The technology-led transformation of competition and consumer agencies:

The Competition and Markets Authority's experience

Discussion paper

Stefan Hunt*

14 June 2022

Abstract

Over the course of the last decade, the innovation and proliferation of digital technologies radically transformed existing markets and gave birth to new ones. Alongside this change the skills that firms need to compete shifted. Competition and consumer agencies similarly have expanded their capabilities to meet the regulatory demands of dynamic, digital markets. 'Technologists' now work at most major agencies, on cases involving data and digital markets and on innovating agency processes; and most of these agencies have reorganised to have a dedicated specialist branch or unit. Bill Kovacic, former Federal Trade Commission Chair, has referred to this technology-led transformation as 'one of the top five developments in competition over the last three decades'. The Competition and Markets Authority (CMA) has been an early adopter and its Data, Technology and Analytics unit now has almost 50 people – and growing – across the disciplines of data science, engineering, technology insight, behavioural science, eDiscovery and digital forensics. This paper highlights how the unit has brought value to cases, and it draws lessons for how similar units can succeed. It describes five roles that technologists play in competition and consumer agencies and explains how their work contributes to cases, drawing on concrete examples primarily from the CMA's work. It also outlines the main issues and design choices when founding a data and technology unit in an agency. The paper concludes by considering how competition and consumer work will be affected, including a potential shift from the agendas of agencies being mostly reactive and set by complaints, to being proactive and increasingly set by pre-emptive data gathering and monitoring.

* Chief Data and Technology Insight Officer, Competition and Markets Authority, stefan.hunt@cma.gov.uk. The views expressed in this paper are my own and not those of the CMA. I thank Tim Capel, David Dorrell, Matthew Hunt, Andreas Jespersen, George Lusty, Tammy Masterson, Joe Perkins, Laura Smart, Ollie Sugg, Chris Tynan, Lindsay Taylor, Gordon Wai, Mike Walker and additional CMA colleagues for their helpful comments. All errors are my own.

Contents

1. Introduction and summary

Over the past few years, the global competition and consumer protection community has increasingly focused on issues in digital markets. Each week now brings major developments in antitrust cases or the progress of landmark pieces of legislation aimed at regulating 'big tech'. The underlying forces driving the changes in markets are technological. The prevalence of digital hardware and the digitally integrated nature of modern life now means data is being collected at both a rate and scale unlike that of a decade or two ago. The use of massive data sets, complex machine learning and artificial intelligence (AI) algorithms, user experience testing, and a raft of other technologies has caused the information asymmetry between firms and agencies to grow.

In the face of such change, agencies must bring their skills up to date for two purposes. First, they need to deal with challenges in digital markets. To detect, understand and remedy these challenges effectively, they need staff members that know, for example, how to audit data flows in organisations, how algorithms are developed, how privacy enhancing technologies can be used in data sharing and interoperability remedies, and much more. Second, agencies can use technology to radically improve their operational performance, so-called digital transformation. They can develop their capability to monitor markets, detect cartels, process and search documents, disclose while ensuring confidentiality, and more. I call these two purposes together technology-led transformation.

Digital transformation has been a hot topic in industry for 10 to 15 years.[1] In the UK, sectoral regulators, who have much more data than competition agencies, began embracing similar innovation four to five years ago.[2] Competition and consumer agencies are beginning to embrace technology-led transformation.

In the last two to three years, many of the larger agencies have set up dedicated specialist teams and started hiring technologists – an umbrella term to cover a range of data and technology skills. The Competition and Markets Authority (CMA) launched its Data, Technology and Analytics (DaTA) unit in February 2019 – which I founded and lead. Other authorities have also formed their own units including (non-exhaustively)

- The US Federal Trade Commission (FTC),

---

[1] For example, see George Westerman, Didier Bonnet and Andrew McAfee (2014), The nine elements of digital transformation, MIT Sloan Management Review.

[2] For example, the Financial Conduct Authority. See https://www.fca.org.uk/news/speeches/innovation-hub-innovation-culture.

- French Autorité de la Concurrence (AdlC),
- Canadian Bureau of Competition (CBC),
- Australian Competition and Consumer Commission (ACCC),
- the EU's Directorate General for Competition (DG Comp) and
- Netherlands Authority for Consumers and Markets (ACM).[3]

Bill Kovacic – the Global Competition Professor of Law and Policy at George Washington University, former FTC Chair and until recently Non-Executive Director at the CMA – has described technology-led transformation as 'one of the top five developments in competition over the last three decades, up there with the introduction of leniency programs'.[4]

What makes this such an important development? How exactly are these new data and technology units (henceforth data units) working in competition and consumer agencies? How might a data unit best be configured? This paper aims to unpack these questions. Overall, the paper should be of interest to two principal audiences: agencies that have a dedicated unit, are creating one or are considering doing so; and stakeholders of agencies including firms, their legal and economic advisers, and consumer groups. Others who may be interested include governments seeking to understand the benefits of technology-led transformation, sectoral regulators and data protection agencies, and academics.

Data units undertake five roles that contribute to a competition or consumer agency and its goals. The first four of these roles involve using data and technology skills to deliver directly for cases.

1. The most important role in terms of impact per unit resource is providing bespoke **expert data and technology advice**. Having a thorough grasp of the data and technology firms use is increasingly pivotal to cases across the CMA's markets, antitrust, consumer enforcement and mergers functions. For example, at the time of writing the DaTA unit has three team members

---

[3] See Competition and Markets Authority (2021). Compendium of approaches to improving competition in digital markets, published following the UK's presidency of the G7. The list is non-exhaustive and other countries have their own units, e.g. Spain. The FTC has had a Chief Technologist since 2011 and have other technology capabilities across the agency. The AdlC formed the five-person dedicated digital economy unit in January 2020. The ACCC restructured to create its Data and Intelligence branch in early 2021 including the now 19-person Strategic Data Analysis Unit. The CBC formed the position of Chief Digital Enforcement Officer and has formed the new Digital Enforcement and Intelligence Branch. And DG Comp announced the formation of its Intelligence, Analysis and Forensic IT Support unit in October 2021. The ACM hired its first data scientist in 2016 and has around 20 data scientists; they were spread through the agency but are now forming a single unit.

[4] Bill Kovacic (2022), 'A US Perspective on the Regulation of Digital Markets', 17th Annual Symposium on Competition Amongst Retailers and Suppliers, University of Oxford, May 13, 2022.

embedded in the CMA's antitrust team working on Google's Privacy Sandbox, the removal of third-party cookies on Chrome and introduction of new technology to allow targeted advertising.

2. Another crucial role is **data acquisition and data science**.

   - Bespoke big data handling and data science for individual cases can deliver new insight. For example, to understand economies of scale in search and the value of targeted advertising in the CMA's market study on digital advertising, we requested and analysed over 4TB of data from Google and Bing.

   - Agencies can build their own data, through scraping – the DaTA unit has scraped extensively – or through creating data pipelines.[5] They can then use machine learning, especially methods to analyse and understand language (called natural language processing), to detect problems.

3. Through **data-driven tool development**, data units drive efficiency and new capabilities within agencies. There are many aspects of agency work that can be automated and turned into tools/software – analogous to the growing use of legaltech in law firms. Yet there is insufficient demand such that external markets provide the required products, so internal data units can step in. For example, the CMA's Evidence Submission Portal automatically checks documents from firms are in the right format, saving staff time, and provides bespoke natural language processing capability for merger cases.

4. Complementary to the other roles is **behavioural science**. Consumer behaviour often plays a crucial role in digital competition and consumer cases. Digital firms have the infrastructure and capacity to test multiple designs of their user interfaces and analyse consumer responses, allowing for rapid optimisation that is not always beneficial to consumers. Behavioural teams conduct primary research and apply existing insights to cases involving consumer behaviour. For example, the market study on Apple's and Google's mobile ecosystems has drawn heavily on behavioural insight as have many consumer cases, e.g. on contract autorenewal or subscription traps.

5. A final crucial role is **research, horizon scanning and case pipeline[6] development**. Horizon scanning allows data units to identify new

---

[5] A data pipeline is a set of data processing steps from a data source to a destination data set, with the output of one step being the input of the next. It is valuable whenever data sources are used repeatedly.

[6] 'The case pipeline' and 'data pipelines' are both commonly used terms and so used in this document but are distinct. Data pipeline is defined in the previous footnote. Case pipeline refers to potential future cases that could be launched.

developments in technology and markets. Research allows for understanding the potential implications for competition and consumers. These activities help the CMA prepare to intervene if necessary and can feed directly into the identification of potential cases. Two research areas that we have been particularly active in investigating are algorithms – which have become an important issue in case work – and online choice architecture, the powerful effects of digital design on competition and consumers.

There are many design choices and trade-offs when founding a data unit, including

- How to structure a unit;
- What is the right allocation of resource between immediate impact on casework versus longer-term innovation; and
- How to hire and retain data scientists and engineers.

There is no one-size-fits-all answer, but many of the underlying issues are similar across organisations. Drawing on the experience of the DaTA unit, the paper illustrates likely challenges, lessons learned and examples of good practice.

The paper discusses how acquiring new capabilities may change competition and consumer agencies. One part of the logic of setting up a data unit is efficiency, either in understanding technology in cases or providing tools to deliver on cases. Another change is in the increased quality of agency outputs, e.g. improved pipeline ideas, analysis of a case, or the construction of remedies.

These attributes alone are valuable, but there are at least a couple of impacts from data units that can be described as potential game changers. First, a substantial part of what data units do is to code – coding is the major activity underpinning data acquisition and data science (role 2), and data-driven tool development (role 3). The marginal cost of copying code is close to zero. To the extent that agencies can share code with each other, or even develop code together, they can benefit from some of the same digital forces that are reshaping markets. The benefits to international collaboration in the data and tech space have the potential to be very high. Second, competition and consumer agencies have largely had their portfolio of cases driven by complaints or leniency and so the work is quite reactive. However, many issues in markets – e.g. the use of concerning online choice architecture practices – are inherently detectable from public information. If agencies can monitor markets regularly and systematically, that would be a radical change for competition

and consumer enforcement and move further toward the proactive identification of issues, previously the domain of sectoral regulators.[7]

Section 2 goes through the five different roles in turn. It explains each and gives concrete examples of how technologists substantively and operationally contribute to cases or projects, including what the impact was, primarily drawn from the CMA's experience. Section 3 discusses how to set up a data unit. It describes the skills that different technologists have and key topics to consider in creating a unit. This section will principally be of interest to other agencies, though law firms and consultancies seeking to build such units may also be interested. Section 4 reflects on how acquiring these new data and technology capabilities may change agencies in the medium-to-long-run (expanding on the comments in the previous two paragraphs). Section 5 briefly concludes.

[7] On the use of machine learning for proactive problem detection by sectoral regulators, see Stefan Hunt (2017), From Maps to Apps: the Power of Machine Learning and Artificial Intelligence for Regulators, Beesley lecture series.

2. What do data units do?

How can these new skills be deployed to make a difference? This section goes systematically through the five roles mentioned previously providing non-exhaustive examples of the DaTA unit's work.

There are four types of new skill, which this section refers to at various points: data science, data engineering, technology insight and behavioural insight.[8] Data engineering and data science are primarily quantitative roles, while 'technology insight' professionals provide a detailed qualitative understanding of how relevant technologies work and their implications. Section 3.1 discusses each of these skills in turn in more depth, for those interested.

2.1 Expert data and technology advice

At the CMA providing expert data and technology advice is, currently, the DaTA unit's highest return to resource. We can have high impact, often with just one or two team members providing specialist advice. Other data units provide a similar function. The DaTA unit has provided advice across market studies and investigations, antitrust, consumer enforcement and merger cases. This subsection goes through these four areas in turn explaining the roles that we took on individual cases and how we had impact, before discussing broader themes across our cases.

Market studies and investigations (analysing markets and rule-setting)

One of the first cases that we advised on was the digital advertising market study, which ran for 12 months from July 2019.[9] Digital advertising markets have evolved to use complex tracking technologies that allow firms to follow people as they use the internet, on both desktops and mobiles and across devices. Using these technologies, platforms disseminate information to advertisers, and advertisers decide how much to bid for advertising in real-time auctions. The market study team needed to understand these technologies as well as solutions that could enable privacy while allowing targeted advertising, such as privacy enhancing technologies. A data engineer already had some understanding of tracking and she became a core part of the market study team. She proposed a new area for investigation and wrote

---

[8] In addition, the DaTA unit has expertise that is more commonly found in agencies: in digital forensics – the recovery, investigation, examination and analysis of material found in digital devices – and eDiscovery – the process of identifying, collecting and sorting through the array of electronic evidence in cases. This paper focuses on the new skills that data units have.

[9] See https://www.gov.uk/cma-cases/online-platforms-and-digital-advertising-market-study

annex G – on the pivotal role of tracking – which has become an important resource for related CMA cases.

In the mobile ecosystems market study, we played a different advisory role.[10] Mobile platform operators retain a large degree of control over their platforms with restrictions in place that limit competition, which the operators argue are necessary to protect users against cyberthreats. The technology insight team investigated several aspects: security concerns created by opening up app distribution and web browsers to greater competition, mitigations that can help protect against threats, and challenges around technical implementation of interventions to improve competition.

Antitrust

Through the digital advertising market study, the CMA became aware of Google's proposed phasing out of third-party cookies on Chrome, its web browser, and introduction of new technologies that would support targeted advertising while preventing cross-site tracking: Google's Privacy Sandbox. The CMA had a concern that the proposed changes could undermine the ability of publishers to generate revenue through advertising, undermine competition in digital advertising and allow Google to entrench its market power as both a publisher providing advertising inventory and a provider of adtech services. In January 2021, the CMA launched an antitrust case in advance of the changes, working closely with the UK Information Commissioner's Office. Understanding the implications of the changes for the many parties affected has required getting deep into the technologies, with significantly increased involvement from the DaTA unit.[11] The CMA accepted commitments from Google setting out how the technology will be developed, with the CMA having an ongoing monitoring role.

This case is important. Google will be making these changes globally and so the CMA is taking action that will affect billions of people. At least $150 billion is spent on open display advertising through Chrome annually.[12] Even small impacts on changes to this market, e.g. on how efficiently ads are provided, can have large effects. And open display advertising is an important alternative to search advertising (mostly Google) or vertically integrated display advertising (such as Facebook). See Box 1 for further details. Technologists have allowed and are allowing the case to progress much more rapidly and more robustly.

---

[10] https://www.gov.uk/cma-cases/mobile-ecosystems-market-study

[11] https://www.gov.uk/cma-cases/investigation-into-googles-privacy-sandbox-browser-changes

[12] Source: author's calculations based on estimates of global internet advertising worldwide and Chrome usage on all devices.

Box 1: The Privacy Sandbox case, the technologies involved and the role of the DaTA unit

Google's Privacy Sandbox changes involve

1. removing support for third-party cookies (commonly used to track users across the web) and introducing other limits on tracking;

2. introducing new technology that replaces the functionality currently served by third-party cookies for certain use-cases.

The DaTA unit has attended key meetings with the parties and driven the CMA's understanding of the technology. The Privacy Sandbox includes a testing and trialling programme, and technologists are working closely with economists to make sure that the programme will adequately inform the CMA of the impact of the Privacy Sandbox.

To go into further detail, Google's proposed changes would

(i) permit the targeting of advertising with a very low risk that individuals could be identified (targeting),

(ii) allow measurement of the impact and success of advertising (measurement and attribution), and

(iii) prevent various backdoor routes that could allow actors to continue to track and identify users across the internet (strengthening of privacy boundaries).

Targeting happens through two means. First, a user's primary set of interests are determined from their browsing history, using machine learning. Second, a separate technology would allow re-targeting, i.e., if you have visited a website that indicates a specific interest, e.g. purchasing a particular product, then firms could target advertising based on that interest when you visit a second website. There will be a wide range of supporting infrastructure, policies and new technology to ensure that neither website can learn or infer that the user visited the other website.

Measurement and attribution: Advertisers need to know whether their advertising is successful or not, or their conversion. They need to i) measure their return on advertising spend and ii) further refine their targeting. This requires following users and observing the coincidence of advertising and product purchase or another conversion event. The basic insight behind the new measurement and attribution APIs is that advertisers do not need to know that a specific user saw an ad and

converted, only that someone saw an ad and converted. The new APIs would use a combination of aggregation, delay and adding noise to allow advertisers to obtain reports about ad exposures and conversions (at various levels of granularity) whilst preserving privacy.

Strengthening of privacy boundaries includes the restriction of information from users' devices that could be used for tracking (fingerprinting). The core rationale for these changes is that Chrome needs to take steps to prevent 'workarounds' that re-establish cross-site tracking in the absence of third-party cookies. However, some of this information fulfils other functions such as preventing fraud, complying with local and regional regulations (based on geolocation), tailoring web content to the user's browser and device, and more. This workstream involves careful design and judgement to ensure that these restrictions achieve the objective of preventing cross-site tracking, without undue negative impacts for websites.

A key aspect of the proposals is the use of differential privacy, a privacy-enhancing technology or PET, which uses a 'privacy budget'. The budget aims to ensure that even though there remain several streams of information about a user – through targeting, attribution and other functionality – it will be statistically very unlikely to be able to combine the information and identify the user.

The Commitments agreed with Google also include internal data separation between Chrome and Ads – so Google cannot self-preference. Before third-party cookies are removed, a standstill period is entered, during which the CMA makes a final assessment on whether to re-open an antitrust case or not.

Could an economist or lawyer understand all this technology? Yes, absolutely. But a background in the wide variety of technologies at play significantly accelerates how fast and how well the details and their implications can be understood.

Another antitrust case that the DaTA unit has worked on is the investigation into Meta's gathering and use of advertising data for its online classified ads and online dating services.[13] Initially the DaTA unit deployed its technology insight advisers. But it quickly became clear that understanding the use of data was going to require grappling with how Meta's algorithmic systems function and so we deployed a data scientist with many years of experience of using machine learning algorithms, including within tech firms. The data scientist made a large difference: he enabled

[13] https://www.gov.uk/cma-cases/investigation-into-facebooks-use-of-data

the case team to pursue important lines of inquiry that they otherwise would not have.

Consumer enforcement

Consumer enforcement has been a significant area of focus with the unit's biggest and longest engagement being the CMA's series of cases on online reviews. The first phase of the work focused on the platforms where fake and misleading reviews are traded.[14] This involved meeting, as part of the case team, with eBay and Facebook on their proposals to detect and remove the trading of reviews. These proposals involved algorithms and there were marked technical elements, including assessing the reasons the parties used to justify limitations with the implementation of remedies.

The second phase of the work involved the identification of which review businesses to prioritise for in-depth investigation and the subsequent launch of investigations into Google's and Amazon's approaches to identifying and moderating fake and misleading online reviews.[15] As fully embedded case team members, the DaTA unit wrote substantial parts of the detailed requests for information, including requests for considerable data and details of the algorithmic systems that Google and Amazon have built.

Mergers

Mergers have also been a considerable focus and the DaTA unit can play an important role as part of the case team in three different roles.

First, we help develop specific data or technology focused theories of harm. Two examples are the Amazon/ Deliveroo case and the Meta/Giphy case, which Box 2 provides more detail on.[16]

Second, we aid the case team in understanding the overall market or specific aspects of products, in cases where the DaTA unit has pre-existing knowledge of the merging sector. Cases here include Google/Looker and Salesforce/Tableau where cloud technology and the variety of different services on the cloud was a major focus.[17] The Nvidia/ARM proposed merger involved computer chips and several

---

[14] https://www.gov.uk/cma-cases/fake-and-misleading-online-reviews

[15] https://www.gov.uk/cma-cases/online-reviews

[16] See https://www.gov.uk/cma-cases/amazon-deliveroo-merger-inquiry and https://www.gov.uk/cma-cases/facebook-inc-giphy-inc-merger-inquiry

[17] See https://www.gov.uk/cma-cases/google-llc-looker-data-sciences-inc-merger-inquiry and https://www.gov.uk/cma-cases/salesforce-com-inc-tableau-software-inc-merger-inquiry

team members had good knowledge of how chips function and so could help the case team.[18]

Third, we help assess technical remedies, particularly relevant for intellectual property remedies. One example is Tobii/Smartbox where a remedy was proposed that would make the code from Tobii for their products available on an open-source basis.[19] We needed to assess how the remedy would likely perform over time. We also provided input on the Viagogo/StubHub merger on the feasibility of the parties' proposed remedy that involved StubHub selling its international business, including transferring a copy of its platform, mobile app, data, and employees to a new buyer.[20]

[18] https://www.gov.uk/cma-cases/nvidia-slash-arm-merger-inquiry

[19] https://www.gov.uk/cma-cases/tobii-ab-smartbox-assistive-technology-limited-and-sensory-software-international-ltd-merger-inquiry

[20] https://www.gov.uk/cma-cases/viagogo-stubhub-merger-inquiry

Box 2: Meta/ Giphy merger

The CMA's review of Meta's acquisition of Giphy involved assessing a vertical theory of harm that was focused on Giphy's data collection capabilities and the potential for Meta to use that data to disadvantage competing social media businesses. One of the DaTA unit's data scientists spent a considerable proportion of his time on the Phase 2 Investigation to make sure that the mergers team had a full and deep understanding of how this theory of harm might play out in practice.

Assessing whether Meta had the ability to foreclose required understanding the complexities of Giphy's business and how it collects (or is technically able to collect) data and the potential advantages of this data to Facebook. Giphy has an API that is a programmatic interface for its partners to request Giphy content and it has an SDK (Software Development Kit) with a richer set of development tools that provide broader functionality and allow more detailed, richer data collection.

To assess how Facebook could benefit from Giphy's data required understanding in detail the different types of user-level data that Giphy has access to such as advertising IDs, IP addresses and cookies. With an understanding of these technologies and internal Giphy documents, we then assessed how the data could be used to track individuals across different social media sites and augment Facebook's existing user profiles. Giphy also was able to analyse overall usage statistics for each API/ SDK partner. Using this information and again using internal documents we established that Facebook might be able to monitor, with some inherent imprecision, usage trends on individual, competitor apps in real time.

Assessing the theory of harm required getting into substantial detail on the types of data, understanding and assessing different technologies and making an overall judgement as to the value to Facebook of Giphy's data and the implications for Facebook's competitors. It was of substantial value to the merger case team and, ultimately, the inquiry group responsible for determining the outcome of the inquiry to be able to draw upon the expertise of a DaTA team member to provide deeper data and technological understanding.

Discussion

On all these cases, from market studies to mergers, data units can contribute in two ways. The first contribution is in getting deep into the data and technology at play in the case and understanding the details and the implications more quickly and thoroughly than other non-technical colleagues could alone. Technologists draft requests for information, working with the case team to write clear, interpretable questions that facilitate good evidence gathering on technical matters. And they are likely to assess the responses more effectively. They do not merely examine the specifications of the technical systems but also help interpret and contextualise a firm's internal documents and make judgements on how to interpret them. They then work to help inform the rest of the case team about the pertinent aspects of the technology and, together, assess the implications for the case.

The second contribution is in discussions with parties directly. There have been numerous instances where DaTA team members have attended meetings and assessed information provided in real time, challenging the interpretation, or asking follow-on questions. In some cases, team members have enabled the CMA to determine in real-time that parties have been technically misleading; and our presence has ensured a more accurate discussion. Team members have created important new lines of work in a case and significantly influenced the thinking of the case team. Of benefit to firms under investigation, is that these technical team members allow agencies to distinguish firms' good offers or outcomes in a negotiation from bad ones.

Many of these contributions were made possible by the private sector experience of some members of the DaTA unit as data scientists and engineers, including in technology firms. They understand how firms' systems can be structured and evolve over time; they have good intuition as to what is possible for firms to do and what is not. More generally, having staff members in a competition and consumer agency that are fundamentally interested in data and technology, and less so the novelties of a case from a legal or economic perspective, is a very useful complement. Given that data-driven technological change is happening across all industries – e.g. in health or entertainment – the technological perspective will continue to be crucial.

Another theme that runs across these cases is the importance and role of firms' algorithms. When we published our January 2021 research paper on algorithms (discussed in Section 2.5 on research and development), we were only just starting to realise how much grappling with algorithms would be important in cases. We realised that much of what people see on their mobile phones or online is curated and delivered by algorithms. And we realised that there were many theories of harm involving algorithms: undue self-preferencing, unfair personalised ranking, ineffective

use of algorithms to remove harms from platforms, and much more. Given what we have learned, I expect this work to continue to grow and deepen in complexity.

Other important thematic areas for understand technology are just beginning to emerge. I expand on these in the discussion on research and development.

2.2 Data acquisition and data science

Acquiring and analysing data to monitor markets and detect issues or to provide insights for cases is a mainstay of data units. In addition to more traditional methods for data capture such as formal requests to parties, seizing evidence through dawn raids and intelligence gathering, agencies now create their own data.

The DaTA unit has spent much time building data acquisition capability. First, we have scaled the CMA's analytical capability vertically, getting much more data through bespoke data requests to firms using information gathering powers than was previously possible. Analysis of these large datasets provides deeper understanding of markets and actionable insight. Second, we have scaled analytical capability horizontally, assembling new types of data by accessing public sources and some non-public sources of data, through creating data pipelines or scraping. The unit – as detailed below – has spent more time on developing this second type of capability.

This shift to more advanced data acquisition capabilities has the potential to cause significant change in how agencies work, especially in identifying new cases. Other agencies also have developed significant similar capabilities, for example the Strategic Data Analysis Unit of the ACCC and the AdIC.[21]

Big data handling and data science for cases

Data scientists and engineers allow agencies to request and handle much larger data sets. An example is the digital advertising market study mentioned earlier. In the study, we needed to assess Google's and Bing's relative position in search and search advertising. The market study team devised three pieces of bespoke analysis. One was to assess how much better Google's data on searches was compared to Bing's. Another was to quantify the impact of targeting on the price of advertising. And the third was a granular analysis of Google's fees using transaction-level data from Google Ad Manager. These analyses required a large data request of around 4TB of data. Together they provided crucial insights. I will focus my discussion on the first two analyses.

---

[21] See Competition and Markets Authority (2021). Compendium of approaches to improving competition in digital markets

The first analysis was on scale in search and the implications for the quality of search results. By getting and matching search data from Google and Bing we showed that Google's data on less common search terms was 30 times better than Bing's. The work influenced the case team to think about 'click and query' data opening remedies as a possible measure to increase competition in search. The second analysis demonstrated that targeting advertising led to a substantial increase of roughly 70% in the price of advertising – a significant impact of targeting, which was disputed by some. Box 3 explains the details of what the DaTA unit did on the case. Without advanced data handling and analysis skills, the CMA could not have undertaken this work.

Box 3: Big data analysis for the Digital Advertising Market Study

The market study conducted two analyses using a large amount of data from Google and Bing.[22]

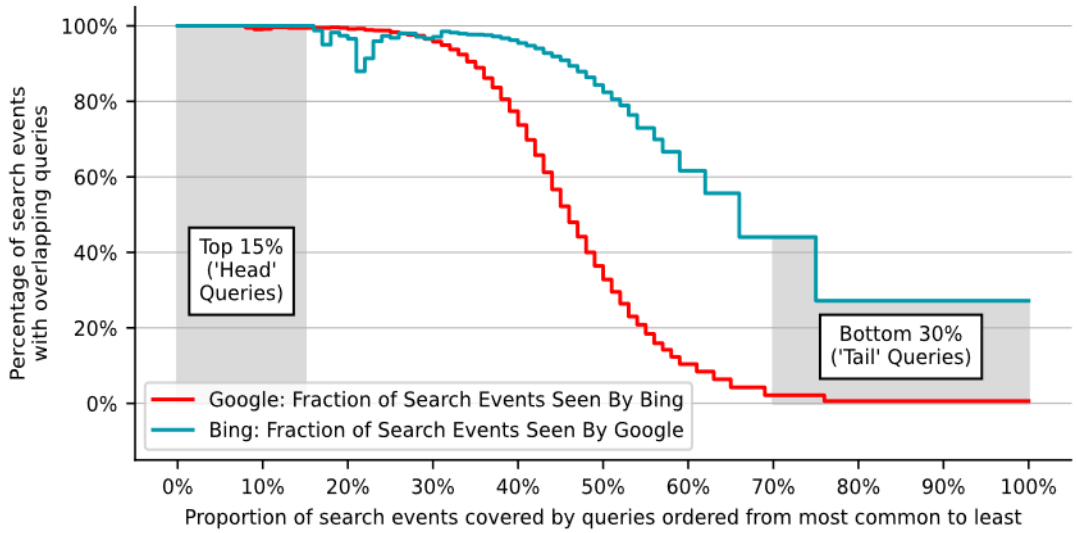
Scale in search: We received 1 week of all the searches on Google and Bing in the UK. As search terms can contain private information, e.g. home addresses, the text fields were hashed (a cryptographic function) so that CMA could not see them. But we could match the searches across both companies and observe if the two firms saw the same search term.

Figure 1 below shows the result. The red line shows the fraction of Google's searches that Bing also sees. From left to right, search terms are ordered from high frequency – searches that are seen a lot – to low frequency – more unusual searches. We can see that for the median search terms at Google, Bing sees approximately 25% of those searches, but for less frequent searches, in the bottom 30% – known as tail queries – Bing only sees 0.95% of the exact searches that Google sees.

The results are strongly asymmetric. Google sees over 80% of the median search terms that Bing receives. For Bing's tail queries, Google sees over 30% of these exact queries. So relatively it has over 30 times better data on tail queries. Overall while tail queries are less frequent, they are still 30% of queries and so performance on these still has a strong effect on consumer experience. Google has much better data than Bing that can allow it to provide better tail query results.

[22] See Simeon Thornton, Chris Jenkins, Giacomo Mason and Dan Griffiths (2020), Opening the Black Box: An Analysis of Google's Behavior in Search and Display Advertising Using Large-Scale Datasets, *Competition Policy International Antitrust Chronicle* October 2020

Figure 1:



Advertising effectiveness: using data from a large experiment by Google to test the impact on advertising publisher revenues when third-party cookies are disabled, the DaTA unit analysed how the extensive access to users' past browsing data enjoyed by large platforms could give those platforms a competitive advantage over other publishers. The dataset consisted of millions of rows of user data (browser, cookie age etc.) and the price and publisher revenue for the winning advertisement bid.[23]

The analysis found that blocking access to user cookie information reduces the advertisement publisher's revenue by an estimated 70% percent. This negative effect was larger for users with older cookies (i.e., more browsing data over a larger period) and was smaller for users who use browsers with anti-tracking technologies (i.e., Safari and Firefox). Therefore, the extensive access to user browser data enjoyed by large platforms could have a deleterious effect on competition.

In the mobile ecosystems market study, the CMA requested around 500GB of data from Apple and Google on app store usage. The market study team wanted to understand the revenue that Apple and Google made from app stores and how consumers used the stores to access apps. We used the data to understand high-level breakdowns and trends. We showed (among other findings) that games have a completely different profile to other categories of app. And they are the substantial

[23] See Appendix F of the digital advertising market study, from page 39

majority of overall billings for both Apple and Google stores through in-app purchases. And we were able to determine the relative importance of search algorithms in app stores, versus other channels such as third-party referrals, in driving app discovery and downloads.

Contributing to cases is not always about large datasets but also about using data science techniques to extract insight more effectively, especially from new forms of data such as documents or other passages of text (though it could be from pictures, video, satellite images and much more). Two examples come from the economic consultancy Compass Lexecon.[24] Both involve the use of natural language processing to get information from news articles. The first example involves identifying articles that talk about people transferring from one firm to another and extracting the relevant information, a tricky task to automate and one that required specialist data science expertise. Where human resource is an integral element of the quality of firms' products, similarities in the workforce between different companies could be relevant to an assessment of closeness of competition. The second example uses news articles to judge how a proposed merger affected the quality of a firms' product by assessing how positive discussion was about the product, pre- and post-announcement of the merger.

Data pipelines

The building and maintenance of a data pipeline – a set of data processing steps from a source to a destination, with the output of one step being the input of the next – is valuable whenever data sources are used repeatedly. The DaTA unit has created several data pipelines so far and I discuss three notable examples. I would expect to see a significant increase in use of data pipelines, given an expansion in the use of data for enforcement and the proposed ex-ante regulation of digital platforms (which may need ongoing data access to appraise compliance).

The first example is the data pipeline created for the CMA's Covid Taskforce, from March 2020.[25] The context was that markets were changing rapidly given the shock of lockdown and the introduction of Covid regulations, leading to price spikes and the unexpected cancellation of travel and pre-booked events. The CMA needed up-to-date information and intelligence on where there were problems to be tackled, such as companies not meeting their obligations to consumers. We launched a webform to collect complaints. In order not to limit the topics of the complaints, many fields of the form were free text boxes. This flexibility however raised issues given the large

[24] https://www.compasslexecon.com/the-analysis/using-natural-language-processing-in-competition-cases/03-22-2022/
[25] See https://www.gov.uk/government/news/latest-update-from-cma-covid-19-taskforce

numbers of complaints we received, above 5,000 per week for several weeks and over 15,000 in the busiest week, considerably more than humans can handle.

The CMA needed to understand what was in each complaint – the issue, the company, the sector. Creating a data pipeline allowed us to take in the complaints (the data) from the webform, perform many steps to clean them, infer their content and turn text into actionable data, using machine learning. There were several technical challenges we faced, and we solved them using natural language processing techniques.[26]

The data pipeline allowed weekly (or more frequent) internal reporting on the key markets and firms that people were worried about, and it allowed us to track these issues over time as the pandemic evolved. It led specifically to the launch of several consumer enforcement cases and enable us to work out which sectors and traders to prioritise. And it allowed us to check that complaints had significantly decreased after our interventions.

The second pipeline example is the Evidence Submission Portal (ESP), which the CMA now uses for almost all document submissions for merger investigations. The ESP can take in millions of documents, check that they are in the right format and process them. ESP is also a tool, as well as a pipeline, and I discuss it more fully in subsection 2.3.

The third pipeline example is currently in an advanced stage of development. It takes records from all registered limited companies from Companies House, the UK's registrar. These data are frequently needed by the CMA for many reasons, including i) getting intelligence on companies with respect to suspected cartel activity, ii) understanding ownership structures in markets and iii) understanding the state of markets, especially concentration and profitability levels. Ordinarily CMA staff navigate the publicly available search tool and download the data by hand. But this is time-consuming, and the manual nature of the process could lead to errors. The pipeline will regularly take in all the data, clean it, deduplicate it, and make it available in an easy-to-use tool, designed for the CMA's needs (with the pipeline also available more widely). This is a significant engineering challenge, especially as the dataset is reasonably large.

[26] For example, to identify which products were the object of the complaint, we used Named Entity Recognition and a neural network-based algorithm. And to understanding which sectors businesses were operating in, we used a supervised machine learning classifier – a gradient-boosted tree trained on hand-labelled data according to a predetermined classification – to allocate complaints to industry sectors

Scraping

Web scraping is an important way to get data to monitor and detect issues on websites. If problems – or signals of potential problems – can be directly observed, then scraping can be useful. This is often the case with respect to consumer law (e.g. worrying choice architecture practices) and can be true of competition law (e.g. mergers that the authority might want to investigate in more detail, or patterns that might signal resale price maintenance). Using machine learning, with humans checking and providing oversight, we can often analyse the data and assess whether there is a problem, frequently with high accuracy.

Scraping has been widely used in academic research, especially in computer science, to identify problems in markets, e.g. scraping surge prices on ride sharing services to assess their fairness and transparency, or the use of 'dark patterns' by firms to negative influence consumers.[27]

The DaTA unit has conducted extensive web scraping on online reviews and used techniques to detect suspicious patterns indicative of fake and misleading reviews. To decide which companies to investigate in depth, we collected evidence from several different platforms. We used techniques both to detect suspicious patterns ourselves, and to ensure that we had a thorough grip on the methods, so in any subsequent investigations we could assess the advantages and drawbacks of firms' approaches.

We also built a software tool to help detect resale price maintenance that scrapes price data, looks for suspicious patterns, and presents it to case officers.[28]

Scraping can also be used to check compliance with CMA remedies or guidance. The payday lending market investigation included a remedy that lenders needed to put a link to a price comparison website on their webpage.[29] We created code to scrape company websites to check that there was a link, writing to any non-compliant firms and telling them to comply. For the CMA's ongoing consumer enforcement case on social media endorsements, we have used scraping to automatically check compliance with the guidance on disclosing commercial relationships.[30]

---

[27] For example, Le Chen, Alan Mislove and Christo Wilson (2015) Peeking Beneath the Hood of Uber in *Proceedings of the Internet Measurement Conference* (IMC 2015). Tokyo, Japan, October, and Arunesh Mathur, Gunes Acar, Michael Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty and Arvind Narayanan (2019), Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites, *ACM Computer Supported Collaborative Work and Social Computing* (CSCW

[28] See Simon Nichols, 29 June 2020, Restricting resale prices: how we're using data to protect customers

[29] See https://www.gov.uk/cma-cases/payday-lending-market-investigation#final-order

[30] See https://www.gov.uk/cma-cases/social-media-endorsements

The DaTA unit has scraped extensively and built a significant codebase – as have teams in some other agencies. The development of this capability means that the CMA can far more easily acquire data through scraping in the future.

Discussion

The capabilities to acquire data from data pipelines or scraping and then manipulate data using machine learning make a big difference for agencies. When we write code that gets or manipulates data, we have created an asset that is at least partially reusable. As we create more and more code, we can create a library of tools – as the CMA has now for web-scraping and for natural language processing – that can be quickly deployed for new problems. We have built considerable knowledge in creating these assets: we are much better and more efficient in how we scrape, organise and clean data. And there are robustness benefits: because our code and data are all carefully version-controlled and we can go back to any previous versions, all our work is easy to check afterwards and is reproducible.

The DaTA unit has started sharing its coding assets with other agencies and we think there is considerable potential for agencies to develop digital assets together.

2.3 Data-driven tool development

Historically, competition and consumer agencies have bought digital products: for document review (e.g. Nuix or Relativity), analysis (e.g. Stata or Excel) or digital forensics (a host of products). These products were hosted on agencies' own hardware (usually on-site servers).

Most businesses also predominantly bought much of the software that they used as well. But as digital skills have proliferated and as organisations have moved to the cloud there has been a digital transformation. Now many firms employ their own data scientists and engineers and develop bespoke software that is unique to the challenges that they face. Cloud services provide many tools that make it easy to acquire data, to build analytical capabilities to extract insight or to create user interfaces.

Agencies can and should be developing their own tools. In many areas of their work there are no software products available commercially which fully meet agencies' needs. For example, the AI capability in commercial document review software is generic and not set up to understand the language used in the context of mergers, nor is there any commercial product to monitor markets for potential law breaches. The opportunity for agencies to develop their own technology is akin to the legaltech opportunity for law firms.

Nonetheless building technology takes time and necessitates an ongoing commitment to maintain and develop the products. Agencies need to consider the trade-offs between buying versus building tools (or doing nothing, or mixing buy and build). The CMA's strategy is to continue to identify where the DaTA unit can develop solutions that deliver casework benefits. We will build solutions in-house when there is no cost-effective commercial solution, the cost-benefit analysis of building is clearly positive, and we can integrate our solution into the wider CMA tech landscape.

One area that we have created a tool is our platform LEDA,[31] for creating a data lake (a system for storing and accessing raw data) and providing analysts with access to the data lake and to cloud computing services. Our data scientists and engineers need flexible access to coding packages, version control of all code and data (using GitHub) so that we can reproduce all analysis, and the ability to build and control information security and data protection.[32] Working closely with our IT department colleagues, we developed a data infrastructure capable of ingesting, curating and processing sensitive data at the scale required of a leading competition authority and with modern advanced functionality. In the last three years we have acquired over 160TB of data across over 130 million objects stored in our cloud-based data lake, at minimal cost and risk. This is important for our quantitative teams, who have excellent alternative employment opportunities and really care about quality of the 'tech stack': they have the tools they need for the job, like the platform, and feel empowered.

The CMA's Evidence Submission Portal (ESP) exemplifies a situation where there was no existing solution to meet our requirements. Our data engineers and eDiscovery specialists built a product that integrates well with our commercial eDiscovery tool, Nuix Discover. We can now take in many millions of documents for cases (for the Nvidia/ARM proposed merger we had to ingest over seven million documents), whereas before we could only deal with much smaller sizes (a few hundred thousand). ESP automatically checks all documents are in the right format, rejecting submissions for those are not. It then automatically processes the documents and makes them available for loading into Nuix Discover, tasks that previously required significant human oversight. We have reduced the amount of time from submission of documents by the parties to when the merger case team

[31] LEDA stands for LEDA is an Environment for Data Analysis. The name is recursive.

[32] We have built information security and data protection into our data lake by design. Data are segregated by case and marked individually with key metadata such as their security classification and associated Information Asset Owner. Expired data will be automatically deleted if no action is taken, and unmarked data will be rendered inaccessible until it is appropriately catalogued. User access to data is temporary, only possible through the CMA network, and is granted through a rigorous business and technical process. All interactions with data are auditable and attributable.

can review the documents from 4 days down to 1 day, which can make a big difference when under time pressure.

In addition, using natural language processing techniques, we have started deploying new capabilities into ESP that provide additional insight on individual documents, helping to speed up the evidence review process. Over time we hope to develop a range of bespoke capabilities, specifically relevant to the context of document review for the different types of cases that a competition authority undertakes (e.g. different types of abuse of dominance case). Through gathering feedback, we are learning considerably from the rollout of these new capabilities.

We have deployed three data science enrichments to Nuix. The first is a 'keywords enrichment', which extracts the most representative words for each document, allowing reviewers to better understand the likely content. The second indicates how likely it is that a document is about competition-related topics (e.g. competitors, prices, market structure, etc.). The third indicates whether a document talks about activity that is within the UK. We are testing this functionality and receiving feedback on how useful it is with some strong feedback of its utility so far.

In addition to these tools, we have created software that i) further automates local area analysis for mergers, ii) aids the detection of resale price maintenance, and iii) simplifies and reduces the scope for human error in the redaction of documents.

Other agencies are building tools as well. For example, the Danish Competition and Consumer Authority has created a tool, BidViewer, to screen for potential cartels.[33] The tool has a variety of different analytical capabilities and uses machine learning that can combine multiple indicators into a single model to identify suspicious behaviour. The tool comes as a piece of software and has been made available to many other national authorities, including the Spanish and Swedish competition authorities. There are several other authorities that have been using the software too and the UK is in the process of starting to use it. As more and more authorities work together and contribute increasing amounts of data, the models become more accurate and detect collusion better. Moreover, there is the potential for agencies to work together on the software development, with many agencies globally benefitting.

Data-driven tool development is inherently a process that requires agencies to invest with benefits over the medium to long-run. To create products that will have very high impact, and then maintain and improve them, it requires an ongoing focus on investment and engagement between users in the 'business' and the developers.

---

[33] See Danish Competition and Consumer Authority (2022), Collusion detection in public procurement using computational methods

This can be hard in organisations that are not used to this kind of investment and where the business is focused on cases with short-run deadlines.

But the gains can be large and sustained, both in terms of removing or speeding up mundane work or in terms of adding new capability. Moreover, to the extent that agencies can share tools and work together in their development – which we are beginning to actively work on – the potential for mutual gains for the world's national competition authorities and ultimately impact for our citizens is great.

2.4 Behavioural science

To diagnose and address demand-side issues effectively, competition and consumer agencies need to update their understanding of consumer behaviour to reflect the latest knowledge from behavioural science. It has been known for a long time that you cannot introspect what drives your own behaviour, let alone other consumers' behaviour. Behavioural science has made enormous progress in building evidence on what drives behaviour. Sophisticated firms use behavioural science skills in abundance, especially among their user experience (UX) researchers, to influence consumers' behaviour. And we know that the impact of firms' design – aka choice architecture – on consumer behaviour can be large and surprising, especially in digital markets.[34]

Agencies therefore need to use behavioural skills and evidence to understand how firms aim to influence consumer behaviour, and how this influence may harm consumers and competition – e.g. if they mislead or pressure consumers into decisions that are not in their best interest, or weaken competitive pressures in the market.

While generalist lawyers and economists in agencies have always had to consider consumer behaviour, particularly as part of consumer protection and enforcement work, there are marked benefits to using behavioural specialists. The discipline has advanced hugely in the past 20 years, and there are vast, high-quality literatures outside economics to draw upon, different research skillsets and more diverse ways of thinking rigorously about problems. The capabilities of firms to influence consumers has also grown, having access to the same growing evidence from behavioural science, greater control over how they design their products and interact with consumers online, and the ability to run experiments at scale: in 2019 Google ran over 464,065 experiments on search alone.[35] Analogous to technology insight

---

[34] For the importance of choice architecture issues in digital markets see Unlocking digital competition: Report of the Digital Competition Expert Panel, March 2019, often referred to as the Furman Report, and Stigler Center (2019), Committee for the Study of Digital Platforms: Market Structure and Antitrust Subcommittee Report, July 2019

[35] See https://www.google.com/search/howsearchworks/mission/users/

specialists, having specialists with in-depth knowledge of behavioural science allows case teams to understand the role and implications of consumer behaviour in a given context more efficiently and impactfully.

The CMA set up a Behavioural Hub in early 2020, to help with identifying problems in markets, diagnosing problems and devising remedies. The Hub was initially just three people, including a secondee, and has recently grown to six permanent staff. Other agencies also have similar teams including the ACM and the FTC – including people with a background in UX – and the CBC is now building its own team.

The Behavioural Hub's first project was the digital advertising market study, focusing on the choice architecture of controls over the use of personal data for targeted advertising.[36] We found that platforms' design could inhibit consumers' ability to exercise informed choice: default settings favoured the platform; there were long and complex privacy policies and terms; and information and choices were presented in ways that could 'nudge' consumers to make decisions most favourable to the platforms. The analysis used a detailed assessment of platforms' choice architecture including academic literature on the impact of these architectures.

Another early project was a series of cases focusing on autorenewal contracts for McAfee and Norton antivirus products and gaming subscriptions for Microsoft Xbox, Sony and Nintendo.[37] Consumer behaviour was central to the investigation. We analysed firms' interactions with consumers – their emails, websites, notifications – and the precise elements of the design. Much is known about the effects of this design – defaults, ease of exit, comprehension of contract terms, provision of risk information – and the behavioural insight advisers built up very detailed evidence on the impact of these elements. The team advised on data collection including what data and information could be requested from firms and how it could be used to support legal arguments. This included advising on how changes introduced by firms could be used to assess the impact on consumers using quasi-experimental methods. As the cases progressed to remedies, the team members were active in assessing the likely impact of the remedies and so what the CMA should accept in terms of undertakings.

More recently, a couple of behavioural scientists worked on the mobile ecosystems market study team, which covered the core elements of operating systems, app stores and web browsers on Apple iOS and Google Android. All these elements are carefully designed, and the team analysed how consumers might react to the use of different choice architecture practices on mobile devices and how this can affect

---

[36] See Appendix Y of the digital advertising market study

[37] See https://www.gov.uk/cma-cases/anti-virus-software and https://www.gov.uk/cma-cases/online-console-video-gaming

competition. This including evaluating the evidence submitted by stakeholders and academic literature on psychological mechanisms. The team assessed pre-installation, default setting, ease of switching default, and other choice architecture for mobile browsers and the design of Apple's App Tracking Transparency framework, among other elements.[38] And they advised on the design of potential remedies to address the concerns identified.

Looking across the CMA's portfolio, the behavioural team provides advice regularly on consumer enforcement cases, which will remain a large share of the team's work. As digital antitrust cases have launched the team has started to work on them, and I expect the team's work going forward to include market studies and investigations, antitrust cases and considerable work on ex-ante digital regulation by the DMU.[39]

Up to now, the team has focused on providing advice for cases, either ad-hoc or embedding into case teams, and providing detailed behavioural analysis using information from stakeholders and existing literature. But going forward, our offering is expanding to include experiments – we are concluding the analysis of our first experiment currently, on the impact of misleading green claims on consumer purchase and search decisions.[40] The green claims experiment is an online experiment where participants make hypothetical decisions, but we also have the capability to conduct field experiments on real consumer decisions, as other regulatory agencies have done.[41] While each case needs to be considered on its own merits, and there are considerable constraints on agencies being able to use experiments effectively (e.g. statutory deadlines), I expect that we will and should see increased use of experiments, especially for testing the impact of consumer-facing remedies.

We also can provide other methodologies, such as eye-tracking, qualitative research/ethnography and the analysis of natural experiments.

2.5 Research, horizon scanning and case pipeline development

This subsection covers a range of research and development activities. I first discuss our working paper research before discussing our work on understanding market trends and emerging technologies. Our research and development is designed to contribute to the pipeline of new cases, to improving analysis within cases or to other goals (e.g. developing standards for AI).

---

[38] See Appendix G and Appendix I of the mobile ecosystems market study interim report

[39] Future work could even include merger cases. See discussion in Amelia Fletcher (2019), The EU Google Decisions: Extreme Enforcement or the Tip of The Behavioral Iceberg?, *Competition Policy International Antitrust Chronicle* January 2019

[40] For the wider case, see https://www.gov.uk/cma-cases/misleading-environmental-claims

[41] For example, see Financial Conduct Authority (2018), When and how we use field trials.

Research

As we grapple with firms' use of a wide range of new technologies, the nature of cases is changing. And the formation of data units, including behavioural science, is new for competition and consumer agencies. We should expect that there is much we do not know about applying these skills to casework. In this context, we need research to build knowledge and capability. To date, the DaTA unit has focused on two streams of research: algorithms and online choice architecture.

We initially focused on algorithms, as they have become increasingly important for markets and in an information-saturated world will (and should) become more so. Firms are increasingly using machine learning algorithms, or AI, in ways that are consumer-facing, e.g. recommending what a customer should purchase. The competition world had been focusing on theories of harm on collusion in prices, and yet we believed that there were already many other concerns that may be having more impact in markets and yet had not received due attention, e.g. undue self-preferencing or unfair personalised ranking. We needed to build the theoretical understanding and empirical toolkit to help identify potential cases and to successfully deliver the analytics for any CMA cases.

Our first research paper was 'Algorithms: How they can reduce competition and harm consumers'.[42] It was essentially a literature review structured to be of practical use to agencies, which included a taxonomy of the main theories of harm, a review of available methods (mostly developed by academics for use with public data, not with agencies' information gathering powers) and the potential role for agencies in addressing harms. The paper aimed to lay out the territory that the CMA would need to cover over the next 5 to 10 years.

We have quickly seen the paper become practically relevant. The CMA's cases on Google and Amazon's systems to deal with fake or misleading online reviews, Google's Privacy Sandbox and Meta's use of advertising data all require substantial knowledge of algorithms. And other cases have needed input, including social media endorsements on Instagram, hotel booking site investigations[43] and the mobile ecosystems of Apple and Google. We have also actively worked on the case pipeline and understanding areas of potential focus for the DMU, in particular on issues of undue self-preferencing. The research paper provided a framework for understanding the different theories of harm and similarities and differences analytically between different cases.

---

[42] Competition and Markets Authority (2021), Algorithms: How they can reduce competition and harm consumers

[43] See https://www.gov.uk/cma-cases/online-hotel-booking

We anticipate publishing a paper later in 2022 on the CMA's practical lessons from cases involving algorithms. We have also conducted additional research on algorithmic harms and algorithmic auditing working with other UK regulatory agencies.[44]

The second line of research is online choice architecture (OCA) and the Behavioural Hub published a discussion paper and an extensive evidence review in April 2022. The aim of the research was similar to the algorithms paper, seeking to lay out a roadmap of an area that we thought would become increasingly important. OCA featured prominently in the Furman report and the Stigler Centre report on digital markets.[45] And OCA is currently an important part of many digital cases, e.g. browser defaults on Apple devices, or app stores in mobile ecosystems, or privacy settings within Google's Privacy Sandbox. We aimed to prepare the CMA systematically for casework over the coming years. The first paper outlines the harms that can arise (while noting that OCA can be and often is hugely beneficial) includes a taxonomy of 21 concerning practices that agencies need to be aware of and alert to.[46] The second paper is a long and detailed review of the practices and cross-cutting themes, which we intend to be a reference document for agencies, providing a trove of evidence to support future cases.[47]

In the first of our two OCA papers, we outlined the next steps for the CMA. In addition to using our knowledge directly in current and upcoming cases, we promised further research to determine the prevalence of harmful OCA practices in different sectors. This will proactively contribute to guidance, input into legislation or input into the case pipeline.[48]

Horizon scanning and emerging technologies

If competition and consumer agencies become aware of potential issues earlier, there can be more options for taking actions to reduce future consumer detriment, which could be less interventionist and potentially have lower costs and high benefits. For example, we might address issues through merger control now as

---

[44] Through the Digital Regulation Cooperation Forum (DRCF) – the group of four UK digital agencies, the CMA, Ofcom, the Information Commissioner's Office, and the Financial Conduct Authority. The first paper – Digital Regulation Cooperation Forum (2022), The benefits and harms of algorithms: a shared perspective from the four digital regulators – sought to take a broad view on algorithmic harm and cover the perspective of all four agencies. The second paper - Digital Regulation Cooperation Forum (2022), Auditing algorithms: the existing landscape, role of regulators and future outlook – focused on algorithmic auditing: what is the state of the market today, where might and should it go in the future, and what role(s) might regulators take.

[45] See Footnote 34.

[46] CMA (2022). Online Choice Architecture: How digital design can harm competition and consumers

[47] CMA (2022). Evidence Review of Online Choice Architecture and Consumer and Competition Harm

[48] The Behavioural Hub has worked closely with the Consumer Enforcement team to i) share the CMA's views on OCA practices with government as part of a consultation on consumer protection law changes, ii) form a view on what guidance might be most useful

opposed to antitrust later; launch antitrust cases to intervene following the announcement of intended conduct, as opposed to active conduct (e.g. Google's Privacy Sandbox); work on developing standards in co-operation with industry (e.g. on AI); or create new ways to ensure safe product development (e.g. facilitating the creation of audit markets for artificial intelligence). Or we could address issues through ex-ante regulation, e.g. using the proposed powers to be given to the Digital Markets Unit (DMU). There are thus benefits to monitoring technological progress and the forefront of innovation, making sure to understand i) the technologies at play, ii) how they may affect markets and iii) potential issues for competition or consumer protection.

For these reasons, the CMA has a horizon scanning, emerging technologies and digital market insights function. Clearly undertaking this kind of work requires a mix of skills. First, you need team members that have a background in and experience of horizon scanning and managing a rigorous process. Second, you need staff that deeply understand the technology and can convey the main issues – from the point of view of agencies – to other team members. Third, you need staff that are steeped in addressing competition and consumer protection issues, to make sure the outputs of the work are actionable. At the CMA, the technology insight team in the DaTA unit (providing the first and second skills) and the DMU (providing the third) jointly operate this function, drawing on technological expertise in data science and engineering.

There are four stages to the process: (1) horizon scanning to develop and update a watch list of emerging technologies or trends;[49] (2) a research stage for prioritised issues, including producing Technology Primers; (3) external engagement to explore and test identified issues; and (4) formal studies, through the CMA case pipeline. The technology insight team in the DaTA unit principally lead the first two stages, working closely with the DMU, while the DMU principally leads the two later stages with input from technology insight.

So far, we have researched cloud technologies, privacy enhancing technologies (PETs), home-based internet of things (IoT) and the metaverse. The process of writing up analysis and engaging a wide variety of people across the organisation – including DMU, DaTA, competition, and consumer protection – and then presenting

[49] Horizon scanning is a systematic process for spotting threats, risks, dynamic change, and opportunities. It is a method for anticipating change and is used across governments both in the UK (for example the Government Office for Science) and internationally, by the European Commission and OECD. See Government Office for Science (2021). A brief guide to futures thinking and foresight or Government Office for Science (2017). The Futures Toolkit: Tools for Futures Thinking and Foresight Across UK Government. It can provide insights such as how might the technologies we identify contribute to shifts in business and how significantly? Or how might developments in AdTech markets impact competition? It can be done using a variety of methods and for different purposes e.g. to spot new technologies, wider societal trends, or developments in particular sectors or markets.

and discussing internally has ensured that the pre-existing knowledge of technical experts as well as new knowledge gained have been disseminated widely.

Way forward on research and development

In addition to the research programmes and horizon scanning, the DaTA unit has also prototyped a process for hackathons,[50] to create a new way of innovating for the CMA, building on work from other parts of the UK government and the FCA.[51] Our first, internal hackathon directly contributed to beginning the Companies House data pipeline project discussed above as well as providing a promising consumer protection pipeline case.

There are many other topics where new research is likely to reward more effort. Our focus on algorithms will continue and there are many different potential avenues. One important area is recommender systems (RecSys), algorithms that provide recommendations to consumers and are core services to facilitate consumer choice. There are many examples: playlist generators for video and music services, product recommenders for online stores, or content recommenders for social media platforms. While obviously these systems are crucial and often beneficial, there are concerns about competition and consumer harm, including how choice architecture interacts with the algorithm. The necessary expertise for understanding the impact of RecSys on consumers requires multiple skills including data science, technology insight and behavioural science.

The importance of PETs is underlined by our Technology Primer and our work related to cases – on Google's Privacy Sandbox and thinking about creating access to Google's click and query data on search. As the internet develops and permits greater privacy, PETs are likely to become an increasingly widespread set of technologies. Given the intricate relationship between privacy and competition, agencies will need to continue to develop their technical expertise in this area.

One final area to note is interoperability. While the importance of interoperability for breaking down barriers to competition has been understood for a long while, authorities now need to build their technical understanding of the possibilities and limitations. Following the March 2022 agreement of the European Parliament on the Digital Markets Act in particular, which mandates interoperability for messaging services, this is an active area that requires research.

---

[50] Short events, typically one or two days, focused on making rapid progress on particular problems, often related to software development but can be used for any problem-solving activity.

[51] See https://www.fca.org.uk/firms/innovation/regtech/techsprints

These topics are a brief sample of areas suitable for further research and development. Going forward we will continue with active research and making sure that our research work directly drives CMA cases, as with the next stage of our OCA research, or meets other goals.

3. How to set up a data unit: reflections from the CMA's experience

Building the DaTA unit and integrating into the CMA has gone well. The team has grown, and has been asked to be involved in, and lead, many projects. There are many design choices when founding a data unit and issues that the DaTA unit has grappled with. There is plenty to be learned from our successes but also from what could have gone better.

This section is primarily intended for other competition and consumer agencies who have set up a data unit, are in the process of setting up one or are considering doing so. It should also be of use to other authorities (e.g. sector regulators) and, in part, to law firms or economic consultancies thinking about setting up a unit. There is no right answer for any question: the aim in sharing our experience is to help others think through what their answers are.

The first subsection goes into further detail on the skills needed for data units. The four other subsections discuss key questions to consider when setting up a unit, and reflections from the CMA's experience, around the topics of

- Unit structure and recruitment;
- Prioritisation;
- Organisation for effective delivery and for retention; and
- Technology requirements.

3.1 Digital skills in agencies

At the CMA, as we have considered how to get the most impact on our case portfolio, we have developed four pools of skill in the DaTA unit that are new for the organisation (in addition to integrating the pre-existing skills of digital forensics and eDiscovery):

- Data science
- Data engineering
- Technology insight

- Behavioural insight

These are also the skills that other competition and consumer agencies have been hiring, and other UK domestic agencies have been hiring for the same skillsets.[52]

A data scientist extracts information from data using a wide range of traditional estimation and modern machine learning or Artificial Intelligence (AI) techniques.[53] In agencies, data scientists can use these skills on individual cases or to create tools that use advanced data techniques. With reference to skills that are more typically in competition agencies, data scientists draw on a wider pool of practical techniques than econometricians. Econometricians are well-suited to become data scientists, and many have made that career transition.

Data engineers are more focused on the infrastructure and data pipelines and deal with issues such as formats, resilience, scaling, and security. At the CMA our data engineers also do much work that would typically be called 'DevOps', a blend of software development (dev) and IT operations (Ops). The DaTA unit's tool development draws on these DevOps skills, for example.

These two types of professionals typically work with data and code (at the CMA, predominantly Python, with some R and other languages). For data professionals' careers, it is important for them to use and develop their coding skills. Yet most of the delivery work usually carried out in competition and consumer cases is qualitative, benefitting from in-depth domain knowledge of specific markets and investing in learning about the specific firms involved. To keep quantitative team members focused on their core skills and gain from the benefits of specialisation, we have found it important to create carefully moulded roles for them on cases, e.g. conducting big data handling and analysis for the digital advertising or mobile ecosystems market studies only, but not joining many case team meetings.

There is considerable qualitative work that is still best done by staff with a strong background in data and technology, as opposed to competition lawyers or economists. This is why, at the CMA, we created a technology insight team. This team undertakes qualitative analysis of technical issues related to a case, and supports the case team to understand the implications for the case objectives. In addition, often it is this team that begins the DaTA unit's engagement with cases and then helps to shape roles as the case develops, when needed, for data scientists

---

[52] For example, at Ofcom, the UK's communications regulator, the Chief Technology Officer – my counterpart – has a team that covers data science, engineering and technology insight. Behavioural insight also exists but sits within economics. The Financial Conduct Authority, the UK's financial regulator, has all four of these skills in different parts of the organisation.

[53] For definitions of data science, machine learning and AI see Stefan Hunt (2017), From Maps to Apps: the Power of Machine Learning and Artificial Intelligence for Regulators, *Beesley lecture series*, pages 4-5

and engineers. The team also leads a range of other work, especially horizon scanning and the understanding of emerging technologies.

The career and academic backgrounds of the technology insight team are more varied than the data science and engineering teams, who typically have quantitative degrees that involve coding: computer science, applied mathematics, physics, the life sciences or applied econometrics.[54] The technology insight team has people with these degrees, but also includes people with a deep interest in technology and other backgrounds, such as law or economics, including some people who were existing CMA employees when the DaTA unit was formed. While some members of technology insight have worked as data scientists and engineers, others have worked in technology policy in the government or civil society.

The final pool of skills is behavioural insight, which applies approaches from a range of related disciplines to analyse people's behaviour and their decision making, including behavioural economics, the behavioural sciences (e.g. social or experimental psychology, or neuroscience), quantitative analysis (e.g. statistics and causal inference) or other related disciplines (e.g. anthropology and ethnography). Although these skills are relevant in many situations involving consumers across the wider economy, in practice over the past couple of years, we have found that digital cases require greater behavioural input.

It may also be useful to be clear what the remit of the team does not involve:

- At the CMA, the DaTA unit is not responsible for the organisation's policies with respect to data use (aka Data Strategy or Data Governance), which sits within the CMA's Legal Service.[55] This split is different to other countries – such as Australia – where this responsibility sits within the data unit.

- We are not responsible for policy work. However, several cases or projects that we have been instrumental in have involved developing policy positions and dialogue with government and other regulators. More generally, members of the unit play an important role in providing technological insights to help inform the CMA's policy development (such as through our work on algorithms).

- We are also not responsible for the intelligence function, nor for the IT function beyond DevOps related to DaTA unit's tools.

---

[54] Of course, data and coding skills can be built through means other than degrees.

[55] Although we are responsible for parts of the implementation of controls on the systems we maintain.

The next four subsections consider key choices to be made in a data unit within an agency, including how to determine the boundaries of the unit.

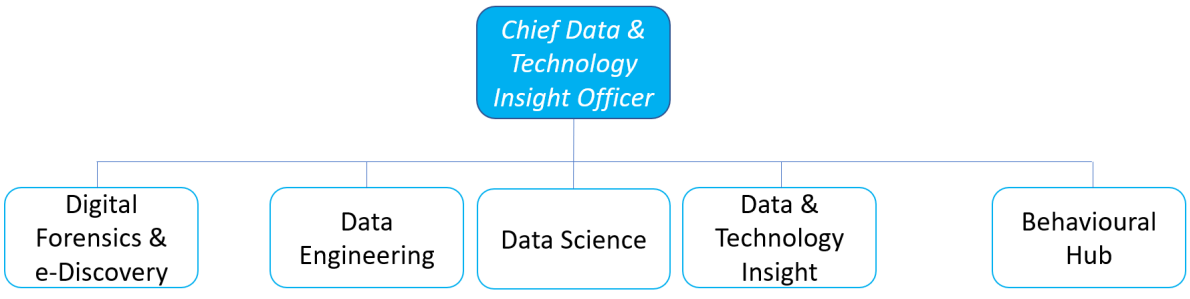
3.2 Unit structure and recruitment

Where should a data unit be placed in an organisation?

At the CMA, prior to the launch of the DaTA unit, the economists would typically lead on technology insight and data analysis. The DaTA unit was formed with its head (me) reporting into the Chief Economist. For four years, this way of organising has worked well for us. Embedding into an existing team reduced the overhead needed to establish an entirely separate division or directorate. Economics at the CMA is still considerably larger than DaTA, and economists are much more hands on and embedded into case teams. This helps us to navigate the wider organisation and target our engagement with the many new pipeline cases on those that are a better match for our skills. Although our core work is largely complementary to economics, some roles remain where our skills can be substitutes, so close working allows both economists and technologists to draw on each other's resource when needed. The shared quantitative nature of professions also means economics is a natural 'home' for technologists with like-minded colleagues.

How should a unit be structured?

There are two primary ways: one is that an agency could structure its data unit around the functions that they need and have multi-disciplinary teams, e.g. intelligence or remedies; another is to have teams within the data unit based on single skills, and as needed put together multi-disciplinary cross-team projects. At the CMA, we have taken the second option, with a data science team, data engineering team, data and technology insight, behavioural insight, eDiscovery and a digital forensics and intelligence service (and operations, a project management team not shown).



At first, we did not have a technology insight team, but quickly realised that we needed to create one. There were specific technical aspects of cases that were qualitative and that technologists were best suited to lead on, e.g. understanding

online tracking technologies. With our casework addressing important societal issues, such as Google's and Facebook's roles in digital advertising, we implicitly assumed that our data scientists and engineers would be interested to embed into these cases. But it quickly became clear that most, understandably, wanted to use their core skills: working with data and coding and did not want to do work that was primarily qualitative. We needed a team of people with technical backgrounds with a deep interest in data and technology but interested in the issues at hand sufficiently that they wanted to work on those issues directly. It works well to have technology insight situated next to data science and engineering, who often have deeper knowledge about technology. All DaTA unit members have a deep interest in the data and technology and technology insight can draw on data science and engineering as needed.

How best to hire and retain technologists?

Other agencies may be surprised and pleased to hear that hiring talented staff has been somewhat easier than we expected. Not all we have learned in the UK may be relevant elsewhere – the ability to hire is going to depend on specific factors for each agency – but some of it will be. We developed marketing material that emphasises the important societal challenges of regulating big tech. For our initial hiring rounds we focused on hiring many people at one time. Not only were people motivated by the chance to be part of a new and growing team, but we benefited from increasing returns to scale in hiring: going out with many different roles at the same time meant we were more likely to have a role that suited each person that saw our campaign and we could generate more interest from our recruitment campaign. We also created interview material that used objective measures of assessment for tasks core to each role – especially coding tests whenever possible – given that research on recruitment suggests that these are the best measures of future performance in the job.

One element that I am often asked about is how we can compete versus big tech firms (and small ones as well). The most important aspect of the jobs that we have to offer is our mission: the CMA exists to protect UK consumers and to serve the public interest. We have found that the mission, when well-articulated, is effective at attracting talented people, especially at more junior stages of their careers. Working for an agency appeals to many people's sense of responsibility as citizens: the cases and policy we deal with shape the economic landscape and are of critical national importance. Agencies also have strong information gathering powers to compel information from players across industries, to get data that no other organisations can get, which can make for attractive, exciting work for technologists. And if we approach elite academics, they are mostly keen to work with us, which many technologists like.

As an aside, while we wouldn't expect regulators to keep up to speed with big tech, we will not necessarily be so far behind in terms of methodological prowess. I have a good sense of the gap as many of my former colleagues from financial regulation now work in big tech. More the issue with keeping up is domain-specific knowledge and the increasing asymmetry of information between firms and regulators. With investment from regulators in data units, good hiring and the in-built advantages that agencies have, we can do well at bridging the gap.

With regards to retention, unlike competition lawyers or competition economists, our technologists mostly do not expect careers in competition, but rather see themselves as professionals in their discipline (although we are working on establishing medium-to long-run career paths). We make sure that our teams focus on the issues that each discipline cares about. This aids the development of technical excellence and, we hope, staff retention. As for all other questions, there is no clear right way and we expect to learn from other agencies who have tried different approaches.

There have been some challenges in hiring. We have found it hard to hire those with sufficiently technical knowledge and skills alongside a desire to do more qualitative work and an ability to quickly grasp competition law and economics. We have refined our recruitment material to emphasise these skills and desires, not only for technology insight but also (though with less emphasis) for data science and engineering.

For roles that include more management, we have found it difficult as well to find the combination of technical knowledge, ability to build an understanding of competition and consumer protection, and leadership and management. This is despite some of our roles such as the recent Director of Algorithm Assessment and Technology Insight receiving very large amounts of interest. There is a thin labour market for the combination of these skills, capabilities and interests.

Where should the boundaries with other teams lie?

Important boundary issues are with IT, intelligence, policy and market research. With IT, the question is whether some or even all of digital development work (or even some IT operations work for servicing and maintain the bespoke tools developed) should sit within the data unit. It has worked for us having DevOps for our tools in DaTA, but it could move to IT at a later date and we continue to actively consider the best location. There are also questions of how to organise traditionally manual intelligence activity (we plan to help them through tool development but keep separate), how much policy should technology insight and others do (some but not too much), and how should market research and behavioural science work together (we think actively). We have largely operated with little interaction with the teams that do econometrics or commission surveys, but I would hope over time that there would be more overlap.

3.3 Prioritisation

Which of the five roles should a data unit prioritise?

There is no doubt, for me, that as it stands the highest return per person is from the unit's expert data and technology advice. There are several cases where a small amount of technology insight, data science or engineering resource – especially technology insight – has enabled whole teams to move much faster or where the direction of the case has been significantly changed. Big data handling, data science for cases, scraping and behavioural insight also have high total impact and have direct impact for cases, but are more resource intensive. Horizon scanning and emerging technology work has a less immediate impact on cases, but carefully designed should bring the CMA good less-interventionist options.

What is the right allocation between immediate case impact and longer-term innovation?

Longer-term innovation work – pipelines, data-driven tool development – requires significant investment from other parts of the organisation to be effective at driving day-to-day work. It takes time to build these assets and data scientists and engineers need regular feedback from 'clients' when doing so, to create outputs that really deliver for case teams. For organisations that are typically focused on case delivery, with immediate deadlines and deliverables, making time for longer-term investment can be tricky. Our colleagues are busy. I return to how best to organise this activity in the next subsection.

Nonetheless I am convinced that as we see more innovation outputs we will see a significant impact, as we are starting to see in document review.

3.4 Organisation for effective delivery

How can data units organise to deliver?

We spend considerable time engaging with other teams responsible for the CMA's pipeline of new cases and assessing the potential role for the DaTA unit's different skills. Our engagement with different cases varies widely, some cases – even digital cases – are focused on a small number of contracts between firms and there is no data, tech or behavioural role. In other cases, again even digital ones, technology may not be much of an issue, but how consumers might respond to a remedy is a key consideration. Some cases may also benefit more from getting and analysing large data sets than others. Technology insight normally leads on the broad DaTA unit interaction with case teams, shaping up roles for data science, engineering or behavioural insight, but if it is clear initially other teams can be involved from the

beginning. Digital forensics, document review and eDiscovery on a case-by-case basis operate reasonably separately with established demand for these skills as a non-embedded service function for cases.

An important decision is whether to embed team members into a case – where they attend all or most of the key case meetings, internally and with parties – or have a more distanced advisory role. Embedding typically requires team members spending much more time getting into the details of a case and can be time consuming. But of course, it means that DaTA unit team members have a much broader understanding of the case. Sometimes there are clear and distinct roles – such as in the Meta/Giphy merger and the focus on the data aspects of a vertical foreclosure theory of harm – so that technologists do not need to embed. If we believe a case will significantly benefit from embedding technologists then we organise ourselves in that way, but we typically try not to, both to get the most from our resource and to keep team members focused on using their core skills, which tends to be better for their interest and motivation. The DaTA unit's prioritisation rubric explicitly covers making sure work is interesting, which is important for retention.

In sum, it took some time to develop and adjust our operational processes and modus operandi for engagement. We meet regularly with and work closely with teams across the organisation to adjust our approach as we learn. We determine our strategy for engaging with case teams depending on the attributes of the case, the match to the skillsets of the DaTA unit, the substitutability of economist or other resource and obviously overall case priority.

What operational processes allow for more effective delivery?

The DaTA unit has always had an operations team staffed with project managers. A talented data scientist and former employee once told me 'data scientists aren't good at project management and they don't like to do it, don't make them'. He advised organising like tech firms and separating project management. That is what we have done at the DaTA unit.

We have instituted a version of 'agile' project management: an iterative approach that helps teams deliver faster. Instead of planning many weeks into the future, agile teams focus on shorter-run deliverables – we focus on planning in two-week 'sprints' – and smaller increments, re-assessing at the end of one sprint and then planning for the next one.

My current assessment is that for our longer-term innovation work, we have not been sufficiently agile and could be more so. We need to make two changes. First, we should focus on clearer and more stretching deliverables. Second, we should make sure to get our products in front of our internal clients earlier – so in a more

rudimentary form – and more frequently. Team members can then adjust to client needs better.

3.5 Technology requirements

How important is access to the cloud?

Access to cloud services, as opposed to on-premises computing, is a crucial part of putting together an effective data unit. Data scientists and engineers will often want access to large amounts of computing power and need access to the latest services. There are many resources on publicly available cloud platforms, e.g. for creating data pipelines – for making sure that data is cleaned efficiently, that data is available quickly and that data processing is robust.

How should IT be structured for technologists?

Data scientists and other analysts will need access to the latest packages for their programming, GitHub and other resources. They cannot effectively work if there are cumbersome IT processes that prevent them from getting the tools for their job. At the DaTA unit we created and maintain our own platform for our team members to access the cloud and all the services they need.

For the inter-agency collaboration mentioned in this paper to be effective, especially developing tools together, agencies will need some type of shared access to cloud infrastructure and Github repositories.

What are other important technology considerations?

There are marked benefits over time of having high-quality codebases. For example, I discussed above, the code libraries that the DaTA unit has compiled for scraping and natural language processing. To make sure that these assets are kept at a sufficiently high quality, data units need to spend enough time creating and maintaining thorough documentation. An important benefit from this is that new team members can more easily get up to speed and able to use and develop the code.

4. How will competition and consumer agencies be affected?

This paper has focused on what data units have done so far. Data units have driven efficiency, both through digital transformation tools like the Evidence Submission Portal, and accelerating cases that require grappling with the technical details of markets. And they have improved the quality of agency outputs, for example, distinct case pipeline ideas, improved analytical outputs or further-refined remedies.

But to understand the potential impact of technology-led transformation on competition and consumer agencies fully, we need to look forwards, incorporate sufficient vision and think about what more data units could do to drive change. It is apposite to mention again that these are my views and not those of the CMA.

Building deeper data capability

While I have no doubt that our current data capabilities provide considerable benefits to the CMA's work and we should continue to do this work, in my view there is the potential for even greater, more transformative impact.

Big data handling, data science and experiments for cases are potential activities with high impact for consumers. We observe that technology firms, when they want to understand the world and how to provide experiences for consumers, use massive granular datasets and quantitative methods copiously. These are fundamentally data-driven organisations, which believe that you learn about the real world by analysing data or testing. We have the necessary capabilities and have started to use some of these quantitative tools, but we could do more to take advantage of the opportunity at scale across our work.

I am somewhat of a data nerd (you would hope given the technical teams that I lead). My beliefs about specific markets have often been most shifted by the results of granular data work (especially when using methods that allow us to infer how different market factors causally impact consumers or firms) or from field experiments (including A/B tests) as opposed to qualitative information or high-level descriptive statistics.

In my previous role at the UK's financial regulator, the Financial Conduct Authority (FCA), my teams often worked with very large datasets with granular information on consumer behaviour and we used techniques that allowed us to infer the causal impact of firm behaviour on consumers. We also regularly used field experiments to test remedies, often involving many tens of thousands of consumers. My colleagues and I learned a lot from this work.

For example, the FCA was told by parliament to put a price cap on high-cost short-term loans (payday loans). The big concern was that people who would be pushed out of the market for these loans – those that had previously been only marginally profitable and would be unprofitable with lower prices – would be worse off, with the most concern being that they would turn to illegal lending ('loan sharks'), with very negative consequences. By using granular data and empirical 'causal identification' techniques using historical data on marginally profitable and unprofitable consumers, we showed rigorously that it is was most likely that these consumers would be much

better off from stopping having access to these loans.[56] The empirical research shone a light on this important issue involving vulnerable people and supported bolder decision-making and a tighter price cap. As another example, we learned from field experiments that many remedies that involved giving information to consumers had small effects or were completely ineffective.[57] While many people had expressed concerns about using these remedies alone, in practice the FCA still did. The experiments that we ran supported, again, bolder interventions in the interests of consumers.

This work does have some drawbacks. It requires a considerable amount of resource, can be slow and often the picture that it paints is quite complicated, with some evidence pointing in one direction and some evidence in another. The real world is messy. And it often does not conform to even well-informed conceptions.

Messy data can be hard to explain in a neat way for a non-specialist – for example a judge reviewing a decision on appeal. Doing something new undoubtedly creates challenges and risks. But like every other part of our case, the data analysis needs to be clear enough and explicable enough to stand up to that sort of scrutiny. I wholeheartedly believe that this kind of data work would allow competition and consumer authorities to develop remedies that have greater impact in the interests of consumers. I believe we should, carefully, integrate more of this work into our cases.

It may be that the legal frameworks for competition and consumer protection will need to continue to develop as data science provides further evidence to support agency's decisions – in the same way that it is has incorporated many economic concepts and methods, for example. As this process progresses, the use of data science is likely to be a more normalised element of the overall competition and consumer toolkit, which would also allow the agencies to draw more on the skills of data scientists and engineers in new and innovative ways.

Other activities that I believe could be expanded with impact are data-driven tool development and data pipelines. The DaTA team have built a few tools and pipelines, but there is the potential to do considerably more and make the CMA's work more founded on regular sources of data and provide much improved tooling to

---

[56] John Gathergood, Benedict Guttman-Kenney, Stefan Hunt, How Do Payday Loans Affect Borrowers? Evidence from the U.K. Market, *The Review of Financial Studies*, Volume 32, Issue 2, February 2019, Pages 496–523

[57] For example, see Paul Adams, Robert Baker, Stefan Hunt, Darragh Kelly and Alessandro Nava (2015). Encouraging consumers to act at renewal: Evidence from field trials in the home and motor insurance markets, *FCA Occasional Paper Series*, No. 12, or Paul Adams, Stefan Hunt, Christopher Palmer and Redis Zaliauskas Attention, Search and Switching: Evidence on Mandated Disclosure from the Savings Market, *FCA Occasional Paper Series*, No. 10 or Paul Adams, Benedict Guttman-Kenney, Lucy Hayes, Stefan Hunt, David Laibson and Neil Stewart (2018) The semblance of success in nudging consumers to pay down credit card debt, *FCA Occasional Paper Series*, No. 45

make people's jobs easier. But it will require ongoing investment and engagement with the lawyers and economists, who will benefit over time, as discussed.

Where might there be under-exploited opportunities?

An area where there is likely significant opportunity, in my view, is the integration of data science into empirical methods to analyse competition problems. The frontier of academic economic research uses a wide variety of non-traditional data sources and incorporates different types of machine learning.[58]

However, developing these new methods is difficult and requires three inputs. First, it requires a discerning understanding of existing methods, their theoretical underpinnings and their drawbacks. Second, it requires good awareness and reasonable understanding of new and different data sources available and machine learning techniques, probably through closely following the existing research literatures. Third, it requires sufficient time to engage with case teams at an early stage and to form analysis that will deliver within case constraints and is not overly legally risky. It is hard to find people or teams that cover all three of these inputs.

I nonetheless expect that we will see innovation in methods that lie in the intersection of data science and econometrics, possibly from newly minted economics or data science PhDs that join competition agencies.

More generally, my view is that the greatest opportunities for improving competition and consumer protection analytical methods are likely to come from the machine learning used practically in technology firms – which are at a high-level covered in this paper, though I do not get into specific techniques (e.g. different types of neural nets, gradient boosted trees etc). I have seen suggestions that agencies should use agent based modelling or other techniques. While obviously these techniques have good uses (e.g. modelling biological ecosystems), there are good reasons they are not used, yet, for practical analysis of markets (such as the difficulty currently of making somewhat accurate assumptions on how people will react to firms' behaviour).

---

[58] See for example example the Compass Lexecon work discussed in Section 2.2. Or Morozov I, Seiler S, Dong X, Hou Let al., 2021, Estimation of preference heterogeneity in markets with costly search, Marketing Science, Vol: 40, Pages: 871-899 on the use of consumer search behaviour on a retailer's website for estimating preferences. Or Rob Donnelly, Francisco R. Ruiz, David Blei, Susan Athey (2019) Counterfactual Inference for Consumer Choice Across Many Product Categories for the use of techniques from the machine learning literature on matrix factorisation.

Bringing data and technology input into the mainstream

I am sometimes asked whether data units are just needed now given the focus on digital cases and whether the need will subsequently diminish. I do not believe that is right. Innovation across all sectors of the economy, in start-ups or within large firms, heavily involves data and technology. The trend is that firms in all sectors are becoming digital. Even in a world where ex-ante regulation of digital players was in a separate entity to a national competition authority, the authority would need digital skills, as almost all organisations now do. I expect technologists to remain within agencies and become a mainstream function.

I see the way that the DaTA unit engages with cases evolving. I expect technologists to have a more integral role over time and increasingly shape the analysis of cases. The CMA's DaTA unit has roughly 25 team members that engage with cases compared to roughly 100 economists who do. Especially as the CMA has a large portfolio of digital cases right now, we are stretched across the cases. We are involved in almost all the digital cases and must allocate our resource carefully, with most cases having only a small amount of input and only two or three having more than that. Our team members are also new to competition and consumer protection work and we continue to develop efficient and effective onboarding and training for colleagues in competition and consumer protection policy. The combination of stretch and need to acquire domain knowledge mean that we have only so much capability to influence the analytical design early on in a case.

Going forward as team members gain more experience and with some reduction in stretch, I expect the DaTA unit to increasingly work with economists and other colleagues to shape the direction of cases.

As part of the mainstreaming of data unit, we will continue to work on creating career paths and establishing competition technologist as a profession, alongside competition lawyer and competition economist.

Potential game changers

There are at least a couple of impacts from data units that might be described as potential game changers for how competition and consumer agencies work. First, a substantial part of what data units do is to code. The marginal cost of copying code is close to zero. The problems that agencies around the world face are very similar, and with multinational firms we are often monitoring, regulating and tackling the same or similar behaviour.

To the extent that agencies can share code with each other – for data pipelines, scraping, software/ tools, analysis – agencies can benefit from some of the same digital forces that are reshaping markets. Imagine if for every piece of code that we

wrote, we could find one other commensurate-sized agency that could use that code, and that the transfer of code and know-how is costless (not true, but costs can be low). That would be a doubling of the return, and potentially the returns could be much higher. And where possible some code (probably not code used for detecting non-compliance, as this would lead to gaming) could be open-sourced. Open-source approaches aid transparency and sharing, whether fully open-source or not, aids repeatability and enable errors to be more quickly identified and corrected. The benefits to international collaboration in the data and tech space have the potential to be very high, much higher than regular knowledge sharing. These forces mean the use of data science and engineering in agencies is likely to increase significantly as it gets easier for new agencies to use existing codebases and for agencies to develop technology together.

There are also benefits of domestic collaboration with other regulatory agencies. Though use-cases differ considerably more across agencies with different remits, so sharing of code may be more related to more abstract capabilities, such as natural language processing, rather than specific pipelines or tools (though the Companies House pipeline could be useful for several other UK agencies and presumably some document review tools could be relevant across agencies).

Second, competition and consumer agencies have largely had their portfolio of cases driven by external events (e.g. mergers), stakeholder complaints or leniency applications and so the work is quite reactive. This is different to sectoral regulators who regularly receive data from the firms they oversee and monitor this data to prioritise proactively. It's an open question as to how cost-effective monitoring will be for agencies to do in the future, but data units can drastically reduce the costs through scraping. And many issues in markets – e.g. use of concerning online choice architecture practices, mergers of interest to authorities, signals of resale price maintenance – are inherently easily detectable from public information.

In addition, agencies might get hold of non-public data sources as well: one source potentially being from the ex-ante regulation of large technology firms; another is smart grids in infrastructure sectors such as water and energy; and presumably potentially many more. If agencies were to monitor regularly and systematically, that would be a radical change for competition and consumer enforcement and further shift toward the proactive identification of harm and potential cases. Investing in proactive market monitoring would not only create better lead generation but also increase deterrence and potentially allow better assessment of the impact of our work, through the better data leading to improved ex-post assessment.

5. Conclusion

Data units in competition and consumer agencies are new, and we are just beginning to learn how to build the skillsets into our organisations. For firms that deal

with competition and consumer agencies and their advisers, we hope that this paper has been useful for understanding the technological capabilities you can expect from agencies going forward.

For authorities that already have a data unit or are in the process of creating a dedicated unit, there is much opportunity to learn from each other, given the novelty of the roles that data units have. Experimentation – in the broadest sense, of trying new things – yields novel results and this knowledge can be a public good about what works. There is a myriad of different things that can be tried, and at the CMA my team and I have only conceived of some and tried fewer, discussed here. We have already learned a lot from our international and domestic colleagues and know there is much more to learn. In addition, the fact that a large amount of what data units do is reflected in code means that the benefits from collaborating have the potential to be an order of magnitude greater than solely best practice learning.

On 15-16 June 2022, starting the day after the publication of this paper, the CMA hosts its Data, Technology and Analytics conference on new and evolving challenges in the tech industry and digital markets, and how competition and consumer agencies are developing technical capabilities and expertise to tackle these challenges. This is the inaugural event, and the CMA has committed to holding the conference every two years. In addition to hearing from external experts, the audience will hear from several agencies and there will be an additional agency-only day, where we will share our experiences. I hope this paper will be of use to other agencies and, speaking on behalf of the CMA, we look forward to actively working together and learning from each other more.