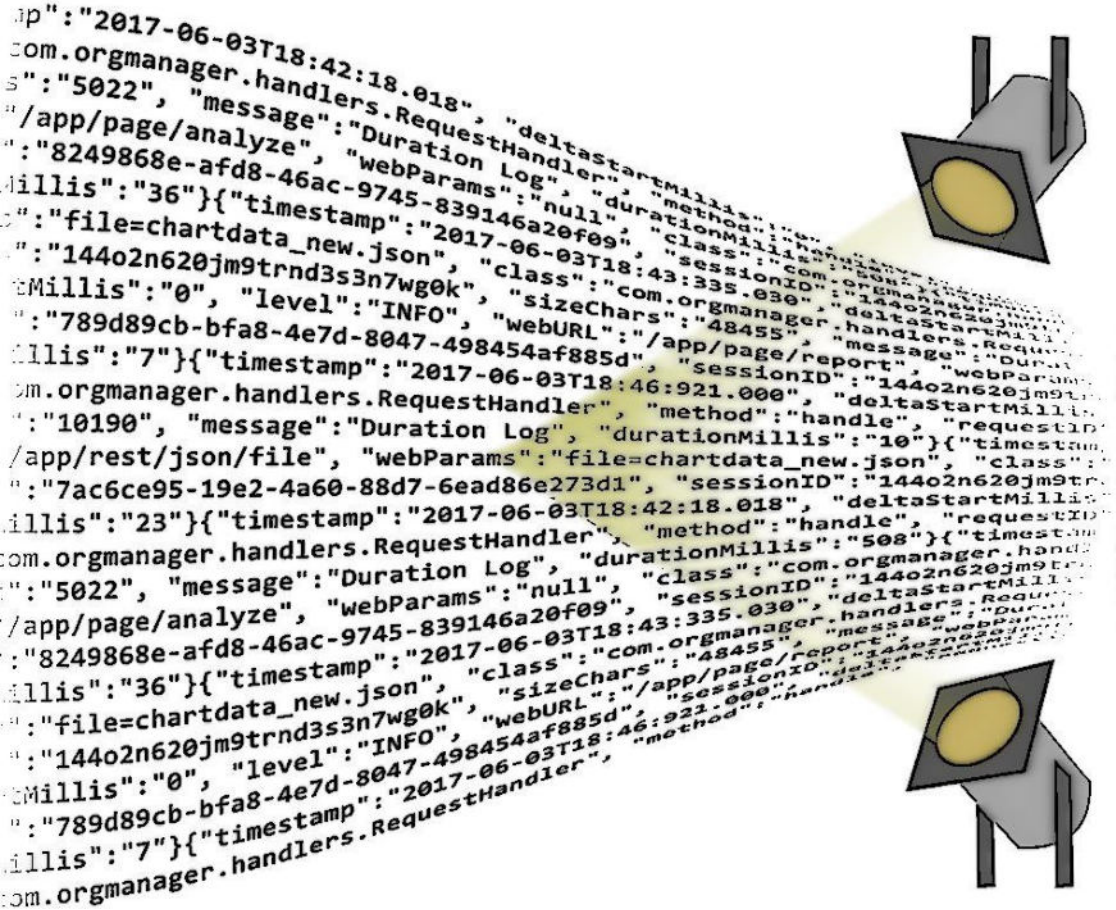
universität innsbruck

Department of Computer Science

Austrian Testing Board
member of ISTQB®

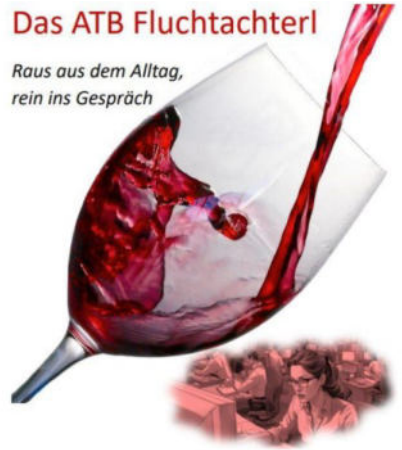


Risikobasiertes Testen von Daten

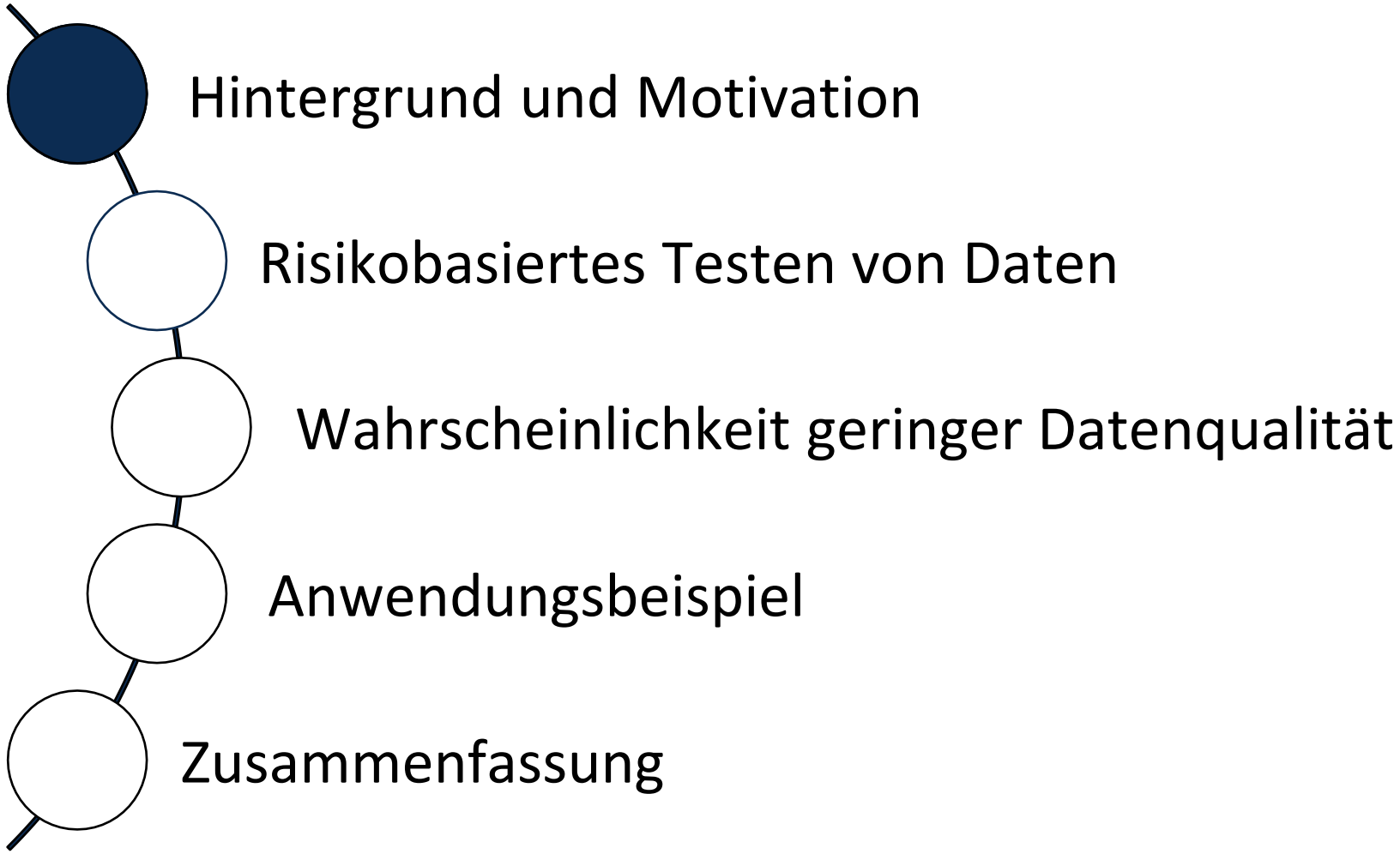
Ing. Harald Foidl, PhD

ATB Fluchtachterl

18.06.2025

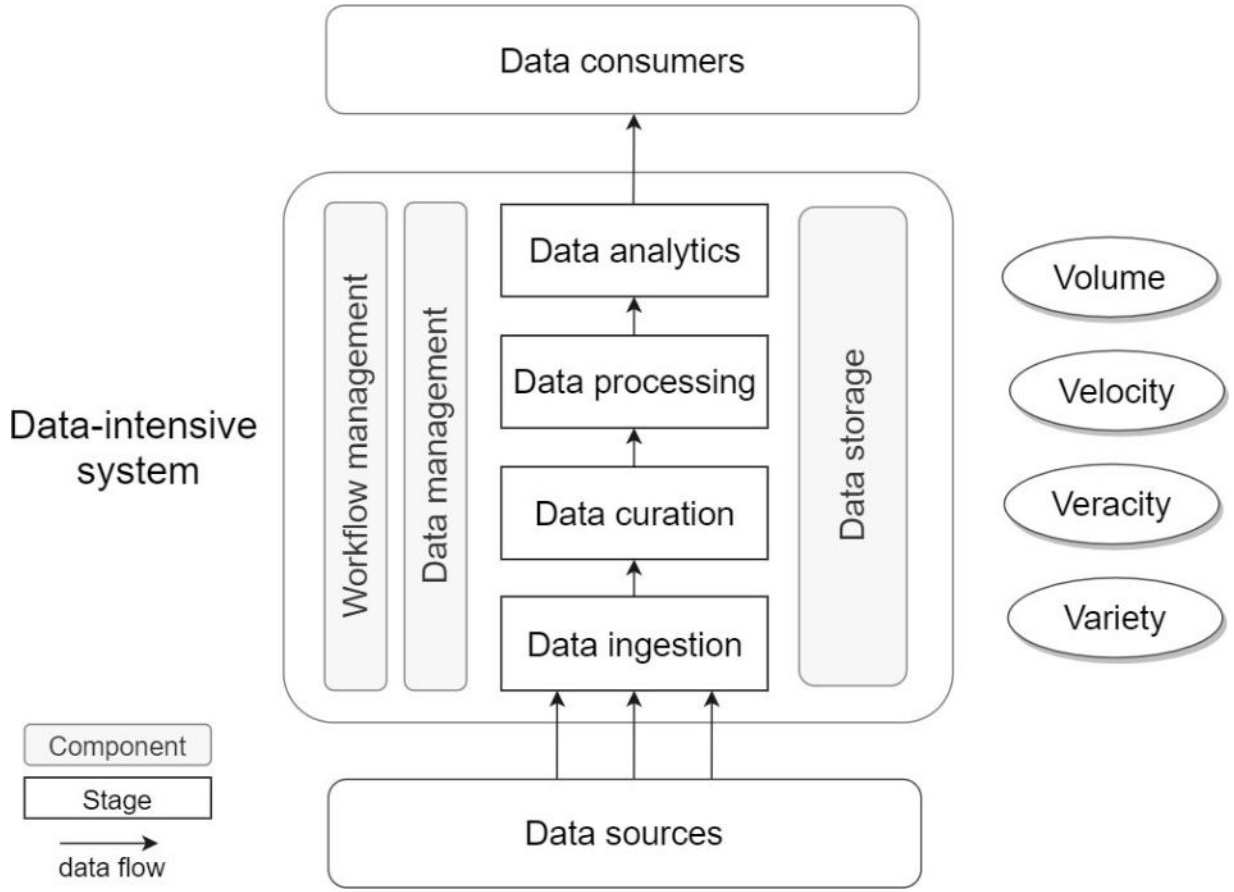


Agenda


Hintergrund und Motivation
Risikobasiertes Testen von Daten
Wahrscheinlichkeit geringer Datenqualität
Anwendungsbeispiel
Zusammenfassung

Datenintensive Systeme (DIS)

- große Datenmengen
- Datenverarbeitung
- Lernen von Daten
- Intelligente Entscheidungen




© Siemens

Zuverlässigkeit von Datenintensiven Systemen

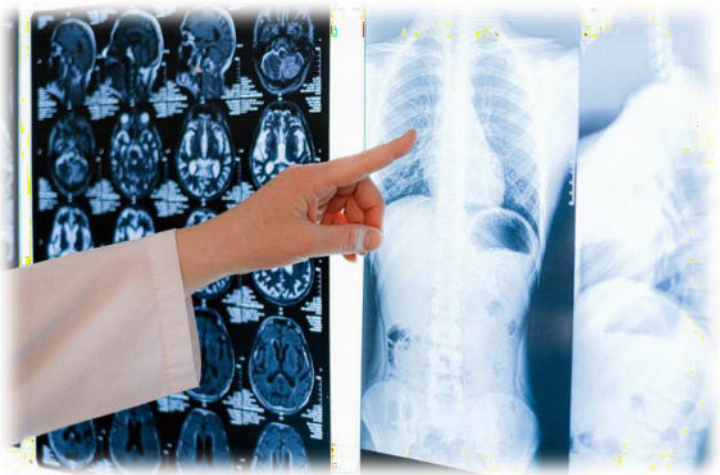
Fehlentscheidungen solcher Systeme können schwerwiegende Folgen haben!

How a Self-Driving Uber Killed a Pedestrian in Arizona

By TROY GRIGGS and DAISUKE WAKABAYASHI UPDATED MARCH 21, 2018

Report: IBM Watson delivered 'unsafe and inaccurate' cancer recommendations

JULY 25, 2018 BY FINK DENSFORD — LEAVE A COMMENT



Datenqualität entscheidend

Lösung: Testen von Daten (Datenvalidierung)

- Kontinuierliche Prüfung/Überwachung von Daten zur Qualitätsbewertung
- Definierte Schemata, Wertebereiche und weitere Kriterien (z. B. Eindeutigkeit, Datentypen)
- Techniken, z.B.:

Hypothesentests

Korrelationsanalysen

Aggregationen

Lageparameter (z.B. Mittelwert)

Jedoch …

Datenvalidierung beeinflusst die Systemperformance

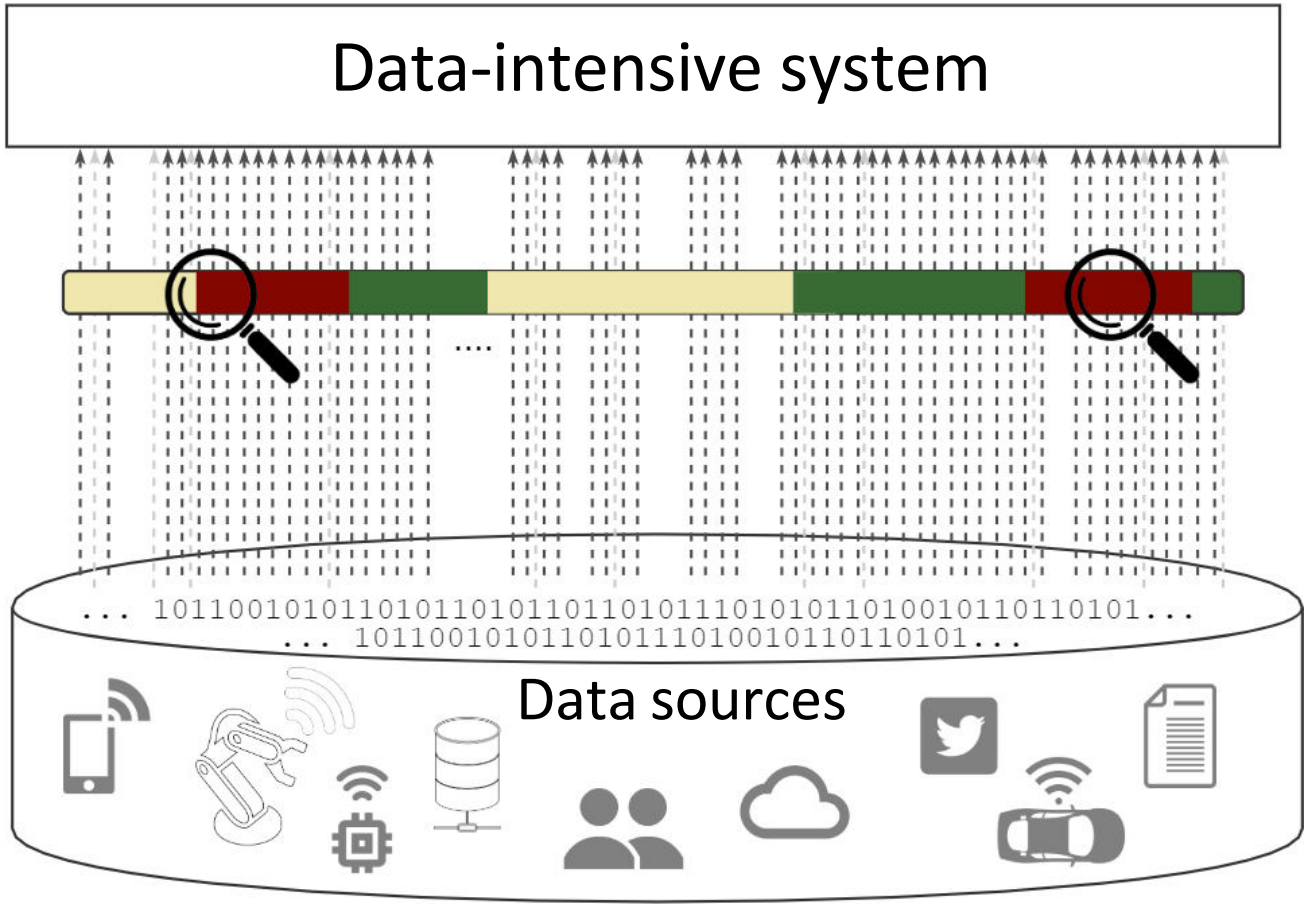
Manuell, zeitaufwendig, erfordert Expertenwissen

Problem der Validierungstiefe/Strenge der Validierung

Vollständige Datenvalidierung ist nicht möglich

Problemstellung

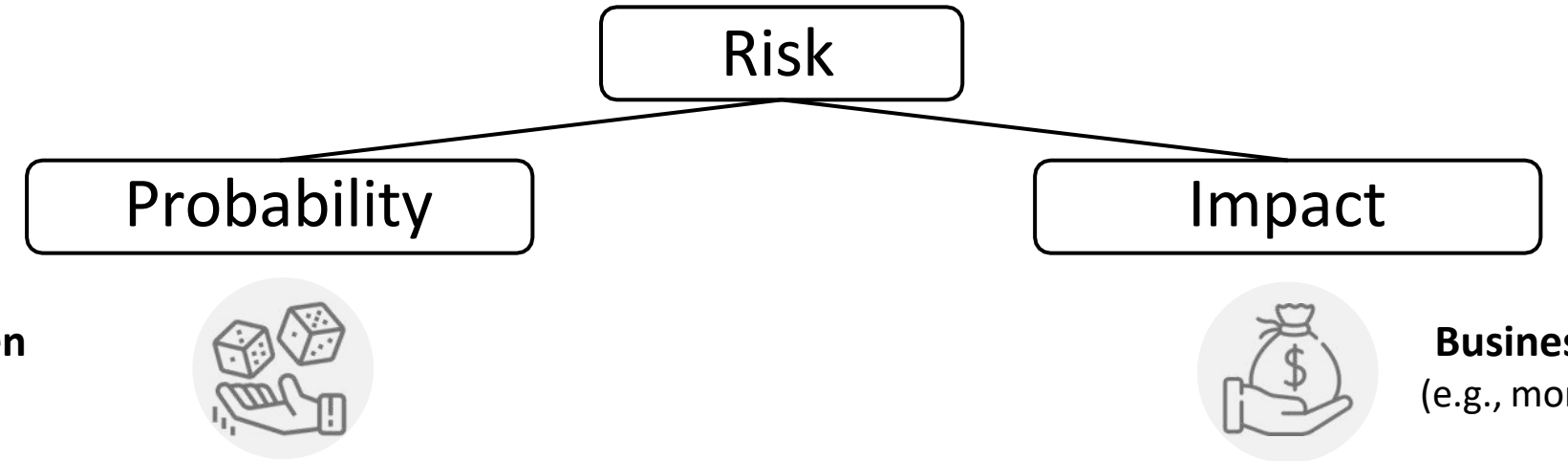
SELEKTIVES
Testen von Daten



Welche Daten sollen validiert werden – und wie *gründlich*?

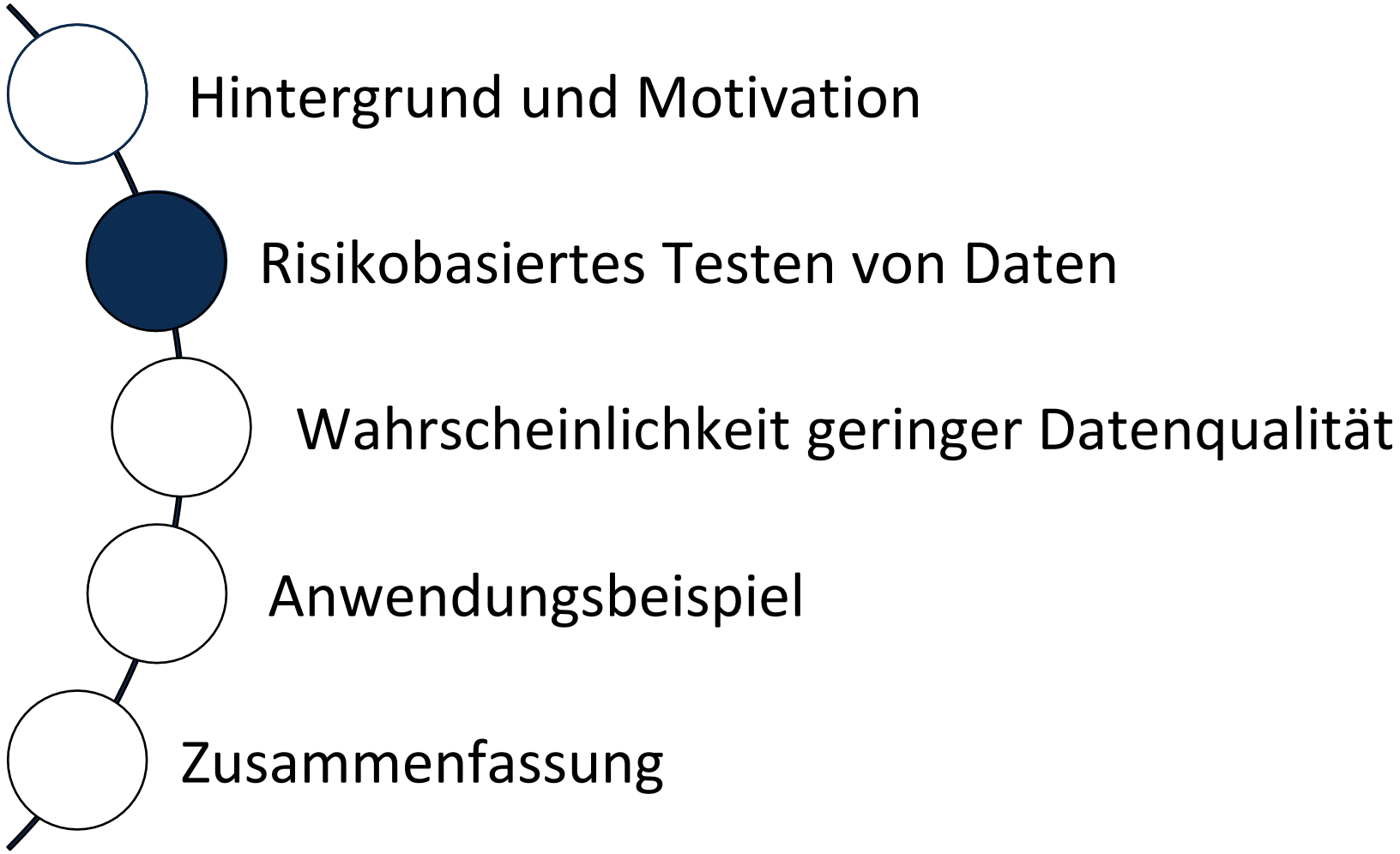
Ansatz

- Softwaretesten: risikobasierte Strategie für vergleichbare Entscheidungsprobleme
- Risikobasiertes Testen: Risiken eines Softwareprodukts steuern den Testprozess



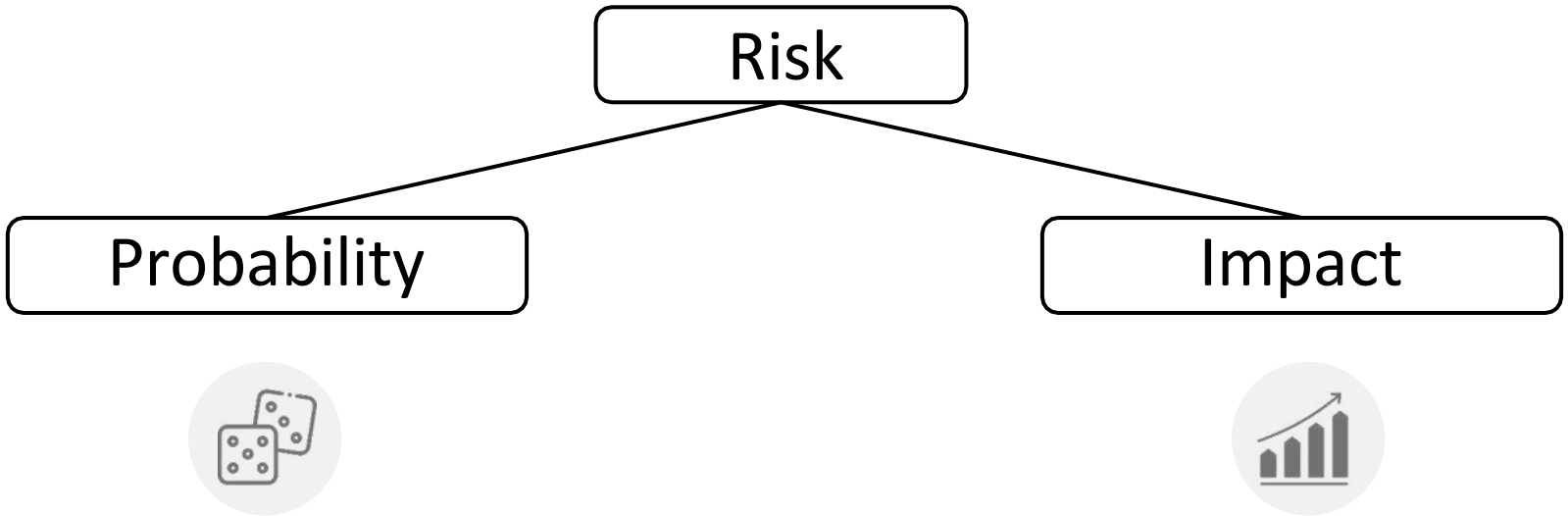
Nutzung des Konzepts des risikobasierten Softwaretestens

Agenda


Hintergrund und Motivation
Risikobasiertes Testen von Daten
Wahrscheinlichkeit geringer Datenqualität
Anwendungsbeispiel
Zusammenfassung

Idee des risikobasierten Testens von Daten

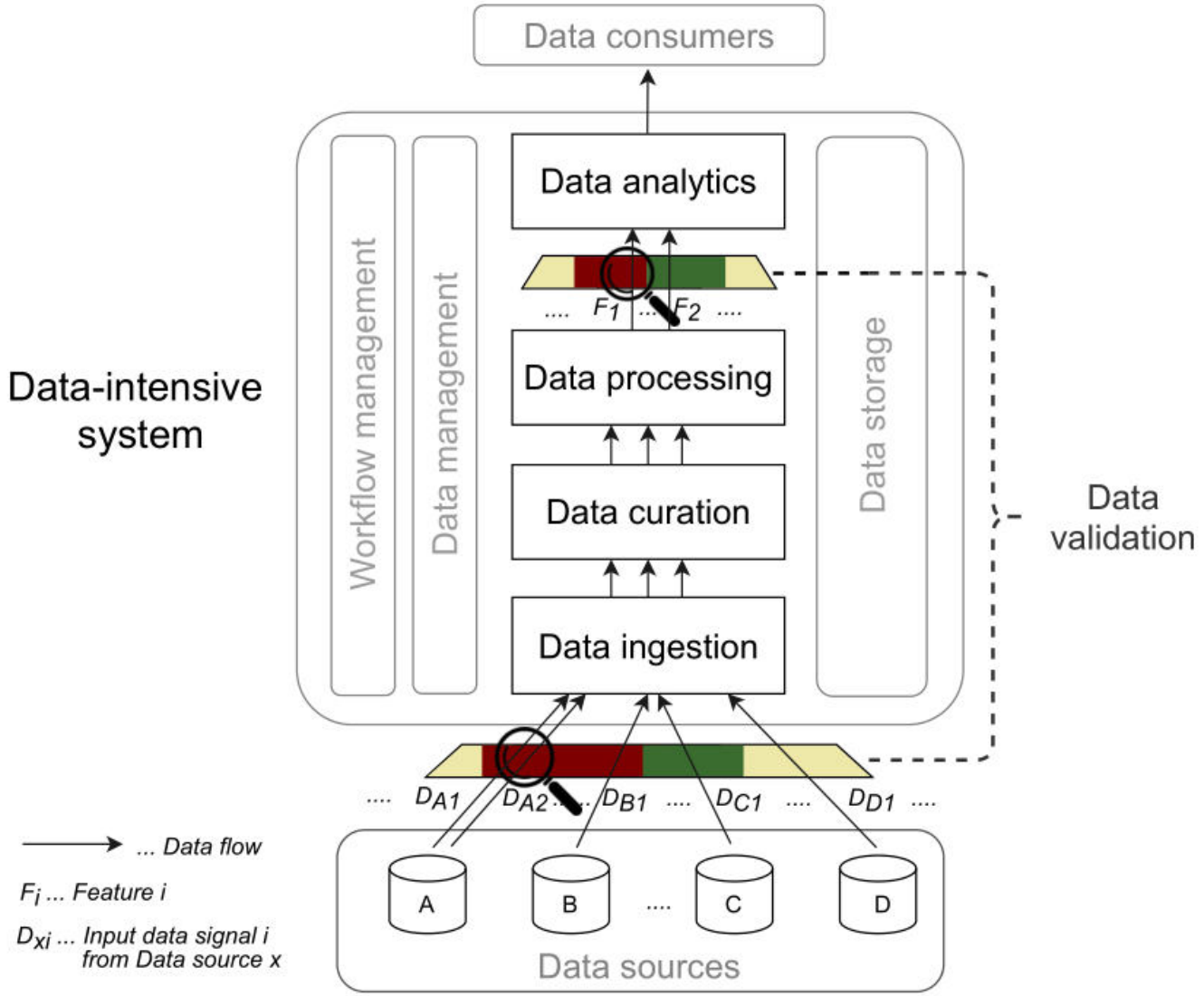
Bestimme **Risiko von geringer Datenqualität,**
um Testen zu steuern



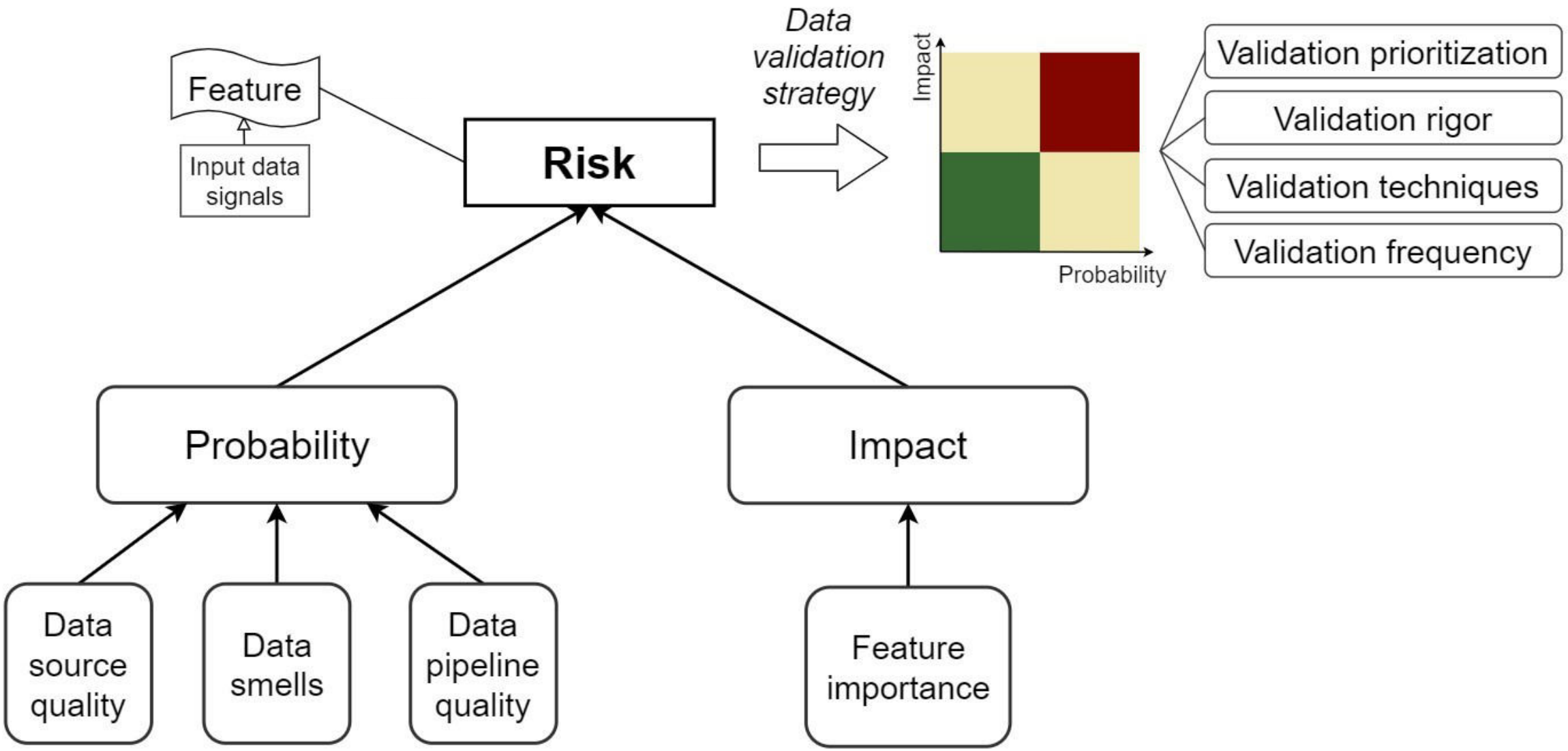
Wahrscheinlichkeit
Daten von geringer Qualität

Auswirkung
geringer Qualität der Daten auf DIS

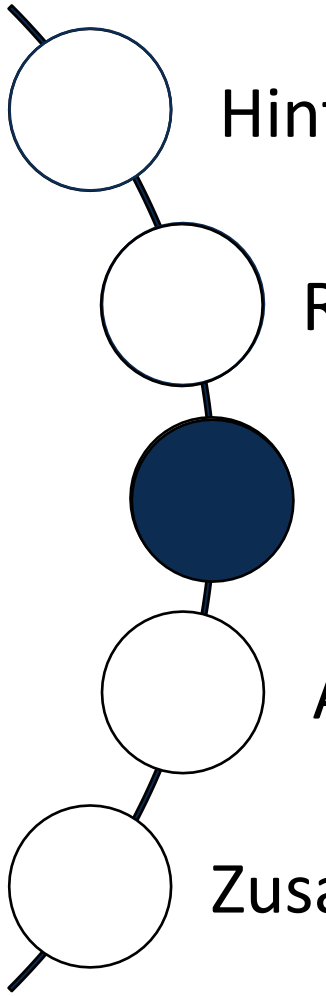
Risikobasiertes Testen von Daten



Risikobasiertes Testen von Daten - Framework



Agenda



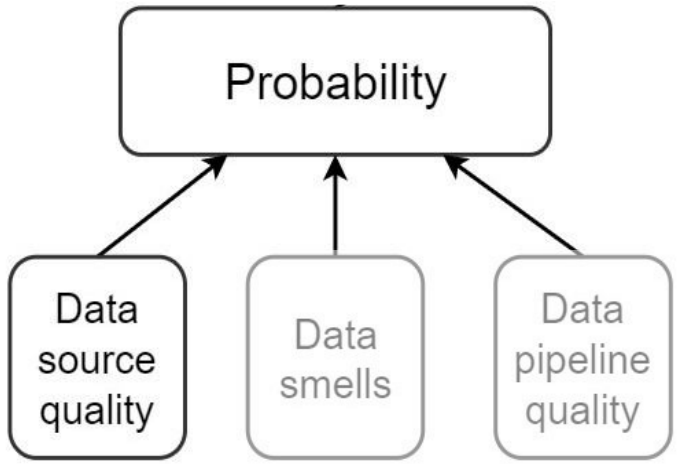
Hintergrund und Motivation

Risikobasiertes Testen von Daten

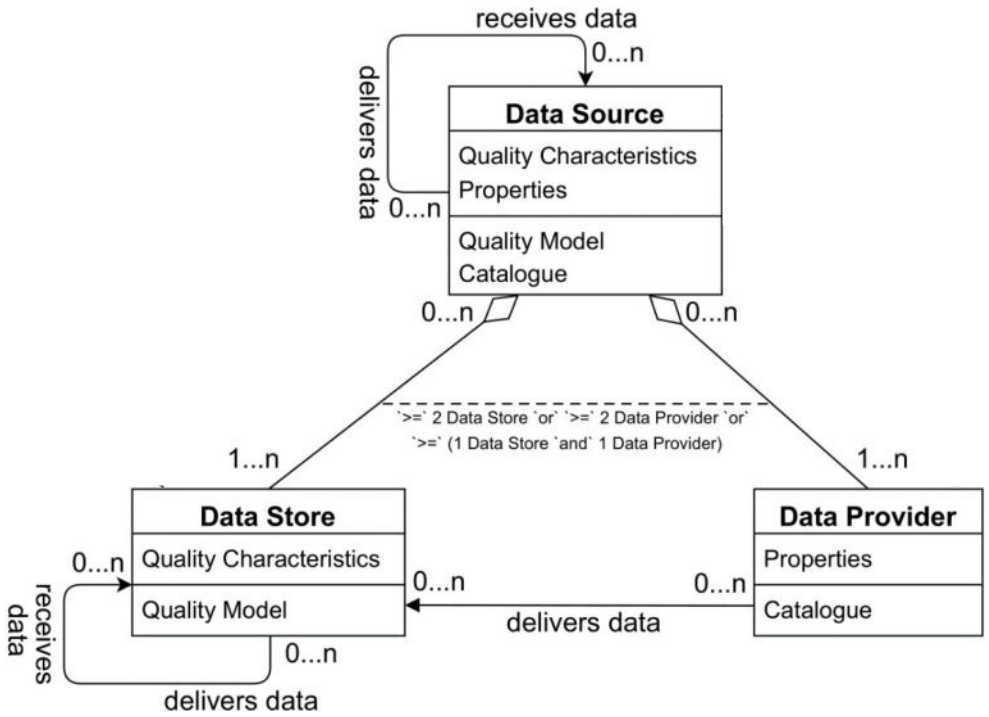
Wahrscheinlichkeit geringer Datenqualität

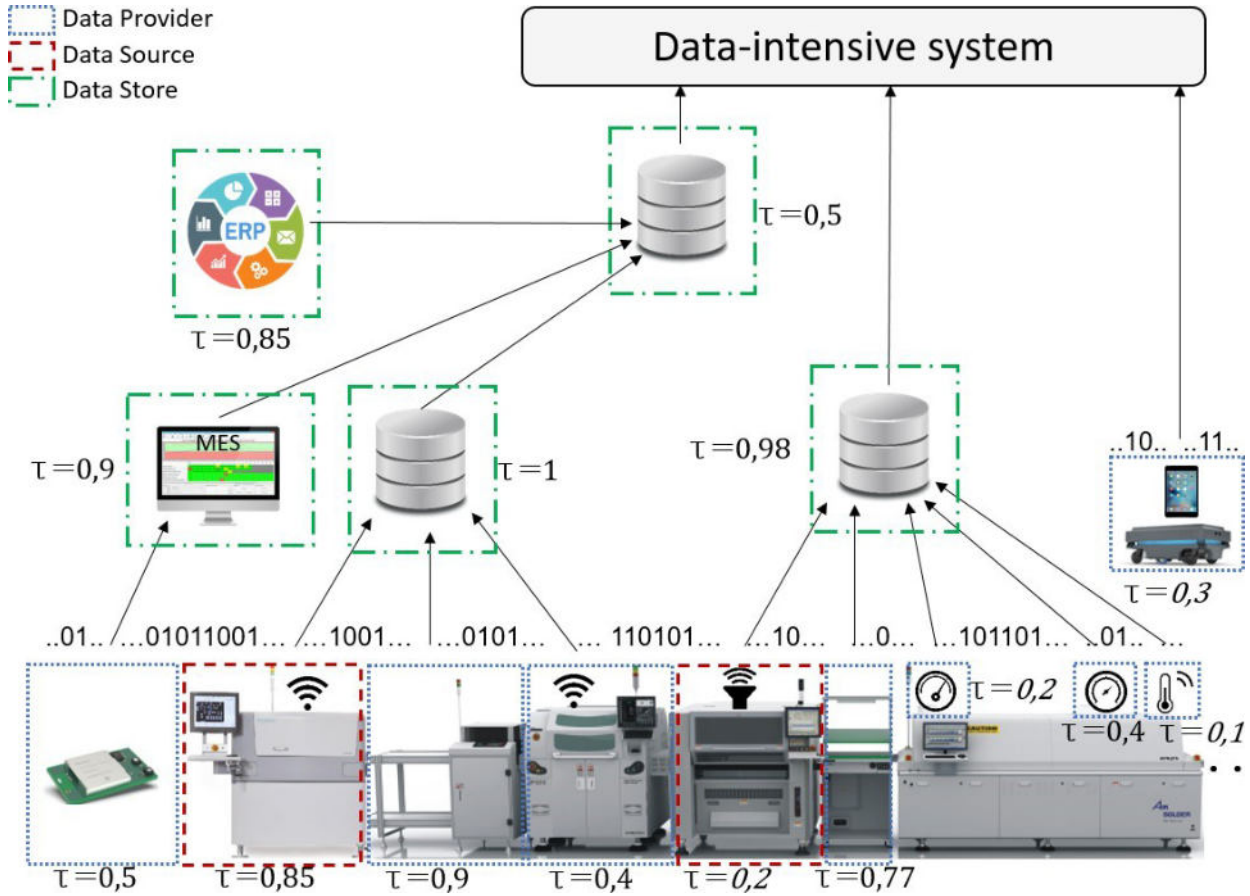
Anwendungsbeispiel

Zusammenfassung



Qualität von Datenquellen





Qualitätsmodel für Datenspeicher

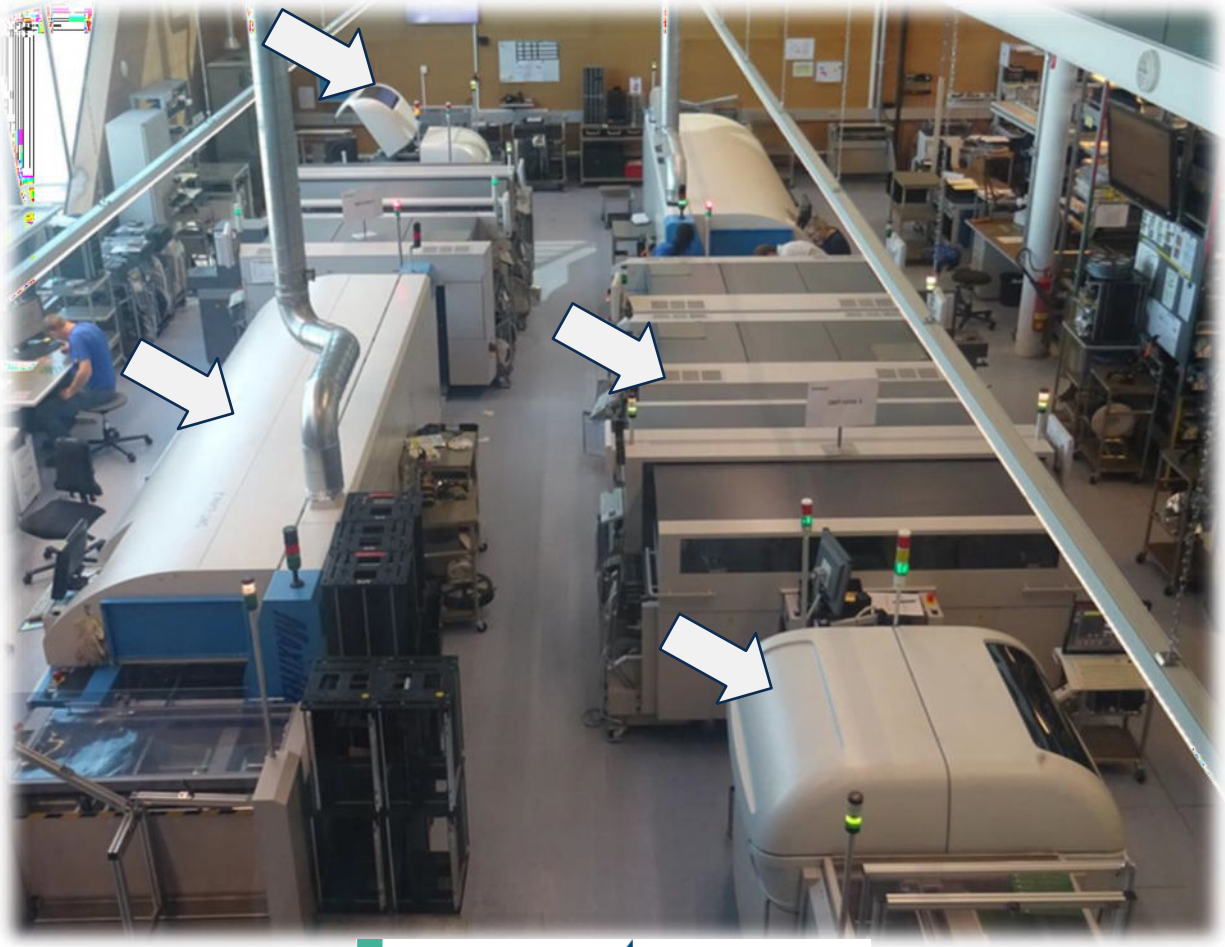
Data Quality Category	Data Quality Characteristic	Description	Properties and Sub-characteristics
Representational Data Store Quality	***Representational Adequacy***	Degree to which a data store presents data in a *concise and organised way*.	**Schema Minimality**
			Schema Normalisation
			Schema Pertinence
	Representational Consistency	Degree to which a data store presents data *always in the same format and compatible with previous data.*	**Data Format Variety**
			Data Type Variety
			Schema Change Proneness
	Understandability	Degree to which *users can understand the data* provided by a data store.	**Data Format Complexity**
			Documentation Degree
			Metadata Quality
			Schema Readability
Dynamical Data Store Quality	***Accessibility***	Degree to which data are *easily and quickly retrievable.*	**Access Maturity**
			Operability
			Retrievability
	Availability	Degree to *which data are available* from a data store.	**Durability**
			Fault Tolerance
			Recoverability
			Scalability
			Uptime
	Security	Degree to which *access to data for unauthorised persons is restricted* by a data store.	**Authentification/Encryption**
			Authorisation Policy
	Timeliness	Degree to which a data store *provides up-to-date data in a timely manner.*	**Refresh Rate**
			Response Time
Statical Data Store Quality	***Completeness***	Degree to which a data store is able to represent *every meaningful state of the real world.*	**Schema Completeness**
			Schema Correctness
	Contactability	Degree to which a data store provides *contact information for further inquiries.*	**Support Degree**
	Trustworthiness	Degree to which a *data store can be trusted*.	**Complexity**
			Data Governance
			Maturity
			Verifiability

Eigenschaftskatalog von Datenanbietern

Category	Property	Option
General	Age	old
		middle
		new
	Provider type	persons involved
		only sensor/device/machine
Location	Environment	harsh, dynamically changing
		moderate, seldom changing
		mild, never changing
	Mobility	always
		sometimes
		never
Energy	Power source	battery
		hardwired
	Usage	low
		medium
		high
Connectivity	Mechanism	wireless
		physical
		wired

Category	Property	Option
Hardware	Quality	low
		medium
		high
	Maintenance/calibration	never
		sporadic
		regulary
	Ease of replacement/repair	hard
		medium
		easy
	Constraints	severe limitations
		moderate limitations
		no limitations
Data pre-processing		high
		medium
		low/no

Fallstudie - Elektronikfertigung



SMT assembly machine

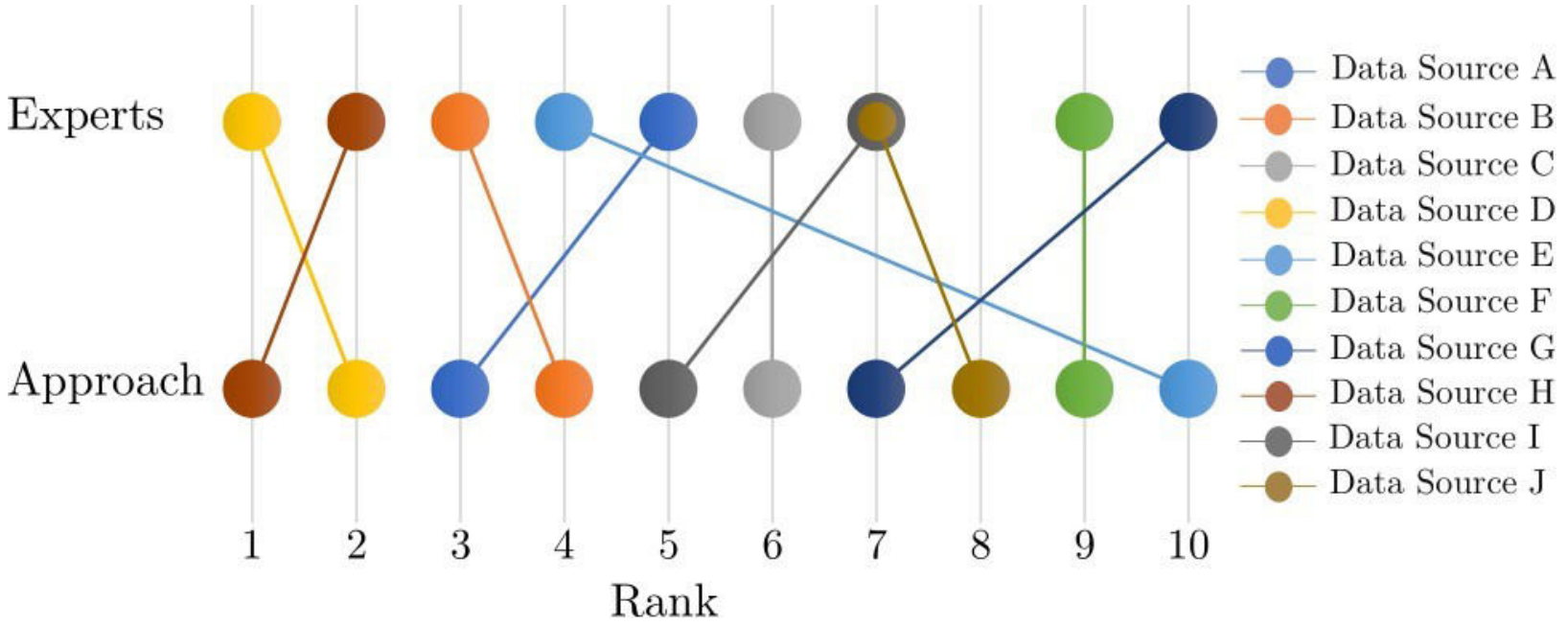
wearable barcode scanner

nitrogen sensor

kontron
The Power of IoT

universität innsbruck

Fallstudie - Ergebnis

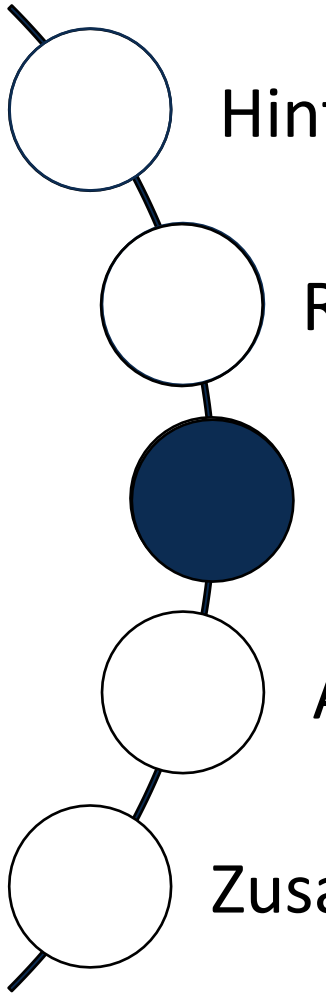


Moderate positive Korrelation zwischen Experten Bewertung und entwickelten Ansatz.

(Ranking von Datenquellen)

Spearman's Rho: 0.64 (p-value = 0.04)
Inter-rater Agreement of Experts: 0.79 (Krippendorff's Alpha)

Agenda



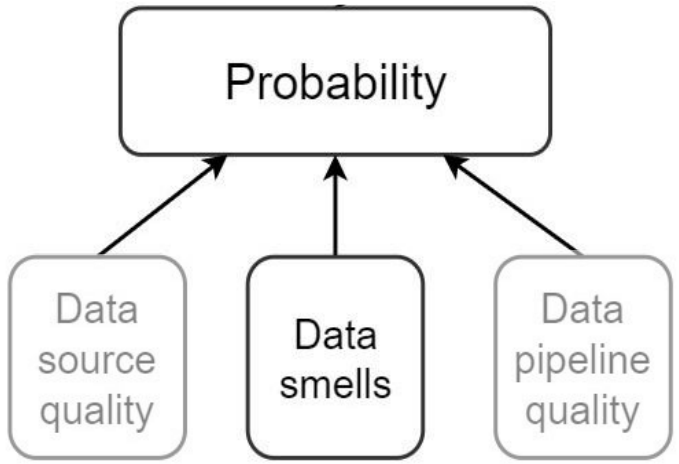
Hintergrund und Motivation

Risikobasiertes Testen von Daten

Wahrscheinlichkeit geringer Datenqualität

Anwendungsbeispiel

Zusammenfassung



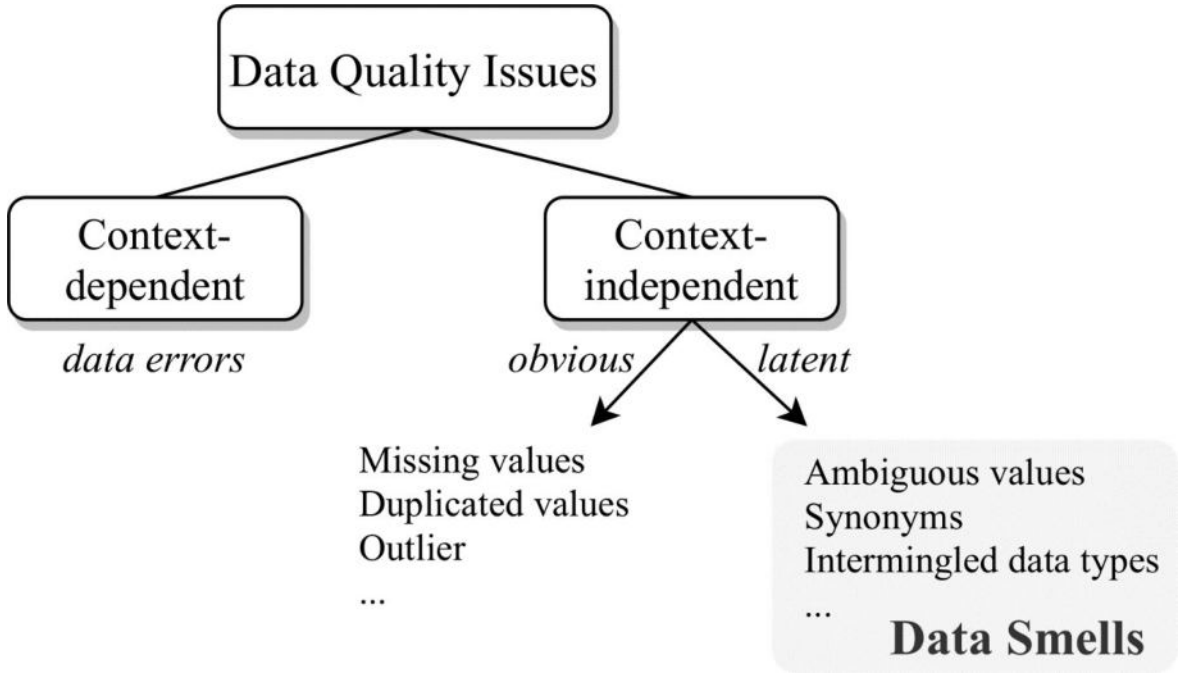
Data Smells

kontextunabhängige, datenwertbasierte Indikatoren für latente Datenqualitätsprobleme, die durch schlechte Praktiken verursacht werden und zu zukünftigen Problemen führen können.

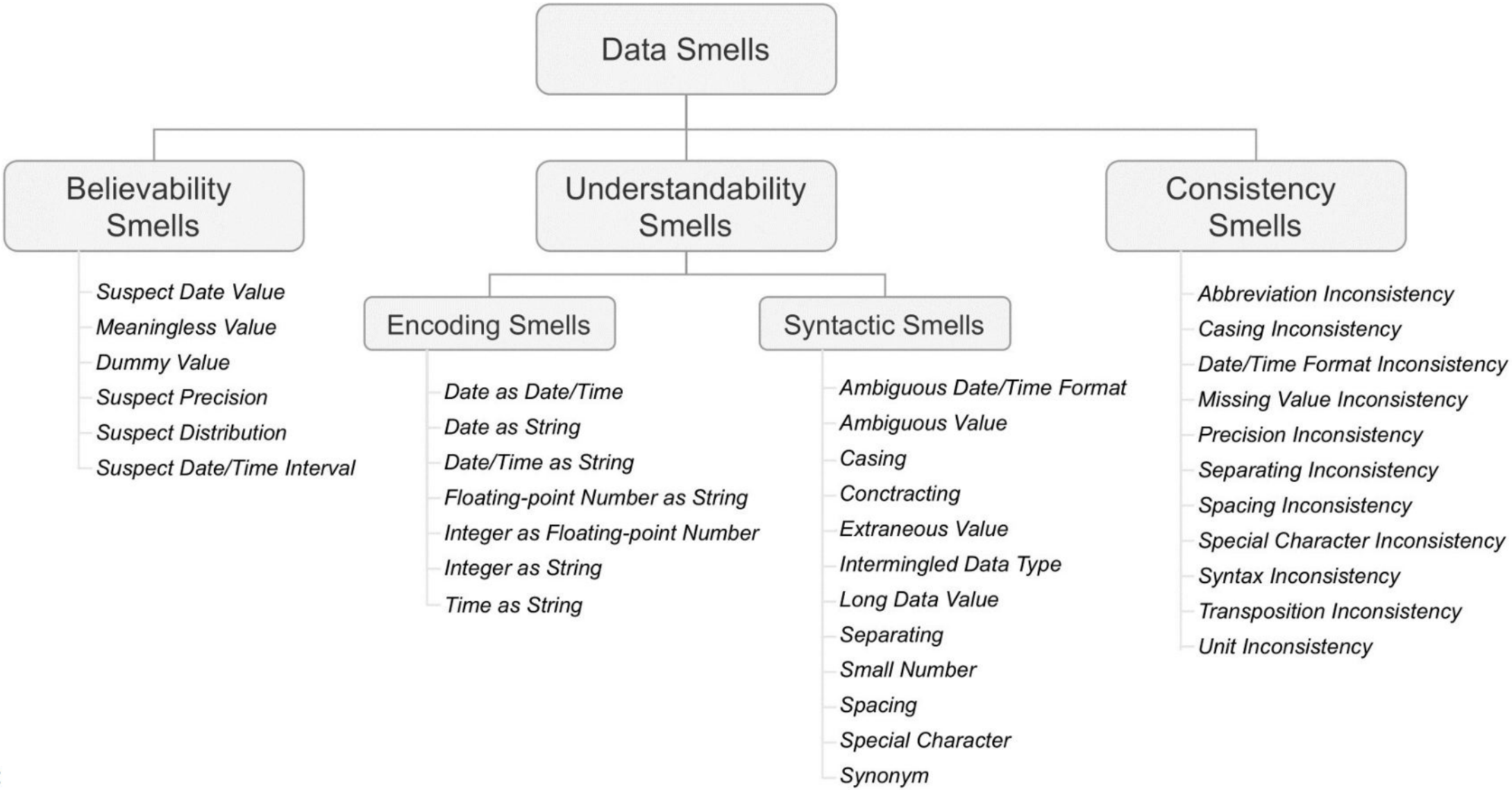
„New York City“

„New York“

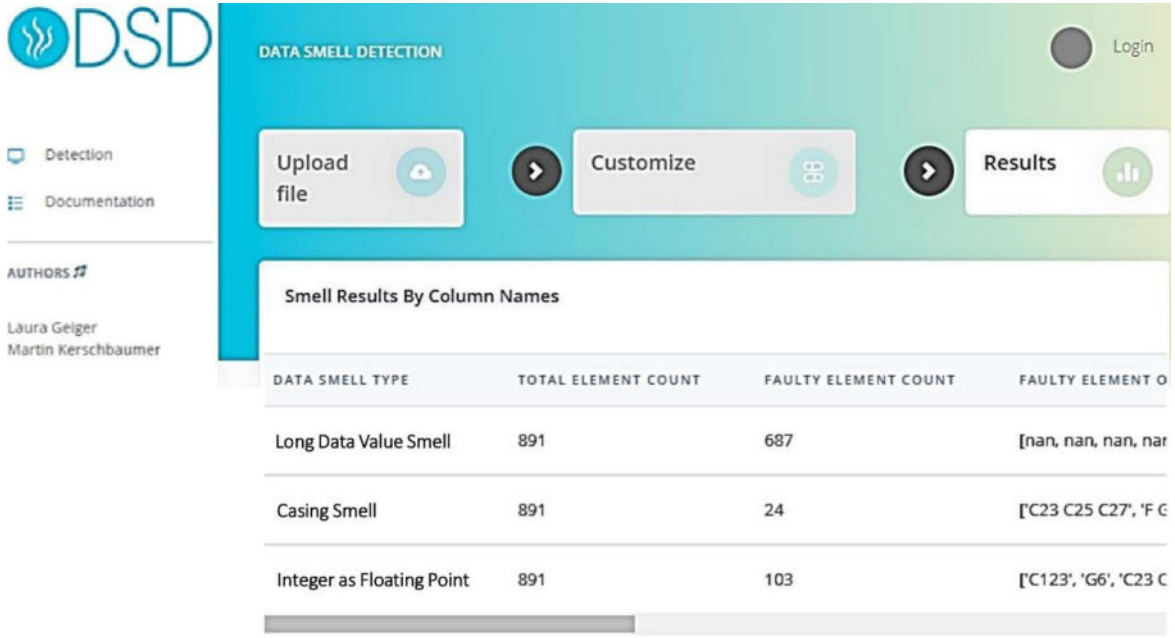
„New Yurc“



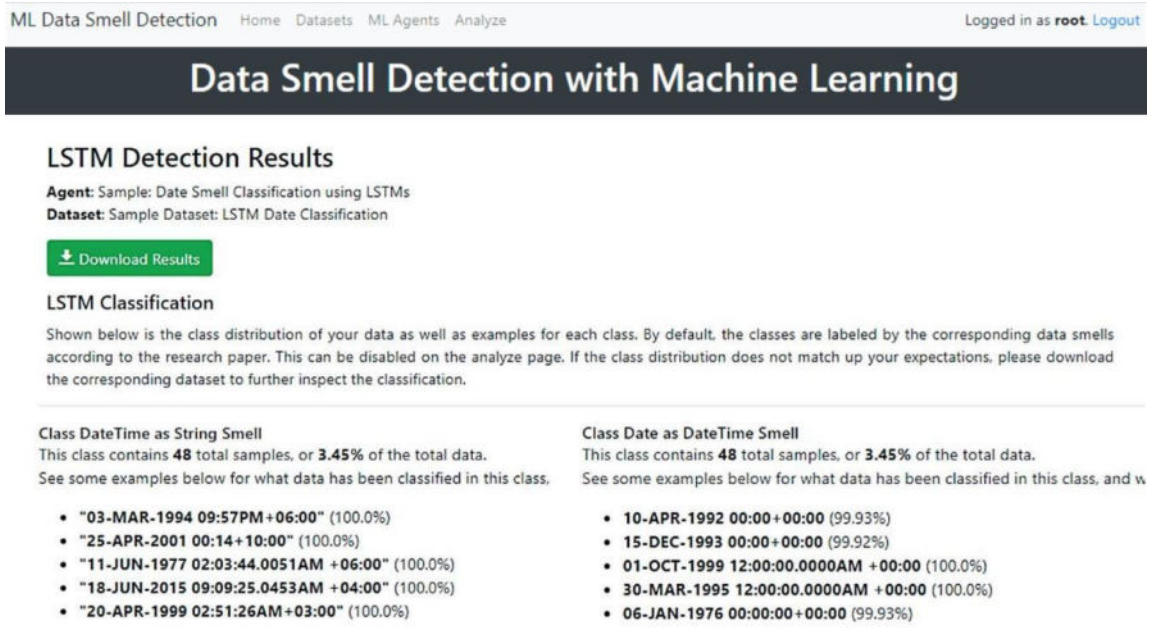
Data Smells Katalog



Data Smells Erkennung



Regelbasierte Erkennung



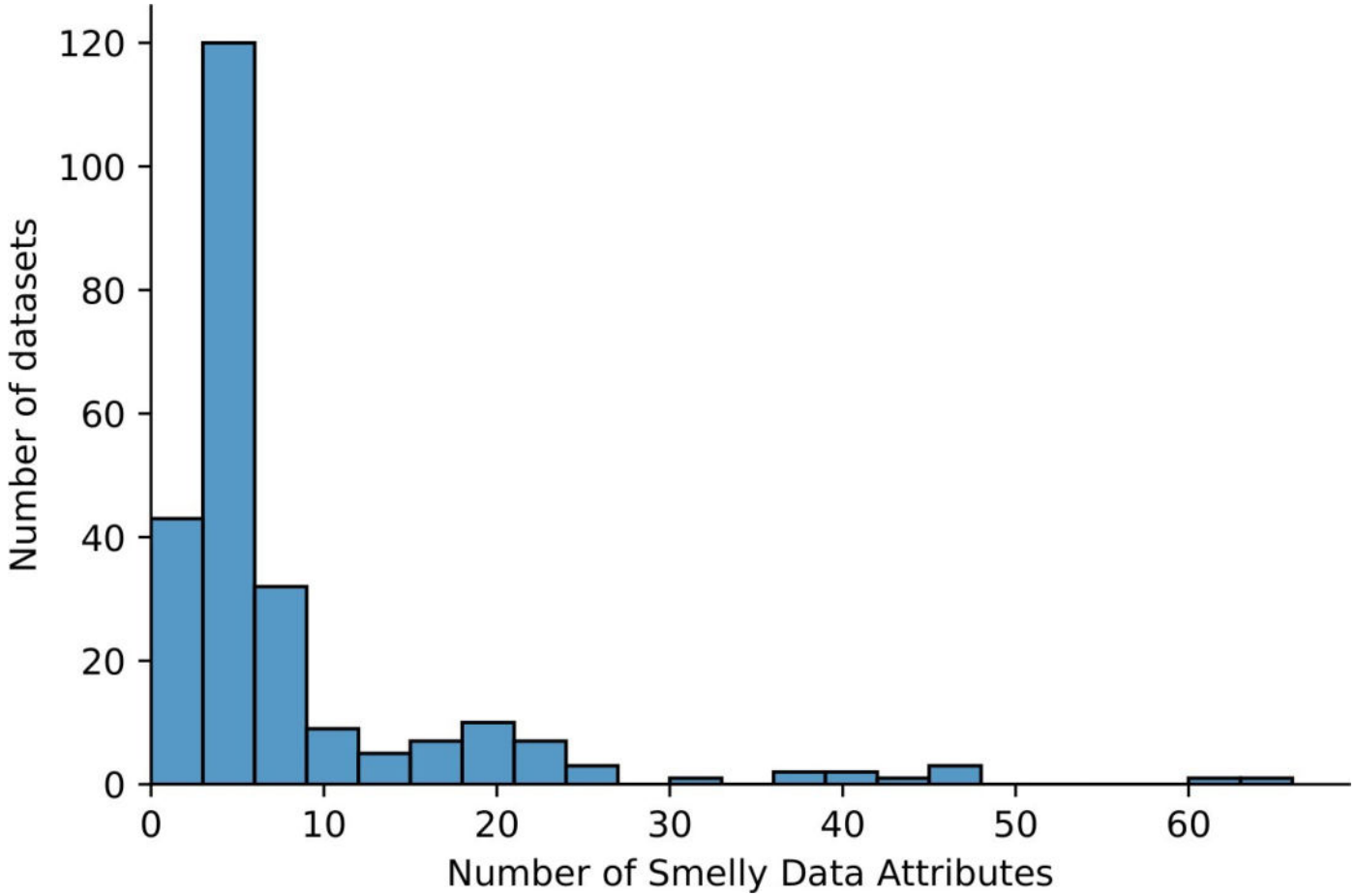
Intelligente Erkennung

Data Smells in der Praxis

246 Kaggle Datensätze

> 2000 Spalten

> 42 Mrd. Zeilen



- Meist drei bis fünf Smells
- Drei Datensätze keine Smells
- 45 Datensätze > 10 Smells
- Ergebnisse handhabbar

Data Smells Beispiele

Date	Daily Confi...	Total Confi...
11-Feb	0	3
12-Feb	0	3
13-Feb	0	3
14-Feb	0	3
15-Feb	0	3
16-Feb	0	3
17-Feb	0	3
18-Feb	0	3

[1]

Ambiguous Date/Time Format

Appointme...	Appointme...	Age
5638447	2016-04-29T00:00:00Z	21
5629123	2016-04-29T00:00:00Z	19
5630213	2016-04-29T00:00:00Z	30
5620163	2016-04-29T00:00:00Z	29
5634718	2016-04-29T00:00:00Z	22
5636249	2016-04-29T00:00:00Z	28

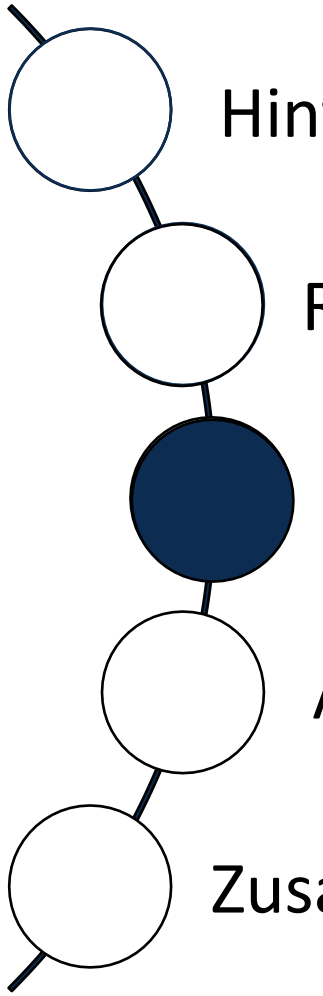
[2]

Date as Date/Time

universität innsbruck

1 https://www.kaggle.com/ravichaubey1506/covid19-india
2 https://www.kaggle.com/joniarroba/noshowappointments

Agenda



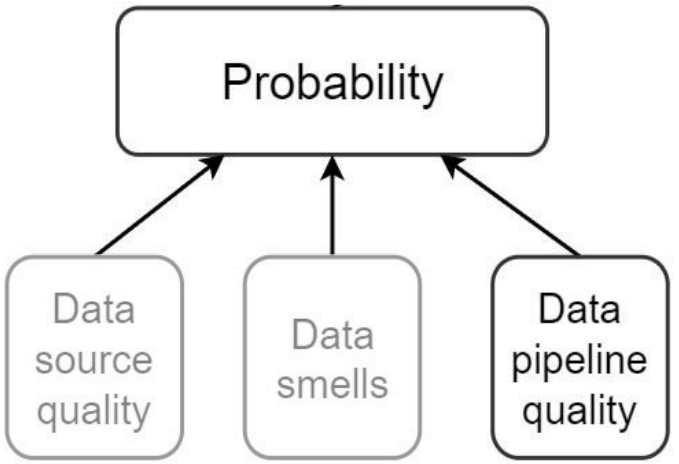
Hintergrund und Motivation

Risikobasiertes Testen von Daten

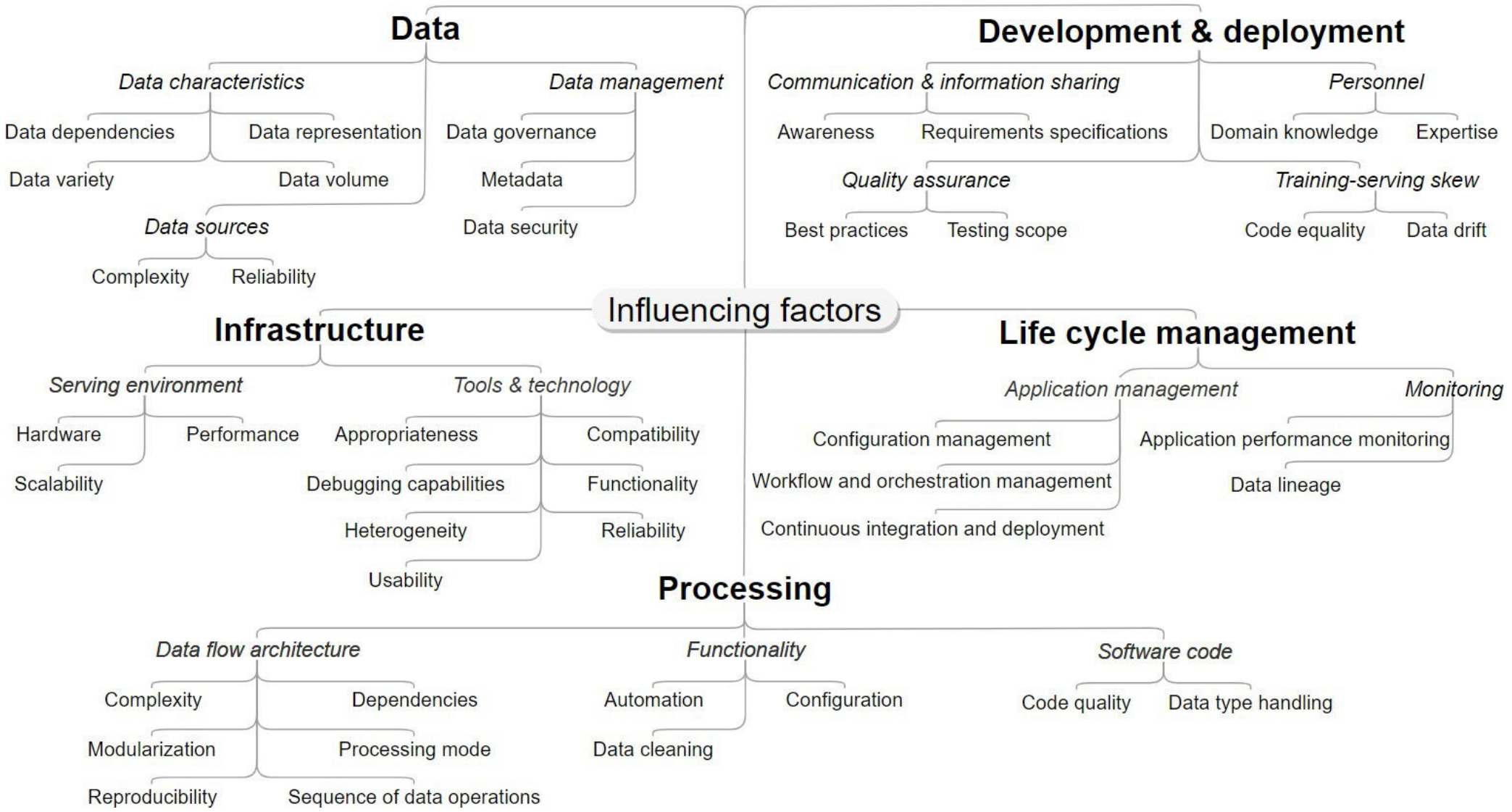
Wahrscheinlichkeit geringer Datenqualität

Anwendungsbeispiel

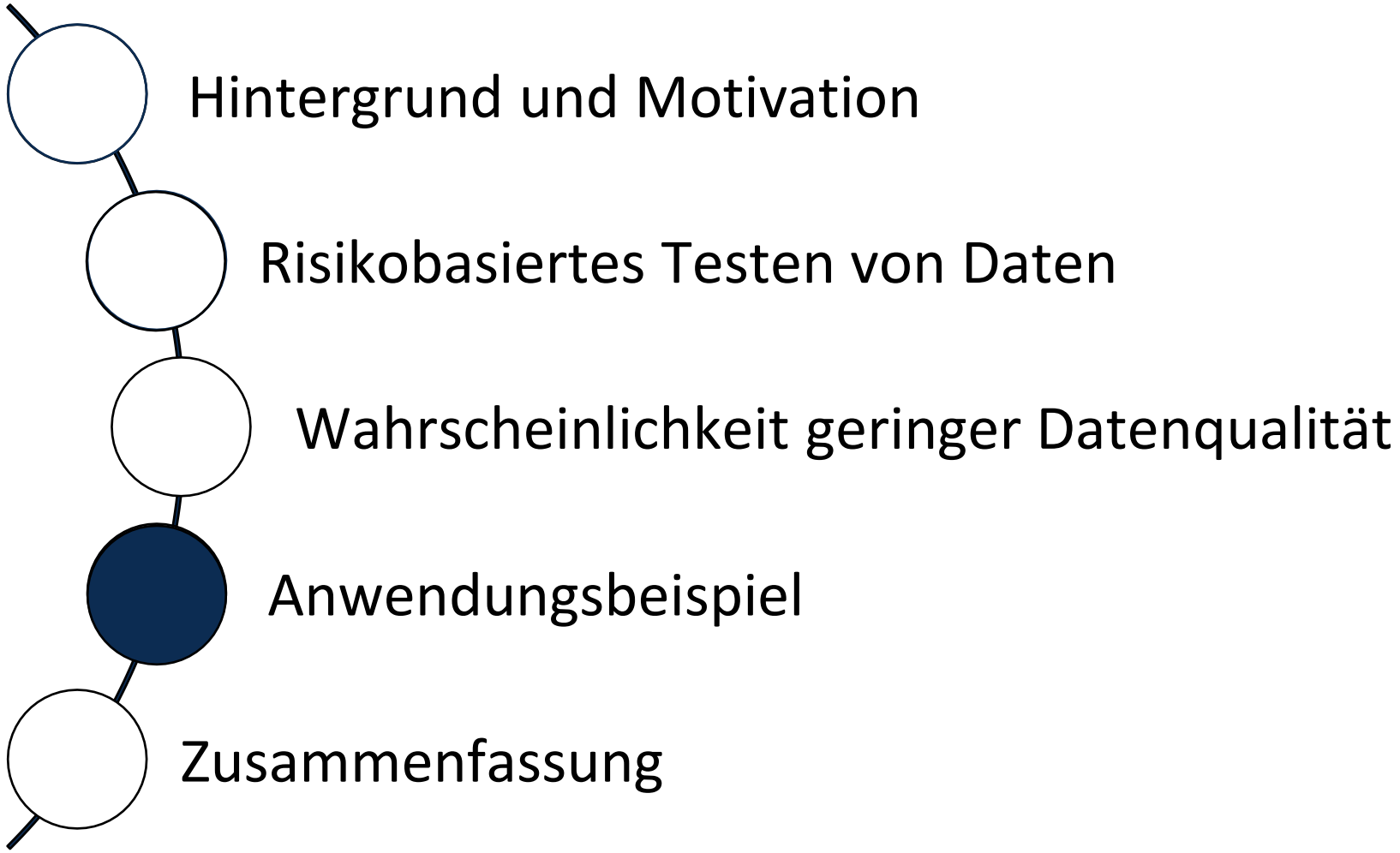
Zusammenfassung



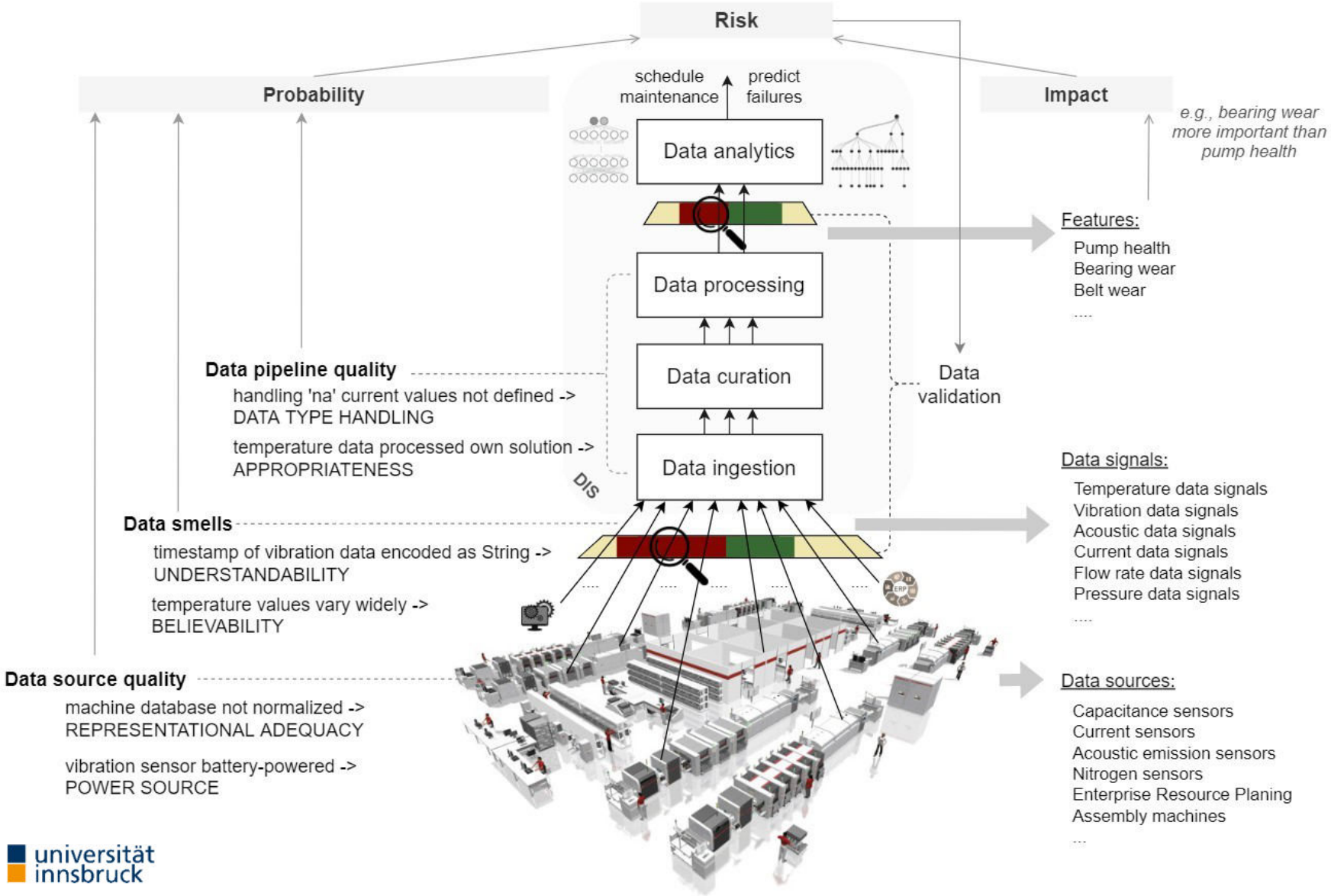
Datenpipeline Qualität

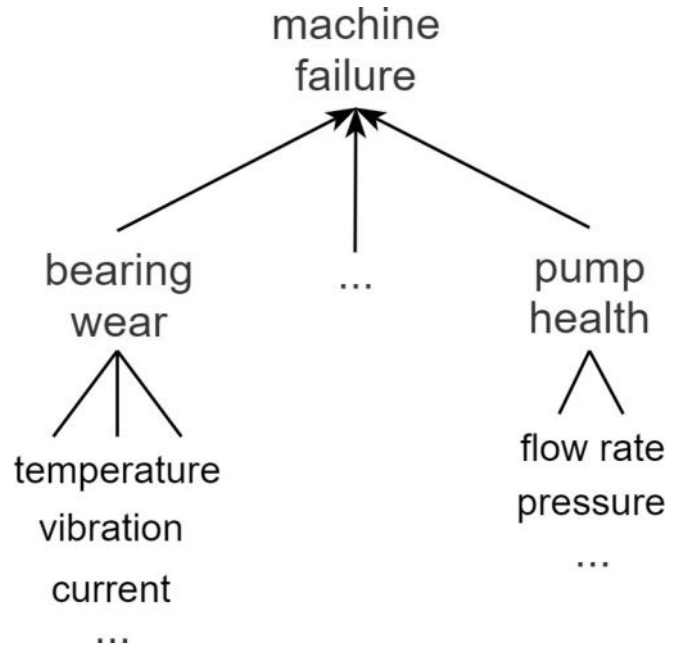


Agenda


Hintergrund und Motivation
Risikobasiertes Testen von Daten
Wahrscheinlichkeit geringer Datenqualität
Anwendungsbeispiel
Zusammenfassung

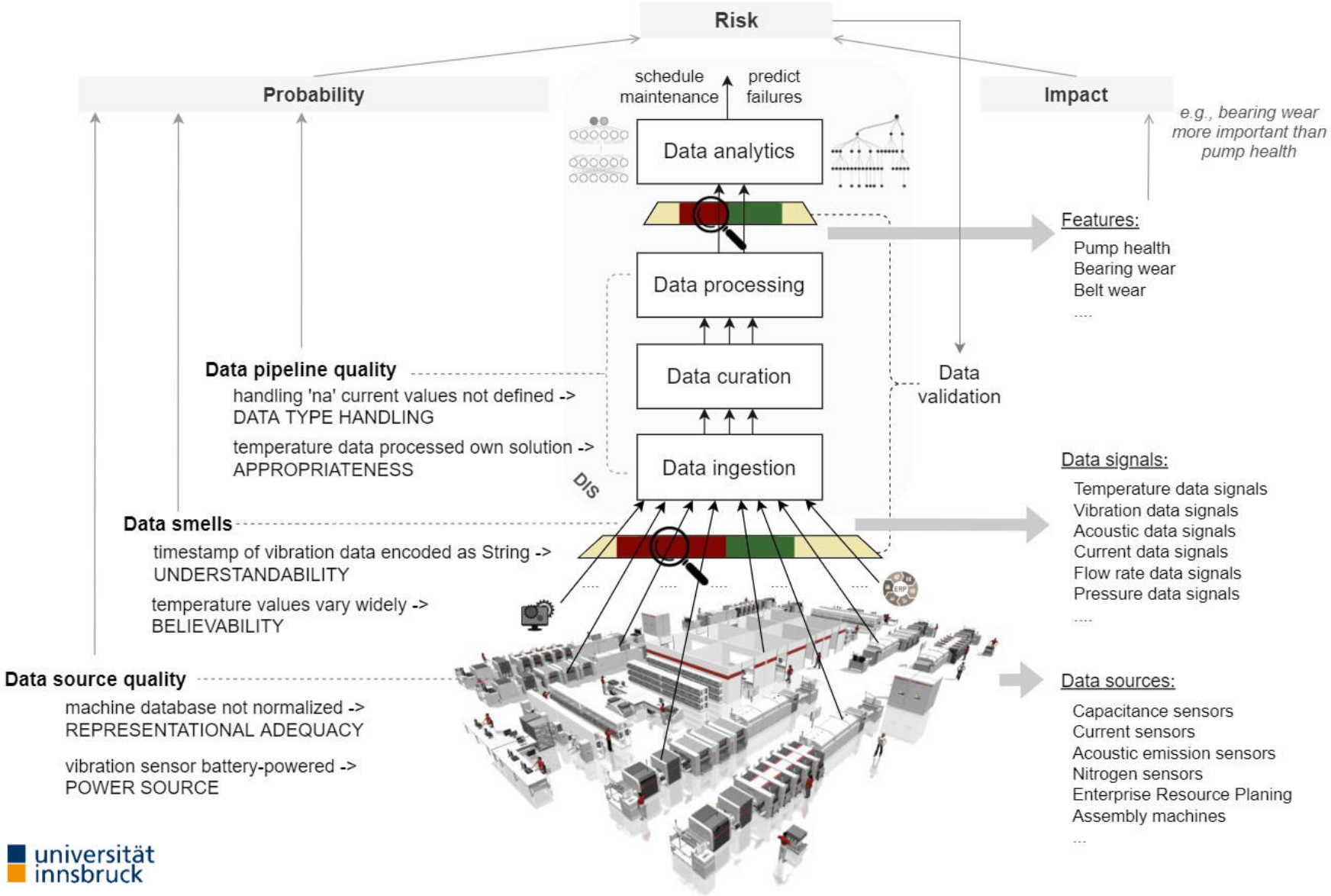
Anwendungsbeispiel

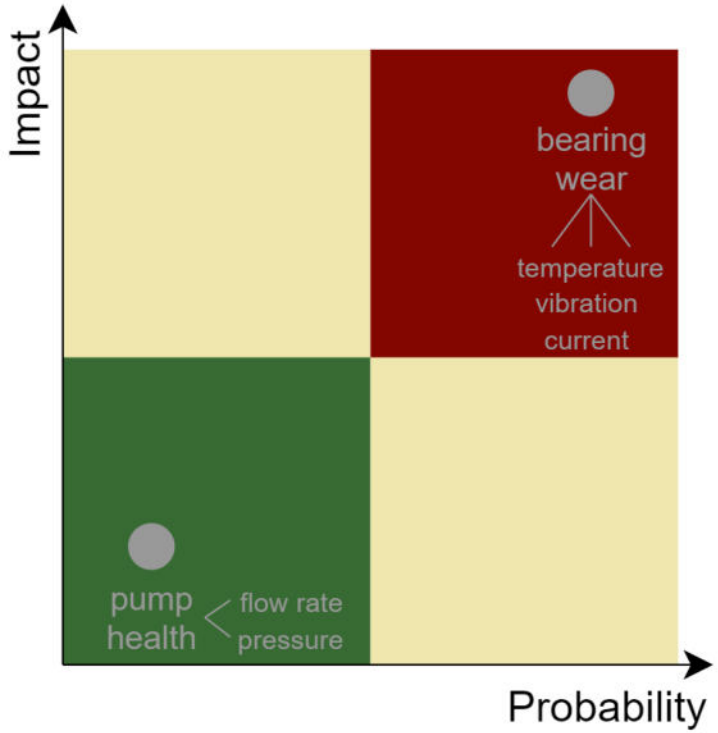




universität innsbruck

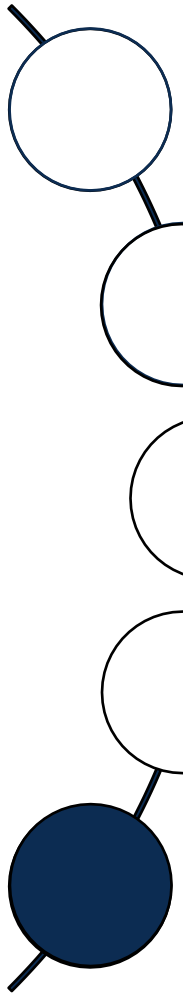
Anwendungsbeispiel





universität innsbruck

Agenda



Hintergrund und Motivation

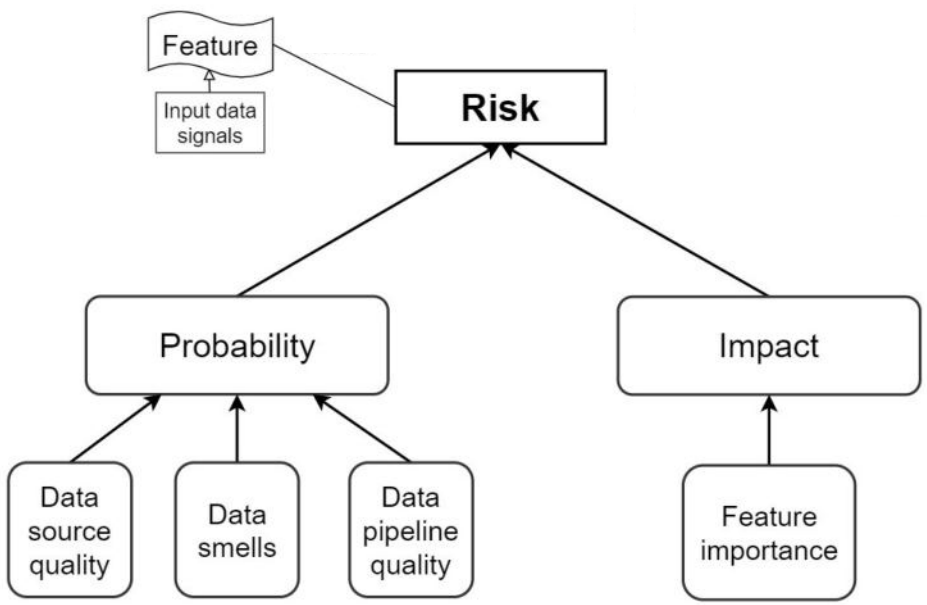
Risikobasiertes Testen von Daten

Wahrscheinlichkeit geringer Datenqualität

Anwendungsbeispiel

Zusammenfassung

Risikobasiertes Testen von Daten



- Ermöglicht selektive Datenvalidierung
- Anwendbar in heterogenen Umgebungen
- Unterstützt fundierte Entscheidungsfindung
- Gezielte Ressourcenzuweisung
- Steigerung von Effizienz und Produktivität

Veröffentlichte Artikel

Foidl, H., & Felderer, M. (2015). Research challenges of industry 4.0 for quality management. In *International Conference on Enterprise Resource Planning Systems*, pp. 121-137. Springer, Cham.

Foidl, H., & Felderer, M. (2016). Data science challenges to improve quality assurance of Internet of Things applications. In *International Symposium on Leveraging Applications of Formal Methods*, pp. 707-726. Springer, Cham.

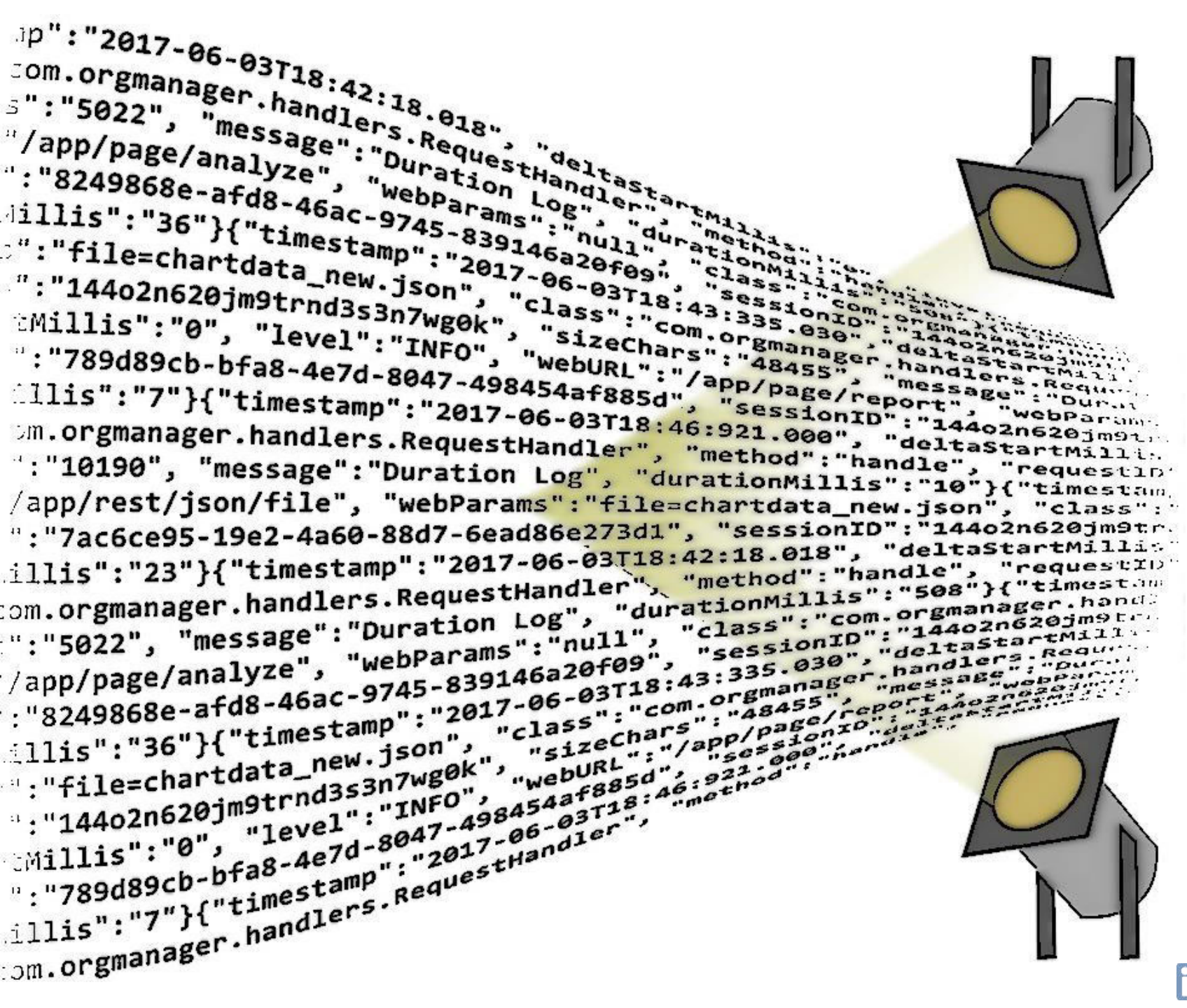
Foidl, H., Felderer, M., & Biffl, S. (2019). Technical debt in data-intensive software systems. In *45th Euromicro Conference on Software Engineering and Advanced Applications* (SEAA), pp. 338-341. IEEE.

Foidl, H., & Felderer, M. (2019). Risk-based data validation in machine learning-based software systems. In P*roceedings of the 3rd ACM SIGSOFT international workshop on machine learning techniques for software quality evaluation*, pp. 13-18.

Foidl, H., Felderer, M., & Ramler, R. (2022). Data smells: categories, causes and consequences, and detection of suspicious data in AI-based systems. In *1st International Conference on AI Engineering - Software Engineering for AI (CAIN)*, pp. 229-239. IEEE/ACM.

Foidl, H., & Felderer, M. (2023). An approach for assessing industrial IoT data sources to determine their data trustworthiness. In *Internet of Things*, 22, 100735.

Foidl, H., Golendukhina, V., Ramler, R., & Felderer, M. (2024). Data pipeline quality: influencing factors, root causes of data-related issues, and processing problem areas for developers. In *Journal of Systems & Software, 207, 111855.*



Danke

Risikobasiertes Testen von Daten

Ing. Harald Foidl, PhD

Universität Innsbruck

harald.foidl@gmail.com

http://www.linkedin.com/in/harald-foidl-7616b790