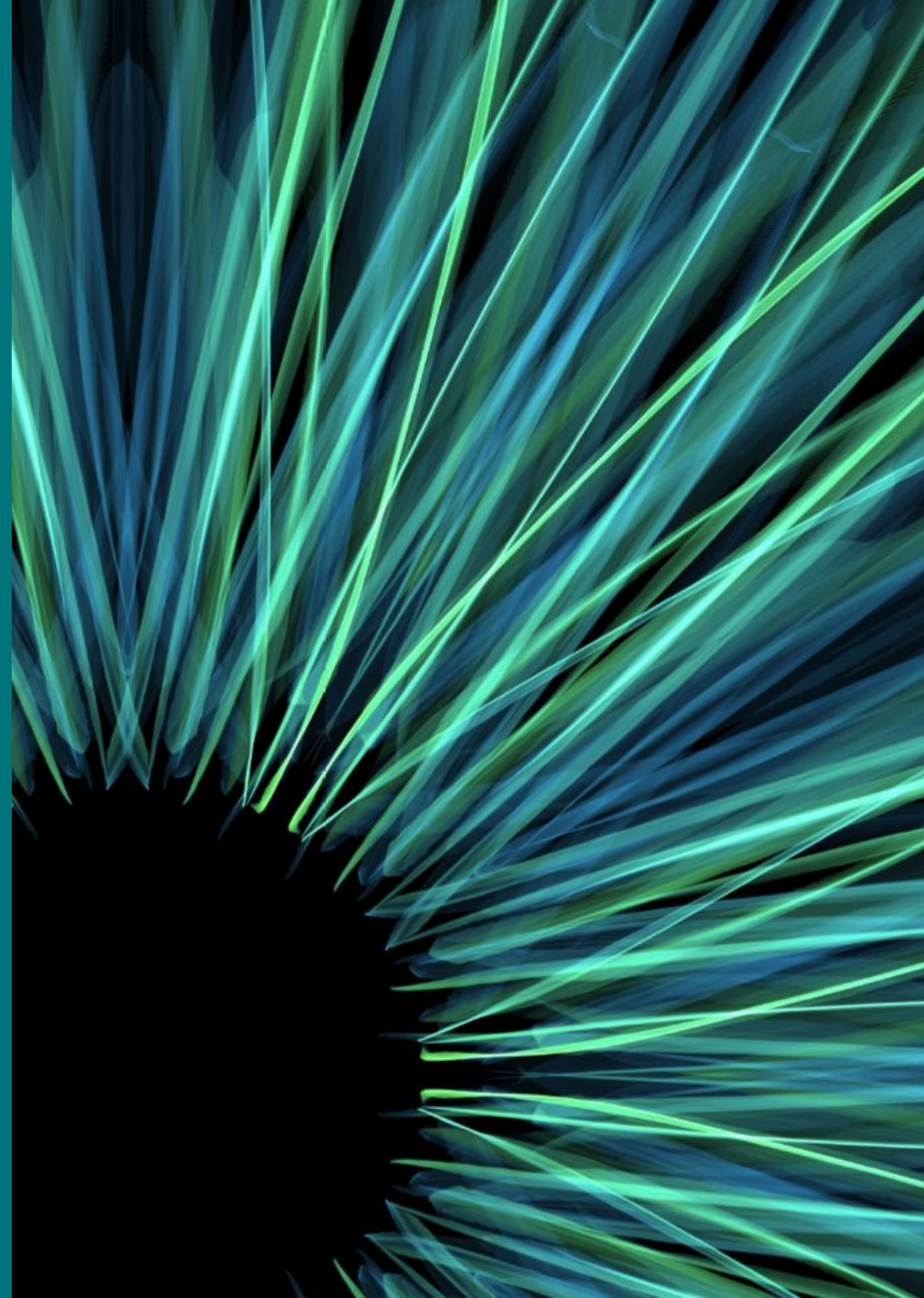




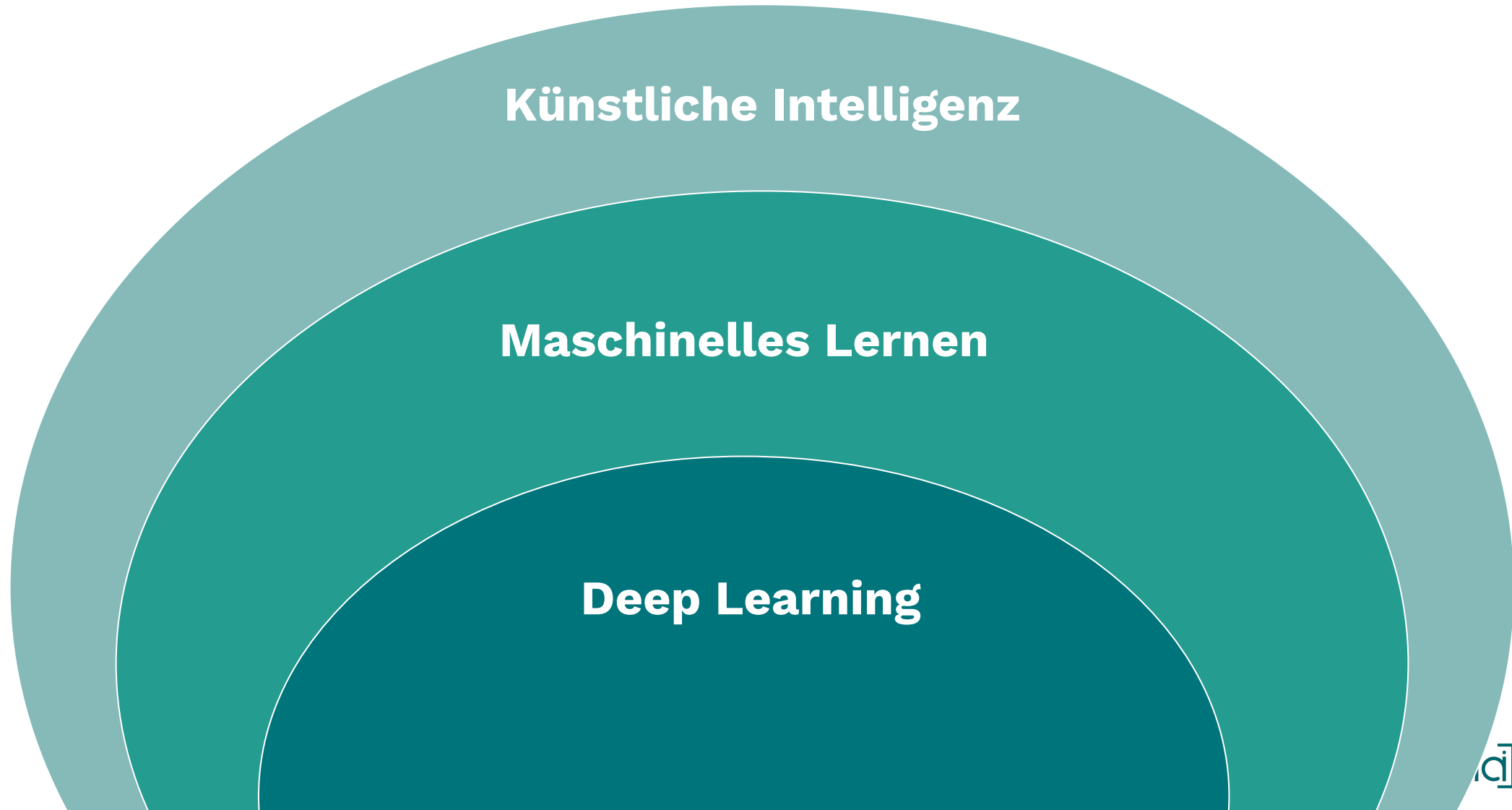
INITIATIVE FOR  
APPLIED ARTIFICIAL  
INTELLIGENCE

# Datenstrategie für KI-Anwendungen

Webinarreihe “KI für den Mittelstand”  
25. November 2020



# KI vs. Maschinelles Lernen vs. Deep Learning

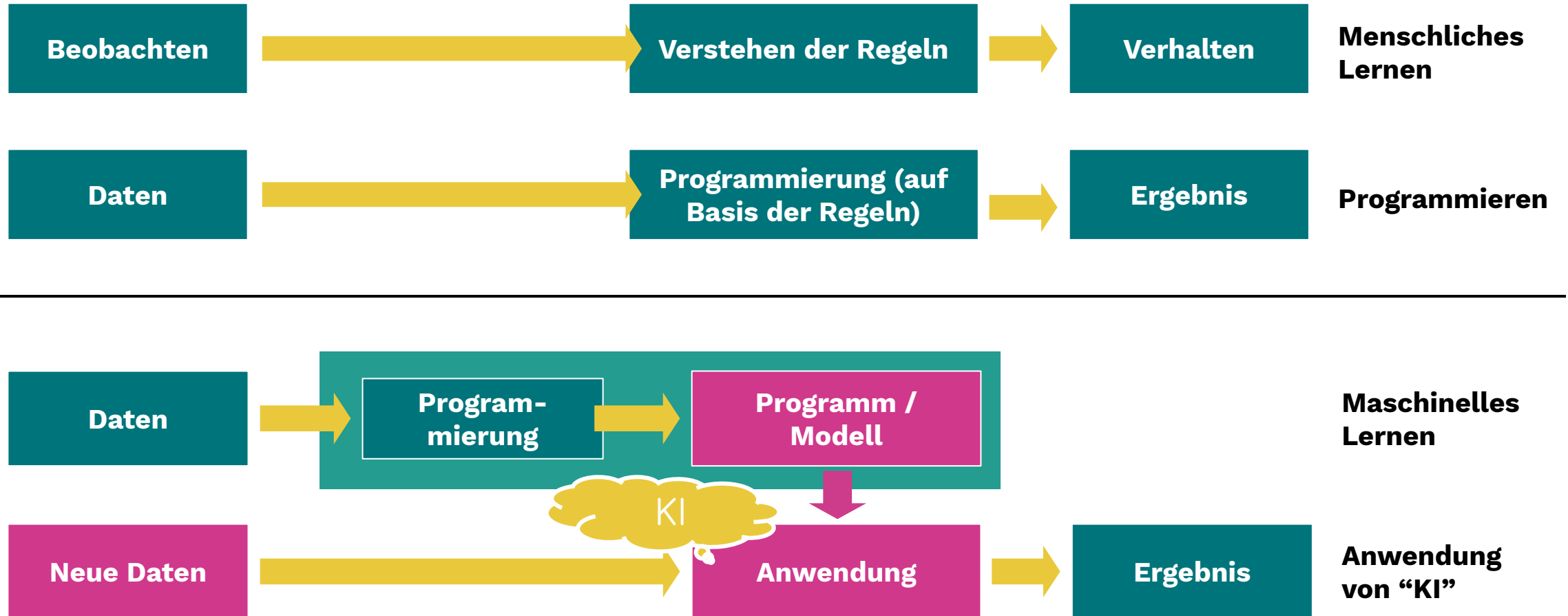


**Künstliche Intelligenz**

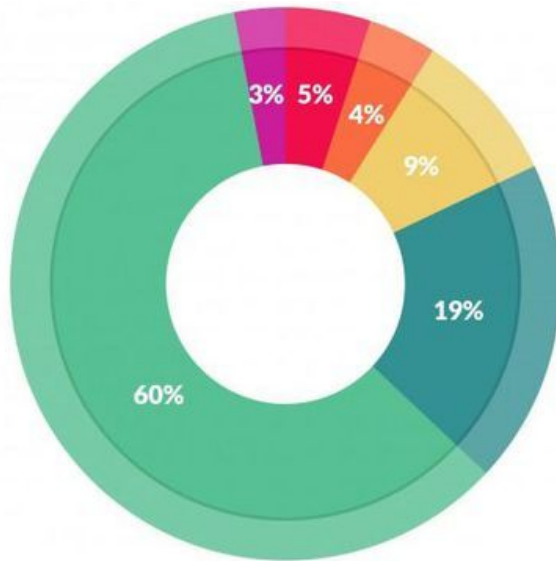
**Maschinelles Lernen**

**Deep Learning**

# Programmierparadigmen verändern sich

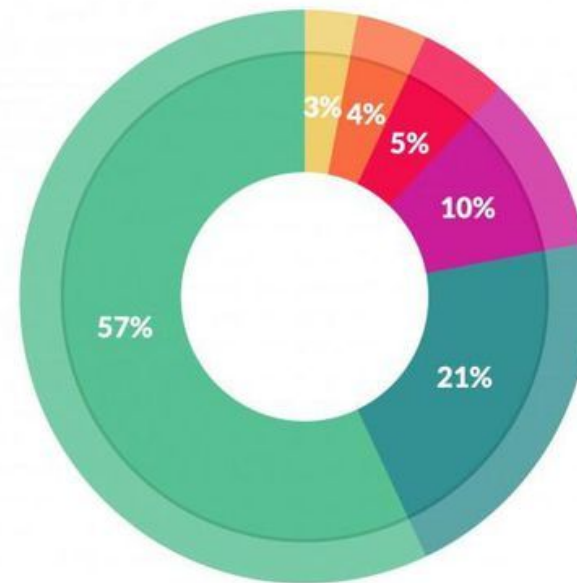


# Sammeln von Trainingsdaten



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%



What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

# Sammeln von Trainingsdaten

” Among leading AI teams, many can likely replicate others’ software in, at most, 1–2 years. But it is exceedingly difficult to get access to someone else’s data. Thus data, **rather than software, is the defensible barrier for many businesses.** ”

Andrew Ng

- Wachsender **Wert von Daten**
- **Digitalisierung Voraussetzung** für skalierte KI-Lösungen
- **Datenstrategie auf Basis von Anwendungspotenzialen** sollte erarbeitet werden!

# Kommt Ihnen das bekannt vor?

Data is the new oil.



Clive Humby (2006)

# Das auch?

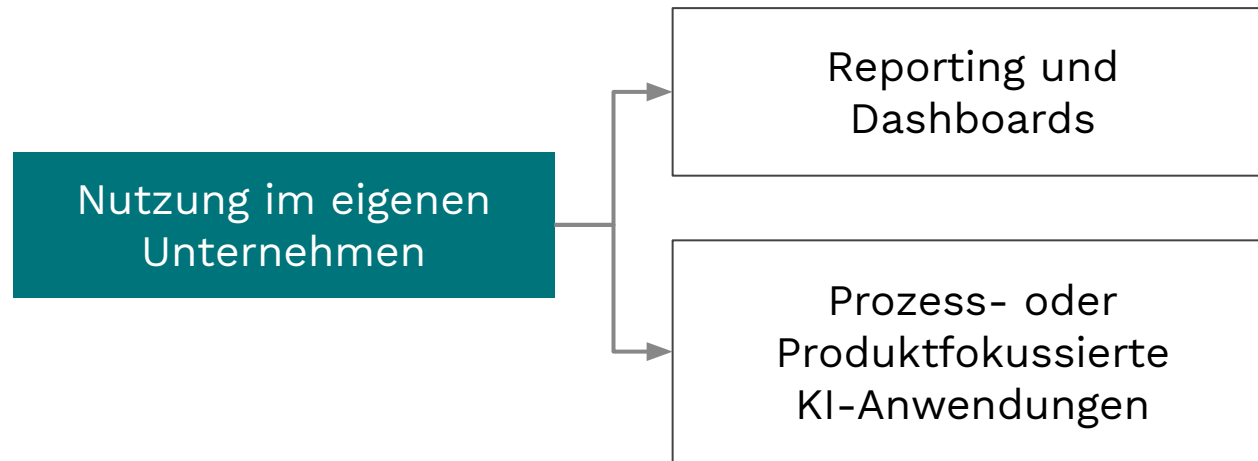
Data is the new oil. It's valuable, but if unrefined it cannot really be used. **It has to be changed** into gas, plastic, chemicals, etc **to create a valuable entity** [...].



Clive Humby (2006)



# Um Wert zu generieren müssen Daten genutzt werden



Nutzung durch andere Unternehmen

## Source of value

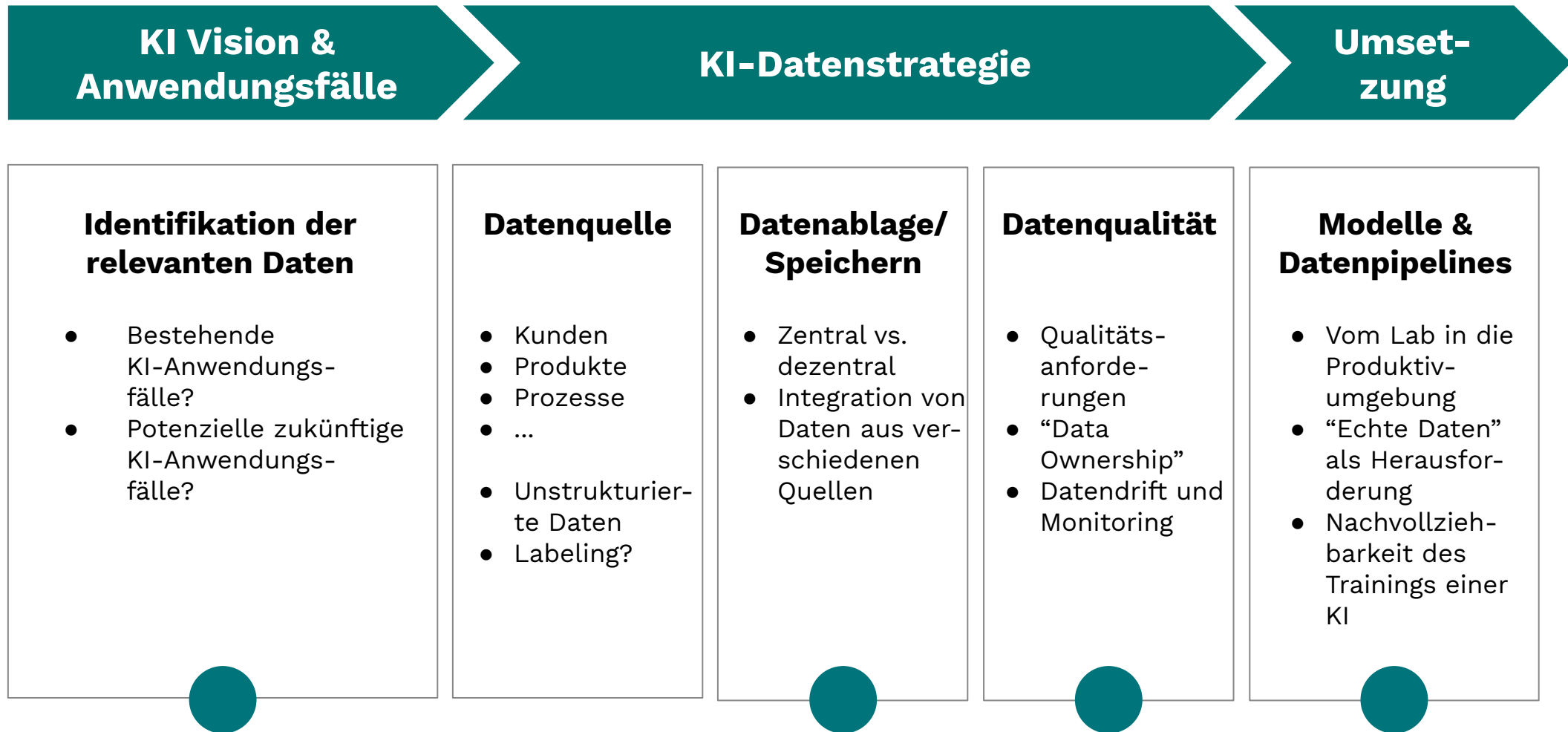
- Verbesserung der Entscheidungsqualität

- Automatisierung oder Beschleunigung, Customer Experience, Image... **Datenstrategie als Grundlage**

- Verkauf von Daten
- Erbringung datenbasierter Dienstleistungen oder bereitstellung von aus Daten gewonnener Informationen



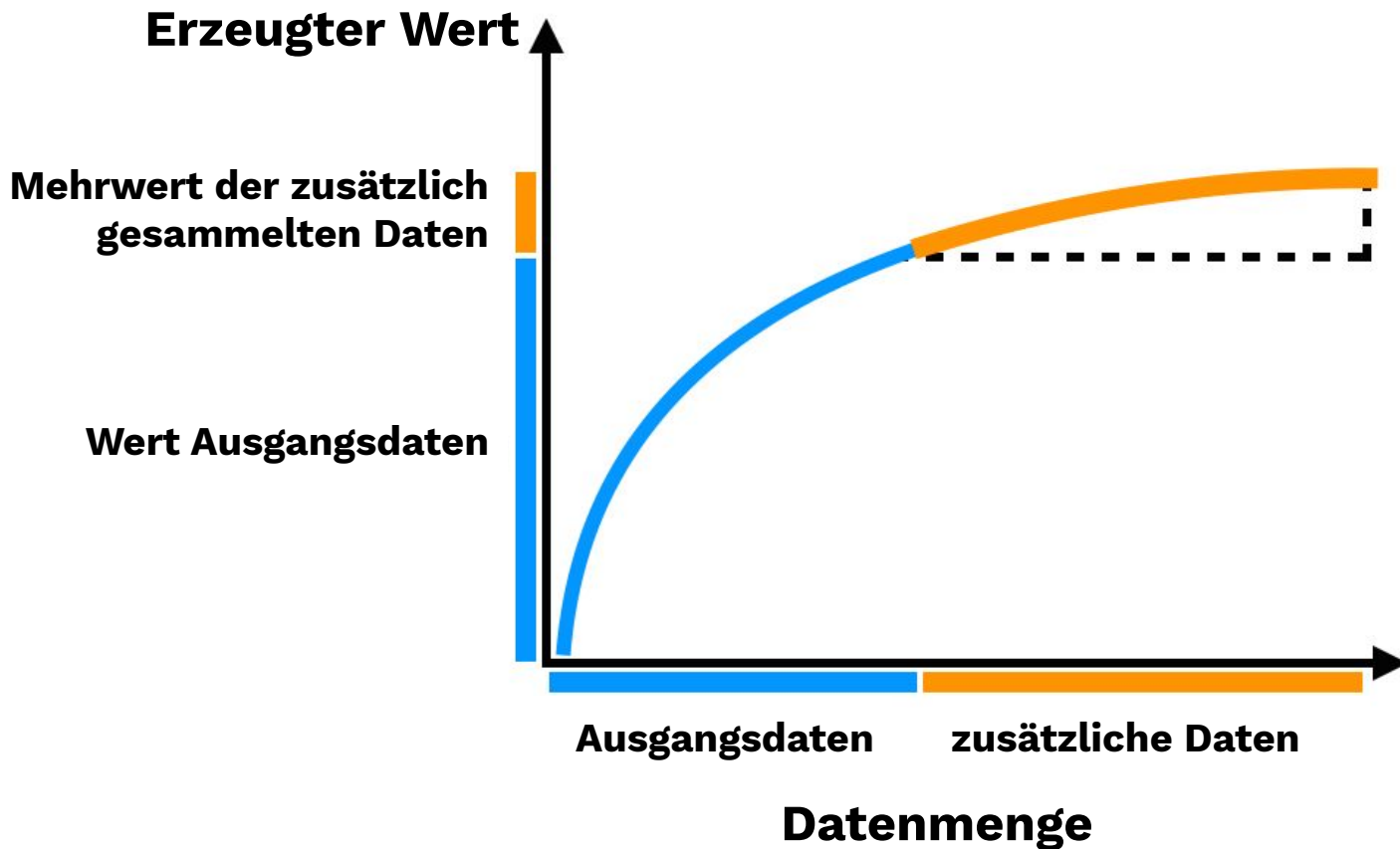
# KI Stellt diverse Herausforderungen an eine Datenstrategie



adressiert im Rahmen des heutigen Webinars

# Identifikation der relevanten Daten

Manchmal kann weniger mehr sein



➤ Auf die **Daten konzentrieren**, die wirklich wertvoll sind (für die also ein **Anwendungsfall vorhanden** ist!)

Das Sammeln von Daten kann **auch negative Konsequenzen** haben:

- Kosten für **Speicherung**
- Aufwand des **Sammelns und Pflege** von Daten
- Kosten für **IT-Security**
- Möglichkeit von **Datenzugriff durch Unbefugte**
- **Rechtliche Risiken** (GDPR)

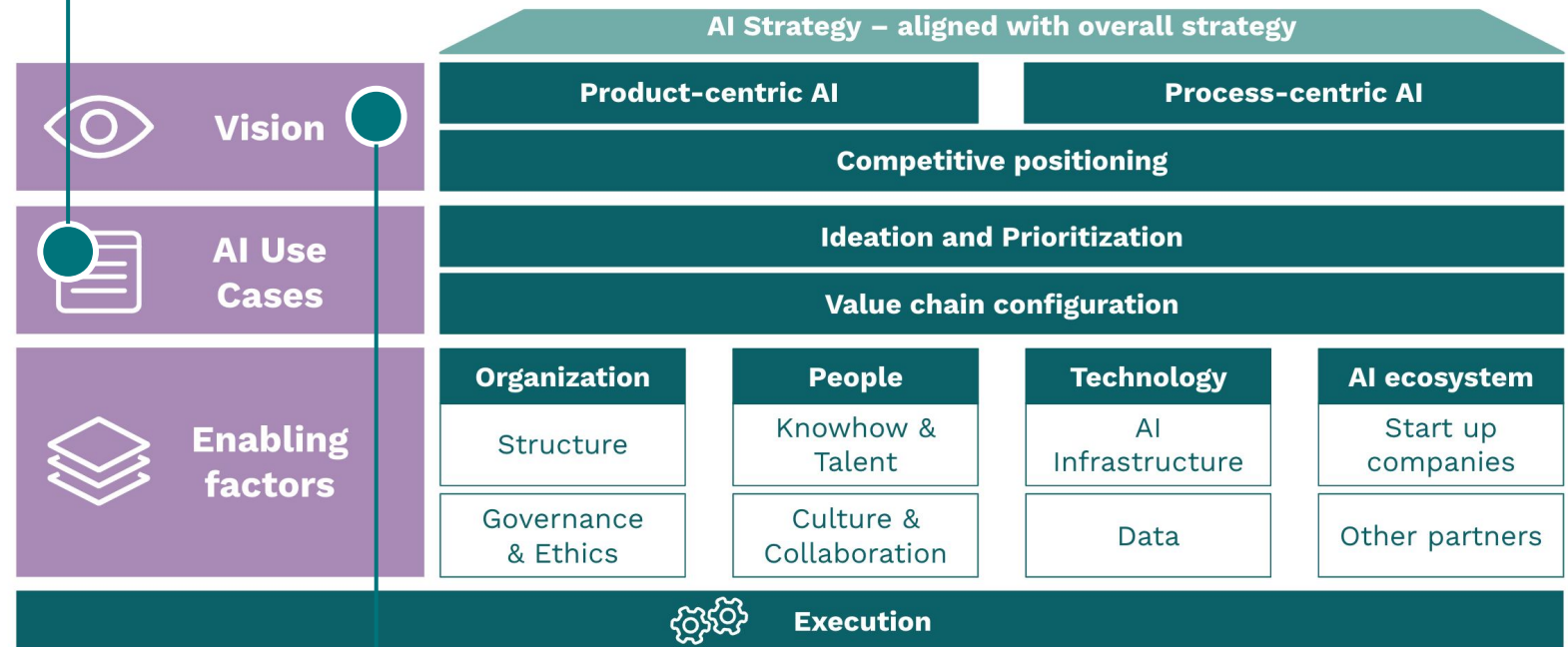
# Identifikation der relevanten Daten

## Anwendungsfälle und Vision helfen bei Priorisierung



➤ Aus bereits identifizierten Anwendungsfällen lassen sich Datenanforderungen ableiten

➤ KI-Vision definiert potenziell wertstiftende breitere Anwendungsfelder, die auf “wertvolle” Daten schließen lassen

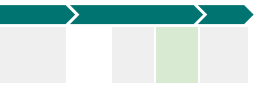


# Speichern von Daten

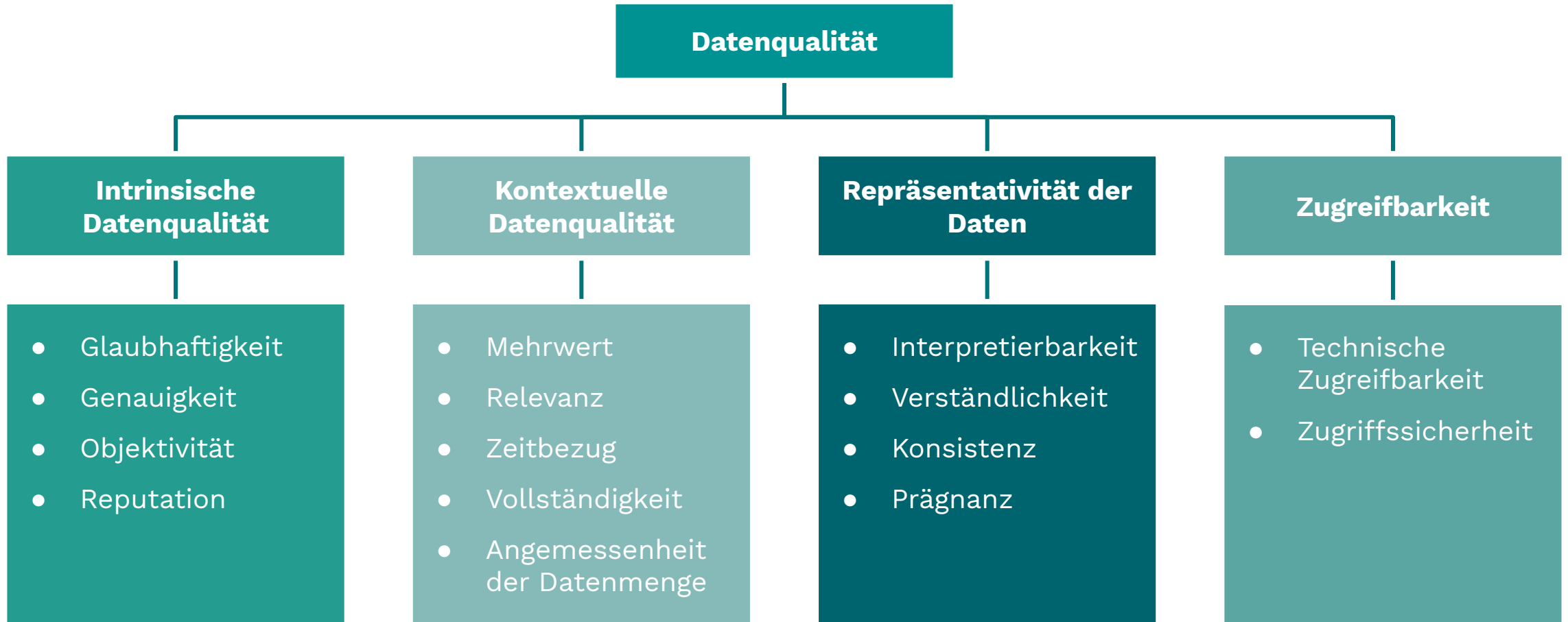
## Vor- und Nachteile verschiedener Ansätze



	Zentrale Datenspeicherung	Hybrider Ansatz*	Dezentrale Datenspeicherung
<b>Sichtbarkeit</b> verfügbarer Daten im Unternehmen	+	+	-
Möglichkeit des <b>zentralen Monitoring</b> von <b>Datenqualität</b>	+	-	-
Einfacher <b>Zugriff und Nutzbarkeit</b> für Gesamtunternehmen	+	+	-
<b>“Single Version of Truth”</b>	+	-	+
<b>Flexibilität</b>	-	~	+



# Datenqualität - Was steckt dahinter?



# Datenqualität - Was steckt dahinter?

## Datenqualität

### Intrinsische Datenqualität

### Kontextuelle Datenqualität

### Repräsentativität der Daten

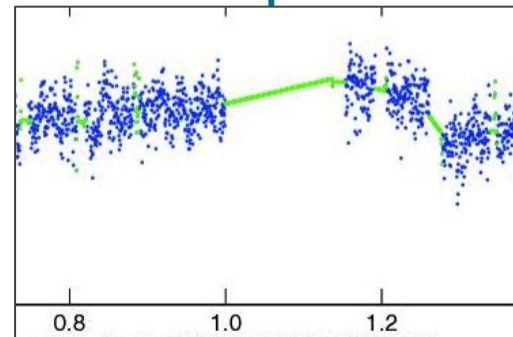
### Zugreifbarkeit

Donald J. Trump  
@realDonaldTrump

There is NO WAY (ZERO!) that Mail-In Ballots will be anything less than substantially fraudulent. Mail boxes will be robbed, ballots will be forged & even illegally printed out & fraudulently signed. The Governor of California is sending Ballots to millions of people, anyone....

Get the facts about mail-in ballots

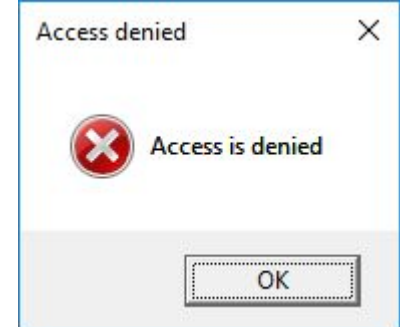
**Glaubhaftigkeit**



**Vollständigkeit**

	A	B	C	D	E	
1	File Name	EX_COEF_355	BA_COEF_355	EX_COEF_532	BA_COEF_532	
2	FILE: r1200i000-0ln140410.LID d=0.01	m = 1.200 - .000E+00	2.02E-06	5.38E-09	1.56E-06	3.00E-09
3	FILE: r1200i100-0ln140410.LID d=0.01	m = 1.200 - .100E+01	1.61E-06	9.46E-09	1.67E-06	9.56E-09
4	FILE: r1200i100-1ln140410.LID d=0.01	m = 1.200 - .100E+00	1.57E-06	5.74E-10	1.41E-06	6.69E-10
5	FILE: r1200i100-2ln140410.LID d=0.01	m = 1.200 - .100E-01	1.94E-06	2.56E-09	1.53E-06	2.06E-09
6	FILE: r1200i100-3ln140410.LID d=0.01	m = 1.200 - .100E-02	2.01E-06	4.96E-09	1.56E-06	2.88E-09
7	FILE: r1200i100-4ln140410.LID d=0.01	m = 1.200 - .100E-03	2.02E-06	5.34E-09	1.56E-06	2.99E-09
8	FILE: r1200i100-5ln140410.LID d=0.01	m = 1.200 - .100E-04	2.02E-06	5.38E-09	1.56E-06	3.00E-09
9	FILE: r1200i100-8ln140410.LID d=0.01	m = 1.200 - .100E-07	2.02E-06	5.38E-09	1.56E-06	3.00E-09
10	FILE: r1200i125-1ln140410.LID d=0.01	m = 1.200 - .125E+00	1.53E-06	6.24E-10	1.40E-06	6.88E-10
11	FILE: r1200i125-2ln140410.LID d=0.01	m = 1.200 - .125E-01	1.92E-06	2.18E-09	1.53E-06	1.89E-09
12	FILE: r1200i150-1ln140410.LID d=0.01	m = 1.200 - .150E+00	1.51E-06	6.93E-10	1.40E-06	7.39E-10
13	FILE: r1200i150-2ln140410.LID d=0.01	m = 1.200 - .150E-01	1.91E-06	1.88E-09	1.52E-06	1.75E-09
14	FILE: r1200i175-1ln140410.LID d=0.01	m = 1.200 - .175E+00	1.49E-06	7.77E-10	1.40E-06	8.13E-10
15	FILE: r1200i175-2ln140410.LID d=0.01	m = 1.200 - .175E-01	1.89E-06	1.63E-09	1.51E-06	1.63E-09
16	FILE: r1200i200-1ln140410.LID d=0.01	m = 1.200 - .200E+00	1.48E-06	8.76E-10	1.41E-06	9.06E-10
17	FILE: r1200i200-2ln140410.LID d=0.01	m = 1.200 - .200E-01	1.87E-06	1.43E-09	1.51E-06	1.52E-09
18	FILE: r1200i225-1ln140410.LID d=0.01	m = 1.200 - .225E+00	1.47E-06	9.88E-10	1.42E-06	1.01E-09

**Verständlichkeit / Interpretierbarkeit**



**Zugriffssicherheit**

Beispiele

# Data Ownership

## Klare Verantwortlichkeiten für Daten und Datenqualität



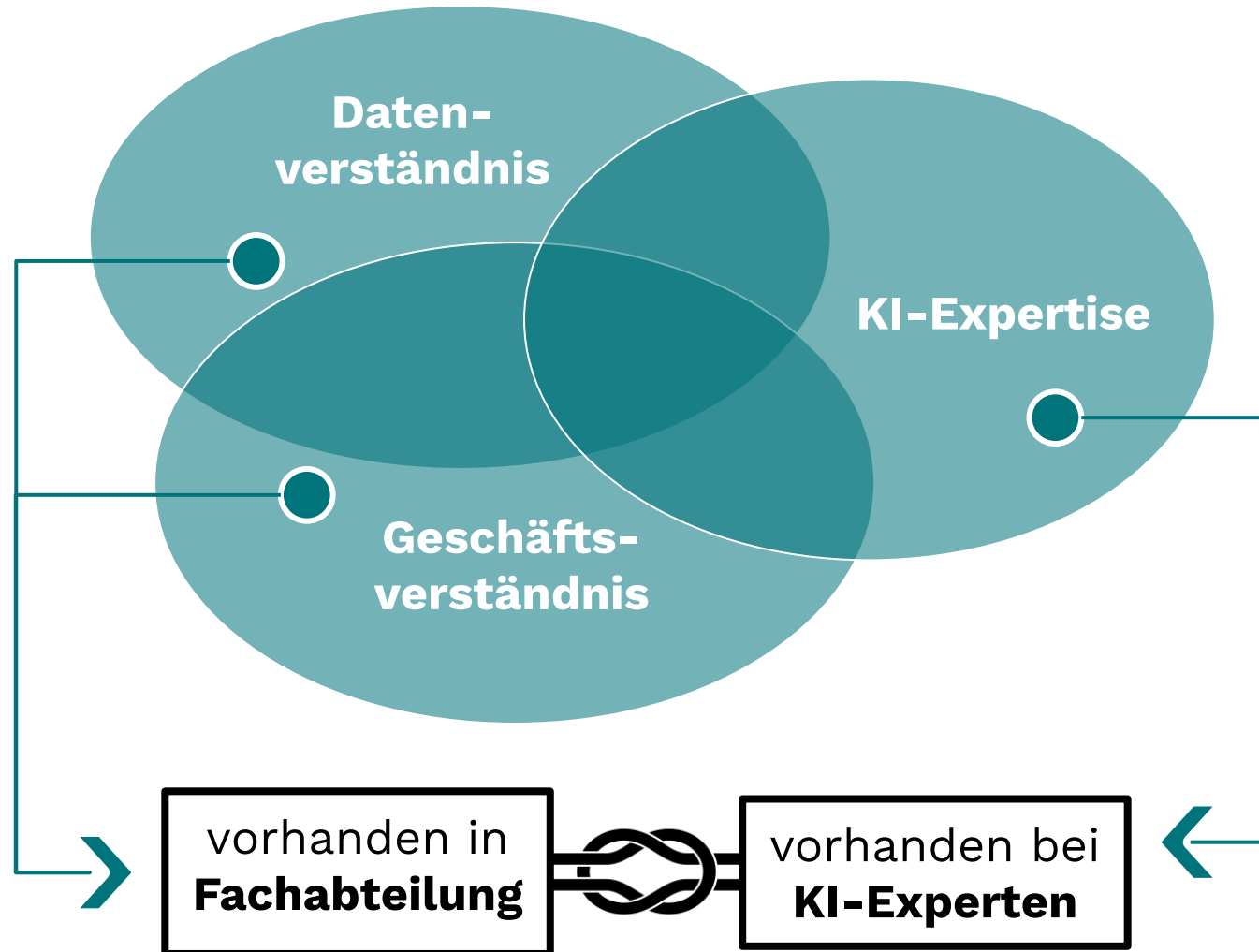
“ Der Data Owner ist **für einen bestimmten Teil der Daten** innerhalb einer Organisation **verantwortlich**. Er hat **volle Kontrolle** über die Daten und stellt unter anderem deren **Schutz und Qualität** sicher. Zudem übt er die Kontrolle über **Datenzugriffe** aus.

“

<https://www.bigdata-insider.de/>

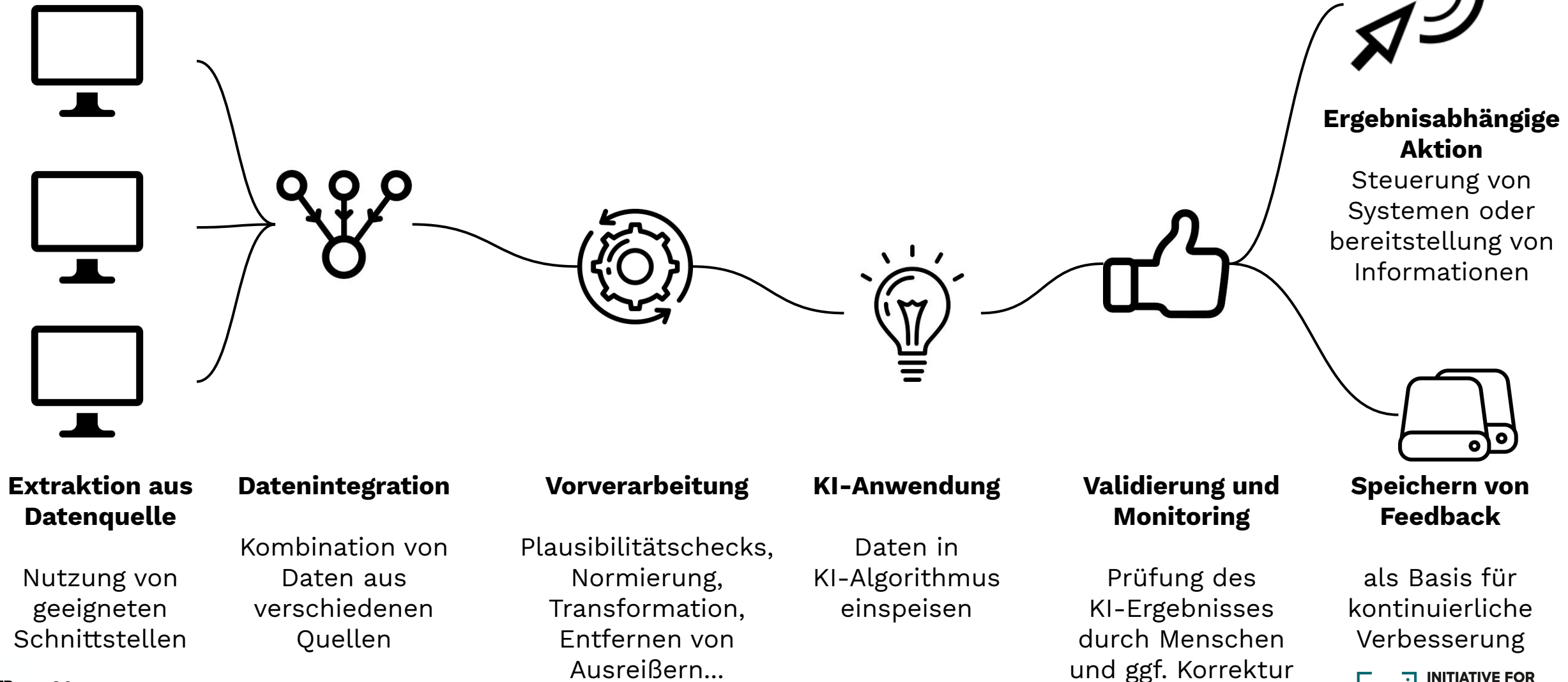


# Fachabteilungen und KI-Experten müssen eng zusammenarbeiten, um Datenqualität sicherzustellen



# KI in der Anwendung

## Datenpipelines für Echtzeitanwendungen



# Sinnvolle erste Schritte

- **Erst** über **Anwendungsfälle** nachdenken, dann über die dafür notwendigen Daten
- Daten **priorisieren** und realistische Ziele setzen
- **Früh starten** - der Aufbau von umfassenden Datensätzen kann viel Zeit in Anspruch nehmen
- Frühzeitig **KI-Experten einbinden** um sicherzustellen, dass Daten in geeigneter Form (Qualität) erfasst werden
- Für wichtige Daten **Qualitätsstandards festlegen**, dokumentieren und durchsetzen

# Vielen Dank für die Aufmerksamkeit

**Dr. Hendrik Brakemeier**

Head of AI Journey

[brakemeier@unternehmertum.de](mailto:brakemeier@unternehmertum.de)

<https://www.linkedin.com/in/hbrakemeier>

**appliedAI Initiative**

UnternehmerTUM GmbH

Lichtenbergstraße 6

685748 Garching

**[info@appliedAI.de](mailto:info@appliedAI.de)**

**<https://appliedAI.de>**