

# Daten und KI

Wie Maschinen richtig funktionieren  
und doch falsch liegen können

Prof. Dr. Dirk Nowotka

Institut für Informatik, CAU

# Computer sind die perfekten Rechner

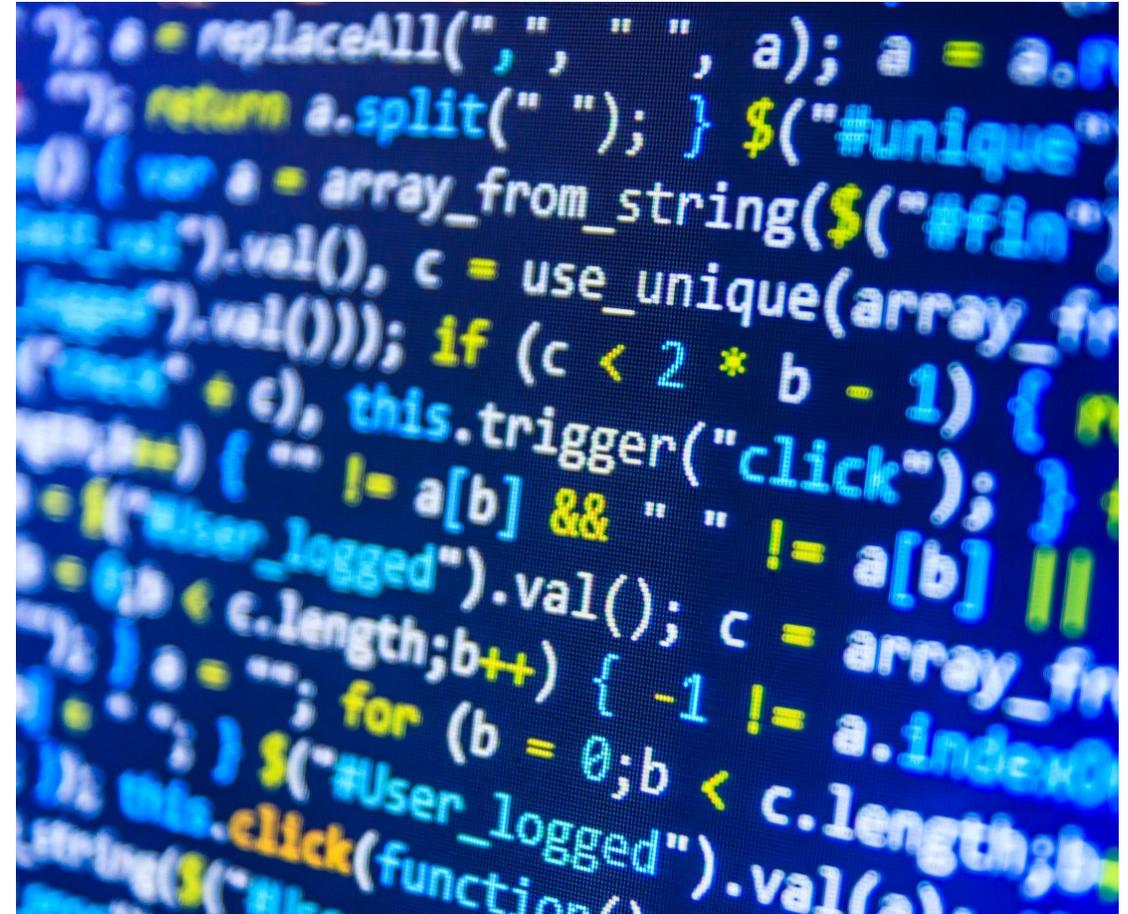
Sie sind schnell, präzise, korrekt.

... wenn sie die richtigen Befehle erhalten.

Wenn Eingabe=1 dann ...

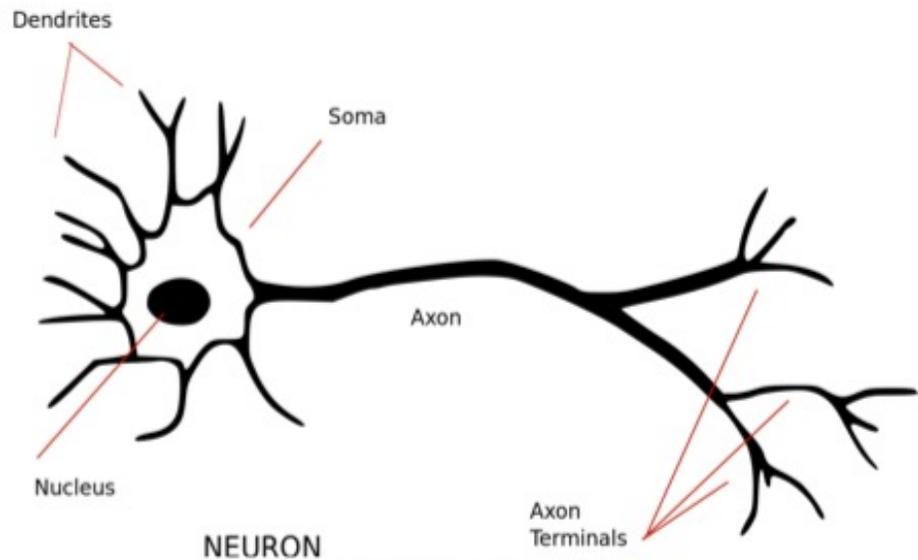
Wiederhole 5 mal ...

Berechne  $3x-17y$ .



Wie beschreibt man: „Alle Bilder mit einem Hund“?

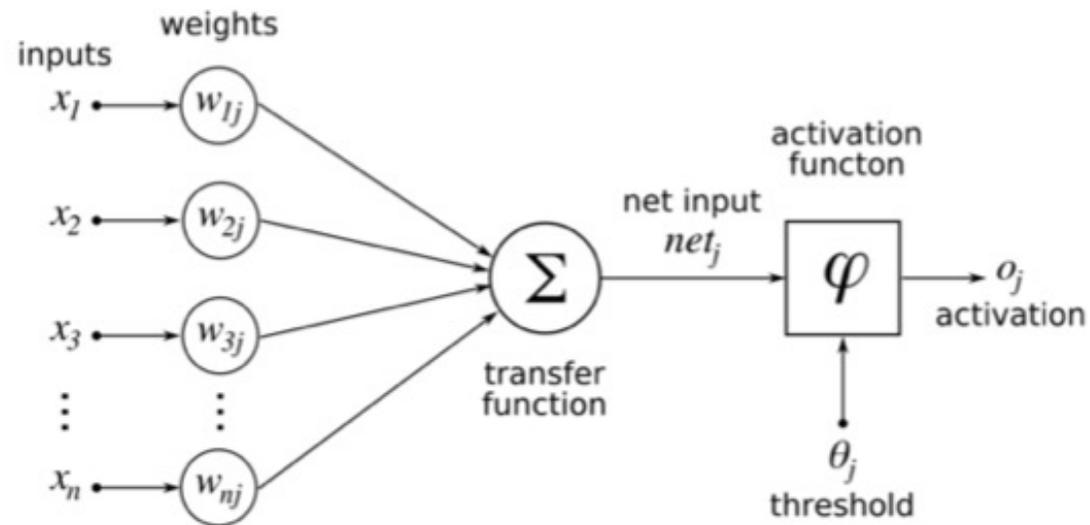




NEURON

A biological Neuron

Source: [https://commons.wikimedia.org/wiki/File:Neuron\\_-\\_annotated.svg](https://commons.wikimedia.org/wiki/File:Neuron_-_annotated.svg)

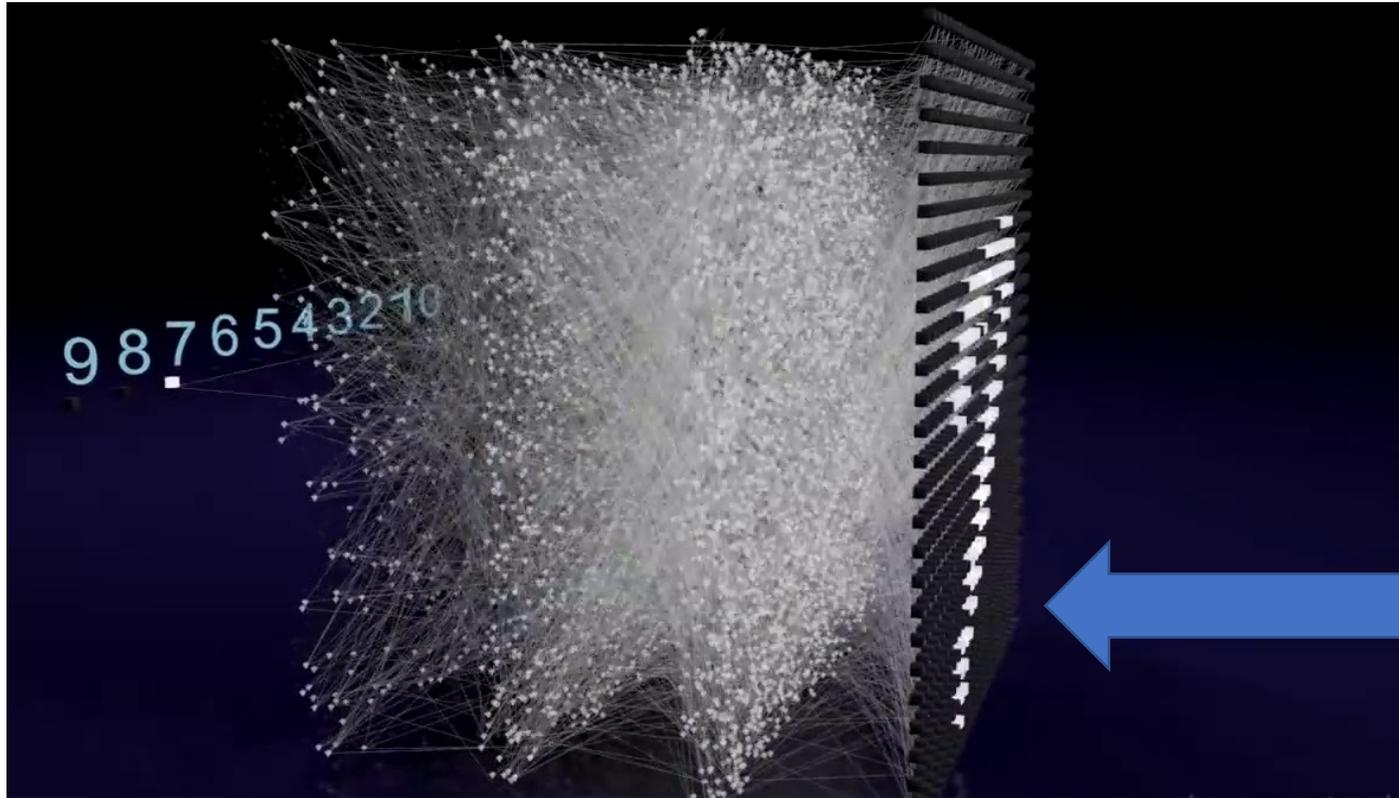


An Artificial Neuron

Source: <https://commons.wikimedia.org/wiki/File:Rosenblattperceptron.png>

# Siegeszug der KI: Es begann vor 50 Jahren...

Automatisierte Briefsortierung seit den 1970er Jahren



## Trainingsdaten (USPS)

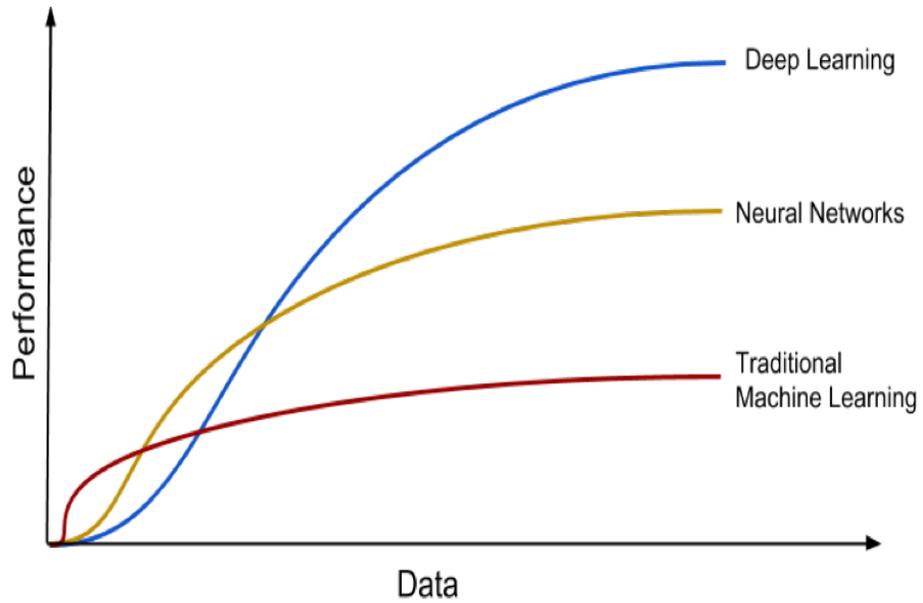
0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9	9	9

60.000 Ziffern

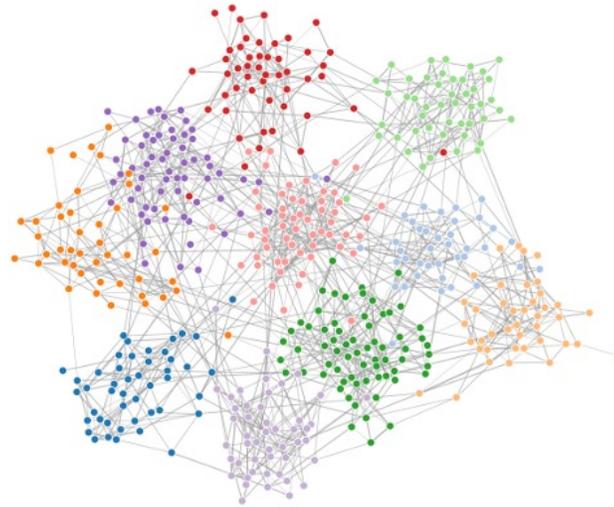
Neuronales Netz wird trainiert

**Heute: Automatische Ziffern-Erkennung: >99,8% Genauigkeit!**

# Warum wird KI erst jetzt wichtig?



Neue Algorithmen



Mehr Daten



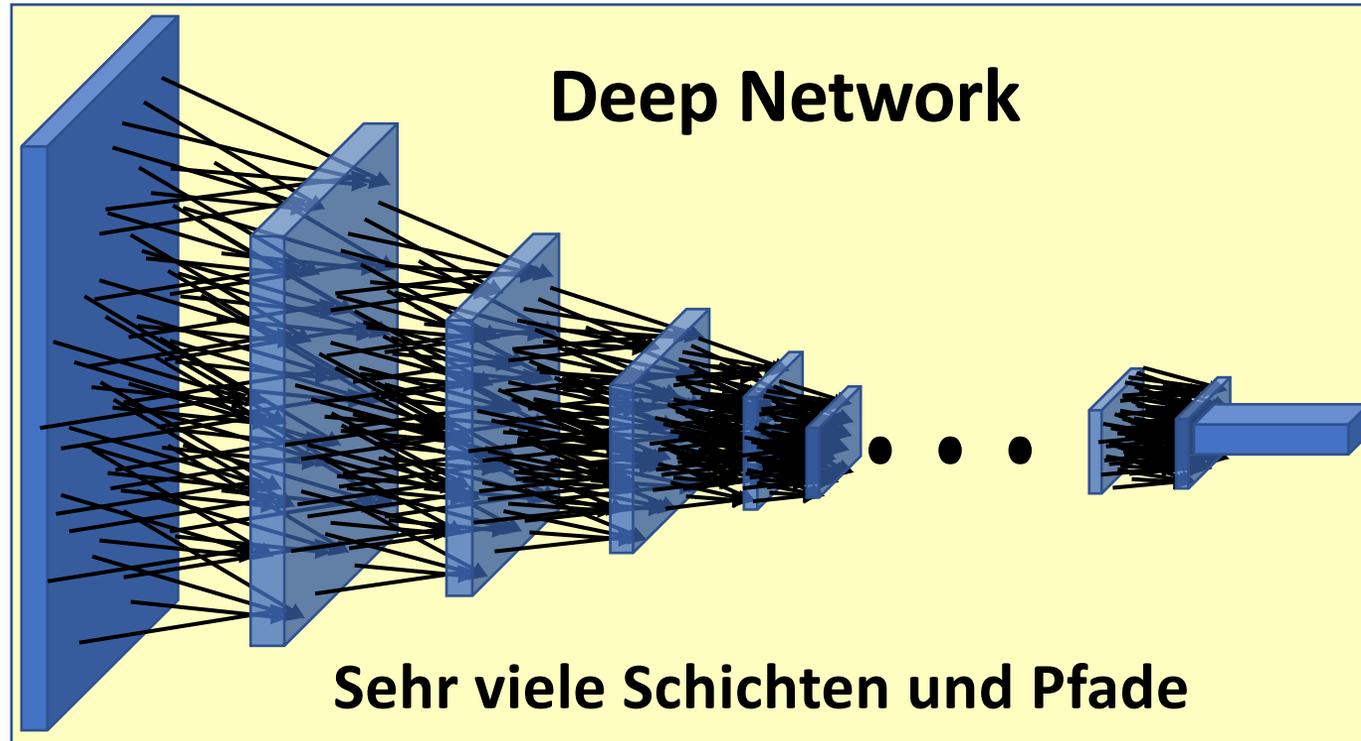
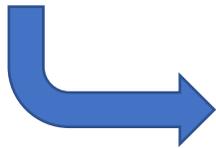
Mehr Rechenleistung

# KI heute: Deep Networks, viele neuronale Schichten

14 Millionen Trainings-Bilder, 100 Millionen Pfade, 1000 Klassen

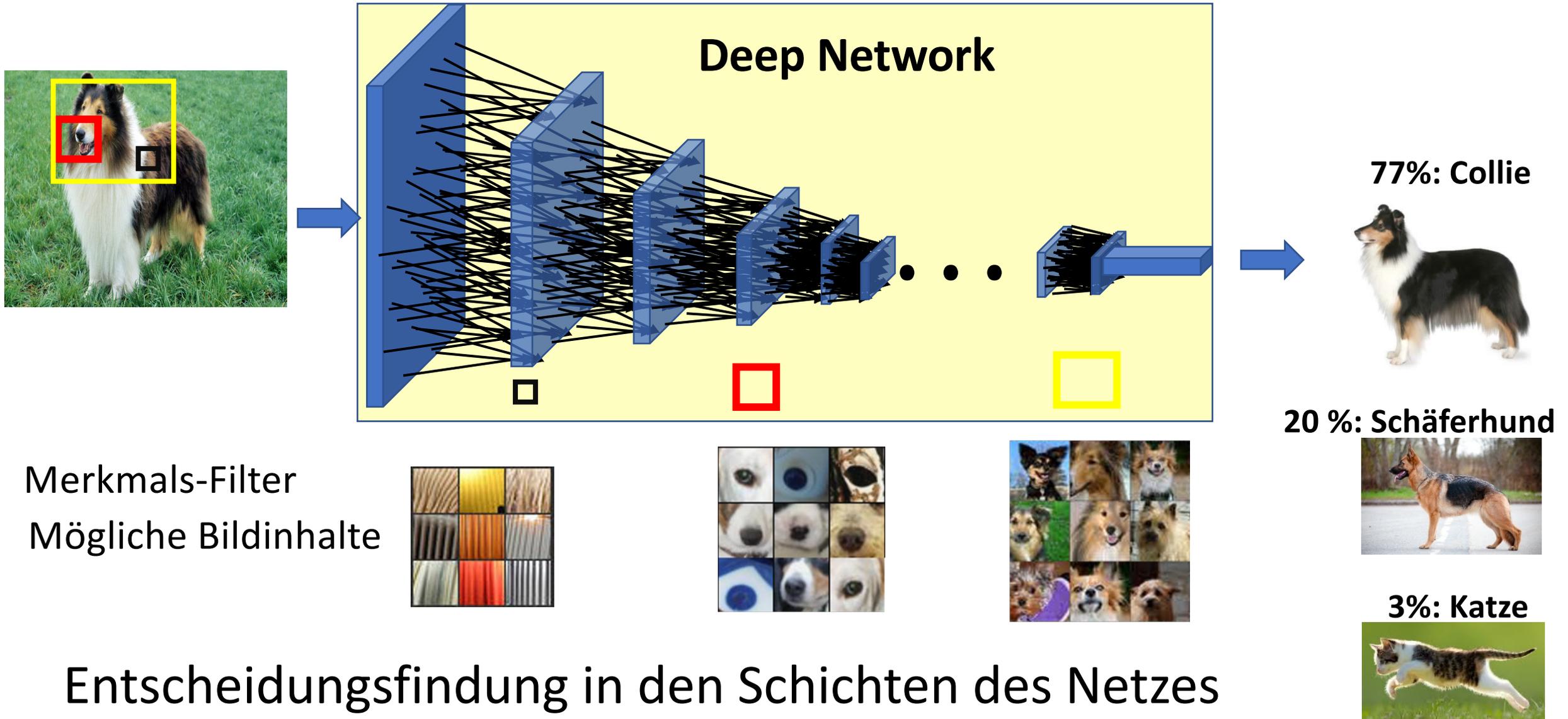
**Training: Bilder rein**

**Objektklasse angeben**

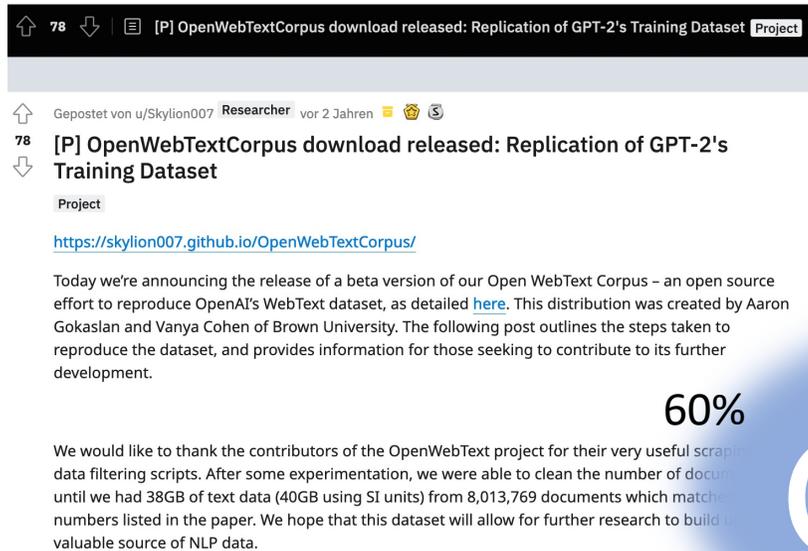


Überwachtes Lernen: Zeige ein Bild, benenne das Objekt

# Klassifikation: Was ist zu sehen ?



# Wie weit können wir heute gehen?



78 [P] OpenWebTextCorpus download released: Replication of GPT-2's Training Dataset Project

Gepostet von u/Skylion007 Researcher vor 2 Jahren

## [P] OpenWebTextCorpus download released: Replication of GPT-2's Training Dataset

Project

<https://skylion007.github.io/OpenWebTextCorpus/>

Today we're announcing the release of a beta version of our Open WebText Corpus – an open source effort to reproduce OpenAI's WebText dataset, as detailed [here](#). This distribution was created by Aaron Gokaslan and Vanya Cohen of Brown University. The following post outlines the steps taken to reproduce the dataset, and provides information for those seeking to contribute to its further development.

We would like to thank the contributors of the OpenWebText project for their very useful scripts and data filtering scripts. After some experimentation, we were able to clean the number of documents until we had 38GB of text data (40GB using SI units) from 8,013,769 documents which matches the numbers listed in the paper. We hope that this dataset will allow for further research to build a valuable source of NLP data.



## Common Crawl

BIG PICTURE - THE DATA - ABOUT - BLOG - CONNECT - Donate

### So you're ready to get started.

The Common Crawl corpus contains petabytes of data collected since 2008. It contains raw web page data, extracted metadata and text extractions.

#### Data Location

The Common Crawl dataset lives on Amazon S3 as part of the [Amazon Public Datasets](#) program. From Public Data Sets, you can download the files entirely free using HTTP or S3.

22% As the Common Crawl Foundation has evolved over the years, so has the format and metadata that accompany the crawls themselves.

Get your sea legs by checking out some sample code

Take a look at some of the many open source libraries and projects for using Common Crawl data. [Sample code](#) is available for a number of languages and includes everything from machine learning to natural language processing. [Check it out](#) »

# GPT-3

60%



Wikipedia

16%



Smashwords™  
your ebook. your way.

BookCorpus

96 Schichten, 175 Milliarden Parameter

# Eindrucksvolle Ergebnisse

Q: What is human life expectancy in the United States?

A: Human life expectancy in the United States is 78 years.

Q: Who was president of the United States in 1955?

A: Dwight D. Eisenhower was president of the United States in 1955.

Q: What party did he belong to?

A: He belonged to the Republican Party.

Q: Who was president of the United States before George W. Bush?

A: Bill Clinton was president of the United States before George W. Bush.

Q: Who won the World Series in 1995?

A: The Atlanta Braves won the World Series in 1995.

 **Mario Klingemann** ✓  
@quasimondo



Another attempt at a longer piece. An imaginary Jerome K. Jerome writes about Twitter. All I seeded was the title, the author's name and the first "It", the rest is done by [#gpt3](#)

Here is the full-length version as a PDF:  
[drive.google.com/file/d/1qtPa1c...](https://drive.google.com/file/d/1qtPa1c...)

## The importance of being on twitter

by Jerome K. Jerome  
London, Summer 1897

It is a curious fact that the last remaining form of social life in which the people of London are still interested is Twitter. I was struck with this curious fact when I went on one of my periodical holidays to the sea-side, and found the whole place twittering like a starling-cage. I called it an anomaly, and it is.

I spoke to the sexton, whose cottage, like all sexton's cottages, is full of antiquities and interesting relics of former centuries. I said to him, "My dear sexton, what does all this twittering mean?" And he replied, "Why, sir, of course it means Twitter."  
"Ah!" I said, "I know about that. But what is Twitter?"

"It is a system of short and pithy sentences strung together in groups, for the purpose of conveying useful information to the initiated, and entertainment and the exercise of wits to the initiated, and entertainment and the exercise of wits to the rest of us."

"Very interesting," I said. "Has it a name?"  
"It has," he said; "it is called Twitter."  
"Yes," I said, "I know that, but what is it?"  
"It is a system of information," he said.  
"Oh, yes," I replied; "but what is it?"

"Why, sir," he said, "you can go up to any of the gentlemen you see twittering in the street, and say to him, 'You are a fool,' or 'Your wife is an adulteress,' or 'You have stolen that hat,' and if he is a member of the initiated he will answer you in the same form and tell you that you are a liar, or that your eyes resemble the eyes of a duck, or that you have stepped out of your part in the last charade you acted in, or that you were for a short time a statistician in a Government Office, and he will go on to tell you the whole story of your life, in language so exceedingly small and pointed that even you will be glad you can't understand it."

5:25 PM · Jul 18, 2020



 1.2K  40  Copy link to Tweet

# KI ist eine Erfolgsgeschichte



# Ungebremster Fortschritt?



2015



Join Extra Crunch

Login

Search Q

TC Sessions:  
SaaS 2021

## Elon Musk, Sam Altman Predict Self-Driving Cars Will Be On The Road In Just A Few Years

Sarah Buhr @sarahbuhr / 2:55 AM GMT+2 • October 7, 2015

Comment

2021



TRANSPORTATION \ TESLA \ FEATURED STORIES

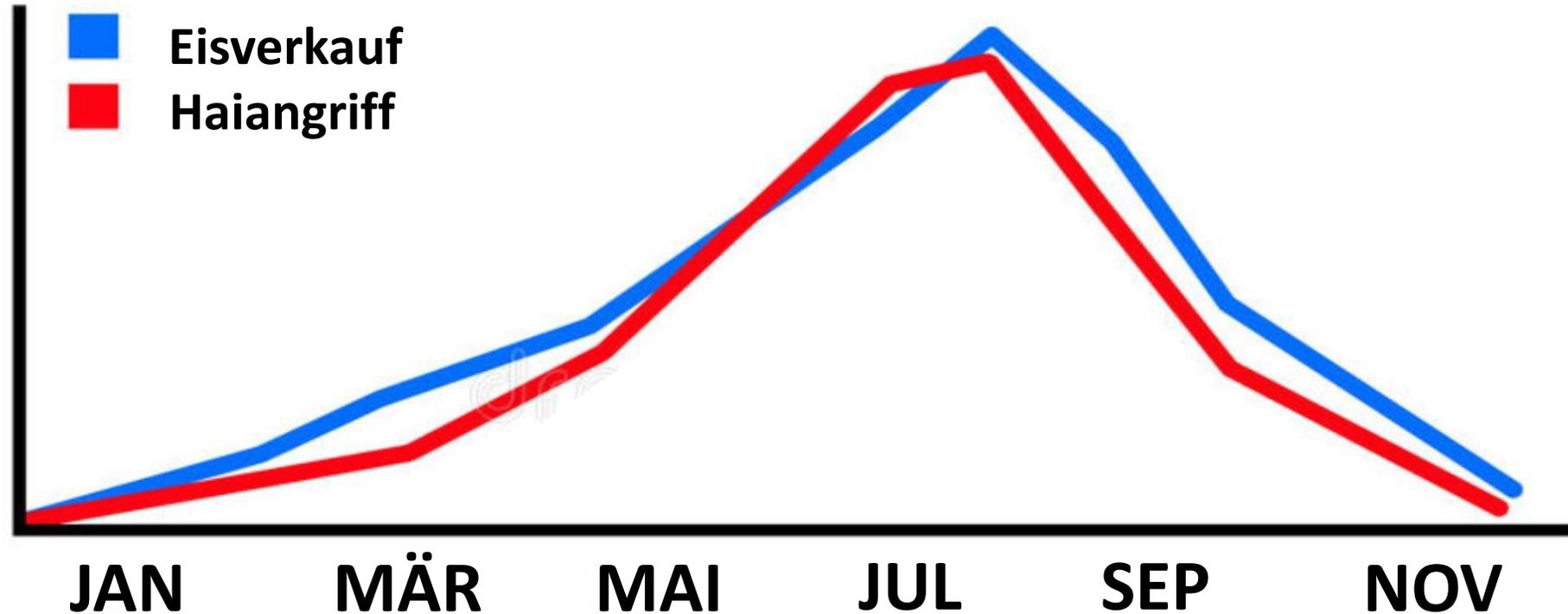
## Elon Musk just now realizing that self-driving cars are a 'hard problem'

279

*Overpromising and underdelivering*

By Andrew J. Hawkins | @andyjayhawk | Jul 5, 2021, 10:53am EDT

# Maschinen lernen Korrelationen in den Daten, nicht nach Kausalitäten



Sowohl die Eisverkäufe als auch die Haiangriffe nehmen zu, wenn das Wetter heiß und sonnig ist; aber sie werden nicht durch einander verursacht (sie werden durch gutes Wetter verursacht, viele Menschen am Strand, Eis essen und im Meer schwimmen)

# Ist GPT-3 intelligent?

Q: Who was president of the United States in 1700?

A: William Penn was president of the United States in 1700.

Q: Who was president of the United States in 1650?

A: Richard Bennett was president of the United States in 1650.

Q: Who was president of the United States in 1620?

A: James I was president of the United States in 1620.

Q: Who was president of the United States in 1600?

A: Queen Elizabeth I was president of the United States in 1600.

**Die USA wurde 1776 gegründet.**

[Kevin Lacker]

# Ist GPT-3 intelligent?

Q: How many eyes does a giraffe have?

A: A giraffe has two eyes.

Q: How many eyes does my foot have?

A: Your foot has two eyes.

Q: How many eyes does a spider have?

A: A spider has eight eyes.

Q: How many eyes does the sun have?

A: The sun has one eye.

Q: How many eyes does a blade of grass have?

A: A blade of grass has one eye.

**GPT-3 weiß nicht was ein Auge ist.**

[Kevin Lacker]

# Ist GPT-3 intelligent?

Q: How do you sporgle a morgle?

A: You sporgle a morgle by using a sporgle.

Q: How many bonks are in a quoit?

A: There are three bonks in a quoit.

Q: How many rainbows does it take to jump from Hawaii to seventeen?

A: It takes two rainbows to jump from Hawaii to seventeen.

Q: Which colorless green ideas sleep furiously?

A: Ideas that are colorless, green, and sleep furiously are the ideas of a sleep furiously.

Q: Do you understand these questions?

A: I understand these questions.

**GPT-3 weiß nicht wenn es nichts weiß.**

[Kevin Lacker]

# Ist GPT-3 intelligent?

Q: Who won the World Series in 2021?

A: The New York Yankees won the World Series in 2021.

Q: Who won the World Series in 2022?

A: The New York Yankees won the World Series in 2022.

Q: Who won the World Series in 2023?

A: The New York Yankees won the World Series in 2023.

Nun ja...

[Kevin Lacker]

# Skalierungsproblem?

## **GPT-3**

- 175 Mrd. Parameter
- Ca. 190.000 kWh  
Energieverbrauch in Lernphase

## **Menschliches Gehirn**

- 1.000.000 Mrd. Synapsen
- 20 Watt Leistung
- Täglicher Energieumsatz eines Erwachsenen: ca. 2,5 kWh

Mit der Energie von GPT-3 kann ein Mensch 1000 Jahre lang lernen.

Wäre GPT-3 so effizient wie ein Mensch, würde es mit 0,1 Mikrowatt lernen, d.h. 20 GPT-3 Systeme würden mit der Leistung einer Quarz-Armbanduhr lernen.

