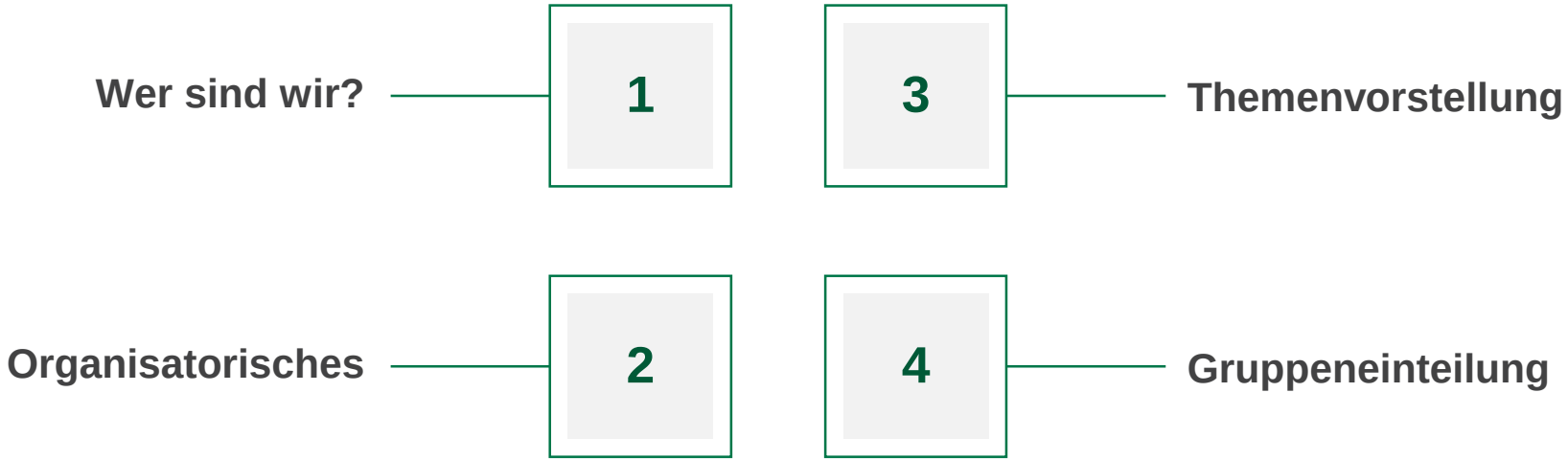
Seminar Real-World Challenges in Data Science und Analytics

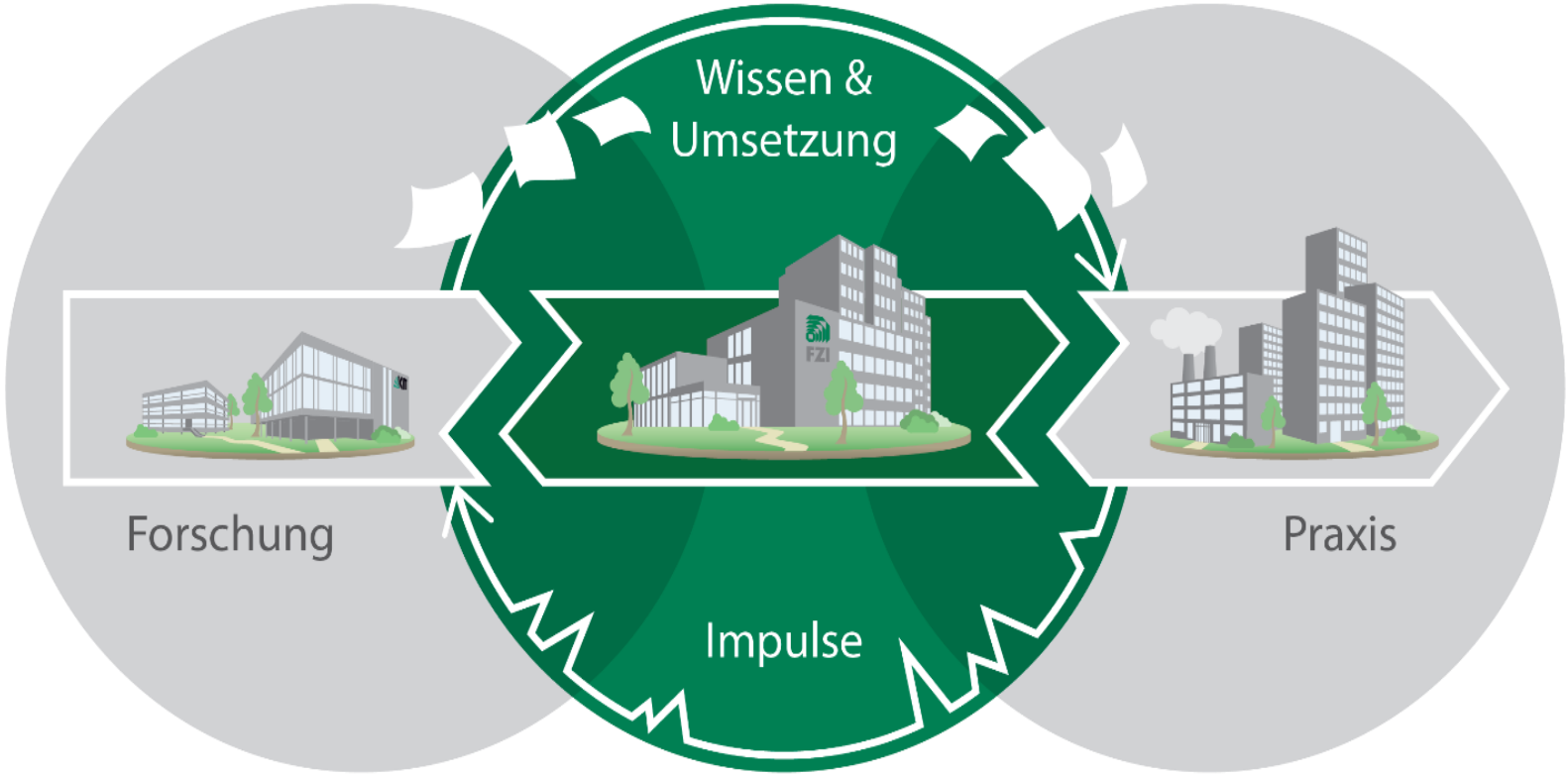
Kick Off

Agenda


Wer sind wir?
1
3
Themenvorstellung
Organisatorisches
2
4
Gruppeneinteilung

01 Wer sind wir?

Was ist das FZI?



FZI-Abteilung Wissensmanagement

Ziel

Auswertung von **großen Datenmengen** aus **heterogenen Quellen** mit intelligenten Verarbeitungspipelines unter Einsatz von maschinellen Lernverfahren zur **Entscheidungsunterstützung** und **-automatisierung**.

Technik

Maschinelles Lernen

- Natural Language Processing
- Explainable AI
- Neuro-symbolic AI
- Computer Vision
- Automated Machine Learning
- Interactive Machine Learning
- Reinforcement Learning

Datenanbindung und -visualisierung

- Stream Processing
- Web Application
- Data Parsing / Extraction

Datengrundlage

- Zeitreihen
- Text
- Ton
- Bilder
- Wissensgraphen

Ansprechpartner

FZI

FZI

Dr. Steffen Thoma,

Abteilungsleitung

Felix Hertlein,

Computervision

Jacqueline Höllig,

Explainable AI

Jin Liu

NLP

Lucas Cazzonelli,

Online Learning

Anastasia Slobodyanik

Walter Laurito,

NLP

AIFB

Prof. Dr. York Sure-Vetter

Direktor FZI, Professor,
Web Science, AIFB

ByteFabrik.AI

Dr. Philipp Zehnder

Industrial Data Analytics

Dr. Dominik Riemer

Industrial Data Analytics

02 Organisatorisches

Idee des Seminars	Die Studierenden bearbeiten in Gruppen (≤ 3 Pers.) ein Themen mit bereitgestellten Daten. Hierbei wird der typische Ablauf eines **Data Science Projektes** abgebildet.
Ergebnis der Arbeit	Auf Grundlage der jeweils vorgegebenen Datenbasis soll eine **einsatz- und auslieferungsfähige Softwarlösung** entwickelt werden.
Lernziel ist…	…den **strukturierten** Umgang mit den Herausforderungen **realer Daten** und den Möglichkeiten **moderner Technologien** erlernen, um Potentiale für bessere unternehmerische Entscheidungen zu heben.

Organisatorisches

- Die Case Challenge bietet die Möglichkeit zur „freien“ Gestaltung
  - des **Lösungswegs** und
  - des konkreten **Deliverables**.

- Was wird bewertet?
  - Die entwickelte Lösung, insb. Der **Lösungsweg**
  - Schriftliche Ausarbeitung
  - Präsentation

Wissenschaftliches Arbeiten

Was macht eine gute wissenschaftliche Arbeit aus?

Objektivität

- Unvoreingenommenes wissenschaftliches Vorgehen
- eine sachliche Argumentation und neutrale Darstellung der Ergebnisse

Nachvollziehbarkeit

- Nachvollziehbare Aussagen und Ergebnisse
- Richtige Geltungsbereiche

Relevanz

- Informationswert
- Neue Verfahren, die helfen, Probleme

Verständlichkeit

- Definition relevanter Begriffe
- Klare, präzise Sprache
- Wahrnehmungshilfen

Überprüfbarkeit

- Belegen von Aussagen
- Vollständige Quellenangaben

Ehrlichkeit

- Fremdwissen & Eigenleistung
- Primärquellen
- Künstliches aufblähen von Quellenangaben

Leitfaden: https://www.aifb.kit.edu/images/3/39/AIFB_Leitfaden_Abschlussarbeiten.pdf

Zeitplan

Datum	Uhrzeit	Ort	Gegenstand
06. November 2024	15:00 – 17:00	Raum Hamburg, FZI	Kick-Off & initiales Gruppenmeeting
11. Dezember 2023	15:00 – 17:00	Raum ?, FZI	Zwischenpräsentation 5 min Präsentation
09. Februar 2024	23:59 Uhr	(elektronisch)	Abgabe der Ausarbeitung
29. Januar 2024	16:00 – 18:00	Raum ?, FZI	Abschlusspräsentation 10 min Präsentation

Ablauf und Formales

- Initiales Gruppenmeeting mit Betreuer (im Anschluss & meldet euch bitte bei euren zugeteilten Betreuern)
- Anmeldung: https://campus.studium.kit.edu/exams/registration.php (2513315)
- Treffen mit Gruppenbetreuer nach Vereinbarung
- Midterm Meeting
  - **Ziel:** Vorstellung Halbzeitstand, bisherige Erkenntnisse, Herausforderungen und weiteres Vorgehen
  - Präsentation (5 Min.) Vorstellung der Gruppe, Thema und bisherige Erkenntnisse
  - Diskussion (5 Min.)
    - Austausch Betreuer / Seminarteilnehmer
- Abschlusspräsentation (10 Min. Vortrag + 10 Min Diskussion)
- Deliverable:
  - Umfang: 10 – 15 Seiten (Inhalt)
  - Inhalt: Dokumentation der Projektarbeit

UniCluster

- Für die Laufzeit des Seminars (bis 01.03)
- Registrierung notwendig: https://bwidm.scc.kit.edu
- Danach können Sie sich auf einem Login-Knoten des Clusters mit Ihrem KIT-Account einloggen (uc2.scc.kit.edu).
- Weitere Informationen:
  - https://wiki.bwhpc.de/e/BwUniCluster_2.0_User_Access
  - http://www.scc.kit.edu/dienste/hpc.php

03 Themenvorstellung

Themenübersicht

Gruppe	Thema	Ansprechpartner
1	Integration von Python Jupyter Notebooks in eine Webanwendung	Philipp Zehnder (philipp.zehnder@bytefabrik.ai) Dominik Riemer (dominik.riemer@bytefabrik.ai)
2	Explaining Text Classification Models using LLM generated counterfactuals	Anastasia Slobodyanik (slobodyanik@fzi.de)
3	Real-World Fact Checking	Jin Liu (Liu2@fzi.de)
4	Forward-Forward Algorithmus für Online Learning	Lucas Cazzonelli (cazzonelli@fzi.de)
5	Explainable Machine Learning for Time Series	Jacqueline Höllig (hoellig@fzi.de)
6	ChemChat	Walter Laurit (laurito@fzi.de)

Bytefabrik.AI

BYTEFABRIK.AI

Seminar Real-World Challenges in Data Science und Analytics

Bytefabrik.AI

Bytefabrik.AI: Experten an der Schnittstelle zwischen OT, IT und KI

Unternehmen

IIoT/KI-Analyselösungen für automatisierte Anlagen

Spin-Off des Karlsruher Instituts für Technologie (KIT)

Open-Core-Geschäftsmodell

Produkte

Bytefabrik Manufacturing Insights
Analyselösung für automatisierte Anlagen

Bytefabrik IIoT-Anwendungsplattform
Beschleunigte Entwicklung von KI-Anwendungen für die Industrie

Apache StreamPipes
Open Source

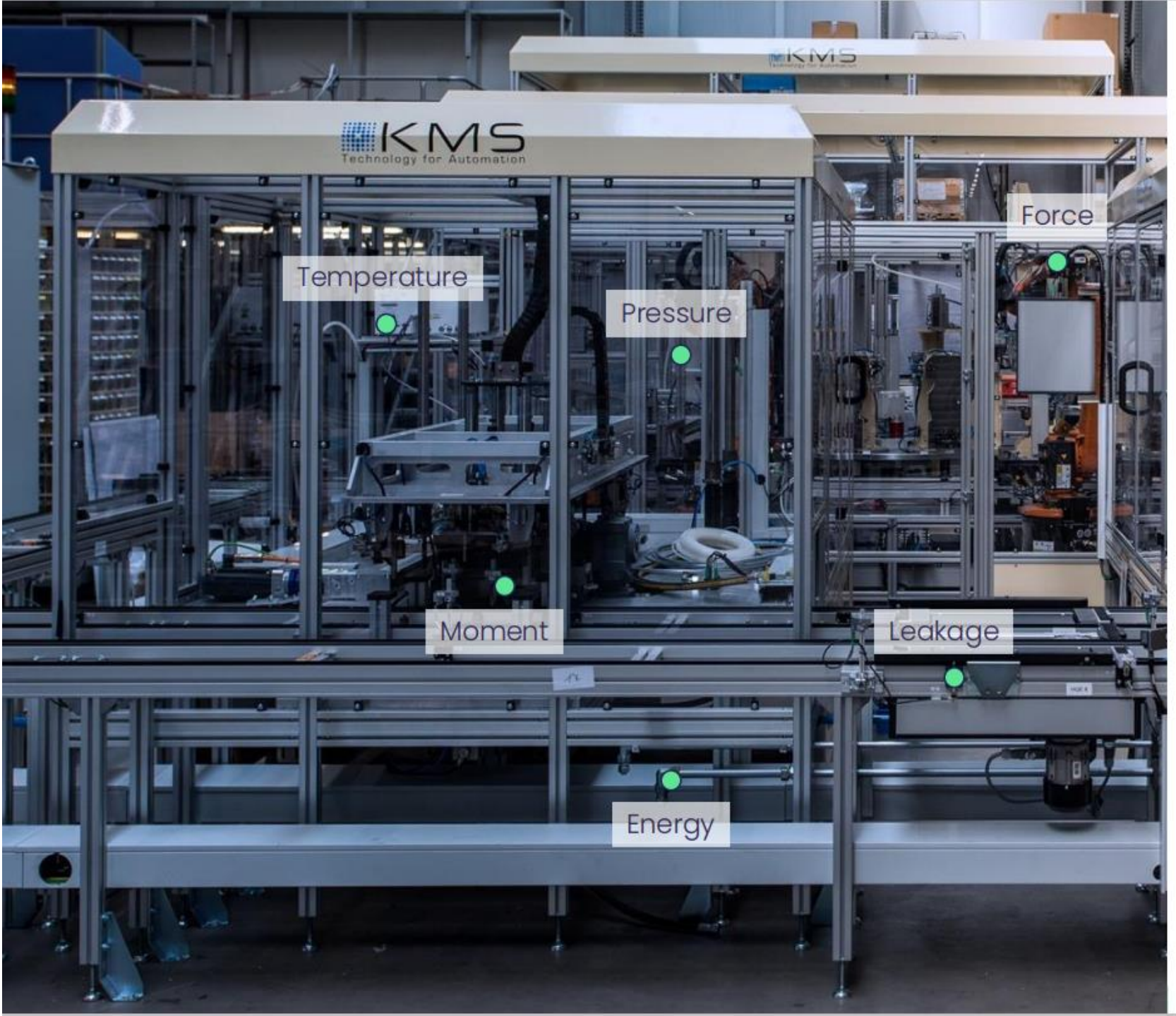
Dienstleistungen

Entwicklung

Beratung

Training
IT/OT-Integration, KI-Anwendungen, Apache StreamPipes

Bytefabrik.AI



CONTEXT

Automated production lines in manufacturing and intralogistics

Bytefabrik.AI

PROBLEM

Many factors cause **reduced performance** and **unforeseen failures** of automated production lines.

Root cause analysis is often based on experience and gut feeling.

Bytefabrik.AI

SOLUTION

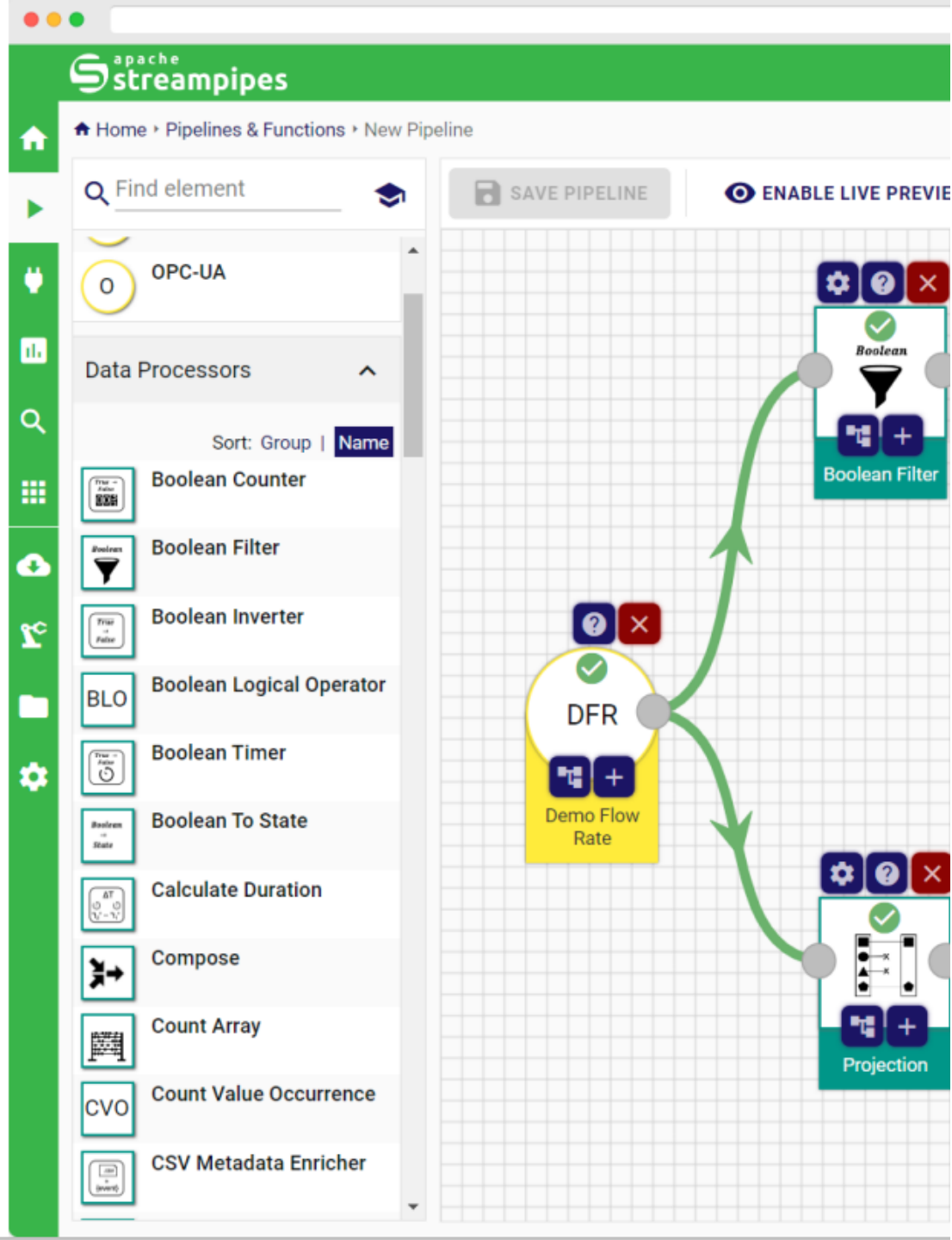
Bytefabrik.AI feels the pulse of your production line - through **AI-based evaluation of control and sensor data**, you get a **360-degree analysis of all processes and process parameters**.

You save costs through **higher plant efficiency**, **lower maintenance costs** and **consistent product quality.**

Live Data from PLCs

OPEN CORE

apache streampipes



Thema: Integration von Python Jupyter Notebooks in eine Webanwendung

Seminarinhalt: Integration von **Jupyter Notebooks** in eine Webanwendung

- **Fokus**: Entwicklung einer **Plug-and-Play-Lösung** zur Verwaltung und Bearbeitung von Python-Programmen im Browser
- Ziel: **Nahtlose Nutzung der Python-Schnittstelle** für IIoT-Daten direkt im Web

Praktische Umsetzung:

- **Technische Integration von Jupyter Notebooks** für eine webbasierte, benutzerfreundliche Python-Entwicklungsumgebung
- Nutzer können eigene **KI-Modelle entwickeln und ausführen**

Ergebnisse: Studierende entwickeln eine Lösung, die Jupyter Notebooks mit IIoT-Daten verbindet, um eine komfortable und integrierte Plattform für datengetriebene Anwendungen und KI-Entwicklung bereitzustellen.

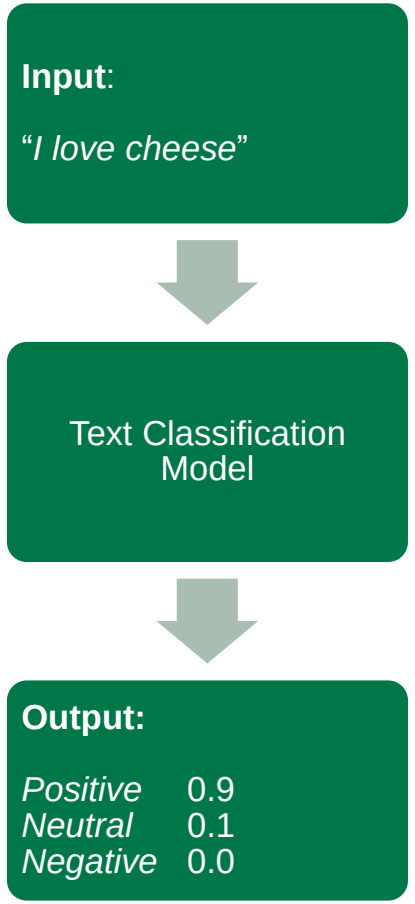
Explaining Text Classification Models Using LLM-generated Counterfactuals

Text Classification

- Assign a label or class to a given text or a relationship between two given texts

- Sentiment Analysis (SA)
  - Positive, Negative, (Neutral)

- Natural Language Inference (NLI)
  - Entailment, Contraction, Neutral



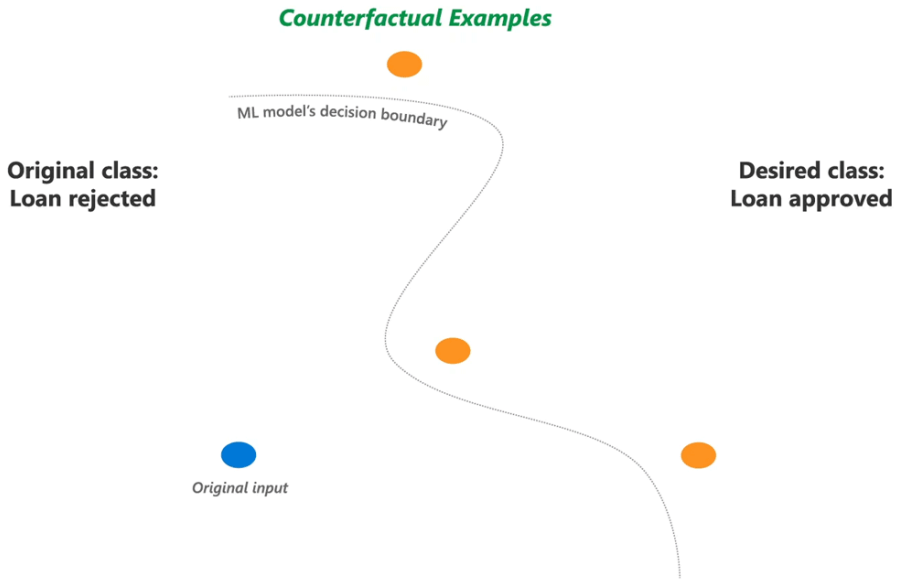
Natural Language Inference

Determine whether a "hypothesis" is true (entailment), false (contradiction), or undetermined (neutral) given a "premise".

Premise	Hypothesis	Label
A man inspects the uniform of a figure in some East Asian country.	The man is sleeping.	contradiction
An older and younger man smiling.	Two men are smiling and laughing at the cats playing on the floor.	neutral
A soccer game with multiple males playing.	Some men are playing a sport.	entailment

Counterfactual Explanations

Explain predictions of individual instances by finding the smallest change to the feature values that changes the prediction to a predefined output

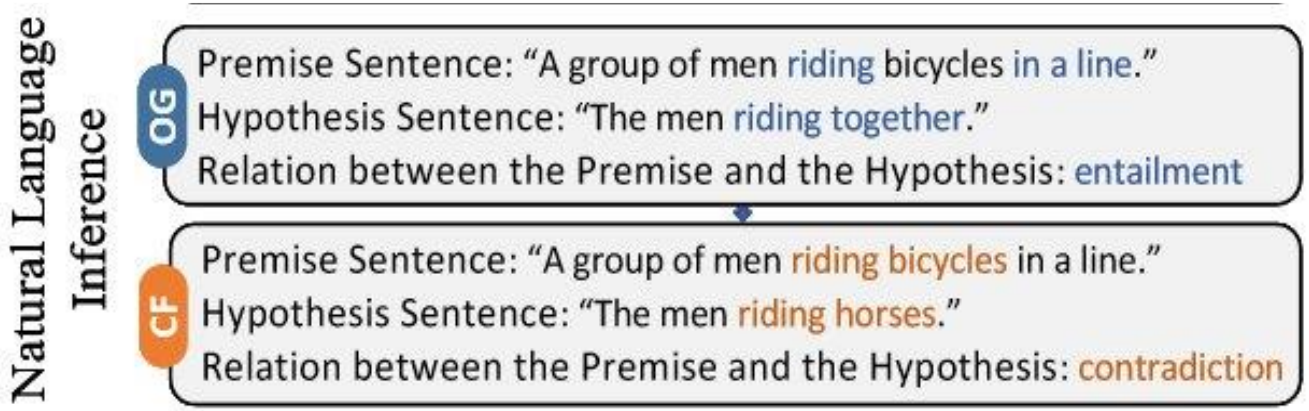


Task: LLM-Generated Counterfactuals for NLI

- Literature review
- Chose a pretrained model and a dataset
- Implement a baseline
  - Existing approaches
  - Heuristics
- Prompt a LLM to
  - Change the premise
  - Change the hypothesis
  - Change both
- Evaluate results

Hugging Face

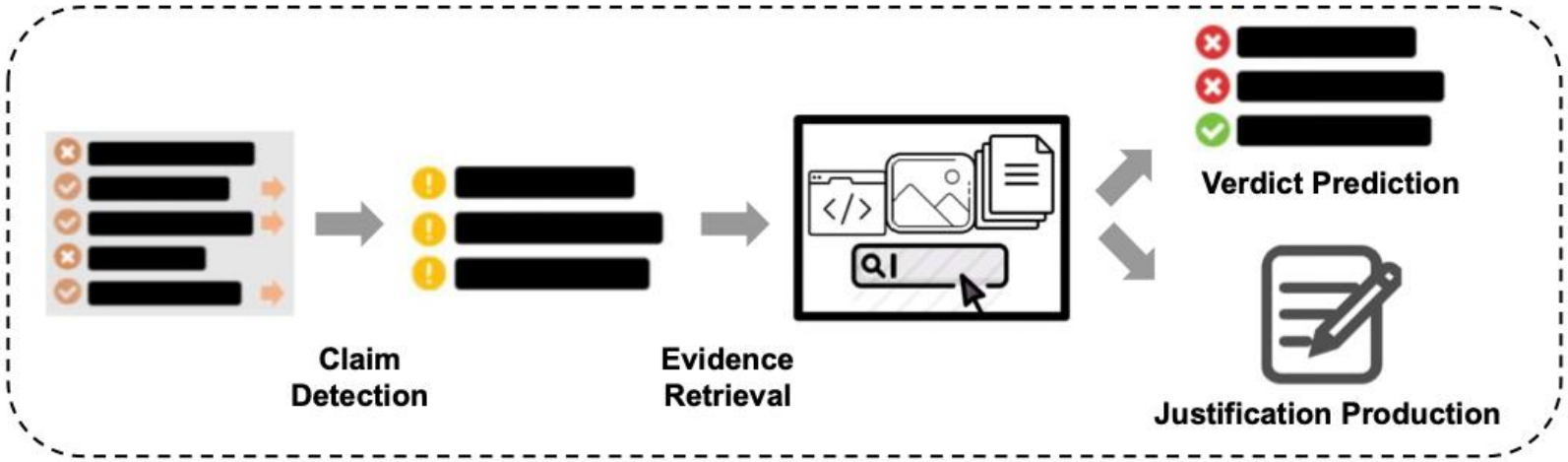
ChatGPT



Real-World Fact-Checking

- Automated fact-checking
  - Knowledge-intensive NLP tasks
  - Information retrieval
  - Natural language inference (reasoning): (Premise, Hypothesis) Verdict (Supported, Refuted, Not Enough Information)

- Pipeline [1]



Real-World Fact-Checking – AVeriTeC [2]

- Tasks:
  - Question Generation with LLM
  - Evidence Retrieval
    - Given the claim and corresponding question, retrieve relevant documents with Google Search API
    - Question-Answering with LLM
  - Claim Verification

Claim: Donald Trump has kept his promises to voters.
Claim type: Event/Property Claim
Speaker: None
Claim date: 24-8-2020

Question 1: What promises did Donald Trump make to voters?
Answer 1 (Extractive & Abstractive): During the 2016 campaign, Donald Trump made more than 280 promises, though many were contradictory or just uttered in a single campaign event. By 2020 Trump had made a number of promises, 6 of which he had not fulfilled, including ...
Question 2: Of the promises Donald Trump made, did he fulfil any of them?
Answer 2 (Boolean): Yes.
Question 3: Has President Donald Trump kept his campaign promises to voters?
Answer 3 (Abstractive): President Trump has only kept a few of his promises.

Verification: Conflicting Evidence/Cherrypicking
Justification: QA pairs state promises kept and not kept. Claim does not state he kept all promises.

References

1. Guo, Z. et al., A Survey on Automated Fact-Checking, TACL 2022
2. Schlichtkrull, M. et al., AVeriTeC: A Dataset for Real-world Claim Verification with Evidence from the Web. NeurIPS 2024

Forward-Forward Algorithmus für Online Learning

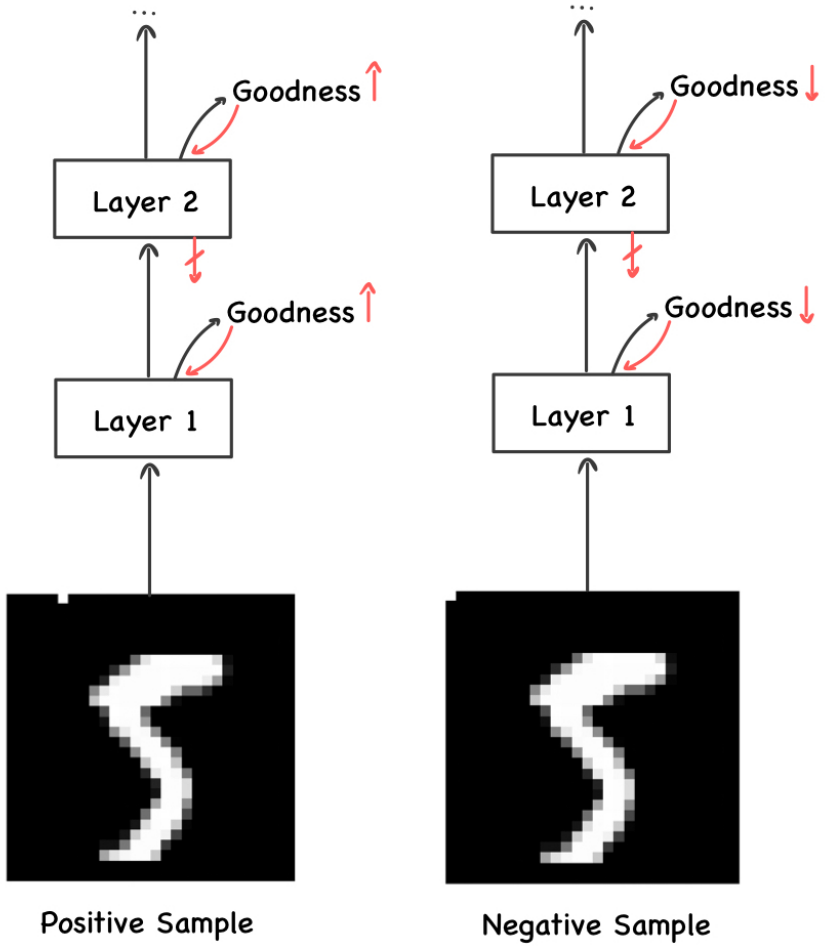
- Neuartiges, biologisch plausibles Lernverfahren für neuronale Netze
- Potentiell besonders für Online Learning geeignet, da genaue Anpassung des Modells an Einzelbeispiele möglich sind

Ziele:

- Implementierung des Forward-Forward Algorithmus für Online Learning
- Vergleichende Evaluation mit Backpropagation und ggf. weiteren Lernverfahren

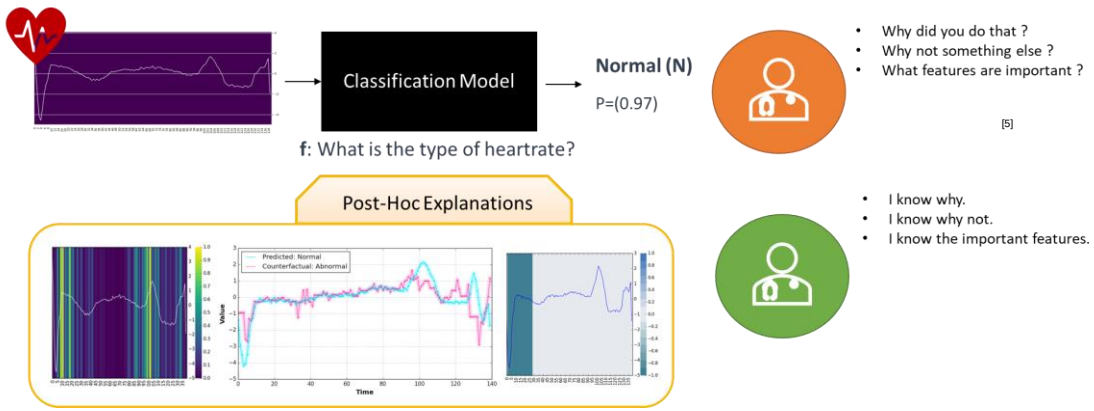
Erwarteter Nutzen:

- Bessere Anpassungsfähigkeit und Effizienz bei dynamischen Datenströmen

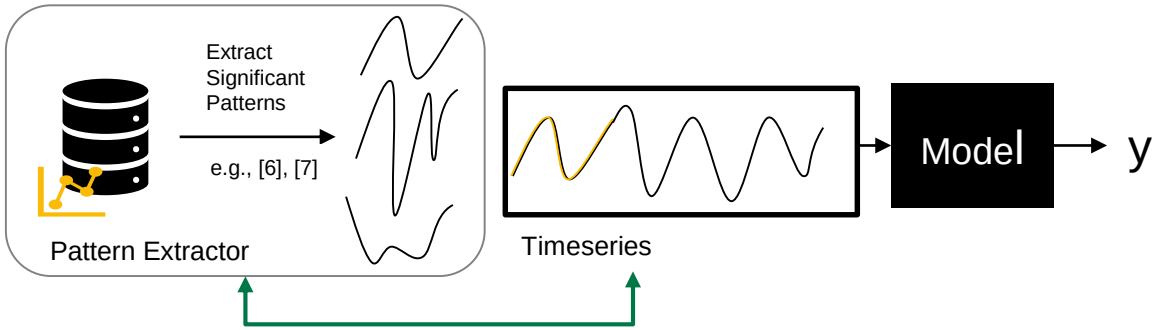


Hinton, 2022

Explainability for Time Series Classification and (Regression)



Subsequence-Based Explainers for Time Series Models



Which parts of the model input are significant for the prediction ?

Can subsequence and patter mining algorithms be used to find prediction Anchors [2]?

1. Research Existing Algorithms
2. Implement Algorithm.
3. Evaluate with the help of [3].

[2] https://christophm.github.io/interpretable-ml-book/anchors.html

[3] Höllig, J., Thoma, S., and Grimm, F., “XTSC-Bench: Quantitative Benchmarking for Explainers on Time Series Classification”, arXiv e-prints, 2023.

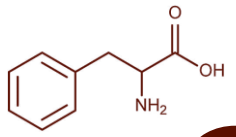
[1]

[4] Raykar, Vikas C., et al. "TsSHAP: Robust model agnostic feature-based explainability for time series forecasting." arXiv preprint arXiv:2303.12316 (2023)..

[5] Spinnato, Francesco, et al. "Understanding Any Time Series Classifier with a Subsequence-based Explainer." ACM Transactions on Knowledge Discovery from Data (2023).

[6] Grabocka, Josif, et al. "Learning time-series shapelets." Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014.

[7] Bostrom, Aaron, and Anthony Bagnall. "Binary shapelet transform for multiclass time series classification." Transactions on Large-Scale Data-and Knowledge-Centered Systems XXXII: Special Issue on Big Data Analytics and Knowledge Discovery (2017): 24-46.



ChemChat

- Chatbot to support prioritisation of identified unknown chemicals in water quality studies by retrieving contextual information not directly available

Which chemicals are pesticides?
Are chemicals neurotoxic?

6-chloro-N-ethyl-N'-(1-methylethyl)-1,3,5-triazine-2,4-diamine

O,O-diethyl O-(3,5,6-trichloro-2-pyridyl) phosphorothioate

N-(4-hydroxyphenyl)acetamide

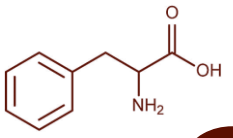
… often > 1K

Public Databases

PubChem

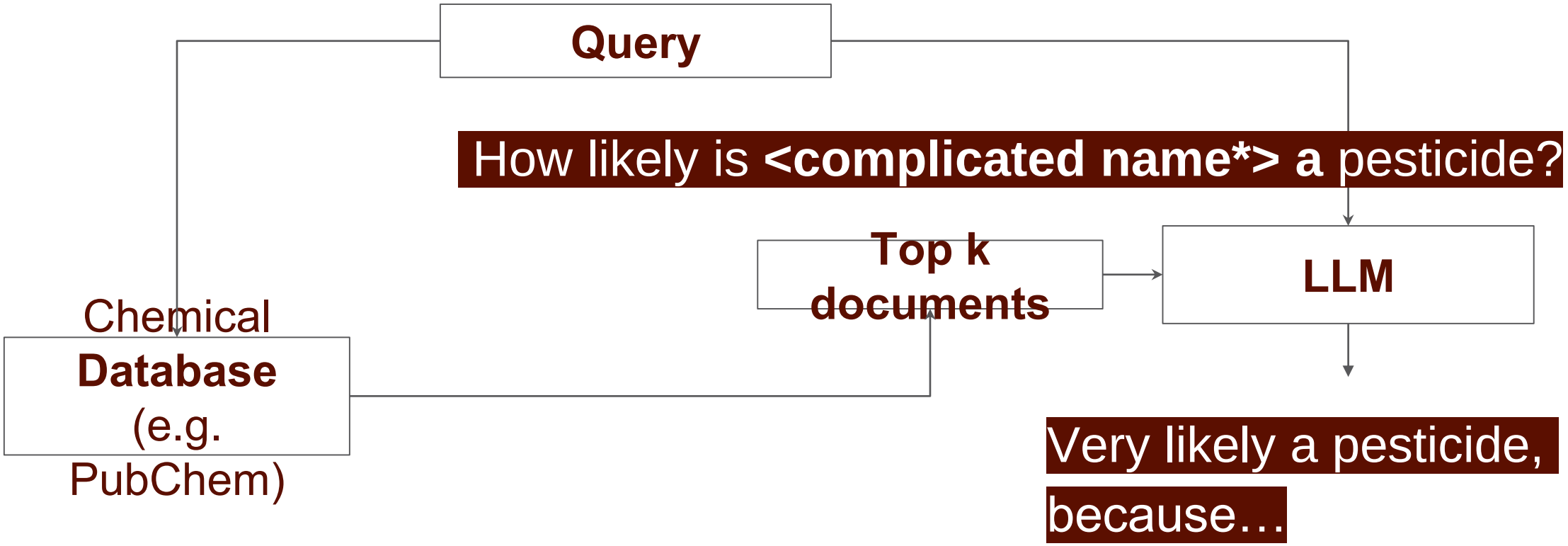
ChEMBL

ChemSpider
Search and share chemistry

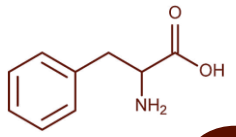


ChemChat

- Chatbot concept



* [2-[2-(2,6-dichloroanilino)phenyl]acetic acid



ChemChat

FZI

- **Challenges:**
  - Database can be quite big
  - LLM can still say false things
  - …
- **Tools we might use:**
  - Python
  - Haystack
  - HuggingFace
  - PubChem

FZI-Abteilung Wissensmanagement

Ziel

Auswertung von **großen Datenmengen** aus **heterogenen Quellen** mit intelligenten Verarbeitungspipelines unter Einsatz von maschinellen Lernverfahren zur **Entscheidungsunterstützung** und **-automatisierung**.

- ChemChat
- Real-World Fact Checking
- Explaining Text Classification Models using LLM generated counterfactuals
- Explainers for Time Series Classification
- Integration von Jupyter Notebooks in eine Webanwendung
- Forward-Forward Algorithmus für Online Learning

Maschinelles Lernen

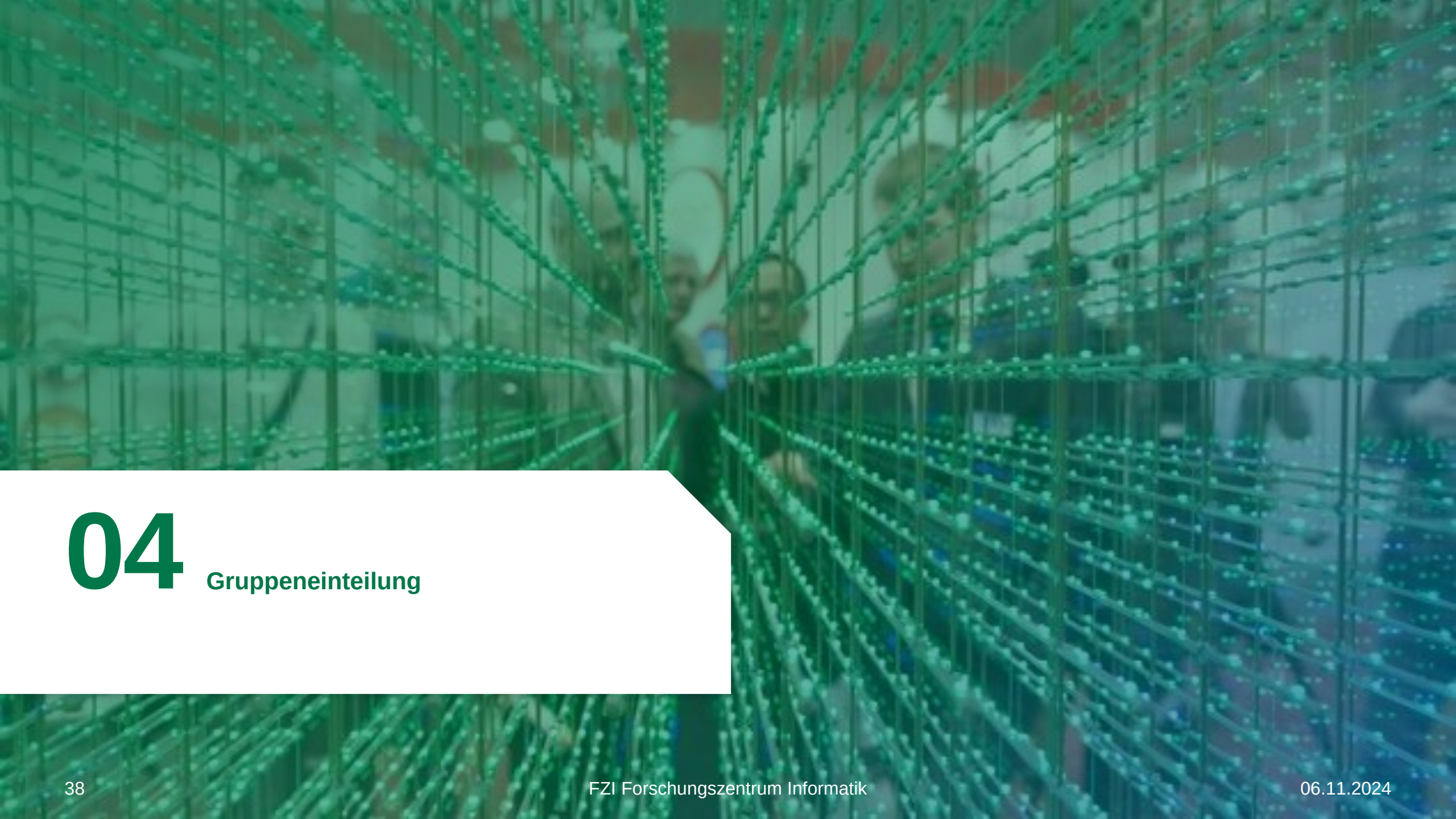
- Natural Language Processing
- Explainable AI
- Neuro-symbolic AI
- Computer Vision
- Automated Machine Learning
- Interactive Machine Learning
- Reinforcement Learning

Datenanbindung und -visualisierung

- Stream Processing
- Web Application
- Data Parsing / Extraction

Datengrundlage

- Zeitreihen
- Text
- Ton
- Bilder
- Wissensgraphen



04 Gruppeneinteilung

Themenübersicht

Gruppe	Thema	Ansprechpartner	Gruppe
1	Integration von Python Jupyter Notebooks in eine Webanwendung	Philipp Zehnder (philipp.zehnder@bytefabrik.ai) Dominik Riemer (dominik.riemer@bytefabrik.ai)	Felix Marschall
2	Explaining Text Classification Models using LLM generated counterfactuals	Anastasia Slobodyanik (slobodyanik@fzi.de)	
3	Real-World Fact Checking	Jin Liu (Liu2@fzi.de)	
4	Forward-Forward Algorithmus für Online Learning	Lucas Cazzonelli (cazzonelli@fzi.de)	Bilal Akdag Tim Fridtjof Tenning
5	Explainable Machine Learning for Time Series	Jacqueline Höllig (hoellig@fzi.de)	Dominik Müller Fabian Wylczoch
6	ChemChat	Walter Laurito (laurito@fzi.de)	Jasper Richter Jeremias Hohner

Zeitplan

Datum	Uhrzeit	Ort	Gegenstand
06. November 2024	15:00 – 17:00	Raum Hamburg, FZI	Kick-Off & initiales Gruppenmeeting
11. Dezember 2023	15:00 – 17:00	Raum ?, FZI	Zwischenpräsentation 5 min Präsentation
09. Februar 2024	23:59 Uhr	(elektronisch)	Abgabe der Ausarbeitung
29. Januar 2024	16:00 – 18:00	Raum ?, FZI	Abschlusspräsentation 10 min Präsentation


VIELEN
DANK