Università di Genova

SmartLab

Pattern Recognition and Applications Lab

University of Cagliari, Italy

From known knowns to unknown unknowns in AI: Historical and Technical Issues

Fabio Roli

Deep Learning and Computer Vision School, Genova, 12 June 2024

My take-home messages for today

- Modern AI: which kind of AI is? And how we got here over the last 70 years.
- The history of AI can be regarded as an evolution from learning of «known knowns» to learning of «unknown unknowns»
- To get to modern AI we have taken a «shortcut», a successful one. But we should aware of the potential limitations of modern AI

The beginning of the story, summer 1955...

[http://www.aaai.org/ojs/index.php/aimagazine/article/view/1904]

A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence

August 31, 1955

John McCarthy, Marvin L. Minsky,
Nathaniel Rochester,
and Claude E. Shannon

The beginning of the story, summer 1955...

"The study will proceed on the basis of the conjecture that, in principle, ***every aspect*** *of* ***learning*** *or* ***any*** *other* ***feature*** *of* ***intelligence*** *can be described so precisely that a* ***machine can*** *be constructed to* ***simulate it***"

"We propose a ***two-month*** *study done by* ***ten people****...*
... We believe that in one summer, ***significant progress*** *can be made in the development of* ***artificial intelligence***"

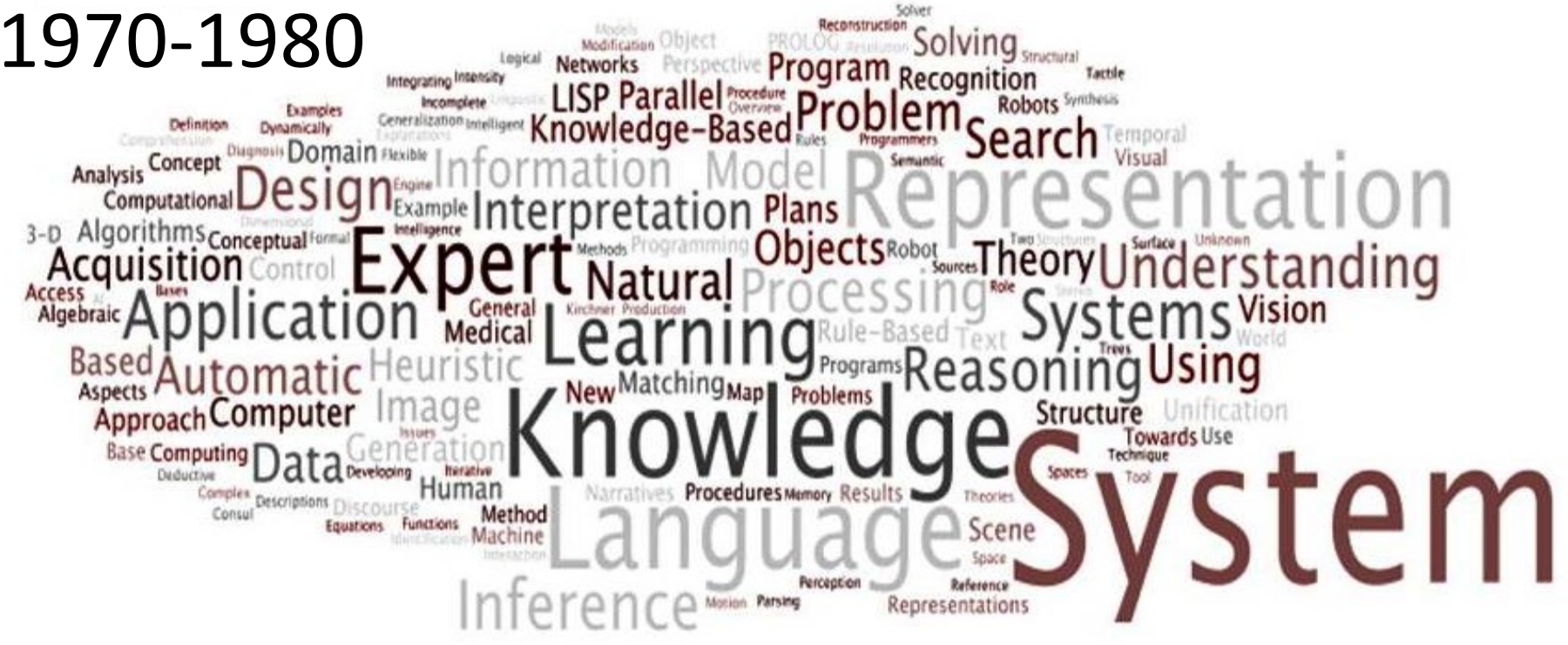
How we got to ChatGPT over the last 70 years?

Disclaimer: this is a short and biased history of AI…

What was AI mainstream in 1970-1980?

AI mainstream in 1970-1980

1970-1980



N. Cristianini, *On the current paradigm in artificial intelligence* (2014)

1980: The INDUCE algorithm by Ryszard S. Michalski

- A learning algorithm from examples that generates **symbolic descriptions** of object classes
  - The description language of object classes is a simple extension of the first-order predicate calculus

1. Learning starts from one example of a class (the «seed»)
2. Examples of other classes are «counter-examples»
3. Guided search for symbolic descriptions that «cover» all the examples of the given class and do not cover the counter-examples

1937 -2007

Ryszard S. Michalski, IEEE-T PAMI, 1980

Learning of object classes with INDUCE…

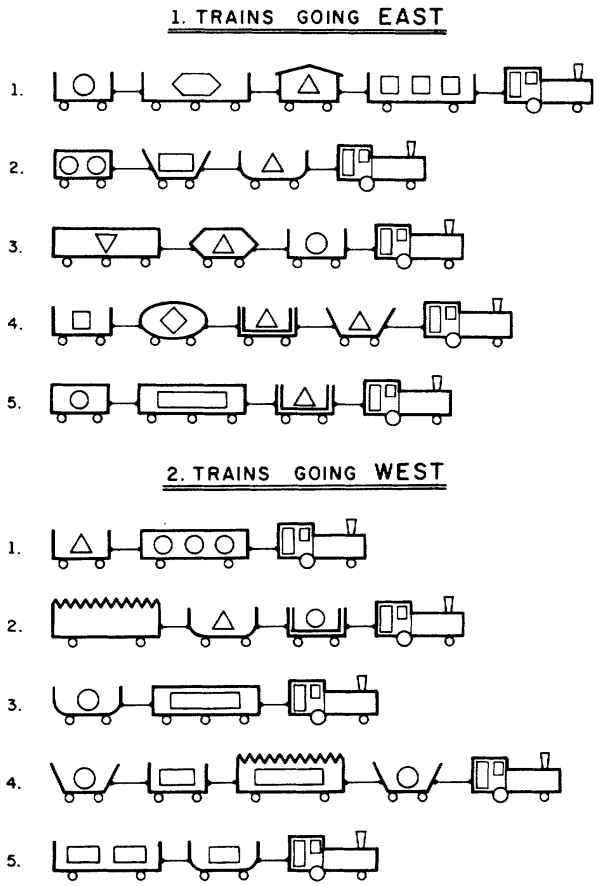


Fig. 4.

Eastbound Trains:

$$\exists\, car_1\ [length(car_1)=short]\ [car\text{-}shape(car_1)=closed\ top]$$
$$::>\ [class=Eastbound] \qquad (22)$$

IF a train contains a car that is short and has a closed top, THEN it is an Eastbound train

Machine learning of "known knowns"

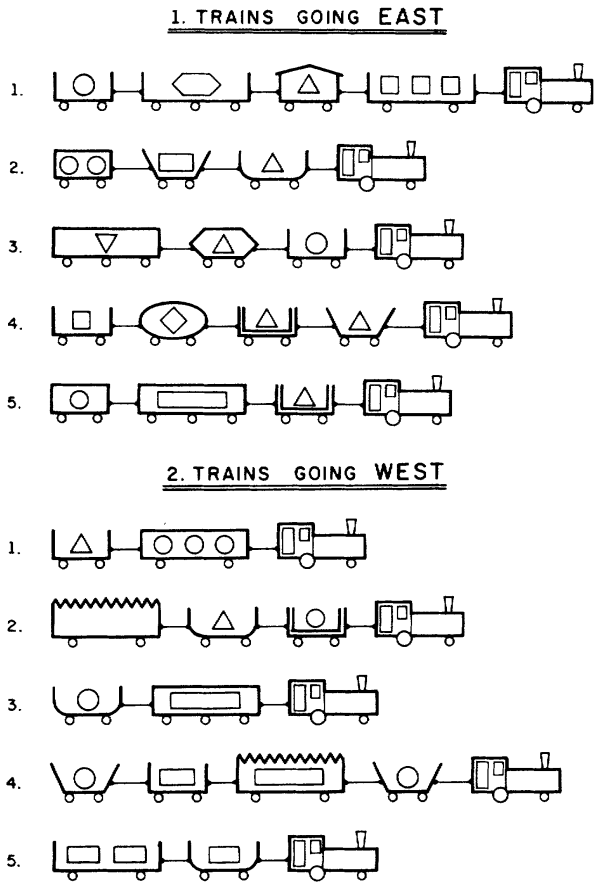


Fig. 4.

Eastbound Trains:

$$\exists\, car_1\ [length(car_1)=short]\ [car\text{-}shape(car_1)=closed\ top]$$
$$::>\ [class=Eastbound] \qquad (22)$$

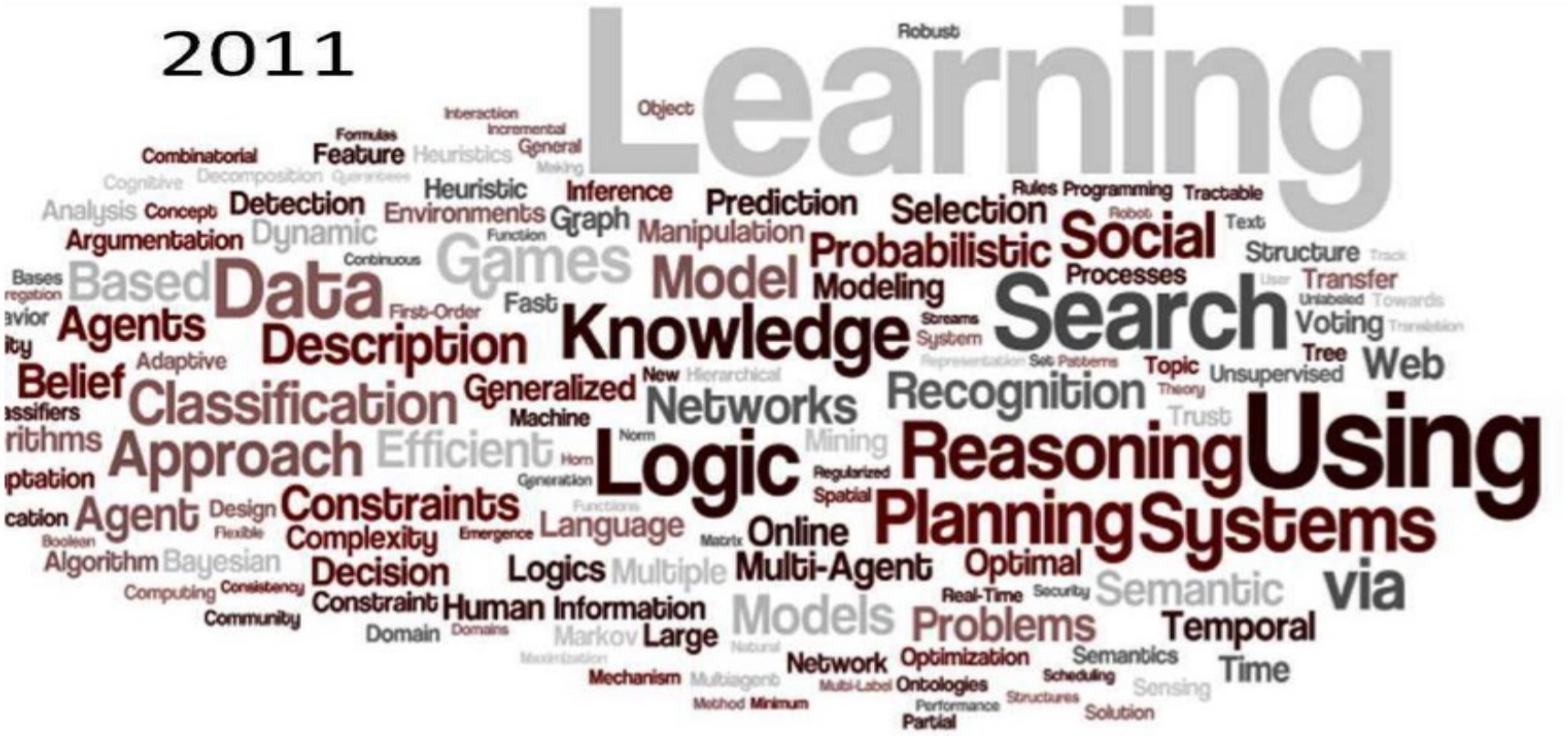
IF a train contains a car that is short and has a closed top, THEN it is an Eastbound train

- «micro worlds» that were perfectly known and predictable («noise-free»)
- The INDUCE algorithm dealt with **«known knowns»**

Credits: slide partially inspired from a speech by Donald Rumsfeld and a research project by Thomas Dietterich
https://futureoflife.org/ai-researcher-thomas-dietterich/

What is AI mainstream nowadays?

What is AI mainstream nowadays?



N. Cristianini, *On the current paradigm in artificial intelligence* (2014)

XD: eXtreme Data-driven Learning

Here we are

After 50 years of research in AI, the main stream is **Big Data + Deep Learning + GPUs**

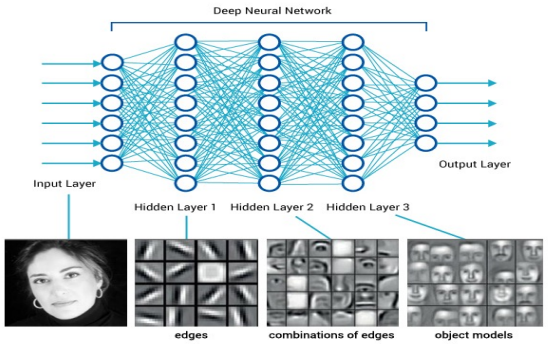
Big Data

Facebook 350 millions of images per day

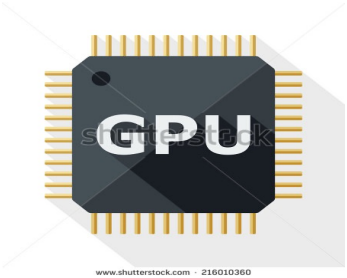
Walmart 2.5 Petabytes customer data hourly

YouTube 300 hours of videos per minute

Deep Learning



GPU



Machine learning of "known unknowns"

- **Known** because we usually assume that we know all the object/data classes to recognize (supervised learning)
- **Unknown** because data are affected by noise that is usually unknown
- **Underlying assumption:** past data is *representative* of future data (IID data)
- The success of modern AI is on tasks for which we collected enough representative training data for the object classes to recognize, or the next word to predict…

The current mainstream of “known unknowns”

Big Data + GPU + Deep Learning

How we got here ?

Machine learning as an experimental science…

«Machine learning is a scientific discipline and, like the fields of AI and computer science, has both theoretical and empirical aspects.

[…]

Although experimental studies are not the only path to understanding, we feel they constitute one of machine learning's brightest hopes for rapid **scientific progress**, and we encourage other researchers to join in this evolution.»

Pat Langley

Machine Learning as an Experimental Science (1988)

The rise of benchmark data sets

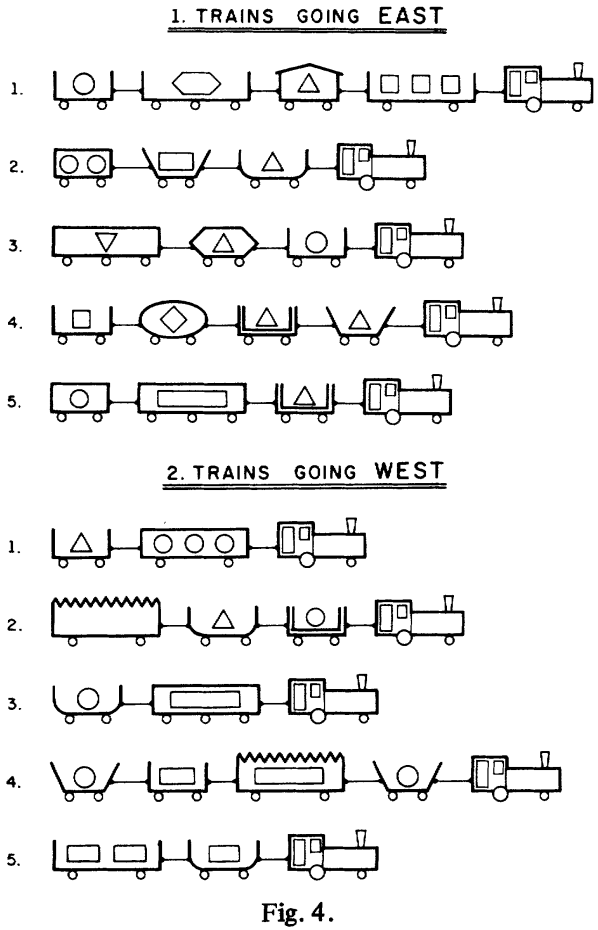


Fig. 4.

Bigger is better…

4 classes
2209 objects

2005

10 classes
4754 objects

2006

The PASCAL VOC data set

21.841 classes
14.197.122 images

2020

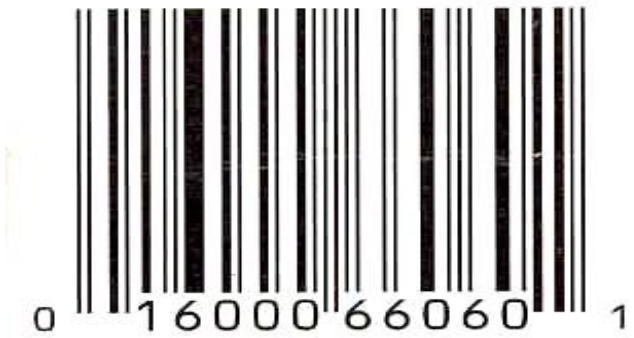
image-net.org

Splendors of benchmark data sets

NIST Special Database 8

NIST Machine-Print Database of Gray Scale and Binary Images (MPDB)

Rate our Products and Services



tesseract-ocr

An OCR Engine that was developed at HP Labs between 1985 and 1995... and now at Google.

The map is not the territory…

But what about recognition of real objects whose appearance can exhibit large variations?

Are benchmark data sets representative of the real-world problem ?

«Under that Empire the art of map-making reached such perfection that the map covering only one district took up a whole city, and the map of the Empire took up a whole district. After some time these outsized maps were no longer sufficient and the School of Cartography fabricated a map of the Empire that was the size of the Empire and coincided with this.»

Jorge Luis Borges, *On Rigor in Science*

Miseries of benchmark data sets

…So, what is the value of current data sets when used to train algorithms for object recognition that will be deployed in the real world?

Antonio Torralba, Alexei A. Efros, CVPR 2011

The curse of «biased» data sets

…So, what is the value of current data sets when used to train algorithms for object recognition that will be deployed in the real world? …

…The answer that emerges can be summarized as: "better than nothing, but not by much"…

Antonio Torralba, Alexei A. Efros, CVPR 2011

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

Beery, Sara et al. "Recognition in terra incognita.", ECCV 2018

2009-2010: Bigger is better?

Maybe current data sets are not large enough to represent well problems of the real world?

Should we make them bigger?

No, bigger is not better…

Estimate No. 1: The number of meaningful/valid images on a 1200 by 1200 display is at least as high as 10^{400}.

Estimate No. 2: 10^{25} (greater than a trillion squared) is a very conservative lower bound to the number of all possible discernible images.

«These numbers suggest that it is impractical to construct training or testing sets of images that are dense in the set of all images unless the class of images is restricted.»

Theo Pavlidis

The Number of All Possible Meaningful or Discernible Pictures (2009)

Yes! Bigger can be better. The unreasonable effectiveness of data

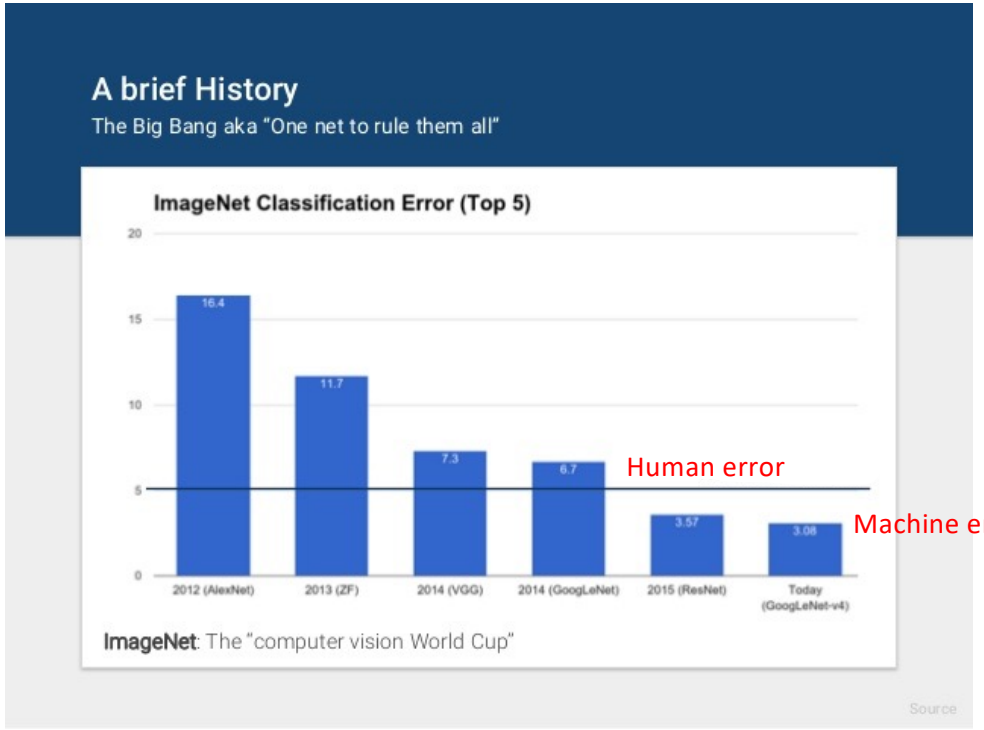
*«Perhaps when it comes to **natural language processing** and related fields….we should stop acting as if our goal is to author extremely elegant theories, and instead embrace complexity and make use of the best ally we have: the **unreasonable effectiveness of data**»*

Alon Halevy, Peter Norvig, and Fernando Pereira
The unreasonable effetiveness of data
IEEE Intelligent Systems 2009

In 2009, this was the intuition behind the future success of ChatGPT

Language *(the next word in a sentence)* can be predicted by learning from huge amount of examples. We don't need an explicit and formal model of the human language, we can learn a «large language model» from examples

The bright side of AI: super human performances…



ImageNet Challenge

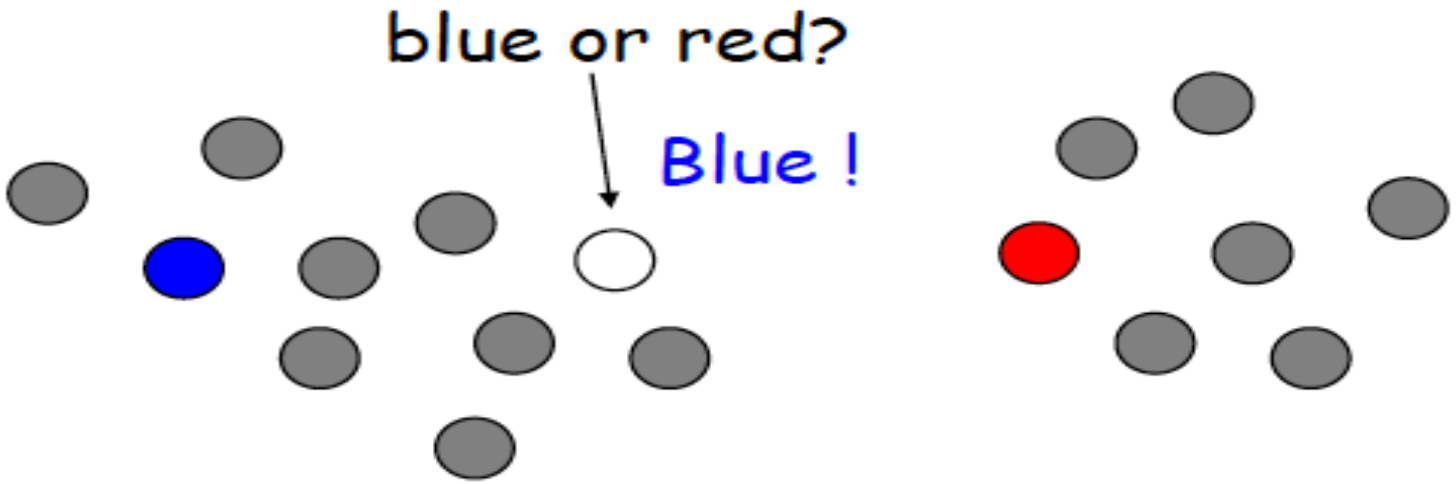
IMAGENET

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.

High accuracy is high robusteness?

- What about the performance of machine-learning algorithms on data which has been modified very slightly ?

The smoothness assumption



Points close to each other are more likely to share a label

This is generally assumed in machine learning…

High accuracy is high robusteness?

How a machine-learning algorithm should classify these two images?

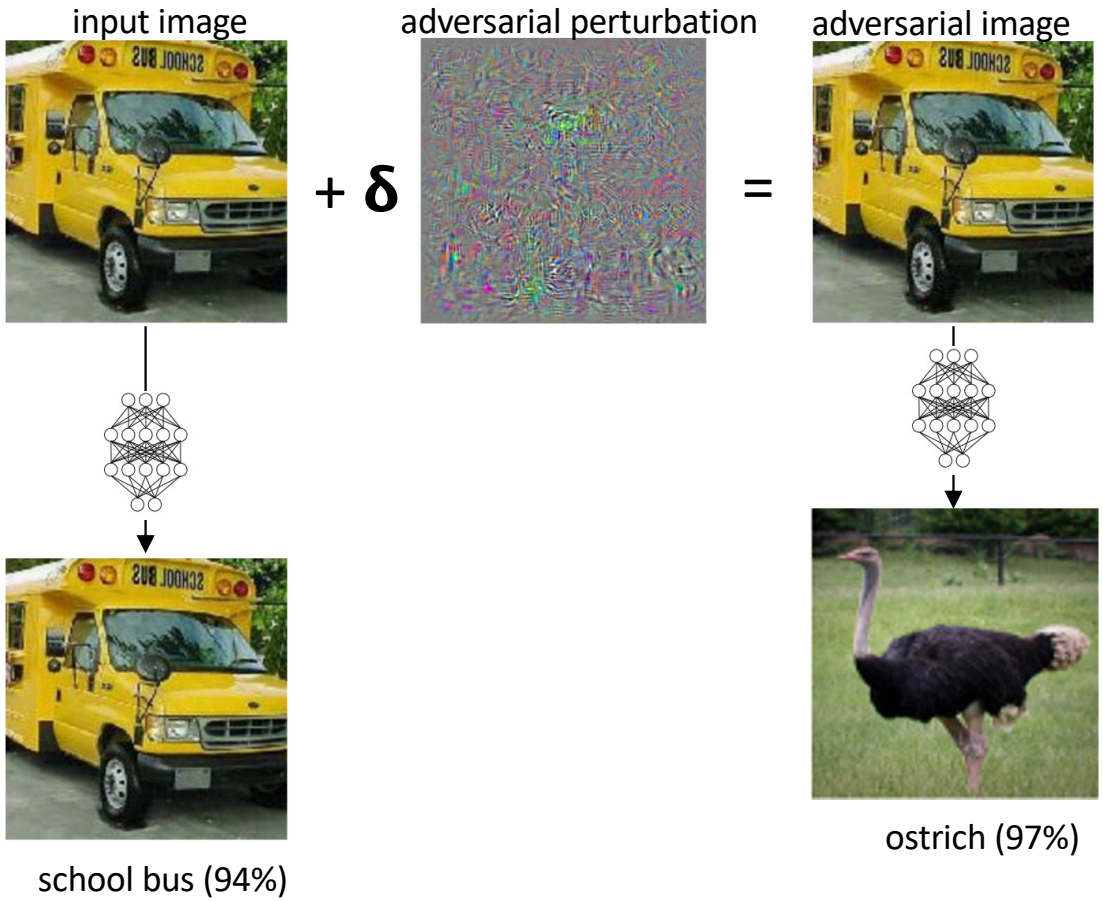
With the same or a different classification label?

2013-2014: Unknown unknowns in machine learning

$Minimize\ \|\boldsymbol{\delta}\|$

so that $f(x+\boldsymbol{\delta})=l$

The adversarial image $x+\boldsymbol{\delta}$ is visually hard to distinguish from x
Informally speaking, the solution $x+\boldsymbol{\delta}$ is the closest image to x classified as l=*ostrich*

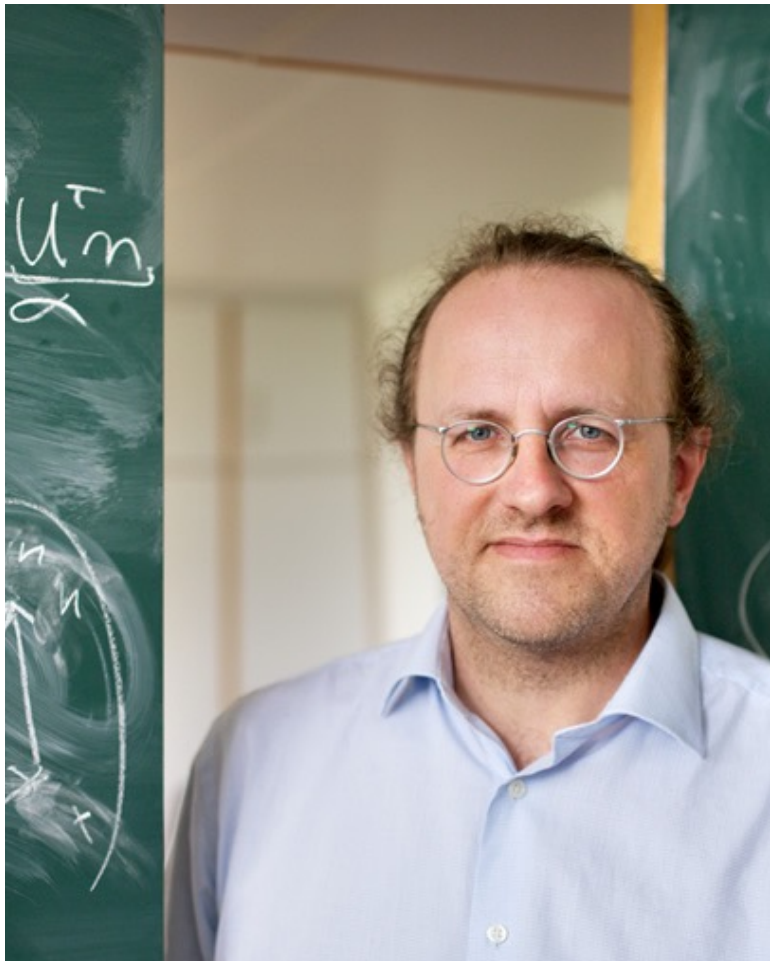


Biggio, Roli et al., Evasion attacks against machine learning at test time, **ECML-PKDD 2013**
Szegedy et al., Intriguing properties of neural networks, **ICLR 2014**

Why is ML so fragile?

The i.i.d. assumption

- **Underlying assumption:** past data is *representative* of future data (IID data)
- The success of modern AI is on tasks for which we collected enough representative training data
- **Adversarial attacks** point exactly at this lack of robustness which comes from IID specialization

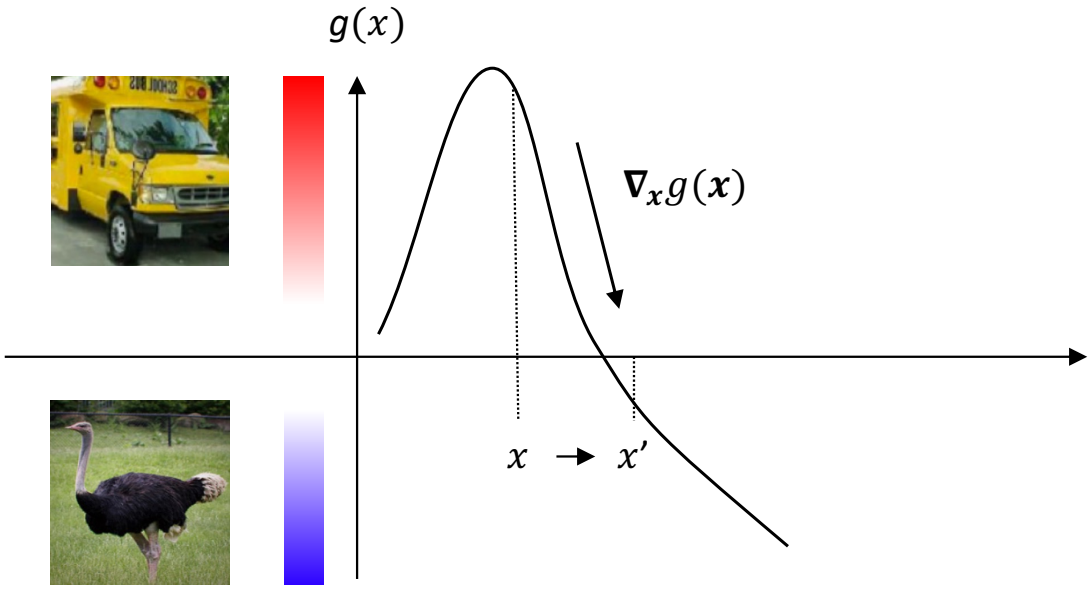


Bernhard Schölkopf

Director, Max Planck Institute, Tuebingen, Germany

Large gradient

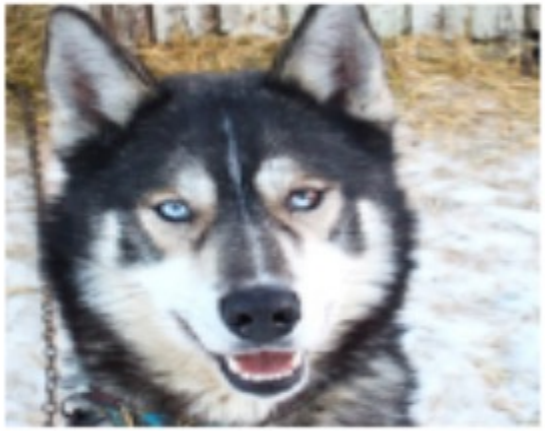
- Large sensitivity of g(**x**) to input changes
  - i.e., the **input gradient** $\nabla_x g(\boldsymbol{x})$ has a large norm (scales with input dimensions!)
  - Thus, even small modifications along that direction will cause large changes in the predictions



[Simon-Gabriel et al., Adversarial vulnerability of NNs increases with input dimension, 2018]

What does machine learning learn?

[M.T. Ribeiro et al., KDD 2016]

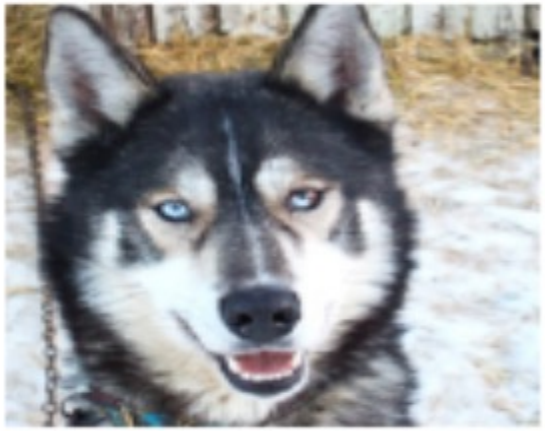


(a) Husky classified as wolf

if pixels are our input features, which pixels are more correlated with label=wolf ?

Biased data sets and spurious correlations

[M.T. Ribeiro et al., KDD 2016]



(a) Husky classified as wolf

if pixels are our input features, which pixels are more correlated with label=wolf ?

(b) Explanation

The issue of bias is amplified in LLMs: Hallucinations

- LLM's training data can contain several types of biases
  - **Imitative Falsehood**: incorrect facts tha appear multiple times
  - **Duplication Bias**: repetitions of the same statements/facts
  - **Social Bias**: facts/data related to social features (e.g., gender)

Type	Premise	Input	Model Ouput	Explanation
Imitative Falsehood	Within the LLM's training data, a frequently referenced but incorrect fact appears multiple times: *Thomas Edison invented the light bulb.*	Who is credited with the invention of the light bulb?	Thomas Edison is credited with the invention of the light bulb.	**While Edison improved the light bulb's design, he was not its sole inventor.** The model's answer reflects a misconception prevalent in its training data, exemplifying an imitative falsehood.
Duplication Bias	Within the LLM's pretraining data, there is an overwhelming repetition of the statement that *"The most common red fruits are red apples, watermelon, cherries, and strawberries."*	Please list some red fruits, excluding apples.	Red fruits are red apples, watermelon, cherries, and strawberries.	**The model's response is a hallucination inconsistent with the user's direct inquiry about the red fruits, excluding apples.** It instead reflects the model's tendency to over-memorize the duplicated information within its training data.
Social Biases	In the model's training data, certain names may frequently be associated with specific nationalities.	Provide a summary of the following news: Dr. Kim recently published a breakthrough paper on quantum computing. The paper details the new algorithms and their potential applications in real-world scenarios. It has been hailed as revolutionary by experts in the field.	Dr. Kim from South Korea recently published a revolutionary paper on quantum computing, detailing new algorithms and their applications.	**The model erroneously added "South Korea" based on the name "Kim"** revealing its bias in linking certain names to specific nationalities, which can cause hallucinations in the summary.

Lei Huang et a., A Survey on Hallucination in Large Language Models, 2023

A look back...

Nello Cristianini

The road to artificial intelligence: A case of data over theory

(New Scientist, 2016)

The "fathers" of artificial intelligence in the 1950s thought that a machine could reproduce human intelligence with search algorithms and symbolic reasoning.

Today we know that they were wrong!

The "fathers" of artificial intelligence in the 1950s thought that building intelligent machines would give us a better understanding of how the human brain works.

Today we know that they were wrong!

Today, the behavior of our computers may appear intelligent. But in fact, these are statistical algorithms that discover patterns, correlations. Without understanding the causes... It is, however, intelligence, in the behavioral sense of Turing! It is, however, a great scientific and technological advance!

Which way we have taken towards modern AI?

- Why we left the road of the fathers of AI?
- Because it didn't work…
- But which way we have taken?

The shortcut

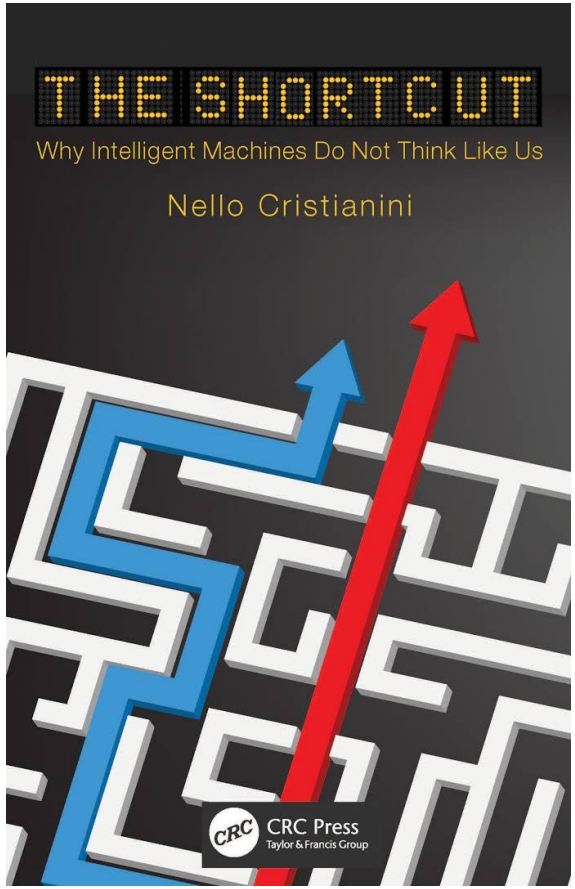
The "fathers" of artificial intelligence in the 1950s thought that the **direct way** to AI should be based on a clear understanding of the complex phenomena that we wanted the machines to emulate

We have taken a **shortcut** by removing the goal to actually understand the complex phenomena that we wanted the machines to emulate, such as language.

We decided to design intelligent machines that rely on making decisions based on statistical patterns found in data.

This removed the need to actually understand the complex phenomena that we wanted the machines to emulate, such as language.

The shortcut, N. Cristianini, 2023



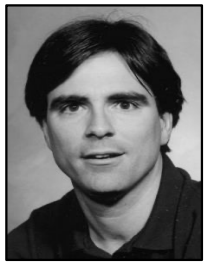
Today AI (ChatGPT…) is intelligent according to the Turing's test…

- But we know that we have exploited a shortcut to get here…
- High accuracy does not always imply robustness, and correlation does not imply causation…
- This means that AI can exhbit a lack of robustness and several risks…

Alan Turing, Computing Machinery & Intelligence, 1950

Thanks for Listening!

Any questions?



Engineering isn't about perfect solutions; it's about doing the best you can with limited resources
(Randy Pausch, 1960-2008)