Data, technology and analytics in competition enforcement:
building a new professional capability and offering

Stefan Hunt

Autoridade da Concorrência, 11th December, 2019

Digitisation had led to many competition issues

- Technology and markets continue to rapidly develop, yielding many benefits
- However concerns regarding digital competition. 2019: UK's "Furman Review", EU's Digital Competition Expert report, Australia's ACCC Report into Digital Platforms, Stigler Centre report…
- Three Furman recommendations particularly relevant to analytics:
  - Digital market unit (Strategic recommendation A)
  - Information gathering powers (Recommended Action 15)
  - How use of machine learning algorithms and AI evolves (Strategic recommendation D)
- Technology concerns broader: Online Harms, news etc

… and also creating opportunities for the public sector

CMA
Competition & Markets Authority

Overview

- Team structure and infrastructure
- What we offer the CMA
- Next steps

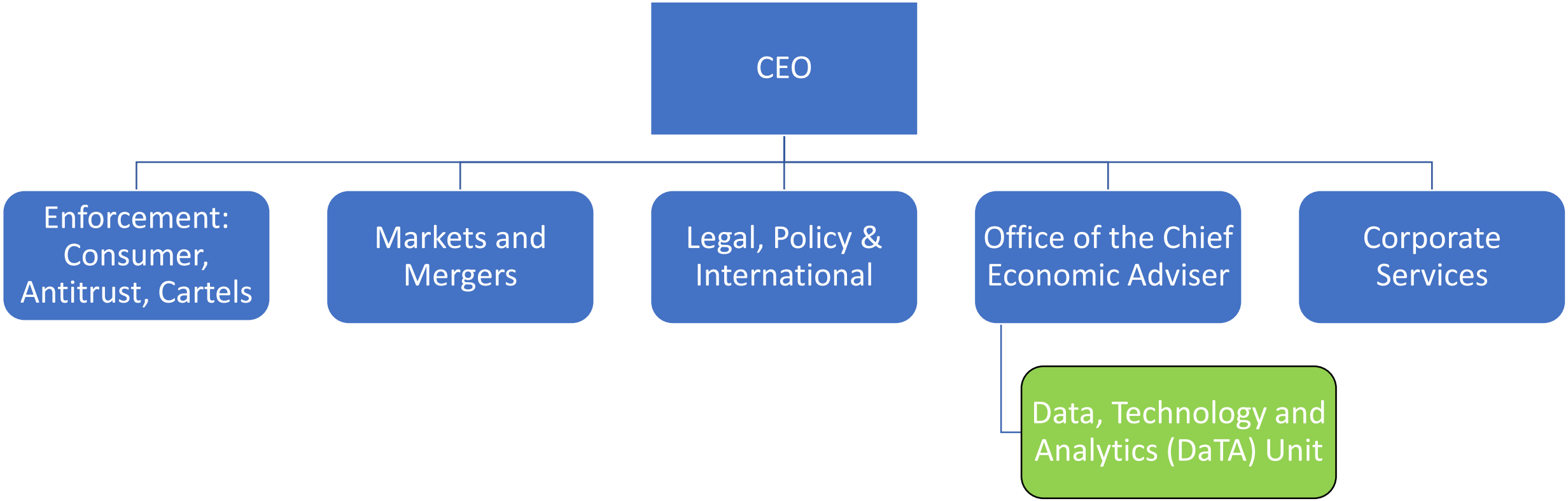
Team structure and infrastructure

A new capability from scratch

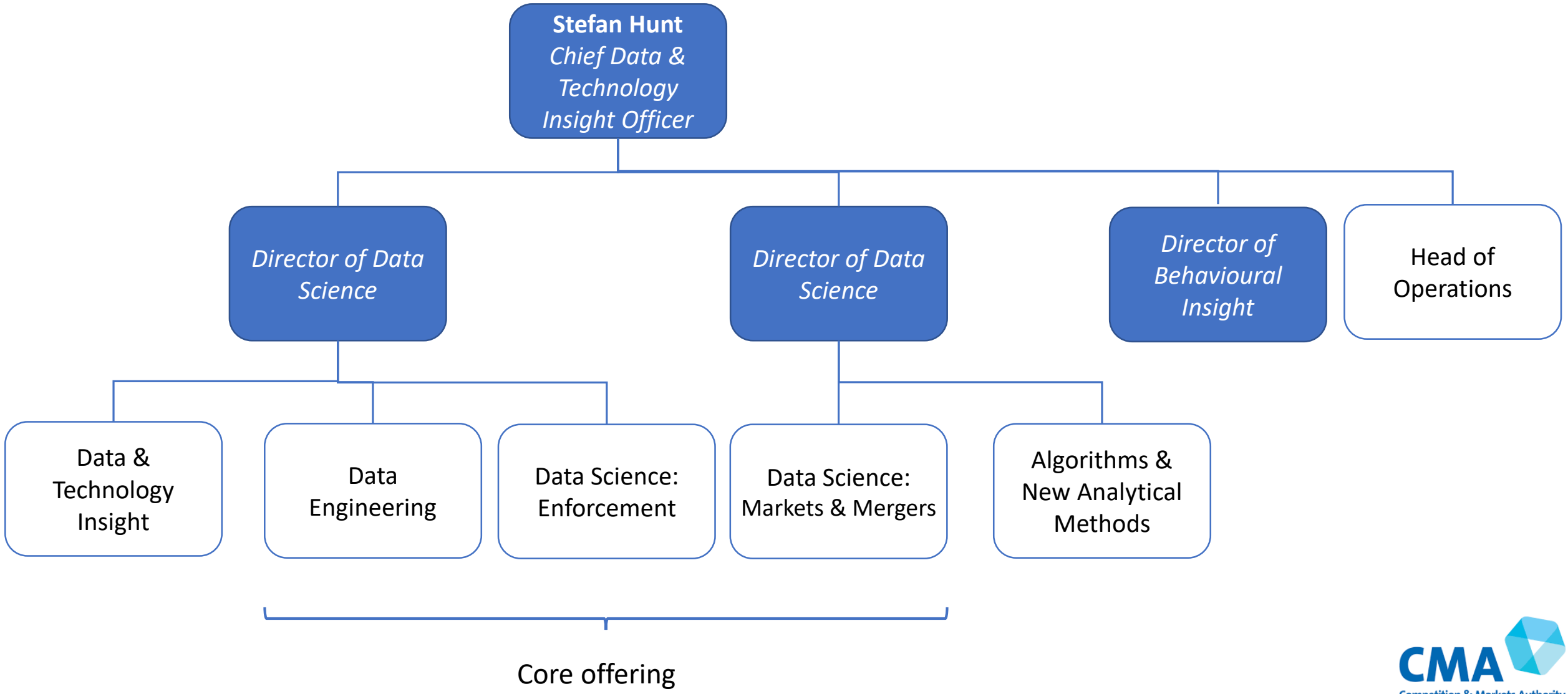
- **Team arrived in February**, 22 now, 30 by early 2020
- **Broad technical and commercial experience**

CMA
Competition & Markets Authority

The CMA


CEO
Enforcement: Consumer, Antitrust, Cartels
Markets and Mergers
Legal, Policy & International
Office of the Chief Economic Adviser
Corporate Services
Data, Technology and Analytics (DaTA) Unit

How the team is structured and key capabilities

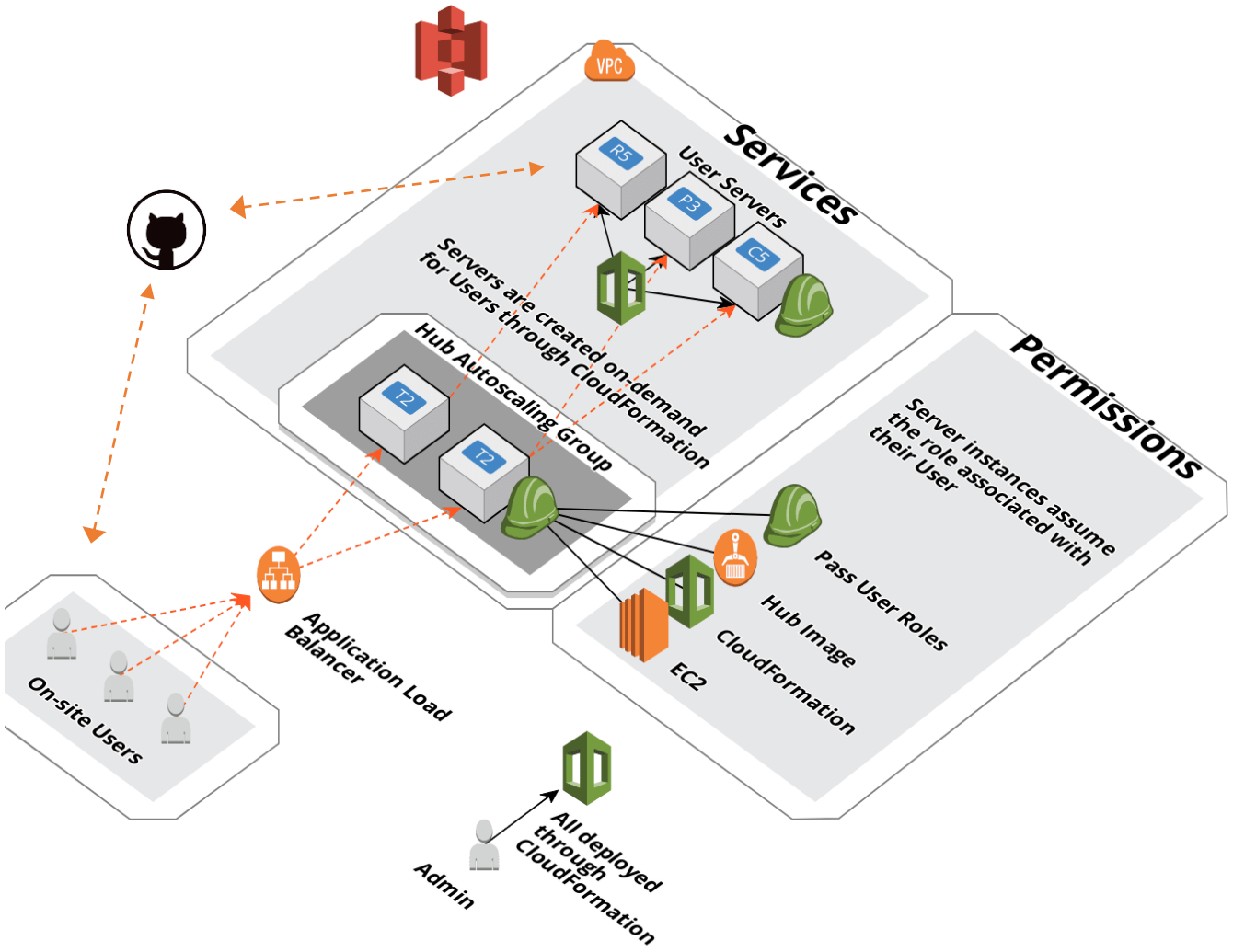


CMA
Competition & Markets Authority

Developing a range of data, tech and algorithmic capabilities

	Firms	CMA
Data (engineering)	• How firms store data, how to request it • Who controls data? Who gets access?	• Defining, collecting, receiving, ingesting, extracting and cleaning • Handling large admin datasets • Maps, web scraping, images, cookies, APIs…
Technology (engineering)	• Data privacy, access, privacy-protecting data matching • APIs, interoperability protocols	• Developing software • Developing necessary pipelines for data, and analytical platforms
Algorithms (data science/ ML/ AI)	• Understand firms' use of algorithms, the impact and any issues • Fairness, transparency and explainability	• Predictive analytics • Descriptive analytics • Natural Language Processing • Visualisation • Coding quality

Setting up our IT Platform



- Developed an easy-to-use “serverless” cloud platform – very low cost
- Can deal with essentially any size of data
- Users access Jupyter and RStudio through a web browser
- Software engineering levels of quality assurance/ version control

CMA
Competition & Markets Authority

The Three Phase Delivery Model

1. Scoping

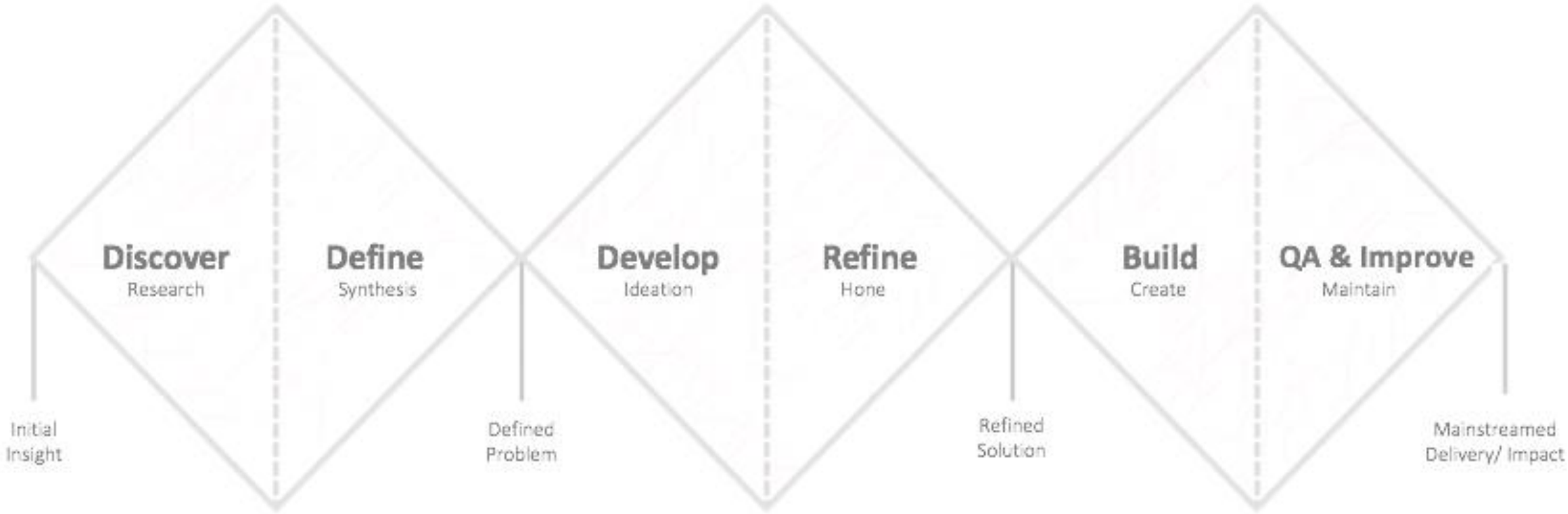
Designing the Right Thing

2. Develop & Refine

Creating the Thing Right

3. Product Delivery

Implementing the Thing



What we offer the CMA

Data science and engineering, five offerings:

1. Building software tools, to tackle repetitive or slow tasks

2. Data gathering and manipulation, at scale and pace

3. Analysis and insight, using data science

4. Understanding and explaining technology

5. Analysing algorithms, and how they are used

CMA
Competition & Markets Authority

1. Building tools: local competition assessment

Competition authorities often analyse locations of competitors

euro CAR PARTS

Andrew Page

JD

GO Outdoors

CMA
Competition & Markets Authority

This analysis can be particularly cumbersome

- Many steps in local assessment for mergers
  - Step 1: Get postcodes for all locations of merging parties + competitors.
  - Step 2: Geocode to convert to lat/lon coordinates.
  - Step 3: Generate a matrix of drive-times.
  - Step 4: Define geographic market.
  - Step 5: Find overlaps.
  - Step 6: Generate candidate SLC areas.
  - Step 7: Make decisions

- Takes long time, and uses valuable economist resource and software

ArcGIS

STATA®

Excel

CMA
Competition & Markets Authority

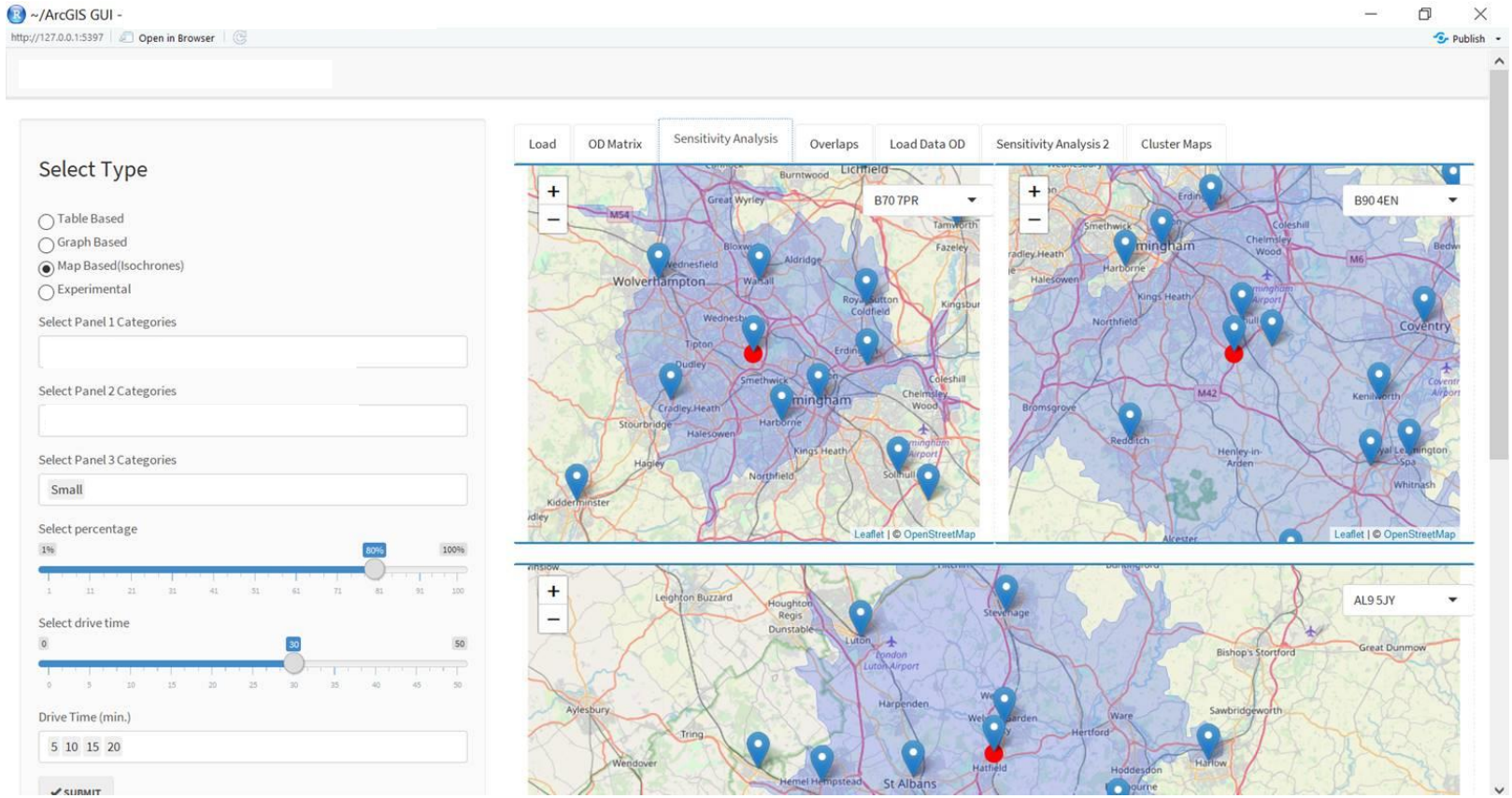
We built a tool that is more than 90% quicker

	Current	NEW tool	Saving
Step 2: Geocoding 1,000 postcode	5.0-6.0s	0.5s	92% less time
Step 3: OD matrix 300 x 200 locations, cut-off 12km	8.0-13.0m	0.5m	94% less time

- On large datasets, some tasks could take 10+ hours
- Economists needed to constantly monitor
- Cost not one-off, as Parties frequently update list of locations

CMA
Competition & Markets Authority

And has a user-friendly, bespoke-designed interface



- Generate isochrones and sensitivities
- Even more quality control
- Piloted
- Released Nov 19

… and many other opportunities for new products

CMA
Competition & Markets Authority

2. Data gathering and manipulation: Funerals

Markets required a definitive list of funeral directors in UK

- More than 7k funeral directors and colleagues were trying to cleanse and merge seven different lists
- We created a rigorous method for de-duplication:
  - Text pre-processing
  - Simplifying the problem: identify potential duplicates
  - Creating matches, using fuzzy matching on individual fields e.g. branch name, company name and address fields
  - Determining which records are duplicates

CMA
Competition & Markets Authority

Analytical methods

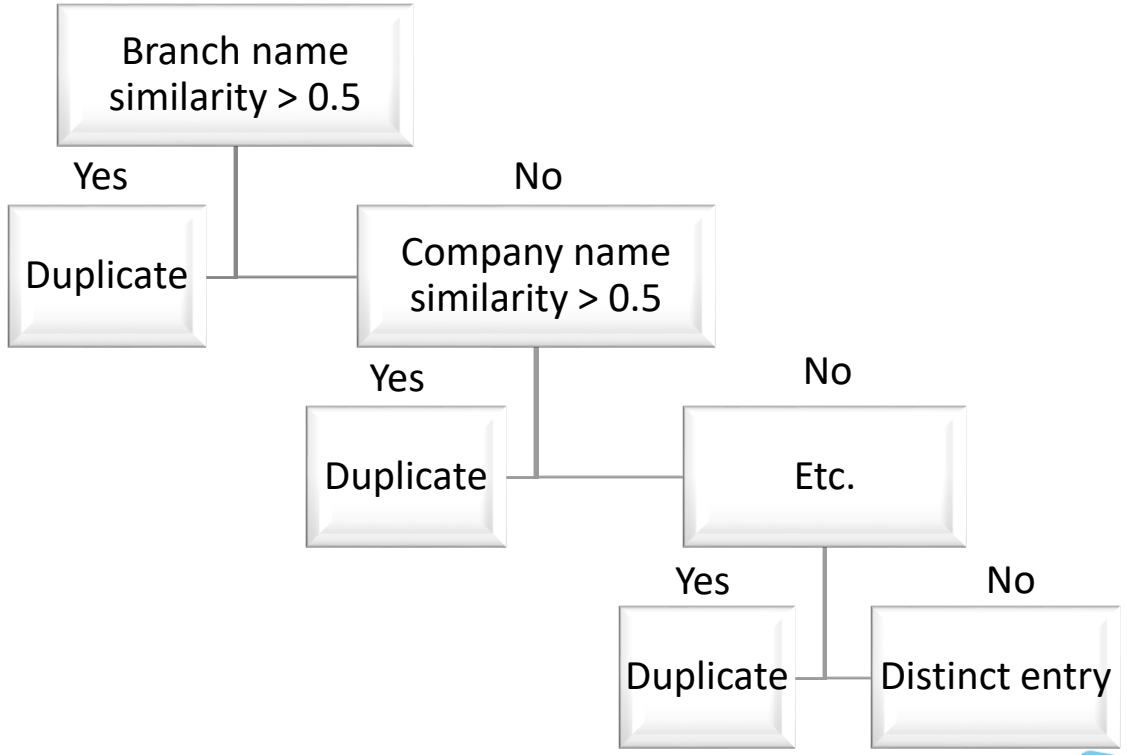
Fuzzy Matching of Text Fields

Similarity values are calculated for the branch names, company names, and other data such as addresses.

	Definition	**Example: Adam vs Alan**
Levenshtein edit distance	the number of substitutions between two strings (the higher the value, the less similar the strings are)	2
Levenshtein similarity	1 - edit distance normalised by the length of the strings (the higher the value, the more similar the strings are)	1-2/4 = 0.5

Classification: Decision Tree

Records are fed through a simple hierarchical decision tree to identify duplicates.



CMA
Competition & Markets Authority

We now have a reusable asset for other de-duplications

- Uses a well-documented and easy-to-quality-assure process
- Econometricians saved from low value data cleansing and processing
- Process and code can be applied to future deduplication and cleansing

- We are building many re-usable pieces of code to make projects more efficient and effective

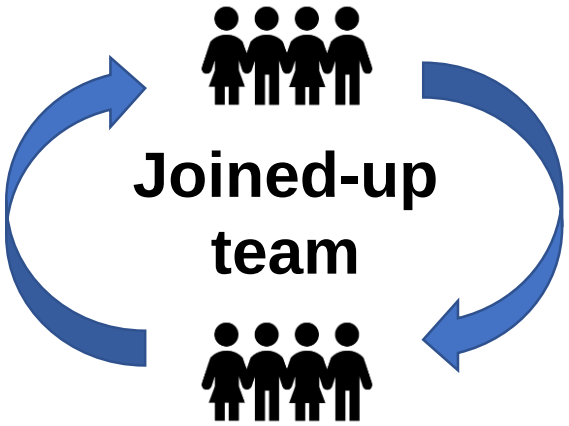
… more generally, great potential for supporting use of information gathering powers

CMA
Competition & Markets Authority

2. Data gathering and manipulation: E-discovery

We aim to make the mergers document review more efficient

Requesting → **Ingesting and processing** → **Reviewing** → Presenting outputs



- **Case officers**: provide domain expertise
- **DaTA unit**: investigate technology-assisted review

CMA Competition & Markets Authority

Three example areas where data science can add value

Ingesting and **processing**

- Taking in documents efficiently
- Summarising documents
- Tagging documents / Extracting topics

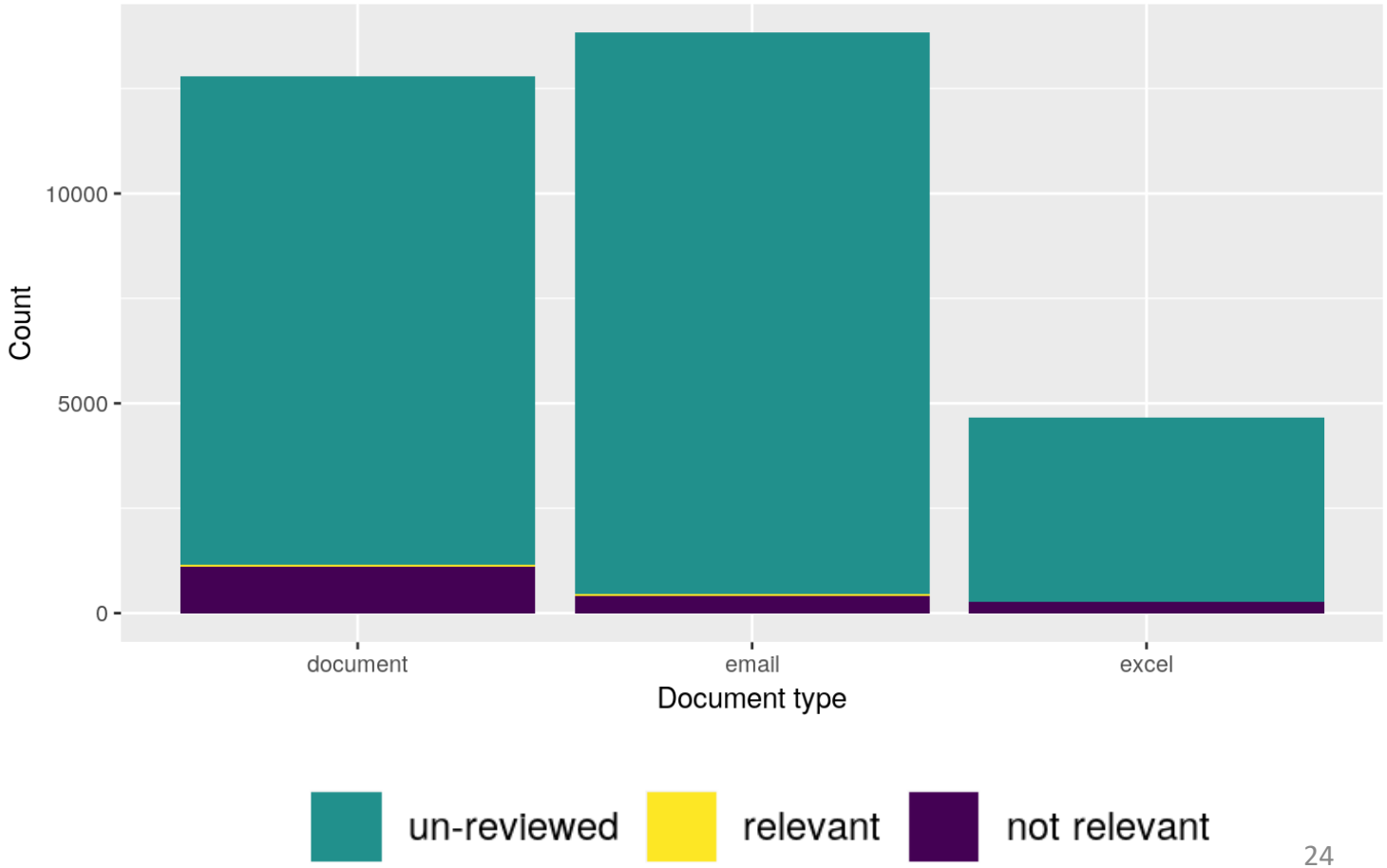
Reviewing

- Prioritising documents according to their relevance

CMA
Competition & Markets Authority

Identifying relevant documents is time-consuming

- We receive a very large number of documents in some cases
- Limited resources mean many are not reviewed
- Proportion of relevant documents tends to be small
- Objective: make the review more efficient
  - Saving time
  - Reviewing more documents



CMA
Competition & Markets Authority

There are several challenges to deploying algorithms

- Using text as feature for a machine learning model raises challenges that a human reviewer would not encounter

1. Noisy data
2. Duplicated content

→ 3. Inconsistent ground truth (from duplicated content and email threads)

- Lots of time spent correcting data errors, identifying and removing duplicates, pre-processing email chains.

CMA
Competition & Markets Authority

3. Analysis & Insight: Document review for mergers

We are testing open source algorithms for summarisation

- Reduce time needed to parse documents during first review. Decide whether document should be read in more detail
- Unsupervised, i.e. we don’t need training data from past cases
- Very few parameters needed = less fine tuning to make
- Does both keyword and sentence highlighting
- Treats each document as stand-alone (no need to parse whole case corpus before producing summaries)

CMA
Competition & Markets Authority

What does the TextRank algorithm do?

The CMA has found that JD Sports' takeover of close competitor Footasylum could be bad for shoppers.

The Competition and Markets Authority (CMA) is concerned that the loss of competition brought about by the merger could result in a worse deal for customers, both in-store and online, through higher prices, worse choice in stores or reductions in service quality. JD Sports must now address the concerns identified or face a further, more in-depth, investigation.

JD Sports and Footasylum both sell sports-inspired casual clothing and footwear in stores across the UK, and through their apps and websites. JD Sports agreed to buy Footasylum in a £90 million deal announced earlier this year.

In 2018, UK consumers spent more than £5 billion on sports clothing and footwear. Sportswear is currently a significant influence on fashion trends, with sports styles replacing traditional casualwear, particularly among younger shoppers. Retailers carefully curate the selection of brands that they offer, and develop attention-grabbing offerings in-store and online, in order to compete for customers.

With over 400 stores, JD Sports is well-established as the leading UK retailer of sports fashion footwear and clothing. It already owns several well-known sports fashion brands on the UK high street – such as Size?, Scotts, Tessuti and Footpatrol, in addition to its signature JD brand – and generated revenues of over £2.14 billion in 2018 in the UK.

Since opening its first store in 2006, Footasylum has experienced strong market share growth and now operates around 70 stores across the UK. Footasylum generated revenues of close to £200 million in 2018.

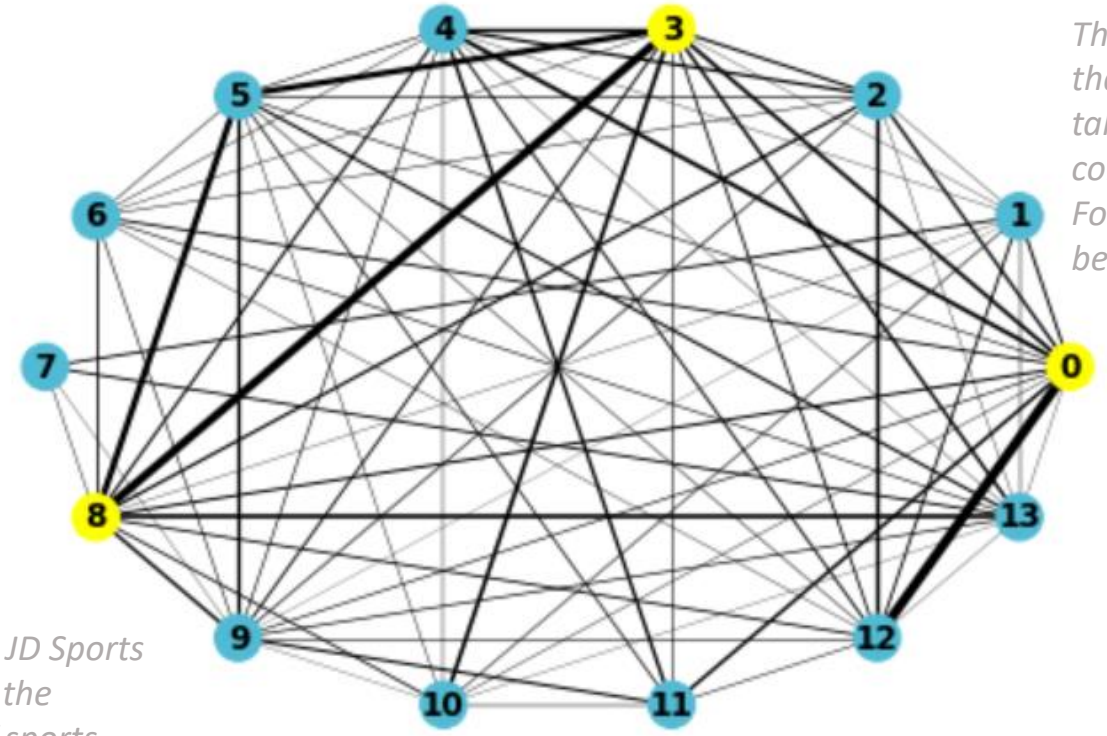
The CMA's initial, Phase 1, investigation has found that the merger could remove one of JD Sports' closest competitors. While a wide variety of retailers sell sports clothing and footwear, the merging businesses are 2 of a smaller number of firms who have the brand relationships and market presence to be able to credibly meet the demands of sports fashion customers.

CMA
Competition & Markets Authority

TextRank applied to sentences

- TextRank is derived from Google's PageRank algorithm
- View text as a graph, where each sentence is a node
- Compute similarity between all sentence pairs
- Sentences that are most central to the network (document) tend to be similar to several other sentences

JD Sports and Footasylum both sell sports-inspired casual clothing and footwear in stores across the UK, and through their apps and websites



The CMA has found that JD Sports' takeover of close competitor Footasylum could be bad for shoppers.

With over 400 stores, JD Sports is well-established as the leading UK retailer of sports fashion footwear and clothing

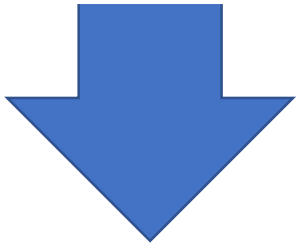
CMA
Competition & Markets Authority

We are evaluating different approaches to predictive coding

- No hard classification into **“read”** vs. **“do not read”** groups
- Prioritisation of documents to focus resources, based on a **relevance metric**
- Relative rather than absolute
- Which documents are to be reviewed remains at the case team’s discretion

Document ID	Type	Relevance
15234	Email	0.391
54231	Email	0.376
65352	Document	0.281
65797	Email	0.271
54235	Document	0.241

Documents ranked in decreasing order – actual value not important
(and could be hidden)



CMA
Competition & Markets Authority

We need to augment our document review process

- Validating the results using other more recent cases. Scope for testing alternative algorithms and tuning models

- Developed **in-depth understanding of the issues and the data.** Promising early results and we've learnt a lot but..

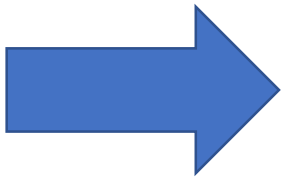
  - Labels are not high enough quality
    - Whole document review process could be overhauled: working with case teams to standardise tagging

  - Merger cases may be too unique to use pre-trained algorithm to identity relevant documents
    - Looking into automated tagging / real-time learning

CMA
Competition & Markets Authority

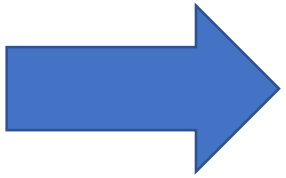
3. Analysis & Insight: Online reviews

DaTA unit complementary to existing case team



Marketplaces

Sites where there has been buying and selling fake reviews

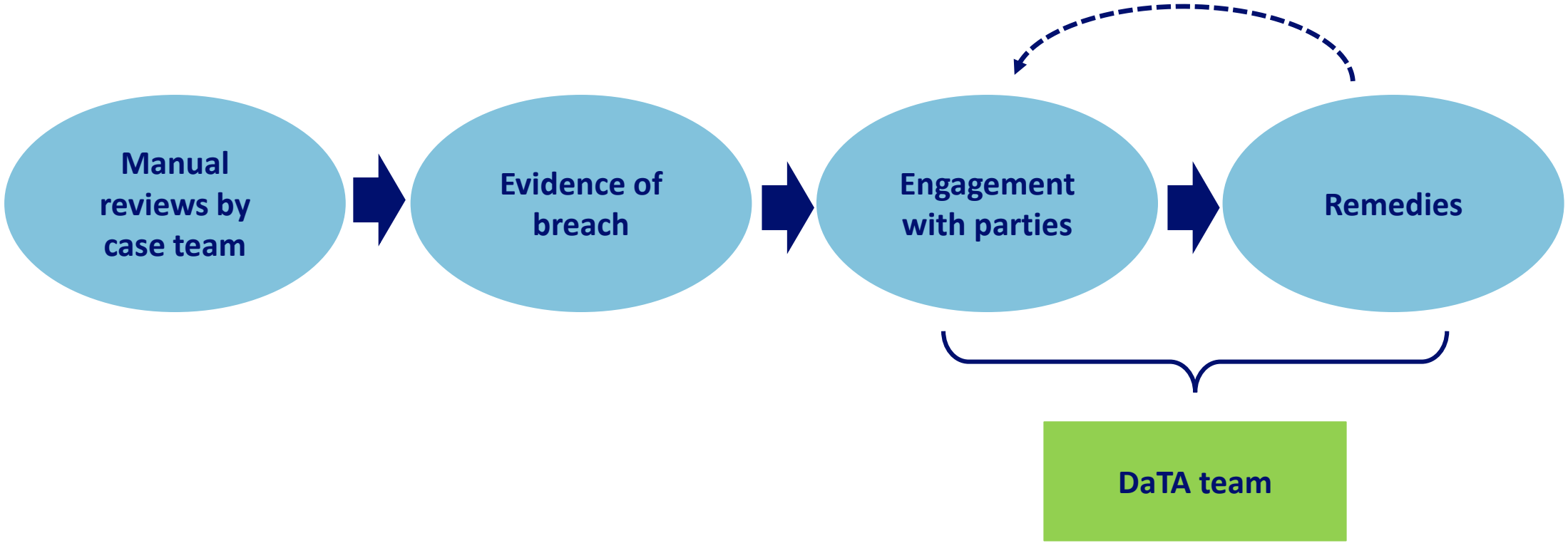


Review sites

Review sites host reviews either as core activity or ancillary to retail

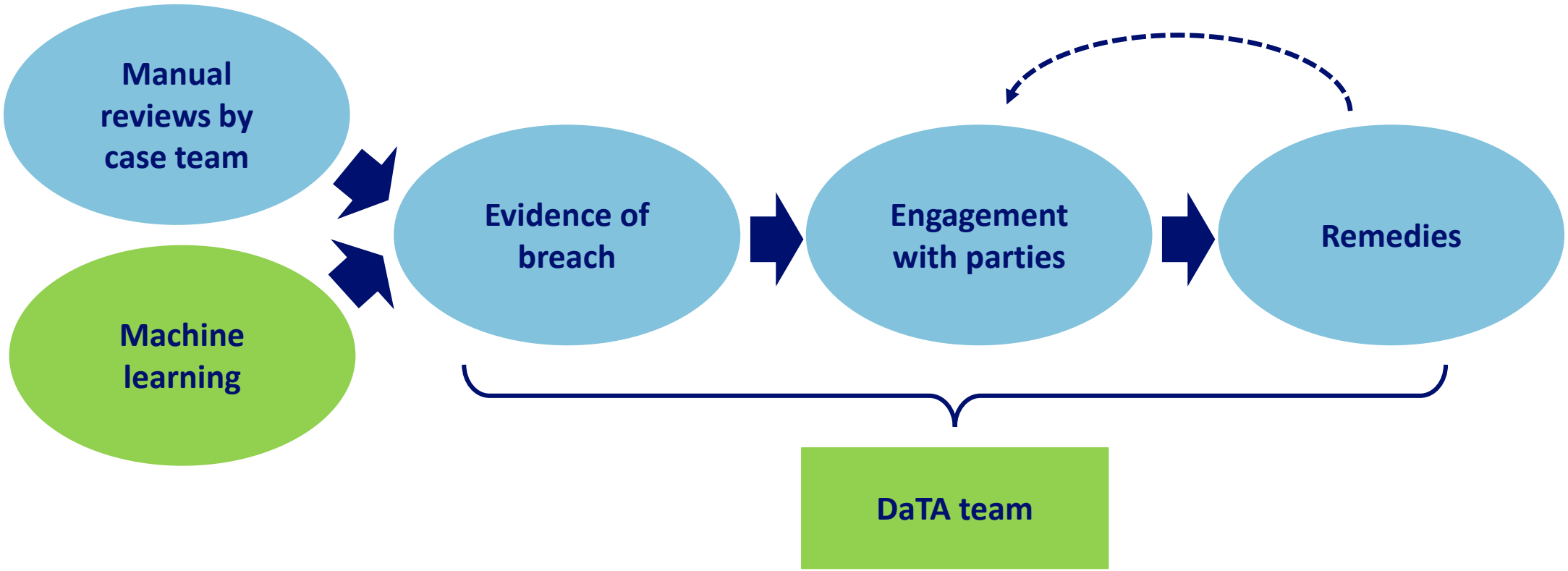
Supporting marketplace engagement

Sites used to trade reviews, either by creating private groups or by holding direct sales and auctions



Review sites: building up evidence

Sites that host reviews either as core activity or ancillary to retail



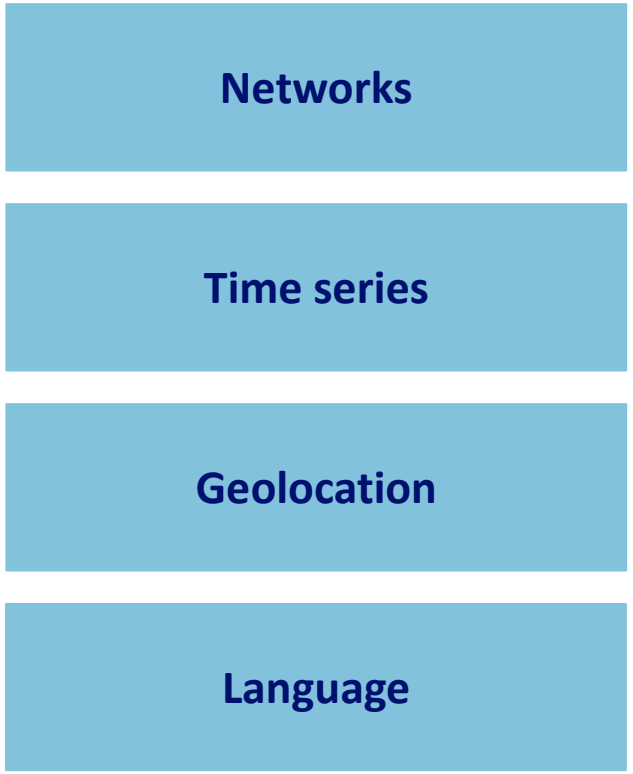
Finding fake reviews is hard

- Detecting each single fake review is hard
- Lack of reliable labelled datasets to learn from
- Unsupervised machine learning: a single technique may not provide strong enough evidence
- Fake vs. incentivised reviews

CMA
Competition & Markets Authority

Focus on detecting signs of irregular activity

- Experimented with four unsupervised techniques
- Layered results from different approaches to provide stronger evidence
- Build up evidence, but also…
- …understand problem better, the challenges and possible approaches, to engage with parties more effectively



CMA
Competition & Markets Authority

Web-scraping enables scaling up of evidence gathering

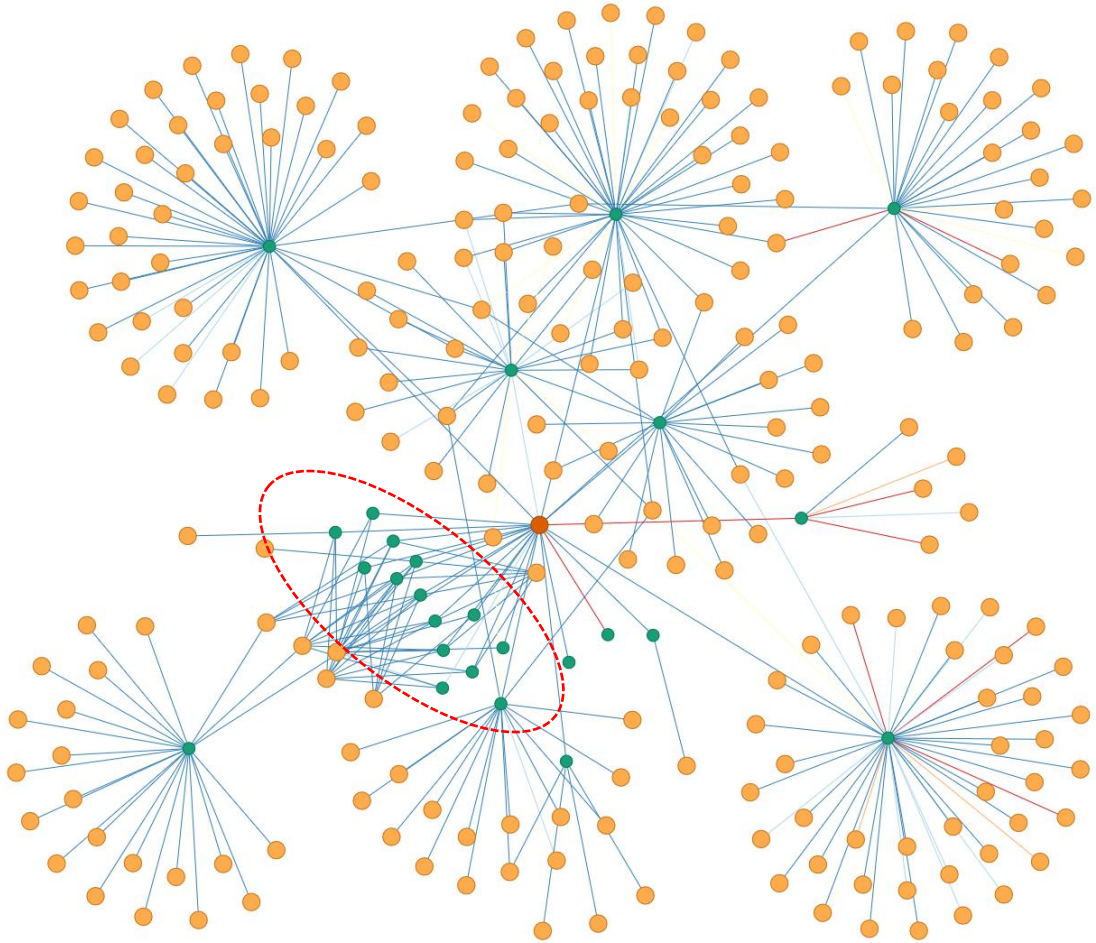
- Consumer team identified methods to spot fake reviews, manually recording information in Excel

- DaTA unit
  - automates data gathering
  - helps create stronger evidence by gathering much more data (c. 100,000x)

CMA
Competition & Markets Authority

Network analysis

Start with a suspicious seed company at centre



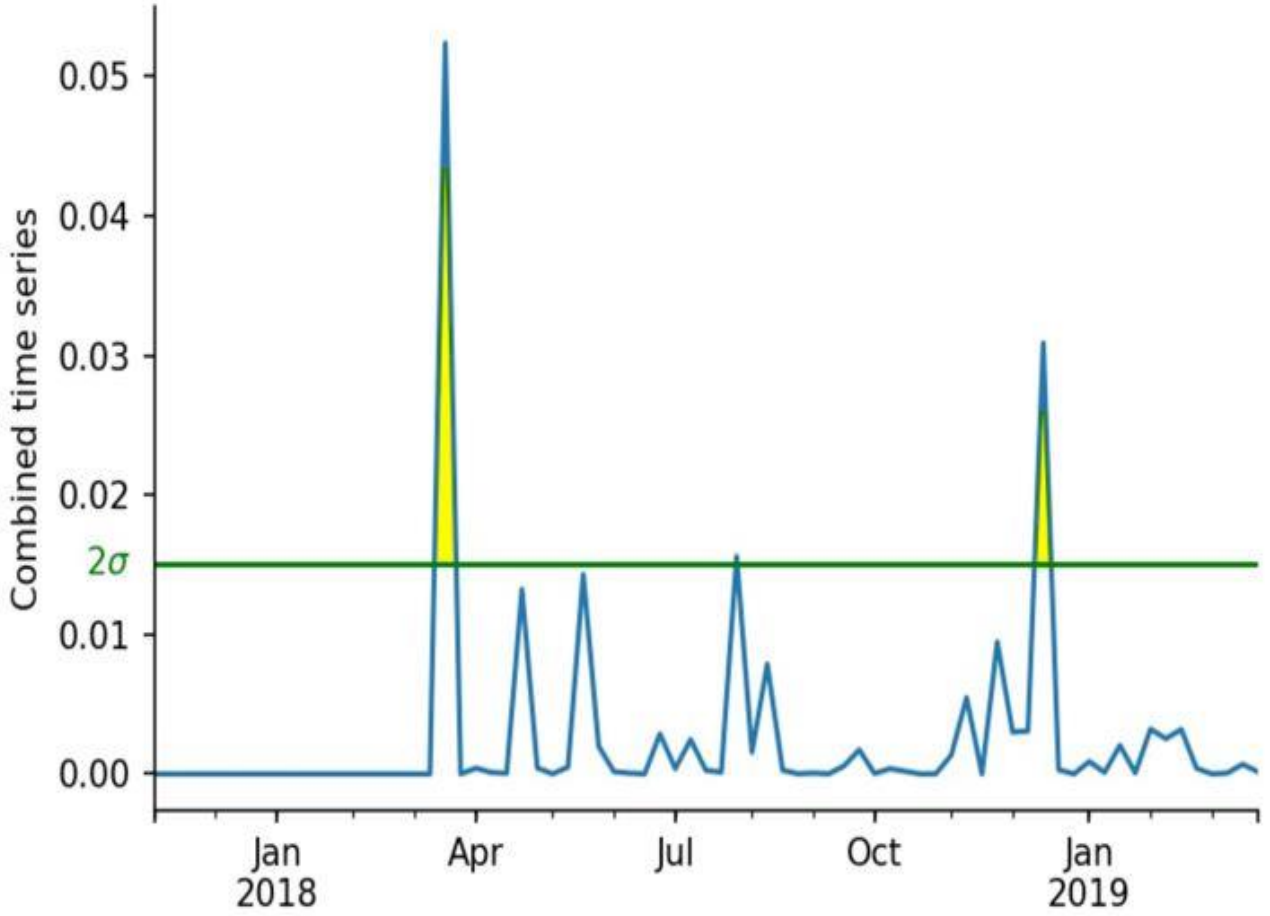
Automatically scrape Reviewers of the seed company and all the Companies reviewed by those reviewers

Network analysis to find clusters of reviewers that have all reviewed the same businesses

CMA
Competition & Markets Authority

Time series

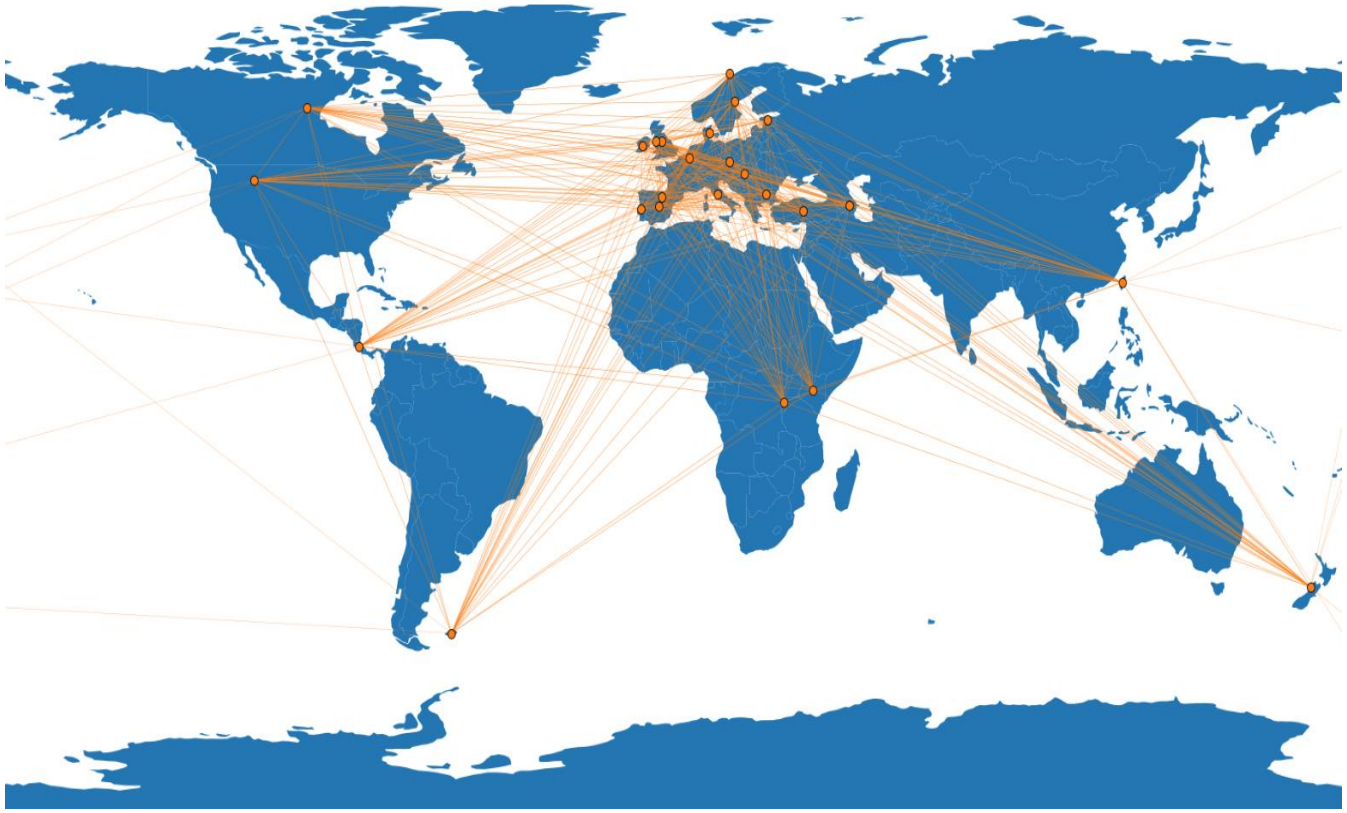
- Look for bursts in posting
- Fake review ‘campaigns’ might be concentrated in time
- **Possible metric**: index taking into account the number of bursts above a certain magnitude, their actual level as well as their duration
- **Caveat:** ‘spammers’ can drip-feed to avoid detection



CMA
Competition & Markets Authority

Geolocation

- Fake reviews obtained from sellers all over world
- **Possible metric:** index capturing geographic spread of reviewers for each company
- **Caveats:**
  - IP address / location spoofing
  - IP address not often available



CMA
Competition & Markets Authority

Language

- Look for repetitive, not original language
- Consider overall ‘same-ness’ of reviews
- **Possible metric:** index capturing proportion of reviews with similarity over certain threshold
- **Caveats:**
  - Measuring meaning vs. formal similarity
  - Sophisticated spammer mimicking original language

CMA
Competition & Markets Authority

4. Understanding technology: Mergers

Specialists in the unit are providing insight on tech mergers

- Members of team have practical experience of these markets and using the tools
- Working alongside Mergers teams to provide specialist insight to live cases
- Attending meetings with parties and providing advice to panel members

… and more in our Consumer and Markets work

CMA Competition & Markets Authority

5. Analysing algorithms: towards an agenda

Analysing algorithms: pipeline proposals

- Cases featuring algorithms will require different analysis
- Use of algorithms becoming prevalent, with a variety of potential issues: so expect these cases to be a significant feature in future
- Many options for analysis and new skills needed. Aim to build up institutional knowledge, understanding and capability

Potential issues

- **Collusive behaviour:** pricing algorithms and intermediaries selling them
- **Choice architecture:** ranking and listing algorithms and their influence on choice
- **Exacerbation of "addictive" behaviours**
- **Discrimination:** what is shown to whom

CMA
Competition & Markets Authority

Next steps

We are adding new capabilities to the unit

Data Science & Engineering

We improve cases by creating better evidence and providing confidence to challenge firms

1

Data & Tech Insight

Through our research and ecosystem we deliver digital market insights and translate these into competition – related policy and case understanding

2

New Analytical Methods & Tools

We work with colleagues across the CMA to develop innovative analytical methods and tools that can be used on cases.

3

Behavioural Insight

Our behavioural science hub will work across the CMA to offer insight on remedy design and consumer engagement.

CMA
Competition & Markets Authority

Conclusions

- Huge opportunity – especially in context of recent reviews – to use more advanced data, technology and analytics skills in competition enforcement
- The exact balance of tools needed depends on the institution's size, remit, sources of data and IT estate
- Grow advanced analytics capability iteratively using "agile" processes

- Great opportunity for domestic and international public bodies to work together, e.g. learn from each other, share code, or resources

CMA
Competition & Markets Authority

Thank you

CMA
Competition & Markets Authority