Big Data and Security

Jeffrey Borowitz, PhD

Lecturer

Sam Nunn School of International Affairs

Google Flu Trends

Google Flu Trends

Illustrative example

- Why is it big data?
- What is the new benefit?
- What are the key assumptions?
- How did the assumptions work out?

Flu Trends Product

- Predicted flu based on search queries

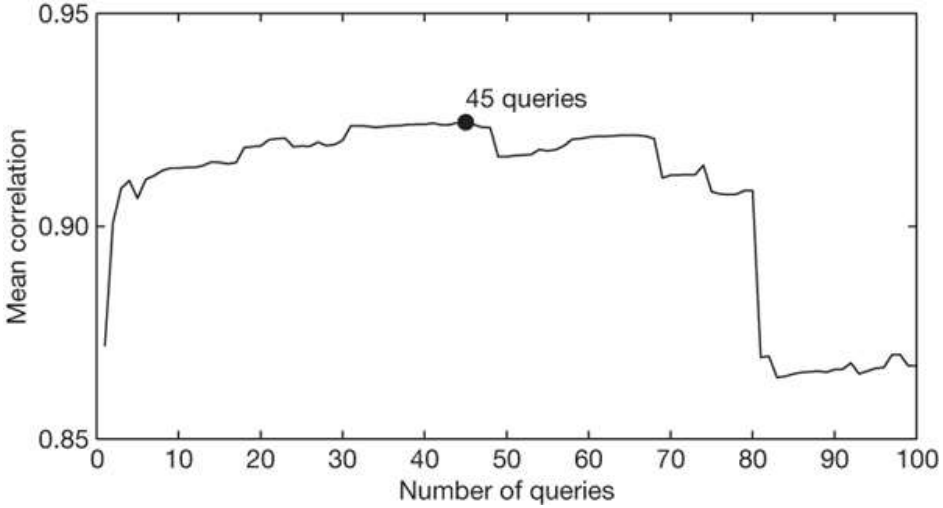
How It Works

- Original paper (2009)
- Start with CDC’s system of cataloguing Influenza-Like-Illness (ILI)
  - Weekly, state-level number of incident counts
- Calculate weekly, state-level search term prevalence
  - Used top 50 million terms
- Tried a lot of combinations of these terms to pick the group which best matched inputs

Georgia Tech

Number of Top Queries and Correlation

An evaluation of how many top-scoring queries to include in the ILI-related query fraction.



J Ginsberg *et al. Nature* **000**, 1-3 (2008) doi:10.1038/nature07634
https://www.nature.com/articles/nature07634#MOESM267

nature

Georgia Tech

Issues

- Had trouble in 2008-2009 with the onset of H1N1, and model was refit
- Had trouble in 2012-2013 season due to high press coverage of flu

Assumptions and Issues

- Flu trends depends on CDC's ILI reporting
  - There's no Google Flu Trends without actual measured flu trends
- The relationship between search behavior and flu is the same over time
  - Google grew quickly from 2003-2008

Year	Searches
2000	22,000,000,000
2007	438,000,000,000
2008	637,200,000,000
2009	953,700,000,000

- A more detailed discussion of issues with flu trends: Big Data Traps.pdf in this week's readings

Georgia Tech

Epilogue

- In 2015, it was shut down and data was given to researchers
- There are continuing experiments to incorporate this, but CDC doesn’t use it

Summary

- Flu Trends
  - Why is it big data?
  - What is the new benefit?
  - What are the key assumptions?
  - How did the assumptions work out?
- Flu Trends represents an interesting and useful new product of Big Data, but shows how insights and products built on big data might be fragile