

Datenanalyse mit KI

Vom Laptop zum Supercomputer

10 April 2025
ONLINE



Datenanalyse mit KI

vom Laptop zum Supercomputer

Simeon Harrison

EuroCC Austria, VSC Research Center, TU Wien

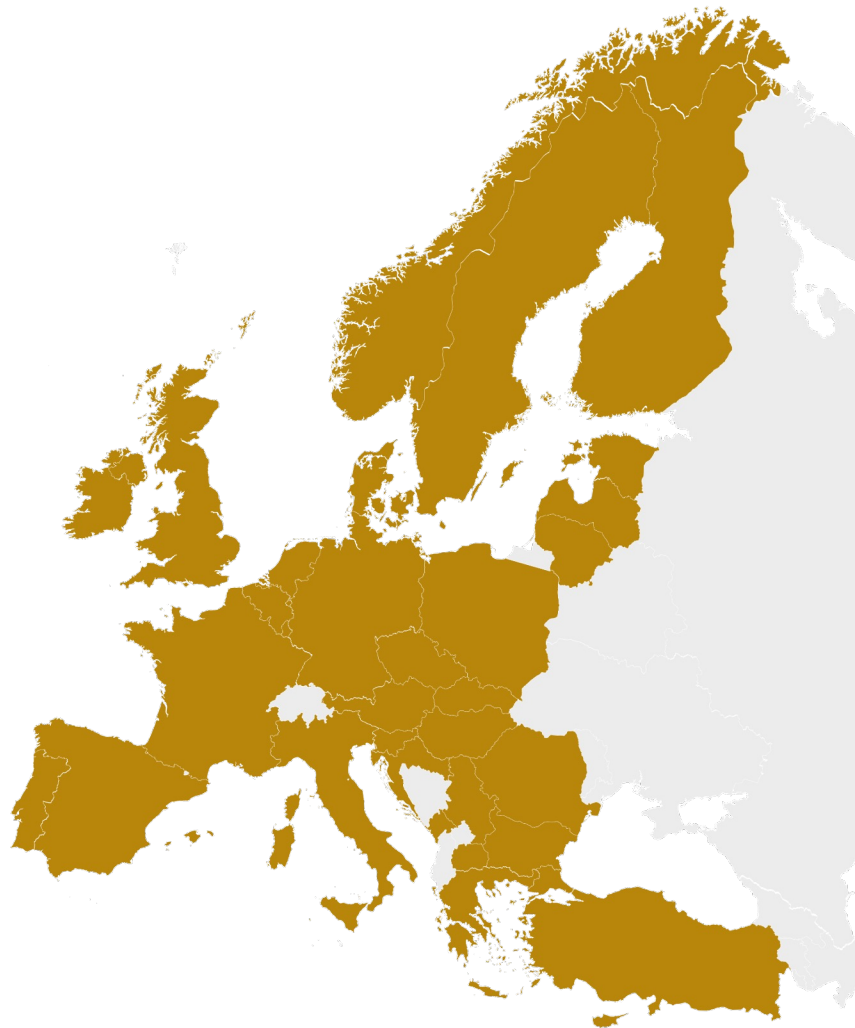
Martin Pfister

EuroCC Austria, Uni Wien

EuroCC

Fully funded EU project

- EuroCC is EU-funded international initiative aimed to support the uptake of AI and High-Performance Computing (HPC) in Europe
- Set up of 32 National Competence Centres (NCCs) across Europe
- EuroCC Austria is one of them
- Service Provider for AI, HPC and HPDA

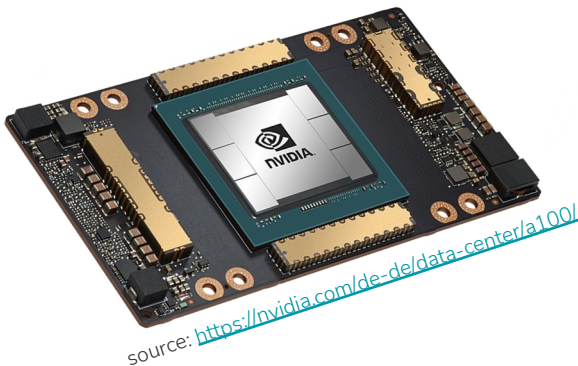


The Vienna Scientific Cluster

VSC-4 (2019)

790 CPU nodes

- 2x Intel Skylake Platinum CPUs
- 2x 24 cores per CPU
- 96 GB of memory per node



source: <https://nvidia.com/de-de/data-center/a100/>

VSC-5 (2022)

770 CPU nodes

- 2x AMD EPYC Milan
- 2x 64 cores per CPU
- 512 GB of memory per node

60 GPU nodes 2x NVIDIA A100,

- 40 GB memory per GPU

40 GPU nodes 2x NVIDIA A40

- 40 GB memory per GPU

Need More Compute-Power?

LUMI

- Third most powerful supercomputer in Europe and the 8th globally (Nov 2024)
- Sustained computing power (HPL) is 380 petaflops
- Over 262 000 AMD EPYC CPU cores
- Equipped with AMD Radeon Instinct MI250X GPUs

<https://www.lumi-supercomputer.eu/>

Leonardo

- 4th most powerful supercomputer in Europe and the 9th globally (Nov 24)
- Sustained computing power (HPL) is 239 petaflops
- Intel new gen Sapphire Rapids 56 cores
- Equipped with custom NVIDIA A100 SXM6 64GB GPUs

<https://leonardo-supercomputer.cineca.eu/>

European HPC Landscape

EuroHPC JU systems

Different access modes:
Calls for Proposals

EuroHPC development access:
Opportunity to test the system

Applicants can request a small number of node hours to get acquainted with the supercomputers to further develop their software.



AI Factory

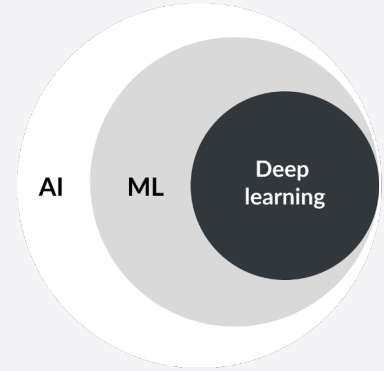
Coming to Austria very soon!

Datenanalyse mit KI

Was verstehen wir unter KI

Gemeint ist dabei maschinelles Lernen (ML) und Deep Learning (DL)

Wir betrachten in diesem Webinar „supervised“ und „unsupervised“ Methoden



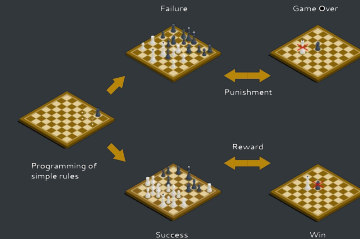
supervised learning



unsupervised learning



reinforcement learning



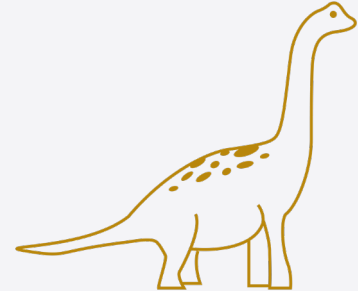
Datenanalyse mit KI

Was verstehen wir unter KI?

- Interaktiver, mehrstufiger Informationsdestillationsprozess
- Am Ende steht ein komplexes mathematisches Modell
- Funktionsparameter werden vom Algorithmus gelernt
- Lernen geschieht durch Beispiele, ähnlich dem menschlichen Lernen

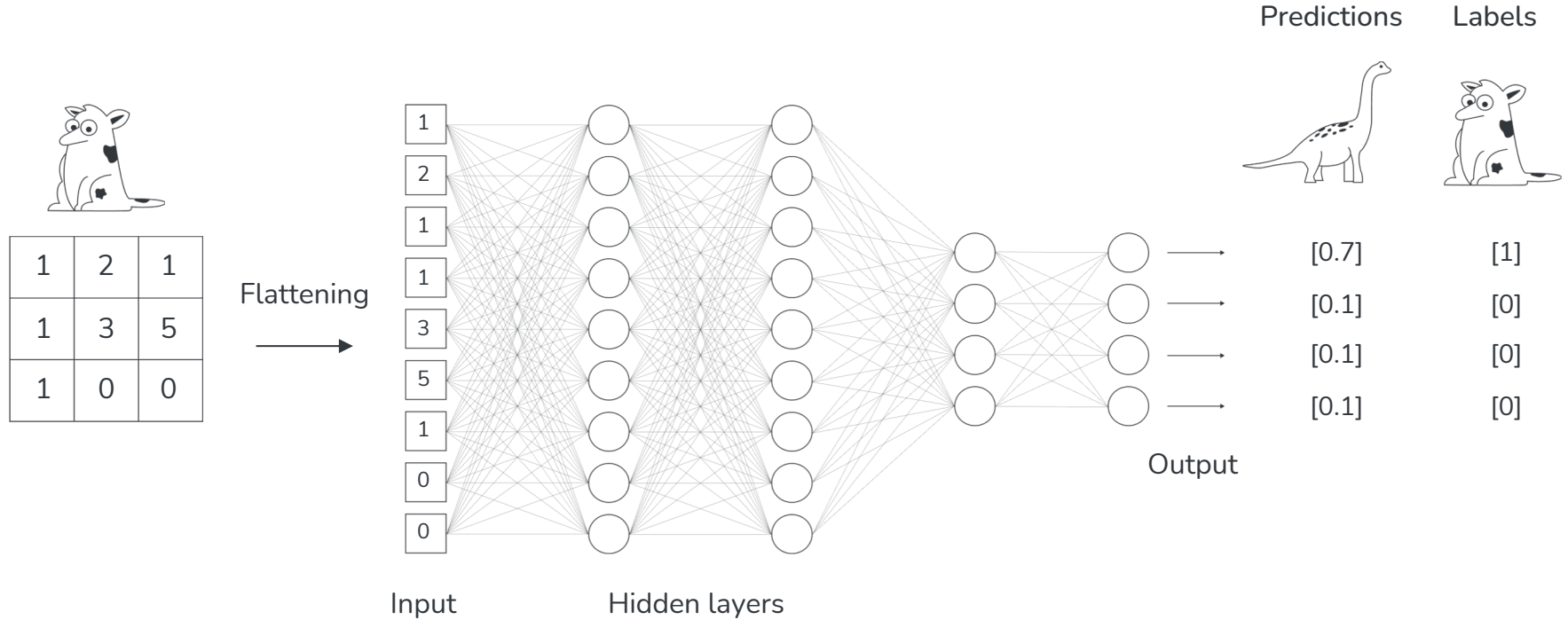


Das ist ein Hund

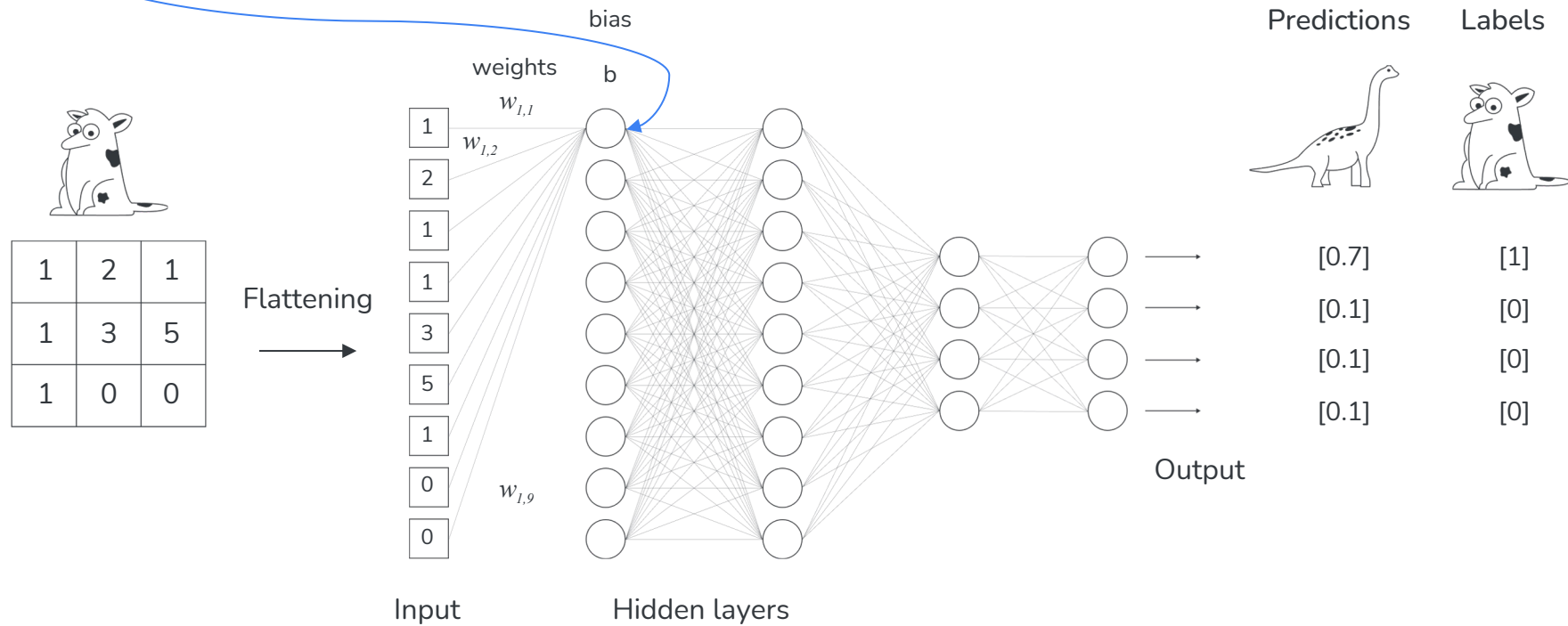


Das ist
(höchstwahrscheinlich)
kein Hund

Datenanalyse mit KI



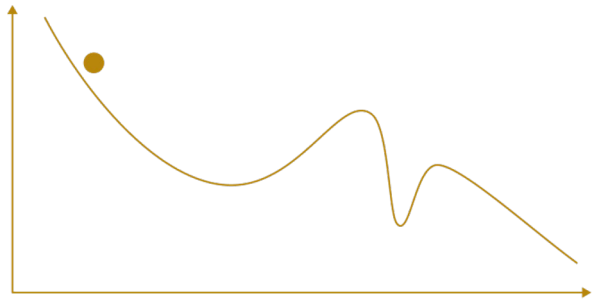
$$neuron_{layer\ 1,1} = activation_function(x_1 \times w_{1,1} + x_2 \times w_{1,2} + x_3 \times w_{1,3} + \dots + x_n \times w_{1,n} + b)$$



Datenanalyse mit KI

Minimierung der Lossfunction

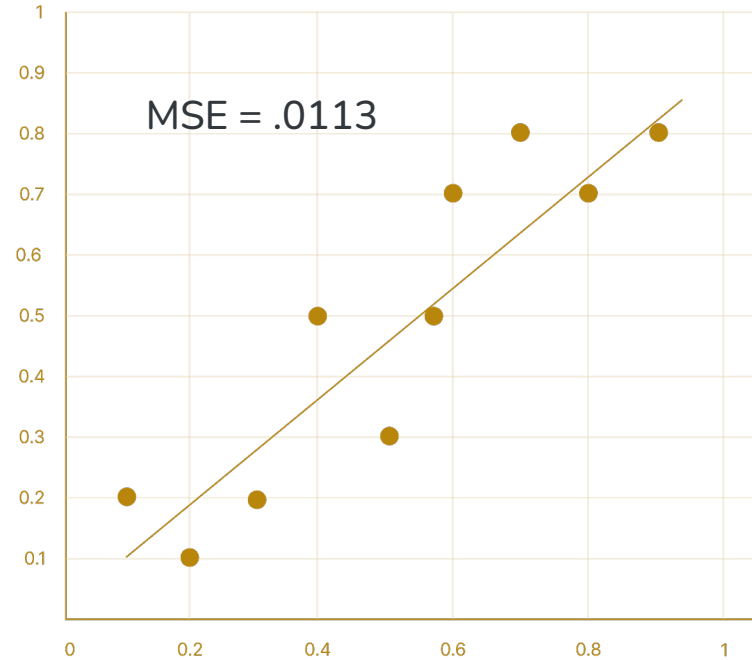
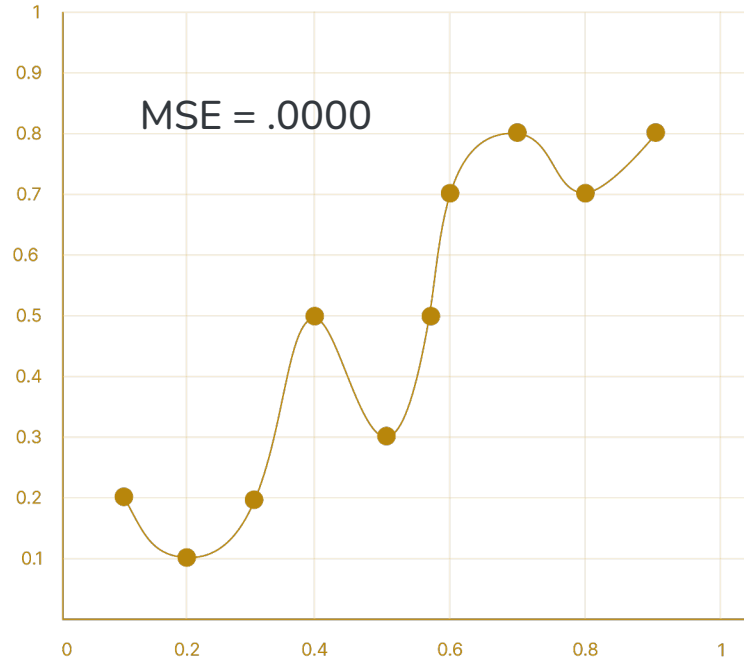
- Optimierung durch eine Form des stochastischen Gradientenabstiegs
- Iterativer Prozess (in kleinen Schritten) in Richtung des negativen Gradienten
- Ziel: Das (globale) Minimum der Lossfunction zu finden



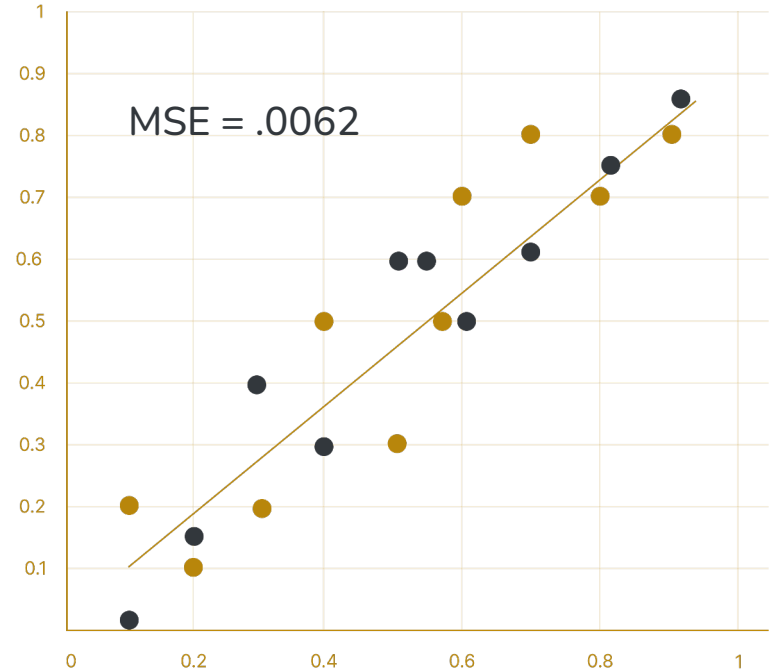
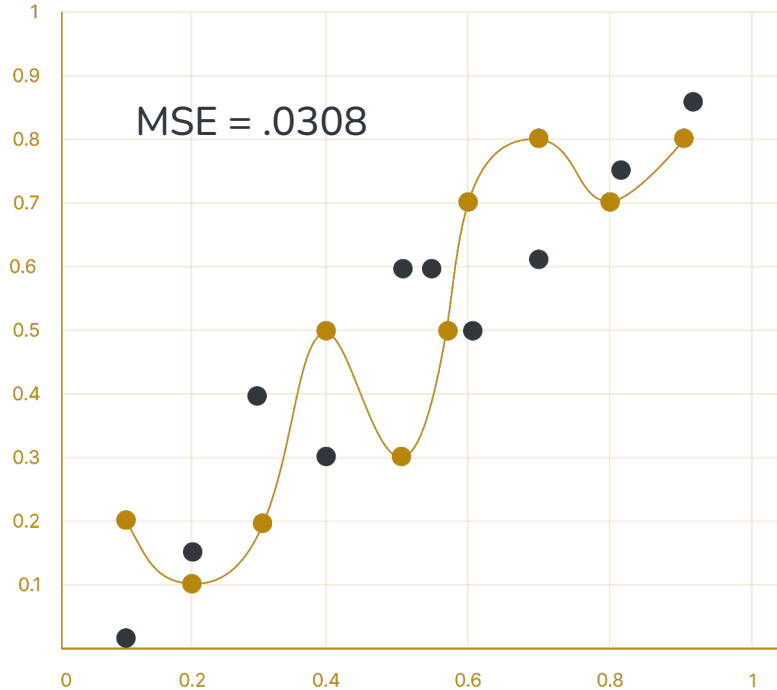
$$\nabla MSE(\theta) = \begin{pmatrix} \frac{\partial}{\partial \theta_0} \\ \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_n} \end{pmatrix} MSE(\theta)$$

$$\theta^{(next\ step)} = \theta - \lambda \nabla_{\theta} MSE(\theta)$$

Datenanalyse mit KI



Datenanalyse mit KI



Training data

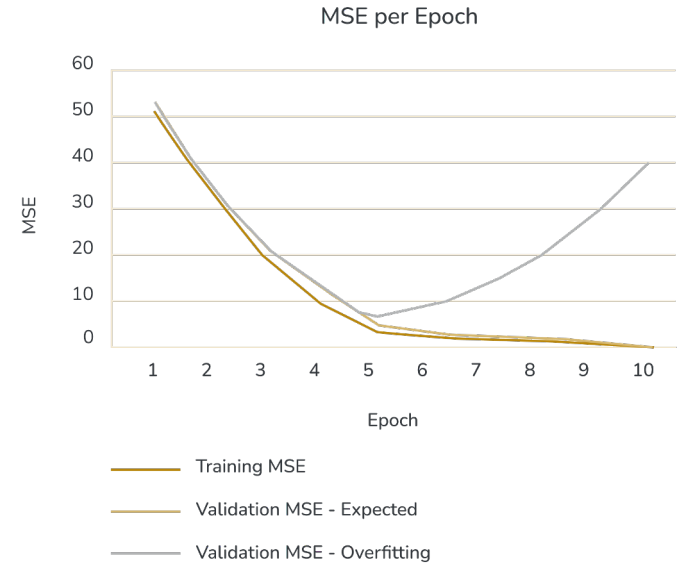
Datensatz zum Trainieren des Modells

Validation data

Validationsdatensatz um Overfitting auszuschließen

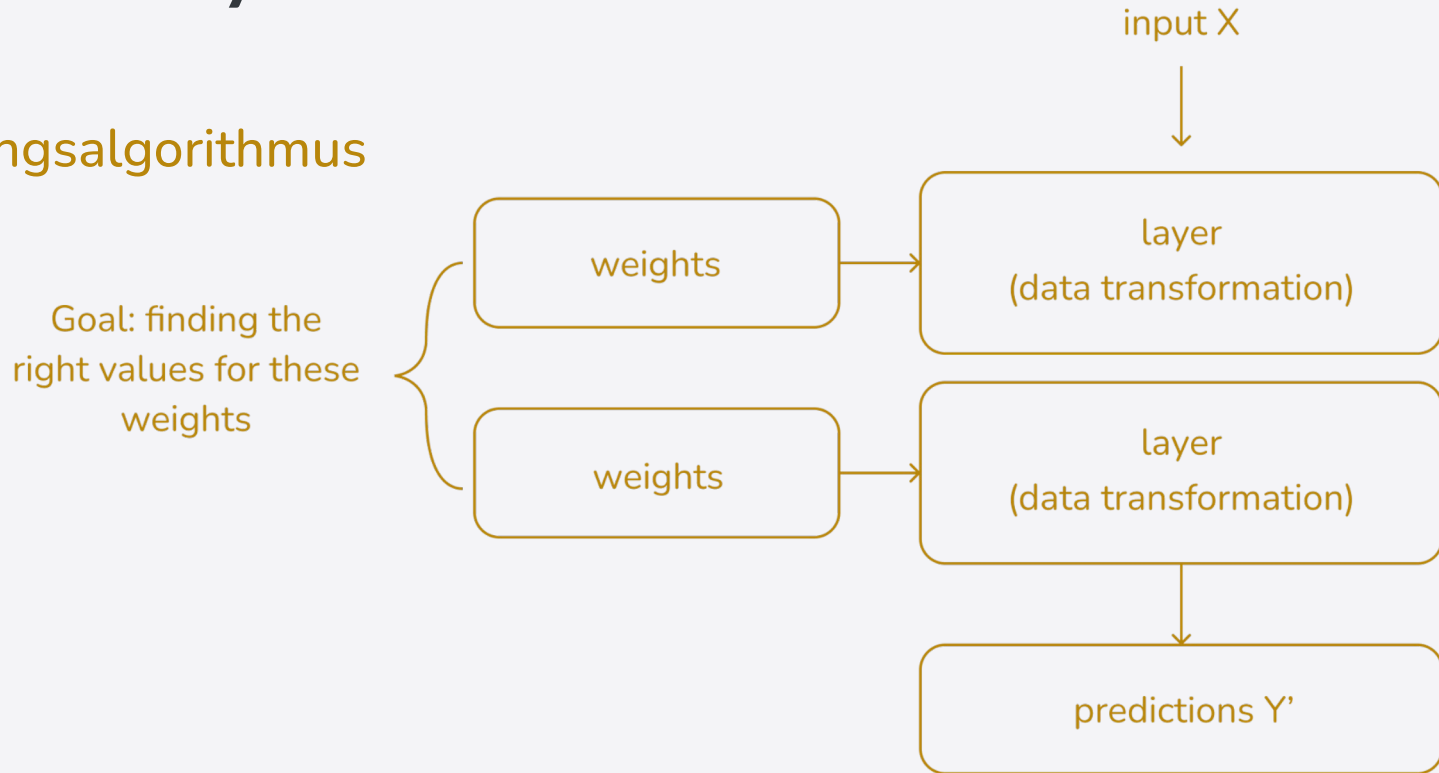
Test data

Testdatensatz wird vor Veröffentlichung angewandt

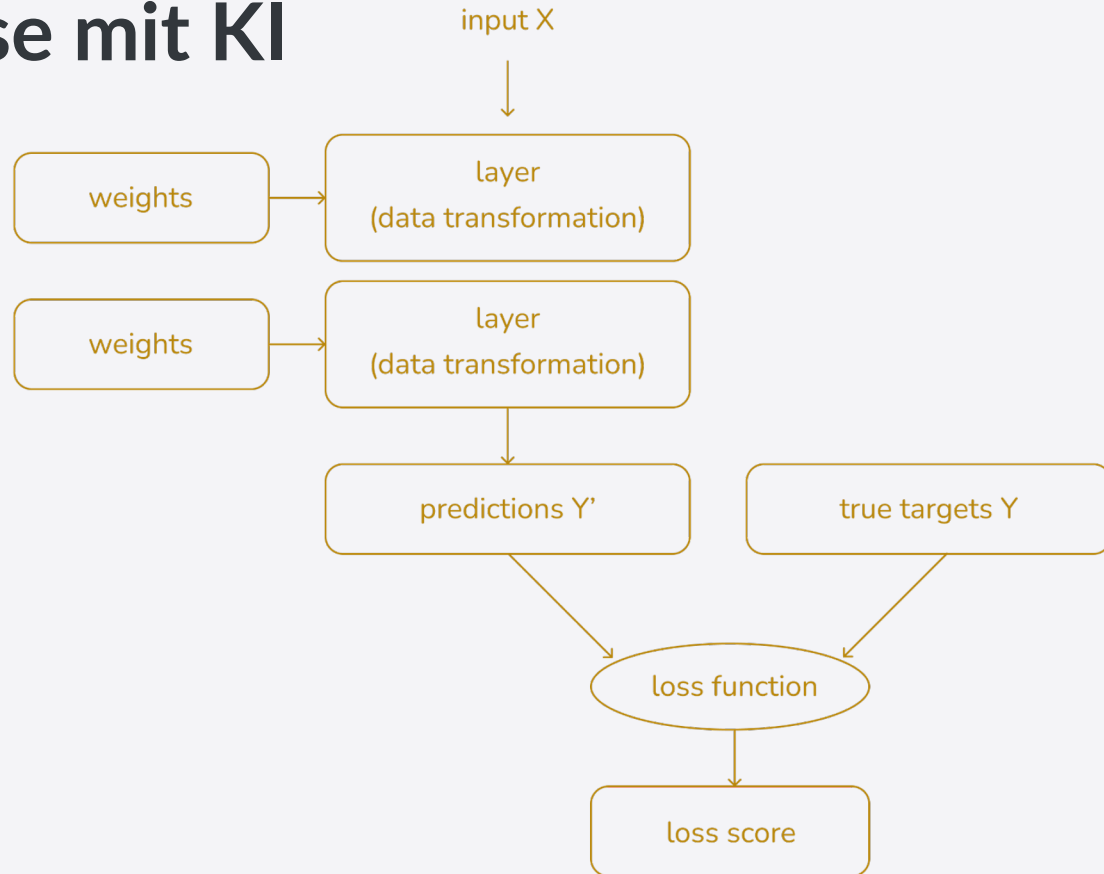


Datenanalyse mit KI

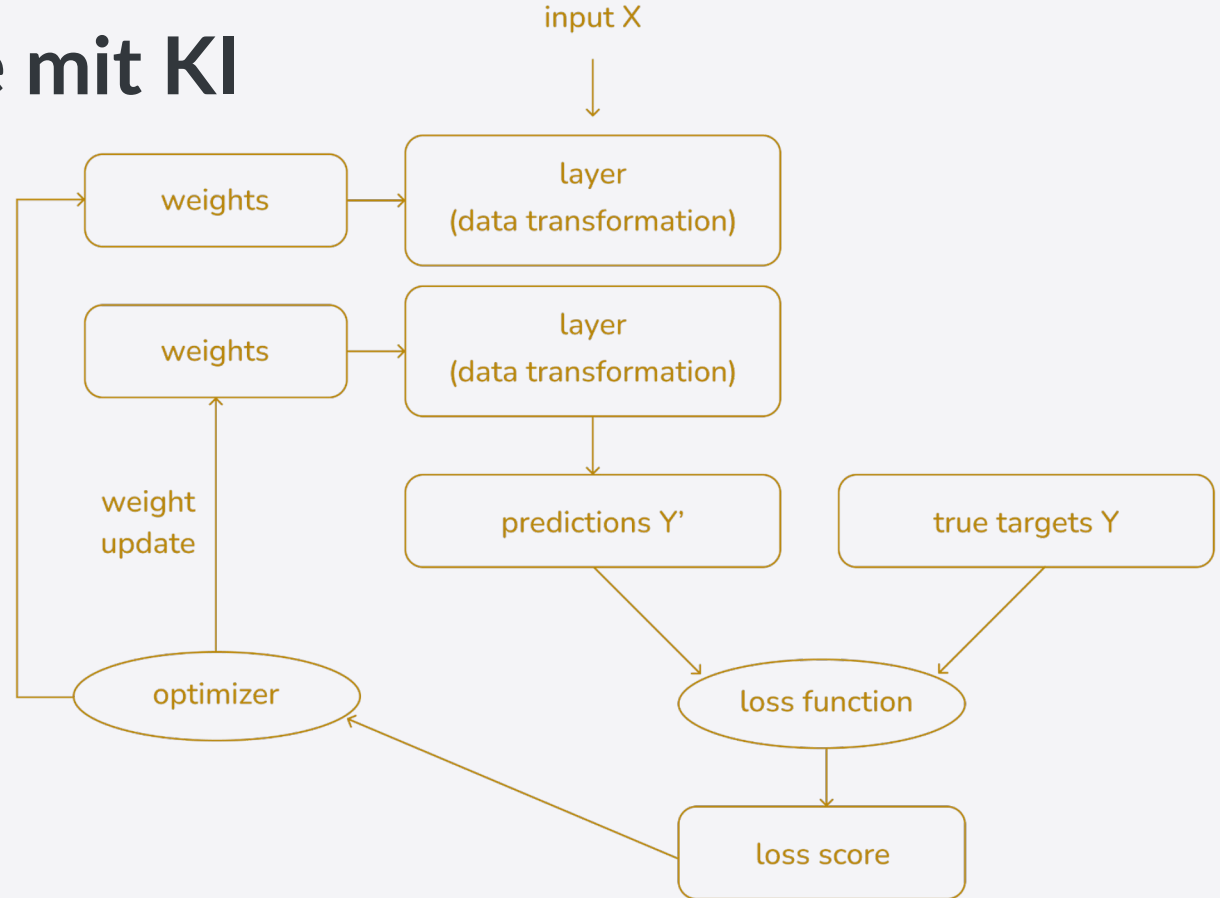
Trainingsalgorithmus



Datenanalyse mit KI



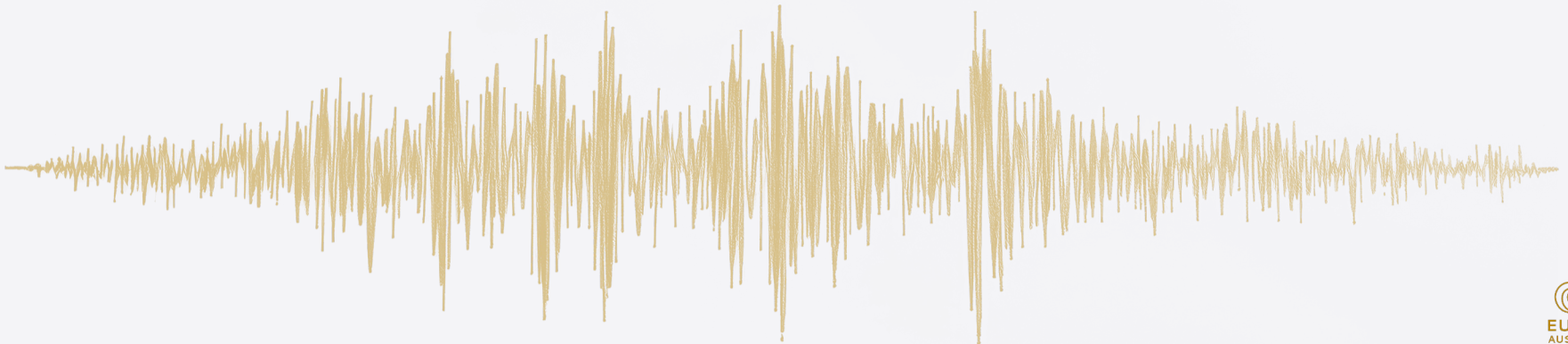
Datenanalyse mit KI



Zeitreihenanalyse

Was sind Zeitreihen?

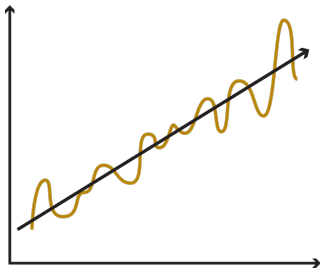
- Zeitlich geordnete Datenpunkte
- Zeitreihen sind **NICHT** unabhängig – aktuelle Werte hängen normalerweise von vorherigen ab
- Beispiele: Umsatz pro Woche
Temperatur pro Tag
Sensorwert pro Sekunde



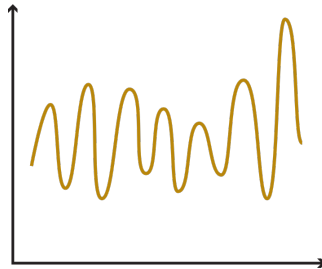
Zeitreihenanalyse

Eigenschaften von Zeitreihen

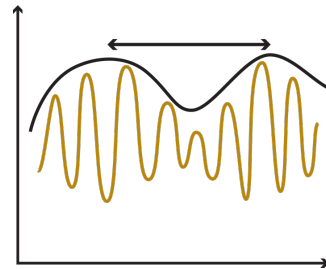
Trend:
langfristige Richtung



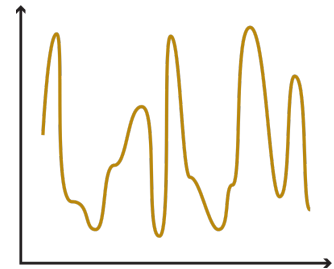
Saisonalität:
wiederkehrende
Muster



Zyklische
Schwankungen:
unregelmäßige
Veränderungen



Rauschen:
zufällige Ausschläge



Zeitreihenanalyse

Was wollen wir erreichen?

- Muster erkennen und Entscheidungen datenbasiert treffen
- Zukunft vorhersagen (forecasting)

z.B. Prognose von Verkaufszahlen

- Anomalien entdecken

z.B. Maschinenfehler



Zeitreihenanalyse

Warum sind Zeitreihen wichtig?

- Lawine an Zeitdaten von
Sensoren, IoT, Transaktionen, Logdaten...
- Zeitreihen entstehen in jeder Branche
z.B. Tourismus, Verkauf, Fertigung
- Infrastruktur & KI ermöglichen völlig neue Auswertung



Zeitreihenanalyse

Warum Zeitreihenanalyse mit KI?

Klassische Methoden funktionieren gut bei einfachen, regelmäßigen Mustern.

Probleme entstehen bei:

- Vielen parallelen Zeitreihen
- Unregelmäßigen Abständen oder Ausreißern
- Externen Einflussfaktoren



Zeitreihenanalyse

Statistische Methoden

Eigenschaften:

Gute Wahl bei klaren, saisonalen Mustern und Trends

Ideal für den Laptop

Keine Automatische Feature-Erkennung oder Externe Einflüsse

Typische Vertreter:

Seasonal Naive: Verwendet Wert der letzten Saison direkt als Prognose

ARIMA: aktuelle Werte werden durch vergangene Werte und Fehler erklärt

DynamicOptimizedTheta: Zerlegung der zeitreihe in Mittelwert, Trend, Saisonalität und Rauschen

Zeitreihenanalyse

ML Methoden

Eigenschaften:

Zeitreihe wird in Features umgewandelt

Schnell trainierbar

Gut interpretierbar

Typische Vertreter:

Recursive Tabular

Direct Tabular

Zeitreihenanalyse

Deep Learning Methoden

Eigenschaften:

Lernt Zeitreihen aus Rohdaten

Gut bei vielen ähnlichen Serien

Nutzt neuronale Netze (RNN, MLP)

Typische Vertreter:

DeepAR: autoregressives RNN

Chronos: leichtgewichtiges MLP-Modell

TiDE: MLP-Modell

N-BEATS: MLP-Modell

Zeitreihenanalyse

Transformer Methoden

Eigenschaften:

State-of-the-art Architektur aus NLP

Stark bei langen Abhängigkeiten

Nutzen Attention statt rekursiver Verbindungen

Typische Vertreter:

TemporalFusionTransformer:
hochkomplexe Architektur bestehend aus LSTM-, Attention- und MLP Blöcken

PatchTST: verarbeitet Zeitreihen in Patches

TiDE: MLP mit Zeit-Fokus

Modell	Kategorie	Externe	Multi-Input	Multi-Output	Laptop	GPU	HPC
Seasonal Naive	Klassisch	Nein	Nein	Nein	✓	✗	
AutoETS	Klassisch	Nein	Nein	Nein	✓	✗	
DynamicOptimTheta	Klassisch	Nein	Nein	Nein	✓	✗	
NPTS	Klassisch	Nein	~	Nein	✓	✗	
ARIMA / SARIMA	Klassisch	~	Nein	Nein	✓	✗	
Facebook Prophet	Klassisch	Ja	~	Nein	✓	✗	
RecursiveTabular	ML	Ja	Ja	Nein	✓	✗	
DirectTabular	ML	Ja	Ja	Ja	✓	✗	
Chronos	Deep Learning	Ja	Ja	~	✓	~	
TiDE	Deep Learning	Ja	Ja	Ja	⚠	✓	⚠
DeepAR	Deep Learning	Ja	Ja	Nein	⚠	✓	⚠
N-BEATS	Deep Learning	Nein	Ja	Nein	⚠	✓	⚠
PatchTST	Transf	Nein	Ja	Nein	⚠	✓	✓
TFT	Transf	Ja	Ja	Ja	⚠	✓	✓

Zeitreihenanalyse

AutoML für Zeitreihen

Eigenschaften:

Kombiniert die genannten Methoden

Automatische

- Feature Auswahl
- Hyperparameter-Tuning
- Modellvergleich

Ideal für Einsteiger und Produktion

Alle Modelle werden dabei getestet und eine Auswahl der besten zu einem Ensemble kombiniert.

AutoGluon lässt sich lokal starten und später auf HPC skalieren

Open Source – kostenfrei nutzbar

Unsupervised Learning - Clustering

Was ist Clustering?

Unüberwachtes Lernverfahren zur Gruppierung ähnlicher Datenpunkte

Kein Vorwissen über Klassen nötig

Ziel: „ähnliches zueinander, verschiedenes getrennt“

Arbeitet häufig mit Distanzmaßen

Typische Anwendung:
Kundensegmentierung, Anomalieerkennung



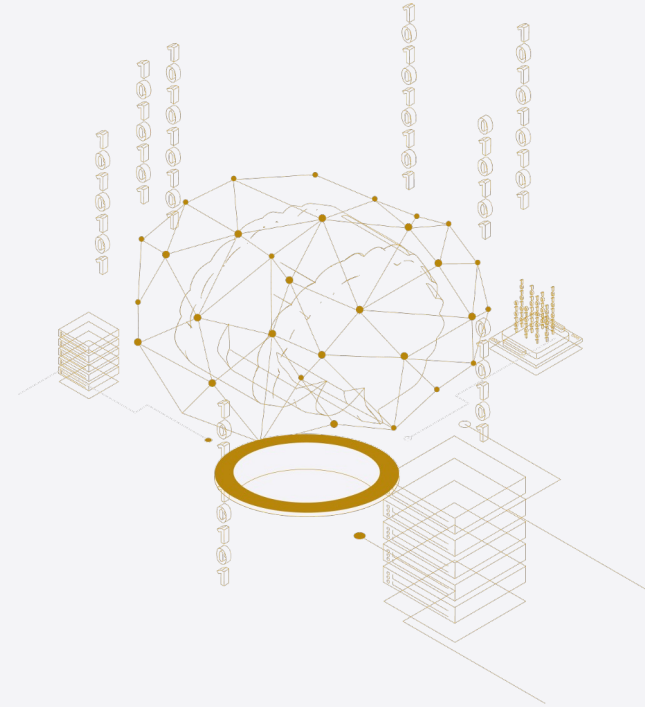
Unsupervised Learning - Clustering

k-means & k-prototypes

k-means gruppiert nach minimalem Abstand zum Clusterzentrum (numerische Daten)

k-prototypes wie k-means, erweitert um kategoriale und binäre Daten

Anzahl der Cluster muss vorgegeben werden (k)



Unsupervised Learning - Clustering

DBSCAN

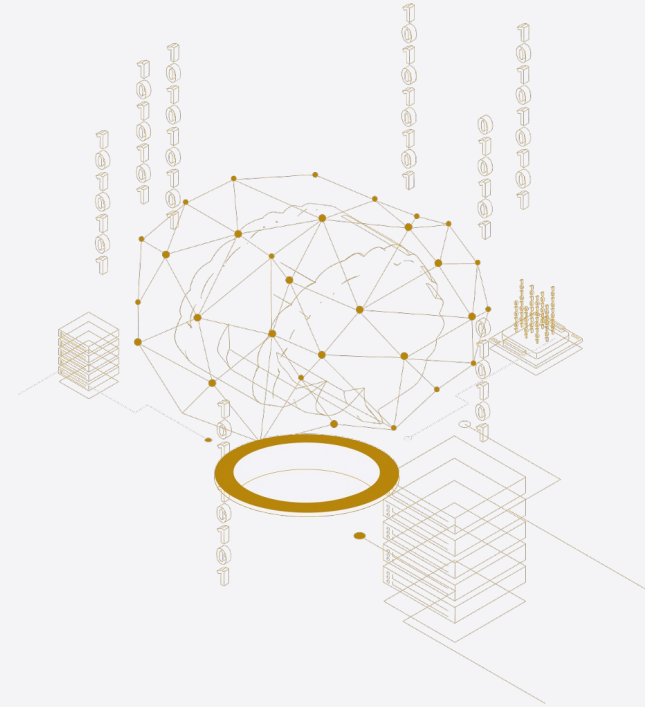
Dichtebasiertes Clustering – erkennt dichte Regionen

Keine Cluseranzahl (k) nötig. Anzahl der Cluster ergibt sich automatisch

Erfasst auch unregelmäßige Formen

Zählt Datenpunkte innerhalb eines vorgegebenen Abstandes (ϵ).

Parameter: ϵ , min_samples



Unsupervised Learning - Clustering

Distanzmaße & Metriken

Euklidische Norm zur Abstandsmessung numerischer Datenpunkte

Gower Distance zur Abstandsmessung von Datenpunkten gemischten Typs

Silhouette Score misst wie gut ein Punkt zu einem Cluster passt. Werte von -1 bis +1 (+1 perfekt gruppiert)

Clusterplots helfen in der Beurteilung



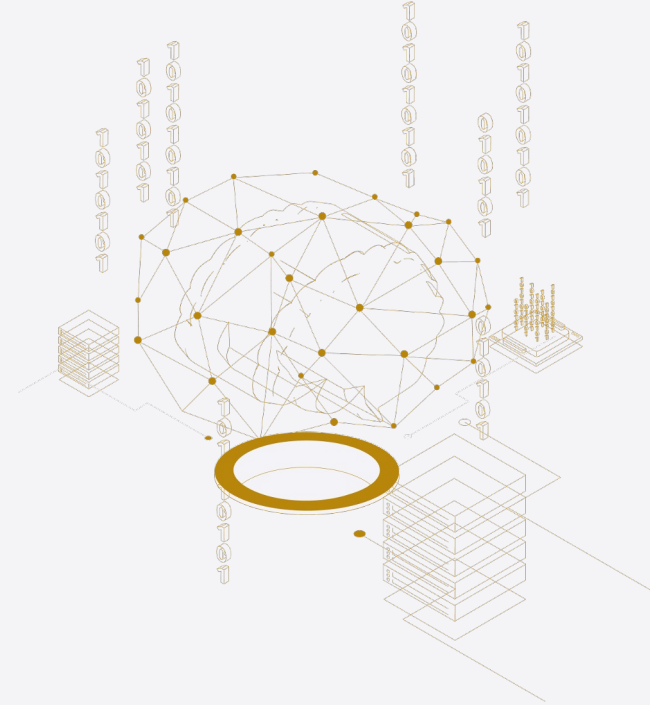
Unsupervised Learning - Clustering

Dimensionsreduktion

Dient dazu hochdimensionale Daten in 2D oder 3D abzubilden. Dadurch wird eine visuelle Beurteilung eventuell vorhandener Cluster ermöglicht.

UMAP: Methode, die Datenpunkte in hoher Dimension mit Graphenstruktur verbindet. Globale Distanzen werden möglichst getreu in der niedrigen Dimension abgebildet.

PCA: Lineare, rechengünstige Methode



Unsupervised Learning - Clustering

Autoencoder

Ein Autoencoder ist ein Deep Learning Modell, das zur Erstellung numerischer Repräsentationen (Embeddings) dient.

Wird als Vorstufe zum Clustering verwendet (Dimensionsreduktion)

Anwendung in der Anomalieerkennung und extraktion relevanter Merkmale (Feature Learning)



STAY IN TOUCH



eurocc-austria.at



vsc.ac.at



THANK YOU



This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101101903. The JU receives support from the Digital Europe Programme and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Türkiye, Republic of North Macedonia, Iceland, Montenegro, Serbia