

› Data Science Grundlagen Einführung

Prof. Dr. Carsten Lanquillon | Fakultät Wirtschaft und Verkehr | Wirtschaftsinformatik

90 %

der in der Welt verfügbaren Daten
nicht älter als

2 Jahre

täglich

2.500.000.000.000 MB

neue Daten

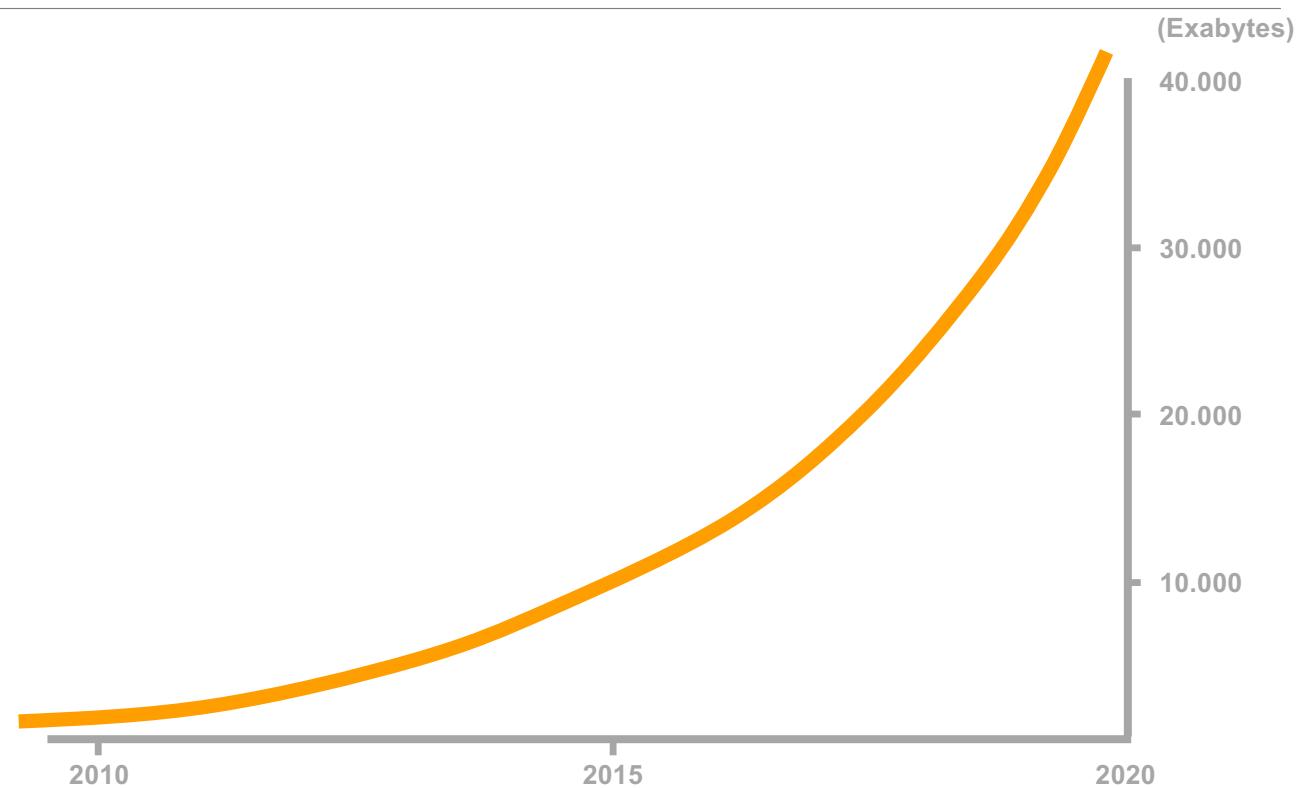
Quelle: IBM & Leitl, M: Was ist Big Data, HBM, 04/2014

Exponentielles Datenwachstum

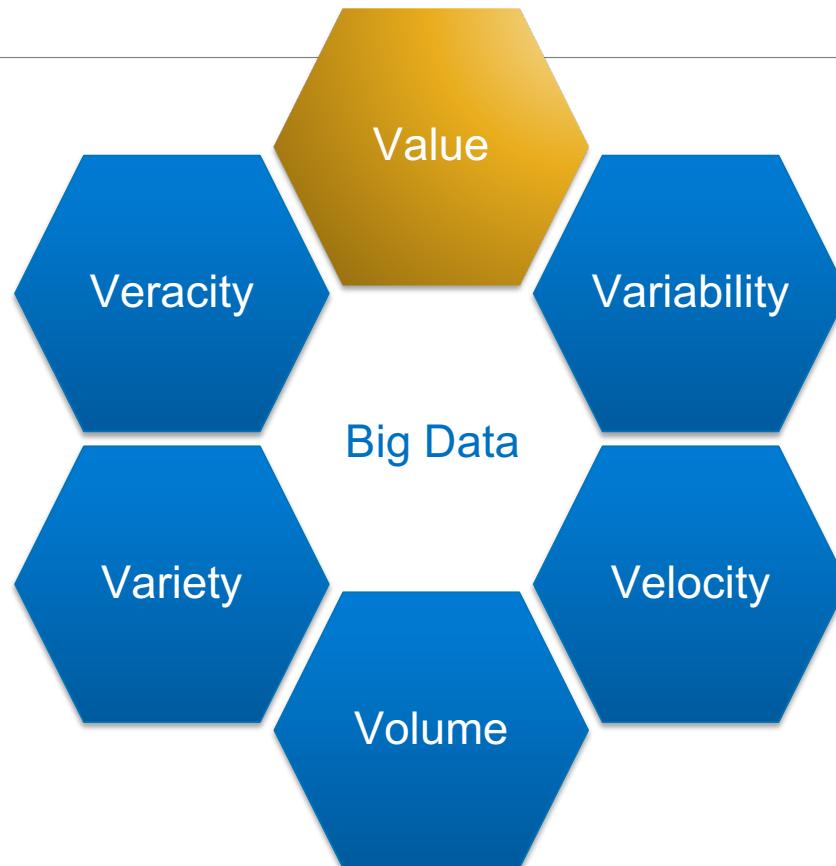
Datenquellen

- > Interne Geschäftsdaten
- > Web 2.0 / Soziale Medien
- > Öffentliche Datenquellen
- > Mobile / Web-Daten
- > Rich Media (Bild, Video, Audio)
- > Sensordaten
- > M2M-Kommunikation

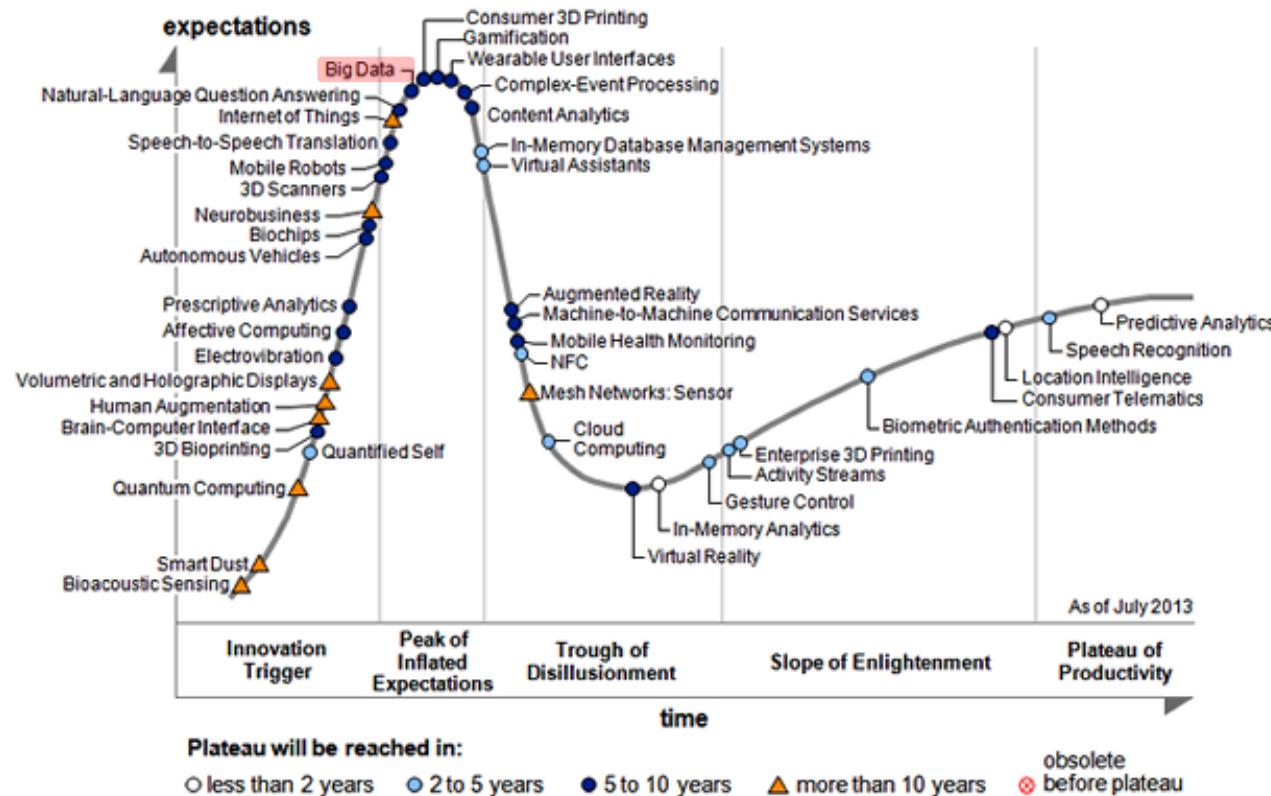
→ Big Data



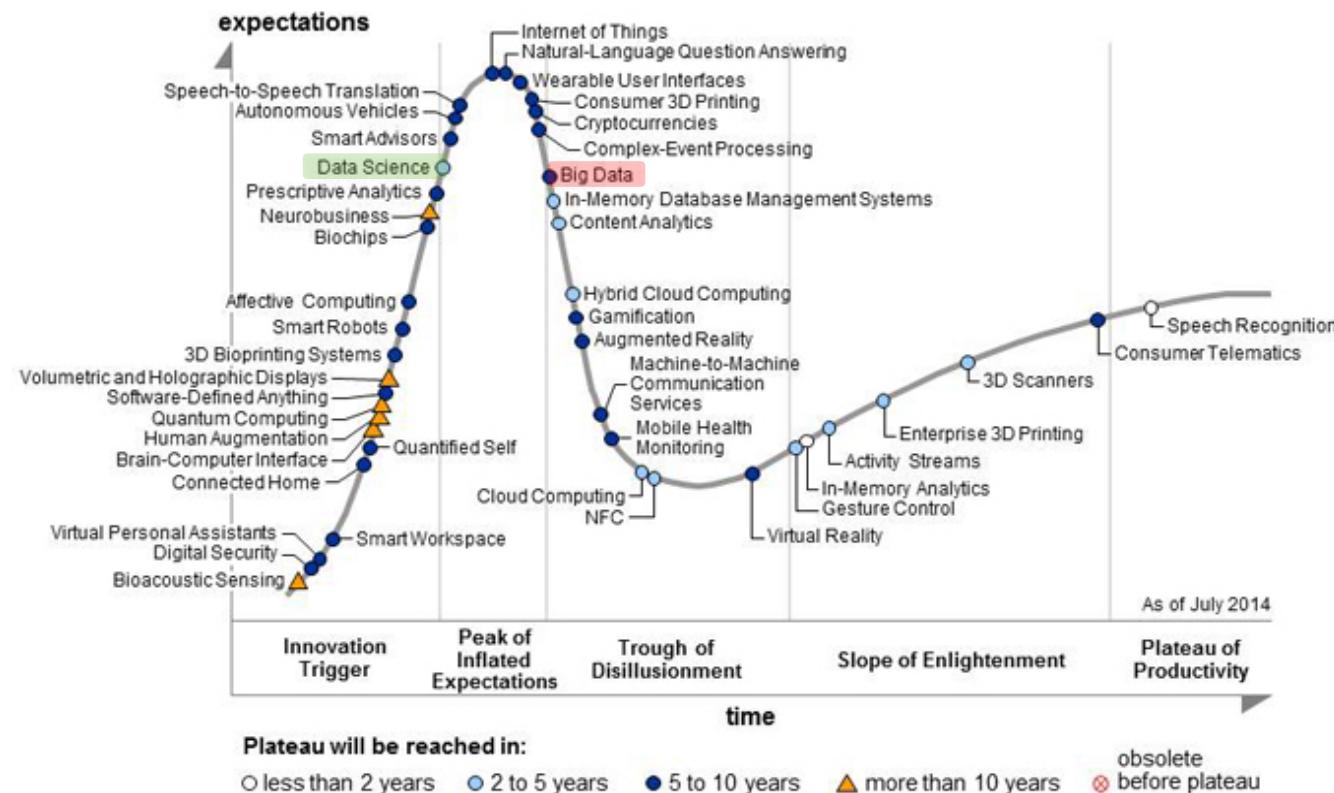
Big Data Eigenschaften



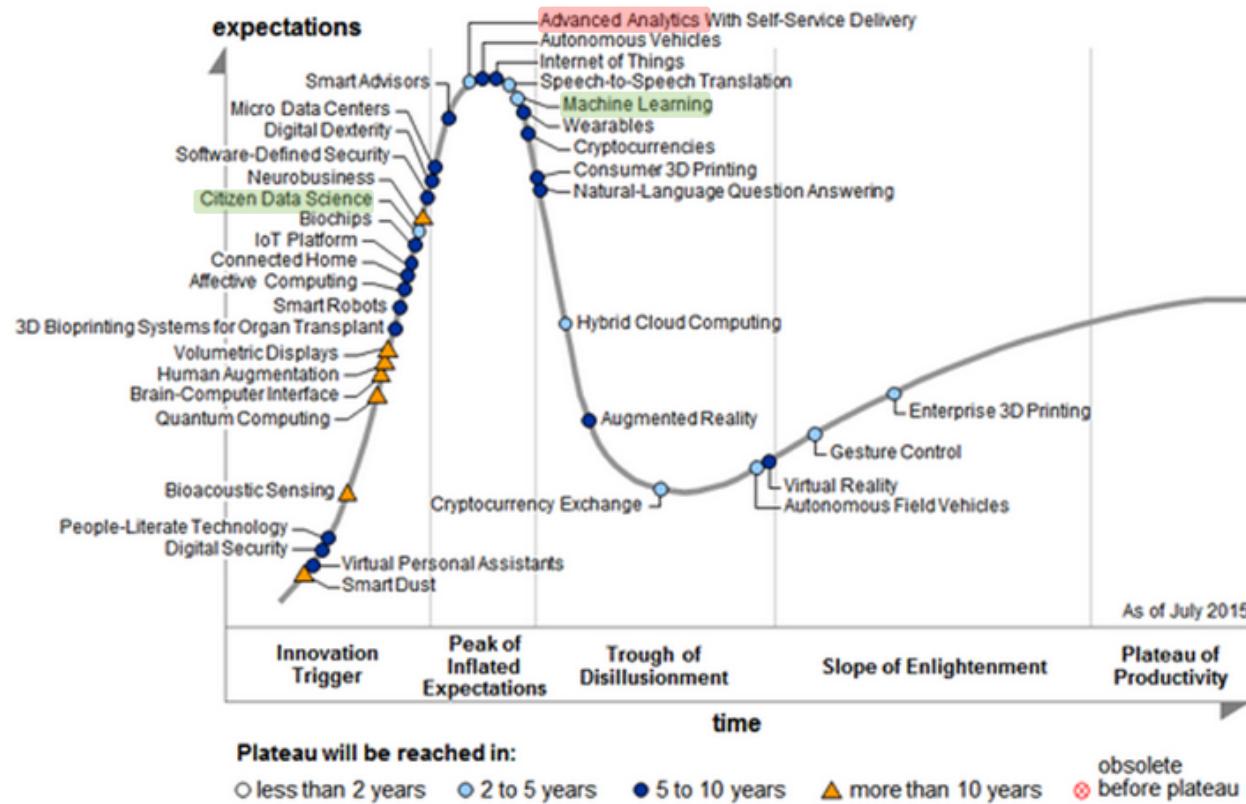
Big Data – Hype oder Realität? (2013)



Big Data – Hype oder Realität? (2014)



Big Data – Hype oder Realität? (2015)



- > Big Data als **Megatrend** bedeutend in vielen Bereichen unseres Lebens, daher aus Hype Cycle entfernt
- > Anteile von Big Data z.B. in
 - Machine Learning
 - Advanced Analytics
 - Citizen Data Science

Daten – Information – Wissen (Perspektive der Wirtschaftsinformatik)

Daten

Zum Zweck der Verarbeitung zusammengefasste Zeichen, die aufgrund bekannter oder unterstellter Abmachungen Informationen (d.h. Angaben über Sachverhalte und Vorgänge) darstellen.

Daten – Information – Wissen (Perspektive der Wirtschaftsinformatik)

Entscheidung

Pragmatik / Vernetzung

Semantik / Kontext

Syntax



Analysis vs. Analytics

Begriffe werden oft synonym verwendet, aber genau genommen ...

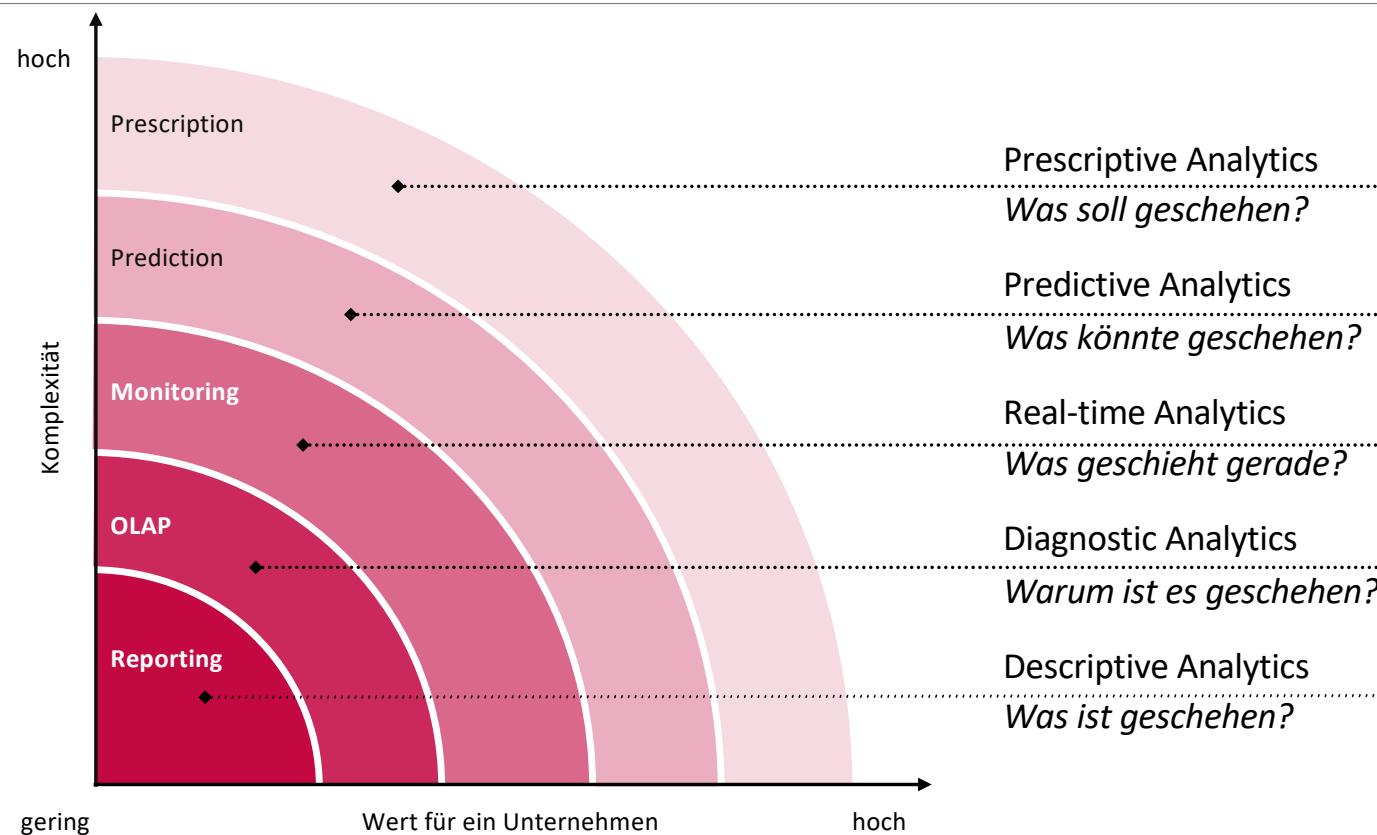
Analysis: Konkrete (individuelle) Analyse von Dingen (Anwendungsfall)

Data Analysis: Analyse von Daten

Analytics: Prozess und Methoden für die Durchführung von Analysen (Disziplin)

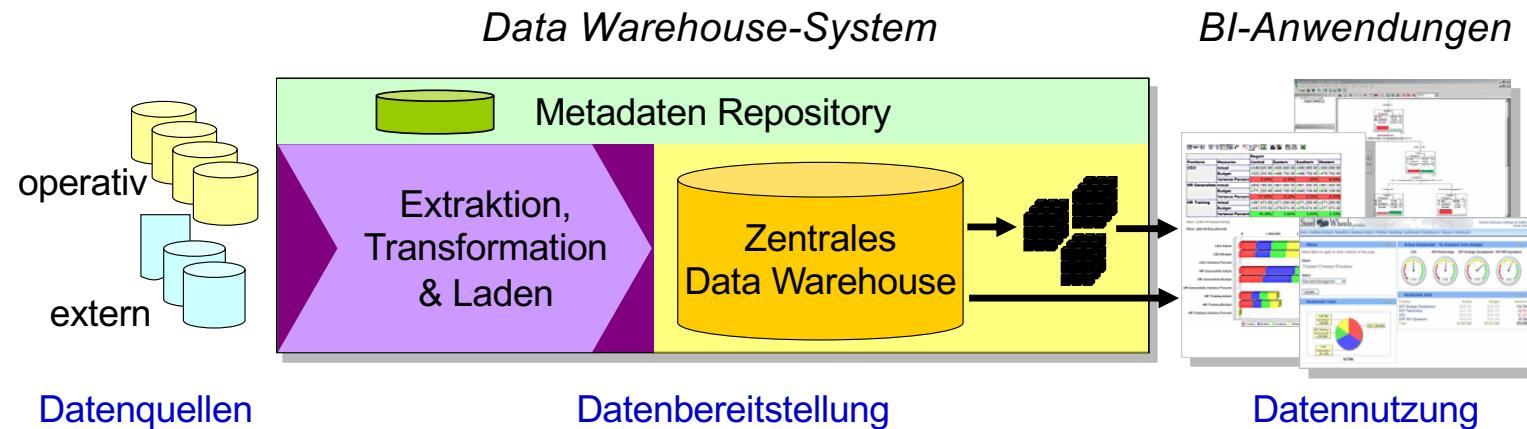
Data Analytics: Prozess und Methoden für die Durchführung von Datenanalysen

Analysespektrum



Business Intelligence: Ziel und Architektur

Umwandlung von Daten in Erkenntnisse, die bezüglich der Unternehmensziele bessere operative, taktische oder strategische Entscheidungen ermöglichen.



Big Data Verarbeitungsschritte



Datenquellen

Eigenschaften

Volumen
Geschwindigkeit
Vielfalt
Heterogenität
Unsicherheit
Glaubwürdigkeit

Big Data Management

Anforderung

Skalierbare
Echtzeitverarbeitung
komplexer Daten

Big Data Analytics

Anforderung

Skalierbare Analysen
und Visualisierungen
komplexer Daten

Ergebnisse

Zu beachten

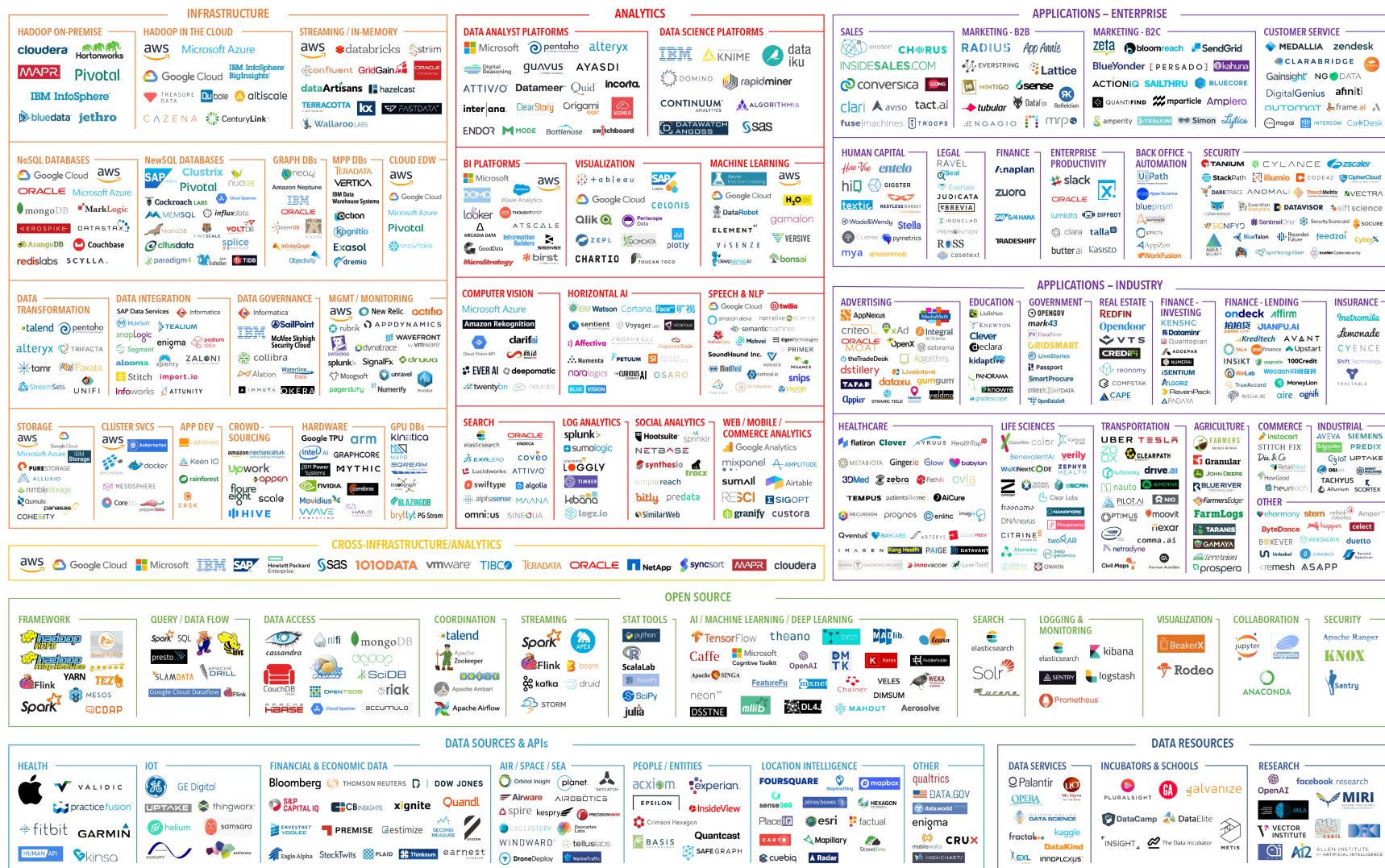
Unsicherheit
Effekt oder Zufall
Kausalität

Big Data Governance

Anforderung

Ermöglichung einer einfachen, korrekten und
rechtlich zulässigen Verwendung der Daten

BIG DATA & AI LANDSCAPE 2018



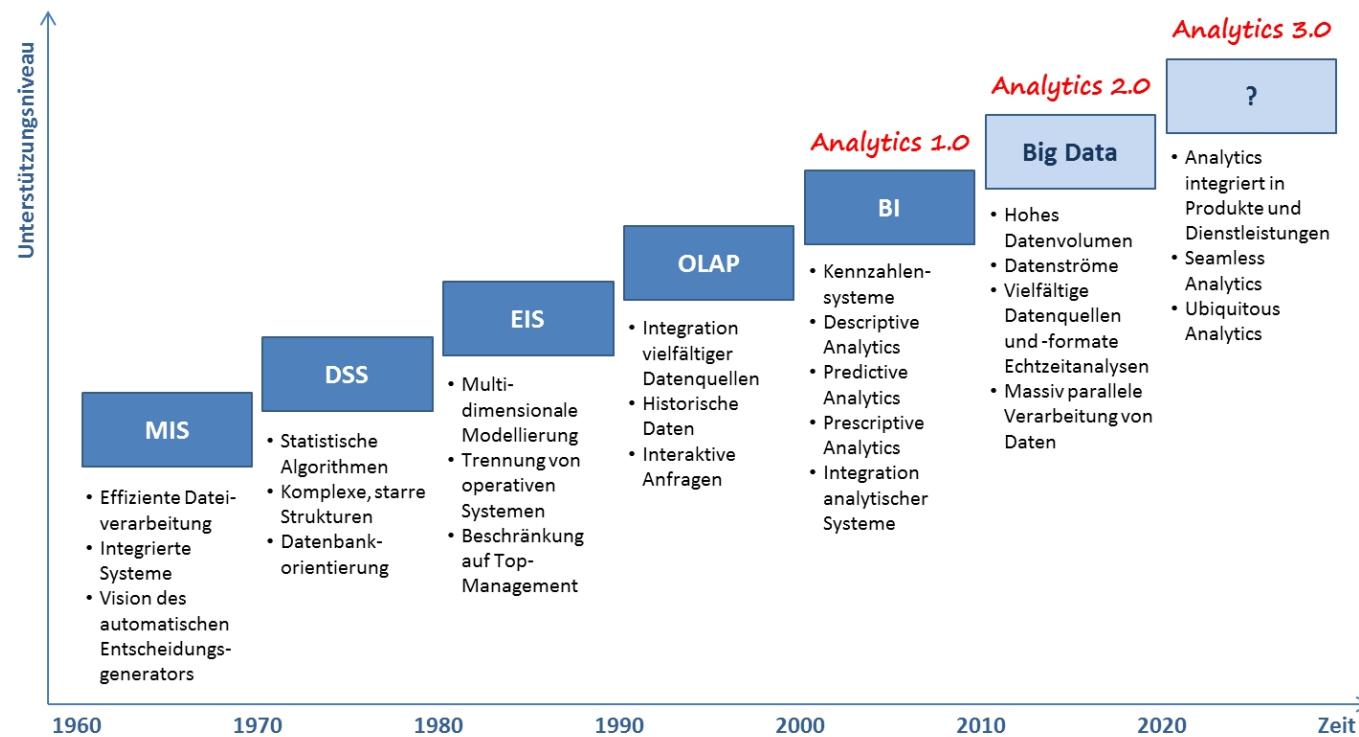
Final 2018 version, updated 07/15/2018

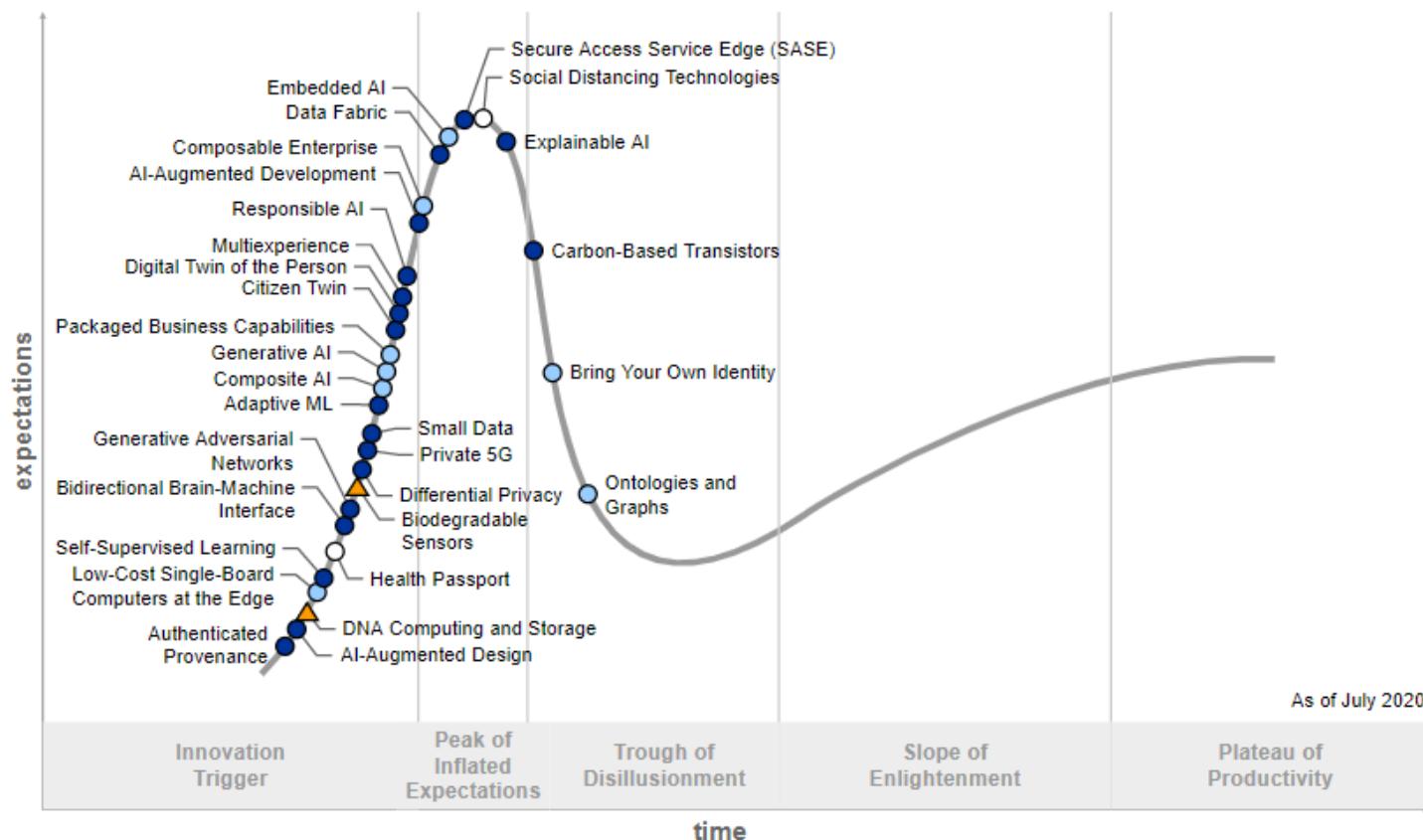
© Matt Turck (@mattturck), Demi Obavomi (@demi_ obavomi), & FirstMark (@firstmarkcap)

mattturck.com/bigdata2018

FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Evolution der Entscheidungsunterstützung





Plateau will be reached:

- less than 2 years
- 2 to 5 years
- 5 to 10 years
- ▲ more than 10 years
- ✖ obsolete before plateau

Lernen aus Daten

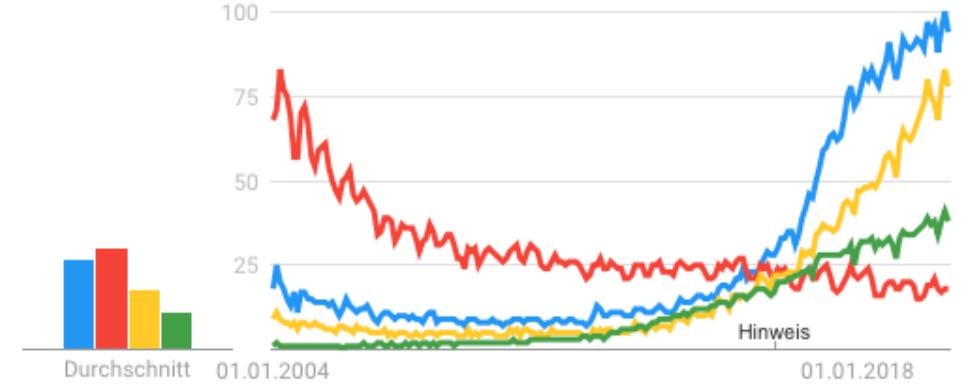
Disziplinen mit Fokus „*Datenanalyse*“:

- > (Applied) Statistics (1830 -)
- > Pattern Recognition (1955 -)
- > **Machine Learning** (1959 -)
- > **Data Mining** (1980 -)
- > Knowledge Discovery in Data (1989 -)
- > Business Analytics (1997 -)
- > Predictive Analytics (2002-)
- > Visual Analytics (2004-)
- > **Data Analytics** (2011-)
- > **Data Science** (2011 -)

Interesse im zeitlichen Verlauf

Google Trends

● machine learning ● data mining ● data science ● data analytics



Weltweit. 01.01.04 bis 12.03.20. Websuche.

Was sind Gemeinsamkeiten bzw. Unterschiede? → Motivation, Ziele, Hintergrund, Kontext, ...

Was ist Data Mining?

- > Unter **Data Mining** versteht man die systematische Anwendung statistischer Methoden auf große Datenbestände mit dem Ziel, neue Querverbindungen und Trends zu erkennen.

Quelle: de.wikipedia.org/wiki/Data-Mining

- > **Data Mining** ist ein Prozess zur Identifizierung von gültigem, neuen, potentiell nützlichem und letztlich verständlichem Wissen aus Daten.

Quelle: Fayyad, Piatetsky-Shapiro, Smyth (1996): From Data Mining to Knowledge Discovery: An Overview. In Advances In Knowledge Discovery and Data Mining.

- > **Data Mining** ist die Analyse meist großer Datenbestände durch automatische oder halbautomatische Methoden, um aussagefähige Muster oder Modelle zu finden.
- > Anfangs nur ein Schritt beim **Knowledge Discovery in Databases (KDD)**, inzwischen jedoch meist **synonym** dazu verwendet.

Was ist Visual Analytics?

"Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces."

Quelle: James J. Thomas und Kristin A. Cook

"Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets."

Quelle: Keim et al.

Visual Analytics soll aus großen und komplexen Datenbeständen Erkenntnisse gewinnen. Durch die Kombination automatischer Datenanalyse und visueller Interaktion soll die Fähigkeit des Menschen ausgenutzt werden, Muster und Trends visuell schnell erfassen zu können.

Beispiel: Anscombe Datensätze – analytisch betrachtet

A		B		C		D	
X	Y	X	Y	X	Y	X	Y
10,00	8,04	10,00	9,14	10,00	7,46	8,00	6,58
8,00	6,95	8,00	8,14	8,00	6,77	8,00	5,76
13,00	7,58	13,00	8,74	13,00	12,74	8,00	7,71
9,00	8,81	9,00	8,77	9,00	7,11	8,00	8,84
11,00	8,33	11,00	9,26	11,00	7,81	8,00	8,47
14,00	9,96	14,00	8,10	14,00	8,84	8,00	7,04
6,00	7,24	6,00	6,13	6,00	6,08	8,00	5,25
4,00	4,26	4,00	3,10	4,00	5,39	19,00	12,50
12,00	10,84	12,00	9,13	12,00	8,15	8,00	5,56
7,00	4,82	7,00	7,26	7,00	6,42	8,00	7,91
5,00	5,68	5,00	4,74	5,00	5,73	8,00	6,89



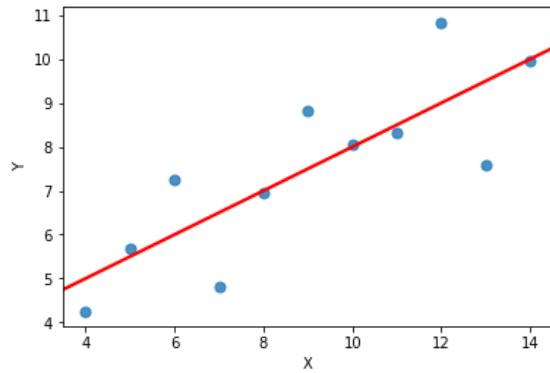
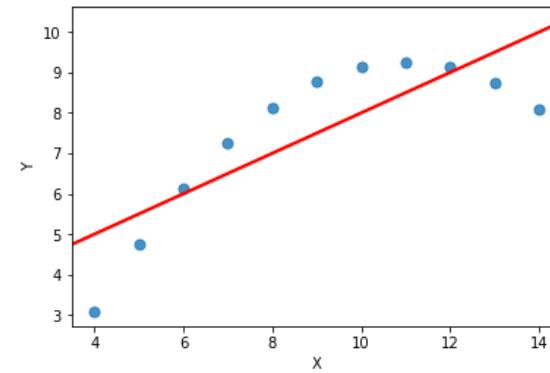
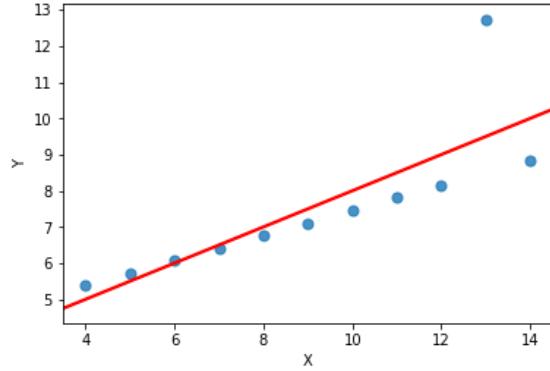
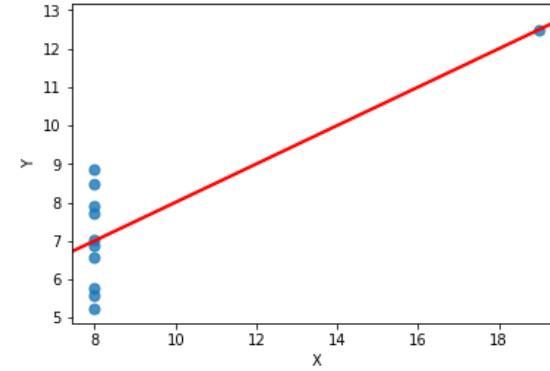
	A	B	C	D
Mittelwert X	9,000	9,000	9,000	9,000
Mittelwert Y	7,501	7,501	7,500	7,501
Varianz X	11,000	11,000	11,000	11,000
Varianz Y	4,127	4,128	4,123	4,123
Kovarianz	5,501	5,500	5,497	5,499
Korrelation	0,816	0,816	0,816	0,817



$$\hat{y} = f(x) = 3 + 0,5x$$

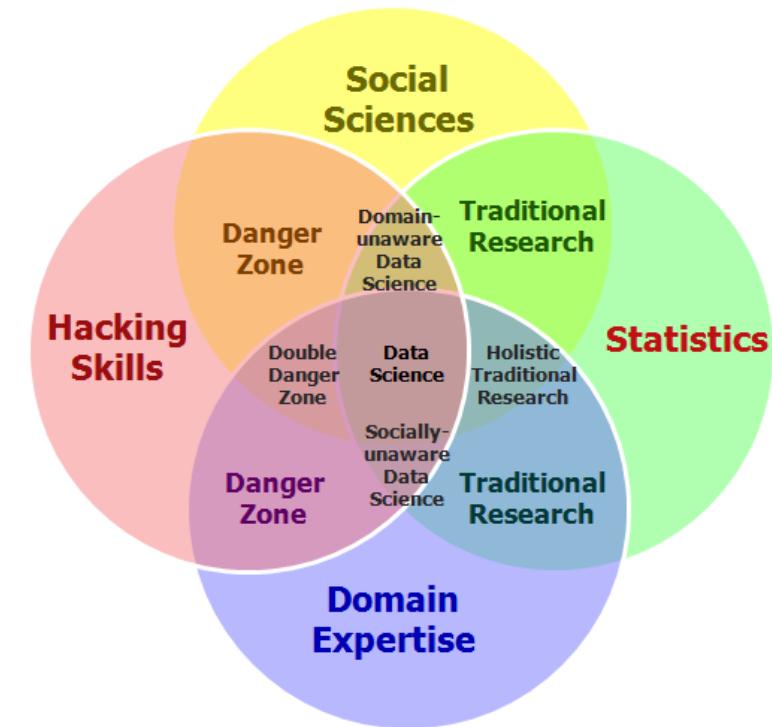
Dieselbe Regressionsgerade bei allen Datensätzen!

Beispiel: Anscombe Datensätze – visuell betrachtet

A**B****C****D**

Was ist Data Science?

- > **Fragen mit Hilfe von Daten beantworten**
- > Erkenntnisse aus Daten gewinnen
- > **Data Science** is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured.
- > It is a continuation of some of the data analysis fields such as statistics, machine learning, data mining, and predictive analytics, similar to Knowledge Discovery in Databases (KDD).



Quelle: datascienceassn.org/content/fourth-bubble-data-science-venn-diagram-social-sciences

Data Science im weiteren Sinne

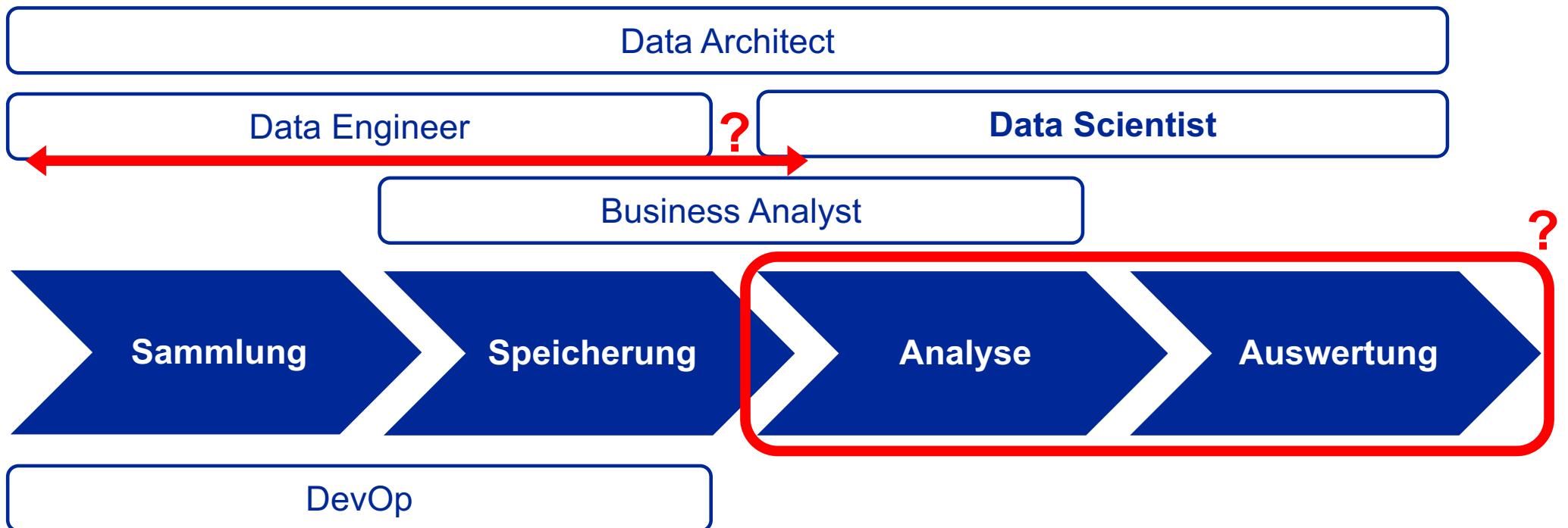
Data science includes **data extraction, data preparation**, data exploration, data transformation, **storage and retrieval, computing infrastructures**, various types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects.

Quelle: Van der Aalst (2016): Process Mining, S. 10.

The term data science is often used to include all areas to do with the **capturing and processing of data including cleaning, warehousing, and converting unstructured data to structured data**—before its subsequent analysis and output of results.

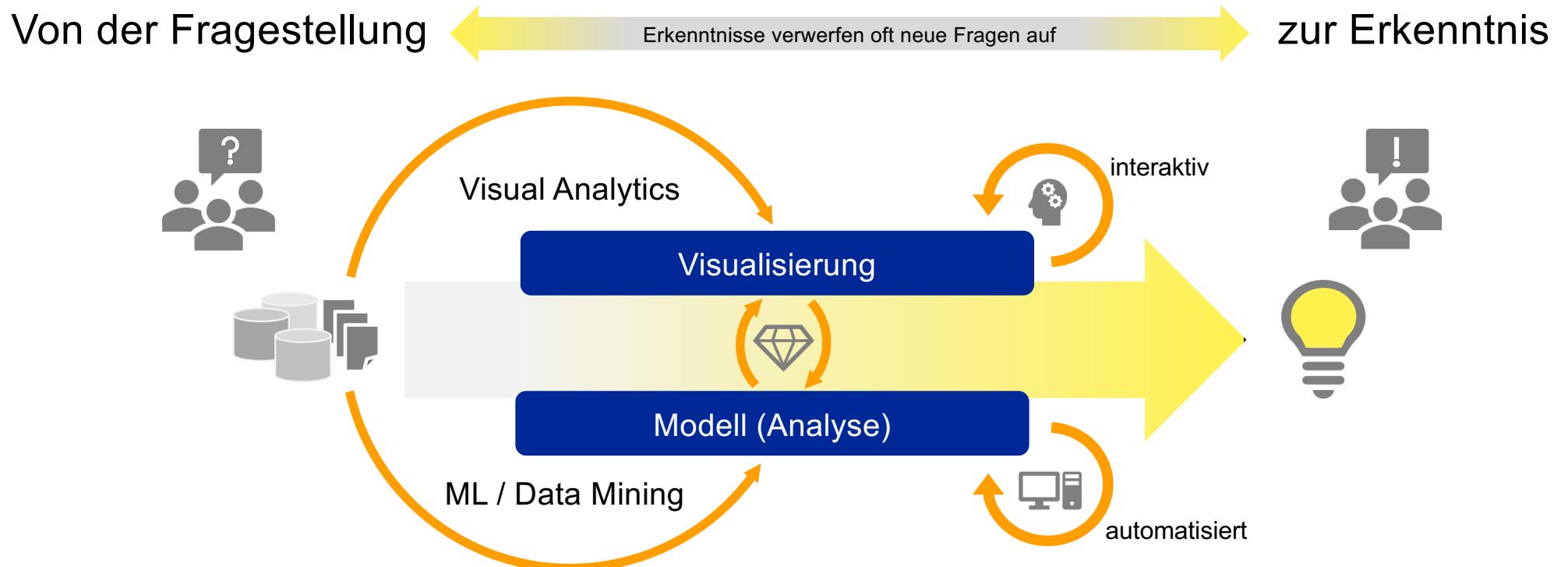
Quelle: Earnshaw (2019): Data Science and Visual Computing, S. 3.

Data Science und angrenzende Aufgabenprofile



Quelle: In Anlehnung an Papp et al. (2019): Handbuch Data Science, S. 10.

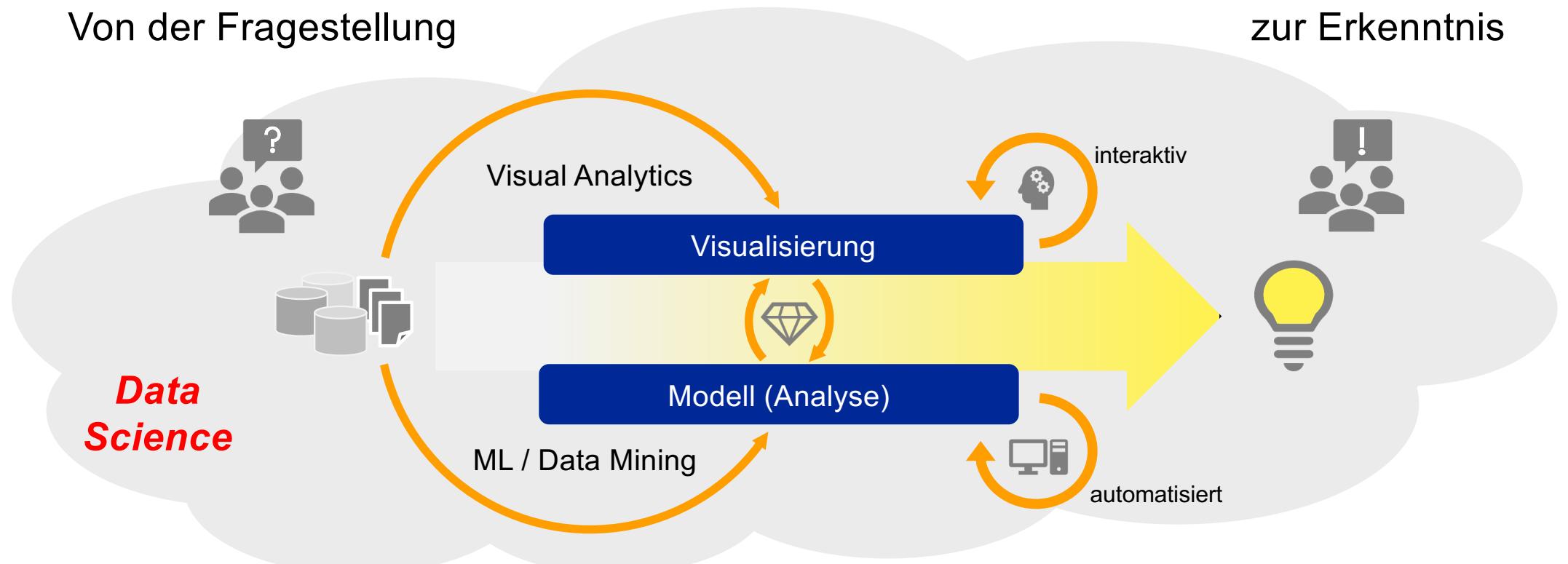
Erkenntnisse aus Daten gewinnen



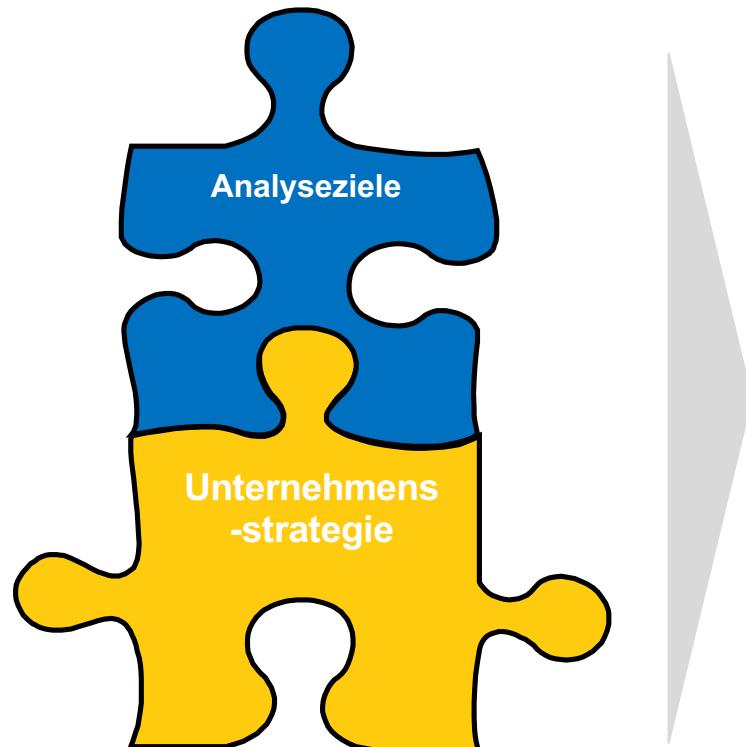
Data Science als Oberbegriff

Von der Fragestellung

zur Erkenntnis



Analyseziele und Unternehmensstrategie



- > Produkte/Dienstleistungen verbessern
- > Sicherheit erhöhen
- > Geschäftsprozesse optimieren
- > Kosten senken
- > Kunden besser verstehen und ansprechen
- > Betrugsfälle aufdecken
- > **Neue Geschäftsideen entwickeln**

Erfolgsfaktoren für Data-Science-Vorhaben

Das intelligente Unternehmen braucht mehr als **Technologie**

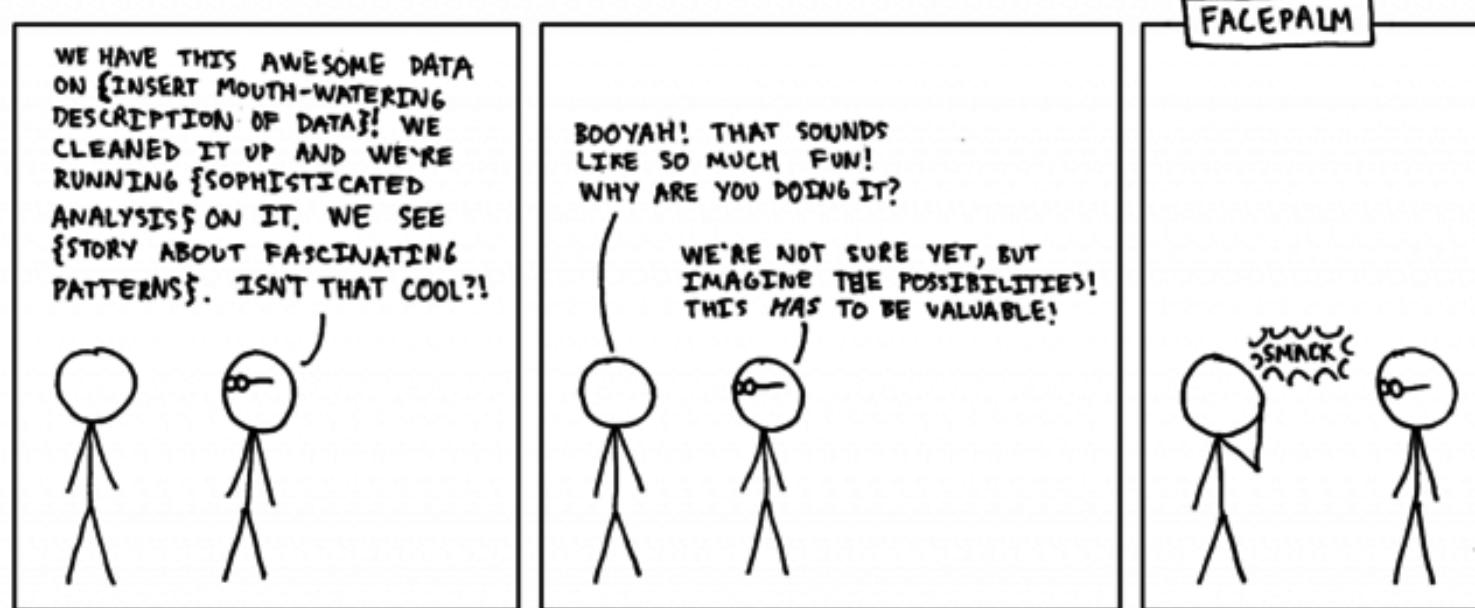
- > Verfügbarkeit relevanter Daten
- > Integration interner und externer Datenquellen
- > Data Governance

- > **Offene Unternehmenskultur**
- > **Fähiger Mitarbeiter/-innen**



→ Parallel Veranstaltung „Data Science Management“

Lasst die Daten sprechen ...



Aber: Die Frage ist wichtiger als die Antwort!

Die richtigen Fragen stellen

- > Bei der Datenanalyse ist die Fragestellung das wichtigste!
- > An zweiter Stelle kommen die Daten
- > Die verfügbaren Daten sind zwar entscheidend dafür, was beantwortet werden kann ...
- > ... aber die Verfügbarkeit von Daten erspart dennoch nicht das Stellen der Frage!

Quelle: Jeff Leek: The Elements of Data Analytic Style

- > Elementare Fähigkeiten:
 - > Die richtigen fachlichen Fragen stellen
 - > Fachliche Fragen abbilden auf analytische Fragen, die sich mit verfügbaren Daten beantworten lassen
- > **Diskussion: Welche Arten von Fragen lassen sich beantworten?**

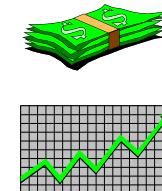
Klassische Analyse-Aufgaben

Vorhersagen

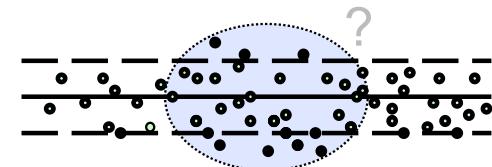
Klassifikation



Regression (Numerische Prognose)



Anomalieerkennung



Konzeptbeschreibung

Typischer
Mercedes-Kunde ?

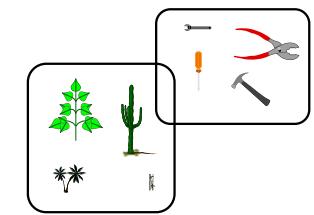


Assoziationsanalyse

Mit einer Wahrscheinlichkeit
von 90% werden
Kaffee und Milch
zusammen gekauft



Clusteranalyse



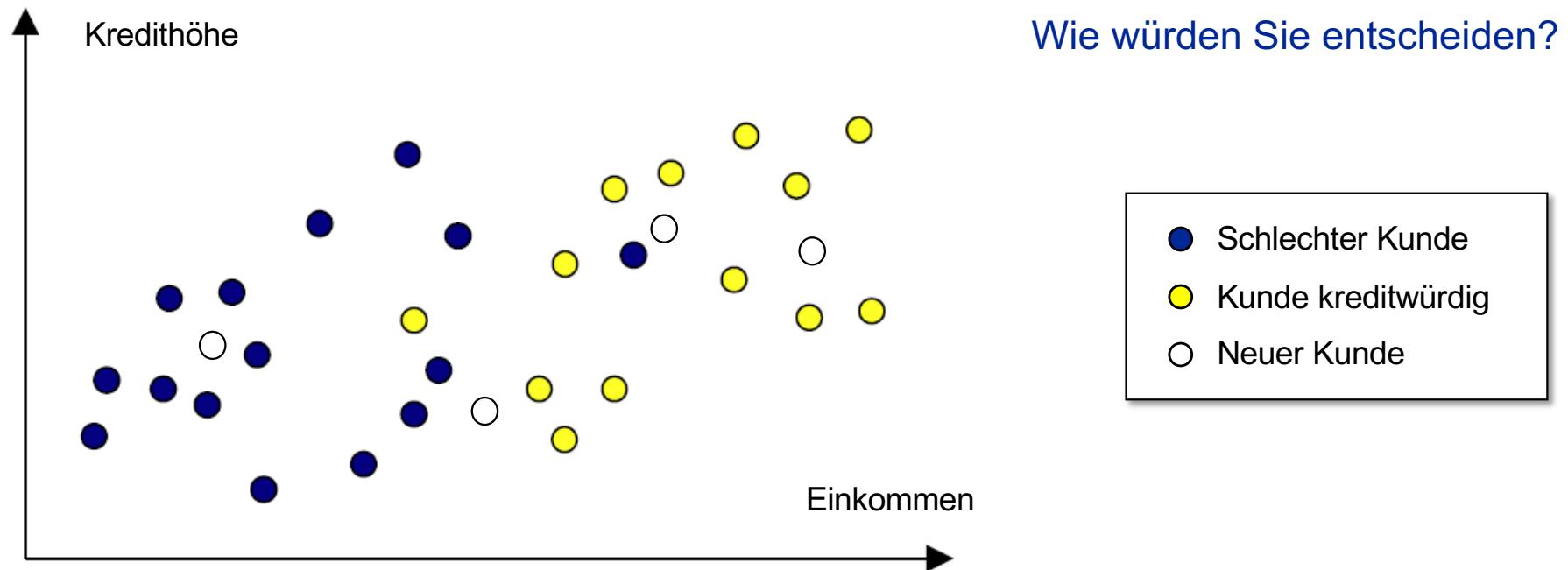
Beschreibungen

Beispiele für Analyse-Aufgaben

-
- > Anomalieerkennung Ist eine Kreditkartentransaktion legitim oder Betrug?
 - > Assoziationsanalyse Welche Waren werden häufig zusammen gekauft?
 - > Clusteranalyse Gibt es Kundengruppen mit ähnlichem Verhalten?
 - > Klassifikation Ist ein neuer Kunde kreditwürdig oder nicht?
 - > Konzeptbeschreibung Was kennzeichnet einen Kunden der kündigt?
 - > Regression Wie viel Umsatz machen meine Kunden?

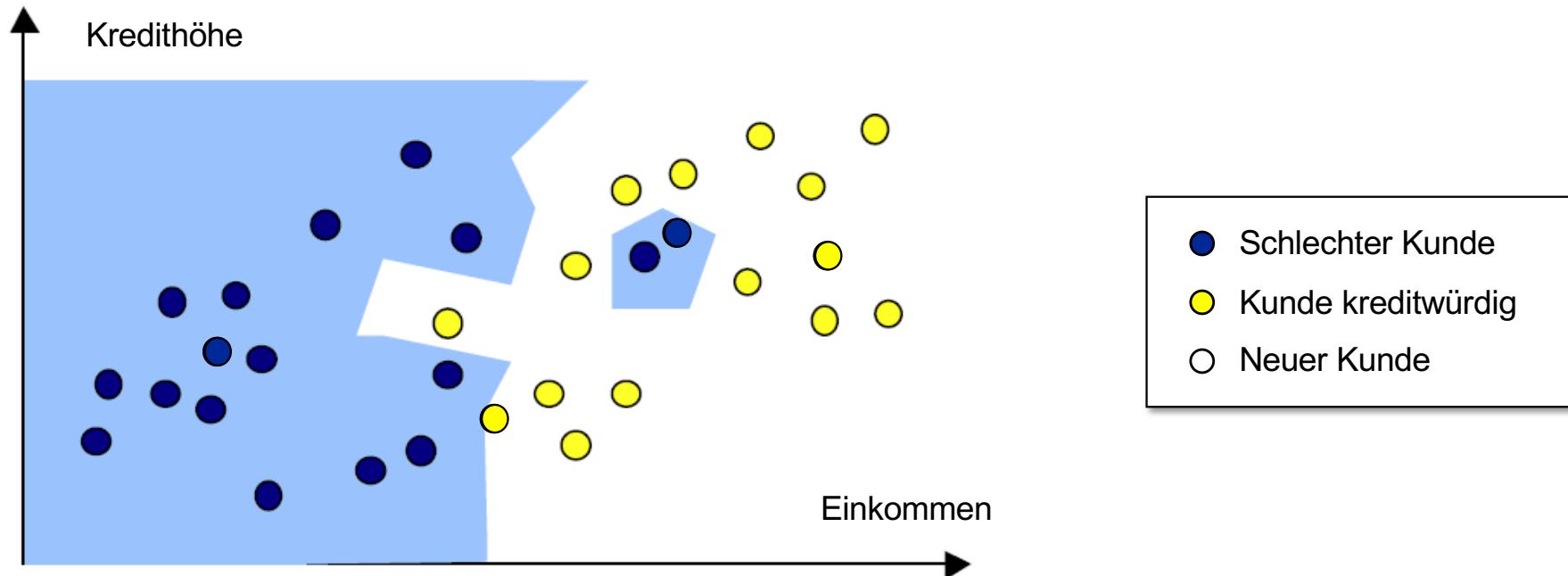
Klassifikation als Aufgabe

Ziel: Zuordnung von Objekten zu bekannten Klassen anhand von beobachtbaren Merkmalen



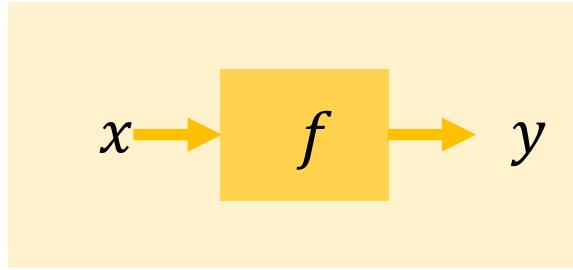
1-NN: Klassifikation mit Hilfe des nächsten Nachbarn

Einfache Lösung: Ordne neue Objekte der Klasse zu, die der nächste Nachbar hat!

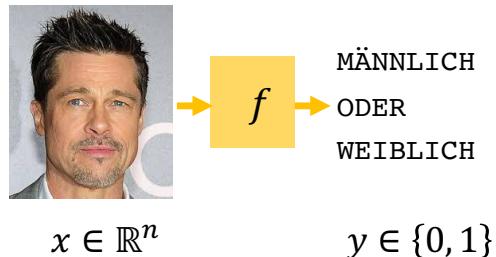


Wie löst man ein Problem in der Informatik?

Eingabevektor
zur Beschreibung
des Problems



Ausgabevektor
zur Beschreibung
der Lösung



Optimal Sample Complexity of M -wise Data for Top- K Ranking

Algorithm 1 Rank Centrality (Nugueline et al., 2022)

Input: the collection of statistics $\alpha = \{\alpha_{\mathcal{T}}, \mathcal{T} \in \mathcal{G}^{(M)}\}$.
Convert the M -wise sample for each hyper-edge \mathcal{T} into
3. If $\alpha_{\mathcal{T}} < 0$, then

1. Choose a circular permutation of the items in \mathcal{T} uniformly at random;
2. Break it into M pairs of adjacent items, and determine the set of pairs by $\pi(\mathcal{T})$;
3. Use the pairwise data of $\pi(\mathcal{T})$.

Compute the transition matrix $P = (P_{ij})_{i,j \in [M]}$:

$$P_{ij} = \begin{cases} \frac{d_{max}}{d_{max} - d_{min}} P_{\mathcal{T}_j} & i \neq j \\ 0 & \text{otherwise,} \end{cases}$$

where d_{max} is the maximum out-degree of vertices in \mathcal{T} . Output the stationary distribution of matrix P :

$$\hat{w} := \sum_{\mathcal{T} \in \mathcal{G}^{(M)}} \frac{1}{M} \sum_{i,j \in \mathcal{T}} \frac{\alpha_{\mathcal{T}}}{P_{ij}}. \quad (16)$$

$w_0 := w_{M+1} = 1$

f

$y \in \{0, 1\}$

Output: $\hat{w} = w_M = w_{M+1} = 1$ for ease of demonstration.

To outline the proof of Theorem 1, let us introduce Theorem 2 and Theorem 3 which lead to Theorem 1.

Theorem 3. When Rank Centrality is explored, with high probability, the ℓ_∞ norm estimation error is upper-bounded by

$$\frac{\|\hat{w} - w\|_\infty}{\|\hat{w}\|_\infty} \leq \sqrt{\frac{n \log n}{(2\pi)^2 M^2}} \sqrt{\frac{1}{M}}, \quad (18)$$

where $p \geq c(M-1) \sqrt{\frac{n \log n}{(2\pi)^2}}$ and c is some numerical constant.

Let $\|w\|_\infty = w_{M+1} = 1$ for ease of demonstration. Suppose $\Delta x = w\hat{x} - w\hat{w} \geq \sqrt{\frac{n \log n}{(2\pi)^2}} \sqrt{B}$. Then,

$$\hat{w}_i - \hat{w}_j \geq w_i - w_j = |w_i - w_j| \quad (19)$$

for all $1 \leq i \leq K$ and $j \geq K+1$. That is, the top- K items are identified correctly. However, as long as $\Delta x \geq \sqrt{\frac{n \log n}{(2\pi)^2}} \sqrt{B}$, i.e., $(2\pi)^2 p \geq \frac{n \log n}{B}$, we can ignore the effect of $\frac{n \log n}{B}$.

Now, let us prove Theorem 3. To find a ℓ_∞ cover bound, we first derive an upper bound on the point-wise error between the score estimate of item i and its true score, which corresponds to $\hat{w}_i - w_i$:

$$|\hat{w}_i - w_i| \leq |\hat{w}_i - w_K| P_{K,i} + \sum_{j \neq K} |\hat{w}_j - w_j| P_{j,i} \quad (20)$$

+ $\left| \sum_{j \neq K} (w_j - w_K) \left(P_{j,i} - P_{K,i} \right) \right|.$

This can be obtained applying $\hat{w} = P\hat{w}$ and $w = Pw$. We obtain upper bounds on these terms as follows.

$P_{K,i} < 1, \quad (21)$

$$\left| \sum_{j \neq K} (w_j - w_K) \left(P_{j,i} - P_{K,i} \right) \right| \leq \sqrt{\frac{n \log n}{(2\pi)^2} M^2} \sqrt{\frac{1}{M^2}}, \quad (22)$$

$$\sum_{j \neq K} |\hat{w}_j - w_j| P_{j,i} \leq \sqrt{\frac{n \log n}{(2\pi)^2} M^2} \sqrt{\frac{1}{M^2}}, \quad (23)$$

with high probability (see Lemmas 1, 2 and 3 in the supplementary for details). One can see that the inequalities (21)

Abstract

Given a sample of instances with binary labels, the top ranking problem is to produce a ranked list of instances where the *head* of the list is dominated by positives. Popular existing approaches to this problem are based on surrogates to a performance measure known as the fraction of positives of the top (PTop). In this paper, we show that the measure and its surrogates have an undesirable property: for certain noisy distributions, it is optimal to trivially predict the *same score for all instances*. We propose a simple rectification of the measure which avoids such trivial solutions, while still focussing on the head of the ranked list and being as easy to optimise.

Was ist Machine Learning

Machine learning is the subfield of computer science that gives "computers the ability to learn without being explicitly programmed."

– Arthur Samuel (1959)

```
function GetMin(var a: TList)
var
  i, min, mini: integer;
begin
  min := MaxInt;
  mini := 0;
  for i := 1 to a.len do
    if a.arr[i].G < min then
      begin
        min := a.arr[i].G;
        mini := i;
      end;
  GetMin := mini;
end;
```

```
mann(adam).
mann(tobias).
mann(frank).
frau(eva).
frau(daniela).
frau(ulrike).
vater(adam,tobias).
vater(tobias,frank).
vater(tobias,ulrike).
mutter(eva,tobias).
mutter(daniela,frank).
mutter(daniela,ulrike).
```

```
# Spot Check Algorithms
models = []
models.append('LR', LogisticRegression)
models.append('LDA', LinearDiscriminantAnalysis)
models.append('KNN', KNeighborsClassifier)
models.append('CART', DecisionTreeClassifier)
models.append('NB', GaussianNB)
models.append('SVM', SVC)
# evaluate each model in turn
results = []
names = []
for name, model in models:
    kfold = model_selection.KFold(n_splits=10)
    cv_results = model_selection.cross_val_score(model, X, y, cv=kfold)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, np.mean(cv_results), np.std(cv_results))
    print(msg)
```



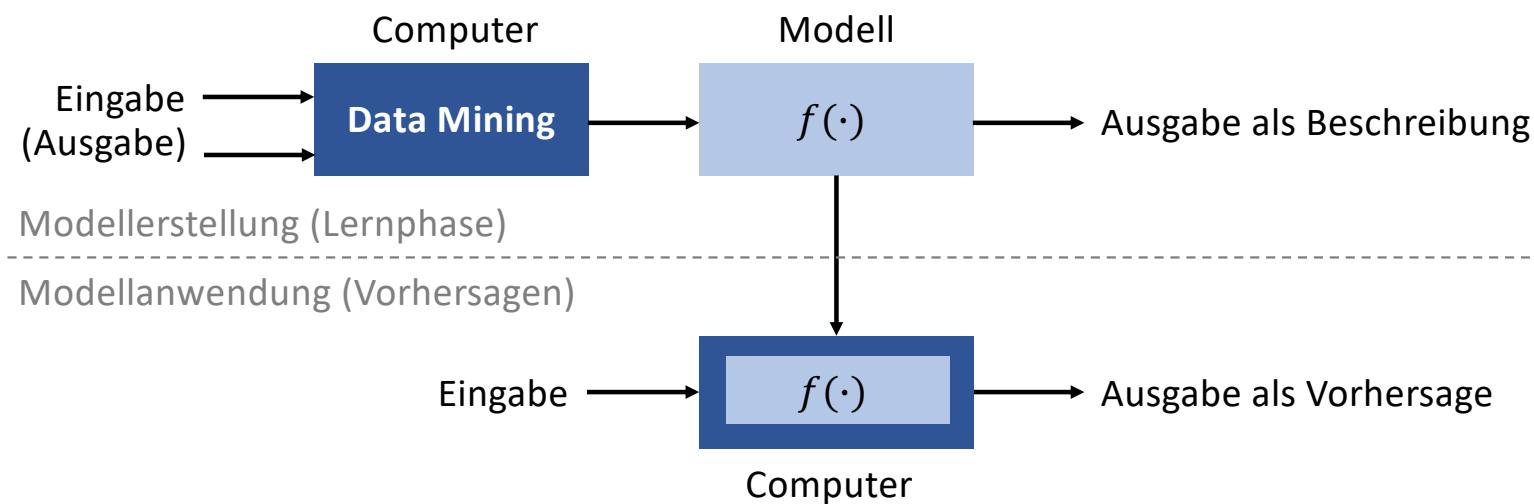
*classical
programming*
... schwierig
für komplexe
Probleme

*knowledge-based
programming*
... aufwendige
Wissens-
beschaffung

*“implicit”
programming*

Quelle: Eyke Hüllermeier, 2018

Modellerstellung vs. Modellanwendung

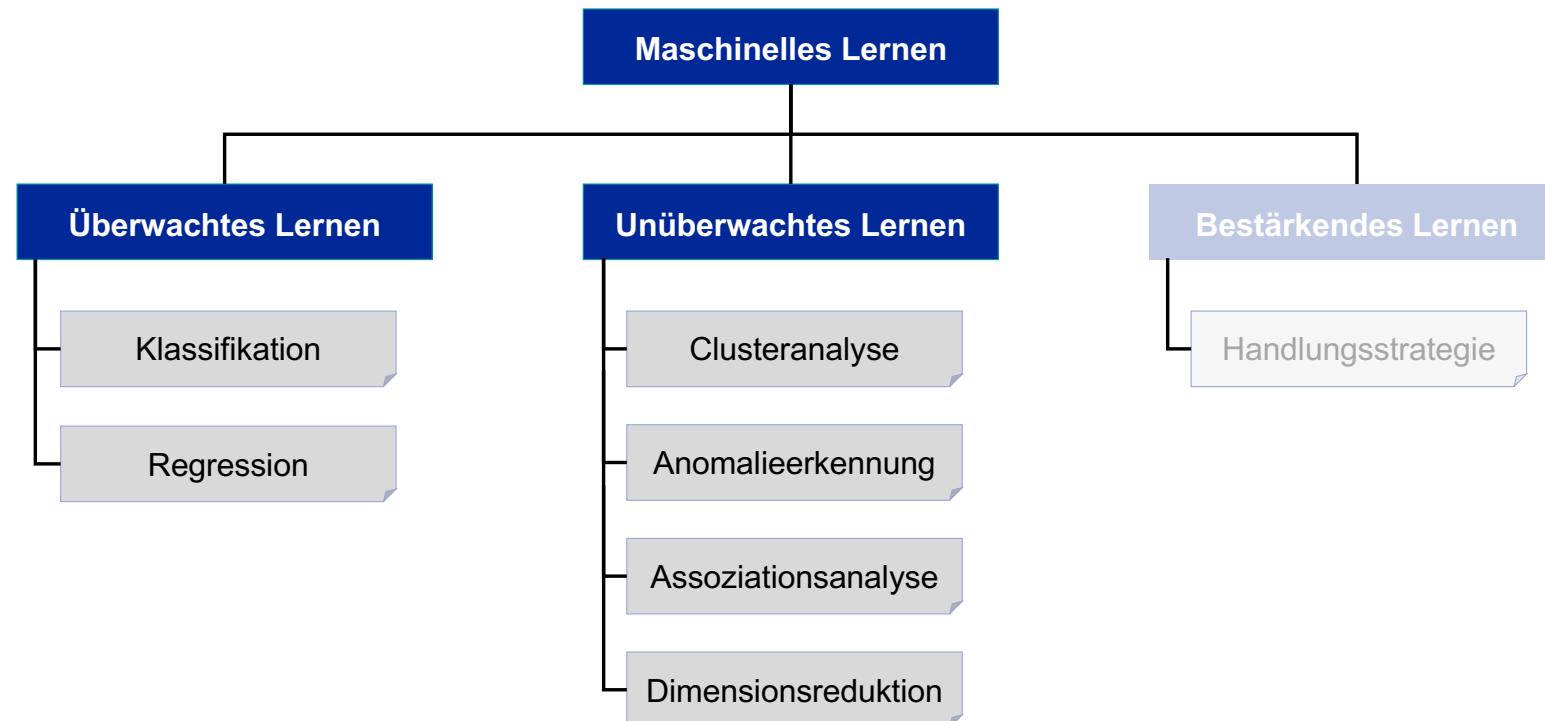


Lernformen beim Machine Learning

- > Überwachtes Lernen (Supervised Learning)
 - **Generalisieren basierend auf Beispielen**
- > Unüberwachtes Lernen (Unsupervised Learning)
 - **Strukturentdeckend ohne Vorgaben/Feedback**
- > Bestärkendes Lernen (Reinforcement Learning)
 - **Lernen durch Versuch und Irrtum**



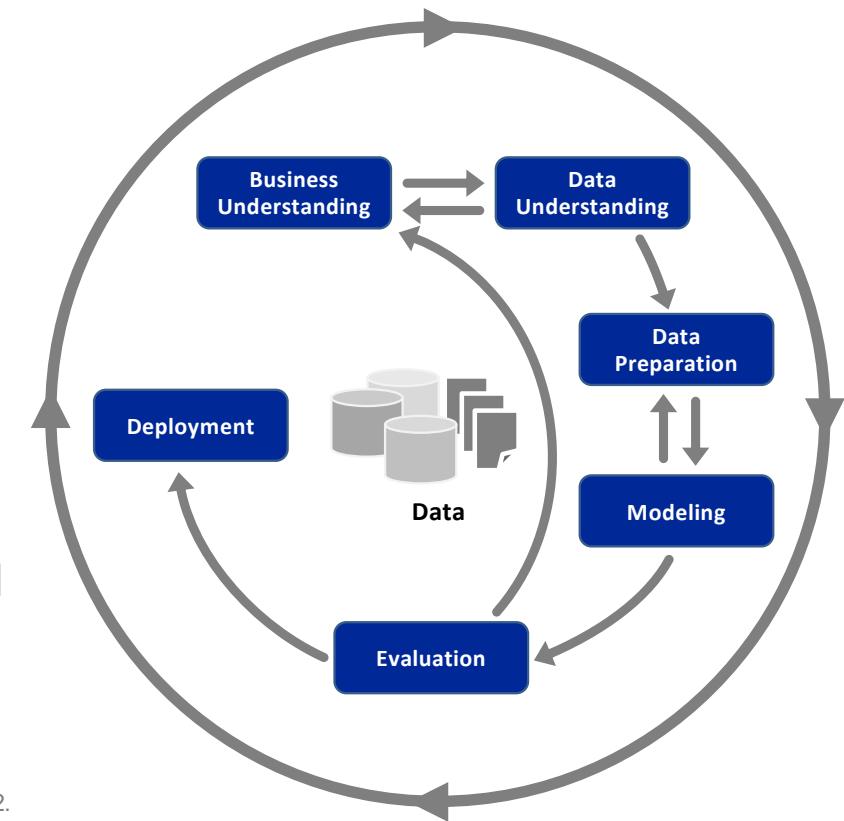
Lernformen und Data-Mining-Aufgaben



CRISP-DM: Ein Prozessmodell für Analyseprojekte

CRoss Industry Standard Process for Data Mining

- > Sehr weit verbreitetes Referenzmodell für die Durchführung von Data-Mining-Projekten
- > Ermöglicht eine strukturierte, effektive und gut kommunizierbare Durchführung von Analyseprojekten
- > Besteht aus 6 Phasen mit jeweils mehreren Aufgaben
- > Betont den iterativen Charakter mit beliebigen Sprüngen zwischen den Phasen, die sich aufgrund von Fragen und Erkenntnissen in den einzelnen Phasen ergeben können



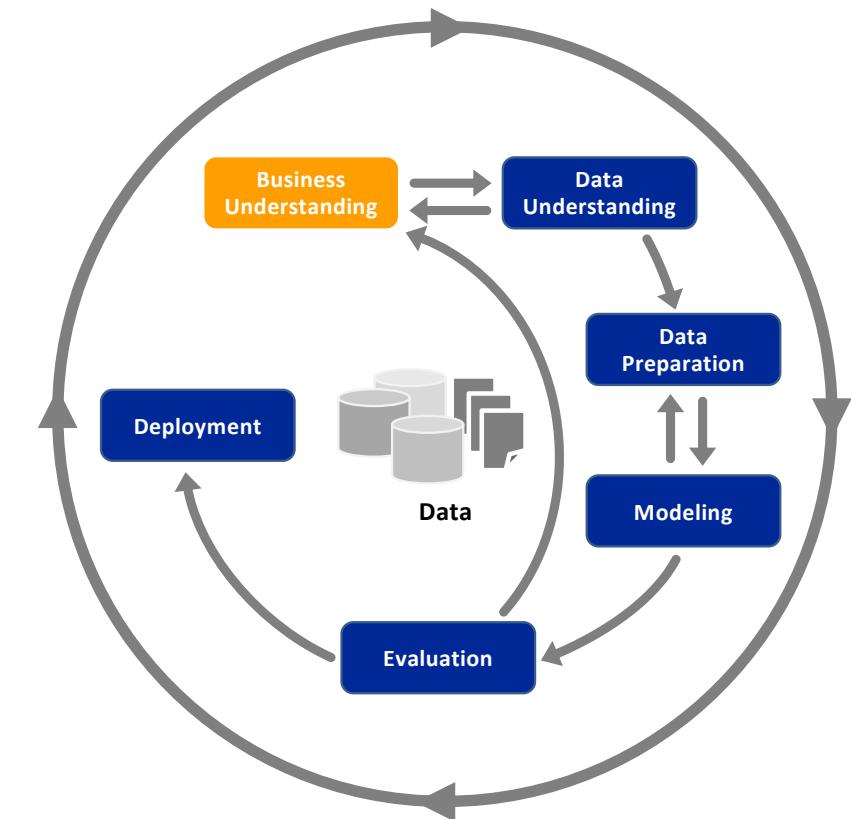
Shearer C., The CRISP-DM Model: The new blueprint for data mining, Journal of Data Warehousing (2000); 5:13—22.

CRISP-DM Phase 1: Business Understanding

Ziel: Festlegen Aufgabenstellung und Erfolgskriterien

Aufgaben

- > Verstehen des Hintergrunds der Anwendung
- > Einbinden von Vorwissen (Experten, bestehende Lösungen)
- > Definition der fachlichen Aufgabenstellung (Projektziel) im Kontext der Anwendung (Business)
- > Abbildung des Projektziels auf eine Analyseaufgabe
- > Festlegen von Erfolgskriterien (Erwartungshaltung)
- > Projektmanagement



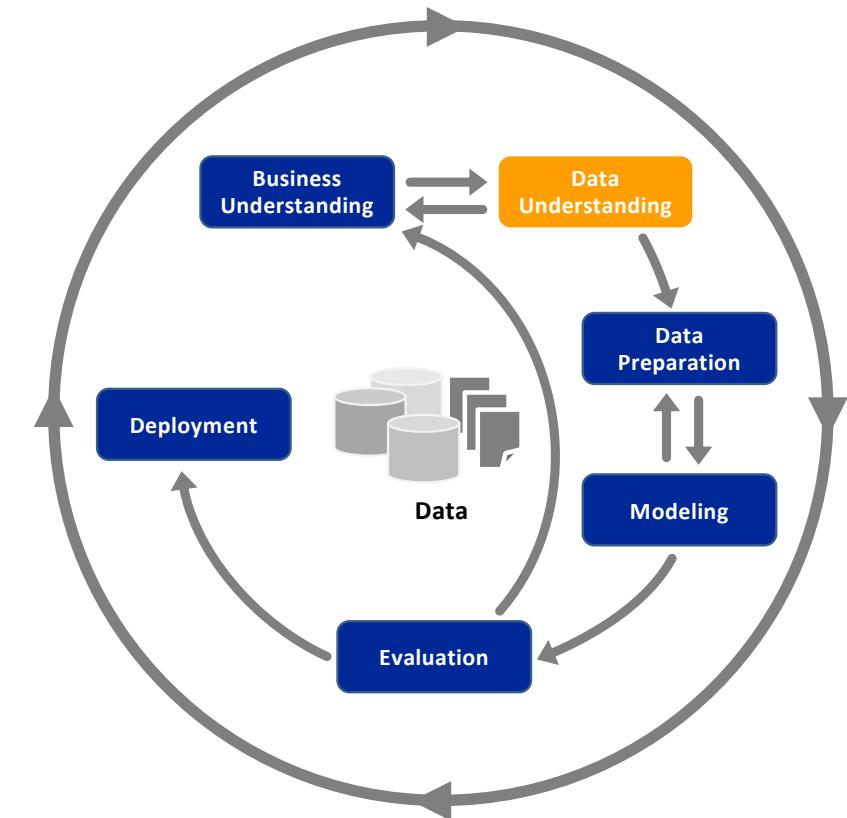
CRISP-DM Phase 2: Data Understanding

Aufgaben

- > Sammlung von Daten
- > Erkunden der Daten (Explorative Datenanalyse)
- > Bewertung der Datenqualität

Erkenntnisse

- > Datenqualität ist sehr wichtig: *Garbage in, garbage out!*
- > Big Data als wesentlicher Faktor für den Erfolg
- > *The „unreasonable effectiveness of data“:*
Mehr Daten schlagen den besseren Algorithmus!

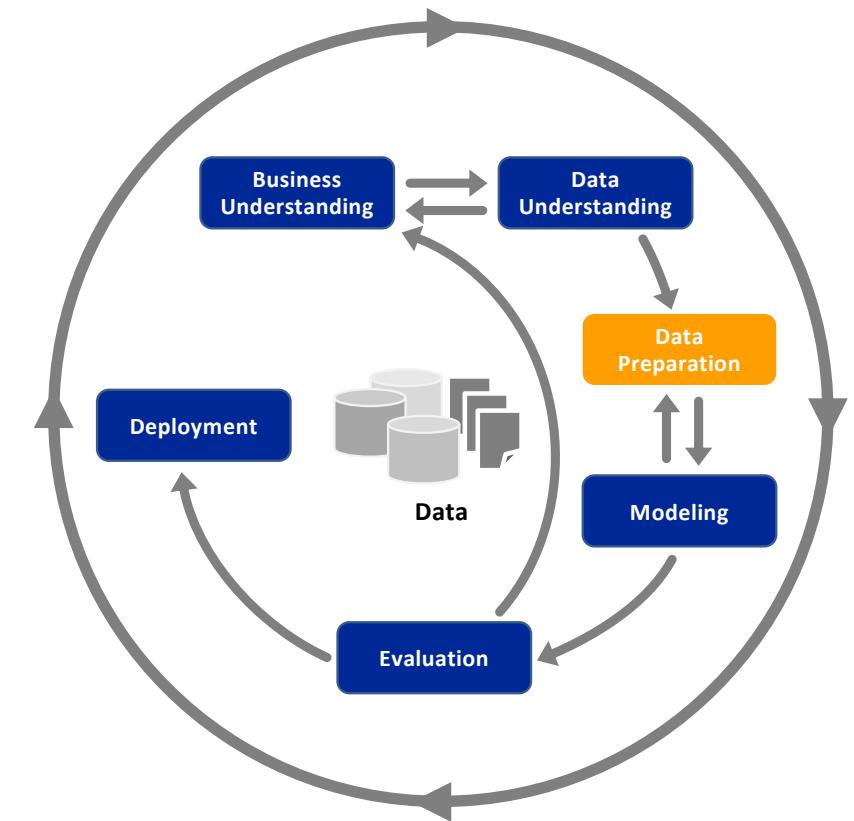


CRISP-DM Phase 3: Data Preparation

Ziel: Aufbereitung der Daten (oft als Datenmatrix) für die Modellierungsphase

Aufgaben

- > Integration unterschiedlicher Datenquellen
- > Datenbereinigung (Fehler, fehlende Werte)
- > Formatierung, Codierung, Skalierung
- > Auswahl relevanter Merkmale (Feature Selection)
- > Konstruktion neuer Merkmale (Feature Engineering)

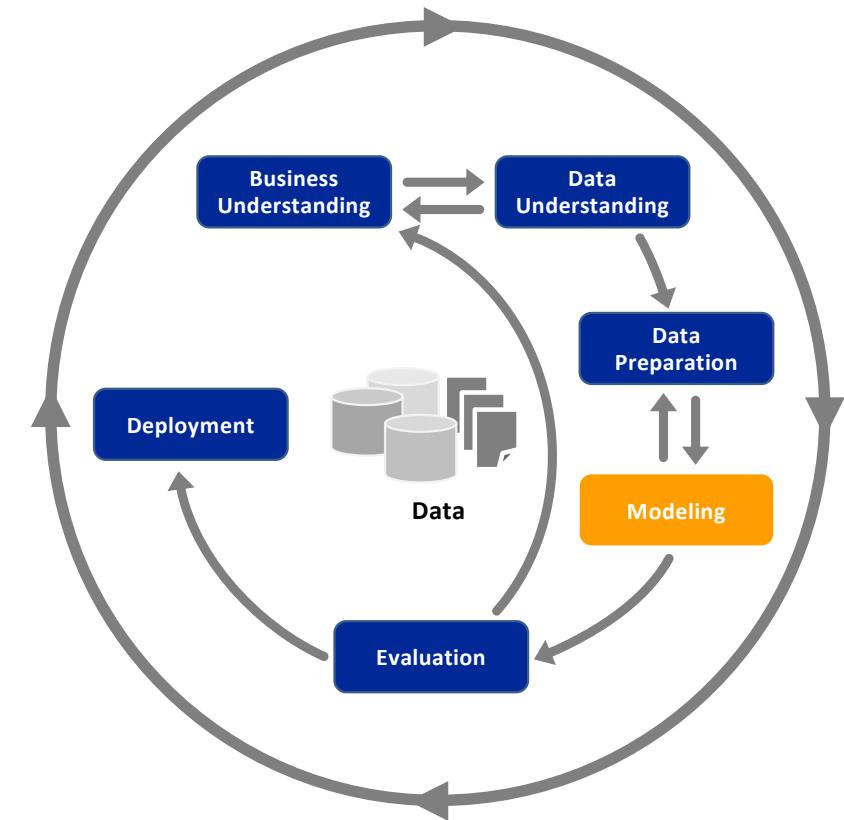


CRISP-DM Phase 4: Modeling

Ziel: Modellerstellung durch maschinelle Lernverfahren

Aufgaben

- > Auswahl geeigneter Lernverfahren (Model Selection)
- > Nicht auf einen Lösungsansatz (Lernverfahren) festlegen („no free lunch“)
- > Festlegen von Modellparametern
- > Aufbau verschiedener Modelle (Lernen aus Daten)
- > Auswahl eines Modells für die Aufgabenstellung

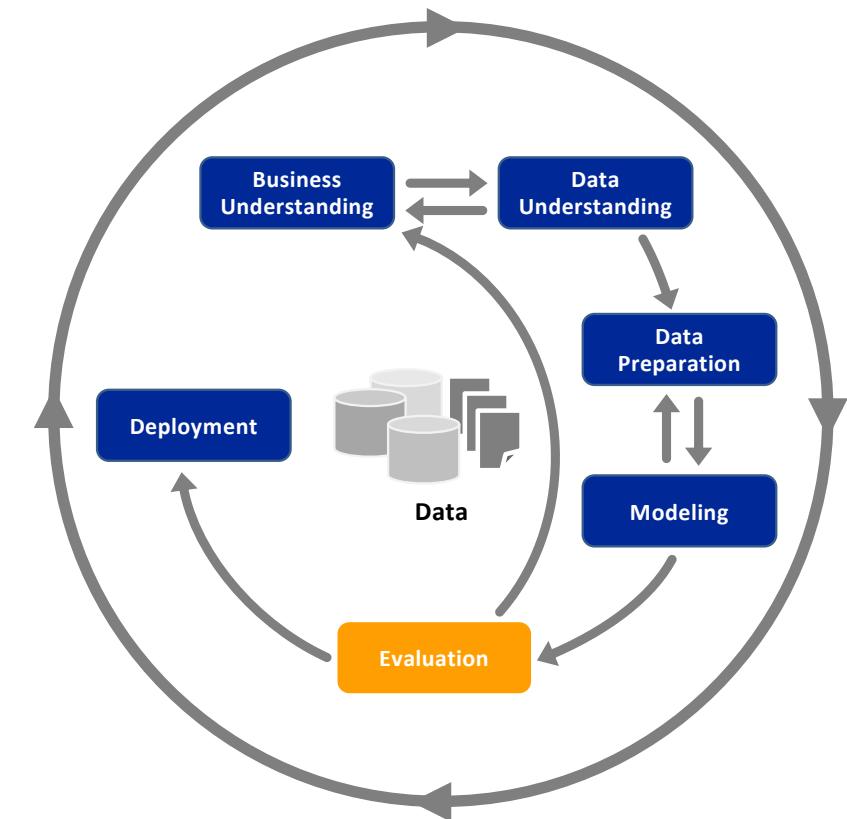


CRISP-DM Phase 5: Evaluation

Ziel: Bewertung des Modells anhand der Erfolgskriterien

Aufgaben

- > Überprüfen der Leistungsfähigkeit mit Testdaten
- > Bewertung und Visualisierung der Ergebnisse
 - Plausibilisierung der Erkenntnisse
 - Robustheit
- > Überprüfen, ob das Model die fachliche Aufgabenstellung unter Berücksichtigung der Erfolgskriterien zufriedenstellend lösen kann

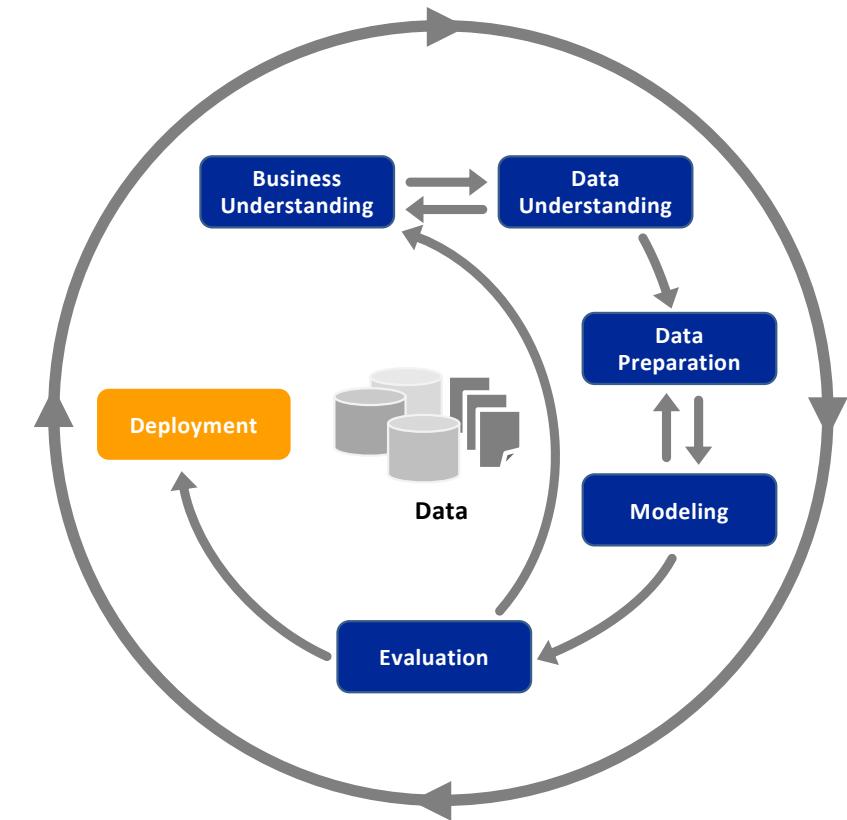


CRISP-DM Phase 6: Deployment

Ziel: Nutzung des Modells im Kontext der Anwendung

Aufgaben

- > Präsentation der Ergebnisse
- > Projektabschluss / Review
- > Einbindung des Modells in den operativen Betrieb
- > Management des Modelllebenszyklus bei regelmäßiger/automatisierter Nutzung
- > Kontinuierliche Beobachtung der Daten- und Ergebnisqualität
- > Konzept zur Modellanpassung

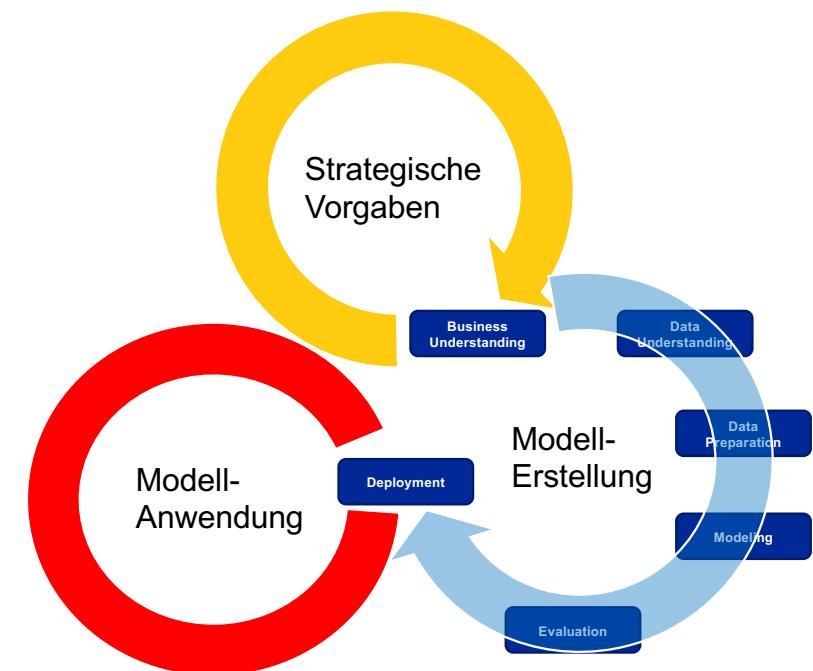


CRISP-DM: Zusammenfassung Phasen und Aufgaben

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives <i>Background Business Objectives Business Success Criteria</i>	Collect Initial Data <i>Initial Data Collection Report</i>	Data Set <i>Data Set Description</i>	Select Modeling Technique <i>Modeling Technique Modeling Assumptions</i>	Evaluate Results <i>Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models</i>	Plan Deployment <i>Deployment Plan</i>
Situation Assessment <i>Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits</i>	Describe Data <i>Data Description Report</i>	Select Data <i>Rationale for Inclusion / Exclusion</i>	Generate Test Design <i>Test Design</i>	Review Process <i>Review of Process</i>	Plan Monitoring and Maintenance <i>Monitoring and Maintenance Plan</i>
Determine Data Mining Goal <i>Data Mining Goals Data Mining Success Criteria</i>	Explore Data <i>Data Exploration Report</i>	Clean Data <i>Data Cleaning Report</i>	Build Model <i>Parameter Settings Models Model Description</i>	Determine Next Steps <i>List of Possible Actions Decision</i>	Produce Final Report <i>Final Report Final Presentation</i>
Produce Project Plan <i>Project Plan Initial Assessment of Tools and Techniques</i>	Verify Data Quality <i>Data Quality Report</i>	Construct Data <i>Derived Attributes Generated Records</i>	Assess Model <i>Model Assessment Revised Parameter Settings</i>		Review Project <i>Experience Documentation</i>

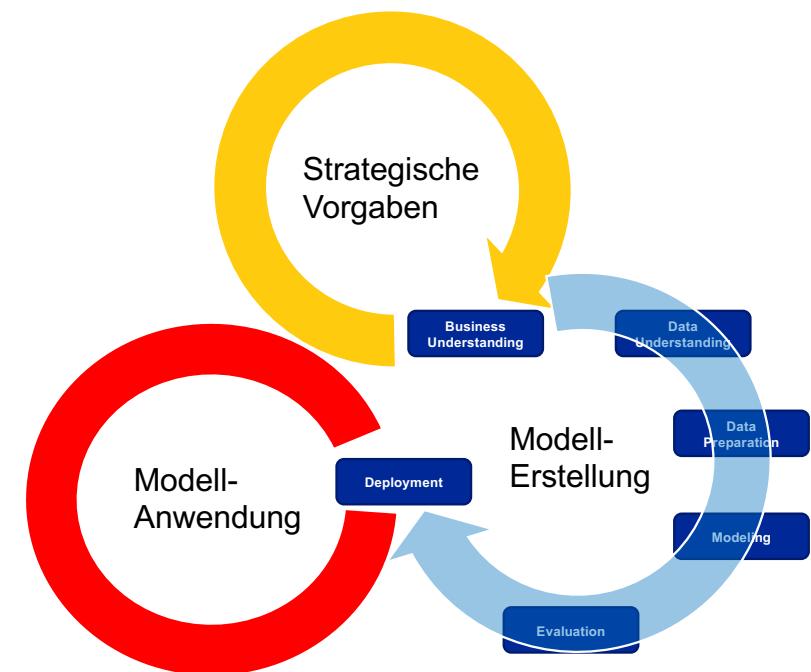
CRISP-DM und das datengetriebene Unternehmen

- > Bei CRISP-DM steht die Durchführung von Analyseprojekten im Fokus
- > Anknüpfung an die Unternehmensstrategie
 - Wie sieht die Strategie eines datengetriebenen Unternehmens aus?
 - Welche Aufgabenstellungen (Analyseziele) sollen verfolgt werden?
- > Anwendung von Modellen im Unternehmenskontext
 - Wie werden Ergebnisse (Modelle) automatisiert und nachhaltig in die Geschäftsprozesse eingebunden?
 - Management Modelllebenszyklus



CRISP-DM im Kontext von Big Data Science (Analytics)

- > Bei CRISP-DM steht die Durchführung von Analyseprojekten im Fokus
- > **Berücksichtigung von Veränderungen notwendig?**
 - Daten: auch semi- und unstrukturiert
 - Infrastruktur: Big-Data-Technologien
 - Modellerstellung: z.B. Deep Learning
 - Organisation: Agilität, Zusammenarbeit



Zusammenfassung

- > Daten sind ein sehr wichtiger Faktor in Unternehmen
- > Viele Disziplinen leiten aus Daten Erkenntnisse ab
- > **Data Science** entwickelt sich als Obergriff
- > Interaktion im Analyseprozess erlaubt Integration von Expertenwissen und fördert den Erkenntnisgewinn
- > Prozessmodelle mit definierten Phasen und Aufgaben strukturieren Analyseprojekte und führen zu einem effizienten und gut kommunizierbaren Ablauf
- > Strategie / Festlegung einer Fragestellung
- > Deployment / Management des Modelllebenszyklus

