

# 01

## DATA SCIENCE

# KURSINHALTE UND TERMINE

Kursinhalte	
Data Science	1-2
Grundlagen, Methoden und Anwendungen Künstlicher Intelligenz (KI)	2-4
Software-Testing	5
Projektmanagement-Ansätze in der Softwareentwicklung	6
Paradigmen der Softwareentwicklung	7
Vom Modell zur Produktion	8

Termine			
Wochentag	Datum	von - bis	Räume
Freitag	04.04.2025	09:00 - 12:15	HAN - Schiffgraben 49-51 - 1.24 Südstadt
Freitag	11.04.2025	09:00 - 12:15	HAN - Schiffgraben 49-51 - 1.27 Bothfeld
Freitag	25.04.2025	09:00 - 12:15	HAN - Schiffgraben 49-51 - 1.24 Südstadt
Freitag	16.05.2025	09:00 - 12:15	HAN - Schiffgraben 49-51 - 1.24 Südstadt
Freitag	23.05.2025	09:00 - 12:15	HAN - Schiffgraben 49-51 - 1.24 Südstadt
Freitag	06.06.2025	09:00 - 12:15	HAN - Schiffgraben 49-51 - 1.24 Südstadt
Freitag	20.06.2025	09:00 - 12:15	HAN - Schiffgraben 49-51 - 1.24 Südstadt
Freitag	04.07.2025	09:00 - 12:15	HAN - Schiffgraben 49-51 - 1.24 Südstadt
Freitag	11.07.2025	09:00 - 12:15	HAN - Schiffgraben 49-51 - 1.24 Südstadt

Nach der Bearbeitung dieser Lektion werdet ihr wissen, ...

- was unter **Data Science, Data Mining und Knowledge Discovery in Databases (KDD)** verstanden wird und wie sich die **Prozesse** unterscheiden.
- welche **Rollen** in einem **Data Science Projekt** unterschieden werden können.
- welche **Tools** und **Plattformen** im Bereich **Data Science** eingesetzt werden.



# DATA SCIENCE - DEFINITIONEN

— „**Data Science** ist ein interdisziplinäres Wissenschaftsfeld, welches durch die Anwendung wissenschaftlich fundierter **Methoden, Prozesse, Algorithmen und Systeme** die Extraktion von **Erkenntnissen, Mustern und Schlüssen** sowohl aus strukturierten als auch aus unstrukturierten Daten ermöglicht.“ (Gesellschaft für Informatik, Arbeitspapier, 2019)

— Vier Kernbereiche (acatech, 2018):

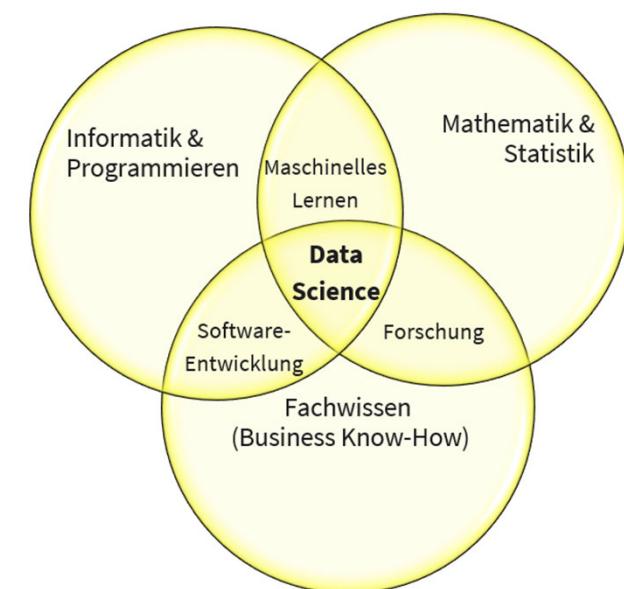
1. **Data Engineering** umfasst alle Methoden und Prozesse, die für die Speicherung, den Zugriff sowie die Rückverfolgbarkeit von Daten nötig sind.
2. **Data Analytics** beschäftigt sich mit der Datenanalyse.
3. **Data Prediction** befasst sich mit der Vorhersage von Themen und Situationen auf Basis von Erfahrungswissen.
4. **Maschinelles Lernen** ist ein Querschnittsbereich zu den anderen drei Bereichen und steht für die Entwicklung von Algorithmen, die aus Daten (Erfahrungswissen) lernen, dabei Muster erkennen, Modelle generieren und darauf aufbauend Themen und Situationen vorhersagen können.



Der **Schwerpunkt der Data Science** liegt dabei nicht auf den Daten selbst, sondern auf der **Art und Weise**, wie diese **verarbeitet, aufbereitet und analysiert** werden. Data Science beschäftigt sich mit einer **zweckorientierten Datenanalyse** und der **systematischen Generierung von Entscheidungshilfen und -grundlagen**, um **Wettbewerbsvorteile** erzielen zu können.

# DATA SCIENCE ALS INTERDISZIPLINÄRES THEMA

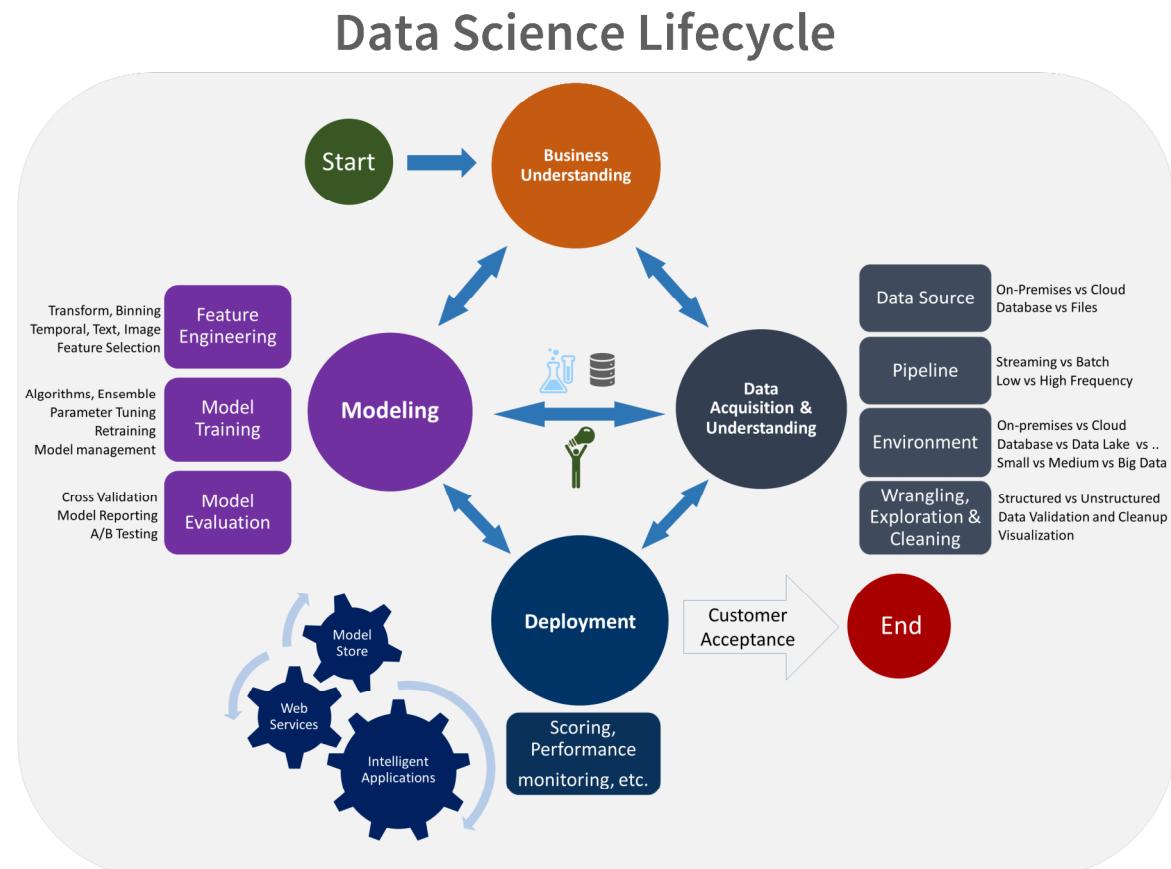
- Data Science ist ein **interdisziplinäres Wissenschaftsfeld**, das sich mit der exakten **digitalen Erfassung, Analyse und Visualisierung** vergangener, aktueller sowie zukünftiger Phänomene unserer realen Welt beschäftigt, um datengetrieben den Prozess der Wissensgenerierung als bestmögliche **Entscheidungsbasis für menschliches Handeln** zu optimieren.
- Bezüge zur **Künstlichen Intelligenz** (definiert Herausforderungen, die es zu lösen gilt, und entwickelt Lösungsansätze)
- Bezüge zum **Maschinellen Lernen** (hier steht das Erlernen der Lösungen im Vordergrund)
- sowohl für die **Unternehmenspraxis** als auch für **Lehre und Forschung** von großer Relevanz
- Im Unternehmensumfeld häufig im Bereich **Business Intelligence** angesiedelt
- **Unternehmen aus allen Branchen suchen** händeringend **Experten** in dem Bereich oder sehen die Notwendigkeit, diese selbst aus- und weiterzubilden.
- In der Wissenschaft beschäftigt sich Data Science mit **unterschiedlichen Teilbereichen** und kann daher vor dem Hintergrund verschiedener akademischer Disziplinen betrieben werden:  
Informatik, Statistik, Mathematik, Natur- oder Wirtschaftswissenschaften, einschließlich des Maschinellen Lernens, des Statistischen Lernens, der Programmierung, der Datentechnik, der Mustererkennung, der Prognostik, der Modellierung von Unsicherheiten und der Datenlagerung.



[https://gi.de/fileadmin/GI/Allgemein/PDF/GI\\_Arbeitspapier\\_Data-Science\\_2019-12.pdf](https://gi.de/fileadmin/GI/Allgemein/PDF/GI_Arbeitspapier_Data-Science_2019-12.pdf)  
[https://www.vde-verlag.de/buecher/leseprobe/9783879077212\\_PROBE\\_01.pdf](https://www.vde-verlag.de/buecher/leseprobe/9783879077212_PROBE_01.pdf)

# TEAM DATA SCIENCE-PROZESS (TDSP)

- Der TDSP-Lebenszyklus besteht aus fünf Hauptphasen, die von Ihrem Team immer wieder wiederholt werden. Diese Phasen umfassen:
  - Geschäftliche Aspekte
  - Datenerfassung und -auswertung
  - Modellierung
  - Bereitstellung
  - Kundenakzeptanz



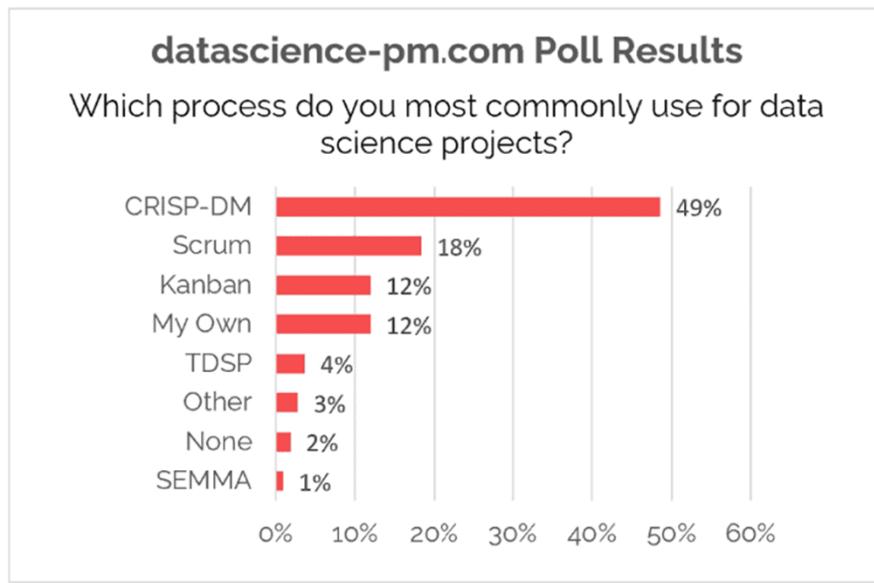
[Was ist der Team Data Science-Prozess \(TDSP\)? - Azure Architecture Center | Microsoft Learn](#)

# VERWENDUNG DER PROZESSE IN DER PRAXIS

## What is the Most Common Data Science Process?

We asked you – our readers – this question in a [poll](#) in August and September 2020. [CRISP-DM](#) was by far the most common response.

This is consistent with similar surveys done by other organizations in years past.

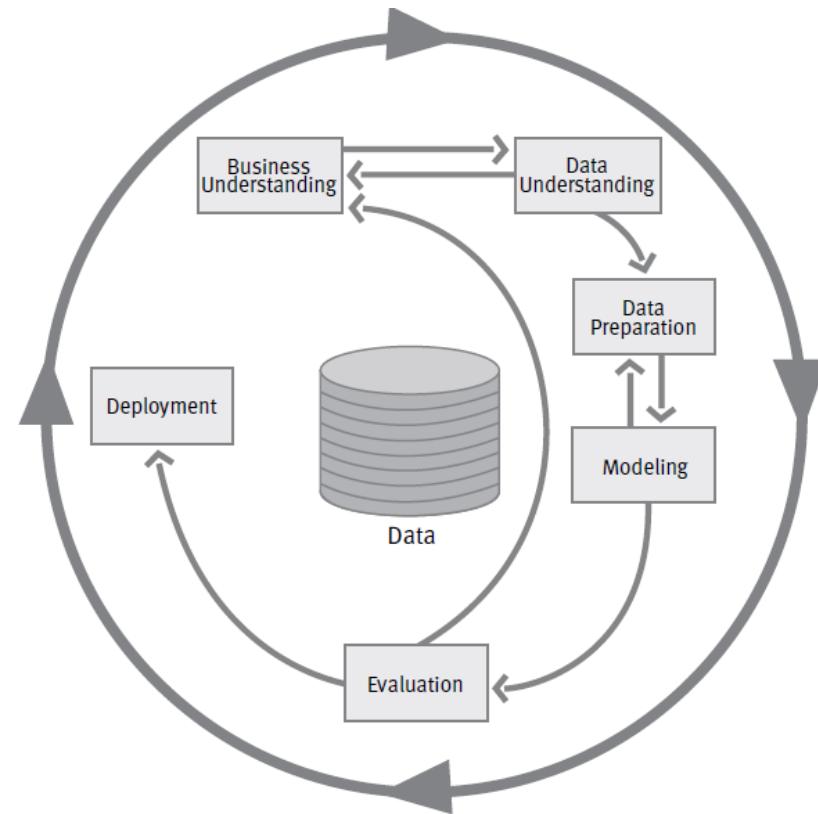


[What is the Data Science Process? - Data Science Process Alliance \(datascience-pm.com\)](#)

# WEITERE ANSÄTZE AUS DER INDUSTRIE

## Cross Industry Standard Process for Data Mining **(CRISP-DM)**

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment



Quelle Grafik: Chapman, P., et al., 2000, CRISP-DM 1.0, S. 10.

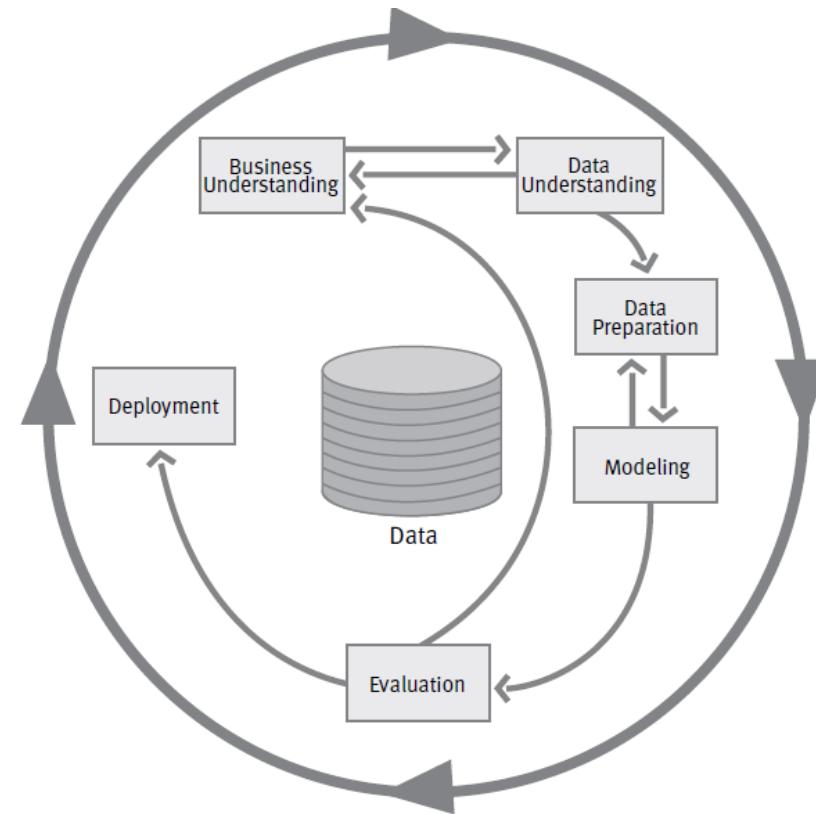
# CRISP-DM

## PHASEN UND AKTIVITÄTEN

Aufgabendefinition	Datenverständnis	Datenaufbereitung
Business Understanding	Data Understanding	Data Preparation
<ul style="list-style-type: none"> <li>Bestimmung der betriebswirtschaftlichen Problemstellung</li> <li>Situationsbewertung</li> <li>Bestimmung analytischer Ziele</li> <li>Erstellung eines Projektplans</li> </ul>	<ul style="list-style-type: none"> <li>Daten sammeln</li> <li>Daten beschreiben</li> <li>Untersuchung der Daten</li> <li>Verifizierung der Datenqualität</li> </ul>	<ul style="list-style-type: none"> <li>Auswahl der Daten</li> <li>Bereinigung der Daten</li> <li>Transformation und Integration der Daten</li> <li>Daten-formatierung</li> </ul>

Die Datenaufbereitung nimmt zwischen 40-70 % der Zeit in Anspruch

Quelle: Datenaufbereitung im Mining-Prozess - IBM Dokumentation

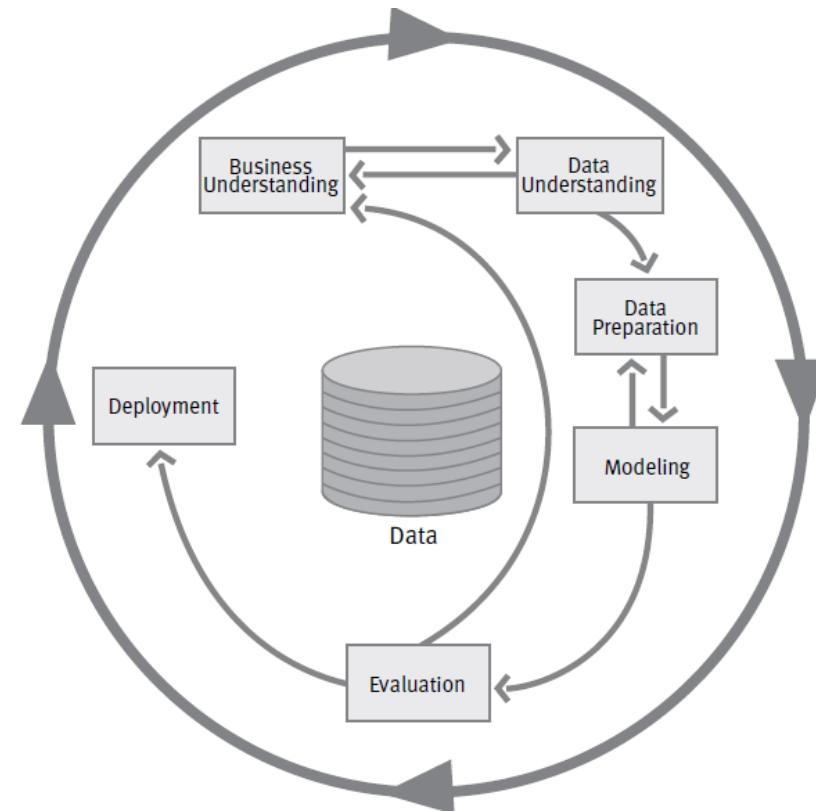


Quelle Grafik: Chapman, P., et al., 2000, CRISP-DM 1.0, S. 10.

# CRISP-DM

## PHASEN UND AKTIVITÄTEN

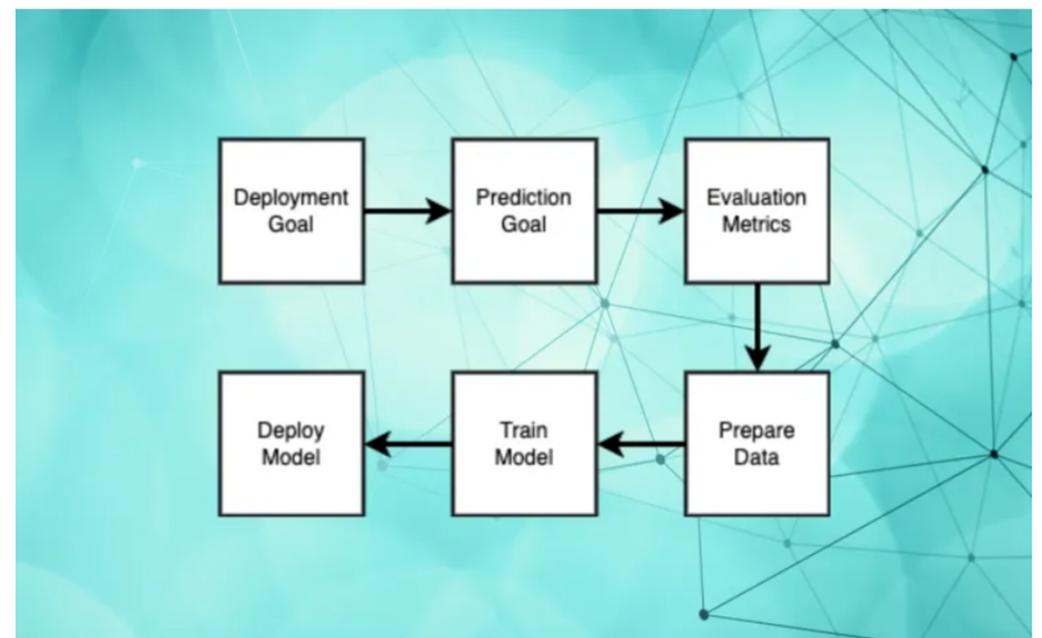
Modellierung	Bewertung	Anwendung
Modeling	Evaluation	Deployment
<ul style="list-style-type: none"> <li>Auswahl des Modells</li> <li>Testmodell erstellen</li> <li>Modell erstellen</li> </ul>	<ul style="list-style-type: none"> <li>Bewertung des Modells</li> <li>Bewertung der Resultate</li> <li>Bewertung des Prozesses</li> <li>Nächste Schritte festlegen</li> </ul>	<ul style="list-style-type: none"> <li>Zusammenfassender Bericht und Präsentation der Ergebnisse</li> <li>Implementierungsstrategie planen</li> <li>Überwachung der Gültigkeit der Modelle planen</li> </ul>



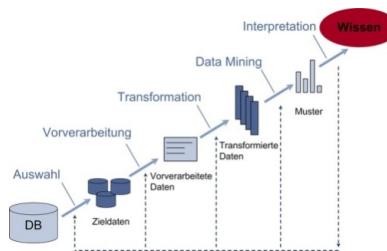
Quelle Grafik: Chapman, P., et al., 2000, CRISP-DM 1.0, S. 10.

Eric Siegel's end-to-end framework for ML projects is called bizML, and as he recommends, starts from the end goal and moves backward:

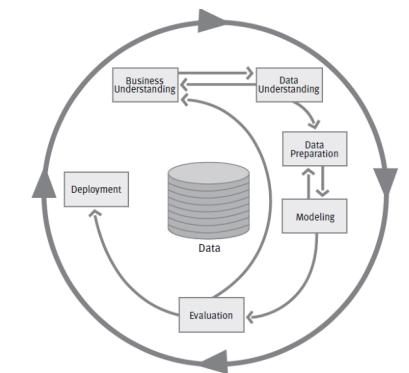
- **1- Establish the deployment goal:** Define clearly how ML will affect your operations
- **2- Establish the prediction goal:** Determine what the model will predict and how it relates to the deployment goal
- **3- Establish the evaluation metrics:** Determine the metrics that matter and the performance level required to achieve the deployment goal
- **4- Prepare the data:** Define what the data must look like and prepare the datasets
- **5- Train the model:** Use the data to train your machine learning model
- **6- Deploy the model:** Integrate the model into your product to make predictions on new data coming from business operations



# VERGLEICH VON KDD UND CRISP-DM



KDD	CRISP-DM
-	<b>Business Understanding</b>
<b>Auswahl</b>	<b>Data Understanding</b>
<b>Vorverarbeitung</b>	<b>Data preparation</b>
<b>Transformation</b>	<b>Modeling</b>
<b>Data Mining</b>	<b>Evaluation</b>
<b>Interpretation / Evaluation</b>	
-	<b>Deployment</b>



Quelle: Azevedo, A. I. R. L., 2008, KDD, SEMMA and CRISP-DM: a parallel overview, S. 5

# KNOWLEDGE DISCOVERY IN DATABASES (KDD) UND DATA MINING - DEFINITIONEN

## — “Knowledge Discovery in Databases (KDD)

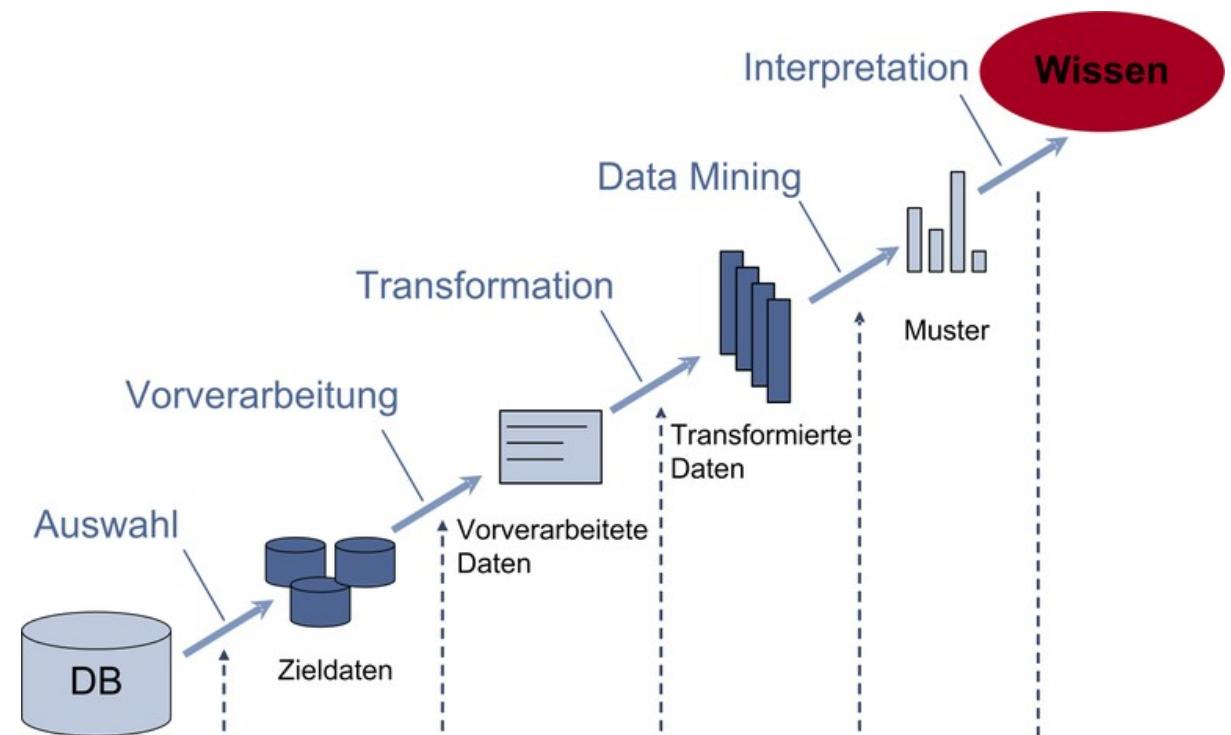
describes the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.”

[Fayyad et al. 1996, S. 6]

## — “Data Mining

is a problem-solving methodology that finds a logical or mathematical description, eventually of a complex nature, of patterns and regularities in a set of data.”

[Decker/Focardi 1995, S. 3]



Quelle Grafik: KDD-Prozess nach Fayyad , 1996.

# Data Science, Data Mining und CRISP-DM

## Übungsfragen



1. **Warum** wird **Data Science** als ein **interdisziplinäres Wissenschaftsfeld** bezeichnet?
2. **Welcher Prozess** wird im Bereich **Data Science** häufig verwendet?
3. **Wie** lauten die **6 Phasen** des „Cross Industry Standard Process for Data Mining“ (**CRISP-DM**).
4. **Welche** der **Phasen** nimmt in der Praxis **40-70% der Zeit** in Anspruch?
5. **Was** wird unter dem Begriff **Data Mining** verstanden?  
Beschreibe den Begriff mit eigenen Worten.

Nach der Bearbeitung dieser Lektion werdet ihr wissen, ...



- was unter **Data Science, Data Mining und Knowledge Discovery in Databases (KDD)** verstanden wird und wie sich die **Prozesse** unterscheiden.
- welche **Rollen** in einem **Data Science Projekt** unterschieden werden können
- welche **Tools** und **Plattformen** im Bereich **Data Science** eingesetzt werden

# VERGLEICH UNTERSCHIEDLICHER DATA-SCIENCE-STRUKTURIERUNGSANSÄTZE UND BENÖTIGTER IT-SKILLS

EDSF V3	ACATECH	CRISP-DM	DATA LITERACY	IT-SKILLS
Business Analytics (DSBA) und Domänen- Spezifika	Datengetriebene Geschäftsmodelle	Business Understanding	-	Business- Domänenwissen
Data Management (DSDM)	Data Engineering	Data Understanding	Datensammlung	Datenintegration und -transformation
		Data Preparation	Datenmanagement	
			Datenanwendung	
Data Analytics (DSDA)	Data Analytics	Modelling	Datenevaluation	Statistik und statistische Programmiersprachen
	Data Prediction	Evaluation		Präsentation und Visualisierung
	Maschinelles Lernen			
Research Methods and Project Management (DSRMP)	Forschung und Entwicklung für Data Analytics	Modelling	Konzeptioneller Rahmen	-
	Open Data	Evaluation	Datenevaluation	
Data Science Engineering (DSENG)	Data Engineering	Deployment	Datenevaluation	Big-Data-Infrastruktur
	Rechtsfragen		Datenanwendung	

[https://gi.de/fileadmin/GI/Allgemein/PDF/GI\\_Arbeitspapier\\_Data-Science\\_2019-12.pdf](https://gi.de/fileadmin/GI/Allgemein/PDF/GI_Arbeitspapier_Data-Science_2019-12.pdf)

## ROLLEN IM DATA SCIENCE UMFELD

### BUSINESS ENGINEER

#### Business Engineer

Als Business Engineer analysieren Sie die Geschäfts- und IT-Prozesse unserer Kunden, führen das Requirements Engineering und Stakeholdermanagement in verschiedenen Projekten durch und fungieren so als Schnittstelle zwischen Fachbereich und IT. Daneben sind Sie verantwortlich für das Projektmanagement in agilem und klassischem Projektumfeld sowie für die Durchführung von Kunden-Workshops vor Ort.



#### Data Analyst

#### Data Engineer

#### Big Data Engineer

#### Data Scientist

Unternehmerisches Denken, ausgezeichnete kommunikative Fähigkeiten und Überzeugungskraft gehören ebenso zu Ihren Merkmalen wie fundierte Kenntnisse im Business-Intelligence-Bereich mit Data Warehouse-Konzepten und Data-and-Analytics-Architekturen. In den Bereichen Business Consulting, Business Analysis und Data and Analysis fühlen Sie sich wohl.

# ROLLEN IM DATA SCIENCE UMFELD

## DATA ANALYST

### Business Engineer

### Data Analyst

### Data Engineer

### Big Data Engineer

### Data Scientist

Als Data Analyst erheben, visualisieren und werten Sie Daten systematisch aus, um daraus geschäftsrelevante Erkenntnisse zu generieren. Sie sind verantwortlich für die Konzeption, Entwicklung, Überwachung und Steuerung von BI-Systemen bei unseren Kundinnen und Kunden und erstellen beziehungsweise konzipieren passende IT-Lösungen.

Als Data Analyst bestechen Sie durch ein hohes Maß an analytischer und vernetzter Denkweise, gepaart mit konzeptionellen Fähigkeiten sowie einer raschen Auffassungsgabe. In den Bereichen Business Consulting, Business Analysis und Data Analysis fühlen Sie sich wohl.



# ROLLEN IM DATA SCIENCE UMFELD

## DATA ENGINEER

### Business Engineer

Als Data Engineer sind Sie verantwortlich für alle Prozesse rund um die Generierung, Speicherung, Pflege, Aufbereitung, Anreicherung und Weitergabe von Daten. Sie beraten und unterstützen unsere Kundinnen und Kunden maßgeblich beim Aufbau und bei der Überwachung der Hardware- und Software-Infrastruktur.



### Data Analyst

### Data Engineer

### Big Data Engineer

### Data Scientist

Als Data Engineer kennen Sie alle Anforderungen an Datenprozesse und können Datenmengen performant skalieren. Begriffe wie Programmierung, Datenbanken und SQL sind für Sie keine Fremdwörter. In den Bereichen Big Data, Konzeption und Softwareentwicklung fühlen Sie sich wohl.

# ROLLEN IM DATA SCIENCE UMFELD

## BIG DATA ENGINEER

Business Engineer

Data Analyst

Data Engineer

Big Data Engineer

Data Scientist

Als Big Data Engineer evaluieren und beraten Sie unsere Kundinnen und Kunden zu modernen Data-Plattform-Architekturen. Darüber hinaus sind Sie verantwortlich für das Design und die Konzeption von komplexen ingest-, process-, und access-Prozessen für datengetriebene Produkte. Maßgeblich beeinflussen werden Sie den Aufbau und die Definition von Big-Data-Plattformen und Data Lakes.

Mit viel Begeisterung eignen Sie sich neues Wissen zu Applikationen, Technologien und Tools im Big-Data-Kontext an. SQL ist für Sie keine Fremdsprache und Sie können einfache Abfragen formulieren. Ihr tiefes Verständnis der aktuellen Big-Data-Landschaft ermöglicht es Ihnen, sich schnell in neue Projekte einzuarbeiten und auf diese Weise eine konkrete Grundlage für Daten-Analyseprozesse zu schaffen. In den Bereichen Big Data, Konzeption und Softwareentwicklung fühlen Sie sich wohl.



# ROLLEN IM DATA SCIENCE UMFELD

## DATA SCIENTIST

Business Engineer

Data Analyst

Data Engineer

Big Data Engineer

Data Scientist

Als Data Scientist analysieren Sie die Daten unserer Kundinnen und Kunden nach Trends und Mustern und unterstützen auf diese Weise die Geschäftsprozessoptimierung. Unter Verwendung moderner Machine-Learning- und Deep-Learning-Verfahren beantworten Sie verschiedenste Fragestellungen aus den Fachbereichen unserer Kundinnen und Kunden und leiten daraus Handlungsempfehlungen ab. Dabei ist eine enge Kooperation und Absprache mit anderen Abteilungen unabdingbar.

Analytisches Denken, ausgezeichnete kommunikative Fähigkeiten sowie Überzeugungskraft zeichnen Sie aus und Machine-Learning- und Deep-Learning-Verfahren sind aus Ihrem Leben nicht mehr wegzudenken. Mit großer Begeisterung arbeiten Sie mit großen Datenmengen und -banken. In den Bereichen Statistik, Business-Analyse und Softwareentwicklung fühlen sie sich wohl.



# ROLLEN IM DATA SCIENCE UMFELD

## WEITERE ROLLEN

Position	Aufgaben	Kenntnisse	Bedeutung
Data Scientist	Auswahl einer Analysestrategie mit passenden statistischen Modellen, Visualisierung der Ergebnisse	Mathematische, statistische Modelle, Fähigkeiten der Informatik	Kommunikation zwischen Fachabteilungen, Bedürfnisse erkennen und umsetzen
Data Engineer	Sammlung, Aufbereitung und Analyse von Daten	Umgang mit Datenbanken, Data Warehousing-Tools und Cloud-Systemen	Technischer Experte im Team, Implementierung der Ergebnisse ins operative Geschäft
Machine Learning Engineer	Anpassung von Machine Learning Modellen, Analyse großer Datenmengen	Mathematik, Computerwissenschaften, Programmierung und Statistik	Experte im Bereich Prozessautomatisierung und Datenverarbeitung
DevOps Engineer	Monitoring, Programmierung, Erstellung von Skripten	Administrative und softwarebasierte Aufgaben	Zusammenführung zwischen Entwicklern und dem IT-Betrieb
Domänenexperte	Branchenwissen, Einschätzung bestimmter Fragestellungen	Fachkenntnisse des jeweiligen Gebiets (Marketing Manager, Ingenieur, Maschinenbau etc.)	Ständiger Austausch mit technischen Experten zur Entwicklung einer Strategie
Chief Analytics Officer	Umgang mit Daten in einem Unternehmen, Personalrekrutierung	Statistische Analysen, Marketing, Finanzen und betriebswirtschaftliche Kenntnisse	Bereitstellung und Rekrutierung von geschultem Personal

Aufbau eines Data Science Teams - datasolut GmbH

# ROLLEN IM DATA SCIENCE PROJEKT



## 4 Kernrollen:

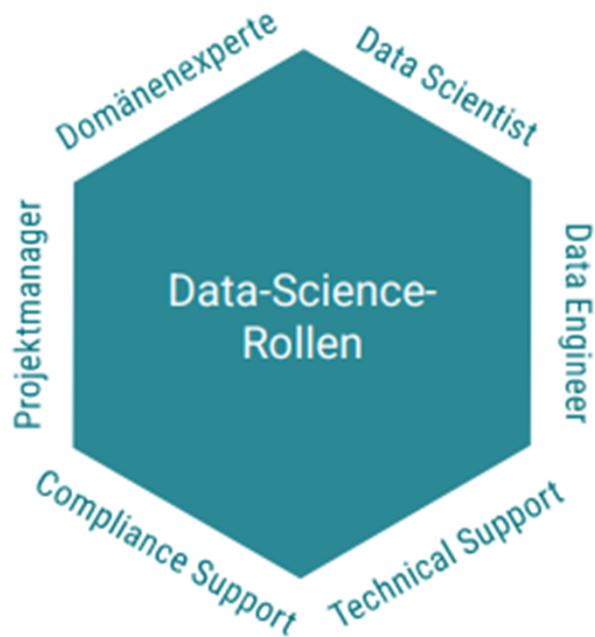
- Data Scientist
- Data Engineer
- Domänenexperte

„Domänenexperten sind Fachanwender oder Vertreter der Fachanwender. Sie verfügen über spezifisches Wissen in Bezug auf die Anwendungsdomäne und besitzen ein inhaltliches Verständnis der Problemstellung bzw. des Anwendungsfalls“

- Projektmanager

„Projektmanager planen, steuern und koordinieren den Gesamtablauf eines Data-Science-Projekts. Insbesondere bei kleineren Projekten wird das Projektmanagement häufig von Personen übernommen, die auch die Rolle eines Data Scientists oder eines Data Engineers ausfüllen“

# ROLLEN IM DATA SCIENCE PROJEKT



## 2 ergänzende Rollen:

### – Technischer Support

„Der technische Support umfasst alle Aufgaben, die erledigt werden müssen, um die technischen Voraussetzungen für die Durchführung des Data-Science-Projekts zu schaffen.“

### – Compliance Support

„Der Compliance-Support ist für die Einhaltung gesetzlicher Vorgaben, die Kompatibilität des DataScience-Projekts mit den organisationsinternen Regelwerken und das korrekte Verhalten der Projektmitarbeiter verantwortlich. Er ist außerdem für das übergreifende Sicherheitsmanagement zuständig und gewährleistet den Datenschutz, insbesondere den Schutz personenbezogener Daten.“

# DATA SCIENCE KARRIEREN UND GEHÄLTER

Career	Overview	Average Salary
Data Scientist	Analyze and interpret complex digital data to create actionable insights for companies, utilizing expertise in computer science, statistics, and mathematics.	\$156k
Data Analyst	Process and perform statistical analysis on large datasets, translating data into insights to inform business decisions, using tools like Excel, SQL, and visualization software.	\$77k
Business Analyst	Identify business needs and challenges, translating them into solutions through data analysis, process improvement, and resource allocation.	\$84k
Database Administrator	Manage, secure, and ensure the availability and performance of database servers, facilitating access to users and maintaining data integrity.	\$77k
Data Engineer	Develop and maintain the infrastructure that allows for the efficient flow and accessibility of data for analysis, utilizing skills in programming and system design.	\$125k
Data Architect	Design the blueprint of data management systems, ensuring scalable and efficient data infrastructure that supports organizational goals.	\$165k
Machine Learning Engineer	Develop algorithms and models for tasks like image recognition and predictive analytics, collaborating with data scientists to apply insights from data.	\$125K to \$187K
Quantitative Analyst	Apply mathematical and statistical methods to financial markets, creating models for trend prediction, investment strategy evaluation, and risk assessment.	\$173k
Data Mining Specialist	Use statistical and machine learning techniques to uncover patterns and insights in large datasets, supporting predictive analytics and decision-making.	\$109k
Data Visualization Engineer	Specialize in converting data into visually appealing graphics, working with data analysts and business teams to tell stories with data.	\$103k

Achtung: Quelle der Gehaltszahlen  
nicht transparent aufgezeigt!

## Rollen im Data Science Projekt

### Übungsfragen



1. Nenne **3 übliche Rollen**, die an einem **Data Science Projekt** beteiligt sind!
2. Was ist die Aufgabe **eines Data Scientist**?
3. Welche Aufgabe hat ein **Data Engineer**?
4. Was macht ein **Domänenexperte**?
5. Welche **Rollen** sind aus deiner Sicht in jedem Data Science Projekt **unbedingt zu besetzen**?

Nach der Bearbeitung dieser Lektion werdet ihr wissen, ...



- was unter **Data Science, Data Mining und Knowledge Discovery in Databases (KDD)** verstanden wird und wie sich die **Prozesse** unterscheiden.
- welche **Rollen** in einem **Data Science Projekt** unterschieden werden können
- welche **Tools** und **Plattformen** im Bereich **Data Science** eingesetzt werden

# GARTNER DATA SCIENCE AND MACHINE LEARNING PLATFORMS

## Data Science und Machine Learning (DSML) Platforms

„Gartner definiert eine **Data-Science- und Machine-Learning-Plattform** als einen integrierten Satz von **codebasierten Bibliotheken** und **Low-Code-Tools**, die die unabhängige Nutzung und Zusammenarbeit zwischen **Data Scientists** und ihren Geschäfts- und IT-Kollegen **in allen Phasen des Data-Science-Lebenszyklus** unterstützen.

Zu diesen Phasen gehören das **Geschäftsverständnis**, der **Datenzugriff** und die **Datenaufbereitung**, das **Experimentieren** und die **Modellerstellung** sowie der **Austausch von Erkenntnissen**.

Sie unterstützen auch **Engineering-Workflows für maschinelles Lernen**, einschließlich der Erstellung von Daten-, Funktions-, Bereitstellungs- und Testpipelines.

Die Plattformen werden über einen **Desktop-Client** oder **Browser** mit unterstützenden Compute-Instanzen und/oder als vollständig verwaltetes **Cloud-Angebot** bereitgestellt.“

(Gartner, 2024)



(US-amerikanischer Anbieter von  
Marktforschungsergebnissen und Analysen  
über die IT. Bekannt für Gartner Hypecycle und  
Magic Quadrant)

# GARTNER MAGIC QUADRANT FOR DATA SCIENCE AND MACHINE LEARNING PLATFORMS

## Gartner Magic Quadrant

- **Leaders** execute well against their current vision and are well positioned for tomorrow.”
- **Visionaries** understand where the market is going or have a vision for changing market rules but do not yet execute well.”
- **Niche Players** focus successfully on a small segment or are unfocused and do not out-innovate or outperform others.”
- **Challengers** execute well today or may dominate a large segment but do not demonstrate an understanding of market direction.”



[Gartner Magic Quadrant for Data Science and Machine Learning Platforms](#)

# GARTNER MAGIC QUADRANT FOR DATA SCIENCE AND MACHINE LEARNING PLATFORMS 2024

Analyse von 18 (vormals 20) Anbietern von Plattformen die insbesondere für **data science** und **machine learning** verwendet werden können:

- |                               |                                    |
|-------------------------------|------------------------------------|
| <b>1. Alibaba Cloud</b>       | <b>11. Alibaba Cloud</b>           |
| <b>2. Altair</b>              | <b>12. H2O.ai</b>                  |
| <b>3. Alteryx</b>             | <b>13. IBM</b>                     |
| <b>4. Amazon Web Services</b> | <b>14. KNIME</b>                   |
| <b>5. Anaconda</b>            | <b>15. MathWorks</b>               |
| <b>6. Cloudera</b>            | <b>16. Microsoft</b>               |
| <b>7. Databricks</b>          | <b>17. RapidMiner =&gt; Altair</b> |
| <b>8. Dataiku</b>             | <b>18. Samsung SDS</b>             |
| <b>9. DataRobot</b>           | <b>19. SAS</b>                     |
| <b>10. Domino</b>             | <b>20. TIBCO Software</b>          |

Figure 1: Magic Quadrant for Data Science and Machine Learning Platforms



[Gartner Magic Quadrant for Data Science and Machine Learning Platforms](#)

# GARTNER MAGIC QUADRANT FOR DATA SCIENCE AND MACHINE LEARNING PLATFORMS 2021 AND 2024



Figure 1: Magic Quadrant for Data Science and Machine Learning Platforms



Source: Gartner (June 2024)

**Gartner**

[Gartner Magic Quadrant for Data Science and Machine Learning Platforms](#)

# GARTNER MAGIC QUADRANT FOR DATA SCIENCE AND MACHINE LEARNING PLATFORMS 2024 – LEADER DETAILS

Provider	Product	Beschreibung	Lizenztyp	Website
<b>Databricks</b>	Data Intelligence Platform,	Cloud-basierte Plattform für Data Engineering, Machine Learning und Analytics, die Spark-basierte Datenverarbeitung und Zusammenarbeit in Echtzeit ermöglicht.	Proprietär / Subscription-Based	<a href="https://databricks.com/">https://databricks.com/</a>
<b>Microsoft</b>	Azure Machine Learning, ...	Microsofts cloud-basierte Machine Learning-Plattform zur Erstellung, Schulung und Bereitstellung von ML-Modellen	Proprietär / Subscription-Based	<a href="https://azure.microsoft.com/en-us/services/machine-learning/">https://azure.microsoft.com/en-us/services/machine-learning/</a>
<b>Google</b>	Cloud AutoML / Dataflow / Datalab, ...	Google Cloud-Service, der es ermöglicht, benutzerdefinierte ML-Modelle ohne tiefgehende Programmierkenntnisse zu erstellen und zu trainieren.	Proprietär / Pay-per-Use	<a href="https://cloud.google.com/automl">https://cloud.google.com/automl</a>
<b>Amazon Web Services</b>	SageMaker, ...	Amazon Web Services (AWS)-Plattform für das Erstellen, Trainieren und Bereitstellen von Machine Learning-Modellen auf skalierbarer Infrastruktur.	Proprietär / Pay-as-you-go	<a href="https://aws.amazon.com/sagemaker/">https://aws.amazon.com/sagemaker/</a>
<b>Dataiko</b>	Dataiko	Plattform für Data Science und maschinelles Lernen, die es Teams ermöglicht, Daten zu analysieren, Modelle zu entwickeln und diese in Produktionsumgebungen zu integrieren.	Proprietär / Subscription-Based	<a href="https://www.dataiku.com/">https://www.dataiku.com/</a>
<b>Altair</b>	RapidMiner	Open-Source-Plattform für Data Science und maschinelles Lernen, mit Funktionen für Modellierung, Datentransformation und Visualisierung.	Open-Source / Subscription-Based	<a href="https://rapidminer.com/">https://rapidminer.com/</a>
<b>SAS</b>	Base SAS / Enterprise Guide, ...	Software-Suite für Analytics, Data Management und Business Intelligence, die fortschrittliche statistische Modelle und maschinelles Lernen umfasst.	Proprietär / Subscription-Based	<a href="https://www.sas.com/">https://www.sas.com/</a>
<b>DataRobot</b>	Enterprise AI Suite	Plattform für Automatisierung von Machine Learning und KI, die es Unternehmen ermöglicht, Modelle schnell zu entwickeln und in Produktionsumgebungen einzusetzen.	Proprietär / Subscription-Based	<a href="https://www.datarobot.com/">https://www.datarobot.com/</a>

[Best Data Science and Machine Learning Platforms Reviews 2025 | Gartner Peer Insights](#)

# OPEN-SOURCE - DATA SCIENCE TOOLS

Development Environment (IDE)		Data Mining and Transformation		Data Analysis and Big Data Tools		Model Deployment		Data Visualization	
Jupyter Notebooks	Web application, host code, data, notes, equations, collaboration tool	Weka	Tool for data mining, pre-processing, classifying data, GUI for classification, association, regression, clustering	KNIME	End-to-end data analysis and integration and reporting, GUI for pre-processing, analysis, model building and visualization	TensorFlow.js	Machine learning framework, models in JavaScript or Node.js, deploy over the web / browser	Orange	Data visualization, GUI-based, beginner friendly, statistical distributions and box plots or decision trees, etc.
Zeppelin Notebooks	Web-based environment, many languages like Python, SQL, Scala, explore, share, analyze, visualize data	Scrapy	Writing spiders that crawl websites and extract data, written in Python	Hadoop	Storage and processing of big data, on distributed model, allows fast processing	MLFlow	Machine learning lifecycle management platform, building, packaging, deploying models	D3.js	Visualize data on web browsers using HTML, SVG and CSS, animation and interactive visuals
R Studio	Integrates R-based tools into single environment, write clean code, execute, manage workflows, debug	Pandas	Data wrangling software, written in Python, good for numerical tables or time-series data (used at Netflix and Spotify as recommendation engine)	Spark	Analytics engine for big data, run large scale workloads of petabytes of data, build, deploy, apps across VMs and container			Ggplot2	Create aesthetically pleasing and elegant visualizations using R
				Neo4J	Graph database management platform				

[Top 15 Open-Source Data Science Tools to Learn \(and Use\) in 2024](#)

# PROJECT JUPYTER (JUPYTER)

## Historie:

- 2014 gegründet
- Non-Profit-Organisation
- Der Name *Jupyter* bezieht sich auf die drei wesentlichen Programmiersprachen:  
Julia (ju), Python (pyt) and R (r).
- Project Jupyter hat die Produkte *Jupyter Notebook*, *JupyterHub* und *JupyterLab* entwickelt.
- **Installationsanleitung:** <https://jupyter.org/install>



## Jupyter Notebook

- Webanwendung zum Erstellen und Bearbeiten von „notebooks“
- Ein jupyter-notebook-dokument ist ein JSON-Dokument
- Dateiendung .“ipynb“
- Unterstützt insbesondere die Programmiersprachen Python

# JUPYTER NOTEBOOK

## WAS VERSTEHT MAN UNTER „NOTEBOOK“?

- A **notebook** is a shareable document that combines computer code, plain language descriptions, data, rich visualizations like 3D models, charts, graphs and figures, and interactive controls. A notebook, along with an editor (like JupyterLab), provides a fast interactive environment for prototyping and explaining code, exploring and visualizing data, and sharing ideas with others.
- A notebook file on their computer
- The idea of combining computer code, explanatory text, images and more into the “notebook format”
- The “Jupyter Notebook” application, used to author and edit digital notebook files
- Jupyter’s .ipynb notebook file format (used to save your notebook files on your computer), which is interpreted by the nbformat software library
- The **Jupyter Notebook interface** is a Web-based application for authoring documents that combine live-code with narrative text, equations and visualizations.



# JUPYTER NOTEBOOK

## WAS VERSTEHT MAN UNTER „NOTEBOOK“?

Ein **Notebook** ist ein teilbares Dokument, das Computer-Code, einfache sprachliche Beschreibungen, Daten, reichhaltige Visualisierungen wie 3D-Modelle, Diagramme, Grafiken und Abbildungen sowie interaktive Steuerungen kombiniert. Ein Notebook bietet zusammen mit einem Editor (wie JupyterLab) eine schnelle, interaktive Umgebung zum Prototypisieren und Erklären von Code, zum Erkunden und Visualisieren von Daten sowie zum Teilen von Ideen mit anderen.



- Ein Notebook-Dateiformat auf ihrem Computer
- Die Idee, Computer-Code, erläuternde Texte, Bilder und mehr in das „Notebook-Format“ zu integrieren
- Die „Jupyter Notebook“-Anwendung, die zum Erstellen und Bearbeiten von digitalen Notebook-Dateien verwendet wird
- Jupyers .ipynb-Notebook-Dateiformat (wird verwendet, um deine Notebook-Dateien auf deinem Computer zu speichern), das von der nbformat-Softwarebibliothek interpretiert wird

# JUPYTER NOTEBOOK

Die **Jupyter Notebook-Oberfläche** ist eine webbasierte Anwendung zum Erstellen von Dokumenten, die Live-Code mit erläuterndem Text, Gleichungen und Visualisierungen kombiniert.

The screenshot shows the Jupyter Notebook interface. At the top, there's a toolbar with File, Edit, View, Run, Kernel, Settings, Help, and a language icon (Python 3). Below the toolbar, the title bar says "jupyter Running Code Last Checkpoint: 10 months ago". The main area is titled "Running Code". It contains text about the notebook being an interactive environment for writing and running code in Python. It shows two code cells:

```
[1]: a = 10
[2]: print(a)
10
```

Below the cells, it says there are two other keyboard shortcuts for running code: Alt-Enter and Ctrl-Enter. A section titled "Managing the Kernel" explains that code is run in a separate process called the Kernel. It shows a third code cell:

```
[3]: import time
time.sleep(10)
```

Text at the bottom of this cell says: "If the Kernel dies you will be prompted to restart it. Here we call the low-level system libc routine with the wrong argument via ctypes to segfault the Python interpreter."

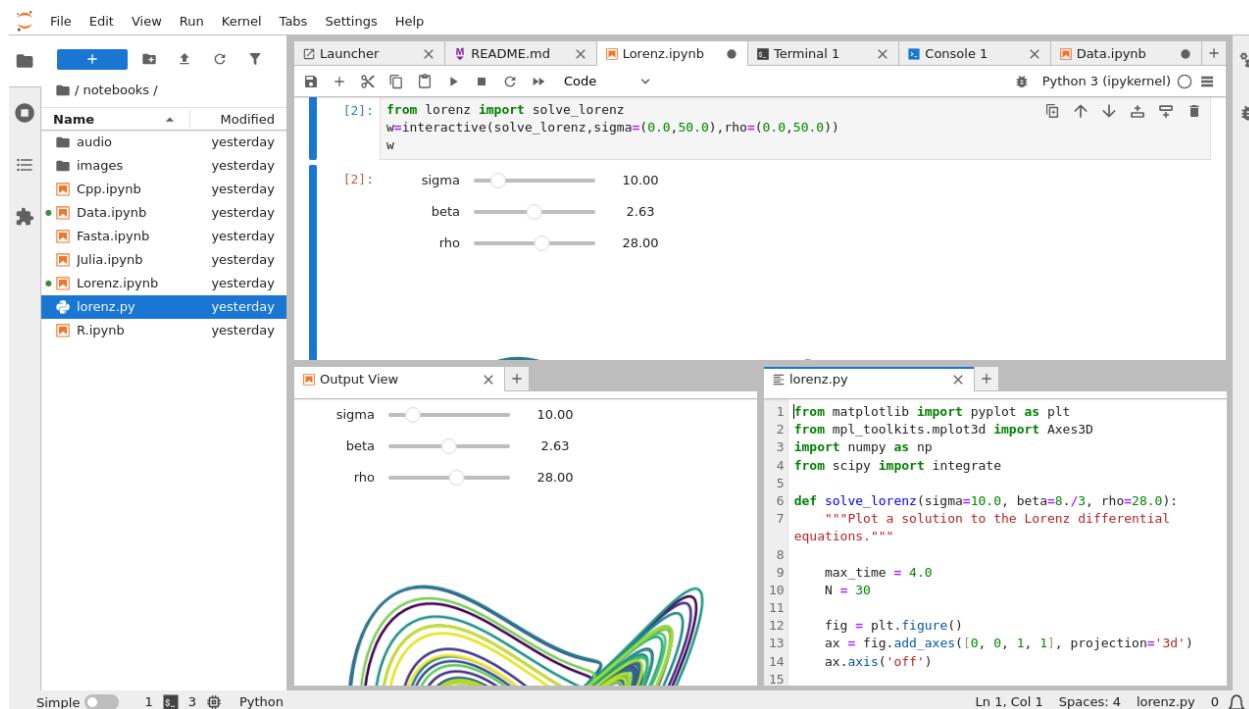


<https://jupyter-notebook.readthedocs.io/en/latest/>

# JUPYTER LAB

## Features:

- Nachfolgeprodukt des Jupyter Notebook



## 3 Schritte um Jupyter Lab nutzen zu können

### Python – PIP – Jupyter Lab



#### 1. Python

— **Installationsanleitung:**

<https://realpython.com/installing-python/>

#### 2. PIP Package Installer

— **Installationsanleitung:**

<https://pip.pypa.io/en/stable/>

#### 3. Jupyter Lab

— **Installationsanleitung:**

<https://jupyter.org/install>



## – Historie:

- 20.02.1991 erschienen
- Ursprünglich von Guido von Rossum in Amsterdam entwickelt
- Nachfolger der Programmier-Lehssprache ABC
- Offenes, gemeinschaftsbasiertes Entwicklungsmodell
- Wird durch die Python Software Foundation gestützt
- Python Software Foundation License

## – Features:

- Aktuelle Version 3.13.2
- Interpretierte, höhere Programmiersprache
- Anspruch gut lesbar zu sein und einen knappen Programmierstil zu fördern
- Unterstützt mehrere Programmierparadigmen (objektorientiert, prozedural, funktional)
- Bietet dynamische Typisierung und wird oft als Skriptsprache genutzt
- Plattformunabhängig

**Installationsanleitung:** <https://realpython.com/installing-python/>



# PIP - PACKAGE INSTALLER FOR PYTHON

## Features:

- Paketmanager (zum Installieren und Verwalten von Paketen)
- Aktuelle Version 25.01
- Rekursives Akronym für „pip installs packages“
- Am weitesten verbreitet, Alternativen sind Conda oder Pipenv
- Umfasst mehr als 380.000 Projekte (<https://pypi.org>)



pip is the package installer for Python.

**Installationsanleitung:** <https://pip.pypa.io/en/stable/>

## Usefull commands for pip:

```
pip install <Paket>
pip uninstall <Paket>
pip freeze / pip freeze > requirements.txt
pip -h / pip --help
```

## Data Science – Tools

### 20 Minuten Zeit



#### Aufgabe:

1. Python auf dem Rechner installieren  
<https://realpython.com/installing-python/>
2. PIP – Package Installer für Python installieren  
<https://pip.pypa.io/en/stable/>
3. Jupyter Lab auf dem Rechner installieren  
<https://jupyter.org/install>



## Weiterführende Links

### Python und Jupyter



#### Python

- Windows Downloads: [Python Releases for Windows | Python.org](#)
- Python Online Tutorial: [The Python Tutorial — Python 3.13.2 documentation](#)
- Python Versions Doku: [3.13.2 Documentation](#)
- Whats new in version: [What's New In Python 3.13 — Python 3.13.2 documentation](#)
- Using python on windows: [4. Using Python on Windows — Python 3.13.2 documentation](#)

#### Jupyter

- Jupyter: [Project Jupyter | Home](#)
- What ist jupyter: [What is Jupyter? — Jupyter Documentation 4.1.1 alpha documentation](#)
- Jupyter Doku: [Project Jupyter Documentation — Jupyter Documentation 4.1.1 alpha documentation](#)
- Jupyter lab doku: [JupyterLab Documentation — JupyterLab 4.3.6 documentation](#)

## Beliebte packages

### pip command



Package	Pip Command	Link
Jupyter lab	pip install jupyterlab	<a href="https://docs.jupyter.org/en/latest/">https://docs.jupyter.org/en/latest/</a>
Jupyter notebook	pip install notebook	
Jupyter Voila	pip install voila	
pandas	pip install pandas	<a href="https://pandas.pydata.org/docs/user_guide/index.html">https://pandas.pydata.org/docs/user_guide/index.html</a>
matplotlib	pip install matplotlib	<a href="https://matplotlib.org/stable/users/index.html">https://matplotlib.org/stable/users/index.html</a>
numpy	pip install numpy	<a href="https://numpy.org/doc/stable/user/basics.html">https://numpy.org/doc/stable/user/basics.html</a>
sklearn	pip install scikit-learn	<a href="https://scikit-learn.org/1.6/index.html">https://scikit-learn.org/1.6/index.html</a>
xgboost	pip install xgboost	<a href="https://xgboost.readthedocs.io/en/latest/index.html">https://xgboost.readthedocs.io/en/latest/index.html</a>
lightgbm	Pip install lightgbm	<a href="https://lightgbm.readthedocs.io/en/latest/index.html">https://lightgbm.readthedocs.io/en/latest/index.html</a>

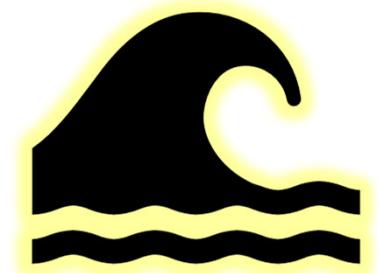
## Anwendungsbeispiel

### Analyse von Passagierdaten der Titanic

#### Szenario:

Das Sinken der Titanic ist eines der berüchtigtesten Schiffsunglücke in der Geschichte. Am 15. April 1912, während ihrer Jungfernreise, sank die weithin als „unsinkbar“ geltende RMS Titanic, nachdem sie mit einem Eisberg kollidiert war. Leider gab es nicht genug Rettungsboote für alle an Bord, was zum Tod von 1502 von 2224 Passagieren und Besatzungsmitgliedern führte.

Obwohl ein gewisses Element des Glücks beim Überleben eine Rolle spielte, scheint es, dass einige Personengruppen eine höhere Überlebenswahrscheinlichkeit hatten als andere.



**Ziel:** Basierend auf den Passagierdaten ein **Vorhersagemodell zu erstellen**, das die Frage beantwortet:

„Welche Art von Menschen hatte eine höhere Wahrscheinlichkeit zu überleben?“

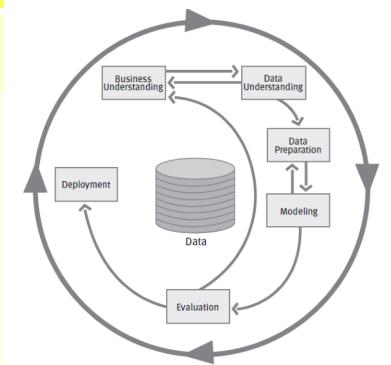
Vorgehen nach dem CRISP-DM Prozess in 6 Schritten:

1. **Business Understanding** (Aufgabendefinition)
2. **Data Understanding** (Auswahl der relevanten Datenbestände)
3. **Data Preparation** (Datenaufbereitung)
4. **Modeling** (Auswahl und Anwendung von Modellen)
5. **Evaluation** (Bewertung und Interpretation der Ergebnisse)
6. **Deployment** (Anwendung der Ergebnisse)

# CRISP-DM

## PHASEN UND AKTIVITÄTEN

Aufgabendefinition	Datenverständnis	Datenaufbereitung	Modellierung	Bewertung	Anwendung
Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
<ul style="list-style-type: none"> <li>Bestimmung der betriebswirtschaftlichen Problemstellung</li> <li>Situationsbewertung</li> <li>Bestimmung analytischer Ziele</li> <li>Erstellung eines Projektplans</li> </ul>	<ul style="list-style-type: none"> <li>Daten sammeln</li> <li>Daten beschreiben</li> <li>Untersuchung der Daten</li> <li>Verifizierung der Datenqualität</li> </ul>	<ul style="list-style-type: none"> <li>Auswahl der Daten</li> <li>Bereinigung der Daten</li> <li>Transformation und Integration der Daten</li> <li>Datenformatierung</li> </ul>	<ul style="list-style-type: none"> <li>Auswahl des Modells</li> <li>Testmodell erstellen</li> <li>Modell erstellen</li> </ul>	<ul style="list-style-type: none"> <li>Bewertung des Modells</li> <li>Bewertung der Resultate</li> <li>Bewertung des Prozesses</li> <li>Nächste Schritte festlegen</li> </ul>	<ul style="list-style-type: none"> <li>Zusammenfassen der Bericht und Präsentation der Ergebnisse</li> <li>Implementierungsstrategie planen</li> <li>Überwachung der Gültigkeit der Modelle planen</li> </ul>



Quelle Grafik: Chapman, P., et al., 2000, CRISP-DM 1.0, S. 10.

## Unternehmen:

- Gegründet: 2010
- Sitz: San Francisco, USA
- Branche: Data Science
- Gehört seit 2017 zur Google LLC
- **Website:** <https://www.kaggle.com/>

## Produkte:

- **Kaggle Online-Community für Data Science**
- Organisation von Data Science Competitions
- Austausch von Data Sets, Modellen, Code
- Diskussionsforum und Lernplattform für

≡ kaggle

+ Create

Home

Competitions

Datasets

Models

Code

Discussions

Learn

More

# KAGGLE - COMPETITIONS

## Competitions

Grow your data science skills by competing in our exciting competitions. Find help in the [documentation](#) or learn about [Community Competitions](#).

[Host a Competition](#) [Your Work](#)



Search competitions

Filters

All Competitions

Everything, past & present

Featured

Premier challenges with prizes

Getting Started

Approachable ML fundamentals

Research

Scientific and scholarly challenges

Community

Created by fellow Kagglers

Playground

Fun practice problems

### Getting Started

[See all](#)

Competitions with approachable ML fundamentals.



#### Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titani...

Getting Started

16428 Teams

**Knowledge**

Ongoing



#### Housing Prices Competition for Kaggle...

Apply what you learned in the Machine ...

Getting Started

7077 Teams

**Knowledge**

Ongoing



#### House Prices - Advanced Regression Techniques

Predict sales prices and practice feature...

Getting Started

4280 Teams

**Knowledge**



#### Spaceship Titanic

Predict which passengers are transport...

Getting Started

1975 Teams

**Knowledge**



<https://www.kaggle.com/competitions>

# KAGGLE – DATA SETS

## Datasets

Explore, analyze, and share quality data. [Learn more](#) about data types, creating, and collaborating.

+ New Dataset
Your Work

Filters

All datasets
Computer Science
Education
Classification
Computer Vision
NLP
Data Visualization
Pre-Trained Model

Trending Datasets
See All



**Users vs bots classification** · :  
Aleksei Zagorskii · Updated 4 days ago  
Usability 10.0 · 112 kB  
1 File (CSV)

▲ 18



**Gold Insights Dataset (2020–2023)** · :  
Vesela Gencheva · Updated 25 days ago  
Usability 10.0 · 6 kB  
5 Files (CSV)

▲ 20



**Fecom Inc. (e-Com Marketplace Orders Data)** · :  
Cem Eragan · Updated a month ago  
Usability 9.4 · 28 MB  
8 Files (CSV)

▲ 17



**Stock Market Simulation Dataset** · :  
Samay Ashar · Updated 22 days ago  
Usability 10.0 · 90 kB  
1 File (CSV)

▲ 12



<https://www.kaggle.com/datasets>

# Jupyter Notebook vs. Kaggle Notebook

## Jupyter Notebook

### Vorteile:

- Flexibel und unterstützt viele Programmiersprachen.
- Interaktive Entwicklung und Visualisierungen.
- Offline nutzbar.

### Nachteile:

- Komplexe Installation und Verwaltung.
- Keine Echtzeit-Kollaboration.
- Leistung bei großen Datenmengen begrenzt.

## Kaggle Notebook

### Vorteile:

- Einfach zu verwenden, keine Installation erforderlich.
- Kostenlose Cloud-Ressourcen (GPU/TPU).
- Einfache Kollaboration und Zugang zu Datensätzen.

### Nachteile:

- Abhängig von Internetverbindung.
- Begrenzte Rechenressourcen.
- Weniger Kontrolle über die Umgebung.

• **Jupyter Notebooks** bieten mehr Flexibilität und Kontrolle über die Umgebung und können lokal ausgeführt werden, was jedoch mehr Setup erfordert und weniger kollaborativ ist.

• **Kaggle Notebooks** sind ideal für einfache Nutzung, Zusammenarbeit und Cloud-basierte Arbeit, bieten jedoch weniger Kontrolle und haben Einschränkungen bei den Rechenressourcen und der Anpassbarkeit.

## Data Science – Tools

20-30 Minuten Zeit



### Aufgabe:

1. Öffne Jupyter Lab (bspw. über MS Power Shell, Befehl `jupyter lab`)
2. Lege einen neuen Ordner „Housing“ an
3. Lade das Notebook „Cali\_Housing.ipynb“ in den Ordner
4. Lade die Datei „cali\_data.csv“ in den Ordner
5. Schaue dir die Beschreibungsseite zu den Daten auf der Platform kaggel an:  
<https://www.kaggle.com/datasets/camnugent/california-housing-prices/data>
6. Gehe die einzelnen Kapitel des Notebooks mit den zum kalifornischen Hausmarkt durch.

