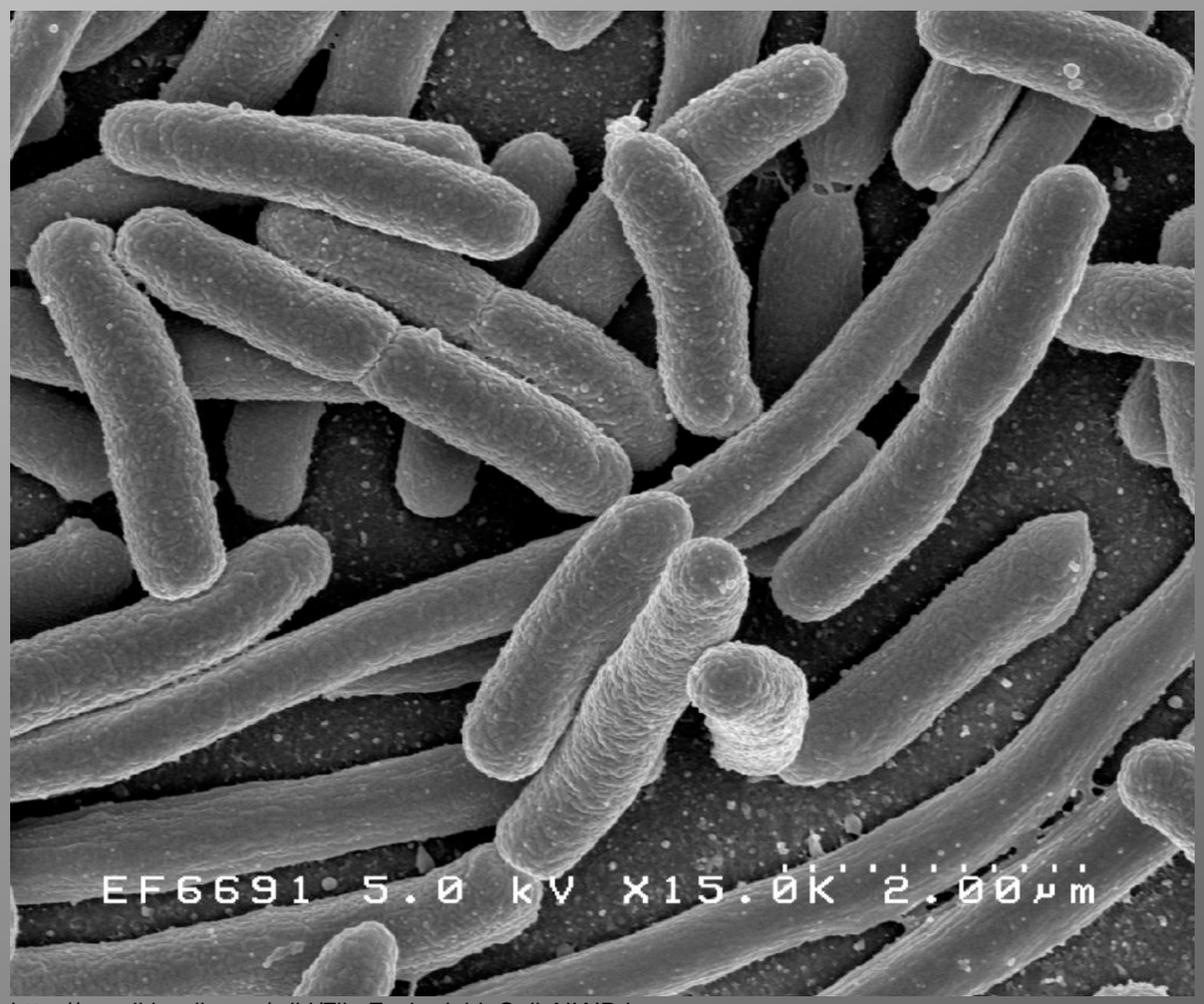
Global Disease Monitoring and Forecasting with Wikipedia + *bonus nerd stuff*

Nicholas Generous
Geoffrey Fairchild
Kyle Hickmann
Alina Deshpande
Sara Y. Del Valle
Reid Priedhorsky

CSCNSI Seminar
June 0x1E, 0x7DF



http://en.wikipedia.org/wiki/File:EscherichiaColi_NIAID.jpg

Los Alamos NATIONAL LABORATORY EST.1943

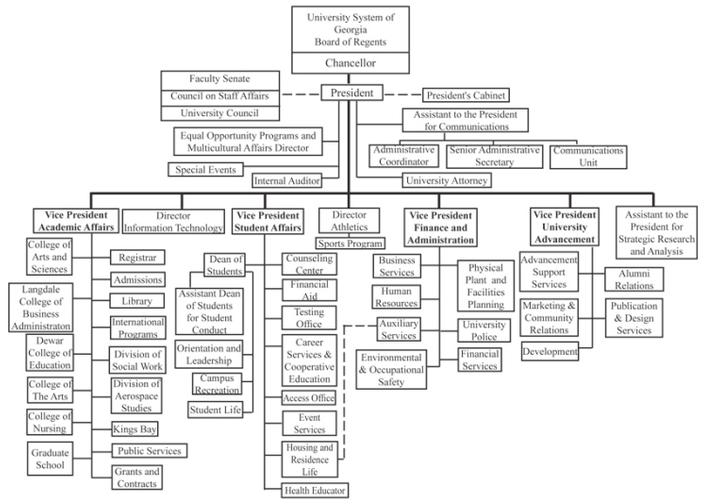
MIDAS

LA-UR 15-24972

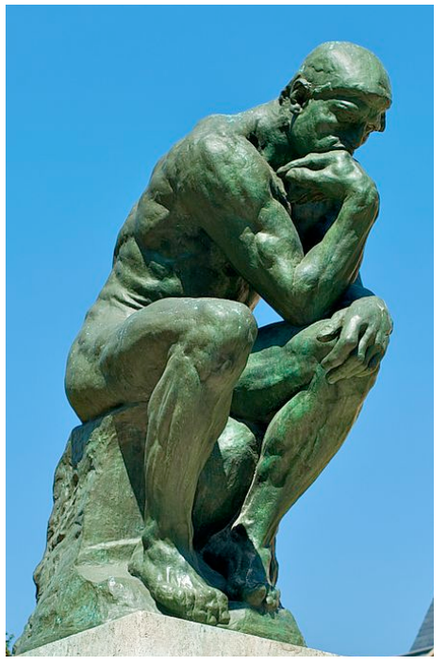
How disease monitoring works

http://en.wikipedia.org/wiki/File:Blooddraw.jpg
http://ww2.valdosta.edu/vsu/org/images/orgchart.jpg
https://en.wikipedia.org/wiki/File:Paris_2010_-_Le_Penseur.jpg
http://en.wikipedia.org/wiki/File:Internet_map_1024_-_transparent.png
http://en.wikipedia.org/wiki/File:Viegas-UserActivityonWikipedia.gif

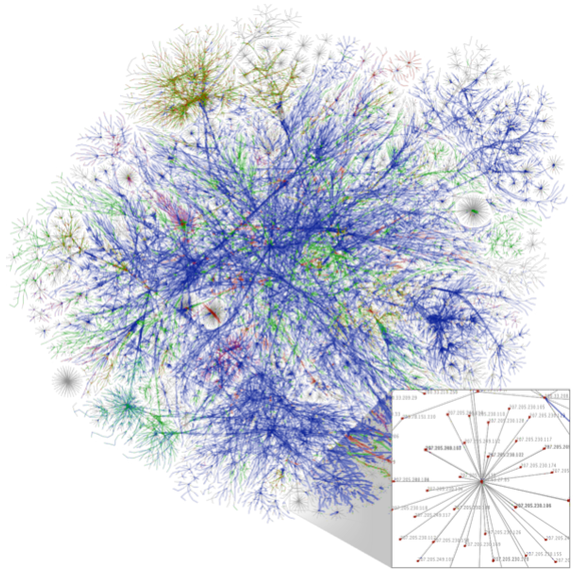
old:

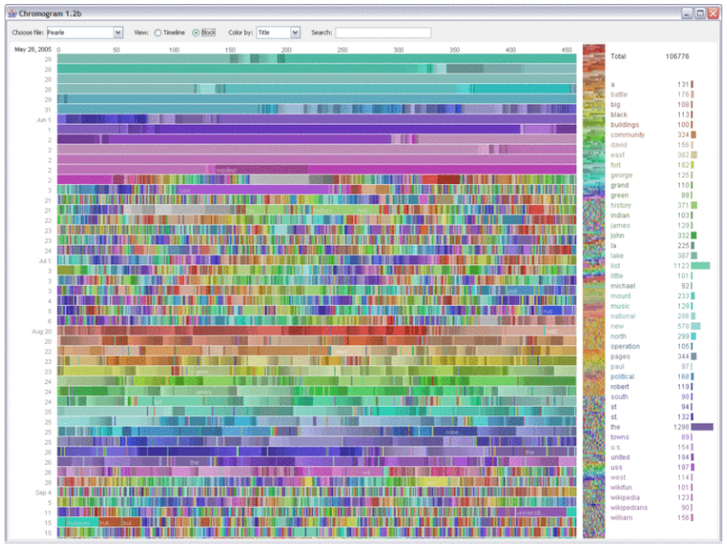


accurate ... costly, slow, & not always available



new:





Fortunately, Google has solved this for us

https://www.google.org/flutrends/us/#US-GA
http://time.com/23782/google-flu-trends-big-data-problems/

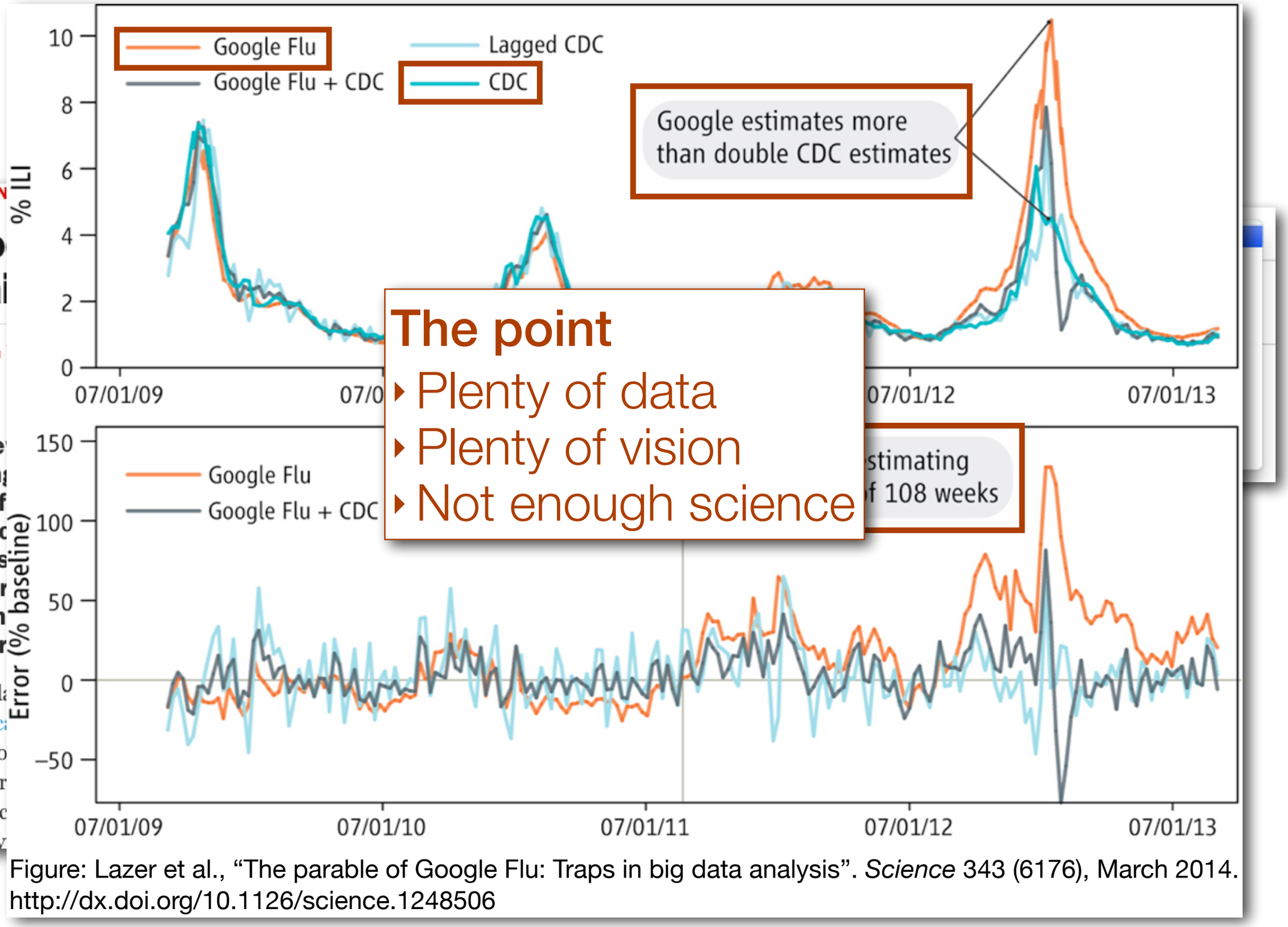


Figure: Lazer et al., “The parable of Google Flu: Traps in big data analysis”. *Science* 343 (6176), March 2014. http://dx.doi.org/10.1126/science.1248506

Related scientific work

Inclusion

- Single-source Internet data
- Biological disease metric

Excluded e.g.'s

- Alerts
- Public perception
- Disease dynamics in model
- Evaluates 3rd party method
- Multiple-source data
- Seasonality metric
- Drug sales metric

McIver & Brownstein

PLOS Comp Bio, April 17, 2014

- Influenza in United States
- Better statistics
- We're better on most other dimensions

The point

- Plenty of science
- Limited vision
- Not enough working systems

Search queries	**27**
Baidu	3
Google	20
Yahoo	2
Yandex	1
medical sites	3
Social media	**16**
Twitter	16
Server logs	**3**
Wikipedia	1
medical sites	2
Total	**46**

avian influenza	1
cancer	1
chicken pox	1
cholera	1
dengue	3
dysentery	1
gastroenteritis	3
gonorrhea	1
HFMD	1
HIV/AIDS	2
influenza	**27**
kidney stones	1
listeriosis	1
malaria	1
MRSA	1
pertussis	1
pneumonia	1
RSV	1
scarlet fever	1
stroke	1
suicide	2
tuberculosis	1
West Nile virus	1
Total	**55**

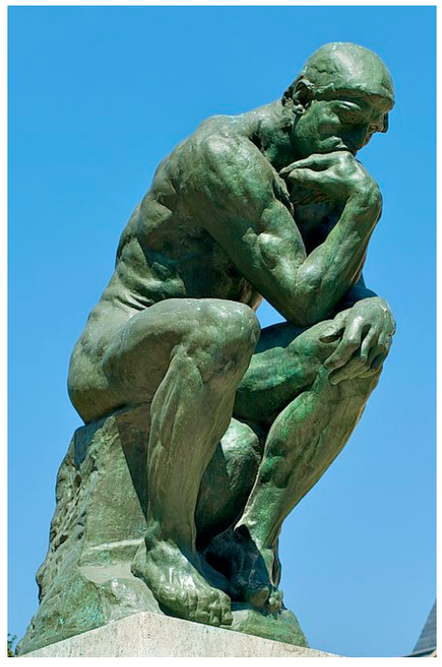
Four challenges for an operational resource

C1. Open data & algorithms

- High quality science
- Continuity & expandability

C2. Breadth

- 100's of disease/country
- Adapt w/o writing code



C3. Transferability

- Missing incidence data
- Transfer w/o re-training

C4. Forecasting

- Simpler = easier to apply

Our goal

Build an operational disease monitoring and forecasting system with open data and open source code.

For now: **proof-of-concept**

WIKIPEDIA
The Free Encyclopedia

What is it?

... online encyclopedia
... 30 million articles in 287 languages
... 850M hits per day (6th)
... top result for many search queries

WIKIPEDIA
The Free Encyclopedia

Here be public servants

How does it work?

... anyone can edit articles
... changes live immediately
... inverts traditional review/publish

What data are available?

This work

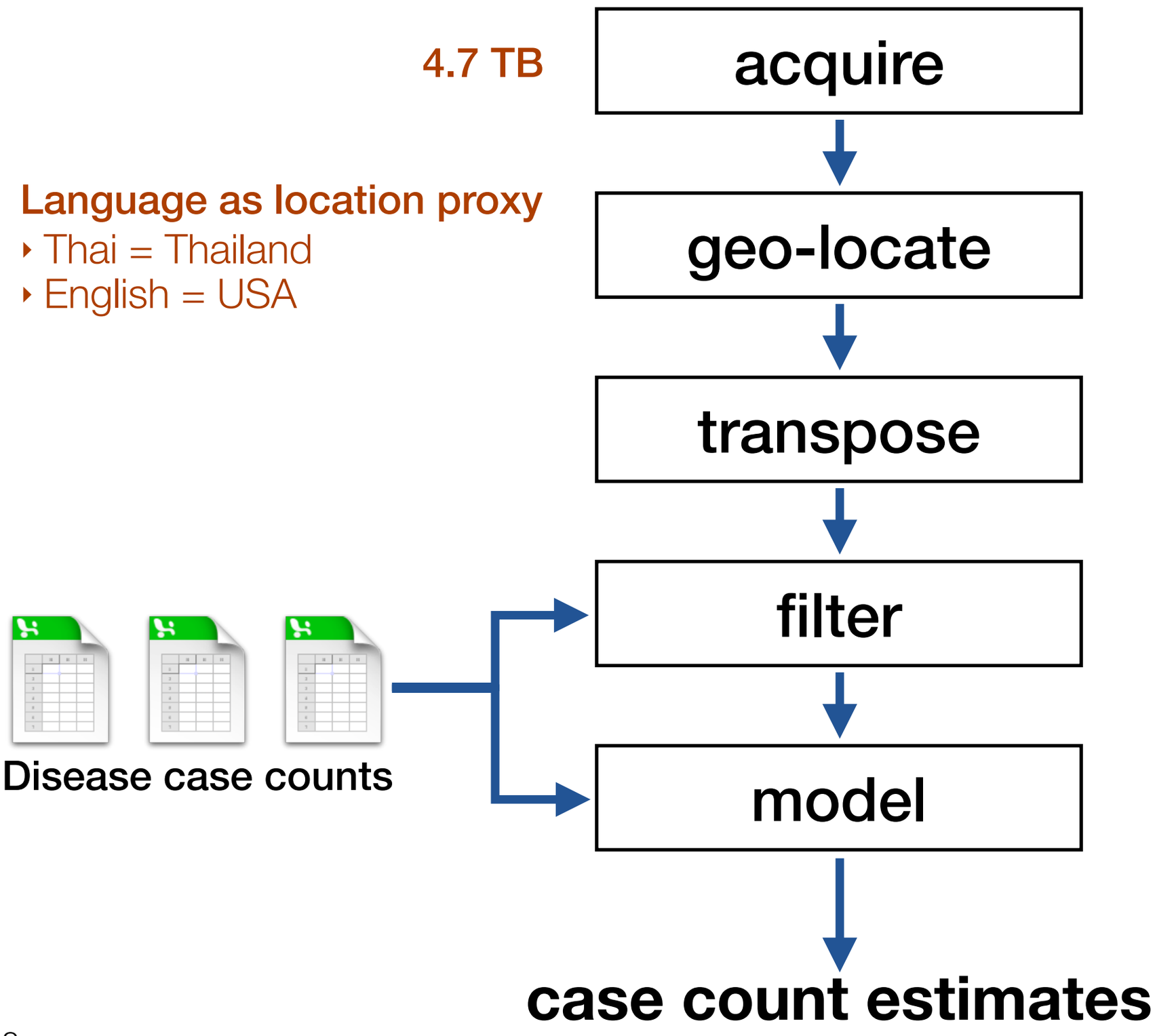
... complete history for every article
... hourly requests for each article ("access logs")
... and lots more

Data pipeline

4.7 TB

Language as location proxy

- Thai = Thailand
- English = USA



Filtering: Candidate articles in English

1. Go to disease article
2. Enumerate linked articles
3. Select articles which are:
   - symptoms
   - syndromes
   - pathogens
   - conditions
   - treatments
   - biological processes
   - epidemiology

e.g.
- Influenza
- Amantadine
- Swine influenza

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikimedia Shop

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

Print/export

Languages
Afrikaans
Alemannisch
العربية
Aragonés
Asturianu

Article | Talk | Read | View source | View history

Influenza

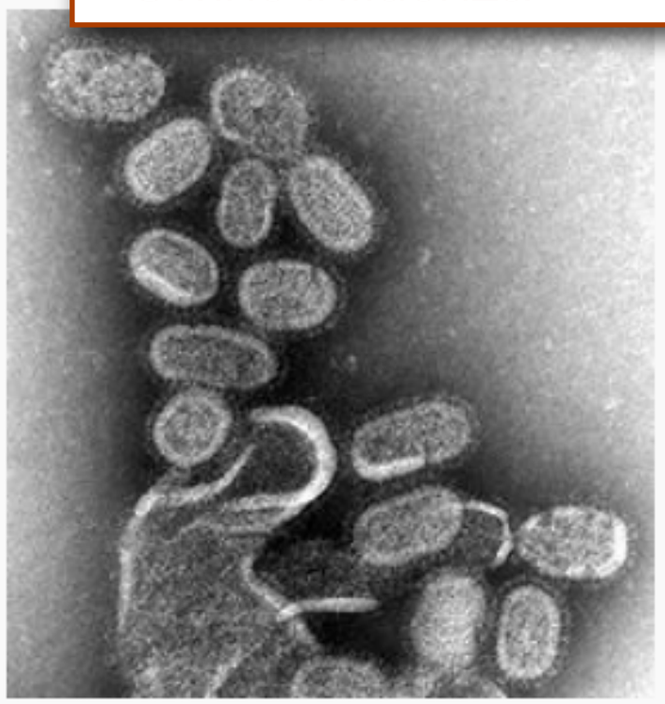
From Wikipedia, the free encyclopedia

"Flu" redirects here. For other uses, see Flu (disambiguati

"Grippe" redirects here. For other uses, see Grippe (disam

Influenza, commonly known as "the **flu**", is an infectious disease of birds and mammals caused by RNA viruses of the family Orthomyxoviridae, the influenza viruses. The most common symptoms are chills, fever, runny nose, sore throat, muscle pains, headache (often severe), coughing, weakness/fatigue and general discomfort. Although it is often confused with other influenza-like illnesses, especially the common cold, influenza is a more severe disease caused by a different type of virus.[1] Influenza may produce nausea and vomiting, particularly in children,[2] but these symptoms are more common in the unrelated gastroenteritis, which is sometimes inaccurately referred to as "stomach flu" or "24-hour flu".[3]

Typically, influenza is transmitted through the air by coughs or sneezes, creating aerosols containing



TEM of negatively stained influenza virions, magnified approximately 100,000 times

ICD-10 J10, J11

Filtering: Articles in all target languages

4. Translate w/ links

- Grypa
- Amantadyna
- Świńska grypa

5. Compare to incidence

- Correlation *r* for each article
- Order by decreasing |*r*|

6. Take top 10 articles

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

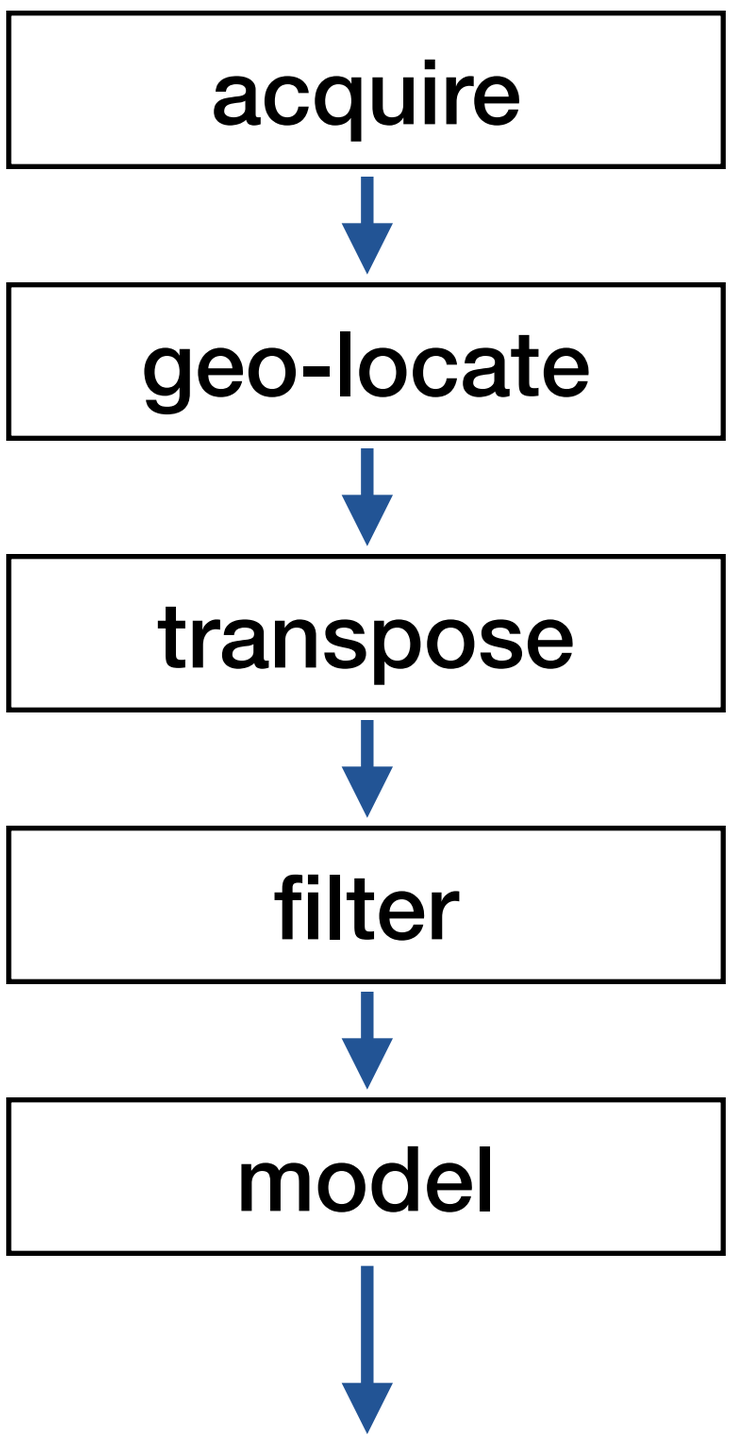
Data pipeline

WIKIPEDIA
The Free Encyclopedia

3.5 TB

Language as location proxy

- Thai = Thailand
- English = USA



Disease case counts

Simple linear model

- article hits ↦ disease incidence
- 10 variables
- ordinary least squares regression

Lag up to ±28 days

Q: Is there a signal?

Disease-location contexts analyzed

7 diseases, 9 countries, 14 total contexts **~3 years**

Disease	Country	Language	Dates	Resolution
Cholera	Haiti	French	2010-12-05 – 2013-12-05	daily
Dengue	Brazil	Portuguese	2010-03-13 – 2013-03-16	weekly
	Thailand	Thai	2010-10-20 – 2013-11-02	weekly
Ebola	Uganda/DRC	English	2011-01-01 – 2013-12-31	daily
HIV/AIDS	China (PRC)	Chinese	2011-01-01 – 2013-12-01	monthly
	Japan	Japanese	2010-10-09 – 2013-10-12	monthly
Influenza	Japan	Japanese	2010-06-26 – 2013-06-29	weekly
	Poland	Polish	2010-10-22 – 2013-10-22	weekly
	Thailand	Thai	2011-01-09 – 2014-01-12	weekly
	United States	English	2011-01-01 – 2014-01-04	weekly
Plague	United States	English	2011-01-22 – 2014-01-25	weekly
Tuberculosis	China (PRC)	Chinese	2010-12-01 – 2013-12-01	monthly
	Norway	Norwegian	2010-12-01 – 2013-12-01	monthly
	Thailand	Thai	2010-12-01 – 2013-12-01	monthly

results

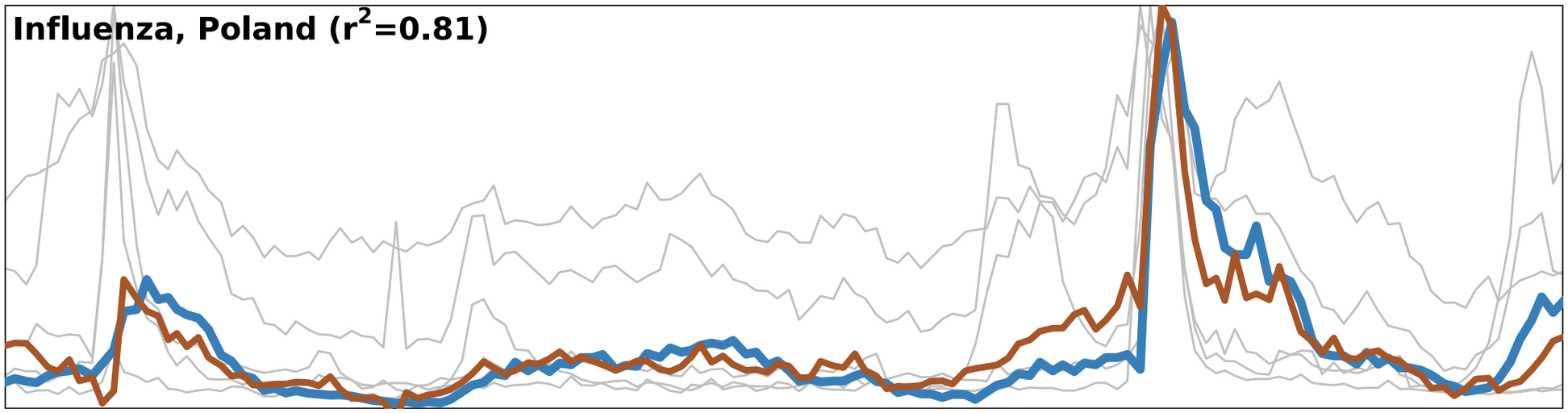
Failure modes

1. Didn’t capture pattern in official data

2. Weak signal-to-noise ratio

Success (8 of 14 contexts)

~3 years



1. Forecasting
... best r^2 a few weeks ahead

2. Transferability
... article weights sometimes match across languages

Summary so far

Wikipedia access logs seem promising
... comparable to other internet data (Twitter, Google)

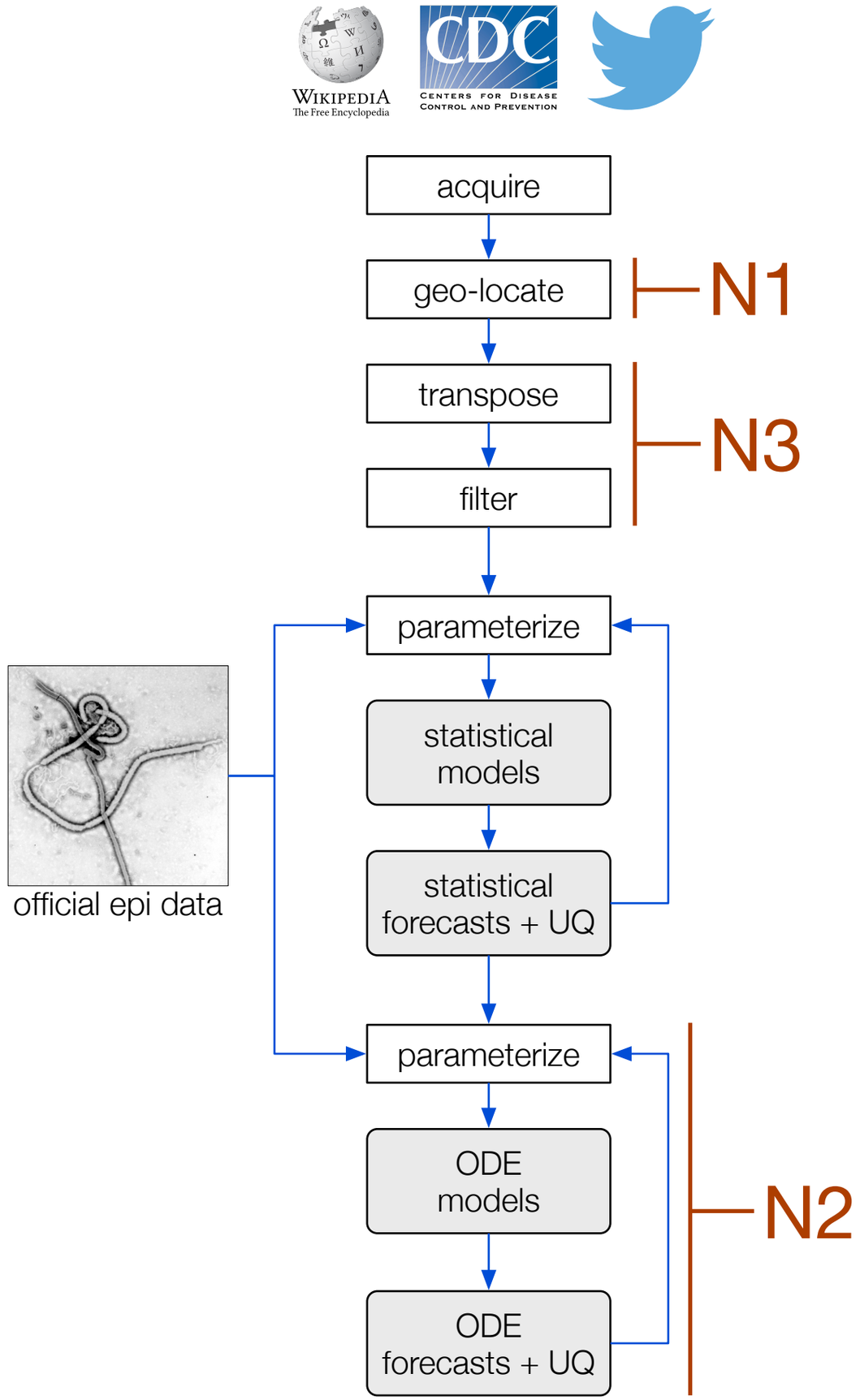
Next steps:

1. Better geo-location
2. Couple with mechanistic models
3. Data-driven article filter

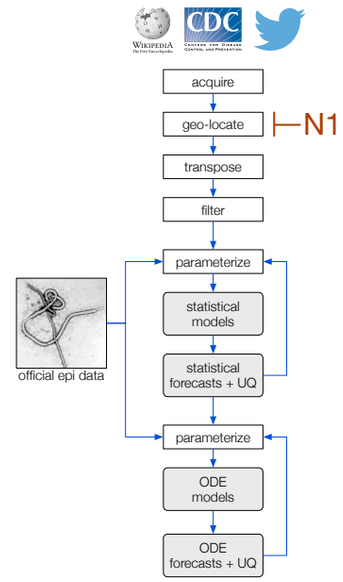
Challenges:

1. Missing ground truth data
2. What is the value-add?

New data flow



C1: Geo-location: Work with WMF



Geo-aggregation of Wikipedia page views:
Maximizing geographic granularity while preserving privacy

Reid Priedhorsky, Geoffrey Fairchild, Sara Del Valle
{reidpr,gfairchild,sdelvall}@lanl.gov
Los Alamos National Laboratory

Proposal and request for feedback
LA-UR 15-20145
Draft 2 — January 6, 2015

1 Motivation

Research shows that analyzing internet traces of health-related activity, such as search queries and social media messages, is an effective means of monitoring the spread of disease and has significant promise to tackle problems that traditional disease monitoring tools cannot. The most well-known example of this is Google Flu Trends.[1]

Recently, two teams have extended these techniques to use the globally-aggregated Wikipedia page view logs currently available,[2] with good results.[3,4] There is even promising evidence that disease forecasting, not simply monitoring, is possible as well using these data. Further, these page view logs are, to our knowledge, the only freely available data source in this class, meaning that the science and utility they can support is significantly greater than proprietary alternatives such as Google queries or Twitter messages.

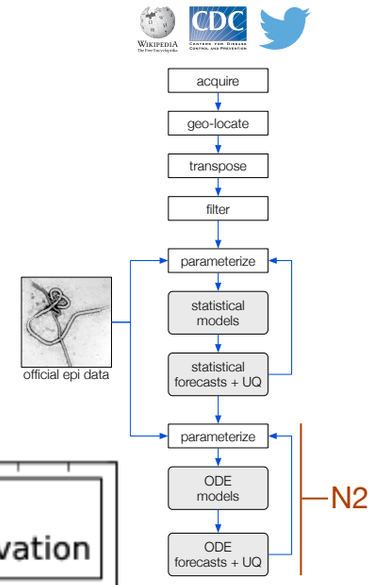
Effective disease monitoring is fundamentally geographic and requires geo-located data sources. Both of the above efforts infer geography at the country level from the wiki language (e.g., views of the Thai Wikipedia are assumed to come from Thailand), whether implicitly or explicitly. Wikimedia traffic statistics generally support this technique.[5] However, it has several problems that make it unsuitable for operational use, including:

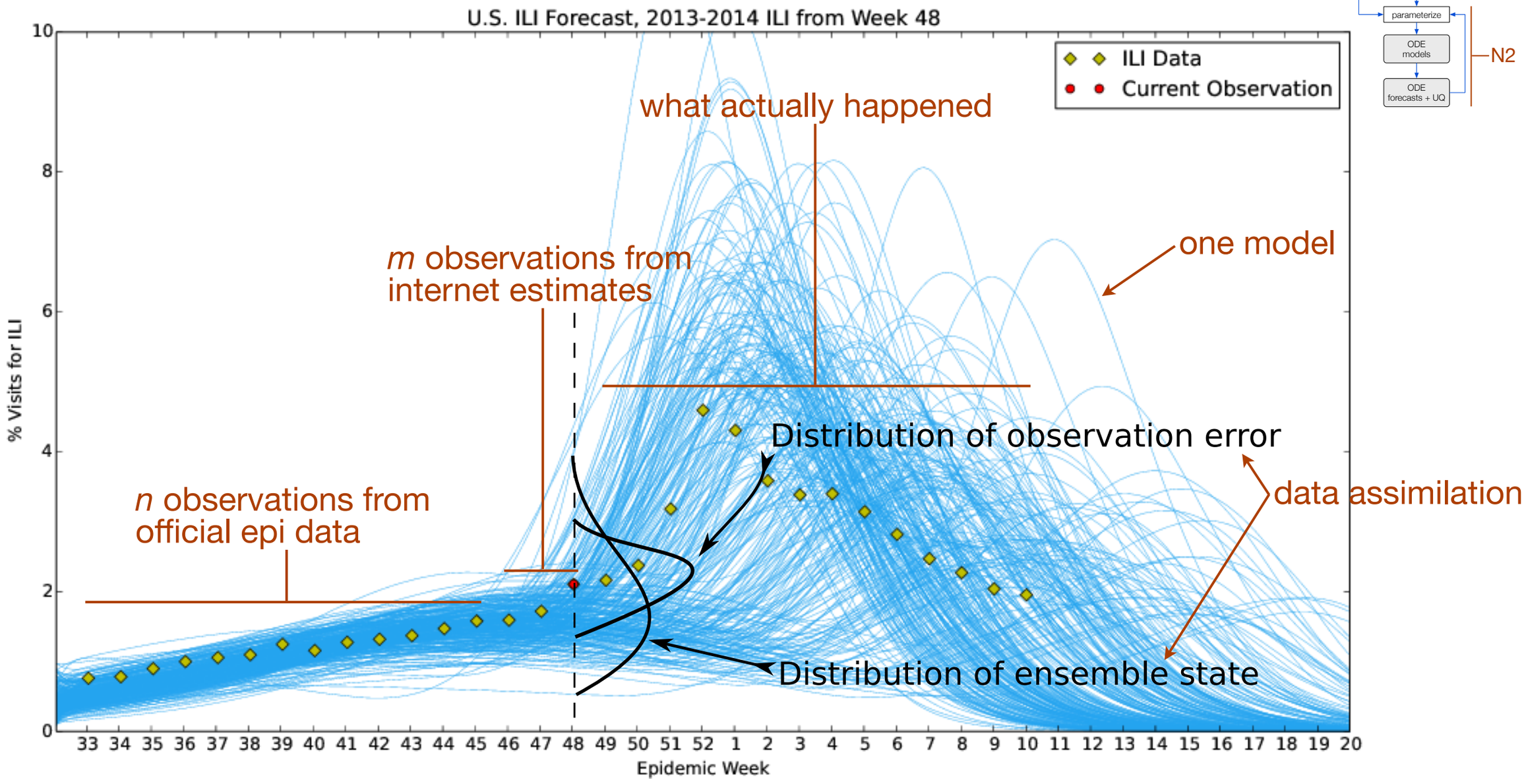
1. It cannot be applied at geographic granularity finer than the country level. However, in many countries, disease monitoring must be carried out at the state or metro-area level in order to be effective.

2. Many or most countries can't be covered at all. For example, this approach is not feasible for Chile, which comprises only 4.4% of Spanish-language requests and whose signal is swamped by other Spanish-speaking countries.

[1] http://www.google.org/flutrends/
[2] http://dumps.wikimedia.org/other/pagecounts-raw/
[3] McIver & Brownstein (http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1003581).
[4] Generous et al. (http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1003892).
[5] http://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm

C2: Model coupling via data assimilation



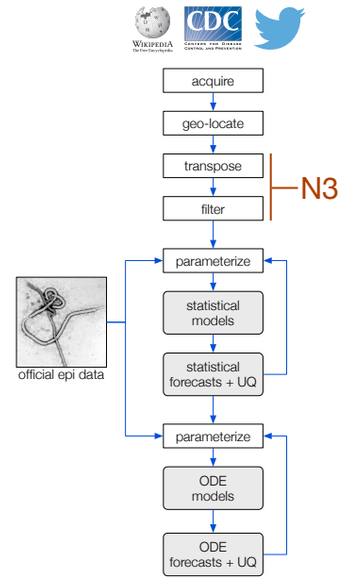


C3: Filter: Raw data file

```
en Lorry_Driver_(film) 1 10720
en Lorry_Greenberg 1 10427
en Lorry_I._Lokey 2 23188
en Lorsch_codex 2 37435
en Lortab 1 29079
en Lortalamine 1 26683
en Lortel_Archives 1 8378
en Lorti 1 6438
en Lorton_(Amtrak_station) 1 10916
en Lorton_(VRE_station) 3 0
en Lorton_Lick 1 6466
en Lortryggen 1 6451
en Lorum 2 10091
en Lorum%2BPiercing&as_qdr=all&filter=0&num=100&complete=0&cr=countryUS&gws_rd=cr,ssl&hl=sv&&ct=clnk 1 6935
en Lorum_ladder 1 6853
en Lorwyn 1 17000
en Lory_Del_Santo 4 37288
en Lory_Park_Zoo 1 6477
en Loryma_callos 1 7845
en Loryn 1 6437
en Loryn_Locklin 1 10859
en Los 3 28330
en Los_%C3%81lamos 1 12498
en Los_%C3%81ngeles,_B%C3%ADo_B%C3%ADo 6 111611
en Los_%C3%81ngeles,_Catamarca 1 8006
en Los_%C3%81ngeles,_Chiriqu%C3%AD 1 45591
en Los_%C3%81ngeles_Negros 1 19265
en Los_%C3%81ntrax 3 38522
en Los_%C3%9Altimos_D%C3%ADas 1 15202
en Los_%C3%BAltimos_de_Filipinas 1 9296
en Los_(Blake) 1 17572
en Los_(rapper) 33 671564
en Los_12_Disc%C3%ADpulos 1 14703
en Los_Adaes 1 22297
en Los_Alamitos,_California 2 57506
en Los_Alamitos_Elementary_School 2 13543
en Los_Alamitos_High_School 2 19528
en Los_Alamos,_California 2 103141
en Los_Alamos,_NM 1 25098
en Los_Alamos,_New_Mexico 11 275203
en Los_Alamos_County,_New_Mexico 5 40740
en Los_Alamos_Laboratory 2 73937
en Los_Alamos_National_Laboratory 15 647836
en Los_Alamos_National_Security,_LLC 1 0
en Los_Alcarrizos 1 14118
en Los_Alfaques_Disaster 1 16868
en Los_Alpargatales 1 6471
en Los_Altos,_California 11 749883
en Los_Altos,_Central_America 1 20379
en Los_Altos_(Jalisco) 1 34741
en Los_Altos_High_School_(Los_Altos,_California) 2 52392
en Los_Altos_Hills 1 23742
en Los_Altos_Hills,_California 12 448753
en Los_Amigos 1 11083
en Los_Amigos_Invisibles 1 43042
en Los_Amigos_River 1 7832
en Los_Andes,_Chile 2 70583
en Los_Andes_Province,_Chile 2 29308
en Los_Andes_de_Lomas_de_Zamora 1 13370
en Los_Angeles 191 22300490
en Los_Angeles,_Bohol 1 73608
en Los_Angeles,_CA 2 782968
en Los_Angeles,_California 25 2970494
en Los_Angeles-Long_Beach-Anaheim,_CA_Metropolitan_Statistical_Area 1 46183
en Los_Angeles-class_submarine 9 365370
```

Not a real article (HTTP 404)

One hour
5M URLs



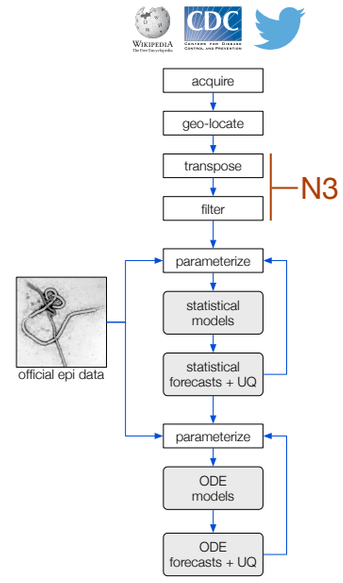
	URLs	Size
One hour:	5M	75MiB
One month:	500M	71GiB
Total (7.5 years):	25G?	4.7TiB

`en Los_Alamos,_New_Mexico 11 275203`

language URL hits bytes

66,239 gzipped text files (and counting)

C3: Filter: Goals for data format



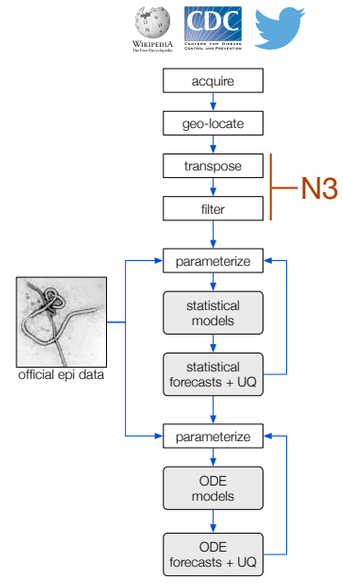
Use cases

- Retrieve ~1 article time series (serial)
- Iterate through all time series (parallel)

Maintenance cases

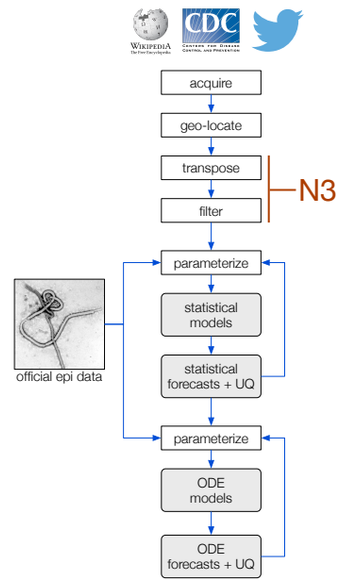
- Bulk transpose all raw data (parallel)
- Update with 1 day of new data (serial)

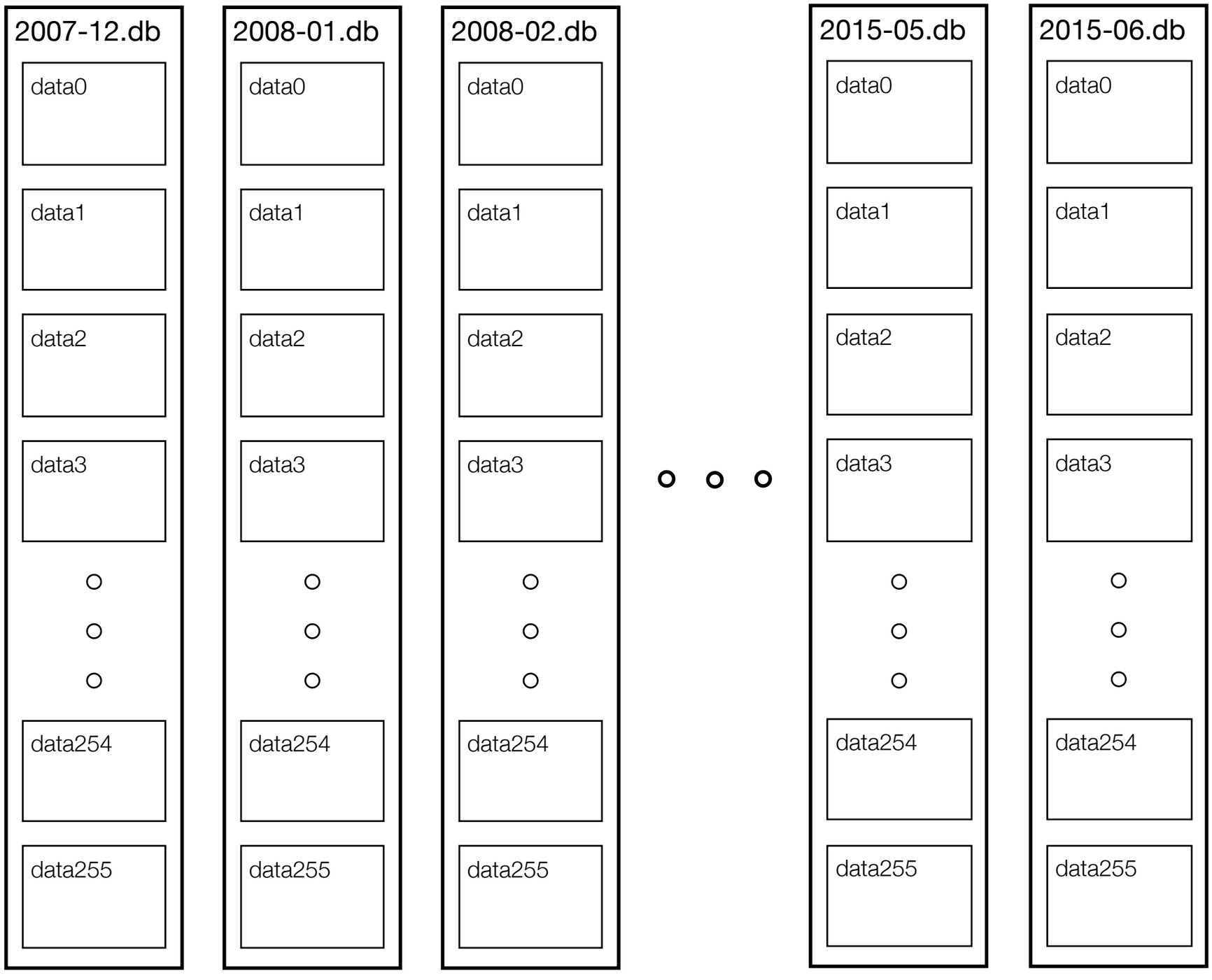
C3: Filter: Data funnel



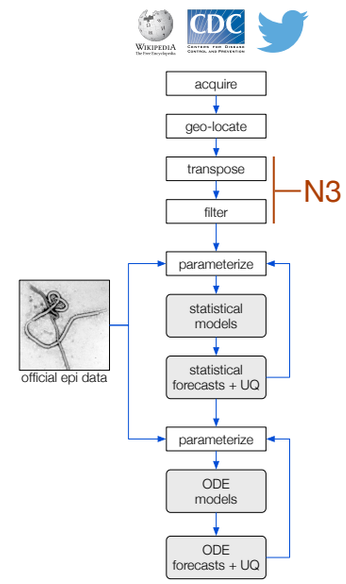
	URLs	
Total	**25G**	
Well-formed	**20G**	*egrep*
Plausible for any model	**37M**	60 hits/month
Plausible for specific model	**300**	correlation
Actually in model	**30**	linear regression

C3: Filter: Data format: Sharded SQLite databases





C3: Filter: Data format: Database schema



2007-12.db
- data0
- data1
- data2
- data3
- ○ ○ ○
- data254
- data256

2008-01.db
- data0
- data1
- data2
- ○ ○
- data254
- data256

2008-02.db
- data0
- data1
- data2
- ○ ○
- data254
- data256

2015-05.db
- data0
- data1
- data2
- ○ ○
- data254
- data256

2015-06.db
- data0
- data1
- data2
- ○ ○
- data254
- data256

```
CREATE TABLE data0 (
  name   TEXT NOT NULL PRIMARY KEY,
  dtype  TEXT NOT NULL,
  total  REAL NOT NULL,
  data   BLOB NOT NULL);
```

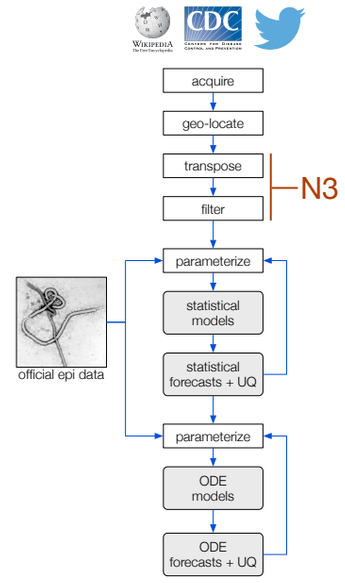
URL

NumPy data type

Vector fragment sum

Vector fragment
NumPy memory dump,
maybe gzipped

C3: Filter: Data format performance



Use cases

- Retrieve ~1 article time series — 1–20 s
- Iterate through all time series — 7 min (256×350/s) [est.]

Maintenance cases

- Bulk transpose all raw data — 24 hours (91× parallel) [est.]
- Update with 1 day of new data — 6–12 hours [est.]

Nicholas Generous
Geoffrey Fairchild
Kyle Hickmann
Alina Deshpande
Sara Y. Del Valle
Reid Priedhorsky / reidpr@lanl.gov

dx.doi.org/10.1371/journal.pcbi.1003892

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Global Disease Monitoring and Forecasting with Wikipedia

Nicholas Generous*, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle, Reid Priedhorsky

Defense Systems and Analysis Division, Los Alamos National Laboratory, Los Alamos, New Mexico, United States of America

Abstract

Infectious disease is a leading threat to public health, economic stability, and other key social structures. Efforts to mitigate these impacts depend on accurate and timely monitoring to measure the risk and progress of disease. Traditional, biologically-focused monitoring techniques are accurate but costly and slow; in response, new techniques based on social internet data, such as social media and search queries, are emerging. These efforts are promising, but important challenges in the areas of scientific peer review, breadth of diseases and countries, and forecasting hamper their operational usefulness. We examine a freely available, open data source for this use: access logs from the online encyclopedia Wikipedia. Using linear models, language as a proxy for location, and a systematic yet simple article selection procedure, we tested 14 location-disease combinations and demonstrate that these data feasibly support an approach that overcomes these challenges. Specifically, our proof-of-concept yields models with r^2 up to 0.92, forecasting value up to the 28 days tested, and several pairs of models similar enough to suggest that transferring models from one location to another without re-training is feasible. Based on these preliminary results, we close with a research agenda designed to overcome these challenges and produce a disease monitoring and forecasting system that is significantly more effective, robust, and globally comprehensive than the current state of the art.

Citation: Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R (2014) Global Disease Monitoring and Forecasting with Wikipedia. PLoS Comput Biol 10(11): e1003892. doi:10.1371/journal.pcbi.1003892

Editor: Marcel Salathé, Pennsylvania State University, United States of America

Received April 18, 2014; **Accepted** August 21, 2014; **Published** November 13, 2014

Data Availability: The authors confirm that all data underlying the findings are fully available without restriction. Data are available in the Supplemental Information. Software is available at http://github.com/reidpr/quac

Funding: This work is supported in part by NIH/NIGMS/MIDAS under grant U01-GM097658-01 and the Defense Threat Reduction Agency (DTRA), Joint Science and Technology Office for Chemical and Biological Defense under project numbers CB3656 and CB10007. Data collected using QUAC; this functionality was supported by the U.S. Department of Energy through the LANL LDRD Program. Computation used HPC resources provided by the LANL Institutional Computing Program. LANL is operated by Los Alamos National Security, LLC for the Department of Energy under contract DE-AC52-06NA25396. Approved for public release: LA-UR~14-22535. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: generous@lanl.gov

Introduction

Motivation and Overview

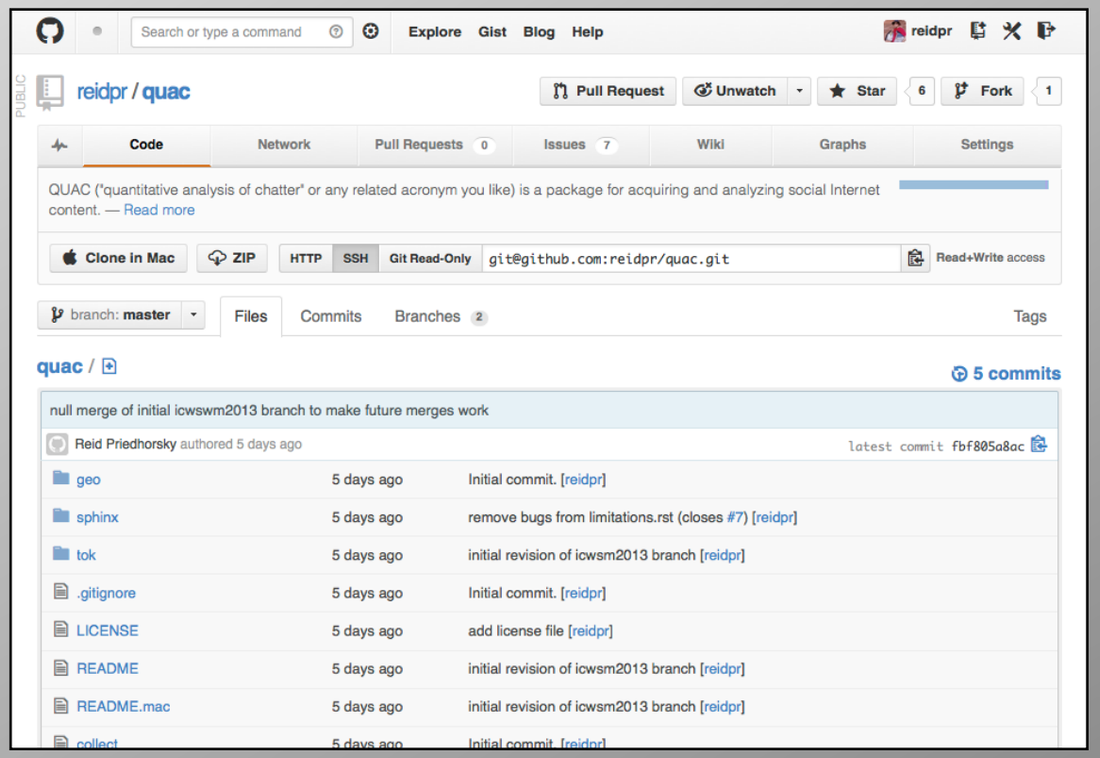
Infectious disease remains extremely costly in both human and economic terms. For example, the majority of global child mortality is due to conditions such as acute respiratory infection, measles, diarrhea, malaria, and HIV/AIDS [1]. Even in developed countries, infectious disease has great impact; for example, each influenza season costs the United States between 3,000 and 49,000 lives [2] and an average of $87 billion in reduced economic output [3].

Effective and timely disease surveillance — that is, detecting, characterizing, and quantifying the incidence of disease — is a critical component of prevention and mitigation strategies that can save lives, reduce suffering, and minimize impact. Traditionally, such monitoring takes the form of patient interviews and/or laboratory tests followed by a bureaucratic reporting chain; while generally considered accurate, this process is costly and introduces a significant lag between observation and reporting.

These problems have motivated new surveillance techniques based upon internet data sources such as search queries and social media posts. Essentially, these methods use large-scale data mining techniques to identify health-related activity traces within the data streams, extract them, and transform them into some useful metric. The basic approach is to train a statistical estimation model against ground truth data, such as ministry of health disease incidence records, and then apply the model to generate estimates when the true data are not available, e.g., when forecasting or when the true data have not yet been published. This has proven effective and has spawned operational systems such as Google Flu Trends (http://www.google.org/flutrends/). However, four key challenges remain before internet-based disease surveillance models can be reliably integrated into an decision-making toolkit:

C1. Openness. Models should afford review, replication, improvement, and deployment by third parties. This guarantees a high-quality scientific basis, continuity of operations, and broad applicability. These requirements imply that model algorithms — in the form of source code, not research papers — must be generally available, and they also imply that complete input data must be available. The latter is the key obstacle, as terms are dictated by the data owner rather than the data user; this motivated our exploration of Wikipedia access logs. To our knowledge, no models exist that use both open data and open algorithms.

C2. Breadth. Dozens of diseases in hundreds of countries have sufficient impact to merit surveillance; however, adapting a



github.com/reidpr/quac

acknowledgements:

- Susan M. Mniszewski
- Funding:
  - NIH/NIGMS: U01-GM097658-01
  - DTRA
  - U.S. Dept. of Energy via LANL/LDRD
- LANL is operated by LANS, LLC for DOE under contract DE-AC52-06NA25396.

Los Alamos NATIONAL LABORATORY EST.1943

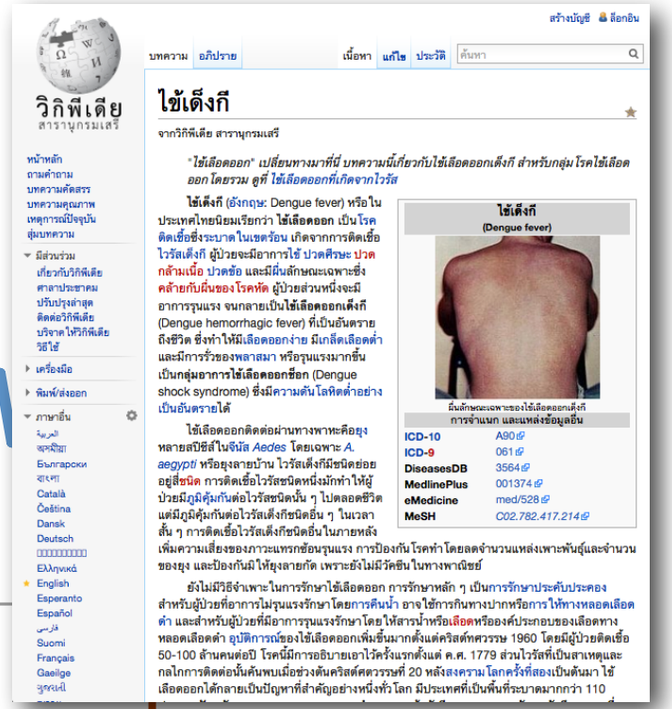
MIDAS

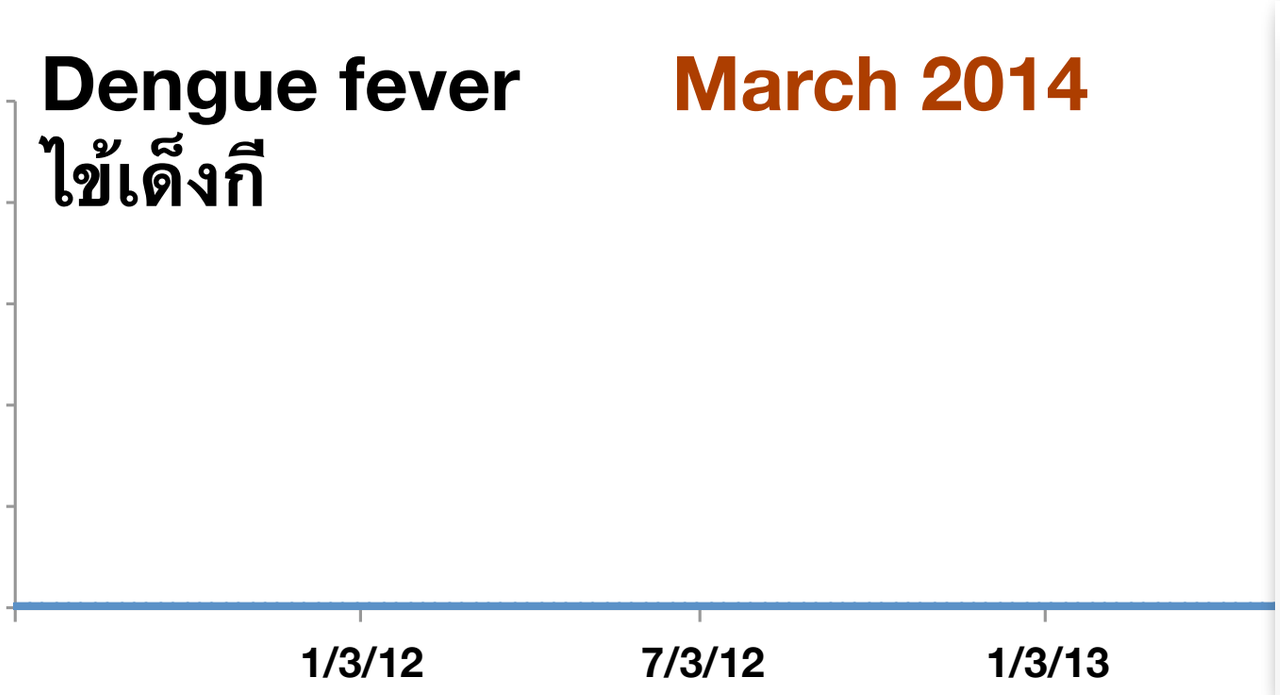

wow
so results
such Big Data Methods™
easy win
much funding

A story about dengue in Thailand

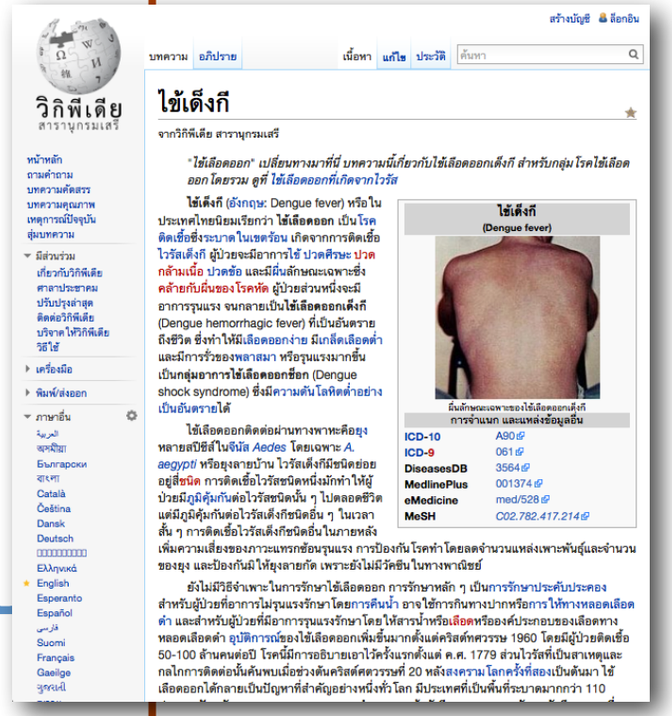
The point: Know your data!

~July 2013





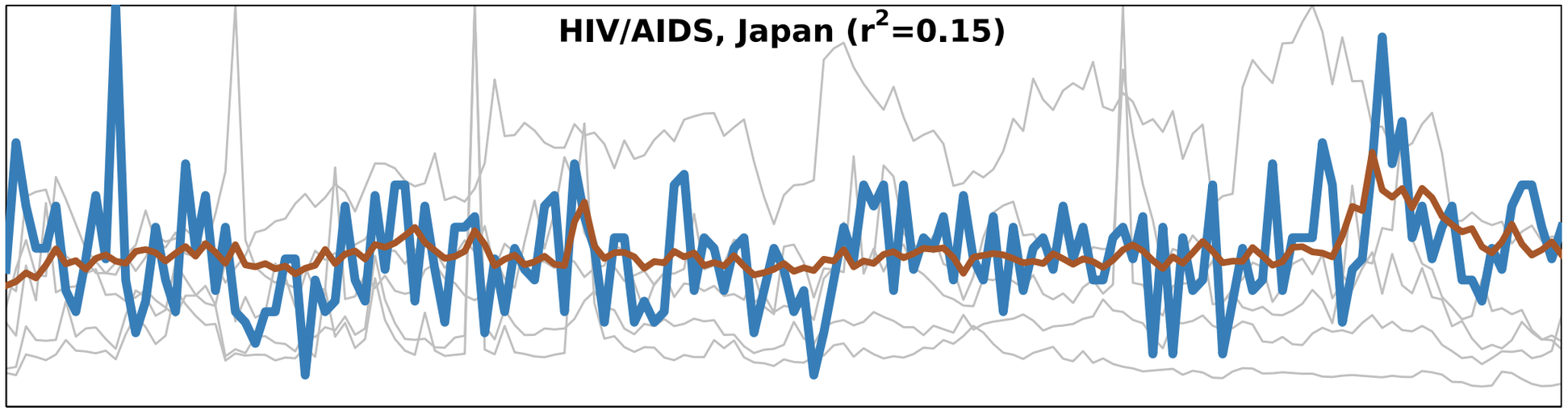
March 2014



July 23, 2013

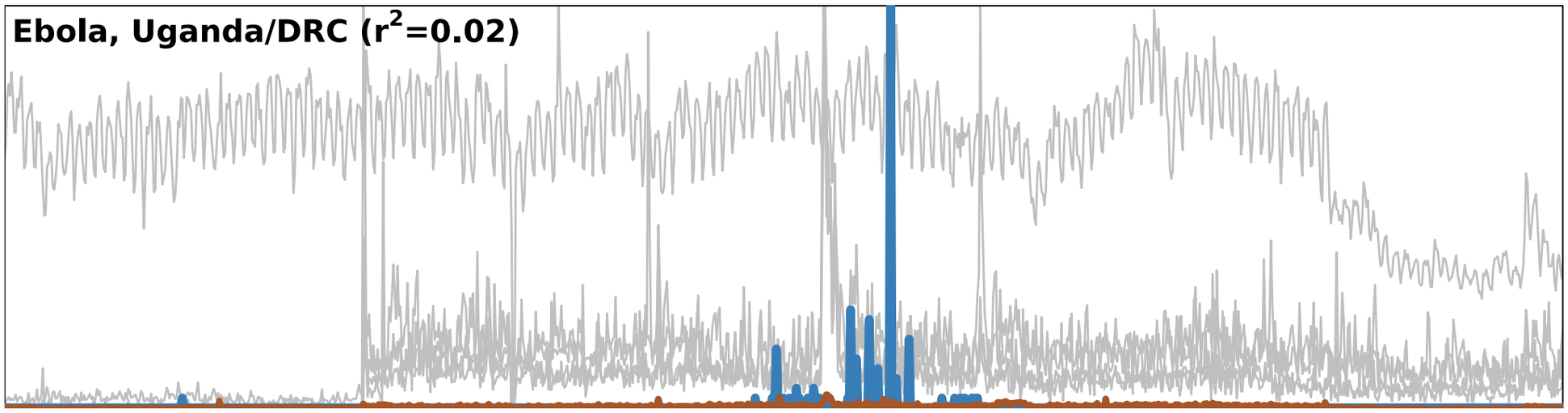
Failed: Didn’t capture data pattern

~3 years



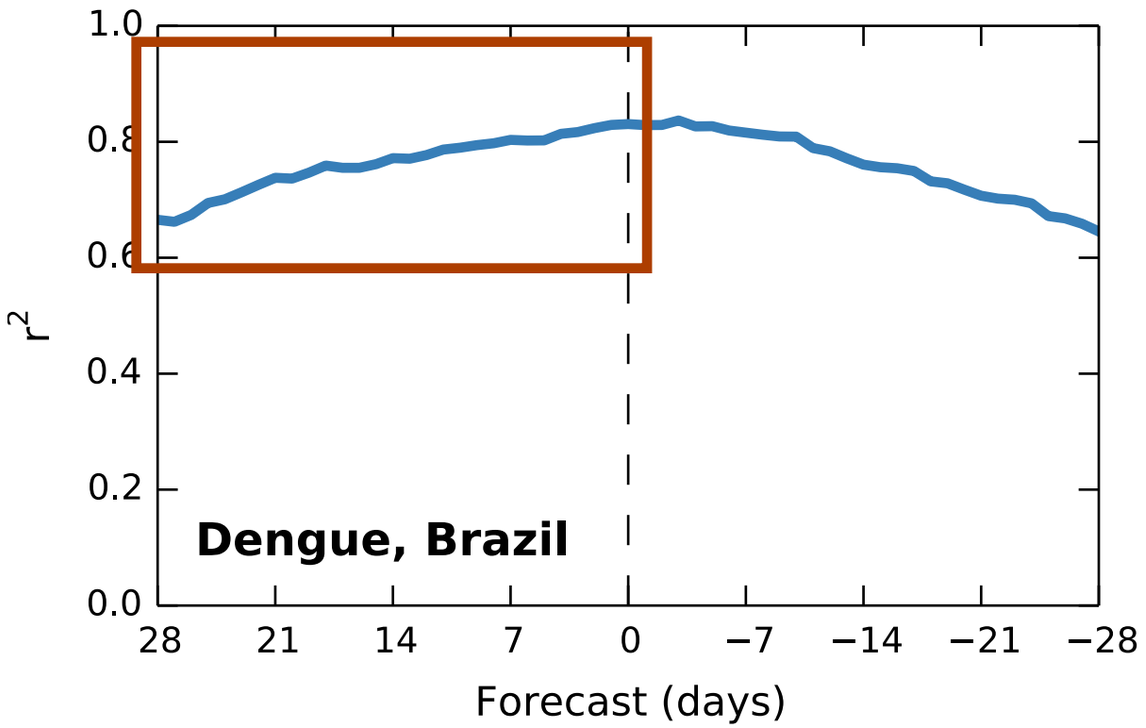
Failed: Weak signal-to-noise ratio (SNR)

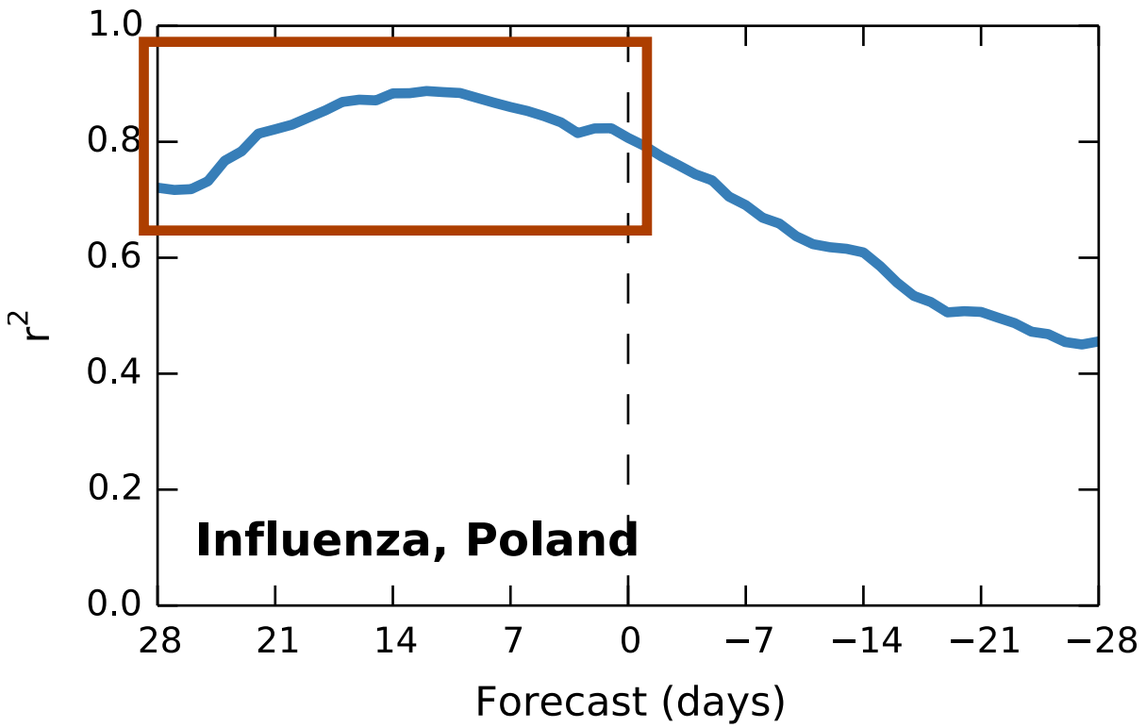
~3 years

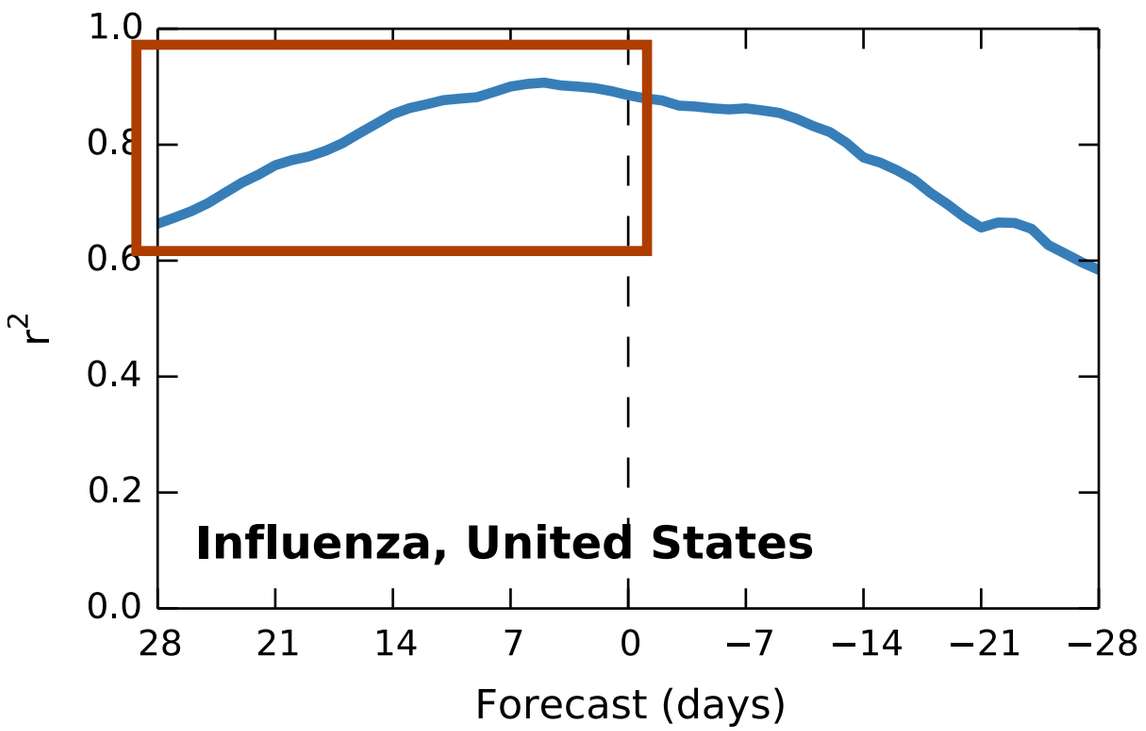


Success

Seasonality? News? Observing infectors?







Other successes

- Dengue: Thailand
- Influenza: Japan, Thailand
- Tuberculosis: China, Thailand

McIver & Brownstein vs. Generous et al.

	McIver	Generous
First use of Wikipedia	yes	yes
Peer reviewed	yes	yes
Contexts	1	14
Analysis period	5 years	3 years
Language as geo-proxy	mentioned	tested
Statistics	better	worse
Goodness of fit	r, MAE	r^2
Compare to GFT	yes	no
Negative results	no	included
Forecasting	no	28 days
Translatability	no	tested
Focus	result	data source
Research agenda	no	proposed