BANK OF ENGLAND

Advanced Analytics at the Bank of England

Presentation to the Riksbank conference on Big Data: Building Strategies for Central Banks in Light of the Data Revolution

9 September 2015

Outline

- Why are we interested?
- What are we interested in?
  - Matched micro data sets
  - Text mining
  - Visualisation
  - Machine learning

- But… there are no free lunches, so what’s the bill?

Why are we interested in Big Data?

- What do we mean by the term?
  - Fuzzy meaning, covering data, techniques and attitude
- Why are we interested?
  - Change of responsibilities
    - The arrival of the PRA
  - Change of opportunity
    - More data, increased computing power, technical advances
  - Change of circumstances
    - Lessons from the financial crisis
  - Change of philosophy
    - Inductive vs deductive reasoning

Why are we interested in Big Data?

- What do we mean by the term
  - Very loose meaning, covering data, techniques and attitude
- Why are we interested?
  - Change of responsibilities
    - The arrival of the PRA
  - Change of opportunity
    - More data, increased computing power, technical advances
  - Change of circumstances
    - Lessons from the financial crisis
  - Change of philosophy
    - Inductive vs deductive reasoning

Why are we interested in Big Data?

- What do we mean by the term
  - Very loose meaning, covering data, techniques and attitude
- Why are we interested?
  - Change of responsibilities
    - The arrival of the PRA
  - Change of opportunity
    - More data, increased computing power, technical advances
  - Change of circumstances
    - Lessons from the financial crisis
  - Change of philosophy
    - Inductive vs deductive reasoning

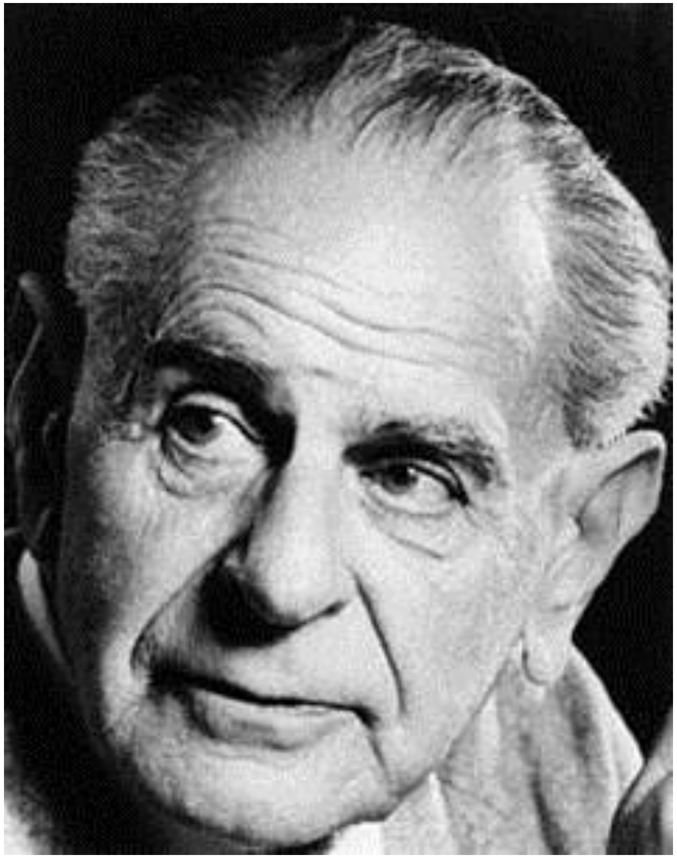
Why are we interested in Big Data?

- What do we mean by the term
  - Very loose meaning, covering data, techniques and attitude
- Why are we interested?
  - Change of responsibilities
    - The arrival of the PRA
  - Change of opportunity
    - More data, increased computing power, technical advances
  - Change of circumstances
    - Lessons from the financial crisis
  - Change of philosophy
    - Inductive vs deductive reasoning

Why are we interested in Big Data?

- What do we mean by the term
  - Very loose meaning, covering data, techniques and attitude
- Why are we interested?
  - Change of responsibilities
    - The arrival of the PRA
  - Change of opportunity
    - More data, increased computing power, technical advances
  - Change of circumstances
    - Lessons from the financial crisis
  - Change of philosophy
    - Inductive vs deductive reasoning

Is Correlation the New Causality?



Karl Popper

(Source: http://en.wikipedia.org/wiki/Karl_Popper)

Hal Varian

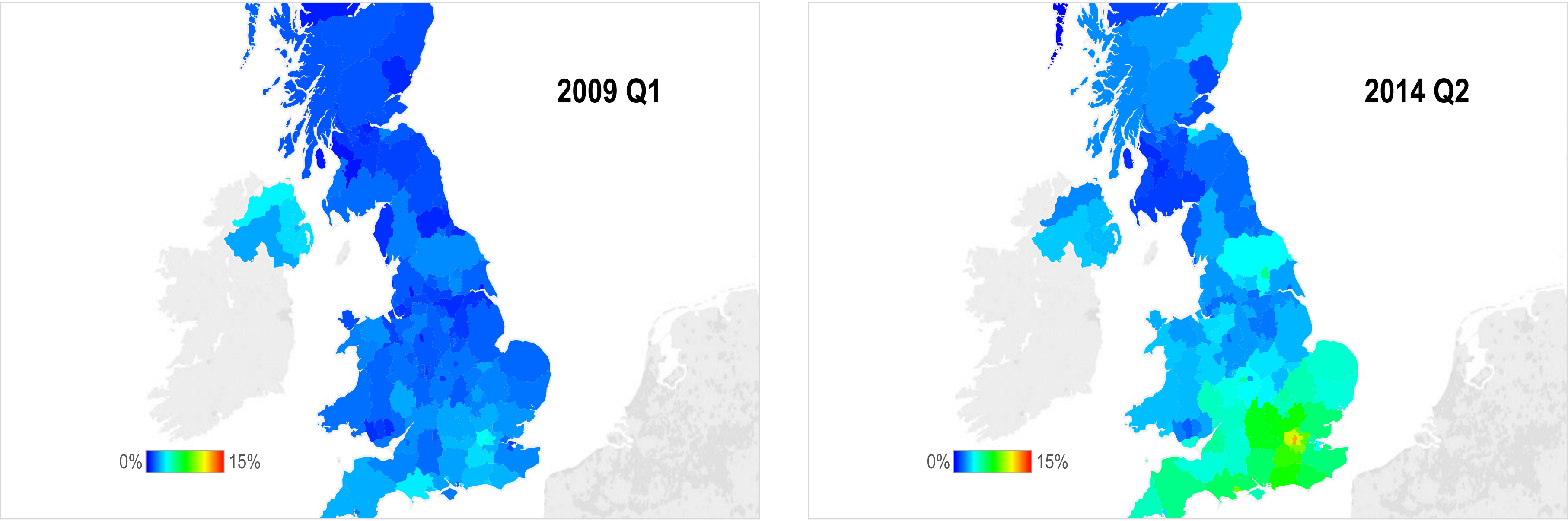
(Source: http://en.wikipedia.org/wiki/Hal_Varian)

What are we interested in?

- Gaining a richer understanding of the phenomenon of interest
  - Can help disentangle cause and effect…
  - …and identify the underlying issue that needs to be addressed
- Getting a speedier reading of developments in the economy and financial system
  - ‘Nowcasting’ and ‘nearcasting’
  - This might be particularly important when the system is undergoing rapid changes
- Quantifying previously purely qualitative data
  - Eg text

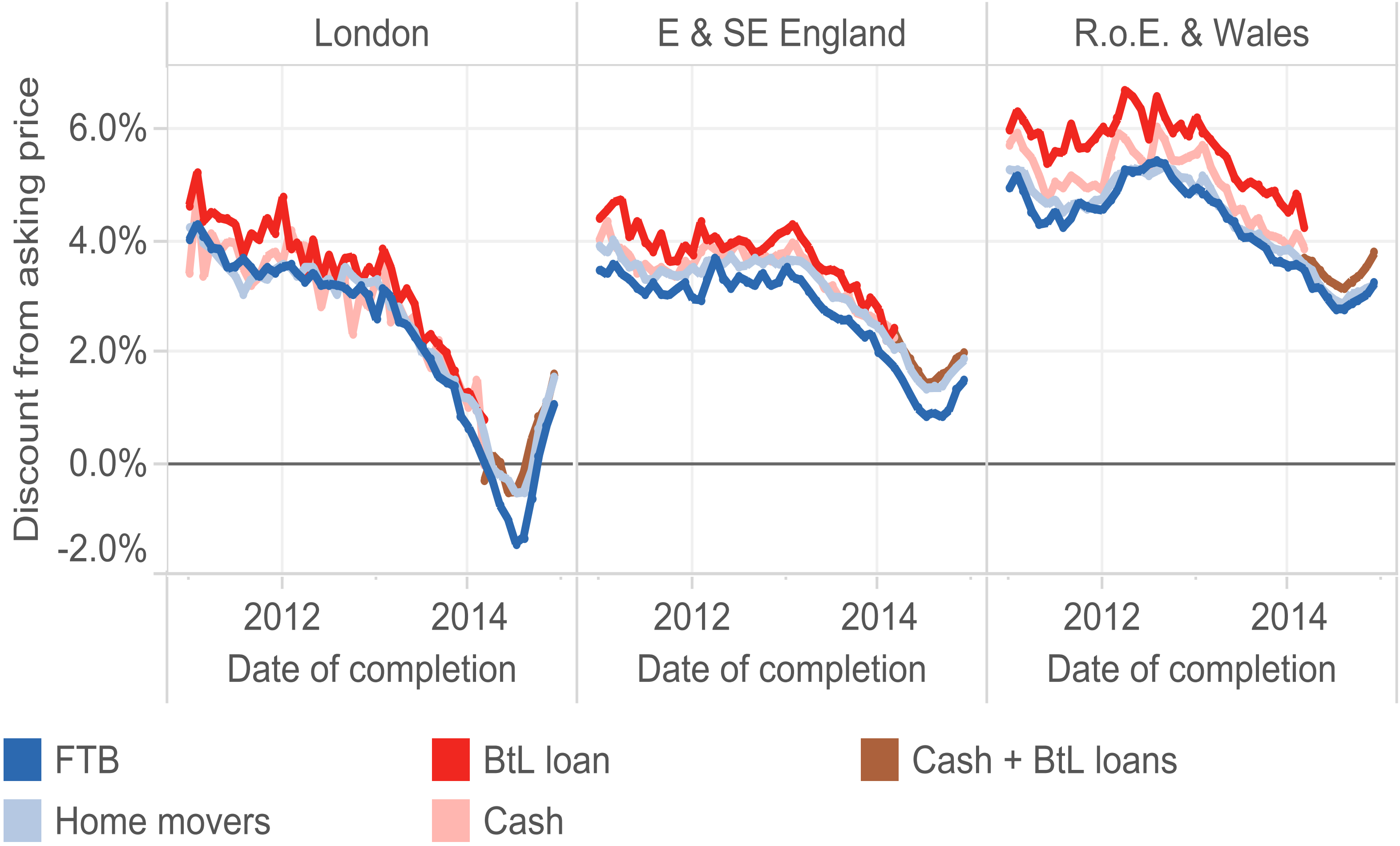
Matched micro data sets

Loan-to-income multiple ≥ 4.5



Source: Data are based on the Bank of England’s internal Product Sales Database collected by the FCA.

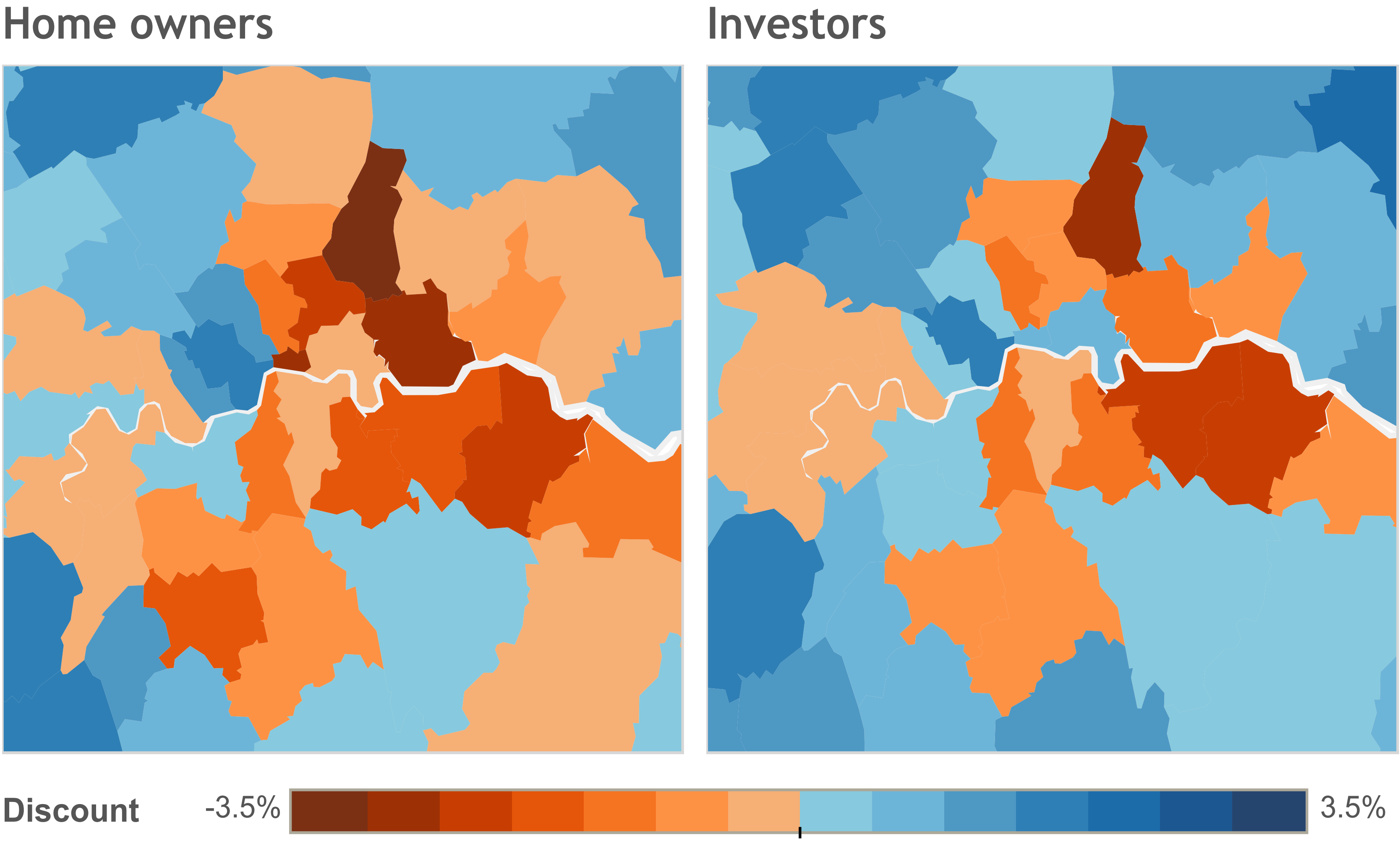
Discount from last asking price by buyer type



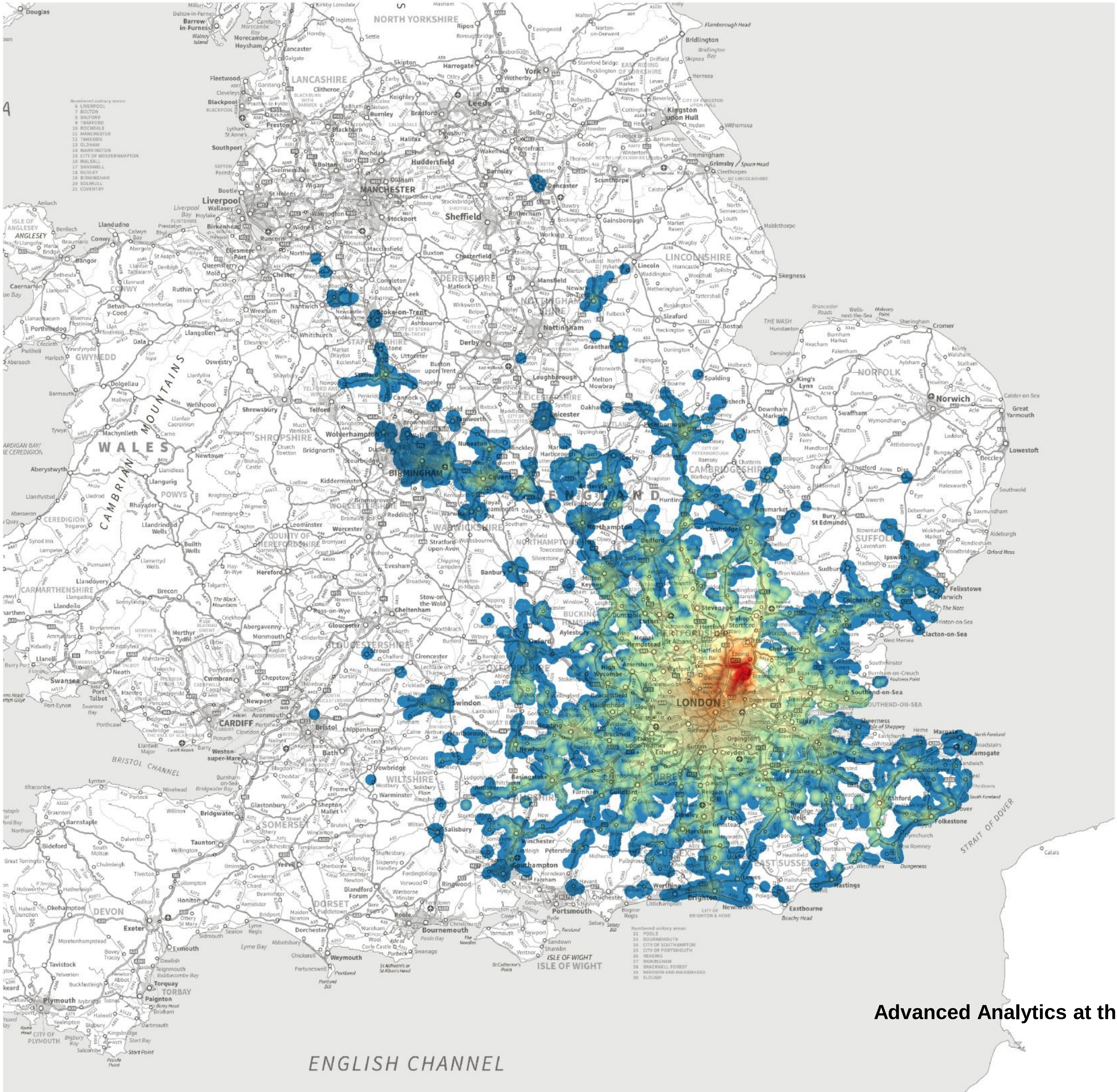
Sources: WhenFresh (Zoopla listings), Land Registry Price Paid, Land Registry Cash/Mortgage data, FCA Product Sales Data on mortgages, ONS Postcode Directory.

Advanced Analytics at the Bank of England

BANK OF ENGLAND



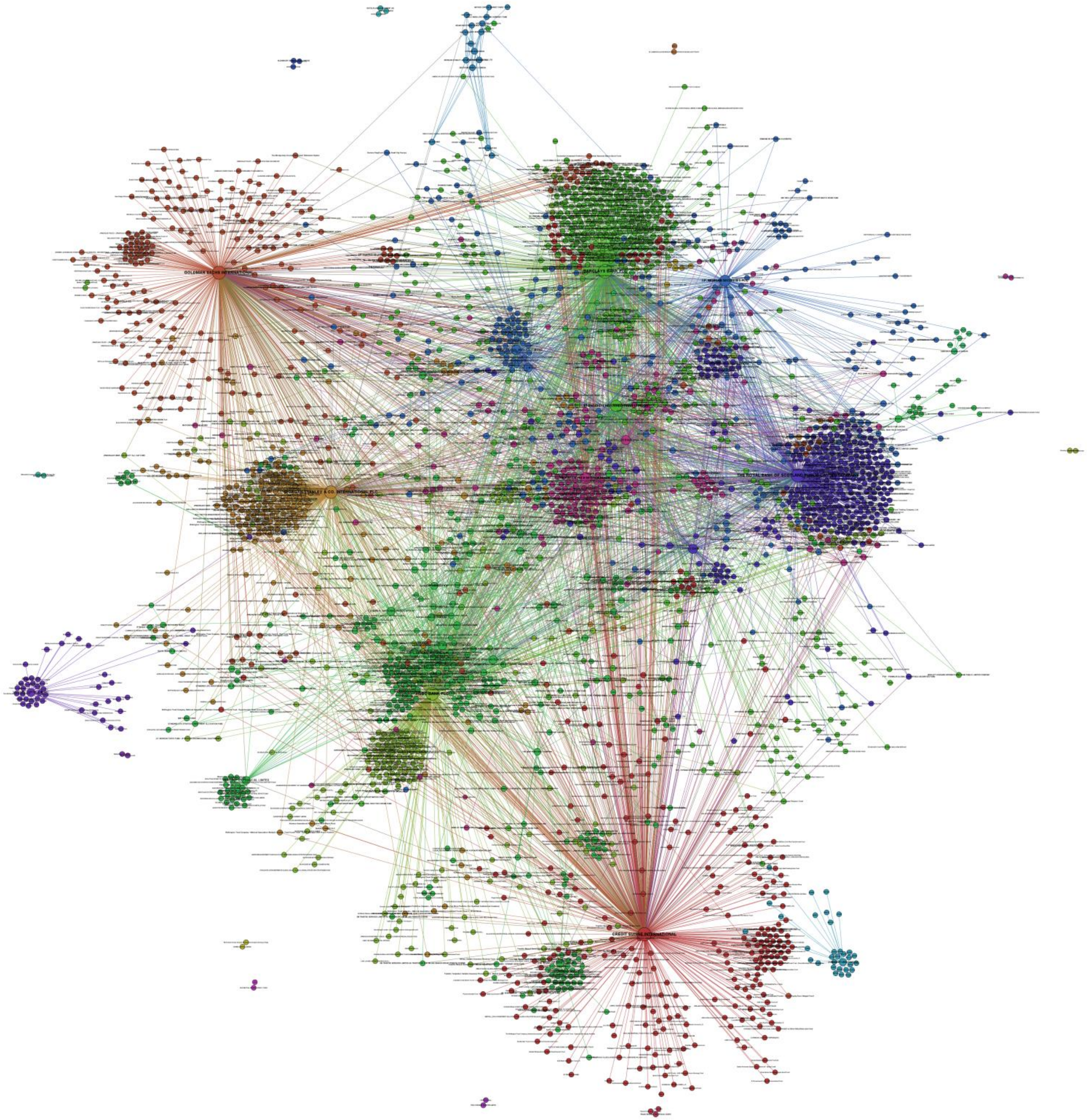
Sources: WhenFresh (Zoopla listings), Land Registry Price Paid, Land Registry Cash/Mortgage data, FCA Product Sales Data on mortgages, ONS Postcode Directory.


NORTH YORKSHIRE
LANCASHIRE
MANCHESTER
Liverpool
Sheffield
Leeds
York
Kingston upon Hull
LINCOLNSHIRE
DERBYSHIRE
NOTTINGHAMSHIRE
Nottingham
Birmingham
Norwich
NORFOLK
SUFFOLK
Cambridge
WALES
CAMBRIAN MOUNTAINS
ENGLAND
LONDON
Bristol
CARDIFF
BRISTOL CHANNEL
DEVON
Plymouth
Exeter
Southampton
Portsmouth
Brighton
ISLE OF WIGHT
STRAIT OF DOVER
ENGLISH CHANNEL

Highly granular data sets

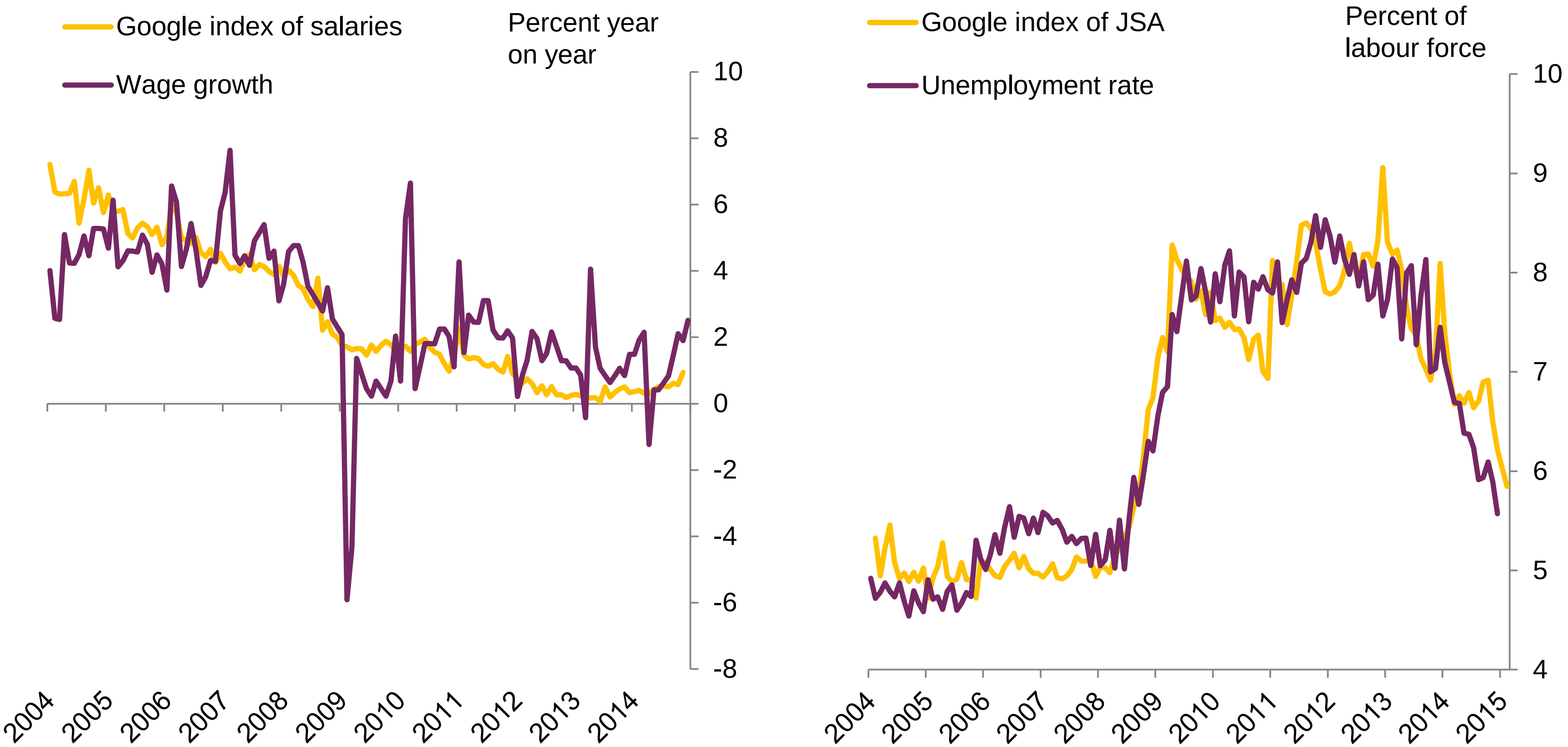
EMIR Data

Positions in outstanding CHF-denominated FX derivatives positions on 15/1/15

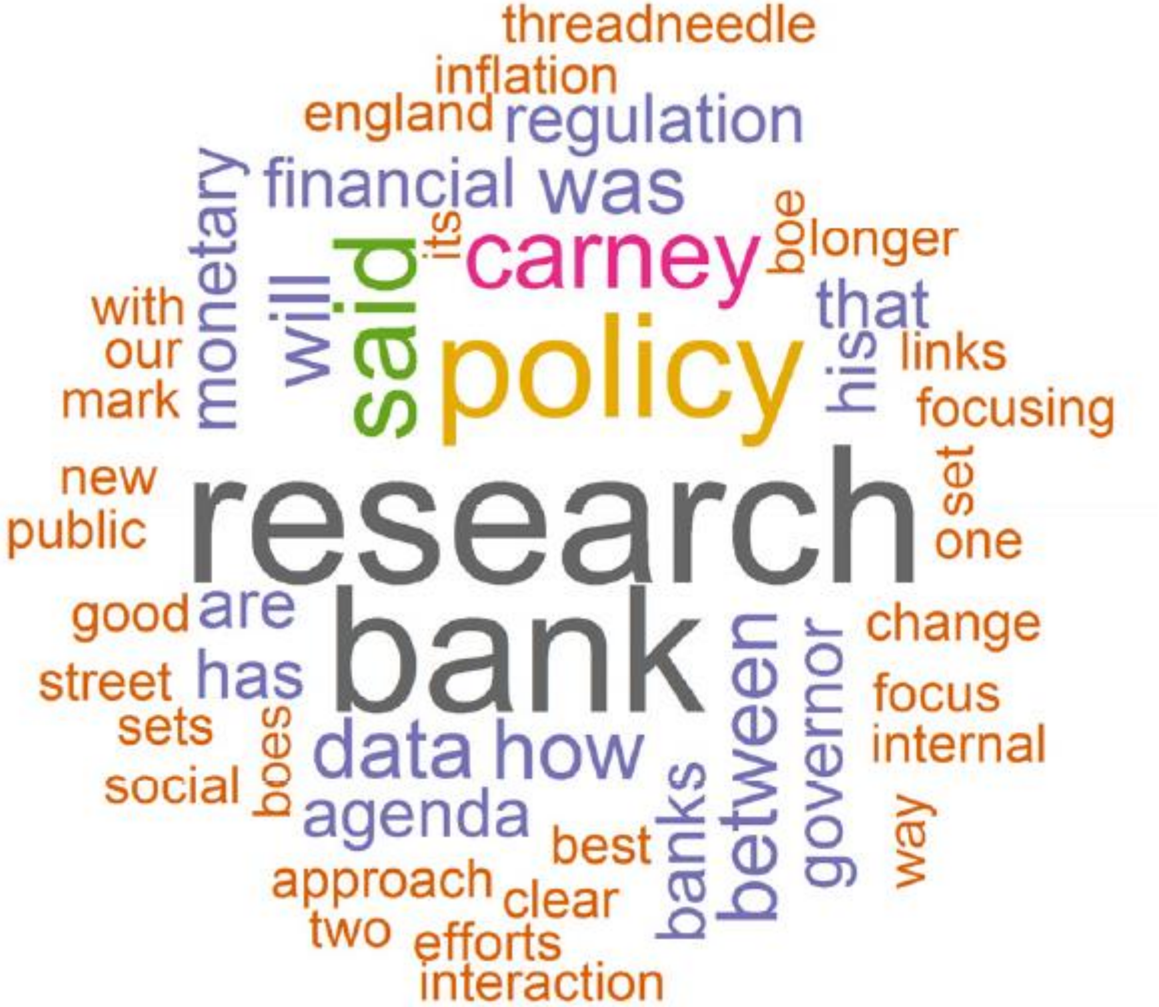

GOLDMAN SACHS INTERNATIONAL
BARCLAYS BANK PLC
J.P. MORGAN SECURITIES PLC
THE ROYAL BANK OF SCOTLAND PUBLIC LIMITED COMPANY
MORGAN STANLEY & CO. INTERNATIONAL PLC
HSBC BANK PLC
CREDIT SUISSE INTERNATIONAL

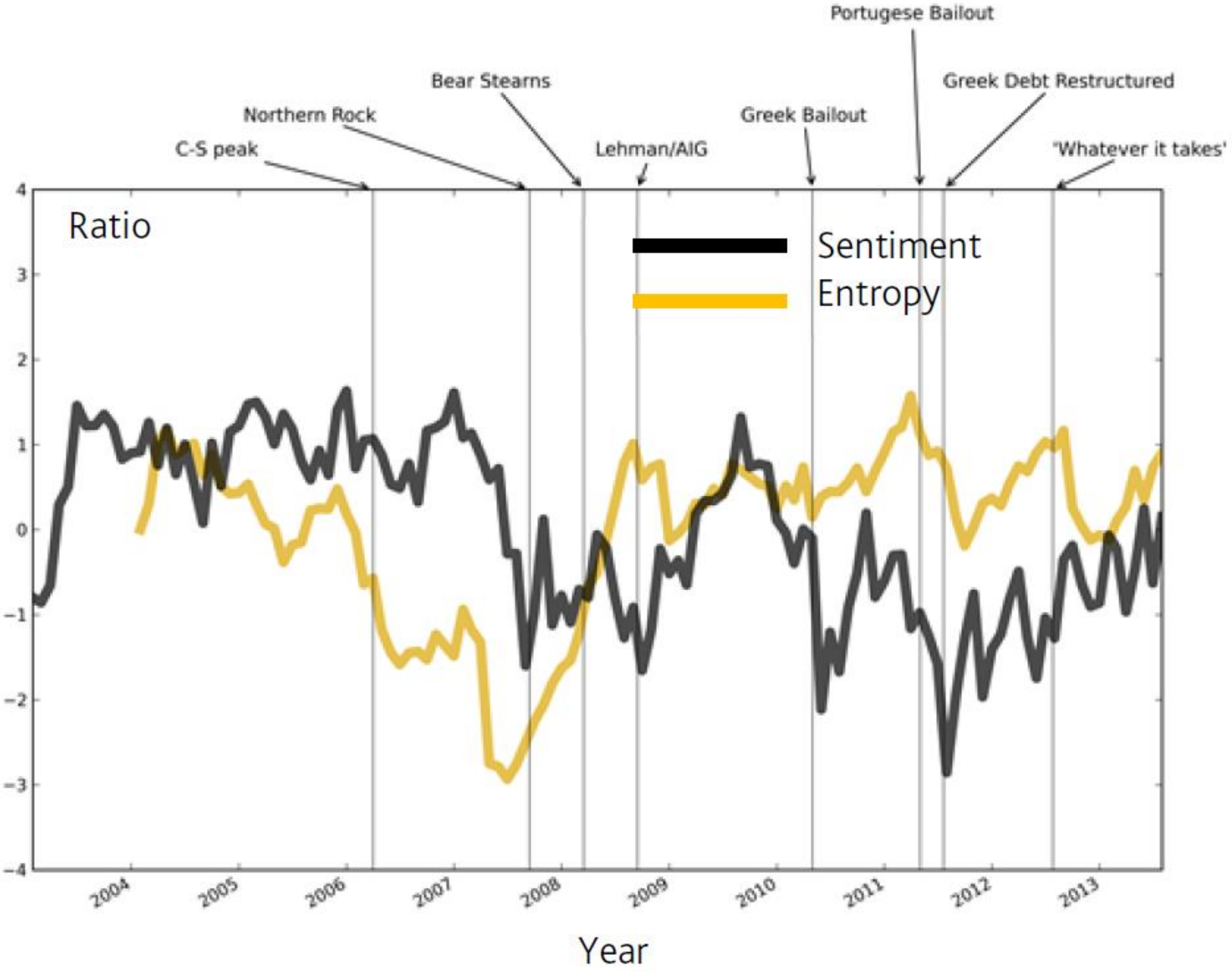
Text Analytics

Googling the Labour Market

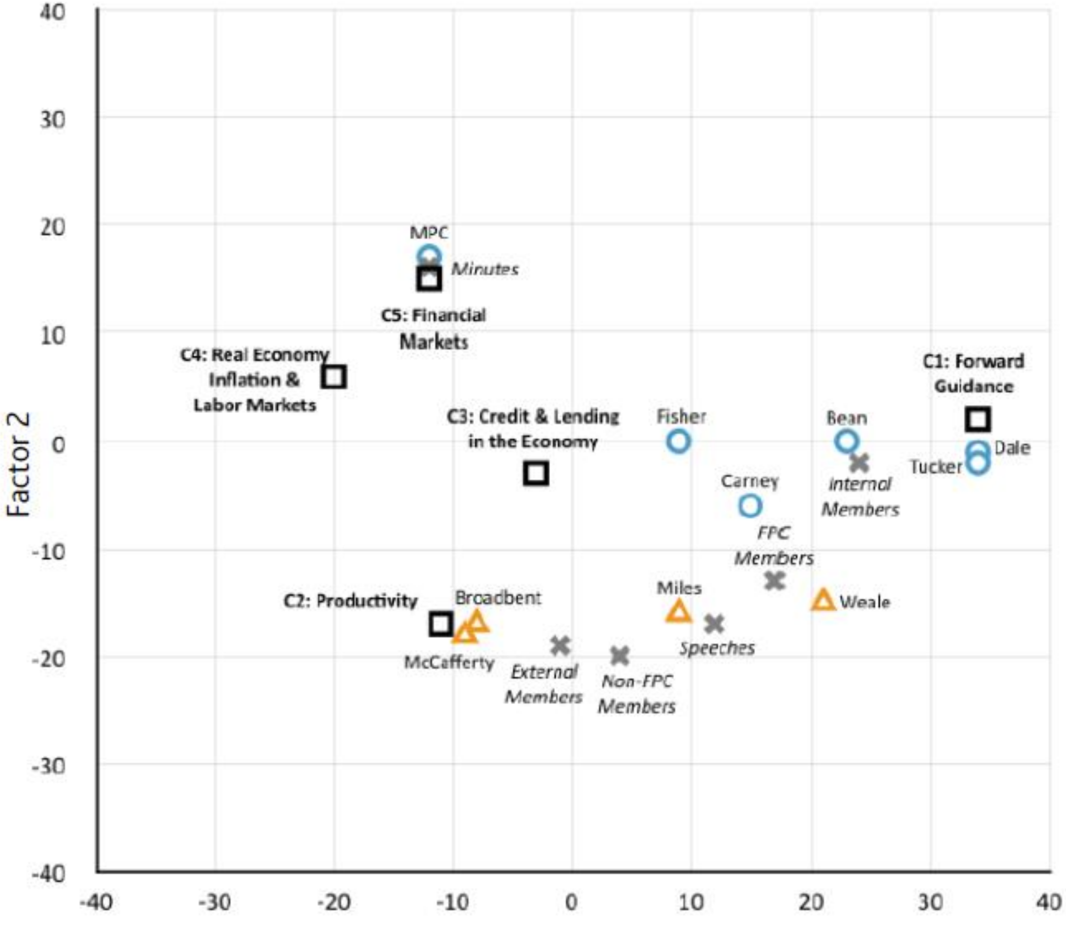


Source: ONS; Google. Notes: The Google indices are mean and variance adjusted to put on the same scale as the unemployment rate and wage growth. The Google indices are drawn from searches containing the terms “salaries” and “job seekers allowance”. See Mclaren and Shanbhogue (2011) for further details.


threadneedle
inflation
england
regulation
monetary
financial
was
its
boe
longer
carney
said
will
with
that
our
policy
his
links
mark
focusing
set
new
research
public
one
good
are
change
bank
street
has
between
governor
focus
sets
internal
data
how
boes
social
agenda
banks
way
best
approach
clear
two
efforts
interaction


Portugese Bailout
Greek Debt Restructured
Bear Stearns
Northern Rock
Greek Bailout
C-S peak
Lehman/AIG
'Whatever it takes'
Ratio
Sentiment
Entropy
4
3
2
1
0
−1
−2
−3
−4
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
Year

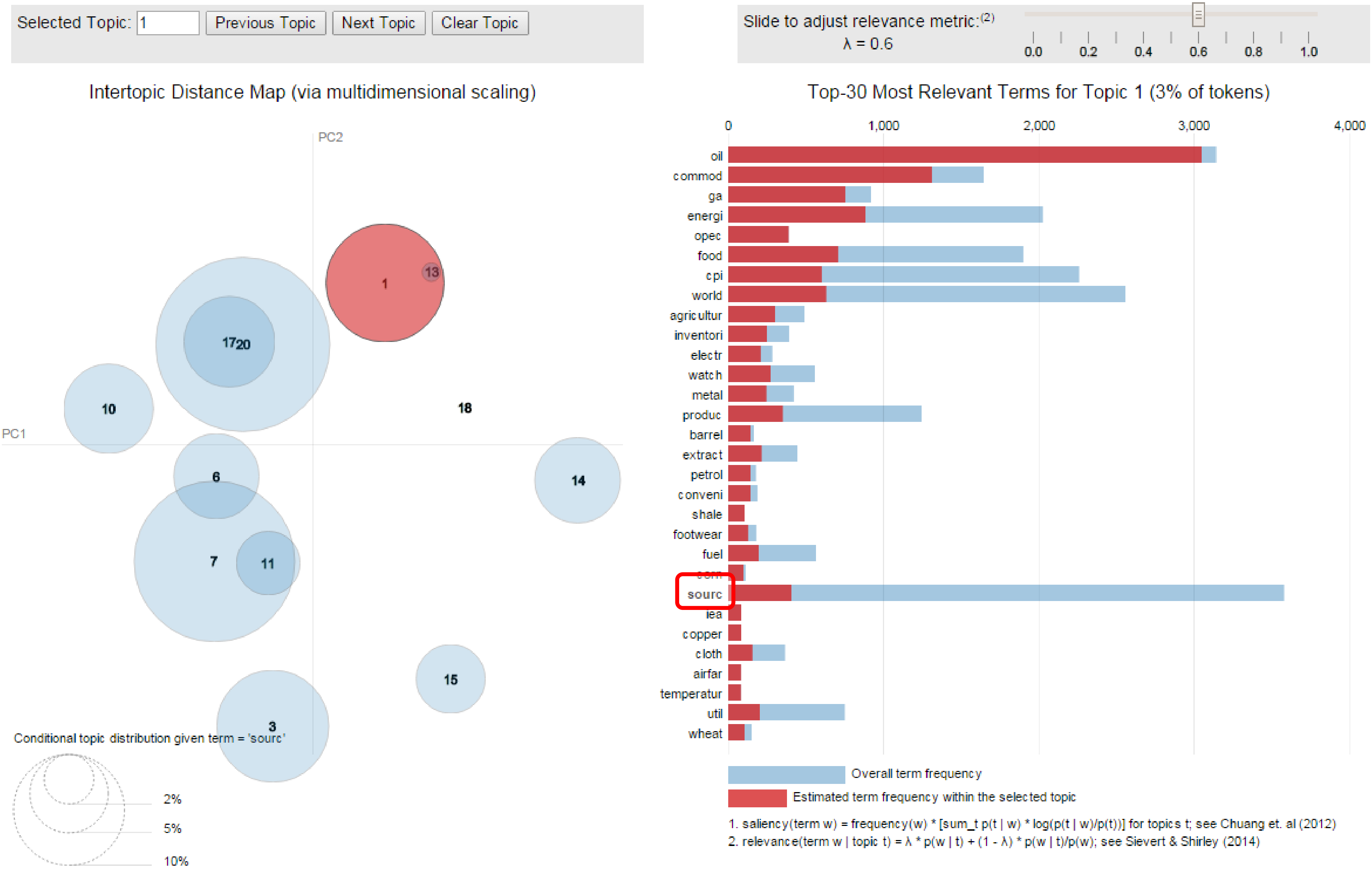
Correspondence Analysis of Themes & Passive Variables
MPC Minutes & Speeches, Carney Governorship (Jul 2013-Jun 2014)



□ Class
× Attribute
○ Internal Member
△ External Member

Factor 1	Association	Cumulative
Factor 1	33.5%	33.5%
Factor 2	26.3%	59.8%

Class 1 (27%) — Forward Guidance
Class 2 (27%) — Productivity
Class 3 (10%) — Credit & Lending in the Economy
Class 4 (24%) — Real Economy, Inflation & Labor Markets
Class 5 (12%) — Financial Markets

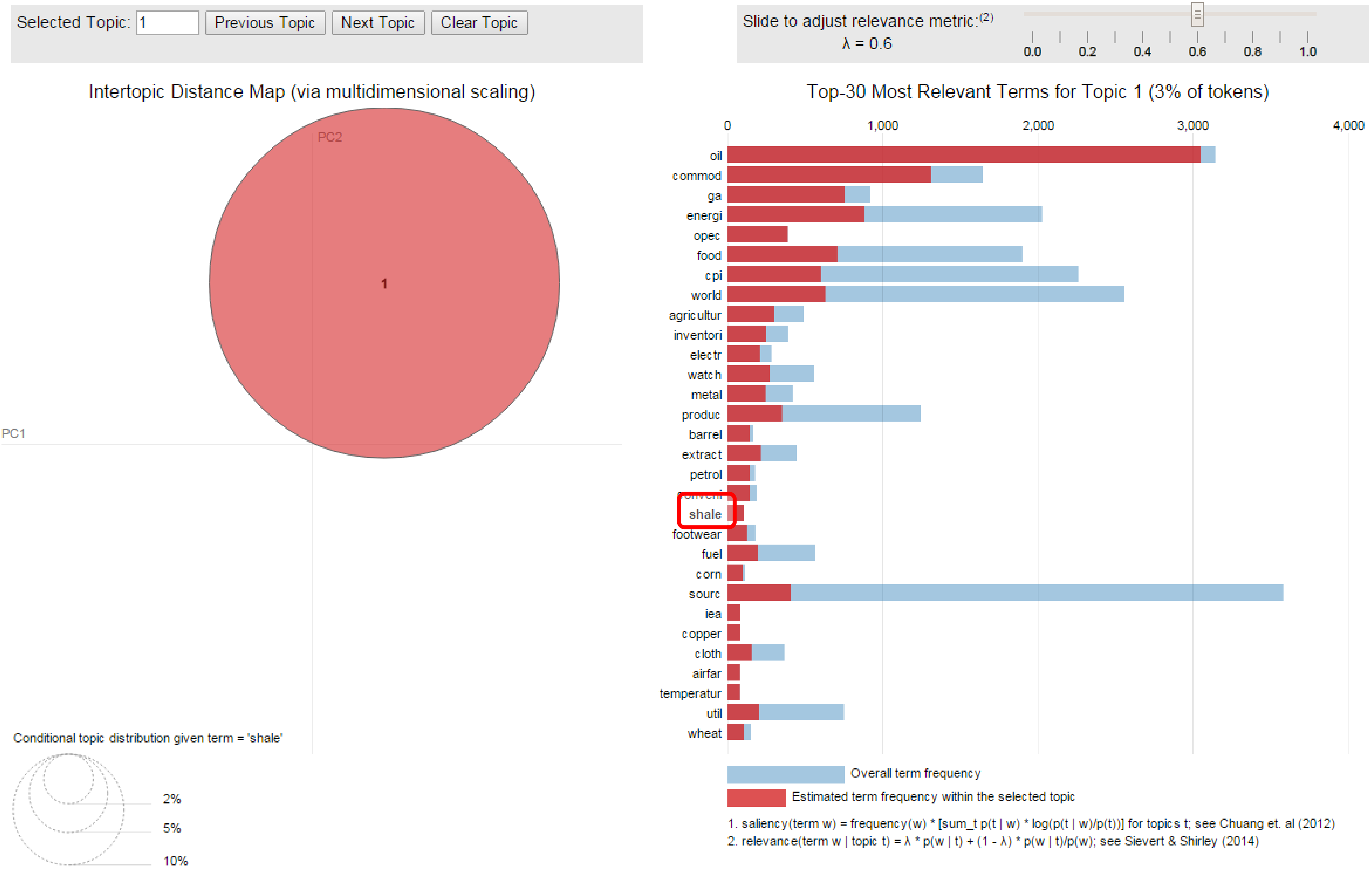


Selected Topic: 1

Slide to adjust relevance metric:(2)
λ = 0.6

Conditional topic distribution given term = 'sourc'

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)



Selected Topic: 1

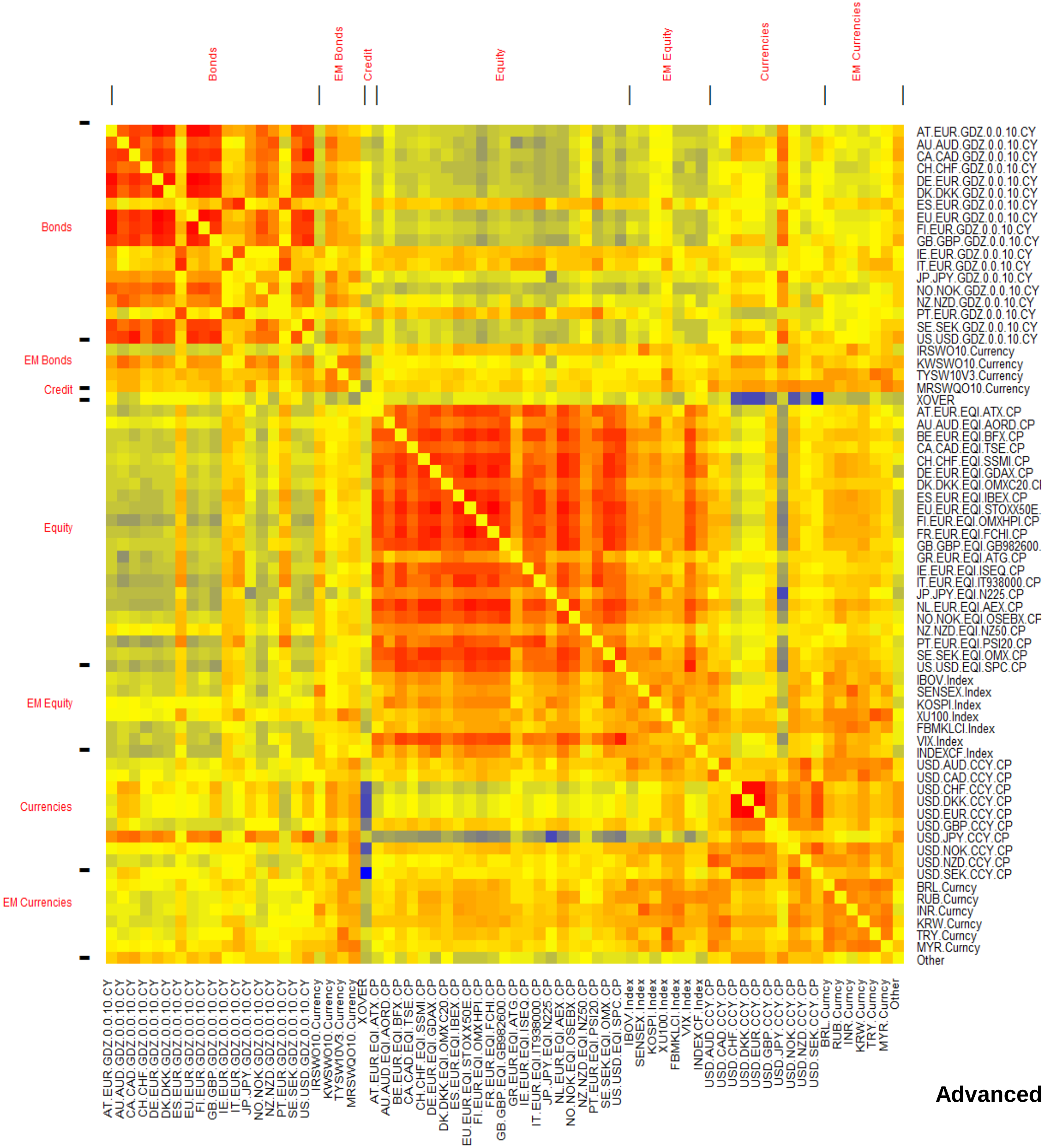
λ = 0.6

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

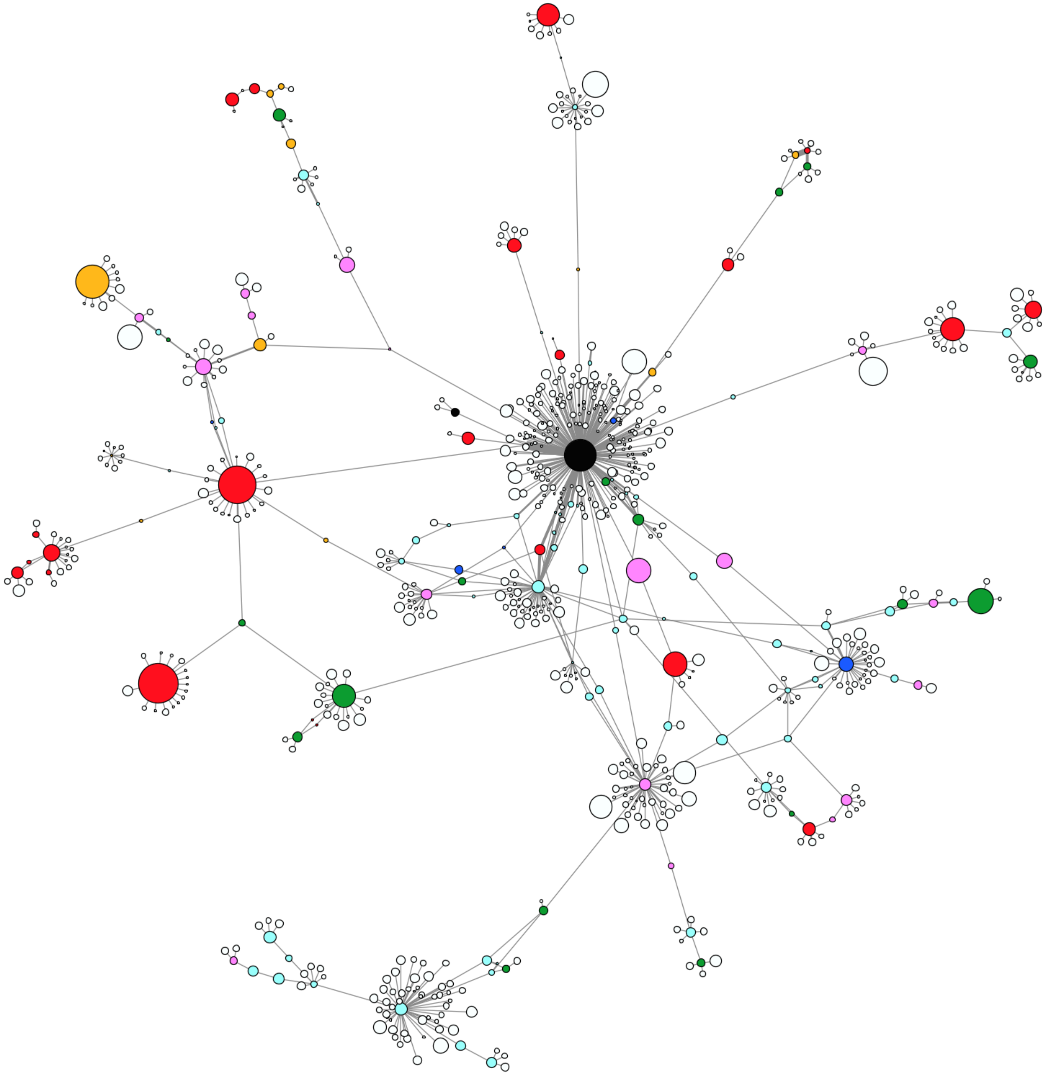
Visualisation

Asset class correlation heatmap



Advanced Analytics at the Bank of England

BoE communication

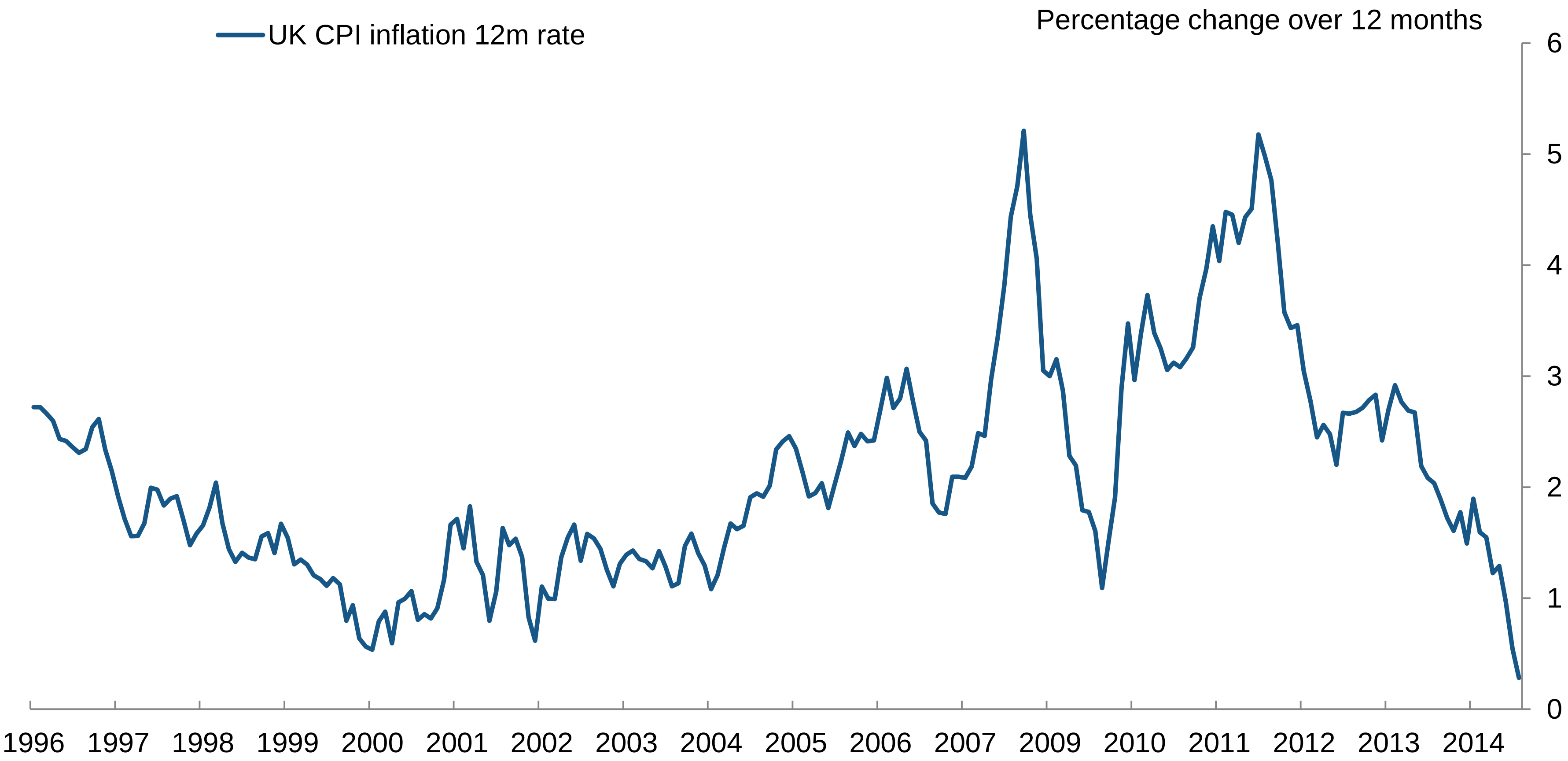


Machine learning

Issues with analysing 'Big Data'

- Example: CPI micro-data
- The ONS has produced a data set comprising:
  - 215 months (Feb 1996-Dec 2013)
  - ~110,000 prices collected per month (not the same number each month)
  - 1,113 items (not the same items each year)
  - 71 COICOP classes
  - various other meta-data (eg type of shop, region etc)
  - in total: 24,442,988 records with 25 fields
  - 611,074,700 pieces of data

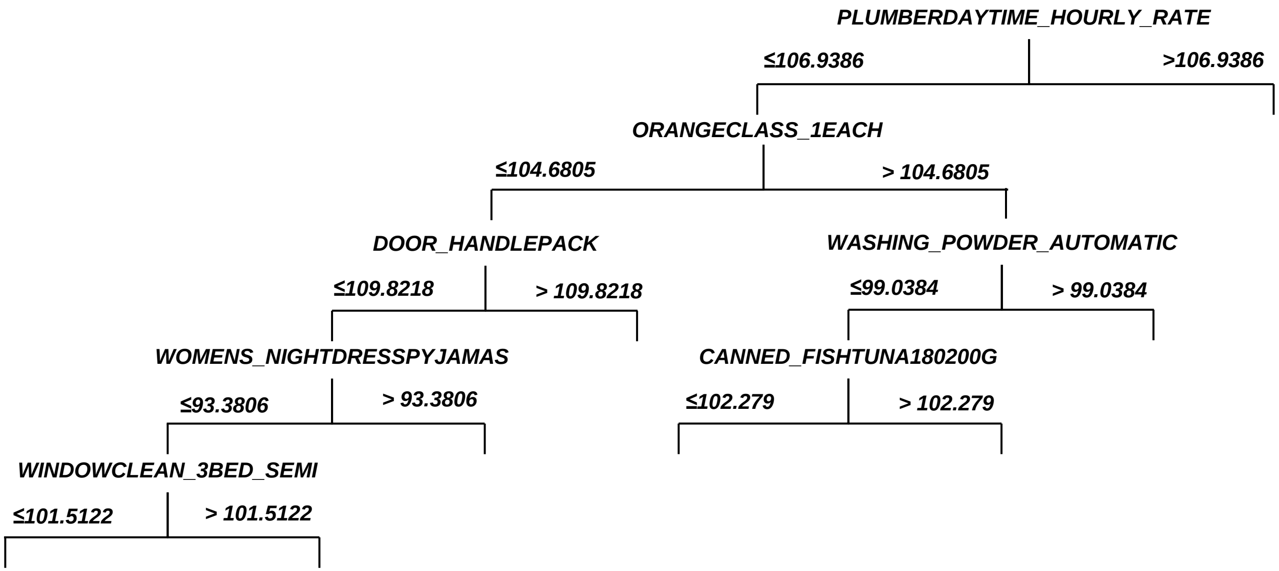
Issue 1: the stability of annual inflation



Issue 2: over-fitting: what to do?

- *T* is normally >> *N*; here *N* >>*T* (indeed this is occasionally used as a definition of ‘big data’ by statistically-minded analysts)
- Aggregate the data? – eg by type of good
  - But that tends to obviate the point of using the micro data
- Shrink the dimensionality of the matrix of explanatory variables:
  - PCA/factor models extract combinations of the variables that explain most of the variance of the dependent variable
  - So you can keep much of the information that is contained in the data set without getting into large statistical problems
  - A key issue here though is how to interpret the resulting model, and the components/factors may be unstable over time
- Penalised regressions

Issue 3: explaining non-linear functions

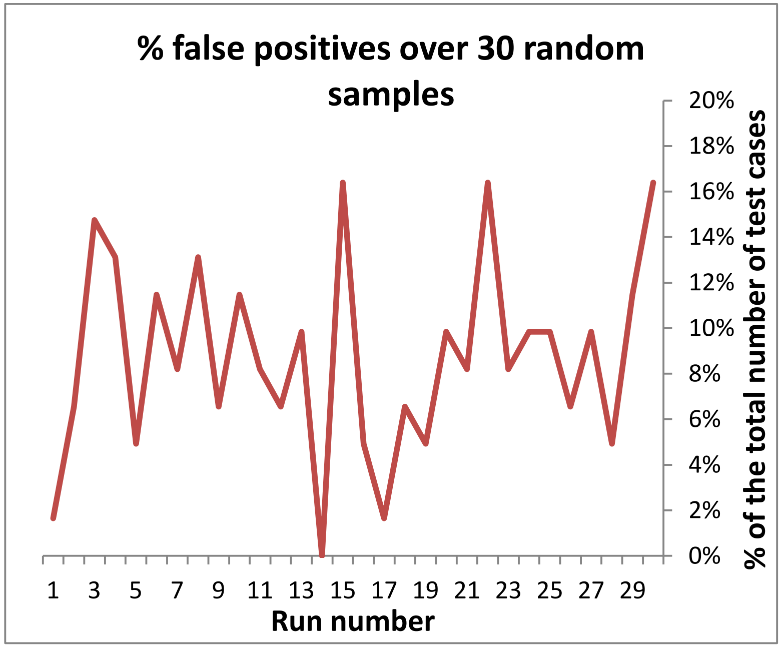


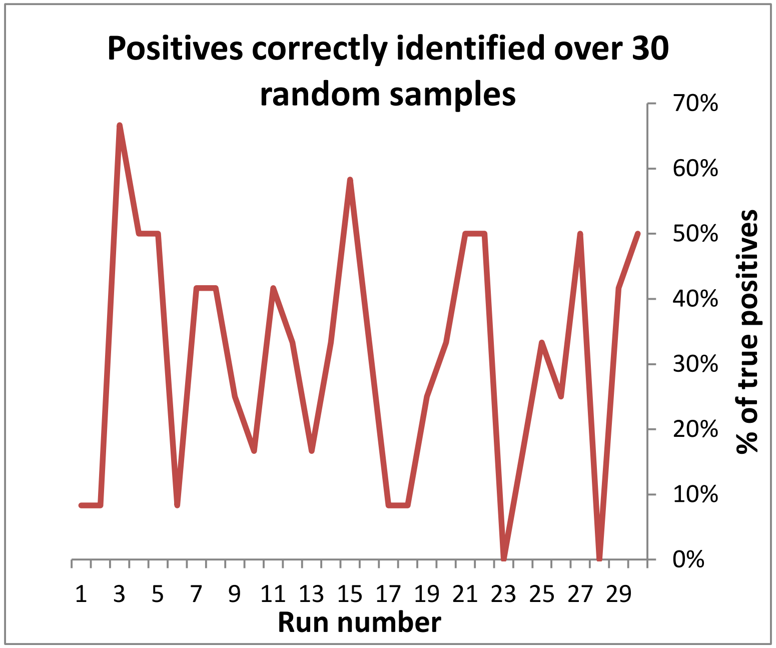
- Try explaining the intuition behind this relationship to busy policy makers…

BANK OF ENGLAND

Issue 4: Stability

- An issue that is closely linked to over-fitting is the stability of the models
- This is a particularly important issues when there is no strong *a priori* reason to think that the world works in this way
- (Though *a priori* thinking can also be misleading at times)





Issue 5: Confidentiality / 'Big Brother' state

- This was not relevant to the CPI work
- In general, the more detailed and granular the data set is, the more likely it is to contain confidential information
- We must ensure that:
  - we only use data for appropriate reasons
  - the minimum number of people are able to see any confidential data given the needs of the situation
  - data are stored securely and professionally

Issue 6: Practical issues

- Hiring!
- IT!

Conclusion

- Do these issues mean that ‘Big Data’ is likely to be a passing fad?
  - No!
- The data exist and the BoE has the responsibility and opportunity to use them to help us understand economic developments and the structure of the economy and financial system
- But the issues do mean that this is no panacea
- Just as with any other empirical work the data need to be cleaned and understood (both of which are more difficult with larger data sets) prior to analysis
- And then analysed carefully using appropriate methods