BANK OF ENGLAND

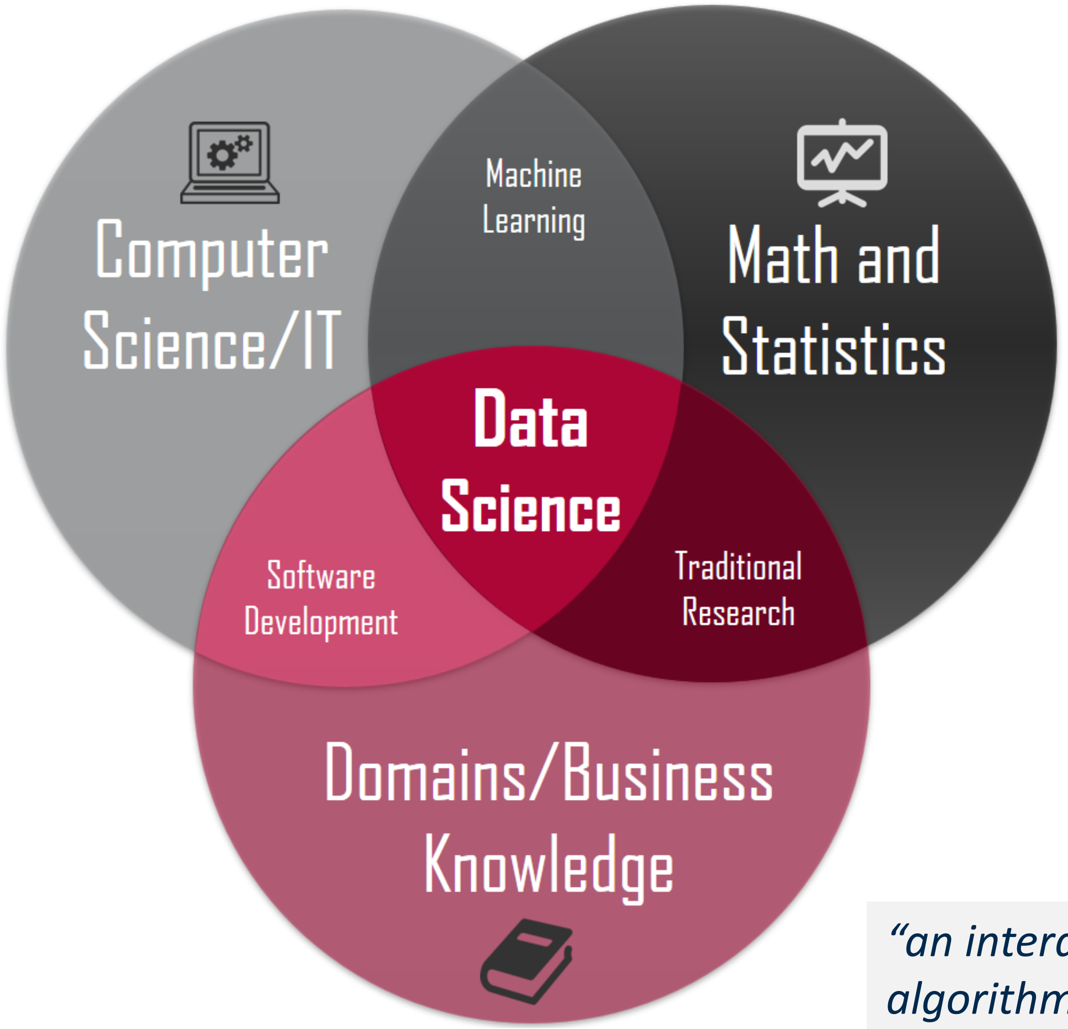
Taras Shevchenko National University of Kyiv
May 23, 2019

Data Science in the Bank of England

Who are we and what do we do?

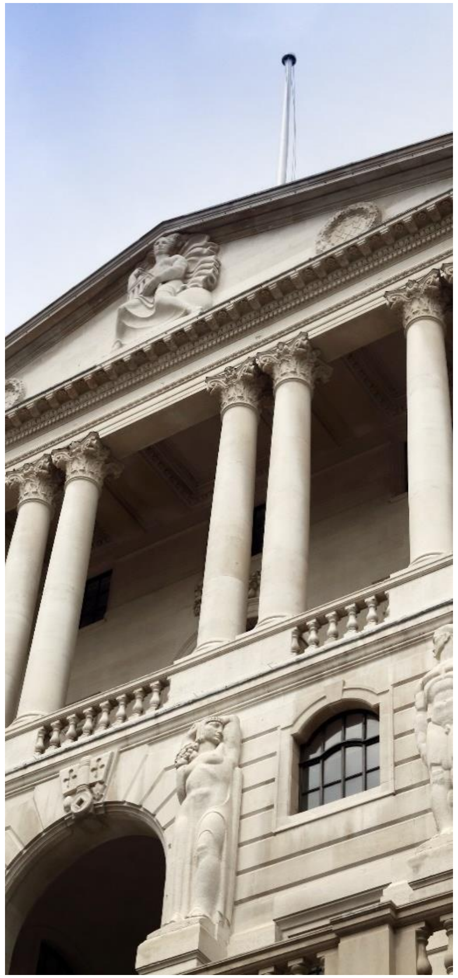
Eryk Walczak
Data Scientist
Bank of England

What is data science?



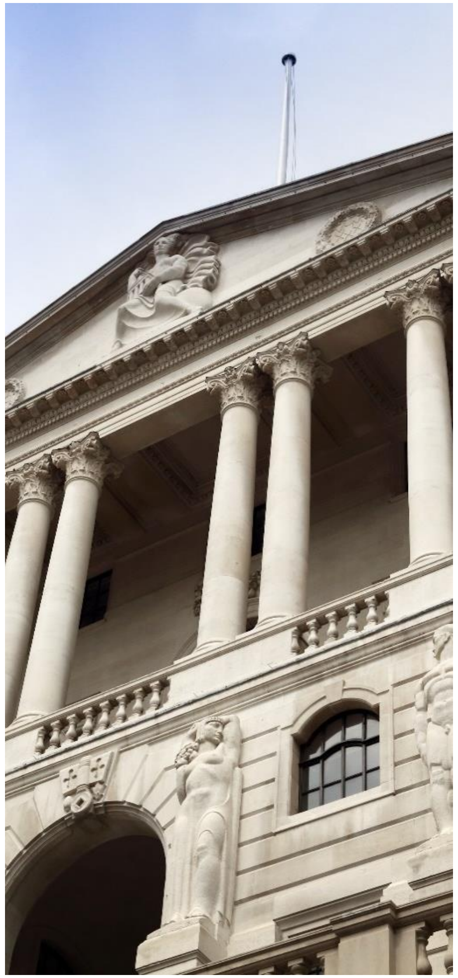
"an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured"

Bank of England



- UK’s central bank
- Our mission is to promote the good of the people by maintaining monetary and financial stability

Bank of England



- Regulate other banks
- Issue banknotes
- Set monetary policy
- Maintain stability

Structure

- **Part I**
- What is Advanced Analytics (AA)?
- What do we do?
- What problems do we like to tackle?

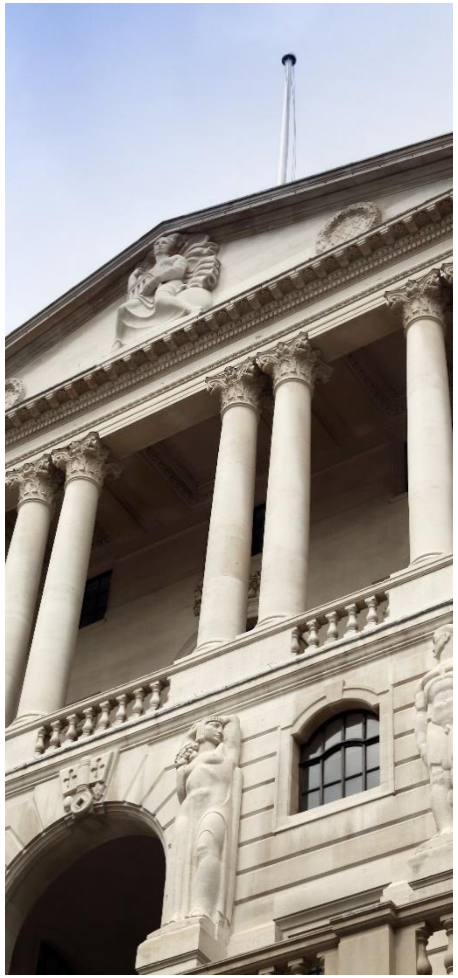
- **Part II**
- Detailed example of AA work

Advanced Analytics' Strategic Goals

- Advanced Analytics exists in order to help the Bank achieve excellence as a cutting-edge research and analytics institution.

- Our mission is to:
  - Use optimal statistical techniques to answer Bank-relevant questions
  - Facilitate more effective use of granular data sets in the Bank
  - Meet computational challenges when using Big/complex data

Advanced Analytics



Applied Data Science and Visualisation Team

- Leads on efficient use of technology for data science, and engagement with Technology
- Leads on defining and applying best practice in data visualisation
- Operationalises data science techniques

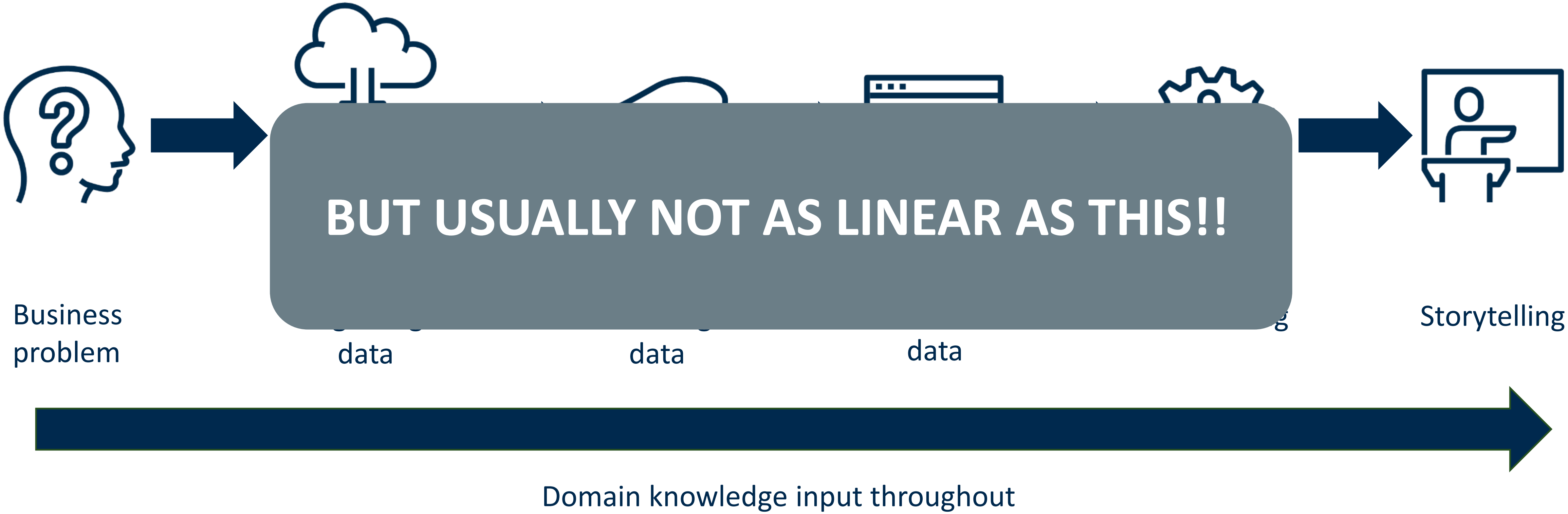
Analytics, Research and Outreach Team

- Leads on research and publications
- Expands the range of quantitative methods used by the Bank
- Promotes collaboration with external researchers and the Bank's research community

How we work

- Collaborating with other areas of the Bank
- Business areas bring **domain understanding**
- AA brings **data science** solutions to bear on business problems

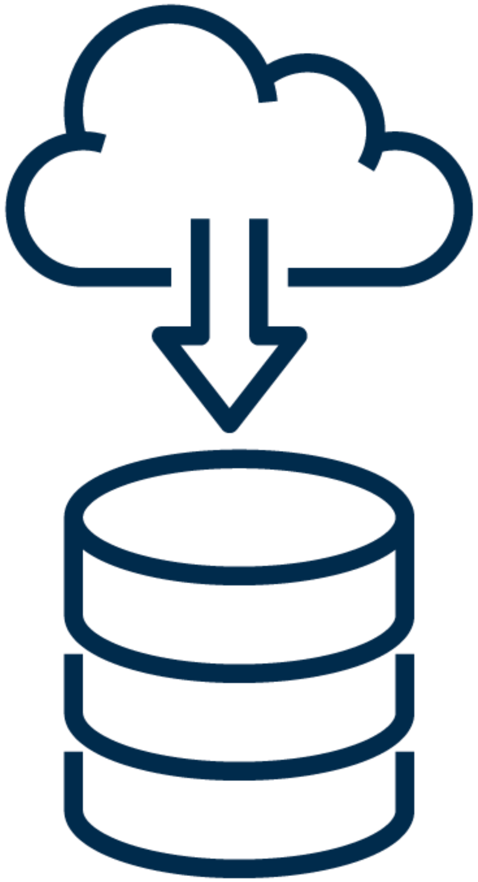
“Ideal” Data Science Pipeline



Business Problem

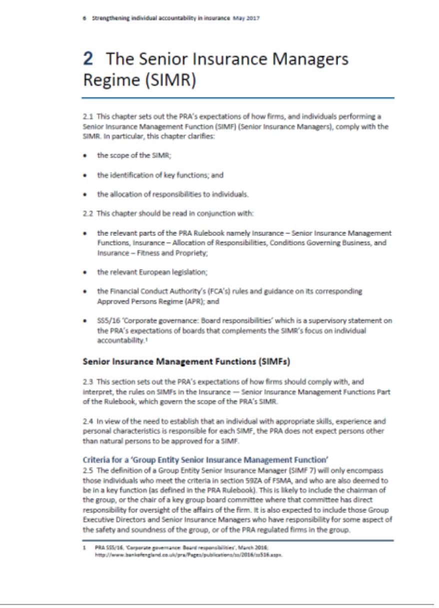
- Are we communicating with firms proportionately?
- Can we spot potentially disruptive technological trends?
- What are the trends in the UK labour market?
- What's the effectiveness of different approaches to central bank communications?
- How complex does our regulation need to be to meet our prudential objectives?

Obtaining Data



- Web scraping
- PDFs
- Images
- Audio
- Matching datasets from multiple sources

Example: PDFs (Supervisory Statements)



File Edit Format View Help

2 The Senior Insurance Managers
Regime (SIMR)

2.1 This chapter sets out the PRA's expectations of how firms,
and individuals performing a
Senior Insurance Management Function (SIMF) (Senior Insurance
Managers), comply with the
SIMR. In particular, this chapter clarifies:

. the scope of the SIMR;
. the identification of key functions; and
. the allocation of responsibilities to individuals.
2.2 This chapter should be read in conjunction with:

. the relevant parts of the PRA Rulebook namely Insurance –
Senior Insurance Management
Functions, Insurance – Allocation of Responsibilities,
Conditions Governing Business, and
Insurance – Fitness and Propriety;
. the relevant European legislation;
. the Financial Conduct Authority's (FCA's) rules and guidance
on its corresponding
Approved Persons Regime (APR); and
. SS5/16 'Corporate governance: Board responsibilities' which
is a supervisory statement on
the PRA's expectations of boards that complements the SIMR's
focus on individual
accountability.1

1 PRA SS5/16, 'Corporate governance: Board responsibilities',
March 2016;
http://www.bankofengland.co.uk/pra/Pages/publications/ss/2016/
ss516.aspx.

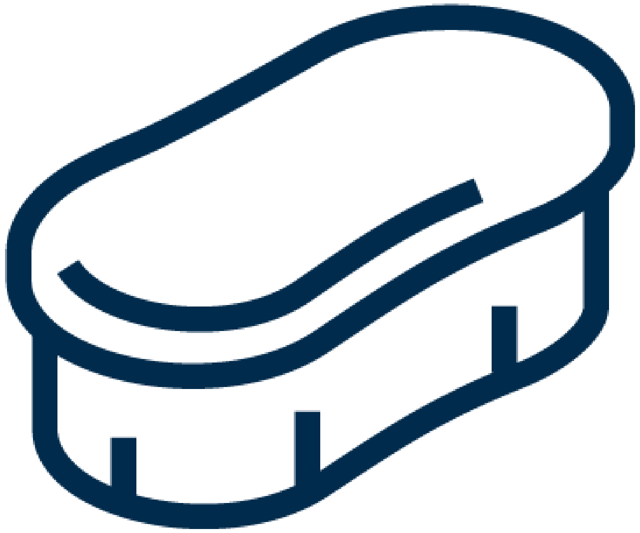
Senior Insurance Management Functions (SIMFs)

2.3 This section sets out the PRA's expectations of how firms
should comply with, and
interpret, the rules on SIMFs in the Insurance – Senior
Insurance Management Functions Part
of the Rulebook, which govern the scope of the PRA's SIMR.
2.4 In view of the need to establish that an individual with
appropriate skills, experience and
personal characteristics is responsible for each SIMF, the PRA
does not expect persons other
than natural persons to be approved for a SIMF.

```r
for (doc in 1:length(text)){
  dcmt <- text[doc]
  all_text <- unlist(strsplit(dcmt, "\f"))
  title_text <- all_text[1]
  body_start <- which((nchar(all_text) > 300 & !grepl("Co
  body_end <- ifelse(identical(which(grepl('Appendix\\:|A
                     which(grepl('Appendix\\:|Annex -', a
  main_text <- all_text[body_start:body_end]


  # get title

  spl1 <- unlist(strsplit(title_text, "\\|"))
  #doc_type <- trimws(spl1[1])
  months <- gsub(" ", "", paste(format(ISOdate(1111,1:12,
  spl_ptn <- substr(months, 1, nchar(months)-1)
  spl2 <- unlist(strsplit(spl1[2], split = spl_ptn))[1]
  issue <- trimws(str_extract(spl2, "SS[:digit:]+\\/[:dig
  title <- tm::stripWhitespace(trimws(unlist(strsplit(spl
  updates_which <- ifelse(is.na(trimws(str_extract(paste(
                          trimws(str_extract(paste(spl1,

  # get sections and section numbers

  main_text <- paste(main_text, collapse = "\n")

  sections <- unlist(str_split(main_text,regex("(  [0-9]+
  sect_nos <- trimws(unlist(str_extract_all(main_text, re


  # put into dataframe

  df <- data.frame(#doc_type = rep(doc_type, length(secti
    issue = rep(issue, length(sections)),
    title = rep(title, length(sections)),
    updates_which = rep(updates_which, length(sections)),
    section_number = sect_nos,
    text = trimws(tm::stripWhitespace(sections)))
```

	issue	title	updates_which	section_number	text
1	SS3/13	Capital requirements for major UK banks and building s...	new document	1.1	The purpose of this supervisory statement is to set out t...
2	SS3/13	Capital requirements for major UK banks and building s...	new document	1.2	Firms will be expected to meet a 7% common equity Tier ...
3	SS3/13	Capital requirements for major UK banks and building s...	new document	1.3	The PRA expects firms to meet the above standards at th...
4	SS3/13	Capital requirements for major UK banks and building s...	new document	1.4	From 1 July 2014, the adjustments to CET1 capital and ris...
5	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	1.1	This supervisory statement is aimed at firms and groups ...
6	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	1.2	The statement expands on the PRA's general approach, a...
7	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	1.3	EIOPA's guidelines are addressed to National Competent...
8	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	1.4	Firms are reminded that they must also continue to com...
9	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	1.5	The PRA believes this statement is compatible with the R...
10	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	1.6	This supervisory statement will apply when EIOPA's guid...
11	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	1.7	The guidelines cover four areas which EIOPA considers f...
12	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	1.8	This statement aligns with the four areas above, and pro...
13	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	1.9	In this supervisory statement, the PRA articulates its exp...
14	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	1.10	The PRA expects EIOPA to launch a Q&A process during ...
15	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	1.11	The PRA supports EIOPA's proportionate and pragmatic ...
16	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	1.12	In considering the PRA's strategic approach against the ...
17	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	1.13	Firms will be expected to apply the guidelines in a way t...
18	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	1.14	Many of the guidelines represent good practice in conf...
19	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	1.15	The guidelines set out the following principles of applic...
20	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	1.16	The PRA will use its work with firms during the preparat...
21	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	2.1	The PRA consulted on this statement in CP9/13. Eightee...
22	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	2.2	Overall, stakeholders welcomed the pragmatic and prop...
23	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	2.3	Some respondents sought clarification of the PRA's expe...
24	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	2.4	Amendments have been made to some sections to enha...
25	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	3.1	This chapter sets out the PRA's expectations in respect of...
26	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	3.2	These guidelines will assist firms in developing their gov...
27	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	3.3	The PRA considers the guidelines to be generally consist...
28	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	3.4	The PRA will not prescribe how these requirements shou...
29	SS4/13	Solvency II: applying EIOPA's preparatory guidelines to P...	new document	3.5	It is anticipated that Solvency II will apply the system of ...

Cleaning data

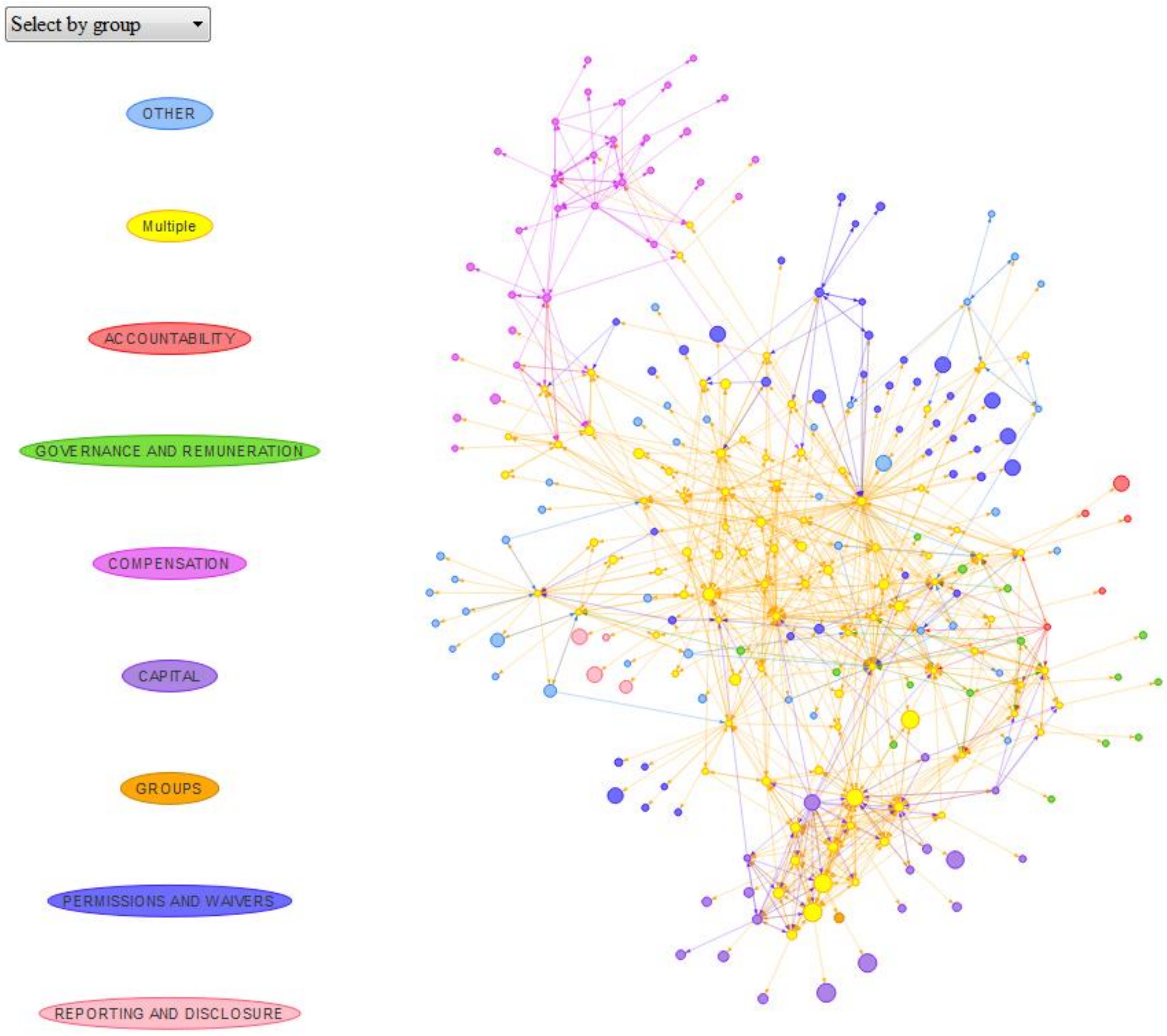


- Removing duplicate records
- Missing data imputation

Data exploration

- Initial probing of data using visualization techniques

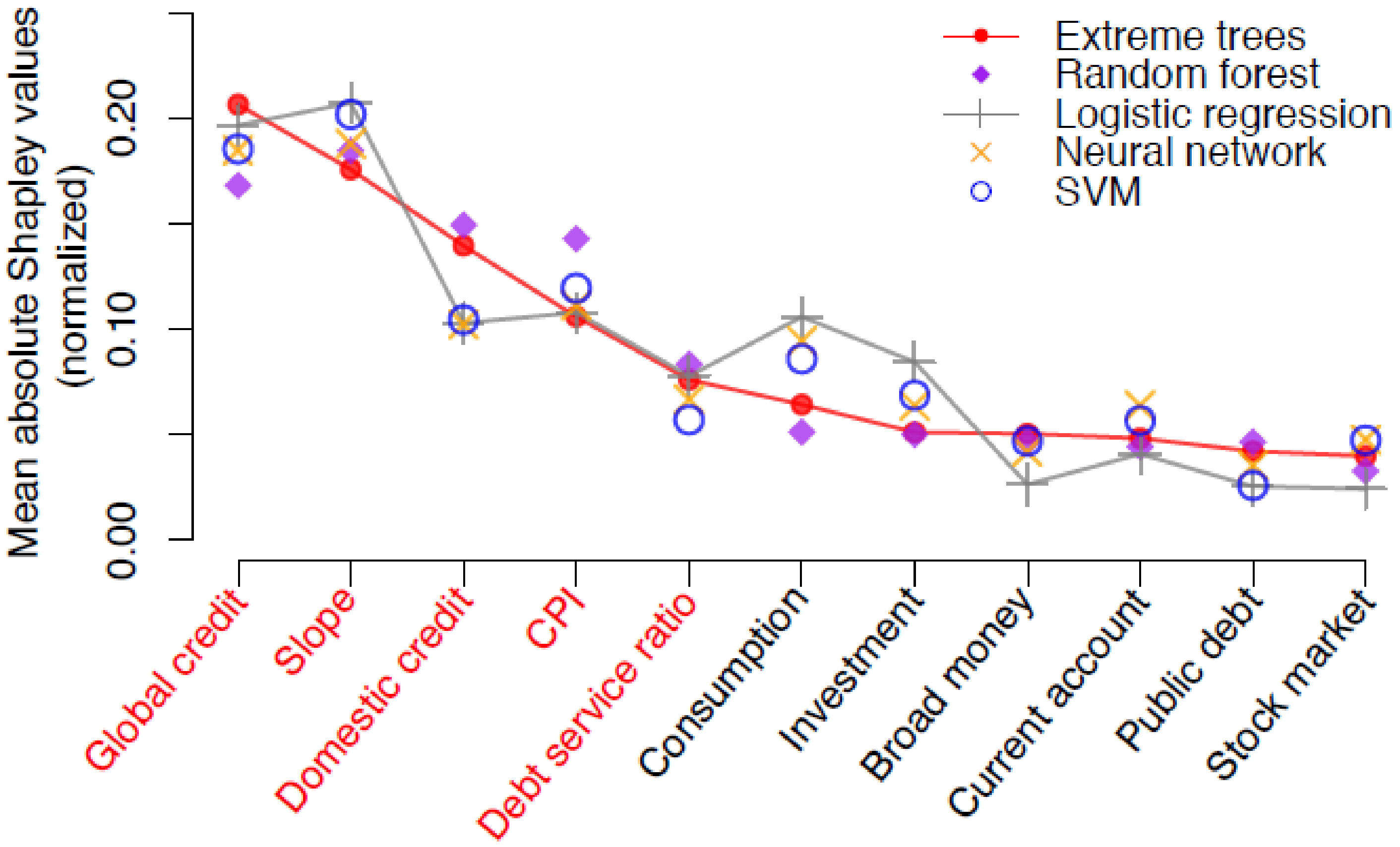
Example: Network plot (FSA Handbook)



Statistical modelling

- Prediction
- Variable interpretation
- Clustering

Example: features that predict financial crises



Storytelling fora

- Data products
- Staff Working Papers
- Bank Underground Posts
- Present at Data Boards
- Senior Management Team meetings
- CCBS events
- Research Showcase
- Dashboards in R Shiny and Tableau

Do I need a data scientist?

- **Problems which a Data Scientist will solve**
  - Do we have big data sets which cannot be handled with standard analyst tools/techniques?
  - Do we need to use new types of data?
  - Do we need extract clean information from messy data?
  - Do we need to automate labour-intensive analytical or data processes?
  - Do we need to make data processes better, more robust, faster?
  - Do we want to use machine learning techniques?
- **Poor uses for a Data Scientist**
  - Standard analysis, on known datasets

Source: Al Firrell

Examples

Forecasting Inflation

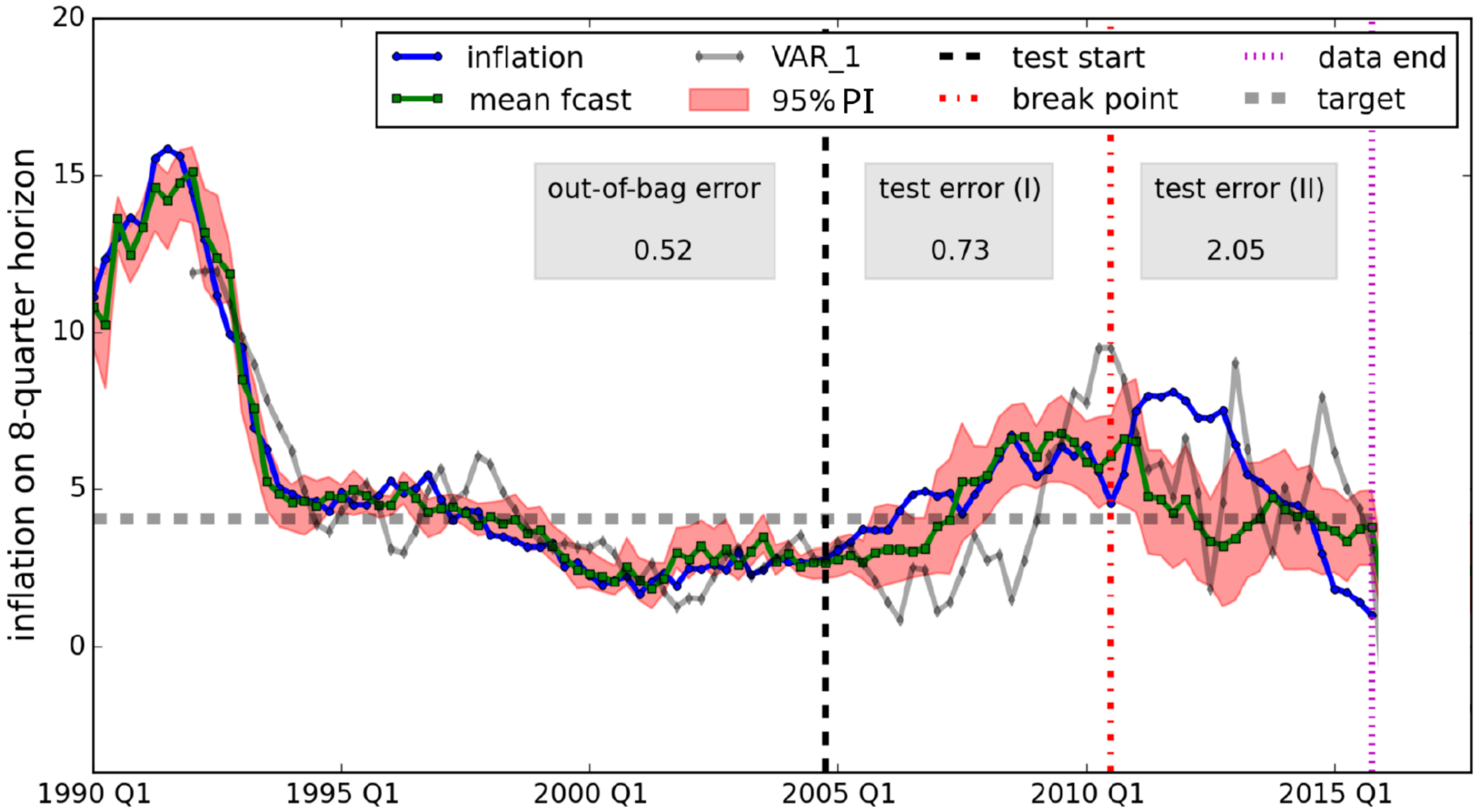
- Quarterly prediction of UK's inflation
- Regression problem
- Predictors: household income, broad money, unemployment, ..., measured two years earlier.
- Time period: 1990-2015

Chakraborty and Joseph, 2017

Forecasting Inflation

Method	Mean absoulute test error (SD) Years: 2005–2015
AR_1	2.02 (0.95)
VAR_1	2.32 (1.42)
Ridge regression	2.79 (2.31)
SVM	1.58 (1.26)
Neural net	1.96 (1.67)
SVM + Neural net	**1.35** (1.21)

Chakraborty and Joseph, 2017



Chakraborty and Joseph, 2017

Clustering job advertisements

- Existing job classification schemes: 21 sectors, 90 occupations

- Carefully designed and well-established
- Analysis over time is possible
- Do not reflect changes in the labour market
- Manual assignment can be inaccurate (Schierholz et al., 2018; Belloni et al., 2014)

Turrell et al., 2018

Clustering job advertisements

Research project by Turrell et al. (2018):

- Bottom-up taxonomy based on 15 million job adverts posted online on reed.co.uk between 2008-2016
- Machine learning to (1) identify topics in adverts, (2) cluster adverts according to topics
- The new taxonomy has explanatory power in predicting offered wages on top of established taxonomies (sector, occupation, region)

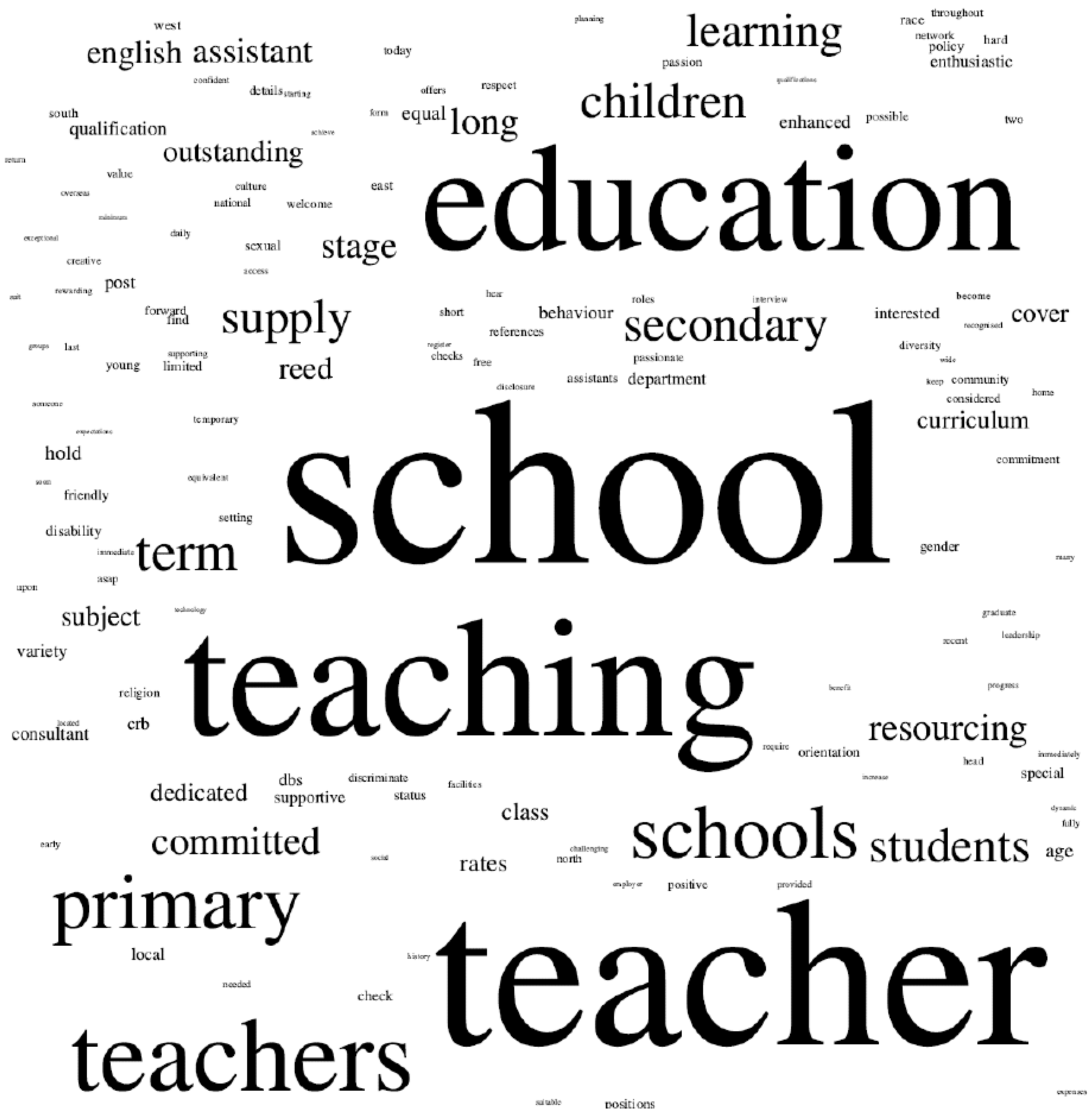
Clustering job advertisements

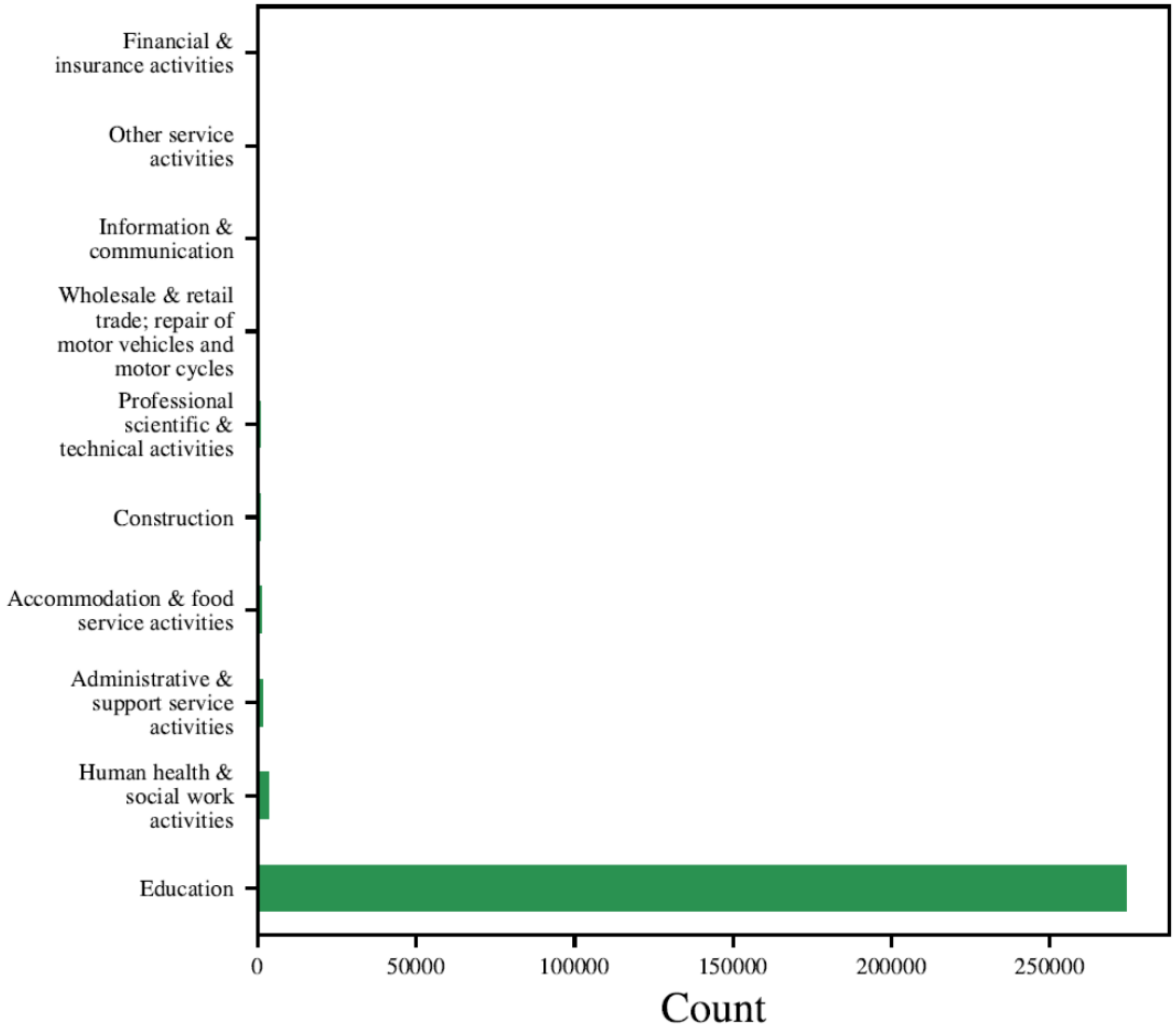
- **Topic modelling**

- A topic is a collection of words that frequently occur together in a document
- Each document covers only few topics

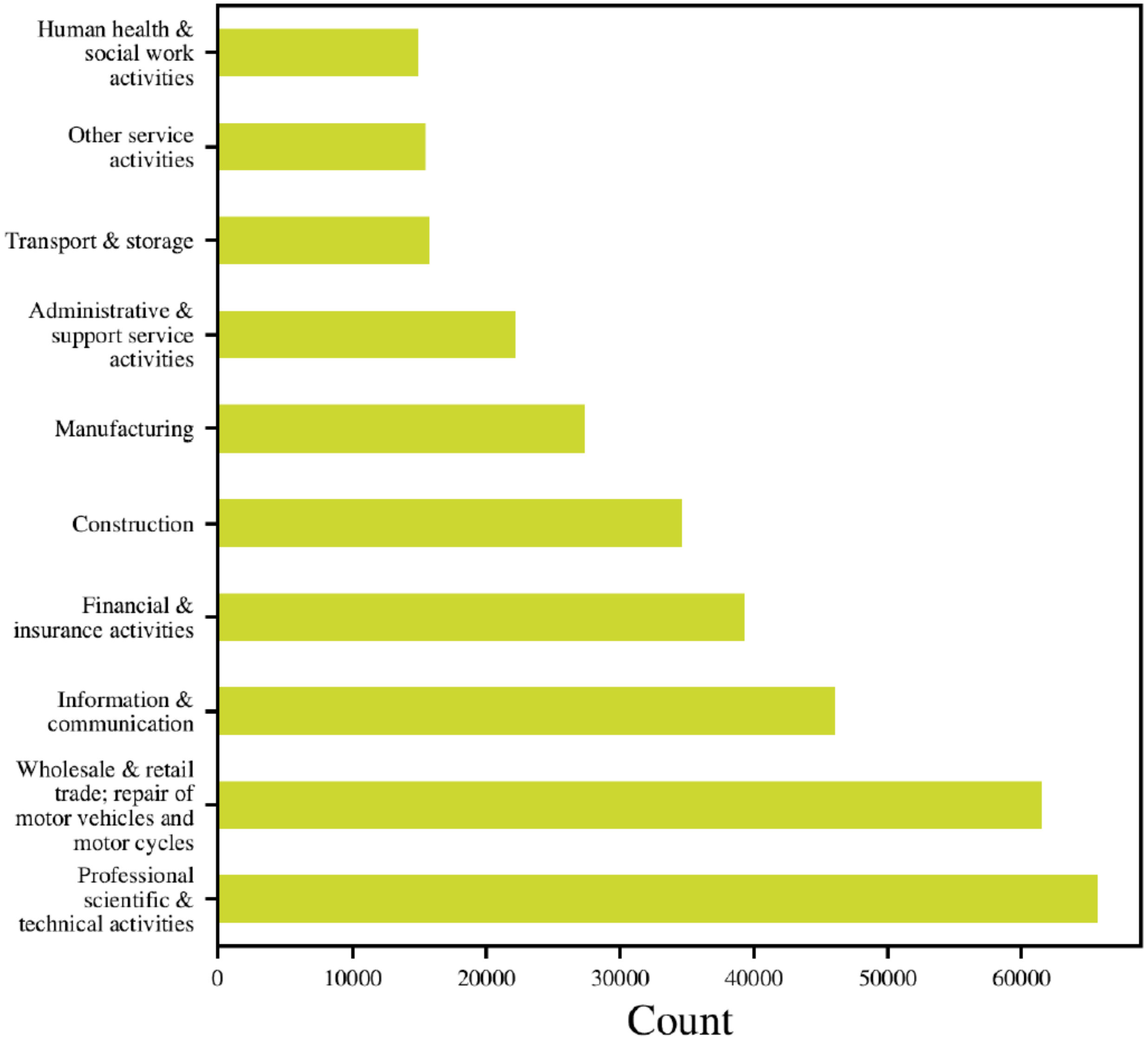
- Topic model learns
  (1) words that characterise topics
  (2) which topics are represented in which document
- Humans have to interpret the topics

- Example: [school, students, learning, teaching, ...] -> Topic Education


school
teacher
teaching
education
teachers
schools
primary
secondary
supply
children
learning
english assistant
students
resourcing
committed
dedicated
term
outstanding
stage
long
reed
curriculum
cover
qualification
subject
class
rates
behaviour
department
enhanced
enthusiastic
equal
hold
friendly
disability
variety
consultant
religion
crb
dbs
supportive
discriminate
status
local
check
positions
positive
orientation
special
age
gender
interested
community
considered
diversity
assistants
references
passionate
checks
young
limited
post
sexual
creative
welcome
national
culture
value
south
west
today
passion
race
throughout
network
policy
hard
possible
two
east
details
respect
offers
form
short
roles
commitment
setting
temporary
equivalent
subject
north
provided
head
free
last
find
forward


Financial & insurance activities
Other service activities
Information & communication
Wholesale & retail trade; repair of motor vehicles and motor cycles
Professional scientific & technical activities
Construction
Accommodation & food service activities
Administrative & support service activities
Human health & social work activities
Education
0
50000
100000
150000
200000
250000
Count

Project management cluster



BANK OF ENGLAND

Sending firm messages: Text mining PSM letters

David Bholat, James Brookes, Chris Cai, Katy Grundy and Jakob Lund

Primary research question and hypotheses

Are PSM letters 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 <u>complex</u>
- more negative in <u>sentiment</u>
- more <u>directive</u>

‘Intrinsic risk’ = Potential Impact = Firm Category

Increasing risk

Category 1	Most significant deposit-takers capable of **very significant disruption**
Category 2	Significant deposit-takers capable of **some disruption**
Category 3	Deposit-takers capable of **minor disruption**
Category 4	Deposit-takers capable of **very little disruption**
Category 5	Deposit-takers capable of **almost no disruption**

‘Imminent risk’ = PIF stage = proximity to resolution

Increasing risk

Stage 1	**Low risk** to viability of firm
Stage 2	**Moderate risk** to viability of firm
Stage 3	**Risk to viability** absent action by the firm
Stage 4	**Imminent risk** to viability of firm
Stage 5	Firms **in resolution** or being actively wound up

Secondary research question and hypotheses

Has supervisory communication measurably changed post-crisis?

- *If so, how do PRA PSM letters differ from FSA ARROW letters?*

Compared to the ARROW letters, we expected the PSM letters to be:

- more <u>complex</u>
- more <u>directive</u>
- more <u>forward-looking</u>

Linguistic features

- **Complexity**

e.g. length of letter, subordinate clauses

- **Sentiment**

e.g. balance of positive versus negative words

- **Directiveness**

e.g. obligative phrases such as *should, must, expect*

- **Formality**

e.g. whether the salutation is handwritten or typed

- **Forward-lookingness**

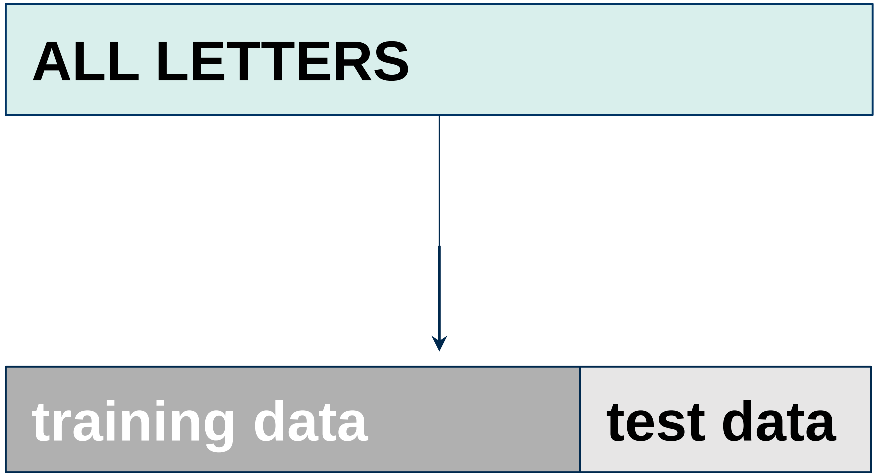
e.g. future-oriented verb tenses

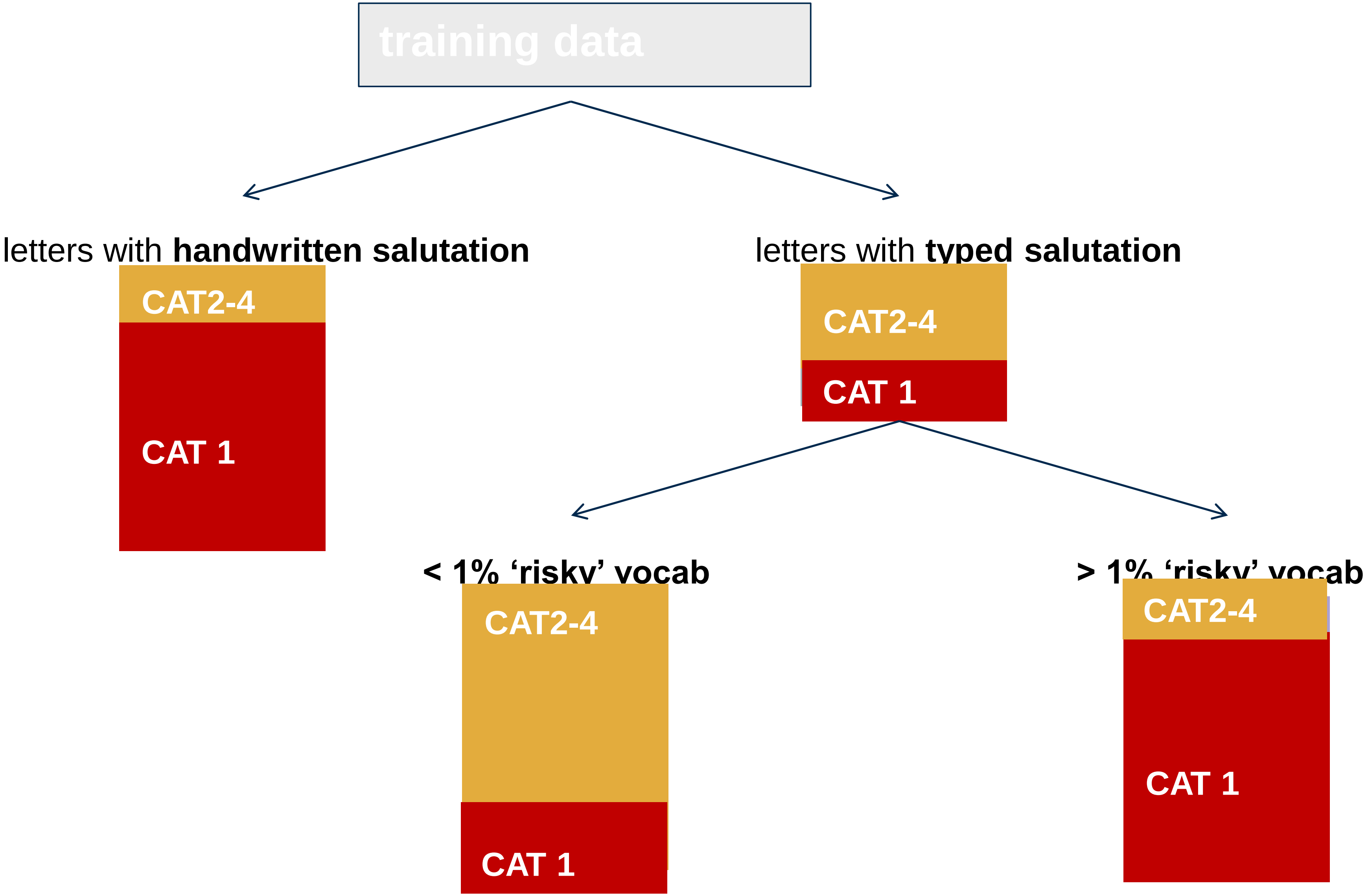
Random Forests

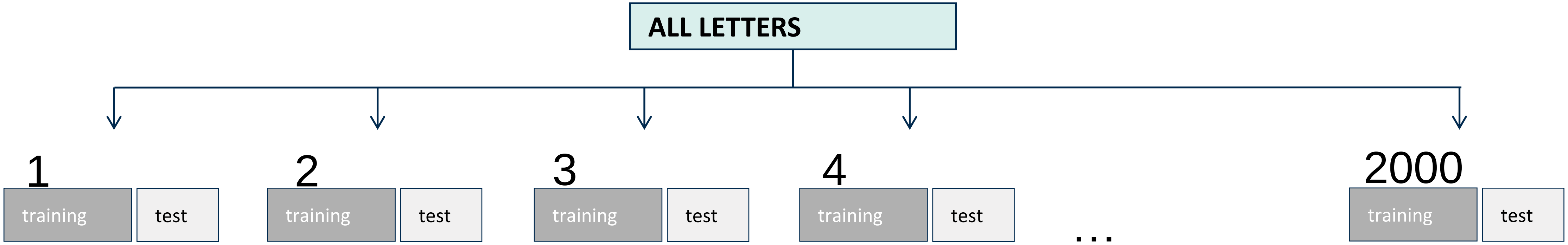
1. Category 1 vs. Category 2-4
2. PIF 1-2 vs. PIF 3-4
3. PSM letter vs. ARROW letter

~ 25 linguistic features

Random Forests




training data
letters with **handwritten salutation**
CAT2-4
CAT 1
letters with **typed salutation**
CAT2-4
CAT 1
< 1% ‘risky’ vocab
CAT2-4
CAT 1
> 1% ‘risky’ vocab
CAT2-4
CAT 1


ALL LETTERS
1
training
test
2
training
test
3
training
test
4
training
test
...
2000
training
test

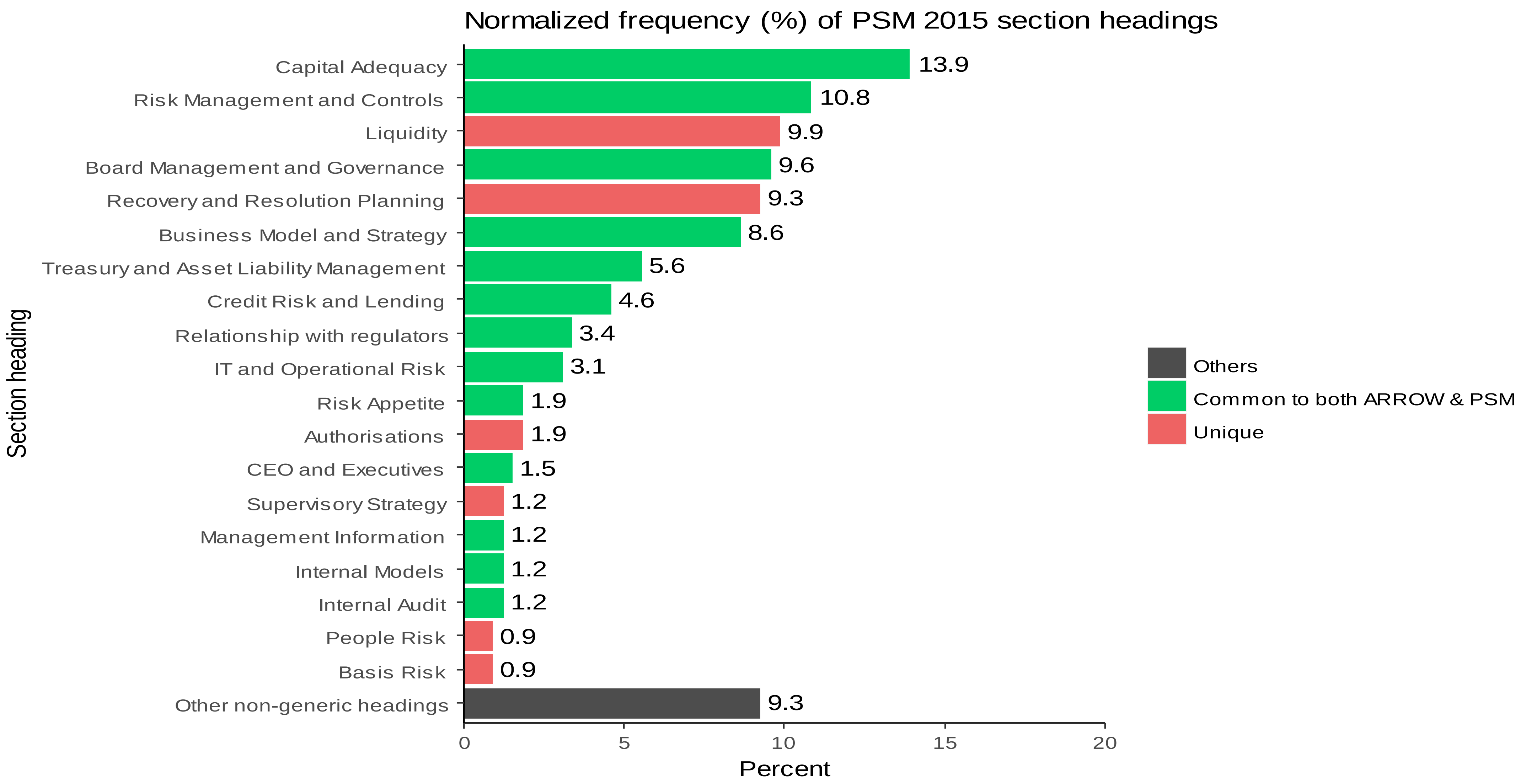
CAT 1 PSM letters different from CAT 2-4 letters

- More complex
- Less directive
- Less formal
- No differences in sentiment

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



Summary

- **Are PSM letters written differently to firms with different risk profiles?**

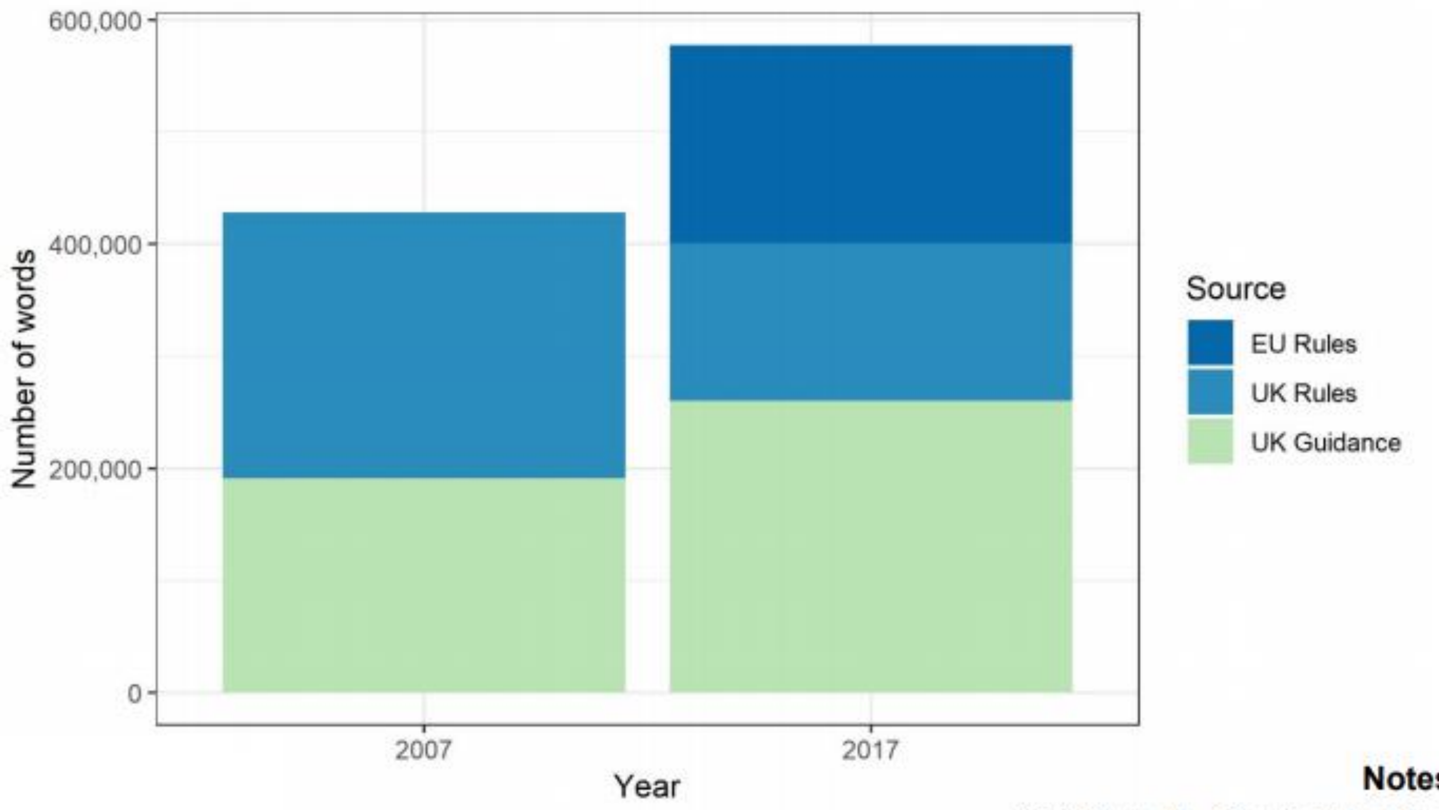
Yes

- **Has supervisory communication measurably changed post-crisis?**

Yes

Textual Complexity in Prudential Regulation

Dataset: Universe of UK banking prudential regulation



Notes on source documents

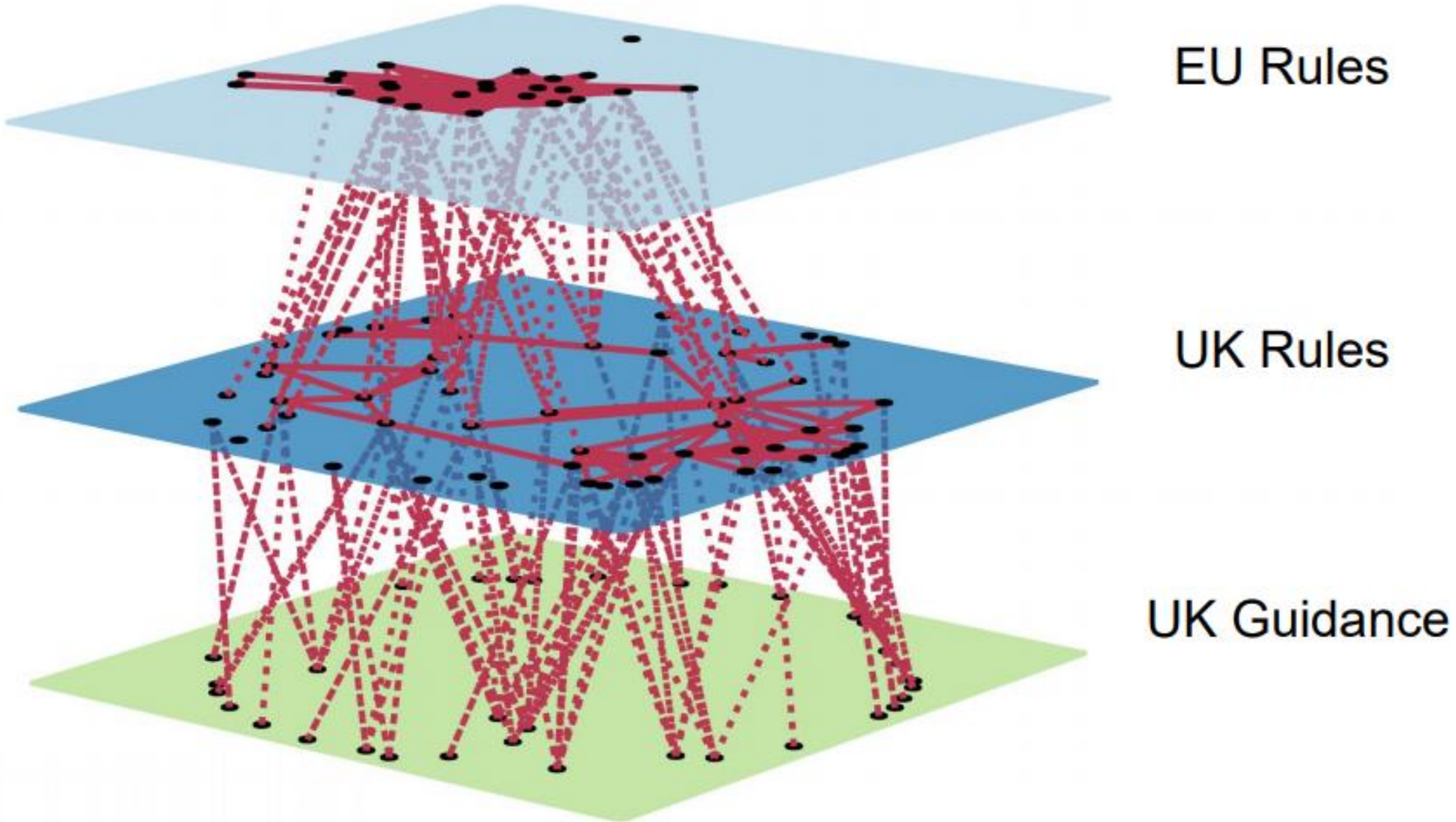
"EU Rules" = Capital Requirements Regulation (2013)

"UK Rules" = PRA Rulebook

"UK Guidance" = PRA Supervisory Statements

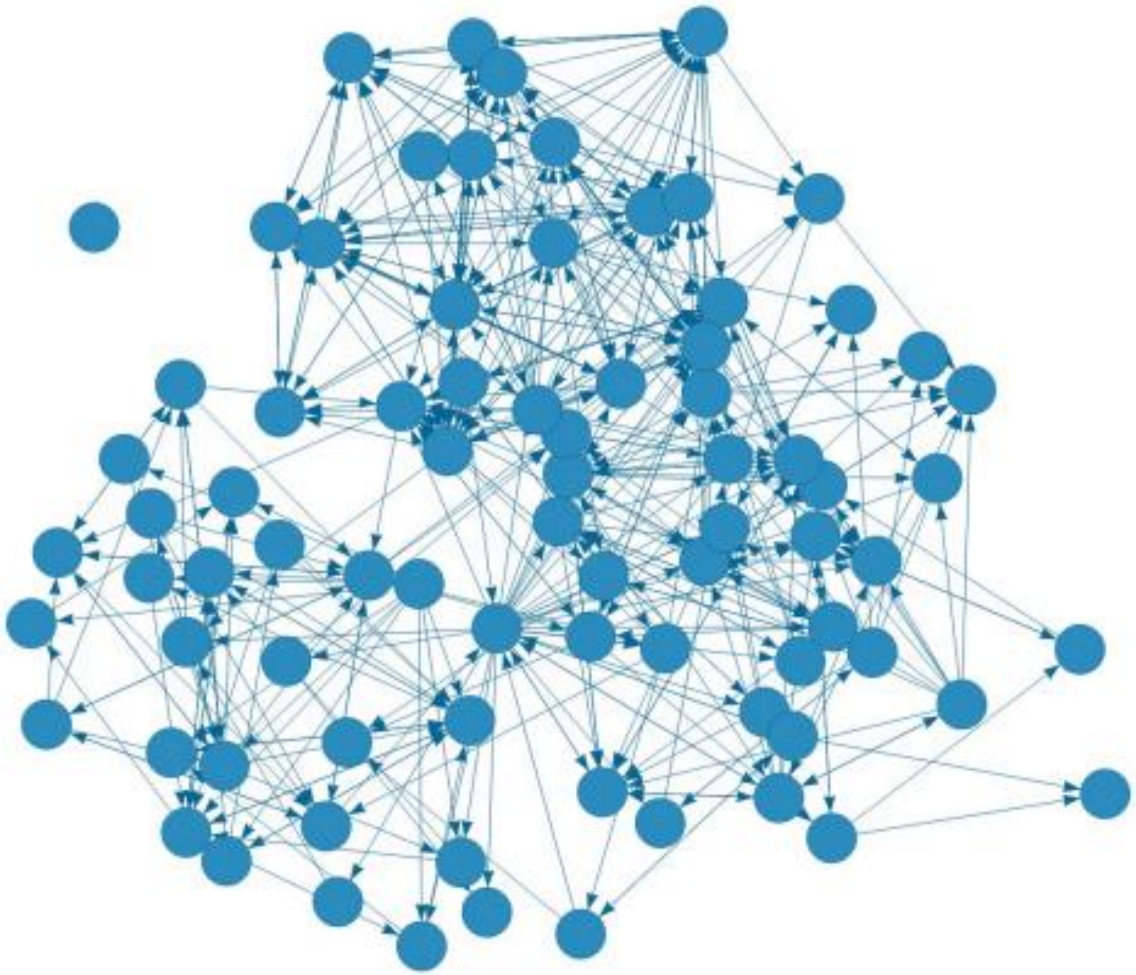
Zahid Amadxarif, James Brookes, Nicola Garbarino, Rajan Patel and Eryk Walczak (forthcoming)

Structure (length isn't everything, part 1)

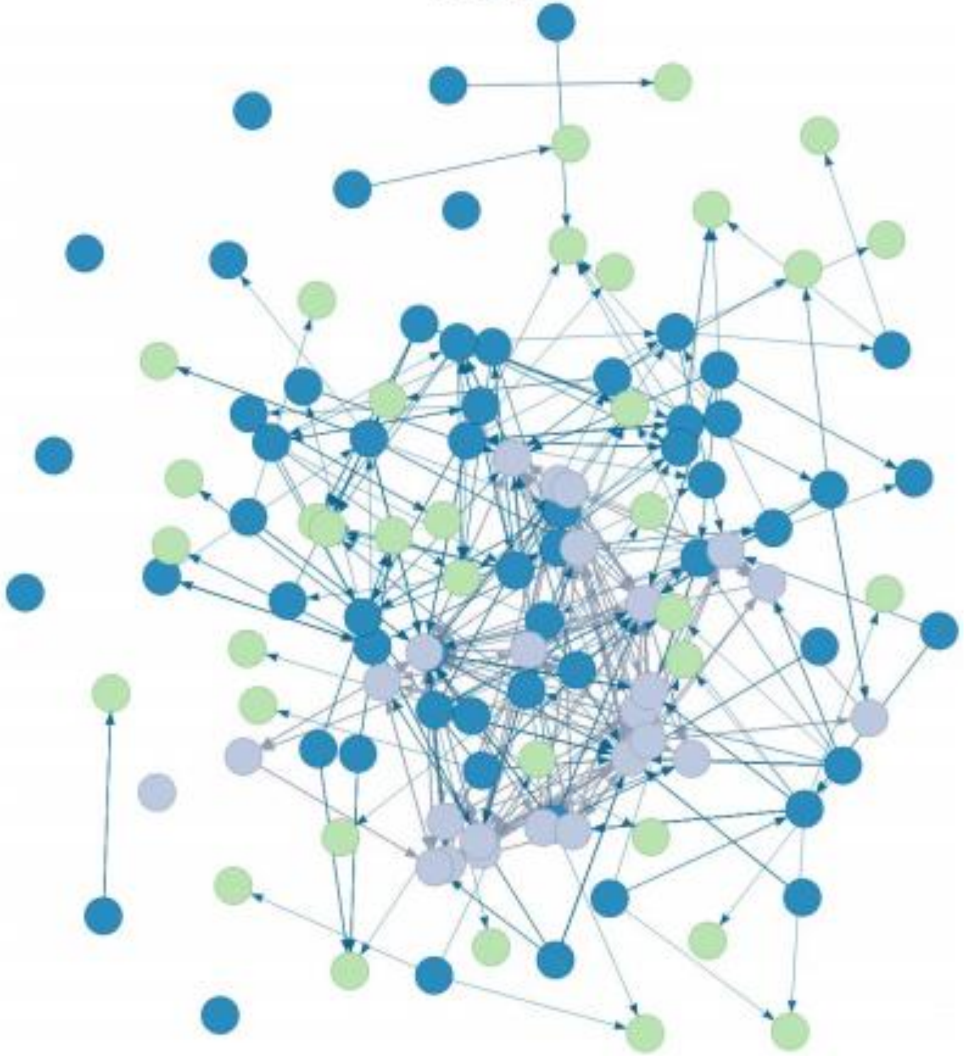


Legal network was reorganised post-crisis…

2007



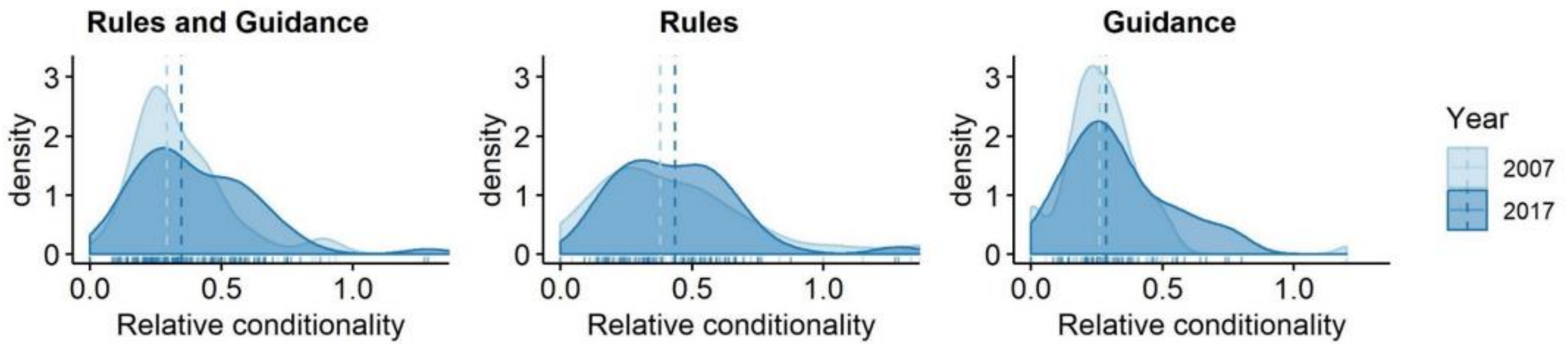
2017



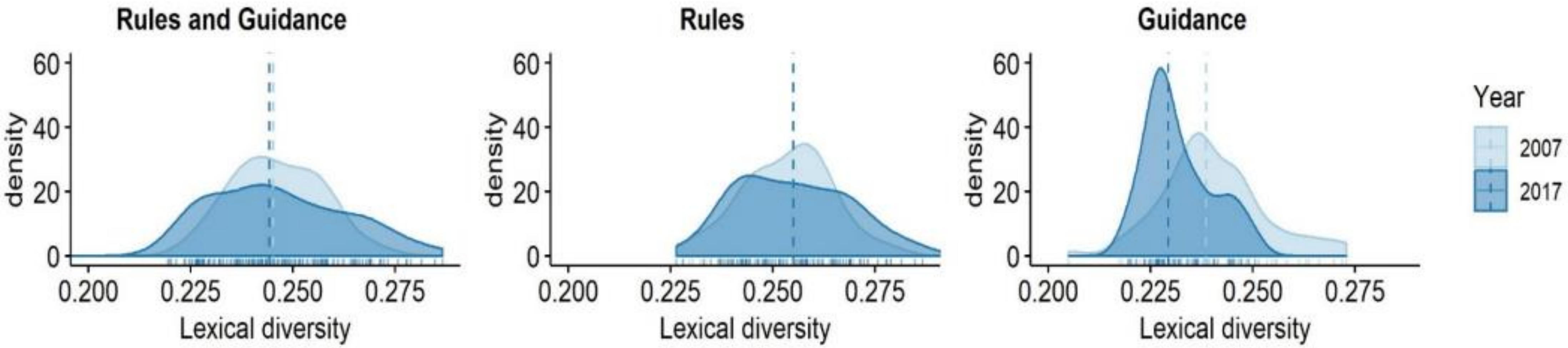
Our suite of linguistic measures

Measure	Calculated as…	Tells you about…
Lexical Diversity	Relative frequency of unique words	Precision (counts concepts)
Conditionality	Relative frequency of conditional statements (e.g. “if”; “but”; …)	# of operative facts (counts exceptions)
Readability	Flesch-Kincaid grade level	Cognitive cost
Length	Number of words	All of the above

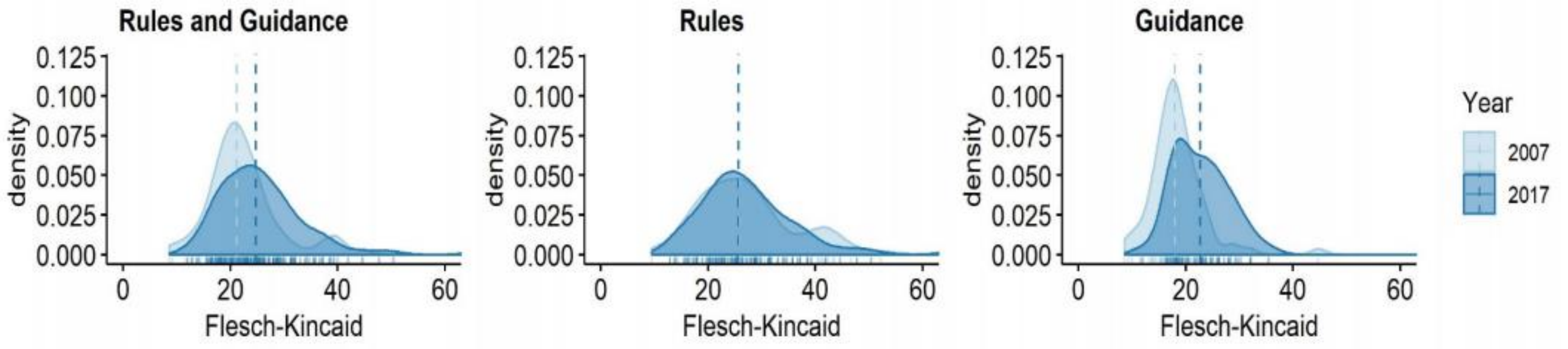
Conditionality has increased: are regulators trying to specify more states of the world?



Lexical diversity is similar: less change in how regulators think about each topic?



The framework is less ‘readable’: has cognitive burden increased?



Summary of results

• Total regulatory text grew by 35% between 2007 and 2017
– Rules and guidance have similar growth rates

• Language has become (on average):
1) more conditional; and
2) less readable. Lexical diversity is similar overall
    – Bigger changes for guidance than for rules

• A 'longer' regulatory network post-crisis.
– Indirect connections have become relatively more important.

Thank you

References

- Amadxarif, Z., Brookes, J., Garbarino, N., Patel, R., Walczak, E. (forthcoming). Textual Complexity in Prudential Regulation. Bank of England Staff Working Paper Series.
- Bholat, D., Brookes, J., Cai, C., Grundy, K., Lund, J. (2017). Sending firm messages: text mining letters from PRA supervisors to banks and building societies they regulate. Bank of England Staff Working Paper No. 688.
- Chakraborty, C. and Joseph, A. (2017). Machine learning at central banks. Bank of England Staff Working Paper No. 674.
- Joseph, A. (2019). Shapley regressions: A universal framework for statistical inference on machine learning models. Bank of England Staff Working Paper No. 784.
- Turrell, A., Thurgood, J., Djumalieva, J., Copple, D., Speigner, B., et al. (2018). Using online job vacancies to understand the UK labour market from the bottom-up. Bank of England Staff Working Paper No. 742.