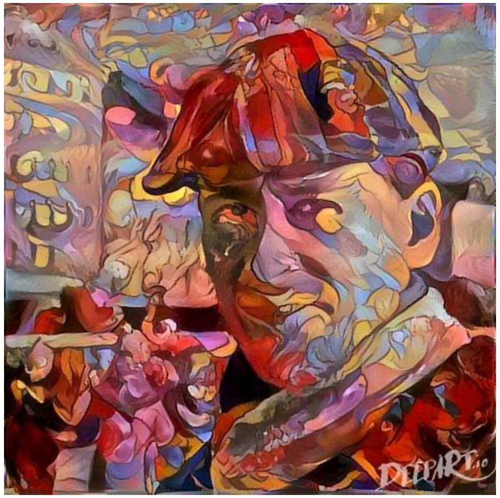
Marcin Kierczak, NBIS

AI, ML & GENETICS

UPPSALA UNIVERSITET



Biology + math/comp. sci.

BSc molecular biology, Warsaw

MSc bioinformatics, Uppsala

PhD bioinformatics, LCB, Uppsala — ML, feature selection

SLU, Carlborg’s Lab, postdoc — ML, GWAS, R

UU, Lindblad-Toh Lab — canine genetics, GWAS, population genetics

Associate professor, bioinformatics, TekNat, UU

NBIS Sweden — ML, AI, reproducible research, aDNA, RaukR

Outside academia: woodworking, electronics, plants

UPPSALA UNIVERSITET

Artificial Intelligence & co.

Artificial Intelligence (AI)

The effort to automate intellectual tasks normally performed by humans.[1]

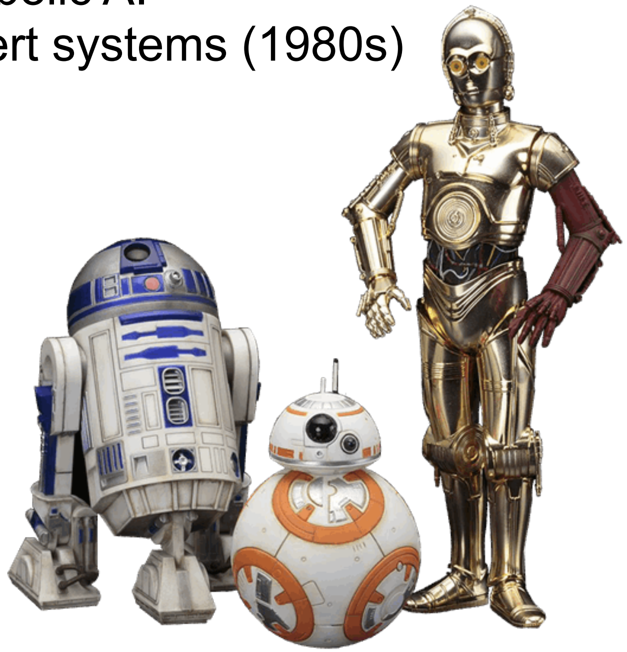
[1] *Chollet & Allaire. Deep learning with R.*

Artificial Intelligence & co.

Artificial Intelligence (AI)

The effort to automate intellectual tasks normally performed by humans.[1]

- 1950s
- symbolic AI
- expert systems (1980s)



[1] *Chollet & Allaire. Deep learning with R.*

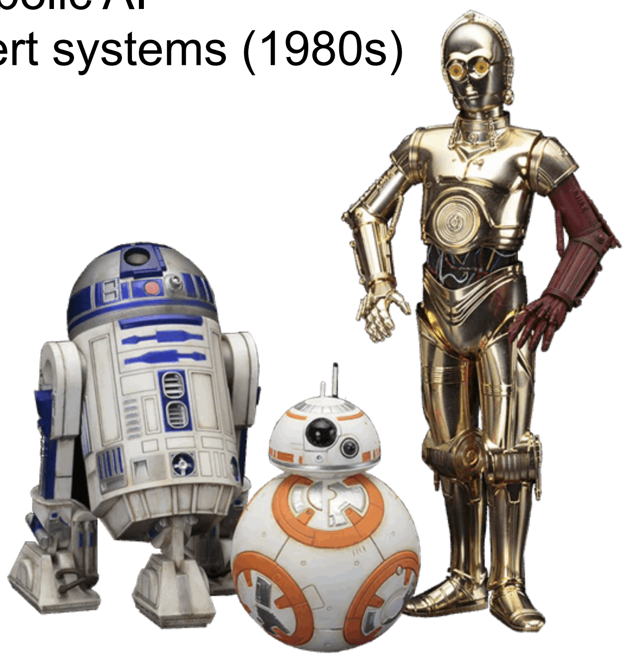
UPPSALA UNIVERSITET

Artificial Intelligence & co.

Artificial Intelligence (AI)

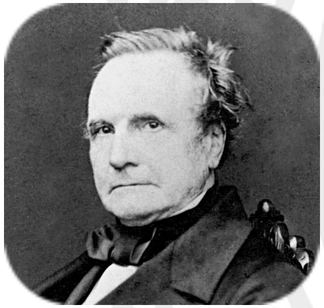
The effort to automate intellectual tasks normally performed by humans.[1]

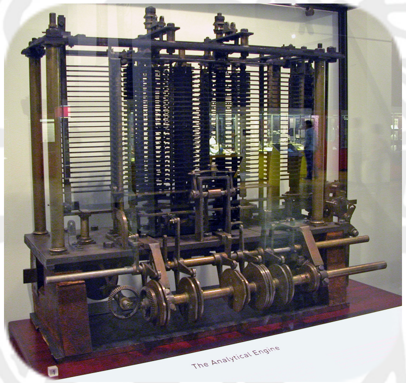
- 1950s
- symbolic AI
- expert systems (1980s)



The Analytical Engine has no pretentions whatever to originate anything. It can do whatever we know how to order it to perform…

Lady Ada Lovelace





[1] *Chollet & Allaire. Deep learning with R.*

images: commons.wikimedia.com

Machine Learning

Artificial Intelligence (AI)

The effort to automate intellectual tasks normally performed by humans.[1]

Allan Turing

- *Computing Machinery and Intelligence*
- Turing test

Data →
Rules →
PROGRAM
→ Answers

[1] *Chollet & Allaire. Deep learning with R.*

images: commons.wikimedia.com

Machine Learning

Artificial Intelligence (AI)

The effort to automate intellectual tasks normally performed by humans.[1]

Allan Turing

- *Computing Machinery and Intelligence*
- Turing test

Data → PROGRAM → **Rules**

Answers →

[1] *Chollet & Allaire. Deep learning with R.*

images: commons.wikimedia.com

Machine Learning

Artificial Intelligence (AI)

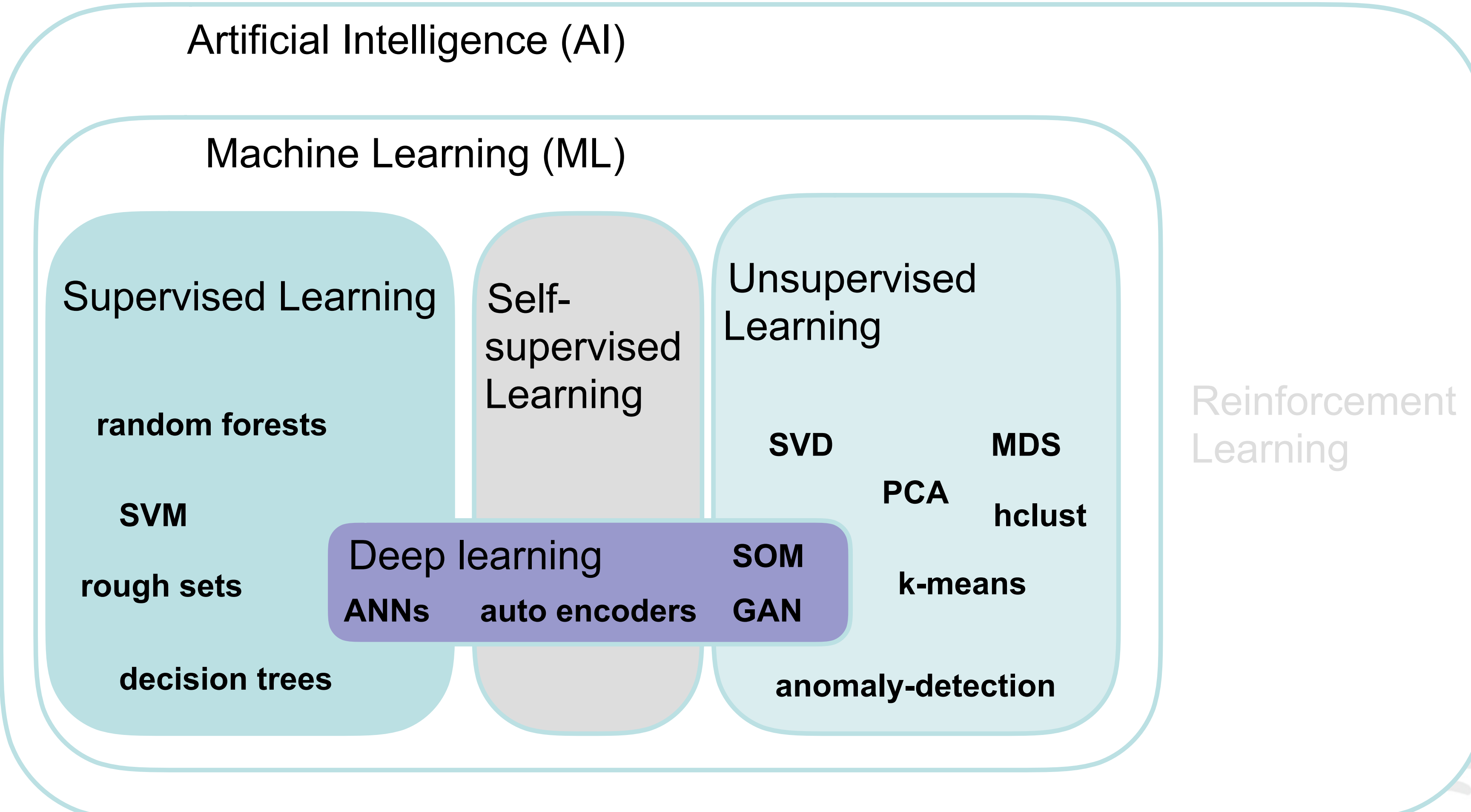
The effort to automate intellectual tasks normally performed by humans.

Machine Learning (ML)

Machine Learning is a field of study that gives computers the ability to learn without being explicitly programmed.

Arthur Samuel, 1959

Overview of the field



SVM - support-vector machines
ANN - artificial neural network
SVD - singular value decomposition
PCA - principal comp. analysis
SOM - self-organising maps
MDS- multidimensional scaling
GAN - generative adversarial nets
hclust- hierarchical clustering

Why deep learning

P variables

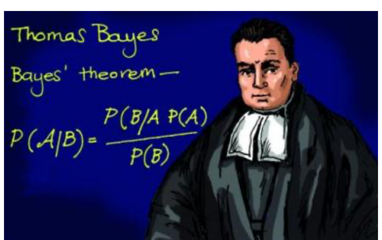
	Attr. 1	Attr. 2	…	Attr. P	d
Ex. 1	AA	big	…	11.54	**case**
Ex. 2	TA	small	…	-5.48	**control**
…	…	…	…	…	**…**
Ex. N	TA	medium	…	7.26	**control**

N

amount of data

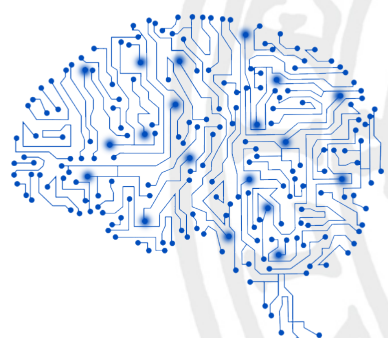
Reinforcement Learning

P >> N



P ~= N

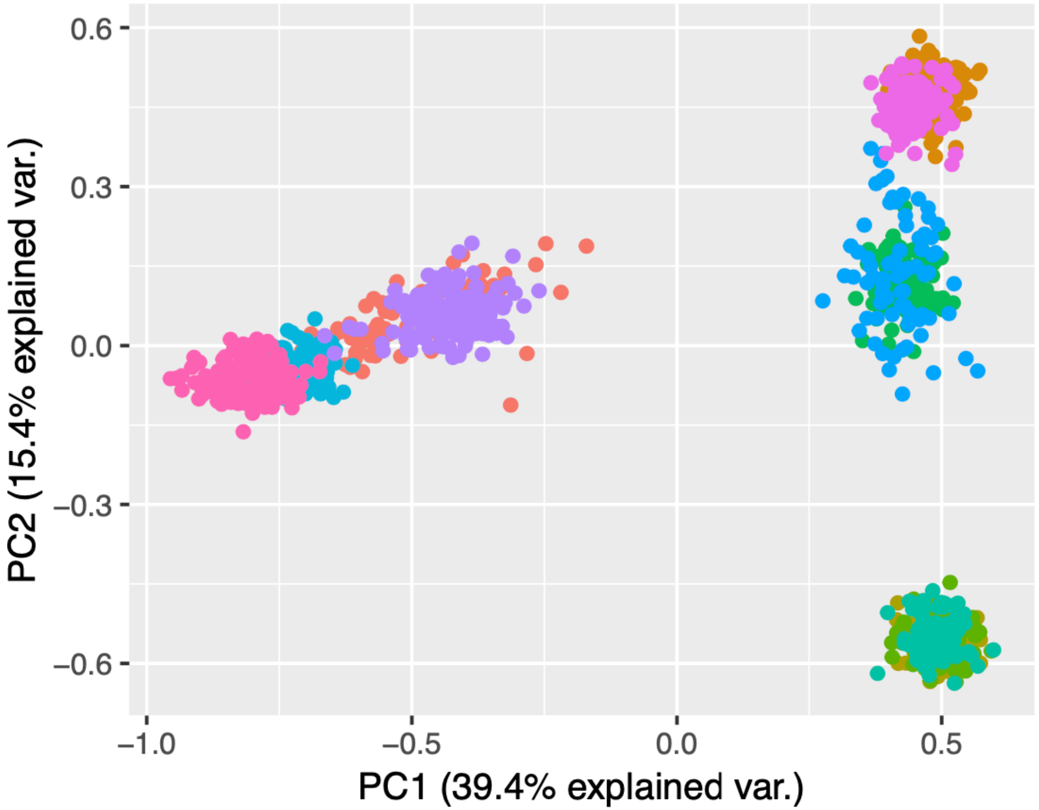
N >> P



UPPSALA UNIVERSITET

Unsupervised learning

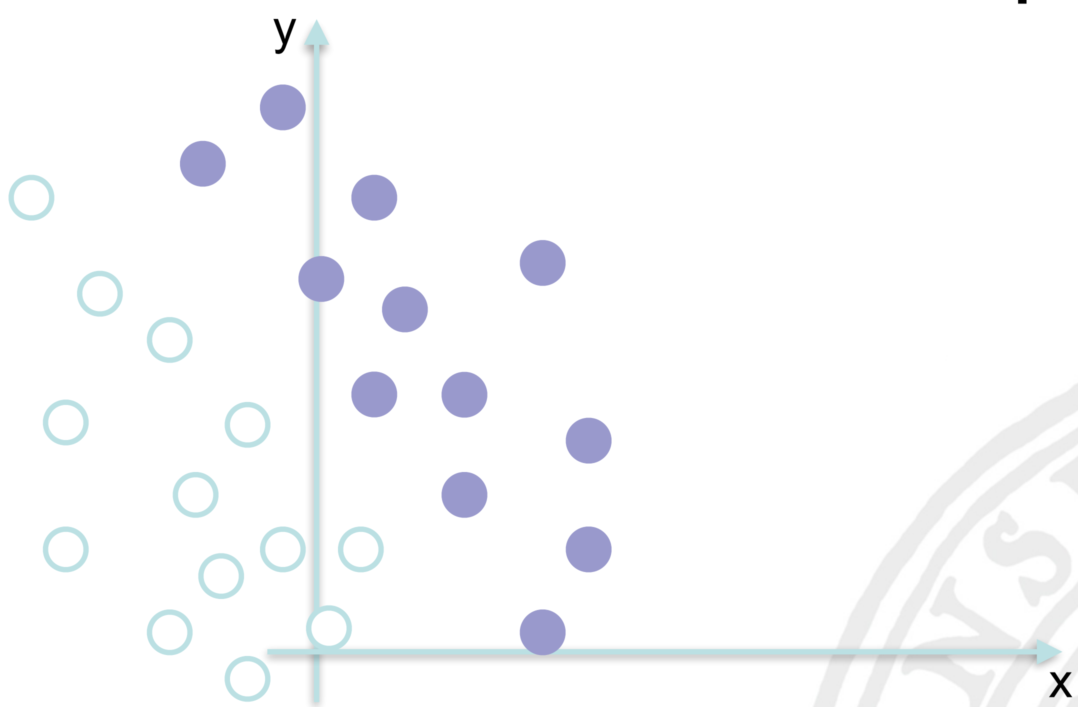
	Attr. 1	Attr. 2	…	Attr. P
Ex. 1	A	G	…	C
Ex. 2	T	A	…	T
…	…	…	…	…
Ex. N	T	G	…	T



Kierczak et al. 2019 in press.

Geometric interpretation

Geometric interpretation



IF x < 0 THEN white ELSE blue

Geometric interpretation

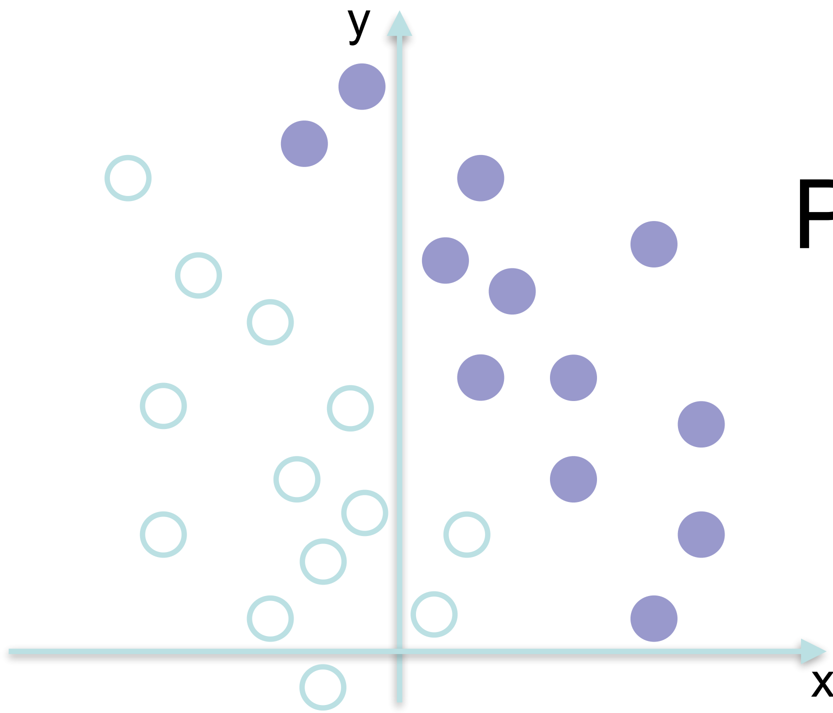
IF x < 0 THEN white ELSE blue

Formal definition of ML

A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with E.

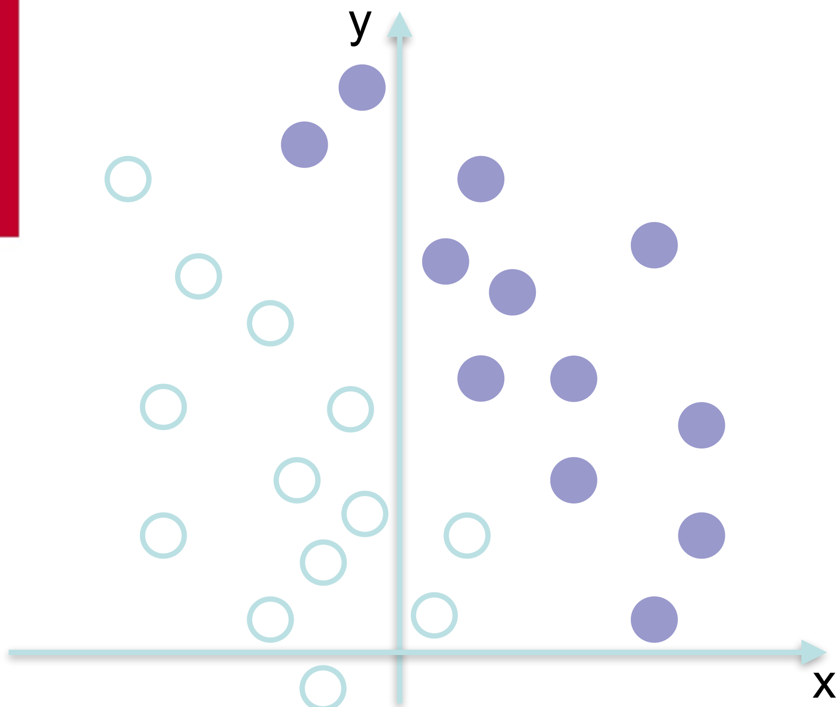
Tom Mitchell, 1997

Performance P



		Actual	
		●	○
Predicted	●	TP	FP
	○	FN	TN

Performance P

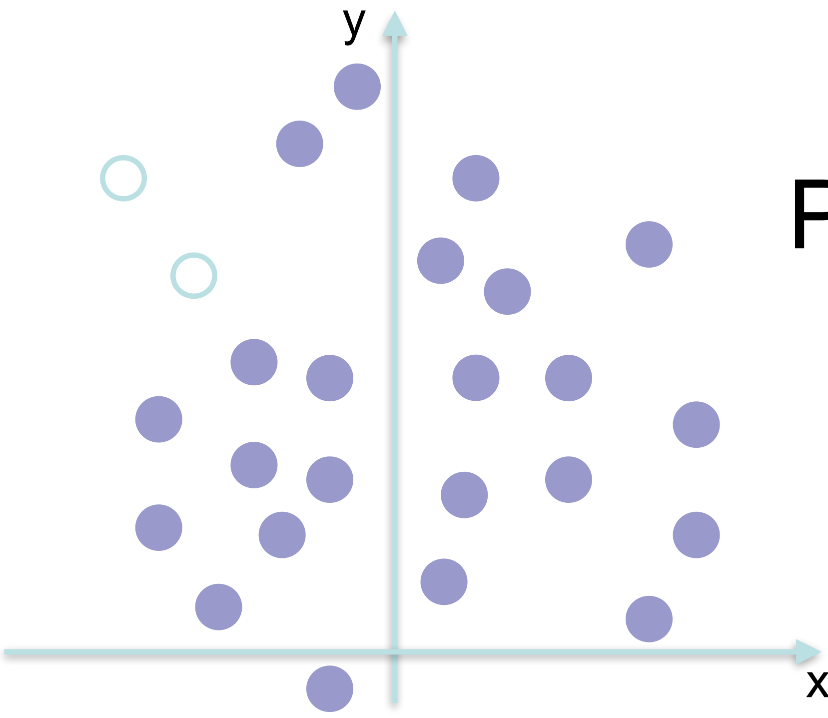


13 white
12 blue

- – class bias
- – misclassification cost

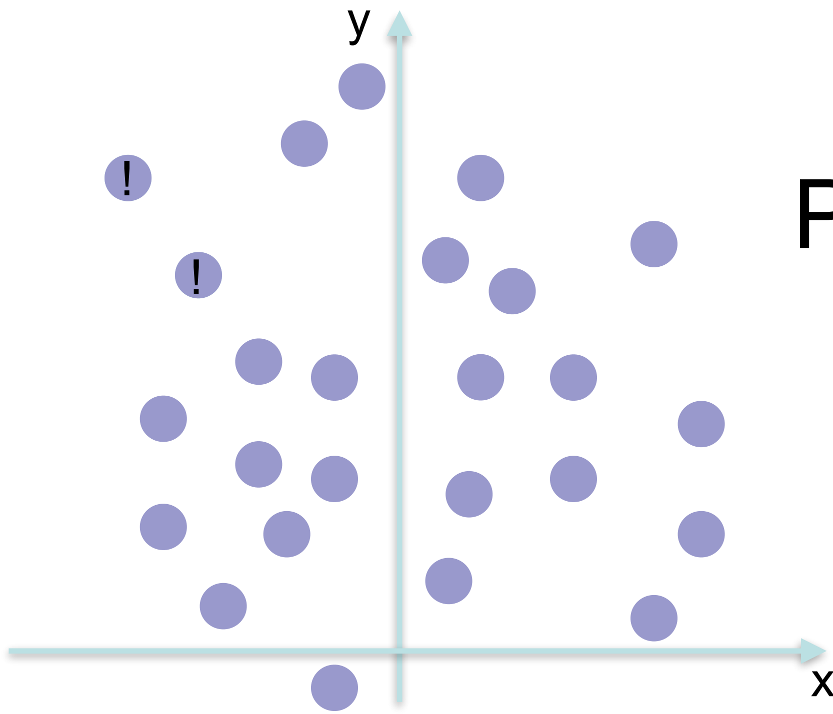
		Actual	
		●	○
Predicted	●	10	2
	○	2	11

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} = \frac{10 + 11}{10 + 11 + 2 + 2} = \frac{21}{25} = 0.84$$



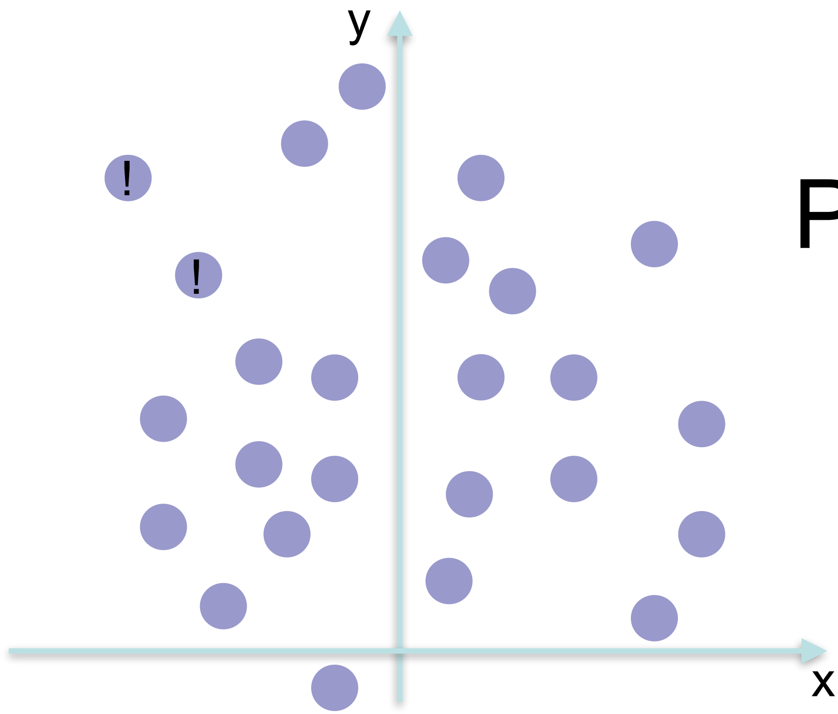
Performance P

Performance P



		Actual	
		●	○
Predicted	●	23	2
	○	0	0

Performance P



		Actual	
		●	○
Predicted	●	23	2
	○	0	0

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} = \frac{23 + 0}{23 + 0 + 2 + 0} = \frac{23}{25} = 0.92$$

UPPSALA UNIVERSITET

Performance P

		Actual	
		●	○
Predicted	●	10	2
	○	11	2

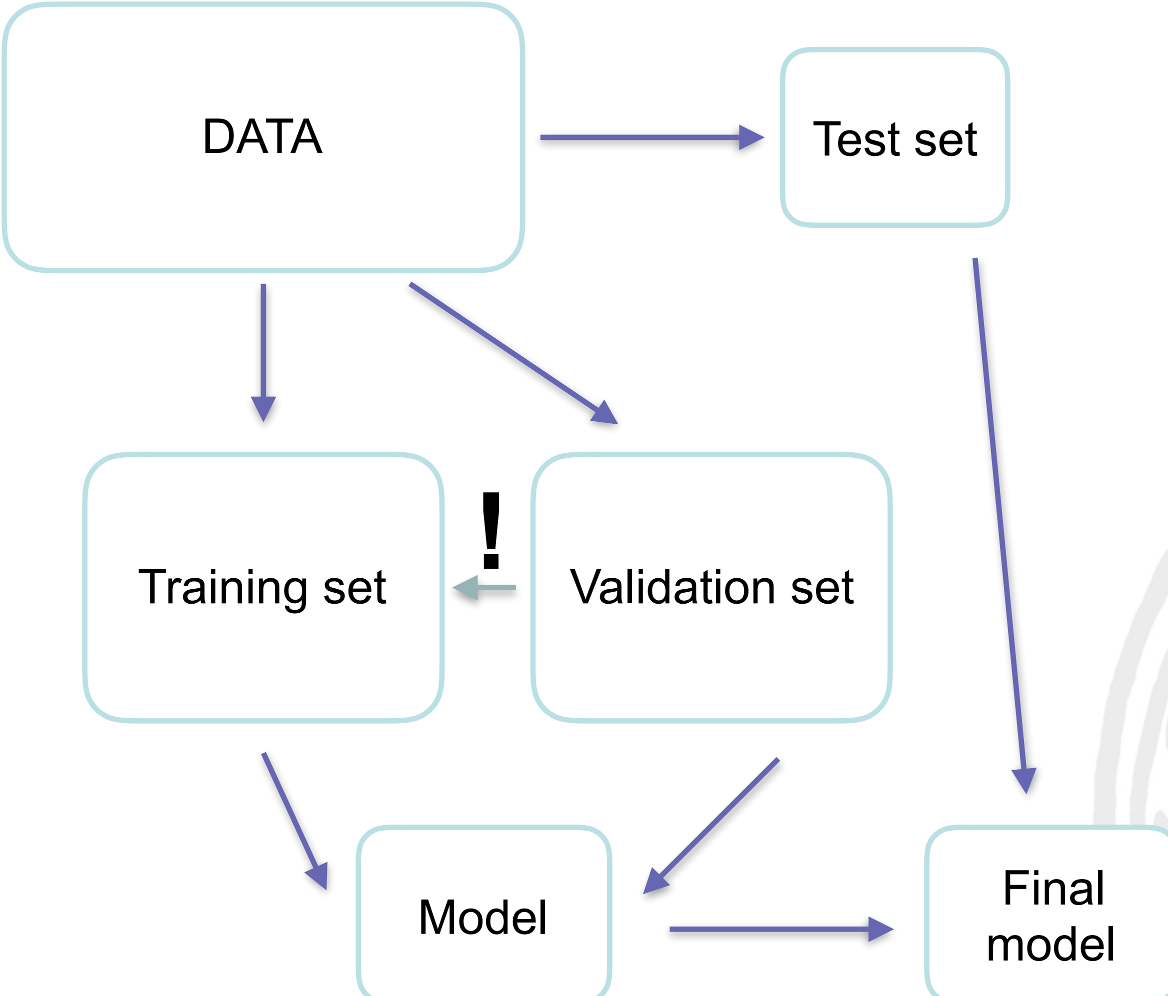
$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$ fraction of correctly predicted sick to all predicted to be sick

$$Sensitivity = \frac{TP}{TP + FN}$$ fraction of correctly predicted sick to all sick

$$Specificity = \frac{TN}{TN + FP}$$ fraction of correctly predicted healthy to all healthy

Measuring performance



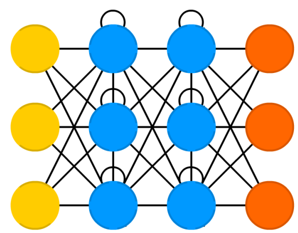
- class distribution
- information leaks
- temporal leaks
- duplication leaks

Performance baseline

IF x < 0 THEN white ELSE blue

Acc = 0.74

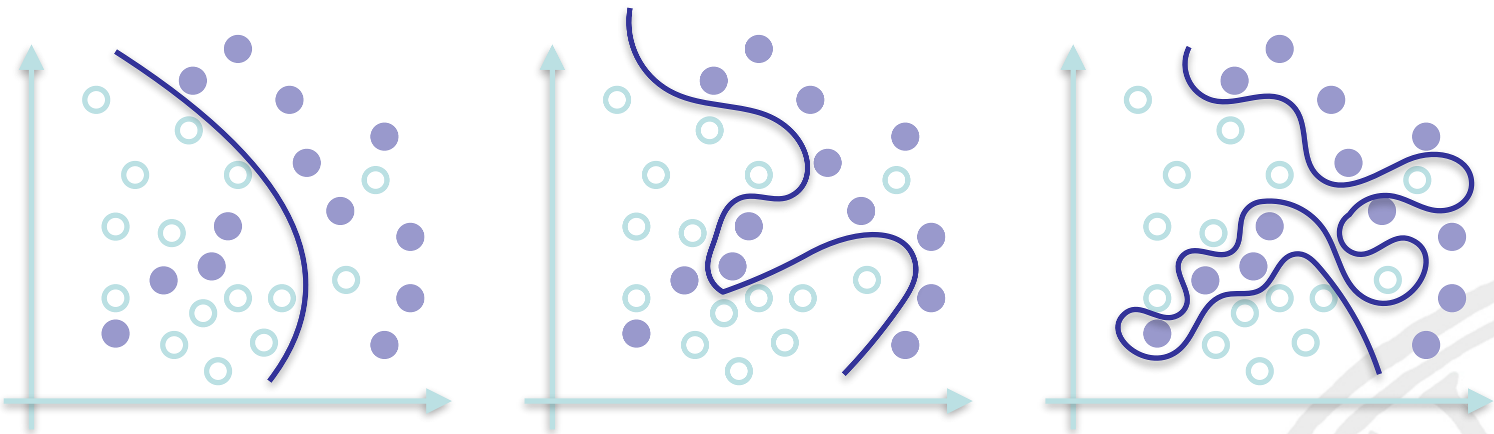
Acc = 0.54



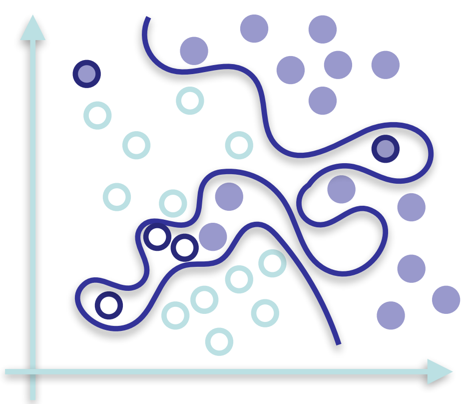
Acc = 0.72

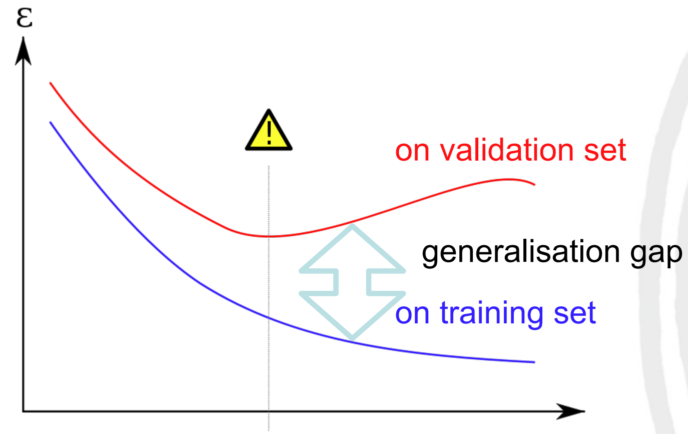
Acc = 0.80

Overfitting

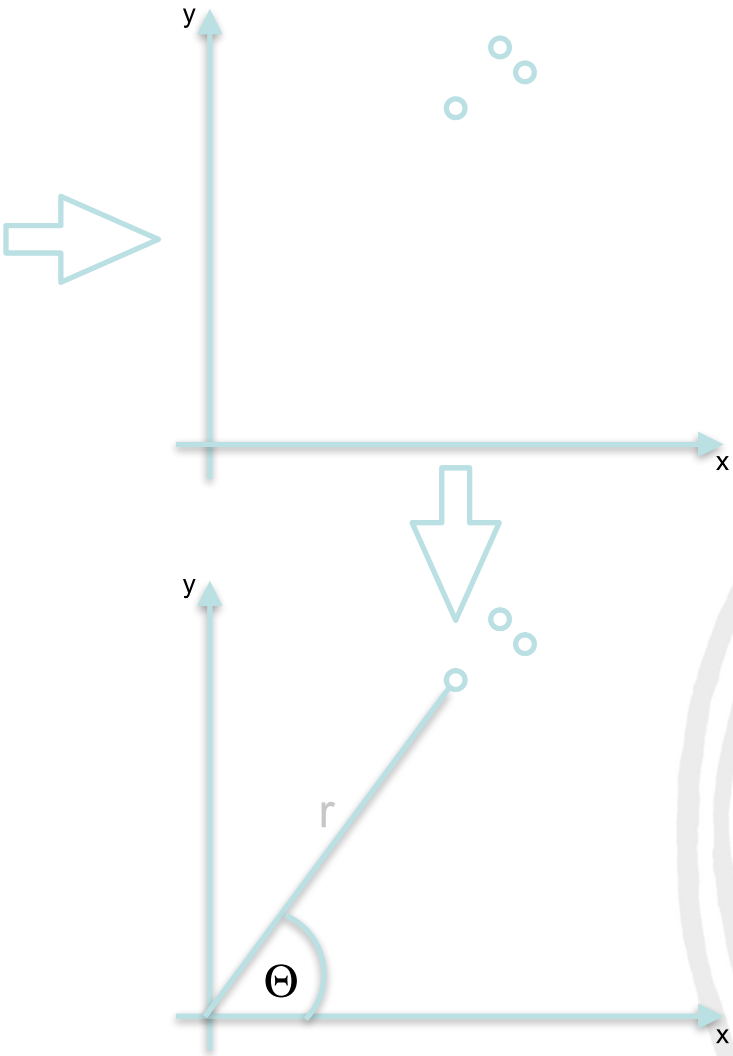


On new data

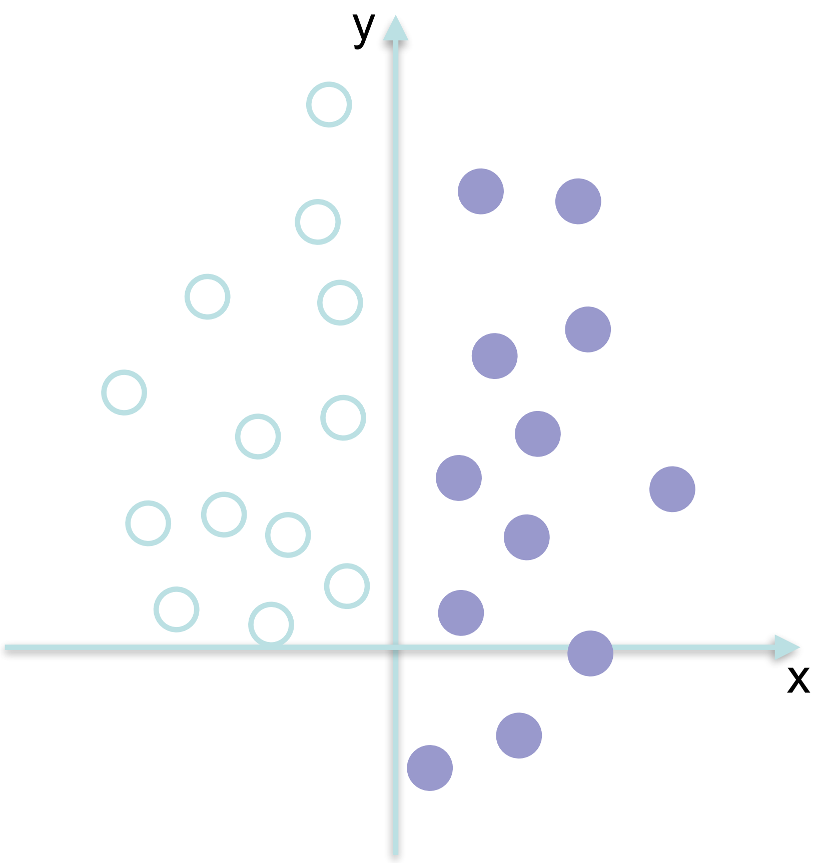




Feature engineering



Feature selection



Feature selection

	a1	a2	d
Ex. 1	1	7.45	**case**
Ex. 2	0	3.24	**control**
…	…	…	**…**
Ex. N-1	1	8.72	**case**
Ex. N	0	11.5	**control**

Acc = 0.98

Ex. 1	1	7.45	**case**
Ex. 2	1	3.24	**control**
…	…	…	**…**
Ex. N-1	0	8.72	**case**
Ex. N	0	11.5	**control**

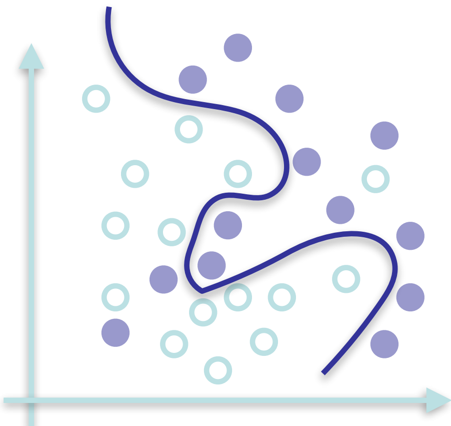
Acc = 0.95

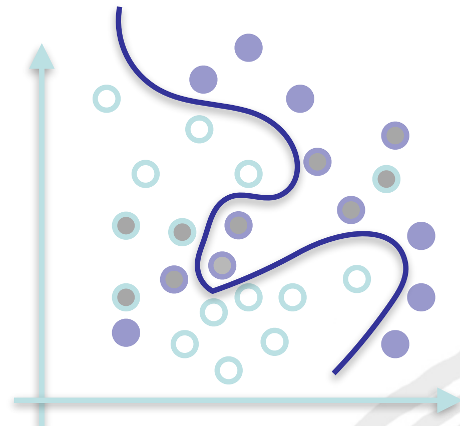
I(a2) > I(a1)

Ex. 1	1	8.72	**case**
Ex. 2	0	11.5	**control**
…	…	…	**…**
Ex. N-1	1	7.45	**case**
Ex. N	0	3.24	**control**

Acc = 0.75

Interpretability





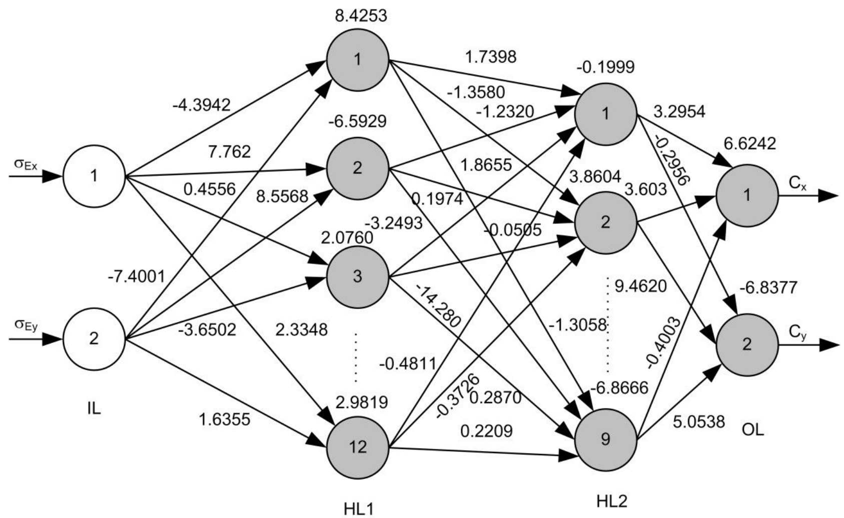
Input image

a couple of bears are standing in a field

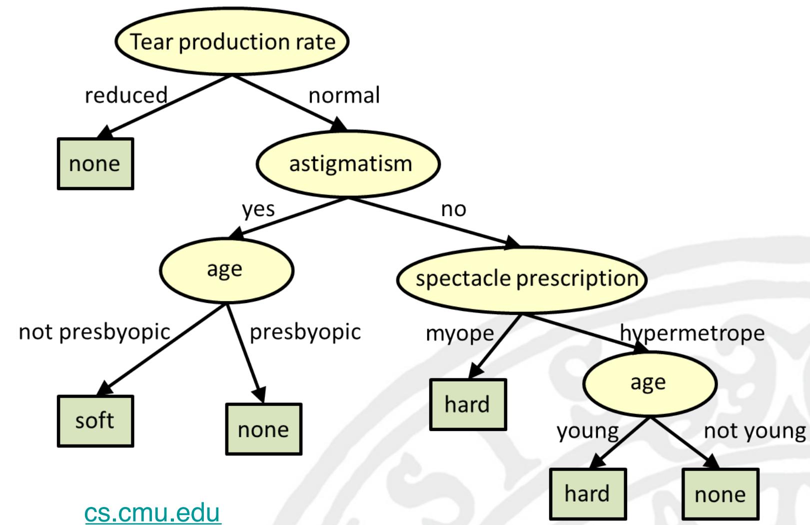
Output caption

Importance map

Explainability



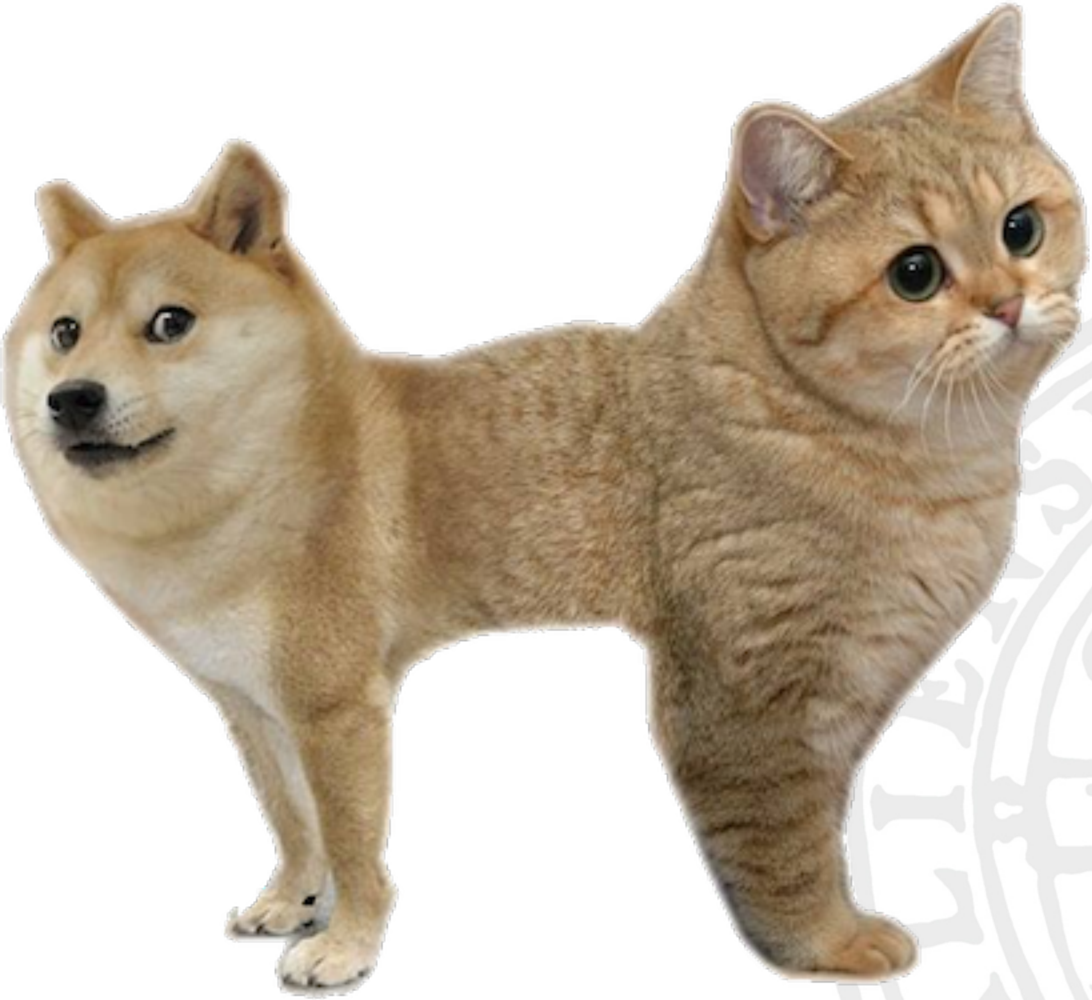
Arockia et al. Journal on Wireless Communications and Networking 2014:160



cs.cmu.edu

IF tear prod. rate = normal & astigmatism = yes & age != presbyopic THEN lenses = soft

Dealing with uncertainty

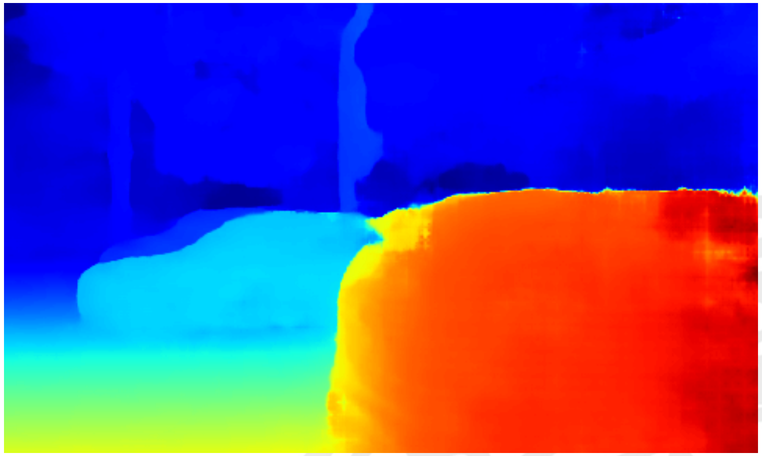


UPPSALA UNIVERSITET

Uncertainty

(1)

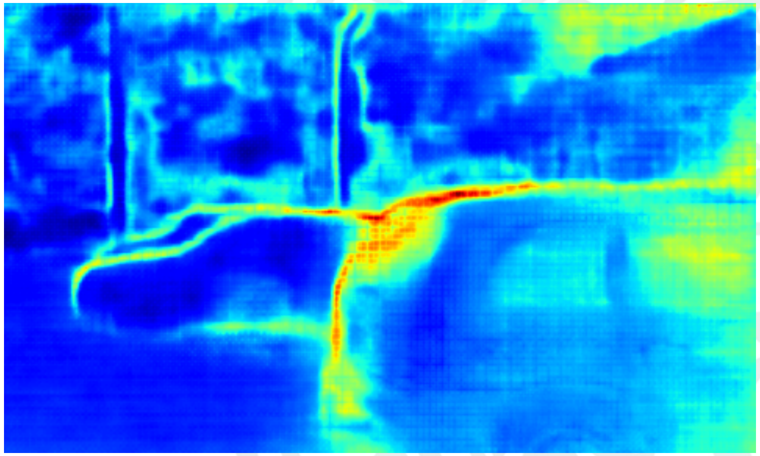
(2)



Uncertainty of depth estimation:

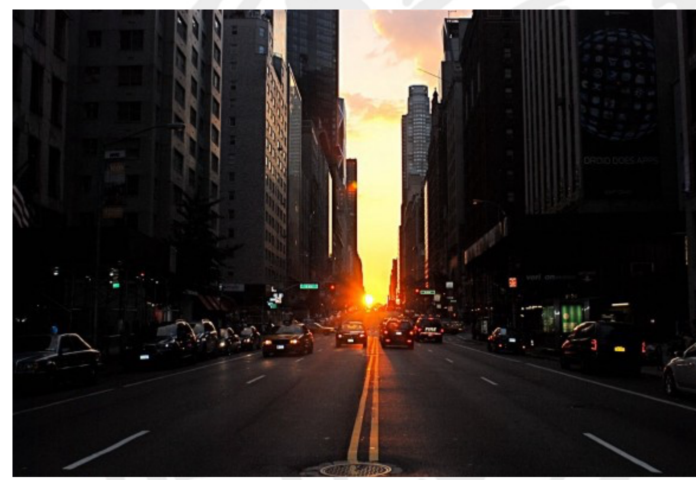
1. Original picture.
2. Estimated depth.
3. Estimation uncertainty.

(3)

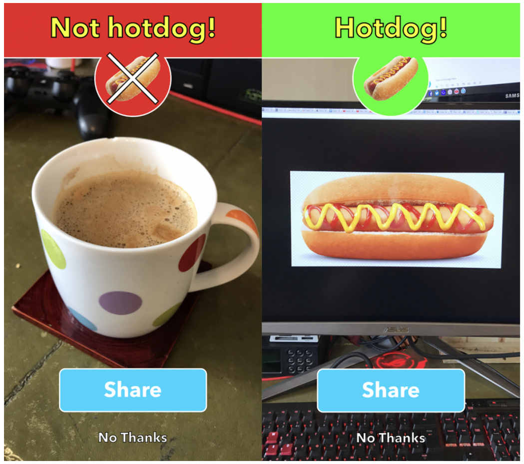


Aleatoric uncertainty

Aleatoric uncertainty is the sensing uncertainty.

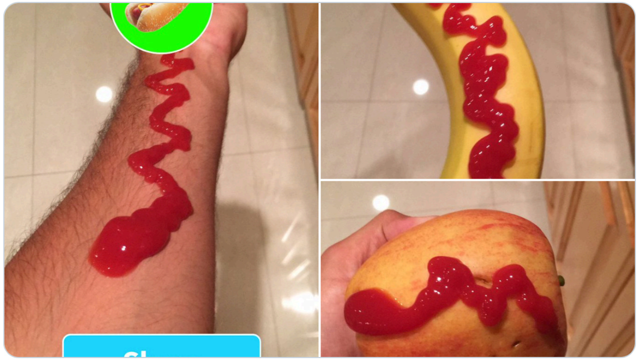


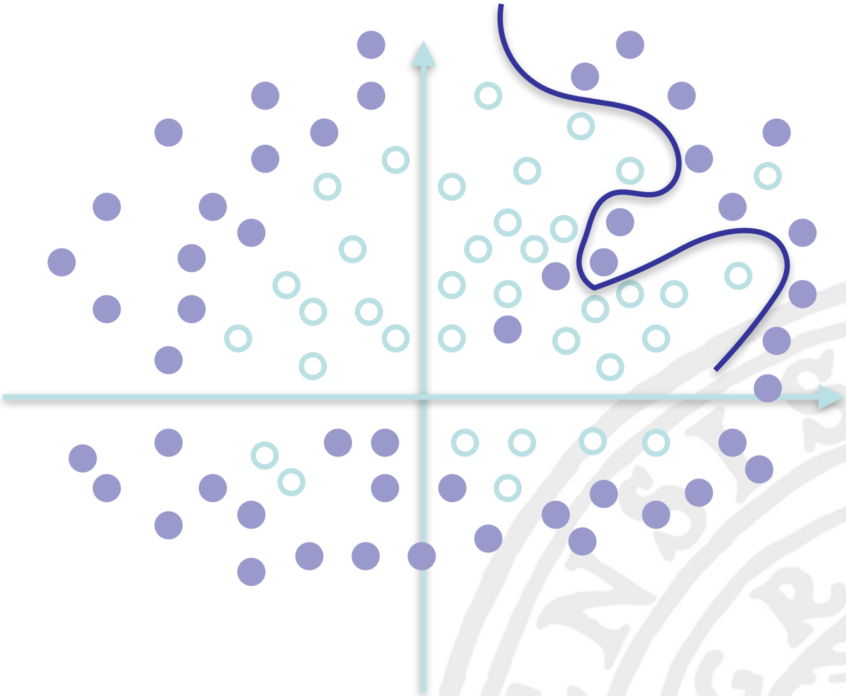
Epistemic uncertainty



David Khachatryan
@david__kha

If there's ketchup, it's a hotdog @FunnyAsianDude #nothotdog #NotHotdogchallenge

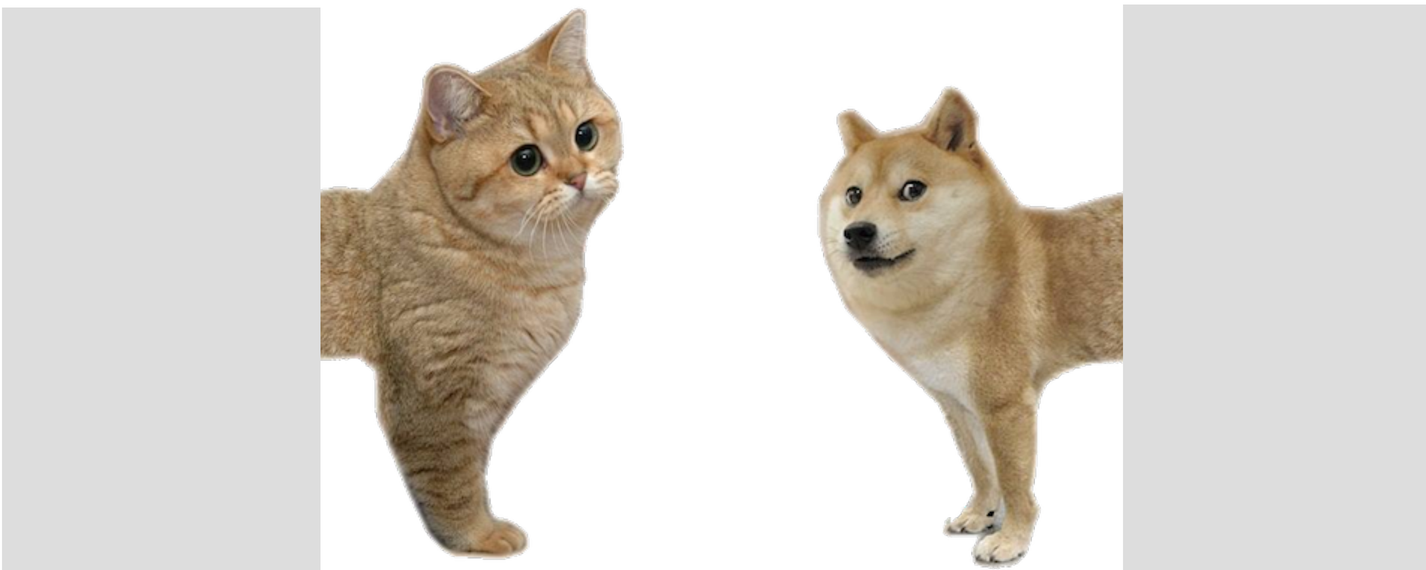




Epistemic uncertainty – related to the model.

- can be fixed given more training data

Assessing uncertainty

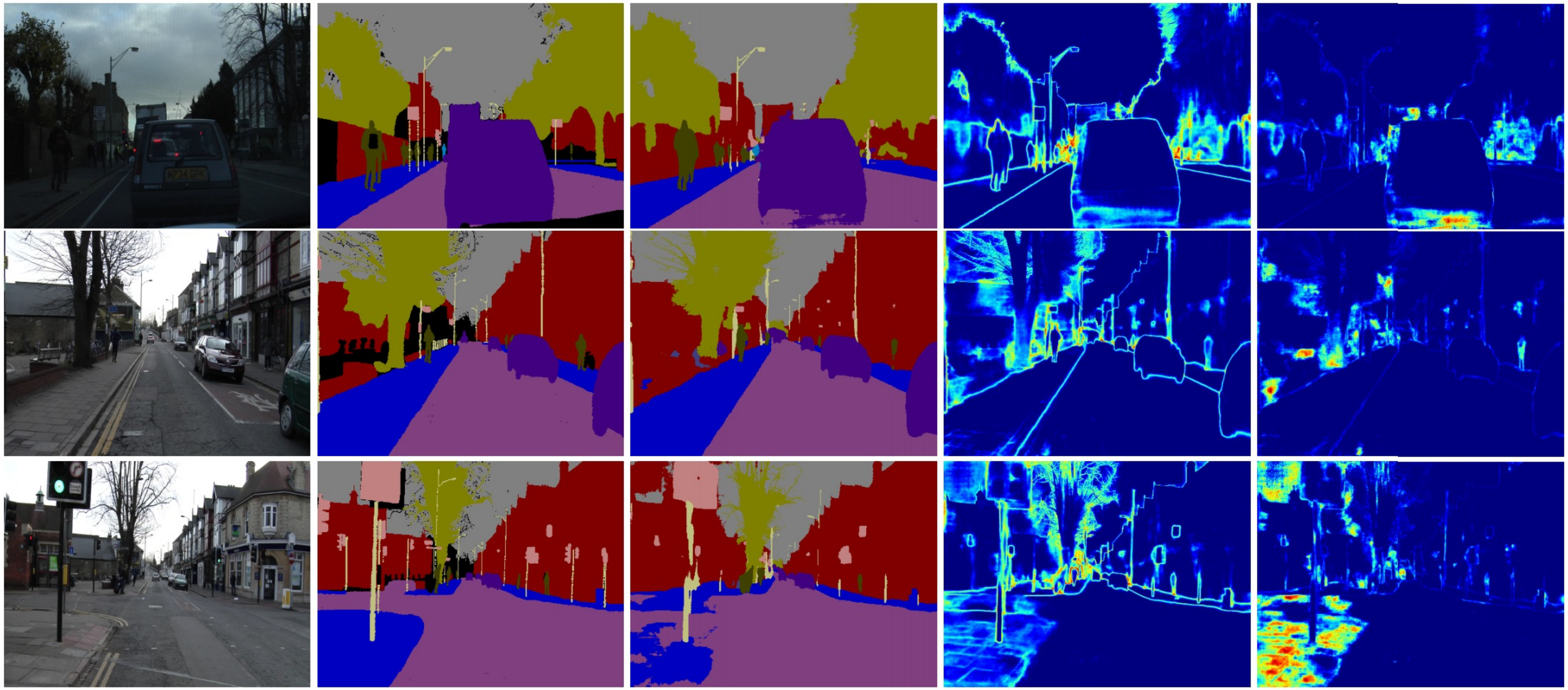


100% cat

100% dog

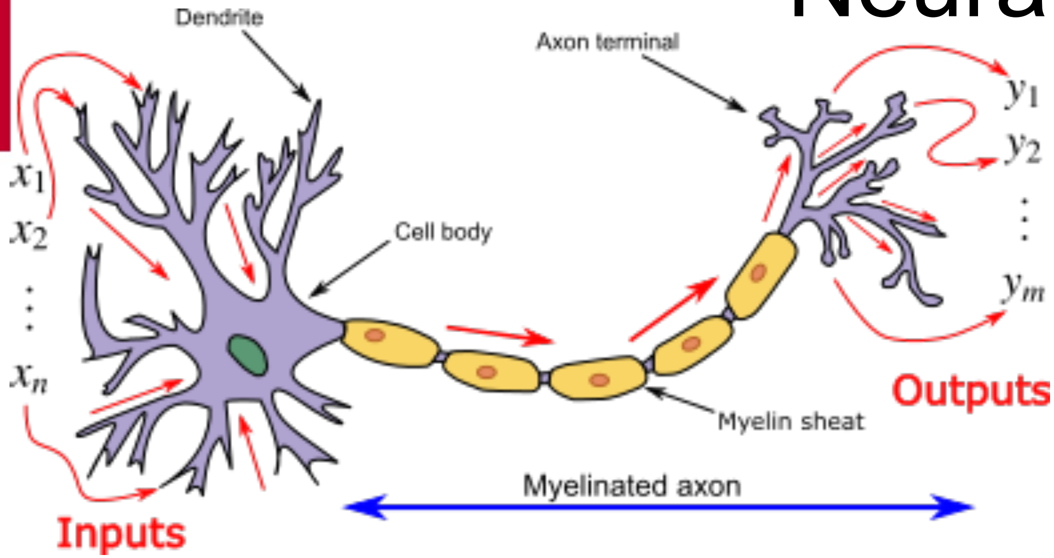
high epistemic uncertainty

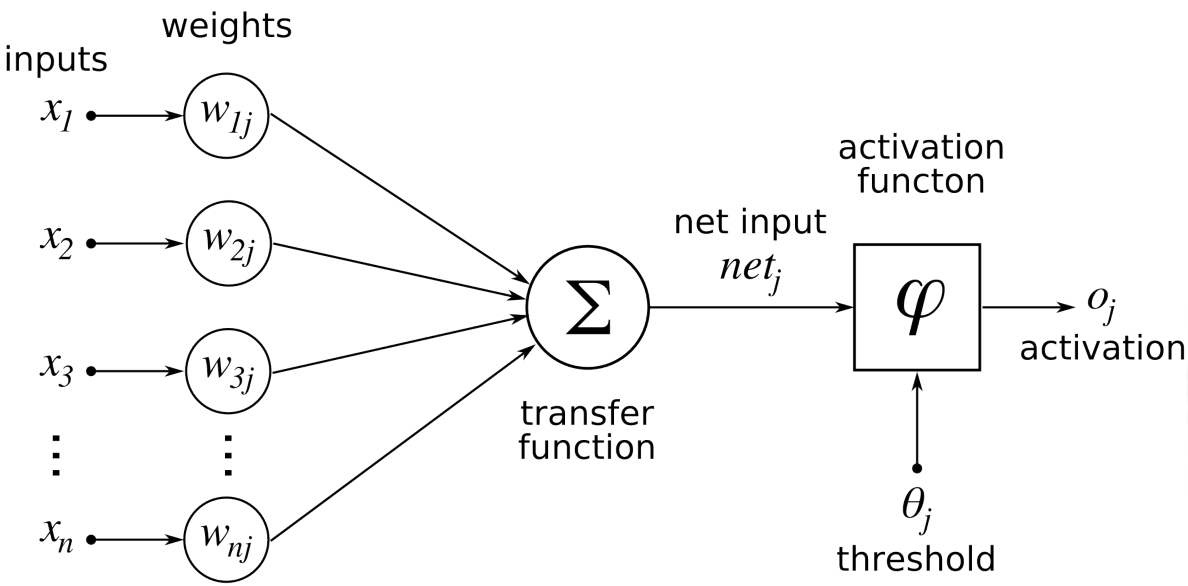
Uncertainty – applications

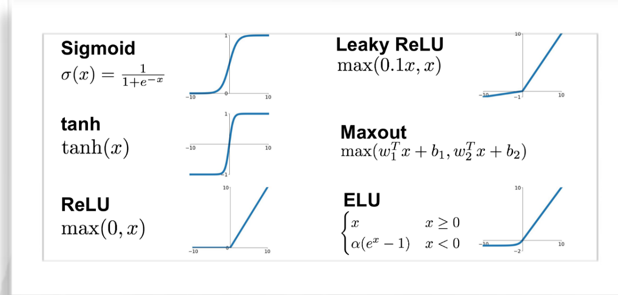


(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

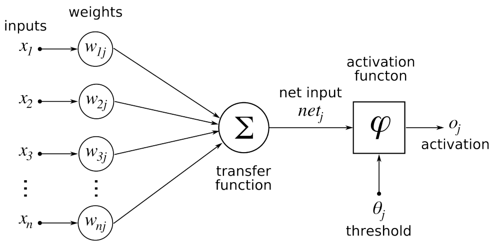
Neural nets - intro

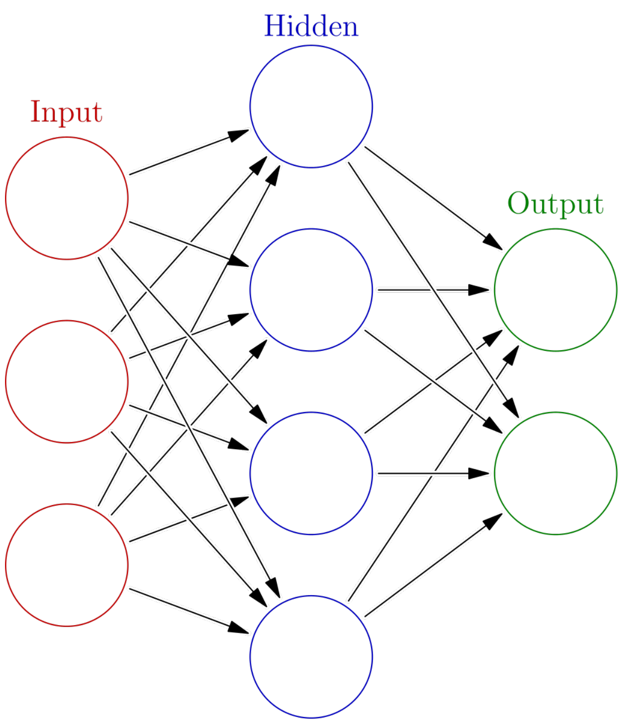


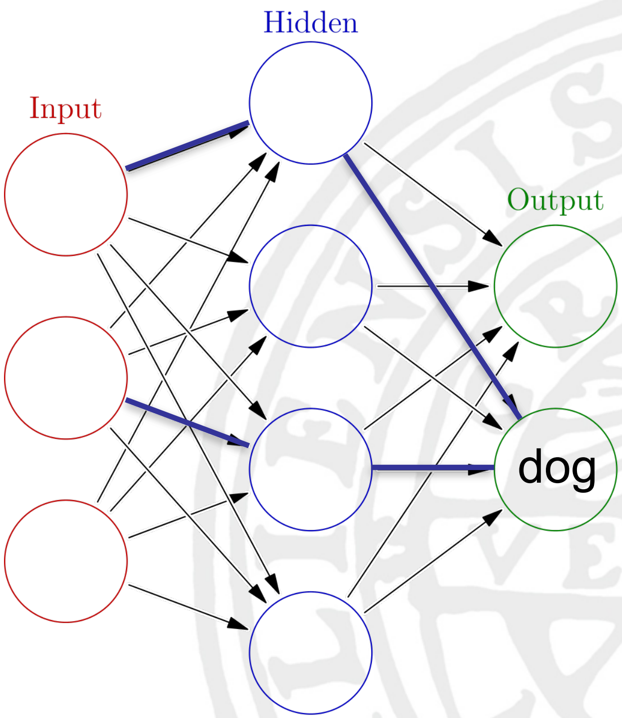


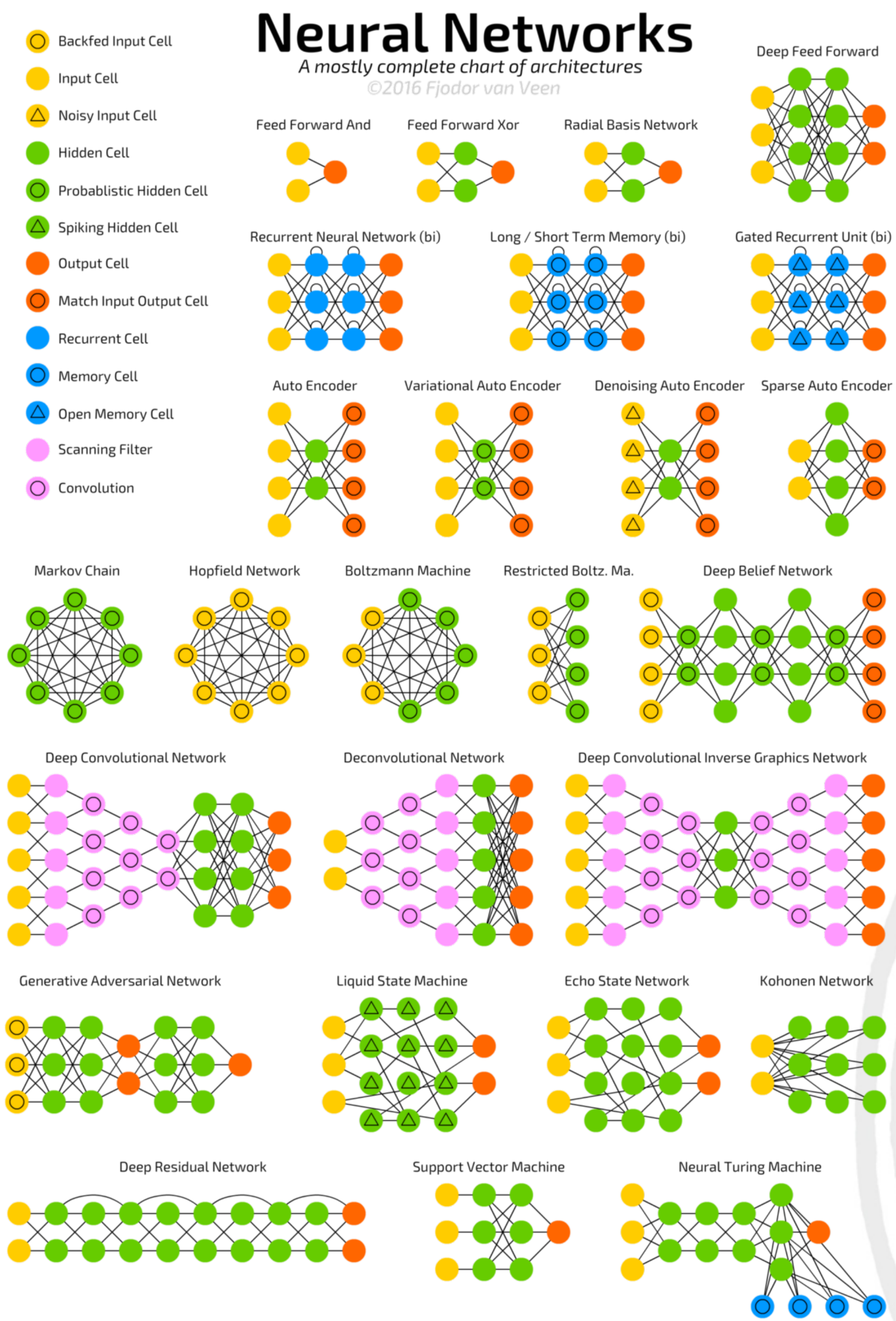


Neural nets - intro



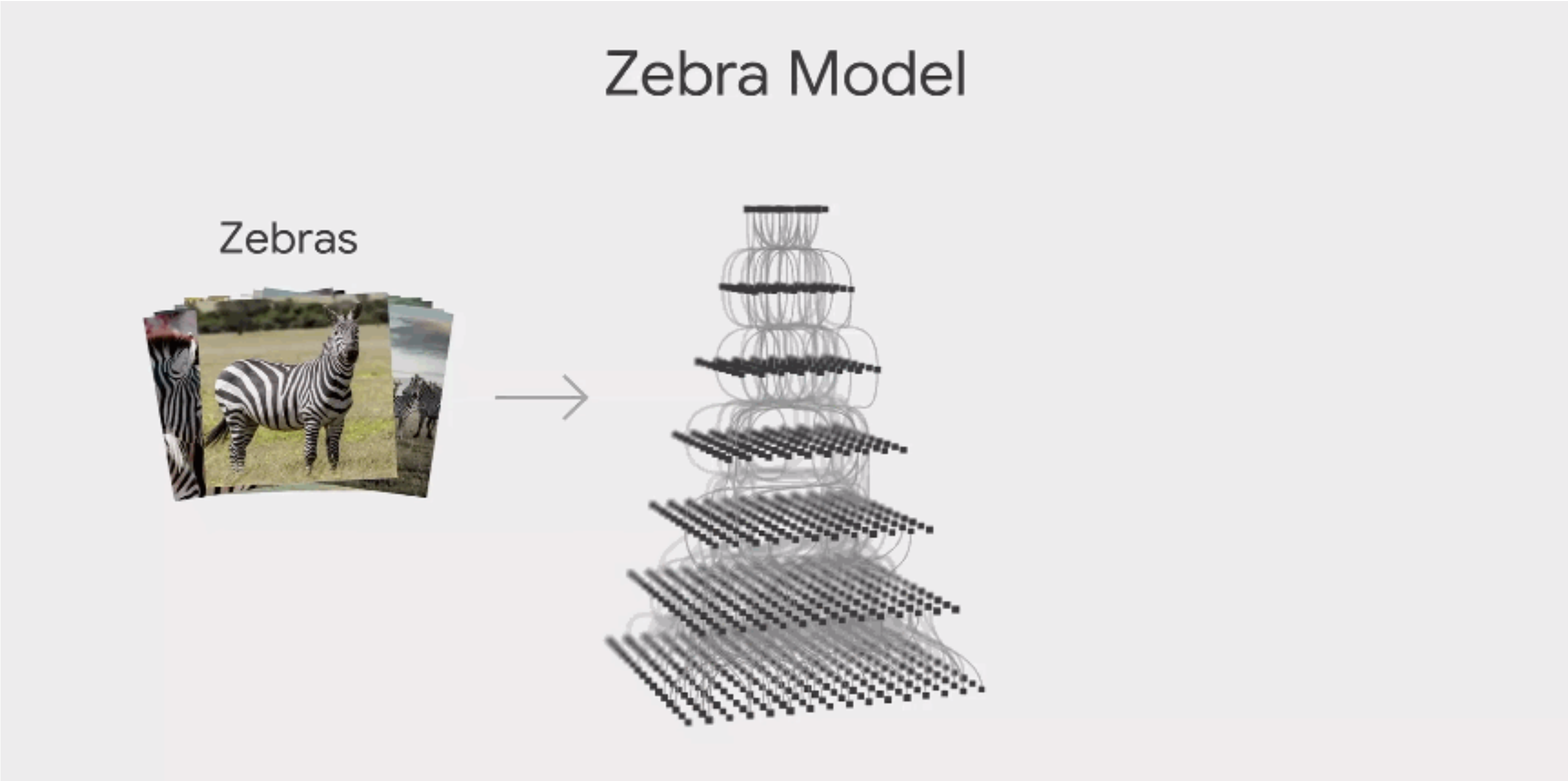




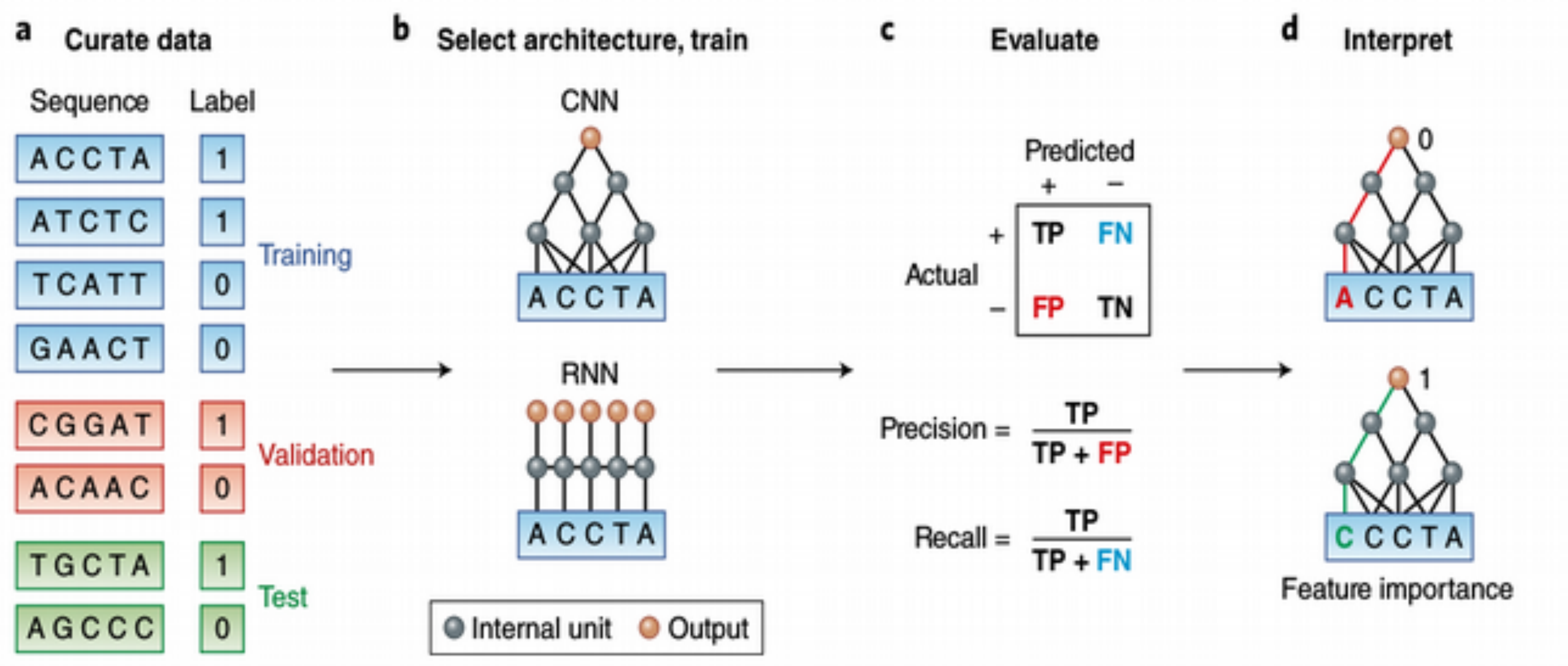


source: https://cdn-images-1.medium.com/max/1200/1*niisN1cdRP9AdWZ9lLkodQ.png

Concept Activation Vectors - TCAV



ANN in genetics



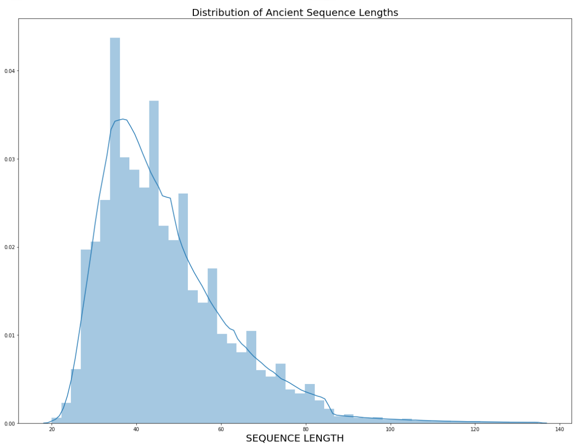
CNN for Genomics from Zou et al. Nature Genetics **51**, pages12–18 (2019)

UPPSALA UNIVERSITET

Nikolay Oskolkov, NBIS

INPUT: draft Neanderthal genome 2010

Modern French genome 76nt long reads

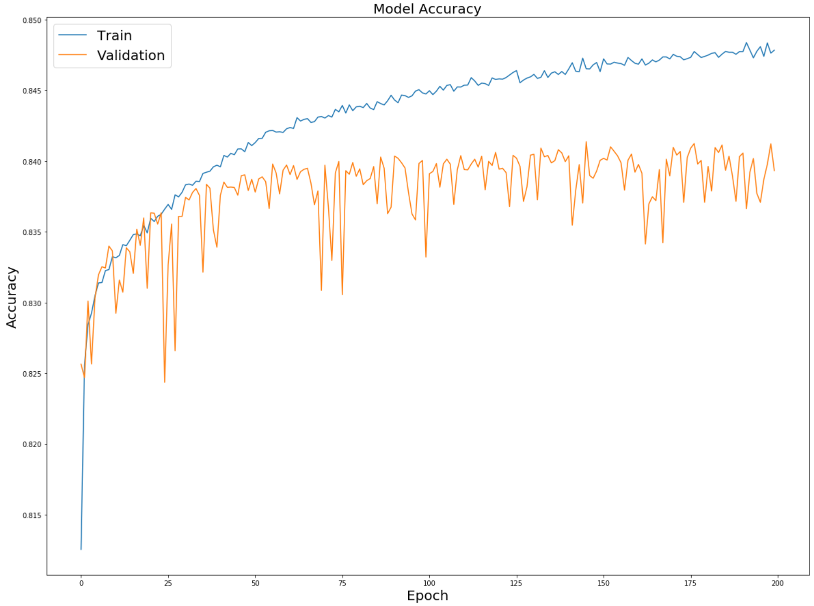


76nt

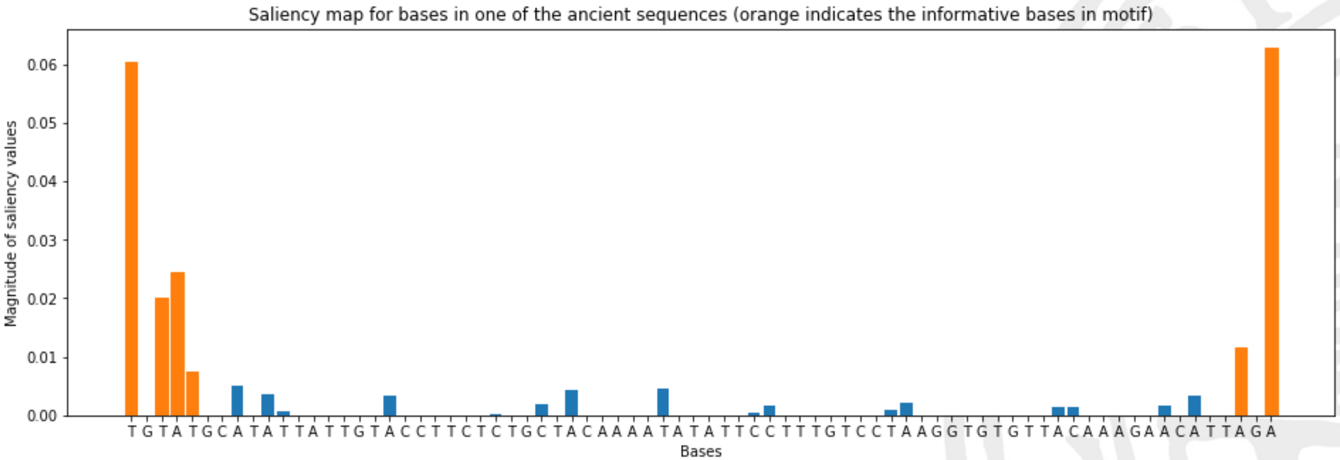
500 000 modern reads

500 000 ancient reads

Nikolay Oskolkov, NBIS

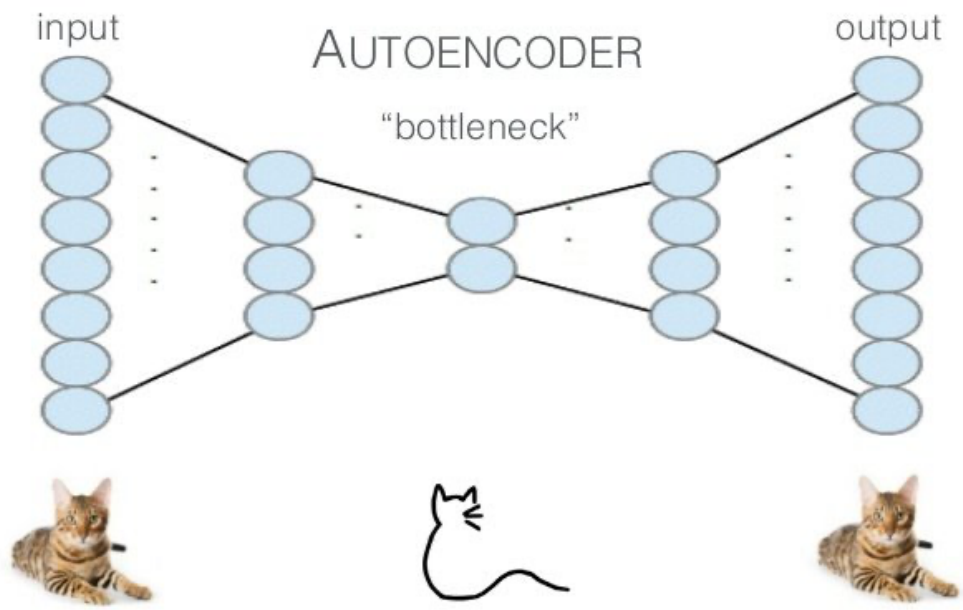


84% Acc



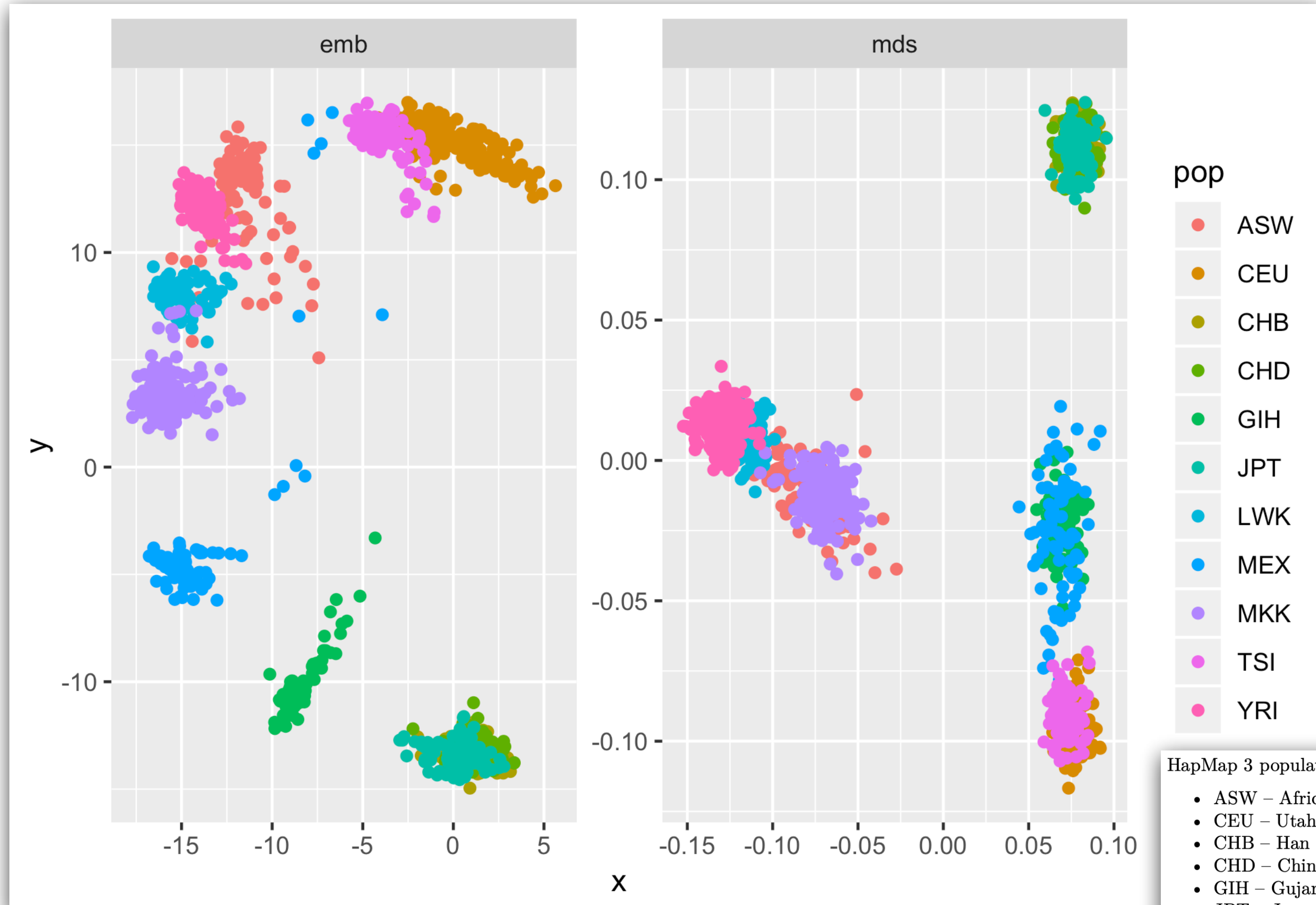
aDNA —> deamination at the ends —> increased C/T & G/A

Autoencoders



https://towardsdatascience.com/tcav-interpretability-beyond-feature-attribution-79b4d3610b4d

Autoencoder



HapMap 3 populations:

- ASW – African ancestry in Southwest USA
- CEU – Utah residents with Northern and Western European ancestry
- CHB – Han Chinese in Beijing, China
- CHD – Chinese in Metropolitan Denver, Colorado
- GIH – Gujarati Indians in Houston, Texas
- JPT – Japanese in Tokyo, Japan
- LWK – Luhya in Webuye, Kenya
- MEX – Mexican ancestry in Los Angeles, California
- MKK – Maasai in Kinyawa, Kenya
- TSI – Toscans in Italy
- YRI – Yoruba in Ibadan, Nigeria

UPPSALA UNIVERSITET

What do we want to achieve?

Genomic data → Model → patterns, denoising, predictions

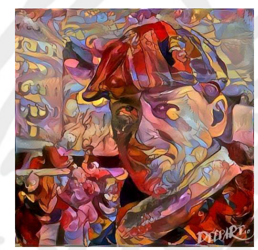
Genomic data → xAI model → insights, how? why?; new examples

Genomic data, Proteomic data → Model → integration

exploration

diagnostics

understanding



multi-omics

KNOWLEDGE

UPPSALA
UNIVERSITET

Thank You!