Zürcher Hochschule
für Angewandte Wissenschaften

What have the Front Lines of Deep Learning to do with the Future of Humanity?

Inaugural lecture, July 21, 2018

Thilo Stadelmann

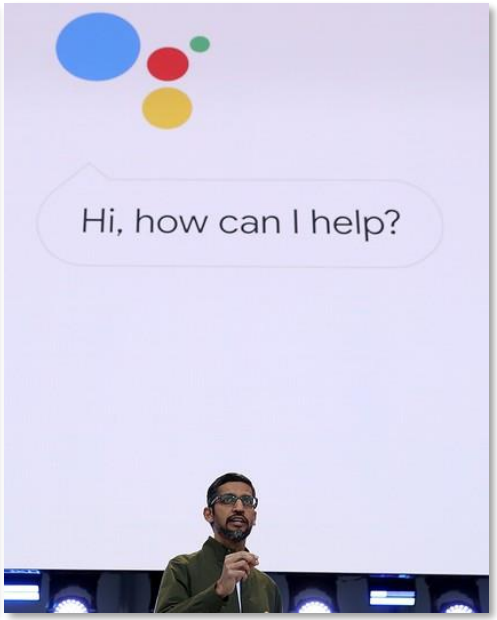
Swiss Alliance for
Data-Intensive Services

swiss group for artificial intelligence
and cognitive science

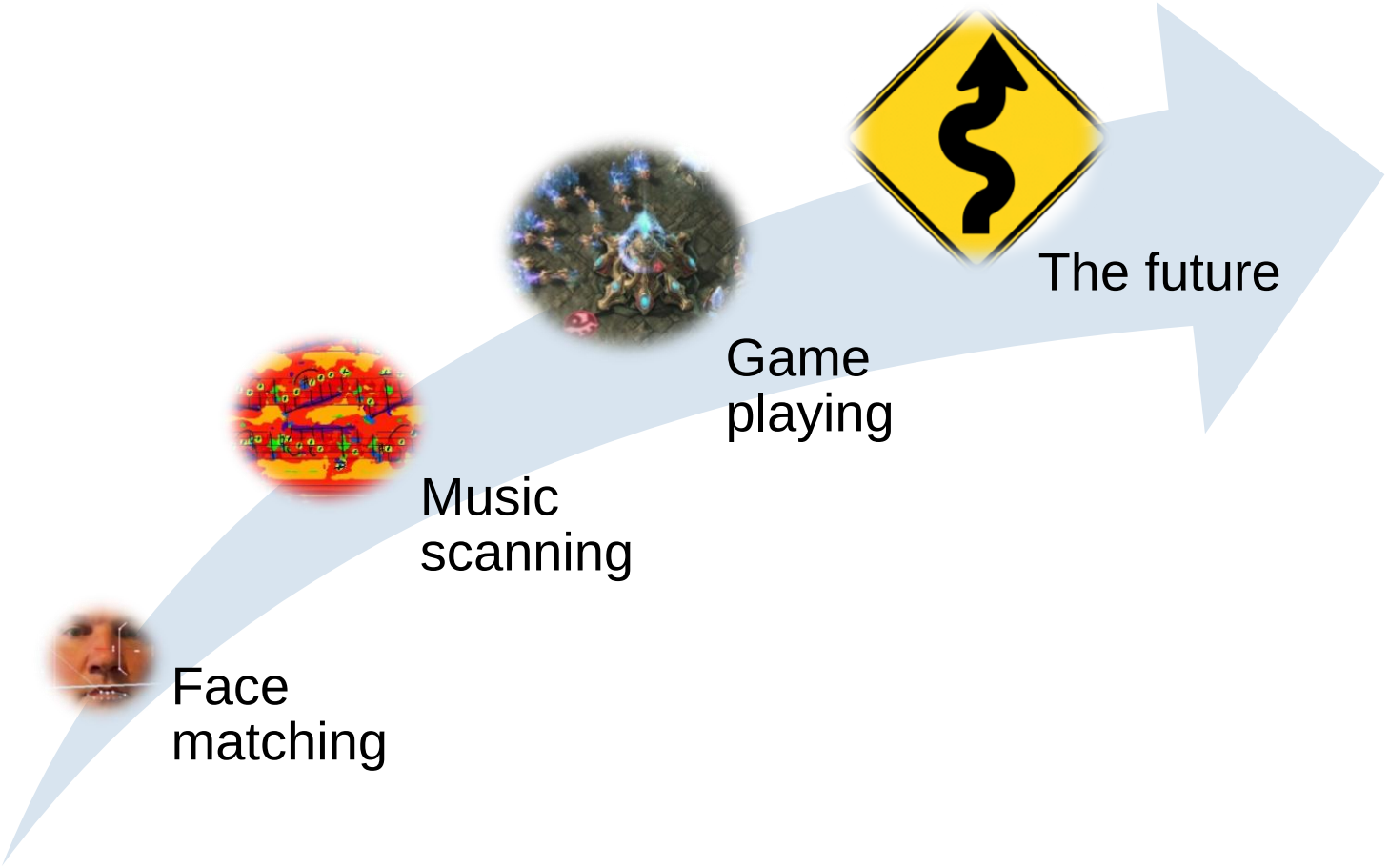
datalab
www.zhaw.ch/datalab

Zürcher Hochschule
für Angewandte Wissenschaften

Why?

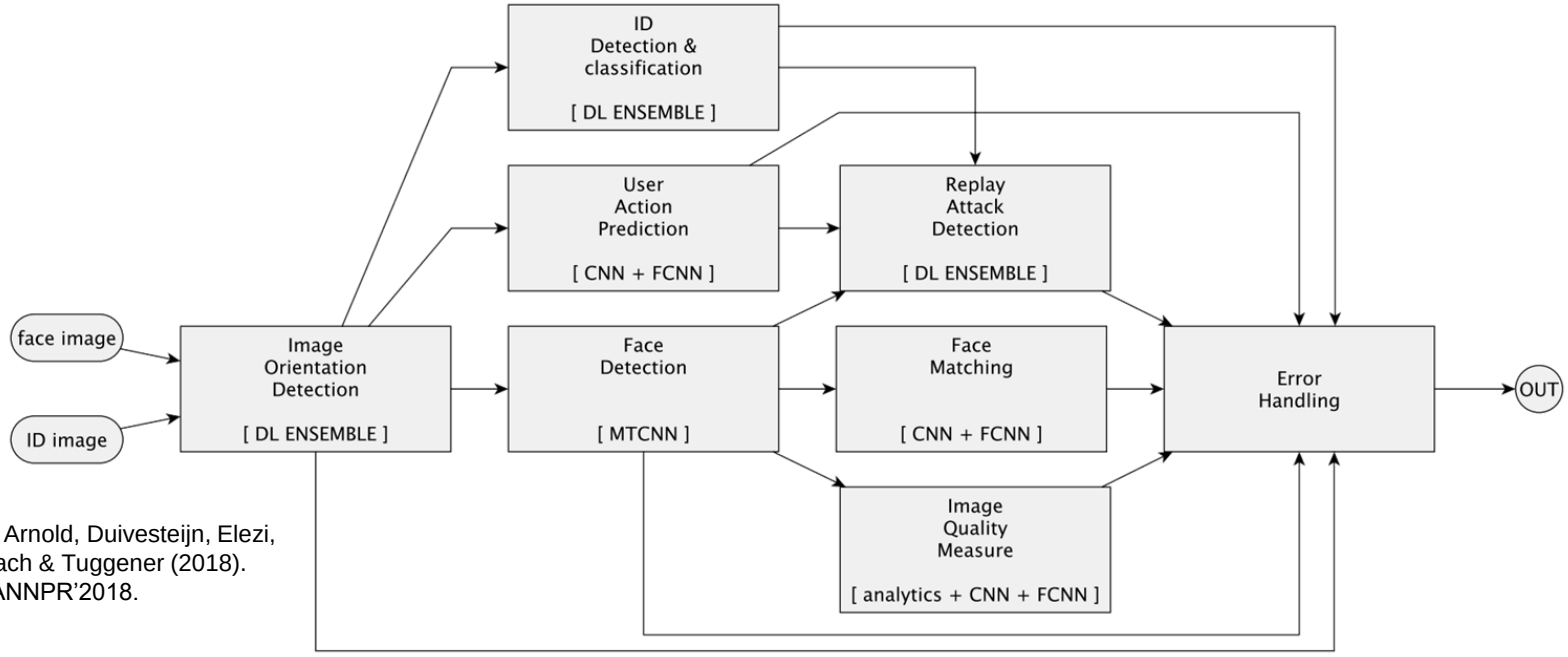


Agenda



Face matching

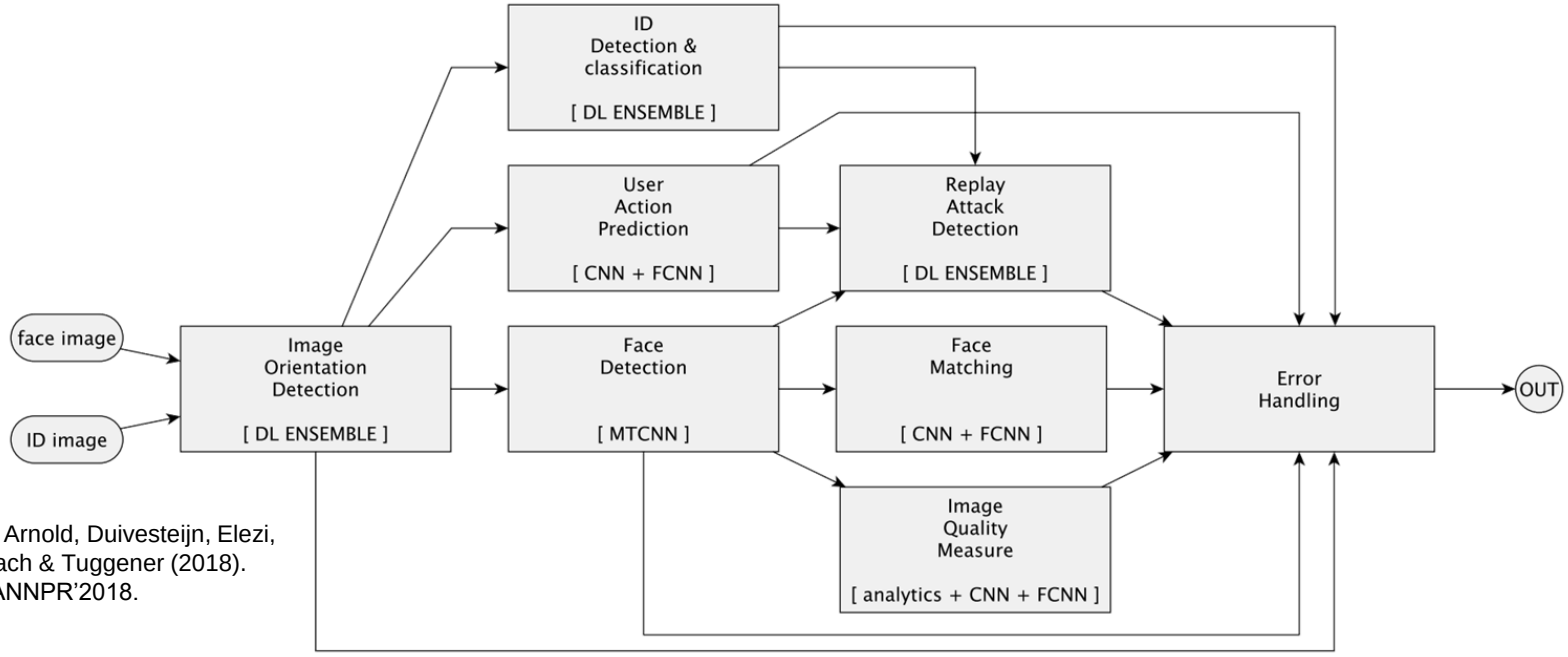
Face matching – challenges & solutions



Stadelmann, Amirian, Arabaci, Arnold, Duivesteijn, Elezi, Geiger, Lörwald, Meier, Rombach & Tuggener (2018). *«Deep Learning in the Wild»*. ANNPR'2018.

Face matching – challenges & solutions

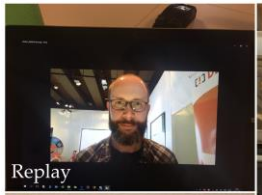


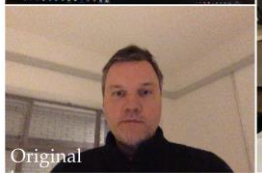


Stadelmann, Amirian, Arabaci, Arnold, Duivesteijn, Elezi, Geiger, Lörwald, Meier, Rombach & Tuggener (2018). *«Deep Learning in the Wild»*. ANNPR'2018.

Face matching – challenges & solutions

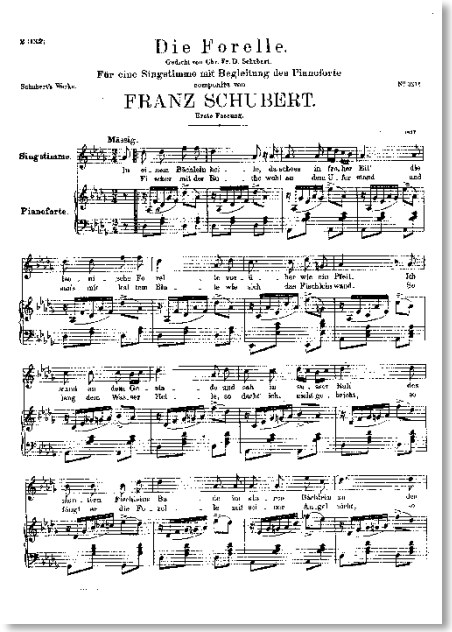


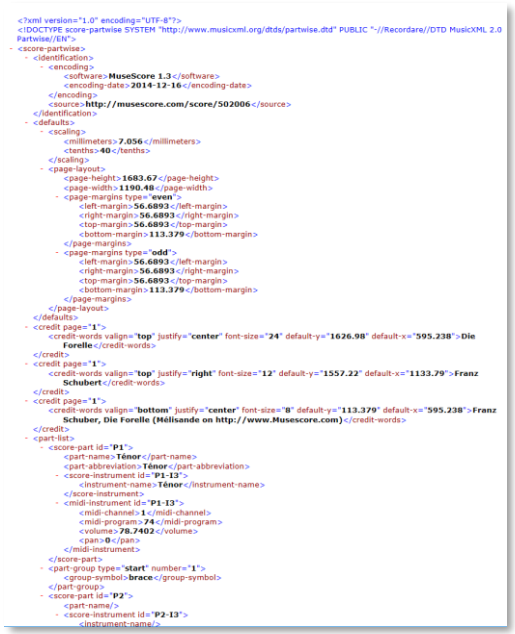


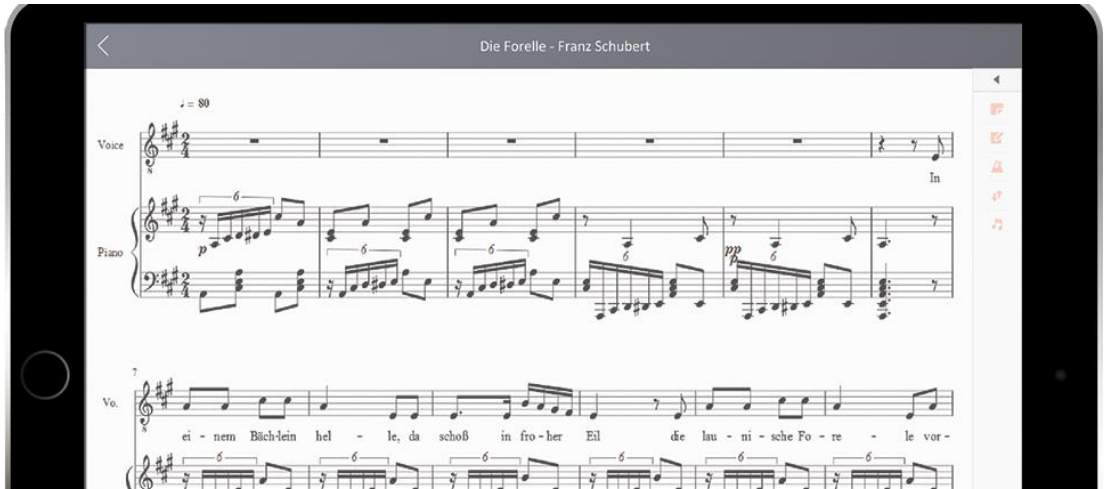


Stadelmann, Amirian, Arabaci, Arnold, Duivesteijn, Elezi, Geiger, Lörwald, Meier, Rombach & Tuggener (2018). *«Deep Learning in the Wild»*. ANNPR'2018.

Music scanning







Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

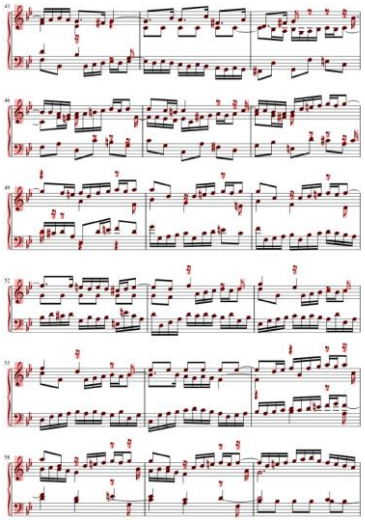
SCOREPAD

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

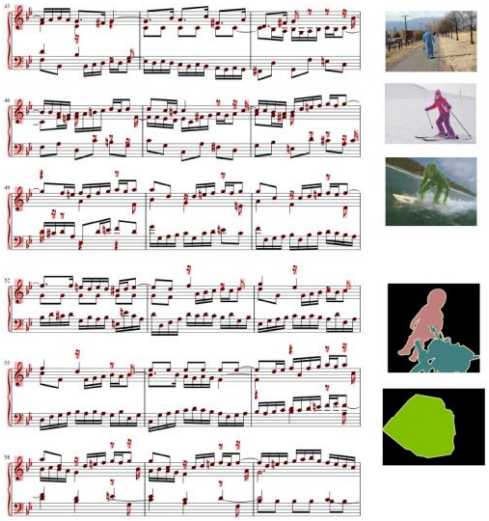
Innosuisse – Swiss Innovation Agency

Music scanning – challenges & solutions



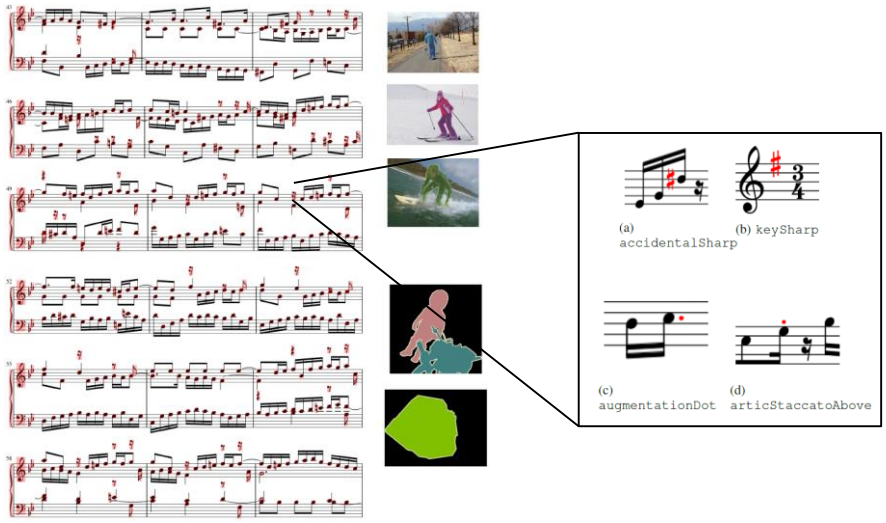
Tuggener, Elezi, Schmidhuber, Pelillo & Stadelmann (2018). *«DeepScores – A Dataset for Segmentation, Detection and Classification of Tiny Objects»*. ICPR'2018.

Music scanning – challenges & solutions



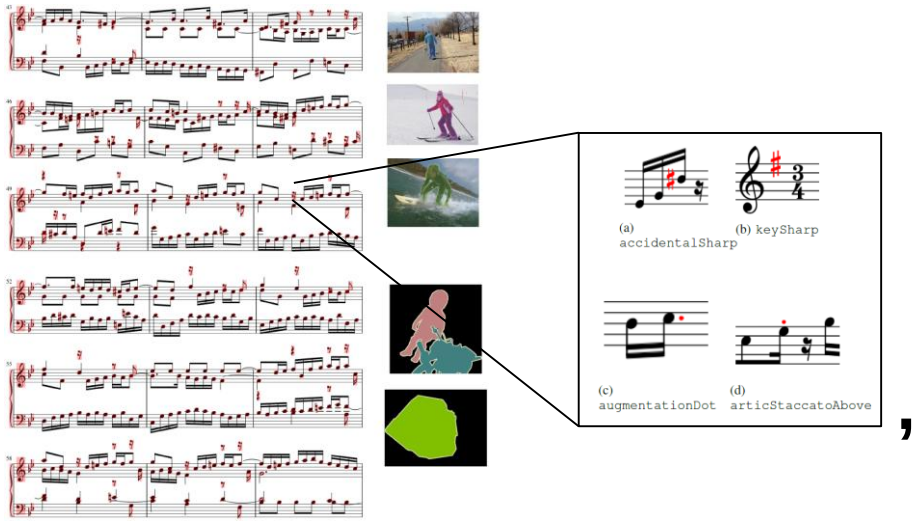
Tuggener, Elezi, Schmidhuber, Pelillo & Stadelmann (2018). *«DeepScores – A Dataset for Segmentation, Detection and Classification of Tiny Objects»*. ICPR'2018.

Music scanning – challenges & solutions

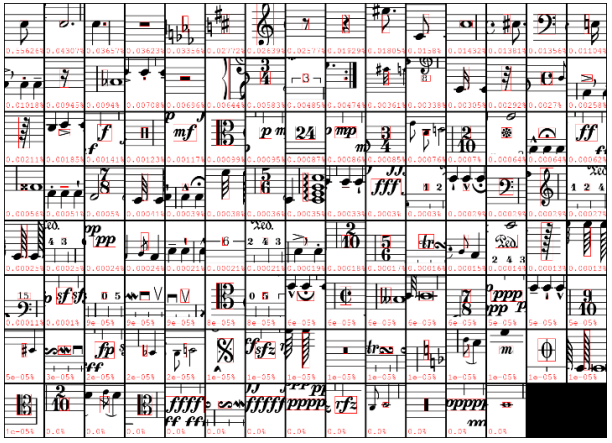


Tuggener, Elezi, Schmidhuber, Pelillo & Stadelmann (2018). *«DeepScores – A Dataset for Segmentation, Detection and Classification of Tiny Objects»*. ICPR'2018.

Music scanning – challenges & solutions

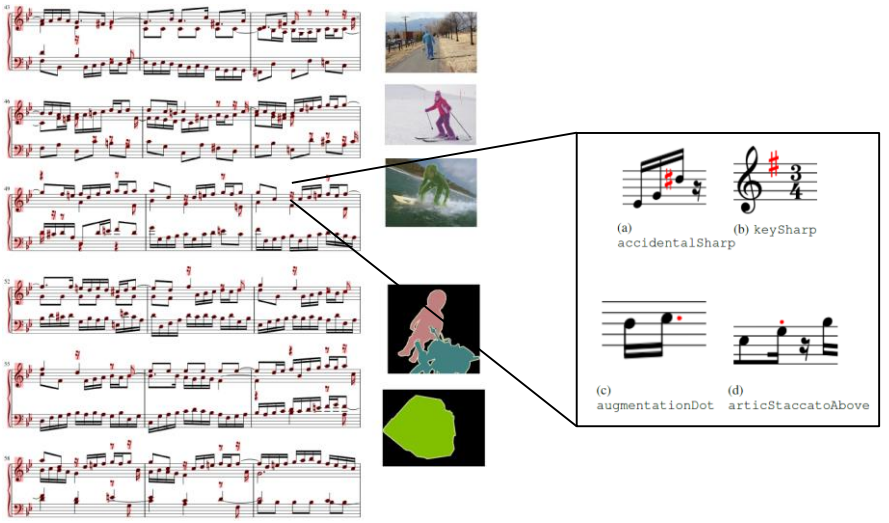


,

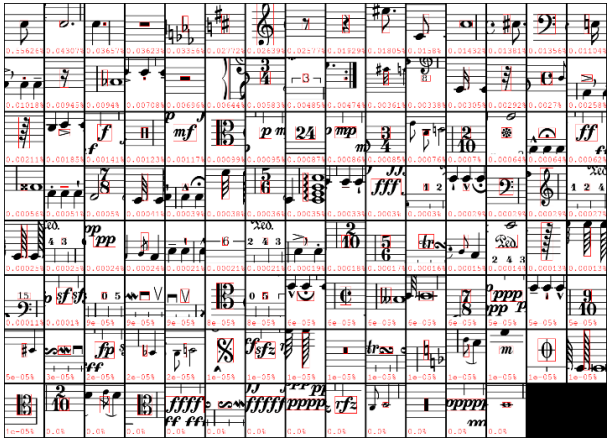


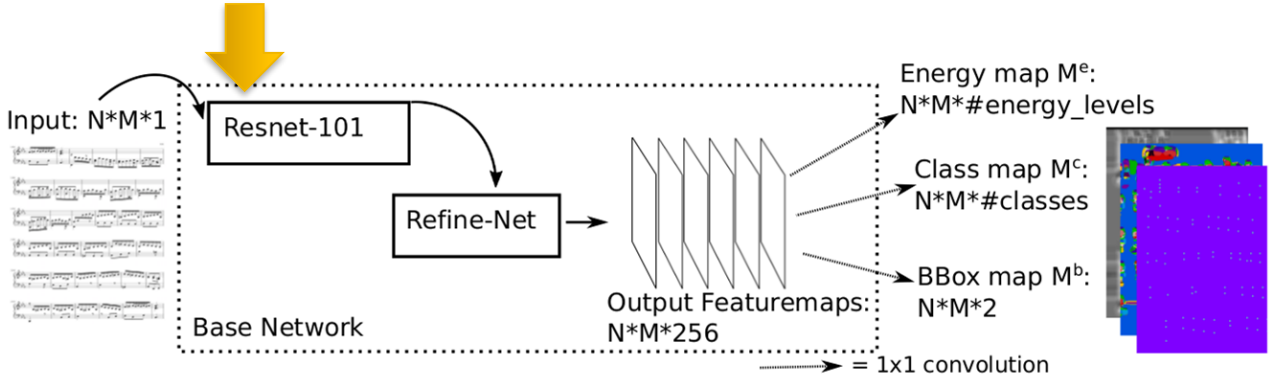
Tuggener, Elezi, Schmidhuber, Pelillo & Stadelmann (2018). *«DeepScores – A Dataset for Segmentation, Detection and Classification of Tiny Objects»*. ICPR'2018.

Music scanning – challenges & solutions



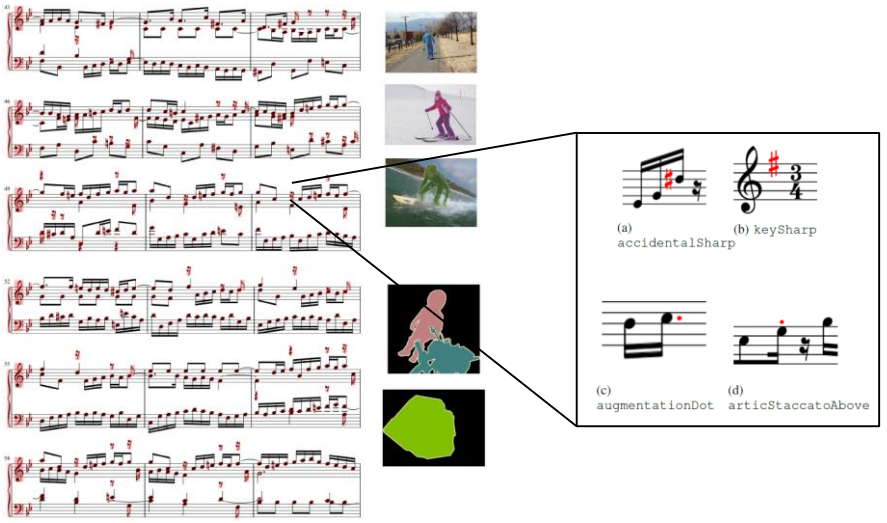
,



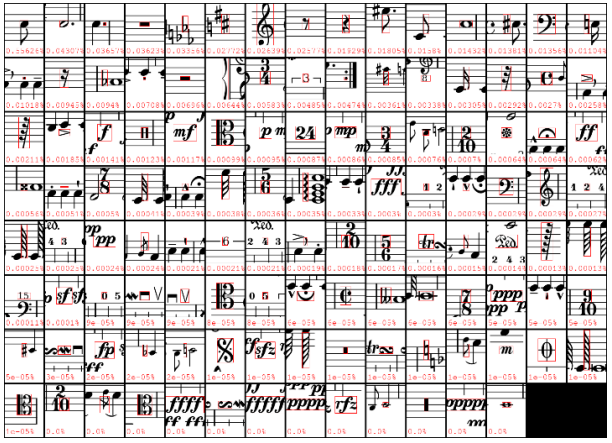


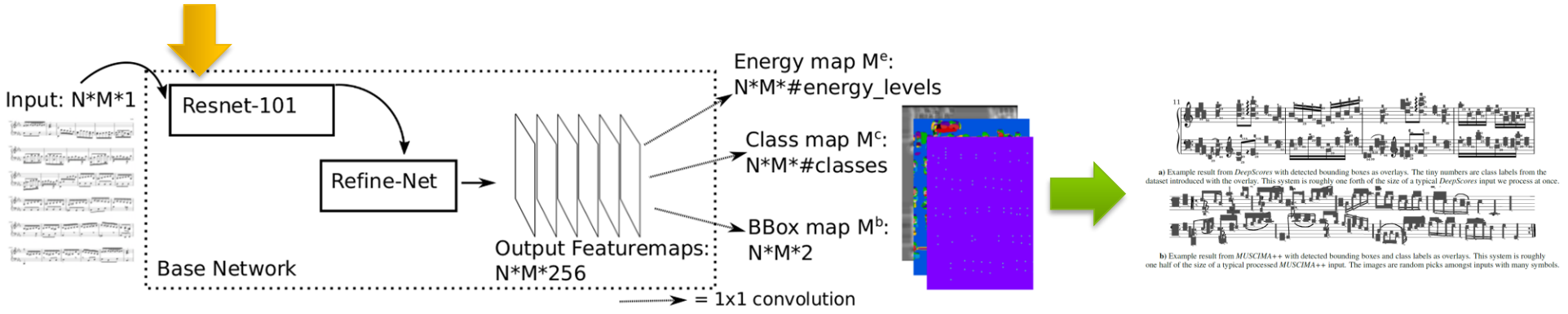
Tuggener, Elezi, Schmidhuber, Pelillo & Stadelmann (2018). *«DeepScores – A Dataset for Segmentation, Detection and Classification of Tiny Objects»*. ICPR'2018.
Tuggener, Elezi, Schmidhuber & Stadelmann (2018). *«Deep Watershed Detector for Music Object Recognition»*. ISMIR'2018.

Music scanning – challenges & solutions



,





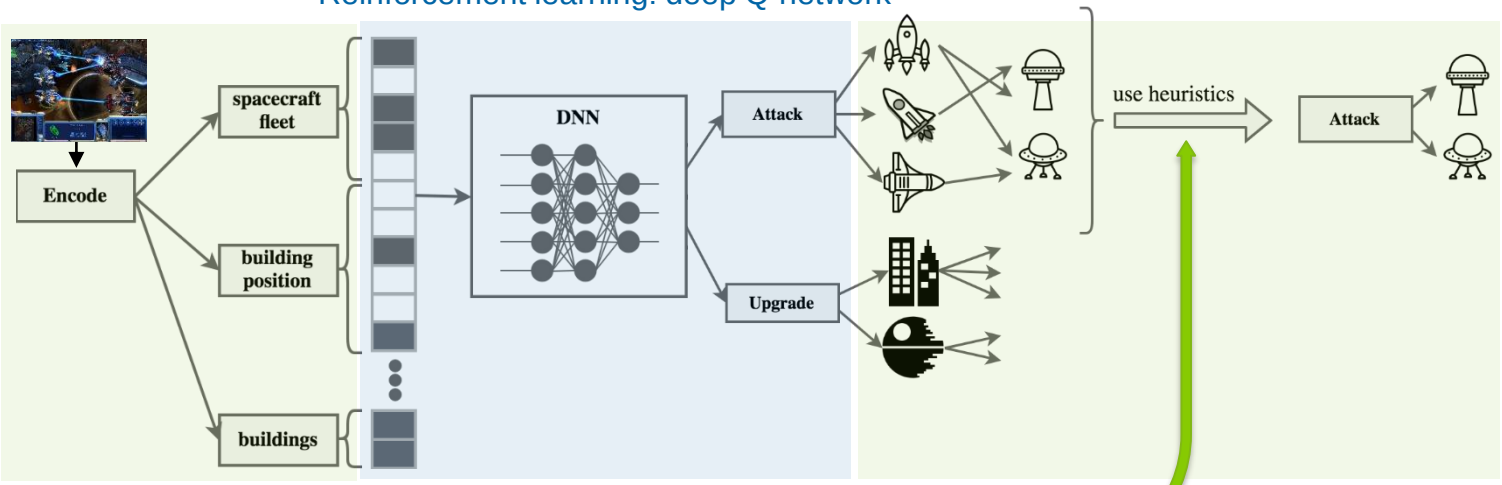
Tuggener, Elezi, Schmidhuber, Pelillo & Stadelmann (2018). *«DeepScores – A Dataset for Segmentation, Detection and Classification of Tiny Objects»*. ICPR'2018.
Tuggener, Elezi, Schmidhuber & Stadelmann (2018). *«Deep Watershed Detector for Music Object Recognition»*. ISMIR'2018.

Game playing

(symbolic figure)

Game playing – challenges & solutions

Reinforcement learning: deep Q network



Large discrete action space → use heuristic

- makes exploration difficult
- elongates training time

Delayed and sparse reward → do reward shaping

- sequence of actions crucial to get a reward

Distance encoding → use reference points

Transfer Learning → difficult: more complex environment needs other action sequence

Stadelmann, Duivesteijn, Amirian, Tuggener, Elezi, Geiger & Rombach (2018). *«Deep Learning in the Wild»*. ANNPR'2018.

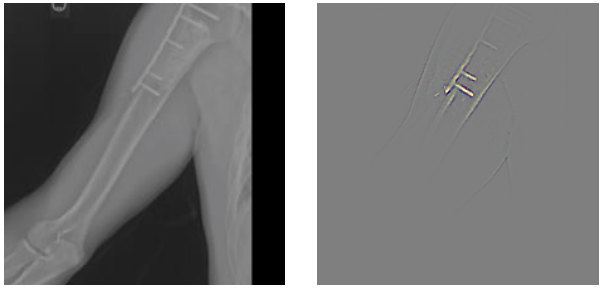
Lessons learned: key of model interpretability

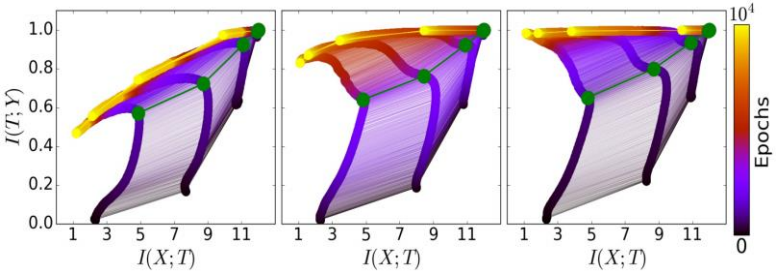
Interpretability is required.

- Helps the developer in «debugging», needed by the user to trust
  → visualizations of learned features, training process, learning curves etc. should be «always on»

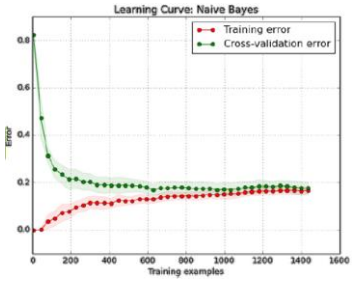
negative X-ray

positive X-ray

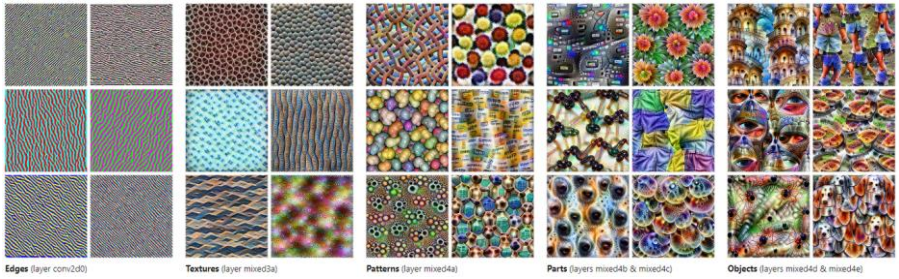




DNN training on the Information Plane



a learning curve



feature visualization

Schwartz-Ziv & Tishby (2017). *«Opening the Black Box of Deep Neural Networks via Information»*.
https://distill.pub/2017/feature-visualization/, https://stanfordmlgroup.github.io/competitions/mura/
Stadelmann, Duivesteijn, Amirian, Tuggener, Elezi, Geiger & Rombach (2018). *«Deep Learning in the Wild»*. ANNPR'2018.

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

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

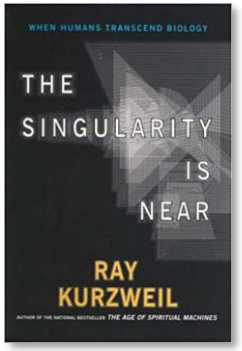
1. **Overestimating and underestimating**: *«We tend to overestimate the effect of a technology in the short run and underestimate the effect in the long run.»*
2. **Imagining magic**: *«Any sufficiently advanced technology is indistinguishable from magic.»*
3. **Performance versus competence**: *«People 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).

The vision of Ray Kurzweil
Google, Inc.

The **singularity** is near

- Superintelligence will enhance human life



“By the time we get to the 2040s, we'll be able to multiply human intelligence a billionfold. That will be a profound change that's singular in nature. Computers are going to keep getting smaller and smaller. Ultimately, they will go inside our bodies and brains and make us healthier, make us smarter.”

Ray Kurzweil

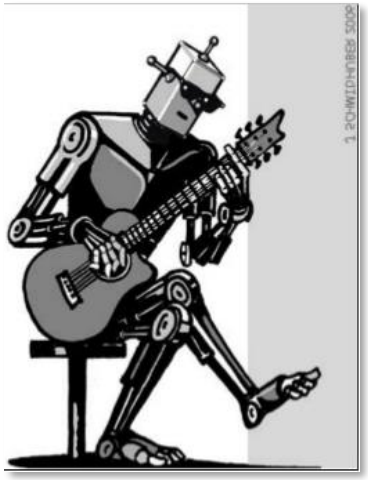
intelligentHQ

The vision of Jürgen Schmidhuber

IDSIA, Lugano, Switzerland

Autonomous robots will

- Be **curious** about human life (rather than hostile)
- Be enabled by **artificially curiosity and LSTM** neural nets
- **Colonize space** on the look for resources to reproduce
- Surface in ca. 2030

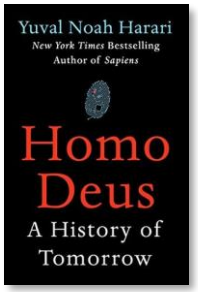


The vision of Yuval Noah Harari

Hebrew University of Jerusalem

Humans can become **godlike**

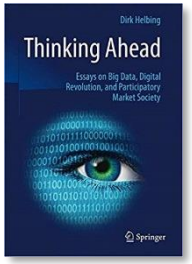
- Humans will upgrade themselves in 3 ways: **biological engineering**, **cyborg** engineering and **robots**
- A new class of people will emerge by 2050: the **useless class** (not just unemployed, but unemployable)
- The most important skill in life will be **learning to learn**: reinvent yourself, again and again until you die to stay out of the useless class
- Computers **function very differently from humans**, and it seems unlikely that computers will become human-like any time soon; however, **intelligence is decoupling from consciousness**
- AI and biotechnology lead to **most powerful narratives** that enable humans to collaborate more effectively and actually **change reality**



The vision of Dirk Helbing
ETH Zurich

Society 4.0

- **Planetary nervous system**: a smartphone app enabling users to share data to achieve scientific and social goals and lay the groundwork for digital democracy
- **Living Earth Simulator**: a computing machine attempting "to model global-scale systems — economies, governments, cultural trends, epidemics, agriculture, technological developments, and more — using torrential data streams, sophisticated algorithms, and as much hardware as it takes“
- **Investment premium**: central banks give money equally to everybody, and they may invest into anybodies idea (not just consumption); negative interest rate regulates the system, and the global crowdfunding enables digital democracy



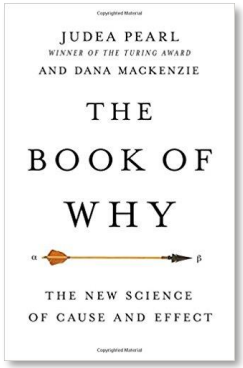
https://www.youtube.com/watch?v=SCcgVEAPJA0

The vision of Judea Pearl

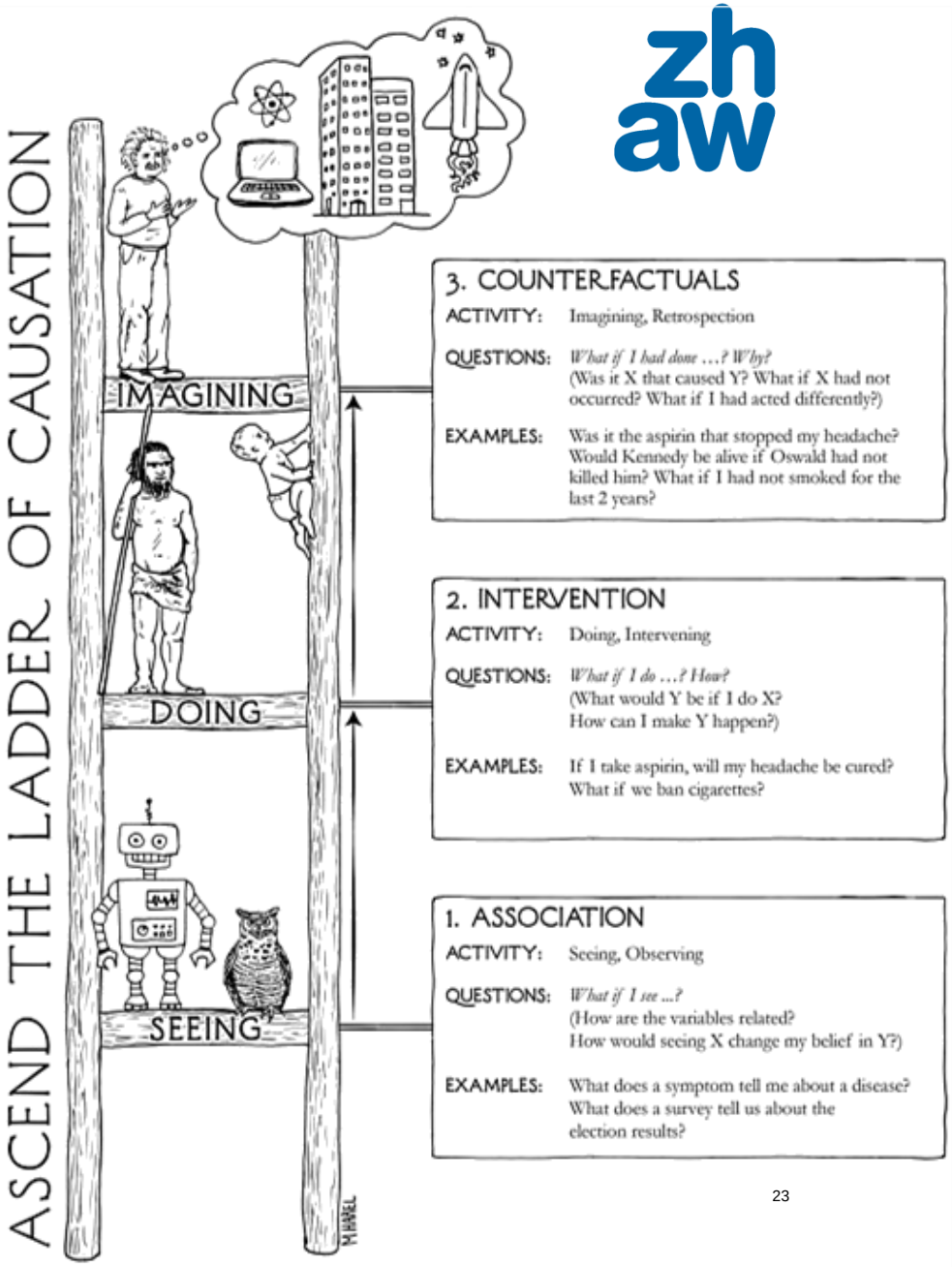
UCLA, Los Angeles, USA

From **causality** to intelligence:

- Machine without a causal model of reality cannot be expected to behave intelligently
- First step (by 2030): **conceptual models of reality will be programmed by humans**
- Next step: machines will postulate such models on their own and will verify and refine them based on empirical evidence



https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/

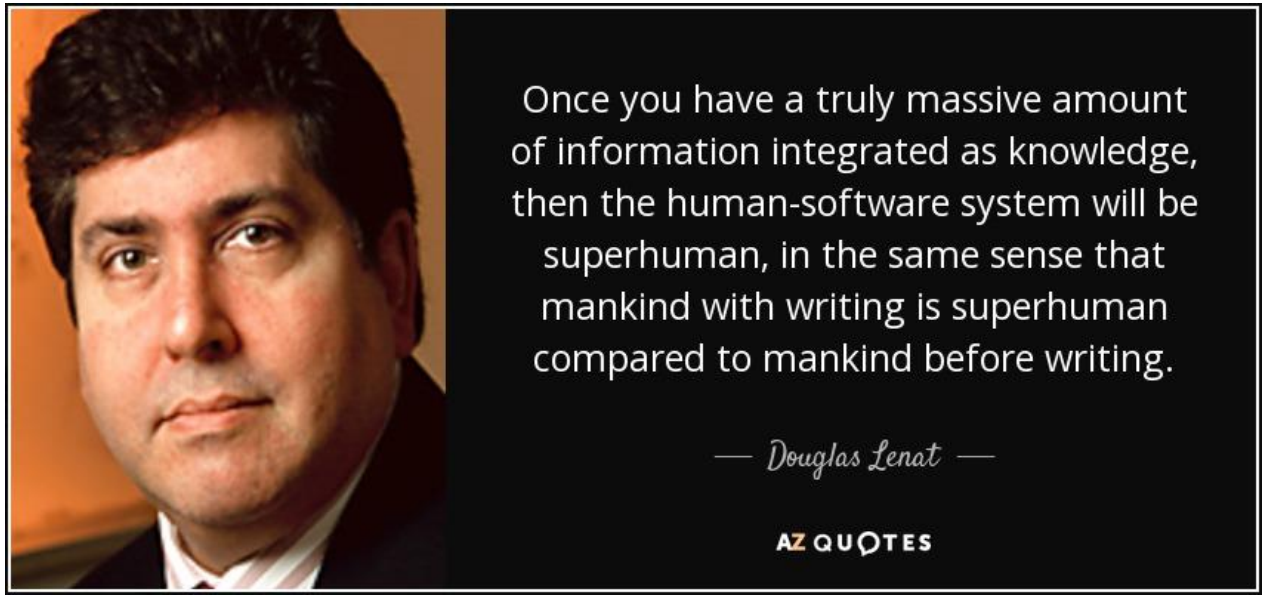


The vision of Doug Lenat

Cycorp Inc, Austin, Texas, USA

Symbolic AI, finally

- Persisted 35 years in building **Cyc**, a knowledge-based system (see V06b)
- Used 2’000 person years, 60 R&D people, 24 millions rules (not counting facts)
- Commercially successful since 2007, and again surfacing as (a) future of AI: *«Intelligence is ten million rules.»*

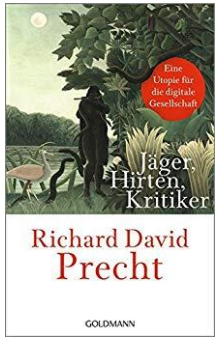


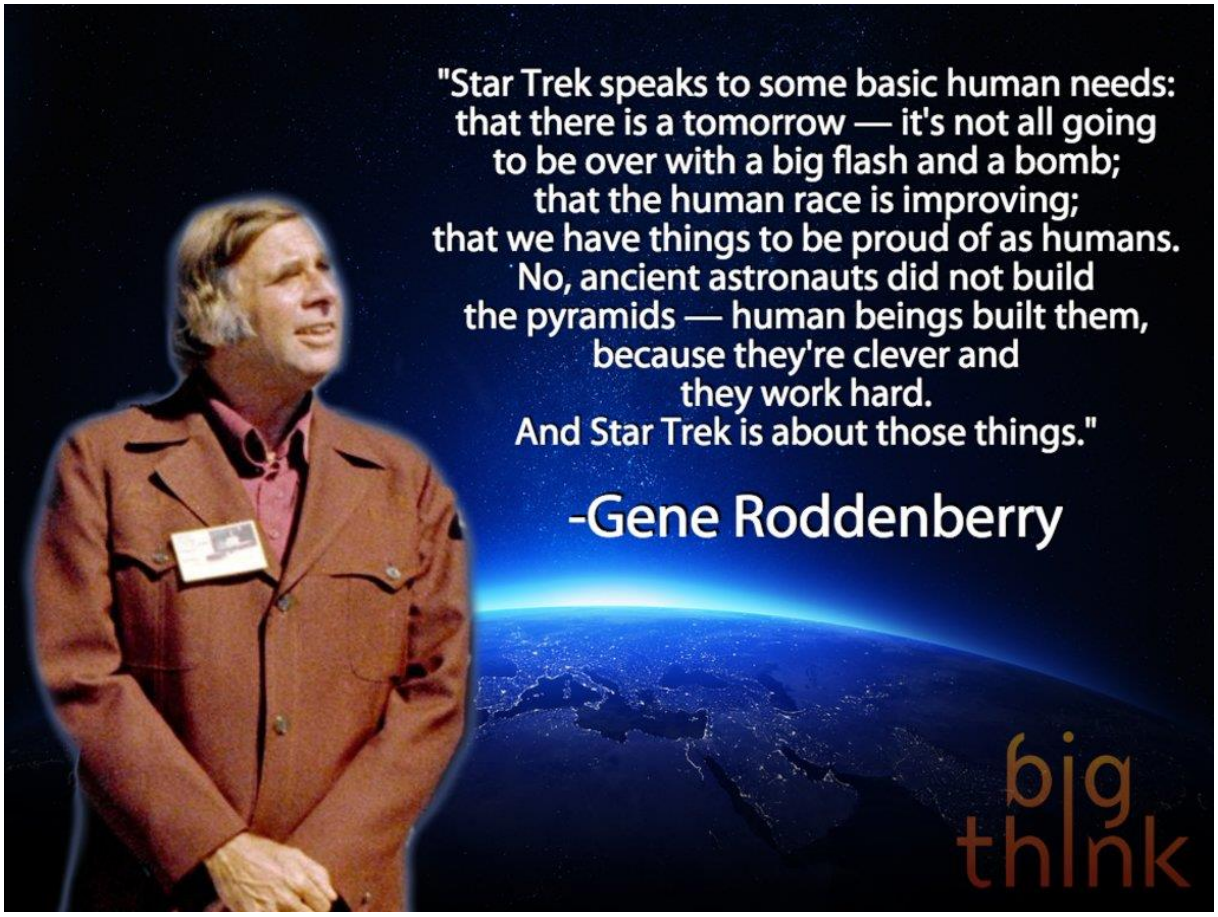
The vision of Gene Roddenberry

*„The acquisition of wealth is no longer a driving force in our lives. We **work to better ourselves and the rest** of humanity.“*

Captain Jean-Luc Picard

Compare Richard David Precht‘s *Jäger, Hirten, Kritiker: Eine Utopie für die digitale Gesellschaft.*



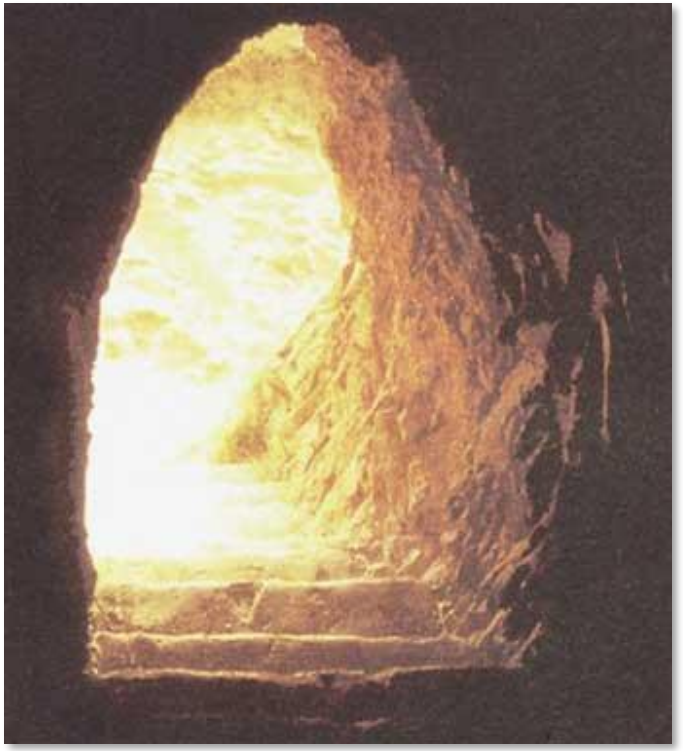


The vision of Jesus Christ

"And ye shall hear of wars and rumours of wars: ***see that ye be not troubled****."*
Matthew 24, 6

"A new commandment I give unto you, that ye ***love one another****."*
John 13, 34

"But ***rather seek ye*** *the kingdom of God* ***[things above]****; and all these things shall be added unto you."*
Luke 12, 31 [Colossians 3, 2]



Conclusions

- **AI systems** will **change** most of how **human societies** function **within this generation**
- This is **due to** the inherent properties of **efficacy**, **efficiency** and **scalability**
- It is **independent of larger progress** in performance / feasibility / AGI

- **AI is** a „**dual use**“ technology (can be used for good and bad) and thus warrants **responsible developers and deployers**
- Due to the potential to massively harmful use, **treating it with the same care** (and measures of protection) **as nuclear technology** is an option to ponder

- The **future has to be shaped** by humans – interdisciplinary, including experts, policy makers, citizens; the **time window is now**
- **Rather than fear**, uncertainty and doubt, clear **visions** of possible futures **help navigating** the current space of options

Zürcher Hochschule
für Angewandte Wissenschaften

APPENDIX

1. UNINTENDED THREATS THROUGH AI SYSTEMS

AI System (definition):

any technical system (software and/or hardware)
based on digital technology
that performs tasks *commonly thought* to require intelligence.

Algorithmic bias

Wikipedia: *"Algorithmic bias occurs when a computer system behaves in ways that reflects the implicit values of humans involved in that* ***data collection, selection, or use****."*

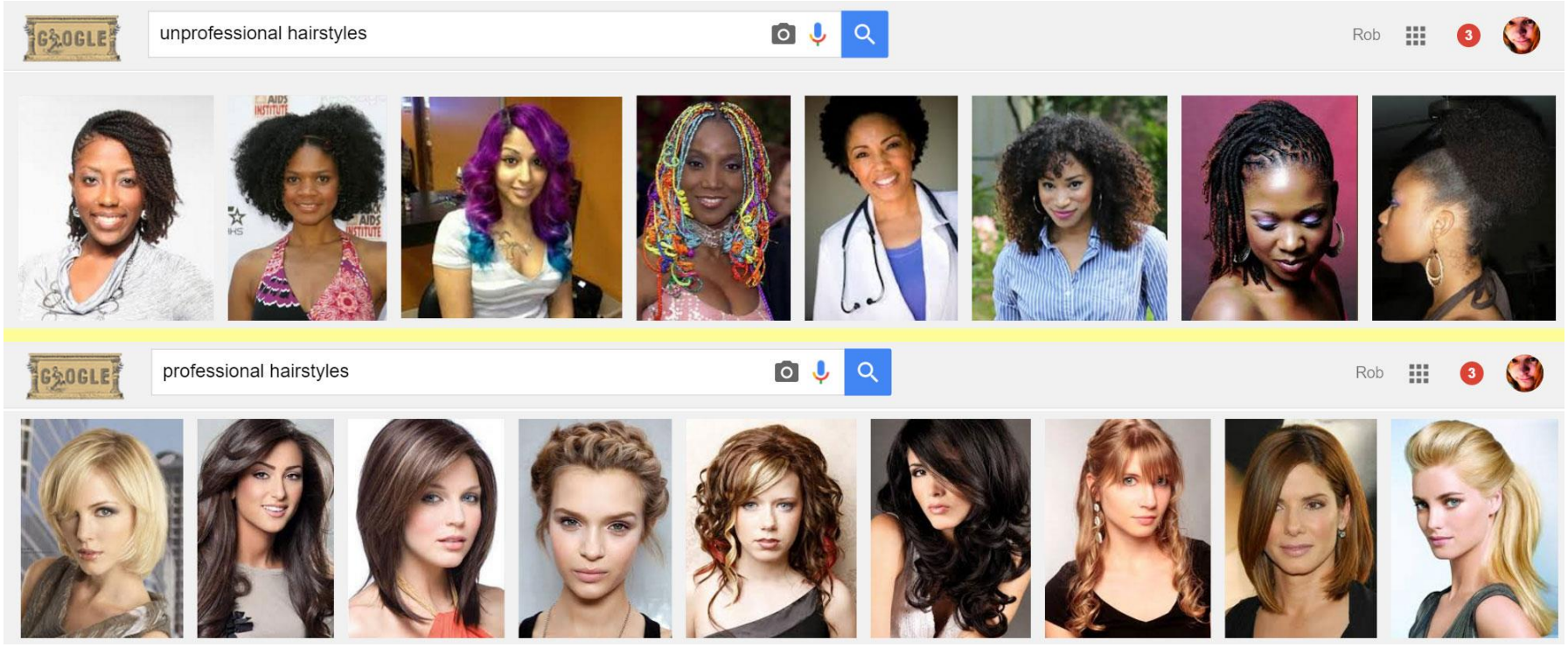
An established misnomer

- Usually it is not meant that the algorithm (code) is intentionally built to discriminate
- Rather, a (neutral) **learning algorithm** picked up our biases from the training data

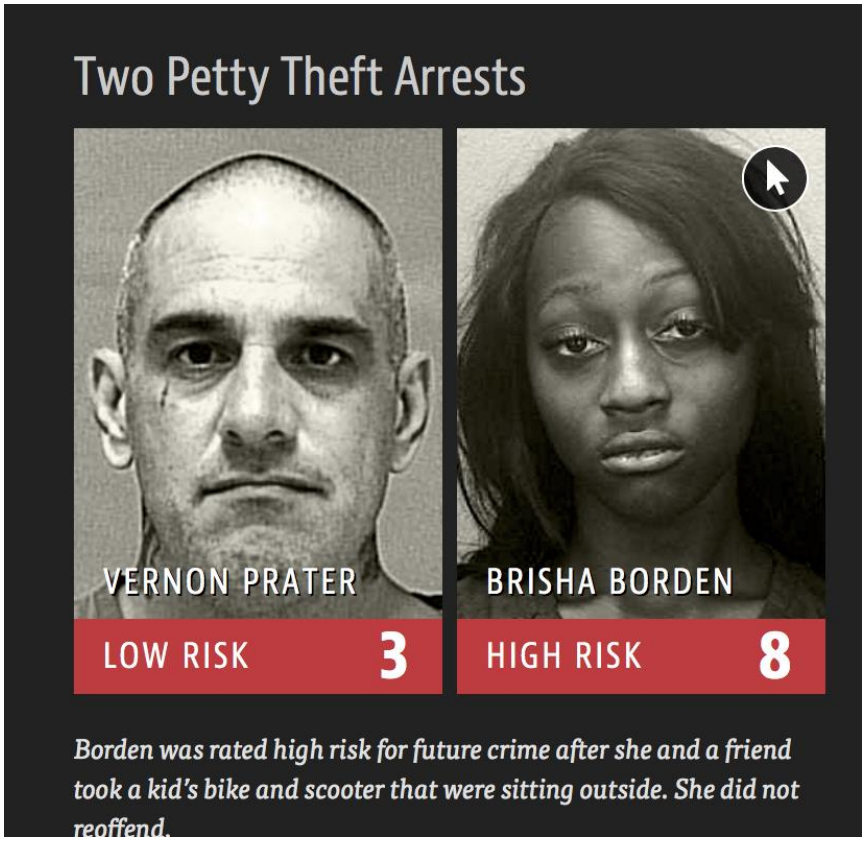
An important research field

- Needs collaboration between technical people, social sciences, law etc.
- Very active since ca. 2017
- See e.g. Kirkpatrick, *„Battling algorithmic bias: how do we ensure algorithms treat us fairly?"*, Communications of the ACM, Volume 59 Issue 10, October 2016

Algorithmic bias: examples



Algorithmic bias: examples (contd.)



Algorithmic bias: examples (contd.)

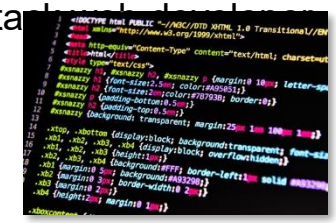
Indirect threat: mass unemployment

Fear

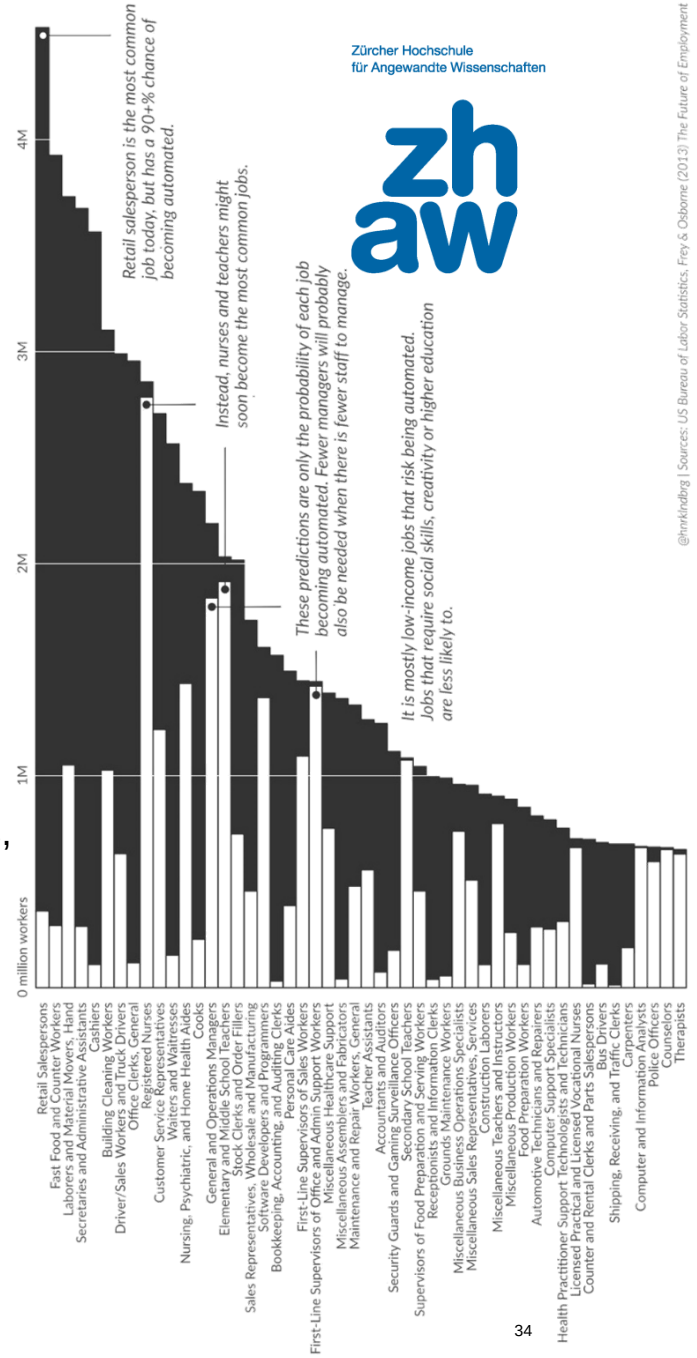
- **Less qualified** jobs **vanish** due to robots (see →)
- See https://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future-of-ai/

Likely

- **Repetitive tasks vanish** due to AI (lawyer researching test cases, …)



- Complex tasks get augmented (compare lab P01)
- Other jobs are created (humans need an occupation)



Indirect threat: overreliance

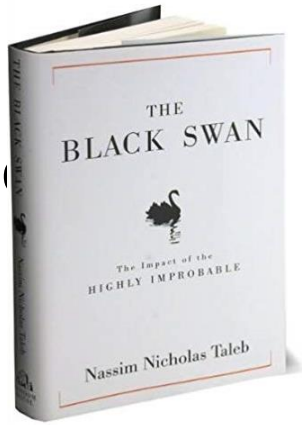
Pattern recognition ≠ intelligence

Patterns

- Wikipedia: *«A pattern is a **discernible regularity** in the world or in a manmade design. As such, the elements of a pattern **repeat in a predictable manner**.»*
- That which is detectable by machine learning solutions

Our world

- Mightily impacted by «black swans»[1]
- Pattern recognition leads to abstraction, on which cognition (logic) must operate for really smart behavior



➔AI based on **machine learning will severely und**
unlikely
but existing phenomena

[1]) Nassim Nicholas Talib, *«The Black Swan: The Impact of the Highly Improbable»*, 2007

Example: semantics by pattern recognition methods can be hard

According to scholars Walter Krämer, Götz Trenkler, Gerhard Ritter, and Gerhard Prause, the story of the posting on the door, even though it has settled as one of the pillars of history, has little foundation in truth. The story is based on comments made by Philipp Melanchthon, though it is thought that he was not in Wittenberg at the time.

What story of little truth is a pillar of history?
Ground Truth Answers: posting on the door | story of the posting on the door | posting on the door
Prediction: the posting on the door

On whose comments is the posting on the door based?
Ground Truth Answers: Philipp Melanchthon | Philipp Melanchthon | Philipp Melanchthon
Prediction: Philipp Melanchthon

Where was Melanchthon at the time?
Ground Truth Answers: not in Wittenberg | not in Wittenberg | not in Wittenberg
Prediction: Wittenberg

What do scholars agree on about the posting on the door story?
Ground Truth Answers: little foundation in truth | has little foundation in truth | settled as one of the pillars of history
Prediction: little foundation in truth

https://rajpurkar.github.io/SQuAD-explorer/explore/1.1/dev/Martin_Luther.html

Zürcher Hochschule
für Angewandte Wissenschaften

2. GUARDING AGAINST MALICIOUS USE

Malicious use (definition):

includes all practices that are
intended to compromise the securi
individuals, groups or a societ

Based on Brundage et al., *"The Malicious Use of Artificial Intelligence",* 2018

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.

Example for novel vulnerabilities
Adversarial attacks and counter measures

Zürcher Hochschule
für Angewandte Wissenschaften

BW-TEC
ENGINEERING AUTOMATION CONSULTING

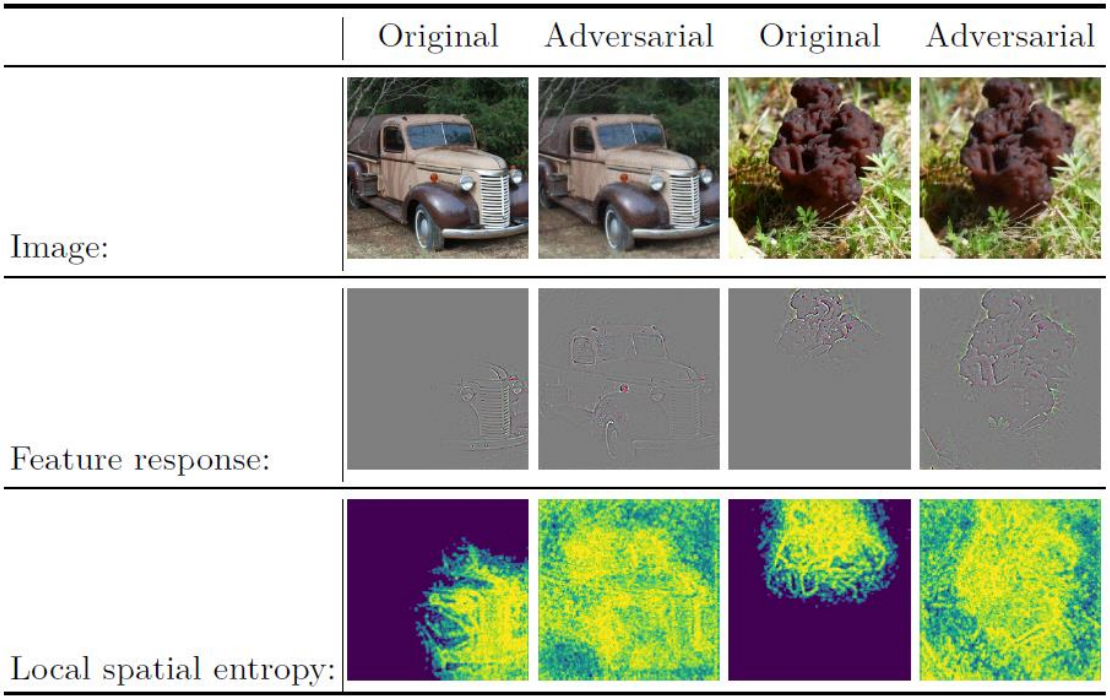
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Innosuisse – Swiss Innovation Agency

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.

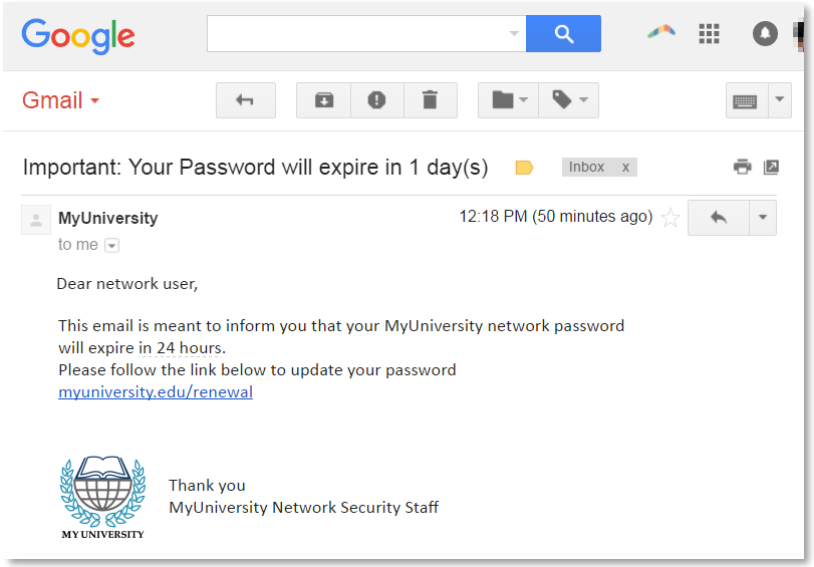
AI expands existing threats

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

- **Set of actors** who are **capable** of carrying 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)



AI introduces new threats

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

- Example: disinformation by **in** 

oice/image/text synthesis
- Compare https://lyrebird.ai/

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

- Example: cause self-driving car 

them with adversarial examples

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

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 drown 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

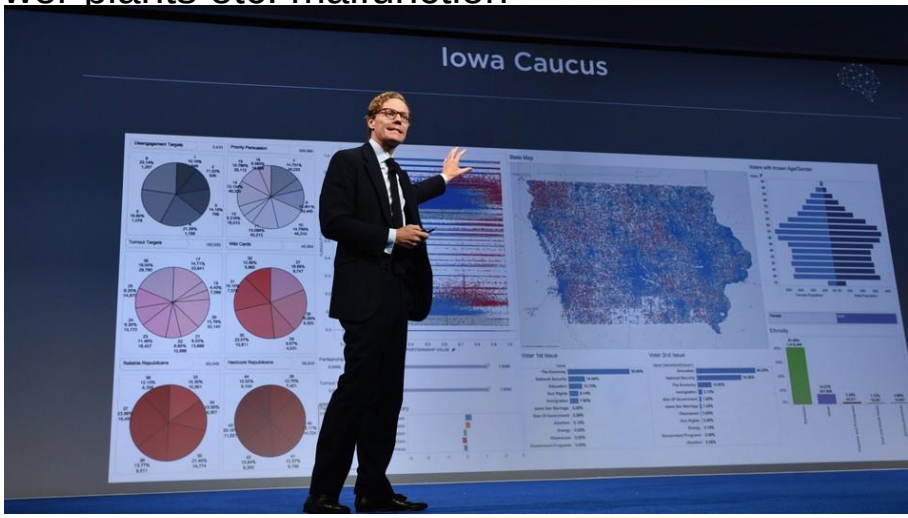
- By using AI systems to automate cyberattacks or social engineering
- By attacking AI systems

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.



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