On promises and pitfalls of big data:
A case study of Google Flu Trends

Paul Ormerod*, Rickard Nyman* & Alexander Bentley**

*Centre for the Study of Decision-Making Uncertainty
University College London

**Department of Anthropology
University of Bristol

The general problem; naïve brute force search vs. theory driven search

- If the hypothesis tests performed are ‘unrestricted’ and evaluated ‘independently’, it is easy to argue that most results, as judged significant, are in fact **incorrect**
  - Major concern even before big data due to a ‘positive publication bias’
  - With big data, this situation is now very much a reality, summary:

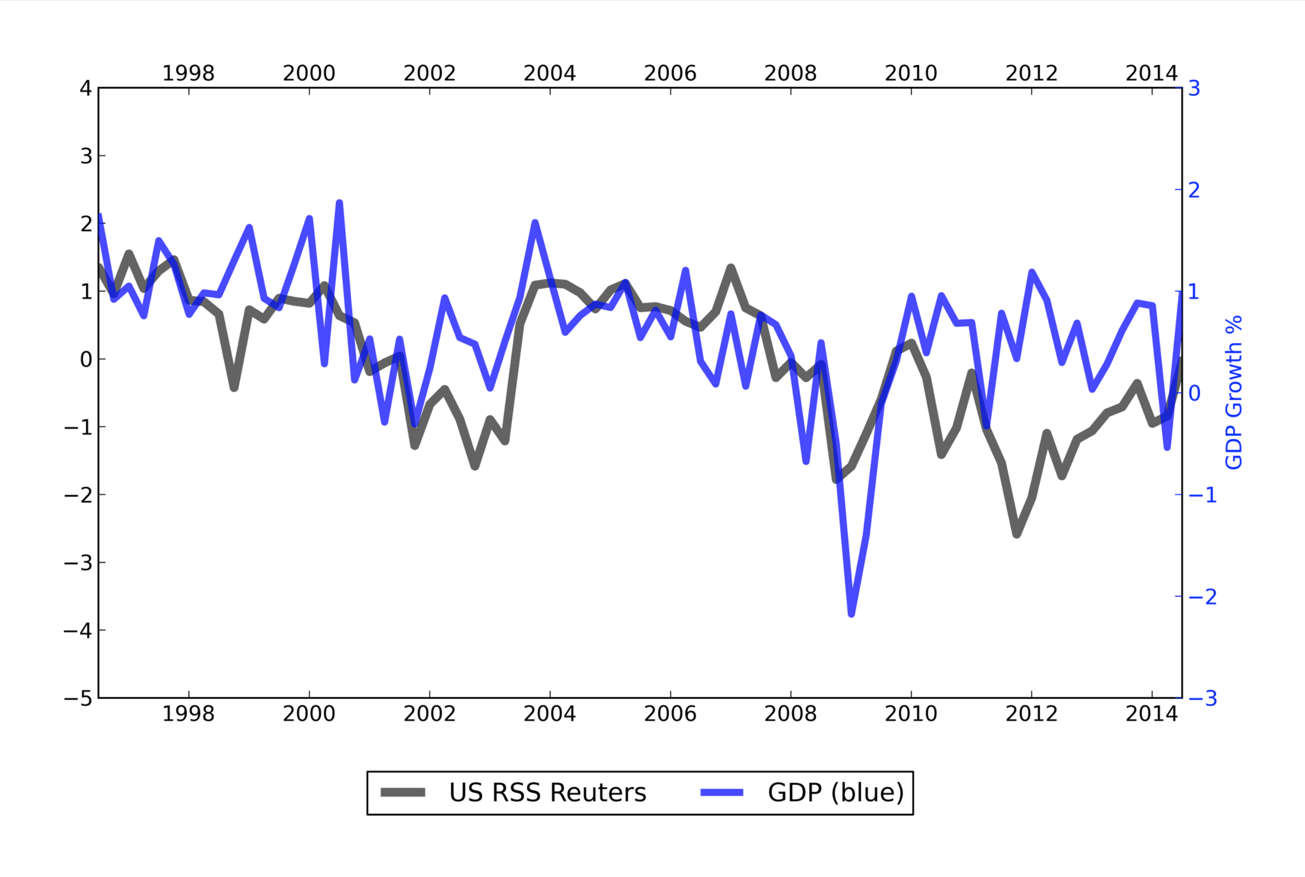
‘Test space’ features	Typical big data approach	Theory driven approach
Unrestricted test/hypothesis space?	**Yes**, ‘unlimited’ number of tests and features on big data. Tests are not filtered by expectations	**No**, effort + time to perform a test. Features and hypotheses are filtered by theory. Tests are not ‘wasted’. Greatly limits false positives
Independent interpretation of results?	**Yes**, in most cases. The ‘negative’ results are discarded even though potentially relevant to ‘positive’ results	**No**, one test result effects interpretation of others as they rely on the validity of the same theory (IT IS a problem if two research teams do not share results)

Illustration 1: Google Flu Trends

- In a recent paper published in Nature, engineers at Google were able to estimate the incidence of flu based on the number of flu related search queries, with astonishing accuracy
- However, later these search queries dramatically overestimated the incidence of flu in several cases
- We consider a number of countries and years when Google estimates were accurate and cases were they overestimated
- Using a well-known non-linear model from the social sciences we identify the relative strength of ‘socially influenced’ and ‘independent’ searches
- We find that the ‘social’ coefficient of the overestimated cases dominated that of the accurate estimates, and vice versa
  - Strongly suggests that overestimates were caused by social influence
- The details can be found in http://arxiv.org/abs/1408.0699

Illustration 2: A theory driven approach

- Forec viction
  Narra
  - The balance
    bet ction
  - We
  - Pre keeping
    the in the
    tex
  - The d UK)
    nev nd UK)
    GD
  - Wit



Conclusion

- ICT and big data offer enormous potential
  - For example, to measure public health concerns in regions where official estimates are uncertain
- A general problem is that of restricting the hypothesis space and evaluating results on a theoretical basis
  - Traditionally with Big Data, the tests performed are less ‘restricted’ and the test results are evaluated ‘independently’
  - Negative findings are ignored because tests are not related via theory, and positive findings are often post-theorized
  - **Theory driven testing limits these two problems with big data**
- Illustration of the problem; estimates using Google Flu Trends sporadically failed
  - A word of caution, we don’t know what search terms Google used
  - Future research might try to detect very early on which motivation is dominating (information about social influence could help)
- Seen an illustration of a theory driven approach