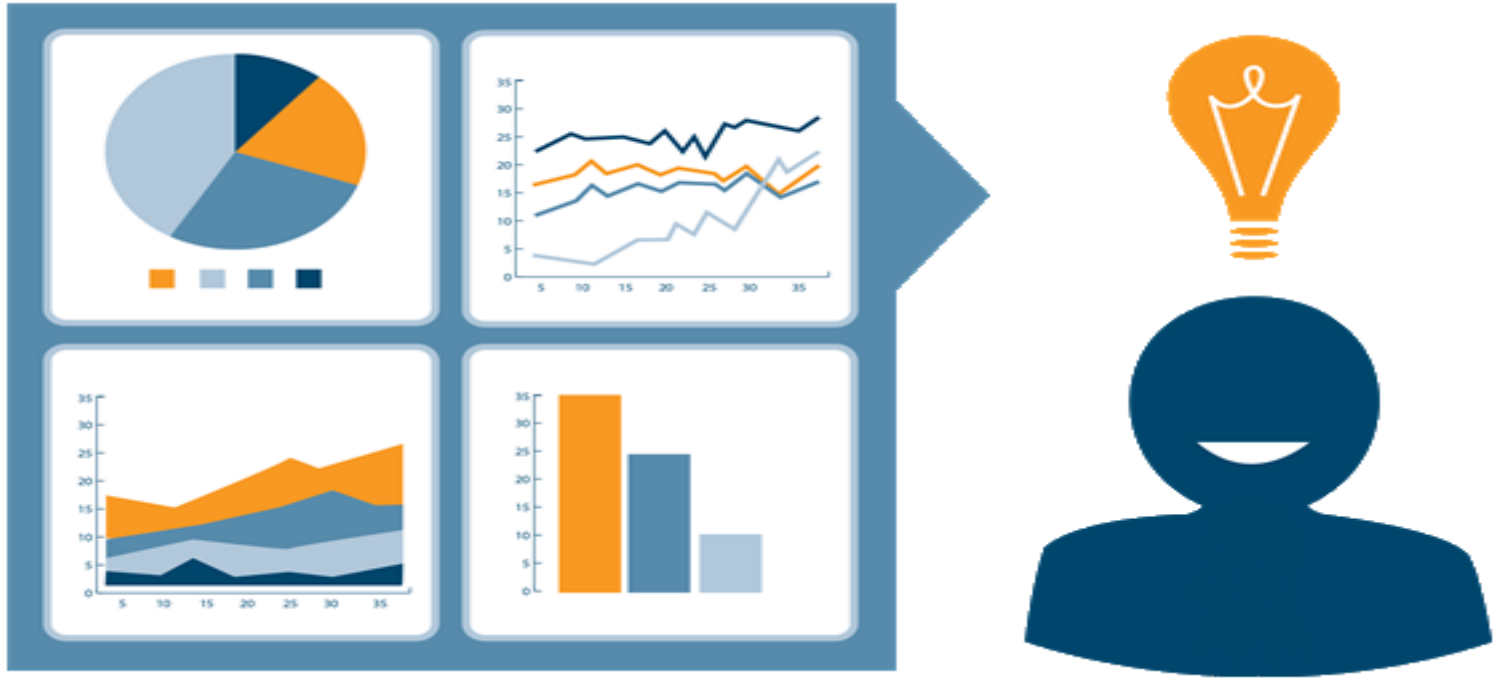
Introduction to Data Science
Prof. Dr. Bernhard Drabant
Duale Hochschule Baden-Württemberg Mannheim
Fakultät Wirtschaft

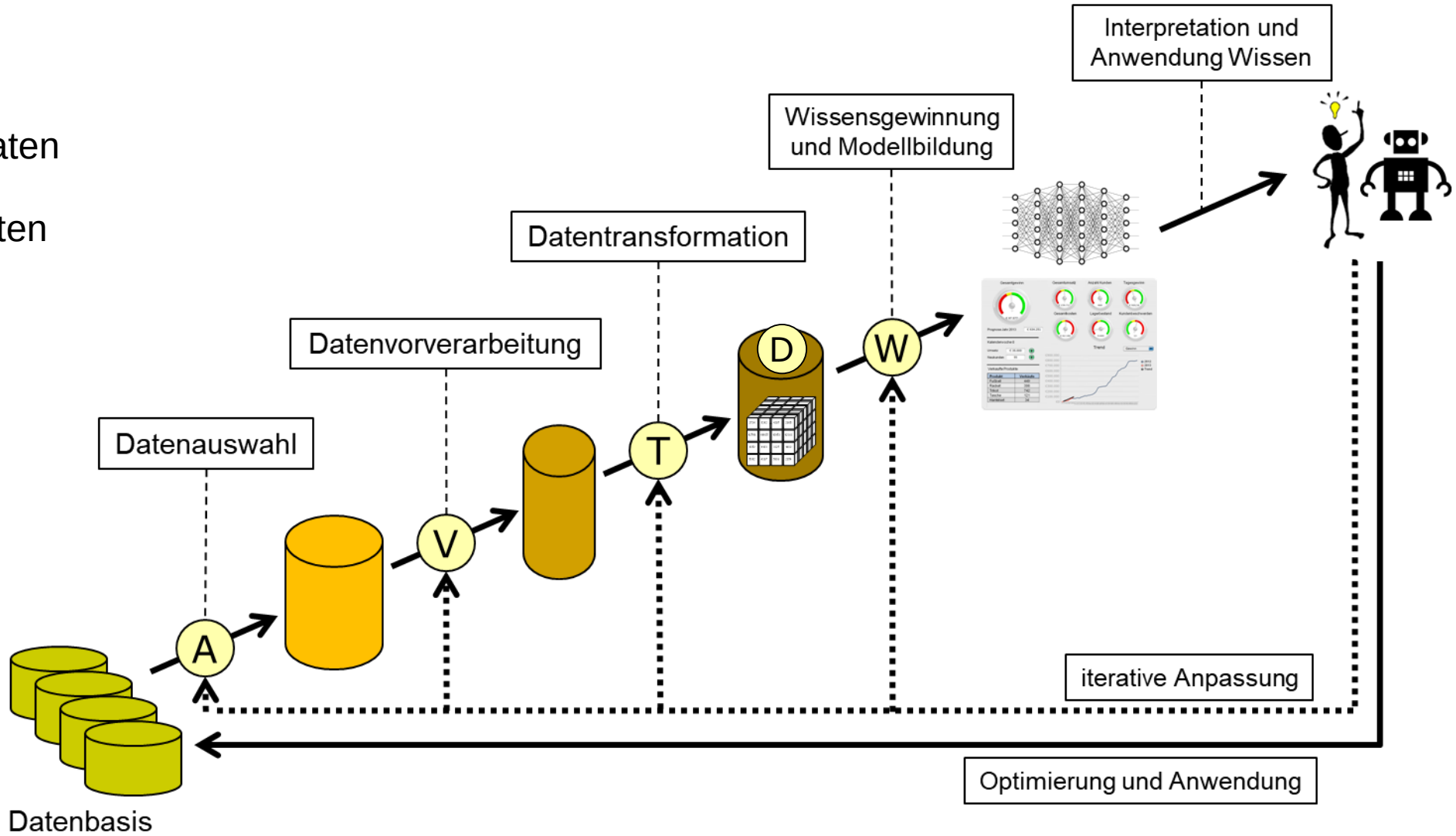
Datenaufbereitung



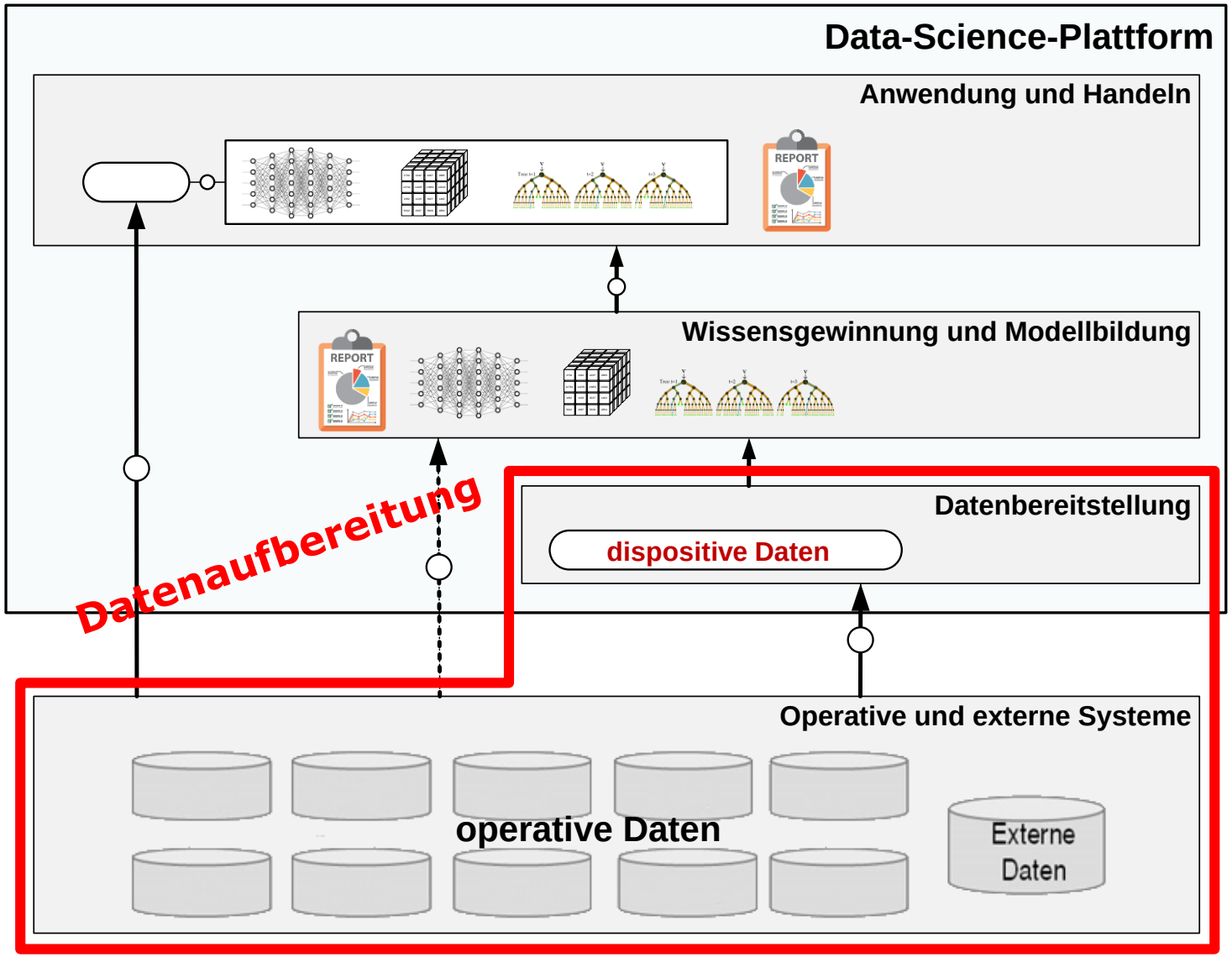
Data Science als Prozess

A V T D = **Datenaufbereitung**

- **A**uswahl der Daten
- **V**orverarbeitung der Daten
- **T**ransformation der Daten
- **D**atenbereitstellung
- **W**issensensgewinnung



Data-Science-Plattform



Operative / externe vs. dispositive Systeme und Daten

	operativ / extern	dispositiv
Ziel	Daten des laufenden Betriebs: Geschäftsprozessdaten, Sensordaten, etc.	Daten auf Basis operativer oder externer Daten. Relevant für Wissensgewinnung, extrahiert von Datenquelle
Ausrichtung	Detaillierte, feingranulare Geschäftsprozessdaten; Datensätze pro Transaktion	Transformierte, angereicherte, ggf. aggregierte Daten, aber auch Streaming-Daten
Zeitbezug	Aktuell; zeitpunktbezogen; transaktionsbasiert	zeitraum- und zeitpunktbezogen; historisiert („Eternal Truth“)
Zustand	Oft redundant; verteilt und heterogen	konsistent, vereinheitlicht, skaliert, bereinigt
Update	laufend, konkurrierend, transaktional (z. B. ACID)	fortschreibend auf jeder Aggregationsstufe (s. o. Zeitbezug)
Anfragen / Queries	oft statisch im Programm, auf Datensätzen	Variabel, gemäß Anforderungen der Wissensgewinnung

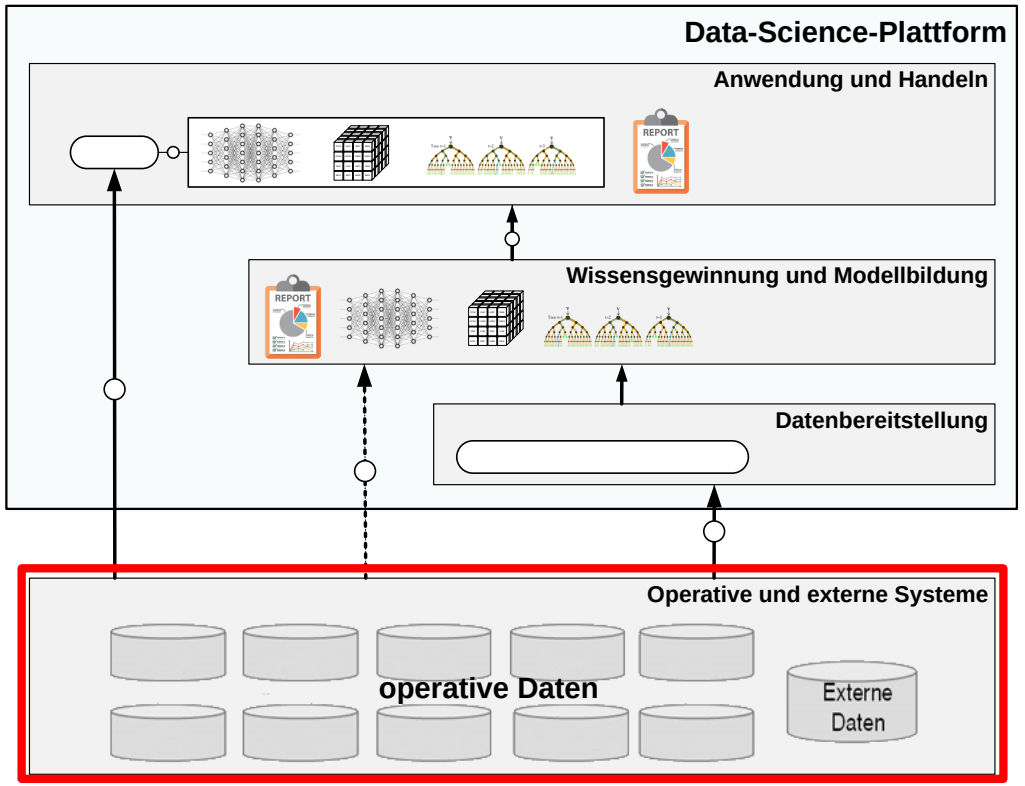
Probleme der Datenaufbereitung

- Verschiedene Quellen und Sichten auf die Daten
- Heterogenität
  - Syntax, Wertebereiche, Schema, Semantik, Pflicht- oder Optional-Felder, Schnittstellen und Formate (Datenbanken, Dateien, Bilder, Streaming), …
- Inkompatibilität
- Inkonsistenz
- Fehlerhafte Daten oder fehlende Genauigkeit der Daten
- Unvollständigkeit, Nicht-Verfügbarkeit, …
- Veraltete Daten
- Redundanz
- Fehlende Vertrauenswürdigkeit
- Fehlende Interpretierbarkeit

Anforderungen

Auswahl der Daten (A)

- Daten aus verschiedenen operativen Systeme und externen Datenquellen
  - Daten aus Geschäftsprozessen betrieblicher Informationssysteme
  - Internet-Daten
  - Sensordaten
  - Maschinendaten
  - …
- Ausrichtung an Bedarfen der Anwendungsdomäne
- Inhaltliche Themenschwerpunkte



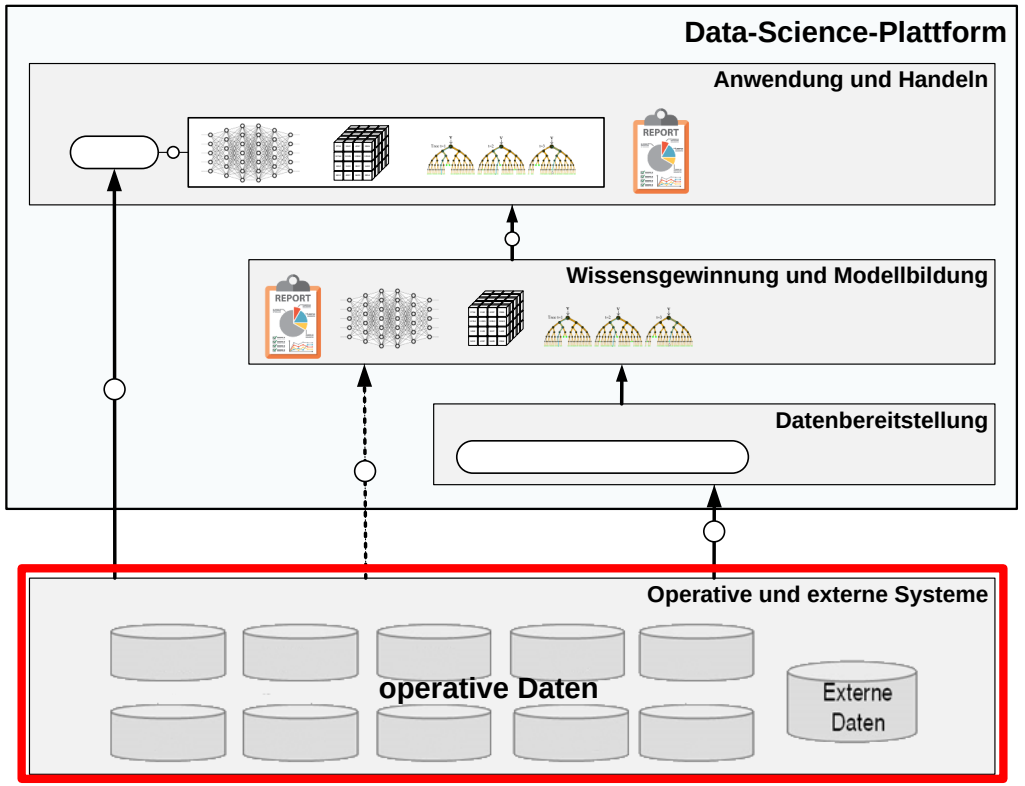
Anforderungen

Vorverarbeitung der Daten (V)

- Qualität der Daten von besonderer Bedeutung
- Bereinigung und Konsolidierung
- Vervollständigung
- aber ggf. auch Original-Daten (Warum?)

Transformation (T)

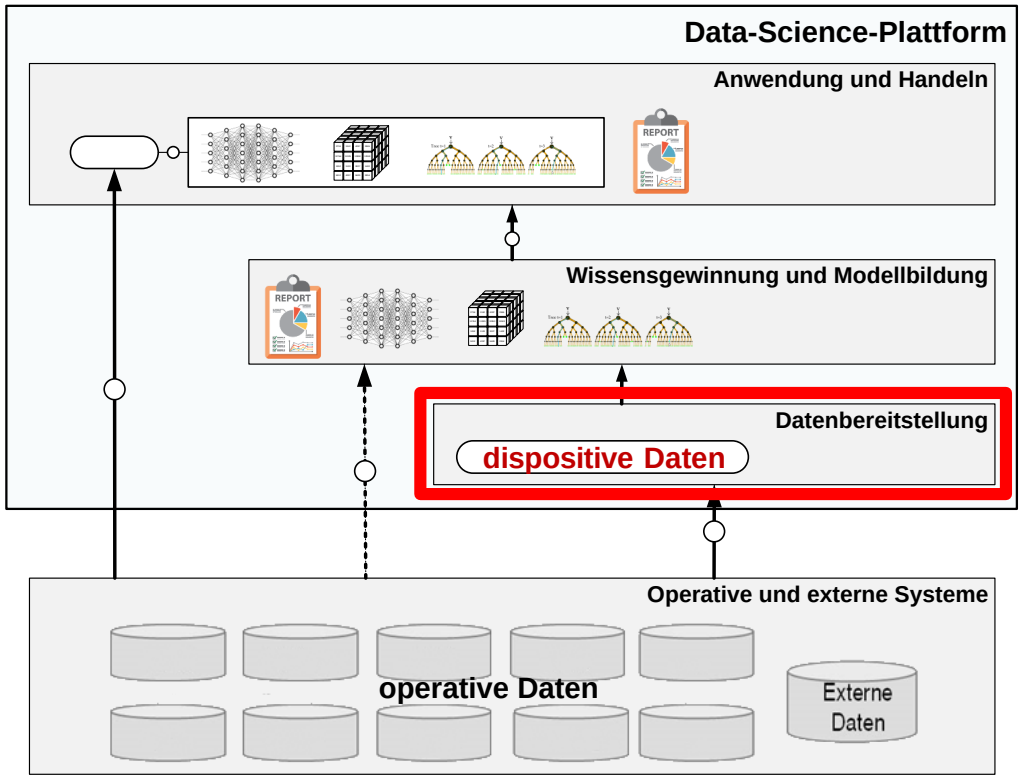
- Erzeugung integrierter Daten
  - vereinheitlicht
  - harmonisiert
  - reduziert, verdichtet
  - abgeleitet



Anforderungen

Datenbereitstellung (D)

- Konsistente dispositive Daten: Basis für Wissensgewinnung, Modellbildung, Anwendung
- Von operativen und externern Quellen unabhängige Datenhaltung
- Zentralisierte oder verteilte dispositive Daten
- Relevanz im Kontext der Belange der Anwendungsdomäne
  - gemäß Fragestellungen & Anforderungen
- Durchführung „beliebiger“ Auswertungen
- Unterstützung individueller Sichten
  - z. B. Zeitrahmen, Struktur, Dimensionen, Verdichtungen etc.



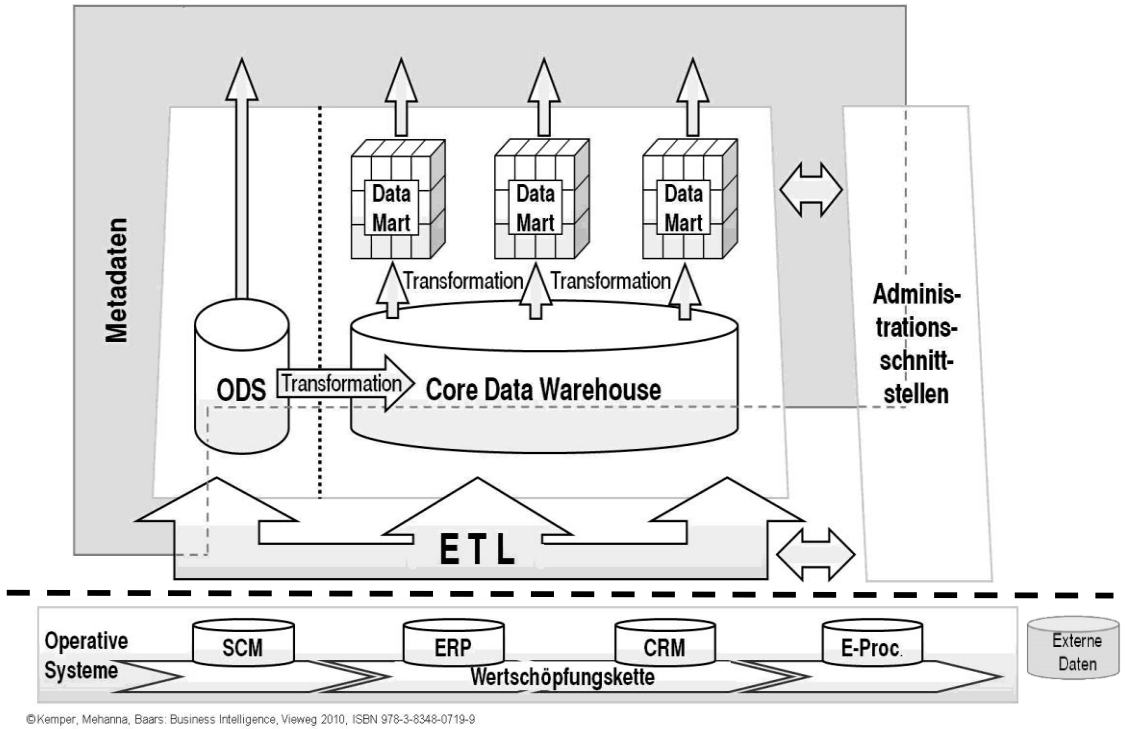
Anforderungen

Generelle Anforderungen (Ende-zu-Ende)

- Automatisierung der Abläufe
- Mehrfachverwendbarkeit und Erweiterbarkeit
  - Integration neuer Quellen und Sichten

Beispiel: Datenaufbereitung in BI-Systemen

- ETL-Prozess (Extract-Transform-Load)



Qualitätsmerkmale der Datenaufbereitung

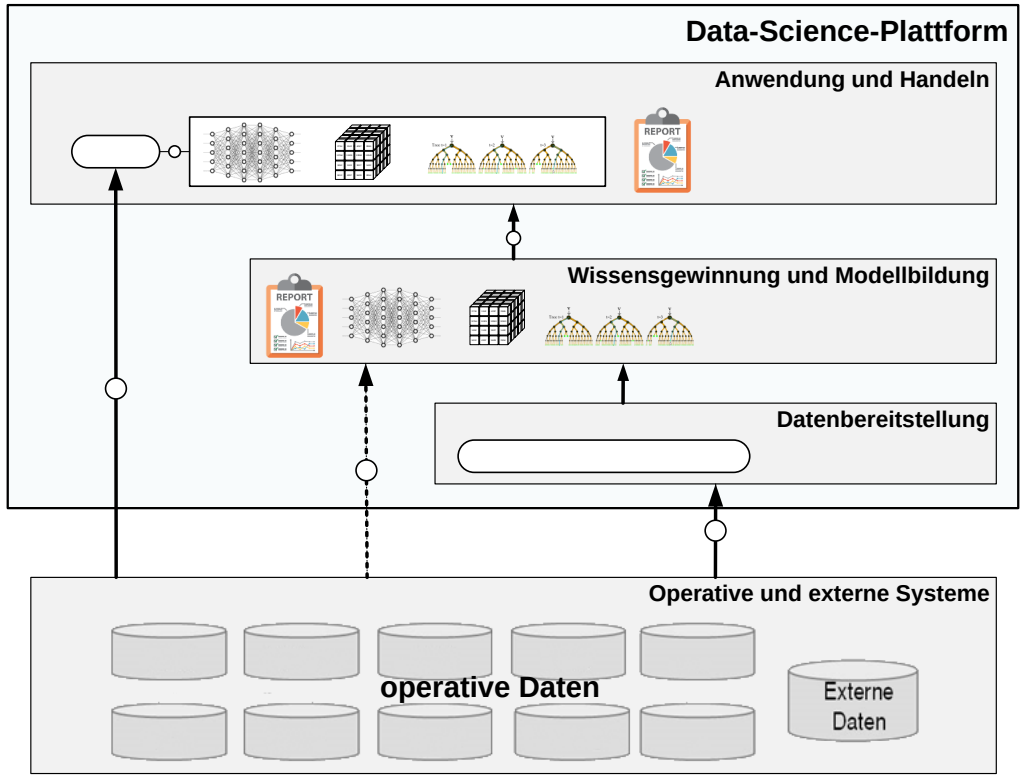
Qualitätsmerkmale

- Relevanz, Zweckdienlichkeit
- Zuverlässigkeit, Glaubwürdigkeit, Nachvollziehbarkeit
- Konsistenz
- Korrektheit
- Verwendbarkeit, geeignetes Format
- Vollständigkeit
  - z. B. keine fehlenden Werte oder Attribute
  - ggf. Originaldaten
- Genauigkeit
  - z. B. Anzahl Nachkommastellen
- Granularität – Tag, Monat, Quartal, Zeitstempel

Konkrete Schritte der Datenaufbereitung

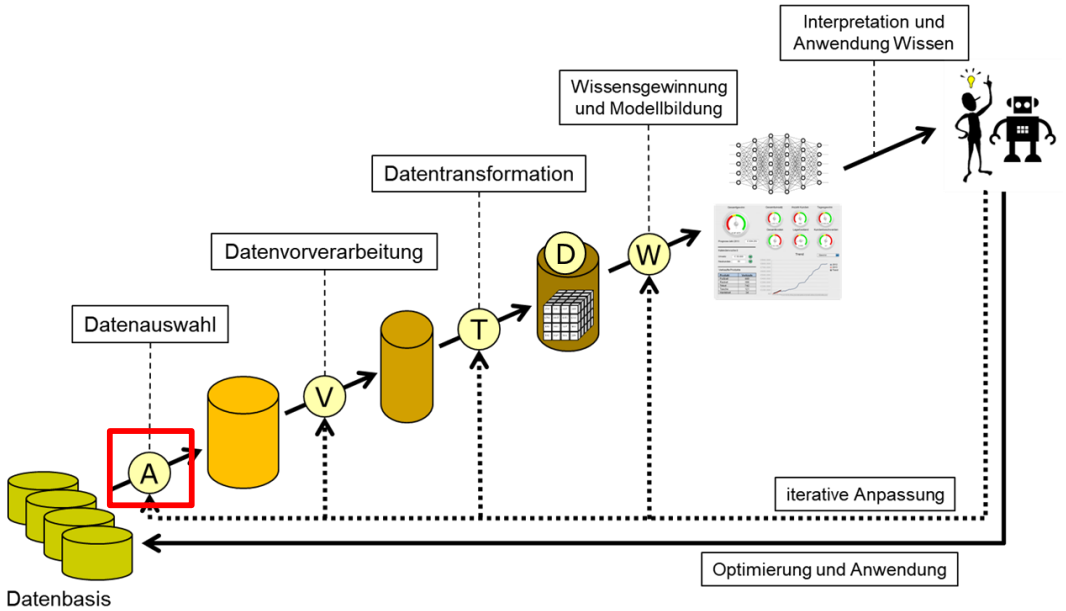
Aufbau der Data-Science-Plattform

- **Abhängig von**
  - Data-Science-Szenario
  - Anforderungen an System
- **Beispiele**
  - λ-Architektur für Big Data und IoT
  - Lokales Data Science Lab



Auswahl der Daten

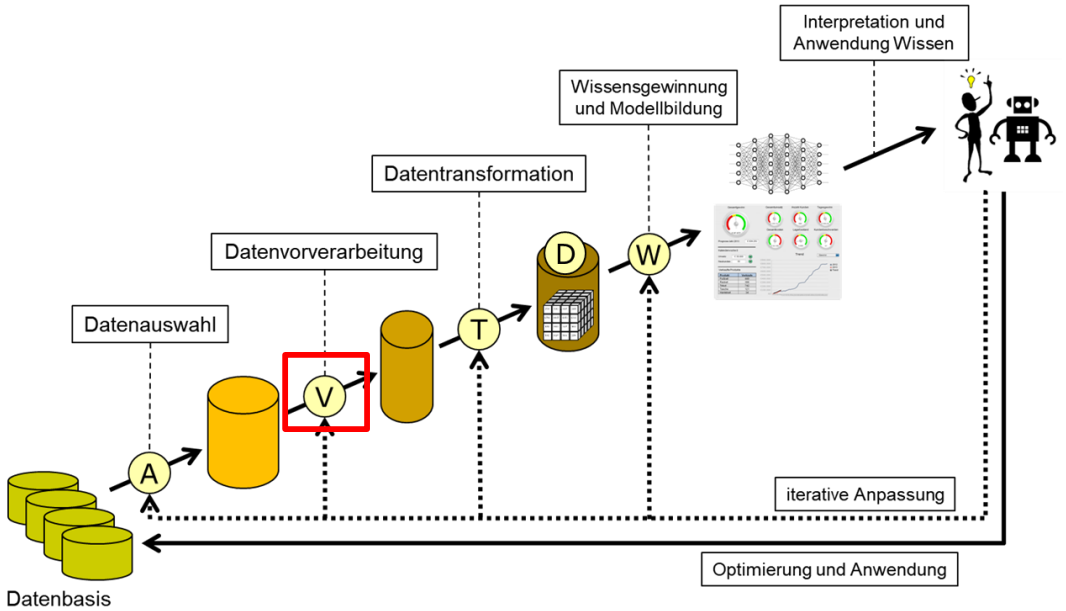
- **Aufgabe**
  - Bestimmung der für das Data-Science-Szenaro relevanten externen Daten und Daten aus den operativen Systemen
  - Übertragung und Zusammenführung dieser Daten in temporäre Arbeitsbereiche (Staging Areas)
- **Strategie** (abhängig von Szenario)
  - periodisch
  - auf Anfrage
  - Ereignisgesteuert (z.B. bei Erreichen einer definierten Anzahl von Änderungen)
  - sofortige Übertragung
- **Realisierung**
  - Nutzung von Standardschnittstellen (ODBC, Streaming, IoT-Protokoll MQTT, …)
  - Ausnahmebehandlung im Fehlerfall



Vorverarbeitung

- **Aufgabe**
  - Extrahierte Daten **bereinigen**
  - Möglicherweise auch zusätzlich: Original-Daten mitführen bis zur Datenbereitstellung
    - Data Lake (Big Data)

- **Realisierung**
  - Nutzung von syntaktischen und semantischen Modellen, Syntax Checker, Rule Engines, etc.
  - ggf. Ausnahmebehandlungen bei nicht automatisiert auflösbaren Konflikten



Vorverarbeitung

- **Bereinigung der Daten**
  - syntaktische Defekte und semantische Defekte
    - Fehlerhafte Werte
    - Fehlende Werte
    - Veraltete Werte
  - Rauschen und Ausreißer (?)
  - …

Vorverarbeitung

- **Syntaktische Defekte**
  - Formale Mängel (Kodierung der Daten)
    - Syntax definiert alle zulässigen Wörter/Zeichenfolgen
- **Semantische Defekte**
  - Mängel, die semantische Inhalte oder die Sinnhaftigkeit betreffen
- **Klassen von Defekten**
  - Klasse 1: Automatisierbare Erkennung, automatisierbare Korrektur
  - Klasse 2: Automatisierbare Erkennung, manuelle Korrektur
  - Klasse 3: Manuelle Erkennung, manuelle Korrektur

Beachte

- Jede Klasse erfordert besondere Behandlung
- Fehler der Klasse 2 und 3 erschweren zeitnahe Datenbereitstellung
  - kritisch bei großen
    Datenmengen und zeitnaher Bereitstellung

Vorverarbeitung

- **Defekte der Klasse 1**
  - Automatisierbare Erkennung und Korrektur
  - Beispiel syntaktischer Defekte:
    - Bestimmte eindeutige fehlerhafte Formate.
      - Behebung durch Mapping-Tabellen, Transformationsregeln
  - Beispiel semantischer Defekte:
    - Fehlende Ist-Werte.
      - Behebung durch Soll-Werte oder Ist-Werte aus Vormonat

Vorverarbeitung

- **Defekte der Klasse 2**
  - Automatisierbare Erkennung und manuelle Korrektur
  - Manuelle Behebung durch technische Spezialisten oder Domänenexperten

  - Beispiele syntaktischer Defekte:
    - Bisher nicht berücksichtigte, eindeutige fehlerhafte Syntaxdefekte
      - Zukünftige Berücksichtigung in Modellen und Algorithmen → wird zu Defekt der Klasse 1
    - Uneindeutige, fehlerhafte Syntaxdefekte
      - Zukünftige Berücksichtigung in Algorithmen und Modellen nicht möglich
      - Konkrete Beispiele?

  - Beispiel semantischer Defekte:
    - Ausreißerwerte oder bisher unbekannte Muster
      - Löschen, separieren oder markieren der Ausreißerwerte und Muster
      - Ggf. zukünftige Berücksichtigung gewisser neuer Muster

Vorverarbeitung

- **Defekte der Klasse 3**
  - Manuelle Erkennung und manuelle Korrektur
  - Manuelles Erkennen und Behebung durch Spezialisten

  - Beispiele semantischer Defekte der Klasse 3:
    - Datenfehler, die nicht durch Plausibilitätsprüfungen, Mustererkennung, etc. identifiziert werden können
    - Ggf. Behebung der Mängel in den Prozessen in den operativen Systemen

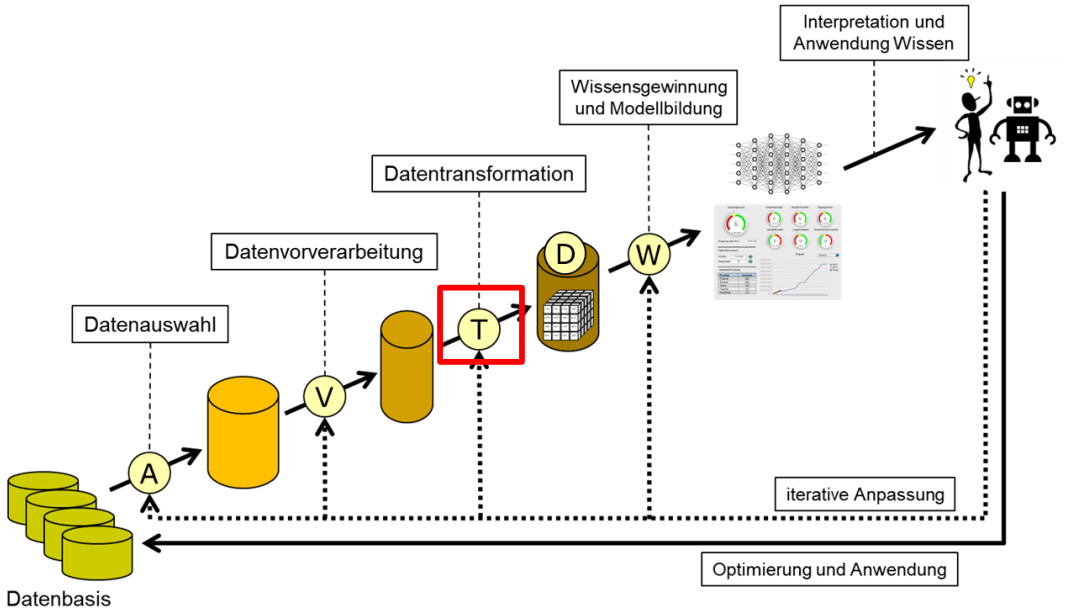
  - Syntaktische Defekte fallen **nicht** in Klasse 3
    - Syntax definiert alle zulässigen Wörter/Zeichenfolgen

Vorverarbeitung

Bereinigung	1. Klasse	2. Klasse	3. Klasse
Syntaktische Defekte	Bekannte Abweichungen und resultierende Anpassungen	Bisher nicht berücksichtigte Syntaxvarianten Nicht eindeutig zu behebende Syntaxfehler	
Semantische Defekte	Fehlende Datenwerte Unstimmige Wertekonstellationen	Fehlende Datenwerte Unstimmige Wertekonstellationen Ausreißer	Nur fachlich erkennbare und korrigierbare Defekte

Transformation

- **Aufgabe**
  - Umwandlung der bereinigten Daten zur Verwendung in den Data-Science-Szenarien
- **Teilprozesse**
  - **Harmonisierung**
    - semantisch und syntaktisch
    - Vergleichbarkeit der Daten und Entitäten
    - Auflösung von Redundanzen
  - **Integration**
    - Reduktion, Verdichtung, Diskretisierung
    - Hierarchiebildung, Aggregation
    - Anreicherung durch weitere Kennzahlen und abgeleitete Daten
    - Erzeugung diverser Sichten, Filterung



Transformation → Harmonisierung

- Bereinigte Daten liegen nach Vorverarbeitung heterogen vor
- Semantische und syntaktische Harmonisierung
  - syntaktische und semantische Konsistenz
  - in feinster Granularität und größtem Umfang
    - feiner und umfänglicher geht es im Folgenden nicht mehr
    - allenfalls können diese Daten noch zu einer gröberen Granularität aggregiert werden
    - Folge: Initiale Granularität und Umfang muss vorab festgelegt werden
- Ggf. Beibehaltung der Original-Daten: Data Lake

Transformation → Harmonisierung

Syntaktische Harmonisierung

- Grund: Operative und externe Daten haben hohe Heterogenität in Form von
  - unterschiedlich kodierten Daten
  - unterschiedlich skalierte Daten
  - Synonymen
  - Homonymen
  - Schemadisharmonien
  - Schlüsseldisharmonien
  - Redundanzen
  - Dimensionsreduktion

  und Kombinationen davon
- Aktion: Auflösung dieser syntaktischen Heterogenitäten und Inkonsistenzen

Transformation → Harmonisierung

Syntaktische Harmonisierung

- **Unterschiedlich kodierte Daten**
  - Daten mit identischen Attributnamen und identischer Bedeutung
  - Unterschiedliche Wertebereiche/Domänen, Einheiten, Formatierungen
- Beispiele:
  - Attribut: `masse` des Produktes
  - Datensätze aus verschiedenen Quellen:
    - `(masse = 10,5)`; Einheit: kg, Domäne: Dezimalzahlen, 1 Nachkommastelle
    - `(masse = 10.512)`; Einheit: kg, Domäne: Dezimalzahlen, 3 Nachpunktstellen
    - `(masse = s)`; Domäne: {**l** < 10kg, 10kg ≤ **m** < 50kg, **s** ≥ 50kg}
    - `(masse = 11.00)`; Einheit: Unze, Domäne: Dezimalzahlen, 2 Nachpunktstellen
- Aktion: Wahl einer gemeinsamen Domäne, Einheit, Formatierung

Transformation → Harmonisierung

Syntaktische Harmonisierung

- **Daten mit gleichem Skalenniveau und unterschiedlicher Skalierung**
  - Attribute von zwei Mengen von Datensätzen haben gleiche Bedeutung und gleiche Skalenniveaus aber unterschiedliche Domänen / Skalierungen

- Beispiel:
  - Einkommen (metrische Verhältnisskalierung mit natürlichem Nullpunkt)
    - Afrika: Zwischen 10 € und 10.000 € pro Jahr
    - Europa: Zwischen 1.000 € und 300.000 € pro Jahr
  - Vergleiche die relativen Einkommensverteilungen der beiden Kontinente
    - Dazu skaliere die Einkommen beispielsweise auf das Intervall [0,100]

Transformation → Harmonisierung

Syntaktische Harmonisierung

- **Synonyme**
  - Daten mit unterschiedlichen Attributnamen und identischer Bedeutung
- Beispiele:
  - Attribut: `Personal`
  - Attribut: `Mitarbeiter`
  - Semantik: jeweils Name der/des Angestellten
- Aktion:
  - Wahl eines gemeinsamen Attributnamens
  - ggf. Wahl einer gemeinsamen Domäne, Einheit, Formatierung

Transformation → Harmonisierung

Syntaktische Harmonisierung

- **Homonyme**
  - Daten mit gleichen Attributnamen und unterschiedlicher Bedeutung
- Beispiele:
  - Attribut: `Partner`,
    - Semantik: Name Kunde
  - Attribut: `Partner`
    - Semantik: Name Lieferant
- Aktion:
  - Differenzierende Attributnamen

Transformation → Harmonisierung

Syntaktische Harmonisierung

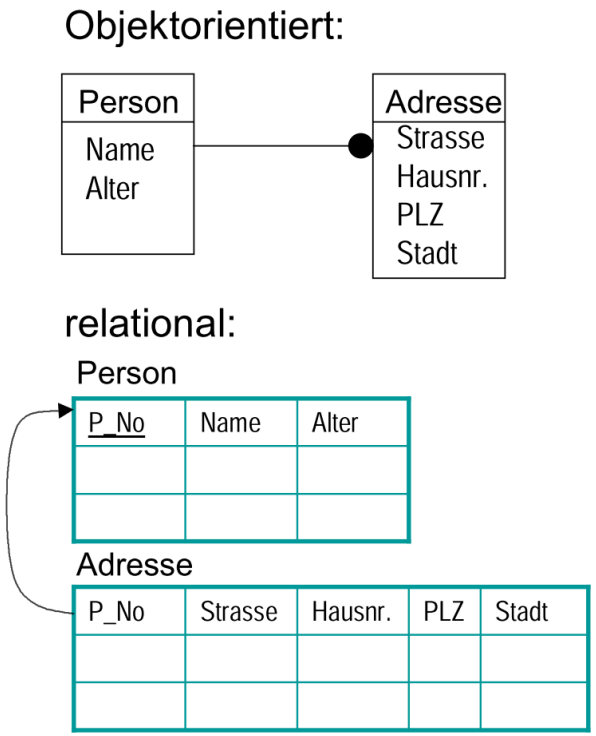
- **Schemadisharmonien**
  - Unterschiedliche Datenmodelle für gleichen Sachverhalt
  - Gleiche Datenmodelle aber unterschiedliche Modellierung
    - Spezialfall: Schlüsseldisharmonien

- Aktion: Wahl eines gemeinsamen, integrierenden Schemas

Transformation → Harmonisierung

Syntaktische Harmonisierung

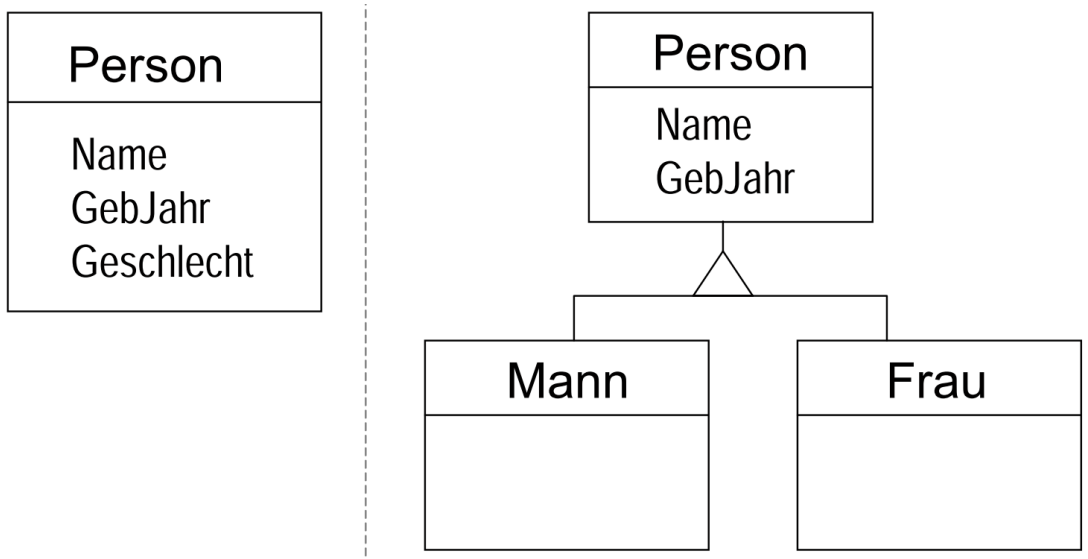
- **Schemadisharmonien**
  - Unterschiedliche Datenmodelle für gleichen Sachverhalt
    - Beispiel: Objektorientiertes vs. relationales Datenmodell



Transformation → Harmonisierung

Syntaktische Harmonisierung

- **Schemadisharmonien**
  - Gleiche Datenmodelle und unterschiedliche Modellierung
    - Beispiel: Klassenmodell mit zwei Modellierungen



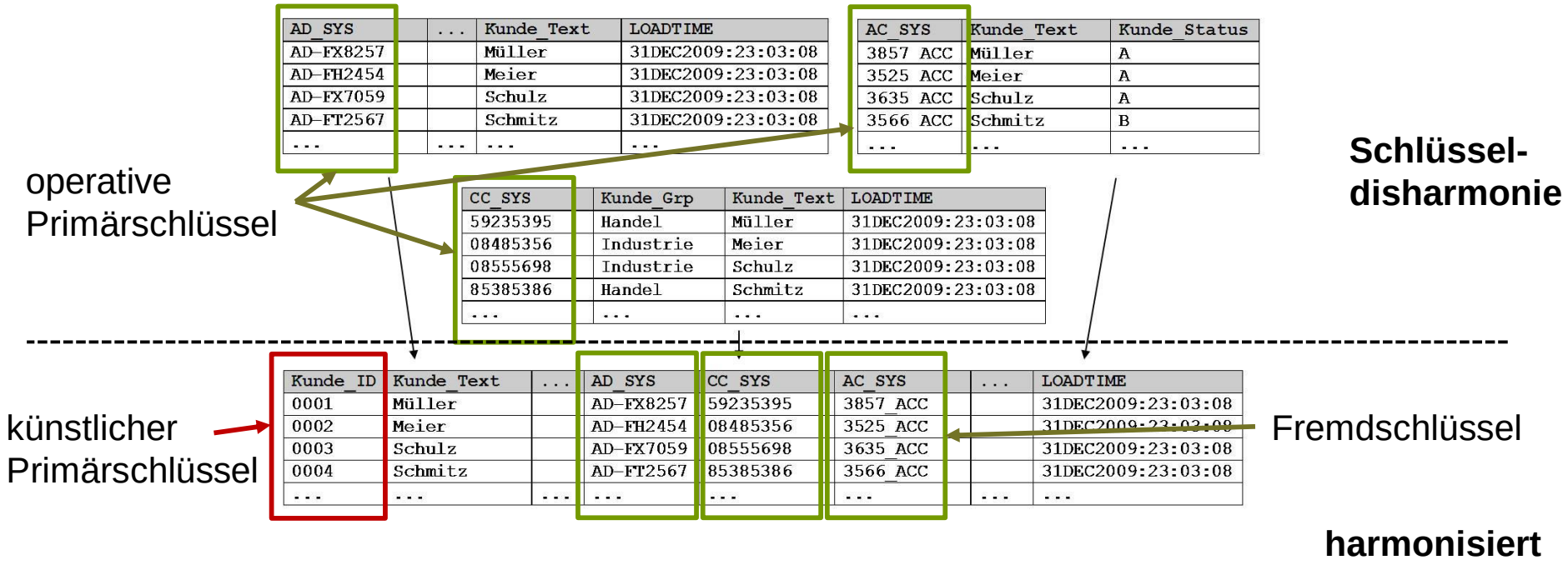
Transformation → Harmonisierung

Syntaktische Harmonisierung

- **Schlüsseldisharmonien**
  - basieren auf Unverträglichkeiten der Primärschlüssel
  - … sind spezielle Schemadisharmonien
- Aktionen:
  - Verschiedene Methoden
  - Beispielsweise:
    - Mapping-Tabellen mit künstlichen Primärschlüsseln
    - Fremdschlüsselbeziehung zu Original-Primärschlüsseln

Syntaktische Harmonisierung

- **Beispiel für Schlüsseldisharmonien**
  - 3 operative Anwendungssysteme mit Kundendaten
  - 3 unterschiedliche Primärschlüssel für die Kunden



Transformation → Harmonisierung

Semantische Harmonisierung

- Syntaktisch harmonisierte Daten werden semantisch harmonisiert

- Gründe
  - Semantische Kennzahlen sind nicht abgestimmt
  - Festlegung der Granularität der bereitzustellenden Daten
  - …

Transformation → Harmonisierung

Semantische Harmonisierung

- **Abgleich der semantischen Kennzahlen**
  - Sicherstellung semantisch konsistenter Daten und Metriken
  - Vereinheitlichung und Abgrenzung der Kennzahlen und deren Berechnung

- Beispiele
  - Umsatz
    - Zusammenführung unterschiedlicher Gebiets- bzw. Ressortgrenzen
    - Einheitliche Periodenzuordnung
  - Prozess-Kennzahlen
    - Abstimmung semantisch nicht einheitlich abgegrenzter Begrifflichkeiten

Transformation → Harmonisierung

Semantische Harmonisierung

- **Festlegung der Granularität der bereitzustellenden Daten**
  - Zusammenführung der Daten
    - auf die im DS-Szenario geforderte feinste Granularität
    - und ggf. für verschiedene Dimensionen

- Beispiel
  - Zusammenfassung der Positionen der Einzelbelege zu tagesaktuellen Werten
  - Sowohl auf Basis von Produktgruppen als auch auf Basis von Kundengruppen

Transformation → Harmonisierung

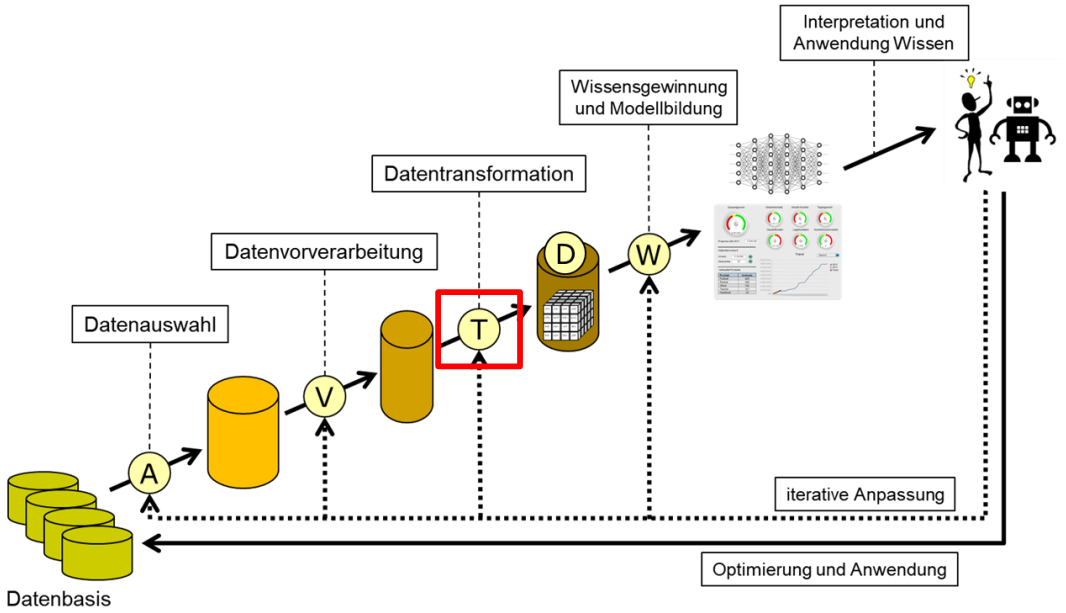
Elimination von Redundanz

- **Daten zu gleichen Sachverhalten werden über mehrere Quellen bereitgestellt**
  - Diese Redundanzen können beseitigt werden
  - Ursprung von Redundanz
    - unsauberes oder gewollt redundantes Design der Anwendung
    - verteilte Datenquellen, die die gleichen Sachverhalte beschreiben

- Beispiel:
  - Datenquelle 1 (vom Verband der Getränkeindustrie):
    - Verbrauchszahlen von Erfrischungsgetränken mit Zeitstempel und Wetterdaten
  - Datenquelle 2 (vom Deutschen Wetterdienst):
    - Wetterdaten mit Zeitstempel

Zur Erinnerung: Transformation

- **Aufgabe**
  - Umwandlung der bereinigten Daten zur Verwendung in den Data-Science-Szenarien
- **Teilprozesse**
  - **Harmonisierung**
    - semantisch und syntaktisch
    - Vergleichbarkeit der Daten und Entitäten
    - Auflösung von Redundanzen
  - **Integration**
    - Reduktion, Verdichtung, Diskretisierung
    - Hierarchienbildung, Aggregation
    - Anreicherung durch weitere Kennzahlen und abgeleitete Daten
    - Erzeugung diverser Sichten, Filterung



Transformation → Integration

Reduktion

- Einschränkung / Projektion auf die tatsächlich relevanten Daten
  - Nicht benötigte Daten können entfernt werden
- Anwendungsfälle
  - Beschränkung auf die wichtigsten Attribute/Parameter: **Principal Component Analysis (PCA)**
  - Verkleinerung des eigentlichen Dateninhalts auf wesentliche Informationen

- Beispiele:
  - Angestelltenverhältnisse der Angestellten der IT-Branche in Deutschland.
    - Dafür werden die Mitarbeiter-Identifikationsnummern der einzelnen Firmen nicht benötigt und können aus den jeweiligen Tabellen entfernt werden
  - Eine Firma interessiert sich für die Einkommensverteilungen ihrer Mitarbeiter
    - Das Attribut Einkommen kann aus den Attributen Grundeinkommen und Boni ermittelt werden und kann aus den Datensätzen entfernt werden
      - Art von Redundanz!

Transformation → Integration

Verdichtung

- Transformation großer Datenmengen zu neuen , kleineren Datenmengen, die für die Belange von Relevanz sind
  - Art Datenreduktion

- Beispiel:
  - Zusammenfassung von Datenreihen zu Balkendiagrammen
  - Transformation von Daten zu Häufigkeitsverteilungen
  - Berechnung von Mittelwerten oder Min-Max-Intervallen
  - …

Transformation → Integration

Diskretisierung

- Ersetzung der eigentlichen Daten durch eine kleine (diskrete) Datenmenge

- Beispiele
  - Intervall-Blöcke
    - Beispiel: Alter wird ersetzt durch Intervalle (oder Labels)
      - A1 ≅ (0,10], A2 ≅ (10,30], A3 ≅ (30,70], A4 ≅ (70,100]
  - Repräsentanten-Werte
    - ersetzen/repräsentieren die Werte der eigentlichen Datenmengen
    - Beispiel: Streaming-Temperaturdaten in 5-min-Zeitfenster
      - Min-Max-Werte, Durchschnittswerte, …
  - Buckets (Labels) fassen Objekte mit bestimmten Datenwerten zusammen
    - thematisch, konzeptionell, ähnlichkeitsbasiert
- Auch Art Datenreduktion

Transformation → Integration

Aggregation

- Erweiterung der Daten um **Dimensionen**
- Daten als Fakten, mit Dimensionen ausgestattet
- Daten aggregierbar entlang der Dimensionen …

➔ **Multidimensionale Daten**

Exkurs – Multidimensionale Daten

Multidimensionale Daten

- Multidimensionale Datenstrukturen bestehen aus
  - Fakt und mehreren Dimensionen
- **Fakt (oder Measure)**
  - operative, betriebswirtschaftliche Daten → Kennzahlen
  - in der Regel numerische Werte, seltener Ordinalwerte oder nominale Werte
  - Beispiele: Kennzahlen, Umsatzerlöse, Einzelkosten, Lagerbestand, etc.
- **Dimensionen**
  - beschreibende Daten für ein Fakt
    - reichern Fakt deskriptiv an
  - ermöglichen unterschiedliche Sichten auf Fakt
  - Beispiele: Lokationen, Regionen, Tage, Quartale, Produkte, Kunden, etc.

Exkurs – Multidimensionale Daten

Multidimensionale Daten

- Beispiele:
  - Datenmodell: **`Bestelldaten`**
    - `Wert`: 10 Euro, `Kunde`: Schmitt, `Stadt`: Mannheim, `Datum`: 2017-03-24
    - **Fakt**: `Wert` (10 Euro)
    - **Dimensionen**: `Kunde` (Schmitt), `Stadt` (Mannheim), `Datum` (2017-03-24)
    - Erweiterung der Dimensionen der Bestelldaten möglich
      - Neue Dimensionen: `Kundenstatus`, `Produktgruppe`, …
  - Datenmodell: **`Sportartikelkette`**
    - Fakt: `Umsatz`
    - Dimensionen: `Ort`, `Zeit`, `Produkt`

Darstellung?

Transformation → Integration

Aggregation

- Aggregierte Abfragen über Fakten entlang einer oder mehrerer Dimensionen
- Mehrere Aggregate für Fakten möglich: sum, avg, median, min, max, …

- Beispiel:
  - Bestelldaten
    - **Wert**: 10 Euro, **Kunde**: Schmitt, **Stadt**: Mannheim, **Datum**: 2017-03-24
    - **Fakt**: Wert (10 Euro)
    - **Dimensionen**: Kunde (Schmitt), Stadt (Mannheim), Datum (2017-03-24)
  - Abfrage auf ursprünglichen Daten: Umsatz durch Kunde Schmitt in Mannheim?
  - Aggregierte Abfragen
    - Umsatz (sum) durch Kunde Schmitt in Baden-Württemberg?
    - Durchschnittlicher Warenkorbeinkaufswert (avg) in Filiale Mannheim

Transformation → Integration

Aggregation

- Aggregate
  - in zusätzlichen Aggregatstabellen zwecks Optimierung der Performanz
  - oder als neue Daten, die die eigentlichen Daten ersetzen/komprimieren
  - oder als Sichten über den eigentlichen Daten

Transformation → Integration

Anreicherung

- Berechnung weiterer Kennzahlen (Fakten) aus den eigentlichen Daten
  - sowohl auf Basis ursprünglichen (harmonisierten) Daten
  - als auch auf Basis der verdichteten, diskreten, aggregierten Daten

- Anreicherung der Datenbasis durch Integration dieser Kennzahlen

- Beispiele
  - Tägliche Umsätze auf Produktebene
    - auf eigentlichen Daten
  - Jährliche Umsätze über alle Produkte auf Regionenebene
    - auf aggregierten Daten
  - Prozesslaufzeit eines Order-to-Casch-Prozesses
    - von Bestellungseingang bis Geldeingang auf Konto
    - oder von Bestellugseingang bis Auslieferung aus Lager

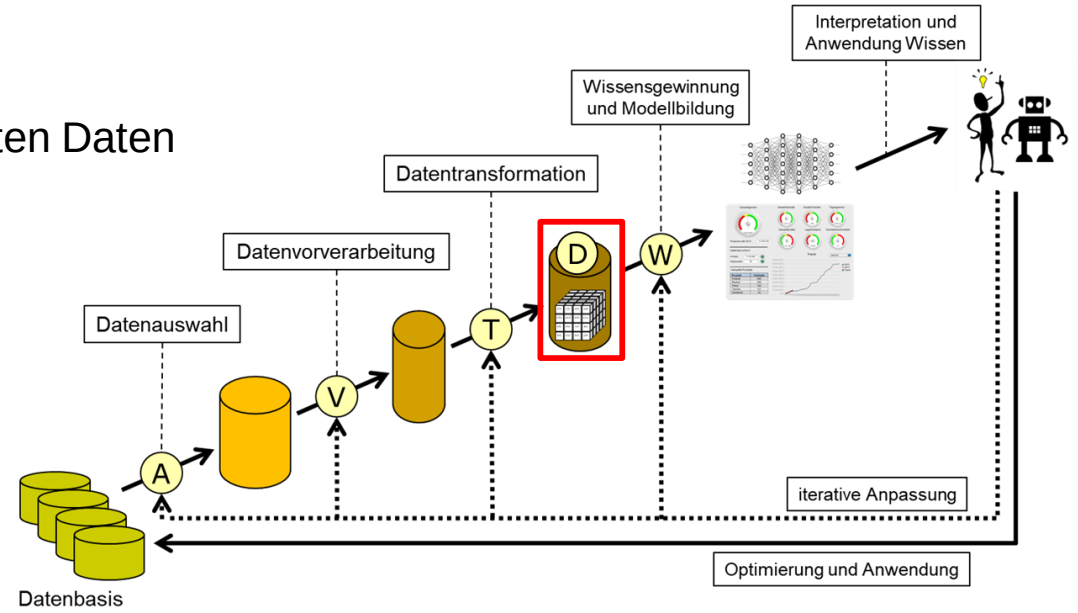
Transformation → Integration

Anreicherung

- Vorteil der Anreicherung / Kennzahlenberechnung
  - Vorberechnung der Kennzahlen erhöht Performanz
    - Kalkulierbares Antwortzeitverhalten bei späteren Abfragen
  - Konsistenz der berechneten Werte
    - nur einmal anwendungs- bzw. szenarioübergreifend gebildet
  - Etablierung gesamtbetrieblicher, semantisch abgestimmter Berechnungsmethoden

Datenbereitstellung

- **Aufgabe**
  - Datenbereitstellung der extrahierten, vorverarbeiteten, harmonisierten Daten
  - für den zentralen Schritt der Wissensgewinnung und Modellbildung
- **Strategie** (abhängig von Szenario)
  - persistent
  - transient
  - Mischform
- **Realisierung** (abhängig von Szenario)
  - persistent
    - Datenbanken: SQL, NO-SQL
    - Data Warehouse
    - in verschiedenen Schichten/Stufen
  - transient
    - Streaming-Kanäle: Online-processing, Windowing, …
    - Messaging: pub-sub, push-pull, …



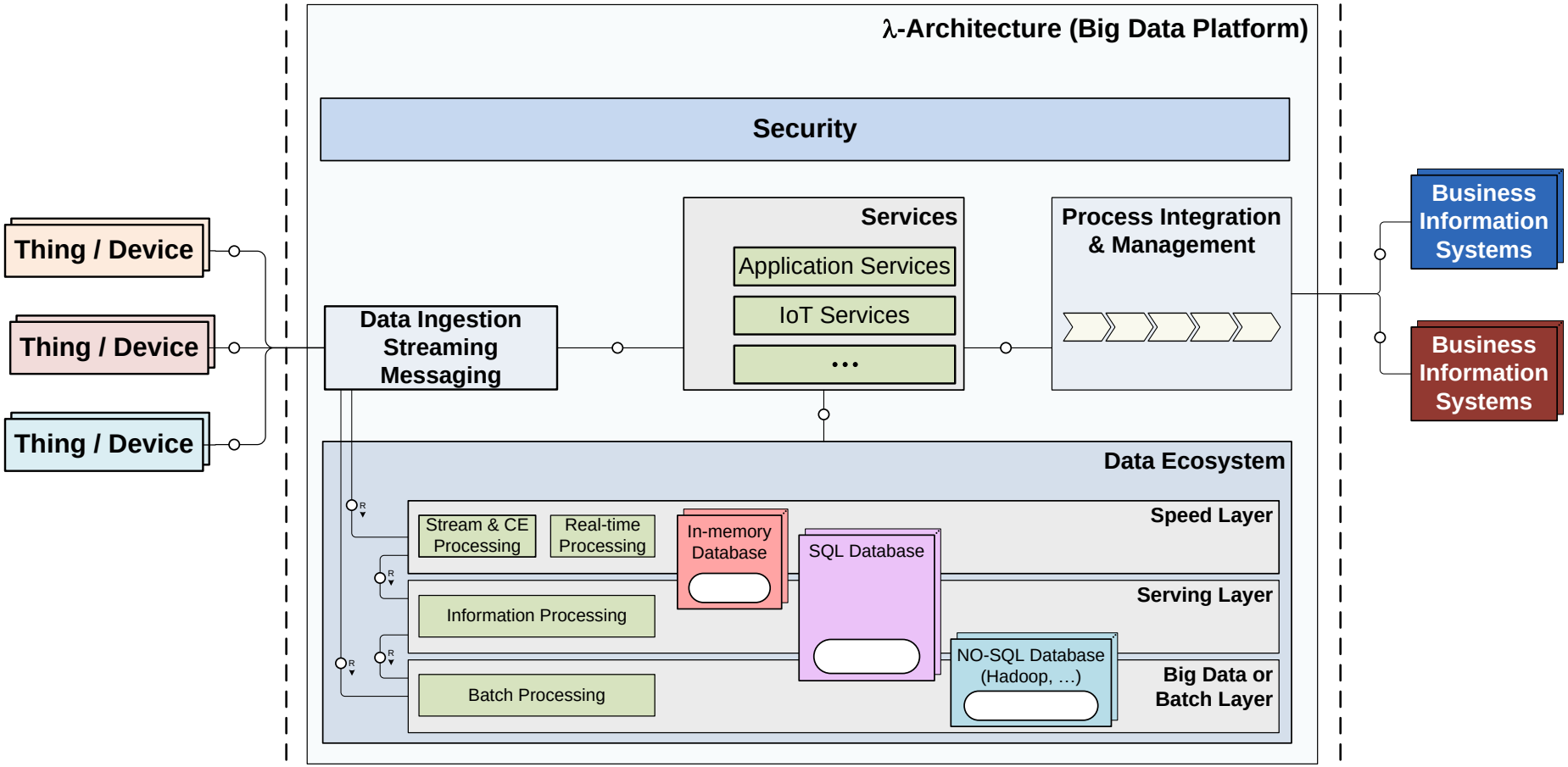
Datenbereitstellung

Konzepte und Architektur der Datenablage – und letztendlich der gesamten Data Science Platform – hängen von verschiedenen Faktoren ab:

- Anforderung an die Verwendung der Daten
  - Re-use der Daten: persistent oder transient
  - Struktur der Daten
  - Alter der Daten im Kontext ihrer Verwendung
    - Stichworte: hot, warm, cold
  - Historisierung und Wiederaufsetzbarkeit, „Eternal Truth“
- Anforderung an
  - nachfolgende Wissensgewinnung und Modellbildung
  - nachfolgende Verwendung von Daten, des Wissens und der Modelle
    - Business Intelligence in Unternehmen
    - Sprachverarbeitung und Bilderkennung
    - Online-Szenarien
      - IoT, Predictive Maintenance und Alerting
      - Gesichtserkennung
      - selbstfahrende oder autonome Fahrzeuge, Robotik, KI

Datenbereitstellung

- **Beispiel**
  - Data Science im Kontext von λ-Architekturen für Big-Data- und IoT-Szenarien



Datenaufbereitung – Abschließende Diskussion

- Warum überhaupt Datenaufbereitung?
- Warum nicht gleich Wissensgewinnung auf operativen und externen Daten?
- Diskussion von Data-Science-Szenarien und möglichen Strategien für die Datenaufbereitung im Kontext dieser Szenarien

Fragen?