Can We Trust AI?

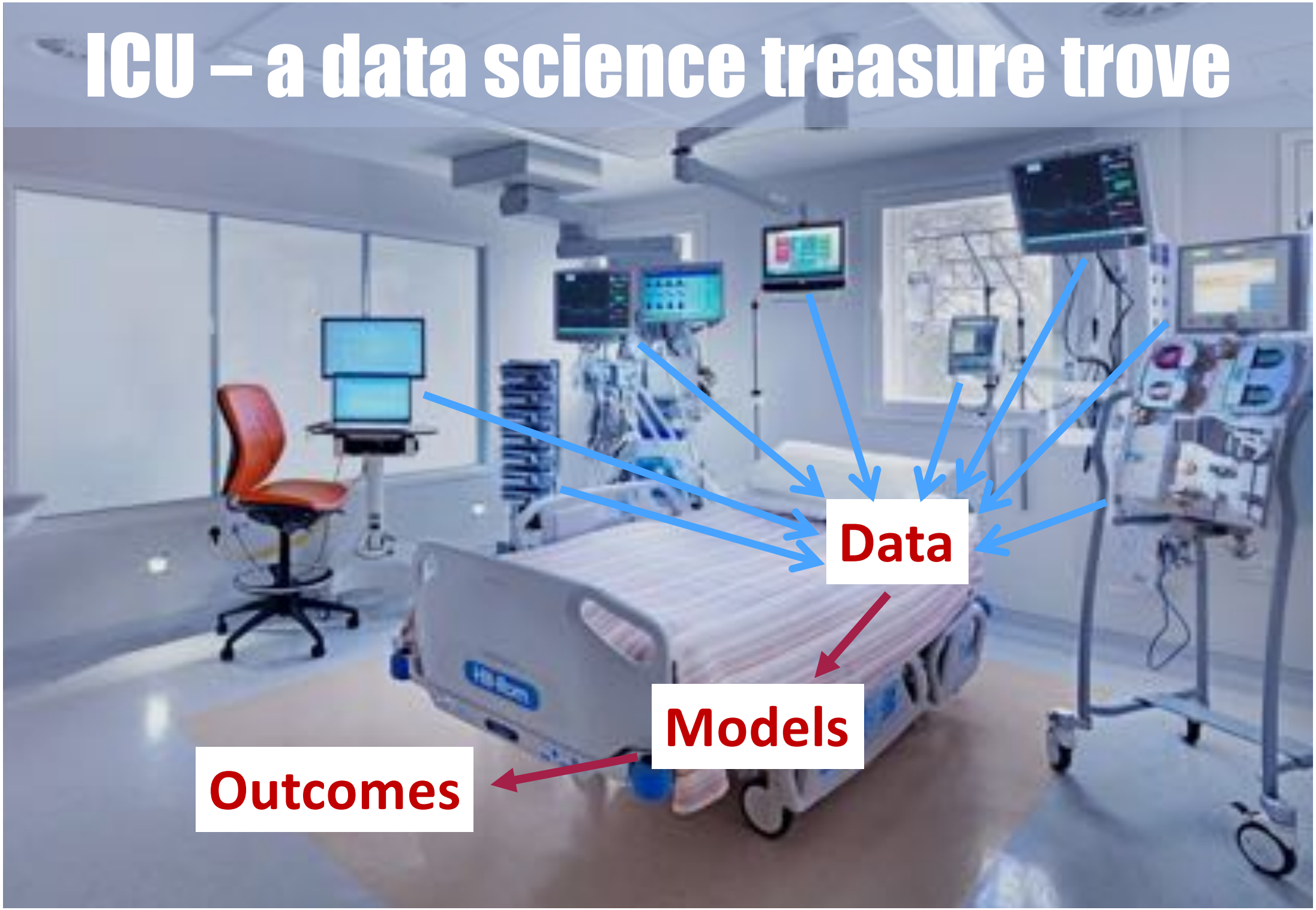
Dr Kelvin Ross
Chairman, KJR
Director, IntelliHQ
kelvin.ross@kjr.com.au

Dr Kelvin Ross is an entrepreneur, technologist and researcher. He currently holds a number of roles, including Founder & Chairman of KJR, a mid-tier IT consultancy, Adjunct Associate Professor in Intelligent and Integrated Systems at Griffith University, and Director at IntelliHQ, a non-profit innovation Centre focused on Artificial Intelligence in healthcare at Gold Coast University Hospital.

He has over 25 years of experience in advanced technology commencing with safety-critical systems in the military, then moving on to transportation, banking, financial markets, government and healthcare systems. He has participated in several successful and unsuccessful technology startups, as well numerous successful and unsuccessful technology implementation programmes in medium and large enterprises.

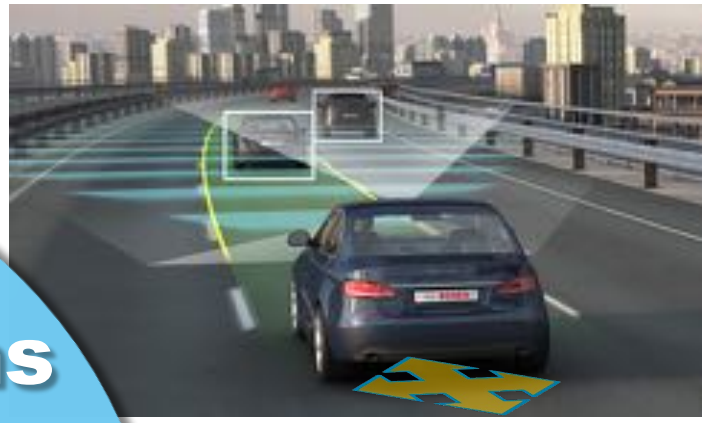

This Is Not About...
Killer Bots
Mass Unemployment
WE NEED WORK
PUP

ICU – a data science treasure trove

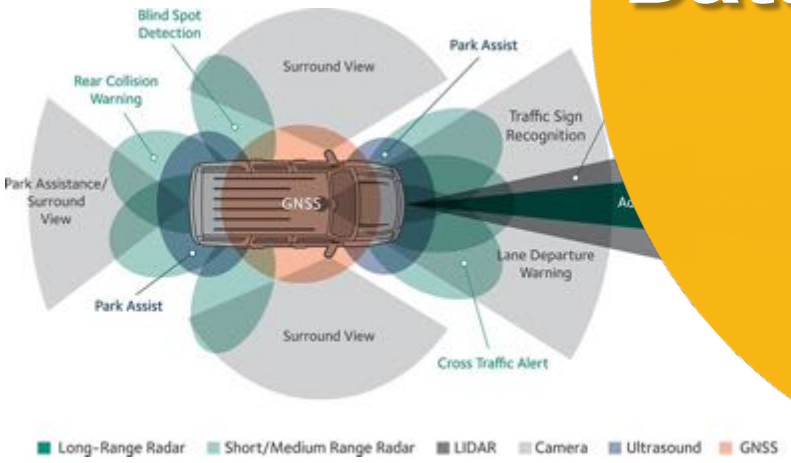


Driverless Car Journey

Interventions



Sensors



Actions

Data

Goals

Outcomes

Precision Medicine Journey

Interventions

Bio-signals

Outcomes


QLD
AI NERD

Fatigue M8
BY AUGMENTED INTELLIGENCE

arterial

Stopped at traffic light

straight

day

PREDICT

Fatigue

+15 mins

+30 mins

+1 hr

+2 hr

Stress

Heart Arythmia

Stroke Risk

Diabetes Risk

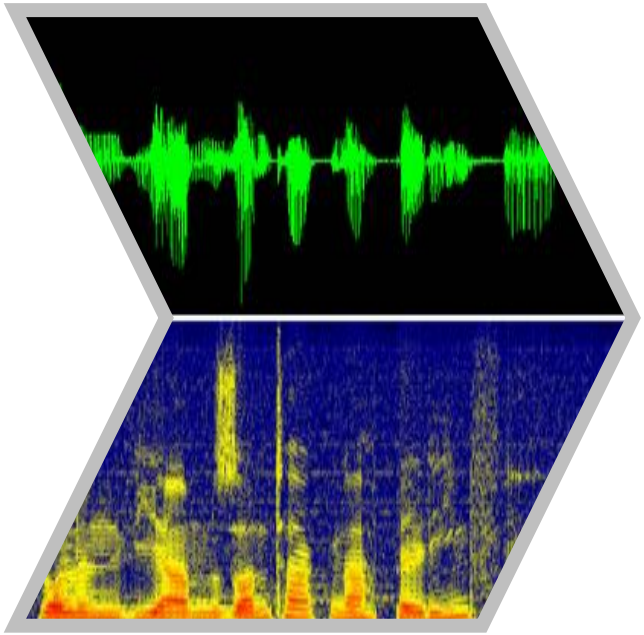
What is Machine Learning

Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed
– Arthur Samuel (1959)

DATA SCIENCE

MACHINE LEARNING

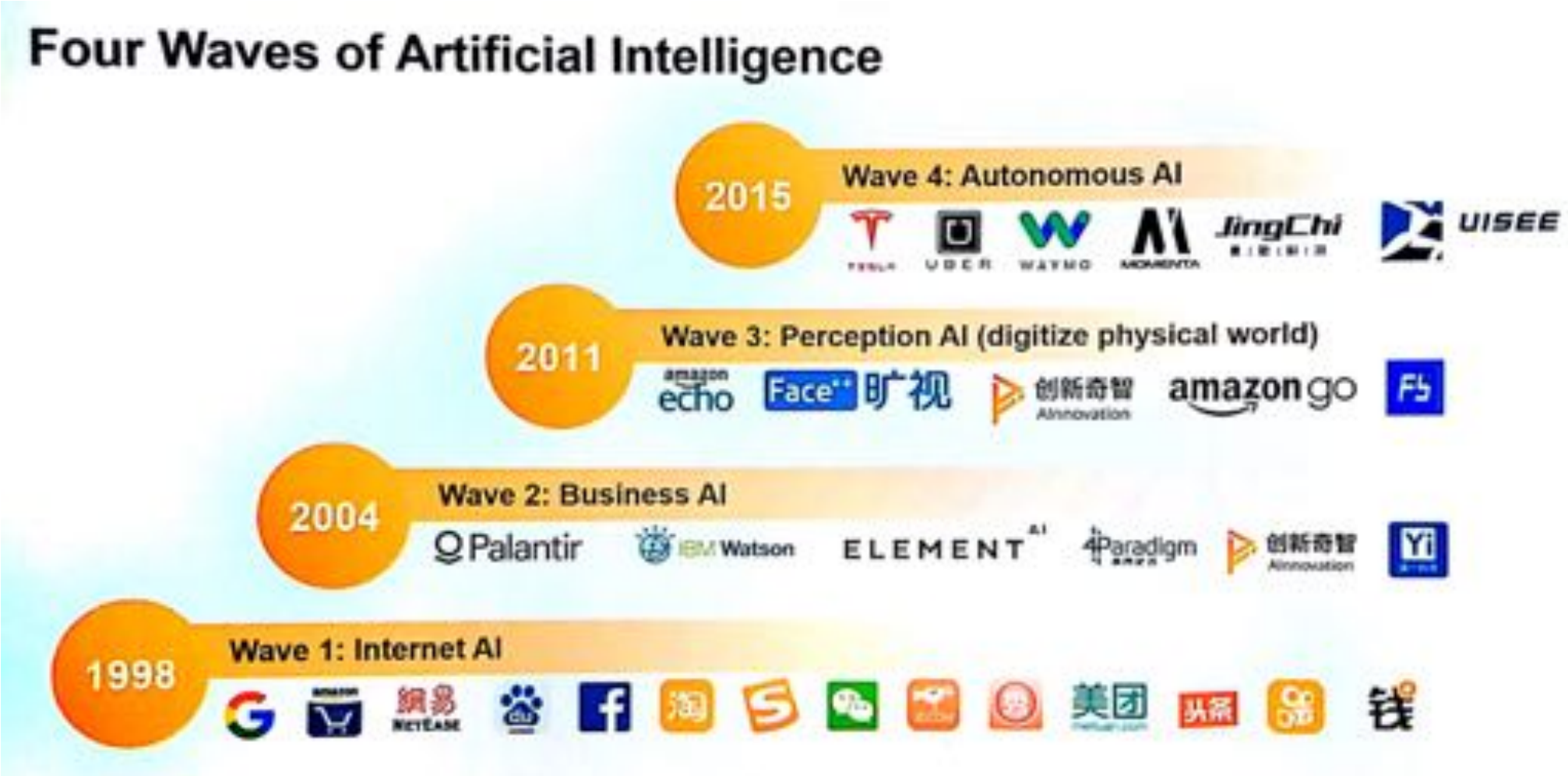
ARTIFICIAL INTELLIGENCE



INSIGHTS

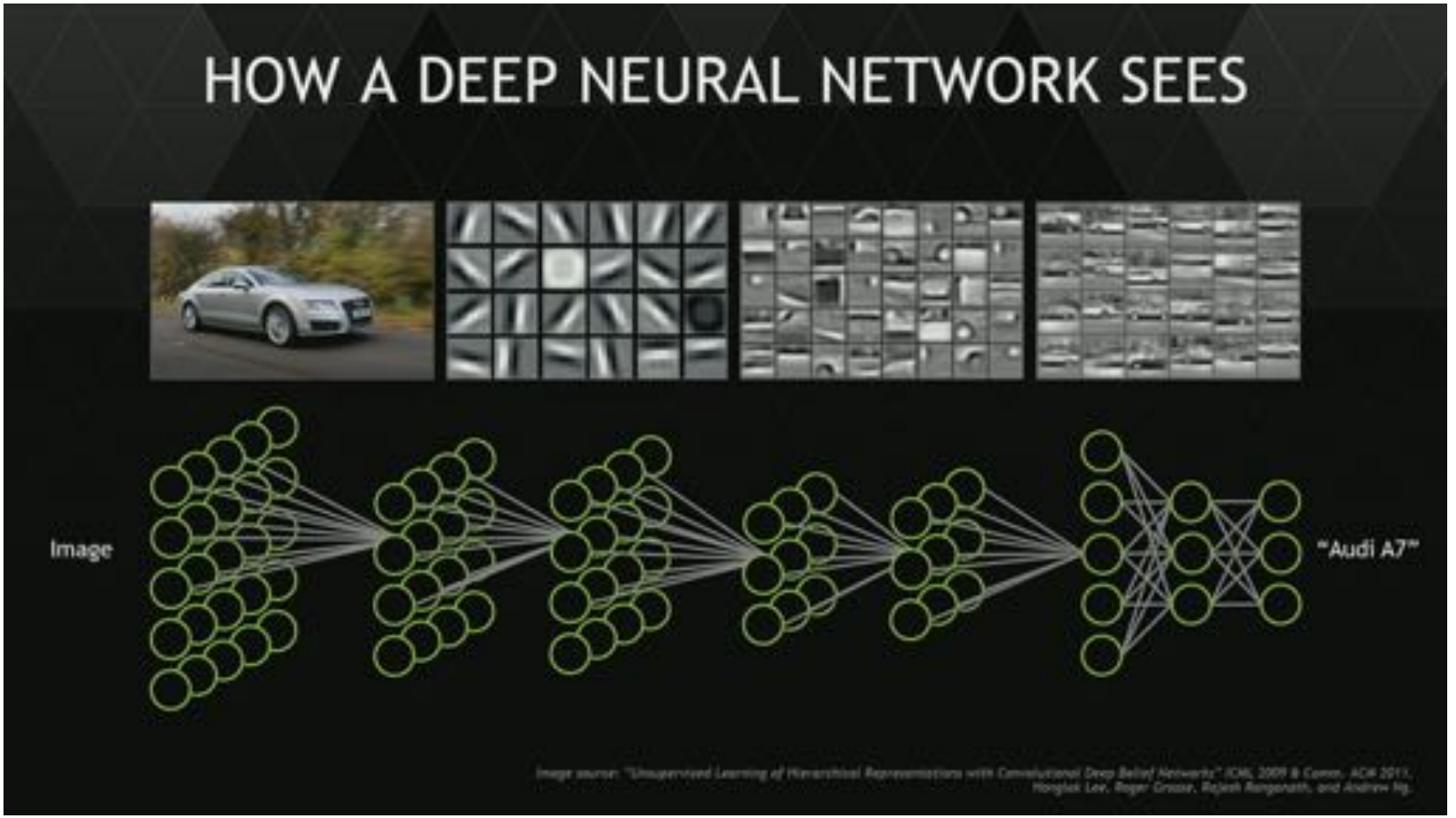
PREDICTIONS

ACTIONS

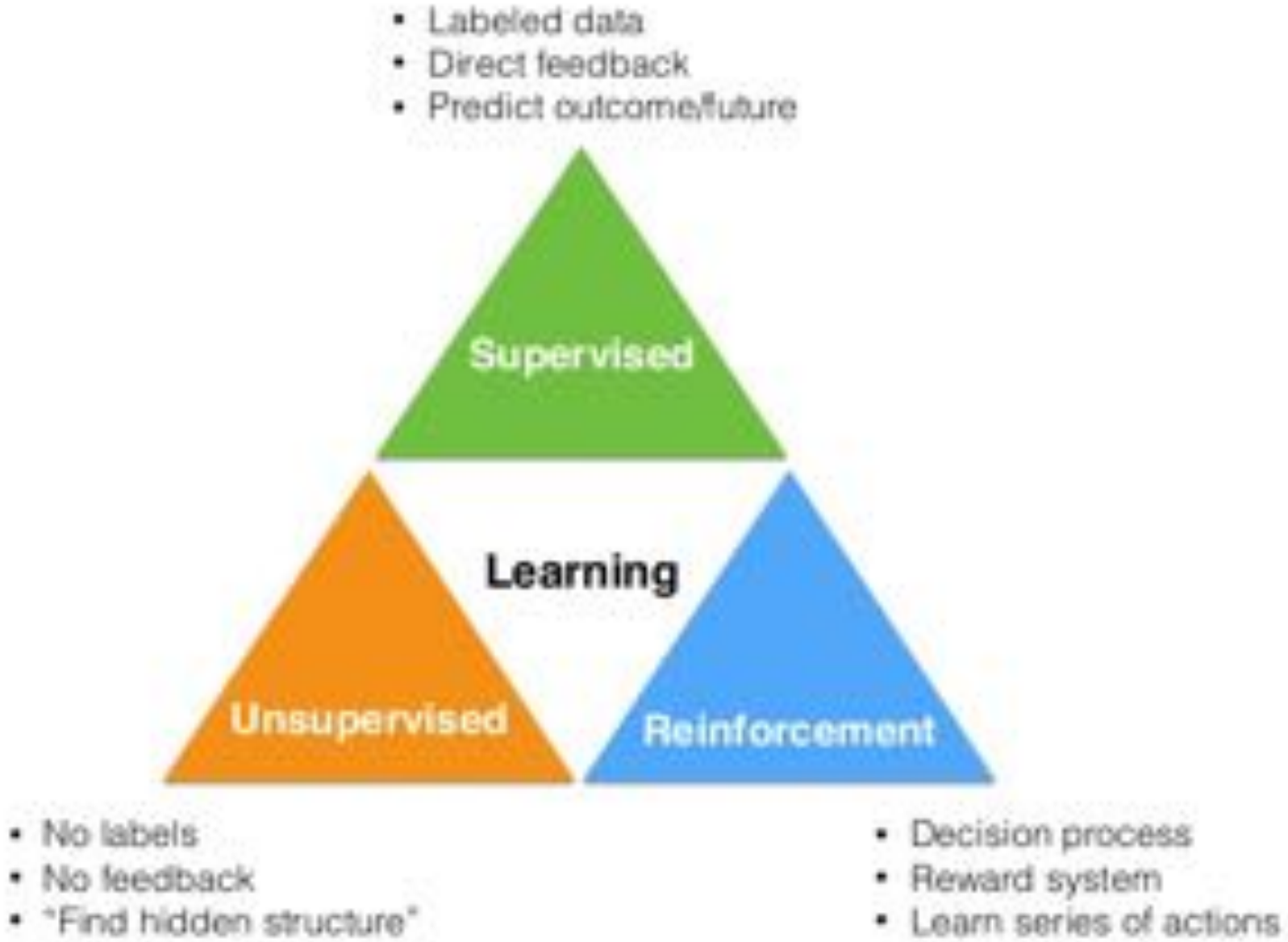


Source: Sinovation Ventures, https://pandaily.com/kai-fu-lee-9-years-of-sinovation-ventures-parallel-tech-universes-and-vc-ai/

Deep Learning

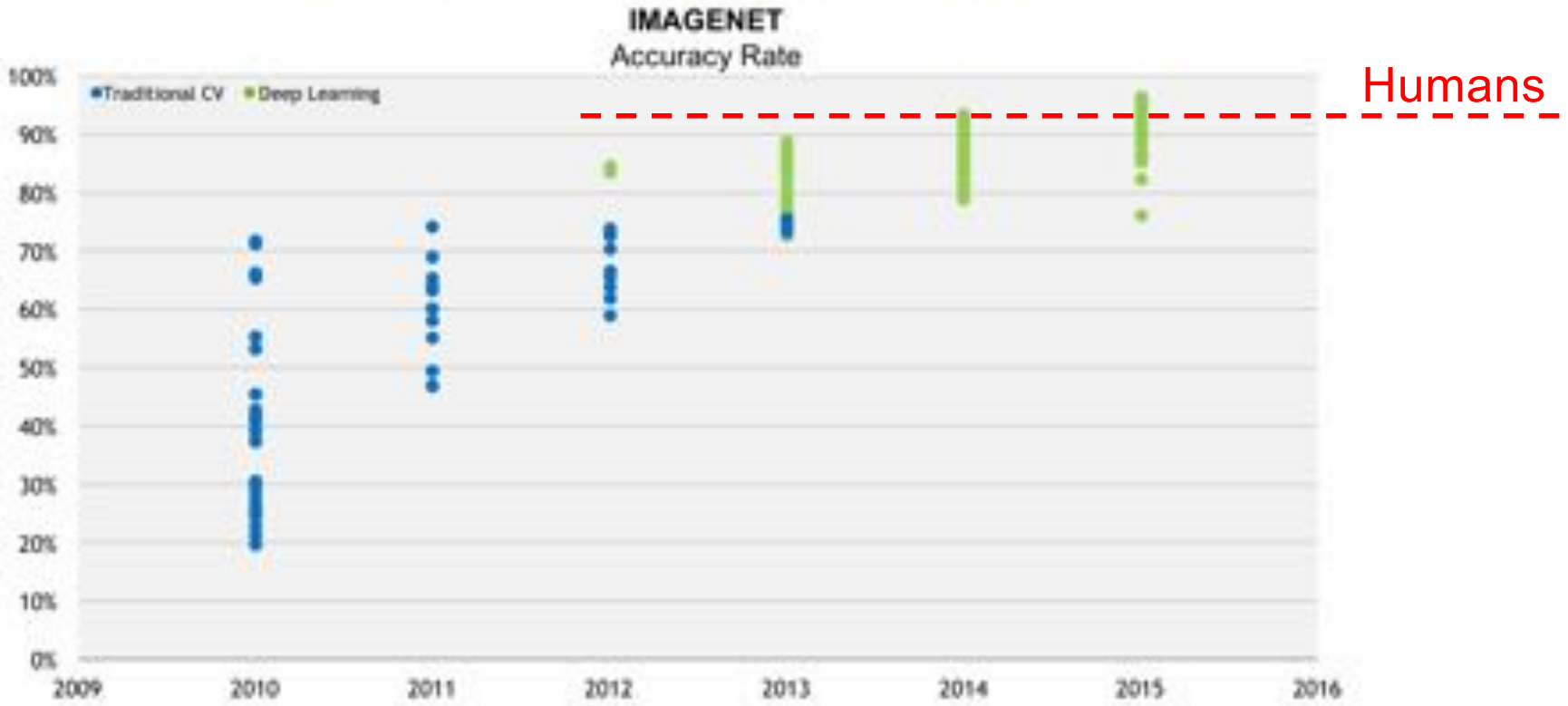


Types of Machine Learning

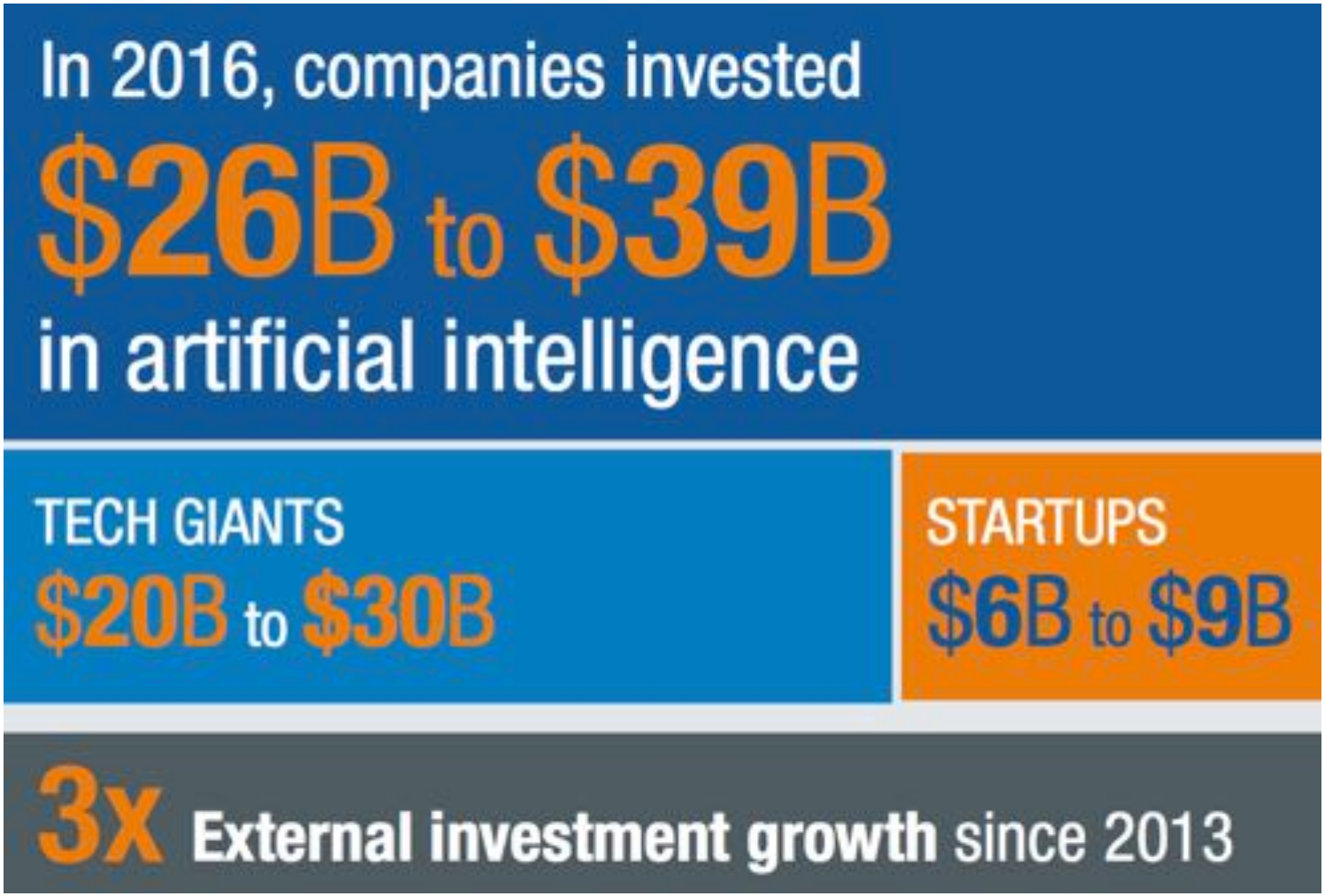


DEEP LEARNING FOR VISUAL PERCEPTION

Going from strength to strength



More info: https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence



Source: ARTIFICIAL INTELLIGENCE THE NEXT DIGITAL FRONTIER?, McKinsey Global Institute, June 2017, http://www.mckinsey.com/mgi

Race To Rule The World

https://www.forbes.com/sites/forbestechcouncil/2017/12/05/these-seven-countries-are-in-a-race-to-rule-the-world-with-ai/

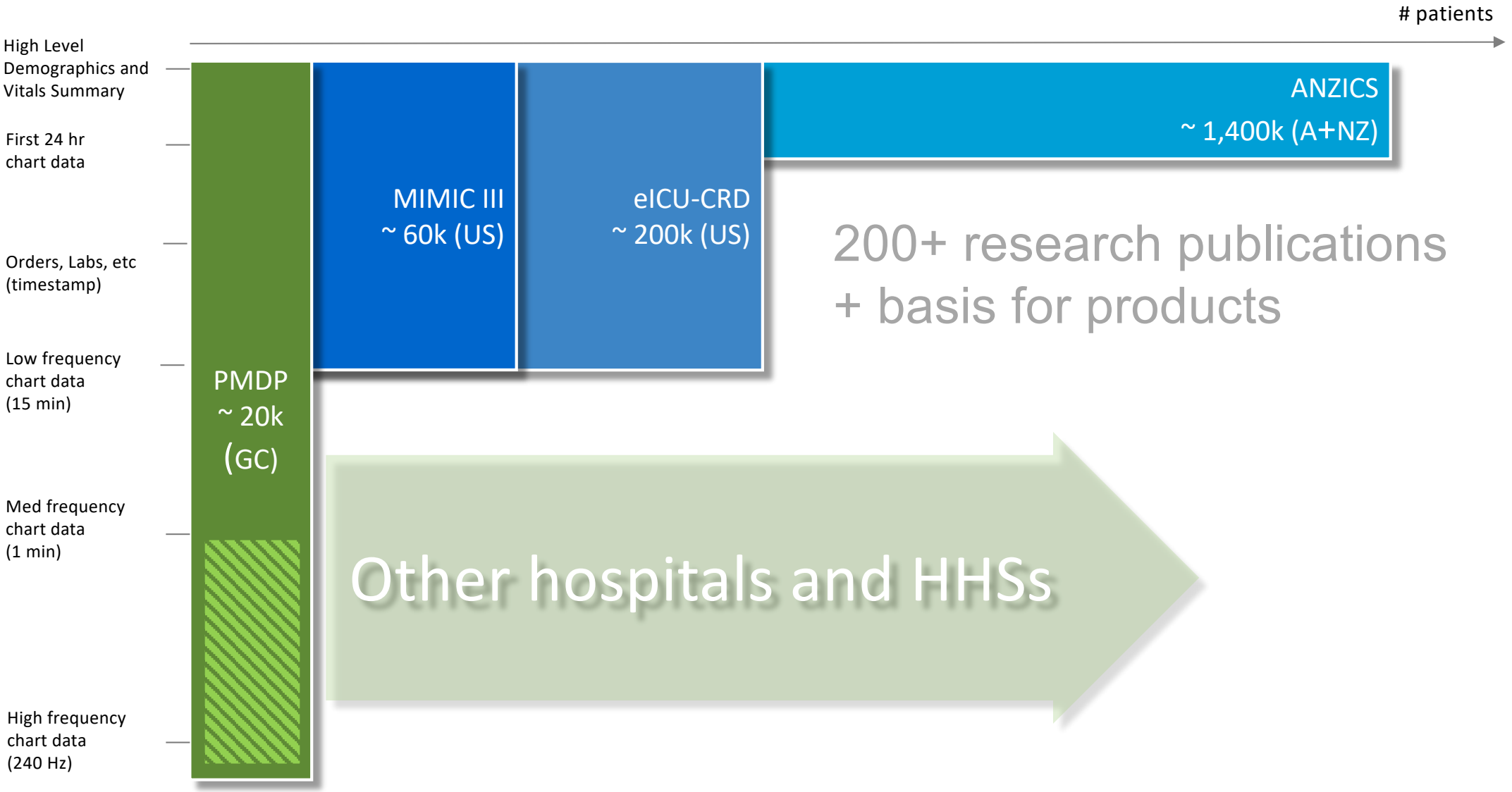
Key to AI Competitiveness

DATA

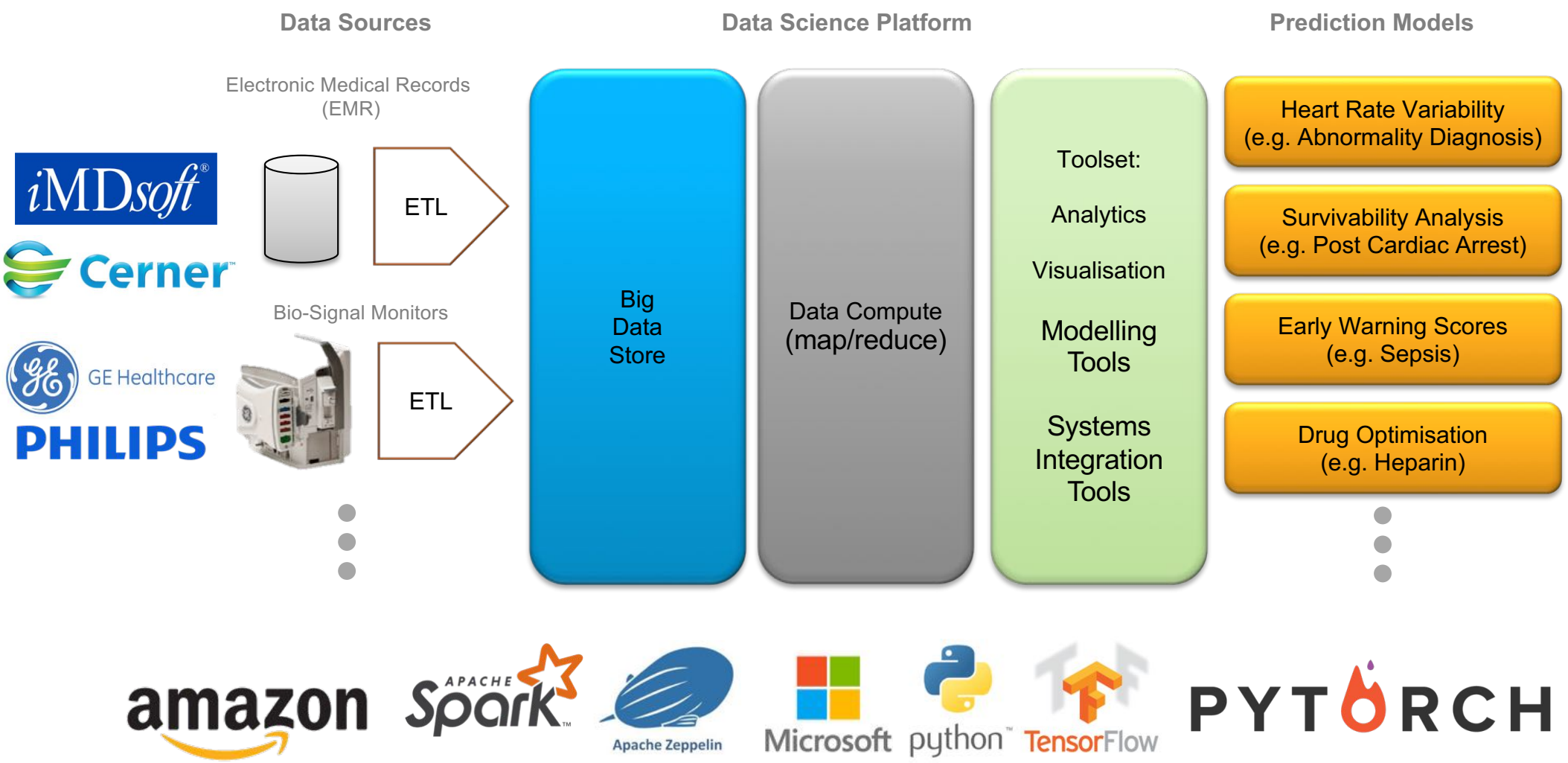
TALENT

Streaming Physiology Data

Deeper ICU Data Sets



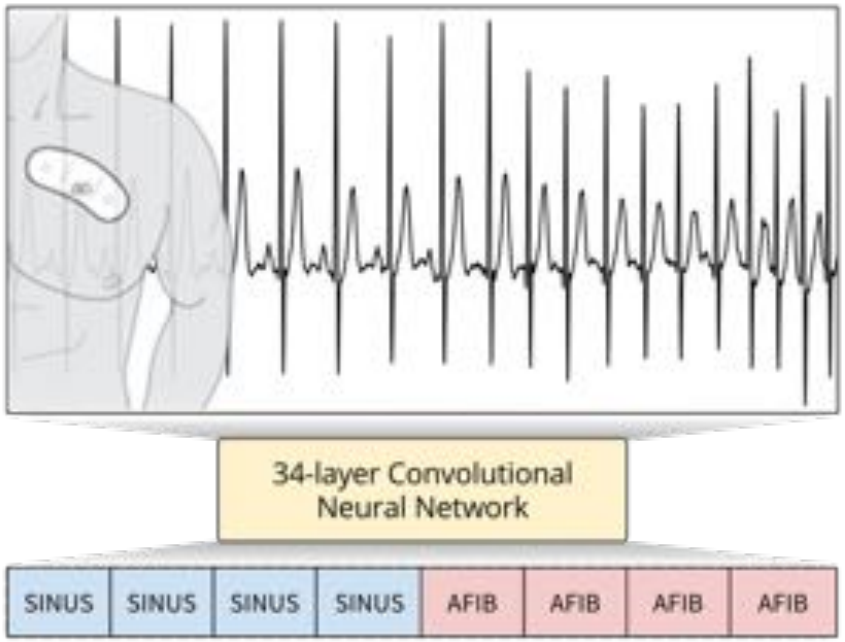
Precision Medicine Data Platform



amazon APACHE Spark Apache Zeppelin Microsoft python TensorFlow PYTORCH

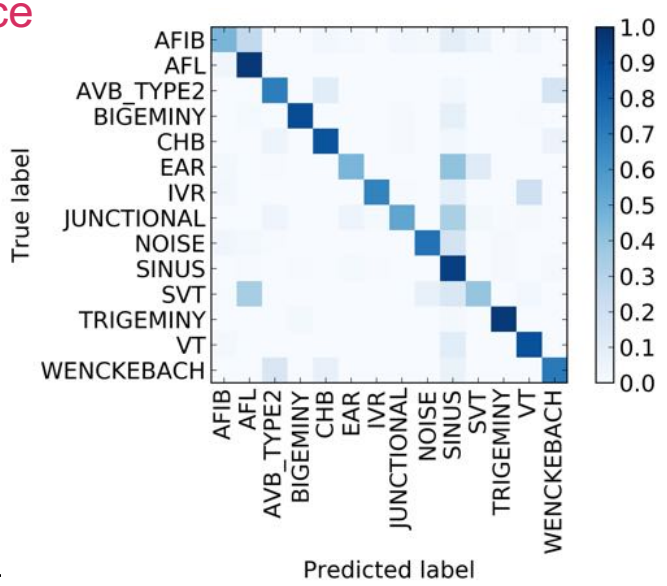
Cardiologist-level arrhythmia detection with CNNs

- Andrew Ng's Stanford group
- 34 layer CNN to diagnose arrhythmias on 64,000 30sec ECGs
- Model outperformed cardiologists



	Seq		Set	
	Model	Cardiol.	Model	Cardiol.
Class-level F1 Score				
AFIB	**0.604**	0.515	**0.667**	0.544
AFL	**0.687**	0.635	**0.679**	0.646
AVB_TYPE2	**0.689**	0.535	**0.656**	0.529
BIGEMINY	**0.897**	0.837	**0.870**	0.849
CHB	**0.843**	0.701	**0.852**	0.685
EAR	**0.519**	0.476	**0.571**	0.529
IVR	**0.761**	0.632	**0.774**	0.720
JUNCTIONAL	0.670	**0.684**	**0.783**	0.674
NOISE	**0.823**	0.768	**0.704**	0.689
SINUS	**0.879**	0.847	**0.939**	0.907
SVT	**0.477**	0.449	**0.658**	0.556
TRIGEMINY	**0.908**	0.843	**0.870**	0.816
VT	0.506	**0.566**	0.694	**0.769**
WENCKEBACH	**0.709**	0.593	**0.806**	0.736
Aggregate Results				
Precision (PPV)	**0.800**	0.723	**0.809**	0.763
Recall (Sensitivity)	**0.784**	0.724	**0.827**	0.744
F1	**0.776**	0.719	**0.809**	0.751

Note – human expert performance



https://stanfordmlgroup.github.io/projects/ecg

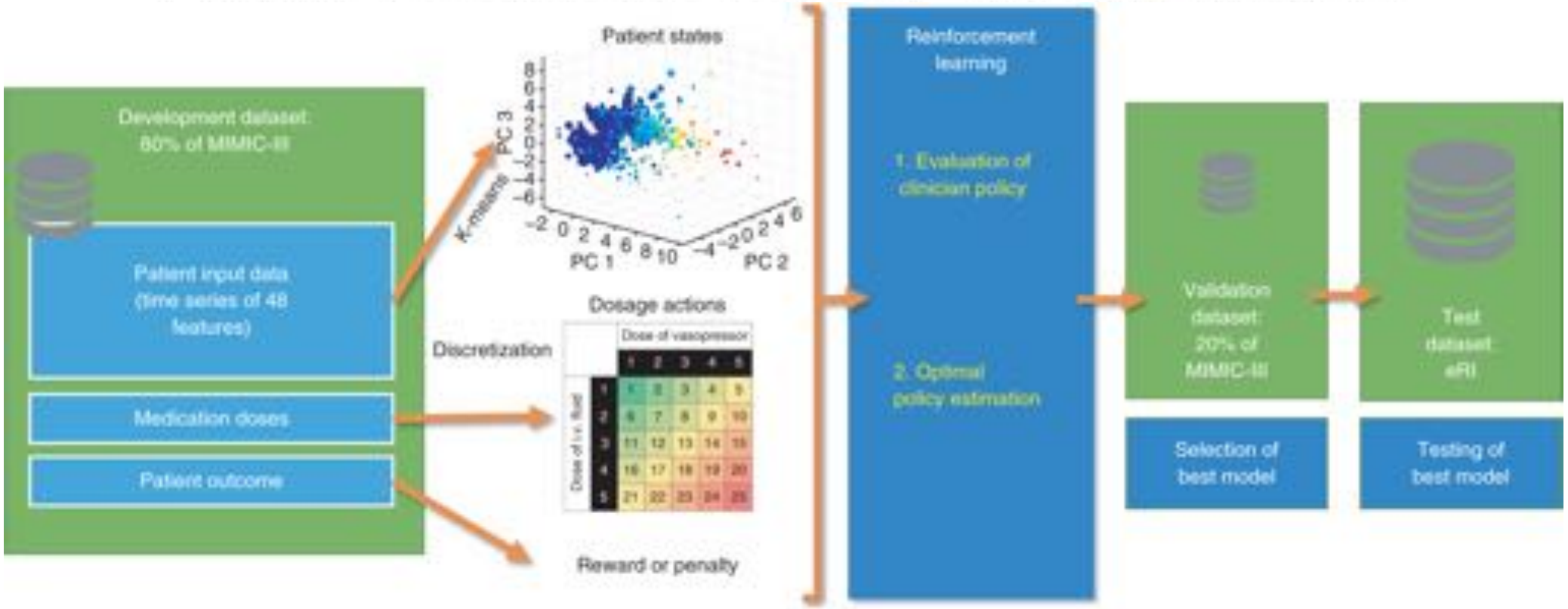
nature medicine

ARTICLES

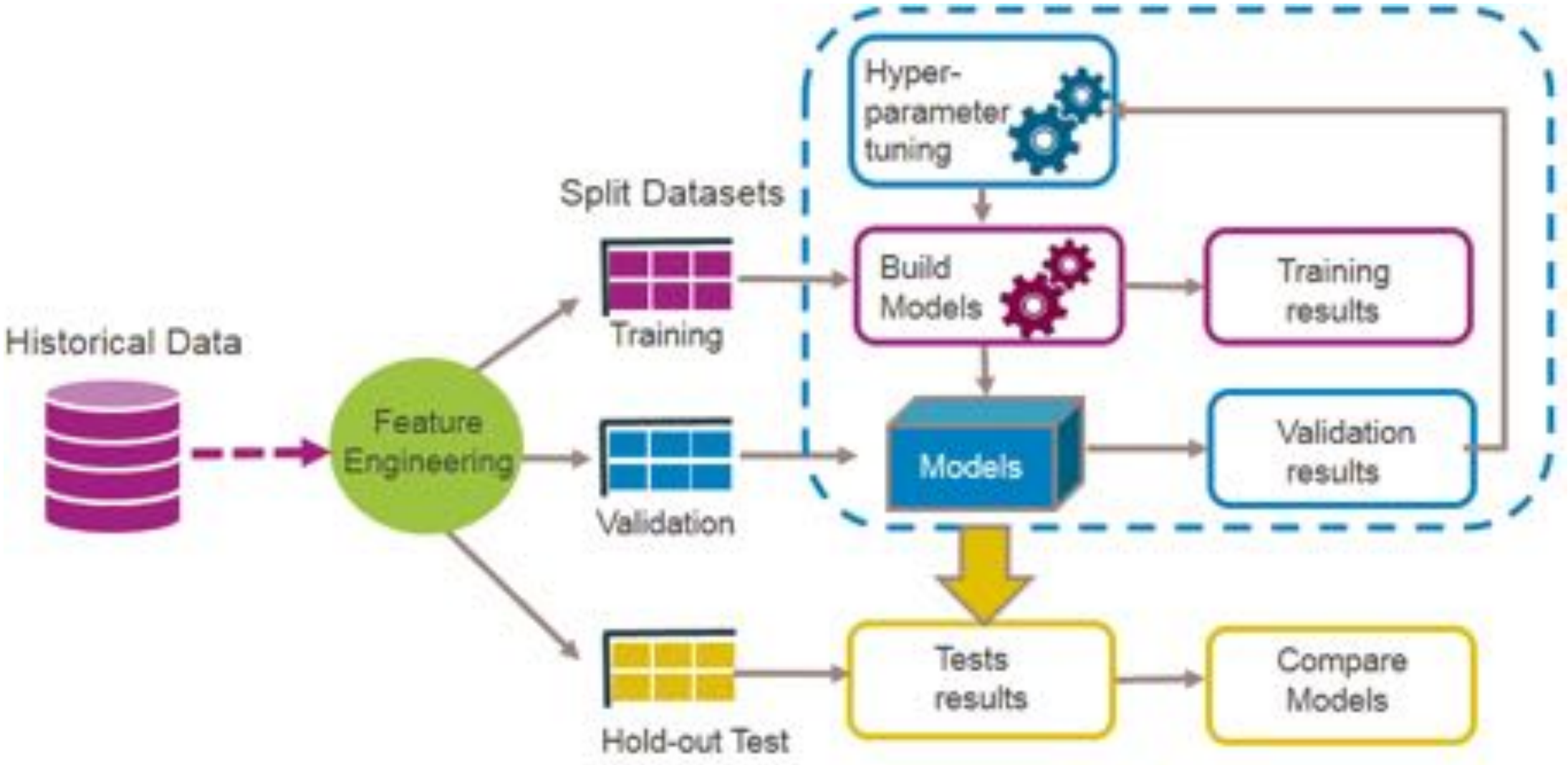
https://doi.org/10.1038/s41591-018-0213-5

The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care

Matthieu Komorowski[1,2,3], Leo A. Celi[3,4], Omar Badawi[3,5,6], Anthony C. Gordon[1*] and A. Aldo Faisal[2,7,8,9*]

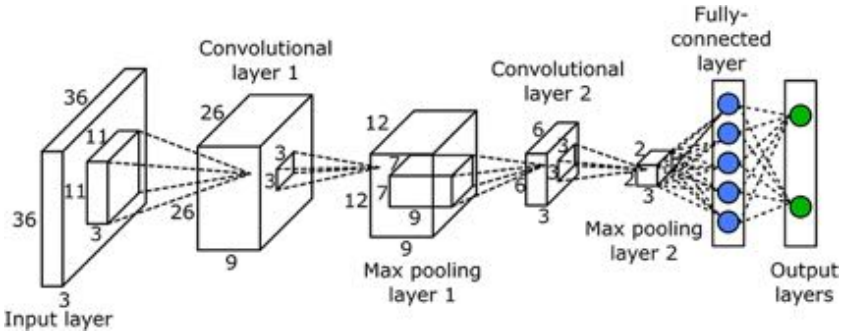


ML Process


Hyper-
parameter
tuning
Split Datasets
Build
Models
Training
results
Historical Data
Training
Feature
Engineering
Validation
Models
Validation
results
Tests
results
Compare
Models
Hold-out Test

Training

inference

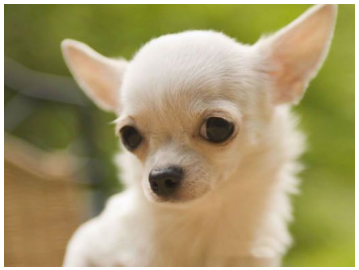


back propagation

Class	Truth	Infer	Loss
Chihuahua	1.0	0.214	+0.786
Labrador	0.0	0.769	-0.769

















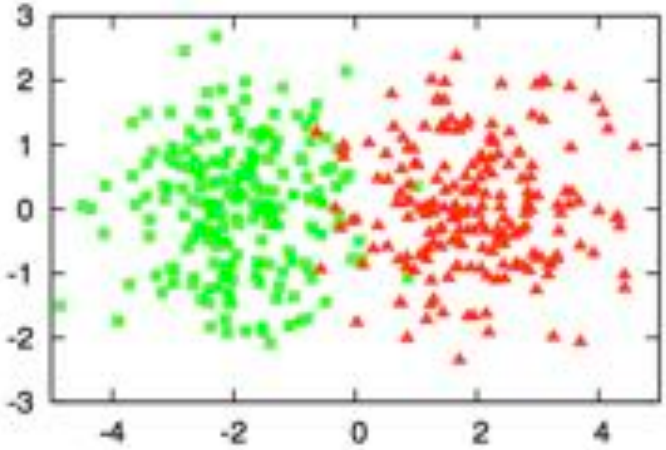
Blingee


myKnitt.com

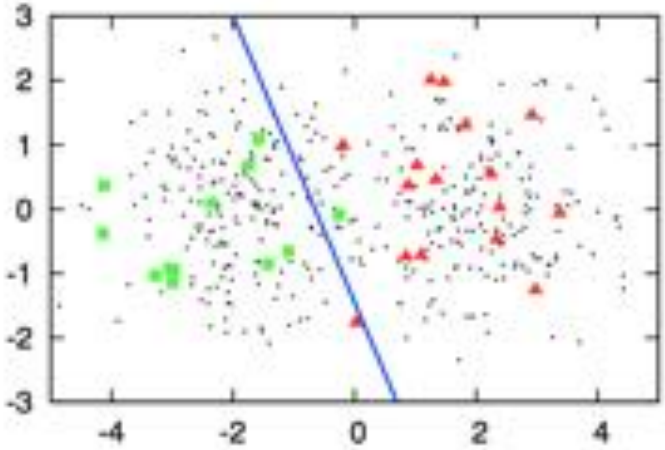

Chihuahua or Muffin


40 days – AlphaGo Zero surpasses all previous versions, becomes the best Go player in the world
36 hours – AlphaGo Zero reaches level of Alpha Go Lee, which beat world champion Lee Sedol in 2016
72 hours – AlphaGo Zero beats Alpha Go Lee, 100:0
Elo Rating
5000
4000
3000
2000
1000
0
-1000
-2000
Training days
0
5
10
15
20
25
30
35
40
AlphaGo Zero 40 blocks
AlphaGo Lee
AlphaGo Master

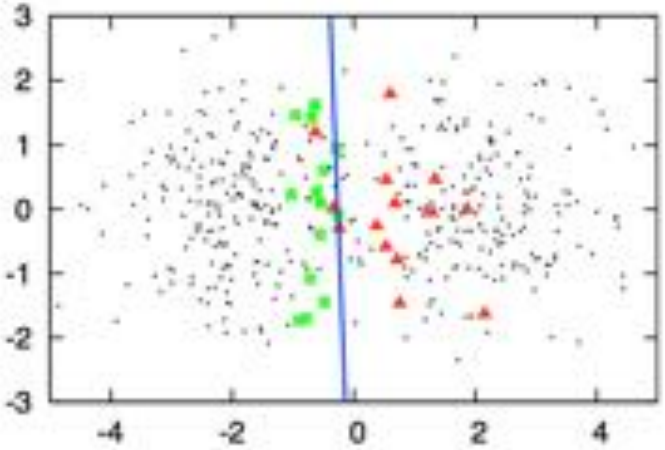
Active Learning



All 400



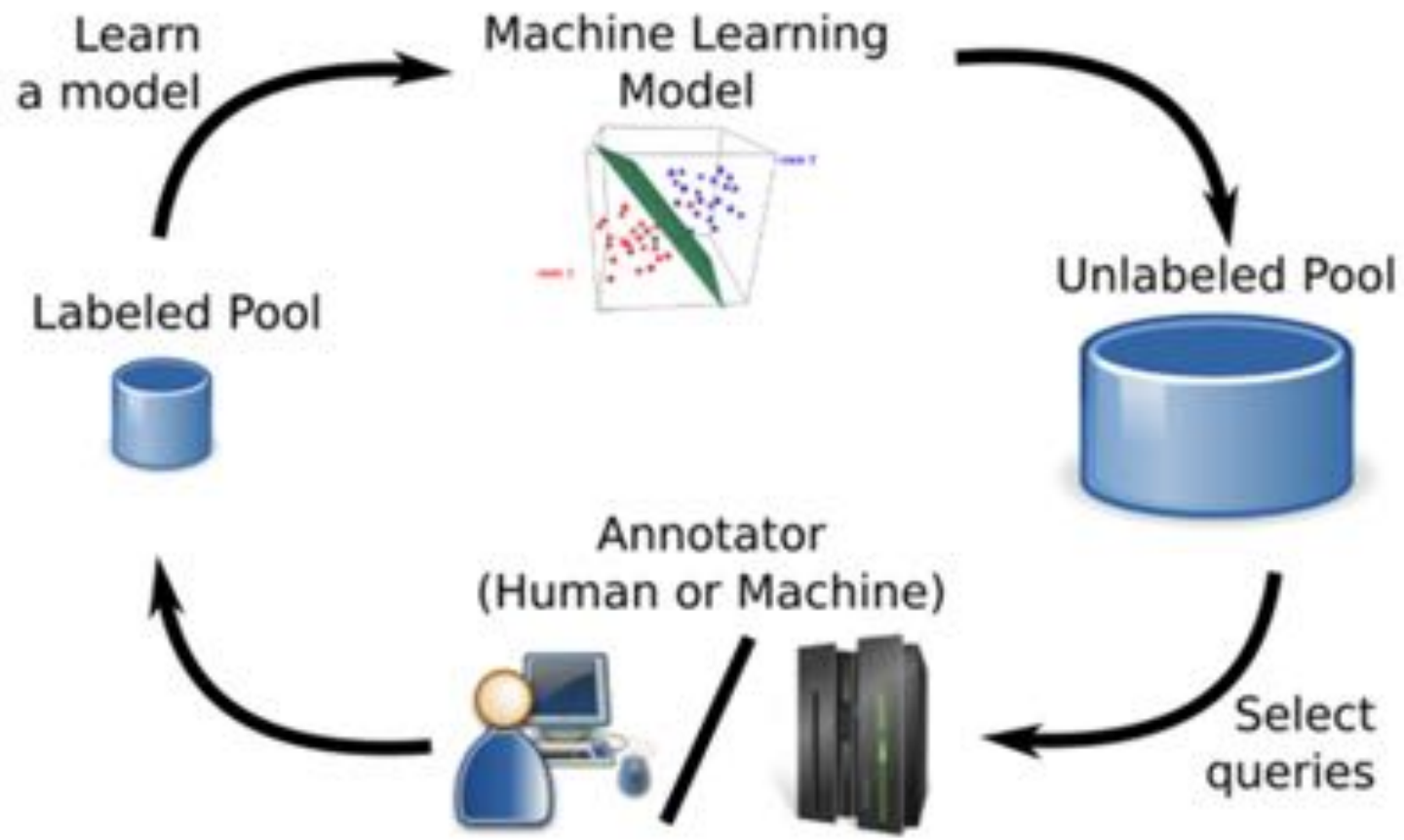
30 Random Selected
70% accuracy



30 Active Selected
90% accuracy

https://www.datacamp.com/community/tutorials/active-learning

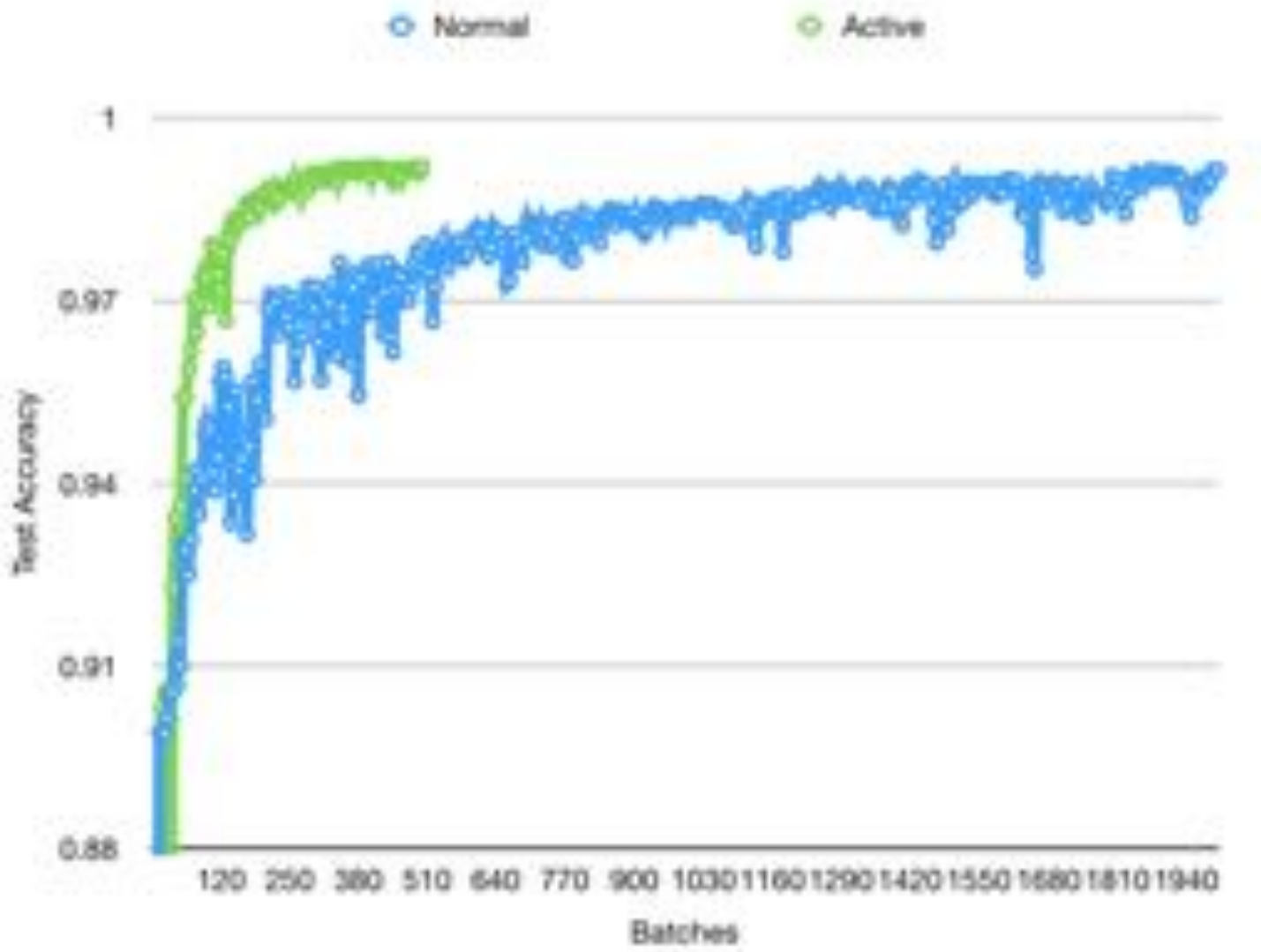
Active Learning



- Uncertainty
- Decision Boundaries
- Similarity

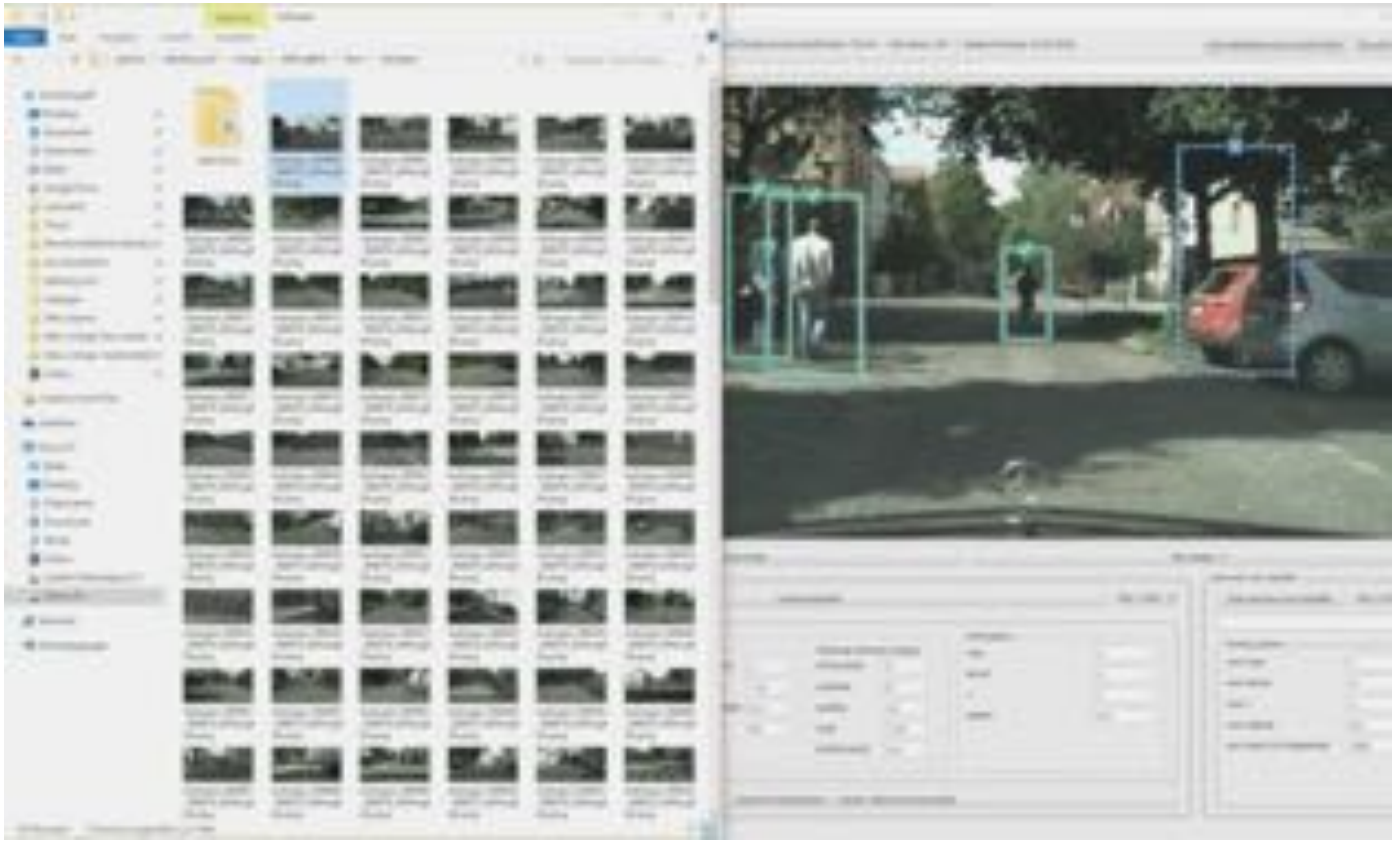
https://www.datacamp.com/community/tutorials/active-learning

Active Learning



https://becominghuman.ai/accelerate-machine-learning-with-active-learning-96cea4b72fdb

Labeling Platforms



appen

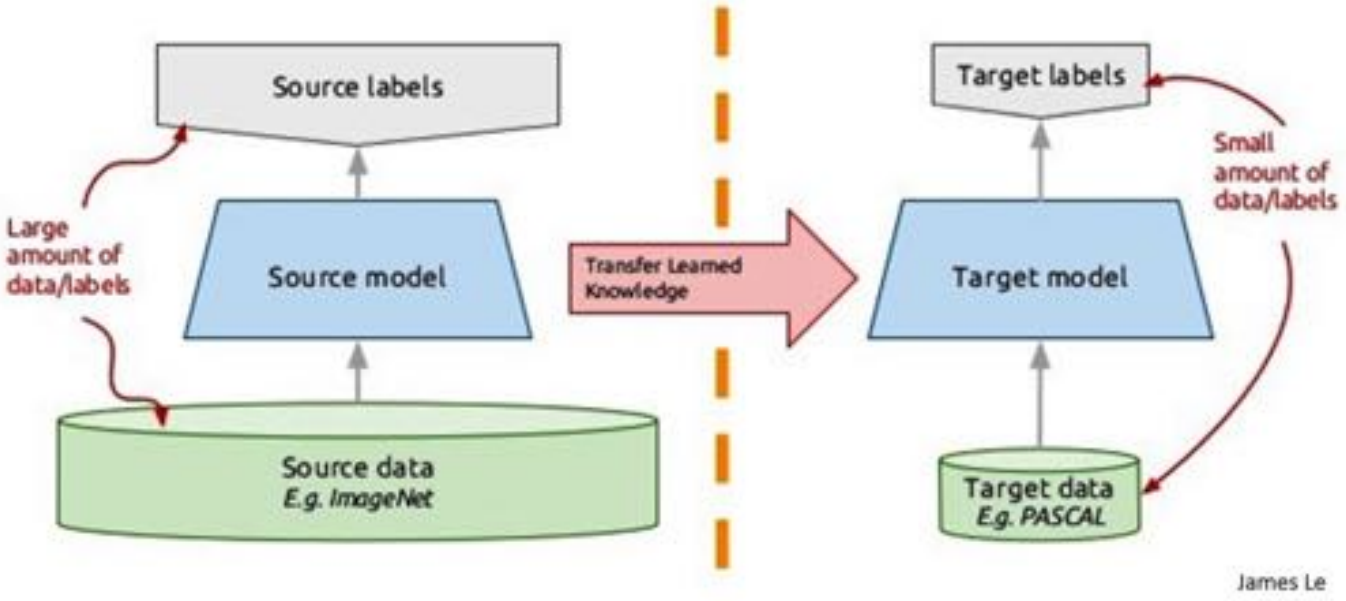
figure eight

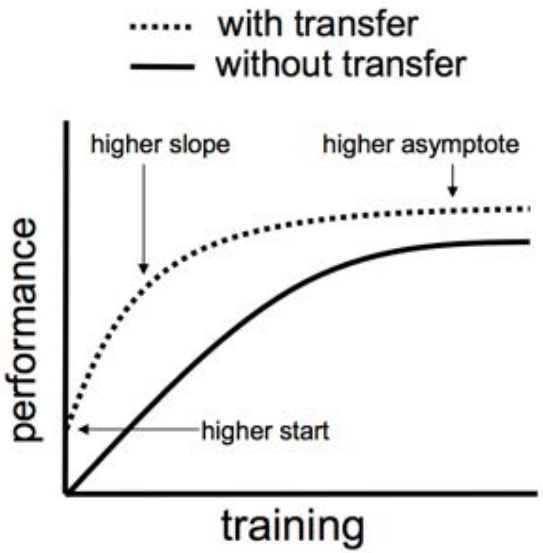
Active Learning

Other references:

- https://sfu-db.github.io/cmpt884-fall16/Lectures/884_presentation_on_active_learning.pdf
- https://www.youtube.com/watch?v=8Jwp4_WbRio
- https://www.datacamp.com/community/tutorials/active-learning
- https://www.slideshare.net/alx_voropaev/introduction-to-active-learning

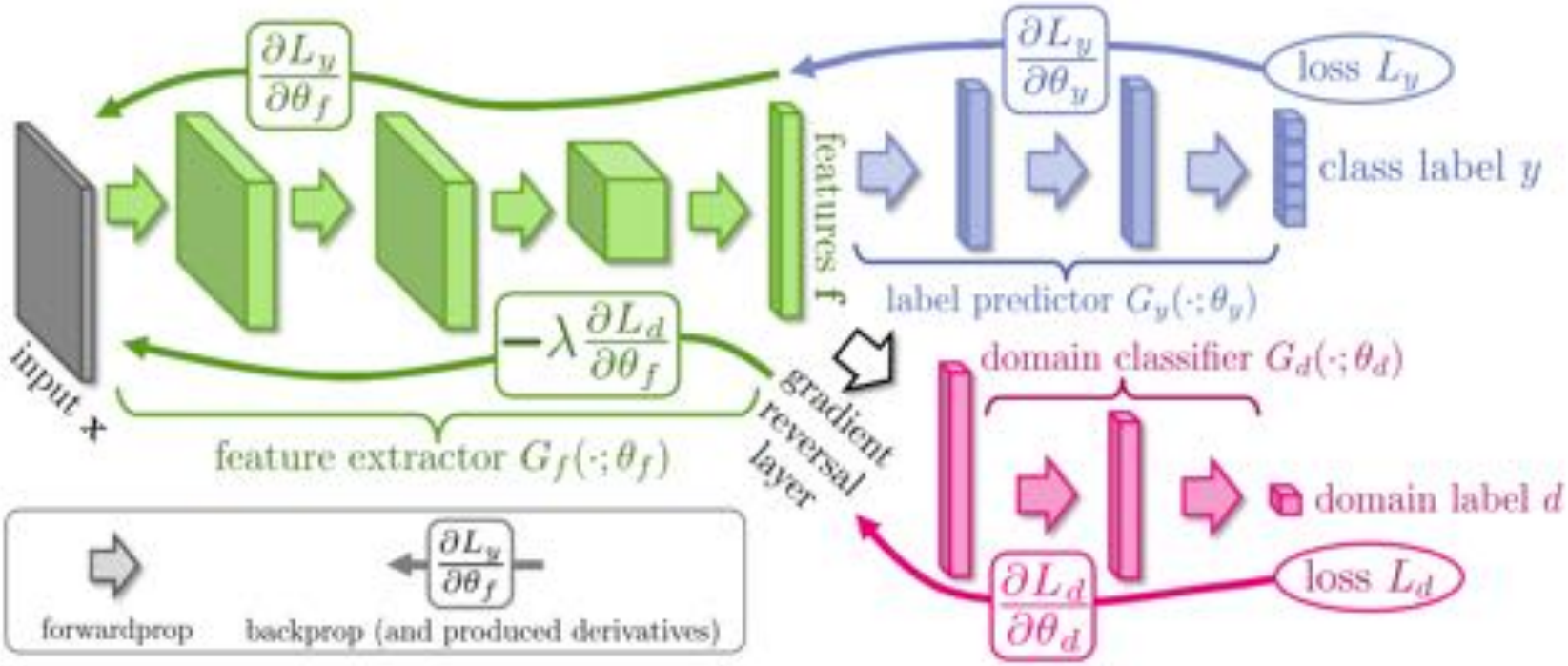
Transfer Learning





https://www.datasciencecentral.com/profiles/blogs/transfer-learning-deep-learning-for-everyone

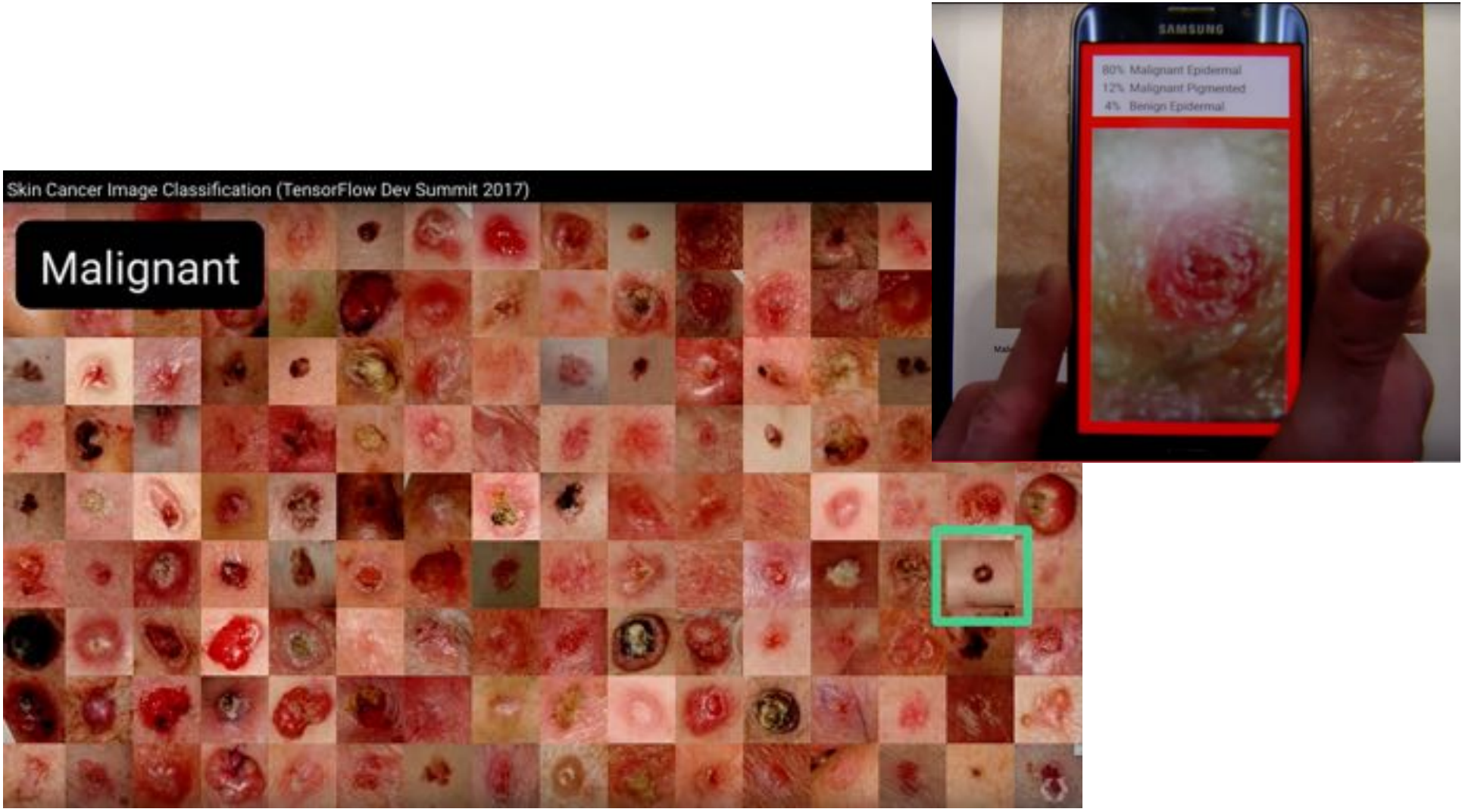
Transfer Learning



http://ruder.io/transfer-learning/

Skin Cancer Detection

Stanford/Tensorflow



https://www.youtube.com/watch?v=toK1OSLep3s

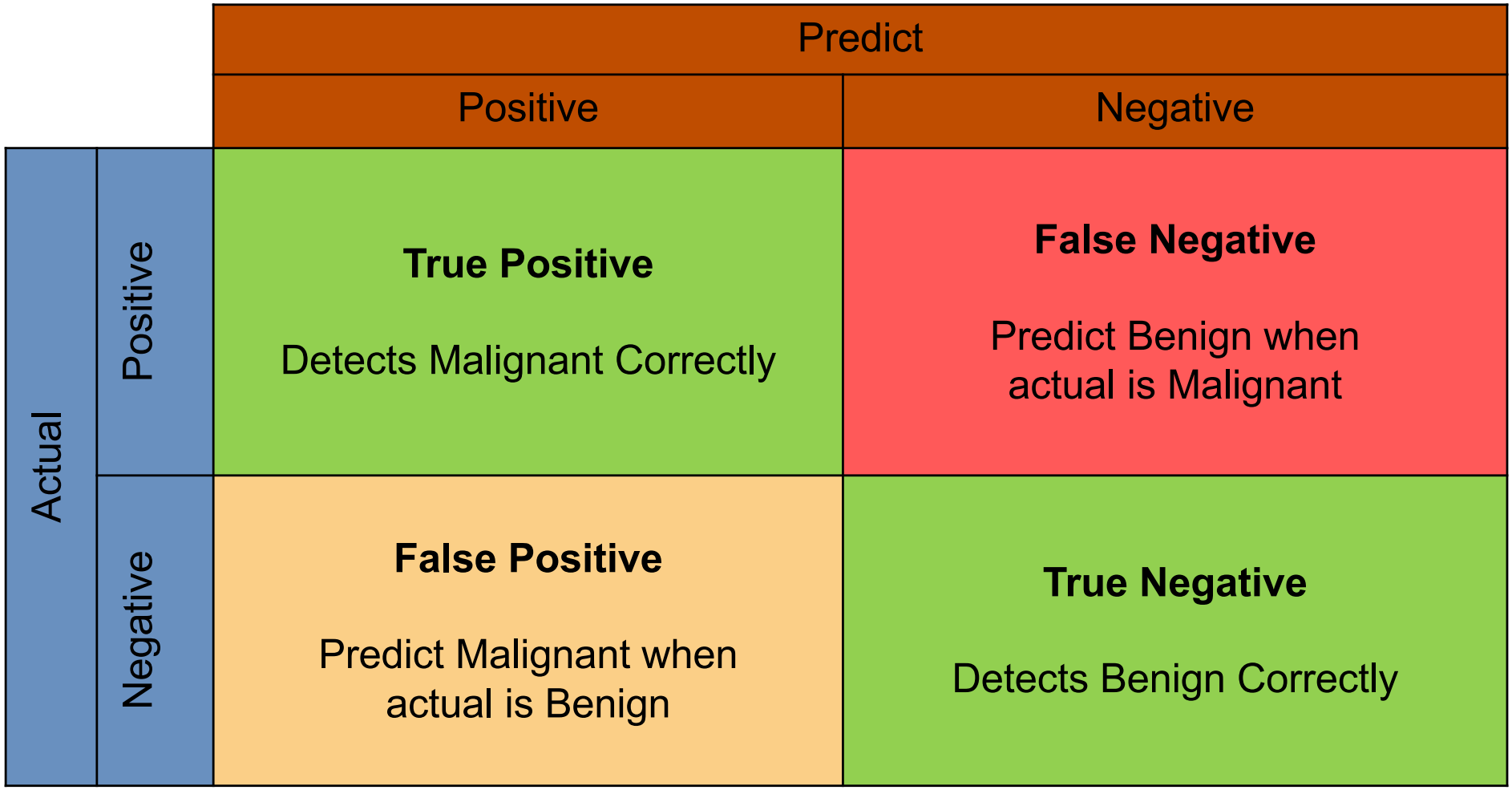
Disease Labels

Taxonomy

(subset)

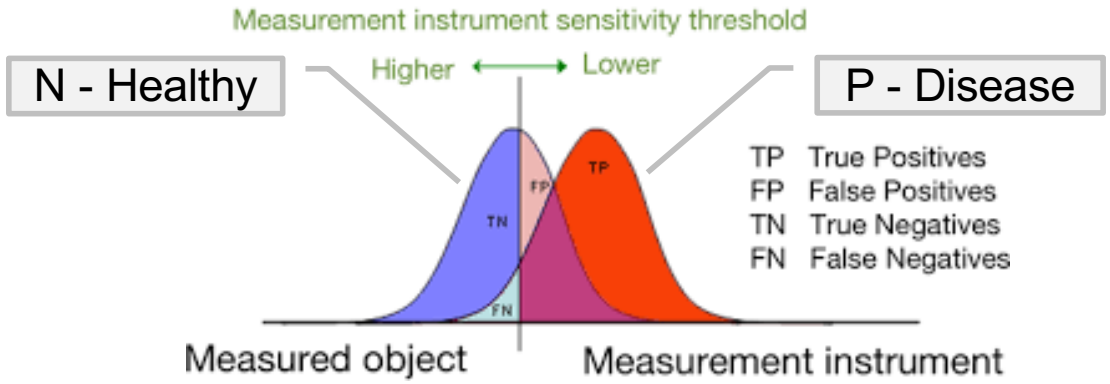
129k images
2k diseases

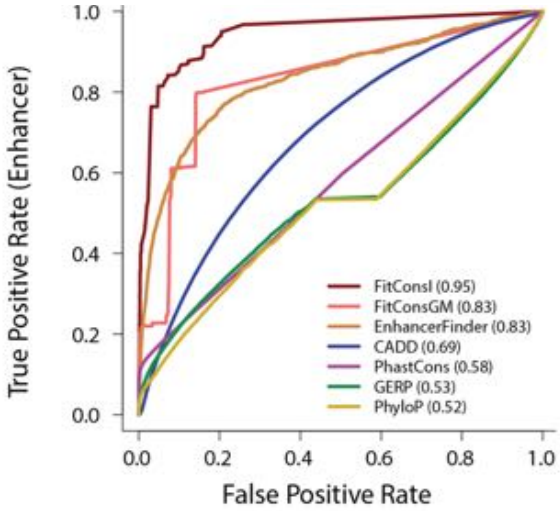
Prediction Accuracy



Accuracy Metrics

		Predicted condition			
	Total population	Predicted Condition positive	Predicted Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
True condition	condition positive	True positive	False Negative (Type II error)	True positive rate (TPR), Sensitivity, Recall, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$
	condition negative	False Positive (Type I error)	True negative	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$
	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Positive likelihood ratio (LR+) = $\frac{TPR}{FPR}$	Diagnostic odds ratio (DOR) = $\frac{LR+}{LR-}$
		False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$	Negative likelihood ratio (LR−) = $\frac{FNR}{TNR}$	

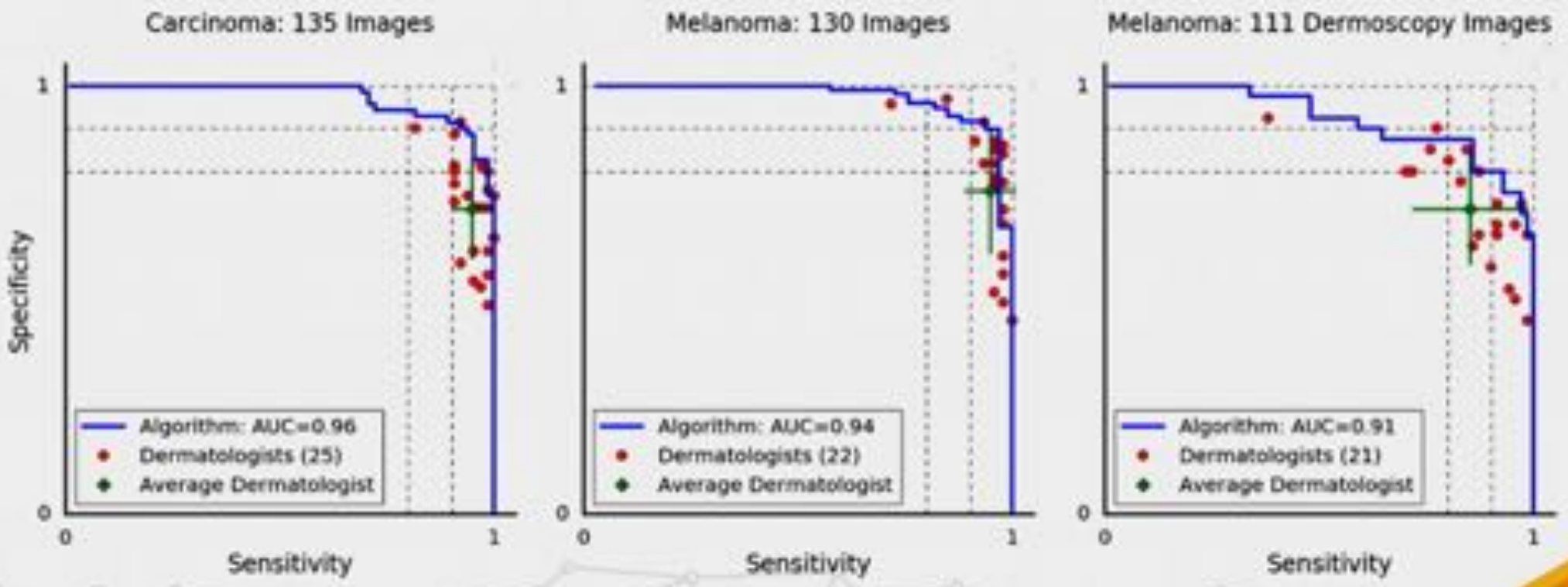




https://en.wikipedia.org/wiki/Receiver_operating_characteristic

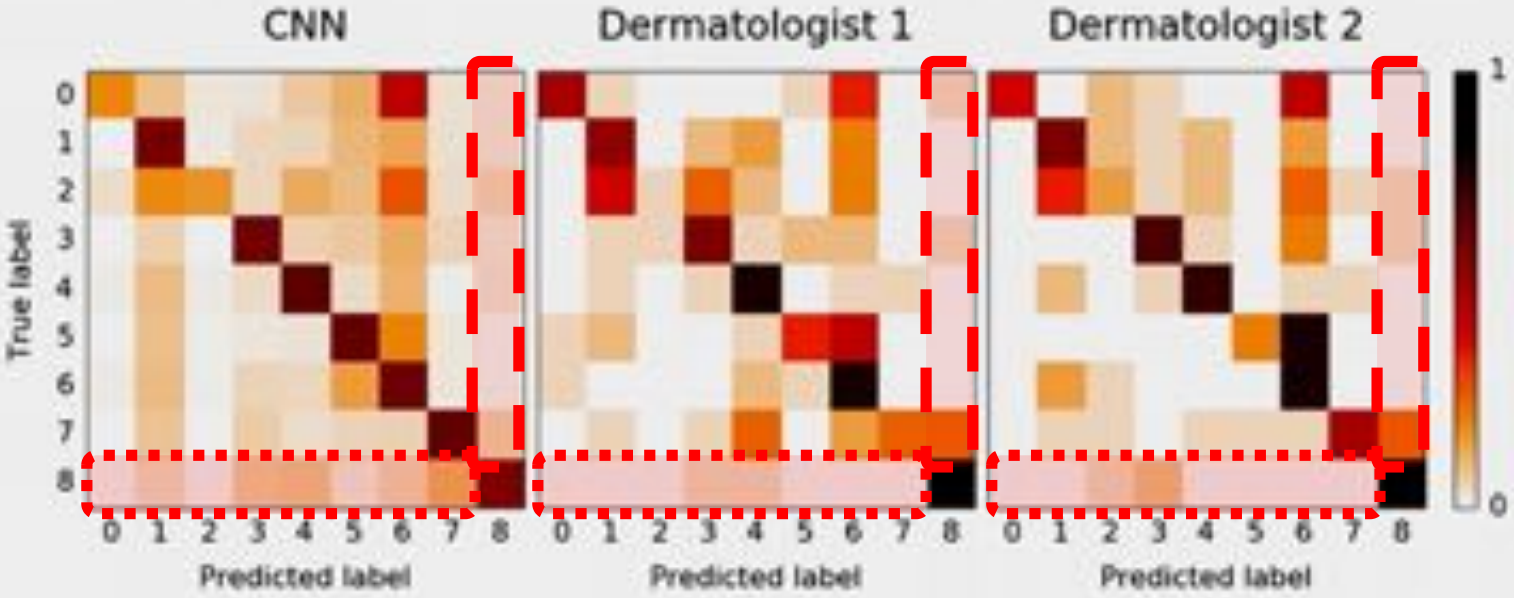
Receiver Operator Curves

Results



Some Errors Worse Than Others

Confusion Matrices



0	Benign Dermal
1	Malignant Dermal
2	Benign Epidermal
3	Malignant Carcinoma
4	Genodermatosis
5	Cutaneous Lymphoma
6	Inflammatory
7	Benign Pigmented
8	Malignant Melanoma

False Positive

False Negative

Alarm Fatigue

It's quite alarming...

» Alarm fatigue occurs when hospital staff become desensitized to alarm alerts causing missed alarms or delayed response

216 deaths

» Between 2005 and 2010, alarm fatigue caused 216 hospital deaths

#1 Hazard

» Voted the top technology hazard of 2012 by ECRI, beating out 2011's radiation exposure

942 alarms each day

» 942 alarms sound off each day in a typical 15-bed unit

1 alarm every 90 seconds

90% are unanswered

» Alarms are reported to be unanswered 90% of the time

» It's no wonder that alarm fatigue is so prevalent

Locate
» Locate the source of each alarm

Limit
» Limit active alarms based on patient needs

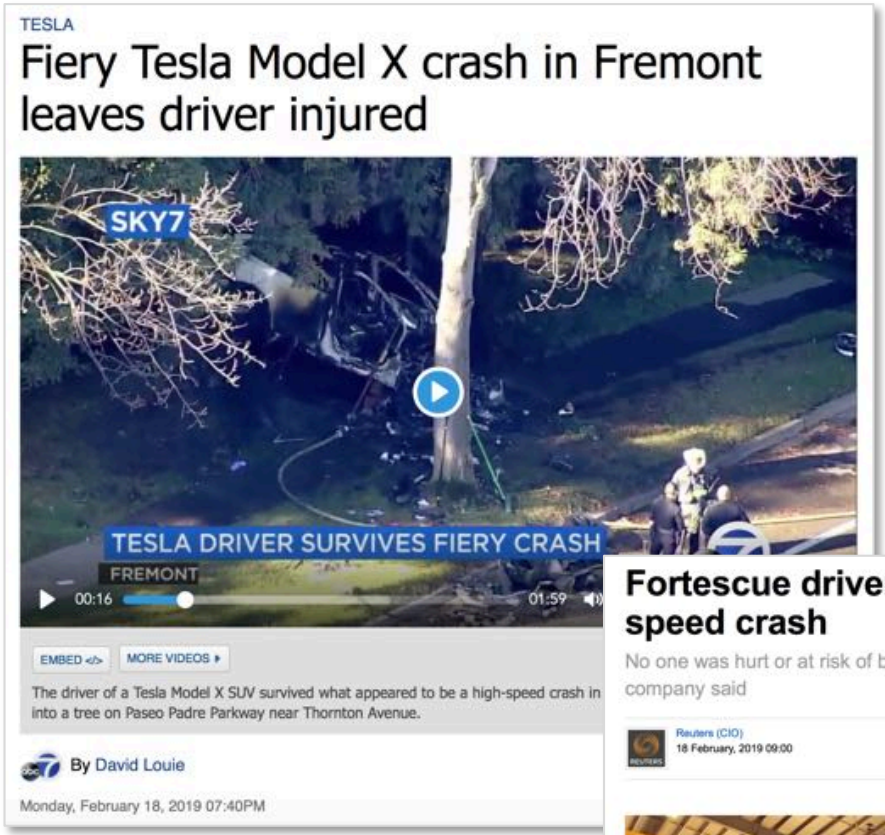
Volume
» Set volume accordingly

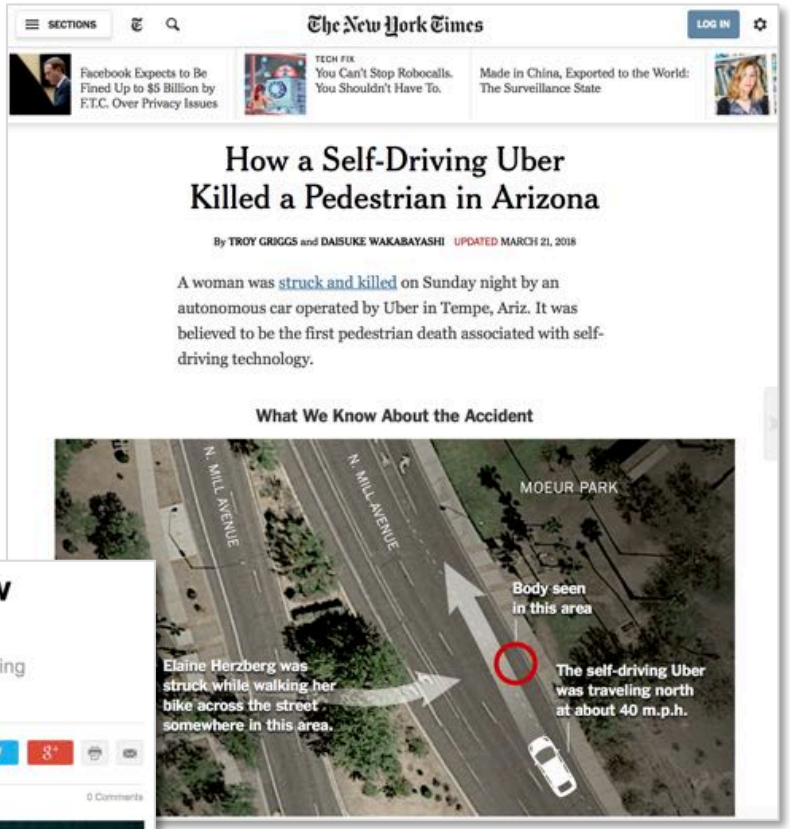
How can alarm fatigue be prevented?

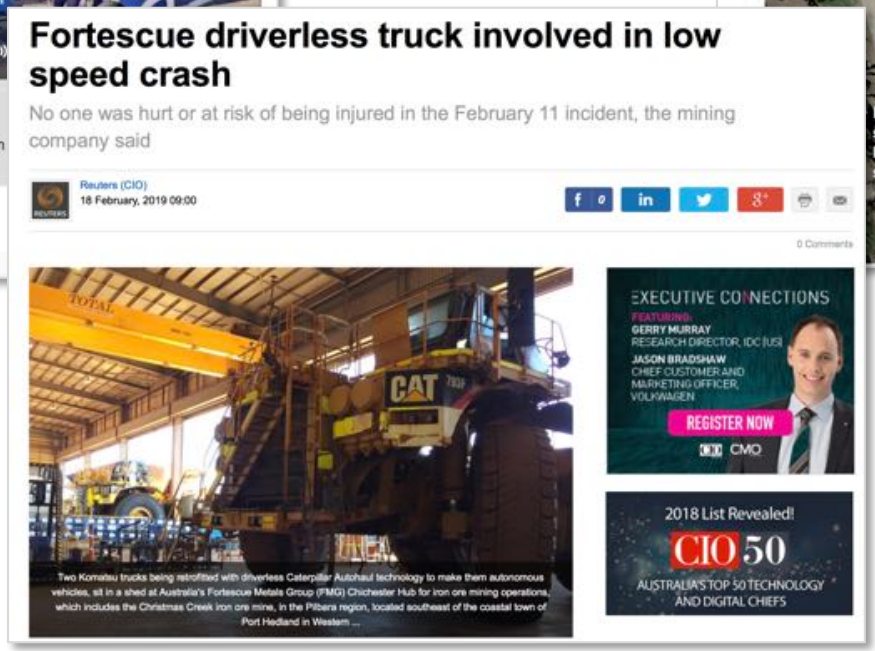
CHG HOSPITAL BEDS
www.chgbeds.com

Sources
The Boston Globe: http://www.boston.com/lifestyle/health/articles/2011/02/13/patient_alarms_often_unheard_unheeded
ECRI: https://www.ecri.org/Forms/Pages/ECRI-Institutes-2012-Top-10-Health-Technology-Hazards.aspx
Macleans: http://www2.macleans.ca/2011/10/12/on-noisy-hospitals-'alarm-fatigue'-and-how-all-those-bells-interfere-with-sleep-and-healing/

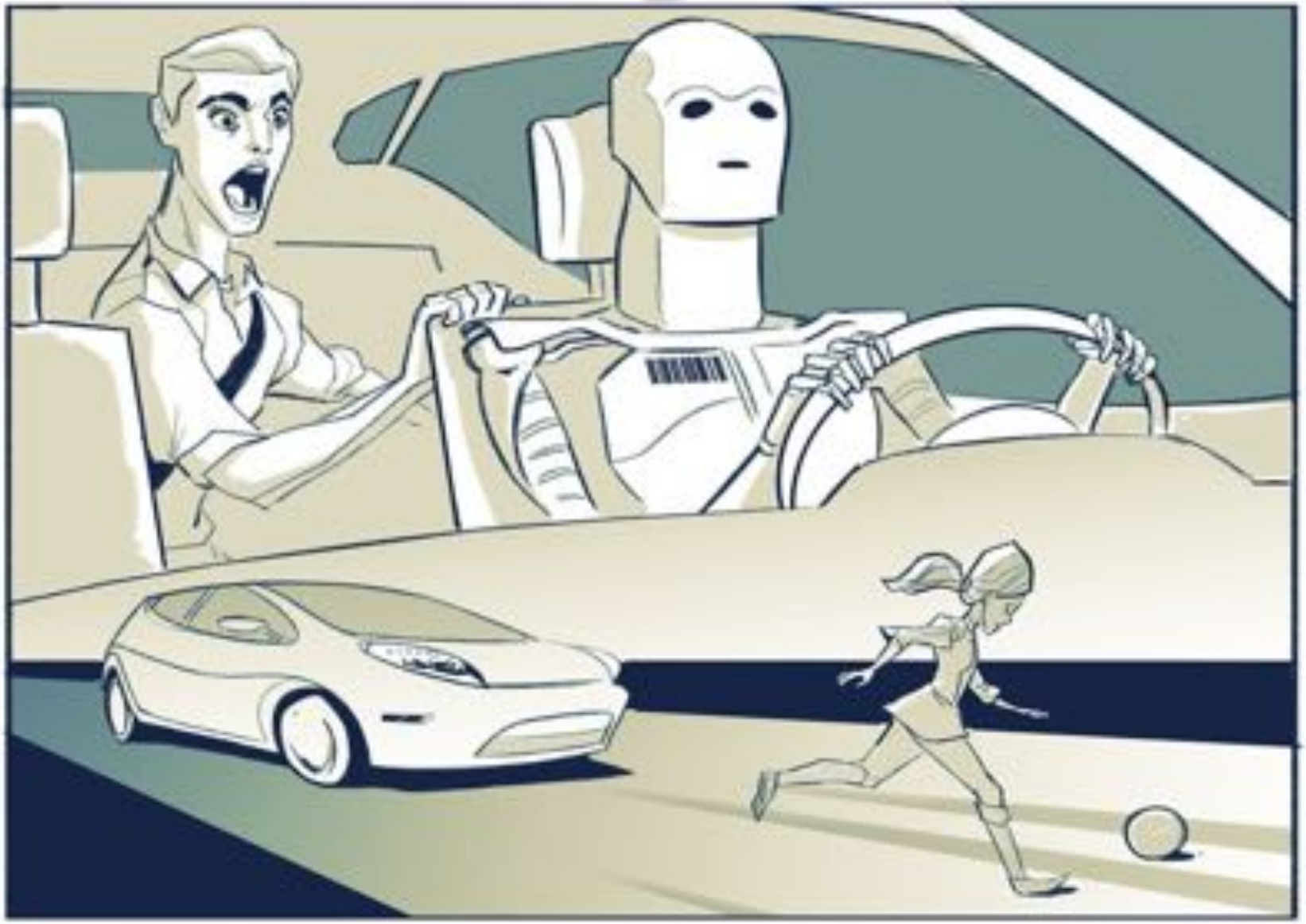
Perfection is the enemy of the good Voltaire





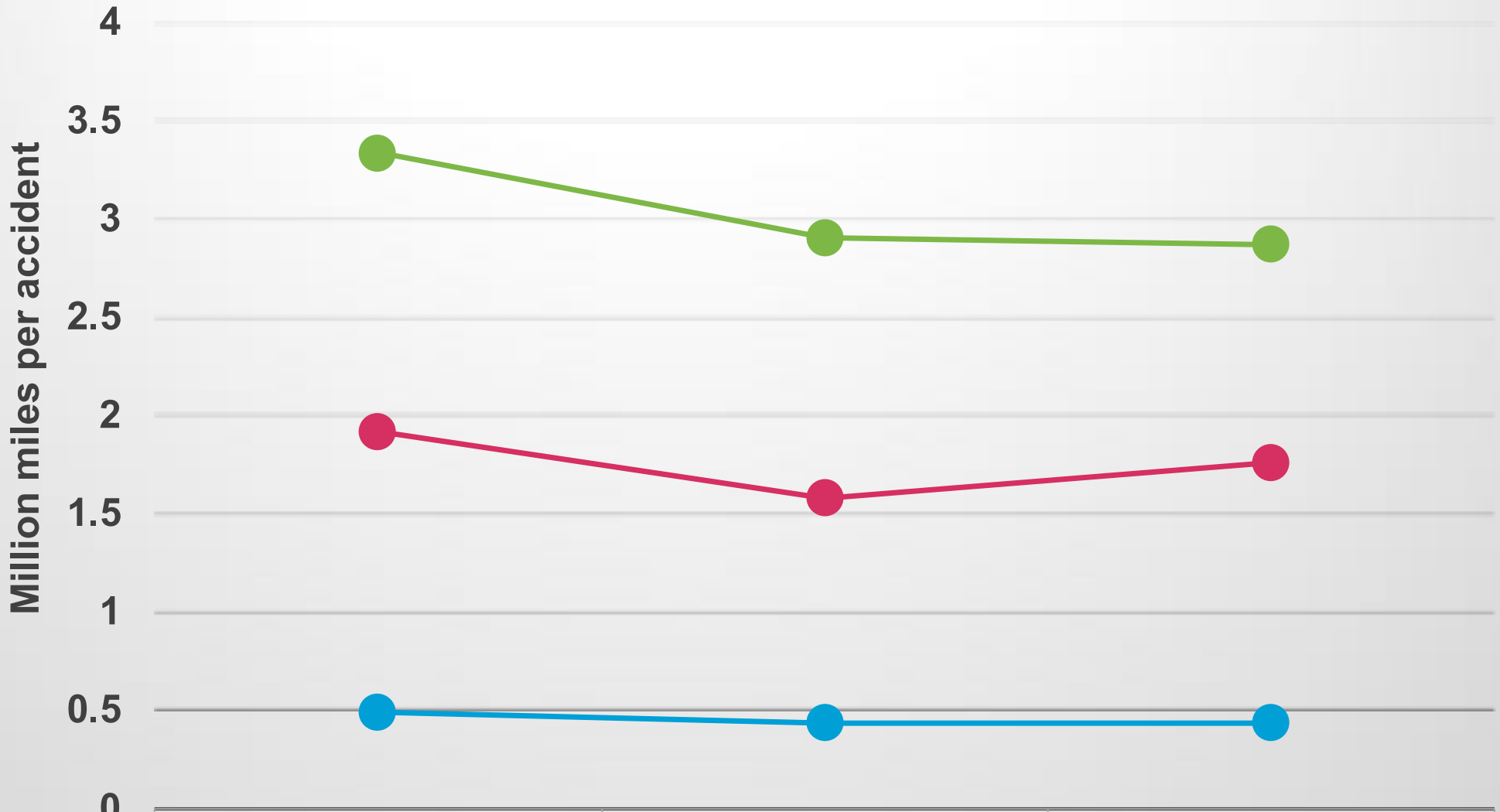


The Trolley Dilemma



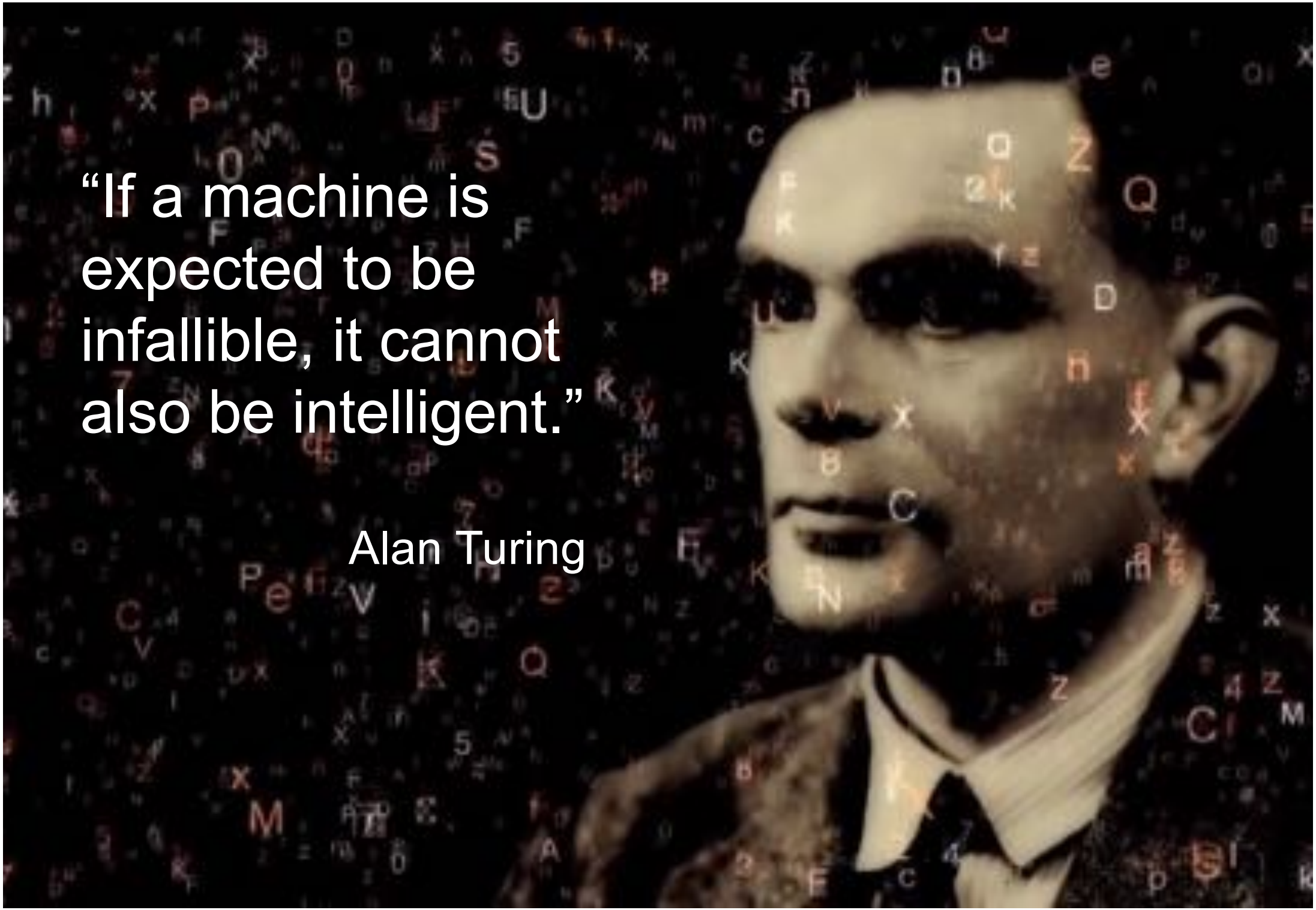
https://www.youtube.com/watch?v=nhCh1pBsS80

Tesla Accident Data

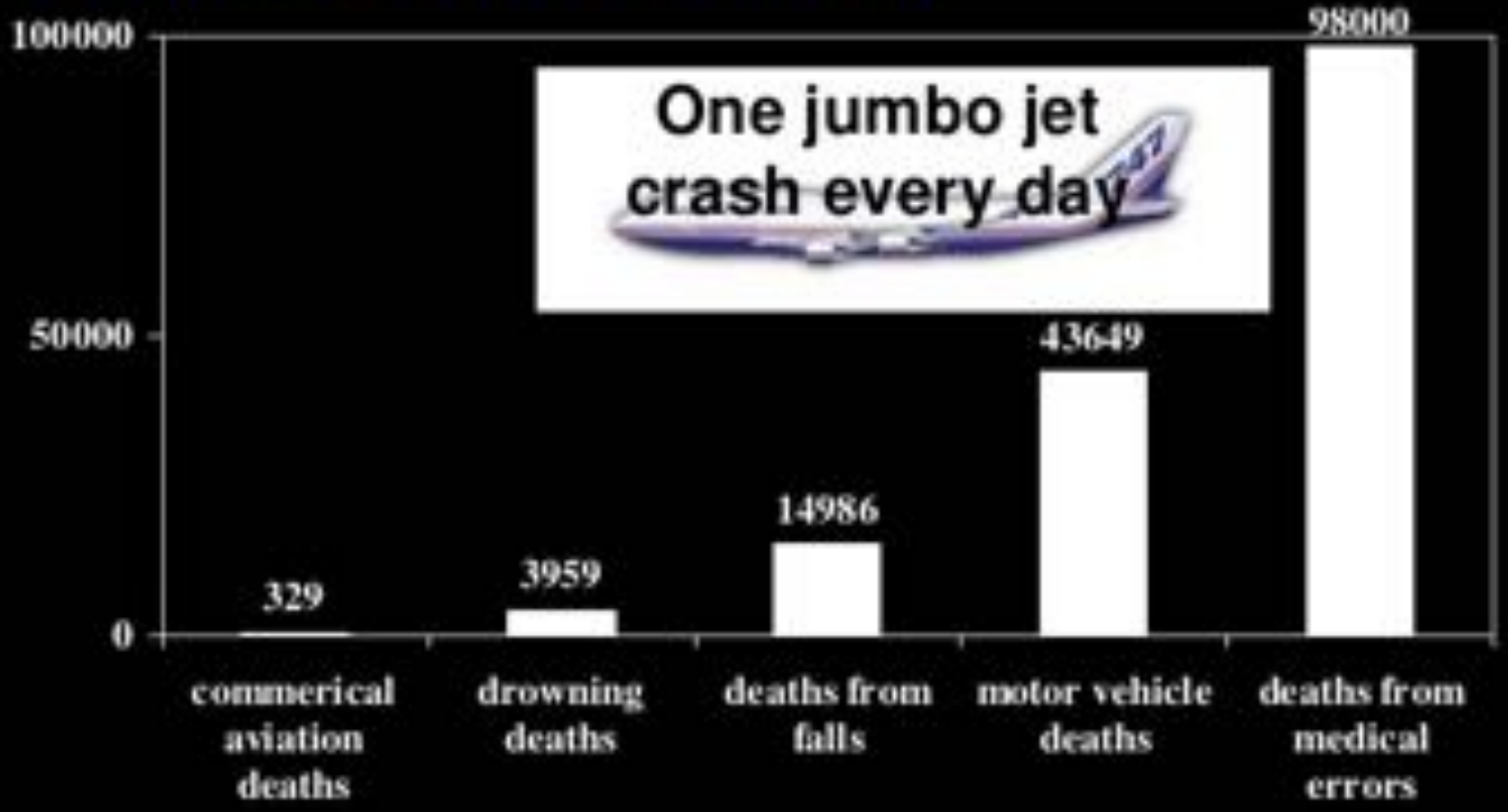


	Q3 - 2018	Q4 - 1018	Q1 - 2019
NHTSA	0.492	0.436	0.436
Tesla Autopilot Off	1.92	1.58	1.76
Telsa - Autopilot On	3.34	2.91	2.87

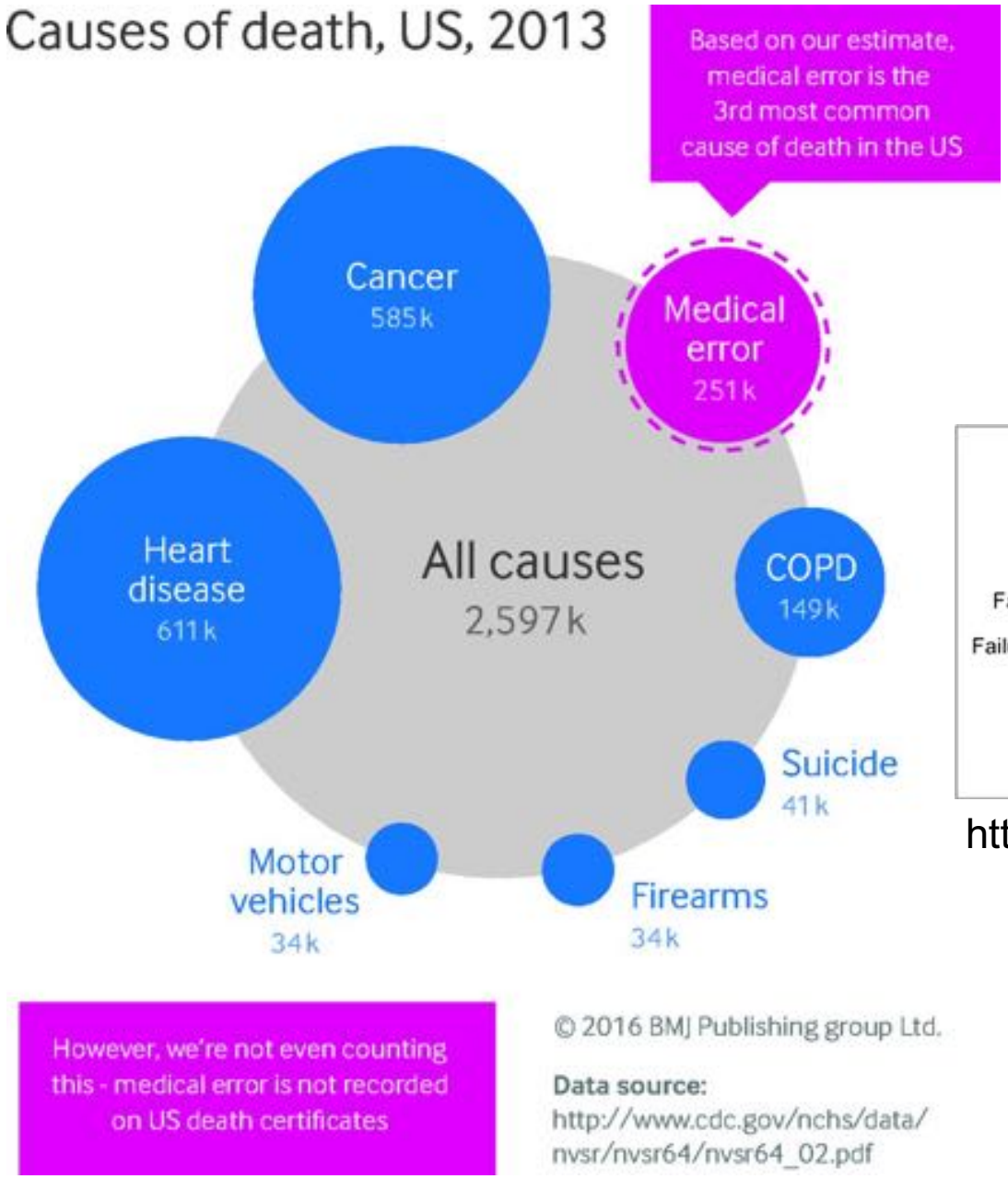
https://www.tesla.com/en_AU/VehicleSafetyReport

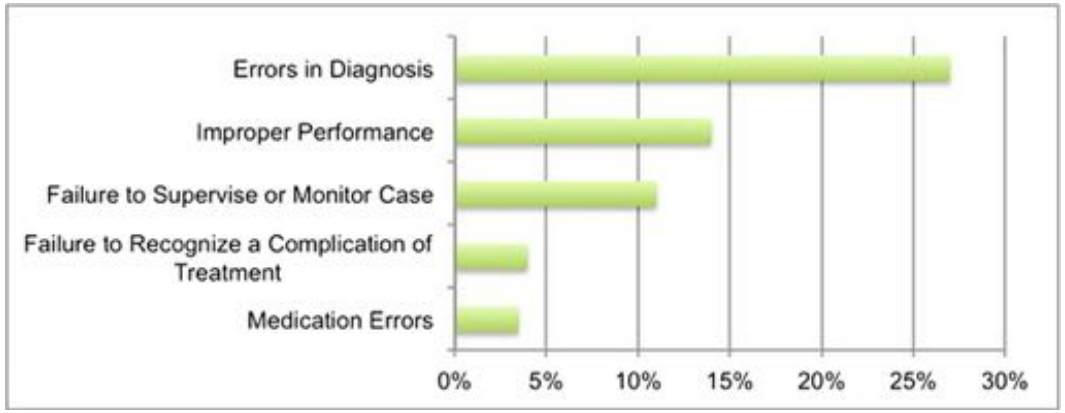

“If a machine is expected to be infallible, it cannot also be intelligent.”
Alan Turing

Annual Death Rates in US



Anna L Cox | University College London | anna.cox@ucl.ac.uk





https://qualitysafety.bmj.com/content/22/Suppl_2/ii21

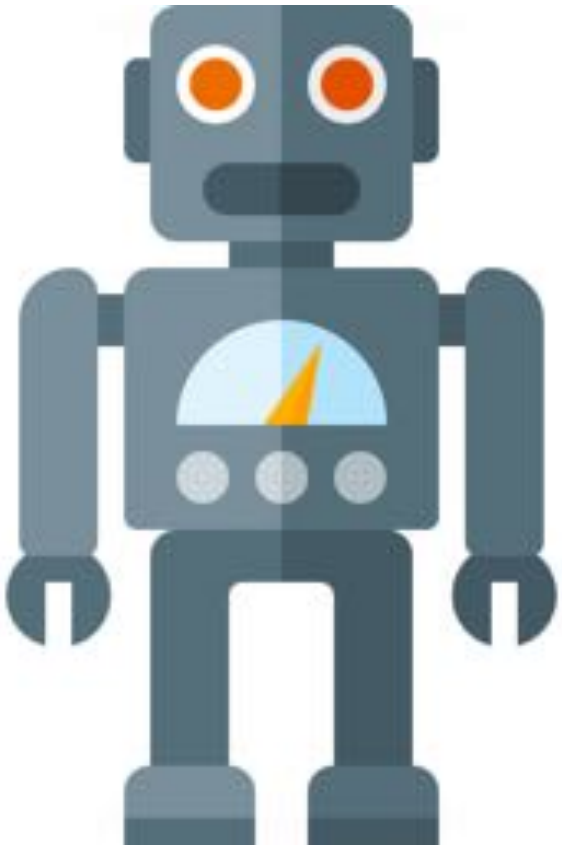
Human Versus Computer

Empathy

Common Sense

Inference

Versatility



Accurate

Consistent

Attentive

Replicable

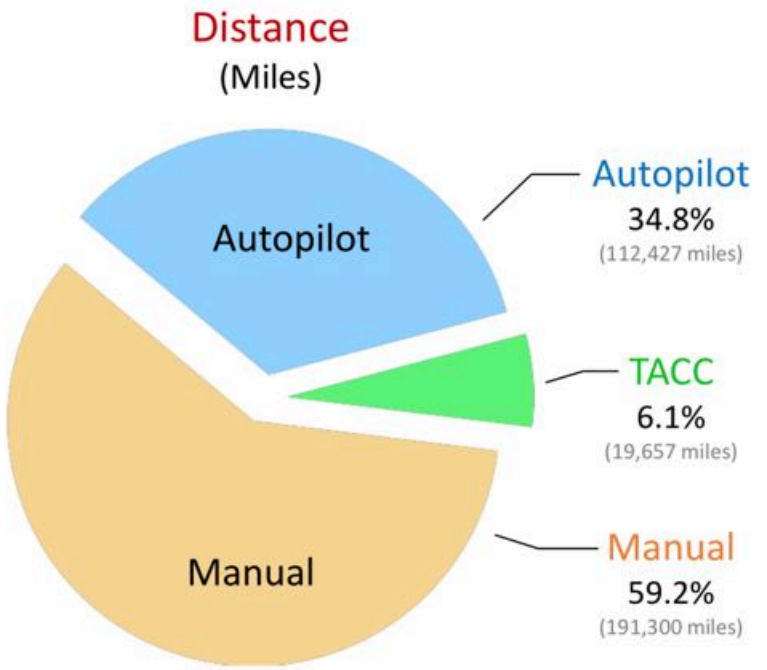
Human Augmentation

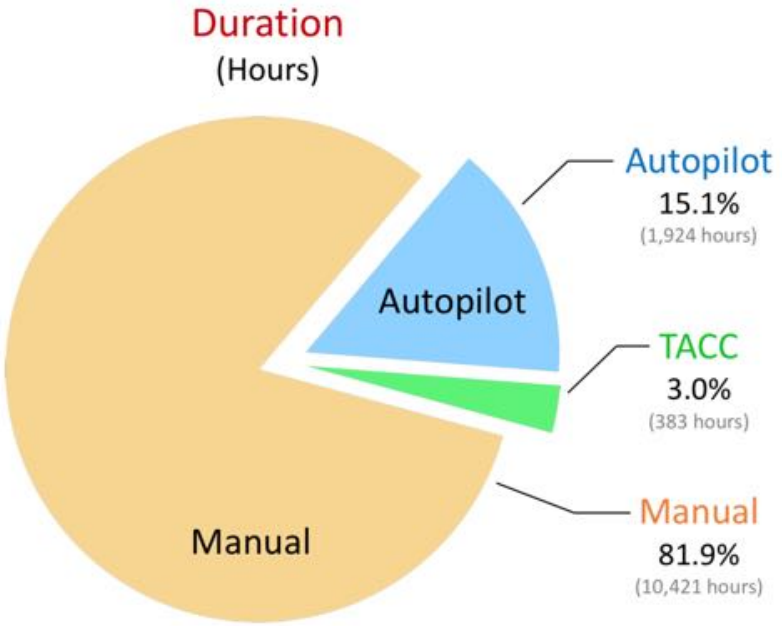
Error Rate in detection of cancer in lymph node cells	Human Pathologist	AI	Human Pathologist + AI
	3.5%	**7.5%**	**0.5%**

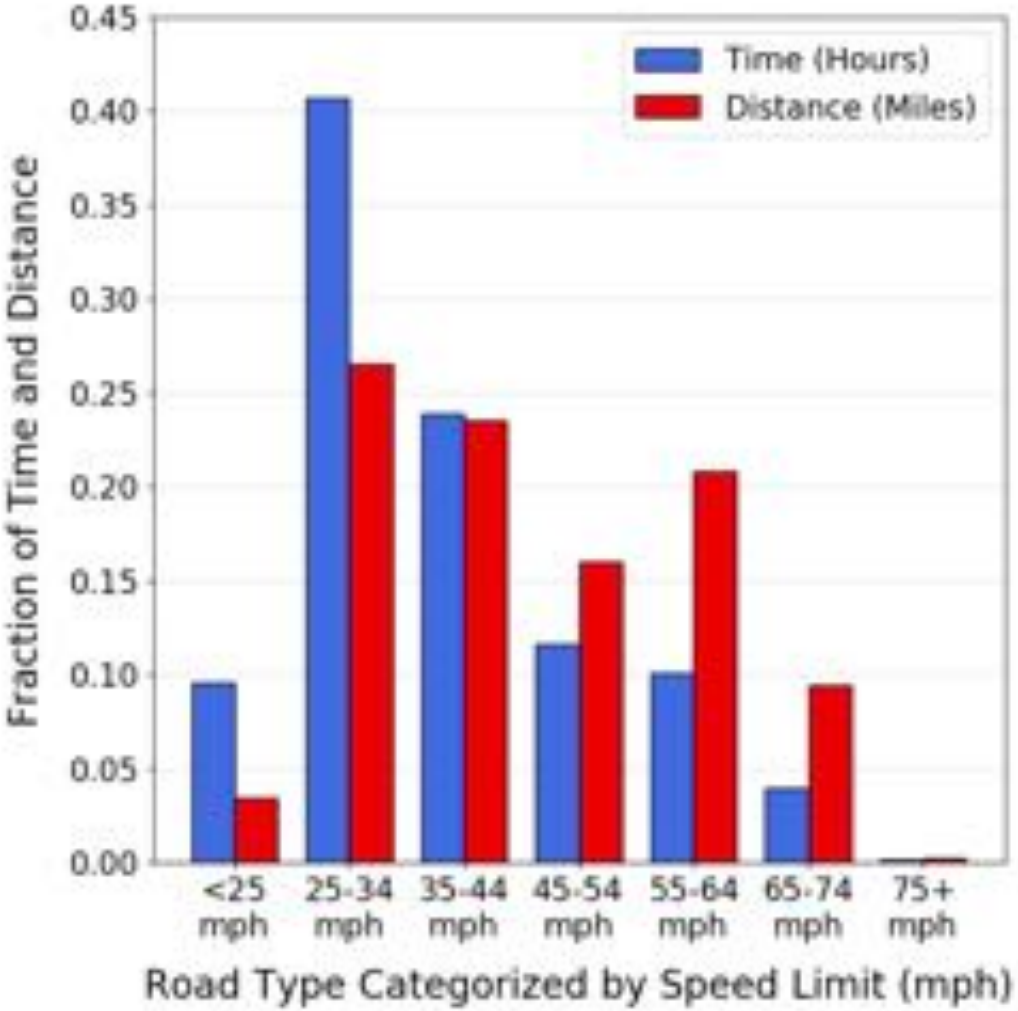
Source: Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, Andrew H. Beck, “Deep Learning for Identifying Metastatic Breast Cancer,” June 18, 2016, https://arxiv.org/pdf/1606.05718v1.pdf.

Humans & Autonomy

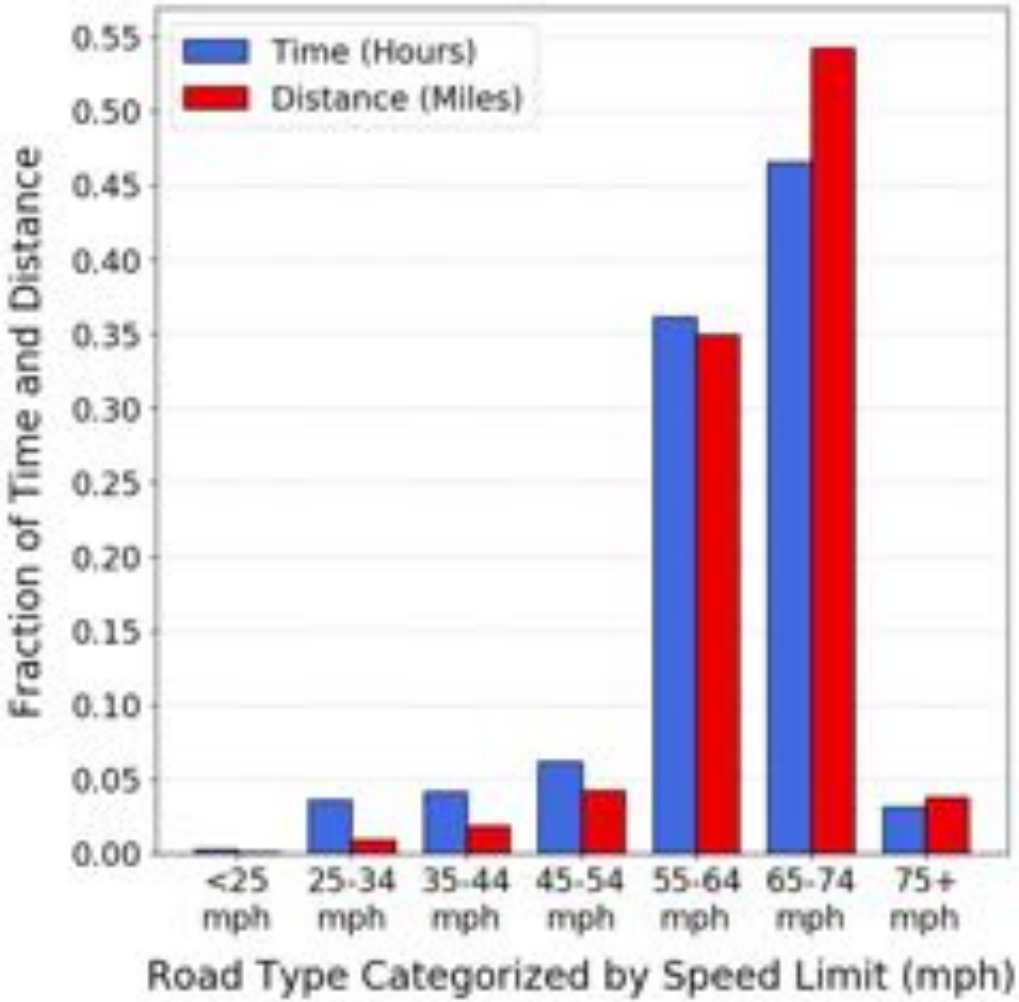
https://hcai.mit.edu/human-side-of-tesla-autopilot/







(a) Vehicle under manual control.



(b) Vehicle under Autopilot control.

Fig. 4: Time spent and distance traveled on different road types categorized by speed limit.

https://hcai.mit.edu/tesla-autopilot-human-side.pdf

Humans & Autonomy - Safe Handover

Critical Event Category	Disengagement Reason	Description	Human Initiated	Machine Initiated
Tricky Situation Present	Act too late after tricky situation	Delayed response to tricky situation (see details in §III-G).	0	0
	Act right after tricky situation	Rapid timely response after a tricky situation arises.	813	47
	Act before tricky situation	Anticipatory action before a tricky situation.	7,869	0
No Tricky Situation Present	Planned Turning or Speed Change	Taking control to make a planned navigation decision.	8,608	68
	Planned Stopping	Stopping for stop sign, yellow/red traffic light.	601	0
	Accidental	Accidentally bumping the wheel or the Autopilot stalk.	38	0
	Annotation Difficult	Image is too bright/dark for accurate annotation.	94	0
	No clear reason	No clearly identifiable reason	777	0
	Hands off wheel	Warning ignored while remaining attentive to the road.	0	13
		Total Annotated Disengagement Epochs:	**18,800**	**128**

Summary of Results:

Type of Driver Response	Percentage of Disengagements
Delayed (Slow responses or missed detections)	0.0%*
Responsive (Rapid timely responses)	4.5%
Anticipatory (Action before T.S. or planned decision)	90.6%
Other (Accidental, not annotatable, etc.)	4.9%

*This value is entered as "0.0%" to reflect the fact that no such events were found in our dataset using the methods described. However, it is possible that some events of this type exist in the dataset but went undiscovered. Future work may lead to new methods that will help identify these, if any exist.

TABLE I: Annotated reasons for disengagement of Autopilot. The annotation process and question details are described in §III-F. Reasons are divided into two categories: those associated with tricky situations and those that are not. The label of "act too late after tricky situation" was designed to locate disengagement epochs associated with high functional vigilance decrement. Of the 18,928 total annotated disengagements, no epochs were labeled in this way by the annotators. The results are summarized in the table on the right with respect to functional vigilance and anticipatory characteristics of the quantitative results.

https://hcai.mit.edu/tesla-autopilot-human-side.pdf

Human-Centered Approach

90% ← No — Human Needed — Yes → 10%

Solve the perception-control problem where **possible**:

And where **not possible**: involve the human

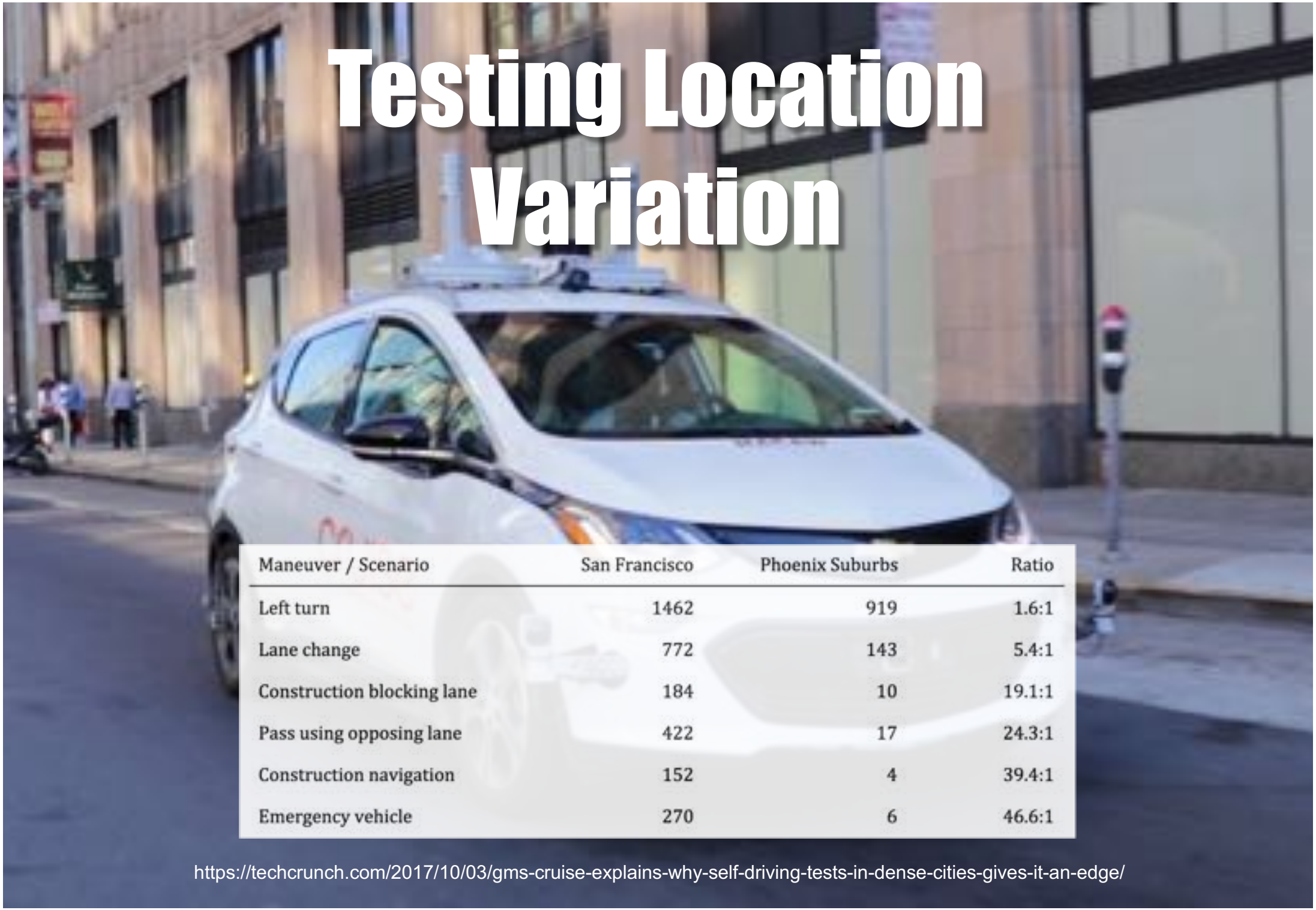
https://youtu.be/sRxaMDDMWQQ


Super
Model
Sample
Bias

Overfitted Data

Training optimised with sample biased data

Testing Location Variation



Maneuver / Scenario	San Francisco	Phoenix Suburbs	Ratio
Left turn	1462	919	1.6:1
Lane change	772	143	5.4:1
Construction blocking lane	184	10	19.1:1
Pass using opposing lane	422	17	24.3:1
Construction navigation	152	4	39.4:1
Emergency vehicle	270	6	46.6:1

https://techcrunch.com/2017/10/03/gms-cruise-explains-why-self-driving-tests-in-dense-cities-gives-it-an-edge/

Simulators

http://metro.co.uk/video/google-s-ai-just-taught-walk-run-itself-1502819/?ito=vjs-link

Simulators

Beyond Grand Theft Auto V for Training, Testing and Enhancing Deep Learning in Self Driving Cars

https://arxiv.org/pdf/1712.01397.pdf
https://arxiv.org/pdf/1610.01983.pdf

Udacity Self Driving Simulator

https://towardsdatascience.com/introduction-to-udacity-self-driving-car-simulator-4d78198d301d

AWS DeepRacer

Distributed DeepRacer RL training with SageMaker and RoboMaker

Introduction

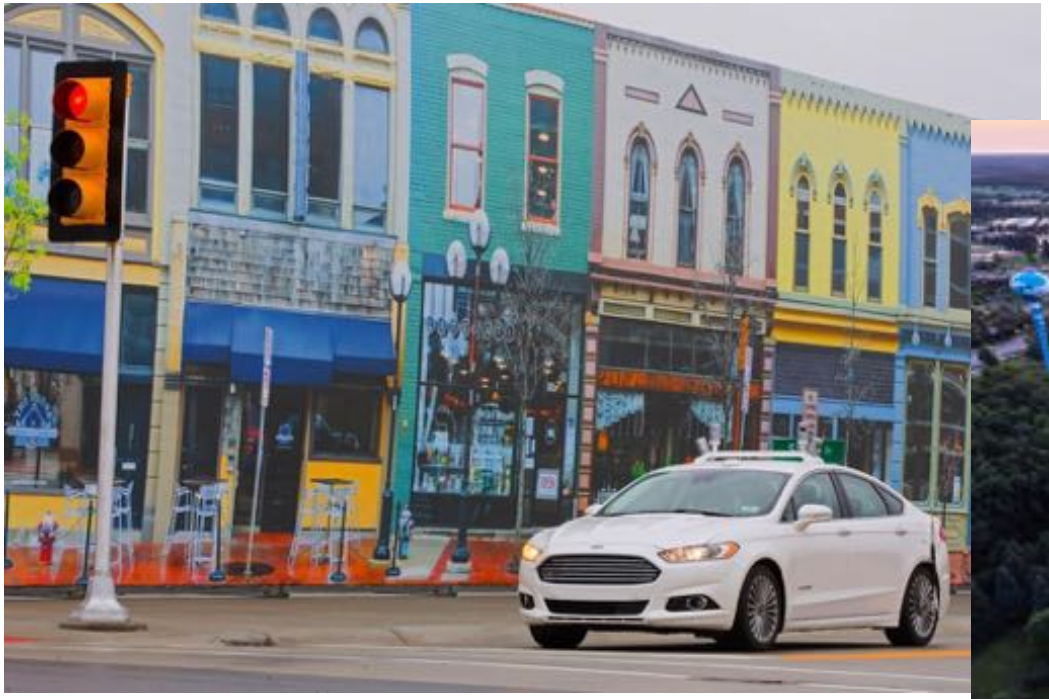
In this notebook, we will train a fully autonomous 1/18th scale race car using reinforcement learning using Amazon SageMaker RL and AWS RoboMaker's 3D driving simulator. AWS RoboMaker is a service that makes it easy for developers to develop, test, and deploy robotics applications.

This notebook provides a jailbreak experience of AWS DeepRacer, giving us more control over the training/simulation process and RL algorithm tuning.

aws DEEPRACER LEAGUE

aws re:Invent

MCity– University of Michigan Driverless Car Test Facility



https://www.youtube.com/watch?v=oktuH5qGXvk

Corner Cases

https://www.youtube.com/watch?v=Ucp0TTmvqOE&t=1h51m05s

Tesla Data Engine

https://www.youtube.com/watch?v=Ucp0TTmvqOE&t=1h51m05s

Model Uncertainty

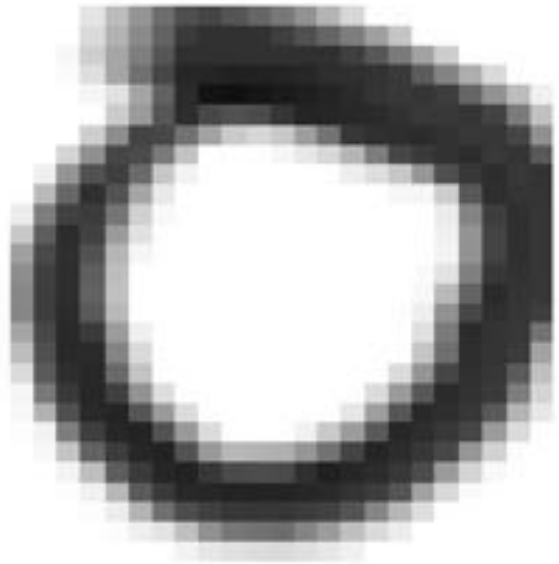
Data & Labels

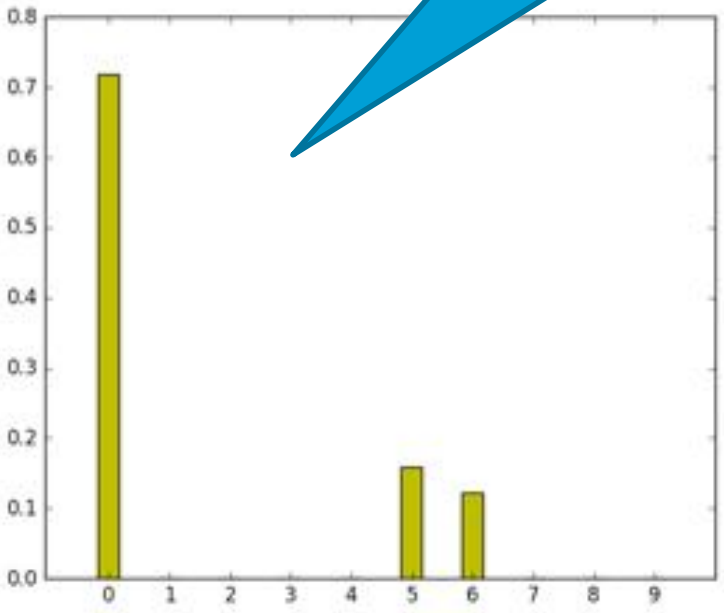
0
1
2
3
4
5
6
7
8
9

http://yann.lecun.com/exdb/mnist/

Model Uncertainty

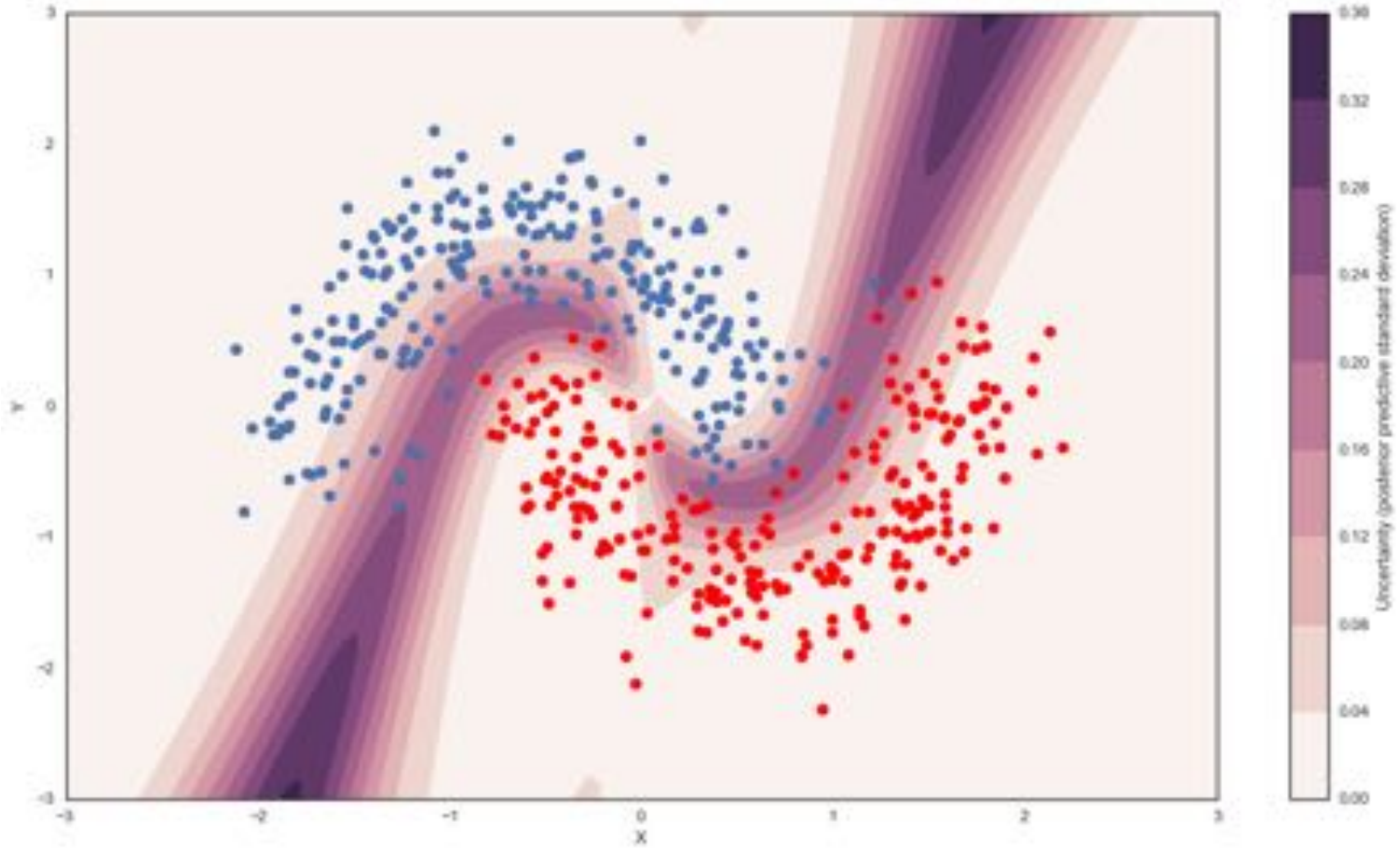
How confident is the model in making these predictions





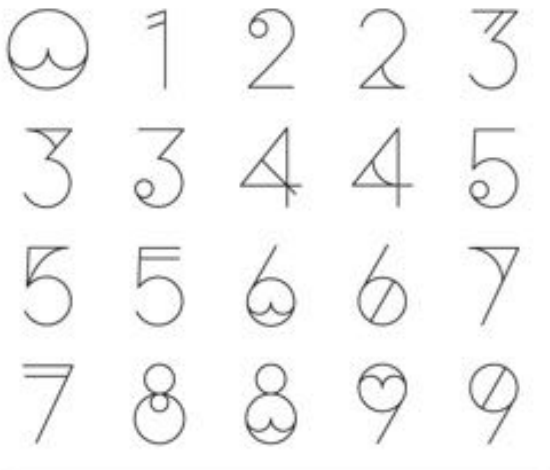
http://yann.lecun.com/exdb/mnist/

Model Uncertainty

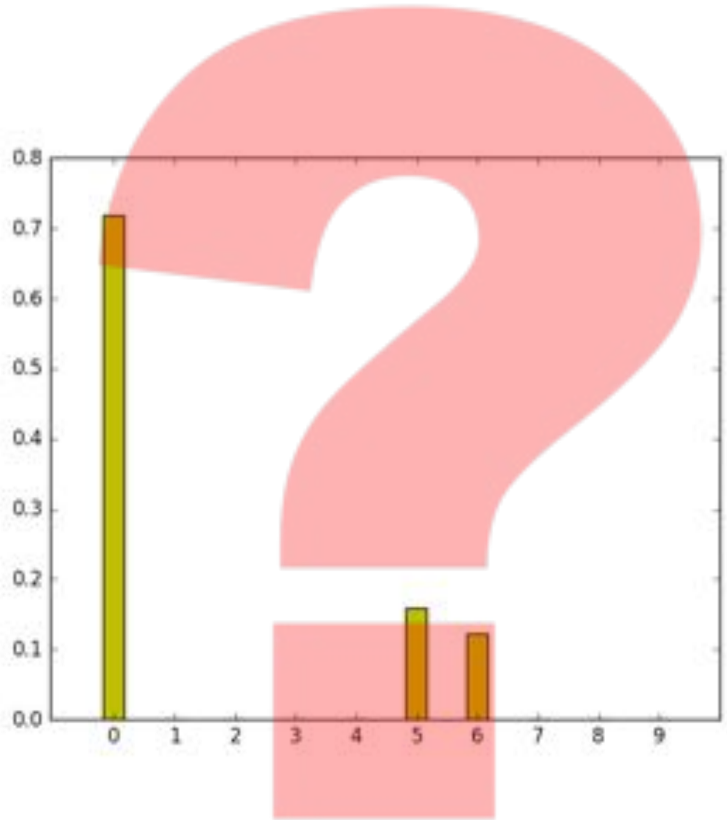


Evaluating Uncertainty Quantification in End-to-End Autonomous Driving Control
https://arxiv.org/pdf/1811.06817.pdf

Out Of Distribution Inputs

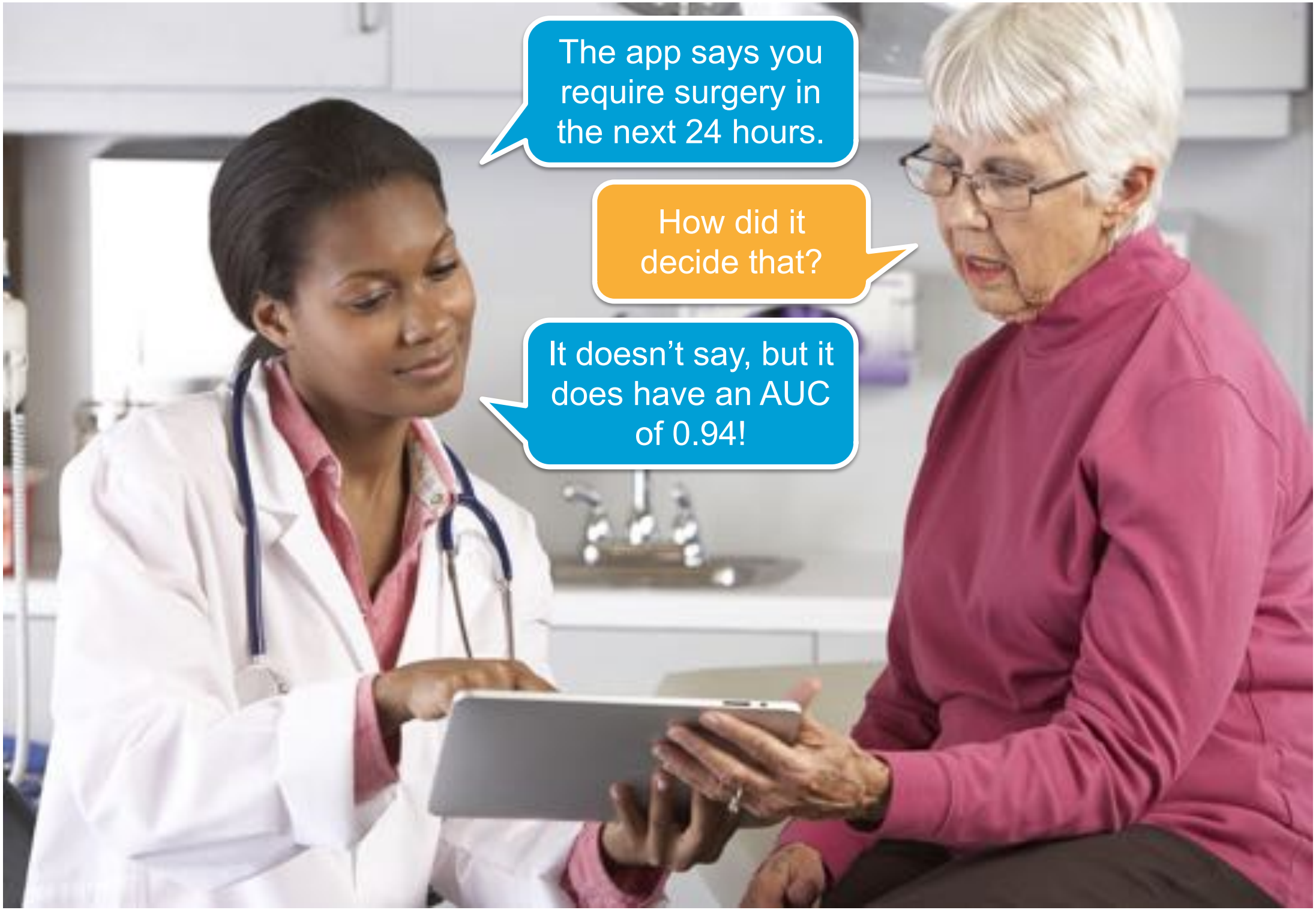


Out Of Distribution Inputs



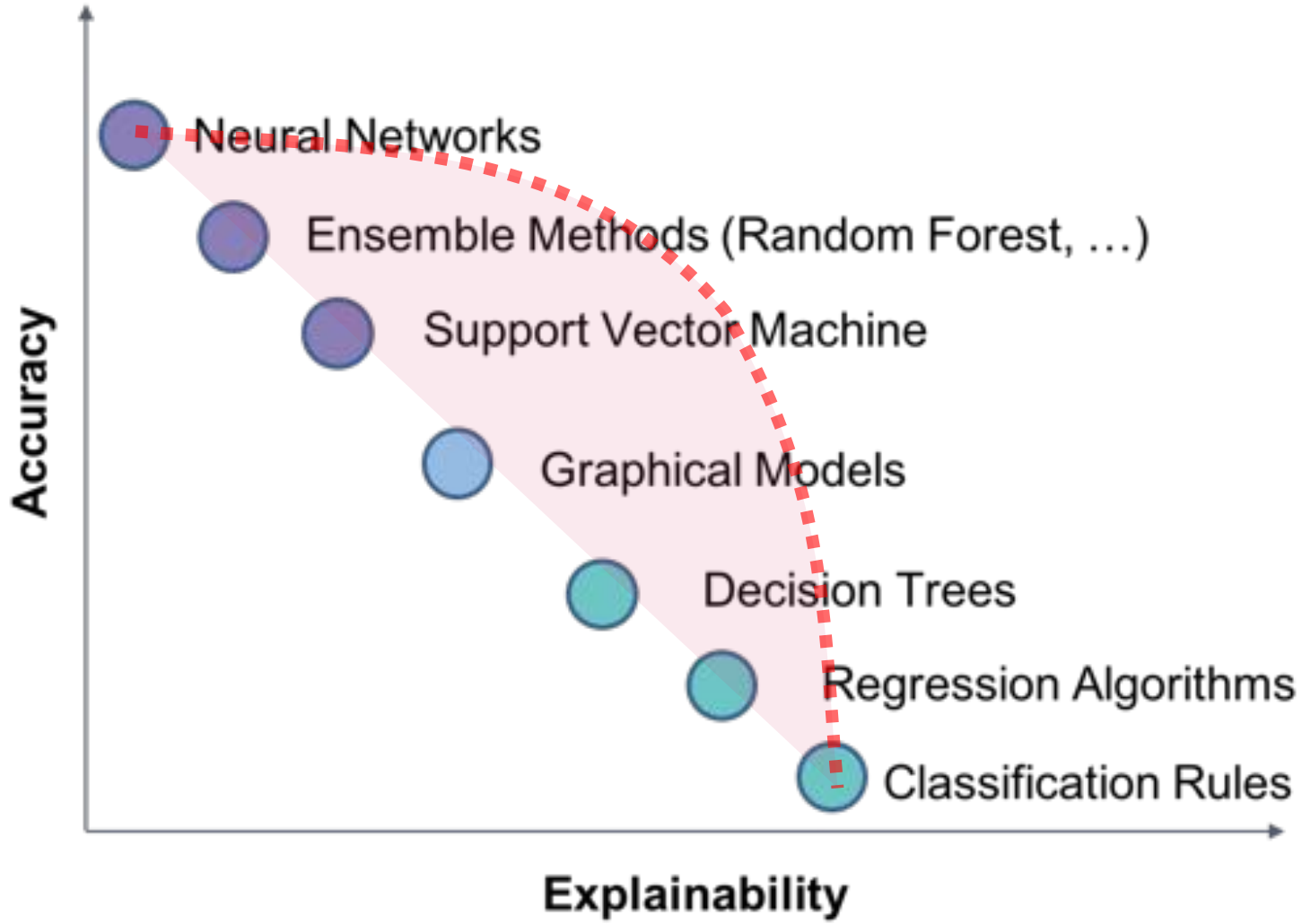
https://www.youtube.com/watch?v=t-9LoAxJeME&list=PLkFD6_40KJIxG6I7MWd4LXAKI-kQO54_8&index=9&t=0s

https://!!!!!!!!!!!!!!!!!/file/d/1PEnhC5YMvilFz60NQGQLyxTnTwZ0nTBp/view?usp=sharing


The app says you require surgery in the next 24 hours.
How did it decide that?
It doesn't say, but it does have an AUC of 0.94!

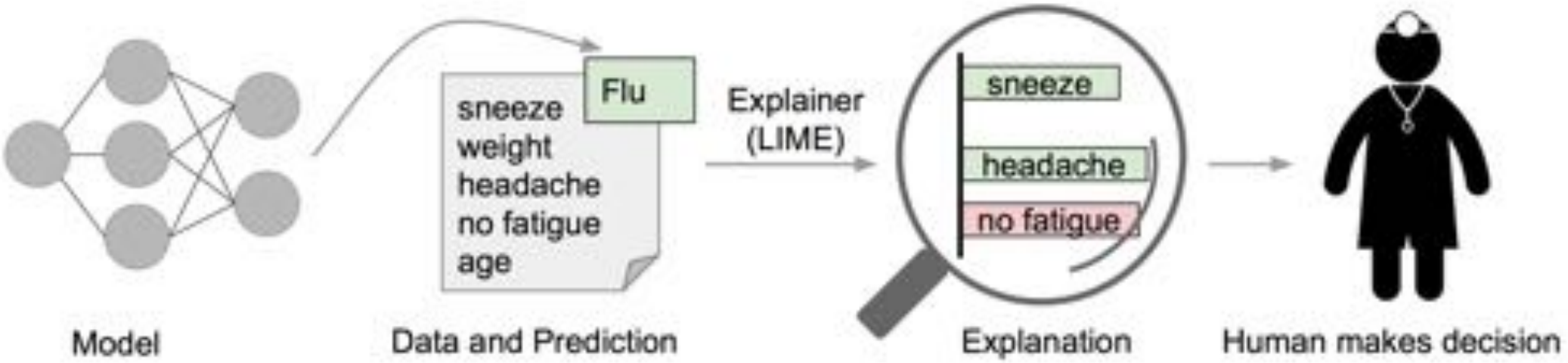

The app says you require surgery in the next 24 hours as your blood pressure is low, ECG rhythm is irregular, and protein test returned positive.

Accuracy Explainability Tradeoff



https://towardsdatascience.com/explainable-ai-the-data-scientists-new-challenge-f7cac935a5b4

Explainability



"Why Should I Trust You?: Explaining the Predictions of Any Classifier", Ribeiro et al, 2016, https://arxiv.org/pdf/1602.04938

Clinical Decision Support

Characterizing risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study

Shane Nanayakkara[1, 2, 3], Sam Fogarty[4,5], Michael Tremeer[4], Kelvin Ross[4, 5], Brent Richards[5, 6], Christoph Bergmeir[7], Sheng Xu[7], Dion Stub[1,2,3], Karen Smith[8, 9], Mark Tacey[10], Danny Liew[10], David Pilcher[10, 11, 12], David M Kaye[1,2,3]

1 Department of Cardiology, Alfred Hospital, Melbourne, Australia
2 Heart Failure Research Group, Baker Heart and Diabetes Institute, Melbourne, Australia
3 Department of Medicine, Nursing and Health Sciences, Monash University, Melbourne, Australia
4 Institute for Integrated and Intelligent Systems, Griffith University, Gold Coast, Australia
5 IntelliHQ, Gold Coast, Australia
6 Gold Coast University Hospital, Gold Coast, Australia
7 Faculty of Information Technology, Monash University, Melbourne, Australia
8 Centre for Research and Evaluation, Ambulance Victoria, Melbourne, Australia
9 Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, Australia
10 School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia.
11 Department of Intensive Care, Alfred Hospital, Melbourne, Australia
12 The Australian and New Zealand Intensive Care Society (ANZICS) Centre for Outcome and Resource Evaluation (CORE), Melbourne, Australia

- Cardiac arrest is a frequent cause of admission to the intensive care unit and has a low survival rate following admission to hospital.
- Current illness severity scores perform poorly in regard to predicting survival for this specific group of patients.
- Estimate risk of adverse outcomes from first 24 hours admission data
- Analysis derived from 39,566 cardiac arrest patients from 186 ICUs
- Use explainability tools to understand reasoning of algorithm's decision for individual patients

Random forests classifier

Correctly predicted survived case

Model prediction:	0.116366362715	LIME prediction:	0.1616781

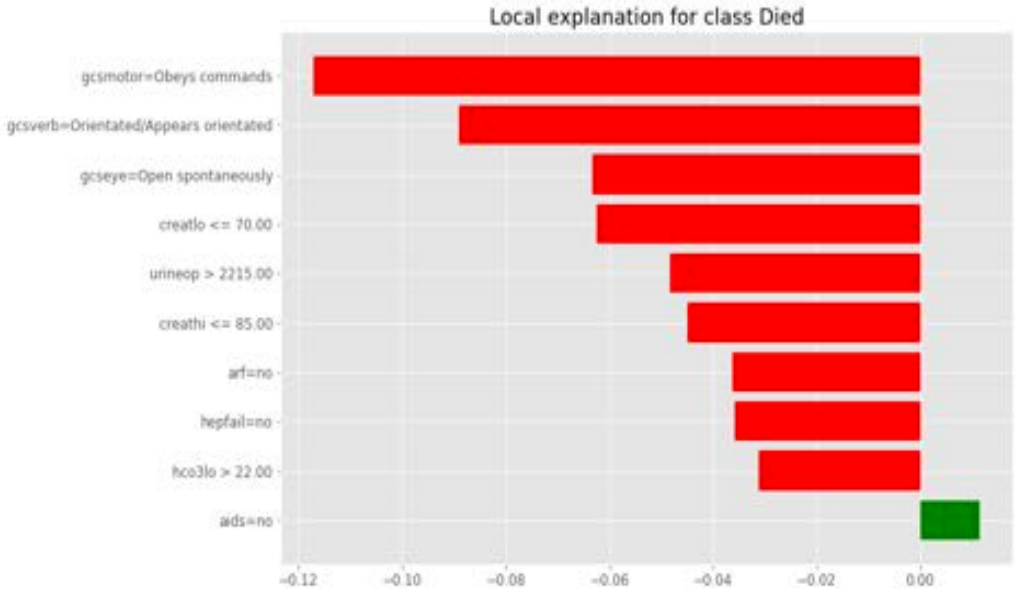
Good case:

The model prediction is quite distinct between died and survived probabilities. Therefore, LIME is also able to provide an approximation that's on the same side of the prediction.

Comment:

The explanations are making sense most part, however, for this particular patient, having no aids contributes to his mortality, and this is contrary to (my) common sense.

LIME explanation:



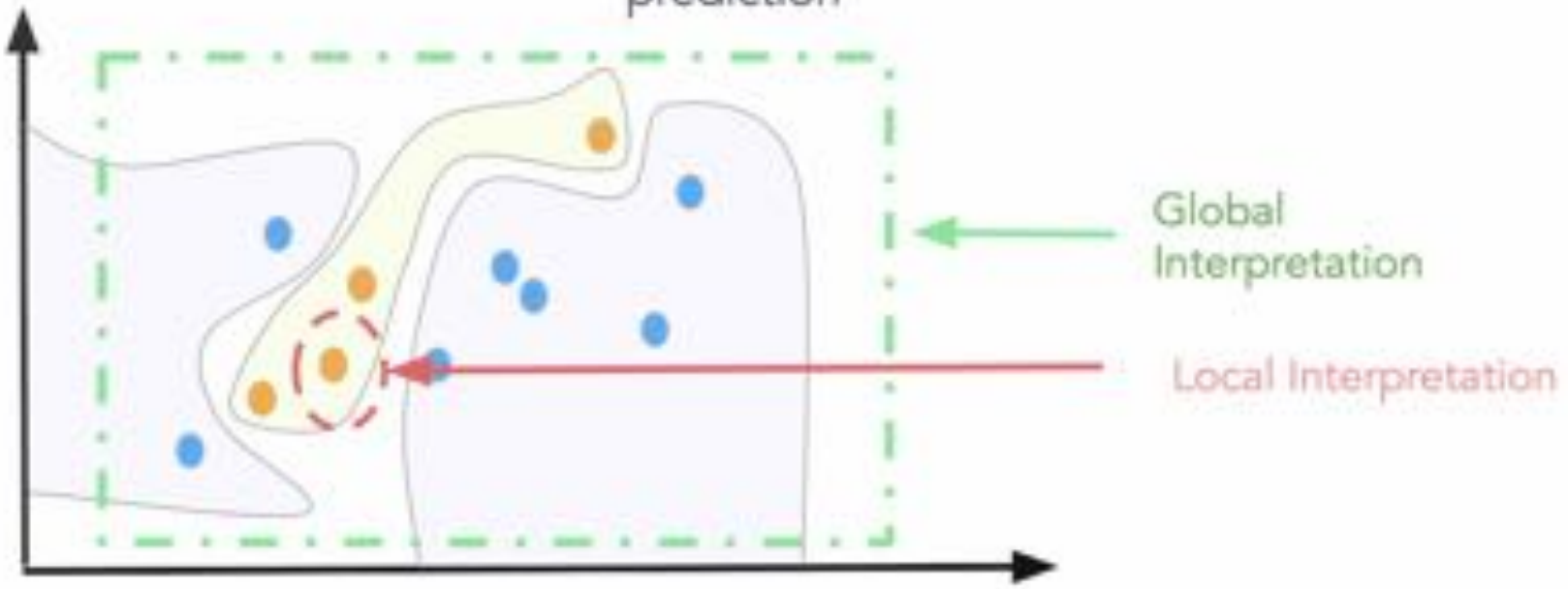
Global V Local

Global Interpretation

Being able to explain the conditional interaction between dependent(*response*) variables and independent(*predictor, or explanatory*) variables based on the complete dataset

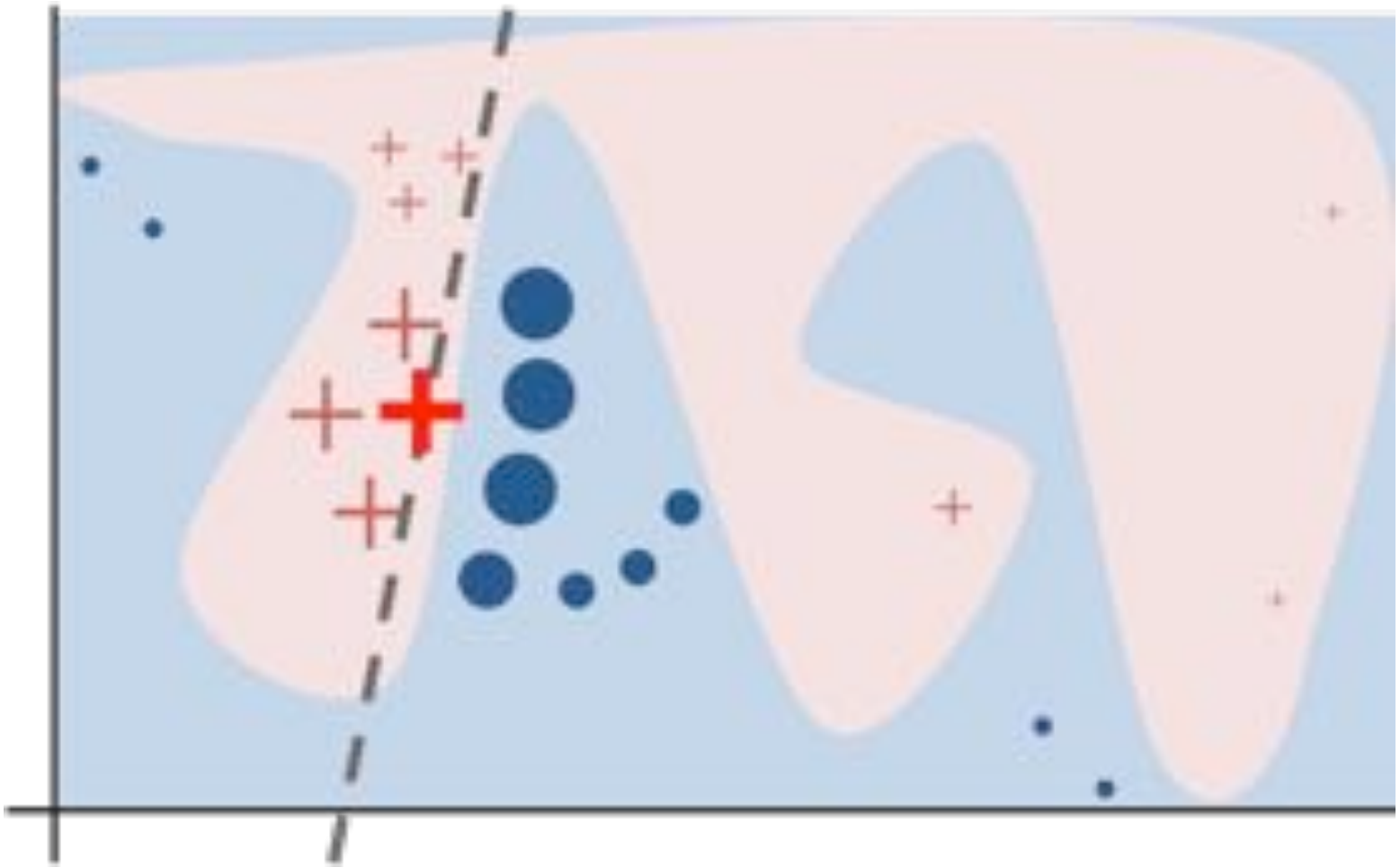
Local Interpretation

Being able to explain the conditional interaction between dependent(*response*) variables and independent(*predictor, or explanatory*) variables wrt to a single prediction



https://towardsdatascience.com/human-interpretable-machine-learning-part-1-the-need-and-importance-of-model-interpretation-2ed758f5f476

Local Explainability

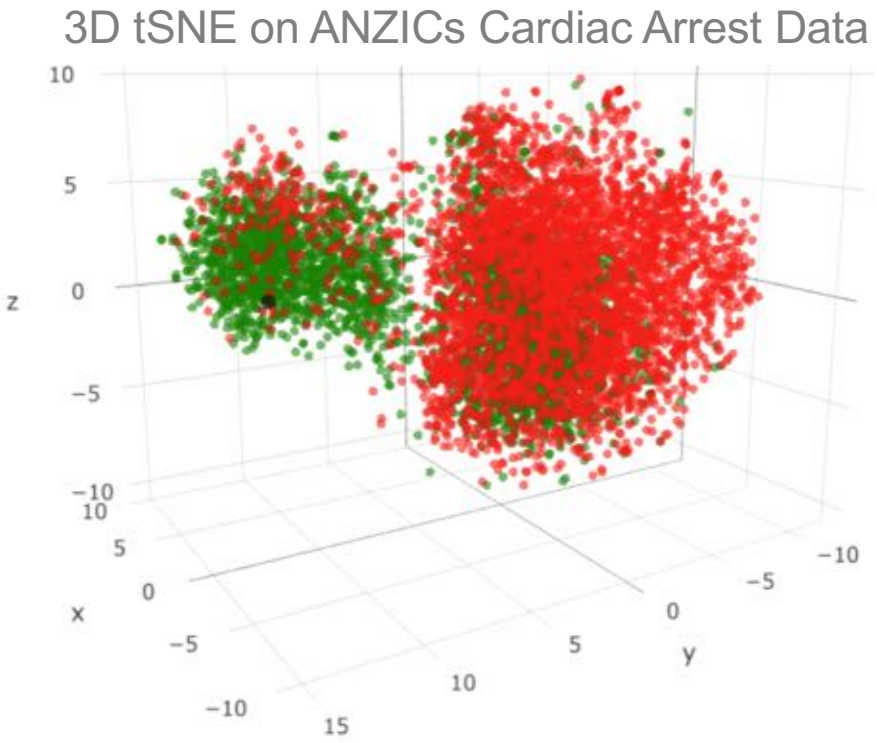


Similarity and Clustering

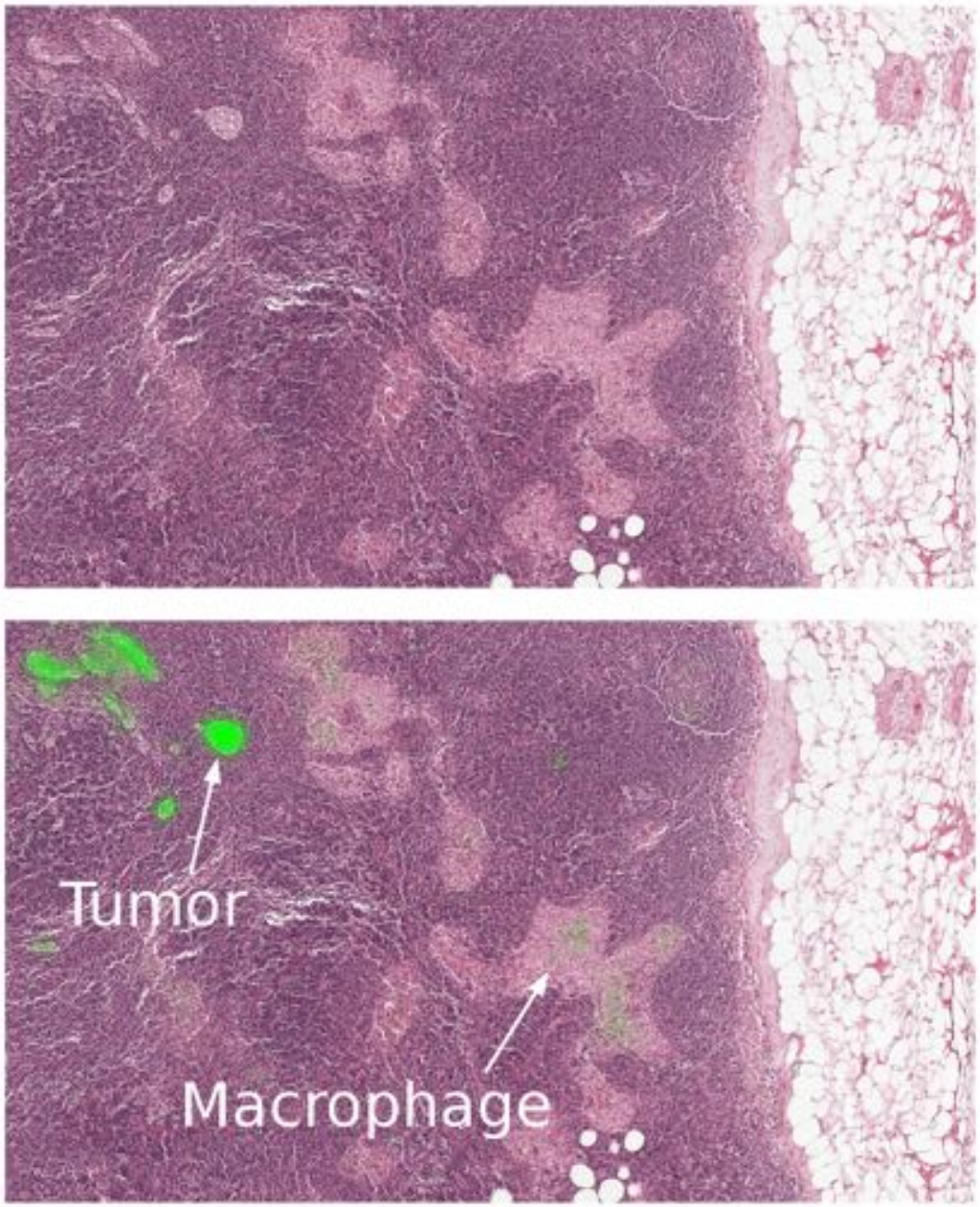
Dimensionally reduced data can be clustered - keeping similar patients near each other. Useful for localised data analysis and active learning.

```
Model prediction = 0.03
Neighbour count chosen = 40
Local avg prediction = 0.07
Local avg outcome = 0.05
Local success = 0.95
Feature columns are sorted by the distance the local avg is from the
```

	gcsverb	gcseye	gcsmotor	hco3lo	templo	maplo	hco3hi	gluchi	dia
0	5.00	4.00	6.00	2.61	0.46	0.52	2.28	-1.02	
Neighbours mean (k=40)	4.97	3.88	6.00	1.09	0.98	0.80	0.77	-0.75	
Training set mean	1.68	1.59	2.09	0.00	0.00	-0.00	-0.00	-0.00	
Means dif in Stds	2.32	2.01	2.00	1.09	0.98	0.80	0.77	-0.75	



Explainability – Lymph Node Biopsy



“Assisting Pathologists in Detecting Cancer with Deep Learning”, Stumpe et al, Mar 2017, https://research.googleblog.com/2017/03/assisting-pathologists-in-detecting.html

Explainability – Driverless Cars

Steer Left or Right?

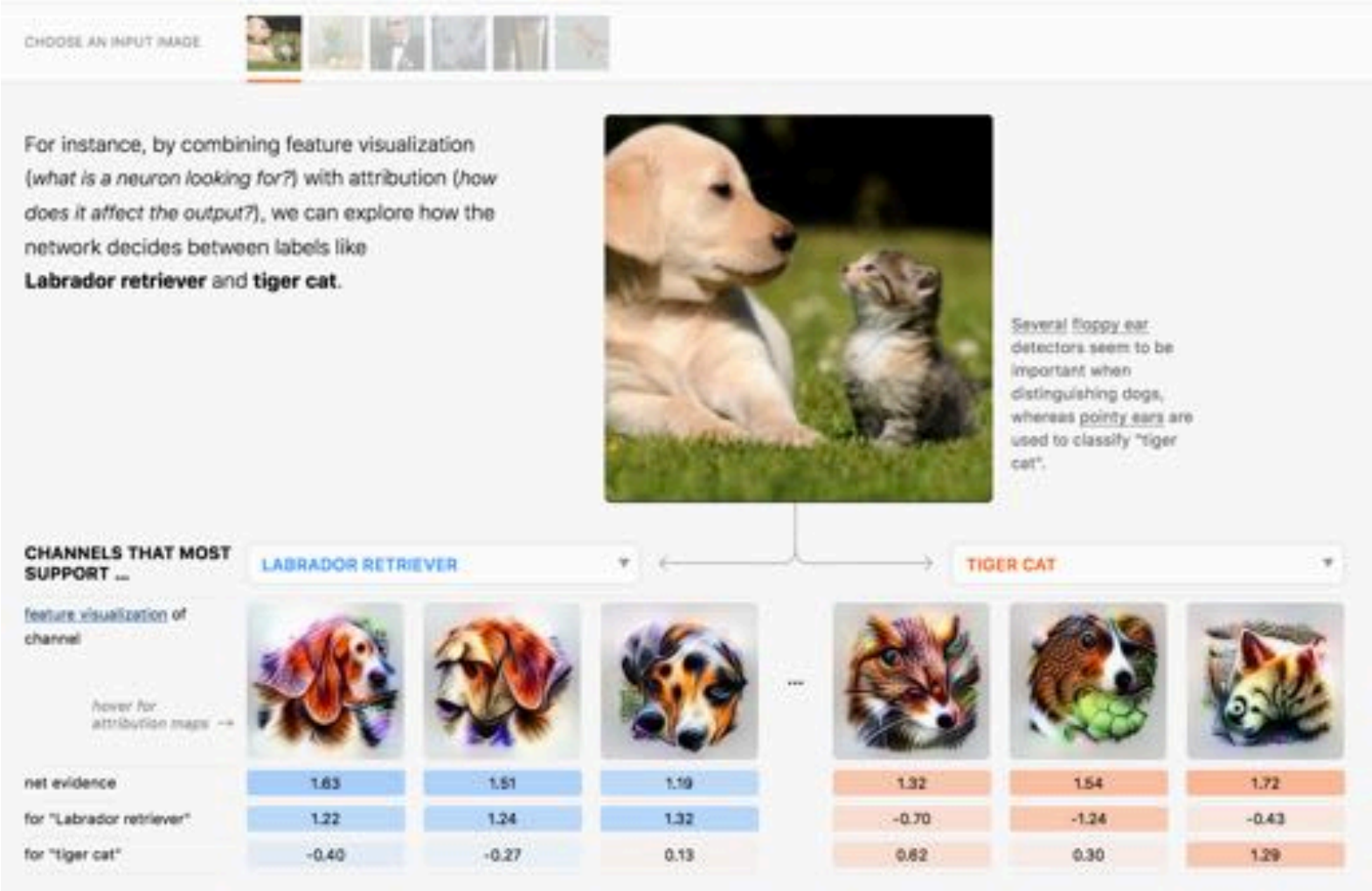
What features are used?

Figure 4: Examples of salient objects for various image inputs.

"Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car", Bojarski et al, 2017, https://arxiv.org/abs/1704.07911

The Building Blocks of Interpretability

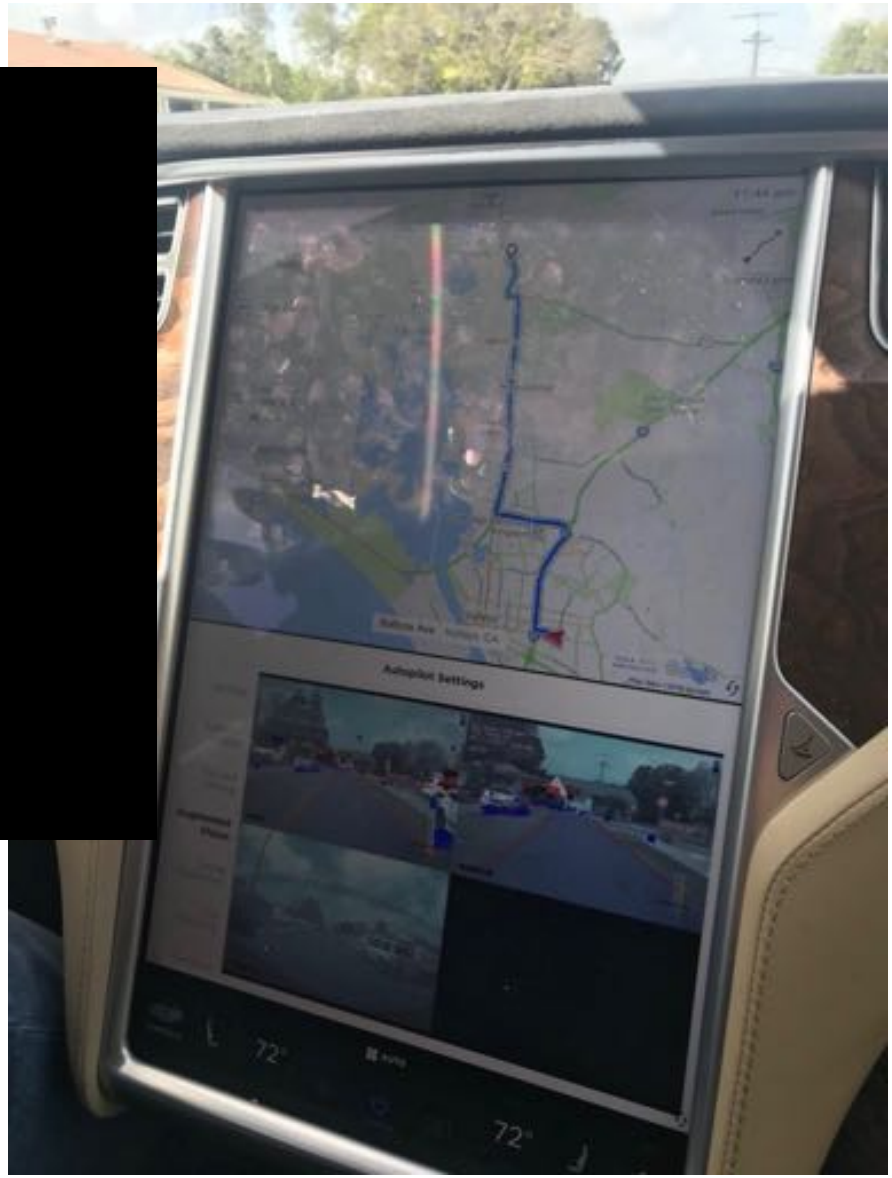
Interpretability techniques are normally studied in isolation.
We explore the powerful interfaces that arise when you combine them —
and the rich structure of this combinatorial space.



https://distill.pub/2018/building-blocks/

Debug Mode

https://youtu.be/VG68SKoG7vE



https://electrek.co/2018/04/10/tesla-autopilot-engineering-car-leaked-picture-full-self-driving-settings/

DARPA DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

EXPLORE BY TAG

ABOUT US / OUR RESEARCH / NEWS / EVENTS / WORK WITH US /

Defense Advanced Research Projects Agency > Program Information

Explainable Artificial Intelligence (XAI)

Mr. David Gunning

RESOURCES

DARPA-BAA-16-53

DARPA-BAA-16-53: Proposers Day Slides

XAI Program Update

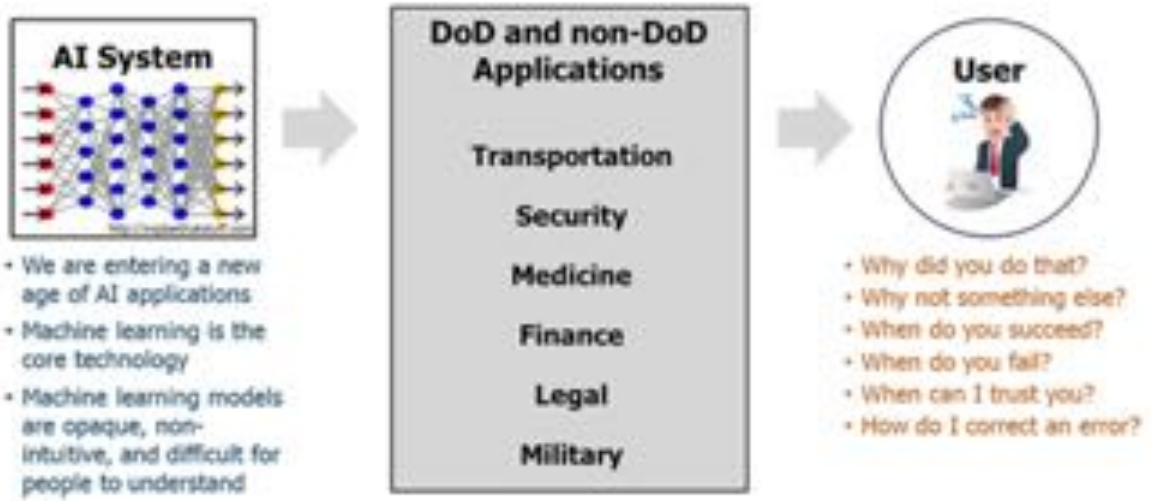


Figure 1. The Need for Explainable AI

https://www.darpa.mil/program/explainable-artificial-intelligence

https://youtu.be/YSsYXAn_L00

Other Links

- "**Attentive Explanations: Justifying Decisions and Pointing to the Evidence**", Park et. al. 14 Dec 2016
  - https://arxiv.org/pdf/1612.04757v1.pdf
- "**Programming your way to explainable AI**", Hammond, O'Reilly AI Conf, NY. Jun 2017
  - https://cdn.oreillystatic.com/en/assets/1/event/258/Programming%20your%20way%20to%20explainable%20AI%20Presentation.pdf
    - Covers RL explainability
- https://blog.goodaudience.com/holy-grail-of-ai-for-enterprise-explainable-ai-xai-6e630902f2a0

Adversarial Attacks

Clean Stop Sign

Real-world Stop Sign in Berkeley

Adversarial Example

Adversarial Example

"Stop sign"

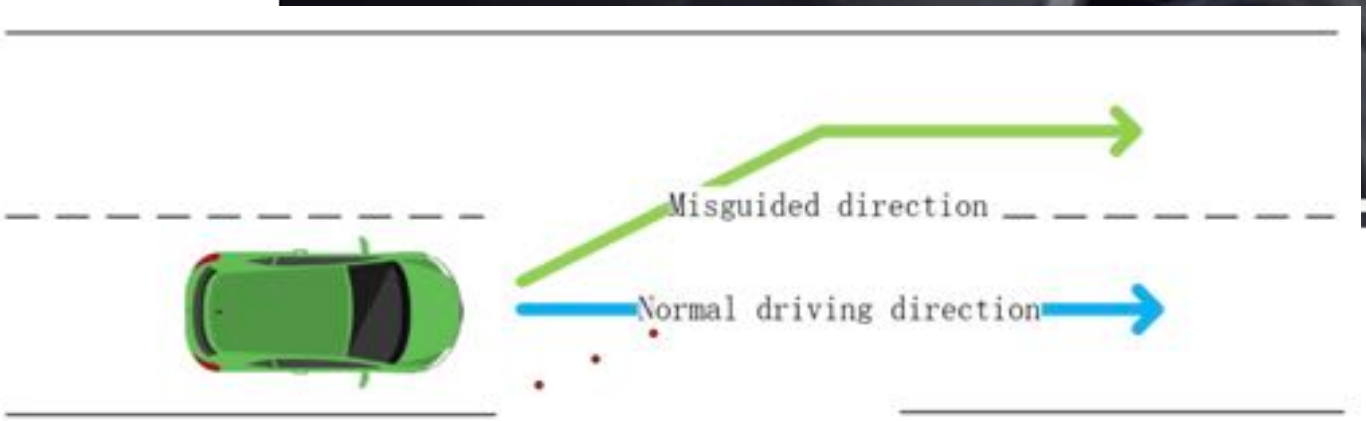
"Stop sign"

"Speed limit sign 45km/h"

"Speed limit sign 45km/h"

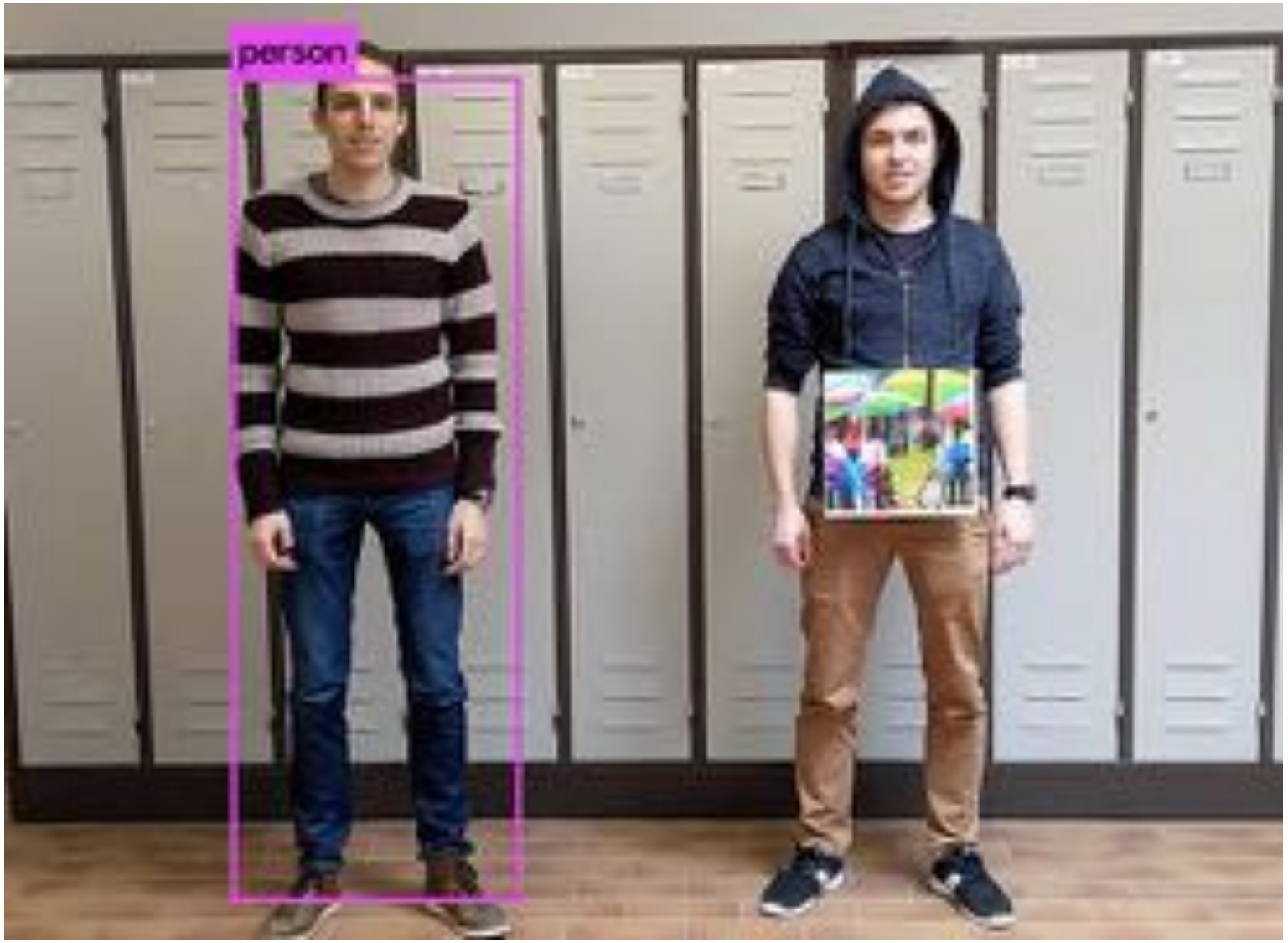
https://www.technologyreview.com/s/613170/emtech-digital-dawn-song-adversarial-machine-learning/

Small stickers on the ground trick Tesla autopilot into steering into opposing traffic lane



https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf

Researchers design patch to make people ‘virtually invisible’ to AI detectors



https://www.computerworld.com.au/article/660283/researchers-design-patch-make-people-virtually-invisible-ai-detectors/

https://www.newscientist.com/article/2111041-glasses-make-face-recognition-tech-think-youre-milla-jovovich/

DARPA DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

EXPLORE BY TAG

ABOUT US / OUR RESEARCH / NEWS / EVENTS / WORK WITH US /

Defense Advanced Research Projects Agency > News And Events

Defending Against Adversarial Artificial Intelligence

DARPA aims to develop a new generation of defenses to thwart attempts to deceive machine learning algorithms

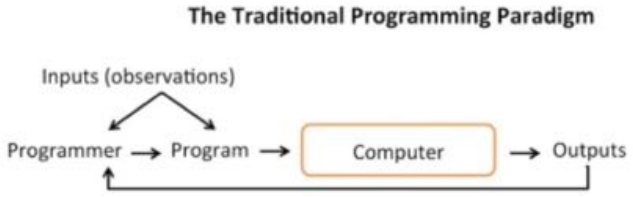
OUTREACH@DARPA.MIL
2/6/2019

https://www.darpa.mil/attachments/GARD_ProposersDay.pdf

Software 2.0 Stack

Andrej Karpathy, Tesla
https://vimeo.com/274274744

Software 1.0

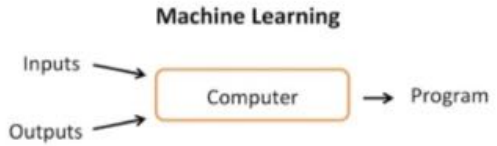


Supporting developers to write rules (programs) to produce outputs from inputs

E.g. IDEs, Test Automation

Software 2.0

Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed
– Arthur Samuel (1959)



Sebastian Raschka, 2016

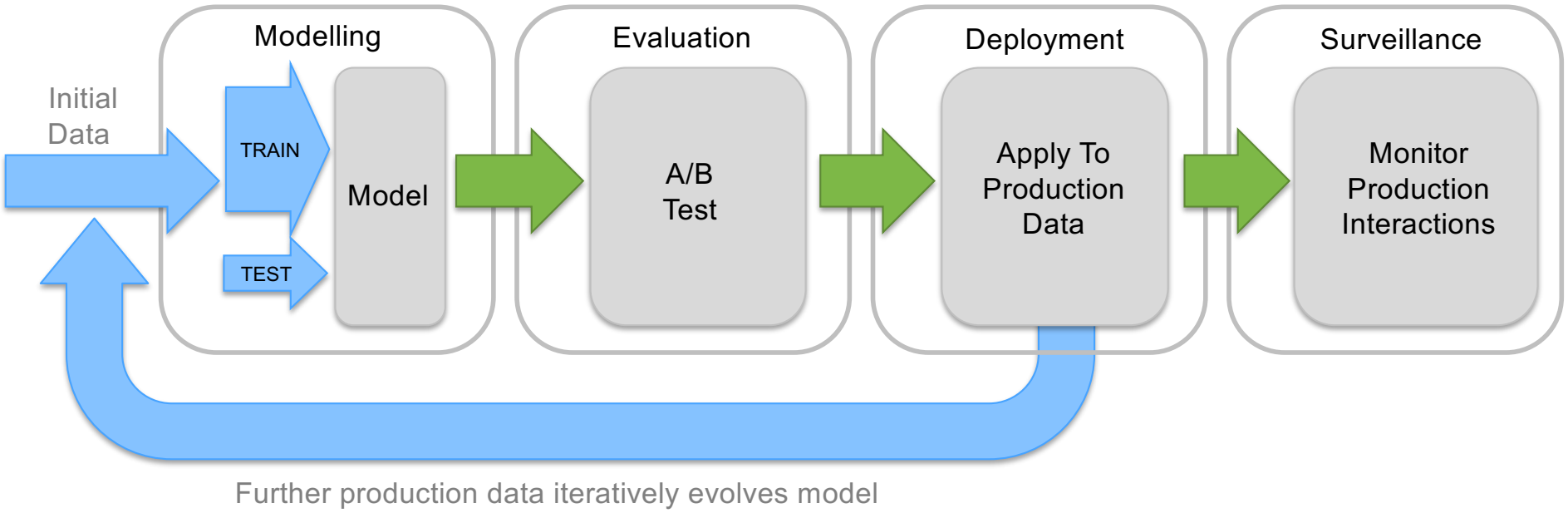
Supporting developers to learn optimal rules (ML architectures and weights) from example inputs and outputs

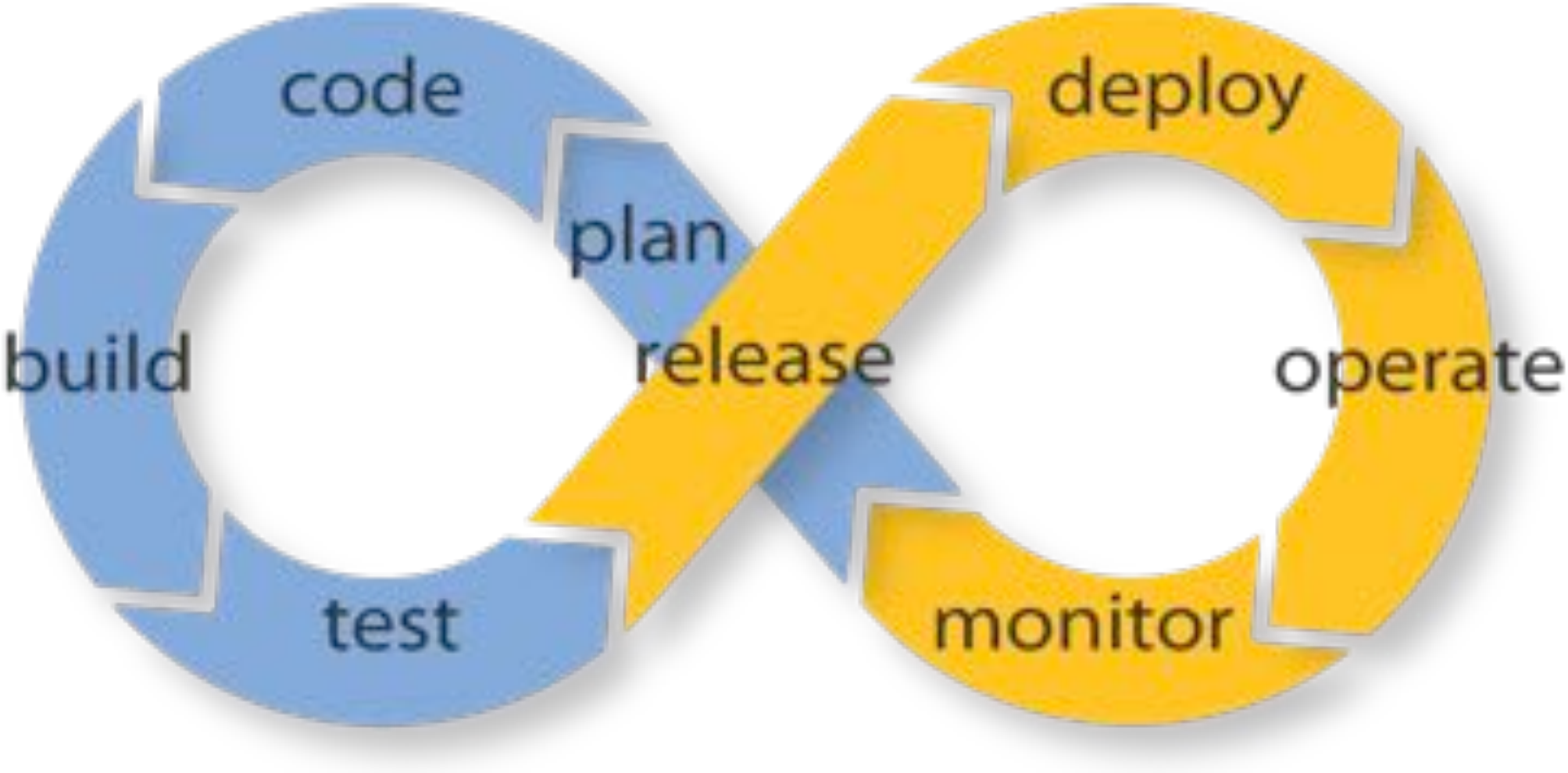
2 main areas supporting teams:

- Label
- Maintain surrounding “Data Infrastructure”
  - Visualise datasets
  - Create/edit labels
  - Bubble up likely mislabeled examples
  - Suggest data to label
  - Flag labeler disagreements
  - ...

Clinical Trial Model

Phase 1	Phase 2	Phase 3	Phase 4
Lab	Single Site	Multi Site	Market Surveillance



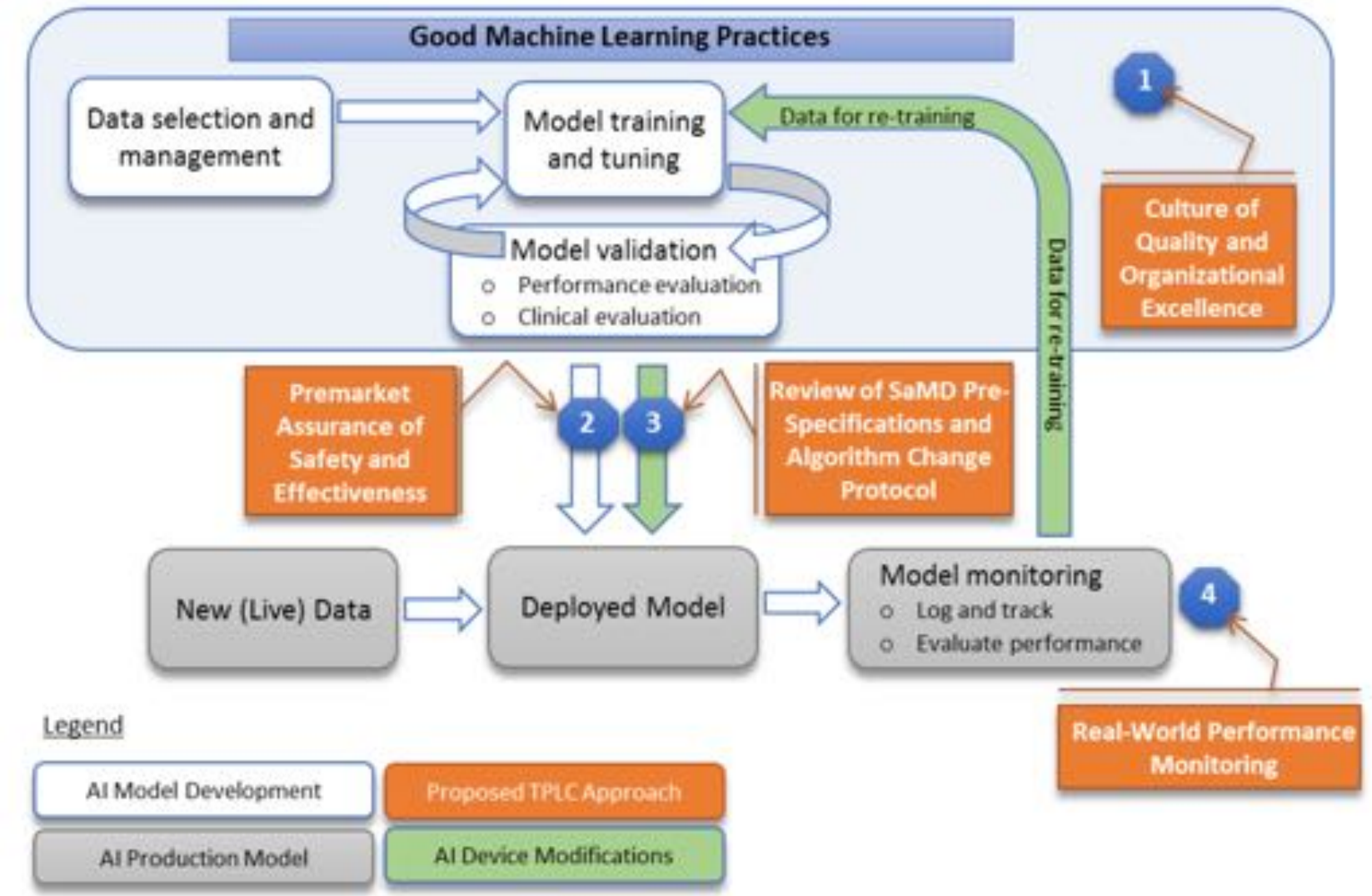


Source: http://readwrite.com/2014/01/01/three-reasons-your-startup-needs-devops-or-else

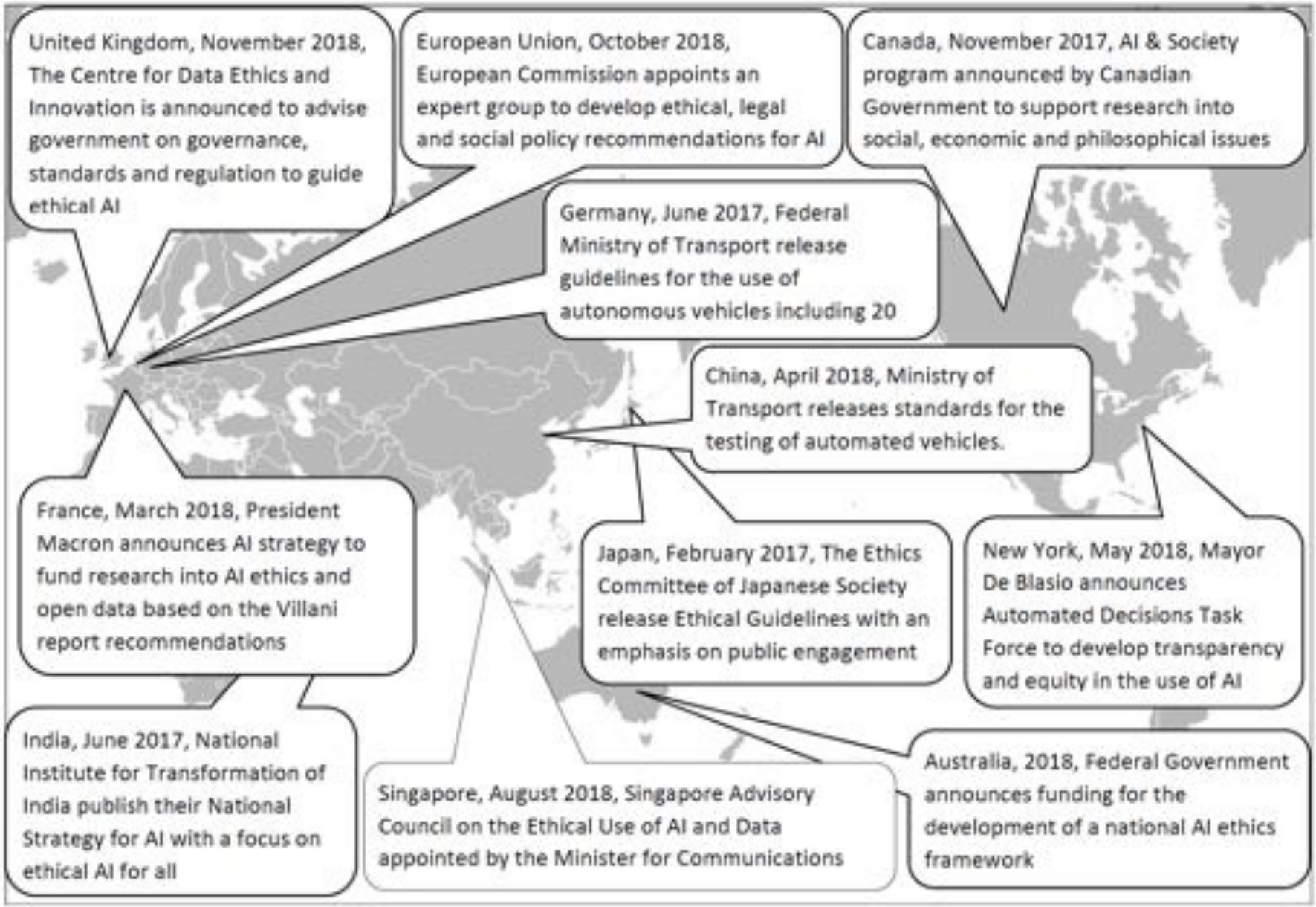
Tesla Data Engine

https://www.youtube.com/watch?v=Ucp0TTmvqOE&t=1h51m05s

FDA: Proposed Regulatory Framework for Modifications to AI/ML based SAMD

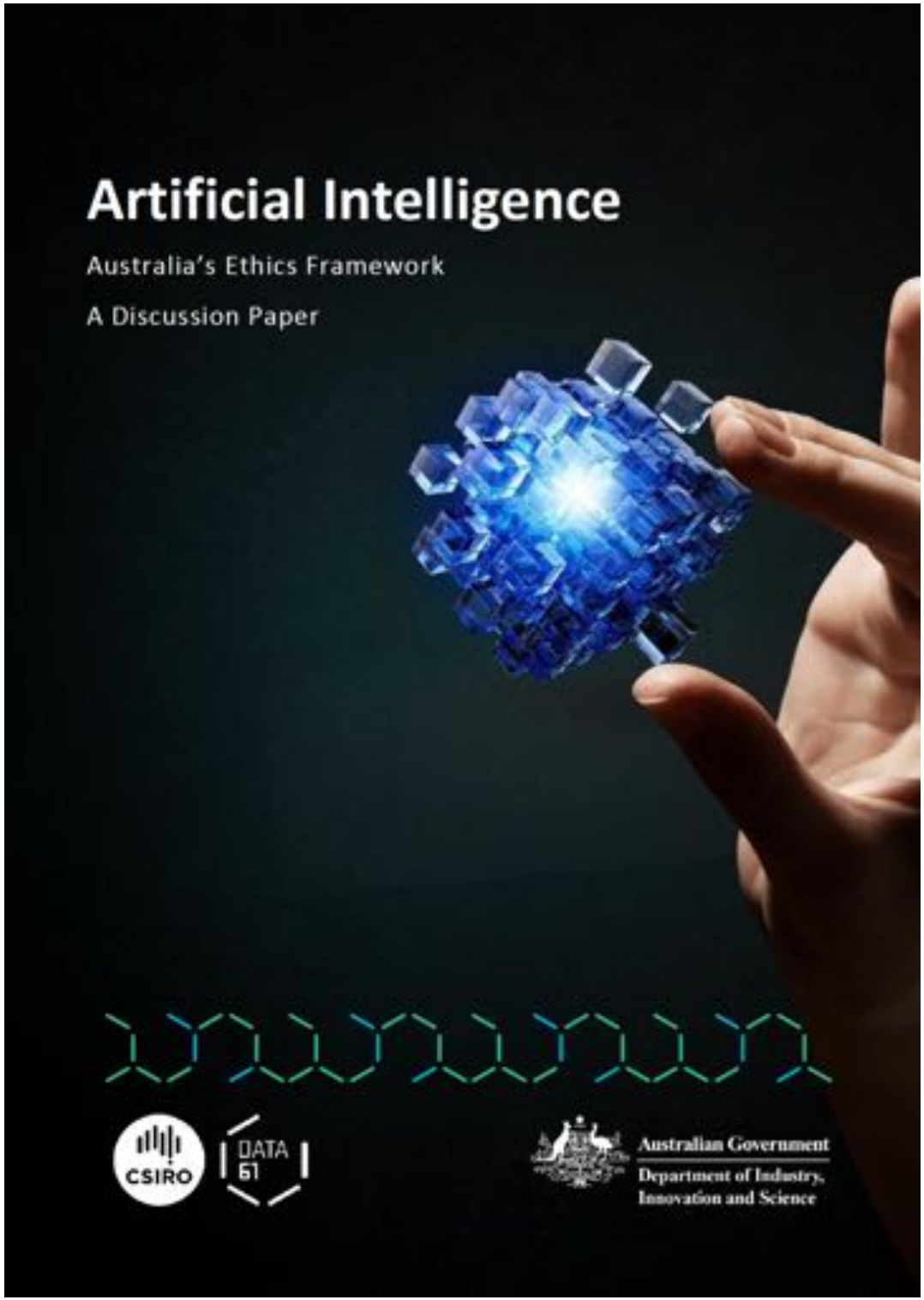


https://www.fda.gov/media/122535/download



https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/

Australia's Response



https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/

Core principles for AI	
1. Generates net-benefits. The AI system must generate benefits for people that are greater than the costs.	**2. Do no harm.** Civilian AI systems must not be designed to harm or deceive people and should be implemented in ways that minimise any negative outcomes.
3. Regulatory and legal compliance. The AI system must comply with all relevant international, Australian Local, State/Territory and Federal government obligations, regulations and laws.	**4. Privacy protection.** Any system, including AI systems, must ensure people's private data is protected and kept confidential plus prevent data breaches which could cause reputational, psychological, financial, professional or other types of harm.
5. Fairness. The development or use of the AI system must not result in unfair discrimination against individuals, communities or groups. This requires particular attention to ensure the "training data" is free from bias or characteristics which may cause the algorithm to behave unfairly.	**6. Transparency & Explainability.** People must be informed when an algorithm is being used that impacts them and they should be provided with information about what information the algorithm uses to make decisions.
7. Contestability. When an algorithm impacts a person there must be an efficient process to allow that person to challenge the use or output of the algorithm.	**8. Accountability.** People and organisations responsible for the creation and implementation of AI algorithms should be identifiable and accountable for the impacts of that algorithm, even if the impacts are unintended.

https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/

Key Takeaways

1. Ability to scale data collection and validation will be key to success
2. A new set of tools required for SDLC (Software 2.0)
3. How to evaluate and build robust models is hot research topic

Questions?

kelvin.ross@kjr.com.au
@kelvinjross

KJR