The Reasonable Ineffectiveness of Data

Thomas Hellström
Department of Computing Science
Umeå University
SweDS2018
2018-11-20
Intelligent robotics @ Umeå University

Professor Thomas Hellström
Department of Computing Science
Umeå University

Machine Learning for
• Robot learning
• Natural Language Processing
• Object identification in images
This talk ... *The Reasonable Ineffectiveness of Data*

Two major approaches to learning about the world

The Model Driven approach

The Data Driven approach

---

**The Unreasonable Effectiveness of Mathematics in Natural Sciences**

*Eugene Wigner*

Hungarian-American

Nobel Prize in Physics in 1963

---

**The Unreasonable Effectiveness of Data**

*Peter Norvig*

American

Computer scientist

Director of research @ Google
The Model Driven Approach

- **Galileo Galilei**
  - One of the first to combine theoretical and experimental physics with mathematics
  - The Scientific Method: A mathematically formulated hypothesis about the world is tested with experiments: collecting and analyzing data
  - “the laws of nature are mathematical”

- **Physics can often be described with very simple equations**
  - \( s = \frac{1}{2}at^2 \)
  - \( f = ma \)
  - \( e = mc^2 \)
The Model Driven Approach

- Eugene Wigner
  - Hungarian-American theoretical physicist
  - Nobel Prize in Physics in 1963
- “The Unreasonable Effectiveness of Mathematics in the Natural Sciences”¹
  - Newton’s law of gravitation is accurate to less than a ten thousandth of a per cent.
  - In quantum mechanics they make fantastic discoveries by generalizing mathematical rules, generated from data
  - “the enormous usefulness of mathematics in the natural sciences is something bordering on the mysterious and that there is no rational explanation for it”.

Limitations With the Model Driven Approach

- Science that include human behavior is often resistant to elegant mathematics
  - Cognitive science
    - Speech recognition
    - Language understanding
      - An (incomplete) English grammar is more than 1700 pages long
  - Economics
  - Ethics
  - …
Machine translation

Traditional (model driven):

Model
1700+ pages of English grammar, and a German grammar

Deep Learning (data driven):

Data
10^9 pages of translated text

Much higher accuracy than state-of-the-art (2015)

I have a small dog
Ich habe einen kleinen Hund
The Data Driven Approach

“The Unreasonable Effectiveness of Data”¹

- State-of-the-art in speech recognition, machine translation, and image analysis are data driven.
- ”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.”

- This view is embraced in machine learning, not least in deep learning

Limitations With the Data Driven Approach

”The Reasonable Ineffectiveness of Data”
The Reasonable **Ineffectiveness** of Data

**Machine translation**

- MUCH better than 10 years ago
- However, the machines make mistakes no human would make
  - Some random Thai characters translates into:
    “There are six sparks in the sky, each with six spheres. The sphere of the sphere is the sphere of the sphere.”

- Do these machines UNDERSTAND language?

The Reasonable **Ineffectiveness** of Data

A system learns to generate image annotations from a database with images & annotations (>1M images)\(^1\)

---

A group of young people playing Frisbee

A person riding a motorcycle on a dirt road

A refrigerators filled with lots of food and drinks

**Much better than state-of-the-art**

But does the program **UNDERSTAND** in any sense?

---

The **Reasonable** Ineffectiveness of Data

- So is this observed ineffectiveness reasonable and even expected?
- Yes, and it is a consequence of a purely data-driven approach which leads to
  - Finding correlations by chance
  - Confusing correlation with causation
  - Inability to identify causation
Finding correlations by chance

- “Data snooping”

- “If you torture data long enough it will confess to anything”

- Correlations and patterns only exist in the examined data
- Especially problematic if data is big AND limited (e.g. economy data)
Finding correlations by chance

- Thousands of statistical time series in economy and politics
- All of them stop at 2018 and don’t go very far back
Confusing correlation with causation

- **Data**: HDL (‘good’) cholesterol is negatively correlated with heart attacks.
- **(incorrect) Conclusion**: Taking medication to raise HDL decreases the risk of getting a heart attack.
- Further research (experiments) showed that
  - Exercise, Genes, Diet,... affect *both* HDL levels and the likelihood of having a heart attack
  - This is manifested as the observed correlation
  - Medication to increase HDL may even increase the risk
- Data alone could not answer what would happen if we increase HDL
- **Randomized Controlled Trials (RCT)** is a common technique in medicine

*NIH stops clinical trial on combination cholesterol treatment, National Institute of Health, 2011.*
Inability to identify causation

- Data alone cannot identify causation and answer questions such as “What if ...”
- Deep Learning normally only works with correlations
- That’s why the program thinks this picture is a “refrigerators filled with lots of food and drinks”
- We need to incorporate understanding in our solutions
  - Judea Pearl introduced do-calculus and uses causal diagrams
    - $X$ causes $Y$ if $P(Y \mid do(X)) > P(Y)$
  - Hybrid solutions

Judea Pearl, Dana Mackenzie, The Book of Why: The New Science of Cause and Effect, 2018
The Reasonable Ineffectiveness of Data

SUMMARY

- Problems with a purely data-driven approach
  - Finding correlations by chance
    - Caused by the huge amount of data
  - Confusing correlation with causation
    - Not so strange since correlations often IS causation
  - Inability to identify causation
    - There is no general way to identify causation from data only
      - Understanding of the problem is required!
    - For this, models AND data are necessary