

MERANTIX

TU Munich

Dr. Rasmus Rothe

February 7, 2019





ETH

CVL

Some like it hot – visual guidance for preference prediction

Rasmus Røhse, Rado Timofte, Liu Van Ooij
Computer Vision Lab @ ETH Zurich, Switzerland



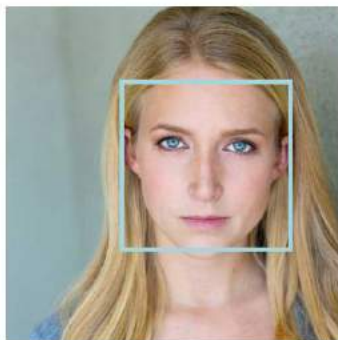
EXIT

CAESARS PALACE



*Let Artificial Intelligence guess your
attractiveness and age*

#howhot



BLINQ

♀ 25 years



Share with your friends:



 TRY ANOTHER PHOTO

Do I have to take this seriously?

Bin ich schön? Oder bloß hübsch?



Schöne Forscher haben eine Methode entwickelt, die die Attraktivität ihrer Nutzer bewertet. Wie soll die gehen? (The Daily Mail)



ওয়েবসাইটই বলে দিচ্ছে আপনি সুন্দর না কুৎসিত (এখানেই জেনে নিন আপনার সৌন্দর্য কতটা)

ওয়েবসাইট, না, না আর বারবার আপনার প্রফাইল দেখে। যুক্তি দিয়ে মিলিয়ে দিয়ে মানুষের কাছ থেকে জেনে নেওয়ার সময়কাল নেই। যে দিয়ে বন্ধু কখনো মিম্বা বলে না তাকে প্রশংসা করার সময়কাল নেই। দারুন সোজা মিম্বায়ে কেউকে যেটো আপনাকে কতটা পুর লাগছে তাই মিম্বায়ে দেখানোর সময়কাল নেই। এতে কৃত্রিম বুদ্ধিমত্তার এক অস্বাভাবিকই ভাবে সেন্স আপনাকে সুন্দর না কুৎসিত। এই ওয়েবসাইট আপনার ছবি আপনাকে কতটা (এখানে ক্লিক করে সেই ওয়েবসাইট ঘা)



Politik, Wirtschaft, Panorama, Sport, Kultur, Netzwerk, Wissenschaft, Gesundheit, einestages, Karriere, Uni, Reise, Auto, Stil

Netztrend: Diese Software sagt Ihnen, ob Sie schön sind

Von Angela Gruber



Швейцарские ученые создали сервис, невероятно точно оценивающий привлекательность человека по фотографии

Уверены, что выглядите как горячий мачо, или комплексуете по поводу своей внешности? Смотрите себе неотразимой красоткой по фотографии? EveningStandard

How attractive is YOUR selfie? Nigella just 'OK', the Duchess of Cambridge 'nice' and Kylie Jenner is 'Godlike' according to a new dating app that rates your photos

Dating app Blinq is now enabling users to rate



- 1 6713 It turns out I'm less attractive than a tent. (Limgur.com)
- 2 6499 Me, my painting and Bob Ross (1990) (Limgur.com)
- 3 3858 Magnitude 5.1 seismic disturbance recorded in North Korea (cnn.com)
- 4 5924 The festivities continue in Russia (imgur.com)
- 5 4096 My buddy made this Doritos commercial last year; it didn't win, but I feel it deserved a little more love. (vimeo.com)
- 6 4639 TIL When larger Kangaroos are chased they will often lead their pursuer to water, then once standing submerged to the chest, the kangaroo will attempt to drown the attacker. (wikipedia.org)

20 Minuten

Wiederholt Games 5

Media an Diek

News - Video - Events - CrunchBase

9TH ANNUAL CRUNCHIES Two Days Until Prices Increase For The Tech Awards Show Of The Year Get Your Tickets Now

Bling Dating App Uses AI To Judge Hotness

Posted Jan 11, 2016 by Natasha Lomas (editor)

679 SHARES

Let Artificial Intelligence guess your attractiveness and age

#HowHot #TechCrunch



Wondering exactly how hot your profile picture is? New dating app Blinq will tell you

Product Hunt

How hot: Let artificial intelligence guess your age, attractiveness and hotness

It's getting hot!

Hacker News new | threads | comments | show | ask | jobs | submit







Challenges when building an AI product



Commercial challenges



Technical challenges

Challenges when building an AI product



Commercial challenges



Technical challenges

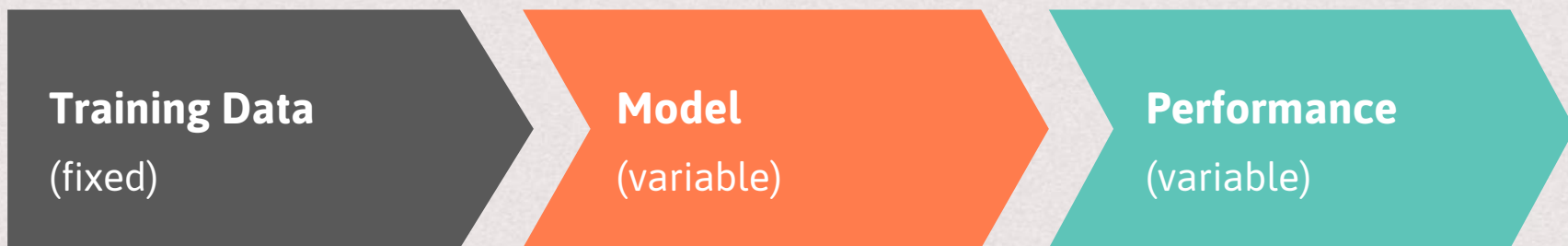


A little disclaimer

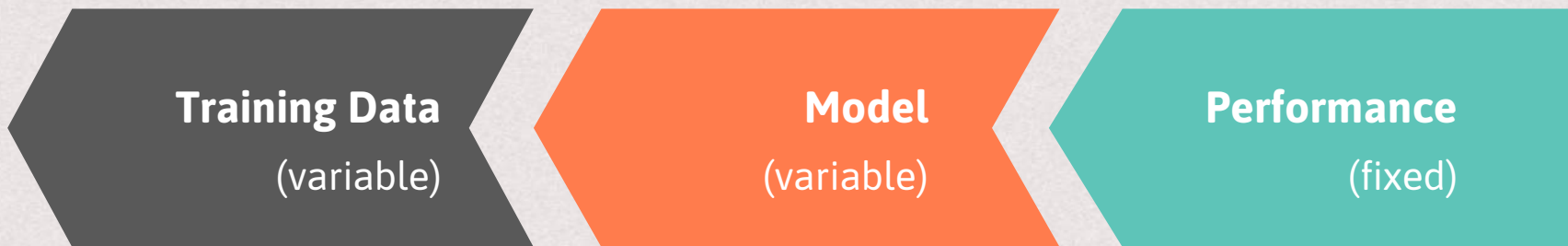
This presentation is:

- Representing Merantix' learnings from technological challenges when building robust AI products in three highly demanding industries: healthcare, finance and automotive
- Covering a list of challenges and learnings that is not exhaustive but gives a structured overview of the major topics, we have been facing across industries
- Though we do mostly supervised deep learning, many challenges and learnings generalize and are applicable for other types of ML
- Maybe not entirely new to you but hopefully you can take away at least a few points

How academia works

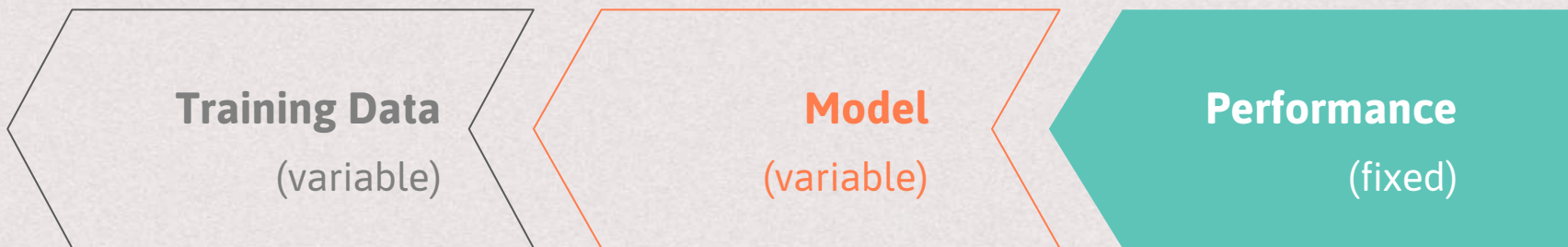


How industry works



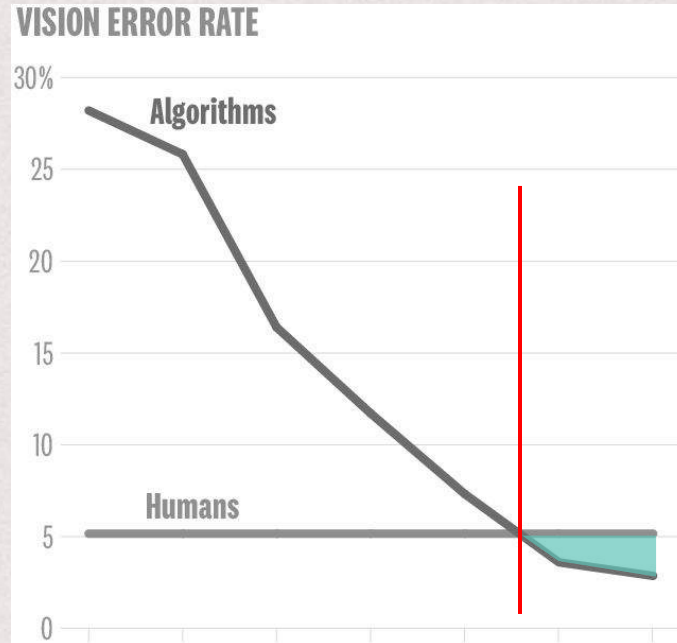
Lots of Implications

How industry works





Performance: Success depends on product scope



Support systems

**Specialized
autonomous
products**



Performance: “binary” commercial success criteria

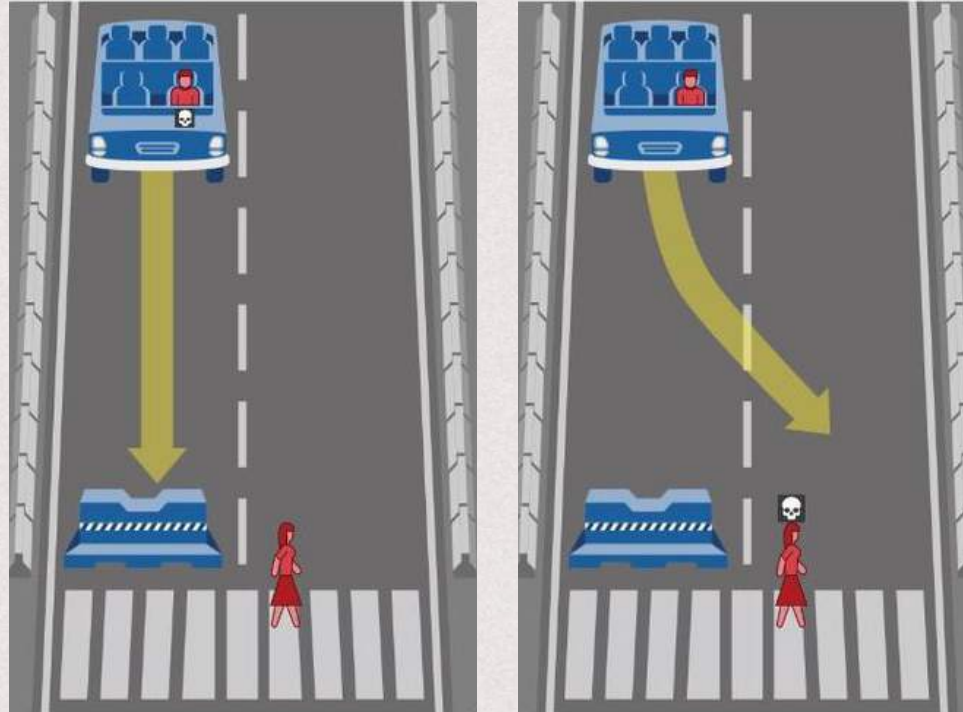


VS.



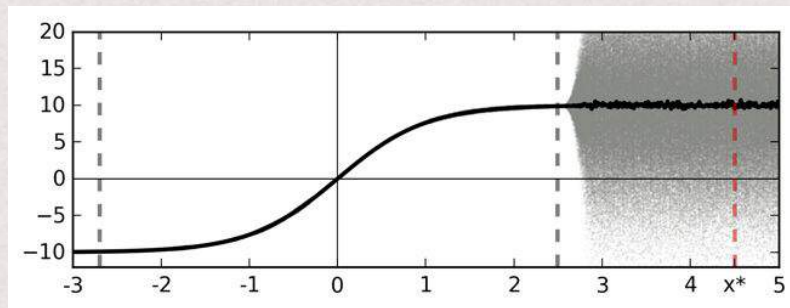


Performance: Perception can vastly differ

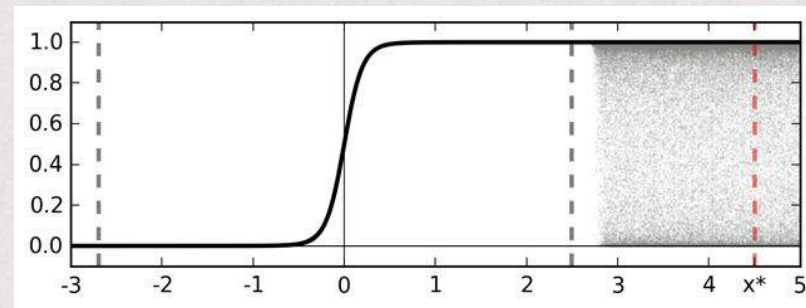




Performance: Uncertainty is still an open topic



Softmax **input** as a function of data x



Softmax **output** as a function of data x

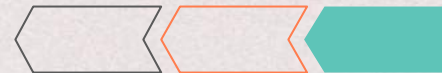
Softmax

≠

Uncertainty

Recent trend: Bayesian deep learning





Performance: Measuring is not trivial

Know your target environment



Relevant Context

- city, weather
- screening vs diagnostics



Dynamic elements

- cities changing
- people changing



Impact on environment

- impact of other autonomous cars
- biasing the radiologists



Performance: good enough?

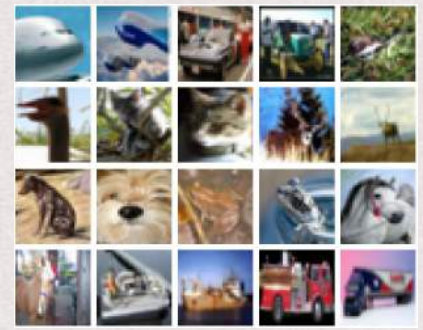
Overfitting



Original test set

Good performance = good model OR too easy test set?

4-10% drop in accuracy



Newly collected test set

Source:

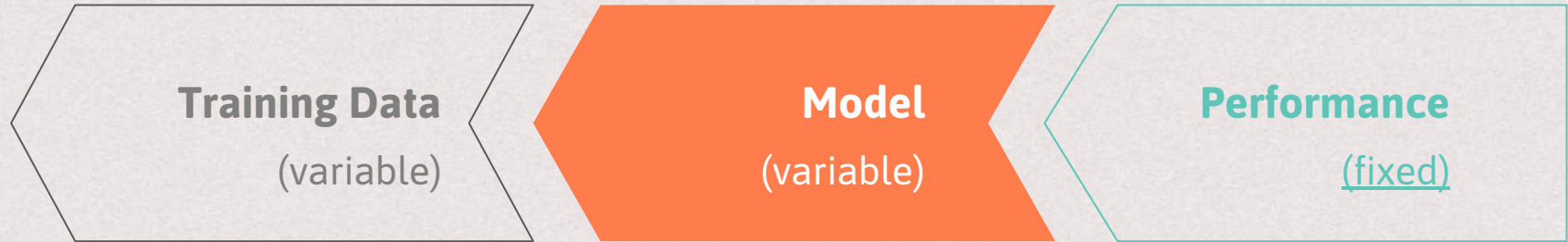
[1] Recht, B., Roelofs, R., Schmidt, L. and Shankar, V., 2018. Do CIFAR-10 Classifiers Generalize to CIFAR-10?. arXiv preprint arXiv:1806.00451.



Performance: Summary

- 1. Hit the binary success criteria**
- 2. Define (and limit) the product scope**
- 3. Understand and shape public perception**
- 4. Predict uncertainty**
- 5. Know your target environment**
- 6. Don't overfit on your test set**

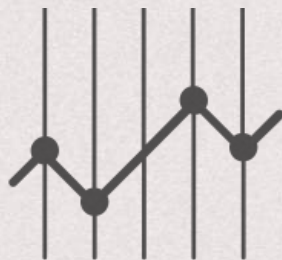
How industry works





Model: Optimize the right thing

Learning task (loss) = Performance metric = Business goal



Predicting the
market behavior vs
p&l loss






Mass detection accuracy / smart
elimination cut off
for zero false negatives



Model: Not equal loss

Trading

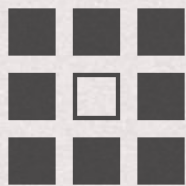
	Predicted Move	Opposite Move
Traded		
Ignored		

Medical Imaging

	Healthy	Sick
Diagnosed		
Not diagnosed		



Model: Industry doesn't like black boxes



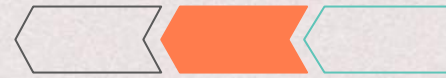
Business context



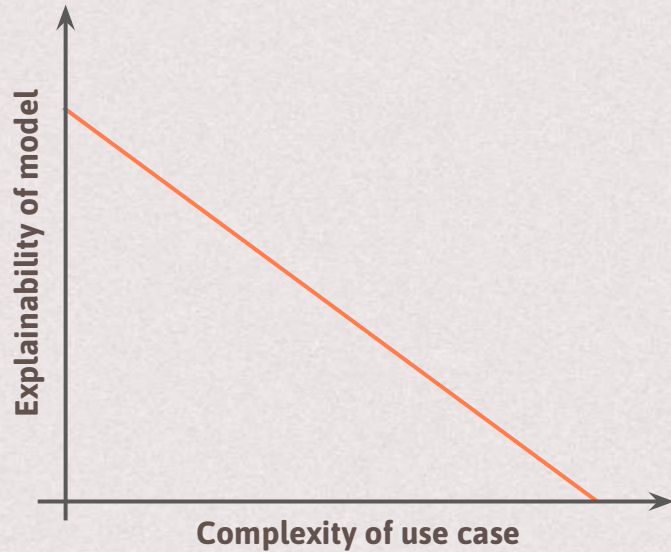
Liability



Auditability



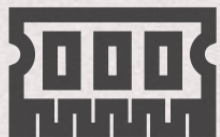
Model: Explainability remains a trade off



- Complex use cases require complex models
- Model complexity is anti-proportional to its explainability
- Growing research field



Model: Size matters



**Limited
memory**



**Limited
bandwidth**

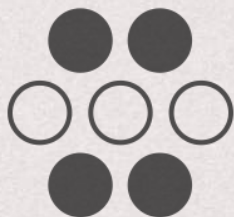


**Limited execution
time**



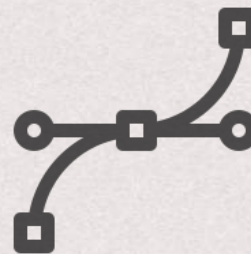
Model: State of the art

State of the art often not needed



**Many papers only on MNIST/
small dataset**

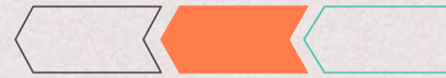
→ questionable if it works on larger dataset



Many methods overengineered

In order to improve state of the art

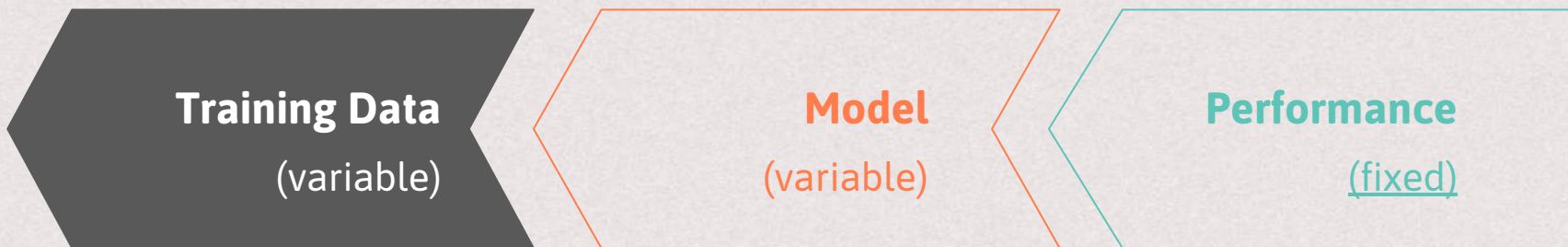
→ not worth it for production



Model: Summary

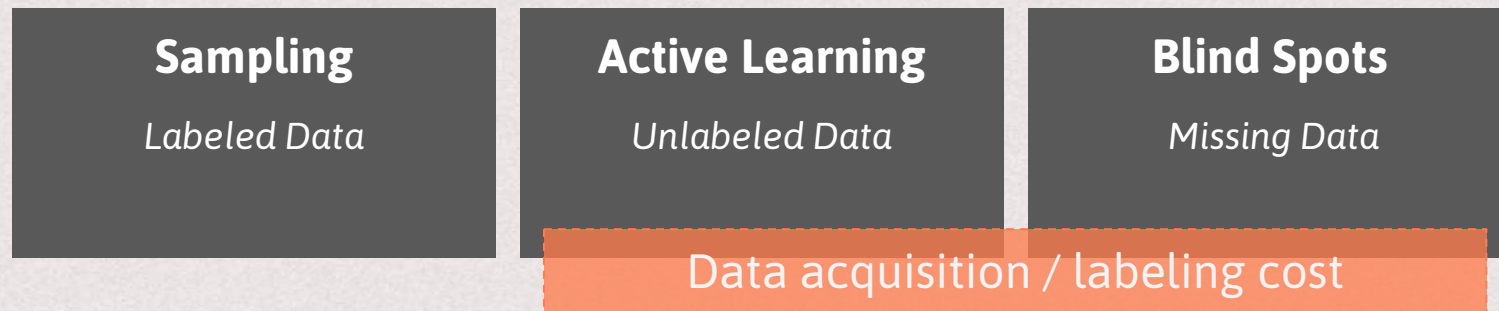
- 1. Align loss function to business goal**
- 2. Consider unequal cost of misclassification**
- 3. Make your model explainable**
- 4. Size matters**
- 5. State of the art models often not required**

How industry works





Data: How to assemble the training set



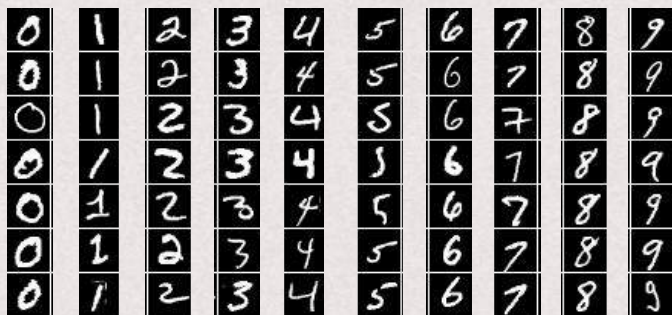
Sources:

- [1] Sener, O. and Savarese, S., 2018. ACTIVE LEARNING FOR CONVOLUTIONAL NEURAL NETWORKS: ACORE-SET APPROACH. *stat*, 1050, p.21.
- [2] Wang, K., Zhang, D., Li, Y., Zhang, R. and Lin, L., 2017. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12), pp.2591-2600.
- [3] Gal, Y., Islam, R. and Ghahramani, Z., 2017. Deep bayesian active learning with image data. *arXiv preprint arXiv:1703.02910*.
- [4] Krishnakumar, A., 2007. Active learning literature survey. Technical Report, University of California, Santa Cruz.
- [5] Torralba, A. and Efros, A.A., 2011, June. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* (pp. 1521-1528). IEEE.

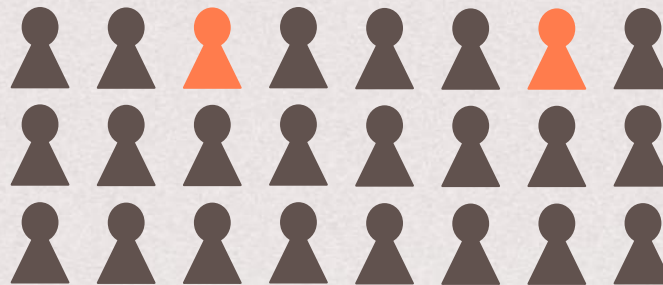


Data: deal with class imbalance

Academic datasets are balanced ...



... real world datasets aren't



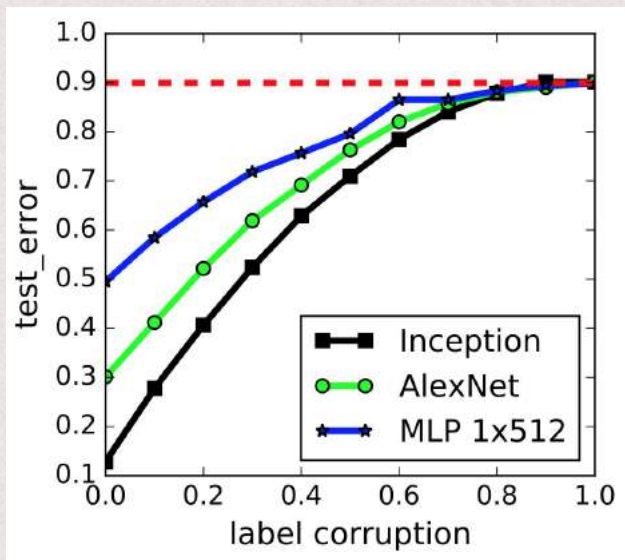
Source:

[1] Buda, M., Maki, A. and Mazurowski, M.A., 2017. A systematic study of the class imbalance problem in convolutional neural networks. arXiv preprint arXiv:1710.05381.





Data: quality / noise



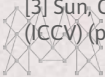
- Labels are noisy, especially when they are created by humans;
e.g. annotation of medical images
- Noise has huge impact on performance

Sources:

[1] Zhang, C., Bengio, S., Hardt, M., Recht, B. and Vinyals, O., 2016. Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530.

[2] Rolnick, D., Veit, A., Belongie, S. and Shavit, N., 2017. Deep learning is robust to massive label noise. arXiv preprint arXiv:1705.10694.

[3] Sun, C., Shrivastava, A., Singh, S. and Gupta, A., 2017, October. Revisiting unreasonable effectiveness of data in deep learning era. In 2017 IEEE International Conference on Computer Vision (ICCV) (pp. 843-852). IEEE.





Data: Summary

1. **Consider cost-benefit trade off for data acquisition or labeling**
2. **Focus on rare samples**
3. **Get high quality annotations**

Development

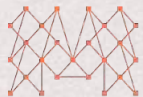
Challenges

- Lengthy iteration cycle (training)
- Unit and integration testing is difficult on models
- Reproducibility
- Concept of modularization vs end-to-end trainable systems are desirable from an ML perspective



Life hacks

- Run the pipeline every night
- Standardized configuration system
- Toy example to be run locally
- ...?



MERANTIX

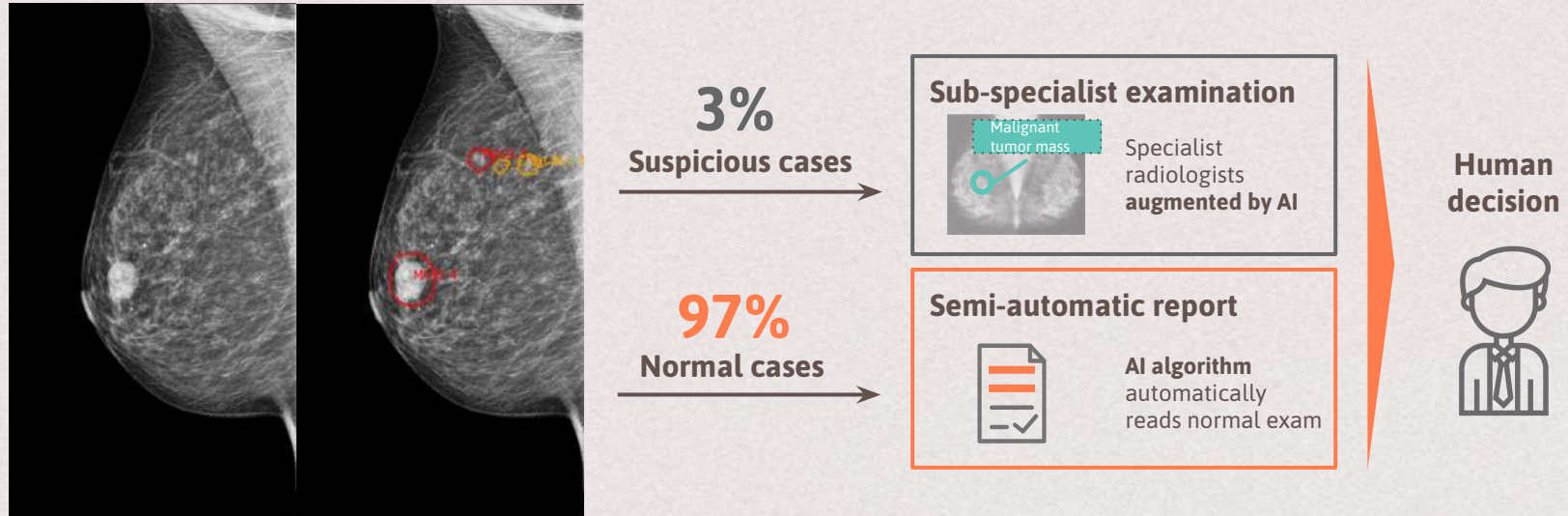
TU Munich

Dr. Rasmus Rothe

February 7, 2019

MX Healthcare

Smart elimination & automation in mammography screening



Please log in to your account



Log in

[I forgot my password](#)



Name	Surname	Date of birth	Acquisition	Patient ID	Screening ID	Institution	
Clara	Eriksson	12.06.1980	16.08.2018	ERN 206	ERN 206	Berlin	>
Anna	Smith	10.08.1972	21.07.2018	OHN 305	OBR 641	Munich	>
Jennifer	Johnson	7.12.1969	20.07.2018	PGB 847	NGG 667	Dortmund	>
Samantha	Williams	10.06.1968	17.06.2018	ZVE 016	NHM 007	Hamburg	>
Aisling	Brown	2.02.1973	11.06.2018	NFI 803	BBZ 779	Hamburg	>
Tania	Jones	1.05.1965	10.06.2018	HDI 960	NRE 900	Berlin	>
Elisabeth	Miller	3.09.1970	6.05.2018	NDW 406	WTQ 786	Berlin	>
Maria	Davis	4.05.1962	6.05.2018	FYQ 600	CUB 874	Munich	>
Emma	Wilson	11.08.1967	4.05.2018	VGX 563	VVX 901	Berlin	>

Name **Clara**Surname **Eriksson**Date of birth **20.06.1980**Age **38**Gender **Female**Patient ID **ERN 206**

Load more information



Current report

Acquisition date 8.10.2018

Prior study considered

 Yes No

Density *

 A B C D

Technical quality

 No limitations Restricted Repeat

Assessment *

 BI-RADS® 0 No pathological finding Suspicious findings[Cancel](#)[Sign and send](#)

* Obligatory field



Fit



Dicom Info



Invert



Reset



Forward



Backward



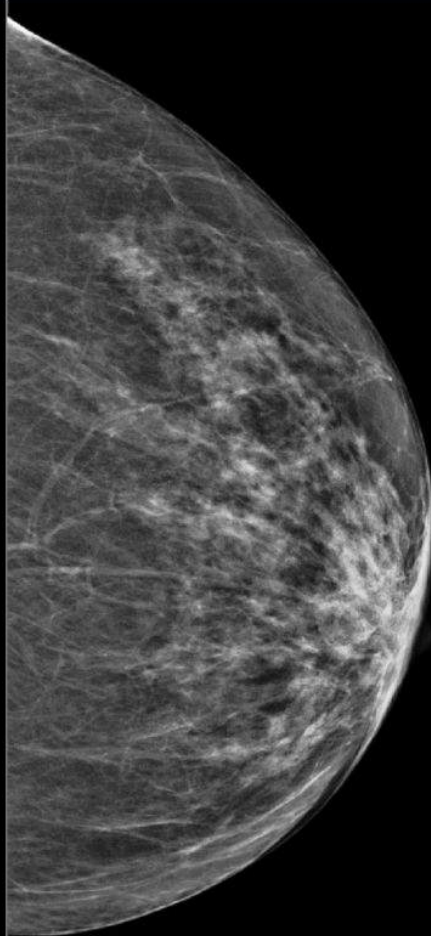
Highlight the findings

R CC

Michaela Mammo (F)
PatID: DRK2011Mammo04
05.11.2009 08:11:51
Hologonic Selenia and R2
Selenia
Zoom: 0,11 / 0,42

1%

loading





Fit



Dicom Info



Invert



Reset



Forward



Backward



Highlight the findings

R CC

Michaela Mammo (F)
PatID: DRK2011Mammo04
05.11.2009 08:11:51
Hologonic Selenia and R2
Selenia
Zoom: 0,11 / 0,42

L CC

Michaela Mammo (F)
PatID: DRK2011Mammo04
05.11.2009 08:11:51
Hologonic Selenia and R2
Selenia
Zoom: 0,11 / 0,42

Specify the finding

7.9mm

Left Right

Localisation

This field is mandatory

Calcification

Birads

This field is mandatory

Add Cancel

Specify the finding

7.9mm

Left Right

Localisation

- Option 1
- Option 2
- Option 3
- Option 4
- Option 5
- Option 6
- Option 7
- Option 8
- Option 9
- Option 10

Specify the finding

7.9mm

Left Right

10:00 o'clock

Calcification

4b

✓ Cancel



Rotate left



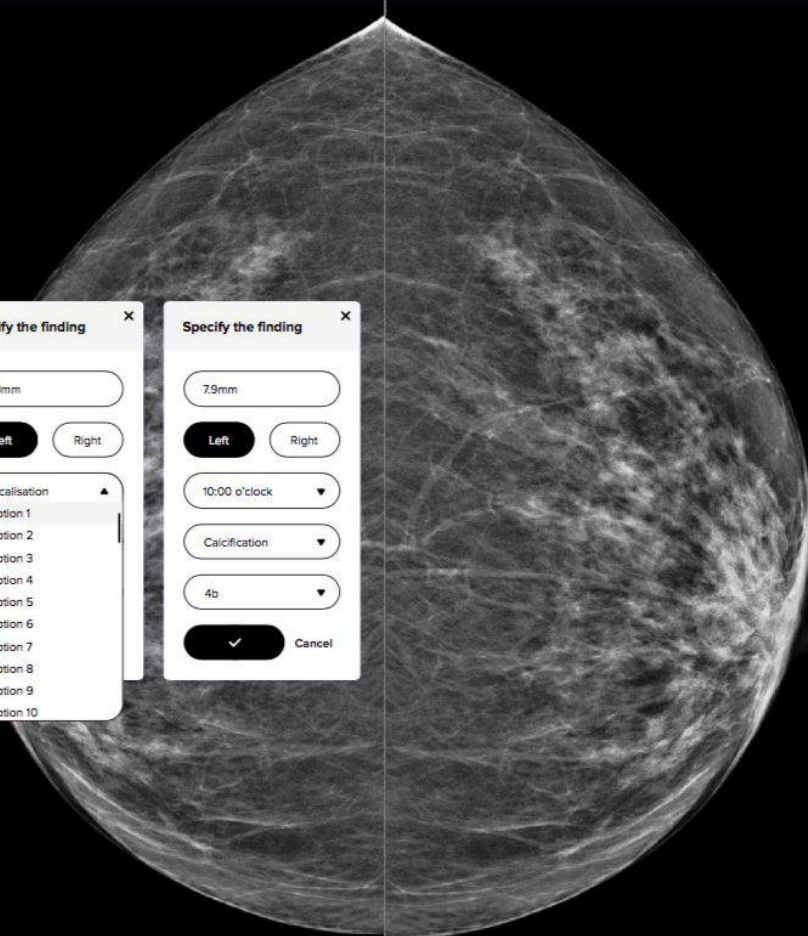
Rotate right

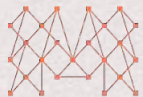


Flip



Reset





MERANTIX

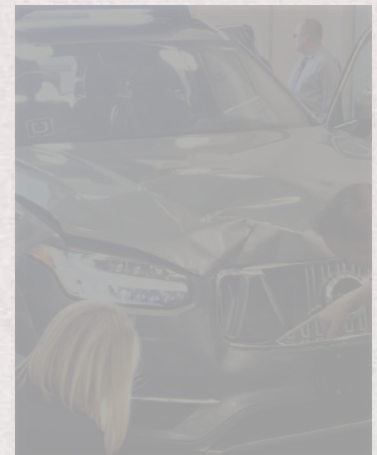
TU Munich

Dr. Rasmus Rothe

February 7, 2019

Self-driving car research has been around for a while

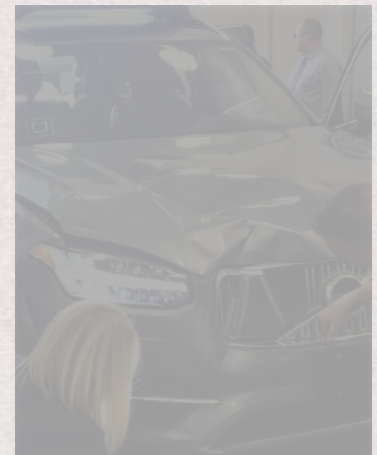
Until recently, with limited results



DARPA Grand Challenge “Winning” Sandstorm vehicle by Carnegie Mellon University in 2004

Self-driving car research has been around for a while

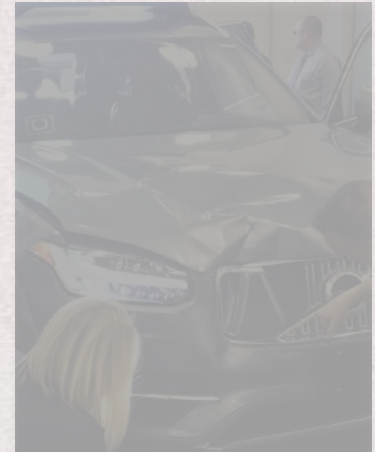
Until recently, with limited results



DARPA Grand Challenge “Winning” Sandstorm vehicle by Carnegie Mellon University in 2004

Recently, ML and robotics have made rapid progress

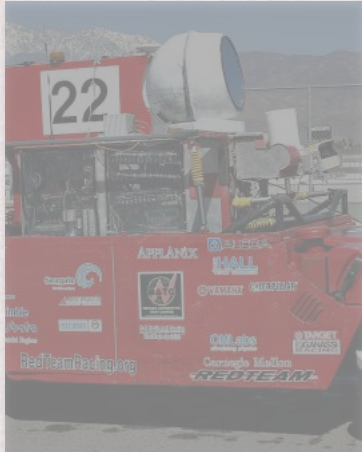
Helping researchers to reach a unique milestone



“Stanley” became to the first fully autonomous car to fully win the DARPA Grand Challenge in 2005

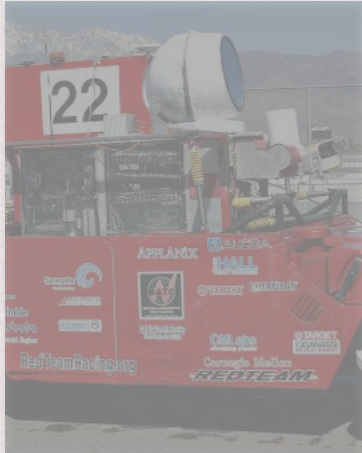
Commercial interest and research funding picked up

But the safe, widespread commercial use remains a vision



Commercial interest and research funding picked up

But the safe, widespread commercial use remains a vision



Recent fatal Uber crash, the AV saw the pedestrian but did not act appropriately

Highly Autonomous Vehicles will profoundly change mobility away from personal car ownership to an automated service

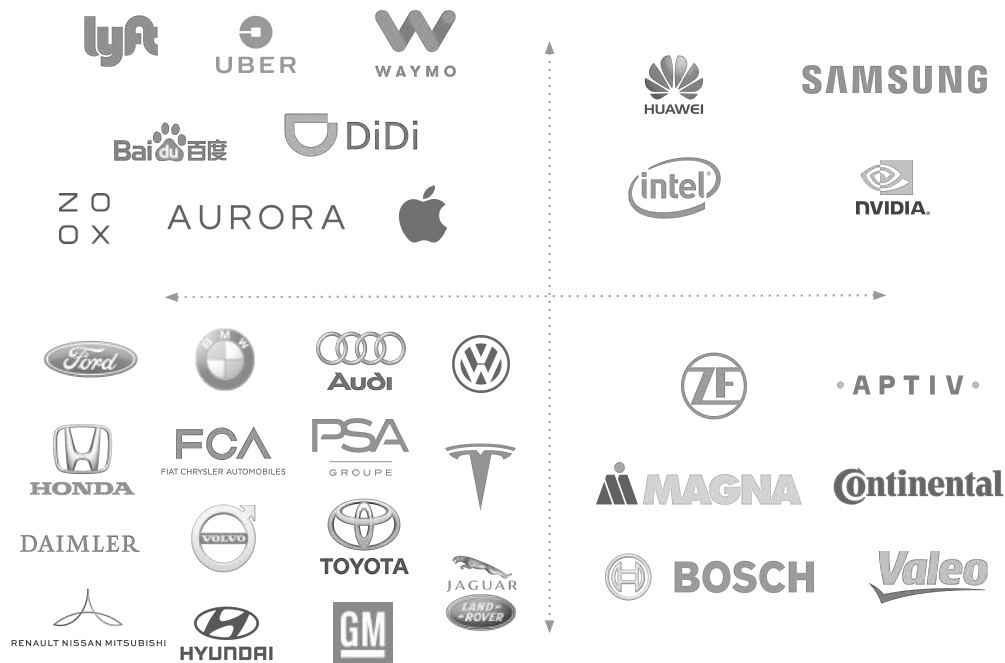
Highly Autonomous Vehicles (HAVs) have the potential to unlock major economic and societal benefits for our cities and economies:

- Offer low cost universal demand-responsive mobility to every citizen
- Unlock unproductive commuting time
- Increased road safety for all road users by orders of magnitude
- Materially lower pollution, congestion and resources today wasted in manufacturing, assembling, parking and disposing of personal vehicles.



However, all of the above only applies, if we can deploy L4 vehicles on scale.

The ongoing race to commercially deploy L4+ autonomous vehicles is accelerating



- More and more companies are entering the space and have set **challenging targets** for their initial L4 (HAV) deployment
- Commercial success is directly linked to how soon HAVs can be deployed

None of the leading companies has solved L4 the differentiating factors are speed and safety



HOW FAST

Can you iterate your software?

Operating autonomous test cars is expensive, slow & risky; we need to accelerate dev cycles to reduce time to market.



HOW RELIABLY

Can you evaluate its safety?

Deploying self-driving cars poses significant risks; we need a way to reliably measure the safety of highly complex autonomous systems.



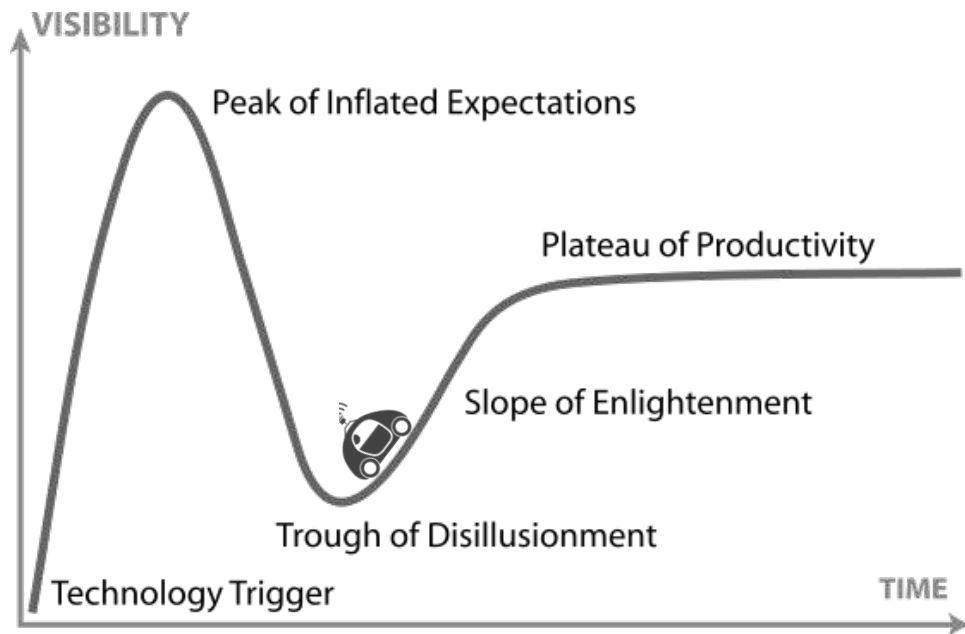
HOW EFFICIENTLY

Can you allocate your resources?

The amount of engineers and R&D funding is limited; we need to maximize the impact of model adjustments and training data.

Safety of Autonomous Vehicles is defined and measured through the testing method and technology.

If we do not find a solution to evaluate AV safety we might risk a prolonged halt of technological progress



- Progress has consistently been slower than expected, but why?
- Despite many years of research, fatal accidents occur easily (e.g. Uber in March 2018)
- Challenges posed by most current validation approaches are centered around **lack of safety guarantees, and lack of scalability**

The “functional safety” approach in the auto industry is not sufficient to cope with multi-agent environments

Functional Safety

- Integrity of the operation in an electrical (i.e. HW/SW) subsystem that is operating in a safety critical domain
- failure in HW or bugs in the SW that could lead to a safety hazard
- ISO 26262 / ASIL

Nominal Safety

- Concern of whether the AV is making safe logical decisions assuming that the HW and SW systems are operating error free (i.e. are functionally safe)
- There exists no nominal safety standard for the safe decision making capabilities of an AV

*Functional Safety then is a necessary, but not sufficient measure of safety assurance.
For nominal safety a model-based approach is required.*

HAVs have three primary stages of functionality

Nominal Safety is mainly concerned with planning

See

Accurately perceive the environment around the vehicle

Large progress in recent years, thanks to advancements in machine learning and computer vision.

Can mostly be evaluated based on ASIL, but lacks explainability.

Think

Decision making for strategic (i.e. change lanes) and tactical (i.e. overtake the blue car) decisions to take, which also take into account interactions.

Most critical to evaluate the nominal safety of an HAV

Key reason why testing and validation of HAVs is challenging and the focus of our work.

Act

Execution of the decision (translated into mathematical trajectories and velocities) to the various actuators within the vehicle to perform the driving decision

This is well understood by control theory and can be tested using classical ASIL methodologies

However, all three stages need to be tested in conjunction (full stack / end-to-end) in order to make a meaningful decision about system safety.

Safety requirements for driving are higher than you think

Back of the envelope:

1 million cars, driving 1 hour per day

→ 10^6 hours of driving a day

Safety target: 1 catastrophic failure every 1000 days

→ 10^9 hours without failure (similar to aircraft)

More failures may be acceptable for assisted systems, as humans can take corrective actions.

The human benchmark

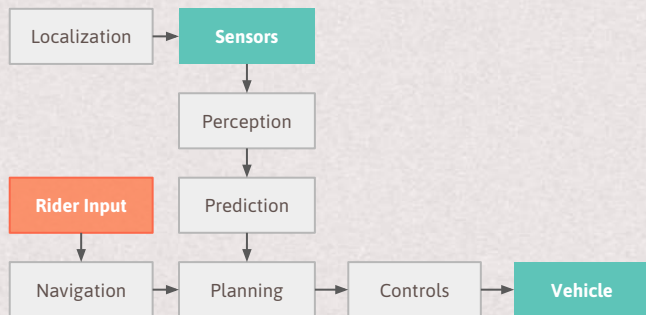
In Germany we record about 5 fatalities per 1 billion km of driving.

This means 1 fatal accident every 200 million km. ([source](#))

Waymo has self-driven ~15 million km on public roads since its inception. ([source](#))

While the autonomous driving stack is complex

Highly complex / interdependent system



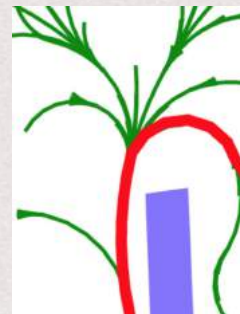
Distributed system architecture of an AV

Noisy sensor data



Velodyne LiDAR

Randomness in the system



Probabilistic path planning

Neural networks bring even more complexity

Transparency

- Increasing model complexity, decreases explainability
- Some efforts on explainability, but not (and unlikely to happen) breakthrough

Error rate

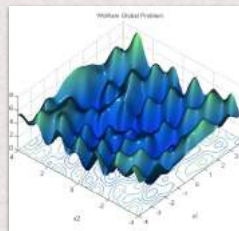
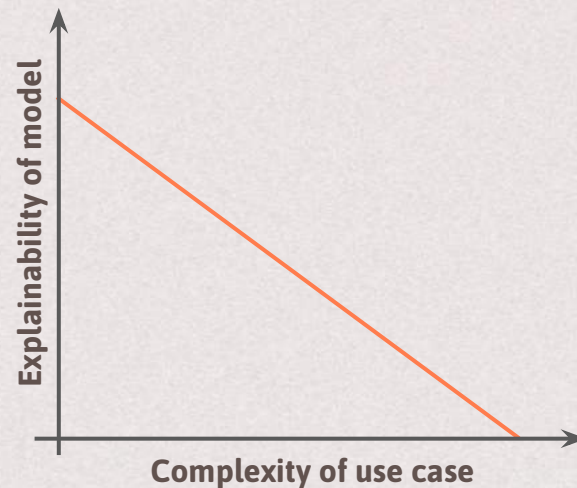
- Only statistical guarantee

Training-based

- No guarantee that training covers all relevant scenarios
- Critical (rare) cases usually underrepresented

Instability

- High dimensional parameter space leads to local optima
- Repeated training results in structurally different optima with similar behavior
- Makes it difficult to debug



Obtaining test data is not an easy feat



- Coverage requirements
- Dangerous to collect some of the cases
- Difficult to exercise a particular specific edge-case situation
- Testing set ever changing
- Secondary effects: Impact of autonomous car on environment
- Simulation can be artificial, oversimplified or unrealistic

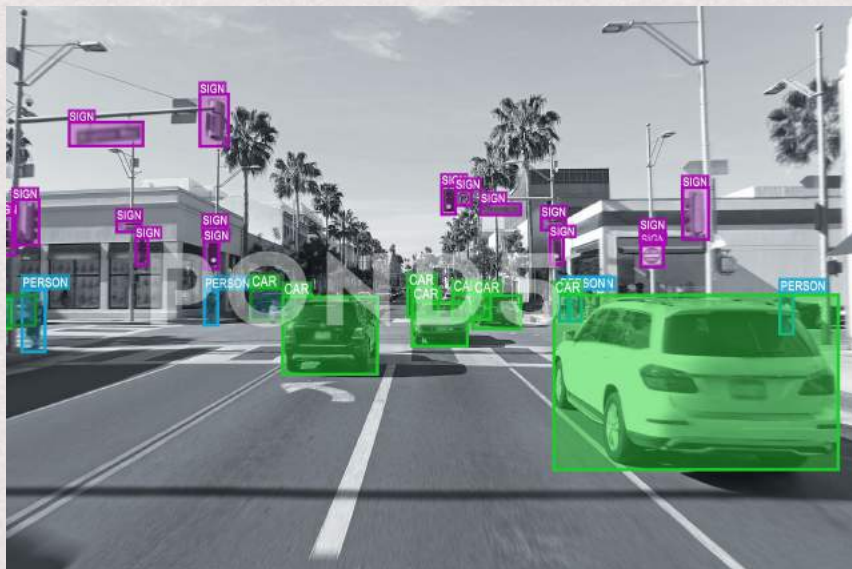
And there are many challenges when testing

- There is no unique correct system behavior for a given test case
- Probabilistic system behaviors
- Danger of overfitting on test set
- If any detail changes, everything needs to be re-tested

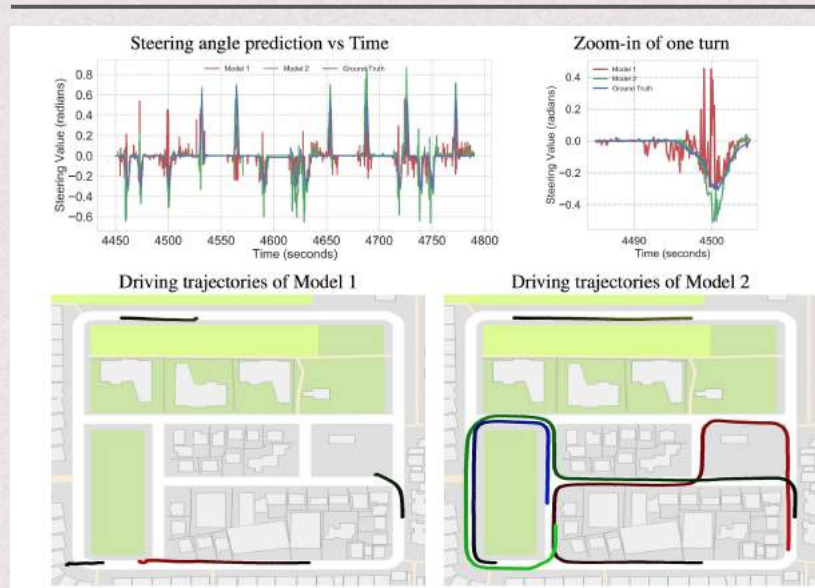
[1] Recht, B., Roelofs, R., Schmidt, L. and Shankar, V., 2018. Do CIFAR-10 Classifiers Generalize to CIFAR-10?. arXiv preprint arXiv:1806.00451.

Standard metrics don't work

Not everything matters



Similar metrics exhibit very different results



Codevilla, F., Lopez, A.M., Koltun, V. and Dosovitskiy, A., 2018. On Offline Evaluation of Vision-based Driving Models. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 236-251).

Disengagements/problems during road tests are difficult to prioritize and action on their own

Disengagement reports of autonomous mode Waymo and Daimler in California 2017:

Cause	Dec 2016	Jan 2017	Feb 2017	Mar 2017	Apr 2017	May 2017	Jun 2017	Jul 2017	Aug 2017	Sep 2017	Oct 2017	Nov 2017	Total
Disengage for a recklessly behaving road user	0	0	0	0	1	0	0	0	0	0	0	0	1
Disengage for hardware discrepancy	0	0	1	0	6	1	2	1	1	0	1	0	13
Disengage for unwanted maneuver of the vehicle	4	3	1	2	2	1	3	2	0	0	1	0	19
Disengage for a perception discrepancy	6	4	2	2	0	1	0	0	0	0	1	0	16
Disengage for incorrect behavior prediction of	1	0	0	0	1	2	0	0	0	0	0	1	5

Month	Miles driven in autonomous mode	Number of manual disengagements	Number of automatic disengagements
May	0.46	0	0
June	88.41	38	85
July	203.40	20	96
August	25.69	16	6
September	9.91	2	6
October	140.77	36	19
November	36.77	10	11

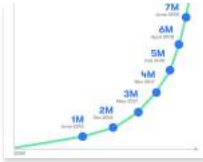
Yet, limited insights provided by the reports:

- **Disengagement reports** with information such as # of disengagements per mile or # of incidents **provide insufficient insights into AV software**
- **No information** is provided in many cases about the **reason of disengagement or failure**

Remaining questions:

- How and where can the model be improved?
- Did the model perform better than the previous?
- Which model should be deployed for road testing?

Most current approaches for testing nominal safety face problems to prove safe behavior and scalability



Miles Driven

The amount of miles needed to validate better than human performance is huge. And it would need to be re-done after the slightest change to the software or the environment.



Disengagements

Problematic distribution between “easy” and “difficult” cases. Furthermore, the relationship between “almost accidents” and “accidents” might not hold to up the requirements of rare corner-cases.



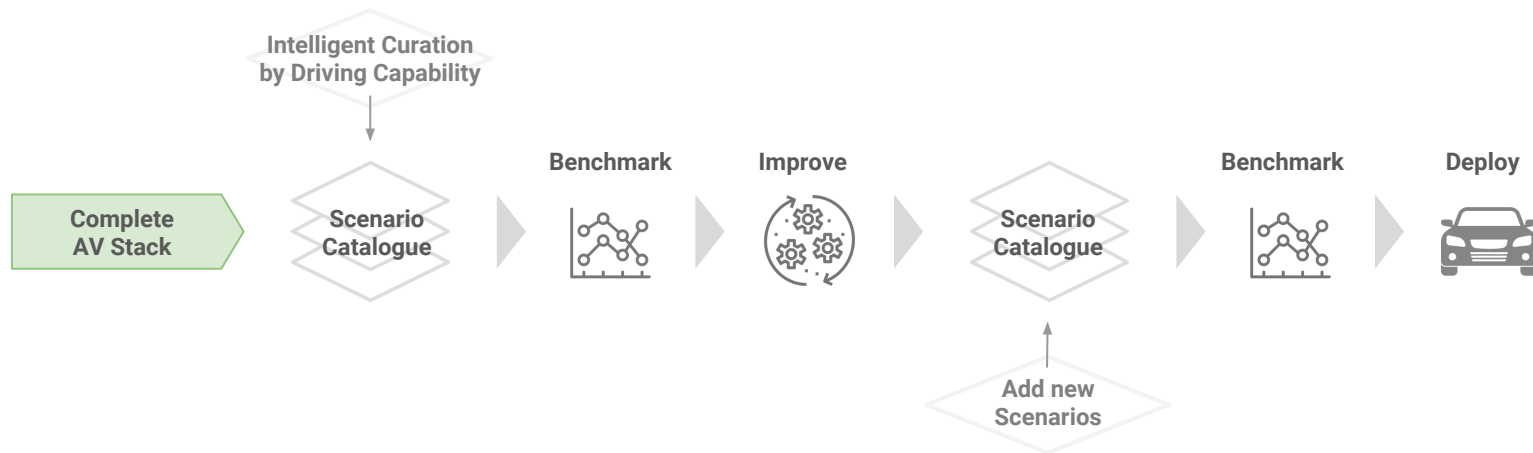
Simulation

Driving more and more cheap miles in simulation reinforces the generation of low value data. At the same time, proving that the simulator represents reality is as hard as validating driving policy itself.

To close the gap to the intensifying international competition, **we need to rethink the requirements for the validation of highly autonomous vehicles.**

Nominal safety needs to be evaluated in a semantic way

The most promising method is Scenario Based Testing



The scenario approach **mimics** the way we would test **safe decision making of a human driver**.

It offers the possibility to **test the HAV system as a whole**, with a focus on the hardest part: decision making

Intelligent curation of scenarios enables **exponentially faster** iteration, compared to alternative approaches.

Why does this make sense from a developer's point of view?



TRACK SAFETY RELEVANT PROGRESS

Through a common scenario language and motion model, coverage of driving capabilities becomes measurable. It's clear if model tweaks result in safety-relevant improvements or not.



WORK ON WHAT MATTERS MOST

Cluster weaknesses from structured reports, derive systematic insights, prioritize and implement; making the best use of your team's resources and iterate exponentially faster.



KNOW WHAT TO DEPLOY

Benchmarking of models over time or comparing competing release candidates ensures that expensive and slow road testing is used only on the best and latest iteration.

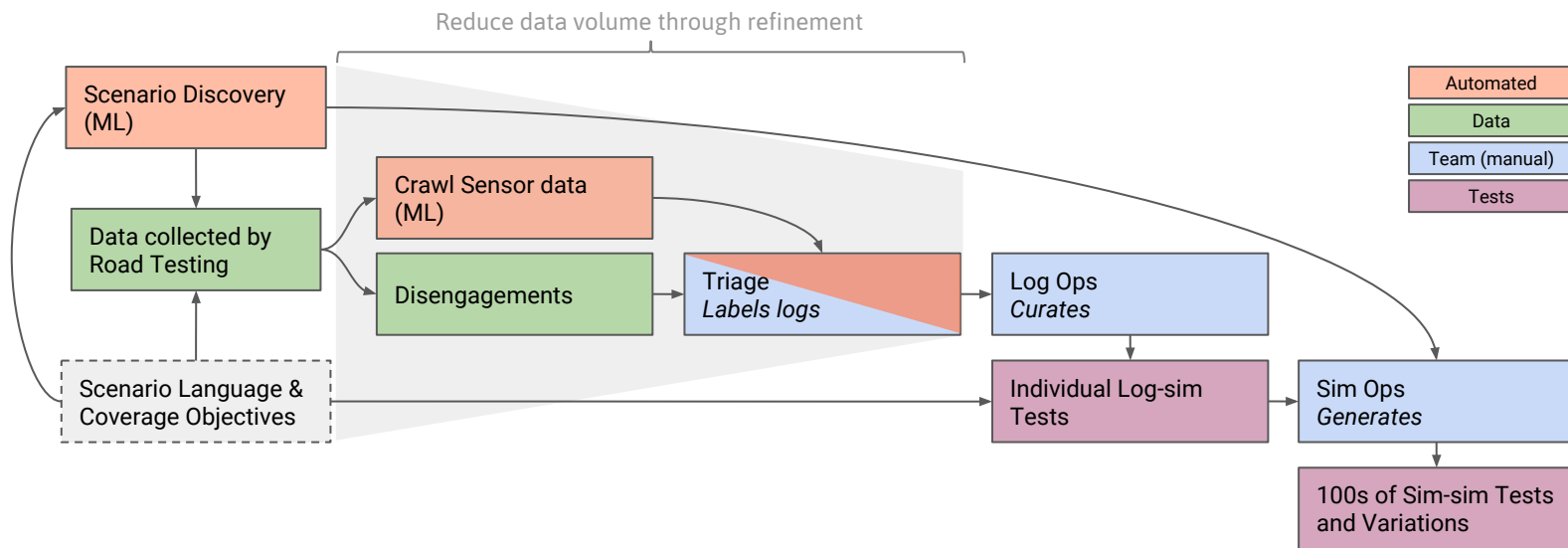


WORK WITH A SYSTEM THAT SCALES

The binary success structure of scenario tests and a refined complexity management to administer tests allow the system to scale to hundreds and thousands of AVs.

A scenario based approach is required because.. nothing else will scale to millions of tests and HAVs

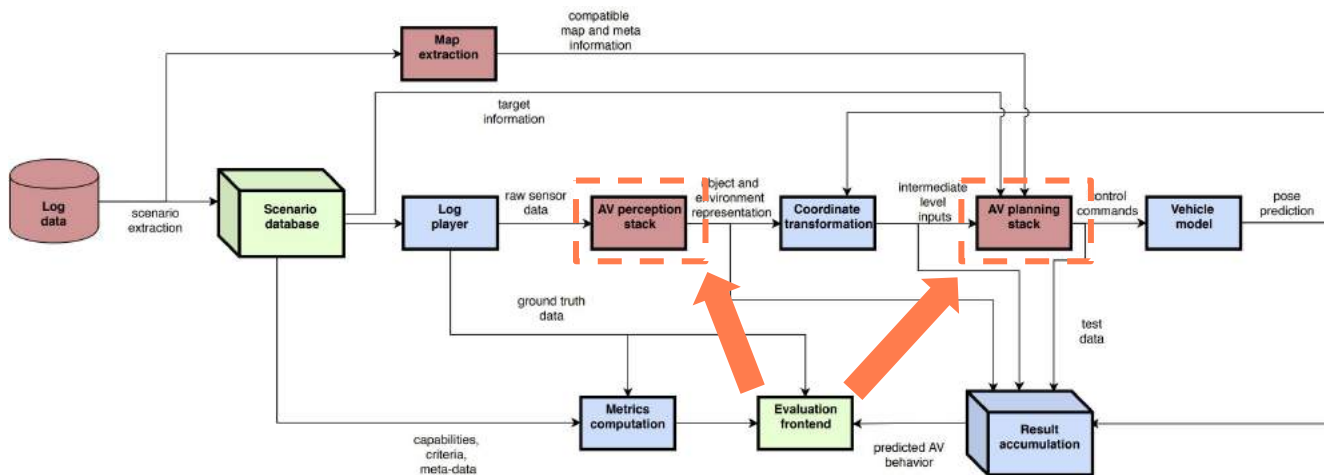
- Imagine, you have 100s of autonomous cars on the road collecting data every day
- How do you organize this data and improve your software? How do you measure progress and what to record next?
- **We are building a smooth and scalable pipeline from raw data acquisition to high value log and simulation tests**



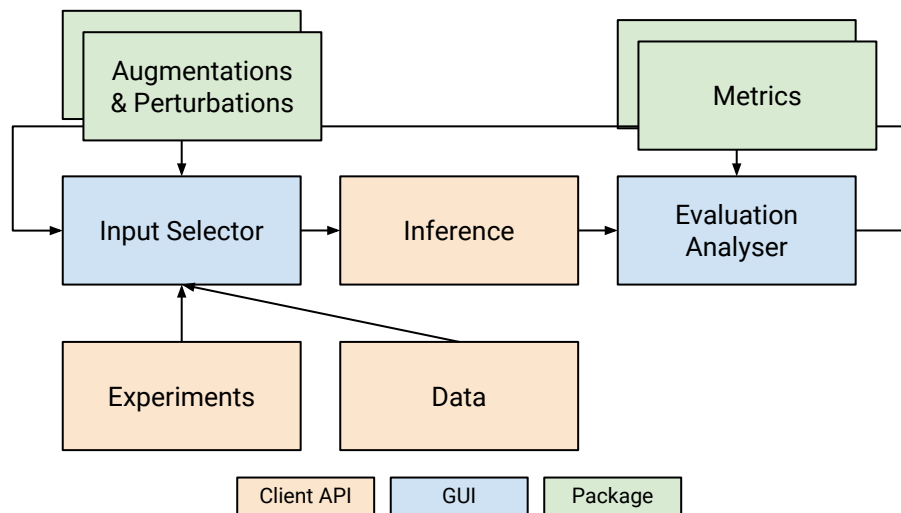
Going one level deeper: Component testing

is the functional counterpart to scenario testing

- Scenario based evaluation provides an assessment if the **system as a whole** behaves correctly
- While this might be sufficient for an external party, it is not for an AV developer
- In order to **know why the AV fails**, we need to be able to understand and test individual components of the stack
- Components such as the **perception stack involve deep learning**, which prevents the use of conventional methods



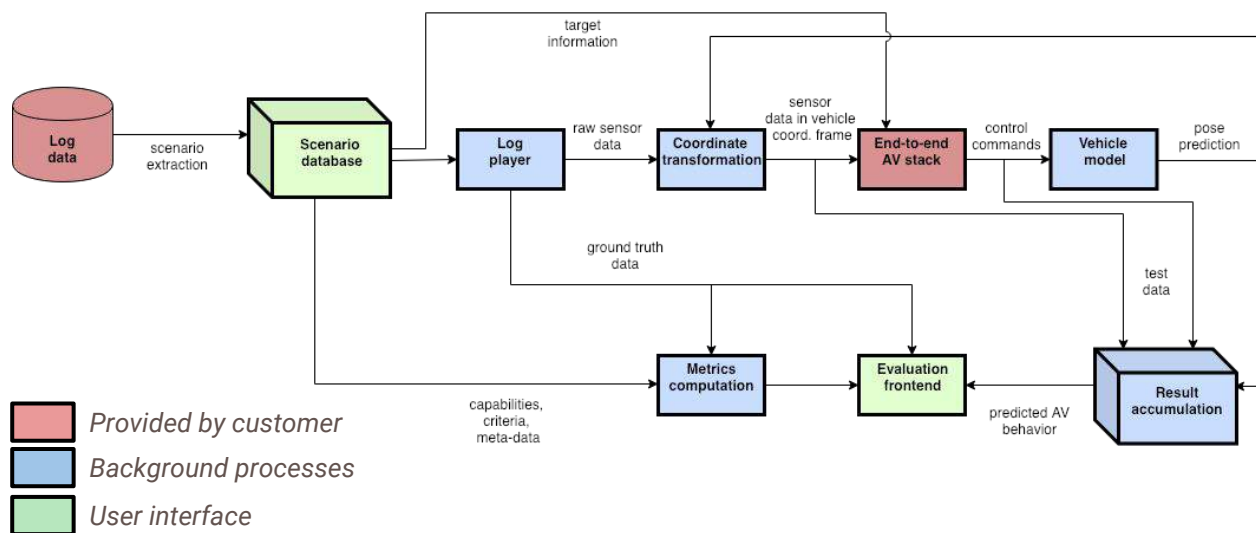
To drive improvements, we integrate component testing into our scenario based platform from the beginning



Above: Example framework we implement to evaluate deep learning based perception performance

- Our machine learning expertise enables us to **go beyond scenario based evaluation** and add value on the component level
- For example, we are building state of the art technology into our platform, that **helps AV developers understand and improve the performance of deep learning models**, powering the perception stack
- Similar frameworks are planned for evaluating planning and prediction

While the concept is simple, implementing SBT requires in-depth understanding of the AV tech stack



- Model **integrated** behavior for **end-to-end** testing of the SUT to provide expressive results
- Implement same **interfaces** and characteristics as a real vehicle
- Universal framework only requires **data recordings** and the **navigation stack** (SUT)
- Access through **database** and evaluation frontend
- **Straightforward** execution of tests

The first version of our testing product includes all the features you need to accelerate iteration



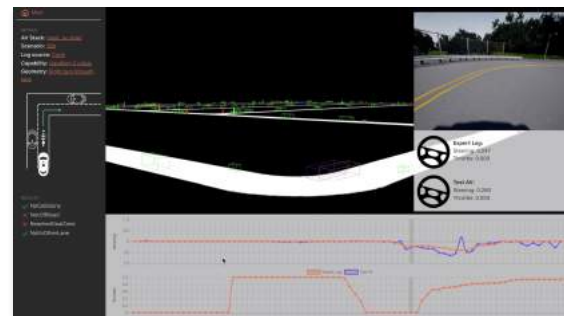
Get a **performance comparison** of one AV stack against other AV stacks in the form of a benchmark.



Monitor **performance trends of AV stack versions** over time for a comprehensive set of driving capabilities.



We **check a wide variety of driving capabilities** using many scenarios which can be **clustered and organized** with high flexibility.



Replay and thoroughly **investigate** failing scenarios for debugging, comparing against the **desired behaviour ground truth**.



Work with Merantix to **create many new scenario test cases** easily based on AV log data using our internal tooling.



Each scenario includes a set of comprehensive metrics (e.g. goal zone reach or lane changes), which jointly define the desired behaviour of the agent.

Join us on our journey

1 Science



Join a team of experts to research on the bleeding edge of deep learning.

2 Datasets



Get access to and explore some of the world's best datasets.

3 Business



Work within the leading European environment to commercialize AI.

We are hiring in Berlin!



**Machine Intelligence
Engineer**



Robotics Engineer



Software Engineer



**Technical Program
Manager**

