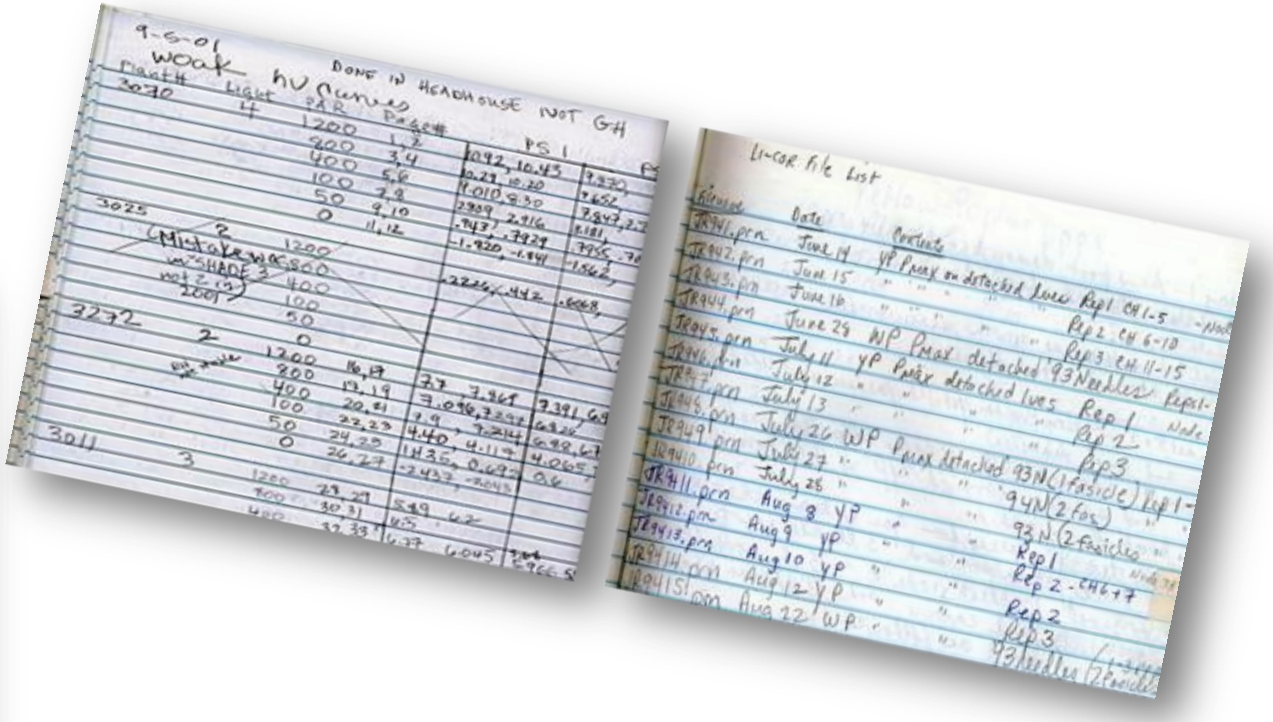
Science is all about the data

scientific method

▸ Definitions

noun

a method of investigation in which a problem is first identified and observations, experiments, or other relevant data are then used to construct or test hypotheses that purport to solve it



So why the scepticism to Data Science?

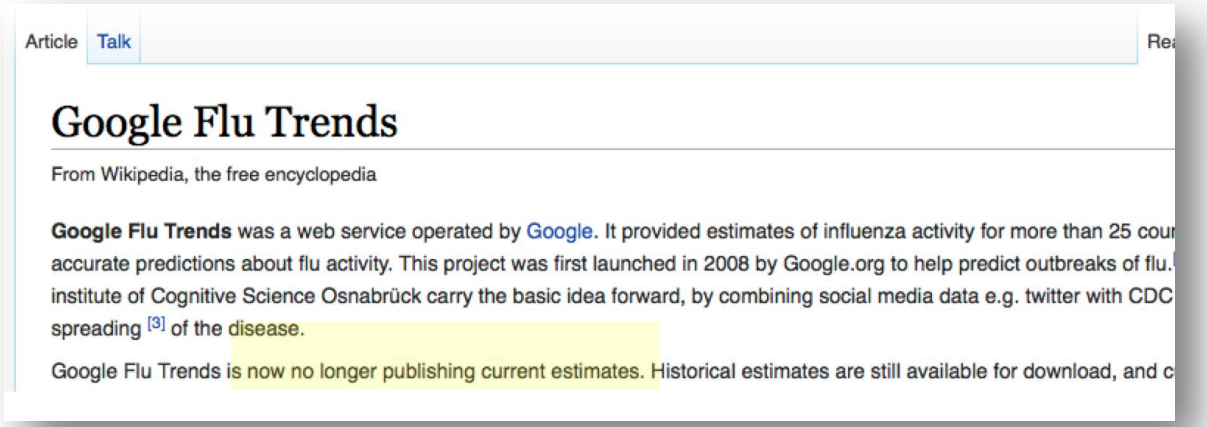
Real scientists do:

- Application of Physics
- Generation of hypotheses
- Experimental Testing
- Statistical analysis

Data scientists do:

- Letting the data speak
- Finding relationships without first creating a hypothesis

Sometimes data science does get it wrong



Article Talk

Google Flu Trends

From Wikipedia, the free encyclopedia

Google Flu Trends was a web service operated by Google. It provided estimates of influenza activity for more than 25 cou
accurate predictions about flu activity. This project was first launched in 2008 by Google.org to help predict outbreaks of flu.
institute of Cognitive Science Osnabrück carry the basic idea forward, by combining social media data e.g. twitter with CDC
spreading [3] of the disease.

Google Flu Trends is now no longer publishing current estimates. Historical estimates are still available for download, and c

Home News Journals Topics Careers

Science Science Advances Science Immunology Science Robotics Science Signaling Science Translational Medicine

SHARE

POLICY FORUM | BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer[1,2,*], Ryan Kennedy[1,3,4], Gary King[3], Alessandro Vespignani[5,6,3]

+ Author Affiliations

*Corresponding author. E-mail: d.lazer@neu.edu.

Science 14 Mar 2014:
Vol. 343, Issue 6176, pp. 1203-1205
DOI: 10.1126/science.1248506

Article Figures & Data Info & Metrics eLetters PDF

You are currently viewing the summary.

View Full Text

Summary

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

“Scientific” data science

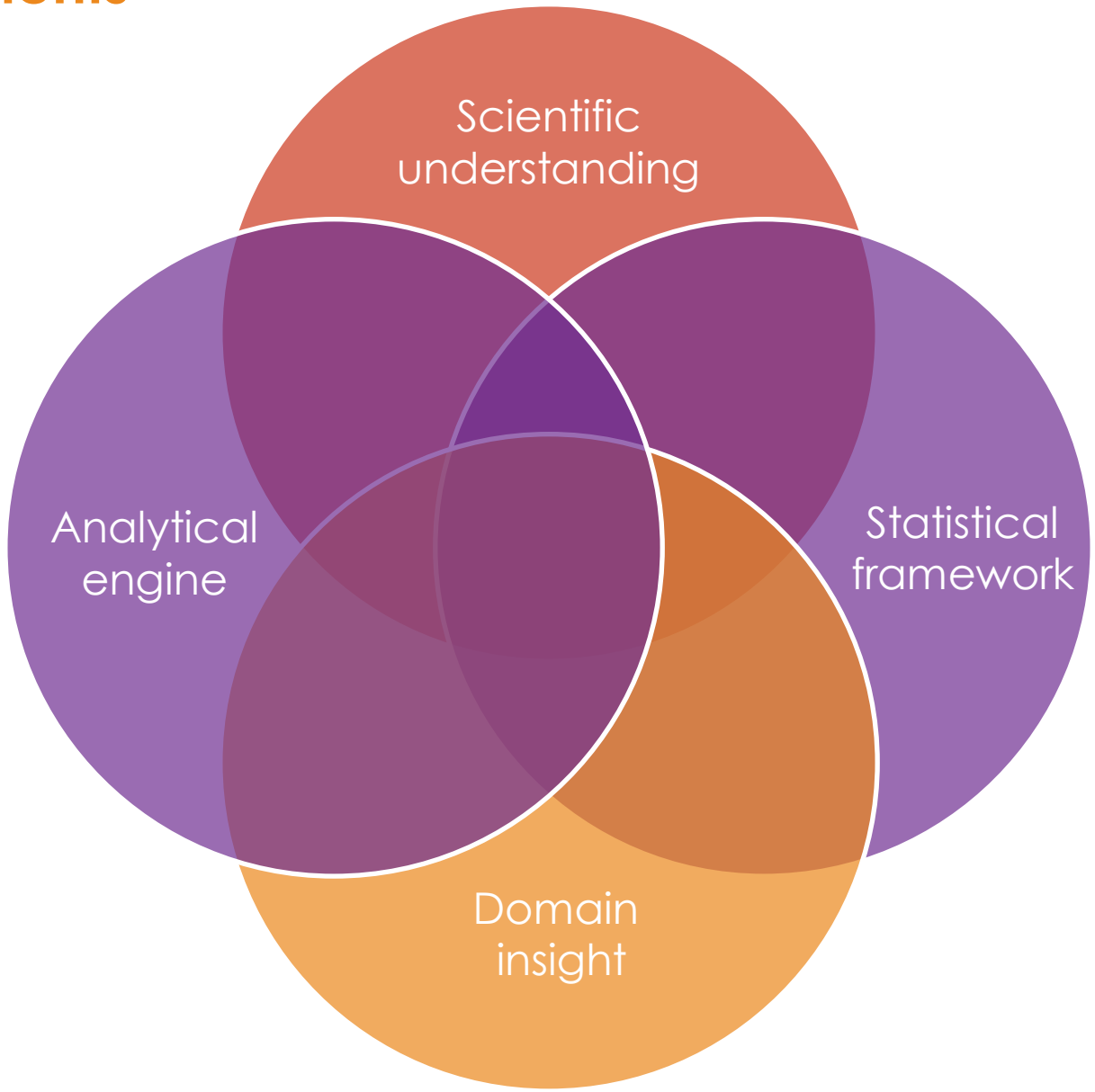
- You **don’t** need to
  - throw physics and causality out the window
- You **do** need to
  - get the data out of the application silos and unwieldy file formats
- You probably need to
  - understand (and improve) your data quality
- You probably have a **lot** of data
  - If so, architecture for your analytics is important
  - Look at best storage and data model approaches
  - Consider how to run your analytics in parallel

“Scientific” data science - 2

- You probably have a **lot** of variables, and a lot of interaction
  - One of your first analytical steps is likely to be to reduce the complexity
  - You have a few options here
    - Defining higher order features manually based on understanding of the data – giving a small enough set to run simple techniques like regression
    - Use a Clustering technique to reduce the number of examples
    - Use Principal Component Analysis or similar to reduce the number of dimensions

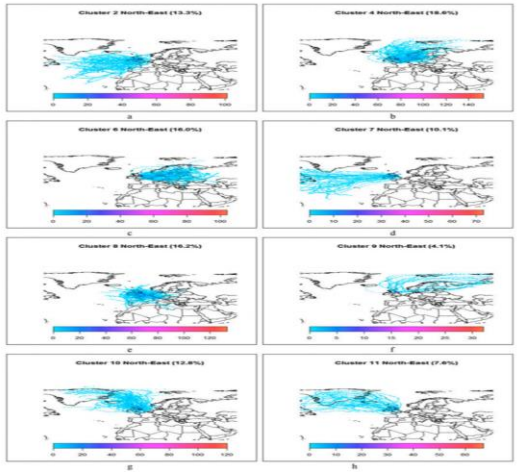
- THEN you can
  - Use analytical techniques to discover relationships in your data
  - Learn new things about your area of research
  - Identify new optimisation or prediction opportunities

The magic ingredients



Predicting the Weather

- Very good science in a supercomputer simulation
- Very good measurements come from sparse weather stations
- Can the science and the observations be compared to generate answers to:
  - When did we last see this?
  - What happened next?
  - And crucially, what was the impact?



Predicting the quality of 4D seismic