



Seminar Applied Machine Learning for Digital Health Kick-off

Florian Borchert, Aadil Rasheed, Dr. Matthieu-P. Schapranow

Applied Machine Learning for Digital Health

Summer 2023

Seminar Organization

Administrative Details Time / Dates

- Format: Seminar
- Scope: 4 SWS (6 graded ECTS)
- Dates and time:
 - Tuesdays + Thursdays 01.30pm - 03.00pm and
 - Individual appointments with your personal supervisor.
- Further details are available on the seminar website:

<https://hpi.de/digital-health-cluster/teaching/summer-term-2023/applied-machine-learning-for-digital-health.html>

The screenshot shows the top navigation bar of the Hasso Plattner Institut website. It includes the HPI logo, the name 'Hasso Plattner-Institut', and social media icons for Facebook, Twitter, Instagram, YouTube, RSS, and LinkedIn. A 'Home' link is visible on the right. Below the navigation bar is a hamburger menu icon. The main content area features the seminar title 'Applied Machine Learning for Digital Health' in a large, bold, dark red font. Underneath is a section titled 'General Information' in a smaller, bold, dark red font. This section contains a list of details: teaching staff (Florian Borchert, Aadil Rasheed, Dr. Matthieu-P. Schapranow), location (Campus III, G1-E.15/16), duration (4 Semesterwochenstunden (SWS) 6 ECTS (graded)), participant limit (Max. number of participants defined by the number of provided topics), dates and times (Tue & Thu 1.30pm-3.00pm s.t.), kickoff courses (Thu Apr 20, 2023 @ 1.30pm s.t.), and instructions for topic selection after the kickoff event.

Applied Machine Learning for Digital Health

General Information

- > Teaching staff: Florian Borchert, Aadil Rasheed, Dr. Matthieu-P. Schapranow
- > Location: Campus III, G1-E.15/16
- > 4 Semesterwochenstunden (SWS) 6 ECTS (graded)
- > Limit: Max. number of participants defined by the number of provided topics.
- > Dates & times: Tue & Thu 1.30pm-3.00pm s.t.
- > Kickoff courses: Thu Apr 20, 2023 @ 1.30pm s.t.
- > 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 Kickoff, Apr
20, 2023**

Applied Machine
Learning for Digital
Health, Summer 2023
3

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 Kickoff, Apr
20, 2023**

Applied Machine
Learning for Digital
Health, Summer 2023

4

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 Kickoff, Apr
20, 2023**

Applied Machine
Learning for Digital
Health, Summer 2023
5

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



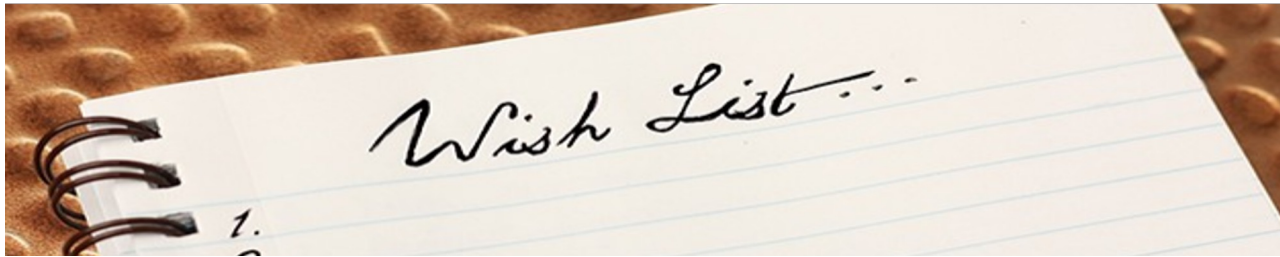
Seminar Kickoff, Apr
20, 2023

Applied Machine
Learning for Digital
Health, Summer 2023
6

Next Steps

Enrollment Process: How to apply for a topic?

- Send prioritized list of top three topics to Aadil Rasheed (Aadil.Rasheed@hpi.de) by **Wed Apr 26, 2023 9am (sharp)**
 - 1st choice: ...
 - 2nd choice: ...
 - 3rd choice: ...
- Assignment of seminar topics: **Wed Apr 26, 2023 by noon**
- Enrollment deadline: **Fri Apr 28, 2023 end of day (via Studienreferat)**



Seminar Kickoff, Apr
20, 2023

Applied Machine
Learning for Digital
Health, Summer 2023
7

Seminar Schedule: Presentations

- **May 23rd (+ May 25th)** Intermediate presentations
 - 10 minutes presentation
 - Introduce your topic, problem/motivation, how you want to solve it
 - Slides due at day of presentation, 9am

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

Seminar Kickoff, Apr
20, 2023

Applied Machine
Learning for Digital
Health, Summer 2023

Seminar Schedule: Paper Writing

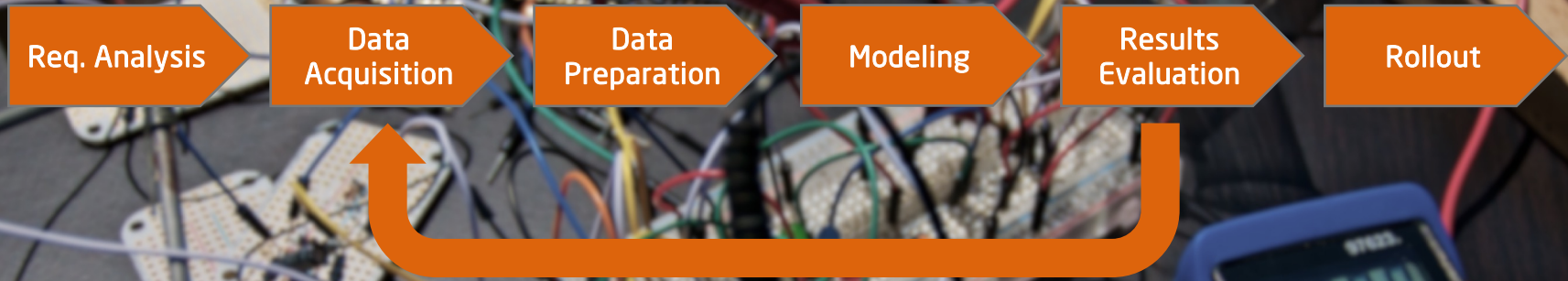
- **July 20th:** Introduction to scientific writing

- **Aug 20th (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

**Seminar Kickoff, Apr
20, 2023**

Applied Machine
Learning for Digital
Health, Summer 2023

Machine Learning for Digital Health Process:



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.

Seminar Kickoff, Apr
20, 2023

Applied Machine
Learning for Digital
Health, Summer 2023
10

Overview of Seminar Topics

- A. Weak Supervision for Detecting Mentions of Genetic Variants in Medical Text
- B. MedNLP-SC Shared Task: Detecting Adverse Drug Events on Social Media
- C. Generating Multilingual Interface Terminologies with Large Language Models
- D. Effect of Feature Drift on Model Quality using MLOps
- E. De-identified Medical Data and Its Impact on Model Performance Metric
- F. Comparison of Automated and Manual Feature Engineering for Training of CPMs

Seminar Kickoff, Apr
20, 2023

Applied Machine
Learning for Digital
Health, Summer 2023

11

A: Weak Supervision for Detecting Mentions of Genetic Variants in Medical Text

- As expert labels for supervised ML are expensive to obtain, we can use „**weak**“ supervision to obtain cheaper, but noisier labels
- Your tasks:
 - Use the **skweak** framework [1] to generate weakly labeled training data
 - Extend our **GGTWEAK** [2] pipeline to handle variants (building on last year's seminar results)
- **Data**: up to 1k manual annotations of variants + much more unlabeled data from GGPONC 2.0
- **Requirements**: Good Python skills, basic ML knowledge, German language might be helpful

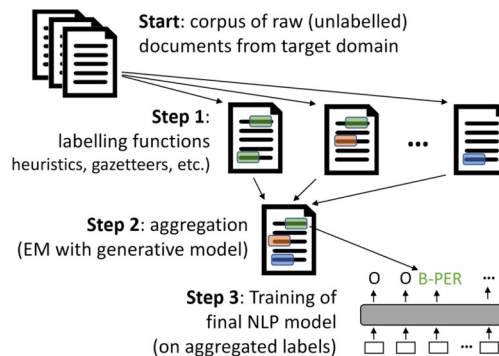
Variant
Translokationen t(1;3)(p36.3;q25) mit einem Fusionsgen

Variant | Transcript Fusion
Gene or Protein
23261 | Gene or Protein
25937 | Gene or Protein
WWTR1-CAMTA1

Variant | Gene Variant, Gain Of Function Variant
Gene or Protein
673 | BRAF-Mutationen

(bislang ausschließlich

Variant | Missense Variant
12 | p.V600E)



Seminar Kickoff, Apr
20, 2023

Applied Machine
Learning for Digital
Health, Summer 2023
12

[1] <https://spacy.io/universe/project/skweak>

[2] https://www.dropbox.com/s/j901c7h0onntvsl/AIME_2023_paper_24.pdf?dl=0

B: MedNLP-SC Shared Task: Detecting Adverse Drug Events on Social Media

- The goal of the MedNLP-SC shared task is to detect **Adverse Drug Events (ADE)** from **tweets** in 4 languages:

<https://sociocom.naist.jp/mednlp-sc/>

- **Your tasks:**

- Participate in the challenge
- Win!

- **Data:** 10k tweets in Japanese, English, German, and French

- **Requirements:** Python, Basic ML knowledge, any of the languages besides English might be helpful

JA アザチオプリンを服用して2ヶ月経ちました。1週間くらいで全身の発疹はなくなり、かゆみもほぼ無くなっていたのですが、麻疹が少し残って怖かったなぁと思います。

EN I've been on Azathioprine for 2 months now, and after about a week the rash all over my body was gone and the itching was almost gone, but I still had a bit of measles and I think it was scary.

DE Ich nehme jetzt seit zwei Monaten Azathioprin, und nach etwa einer Woche war der Ausschlag am ganzen Körper verschwunden und der Juckreiz fast weg, aber ich hatte immer noch ein bisschen Masern, und ich glaube, das war beängstigend.

FR Je prends de l'azathioprine depuis deux mois maintenant, et après environ une semaine, l'éruption cutanée sur tout mon corps avait disparu et les démangeaisons avaient presque disparu, mais j'avais encore un peu de rougeole et je pense que c'était effrayant.

Positive for "Rash" and "Other"

Multi-label classification

Schedule

- March 2023: Dataset release
- (March-June 2023: Dry run)
- June 1, 2023: Registration deadline
- June 24, 2023: Training data (final version) release
- July 10, 2023: Test data release
- July 17, 2023: Result submission
- August 1, 2023: Evaluation result release
- August 1, 2023: Task overview paper release (draft)
- September 1, 2023: Submission due of participant papers (draft)
- November 1, 2023: Camera-ready participant paper due
- December 12-15, 2023: NTCIR-17 Conference in NII, Tokyo (Online presentation will be available)

If you choose the topic, we will register as a team to get access to the full dataset asap.

Seminar Kickoff, Apr 20, 2023

Applied Machine Learning for Digital Health, Summer 2023
13

C: Generating Multilingual Interface Terminologies with Large Language Models

- **Linking entities to IDs** in knowledge bases (UMLS, SNOMED CT) is essential for most downstream tasks
- But most terms in medical terminologies are available in English only!
- **Your tasks:**
 - Use modern **large language models** like GPT-*, BLOOMZ, et al. to generate high-quality aliases in different languages [1]
 - Use these aliases for entity linking and improve over monolingual / cross-lingual baseline
- **Data:** UMLS, annotated Non-English corpora (e.g., Quaero, Mantra GSC)
- **Requirements:** Solid ML + NLP background, Basic understanding of medical ontologies

When emotion and **expression** diverge : The social costs of Parkinson 's disease Patients with Parkinson 's disease are perceived more negatively than their healthy peers , yet it remains unclear what factors contribute to this negative social perception . Based on a cohort of 17 Parkinson 's disease patients and 20 healthy controls , we assessed how naïve raters judge the emotion and emotional intensity displayed in dynamic **facial expressions** as adults with and without Parkinson 's disease watched emotionally evocative films (Experiment 1) , and how age - matched peers naïve to patients ' disease status judge their social desirability along various dimensions from audiovisual stimuli (interview excerpts) recorded after certain films (Experiment 2) . In Experiment 1 , participants with Parkinson 's disease were rated as significantly more **facially expressive** than healthy controls ...

[C001726Z]
Name: Gene Expression
Type: Biologic Function
Qualifiers: expression gene, Gene expression, Expression, Expressed, gene expression

[C0517243]
Name: Facial Expression
Type: Finding
Qualifiers: Facial, observable entity, Facial expression, face expression

Name Count by Language:

Language	Name Count	% of Metathesaurus
ENG	11855838	70.33%
SPA	1806743	10.72%
FRE	443827	2.63%
POR	440312	2.61%
JPN	338265	2.01%
RUS	302797	1.8%
DUT	299344	1.78%
ITA	258724	1.53%
GER	255542	1.52%

Source: <https://aclanthology.org/2021.naacl-main.205/>

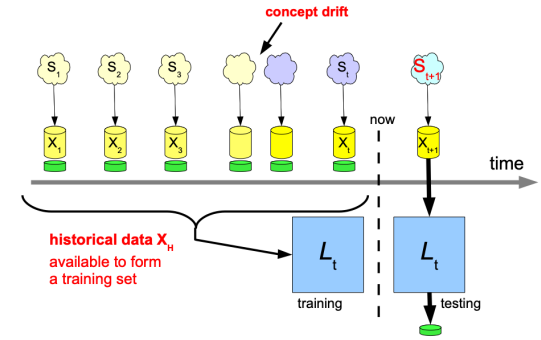
Seminar Kickoff, Apr 20, 2023

Applied Machine Learning for Digital Health, Summer 2023
14

[1] <https://aclanthology.org/2020.multilingualbio-1.3/>

D: Effect of Feature Drift on Model Quality using MLOps

- Changes in data generation process leads to concept drift[1], leading to model **retraining**.
- Modern MLOps technique makes retraining, maintainable and reproducible.
- **Your tasks:**
 - Train different models on drifted datasets
 - Use **MLOps** tools i.e. wandb to version models, analyse performance as the features drift demonstrating goals like maintainability etc.
- **Data:** Scientific Registry of Transplant Recipients(SRTR) dataset [2]
- **Requirements:** Solid ML understanding, good knowledge of Python, SSH, proficiency in working on linux VMs remotely



**Seminar Kickoff, Apr
20, 2023**

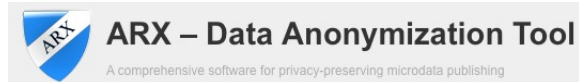
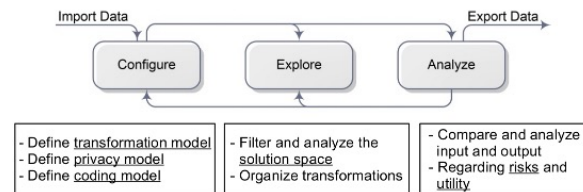
Applied Machine
Learning for Digital
Health, Summer 2023
15

[1] Žliobaitė, Indrė. "Learning under concept drift: an overview." arXiv preprint arXiv:1010.4784 (2010).

[2] <https://www.srtr.org/requesting-srtr-data/saf-data-dictionary/>

E: De-identified Medical Data and Its Impact on Model Performance Metric

- Sharing of data helps reproducibility in ML but **privacy** is main challenge
- De-identification mitigates privacy concern but lead to information loss
- **Your tasks:**
 - Use modern **ARX** tool[1] to create de-identified data using different privacy measures
 - Compare the model performance on transformed datasets vs original dataset
- **Data:** NephroCAGE[2] dataset. Tabular data set with information about transplanted patients
- **Requirements:** ML skills, SSH, proficiency in working on linux VMs remotely



Seminar Kickoff, Apr 20, 2023

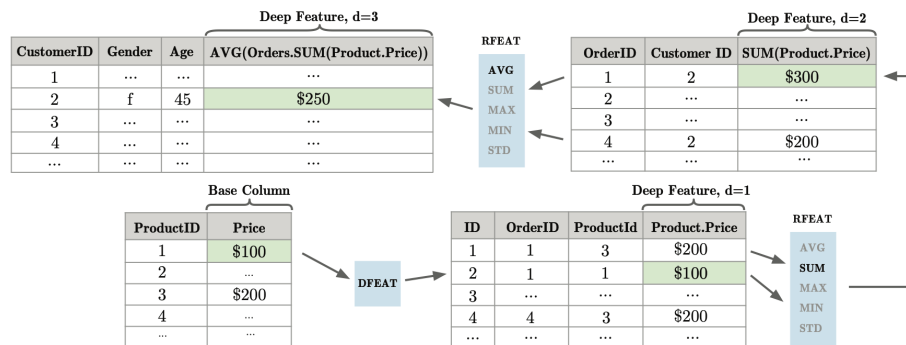
Applied Machine Learning for Digital Health, Summer 2023
16

[1] <https://arx.deidentifier.org>

[2] <https://nephrocage.org>

F: Comparison of Automated and Manual Feature Engineering for Training of CPMs

- Manual feature engineering requires time and expertise
- Deep feature synthesis (DFS) can automate it
- Your tasks:
 - Generate new features using [featuretools\[2\]](#)
 - Compare ML model performance between handpicked and DFS generated feature .
 - Data: SRTR dataset [3]
- Requirements: Solid ML skills, SSH, proficiency in working on linux VMs remotely



**Seminar Kickoff, Apr
20, 2023**

Applied Machine
Learning for Digital
Health, Summer 2023
17

[1] Kanter, James Max, and Kalyan Veeramachaneni. "Deep feature synthesis: Towards automating data science endeavors." 2015 IEEE international conference on data science and advanced analytics (DSAA). IEEE, 2015.

[2] <https://www.featuretools.com>

[3] <https://www.srtr.org/requesting-srtr-data/saf-data-dictionary/>

Do Not Forget to Enroll for the Lecture!



We want you!



**Seminar Kickoff, Apr
20, 2023**

Applied Machine
Learning for Digital
Health, Summer 2023
18

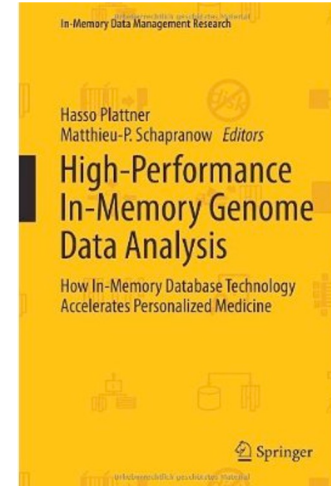
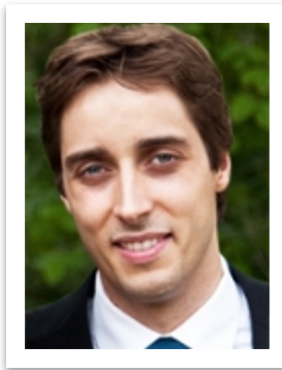
Contacts

- Florian Borchert
- Aadil Rasheed
- 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
🐦 @AnalyzeGenomes



**Seminar Kickoff, Apr
20, 2023**

Applied Machine
Learning for Digital
Health, Summer 2023
19