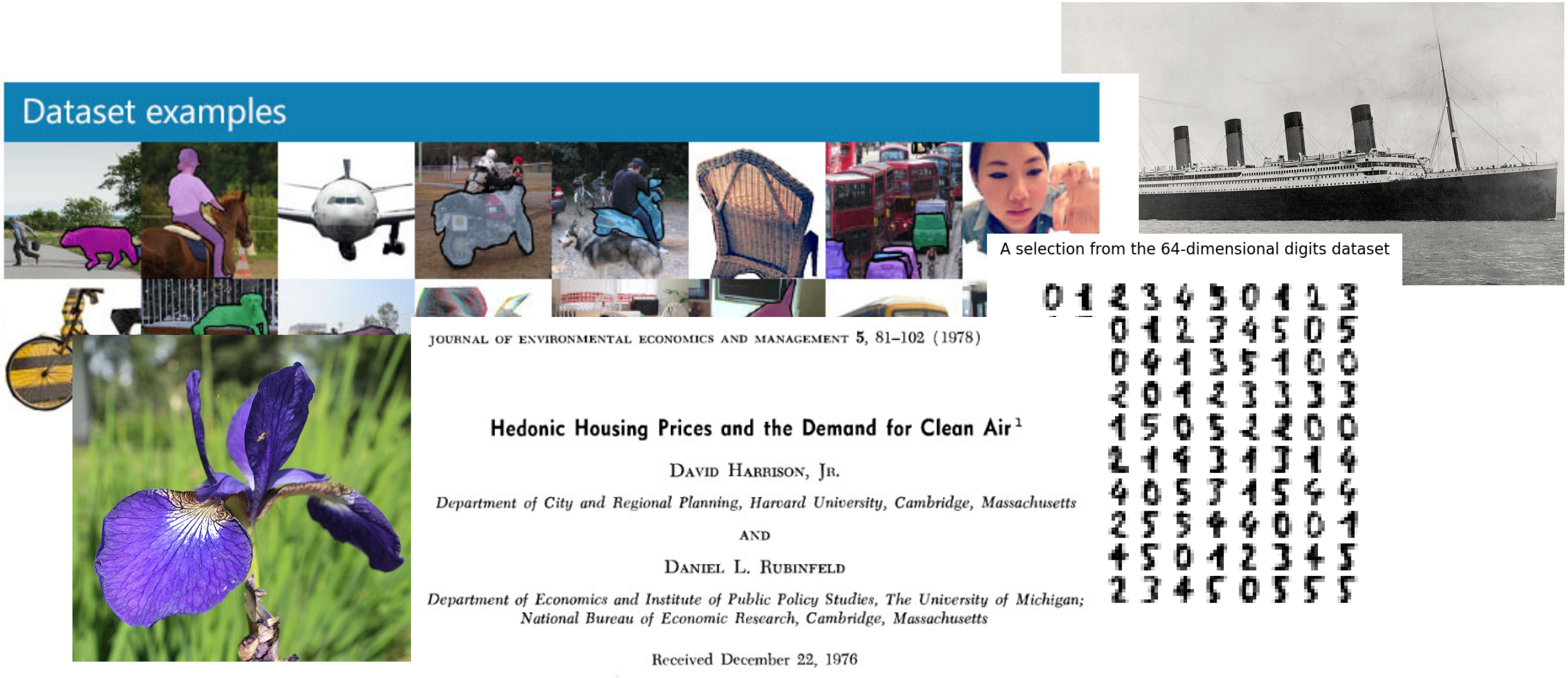
Using multiple data modalities for brain tumor diagnostics and treatment

Sören Lukassen
BIH Digital Health Center
02/02/2023

Berlin Institute of Health @Charité



Dataset examples

A selection from the 64-dimensional digits dataset

JOURNAL OF ENVIRONMENTAL ECONOMICS AND MANAGEMENT 5, 81–102 (1978)

Hedonic Housing Prices and the Demand for Clean Air[1]

DAVID HARRISON, JR.

Department of City and Regional Planning, Harvard University, Cambridge, Massachusetts

AND

DANIEL L. RUBINFELD

Department of Economics and Institute of Public Policy Studies, The University of Michigan; National Bureau of Economic Research, Cambridge, Massachusetts

Received December 22, 1976

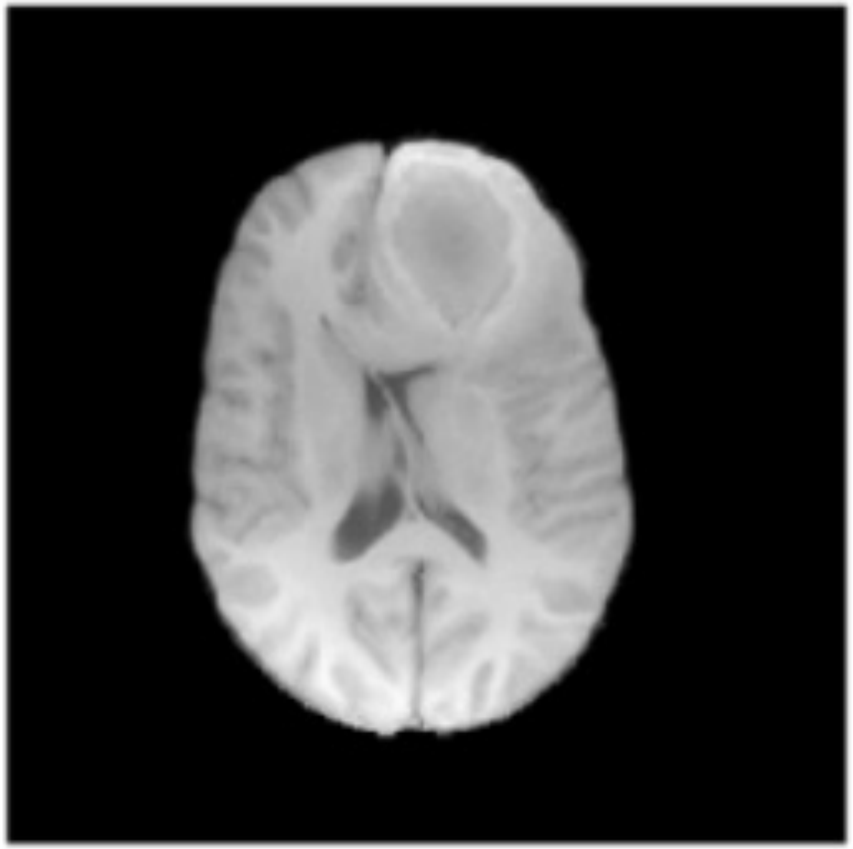
© scikit-learn developers
© COCO Consortium
© Tiia Monto, CC4-SA

MEDIZIN INFORMATIK INITIATIVE

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

BIH Berlin Institute of Health @Charité

Our application: brain tumor diagnostics and treatment



Jabareen & Lukassen, 2022

MEDIZIN INFORMATIK INITIATIVE

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

BIH Berlin Institute of Health @Charité

Tumor diagnostics – common workflow

Step 0: Indication for diagnostics (screening, symptoms, etc.)

Step 1: Imaging

Step 2: Biopsy

Step 3: Surgery + immediate frozen section

Step 4: Histopathology

MEDIZIN INFORMATIK INITIATIVE

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

BIH Berlin Institute of Health @Charité

Tumor diagnostics – common workflow

Step 0: Indication for diagnostics (screening, symptoms, etc.)

Unilateral loss of vision

MEDIZIN INFORMATIK INITIATIVE

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

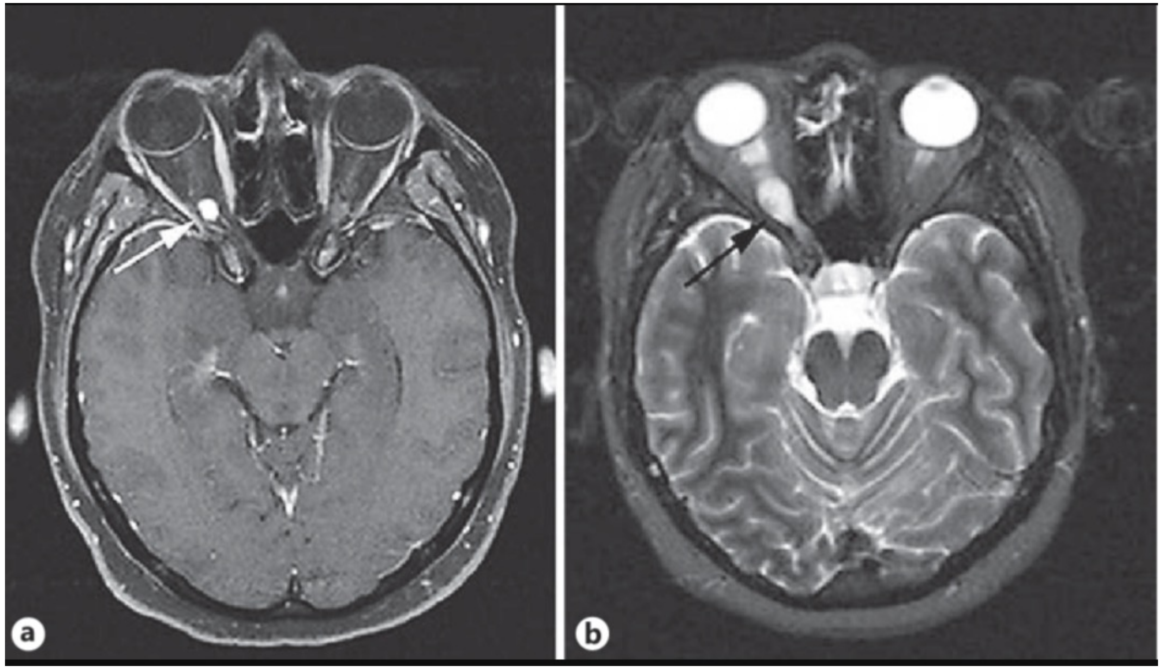
BIH Berlin Institute of Health @Charité

Tumor diagnostics – common workflow

Step 0: Indication for diagnostics (screening, symptoms, etc.)

Unilateral loss of vision

Step 1: Imaging



McGrath et al., 2018

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

MEDIZIN INFORMATIK INITIATIVE

BIH Berlin Institute of Health @Charité

Tumor diagnostics – common workflow

Step 0: Indication for diagnostics (screening, symptoms, etc.)

Unilateral loss of vision

Step 1: Imaging

Lesion at the optic nerve

Step 2: Biopsy

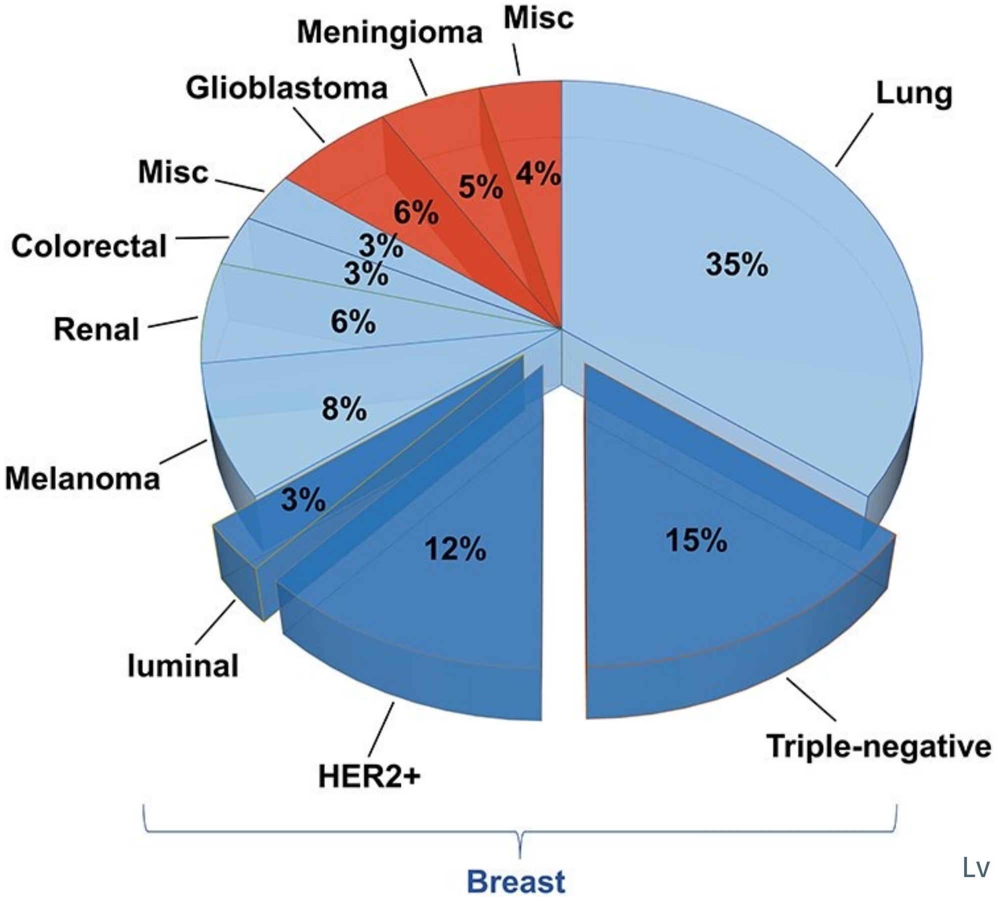
Difficult to reach location, risk of permanent damage to optic nerve

MEDIZIN INFORMATIK INITIATIVE

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

BIH Berlin Institute of Health @Charité

Problem 1: we don‘t know what we‘re looking at



Lv et al., 2021

© Patrick J. Lynch, CC2.5

Problem 2: no fine-needle biopsies

© Patrick J. Lynch, CC2.5

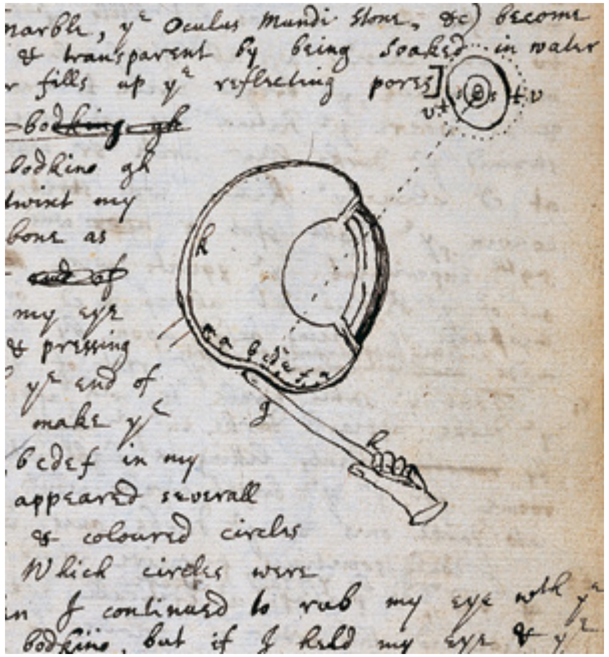
MEDIZIN INFORMATIK INITIATIVE

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

BIH Berlin Institute of Health @Charité

Problem 2: no fine-needle biopsies



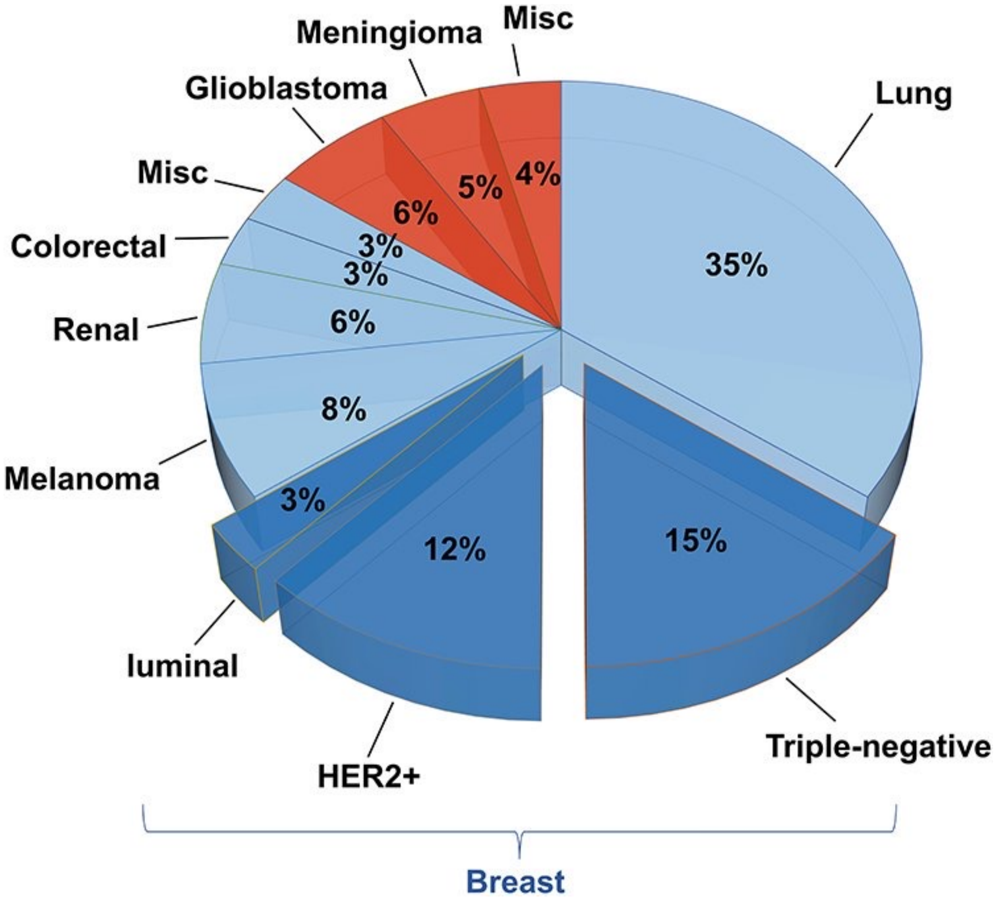
marble, y^e Oculus Mundi Stone, &c) become
& transparent by being soaked in water
r fills up y^e reflecting pores]
~~bodking gh~~
bodkine gh
twixt my
bone as
~~and~~ of
my eye
& pressing
y^e end of
make y^e
bcdef in my
appeared severall
& coloured circles
Which circles were
n I continued to rub my eye with y^e
bodkine, but if I held my eye & y^e

v t
t v
k
a b c d e f
g
h

Isaac Newton

© Patrick J. Lynch, CC2.5

MEDIZIN INFORMATIK INITIATIVE
GEFÖRDERT VOM Bundesministerium für Bildung und Forschung
BIH Berlin Institute of Health @Charité



Lv et al., 2021

© Patrick J. Lynch, CC2.5

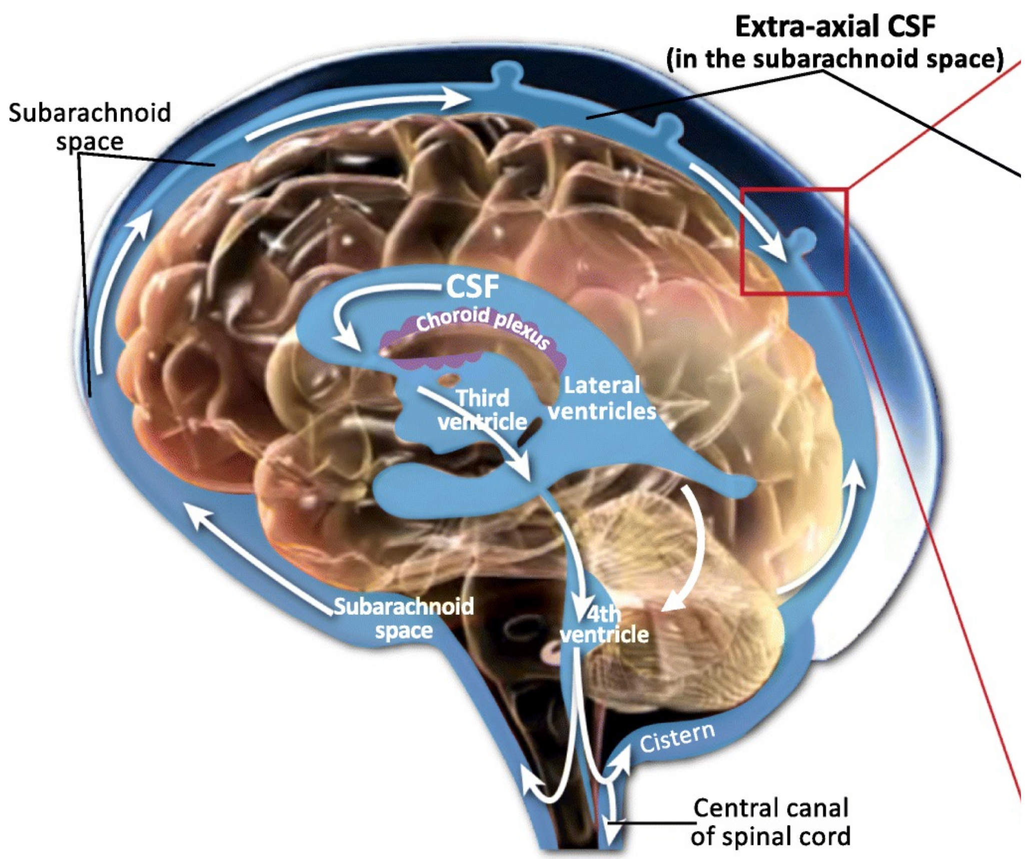
GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

MEDIZIN INFORMATIK INITIATIVE

BIH Berlin Institute of Health @Charité

We can‘t reach the tumor directly, but...



© Mark D. Shen, CC4, cropped

MEDIZIN INFORMATIK INITIATIVE
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
BIH Berlin Institute of Health @Charité

What can we do with CSF?

1. Cytology (stain, identify & count cells)
2. Proteomics
3. Metabolomics
4. Cell-free DNA analysis

MEDIZIN INFORMATIK INITIATIVE

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

BIH Berlin Institute of Health @Charité

What can we do with CSF?

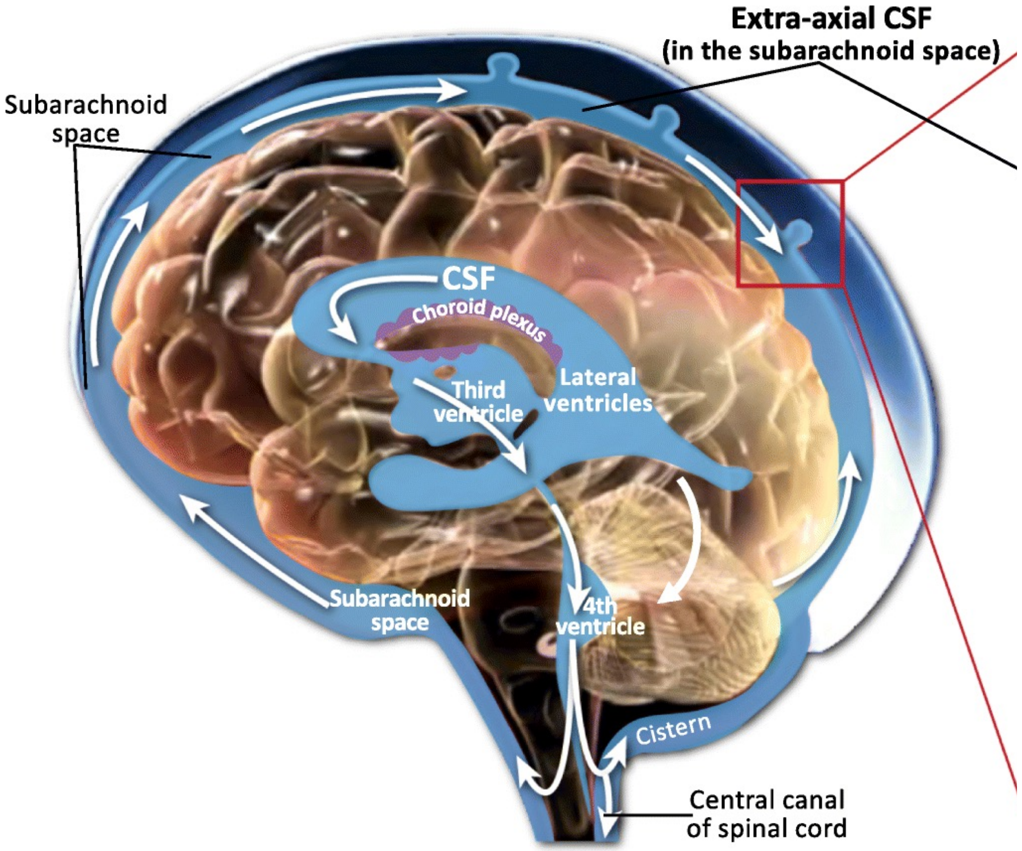
1. Cytology (stain, identify & count cells)
2. Proteomics
3. Metabolomics
4. Cell-free DNA analysis

MEDIZIN INFORMATIK INITIATIVE

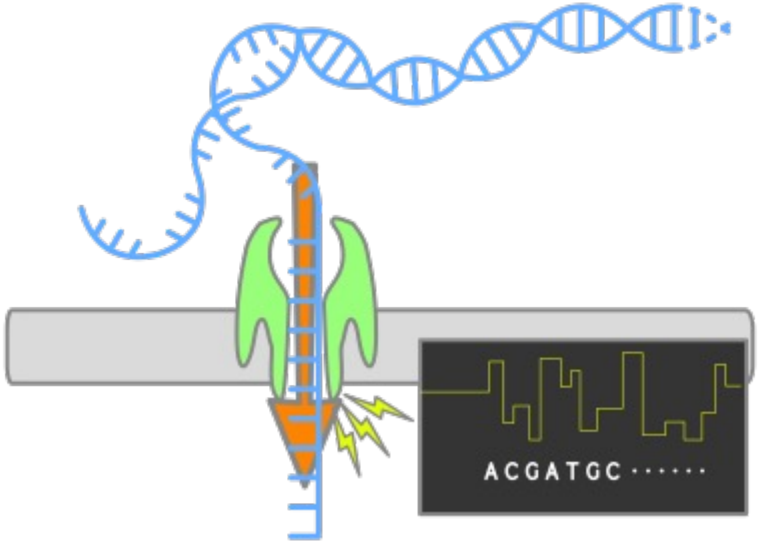
GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

BIH Berlin Institute of Health @Charité

We can‘t reach the tumor directly, but...



© Mark D. Shen, CC4, cropped



© DBCLS, CC4

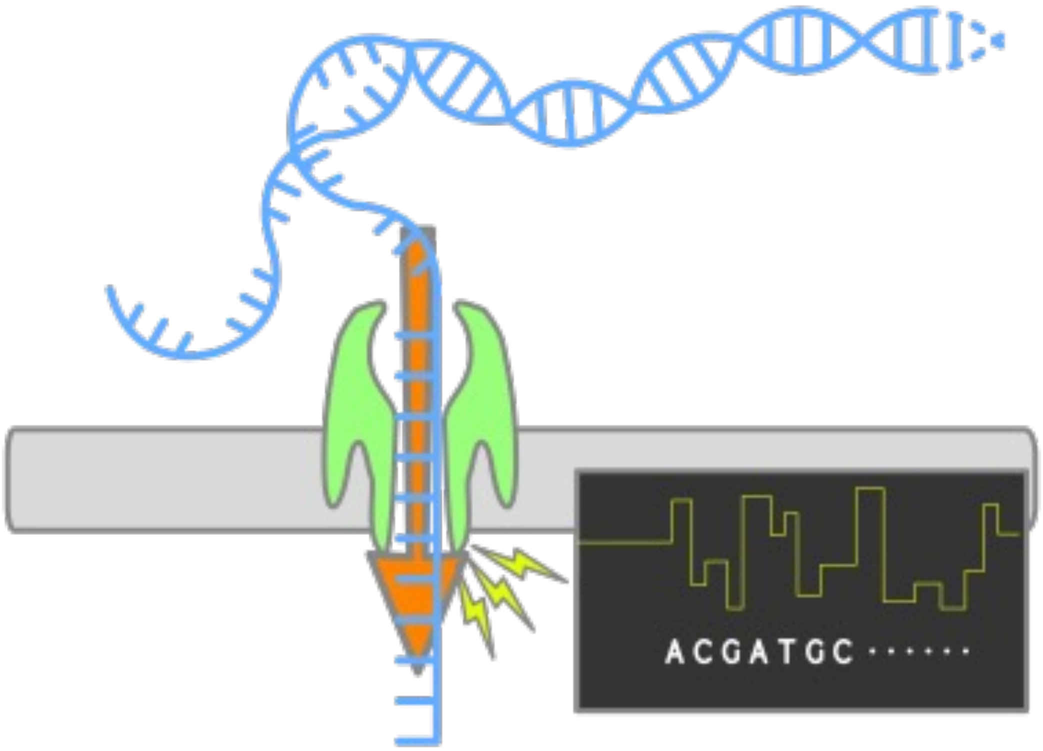
MEDIZIN INFORMATIK INITIATIVE

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

BIH Berlin Institute of Health @Charité

Nanopore sequencing

- Very fast (initial results within minutes)
- Read length can be hundreds of kb
- Produces reads sequentially
- Not very accurate
- Can detect DNA modifications



MEDIZIN INFORMATIK INITIATIVE
GEFÖRDERT VOM Bundesministerium für Bildung und Forschung
BIH Berlin Institute of Health @Charité

Nanopore sequencing

- Very fast (initial results within minutes)
- Read length can be hundreds of kb
- Produces reads sequentially
- Not very accurate
- Can detect DNA modifications

Good:

- We can process while we sequence
- We can stop sequencing once we have enough data
- Read length means we can distinguish cellular/cell-free DNA

Bad:

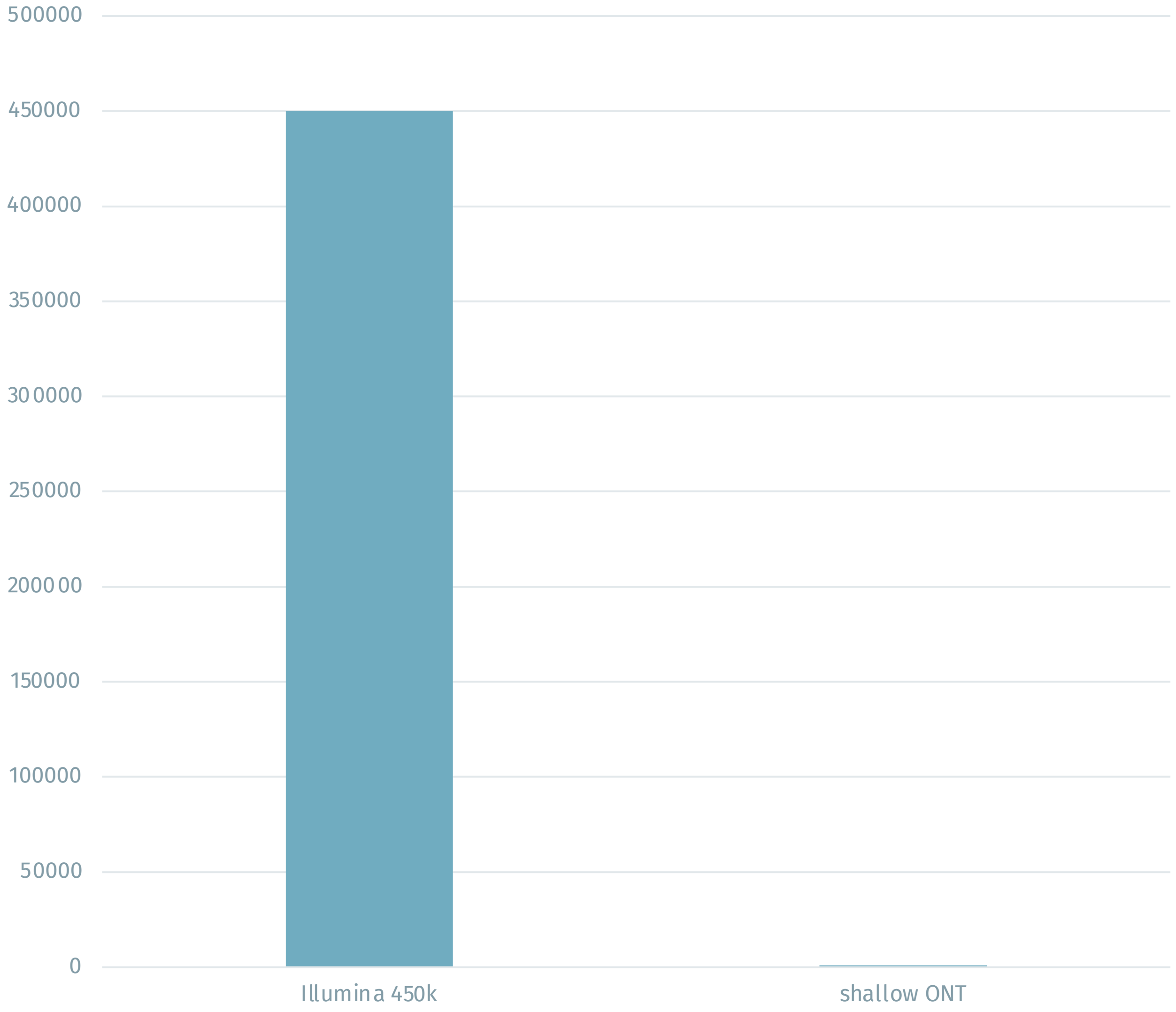
- Few training data (but we can use microarrays)
- Very shallow coverage
- Mutation calling is hard

MEDIZIN INFORMATIK INITIATIVE

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

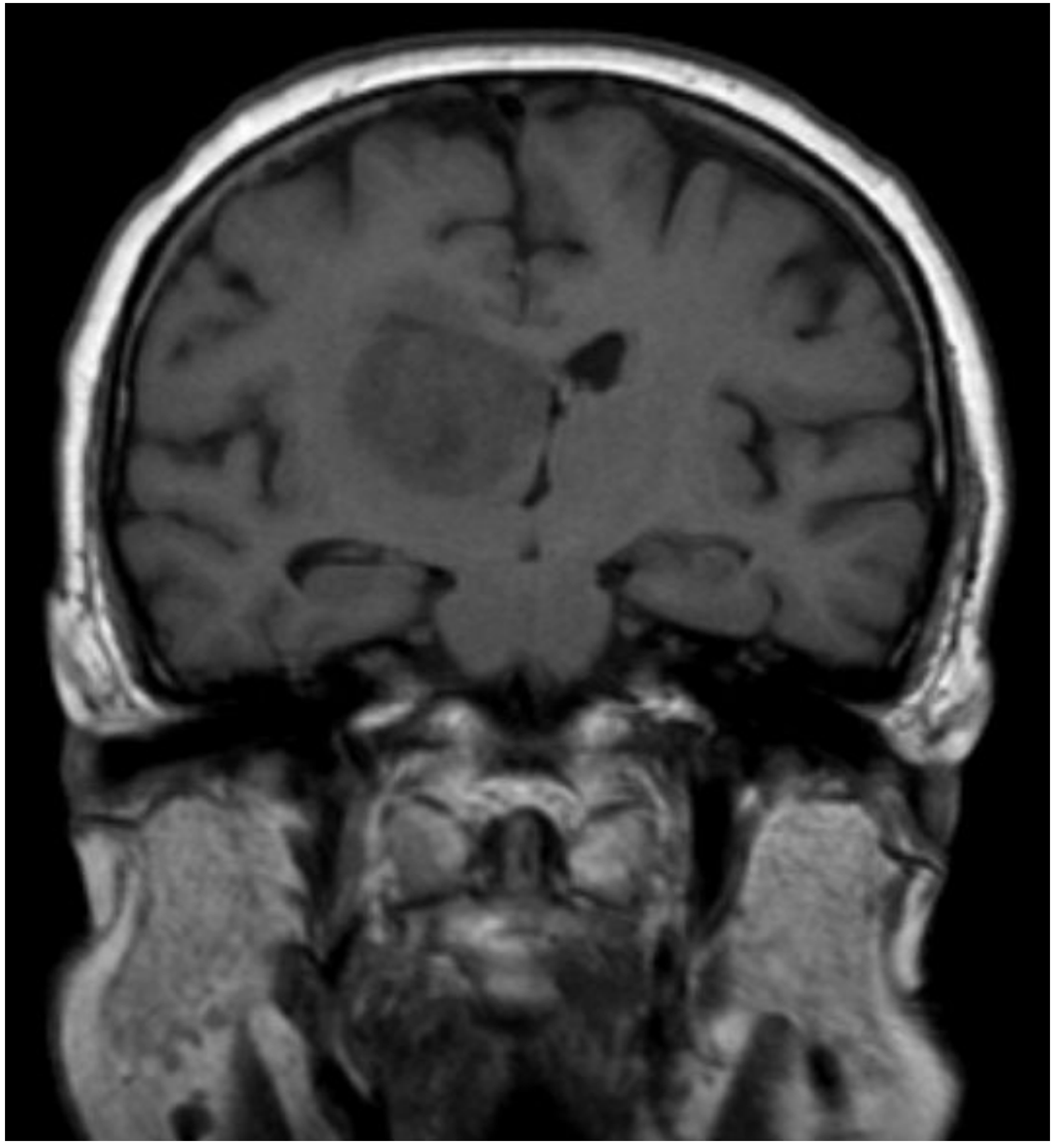
BIH Berlin Institute of Health @Charité

So what‘s shallow?



- Of the 450,000 sites in our microarray training data, as little as 1,000 are covered

So what‘s shallow?



© Hellerhoff, CC3-SA

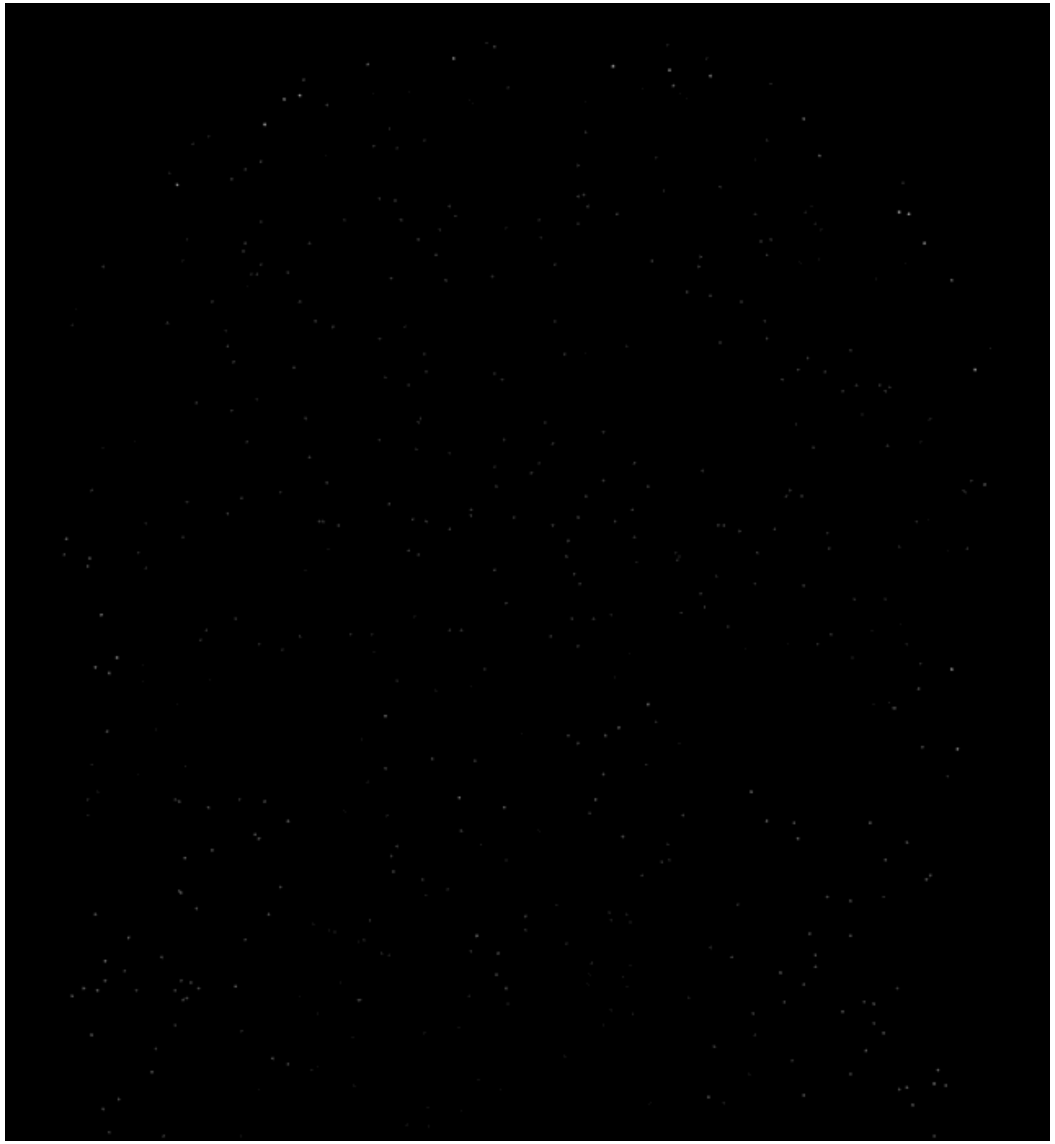
- Of the 450,000 sites in our microarray training data, as little as 1,000 are covered

MEDIZIN INFORMATIK INITIATIVE

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

BIH Berlin Institute of Health @Charité

So what‘s shallow?



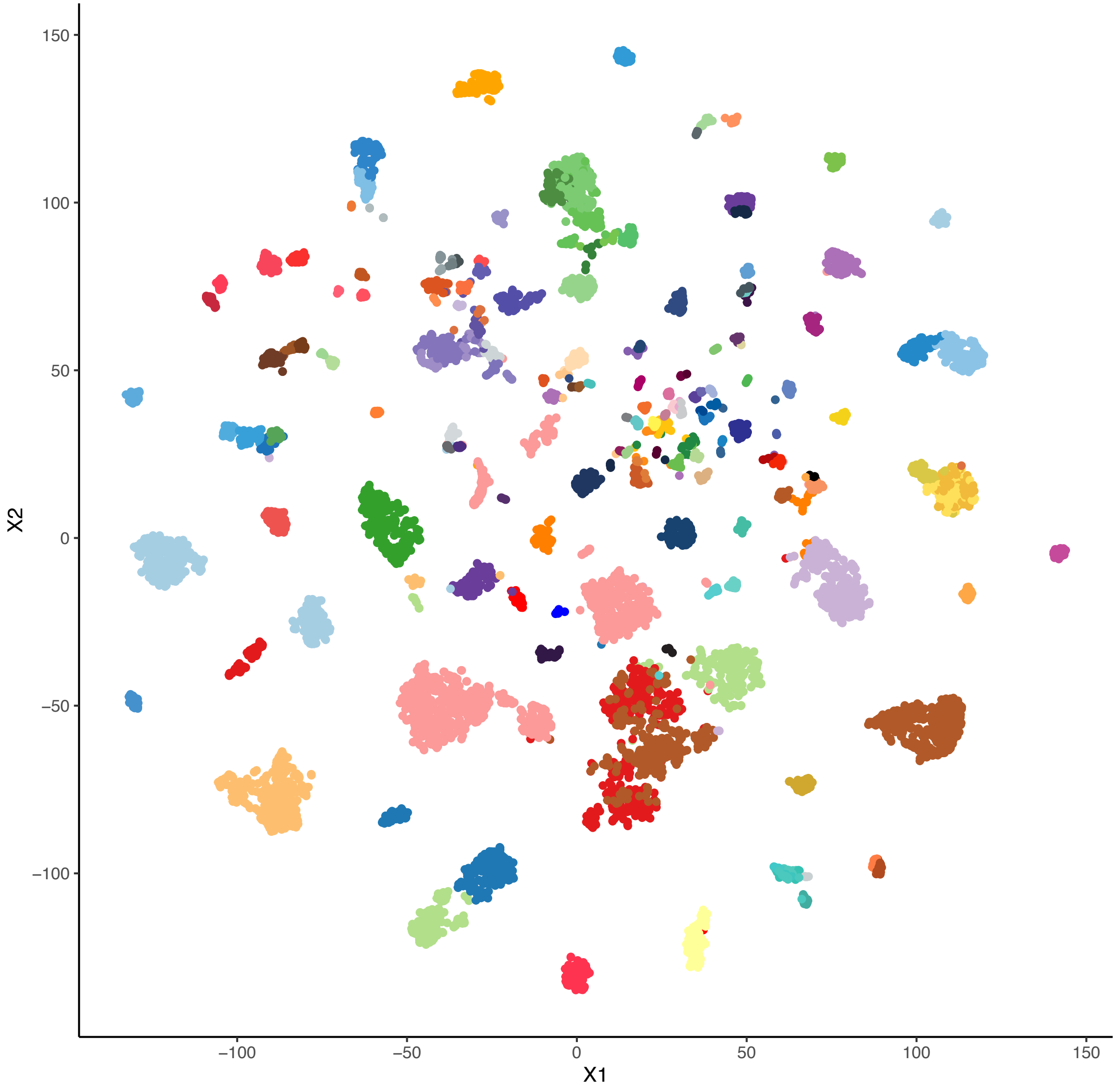
- Of the 450,000 sites in our microarray training data, as little as 1,000 are covered

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

Main problems with our project

t−SNE, perplexity = 30



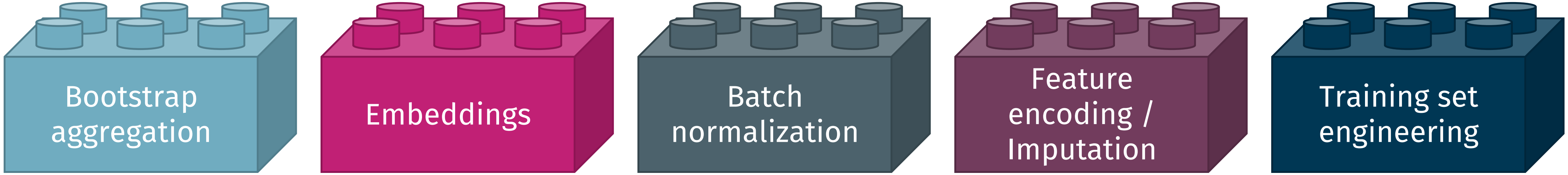
- Predictor missingness
- Relatively few training samples (~8000)
- n << p
- Many classes (~180)
- Severe class imbalance

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

MEDIZIN INFORMATIK INITIATIVE

BIH Berlin Institute of Health @Charité

Building blocks for n << p; predictor missingness



Bootstrap aggregation

To prevent overfitting

Embeddings

Reduces number of predictors

Ideally helps to minimize differences between predictor sets

Batch normalization

Keeps activations in deep neural networks constant when input activation varies

Feature encoding / Imputation

What to do with missing values? Mean? 0.5?

Could make predictor missingness informative

Training set engineering

Reduction of training set features to observed features

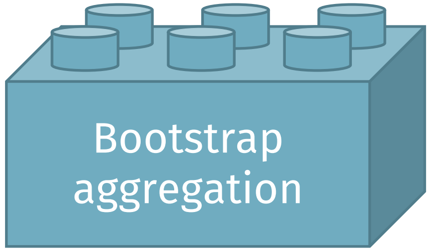
Needs repeated training

MEDIZIN INFORMATIK INITIATIVE

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

BIH Berlin Institute of Health @Charité

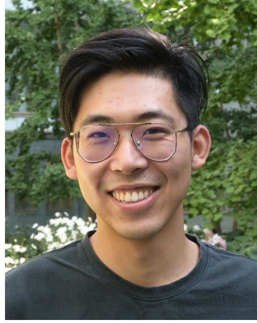
Our journey through algorithms, part I: baseline



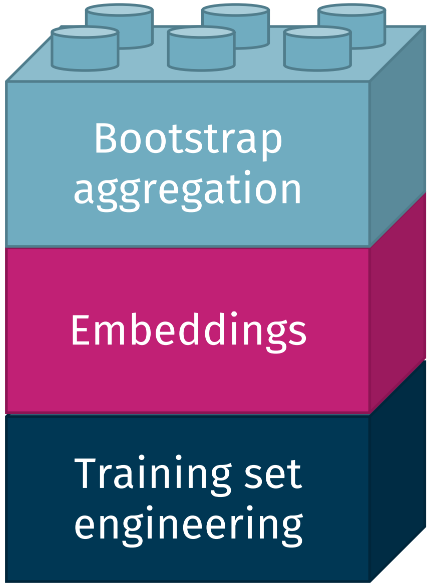
- Random forest: ~ 85% accuracy on tissue samples

MEDIZIN INFORMATIK INITIATIVE
GEFÖRDERT VOM Bundesministerium für Bildung und Forschung
BIH Berlin Institute of Health @Charité

Our journey through algorithms, part II: multi-step



Dongsheng Yuan



- Multi-step model
- Only uses CpGs from the sample to be classified
- Joint embedding of sample and training set; selection of most similar classes

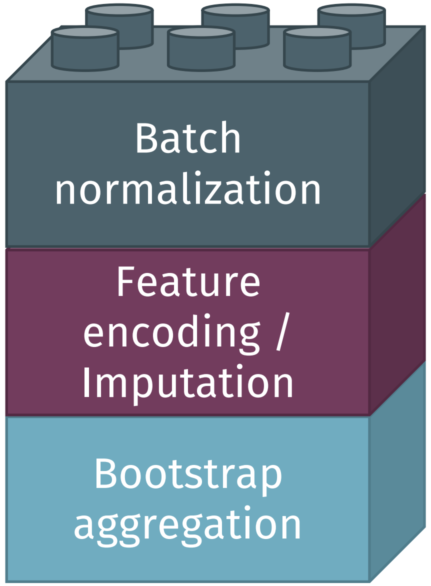
  → retraining for every sample
- Second step: random forest classifier

➔ Accuracy on tissue samples: > 90% (Top1)

➔ with < 1000 CpGs: > 60%

MEDIZIN INFORMATIK INITIATIVE
GEFÖRDERT VOM Bundesministerium für Bildung und Forschung
BIH Berlin Institute of Health @Charité

Our journey through algorithms, part III: train once



- Multilayer Perceptron
- Full training set with random feature selection each epoch
- Data encoded as: methylated(1)/unmethylated(-1)/missing(0)

MEDIZIN INFORMATIK INITIATIVE

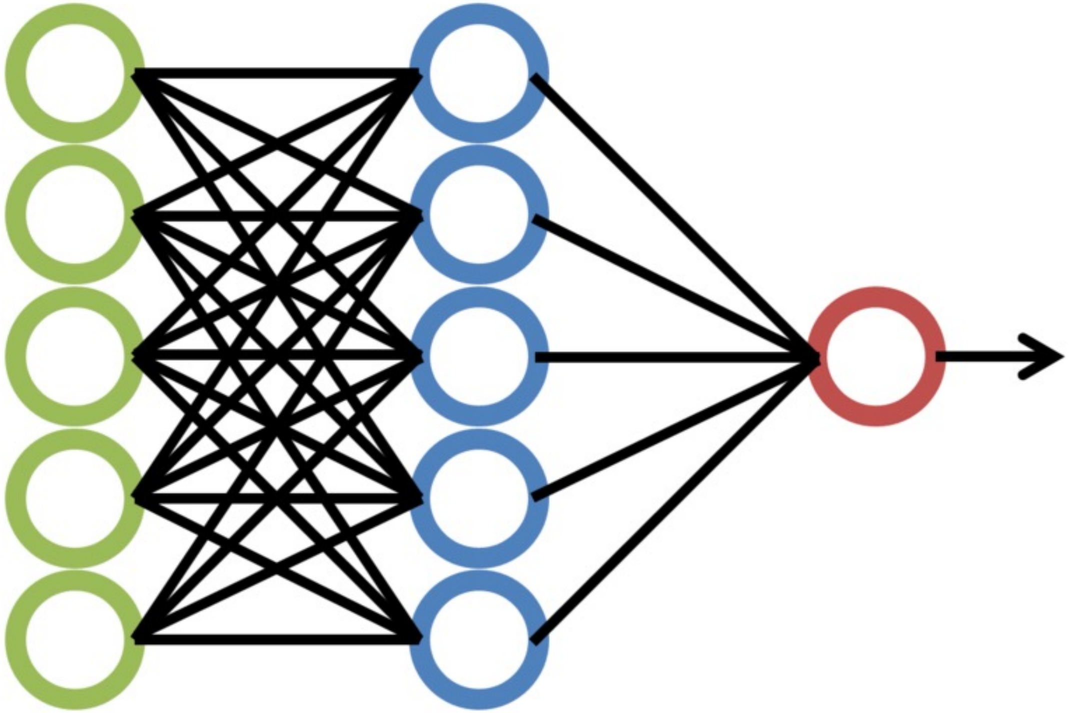
GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

BIH Berlin Institute of Health @Charité

Short reminder: neural networks and encoding

© Marco Verch, CC2

$$y = mx + n$$



MEDIZIN INFORMATIK INITIATIVE

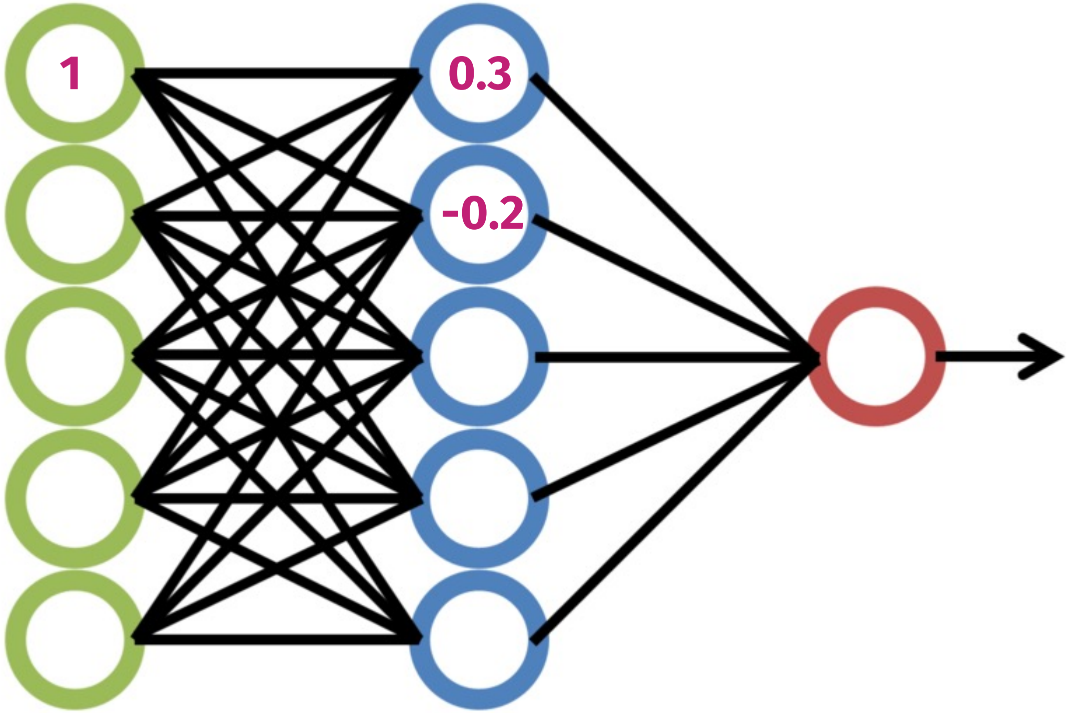
GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

BIH Berlin Institute of Health @Charité

Short reminder: neural networks and encoding

© Marco Verch, CC2

$$y = m * 1 + 0 = m$$



GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

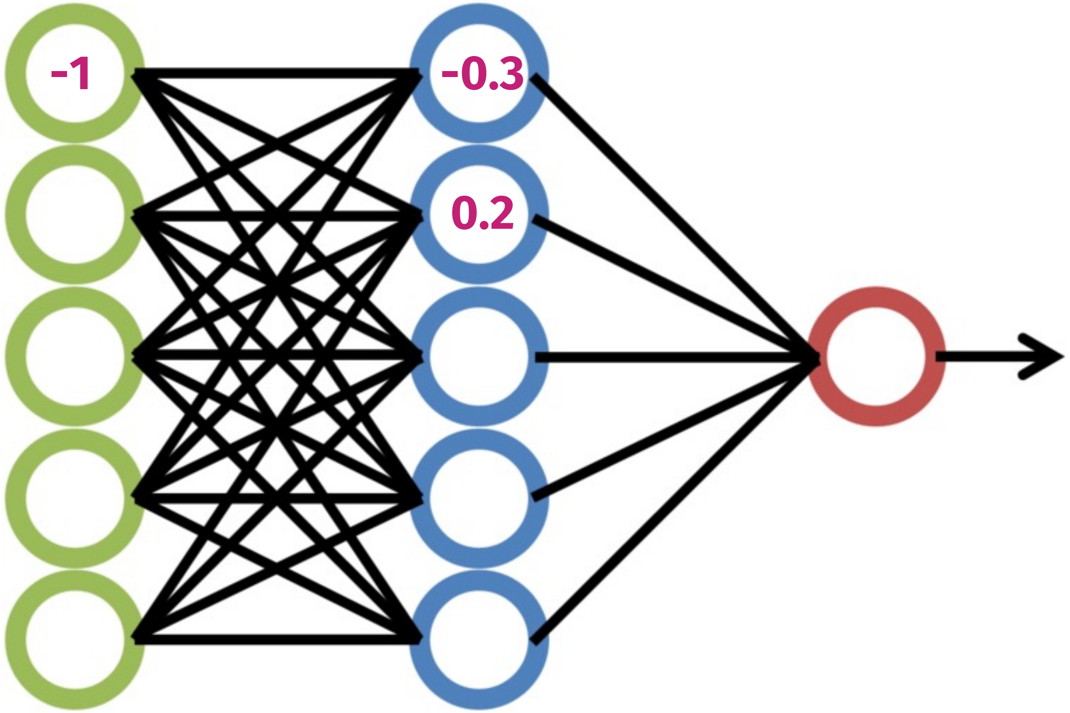
MEDIZIN INFORMATIK INITIATIVE

BIH Berlin Institute of Health @Charité

Short reminder: neural networks and encoding

© Marco Verch, CC2

$$y = m * -1 + 0 = -m$$

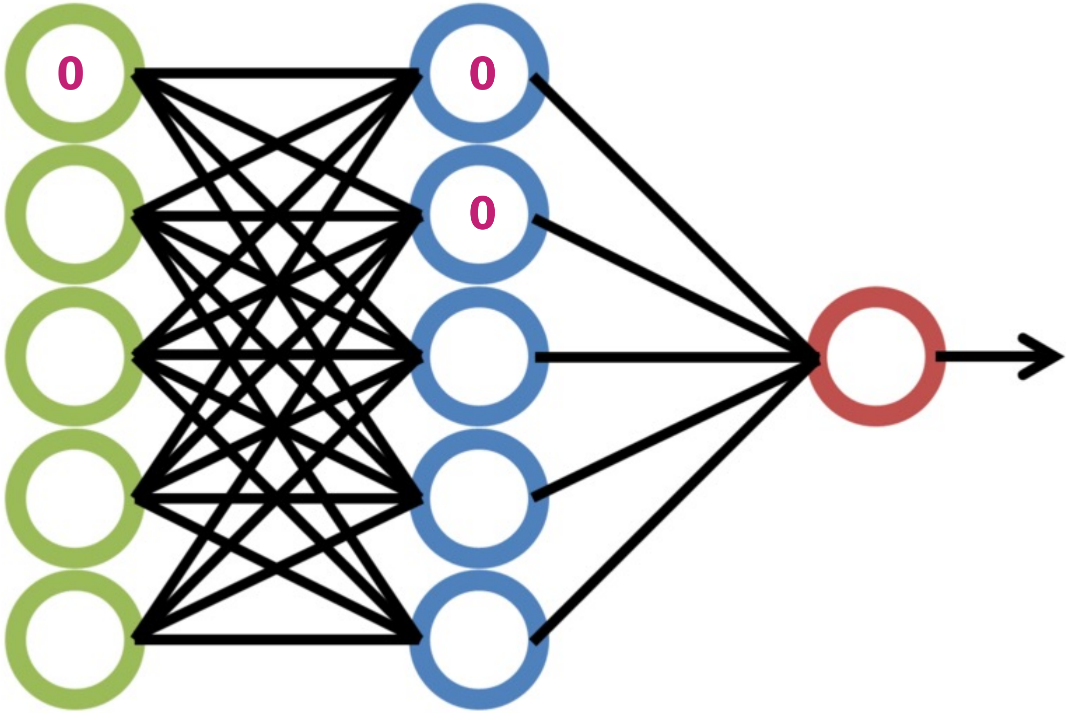


MEDIZIN INFORMATIK INITIATIVE
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
BIH Berlin Institute of Health @Charité

Short reminder: neural networks and encoding

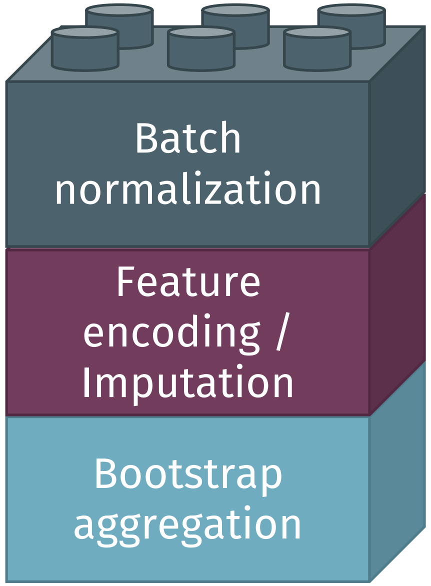
© Marco Verch, CC2

$$y = m * 0 + 0 = 0$$



MEDIZIN INFORMATIK INITIATIVE
GEFÖRDERT VOM Bundesministerium für Bildung und Forschung
BIH Berlin Institute of Health @Charité

Our journey through algorithms, part III: train once

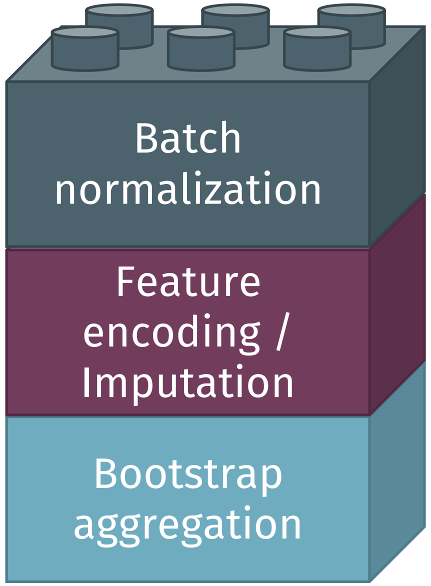


- Multilayer Perceptron
- Full training set with random feature selection each epoch
- Data encoded as: methylated(1)/unmethylated(-1)/missing(0)
- Batch normalization to harmonize weights with different predictor numbers

➔ Accuracy in tissue with >99% missing CpGs: > 80% (Top1)

MEDIZIN INFORMATIK INITIATIVE
GEFÖRDERT VOM Bundesministerium für Bildung und Forschung
BIH Berlin Institute of Health @Charité

Our journey through algorithms, part IV: the KISS principle

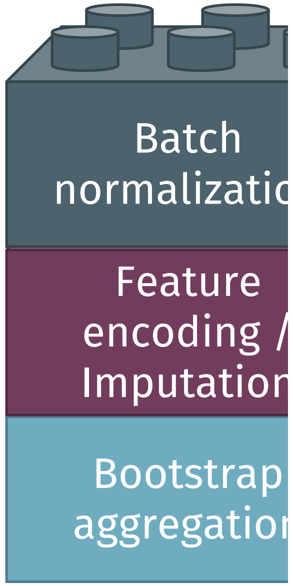


- Once initial results are in, hyperparameters should be tuned
- In our case, the network kept suggesting fewer layers
- And fewer layers...
- And fewer layers...

➔ Our newest model: linear regression...

MEDIZIN INFORMATIK INITIATIVE
GEFÖRDERT VOM Bundesministerium für Bildung und Forschung
BIH Berlin Institute of Health @Charité

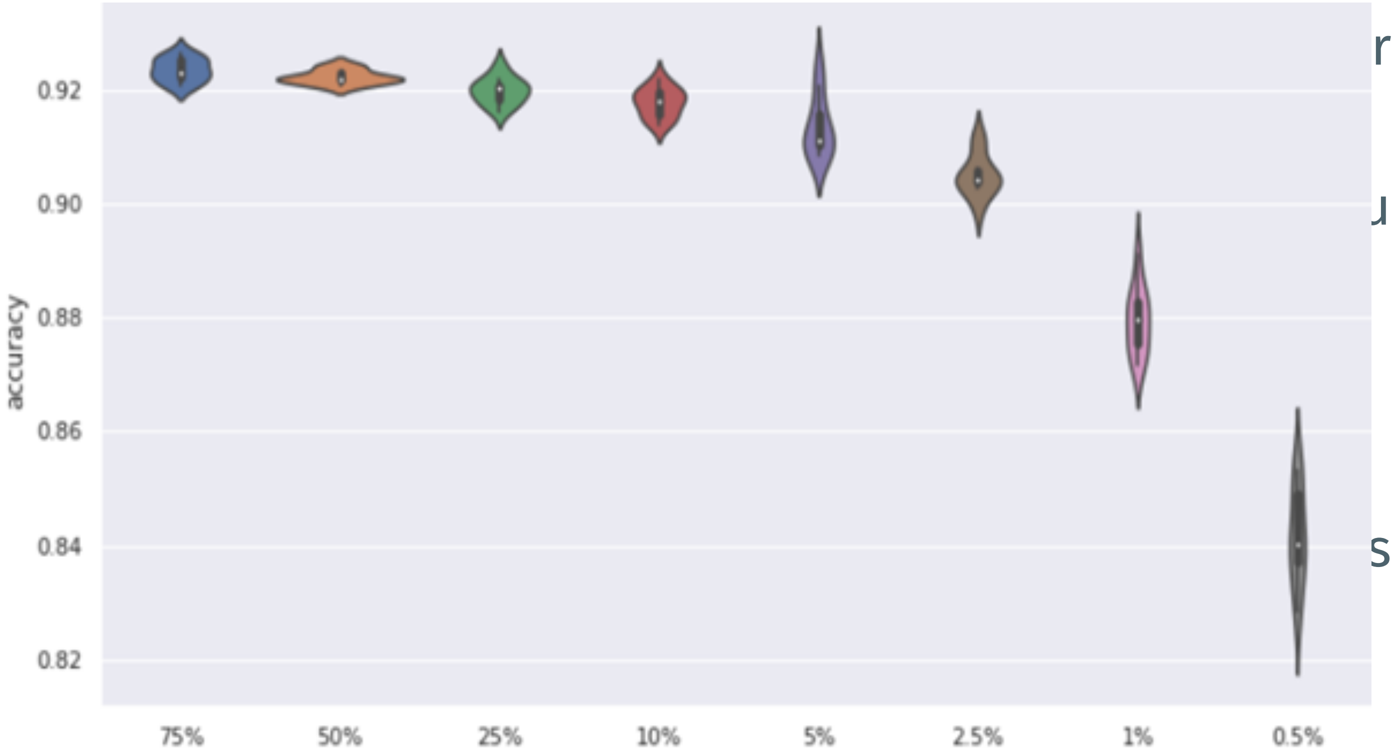
Our journey through algorithms, part IV: the KISS principle



Batch normalizatio

Feature encoding / Imputation

Bootstrap aggregatio



rparameters

uggesting fewer

ssion...

MEDIZIN INFORMATIK INITIATIVE

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

BIH Berlin Institute of Health @Charité

What went wrong?

1. When we took over the project from collaborators, they were already using RFs
2. We tried to improve the solution, rather than working from the ground up
3. Either way, we probably wouldn‘t have tried linear regression – the problem looked too complicated

Beware the deep learning trap: If there‘s an easy solution, complex machine learning models will often give you reasonably good results. Starting with complex models can leave you stuck with overcomplicated pipelines.

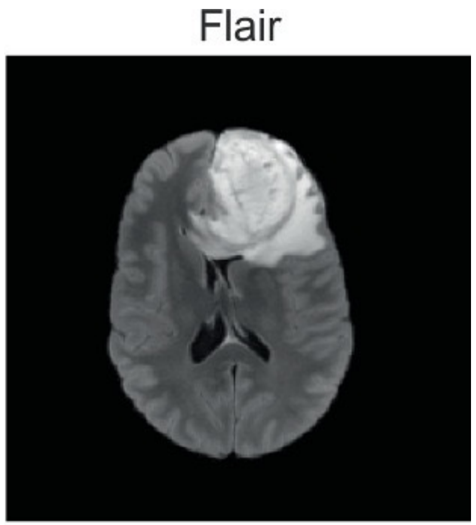
MEDIZIN INFORMATIK INITIATIVE
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
BIH Berlin Institute of Health @Charité

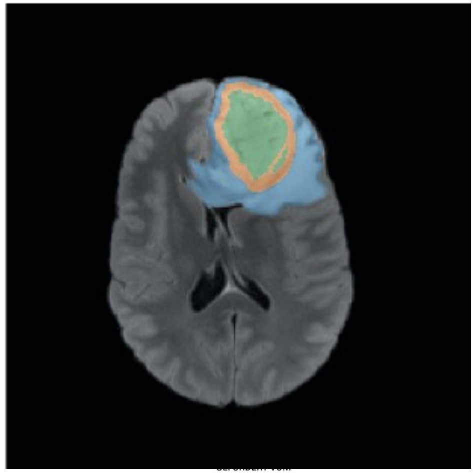
Remaining issues: liquid biopsies

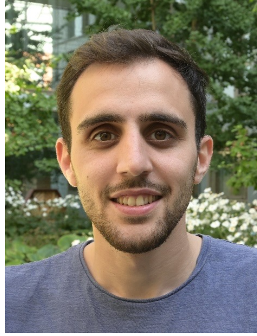
- Liquid biopsies generally contain less tumor DNA than solid tumor biopsies
- The proportion is dependent on size, proliferation, and apoptosis of the tumor

➔ Tuning the sensitivity of the tumor based on imaging

… if we manage to get access to enough samples for which we have both






Nabil Jabareen

MEDIZIN INFORMATIK INITIATIVE

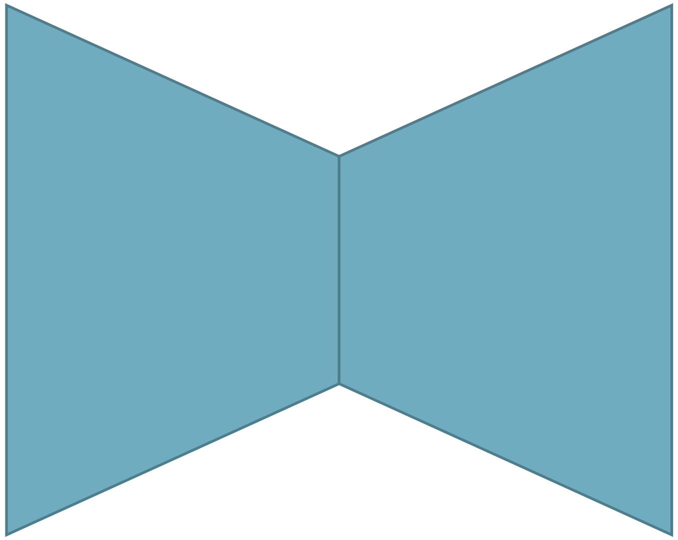
Bundesministerium für Bildung und Forschung

BIH Berlin Institute of Health @Charité

Translating multi-modal data

Foo Wei Ten

Autoencoder



encoder

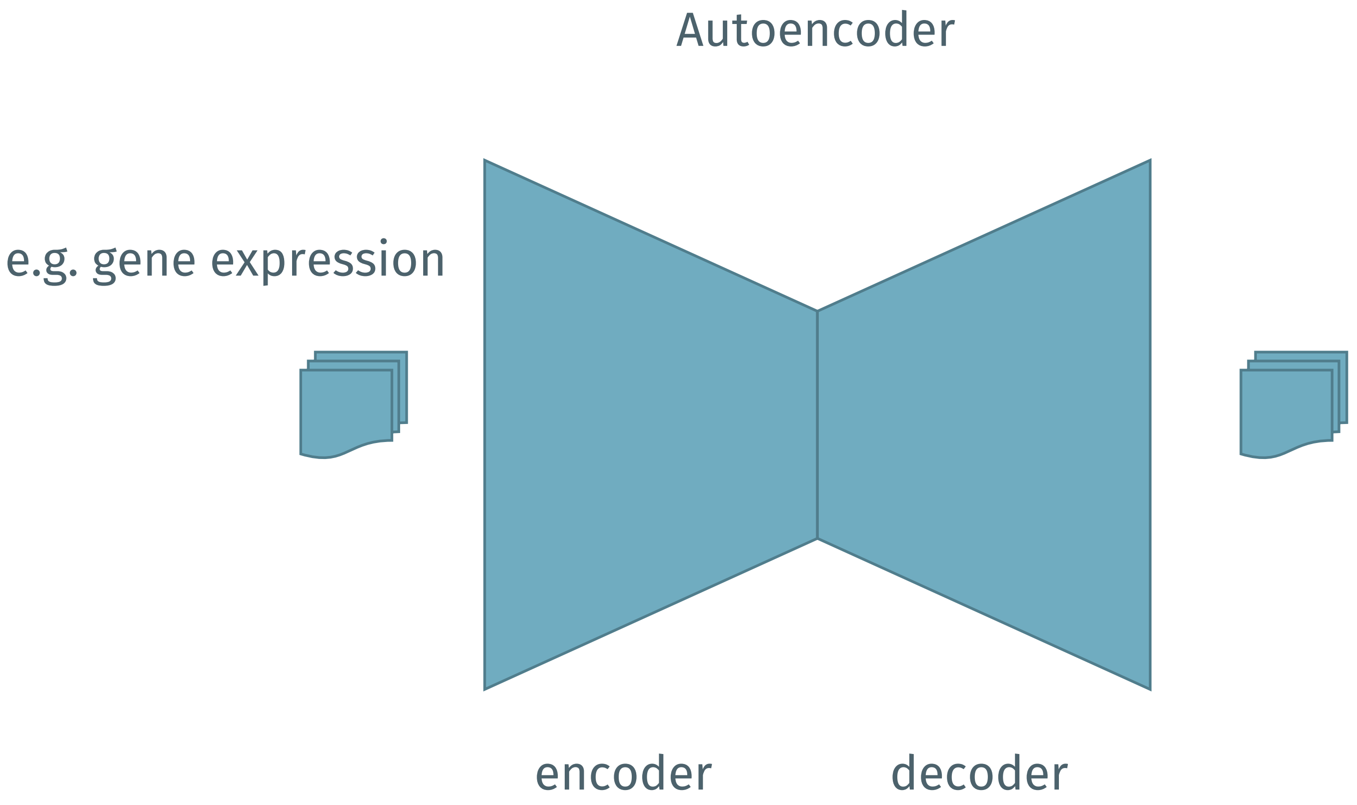
decoder

MEDIZIN INFORMATIK INITIATIVE

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

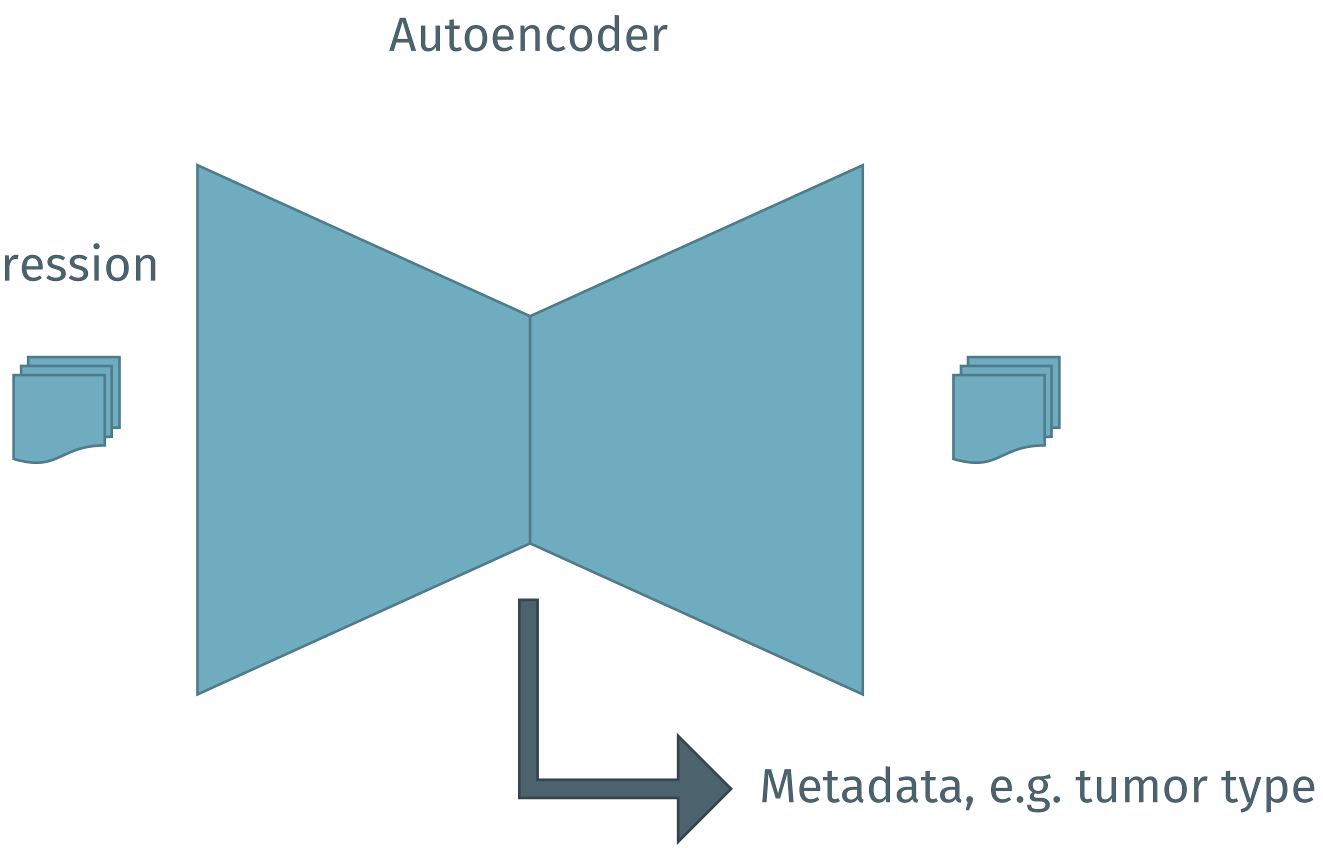
BIH Berlin Institute of Health @Charité

Translating multi-modal data



MEDIZIN INFORMATIK INITIATIVE
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
BIH Berlin Institute of Health @Charité

Translating multi-modal data



Autoencoder

e.g. gene expression

Metadata, e.g. tumor type

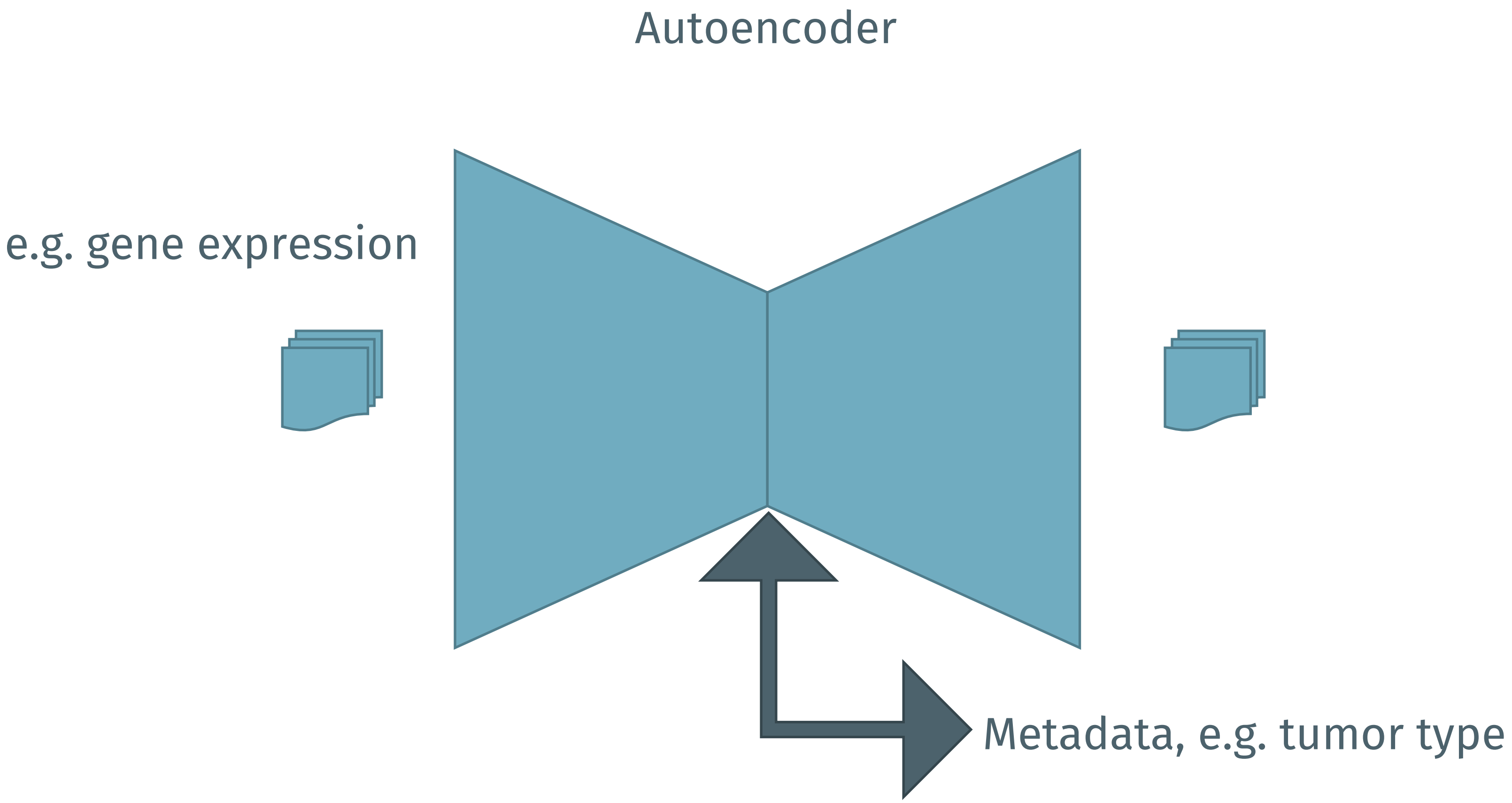
GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

MEDIZIN INFORMATIK INITIATIVE

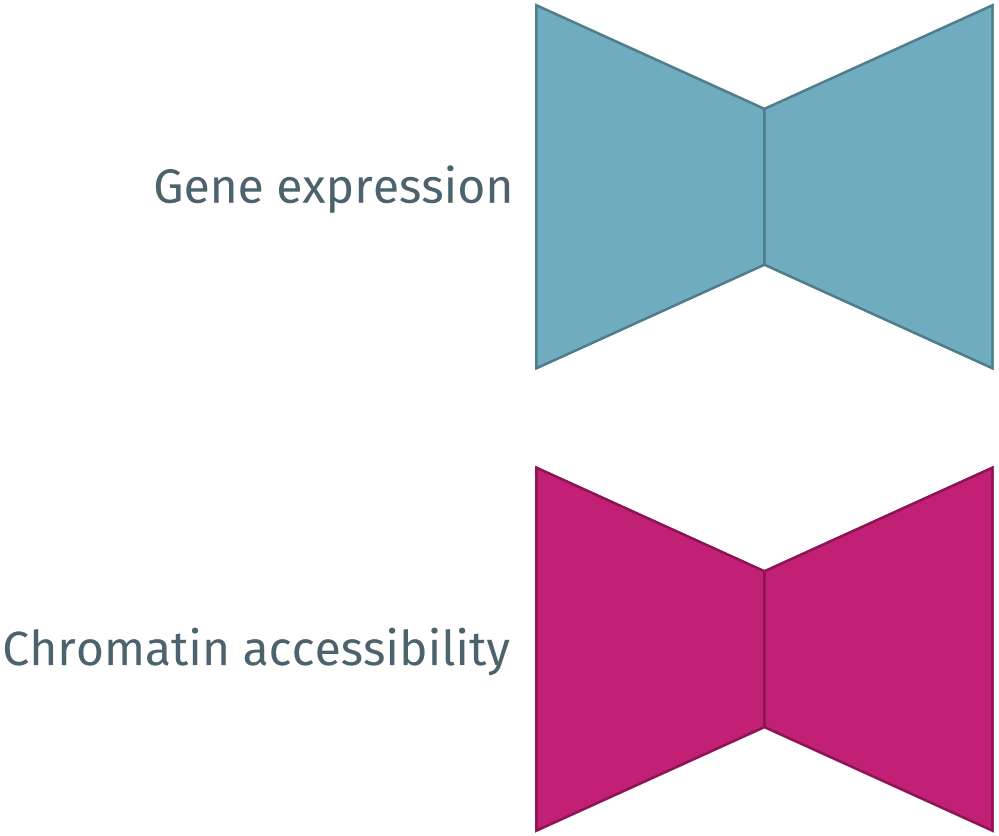
BIH Berlin Institute of Health @Charité

Translating multi-modal data



MEDIZIN INFORMATIK INITIATIVE
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
BIH Berlin Institute of Health @Charité

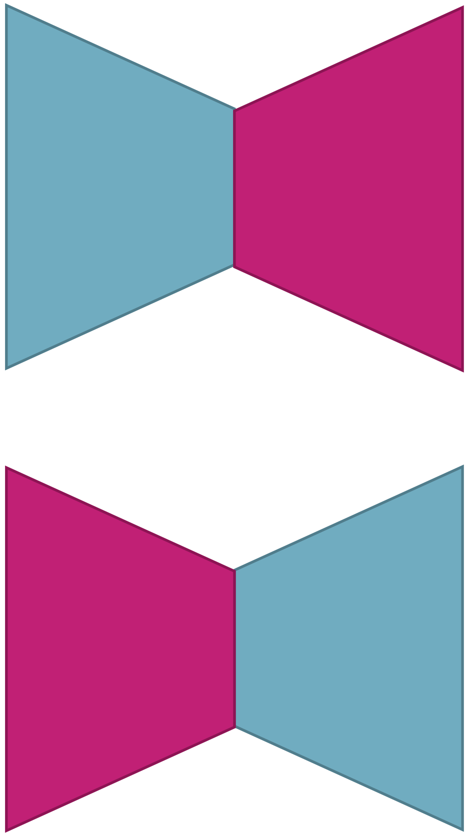
Translating multi-modal data



➔ Similar to a multi-layered non-linear consensus NMF

➔ Can be used similarly: decoder layers capture compentents, e.g. gene sets

Translating multi-modal data



MEDIZIN INFORMATIK INITIATIVE

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

BIH Berlin Institute of Health @Charité

Summary

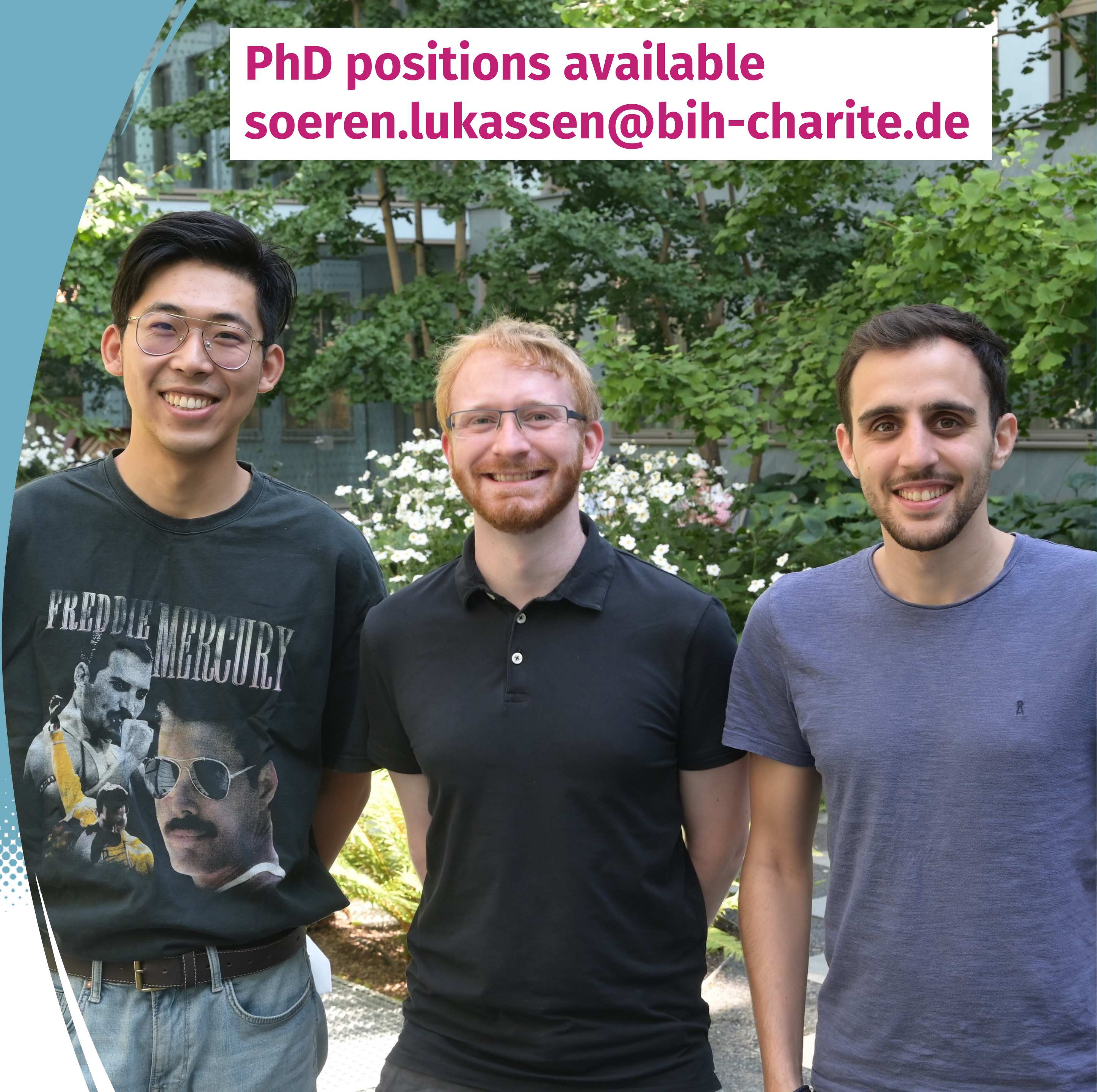
1. Low amounts of training data and high missingness don‘t necessarily doom a ML project
   - If redundancy is high (DNA methylation) or there is a constant structure (medical imaging)
2. Start simple*
3. Real-world data are noisy, incomplete, and hard to get → if possible, try to use methods where samples don‘t have to match
4. When planning a ML project, ~80% of the time is used for data curation even if the data exist already

* Starting complex can give you an idea whether there is something in the data

MEDIZIN INFORMATIK INITIATIVE

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

BIH Berlin Institute of Health @Charité



PHD positions available
soeren.lukassen@bih-charite.de

AG Lukassen, BIH
Nabil Jabareen
Dongsheng Yuan
Chantal Kühn
Sharmilaa Ramakrishnan

AG Ishaque, BIH
Naveed Ishaque

AG Conrad, BIH
Christian Conrad
Foo Wei Ten

Neurology, Charité
Philipp Euskirchen
Luis Kuschel

Radiooncology, Charité
David Kaul
Felix Ehret

Radiology, Südharz Klinikum Nordhausen
Ismini Papageorgiou

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

MEDIZIN
INFORMATIK
INITIATIVE