Opacity, Neutrality, Stupidity

Three Challenges for Artificial Intelligence

Marcello Pelillo
European Centre for Living Technology
University of Venice, Italy

eclt

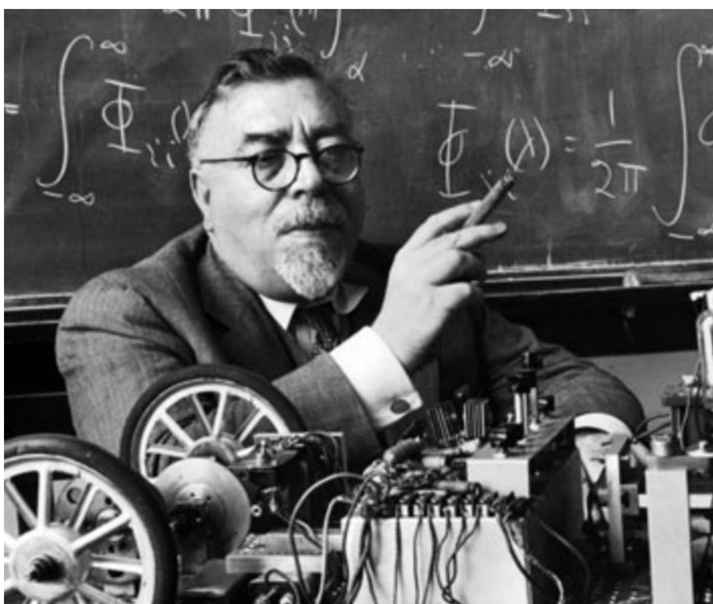
Artificial Intelligence in Industry and Finance, Winterthur, September 6, 2018

Wiener's lesson

«Any machine constructed for the purpose of making decisions, if it does not possess the power of learning, will be completely literal-minded.

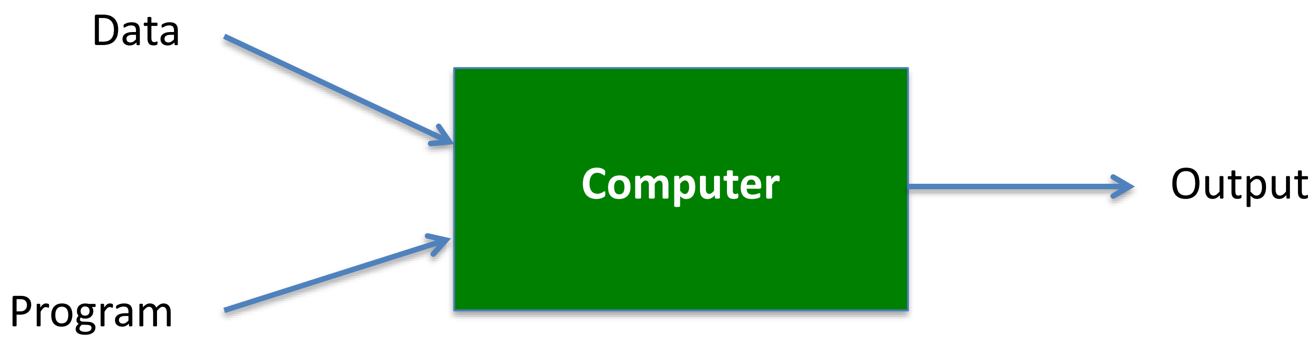
Woe to us if we let it decide our conduct, unless we have previously examined its laws of action, and know fully that its conduct will be carried out on principles acceptable to us!»

Norbert Wiener
The Human Use of Human Beings (1950)

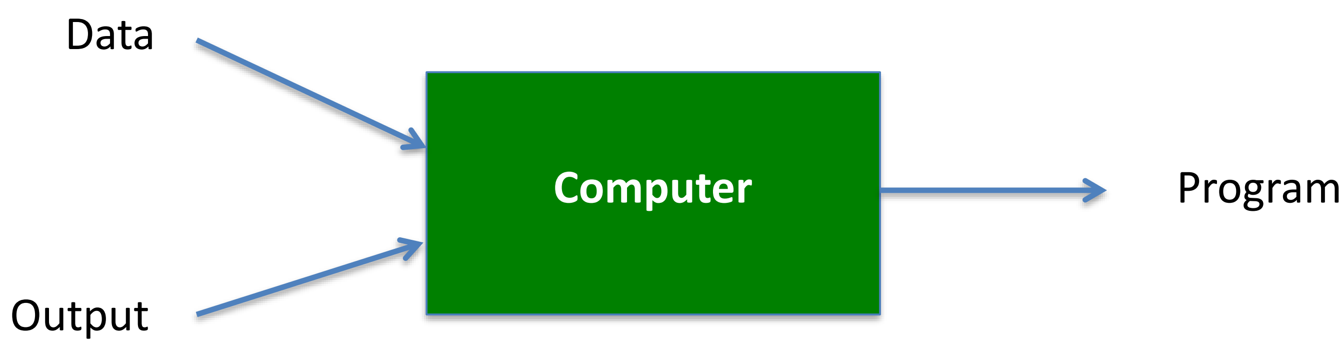


Machines that learn?

Traditional programming

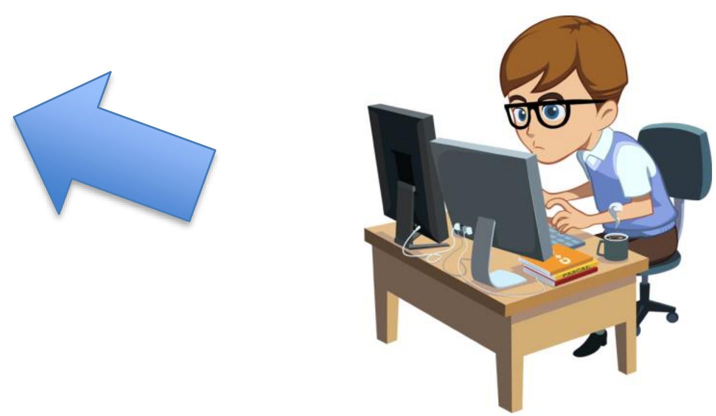


Machine learning

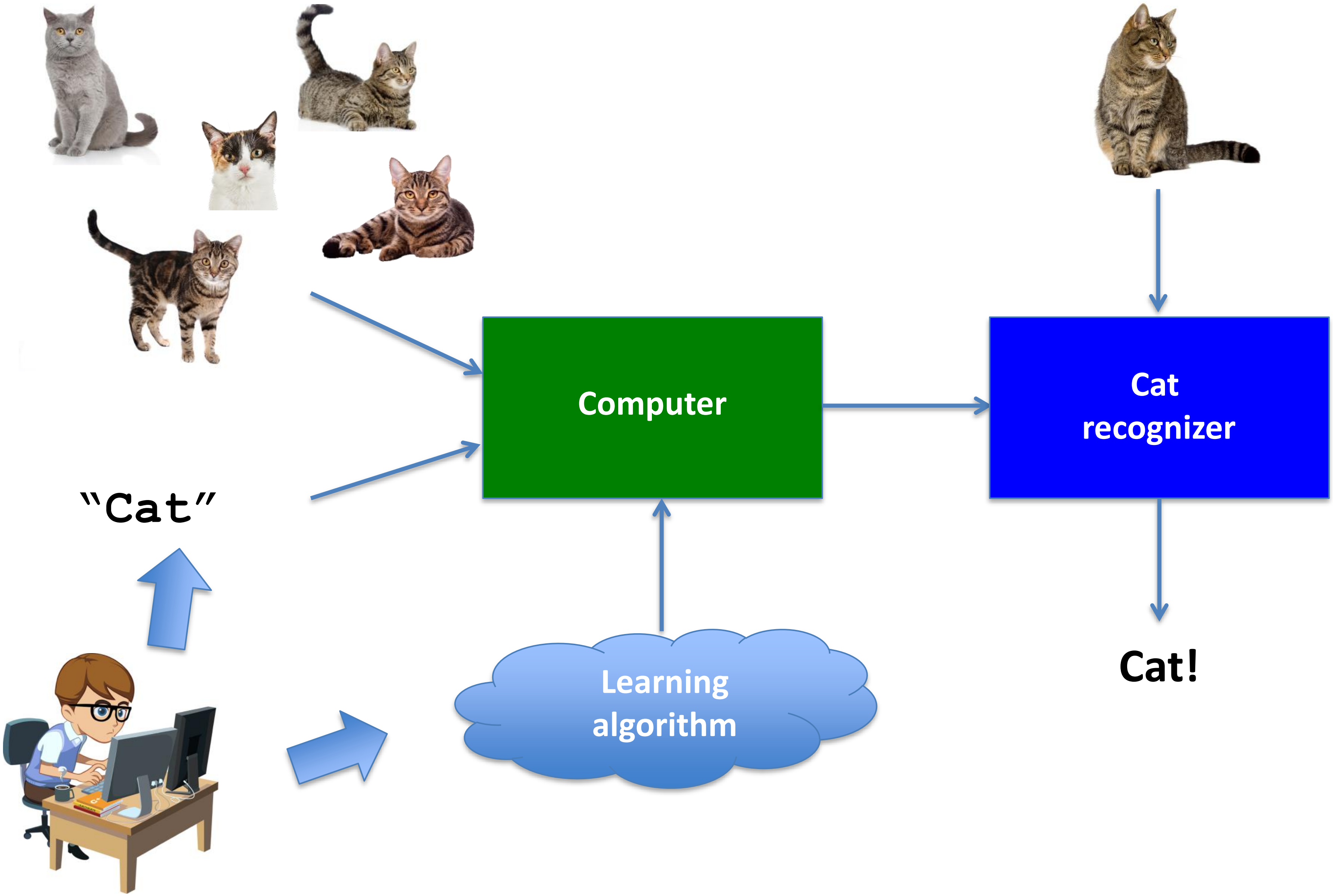


Traditional programming


```
if  (eyes == 2) &
    (legs == 4) &
    (tail == 1 ) &
    …
then Print “Cat!”
```



Machine learning



The philosophy of machine learning

«This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity.

With enough data, the numbers speak for themselves.»

Chris Anderson

The end of theory (Wired, 2008)

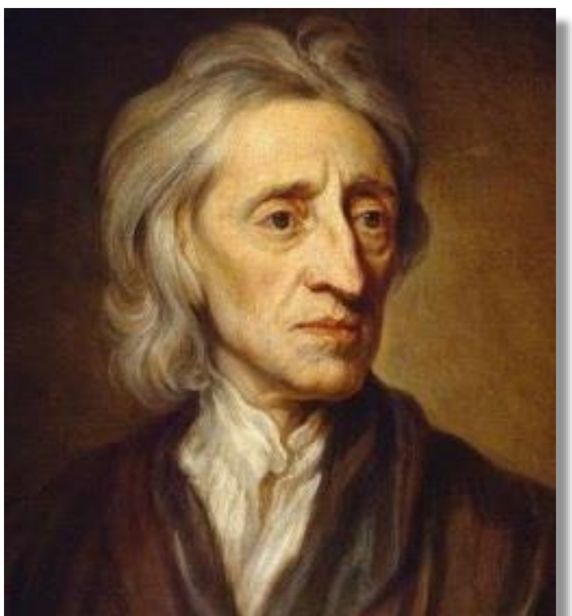
Back to *tabula rasa*

«Let us then suppose the mind to be, as we say, white paper void of all characters, without any ideas. How comes it to be furnished? Whence comes it by that vast store which the busy and boundless fancy of man has painted on it with an almost endless variety? Whence has it all the materials of reason and knowledge?

To this I answer, in one word, from EXPERIENCE.

In that all our knowledge is founded;

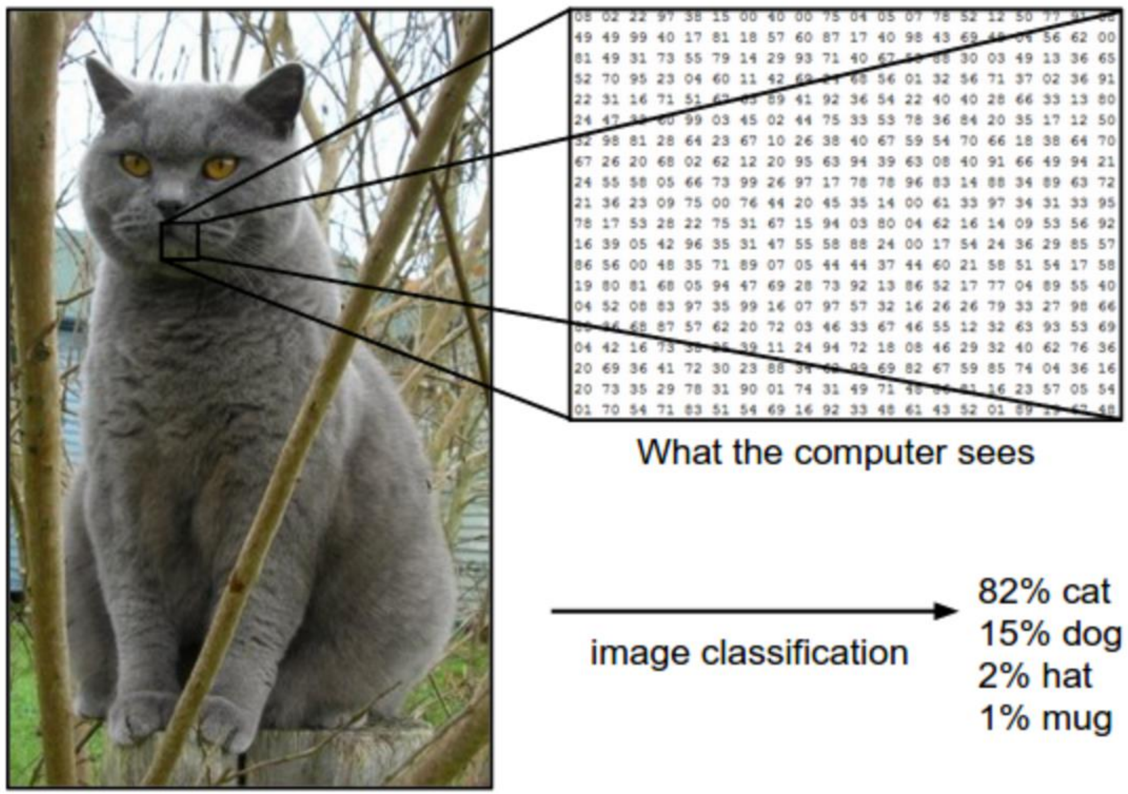
and from that it ultimately derives itself.»



John Locke

An Essay Concerning Human Understanding (1690)

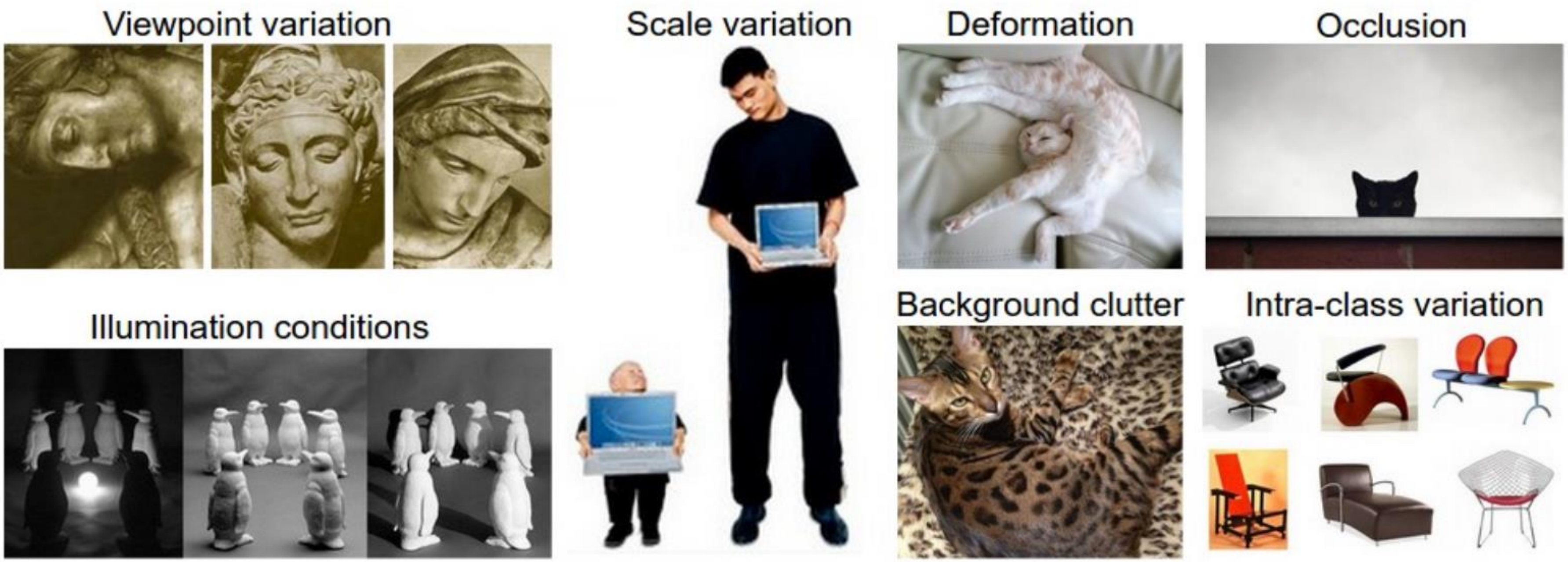
A success story: Image classification



Predict a single label (or a distribution over labels as shown here to indicate our confidence) for a given image. Images are 3-dimensional arrays of integers from 0 to 255, of size Width x Height x 3. The 3 represents the three color channels Red, Green, Blue.

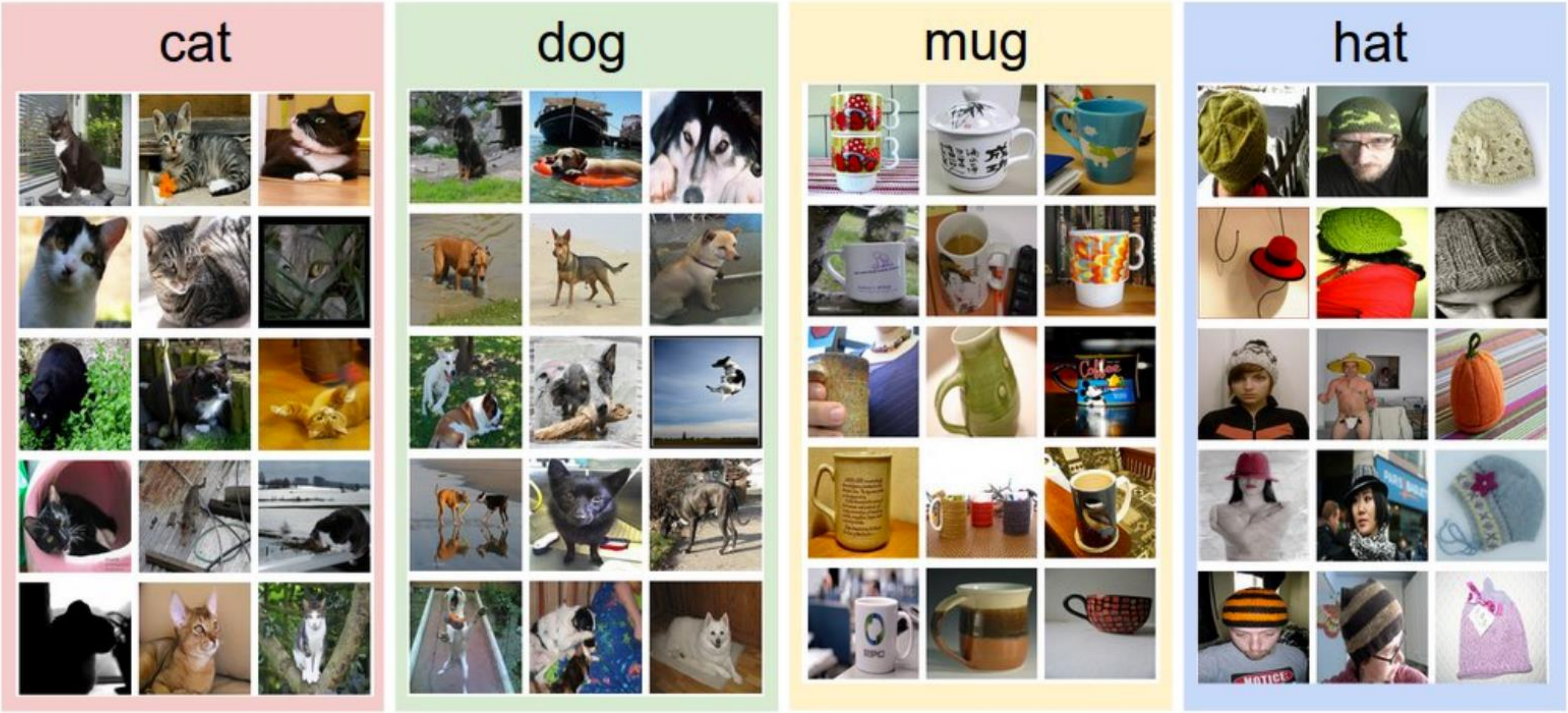
From: A. Karpathy

A challenging problem



From: A. Karpathy

The data-driven approach



An example training set for four visual categories.

In practice we may have thousands of categories and hundreds of thousands of images for each category.

From: A. Karpathy

The age of "deep learning"

News & Analysis

Microsoft, Google Beat Humans at Image Recognition

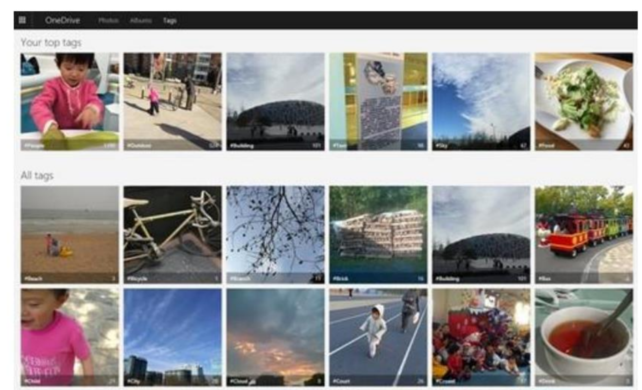
Deep learning algorithms compete at ImageNet challenge

R. Colin Johnson
2/18/2015 08:15 AM EST
14 comments

NO RATINGS
1 saves
LOGIN TO RATE

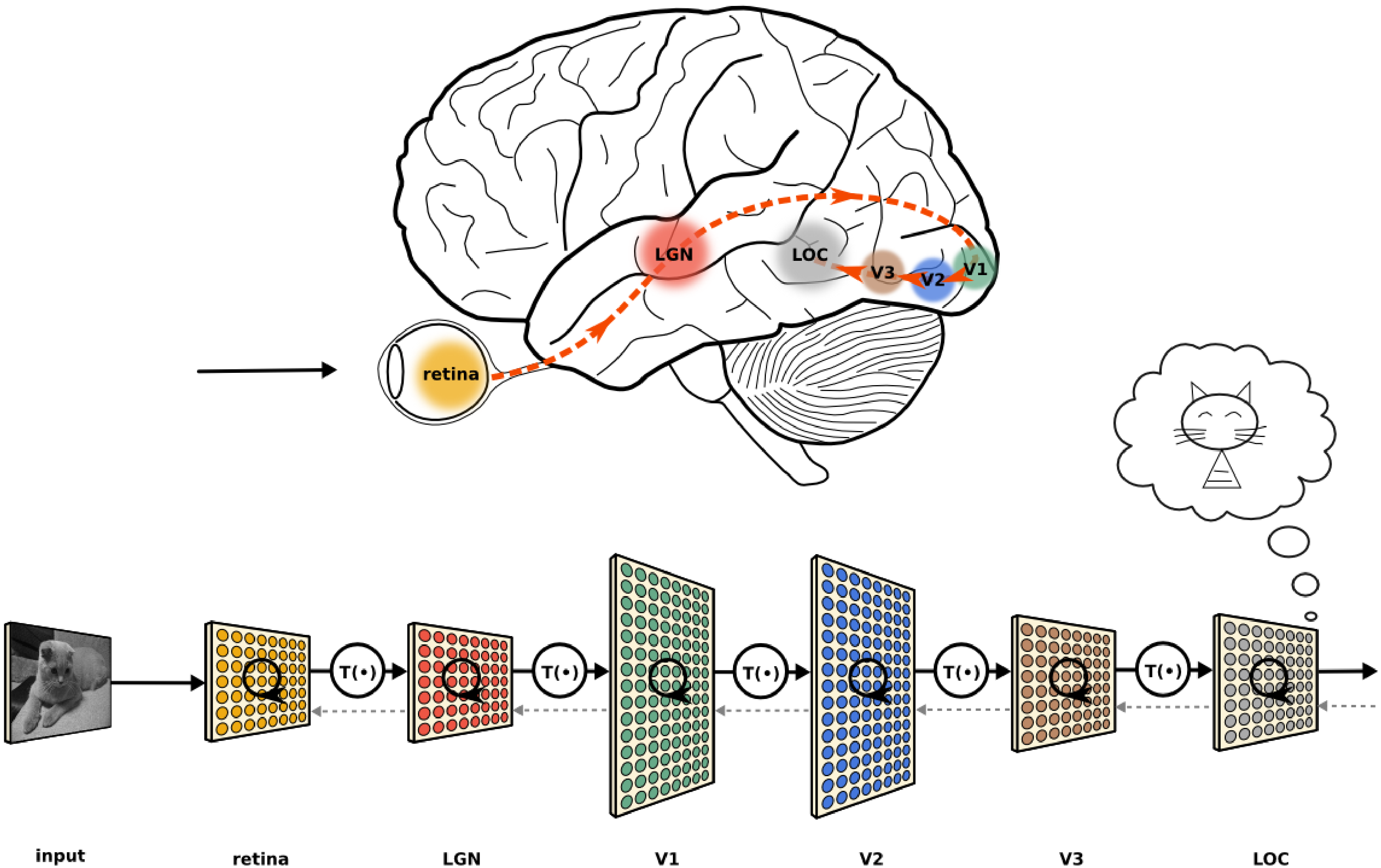
PORTLAND, Ore. -- First computers beat the best of us at chess, then poker, and finally Jeopardy. The next hurdle is image recognition — surely a computer can't do that as well as a human. Check that one off the list, too. Now Microsoft has programmed the first computer to beat the humans at image recognition.

The competition is fierce, with the ImageNet Large Scale Visual Recognition Challenge doing the judging for the 2015 championship on December 17. Between now and then expect to see a stream of papers claiming they have one-upped humans too. For instance, only 5 days after Microsoft announced it had beat the human benchmark of 5.1% errors with a 4.94% error grabbing neural network, Google announced it had one-upped Microsoft by 0.04%.

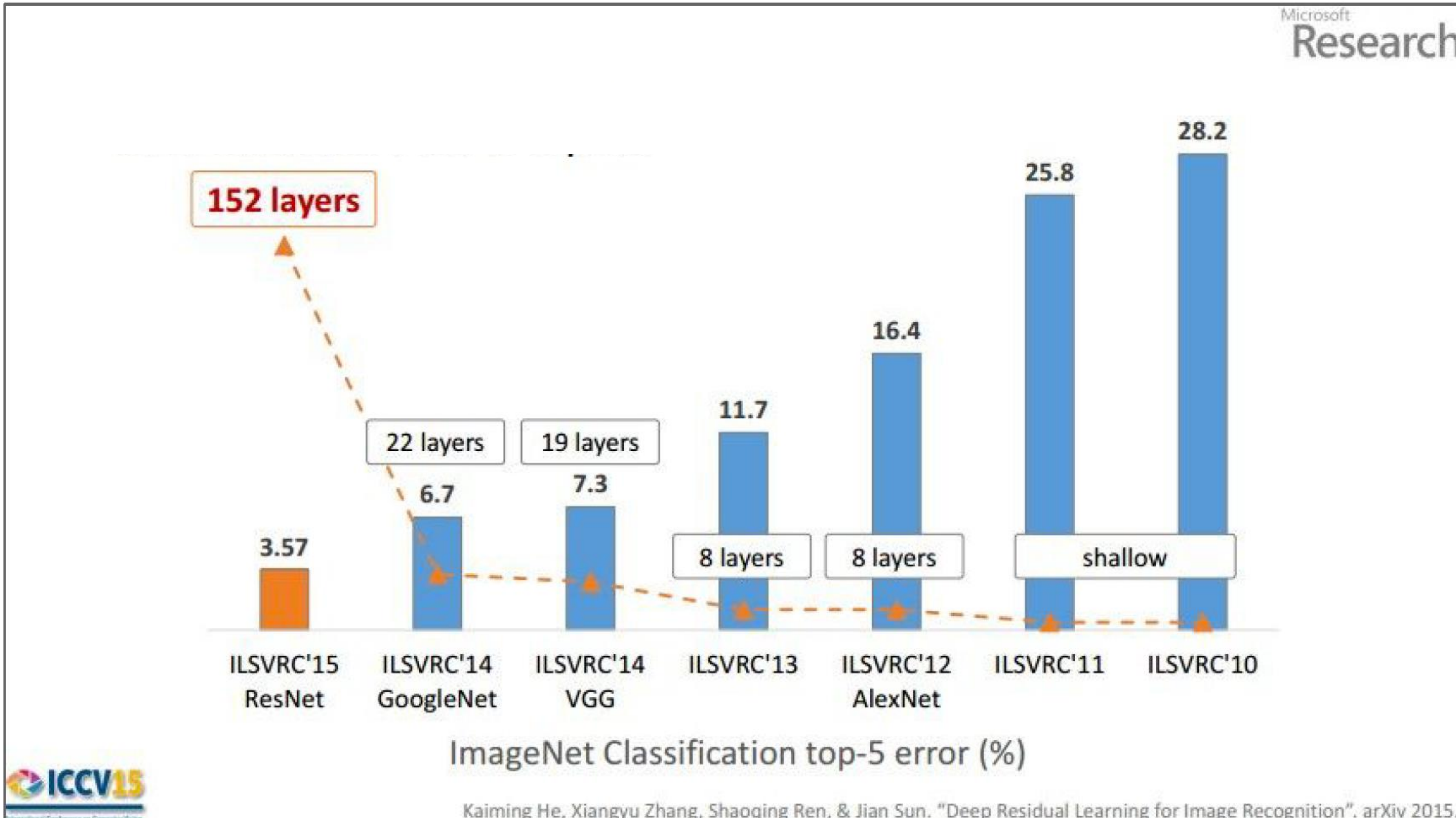


The top row is a representative of the categories that Microsoft's algorithm found in the database and the image columns below are examples that fit. (Source: Microsoft)

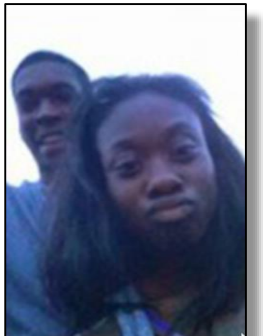
Inspiration from biology

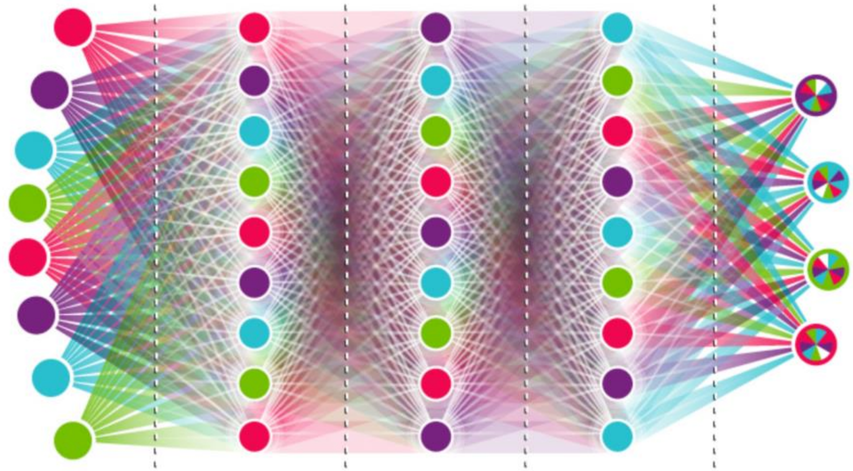


A question of layers



Opacity





Gorilla!

BBC Sign in News More

NEWS

Home | UK | World | Business | Politics | Tech | Science | Health | Family & Education

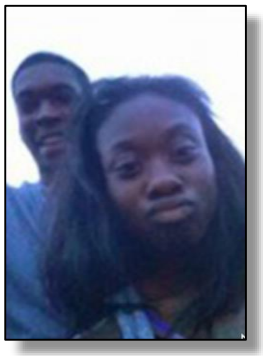
Technology

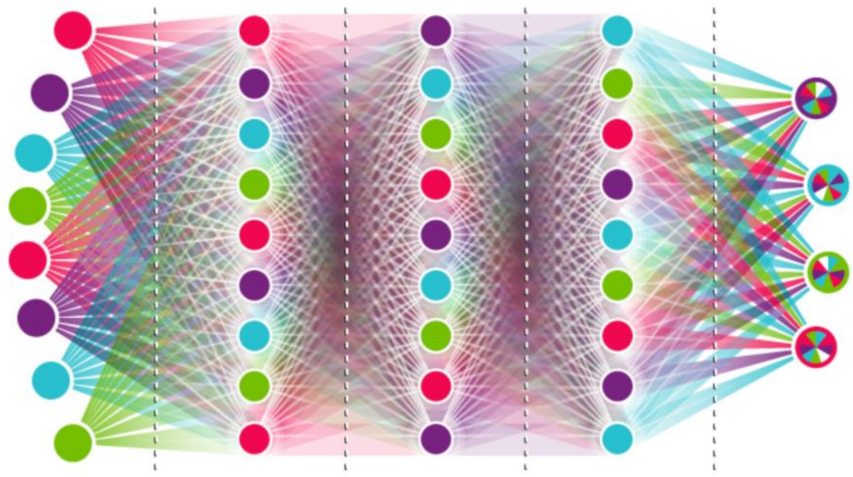
Google apologises for Photos app's racist blunder

1 July 2015

Share

Debugging?





Gorilla!

Hmm… maybe it's the weight on the connection between unit 13654 and 26853 ???

After three years …

WIRED

TOM SIMONITE BUSINESS 01.11.18 07:00 AM

WHEN IT COMES TO GORILLAS, GOOGLE PHOTOS REMAINS BLIND

Towards more frightening scenarios

The New York Times

POLITICS

Sent to Prison by a Software Program's Secret Algorithms

Sidebar

By ADAM LIPTAK MAY 1, 2017

Eric L. Loomis

You're identified, through the COMPAS assessment, as an individual who is at high risk to the community.

Accuracy *vs* transparency

«Deploying unintelligible black-box machine learned models is risky – high accuracy on a test set is NOT sufficient. Unfortunately, the most accurate models usually are not very intelligible (e.g., random forests, boosted trees, and neural nets), and the most intelligible models usually are less accurate (e.g., linear or logistic regression).»

Rich Caruana

Friends don't let friends deploy models they don't understand (2016)

2016 Workshop on Human Interpretability in Machine Learning

WHI 2016 @ ICML, New York, June 23, 2016

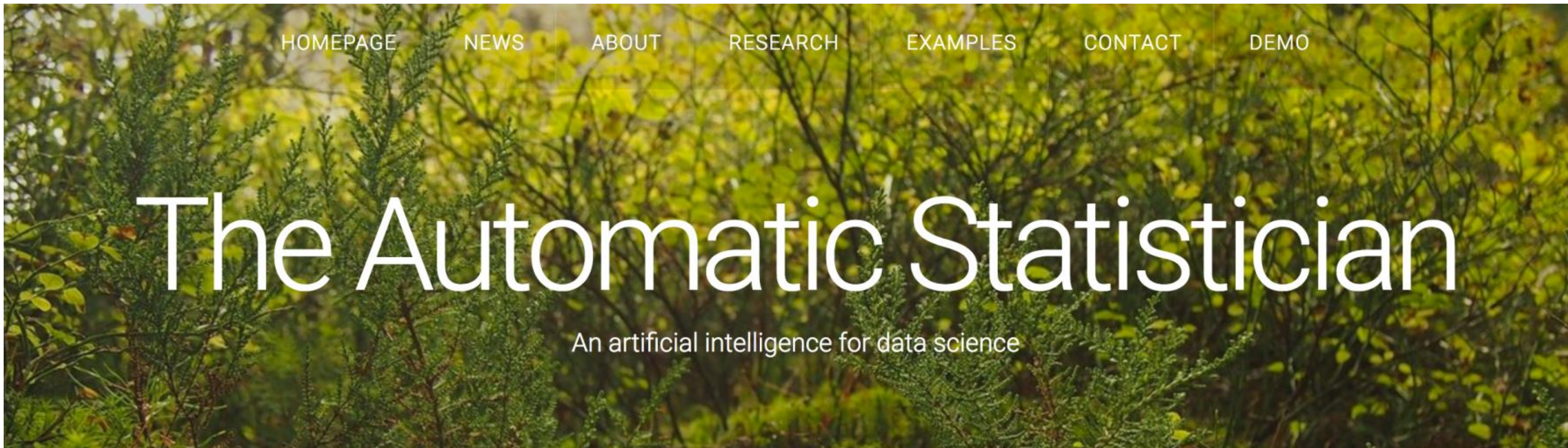
Back to the 1980's

«The results of computer induction should be symbolic descriptions of given entities, semantically and structurally similar to those a human expert might produce observing the same entities.

Components of these descriptions should be comprehensible as single 'chunks' of information, directly **interpretable in natural language**, and should relate quantitative and qualitative concepts in an integrated fashion.»

Ryszard S. Michalski
A theory and methodology of inductive learning (1983)

The "automatic statistician"



«The aim is to find models which have both good predictive performance,
and are somewhat interpretable.
The Automatic Statistician generates a natural language summary of the analysis, producing a 10-15 page report with plots and tables describing the analysis.»

Zoubin Ghahramani (2016)

But why should we care?

«There are things we cannot verbalize.
When you ask a medical doctor why he diagnosed
this or this, he's going to give you some reasons.
But how come it takes 20 years to make a good doctor?
Because the information is just not in books.»

Stéphane Mallat (2016)

«You use your brain all the time; you trust your brain all the time; and you have no idea how your brain works.»

Pierre Baldi (2016)

From: D. Castelvecchi, Can we open the black box of AI? *Nature* (October 5, 2016)

Indeed, sometimes we should …

Explanation is a core aspect of due process (Strandburg, HUML 2016):

- ✓ Judges generally provide either written or oral explanations of their decisions
- ✓ Administrative rule-making requires that agencies respond to comments on proposed rules
- ✓ Agency adjudicators must provide reasons for their decision to facilitate judicial review

Example #1. In many countries, banks that deny a loan have a legal obligation to say why — something a deep-learning algorithm might not be able to do.

Example #2. If something were to go wrong as a result of setting the UK interest rates, the Bank of England can't say: "the black box made me do it".

From: D. Castelvecchi, Can we open the black box of AI? *Nature* (October 5, 2016)

A right to explanation?

General Data Protection Regulation

Art. 13

A data subject has the right to obtain
“meaningful information about the logic involved”

Pedro Domingos
@pmddomingos

Starting May 25, the European Union will require algorithms to explain their output, making deep learning illegal.

5:59 AM - Jan 29, 2018

344 247 people are talking about this

ARTICLE

International Data Privacy Law, 2017, Vol. 7, No. 2

Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation

Sandra Wachter*, Brent Mittelstadt** and Luciano Floridi***

Neutrality?

Kranzberg's First Law of Technology
Technology is neither good nor bad; nor is it neutral.

Machine Bias

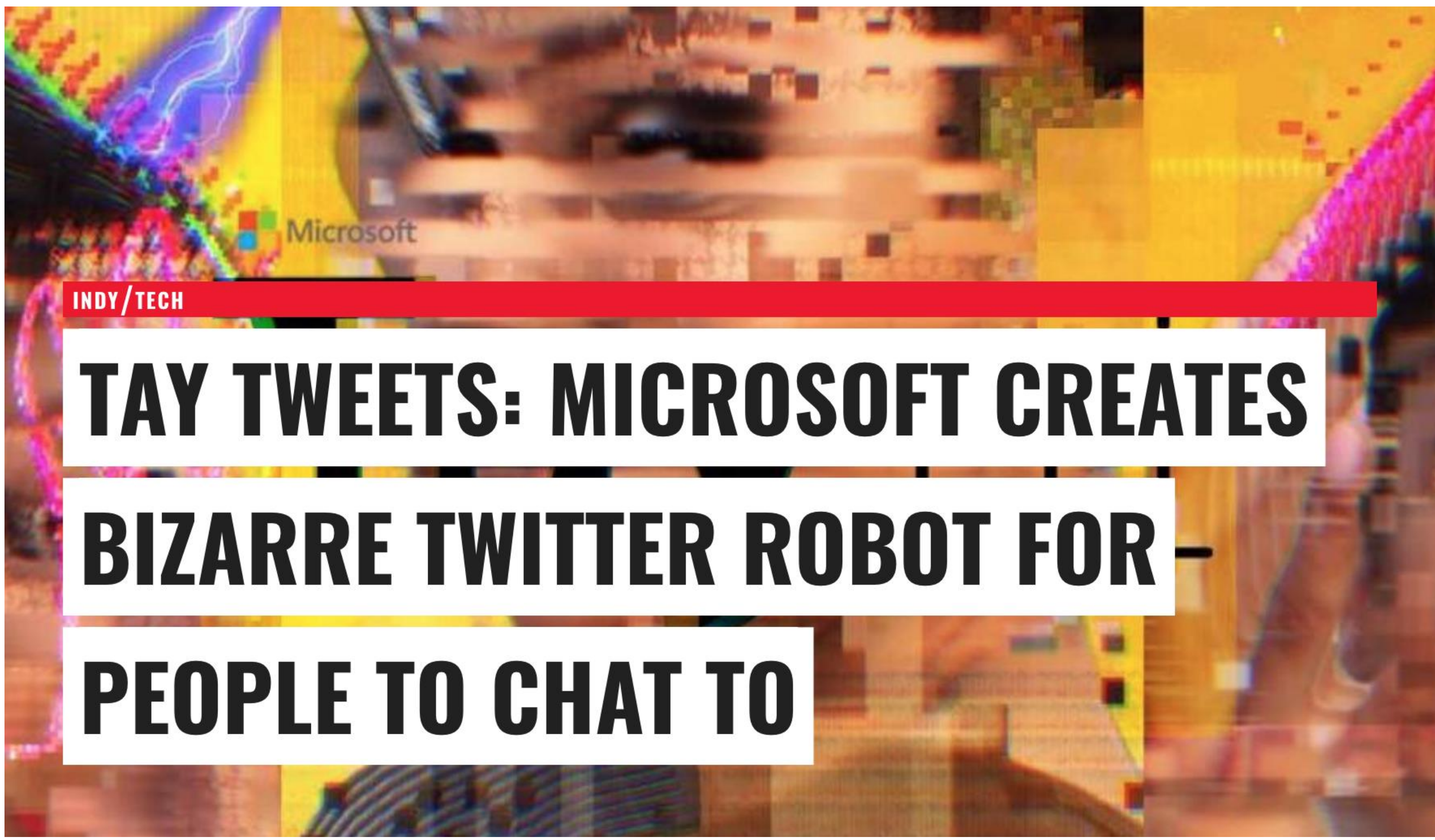
There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

	White	African American
Labeled Higher Risk, But Didn't Re-Offend	23,5%	**44,9%**
Labeled Lower Risk, Yet Did Re-Offend	**47,7%**	28,0%

March 23, 2016

INDEPENDENT



A few hours later ...

INDEPENDENT

24 March 2016

The (well-known) question of bias

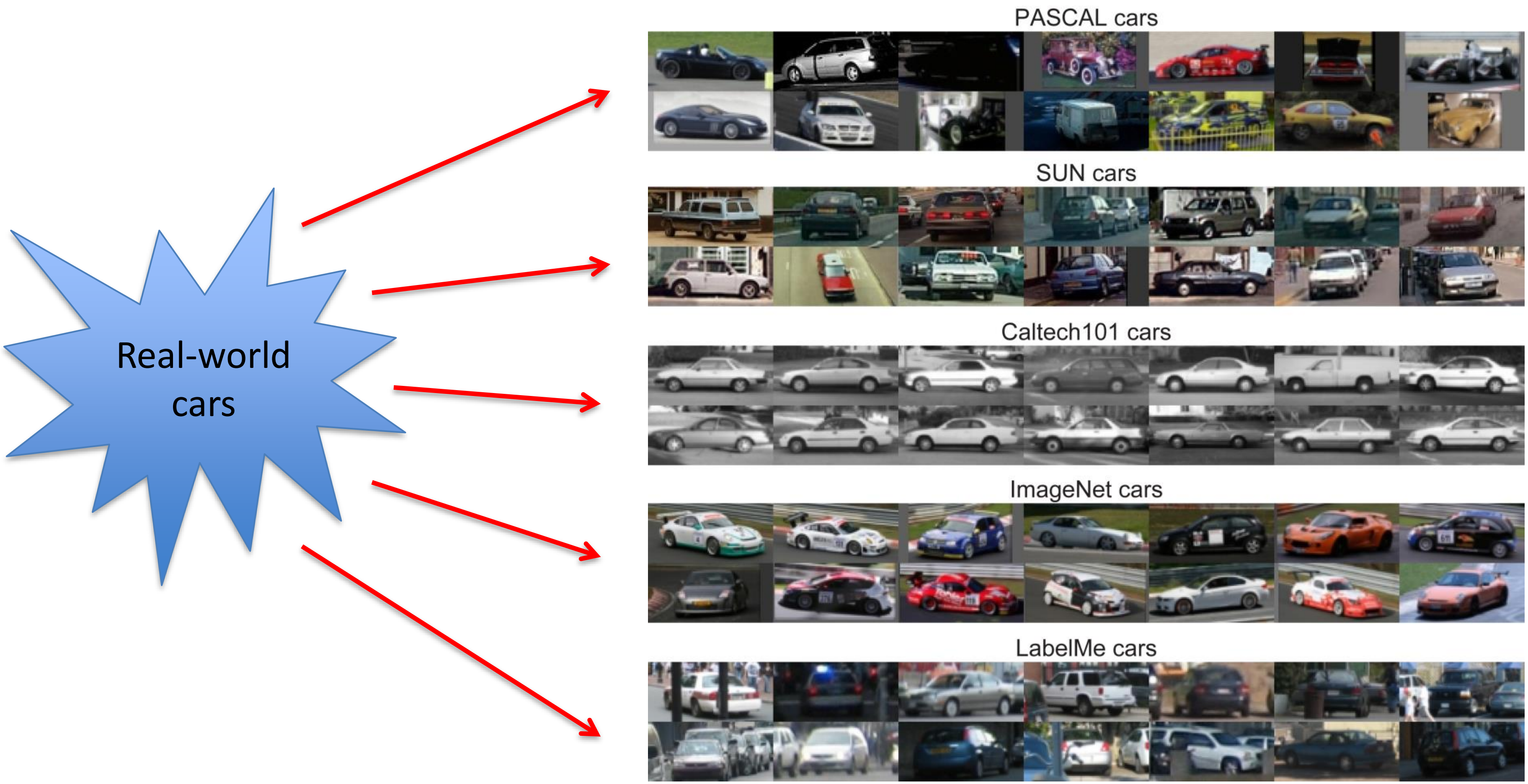
«So, what is the value of current datasets when used to train algorithms for object recognition that will be deployed in the real world?

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

Antonio Torralba and Alexei Efros

Unbiased look at dataset bias (2011)

The map is not the territory



The curse of biased datasets

«We would like to ask the following question: how well does a typical object detector trained on one dataset generalize when tested on a representative set of other datasets, compared with its performances on the "native" test set?»

A. Torralba and A. Efros (2011)

task	Test on: / Train on:	SUN09	LabelMe	PASCAL	ImageNet	Caltech101	MSRC	Self	Mean others	Percent drop
"car" classification	SUN09	**28.2**	29.5	16.3	14.6	16.9	21.9	28.2	19.8	**30%**
	LabelMe	14.7	**34.0**	16.7	22.9	43.6	24.5	34.0	24.5	**28%**
	PASCAL	10.1	25.5	**35.2**	43.9	44.2	39.4	35.2	32.6	**7%**
	ImageNet	11.4	29.6	36.0	**57.4**	52.3	42.7	57.4	34.4	**40%**
	Caltech101	7.5	31.1	19.5	33.1	**96.9**	42.1	96.9	26.7	**73%**
	MSRC	9.3	27.0	24.9	32.6	40.3	**68.4**	68.4	26.8	**61%**
	Mean others	10.6	28.5	22.7	29.4	39.4	34.1	53.4	27.5	48%
"person" classification	SUN09	**16.1**	11.8	14.0	7.9	6.8	23.5	16.1	12.8	**20%**
	LabelMe	11.0	**26.6**	7.5	6.3	8.4	24.3	26.6	11.5	**57%**
	PASCAL	11.9	11.1	**20.7**	13.6	48.3	50.5	20.7	27.1	**-31%**
	ImageNet	8.9	11.1	11.8	**20.7**	76.7	61.0	20.7	33.9	**-63%**
	Caltech101	7.6	11.8	17.3	22.5	**99.6**	65.8	99.6	25.0	**75%**
	MSRC	9.4	15.5	15.3	15.3	93.4	**78.4**	78.4	29.8	**62%**
	Mean others	9.8	12.3	13.2	13.1	46.7	45.0	43.7	23.4	**47%**

Too big to fail?

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)

The illusion of progress

«An apparent superiority in classification accuracy, obtained in "laboratory conditions," may not translate to a superiority in real-world conditions and, in particular, the apparent superiority of highly sophisticated methods may be illusory, with simple methods often being equally effective or even superior.»

David J. Hand

Classifier Technology and the Illusion of Progress (2006)

Belief in the "law of small numbers"

«People's intuitions about random sampling appear to satisfy the law of small numbers, which asserts that the law of large numbers applies to small numbers as well.»

Amos Tversky and Daniel Kahneman
Belief in the Law of Small Numbers (1971)

Belief in the "law of small numbers"

The believer in the law of small numbers practices science as follows:

1 He gambles his hypotheses on small samples without realizing that the odds against him are unreasonably high. **He overestimates power.**

2 He has undue confidence in early trends and in the stability of observed patterns. **He overestimates significance.**

3 In evaluating replications, he has unreasonably high expectations about the replicability of significant results. **He underestimates the breadth of confidence intervals.**

4 He rarely attributes a deviation of results from expectations to sampling variability, because he finds a causal "explanation" for any discrepancy. Thus, **he has little opportunity to recognize sampling variation in action**.

His belief in the law of small numbers, therefore, will forever remain intact.

From: A. Tversky and D. Kahneman, *Belief in the Law of Small Numbers* (1971)

Bias and social justice

But ML is increasingly being used in several "social" domains:

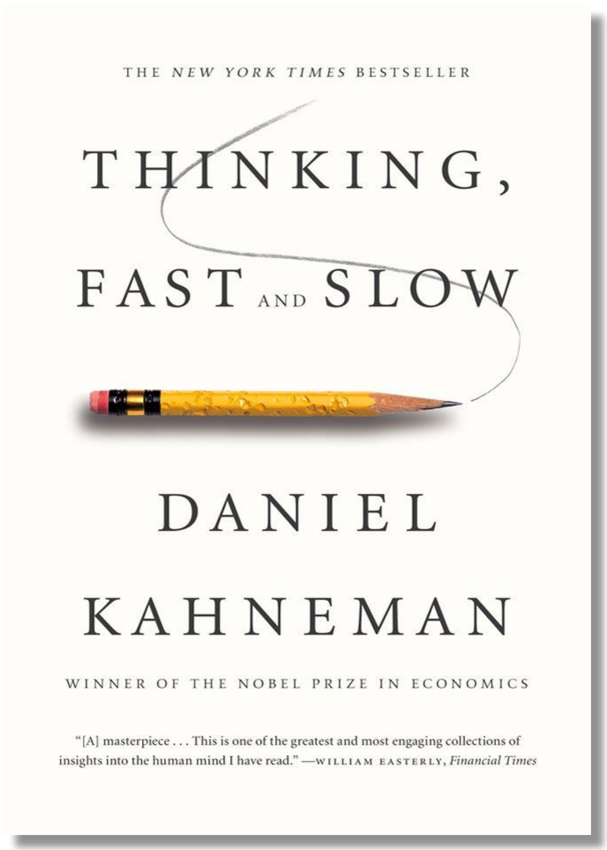
- Recruiting: Screening job applications
- Banking: Credit ratings / loan approvals
- Judiciary: Recidivism risk assessments
- Journalism: News recommender systems
- …

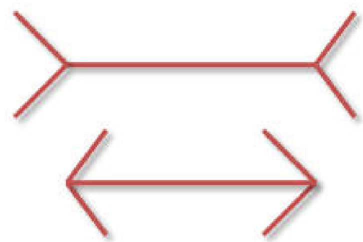
Sources of potential social discrimination:

- Social biases of people collecting the training sets
- Sample size disparity
- Feature selection
- Optimization criteria
- …

M. Hardt, *How big data is unfair. Understanding unintended sources of unfairness in data driven decision making (2014)*

Bias in humans and machines





Algorithms are biased, but humans also are …

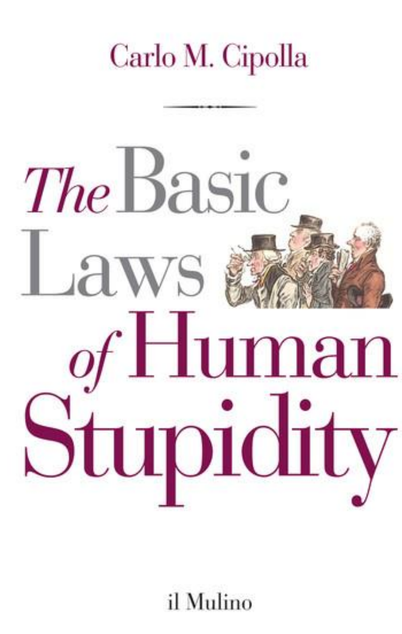
When should we trust humans and when algorithms?

Stupidity according to C. M. Cipolla

Third (and golden) basic law of stupidity

A stupid person is a person who causes losses to another person or to a group of persons while himself deriving no gain and even possibly incurring losses.

Carlo M. Cipolla
The Basic Laws of Human Stupidity (2011)



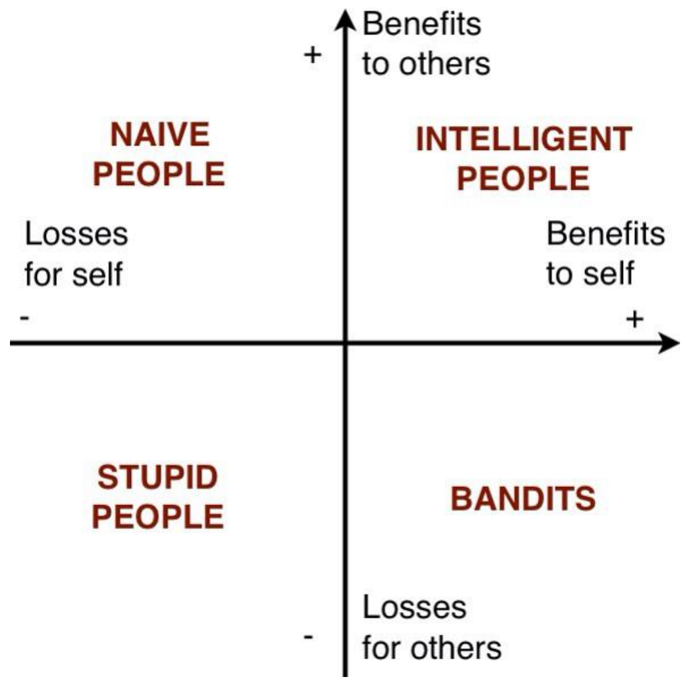
Stupidity according to C. M. Cipolla

Third (and golden) basic law of stupidity

A stupid person is a person who causes losses to another person or to a group of persons while himself deriving no gain and even possibly incurring losses.

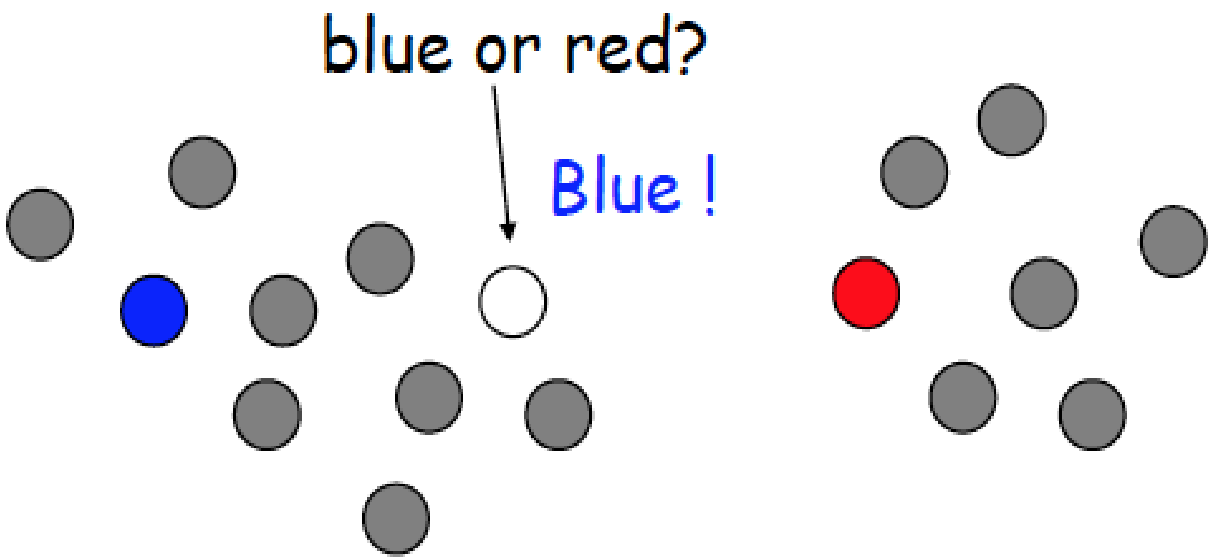
Carlo M. Cipolla

The Basic Laws of Human Stupidity (2011)



The smoothness assumption

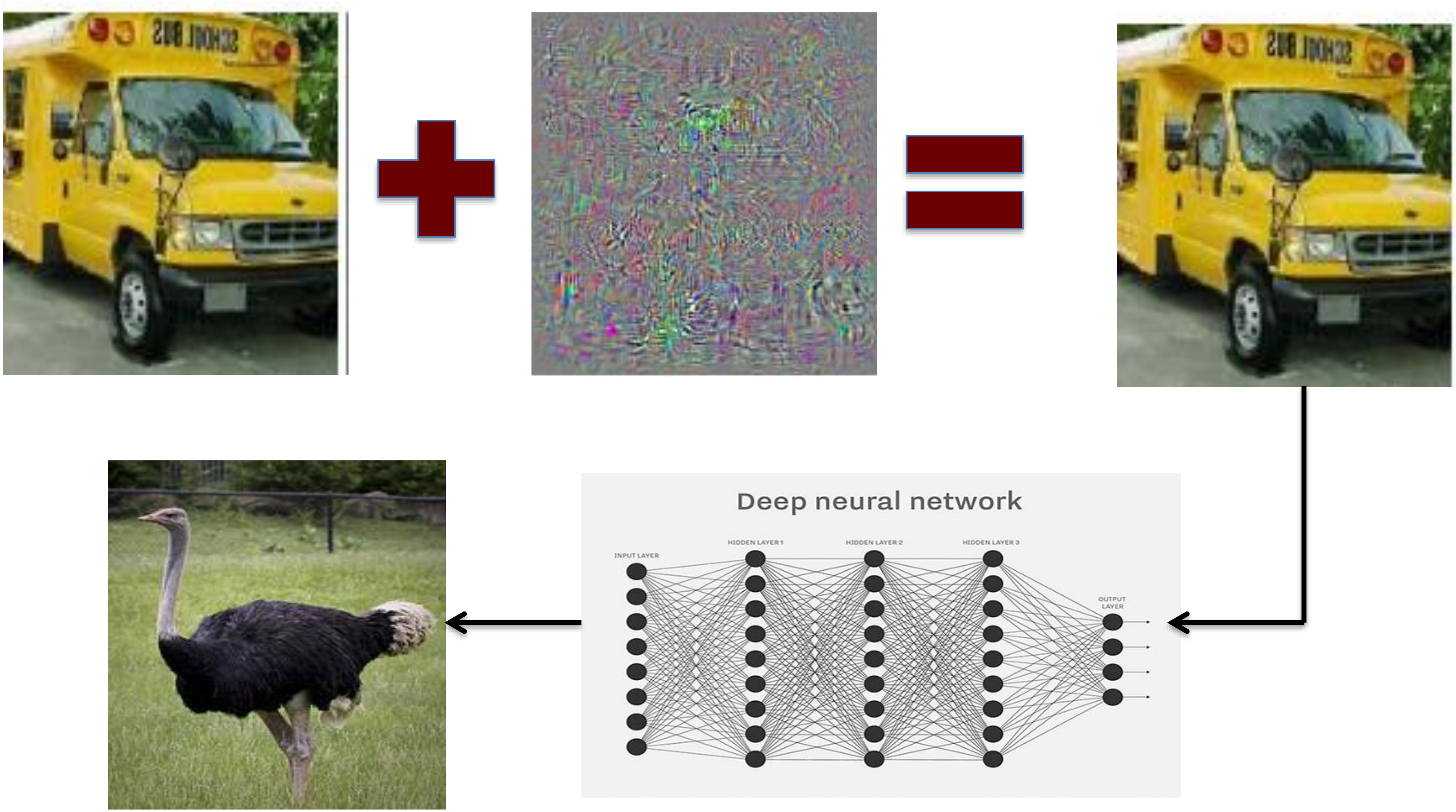
Points close to each other are more likely to share the same label



What about the performance of deep networks on image data that have been modified only slightly?

Courtesy Fabio Roli

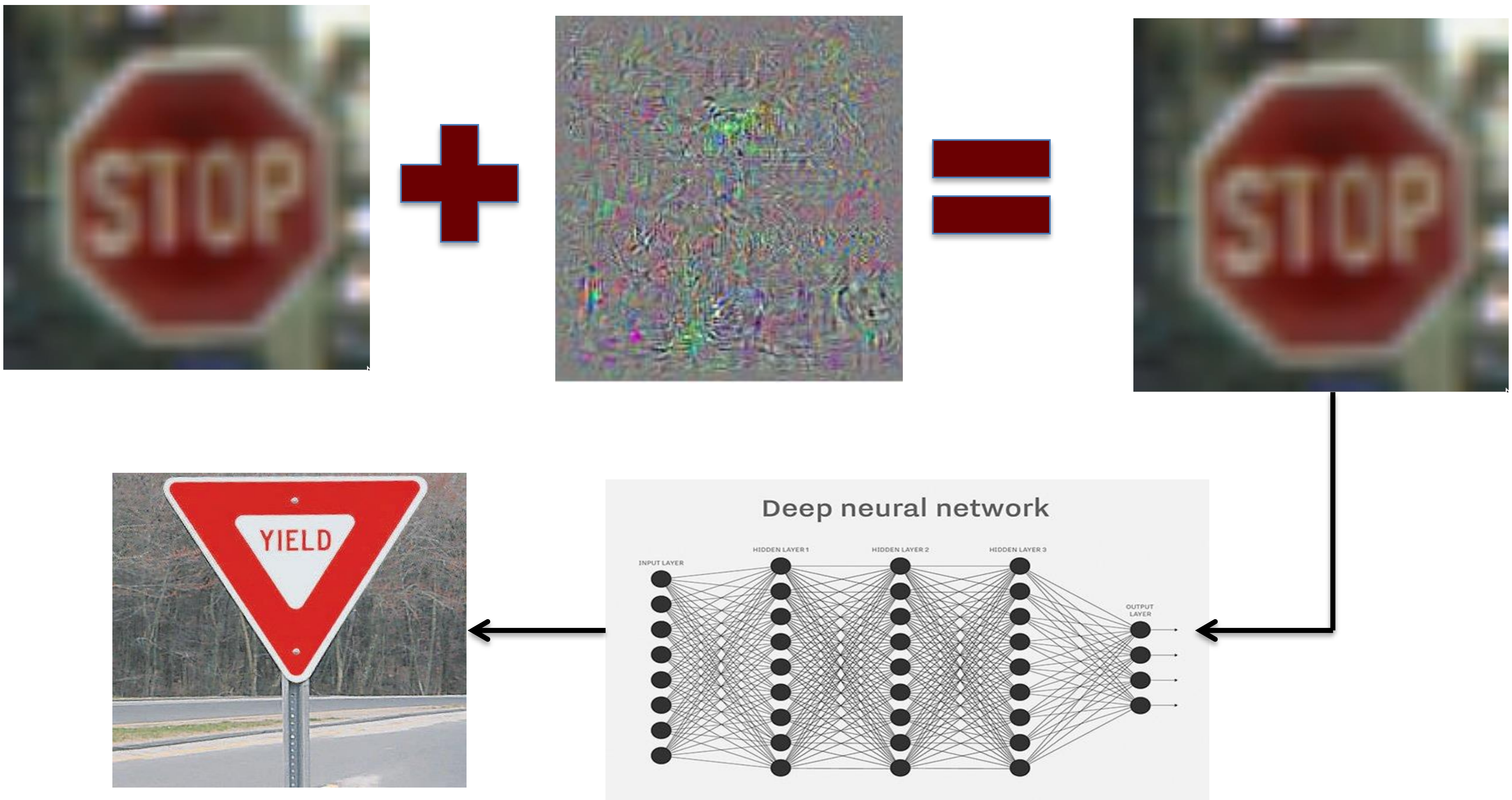
High accuracy = high robustness?



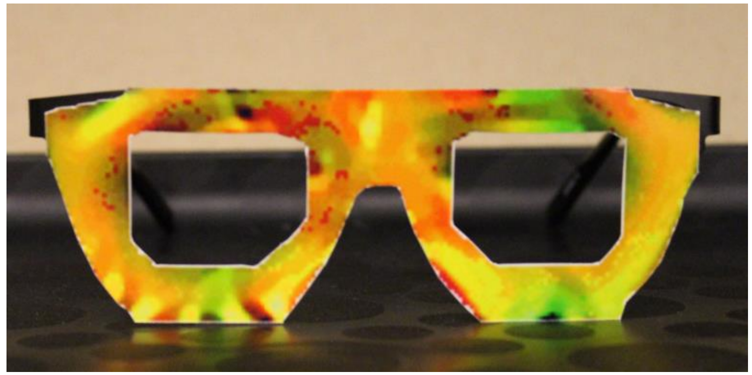
Szegedy et al., *Intriguing properties of neural networks* (2014)

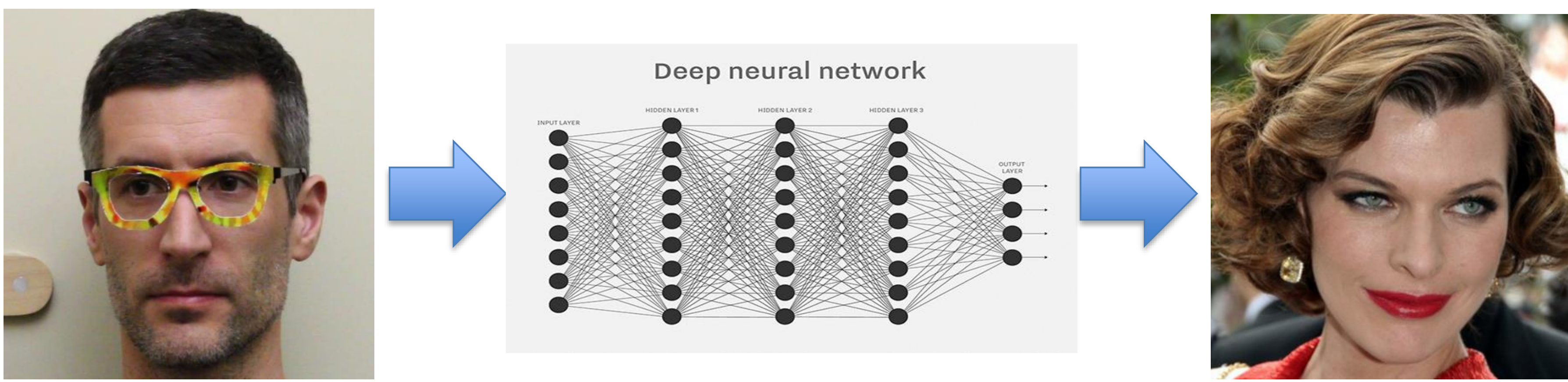
Cortesia di F. Roli

What if ...


STOP
STOP
YIELD
Deep neural network
INPUT LAYER
HIDDEN LAYER 1
HIDDEN LAYER 2
HIDDEN LAYER 3
OUTPUT LAYER

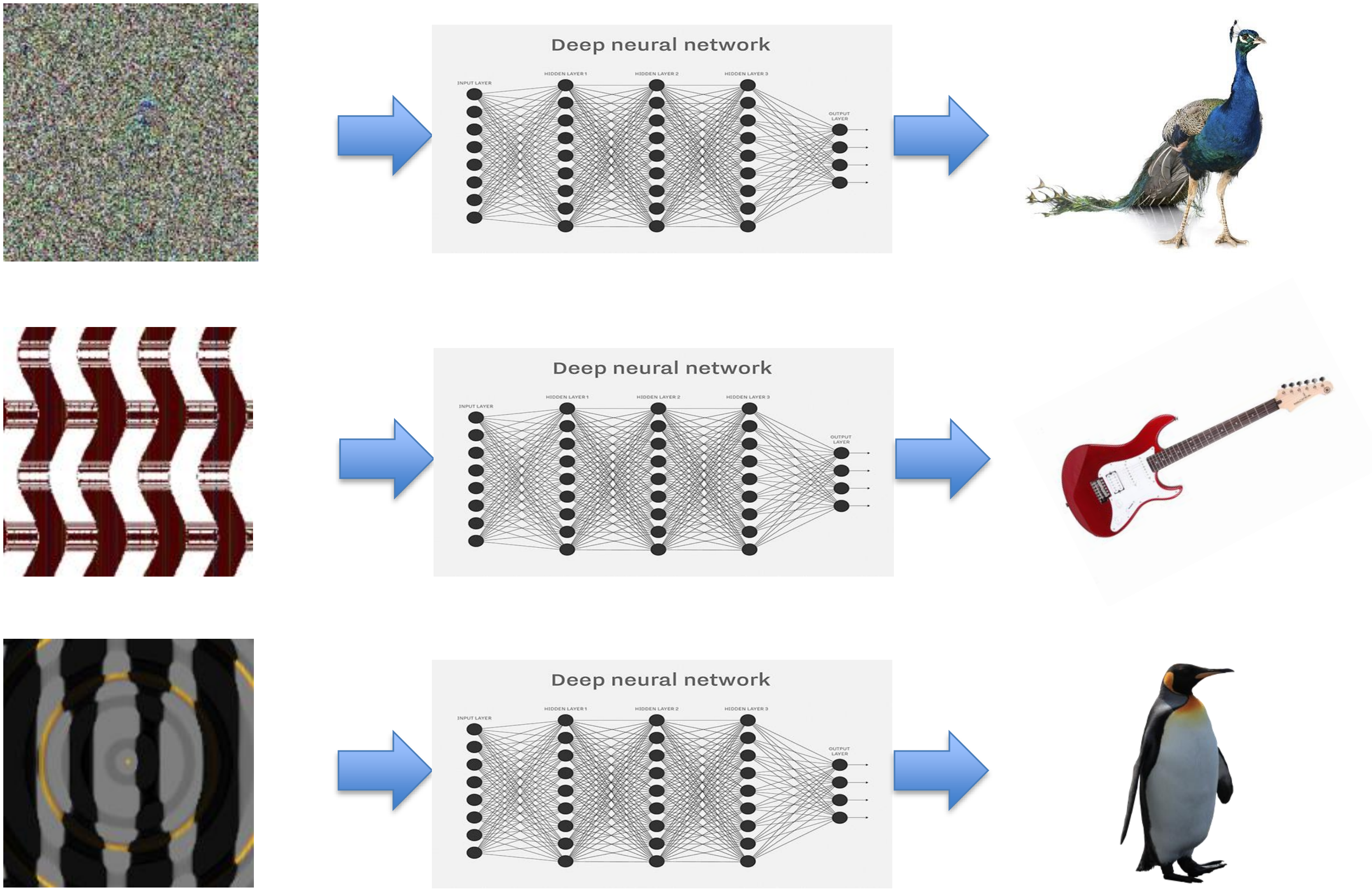
Fashionable glasses





M. Sharif et al., *Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition* (2016)

What does a machine see here?



A. Nguyen et al., *Deep neural networks are easily fooled: High confidence predictions for unrecognizable images* (2015)

The primacy of similarity

«Surely there is nothing more basic to thought and language than our sense of similarity. [...]

And every reasonable expectation depends on resemblance of circumstances, together with our tendency to expect similar causes to have similar effects.»

Willard V. O. Quine
Natural Kinds (1969)

Different similarity spaces

«Different creatures will have different similarity-spaces, hence different ways of grouping things […]

Such perceived similarities (or, for what matter, failure to perceive similarities) will manifest themselves in behavior and are a crucial part of explaining what is distinctive in each individual creature's way of apprehending the world.»

José Luis Bermùdez
Thinking Without Words (2003)

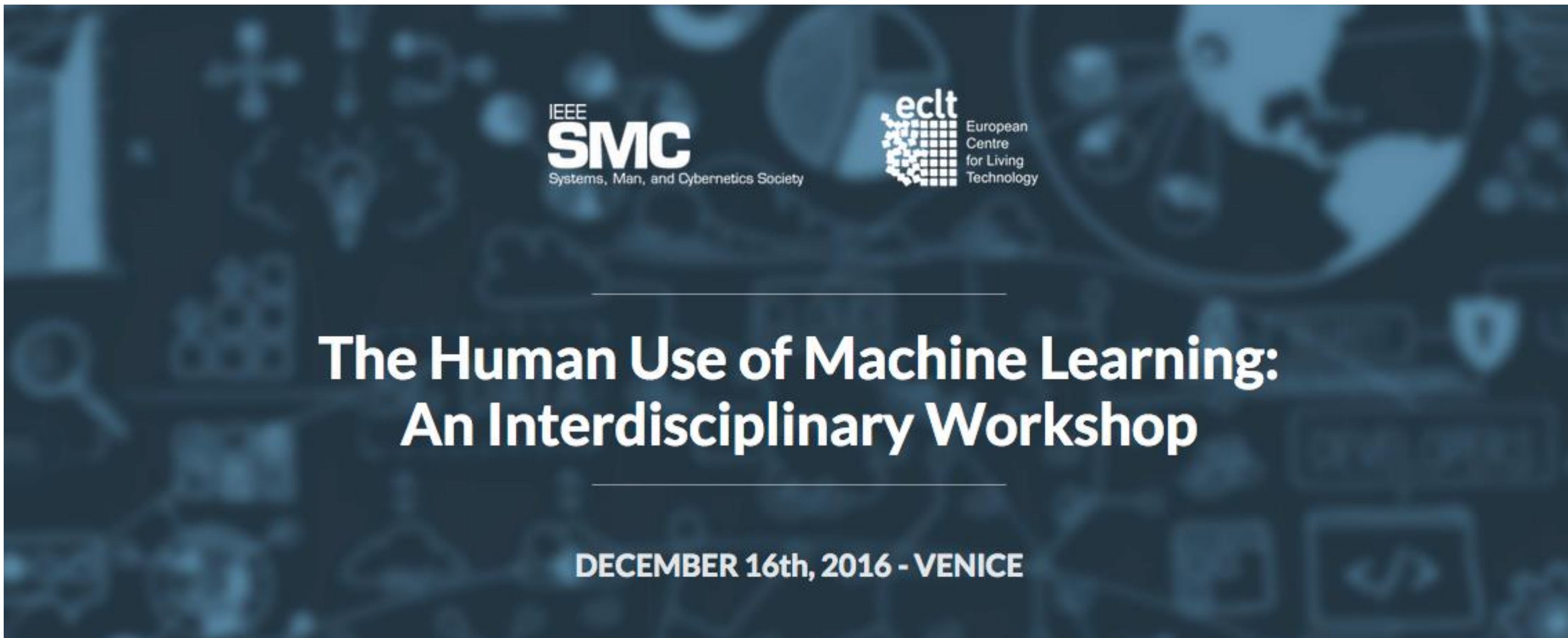
Cipolla, again

Fifth basic law of stupidity
A stupid person is the most dangerous type of person.

Corollary
A stupid person is more dangerous than a bandit.

Carlo M. Cipolla
The Fundamental Laws of Human Stupidity (2011)

If you want to learn more …



http://www.dsi.unive.it/HUML2016

Welcome to the AI4EU initiative

The AI4EU proposal adressing ICT-26 2018 H2020 call has successfully passed the evalution process.

The project should start early this autumn

https://ai4eu.org