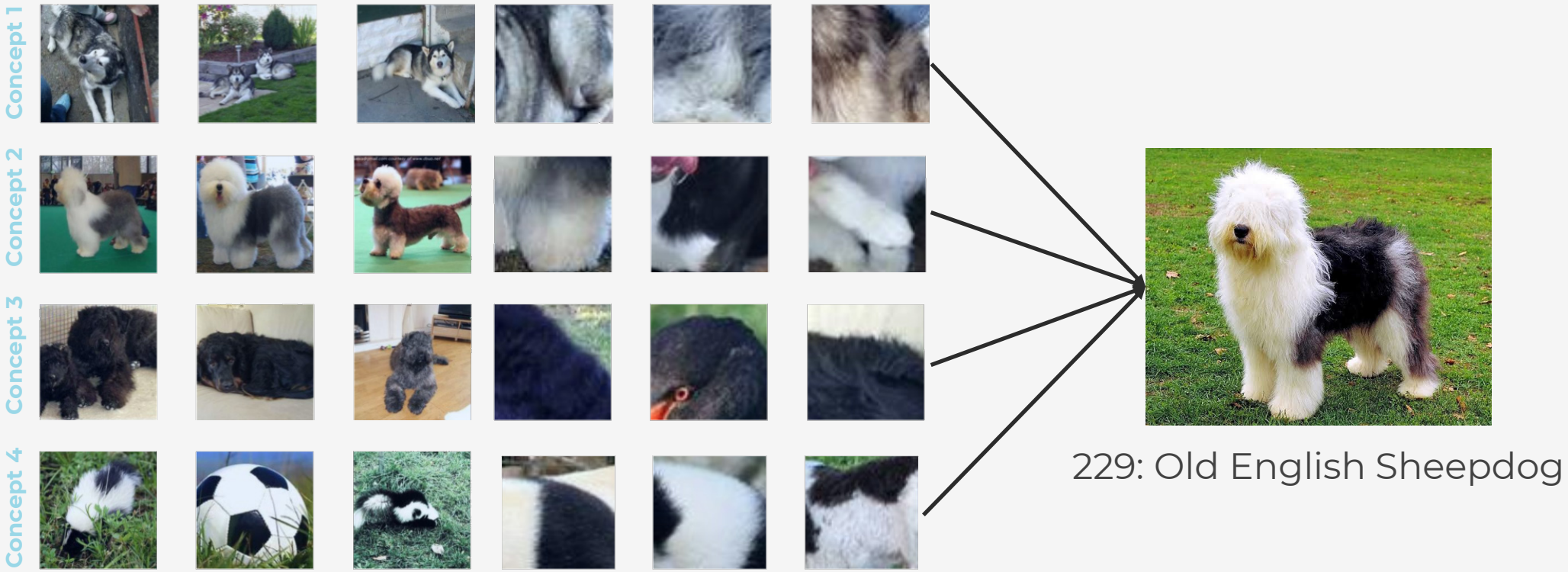


Concept ~~Feature~~ activations								Y
0	1	1	1	1	1	1	1	c_1
1	1	0	0	0	0	1	1	c_1
1	1	1	1	0	1	0	1	c_1
0	1	1	1	0	1	0	0	c_2
0	1	1	1	0	1	0	0	c_2



(Nait Said, work in progress)

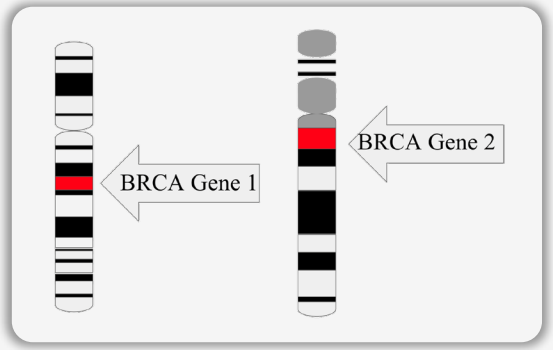
Connecting concepts



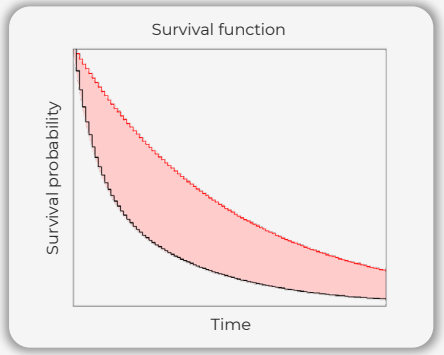
(Nait Said, work in progress)

Conclusion

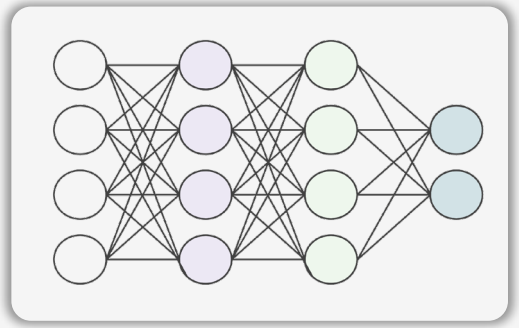
DiffNaps



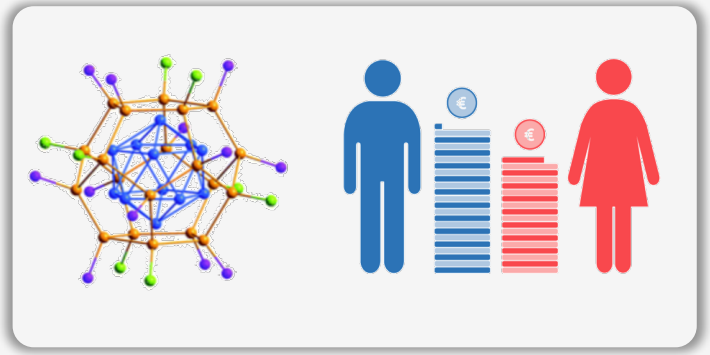
SySurv



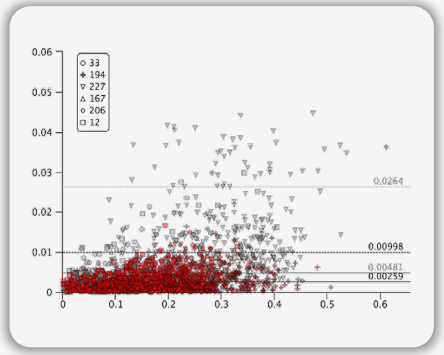
ExplaiNNaps



Data ← → Models



SyFlow



Domains of Applicability