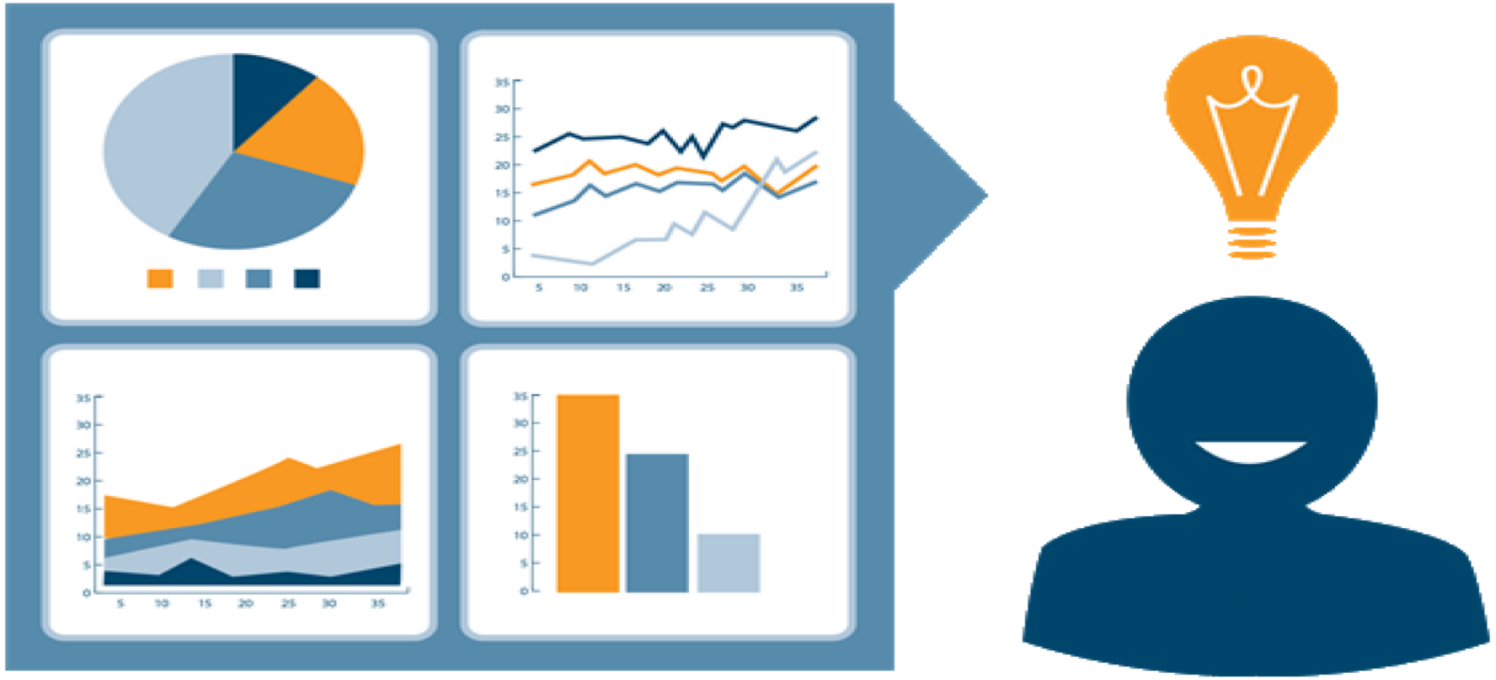
Introduction to Data Science
Prof. Dr. Bernhard Drabant
Duale Hochschule Baden-Württemberg Mannheim
Fakultät Wirtschaft

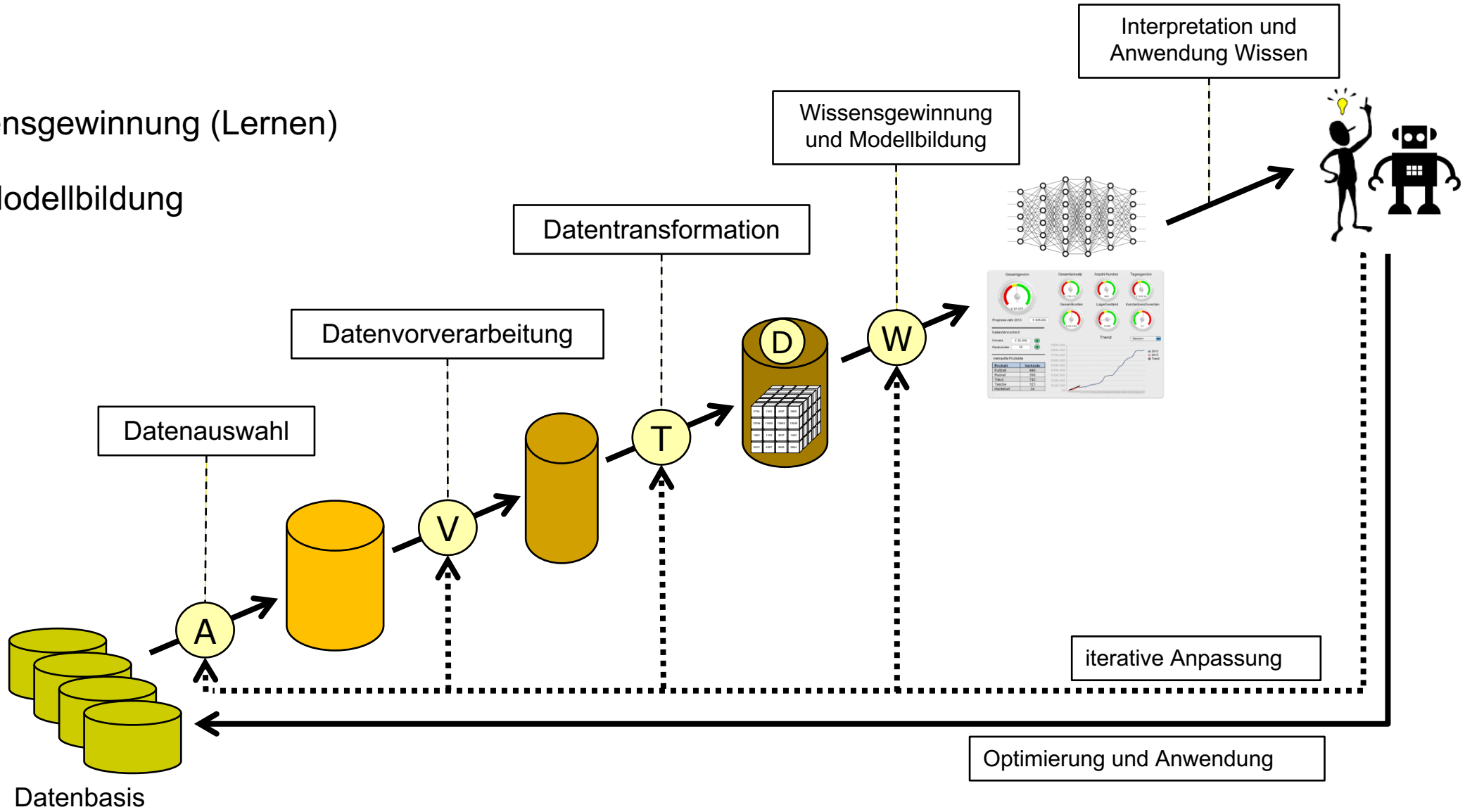
Wissensgewinnung und Modelle



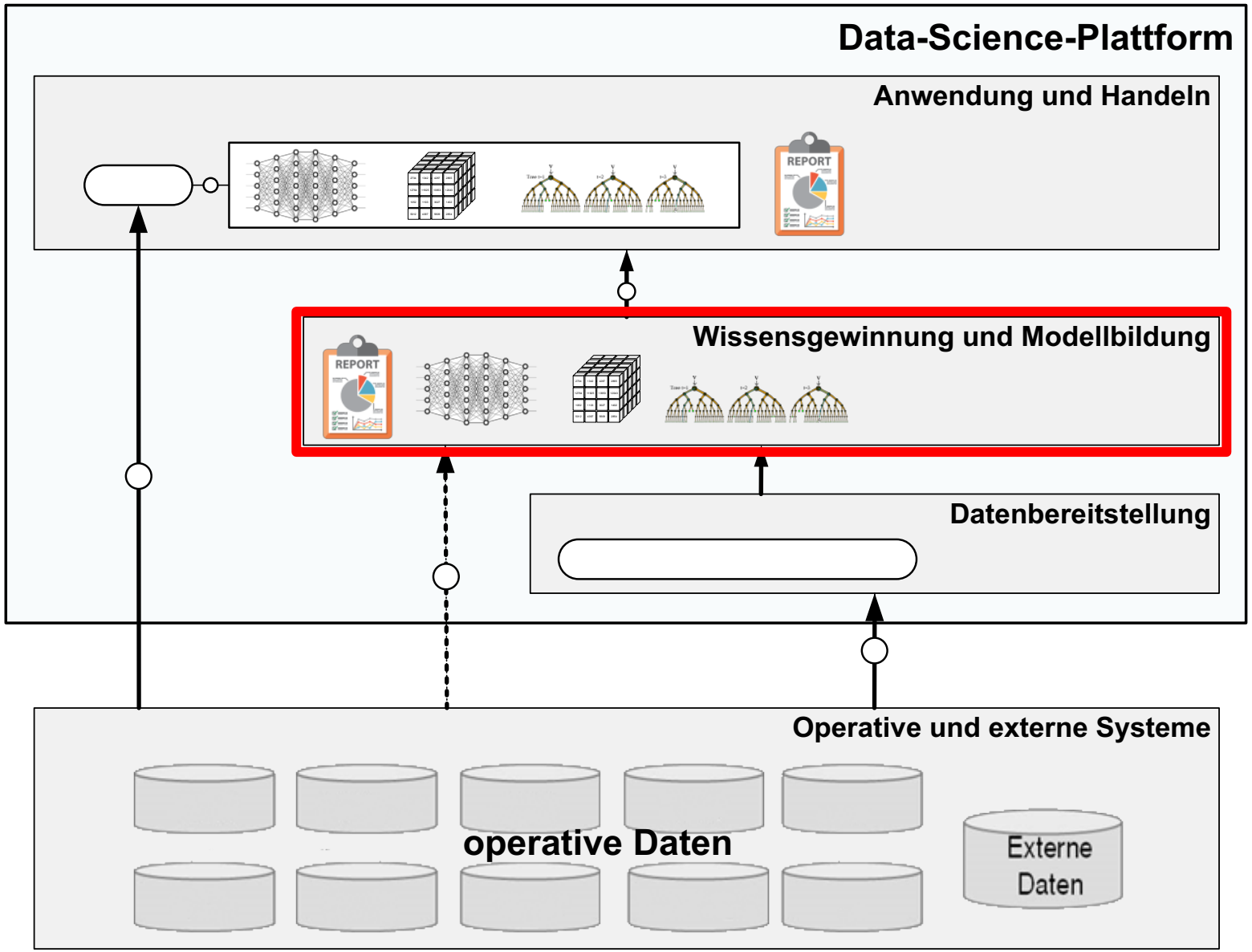
Data Science als Prozess

W

- Wissensgewinnung (Lernen)
- ggf. Modellbildung



Data-Science-Plattform

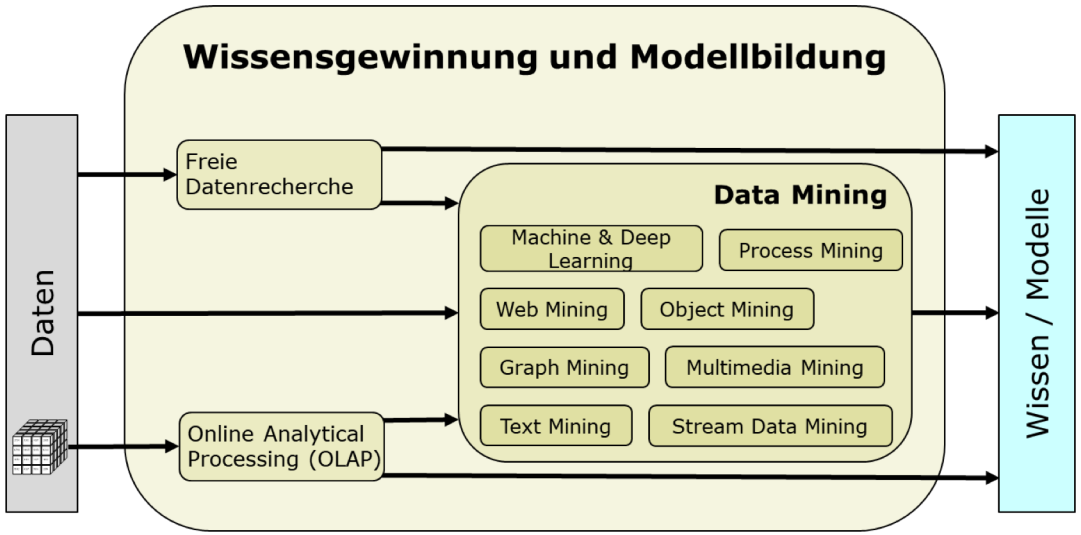


Wissensgewinnung und Modellbildung

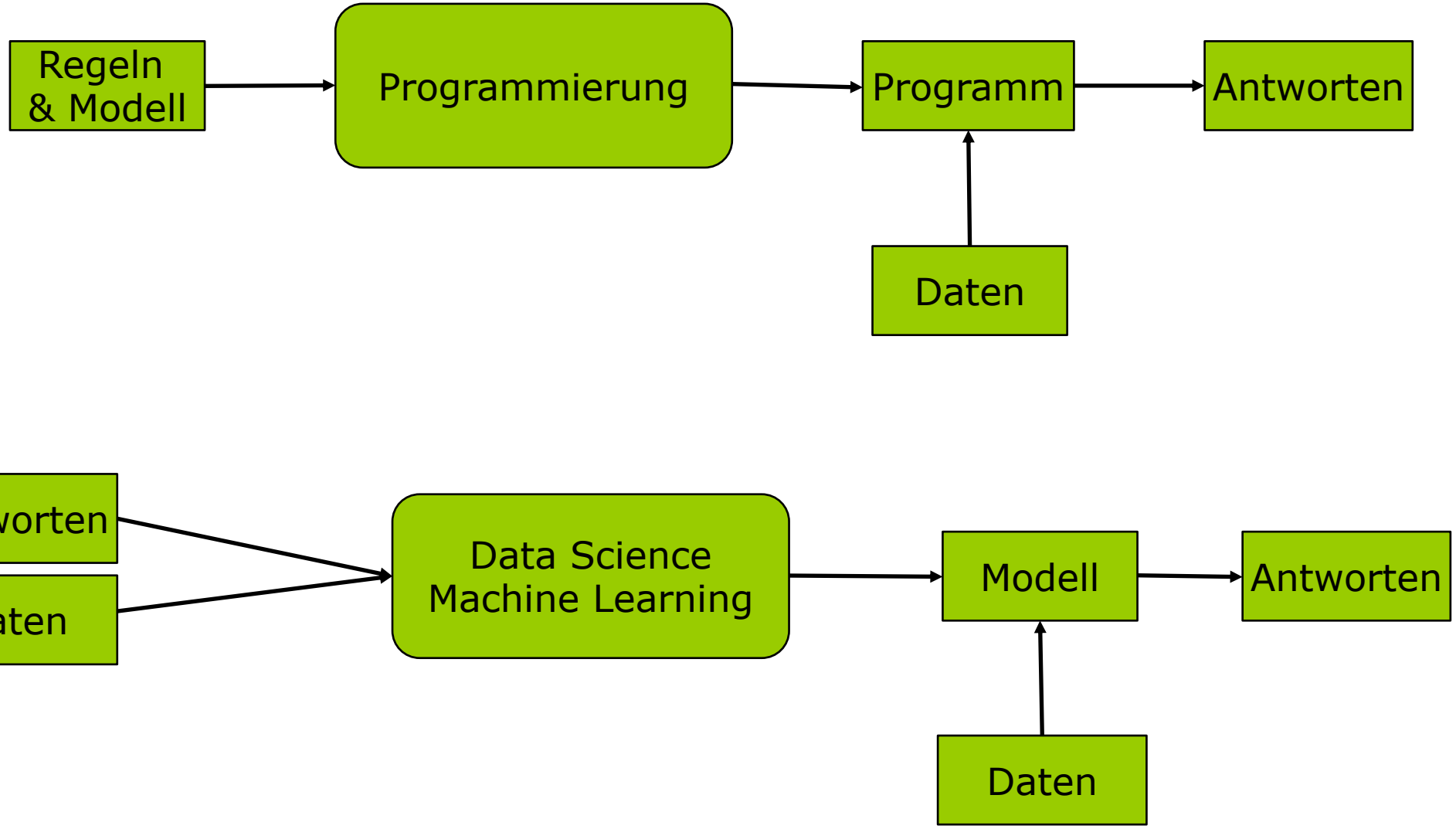
Zentraler Schritt im Data-Science-Prozess

- Ziel
  - Identifizierung von (bisher unbekannten) aussagekräftigen Mustern und Zusammenhängen in (aufbereiteten) Daten
  - Darstellung dieses Wissens in Modellen und Berichten
  - Modelle auf Daten (z. B. in operativen Systemen) anwenden

- Methoden
  - abhängig von Anwendungsfällen, Anforderungen, Datenstrukturen
  - aus den Bereichen
    - Freie Datenrecherche
    - Analytik
    - Data Mining, Machine & Deep Learning, …

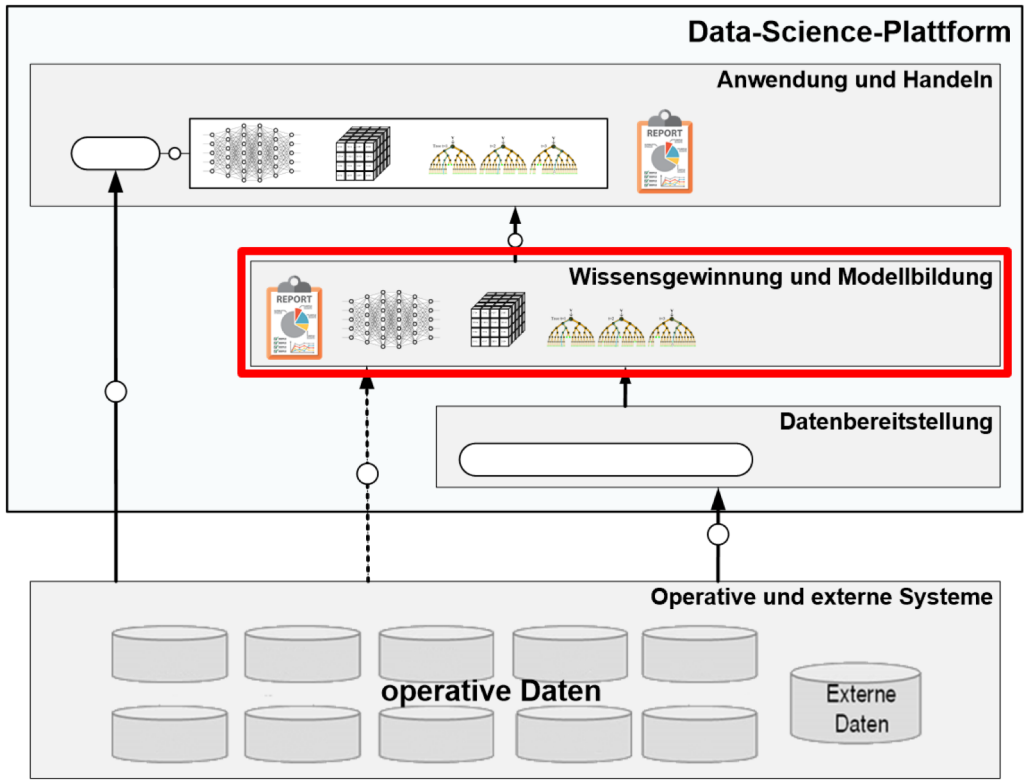


Vergleich: Klassische Programmierung – Maschinelles Lernen



Analysesysteme

- Tools zur Wissensgewinnung und Modellbildung
- Teil der Data-Science-Plattform
- Arten
  - **Freie Datenrecherche**
  - **OLAP**
  - **Data Mining**
- Auswahl gemäß der Anwendungsfälle und Anforderungen



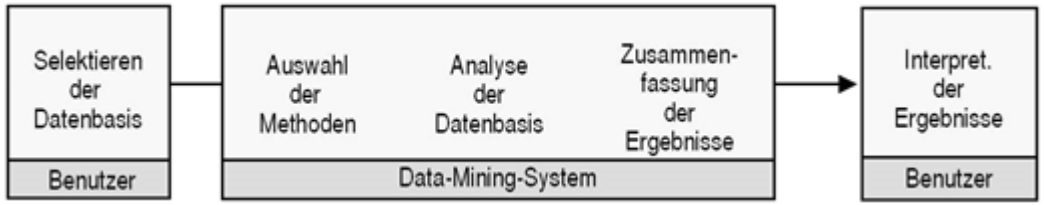
Data Mining

- (Automatisierte) Anwendung von Lernmethoden und Lernalgorithmen, um in (großen) Datenmengen
  - bisher unbekannte, aussagekräftige, potenziell nützliche Strukturen und Beziehungen zwischen Datenobjekten zu identifizieren
  - und diese einem Konsumenten in Form von Modellen zur Verfügung zu stellen
- Sehr häufig: statistische Lernmethoden

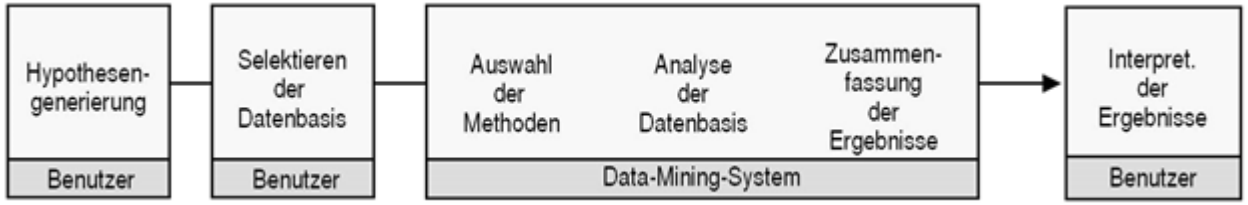
$\Rightarrow$ **(Automatisierte, statistische) Gewinnung von Wissen aus Daten**

Data Mining als Wissenschaft der Daten

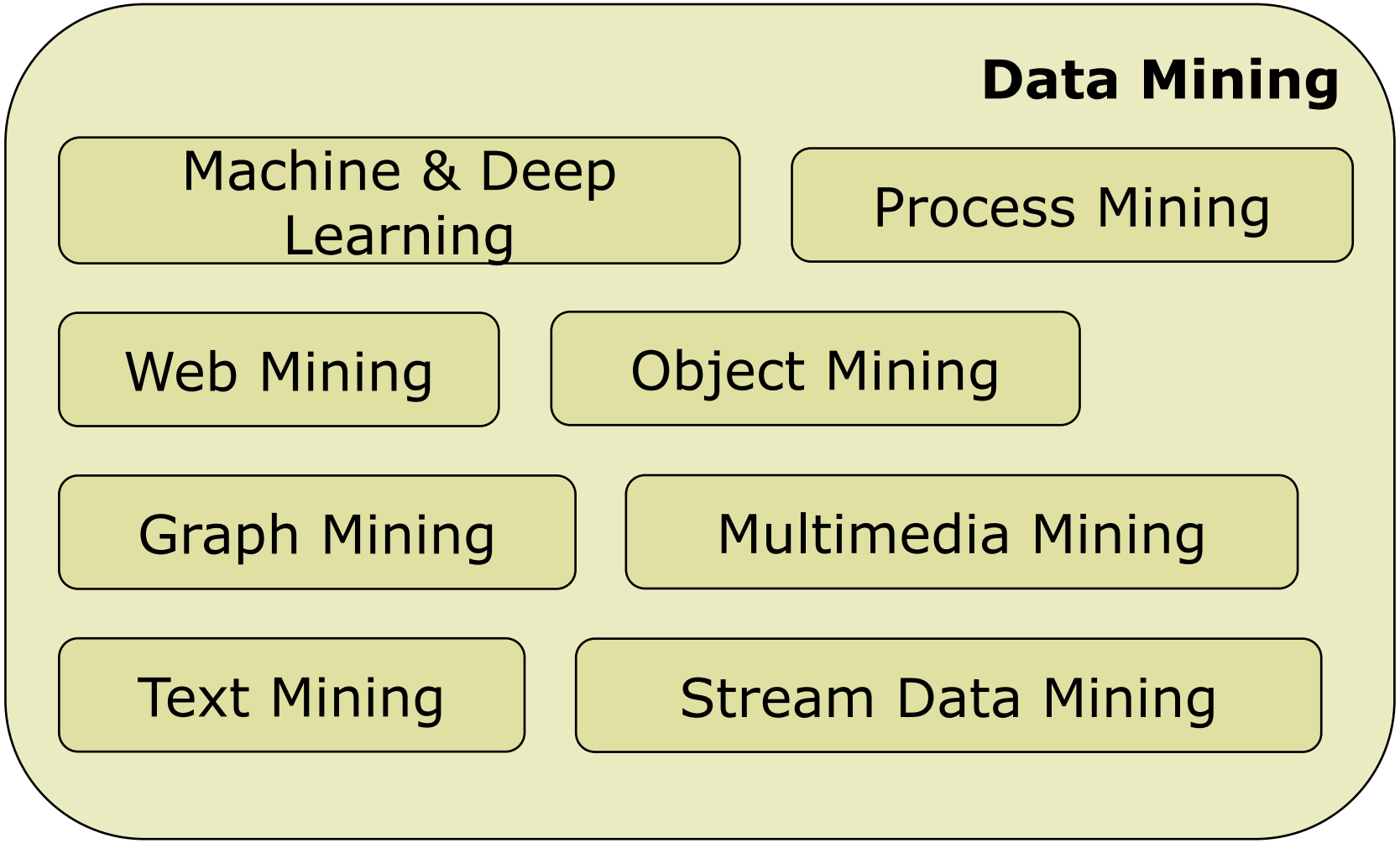
- Hypothesenfreie Erkennung von Strukturen und Zusammenhängen



- Validierung von Hypothesen



Methodiken des Data Mining / der Wissensgewinnung



Taxonomie des Lernens (Data Mining / Machine Learning)

- Kategorien von Lernproblemen
  - Arten des Lernens
    - Klassen / Typen des Lernens
      - Methoden des Lernens
        - » Anwendungsfelder des Lernens

Problemkategorien des Lernens

- **Entscheidungs- und Prognoseprobleme**
- **Beschreibungs- und Strukturierungsprobleme**

Wissensgewinnung/Unterstützung durch Data Mining

- **bei Entscheidungs- und Prognoseproblemen**
  - Unterstützung bei Entscheidungen oder autonome Entscheidungsfindung
  - Prognose und Wahrscheinlichkeiten von Ereignissen, Zuständen, Prozessabläufen bestimmen
  - Prozesse auf Basis erkannter Muster zu optimieren
- **bei Beschreibungs- und Strukturierungsproblemen**
  - Unterstützung bei der Beschreibung oder Bewertung von Eigenschaften von Datenobjekten oder komplexeren Datenzusammenhängen
- **durch Lernalgorithmen**
  - Ermöglichen das Erlernen von Wissen und das Generieren von Modellen anhand von Daten

Data Mining – Problemkategorien

Problemkategorien des Lernens und Wissensgewinnung/Unterstützung durch Data Mining

- Generell: Datenobjekte O durch Variablen $X_1, X_2, \ldots, X_n$ charakterisierbar.

Grundannahme: Durch Data Mining erlerntes Wissen i. d. R. durch Modelle darstellbar

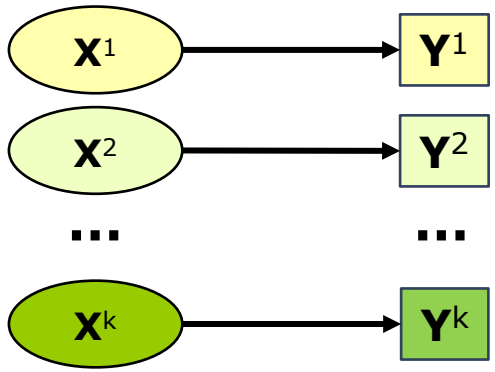
- **Entscheidungs- und Prognoseprobleme**
  - Entweder: Finde Zuordnung von Variablen $X_1, X_2, \ldots, X_n$ eines Datenobjekts O auf Zielvariable Y
  - Oder: Finde Zuordnung von Variablen $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(m)}$ von m Datenobjekten $O^{(1)}, \ldots, O^{(m)}$ (ggf. verschiedenen Typs) auf Zielvariable Y
  - Modell **M** ist eine Funktion: $\mathbf{Y} = \mathbf{M}(X_1, X_2, \ldots, X_n)$ bzw. $\mathbf{Y} = \mathbf{M}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(m)})$
  - Wert **Y** aus Wert **X** ermittelbar (i. d. R. mit gewisser Wahrscheinlichkeit → **Zufallsvariable**)
- **Beschreibungs- und Strukturierungsprobleme**
  - Finde Zusammenhang zwischen den Variablen $X_1, X_2, \ldots, X_n$ bzw. zwischen $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(m)}$
  - Modell **M** ist eine neue Struktur: $\mathbf{M}(X_1, X_2, \ldots, X_n)$ bzw. $\mathbf{M}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(m)})$

Arten des Lernens

- Supervised Learning (Überwachtes Lernen)
- Unsupervised Learning (Unüberwachtes Lernen)
- Self-taught Learning / Semi-supervised Learning

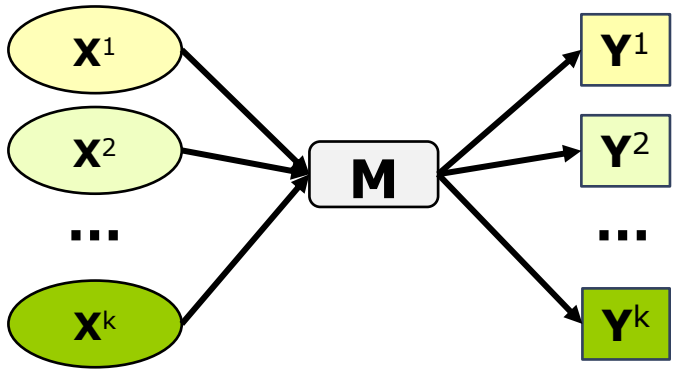
  (Autodidaktisches Lernen / teilüberwachtes Lernen)
- Reinforcement Learning (Verstärkendes Lernen)

Wissensgewinnung: Supervised Learning



- **Ausgangslage / Annahme**
  - Datenobjekte O haben Daten **X** = (X_1, X_2, …, X_n) (= Eingangsvariablen)
  - Jedem Datenobjekt O wird Zielvariable **Y** zugeordnet
  - Wert **Y** aus Wert **X** ermittelbar (Zufallsvariable)

Wissensgewinnung: Supervised Learning



- **Ziel**
  - Erlerne Funktion/Modell **M**, womit aus Wert **X** der Zielwert **Y** bestimmt werden kann:

    Y = M (X)
- **Vorgehensweise**
  - Lernen von Funktion/Modell **M** mit k *Trainingsdatensätzen* (Stichprobe / Sample) { ($\mathbf{X}^1$, $\mathbf{Y}^1$), ($\mathbf{X}^2$, $\mathbf{Y}^2$), … , ($\mathbf{X}^k$, $\mathbf{Y}^k$) }
  - Lernen endet, wenn ein Modell **M** mit ausreichender Güte bestimmt wurde
    - Güte wird in der Regel mit *annotierten Testdaten* gemessen
- **Zentrale Probleme des Supervised Learning**
  - **M** a priori nicht bekannt
  - Welche Methoden wendet man an, um Modell **M** zu bestimmen?
  - Was bedeutet „möglichst genau“ und „ausreichende Güte“?

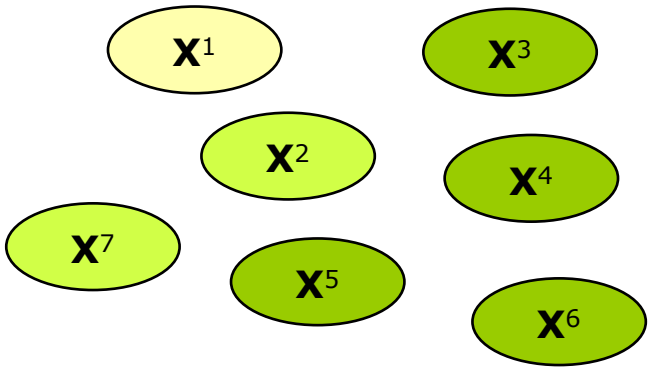
Wissensgewinnung: Supervised Learning

- **Klassen des Supervised Learning**
  - Klassifikation: Zielvariable Y endlich, nominal oder ordinal
  - Regression: Zielvariable Y metrisch, diskret oder kontinuierlich
  - Mischformen möglich

- **Methoden des Supervised Learning**
  - Lineare Regression, logistische Regression, …
  - Entscheidungsbaumverfahren, Random Forest
  - Support-Vektor-Maschinen
  - Neuronale Netze

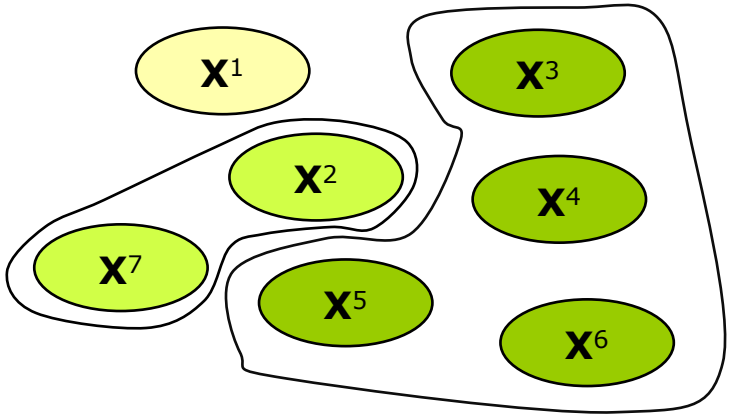
- **Anwendungsfelder des Supervised Learning**
  - Empfehlungen
  - Zeitreihenvorhersage
  - Predictive Analytics / Predictive Maintenance
  - …

Wissensgewinnung: Unsupervised Learning



- **Ausgangslage / Annahme**
  - Datenobjekte O haben Daten **X** (= Eingangsvariablen)
  - Den Datenobjekten sind keine Zielvariablen zugeordnet

Wissensgewinnung: Unsupervised Learning



3 Gruppen:
gelb / hellgrün / grün

- **Ziel**
  - Finde Modell **M** (Struktur, Beziehungen in den Daten)
- **Vorgehensweise**
  - Lernen oder Finden der relevanten Strukturen mit (nicht-annotierten) *Trainingsdaten*
  - Lernen ohne Anleitung!
  - Lernen endet, wenn Strukturen mit ausreichender Güte oder Aussagekraft bestimmt wurde
    - Güte oder Aussagekraft wird unter anderem mit *Testdaten* gemessen
- **Zentrale Probleme des Unsupervised Learning**
  - Welche Methoden wendet man an, um die relevanten Strukturen zu ermitteln?
  - Was bedeutet „ausreichende Güte oder Aussagekraft“?

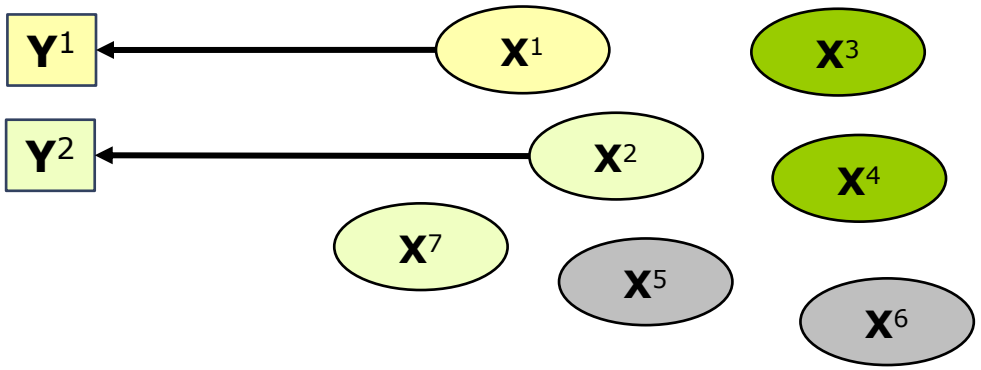
Wissensgewinnung: Unsupervised Learning

- **Klassen des Unsupervised Learning**
  - Gruppierung/Clustering: Gruppierungsmerkmale und -strukturen in Daten ermitteln
    - Ähnlichkeiten von Objekten / Funktionen auf Objekt-Daten
  - Dimensions- oder Datenreduktion
    - Vereinfachung der Datenstruktur oder Reduktion der Datenmenge / Reduktion auf wesentlichen Parameter oder Merkmale
  - Assoziationsanalyse
    - Regeln ermitteln, die Zusammenhänge zwischen Daten beschreiben
  - Muster- und Abweichungserkennung / Anomaly Detection

- **Methoden des Unsupervised Learning**
  - Frequent Item Set Analysis
  - k-means-Methoden
  - Principal Component Analysis (PCA)
  - Neuronale Netze

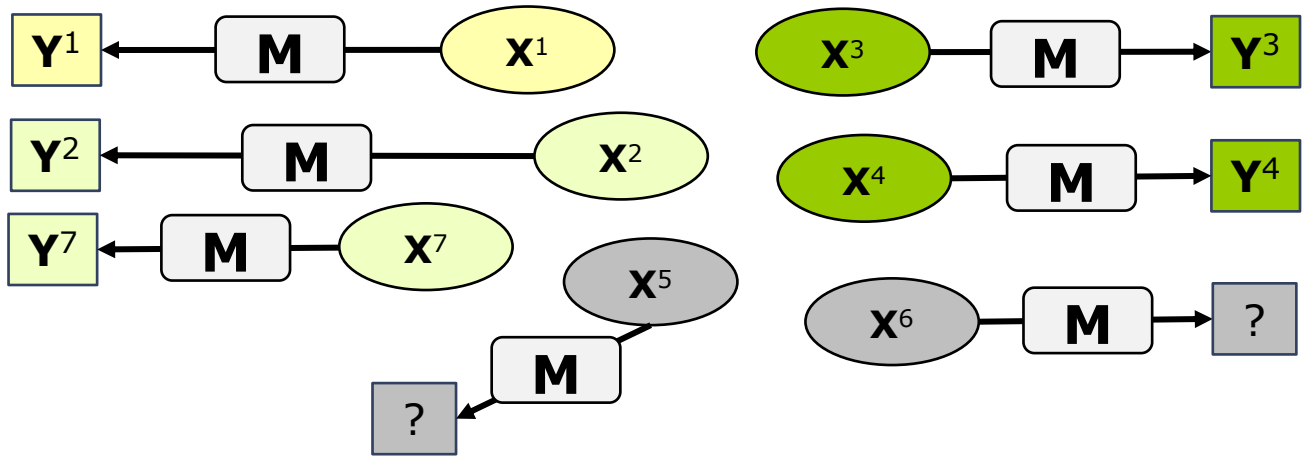
- **Anwendungsfelder des Unsupervised Learning**
  - Warenkorbanalyse / Kunden, die Ware X kaufen, neigen auch dazu, Produkt Y zu kaufen
  - Aufdeckung betrügerischen Verhaltens (Kreditkartenbetrug)
  - Gesichtserkennung

Wissensgewinnung: Self-taught und Semi-supervised Learning



- **Ausgangslage / Annahme**
  - Datenobjekte O haben Daten **X** (= Eingangsvariablen)
  - Trainingsobjekte sind nur teilweise mit Zielvariablenwerten **Y** annotiert
  - In der Praxis: Daten mit annotierten Zielvariablen in der Unterzahl

Wissensgewinnung: Self-taught und Semi-supervised Learning



- **Ziel**
  - Erlerne Funktion/Modell **M**, womit aus **X** der Zielwert **Y** bestimmt werden kann:

    Y = M (X)
- **Vorgehensweise**
  - Lernen der Funktion/Modell **M** erfolgt mit *Trainingsdaten* (in mehreren Schritten)
  - Lernen endet, wenn ein Modell **M** mit ausreichender Güte bestimmt wurde
    - Güte wird in der Regel mit *Testdaten* gemessen
- **Nebenbedingungen (!)**
  - Nicht alle Trainingsobjekte müssen in Lernphase eingebunden werden
  - Nicht alle Daten müssen durch **M** klassifiziert werden können

Wissensgewinnung: Self-taught und Semi-supervised Learning

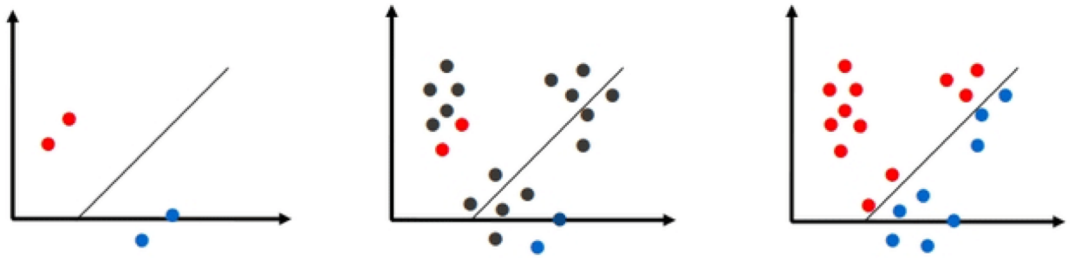
Szenarien

I. Alle Trainingsobjekte sind un-annotiert → Self-taught
   1. Unsupervised Learning mit Gruppenbildung auf den Trainingsdaten
   2. Annotation einiger Trainingsobjekten mit entsprechenden Gruppen-Indizes
   3. Supervised Learning auf allen annotierten Trainingsobjekten → Modell $\mathbf{M_1}$
   4. Annotation restlicher Trainingsobjekte (oder Teilmenge davon) mit Hilfe von Modell $\mathbf{M_1}$
   5. Supervised Learning auf allen annotierten Trainingsobjekten → Modell $\mathbf{M_2}$
   6. Gegebenenfalls Iteration der Schritte (4) und (5)
   7. Gütebestimmung des Modells auf Testdaten
   8. Anwendung des Modells

II. Einige Trainingsobjekte sind annotiert → Semi-supervised
   - Vorgehensweise wie im ersten Szenario ab Schritt (3)

Vergleich: supervised, unsupervised, semi-supervised

supervised



unsupervised

semi-supervised

Wissensgewinnung: Reinforcement Learning

- **Ausgangslage**
  - Lernsystem (= **Agent**) beobachte seine Umgebung
  - System wählt kontextabhängige Aktionen aus und führt diese durch

- **Vorgehensweise**
  - Je nach Auswahl der Aktion wird das System mehr oder weniger belohnt
  - System findet selbst heraus, welche Aktionen in welchem Kontext die höchste Belohnung erbringen
  - System erzeugt im Lernprozess eine Ausführungsstrategie (**Policy**)

- **Ziel**
  - Optimierung der Policy

Wissensgewinnung: Reinforcement Learning

- **Anwendungsfelder**
  - Roboter
  - Autonome oder selbstfahrende Fahrzeuge
    - teilweise, neben anderen Technologien wie NNs, …
  - KI und Spiel-Programme
    - Beispiel: AlphaGo von DeepMind schlägt Go-Weltmeister im Jahr 2017 und erlernt/entdeckt bis dahin unbekannte Strategien

- **Reinforcement Learning nicht Teil der Vorlesung**

Prozesse der Wissensgewinnung

- Batch-Verfahren
- Online-Verfahren

Prozesse der Wissensgewinnung

- **Batch-Verfahren**
  - Alle Trainingsdaten werden für Training verwendet → Daten-Pool
    - in der Regel in separater offline-Umgebung durchgeführt
  - Nach Training (und Test) steht erlerntes Wissen zur Verfügung
    - in Produktivsystemen eingesetzt
    - keine weiteres Lernen in Produktivsystem
  - Alternative Bezeichnung: Offline-Verfahren

Prozesse der Wissensgewinnung

- **Batch-Verfahren**
  - Eigenschaften
    - Zeit- und ressourcenintensives Lernen auf gesamtem Trainingsdaten-Pool
      - Herausforderung! Je nach Größe der Datenmenge und Art der Berechnungen
    - Wie wird aus neuen Daten gelernt?
      - Neuer Trainingslauf in offline-Umgebung mit neuen Trainingsdaten
      - Re-Deployment des Wissens (Modell, …) in Produktivumgebung

Prozesse der Wissensgewinnung

- **Online-Verfahren**
  - Inkrementelles Training durch Hinzufügen
    - einzelner Datensätze
    - kleiner Datenpakete (Mini-Batches)
  - Mögliche Lernszenarien
    - auf Produktivdaten
    - aber auch auf separaten Trainingsdaten
    - in Mischform auf Trainings- und Produktivdaten

Prozesse der Wissensgewinnung

- **Online-Verfahren**
  - Eigenschaften
    - schnelles, inkrementelles Lernen in Echtzeit möglich
    - auf transienten oder auf persistenten Daten möglich – je nach Anwendung
  - Anwendungsszenarien
    - Lernen aus sich schnell ändernden Daten wie Aktienkursen etc.
    - Roboter und autonome Maschinen
    - Umgebungen mit geringer Rechen- und Ressourcenkapazität
    - Batch-Systeme mit riesigem (Trainings-)Daten-Pool
      - Inkrementelles Lernen auf kleineren Batches
        - » z. B. in Form von Out-of-Core Learning (im Hauptspeicher)

Prozesse der Wissensgewinnung

- **Online-Verfahren**
  - Herausforderung beim Online-Verfahren:
    - Allmähliche Verschlechterung der Lernleistung durch fortlaufende Einspeisung minderwertiger Daten
    - Beispiel: Defekter Sensor, Ranking-Bot, …

Trainingsdaten, Testdaten, Batch, Epoche, Lernrate

- Trainingsdatensatz
  - Zufällig gewählter Datensatz (Stichprobe) zum Erlernen eines (statistischen) Wissensmodells mit Hilfe eines gegebenen (parametrisierten) Lernalgorithmus → maschinelles Lernen / Machine Learning

- Testdatensatz
  - Zufällig gewählter Datensatz (Stichprobe) zum Testen der Güte eines (statistischen) Wissensmodells
  - Disjunkt von Trainingsdatensatz

Zusammenfassung: Trainingsdaten, Testdaten, Batch, Epoche, Lernrate

- Training/Lernen des Modells
  - Erlernen eines Modells durch Trainingsdatensatz
  - Lernen kann in mehreren Phasen iterativ erfolgen

- Test/Validierung des Modells
  - Bestimmung der Güte des Modells durch Testdatensatz
  - Gütebestimmung abhängig vom Modell bzw. vom Lernverfahren
    - supervised: Abweichung zwischen Vorhersage und Annotation
    - unsupervised: Abweichung von der in der Lernphase ermittelten Struktur

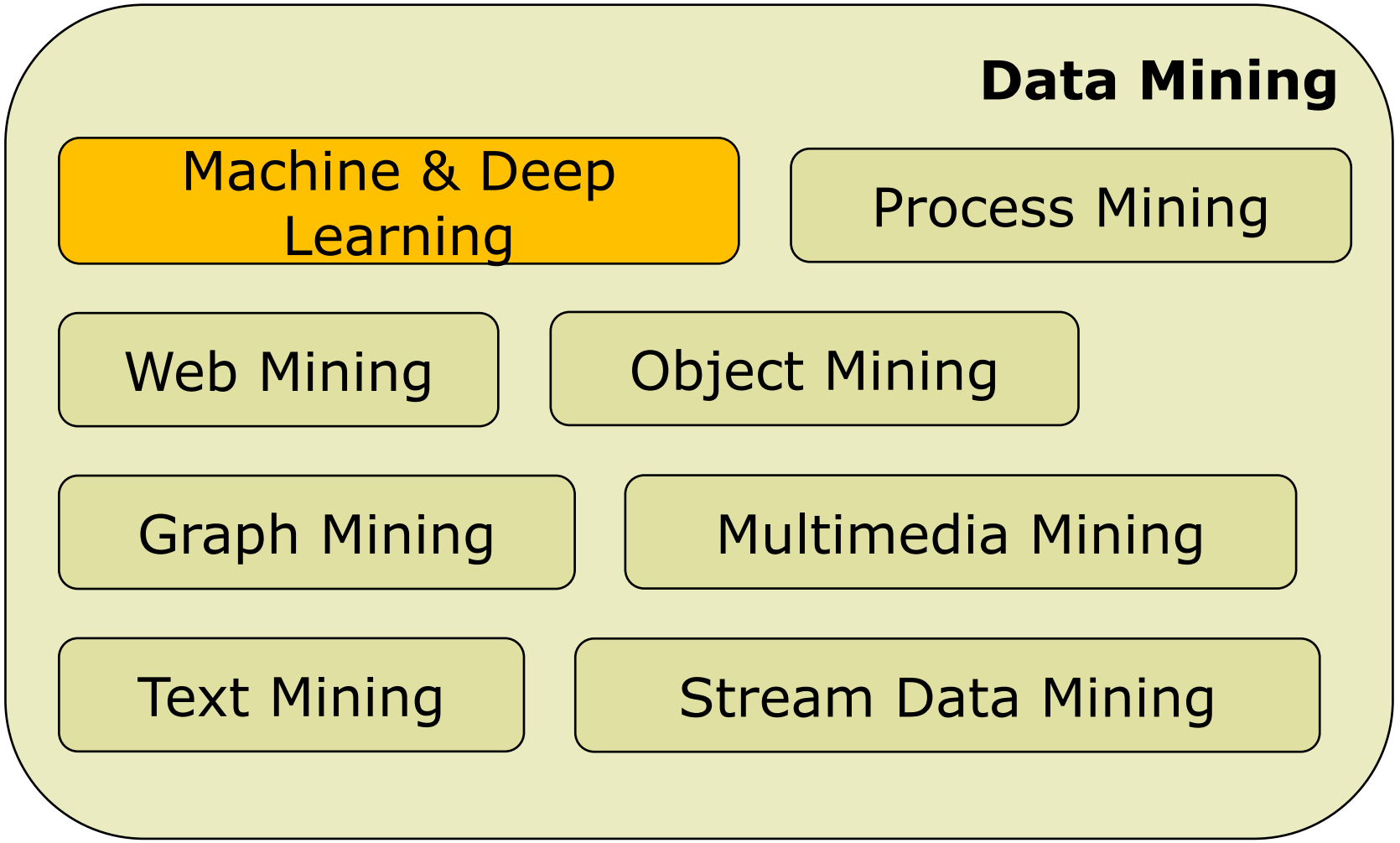
Zusammenfassung: Trainingsdaten, Testdaten, Batch, Epoche, Lernrate

- Batch
  - Batchdatensatz: Untermenge eines Trainingsdatensatzes
  - Batch: Partition des Trainingsdatensatzes in möglichst gleich große Batchdatensätze
  - Verwendung: Sukzessives Training eines Wissensmodells.
    - Nach jedem Batchdatensatz wird die Güte des erlernten Modells ermittelt und die Parameter des Lernalgorithmus / Modells angepasst bzw. optimiert
- Epoche: Training eines Modells mit einem vollständigen Durchlauf des Trainingsdatensatzes
  - Training kann über mehrere Epochen erfolgen – mit dem selben Trainingsdatensatz
  - Gütebestimmung nach jeder Epoche

Zusammenfassung: Trainingsdaten, Testdaten, Batch, Epoche, Lernrate

- Lernrate
  - Ziel: Güte eines Modells durch Anpassung/Optimierung der Parameter des Lernalgorithmus in der nächsten Epoche erhöhen
  - Lernrate: Maß der Anpassung/Optimierung der Parameter
  - Lernrate hoch:
    - System lernt schnell
    - Aber:
      - Gelerntes aus den alten Daten „schneller *vergessen*“
      - System kann „über das Ziel hinaus schießen“
  - Lernrate niedrig:
    - System hat höhere Trägheit beim Lernen und lernt somit langsamer
    - Aber: weniger anfällig für *Rauschen* oder nicht repräsentative Datenobjekte in neuen Daten

Machine Learning (ML)



Machine Learning (ML)

- ML = Wissenschaft und Methodologie, IT-Systeme mithilfe bestimmter ML-Algorithmen so zu programmieren, dass sie anhand vorgegebener Daten **eigenständig** lernen

- Die IT-Systeme können dann auf Basis vorhandener Trainingsdaten
  - Muster und Gesetzmäßigkeiten erlernen
  - Lösungen (z. B. Modelle) entwickeln

- Eigenständiges Lernen der IT-Systeme bedarf der Vorbereitung durch Menschen
  - Systeme müssen mit relevanten Lernalgorithmen und Daten ausgestattet werden
  - Algorithmen müssen für Analyse der Daten und Erkennen der Muster konfiguriert werden
  - Dieses Verfahren muss iterativ „orchestriert“ werden
  - Hohe Kunst des ML!

Machine Learning (ML)

- Ein zentraler Bereich im Data Mining
- Sehr wichtiger Teil der Studienrichtung Data Science @ DHBW
- Deep Learning (DL): Fortgeschrittene Methoden des modernen ML
  - in späteren Semestern …

Fragen?