



# Asking the right questions

Getting the right answers from your data



Data After Dark  
OHSU BD2K Data Science Workshop

**Shannon McWeeney, PhD**

**13<sup>th</sup> January 2016**

# **Importance of Question + Scope**

Guidelines for Study Design

# Study Design

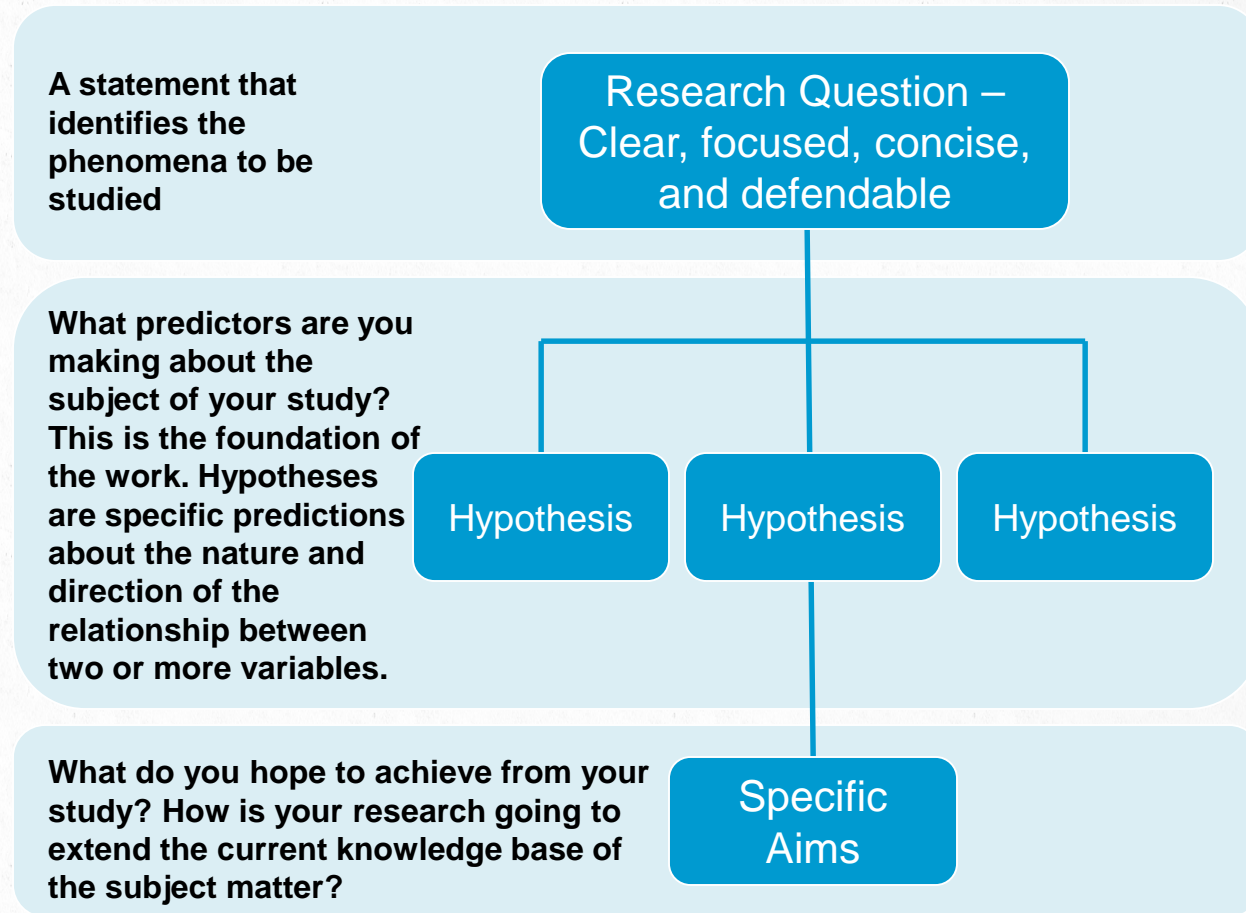
- Define Question
- Type of Study / Method of data Generation
- Sampling Considerations

# Defining Your Question

- Why is this research important?
- What is it that we don't know or fully understand?
- What have other researchers in my field done?
- What areas need further exploration?
- Can my study help fill in these gaps or lead to greater understanding?

# From Question->Hypothesis

- Characteristics of a good hypothesis
  - Gives insight into the proposed research question;
  - Is measurable and testable;
  - Is developed directly from the experiences of the researcher and should have a well-founded rationale for all proposed hypotheses.



# Guidelines for Good Design

- **Clarity**
  - Clear hypothesis.
  - **Does your question match your analysis method?**
- **Simplicity**
  - Multiple Questions
  - Data Snooping

# Guidelines for Good Design

- **Confounding Factors**
  - Distinguish between variation of interest and other sources of variation.
- **Replicates**
  - Type of replicates
  - Can you detect the effect if it is present?



# Assessing Significance

- $P$  value (R.A. Fisher): Informal way to judge whether evidence was “significant” (i.e., worthy of a second look)
  - Formulate ‘Null hypothesis’
  - Set up statistical test assuming null hypothesis is true
  - Calculate the chances of getting results at least as extreme as what was actually observed. This probability =  $P$  value.
  - Smaller  $P$  value = greater the likelihood that the straw-man null hypothesis was false
  - **Context**: Part of the research process / life-cycle



# Evidence Based Decisions

- Rigorous and Objective Framework
  - Key Concepts:
    - Statistical Power
    - Estimation of False positives and False Negatives
    - Explicit statements about effect size and variability

**This framework was incorrectly hybridized with P-value concept.**

**“THE P VALUE WAS NEVER MEANT TO BE USED THE WAY IT’S USED TODAY”**

# What we **should** be asking

- What are the odds that a hypothesis is correct?
  - Depends on how plausible the hypothesis is in the first place.

# Rephrasing the Question

## PROBABLE CAUSE

A  $P$  value measures whether an observed result can be attributed to chance. But it cannot answer a researcher's real question: what are the odds that a hypothesis is correct? Those odds depend on how strong the result was and, most importantly, on how plausible the hypothesis is in the first place.

■ Chance of real effect  
■ Chance of no real effect

### Before the experiment

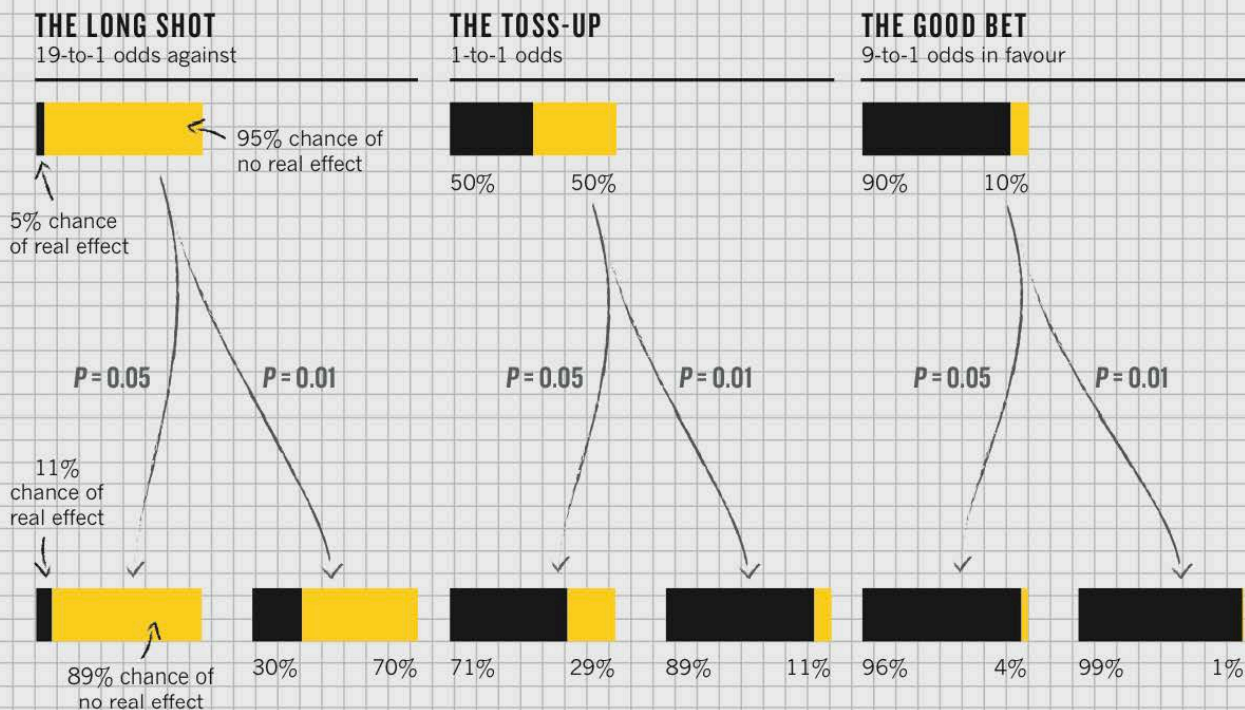
The plausibility of the hypothesis — the odds of it being true — can be estimated from previous experiments, conjectured mechanisms and other expert knowledge. Three examples are shown here.

### The measured $P$ value

A value of 0.05 is conventionally deemed 'statistically significant'; a value of 0.01 is considered 'very significant'.

### After the experiment

A small  $P$  value can make a hypothesis more plausible, but the difference may not be dramatic.



# Key Danger

- “P-hacking” : “is trying multiple things until you get the desired result, even unconsciously
  - Monitoring data while it is being collected
  - Exploratory studies confused with confirmatory studies

# Three Questions

- What is the evidence?
- What should I believe?
- What should I do?



# Case Study

Google Flu



# What Do we have here?

- A patient comes into your office in January with the following complaints:
  - Body aches, muscle and joint pain, headache, a sore throat and a unproductive cough with occasionally harsh breathing
  - Fever, which ranged from 100 to 104 F and lasted for a few days
  - Felt sudden dizziness, weakness and pain while at work
  - Constipation
  - Bloody nose, red mucous membranes
  - Family members have noted he is “not acting like himself”

What is a possible diagnosis?



# What If I told you....

- The year is not 2016 but 1918
  - All of those symptoms were what was being reported in medical literature at the time
  - Excerpts from JAMA, 10/3/1918 and 1/25/1919
  - At time, most basic clinical guideline was the temperature
  - Other Data collected:
    - Pulse rate "the pulse was remarkably slow," (JAMA, 4/12/1919)
    - Respiration rate
    - White blood cells counts



# Estimated Mortality Rate

Untreated Plague 100%

Untreated Anthrax 90%

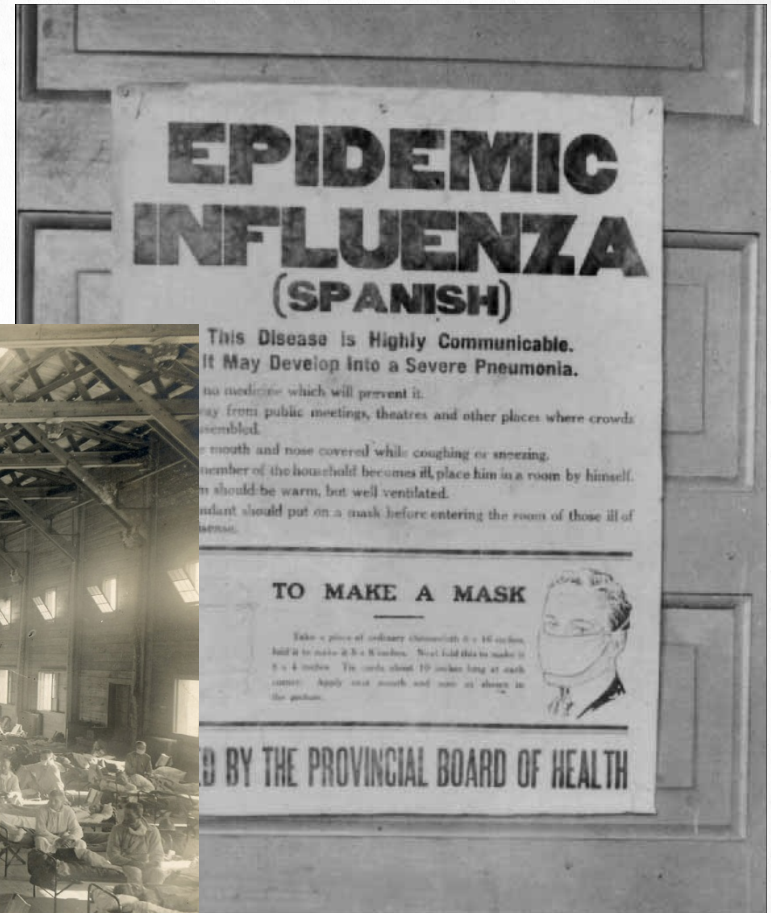
Smallpox 30%

Spanish Influenza 2.5%





# 1918 Spanish Influenza



<https://commons.wikimedia.org/wiki/File:SpanishFluPosterAlberta.jpg>



# Traditional Surveillance

**CDC** Centers for Disease Control and Prevention  
CDC 24/7: Saving Lives, Protecting People™

SEARCH

CDC A-Z INDEX ▾

## Influenza (Flu)

### Seasonal Influenza (Flu)

2014-2015 Flu Season +

Influenza - Flu Basics +

Prevention - Flu Vaccine +

Treatment - Antiviral Drugs +

Specific Groups +

Questions & Answers +

Health Professionals +

Resources for Flu Prevention Partners +

Flu Activity & Surveillance -

Situation Update: Summary of Weekly FluView

Overview of Influenza Surveillance in the United States

Seasonal Influenza (Flu)

### Flu Activity & Surveillance

Recommend Tweet Share

#### FluView Weekly U.S. Influenza Surveillance Report



A weekly influenza surveillance report prepared by the Influenza Division. All data are preliminary and may change as more reports are received.

More >

#### Summary of Weekly FluView



A brief overview of flu activity in the United States highlighting key data points from the weekly influenza surveillance report, FluView.

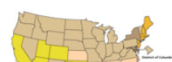
More >

#### FluView Interactive



Influenza surveillance data the way you want it. This series of dynamic

#### Current United States Flu Activity Map



The influenza activity reported by state and territorial epidemiologists indicates geographic spread of

*Flu Activity & Surveillance. (2016). Retrieved May 23, 2016, from <http://www.cdc.gov/flu/weekly/fluactivitysurv.htm>*

- US Centers for Disease Control and Prevention (CDC) and European Influenza Surveillance Scheme (EISS)
  - Rely on both virological and clinical data, including influenza-like illness (ILI) physician visits.

CDC publishes national and regional data from these surveillance systems on a weekly basis, typically with a 1–2-week reporting lag





Season: 2014-15 ◀ 40 50 1 9 ▶

weeks

☒ National ☐ HHS Regions ☐ Census Divisions

Choose components to view

☒ Map ☒ Line Chart ☒ Pie Chart ☒ Column Chart

Download Data

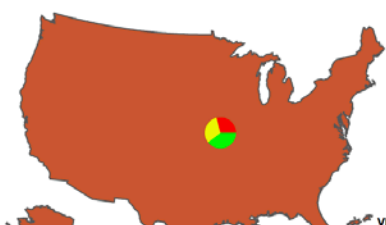
H...

Map

Influenza Positive Tests Reported to CDC and ILI Activity, National Summary, 2014-15 Season, week ending Mar 06, 2015

Reported by: U.S. WHO/NREVSS Collaborating Laboratories and ILINet

Download Image



Piechart Legend

(Piechart includes data from most recent 3 weeks of season)

- A (H1) - 0
- A (Unable to Subtype) - 0
- A (H3) - 2002
- 2009 H1N1 - 25
- A (Subtyping not Performed) - 2151
- B - 2723
- H3N2v - 0

Pie Chart WHO/NREVSS

Influenza Positive Tests Reported to CDC, National Summary, 2014-15 Season, weeks ending Feb 14, 2015 - Mar 06, 2015

Reported by: U.S. WHO/NREVSS Collaborating Laboratories and ILINet

Download Image



Number of Influenza Positive Tests

- A (H1) - 0
- A (Unable to Subtype) - 0
- A (H3) - 2002
- 2009 H1N1 - 25
- A (Subtyping not Performed) - 2151
- B - 2723
- H3N2v - 0
- No Data

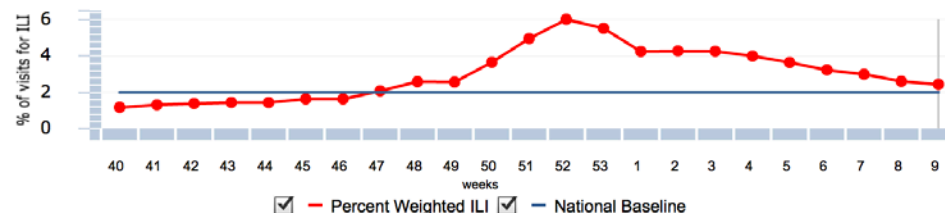
Line Chart ILINet

Percentage of visits for ILI, National Summary, 2014-15 Season through Mar 06, 2015

Reported by: U.S. WHO/NREVSS Collaborating Laboratories and ILINet

Download Image

Download Data



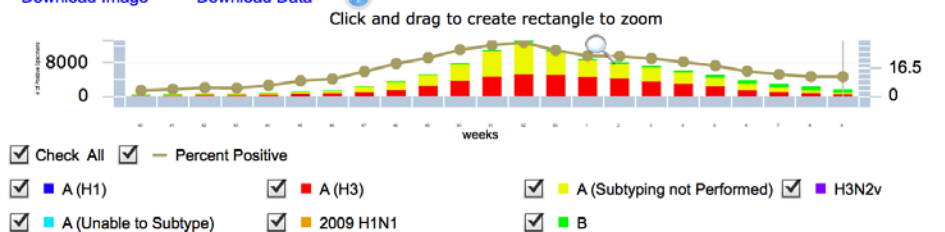
Stacked Column Chart WHO/NREVSS

Influenza Positive Tests Reported to CDC, National Summary, 2014-15 Season through Mar 06, 2015

Reported by: U.S. WHO/NREVSS Collaborating Laboratories and ILINet

Download Image

Download Data



<http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html>





[Google.org home](#)

[Dengue Trends](#)

**Flu Trends**

Home

Select country/region

[How does this work?](#)

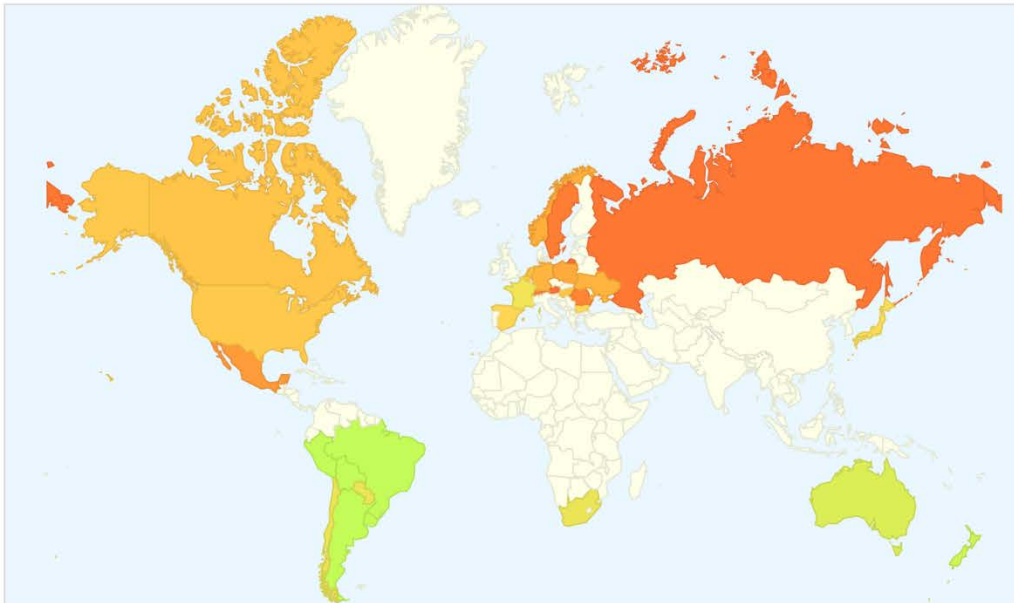
[FAQ](#)

**Flu activity**

Intense  
High  
Moderate  
Low  
Minimal

## Explore flu trends around the world

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. [Learn more »](#)



[Download world flu activity data](#) - [Animated flu trends for Google Earth](#) - [Compare flu trends across regions in Public Data Explorer](#)

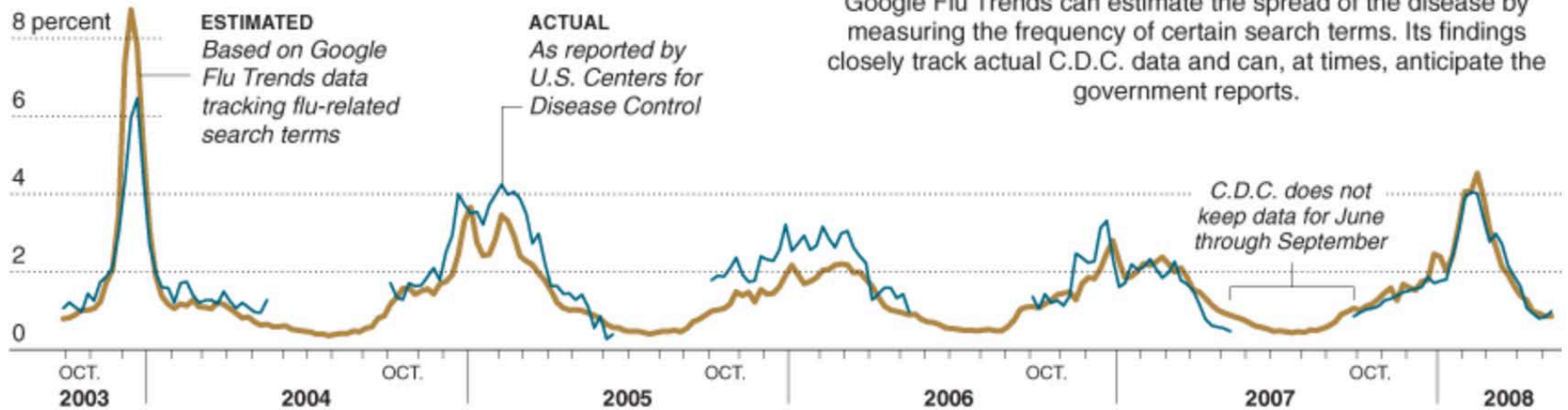
<https://www.google.org/flutrends/>



# Problems Amenable to Analytics

When do we need Big Data

**PERCENT OF HEALTH VISITS FOR FLU-LIKE SYMPTOMS** *Mid-Atlantic region*



Sources: Google; Centers for Disease Control

THE NEW YORK TIMES



# Google Flu Approach

- 5 years (2003-2008) of Google web search logs for modeling
- Time series of weekly counts: 50 million most common search queries (US only)
- No information about the identity of any user was retained
- Normalized by dividing the count for each query in a particular week by the total number of online search queries submitted in that location during the week (query fraction)
- Used the public historical CDC Influenza Sentinel Provider Surveillance Network data
  - Reported influenza-like illness (ILI) physician visits



find this original figure

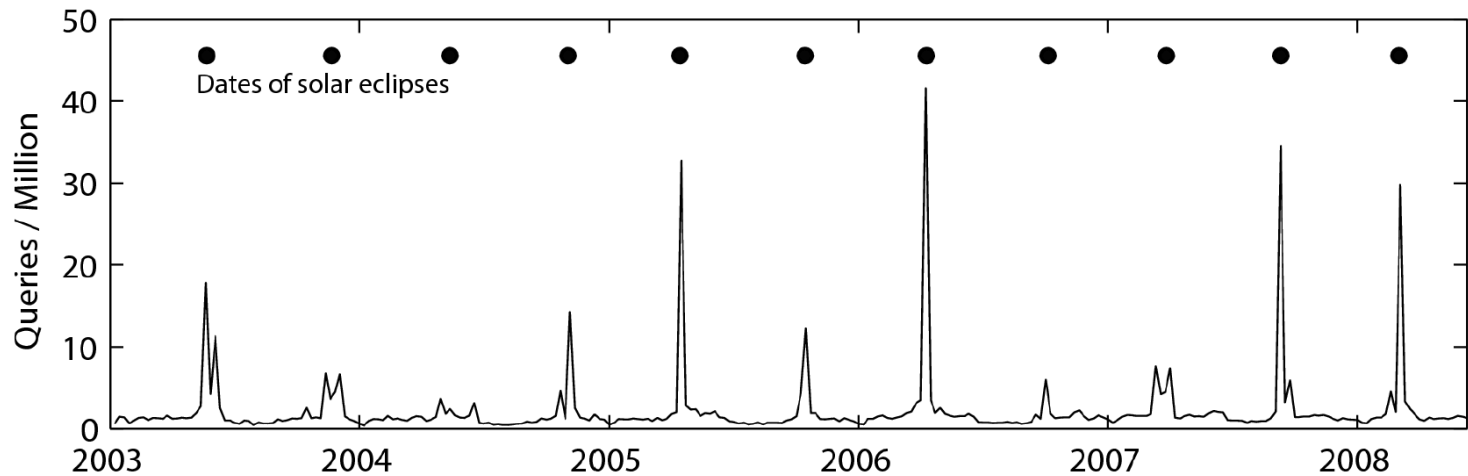


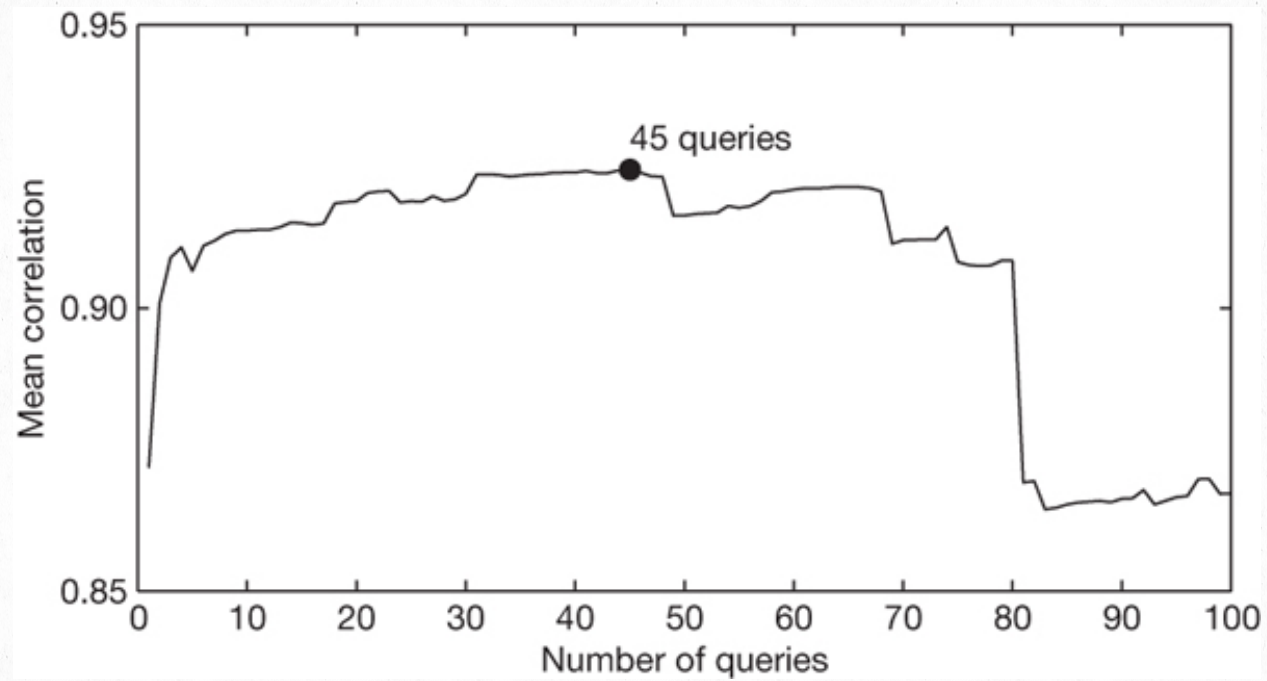
Figure 1: Weekly frequency of the search query “solar eclipse” in the United States from January 2003 to May 2008 and occurrences of solar eclipses, indicated by black dots.

# Automated Approach

- Requires no previous knowledge about influenza
- Measure how effectively model would fit the regional CDC ILI data if they used only a single query as the explanatory variable,  $Q(t)$
- Each of 50 million candidate queries was separately tested, to identify search queries which most accurately modeled the CDC regional ILI visit %
- Approach rewarded queries that showed regional variations similar to the regional variations in CDC ILI data
  - Motivation: the chance that a random search query can fit the ILI percentage in all nine regions is considerably less than the chance that a random search query can fit a single location



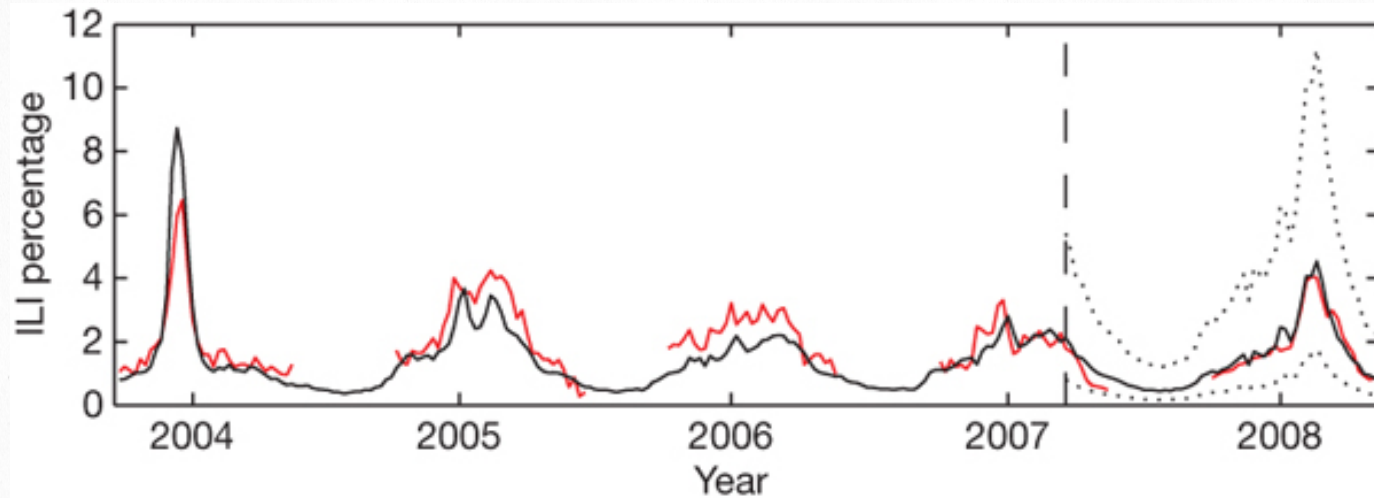




J Ginsberg *et al.* *Nature* **000**, 1-3 (2008) doi:10.1038/nature07634

**nature**





J Ginsberg *et al.* *Nature* **000**, 1-3 (2008) doi:10.1038/nature07634

**nature**

Model estimates for the mid-Atlantic region (black)  
CDC-reported ILI percentages (red)



# Google's Flu Project Shows the Failings of Big Data

<http://time.com/23782/google-flu-trends-big-data-problems/>

## Data Fail! How Google Flu Trends Fell Way Short

Posted: 03/16/2014 8:12 pm EDT | Updated: 03/16/2014 8:59 pm EDT

### Google Flu Trends gets it wrong three years running

- › 18:00 13 March 2014 by [Hal Hodson](#)
- › For similar stories, visit the [Bird Flu](#) Topic Guide

PHARMA & HEALTHCARE 3/23/2014 @ 9:00AM | 47,878 views

## Why Google Flu Is A Failure

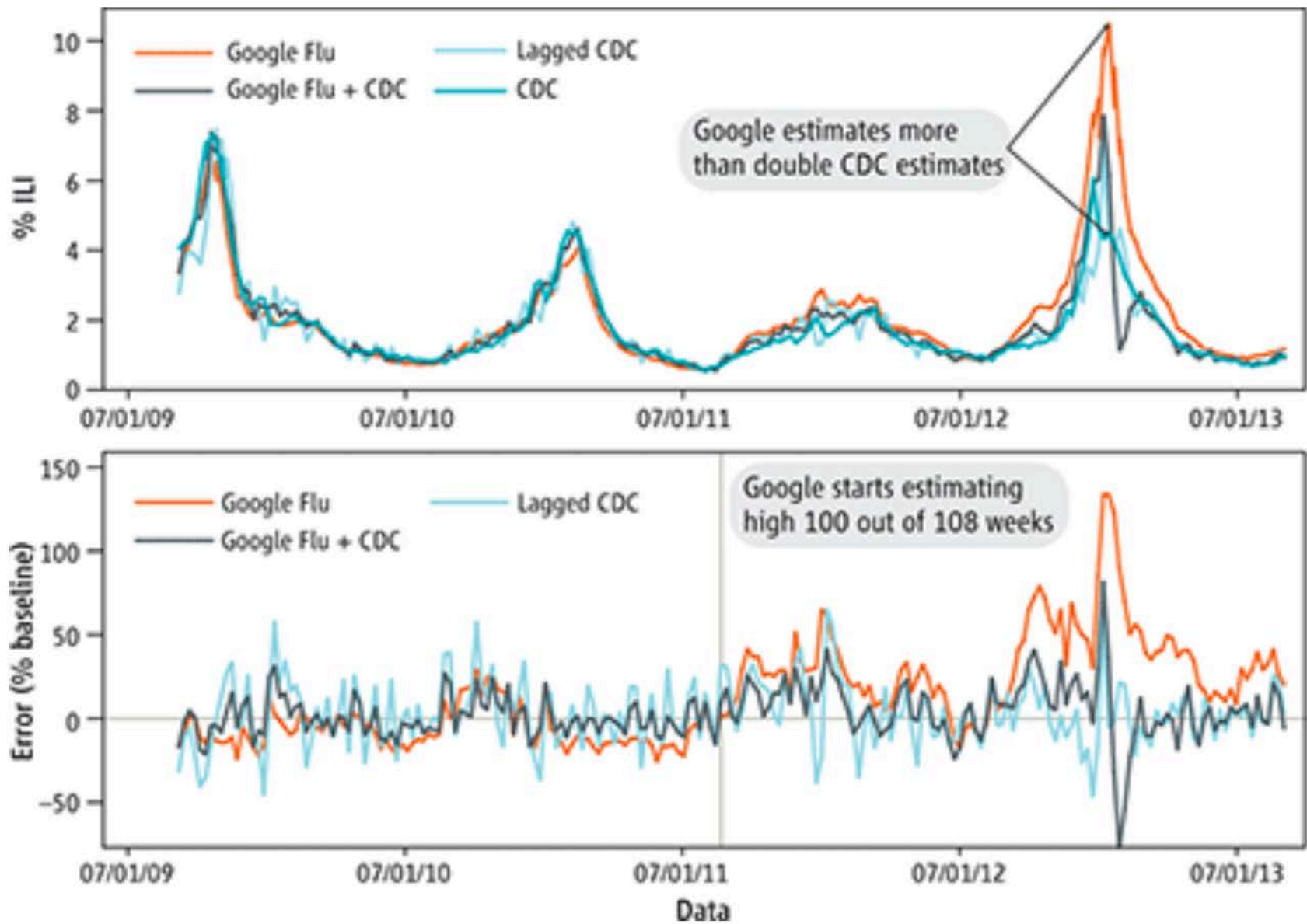


# "Houston we have a problem...."

Combining the  $n = 45$  highest-scoring queries was found to obtain the best fit. These 45 search queries, although selected automatically, appeared to be consistently related to ILIs. Other search queries in the top 100, not included in our model, included topics like 'high school basketball', which tend to coincide with influenza season in the United States (Table 1).

50 million search terms to fit 1152 data points!







# Big data Hubris

- Assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis
- Core challenge: most big data are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.





# Concept of Measurement

- Is the instrumentation actually capturing the theoretical construct of interest?
- Is measurement stable and comparable across cases and over time?
- Are measurement errors systematic?



# Big Data vs Small Data

- Choice depends on question being asked

**“You just brought a tote bag full of David Sedaris books to a knife fight” – Jon Stewart**