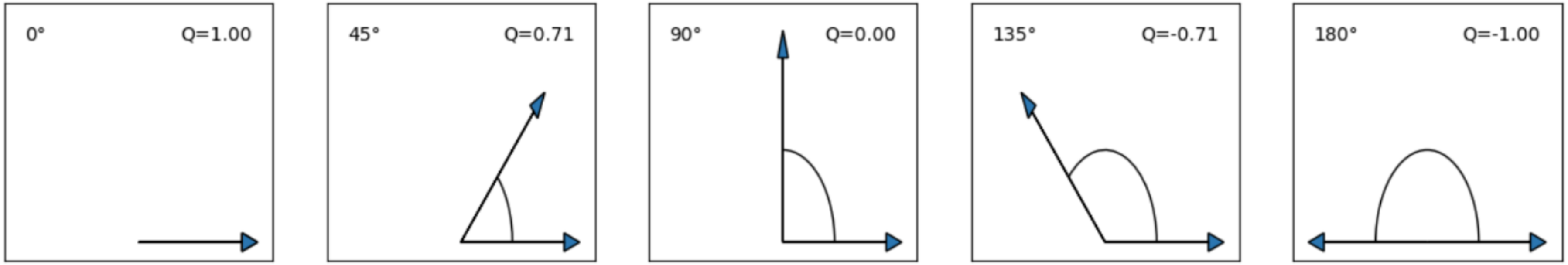
Wiederholung

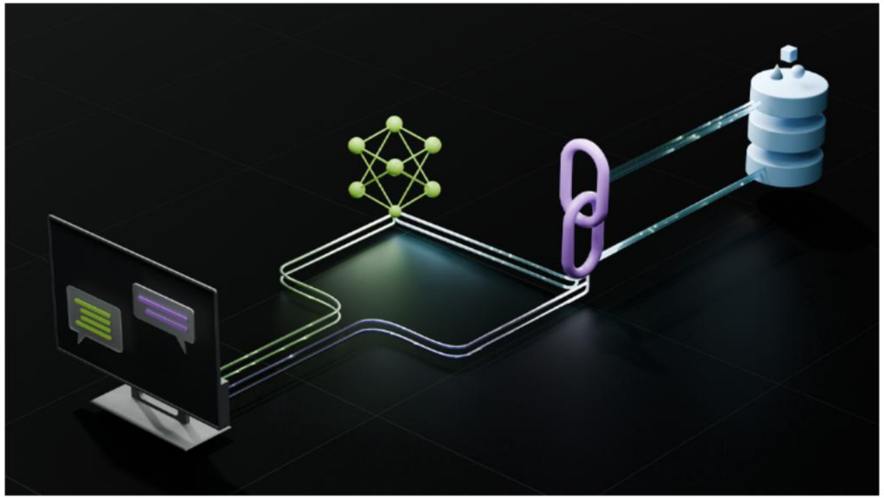


Erinnerung: Für den Winkel $W(u, v)$ zwischen zwei Vektoren u, v mit positiver Länge gilt die Formel

$$\cos(W(u,v)) = \underbrace{\frac{< u, v >}{||u|| \cdot ||v||}}_{\text{Cosine Similarity } Q} =< u/||u||, v/||v|| >$$

wobei $< u, v >= \sum_{i=1}^{d} u_i v_i$ das Skalarprodukt und $||u|| = \sqrt{< u, u >}$ die Länge des Vektors beschreibt.

Die Cosine Similarity wird verwendet z.B. zur Abfrage "ähnlicher" Vektoren in einer Vektordatenbank:



https://developer.nvidia.com/blog/rag-101-demystifying-retrieval-augmented-generation-pipelines/

Hier z.B. für eine RAG (*Retrieval-Augmented Generation*)-Pipeline, um einem Large Language Model passenden Kontext für eine Anfrage zu liefern.

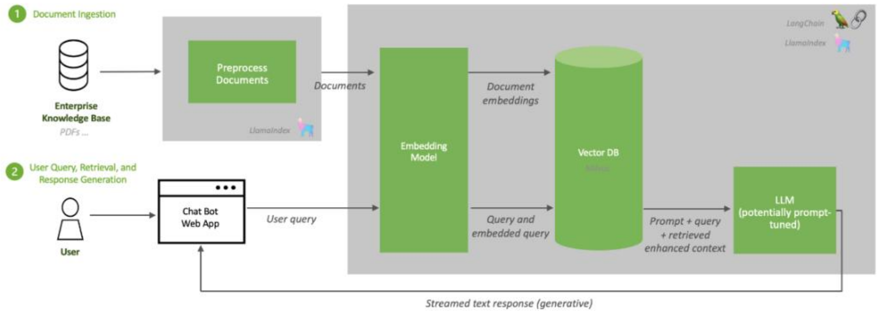


Figure 1. Overview of RAG pipeline components: ingest and query flows

Hier werden (in einem sog. Embedding-Schritt) viele Dokumente ("Hintergrundwissen") in Vektoren übersetzt und in einer Vektordatenbank abgelegt.

Eine Anfrage des Benutzers wird ebenfalls in einen Vektor übersetzt und dann dazu ähnliche (Dokument-)Vektoren in der Vektordatenbank gesucht (z.B. mittels Cosine Similarity).

Die Anfrage wird dann zusammen mit einigen ähnlichen Dokumenten dem LLM zur Beantwortung der Frage übergeben.

Übergang zur Korrelation durch **Zentrierung** der Vektoren (Mittelwert komponentenweise abziehen):

Es seien Datenvektoren $x = (x_1, \ldots, x_n)$ und $y = (y_1, \ldots, y_n)$ gleicher Länge $n \geq 2$ gegeben.

- Die (empirische) **Kovarianz** ist definiert via

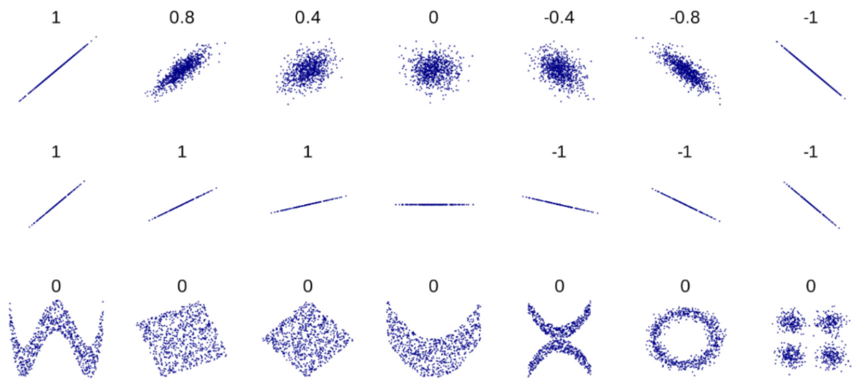
$$\operatorname{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

- Die (empirische) **Pearson-Korrelation** ist definiert via

$$\operatorname{cor}(x, y) = \frac{\operatorname{cov}(x, y)}{\operatorname{sd}(x)\operatorname{sd}(y)}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

Die Korrelation ist ein gutes Maß für den **linearen** Zusammenhang:



https://en.wikipedia.org/wiki/Correlation

Insbesondere impliziert i.A. Korrelation=0 nicht die Unabhängigkeit (Abwesenheit von Kausalität) zwischen zwei Datenquellen.

von: Jepson Taylor, Dataiku
(generiert mit MidJourney)

Least Squares is like a tailor adjusting a suit to fit a customer perfectly. Imagine a customer trying on a suit that doesn't fit quite right - it's too loose in some places and too tight in others. The tailor measures the differences between the suit and the customer's body, then makes adjustments to minimize these discrepancies. By finding the smallest sum of these squared differences, the tailor ensures the suit fits as closely as possible to the customer's measurements. Least Squares algorithm works similarly, adjusting the parameters of a mathematical model to minimize the sum of the squared differences between the predicted values and the actual data, resulting in the best possible fit.

1. Lineare Regression
2. Gesundheitsdaten
3. Gradient Descent

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

1.1 Beispiel einer Einfachen Linearen Regression

In [281]:

```
# Der sklearn diabetes-Datensatz
from sklearn.datasets import load_diabetes
X = load_diabetes(as_frame=True)

# Daten extrahieren
X,y = X["data"], X["target"]
df = pd.concat([X,y],axis=1)

# .T zum transponieren (spiegeln)
# (hier verwendet, um alle Spalten zu sehen)
df.head().T
```

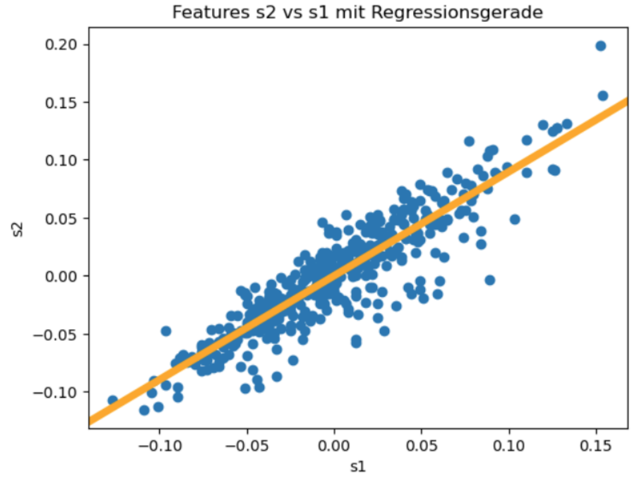
Out[281]:

	0	1	2	3	4
age	0.038076	-0.001882	0.085299	-0.089063	0.005383
sex	0.050680	-0.044642	0.050680	-0.044642	-0.044642
bmi	0.061696	-0.051474	0.044451	-0.011595	-0.036385
bp	0.021872	-0.026328	-0.005670	-0.036656	0.021872
s1	-0.044223	-0.008449	-0.045599	0.012191	0.003935
s2	-0.034821	-0.019163	-0.034194	0.024991	0.015596
s3	-0.043401	0.074412	-0.032356	-0.036038	0.008142
s4	-0.002592	-0.039493	-0.002592	0.034309	-0.002592
s5	0.019907	-0.068332	0.002861	0.022688	-0.031988
s6	-0.017646	-0.092204	-0.025930	-0.009362	-0.046641
target	151.000000	75.000000	141.000000	206.000000	135.000000

In [296]:

```
# Alternative: df.plot.scatter("s1","s2");
plt.scatter ( df["s1"], df["s2"] )
plt.xlabel("s1")
plt.ylabel("s2")

# Berechne die Regressionsgerade
# (Alternative zu sklearn LinearRegression())
slope, intercept = np.polyfit(df["s1"],df["s2"],
                              deg=1)
plt.axline ( xy1=(0,intercept), slope=slope,
             color="orange", linewidth=5 )
plt.title("Features s2 vs s1 mit Regressionsgerade");
```



Lineare Regression:

- **"einfach"**, wenn ein linearer Zusammenhang nur eines Features x_1 zu einer Zielgröße y gesucht wird;
- **"multiple"**, wenn ein linearer Zusammenhang mehrerer Features $x_1, \ldots, x_k$ zu einer Zielgröße y gesucht wird.

Modell:

$$f_{a_0,\ldots,a_k}(x_1, \ldots, x_k) = a_0 + a_1 \cdot x_1 + \ldots + a_k \cdot x_k$$

Methode der kleinsten Quadrate: Die Koeffizienten werden so bestimmt, dass $MSE(y, f_{\ldots}(\ldots))$ minimal wird, wobei

$$MSE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \,.$$

Hierbei kürzen wir ab: $f_{a_0,\ldots,a_k}(x_{i1}, \ldots, x_{ik}) = \hat{y}_i$.

In [10]:
```
from sklearn.metrics import mean_squared_error

y     = (3,4,5)
y_hat = (3.1,4.2,5.3)

mse = lambda y,z: ( (np.array(y)-np.array(z))**2 ) \
                    .mean()

mean_squared_error(y,y_hat), mse (y,y_hat)
```

Out[10]:
```
(0.04666666666666666, 0.04666666666666666)
```

Lineare Regression:

- **"einfach"**, wenn ein linearer Zusammenhang nur eines Features x_1 zu einer Zielgröße y gesucht wird;
- **"multiple"**, wenn ein linearer Zusammenhang mehrerer Features $x_1, \ldots, x_k$ zu einer Zielgröße y gesucht wird.

Modell:

$$f_{a_0,\ldots,a_k}(x_1, \ldots, x_k) = a_0 + a_1 \cdot x_1 + \ldots + a_k \cdot x_k$$

Methode der kleinsten Quadrate: Die Koeffizienten werden so bestimmt, dass $MSE(y, f_{\ldots}(\ldots))$ minimal wird, wobei

$$MSE(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \; .$$

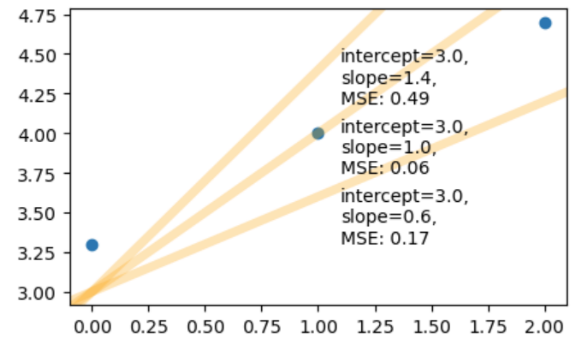
Hierbei kürzen wir ab: $f_{a_0,\ldots,a_k}(x_{i1}, \ldots, x_{ik}) = \hat{y}_i$.

In [351]:

```
# Beispiel für eine einfache lineare Regression
x = np.array([0.0,1.0,2.0])
y = np.array([3.3,4.0,4.7])
plt.scatter ( x, y );
plt.gcf().set_size_inches( (5,3) )

# Kandidaten-Modelle (fixiere intercept=3.0)
intercept=3.0
f = lambda x, intercept, slope: intercept+slope*x

# Variiere den Slope Parameter
for slope in [0.6,1.0,1.4]:
    plt.axline ( xy1=(0,intercept), slope=slope,
        color="orange", alpha=0.3, linewidth=5 )
    plt.text(s="intercept={0},\nslope={1},\nMSE: {2:.2f}"
            .format(intercept, slope,
                    mse(y,f(x,intercept,slope))),
             x=1.1, y=f(1.1,intercept,slope),
             color="black", va="top" )
```



1.2 Herleitung der Koeffizienten

Es seien x, y Vektoren der Länge n.

Ziel: Minimiere $MSE(y, f_{a_0,a_1}(x))$ als Funktion von a_0, a_1.

Ansatz:

$$\frac{\partial}{\partial a_0} MSE(y, f_{a_0,a_1}(x)) = 0 \,,$$
$$\frac{\partial}{\partial a_1} MSE(y, f_{a_0,a_1}(x)) = 0 \,.$$

Ausgeschrieben:

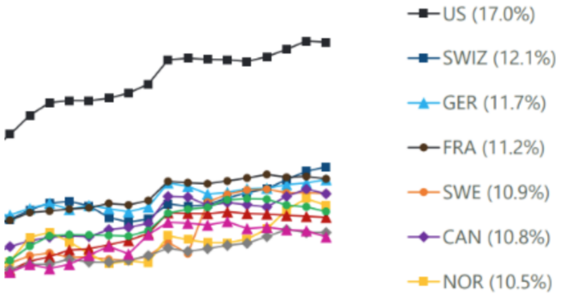
$$\begin{aligned} MSE(y, f_{a_0,a_1}(x)) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - f_{a_0,a_1}(x_i))^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - a_0 - a_1 \cdot x_i)^2 \end{aligned}$$

Ergebnis für eine einfache Lineare Regression:

$$\begin{aligned} a_0 &= \overline{y} - a_1 \cdot \overline{x} \,, \\ a_1 &= \mathrm{cor}(x, y)\frac{\mathrm{sd(y)}}{\mathrm{sd(x)}} \,. \end{aligned}$$

Hierbei ist der *Mittelwert*, die *Standardabweichung* und die *(Pearson-)Korrelation* definiert via

$$\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \,, \\ \mathrm{sd(x)} &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \,, \\ \mathrm{cor}(x, y) &= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\mathrm{sd(x)sd(y)}} \,. \end{aligned}$$



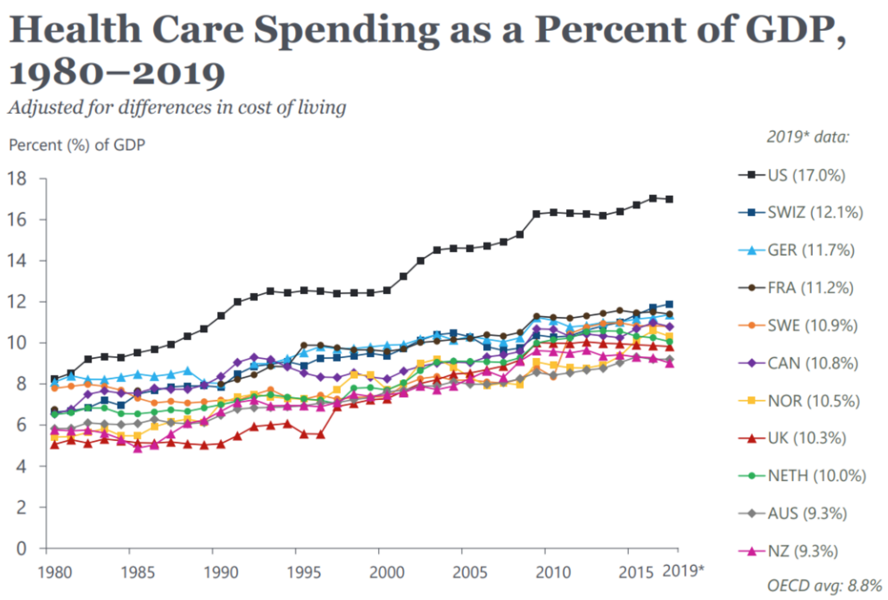
1. Lineare Regression

2. Gesundheitsdaten

3. Gradient Descent

2. Gesundheitsdaten

- Wie hoch sind die Gesundheitsausgaben pro Person?
- Wie gut ist das deutsche Gesundheitssystem im internationalen Vergleich?
- Wie können wir "Erfolg im Gesundheitswesen" quantifizieren?

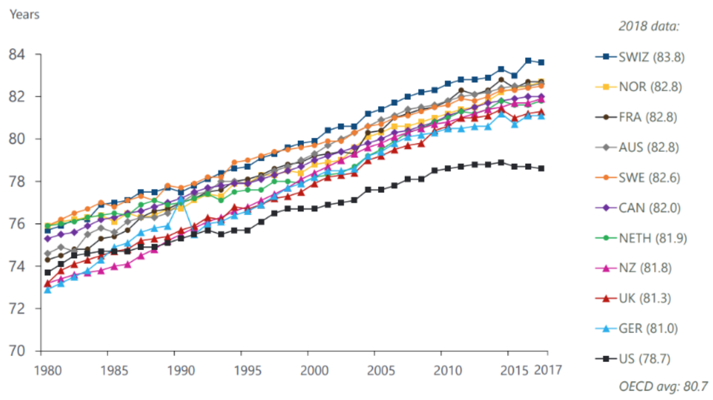


Multinational Comparisons of Health Systems Data, 2020. (The Commonwealth Fund, Datenquelle: OECD)
https://www.commonwealthfund.org/publications/other-publication/2021/feb/multinational-comparisons-health-systems-data-2020

pro Kopf: ca. 11 000$ USA und 6 000$ DE (Durchschnittswert über alle Altersgruppen)

Frage: Was ist eine gute Metrik, um den Nutzen der Ausgaben zu bewerten?

Life Expectancy at Birth, 1980–2018

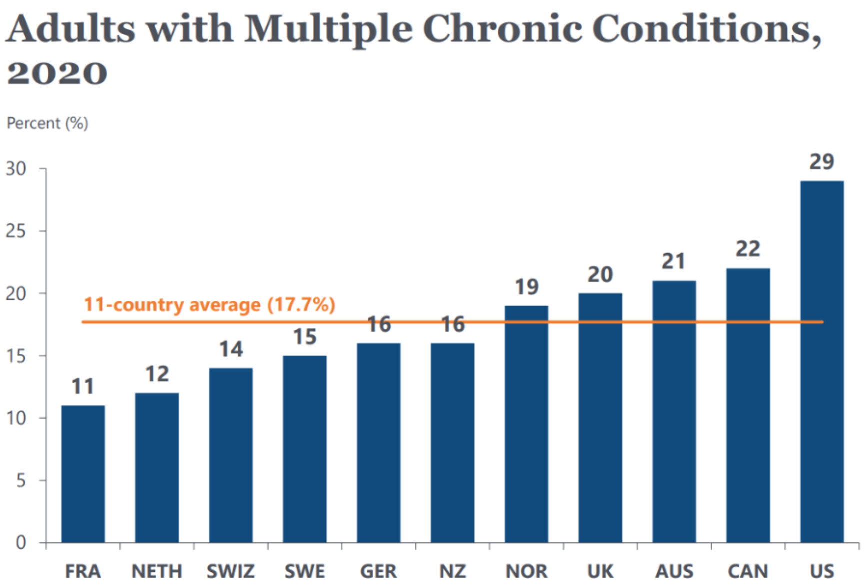


Multinational Comparisons of Health Systems Data, 2020. (The Commonwealth Fund, Datenquelle: OECD)
https://www.commonwealthfund.org/publications/other-publication/2021/feb/multinational-comparisons-health-systems-data-2020 , https://stats.oecd.org/

Zum Vergleich: Südafrika ca. 57 Jahre (Vergleichswert von 2013, gemittelt über m/w/d)

Frage: Warum sind die US-Werte (im Mittel) so schlecht, obwohl die Ausgaben dort (im Mittel) am höchsten sind?

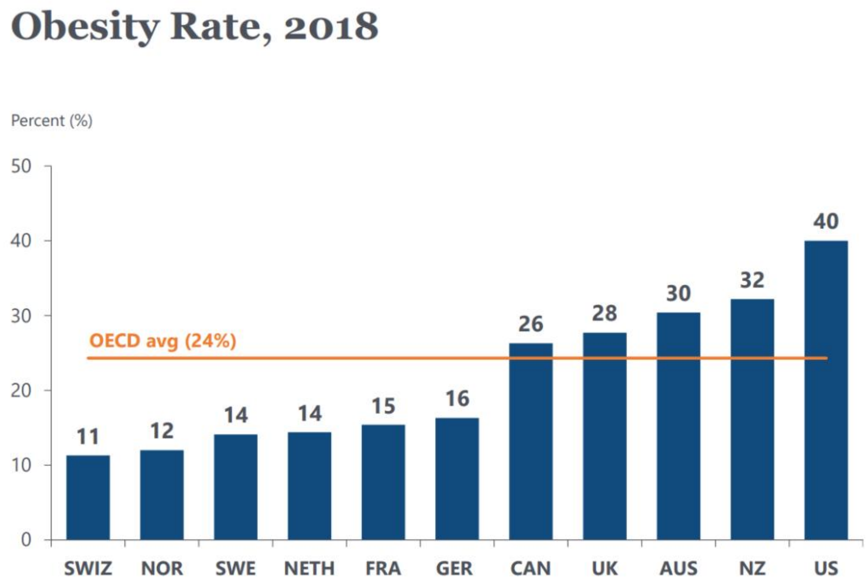
Hypothese: Ein Grund für die hohen Gesundheitsausgaben könnte eine hohe Spreizung "zwischen Arm und Reich" in den USA sein. Der Mittelwert würde durch wenig Reiche mit hohen Ausgaben verzerrt. (Oder es gibt andere US-spezifische Gründe, z.B. Lifestyle.)



Multinational Comparisons of Health Systems Data, 2020. (The Commonwealth Fund, Datenquelle: OECD)
https://www.commonwealthfund.org/publications/other-publication/2021/feb/multinational-comparisons-health-systems-data-2020 , https://stats.oecd.org/

Frage: Bedeutet das, dass das US-Gesundheitssystem schlechter ist?

Nicht unbedingt, denn a) es kann im Hintergrund wirkende Kausalzusammenhänge geben (allgemein: "Confounding Factors", hier z.B. Lifestyle),
b) es könnte sogar ein Hinweis auf ein "gutes" Gesundheitssystem sein (mehr frühzeitig diagnostizierte Krankheiten).

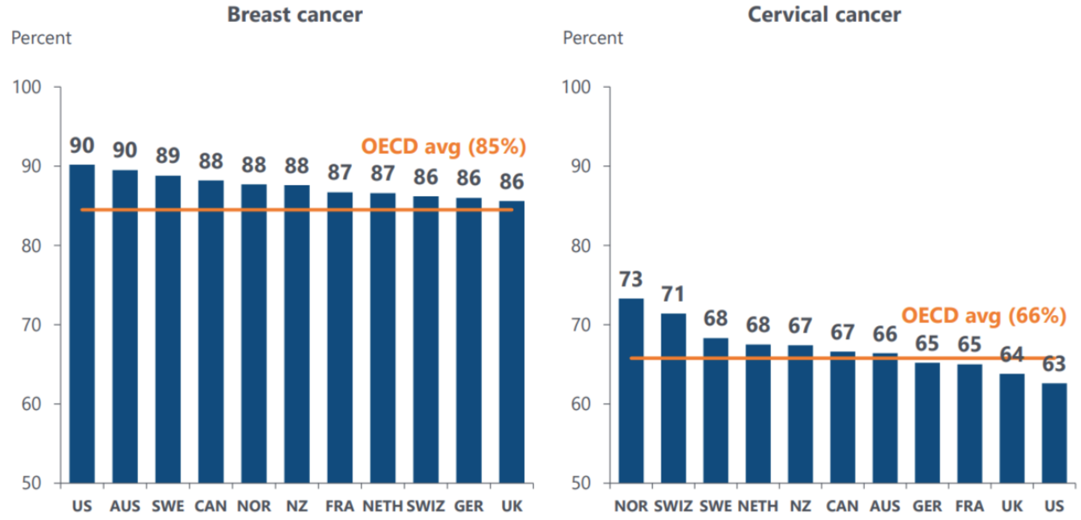


Multinational Comparisons of Health Systems Data, 2020. (The Commonwealth Fund, Datenquelle: OECD)
https://www.commonwealthfund.org/publications/other-publication/2021/feb/multinational-comparisons-health-systems-data-2020 , https://stats.oecd.org/

Eine mögliche Ursache (?) für chronische Krankheiten (möglicher *Confounding Factor*)

Frage: Gibt es neben der Lebenserwartung andere Metriken für "Erfolg im Gesundheitswesen"?

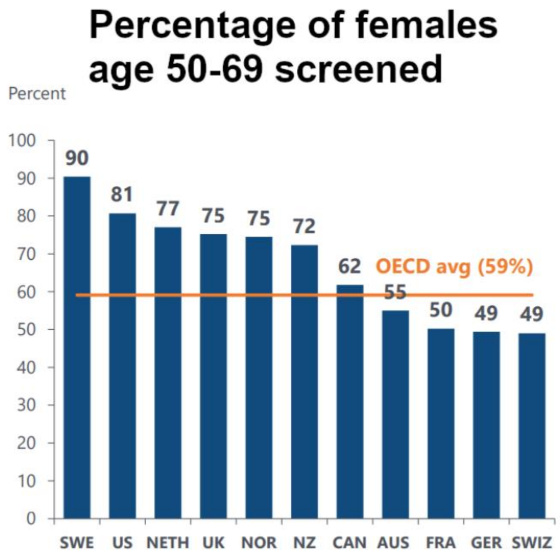
Breast and Cervical Cancer Five-Year Net Survival Rates, 2010–2014

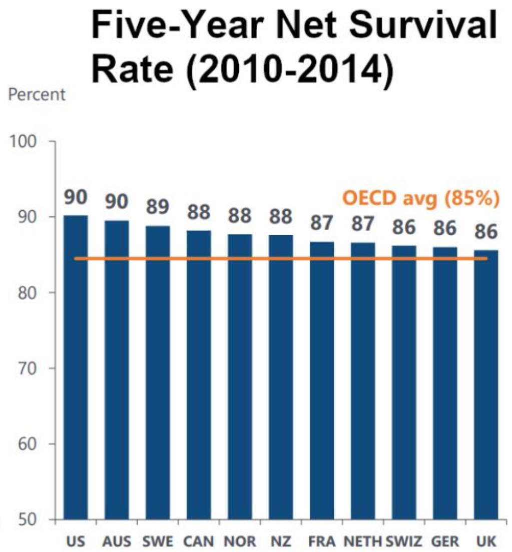


Multinational Comparisons of Health Systems Data, 2020. (The Commonwealth Fund, Datenquelle: OECD)
https://www.commonwealthfund.org/publications/other-publication/2021/feb/multinational-comparisons-health-systems-data-2020 , https://stats.oecd.org/

Frage: Ist die "Five-Year survival rate" eine gute Metrik?

Hier am Beispiel **Brustkrebs**:

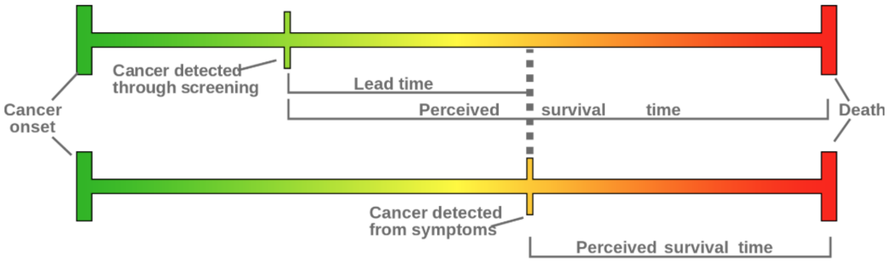




Eigene Darstellung, basierend auf: Multinational Comparisons of Health Systems Data, 2020.
https://www.commonwealthfund.org/publications/other-publication/2021/feb/multinational-comparisons-health-systems-data-2020 , https://stats.oecd.org/

Unklar, ob eine höhere "Five-Year survival rate" nicht einfach auf eine frühere Diagnose zurückzuführen ist ("Lead Time Bias").

Lead Time Bias:



https://en.wikipedia.org/wiki/Lead_time_bias

The Five Year Survival Rate

The five year survival rate fulfils only one of the criteria of a good index, namely simplicity. Apart from this it has many disadvantages. These spring mainly from the false opinions held about human cancer.

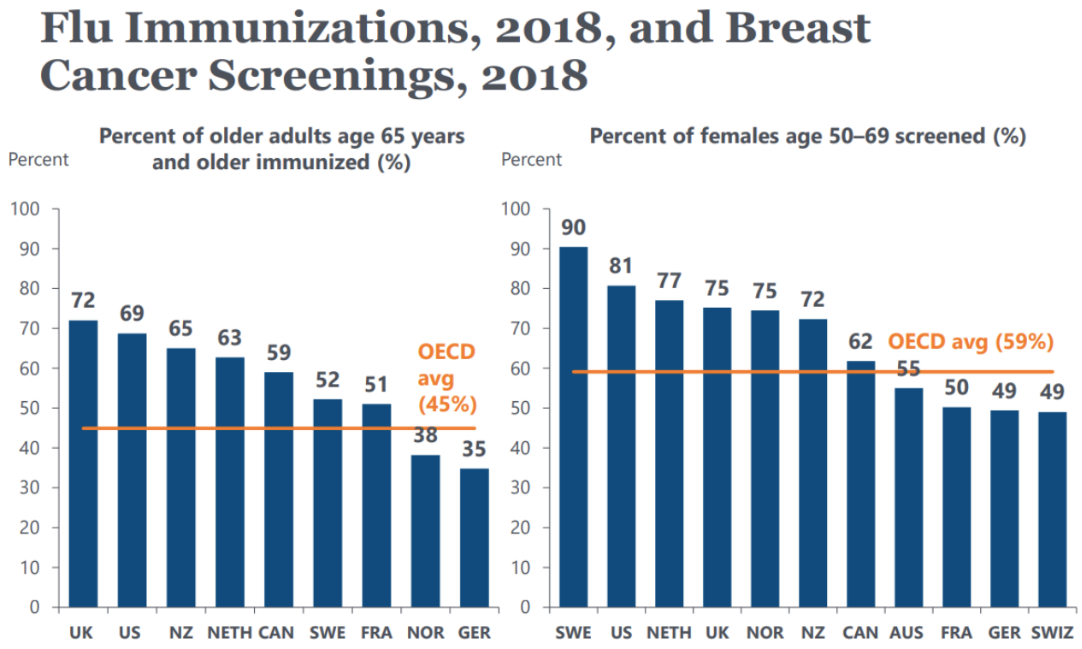
1. The five year survival rate is not a true index of the value of treatment. It implies that in the absence of treatment the survival rate at five years would have been nil.

A five year survival rate after treatment should be contrasted with the survival of similar cancers without treatment. We are interested in the *increase* of survival times due to treatment and at present our idea of the survival of treatable but untreated cancer is very vague.

2. By only recording the number alive at five years it fails to indicate the benefits to those who died within five years, but whose life was prolonged appreciably but not quite for five years. Nor does it indicate the benefit to those who would have lived for five years without treatment.

...

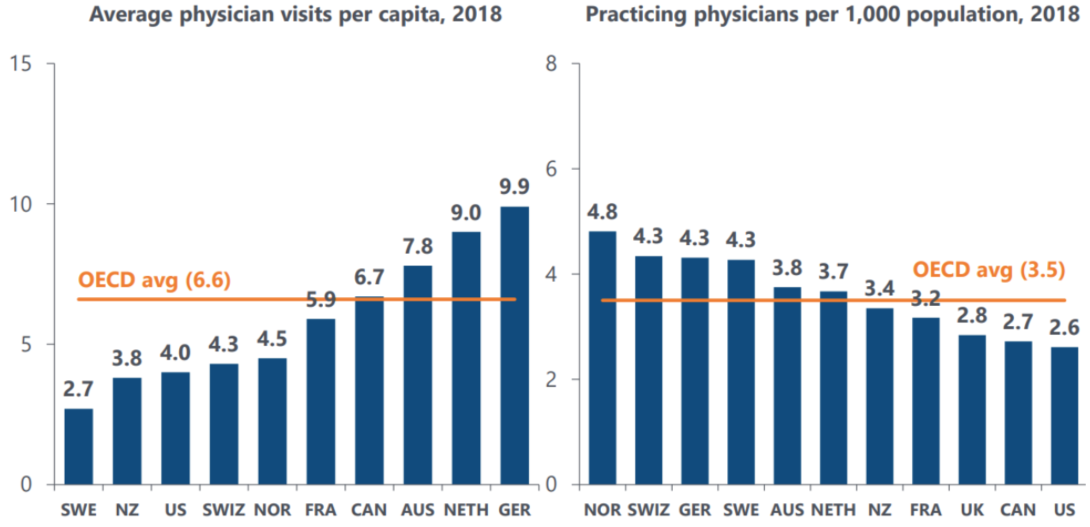
(1956) Criticism of Present Methods of Analysis, Acta Radiologica, 45:sup132, 19-25, DOI: 10.3109/00016925609172299
https://www.tandfonline.com/doi/pdf/10.3109/00016925609172299



Multinational Comparisons of Health Systems Data, 2020. (The Commonwealth Fund, Datenquelle: OECD)
https://www.commonwealthfund.org/publications/other-publication/2021/feb/multinational-comparisons-health-systems-data-2020 , https://stats.oecd.org/

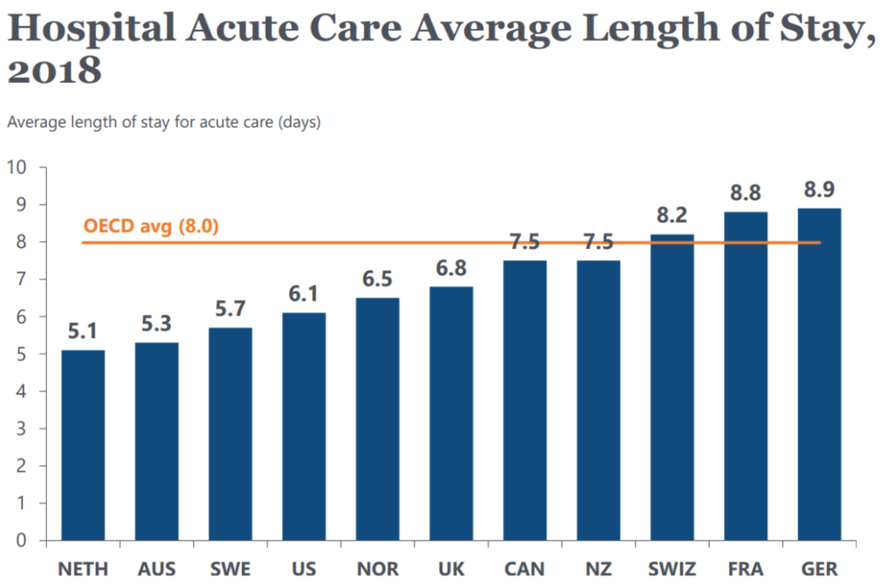
Deutsche Besonderheiten: Niedrige (Grippe-)Impfraten...

Physician Visits, 2018 and Physician Supply, 2018



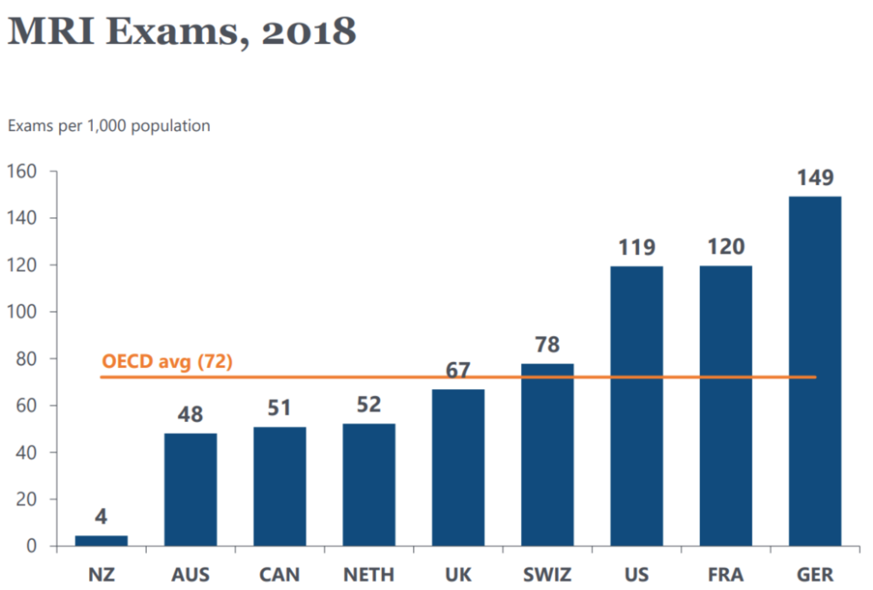
Multinational Comparisons of Health Systems Data, 2020. (The Commonwealth Fund, Datenquelle: OECD)
https://www.commonwealthfund.org/publications/other-publication/2021/feb/multinational-comparisons-health-systems-data-2020 , https://stats.oecd.org/

Deutsche Besonderheiten: ... viele Arztbesuche ...



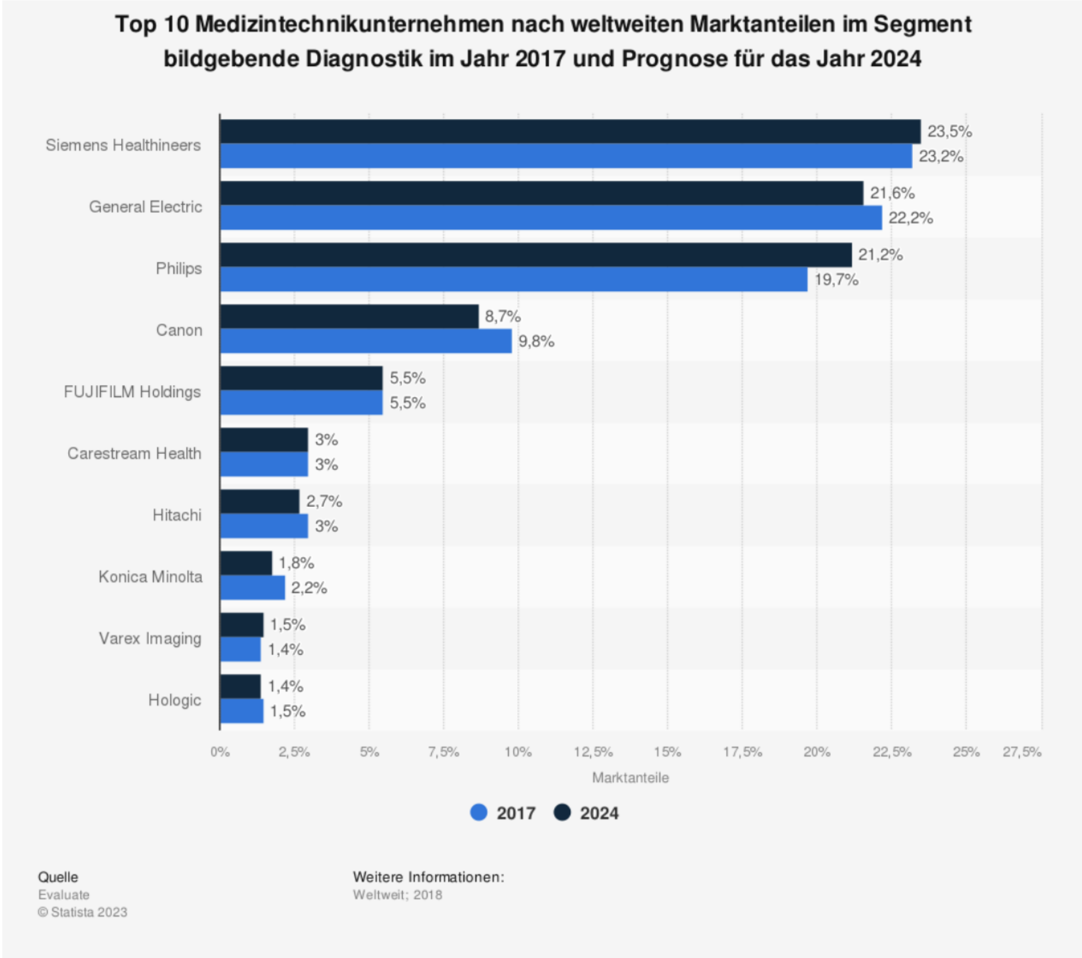
Multinational Comparisons of Health Systems Data, 2020. (The Commonwealth Fund, Datenquelle: OECD)
https://www.commonwealthfund.org/publications/other-publication/2021/feb/multinational-comparisons-health-systems-data-2020 , https://stats.oecd.org/

Deutsche Besonderheiten: ... lange Krankenhausaufenthalte ...



Multinational Comparisons of Health Systems Data, 2020. (The Commonwealth Fund, Datenquelle: OECD)
https://www.commonwealthfund.org/publications/other-publication/2021/feb/multinational-comparisons-health-systems-data-2020 , https://stats.oecd.org/

Deutsche Besonderheiten: ... und viele MRT-Aufnahmen!



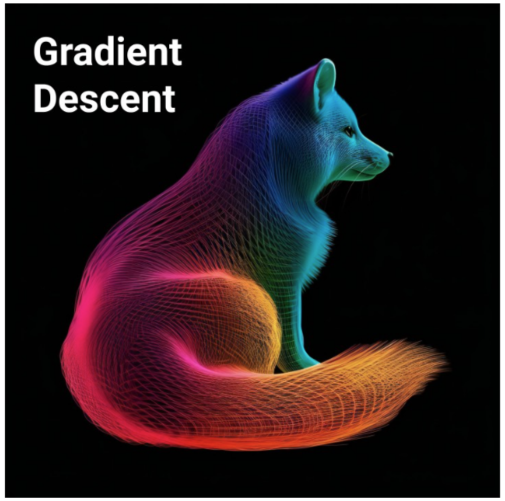
Evaluate. (26. September, 2018). Top 10 Medizintechnikunternehmen nach weltweiten Marktanteilen im Segment bildgebende Diagnostik im Jahr 2017 und Prognose für das Jahr 2024 [Graph]. In Statista. Zugriff am 22. Mai 2023, von https://de.statista.com/statistik/daten/studie/332494/umfrage/fuehrende-medizintechnikunternehmen-nach-weltweiten-marktanteilen-im-segment-bildgebende-diagnostik/

Fazit:

- Der Erfolg einer Behandlung (oder allgemeiner von Gesundheitsausgaben) ist schwer zu messen.
- Populäre Metriken wie "Five-Years Survival Rate" sind kritisch zu sehen;
  können aber sinnvoll sein in kontrollierten Vergleichen (Behandlung A vs Behandlung B).
- Die Bewertung ist schwierig v.a. wegen der vielfältigen Einflussfaktoren, die ein Behandlungsergebnis beeinflussen.
- Ein Einflussfaktor ist der Zeitpunkt der Diagnose. Eine frühere Diagnose kann scheinbar zu einer Erhöhung der "Five-Years Survival Rate" führen (*Lead Time Bias*).

Konsequenzen:

- Früherkennungsverfahren sollten gut durchdacht sein (insb. nur dann anwenden, wenn es gute Behandlungsoptionen gibt); siehe z.B. "Wilson and Jungner"-Kriterien der WHO
- Zur Bewertung von Behandlungen sind **kontrollierte klinische Studien** notwendig ("Phase I/Phase II/Phase III"). Insbesondere notwendig:
  - Ein klar definiertes Ziel ("Endpunkt");
  - eine ähnlich zusammengesetzte Vergleichsgruppe, die die Behandlung nicht bekommt ("Kontrollgruppe");
  - eine zufällige Zuweisung auf Kontroll- und Experimentalgruppe ("Verblindung").



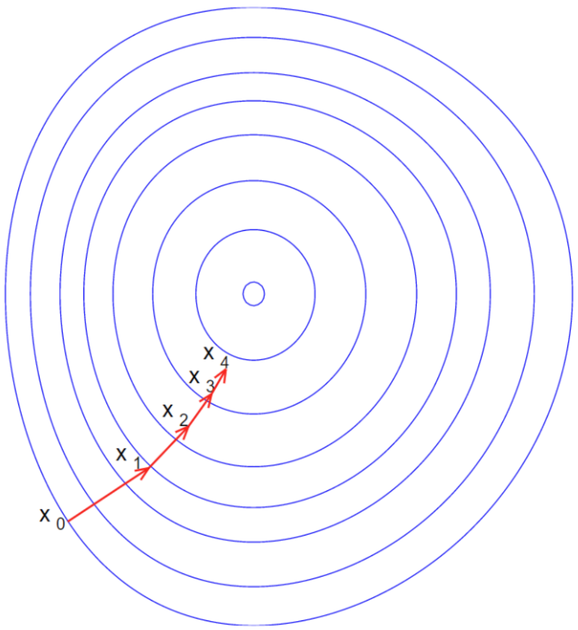
von: Jepson Taylor, Dataiku
(generiert mit MidJourney)

Gradient Descent is like a hiker trying to find the lowest point in a valley by taking steps downhill. The hiker checks which direction is steepest, takes a step in that direction, and repeats the process until they reach the bottom, where the terrain is flat and there's no more descending to do.

1. Lineare Regression

2. Gesundheitsdaten

3. Gradient Descent



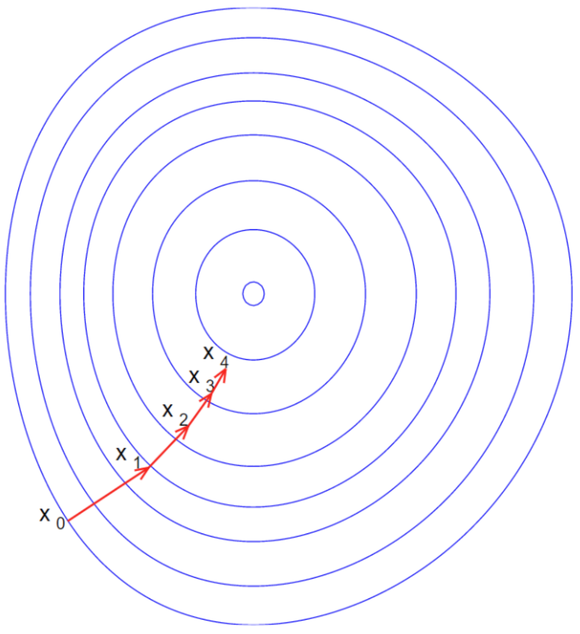
https://upload.wikimedia.org/wikipedia/commons/f/ff/Gradient_descent.svg

Lineare Regression $f_{a_0,a_1}(x) = a_0 + a_1 \cdot x$.

Bisher: Minimiere $MSE(y, f_{a_0,a_1}(x))$ als Funktion von a_0, a_1, indem die partielle Ableitungen $\frac{\partial}{\partial a_i} MSE(y, f_{a_0,a_1}(x))$ Null gesetzt werden.

Alternativ: Iterative Lösung mittels **Gradient Descent**:

1. Fixiere eine Learning Rate α (z.B. 0.1).
2. Wähle Parameter (a_0, a_1) zufällig.
3. Berechne den Fehler $MSE(y, f_{a_0,a_1}(x))$.
   Breche ab, wenn der Fehler (oder die Fehleränderung) kleiner als eine vorgegebene Schranke ist.
4. Berechne die Richtung des stärksten Fehleranstiegs:
   $$v_0 = \partial/\partial a_0 \, MSE(y, f_{a_0,a_1}(x))$$
   und
   $$v_1 = \partial/\partial a_1 \, MSE(y, f_{a_0,a_1}(x))$$
   an der Stelle (a_0, a_1).
5. Ersetze a_0 durch $a_0 - \alpha v_0$
   und a_1 durch $a_1 - \alpha v_1$.
6. Iteriere Schritte 3.-6.



https://upload.wikimedia.org/wikipedia/commons/f/ff/Gradient_descent.svg

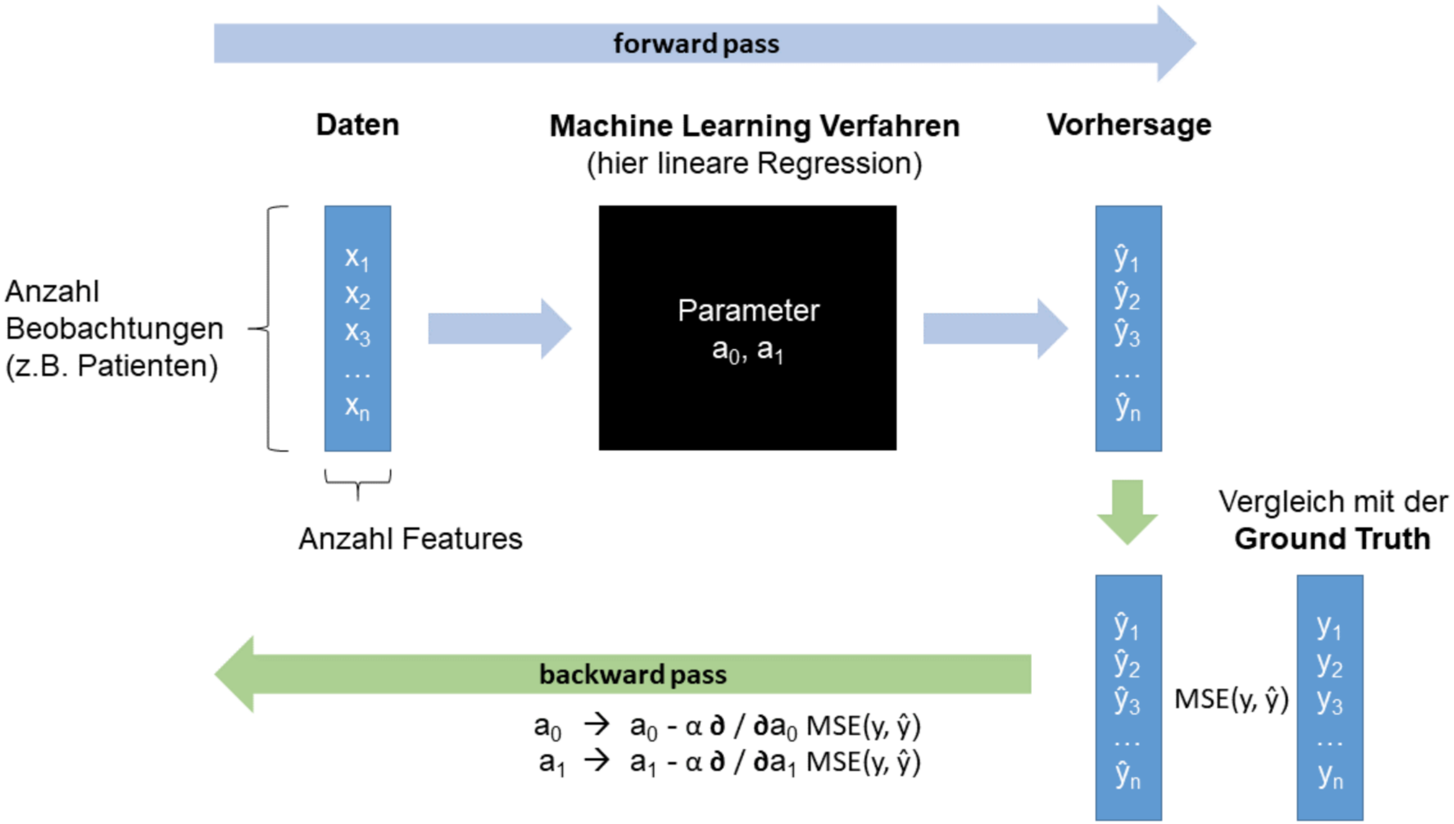
Bei der linearen Regression $f_{a_0,a_1}(x) = a_0 + a_1 x$, wenn Punkte $(x_1, y_1), \ldots, (x_n, y_n)$ gegeben sind:

$$
\begin{aligned}
v_0 &= \partial/\partial a_0\, MSE(\mathbf{y}, f_{a_0,a_1}(\mathbf{x})) \\
&= \partial/\partial a_0 \frac{1}{n} \sum_{i=1}^{n} (y_i - a_0 - a_1 \cdot x_i)^2 \\
&= -\frac{2}{n} \sum_{i=1}^{n} (y_i - a_0 - a_1 \cdot x_i)
\end{aligned}
$$

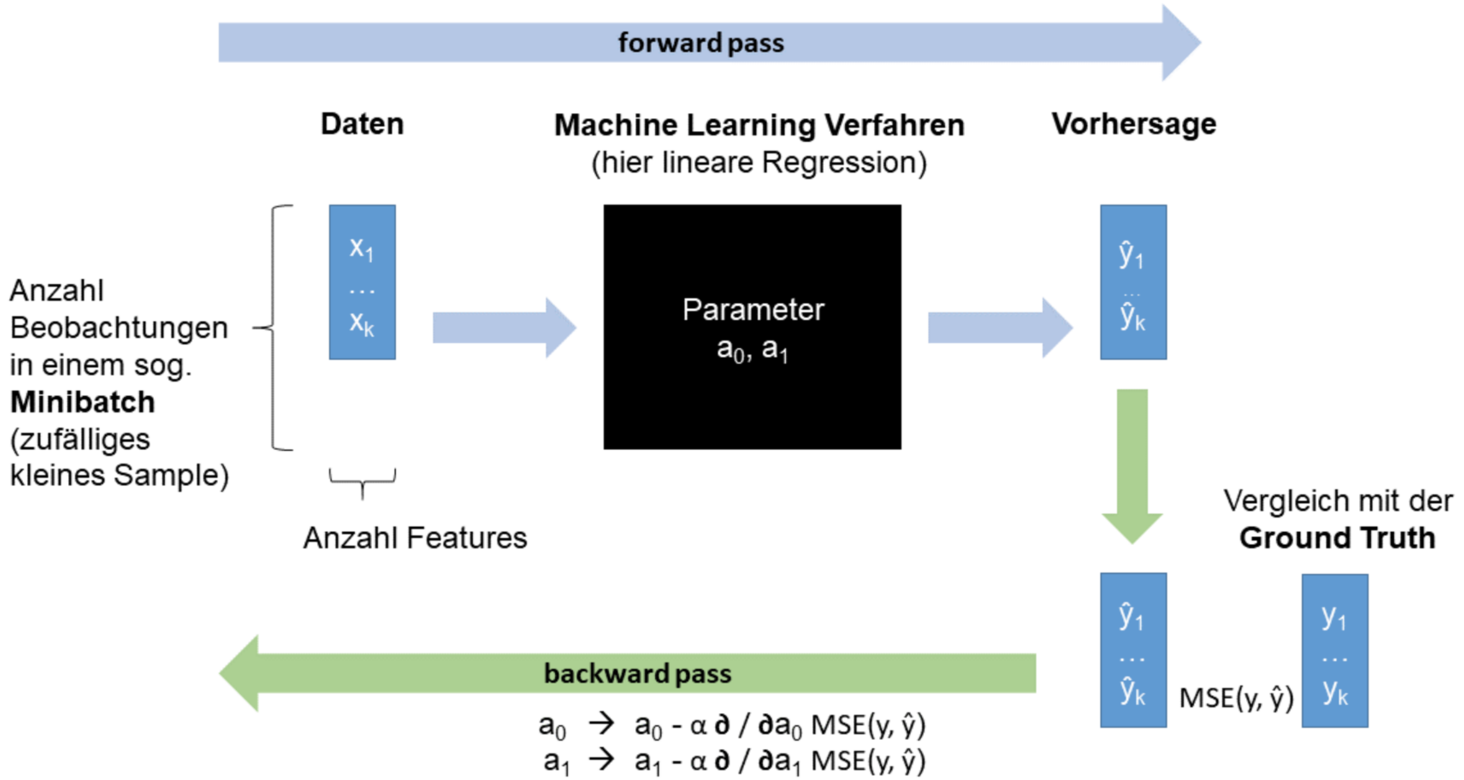
und

$$
\begin{aligned}
v_1 &= \partial/\partial a_1\, MSE(\mathbf{y}, f_{a_0,a_1}(\mathbf{x})) \\
&= -\frac{2}{n} \sum_{i=1}^{n} x_i (y_i - a_0 - a_1 \cdot x_i)
\end{aligned}
$$

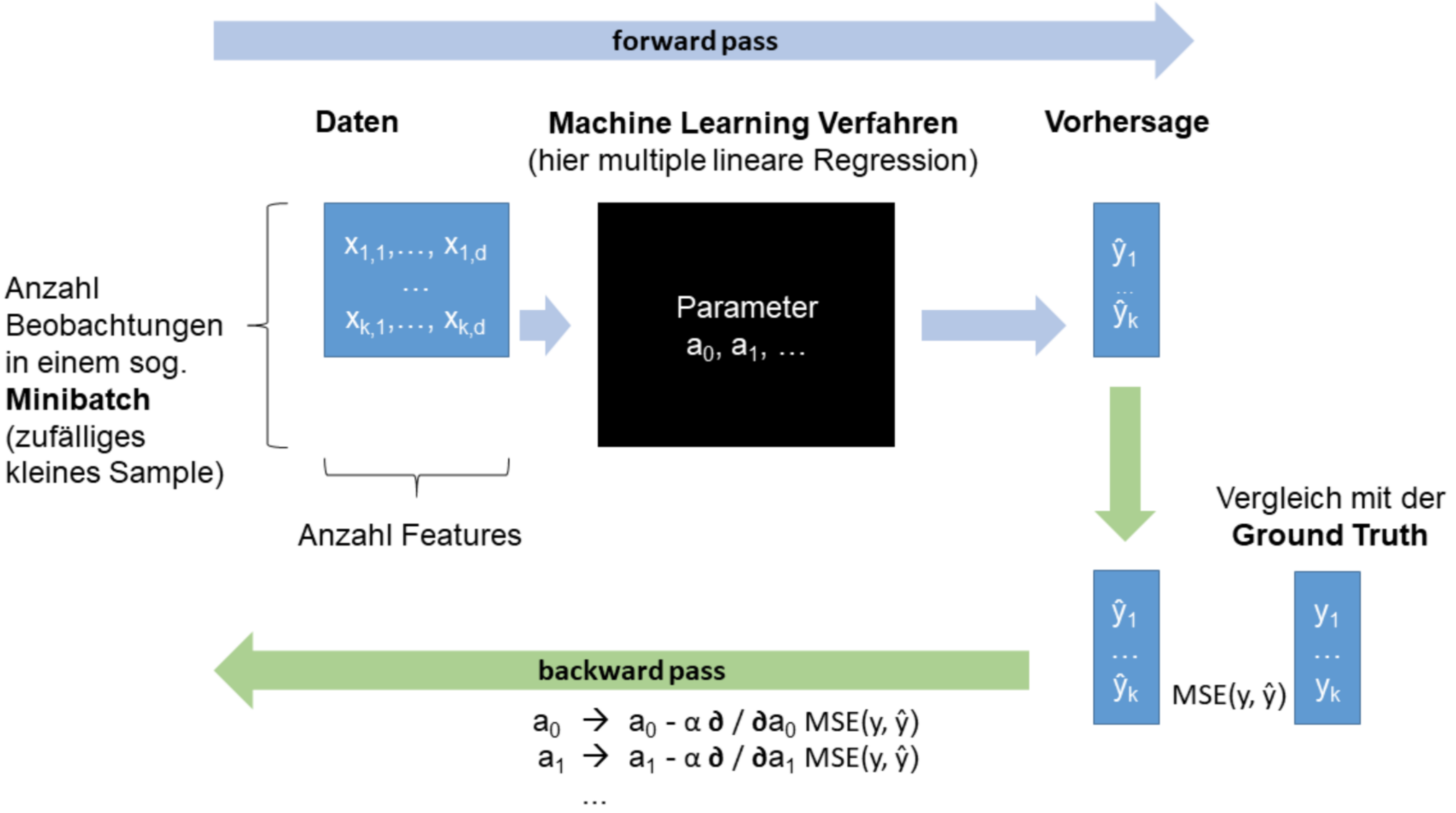
Schematisch für eine Lineare Regression $\hat{y}_i = a_0 + a_1 \cdot x_i$:



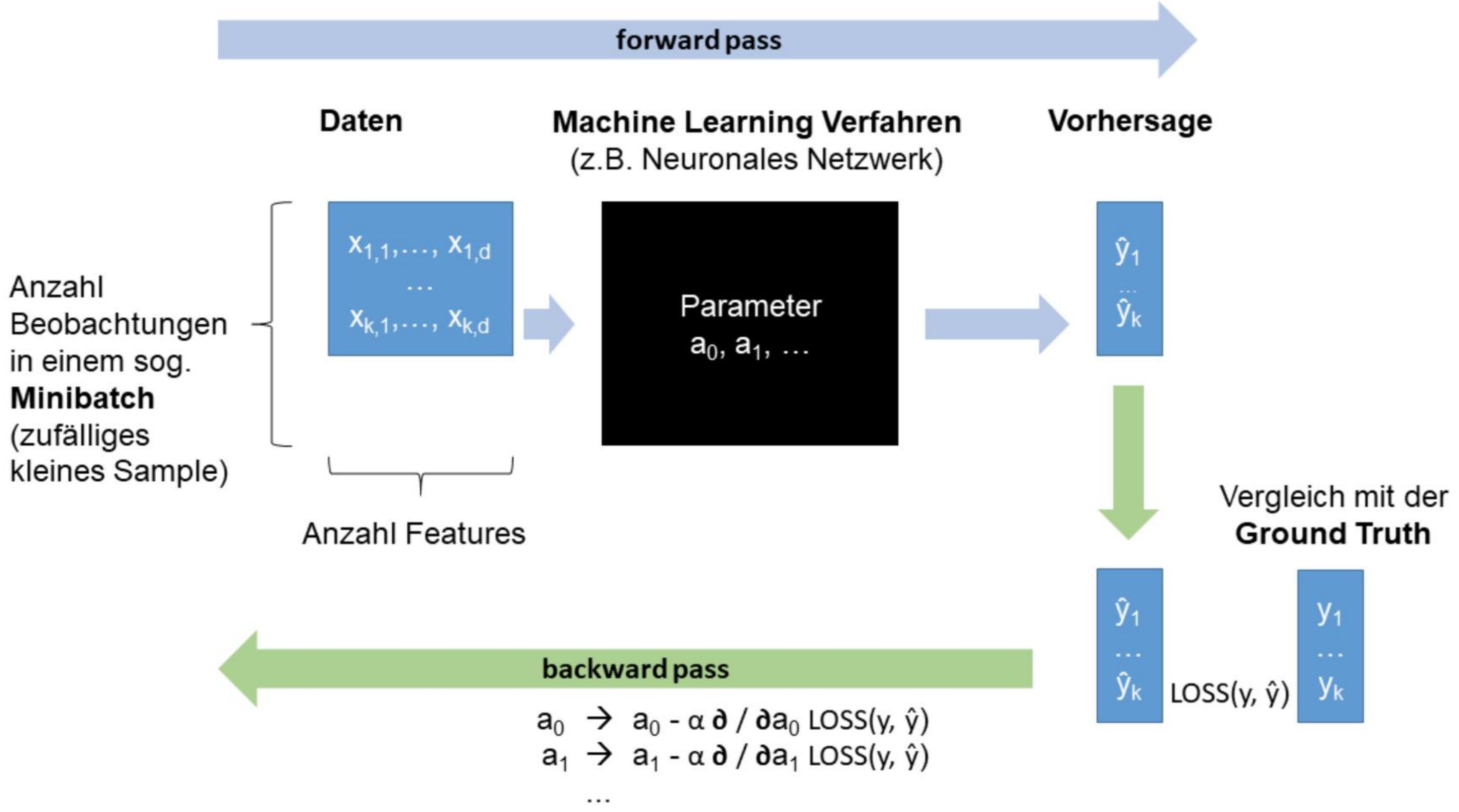
Stochastic Gradient Descent: Berechne die Gradienten nur auf zufällig gewählten kleineren Datenmengen.



Analog, wenn das Verfahren mehr Parameter hat oder mehr als ein Feature vorliegt (z.B. Multiple Lineare Regression $\hat{y}_i = a_0 + a_1 \cdot x_{i1} + \ldots + a_d \cdot x_{id}$):

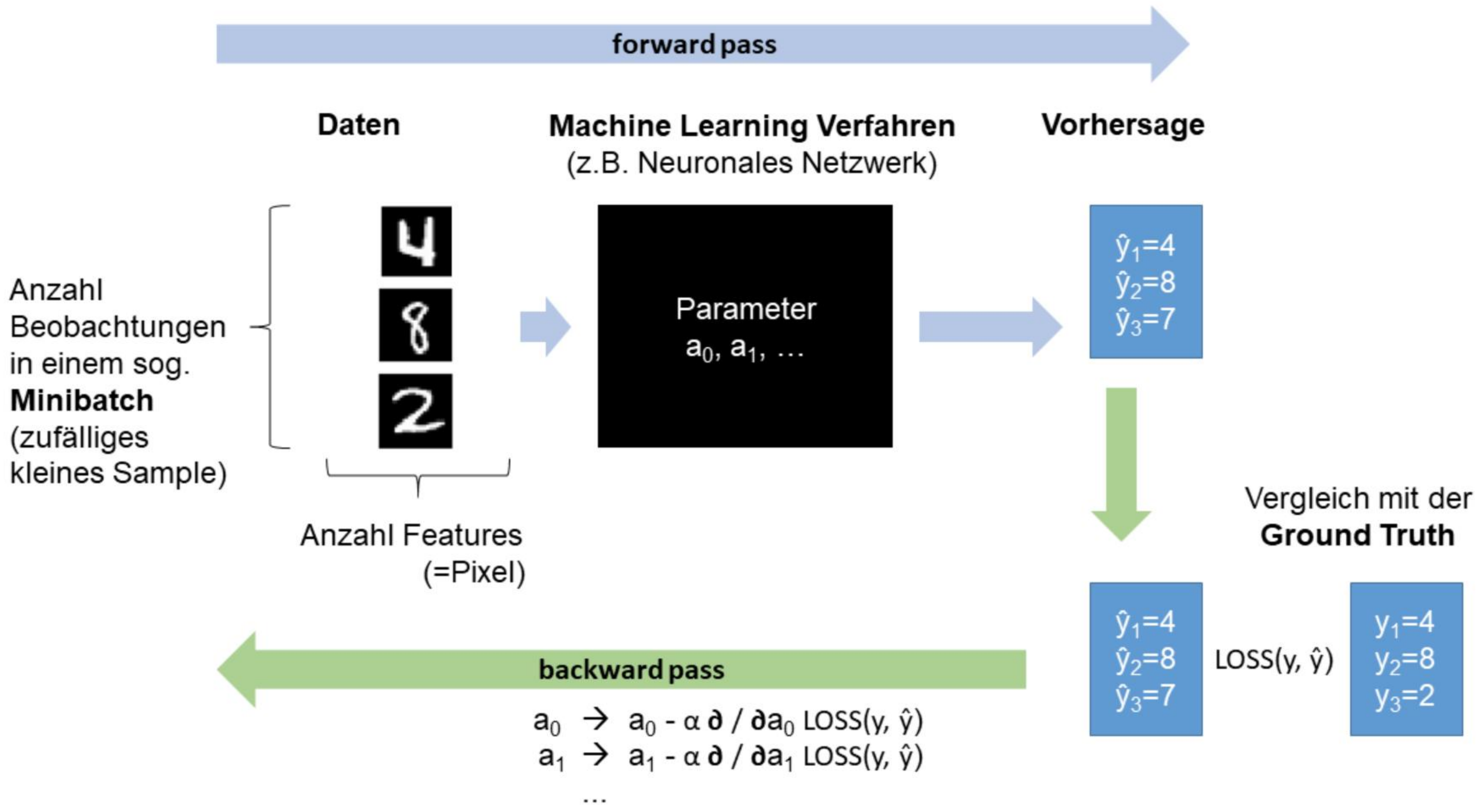


Das Machine Learning Verfahren könnte genauso ein **Neuronales Netzwerk** sein (und bei geeigneter Loss-Funktion auch eine Klassifikation statt Regression):

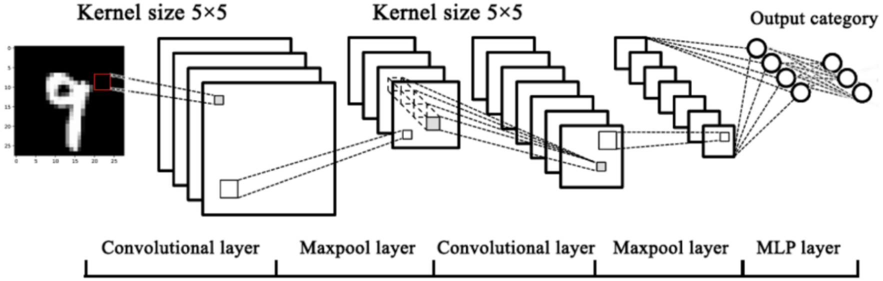


Beispiel: Die Features könnten Pixel sein.

(in diesem Klassifikationstask würde man sich die CrossEntropyLoss anschauen; die Vorhersage wäre ein Wahrscheinlichkeitsvektor)



Neuronale Netzwerke sind typischerweise in Schichten aufgebaut (hier ein CNN=*Convolutional Neural Network*):



Wang, J.J., Hu, S.G., Zhan, X.T. et al. Handwritten-Digit Recognition by Hybrid Convolutional Neural Network based on HfO2 Memristive Spiking-Neuron. Sci Rep 8, 12546 (2018)

backward pass

$a_0 \rightarrow a_0 - \alpha\, \partial / \partial a_0 \, \mathrm{LOSS}(y, \hat{y})$
$a_1 \rightarrow a_1 - \alpha\, \partial / \partial a_1 \, \mathrm{LOSS}(y, \hat{y})$
…

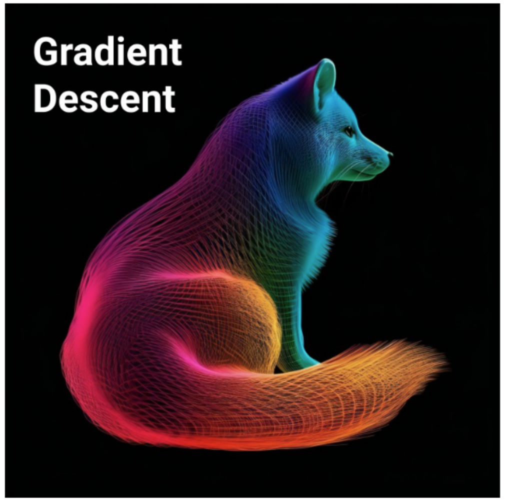
"Schichten": Die Vorhersage $\hat{y}$ kann als Hintereinanderausführung verschiedener Funktionen dargestellt werden:

$$\hat{y} = f_n(f_{n-1}(\ldots(f_1(x))))$$

In der Ableitung $\partial/\partial a_0 \, \mathrm{LOSS}(y, \hat{y})$ kann daher die **Kettenregel** verwendet werden:

$$\partial/\partial a_0 \, f_2(f_1(x; a_0)) = (\partial/\partial a_0 \, f_2)(f_1(x; a_0)) \cdot \partial/\partial a_0 \, f_1(x; a_0)$$

Die Optimierung aller Parameter eines neuronalen Netzwerks mittels der Kettenregel nennt man **Back Propagation**.



von: Jepson Taylor, Dataiku
(generiert mit MidJourney)

Gradient Descent is like a hiker trying to find the lowest point in a valley by taking steps downhill. The hiker checks which direction is steepest, takes a step in that direction, and repeats the process until they reach the bottom, where the terrain is flat and there's no more descending to do.

1. Lineare Regression
2. Gesundheitsdaten
3. Gradient Descent

Vielen Dank für Ihre Aufmerksamkeit!