

TECHNICAL APPLICATIONS AND DATA MANAGEMENT. SS 2022.

VORLESUNG 2

20.09.2022

MÜNCHEN

STUDIENGANG
SUSTAINABILITY
MANAGEMENT &
LEADERSHIP SOWIE
MEDIEN &
KOMMUNIKATION.

GEPLANTE ROADMAP DER VORLESUNG.

ROADMAP	WAS HABEN WIR VOR?
Vorlesung 1	Übersicht und Einführung
Vorlesung 2	Einführung Data Science und Data Science Workflow, Detaillierung Data Engineering
Vorlesung 3 und Vorlesung 4	Deskriptive und explorative Datenanalyse und Vertiefung anhand Case Study
	Vertiefung Datenanalyse anhand Case Study
Vorlesung 5	Aufgabenstellung Data Science, Übersicht und Einführung Machine Learning, unüberwachtes Lernen
Vorlesung 6 und Vorlesung 7	Überwachtes Lernen
	Vertiefung überwachtes Lernen anhand Case Study
Vorlesung 8 und Vorlesung 9	Neuronale Netze und Convolutional Neural Networks (CNN)
	Vertiefung CNN anhand Case Study, Aufgabenstellung AI
Vorlesung 10	Rekurrente Neuronale Netze
Vorlesung 11	Generative AI
Vorlesung 12	Ausblick
Vorlesung 13	„Fragestunde“

Folien der bisherigen Vorlesung verfügbar unter [Link](#)



AGENDA

1. Erklärung Data Science
2. Vorgehensweise Use Case Data Science
3. Fallbeispiel inklusive Data Engineering



1. WAS IST DATA SCIENCE?

what my friends think I do



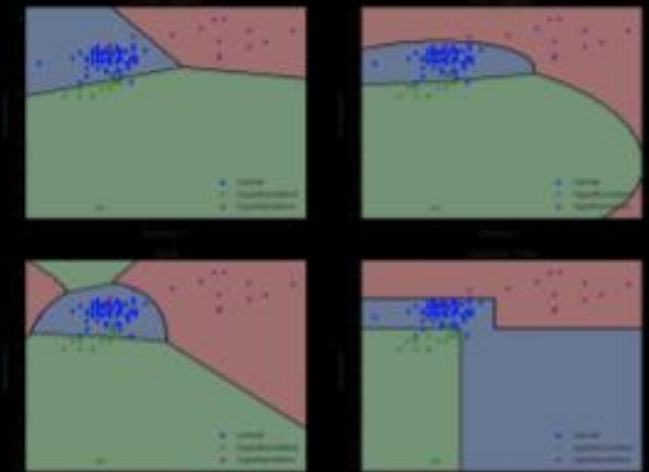
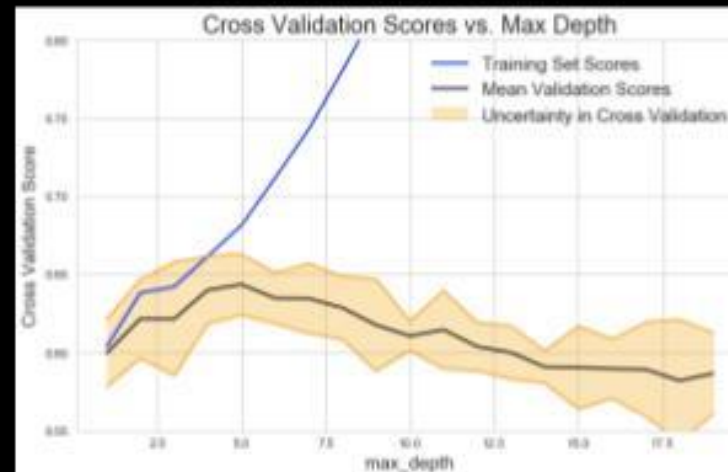
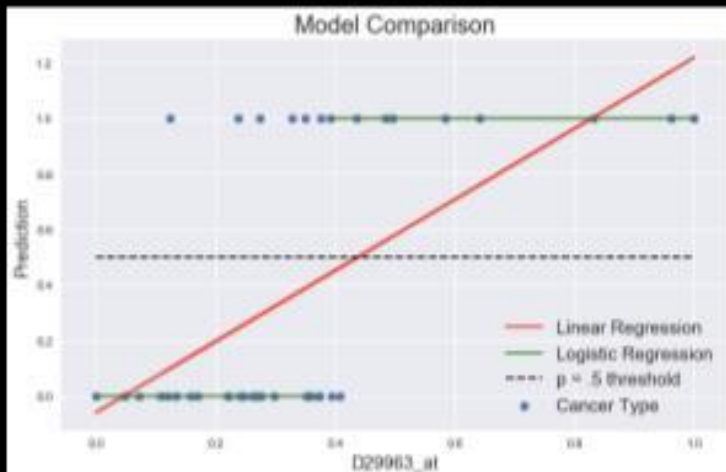
what my family thinks I do



what society thinks I do

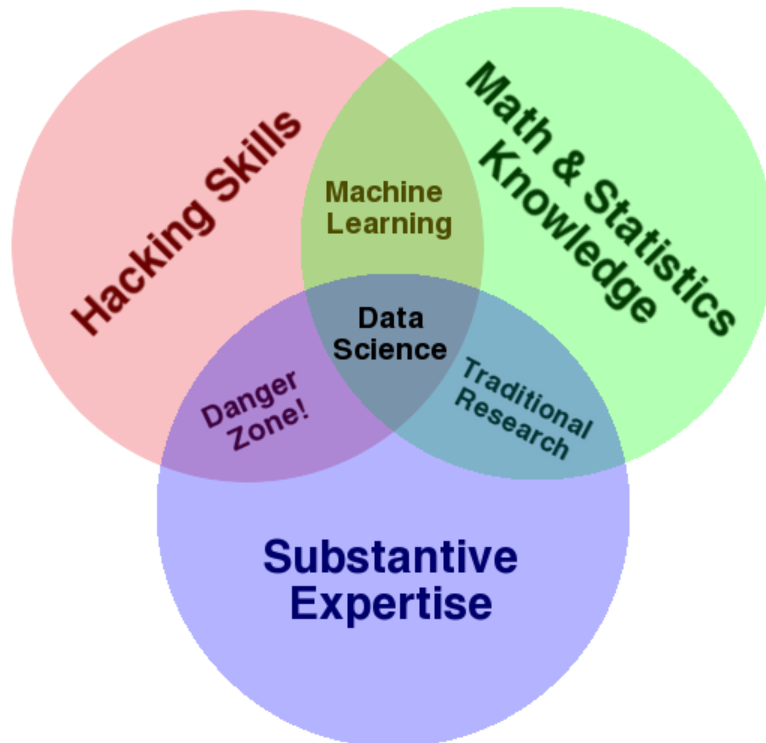


what I actually (will) do in Data Science 1



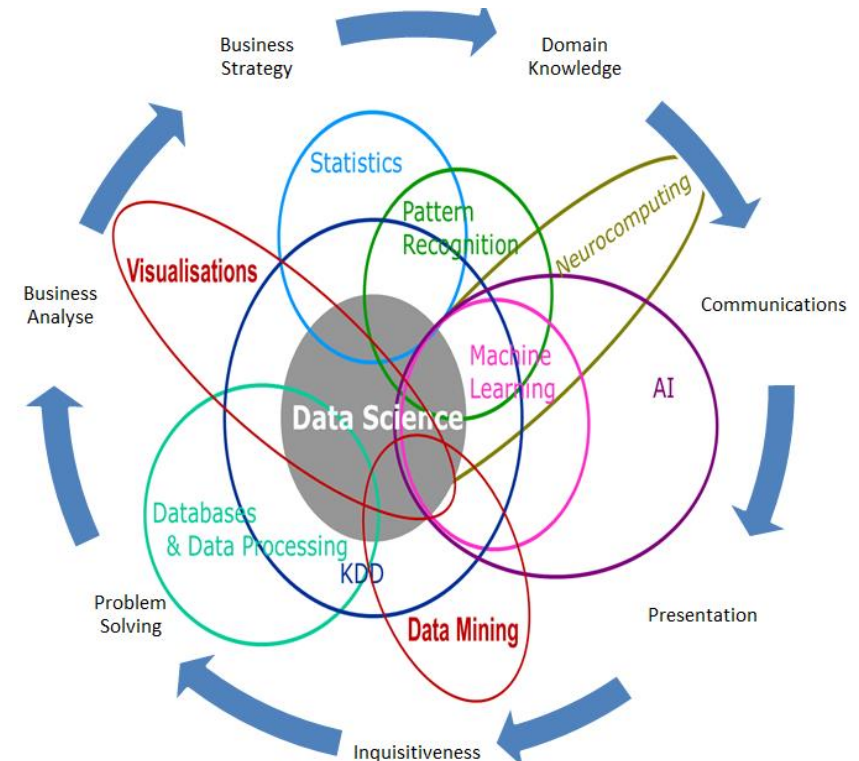
VERSTÄNDNIS FÜR BEGRIFF SOWIE UMFANG DATA SCIENCE HAT SICH STARK GEÄNDERT IN DEN LETZTEN JAHREN.

2010



Quelle: Drew Conway 2010, verfügbar unter: [Link](#)

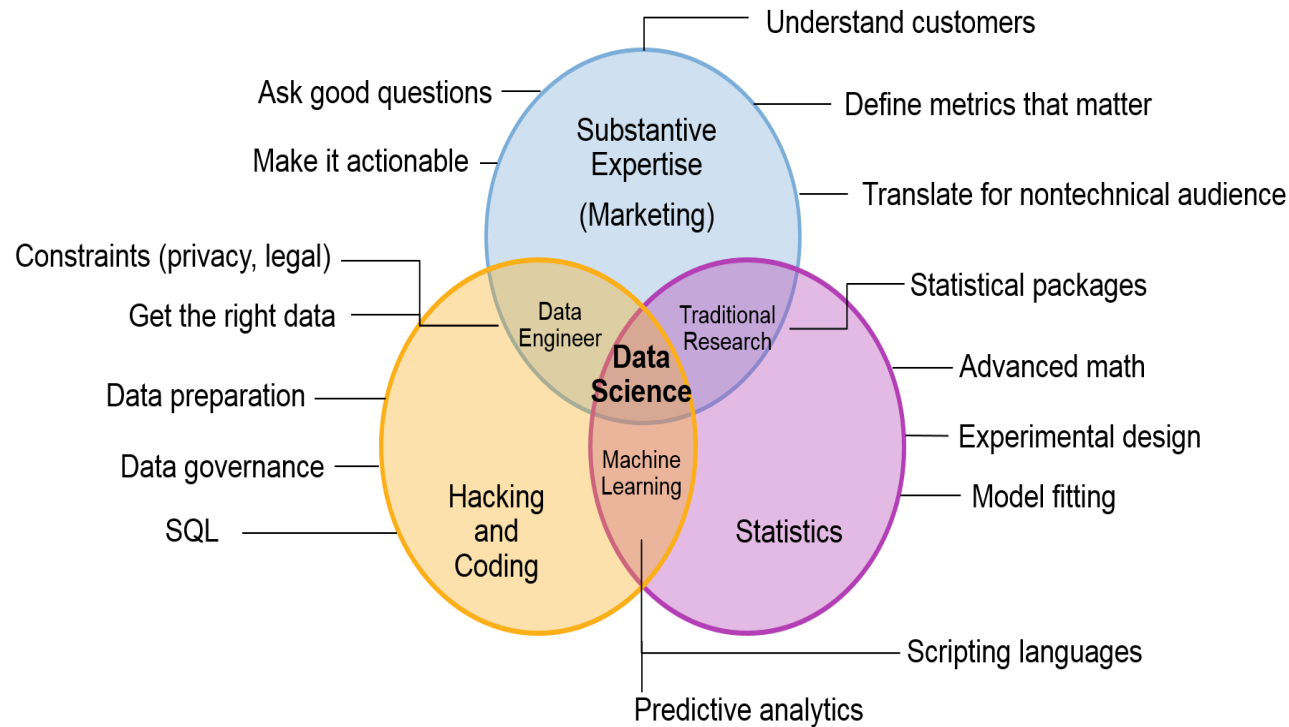
2013



Quelle: B. Tierney, 2013, verfügbar unter: [Link](#)

VERSTÄNDNIS FÜR BEGRIFF SOWIE UMFANG DATA SCIENCE HAT SICH STARK GEÄNDERT IN LETZTEN JAHREN.

2016



Quelle: Gartner 2016, verfügbar unter: [Link](#)

2019



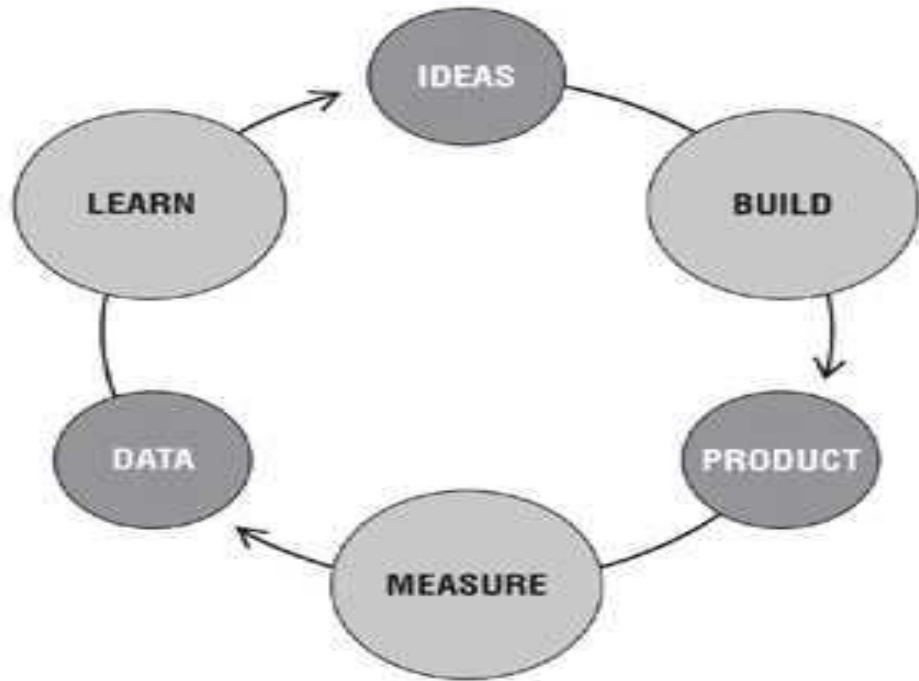
Quelle: NIST big data workgroup, 2019, verfügbar unter: [Link](#)



2. VORGEHENSWEISE BEI EINEM DATA SCIENCE USE CASE

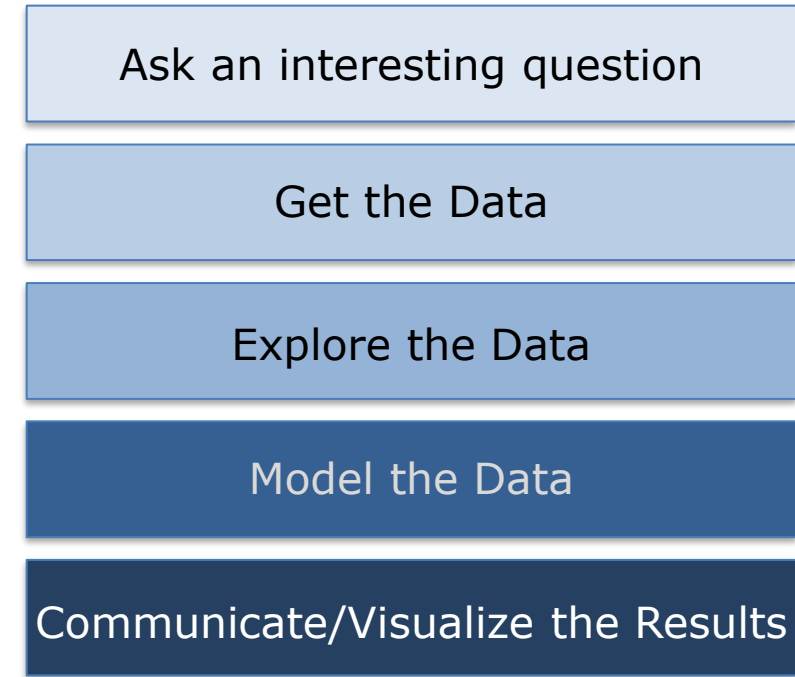
VORGEHENSWEISE USE CASE DATA SCIENCE.

BUILD-MEASURE-LEARN FEEDBACK LOOP



Vorgehensweise einer „data-driven company“

Quelle: E. Ries, „The Lean Start-up“, 2011

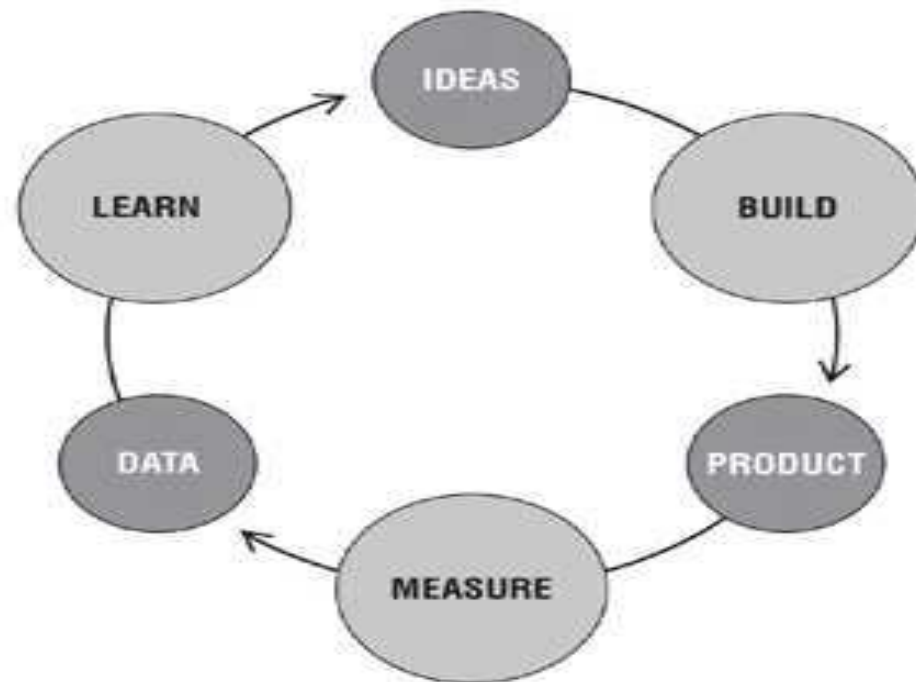


Generische Vorgehensweise Data Science

Quelle: Protopapas, Rader, Tanner, CS109 Data Science, 2020, [Link](#)

DER BUILD-MEASURE-LEARN-FEEDBACK LOOP WIRD SEHR OFT IN STARTUPS, ABER AUCH ANDEREN DIGITALEN FIRMEN EINGESETZT.

BUILD-MEASURE-LEARN FEEDBACK LOOP



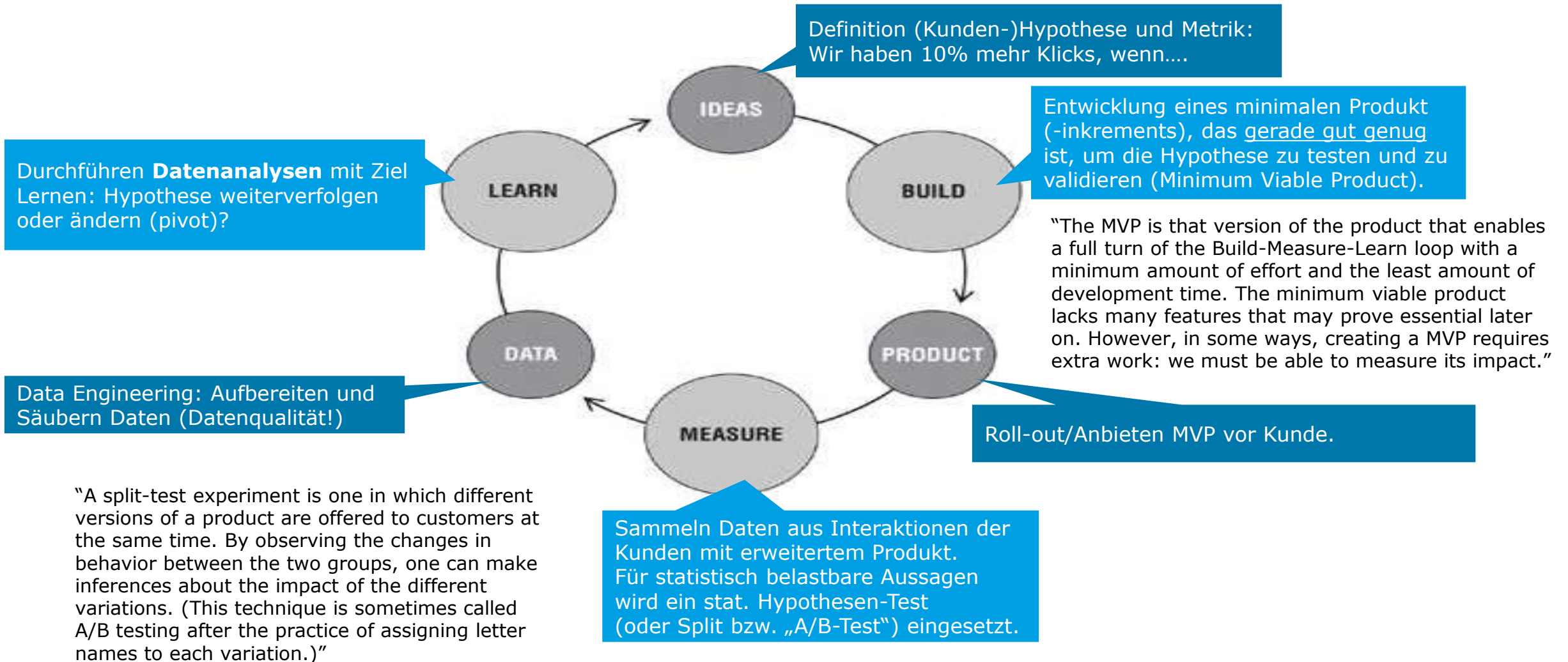
Minimize *TOTAL* time through the loop

“The fundamental activity of a startup is to turn ideas into products, measure how customers respond, and then learn whether to pivot or persevere. All successful startup processes should be geared to accelerate that feedback loop”.

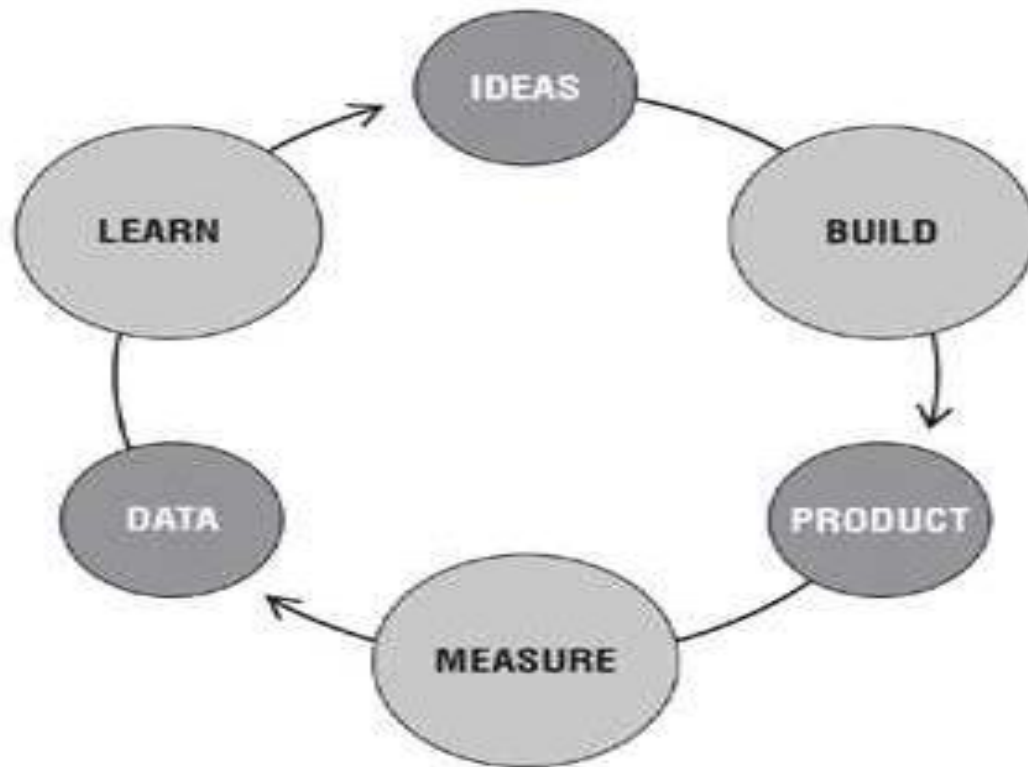
“Startups exist not just to make stuff, make money, or even serve customers. They exist to **learn how to build a sustainable business**. This **learning** can be **validated** scientifically by running **frequent experiments** that allow entrepreneurs to test each element of their vision.”

Iterativer Prozess mit dem Ziel kontinuierliches Lernen

DETAILLIERUNG BUILD-MEASURE-LEARN FEEDBACK LOOP.



VERTIEFUNG BUILD-MEASURE-LEARN FEEDBACK LOOP.



Sie sind verantwortlicher Manager eines Online-Shops/ ...

- Wofür wären Kunden bereit (mehr) zu zahlen? Welche Kundenhypothese haben Sie?
- Was wäre Ihr minimales Produkt (MVP), um diese Hypothese zu testen?
- Was wären (beispielhafte) Metriken für Messen dieser Hypothese?

Am Beispiel WhatsApp:

- Hypothese: Versenden beliebiger Handy-Nachrichten per Internet statt SMS/ MMS liefert Mehrwert für Kunden (für den Kunden auch zahlen¹ würden).
- MVP: eine App, die nur Text versenden kann (Roll-out erst für iPhone um Aufwand zu sparen und mehr Nutzer).
- Metriken: Anzahl Downloads für App, Anzahl versendeter Nachrichten, Anzahl zahlender Kunden, Anzahl Power-User (Kunden mit mehr als x Nachrichten), ...

Definieren Sie einen Durchlauf des Loop für einen Online-Shop oder eine andere digitale Firma

DETAILLIERUNG VORGEHENSWEISE DATA SCIENCE.

Ask an interesting question

Get the Data

Explore the Data

Model the Data

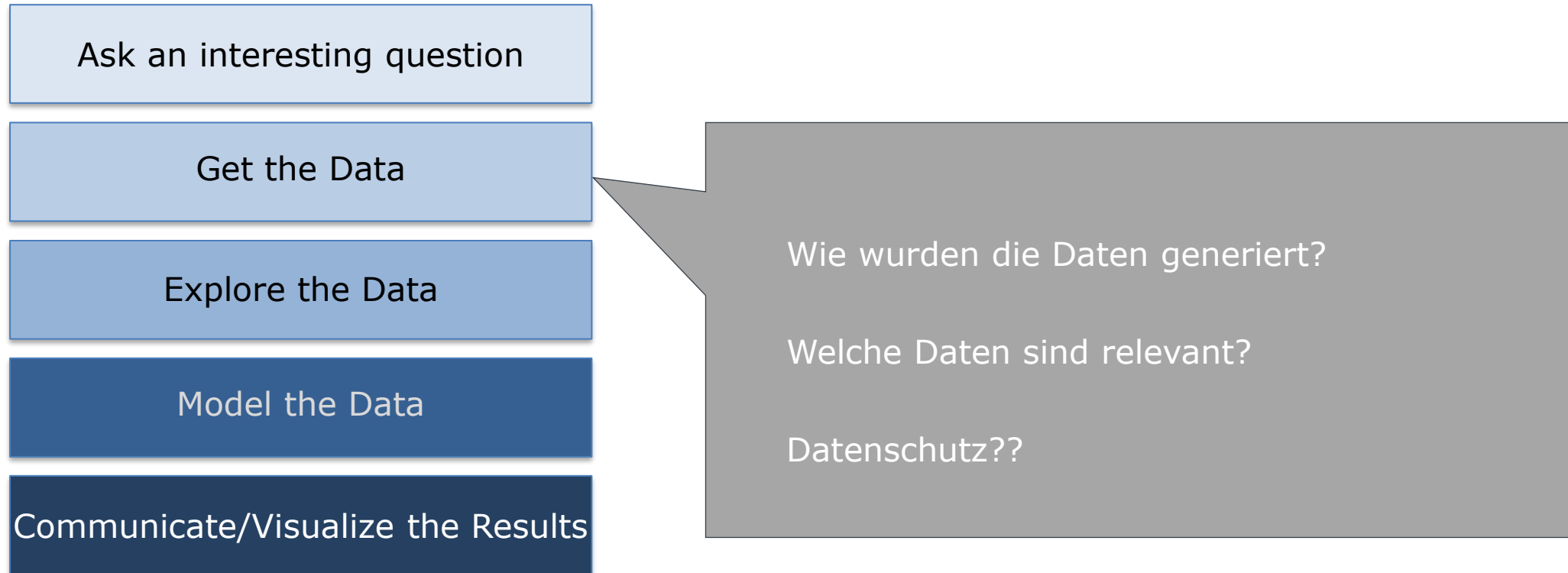
Communicate/Visualize the Results

Was ist die Fragestellung?

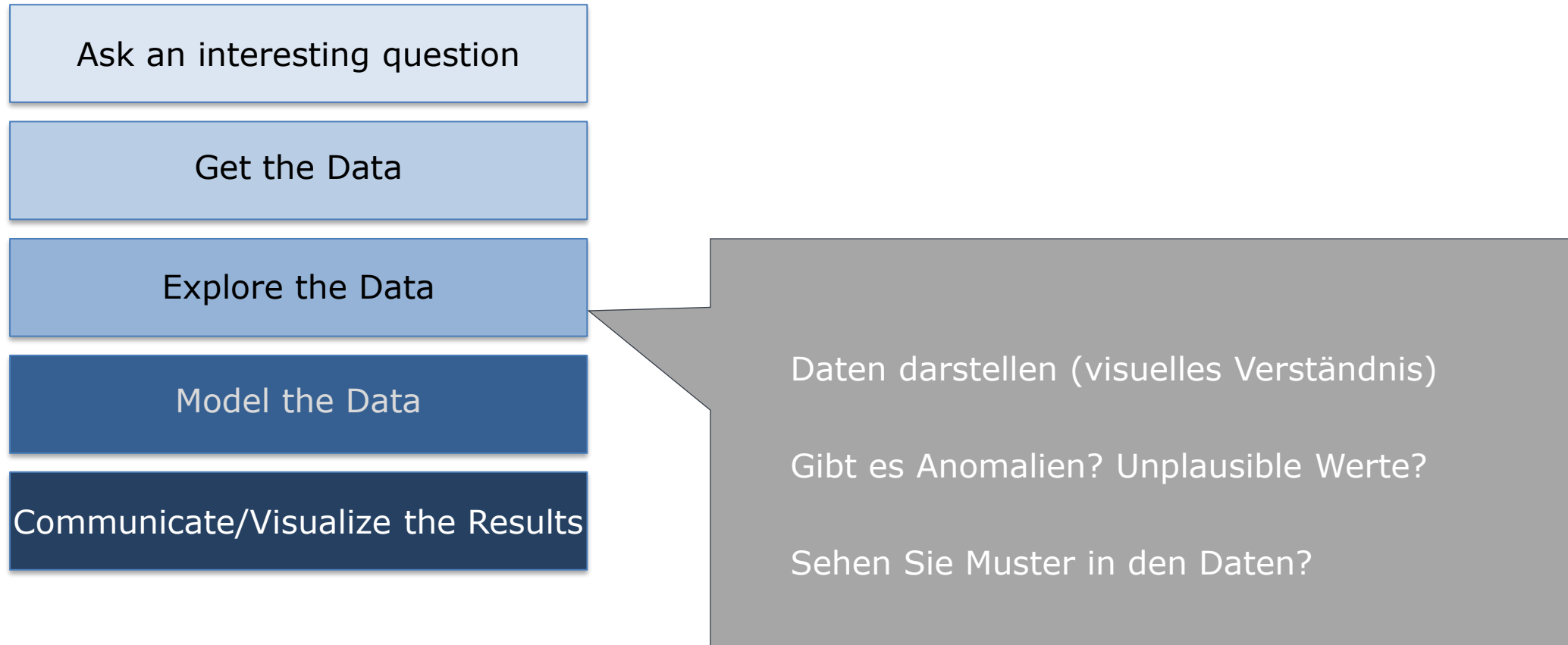
Was würde ich tun, wenn ich alle verfügbare Daten hätte?

Was möchte ich abschätzen/ vorhersagen?

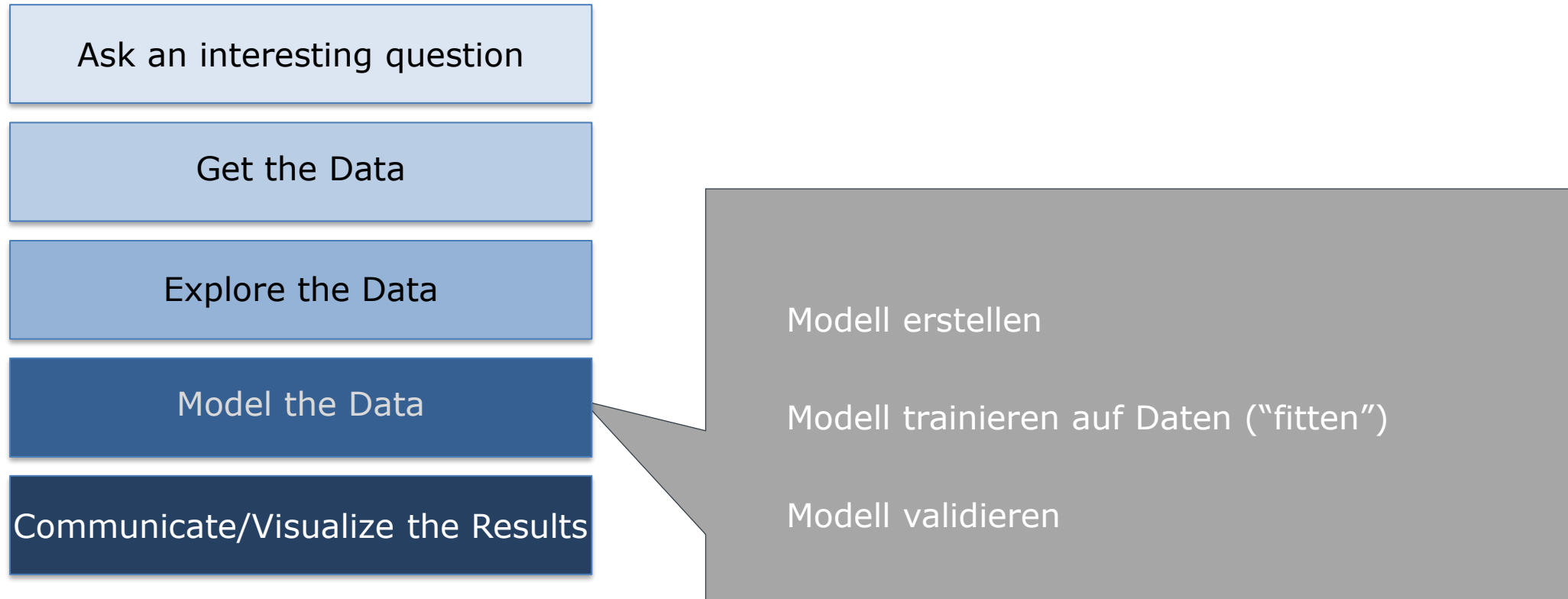
DETAILLIERUNG VORGEHENSWEISE DATA SCIENCE.



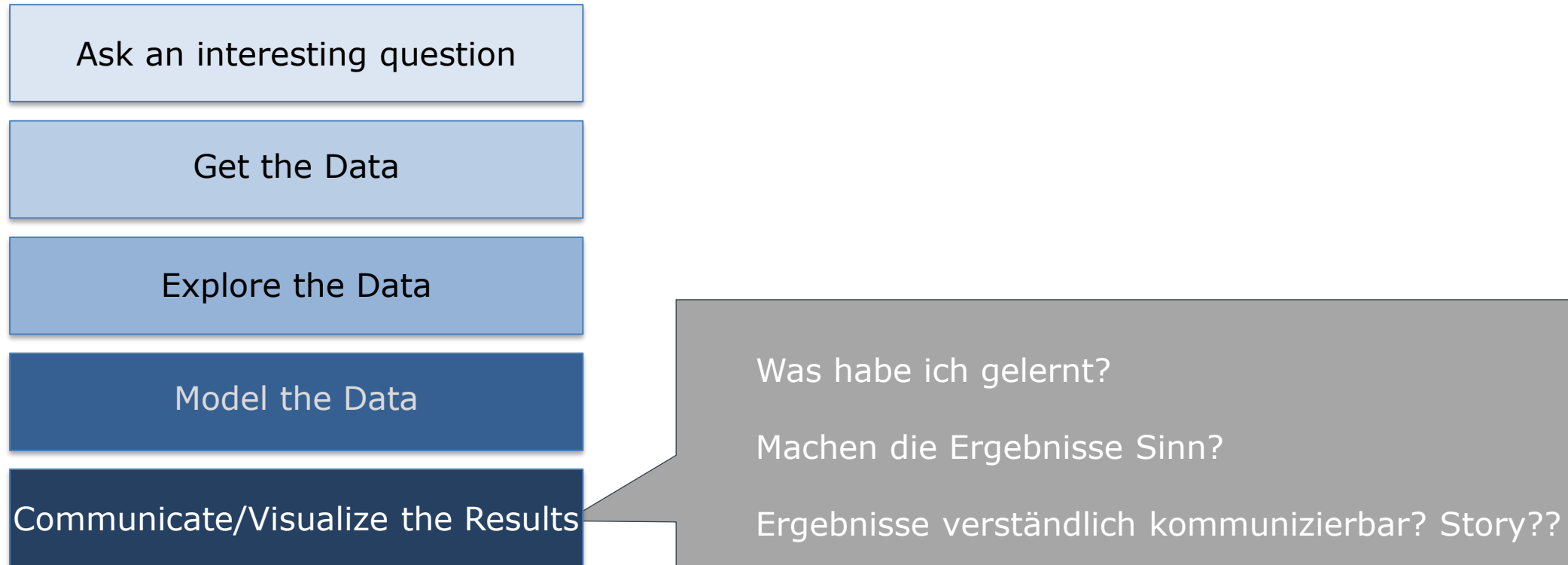
DETAILLIERUNG VORGEHENSWEISE DATA SCIENCE.



DETAILLIERUNG VORGEHENSWEISE DATA SCIENCE.



DETAILLIERUNG VORGEHENSWEISE DATA SCIENCE.

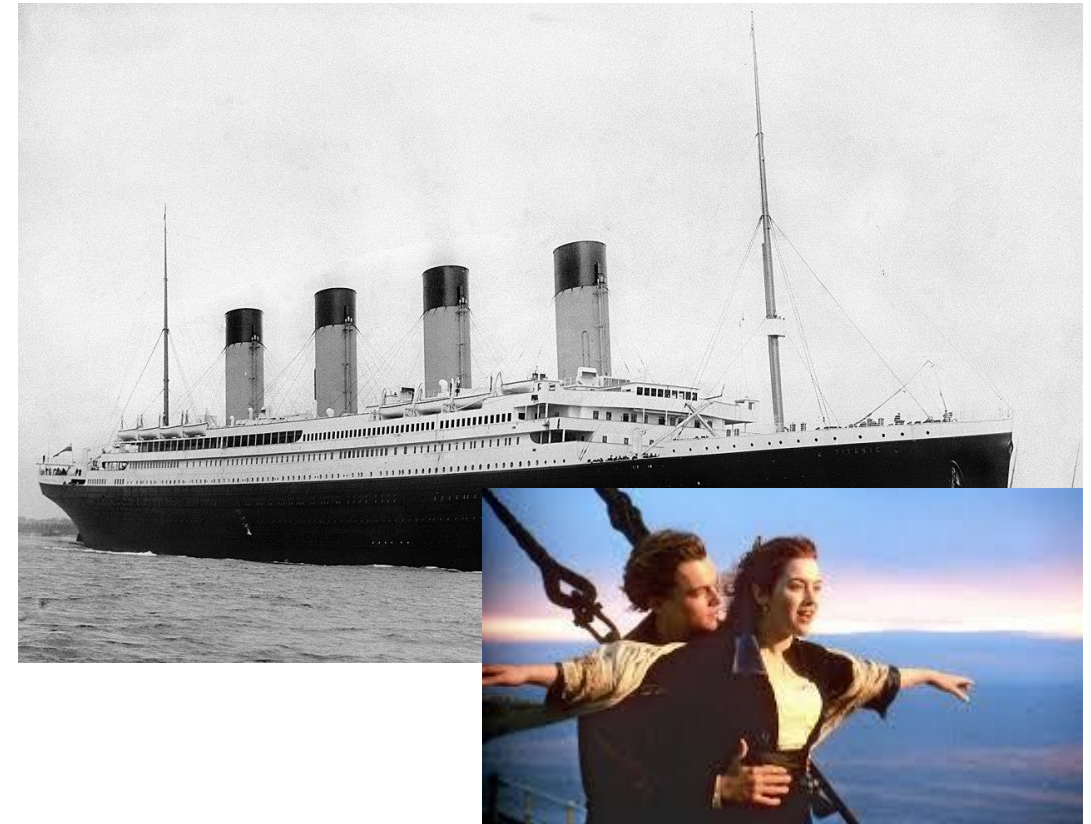




2. FALLBEISPIEL

ÜBERSICHT DATA SCIENCE AM FALLBEISPIEL TITANIC.

- Passagierliste Titanic ist beliebter Datensatz für Data Science:
 - Kleiner Datensatz (1310 Zeilen à 14 Spalten)
 - Deckt ganzen Workflow inkl. üblicher Probleme ab
 - Fragestellung einfach verständlich und interessant
- Im Notebook zur Vorlesung wird folgendes gemacht:
 - Import/ Laden der Daten
 - Data Engineering:
 - Daten säubern und aufbereiten
 - neue Features erstellen
 - Univariate Datenanalysen (Analyse eines Features)
 - Multivariate Datenanalysen (Analyse mehrerer Features)
 - Annäherung Zielvariable per manueller Optimierung

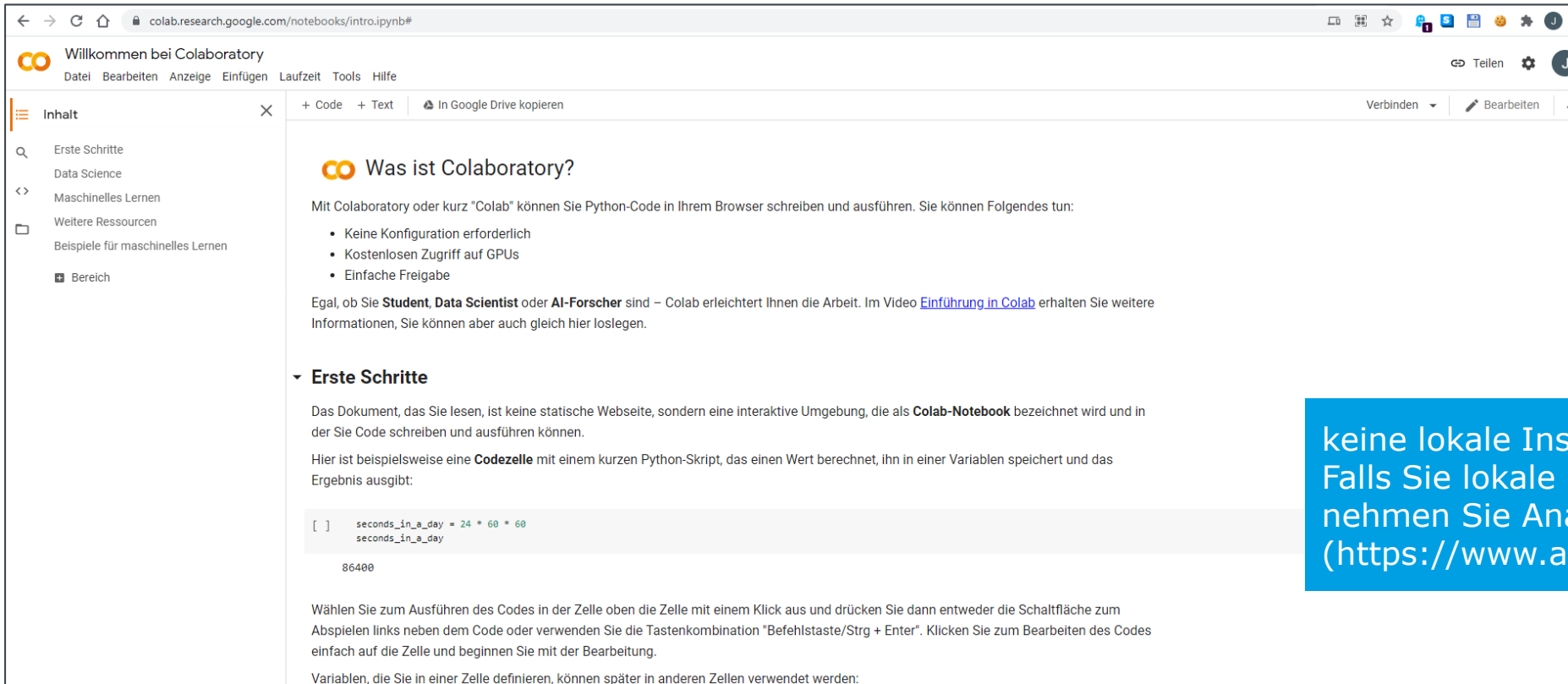


▶ Das Titanic-Notebook ist sehr detailliert, wir werden uns in der heutigen Vorlesung nur die wichtigsten Sachen ansehen

ALS PROGRAMMIERSPRACHE WERDEN WIR PYTHON EINSETZEN.

- Einfach zu erlernen und zu benutzen.
- Kostenfrei verfügbar.
- Sehr viele kostenfreie, leistungsfähige Bibliotheken, die viel Programmierarbeit abnehmen.
- Flexibel und weit einsetzbar.
- Sehr häufig für Data Science und Künstliche Intelligenz eingesetzt.
- Sehr viele frei verfügbare Beispiele und Tutorials.

IM RAHMEN DER VORLESUNG WERDEN SIE PROGRAMMIEREN, EMPFEHLUNG PROGRAMMIERUMGEBUNG IST GOOGLE COLAB.



Willkommen bei Colaboratory

Datei Bearbeiten Anzeige Einfügen Laufzeit Tools Hilfe

Inhalt

- Erste Schritte
- Data Science
- Maschinelles Lernen
- Weitere Ressourcen
- Beispiele für maschinelles Lernen
- Bereich

+ Code + Text In Google Drive kopieren

Verbinden Bearbeiten

Was ist Colaboratory?

Mit Colaboratory oder kurz "Colab" können Sie Python-Code in Ihrem Browser schreiben und ausführen. Sie können Folgendes tun:

- Keine Konfiguration erforderlich
- Kostenlosen Zugriff auf GPUs
- Einfache Freigabe

Egal, ob Sie **Student**, **Data Scientist** oder **AI-Forscher** sind – Colab erleichtert Ihnen die Arbeit. Im Video [Einführung in Colab](#) erhalten Sie weitere Informationen, Sie können aber auch gleich hier loslegen.

Erste Schritte

Das Dokument, das Sie lesen, ist keine statische Webseite, sondern eine interaktive Umgebung, die als **Colab-Notebook** bezeichnet wird und in der Sie Code schreiben und ausführen können.

Hier ist beispielsweise eine **Codezelle** mit einem kurzen Python-Skript, das einen Wert berechnet, ihn in einer Variablen speichert und das Ergebnis ausgibt:

```
[ ] seconds_in_a_day = 24 * 60 * 60
seconds_in_a_day
```

86400

Wählen Sie zum Ausführen des Codes in der Zelle oben die Zelle mit einem Klick aus und drücken Sie dann entweder die Schaltfläche zum Abspielen links neben dem Code oder verwenden Sie die Tastenkombination "Befehlstaste/Strg + Enter". Klicken Sie zum Bearbeiten des Codes einfach auf die Zelle und beginnen Sie mit der Bearbeitung.

Variablen, die Sie in einer Zelle definieren, können später in anderen Zellen verwendet werden:

keine lokale Installation notwendig.
Falls Sie lokale Installation bevorzugen,
nehmen Sie Anaconda
(<https://www.anaconda.com/>)

<https://colab.research.google.com/notebooks/intro.ipynb#>

WIR SCHAUEN UNS DIE EINZELNEN SCHRITTE ANHAND EINES NOTEBOOKS AUF COLAB AN.

The screenshot shows a Google Colab notebook interface. The title bar indicates the notebook is titled "Titanic-Notebook-Update - WS2021/22". The left sidebar contains a table of contents with the following items:

- Titanic Notebook
- Schritt 1: Ask an interesting question
- Schritt 2: Get the Data**
- Schritt 3: Explore the Data
- Übersicht Daten
- Data Management
- Verbesserung Lesbarkeit
- Data Engineering
- Feature Engineering
- Visuelle Datenexploration/ Deskriptive Statistik
- Univariate Analysen
- Multivariate Analysen
- Schritt 4: Model the Data
- Lineare Regression
- Datenaufbereitung
- Training des Algorithmus und Optimierung
- Modellvalidierung
- Ausblick maschinelles Lernen
- Schritt 5: Und wie wendet man das an? Oder: was bringt das alles?
- Bereich

The main content area shows the following text and code:

Bei den ersten beiden Fragen untersuchen wir eine definierte Variable, eine sogenannte **univariate Datenanalyse**. Die Variablen werden auch Features genannt. Damit kann man erste Untersuchungen anstellen, für die meisten interessanten Fragestellungen muß man sich jedoch als ein Feature ansehen. Man spricht da von **multivariaten Datenanalysen**.

▼ Schritt 2: Get the Data

Nachdem wir uns die zu untersuchenden Fragestellungen definiert haben, beschaffen wir uns die Daten.

Für das Laden und Auswerten der Daten gibt es in Python viele Programmibibliotheken für Data Science oder AI, die uns die Detailarbeit abnehmen.

Am Beginn jedes Notebooks laden wir diese Libraries.

Lila markierter Text sind Standard-Befehle der Programmiersprache Python, Kommentare werden mit einer Raute eingeleitet.

Gute, kostenfreie Kurse in Python sind in den Literaturquellen in der Vorlesung angegeben; Sie benötigen diese aber eigentlich nicht für die Umfänge der Vorlesung.

```
[1] import pandas as pd # Importieren Standard-Library für das Bearbeiten und Laden von Daten ("Data Engineering").
import matplotlib.pyplot as plt # Standard-Library für das Plotten von Graphen.
import seaborn as sns # verschönert Matplotlib-Graphiken
import numpy as np # Standard-Library für Rechnen
```

Für das Titanic-Beispiel sind die Daten, die wir bearbeiten wollen, als CSV (Comma Separated Values)-Datei gespeichert.

Nachdem wir die Standard-Libraries geladen haben, laden wir die CSV-Datei mit den Titanic-Passagieren und ihren Daten in ein Standard-Datentyp der **Pandas-Library**, dem sogenannten DataFrame.

Dataframes ist ein sehr häufig genutztes Datenformat in Data Science und AI. Sie können sich das als große Tabelle mit Spalten für die einzelnen Daten vorstellen.

Da wir alles in der Cloud machen, müssen wir die Titanic.csv noch organisieren. Dafür gibt es zwei Möglichkeiten:

- Hochladen von Festplatte
- Laden von einer anderen Internetadresse.

Beide Beispiele sind unten angefügt, das einfachere ist von der Webadresse.

Was müssen Sie tun:

- Datei „Titanic Notebook“ runterladen aus Github.
- Anlegen eines Google-Accounts (falls Sie nicht schon haben).
- Auf Google Colab gehen: [Link](#)
- Hochladen der Datei in Google Colab.
- Starten!

HANDS ON DATA SCIENCE-FALLBEISPIEL

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

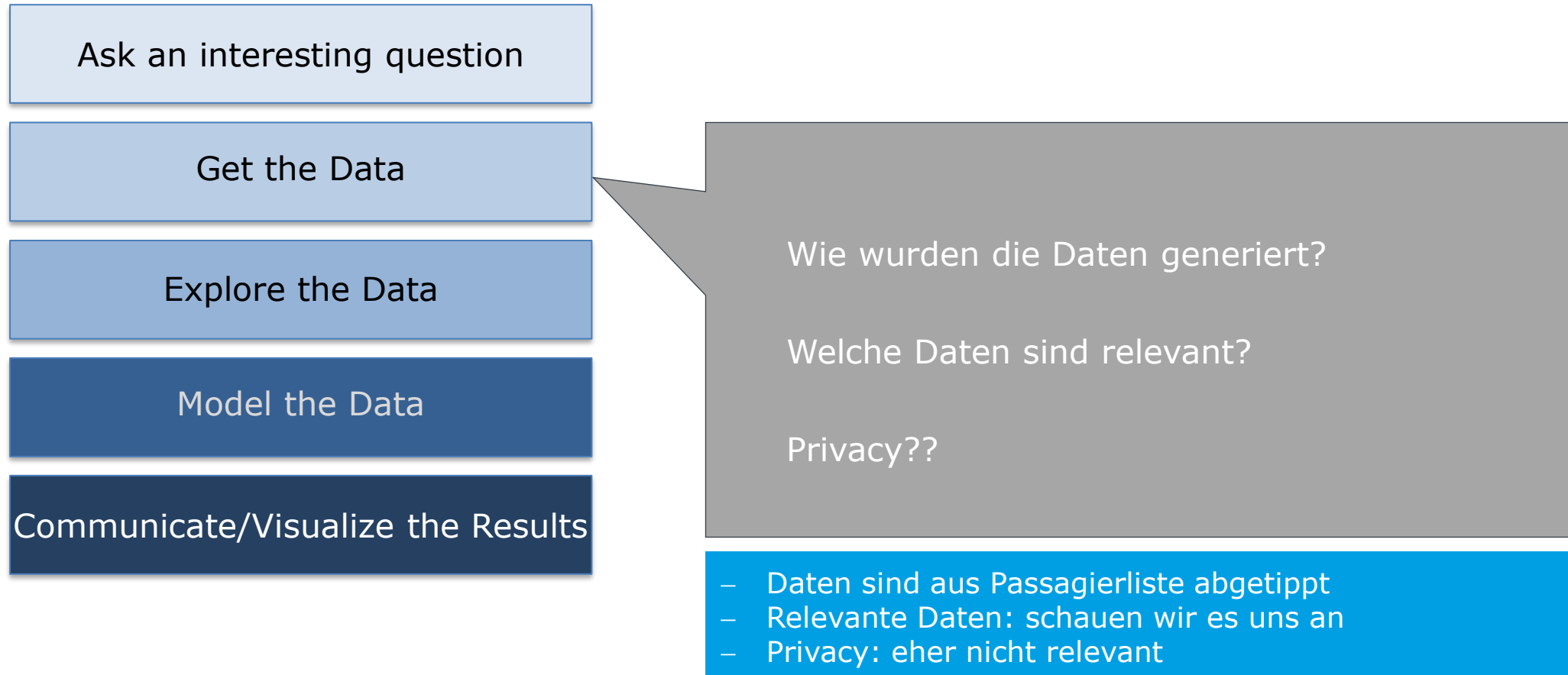
Was ist die Fragestellung?

Was würde ich tun, wenn ich alle Daten hätte?

Was möchte ich abschätzen/ vorhersagen?

- Wie hoch war die Überlebens-Chance eines Passagiers der Titanic?
- Hätte es eine gemeinsame Zukunft für Kate und Leonardo gegeben: oder war es sicherer, in der 1., 2. oder 3. Klasse zu reisen?
- Was ist der sicherste Indikator für das Überleben eines Passagiers?

DETAILLIERUNG VORGEHENSWEISE DATA SCIENCE.



ÜBERSICHT WICHTIGER DATA ENGINEERING TECHNIKEN

- **Verbessern Lesbarkeit/ Erklärbarkeit:** sprechende Namen für Eigenschaften (**Features**) und Werte.
- **Imputation:** fehlende Werte löschen oder ersetzen (bspw. Mittelwert, definierter Wert,)
- **Typumwandlungen:** beim Einlesen der Daten werden numerische Features oft als Text erkannt.
- **Diskretisation:** Einteilen von Werten mit großem Wertebereich in Gruppen, bspw. Alter (Kind, Teenager, Erwachsene, ...)
- **Categorization:** Werte mit beschränktem Wertebereich zusammenfassen (bspw. Farben, Wochentage, Geschlecht).
- **Outliers:** Werte, die sehr unterschiedlich zu restlichen Werten sind löschen oder per Standardwert ersetzen (bspw. Größe)
- **Normalisation/Scaling:** Werte innerhalb gewissen Wertebereichs bringen für Vermeiden Verzerrungen (bspw. Größe)
- **Feature Splitting:** Aufteilen Features für Infogewinn (bspw. Name in Vor-/Nachname, Adresse in Stadt und Straße)
- **One-hot-encoding:** Generieren neuer Features aus einem Feature (bspw. einzelne Tage statt Wochentage).
- **Neue Features:** Bauen neuer Features aus bestehenden oder Berechnungen (bspw. BMI aus Gewicht und Größe).
- **Feature removal:** Unwichtige Features löschen.



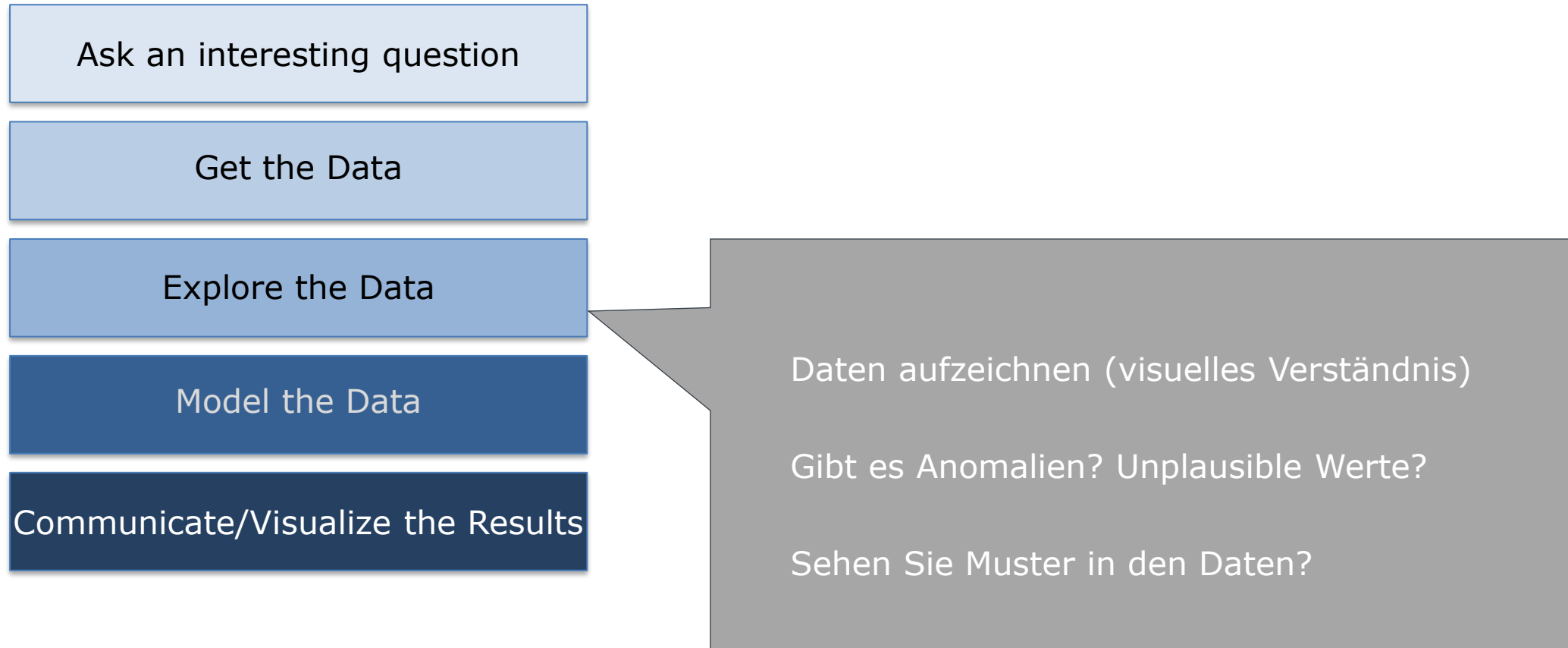
Data Engineering ist ein iterativer Prozess entlang des gesamten Use Cases und umfaßt oft 60 - 80% der Arbeit!

VERANSCHAULICHUNG DATA ENGINEERING ANHAND PASSAGIERLISTE TITANIC

index	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
0	1	1	Allen, Miss. Elisabeth Walton	female	29.0	0	0	24160	211.3375	B5	S	2	NaN	St Louis, MO
1	1	1	Allison, Master. Hudson Trevor	male	0.92	1	2	113781	151.55	C22 C26	S	11	NaN	Montreal, PQ / Chesterville, ON
2	1	0	Allison, Miss. Helen Loraine	female	2.0	1	2	113781	151.55	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
3	1	0	Allison, Mr. Hudson Joshua Creighton	male	30.0	1	2	113781	151.55	C22 C26	S	NaN	135.0	Montreal, PQ / Chesterville, ON
4	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0	1	2	113781	151.55	C22 C26	S	NaN	NaN	Montreal, PQ / Chesterville, ON
5	1	1	Anderson, Mr. Harry	male	48.0	0	0	19952	26.55	E12	S	3	NaN	New York, NY
6	1	1	Andrews, Miss. Kornelia Theodosia	female	63.0	1	0	13502	77.9583	D7	S	10	NaN	Hudson, NY
7	1	0	Andrews, Mr. Thomas Jr	male	39.0	0	0	112050	0.0	A36	S	NaN	NaN	Belfast, NI
8	1	1	Appleton, Mrs. Edward Dale (Charlotte Lamson)	female	53.0	2	0	11769	51.4792	C101	S	D	NaN	Bayside, Queens, NY
9	1	0	Artagaveytia, Mr. Ramon	male	71.0	0	0	PC 17609	49.5042	NaN	C	NaN	22.0	Montevideo, Uruguay

- **Verbessern Lesbarkeit/ Erklärbarkeit:** sprechende Namen für sibsp, parch, pclass, fare.
- **Categorization:** Einsetzen von beschränkten Werten für Embarked wie Southampton, Cherbourg, ...
- **One-hot-encoding:** Generieren neuer Features aus einem Feature (bspw. einzelne Tage statt Wochentage).
- **Diskretisation:** Einteilen von Alter in Altersgruppen wie Kind, Teenager, Erwachsene,
- **Neue Features/ Feature splitting:** Aufbau neues Feature HomeCountry aus Homedest.
- Mehrdeutigkeiten auflösen: cabin oder Name.
- **Imputation:** fehlende Werte löschen oder ersetzen für boat oder body. Aufwendig, sehr oft erfahrungsgetrieben.
- **Normalisation/Scaling:** bspw. für Alter und Ticketpreis.
- **Feature removal:** Löschen bspw. von cabin

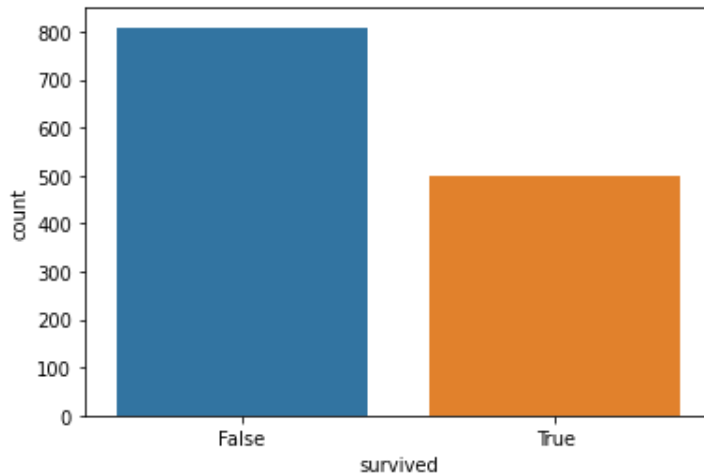
DETAILLIERUNG VORGEHENSWEISE DATA SCIENCE.



DATENEXPLORATION: UNTERSUCHEN EINZELNER MERKMALE (UNIVARIATE ANALYSEN).

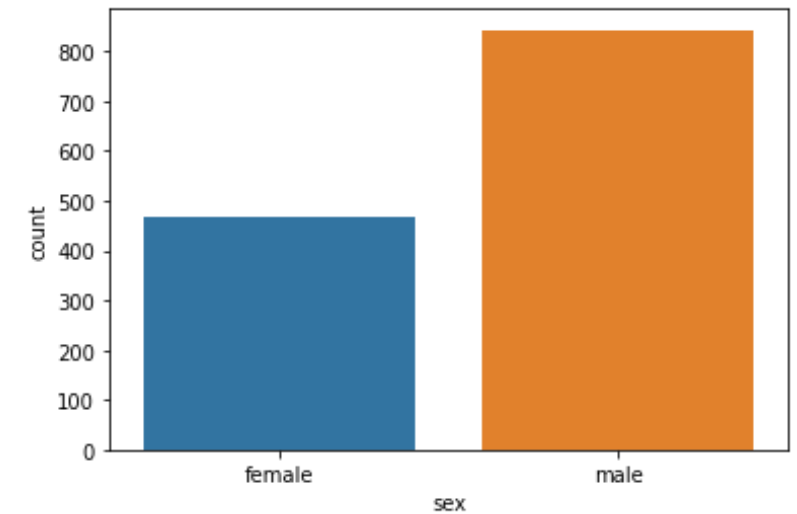
Wie viele Passagiere haben insgesamt überlebt?

500 von 1309



Wie viele Frauen/ Männer waren an Bord?

Frauen: 466
Männer: 843



► Empfehlung: Einsatz univariater Analyse am Anfang jeder Datenanalyse, um Muster oder Anomalien in Daten zu erkennen.

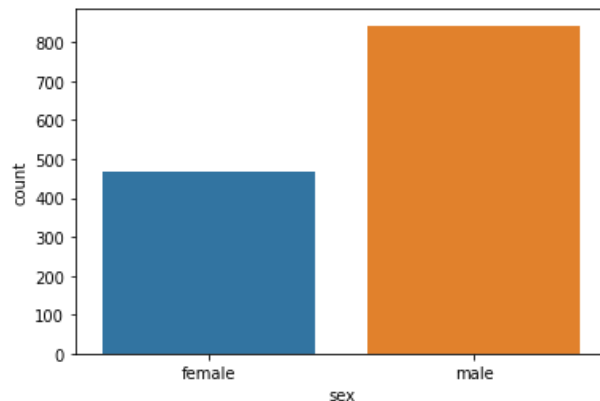
DATENEXPLORATION: UNTERSUCHEN MEHRERER MERKMALE (MULTIVARIATE ANALYSEN).

Zwei Attribute: wie viele Frauen/Männer überlebten?

Geschlecht an Bord: Frauen 466, Männer 843

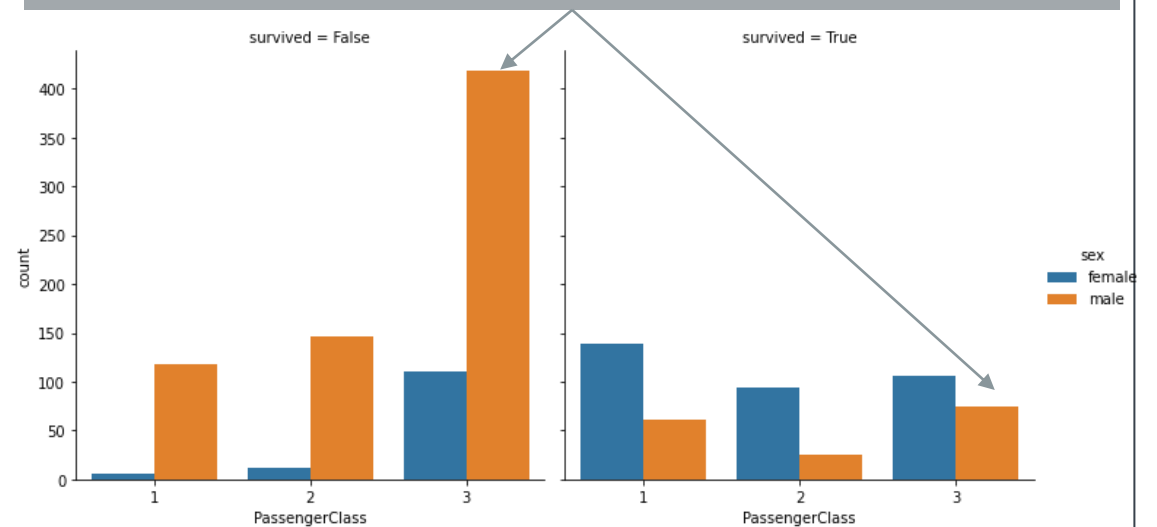
Überlebt:

- Frauen überlebt: $339/466 = 72\%$
- Männer überlebt: $161/843 = 19\%$



Drei Attribute: Wie viele Männer/Frauen überlebten in den verschiedenen Passagierklassen?

Wahrscheinlichkeit, daß ein Mann in Passagierklasse 3 überlebt:
 $\Pr(\text{Mann} \mid \text{PC}=3, \text{überlebt})$

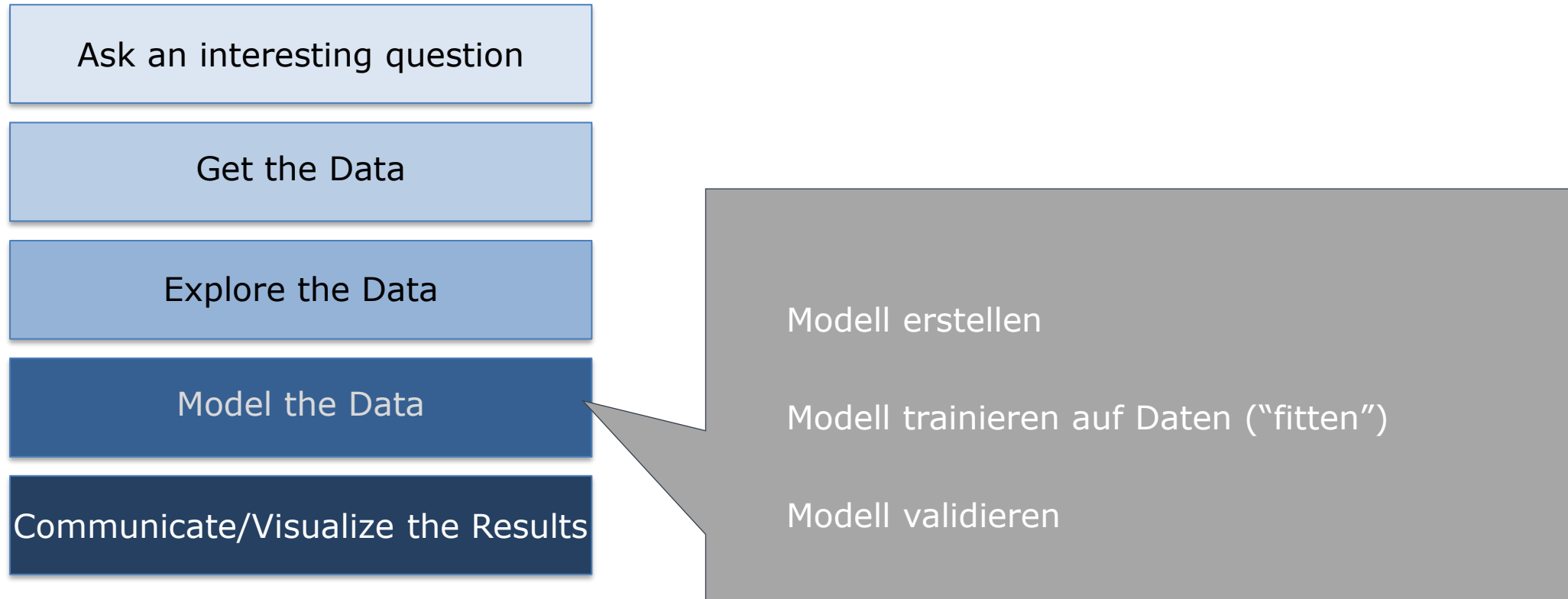


Bedingte Wahrscheinlichkeit daß, gegeben ein Mann, er in Passagierklasse 3 war

Verteilung Geschlechter aus Passagierklassen:

- Anzahl Männer & Passagierklasse: PC1 = 179, PC2 = 171, PC3 = 493)
- Anzahl Frauen & Passagierklasse: PC1 = 144, PC2 = 106, PC3 = 216)
- $\Pr(\text{Mann} \mid \text{PC}=3) = \Pr(\text{Mann und PC3}) / \Pr(\text{Mann}) = 493/843 = 58\%$
- $\Pr(\text{Frau} \mid \text{PC}=1) = \Pr(\text{Frau und PC1}) / \Pr(\text{Frau}) = 144/466 = 30\%$

DETAILLIERUNG VORGEHENSWEISE DATA SCIENCE.



MODEL THE DATA: AUSBLICK AUF DIE SPÄTEREN VORLESUNGEN.

Trainieren Machine Learning Model (vereinfacht)

```
from sklearn.ensemble import RandomForestClassifier

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)

random_forest = RandomForestClassifier(n_estimators=200)
random_forest.fit(X_train, y_train)

Y_prediction = random_forest.predict(X_test)
```

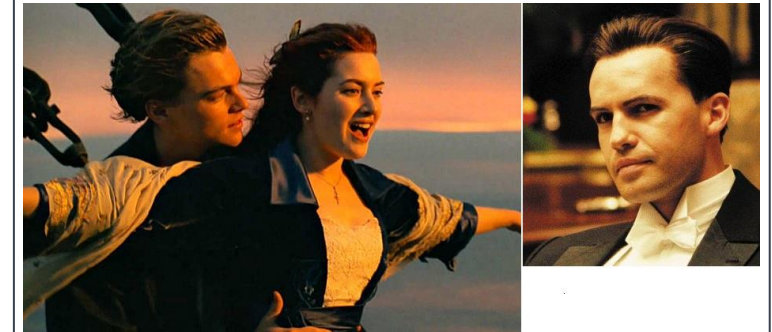
Messen Modellgüte per Metriken

Einsatz Confusion Matrix als Metrik für Klassifikation:

		Predicted	
		Ja	Nein
Tatsächlich	Ja	True Positive	False Negative
	Nein	False Positiv	True Negative

$$\text{Accuracy} = \left(\frac{\text{korrekt vorhergesagt}}{\text{Gesamtzahl}} \right) = 97.49\%$$

Anwenden Modell



```
[ ] # notwendigen Features sind: PassengerClass, Sex (0 Frau, 1 Mann), Age,
    # SiblingSpousesPresent, ParentsChildrenPresent, fare, AgeGroup
Kate = [[1, 0, 17., 0, 1, 150., 2]]
Leo = [[3, 1, 20., 0, 0, 15., 3]]
Billy = [[1, 1, 30., 0, 0, 150., 3]]

# make a prediction
print("Prädiktion Überlebenschance Kate:", logmodel.predict(Kate))
print("Prädiktion Überlebenschance Leo:", logmodel.predict(Leo))
print("Prädiktion Überlebenschance Billy:", logmodel.predict(Billy))

Prädiktion Überlebenschance Kate: [1]
Prädiktion Überlebenschance Leo: [0]
Prädiktion Überlebenschance Billy: [1]
```

Wir werden uns die einzelnen Schritte im Detail in den nächsten Vorlesungen ansehen

LITERATUR UND WEITERE QUELLEN (AUSZUG).

Statistik:

- James, Witten, Hastie and Tibshirani: An Introduction to Statistical Learning (freies Ebuch unter: [Link](#))
- Spiegelhalter: The Art of Statistics: Learning from Data
- Witte: Statistics (10th Edition)
- Silver: The Signal and the noise
- Taleb: Black Swan
- Huff: How to Lie with Statistics
- Wheelan: Naked statistics

Kostenfreie Online-Kurse (bei Interesse):

- Data Science mit Excel ([Link](#))
- Python-Kurse
 - Python for Everybody ([Link](#))
 - Udacity Python Course ([Link](#))
 - Kaggle Courses:
 - Python ([Link](#))
 - Python Library Pandas ([Link](#))
 - PythonData Visualization ([Link](#))

FALLS SIE GERNE WEITERE ERFAHRUNGEN SAMMELN WOLLEN....

- Amazon TOP 50 Books: <https://www.kaggle.com/sootersaalu/amazon-top-50-bestselling-books-2009-2019>
- Credit Card Approval: <https://www.kaggle.com/rikdifos/credit-card-approval-prediction>
- Starbucks Menu: <https://www.kaggle.com/starbucks/starbucks-menu?select=starbucks-menu-nutrition-food.csv>
- Wetter: <https://www.kaggle.com/sudalairajkumar/daily-temperature-of-major-cities>

WEITERE BEISPIELE FÜR MULTIVARIATE DATENANALYSEN ANHAND TITANIC.

Fahrpreis:

- Was war der höchste Fahrpreis, den ein weiblicher Passagier zahlte?
- Schwierig: Was war der durchschnittliche Fahrpreis für Frauen je Passagierklasse?

Zusteigeort:

- Was der häufigste Zusteigeort (Embark) für Passagierklasse 1?
- Gibt es einen Zusammenhang zwischen Zusteigeort (Embark) und der Überlebenschance?

Alter:

- Schwierig: Was ist das Durchschnittsalter in der 2. Klasse? Ist es höher als das für die 3. oder 1. Klasse?
- Was ist das Durchschnittsalter in 2. Klasse für Männer?