Understanding big data: fundamental concepts and framework

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.
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 (e.g., 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

UK house price inflation (% y/y)

Source: Average of HBOS and Nationwide measures
• Key: % of mortgages with loan more than 4.5x income

Understanding Big Data: Fundamental Concepts and Framework
Tracking home movers

Loan 1
Date: 05/01/2008
Type: First time buyer
DoB: 11/12/1880 (Borrower a)
Postcode: XX1 2XX (Property x)

Loan 2
Date: 15/01/2011
Type: Home mover
DoB: 11/12/1880 (Borrower a)
Postcode: YY1 2YY (Property y)

Loan 3
Date: 23/01/2011
Type: Home mover
DoB: 24/04/1887 (Borrower b)
Postcode: XX1 2XX (Property x)

Moving in candidates
Possible matches based on post codes

Moving out candidates
Possible Matches based on date of births
Large-scale data analysis

- EURCHF trades
- 9.30-10.10am

Swiss francs per euro

08:00:00 08:30:00 09:00:00 09:30:00 10:00:00 10:30:00 11:00:00 11:30:00 12:00:00 12:30:00 13:00:00 13:30:00 14:00:00 14:30:00 15:00:00 15:30:00 16:00:00 16:30:00 17:00:00

0.70 0.75 0.80 0.85 0.90 0.95 1.00 1.05 1.10 1.15 1.20 1.25

Understanding Big Data: Fundamental Concepts and Framework
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

Productivity and matching efficiency at 1-digit SOC code

Output per worker (normalised to 100, LHS)  Matching efficiency (RHS)

Managers, Directors And Senior Officials (1)  Professional Occupations (2)  Associate Professional And Technical Occupations (3)  Administrative And Secretarial Occupations (4)  Skilled Trades Occupations (5)  Caring, Leisure And Other Service Occupations (6)  Sales And Customer Service Occupations (7)  Process, Plant And Machine Operatives (8)  Elementary Occupations (9)
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

Understanding Big Data: Fundamental Concepts and Framework
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

Unicorn Firms
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

\[
\theta = \begin{bmatrix}
\theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,N} \\
\theta_{2,1} & \theta_{2,2} & \cdots & \theta_{2,N} \\
\vdots & \vdots & \ddots & \vdots \\
\theta_{D,1} & \theta_{D,2} & \cdots & \theta_{D,N}
\end{bmatrix}
\]

Documents (rows)  Topics (columns)

\[
\beta = \begin{bmatrix}
\beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,N} \\
\beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,N} \\
\vdots & \vdots & \ddots & \vdots \\
\beta_{L,1} & \beta_{L,2} & \cdots & \beta_{L,N}
\end{bmatrix}
\]

Words (rows)  Topics (columns)

Document-topic matrix  Term-topic matrix
The topics

Word clouds of topics found using Latent Dirichlet Allocation.
Sentiment or agreement?

![Graph showing sentiment and entropy over years.](image-url)
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

letters with handwritten salutation

letters with typed salutation

< 1% ‘risky’ vocab

> 1% ‘risky’ vocab

Text mining PSM letters
Random forests and text analytics in a regulatory context

ALL LETTERS

random subset OOB random subset OOB random subset OOB random subset OOB random subset OOB

combined prediction PIF 1-2 PIF 1-2 PIF 1-2 PIF 1-2 PIF 1-2

Prediction == Observed TRUE FALSE TRUE TRUE FALSE

OOB Predictive Accuracy = 90%

Advanced analytics, data and tools
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

Frequency of PSM 2015 section headings

- Capital Adequacy: 45
- Risk Management and Controls: 35
- Liquidity: 32
- Board Management and Governance: 31
- Recovery and Resolution Planning: 30
- Business Model and Strategy: 28
- Treasury and Asset Liability Management: 18
- Credit Risk and Lending: 15
- Relationship with regulators: 11
- IT and Operational Risk: 10
- Risk Appetite: 6
- Authorisations: 6
- CEO and Executives: 5
- Supervisory Strategy: 4
- Management Information: 4
- Internal Models: 4
- Internal Audit: 4
- People Risk: 3
- Basis Risk: 3

Common to both ARROW & PSM

Unique

Section heading

Frequency

Text mining PSM letters
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?

RPI inflation (% change y/y)
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

It is possible to pick out communities from:
- Prudential Legal Unit
- Regulatory Data Group
- Notes Division

Understanding Big Data: Fundamental Concepts and Framework
“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 …