

TECHNICAL APPLICATIONS AND DATA MANAGEMENT. WS 2021/2022.

VORLESUNG 2

25.09.2020

MÜNCHEN

STUDIENGANG
DIGITAL
MANAGEMENT.



AGENDA

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

RÜCKBLICK AUF LETZTE WOCH...

	Kunden-ID	Name	Geboren	Alter	Adresse	Kreditkartennummer	Einkäufe 2020	Umsätze 2020
Regel für Sicherstellen Datenqualität	ID definiert und eindeutig (d.h. darf max. 1 mal vorkommen)	Liegt vor	Datum in europäischem Format: TT.MM.YY., sonst umwandeln	Alter < 120	muß vorliegen	1. $12 \leq \text{Anzahl Ziffern} \leq 16$ 2. Korrekte Prüfsumme (bspw. Luhn-Algorithmus ¹)		Währung in EUR, sonst umwandeln
Relevant für Wertschöpfung per Empfehlung/ Service	-	-	Altersgruppen	Ja, für Empf.. Aber bspw. auch für Ansprache Kunde	Ja, bspw. Wohnort		Ja, für Empfehlungen	Ja, für Empfehlungen

- Wie generiert die gewählte Firma mit Daten Einnahmen?
- Welche Daten benötigt die gewählte Firma hierfür?
- Wie müssen die Daten dann sein? Welche Kriterien für Datenqualität sind dann wichtig?
- Skalieren: Nehmen Sie an, Sie haben 100 000 oder mehr Kunden/ User.
 - Können Sie Regeln für das Erfassen, Prüfen, Auswerten der Daten definieren?
 - Wie können Sie –bspw. auf Basis der definierten Regeln – die Vorgänge automatisieren?

WIE GEHT'S WEITER?

Sie haben in Ihren Fallbeispielen Sixt, H&M, Amazon genommen und wollten per personalisierten Empfehlungen an Ihre Kunden Einnahmen generieren.

Aber:

- wie stellen Sie fest, was für Ihre Kunden denn passende personalisierte Empfehlungen sind?
- Und was zeichnet „den“ Kunden aus? Wer ist denn „der“ Kunde?
- Und wie kann man statistisch verlässliche Aussagen treffen? (Nächste Woche)

1. WAS IST DATA SCIENCE?

what my friends think I do



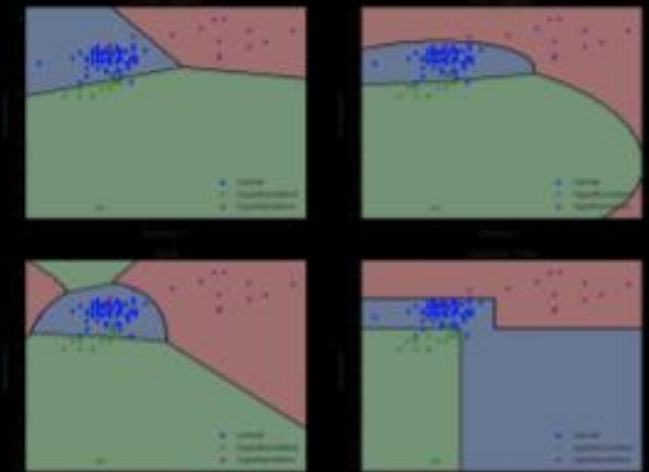
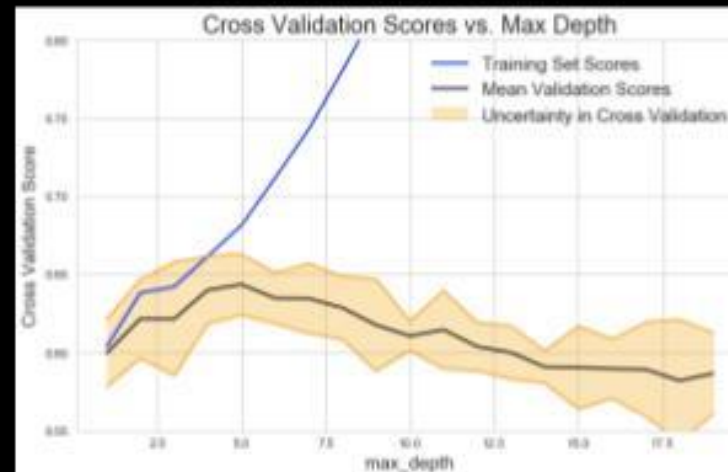
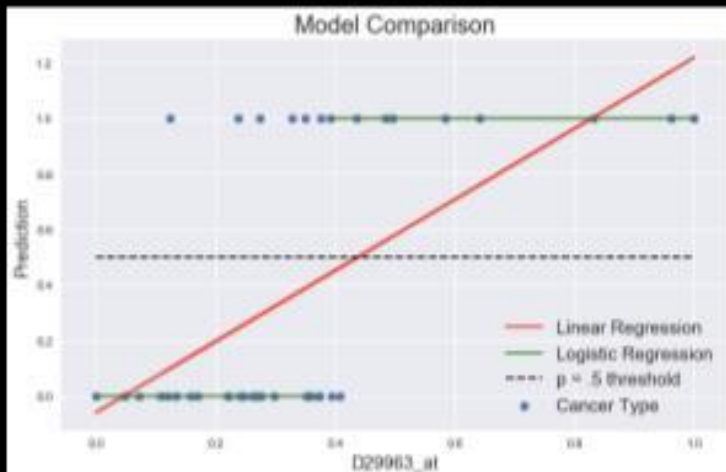
what my family thinks I do



what society thinks I do

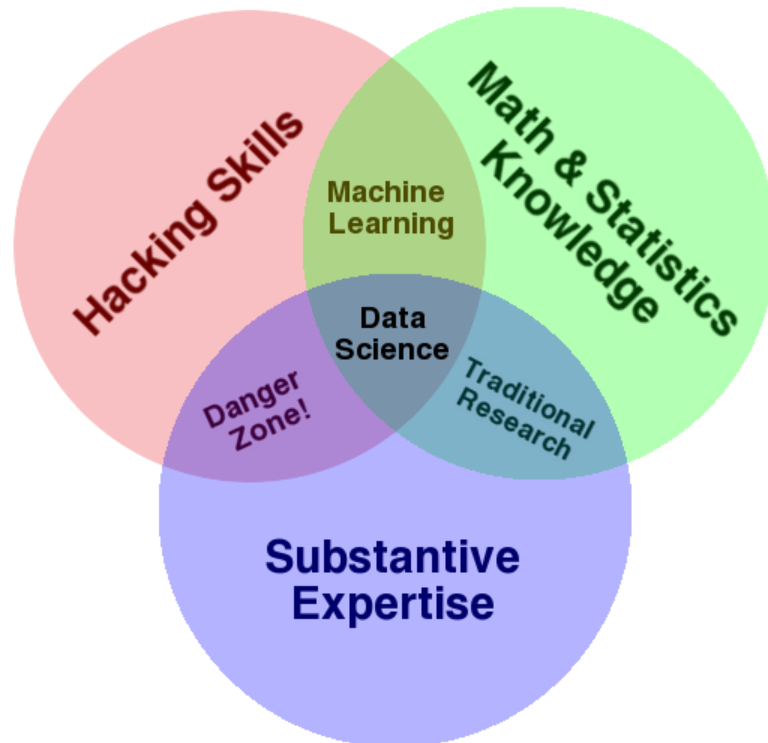


what I actually (will) do in Data Science 1



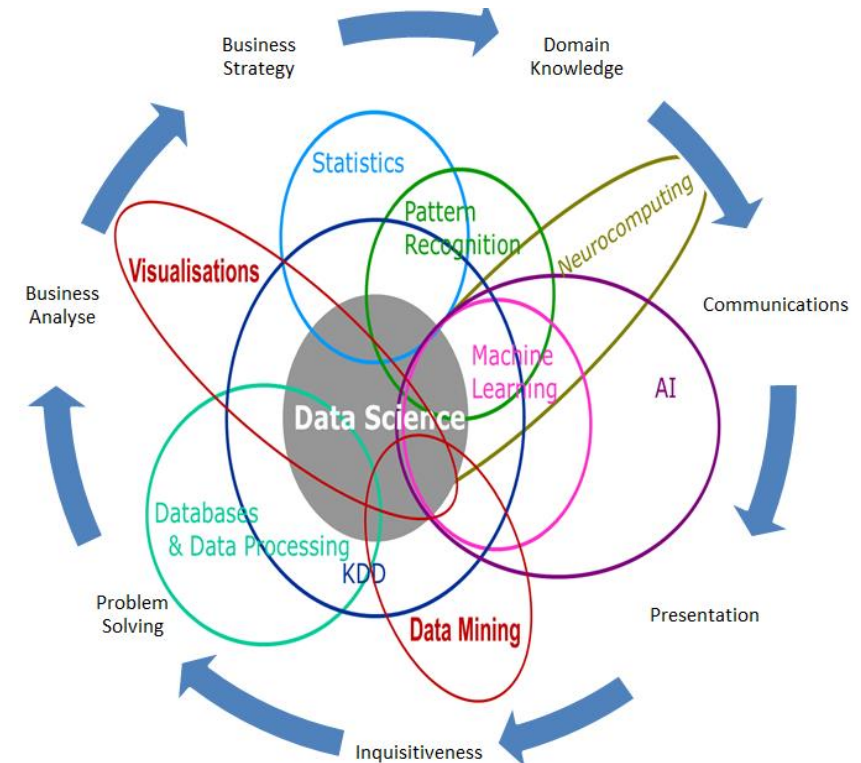
VERSTÄNDNIS FÜR BEGRIFF SOWIE UMFANG DATA SCIENCE HAT SICH STARK GEÄNDERT IN DEN LETZTEN JAHREN.

2010



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

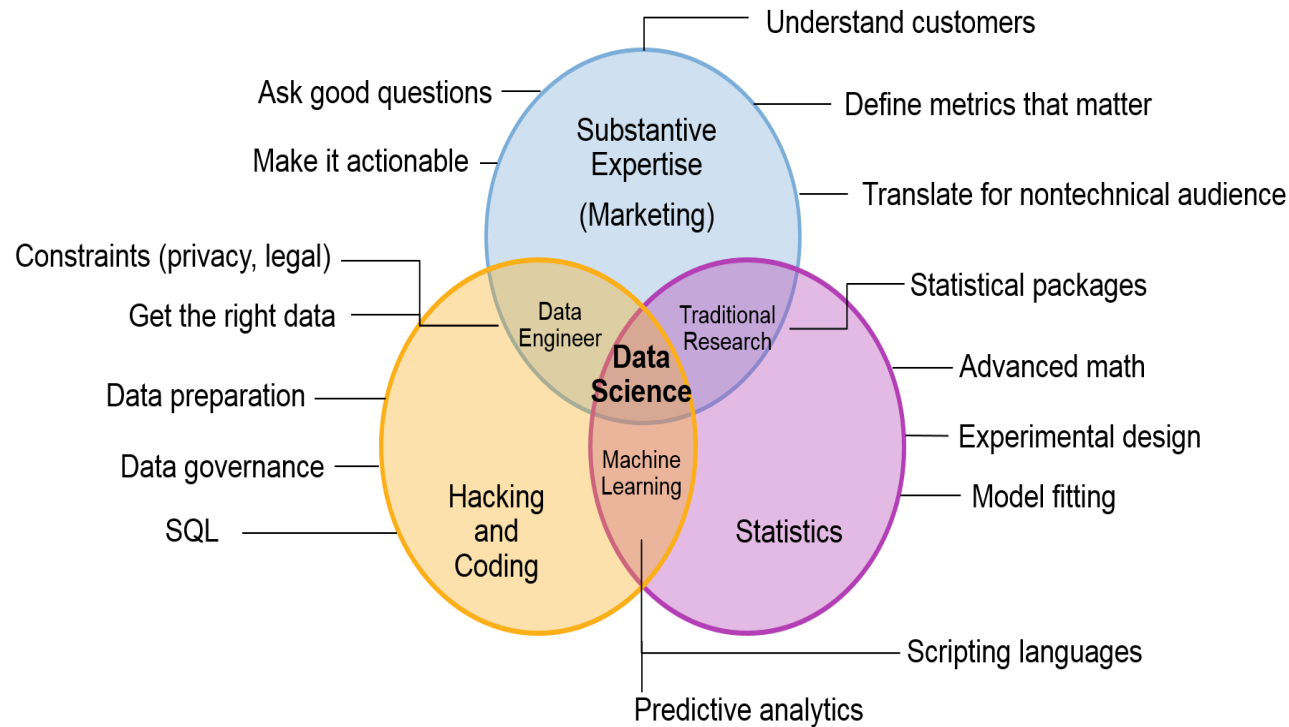
2013



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

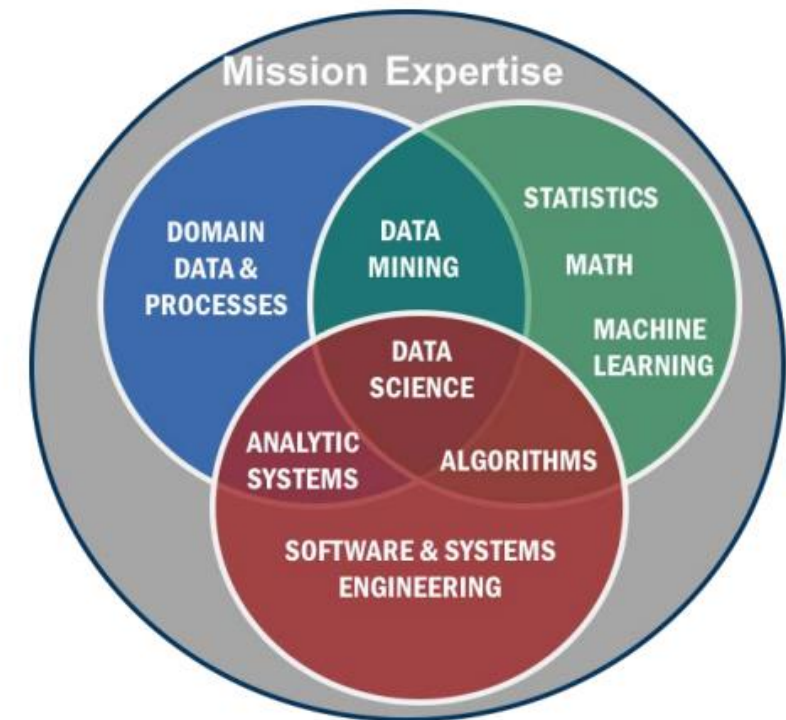
VERSTÄNDNIS FÜR BEGRIFF SOWIE UMFANG DATA SCIENCE HAT SICH STARK GEÄNDERT IN LETZTEN JAHREN.

2016



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

2019



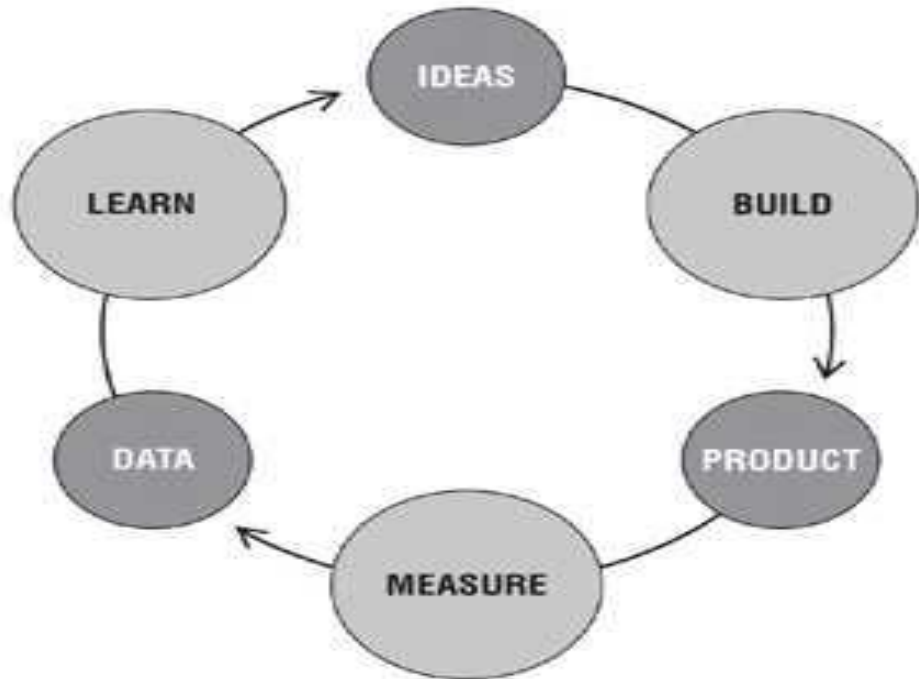
Quelle: NIST big data workgroup, 2019, verfügbar unter: [Link](#)



2. VORGEHENSWEISE BEI EINEM DATA SCIENCE USE CASE

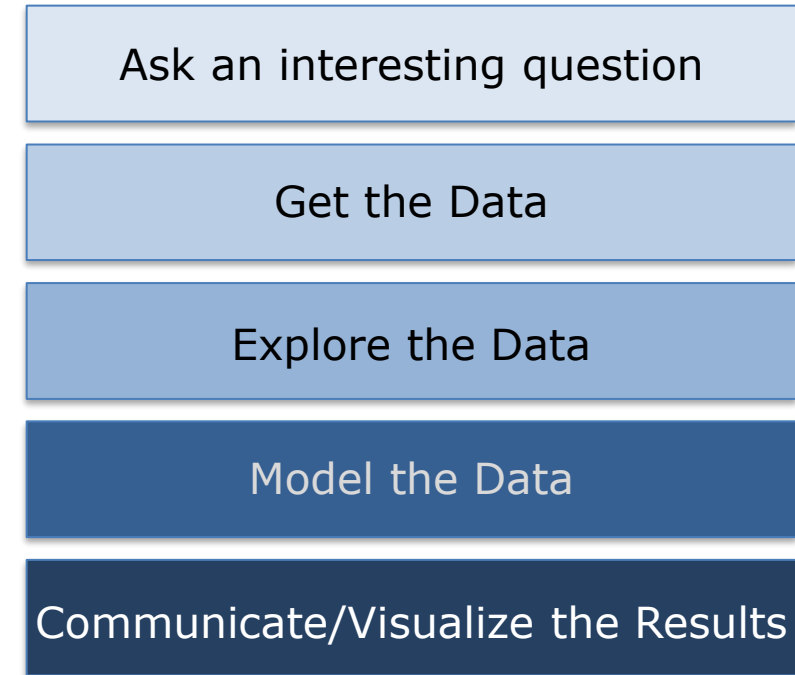
VORGEHENSWEISE USE CASE DATA SCIENCE.

BUILD-MEASURE-LEARN FEEDBACK LOOP



Vorgehensweise einer „data-driven company“

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

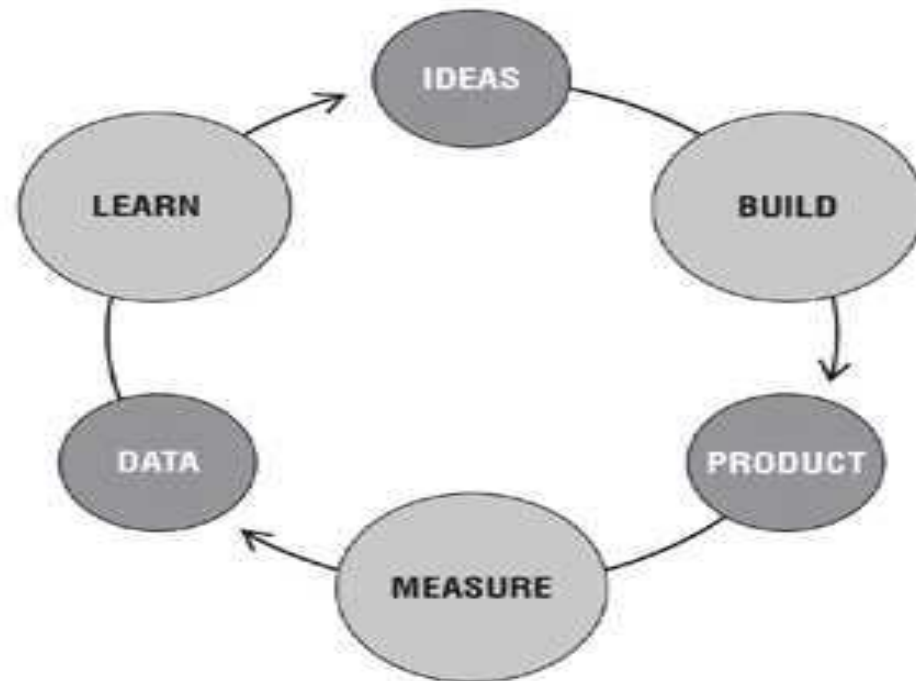


Generische Vorgehensweise Data Science

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



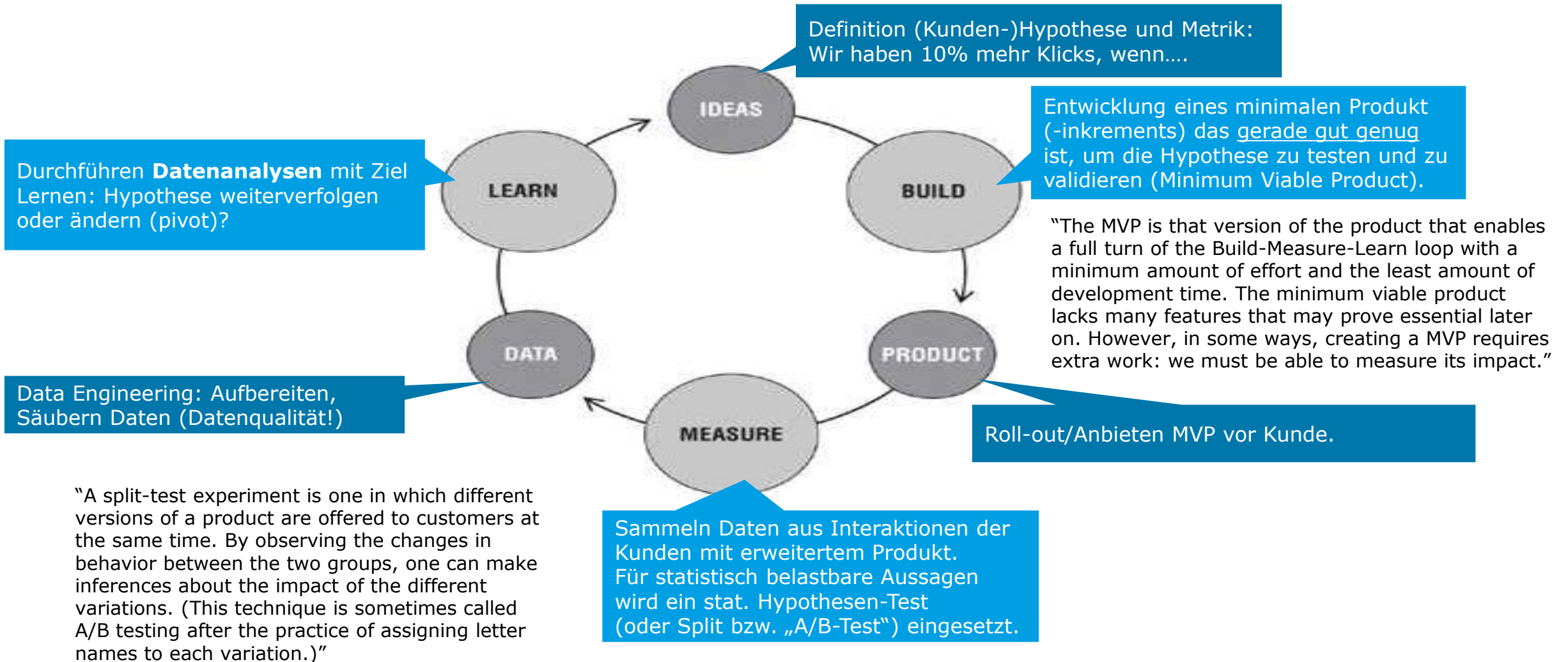
Minimize *TOTAL* time through the loop

“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”.

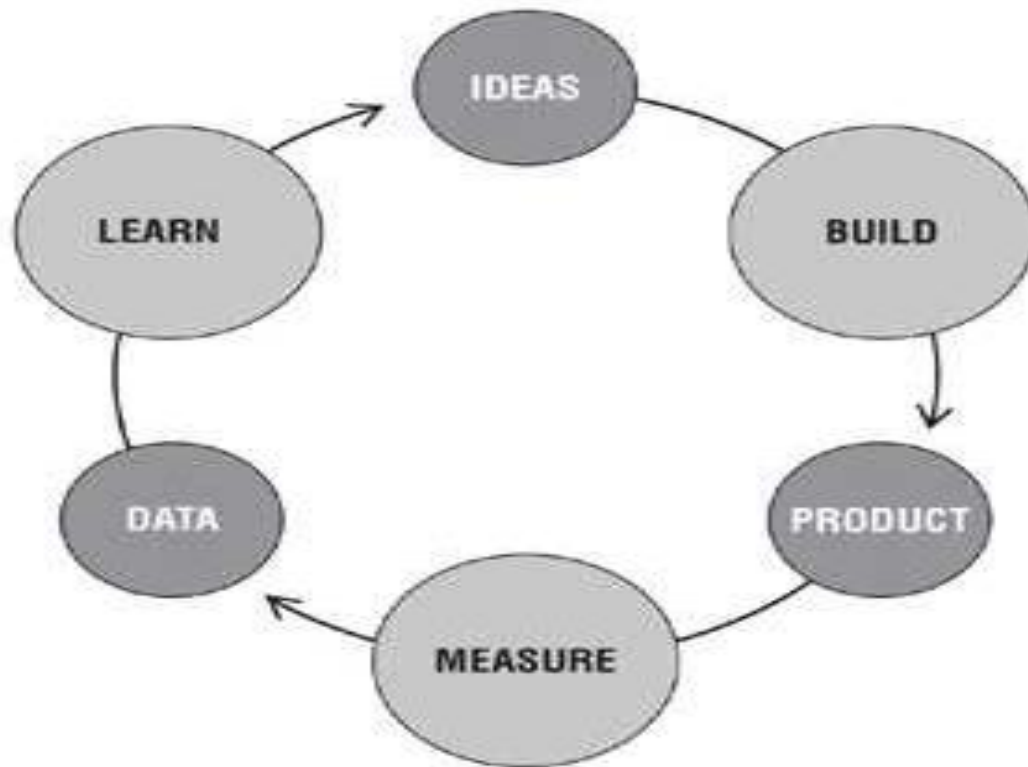
“Startups exist not just to make stuff, make money, or even serve customers. They exist to **learn how to build a sustainable business**. This **learning** can be **validated** scientifically by running **frequent experiments** that allow entrepreneurs to test each element of their vision.”

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 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 in Gruppenarbeit 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

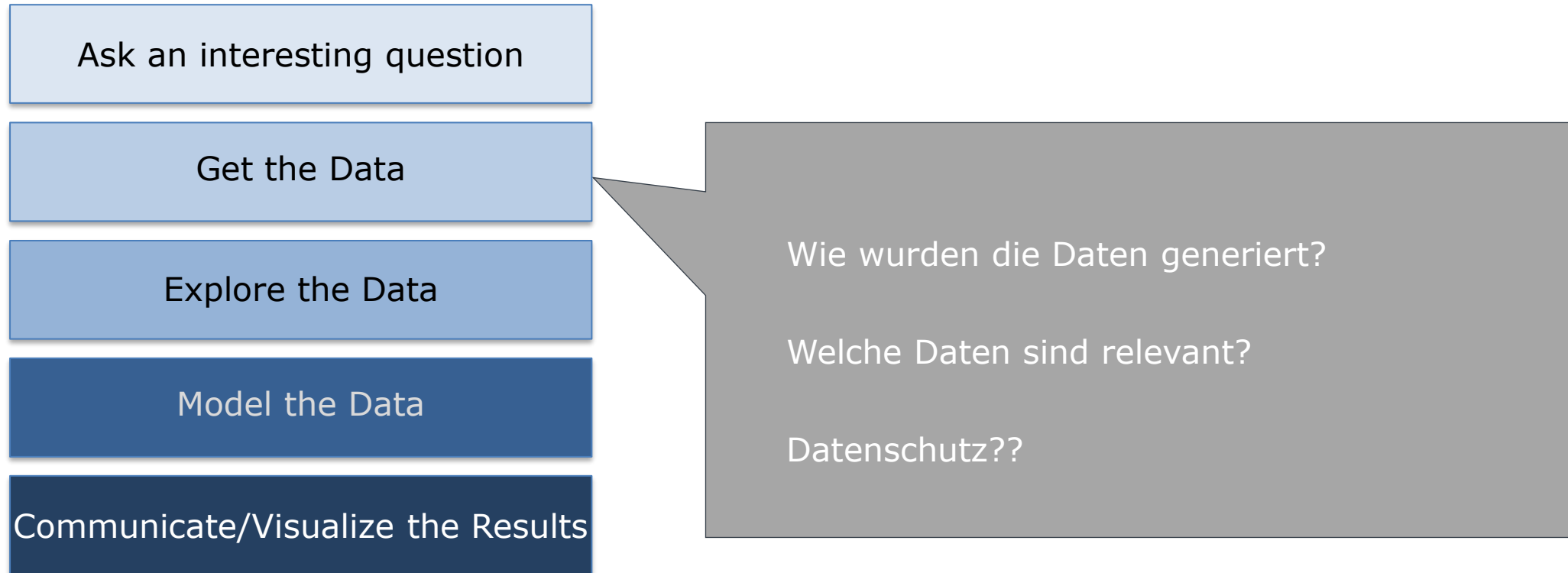
Communicate/Visualize the Results

Was ist die Fragestellung?

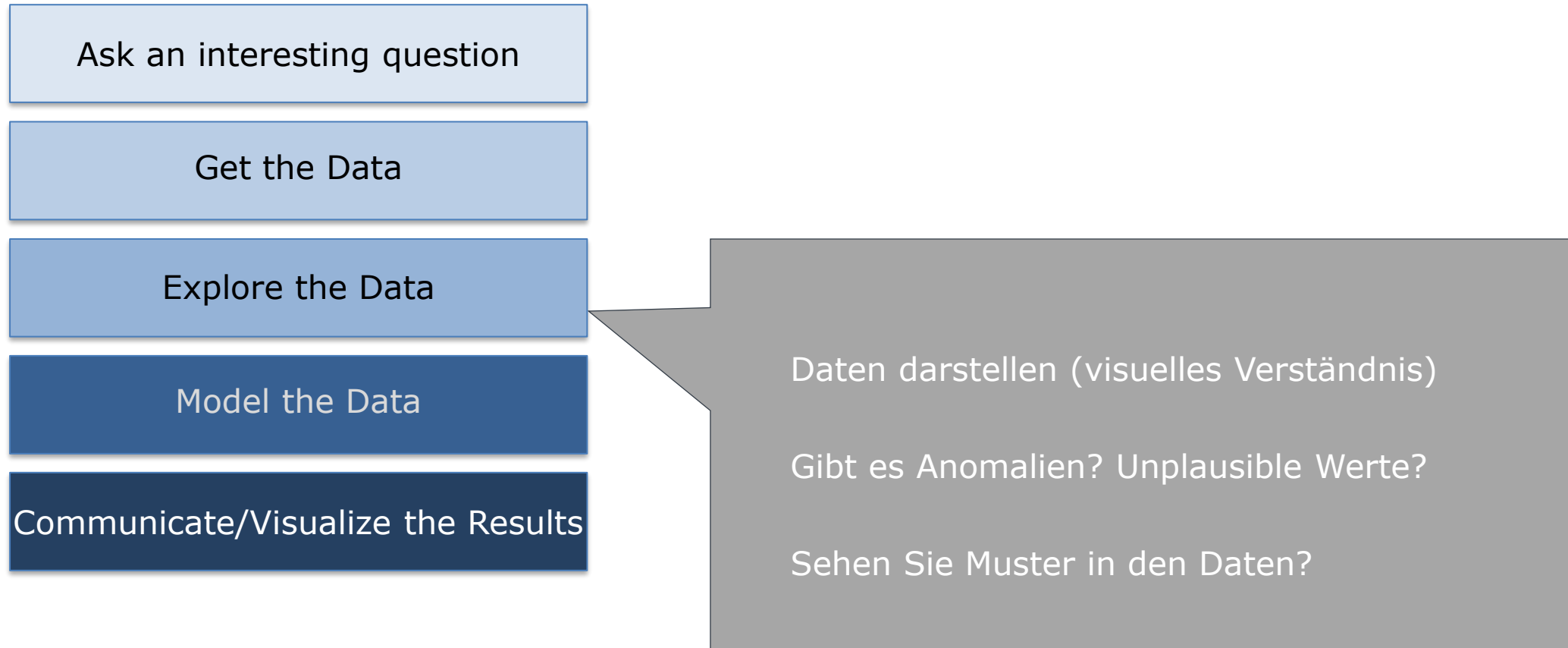
Was würde ich tun, wenn ich alle verfügbare Daten hätte?

Was möchte ich abschätzen/ vorhersagen?

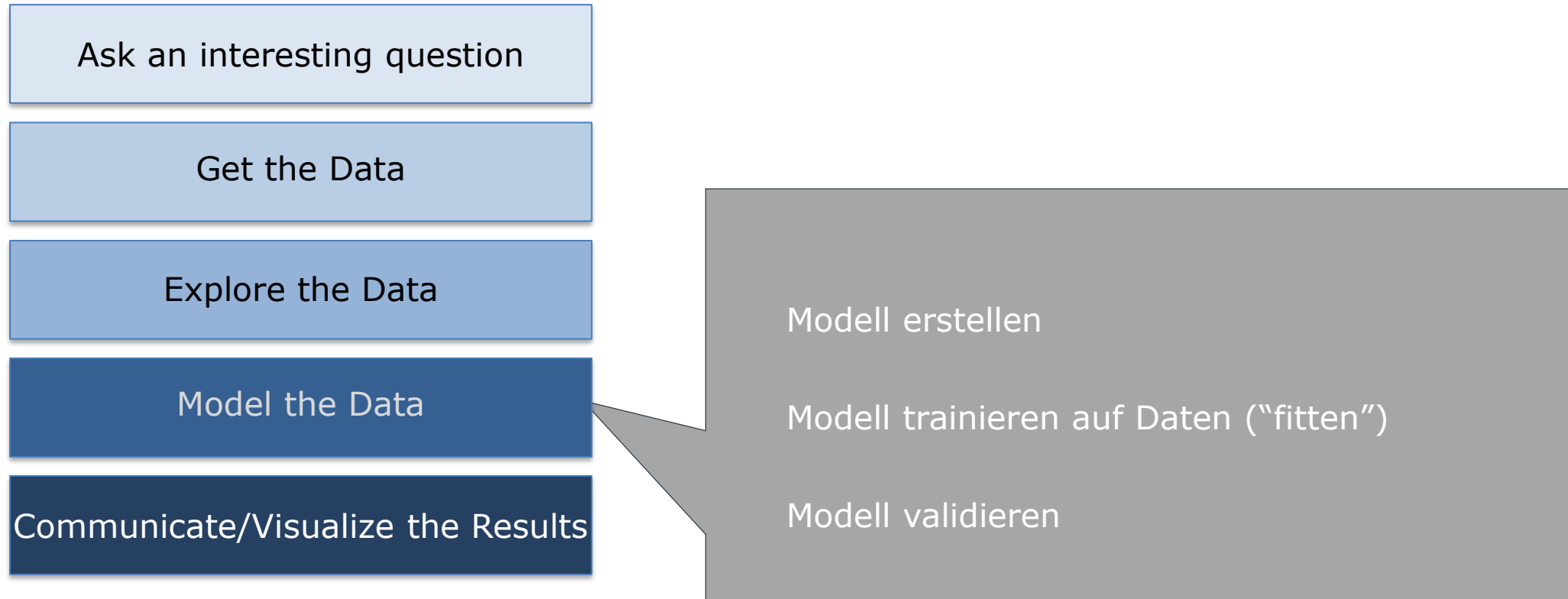
DETAILLIERUNG VORGEHENSWEISE DATA SCIENCE.



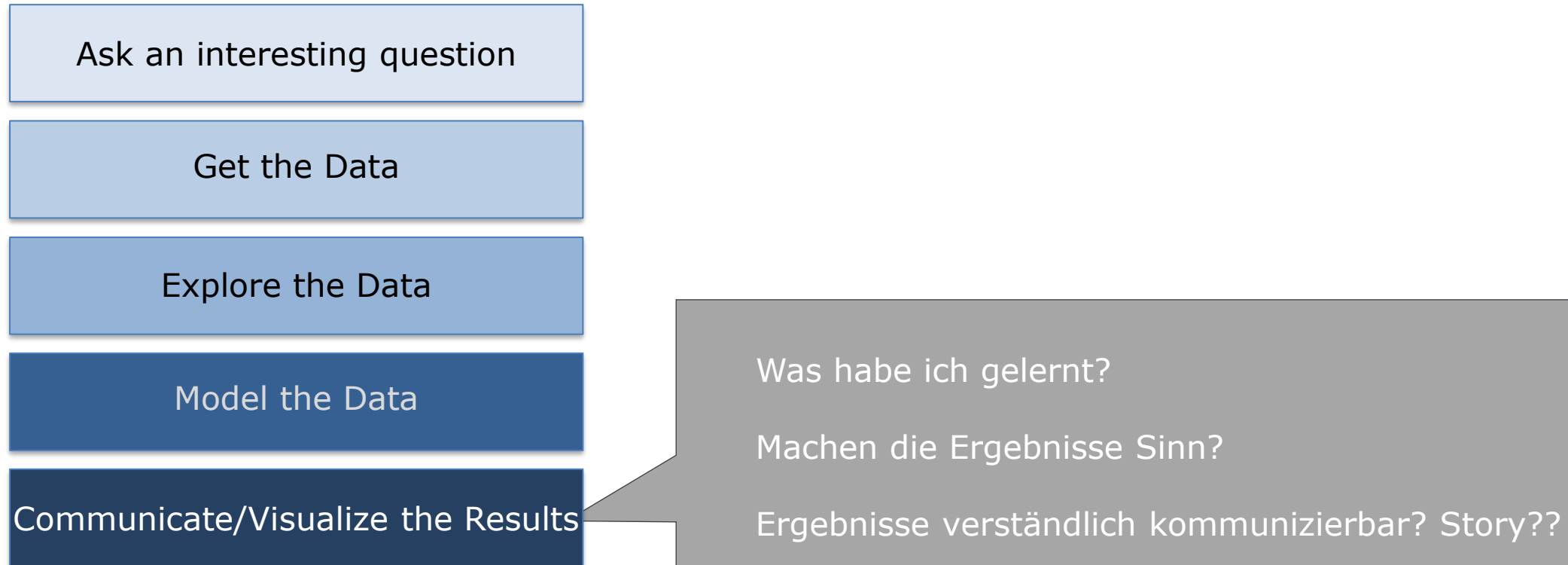
DETAILLIERUNG VORGEHENSWEISE DATA SCIENCE.



DETAILLIERUNG VORGEHENSWEISE DATA SCIENCE.



DETAILLIERUNG VORGEHENSWEISE DATA SCIENCE.

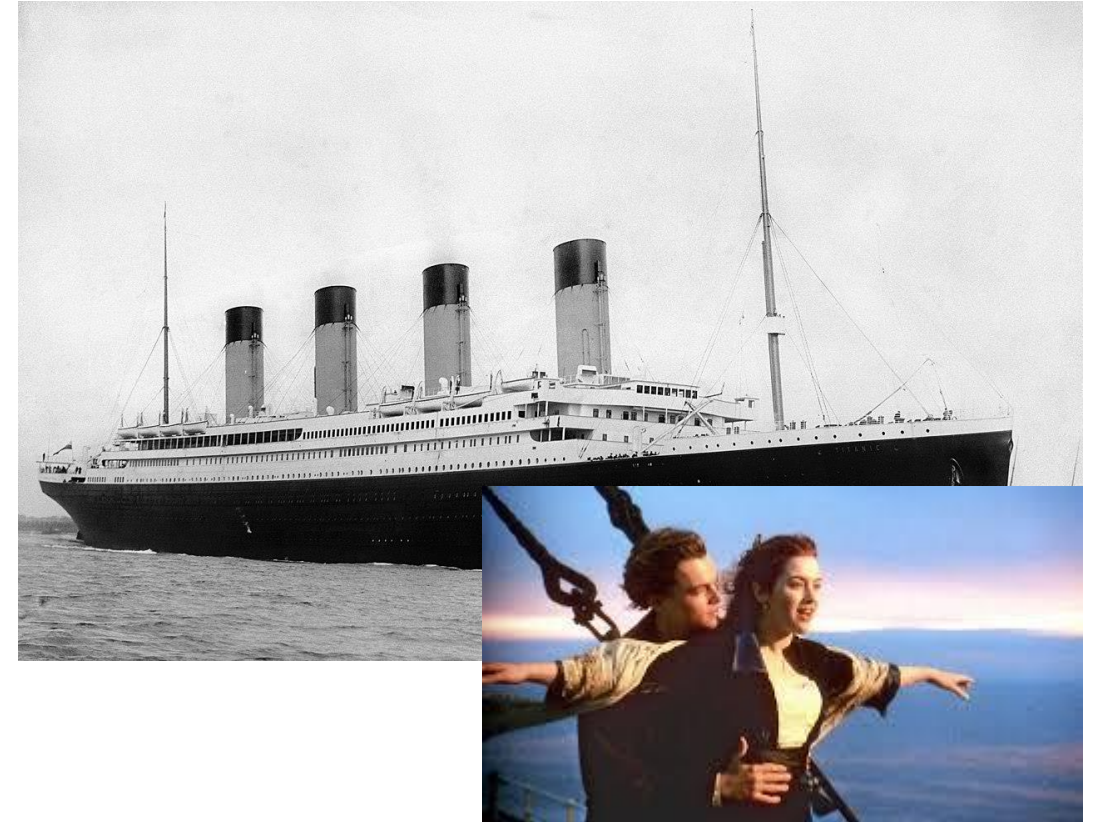




2. FALLBEISPIEL

HANDS ON 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
- Was werden wir machen:
 - Import/ Laden der Daten
 - Data Engineering: säubern, aufbereiten, neue Features
 - Univariate Datenanalysen (Analyse eines Features)
 - Multivariate Datenanalysen (Analyse mehrerer Features)
 - Annäherung Zielvariable per manueller Optimierung

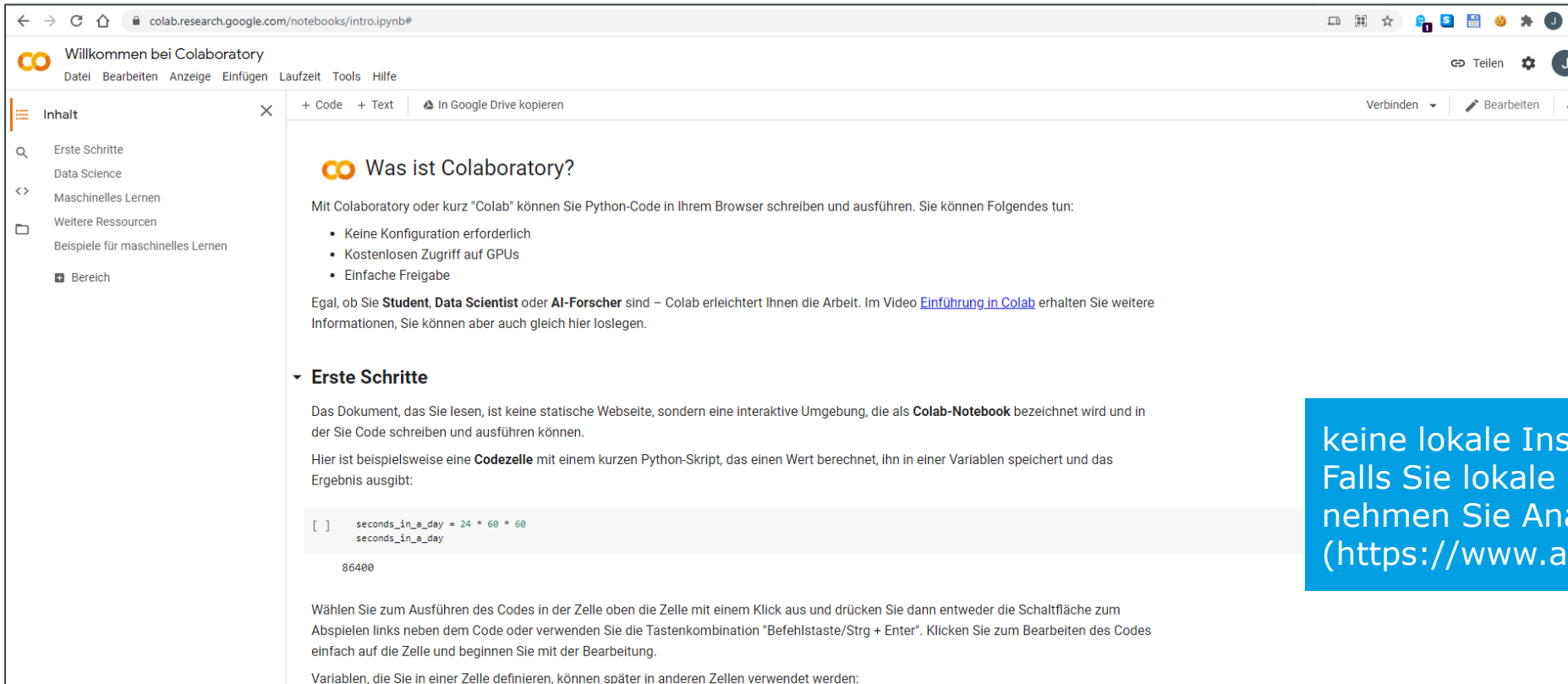


▶ Später werden wir Libraries einsetzen, um viele der o.a. Tätigkeiten zu automatisieren.
Für den Anfang sollten Sie aber diese Schritte im Detail mal gesehen haben.

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

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.

colab.research.google.com/drive/1kTODv6reK7C5N_WbRIVjeXN4aW5jIwgU#scrollTo=NZur1W7XRaST

Titanic-Notebook-Update - WS2021/22

Datei Bearbeiten Anzeige Einfügen Laufzeit Tools Hilfe Alle Änderungen wurden gespeichert

Inhalt

- 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

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.

0 s Abgeschlossen um 10:11

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

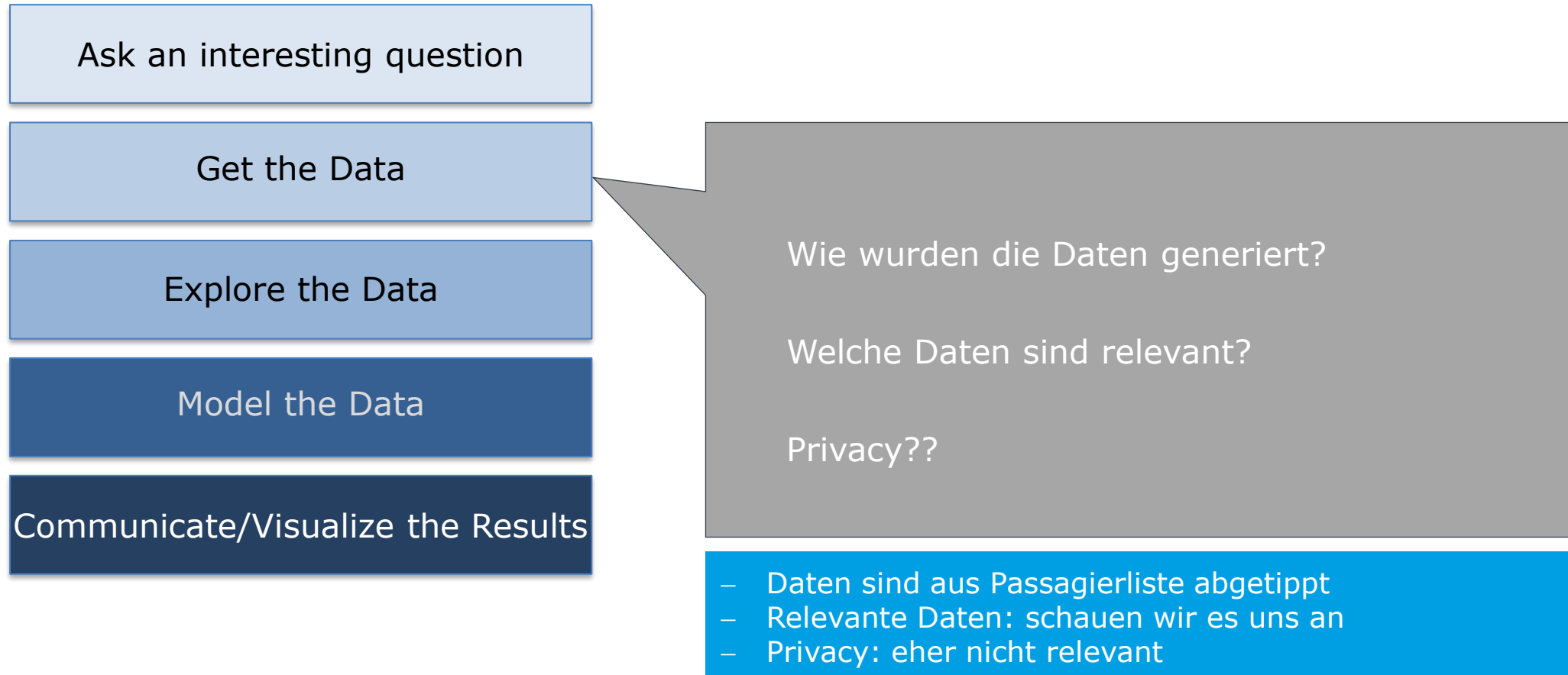
Was ist die Fragestellung?

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.



DATENMANAGEMENT.

- Aufbereiten Daten für bessere Lesbarkeit:
 - Spaltennamen ändern:
 - Sibsp: Number of Siblings/Spouses Aboard
 - Parch: Number of Parents/Children Aboard
 - Werte statt Abkürzungen (**categorical Variables**) für Embarked: C = Cherbourg; Q = Queenstown; S = Southampton
- Datenqualität verbessern:
 - Null-, Leere Werte: welchen Wert nehmen? (**Imputing**)
 - Zielvariable definieren und an richtiger Stelle: Survived
- Neue Spalten bauen (**Feature Engineering**)
 - Altersgruppen (KIND, TEENAGER, ERWACHSEN)
 - Boat_corrected

Sehr aufwendig, sehr oft erfahrungsgetrieben.
Umsetzung mit Programmierkenntnissen empfohlen,
da man so schnell die Auswirkungen auswerten kann.

Benötigen wir vor allem bei Machine Learning

DETAILLIERUNG VORGEHENSWEISE DATA SCIENCE.

Ask an interesting question

Get the Data

Explore the Data

Model the Data

Communicate/Visualize the Results

Daten aufzeichnen (visuelles Verständnis)

Gibt es Anomalien? Unplausible Werte?

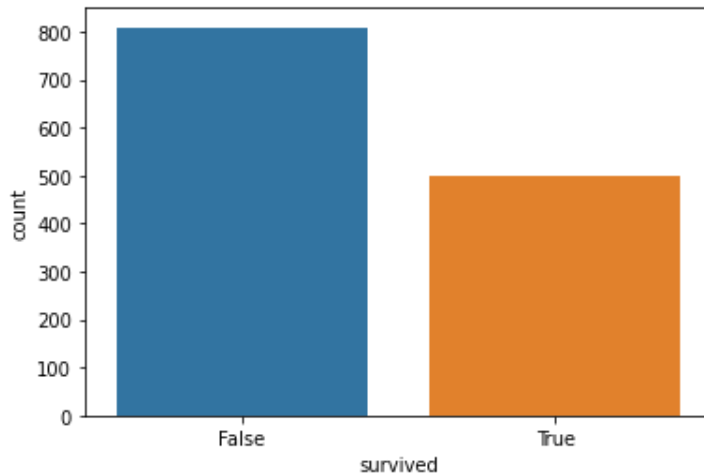
Sehen Sie Muster in den Daten?

Gehen wir's an!
Programm-Code für die jeweiligen Auswertungen
finden Sie im Titanic-Notebook ([Link](#))

DATENEXPLORATION: UNTERSUCHEN EINZELNER MERKMALE (UNIVARIATE ANALYSEN).

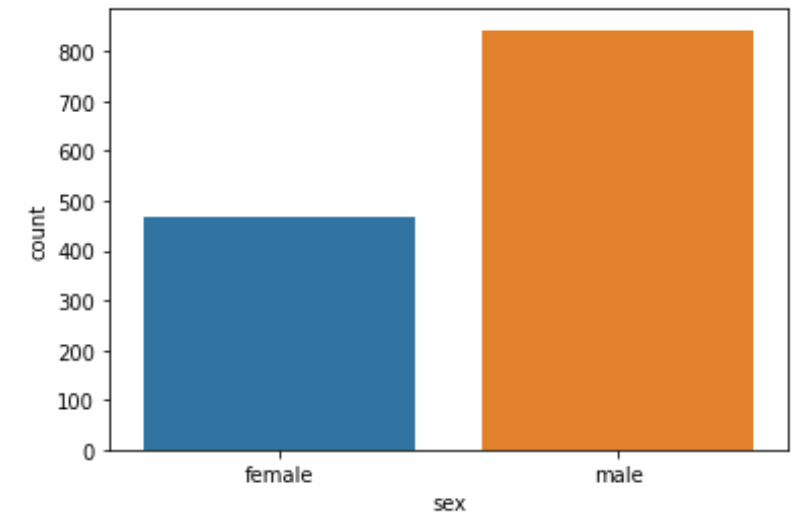
Wie viele Passagiere haben insgesamt überlebt?

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.

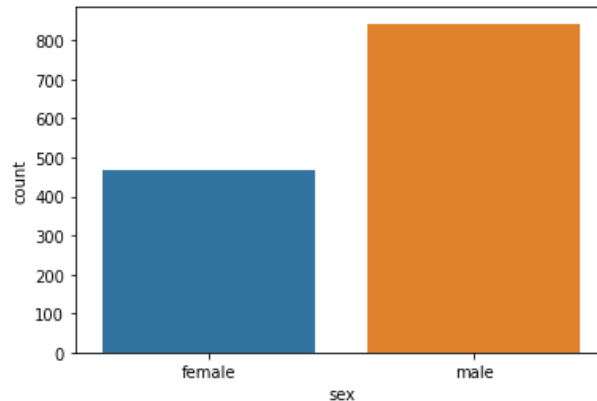
DATENEXPLORATION: UNTERSUCHEN MEHRERER MERKMALE (MULTIVARIATE ANALYSEN).

Zwei Attribute: wie viele Frauen/Männer überlebten?

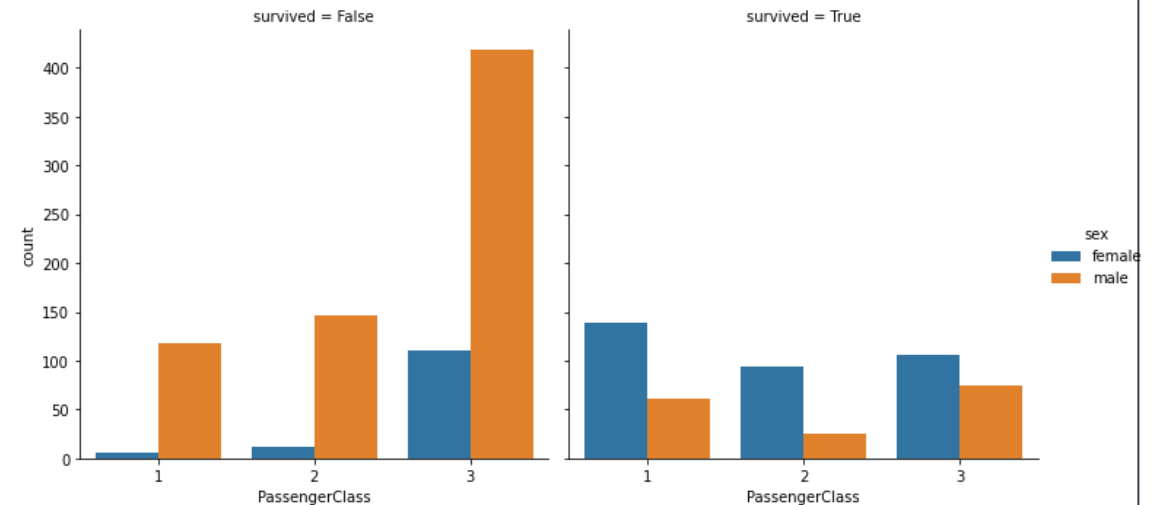
Geschlecht an Bord: Frauen 466, Männer 843

Überlebt:

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



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



DATENEXPLORATION: UNTERSUCHEN MEHRERER MERKMALE (MULTIVARIATE ANALYSEN).

Verteilung Geschlechter aus Passagierklassen:

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

Bedingte Wahrscheinlichkeit
daß gegeben ein Mann er in
Passagierklasse 3 war

Gibt es Unterschiede für die Anzahl Überlebende/ Überlebensrate abhängig von Geschlecht und Passagierklasse

$\Pr(\text{Überlebensrate} \mid \text{Männlich, Passagierklasse} = X) = \frac{\Pr(\text{Überlebensrate und Männlich und Passagierklasse X})}{\Pr(\text{Überlebensrate})}$

Kettenregel, kommt in späterer
Vorlesung

▶ Gerne zum Ausprobieren: Gegeben alle Attribute, welches ist das mit der geringsten Indikation fürs Überleben?

ERSTE DATENANALYSEN IN GRUPPENARBEIT.

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?

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>

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](#))