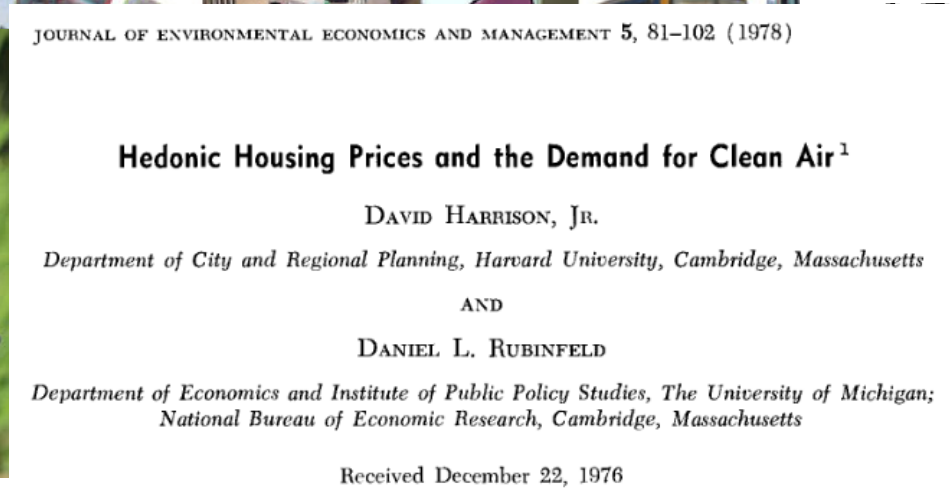
# Kaum Daten und trotzdem genau? Tumorklassifikation mittels shallow nanopore sequencing

**Sören Lukassen**
BIH Digital Health Center
6. Oktober 2022

Berlin Institute of Health @Charité

# Dataset examples



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

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

DAVID HARRISON, JR.

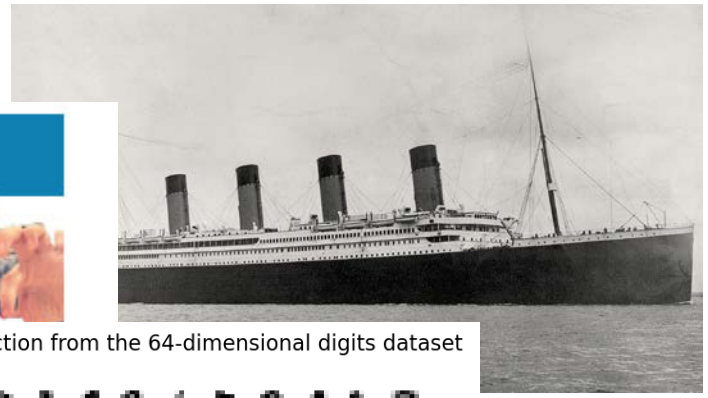
*Department of City and Regional Planning, Harvard University, Cambridge, Massachusetts*

AND

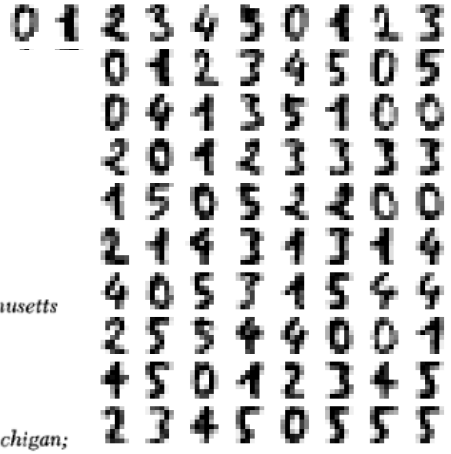
DANIEL L. RUBINFELD

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

Received December 22, 1976

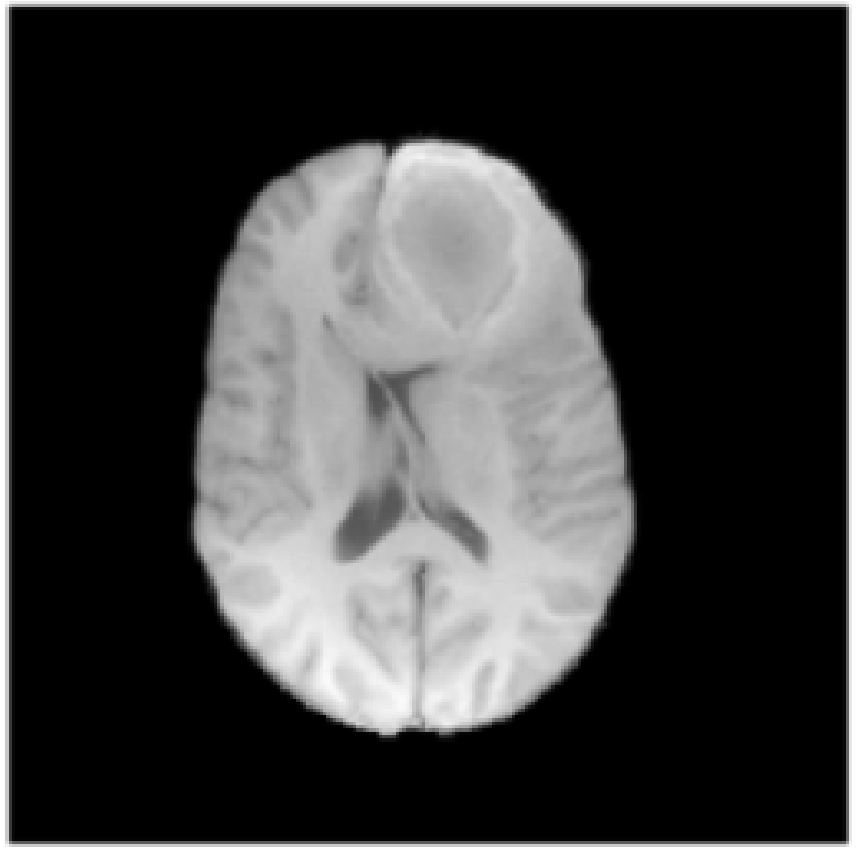


A selection from the 64-dimensional digits dataset



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

# Unser Feld: (molekulare) Hirntumordiagnostik

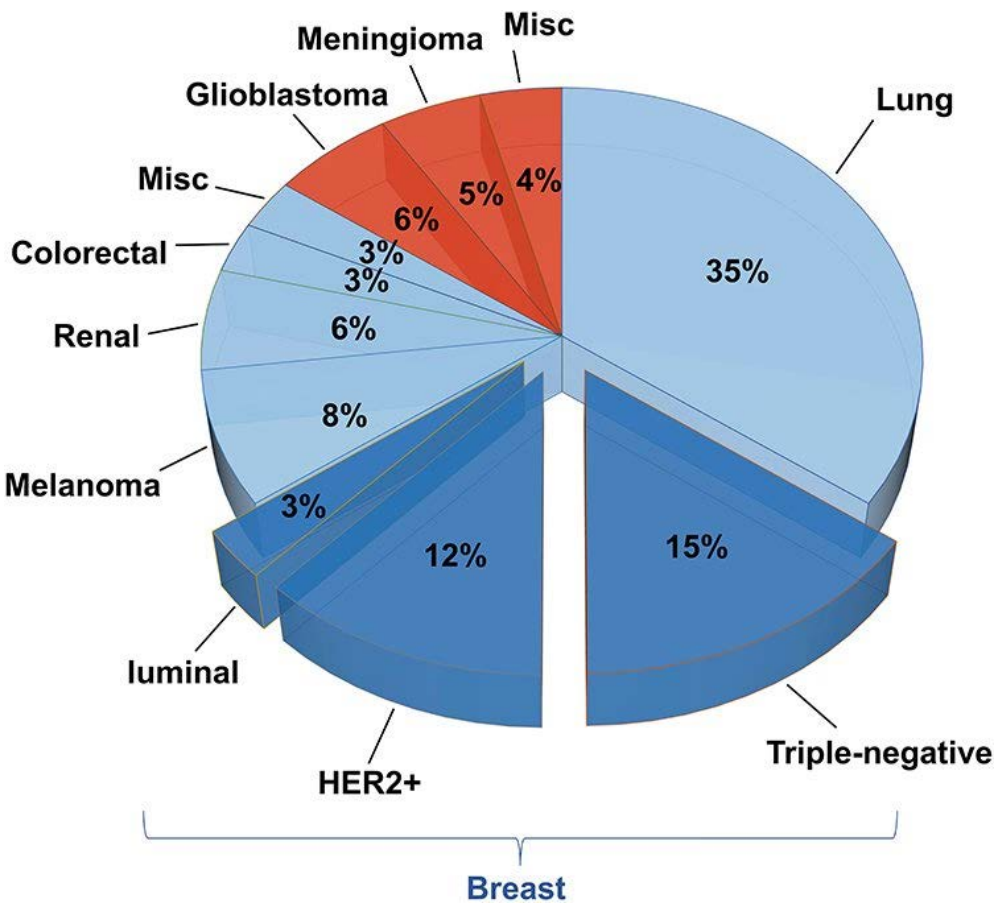


Jabareen & Lukassen, 2022



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

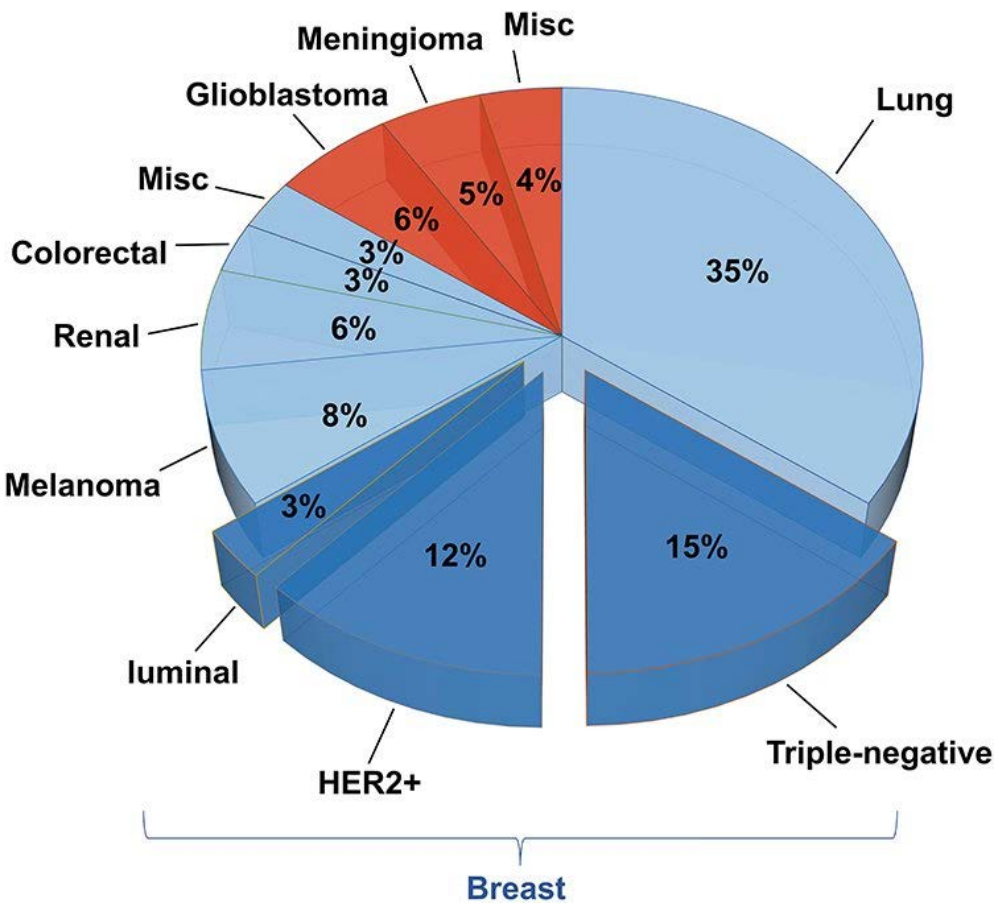
# Unser Feld: (molekulare) Hirntumordiagnostik



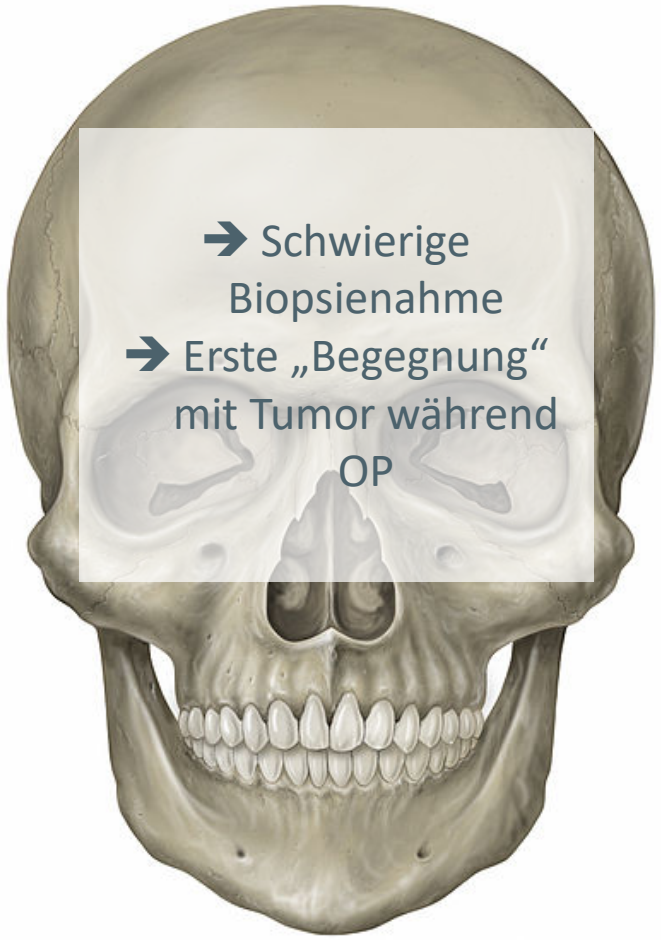

Lv et al., 2021

© Patrick J. Lynch, CC2.5

# Unser Feld: (molekulare) Hirntumordiagnostik



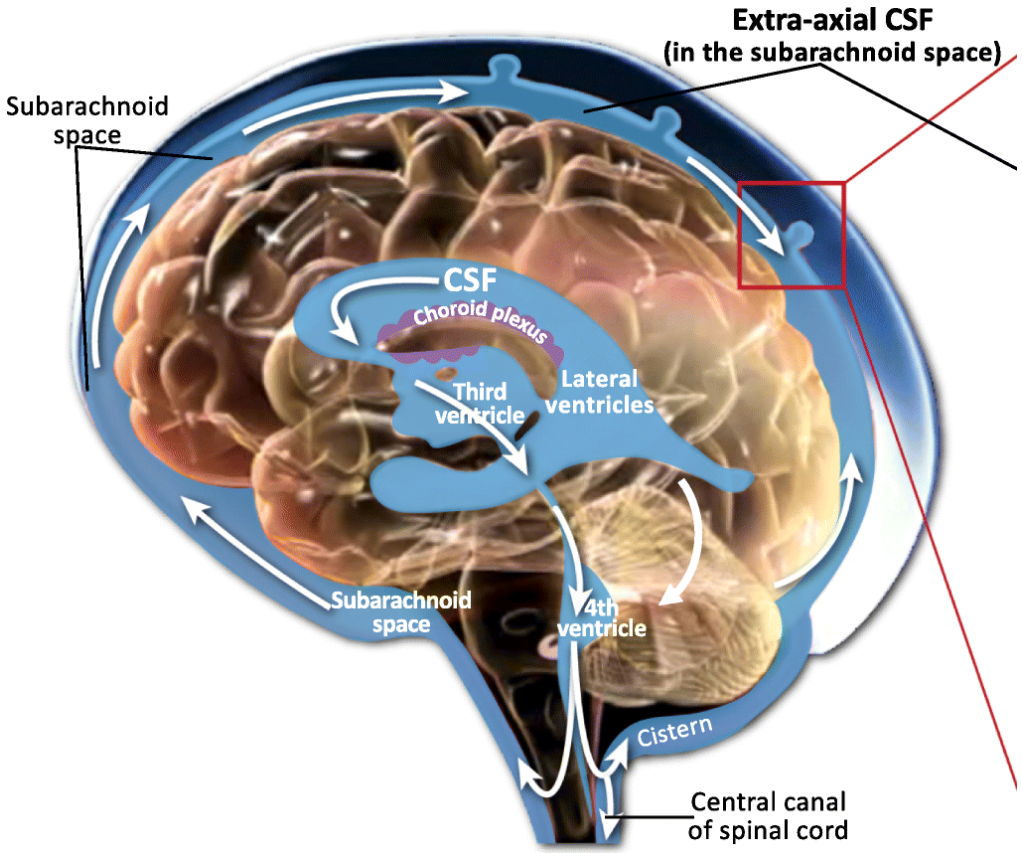

Lv et al., 2021



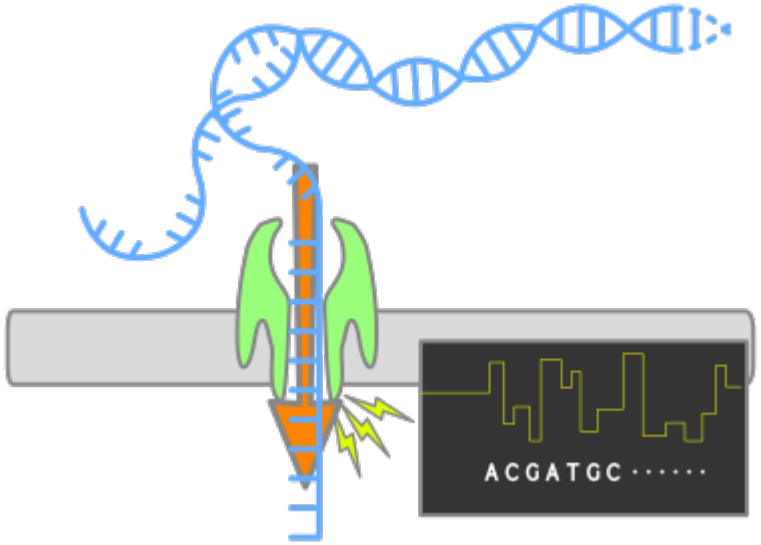

© Patrick J. Lynch, CC2.5

© Patrick J. Lynch, CC2.5

# Einfach & schnell: Nanopore sequencing von Liquor




© Mark D. Shen, CC4, cropped




© DBCLS, CC4

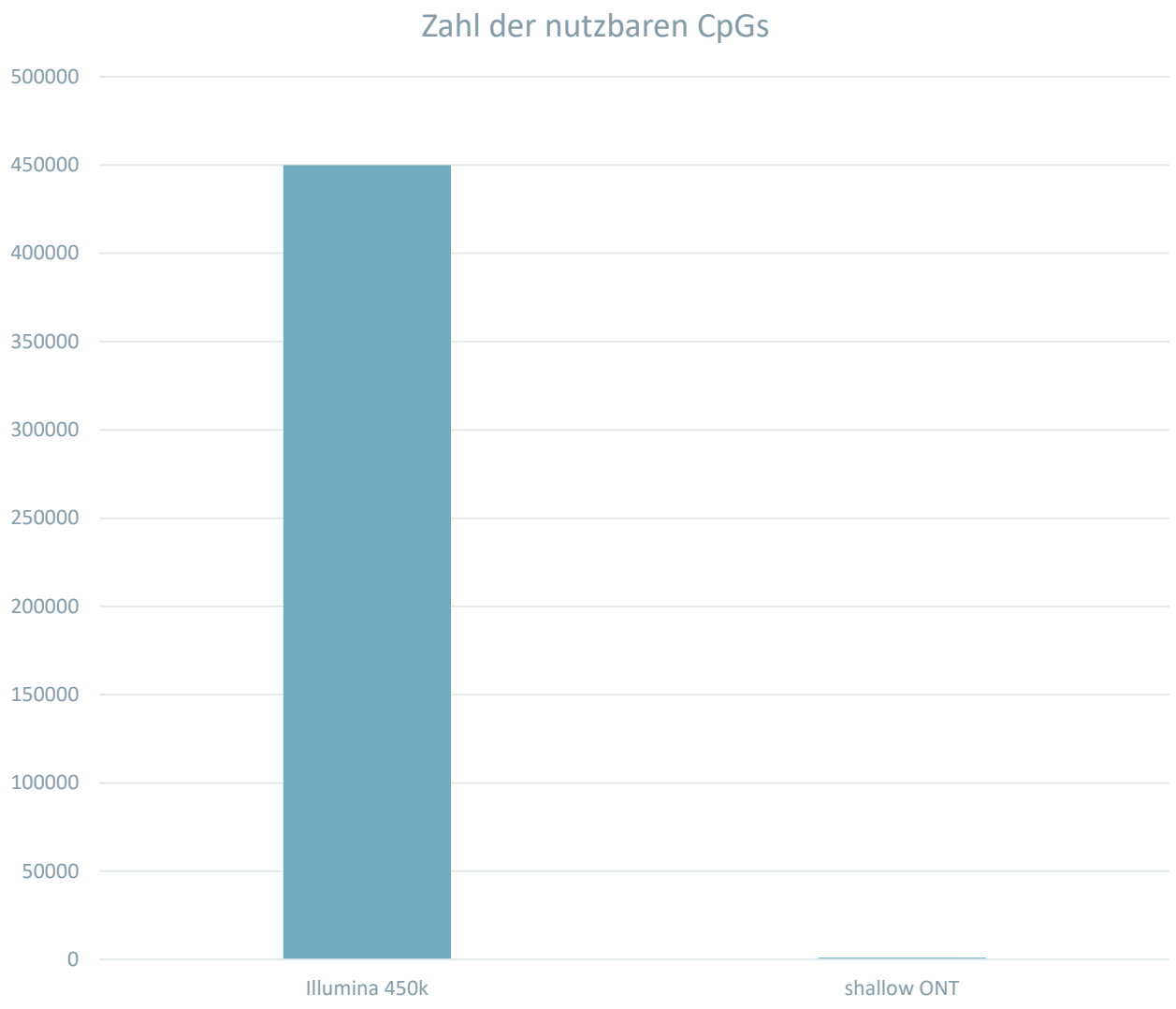


MEDIZIN INFORMATIK INITIATIVE

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

BIH Berlin Institute of Health @Charité

# Die Probleme…




- Viele Prädiktoren, aber nur wenige verfügbar

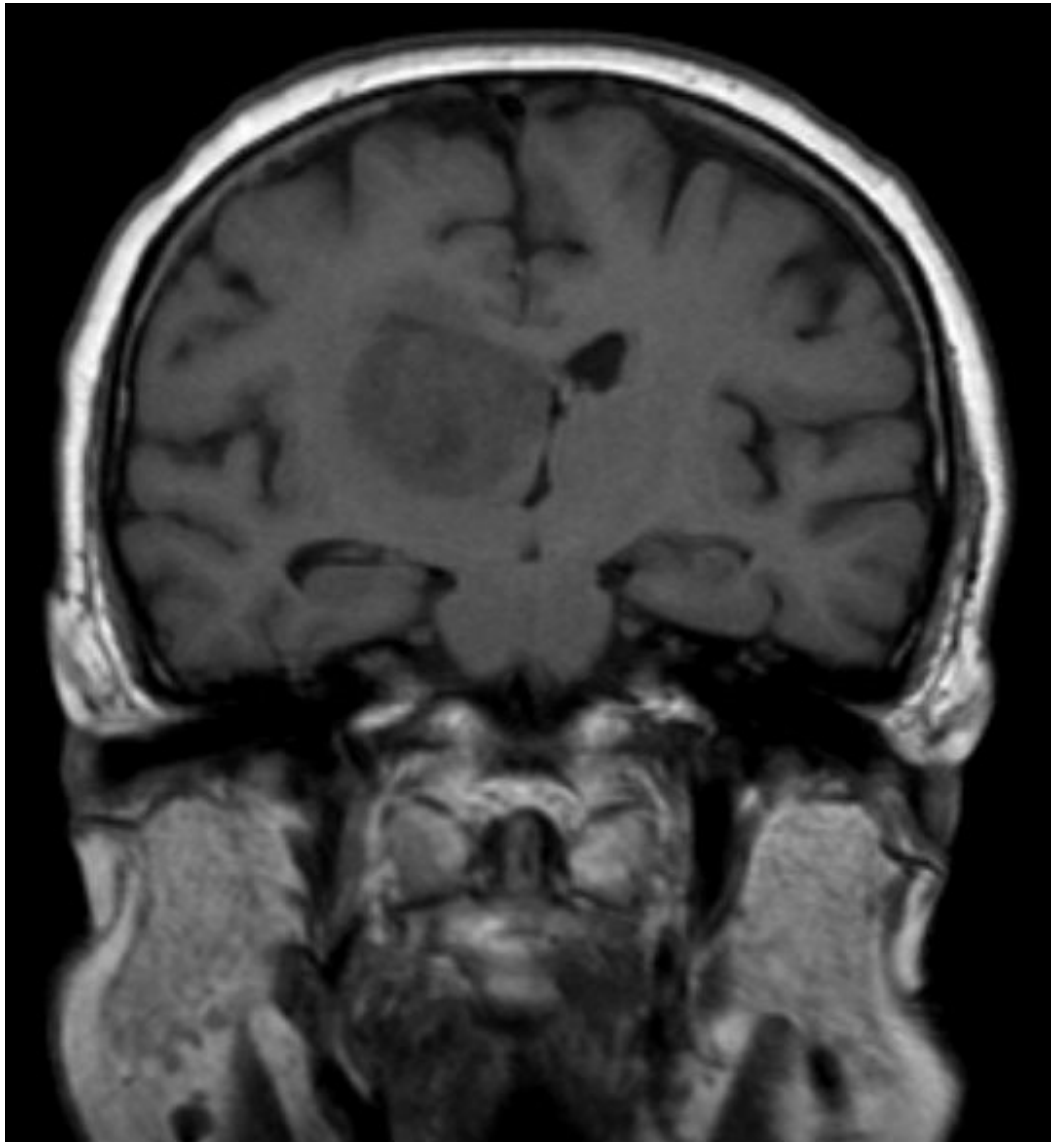


MEDIZIN INFORMATIK INITIATIVE

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

BIH Berlin Institute of Health @Charité

# Die Probleme…



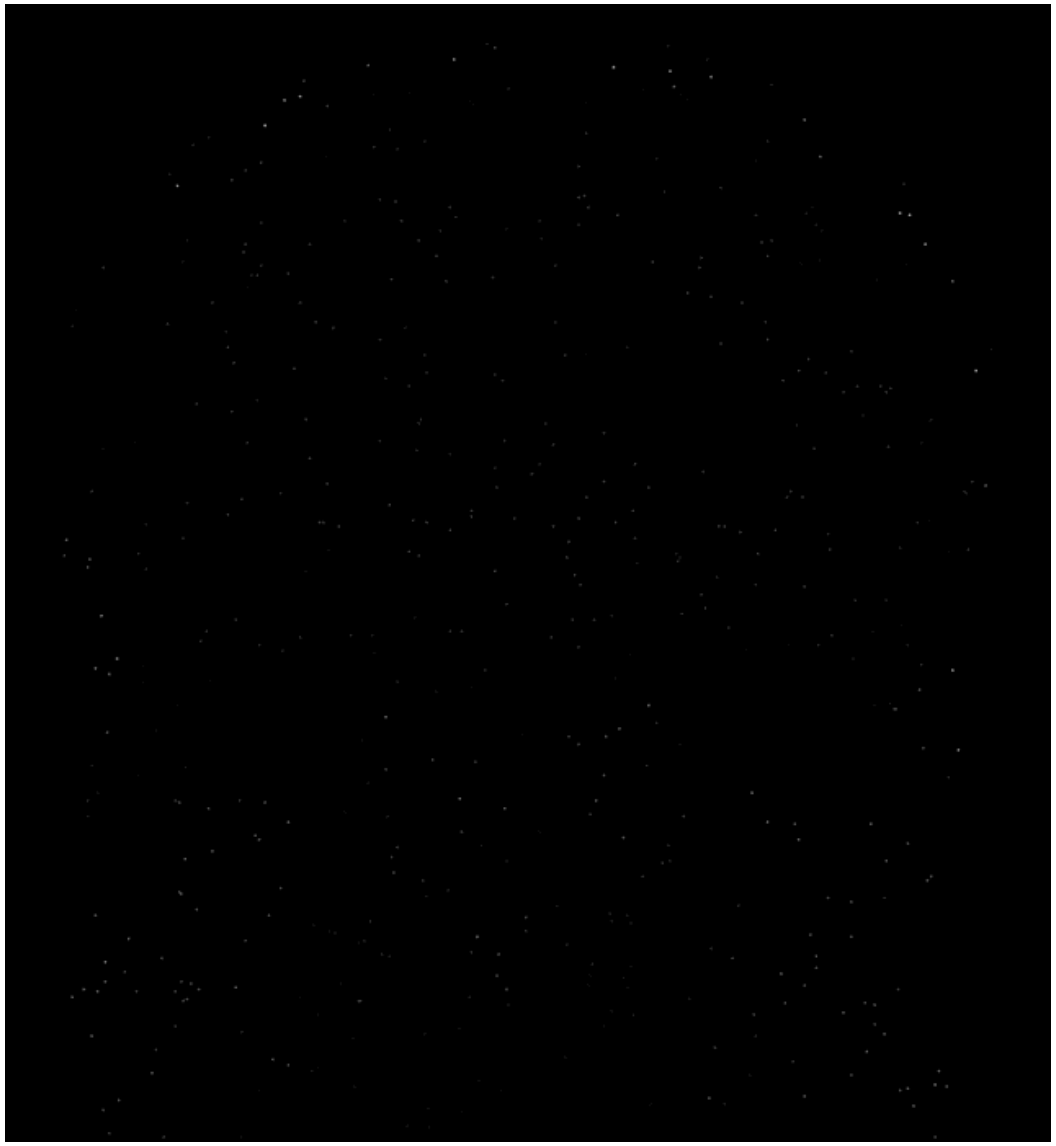
- Viele Prädiktoren, aber nur wenige verfügbar

© Hellerhoff, CC3-SA



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

# Die Probleme…

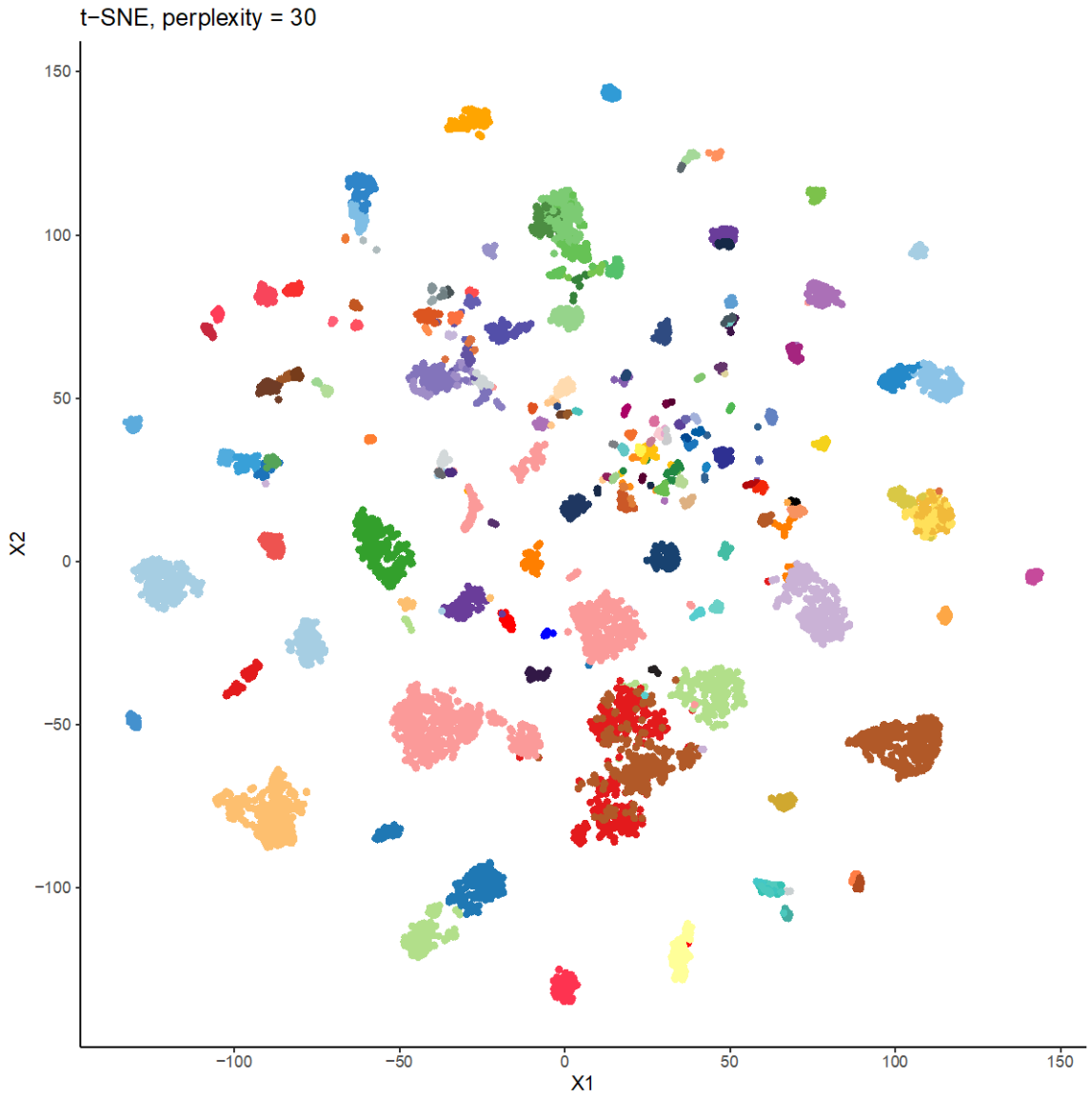


- Viele Prädiktoren, aber nur wenige verfügbar



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

# Die Probleme…




- Viele Prädiktoren, aber nur wenige verfügbar
- Relativ wenige Datenpunkte (~8000)
- n << p
- Viele Klassen (~180)
- Klassen sind nicht balanciert

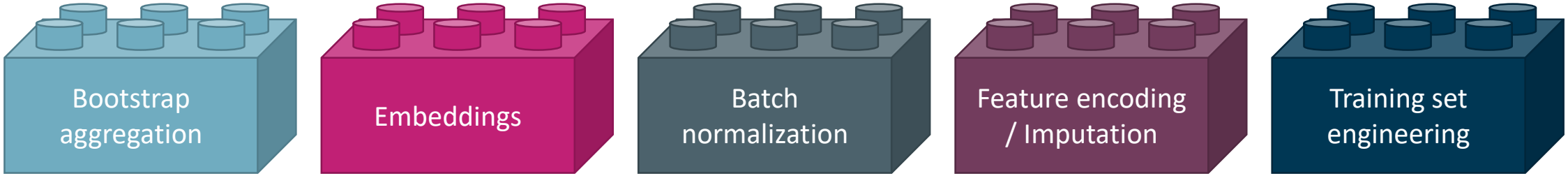
**→ Unmöglich**

**Oder doch nicht?**



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

# Die Bausteine




Verhindert Overfitting

Embeddings reduzieren die Zahl der Prädiktoren

Idealerweise werden Unterschiede zwischen Prädiktorsets ausgeglichen

Hält bei Deep Learning die Verteilung der Aktivierungen trotz variabler Prädiktoren konstant

Umgang mit fehlenden Werten: Mittelwert? 0? 0.5?

Kann potenziell Fehlen von Prädiktoren in Information verwandeln

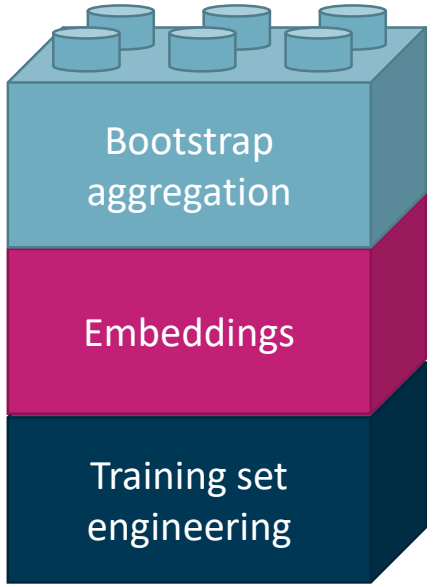
Reduktion der Prädiktoren des Trainingssets auf die in der zu klassifizierenden Probe vorhandenen

Erfordert wiederholtes Training



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

# Strategie 1: Beste Ergebnisse, großer Aufwand




- Verwendung nur von CpGs, die in der zu klassifizierenden Probe enthalten sind
- Gemeinsames Embedding von Probe und Trainingsset; Auswahl der ähnlichsten Klassen

  → Neues Training pro Probe
- Random forest als Klassifikator

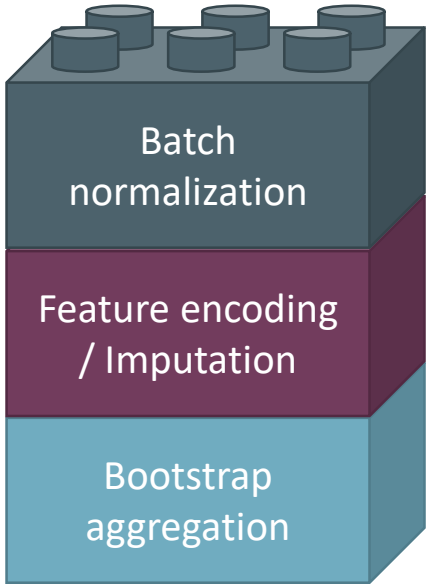
➔ Genauigkeit auf Gewebsbiopsie: > 90% (Top1)

➔ Bei < 1000 CpGs: > 60%



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

# Strategie 1: Beste Ergebnisse, großer Aufwand



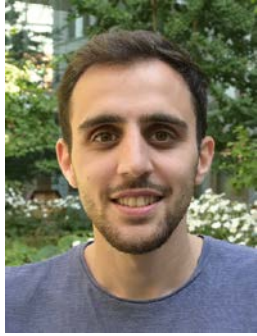

- Multilayer Perceptron
- Volles Trainingsset, Prädiktoren werden pro Epoche zufällig gewählt
- Encoding: methyliert/unmethyliert/fehlend
- Batch normalization, um verschiedene Prädiktorzahlen auszugleichen

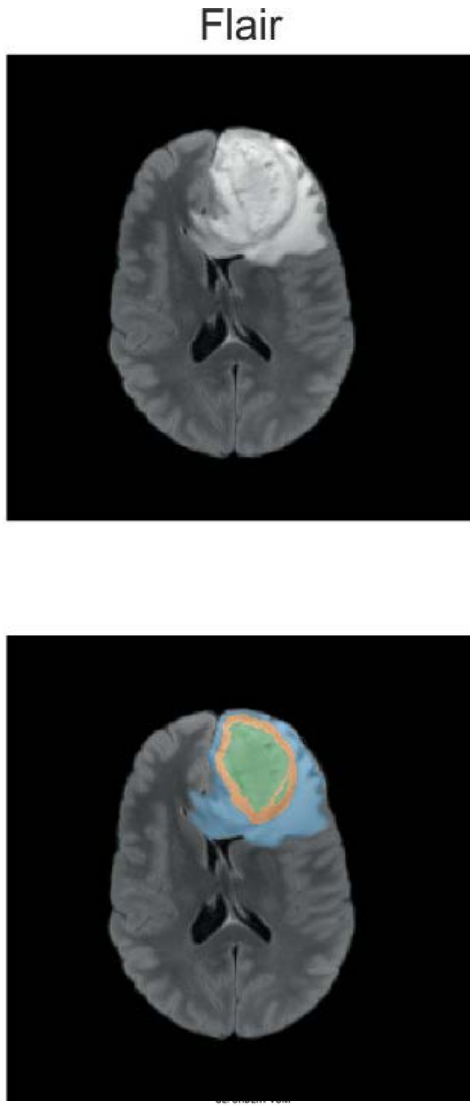
➔ Genauigkeit auf Gewebsbiopsie bei >99% fehlenden CpGs: > 80% (Top1)



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

# Liquid biopsies: das Problem mit dem Tumoranteil






- Liquid biopsies enthalten i.d.R. einen kleineren Anteil Tumor-DNA als Gewebsbiopsien. Darunter leidet die Genauigkeit
- Der Anteil ist abhängig von Größe, Proliferations- und Apoptoserate des Tumors

➔ Tuning des Klassifikators anhand dieser Parameter, ermittelt aus Bildgebung

… falls genug Patientendaten mit gepaarter Bildgebung und DNA-Methylierungsanalyse zugänglich sind



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

# Was lernen wir daraus?

**Maschinelles Lernen ist auch mit sehr kleinen, und höchst unvollständigen Datensätzen mit vielen Klassen möglich, wenn…**

**… die Parameter ausreichend redundant sind**

**… die Klassen sich auf Basis der Parameter gut unterscheiden lassen**

**… man ausreichende Ressourcen hat**

**Die Kombination von verschiedenen Datentypen (hier: DNA-Methylierung und Bildgebung) kann weitere Anwendungsbereiche erschließen. Aber:**

**➔ Datensilos sind große Hindernisse**



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

# Vielen Dank!

*AG Lukassen, BIH*

Nabil Jabareen
Dongsheng Yuan
Chantal Kühn
Sharmilaa Ramakrishnan

*AG Ishaque, BIH*

Naveed Ishaque

*AG Conrad, BIH*

Christian Conrad
Foo Wei Ten

*Neurologie, Charité*

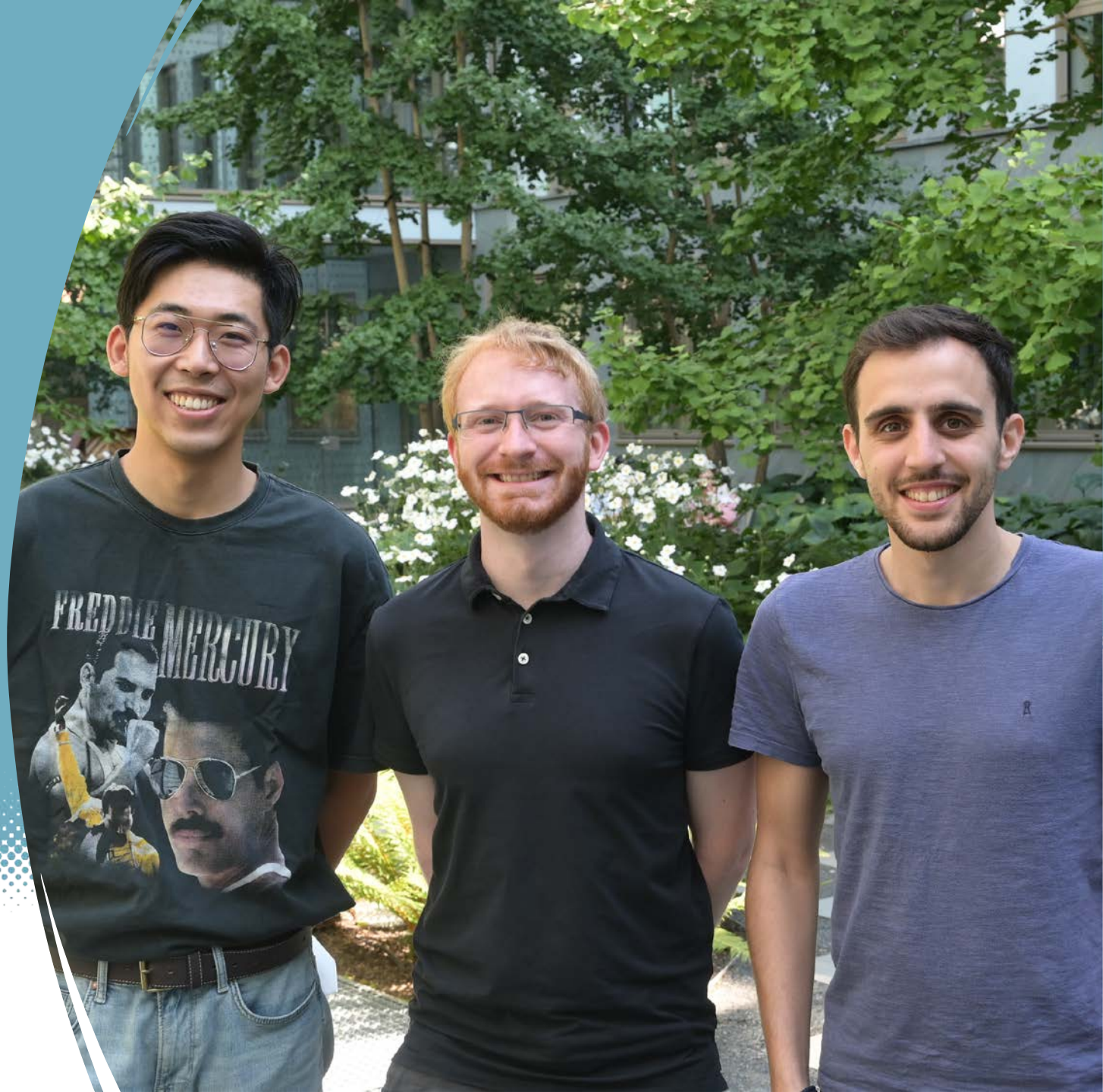
Philipp Euskirchen
Luis Kuschel

*Radioonkologie, Charité*

David Kaul
Felix Ehret

*Radiologie, Südharz Klinikum Nordhausen*

Ismini Papageorgiou




GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

MEDIZIN
INFORMATIK
INITIATIVE