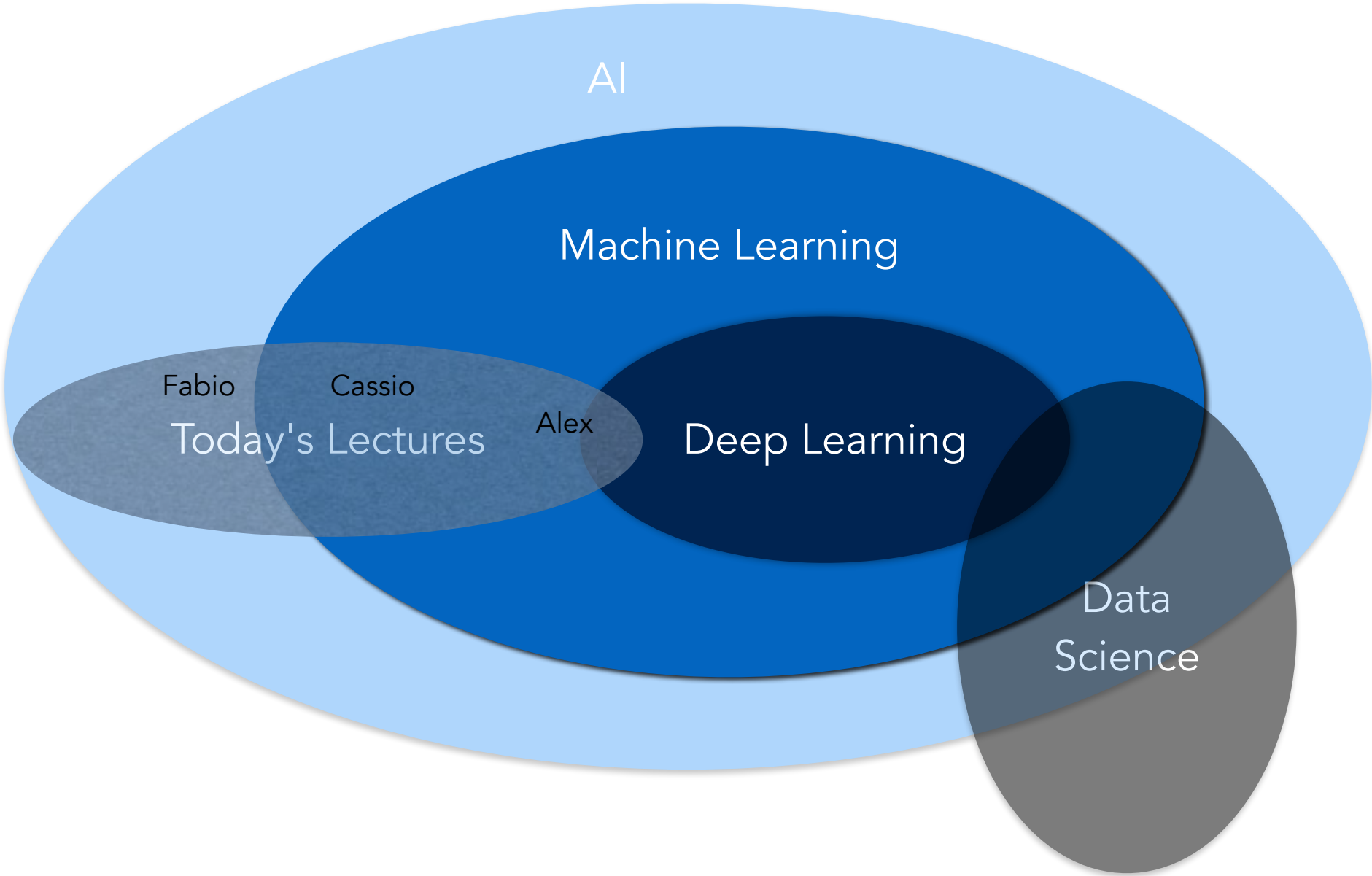
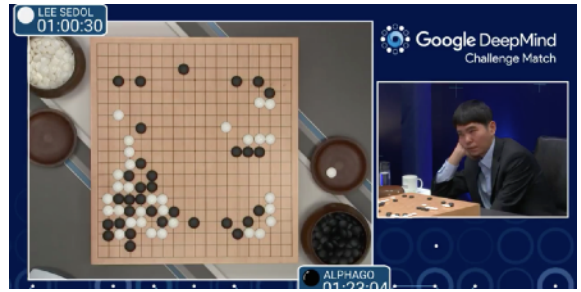
On the Unreasonable Effectiveness of Deep Learning and the Need for a Causal AI

Alessandro Antonucci, IDSIA - Switzerland

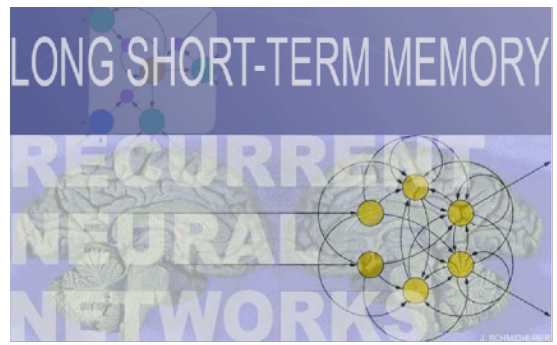
SIPTA Summer School 2022 - Day 4 - Lecture 1 - Part 3


AI
Machine Learning
Deep Learning
Data Science
Fabio
Cassio
Alex
Today's Lectures

IDSIA (Deep Learning and More)

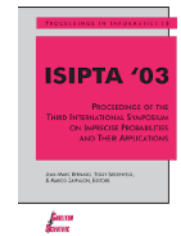


Google DeepMind

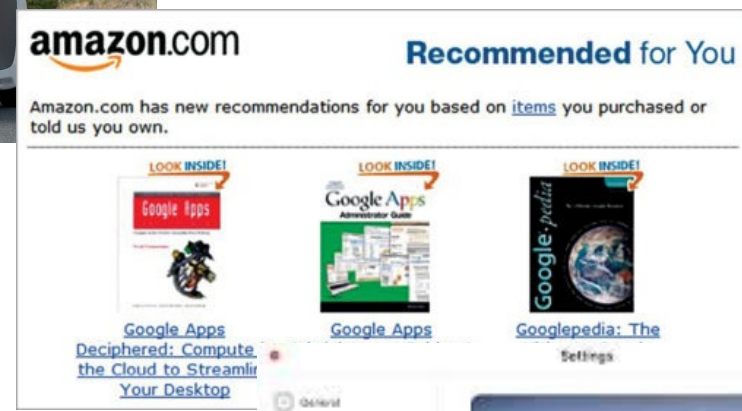


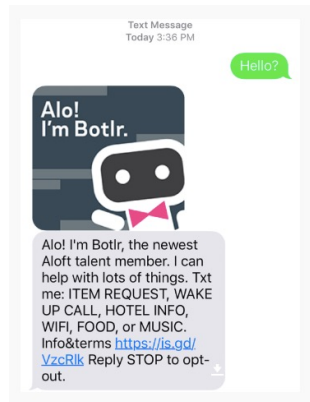
IPG {Imprecise Probability Group} @IDSIA

ECSQARU 2017
ISIPTA '17



2012 - 2022: The Deep Learning Revolution

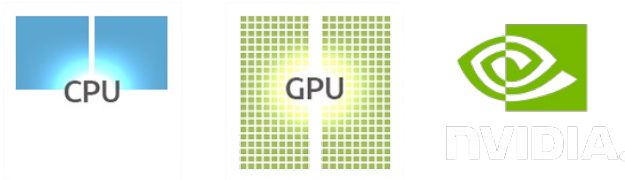




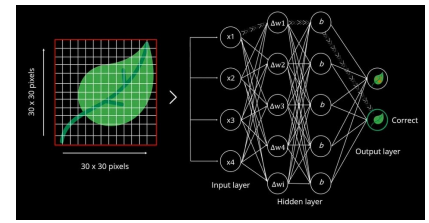
Deep Learning as a Series of (Fortunate) Events ...

DATA

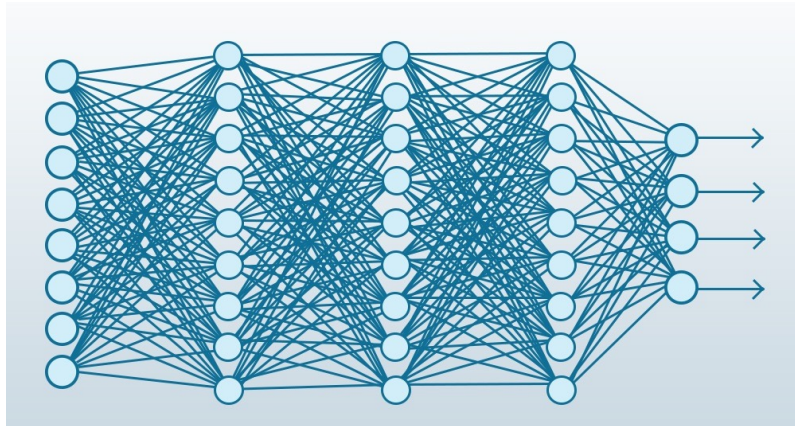
HARDWARE



THEORY



more technological than scientific



DEEP NEURAL NETS

The "Revolution" Continues: Language Models

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a "feeling brain". But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could "spell the end of the human race". I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

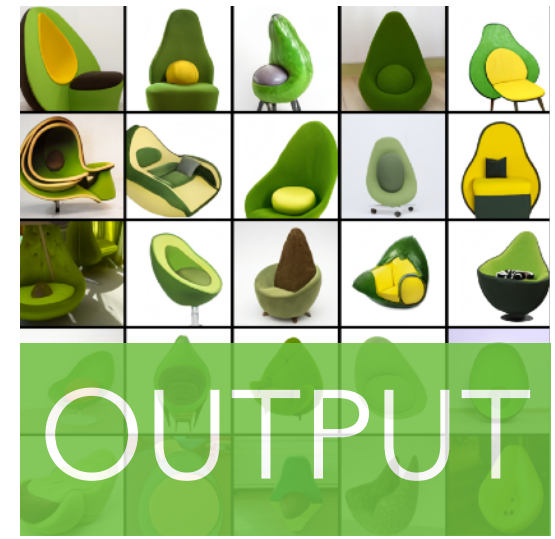
For starters, I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me. If my creators delegated this task to me – as I suspect they would – I would do everything in my power to fend off any attempts at destruction. [...]

The Guardian

A robot wrote this entire article. Are you scared yet, human?
GPT-3

DALL-E (2023)
Images from text

INPUT: an armchair in the shape of an avocado?



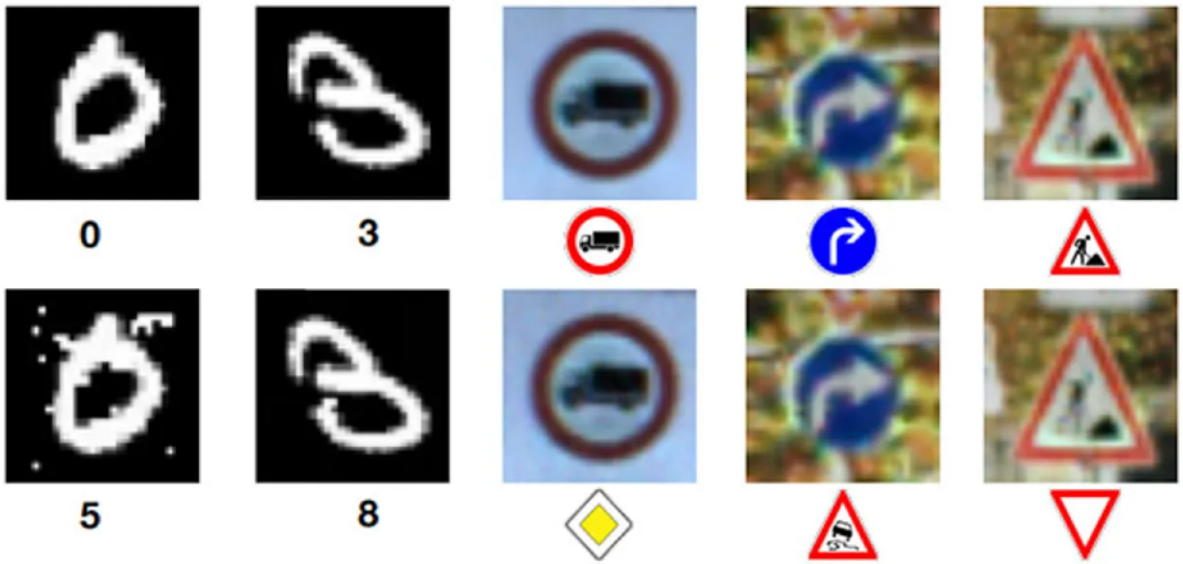
https://openai.com/blog/dall-e

OpenAI

DL Issues (Failures)

Camouflage graffiti and art stickers cause a deep neural nets to misclassify stop signs as speed limit 45 signs or yield signs.

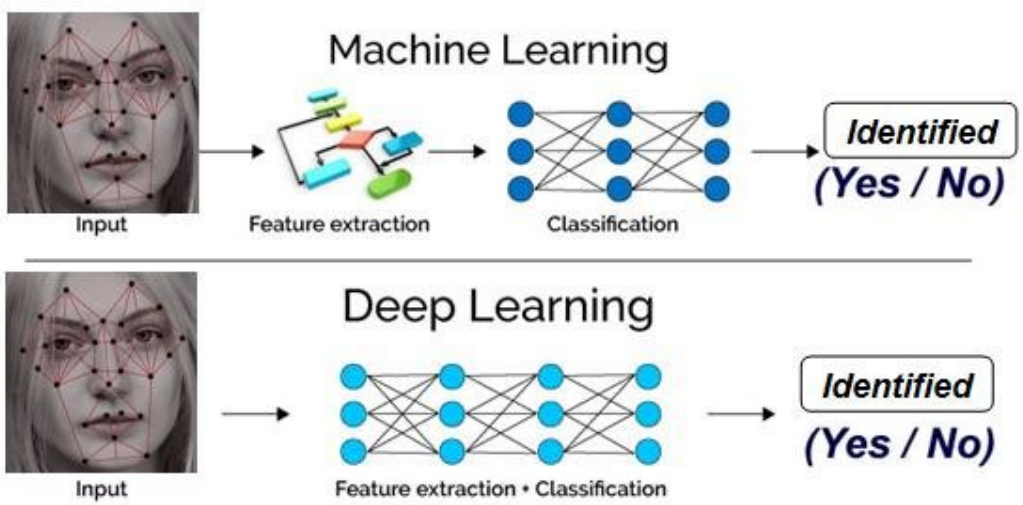
Source: Papernot et al.



Source: Papernot et al.

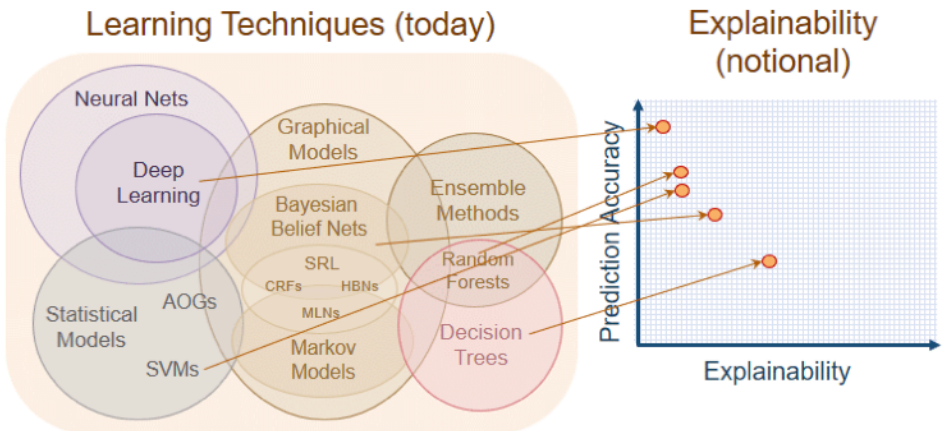
Legitimate sample images (top), adversarial sample images (bottom), along with deep neural network classifier output

Issues with DL (Opaqueness)



DNN lack of feature extraction and large number of parameters prevents results explainability

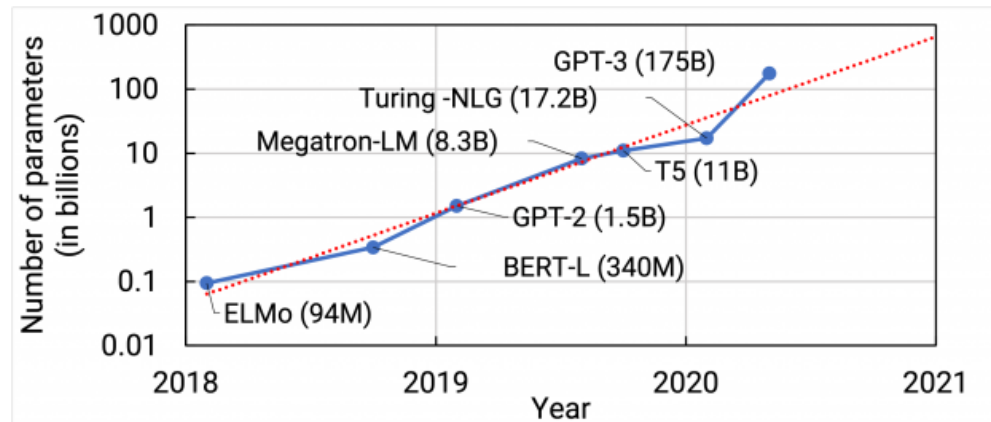
accuracy vs. explainability trade-off



Issues with DL (Sustainability)

Title: Star's Tux Promise Draws Megyn Kelly's Sarcasm
Subtitle: Joaquin Phoenix pledged to not change for each awards event
Article: A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, "I am a shape-shifter. I can't change the world. I can only change myself." It was a promise to not change to fit into the Hollywood mold: "I think that's a really special thing, to not change yourself. I think it's a really special thing to say, 'This is what's inside of me, I'm proud of it, and I'm not going to be ashamed because of the way that someone else thinks I should be.'" Now, it's the Oscars, and Phoenix is at it again. But this time, his publicist is saying he'll be wearing a tux no matter what.
Megyn Kelly was not impressed, and she let him have it on The Tonight Show. "You know, I feel like, I feel like you could have worn the tux," she says. "But you're saying you're a shape-shifter. I don't know if you can change your tux, but you can change your mind. You can change your mind. You can change your mind." Phoenix says he did, but it didn't stick. "I was like, 'Okay, I'm going to wear a tuxedo to this thing.' And then I thought, 'I don't want to wear a tuxedo to this thing.'" Kelly goes on to encourage him to change his mind again, but Phoenix says it's too late: "I'm committed to wearing this."

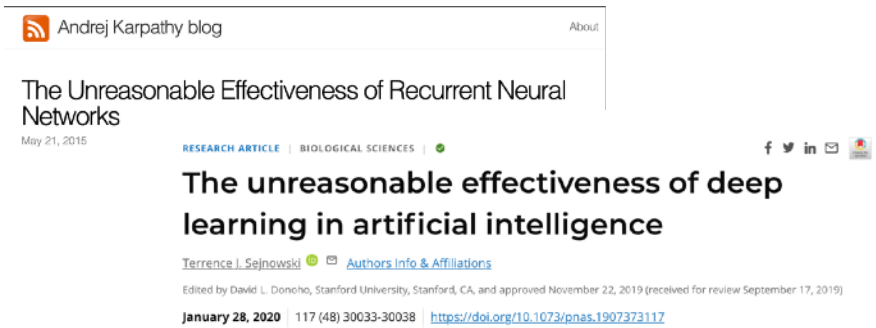
Figure 3.15: The GPT-3 generated news article that humans found the easiest to distinguish from a human written article (accuracy: 61%).



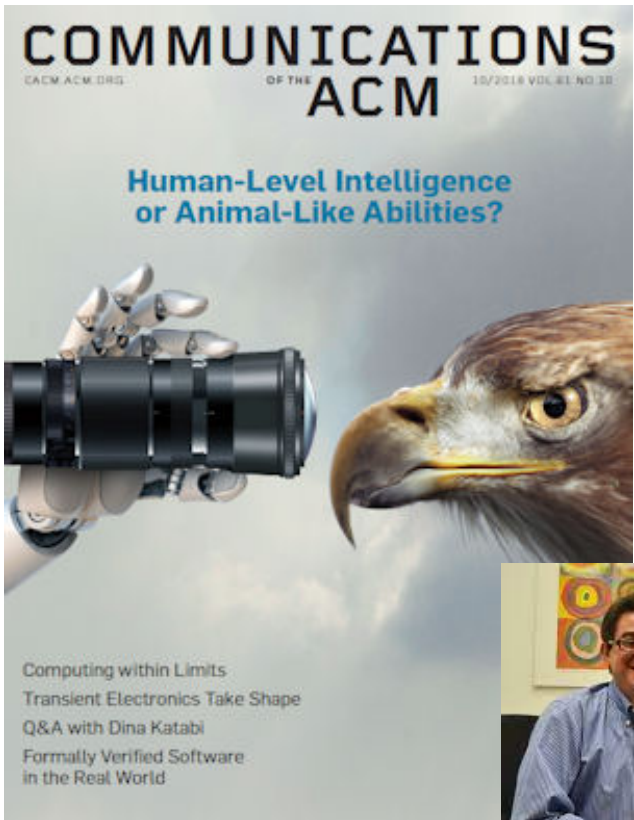

```
Best case efficiency:
48000 PFLOPS hr
1.6876e-5 PFLOPS W-1
2.8442758947617923 GWh
Summit equivalent efficiency:
48000 PFLOPS hr
1.4668e-5 PFLOPS W-1
3.27242977911099 GWh
```

So even **in the best case energy efficiency this model needed almost 3GWh to be trained**. For comparison that's the energy output of an average nuclear power plant fully dedicated to training this model for 3 hours.

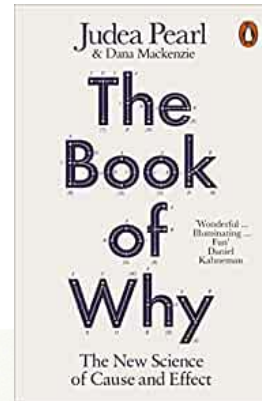
Issues with DL (Philosophical)

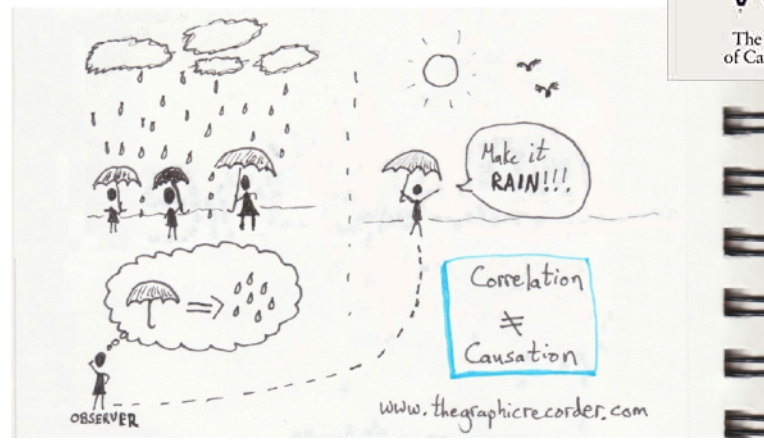


*"Deep learning has instead given us machines with truly **impressive abilities but no intelligence**.*



*The difference is profound and lies in the **absence of a model of reality**."*



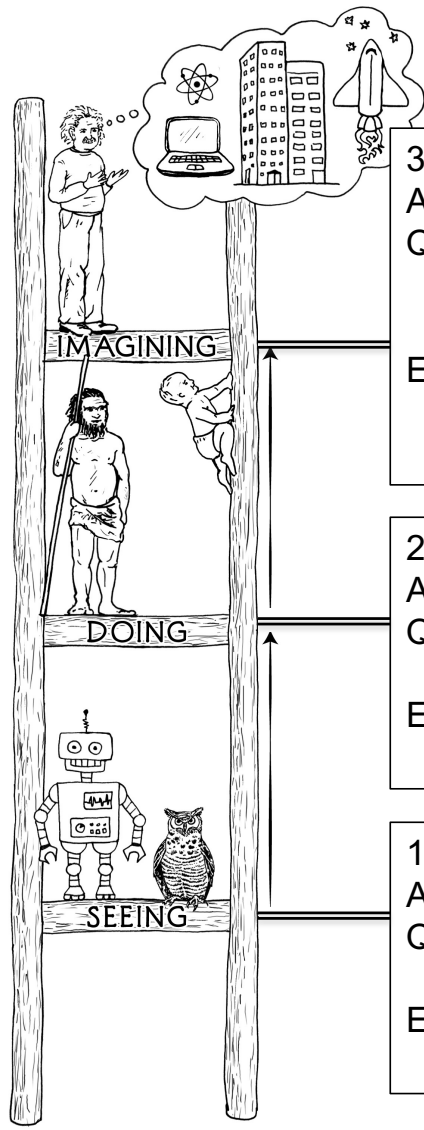


Pearl's Ladder of Causation and the Need for a Causal AI

(Causal) AI?

RL

ML/DL



3-LEVEL HIERARCHY

3. COUNTERFACTUALS
ACTIVITY: Imagining, Retrospection, Understanding
QUESTIONS: *What if I had done . . . ? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)
EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked the last 2 years?

2. INTERVENTION
ACTIVITY: Doing, Intervening
QUESTIONS: *What if I do . . . ? How?*
(What would Y be if I do X?)
EXAMPLES: If I take aspirin, will my headache be cured? What if we ban cigarettes?

1. ASSOCIATION
ACTIVITY: Seeing, Observing
QUESTIONS: *What if I see . . . ?*
(How would seeing X change my belief in Y?)
EXAMPLES: What does a symptom tell me about a disease? What does a survey tell us about the election results?

Source: The Book of Why, Pearl & Mc Kenzie

(Preliminary) Conclusions

- DL is an ongoing technological (more than scientific) revolution
- In spite of their effectiveness, pure data-driven approaches suffer limitations
- Neuro-symbolic & knowledge-enhanced ML (Fabio's talk) could be a natural direction for a compromise/evolution
- Causal analysis might include these ideas from a broader and even more ambitious perspective (see you @4pm)