Statistical Problems in Big Data

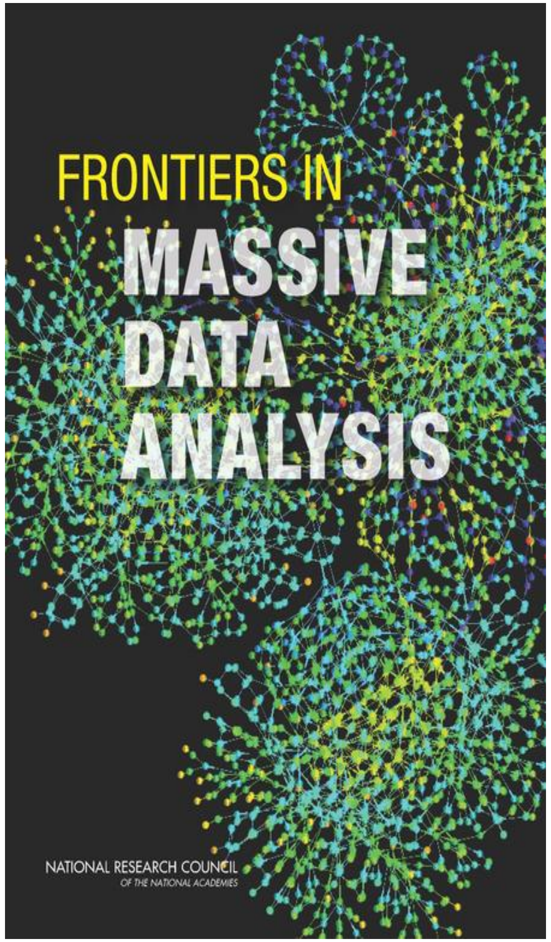
Prof Elena Kulinskaya (UEA)

Big Data: Fields and Disciplines

Actuaries are the founders of the Big Data: the CMI!
Big Data also arises in such fields as genomics, public health, environmental sciences, neuroscience, government and business.

Basic issues of management and storage have primarily implicated computer science, underpinning initiatives sometimes described as business analytics or data science.

Statistical science has not played a prominent role. Because practical problem solving has proceeded rapidly, the science has lagged behind and work to identify the statistical features associated with Big Data has been largely ad hoc.



The Alan Turing Institute: Cambridge, Edinburgh, Oxford, Warwick and UCL.

Statistical Problems in Big Data

Big Data is observational data!

Methodology of observational data analysis/ meta-analysis:

False positives arising from multiple exploratory analyses

Biases due to peculiarities of units, outcomes or settings

Missing data

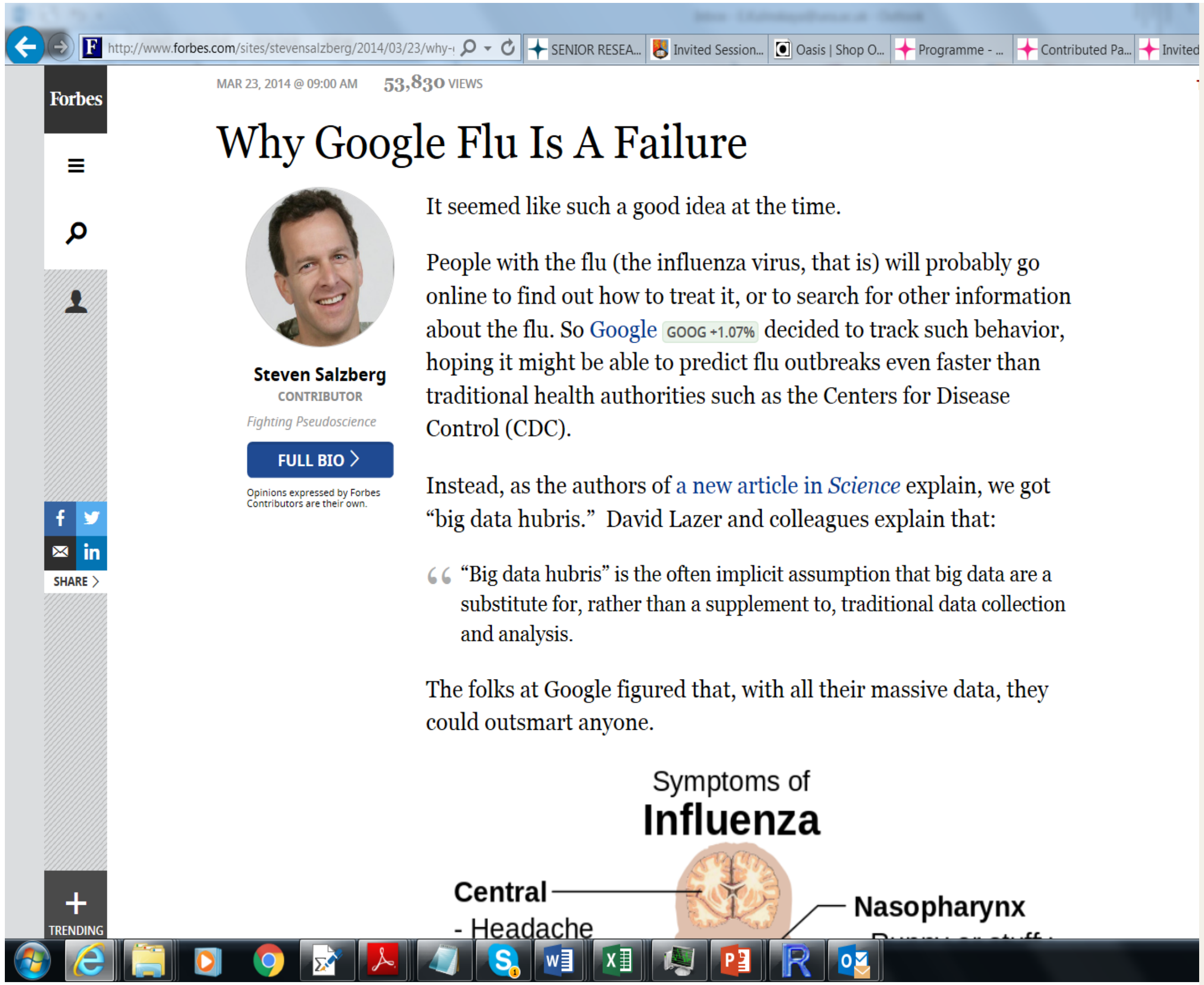
Inadequate linkage strategies

Evidence synthesis for data at varied levels of aggregation such as transaction, person, organization, community, and state

Causal Inference from (mostly) correlational data

Modelling heterogeneity

Google Flu example



MAR 23, 2014 @ 09:00 AM 53,830 VIEWS

Why Google Flu Is A Failure

Steven Salzberg
CONTRIBUTOR
Fighting Pseudoscience
FULL BIO
Opinions expressed by Forbes Contributors are their own.

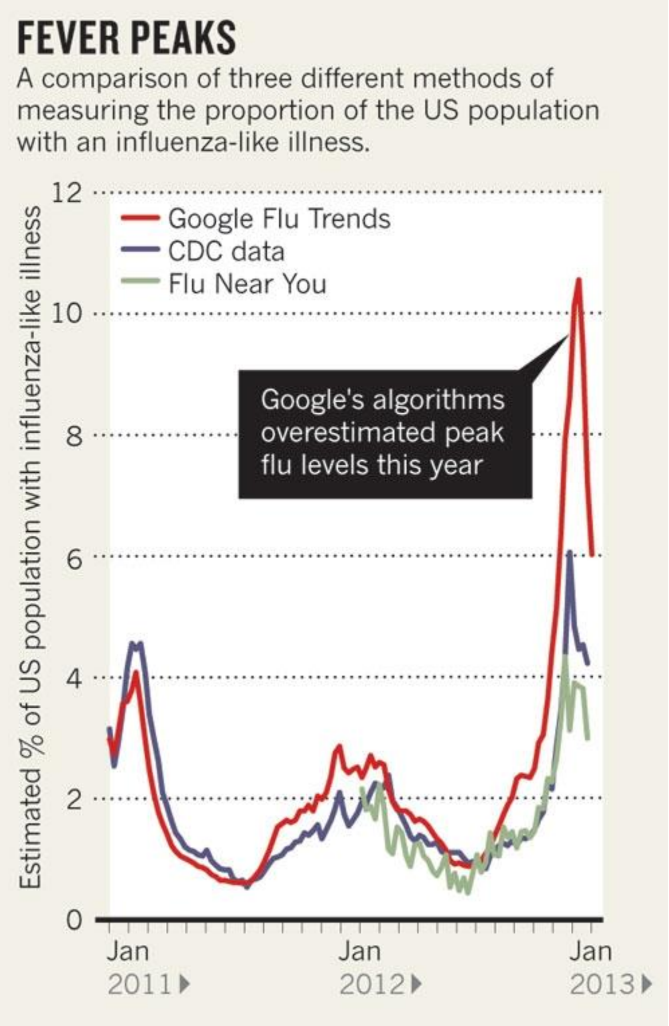
It seemed like such a good idea at the time.

People with the flu (the influenza virus, that is) will probably go online to find out how to treat it, or to search for other information about the flu. So Google GOOG +1.07% decided to track such behavior, hoping it might be able to predict flu outbreaks even faster than traditional health authorities such as the Centers for Disease Control (CDC).

Instead, as the authors of a new article in *Science* explain, we got "big data hubris." David Lazer and colleagues explain that:

> "Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis.

The folks at Google figured that, with all their massive data, they could outsmart anyone.


FEVER PEAKS

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

Google's algorithms overestimated peak flu levels this year

Use Of Big Health And Actuarial Data For Understanding Longevity And Morbidity Risks, IFoA 2016-2020

Development of novel statistical and actuarial methods for:

modelling mortality

modelling trends in morbidity

assessing basis risk

evaluating longevity improvement based on Big Health and Actuarial Data

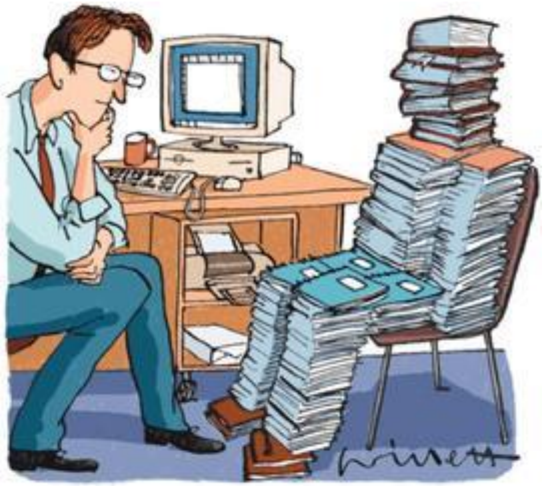
tools to forecast longevity risk of a book

Science

Scientists and insurers develop 'death clock' to predict when customers will die

A new computer algorithm will predict how long people will live CREDIT: WALES NEWS SERVICE LTD.

Data



The Health Improvement Network (THIN) data

  - Medical records from primary care
  - Representative of the UK when adjusted for deprivation

- All patients born before 1960 and followed to 01.01.2015, this includes 3.4 million patients

- Added various social economic status variables such as IMD and Mosaic

- The Continuing Mortality Investigation (CMI) data

Design and methods

The most efficient way to analyse the data of variable quality is to delete “bad” data.

For a particular condition we design a population-based prospective cohort study using an appropriate extract of the primary care data.

We intend to use a case-control design with cases matched with several controls from the same GP practice. This provides balanced and comparable cohorts of cases and controls and simplifies the study of comparatively rare conditions without loss of efficiency.

To account for the interdependence of patients from the same GP practice, we use multilevel modelling and multiple imputation.