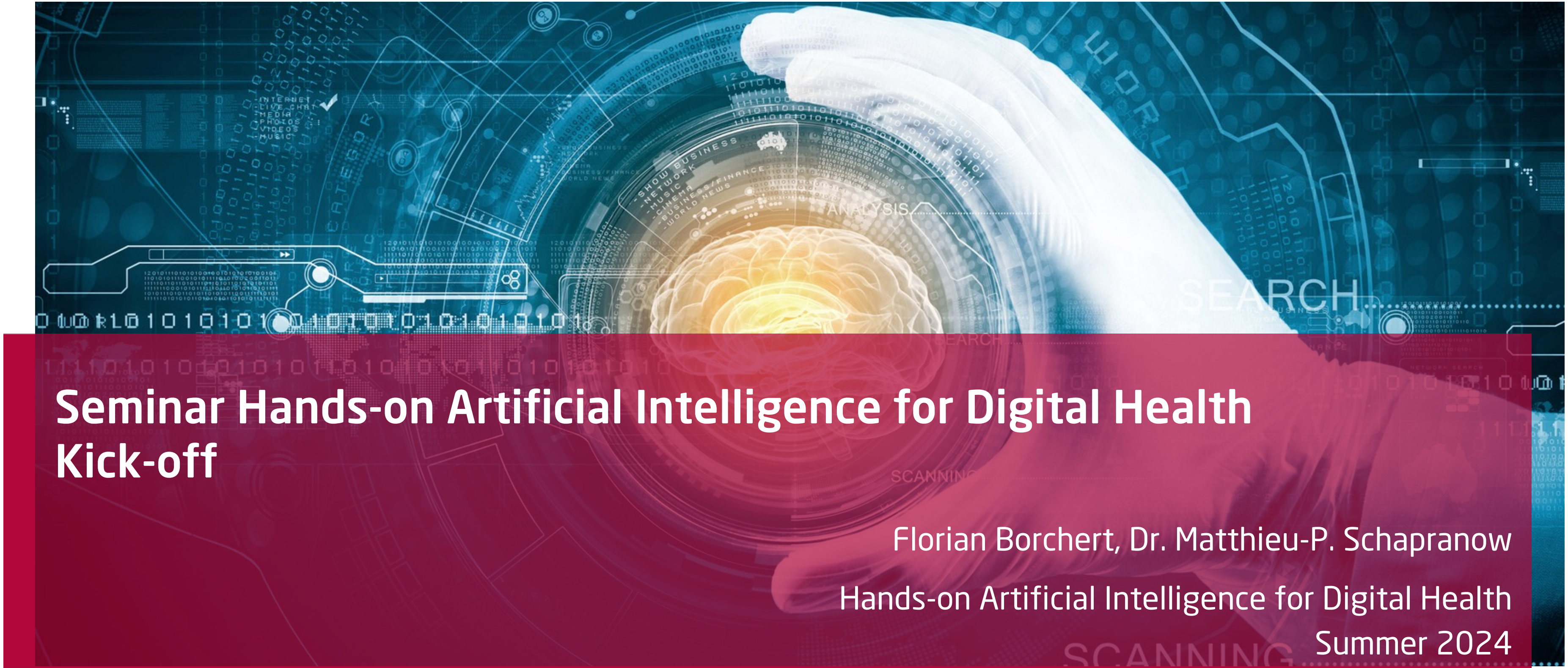


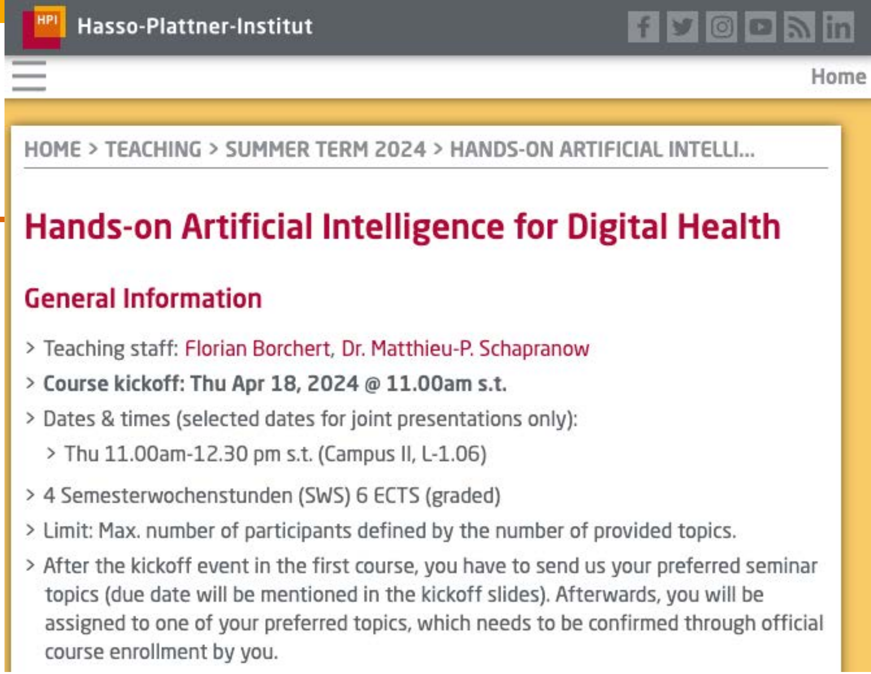
Seminar Hands-on Artificial Intelligence for Digital Health
Kick-off

Florian Borchert, Dr. Matthieu-P. Schapranow

Hands-on Artificial Intelligence for Digital Health
Summer 2024

Seminar Organization
Administrative Details Time / Dates

- Format: Seminar
- Scope: 4 SWS (6 graded ECTS)
- Dates and time:
  - Thursdays 11.00am - 12.30pm and
  - Individual appointments with your supervisor
- Further details are available on the seminar website: https://hpi.de/digital-health-cluster/teaching/summer-term-2024/hands-on-artificial-intelligence-for-digital-health.html



Hands-on Artificial Intelligence for Digital Health

General Information

- Teaching staff: Florian Borchert, Dr. Matthieu-P. Schapranow
- **Course kickoff: Thu Apr 18, 2024 @ 11.00am s.t.**
- Dates & times (selected dates for joint presentations only):
  - Thu 11.00am-12.30 pm s.t. (Campus II, L-1.06)
- 4 Semesterwochenstunden (SWS) 6 ECTS (graded)
- Limit: Max. number of participants defined by the number of provided topics.
- After the kickoff event in the first course, you have to send us your preferred seminar topics (due date will be mentioned in the kickoff slides). Afterwards, you will be assigned to one of your preferred topics, which needs to be confirmed through official course enrollment by you.

Seminar Organization
What You Can Expect from Us

- Broaden your horizons in the fields of
  - Digital Health,
  - Life sciences, as well as
  - Data challenges and opportunities
- Work with real-world data, real-world use cases
- Hands-on experiments of selected tools
- Experience in scientific writing and presentation skills

https://www.haselden.com/wp-content/uploads/2019/01/65298106_s.jpg

Seminar Organization
What We Expect from You

- Commitment to your seminar topic
- Regular participation in all presentations and update meetings
- Active participation in group discussions
- Perform autonomous research to dig deeper into the topics
- Contribute with your expertise also to your colleagues
- Update supervisors on any issues you might encounter

https://www.haselden.com/wp-content/uploads/2019/01/65298106_s.jpg

Seminar Organization
Grading

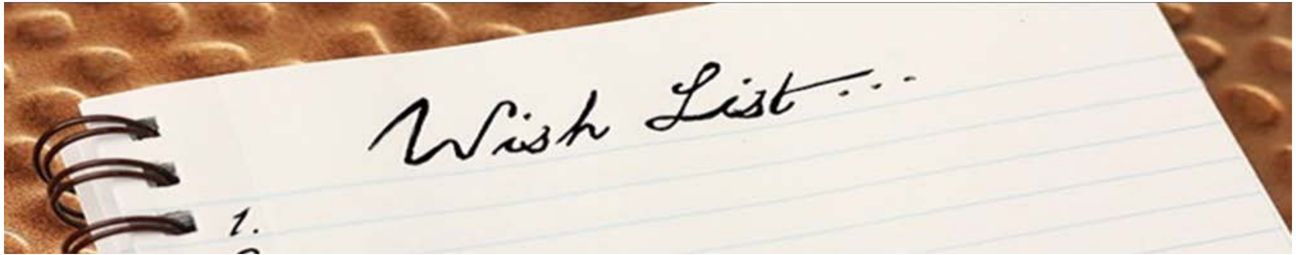
- The grading of the seminar works as follows (aka “Leistungserfassungsprozess”):
  - 40% Seminar results, i.e.
    - Intermediate & final presentation conducted during seminar slots
    - Research prototype
  - 40% Scientific research article about your individual contribution submitted by the end of the seminar
  - 20% Individual commitment throughout the seminar

- **All individual parts have to be passed to pass the complete seminar**

https://hpi.de/fileadmin/user_upload/hpi/bilder/galerien/2022/Eröffnung_Haus_L/KayH20220401-1M7A3644.jpg

Next Steps
Enrollment Process: How to apply for a topic?

- Send prioritized list of top three topics to Florian Borchert (Florian.Borchert@hpi.de) by **Mon Apr 22, 2024 9am (sharp)**
  - 1st choice: …
  - 2nd choice: …
  - 3rd choice: …
- Assignment of seminar topics: **Mon Apr 22, 2024 by noon**
- Enrollment deadline: **Mon Apr 29, 2024 end of day (via Studienreferat)**



Seminar Schedule: Presentations

- **May 23rd (+ May 30th)** Intermediate presentations
  - 10 minutes presentation
  - Introduce your topic, problem/motivation, how you want to solve it
  - Slides due at day of presentation by 9am

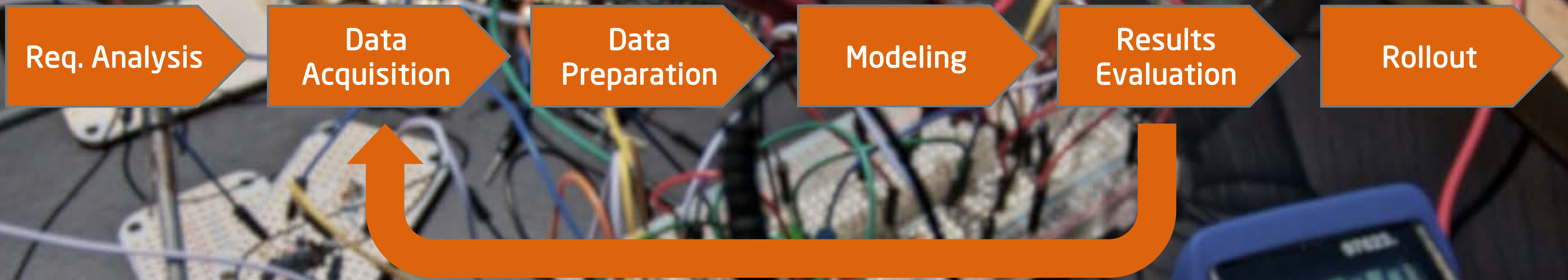
- **July 11th** (+ **July 18th**) Final presentations
  - 20 minutes presentation
  - Slides due at day of presentation by 9am
  - Present your approach and results achieved

Seminar Schedule: Paper Writing

- **July 18th**: Introduction to scientific writing

- **Aug 18th (end of day):** Project results submission
  - One paper per topic
  - Max. 6 pages excluding appendix
  - Iterate regularly with your supervisor prior to submission
  - Source code for project and paper (LaTeX) including documentation how to build it

Machine Learning for Digital Health Process:


Req. Analysis
Data Acquisition
Data Preparation
Modeling
Results Evaluation
Rollout

While designing the following topics, we had in mind, that each of them should cover selected steps of the ML process to allow you to broaden your competency.

Photo by Nicolas Thomas

Overview of Seminar Topics

A. LLM-based Named Entity Recognition for Genes and Variants

B. BioASQ MultiCardioNER (Shared Task)

C. Entity Normalization for Tumor Morphologies

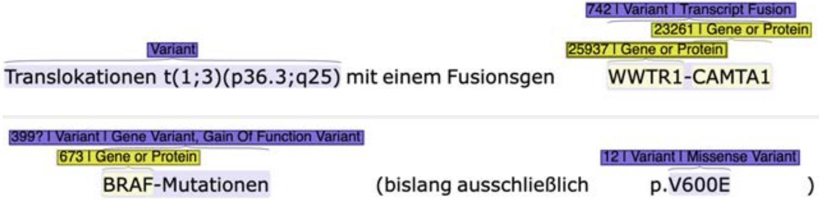
D. Gene Name Normalization

E. Explore ML applications for *omics data

F. Can ML support discovery new cancer subtypes?

A: LLM-based Named Entity Recognition for Genes and Variants

- Automated extraction of gene and variant mentions is important for precision medicine
- **Your tasks:**
  - ☐ Train an NER tagger based on a manually labelled German dataset of gene / variant mentions
  - ☐ Compare with an LLM-based zero-shot / few-shot prompting approach (and investigate different prompting techniques, e.g., self-verification [1])



Translokationen t(1;3)(p36.3;q25) mit einem Fusionsgen WWTR1-CAMTA1

BRAF-Mutationen (bislang ausschließlich p.V600E)

Examples of gene and variant mentions in GGPONC

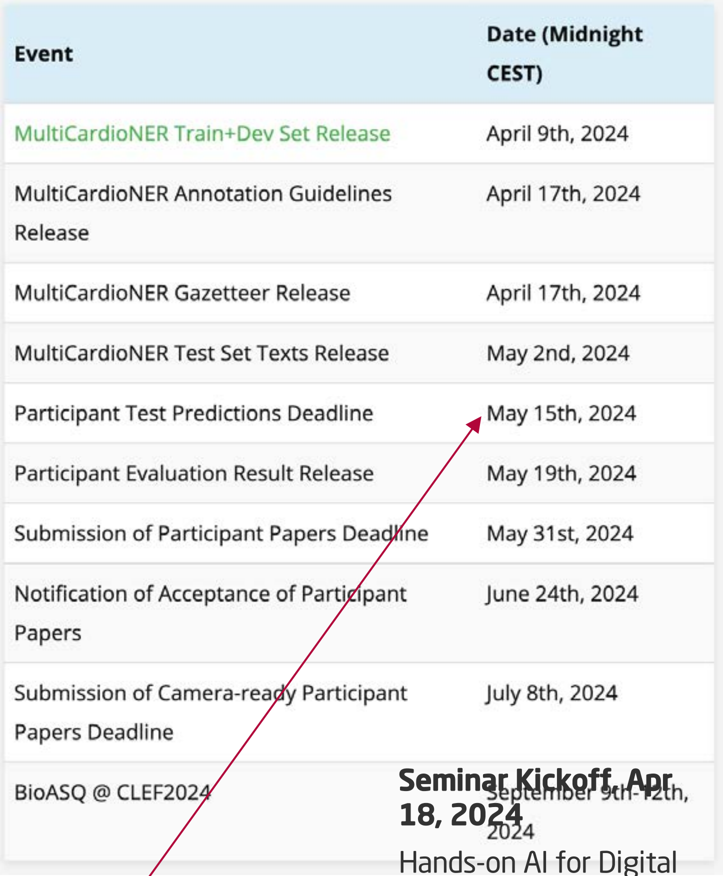
- **Data**: > 1k manual annotations of genes and variants in GGPONC + much more unlabeled data
- **Requirements**: Good Python skills, basic ML knowledge, German language might be helpful

---

[1] https://openreview.net/pdf?id=SBbJICrglS

B: BioASQ MultiCardioNER (Shared Task)

- Focus of the MultiCardioNER shared task is the domain adaptation of NER systems: https://temu.bsc.es/multicardioner/
- Two tracks: (1) Spanish, (2) Multi-lingual
- **Your tasks:**
  - ☐ Develop NER system(s) for any of the tracks, building upon our prior work [1]
  - ☐ Optional: Participate in the challenge
  - ☐ Win!
- **Data:** MultiCardioNER corpus
- **Requirements**: Prior NLP knowledge, any of the languages besides English might be helpful



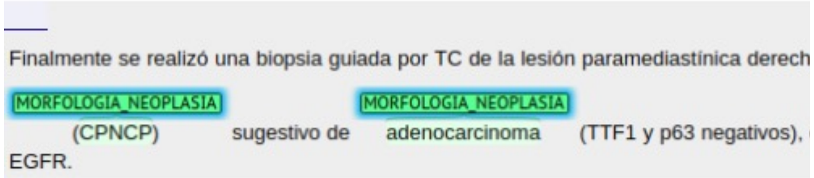
Event	Date (Midnight CEST)
MultiCardioNER Train+Dev Set Release	April 9th, 2024
MultiCardioNER Annotation Guidelines Release	April 17th, 2024
MultiCardioNER Gazetteer Release	April 17th, 2024
MultiCardioNER Test Set Texts Release	May 2nd, 2024
Participant Test Predictions Deadline	May 15th, 2024
Participant Evaluation Result Release	May 19th, 2024
Submission of Participant Papers Deadline	May 31st, 2024
Notification of Acceptance of Participant Papers	June 24th, 2024
Submission of Camera-ready Participant Papers Deadline	July 8th, 2024
BioASQ @ CLEF2024	September 9th-12th, 2024

If you choose the topic, we will register as a team asap.

[1] https://github.com/hpi-dhc/symptemist_biocreative_2023

C: Entity Normalization for Tumor Morphologies

- Entity normalization refers to the task of mapping named entity mentions to canonical identifiers in a terminology / knowledgebase
- In medicine, specialized terminologies exist for different sub-fields, e.g., ICD-O for tumor morphology
- **Your task:**
  - Use the xMEN toolkit [1] to build a state-of-the-art entity normalization pipeline for ICD-O normalization
- **Data**: CANTEMIST dataset (Spanish) [2]
- **Requirements**: Good Python skills, Spanish language might be helpful (but not necessary)

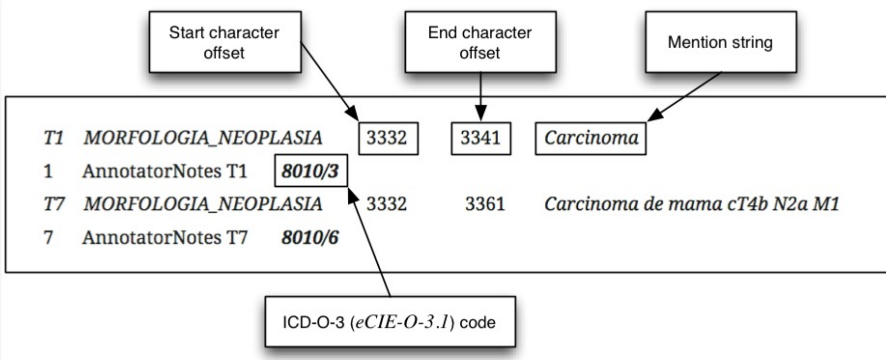


Finalmente se realizó una biopsia guiada por TC de la lesión paramediastínica derech

MORFOLOGIA_NEOPLASIA MORFOLOGIA_NEOPLASIA

(CPNCP) sugestivo de adenocarcinoma (TTF1 y p63 negativos),

EGFR.

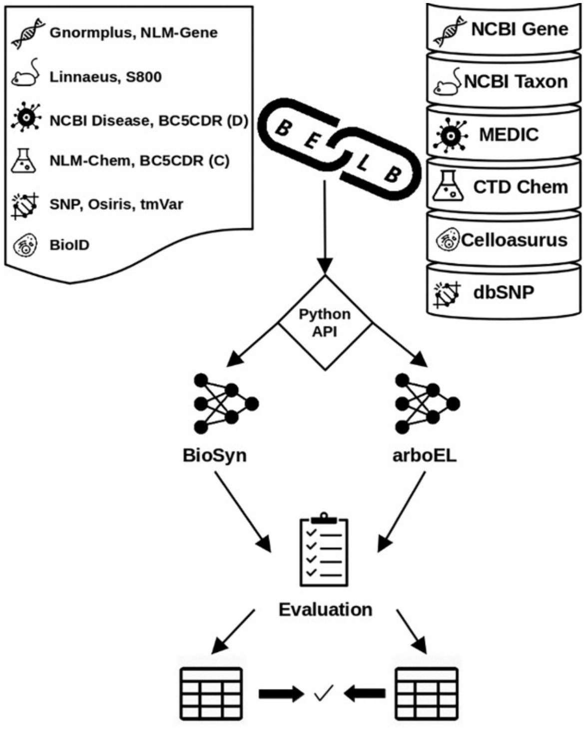


```
T1  MORFOLOGIA_NEOPLASIA  3332  3341  Carcinoma
1   AnnotatorNotes T1  8010/3
T7  MORFOLOGIA_NEOPLASIA  3332  3361  Carcinoma de mama cT4b N2a M1
7   AnnotatorNotes T7  8010/6
```

[1] https://github.com/hpi-dhc/xmen [2] https://temu.bsc.es/cantemist/

D: Gene Name Normalization

- For English, a variety of gene normalization benchmarks exist
- BELB [1] integrates these benchmarks in a unified format
- **Your tasks:**
  - Integrate BELB + xMEN Python toolkit [2]
  - Compare results to reported state-of-the-art
- **Data**: 11 corpora linked to 7 knowledge bases
- **Requirements**: Very good command of Python, data engineering and bioinformatics skills will be helpful

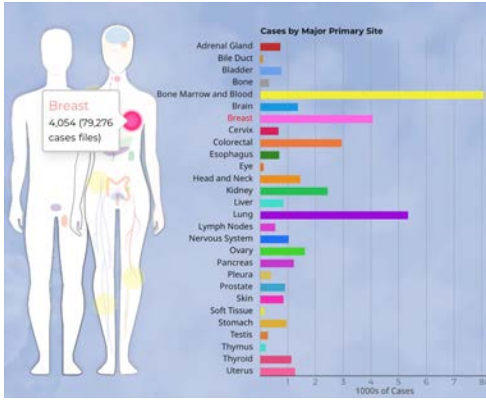


[1] https://academic.oup.com/bioinformatics/article/39/11/btad698/7425450

[2] https://github.com/hpi-dhc/xmen

E: Explore ML applications for *omics data

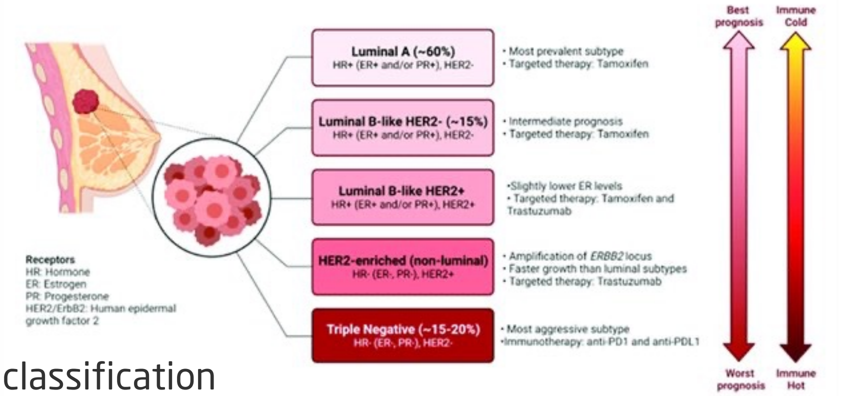
- Genetic changes are a root source for many complex diseases, e.g. cancer
- **Your tasks:**
  - Gather an overview of ML-based tools for predicting the effect of genetic changes
  - Select a subset of tools and compare the approach
  - Apply your selected tools for a selected cancer group
  - Evaluate the tools and their results
- **Data**: The Cancer Genome Atlas (TCGA): 44,6k+ cases, 69 primary sites, …
- **Requirements**: Python, bioinformatics skills are helpful
- **Literature**: Alharbi, W.S., Rashid, M. A review of deep learning applications in human genomics using next-generation sequencing data. Hum Genomics 16, 26 (2022).



https://portal.gdc.cancer.gov

F: Can ML support discovery new cancer subtypes?

- Advanced sequencing technologies helped to identify new cancer subtypes
- **Your tasks:**
  - ☐ Select ML tools for classification of cancer *omics / patient characteristics
  - ☐ Extend your approach and compare outcome to existing classification
- **Data**: The Cancer Genome Atlas (TCGA): 44,6k+ cases, 69 primary sites, …
- **Requirements**: Python, *omics skill are required
- **Literature**:
  - ☐ Park J et al. Comparison of cancer subtype identification methods combined with feature selection methods in omics data analysis. BioData Min. 2023 Jul 7;16(1):18
  - ☐ Orrantia-Borunda E et al. Subtypes of Breast Cancer. In: Mayrovitz HN, editor. Breast Cancer. Brisbane (AU): Exon Publications; 2022 Aug 6. Chapter 3.



Attalla, Sherif & Taifour, Tarek & Muller, William. (2023). Tailoring therapies to counter the divergent immune landscapes of breast cancer. Frontiers in Cell and Developmental Biology. 11. 1111796. 10.3389/fcell.2023.1111796.

Do Not Forget to Enroll for the Lecture!

We want you!

Enroll by Apr 29, 2024

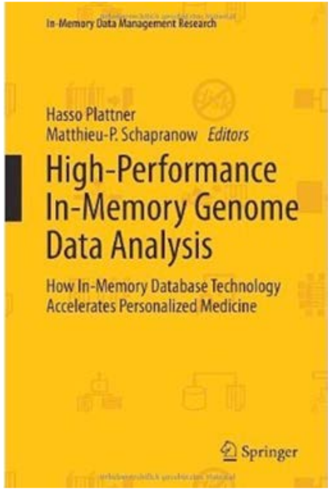
Contacts

- Florian Borchert
- Dr. Matthieu-P. Schapranow

✉<First Name>.<Last Name>@hpi.de

Digital Health Cluster
Hasso Plattner Institute Campus III
Rudolf-Breitscheid-Str. 187
14482 Potsdam, Germany

we.analyzegenomes.com



In-Memory Data Management Research

Hasso Plattner
Matthieu-P. Schapranow *Editors*

High-Performance In-Memory Genome Data Analysis

How In-Memory Database Technology Accelerates Personalized Medicine

Springer