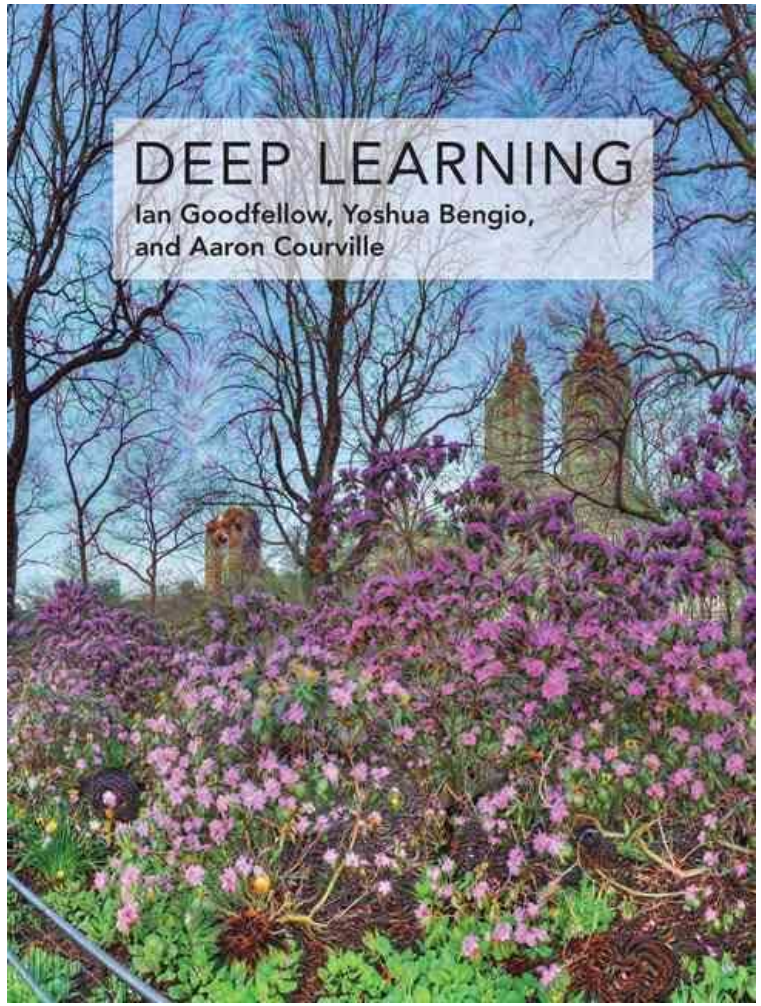
When to trust AI...

Prof. Marta Kwiatkowska
Department of Computer Science
University of Oxford

The changing the face of AI: the rise of deep learning

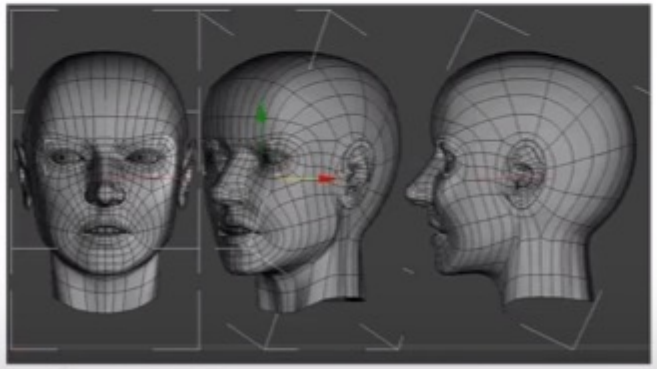
- Neural networks timeline
  - 1940s First proposed
  - 1998 Convolutional nets
  - 2006 Deep nets trained
  - 2011 Rectifier units
  - 2015 Vision breakthrough
  - 2016 Win at Go
  - 2019 Turing Award

- Enabled by
  - Big data
  - Flexible, easy to build models
  - Availability of GPUs
  - Efficient inference



Deep learning with everything…

DeepFace
Closing the Gap to Human-Level Performance in Face Verification



Yaniv Taigman
Ming Yang
Marc'Aurelio Ranzato
Lior Wolf
- 2014

97.35% accuracy
Trained on the largest facial dataset – 4M facial images belonging to more than 4,000 identities.

A Tesla Model S

A clinically applicable approach to continuous prediction of future acute kidney injury

Nenad Tomašev[1]*, Xavier Glorot[1], Jack W. Rae[1,2], Michal Zielinski[1], Harry Askham[1], Andre Saraiva[1], Anne Mottram[1], Clemens Meyer[1], Suman Ravuri[1], Ivan Protsyuk[1], Alistair Connell[1], Cían O. Hughes[1], Alan Karthikesalingam[1], Julien Cornebise[1,12], Hugh Montgomery[3], Geraint Rees[4], Chris Laing[5], Clifton R. Baker[6], Kelly Peterson[7,8], Ruth Reeves[9], Demis Hassabis[1], Dominic King[1], Mustafa Suleyman[1], Trevor Back[1,13], Christopher Nielson[10,11,13], Joseph R. Ledsam[1,13]* & Shakir Mohamed[1,13]

The early prediction of deterioration could have an important role in supporting healthcare professionals, as an estimated 11% of

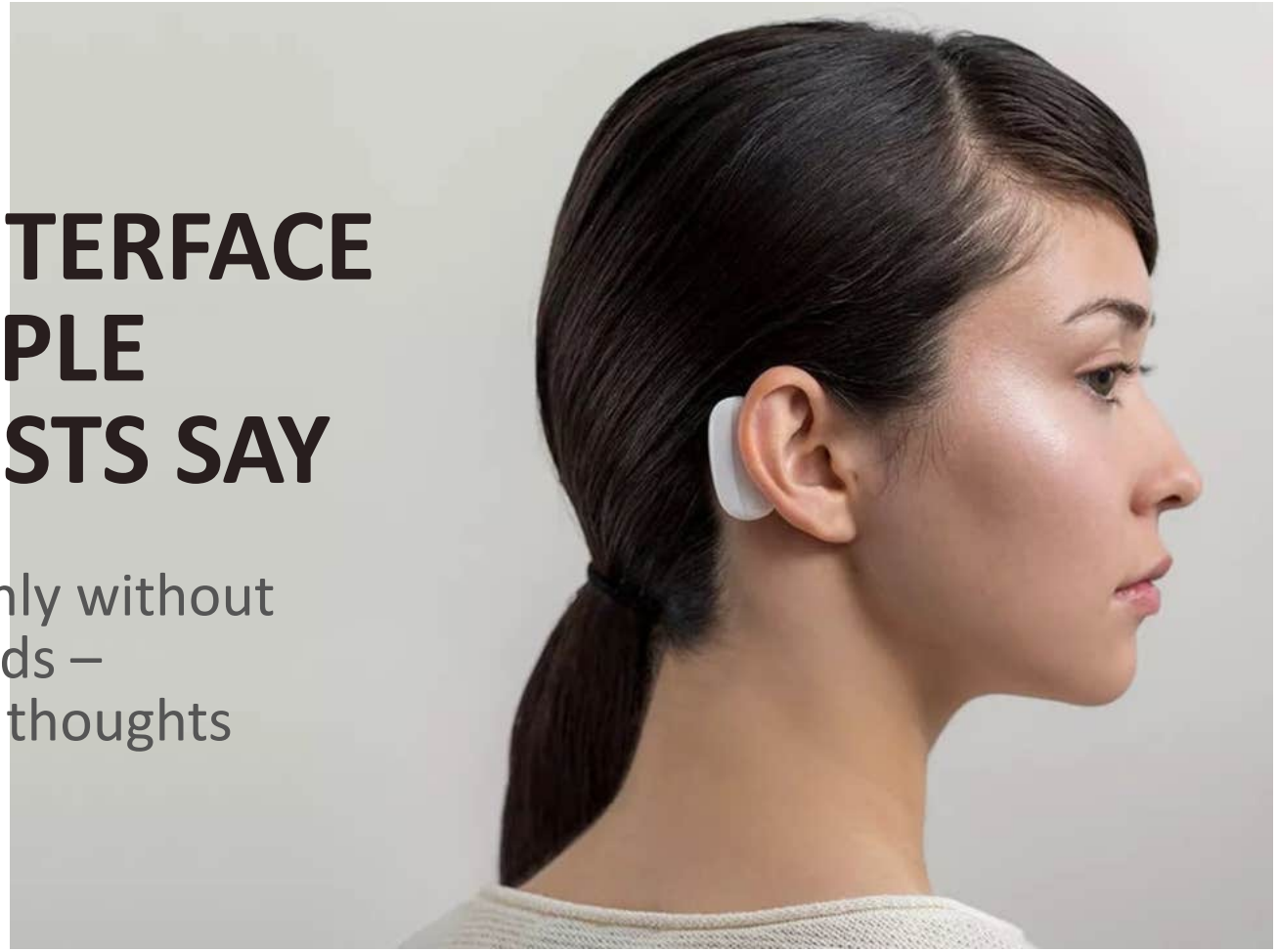
Promising recent work on modelling adverse events from electronic health records[2–17] suggests that the incorporation of machine learning

The stakes are rising even higher…

Brain implants approved…

BRAIN-COMPUTER INTERFACE WILL MAKE PEOPLE TELEPATHIC, SCIENTISTS SAY

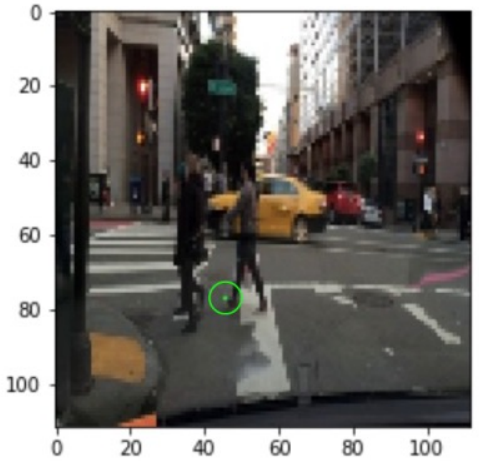
People will communicate 'not only without speaking but without words – through access to each other's thoughts at a conceptual level'



Neuralink says learning to use the device is “like learning to touch type or play the piano”

Should we worry about AI safety?

- Neural networks are unstable to adversarial perturbations!



Red light classified as green
after one pixel change

Physical attack

Real traffic sign

- Need safety assurance, possibly new failure modes

Feature-Guided Black-Box Safety Testing of Deep Neural Networks. Wicker *et al*, In Proc. TACAS, 2018.

The bumpy road to self-driving cars…

- Level 5 autonomous driving promised by 2020, but…

Tesla Says Crashed Vehicle Had Been on Autopilot Before Fatal Accident

By **GREGORY SCHMIDT** MARCH 31, 2018

RELATED COVERAGE

Tesla Looked Like the Fu
Ask if It Has One. MARC

- Now manufacturers scaling back autonomy…
- San Francisco robotaxis have had a mixed reception

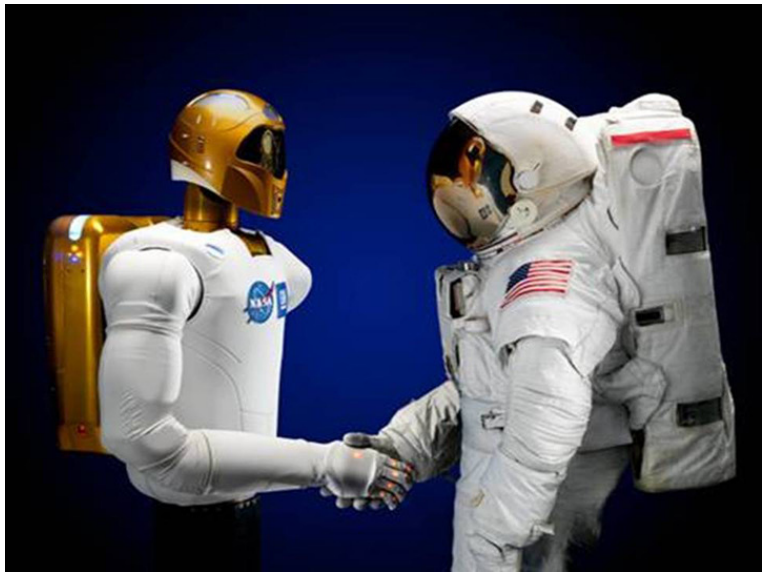
But isn't bumpy for all new technologies?

- Red flag traffic law: safety regulation for automobiles in late 19th century
- Both vehicles and regulation need to be fit for purpose for us to trust them…
- … especially in complex urban environments

To trust AI, or not to trust…

- What is trust?
  - willingness to depend on and be vulnerable to the actions of another party
  - subjective, multifaceted concept
  - can be quickly lost, difficult to rebuild
- Trust what/who? who is responsible?
  - need accountability, appropriate regulation, evidence
- Trust is guides reliance, avoids ovrr-trust

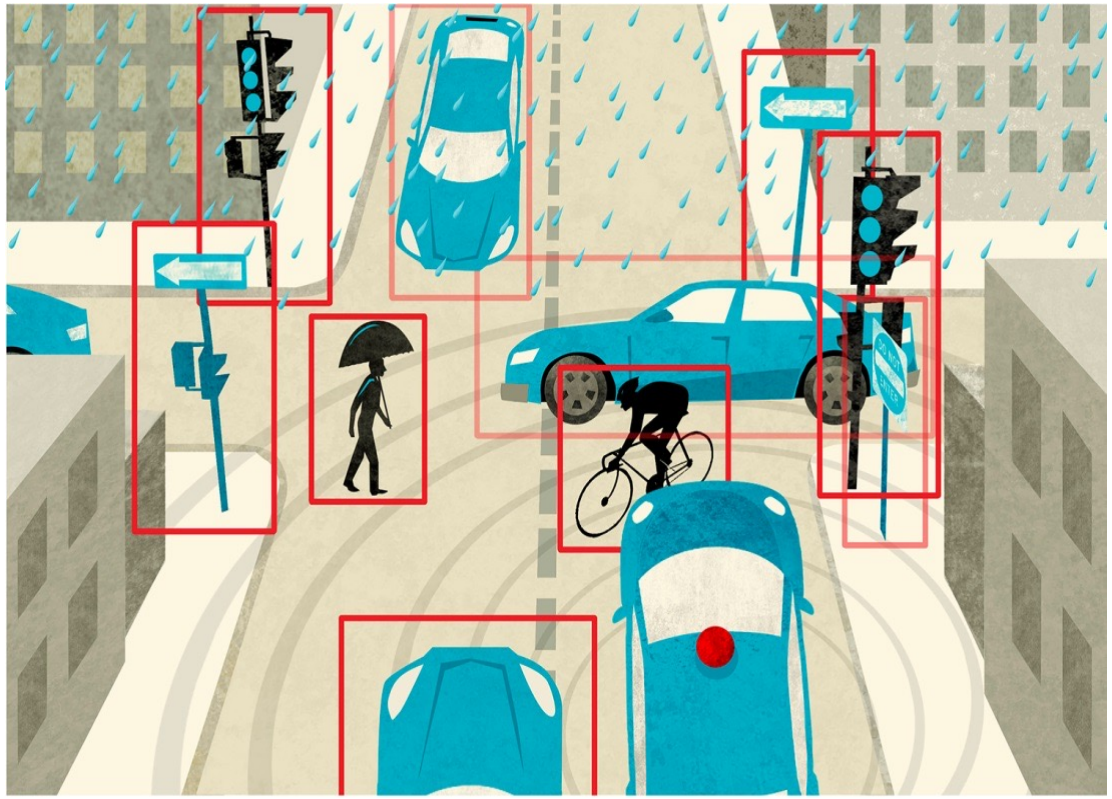
  and is increased through sound engineering: rigour, standards, compliance, safety assurance processes



TRUST is key factor to forming relationships

An AI safety problem…

- Complex scenarios
  - goals and perception
  - autonomy
  - situation awareness
  - context (social, regulatory)
- Safety-critical, need guarantees
- Should failure occur,
  - evidence is needed
  - accountability must to be established



This lecture: focus on trust via rigorous engineering

- Brief recap of robustness for neural networks and highlighting formal verification methodologies being developed to provide provable guarantees
  - focus on adversarial perturbations
  - broad range of applications, not just image classification
  - robustness affects safety, fairness, correctness, conformance, etc

- A snapshot of recent directions beyond robustness to (bounded) perturbations in data-rich settings
  - robustness guarantees for explanations

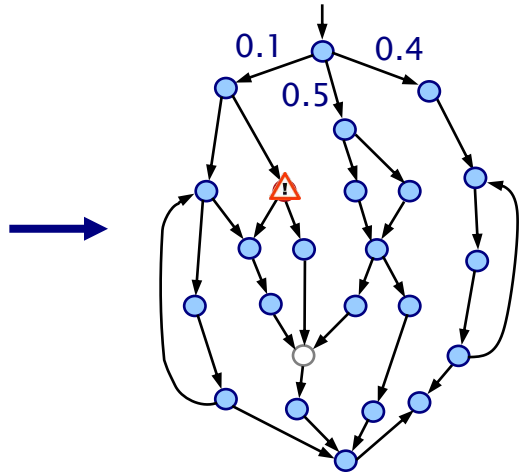
- Perspective on future challenges
- Conclusions and future directions

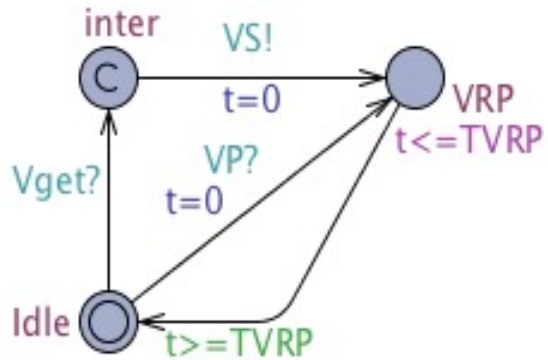
When to Trust AI: Advances and Challenges for Certification of Neural Networks. Kwiatkowska and Zhang, *Proc FedCSIS 2023*

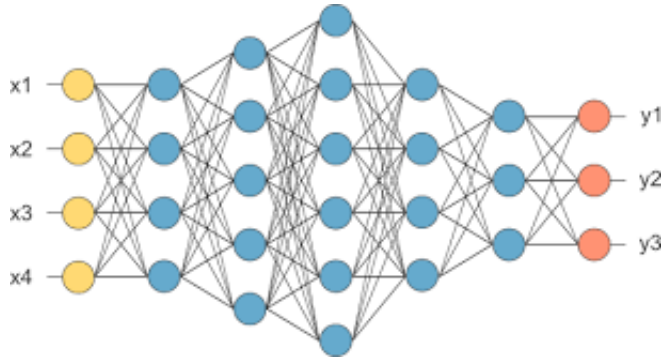
Software verification offers provable guarantees

```
void add(Object o) {
buffer[head] = o;
head = (head+1)%size;
}

Object take() {
...
tail=(tail+1)%size;
return buffer[tail];
}
```







Program

Model

- Modelling = rigorous, mathematical abstraction
- Verification = proof that the model satisfies specification
- Synthesis = correct-by-construction model from specification
- Automated = algorithmic, implemented in software

Formal verification for neural networks

- Neural network models more challenging
  - black box, lacks interpretability
  - programming by pattern matching, not logic
  - corner cases are unseen examples, not missed conditions
  - non-linearity and scale
  - accuracy can be misleading

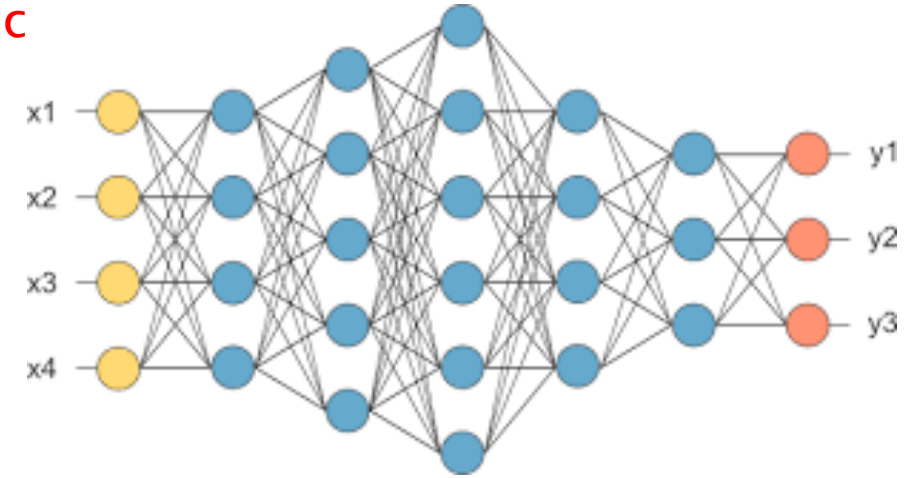


Image classifier is a function f: $R^n \rightarrow \{c_1,\ldots c_k\}$
Learnable weights and bias

Approximates human perception from M training examples

- Formal, automated verification
  - can provide guarantees
  - enables certification
    - testing insufficient
  - and correct-by-construction synthesis (of models)

Safety Verification of Deep Neural Networks. CAV 2017 keynote

How to specify correct behaviour for a neural network?

- For image classification, similar inputs should be mapped to the same class

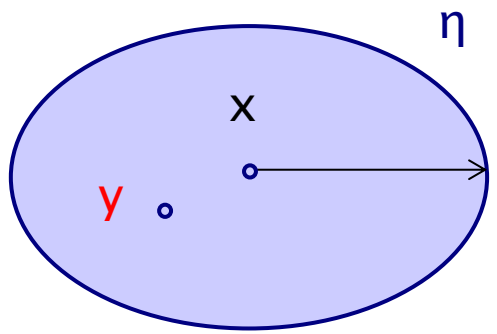
- Problem-specific similarity measures needed to capture key features
  - for images, Mahalanobis distance challenging to work with
  - typically work with L^p norms as a proxy

- Need semantic robustness…

The King is Naked: on the Notion of Robustness for Natural Language Processing. La Malfa et al, *Proc AAAI 2022*

Safety of classification decisions

- Safety assurance process is complex
- Here focus on decision safety as part of such a process
  - local robustness, focus on a specific point x…



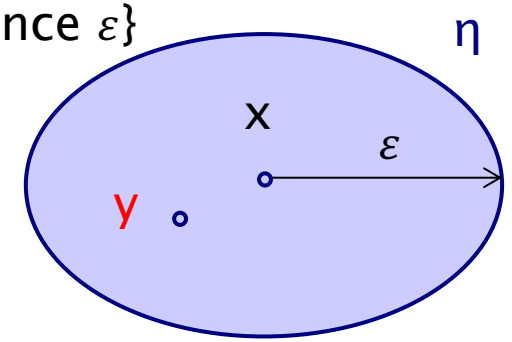
- Assume given
  - trained neural network f : $R^m \rightarrow \{c_1,\ldots c_k\}$
  - support region η for x
  - distance function, e.g. L^2, L^∞

- Define safety as invariance (stability) of classification decision over η
  - i.e. $\nexists y \in \eta$ such that $f(x) \neq f(y)$
- Also wrt family of safe manipulations, e.g. a group of operations
  - e.g. scratches, weather conditions, camera angle, etc

Safety Verification of Deep Neural Networks. CAV 2017 keynote

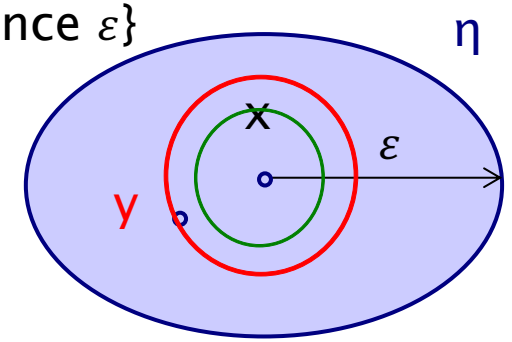
Maximum safe radius

- Measure of safety/robustness, can be average over input distribution
- Define maximum safe radius (MSR)
  - $\mathrm{MSR}(x) = \inf\{\varepsilon > 0 \mid \nexists \text{ adversarial example at distance } \varepsilon\}$
- i.e. the minimum distance from x to the decision boundary



Maximum safe radius

- Measure of safety/robustness, can be average over input distribution
- Define maximum safe radius (MSR)
  - $\text{MSR}(x) = \inf \{\varepsilon > 0 \mid \nexists \text{ adversarial example at distance } \varepsilon\}$
- i.e. the minimum distance from x to the decision boundary
- Difficult to compute, so lower/upper bound



- Intuitively,
  - unsafe: finding an adversarial example at ε_{ub} gives an upper bound on MSR
  - safe: ruling out adversarial examples for all $0 \leq \varepsilon \leq \varepsilon_{lb}$ gives a lower bound on MSR

- But the region contains infinitely many points!

Search-based safety verification

- Take as a specification an input x and region η (and family of manipulations)
  - focus on safety wrt a set of manipulations, e.g. bounded perturbations
  - exhaustively search the region for misclassifications
- Challenges
  - high dimensionality, nonlinearity, infinite region, huge scale
- Automated verification (= ruling out adversarial examples)
  - need to ensure finiteness of search, e.g. Lipschitz
  - guarantee of decision safety if adversarial example not found
- Falsification (= searching for adversarial examples)
  - good for attacks, no safety guarantees

Searching for adversarial examples…

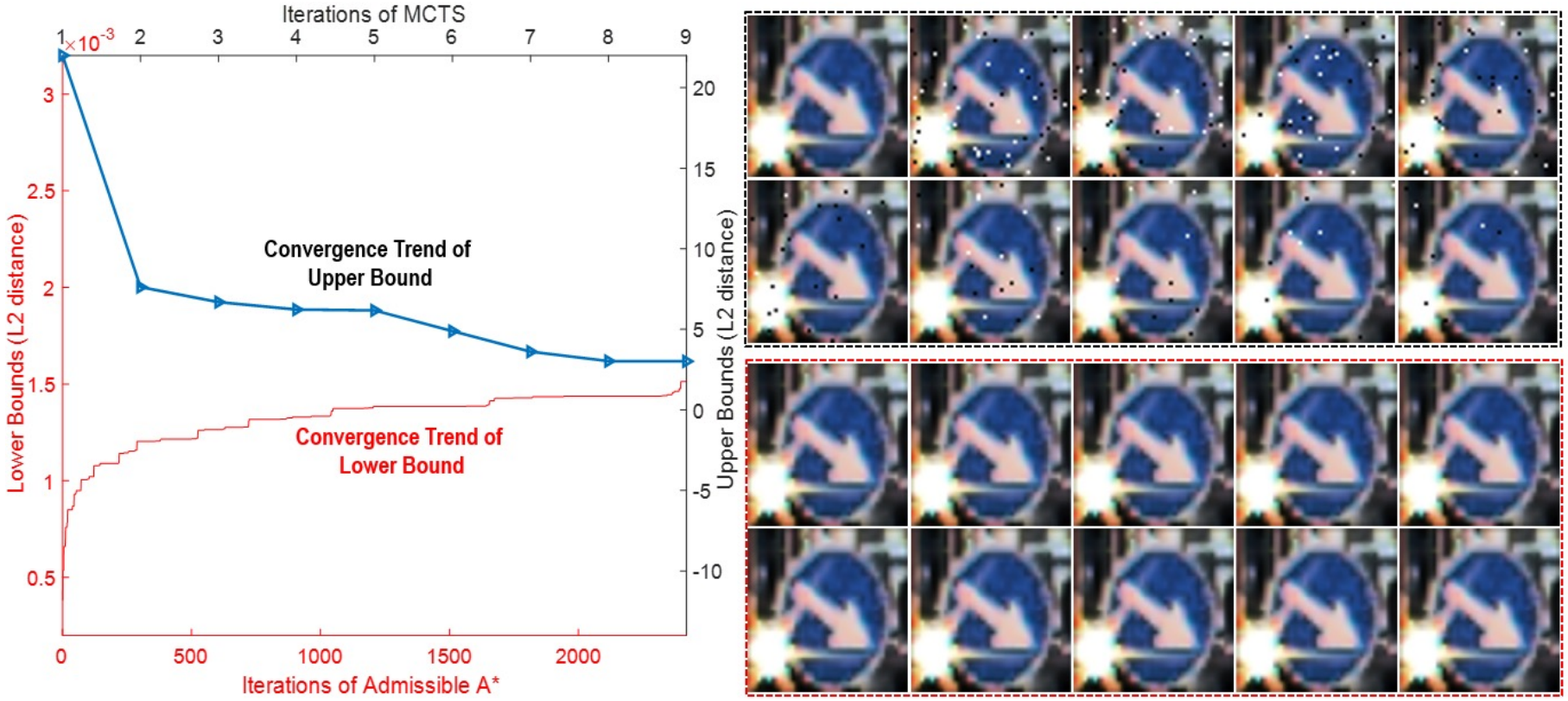
- Input space for most neural networks is high-dimensional and non-linear

- Image of a tree has 4,000 x 2,000 x 3 dimensions = 24,000,000 dimensions
- We would like to find a very small change in these dimensions

- Progress through
  - apply feature-based exploration to reduce dimensionality
  - rely on Lipschitzness of neural networks
  - smart search via a two player game and Monte Carlo Tree Search

MSR robustness in action: image classification

- Consider robustness wrt pixel manipulations
- Convergence of lower and upper bounds on maximum safe radius



A game-based approximate verification of deep neural networks with provable guarantees. Wu *et al, Theor. Comput. Sci.* 807: 298-329 (2020).

MSR not just for images, also videos…

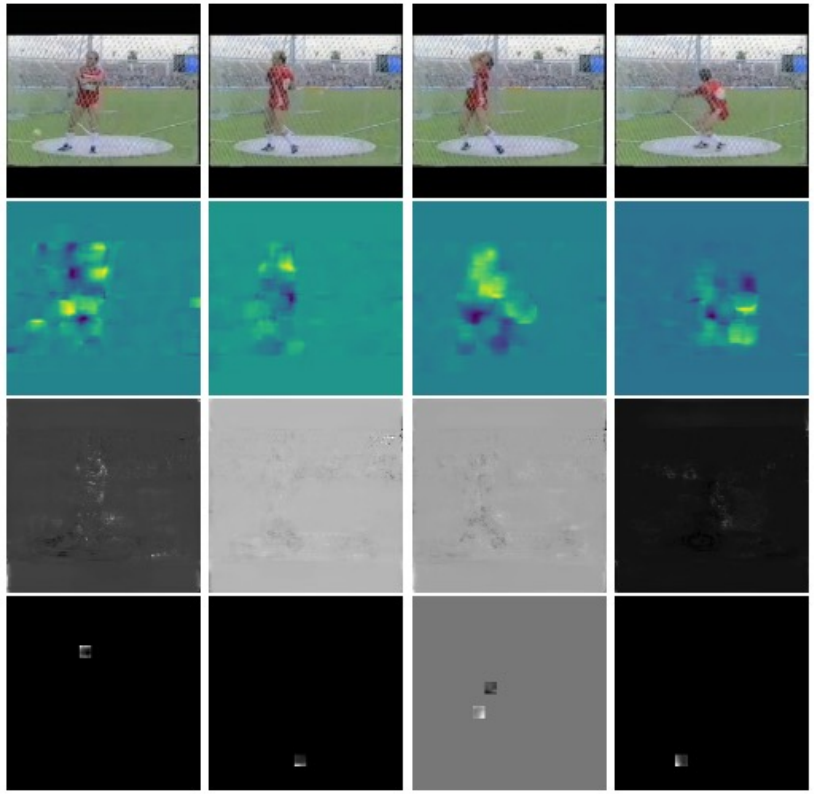
- Guarantees against perturbations of optical flow, extracted from consecutive frames

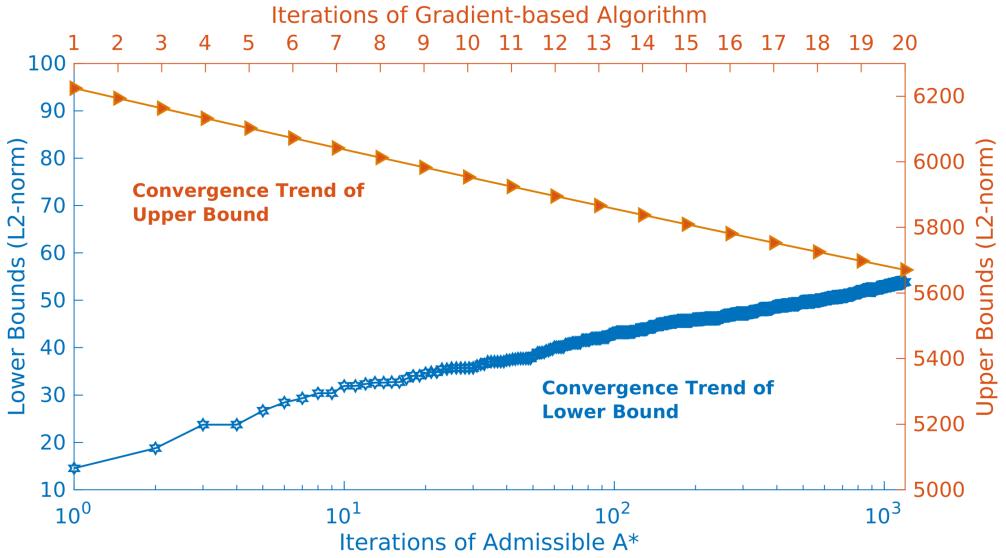
Video frame

Optical flow

Unsafe perturbation

Safe perturbation





Convergence of lower/upper bounds CNN+RNN architecture (VGG16+LSTM, UCF101 dataset)

Many natural distortions: brightness change, angular rotation, …

Robustness Guarantees for Deep Neural Networks on Videos. Wu *et al*, In Proc. CVPR 2020.

And MSR for text classification…

- Consider robustness wrt word substitution, e.g. replacement with a synonym

MCTS ATTACKS - AG DATASET

AG Test Set n° 47, Model Prediction = CLASS "sci-tech", Confidence = 0.53, Words Perturbed = 47/48

ORIGINAL	:	dell	exits	lowend	china	consumer	pc	market [..]
REPLACEMENT	:	parsons	misses	founds	u.s.	benefits	parsons	wall

AG Test Set n° 12, Model Prediction = CLASS "sci-tech", Confidence = 0.86, Words Perturbed = 0/42

ORIGINAL	:	dutch	retailer	beats	local	download	market	[..]
REPLACEMENT	:	-	-	-	-	-	-	

AG Test Set n° 49, Model Prediction = CLASS "sport", Confidence = 0.75, Words Perturbed = 3/33

ORIGINAL	:	ranked	player	who	has	not	won	a	major	champ.	since	his [..]
REPLACEMENT	:	-	replacements	-	-	-	-	-	-	wrestling	-	joke

green: meaningful replacement red: replacement (grammatically inconsistent) - : no replacement found

- NB in some cases no replacement found to change classification

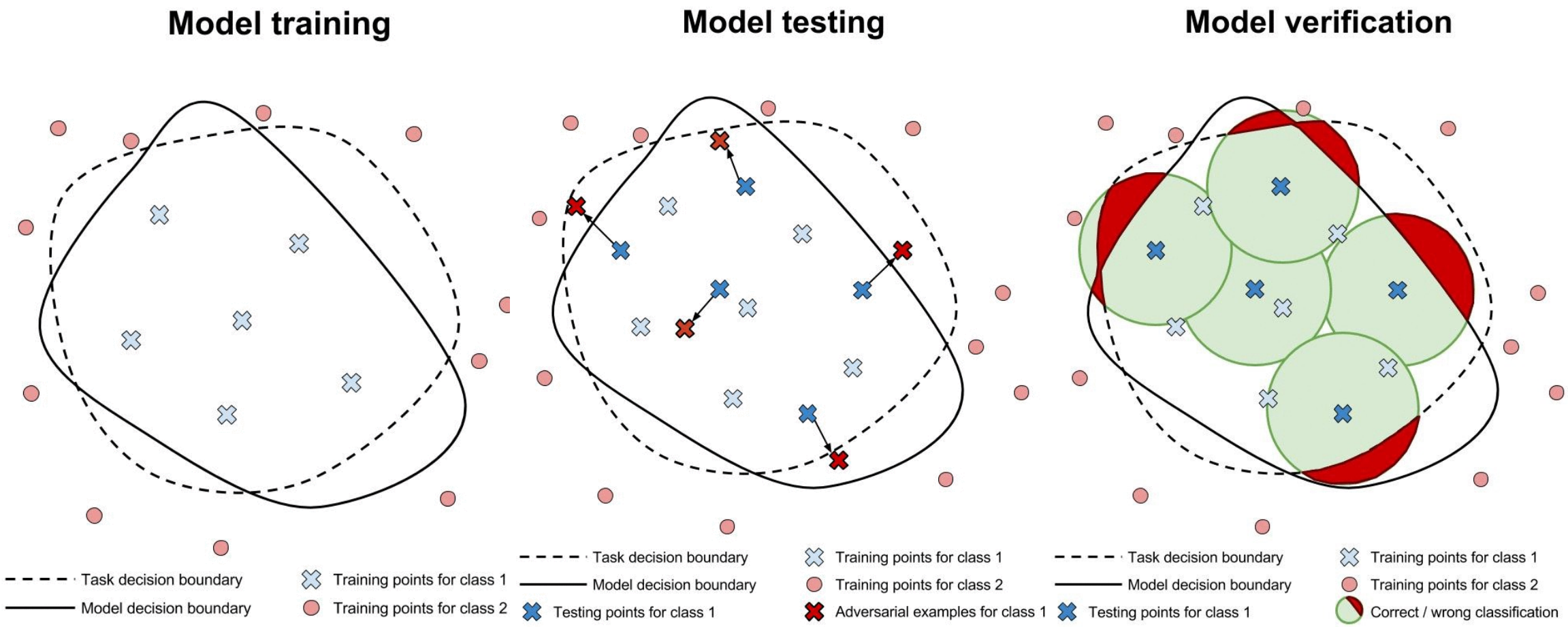
Assessing Robustness of Text Classification through Maximal Safe Radius Computation. La Malfa *et al,* Proc EMNLP 2020.

Evaluating safety-critical scenarios: Nexar

- Using the game-based method we were able to reduce the accuracy of the network from 95% to 0%
- On average, each input took less than a second to manipulate (.304 seconds)
- On average each image was vulnerable to 3 pixel changes
- Part of safety assurance

Feature-Guided Black-Box Safety Testing of Deep Neural Networks. Wicker *et al*, *Proc TACAS 2018*

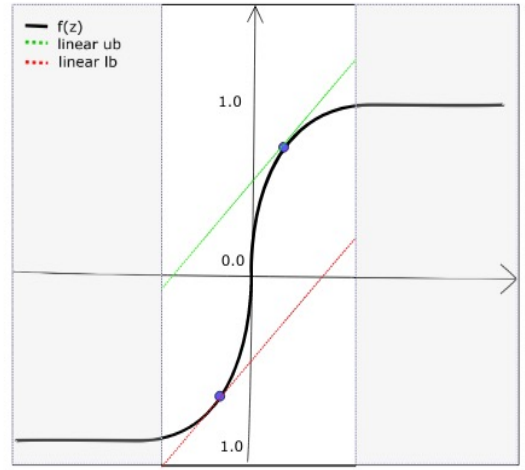
Training vs testing vs verification



Credits: Cleverhans, http://www.cleverhans.io/

Alternative approaches: reachability analysis

- Rather than search, consider input–output relationship and compute (or overapproximate) the reachable values
  - for $x \in \eta$, compute maximum/minimum value of $f(\eta)$
- Methods include exact/approximate
  - constraint solving/SMT, e.g., Reluplex
  - convex relaxation, e.g., linear bound propagation, as in CROWN
  - abstract interpretation, e.g., DeepPoly
  - global optimisation, under assumption of Lipschitz continuity, e.g., DeepGO
- Gives provable guarantees
  - best/worst case confidence values
  - can average over input distribution
- Hard problems, typically NP-completeness, scaling an issue



(a)

How good are MSR guarantees?

- Game-based method practical, enables robustness evaluation
  - model-agnostic and versatile, can be applied to a range of tasks, such as text classification, video processing, etc
  - can be configured with different norms/metrics
  - applicable to continuous (via convex relaxation) and discrete search spaces
- but
  - also need to consider explainability
  - robustness to perturbations too limiting, need robustness to interventions
  - need optimality of learnt policies, not just robustness guarantees
  - ideally, want to learn robust models

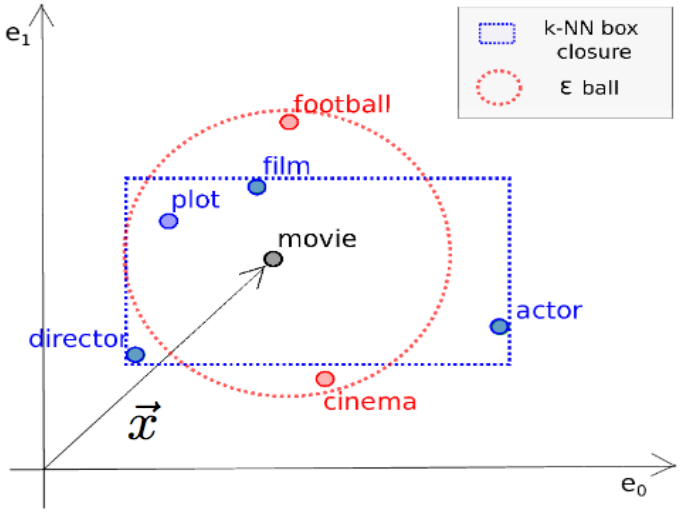
- More effort needed to study topics beyond MSR robustness

On the Hardness of Robust Classification. Gourdeau *et al*, In Proc. *NeurIPS 2019*, extended *JMLR, 22(273)* 2021
When are Local Queries Useful for Robust Learning? Gourdeau *et al*, In Proc. *NeurIPS 2022*

Beyond MSR: explainability and robustness

- Deep learning models are black-box and explaining their decisions helps
- Explanations are sets of features that justify the network decision
  - various tools (LIME, Anchors, ABE, etc) exist
- but explanations lack robustness

- Propose optimal robust explanations (OREs)
  - an ORE is a sufficient reason behind the model's prediction
  - that is robust, i.e., guaranteed to be invariant to perturbations to all the other features
  - and optimal wrt user-defined function
  - consider different norms (L_p, k-NN box closure) for bounding box and uniform/non-uniform cost functions



On Guaranteed Optimal Robust Explanations for NLP Models. La Malfa *et al*, In Proc. IJCAI 2021.

Examples of robust explanations, for sentiment analysis

- Can produce compact robust explanations, in general

'# this movie is really stupid and very boring most of the time there are almost no ghoulies in it at all there is nothing good about this movie on any level just more bad actors pathetically attempting to make a movie so they can get enough money to eat avoid at all costs.' (**IMDB**)	'# well I am the target market I loved it furthermore my husband also a boomer with strong memories of the 60s liked it a lot too i haven't read the book so i went into it neutral i was very pleasantly surprised its now on our highly recommended video list br br.' (**IMDB**)
'The main story ... is compelling enough but it is difficult to shrug off the annoyance of that chatty fish.' (**SST**)	'Still this flick is fun and host to some truly excellent sequences.' (**SST**)
'i couldn't bear to watch it and I thought the UA loss was embarrassing ...' (**Twitter**)	'Is delighted by the beautiful weather.' (**Twitter**)

- Accurate models are over-sensitive to polarized terms or trivial tokens

'# I gave this a 2 and it only avoided a 1 because of the occasional unintentional laugh the film is excruciatingly. Boring and incredibly cheap its even worse if you know anything at all about the Fantastic Four.', (**IMDB, predicted as negative**)

On Guaranteed Optimal Robust Explanations for NLP Models. La Malfa *et al*, In Proc. IJCAI 2021.

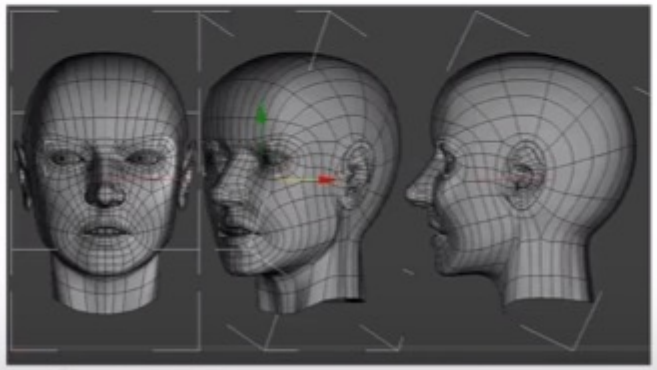
Challenges for AI

- To imbue trust in AI, rigorous verification methodologies can help
- Some successes beyond simplistic MSR robustness, but many limitations

- To make progress, consider:
- Beyond supervised robustness: need robustness evaluation for unsupervised or semi-supervised setting
- Compositionality: formulate a compositional framework based on assume-guarantee interfaces
- Scalability, efficiency and precision: needs to be improved
- Calibrating uncertainty: the models need to be able to say when they don't know, such as Bayesian neural networks
- Reasoning vs statistics: neural network learn statistical features of input data and lack (deductive) reasoning ability, need smart integration of the two

Summarising, much excitement about AI deployment!

DeepFace
Closing the Gap to Human-Level Performance in Face Verification



Yaniv Taigman
Ming Yang
Marc'Aurelio Ranzato
Lior Wolf
- 2014

97.35% accuracy
Trained on the largest facial dataset – 4M facial images belonging to more than 4,000 identities.

nature International weekly journal of science

Article metrics for:

Dermatologist-level classification of skin cancer with deep neural networks

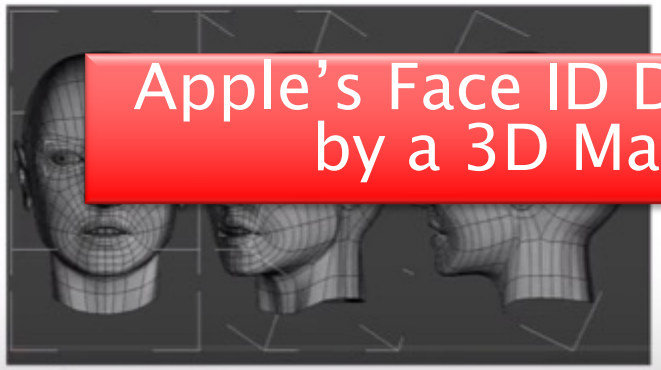
Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun

Nature **542**, 115–118 (02 February 2017) | doi:10.1038/nature21056

Last updated: 24 July 2017 10:10:28 EDT

Continuing concerns about risks and failures…

DeepFace
Closing the Gap to Human-Level Performance in Face Verification



Yaniv Taigman

Trained on the largest facial dataset – 4M facial images belonging to more than 4,000 identities.

Apple's Face ID Defeated by a 3D Mask

nature

Home | News | Re

Archive | Volume

Article metrics fo

Dermatologis… networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun

Nature **542**, 115–118 (02 February 2017) | doi:10.1038/nature21056

Last updated: 24 July 2017 10:10:28 EDT

IBM's "Watson for Oncology" Cancelled After $62 million and Unsafe Treatment Recommendations

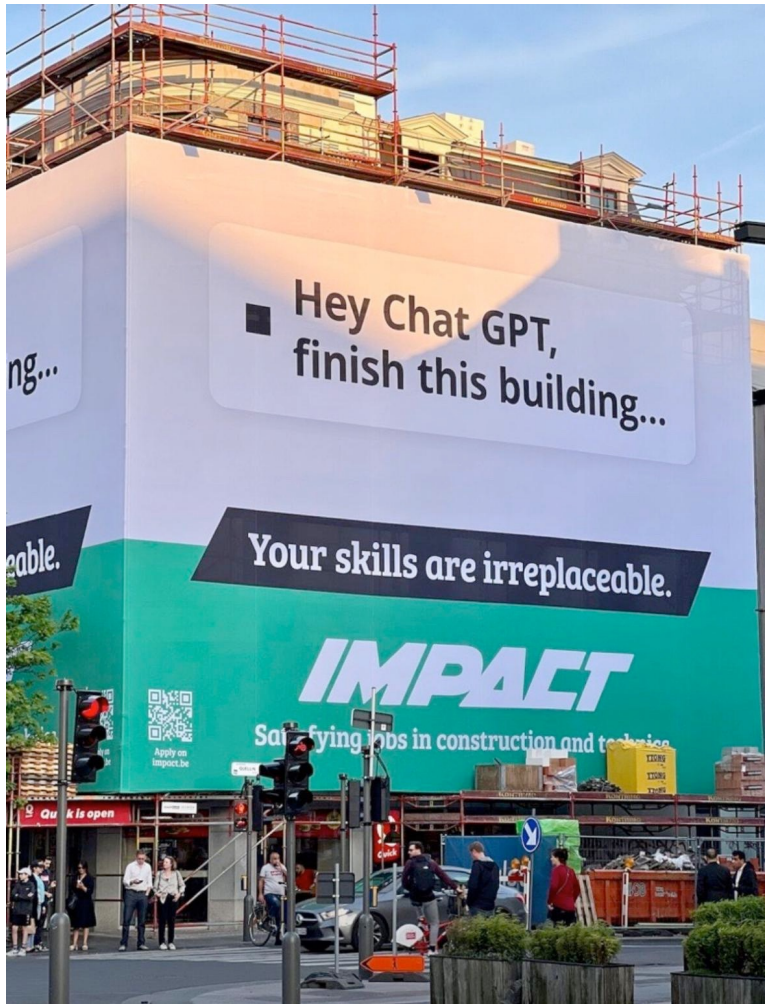
Bu…

Amazon bolsters Alexa privacy after user trust takes a hit

Learn more

amazon alexa

AI is a tool but will not solve all our problems…



Concluding remarks

- Range of techniques developed, active research programmes in the AI, learning and formal verification communities
- Despite progress, major challenges remain
  - complex scenarios and properties, ambiguity, scalability, human involvement!
  - foundational understanding needed
  - neuro-symbolic models
  - robust learning for correct-by-construction models and policies
- Need integrated processes for validation and safety assurance, not just (probabilistic) verification
- Deployment in the wild poses accountability challenges
  - aligning predictive goals with outcomes, e.g., measuring potential for defaulting
  - reason giving, or explainability of decisions

Acknowledgements

- My group and collaborators in this work

- Project funding
  - ERC Advanced Grant fun2model
  - EPSRC project FAIR: Framework for responsible adoption of artificial intelligence in the financial services industry, https://www.turing.ac.uk/research/research-projects/project-fair-framework-responsible-adoption-artificial-intelligence
  - ELSA European Lighthouse on Secure and Safe AI, https://www.elsa-ai.eu/

- See also
  - PRISM www.prismmodelchecker.org