Karlsruher Institut für Technologie

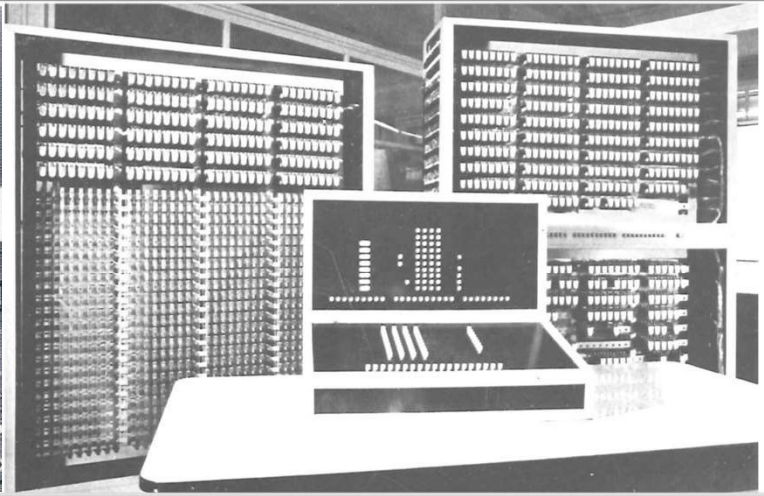
Institut für Technik der Informationsverarbeitung

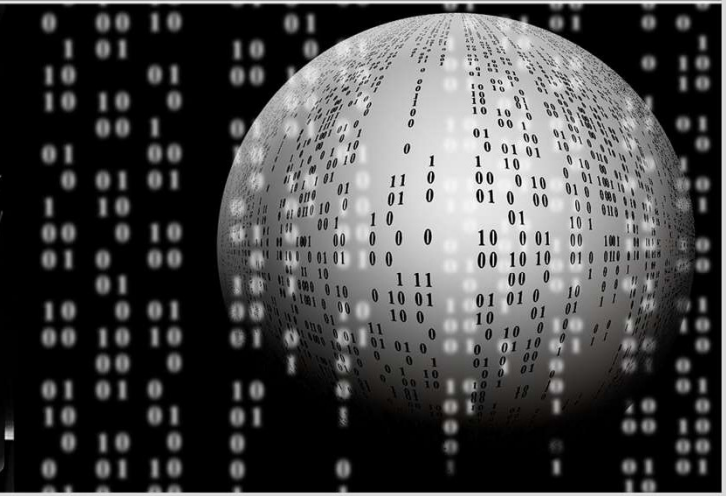
Institutsleitung
Prof. Dr.-Ing. J. Becker
Prof. Dr.-Ing. E. Sax
Prof. Dr. rer. nat. W. Stork

7. Beispiele für Big Data Anwendungen

Informationstechnik II und Automatisierungstechnik

Prof. Dr.-Ing. Eric Sax



Inhalt IT2

7. Beispiele für Big Data Anwendungen

- Mustererkennung auf Fahrzeugdaten
- Anomalieerkennung im Fahrzeug
- Rekonstruktion und Klassifikation von Öl-Daten

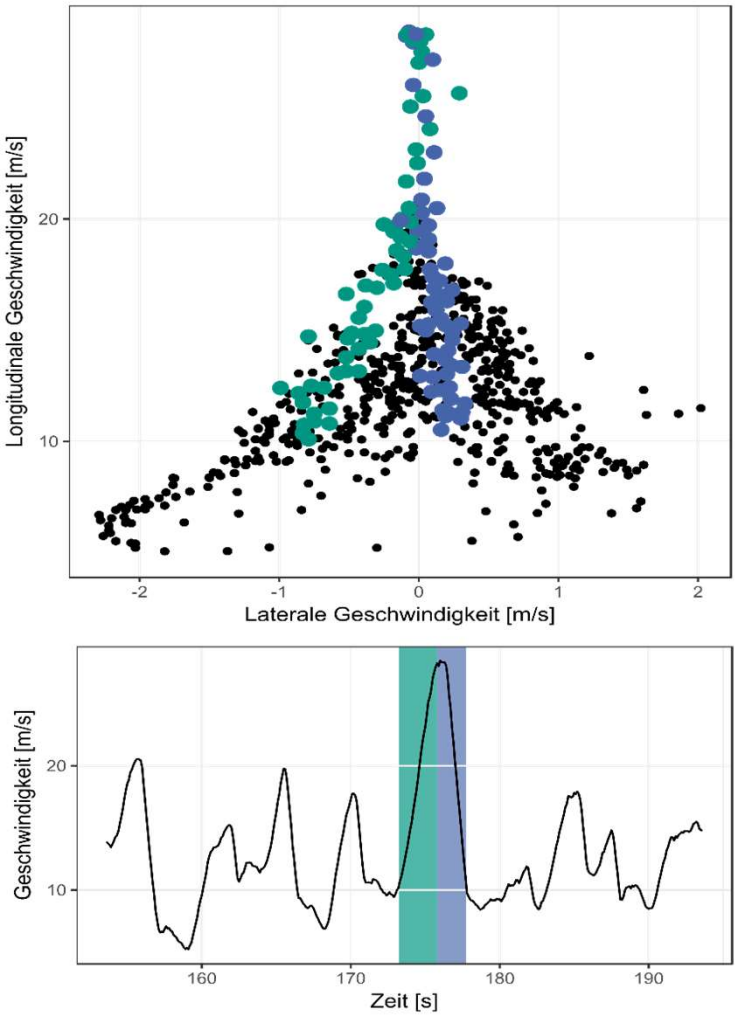
8. Cyber Security, Datenschutz

- Definition, Begriffe
- Angreifertypen- und ziele
- Schutzziele
- Kryptographie
  - Symmetrische Verschlüsselung
  - Asymmetrische Verschlüsselung
- Datenschutz

Mustererkennung in Zeitreihen

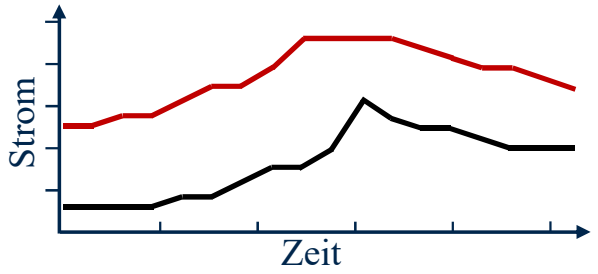
- Im modernen Automobil werden nahezu alle Fahrzeugkomponenten von insgesamt ca. 100 Sensoren und Steuergeräten überwacht
- Jede mögliche Fahrsituation bildet sich in einem bestimmten Signalverlauf ab:
  - Bremsmanöver
  - Fahrweise des Fahrers

Wie können diese Signalverläufe in großen Datensätzen gefunden und für eine weitere Analyse nutzbar gemacht werden?

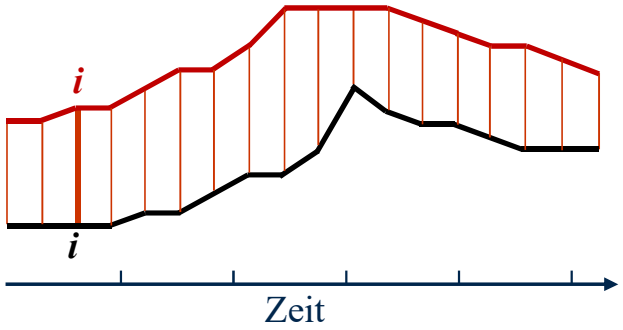


Elastische Abstände - Dynamic Time Warping

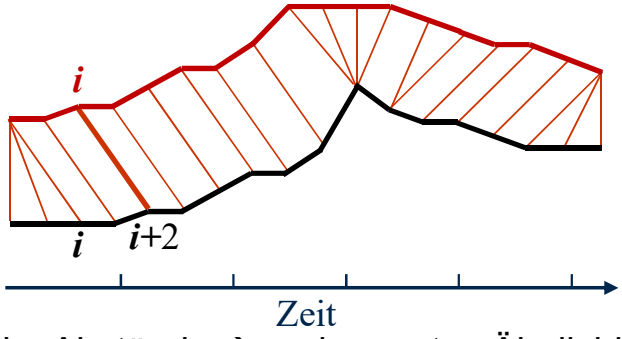
- Erkennung von Fehlerfällen anhand des Signalverlaufs
  - Herausforderung: ähnliche, aber unterschiedliche Verläufe
    - Z.B. zwei verschiedene Öffnungsvorgänge



- Methode: Dynamic Time Warping (DTW)



Euklidische Abstände → schlechtes Ähnlichkeitsmaß

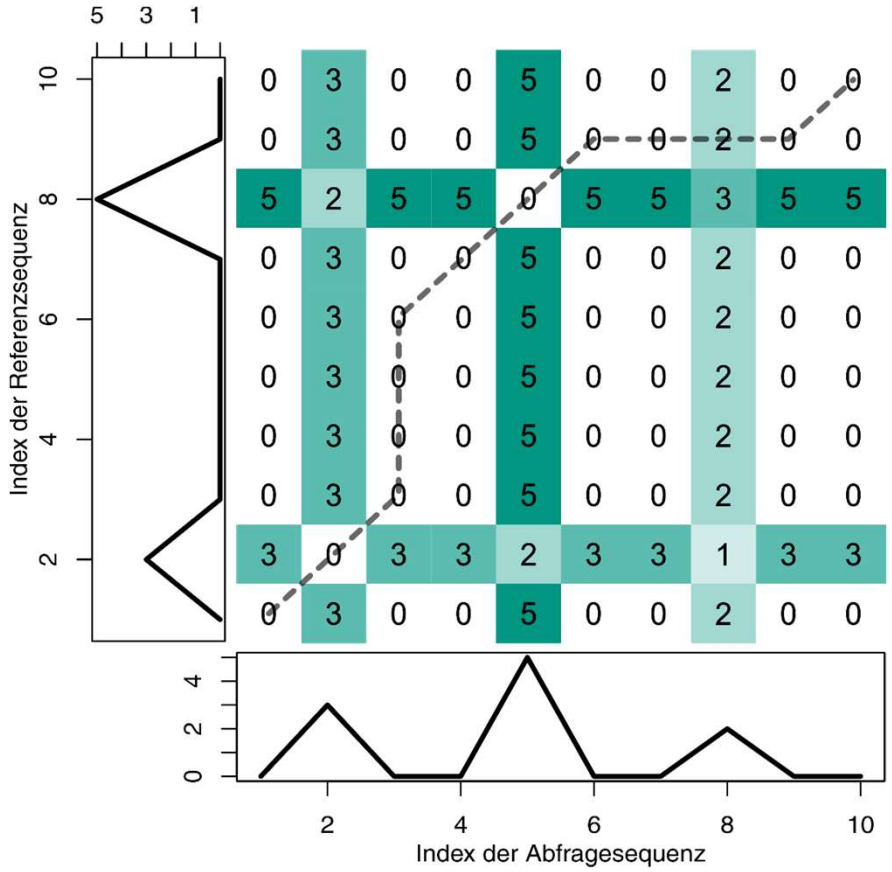


Elastische Abstände → verbessertes Ähnlichkeitsmaß

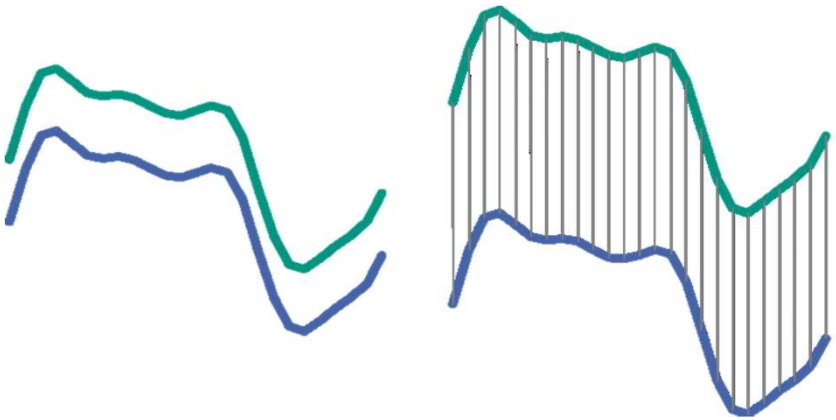
Dynamic Time Warping

- Ursprung:
  - Mathematisches Paper von H. Sakoe und S. Chiba (1978)
  - Dynamischer Ansatz der Spracherkennung
- Ansatz:
  - Suche nach geringster Distanz mit Hilfe von zeitlicher Streckung bzw. Stauchung
  - Aufstellen einer $n \times m$ Distanzmatrix
  - Warping-Kurve:

    kostengünstigster Weg von Element [1,1] zu Element [n,m]



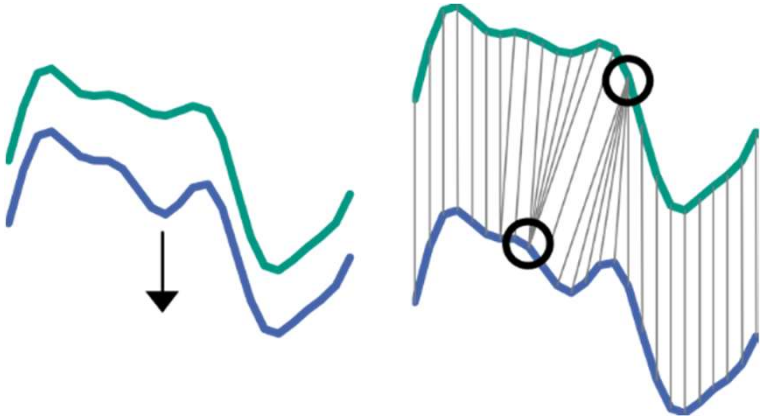
Schwäche des DTW-Algorithmus



Identische Sequenzen

↓

One-to-One-Matching



Leicht veränderte Abfragesequenz

↓

stark ausgeprägte Singularitäten

Der DTW-Algorithmus kann nur mit zeitlichen Veränderungen der Referenzsequenz umgehen

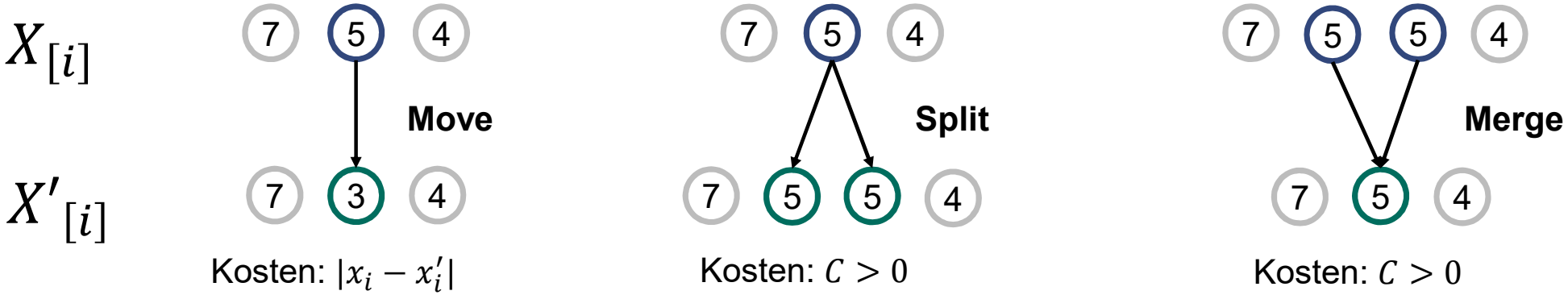
↓

Anpassungen: Verwenden der Ableitung, nutzen einer Strafgewichtung

Bilder abgeändert nach: Keogh, E. & Pazzani, M. (2001). Derivative Dynamic Time Warping. In First SIAM International Conference on Data Mining (SDM'2001), Chicago, USA.

Move-Split-Merge (MSM)

- Ansatz:
  - Wie stark müssen die beiden Sequenzen verändert um sie einander vollständig anzugleichen?
- Kostengünstigste Kombinationen aus den 3 Operationen:



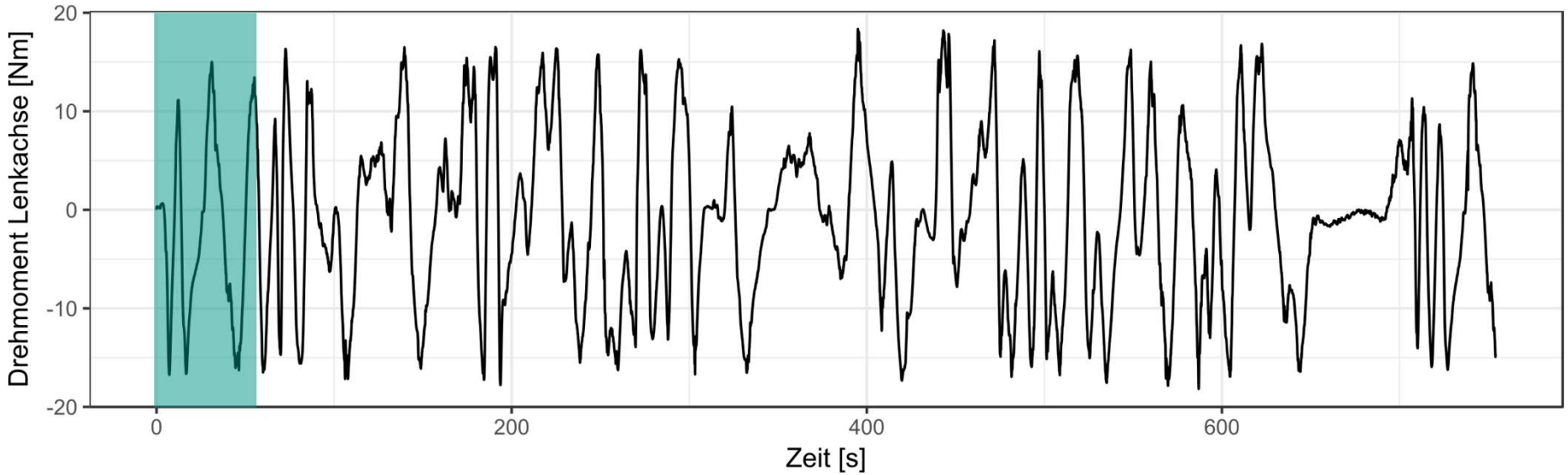
MSM stellt für diskrete Zeitreihen eine vollwertige Metrik dar

Überblick

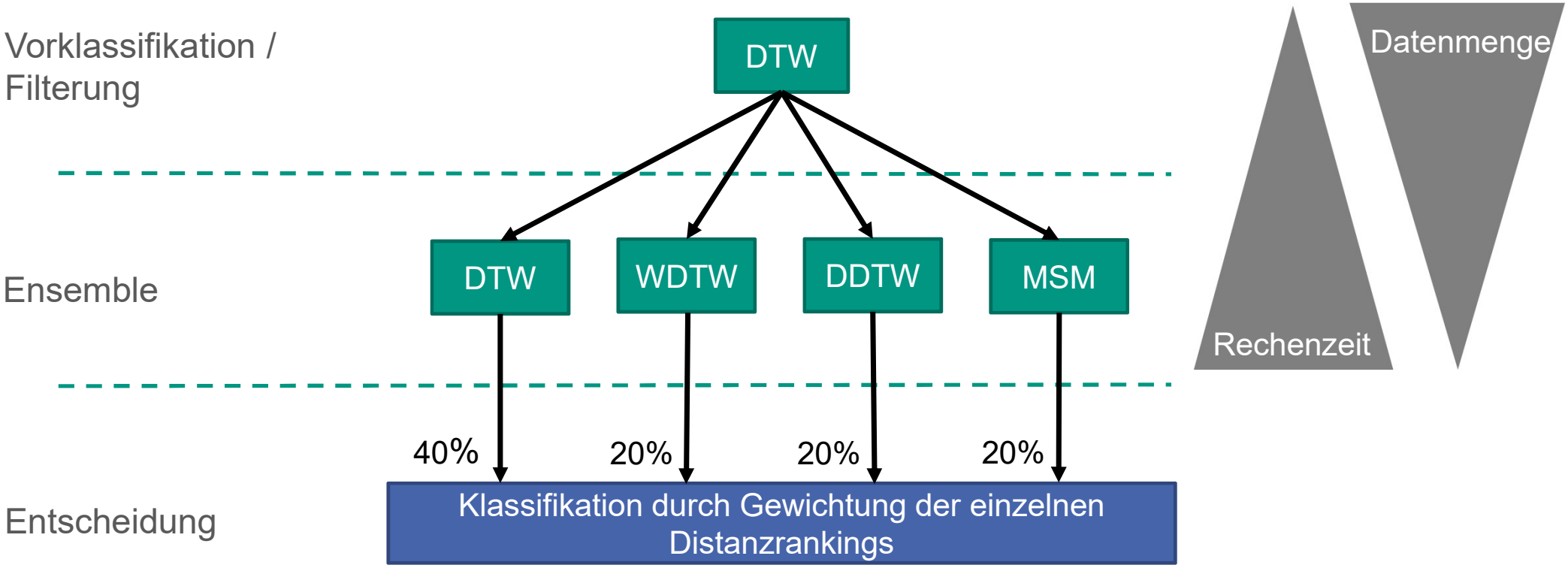
Algorithmus	Besonderheit / Vorteil	Metrik
Dynamic Time Warping (DTW)	Kann durch (starke) zeitliche Verzerrung die Zuordnung zweier Sequenzen finden	✗
Derivative DTW (DDTW)	Zuordnung über Ableitung, findet zeitliche Zuordnung der Extremstellen, schwächere Singularitäten	✗
Weighted DTW (WDTW)	Nutzt Strafgewichtung um Bereich der zeitlichen Zuordnung einzuschränken, weniger rauschempfindlich	✗
Move-Split-Merge (MSM)	Erfüllt die Dreiecksungleichung	✓

Sliding-Window

- Nur bestimmte Intervalle der Daten sind von Interesse
  - Sliding-Window mit ungefährer zeitlicher Breite der gewählten Referenz
  - Datensatz in viele Fenster unterteilen und diese einzeln betrachten



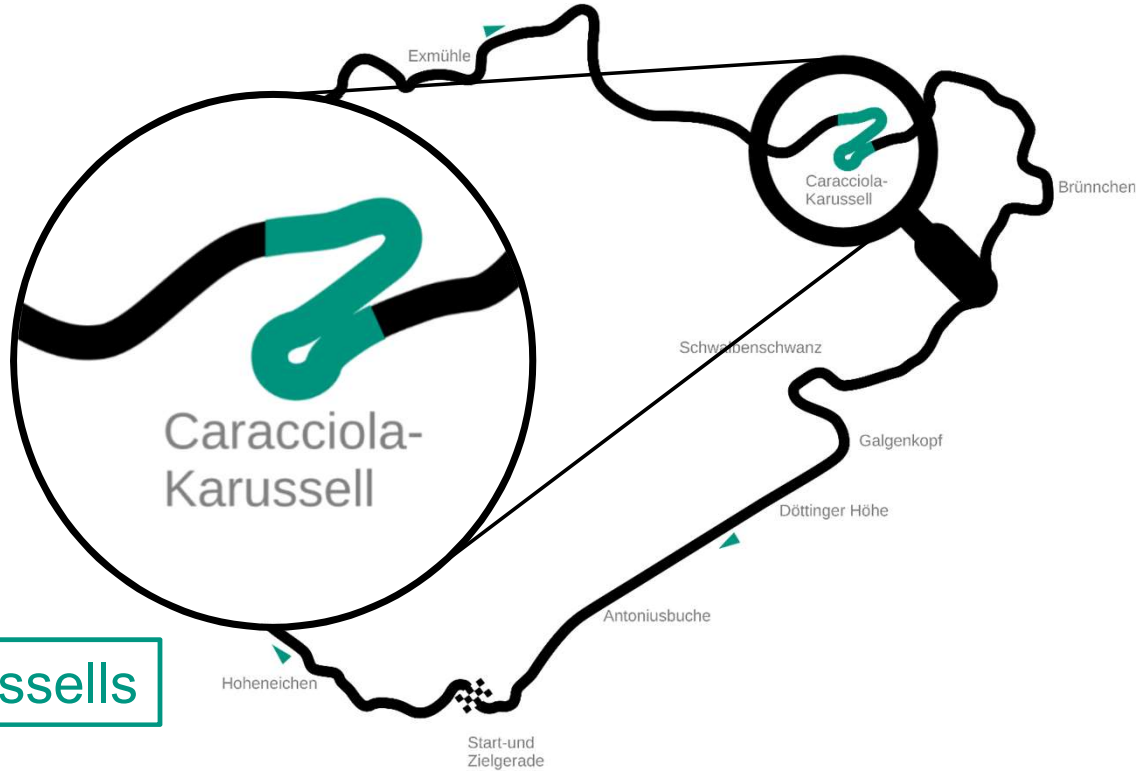
Konzept für Ensemble elastischer Distanzmaße



Testszenario für das Ensemble

- Testdatensätze wurden durch Simulationen in CarMaker® erstellt

- **Teststrecke:** Nürburgring-Nordschleife
- **Testfahrzeuge:**
  - Audi R8
  - BMW 5er
  - Lexus NX300
  - Mercedes C350
  - Peugeot 207CC



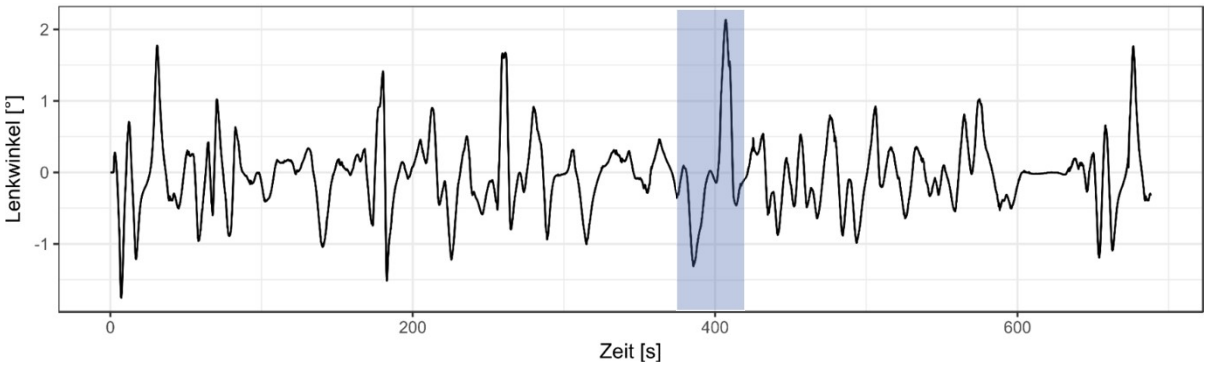
Ziel: Erkennung des Caracciola-Karussells

1. CarMaker® ist eine eingetragene Marke der IPG Automotive GmbH
2. Bild abgeändert nach: PitLane02. *Nürburgring, Streckenführung ab 2002, die Nordschleife (Namen 2013; inkl. Stefan-Boloff-S)*. Lizenz: CC BY-SA 3.0. 2013. URL: https://commons.wikimedia.org/wiki/File:Circuit_N%C3%BCrburgring-2013-Nordschleife.svg (besucht am 05.09.2018)

Referenzsequenzen Caracciola-Karussell

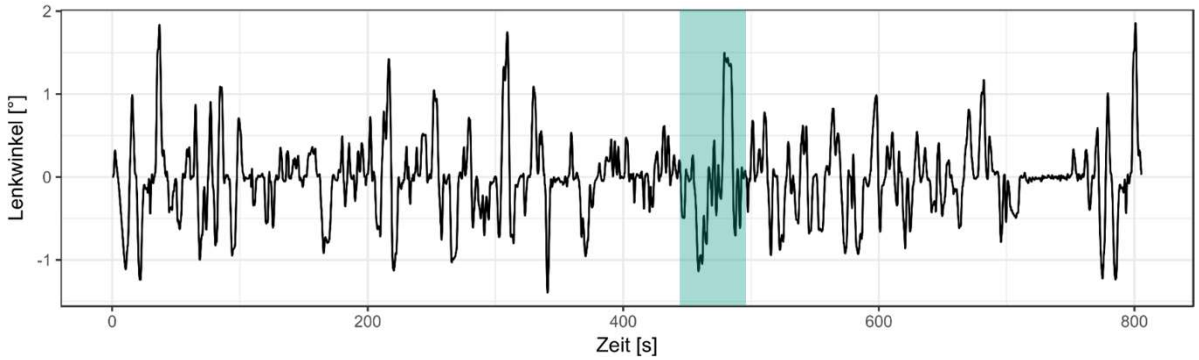
Untersuchte Fahrt:

Durchfahrt auf Ideallinie mit Audi R8



Referenz:

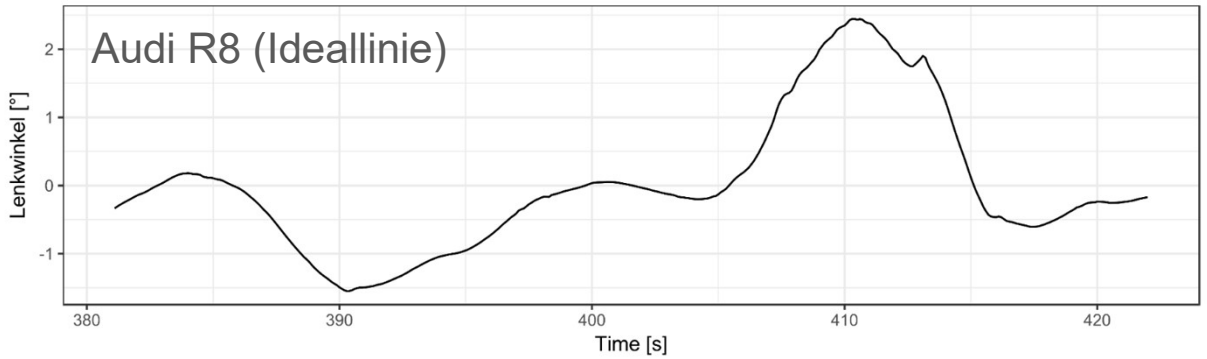
Durchfahrt auf Fahrbahnmitte mit BMW 5er



Referenzsequenzen Caracciola-Karussell

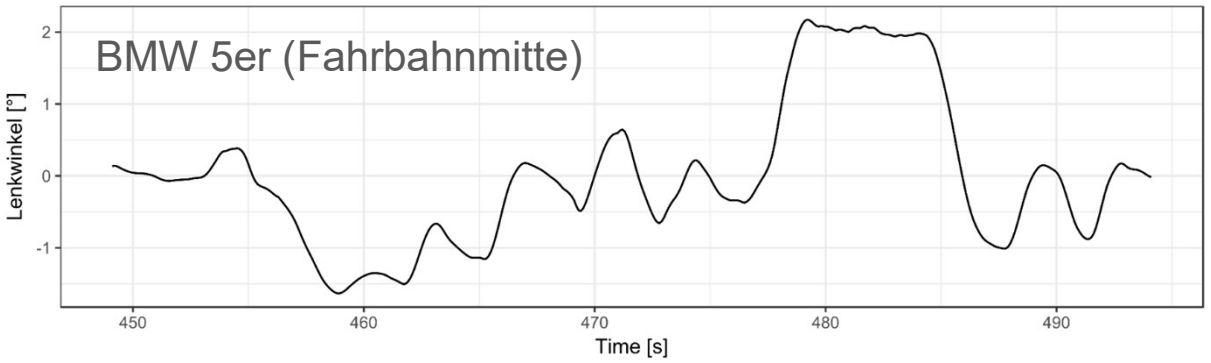
Untersuchte Fahrt:

Durchfahrt auf Ideallinie mit Audi R8



Referenz:

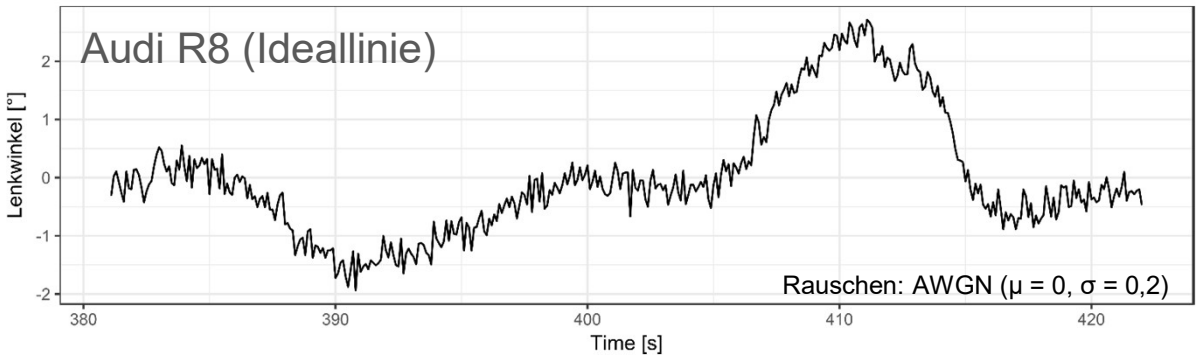
Durchfahrt auf Fahrbahnmitte mit BMW 5er



Referenzsequenzen Caracciola-Karussell mit Rauschen

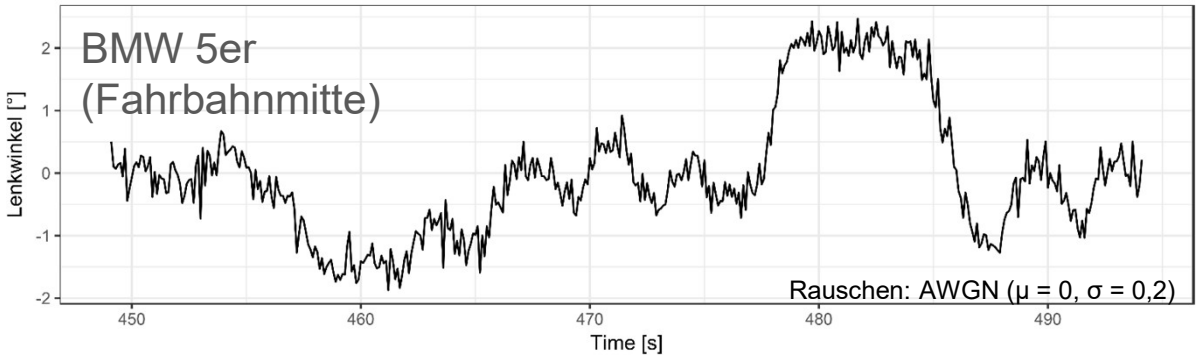
Untersuchte Fahrt:

Durchfahrt auf Ideallinie mit Audi R8 mit Rauschen

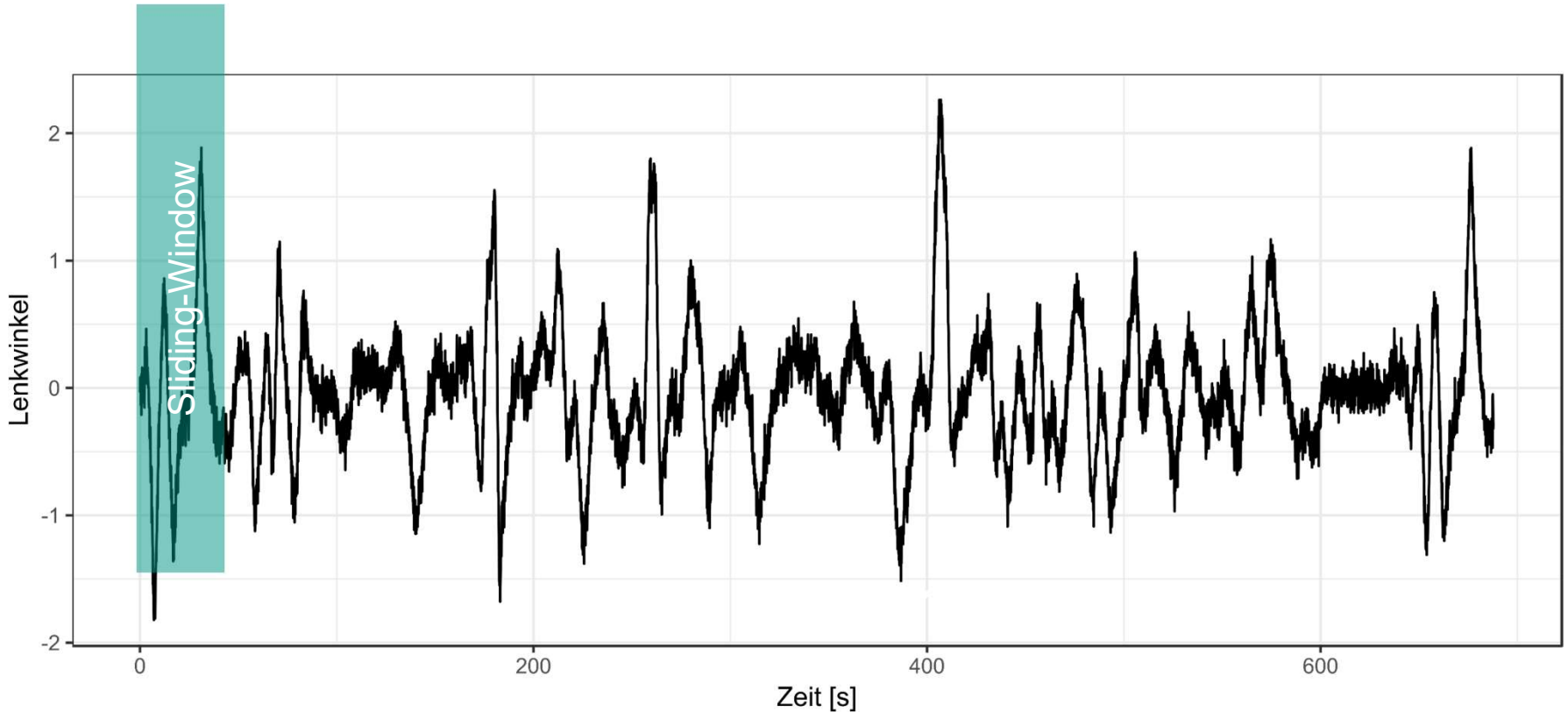


Referenz:

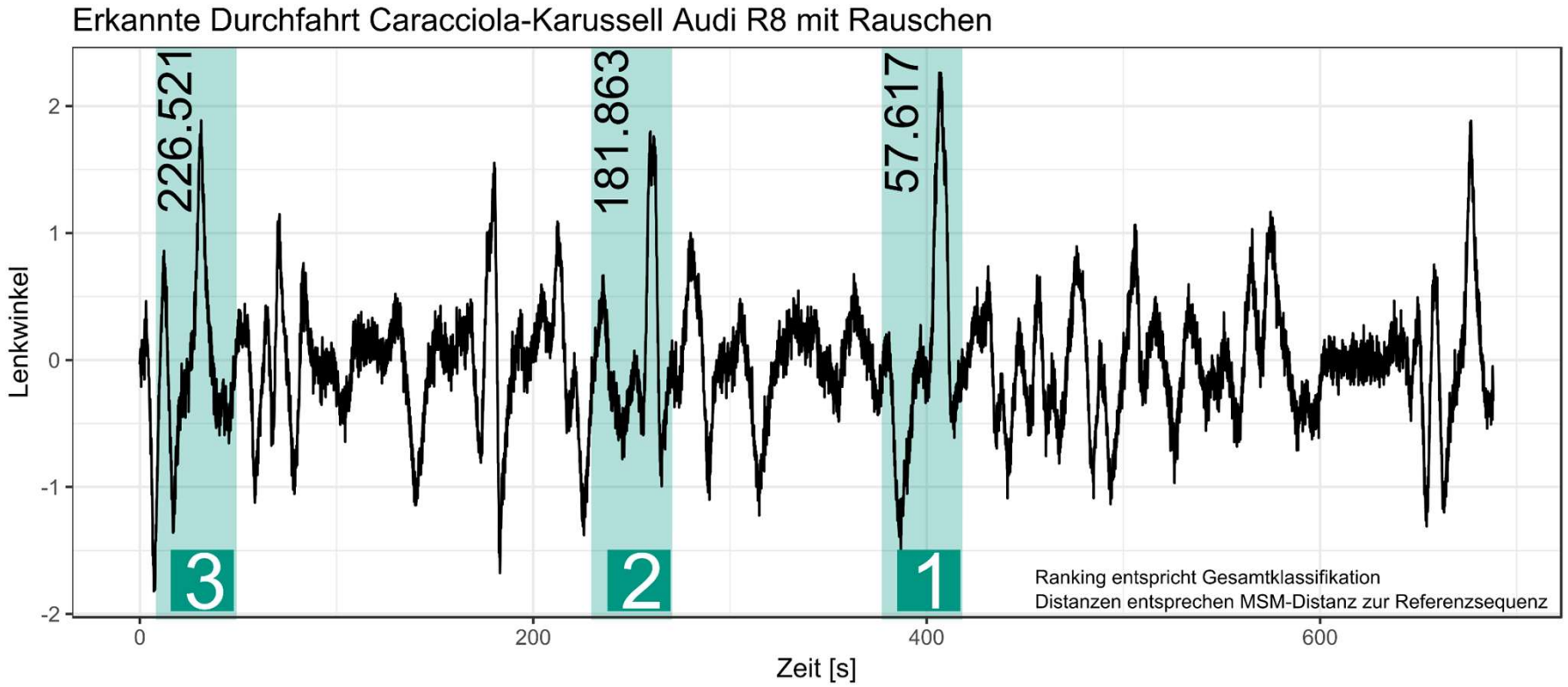
Durchfahrt auf Fahrbahnmitte mit BMW 5er mit Rauschen



Ergebnisse der Ensemble-Klassifikation (Audi R8)

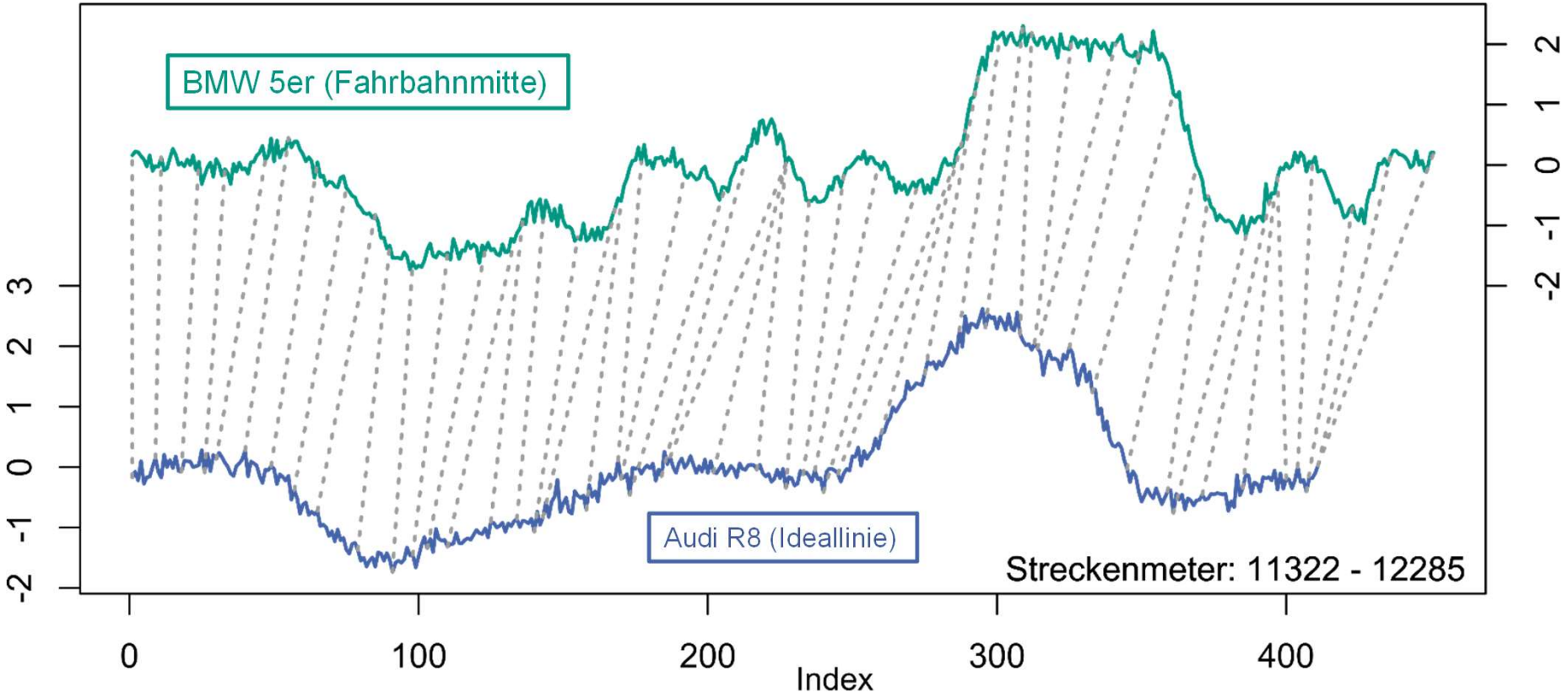


Ergebnisse der Ensemble-Klassifikation (Audi R8)



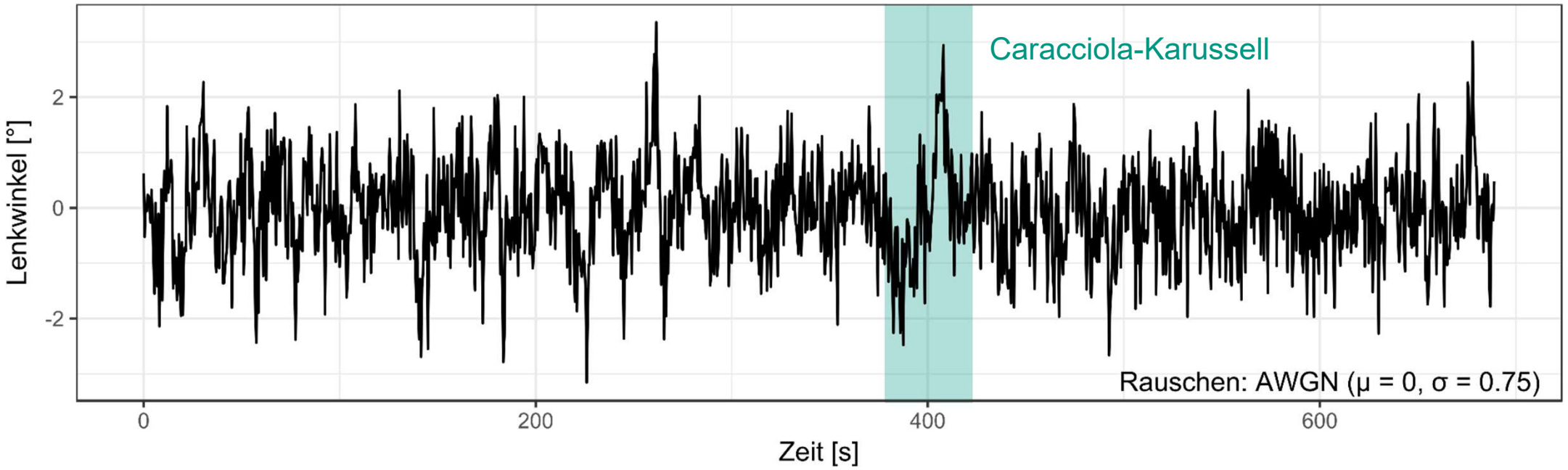
Gesuchter Streckenabschnitt kann mit großer Sicherheit extrahiert werden

Auswertung der Klassifikation



Glatte Zuordnung mit nur schwach ausgeprägte Singularitäten

Robustheit gegenüber Rauschen



Hohe Klassifikationsgüte selbst bei starkem Rauschen

Inhalt IT2

7. Beispiele für Big Data Anwendungen

- Mustererkennung auf Fahrzeugdaten
- Anomalieerkennung im Fahrzeug
- Rekonstruktion und Klassifikation von Öl-Daten

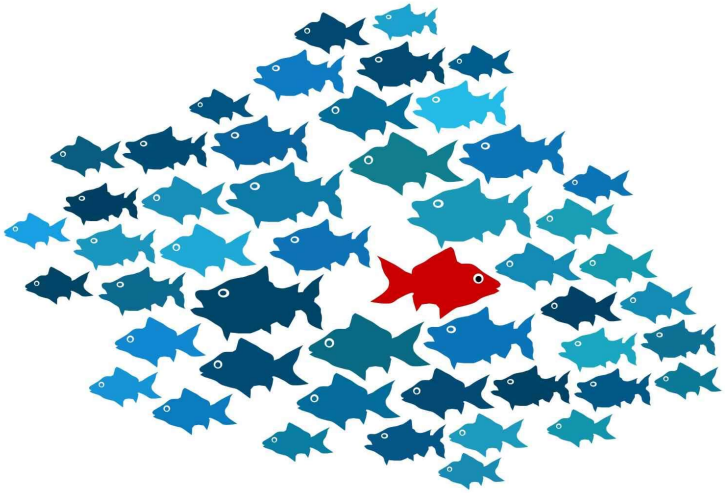
8. Cyber Security, Datenschutz

- Definition, Begriffe
- Angreifertypen- und ziele
- Schutzziele
- Kryptographie
  - Symmetrische Verschlüsselung
  - Asymmetrische Verschlüsselung
- Datenschutz

Anomalie-Erkennung
Motivation

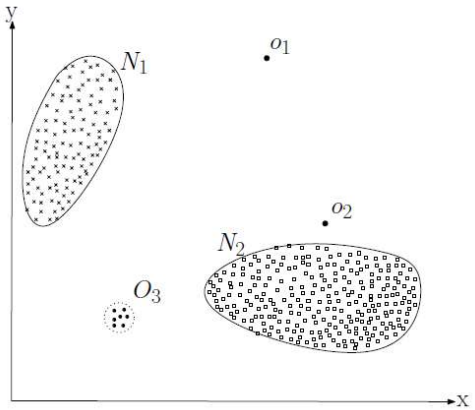
- Große Datenmengen und schnelle Abfertigungszeiten erfordern Unterstützung durch intelligente Algorithmen
  - Bieten objektive Entscheidung welche Werte anormal
  - Ermöglichen quantitative Analysen
- Erkennung von Anomalien in Daten führen zu wichtigen verwertbaren Informationen
  - Betrugsfallerkennung Kreditkarten, Versicherungen, …
  - Fehlererkennung in sicherheitskritischen System
  - Anomalieerkennung in Fahrzeugdaten

- Anomalie-Erkennung durch Machine Learning und Predictive Maintenance zur frühzeitigen Erkennung von Maschinenausfällen kann zu enormen Einsparungen führen

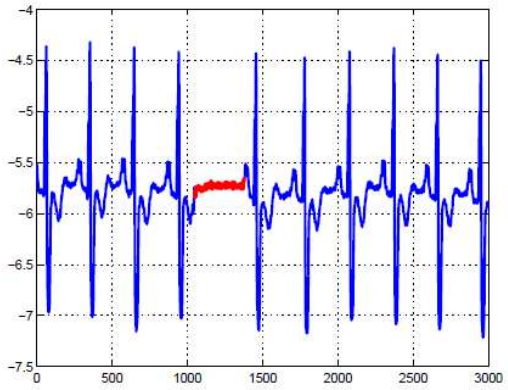


Anomalieerkennung

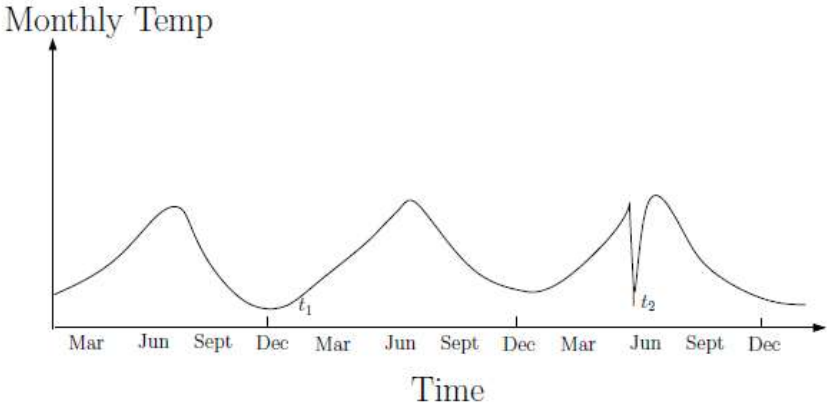
- beschreibt das Problem, Muster in Daten zu finden, die nicht mit dem erwarteten Verhalten übereinstimmen. [1]



Punkt-Anomalien



Kollektive Anomalie in EKG



Kontextuelle Anomalie t_2 in a Temperatur-Zeitserie. Temperatur bei t_1 ist die gleiche wie bei t_2, erscheint aber in anderem Kontext.

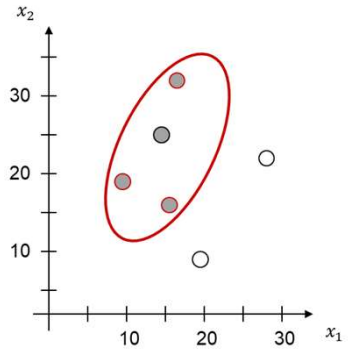
[1]: V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM Computing Surveys*, vol. 41, no. 3, 2009.

Anomalieerkennung in Fahrzeugdaten

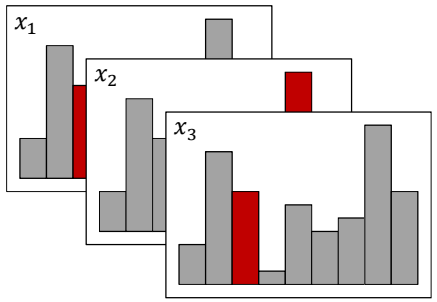
Analyse und Bewertung von Algorithmen

- Randbedingungen:
  - Nur normale Trainingsdaten
  - Steuergeräte-Ressourcen (Kosten)
  - Adaption zur Laufzeit

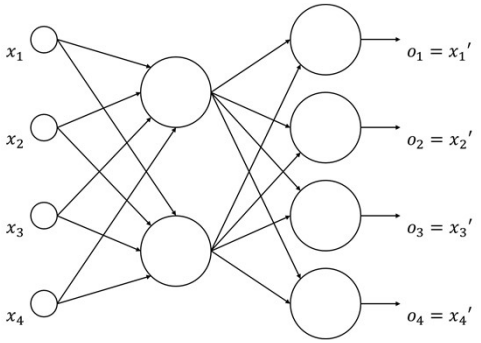
One Class Support Vector Machine [5]
→ *Kein Lernen zur Laufzeit*



Lightweight On-line Detector of Anomalies [6]



Autoencoder [7]
→ *Weniger Ressourcenverbrauch*



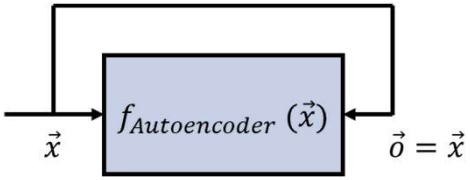
[5] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," in Advances in Neural Information Processing Systems 12. Cambridge, MA, USA: MIT Press, 2000, pp. 582–588.
[6] T. Pevný, "Loda: Lightweight on-line detector of anomalies," Machine Learning, vol. 102, no. 2, pp. 275–304, 2016.
[7] S. Hawkins, H. He, G. Williams, and R. Baxter, "Outlier detection using replicator neural networks," in Data Warehousing and Knowledge Discovery, ser. Lecture Notes in Computer Science, Y. Kambayashi,M. Arikawa, and W. Winiwarter, Eds. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2002.

Anomalieerkennung in Fahrzeugdaten

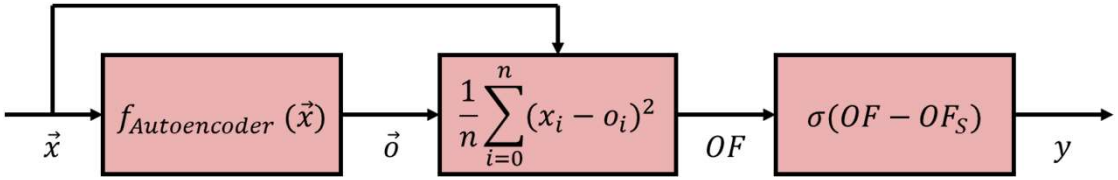
Autoencoder zur Signalplausibilisierung

- Funktionsweise

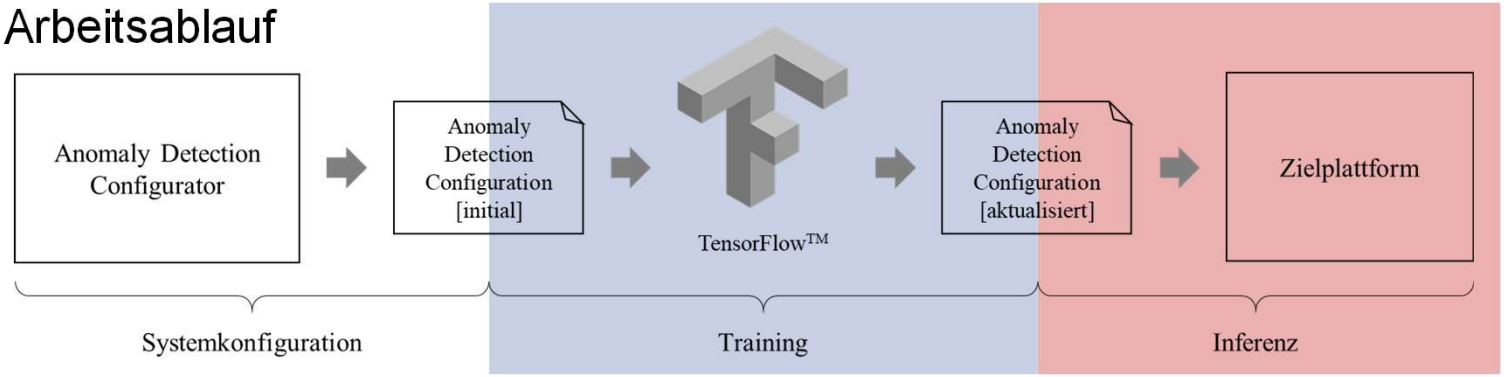
1. Training



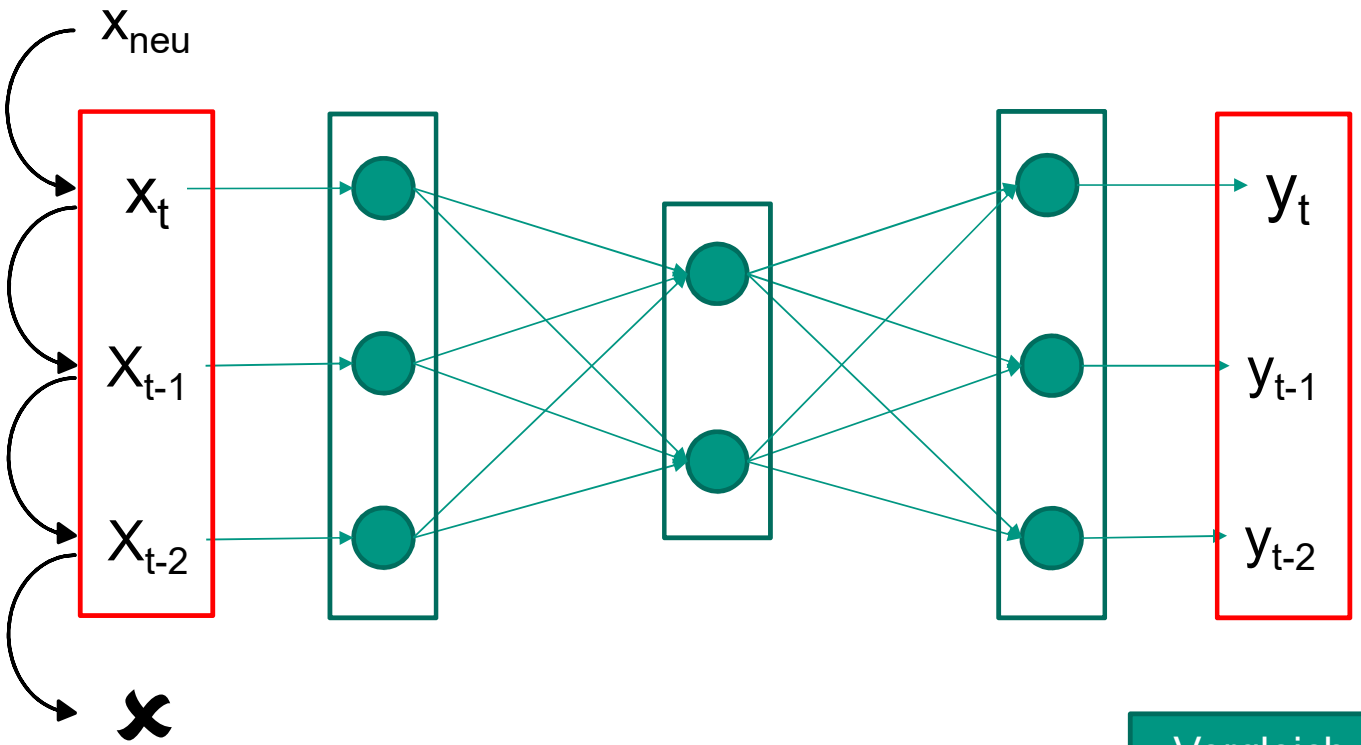
2. Inferenz



- Eingangsvektor enthält die letzten Werte eines Signals (Sliding Window Ansatz)
- Standard Autoencoder; andere Varianten benötigen im Allgemeinen mehr Ressourcen
- Definierter Arbeitsablauf



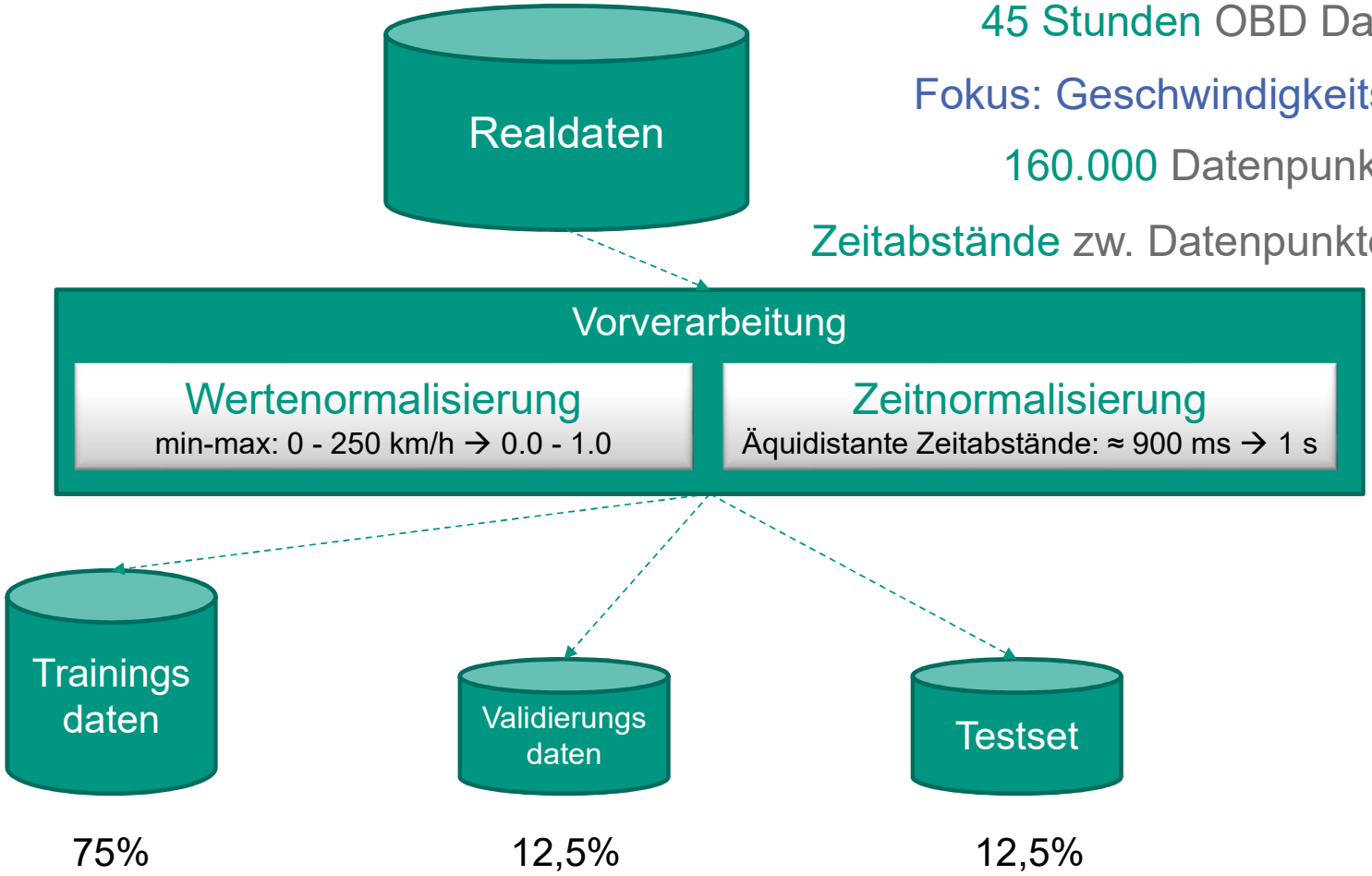
Anomalie-Erkennung mit Autoencoder



$$OF = \left| \begin{pmatrix} x_t \\ x_{t-1} \\ x_{t-2} \end{pmatrix} - \begin{pmatrix} y_t \\ y_{t-1} \\ y_{t-2} \end{pmatrix} \right| = \sqrt{(x_t - y_t)^2 + (x_{t-1} - y_{t-1})^2 + (x_{t-2} - y_{t-2})^2}$$

Vergleich mit Schwellwert → normal / Anomalie

Datensatz



45 Stunden OBD Daten

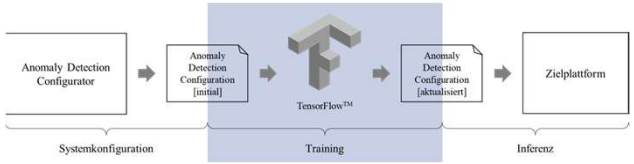
Fokus: Geschwindigkeitssignal

160.000 Datenpunkte

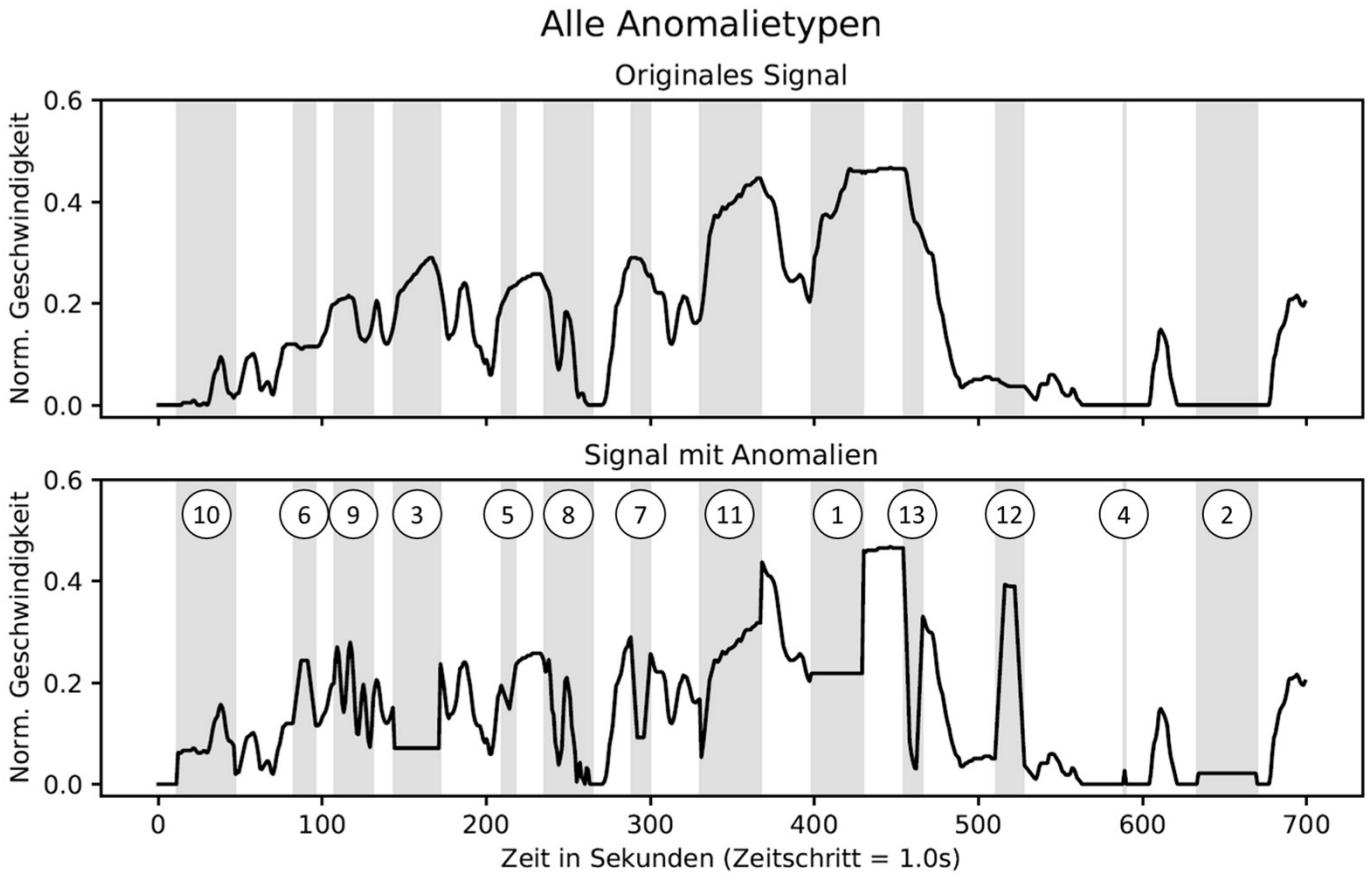
Zeitabstände zw. Datenpunkten ≈ 900 ms

Evaluierung von Autoencodern

Definition von Anomalie-Typen



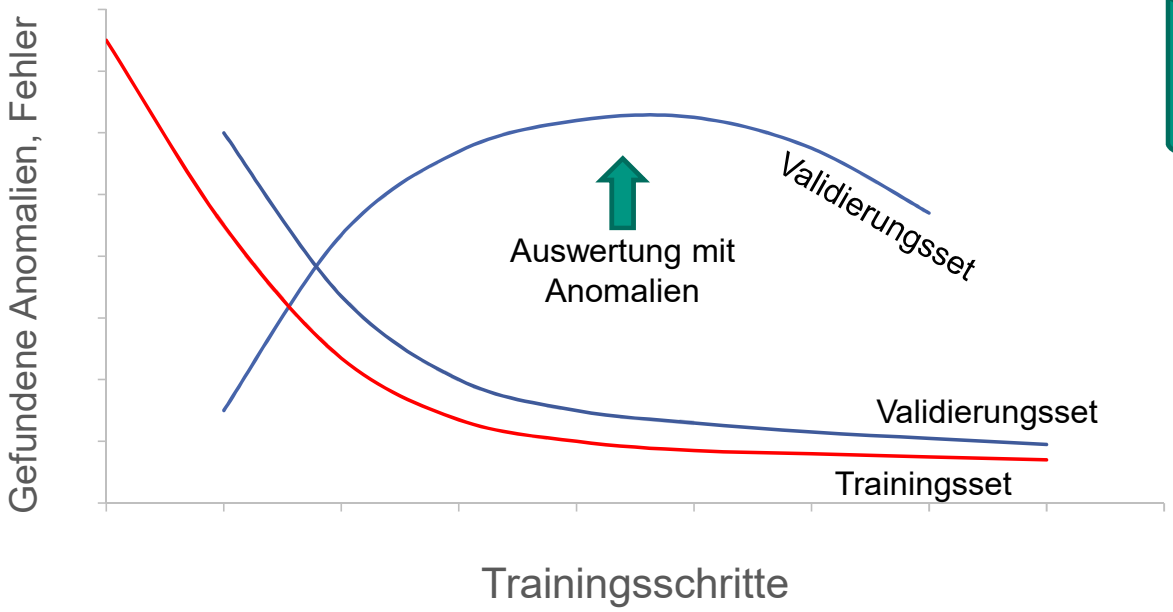
- Selektion und Bewertung des besten Autoencoders anhand von normalen Daten nicht möglich
  - Anreicherung mit Anomalien
  - Definition von 13 Anomalie-Typen (nach ISO26262)
  - Zufällig Synthese basierend auf Parametrierung

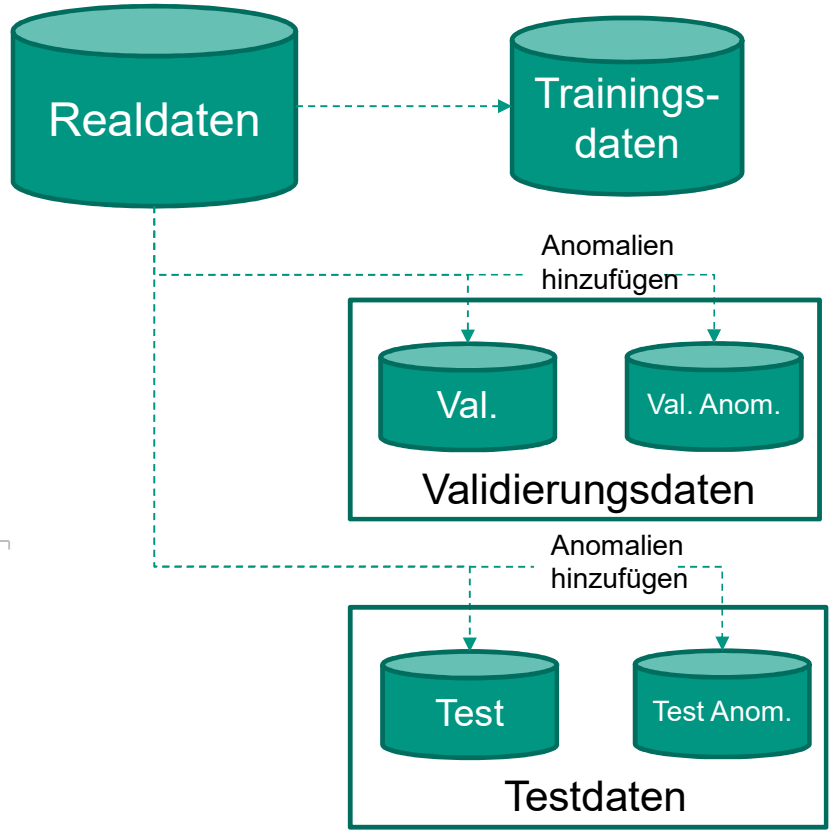


M. Weber, F. Pistorius, E. Sax, J. Maas and B. zimmer, „A hybrid anomaly detection system for electronic control units featuring Replicator Neural Networks“

Bestimmung des Stop-Punktes

Validierung mittels synthetischer Anomalien

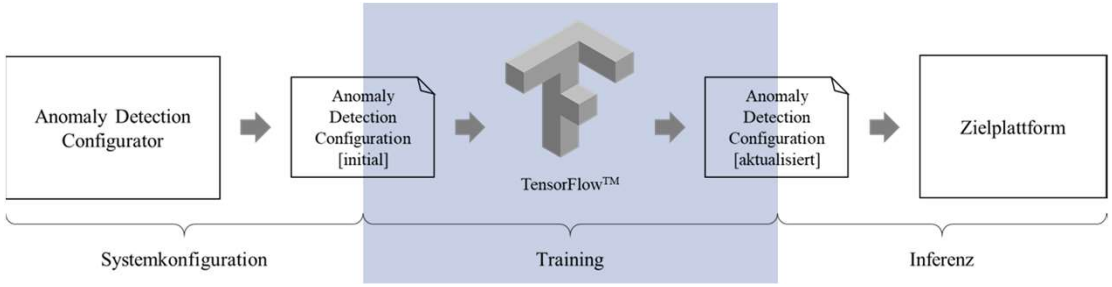


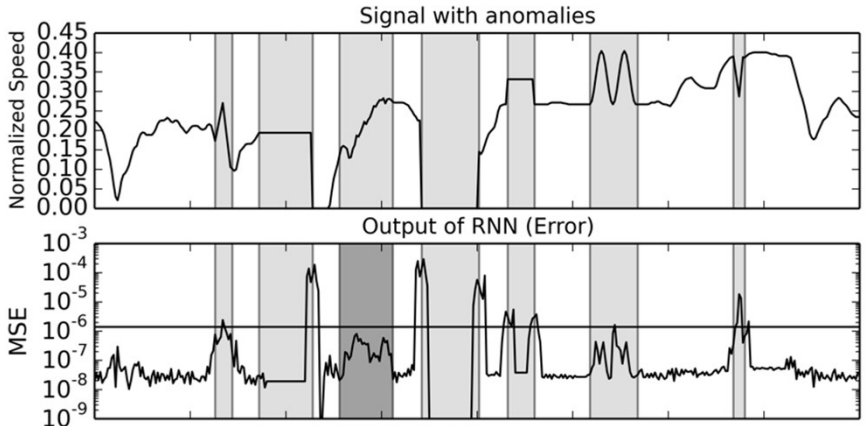


M. Weber, G. Wolf, E. Sax and B. Zimmer, „Online Detection of Anomalies in Vehicle Signals using Replicator Neural Networks“, TBP

Evaluierung von Autoencodern

Ergebnisse





- Autoencoder mit 8-64-8 Architektur
  - 8 Eingangs-, 64 versteckte und 8 Ausgangsneuronen
- Erkennungsrate: 68/101
  - Falschalarmrate 0,065%
  - Ca. 15% besser als eine Erkennung anhand der 1. und 2. Ableitung

Motivation - Warum Anomalieerkenung?

http://www.auto.de/magazin/hackerangriffe-abgewehrt-dacia-ist-sicherstes-auto/

Inhalt IT2

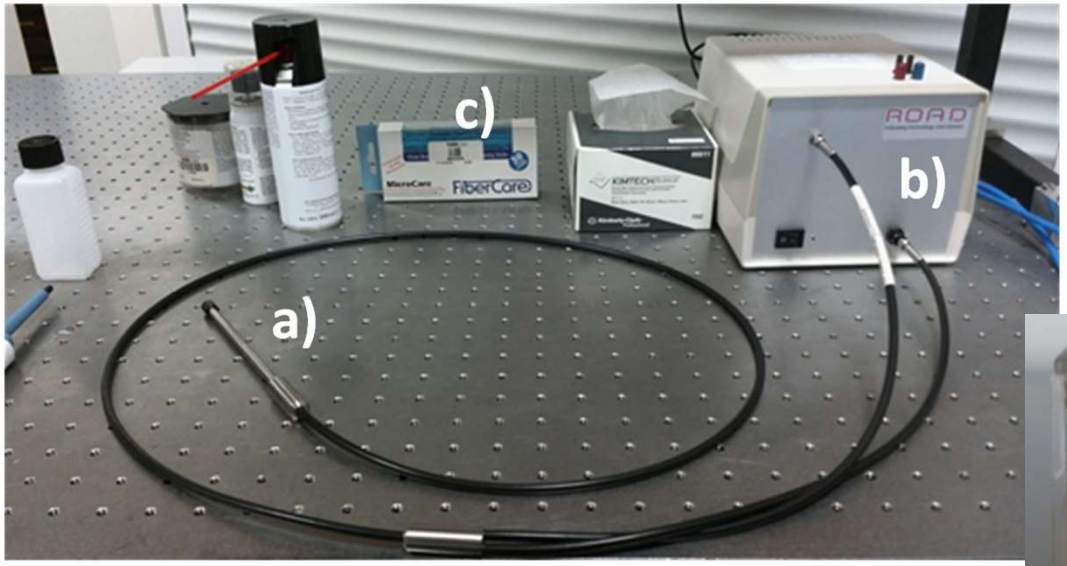
7. Beispiele für Big Data Anwendungen

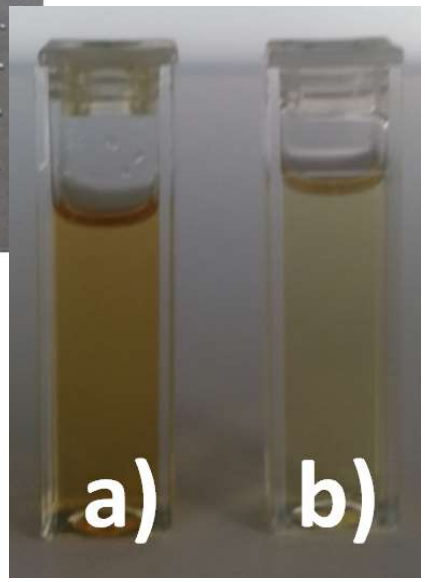
- Mustererkennung auf Fahrzeugdaten
- Anomalieerkennung im Fahrzeug
- Rekonstruktion und Klassifikation von Öl-Daten

8. Cyber Security, Datenschutz

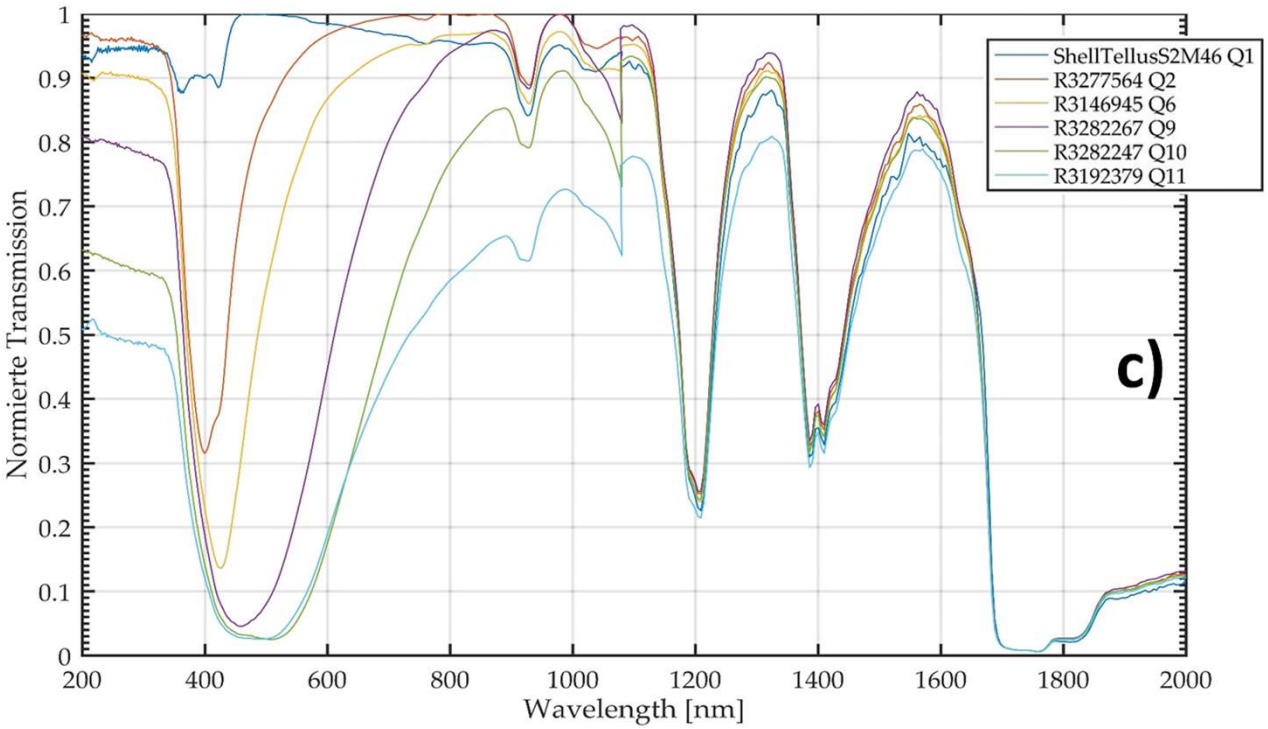
- Definition, Begriffe
- Angreifertypen- und ziele
- Schutzziele
- Kryptographie
  - Symmetrische Verschlüsselung
  - Asymmetrische Verschlüsselung
- Datenschutz

Datenaufzeichnung im KIT- Mixed Signal Labor



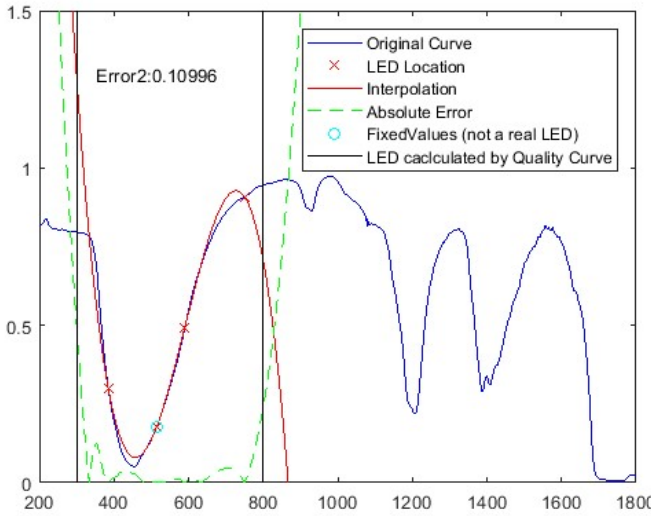


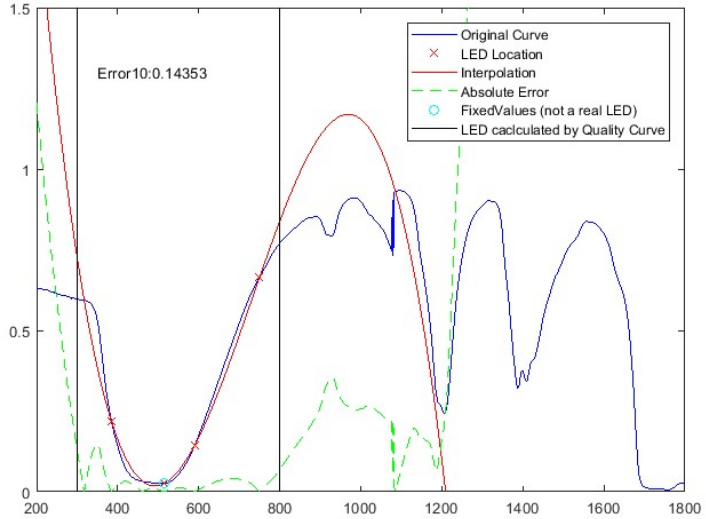
Analyse der aufgezeichneten Daten



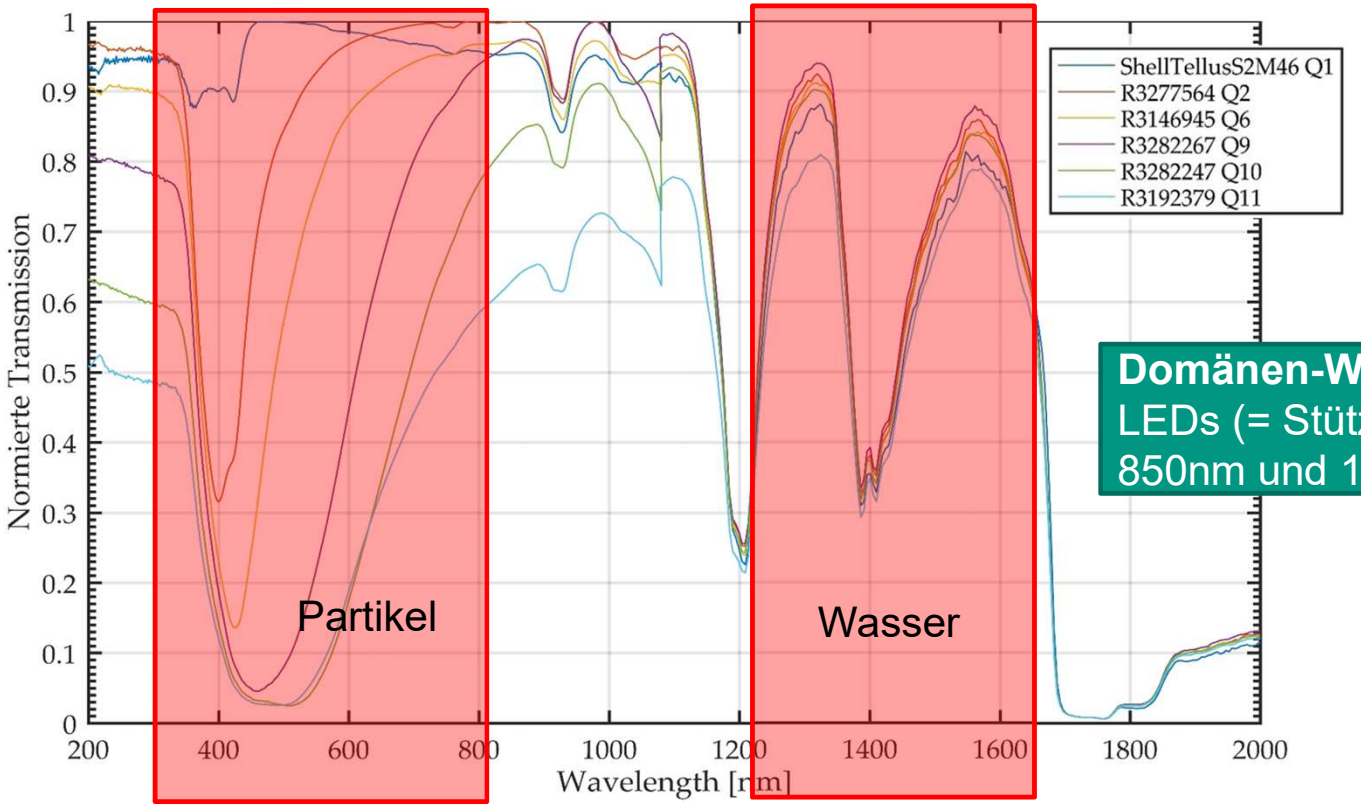
Randbedingungen und Fragestellung

- Constraints:
  - kein Bread Board Model im Feld verwenden
  - Optischer Sensor arbeitet mit wenigen LEDs mit unterschiedlichen Wellenlängen
    - Aber welche Wellenlängen sind am besten?



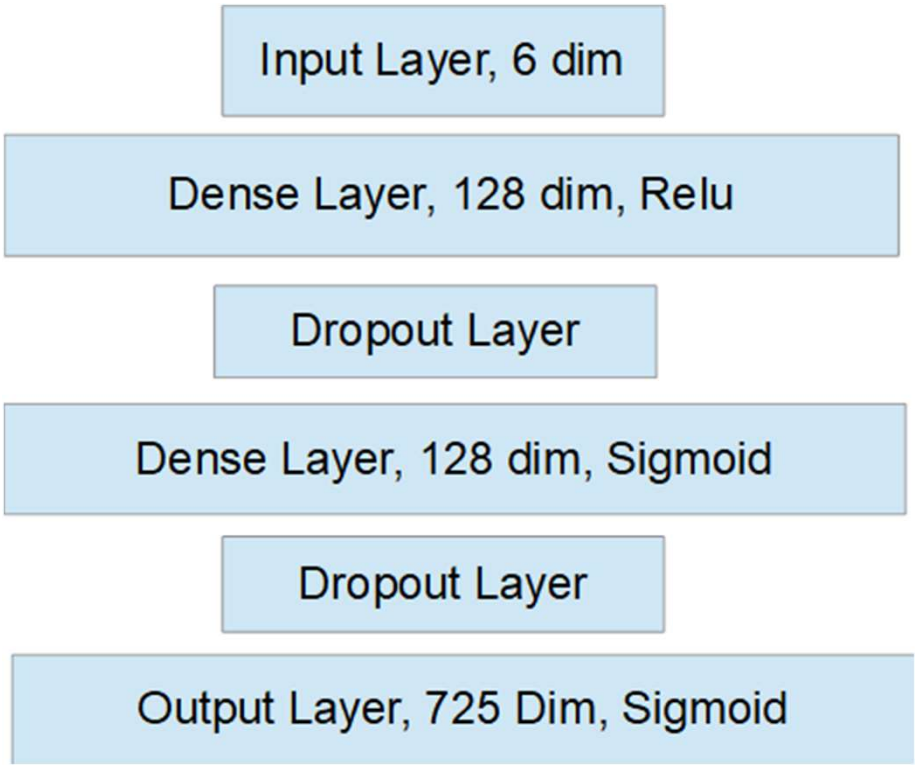


KDD Prozess- Rekonstruktion- Selektion

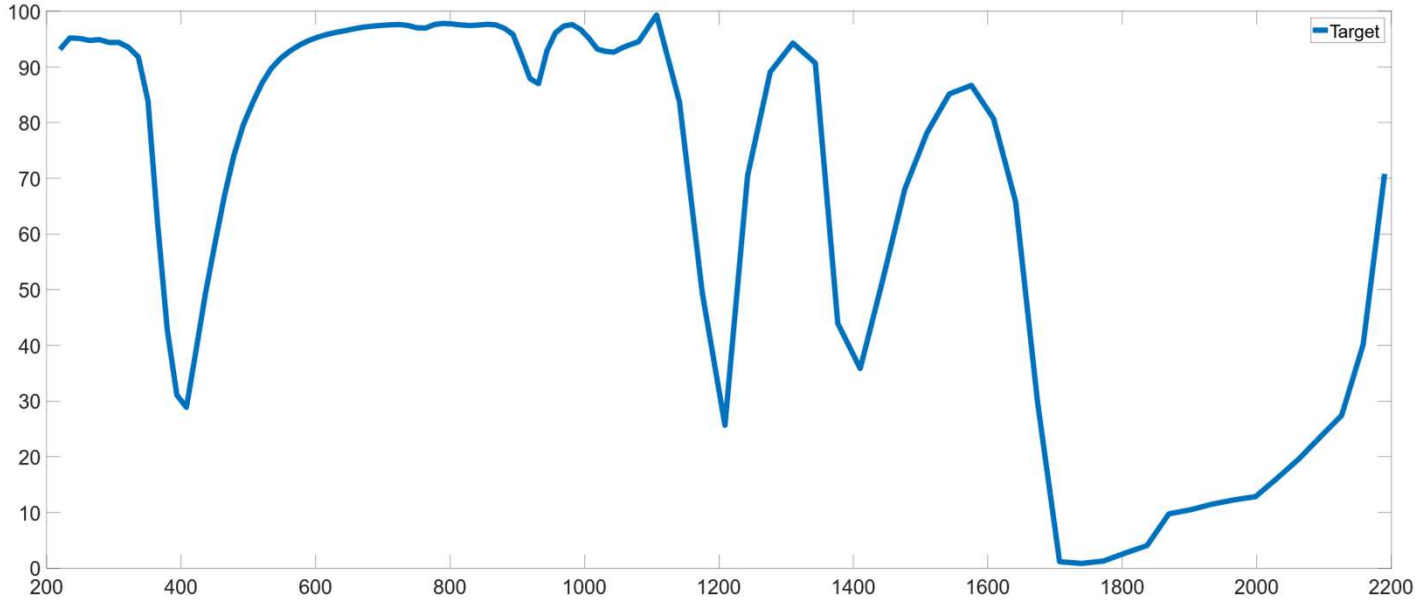


Domänen-Wissen:
LEDs (= Stützstellen) bei 300nm, 850nm und 1300nm

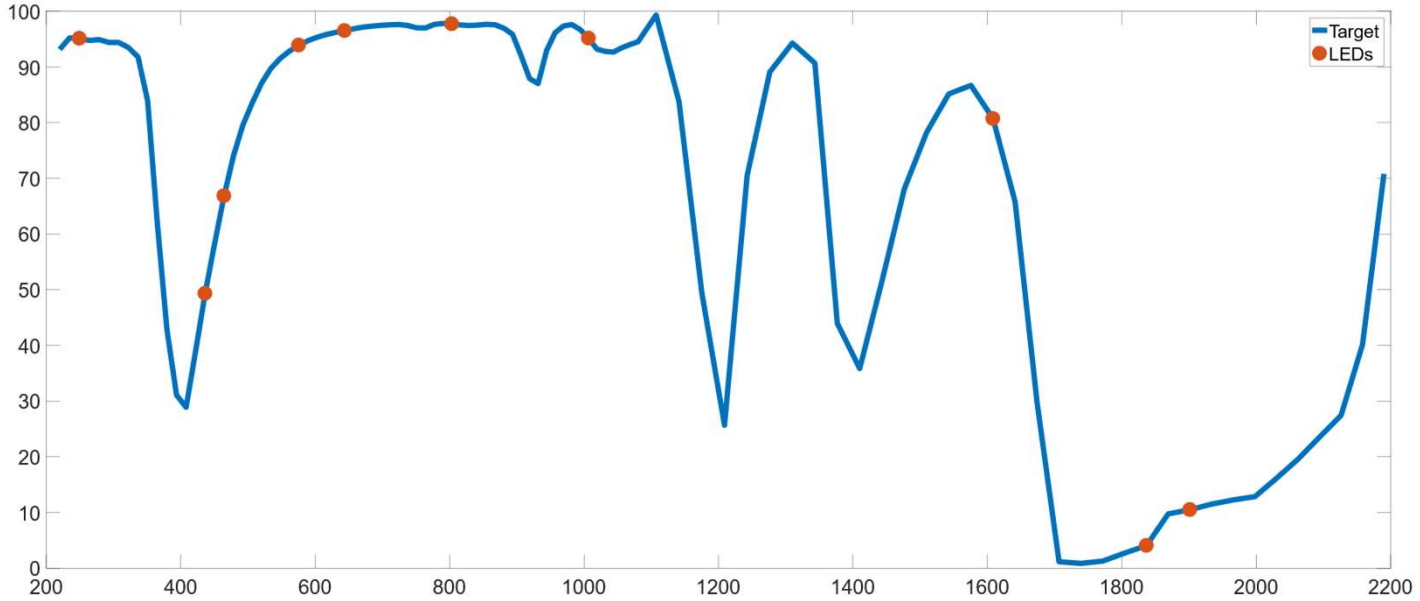
KNN Aufbau


Input Layer, 6 dim
Dense Layer, 128 dim, Relu
Dropout Layer
Dense Layer, 128 dim, Sigmoid
Dropout Layer
Output Layer, 725 Dim, Sigmoid

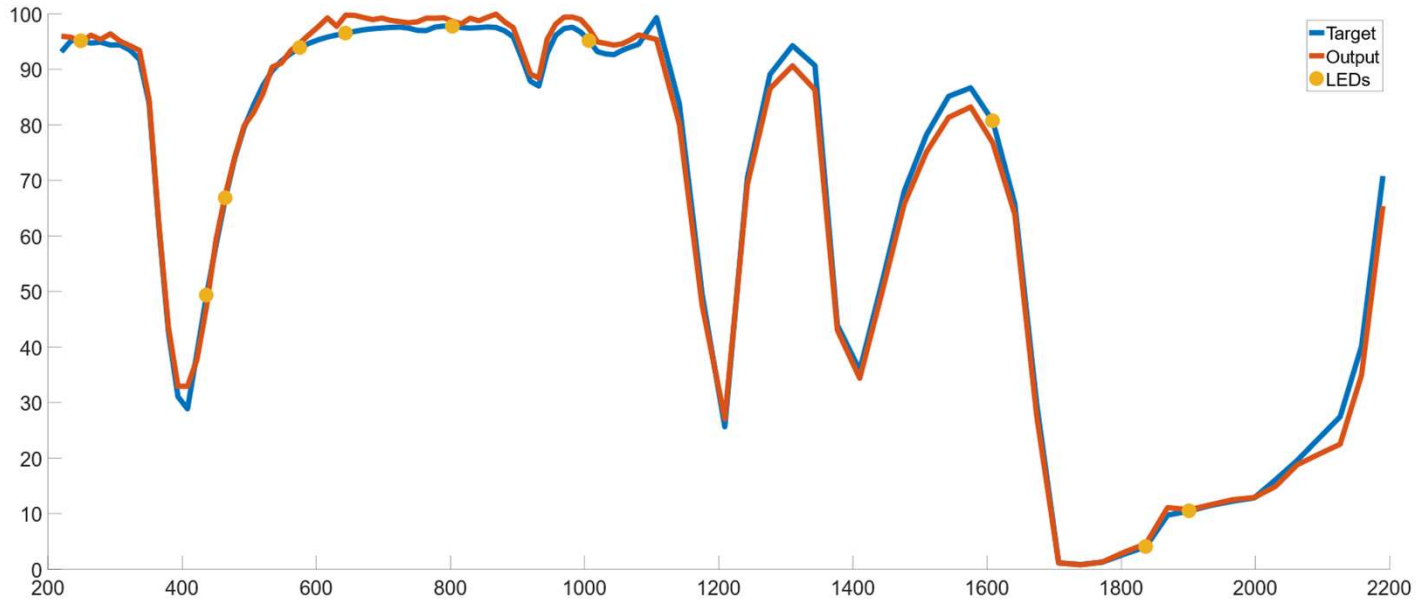
Zusammenfassung
Rekonstruktion durch Neuronale Netze (1/3)



Zusammenfassung
Rekonstruktion durch Neuronale Netze (1/3)

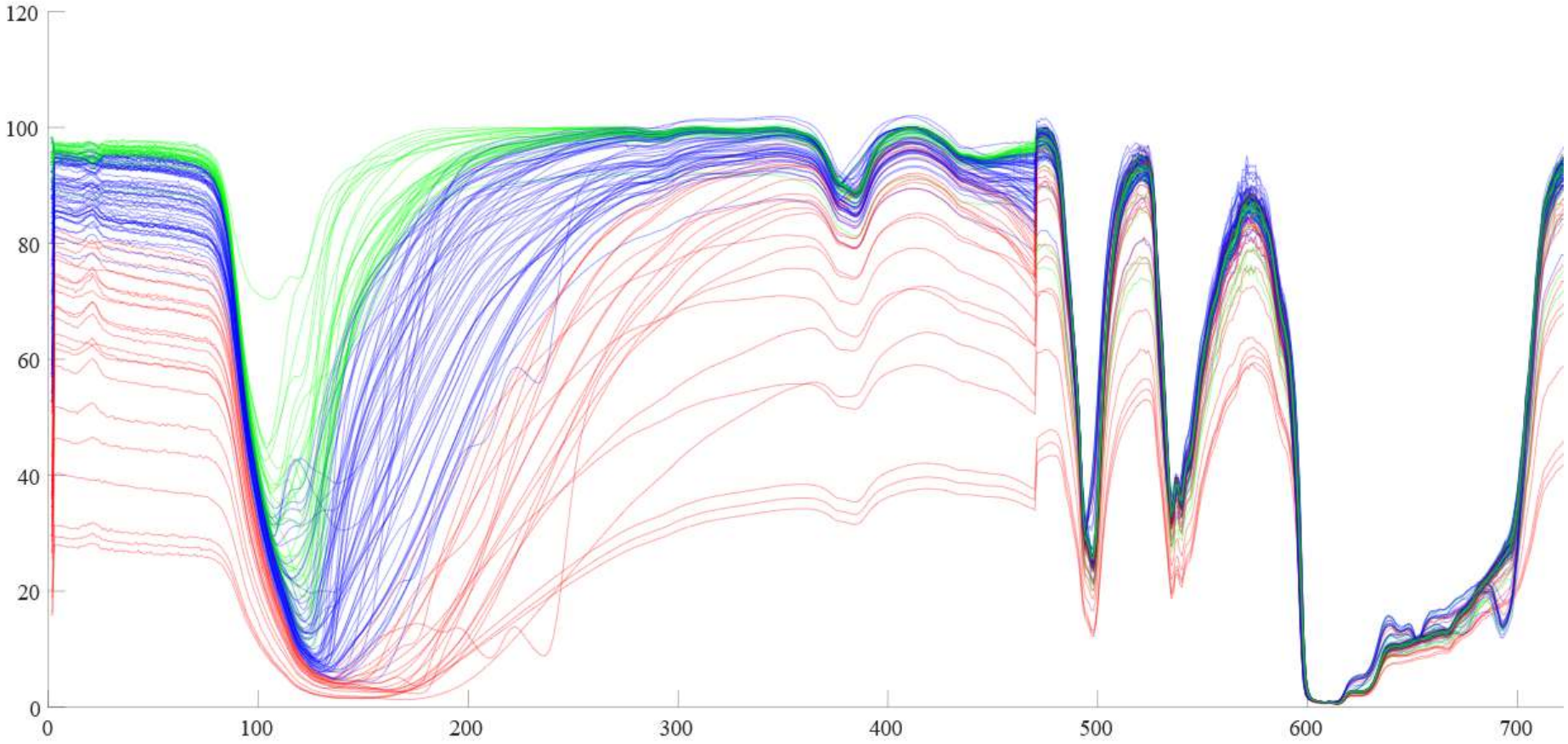


Zusammenfassung
Rekonstruktion durch Neuronale Netze (1/3)



- Ergebnisse der Tests zeigen:
  - Akkurate Resultate ergeben sich schon bei vier Layers mit 200 Neuronen
  - Weitere Neuronen und Layers scheinen nur bedingt Einfluss zu nehmen

Klassifizieren von Ölqualität



Karlsruher Institut für Technologie

Fragen?