Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

IFC – Bank Indonesia International Workshop and Seminar on *"Big Data for Central Bank Policies / Building Pathways for Policy Making with Big Data"*

Bali, Indonesia, 23-26 July 2018

Understanding big data: fundamental concepts and framework[1]

Paul Robinson,
Bank of England

[1] This presentation was prepared for the meeting. The views expressed are those of the author and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

BANK OF ENGLAND

Understanding Big Data: Fundamental Concepts and Framework

International Workshop on Big Data for Central Bank Policies

Paul Robinson, Bank of England

23 July 2018

Outline

- What do we mean by 'Big Data'?
- Several different dimensions that we can classify its use:
  - Different types of data
  - Different uses of the data sets
  - Different analytical techniques
- Are central bank needs' different from other organisations?
- Lots of opportunities but also challenges

What do we mean by ‘Big Data’?

- First page of a Google search for “V’s of big data” included:
  - Infographic: The Four V's of Big Data | IBM Big Data & Analytics Hub
  - The 10 Vs of Big Data | Transforming Data with Intelligence
  - Understanding the 3 Vs of Big Data - Volume, Velocity and Variety
  - The 42 V's of Big Data and Data Science - Elder Research
  - The five V's of big data | BBVA
  - How many V's are in big data?

What do we mean by ‘Big Data’?

- First page of a Google search for “V’s of big data” included:
  - Infographic: The Four V's of Big Data | IBM Big Data & Analytics Hub
  - The 10 Vs of Big Data | Transforming Data with Intelligence
  - Understanding the 3 Vs of Big Data - Volume, Velocity and Variety
  - The 42 V's of Big Data and Data Science - Elder Research
  - The five V's of big data | BBVA
  - How many V's are in big data?

Different types of data

- Despite the confusion and hype the ‘V’s structure does offer a framework to consider the opportunities and challenges
- In particular, the following 5 ‘V’s set up is useful:
  - Volume
  - Velocity
  - Variety
  - Value
  - Veracity

What central banks do

- Regulate important institutions
  - Banks, insurance companies, FMIs, …
- Set policy
  - Monetary policy, macroprudential policy, microprudential policy
  - Engage in international policy setting
- Implement policy
  - Markets, PRA, …
- Run important functions
  - Payment systems, currency issuance …
  - Manage national reserves
  - Act as a bank to key institutions (eg the government)
- Run a large, (singular) institution

- Most central banks have similar responsibilities

How do central banks go about discharging these responsibilities?

- Understand the current situation
  - Combine information with an understanding of how it fits together
- Forecast what would happen holding policy unchanged
- Consider possible policy changes
- Model how they would affect the economy/financial system, …
- Set policy
- Monitor the effects of policy
  - Update our understanding of the current situation and the structure of the system

Why it's difficult

- Imperfect measurement
  - Noise, biases, blind spots, out of date information, (near) simultaneity of cause and effect
- "Too much" data, too little information
- Imperfect theory
- Complex, adaptive system with lots of feedback
  - Leads to "chaotic" behaviour
- Internal frictions

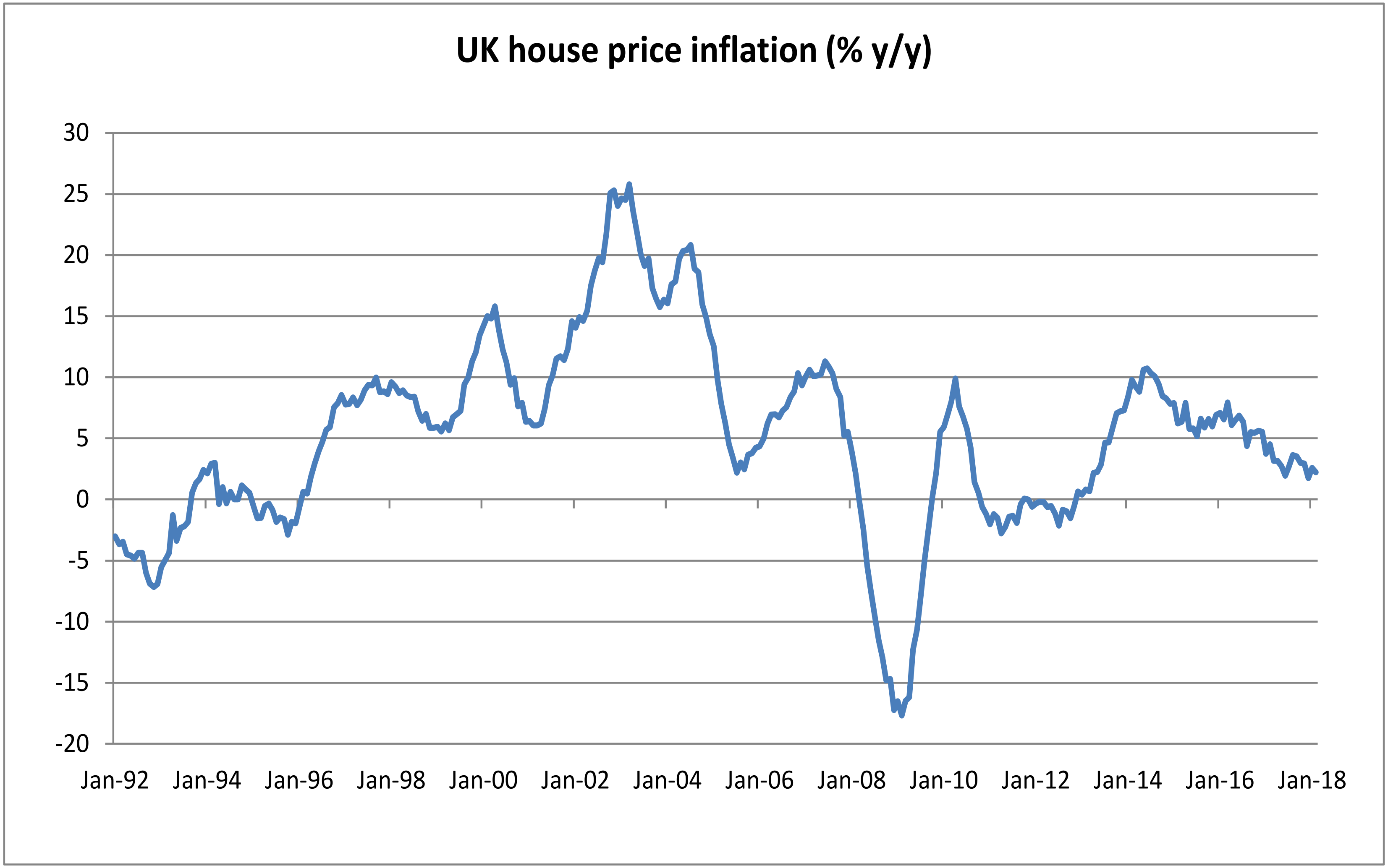
How can Big Data help?

- Imperfect measurement
  - Insight into previously hidden phenomena
  - Combining different types of data
  - Speed and completeness of coverage
- “Too much” data, too little information. Use data science methods to:
  - Improve processing large data sets
  - Help separate the signal from the noise
- Imperfect theory
  - Hypothesis generation
  - Alternative modelling approaches (eg Agent-based models)
- Complex, adaptive system with lots of feedback
  - Difficult to cope with, but more accurate understanding of initial conditions and more frequent updating help a lot
- Internal frictions
  - Improved management information

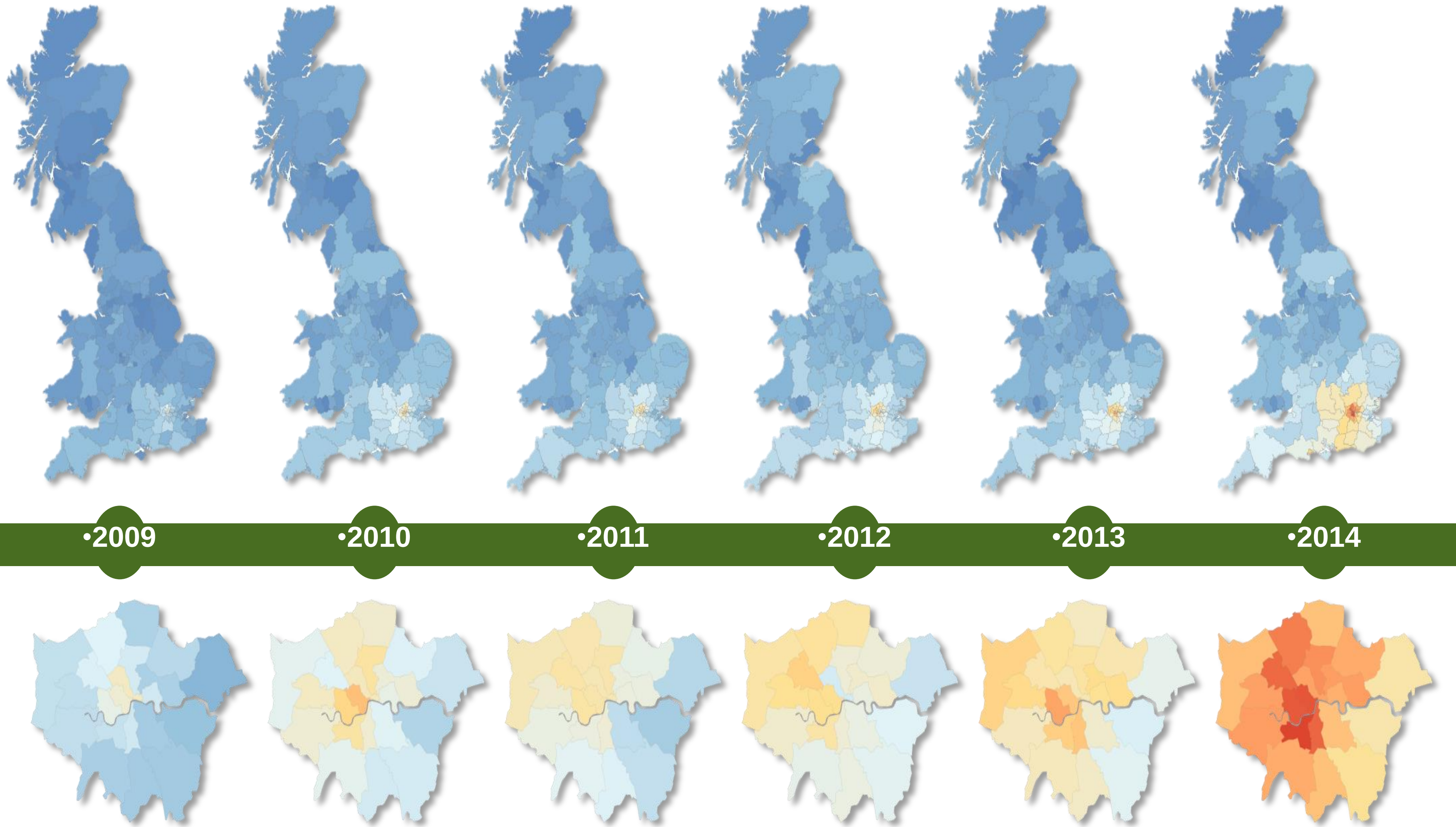
Big data sets offer significant potential advantages

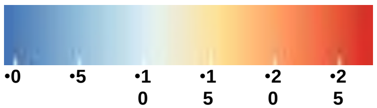
- Greater **detail** (Volume, Velocity, Variety)
- Allow insights that aggregate numbers might obscure
- Examples:
  – UK housing market
  – Market dynamics around the abolition of the EUR/CHF floor
  – Market liquidity around large market moves

UK housing market



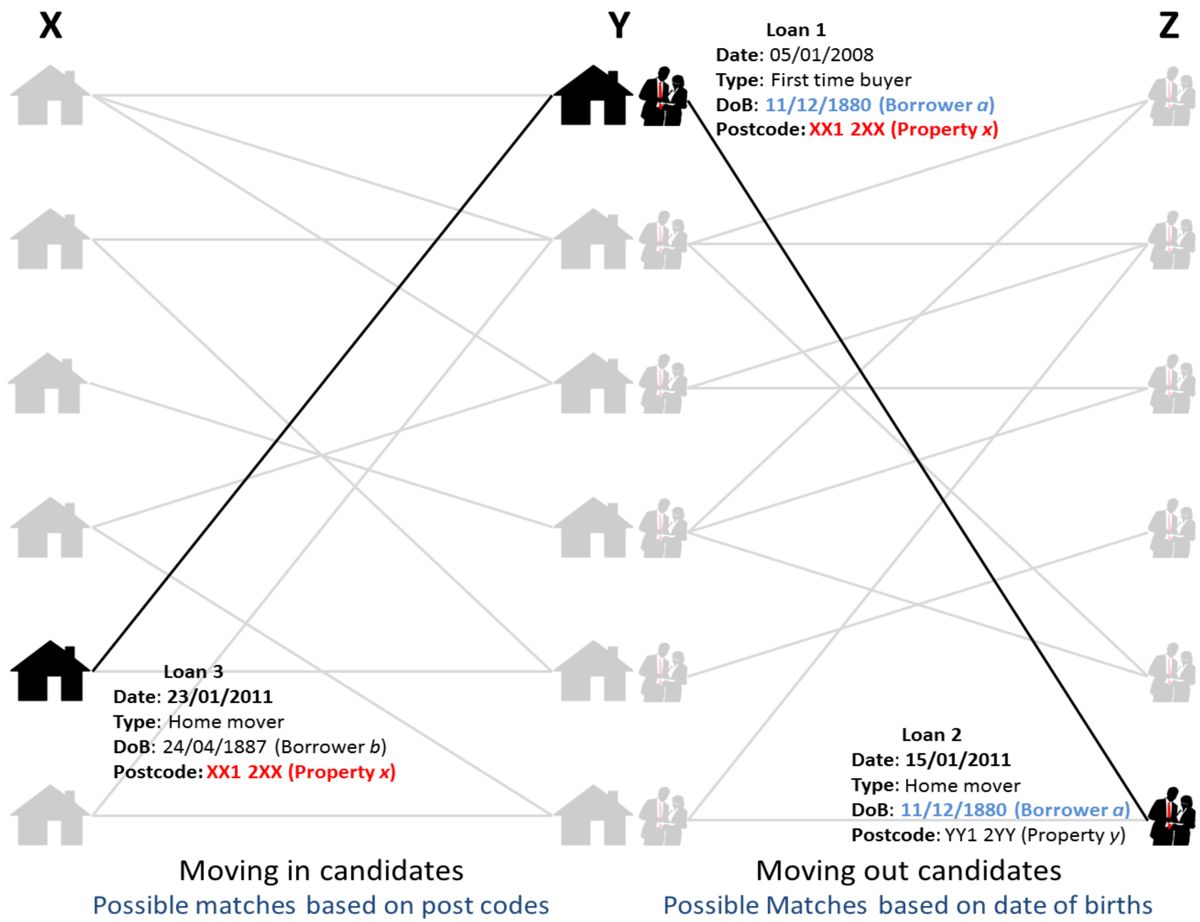
Source: Average of HBOS and Nationwide measures





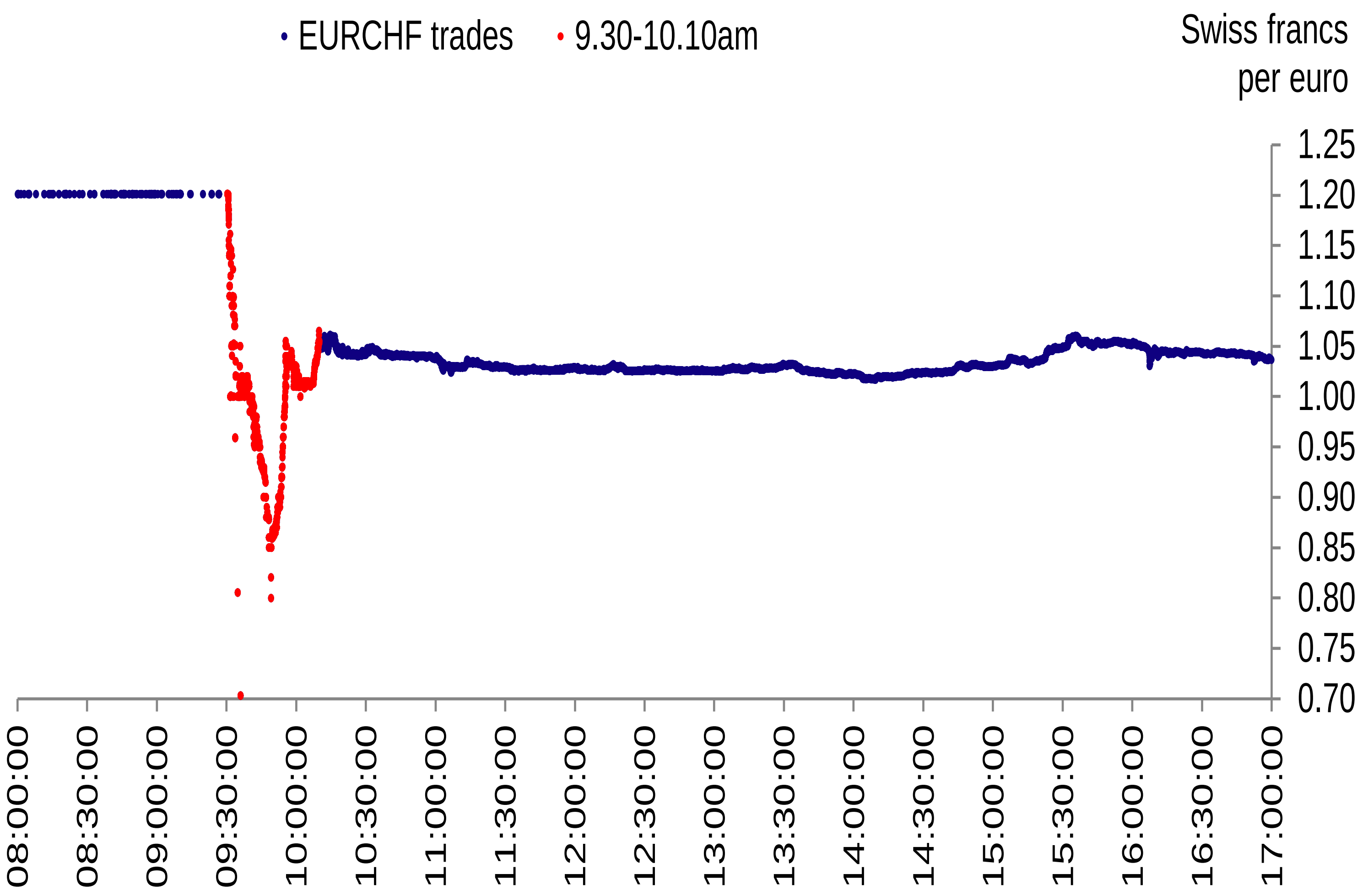
•Key: % of mortgages with loan more than 4.5x income

Tracking home movers

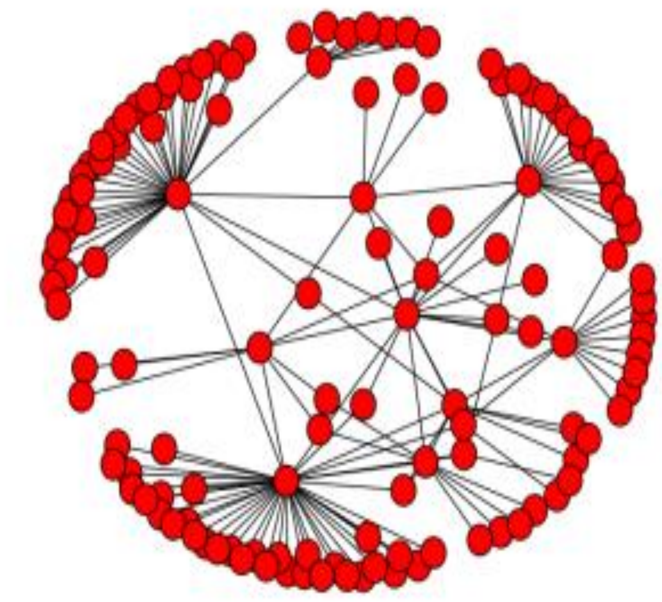


BANK OF ENGLAND

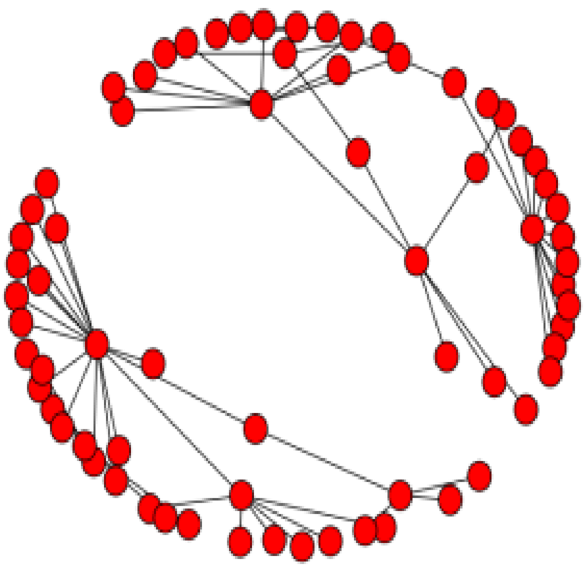
Large-scale data analysis



Network of CHF derivatives contracts

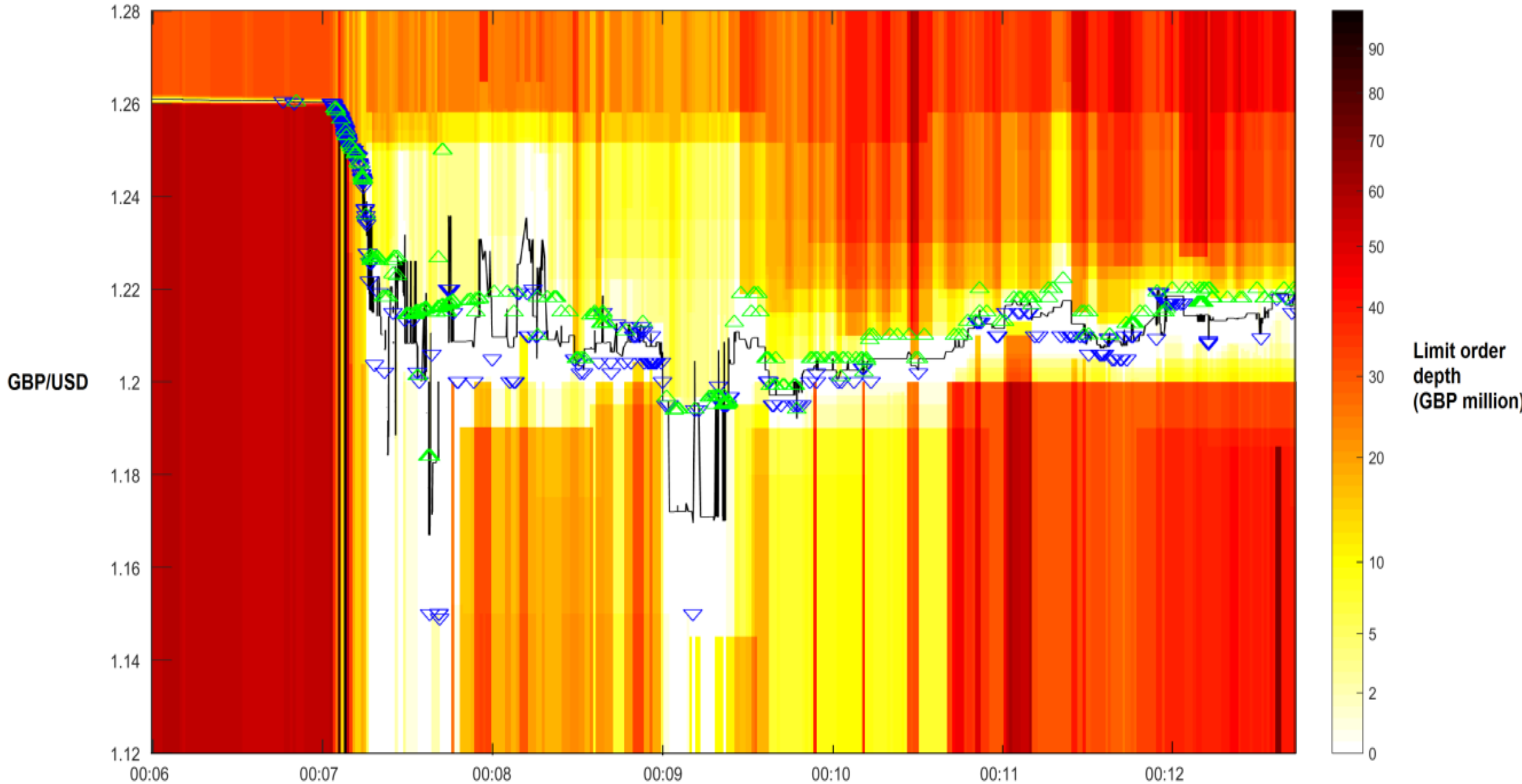


15 January 2015



22 January 2015

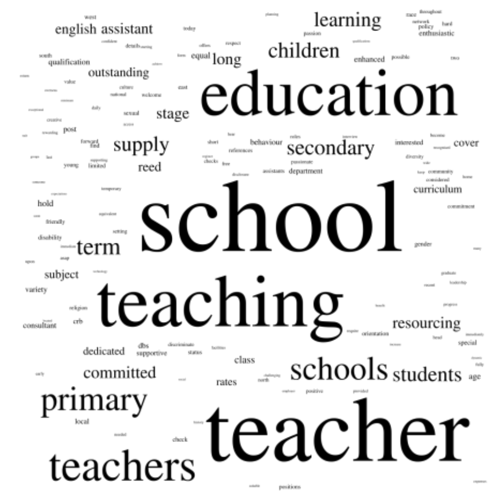
Market depth around sterling "flash crash" episode (7 Oct 2016)

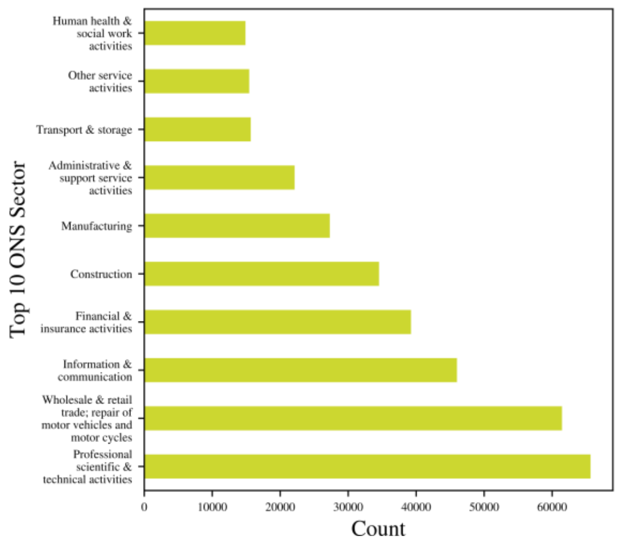


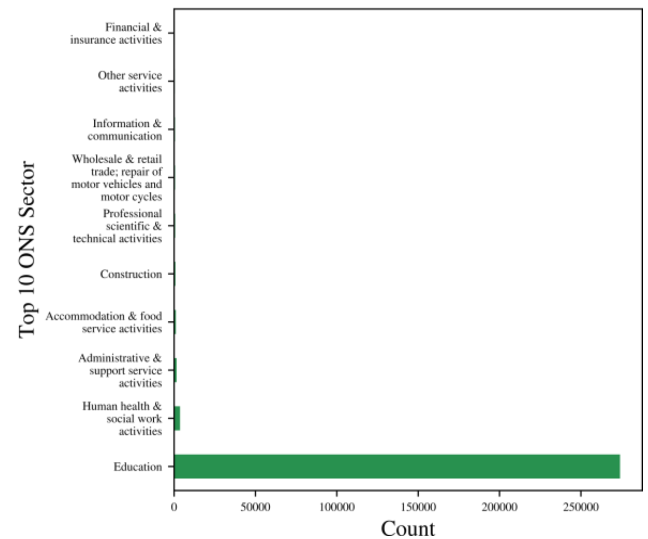
Big data sets offer significant potential advantages

- Greater **flexibility** (Velocity, Variety)
- Gives a window into changing structure of the economy
- Example:
  - Using job adverts to understand changing labour market dynamics

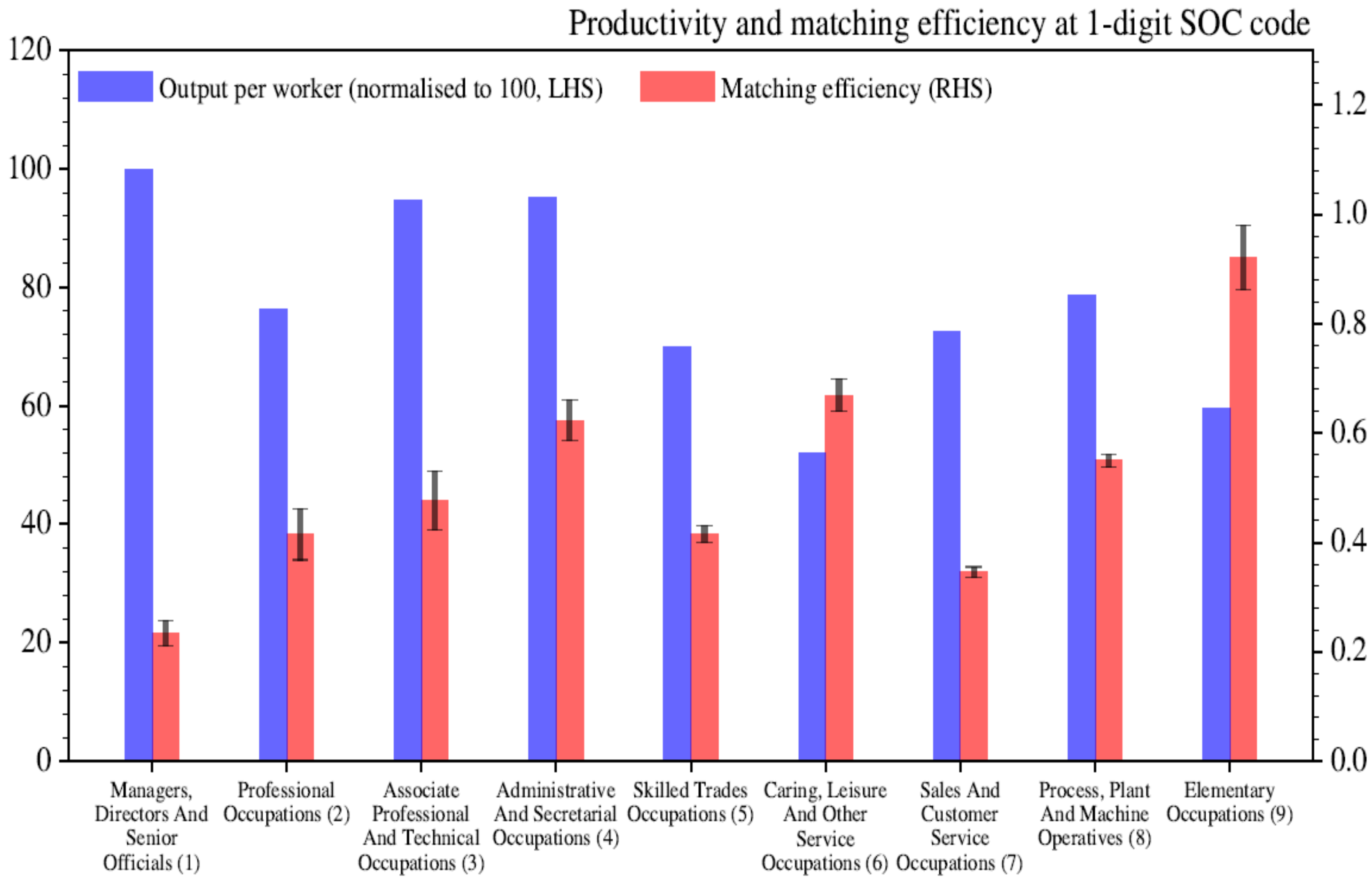
Understanding the labour market using job ads







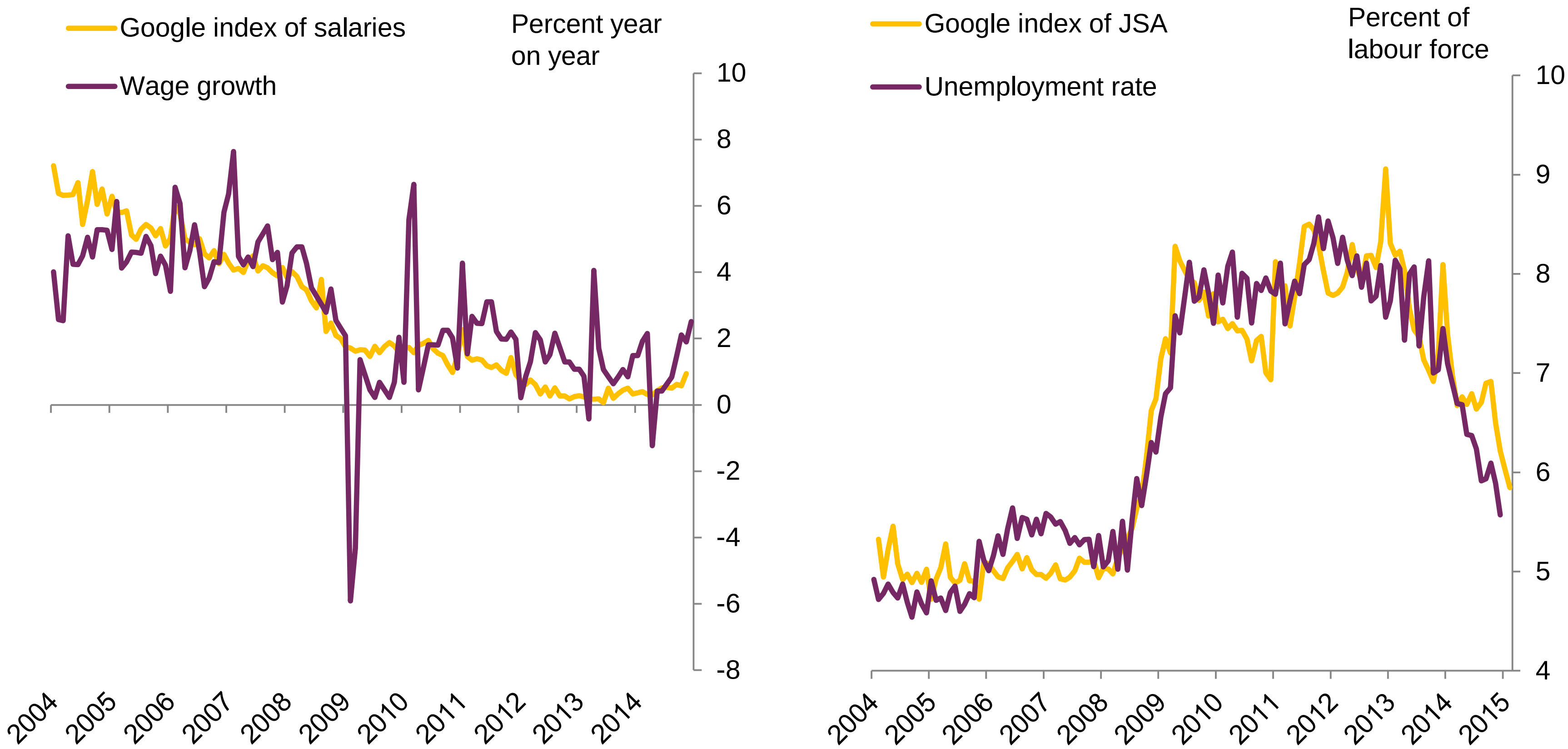
Understanding the labour market using job ads



Big data sets offer significant potential advantages

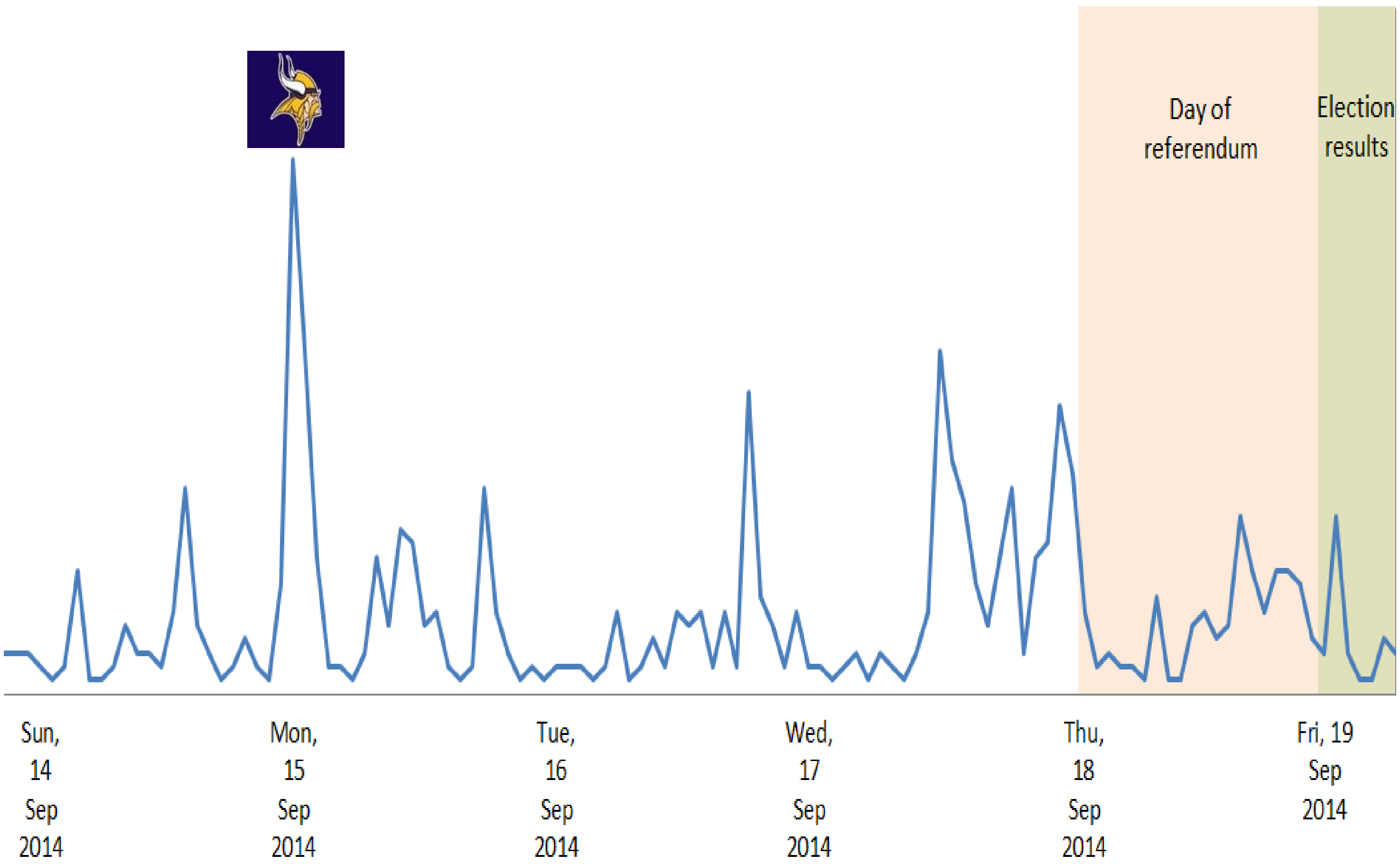
- Greater **timeliness** (Velocity)
  - ‘Nowcasting’ and ‘nearcasting’
  - Always important, especially in times of crisis
- Greater **efficiency** / value for money (Value)
  - Using administrative data
  - ‘Found’ data

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.

Exploiting novel datasets



Big data sets offer significant potential advantages

- New statistical / modelling approaches:
  - Machine learning
  - Network analysis
  - Agent-based modelling

Machine learning

- Different flavours:
  - Supervised
  - Unsupervised
  - Reinforcement learning
- Differences from conventional econometrics:
  - Typically focussed on prediction rather than identifying causal relationships
    - Individual parameter values are generally of limited interest
  - Use the algorithm and data to choose the model rather than theory
  - Use goodness of fit outside the ‘training set’ to determine the quality of the model rather than the familiar statistical tests
- Some key issues:
  - Feature selection
  - Regularisation
  - Researcher judgement vs ‘letting the data speak’
    - ‘Pure’ objectivity is unusual

Machine learning models: supervised learning

- Typical approach:
  - Partition data into three sets:
    - Training set – used to choose the model
    - Validation – used to calibrate it
    - Testing – used to assess it
  - Often repeat the process many times

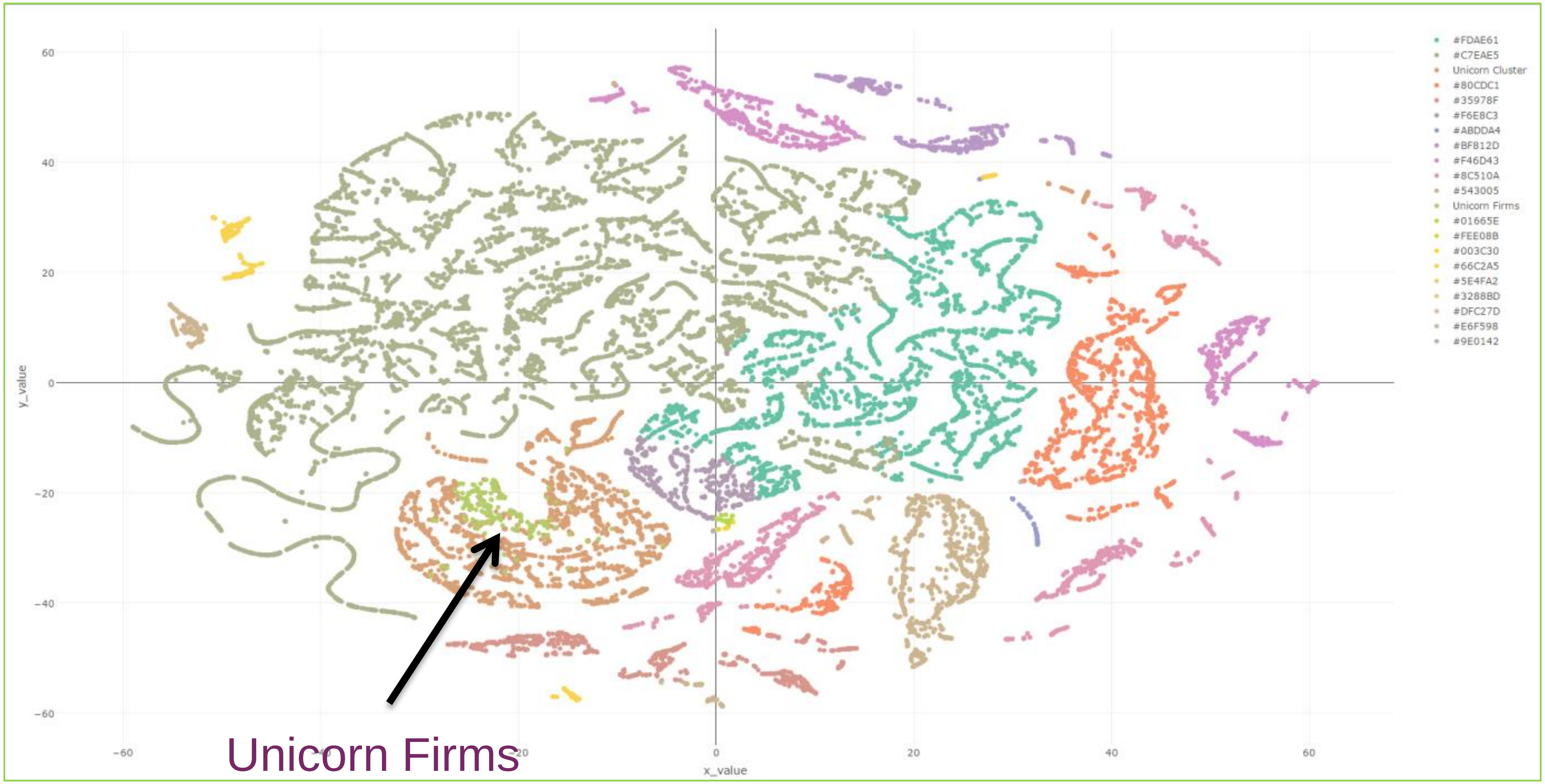
Machine learning models: supervised learning

- Some common models:
  - Linear regression-based:
    - Numerical solution of high dimensional models
    - Penalised regressions where number of explanatory variables is large relative to the number of observations (eg LASSO, Ridge, Elastic Net)
  - Non-linear regression:
    - Support vector machines
    - *K*-nearest neighbours
  - Tree-based:
    - Decision trees
    - Random forests
  - Neural networks

Machine learning models: unsupervised learning

- Classification and pattern identification
- Examples:
  - K-means
  - Hierarchical clustering
  - Neural networks (again)
  - Topic modelling

Cluster Analysis – Identifying potential financial disruptors



Identifying occupations

Three steps for grouping jobs based on the demand expressed in individual vacancies:

1. The text associated with each job vacancy is **cleaned** and the title and job description are combined into a single ‘document’ per vacancy

2. A **topic model** creates N topics to help determine **type** of segmentation

3. Group vacancies into K clusters (final sub-market **types**) using the K-means algorithm

Topic models and the LDA

- We model sectors using a topic model based on the Latent Dirichlet Allocation (LDA)
- Topics are identified by the use of common words and phrases
- Sectors are identified by being made up of common topics

$$\boldsymbol{\theta} = \left.\underbrace{\begin{bmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,N} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{D,1} & \theta_{D,2} & \cdots & \theta_{D,N} \end{bmatrix}}_{\text{Topics (columns)}}\right\}\text{Documents (rows)}$$

Document-topic matrix

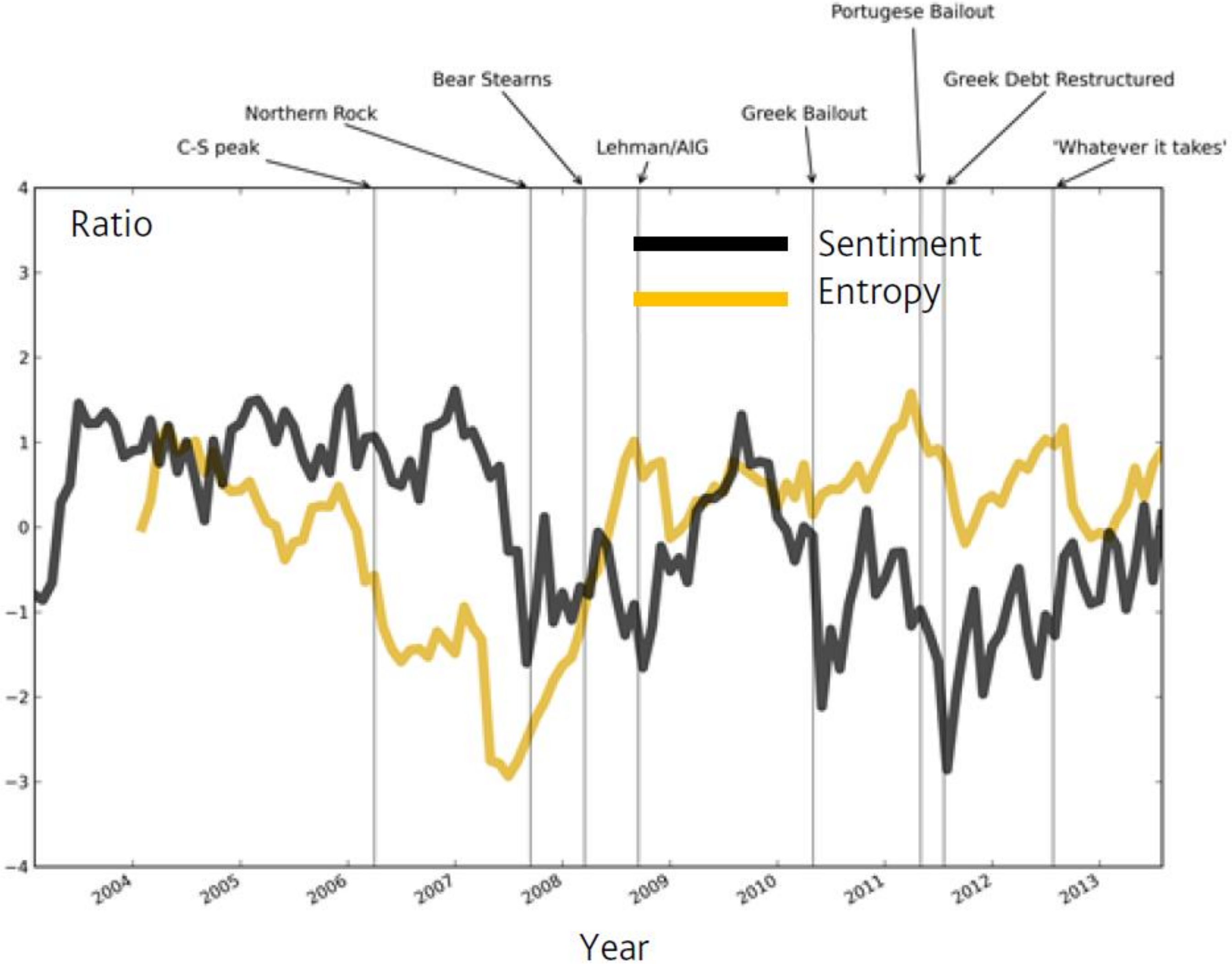
$$\boldsymbol{\beta} = \left.\underbrace{\begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,N} \\ \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{L,1} & \beta_{L,2} & \cdots & \beta_{L,N} \end{bmatrix}}_{\text{Topics (columns)}}\right\}\text{Words (rows)}$$

Term-topic matrix

The topics

Word clouds of topics found using Latent Dirichlet Allocation.

Sentiment or agreement?

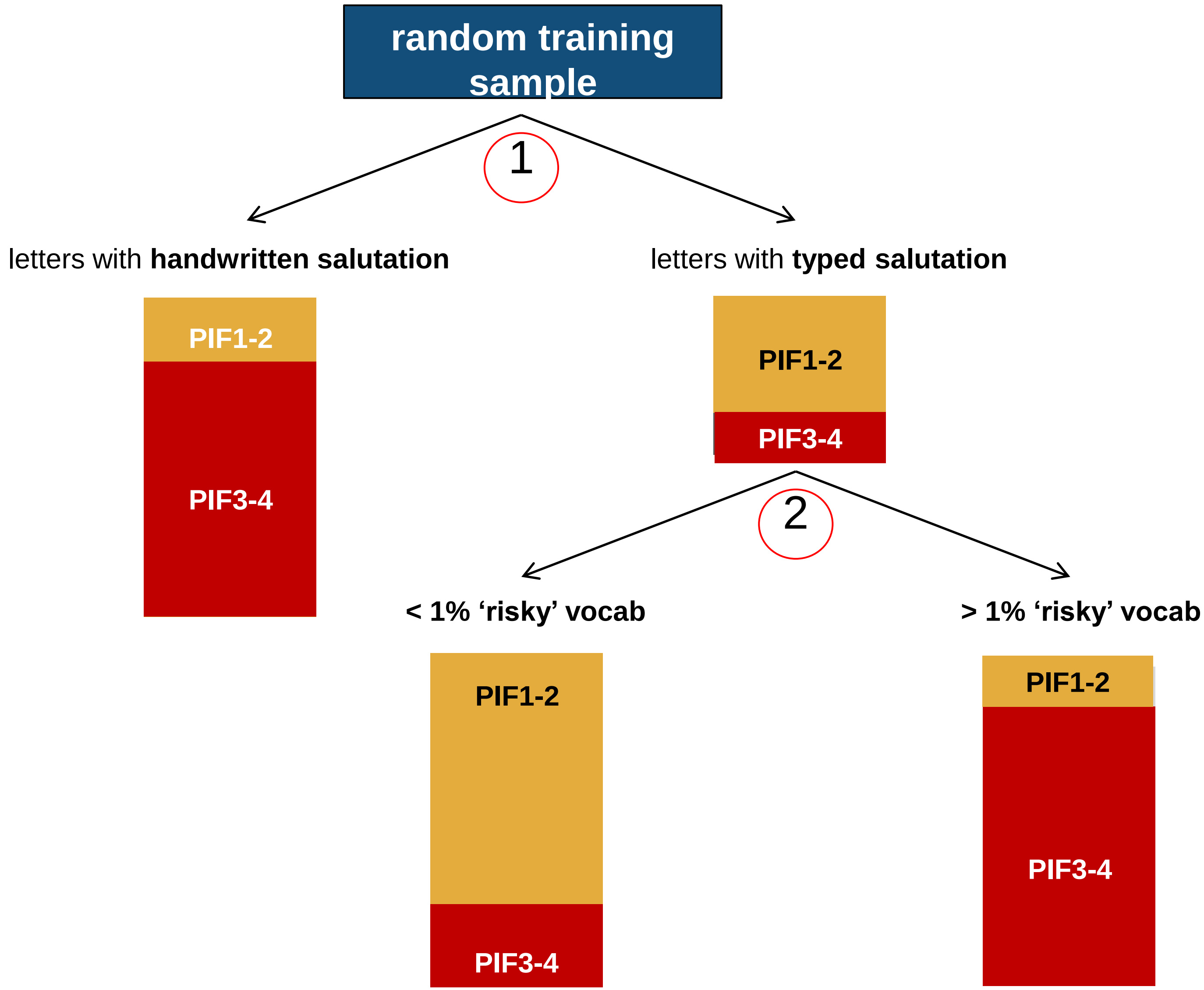


Using text and random forests to understand our own communications

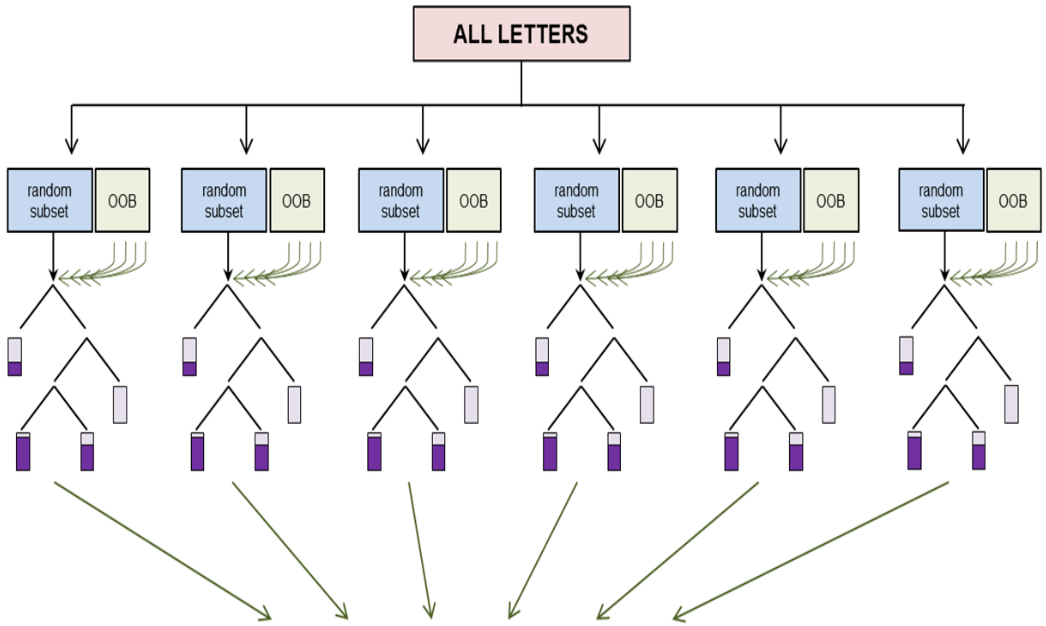
- Analysed periodic summary meeting (PSM) letters from the PRA to the supervised firms
- Are they written differently to firms with different risk profiles?
  - If so, what linguistic features distinguish sub-genres of PSM letters?
- We expected PSM letters to vary depending on firm riskiness
  - consistent with the PRA's principle of proportionality
- We expected higher risk firms to receive letters that were:
  - more complex
  - more negative in sentiment
  - more directive

Linguistic features considered

- Sentiment
  - Positive vs negative words
- Complexity
  - Length of sentences, number of subordinated clauses
- Directiveness
  - Instructions vs suggestions
- Formality
  - Eg “To Whom it may concern” [typed] vs “Dear Jane” [hand-written]
- Forward-lookingness
  - Future focus vs discussion of past developments


random training sample
1
letters with handwritten salutation
PIF1-2
PIF3-4
letters with typed salutation
PIF1-2
PIF3-4
2
< 1% ‘risky’ vocab
PIF1-2
PIF3-4
> 1% ‘risky’ vocab
PIF1-2
PIF3-4

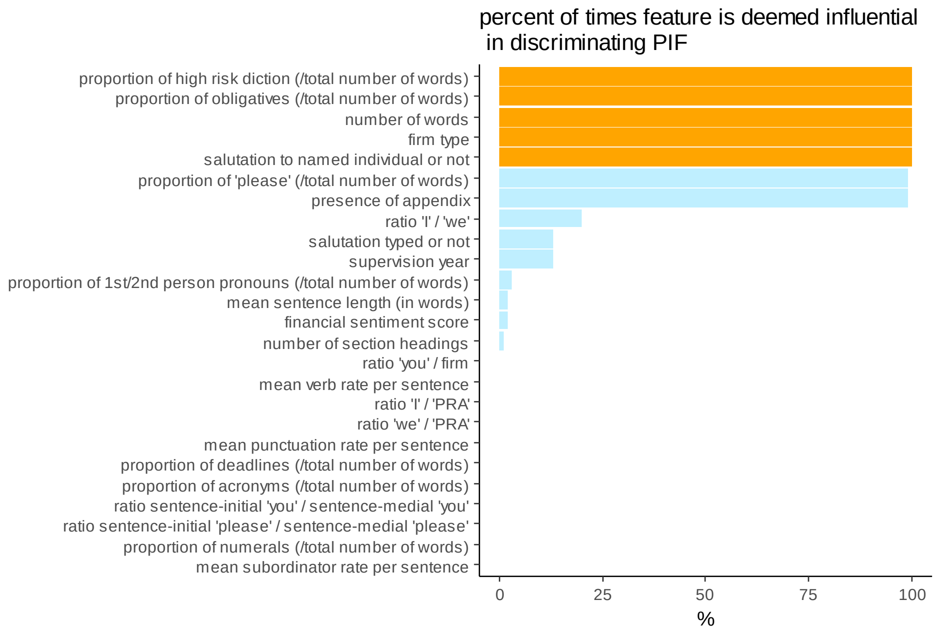
Random forests and text analytics in a regulatory context



Obs.	Obs. Class	Tree 1	Tree 2	Tree 3	Tree 4	Tree 5	Tree 6	Combined Prediction	Prediction == Observed
1	PIF 1-2		PIF 1-2		PIF1-2		PIF 3-4	PIF 1-2	TRUE
2	PIF 1-2	PIF 1-2				PIF 3-4	PIF 3-4	PIF 3-4	FALSE
3	PIF 1-2			PIF 1-2		PIF 1-2		PIF 1-2	TRUE
4	PIF 1-2		PIF 1-2		PIF 3-4		PIF 1-2	PIF 1-2	TRUE
5	PIF 1-2			PIF 1-2	PIF 1-2			PIF 1-2	TRUE
6	PIF 3-4	PIF 1-2	PIF 3-4				PIF 1-2	PIF 1-2	FALSE
7	PIF 3-4			PIF 1-2	PIF 1-2			PIF 1-2	FALSE
8	PIF 3-4		PIF 3-4			PIF 3-4		PIF 3-4	TRUE
9	PIF 3-4	PIF 3-4		PIF 1-2		PIF 3-4		PIF 3-4	TRUE
10	PIF 3-4	PIF 1-2						PIF 1-2	FALSE

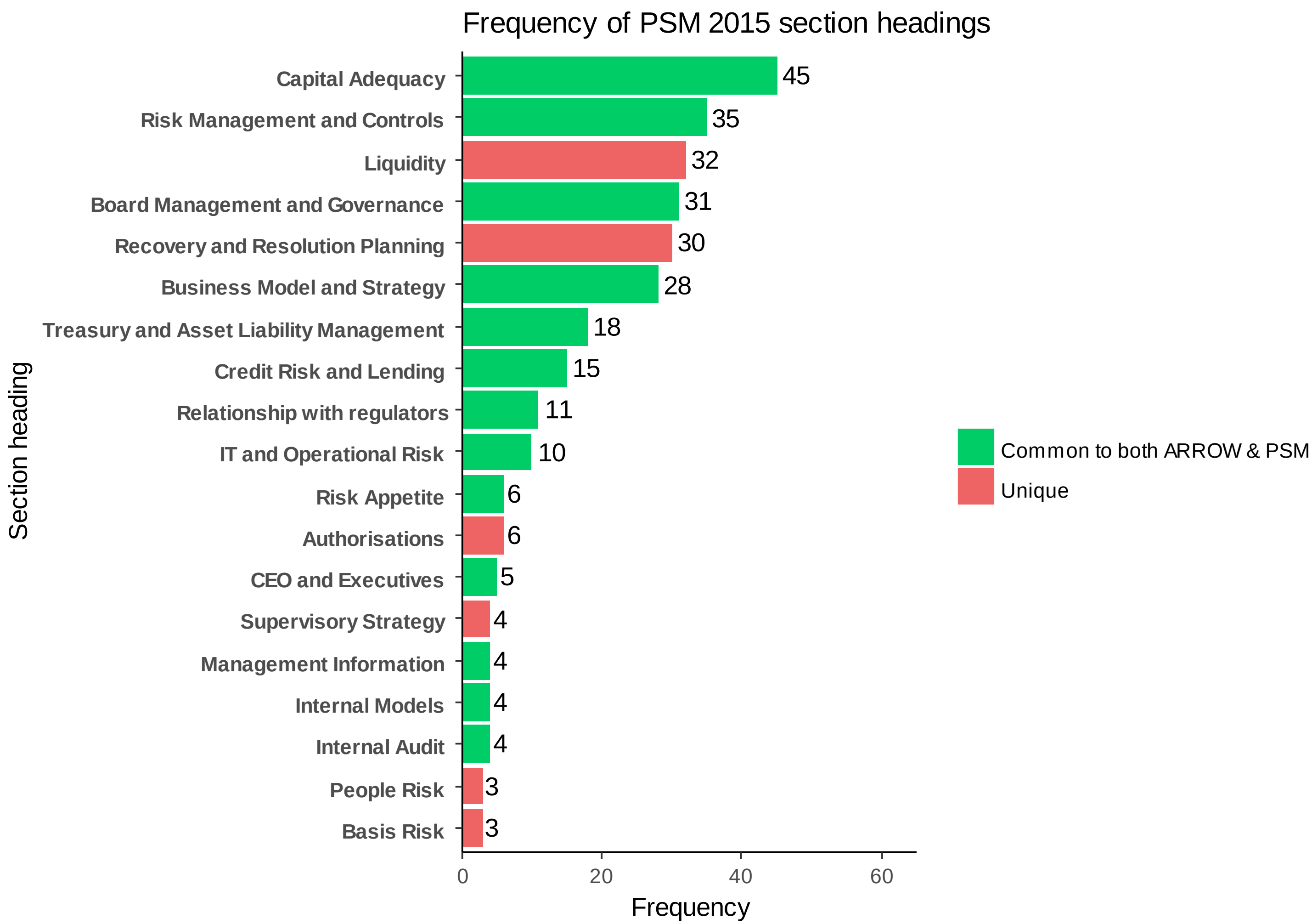
OOB Predictive Accuracy = 60%

PIF 3-4 PSM letters different from PIF 1-2 letters



- More complex
- More ‘high-risk’ vocabulary
- Less directive
- Less formal

PSM letters different from ARROW letters in content



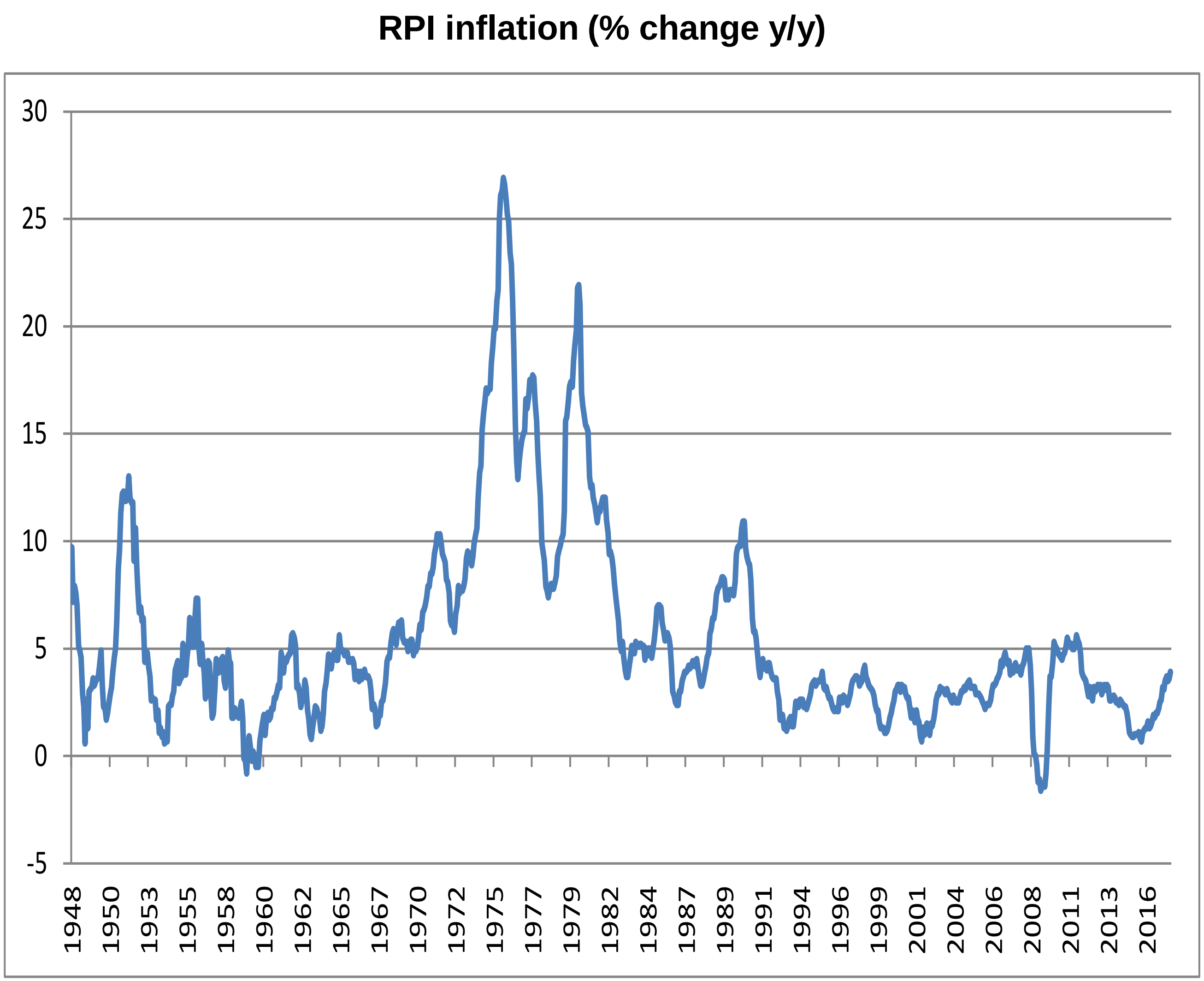
BANK OF ENGLAND

But there is no such thing as a free lunch …

Lots of data == lots of information?

- 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

Lots of data == lots of information?

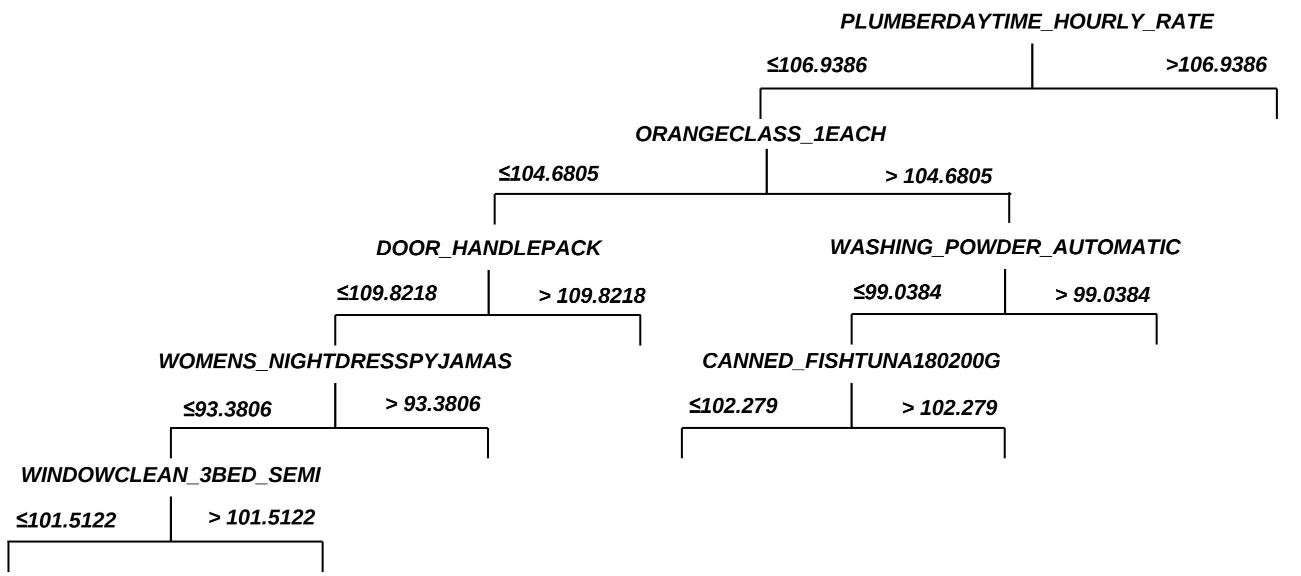


Correlation versus Causality

- ML focuses on prediction
  - Not on structural models
  - But central banks set policy and a policy intervention may change the structure of the economy
  - Beware the ‘Lucas critique’ (and structural breaks)
- This does not mean that ML is not a good fit for central banks
  - Forecasts often matter
  - Intermediate targets can be useful

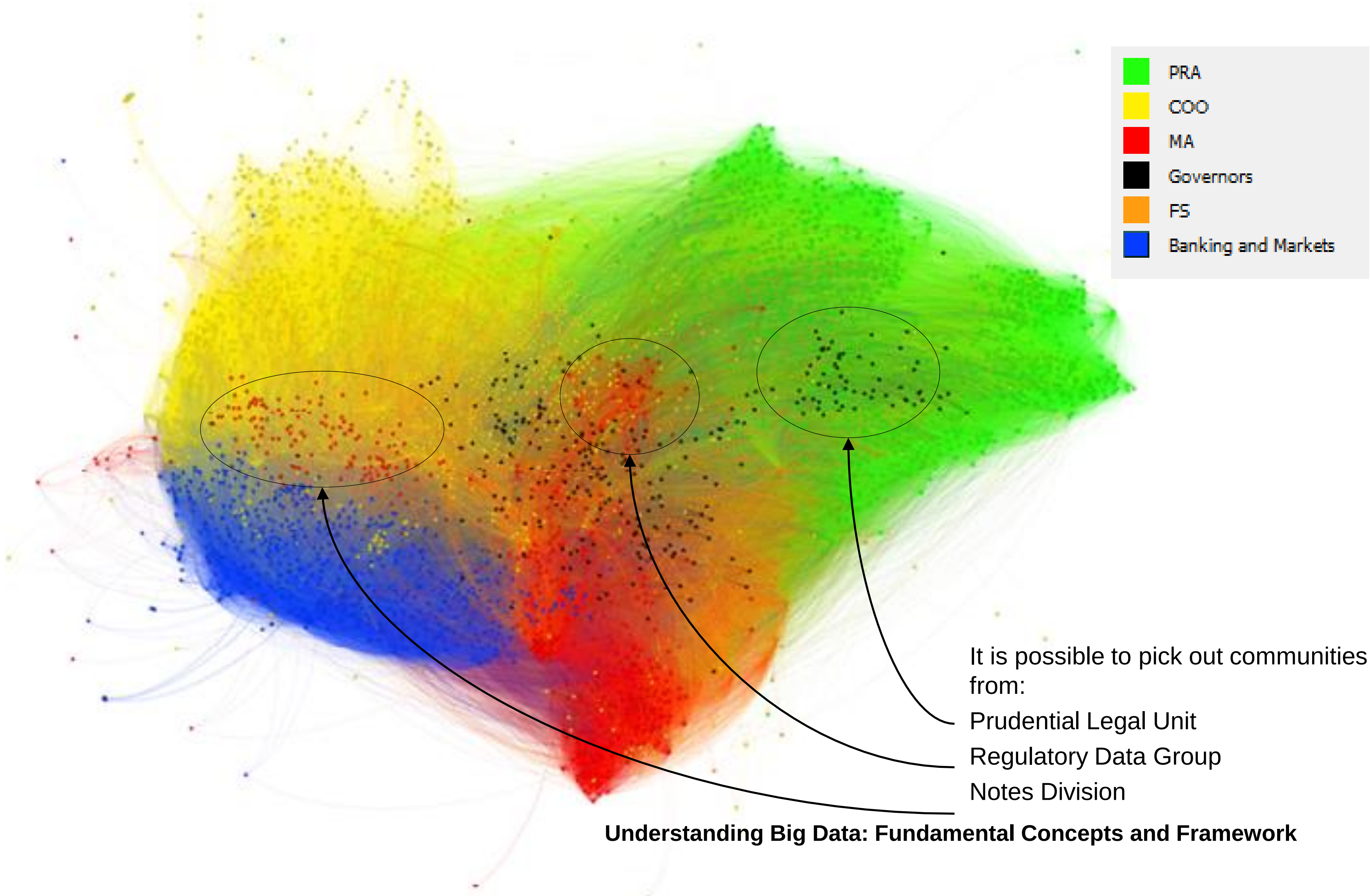
Overfitting

- Large data sets contain huge numbers of correlations …



Interpreting complicated, often highly non-linear relationships

Email connections: January 2015



“Veracity”

- Big data sets are often populations, not samples
  - Therefore no sampling error
- But the observed population characteristics may not be typical of the underlying data generating process
- Or it may be biased relative to the true population of interest

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

Engage with AA, but there are no free lunches …