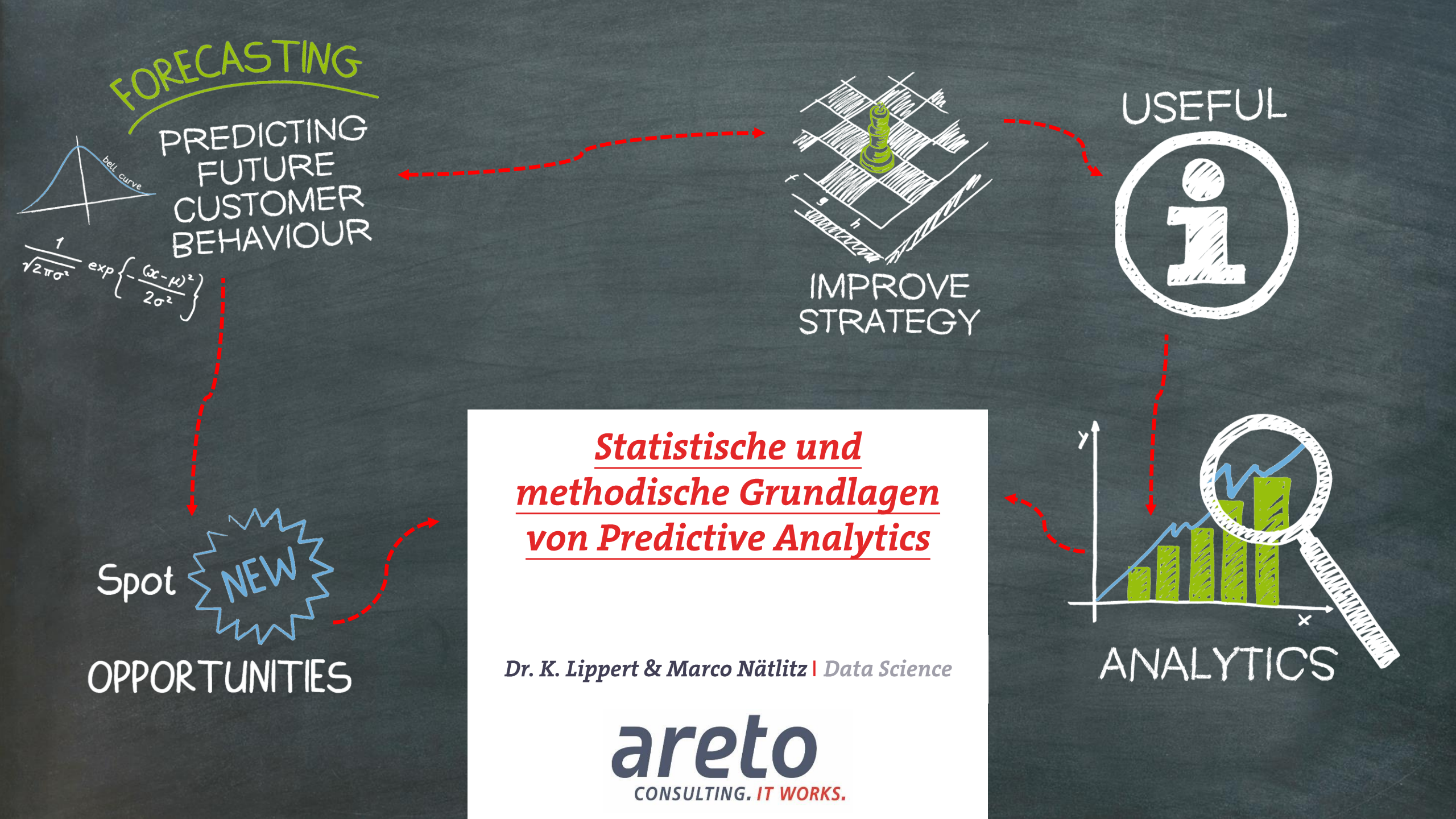


Statistische und methodische Grundlagen von Predictive Analytics

Dr. K. Lippert & Marco Nätlitz | Data Science

areto
CONSULTING. IT WORKS.

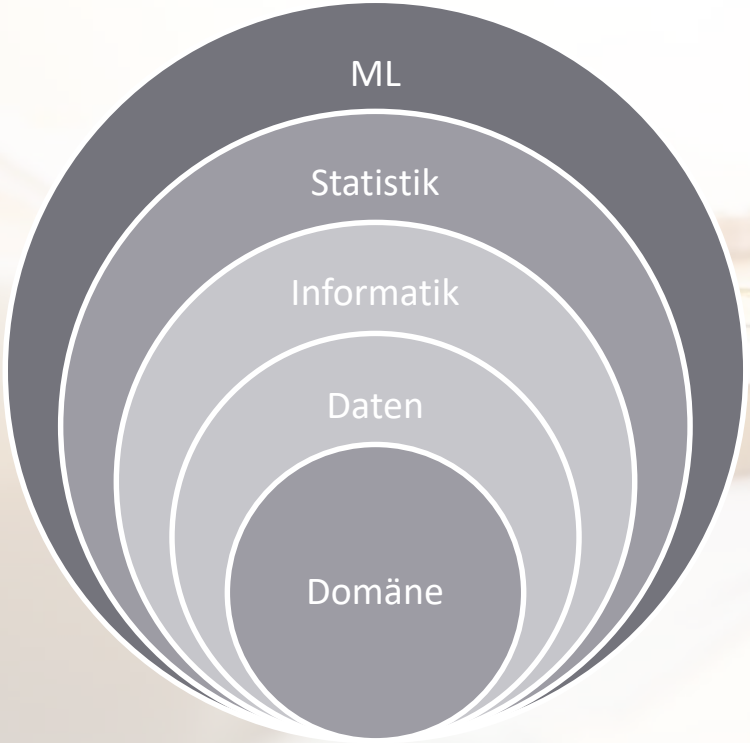
Agenda

Statistische und methodische Grundlagen von Predictive Analytics

- Einordnung von Data Science und Predictive Analytics
- Der Weg zum Vorhersagemodell anhand konkreter Beispiele:
  - Vorhersage des Preises von Bordeaux-Weinen (Regression)
  - Vorhersage der Überlebenschance auf der Titanic (Klassifikation)
  - Umsatzprognose mit Einzelhandel (Autoregression)
- Beispiele mit R und SQL
- Fazit & Ausblick

Was ist eigentlich Data Science?

Data Science beschreibt die Extraktion von Wissen aus Daten, Begriff geprägt durch Naur (1974)



Warum Data Science?

- knappe Ressourcen
- schnelle Märkte
- Wettbewerbsvorteile

Wettbewerbsvorteile durch Data Science

"The median ROI of Predictive Analytics projects is close to three times higher than that of non-predictive projects" (IDC)

Methoden-Dschungel

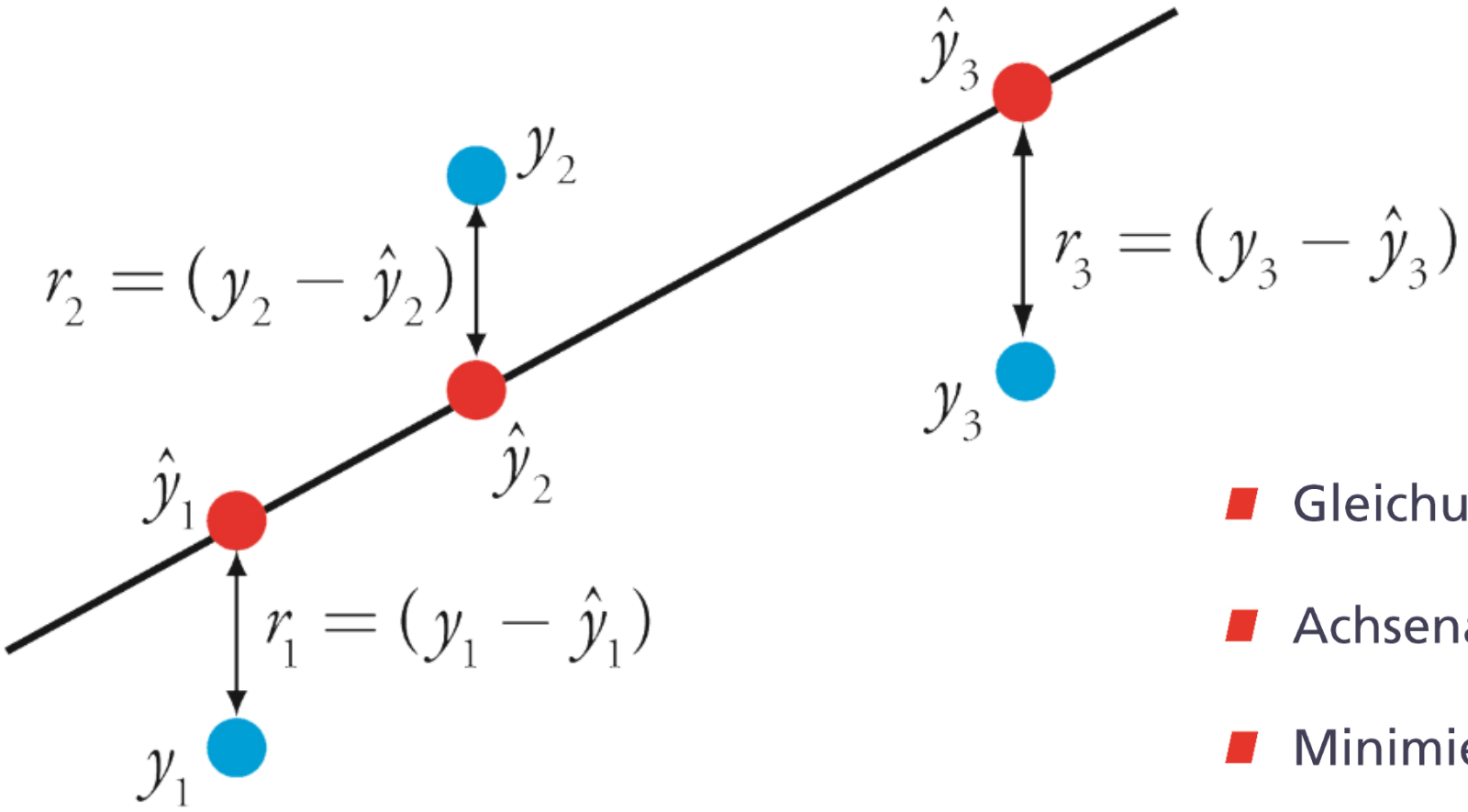
- Business vor Methodik
- Pareto-Prinzip
- Schnell erste Ergebnisse

Wie viel kostet ein Bordeaux? (Regression)

Foto: © Grecaud Paul | fotolia.de

Exkurs: lineare Regression

Wie bestimmt man die beste Gerade in einer Punktwolke?



- Gleichung: $y = ß_0 + ß_1 x$
- Achsenabschnitt $ß_0$, Steigung $ß_1$
- Minimierungsproblem

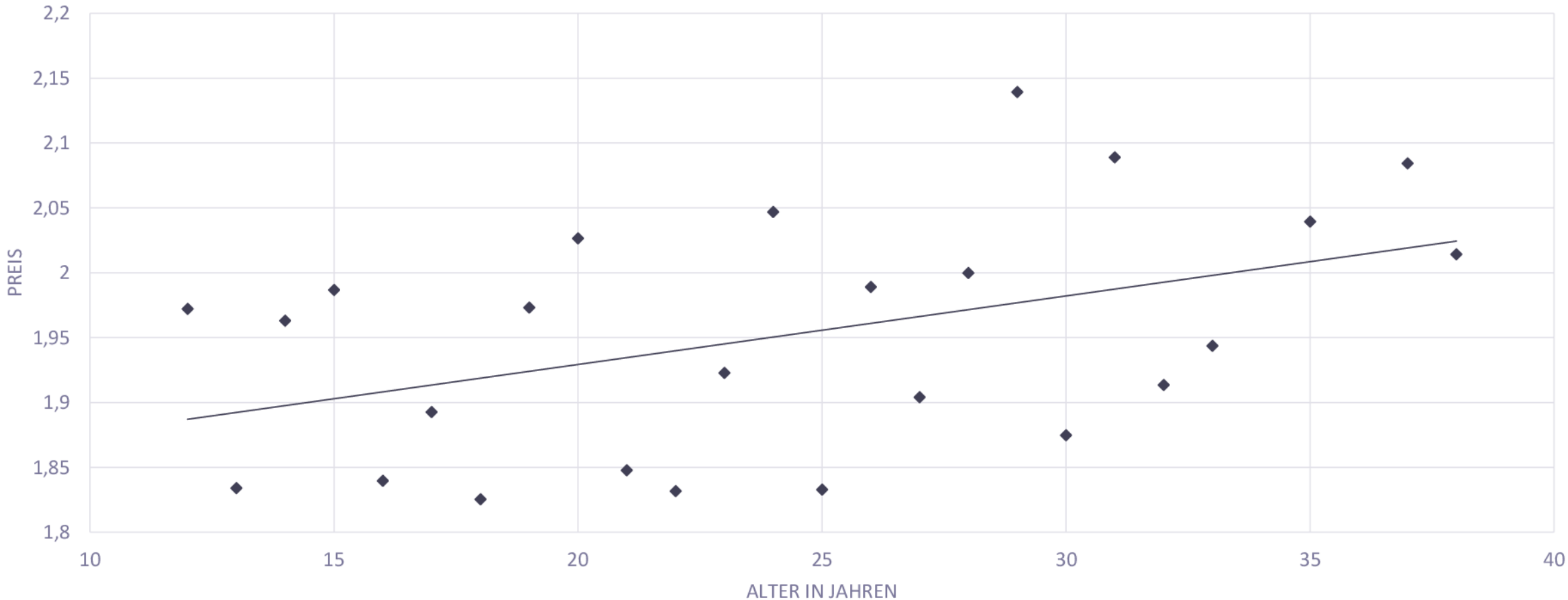
	YEAR	PRICE	WINTERRAIN	AGST	HARVESTRAIN	AGE	FRANCEPOP
1	1952	7,495	600	17,117	160	31	43183,569
2	1953	8,039	690	16,733	80	30	43495,03
3	1955	7,686	502	17,15	130	28	44217,857
4	1957	6,985	420	16,133	110	26	45152,252
5	1958	6,777	582	16,417	187	25	45653,805
6	1959	8,076	485	17,483	187	24	46128,638
7	1960	6,519	763	16,417	290	23	46583,995
8	1961	8,494	830	17,333	38	22	47128,005
9	1962	7,388	697	16,3	52	21	48088,673
10	1963	6,713	608	15,717	155	20	48798,99
11	1964	7,309	402	17,267	96	19	49356,943
12	1965	6,252	602	15,367	267	18	49801,821
13	1966	7,744	819	16,533	86	17	50254,966
14	1967	6,84	714	16,233	118	16	50650,406
15	1968	6,244	610	16,2	292	15	51034,413
16	1969	6,346	575	16,55	244	14	51470,276
17	1970	7,588	622	16,667	89	13	51918,389
18	1971	7,193	551	16,767	112	12	52431,647
19	1972	6,205	536	14,983	158	11	52894,183
20	1973	6,637	376	17,067	123	10	53332,805
21	1974	6,294	574	16,3	184	9	53689,61
22	1975	7,292	572	16,95	171	8	53955,042
23	1976	7,121	418	17,65	247	7	54159,049
24	1977	6,259	821	15,583	87	6	54378,362
25	1978	7,186	763	15,817	51	5	54602,193
26	1979	(null)	717	16,167	122	4	54835,832
27	1980	(null)	578	16	74	3	55110,236

Die Tabelle „wine“

Weinauktionen von 1952 bis 1980

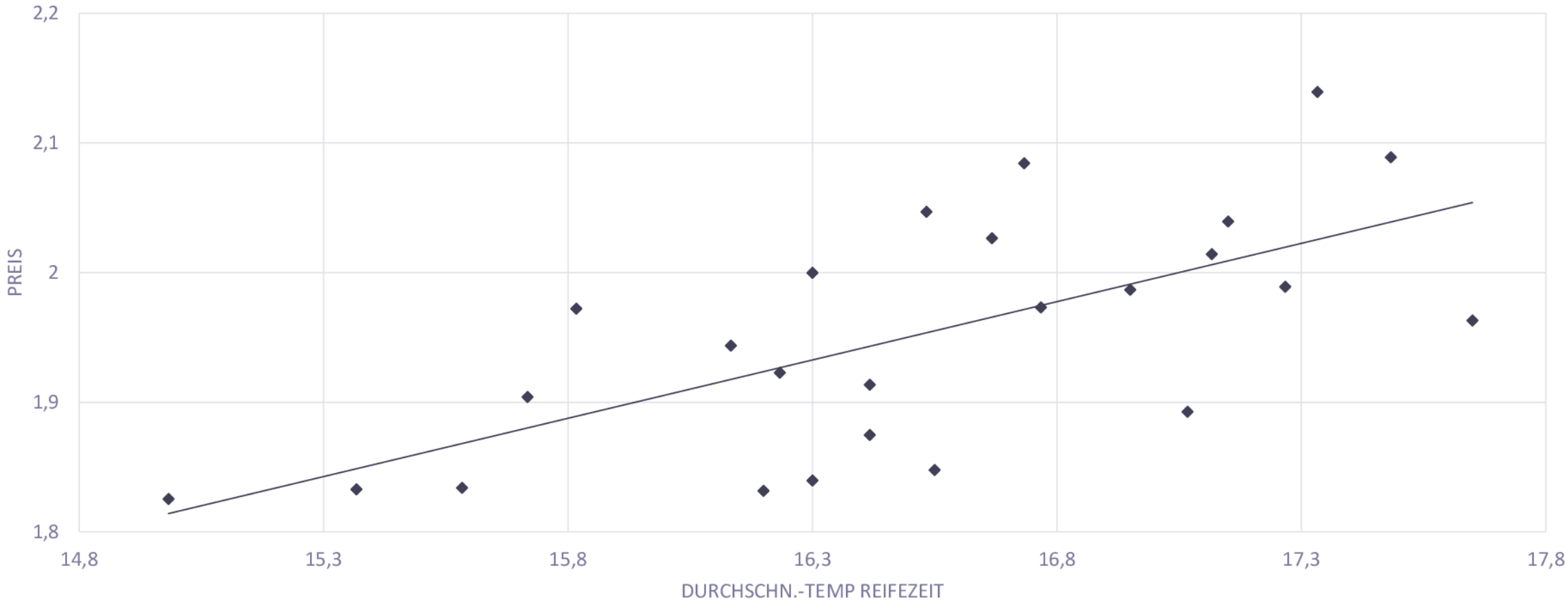
Es besteht ein linearer Zusammenhang zwischen Preis und Alter

Die Wein-Formel im Detail



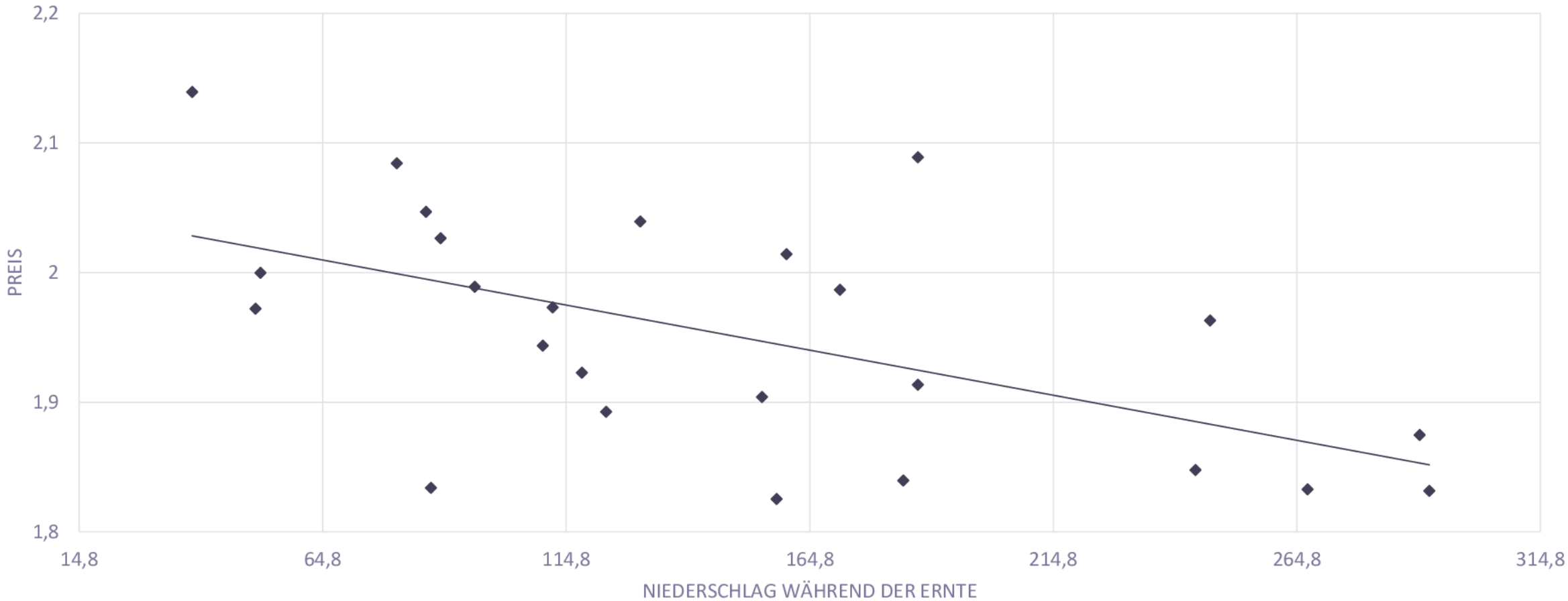
Heiße und trockene Sommer garantieren den typischen Geschmack

Die Wein-Formel im Detail



Wenig Niederschlag während der Ernte erhöht die Qualität

Die Wein-Formel im Detail



```
select
    year
  , (...)
  , price
  , prediction(
        for
          price
        using
            agst
          , winterrain
          , harvestrain
          , age
    ) over ( ) price_pred
from
   wine
```

Regression mit „prediction“

Bei numerischen Zielvariablen

	YEAR	WINTERRAIN	AGST	HARVESTRAIN	AGE	FRANCEPOP	PRICE	PRICE_PRED
1	1952	600	17,117	160	31	43183,569	7,495	7,686
2	1953	690	16,733	80	30	43495,03	8,039	7,849
3	1955	502	17,15	130	28	44217,857	7,686	7,793
4	1957	420	16,133	110	26	45152,252	6,985	7,091
5	1958	582	16,417	187	25	45653,805	6,777	6,969
6	1959	485	17,483	187	24	46128,638	8,076	7,886
7	1960	763	16,417	290	23	46583,995	6,519	6,709
8	1961	830	17,333	38	22	47128,005	8,494	8,303
9	1962	697	16,3	52	21	48088,673	7,388	7,58
10	1963	608	15,717	155	20	48798,99	6,713	6,521
11	1964	402	17,267	96	19	49356,943	7,309	7,389
12	1965	602	15,367	267	18	49801,821	6,252	6,336
13	1966	819	16,533	86	17	50254,966	7,744	7,554
14	1967	714	16,233	118	16	50650,406	6,84	7,001
15	1968	610	16,2	292	15	51034,413	6,244	6,366
16	1969	575	16,55	244	14	51470,276	6,346	6,538
17	1970	622	16,667	89	13	51918,389	7,588	7,398
18	1971	551	16,767	112	12	52431,647	7,193	7,186
19	1972	536	14,983	158	11	52894,183	6,205	6,396
20	1973	376	17,067	123	10	53332,805	6,637	6,828
21	1974	574	16,3	184	9	53689,61	6,294	6,483
22	1975	572	16,95	171	8	53955,042	7,292	7,104
23	1976	418	17,65	247	7	54159,049	7,121	7,197
24	1977	821	15,583	87	6	54378,362	6,259	6,589
25	1978	763	15,817	51	5	54602,193	7,186	6,996
26	1979	717	16,167	122	4	54835,832	(null)	6,84
27	1980	578	16	74	3	55110,236	(null)	7,013

Ergebnisse

Prognosewerte und Realwerte

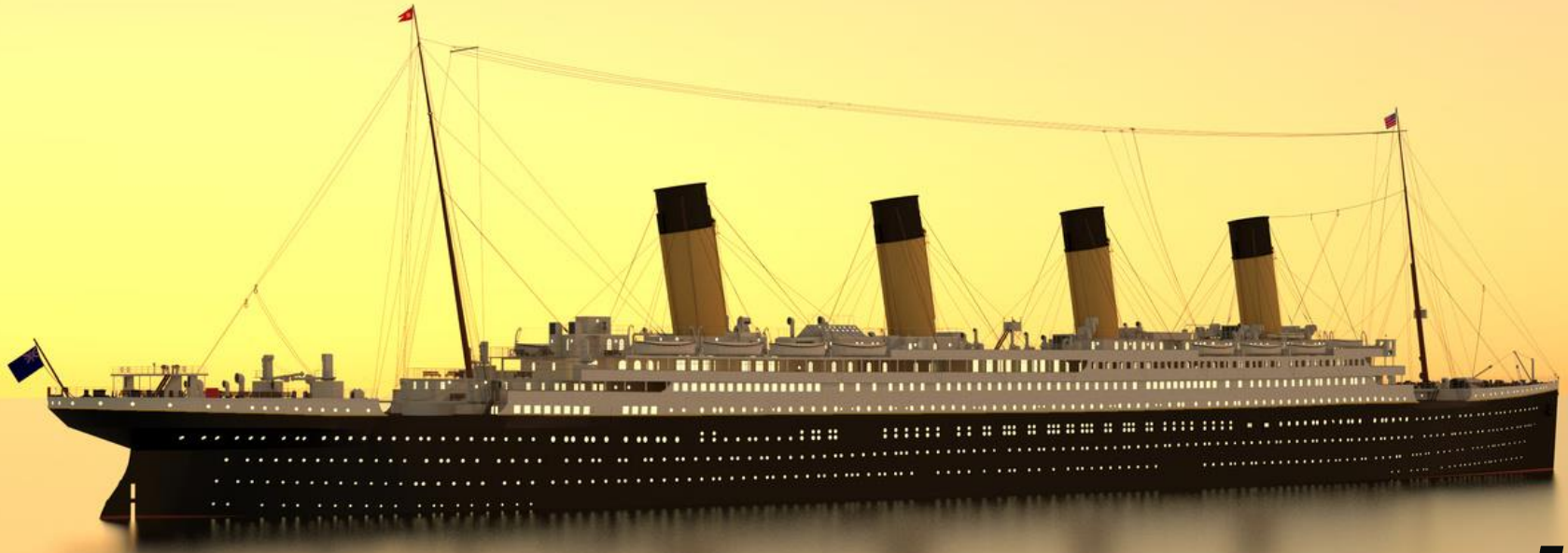
```
select
    sum( power( price_pred - x.ap, 2 ) )
    /
    sum( power( price - x.ap, 2 ) ) r2
  from
      wine_model
    , ( select
          avg( price ) ap
        from
          wine_model
      ) x
```

$R^2 \sim 82\%$

Wie gut ist mein Modell?

$$\frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Via https://de.wikipedia.org/wiki/Bestimmtheitsmass



Vorhersage der Überlebenschance auf der Titanic (Klassifikation)

	PA...	SURVIVED	PCLASS	NAME	SEX	AGE	SIBSP	PARCH	TICKET	FARE	CABIN
1	409	0	3	Birkeland,...	male	21	0	0	312992	7,775	(null)
2	410	0	3	Lefebre, M...	female	(null)	3	1	4133	25,4667	(null)
3	411	0	3	Sdycoff, M...	male	(null)	0	0	349222	7,8958	(null)
4	412	0	3	Hart, Mr. ...	male	(null)	0	0	394140	6,8583	(null)
5	413	1	1	Minahan, M...	female	33	1	0	19928	90	C78
6	414	0	2	Cunningham...	male	(null)	0	0	239853	0	(null)
7	415	1	3	Sundman, M...	male	44	0	0	STON/O 2. 3101269	7,925	(null)
8	416	0	3	Meek, Mrs....	female	(null)	0	0	343095	8,05	(null)
9	417	1	2	Drew, Mrs....	female	34	1	1	28220	32,5	(null)
10	418	1	2	Silven, Mi...	female	18	0	2	250652	13	(null)
11	419	0	2	Matthews, ...	male	30	0	0	28228	13	(null)
12	420	0	3	Van Impe, ...	female	10	0	2	345773	24,15	(null)
13	421	0	3	Gheorgheff...	male	(null)	0	0	349254	7,8958	(null)
14	422	0	3	Charters, ...	male	21	0	0	A/5. 13032	7,7333	(null)
15	423	0	3	Zimmerman,...	male	29	0	0	315082	7,875	(null)
16	424	0	3	Danbom, Mr...	female	28	1	1	347080	14,4	(null)
17	425	0	3	Rosblom, M...	male	18	1	1	370129	20,2125	(null)
18	426	0	3	Wiseman, M...	male	(null)	0	0	A/4. 34244	7,25	(null)
19	427	1	2	Clarke, Mr...	female	28	1	0	2003	26	(null)
20	428	1	2	Phillips, ...	female	19	0	0	250655	26	(null)
21	429	0	3	Flynn, Mr....	male	(null)	0	0	364851	7,75	(null)
22	430	1	3	Pickard, M...	male	32	0	0	SOTON/O.Q. 392078	8,05	E10
23	431	1	1	Bjornstrom...	male	28	0	0	110564	26,55	C52
24	432	1	3	Thorneycro...	female	(null)	1	0	376564	16,1	(null)
25	433	1	2	Louch, Mrs...	female	42	1	0	SC/AH 3085	26	(null)
26	434	0	3	Kallio, Mr...	male	17	0	0	STON/O 2. 3101274	7,125	(null)
27	435	0	1	Silvey, Mr...	male	50	1	0	13507	55,9	E44
28	436	1	1	Carter, Mi...	female	14	1	2	113760	120	B96 B98
29	437	0	3	Ford, Miss...	female	21	2	2	W./C. 6608	34,375	(null)

Der Titanic-Datensatz

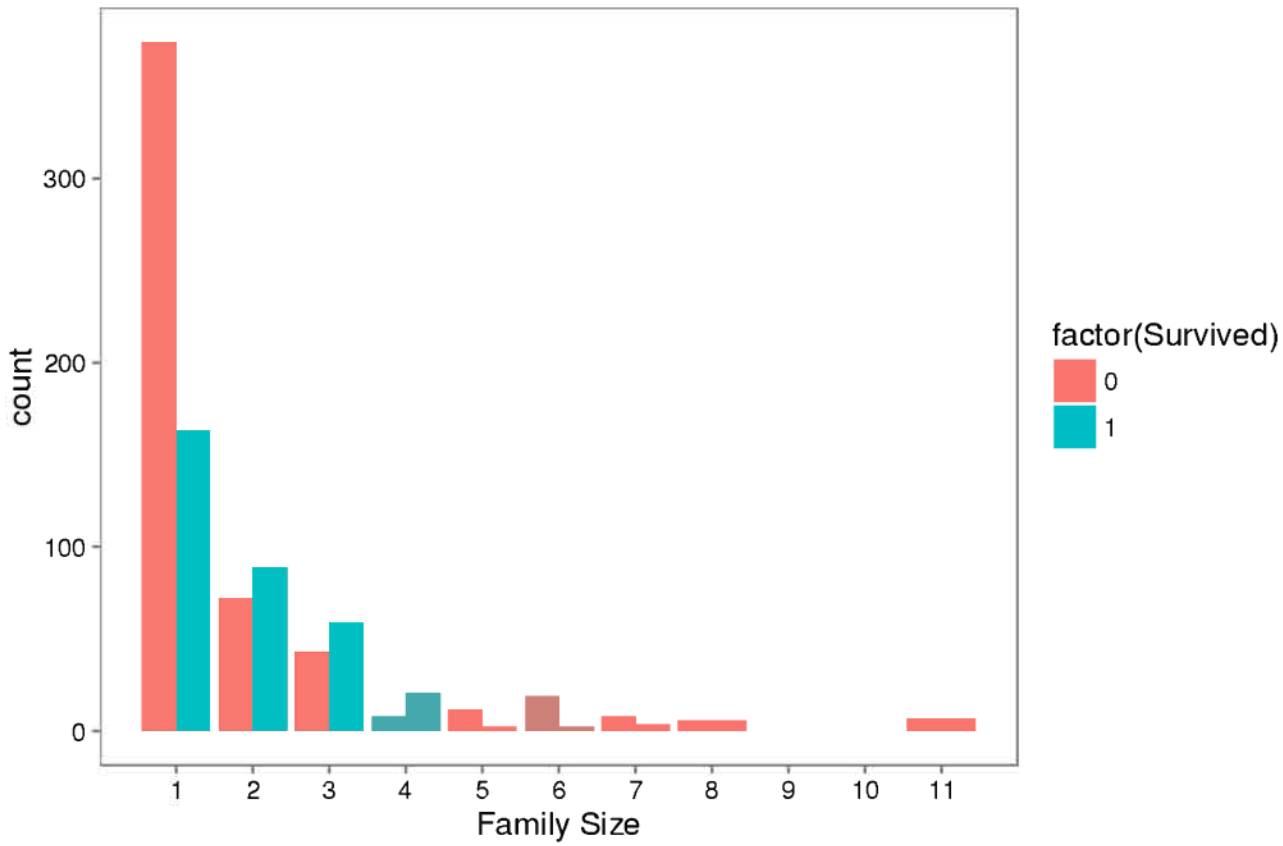
Machine Learning from Disaster

Via https://www.kaggle.com/c/titanic

Gruppe	Gesamt	Opfer	Anteil
1. Klasse insgesamt	325	123	38 %
2. Klasse insgesamt	285	167	59 %
3. Klasse insgesamt	706	528	75 %
Besatzung insgesamt	**908**	**696**	**77 %**
Frauen 1. Klasse	144	4	3 %
Frauen 2. Klasse	93	13	14 %
Frauen 3. Klasse	165	89	54 %
Frauen Besatzung	23	3	13 %
Frauen insgesamt	**425**	**109**	**26 %**
Kinder 1. Klasse	6	1	17 %
Kinder 2. Klasse	24	0	0 %
Kinder 3. Klasse	79	52	66 %
Kinder insgesamt	**109**	**53**	**49 %**
Männer 1. Klasse	175	118	68 %
Männer 2. Klasse	168	154	92 %
Männer 3. Klasse	462	387	84 %
Männer Besatzung	885	693	78 %
Männer insgesamt	**1690**	**1352**	**80 %**
Gesamt	**2224**	**1514**	**68 %**

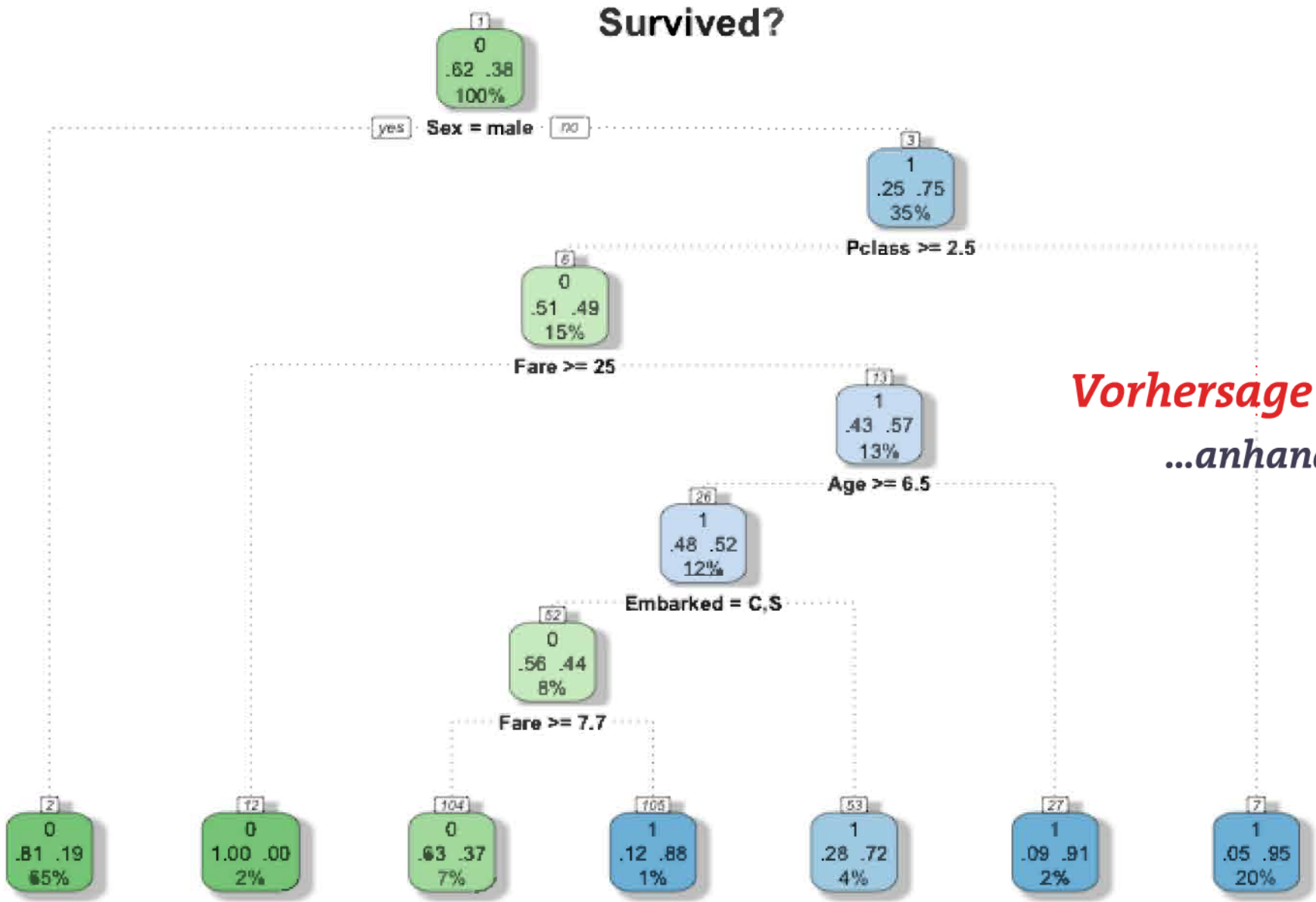
Frauen und Kinder zuerst

Erste Analyse der Passagierdaten der Titanic



Familien wurden bevorzugt

Erste Analyse der Passagierdaten der Titanic



Vorhersage der Überlebenschance...

...anhand eines Entscheidungsbaums

Via https://www.kaggle.com/c/titanic

Modellberechnung

Prognostiziere „Survived? (1/0)“

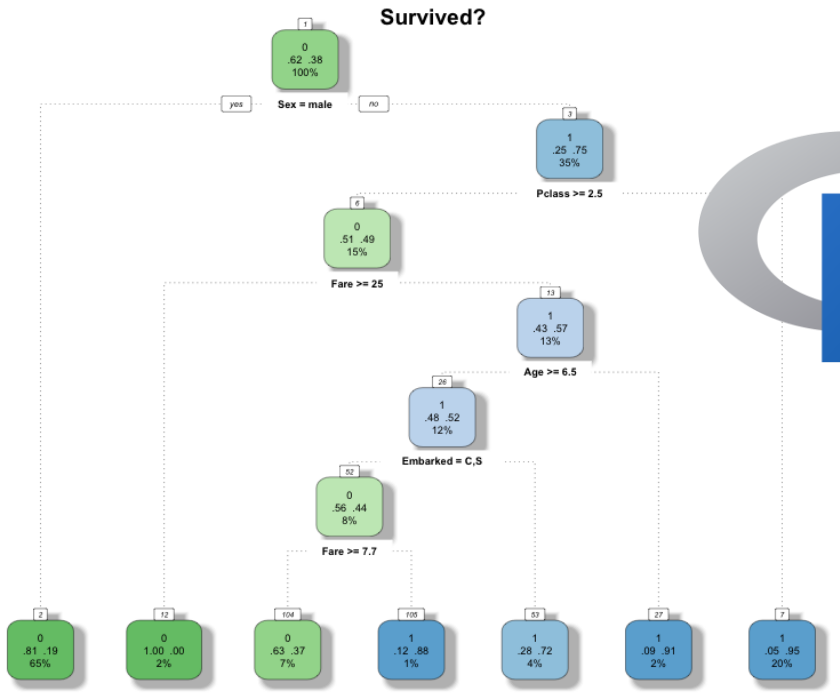
891 Passagierdaten

Training
80%
713

„trainieren“

Test
20%
178

„Vorhersage“



Entscheidungsbaum mit R

Entwicklung des Entscheidungsbaums

...mit R

```
data = read.csv("titanic.csv", stringsAsFactors = FALSE)

set.seed(4711)

split = sample.split(data$Survived, SplitRatio = 0.8)

train = subset(data, split == TRUE)
test = subset(data, split == FALSE)

## Entscheidungsbaum
formula = Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked

tree = rpart(formula, data=train, type="class")

## Prognose
pred = predict(tree, newdata=test, type="class")
```

Wie gut ist das Modell?

Prognose und Qualitätskontrolle

```
## Prognose
pred = predict(tree, newdata=test, type="class")

## Wie gut ist das Modell?
table(test$Survived, pred)
```

Survived?	0	1
0	106	4
1	34	34

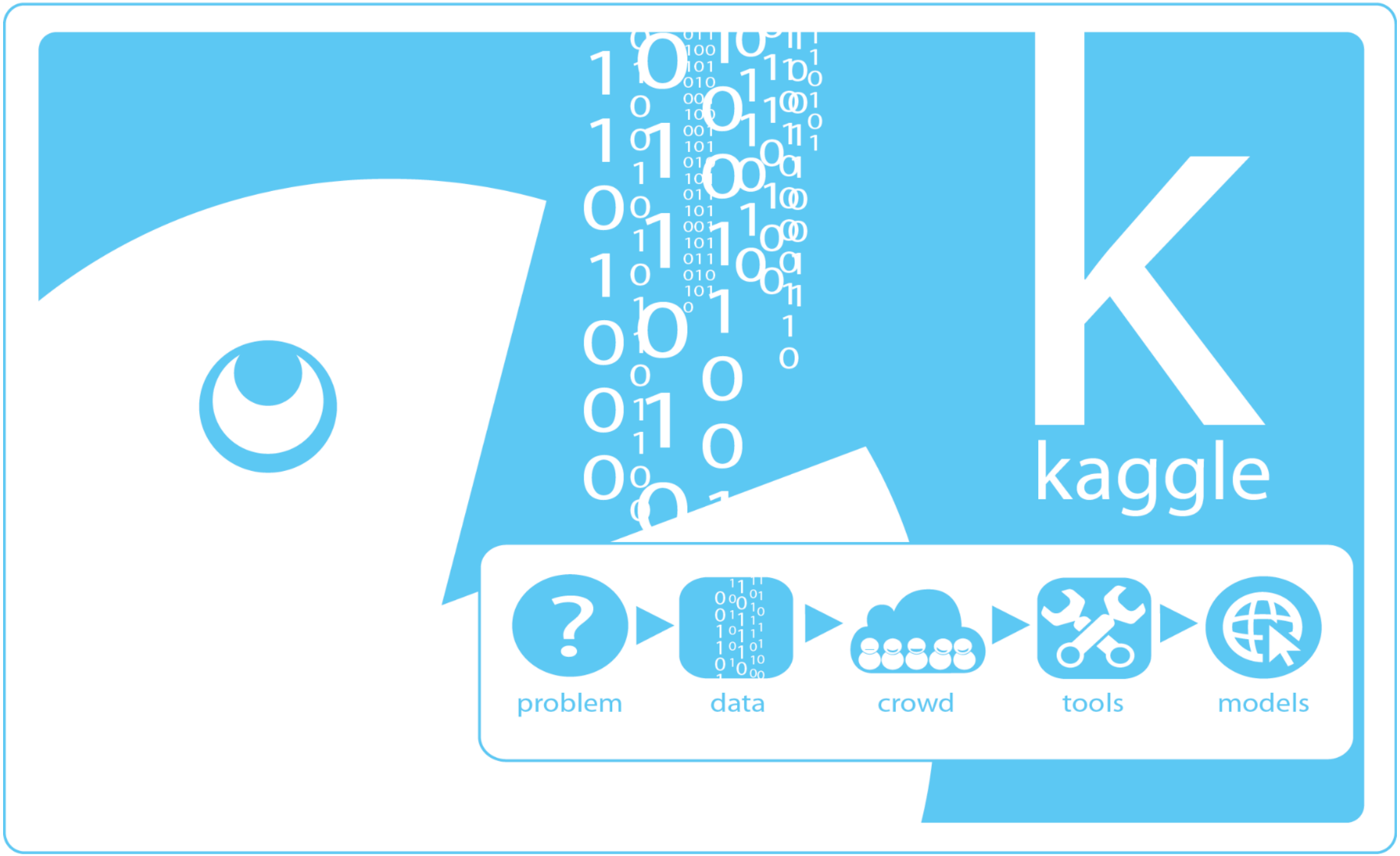
(106 + 34) / (106 + 34 + 34 + 4)

78,7%

Umsatzprognose im Einzelhandel (Autoregression)

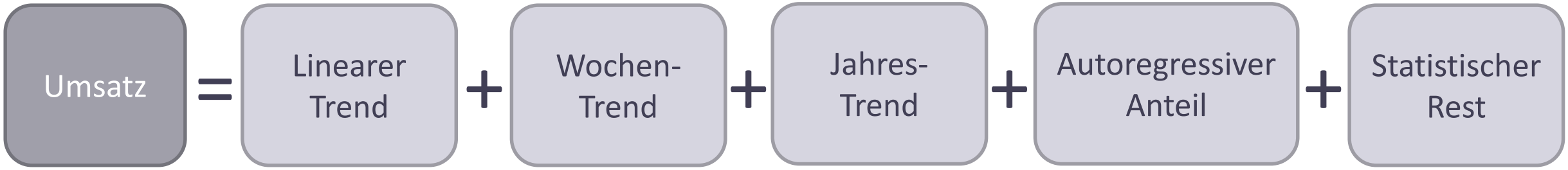
Foto: © Dirk Rossmann GmbH

Wikipedia

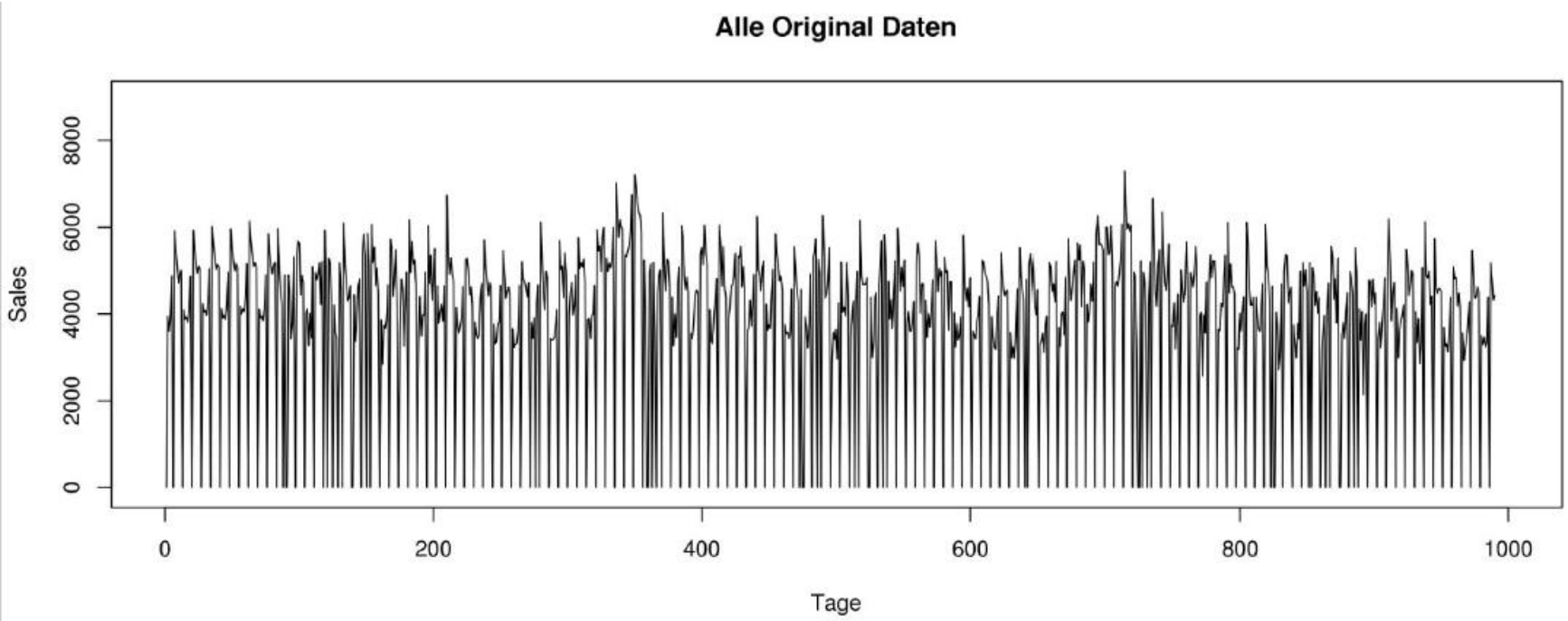

kaggle
problem
data
crowd
tools
models

Vorhersage mit Autoregression

Konzept



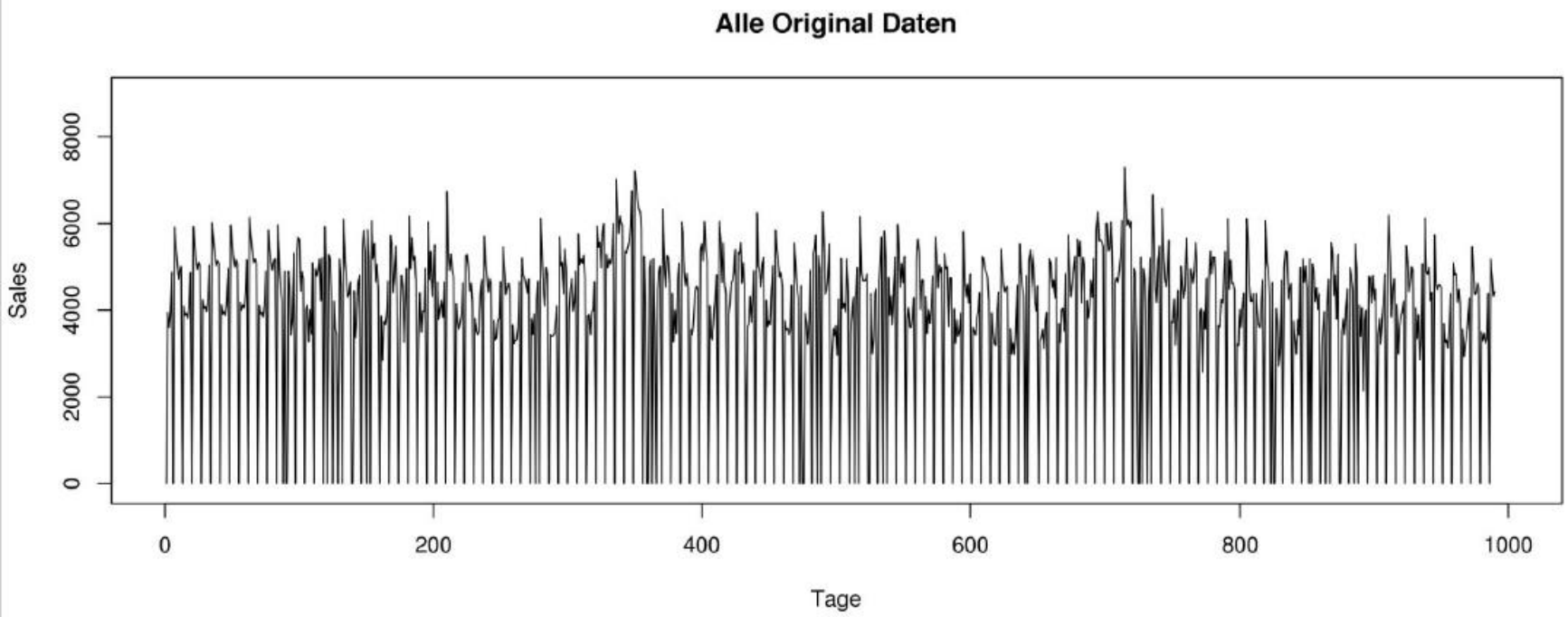
Rossmann-Daten



STORE	1115
STORETYPE	A B C D
ASSORTMENT	A B C
COMPETITION DISTANCE	
COMPETITION OPEN	
DATUM	**1.1.2013 - 31.7.2015**
OPEN	
PROMO	
PROMO2	
DAY OF WEEK	
HOLIDAY (State / School)	
CUSTOMERS	
SALES	

usw.

Rossmann-Daten

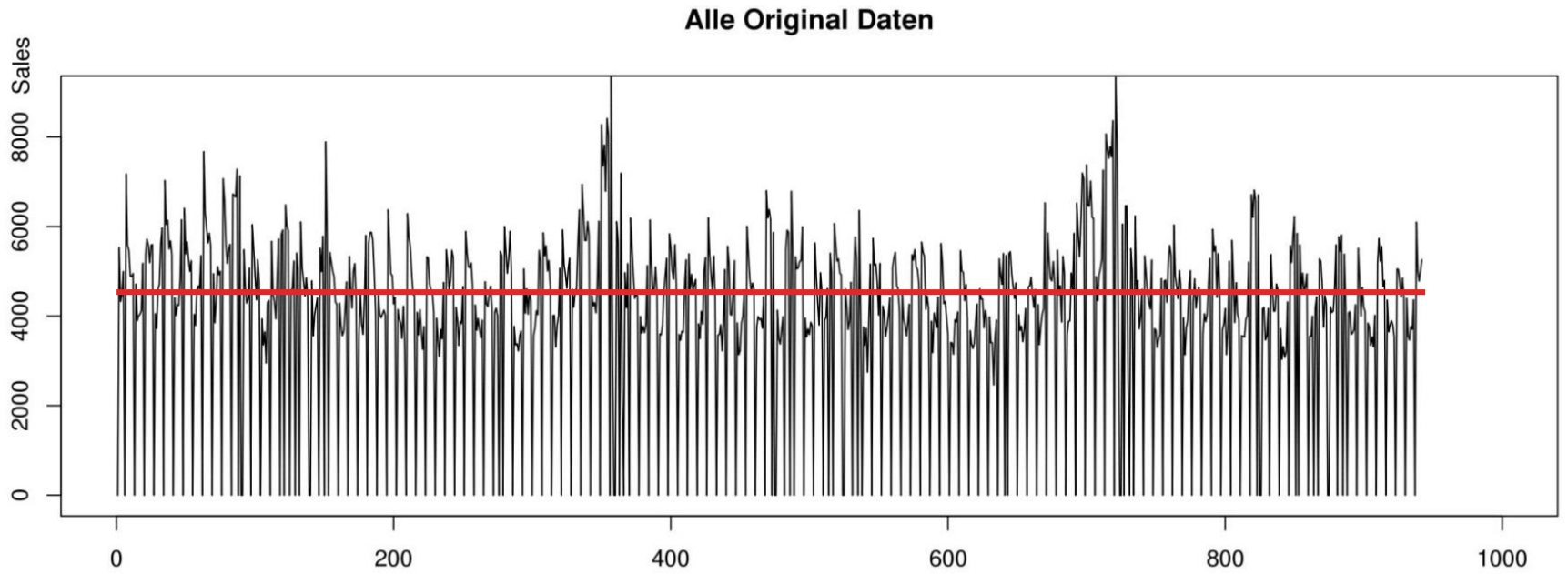


STORE	1115
STORETYPE	A B C D
ASSORTMENT	A B C
COMPETITION DISTANCE	
COMPETITION OPEN	
DATUM	**1.1.2013 - 31.7.2015**
OPEN	
PROMO	
PROMO2	
DAY OF WEEK	
HOLIDAY (State / School)	
CUSTOMERS	
SALES	

Umsatz = Linearer Trend + Wochen-Trend + Jahres-Trend + Autoregressiver Anteil + Statistischer Rest

Vorhersage mit Autoregression

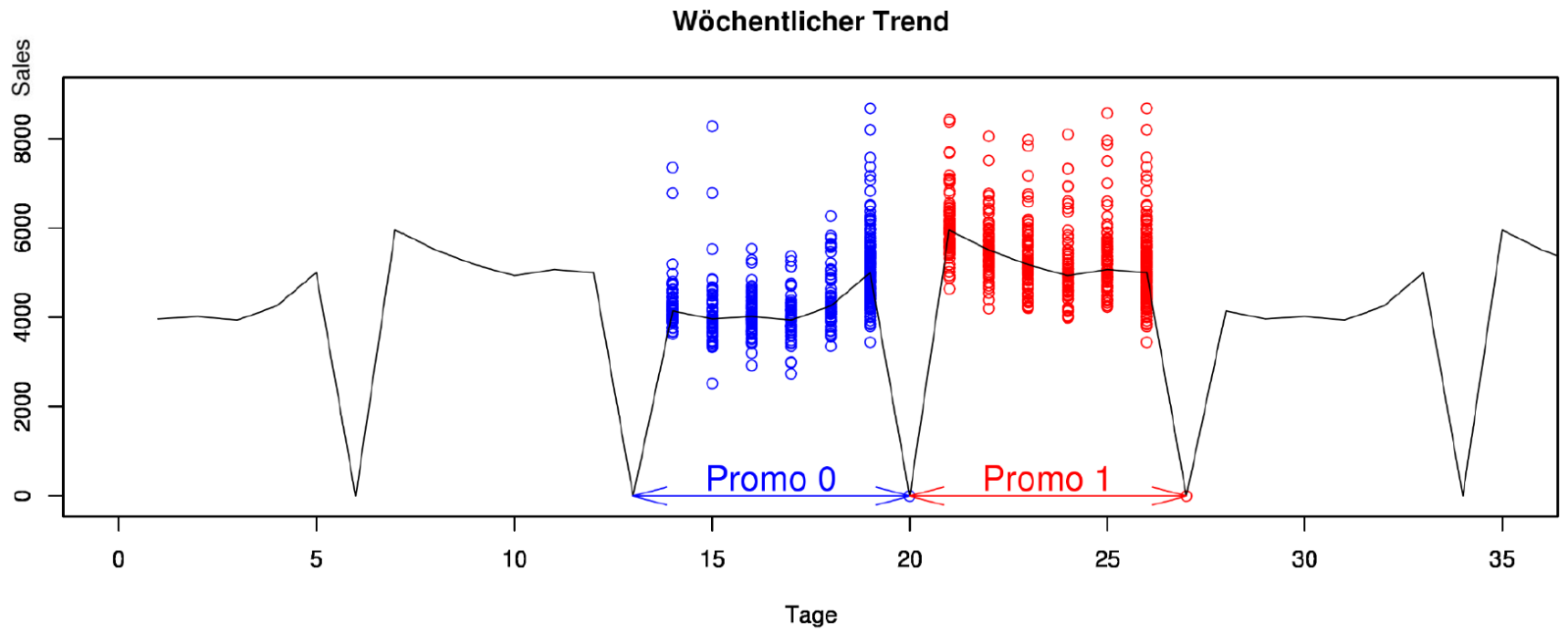
Linearer Trend



Umsatz = Linearer Trend + Wochen-Trend + Jahres-Trend + Autoregressiver Anteil + Statistischer Rest

Vorhersage mit Autoregression

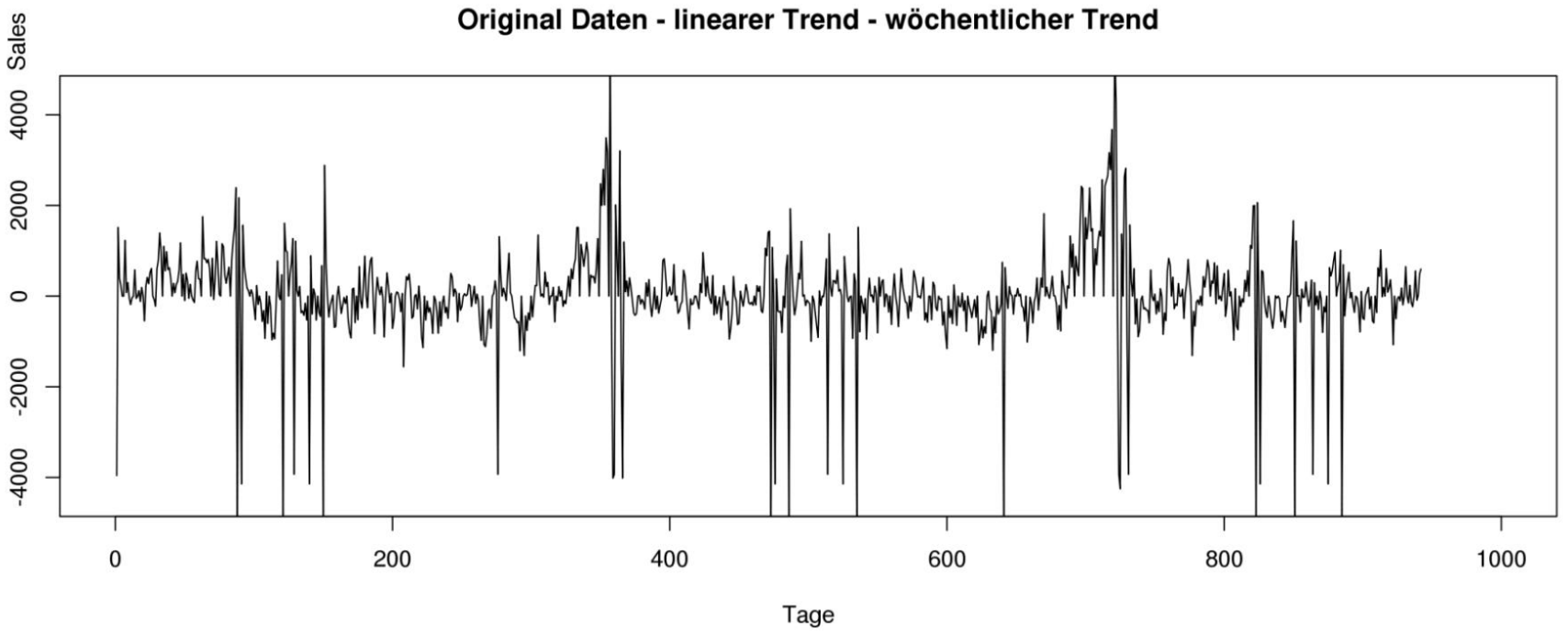
Wöchentlicher Trend



Umsatz = Linearer Trend + Wochen-Trend + Jahres-Trend + Autoregressiver Anteil + Statistischer Rest

Vorhersage mit Autoregression

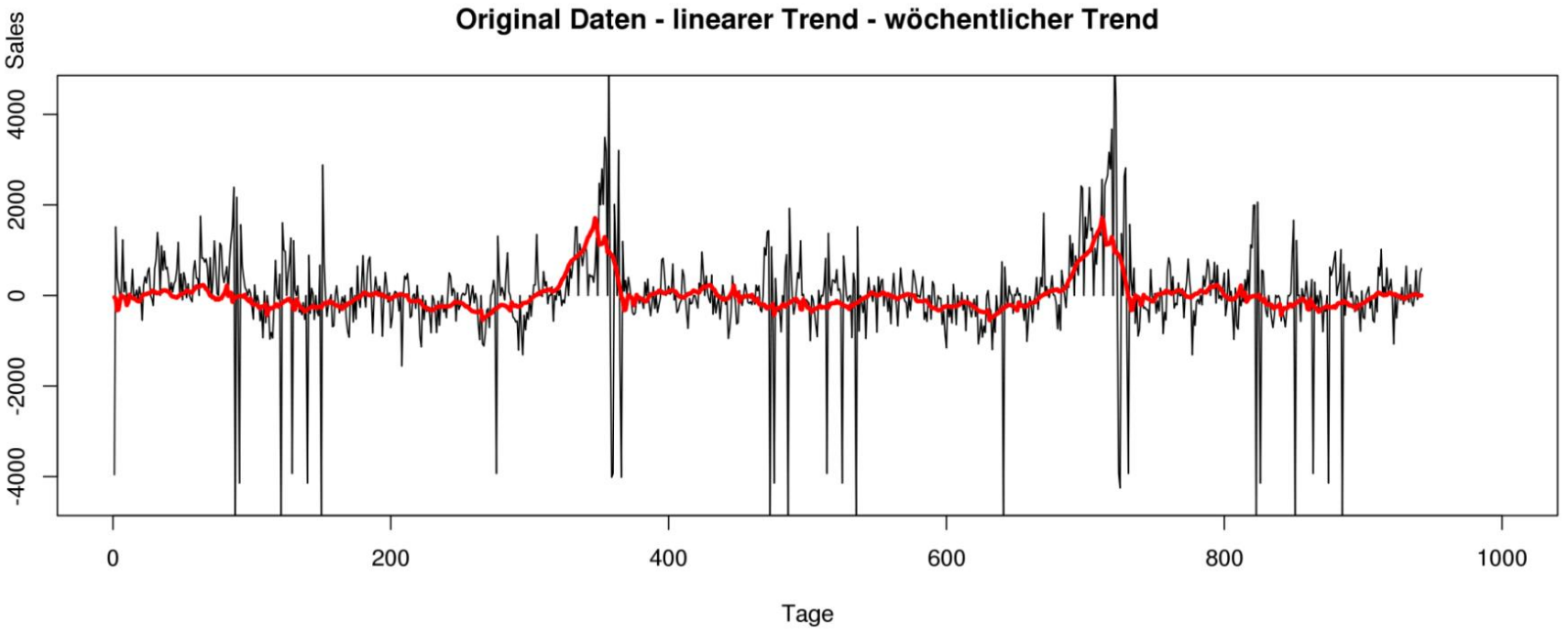
Wöchentlicher Trend



Umsatz = Linearer Trend + Wochen-Trend + Jahres-Trend + Autoregressiver Anteil + Statistischer Rest

Vorhersage mit Autoregression

Jährlicher Trend



Umsatz = Linearer Trend + Wochen-Trend + Jahres-Trend + Autoregressiver Anteil + Statistischer Rest

Vorhersage mit Autoregression

Autoregression

$ax_1+bx_2+cx_3+dx_4=$ y

Vorhersage mit Autoregression

Autoregression

Vorhersage mit Autoregression

Autoregression

$$X \cdot \vec{a} = \vec{p}$$

```
lag=3*4*7
p <- NULL

for (i in 1:sp)
  { xx[i,] <- data[i:(i+(lag-1))] }

for (i in (lag+1):dim(data)[1])
  { p[i-lag] <- data[i] }
```

```
sp <- dim(data)[1]-lag
xx <- matrix(0,sp,lag)
```

Vorhersage mit Autoregression

Autoregression

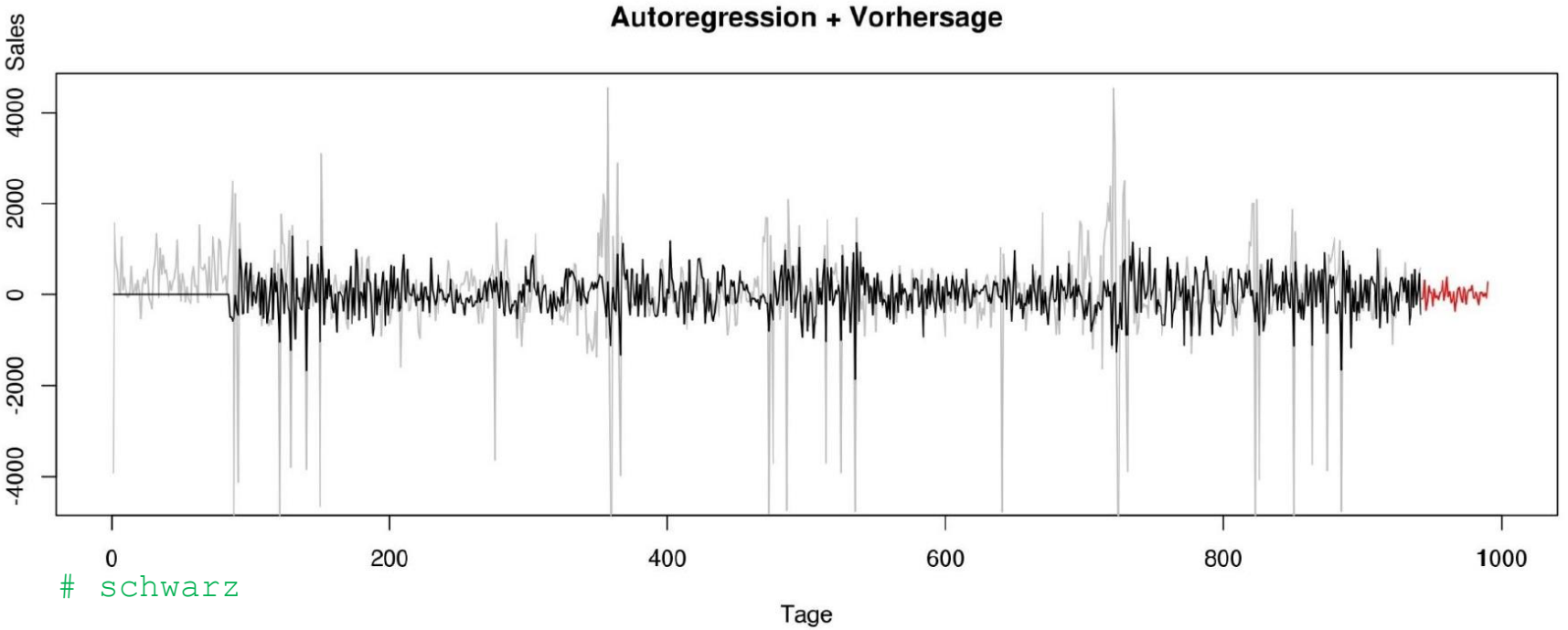
$$X \cdot \vec{a} = \vec{p}$$

$$\vec{a} = (X^T \cdot X)^{-1} \cdot X^T \cdot \vec{p}$$

```
a=solve(t(xx)%*%xx)%*%t(xx)%*%p
```

Vorhersage mit Autoregression

Vorhersage

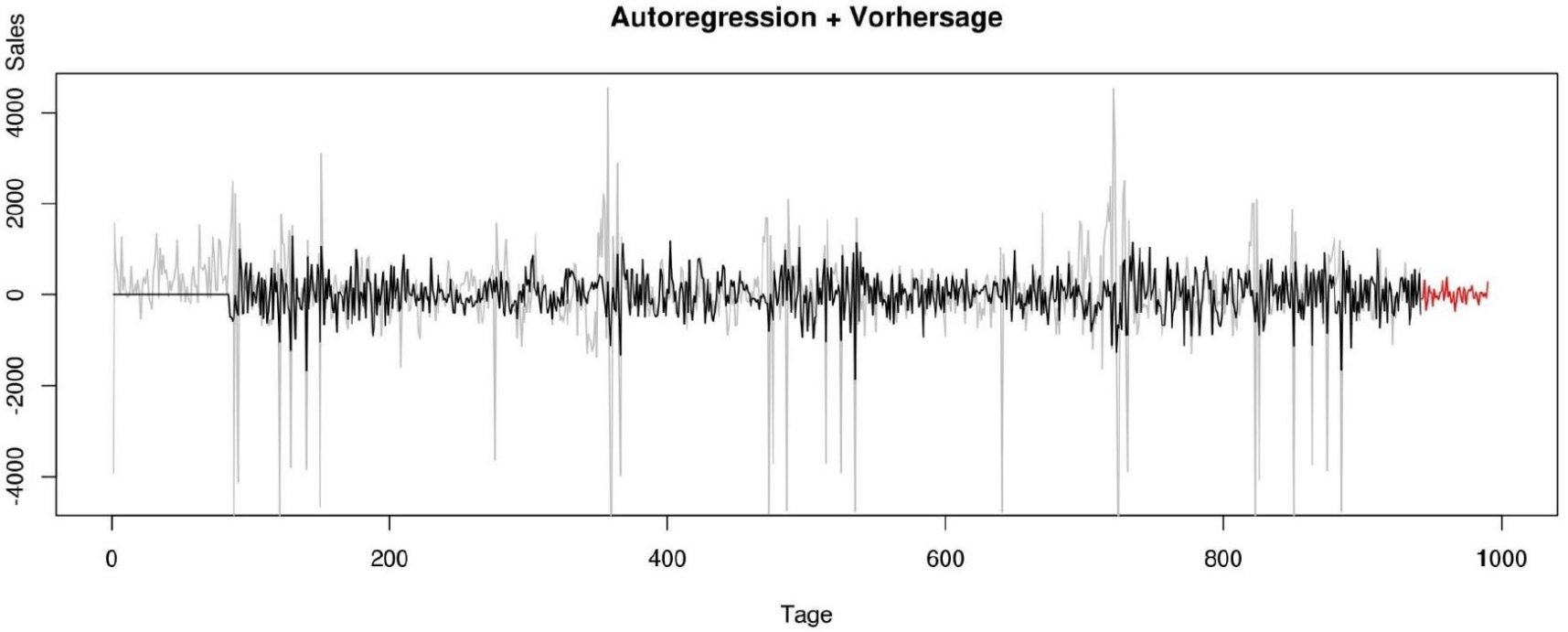



```
# schwarz
for (i in (lag+1):dim(train)[1])            { autoreg[i] <- xx[i-lag,]%*%a }


#rot
for (i in 1:dim(test)[1]+1)
{ pointer <- dim(data)[1]+i-1
        autoreg[pointer] <- t( autoreg[(pointer-lag):(pointer-1)])%*%a }
```

Vorhersage mit Autoregression

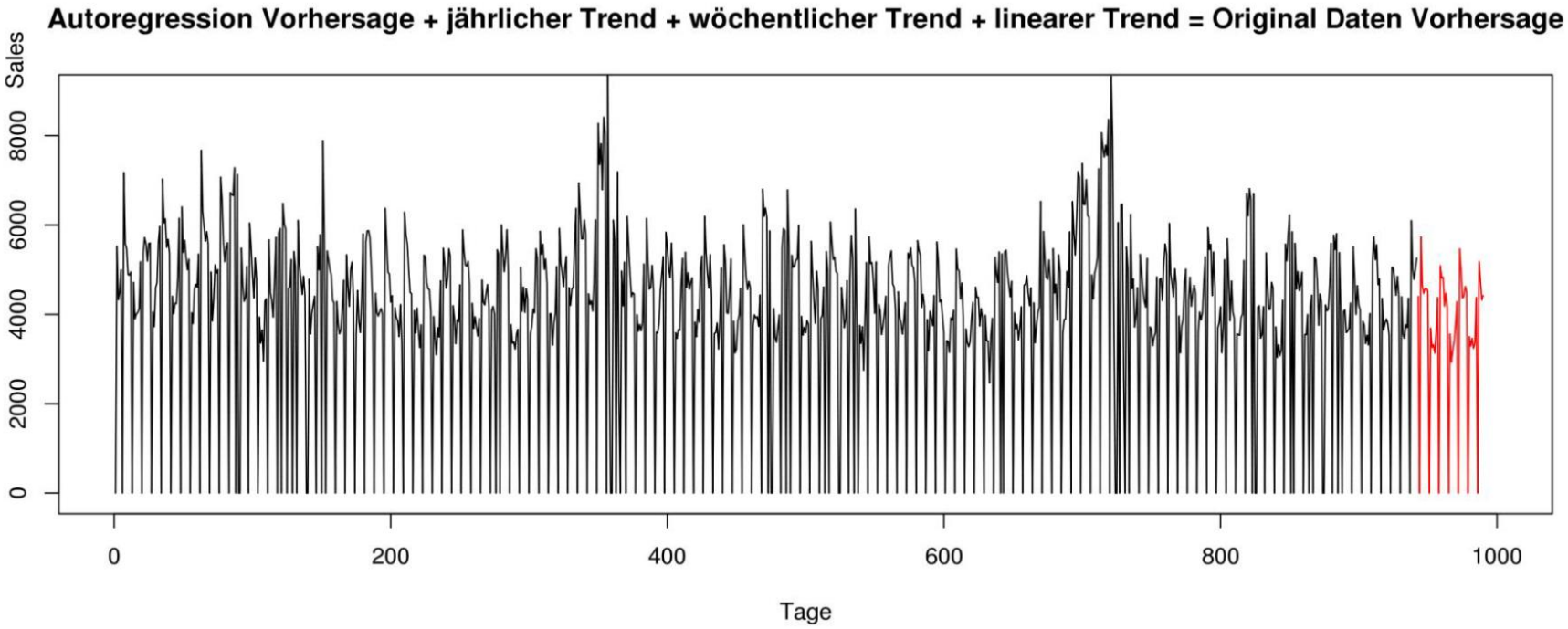
Vorhersage



Umsatz = Linearer Trend + Wochen-Trend + Jahres-Trend + Autoregressiver Anteil + Statistischer Rest

Vorhersage mit Autoregression

Vorhersage



Umsatz = Linearer Trend + Wochen-Trend + Jahres-Trend + Autoregressiver Anteil + Statistischer Rest

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}$$

kaggle *Highscore*

Autoregression (Zeitreihenanalyse):
nur Umsatz- / Besucherzahlen

RMS 0,13

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}$$

kaggle **Highscore**

Nur Nullen	RMS 1,00
Nur mittlere Tageswerte	RMS 0,24
Autoregression (Zeitreihenanalyse): nur Umsatz- / Besucherzahlen	RMS 0,13

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}$$

kaggle *Highscore*

Nur Nullen	RMS 1,00
Nur mittlere Tageswerte	RMS 0,24
Autoregression (Zeitreihenanalyse): nur Umsatz- / Besucherzahlen	RMS 0,13
Random Forest (Machine Learning): alle Variablen (Umsatz, Filialen-Typ, Entfernung zur Konkurrenz, …)	RMS 0,12

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}$$

kaggle **Highscore**

Nur Nullen	RMS 1,00
Nur mittlere Tageswerte	RMS 0,24
Autoregression (Zeitreihenanalyse): nur Umsatz- / Besucherzahlen	RMS 0,13
Random Forest (Machine Learning): alle Variablen (Umsatz, Filialen-Typ, Entfernung zur Konkurrenz, …)	RMS 0,12
Gewinner (Gert, NL): Random Forest: alle Variablen + externe Daten (Wetter, Großereignisse, …)	RMS 0,10

Fazit & Ausblick

Autoregression

Einfache Regression

Klassifikation

Einblick in Data Science

Wie gehe ich an solche Probleme ran?

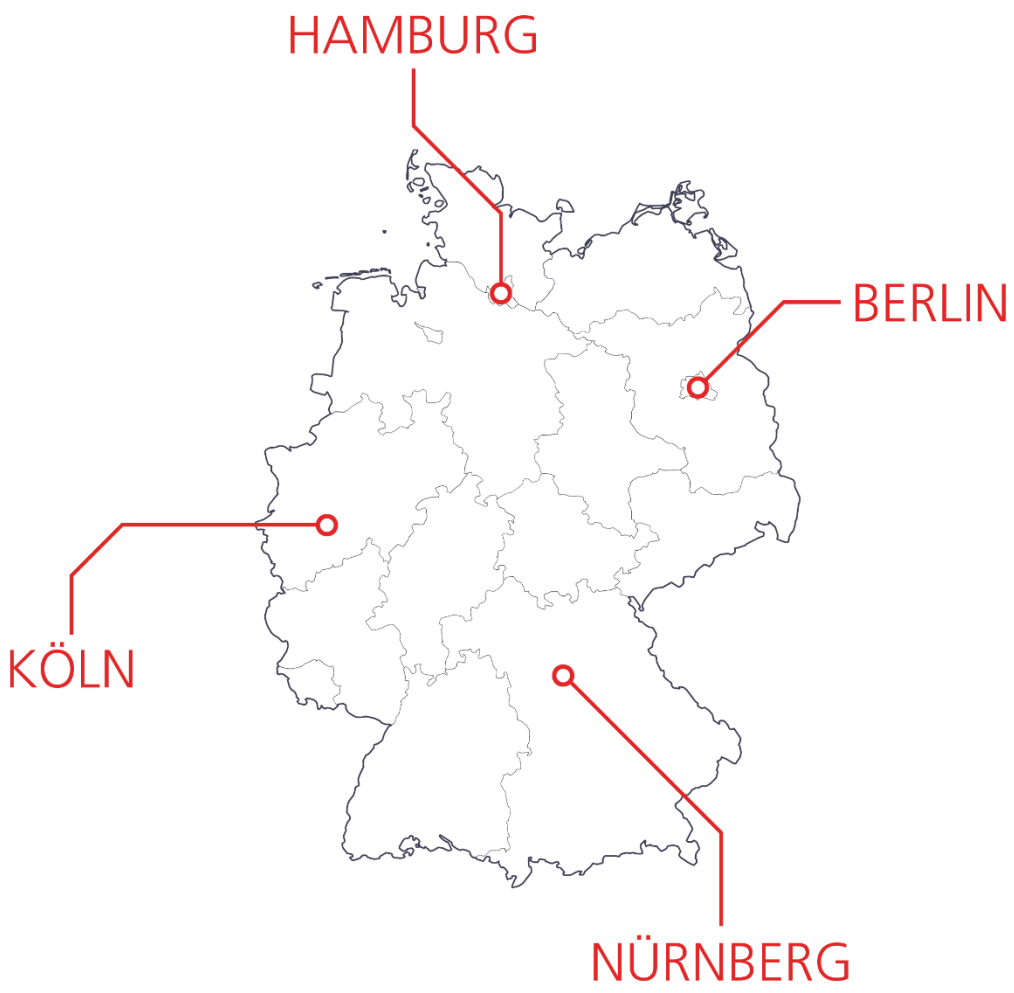
Pareto-Prinzip!

Einfache Mittel zeigen große Wirkung

Dr. Klaus Lippert / Marco Nätlitz
Data Science

areto consulting gmbh
Data Warehouse
Business Intelligence

areto
CONSULTING. IT WORKS.



Visit us at our Branches:

KÖLN

Schanzenstraße 6-20
51063 Köln
Telefon: +49 221 66 95 75-0
Telefax: +49 221 66 95 75-99
E-Mail: info@areto.de

HAMBURG

Rothenbaumchaussee 27
20148 Hamburg
Telefon: +49 40 22 86 53 20
Telefax: +49 40 22 86 53 23
E-Mail: info@areto.de

NÜRNBERG

Neumeyerstraße 24
90411 Nürnberg
Telefon: +49 9 11 14 88 66 77
Telefax: +49 9 11 14 88 66 79
E-Mail: info@areto.de

BERLIN

Weichselstraße 34a
10247 Berlin
Telefon: +49 30 54 90 87 51
Telefax: +49 30 54 90 87 52
E-Mail: info@areto.de

Henning Schwass
Leiter Marketing
henning.schwass@areto.de

areto consulting gmbh
Data Warehouse
Business Intelligence

areto
CONSULTING. IT WORKS.