

# Data Challenges of AI

## Industry Experiences and Data Ecosystem Approach

16th Symposium and Summer School On Service-Oriented Computing, Crete Online, 08.07.22

Dr.-Ing. Christoph Gröger  
Robert Bosch GmbH  
[Christoph.Groeger@de.bosch.com](mailto:Christoph.Groeger@de.bosch.com)

# Overview

## Agenda

- Introduction: Industrial Analytics & AI
- Data Challenges
- Data Ecosystem for Industrial Enterprises

## Key References

Gröger, C. (2022): *Industrial analytics – An overview*. In: it – Information Technology, 64(1-2)

[https://www.christophgroeger.de/download/Groeger\\_Industrial\\_Analytics.pdf](https://www.christophgroeger.de/download/Groeger_Industrial_Analytics.pdf)

Gröger (2021): *There is No AI without Data*, Communications of the ACM, 64(11)

<https://doi.org/10.1145/3448247>

# Introduction

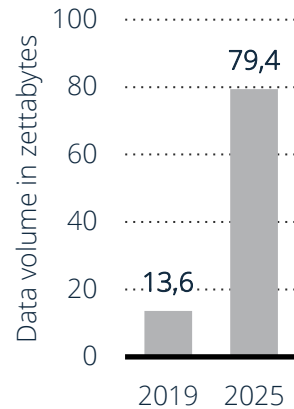
## Industrial Data

“1 TB of production data is created daily by the average factory, but less than 1% of that data is actually being analyzed by manufacturers.”

[Frost & Sullivan 2019]



Data volume of IoT connected devices worldwide 2019 and forecast 2025



[<https://www.statista.com/statistics/1017863/worldwide-iot-connected-devices-data-size>]

## Vehicle to Edge Data Transmission

- Edge data offloading scenarios for autonomous driving vehicles
  - High: 5.17 TB/hr/vehicle
  - Mid-Range: 0.945 TB/hr/vehicle
  - Low: 0.383 TB/hr/vehicle
- Includes: Total video, CAN/GPS/IMU and in-vehicle compute data
- Lower range is closer to connected vehicle levels

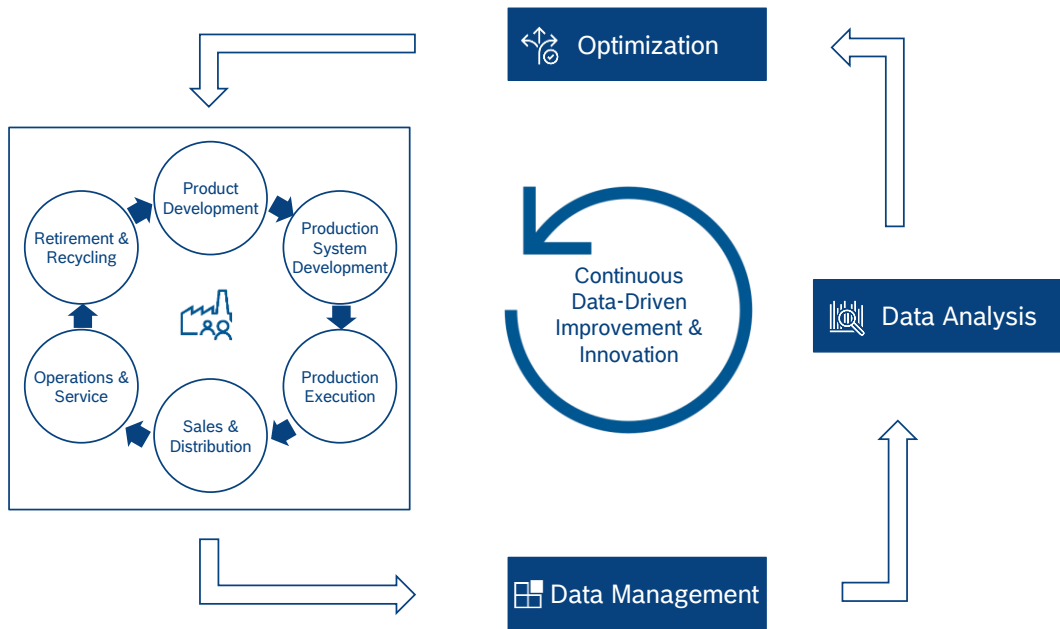


[<https://aecc.org/driving-data-to-deliver-addressing-the-connected-vehicle-data-challenge/>]

- **Data are diverse:** structured, semi-structured or unstructured, batch or stream, operational or process-related, ...
- **Data are valuable:** insights for process optimization, product enhancement, new services, ...

# Introduction

## Industrial Analytics & AI (I)



### Defining “Industrial Analytics”

- Data analytics for industrial value creation
- Industrial value creation as application domain of data analytics
- Sometimes also called “Industrial Intelligence” or “Industrie 4.0 Analytics”

**Note: industrial analytics refers to the entire industrial value chain, not only single phases such as production.**

[Gröger 2022]

# Introduction

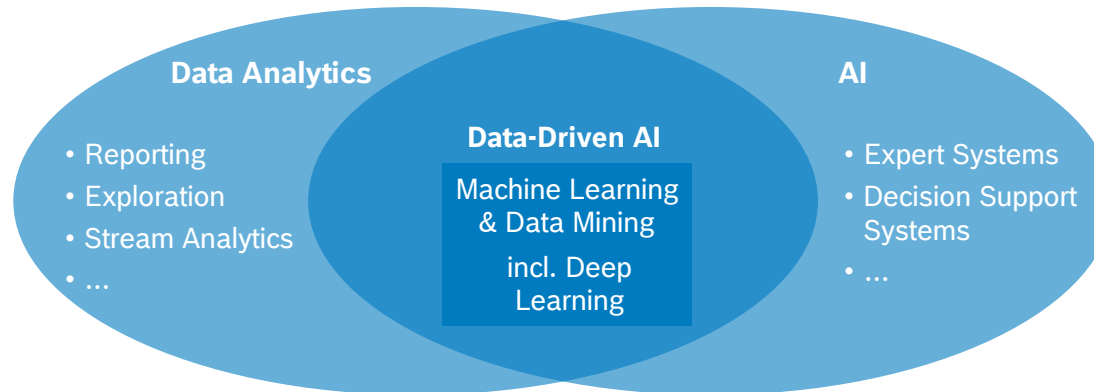
## Industrial Analytics & AI (II)

### Data Analytics

In-depth understanding and discovery of actionable insights from data, comprising descriptive (e.g. reporting), diagnostic, predictive (e.g. machine learning) and prescriptive analytics.

### Artificial Intelligence (AI)

Fuzzy term referring to the ability of a machine to perform cognitive functions. Differentiated are model-driven/deductive AI, e.g., expert systems, and data-driven/inductive AI, e.g. machine learning.



**Note: In the following, we focus on data analytics, especially data-driven AI.**

[Cao 2017, Everitt/Hutter 2018]

# Introduction

## Industrial Analytics Use Cases (Examples)

	Predictive Machine Maintenance	Predictive Process Quality	Engineering in the Loop
Goal	Optimizing machine maintenance	Reducing scrap in manufacturing	Improving product design based on real-world product usage
Object of Analysis	Machine	Process	Product
Product Lifecycle Phases	Production system development, production execution	Production system development, production execution	Product development, operations & service
Source Data	Maintenance data, machine data	Material data, quality data, process data, machine data	Engineering simulation data, master data, product usage data
Analytics Types	Predictive	Diagnostic, predictive	Descriptive
Techniques	Data mining & machine learning	Data mining & machine learning	Reporting & OLAP, exploration
Challenges in Practice	Data availability, data quality, imbalanced data	Data integration, imbalanced data	Data availability, data quality

# Data Challenges

## Current State: Insular Analytics & AI

- Many use cases, e.g., predictive maintenance, per se not new and known for years
- Challenge lies in concrete implementation in individual case
- In industry practice, each implementation is typically case-specific and tailored:
  - Same source data, e.g., ERP data, are extracted multiple times creating high load on business-critical source systems
  - Different data models are developed for the same conceptual data entities, such as 'machine'
  - Heterogeneous data models and different data lake storage technologies used lead to heterogeneous data pipelines for pivoting the same type of source data, e.g., MES tables with sensor data
  - Case- and user-specific analytics tools are used to generate insights



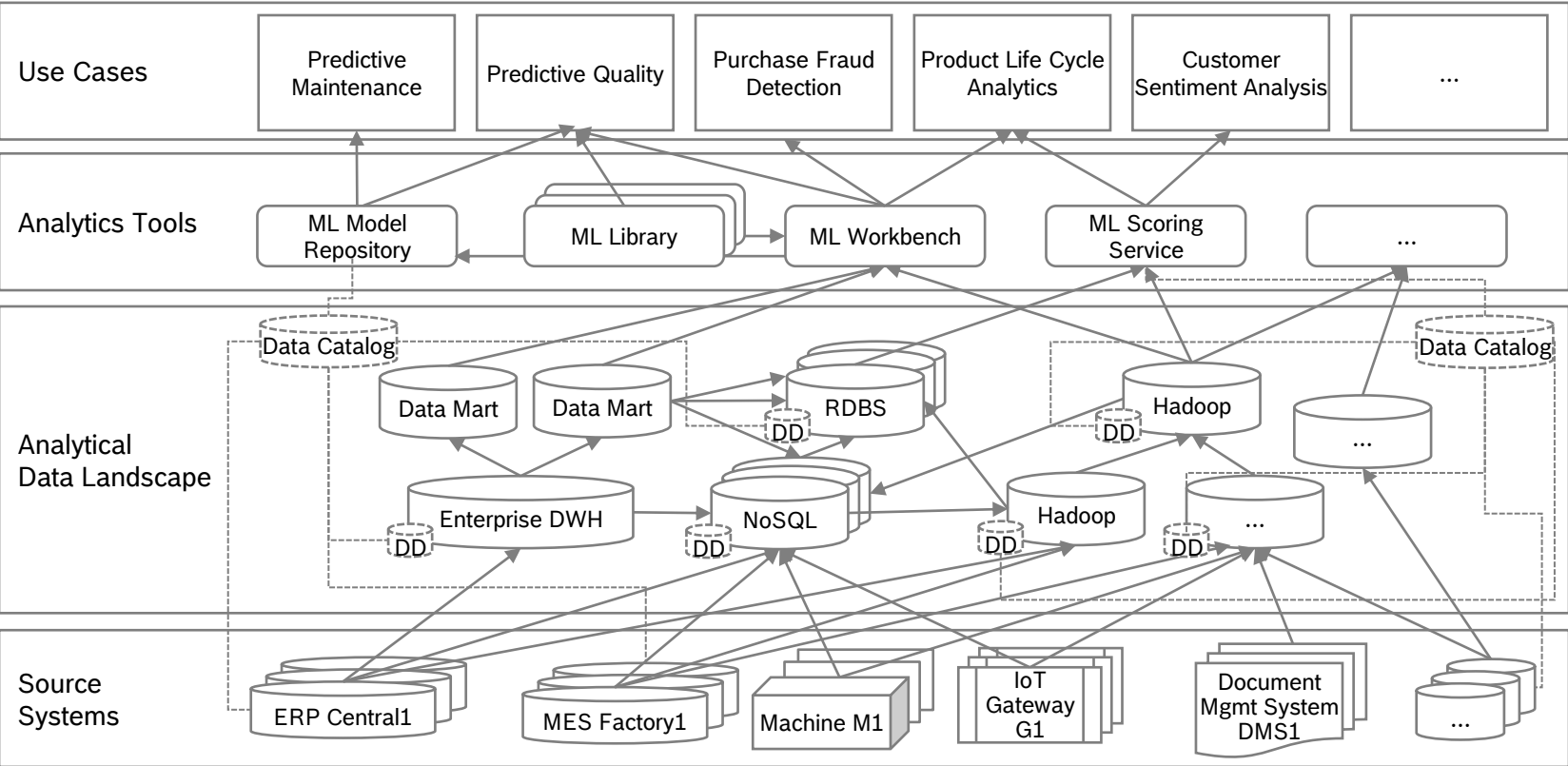
### Result: insular analytics & AI

In industry practice, analytics is done in isolated islands leading to a heterogeneous data landscape with data preparation accounting for 60-80% of use case implementation efforts.



# Data Challenges

## Current State: Heterogeneous Data Landscape



DD: Data Dictionary, DWH: Data Warehouse, ERP: Enterprise Ressource Planning, MES: Manufacturing Execution System, ML: Machine Learning, RDBS: Relational Database System

[Modified from Gröger 2021]



# Data Challenges

## Data Challenges in Practice

### Data challenges of industrial analytics go far beyond ensuring data quality

- Data management: processing, provisioning and controlling data throughout its lifecycle
- Data democratization: facilitating the use of data by everyone in an organization (taking into account data security and data privacy)
- Data governance: organizational structures to treat data as an enterprise asset, especially roles, decision rights and responsibilities

**Remark:** ensuring data quality is a general prerequisite (not detailed here)

#### Data Management Challenge

Comprehensive data management for AI in a heterogeneous enterprise data landscape:

- Data Modelling
- Metadata Management
- Data Architecture



#### Data Democratization Challenge

Making all kinds of data available for AI for all kinds of end users:

- Data Provisioning
- Data Engineering
- Data Discovery & Exploration



#### Data Governance Challenge

Defining roles, decision rights and responsibilities for the effective and compliant use of data for AI:

- Data Ownership
- Data Stewardship



#### Data Quality Challenge (not detailed here)

# Data Challenges

## Data Management Challenge

### Managing data in a heterogeneous and polyglot data landscape

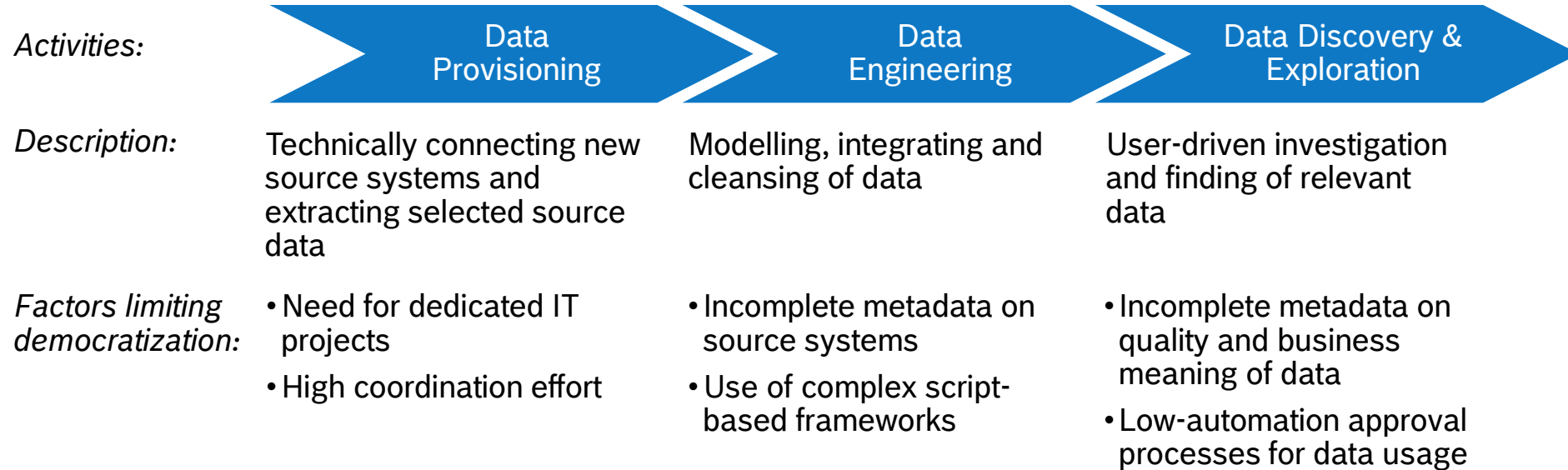
- Data modelling
  - No common data modelling approaches on how to model data across the data landscape complicating data integration and reuse of data and data pipelines
  - Different data modelling techniques, e.g., data vault or dimensional modelling, used for the same kinds of data
- Metadata management
  - No overall metadata management to maintain metadata across the data landscape resulting in complex data usage for analytics
  - Technical metadata, e.g., names of columns and attributes, are mostly stored in system-internal data dictionaries of individual storage systems and are not generally accessible, thus data lineage and impact analyses are hampered
  - Business metadata on the meaning of data, e.g., the meaning of KPIs, are often not systematically managed
- Data architecture
  - No overarching data architecture that structures the data landscape resulting in high development and maintenance costs
  - Enterprise data architecture to orchestrate the various isolated data lakes missing:  
No common zone model across all data lakes complicating data integration and exchange;  
integration of the existing enterprise data warehouse containing KPIs unclear
  - Platform data architecture to systematically design a data lake is lacking



# Data Challenges

## Data Democratization Challenge





**Making all kinds of data available for all kinds of end users**



# Data Challenges

## Data Governance Challenge (I)

### Basic data roles

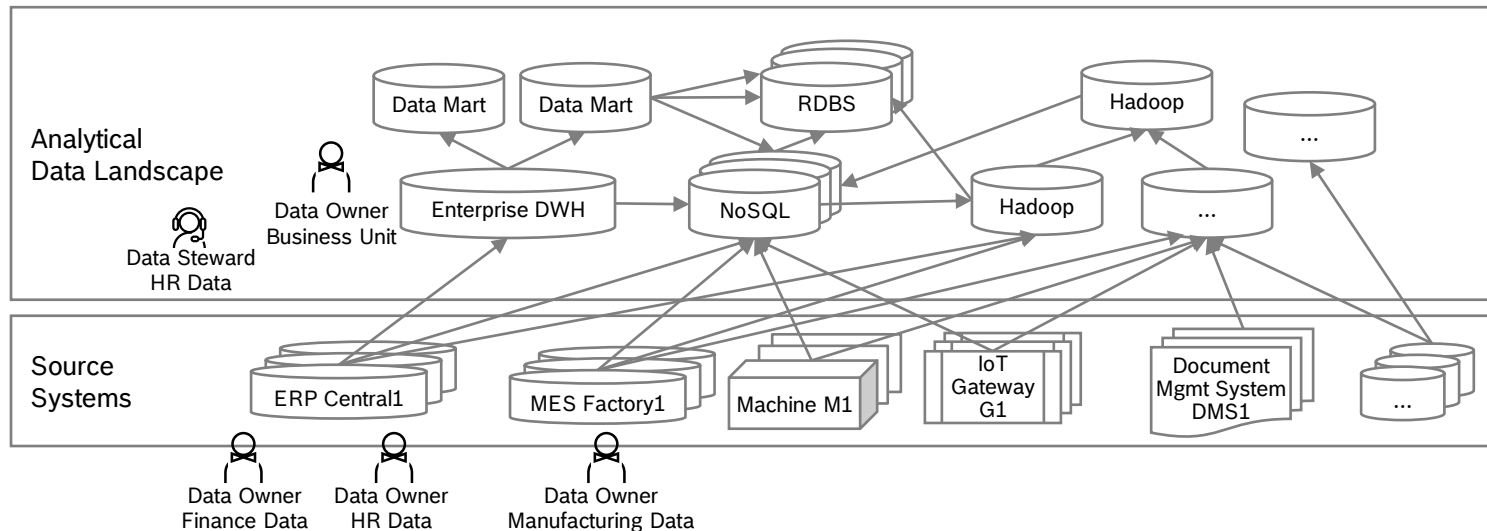
- Data owners 
  - Have overall legal and commercial responsibility for certain kinds of data, e.g. all data on a certain product
  - Assigned to the business, not the IT
  - Responsible for quality, security and compliance of data from a business point of view
- Data stewards 
  - Manage data on behalf of data owners
  - Responsible for realizing necessary policies and procedures from a business and from a technical point of view
- Data engineers 
  - Responsible for developing data pipelines to provide the data basis for further analyses by integrating and cleansing of data
- Data scientists 
  - Focus on the actual analysis of data by feature engineering and applying various data analytics techniques, e.g. different data mining algorithms, to derive insights from data

# Data Challenges

## Data Governance Challenge (II)

### Data governance challenge: defining roles, rights and responsibilities for the effective and compliant use of data

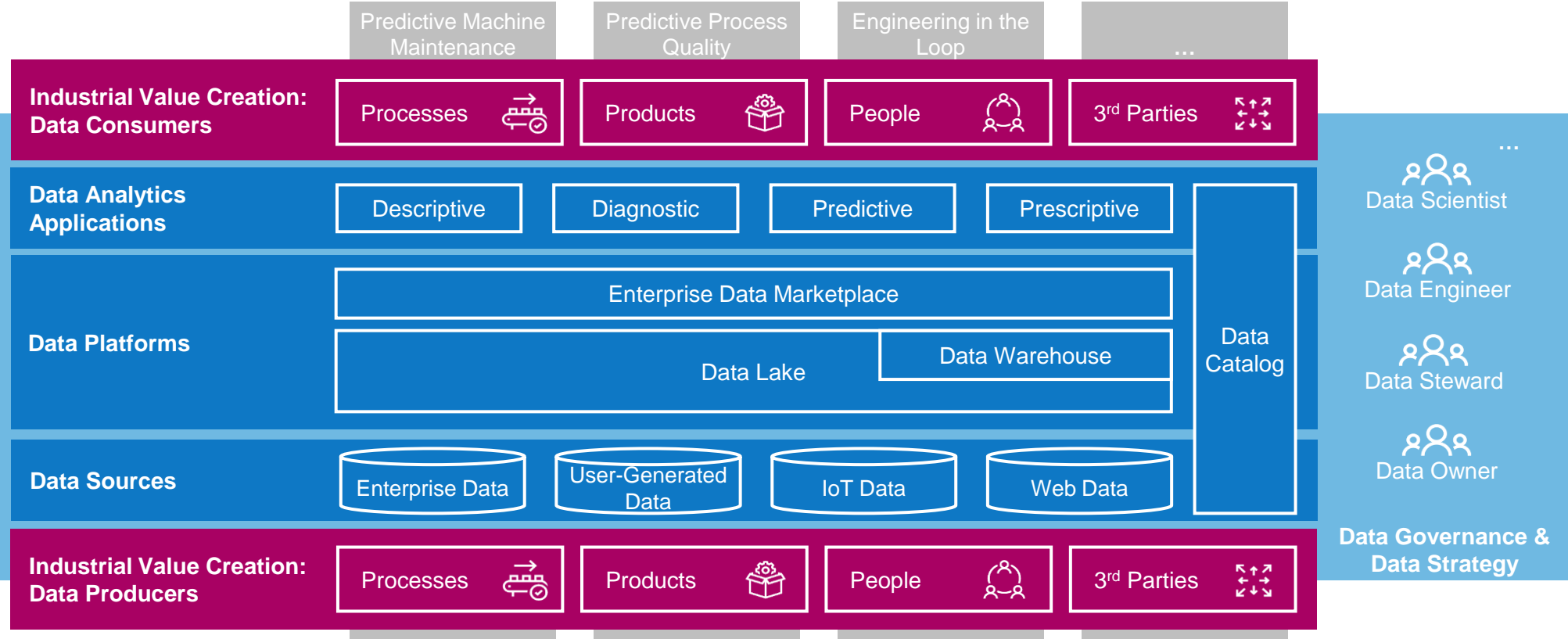
- No uniform data ownership organization: heterogeneous and overlapping data ownership structures, e.g., when data ownership is organized by business function in source systems and by business unit in the data lake
- No overall data stewardship organization to establish common policies and standards for data: data stewardship established mainly for master data, further data categories missing
- Analytics roles (data scientist, data engineer) overlap with classical data governance roles



# Data Ecosystem for Industrial Enterprises

## Overview

Socio-technical, loosely coupled, self-organizing system for the sharing of data



Use Cases

[Gröger 2022]

# References

- Abraham, Schneider, Brocke (2019): *Data governance*, International Journal of Information Management 49
- Cao (2017): *Data Science: A Comprehensive Overview*, ACM Computing Surveys 50(3)
- Everitt, Hutter (2018): *Universal Artificial Intelligence*, in: Foundations of Trusted Autonomy, Springer
- Frost & Sullivan (2019): *Industry 4.0, the Fourth Revolution*, Whitepaper
- Gröger (2021): *There is No AI Without Data*, Communications of the ACM, 64(11)
- Gröger (2022): *Industrial analytics – An overview, it – Information Technology*, 64(1-2)
- Lyon, Mattern (2016): *Education for real-world data science roles (part 2)*, International Journal of Digital Curation 11(2)