Big Data

Einführung

Lernziele

Lerneinheiten und Inhalte		
Lehr- und Lerneinheiten	**Präsenzzeit**	**Selbststudium**
Big Data	**36,0**	**39,0**
Big Data Programming - Einführung in das Themengebiet Big Data-Programmierung - Erläuterung der horizontalen Skalierung von Systemen bei der Verarbeitung digitaler Massendaten - Einführung in die verteilte Verarbeitung digitaler Massendaten - Einführung in Batch- und Stromverarbeitung - Vorstellung aktueller Frameworks, Bibliotheken, Programmiersprachen, etc. - Umsetzung von Praxisbeispielen Big Data Storage - Einführung in das Themengebiet Big Data-Storage - Erläuterung der horizontalen Skalierung von Systemen bei der Speicherung digitaler Massendaten - Einführung in die Speicherung digitaler Massendaten unter Nutzung verschiedener Speicher- und Zugriffsarten (Dateisysteme, Datenbanken, etc.) - Vorstellung aktueller Frameworks, Bibliotheken, Programmier- und Abfragesprachen, etc. - Umsetzung von Praxisbeispielen		
Internet of Things	**36,0**	**39,0**
- Einführung in IoT - Anwendungsgebiete - Technologien (auf einer aktuellen IoT-Plattform) - Kommunikationsprotokolle - Sensorik und Datenerfassung - Plattformen		

Literatur

1. Earl, Th., Khattak, W. , Buhler, P. (2016). *Big Data Fundamentals: Concepts, Drivers & Techniques.* Boston, Columbus, Indianapolis: Prentice Hall
2. Freiknecht, J., Papp, S. (2018). *Big Data in der Praxis.* München: Carl Hanser Verlag
3. D'Onfrio, S., Meier, A. (Hrsg.) (2019). *Big Data Analytics*, HMD: Praxis der Wirtschaftsinformatik, Band 56. Wiesbaden: Spinger Vieweg
4. Ghavami, P. (2020) *Big data analytics methods: analytics techniques in data mining, deep learning and natural language processing.* Boston, Berlin: De Gruyter
5. Burk, S., Miner, G. (2020). *It's all analytics!: the foundations of AI, big data, and data science landscape for professionals in healthcare, business, and goverment.* Boca Raton; CRC Press
6. Otte, R., Wippermann, B, Schade, S., Otte, V. (2020). *Von Data Mining bis Big Data: Handbuch für die Industrielle Praxis.* München: Hanser
7. Borgmeier, A., Grohmann, A., Gross, S. (2022). *Smart Services und Internet der Dinge: Geschäftsmodelle, Umsetzung und Best Practices*. München: Hanser

Big Data – Wozu?

Wofür brauchen wir Daten?

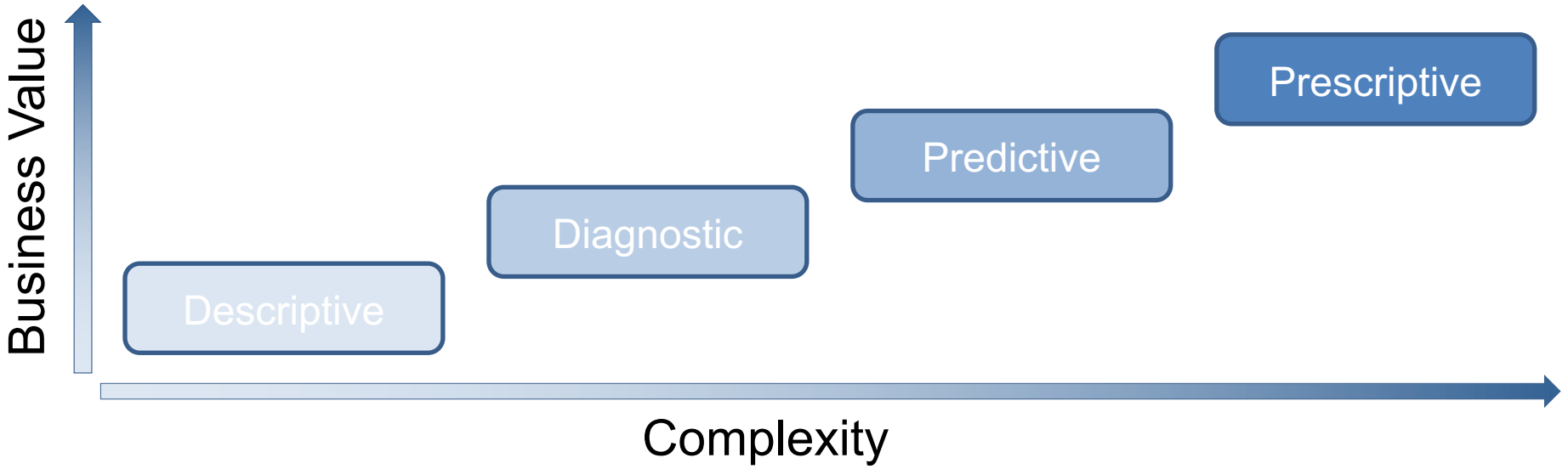
Big Data – Wozu?

Wofür brauchen wir Daten?

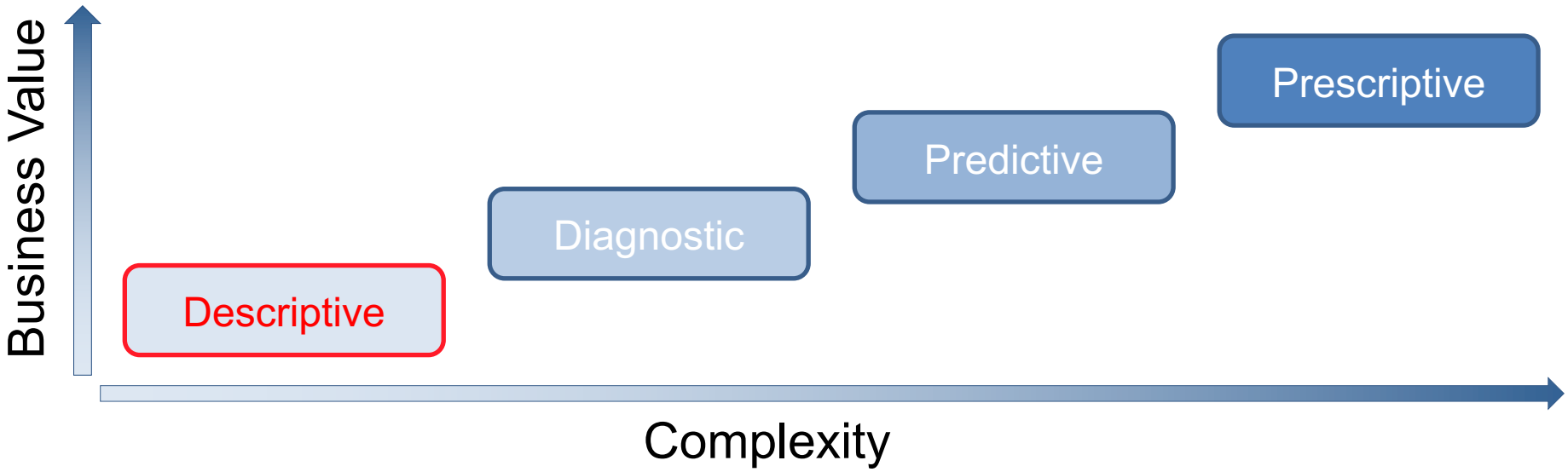
1. Daten sind das Produkt (Wetterdienst, Google & Co, interne R&D)
2. Daten werden für „Data Driven Decision Making“ benötigt, z.B.
   a) Prozessoptimierung
   b) Investitionsentscheidungen
   c) Neue Produkte
   d) Precision Farming
   e) Precision Medicine

Von Descriptive bis Prescriptive

- Um bei der Entscheidungsfindung helfen zu können, müssen Daten analysiert d.h. in Zusammenhang gesetzt werden
- Dabei ist es wichtig, eine möglichst klare Fragestellung zu haben!

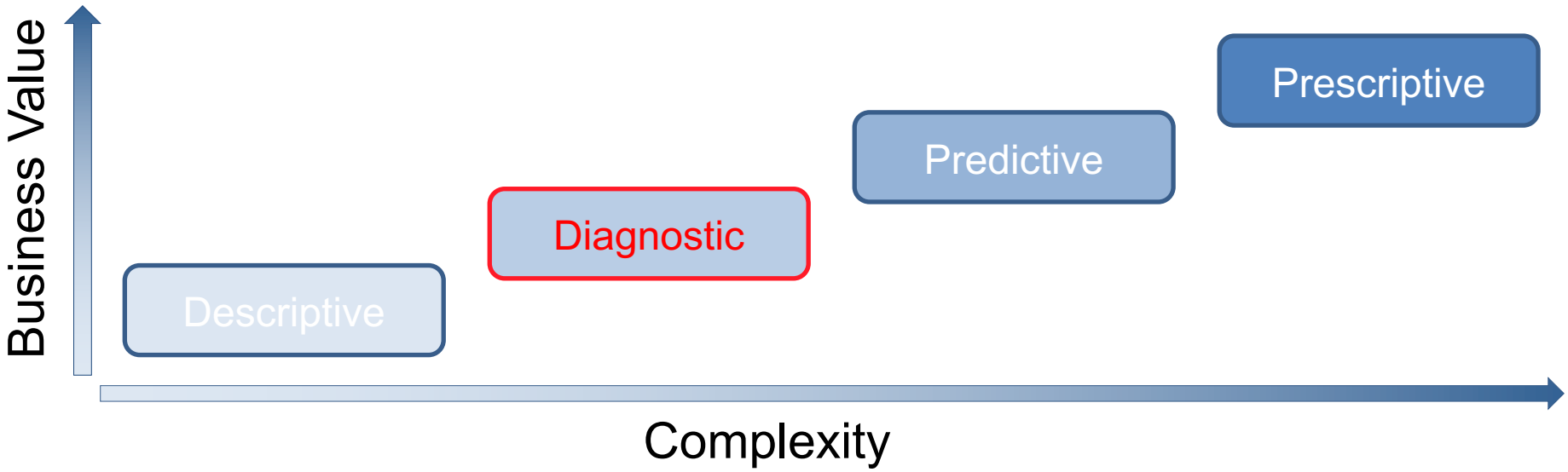


Von Descriptive bis Prescriptive



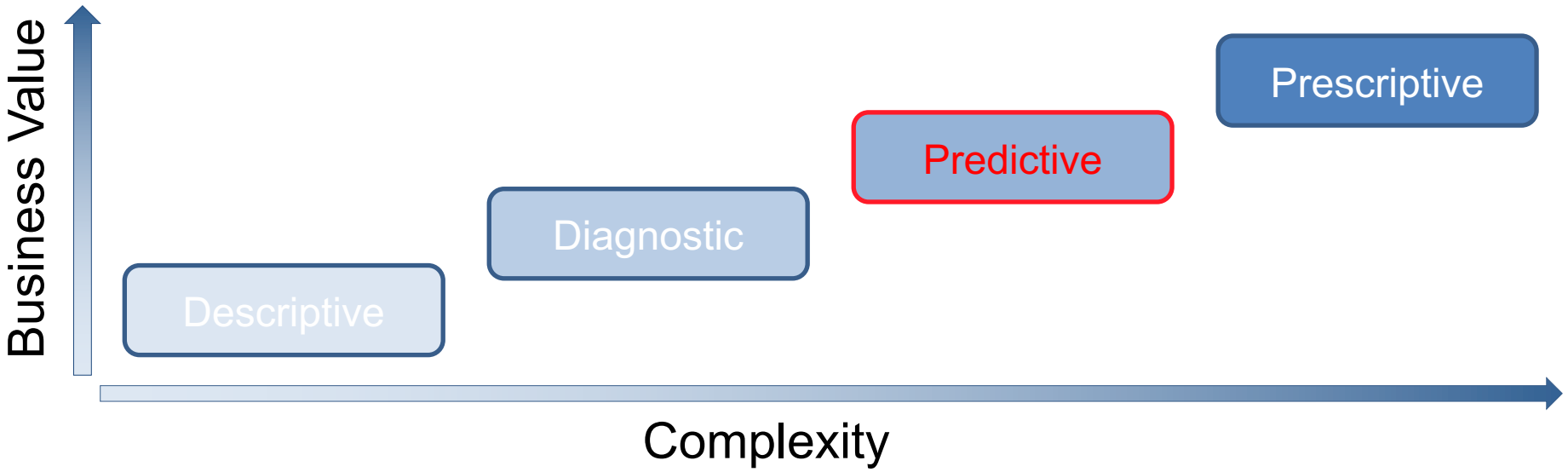
- Beantworten Fragen zu Ereignissen, die in der Vergangenheit stattgefunden haben
- Großteil aller Analytics-Fragestellungen
  - “Was war der Umsatz im Q4/2019?”
  - “Wie viele Support Calls hatten wir von den 10 wichtigsten Kunden?”

Von Descriptive bis Prescriptive



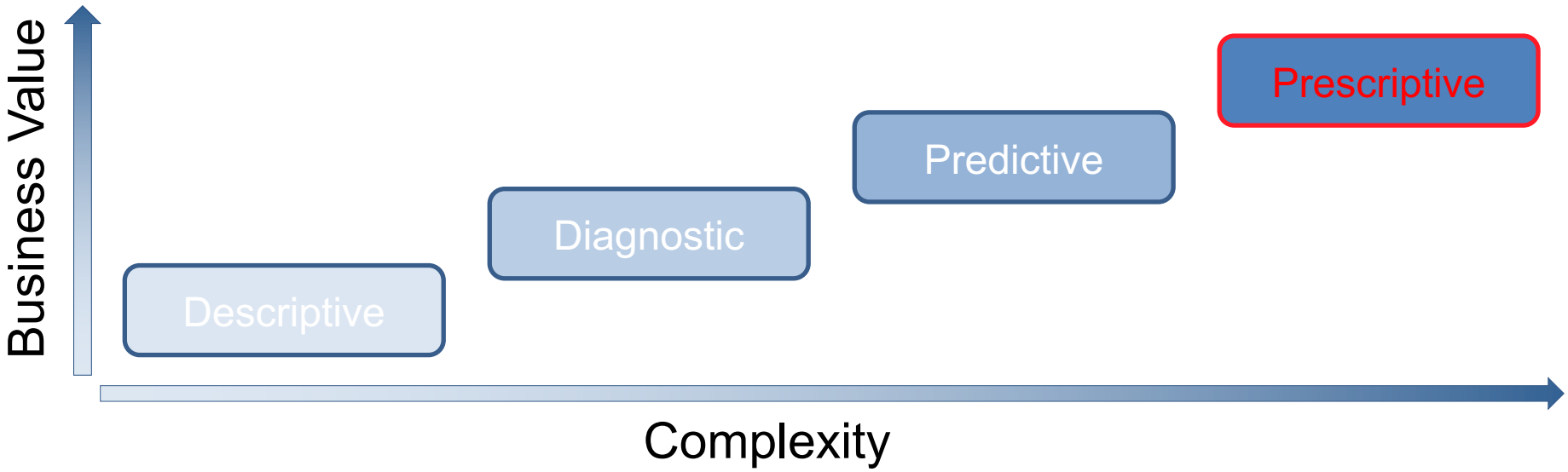
- Untersucht Ursache – Wirkungs –Beziehungen von Ereignissen in der Vergangenheit
- Erfordert oft die Kombination mehrerer Datenquellen
  - “Warum sind die Umsätze im 2. Quartal zurückgegangen?”
  - ”Warum ist die Lebensdauer der Förderpumpen in der Produktion gesunken?”

Von Descriptive bis Prescriptive



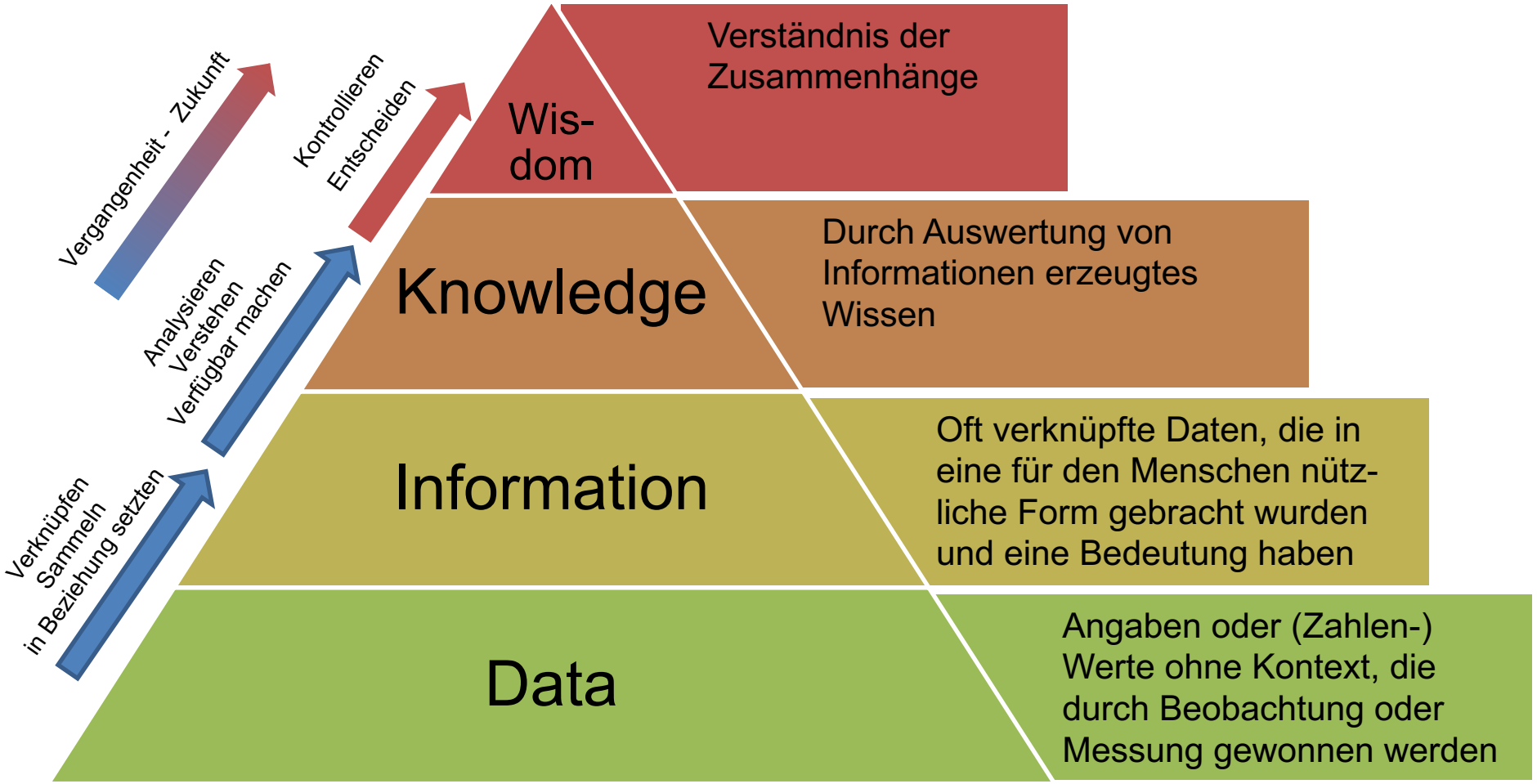
- Versucht den Ausgang eines Ereignisses in der Zukunft vorauszusagen
- Vorhersagen basieren auf Mustern und Trends, die in historischen oder aktuellen Daten gefunden werden.
  - “Wie groß ist das Risiko, dass dieser Kunde seinen Kredit nicht zurückzahlt?”
  - ”Wann wird die Förderpumpe in der Produktion kaputt gehen?”

Von Descriptive bis Prescriptive



- Kann eine von mehreren Handlungsoptionen empfehlen und diese Empfehlung begründen. Basiert oft auf dem Durchspielen mehrerer Szenarien unter Kenntnis der situativen Zusammenhänge.
  - „Wann sollten wir das neue Produkt auf dem Markt bringen?“
  - Welches Medikament verspricht den höchsten Umsatz?“

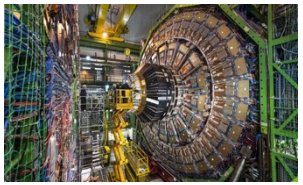
Die Knowledge Management Pyramide



Die 3 bis 5 V

Was ist anders bei

Big Data?

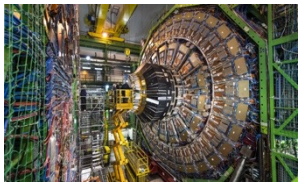


Die 3 bis 5 V

Was ist anders bei

Big Data?

Volume



Velocity

Variety

Veracity

Value

Die 3 bis 5 V - Volume

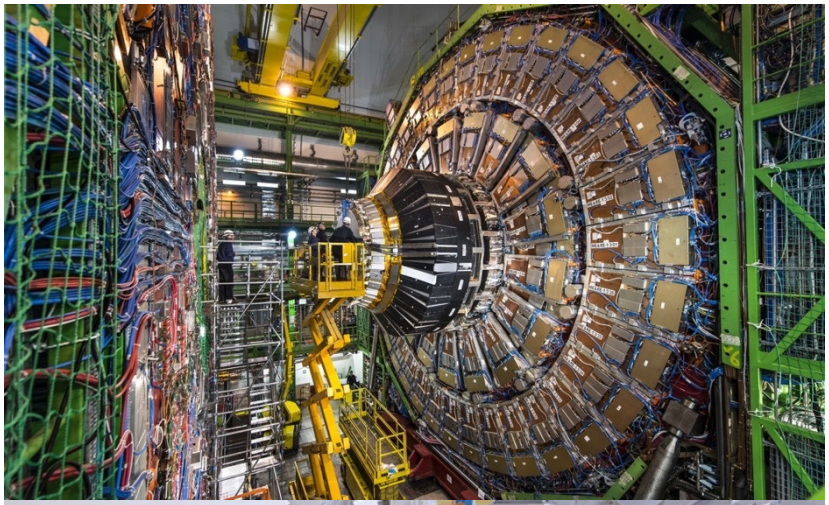
Großes Datenvolumen

- 250 Milliarden Fotos in Facebook, 4 PByte werden täglich von Facebook-Nutzern generiert
- Sensoren in Rolls-Royce Triebwerke erzeugen 70 Billionen Datenpunkte pro Jahr
- DWD speichert mehrere Pbyte an Wetterdaten
- Insgesamt 175 ZByte bis 2025

Die 3 bis 5 V - Velocity

Großes Datengeschwindigkeit

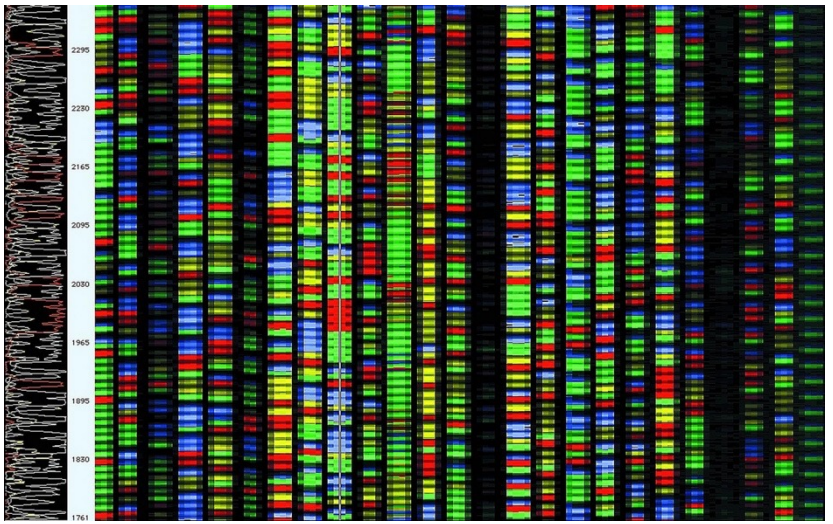
- LHC im Cern erzeugt 25 GB pro Sekunde
- Twitter User senden 6000 tweets pro Sekunde
- 50 Stunden Videos werden pro Sekunde auf YouTube geladen
- 2.8 Millionen Emails werden jede Sekunde versendet



Die 3 bis 5 V - Variety

Unterschiedliche Formate und Datentypen

- (Firmeninterner) Datenbanken wie ERM, CRM, MES etc.
- Texte wie Emails, Dokumente, Tweets, Messages
- Mediendaten wie Fotos, Videos, Audio-Dateien
- Logfiles
- Sensor-Daten
- Wissenschaftliche Daten (Wetter, Genom, ...)



Die 3 bis 5 V - Veracity

Unterschiedliche Qualität bzw. Glaubwürdigkeit

- Hängt von der Quelle ab und wie die Daten erhoben wurden (z.B: Peer-Reviewed Scientific Paper gegenüber anonymen Blogbeitrag)
- Je nach Qualität können zusätzliche Daten das Signal oder das Rauschen einer Analyse verstärken

Die 3 bis 5 V - Value

Wert – Wie groß nützlich sind die Daten bzw. die aus der Analyse gewonnenen Erkenntnisse?

Abhängig u. A. von

- Qualität (Veracity) der Daten
- Wie schnell / zeitnah die Analyse durchgeführt wird
- Ob die Daten für die Fragestellung angemessen sind
- Ob die Erkenntnisse in die Entscheidungsfindung im Unternehmen einfließen

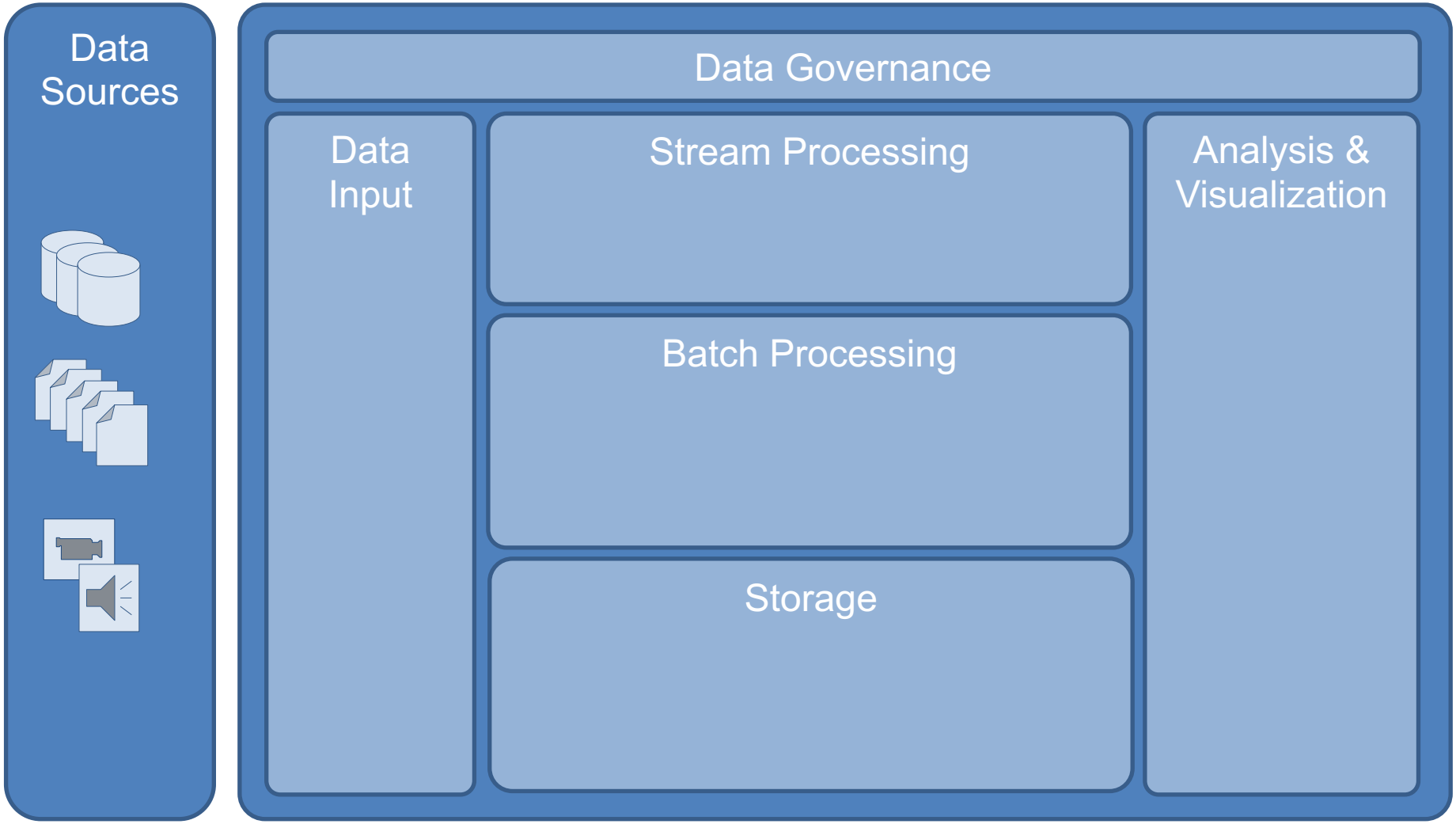
Was ist Data Analytics

Data Analytics (Überbegriff)

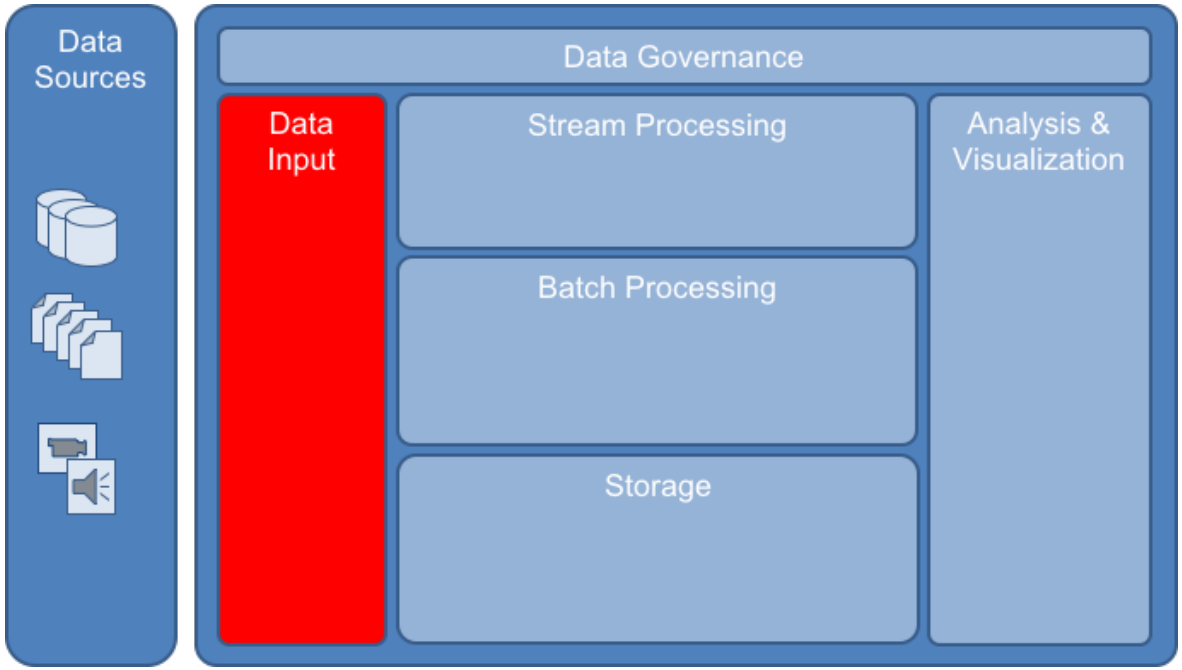
Management des gesamten Lebenszyklus der Daten (Data Lifecycle)

1. Sammeln
2. Bereinigen (cleansing)
3. Organisieren
4. Speichern
5. Analysieren (Data Analysis)
6. Governance

Big-Data Referenzarchitektur



Data Input

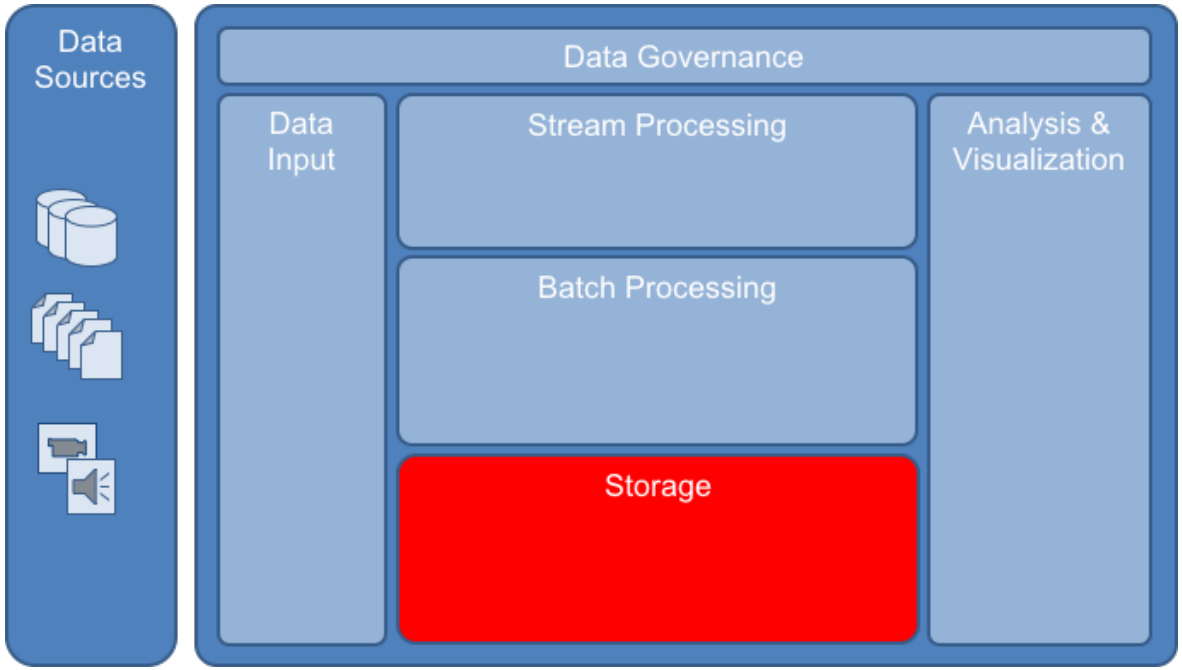


Tools

- Kafka
- Flume
- Sqoop

- Welche Aktivitäten müssen beim Daten-Input erledigt werden?
- ETL vs. ELT
- Wie geht man mit den unterschiedlichen Datenquellen um?

Data Storage

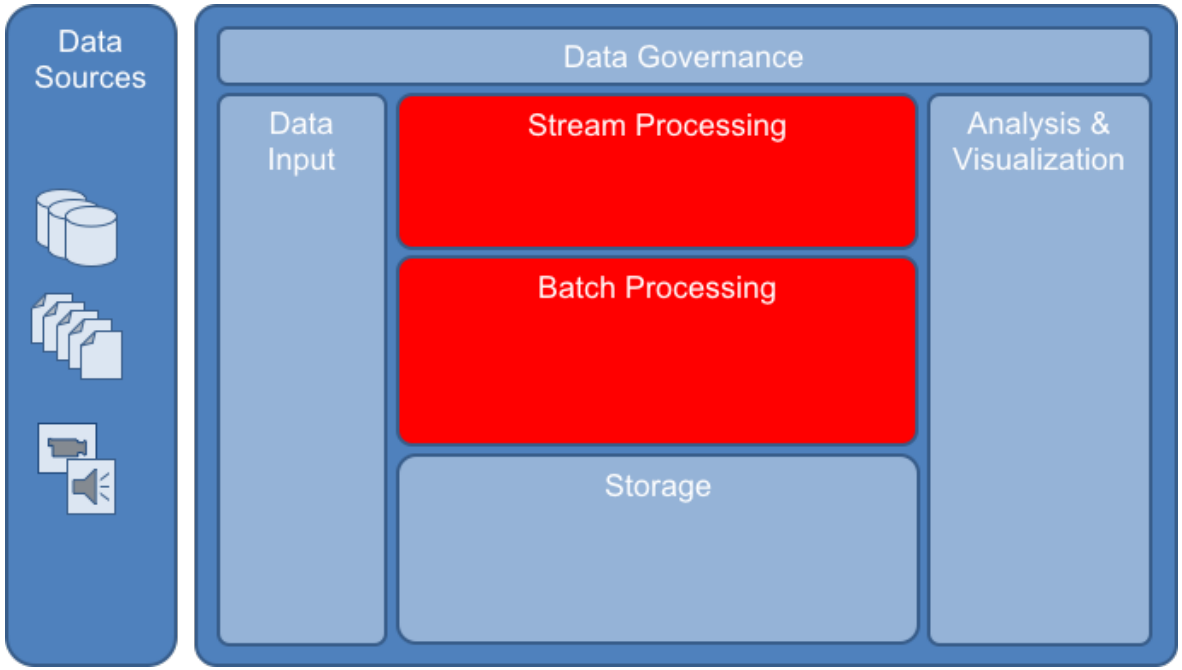


Tools

- HDFS
- HBase
- Hive

- Wie können Speicherlösungen skaliert werden?
- Speichertypen
- Prinzipien: CAP, ACID, BASE
- On-Disk vs. In-Memory

Data Processing

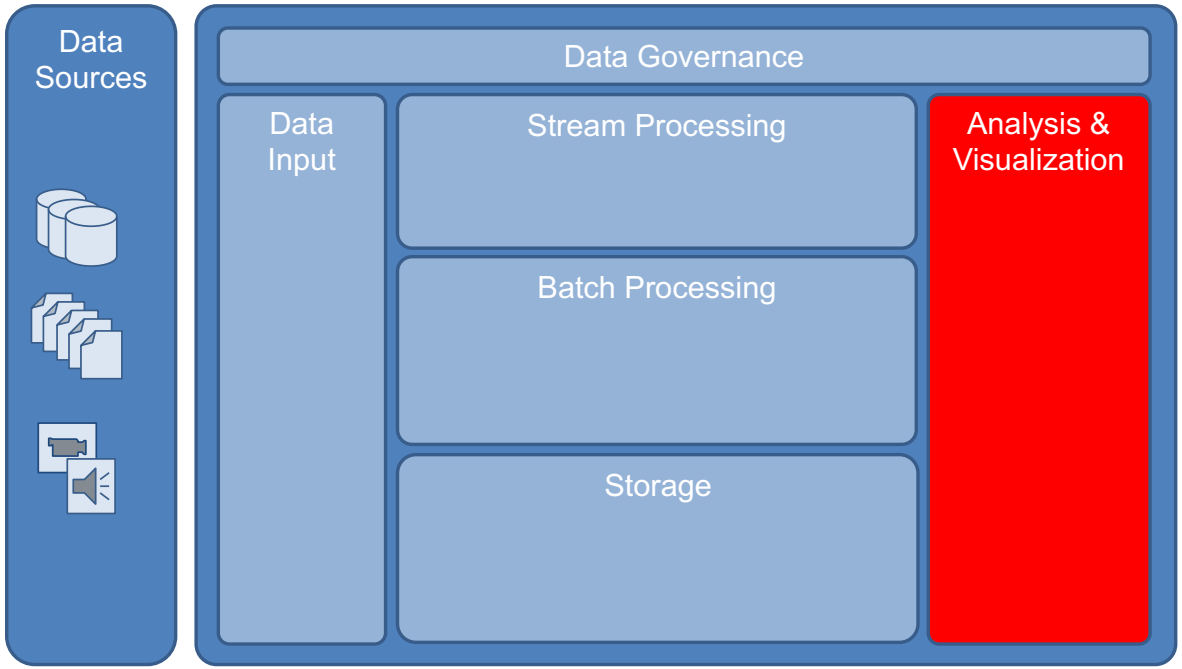


Tools

- MapReduce
- Mahout
- Spark

- Wie kann das Data Processing skaliert werden?
- Batch vs. Stream Processing
- MapReduce
- Statistische Methoden, Semantische Methoden, Machine Learning

Data Analysis & Visualisation

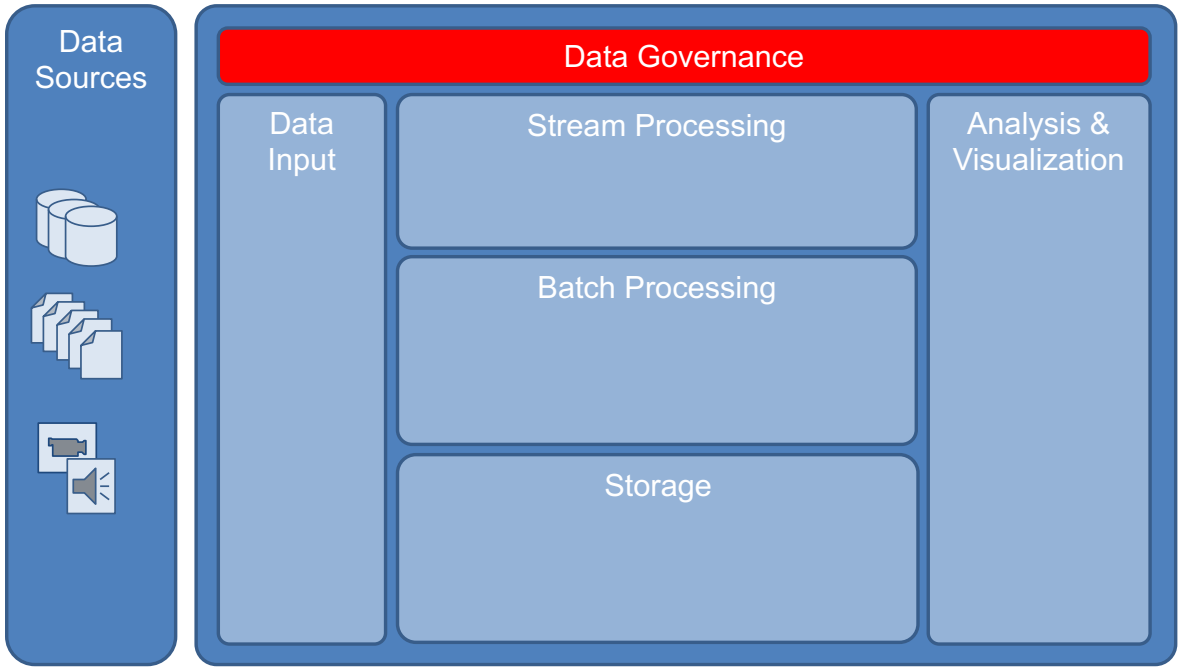


Tools

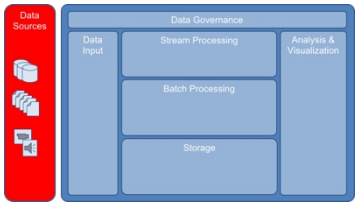
- Power BI
- QlikView
- Spotfire
- R

- Wie kann man Daten so darstellen, dass Zusammenhänge für Menschen schnell ersichtlich sind?
- Wie kann man Benutzer bzw. die Entscheidungsträger im Unternehmen befähigen, mit den Daten zu spielen / zu arbeiten?

Data Governance

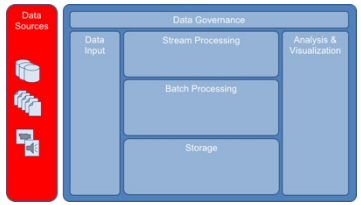


- Was ist (Data) Governance?
- Business-Wert aus Daten, Einführung im Unternehmen
- Daten Qualität, Master Data, Meta Daten
- Zugriffsrechte, Datenschutz, Sicherheit

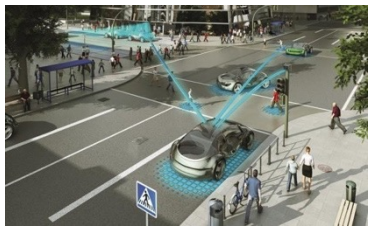


DATA SOURCES

Typen von Daten



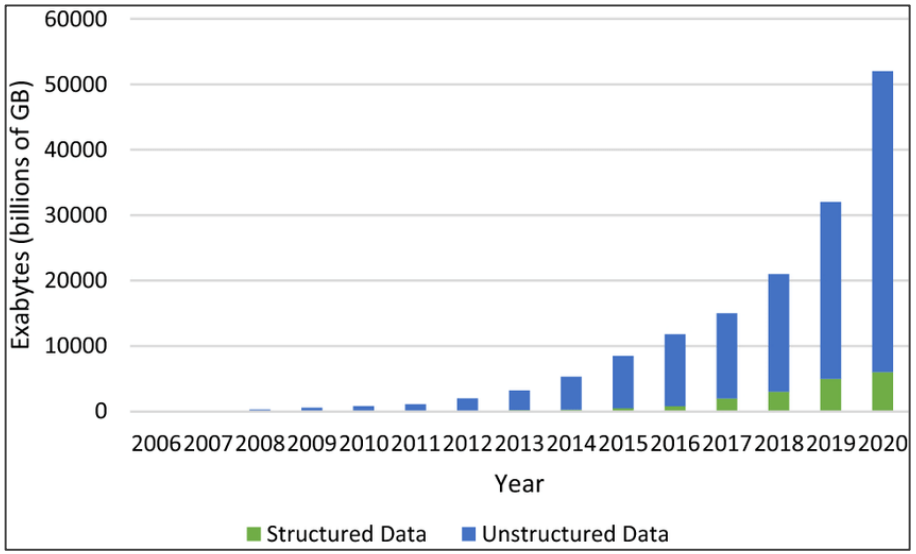
Structured Data



Semistructured Data

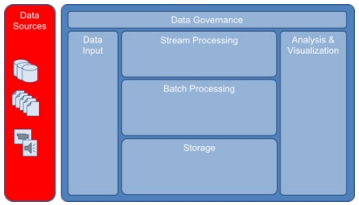
Unstructured Data

Metadata

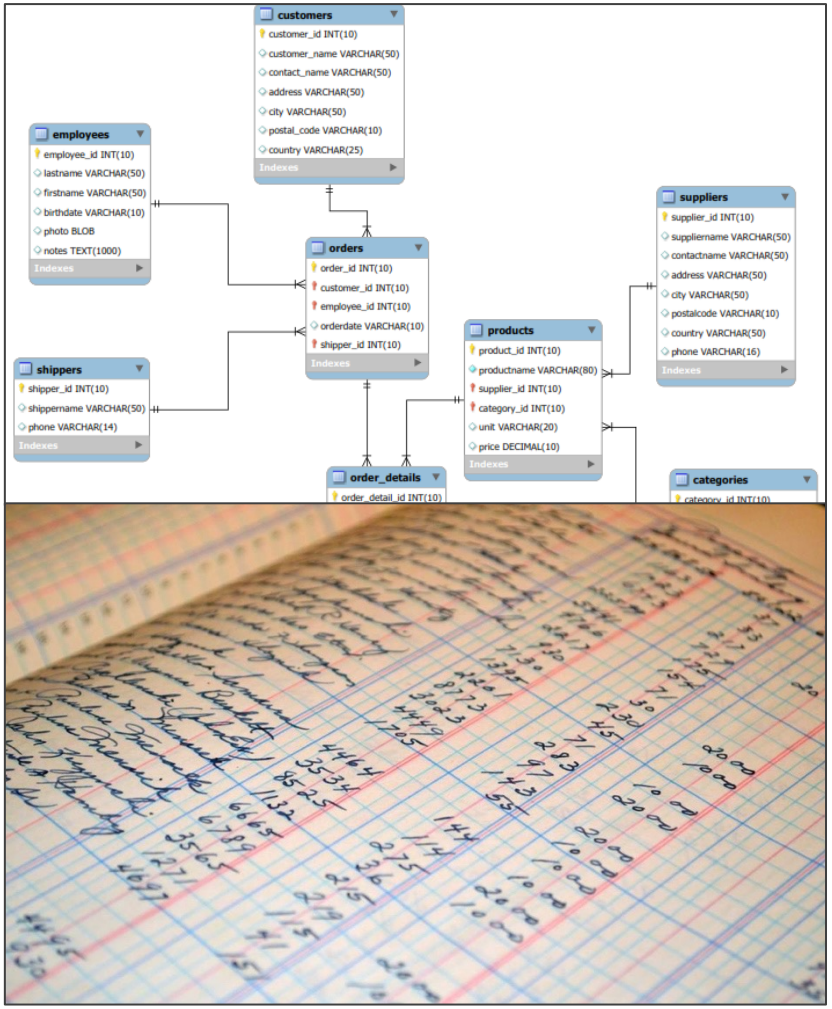


Rizzatti, L.: Digital data storage is undergoing mind-boggling growth. 14th September 2016

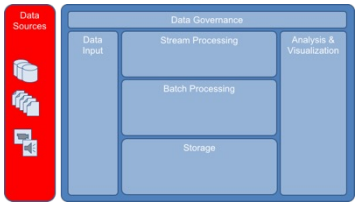
Typen von Daten – Structured Data



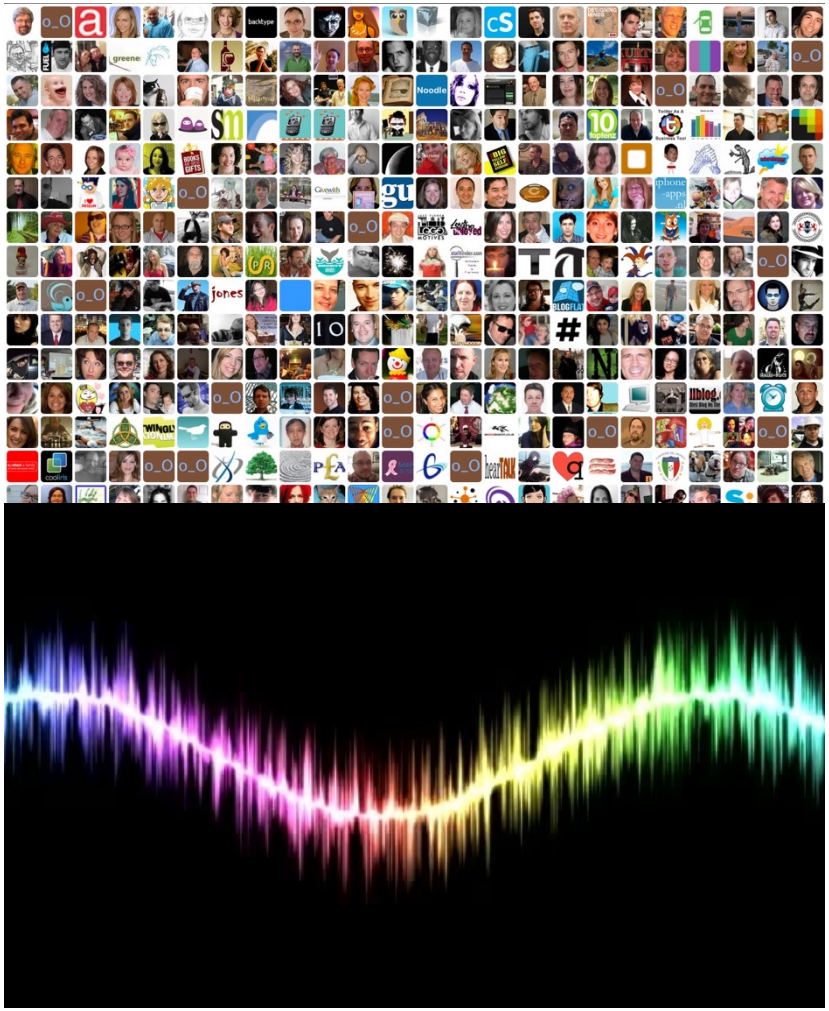
- Folgt einem Daten-Model oder Daten-Schema
- Oft in Tabellen abgelegt
- Bildet Beziehungen zwischen Entitäten ab
- Oft in relationalen Datenbanken abgelegt
- Beispiele:
  - Buchhaltungsdaten
  - Kundendaten
  - Adresslisten



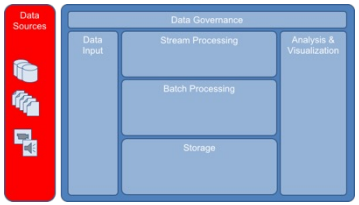
Typen von Daten – unstructured Data



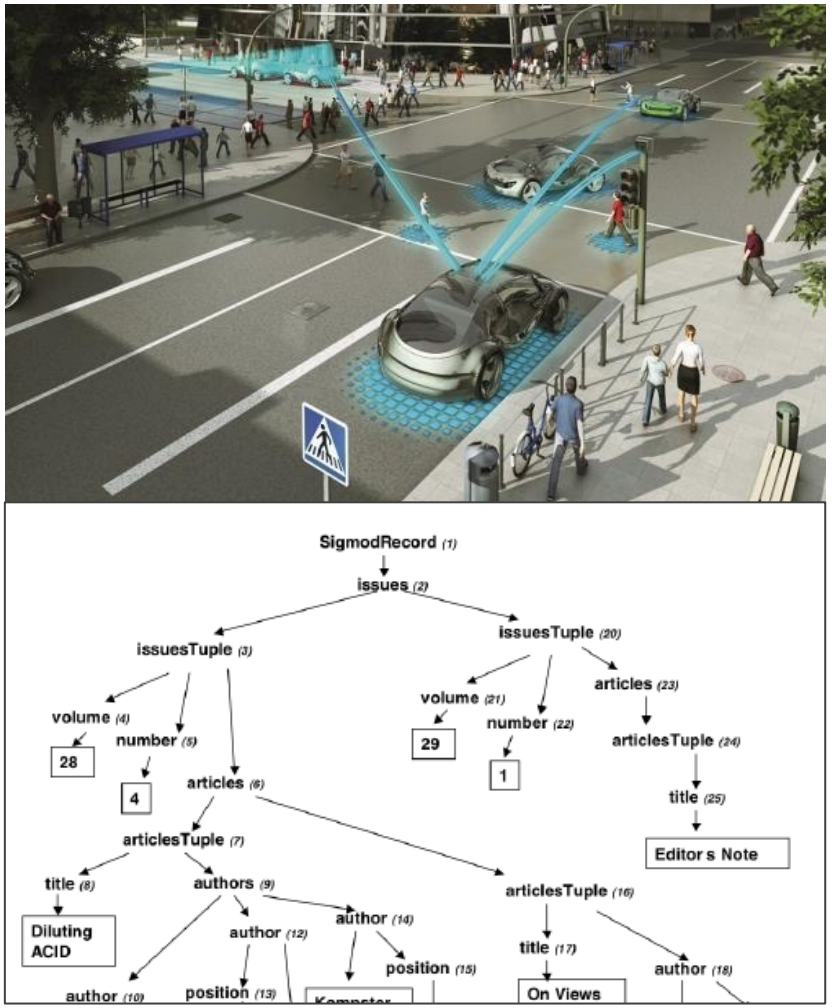
- Folgt keinem Daten-Model (außer Format-Spezifikationen)
- Kann Text oder Binärdaten sein, oft Medienfiles
- Oft sind spezielle Algorithmen nötig, um auf den Content zuzugreifen oder ihn zu analysieren
- Speicherung in FS, NoSQL Datenbanken, als BLOB inrelationalen Datenbanken
- Beispiele:
  - Videos, Audio-Files
  - Bilder



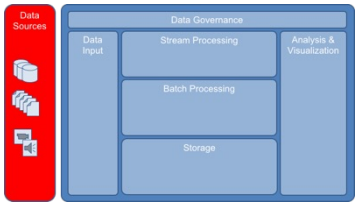
Typen von Daten – semistructured Data



- Hat eine gewisse Struktur aber ist nicht relational
- Oft hierarchisch oder Graph-basiert
- Durch die definierte Struktur leichter zu verarbeiten als unstrukturierte Daten
- Gängige Formate sind XML oder JSON
- Beispiele:
  - Electronic Date interchange files
  - RSS feeds
  - Sensor Data

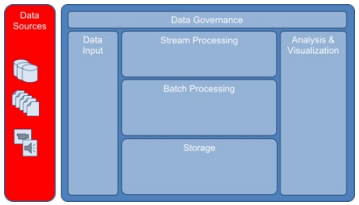


Typen von Daten – Metadata



- Daten, die Daten beschreiben
- i.d.R. maschinengeneriert
- Besonders wichtig auch für Big Data Analytics, da Metadaten Rückschlüsse auf die Herkunft der Daten, ihre Qualität etc. erlauben
- Beispiele:
  - GPS Daten in Fotodateien
  - Autor und Erzeugungsdatum in Dokumenten

Datenquellen



TikTok

amazon.com

Pay

PayPal

ebay

PAYBACK

Google Maps

opendata.swiss

LinkedIn

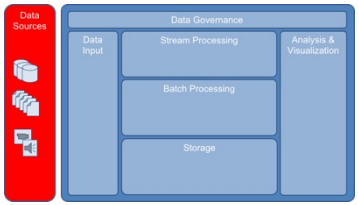
DWD

ORACLE

All of Us RESEARCH PROGRAM

SAP

Datenquellen



Was machen eigentlich Data Broker?

- Welche Daten haben sie?
- Woher bekommen sie die Daten?
- Was machen sie damit?
- Wieviel Umsatz machen sie?
- Gut oder schlecht?

EPSILON

CoreLogic®

ACXIOM

experian™