Northeastern

STATISTICS AND MODELING WITH NOVEL DATA STREAMS

Alessandro Vespignani
@alexvespi

Prediction= is based on what is known (mostly influenced by initial conditions) and provide a single outcome (although it may be probabilistic).

Forecast = given the best current knowledge on the system (knowledge often with uncertainties, model's parameters inferred statistically, etc) there is an ensemble of predictions that are analyzed statistically.

Projection = attempt to describe what would happen under certain assumptions and hypotheses (what if)

THE SOCIAL ATOM

The problem : technosocial systems are unpredictable because include the infinite psychological and cognitive reactions of individuals. We're too complex and unpredictable.

The vision: From the "social atom" or "social molecule" (i.e. small social groups) to the quantitative analysis of "social aggregate states" (Lundberg, Moreno).

THE SOCIAL ATOM

The problem : technosocial systems are unpredictable because include the infinite psychological and cognitive reactions of individuals. We're too complex and unpredictable.

The vision: From the "social atom" or "social molecule" (i.e. small social groups) to the quantitative analysis of "social aggregate states" (Lundberg, Moreno).

THE SOCIAL ATOM

The problem : technosocial systems are unpredictable because include the infinite psychological and cognitive reactions of individuals. We're too complex and unpredictable.

The vision: From the "social atom" or "social molecule" (i.e. small social groups) to the quantitative analysis of "social aggregate states" (Lundberg, Moreno).

MATHEMATICAL -> COMPUTATIONAL

Numerical Weather models

- 1920 Richardson integrate manually equations of the atmosphere
- 1950 First numerical weather forecast (24h computation for a 24h forecast)
- 1955 Numerical weather prediction models became operational by the USWB
- 2015 Government and Commercial entities routinely forecast up to three weeks

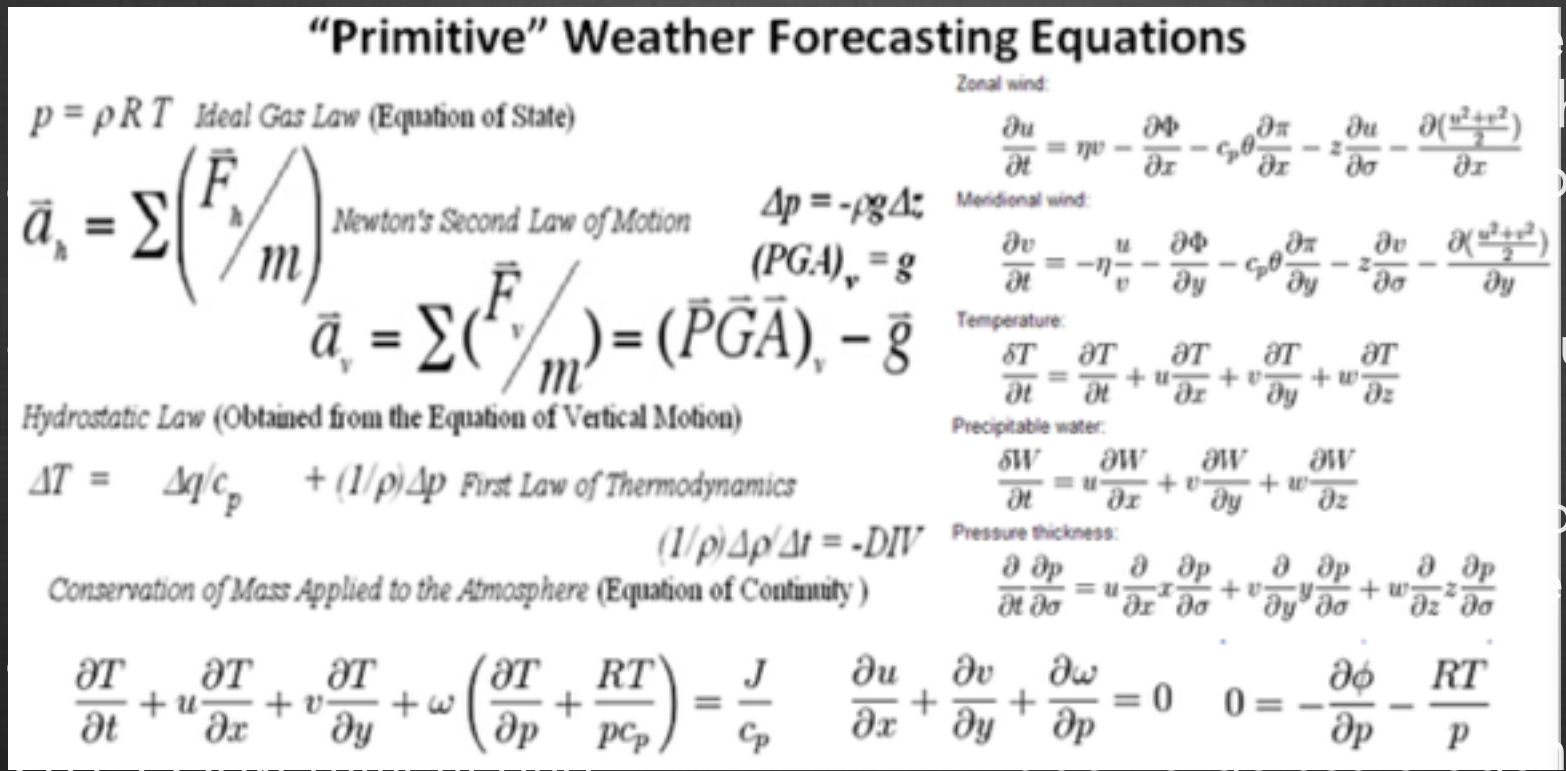
Numerical Epidemic models

- 1930 Reed-Frost define a simple chain binomial model that they integrate with a "sandbox' computer
- 1952 First Reed-Frost numerical implementation
- 1980-2000 progress toward the definition of large-scale individual models
- 2005 Large scale agent-based models
- 2105 ???

MATHEMATICAL -> COMPUTATIONAL

Numerical Weather models

- a simple
- hat they
- box'
- umerical
- oward the
- t-based



- 2015 Governmental and Commercial entities routinely forecast up to three weeks

Numerical Epidemic models

- models
- 2105 ???

MATHEMATICAL -> COMPUTATIONAL

Numerical Weather models

- 1920 Richardson integrate manually equations of the atmosphere
- 1950 First numerical weather forecast (24h computation for a 24h forecast)
- 1955 Numerical weather prediction models became operational by the USWB
- 2015 Government and Commercial entities routinely forecast up to three weeks

Numerical Epidemic models

- 1930 Reed-Frost define a simple chain binomial model that they integrate with a "sandbox' computer
- 1952 First Reed-Frost numerical implementation
- 1980-2000 progress toward the definition of large-scale individual models
- 2005 Large scale agent-based models
- 2105 ???

MATHEMATICAL -> COMPUTATIONAL

Numerical Weather models

- 1920 Richardson integrate manually equations of the atmosphere
- 1950 First numerical weather forecast (24h computation for a 24h forecast)
- 1955 Numerical weather prediction models became operational by the USWB
- 2015 Government and Commercial entities routinely forecast up to three weeks

Numerical Epidemic models

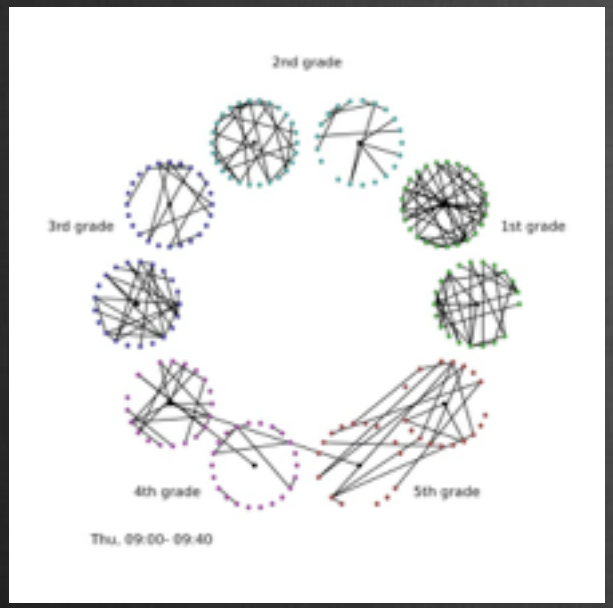
- 1930 Reed-Frost define a simple chain binomial model that they integrate with a "sandbox' computer
- 1952 First Reed-Frost numerical implementation
- 1980-2000 progress toward the definition of large-scale individual models
- 2005 Large scale agent-based models
- 2105 ???

BIG DATA

- Data has fueled spectacular advances in the natural sciences over the last 100 years.
- So what is different now?
- Every 1.2 years, more human-driven socioeconomic data is produced than during all previous history
- Embedded within the data are the raw ingredients for understanding socio-technical and socio-economic systems
- The focus is on understanding these data sets in a statistical sense and more deeply the real world processes which produced the data

DATA ARE CENTRAL IN THE ANALYSIS OF CONTAGION PROCESSES

Human interactions

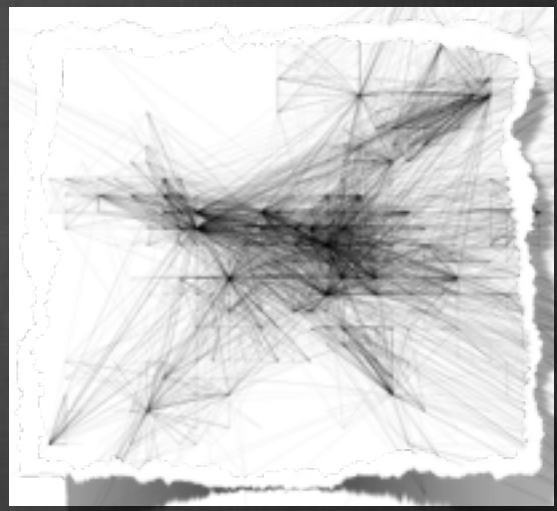


Within school contact patterns

Human Mobility

Multiscale integration of mobility networks in the analysis of potentially pandemic pathogens spread.

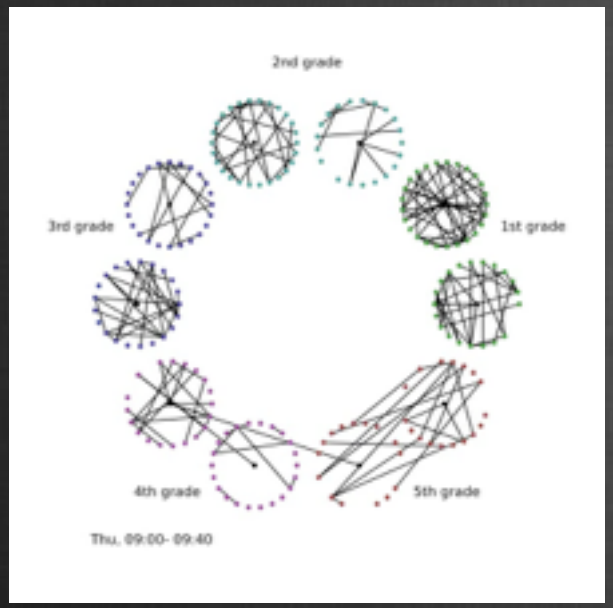
Social networks heterogeneities



Hubs, community, clustering, heavy tails, …

DATA ARE CENTRAL IN THE ANALYSIS OF CONTAGION PROCESSES

Human interactions

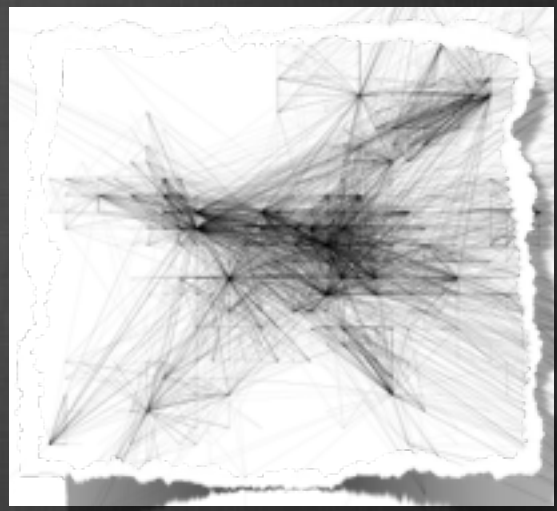


Within school contact patterns

Human Mobility

Multiscale integration of mobility networks in the analysis of potentially pandemic pathogens spread.

Social networks heterogeneities



Hubs, community, clustering, heavy tails, …

IN THE BEGINNING WAS DATA....THE OLD DATA.
LET THERE BE NOVEL DATA SOURCES: ACTIVE-VS-PASSIVE DATA COLLECTIONS

INFLUWEB

flusurvey

grippenet.fr

flu near you
do you have it in you?

FluTracking.net

Influenzanet

Participatory platforms

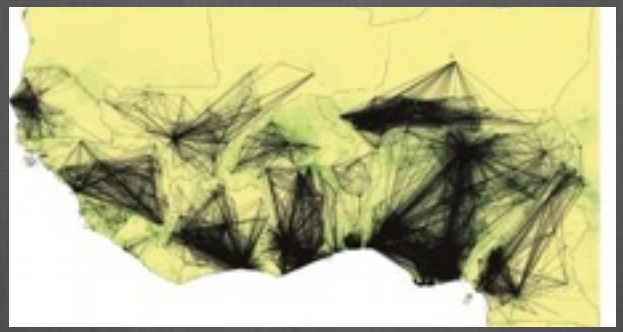
Data harvesting

ACTIVE DATA COLLECTION

- Web participatory platforms
- RAFID tags, SocioBadges
- Pervasive technologies
- Crowdsourced or manually curated data
- .....

PASSIVE DATA COLLECTION: MOBILITY AT THE "POPULATION" SCALE

Bengtsson et al. Sci. Rep 2015



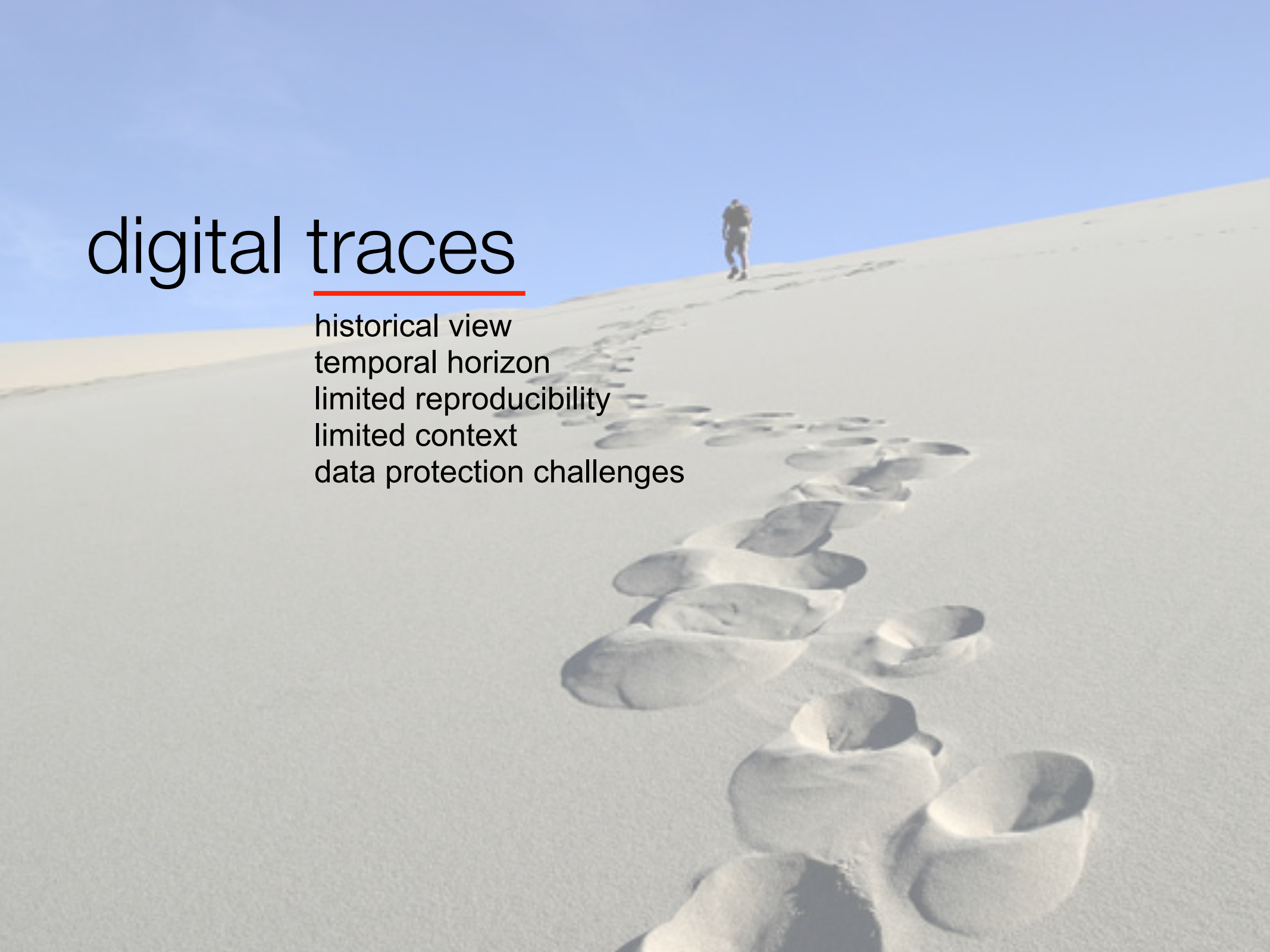
Wesolowsky et al. PLOS Curr. Out. 2014

Pedreschi et al. et al. PLOS Curr. Out. 2014

PASSIVE DATA ANALYSIS

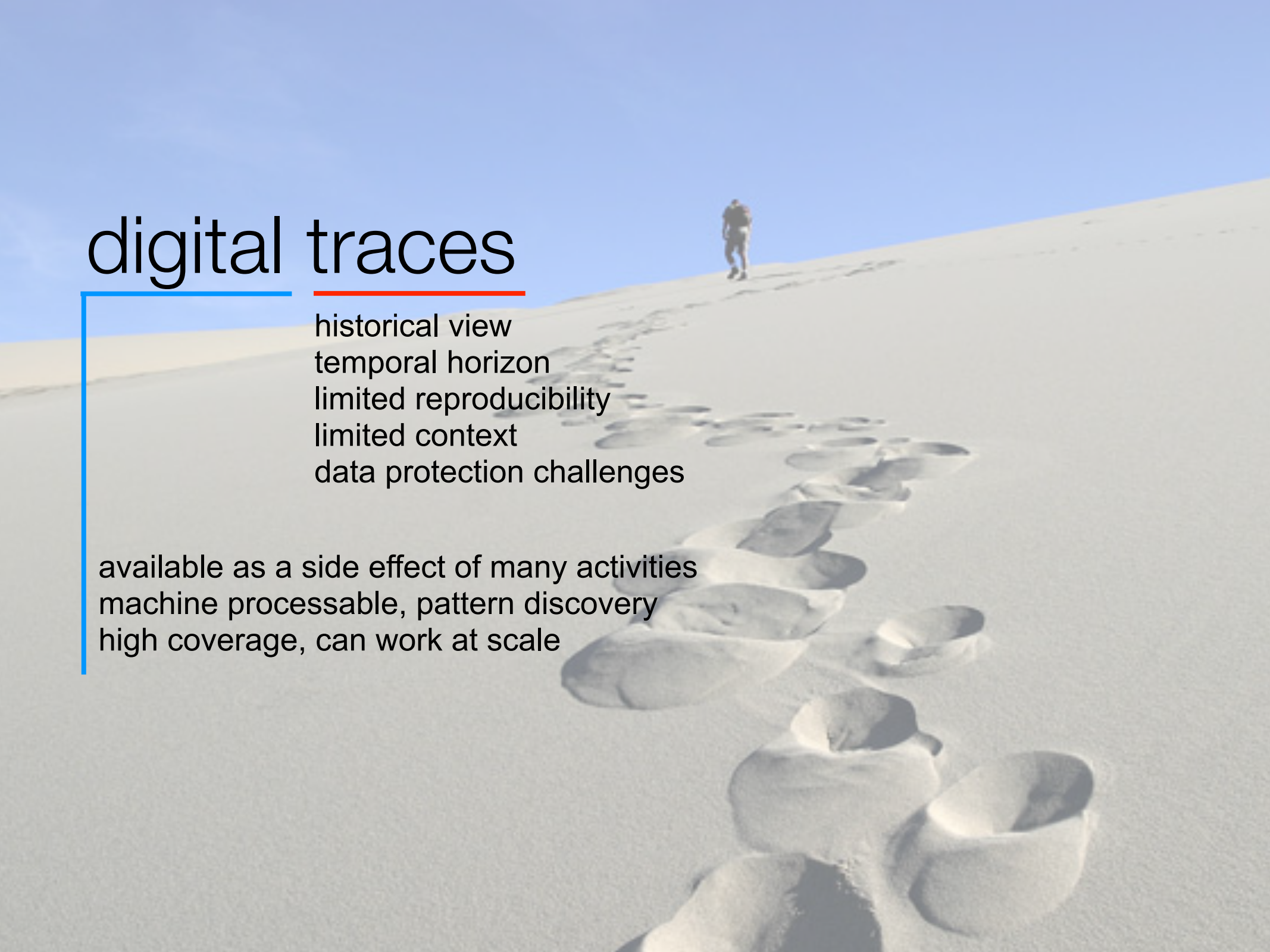
- Search engine queries (GFT)
- Twitter conversation
- Wikipedia logs
- Over the counter drugs
- Facebook postings
- Restaurant reservation cancellations
- Hospitals parking imagery.


digital traces



digital traces

historical view
temporal horizon
limited reproducibility
limited context
data protection challenges



digital traces

historical view
temporal horizon
limited reproducibility
limited context
data protection challenges

available as a side effect of many activities
machine processable, pattern discovery
high coverage, can work at scale

Google

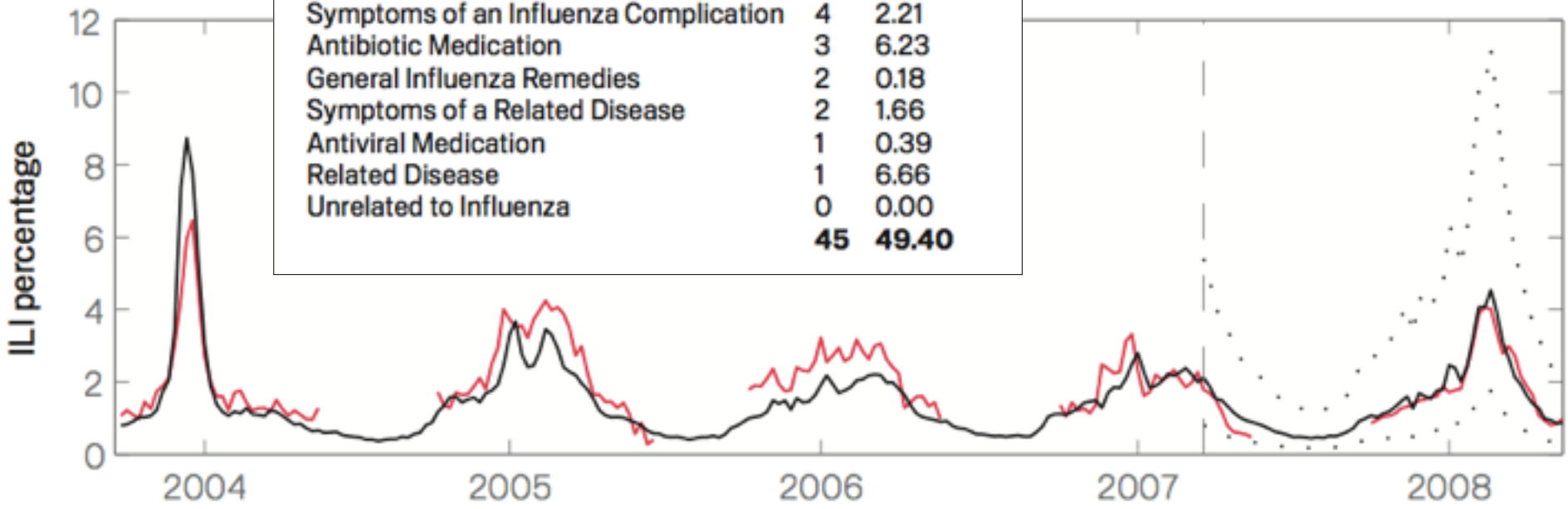
Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

[1]*Google Inc.* [2]*Centers for Disease Control and Prevention*

Search Query Topic	Top 45 Queries N	Weighted
Influenza Complication	11	18.15
Cold/Flu Remedy	8	5.05
General Influenza Symptoms	5	2.60
Term for Influenza	4	3.74
Specific Influenza Symptom	4	2.54
Symptoms of an Influenza Complication	4	2.21
Antibiotic Medication	3	6.23
General Influenza Remedies	2	0.18
Symptoms of a Related Disease	2	1.66
Antiviral Medication	1	0.39
Related Disease	1	6.66
Unrelated to Influenza	0	0.00
	45	**49.40**

google.org/flutrends



J. Ginsberg *et al.*, Nature 457, 1012 (2009)

nature
International weekly journal of science

NATURE | NEWS

عربي

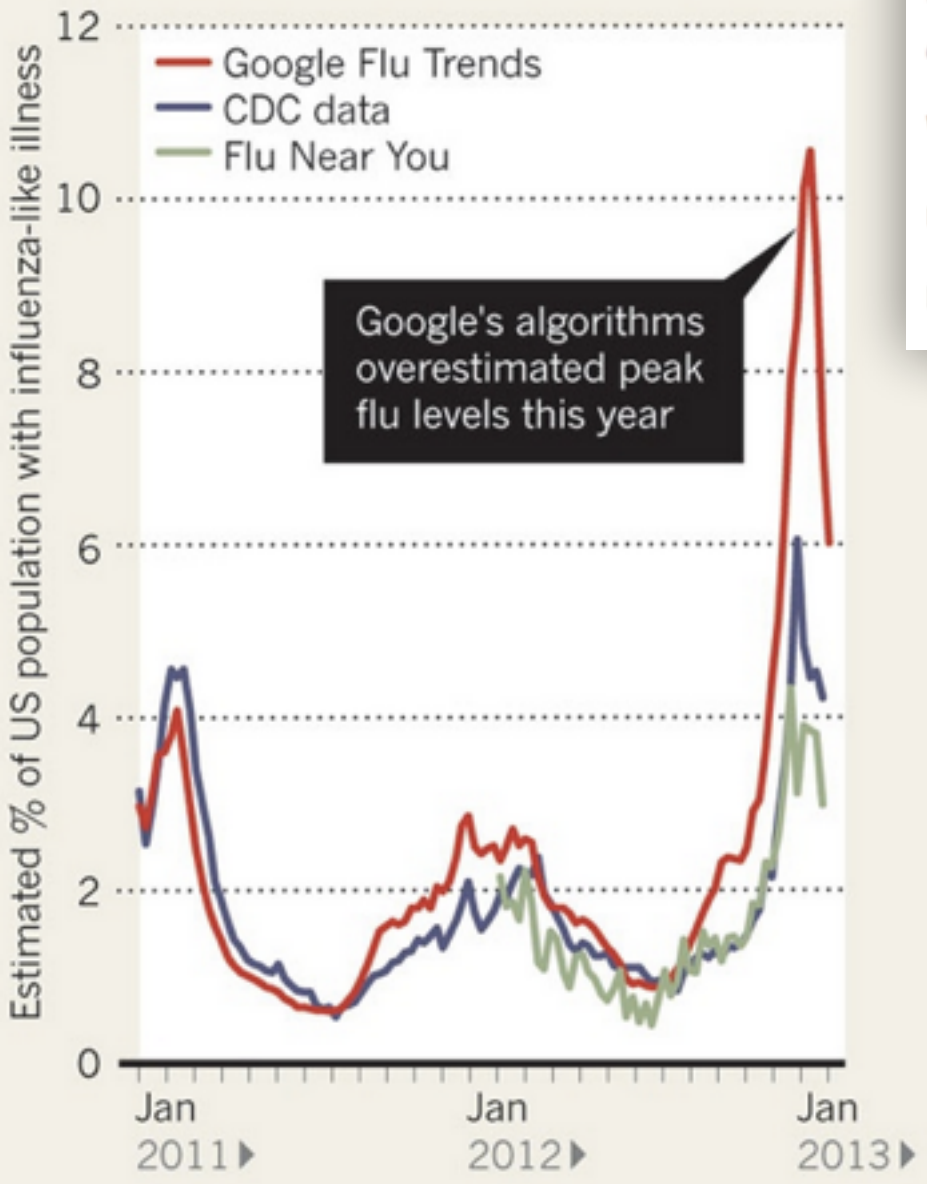
When Google got flu wrong

US outbreak foxes a leading web-based method for tracking seasonal flu.

Declan Butler

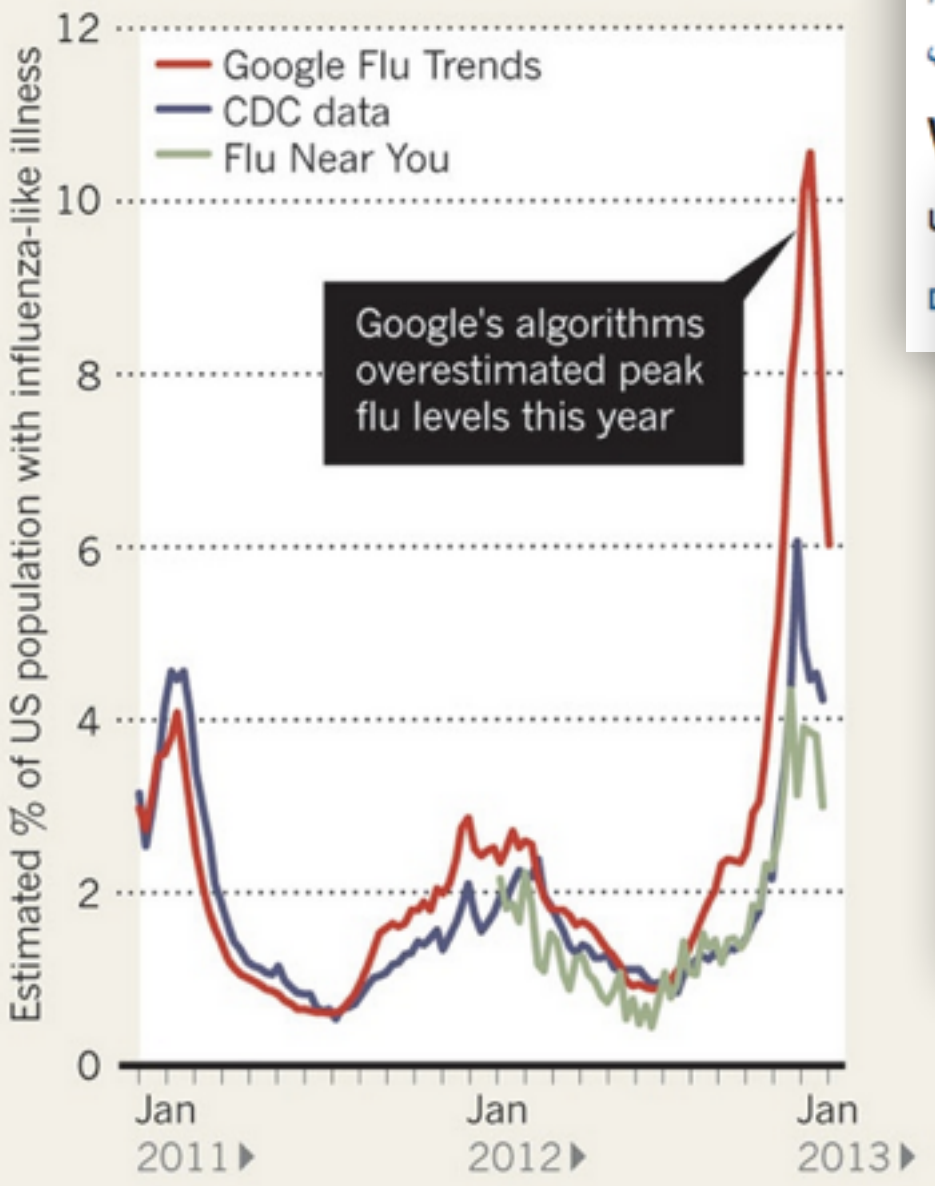
FEVER PEAKS

A comparison of three different methods of measuring the proportion of the US population with an influenza-like illness.



FEVER PEAKS

A comparison of three different methods of measuring the proportion of the US population with an influenza-like illness.



nature International weekly journal of science

When Google got flu wrong

Science 14 March 2014:
Vol. 343 no. 6176 pp. 1203–1205
DOI: 10.1126/science.1248506

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]

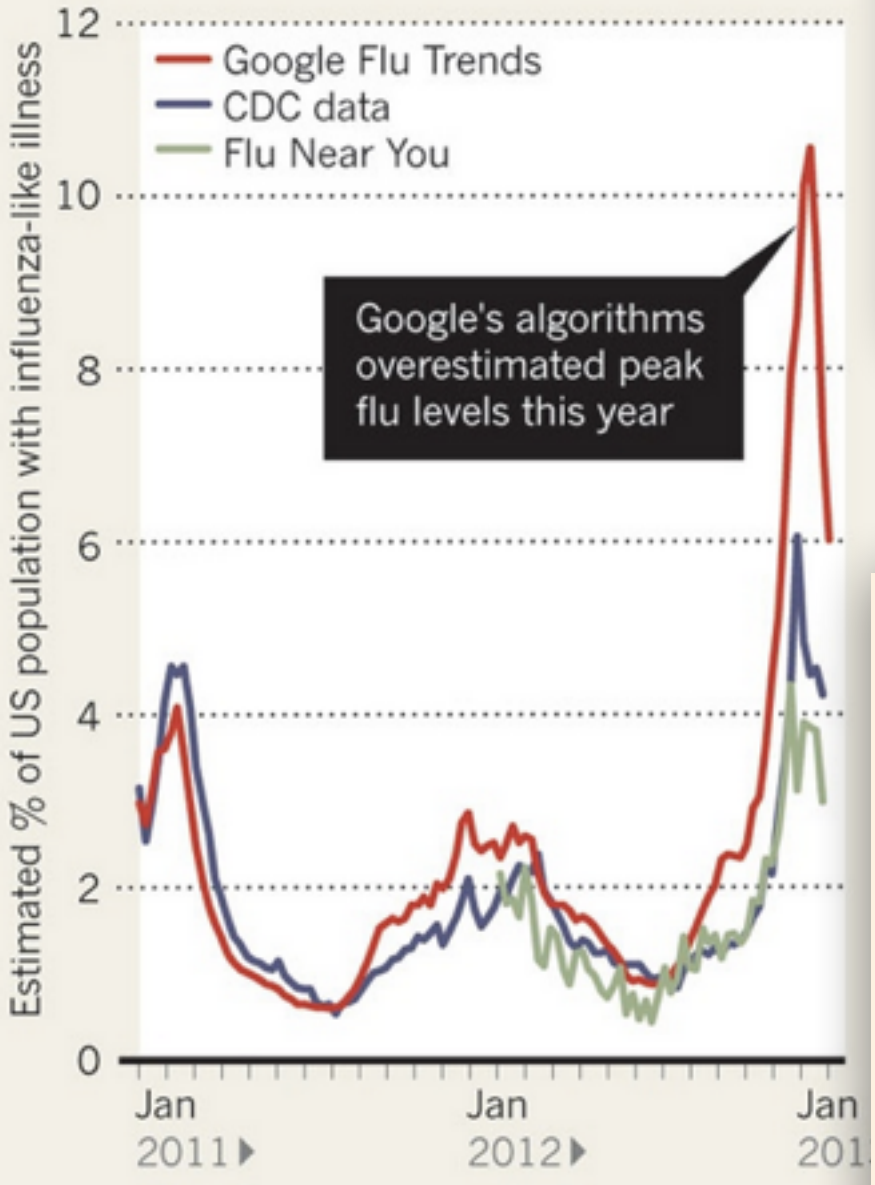
± Author Affiliations

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

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

FEVER PEAKS

A comparison of three different methods of measuring the proportion of the US population with an influenza-like illness.



nature
International weekly journal of science
Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive
Archive | Volume 494 | Issue 7436 | News | Article
NATURE | NEWS
عربي

When Google got flu wrong

Science 14 March 2014:
Vol. 343 no. 6176 pp. 1203–1205
DOI: 10.1126/science.1248506
‹ Prev | T
Read
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]

FT Magazine
Home | World | Companies | Markets | Global Economy
Arts | Magazine | Food & Drink | House & Home | Lunch with the FT | Style | Book
March 28, 2014 11:38 am

Big data: are we making a big mistake?

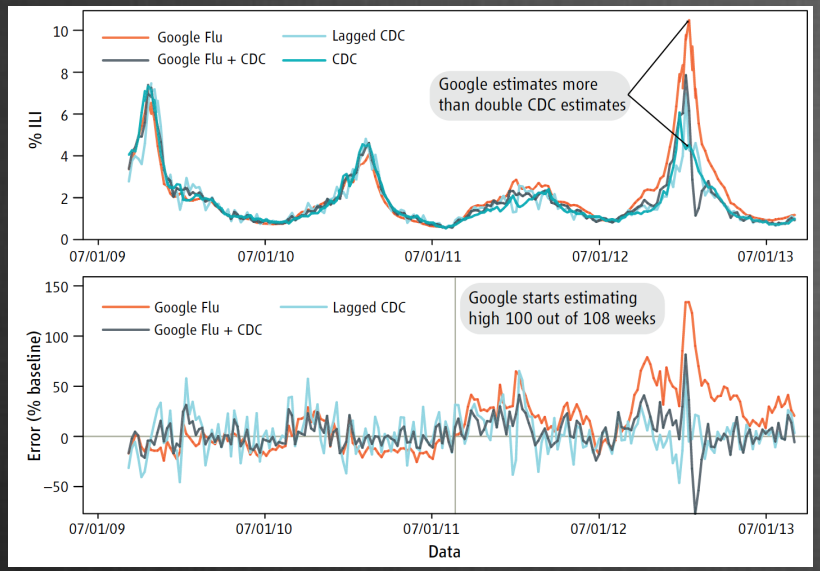
By Tim Harford

Big data is a vague term for a massive phenomenon that ha
rapidly become an obsession with entrepreneurs, scientists
governments and the media

DIFFERENT FORECAST PHILOSOPHIES

Use surrogate signal in algorithm trained on historical data (generally CDC time series) to achieve lead time (real time data collection, time-series extrapolation)

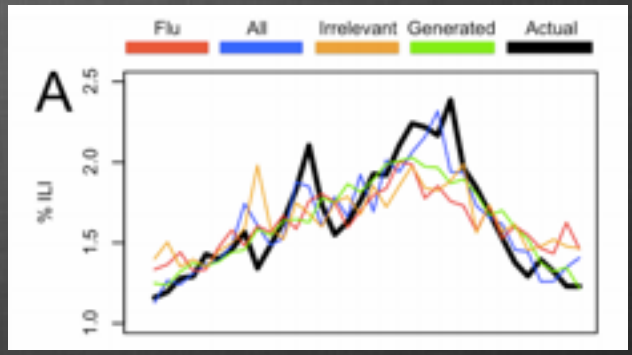
Google Flu Trend (paradigm)



Case study on GFT and other non-generative models simple Lagged regression can be 90% "good" (Lazer et al. Science 2014).

Twitter based approaches

- Salathe'; Culotta; Dredze etc. Etc. (since the first paper by Signorini et al.);
- Word selection
- Linear regression, Multiple linear regression; SVM Regresssion; EFS
- High-level geographical resolution
- Full natural language processing



Statistical biases, "zombies" etc

- Well discussed in the literature
- Similarities & difference with GFT.

THIS IS ~~FORECAST~~ NOWCAST !!!

We miss the microscopic generative foundations and models

MICRO/INDIVIDUAL BASED LARGE-SCALE COMPUTATIONAL MODELS FOR INFECTIOUS DISEASE SPREADING

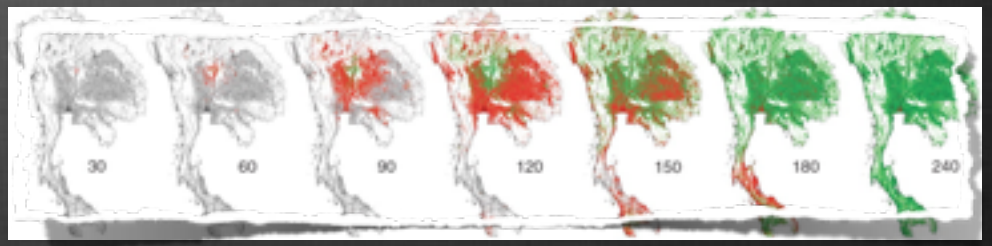
Individual based,structured metapopulation models

- Collection of sub-populations coupled by traveling individuals. [Ravchev, Longini. Mathematical Biosciences (1985) 50 urban areas worldwide;Viboud etal, Valleron et al., Brockmann et al., Colizza et al., Balcan et al.]

Agent based models

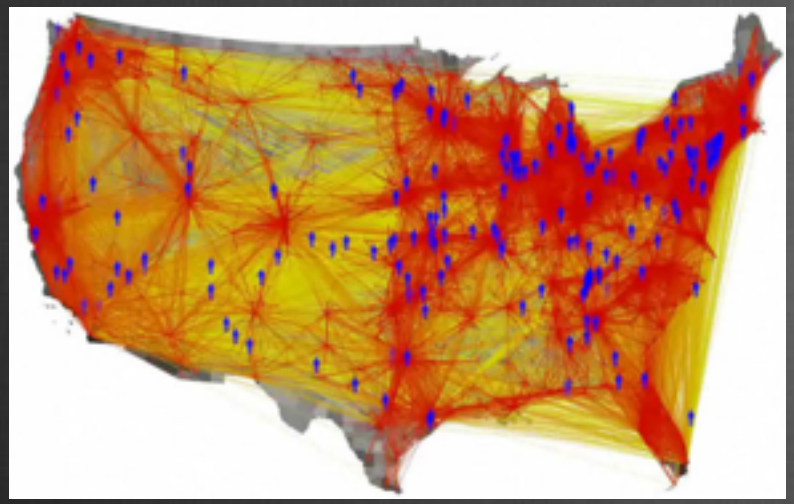
- (highly detailed individual based simulation– household+work+mobility detailed representation)
  - Eubank&Barrett, Longini, Halloran, Ferguson, Burke, Merler and the FBK

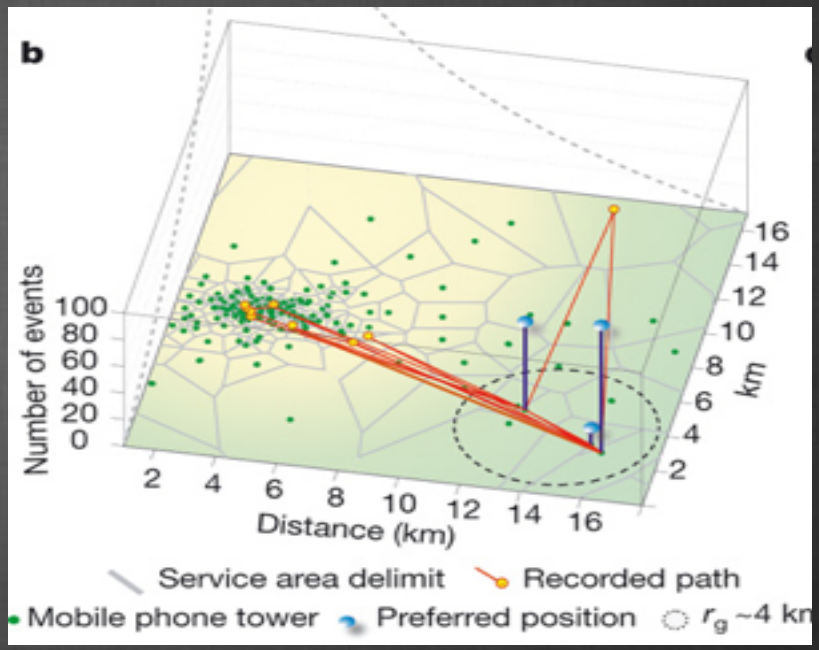
Challenges



- Data hungry models,
- Lack of information=assumption
- Set of parameters usually not available initially (vaccination rate, pre-exposure immunity etc.)

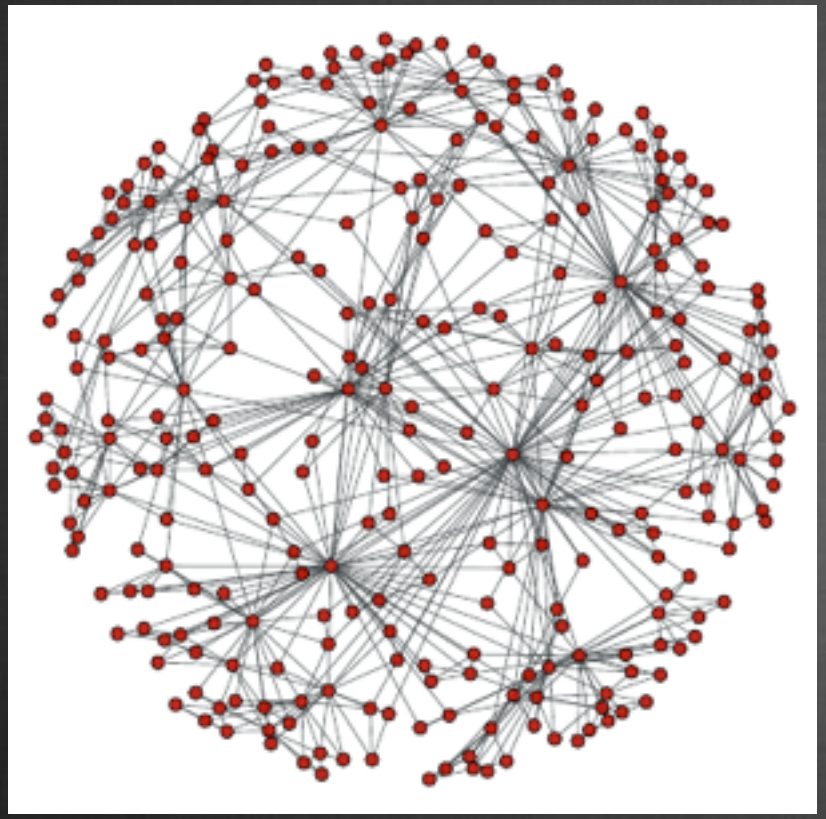
FROM GEOGRAPHY TO SOCIAL SPACE

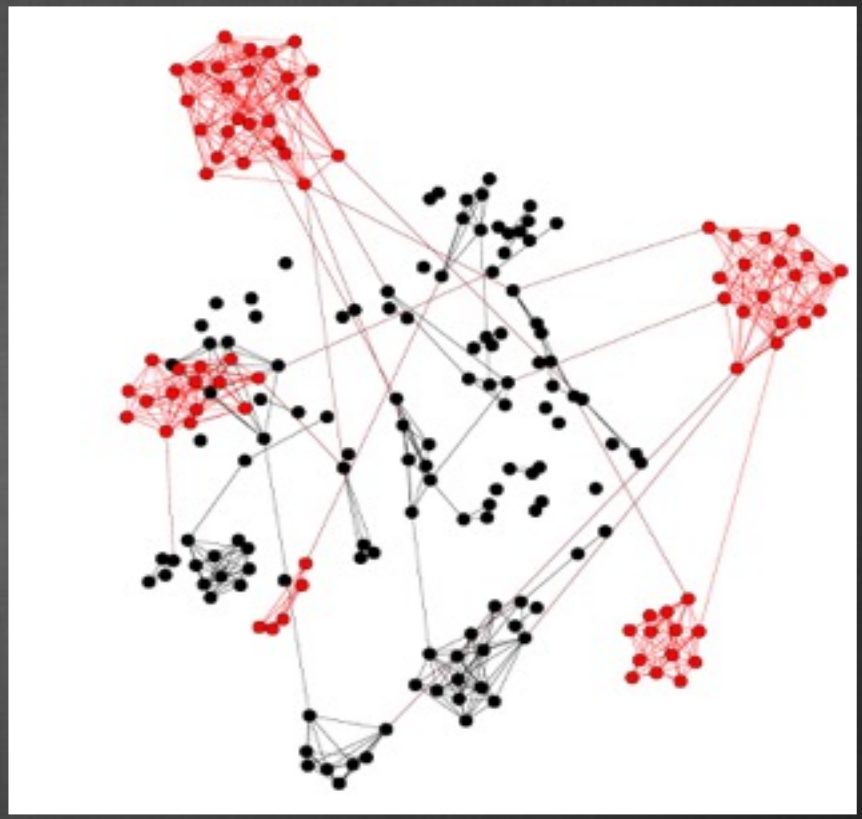




Geographical areas/census
Mobility

FROM GEOGRAPHY TO SOCIAL SPACE





Structured communities in the abstract social space define by knowledge and information

TWITTER NETWORK ON MAY 15TH PROTEST MOVEMENT IN SPAIN

Moreno et al. (BIFI, Universidad de Zaragoza)

TWITTER NETWORK ON MAY 15TH PROTEST MOVEMENT IN SPAIN

Moreno et al. (BIFI, Universidad de Zaragoza)

TRANSFORMING THE WAY WE APPROACH SOCIO-TECHNICAL SYSTEMS

- Social computational science
- Computational epidemiology
- Science of Science
- Information systems and data science
  - General classes of tools methods and models generally applicable to complex techno-social systems:
  - New techniques for the generation of models, social analytics, real time empirical data generation.
  - Scenario analysis: systemic risk, economic impact etc.