FROM DATA TO VALUE

KNOW Center

FFG Forschung wirkt.

Bundesministerium
Klimaschutz, Umwelt,
Energie, Mobilität,
Innovation und Technologie

SINUS Projekt

Machine Learning Modell

Philipp Danninger
pdanninger@know-center.at

GL_Salzburg 2022

05.07.2022

Agenda

AUFGABE
Das Machine Learning Problem

BAUSTEINE
Die Daten

ANSATZ
Supervised Learning

ZUSAMMENFASSUNG
SINUS ML-Modell

AUFGABE

Das Machine Learning Problem

AUFGABE

Das Machine Learning Problem

- **Aufgabe d. Modells:** Vorhersage von **Risiko** für ungeschützte Verkehrsteilnehmer (= vulnerable road users, **VRUs**)

- **Risiko:** Dokumentierter **Verkehrsunfall** in unmittelbarer Nähe einer geografischen Einheit (= Straße)

- **Vorhersage:** Identifikation von risikosteigernder Faktoren für VRUs (Literaturrecherche)

- **Herausforderung:** Kombination verschiedener Datensätze auf räumlicher und zeitlicher Basis

AUFGABE

Unfalldaten und risikobehaftete Bereiche

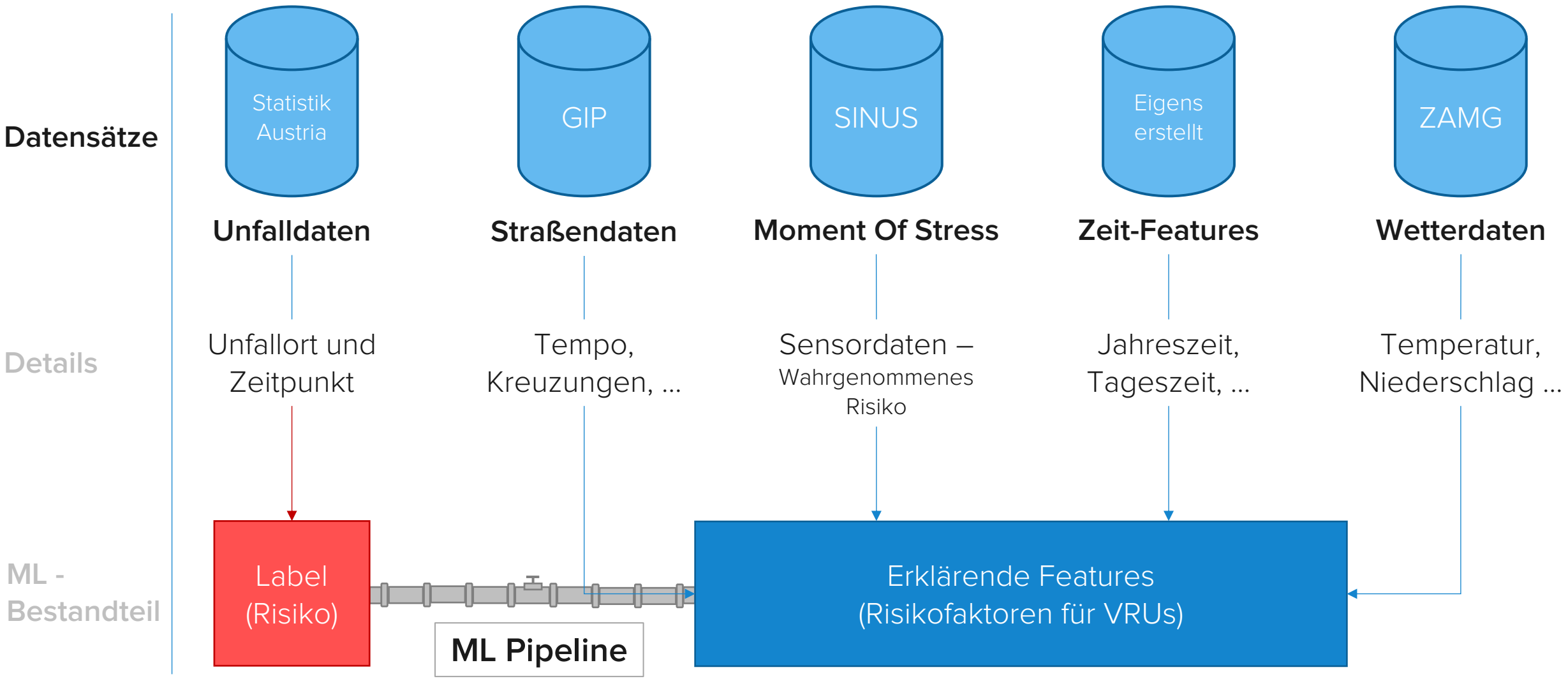
Unfalldaten in der Stadt Salzburg für verschiedene Orte und Zeitpunkte

A: Die fünf dem Unfall am nächsten gelegenen Straßensegmente
B: Als risiko-behaftetes Straßensgement

BAUSTEINE

Die Daten

Zur Verfügung stehende Datensätze



ANSATZ

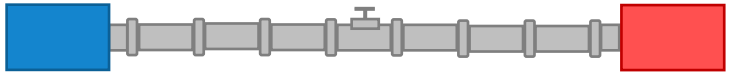
Supervised learning

Was ist Supervised Learning?

- ML: *Durch das **Erkennen von Mustern** in vorliegenden Datenbeständen sind IT-Systeme in der Lage, eigenständig Lösungen für Probleme zu finden.*

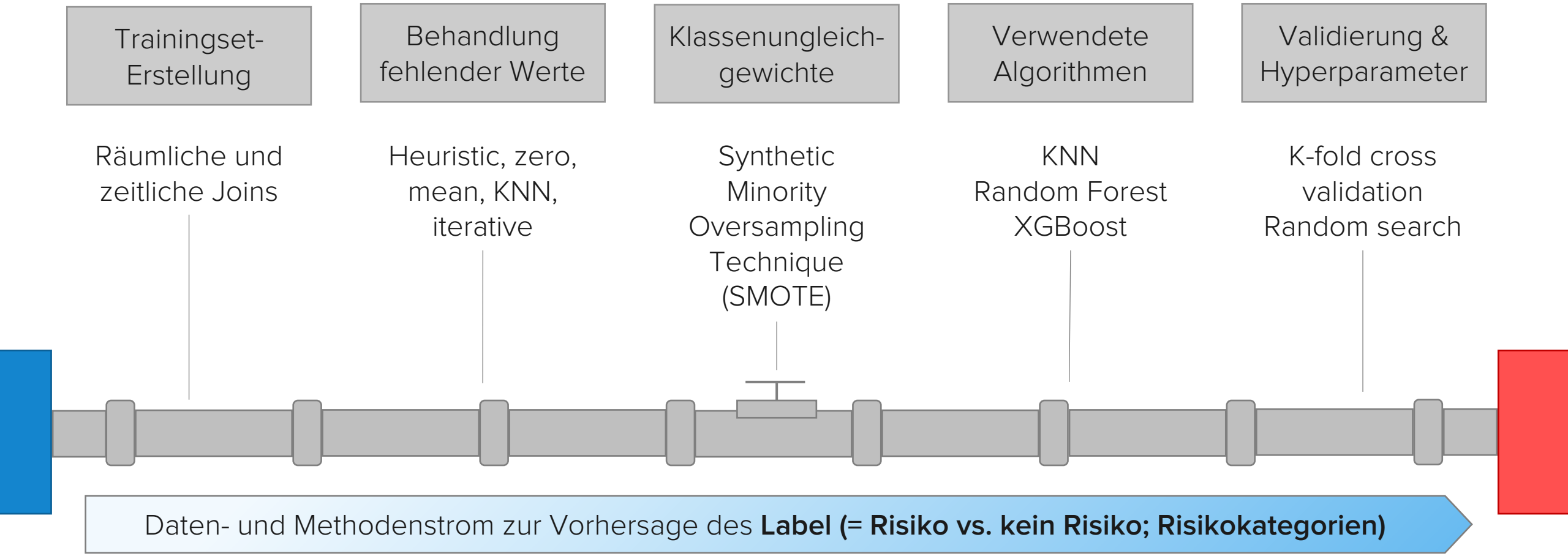
- Daten besitzen ein „**Label**“
  - „Risiko“ vs. kein „Risiko“
  - „Unfall nahe“ vs. „Unfall entfernt“
  - Mehrere Risiko Kategorien (kein, leichtes, mittleres, hohes Risiko)

- Erstellung eines Modells
  - **Risiko-Label** = **Erlaubte Geschwindigkeit + Straße + Kreuzung? + Zeitpunkt + … + Risikofaktor$_n$**

- Modell lernt Muster zwischen den Labels (= **Training**)
  - Machine Learning Pipeline 

- Vom Modell noch ungesehene Daten werden von trainiertem Modell „gelabelt“ (= **Vorhersage**)

Supervised Learning – ML Pipeline



Supervised Learning - Resultate

Run nr.	Algorithm	Imputation	Sampling	Accuracy	Precision	Recall	F1 Score	Roc_Auc	
1	RF	Zero	No sampling	0,80853	0,78550	0,72186	0,75234	0,87060	
2	KNN	Zero	No sampling	0,73924	0,56047	0,67917	0,61414	0,80797	
3	XGBoost	Zero	No sampling	0,79774	0,75324	0,71549	0,73388	0,86808	
4	RF	KNN	No sampling	0,80980	0,78214	0,72554	0,75278	0,87059	
5	KNN	KNN	No sampling	0,73758	0,56093	0,67530	0,61282	0,80673	
6	XGBoost	KNN	No sampling	0,79989	0,76430	0,71491	0,73878	0,87011	
7	RF	Iterative	No sampling	0,77295	0,69918	0,69117	0,69515	0,84891	
8	KNN	Iterative	No sampling	0,62597	0,41460	0,49391	0,45079	0,64697	
9	XGBoost	Iterative	No sampling	0,74963	0,66921	0,65955	0,66435	0,82634	
10	RF	Zero	50:50	0,63238	0,01151	0,72249	0,02267	0,89563	
11	KNN	Zero	50:50	0,56251	0,57397	0,43169	0,49277	0,85339	
12	XGBoost	Zero	50:50	0,63252	0,01289	0,70417	0,02531	0,89712	
13	RF	KNN	50:50	0,62976	0,00000	na	na	0,89562	
14	KNN	KNN	50:50	0,62075	0,03248	0,36379	0,05964	0,85396	
15	XGBoost	KNN	50:50	0,63058	0,00290	0,80851	0,00577	0,90162	⇦
16	RF	Iterative	50:50	0,64593	0,12490	0,60599	0,20712	0,90540	⇦
17	KNN	Iterative	50:50	0,56911	0,29587	0,39160	0,33707	0,81783	
18	XGBoost	Iterative	50:50	0,63809	0,07290	0,59122	0,12979	0,88555	

Greater than	0,85
Smaller than	0,5

Bessere Algorithmen, Imputationen und Sampling verbessern Performance

ZUSAMMENFASSUNG

SINUS ML-Modell

Von der Vorbereitung bis zur Vorhersage

- **Verbindung** zwischen unterschiedlichen Datensätzen
- **Bereinigung** und Erstellung von Trainingsets
- **Wahl** eines Supervised Learning Ansatzes
  - Unsupervised Learning getestet
- **Experimente** mit…
  - … fehlenden Werten
  - … Sampling
  - … ML-Algorithmen
- **Einbettung** des Modells in informatives Dashboard