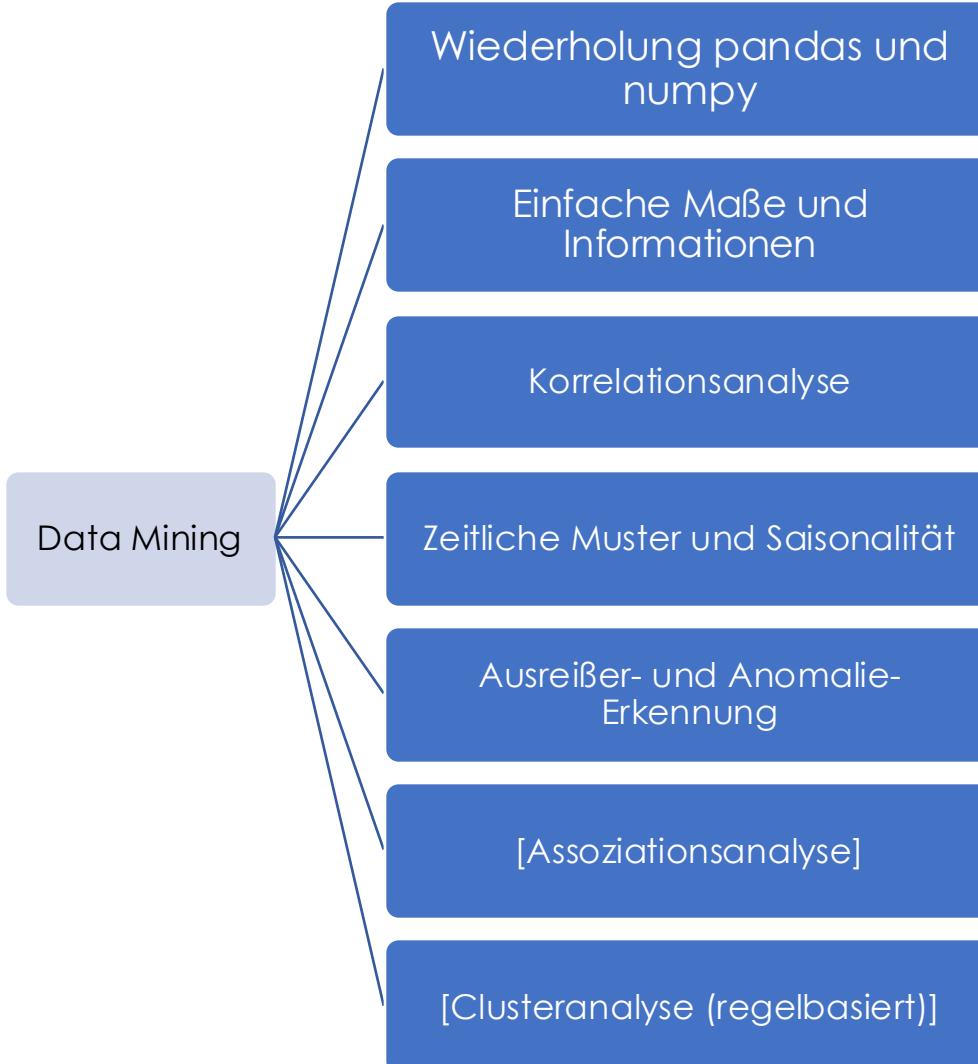


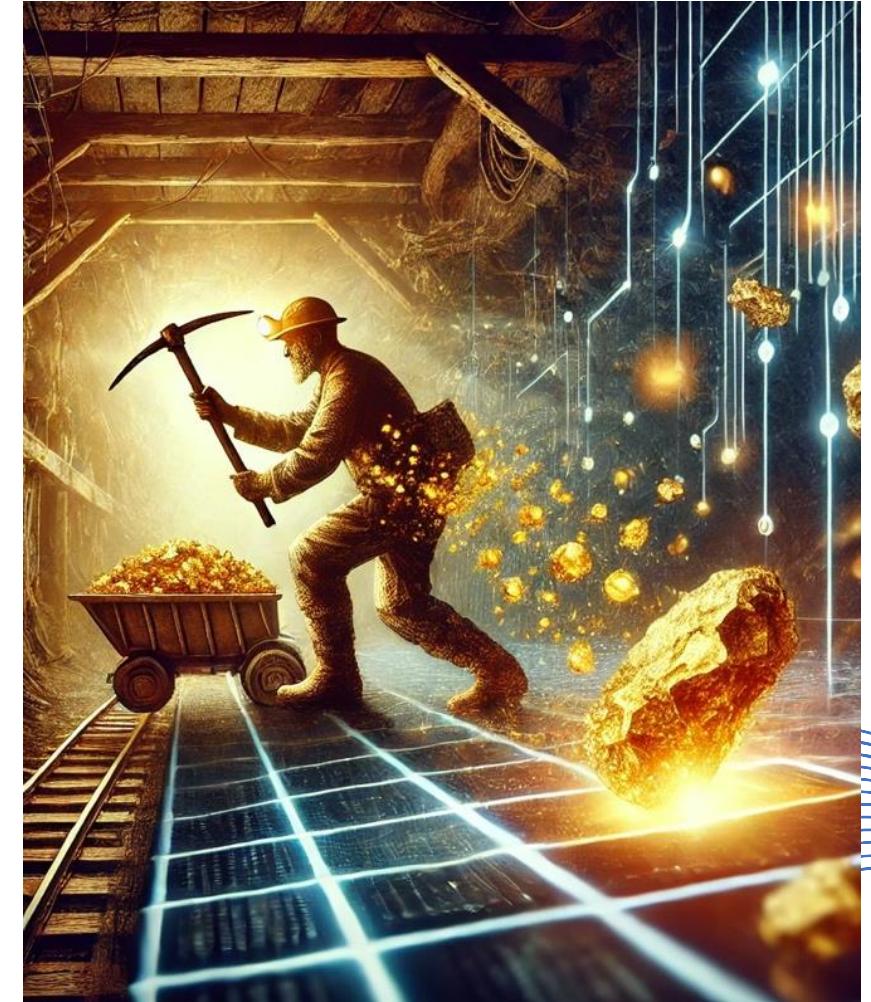
# Agenda

---

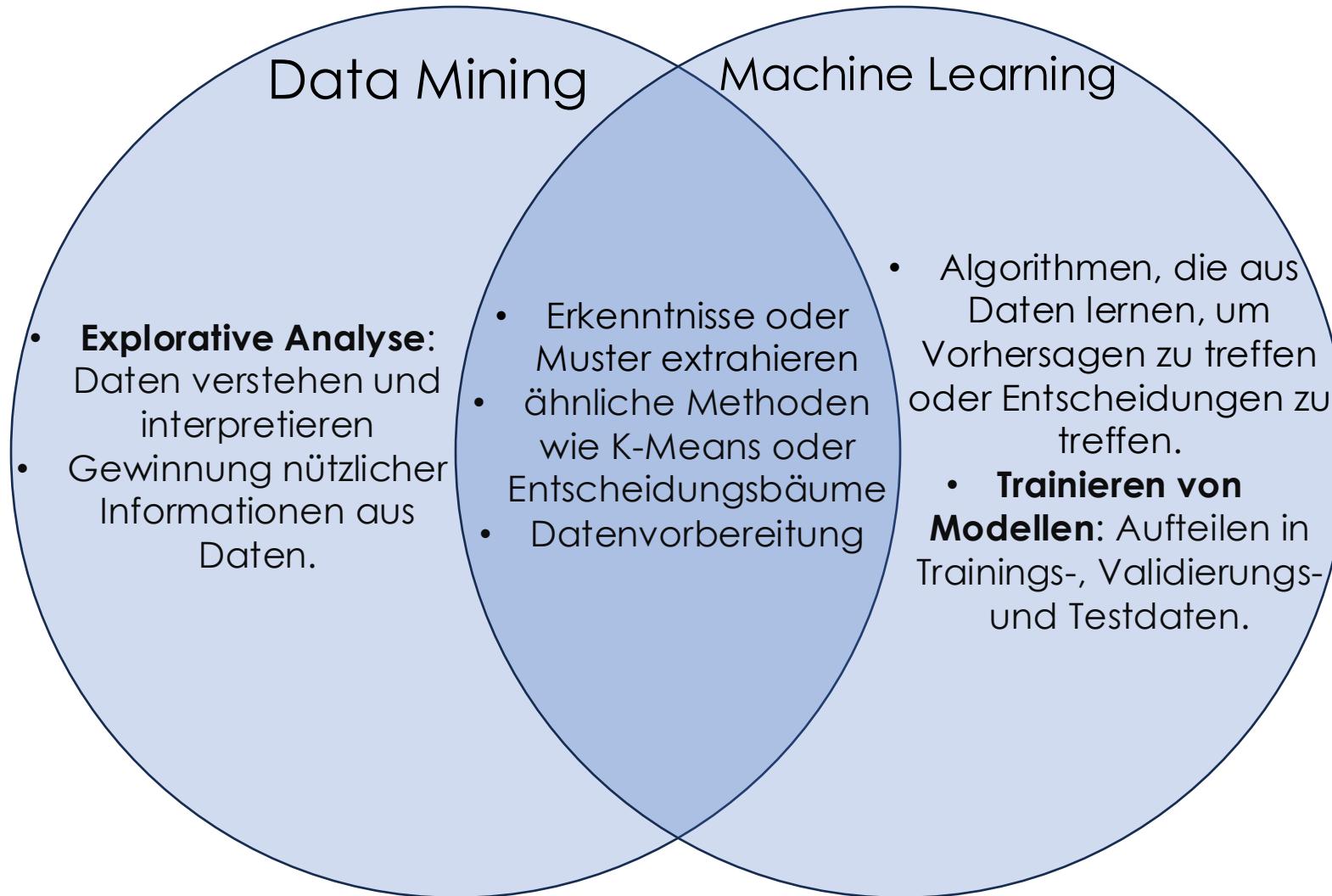


# Data Mining vs. Data Analytics

- **Data Mining** zielt darauf ab, **verborgene Muster und Beziehungen** in großen Datenmengen zu entdecken. Der Fokus liegt auf der Entdeckung neuer Erkenntnisse, die **nicht unmittelbar offensichtlich** sind, und wird oft als Explorationsprozess betrachtet.
- **Data Analytics** hingegen konzentriert sich stärker auf die Analyse und Interpretation von Daten zur **Beantwortung spezifischer Fragen oder zur Entscheidungsfindung**. Es geht oft darum, **bekannte Informationen besser zu verstehen**, Trends zu erkennen und basierend auf den Daten Erkenntnisse für GeschäftSENTscheidungen zu gewinnen.

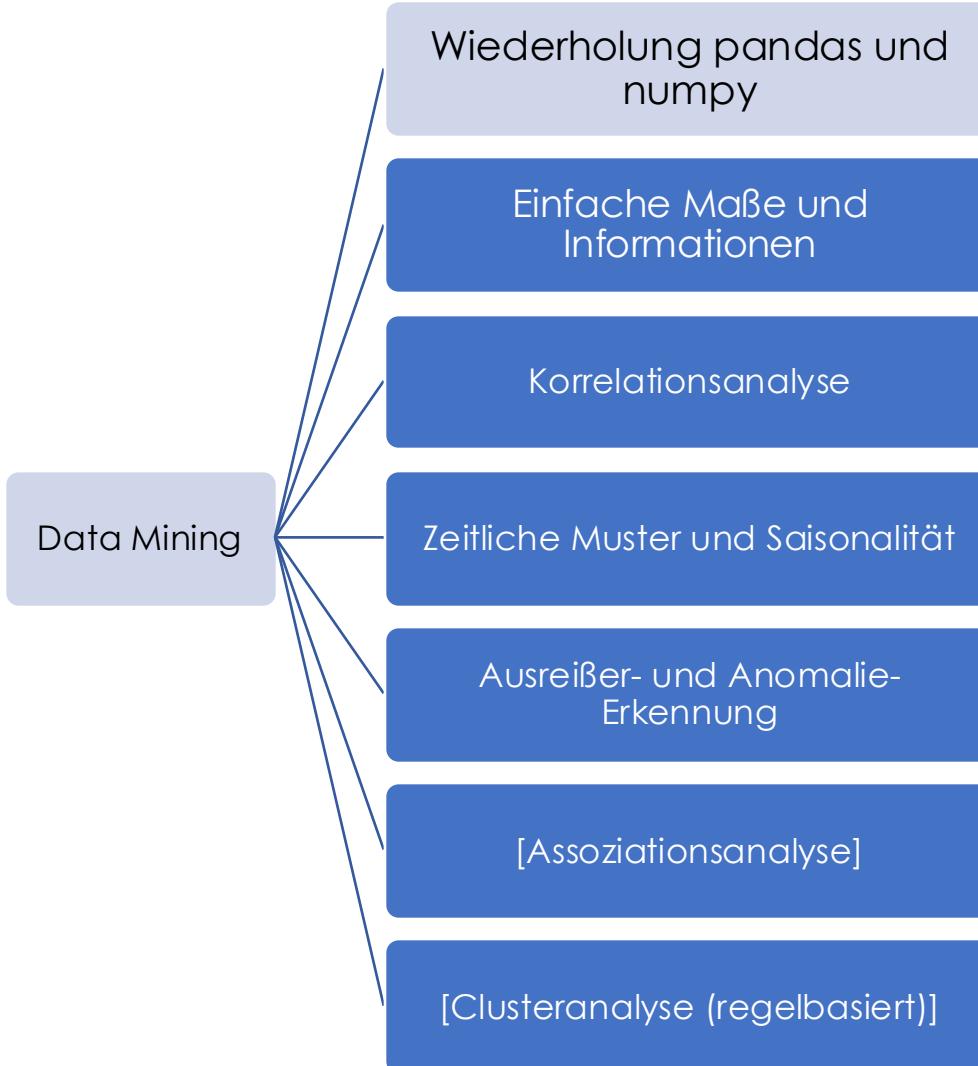


# Data Mining und Machine Learning



# Agenda

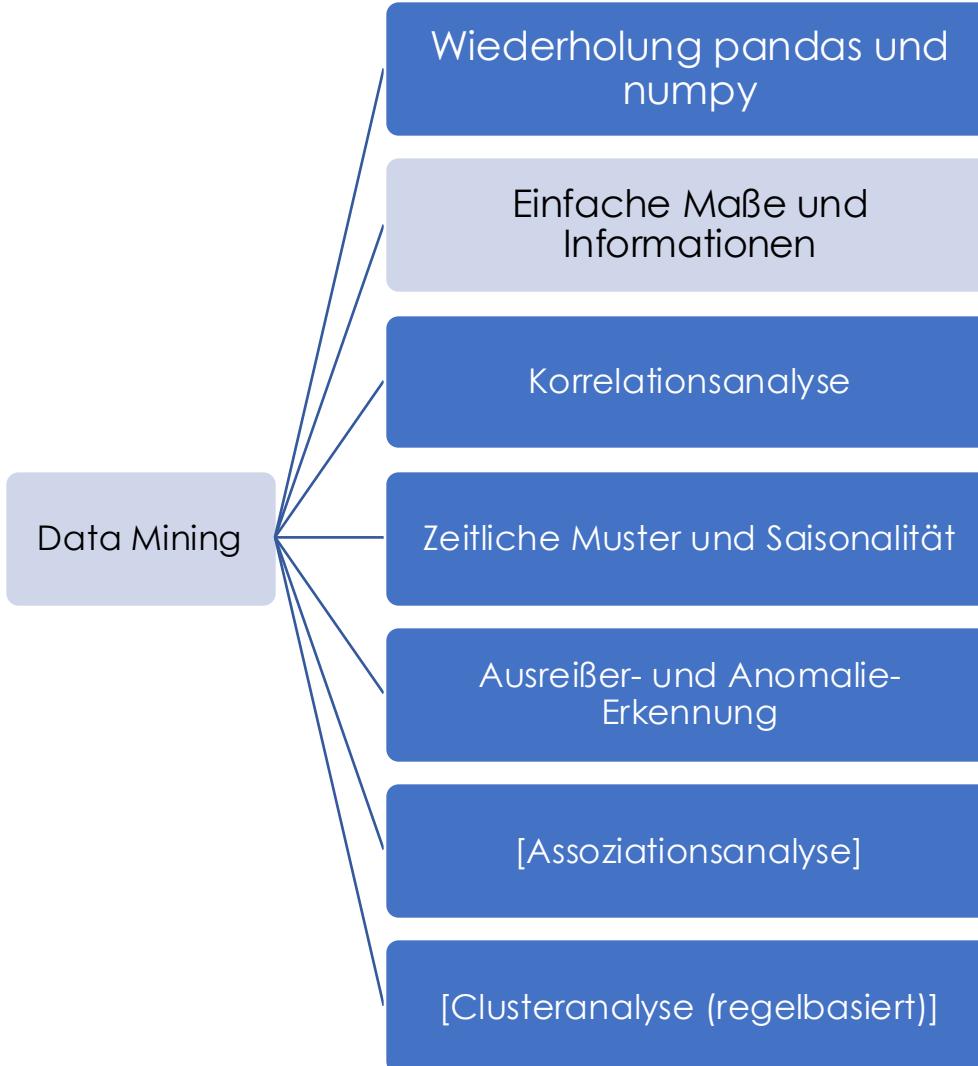
---



# Demo mit Bitcoin Daten

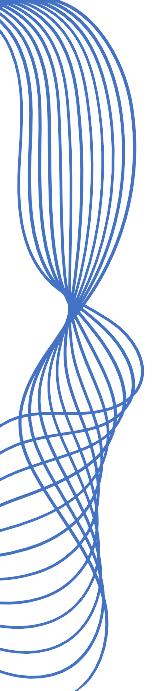
# Agenda

---



# Einfache Maße und Informationen

---



Mittelwert / Durchschnitt

Standardabweichung  
= Streuungsmaß



Median

= nach Sortierung der mittlere Wert

Perzentile und Quartile

= Wertschranke nach Sortierung, z.B.  
Quartil: 25% der Werte unterhalb der Wertschranke

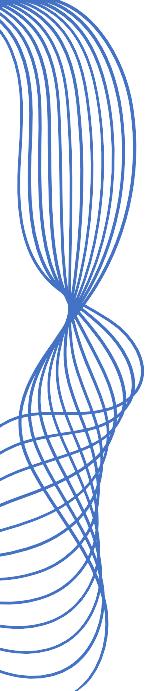
Korrelation

= Zusammenhang zwischen zwei Variablen

Fehlende Werte

# Einfache Maße und Informationen

---



Boxplots



Value Counts und  
Histogramme

Aggregationen

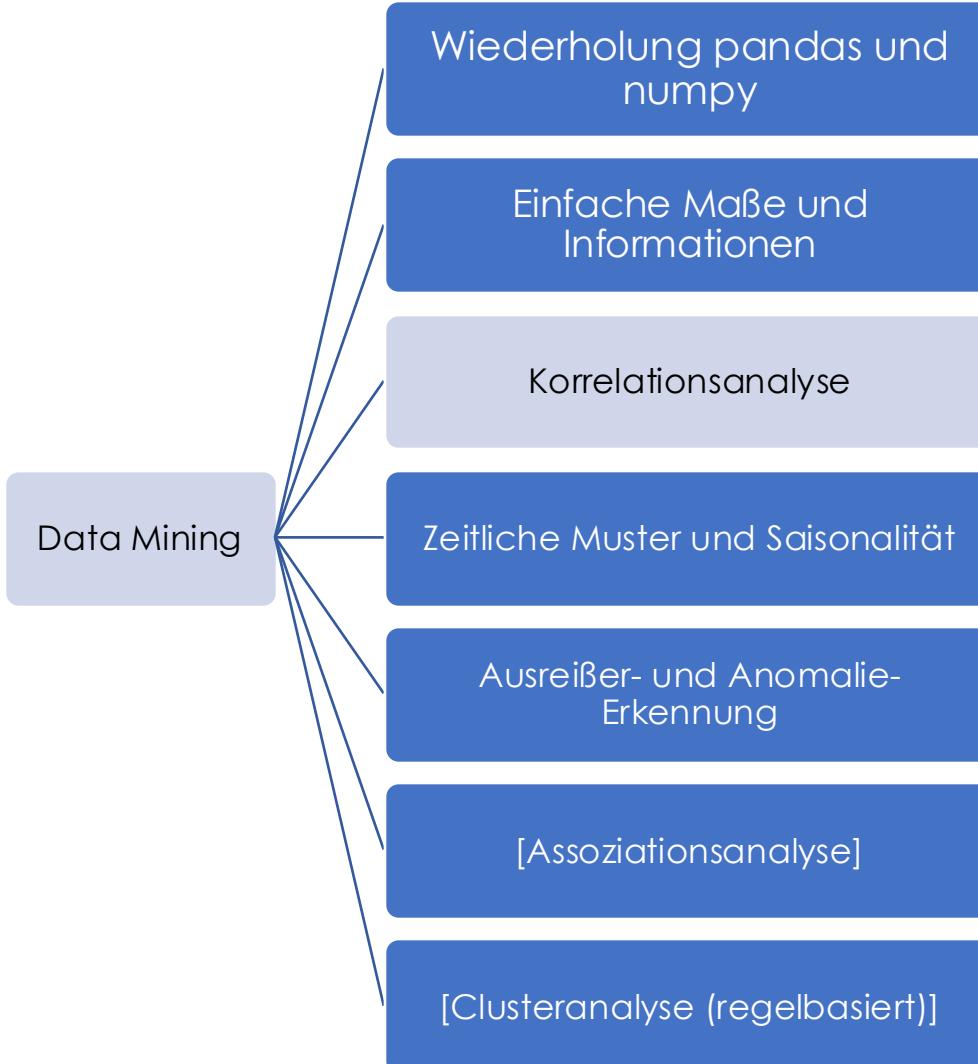
# Demo: Analyse der Wetterdaten

# Übung

- Wo in Australien muss ich hinziehen, wenn ich möglichst gleichbleibende Temperaturen (wenig Schwankungen) haben möchte?
  - Welches ist der kälteste Monat? Wann fällt am meisten Regen?
- Lässt sich eine Erhöhung der Temperatur im Laufe der Jahre feststellen?
  - Was gibt es noch für Muster?

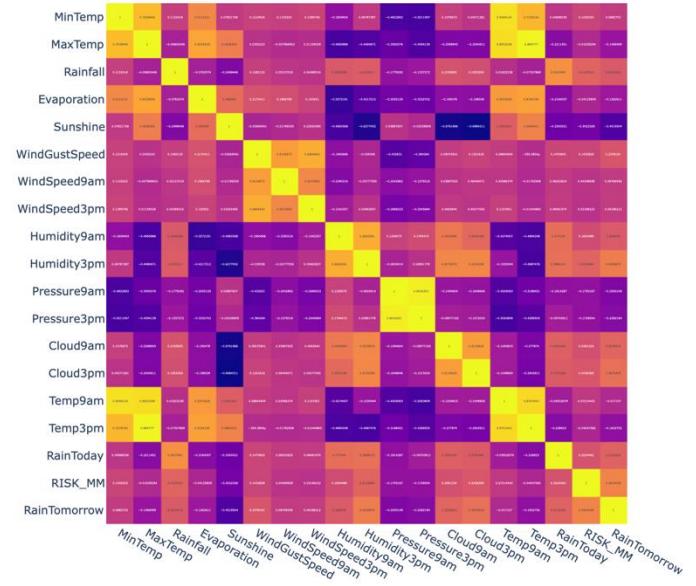
# Agenda

---



# Korrelationsanalyse

- Analyse des statistischen Zusammenhangs zwischen zwei oder mehr Variablen.
- Wertebereich: -1 bis +1
- Interpretation:
  - 1 = negativer Zusammenhang
  - 0 = kein Zusammenhang
  - +1 = positiver Zusammenhang





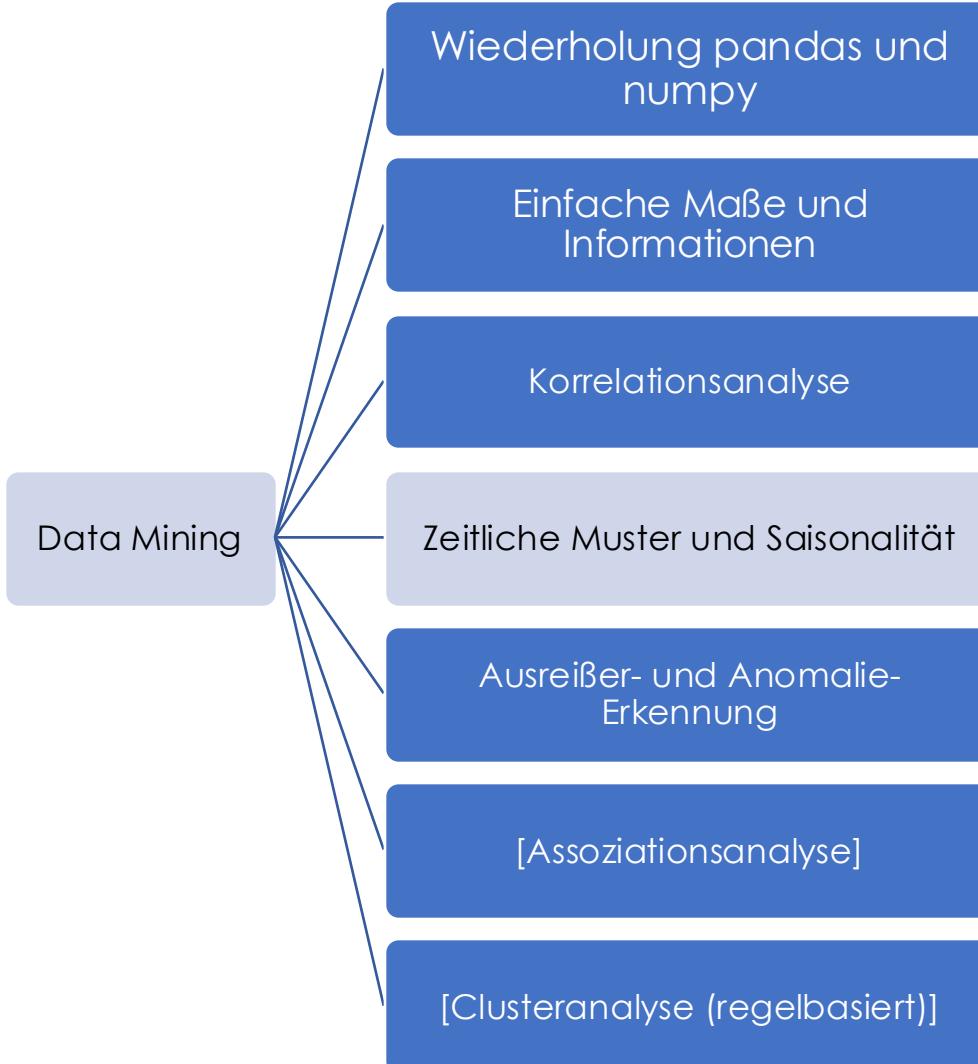
# Demo: Korrelationsanalyse der Wetterdaten

# Übung

Erstelle eine Korrelationsanalyse für den Titanic Datensatz.  
Entferne Spalten, die nicht relevant sind und wandle  
gegebenenfalls Spalten in numerische Werte um.  
Welche Variable korreliert am stärksten mit „Survived“?

# Agenda

---



# Zeitliche Muster und Saisonalität

---

- Liniendiagramme
- Gleitende Durchschnitte
- Aggregationen auf Basis des Datums

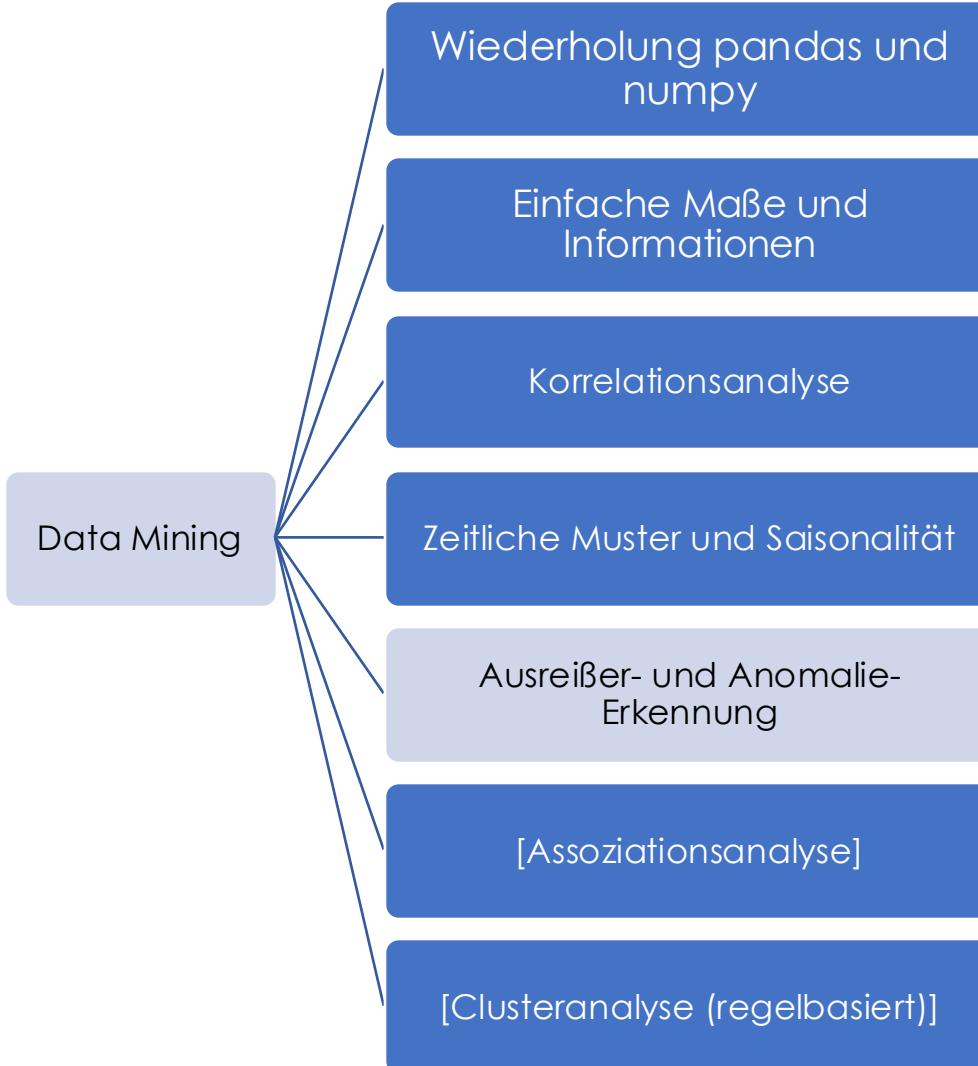
# Demo Saisonalität

# Übung

Wie unterscheidet sich die Temperatur je Jahreszeit (Winter, Frühling, Sommer, Herbst) in Australien?

# Agenda

---



# Ausreißer- und Anomalie Erkennung

---

= Datenpunkte, die erheblich von anderen abweichen.

## **Relevanz:**

- Erkennung von Fehlern, Betrug oder ungewöhnlichen Mustern.
- Verbessert Datenqualität und Entscheidungsfindung.

## **Methoden:**

- Statistische Ansätze: Z-Score, Interquartilsabstand.
- Maschinelles Lernen: Clustering (z. B. k-Means), Neuronale Netze.
- Zeitreihenanalyse: Erkennung von plötzlichen Veränderungen.

## **Herausforderungen:**

- Hohe Dimensionalität der Daten.
- Unterscheidung zwischen seltenen, aber legitimen Werten und echten Anomalien.

# Demo

Finde im Bitcoin Datensatz einen Tag mit ungewöhnlich hohem Volumen.

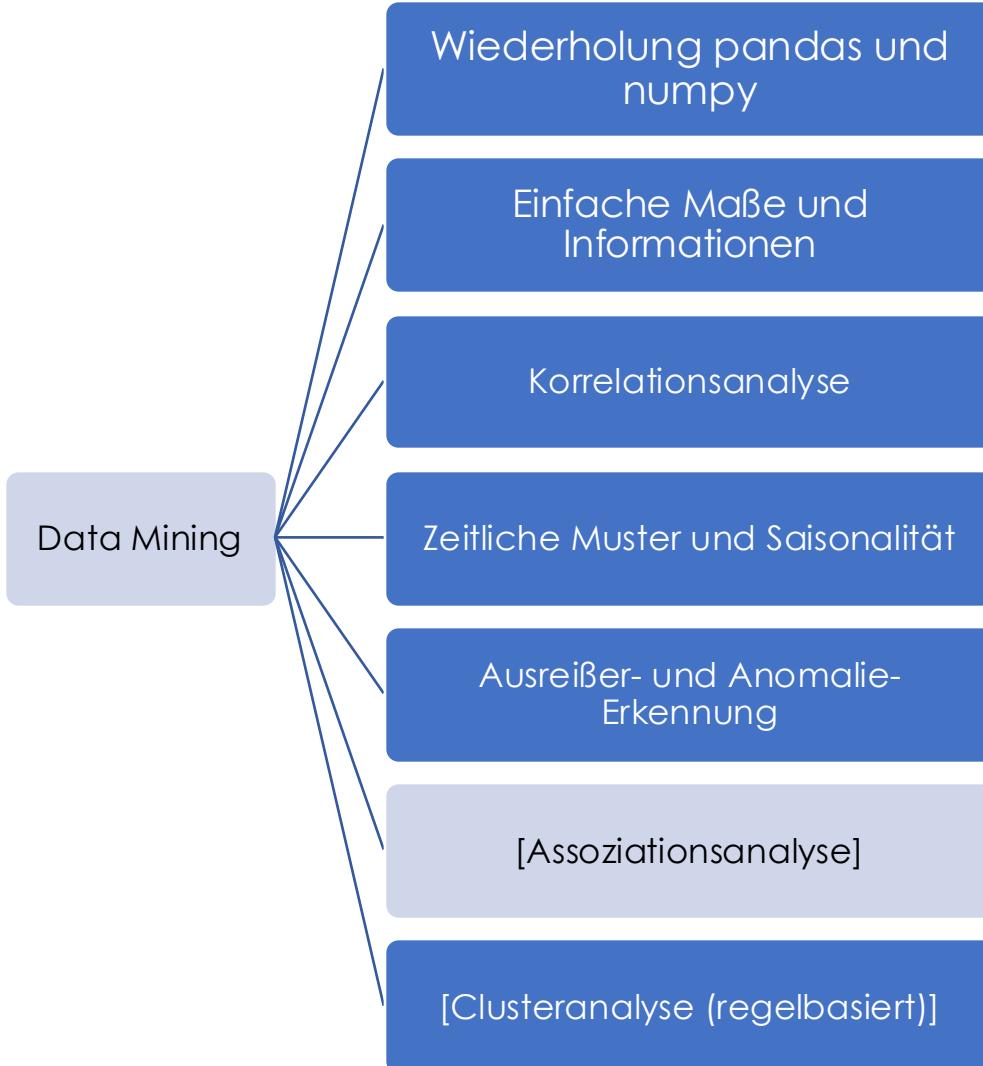
Herausforderung: Volumen nimmt im Laufe der Zeit zu.

# Übung

Im Wetter Datensatz ist in der Spalte „MinTemp“ ein fehlerhafter Messwert. An welchem Tag und in welchem Ort?

# Agenda

---



# Assoziationsanalyse

- Ziel: Identifikation von häufig auftretenden Item-Kombinationen (z. B. in Warenkörben).
- Grundlage: „Wenn-Dann“-Regeln (z. B. „Wenn Artikel A, dann Artikel B“).
- Anwendung in Marktanalyse, Empfehlungsalgorithmen, Betrugserkennung.



# Assoziationsanalyse

Kernmetriken in der Assoziationsanalyse:

- Support: Häufigkeit eines Itemsets in den Transaktionen.
- Confidence: Wahrscheinlichkeit, dass Item B gekauft wird, wenn Item A gekauft wurde.
- Lift: Stärke der Regel im Vergleich zur Zufallswahrscheinlichkeit.

## Beispiel:

T1: {Milch, Brot, Butter}  
T2: {Brot, Käse}  
T3: {Milch, Brot, Butter, Käse}  
T4: {Milch, Butter}  
T5: {Brot, Butter}

- Mindest-Support: 60%
- Häufige Itemsets:
  - {Brot, Butter} (Support: 60%)
  - {Milch, Butter} (Support: 60%)
- Regel: {Milch} → {Butter} (Confidence: 100%, Lift: 1.67)

# Demo + Übung

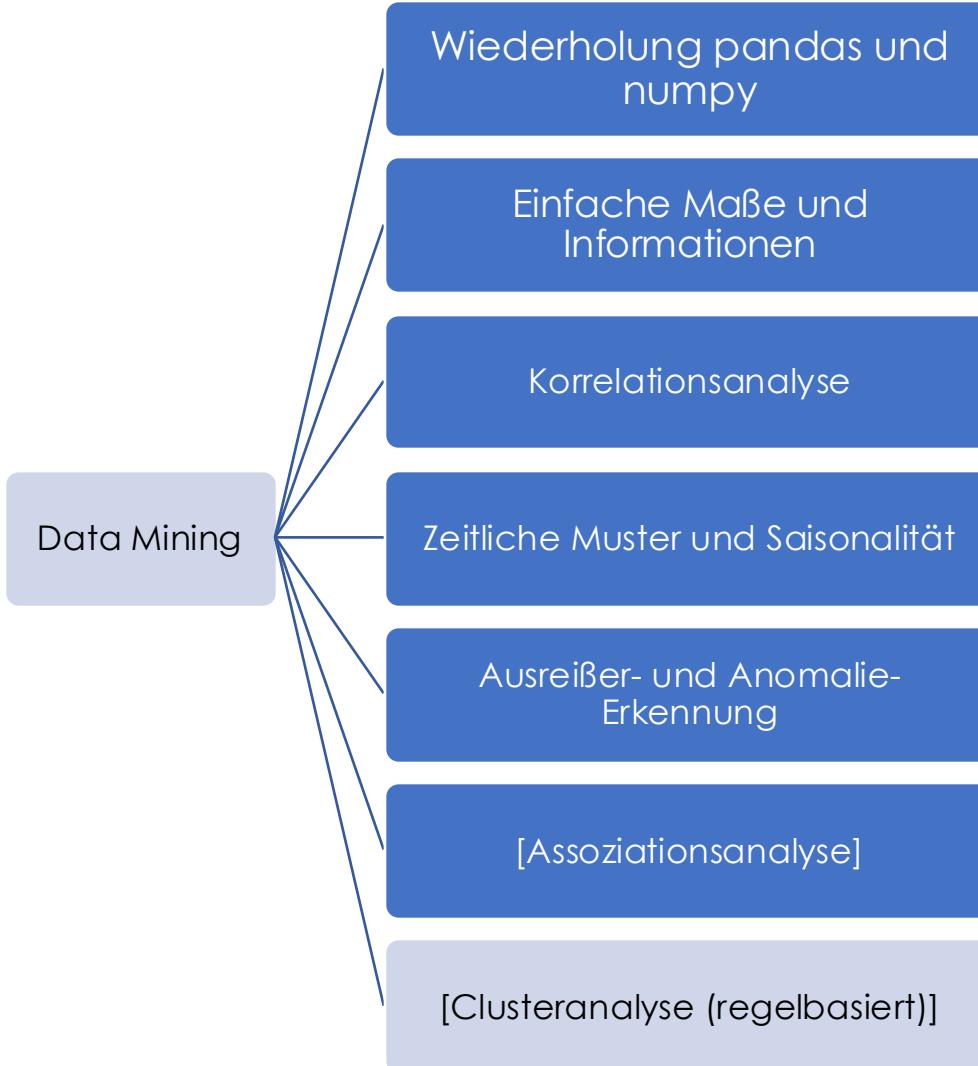
- Wie verändern sich die Regeln nach dem Hinzufügen von T6 = [Milch, Butter]
- Wie verändern sich die Regeln nach dem Verdoppeln der Transaktionslisten? Wieso?

# Sinnvoller Umgang mit Ausreißern

---

- Ist der Wert valide oder ein Fehler?
- Nachverfolgen des Grunds (um zukünftige Fehler zu vermeiden)
- Werte verbessern oder löschen
- Falls keine Daten gelöscht werden können: Imputation, z.B. Ersetzen mit Durchschnittswert

# Agenda



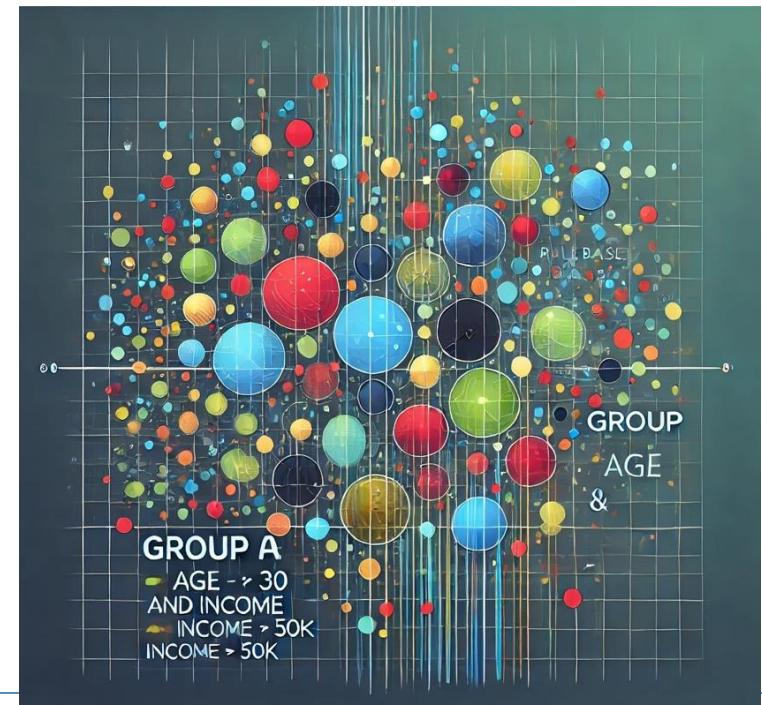
# Clusteranalyse (regelbasiert)

= Analyse von Daten basierend auf vordefinierten Regeln zur Gruppierung

- Nutzt logische Bedingungen wie „Alter > 30 UND Einkommen < 50.000“.

## Anwendungsbeispiele:

- Kundenklassifizierung im Marketing.
- Betrugserkennung in Finanzsystemen.



# Demo: Analyse Auto Daten

- Unterteilung der Autos in Klassen und Analyse der Eigenschaften

# Übung

Unterteile die Autos in 3 Cluster nach PS (Spalte „HP“).  
Welche Unterschiede stellst du fest?

# Ausblick – folgende Schulungen

---

Schulung	Datum	Ort
Python – Machine Learning	06.01.25 – 08.01.25	Köln
Deep Learning und Text Mining – Textanalyse und - generierung	20.01.25 – 24.01.25	Köln

# Ausblick – wie weiter lernen?

---

- Kaggle: <https://www.kaggle.com>
- Datensätze downloaden und analysieren – Übung macht den Meister: <https://archive.ics.uci.edu>

kaggle





# Ende

Bei Rückfragen: [nk@data-convolution.de](mailto:nk@data-convolution.de)

# Glossar

# Glossar I

---

**ACID:** ACID steht für Atomarität, Konsistenz, Isolation und Dauerhaftigkeit und beschreibt die Eigenschaften von Transaktionen in Datenbanken.

- Atomarität: Operationen sind unteilbar und werden entweder vollständig oder gar nicht ausgeführt.
- Konsistenz: Die Datenbank bleibt vor und nach einer Transaktion in einem konsistenten Zustand.
- Isolation: Gleichzeitige Transaktionen beeinflussen sich nicht gegenseitig.
- Dauerhaftigkeit: Änderungen bleiben nach Abschluss einer Transaktion auch bei Systemausfällen erhalten.

**Assoziationsanalyse:** Die Assoziationsanalyse dient der Identifizierung von häufig auftretenden Item-Kombinationen (z. B. in Warenkörben). Sie basiert auf "Wenn-Dann"-Regeln, die die Wahrscheinlichkeit angeben, dass ein bestimmtes Item gekauft wird, wenn ein anderes Item bereits gekauft wurde.

**Asynchrones Programmieren:** Asynchrones Programmieren ermöglicht die gleichzeitige Bearbeitung mehrerer Aufgaben, ohne auf die Fertigstellung einzelner Aufgaben warten zu müssen. Dies führt zu einer erhöhten Performance und Reaktionsfähigkeit von Anwendungen, insbesondere bei I/O-intensiven Operationen wie Datenbankzugriffen oder Netzwerkkommunikation.

**Ausreißer- und Anomalie-Erkennung:** Ausreißer und Anomalien sind Datenpunkte, die von den übrigen Daten abweichen. Sie können auf Messfehler, Betrug oder ungewöhnliche Ereignisse hinweisen.

# Glossar II

---

**Authentifizierung:** Authentifizierungsmechanismen dienen dazu, den Zugriff auf APIs zu kontrollieren und zu sichern. Sie stellen sicher, dass nur autorisierte Benutzer auf die Ressourcen der API zugreifen können. Verschiedene Methoden wie API-Schlüssel, Basic Authentication, OAuth 2.0 und Bearer Tokens können verwendet werden.

**Backups und Wiederherstellung:** Backups und Wiederherstellungsmechanismen sind essenziell für die Sicherheit und Verfügbarkeit von Daten. Backups erstellen Kopien der Daten, die im Falle eines Datenverlustes wiederhergestellt werden können. Die Wiederherstellung kann mithilfe verschiedener Methoden erfolgen, beispielsweise durch Point-in-Time Recovery (PITR), bei der die Datenbank zu einem bestimmten Zeitpunkt wiederhergestellt wird.

**Clusteranalyse:** Die Clusteranalyse dient der Gruppierung von Datenpunkten in Cluster, die ähnliche Eigenschaften aufweisen. Es gibt verschiedene Verfahren der Clusteranalyse, darunter regelbasierte Verfahren, die auf vordefinierten Regeln basieren, und Machine-Learning-Verfahren, die die Cluster automatisch aus den Daten lernen.

**CRUD:** CRUD steht für Create, Read, Update, Delete und beschreibt die vier grundlegenden Operationen, die in den meisten Datenbanksystemen und APIs verwendet werden.

**Data Analytics:** Data Analytics konzentriert sich auf die Analyse und Interpretation von Daten, um konkrete Fragen zu beantworten oder fundierte Entscheidungen zu treffen. Ziel ist es, bereits vorhandene Informationen besser zu verstehen, Trends zu erkennen und datengestützte Entscheidungen zu treffen.

# Glossar III

---

**Data Mining:** Data Mining befasst sich mit dem Aufspüren verborgener Muster und Zusammenhänge in umfangreichen Datensätzen. Der Schwerpunkt liegt auf der Gewinnung neuer Erkenntnisse, die nicht auf den ersten Blick erkennbar sind. Dieser Prozess wird oft als explorativ bezeichnet. Data Mining verwendet ähnliche Verfahren wie Machine Learning, zum Beispiel K-Means oder Entscheidungsbäume. Es dient auch der Vorbereitung von Daten für Machine Learning.

**Data Storing und -Zugriff:** Dieser Bereich umfasst verschiedene Methoden zur Speicherung und zum Zugriff auf Daten:

- Einfache Dateiformate: CSV, JSON, Protobuf, XML, YAML, Avro, Thrift.
- Datenbanken: ZODB, SQL-Datenbanken (PostgreSQL, SQLite), NoSQL-Datenbanken (MongoDB).
- APIs: Schnittstellen zur Kommunikation zwischen Anwendungen.

**Datenbank-Skalierung:** Die Skalierung von Datenbanken ermöglicht es, die Leistung und Verfügbarkeit von Datenbanken zu verbessern, indem die Daten auf mehrere Server verteilt werden. Dies kann durch Partitionierung (Aufteilung einer Tabelle in kleinere Teile) oder Replikation (Kopieren der Datenbank auf mehrere Server) erreicht werden.

**Dependency Injection:** Dependency Injection ist ein Entwurfsmuster, das die Entkopplung von Softwarekomponenten fördert, indem Abhängigkeiten von außen bereitgestellt werden. Dies erhöht die Modularität, Flexibilität und Wartbarkeit von Software und erleichtert das Testen von Komponenten.

**Fehlende Werte:** Fehlende Werte sind ein häufiges Problem in Datensätzen. Sie können verschiedene Ursachen haben, z. B. Messfehler, fehlende Informationen oder bewusstes Auslassen von Daten.

# Glossar III

---

**gRPC:** gRPC ist ein Framework für Remote Procedure Calls (RPC), das eine effiziente Kommunikation zwischen Systemen ermöglicht. gRPC verwendet Protocol Buffers (Protobuf) zur Serialisierung von Daten und unterstützt verschiedene Programmiersprachen.

**Indizes:** Indizes dienen zur Beschleunigung von Datenabfragen in Datenbanken. Sie funktionieren ähnlich wie ein Inhaltsverzeichnis in einem Buch und ermöglichen es, schnell auf bestimmte Datensätze zuzugreifen, ohne die gesamte Tabelle durchsuchen zu müssen.

**Korrelation:** Die Korrelation beschreibt den linearen Zusammenhang zwischen zwei Variablen. Sie kann Werte zwischen -1 und +1 annehmen, wobei -1 einen perfekten negativen Zusammenhang, 0 keinen Zusammenhang und +1 einen perfekten positiven Zusammenhang darstellt.

**Machine Learning:** Bei Machine Learning kommen Algorithmen zum Einsatz, die aus Daten lernen, um Vorhersagen zu treffen oder Entscheidungen zu fällen. Für das Training von Modellen werden die Daten in Trainings-, Validierungs- und Testdaten unterteilt.

**Median:** Der Median ist ein weiteres zentrales Lagemaß, das den Wert angibt, der die Daten in zwei gleich große Hälften teilt.

**Mittelwert / Durchschnitt:** Der Mittelwert (auch Durchschnitt genannt) ist ein zentrales Lagemaß in der Statistik. Er wird berechnet, indem die Summe aller Werte durch die Anzahl der Werte geteilt wird.

# Glossar V

---

**Primär- und Fremdschlüssel:** In relationalen Datenbanken dienen Primär- und Fremdschlüssel zur Verknüpfung von Tabellen. Der Primärschlüssel ist eine eindeutige Kennung für jede Zeile in einer Tabelle, während der Fremdschlüssel auf den Primärschlüssel einer anderen Tabelle verweist und so die Beziehung zwischen den Tabellen herstellt.

**ORM (Object-Relational Mapping):** ORM ermöglicht die Verbindung von Objekten in einer Programmiersprache mit Datenbanktabellen. Dadurch können Entwickler mit Datenbanken arbeiten, ohne direkt SQL schreiben zu müssen, was die Entwicklung vereinfacht und den Code lesbarer macht.

**Perzentile und Quartile:** Perzentile und Quartile sind Wertschranken, die die Daten in bestimmte Anteile aufteilen. Das 25%-Perzentil (auch unteres Quartil genannt) ist der Wert, unterhalb dessen 25% der Daten liegen.

**PostGIS:** PostGIS ist eine Erweiterung für die relationale Datenbank PostgreSQL, die die Verarbeitung von Geodaten ermöglicht. Mit PostGIS können räumliche Daten wie Punkte, Linien und Polygone in der Datenbank gespeichert und abgefragt werden. PostGIS bietet Funktionen zur Berechnung von Distanzen, Flächen und Schnittmengen sowie zur Transformation von Koordinatensystemen.

**Pydantic:** Pydantic ist eine Python-Bibliothek, die Datenvielfältigung und -parsing ermöglicht. Pydantic verwendet Typannotationen, um Datenmodelle zu definieren, und validiert automatisch die Eingabedaten anhand dieser Modelle.

**Python-Bibliotheken:** pandas, numpy, intake, SQLAlchemy, requests, FastAPI.

# Glossar VI

---

**RESTful API:** RESTful APIs sind ressourcenorientierte Architekturen für APIs, die auf dem HTTP-Protokoll basieren. Jede Ressource (z. B. ein Benutzer, ein Artikel) wird durch eine eindeutige URL adressiert. Zur Interaktion mit den Ressourcen werden die HTTP-Methoden GET, POST, PUT und DELETE verwendet.

**Saisonalität:** Saisonalität beschreibt regelmäßige Schwankungen in Daten über die Zeit, die auf wiederkehrende Ereignisse wie Jahreszeiten, Feiertage oder Wochentage zurückzuführen sind.

**Serialisierung:** Die Serialisierung beschreibt die Umwandlung eines Objekts in ein Format, das gespeichert oder übertragen werden kann. Dies ist wichtig für die Speicherung von Daten in Datenbanken oder Dateien sowie für die Übertragung von Daten über Netzwerke, beispielsweise in APIs oder Remote-Verbindungen.

**Standardabweichung:** Die Standardabweichung ist ein Streuungsmaß, das die durchschnittliche Abweichung der Werte vom Mittelwert angibt.