XAI

Sprechen Sie mit der Black Box

Universität
Konstanz

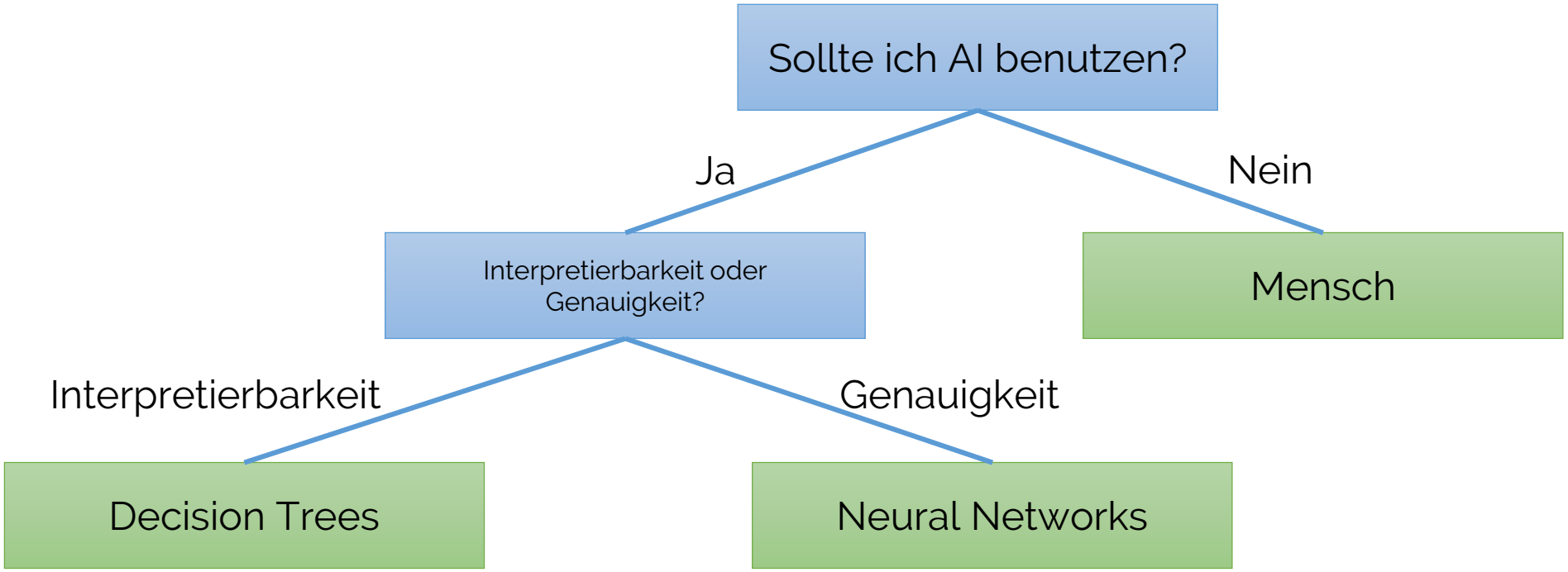
Geschichte und Entwicklung

Von Entscheidungsbäumen zu tiefen Neuronalen Netzwerken

From Decision Trees to Deep Neural Networks

Decision Trees – Einfache Entscheidungen

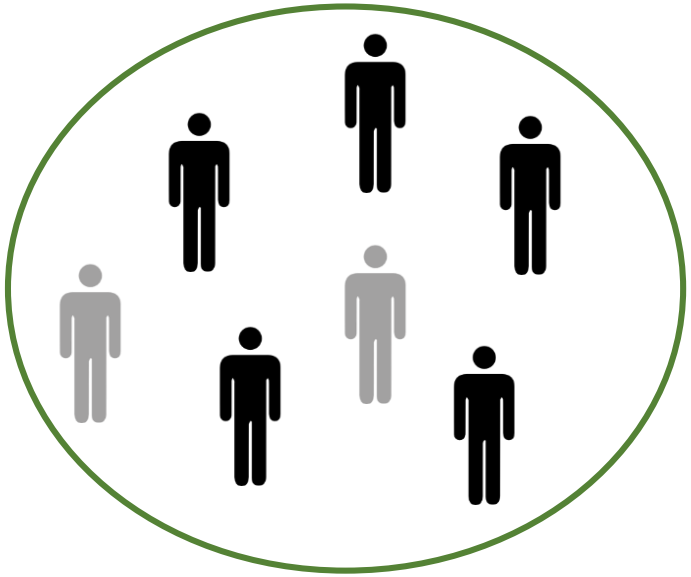
Das interpretierbare Modelle zur Entscheidungsfindung



Random Forests – Vom Baum zum Wald

Mehrere Decision Trees führen zu besseren Ergebnissen – Wisdom of the Crowd Theorem

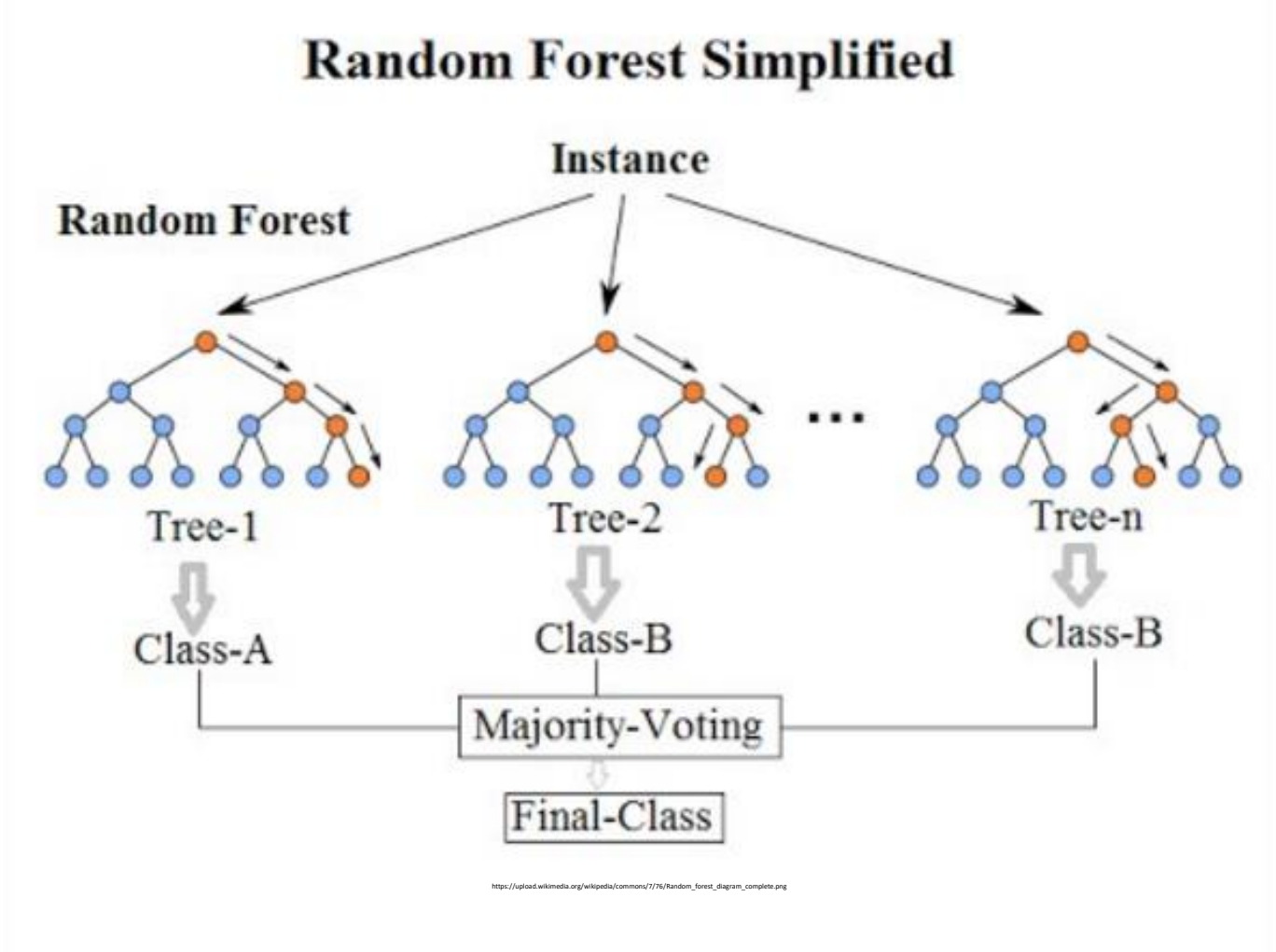
Entscheidung eines Experten
Decision Trees



Entscheidung mehrerer Experten
und weiteren
Decision Trees Forest

Random Forests – Decision Forest

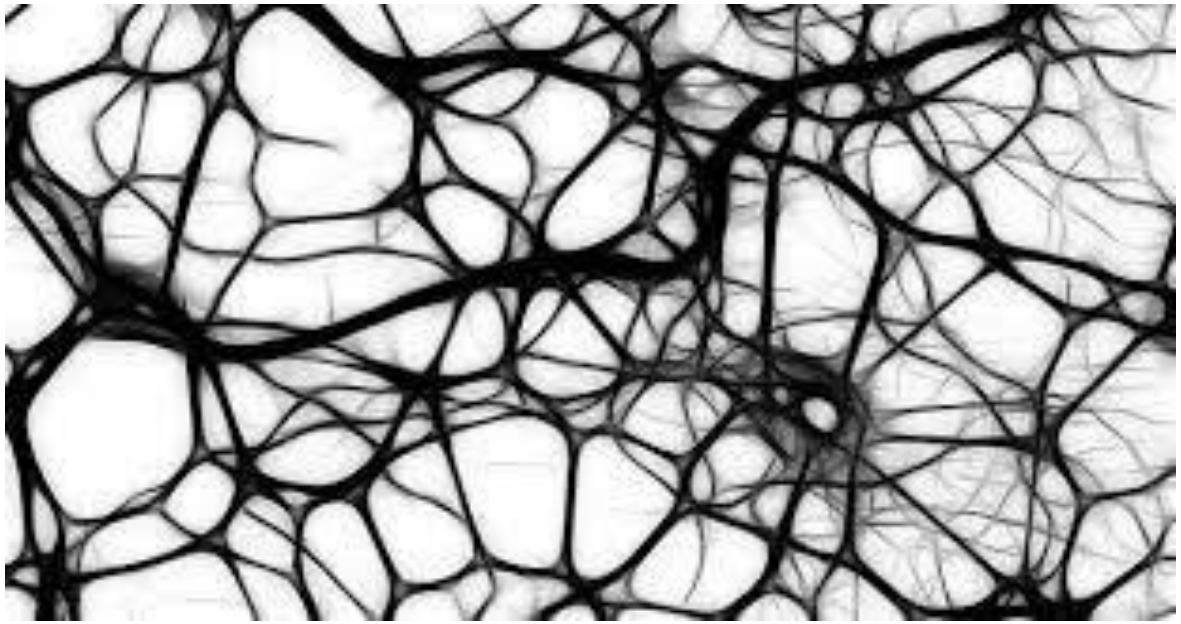
Höhere Genauigkeit – niedrigere Interpretierbarkeit



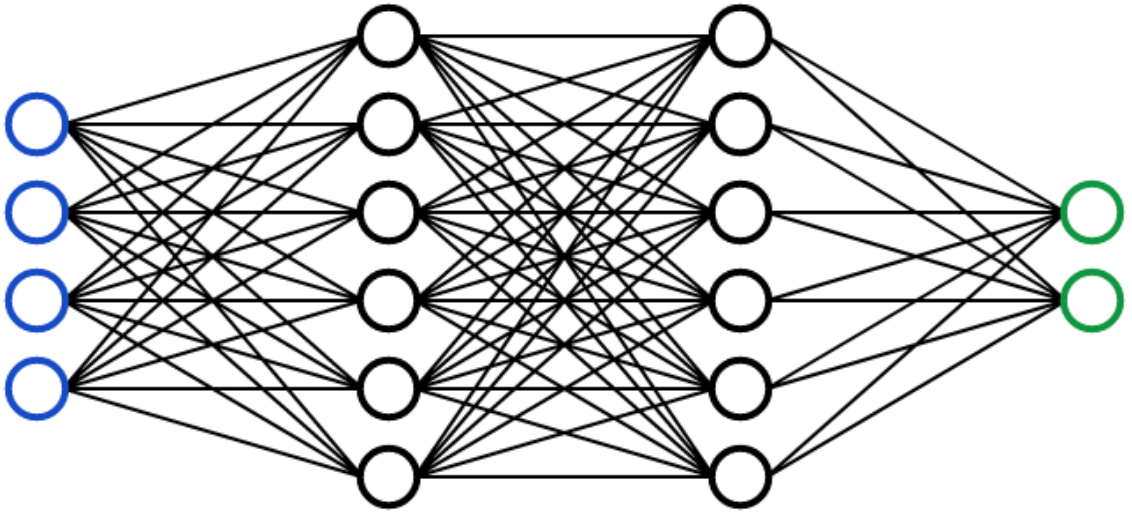
https://upload.wikimedia.org/wikipedia/commons/7/76/Random_forest_diagram_complete.png

Neural Networks – Die nächste Stufe

Modellierung des menschlichen Gehirns - Geflecht aus Neuronen und Synapsen – **State-of-the-art Genauigkeit**



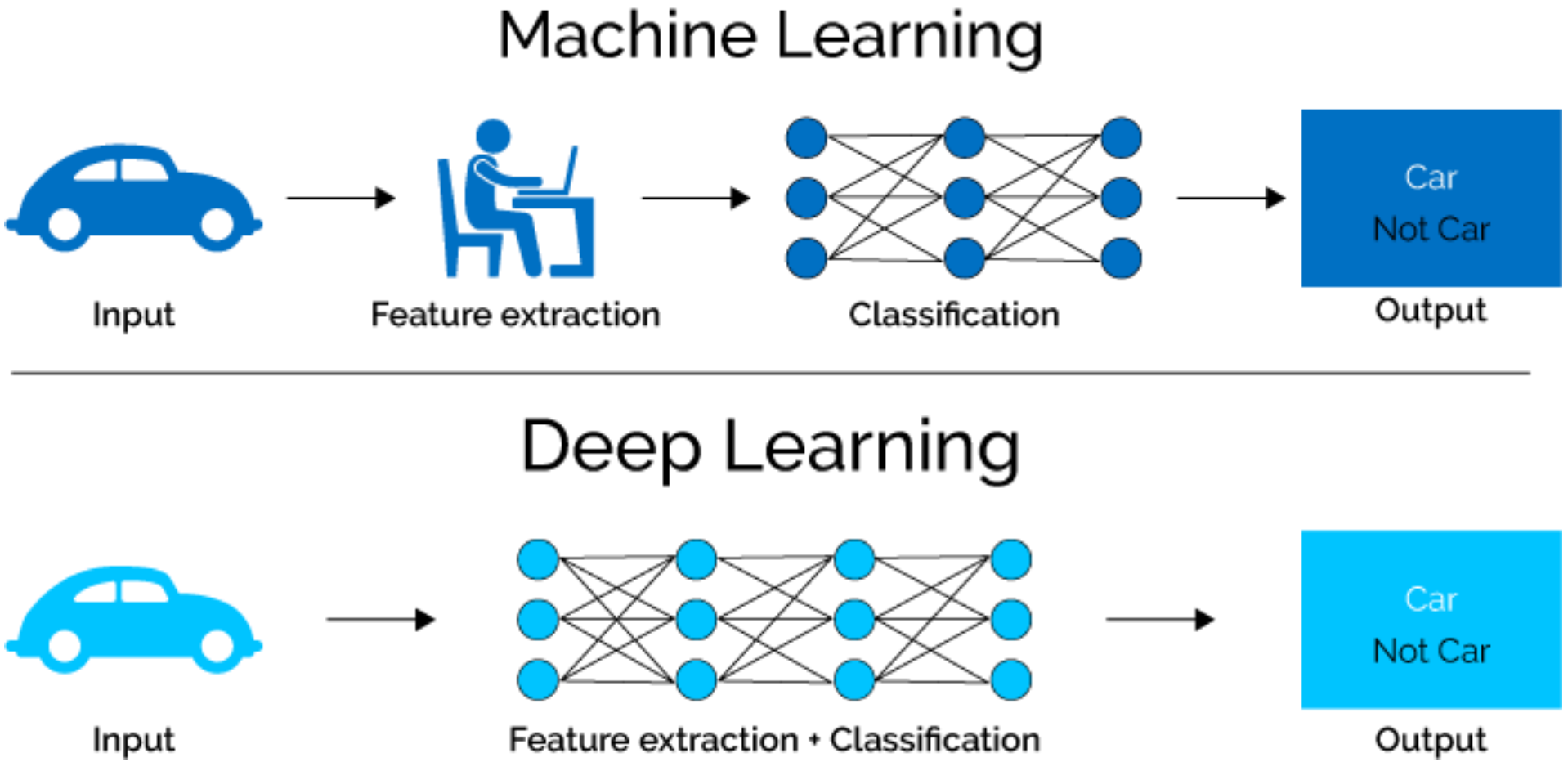
https://i.pinimg.com/originals/4a/9c/78/4a9c78ba5228069fb30afe9e2e1bd559.jpg



https://victorzhou.com/media/nn-series/network.svg

Deep Neural Revolution

Deep Learning ersetzt Domain Wissen und extrahierte Features des Menschen – Schwierige Interpretierbarkeit



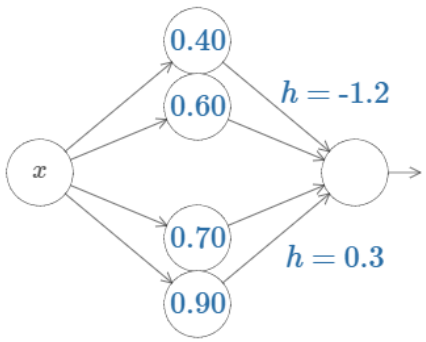
https://lawtomated.com/wp-content/uploads/2019/04/MLvsDL.png

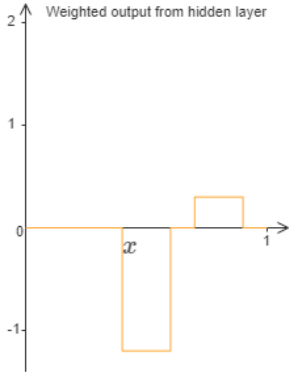
Von Erfolgen und Enttäuschungen

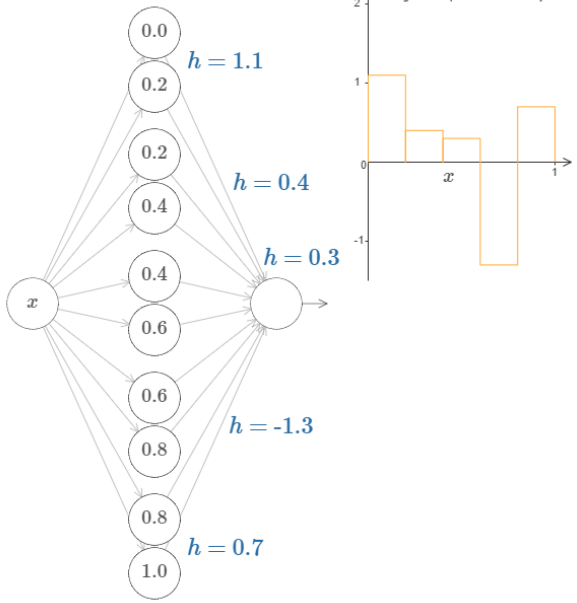
Deep Learning und dessen Probleme

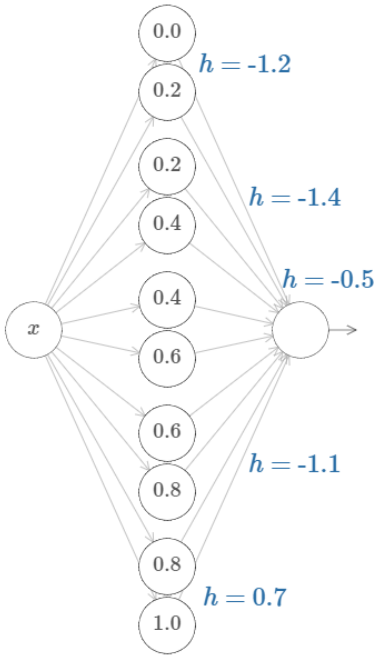
Universal Function Approximator

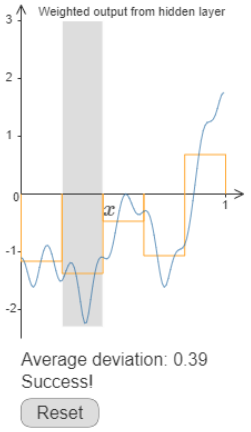
Neural Networks können theoretisch jede Funktion abbilden – somit alles lernen und lösen – Die Hoffnung











http://neuralnetworksanddeeplearning.com/chap4.html

Machine Translation

Traum des Menschen alle Sprachen zu verstehen und zu sprechen – Neural Networks als Hilfsmittel dazu

https://www.topbots.com/wp-content/uploads/2019/02/MT_feature_1600px.jpg

ALPAC Report

(Automatic Language Processing Advisory Committee) zeigt herbe Ernüchterung über AI und Neural Networks 1966
Dadurch nur wenige Forschungsgelder für AI in späteren Jahren

The ALPAC Report of 1966

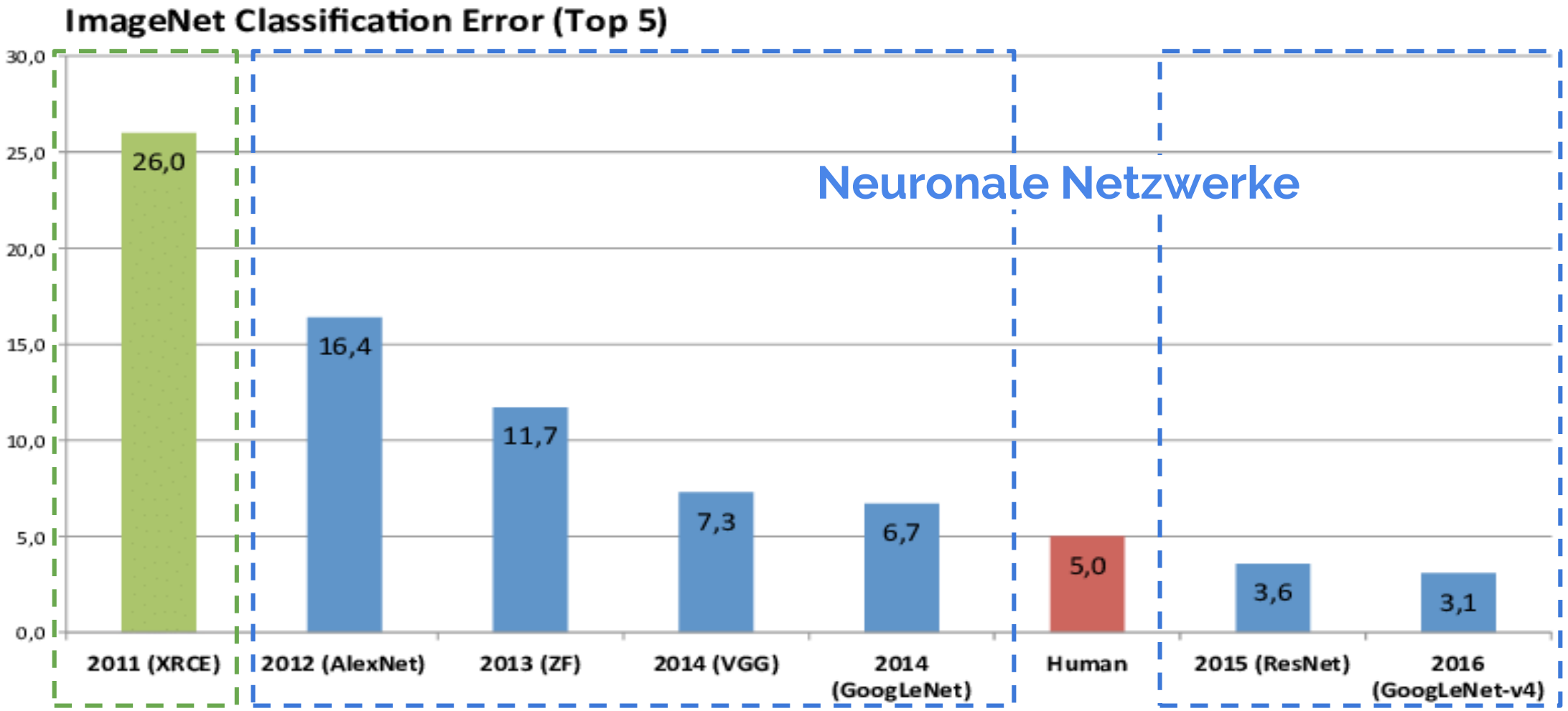
"They concluded, in a famous 1966 report, that machine translation was more expensive, less accurate and slower than human translation."

Photo: Eldon Lyttle, https://commons.wikimedia.org/wiki/File:Computer-translation_Briefing_for_Gerald_Ford.jpg

Nicht interpretierbar (Neural Networks)
Kostenintensive Analyse (Viel Computerleistung nötig)
Kostenintensive Forschung (Grundlagen fehlten)

Der Durchbruch von Deep Learning

Deep Learning schlägt traditionelle Machine Learning Verfahren in der Bildklassifizierung 2012



https://www.researchgate.net/profile/Gustav_Von_Zitzewitz/publication/324476862/figure/fig7/AS:614545865310213@1523530560584/Winner-results-of-the-ImageNet-large-scale-visual-recognition-challenge-LSVRC-of-the.png

The ImageNet winner: AlexNet

Deep Learning gewinnt Bildklassifizierung 2012 mittels Big Data und GPUs

Bis 2012 nur **gewonnen von Menschen-extrahierten Features**

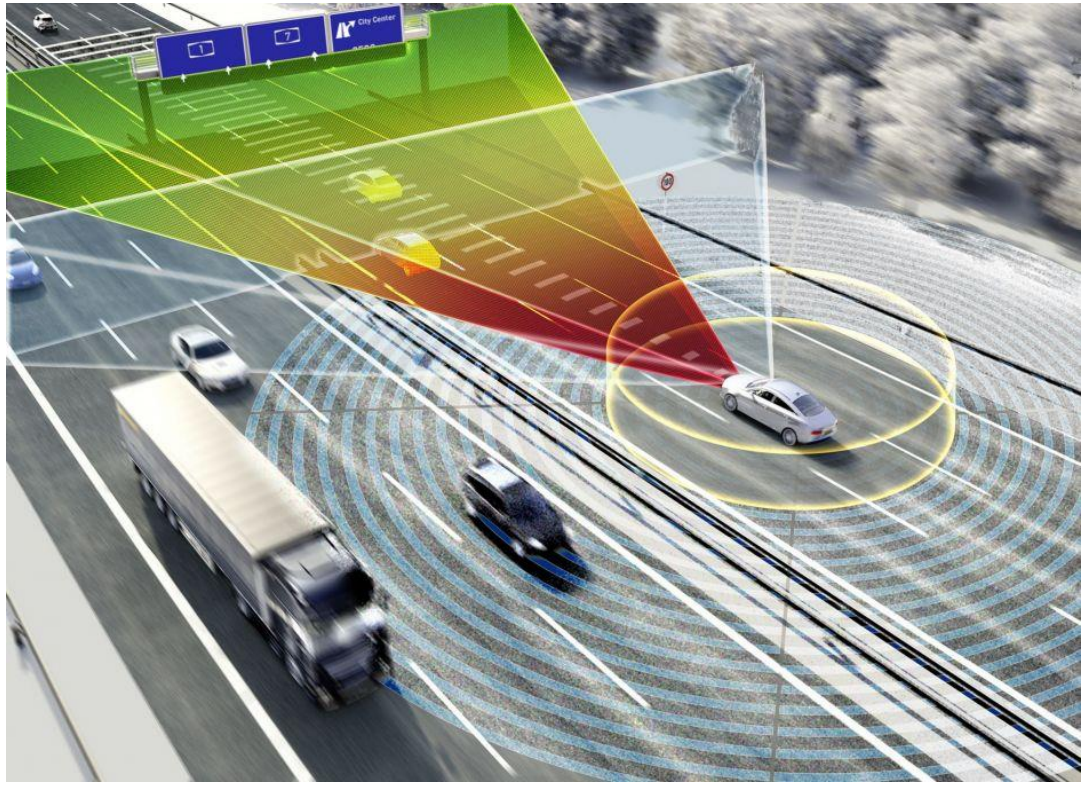
GPUs ermöglichen schnelles Trainieren der Neural Networks

Big Data stellt die Lernmittel bereit

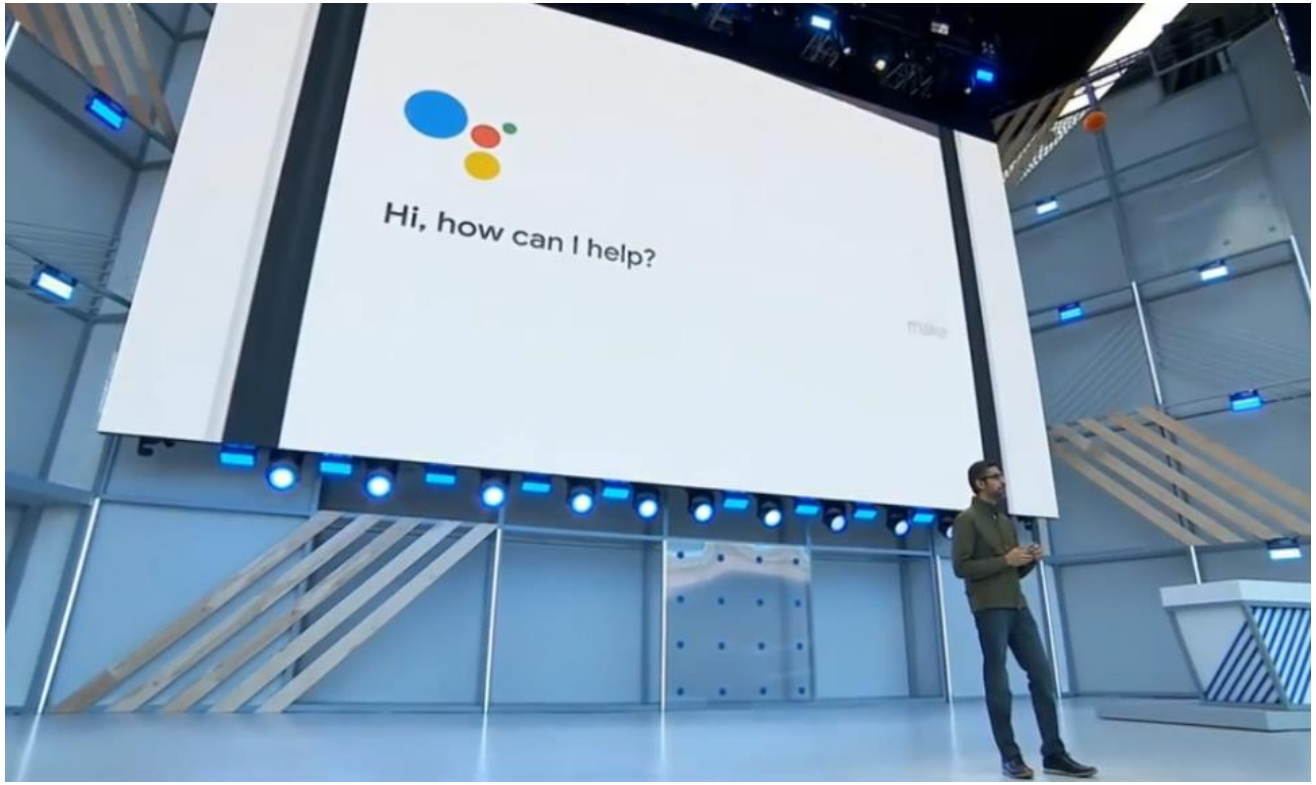
https://miro.medium.com/max/6328/1*WBMLfNKon41AFt8IpwbFFw.png

The neural revolution

Nach dem Durchbruch von Neural Networks in Bildklassifizierung, Einzug in viele andere Felder



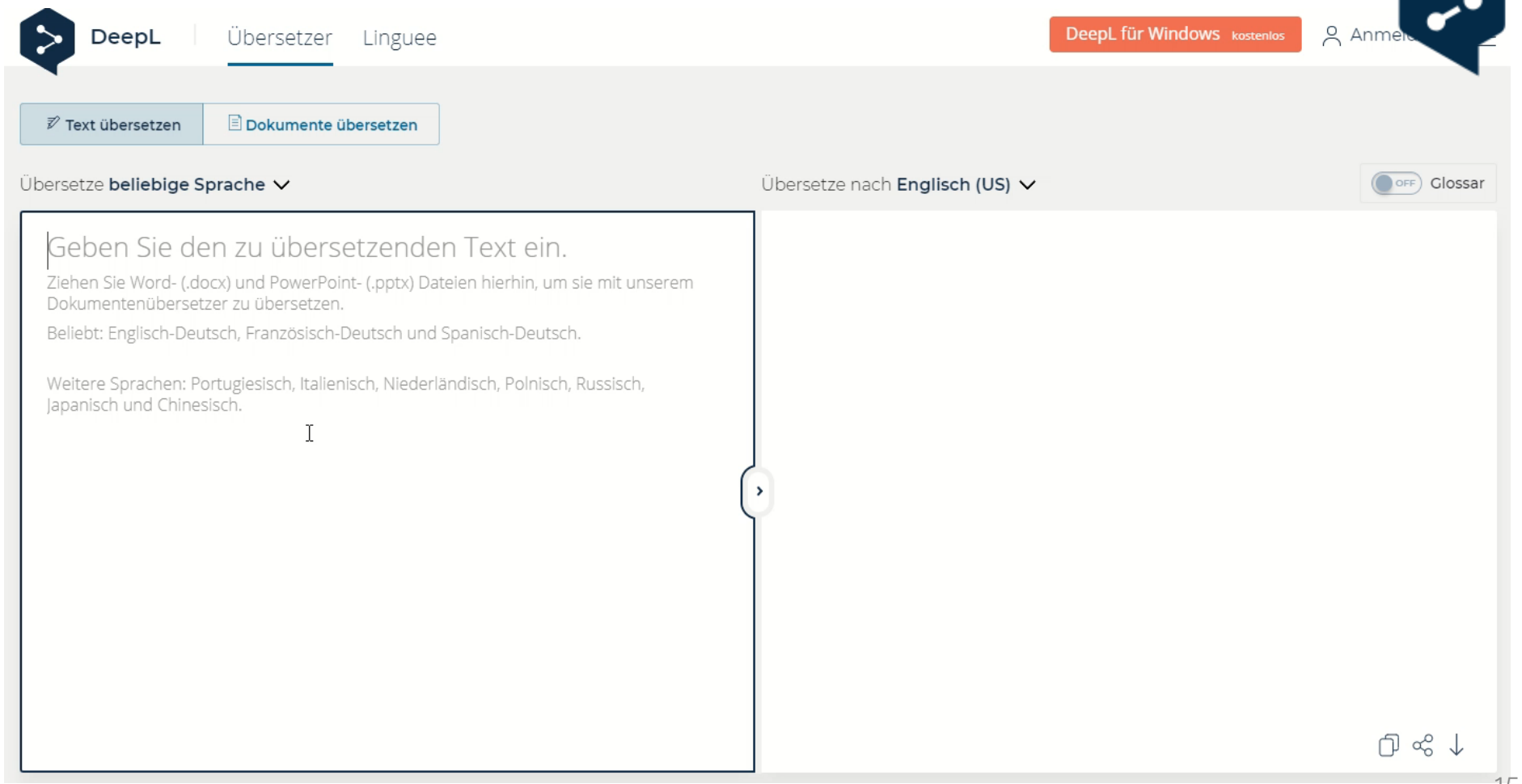
https://www.automobil-produktion.de/assets/images/b/continental_pp_hfl_comprehensiveenvironmentmodel-d5107b5b.jpg



https://www.youtube.com/watch?v=D5VN56jQMWM

Back to Machine Translation

Der Durchbruch in Machine Translation mit überzeugenden Ergebnissen



DeepL

Universität Konstanz

The novel problems

Mit neuem Erfolg kommen neue Probleme – Aktuelle Probleme die bisher unter den Tisch gefallen sind

- Fairness — Kann AI **fair** entscheiden?
- Bias — Kann AI **unvoreingenommen** entscheiden?
- Privacy — Kann AI **private Daten** sicher verwahren?
- Robustness — Kann AI **robust und zuverlässig** entscheiden?

Amazon Fairness / Bias

Neue Probleme schleichen sich in alle Bereiche mit ein

AI lernt auf Basis von Amazon Bewerbungsdaten

Problem:

Databias
(Mehr Jobs mit Männern besetzt als Frauen)

Fairness nicht berücksichtigt

Künstliche Intelligenz diskriminiert (noch)

Der Bewerbungsroboter von Amazon hat Frauen diskriminiert. Wie konnte das passieren? Und wie können Algorithmen geeignete Kandidaten für einen Job erkennen?

Von Felicitas Wilke, 18. Oktober 2018, 20:05 Uhr / 194 Kommentare

https://www.zeit.de/arbeit/2018-10/bewerbungsroboter-kuenstliche-intelligenz-amazon-frauen-diskriminierung

Deep Fakes

Nicht nur Fairness Probleme treten auf – Fälschungen und Manipulation wird einfacher

Gesichtsänderung in Videos mittels Deep Learning

Beispiel Facebook-Gründer Mark Zuckerberg der seinen Datenklau zugibt

XAI als mögliche Lösung

Wie man künstliche Intelligenz erklären kann

Wie Menschen ihr Entscheidungen begründen

Menschen können Entscheidungen auf Basis von Argumenten begründen – Verbalisierung steht im Vordergrund

Das sind **Züge**. Man sieht es an den **Gleisen**.

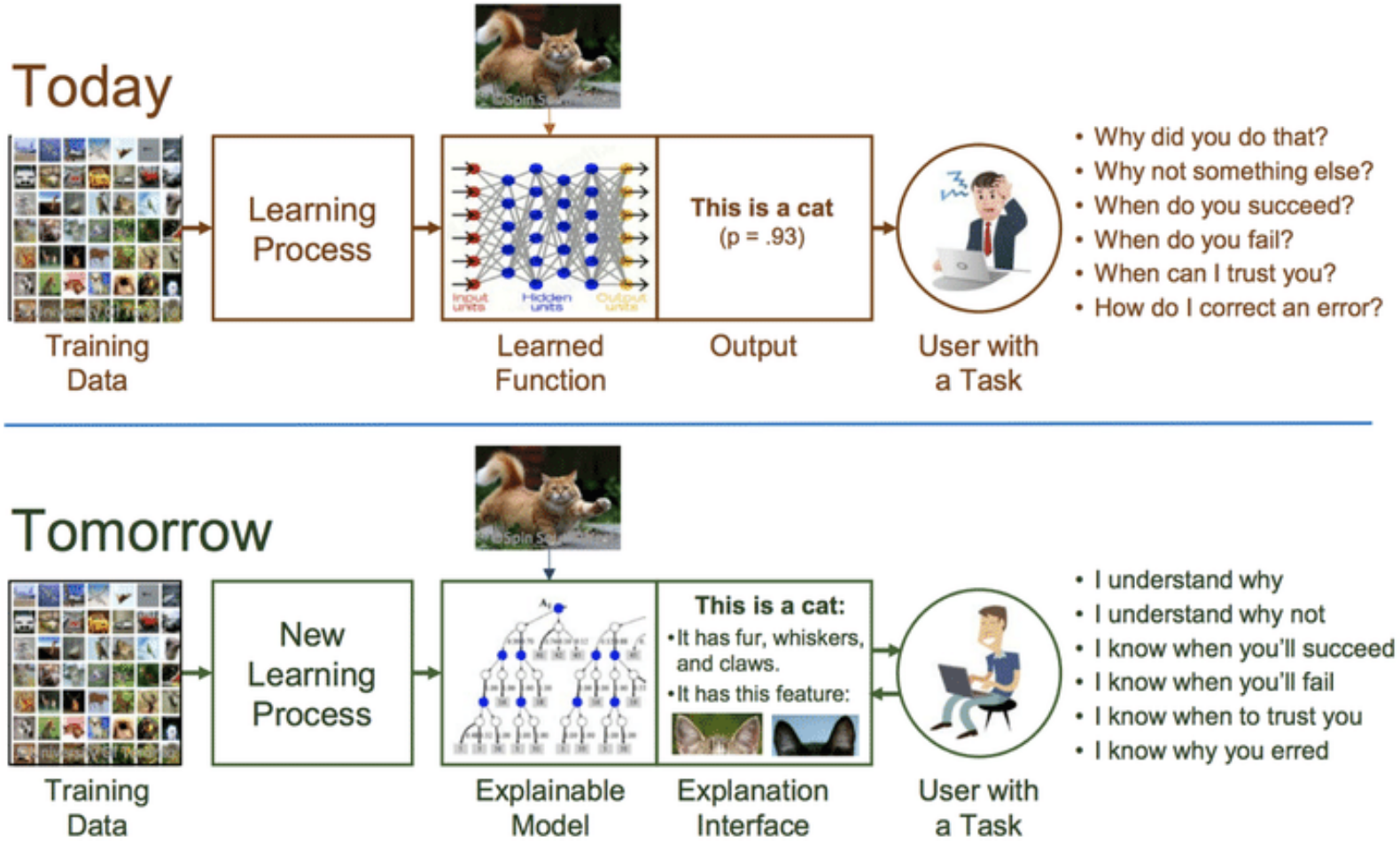
http://xai.unist.ac.kr/static/img/event/ICCV_2019_VXAI_Samek_Talk.pdf

Das ist eine **Reiterin mit Pferd**. Man sieht es an dem **Pferd auf dem sie sitzt**.

http://xai.unist.ac.kr/static/img/event/ICCV_2019_VXAI_Samek_Talk.pdf

Wie AI ihre Entscheidungen begründen

XAI heute und XAI bzw. AI als Wunsch in der Zukunft



https://www.researchgate.net/profile/Feiyu_Xu/publication/336131051/figure/fig12/AS:850956124180484@1579895158548/Explainable-AI-XAI-Concept-presented-by-DARPA-Image-courtesy-of-DARPA-XAI-Program.ppm

XAI als Trust Building in Predictive Maintenance

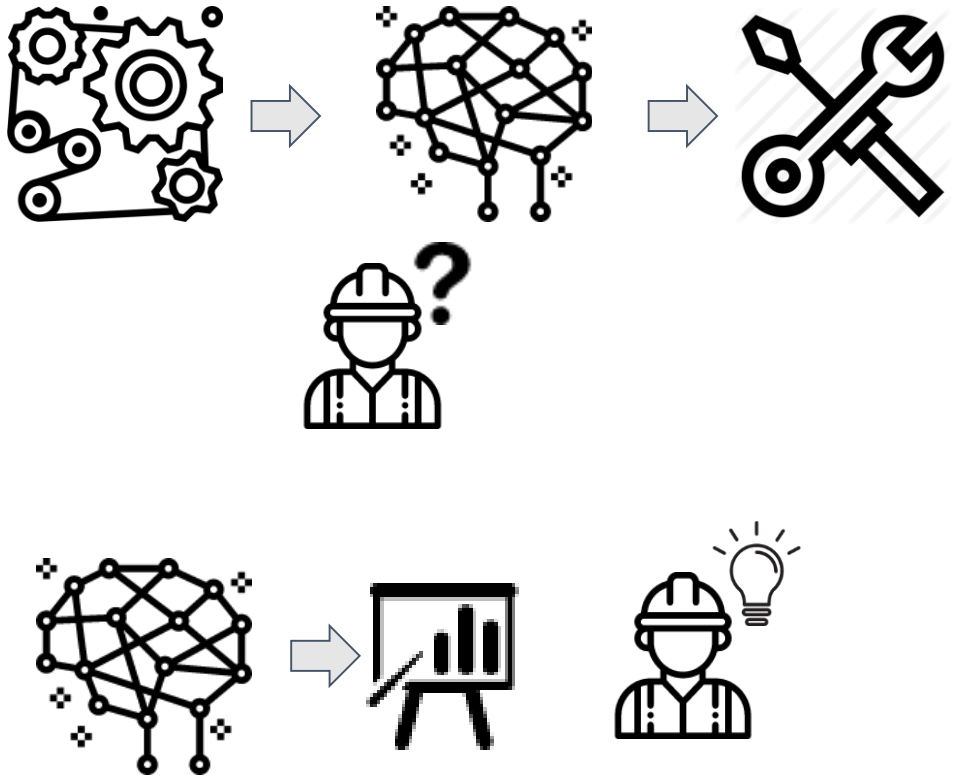
Predictive Maintenance als Beispiel für XAI

Predictive Maintenance um **vorherzusagen wann ein Motor gewartet werden muss**

Problem: **Wartungsarbeiten glauben der Vorhersage nicht**, da der Entscheidungsprozess nicht klar ist

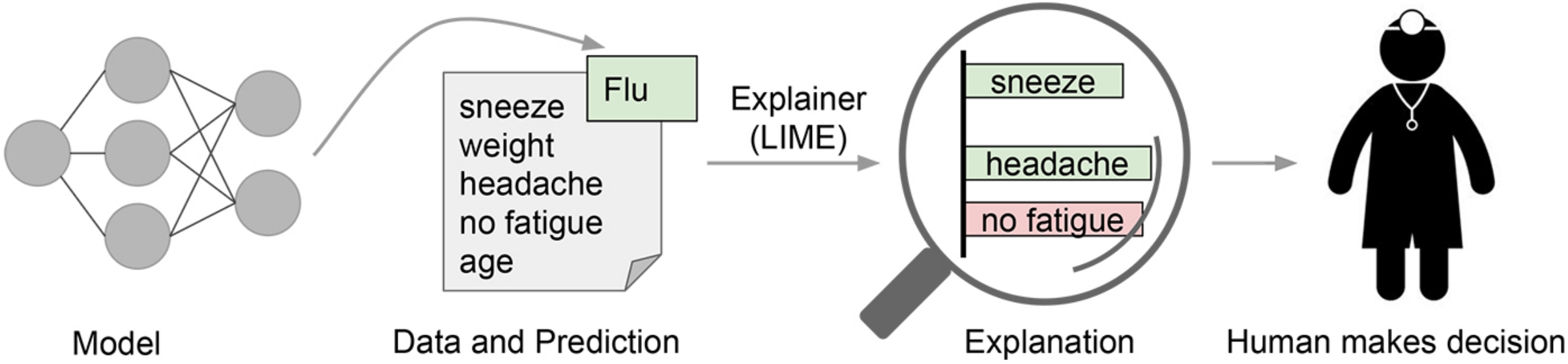
Lösung: XAI Methoden um zu zeigen **auf welcher Basis die Entscheidung getroffen wurde**

Ziel: **Wartungsarbeiten von der Vorhersage überzeugen**



XAI als Mittel zur Erklärbarkeit

XAI als Technik um Black Boxen zu erklären



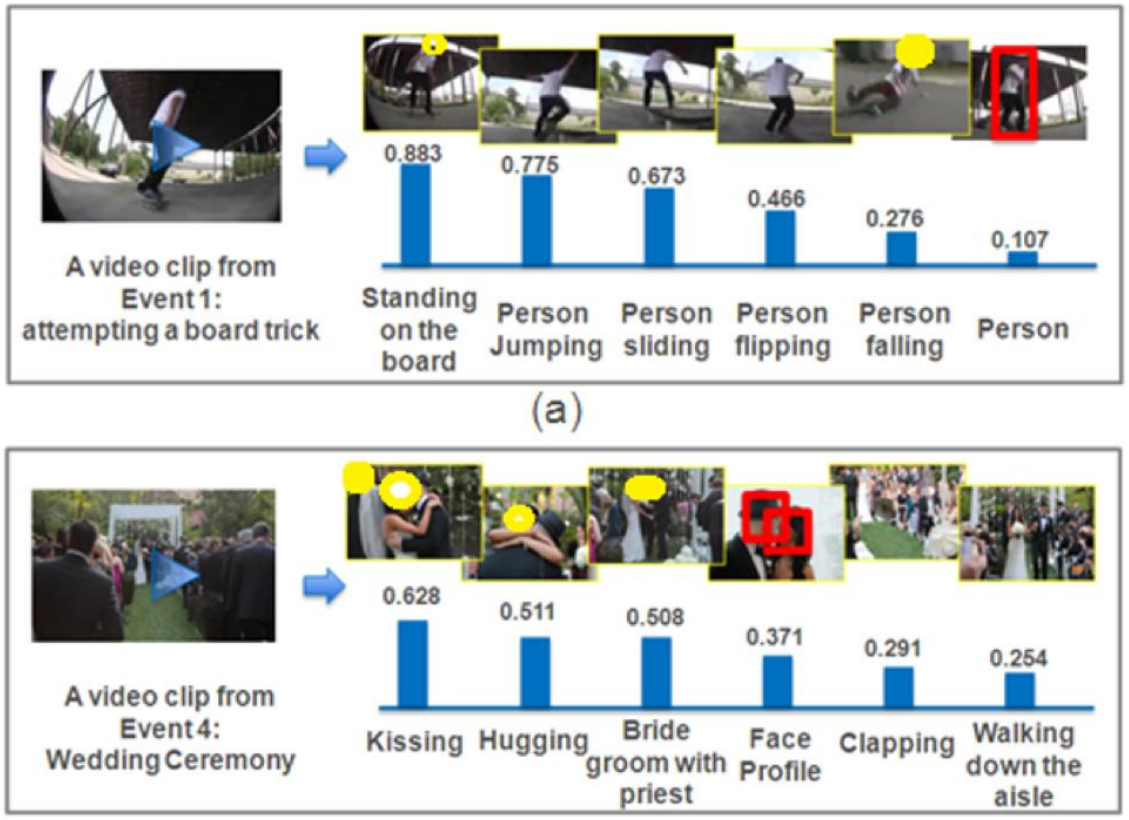
https://d3ansictanv2wj.cloudfront.net/figure1-a9533a3fb9bb9ace6ee96b4cdc9b6bcb.jpg

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

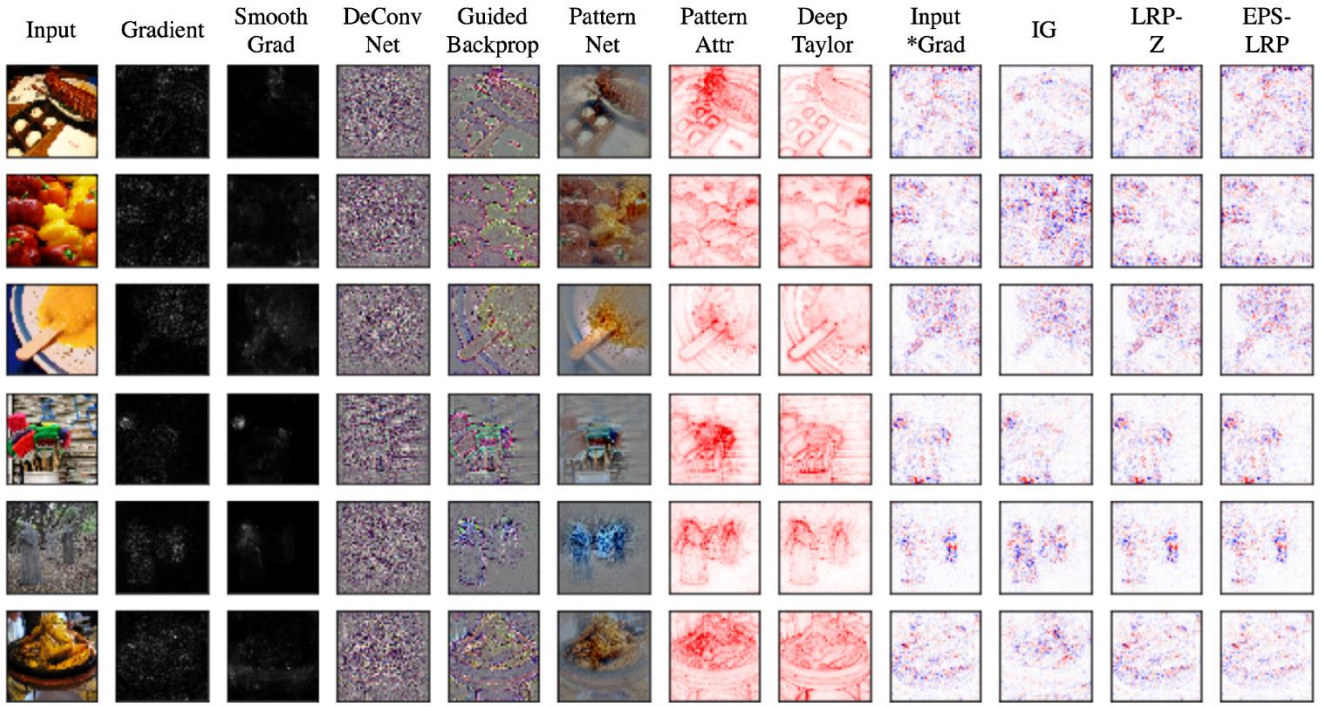
Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

XAI State-of-the-art

XAIs derzeitige Lösung um die Entscheidungen von Neural Networks zu verstehen – Feature Attributions and Heatmaps



https://miro.medium.com/max/2800/1*IFvEbOTh2kyM-kfmXFEL2Q.png

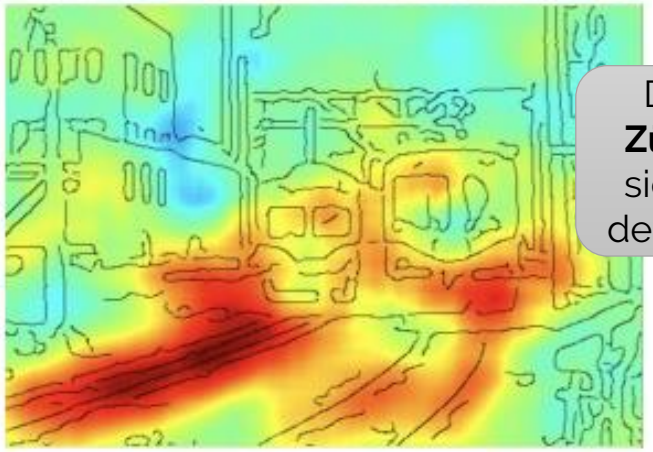


https://www.mdpi.com/jimaging/jimaging-06-00052/article_deploy/html/images/jimaging-06-00052-g003.png

Universität Konstanz

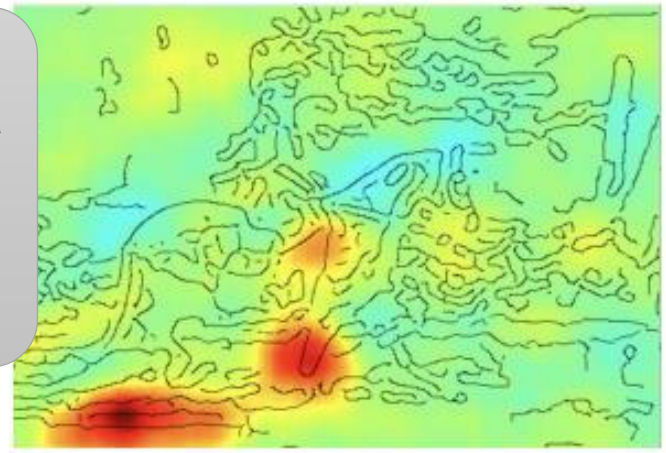
Wie XAI ihre Entscheidungen begründen

Heatmaps helfen AI Systeme zu verstehen – Jedoch lernen diese oft nicht das was gelernt werden soll



http://xai.unist.ac.kr/static/img/event/ICCV_2019_VXAI_Samek_Talk.pdf

Das sind **Züge**. Man sieht es an den **Gleisen**.



http://xai.unist.ac.kr/static/img/event/ICCV_2019_VXAI_Samek_Talk.pdf

Das ist eine **Reiterin mit Pferd**. Man sieht es an der **Caption unten links**.

Der Mensch als Vorbild für XAI (Zukunft)

Die Konversation mit der AI als System der Erklärbarkeit für die Zukunft

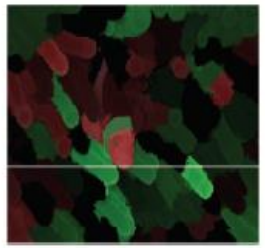
For illustration, the questions and answers are shown in English language text, but our use of a 'dialog' is for illustration only. An interactive GUI, for example, building on the ideas of Krause et al.,[20] would likely be a better realization.

1

ML Classifier

C: I predict FISH

2 H: Why?
C: See below:

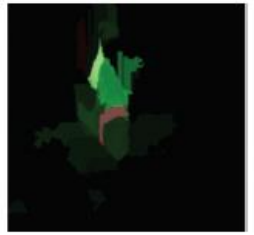


Green regions argue for FISH, while RED pushes toward DOG. There's more green.

3 H: (Hmm. Seems like it might be just recognizing anemone texture!) Which training examples are most influential to the prediction?
C: These ones:

4 H: What happens if the background anemones are removed? E.g.,

C: I still predict FISH, because of these green superpixels:



https://dl.acm.org/cms/attachment/61cab0eb-e769-4a90-8ff9-3920d54f963c/f7.jpg

Take Home Message

- XAI kann helfen **Fairness, Bias, Privacy, Robustness, Trust** und vieles weitere zu ermöglichen

- XAI ist jedoch noch in den **Kinderschuhen** (erste Paper 2004) Erklärungen über Heatmaps oder statistischen Werten

- XAI wird in **Zukunft immer wichtiger** (DSGVO, EU AI Law, DARPA US Research)

End

Udo Schlegel

- 2012 – 2016 B.Sc. Information Engineering
  - Interactive Support for Malware Classification
- 2016 – 2018 M.Sc. Computer and Information Science
  - Towards Crime Forecasting Using Deep Learning
- 2018 – 2022 Ph.D. Computer Science
  - Explainable AI for Time Series

Forschungsgebiete

- Explainable AI
- Visual Analytics
- Deep Learning
- Visualization
- Crime Forecasting (LKA NRW)
- Predictive Maintenance (Siemens)
- Dynamic Networks (Collective Behavior Cluster)

Kontakt

- https://www.vis.uni-konstanz.de/mitglieder/schlegel/
- https://www.linkedin.com/in/udo-schlegel/
- http://merowech.github.io/
- u.schlegel@uni-konstanz.de

Datenanalyse
und Visualisierung

Universitätsstraße 10
Universität Konstanz

Gebäude D, Raum 215

Postfach 78

Tel.: +49 (0) 7531 88-3583
Fax: +49 (0) 7531 88-3268

Backup Slides

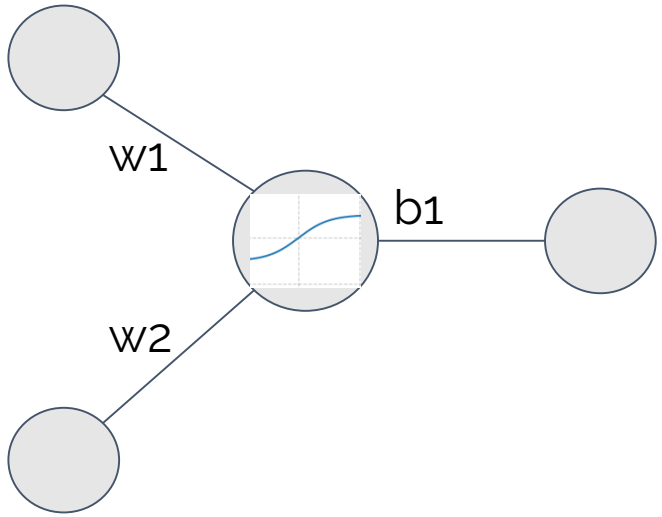
Was ist eine künstliche Synapse? (Perceptron)

Nachbau einer Synapse (Perceptron)

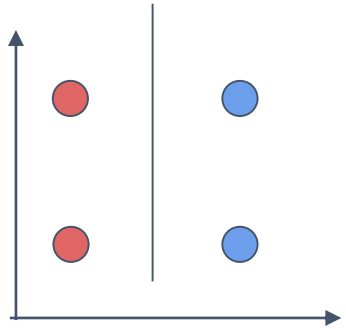
Zwei Eingänge

Eine Aktivierungsfunktion mit Schwellenwert

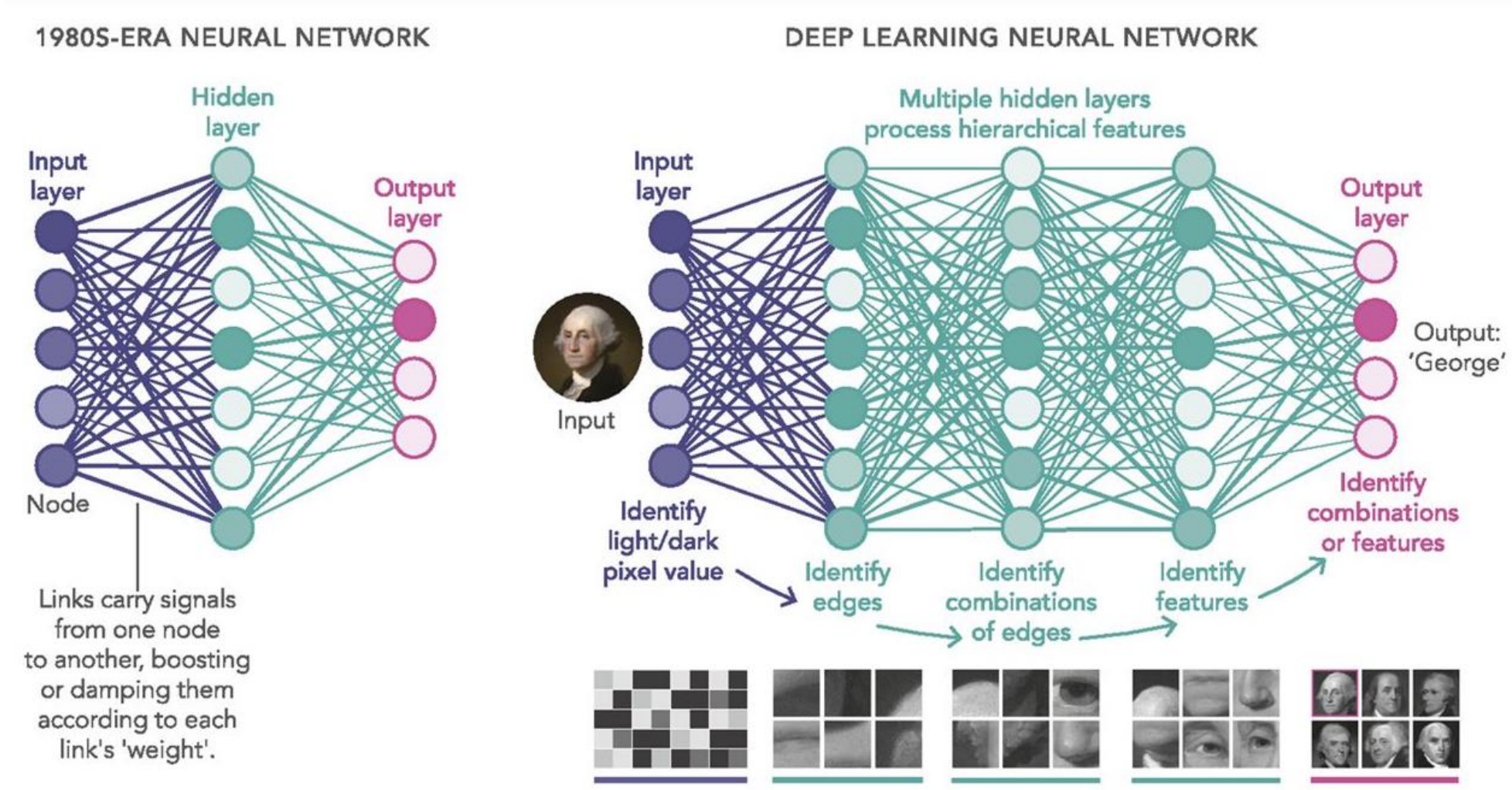
Ein Ausgang



Technik zum linearen Separieren von Daten



Deep Neural Networks



https://www.pnas.org/content/pnas/116/4/1074/F2.large.jpg