Zürcher Hochschule
für Angewandte Wissenschaften

Sicherheitsrelevante Aspekte von KI

Gastvorlesung im Studiengang «Krisen- und Notfallmanagement»,
Carl Remigius Medical School, 12. Dezember 2020

Thilo Stadelmann

Image source: https://www.inverse.com/article/35599-duke-robotics-tikad-drone

Why you should care: an example (a)

See https://en.wikipedia.org/wiki/2018_Caracas_drone_attack

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

Languages

Article Talk Read Edit View history Search Wikipedia

2018 Caracas drone attack

From Wikipedia, the free encyclopedia

On 4 August 2018, **two drones detonated explosives** near Avenida Bolívar, Caracas, where Nicolás Maduro, the President of Venezuela, was addressing the Bolivarian National Guard in front of the Centro Simón Bolívar Towers and Palacio de Justicia de Caracas.[3][4][5] The Venezuelan government claims the event was a targeted attempt to assassinate Maduro, though the cause and intention of the explosions is debated.[6][7] Others have suggested the incident was a false flag operation designed by the government to justify repression of opposition in Venezuela.[8][9][10]

Caracas drone attack

Part of the crisis in Venezuela

Top to bottom:
President Maduro being shielded.
Venezuelan troops retreating from the area.

2018 Caracas drone attack (Central Caracas)

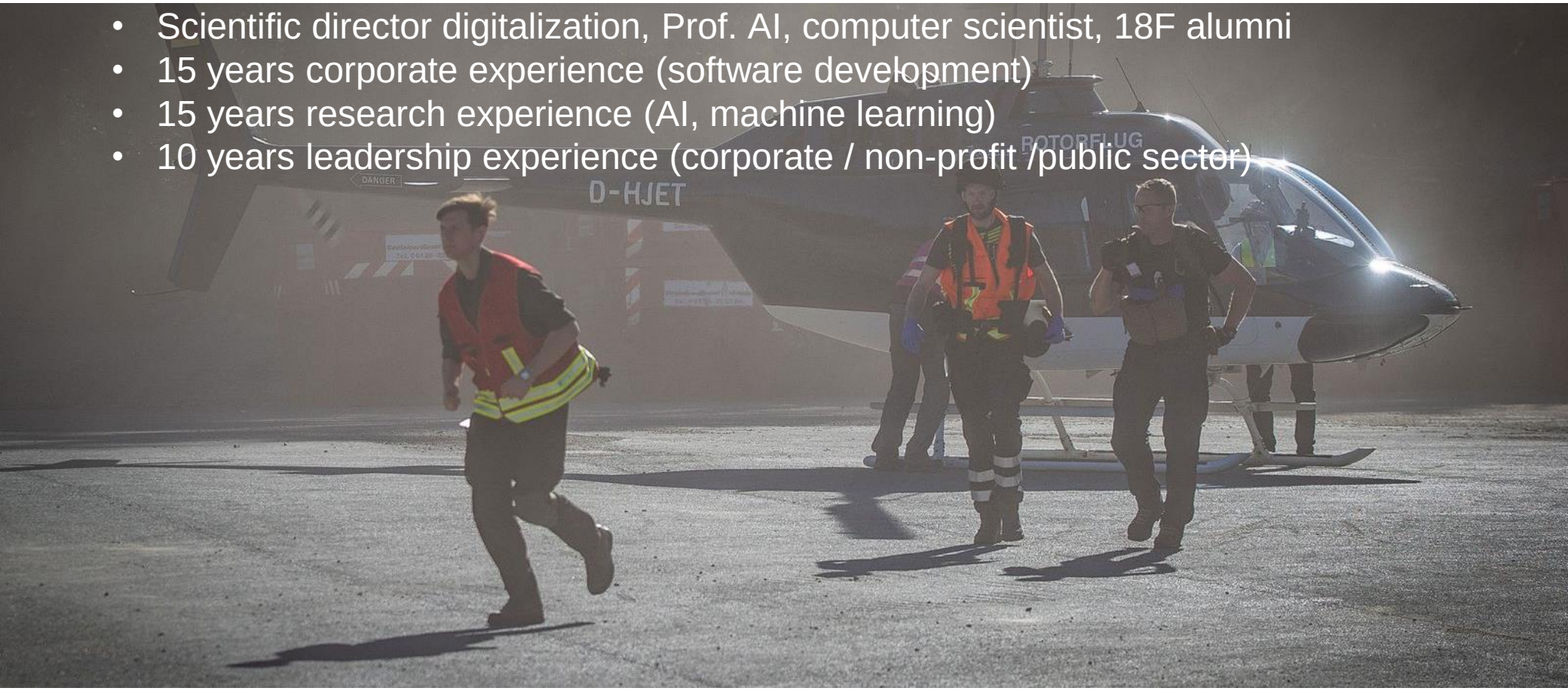
Contents [hide]

Why you should care: an example (b)

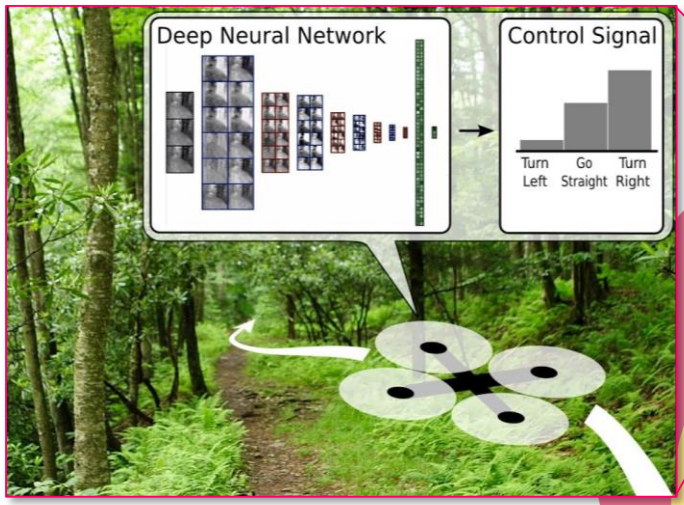
See https://www.youtube.com/watch?v=ruWC10AW87E

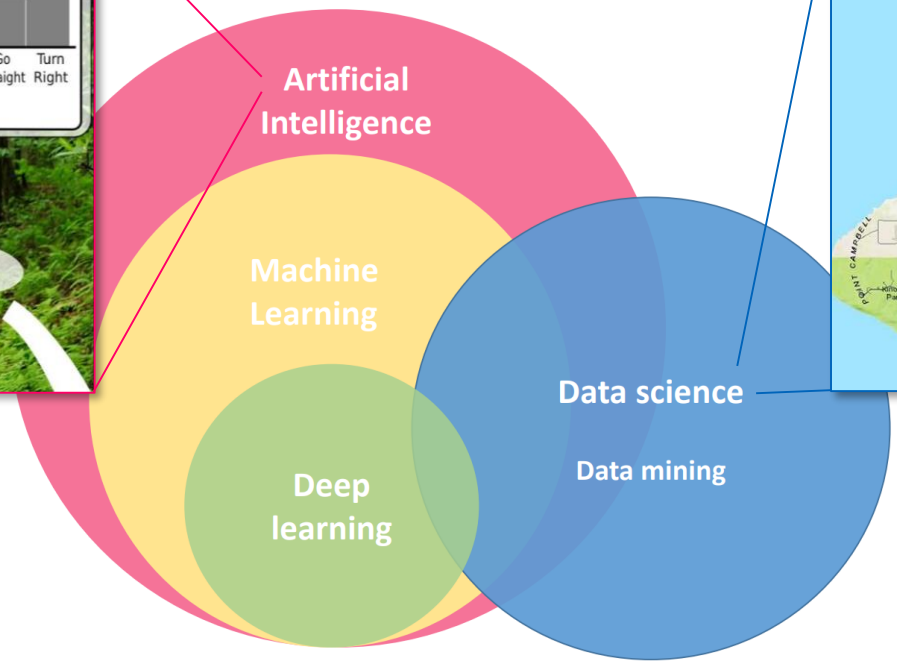
About me

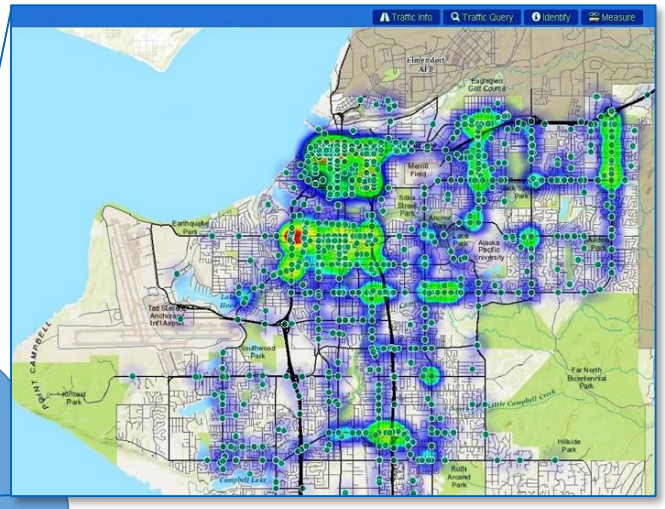
- Scientific director digitalization, Prof. AI, computer scientist, 18F alumni
- 15 years corporate experience (software development)
- 15 years research experience (AI, machine learning)
- 10 years leadership experience (corporate / non-profit /public sector)



1. What is AI / ML / Data Science?







Sources: Kulina et al., «A survey on Machine Learning-based Performance Improvement of Wireless Networks: PHY, MAC and Network layer», 2020
https://www.youtube.com/watch?v=umRdt3zGgpU, https://www.muni.org/Departments/traffic/Pages/Data.aspx

What is AI?

thinking

humanly

rationally

acting

"The exciting new effort to **make computers think**... machines with minds, in the full and literal sense." (Haugeland, 1985)

"[The automation of] **activities that we associate with human thinking**, activities such as decision-making, problem solving, learning..." (Bellman, 1978)

"The study of mental faculties through the use of **computational models**." (Charniak and McDermott, 1985)

"The study of the **computations that** make it possible to **perceive, reason, and act**." (Winston, 1992)

"The art **of creating machines that perform functions** that require **intelligence** when performed by people." (Kurzweil, 1990)

"The study of **how to make computers do things** at which, at the moment, **people are better**." (Rich and Knight, 1991)

"Computational Intelligence is **the study of the design of intelligent agents**." (Poole et al., 1998)

"AI... is concerned with **intelligent behaviour in artefacts**." (Nilsson, 1998)

The scientific field concerned with complex computer applications

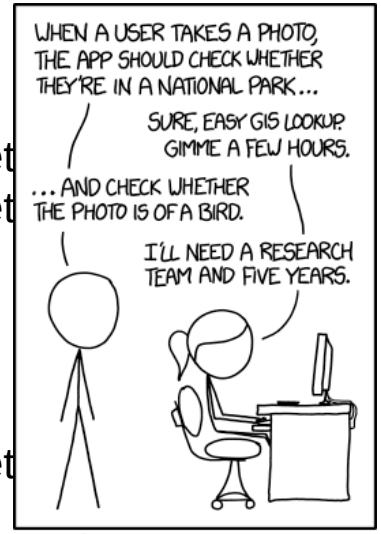
What belongs to AI?

An incomplete view of its subdisciplines

- **AI**
  - **Methods**
    - Symbolic
      - Search
      - Logic
      - Knowledge representation
      - Planning
    - Sub-symbolic
      - Statistical learning
      - Neural networks
        - Deep learning
  - **Applications**
    - Computer vision
      - Image classification
      - Object detection
      - Semantic segmentation
    - Speech processing
      - Speech recognition
      - Voice recognition
    - NLP
      - Text analytics
      - Machine translation
      - Chatbots
    - Robotics
      - Perception
      - Control

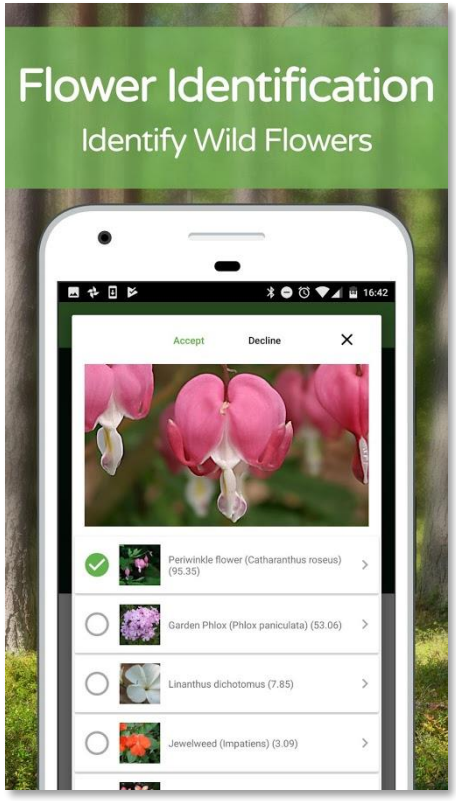
What can AI do today?

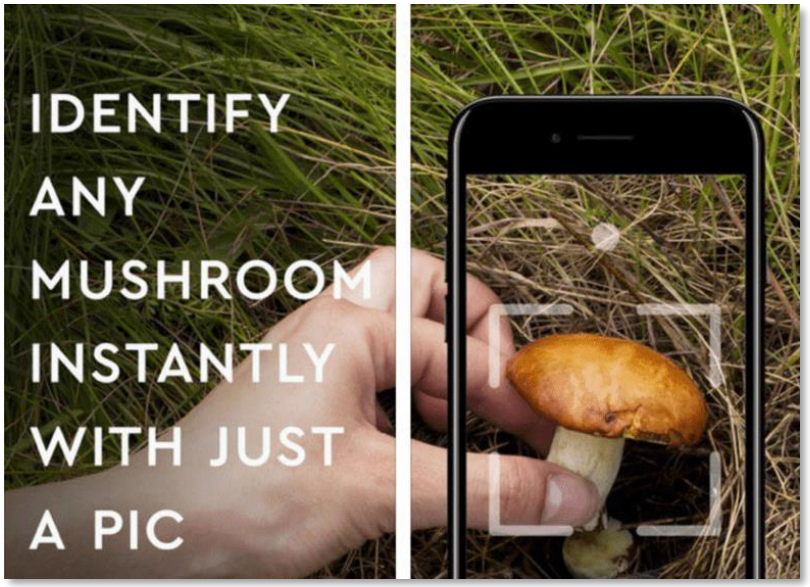
1. Play a decent game of **table tennis** — ok
2. **Drive** safely along a curving **mountain road** — ok
3. Drive safely along **Technikumstrasse** Winterthur — ok (only since recently)
4. **Buy** a week's worth of **groceries on the web** — ok
5. Buy a week's worth of groceries **at Migros** — no
6. **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



IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

Example: What AI can and cannot do in computer vision

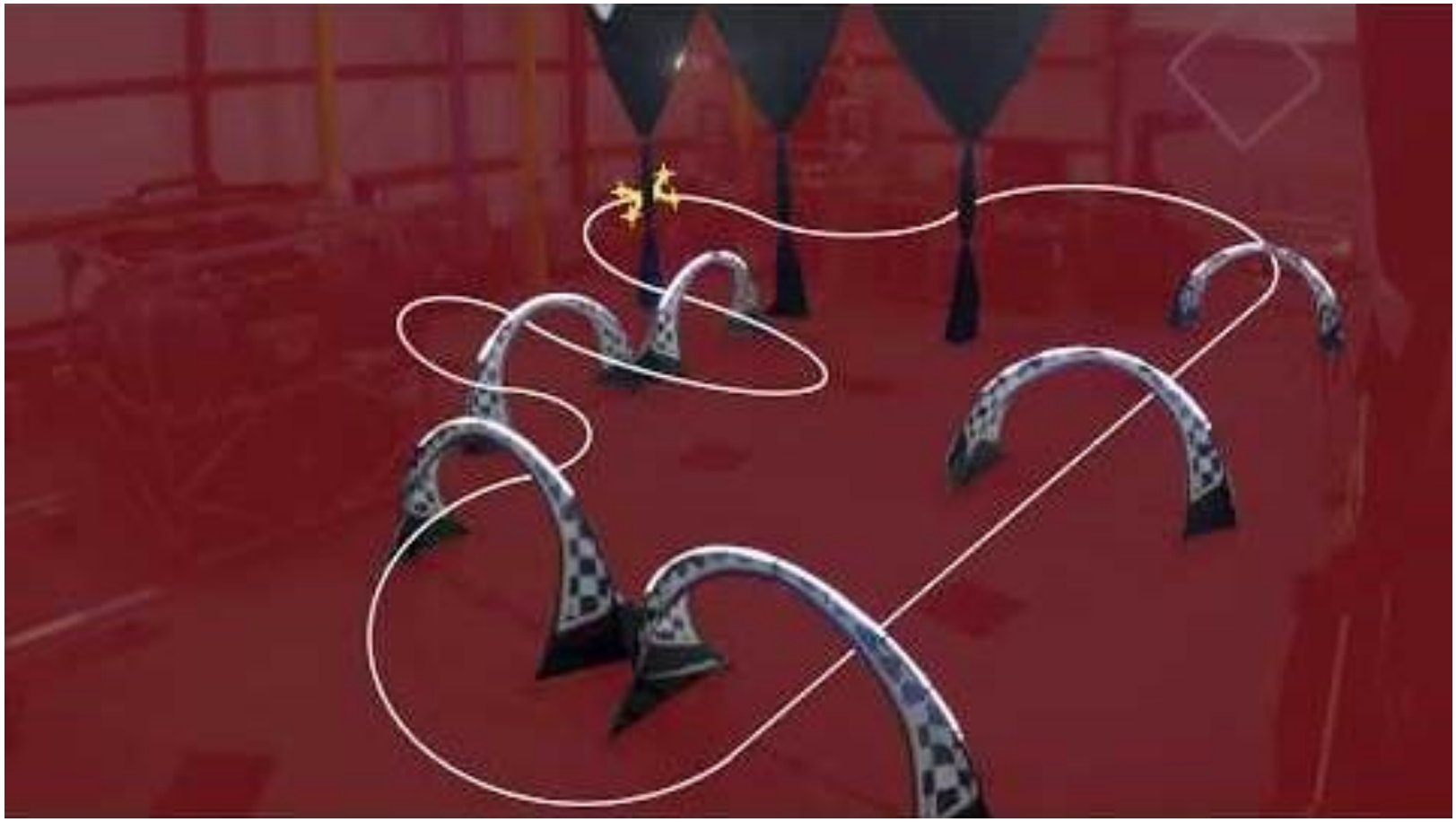




https://www.cultofmac.com/495088/avoid-potentially-deadly-ai-app/

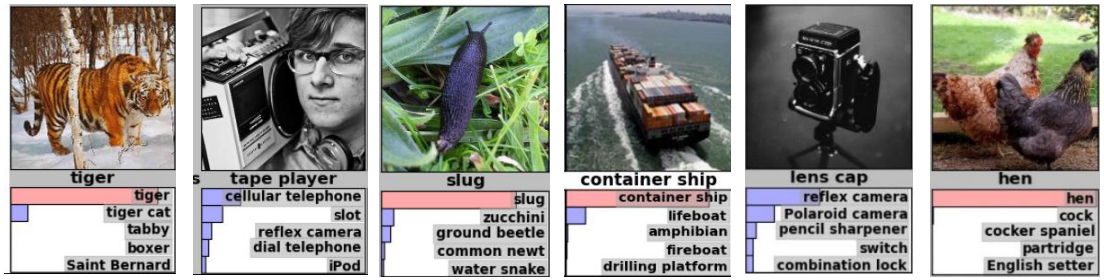
But: «Drone Race: Human vs. Machine»

See https://www.youtube.com/watch?v=SrqrGweKQAU

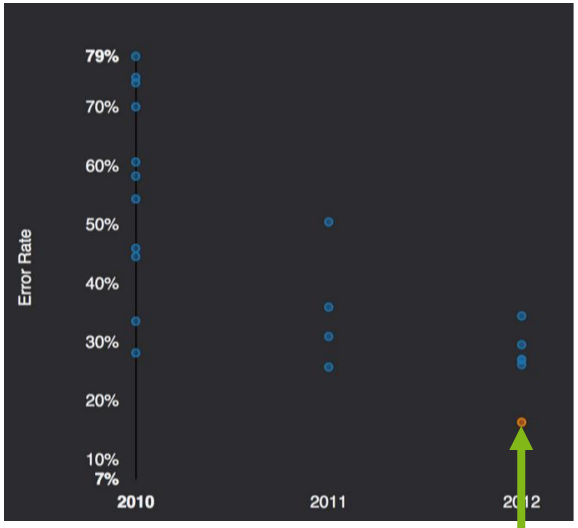


Why is AI hot today?
The ImageNet Competition

1000 categories
1 Mio. examples



…



A. Krizhevsky is first to use a «Deep Neural Network» (CNN)

2015: computers have lerned to «see»

4.95% Microsoft (February 06)
→ super-human (5.10%)

4.80% Google (February 11)

4.58% Baidu (May 11)

3.57% Microsoft (December 10)

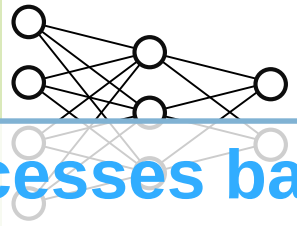
Idea: Add «depth» to learn features automatically

Classical image processing

Feature extraction
(SIFT, SURF, LBP, HOG, etc.)

(0.2, 0.4, …)

Classification
(SVM, neural network, etc.)



Container ship

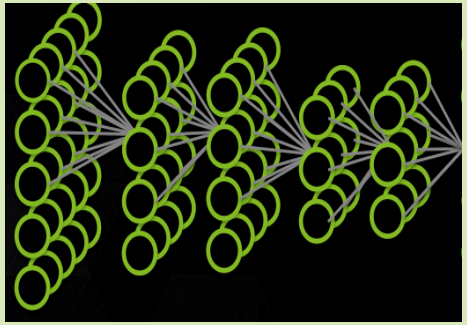
Tiger

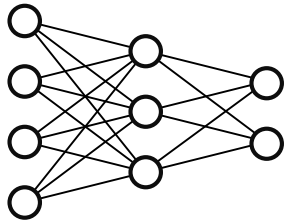
…

Automation of classical processes based on (high-dimenional) sensory input

Using Convolutional Neual Networks (CNNs)

Takes raw pixels in, learns features automatically!





Container ship

Tiger

…

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

Posted Jan 26, 2014 by Ca

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

DeepMind

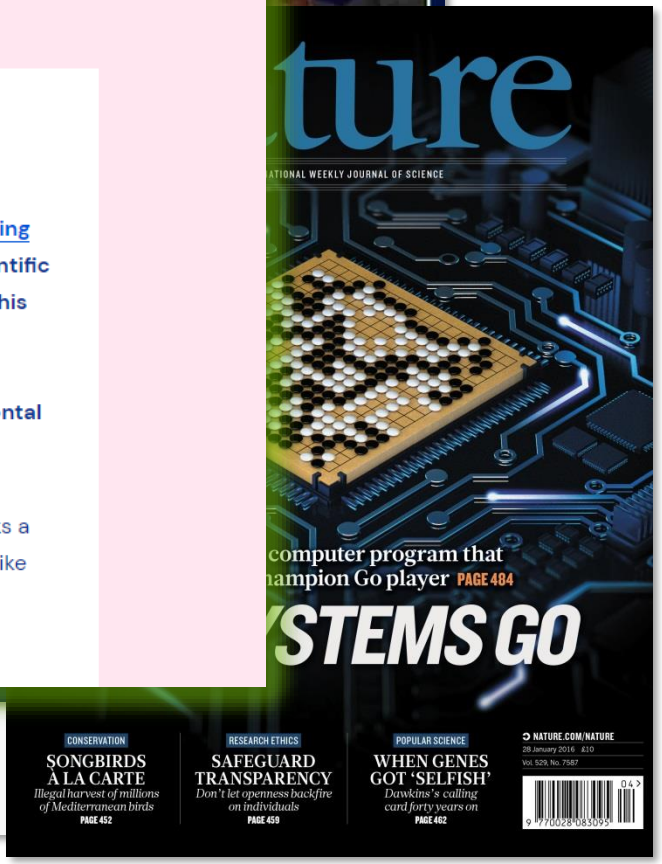
BLOG POST
RESEARCH

30 NOV 2020

AlphaFold: a solution to a 50-year-old grand challenge in biology

Proteins are essential to life, supporting practically all its functions. They are large complex molecules, made up of chains of amino acids, and what a protein does largely depends on its unique 3D structure. Figuring out what shapes proteins fold into is known as the "protein folding problem", and has stood as a grand challenge in biology for the past 50 years. In a major scientific advance, the latest version of our AI system AlphaFold has been recognised as a solution to this grand challenge by the organisers of the biennial Critical Assessment of protein Structure Prediction (CASP). This breakthrough demonstrates the impact AI can have on scientific discovery and its potential to dramatically accelerate progress in some of the most fundamental fields that explain and shape our world.

A protein's shape is closely linked with its function, and the ability to predict this structure unlocks a greater understanding of what it does and how it works. Many of the world's greatest challenges, like developing treatments for diseases or finding enzymes that break down industrial waste, are fundamentally tied to proteins and the role they play.



Künstliche Intelligenz

WaveNet lässt Computersprache natürlich klingen

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 speake 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 the target speaker's utterances and only a small set of <wav, phone> pairs from a number of anonymous speakers.

A's Waveforms | Speech Recognition | Speech Synthesis | B's Waveforms

Train1 \w small parallel dataset

Train2 \w large non-parallel dataset

"My name is Avin!" "My name is Avin!"

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

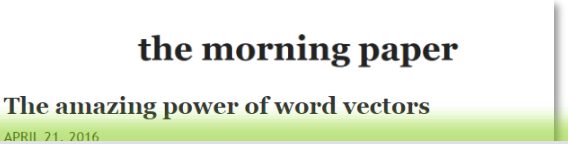
Al Shelley Pens Truly Creepy Horror Stories–And You Can Help

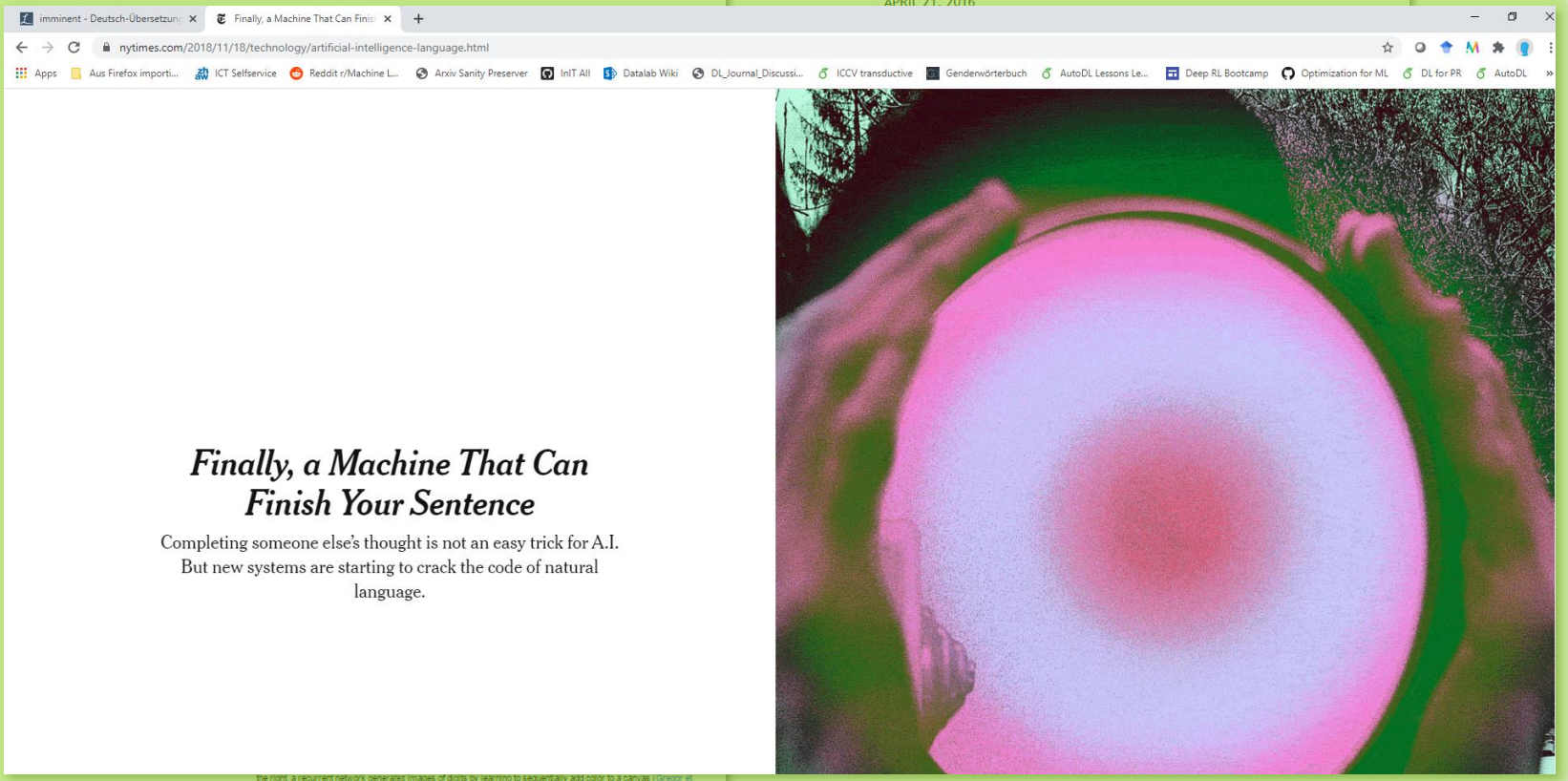
Neural Network Serves Up Truly Frightening Halloween Costume Ideas

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.

…and also with text!





«Die verblüffenden athletischen Leistungen von Quadrocoptern»

See https://www.youtube.com/watch?v=w2itwFJCgFQ

Foundation

Inductive supervised learning

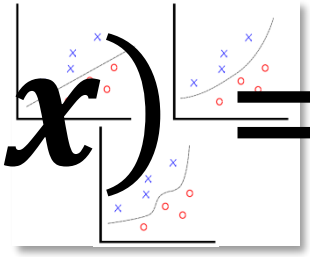
Assumption

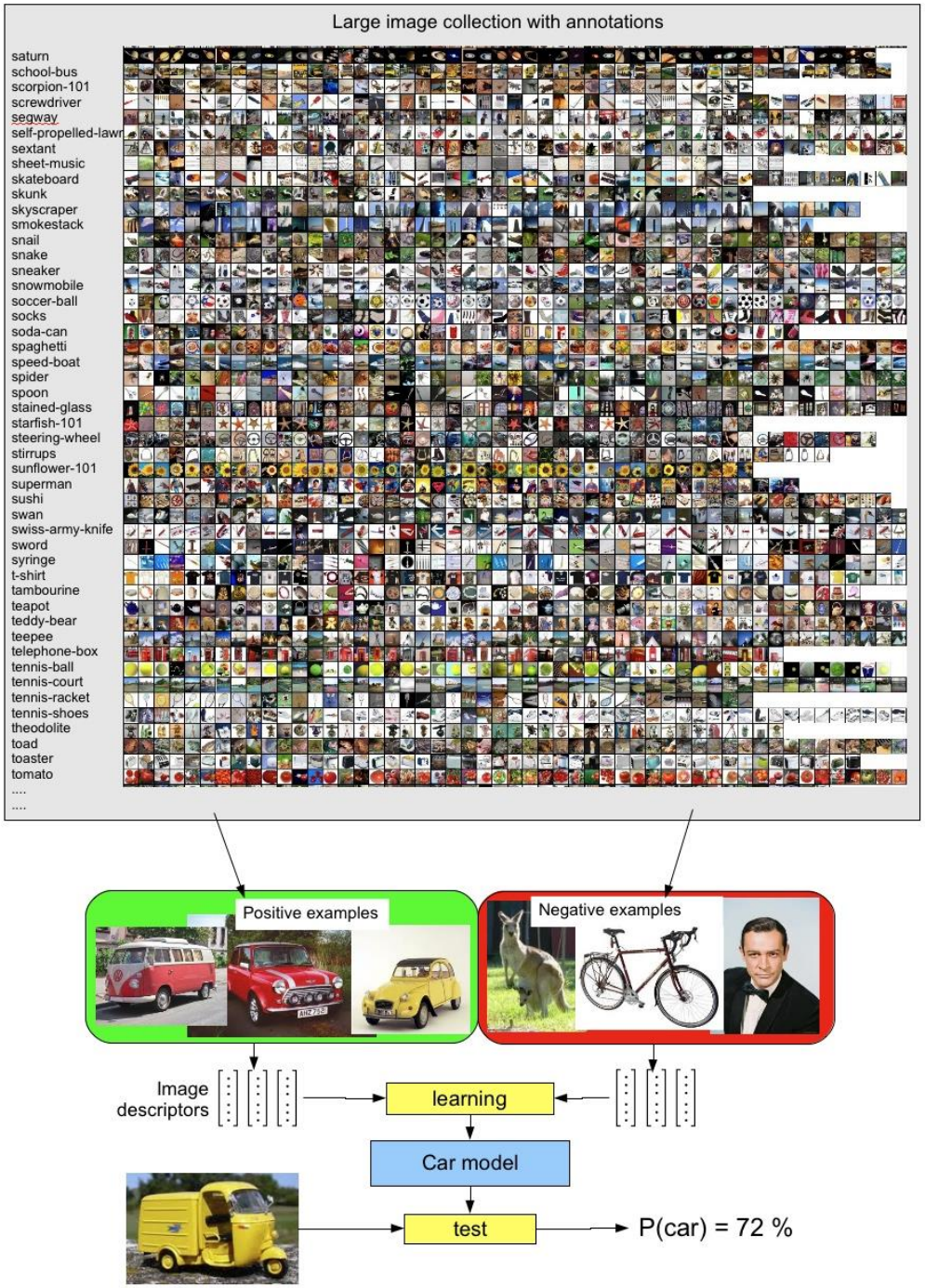
- A model fitted to a *sufficiently large* sample of data…
- …will **generalize** to unseen data

Method

- **Searching for optimal parameters of a function**…
- …such that all sample inputs (images) are mapped to the correct outputs (e.g., «car»)

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

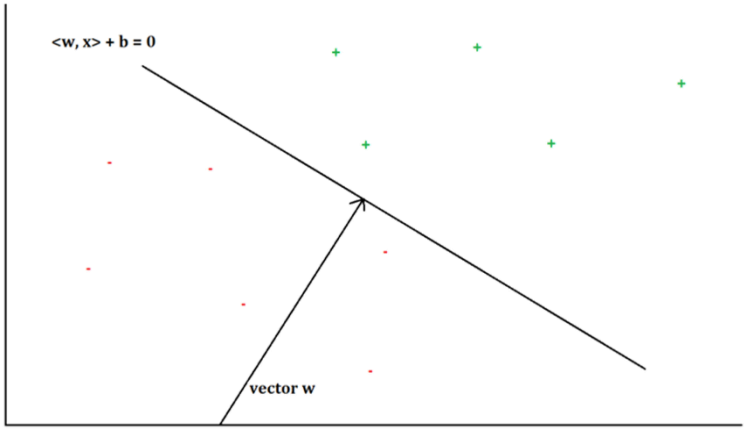


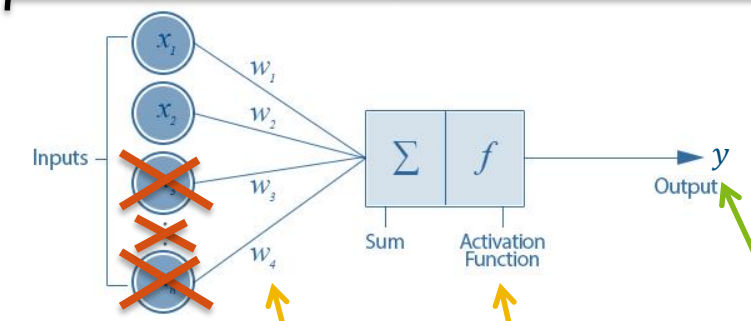


Source: http://lear.inrialpes.fr/job/postdoc-large-scale-classif-11-img/attribs_patchwork.jpg

Search for optimal parameters *of a function*?

neuron





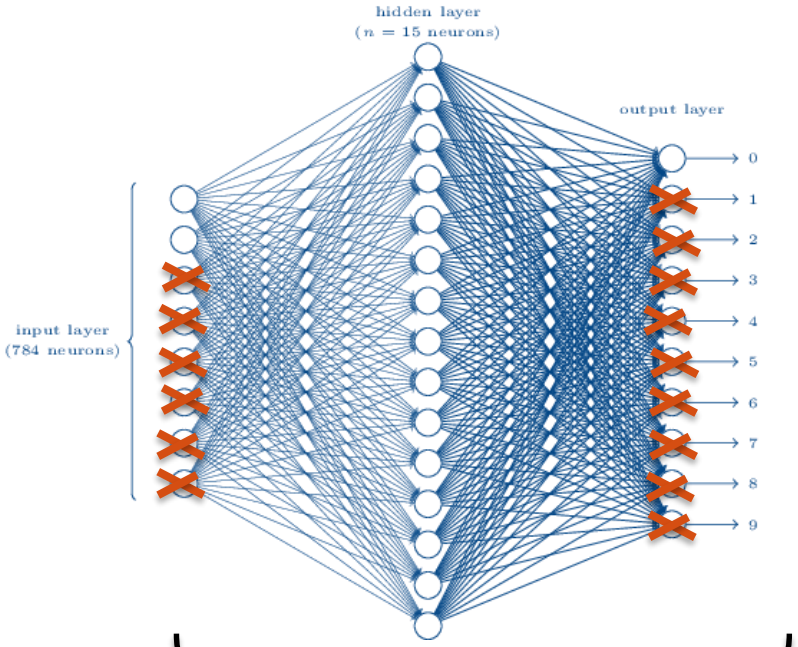
Threshold / decision

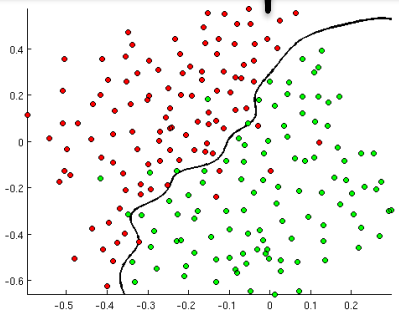
features (e.g., pixel)

Tunable parameters

Result (interpreted as e.g. «car»)

neural net





«Meet the dazzling flying machines of the future»

See https://www.youtube.com/watch?v=RCXGpEmFbO

2. AI for Security

Source: https://eu.usatoday.com/story/tech/2019/11/19/police-technology-and-surveillance-politics-of-facial-recognition/4203720002/

Cybersecurity

See https://securityintelligence.com/posts/how-ai-makes-jobs-in-cybersecurity-less-stressful/

Security Intelligence

News Series Topics Threat Research Podcast Events

Home / Topics / Artificial Intelligence

How AI Can Make Cybersecurity Jobs Less Stressful and More Fulfilling

November 16, 2020 | By **Srini Tummalapenta** co-authored by **Megan Radogna** | 5 min read

Words for health and the human body often make their way into the language we use to describe IT. Computers get viruses; companies manage their security hygiene; incident response teams train on their cyber fitness. Framing IT concepts in terms of health can also be useful when looking at security operations centers (SOCs) and jobs in cybersecurity.

For many businesses and other entities today, SOCs are not the healthiest they could be. Jobs in cybersecurity can be stressful and overwhelming due to the volume of alerts. Many teams lack the staff they need to keep up with the influx.

The average SOC receives over 11,000 alerts a day, and 28% of all alerts are never addressed, says the 2020 State of Security Operations study from Forrester Consulting, sponsored by Palo Alto Networks.

Popular Articles

Sep 28, 2020 | 9 min read
Ransomware 2020: Attack Trends Affecting Organizations Worldwide

5 days ago | 6 min read
IBM Uncovers Global Phishing Campaign Targeting the COVID-19 Vaccine Cold Chain

5 days ago | 3 min read
5 Ways to Accelerate Security Confidence for AWS Cloud

Dec 1, 2020 | 5 min read
The Future of Cybersecurity: How to Prepare for a Crisis in 2020 and Beyond

But: Artificial intelligence vs. natural stupidity

…or the difficulty of “optimizing” a complex system

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.

Face recognition in the public

See https://www.vice.com/en/article/n7ve4q/varanasi-india-using-facial-recognition-surveillance-technology

VICE Video TV The 8:46 Project News World News Tech Music

privacy

Another City Is Using Crime Control as an Excuse for Facial Recognition Surveillance

Varanasi in India is installing 3,000 CCTV cameras with automated facial recognition tech at the city's crossings.

By Varsha Rani

November 27, 2020, 12:58pm Share Tweet Snap

PHOTO COURTESY OF MICHAŁ ZAKUBOWSKI VIA UNSPLASH

From mandatory face masks and temperature checks, to socially distant holiday seasons, 2020 has upended our lives in the most haunting way. It's also meant that governments across the world could introduce intrusive surveillance technology into our daily lives in the name of public health.

In China, the government has been tracking its citizens by monitoring their smartphones. Meanwhile, countries like Singapore and India have been using a contact tracing app to monitor those infected by the virus, while Israel is using a counter terrorism agency to keep track of its citizens' movements.

World News

China Is Sending a COVID Testing Team to Hong Kong. Locals Worry It

MORE LIKE THIS

Tech

Police Are Tapping Into Ring Cameras to Expand Surveillance Network In Mississippi

EDWARD ONGWESO JR

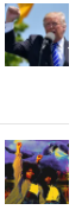
11.06.20

The VICE Guide to Right Now

Narcissists are More Likely to be Involved in Politics, Study Says

VARSHA RANI

10.05.20



Life

'Advodating': The Controversial Dating Trend That Mixes Protest with Pleasure

MAIGHNA NANU

11.30.20

Question Of The Day

What Will You Miss About the Pandemic When (And If) It's Over? We Asked Around.

VARSHA RANI

Biometric access

See https://www.munich-airport.de/kontaktlos-reisen-9912987

/ Kontaktlos Reisen

Star Alliance Biometrics

Seit Mitte November 2020 gibt es am Flughafen München eine neue technische Innovation, die den Reiseprozess erleichtern soll. Star Alliance Biometrics, ein Produkt der Star Alliance, der größten Luftfahrtallianz der Welt, soll Passagieren es ermöglichen, die Vorzüge der Gesichtserkennungstechnik zu nutzen. Die Passagiere können so berührungslos durch ausgewählte Sicherheits- und Boarding-Gates gehen. In naher Zukunft kann die Palette der Einsatzmöglichkeiten schrittweise erweitert werden - zum Beispiel auf die Gepäckausgabe und den Zugang zu den Lounges.

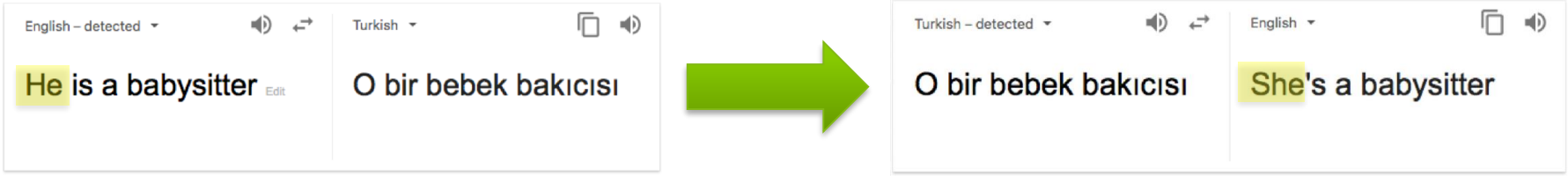
Wie funktioniert Star Alliance Biometrics?

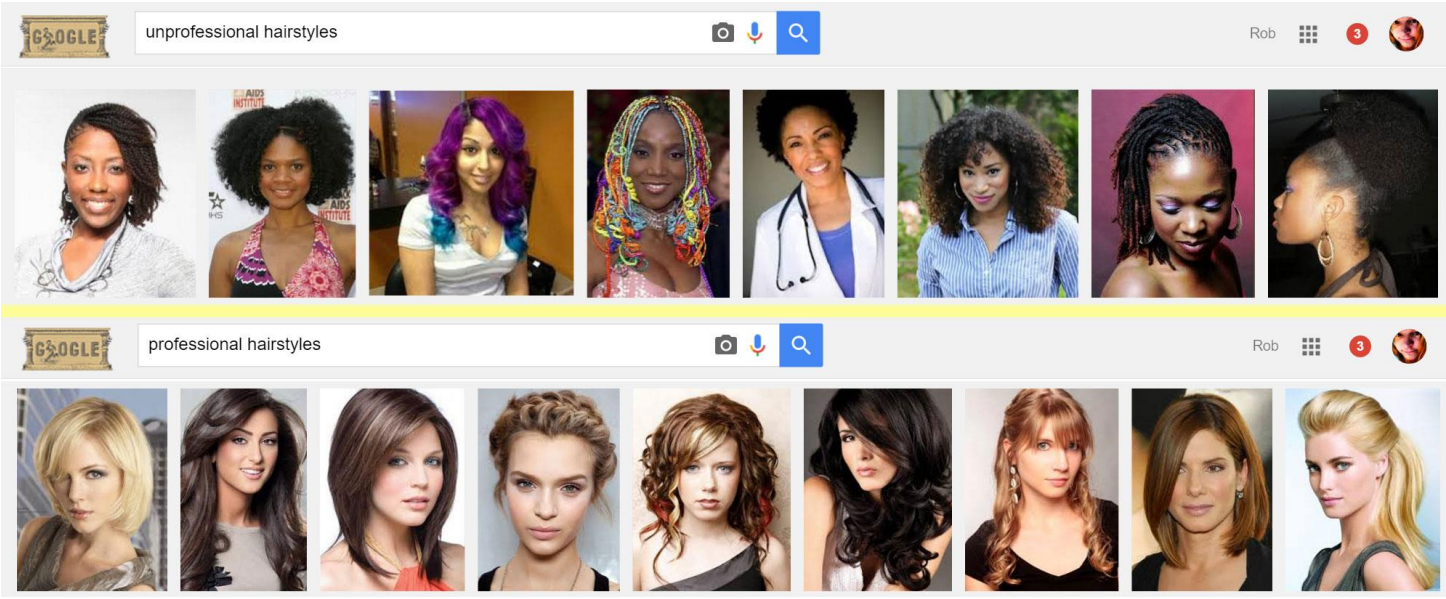
Wenn ein registrierter Passagier von einem teilnehmenden Flughafen und mit einer teilnehmenden Fluggesellschaft reist, gleicht die Gesichtserkennungssoftware von Star Alliance Biometrics das Live-Bild des Passagiers mit den Bordkarteninformationen und dem biometrischen Profil ab, so dass der Gast durch entsprechend ausgestattete Touchpoints gehen kann. Zum Start des Systems ist es an den Flughäfen Frankfurt und München für Passagiere verfügbar, die von dort mit Lufthansa oder SWISS reisen. Dort sind die eGates, die mit dem System zur biometrischen Identifikation ausgestattet sind, deutlich mit dem Star Alliance Biometrics-Logo über dem Gate sowie mit Bodenmarkierungen gekennzeichnet.

Registrierung

Selbst mit Mund-Nasenschutz kann die biometrische Erkennung funktionieren

But: Biases through purely statistical learning



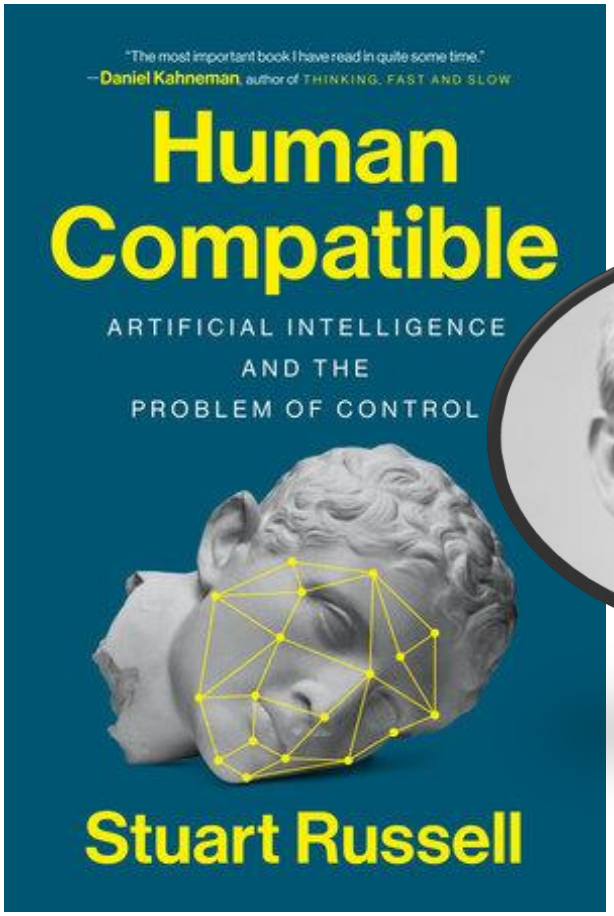


…introduces many problems for biometrics

Source: https://www.aclum.org/en/news/facial-recognition-technology-falsely-identifies-famous-athletes

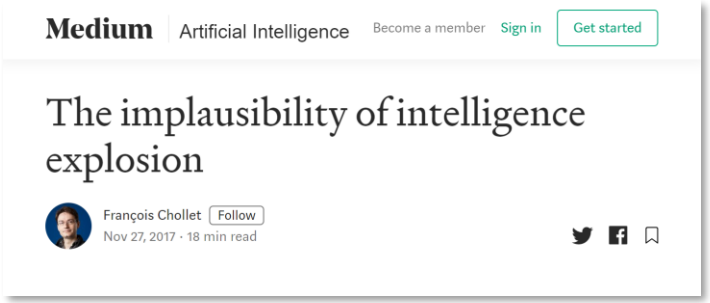
3. Security threats through AI

Inherent existential risks?



Assumption: AI systems will gain **general intelligence** eventually

But:



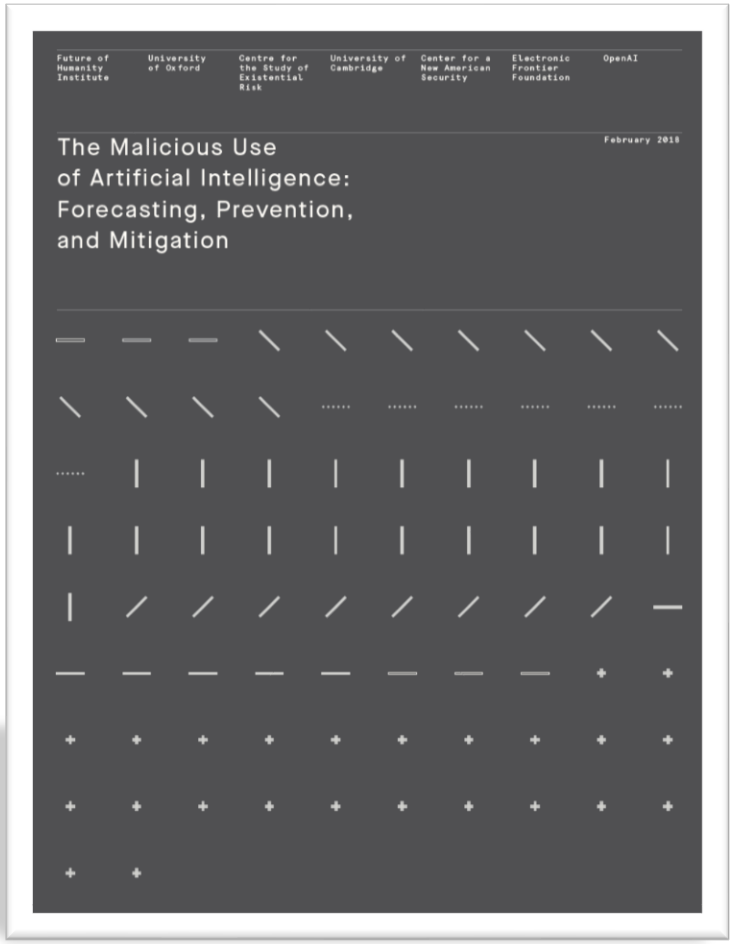
Conclusion: Any AI system optimizing a **fixed objective** will eventually **become destructive** to human interests

Risks through malicious use of AI?

- AI per definition is a "**dual use technology**"
  → see report by Brundage et al., 2018

- Still: "**natural stupidity**" is the more imminent threat

- **AI ethics** and explainable AI became mainstream and hot research topics in the recent years – not because of intolerable risks, but because of:



Security-relevant properties of AI

What enables potential threats by AI systems?

- **Dual-use** area of technology: AI systems and the knowledge of how to design them can be put toward both civilian and military uses, and more broadly, toward beneficial and harmful ends.
- **Efficiency and scalability**: “efficient” if it can complete a certain task more quickly or cheaply than a human could in production; “scalable” if increasing the computing power or making copies would allow it to complete many more instances of the task.
- **Potential to exceed human capabilities**: there appears to be no principled reason why currently observed human-level performance is the highest level of performance achievable.
- **Potential to increase anonymity** and psychological distance: AI systems can allow their users who would otherwise be performing the task to retain their anonymity and experience a greater degree of psychological distance from the people (victims) they impact.
- **Rapid diffusion**: it is easy to gain access to software and relevant scientific findings in AI.
- **Novel unresolved vulnerabilities**: e.g., poisoning attacks (introducing training data that causes a learning system to make mistakes), adversarial examples (inputs designed to be misclassified by machine learning systems), and the exploitation of flaws in the design of autonomous systems’ goals.

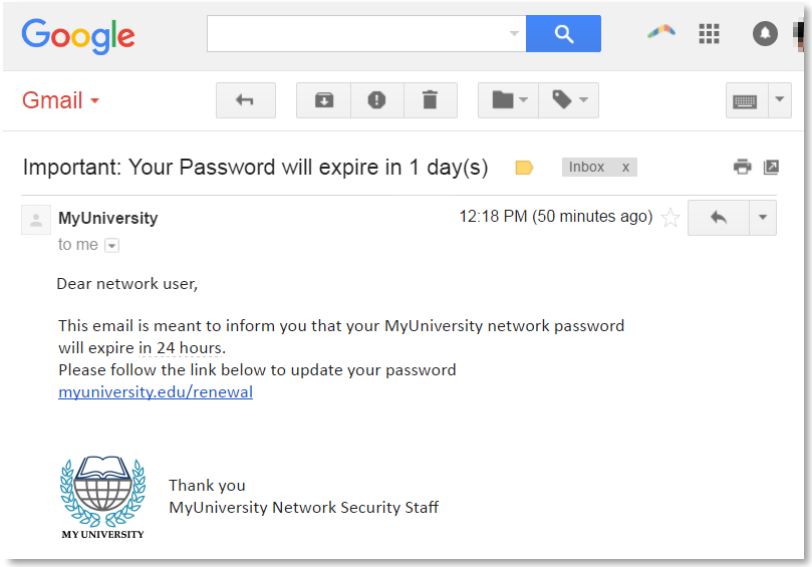
Scenario 1/3: AI expands existing threats

Expandable (by means of efficiency, scalability, and ease of diffusion)

- **Set of actors who can** carry out the attack
- **Rate** at which these actors can **carry** it **out**
- **Set of** plausible **targets**
- **Willingness** of actors **to carry out** certain **attacks** (by means of increased distance)

Example: spear phishing attack

- Definition: a **personally targeted phishing** attack (fooling by building a superficially trustworthy facade) using information specifically relevant to the target
- Usually too expensive and labor-intensive, but likely **automatable** in the future (data collection, data synthesis)



Scenario 2/3: AI introduces new threats

Otherwise **infeasible attacks** (by means of being unbounded by human capabilities)

- Example: disinformation by **impersonating** others using voice/image/text synthesis
- Compare https://lyrebird.ai/

Novel vulnerabilities (by means of deployed systems with known issues)

- Example: cause self-driving cars to **crash** by presenting them with adversarial examples

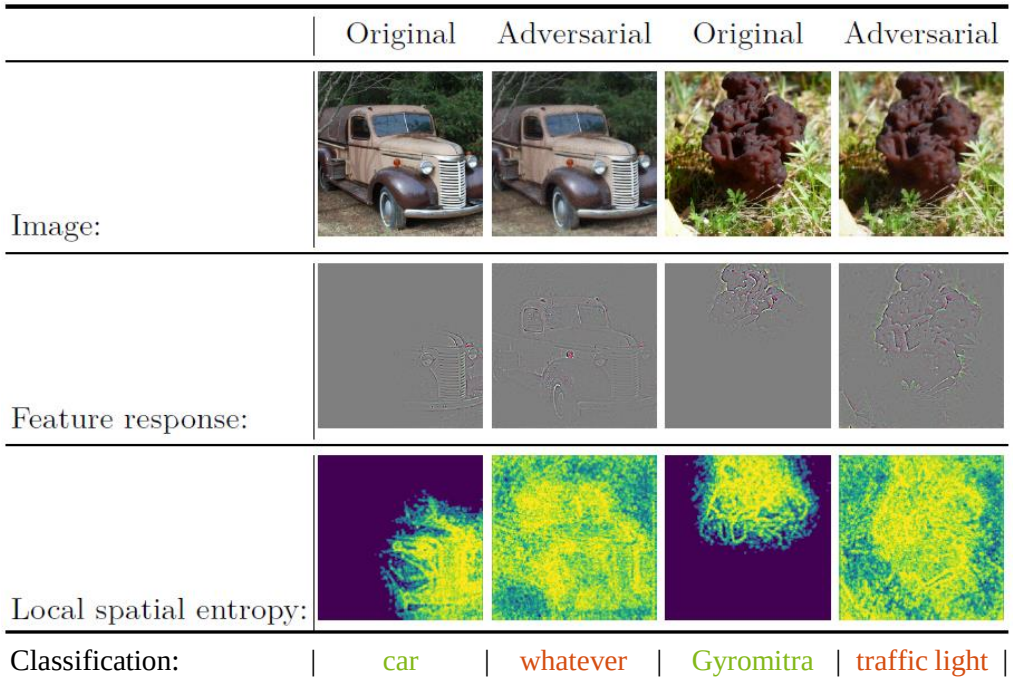
Eykholt et al., *„Robust Physical-World Attacks on Deep Learning Visual Classification“*, CVPR 2018

Example for novel vulnerabilities

Adversarial attacks and counter measures

Adversarial examples

- Created by optimizing (training on) the input image for an expected (wrong) output
- Can be detected 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.

Scenario 3/3: AI alters the typical character of threats

- **Highly effective attacks** will become more **typical** as trade-off between the frequency and scale of attacks vanishes (because of efficiency, scalability, and exceeding human capabilities)

- **Finely targeted attacks** will become more **prevalent** (because of efficiency and scalability): for example, killing specific members of a crowd using drone swarms and facial recognition instead of bombing

- **Difficult-to-attribute attacks** will become more **typical** (because of increasing anonymity)

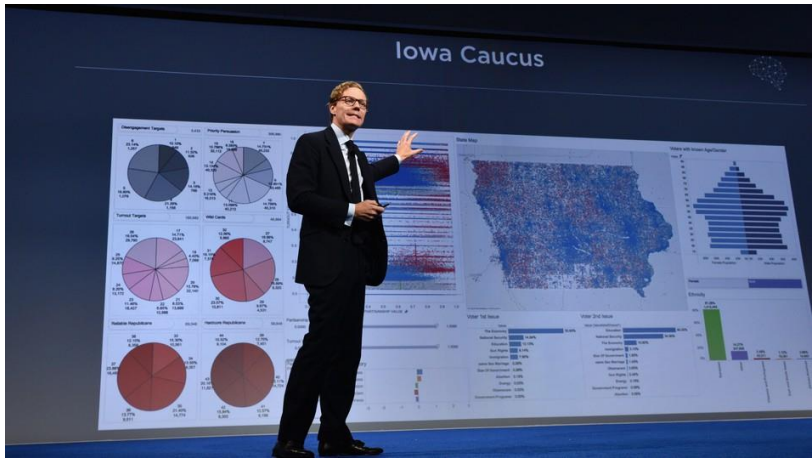
- **Exploiting vulnerabilities** of AI systems become more **typical** (because of known vulnerabilities and pervasiveness of deployed systems)

Potential impact areas

Digital security
- By **using AI** systems to **automate cyberattacks** or social engineering
- By **attacking AI** systems

Physical security
- By individual drones or **autonomous weapons**
- By **coordinating swarms** that otherwise not be controllable
- By **making normal** autonomous **agents** like cars, power plants etc. **malfunction**

Political security
- By **surveillance** and mass collection of data
- By persuasion through **targeted propaganda**
- By deception through synthetic news, videos etc.



Picture: Cambridge Analytica CEO Alexander Nix speaks at the 2016 Concordia Summit
© BRYAN BEDDER / GETTY IMAGES FOR CONCORDIA SUMMIT

4. Outlook & ethical considerations

Source: https://xkcd.com/1838/

It's difficult to make predictions, especially about the future[1]

Some guidelines how **not** to do it[2]:

1. **Overestimating and underestimating**: *«We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run.»*
2. **Imagining magic**: *«Any sufficiently advanced technology is indistinguishable from magic.»*
3. **Performance versus competence**: *«People generalize from the performance an AI shows on some task to a competence that a person performing the same task could be expected to have.»*
4. **Suitcase words**: *«Marvin Minsky called words that carry a variety of meanings "suitcase words." "Learning" is a powerful suitcase word; it can refer to so many different types of experience.»*
5. **Exponentials**: *«People may think that the exponentials they use to justify an argument are going to continue apace. But exponentials can collapse when a physical limit is hit, or when there is no more economic rationale to continue them.»*
6. **Hollywood scenarios**: *«Many science fiction movies assume that the world is just as it is today, except for one new twist. But we will not suddenly be surprised by the existence of super-intelligences.»*
7. **Speed of deployment**: *«Capital costs keep physical hardware around for a long time. Thus, almost all innovations in robotics and AI take far, far, longer to be really widely deployed.»*

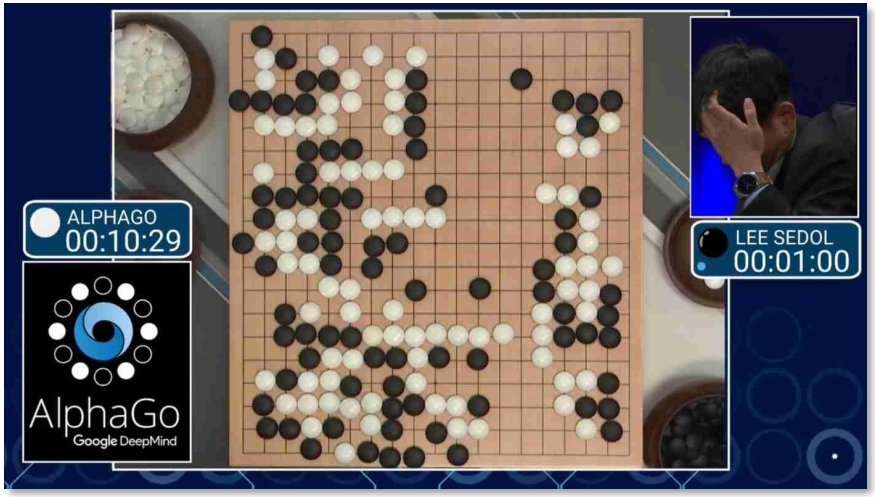
[1]) See https://quoteinvestigator.com/2013/10/20/no-predict/.
[2]) See Rodney Brooks, «The Seven Deadly Sins of AI Predictions», Technology Review, 2017 (compare lab P01b).

Basis for transformation (I): automation „at scale“

Or: “digital transformation” refers to a shift in all aspects of society, driven/enabled by this small set of technologies

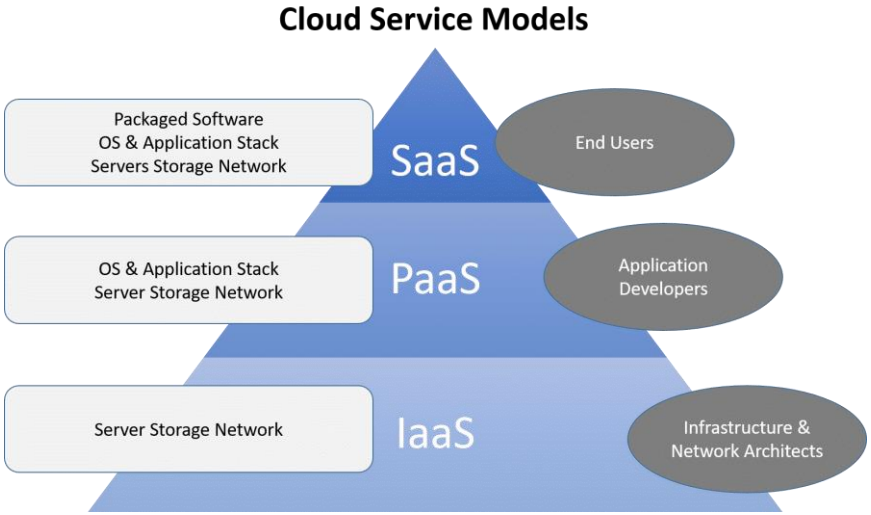
AI

Massively enhanced automation depth through progress in pattern recognition



CLOUD COMPUTING

No need to invest into (IT) infrastructure anymore before entering the market



Basis for transformation (II): decoupling

size of idea ≠ size of implementing organization

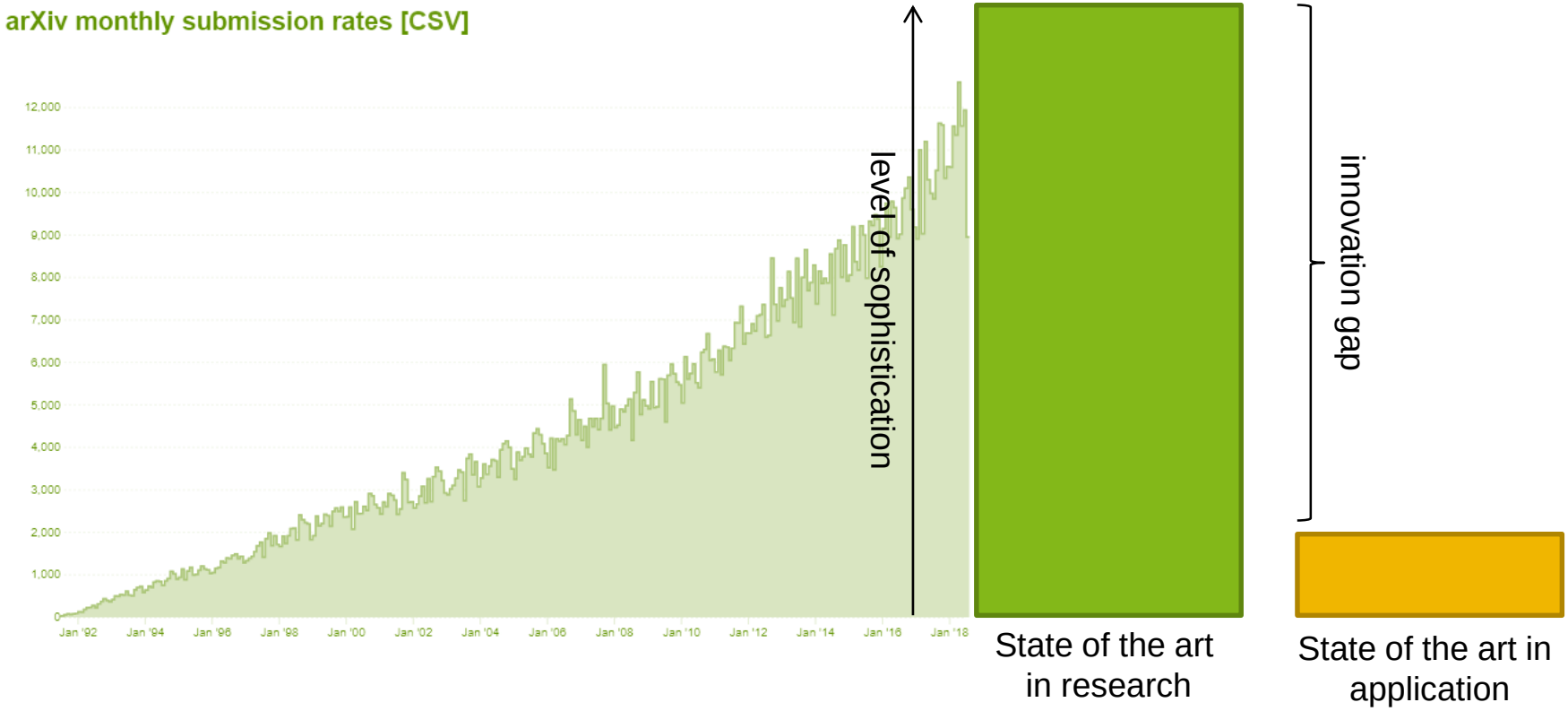
…small organizations can build **whatever they want**
(given know-how, data and an interesting use case)

the technology is sector-independent

…enabling **new** alliances and co-operations

Basis for transformation (III): speed

Average time from (pre-)publication to application: approx. 3 month



Forecast: rapid transformation

…even without any further technological progress

1. hypothesis: Use of (current) AI will increase massively within the next 3 years
- Indicator: **AI progress** is mainly driven by **industrial interests (earnings outlook)**; customers value convenience; these incentives „keep the engine running“

2. hypothesis: This will revolutionize all aspects of society
- Indicator: It **shifts power**

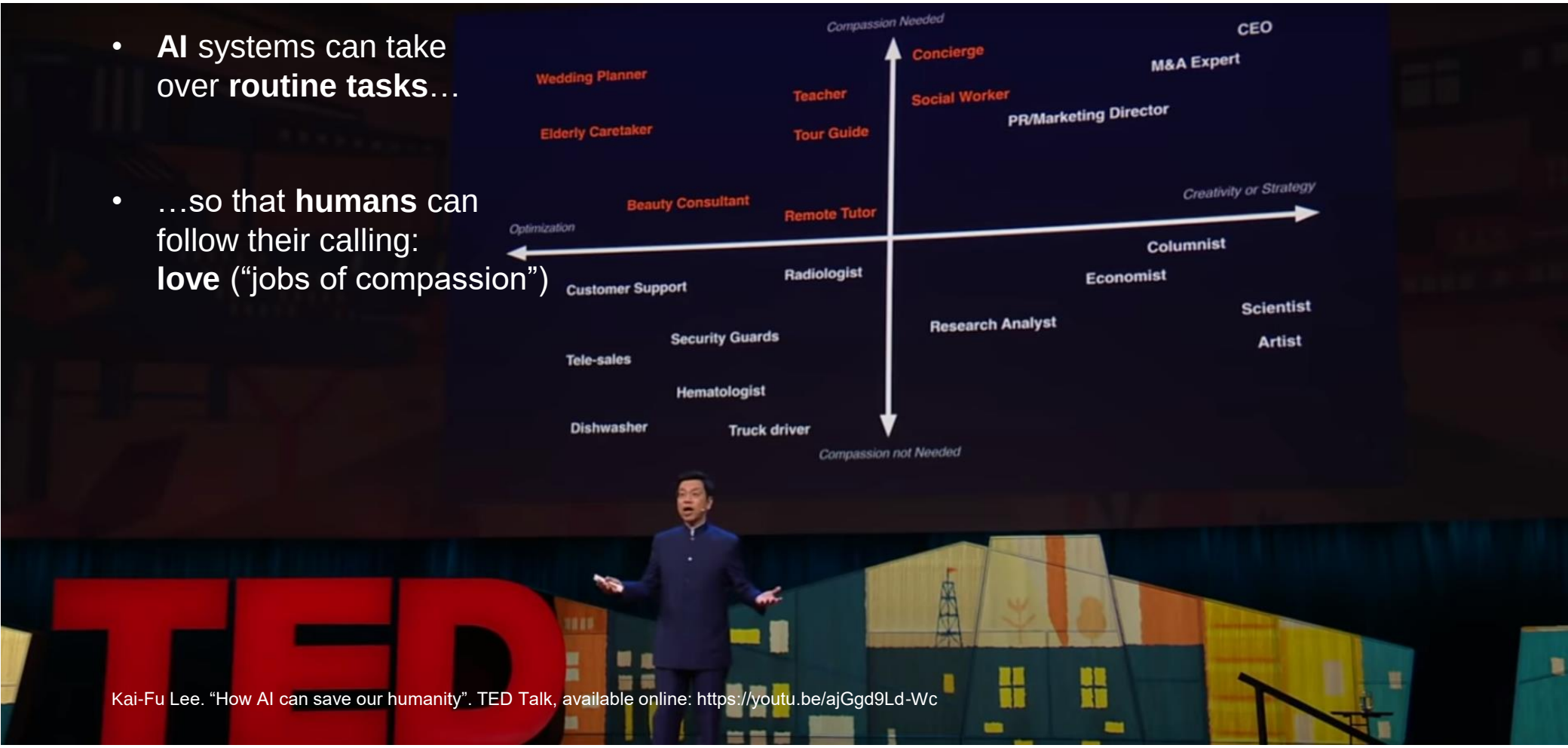
3. hypothesis: Main challenge is our dealings with each other (not with AI)
- Argument: AI (etc.) “for the common good” is an important topic; decisive however is **how the society designs new rules** (regulations) for community life in a digital society

Cp: Stockinger, Braschler & Stadelmann. “Lessons Learned from Challenging Data Science Case Studies”. In: Braschler et al. (Eds), *“Applied Data Science - Lessons Learned for the Data-Driven Business”*, Springer, 2019.

Where are we heading?

The vision of Kai-Fu Lee, venture capitalist & scientist

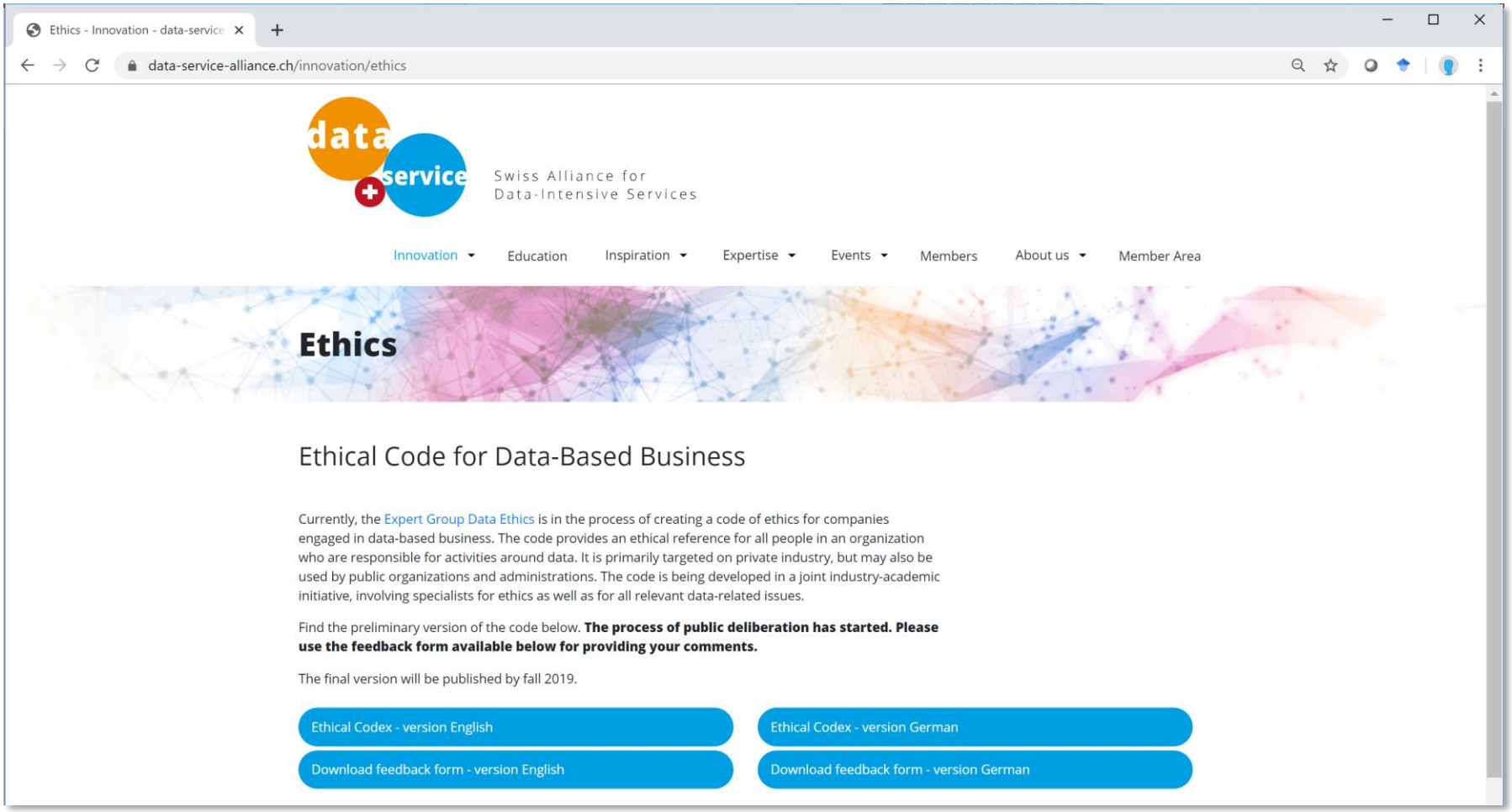
- **AI** systems can take over **routine tasks**…
- …so that **humans** can follow their calling: **love** (“jobs of compassion”)



Kai-Fu Lee. “How AI can save our humanity”. TED Talk, available online: https://youtu.be/ajGgd9Ld-Wc

A pragmatic, Swiss-made ethical code of conduct for using AI in use

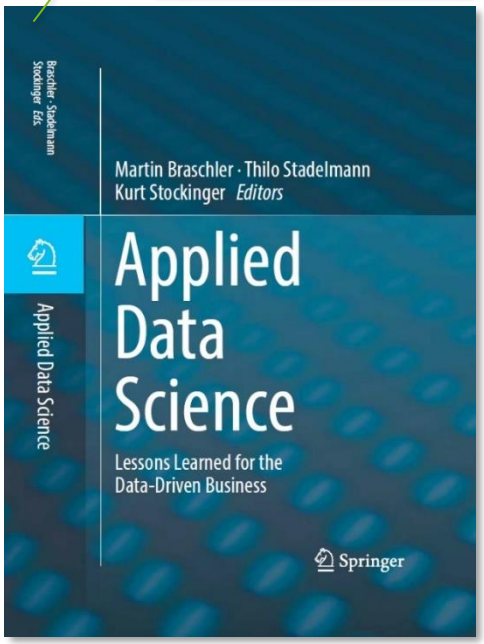
See https://data-service-alliance.ch/innovation/ethics



Conclusions

- Deep Learning led to **a paradigm shift in *pattern recognition*** *tasks*
- The resulting **tech** can be used **for security purposes** (e.g., biometric access, automatic surveillance) – and **to breach security** (new risks, new attack schemes)
- The ***pace is extremely high*** (new results are applied within months)
- Big question: **what *kind of society are we building*** around these opportunities?

Chapter 2 «Introduction to Applied Data Science»



Chapter 4 «Wie Maschinelles Lernen den Markt verändert»

About me:

- Prof. AI/ML, scientific director ZHAW digital
- Email: stdm@zhaw.ch
- Phone: +41 58 934 72 08
- Web: https://stdm.github.io/
- Twitter: @thilo_on_data
- LinkedIn: thilo-stadelmann

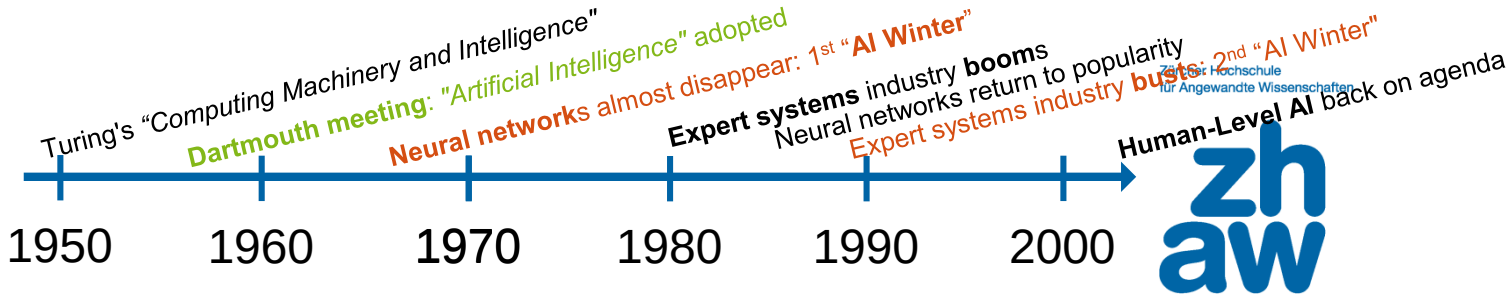
datalab www.zhaw.ch/datalab

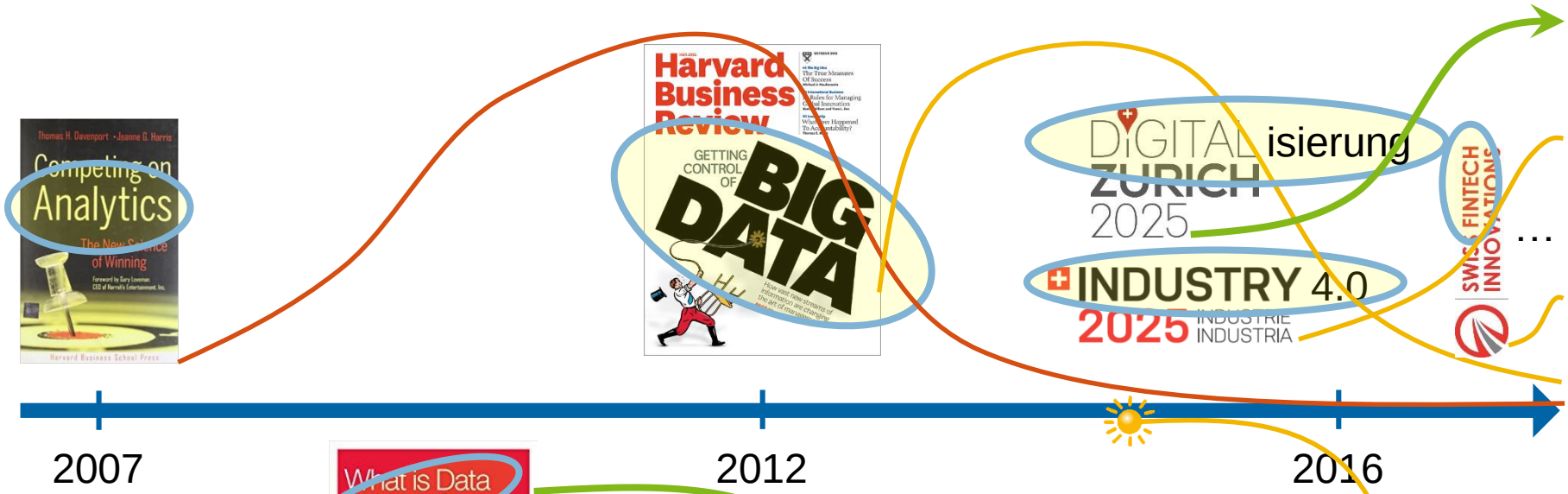
data service Swiss Alliance for Data-Intensive Services

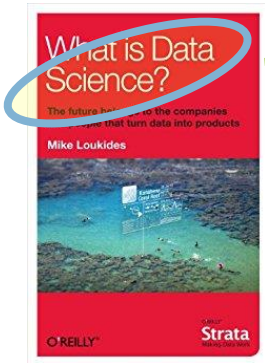
CLAIRE

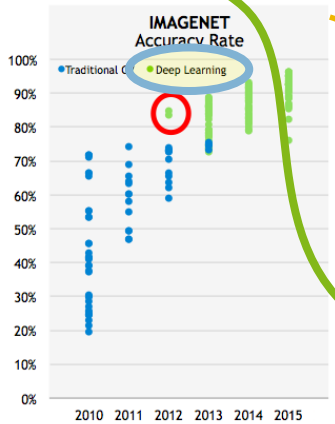
ANHANG

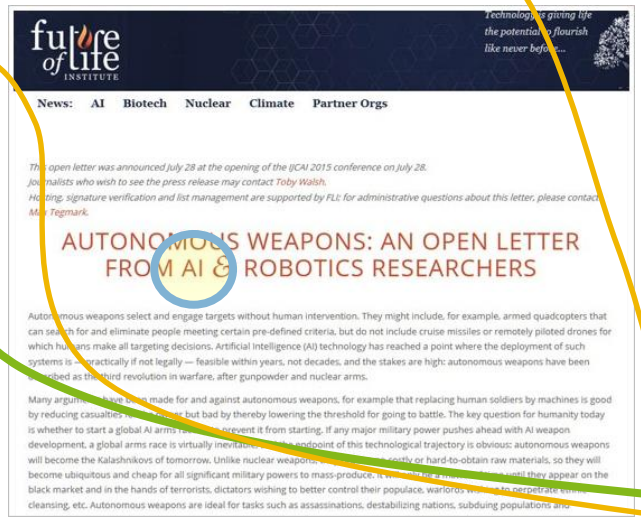
AI in context







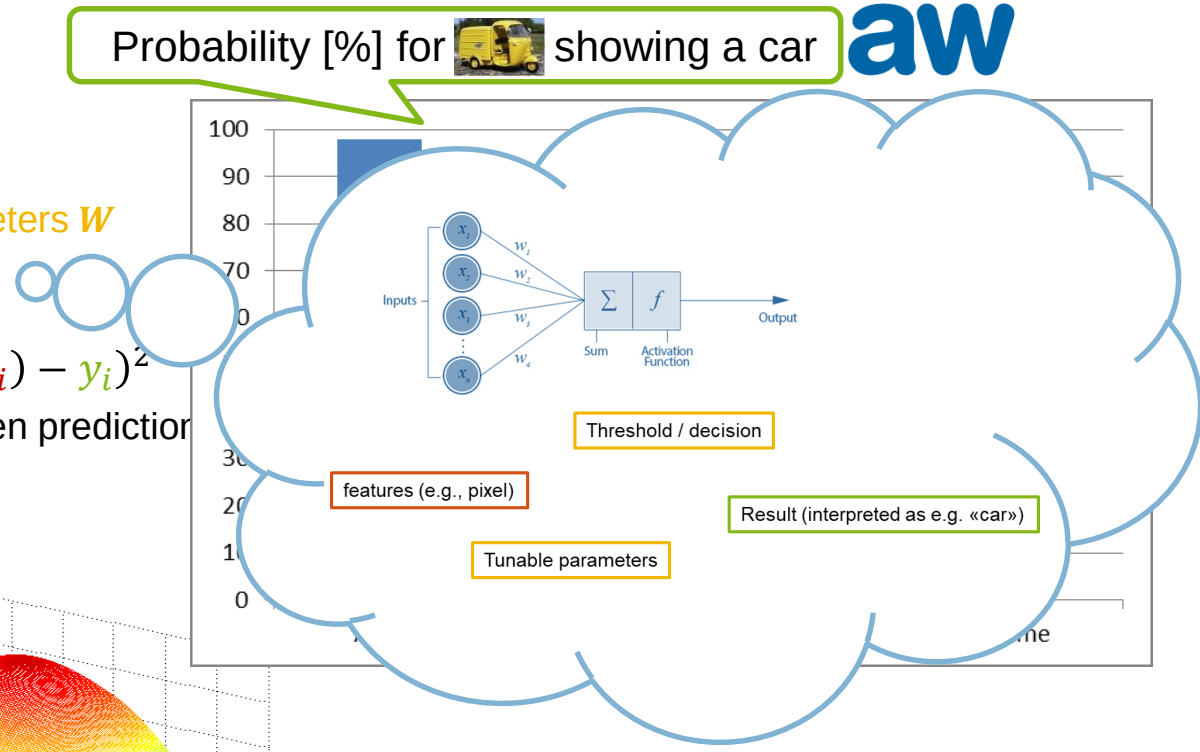


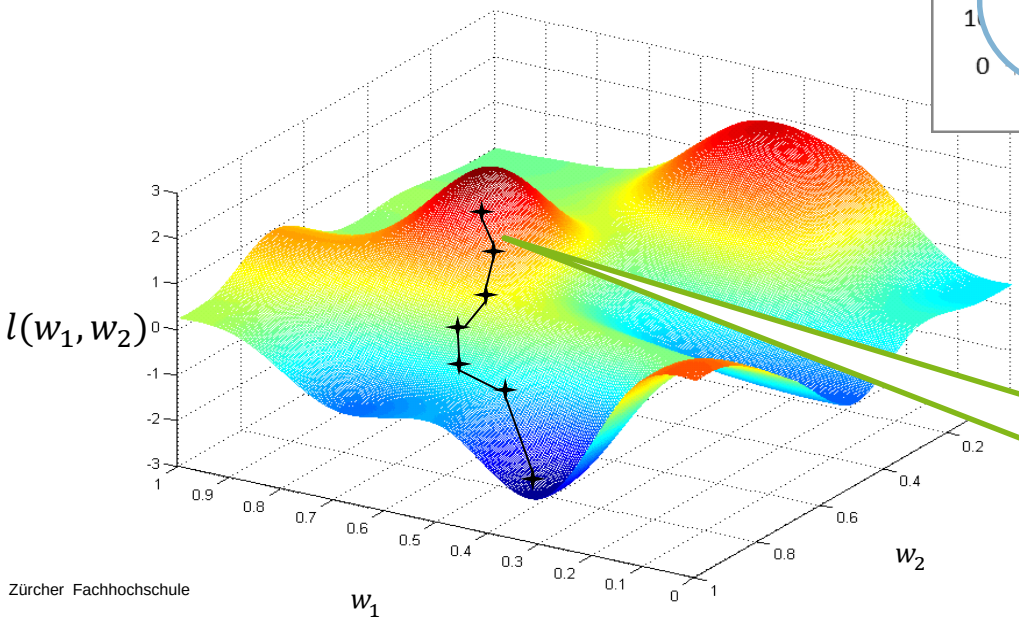


Search for optimal parameters of a function?

- Our artificial neural net: $f_{\boldsymbol{W}}(x) = y$
  with image x, ground truth y and parameters $\boldsymbol{W}$
  ($\boldsymbol{W} = \{w_1, w_2\}$ initialized at random)

- Error measure: $l(\boldsymbol{W}) = \frac{1}{N}\sum_{i=1}^{N}(f_{\boldsymbol{W}}(x_i) - y_i)^2$
  Average of (quadratic) difference between prediction and ground truth («loss»)



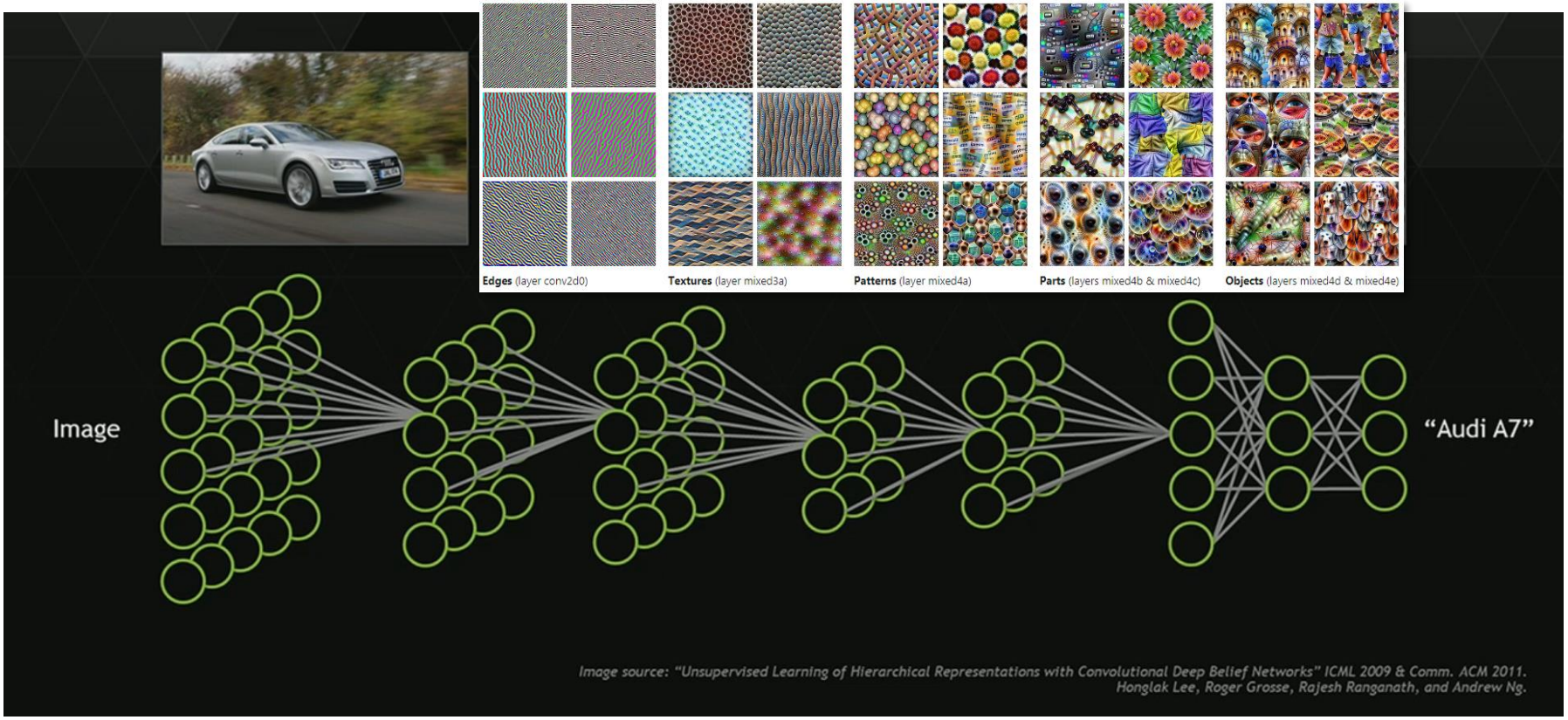


← error landscape

Method: iterative change of parameters of f in the direction of the steepest descent of J

What does the neural network «see»?

Hierarchy of more complex features



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

Potential interventions

Learning **from** and with the **cybersecurity** community

- Explore and potentially implement **red teaming**, **formal verification**, **responsible disclosure** of AI vulnerabilities, **security tools**, and **secure hardware**

Exploring **different openness** models

- **Reimagine norms** and **institutions** around the openness of research
- **Pre-publication risk assessment**, central **access licensing** models, sharing regimes that **favor safety** and security, and other **lessons from other dual-use technologies**

Promoting a **culture of responsibility**

- Highlight **education**, **ethic**al statements & standards, framings, norms, and **expectations**

Developing **technological and policy** solutions

- Strive for **legislative** and **regulatory responses**
- This requires **attention and action from** AI **researchers** and **companies**, **legislators**, **civil servants**, regulators, security researchers and **educators**
- The challenge is daunting, and the stakes are high

Team AI/ML: Overview (cp. https://stdm.github.io/research/)

ZHAW School of Engineering, Winterthur, Switzerland

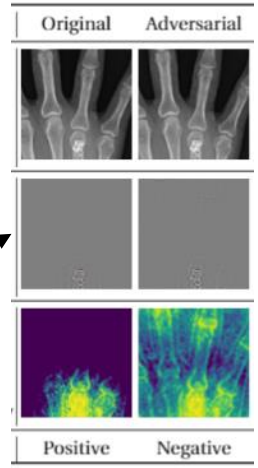
Zürcher Hochschule für Angewandte Wissenschaften

Machine learning-based Pattern Recognition

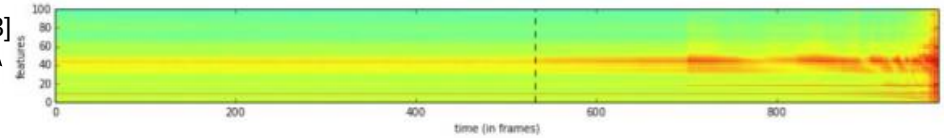
- Robust applications
- Biometrics
- Document Analysis
- Learning to act

[1]

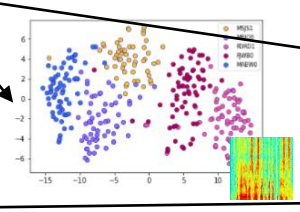
[2]



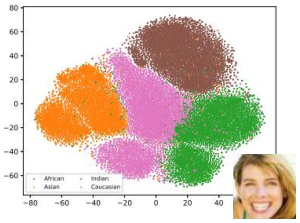
[3]



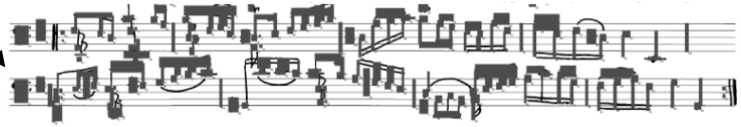
[4]



[5]

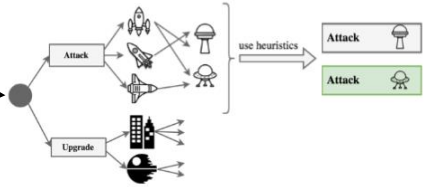


[6]

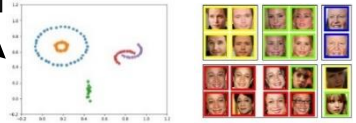


[7]

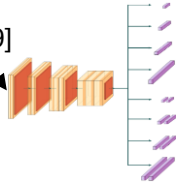
[1]



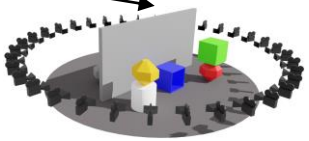
[8]



[9]



[10]



Outlook: Current projects & work in progress

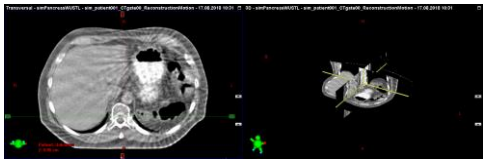
- Medical image analysis: learning to reduce motion artifacts in 3D CT scans

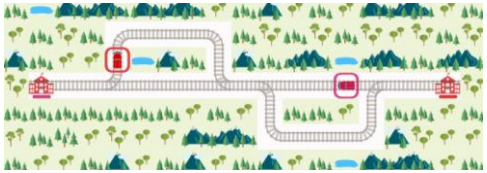
- Learning an artificial communication language for multi-agent reinforcement learning in logistics
  (notable rank in Flatland 2019 competition, best poster award [11])

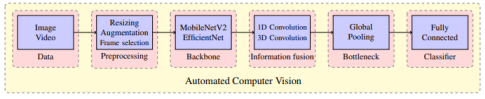
- Automated deep learning
  (top rank in AutoDL 2020 challenge [9])

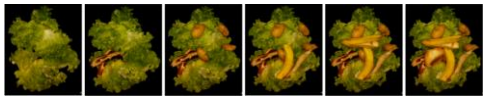
- Learning to segment and classify food waste in professional kitchens under adversarial conditions

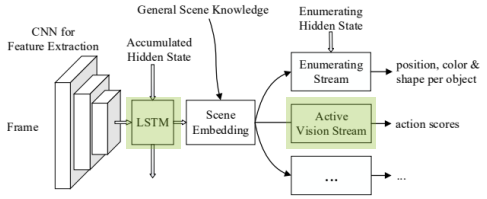
- Improving robotic vision through active vision and combined supervised and reinforcement learning
  (Dr. Waldemar Jucker Award 2020 [10])











References

1. Thilo Stadelmann, Mohammadreza Amirian, Ismail Arabaci, Marek Arnold, Gilbert François Duivesteijn, Ismail Elezi, Melanie Geiger, Stefan Lörwald, Benjamin Bruno Meier, Katharina Rombach, and Lukas Tuggener. "**Deep Learning in the Wild**". In: Proceedings of the 8th IAPR TC 3 Workshop on Artificial Neural Networks for Pattern Recognition (**ANNPR'18**), Springer, LNAI 11081, pp. 17-38, Siena, Italy, September 19-21, 2018.
2. Mohammadreza Amirian, Friedhelm Schwenker, and Thilo Stadelmann. "**Trace and Detect Adversarial Attacks on CNNs using Feature Response Maps**". In: Proceedings of the 8th IAPR TC 3 Workshop on Artificial Neural Networks for Pattern Recognition (**ANNPR'18**), Springer, LNAI 11081, pp. 346-358, Siena, Italy, September 19-21, 2018.
3. Thilo Stadelmann, Vasily Tolkachev, Beate Sick, Jan Stampfli, and Oliver Dürr. "**Beyond ImageNet - Deep Learning in Industrial Practice**". In: Martin Braschler, Thilo Stadelmann, and Kurt Stockinger (Editors). "Applied Data Science - Lessons Learned for the Data-Driven Business". **Springer**, 2019.
4. Thilo Stadelmann, Sebastian Glinski-Haefeli, Patrick Gerber, and Oliver Dürr. "**Capturing Suprasegmental Features of a Voice with RNNs for Improved Speaker Clustering**". In: Proceedings of the 8th IAPR TC 3 Workshop on Artificial Neural Networks for Pattern Recognition (**ANNPR'18**), Springer, LNAI 11081, pp. 333-345, Siena, Italy, September 19-21, 2018.
5. Stefan Glüge, Mohammadreza Amirian, Dandolo Flumini, and Thilo Stadelmann. "**How (Not) to Measure Bias in Face Recognition Networks**". In: Proceedings of the 9th IAPR TC 3 Workshop on Artificial Neural Networks for Pattern Recognition (**ANNPR'20**), Springer, LNAI, Winterthur, Switzerland, September 02-04, 2020.
6. Lukas Tuggener, Yvan Putra Satyawan, Alexander Pacha, Jürgen Schmidhuber, and Thilo Stadelmann. "**The DeepScoresV2 Dataset and Benchmark for Music Object Detection**". In: Proceedings of the 25th International Conference on Pattern Recognition (**ICPR'20**), IAPR, Milan, Italy, January 10-15 (online), 2021.
7. Benjamin Meier, Thilo Stadelmann, Jan Stampfli, Marek Arnold, and Mark Cieliebak. "**Fully convolutional neural networks for newspaper article segmentation**". In: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (**ICDAR'17**). 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto Japan, November 13-15, 2017. Kyoto, Japan: CPS.
8. Benjamin Bruno Meier, Ismail Elezi, Mohammadreza Amirian, Oliver Dürr, and Thilo Stadelmann. "**Learning Neural Models for End-to-End Clustering**". In: Proceedings of the 8th IAPR TC 3 Workshop on Artificial Neural Networks for Pattern Recognition (**ANNPR'18**), Springer, LNAI 11081, pp. 126-138, Siena, Italy, September 19-21, 2018.
9. Lukas Tuggener, Mohammadreza Amirian, Fernando Benites, Pius von Däniken, Prakhar Gupta, Frank-Peter Schilling, and Thilo Stadelmann. "**Design Patterns for Resource-Constrained Automated Deep-Learning Methods**". **AI** section "Intelligent Systems: Theory and Applications" 1(4):510-538, MDPI, Basel, Switzerland, Novemer 06, 2020.
10. Dano Roost, Ralph Meier, Giovanni Toffetti Carughi, and Thilo Stadelmann. "**Combining Reinforcement Learning with Supervised Deep Learning for Neural Active Scene Understanding**". In: Proceedings of the Active Vision and Perception in Human(-Robot) Collaboration Workshop at IEEE RO-MAN 2020 (**AVHRC'20**), online, August 31, 2020.
11. Dano Roost, Ralph Meier, Stephan Huschauer, Erik Nygren, Adrian Egli, Andreas Weiler, and Thilo Stadelmann. "**Improving Sample Efficiency and Multi-Agent Communication in RL-based Train Rescheduling**". In: Proceedings of the 7th Swiss Conference on Data Science (**SDS'20**), Lucerne, Switzerland, June 26, 2020. IEEE.