The ever changing landscape of statistical data collection

Jelke Bethlehem

Leiden University, the Netherlands

The ever changing landscape of official statistics

The past

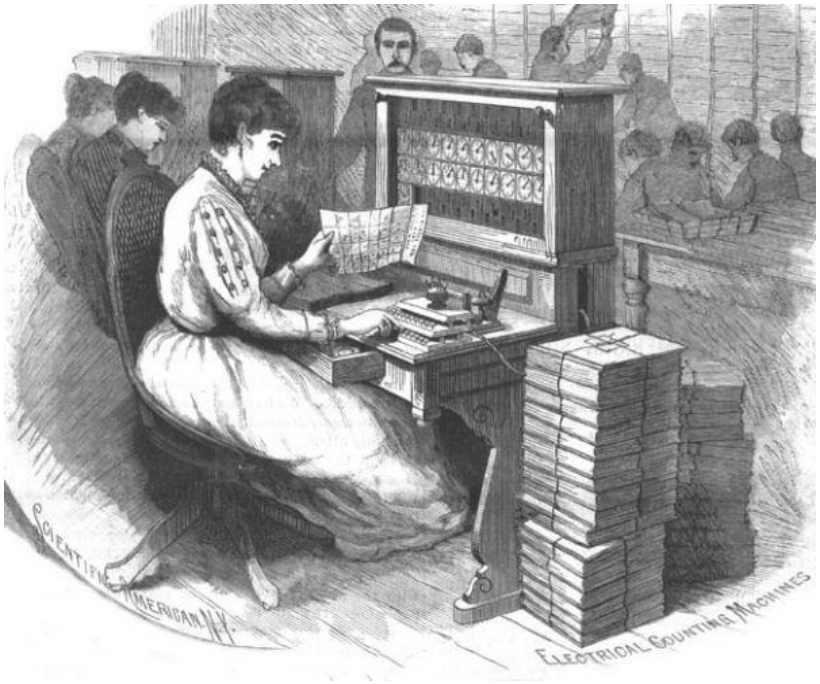
- There have always been official statistics
- The rise of survey sampling
- The role of computers

The present

- Challenges
- Online data collection

The future

- Some new approaches
- Big data



Some history

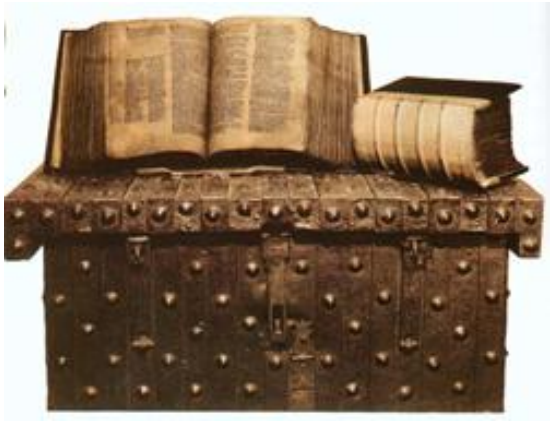
Old empires already needed statistical information

- Always complete enumeration (censuses).
- China and Egypt (1000 BC):
  Overviews for taxation and military affairs.
- Roman Empire (8 BC):
  Counts of people and their possessions.
- Example:
  Census in Bethlehem
  (Pieter Bruegel, 1566)

Some history

The Domesday Book

- Commissioned in 1086 by William the Conqueror after he conquered England from Normandy in 1066.
- Data about landowners, slaves, free people, woodland, pasture, mills, fish ponds, estimated value of the property.



The Quipucamayoc

- Statistician in the Inca Empire (1000-1500 AD).
- Data recorded on quipu's. System of knots in coloured ropes. Decimal system was used.
- RAPI: Rope-assisted personal interviewing.

Some history

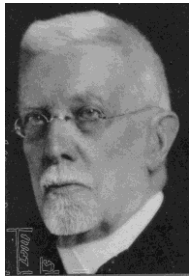
The first modern censuses

- Standardized questionnaires.
- Legal obligation to participate.
- New France (Canada): 1666, Jean Talon, N = 3215.
- Sweden: 1748, Denmark: 1769.
- Netherlands: 1795, new system of electoral constituencies in the Batavian Republic.

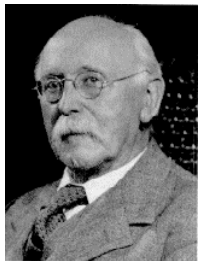
Some history

The rise of sampling

- 1895: Anders Kiaer proposed his 'Representative Method'. A kind of quota sampling. He could not compute the accuracy of his estimates.



- 1906: Arthur Bowley proposed random sampling. Probability Theory can be applied. Estimators have a normal distribution. Variances can be computed.



- 1934: Jerzy Neyman introduced the confidence interval. He also showed that quota sampling (purposive sampling) does not work.

Some history

The fundamental principles of survey sampling

- Sample selection by means of probability sampling.
- Every element must have a positive probability of selection.
- All selection probabilities must be known.

Consequences

- It is always possible to construct an unbiased estimator.
- Estimators often have a (approximately) normal distribution.
- Accuracy of estimators can be computed (confidence intervals).

Warning

- Accurate outcomes are not guaranteed for other forms of sampling (e.g. quota sampling and self-selection).

Some history

Traditional population surveys

- Situation in the Netherlands.
- From 1950: Face-to-face interviewing.
- Sample selection from population register.
- Large teams of interviewers.
- High response rates.
- Expensive and time-consuming.
- From 1980: telephone surveys.

Population register, 1946

Some history

Computer-assisted interviewing

- Since the 1980s.
- Paper questionnaires were replaced by electronic questionnaires.
- CATI: Computer-assisted telephone interviewing.
- CAPI: Computer-assisted personal interviewing.
- CASI: Computer-assisted self- interviewing.

Advantages

- Higher data quality.
- Faster data processing.
- Easier for interviewers.

The present

The rapid rise of web surveys

- Easy: simple access to large group of potential respondents.
- Cheap: no interviewers, no printing, no mailing.
- Fast: a survey can be launched very quickly.
- Everybody can do it!

The methodological challenges

- Under-coverage.
- Sample selection.
- Measurement errors.
- Nonresponse.

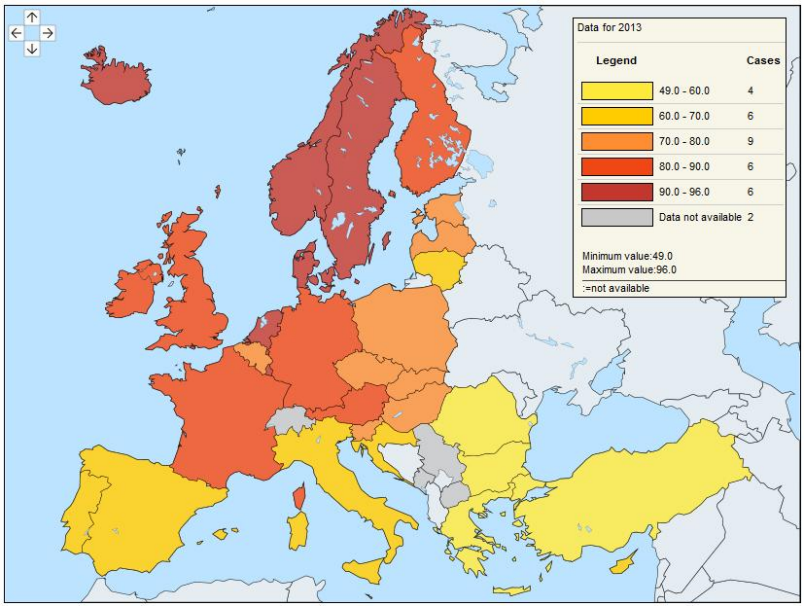
The present

Under-coverage in web surveys

- Problem: not everyone has internet.
- Elderly, low-educated and non-natives are under-represented.
- Result: biased outcomes.

Solutions

- Mixed-mode surveys.
- Supply free internet access (e.g. tablets).
- Weighting adjustment.
- Problem will disappear in future?



Top 3:
Iceland (96%)
Netherlands (95%)
Norway (94%)

Bottom 3:
Greece (56%)
Bulgaria (54%)
Turkey (49%)

Source: Eurostat, 2013

The present

Sample selection for web surveys

- How to apply probability sampling?
- No sampling frame of e-mail addresses available.
- Other modes of recruitment are expensive and time consuming.

Dangers of self-selection

- Unknown selection probabilities: no unbiased estimators.
- Participants from outside target population.
- Risk of manipulation.

Local elections in Amsterdam. Who won the debate (Jan. 2014)?

The present

Measurement errors in web surveys

- There are no interviewers. Respondents are on their own.
- Respondents are not interested in the survey.
- Participating is not important for them.
- They do not read the questions, but just scan through them.
- They know there is no penalty for giving a wrong answer.

Satisficing

- Respondents do not give the optimal answer, but the first more or less acceptable answer that comes into mind.
- For example: primacy effect, selecting *don't know*, selecting the neutral, middle option.

The present

Budget cuts

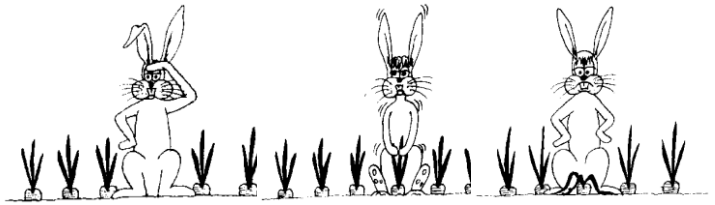
- Interviewer-assisted surveys (CAPI, CATI) become too expensive.
- Can we change to online surveys without sacrificing quality?

Lack of sampling frames

- There are no proper sampling frames for online surveys.
- It becomes more and more difficult to select a sample for a telephone survey.

Increasing nonresponse problems

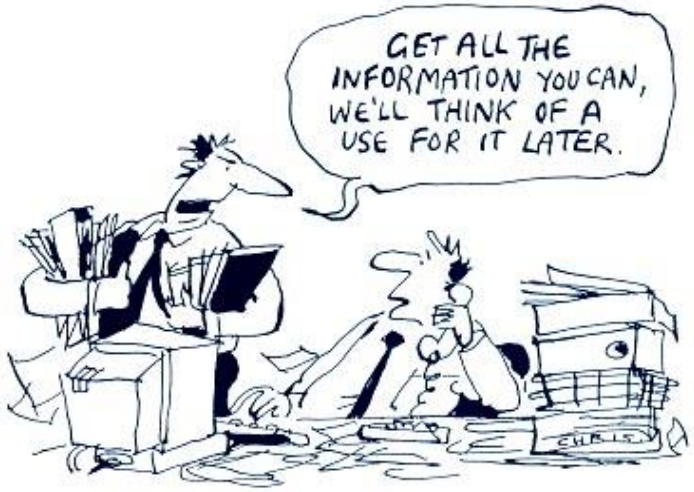
- Response rates < 10% for telephone surveys (RDD, US).
- Response rates < 40% for online surveys.
- Do the principles of probability sampling still apply?



The future

How to collect data in the future?

- Abandon probability sampling. Use self-selection sampling.
- Abandon probability sampling. Use model-based estimation.
- Abandon surveys. Use big data.
- Continue with probability sampling. Invest in correction techniques.



The future

Self-selection sampling

- Replace probability sampling by *self-selection sampling*.
- It is much easier to collect data with self-selection surveys.
- Correct the lack of representativity by adjustment weighting.
- Next step:
  A large self-selection web panel.

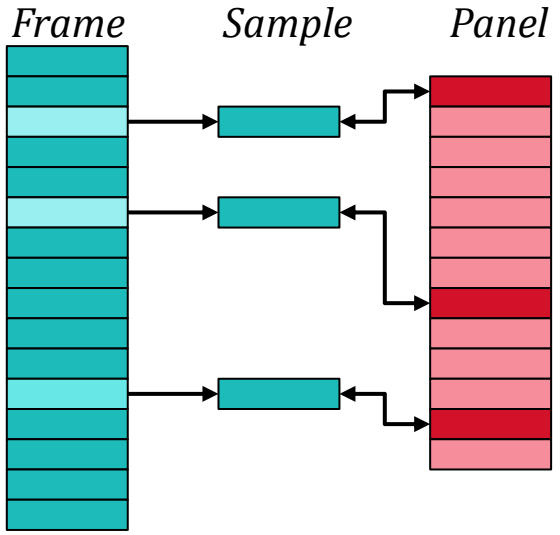
However …

- The representativity problems of self-selection surveys are much bigger than those of probability surveys + nonresponse.
- Is it really possible to remove the bias of the estimates? Not, if specific subpopulations are missing completely.

The future

Self-selection sampling

- Is *sample matching* the solution?
- Random sample from sampling frame (population register).
- Locate similar people in a large self-selection panel.
- Interview these people (and not the people in the sampling frame).
- No non-response.



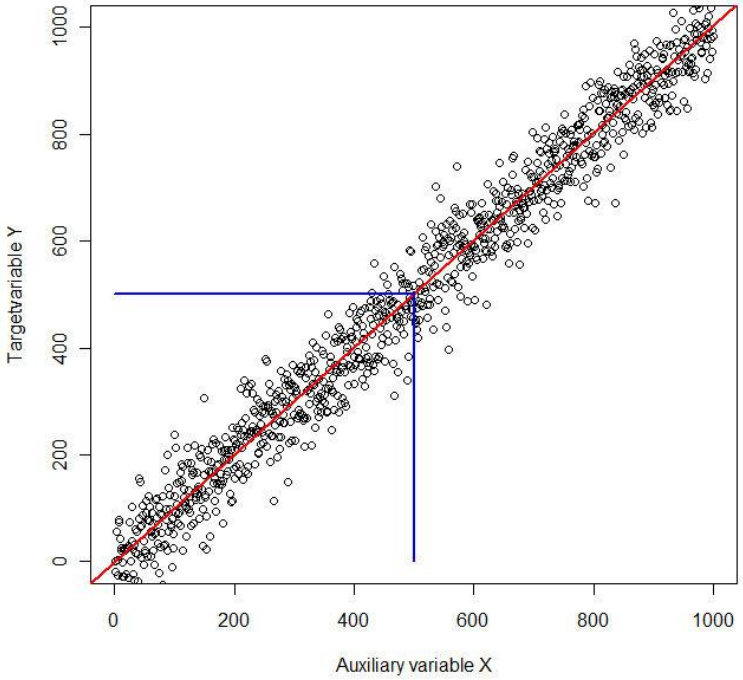
However …

- Estimates are similar to weighting a sample from a self-selection panel.
- Only effective if proper auxiliary variables are available.

The future

Model-based estimation

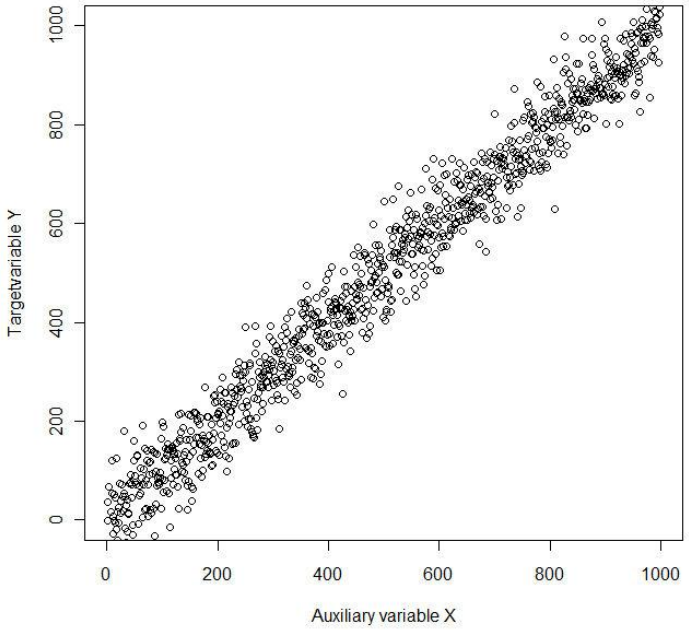
- Traditional approach: design-based approach.
- Assume a linear relationship between target variable and auxiliary variable.
- Draw a random sample.
- Estimate regression model.
- Use the regression estimator:
  $\bar{y}_{REG} = \bar{y} - b\left(\bar{x} - \bar{X}\right)$
- Robust estimator. Also unbiased if model does not hold
- Less precise if wrong model is assumed.



The future

Model-based estimation

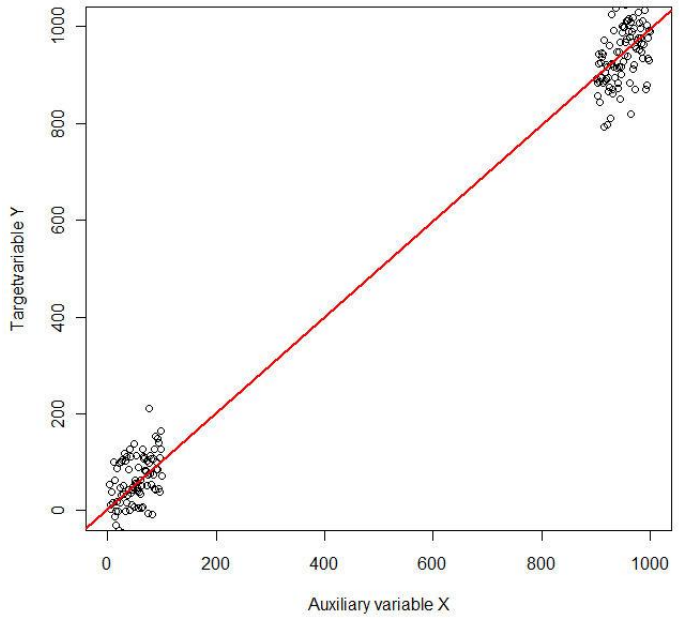
- Model-based approach: forget about sampling.
- Fit a model that explains target variable from a set of auxiliary variables. For example: $Y_k = \alpha + \beta X_k + \varepsilon_k$, with $\varepsilon_k \sim N(0, \sigma)$.
- Predict unknown values of Y by model.
- Prediction of population mean: take mean of known and predicted values of Y.



The future

Model-based estimation

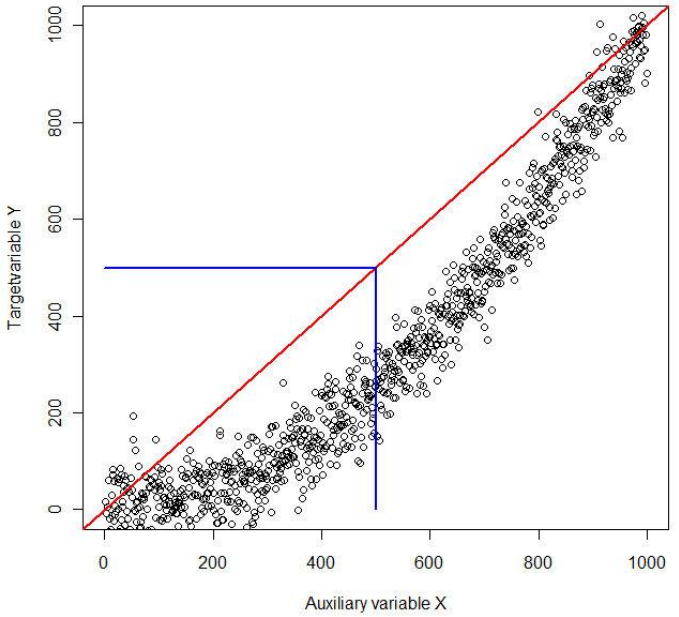
- Model-based approach: forget about sampling.
- Fit a model that explains target variable from a set of auxiliary variables. For example: $Y_k = \alpha + \beta X_k + \varepsilon_k$, with $\varepsilon_k \sim N(0, \sigma)$.
- Predict unknown values of Y by model.
- Prediction of population mean: take mean of known and predicted values of Y.
- Prediction is accurate for observations near upper and lower bound.
- But prediction fails if model does not hold a any more.



The future

Model-based estimation

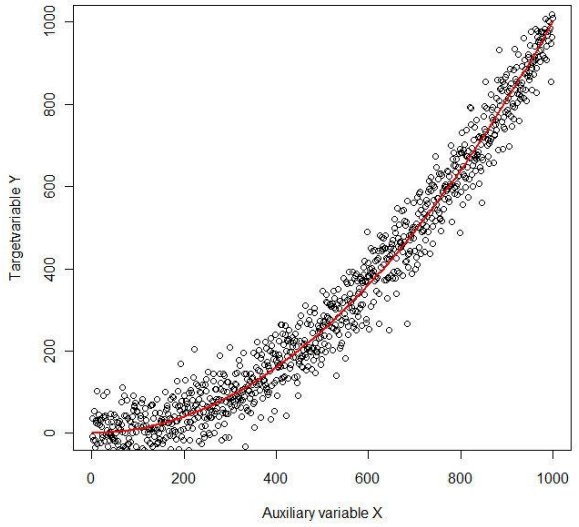
- Model-based approach: forget about sampling.
- Fit a model that explains target variable from a set of auxiliary variables. For example: $Y_k = \alpha + \beta X_k + \varepsilon_k$, with $\varepsilon_k \sim N(0, \sigma)$.
- Predict unknown values of Y by model.
- Prediction of population mean: take mean of known and predicted values of Y.
- Prediction is accurate for observations near upper and lower bound.
- But prediction fails if model does not hold a any more.



The future

Model-based estimation

- Model-based estimates can produce very accurate estimates, but only if the model is correct.
- Model-based estimates may not be robust against misspecification of models.
- In practice, it should regularly be checked whether the model is still valid. This requires sampling.
- Protection against misspecification is possible, but this also requires sampling.



The future

Use of big data

- Big data: very large data sets that are difficult to analyse with traditional statistical tools.
- Big data have always been here. Only it was called differently: data mining (2000).
- Is big data a hype, or a marketing trick, or is it useful new approach?
- Limited applications.
  Is it a lot of data looking for a problem,
  or is it a problem looking for data?

The future

Use of big data

- Many national statistical institutes already use big data sets: register data, and other data from administrative sources.
- Multipurpose population register: data source, sampling frame, and source of weighting adjustment variables.

Issues

- Owned by different organisation.
- Different purpose, different definitions.
- No control over data collection.
- Questions may change or disappear.
- Registers are not without errors.
- Sufficient quality control?

The future

Big data

- Gartner (2001): large data sets that become available at high speed, and that are of a diverse nature.
- Tim Harford (2014): "*Big data is like teenager sex. Everyone is talking about it. Nobody knows how to do it. Everybody claims they are doing it. Everybody assumes everybody else is doing it*".

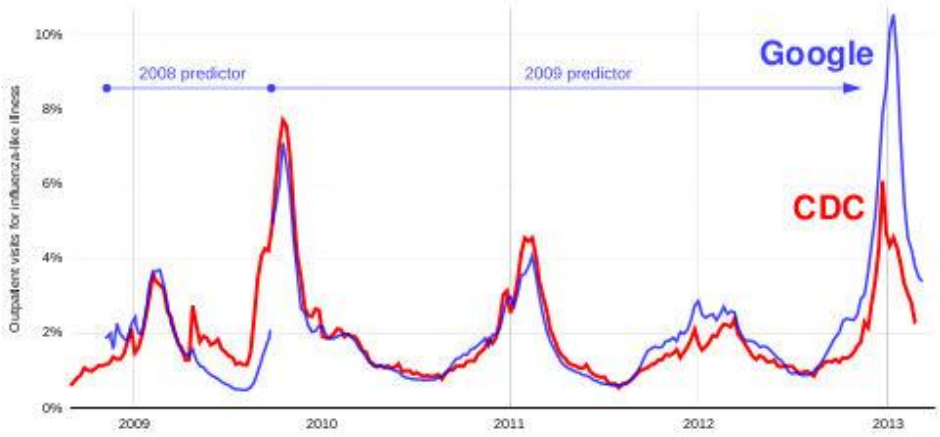
The future

Big data – No theory required

- With enough data, the number speak for themselves (Wired, 2008).
- No theory is needed. Just use the data to build a prediction model based on detected correlations.
- But beware: models may fail!

Example: Google Flu Trends (GFT)

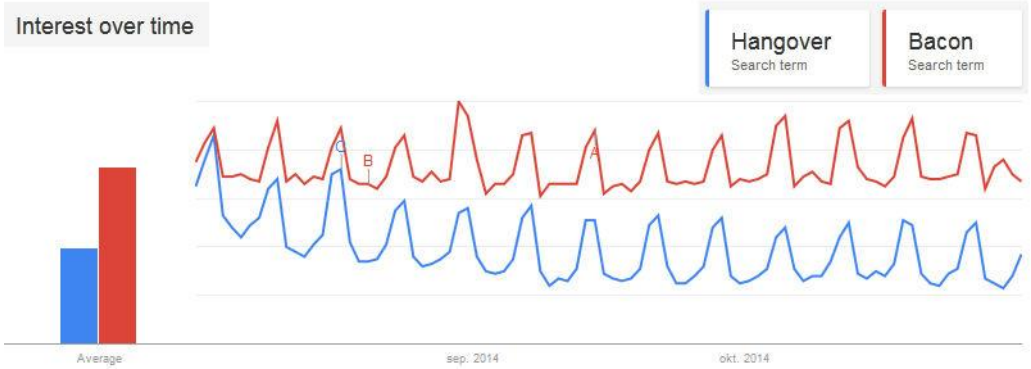
- Model based on search behaviour in Google.
- Model worked well for three years.
- In 2013, the model proved wrong by a factor 2.



The future

Big data – Correlation does not imply causation

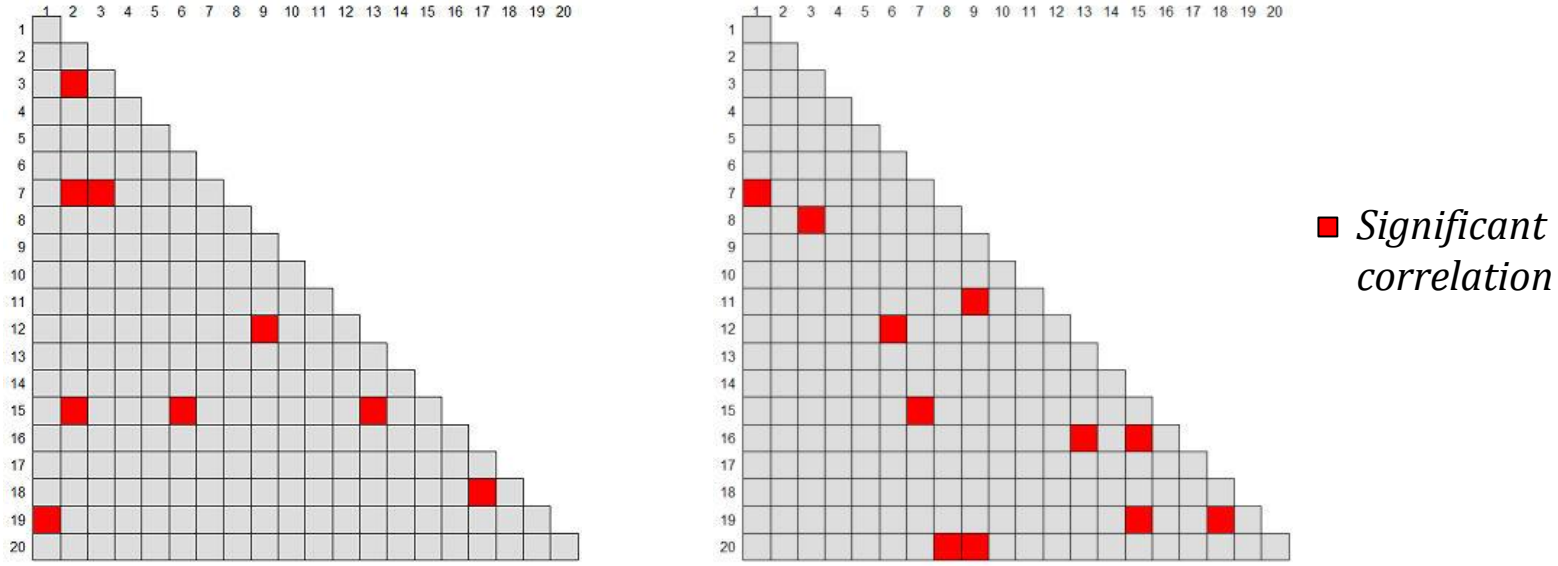
- Big data use seems to focus on detecting correlations, and not on explaining why thing are happening (causal relationships).
- If two trends fluctuate in exactly the same way, this does not mean that one trend is caused by the other.
- There can be a spurious relationship: there is a third (unobserved) variable causing both observed variables.
- Example: the correlation between searching for ‘hangover’ and ‘bacon’:



The future

Big data – Fake correlations

- Even for random noise, 5% of the correlations are significant.
- Data should be split in two portions: one for exploration, and one for hypothesis testing.
- Example: random, independent drawings from normal distribution.

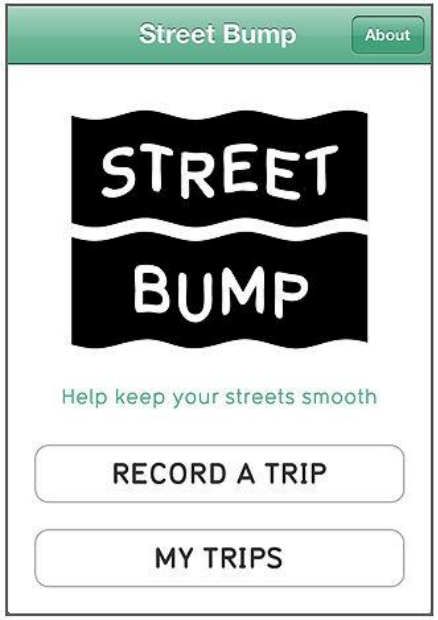


The future

Big data – Lack of representativity

- We do not need big data! We need representative data.
- Big data sets often cover only part of the population. We should not forget the rest of it.

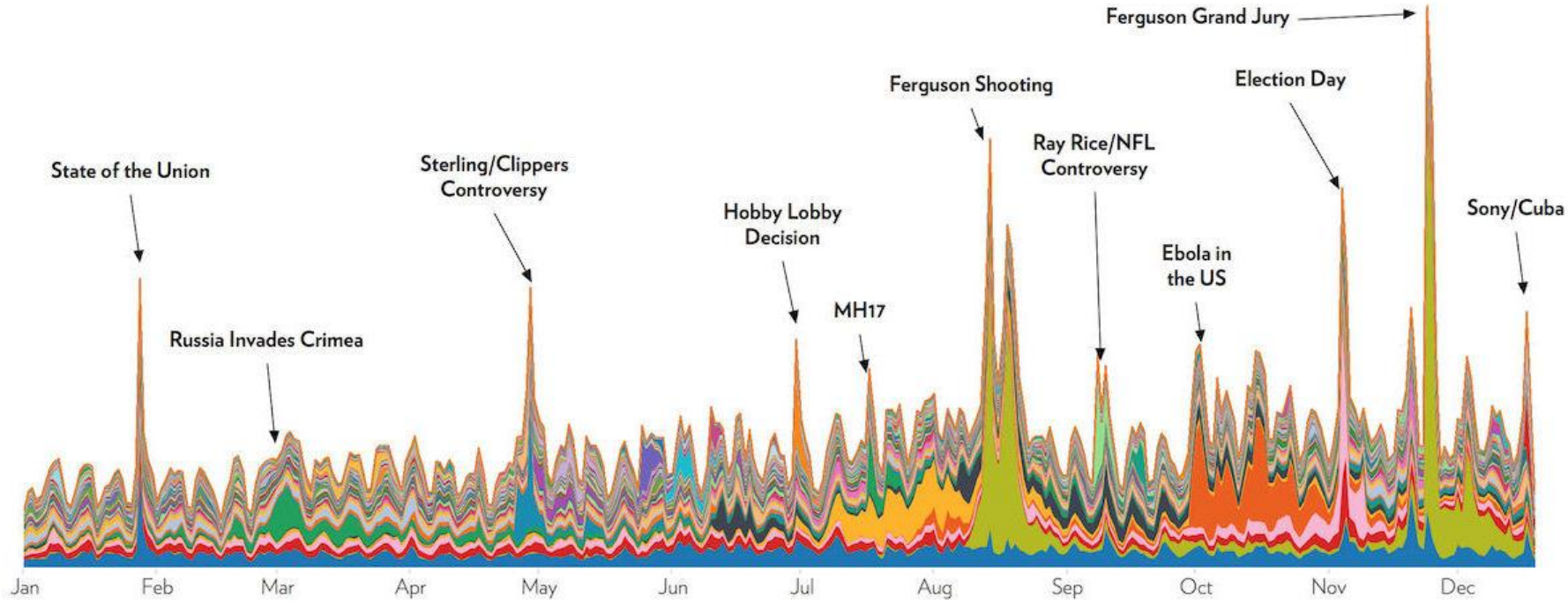
- Example 1: the Boston Street Bump
- Clever idea: smartphone records potholes in the roads. Cheap and fast.
- Unfortunately, not everyone had a smartphone. So only potholes in the richer neighbourhoods were detected.



The future

Big data – Lack of representativity

- Topics of 184.5 million tweets in 2014 (from Echelon Insights).
- Which population is described here?
- A lot of data, but is it representative?



The future

Big data – Lack of representativity

- We do not need big data! We need representative data.

- Example 2: the presidential elections in the U.S. in 1936. Candidates: Alf Landon (R) and Franklin Roosevelt (D).
- The Literary Digest poll. A sample of more than 2 million (lists of car owners and telephone directories).
- The Gallup poll: A (quota) sample of 50,000.
- The Literary Digest poll was wrong (Landon). Republicans were over-represented in the sample.


Today

AMERICA SPEAKS

THE NATIONAL WEEKLY POLL of PUBLIC OPINION

Next Sunday

Institute Forecasts the Re-election of Franklin D. Roosevelt, Gives Him 54% of Popular Vote, Minimum of 315 Electors

Major Party Percent Is 55.7; New York in F.D.R. 'Sure' Column

Election Forecast

Election Will Test Clashing Poll Methods

The future

There still is a future van probability-based surveys

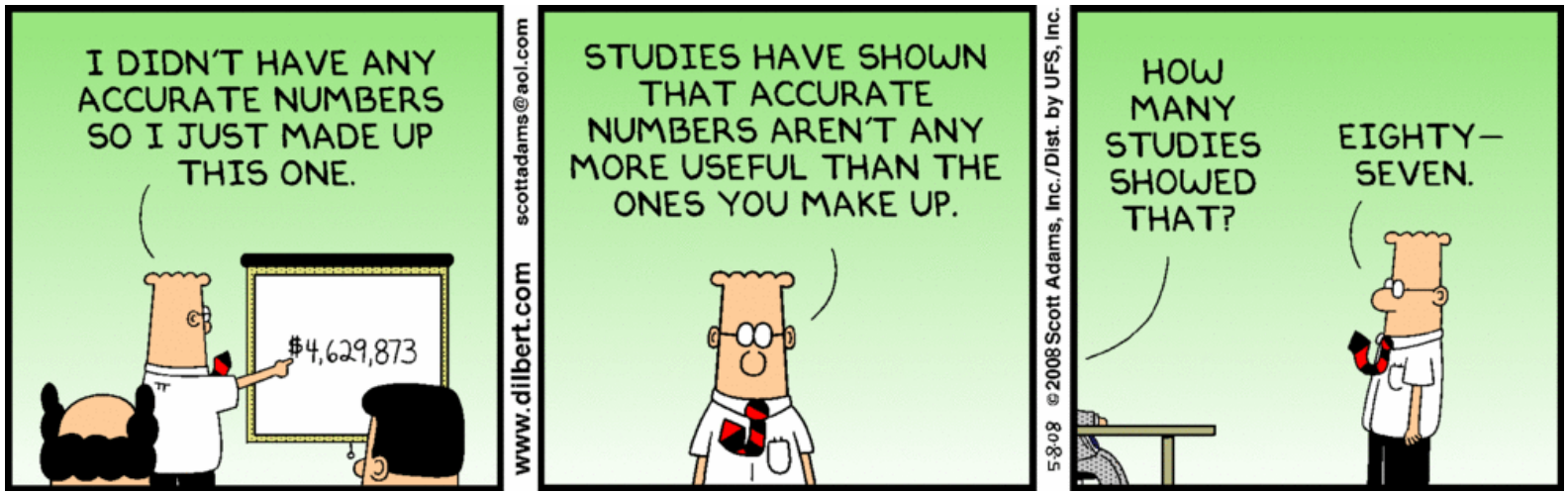
- Do not throw out the baby with the bath water!

We need surveys for …

- Topics that are not covered by other data sets..
- Checking models.
- Quality control of registers.

Invest in …

- Better correction techniques.
- Better (more effective) auxiliary variables.


I DIDN'T HAVE ANY ACCURATE NUMBERS SO I JUST MADE UP THIS ONE.
$4,629,873
www.dilbert.com scottadams@aol.com
STUDIES HAVE SHOWN THAT ACCURATE NUMBERS AREN'T ANY MORE USEFUL THAN THE ONES YOU MAKE UP.
5-8-08 ©2008 Scott Adams, Inc./Dist. by UFS, Inc.
HOW MANY STUDIES SHOWED THAT?
EIGHTY-SEVEN.

The end