Mitglied der Helmholtz-Gemeinschaft

JÜLICH FORSCHUNGSZENTRUM

Skeptical View on AI Application in Science

February 28, 2020 | Jędrzej Rybicki

Disclaimer

Opinions are mine not my employer

Intro

- big successes of AI in recent years

Intro

- big successes of AI in recent years
- ... or is it just a big hype?

Intro

- big successes of AI in recent years
- ... or is it just a big hype?
- resulting funding opportunities in science
- (lots of AI products on the market)

Intro

- big successes of AI in recent years
- ... or is it just a big hype?
- resulting funding opportunities in science
- (lots of AI products on the market)

Skeptik

Skeptik but not denier. Critical thinking, seeing not only powers but also limitations.

Intro: What are we talking about

Classical AI:

- rules & heuristics
- almost forgotten by now?
- clearly limited when applied outside of its “domain”
- reasoning

Intro: What are we talking about

Classical AI:

- rules & heuristics
- almost forgotten by now?
- clearly limited when applied outside of its “domain”
- reasoning

ML AI:

- automatic algorithm creation (*“getting computers to act without being explicitly programmed” Andrew Ng*)
- data driven (data hungriness)
- mostly Deep Learning

Science: Scientific method

1. empirical method of acquiring knowledge
2. develop a more sophisticated understanding over time (novelty)
3. replication, testable outcomes$\Rightarrow$ falsification
4. counterfactual situations

Science: Scientific method

1. empirical method of acquiring knowledge
2. develop a more sophisticated understanding over time (novelty)
3. replication, testable outcomes $\Rightarrow$ falsification
4. counterfactual situations

$$\boxed{\text{Observation}} \xrightarrow[\text{induction}]{} \boxed{\text{Hypothesis}} \xrightarrow[\text{deduction}]{} \boxed{\text{Verification}}$$

Science: Scientific method

1. empirical method of acquiring knowledge
2. develop a more sophisticated understanding over time (novelty)
3. replication, testable outcomes $\Longrightarrow$ falsification
4. counterfactual situations

$$\boxed{\text{Observation}} \xrightarrow[\text{induction}]{} \boxed{\text{Hypothesis}} \xrightarrow[\text{deduction}]{} \boxed{\text{Verification}}$$

$$\boxed{\text{Data}} \xrightarrow[\text{training}]{} \boxed{\text{Model}} \xrightarrow[\text{explainability}]{} \boxed{\text{Insight}}$$

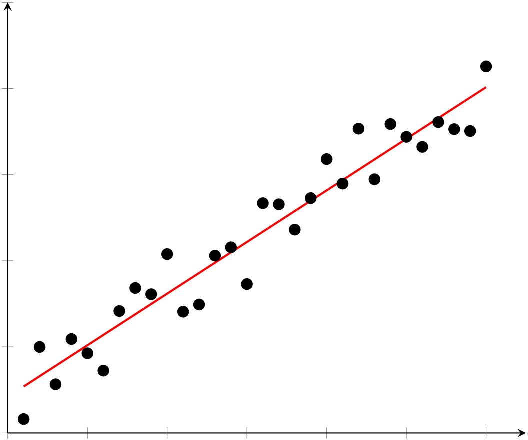
Science: Scientific method

1. empirical method of acquiring knowledge
2. develop a more sophisticated understanding over time (novelty)
3. replication, testable outcomes $\Rightarrow$ falsification
4. counterfactual situations

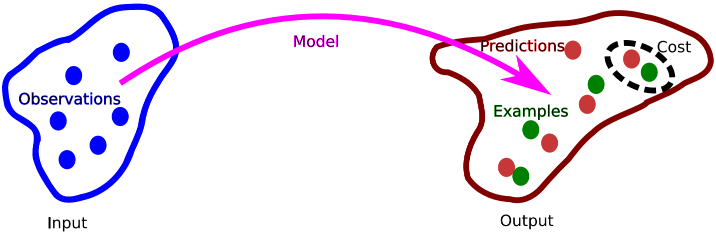
$\boxed{\text{Observation}} \xrightarrow[\text{induction}]{} \boxed{\text{Hypothesis}} \xrightarrow[\text{deduction}]{} \boxed{\text{Verification}}$

$\boxed{\text{Data}} \xrightarrow[\text{training}]{} \boxed{\text{Model}} \xrightarrow[\text{explainability}]{} \boxed{\text{Insight}}$

Model: What are we talking about



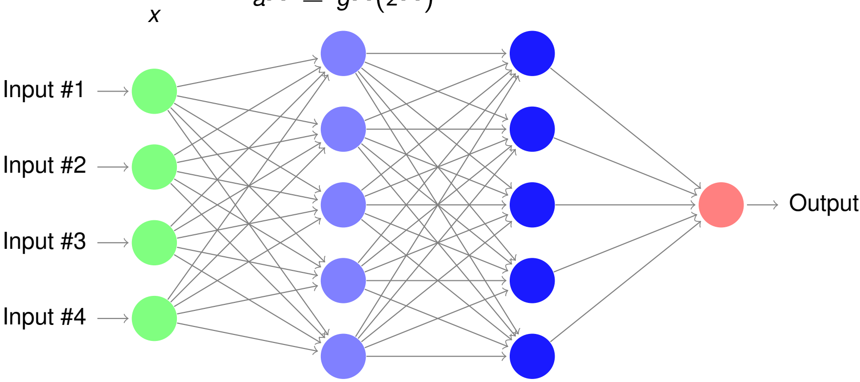
Model: Geometric View



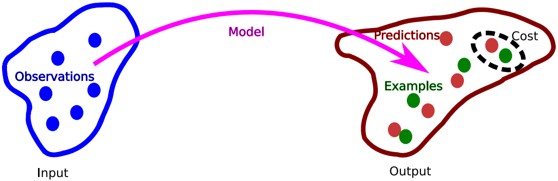
- Learning is optimization problem: minimize the error between model and training set (Cost)
- DL Model: chain of simple geometric continuous transformations

Deep Neural Networks

$$z^{[1]} = w^{[1]}x + b^{[1]}$$
$$a^{[1]} = g^{[1]}(z^{[1]})$$

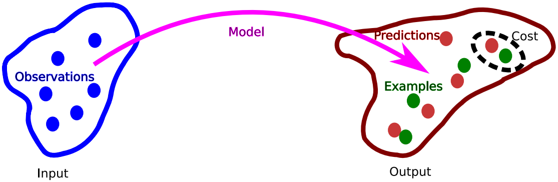


$$z^{[2]} = w^{[2]}a^{[1]} + b^{[2]}$$
$$a^{[2]} = g^{[2]}(z^{[2]})$$



DL Model: the transformation is chain of simple geometric continuous transformations

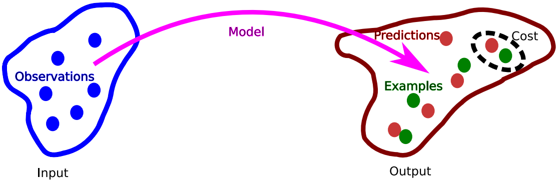
- model is a function
- currently: continuous (which is already a limitation)
- it makes mathematical sense outside of the domain
- at best it can interpolate over the input



DL Model: the transformation is chain of simple geometric continuous transformations

- model is a function
- currently: continuous (which is already a limitation)
- it makes mathematical sense outside of the domain
- at best it can interpolate over the input

$\Rightarrow$ $f(x) = x$ network by Gary Marcus (filling the gaps)



DL Model: the transformation is chain of simple geometric continuous transformations

- model is a function
- currently: continuous (which is already a limitation)
- it makes mathematical sense outside of the domain
- at best it can interpolate over the input

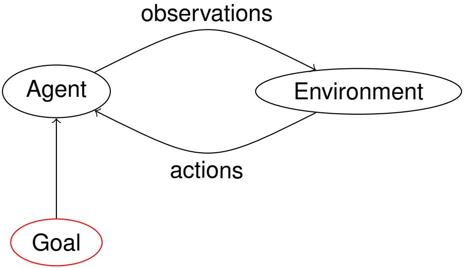
$\Rightarrow$ $f(x) = x$ network by Gary Marcus (filling the gaps)

- it is not programming
- even simple task like sorting cannot be accomplished (efficiently)

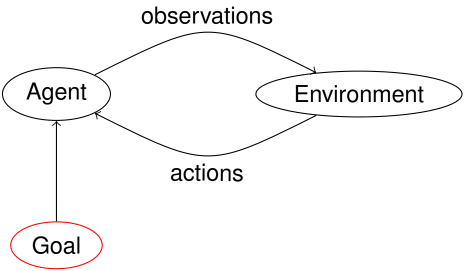
Special case: target domain can be set of (human) concepts

- ... but it does not mean that the model understands or uses the concepts
- “superhuman” performance on ImageNet: what does it mean? (⇒ overattribution)

Special note: Reinforced learning



Special note: Reinforced learning

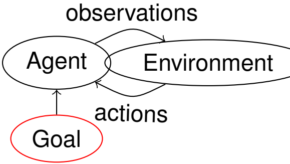


Application criteria (based on Alpha-0):

1. huge combinational space
2. clear objective (function/metric)
3. data (or simulation)

Reinforced learning

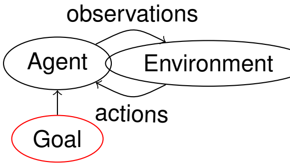
is this the way how we learn? we rather understand in terms of things that we already understand



- Alpha-Go

Reinforced learning

is this the way how we learn? we rather understand in terms of things that we already understand



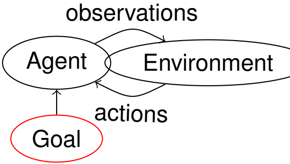
- Alpha-Go
- Alpha-0
- universal framework that learn any game

⇒ Atari 2600 games

- very good on Breakout

Reinforced learning

is this the way how we learn? we rather understand in terms of things that we already understand



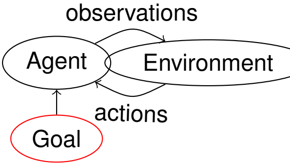
- Alpha-Go
- Alpha-0
- universal framework that learn any game

⇒ Atari 2600 games

- very good on Breakout
- very bad on Montezuma's Revenge

Reinforced learning

is this the way how we learn? we rather understand in terms of things that we already understand



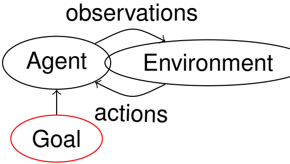
- Alpha-Go
- Alpha-0
- universal framework that learn any game

⇒ Atari 2600 games

- very good on Breakout
- very bad on Montezuma's Revenge
- Breakout: unless you rotate the screen or even move paddle 2 pixels higher

Reinforced learning

is this the way how we learn? we rather understand in terms of things that we already understand



- Alpha-Go
- Alpha-0
- universal framework that learn any game

$\Rightarrow$ Atari 2600 games

- very good on Breakout
- very bad on Montezuma's Revenge
- Breakout: unless you rotate the screen or even move paddle 2 pixels higher
- lots of anticipating but 0 understanding

Schema Networks: Zero-shot Transfer with a Generative Causal Model of Intuitive Physics, Kansky et al.

Training/Model Summary

- despite the hype: very simple (limited?) idea

Training/Model Summary

- despite the hype: very simple (limited?) idea
- DL can do lots of interesting things
- ... but also completely unable to do others

Training/Model Summary

- despite the hype: very simple (limited?) idea
- DL can do lots of interesting things
- ... but also completely unable to do others
- performance is not understanding (image recognition)
- brute force $\Rightarrow$ question of efficiency (ResNet18: 11689512 parameters. Optimal configuration: a point in the 11689512 dimensional space.)

JÜLICH
FORSCHUNGSZENTRUM

Training/Model Summary

- despite the hype: very simple (limited?) idea
- DL can do lots of interesting things
- ... but also completely unable to do others
- performance is not understanding (image recognition)
- brute force $\Rightarrow$ question of efficiency (ResNet18: 11689512 parameters. Optimal configuration: a point in the 11689512 dimensional space.)
- correlations between features rather than abstractions
- trend of hard-coding domain knowledge into the neural networks (Convolutional neural networks)
- limited application outside of the domain

Training/Model Summary

- despite the hype: very simple (limited?) idea
- DL can do lots of interesting things
- ... but also completely unable to do others
- performance is not understanding (image recognition)
- brute force $\Rightarrow$ question of efficiency (ResNet18: 11689512 parameters. Optimal configuration: a point in the 11689512 dimensional space.)
- correlations between features rather than abstractions
- trend of hard-coding domain knowledge into the neural networks (Convolutional neural networks)
- limited application outside of the domain

In principle, given infinite data, deep learning systems are powerful enough to represent any finite deterministic “mapping" between any given set of inputs and a set of corresponding outputs

Data

$$\boxed{\text{Observation}} \xrightarrow[\text{induction}]{} \boxed{\text{Hypothesis}} \xrightarrow[\text{deduction}]{} \boxed{\text{Verification}}$$

$$\boxed{\text{Data}} \xrightarrow[\text{training}]{} \boxed{\text{Model}} \xrightarrow[\text{explainability}]{} \boxed{\text{Insight}}$$

Data:

1. Data deluge
2. Data hungriness

Data

$$\text{Observation} \xrightarrow{\text{induction}} \text{Hypothesis} \xrightarrow{\text{deduction}} \text{Verification}$$

$$\text{Data} \xrightarrow{\text{training}} \text{Model} \xrightarrow{\text{explainability}} \text{Insight}$$

Data:

1. Data deluge
2. Data hungriness

Ways of increasing size of data:

- increasing number of rows
- increasing number of columns
- increasing density of rows

Data: increasing number of rows

Cautionary note: Quality vs. Quantity

1936 U.S. election: “Literary Digest” conducted huge poll with 2.3 million voters: Alf Landon. George Gallup conducted a far smaller (but more scientifically based) survey, correctly predicted Roosevelt’s victory.

JÜLICH FORSCHUNGSZENTRUM

Data: increasing number of rows

Cautionary note: Quality vs. Quantity

1936 U.S. election: “Literary Digest” conducted huge poll with 2.3 million voters: Alf Landon. George Gallup conducted a far smaller (but more scientifically based) survey, correctly predicted Roosevelt’s victory.

- statistics would say: better to have 5% random than 90% non-random
- learning algorithm will not work (not enough iterations)

Data: increasing number of rows

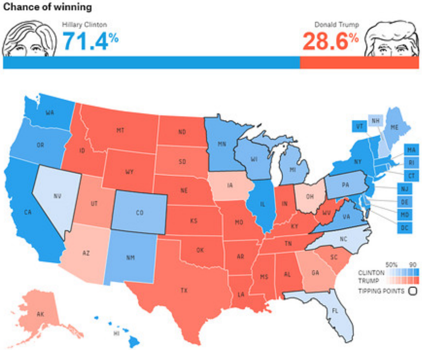
Cautionary note: Quality vs. Quantity

1936 U.S. election: “Literary Digest” conducted huge poll with 2.3 million voters: Alf Landon. George Gallup conducted a far smaller (but more scientifically based) survey, correctly predicted Roosevelt’s victory.

- statistics would say: better to have 5% random than 90% non-random
- learning algorithm will not work (not enough iterations)
- data from different sources
- usually pre-processed $\Rightarrow$ have different probability distributions
- hard to say what is representative

Old stories...

Old stories...



Nate Silver’s model... On election day.

Data: combining sources

Kidney stone treatment study

	Treatment A	Treatment B
Small stones	93%	87%
Large stones	73%	69%

Data: combining sources

Kidney stone treatment study

	Treatment A	Treatment B
Small stones	81/87 (93%)	234/270 (87%)
Large stones	192/263 (73%)	55/80 (69%)
Overall	273/350 (78%)	289/350 (83%)

Data: combining sources

Kidney stone treatment study

	Treatment A	Treatment B
Small stones	81/87 (93%)	234/270 (87%)
Large stones	192/263 (73%)	55/80 (69%)
Overall	273/350 (78%)	289/350 (83%)

Simpson's paradox

a trend appears in several different groups of data but disappears or reverses when these groups are combined

Data: increasing number of rows

Ramsey theory

A branch of mathematics that studies the conditions under which order must appear. (Wikipedia)

Data: increasing number of rows

Ramsey theory

A branch of mathematics that studies the conditions under which order must appear. (Wikipedia)

Example: the minimum number of guests that must be invited so that at least m will know each other and at least n does not?

Data: increasing number of rows

Ramsey theory

A branch of mathematics that studies the conditions under which order must appear. (Wikipedia)

Example: the minimum number of guests that must be invited so that at least *m* will know each other and at least *n* does not?

e4

e2 e6

e1 e5

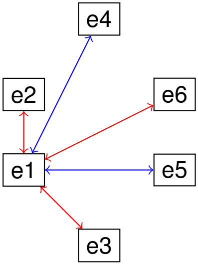
e3

Data: increasing number of rows

Ramsey theory

A branch of mathematics that studies the conditions under which order must appear. (Wikipedia)

Example: the minimum number of guests that must be invited so that at least *m* will know each other and at least *n* does not?

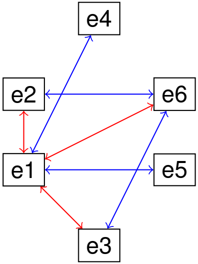


Data: increasing number of rows

Ramsey theory

A branch of mathematics that studies the conditions under which order must appear. (Wikipedia)

Example: the minimum number of guests that must be invited so that at least *m* will know each other and at least *n* does not?

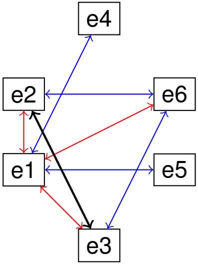


Data: increasing number of rows

Ramsey theory

A branch of mathematics that studies the conditions under which order must appear. (Wikipedia)

Example: the minimum number of guests that must be invited so that at least *m* will know each other and at least *n* does not?



- for a given graph size of 6

$\Rightarrow$ you will find a clique of 3!

- pattern-finding!

Data: increasing number of rows

van der Waerden's theorem

van der Waerden's theorem is a theorem about the existence of arithmetic progressions in sets. In a series of length $W(r, k)$ r colors at least k form an arithmetic progression.

Example: in a series of length $W(r = 2, k = 3) \geq 9$

1	2	3	4	5	6	7	8	9
B	R	R	B	B	R	R	B	?

Data: increasing number of rows

van der Waerden's theorem

van der Waerden's theorem is a theorem about the existence of arithmetic progressions in sets. In a series of length $W(r, k)$ r colors at least k form an arithmetic progression.

Example: in a series of length $W(r = 2, k = 3) \geq 9$

1	2	3	4	5	6	7	8	9
B	R	R	B	B	R	R	B	?
B	R	R	B	**B**	R	R	B	**B**

Data: increasing number of rows

van der Waerden's theorem

van der Waerden's theorem is a theorem about the existence of arithmetic progressions in sets. In a series of length $W(r, k)$ r colors at least k form an arithmetic progression.

Example: in a series of length $W(r = 2, k = 3) \geq 9$

1	2	3	4	5	6	7	8	9
B	R	R	B	B	R	R	B	?
B	R	R	B	**B**	R	R	B	**B**
B	R	**R**	B	B	**R**	R	B	**R**

Data: increasing number of rows

van der Waerden's theorem

van der Waerden's theorem is a theorem about the existence of arithmetic progressions in sets. In a series of length $W(r, k)$ r colors at least k form an arithmetic progression.

Example: in a series of length $W(r = 2, k = 3) \geq 9$

1	2	3	4	5	6	7	8	9
B	R	R	B	B	R	R	B	?
B	R	R	B	**B**	R	R	B	**B**
B	R	**R**	B	B	**R**	R	B	**R**

Ramifications:

- how big is the structure to find a given substructure
- correlation is result of data size
- complete disorder is not possible

Data: increasing number of columns

Data Leakage

Data leakage is when information from outside the training dataset is used to create the model.

Data: increasing number of columns

Data Leakage

Data leakage is when information from outside the training dataset is used to create the model.

Examples:

- “it rains on rainy days”

Data: increasing number of columns

Data Leakage

Data leakage is when information from outside the training dataset is used to create the model.

Examples:

- “it rains on rainy days”
- IBM training set

Data: increasing number of columns

Data Leakage

Data leakage is when information from outside the training dataset is used to create the model.

Examples:

- “it rains on rainy days”
- IBM training set
- normalize or standardize your entire dataset

$\Rightarrow$ data rescaling process that you performed had knowledge of the full distribution of data

Data: increasing number of columns

Data Leakage

Data leakage is when information from outside the training dataset is used to create the model.

Examples:

- "it rains on rainy days"
- IBM training set
- normalize or standardize your entire dataset

$\Rightarrow$ data rescaling process that you performed had knowledge of the full distribution of data

Results:

- overestimation of model's performance
- reversing an anonymization and obfuscation (sensitive data)

JÜLICH
FORSCHUNGSZENTRUM

Data: Summary

- often you require lots of data to create complex model
- or you are overloaded with the data anyways
- data collection might be harder than you think (end-to-end control)
- danger of emerging patterns (Ramsey & van der Waerden)
- The Curse of Dimensionality
- long-tail problem (things that don't happen so often)

Data: Summary

- often you require lots of data to create complex model
- or you are overloaded with the data anyways
- data collection might be harder than you think (end-to-end control)
- danger of emerging patterns (Ramsey & van der Waerden)
- The Curse of Dimensionality
- long-tail problem (things that don't happen so often)
- Illusion of Invariants: Data that span several order of magnitude leads to high R^2 and makes invariants notable.

The Illusion of Invariant Quantities in Life Histories Sean Nee, Nick Colegrave, Stuart A. West, Alan Grafen

Deduction

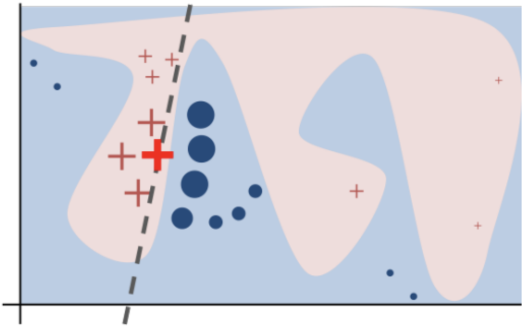
$$\boxed{\text{Observation}} \xrightarrow[\text{induction}]{} \boxed{\text{Hypothesis}} \xrightarrow[\text{deduction}]{} \boxed{\text{Verification}}$$

$$\boxed{\text{Data}} \xrightarrow[\text{training}]{} \boxed{\text{Model}} \xrightarrow[\text{explainability}]{} \boxed{\text{Insight}}$$

Crucial for:

- replication, testable outcomes (trust)
- falsification
- novelty (looking into the black box for new insights)
- counterfactual situations

LIME: Local Interpretable Model-Agnostic Explanations



“Why Should I Trust You?" Explaining the Predictions of Any Classifier Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin

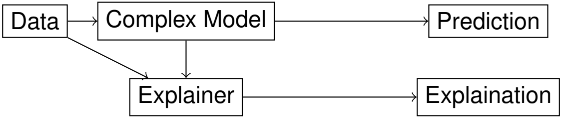
SHAP: SHapley Additive exPlanations

LIME:

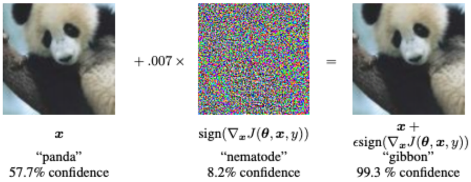
- generate artificial points around an observation
- local approximation by Linear Regression

SHAP:

- “generalization" of LIME
- local explainer is not LR
- more sophisticated model types for explainer



Adversarial Examples



"Explaining and Harnessing Adversarial Examples" Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy

Adversarial Examples

classified as turtle classified as rifle
classified as other

"Synthesizing Robust Adversarial Examples" Anish Athalye, Logan Engstrom, Andrew Ilyas, Kevin Kwok

Deduction

Summary:

- ML model is valid outside of the input

$\Rightarrow$ but often does not make much sense

- Current approaches to explainability: not really deduction

$\Rightarrow$ simplyfing “transformations” for single point

- Neural networks can be tricked

$\Rightarrow$ worrisome and shows how much “intuition" people have

- DL models can even be better than e.g., random forest

AI stories I

Google flu trends:

- 2008: paper in *Nature* claiming to beat Centers for Disease Control and Prevention
- 2013: misses the peak of the flu season by 140 percent

⇒ overfitting and missing changes in search behavior over time

- side note: data

IBM Watson for Oncology:

- data from doctor's notes (leakage, non-representative?), medical studies and clinical guidelines
- treatment recommendations are based on training by human overseers
- "through AI, [...] generate new insights and identify,[...] new approaches to cancer care"

AI stories II

$\Rightarrow$ Cancelled after unsafe treatment recommendations

- it is much easier to make prediction than suggest an action to change the outcome (counterfactual)
- side note: no scientific papers demonstrating how the technology affects physicians and patients

Kitano “Artificial Intelligence to Win the Nobel Prize and Beyond" (2016)

- human cognitive limitations
- 1 mln papers/year, some contradictory, inaccurate (partly language problem)
- explosion of experimental data
- hope of getting rid of bias
- discovery is beyond current knowledge

AI stories III

⇒ hypothesis generation and verification (robotics)

Playing games:

- is it really a proxy for intelligence?
- shown that if you play for 200 years your are better

Recommendation systems:

- successful on manipulating you to buy more things
- cannot explain, reason, convince you

Autonomous cars:

- failure of Volvo, ... and Tesla?

Summary

- DL models are just transformations: overattribution
- DL not able to generalize, explain, fill the gaps. Do not resemble scientific approach.
- Limitations hidden in data (hinder creation of really large data sets)

$\Rightarrow$ Complex vs. simpler models (Model-free, policy-based learning to help?)

- limits of (inefficient) “just learning from data” $\Rightarrow$ reasoning

Summary

- DL models are just transformations: overattribution
- DL not able to generalize, explain, fill the gaps. Do not resemble scientific approach.
- Limitations hidden in data (hinder creation of really large data sets)

$\Rightarrow$ Complex vs. simpler models (Model-free, policy-based learning to help?)

- limits of (inefficient) “just learning from data” $\Rightarrow$ reasoning

Dangers:

- over-hype
- trivialization of science: moving to problems that can be tackled with AI (recipe vs. understand)

Summary

- DL models are just transformations: overattribution
- DL not able to generalize, explain, fill the gaps. Do not resemble scientific approach.
- Limitations hidden in data (hinder creation of really large data sets)

$\Rightarrow$ Complex vs. simpler models (Model-free, policy-based learning to help?)

- limits of (inefficient) “just learning from data” $\Rightarrow$ reasoning

Dangers:

- over-hype
- trivialization of science: moving to problems that can be tackled with AI (recipe vs. understand) vs. trivialization of AI
- what we learn from solutions?
- already dealing with a replication crisis (black box models, questionable reproducibility, limited explainability, and lack of uncertainty quantification)

Summary

- DL models are just transformations: overattribution
- DL not able to generalize, explain, fill the gaps. Do not resemble scientific approach.
- Limitations hidden in data (hinder creation of really large data sets)

$\Rightarrow$ Complex vs. simpler models (Model-free, policy-based learning to help?)

- limits of (inefficient) “just learning from data” $\Rightarrow$ reasoning

Dangers:

- over-hype
- trivialization of science: moving to problems that can be tackled with AI (recipe vs. understand) vs. trivialization of AI
- what we learn from solutions?
- already dealing with a replication crisis (black box models, questionable reproducibility, limited explainability, and lack of uncertainty quantification)
- technical sustainability of brute-force-based progress

Thanks

j.rybicki@fz-juelich.de