A three-level modeling for identifying important predictor variables in GWAS studies suffering from p >> n.

Jakub Liu

CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN

Freie Universität Berlin

Presentation outline

- introduction to the topic
- brief overview of the dataset
- methodology
- summary

Introduction to the p>>n problem

$X_{1,1}$	$X_{1,2}$	...	$X_{1,p}$
$X_{2,1}$	$X_{2,2}$	...	$X_{2,p}$
...	...	...	...
$X_{n,1}$	$X_{n,2}$	...	$X_{n,p}$

Example: rare disease data

- few patients
- genotyped at many locations

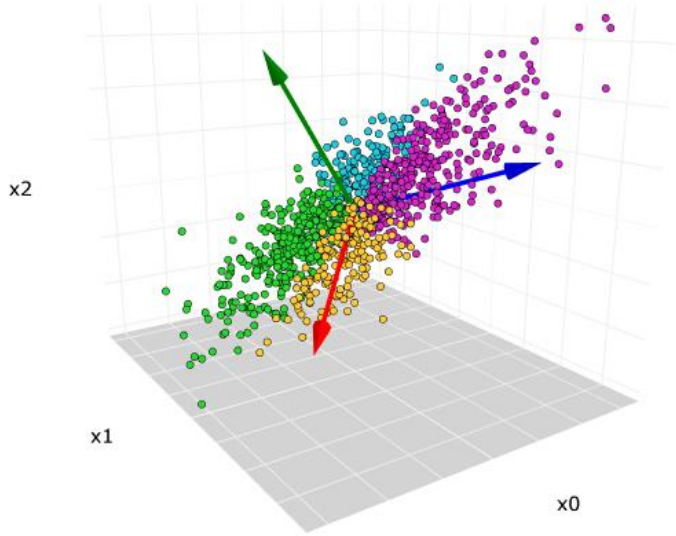
Patient ID	*SNP 1*	*SNP 2*	*SNP 3*	*SNP 4*	*…*	*SNP 3e^6*
0001	*0/0*	*0/0*	*0/0*	*0/0*	*…*	*0/1*
0002	*1/0*	*0/0*	*0/0*	*1/0*	*…*	*0/0*

Problems with p>>n

- **statistical**
  - too low bias, too high variance (overfitting)
  - loss of degrees of freedom (in context of hypothesis testing)
- **mathematical**
  - for example, in OLS, matrix inversion is impossible
- **interpretational**
- **computational**

Common approaches

1.)



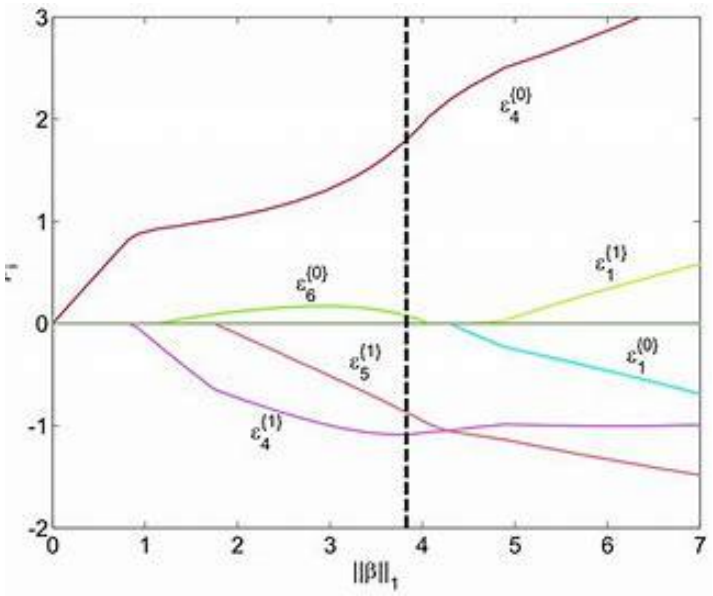
Dimensionality reduction

Pseudo inverse

2.)

$$X^{+} = (X^{t}X)^{-1}X^{t}$$

Regularization



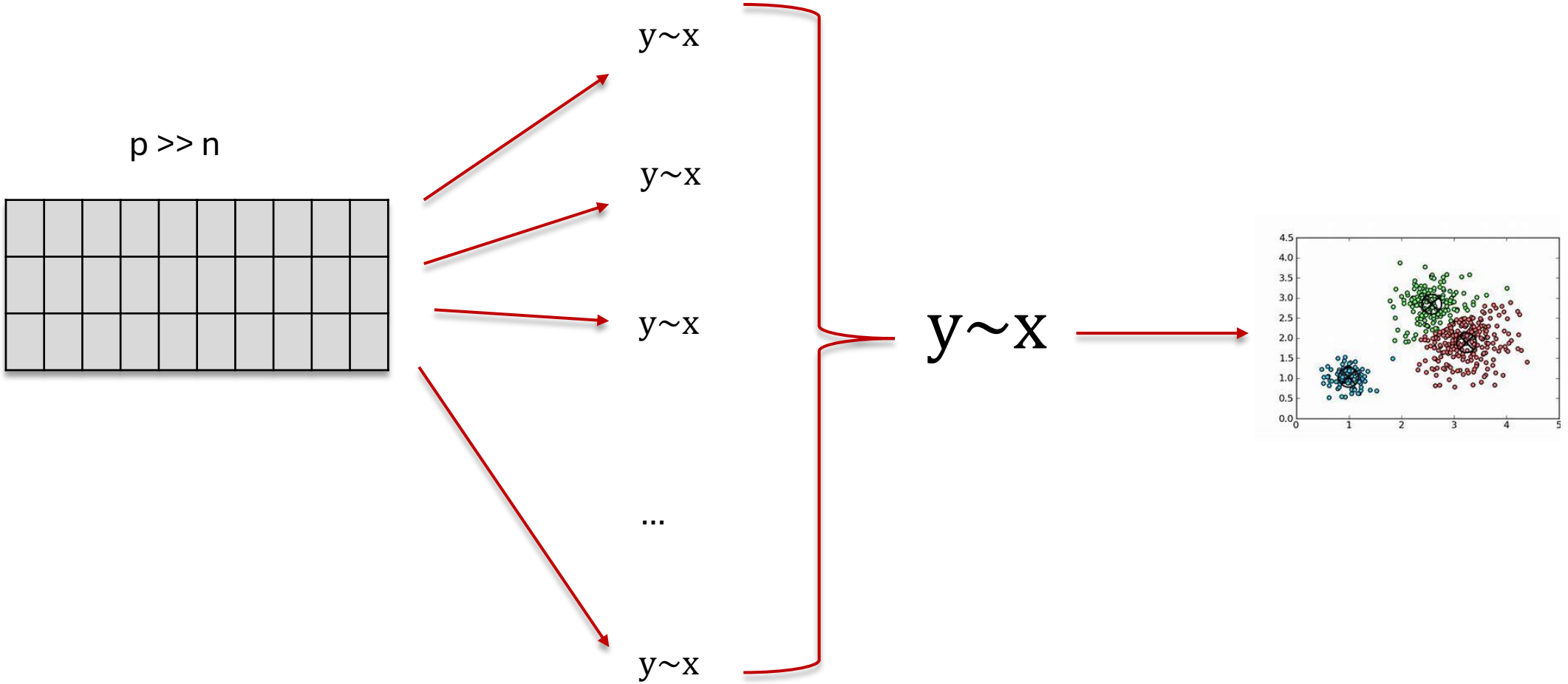
3.)

1.)https://www.researchgate.net/figure/D-scatter-plot-of-the-DLBCL-data-with-colors-representing-the-true-clustering-labels_fig2_346052105

2.)https://fractalytics.io/moore-penrose-matrix-optimization-cuda-c

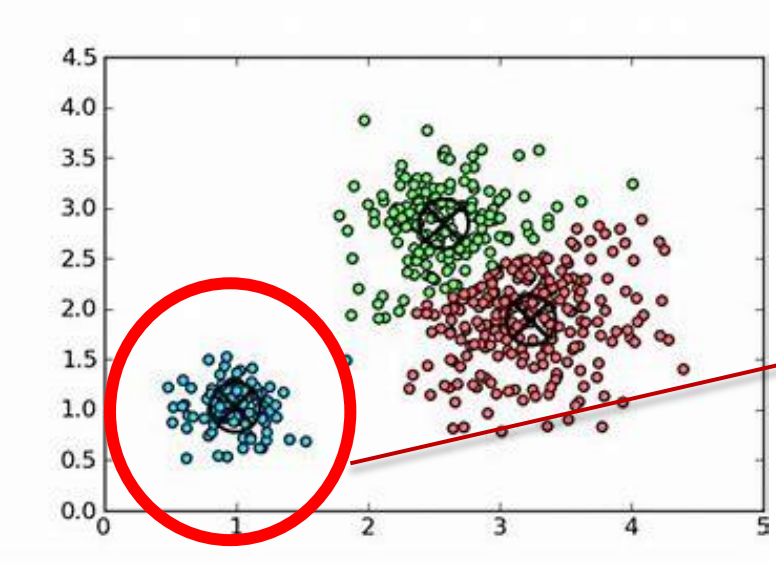
3.)Regularization path of the regularized logistic regression model with... | Download Scientific Diagram (researchgate.net)

Procedure outline

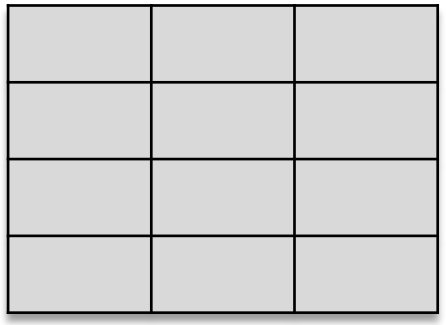


https://datascience.stackexchange.com/questions/97963/visualise-kmeans-clusters-in-2d-when-number-of-input-features-is-greater-than-2-true-clustering-labels_fig2_346052105

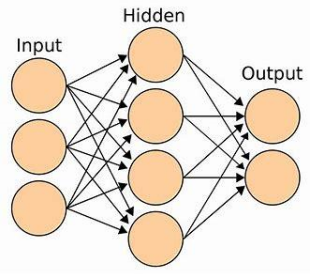
Procedure outline



p < n



analyze the dataset



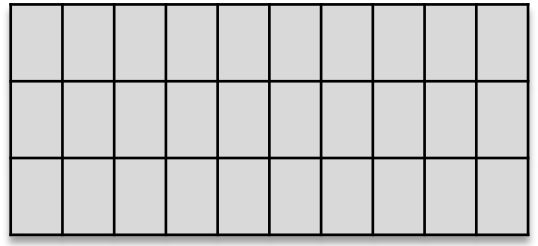
validate the procedure

https://stackabuse.com/introduction-to-neural-networks-with-scikit-learn/

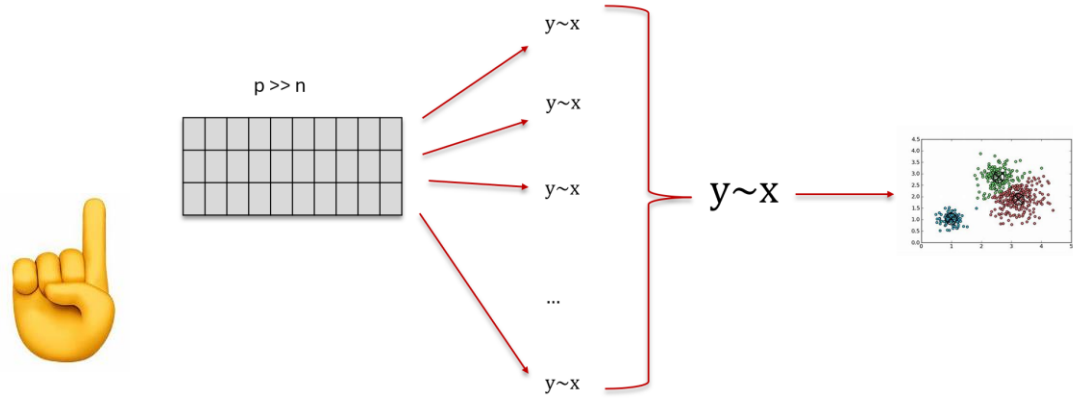
The dataset

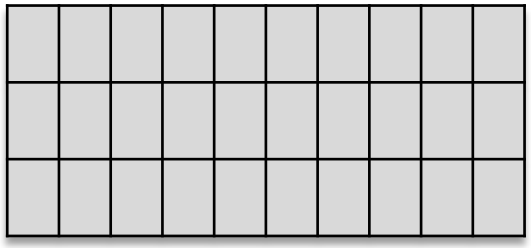
Kaja, et al. **"The Thousand Polish Genomes—a database of Polish variant allele frequencies."** *International Journal of Molecular Sciences* 23.9 (2022): 4532.

- 1222 individuals of Polish origin
- Variant Call Format (VCF) file
- 41836187 SNP (~ 42 million)
- WGS genotype data
- phenotype data: severe/not-severe COVID-19 illness



Data preparation



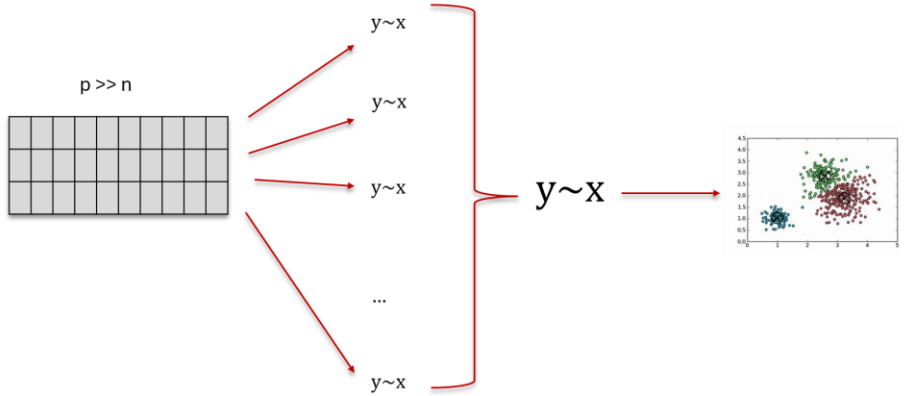


VCFtools
- MAF > 0.1

PLINK

- big data problems
- ~42 million SNPs → ~7 million SNPs

Multiple logistic regression models



	SNP 1	SNP 2	...	SNP p	phenotype
patient 1	0	1	...	2	0
patient 2	2	2	...	1	1
patient 3	0	0	...	0	1
...	...	...	...	...	...
patient n	2	1	...	1	0

scheme

$$p >> n$$

	SNP 1	SNP 2	...	SNP p	phenotype
patient 1	0	1	...	2	0
patient 2	2	2	...	1	1
patient 3	0	0	...	0	1
...	...	...	...	...	...
patient n	2	1	...	1	0

scheme

p>>n

	SNP 1	SNP 2	...	SNP p	phenotype
patient 1	0	1	...	2	0
patient 2	2	2	...	1	1
patient 3	0	0	...	0	1
...	...	...	...	...	...
patient n	2	1	...	1	0

p<n

SNP 1	SNP 2	phenotype
0	1	0
2	2	1
0	0	1
...	...	...
2	1	0

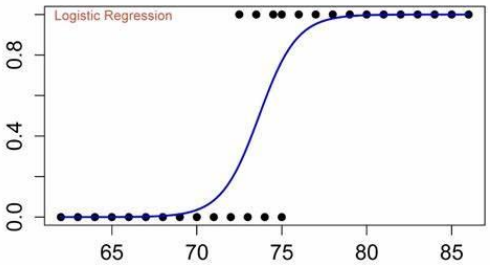
scheme

p>>n

	SNP 1	SNP 2	...	SNP p	phenotype
patient 1	0	1	...	2	0
patient 2	2	2	...	1	1
patient 3	0	0	...	0	1
...	...	...	...	...	...
patient n	2	1	...	1	0

p<n

SNP 1	SNP 2	phenotype
0	1	0
2	2	1
0	0	1
...	...	...
2	1	0

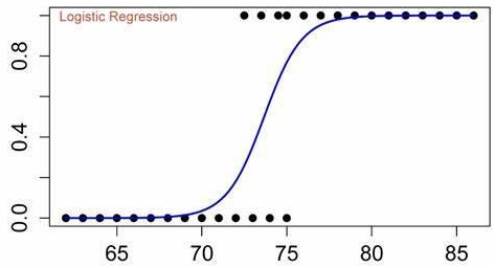


p>>n

	SNP 1	SNP 2	...	SNP p	phenotype
patient 1	0	1	...	2	0
patient 2	2	2	...	1	1
patient 3	0	0	...	0	1
...	...	...	...	...	...
patient n	2	1	...	1	0

p<n

SNP 1	SNP 2	phenotype
0	1	0
2	2	1
0	0	1
...	...	...
2	1	0



SNP 1	SNP 2	SNP 3	...	SNP p	deviance
-0.3	1.3	0	...	0	0.9

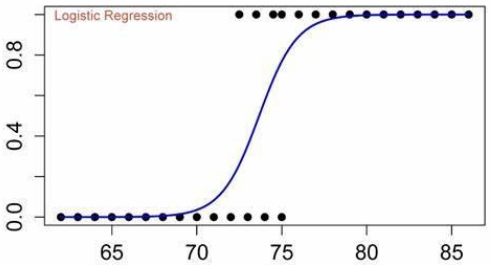
scheme

p<n

p>>n

	SNP 1	SNP 2	...	SNP p	phenotype
patient 1	0	1	...	2	0
patient 2	2	2	...	1	1
patient 3	0	0	...	0	1
...	...	...	...	...	...
patient n	2	1	...	1	0

SNP 1	SNP 2	SNP 3	...	SNP p	deviance
-0.3	1.3	0	...	0	0.9



scheme

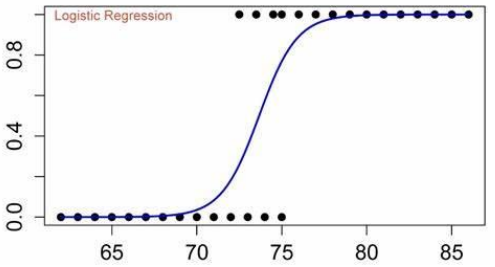
p>>n

	SNP 1	SNP 2		...	SNP p	phenotype
patient 1	0	1		...	2	0
patient 2	2	2		...	1	1
patient 3	0	0		...	0	1
...	...	...		...	...	...
patient n	2	1		...	1	0

p<n

SNP p-1	SNP p	phenotype
...	2	0
...	1	1
...	0	1
...	...	...
...	1	0

SNP 1	SNP 2	SNP 3	...	SNP p	deviance
-0.3	1.3	0	...	0	0.9



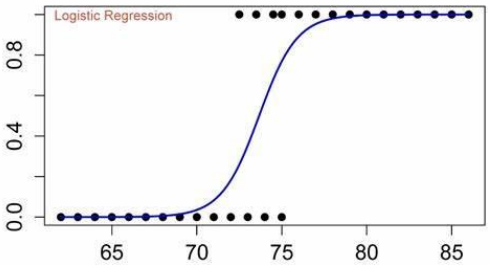
scheme

p>>n

	SNP 1	SNP 2	...	SNP p	phenotype
patient 1	0	1	...	2	0
patient 2	2	2	...	1	1
patient 3	0	0	...	0	1
...	...	...	...	...	...
patient n	2	1	...	1	0

p<n

SNP p-1	SNP p	phenotype
...	2	0
...	1	1
...	0	1
...	...	...
...	1	0



SNP 1	SNP 2	SNP 3	...	SNP p	deviance
-0.3	1.3	0	...	0	0.9

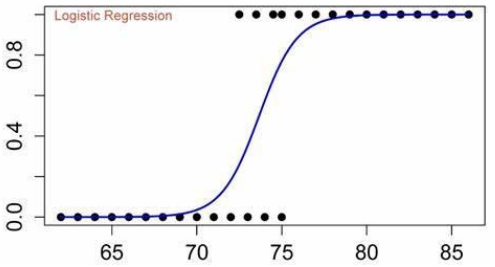
scheme

p>>n

	SNP 1	SNP 2	...	SNP p	phenotype
patient 1	0	1	...	2	0
patient 2	2	2	...	1	1
patient 3	0	0	...	0	1
...	...	...	...	...	...
patient n	2	1	...	1	0

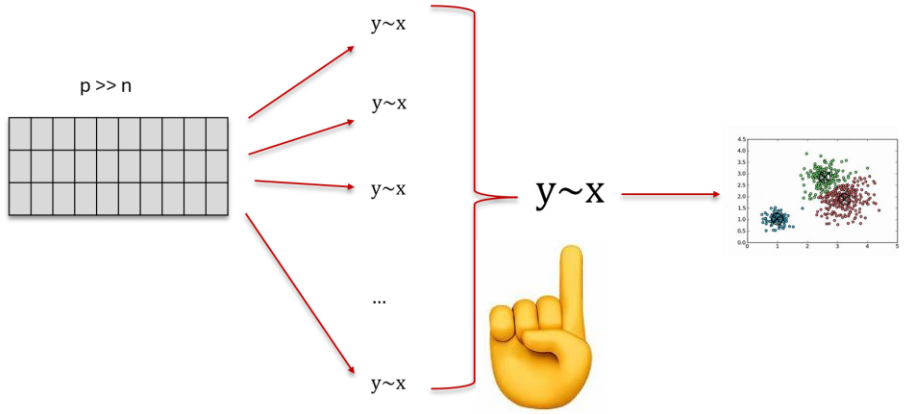
p<n

...	SNP p	phenotype
...	2	0
...	1	1
...	0	1
...	...	...
...	1	0



SNP 1	SNP 2	SNP 3	...	SNP p	deviance
-0.3	1.3	0	...	0.8	0.9
0	0	0	...	1.1	0.6

mixed effect model



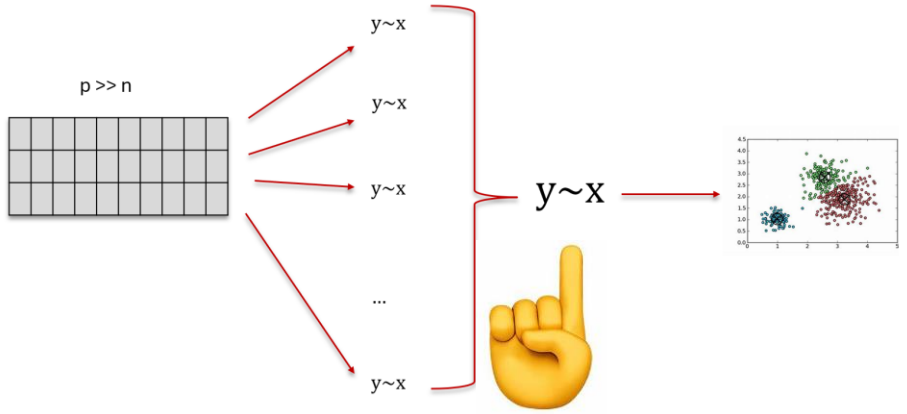
n logistic regression models

SNP 1	SNP 2	SNP 3	...	SNP p	deviance
-0.3	1.3	0	...	0.8	0.9
0	0	0	...	1.1	0.6
...	...	...	...	...	...
-1.2	0.7	1.1	...	0.4	0.7

mixed effects linear model

$$y = \mu + \boldsymbol{Za} + \boldsymbol{\varepsilon},$$

mixed effect model



n logistic regression models

SNP 1	SNP 2	SNP 3	...	SNP p	deviance
-0.3	1.3	0	...	0.8	0.9
0	0	0	...	1.1	0.6
...	...	...	...	...	...
-1.2	0.7	1.1	...	0.4	0.7

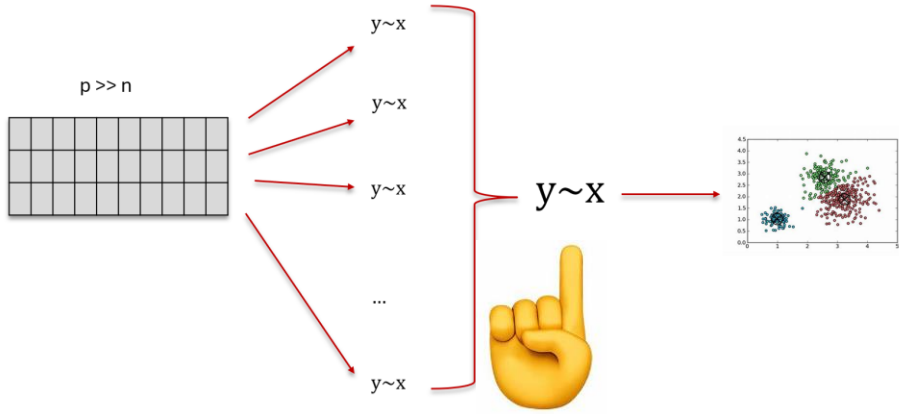
mixed effects linear model

$$y = \mu + Za + \varepsilon$$

y → model deviance

Z → design matrix of SNP effects

mixed effect model



n logistic regression models

SNP 1	SNP 2	SNP 3	...	SNP p	deviance
-0.3	1.3	0	...	0.8	0.9
0	0	0	...	1.1	0.6
...	...	...	...	...	...
-1.2	0.7	1.1	...	0.4	0.7

mixed effects linear model

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{Za} + \boldsymbol{\varepsilon}$$

y → model deviance

Z → design matrix of SNP effects

Mix99 software (Fortran90)

mixed effect model

- SNP effect treated as random effect

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{Za} + \boldsymbol{\varepsilon}$$

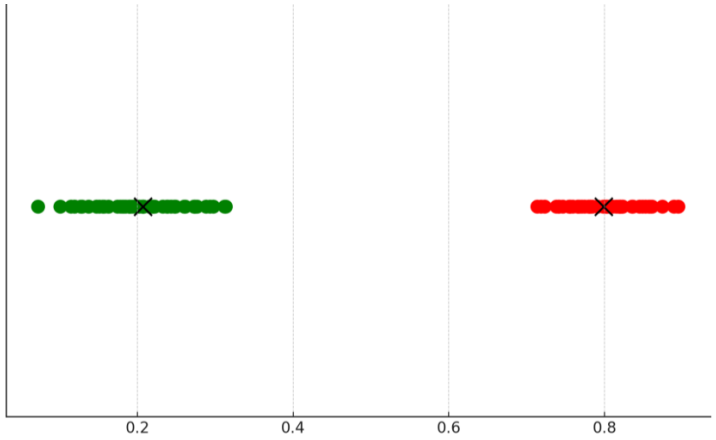
SNP ID	SNP effect
SNP 1	0.76
SNP 2	0.08
…	…
SNP p	0.02

1D clustering

SNP ID	SNP effect
SNP 1	0.76
SNP 2	0.08
…	…
SNP p	0.02

1D clustering

SNP ID	SNP effect
SNP 1	0.76
SNP 2	0.08
…	…
SNP p	0.02

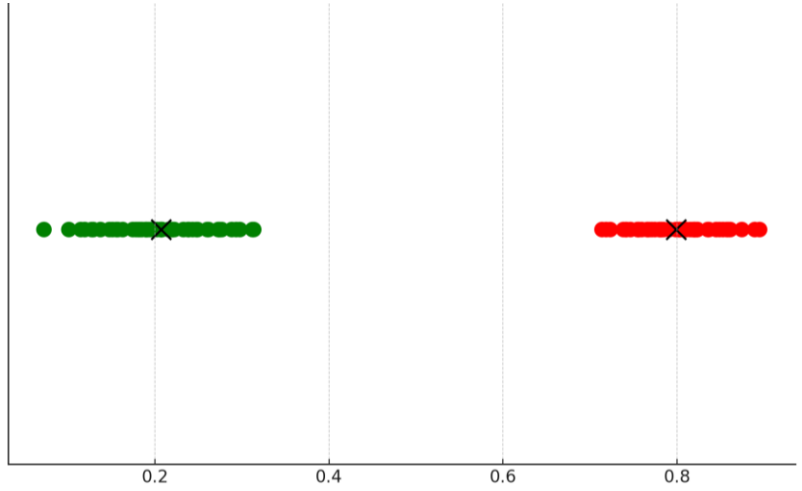


- 2 clusters (significant and non-significant SNPs)

https://chatgpt.com/

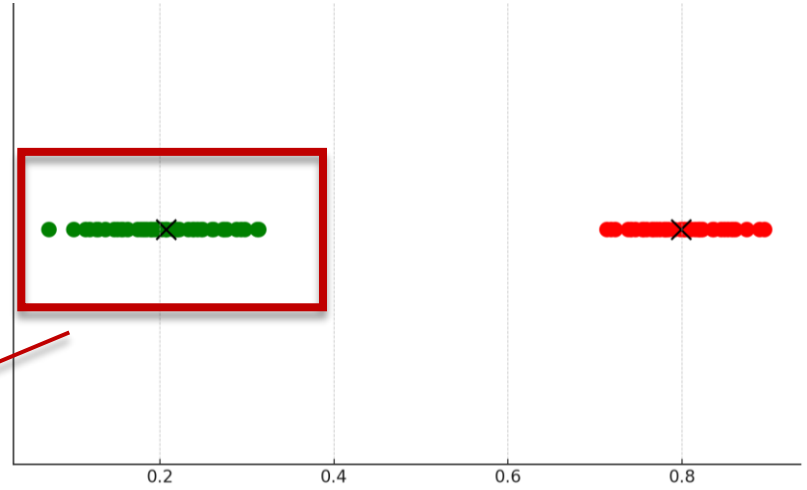
1D clustering

Which cluster corresponds to the significant SNPs?



1D clustering

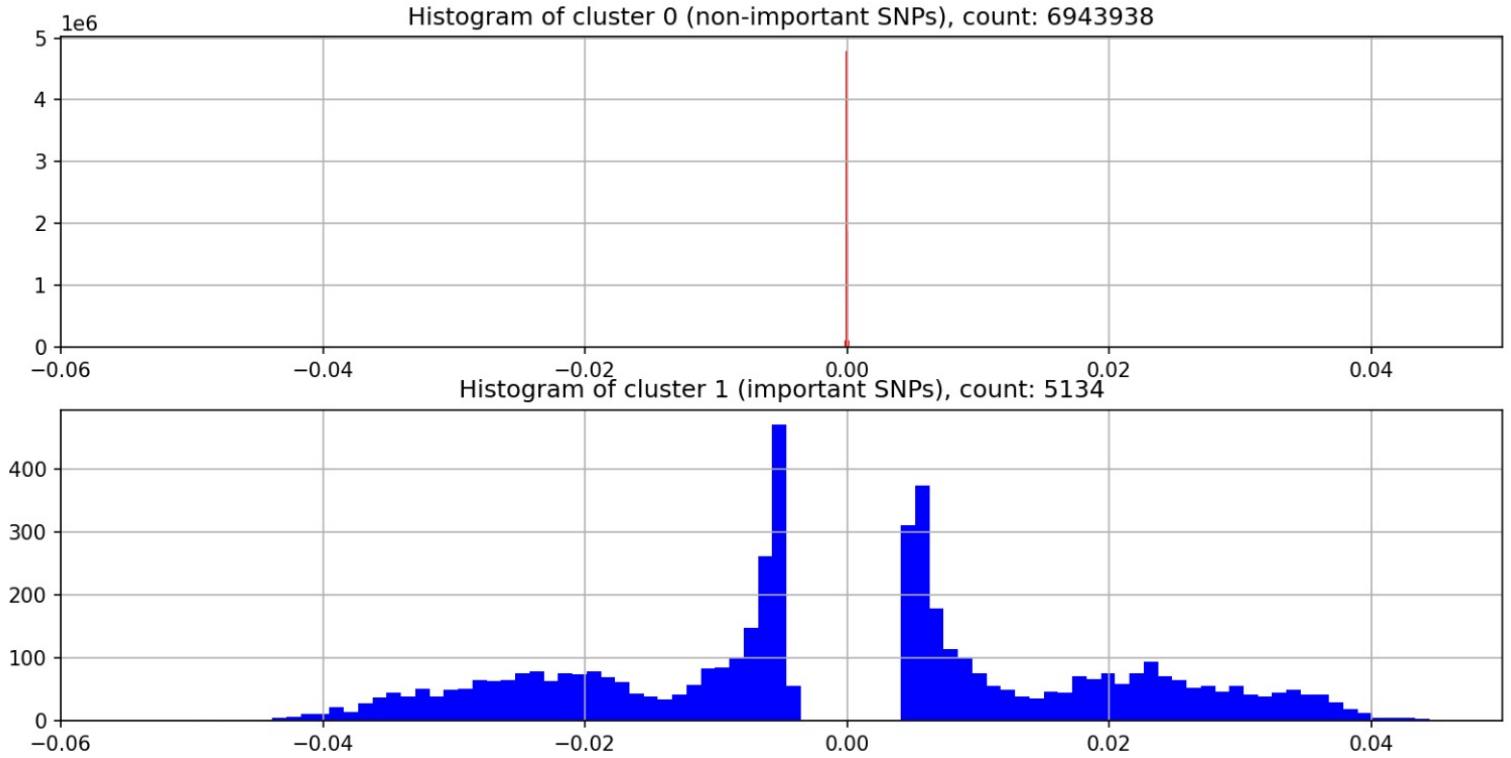
Which cluster corresponds to the important SNPs?



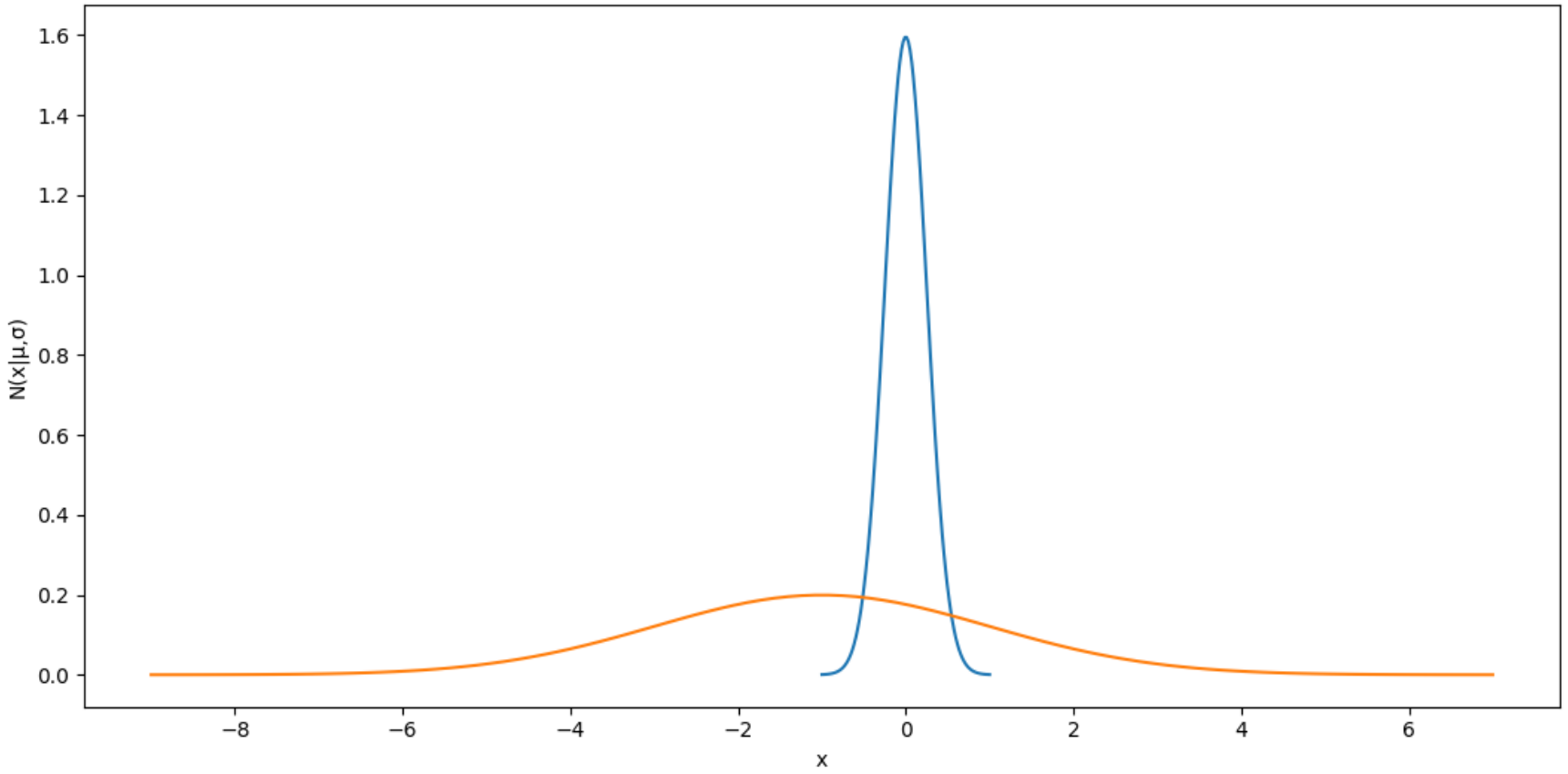
- lower effect
- ‘decreases’ the response variable (deviance)
- a lower deviance indicates a better fit

1D clustering

- approach 1: Gaussian Mixture Model clustering

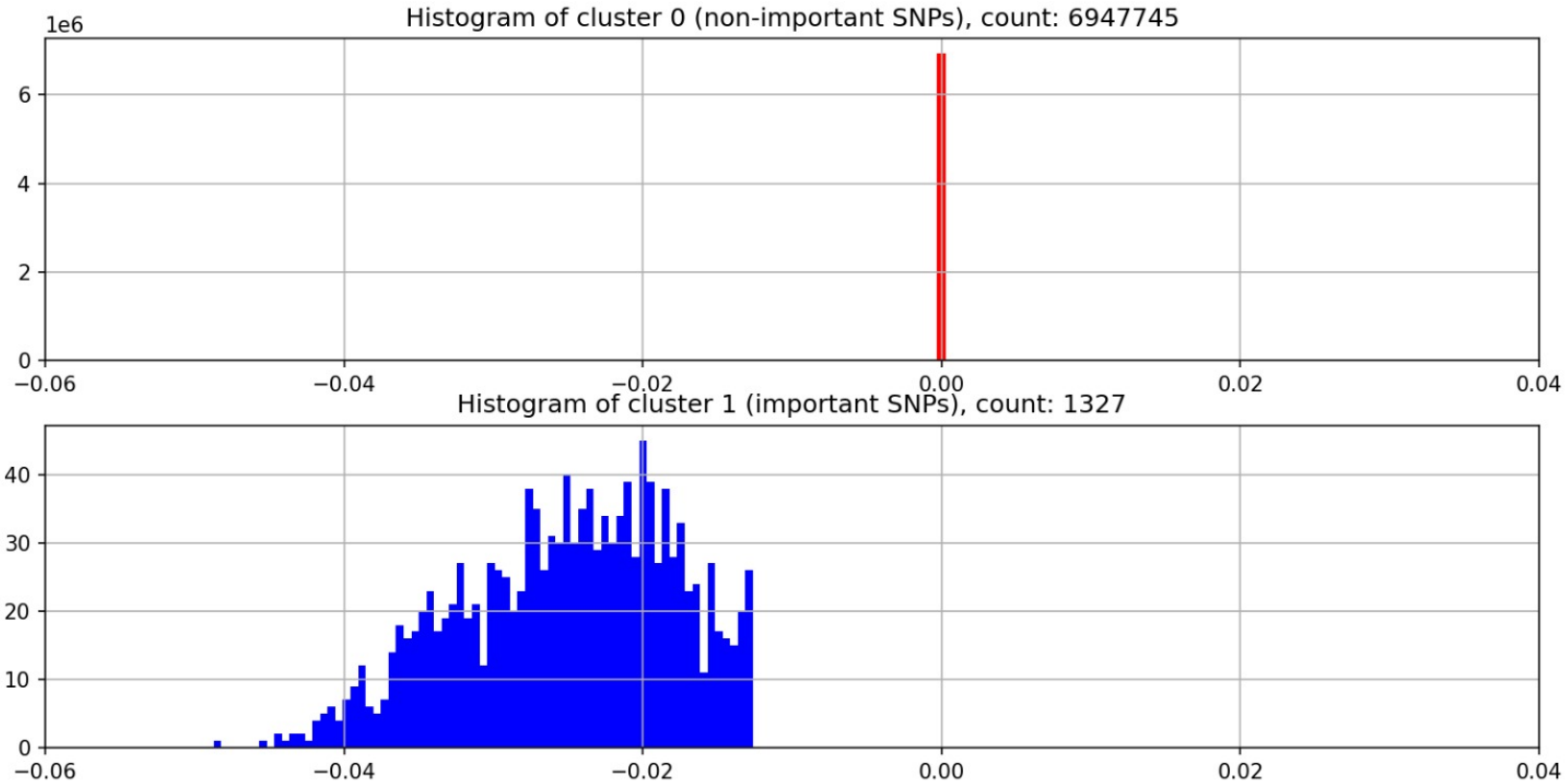


1D clustering



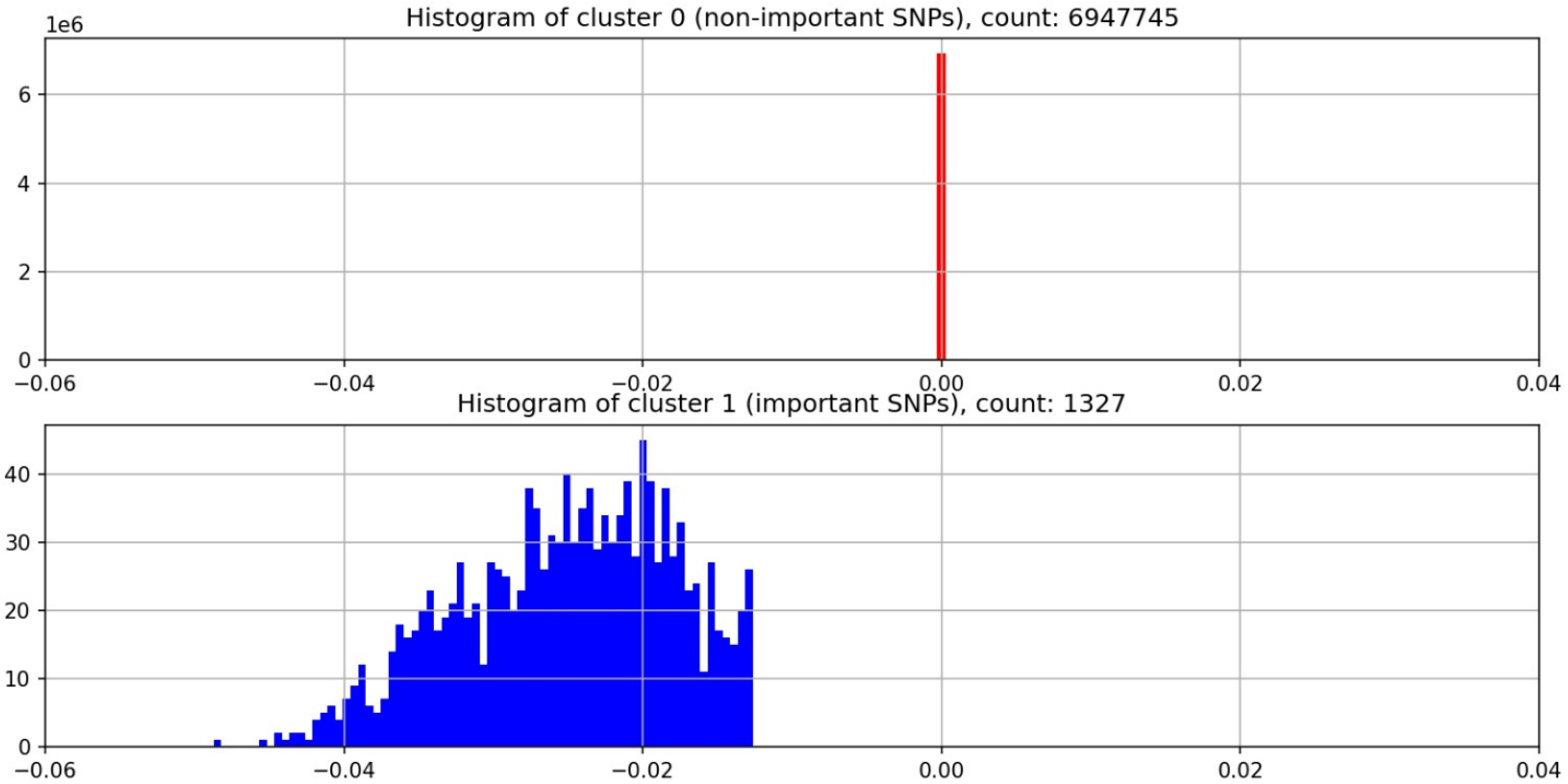
1D clustering

- approach 2: K-Means clustering



1D clustering

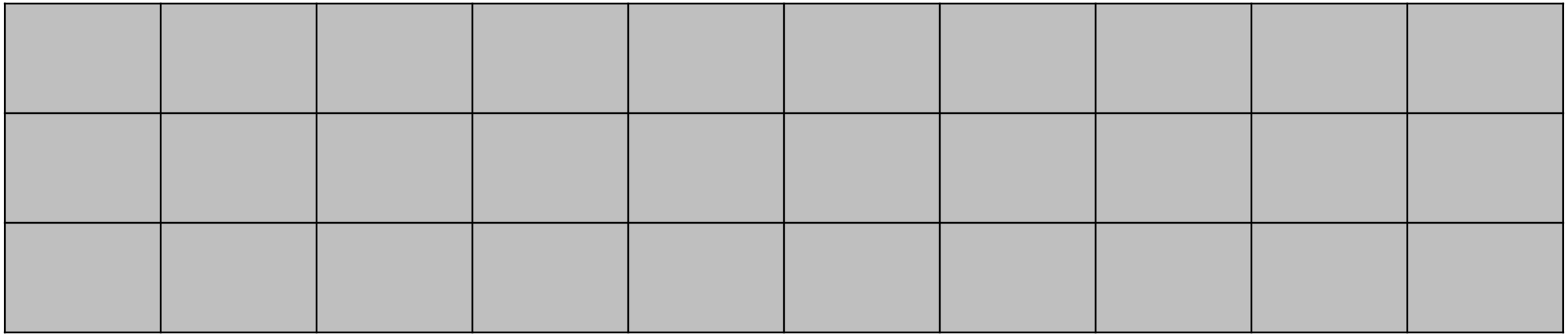
- approach 2: K-Means clustering



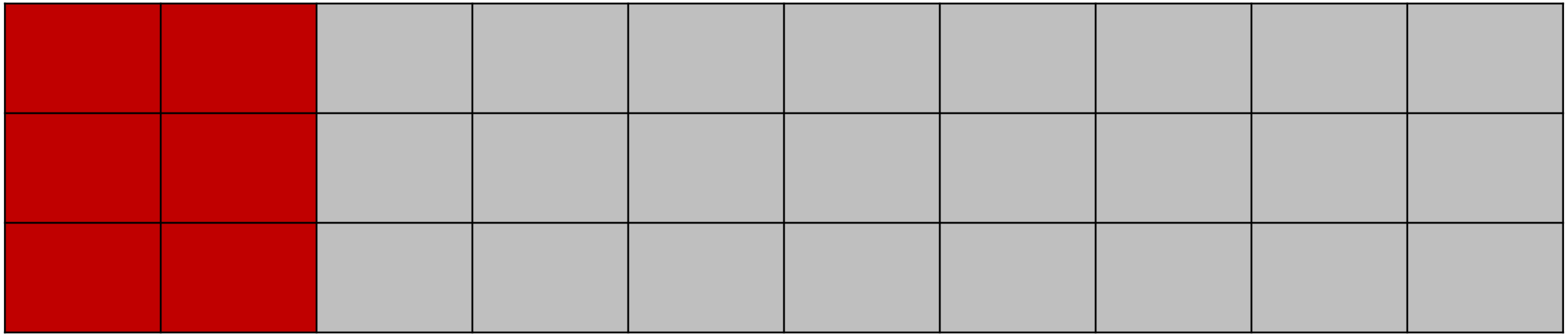
- non-important SNPs:

  ~6.9 million
- **important SNPs: 1327**

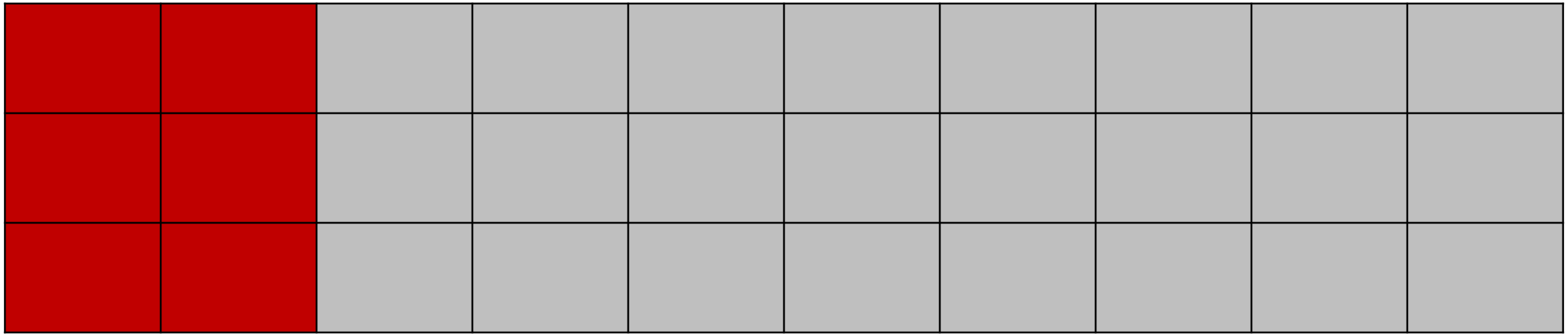
p >> n , ~7 million SNPs



p >> n, 1327 SNPs

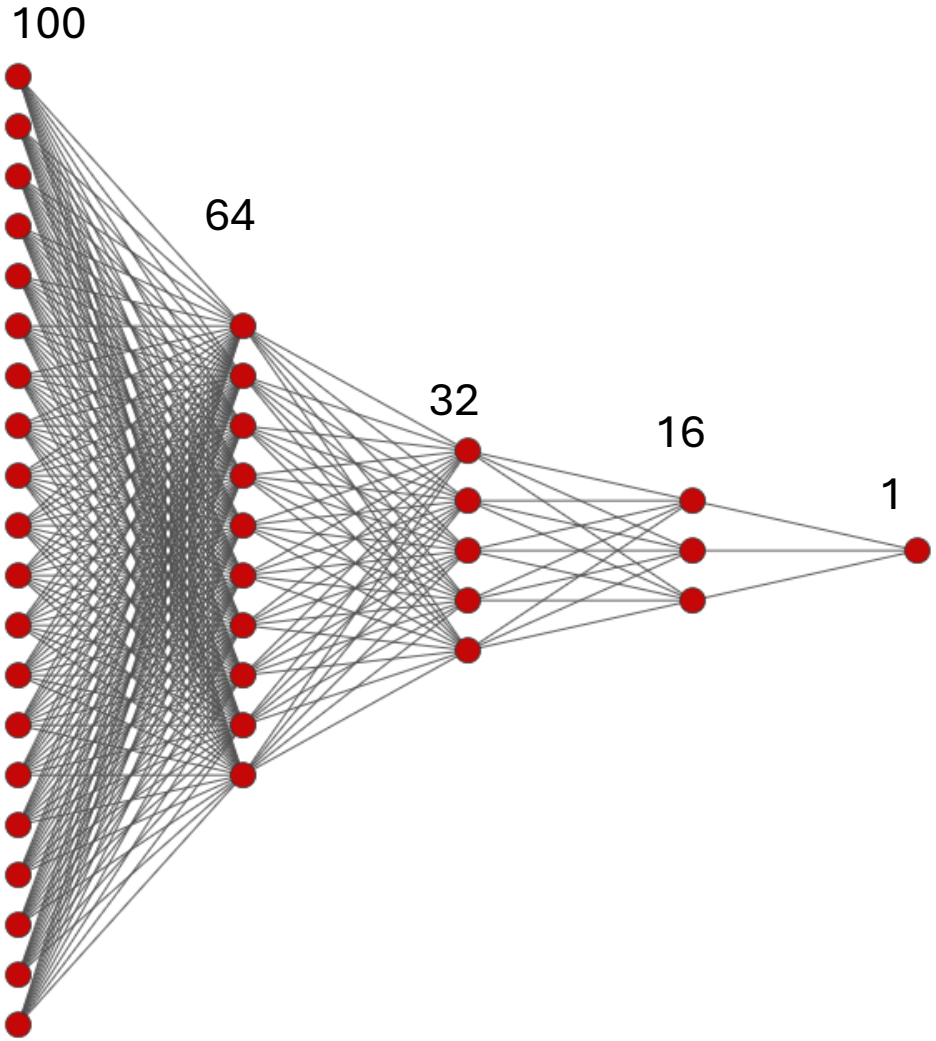


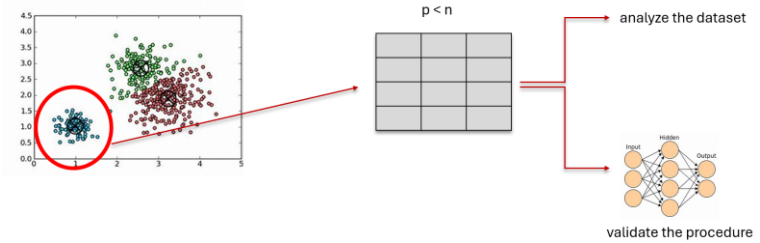
p >> n, 1327 SNPs



Are these predictors actually good?

validation

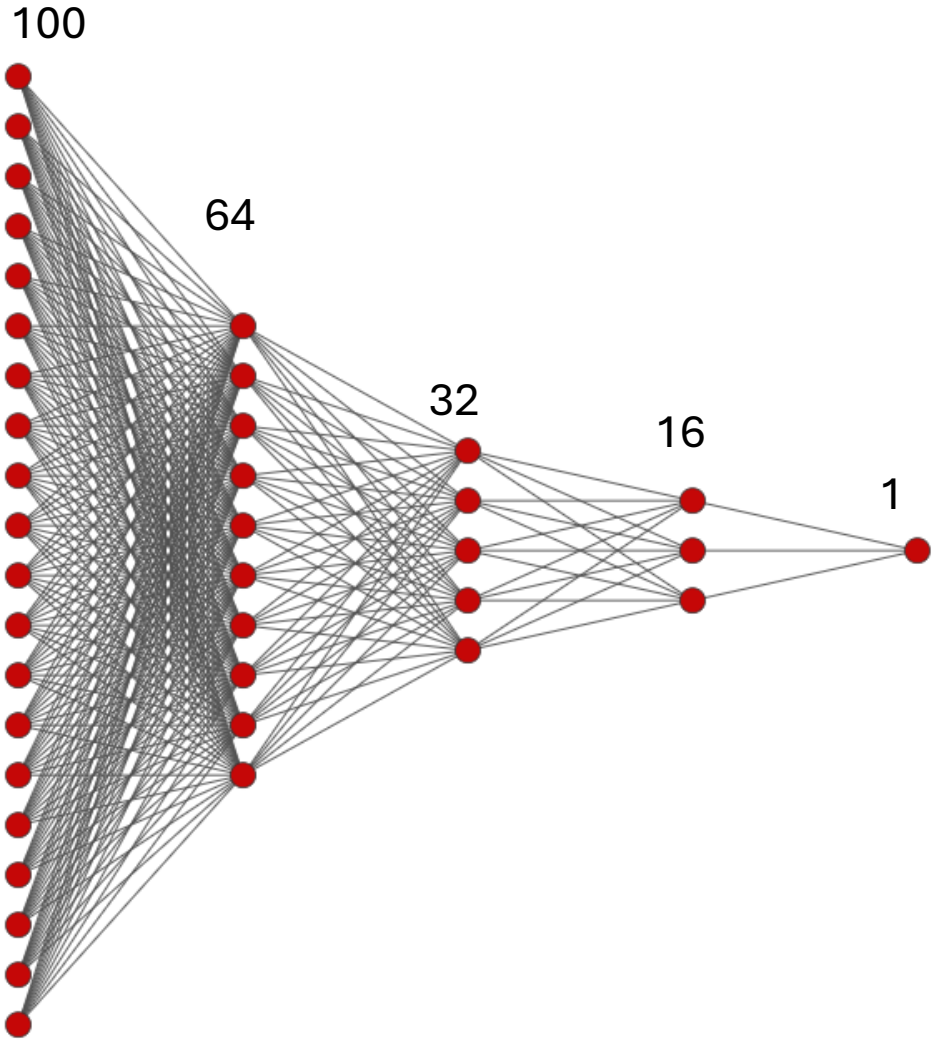


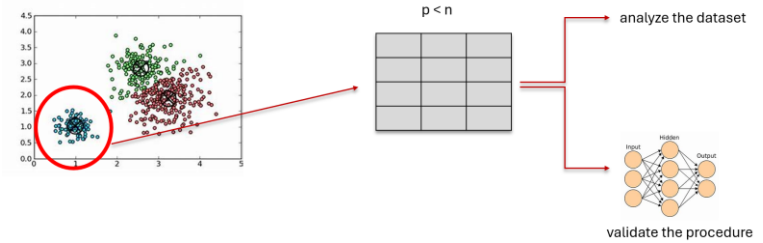


- input: PCA of original feature space (1st 100 PCs)
- binary output

NN SVG (alexlenail.me)

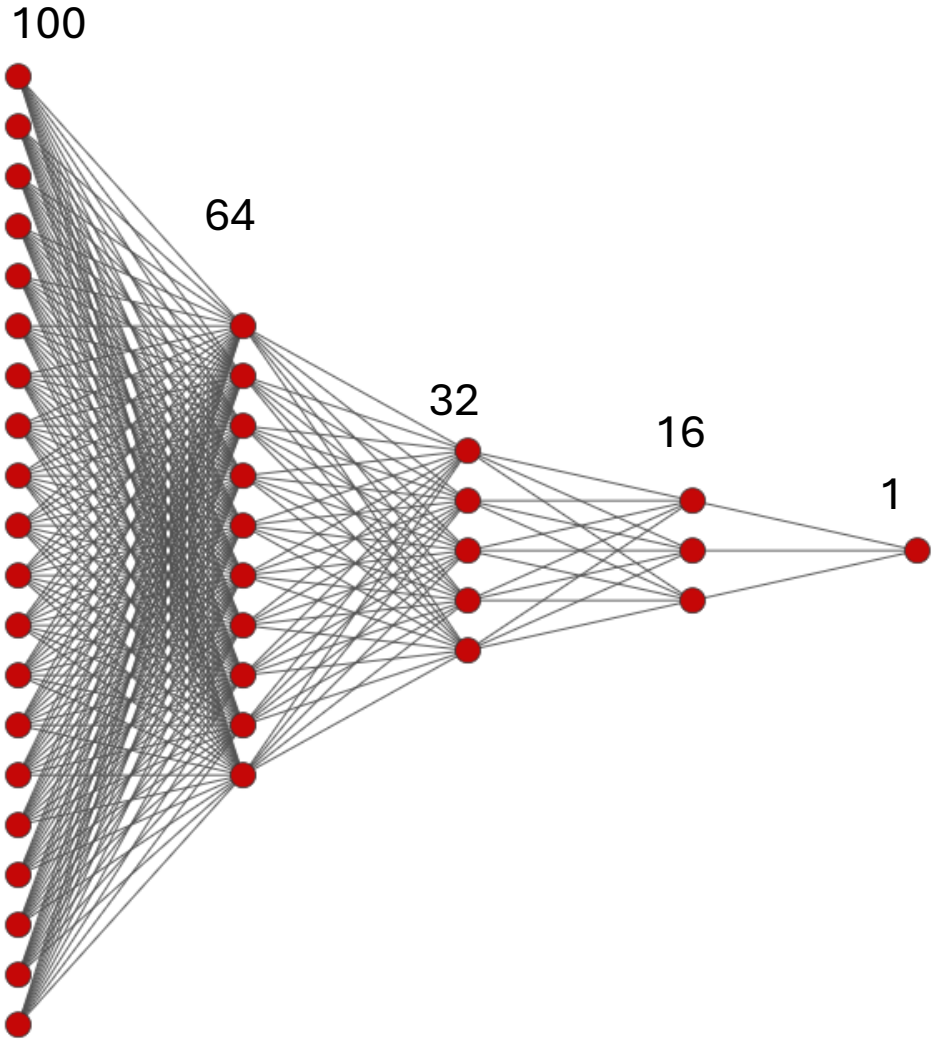
validation

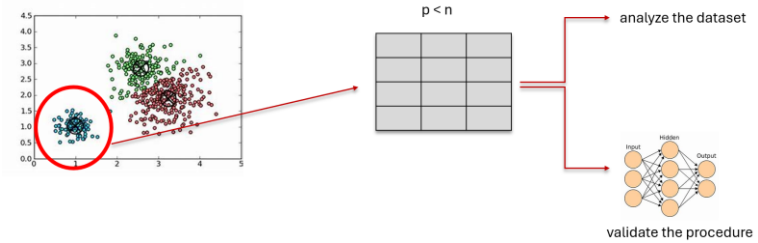




- input: PCA of original feature space (1st 100 PCs) ☝
- binary output
- 4 fully connected layers

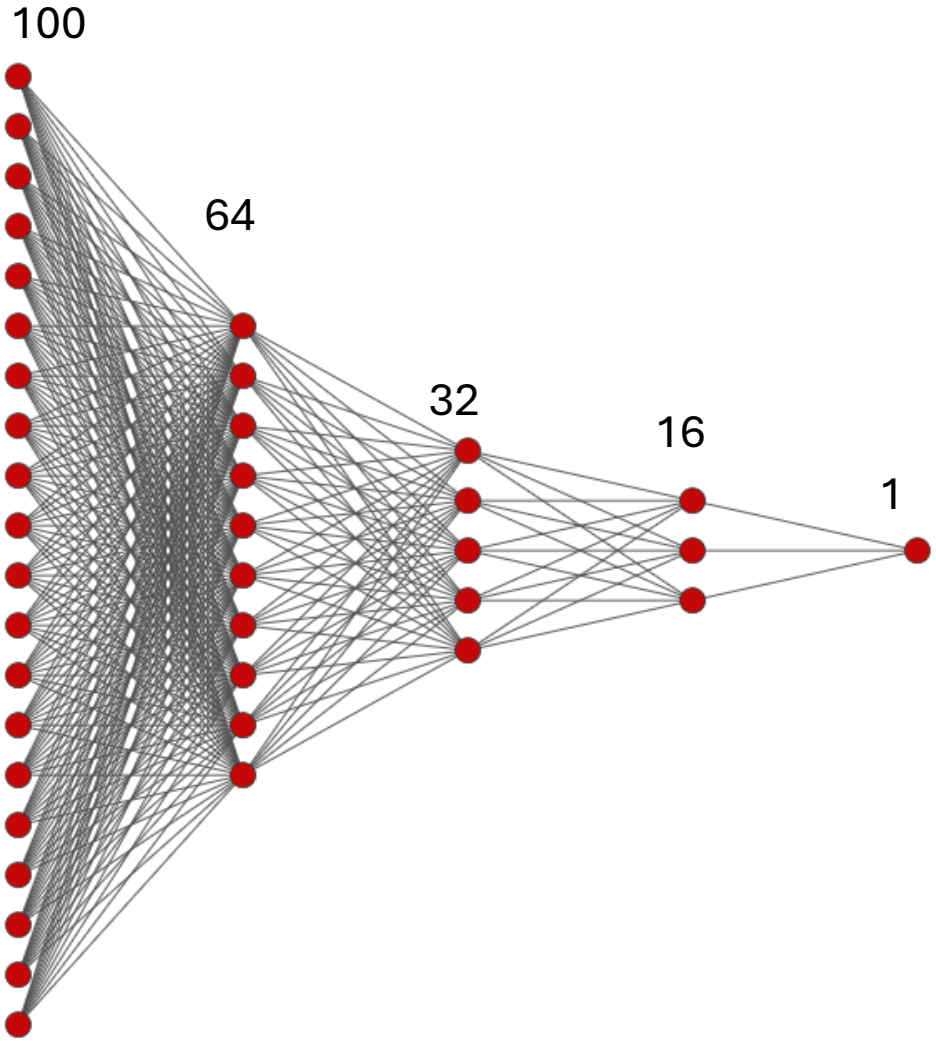
validation

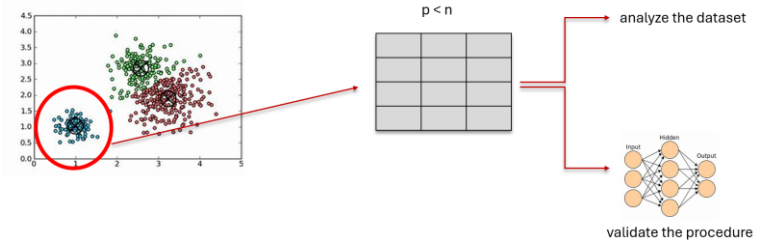




- input: PCA of original feature space (1st 100 PCs)
- binary output
- 4 fully connected layers
- Sigmoid activation function after output layer
- ReLU elsewhere

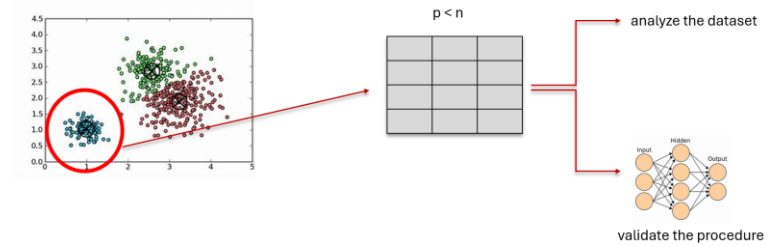
validation

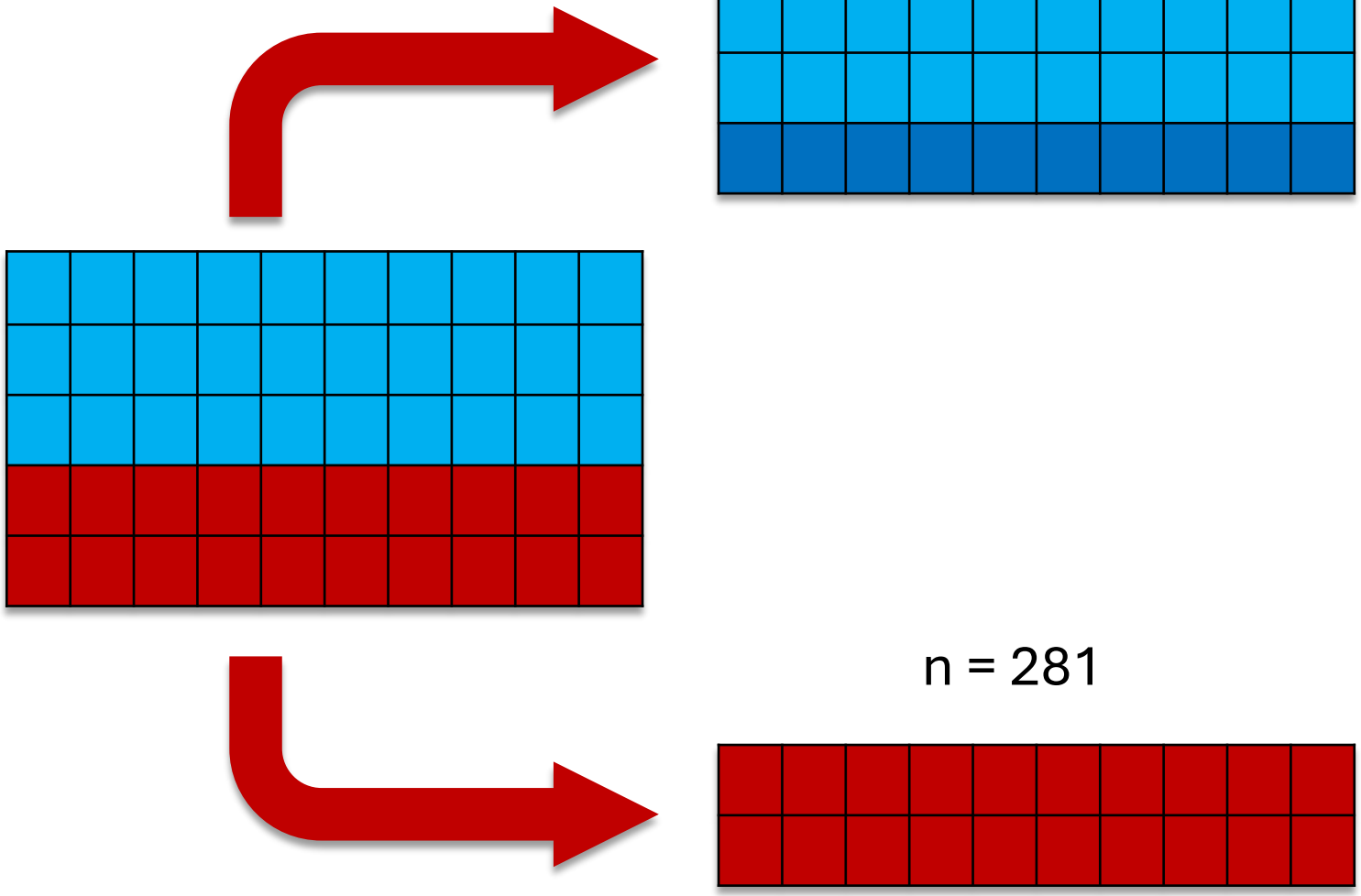




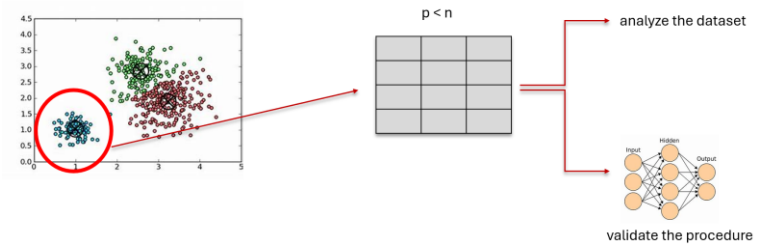
- input: PCA of original feature space (1st 100 PCs) ☝
- binary output
- 4 fully connected layers
- Sigmoid activation function after output layer
- ReLU elsewhere
- Adam
- Binary Cross Entropy loss

validation

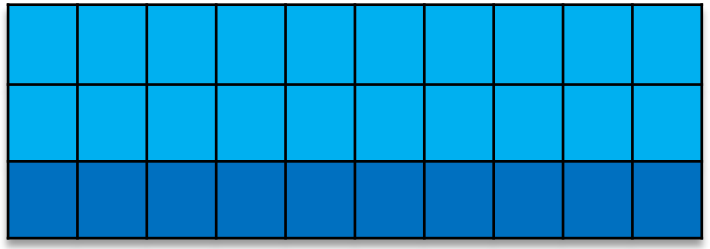


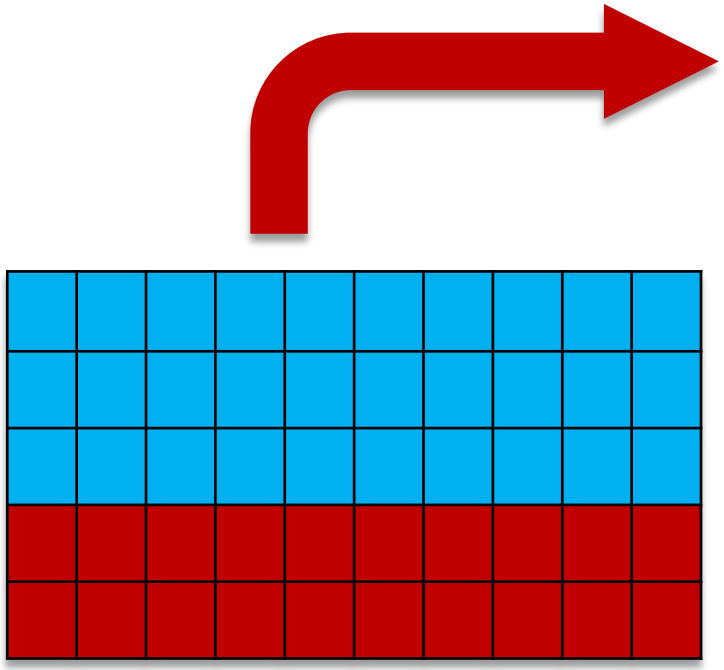


validation

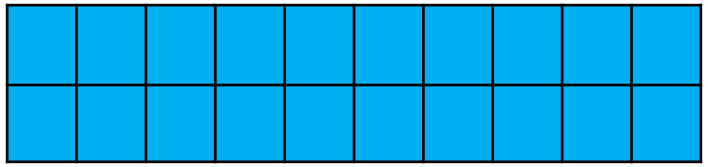


n = 844



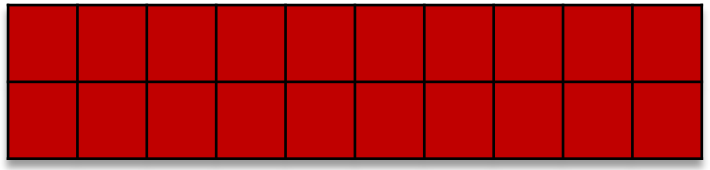


n = 566

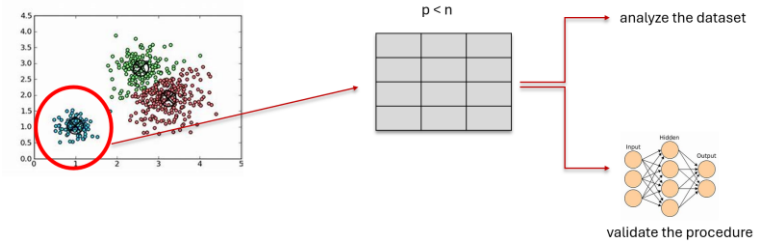


n = 278

n = 281



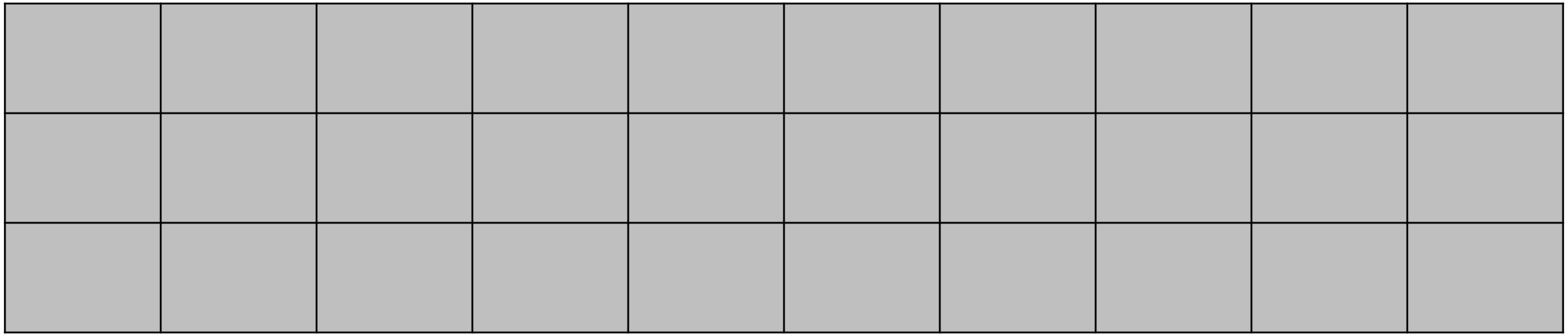
validation



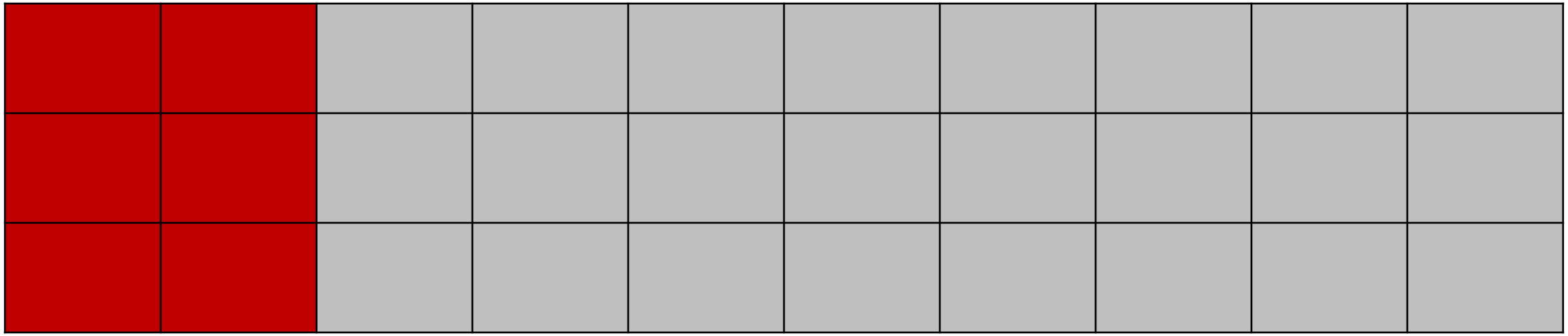
- F1 score (validation set) = 0.25

		true	
		severe COVID-19	**not-severe COVID-19**
predicted	**severe COVID-19**	**21**	**26**
	not-severe COVID-19	**99**	**135**

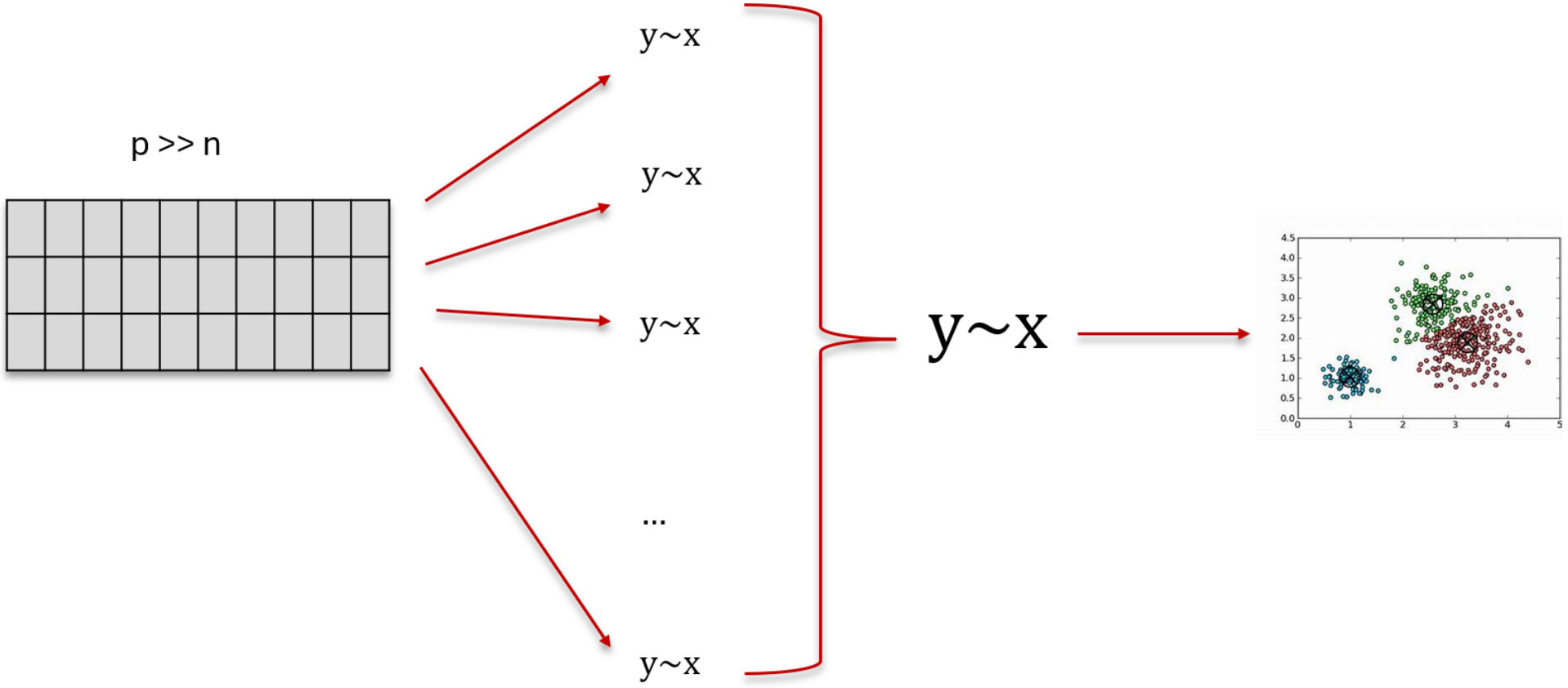
recap



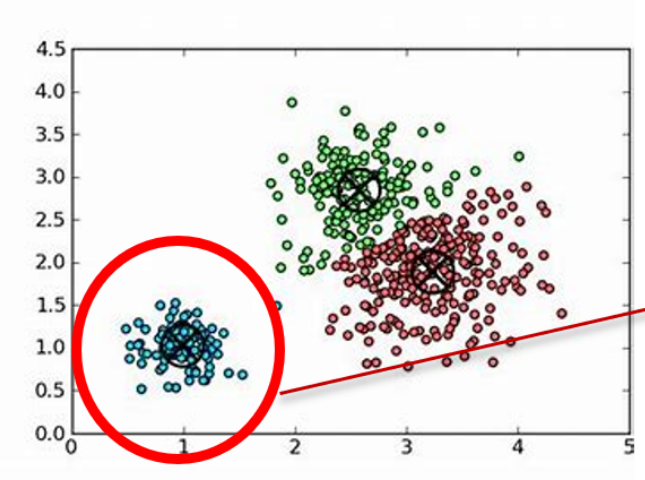
recap



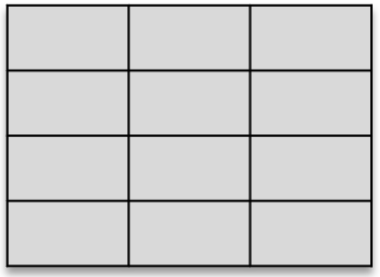
recap



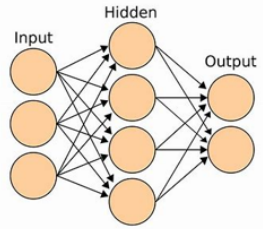
recap



p < n



analyze the dataset



validate the procedure

conclusions

- the p >> n setting is a common problem in data analysis

conclusions

- the p >> n setting is a common problem in data analysis
- a modeling scheme to select a subset of most important predictors was presented

conclusions

- the p >> n setting is a common problem in data analysis
- a modeling scheme to select a subset of most important predictors was presented
- applicable to various fields beyond bioinformatics

conclusions

- the p >> n setting is a common problem in data analysis
- a modeling scheme to select a subset of most important predictors was presented
- applicable to various fields beyond bioinformatics
- validation of the procedure needs tuning

thanks

- Joanna Szyda
- Paula Dobosz
- Dawid Słomian
- Krzysztof Kotlarz

thank you for your attention

WROCŁAW UNIVERSITY OF ENVIRONMENTAL AND LIFE SCIENCES

CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN

Freie Universität Berlin

UNIWERSYTET MEDYCZNY IM. KAROLA MARCINKOWSKIEGO W POZNANIU