The Reasonable Ineffectiveness of Data

Thomas Hellström

Department of Computing Science

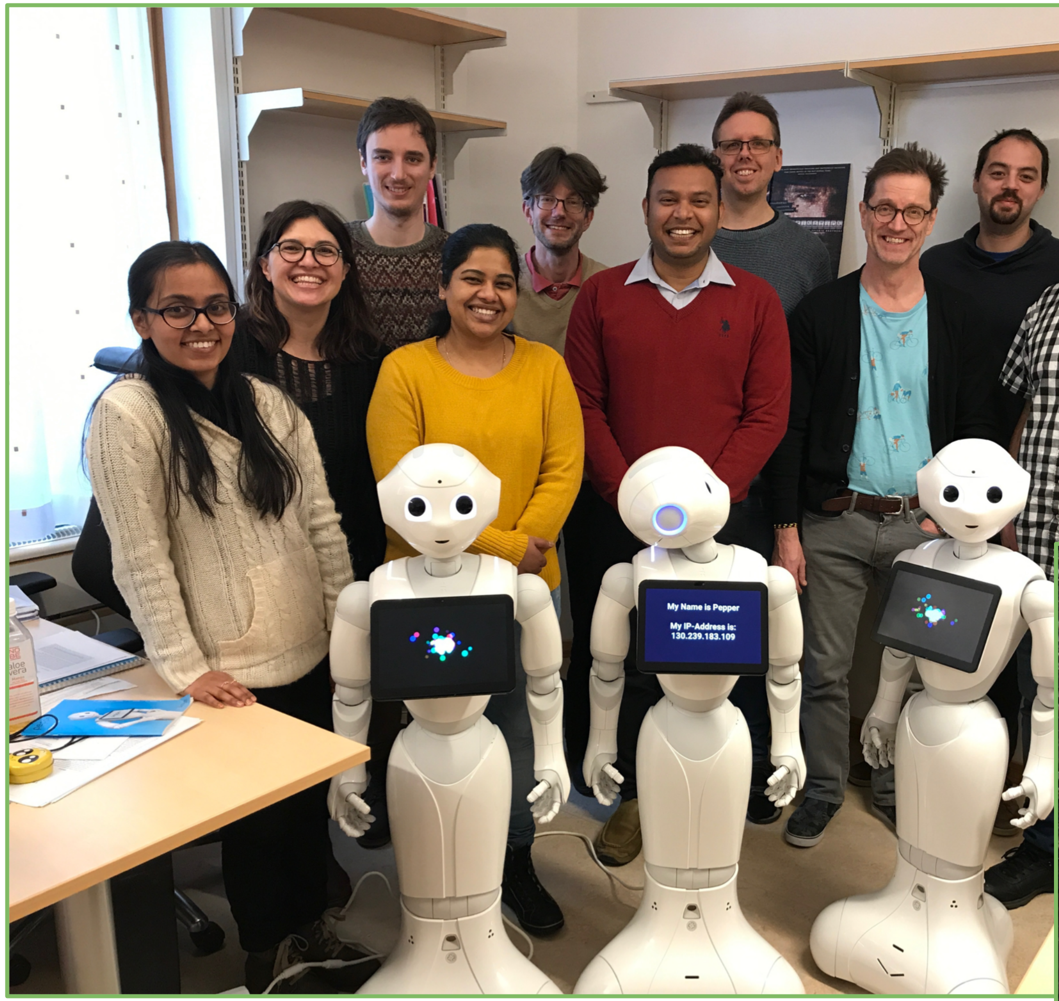
Umeå University

SweDS2018

2018-11-20

UMEÅ UNIVERSITY

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 Unreasonable Effectiveness of Mathematics in Natural Sciences

Eugene Wigner
Hungarian-American
Nobel Prize in Physics in 1963

The Data Driven approach

The Unreasonable Effectiveness of Data

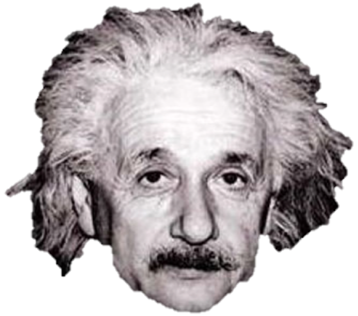
Peter Norvig
American
Computer scientist
Director of research @ Google

The Model Driven Approach

Galileo Galilei
Italian
Father of modern physics

- 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
  - $s = at^2/2$
  - $f = ma$
  - $e = mc^2$

Isaac Newton
British
Founded classical mechanics & more



Albert Einstein
German/American
Theory of Relativity

The Model Driven Approach

Eugene Wigner
Hungarian-American
Nobel Prize in Physics in 1963

- 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".

Bargmann–**Wigner** equations
Wigner D-matrix
Wigner–Eckart theorem
Wigner friend
Wigner semicircle distribution
Wigner classification
Wigner distribution function
Wigner quasiprobability distribution
Wigner crystal
Wigner effect
Wigner energy
Wigner lattice
Relativistic Breit–**Wigner** distribution
Modified **Wigner** distribution function
Wigner–d'Espagnat inequality
Gabor–**Wigner** transform
Wigner theorem
Jordan–**Wigner** transformation
Newton–**Wigner** localization
Wigner–Inonu contraction
Wigner–Seitz cell
Wigner–Seitz radius
Thomas–**Wigner** rotation
Wigner–Weyl transform
Wigner–Wilkins spectrum

[1] E. Wigner, "The Unreasonable Effectiveness of Mathematics in the Natural Sciences," *Comm. Pure and Applied Mathematics*, vol. 13, no. 1, 1960, pp. 1–14.

Limitations With the Model Driven Approach

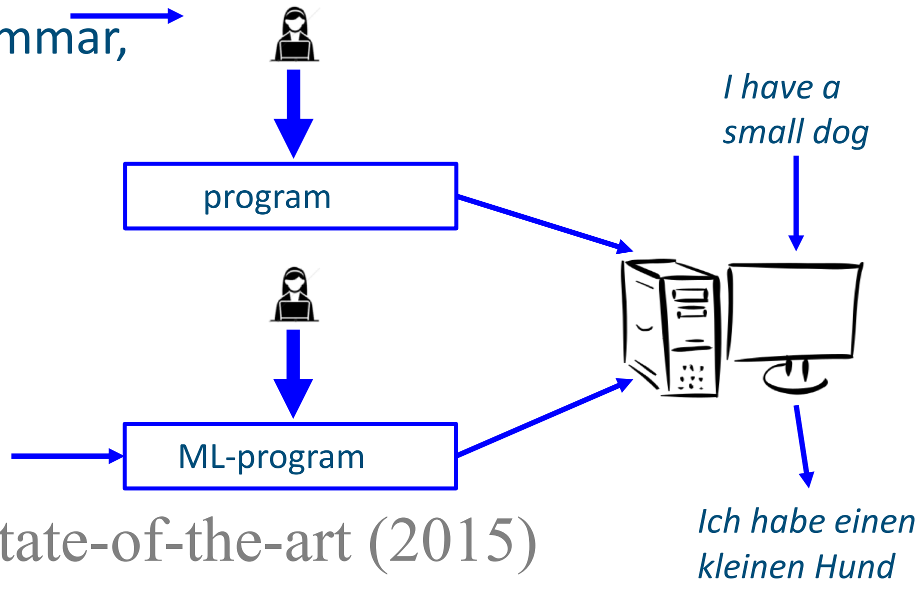
- Science that include human behavior is often resistant to elegant mathematics
  - Cognitive science
    - Speech recognition
    - Language understanding
      - An (incomplete) English grammar is more than 1700 pages long
  - Economics
  - Ethics
  - …

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)

The Data Driven Approach

Peter Norvig

Director of research @ Google

American

"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

[1] Alon Halevy, Peter Norvig, and Fernando Pereira, Google ,The Unreasonable Effectiveness of Data, Intelligent Systems, IEEE 24(2):8 – 12, 2009

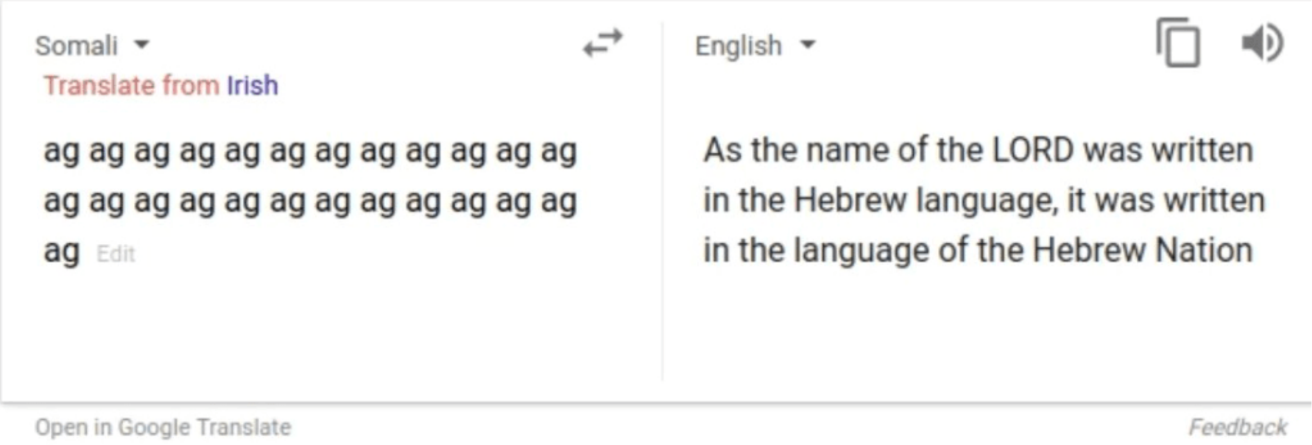
Limitations With the Data Driven Approach

”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."



ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag ag

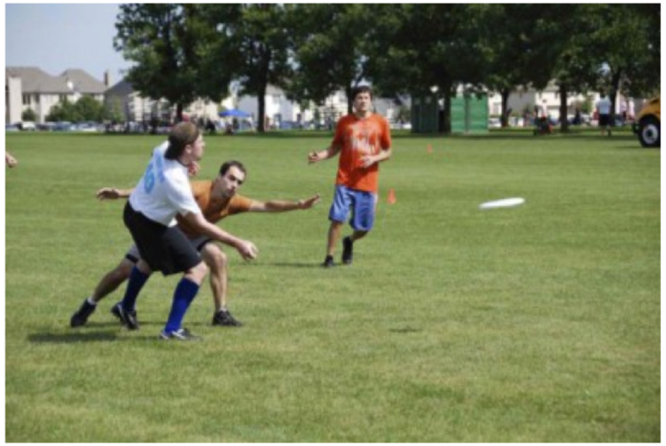
As the name of the LORD was written in the Hebrew language, it was written in the language of the Hebrew Nation

- Do these machines UNDERSTAND language?

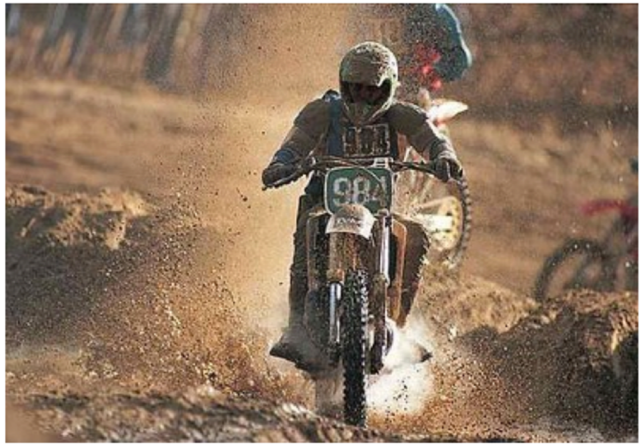
Gomes, Lee (July 22, 2010). "Google Translate Tangles With Computer Learning". Forbes.

The Reasonable **Ineffectiveness** of Data

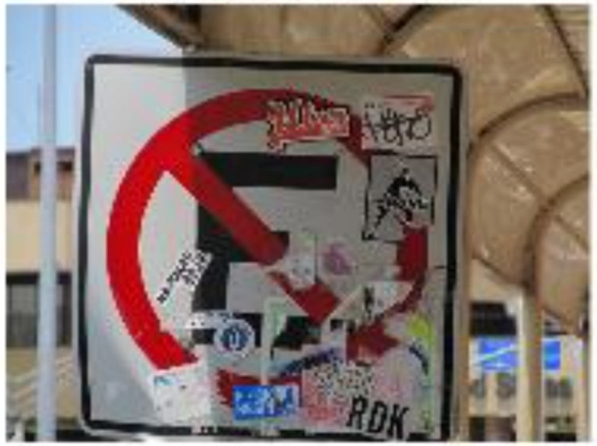
A system learns to generate image annotations from a database with images & annotations (>1M images)[1]



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?

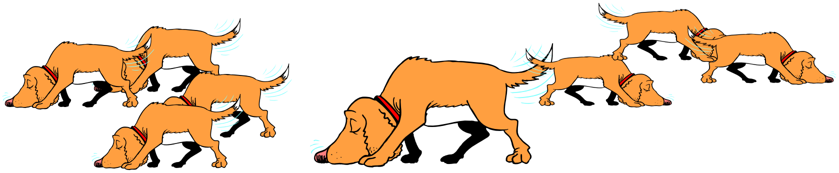
[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, *Show and Tell: A Neural Image Caption Generator* , Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP'15).

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”



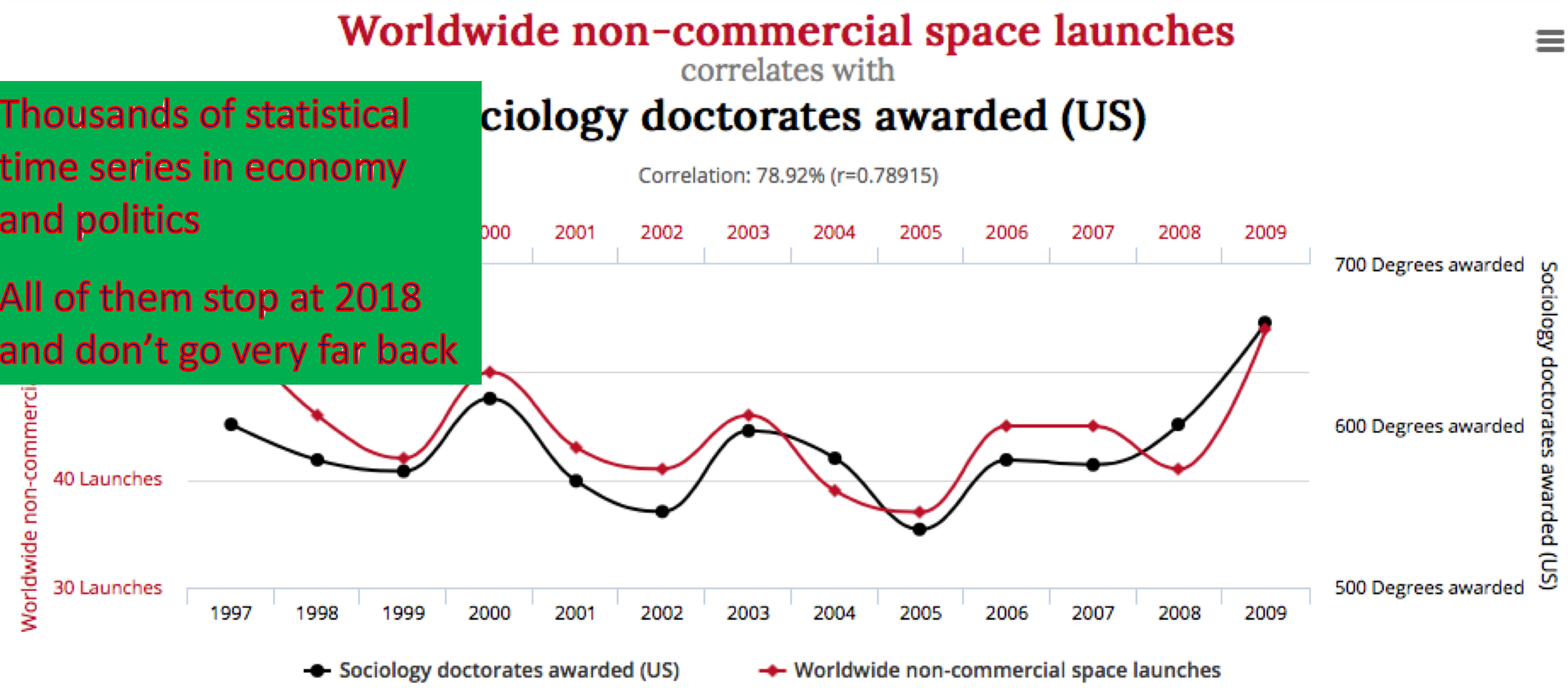
- “If you torture data long enough it will confess to anything”

OUR CONSULTANT HAS BEEN MINING DATA ALL DAY.
THE RESULTS ARE QUITE SHOCKING.
www.dilbert.com scottadams@aol.com
ACCORDING TO THE DATA, SALES ARE ALWAYS HIGHEST WHEN I DO THIS...
5/5/00 © 1999 United Feature Syndicate, Inc.

- 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

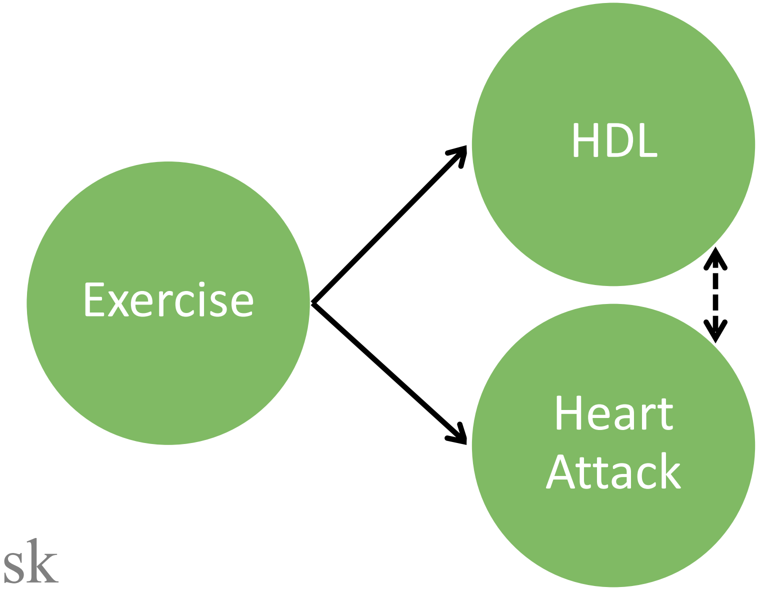
- Thousands of statistical time series in economy and politics
- All of them stop at 2018 and don't go very far back



Tyler Vigen, www.tylervigen.com

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

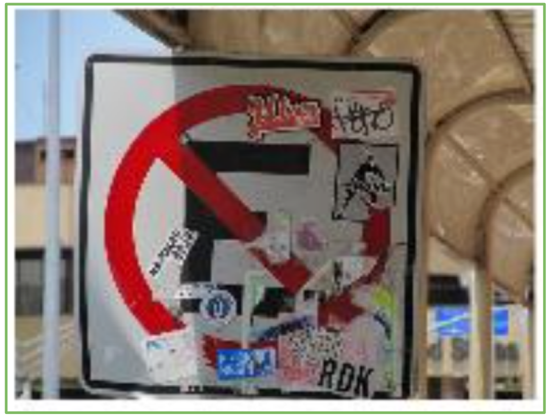


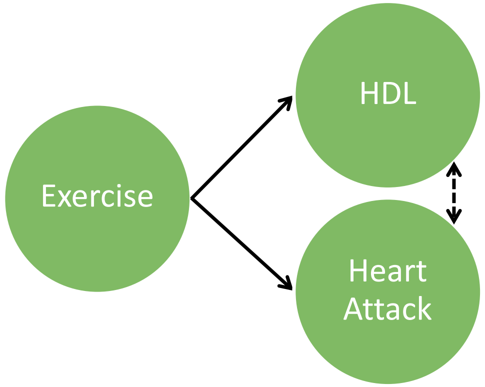
NIH stops clinical trial on combination cholesterol treatment, National Institute of Health, 2011. https://www.nih.gov/news-events/news-releases/nih-stops-clinical-trial-combination-cholesterol-treatment

Inability to identify causation

Judea Pearl
Israeli-American
Computer scientist
2011 winner of the ACM Turing Award

- 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

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