Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Datenmanagement & -analyse

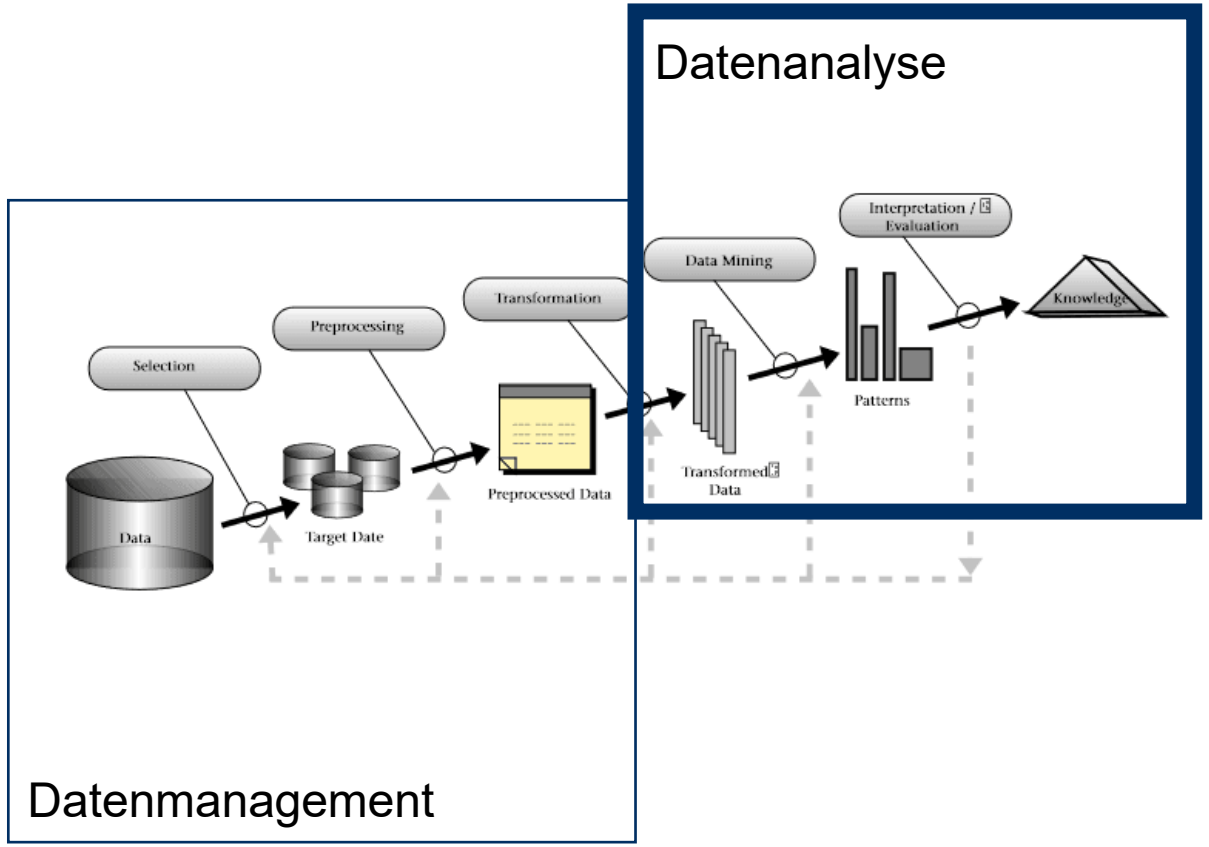
Strukturierte Datenanalyseprozesse

Prof. Dr. Christoph M. Flath
Lehrstuhl für WI & BA
Julius-Maximilians-Universität Würzburg

Sommersemester 2021

Rückblick VL1: Überblick der Themen

- Konzepte klassischer Datenbankarchitekturen
- Datenmodellierung und Normalisierung
- Einführung in eine Anfragesprache
- Nutzung von Datenbanken
- **Hypothesengetriebene und modellbildende Datenanalyse**
- Datenanalyseprozesse und deren Vergleich
- Überwachte und unüberwachte Lernverfahren
- Konzeption und Umsetzung (komplexer) Datenanalysen



Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Agenda

1. **Motivation**
2. **Ein berühmter Brite**
3. **Zahlendetektive**
4. **Ein Unglück erklärbar machen**

Datenfähigkeiten sind notwendig um die Welt (besser) zu verstehen

Höchster Stand seit 2011

Inflation steigt auf 2,5 Prozent

Stand: 31.05.2021 16:00 Uhr

Vor allem wegen des Preisschubs bei Öl und Gas ist die Inflationsrate in Deutschland im Mai auf 2,5 Prozent geklettert. So stark war die Teuerung zuletzt vor knapp zehn Jahren. Wie sind die Prognosen für die kommenden Monate?

MONTAG, 31. MAI 2021

NRW meldet meiste Verfahren

Fünf Länder ermitteln wegen Testbetrug

DONNERSTAG, 06. MAI 2021

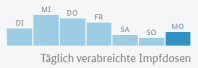
Partner sagt vor Ausschuss aus

EY hatte Hinweise auf Betrug bei Wirecard

> *"Numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning."*
>
> Nate Silver – The Signal and the Noise

43,3 %
Mindestens eine Impfdosis

50,5 Mio.
Verabreichte Impfdosen



Täglich verabreichte Impfdosen

+511.097
Montag 31.05.21

Wie ist der Fortschritt der COVID-19-Impfung?

Aktueller Impfstatus

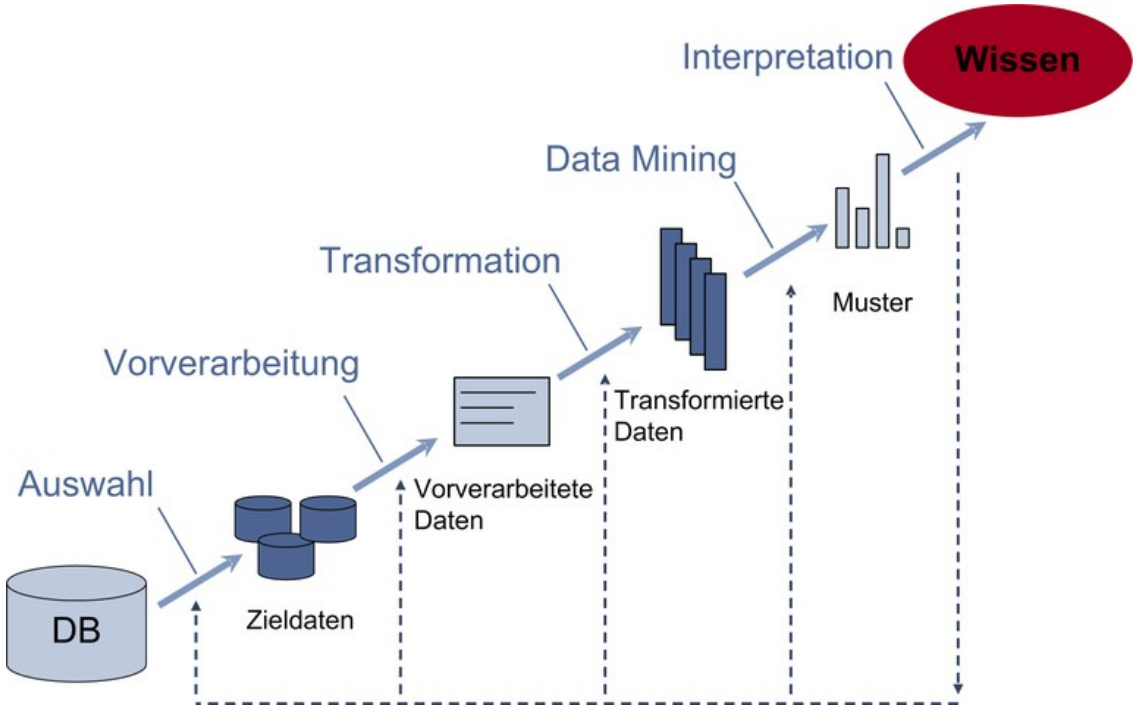
Am 31. Mai 2021 wurden in Deutschland 511.097 Impfdosen verabreicht. Damit sind nun 15.009.970 Personen (18,0% der Gesamtbevölkerung) vollständig geimpft. Insgesamt haben 36.004.055 Personen (43,3%) mindestens eine Impfdosis erhalten.

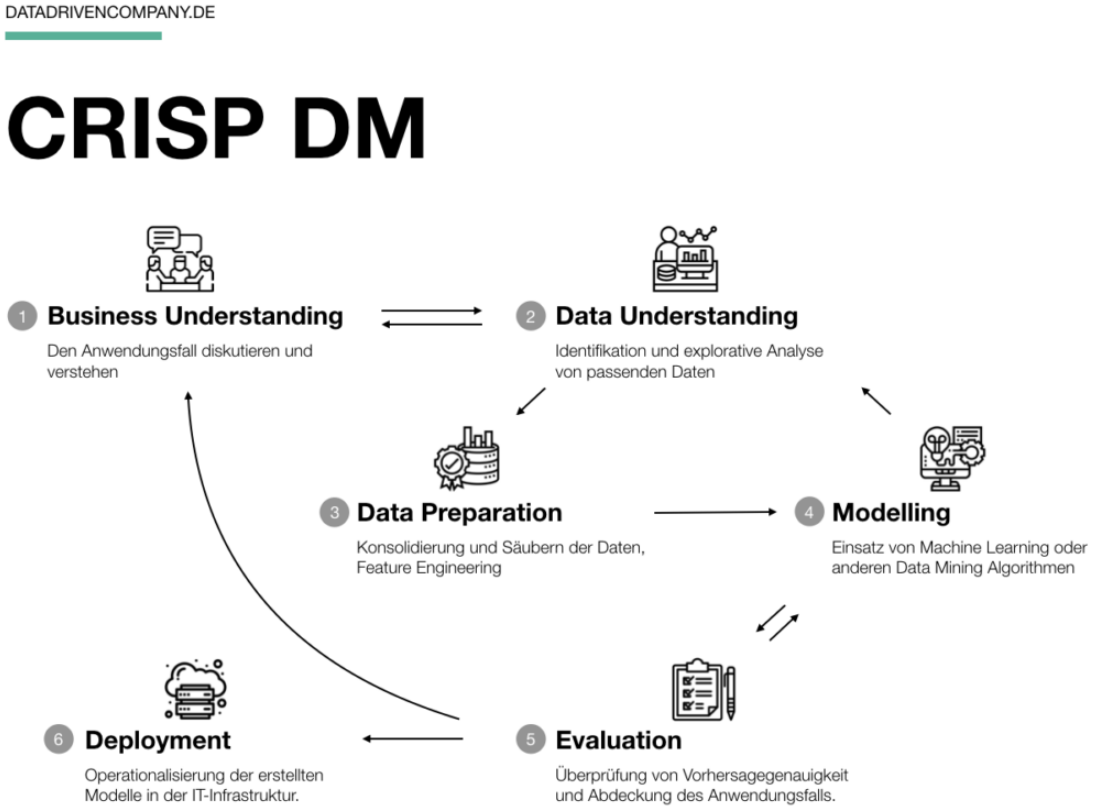
Arbeitsmarkt

Arbeitsmarkt erholt sich: Trotzdem viele Langzeitarbeitslose

Data Mining Vorgehensmodelle

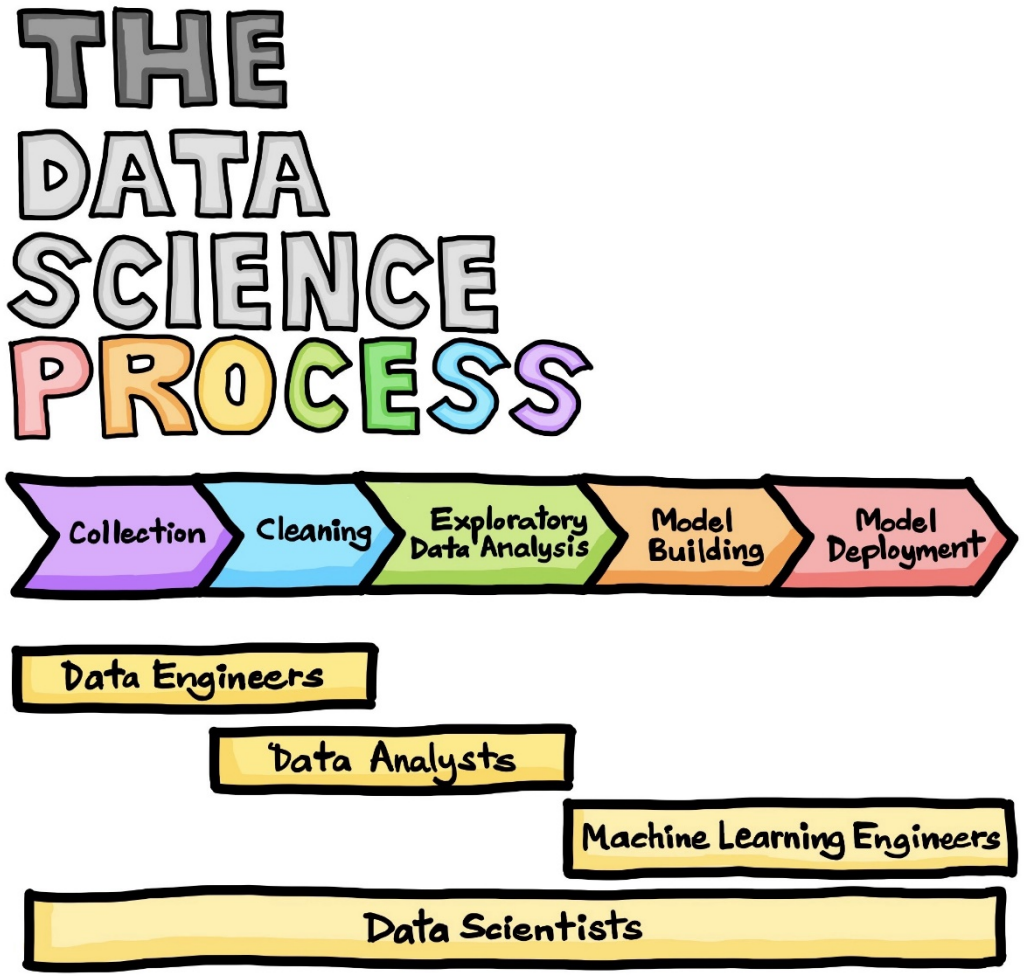
KDD – Knowledge Discovery in Databases

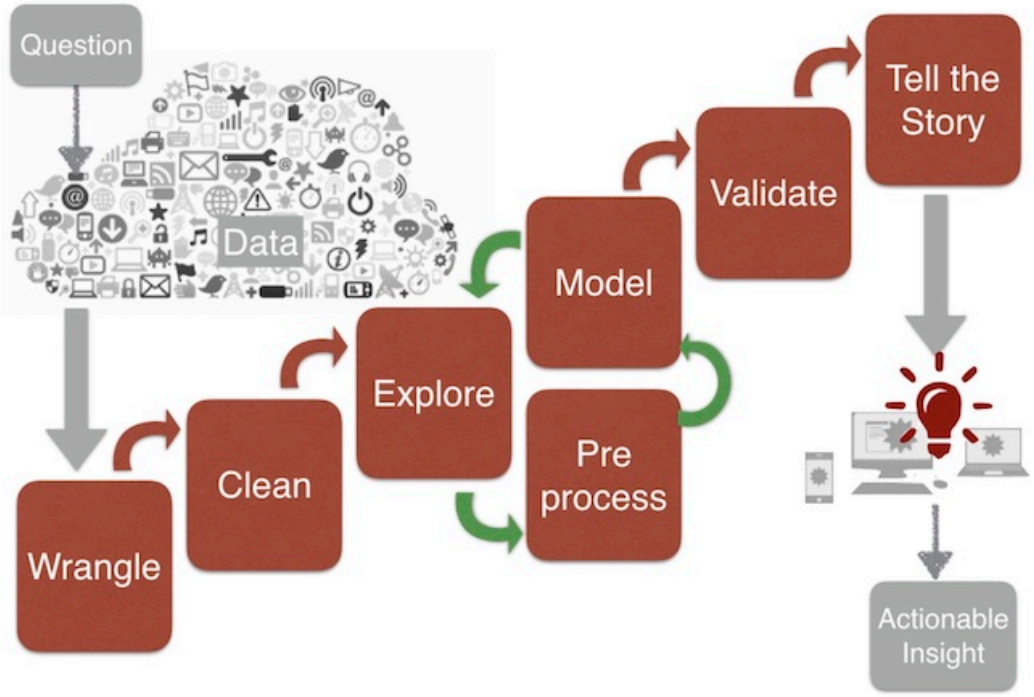




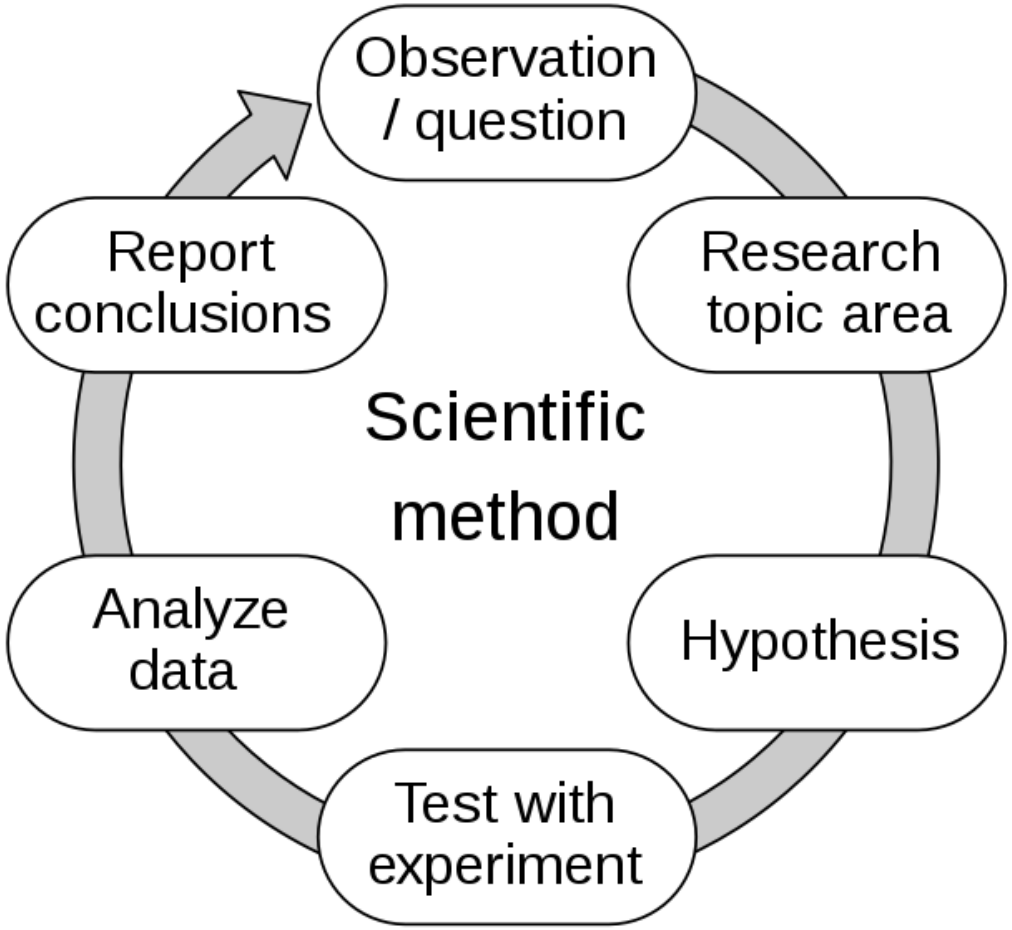
CRISP-DM – CRoss Industry Standard Process for Data Mining

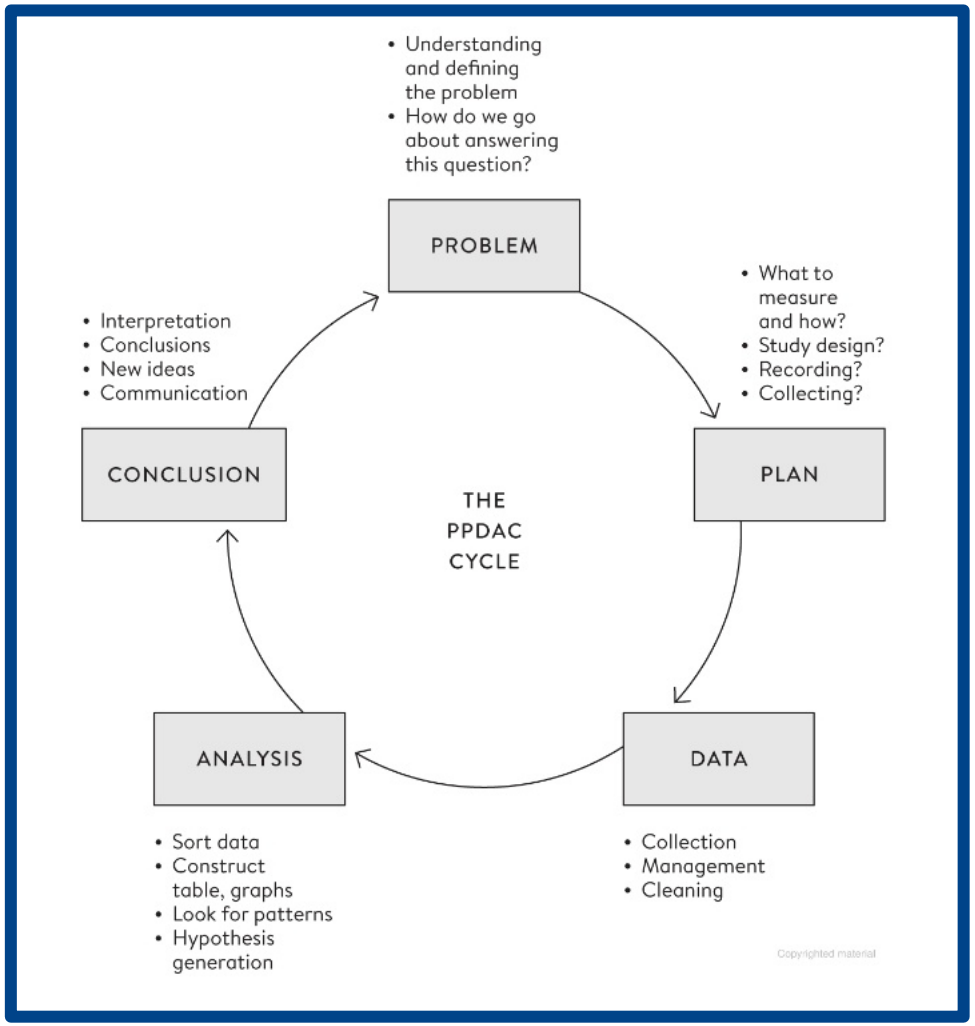
Data Science Pipelines





Die wissenschaftliche Methode





Abgrenzung zu klassicher Statistikausbildung

- Der traditionelle Statistikkurs
  - Beschreiben von Daten mit zusammenfassender Statistik
  - Wahrscheinlichkeitstheorie zur Ziehung von Zufallsbeobachtungen aus einer Populationsverteilung
  - Wahrscheinlichkeitsrechnung für Verteilungen der zusammenfassenden Statistik
  - Formeln für statistische Tests
  - Wenn die Zeit reicht: Beispiele für die Anwendung statistischer Modelle im realen Leben

- Unser Fokus
  - Motivieren durch Problemlösung
  - Visualisierung und Datenexploration als Ausgangspunkt
  - Fokus auf das, was man vernünftigerweise aus Daten lernen kann und die dafür nötigen Modelle und Algorithmen

Wir satteln das Pferd vom Problem aus auf

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Agenda

1. Motivation
2. Ein berühmter Brite
3. Zahlendetektive
4. Ein Unglück erklärbar machen

Wofür ist dieser Mann berühmt?

- Fernsehstar der 80er
- Pionier der statistischen Inferenz
- Massenmörder
- Erbe einer berühmten Milliardärsfamilie

Wofür ist dieser Mann berühmt?

Harold Shipman

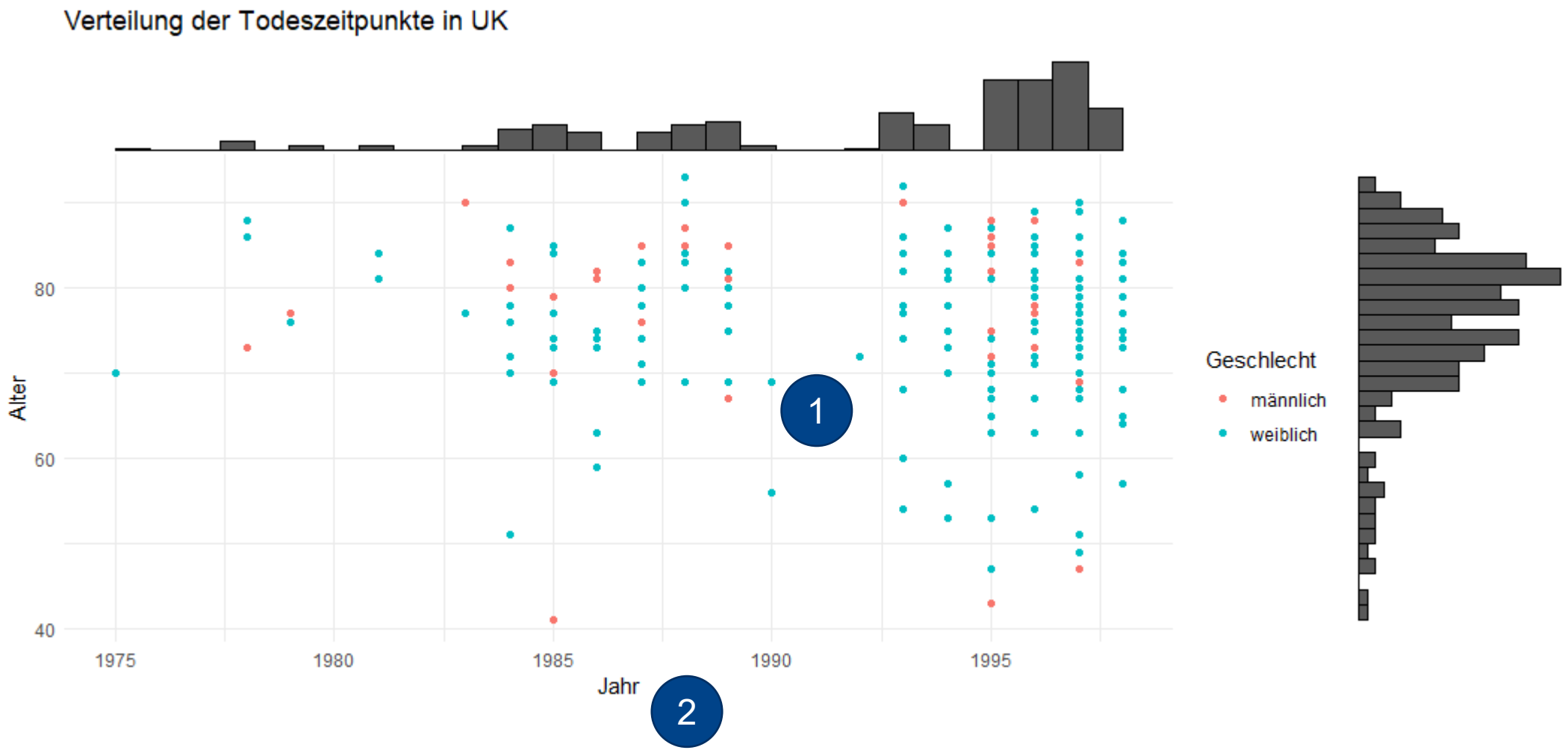
Harold Frederick Shipman (* 14. Januar 1946 in Nottingham; † 13. Januar 2004 in Wakefield) war ein britischer Hausarzt und 215–250-facher[1] Serienmörder.

1 Wer waren die Opfer?

2 Wann?

3 Wie ging er vor?

Ein Bild sagt mehr als tausend Worte



Wie kann Harold Shipman datenbasiert überführt werden

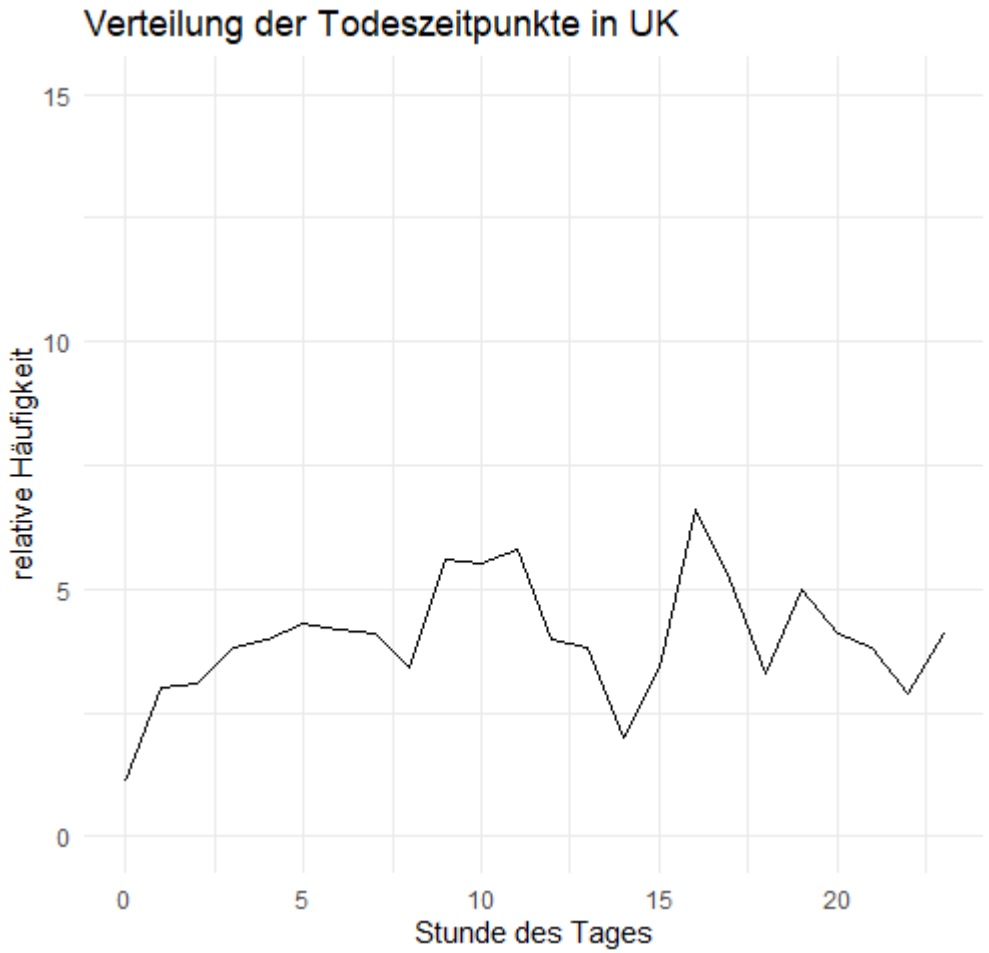
Hypothese: Seine Morde erfolgen im Rahmen seiner Tätigkeit als Allgemeinarzt

- Problem: Wie können wir sein Vorgehen nachweisen
- Plan: ???
- Daten: Sterbeurkunden der Patienten von Harold Shipman und anderen Ärzten in UK
- Analyse: ???

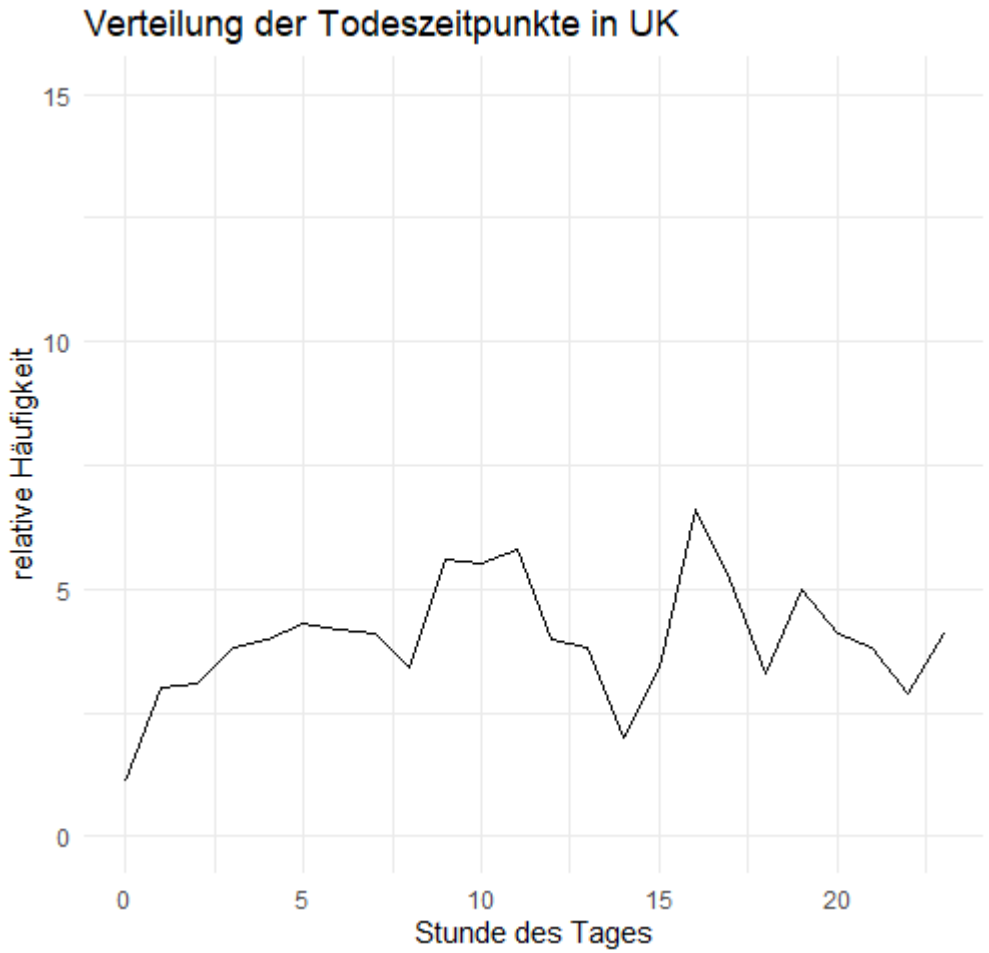
Wie kann Harold Shipman datenbasiert überführt werden

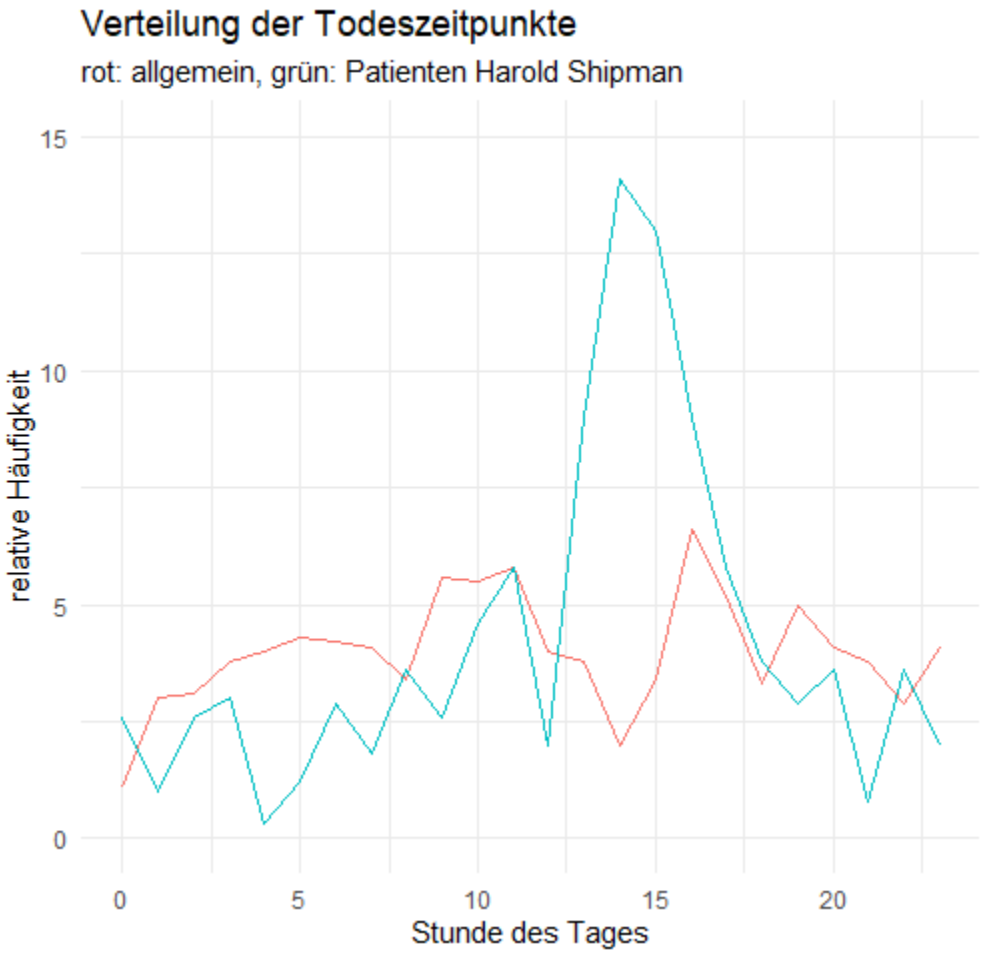
Hypothese: Seine Morde erfolgen im Rahmen seiner Tätigkeit als Allgemeinarzt

- Problem: Wie können wir sein Vorgehen nachweisen
- Plan: Vergleich der tatsächlichen Zeiten, zu denen seine Patienten starben, mit den Zeiten mit den Sterbezeiten anderer lokaler Hausärzte
- Daten: Sterbeurkunden der Patienten von Harold Shipman und anderen Ärzten in UK
- Analyse: einfaches Plotten..... (!)


Verteilung der Todeszeitpunkte in UK
relative Häufigkeit
0
5
10
15
0
5
10
15
20
Stunde des Tages

Ein Bild sagt mehr als tausend Worte (2)



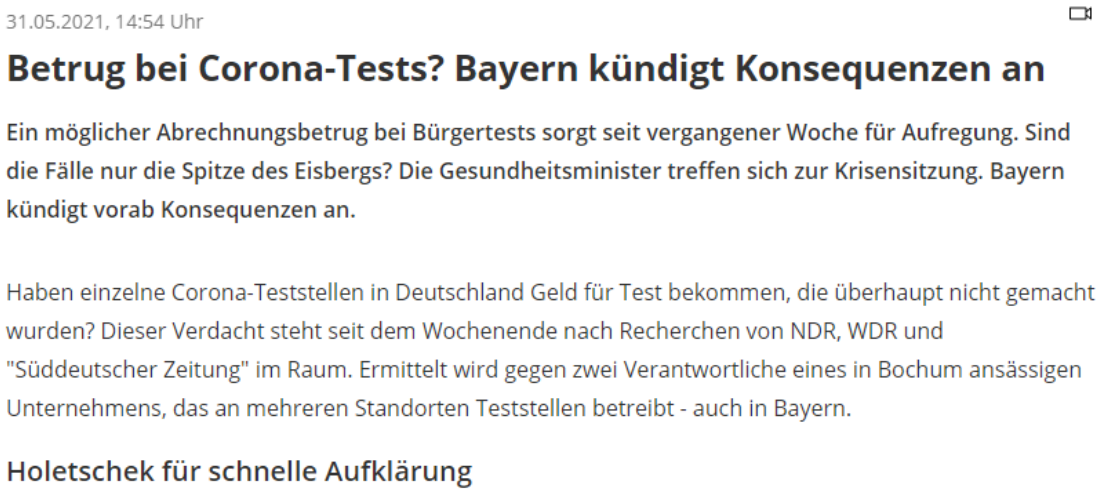


Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Agenda

1. Motivation
2. Ein berühmter Brite
3. Zahlendetektive
4. Ein Unglück erklärbar machen

Wie könnte man hier vorgehen?



31.05.2021, 14:54 Uhr

Betrug bei Corona-Tests? Bayern kündigt Konsequenzen an

Ein möglicher Abrechnungsbetrug bei Bürgertests sorgt seit vergangener Woche für Aufregung. Sind die Fälle nur die Spitze des Eisbergs? Die Gesundheitsminister treffen sich zur Krisensitzung. Bayern kündigt vorab Konsequenzen an.

Haben einzelne Corona-Teststellen in Deutschland Geld für Test bekommen, die überhaupt nicht gemacht wurden? Dieser Verdacht steht seit dem Wochenende nach Recherchen von NDR, WDR und "Süddeutscher Zeitung" im Raum. Ermittelt wird gegen zwei Verantwortliche eines in Bochum ansässigen Unternehmens, das an mehreren Standorten Teststellen betreibt - auch in Bayern.

Holetschek für schnelle Aufklärung

- Problem: Identifikation von Testbetrug
- Plan: Erkennen von Anomalien der berichteten Testzahlen
- Daten: Testzahlen für verschiedene Tage
- Analyse: ???

Ein Beispiel für das Beispiel

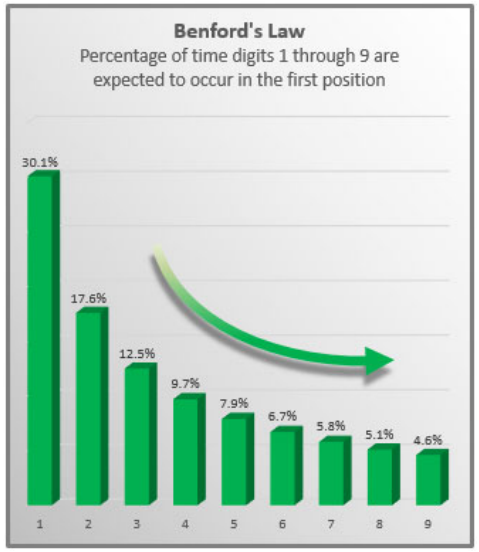
Regionalschlüssel	Gemeinde_Gemeindeverband	Schulden des öffentlichen Bereichs insgesamt	Erste Ziffer
091610000000	Ingolstadt	646.749.649	6
091620000000	München, Landeshauptstadt	6.051.323.887	6
091630000000	Rosenheim	244.405.979	2
09171	Altötting	53.684.915	5
091710111111	Altötting, St	26.276.311	2
091710112112	Burghausen, St	27.402.089	2
091710113113	Burgkirchen a.d.Alz	12.574.779	1
091710117117	Garching a.d.Alz	2.305.204	2
091710118118	Haiming	106.936	1
091710125125	Neuötting, St	15.369.158	1
091710127127	Pleiskirchen	1.722.537	1
091710131131	Teising	180.637	1
091710132132	Töging a.Inn, St	10.666.305	1
091710133133	Tüßling, M	3.093.093	3
⋮	⋮	⋮	⋮
097805745144	Weitnau, M	9.912.028	9
Anzahl		2.051	2.051
Minimum		660	1
Maximum		6.051.323.887	9

- Wie oft erwarten wir die Ziffer 6 im Datensatz mit 2051 Einträgen als erste Ziffer?
- Wie oft erwarten wir die Ziffer 1 als erste Ziffer?

- 228
- Häufiger
- Seltener
- Nicht beantwortbar

Gesetze für Zahlen

- Das **Benfordsche Gesetz** besagt, dass die Auftretenswahrscheinlichkeit der Ziffernsequenzen in den Zahlen nicht gleichverteilt ist, sondern logarithmischen Gesetzen folgt
- Das bedeutet, dass die Auftretenswahrscheinlichkeit einer Ziffernsequenz umso höher ist, je kleiner sie wertmäßig ist und je weiter links sie in der Zahl beginnt. Am häufigsten ist die Anfangssequenz „1“ mit theoretisch 30,103 %.



- Es gilt für reale Datensätze, die **genügend umfangreich sind** und Zahlen in der Größenordnung von **x bis mindestens 10000x** aufweisen

Was bedeutet das für unser Ausgangsproblem?

Schuldenbeispiel

	Naive Erwartung	Tatsächliche Häufigkeiten	Benford's Law	Benford p
1	228	611	617	30,1 %
2	228	381	361	17,6 %
3	228	245	256	12,5 %
4	228	190	199	9,7 %
5	228	161	162	7,9 %
6	228	138	137	6,7 %
7	228	127	119	5,8 %
8	228	114	105	5,1 %
9	228	84	94	4,6 %

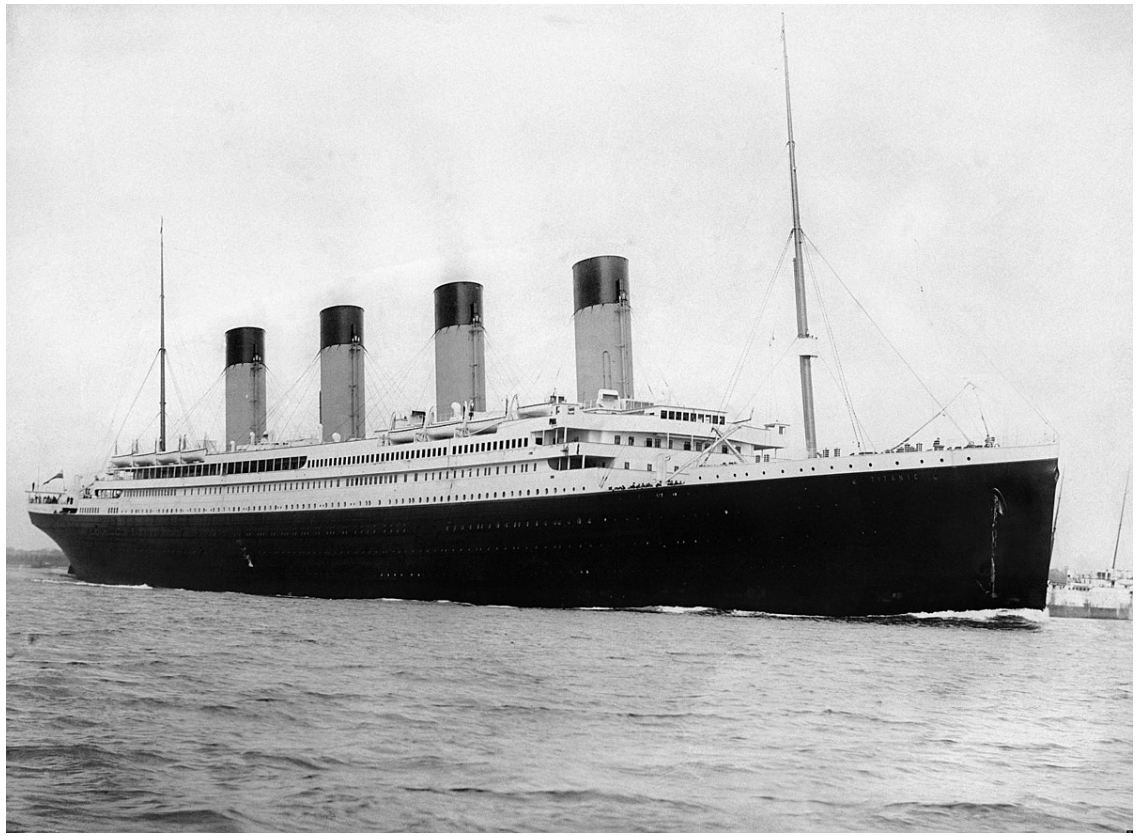
Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Agenda

1. **Motivation**
2. **Ein berühmter Brite**
3. **Zahlendetektive**
4. **Ein Unglück erklärbar machen**

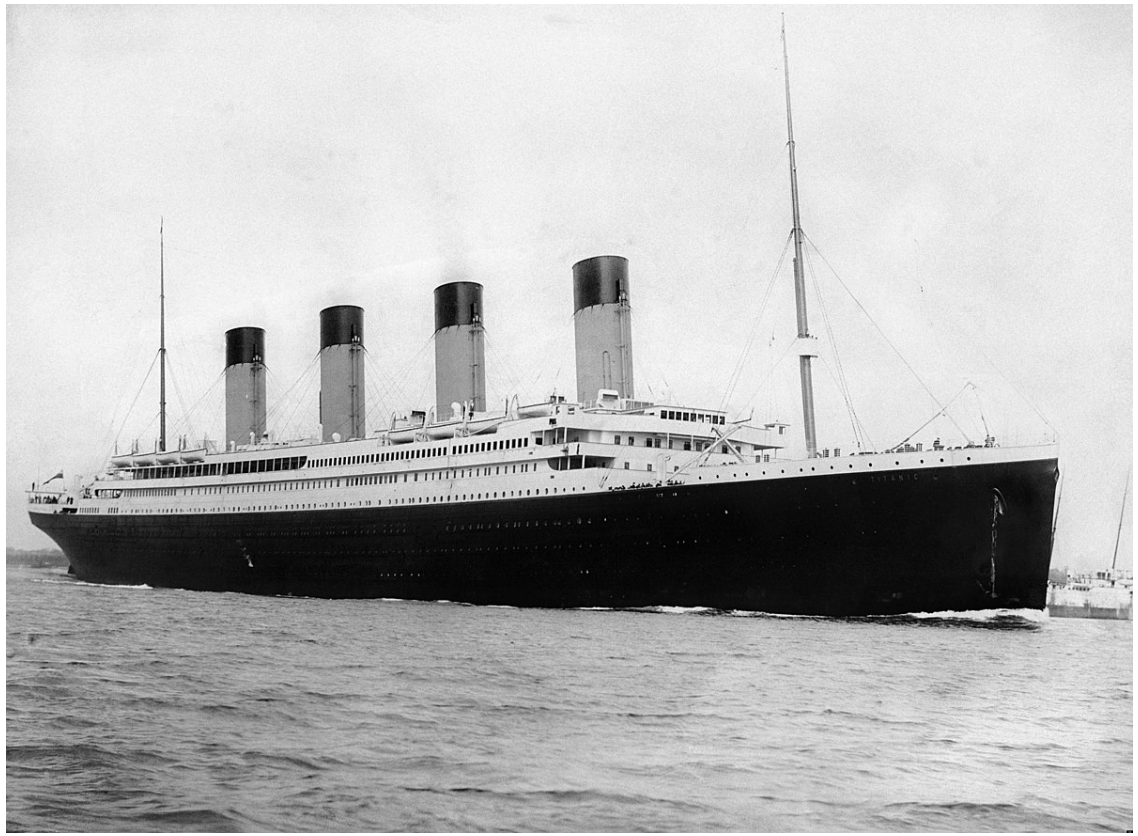
1912, ein Eisberg und die Katstrophe

- Wer hat das Unglück überlebt?



1912, ein Eisberg und die Katstrophe

- Wer hat das Unglück überlebt?



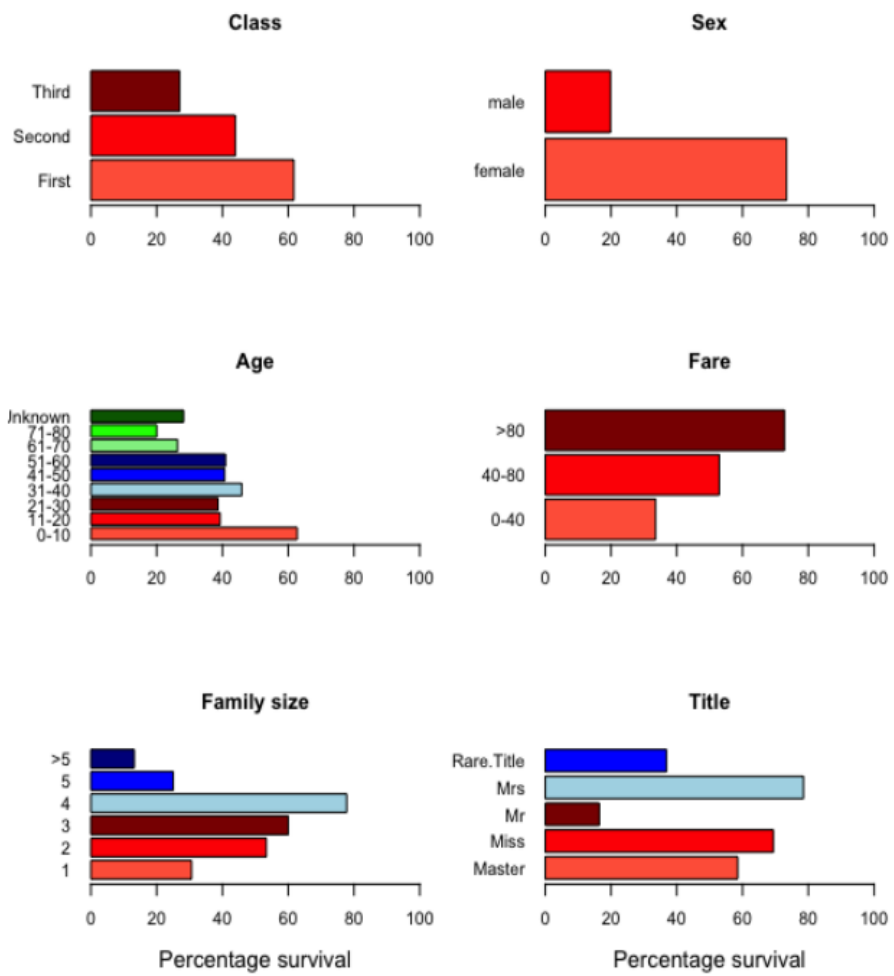
Erzählt Hollywood die Wahrheit?

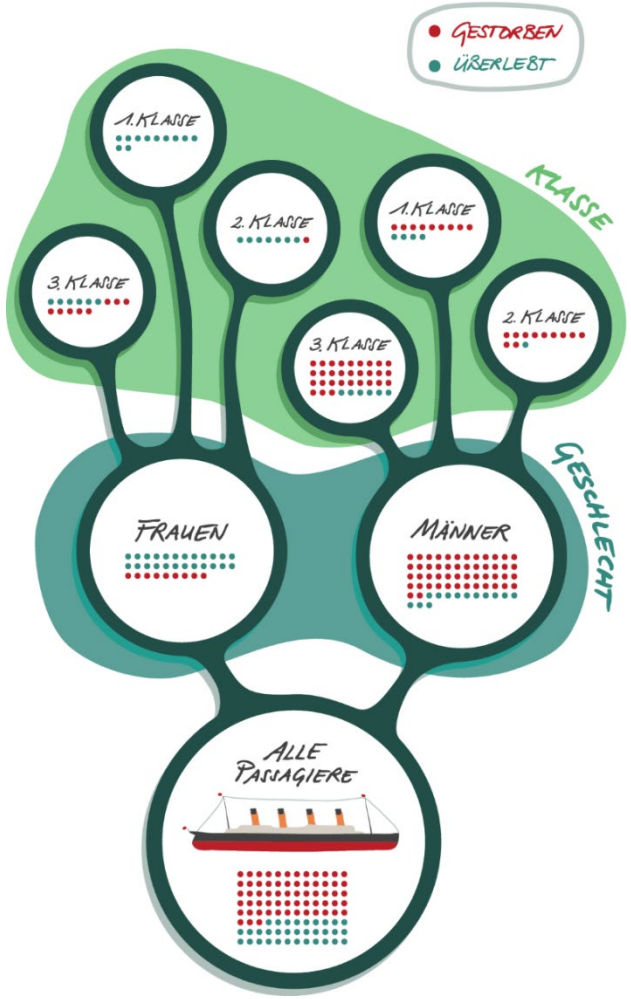
PPDAC

- Problem: Was beeinflusst die Überlebenswkt. auf der Titanic?
- Plan: Überlebenswkt. verschiedener Merkmalskombinationen bestimmen
- Daten: Passagierdatenbank der Titanic
  - Teilweise unvollständig, teilweise weitere relevante Daten ergänzbar
  - https://github.com/codebasics/py/blob/master/ML/9_decision_tree/Exercise/titanic.csv
- Analyse: Regression, Entscheidungsbaum, …

pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body
3	0	Somerton, Mr. Francis William	male	30	0	0	A.5. 18509	8.0500		S		
3	0	Spector, Mr. Woolf	male		0	0	A.5. 3236	8.0500		S		
3	0	Spinner, Mr. Henry John	male	32	0	0	STON/OQ. 369943	8.0500		S		
3	0	Staneff, Mr. Ivan	male		0	0	349208	7.8958		S		
3	0	Stankovic, Mr. Ivan	male	33	0	0	349239	8.6625		C		
3	1	Stanley, Miss. Amy Zillah Elsie	female	23	0	0	CA. 2314	7.5500		S	C	
3	0	Stanley, Mr. Edward Roland	male	21	0	0	A/4 45380	8.0500		S		

Stereotypen scheinen korrekt

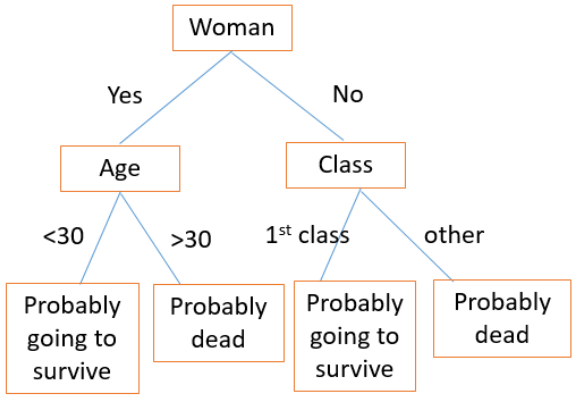


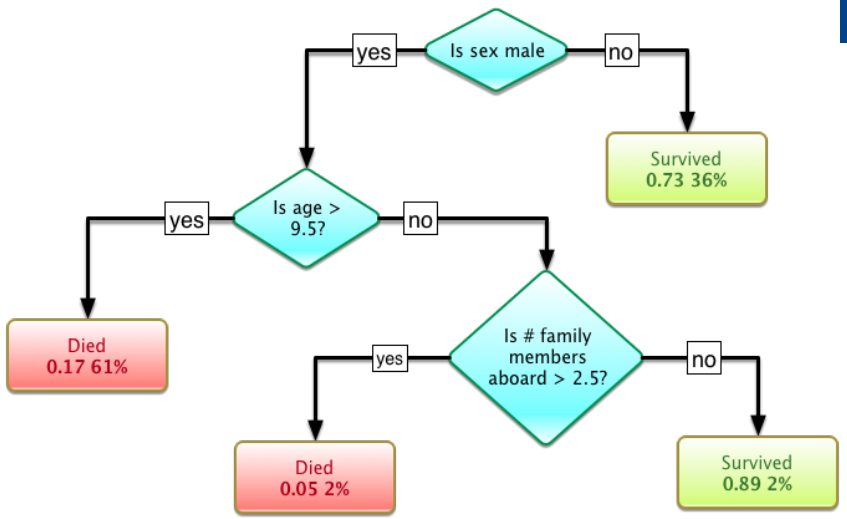


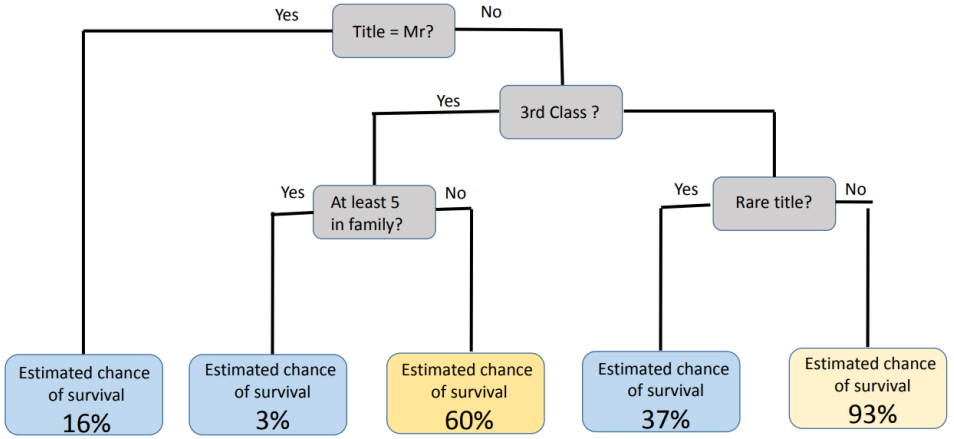
Viele Bäume sind möglich…

A normal tree

A decision tree

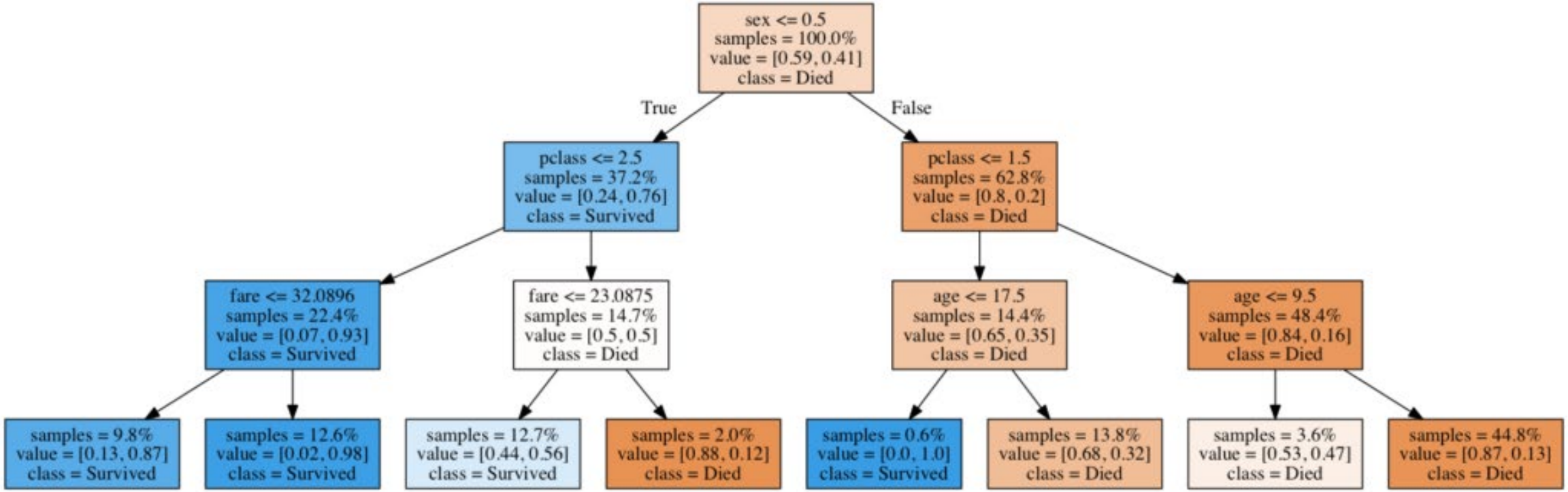






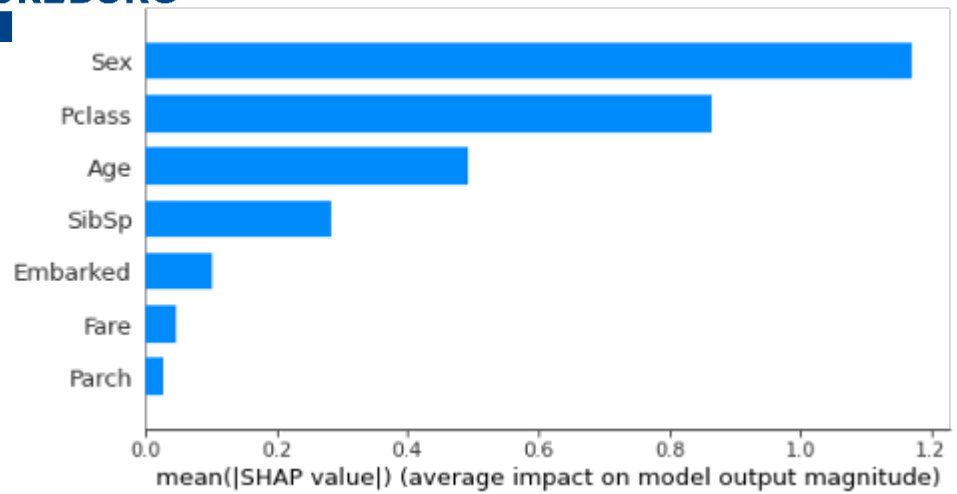
Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

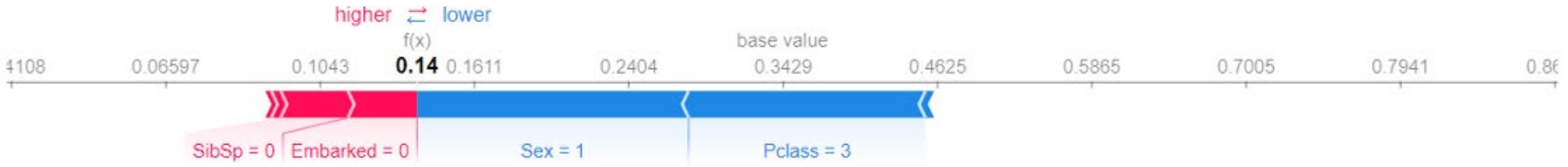
Etwas technischer geht's auch

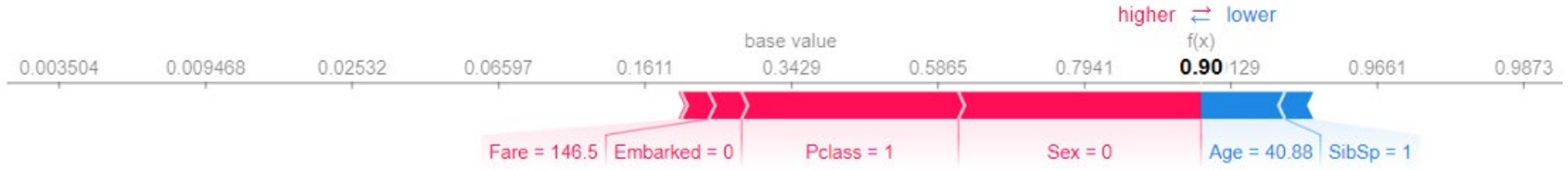


Geht es um Erklärung oder Vorhersage?

Erklärbarkeit als wichtiges Ziel von Datenanalyse (C!)







Bonusfrage: Wer war der größte Glückspilz auf der Titanic?

- Problem: Größten Glückspilz identifizieren
- Plan: ?
- Daten: ?
- Analyse: ?

Bonusfrage: Wer war der größte Glückspilz auf der Titanic?

- Vorbemerkung 1: Sicherlich nicht Leonardo
- Vorbemerkung 2: Glückspilz sollte überlebt haben

- Problem: Größten Glückspilz identifizieren
- Plan: ?
- Daten: ?
- Analyse: ?

```
> titanic %>% filter(Survived==1) %>%
+   filter(Sex == "male") %>%
+   filter(Age > 40) %>%
+   filter(Pclass == 3) %>% glimpse()
Rows: 2
Columns: 12
$ PassengerId <dbl> 339, 415
$ Survived    <dbl> 1, 1
$ Pclass      <dbl> 3, 3
$ Name        <chr> "Dahl, Mr. Karl Edwart", "Sundman, Mr. J...
$ Sex         <chr> "male", "male"
$ Age         <dbl> 45, 44
$ SibSp       <dbl> 0, 0
$ Parch       <dbl> 0, 0
$ Ticket      <chr> "7598", "STON/O 2. 3101269"
$ Fare        <dbl> 8.050, 7.925
$ Cabin       <chr> NA, NA
$ Embarked    <chr> "S", "S"
```