Zürcher Hochschule
für Angewandte Wissenschaften

Was kann Künstliche Intelligenz leisten?

CAS Digitale Technologien und Innovation, 13. September 2019

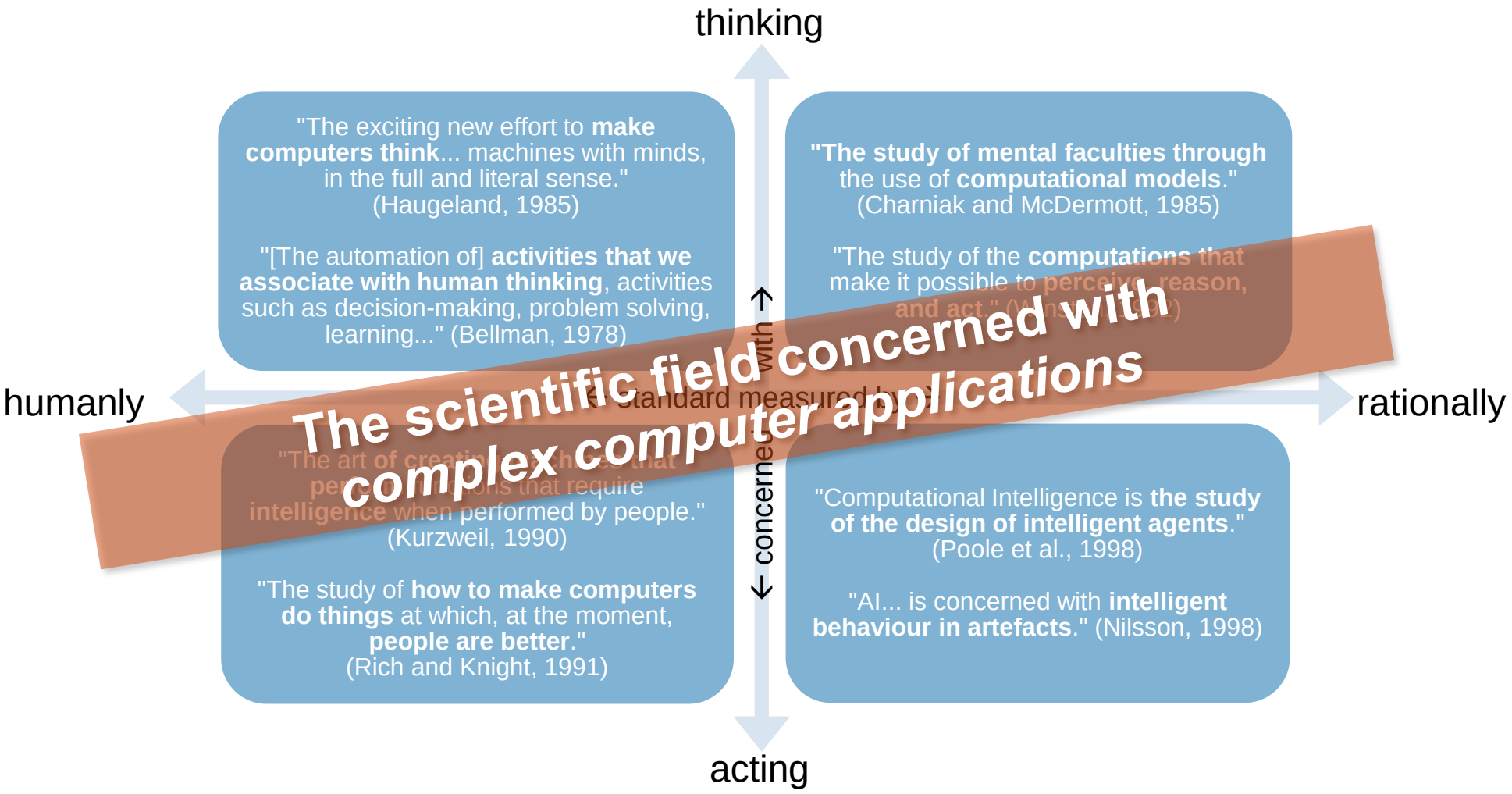
Thilo Stadelmann

Zürcher Hochschule
für Angewandte Wissenschaften

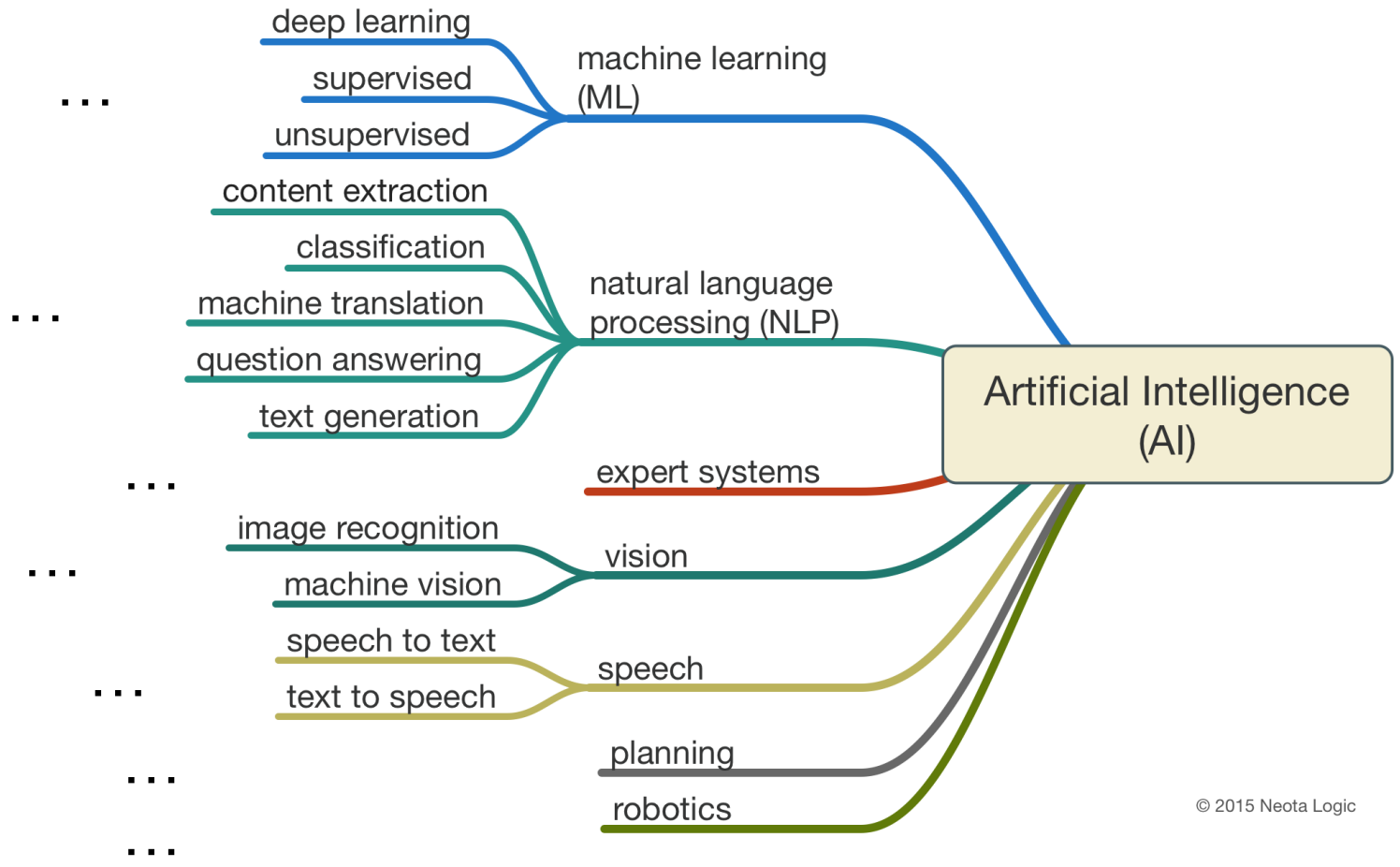
1

Was ist Künstliche Intelligenz?

Was ist künstliche Intelligenz?



Was gehört zu künstlicher Intelligenz?



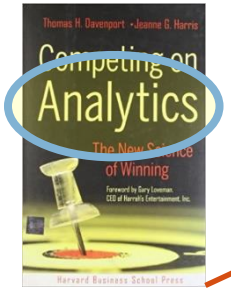
KI im Kontext

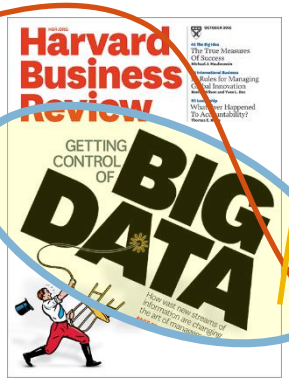
Turing's "Computing Machinery and Intelligence"
Dartmouth meeting: "Artificial Intelligence" adopted
Neural networks almost disappear: 1st "AI Winter"
Expert systems industry booms
Neural networks return to popularity
Expert systems industry busts: 2nd "AI Winter"
Human-Level AI back on agenda

1950 1960 1970 1980 1990 2000

Zürcher Hochschule für Angewandte Wissenschaften

zh aw





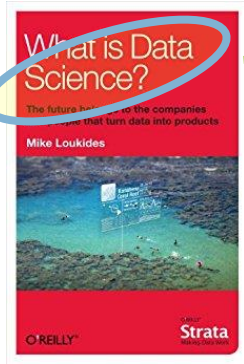
DIGITALisierung ZURICH 2025

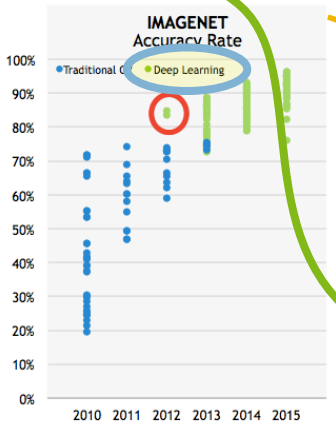
INDUSTRY 4.0 2025 INDUSTRIE INDUSTRIA

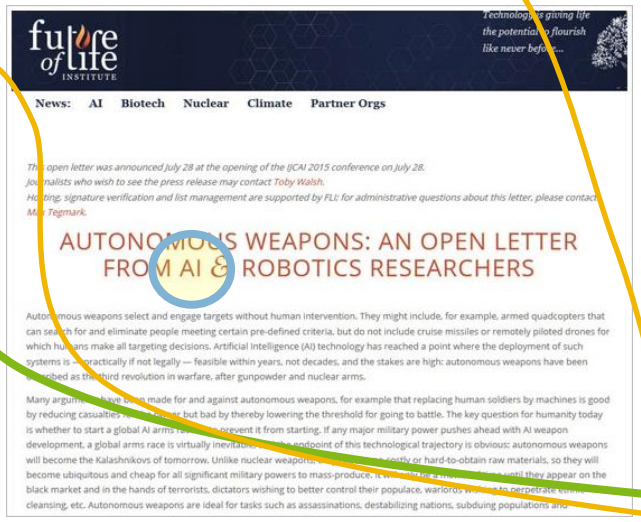
SWISS FINTECH INNOVATIONS

...

2007 2012 2016





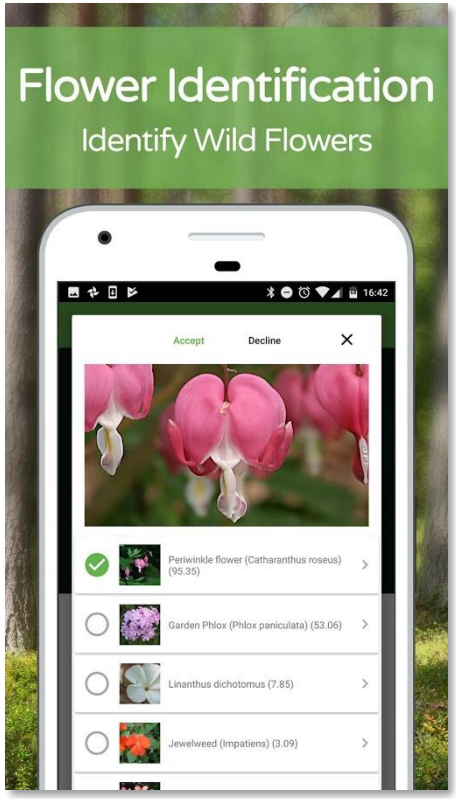


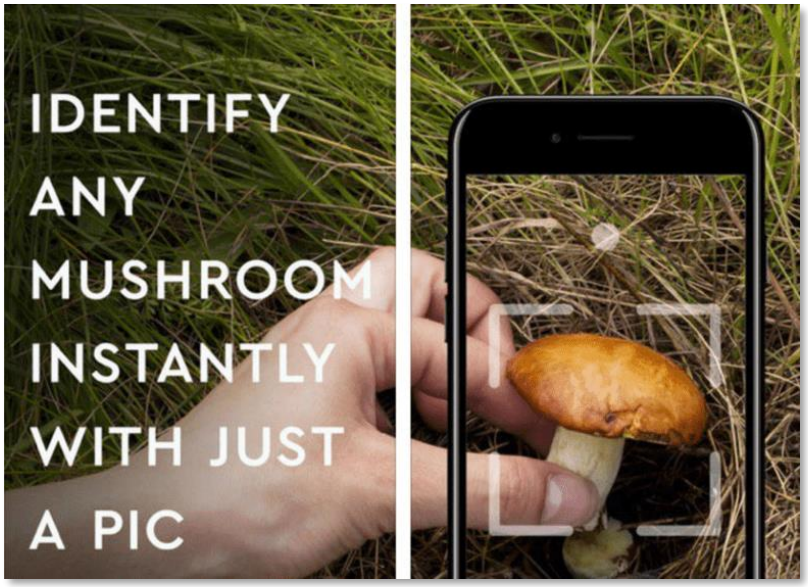
Was kann KI bereits heute?

1.	Play a decent game of **table tennis**	ok
2.	**Drive** safely along a curving **mountain road**	ok
3.	Drive safely along **Technikumstrasse** Winterthur	ok (only since recently)
4.	**Buy** a week's worth of **groceries on the web**	ok
5.	Buy a week's worth of groceries **at Migros**	no
6.	**Play** a decent game of **bridge**	ok
7.	**Discover** and prove a new mathematical **theorem**	not complet
8.	**Design** and execute a **research program** in molecular biology	not complet
9.	Write an **intentionally funny** story	no
10.	Give competent **legal advice** in a specialized area of law	ok
11.	**Translate** spoken English **into spoken** Swedish in real time	ok
12.	**Converse** successfully with another person for an hour	no
13.	Perform a complex **surgical operation**	not complet
14.	**Unload** any **dishwasher** and put everything away	no
15.	Compete in the game show **Jeopardy!**	ok
16.	**Write clickbait** articles fully automatized	ok

Beispiel: Machbar vs. gefährlich

Technologie: Computer Vision mit Deep Learning

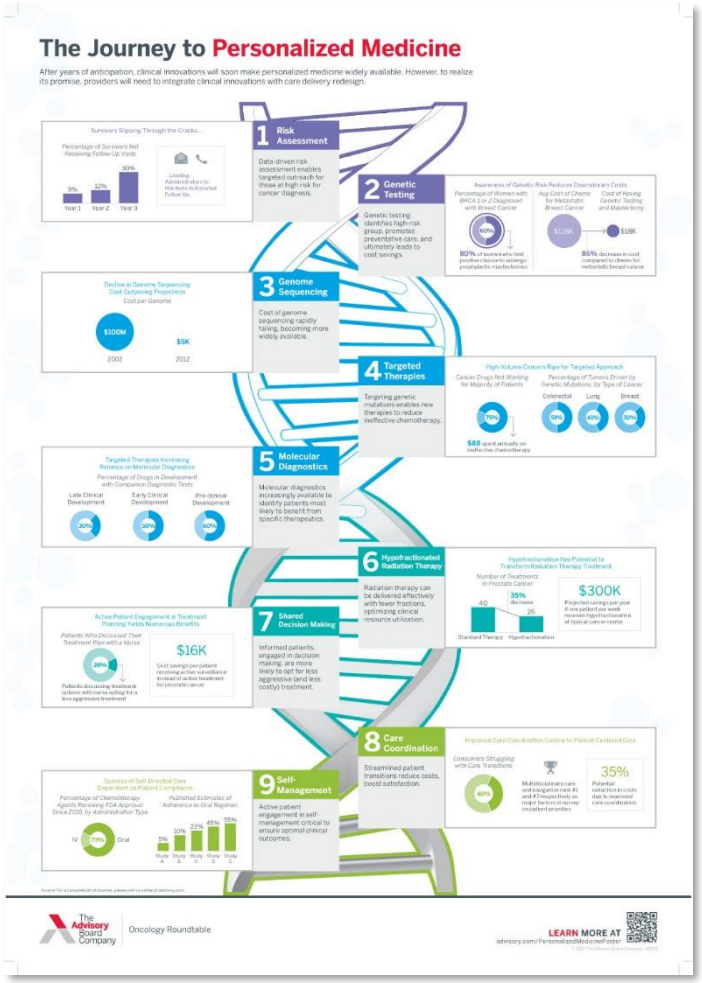




https://www.cultofmac.com/495088/avoid-potentially-deadly-ai-app/

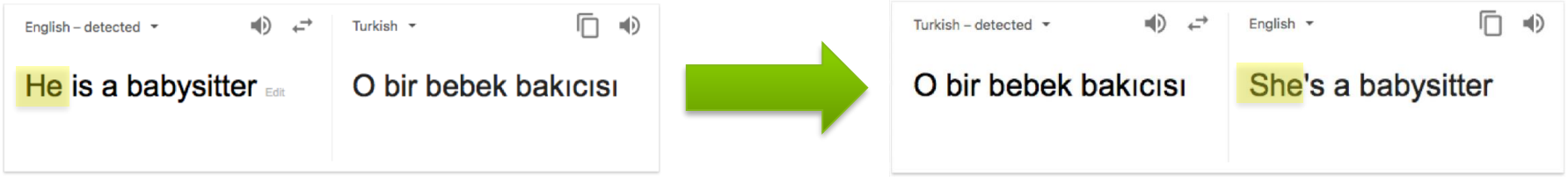
Beispiel: Markterfolg vs. regulatorische Hürden

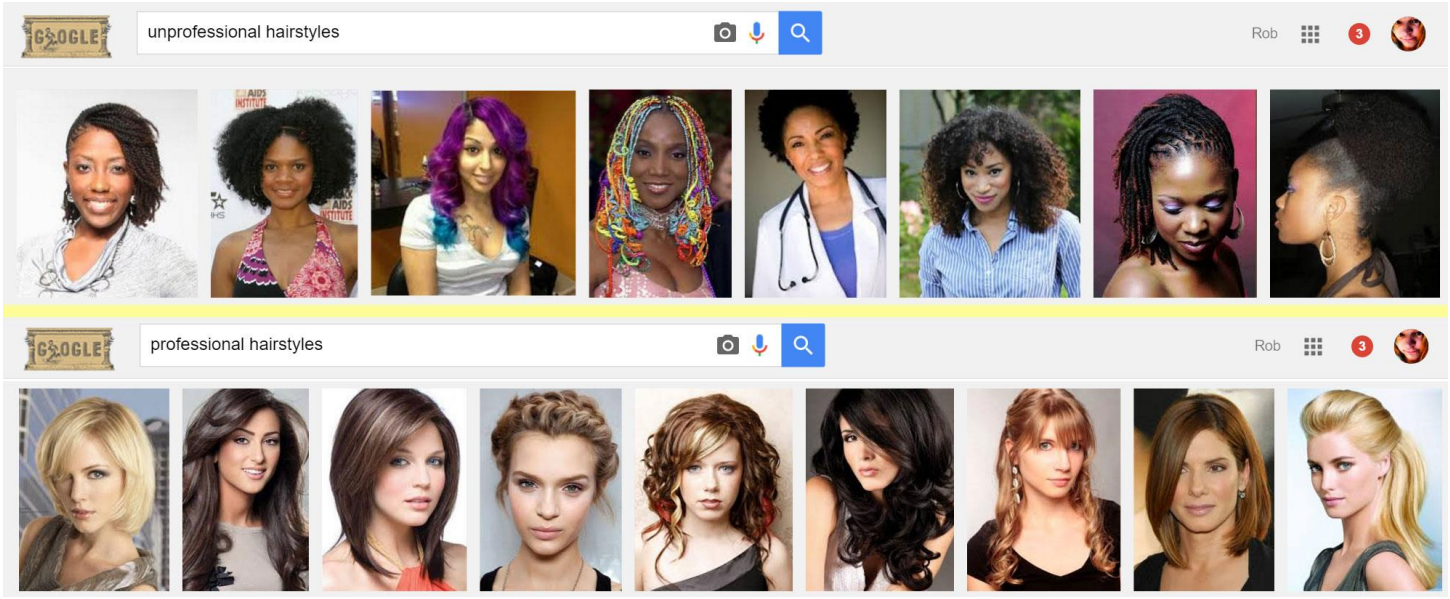
Technologie: Recommender Systems



Beispiel: Statistik vs. Bias

Technologie: Machine Learning

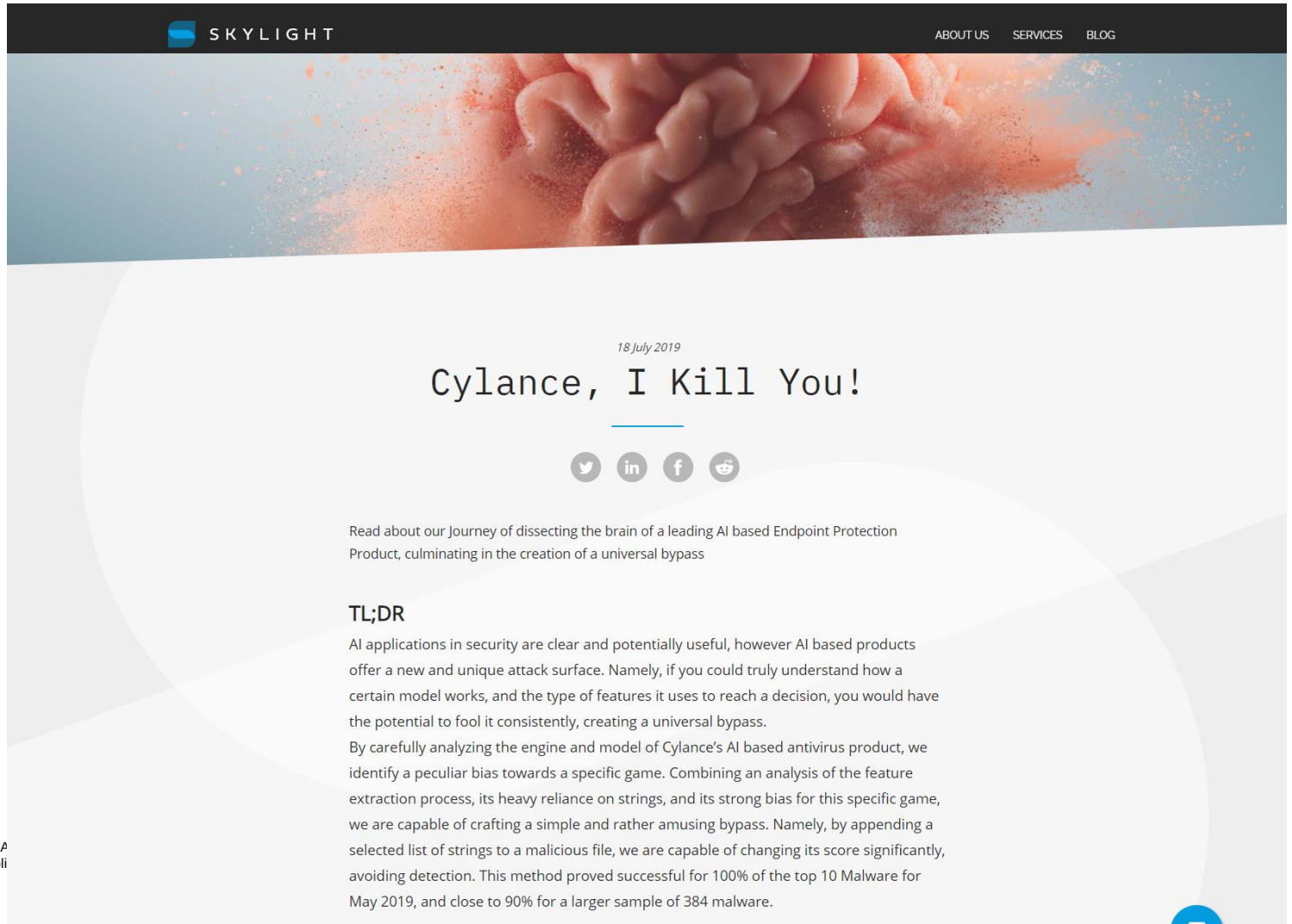




See also: Nassim Nicholas Talib, *«The Black Swan: The Impact of the Highly Improbable»*, 2007

Beispiel: künstl. Intelligenz vs. natürl. Dummheit

Technologie: Machine Learning mit nachgelagerten Regeln



SKYLIGHT

ABOUT US SERVICES BLOG

18 July 2019

Cylance, I Kill You!

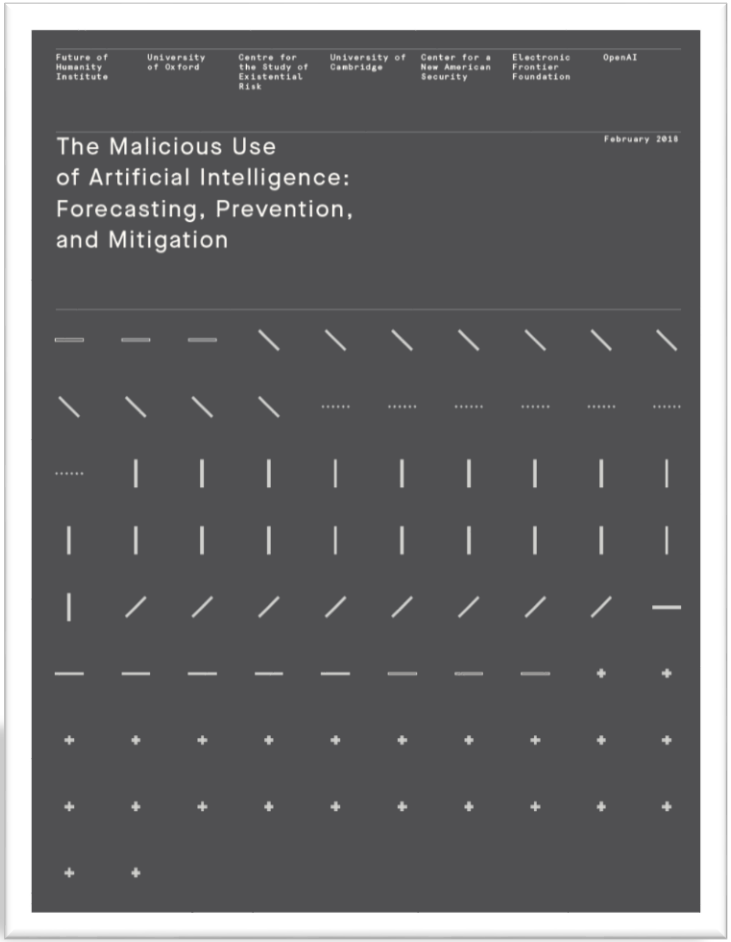
Read about our Journey of dissecting the brain of a leading AI based Endpoint Protection Product, culminating in the creation of a universal bypass

TL;DR

AI applications in security are clear and potentially useful, however AI based products offer a new and unique attack surface. Namely, if you could truly understand how a certain model works, and the type of features it uses to reach a decision, you would have the potential to fool it consistently, creating a universal bypass.
By carefully analyzing the engine and model of Cylance's AI based antivirus product, we identify a peculiar bias towards a specific game. Combining an analysis of the feature extraction process, its heavy reliance on strings, and its strong bias for this specific game, we are capable of crafting a simple and rather amusing bypass. Namely, by appending a selected list of strings to a malicious file, we are capable of changing its score significantly, avoiding detection. This method proved successful for 100% of the top 10 Malware for May 2019, and close to 90% for a larger sample of 384 malware.

Gefahren durch KI?

- KI ist per Definition eine “**dual use Technology**”
  → siehe Report von Brundage et al., 2018
- Aber: “**natürliche Dummheit**” ist die grössere Bedrohung
- **Algorithmische Ethik** und **erklärbare KI** sind in den letzten Jahren zu einem top Forschungsfeld geworden – nicht wegen der unkalkulierbaren Risiken per se, sondern:



Zürcher Hochschule
für Angewandte Wissenschaften

2

Warum ist das jetzt aktuell?
(Eine kurze Geschichte der letzten Jahre)

Google Acquires Artificial Intelligence Startup DeepMind For More Than $500M

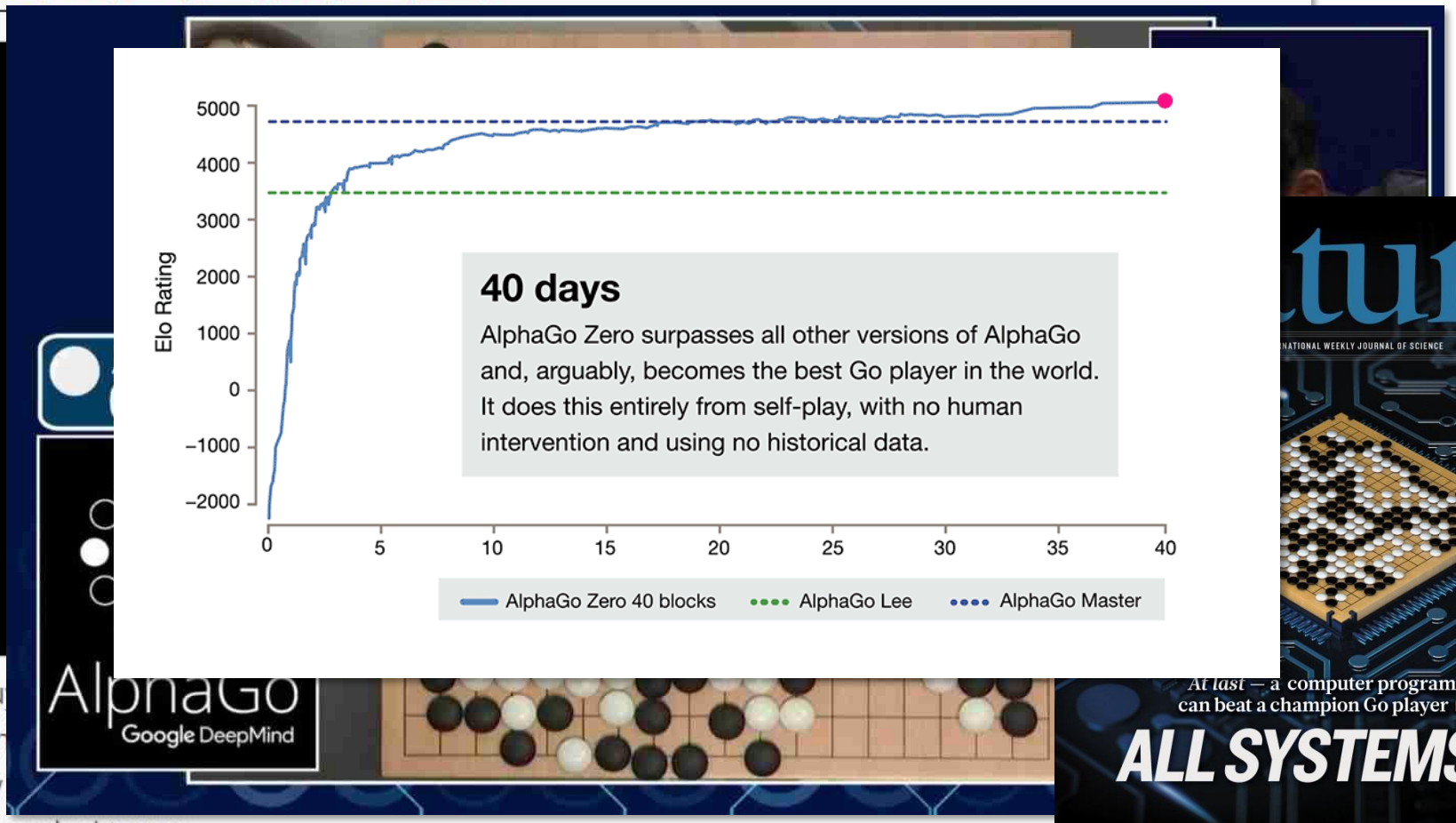
Posted Jan 26, 2014 by Catherine Shu (@catherineshu)

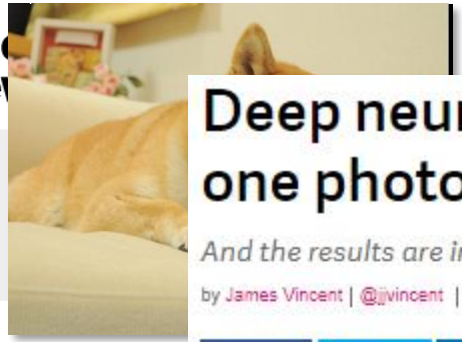
Google will bu
reports that th
in talks to buy
couldn't disclose deal terms.

The acquisition was originally confirmed by Google to Re/code.

Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw





Deep neural networks can now transfer the style of one photo onto another

And the results are impressive

by James Vincent | @jjvincent | Mar 30, 2017, 1:53pm EDT

You've probably heard of an AI technique known as "style transfer" — or, if you haven't heard of it, you've seen it. The process uses neural networks to apply the look and feel of one image to another, and appears in apps like Prisma and Facebook. These style transfers, however, are stylistic, not photorealistic. They look good because they look like they've been painted. Now a group of researchers from Cornell University and Adobe have augmented

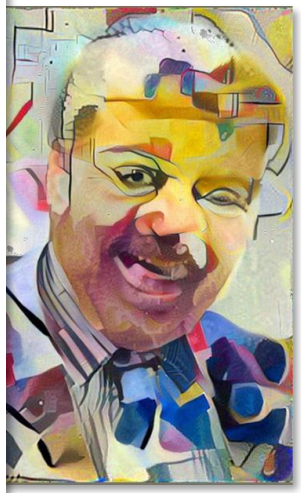
Computing

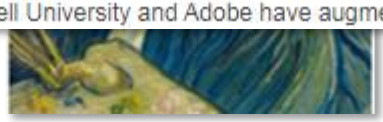
Algorith
Artistic
Other In

A deep neural n
other images.

by Emerging Tech

The nature of arti
of Vincent Van G
Edvard Munch's
humans recogniz



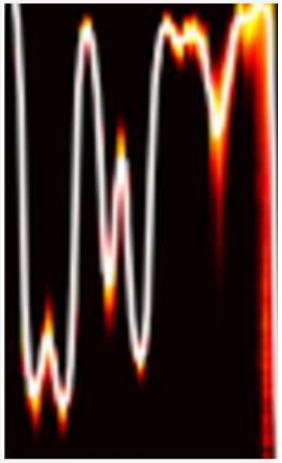


Künstliche Intelligenz

WaveNet lässt Computersprache natürlich klingen

von Henning Steier / 12.9.201

Die Google-Tochter DeepM
macht auch Musik.



DeepMind lässt WaveNet Spra

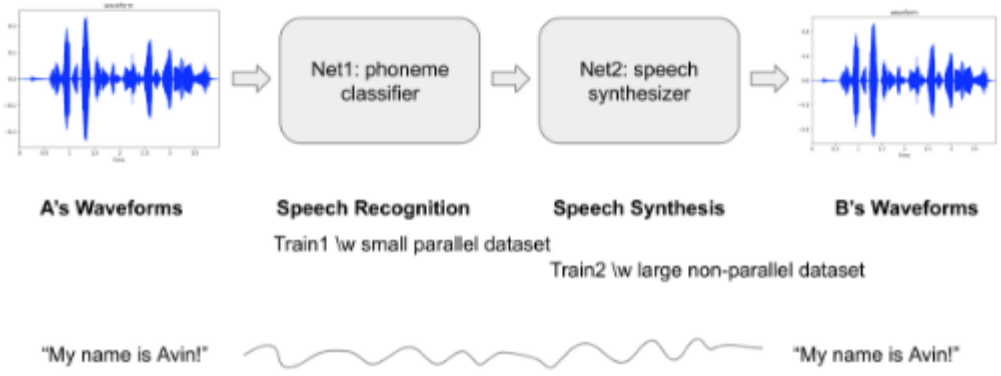
Die Google-Tochter Deep
Spiel «Go» Schlagzeilen:
einen der besten mensch
Londoner Unternehmen
erzeugt Sprache, die seh
im Blogeintrag des Unter
Massstab nimmt. Man ha

Intro

What if you could imitate a famous celebrity's voice or sing like a famous singer? This project started with a goal to convert someone's voice to a specific target voice. So called, it's voice style transfer. We worked on this project that aims to convert someone's voice to a famous English actress Kate Winslet's voice. We implemented a deep neural networks to achieve that and more than 2 hours of audio book sentences read by Kate Winslet are used as a dataset.

Model Architecture

This is a many-to-one voice conversion system. The main significance of this work is that we could generate a target speaker's utterances without parallel data like <source's wav, target's wav>, <wav, text> or <wav, phone>, but only waveforms of the target speaker. (To make these parallel datasets needs a lot of effort.) All we need in this project is a number of waveforms of the target speaker's utterances and only a small set of <wav, phone> pairs from a number of anonymous speakers.



Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

nerierte Sprache
s Texteingabe»

nerierte Musik
ne Inhaltsvorgabe»

1 Second

Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

…und die Liste liesse sich fortsetzen!

Brandon Amos About Blog

Image Completion with Deep Learning in TensorFlow

August 9, 2016

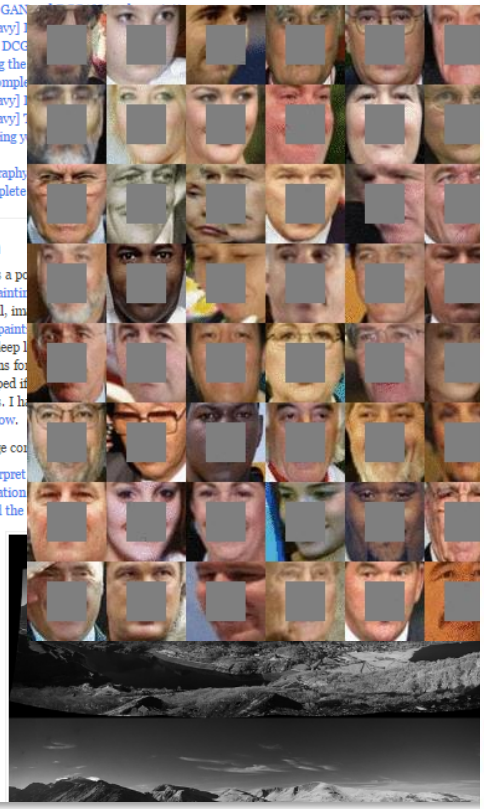
- Introduction
- Step 1: Interpreting images as samples from a probability distribution
  - How would you fill in the missing information?
  - But where does statistics fit in? These are images.
  - So how can we complete images?
- Step 2: Quickly generating fake images
  - Learning to generate new samples from an unknown probability distribution
  - [ML-Heavy] Generative Adversarial Net (GAN) building blocks
  - Using $G(z)$ to produce fake images
  - [ML-Heavy] Training DCGANs
  - Existing GAN
  - [ML-Heavy]
  - Running DC
- Step 3: Finding the
  - Image comple
  - [ML-Heavy]
  - [ML-Heavy]
  - Completing y
- Conclusion
- Partial bibliography
- Bonus: Incomplete

Introduction

Content-aware fill is a po
completion and inpaintin
do content-aware fill, im
"Semantic Image Inpaint
shows how to use deep l
some deeper portions fo
section can be skipped i
from images of faces. I h
completion.tensorflow.

We'll approach image co

1. We'll first interpret
2. This interpretation
3. Then we'll find the



Andrej Karpathy blog About Hacker's guide to Neural Networks

The Unreasonable Effectiveness of Recurrent Neural Networks

GEEK.COM

TECH

Nvidia AI Generates Fake Faces Based On Real Celebs

BY STEPHANIE MLOT 10.31.2017 :: 10:00AM EST

32 SHARES

I'm getting a distinctly mid-90s "The Rachel" vibe from the woman in the top left corner (via Nvidia)

STAY ON TARGET

AI Shelley Pens Truly Creepy Horror Stories–And You Can Help

Neural Network Serves Up Truly Frightening Halloween Costume Ideas

Celebrity scandals are about to get a lot more complicated.

Nvidia has **developed** a way of producing photo-quality, AI-generated human profiles—by using famous faces.

hand,

law,

ds,

the morning paper

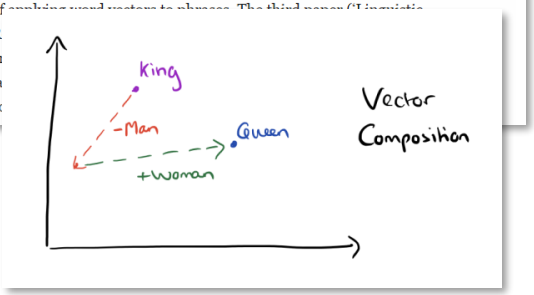
The amazing power of word vectors

APRIL 21, 2016

For today's post, I've drawn material not just from one paper, but from five! The subject matter is 'word2vec' – the work of Mikolov et al. at Google on efficient vector representations of words (and what you can do with them). The papers are:

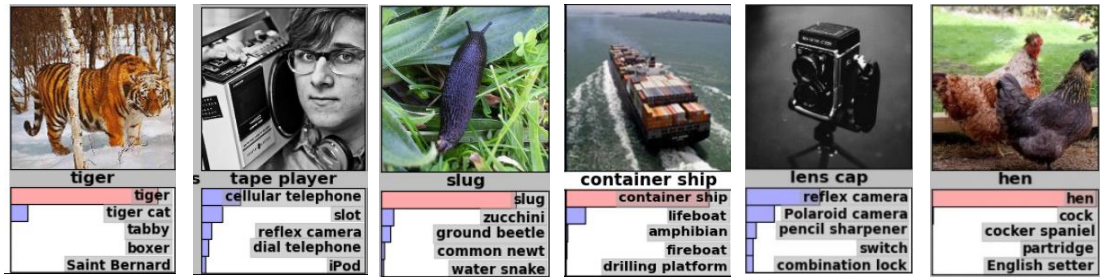
- **Efficient Estimation of Word Representations in Vector Space** – Mikolov et al. 2013
- **Distributed Representations of Words and Phrases and their Compositionality** – Mikolov et al. 2013
- **Linguistic Regularities in Continuous Space Word Representations** – Mikolov et al. 2013
- **word2vec Parameter Learning Explained** – Rong 2014
- **word2vec Explained: Deriving Mikolov et al's Negative Sampling Word-Embedding Method** – Goldberg and Levy 2014

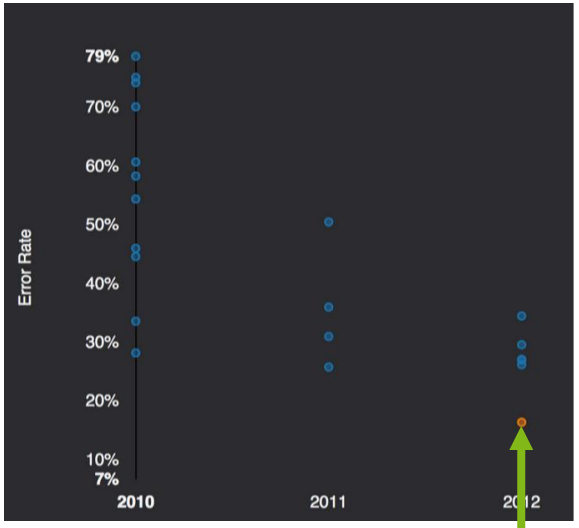
From the first of these papers ('Efficient estimation…') we get a description of the *Continuous Bag-of-Words* and *Continuous Skip-gram* models for learning word vectors (we'll talk about what a word vector is in a moment…). From the second paper we get more illustrations of the power of word vectors, some additional information on optimisations for the skip-gram model (hierarchical softmax and negative sampling), and a discussion



Was ist passiert?
Der ImageNet Wettbewerb

1000 Kategorien
1 Mio. Beispiele





A. Krizhevsky verwendet als erster ein sog. «Deep Neural Network» (CNN)

2015: Computer *haben "Sehen" gelernt*

4.95% Microsoft (06. Februar)
→ Besser als Menschen (5.10%)

4.80% Google (11. Februar)

4.58% Baidu (11. Mai)

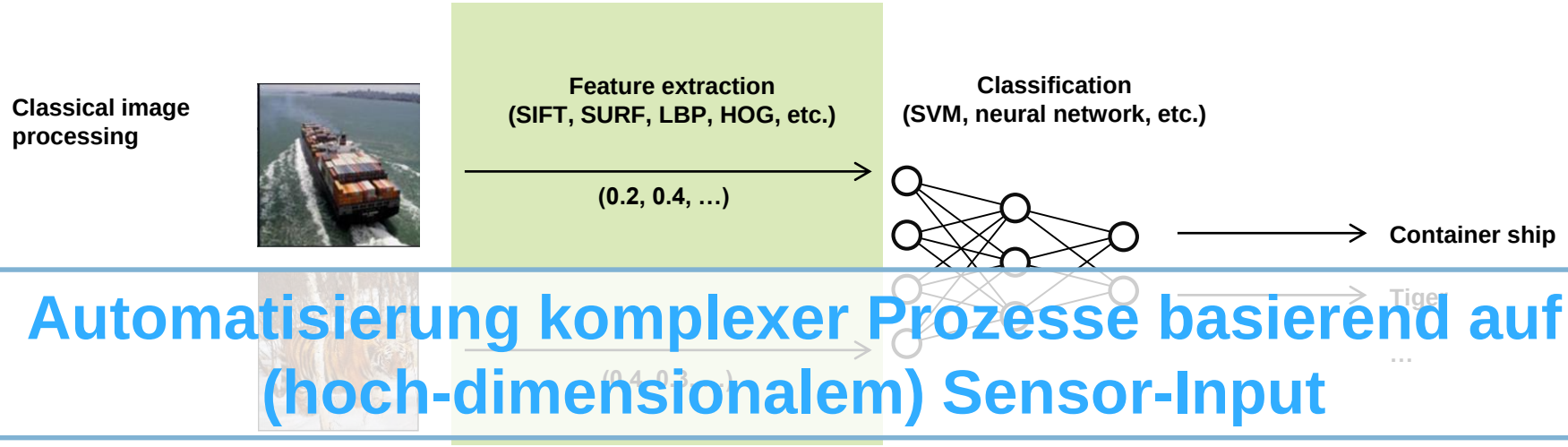
3.57% Microsoft (10. Dezember)

Zürcher Hochschule
für Angewandte Wissenschaften

3

Wie geht das?

Idee: Mehr «Tiefe» um Merkmale automatisch zu lernen



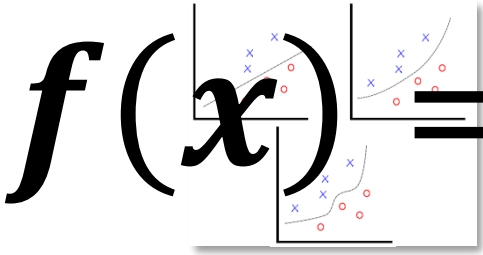
Automatisierung komplexer Prozesse basierend auf (hoch-dimensionalem) Sensor-Input

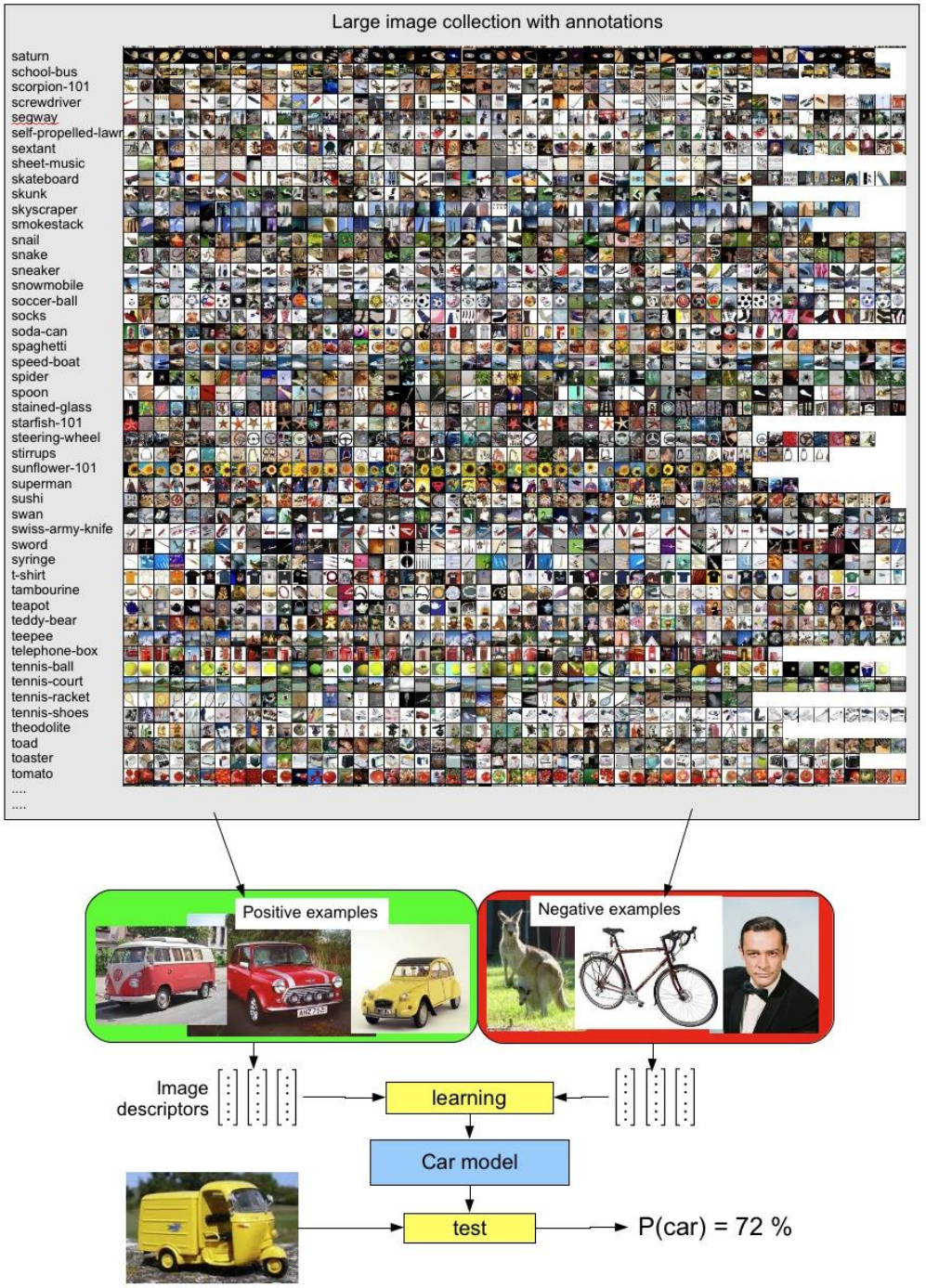
Grundlage
Induktives überwachtes Lernen

Annahme
- Ein an *genügend viele* Beispiele angepasstes Modell…
- …wird auch auf unbekannte Daten **generalisieren**

Methode
- **Suchen der Parameter einer gegebenen Funktion**…
- …so dass für alle Beispiele Eingabe (Bild) auf Ausgabe («Auto») abgebildet wird

$$f(\boldsymbol{x}) = y$$

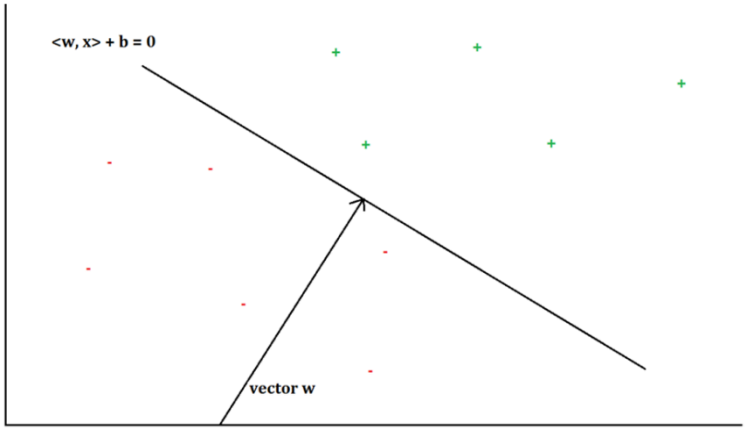


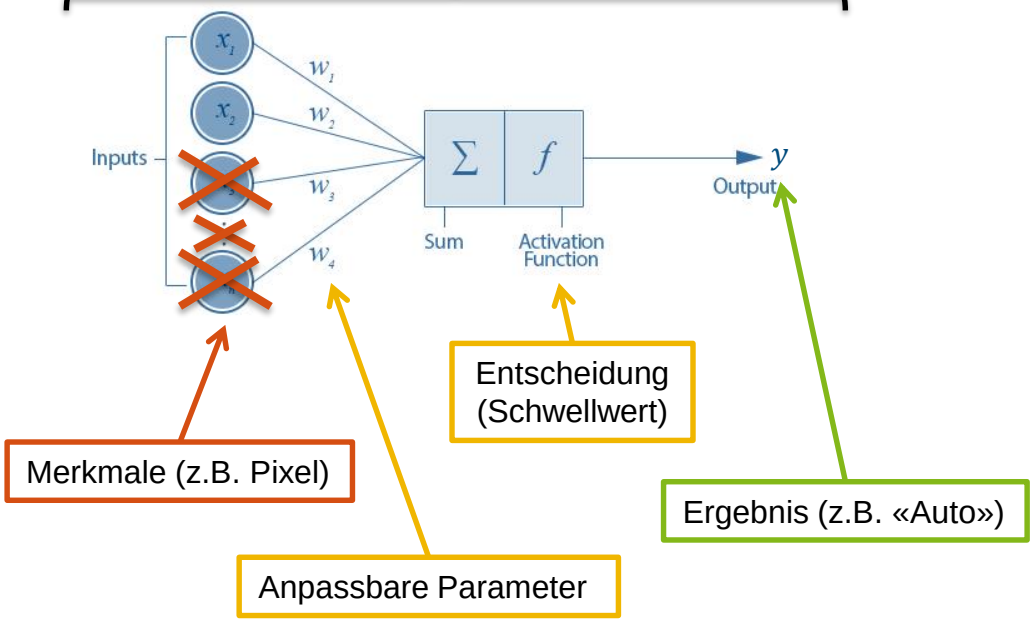


Quelle: http://lear.inrialpes.fr/job/postdoc-large-scale-classif-11-img/attribs_patchwork.jpg

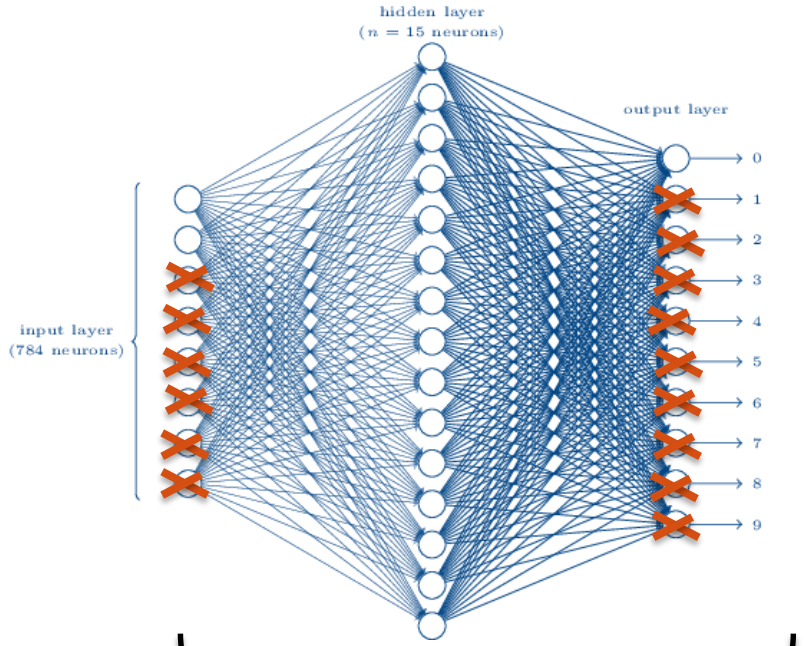
Suche der Parameter *einer Funktion*?

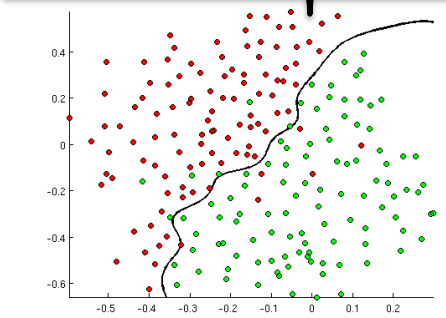
Neuron





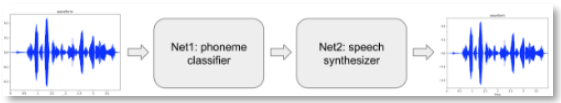
Neuronales Netz





Ausblick: Innovation aus neuronalen Netzen?

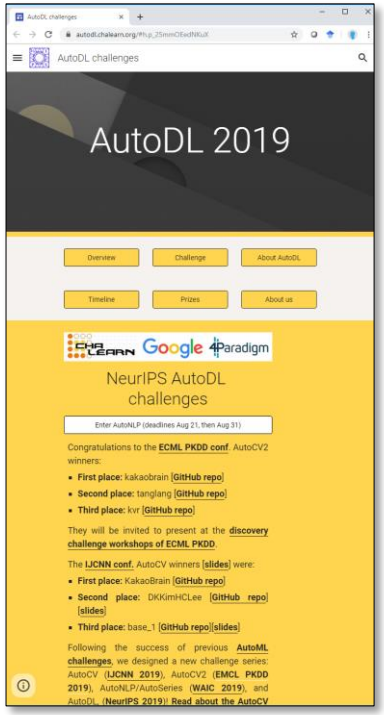


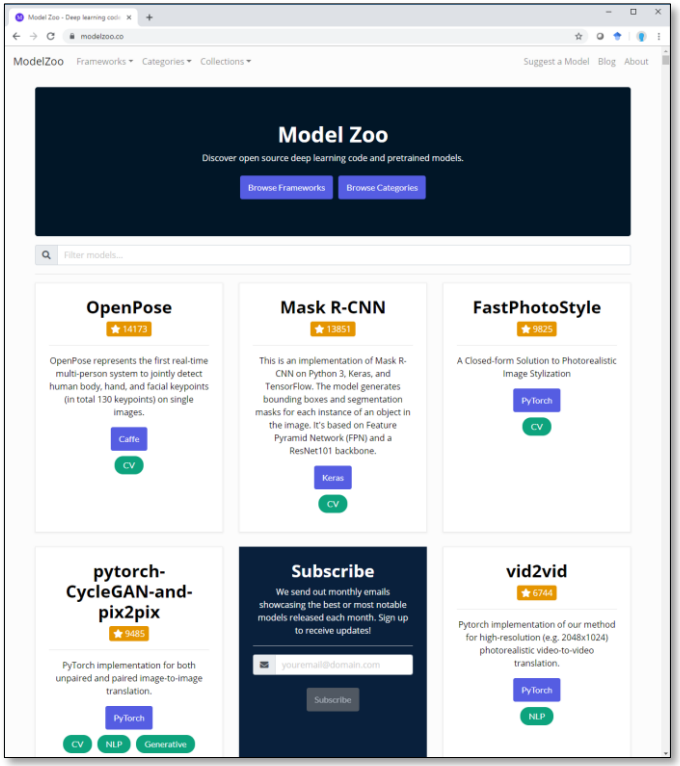


- *Deep Learning* ist der Kern vieler erstaunlicher Anwendungen im Bereich der Automatisierung von Wahrnehmung (Synthese & Analyse)
  ➔ Code & Datensätze laden “Maker” zum Ausprobieren ein

- *Automated Deep Learning* macht in Zukunft vielleicht manche Anwendung einfacher

- Aktuell braucht es ein gutes Verständnis der Details der Methode, um kreativ neue Szenarien zu erdenken (Stichwort “mTrainer” anstatt Programmierer?)





Schlussfolgerungen

- Deep Learning hat zu Paradigmenwechsel in *Mustererkennungsaufgaben* geführt
- Die Zeit vom Grundlagenresultat zur praktischer Anwendung beträgt wenige Monate
- Es gibt Methoden zum Hineinschauen in neuronale Black Boxes (siehe Anhang)
- Spezifische Aufgaben lassen sich sehr gut automatisieren (z.B. Ähnlichkeitssuche)

datalab
www.zhaw.ch/datalab

swiss group for artificial intelligence and cognitive science
sgaico

data service
Swiss Alliance for Data-Intensive Services

Zu mir:

- Prof. KI/ML, Scientific Director ZHAW digital
- Email: stdm@zhaw.ch
- Telefon: 058 934 72 08
- Web: https://stdm.github.io/
- Twitter: @thilo_on_data
- LinkedIn: thilo-stadelmann

Mehr zum Thema:

- Data+Service Alliance: www.data-service-alliance.ch
- KI: https://sgaico.swissinformatics.org/
- Zusammenarbeit: datalab@zhaw.ch

Zürcher Hochschule
für Angewandte Wissenschaften

ANHANG

Developing for algorithmic fairness

The FAT ML code of conduct

See http://www.fatml.org/resources/principles-for-accountable-algorithms

Purpose

- Help developers to **build algorithmic systems in publicly accountable ways**
- Accountability: the **obligation to report, explain, or justify** algorithmic decision-making & **mitigate** any **negative** social **impacts** or potential harms

Premise

- *A **human ultimately responsible** for decisions made/informed by an algorithm*

Principles

- **Responsibility**, **Explainability**, **Accuracy**, **Auditability**, **Fairness**

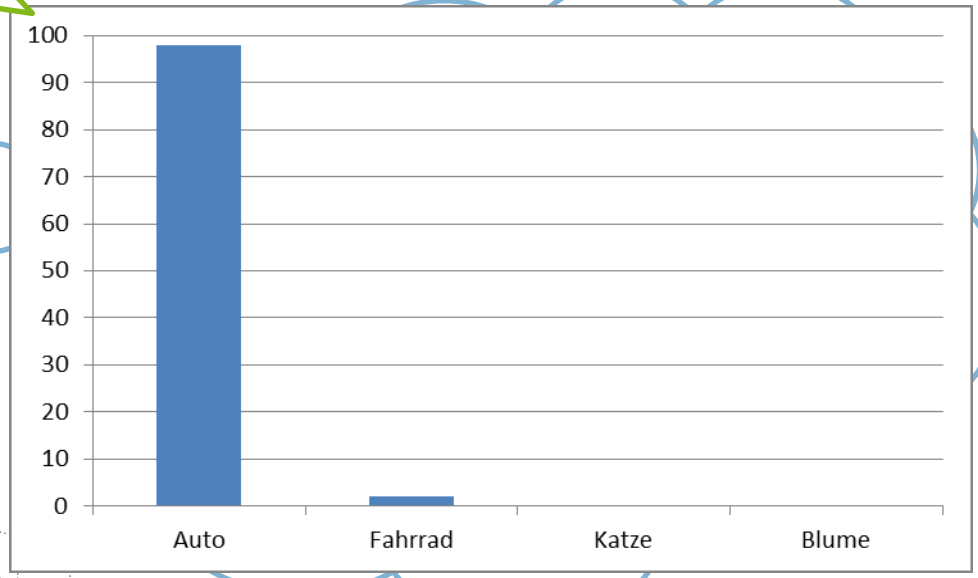
Make available somebody who will take care of adverse individual / societal effects	Explain any **algorithmic decision** in non-technical terms to end users	**Report** all **sources of uncertainty** / error in algorithms & data	Enable 3rd parties to **probe & understand** system **behavior**	Ensure algorithmic **decisions are not discriminatory** w.r.t. to people groups

Making it actionable

- **Publish** a **Social Impact Statement**
- …use above **principles as a guiding structure**
- …**revisit three times** during development process: design stage, pre-launch, post-launch

Suche der Parameter einer Funktion?

Wahrscheinlichkeit [%] für bestimmtes Ergebnis



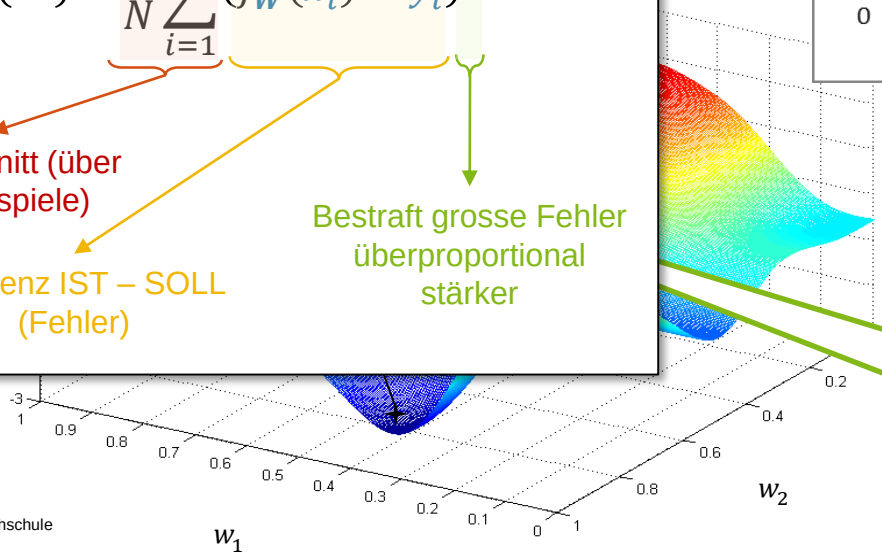
- Unser Neuronales Netz: $f_W(x) = y$
  mit Bild x, echtem Resultat y und Parametern W
  ($W = \{w_1, w_2, \ldots\}$ anfangs zufällig gewählt)
- Fehlermass: $l(W) = \frac{1}{N}\sum_{i=1}^{N}(f_W(x_i) - y_i)^2$
  Durchschnitt der quadratischen Abweichungen
  über alle Bilder (Loss)

$$l(W) = \frac{1}{N}\sum_{i=1}^{N}(f_W(x_i) - y_i)^2$$

Durchschnitt (über alle Beispiele)

Differenz IST – SOLL (Fehler)

Bestraft grosse Fehler überproportional stärker

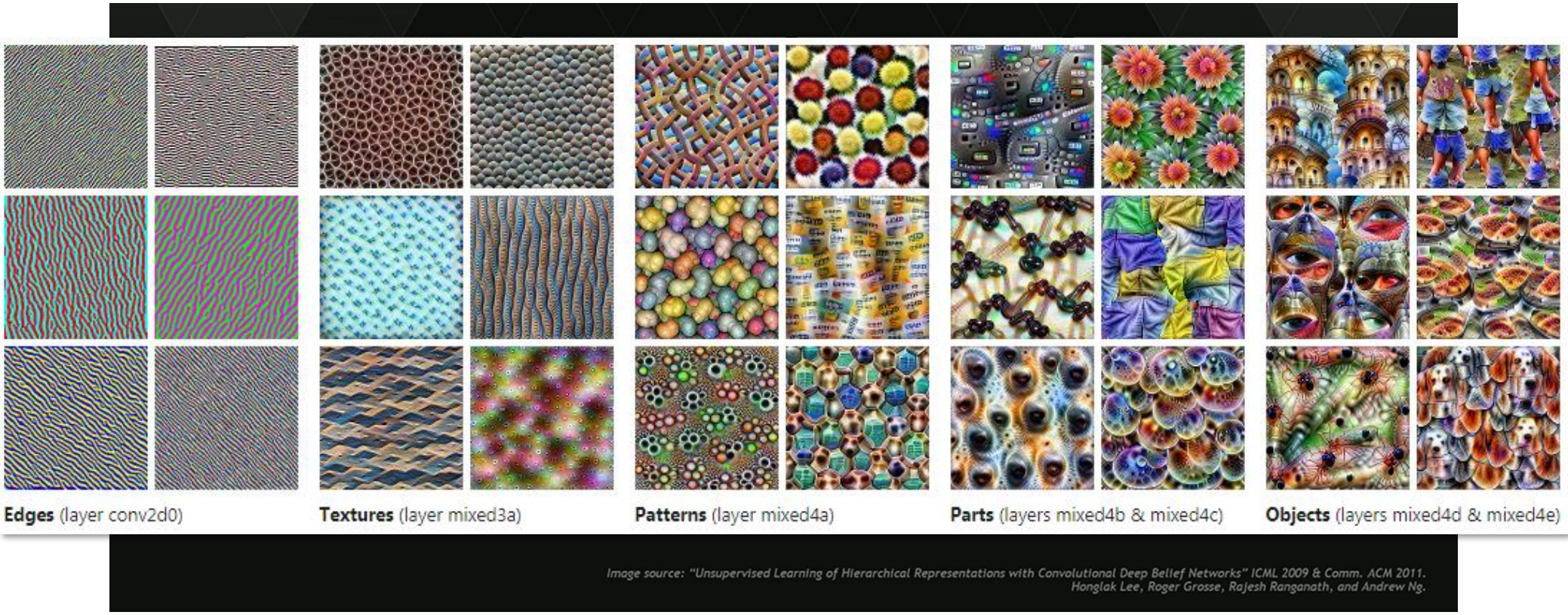


← Fehlerlandschaft

Methode: Anpassung der Gewichte von f in Richtung der steilsten Steigung (abwärts) von J

Was «sieht» das Neuronale Netz?

Hierarchien komplexer werdender Merkmale



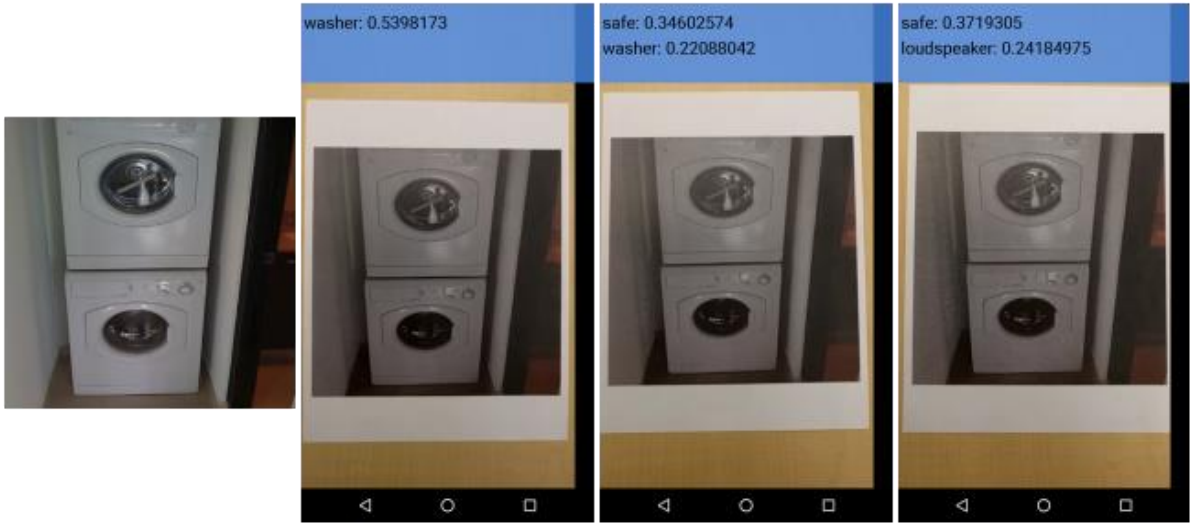
Quellen: https://www.pinterest.com/explore/artificial-neural-network/
Olah, et al., "Feature Visualization", Distill, 2017, https://distill.pub/2017/feature-visualization/.

Wie schlussfolgert die Maschine?

«Debugging» für Einblicke in die vermeintliche «Black Box»

Verdeutlichen ein Problem:

- Adversarial Examples

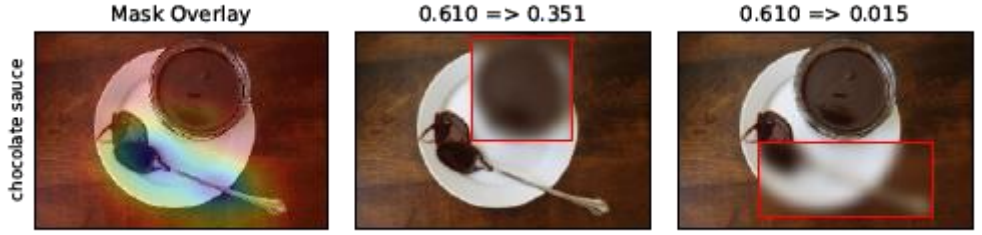


(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

https://blog.openai.com/adversarial-example-research/

Bieten eine Lösung:

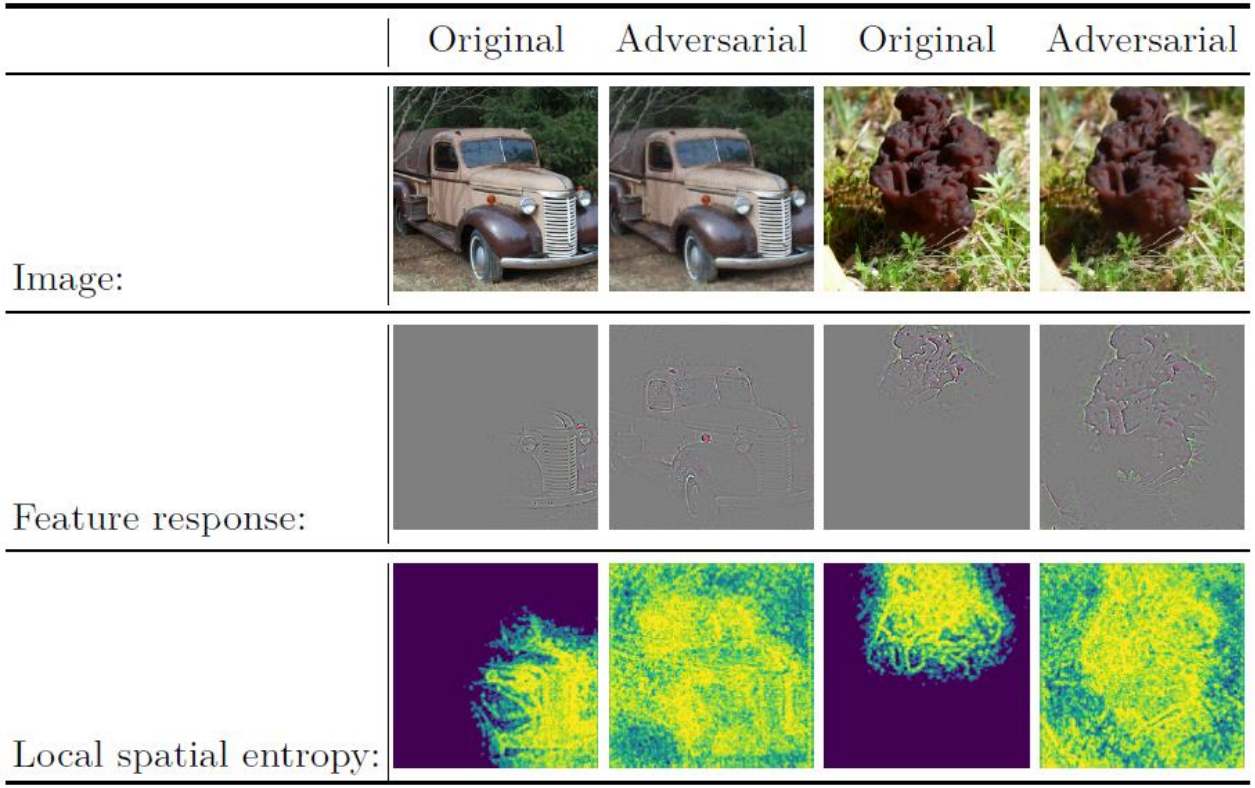
- Saliency Maps



Ruth C. Fong & Andrea Vedaldi, «Interpretable Explanations of Black Boxes by Meaningful Perturbation», 2017

Trace & detect adversarial attacks

…using average local spatial entropy of feature response maps



Amirian, Schwenker & Stadelmann (2018). *«Trace and Detect Adversarial Attacks on CNNs using Feature Response Maps»*. ANNPR'2018.