

CONSULTING



REPORT FOR OFCOM

THE USE OF DATA BY ONLINE SERVICES

MAY 2019

[analysysmason.com](https://www.analysysmason.com)

Contents

1	Executive summary	1
2	Introduction	3
2.1	Background	3
2.2	Scope and objectives of the study	3
2.3	Approach to the study	4
2.4	Structure of this document	5
3	Types of data	6
3.1	The data OSPs hold about users can be either collected from users and third parties, or created by the OSP	6
3.2	The data that OSPs collect can be actively provided by users, observed by the OSP, or received from third parties	7
3.3	Personal data and sensitive data are subject to specific regulation	9
3.4	The value of third-party data is closely linked to whether an OSP derives its revenue from advertising or from subscriptions	10
4	Data-gathering processes	12
4.1	Provided data is a combination of direct user inputs, and permissions granted to access existing data	12
4.2	Observed data is typically collected through telemetry and automated processes, with many processes enabled by mobile devices	14
4.3	OSPs can track users around the Internet using cookies, web beacons or digital fingerprints	14
4.4	Third-party data can be purchased, received through partnerships or through intermediaries	15
5	How data is used	17
5.1	Data is an increasingly central input into companies' operations and efficiency	17
5.2	Data on users enables OSPs to increase the value of their advertising inventory	21
5.3	Data is an important input for AI/ML	23
6	Conclusions	26

This report is the output of a study commissioned by Ofcom but carried out independently by Analysys Mason. The content of this report is based on analysis and research conducted by Analysys Mason and discussions held with industry participants, and does not necessarily represent the view of Ofcom.

For any queries relating to this report please contact the authors David Abecassis and Richard Morgan (david.abecassis@analysysmason.com; richard.morgan@analysysmason.com).

Copyright © 2019. Analysys Mason has produced the information contained herein for Ofcom. The ownership, use and disclosure of this information are subject to the Commercial Terms contained in the contract between Analysys Mason and Ofcom.

Analysys Mason Limited
North West Wing, Bush House
Aldwych
London WC2B 4PJ
UK
Tel: +44 (0)20 7395 9000
london@analysysmason.com
www.analysysmason.com
Registered in England and Wales No. 5177472

1 Executive summary

Online services are becoming increasingly central to how people meet their communications needs, find information and access entertainment. However, as more of our daily interactions move online, this is being accompanied by growing public concern about online privacy and the collection and use of personal data.

This study aims to provide factual evidence on what data is gathered and used by online services and how. We examine the types of data used by online service providers (OSPs), discuss the mechanisms through which it is gathered and processed, and the use cases for data that translate into business benefits for OSPs.

The study has considered these questions through a broad review of publicly available information, supplemented by 19 interviews with global and UK players across six categories of online services:

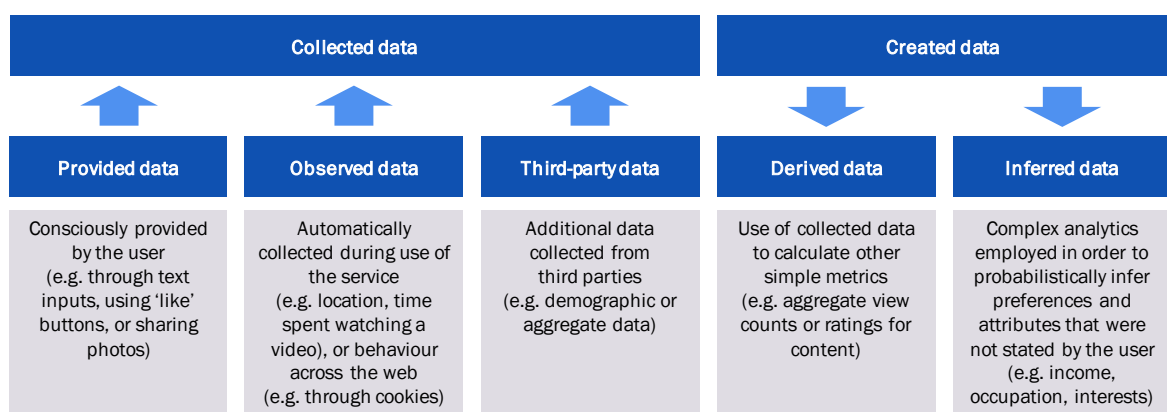
- online advertising
- social networks
- video-sharing platforms
- messaging apps
- news aggregators
- on-demand audiovisual platforms.

For these service categories, the study considers the types of data collected, the data gathering processes, and how the data is used by OSPs in order to support business outcomes. These are summarised in turn below.

Types of data

The data OSPs hold about users can be either collected (provided by users, observed by OSPs, or obtained from third parties), or created by the OSP using calculations or inferences.

Figure 1.1: Taxonomy of data collected and created by OSPs [Source: Analysys Mason, Information Accountability Foundation, 2019]



Interviews with OSPs highlighted that they typically place the most value on observed data that is collected automatically as users interact with the service. This internally-generated data is specific to each OSP, and is key to enabling service improvement and personalisation for individual users.

OSPs using an advertising-funded revenue model are more likely to place high value on data about factors such as demographics and interests, as these are typically used to target advertising. For similar reasons, third-party data is primarily used to support online advertising, as OSPs and other players in the value chain can combine data from multiple sources to create a more detailed profile of users which can increase their value to advertisers.

Data gathering process

In this section we consider the mechanisms used by OSPs to collect data, broadly categorised as:

- Provided data is a combination of direct user inputs, and permissions granted to access existing data
- Observed data is typically collected through telemetry and automated processes, with many processes enabled by mobile devices
- OSPs can track users around the Internet using cookies, web beacons or digital fingerprints
- Third-party data can be purchased, received through partnerships or through intermediaries.

How data is used

The data collected by OSPs is used in a wide variety of ways, and is an increasingly central input into companies' operations and efficiency. These can be broadly categorised under five use cases:

- support for the **design and launch of new products**
- support for **product development and improvement**, particularly observed usage data
- **personalisation of services** to individuals, using data about them and the wider user base
- optimisation of **operational efficiency** and cost reduction
- more **effective marketing** of OSPs' own services

In addition to these use cases, data about users' individual characteristics can enable OSPs to improve advertising revenue opportunity by supporting targeted advertising. OSPs are able to facilitate the targeting of adverts towards users with specific demographic characteristics, interests or behaviours seen as relevant by advertisers, and therefore increase the value of their advertising inventory. Such OSPs often use data from third parties to supplement what they can collect or infer from the use of their services.

Finally, artificial intelligence and machine learning (AI/ML) are being increasingly applied by OSPs in ways that cut across all of the data use cases described above. Interviews with market participants consistently illustrated that AI/ML techniques are seen as valuable tools in harnessing the benefits of large-scale data collection.

2 Introduction

2.1 Background

Online communications services are becoming increasingly central to how people meet their everyday communications needs. They are increasingly used as substitutes for more traditional communications channels such as telecoms and broadcasting. A significant part of these online communications services – including social media, messaging and video-sharing platforms – is funded through revenue generated through online advertising, which is maturing rapidly. At the same time, subscription services are also gaining in popularity, as evidenced by the growth in paid online audiovisual media services alongside free services.

The digital nature of online services enables a great wealth of data to be collected about users and how they interact with services. This data is being used extensively to help improve services and drive innovation, serve targeted advertising, and to develop artificial intelligence and machine learning (AI/ML) techniques.

Ofcom, in its role as communications regulator, has commissioned this study in order to further its understanding of how data is gathered and used by online service providers (OSPs).







2.2 Scope and objectives of the study

This report provides insights into the role of data across six categories of online services specified by Ofcom:

- online advertising
- social networks
- video-sharing platforms
- messaging apps
- news aggregators
- on-demand audiovisual platforms.

Figure 2.1 below provides a brief overview of each service category.

Figure 2.1: Overview of online service categories covered by the study [Source: Analysys Mason, 2019]

Service category	Description	Market structure
Online advertising 	<ul style="list-style-type: none"> • Sale of advertising inventory on online channels (e.g. social media feeds, sponsored search, video platforms, publishers' websites) • Data on online consumers allows far more targeted advertising than traditional channels (e.g. TV and press) 	<ul style="list-style-type: none"> • Led by Google and Facebook, with Amazon gradually increasing its presence • Complex value chain with many data intermediaries, including large data brokers such as The Trade Desk, Acxiom and Experian
Social networks 	<ul style="list-style-type: none"> • Allow users to create online profiles and interact with each other • High levels of user-generated content (UGC), such as posts, blogs, photos and videos 	<ul style="list-style-type: none"> • Consumer market led by Facebook and Instagram, while LinkedIn is focused on business networks • Numerous smaller players such as Twitter, Reddit and Pinterest
Video-sharing platforms 	<ul style="list-style-type: none"> • Enable users to upload, watch and comment on videos • Contain features to help users find content and channels of interest • Feature both UGC and professionally produced content 	<ul style="list-style-type: none"> • Market led by YouTube, with smaller players including Dailymotion and Vimeo • Live streaming increasingly important (e.g. Twitch.tv) • Video increasingly shared on social networks
Messaging apps 	<ul style="list-style-type: none"> • Instant messaging is used for real-time communication and feedback, typically within a closed platform • Email apps allow messaging to other email platforms 	<ul style="list-style-type: none"> • Facebook Messenger, WhatsApp and iMessage are the biggest instant messaging apps, with Gmail and Outlook leading the email market • Snapchat has seen growth stagnate • Emergence of niche, privacy-focused apps like Signal and Threema
News aggregators 	<ul style="list-style-type: none"> • Users can find and access news content from a variety of publishers and news creators • Display content within the service or drive traffic to publisher sites 	<ul style="list-style-type: none"> • Google News and Apple News are the largest aggregators • Users also consume news on social networks such as Facebook (not considered a news aggregator)
Video on demand (VoD) 	<ul style="list-style-type: none"> • Online streaming of professionally produced video content, either through subscription, advertising-funded or through public service broadcasting (PSB) models 	<ul style="list-style-type: none"> • Strong growth among global subscription players (e.g. Netflix, Amazon Prime) • UK players (e.g. BBC iPlayer, All 4 and ITV Player) • Niche players (e.g. Curzon, MUBI, DAZN)

For these services, the study aims to provide factual evidence on how data is gathered and used. We examine the types of data used by OSPs, discuss the mechanisms through which it is gathered and processed, and the use cases for data that translate into business benefits for OSPs.

2.3 Approach to the study

The study has involved a broad review of publicly available information, including published materials from OSPs, news articles and industry commentary, and academic literature.

This research has been supplemented by 19 interviews with players from across the focus categories of online services, which were used to explore and test hypotheses developed through desk research, and to provide direct examples that we draw upon throughout the report. The interviews were conducted on an anonymous basis, but covered a range of both large global players and smaller UK companies. As many of these players operate across multiple categories of online service, the 19 interviews enabled good coverage of each service category, as shown in Figure 2.2 below.

Online service category	Number of relevant interviews
Online advertising	9
Social networks	4
Video-sharing platforms	3
Messaging applications	4
News aggregators	7
Video on demand	6

Figure 2.2: Number of interviews relating to each service category
[Source: Analysys Mason, 2019]

This report focuses on identifying the key issues and messages from the research based on frameworks, drawing on examples from the six focus sectors, and highlighting common themes. It should be noted that many complex technologies and data science techniques are used in the collection, processing and analysis of data; and while the report refers to these techniques in Section 4, it focuses primarily on the operational and commercial aspects of data use.

2.4 Structure of this document

The remainder of this document is laid out as follows:

- Section 3 discusses the types of data used by OSPs
- Section 4 provides an overview of the processes and techniques used to gather data
- Section 5 explores how this data is used by OSPs.

3 Types of data

The types of data used by OSPs in the context of the six online service categories discussed in this study can be considered from four perspectives:

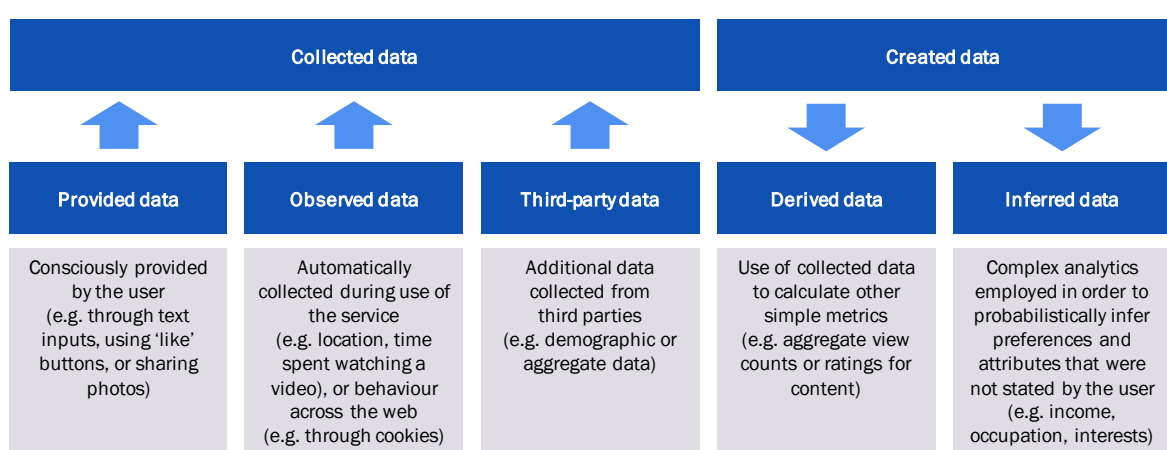
- the data OSPs hold about users can be either collected from users and third parties, or created by the OSP
- the data that OSPs collect can be actively provided by users, observed by the OSP, or received from third parties
- personal data and sensitive data are subject to specific regulation
- the value of third-party data is closely linked to whether an OSP derives its revenue from advertising or from subscriptions.

Each of these points is considered in turn below.

3.1 The data OSPs hold about users can be either collected from users and third parties, or created by the OSP

A taxonomy based on that adopted by the Information Accountability Foundation classifies data as either provided, observed, third-party, derived or inferred. These categories can be grouped into those which the OSP *collects* the data, and those where the OSP *creates* the data, as shown in Figure 3.1 below.

Figure 3.1: Taxonomy of data collected and created by OSPs [Source: Analysys Mason, Information Accountability Foundation, 2019]



Collected data can be based on a range of input mechanisms (broadly provided by users, observed by OSPs, or received from third parties), as discussed further below and in Section 4 on the data-gathering process. Created data relates to how OSPs use and build upon the data they have collected, whereas collected data is raw and provides individual data points that have to be processed and analysed by the OSP to generate insights.

For example, observed data might be an individual data point about a video a user has watched, which in isolation may not provide information that is of value to the OSP. To obtain further insights, the OSP must create its own derived data, such as the aggregate number of view counts of a video for a user or an audience segment, or infer data about the user, such as their preference for a certain kind of video content.

Derived data tends to be based on simple calculations, whereas inferred data is typically probabilistic and based on more complex analysis. Inferred data could include, for example, a user's preference for documentaries based on their viewing history, or an estimate of their income level based on stated interests and purchase history. Both derived and inferred data feed into the data use cases discussed further in Section 5.

Inferred data is most commonly used to support online advertising, whereby inferences are made about a user's interests and characteristics, which OSPs can then use to segment and describe their user base in terms that are relevant to purchasers of advertising inventory. For example, an OSP might use data on a user's location and travel patterns to infer that they are a business traveller, which in turn can be used by a manufacturer of business travel luggage to better target likely customers. However, inferred data is probabilistic, and it is not possible for third parties to verify the accuracy of inferred data about users. Interviews and research have suggested that many players in the advertising value chain are incentivised to maximise scale of campaigns rather than to ensure the accuracy of segmentation characteristics. However, those that place greater emphasis on accuracy are more likely to trust inferred data from larger players (such as Facebook and Google) that have access to more raw data about users.¹

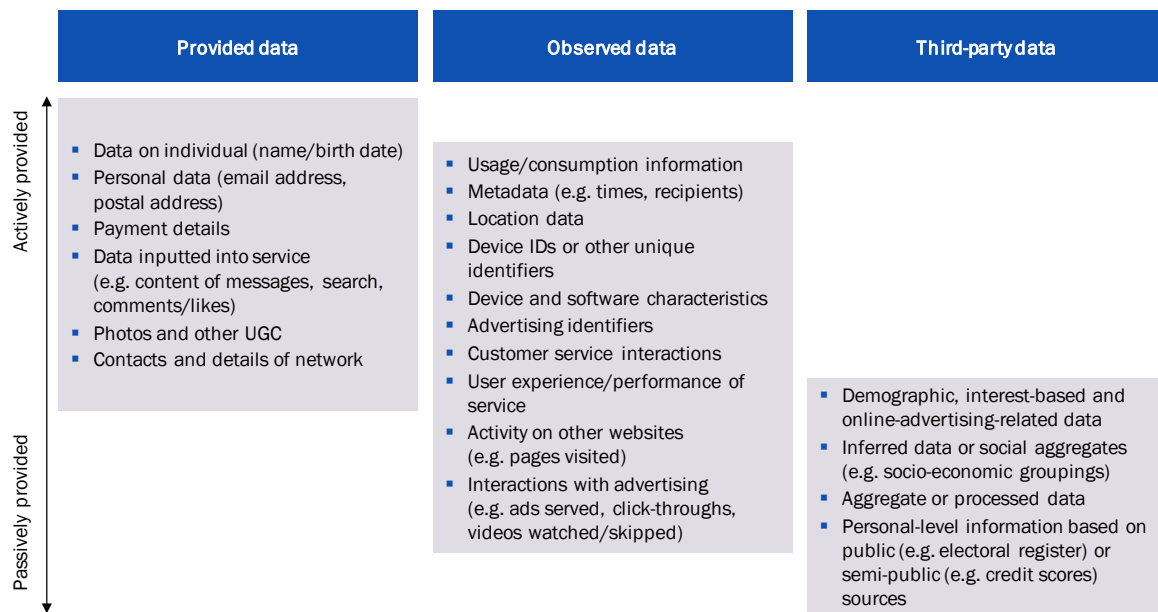
3.2 The data that OSPs collect can be actively provided by users, observed by the OSP, or received from third parties

The categorisation of collected data between provided, observed and third-party data is a common taxonomy adopted by OSPs, as seen from a review of privacy policies from OSPs operating within each of the service categories.² This taxonomy is illustrated in Figure 3.2 below.

¹ See: <https://digiday.com/marketing/data-vendors-struggle-gender>

² While terminology varies, many OSPs adopt the same three broad categories when classifying data. See: Google (<https://policies.google.com/privacy?hl=en#infocollect>); Facebook (<https://www.facebook.com/policy.php>); Instagram (<https://help.instagram.com/155833707900388>); Snap (<https://www.snap.com/en-GB/privacy/privacy-policy/>); Twitter (<https://twitter.com/en/privacy>); Pinterest (<https://policy.pinterest.com/en-gb/privacy-policy/>); Twitch (<https://www.twitch.tv/p/en-gb/legal/privacy-policy/>); Netflix (<https://help.netflix.com/legal/privacy>).

Figure 3.2: Overview of data collected by OSPs by source of data [Source: Analysys Mason, 2019]



This taxonomy distinguishes between:

- Provided data:** information that is provided by the user to an OSP, either through direct inputs or by granting access to data on their devices. For example, many services will collect personal data about a user's identity as part of service registration, or to enable financial transactions. OSPs can also collect data based on the content that is generated by the user. For example, a messaging app can potentially receive and process the content of all of the user's messages (if not encrypted), while a service that allows the sharing of photos can gain access to a user's uploaded images and stored photo gallery to access and analyse metadata on, for instance, location. Once permission has been granted, apps can make extensive use of the features covered by the permission, which the user has to actively withdraw if they wish to no longer provide data to the OSP.
- Observed data:** data that is collected automatically by OSPs when users access the service. Observed data includes a variety of metadata and 'telemetry' data, enabling measurement of various contextual and performance metrics relating to the delivery of a service. Information directly related to the use of services includes, for example, information on time spent using a messaging app, the number of messages sent by a given user, or the time at which a user pauses playback on a particular piece of video content on a video-sharing service. Other information is more indirect or 'passive' in nature, such as location information, device and software characteristics, or data on the performance of a service (e.g. buffering in a VoD service). Beyond this, other data collected automatically can include data from cookies, web beacons and other advertising technology (discussed in Section 4).
- Third-party data:** information an OSP collects from third-party sources, obtained outside of direct interactions with its users. The user has limited input into this type of data and may have

little knowledge of what data is being collected, inferred or shared by third parties.³ The basis of this data may be collected from public sources, such as electoral rolls and land registers, or from semi-public providers such as credit-rating agencies. Combining these various types of data allows third-party data providers to build up profiles on users and aggregate segments, including inferred data such as demographic factors or interests that can be used to support targeted online advertising. For example, a VoD provider such as All 4 might buy data from a data broker to provide insights into the socio-economic make-up of its viewers, to help advertisers target certain segments and thereby increase the value of this inventory.⁴

The typical user has different levels of awareness of OSPs' collection and use of the different types of data. This is largely related to the degree to which the user may have provided the data **actively** or **passively**. Users will typically be more aware of what actively provided information an OSP has access to, such as name, date of birth or gender, while being less aware of data collected automatically or from third parties. In part, this is due to users having a low level of awareness of privacy policies, as well as privacy policies generally being too lengthy and difficult to read for the average consumer.⁵

3.3 Personal data and sensitive data are subject to specific regulation

Much of the provided data discussed in the previous sections is 'personal data' as defined under the EU's General Data Protection Regulation (GDPR). This places specific constraints on how OSPs process this data. Personal data is defined in the GDPR as "*any information that relates to an identified or identifiable natural person*" and can include data that allows the direct or indirect identification of an individual through an identifier such as a name, an ID number, location data, an online identifier or similar. If the information held by an OSP contains personal data, then use or processing of this data is conditional on meeting one of six specific legal justifications, which must be clearly spelled out in privacy policies.⁶

Additionally, personal data is considered 'sensitive' and subject to further specific processing constraints if it includes information relating to racial or ethnic origin, political opinions or memberships, genetic, biometric or other health-related data, as well as data concerning a person's sex life or sexual orientation.⁷ This data can only be processed if the subject has expressed consent

³ Information on the use of data is more likely to be available to users as a result of the General Data Protection Regulation (GDPR), but awareness of what uses their data is being put to is likely to remain low given the complexity of privacy and consent policies.

⁴ See: <https://www.campaignlive.co.uk/article/channel-4-makes-4-platform-100-addressable-programmatic-drive/1450003>

⁵ See: <https://www.varonis.com/blog/gdpr-privacy-policy/>

⁶ For information on the six available lawful bases for processing, see: <https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/>

⁷ Other justifications include the existence of a legal obligation, a medical situation, and the carrying out of a public function. See: <https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/>

explicitly for the OSP to do so, or a set of strict legal conditions has been met.⁸

OSPs can choose to anonymise personal data by removing personal identifiers, both direct and indirect, after which the data no longer falls within the GDPR restrictions and can be used more freely.⁹ By contrast, pseudonymisation refers to the process by which personal data is disguised so that no data is attributable to a unique individual without the use of additional information, such as a key. In these cases, the additional information must be stored separately, and measures put in place to ensure the individual cannot be identified. Pseudonymised data still falls within the ambit of the GDPR, although it helps OSPs decrease the risk of identification and protect the identity of their users when using data internally.¹⁰ The grounds and circumstances in which pseudonymised data can be processed outside of originally envisaged purposes are not yet well defined and are likely to be subject to debate – this is important in the context of data being used for initially unforeseen purposes to support innovation.

The GDPR implies that OSPs cannot use sensitive data on individuals without their express consent. This includes data that may have been inferred by an OSP. An OSP may infer information such as a user's ethnicity, sexual orientation or political affiliation based on their use of the service, but compliance with GDPR prevents this information from being used for purposes such as advertising targeting without the express and specific consent of the user.¹¹ Facebook recently announced that it will no longer allow advertisers to target users using sensitive data such as sexual orientation, although it still allows targeting based on interests in subjects related to sexual orientation.¹²

3.4 The value of third-party data is closely linked to whether an OSP derives its revenue from advertising or from subscriptions

Across the six service categories discussed in this report, the use of third-party data is largely shaped by the extent to which the OSP relies on advertising *versus* subscription revenue. As illustrated in Figure 3.3, OSPs that generate their revenue from online advertising are more likely to gather information on a user's demographic profile, interests and wider online behaviour to increase the attractiveness of the OSP's services for an advertiser. By contrast, subscription-based services use the data generated by users during the use of their services to improve their products and drive internal business improvements (see Section 5).

For example, Google and Microsoft take very different approaches to data. Microsoft's revenue is primarily based on the sale of products and services. It sees itself as a data processor, rather than a controller, and uses data to improve its services rather than treating it as a key revenue driver. Similarly, subscription service Netflix has almost no interest in demographic data as it has found it to be of little use for programming or personalisation and prefers to work with data on revealed

⁸ For more information, see: <https://ico.org.uk/for-organisations/guide-to-law-enforcement-processing-part-3-of-the-dp-act-2018/conditions-for-sensitive-processing/>

⁹ See: <https://www.ucl.ac.uk/legal-services/guidance/gdpr-anonymisation-pseudonymisation>

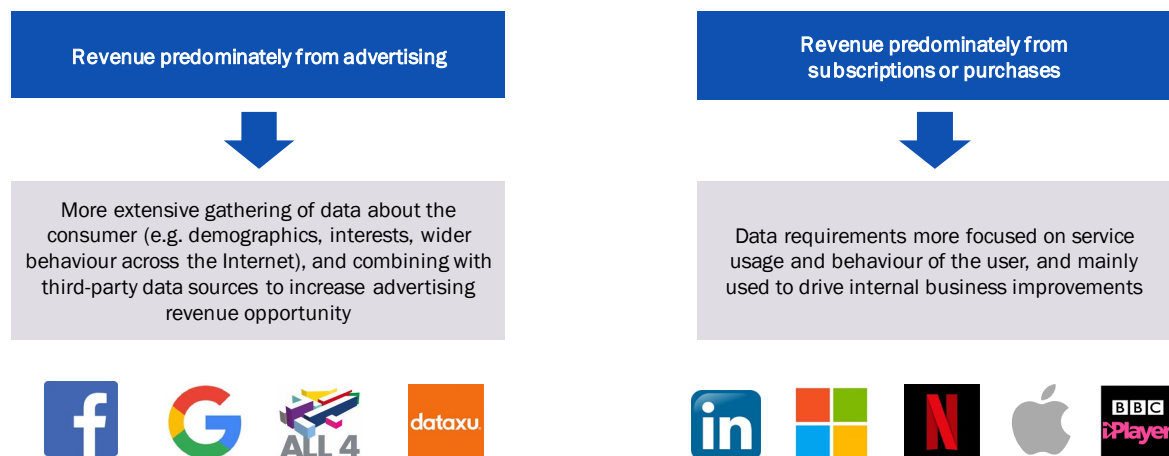
¹⁰ See: <https://www.i-scoop.eu/gdpr/pseudonymization/>

¹¹ See: <https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/>

¹² See: <https://www.buzzfeednews.com/article/alexkanrowitz/facebook-has-blocked-ad-targeting-by-sexual-orientation>

preferences. Channel 4 uses observed user data to drive service personalisation, but also values demographic data to enable targeting of user segments by advertisers.

Figure 3.3: The impact of revenue model on data use [Source: Analysys Mason, 2019]



Major players such as Google and Facebook, whose product offerings run across all six categories, collate all information about a given user into a profile that can be used for ad targeting, collecting user-inputted information on personal data (such as age, date of birth and gender), in addition to data that is collected automatically via their services (such as location data, activities in apps and browsing behaviour).

Large OSPs such as Google and Facebook can develop detailed user profiles based on provided, observed and inferred data, which they can make available to advertisers for targeting purposes. Smaller OSPs can use data provided by third parties (such as data brokers) to enrich the data they hold on their users. Data brokers combine information from various sources (such as voter registration and consumer purchasing data) and use methods such as predictive or inferential modelling to build up a profile of a consumer.

Several interviewees stated that third-party data is mostly used for analyses related to online advertising and has limited usefulness for external applications. However, provided and observed data also play an important role in the advertising value chain, as they allow for the collection of data on the identity of a user as well as behavioural data in response to advertising. Location data can, for example, provide insights on a user's lifestyle preferences, which can be used when classifying users probabilistically into audiences based on shared characteristics against which ads are shown. Tracking technology allows advertisers to observe data on the ads a user has been exposed to, user reactions and conversion rates, which can in turn be used to optimise the types of ads that are served. Research suggests that behaviourally targeted ads can be more effective than ads relying on demographic data, although in practice agencies tend to closely rely on both types of targeting.¹³

¹³ See: <https://hbr.org/2016/04/targeted-ads-dont-just-make-you-more-likely-to-buy-they-can-change-how-you-think-about-yourself>

4 Data-gathering processes

In this section we consider the mechanisms used by OSPs to collect data, broadly categorised as:

- Provided data is a combination of direct user inputs, and permissions granted to access existing data
- Observed data is typically collected through telemetry and automated processes, with many processes enabled by mobile devices
- OSPs can track users around the Internet using cookies, web beacons or digital fingerprints
- Third-party data can be purchased, received through partnerships or through intermediaries.

4.1 Provided data is a combination of direct user inputs, and permissions granted to access existing data

Users of online services provide data to OSPs at various points, including when first downloading apps, registering for services, accepting privacy terms and granting access permissions to the OSP, and then using the services.

Users first interact with a service through a fixed device such as a desktop PC or a smart TV, or a mobile device such as a smartphone or a tablet. Some service categories (such as messaging) are delivered, accessed and consumed primarily through mobile apps; others (notably VoD) are more commonly accessed through larger devices, including desktop computers (in browsers) or connected TVs/set-top boxes (through apps). As fixed services are often shared by multiple users, data derived from fixed services cannot be directly attributed to a single individual; mobile devices typically belong to a single individual, and data on their habits and usage can be tracked directly (e.g. through device IDs and telephone numbers).

In the case of mobile apps, the download of the app is often the first source of user-provided data for the OSP. Some data will be available from the app store (e.g. Google Play or Apple's App Store), such as statistics on searches, page impressions, downloads and technical reports. The app stores provide analytics on this data which can be valuable to smaller app developers that lack in-house analytics capabilities or resources.

Registration allows the gathering of information on the user and provides the means to track the user across devices. When registering for services, OSPs typically require an email address or a mobile number to set up an account, and can also ask for additional information, such as name, date of birth, gender and payment information. Subscription-based services such as Netflix, DAZN or Mubi, by necessity, require that a user registers and logs into an account before they can access and view content. Other advertising-funded services (including YouTube, Instagram, Facebook and Reddit) allow some basic usage through a browser without registering, but encourage users to create accounts to access the services' full functionality. Neither BBC iPlayer nor All 4 relies on

subscription revenue (although the BBC is funded by a licence fee), but both introduced mandatory sign-in to enable the collection of user data and service personalisation.

In addition to setting up an account, use of a service is typically dependent on the acceptance of an OSP's policies and granting its mobile apps permission to access certain features of a user's mobile device. For example, Facebook, Instagram and Snapchat allow a user to take photos within their apps, which requires the user to give their permission for the OSP to access a device's camera and camera roll. Typically, these permissions are only given when a user accesses a feature for the first time, i.e. when they first wish to take a picture or video. However, unless a user revokes a permission, they generally persist indefinitely. Furthermore, apps can also request access to data such as location, address books and app accounts, which gives them access to a broad range of data which can be used to analyse user behaviour. An analysis of the top 100 apps on Google Play and Apple's App Store in 2018 revealed that 45% of Android and 25% of iOS apps requested location data, while 46% and 25% respectively requested access to the camera. On Android, 15% of apps even requested the ability to read SMS messages, a feature which is not available for iOS.¹⁴

Once a user has started interacting with a service, the amount of provided data an OSP can collect depends on the service's functions. Social networks, by their nature, collect a vast amount of actively provided user data, although some social networks are able to collect a broader range of data than others. Facebook's service features enable users to post pictures, videos and live-streams, as well as 'likes' of pages of brands, sports teams and bands, among many other things.

In principle, messaging platforms enable a large volume of user-inputted data, although in practice features such as end-to-end encryption limit many OSPs' ability to gather message content. WhatsApp is only able to collect metadata rather than message content, while Facebook Messenger has much broader access to messages, and states it uses automated systems to scan messages for malware or viruses, as well as comparing photos against databases of child exploitation imagery. Users have the option of enabling encryption to avoid such filtering of their messages.¹⁵ Snapchat collects materially less data than Facebook Messenger, for example, as it does not store and analyse chats, and limits the amount of time for which it stores location data. This is due in part to the ephemeral design of the product whereby most data does not need to be accessed by users in future, in part to the founders' philosophy for the company, and in part to cost considerations.¹⁶

In contrast, services such as audiovisual sharing platforms, news aggregators and VoD platforms mostly allow users to consume rather than generate content and therefore have access to less directly provided data from users. Users can upload content on video-sharing platforms, search for items to consume or leave reviews and comments, however, this directly provided data is far more limited than for a social network such as Facebook.

¹⁴ See: <https://www.symantec.com/blogs/threat-intelligence/mobile-privacy-apps>

¹⁵ See: <https://www.bloomberg.com/news/articles/2018-04-04/facebook-scans-what-you-send-to-other-people-on-messenger-app>

¹⁶ See: <https://www.theguardian.com/technology/2013/nov/13/snapchat-app-sexting-lawsuits-valuation>

4.2 Observed data is typically collected through telemetry and automated processes, with many processes enabled by mobile devices

Observed data is primarily collected through automated processes, enabling OSPs to observe how a service is used. Services are designed using telemetry to collect data such as when a user opens an app, what content they consume and for how long. Much of this data can be collected via desktop browsers, but mobile devices allow for more precise tracking of individual users, as well as the use of additional hardware features.

OSPs can automatically collect information relating to a mobile device that helps identify it and link it to a registered user. These unique identifiers depend on the operating system of the device that the user is using. For example, Google's Android operating system enables app developers to collect hardware identifiers such as SSAID (Android ID) and the International Mobile Equipment Identity (IMEI), neither of which can be reset by the user.¹⁷ Advertising IDs, by contrast, are unique, user-resettable identifiers used by apps that contain advertising to help generate revenue from their apps. Instance IDs are a further means of identifying a user but are short-lived as they are only generated when an app comes online. The lack of user control means that developers are discouraged by Google from using hardware identifiers, and are encouraged to use, for example, Instance IDs to track the preferences of signed-out users, or advertising IDs to track users across different apps or sessions on the same device. In contrast to Google (which allows for greater transfer of data back to its or the OSP's servers), many of Apple's services are designed so that such data is anonymised or stored on the device.¹⁸

Mobile devices contain several technical features that can provide OSPs with detailed data, as long as users provide relevant permissions to apps. For example, accelerometers can help an OSP understand if the user is walking, running or travelling in a vehicle, while the gyroscope can detect whether a handset is being tilted. Magnetometers provide information on the location of the phone in relation to the North Pole, while Global Positioning System (GPS) technology provides information on the location of the phone. How much of these technologies an OSP can use for automated data collection depends on the permissions it obtains from the user. For example, the first time a user wishes to share his location on WhatsApp, the service requests permission, which can either be limited to information collected when the app is running in the foreground or can be more expansive and constantly collect location data.

4.3 OSPs can track users around the Internet using cookies, web beacons or digital fingerprints

Besides these technologies linked to mobile services, OSPs collect data from users when they access websites belonging to parties other than the OSP. The techniques described below are generally used by OSPs involved in online advertising or those conducting their own marketing campaigns.

¹⁷ See: https://developer.android.com/training/articles/user-data-ids#working_with_instance_ids_&_guids

¹⁸ See: <https://gizmodo.com/all-the-ways-your-smartphone-and-its-apps-can-track-you-1821213704>

Cookies are small text files used as identifiers and stored on individuals' devices for a certain amount of time to track usage. When a user visits a site, the server provides the user with a cookie that serves as a form of identification. If the user visits the site again later on, the information contained in the cookie is passed back on to the server, which can then adapt its content to the information it had about the user. Cookies are categorised as being either first- or third-party cookies. A first-party cookie's domain belongs to the OSP, i.e. the site being accessed, whereas a third-party cookie has a domain belonging to a separate entity other than the service provider. First-party cookies usually keep track of log-in details, personal preferences on a website and payment information. Many websites cannot be accessed if first-party cookies are disabled by the user as they provide essential features. However, they can also be used purely for marketing purposes and website performance tracking, functions also performed by third-party cookies.¹⁹ Cookies differ in the length of time over which they are stored on a device. Session cookies are stored in a device's temporary memory and erased when a user closes a web browser, while persistent cookies have an expiration time set by the provider and may last up to two years.²⁰

Web beacons represent short lines of code that enable graphic images to appear on websites and emails.²¹ Also called GIFs or pixel tags, web beacons are usually transparent and small in size (1×1 pixels), so that they are not visible to a user.²² Whenever a website or email is downloaded, the web beacon sends the server a request and collects data about the user such as access time, location or IP address. As with cookies, there are both first-party (domain belonging to the website or email provider) or third-party web beacons.

The rise of ad blockers and other technologies aimed at preventing tracking through cookies or web beacons has given rise to alternative tracking technologies such as fingerprinting, which involves various methods to single out a user or a device and is designed to work even when cookies are disabled. It includes inferring the identity of a specific user or a device based on data such as browser configurations, HTTP header information, installed fonts and add-ins.²³ Fingerprinting can be used for various purposes, such as the evaluation of a website's performance for a user, or adapting the user interface based on the device type.

4.4 Third-party data can be purchased, received through partnerships or through intermediaries

In addition to data collected via their services, OSPs also use data provided by third parties. Third-party data can be obtained from a diverse range of sources including national statistics databases, credit agencies or data brokers. Companies may use aggregate data (such as demographics or

¹⁹ See: <https://www.opentracker.net/article/third-party-cookies-vs-first-party-cookies>

²⁰ See: <https://digiday.com/media/know-cookies-guide-internet-ad-trackers/>

²¹ See: https://www.networkadvertising.org/pdfs/Web_Beacons_11-1-04.pdf

²² See: <https://medium.freecodecamp.org/what-you-should-know-about-web-tracking-and-how-it-affects-your-online-privacy-4293535525>

²³ See: <https://www.twobirds.com/en/news/articles/2015/global/device-fingerprinting>

income) to inform marketing decisions, or rely on data brokers for more granular data. These companies typically gather and combine data on individuals from a variety of sources. In addition to online data obtained via the web tracking systems discussed in the previous section, offline data sources can include court records, magazine subscriptions, vehicle and property records, as well as telephone directories and purchase data.^{24,25} Data brokers such as Experian, Acxiom or Oracle combine data from online and offline sources to build user profiles with data on addresses, income and interests, often without a user's direct knowledge.^{26,27} Companies such as LiveRamp match offline data to online data, using a system of unique authentication records and matched cookies, which allows them to create a single identifier that can be purchased by interested parties.²⁸

In addition to data made freely available and traded by data brokers and others, first-party data on individuals held by two different companies can be matched using unique identifiers, such as email addresses or names. However, the process by which this can occur is complicated and requires high levels of co-operation between both parties. For example, Sky UK owns detailed advertising data, which it wanted to combine with the search and purchasing data of one of its customers in order to create an enriched dataset. In order to ensure that data remained encrypted and no party had first-hand access to proprietary data, the two companies used KCOM as a provider of an escrow environment.²⁹ While interviewees spoke favourably of the mutual benefits that parties could derive from such exchanges, they noted that the requisite high levels of trust and co-operation constituted considerable barriers that were rarely overcome in practice.

²⁴ Offline data refers to data that is not related to interactions with OSPs. This offline data can often be accessed through online channels (e.g. court records). See: <https://crackedlabs.org/en/corporate-surveillance>

²⁵ Englehardt, S. and Narayanan, A. (2016), *Online Tracking: A 1-million-site Measurement and Analysis*. Available at: http://randomwalker.info/publications/OpenWPM_1_million_site_tracking_measurement.pdf

²⁶ Although, technically, users have a right to access and correct the data these companies hold on them, this is sometimes packaged in a way that makes users disclose even more data to brokers. See, for example: <https://www.experian.co.uk/>

²⁷ See, for example: <https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf>

²⁸ See: <https://liveramp.com/discover-identitylink/identitylink-features/identity-graph/>

²⁹ See: <https://www.kcom.com/connected-thinking/opinion/providing-large-scale-secure-escrow-data-processing-for-sky-uk/>

5 How data is used

The data collected by OSPs is used in a wide variety of ways, which we explore in this section: operational efficiency (Section 5.1), advertising (Section 5.2), as an input to develop and apply AI/ML techniques (Section 5.3).

Figure 5.1 below illustrates how each of the data use cases identified and discussed in this section relate to business outcomes. The impact of data use cases on these business outcomes can be either direct (e.g. product development and improvement, which helps to improve customer retention), or indirect (e.g. supporting customer acquisition through word of mouth and network effects). It should be noted that in the case of increasing advertising revenue opportunity, this can be to the detriment of user retention and engagement if advertising is overly intrusive and negatively affects the user experience. The application of AI/ML techniques is not shown in the diagram as it does not directly impact business outcomes, but rather supports the effective use of data in relation to the other use cases identified.

Figure 5.1: Relationship between data use cases and business outcomes [Source: Analysys Mason, 2019]

Data use case		Business outcomes				
		User acquisition	User retention	Increased engagement	Maximise revenue per user	Reduced costs
Improving operations and efficiency	New product design and launch	✓	✓	✓	✓	
	Product development and improvement	✓	✓	✓	✓	
	Service personalisation		✓	✓	✓	
	Operational efficiency		✓	✓		✓
	Marketing of own services	✓				✓
Improving advertising revenue opportunity			✗	✗	✓	

Direct impact
Indirect impact
Negative impact

5.1 Data is an increasingly central input into companies' operations and efficiency

Data uses related to improvements in the efficiency and effectiveness of companies' business can be broadly categorised under the six 'data use cases' discussed in this section:

- support for the **design and launch of new products**
- support for **product development and improvement**, particularly observed usage data
- **personalisation of services** to individuals, using data about them and the wider user base

- optimisation of **operational efficiency** and cost reduction
- more effective **marketing of OSPs' own services**.

Data is used to support the design and launch of new products

The first use case of data for an OSP is during the design and launch of new products, where information on potential users can guide product development and help to launch a service that is already tailored towards the needs of its target audience. Information can be collected from third parties in the form of demographic data on habits, income and other aggregate statistics, while OSPs can also commission market studies and surveys to help them explore markets and discover underserved niches.

Hinge, an online dating app, initially had a similar layout and similar features to those of its main competitors (e.g. Tinder and Bumble) but struggled to gain significant traction. In response to customer complaints about competitors' apps and their effect on online dating culture, Hinge sent surveys to 500 000 of its users, analysed the responses and subsequently re-designed its product in order to better target its core audiences' needs.³⁰

Data supports product development and improvement, particularly observed usage data

Data is used for product development and improvement, in order to retain users and increase engagement, as well as attracting new users. Typically, the data used for these improvements is based on internal, observed behavioural data that is unique to the service, as confirmed by all interviewees. Data-driven product improvements are typically based on data on users' behaviour while using a service, such as the time spent interacting with certain pages or content, or how they navigate around a service. Data can be used to test certain features before full introduction. For example, following a survey of visitors to its platform, Netflix considered enabling non-subscribers to browse the available content before signing up to an account.³¹ However, the results of a test in which this feature was rolled out to a subset of potential users showed that sign-ups were actually reduced, and Netflix decided against introducing the product feature.

The process described above is generally known as an A/B test and is a standard technique used to evaluate product changes between two test groups. It is used by almost all the companies we interviewed as part of this study. In addition to A/B tests, OSPs analyse data to detect common patterns among users and to introduce product features in response. Over-the-top (OTT) sports platform DAZN analysed user data and found that users would unsubscribe from its service during the summer, when fewer sports are broadcast.³² It added a feature allowing users to pause their subscription, locating this option on the same page used to unsubscribe. As a result, the number of subscribers who continued using the service after the summer increased by 140%.³³

³⁰ See: <https://www.vanityfair.com/news/2016/10/hinge-relaunch-swipe-dating-apocalypse>

³¹ See: <https://apptimize.com/blog/2015/11/netflix-registration-ab-test/>

³² See: <https://digiday.com/media/dazn-data-add-retain-subscribers/>

³³ *Ibid.*

Data-driven product improvements can also be based on qualitative feedback from users, who can give the initial impetus to develop features which can then be extensively tested using data-driven approaches. Twitter, for instance, developed features such as '@' (used to address users on the site) and '#' (used to categorise topics) in response to user requests, and then rolled them out following A/B testing.³⁴ It adopted a similar technique when increasing the maximum number of characters per message from 140 to 280. This was based on an analysis of what share of tweets reached the maximum of 140 characters per language, which found that languages such as Japanese required fewer characters than English and were less likely to hit the character limit. In response, Twitter increased the character limit to 280 for a pilot sub-set of its users and then rolled out the changes to the entire site after it determined that users with more characters had received more engagement, received more followers and spent more time on Twitter.³⁵

Services can be personalised to individuals, using data about them and the wider user base

Data is also used to support service personalisation across many online services, in order to promote the most relevant content and features for an individual, improve recommendations, and to increase perceived value by the user. This often involves applying techniques such as A/B testing, cluster analysis and AI/ML in order to increase retention, engagement and revenue per user.³⁶ Observed behavioural data is again central to this use case. It cannot be acquired outside of the platform and can only be generated by an OSP's own users.

YouTube has deployed a sophisticated personalisation system that is able to serve users with content that they were previously unfamiliar with, but that is related to their interests.³⁷ Initially, the OSP experimented with so-called 'Channels', which presented users with content based on their selected subscriptions, but this did not lead to any substantial impact on engagement. However, a large engagement impact was seen when YouTube changed its service's recommendation algorithm from one based on view counts to one based on watch time for each video. Previously, content creators had been able to use misleading video titles with key search terms to gather high numbers of clicks from users, which then helped their videos be recommended to further users. The new focus on watch time promoted content that users wanted to see and deemed of high quality, increasing the time users spent on a particular video. This allowed YouTube to improve engagement and retention of users, and also gained higher advertising revenue as a consequence of the longer amount of time spent watching videos.³⁸

In contrast to the other service categories, news aggregators engage in relatively little personalisation, which is typically clearly labelled for the user. Apple relies on a team of editors to

³⁴ See: <http://www.twentify.com/blog/lessons-to-learn-from-twitter-on-product-development>

³⁵ See: https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html

³⁶ See: https://perso.liris.cnrs.fr/omar.hasan/publications/habegger_2014_bigdata.pdf

³⁷ See: <https://www.theverge.com/2017/8/30/16222850/youtube-google-brain-algorithm-video-recommendation-personalized-feed>

³⁸ See: <http://uk.businessinsider.com/youtube-watch-time-vs-views-2015-7?r=US&IR=T>

select articles from other news sources to feature within its Apple News service. A clearly marked ‘personalised’ section offers items chosen to reflect topics the user has previously read.³⁹ Google News relies on an algorithm to pick the articles it features, giving prominence to authoritative sources and frequently read items. Articles displayed to users do not depend on a user’s search history and do not reflect their preferences.⁴⁰ The lack of personalisation appears to reflect a desire on the part of news aggregators to avoid perceptions of bias or to be seen as cultivating ‘filter bubbles’, which could negatively affect their business in the long run.

Analysis of operational data enables OSPs to increase efficiency and save costs

Data is commonly used across all technology companies to support efforts to improve operational efficiency. Data on the technical inputs required to support service delivery can be analysed in order to improve efficiency and reduce costs, while optimising performance service quality. For example, DAZN has invested in monitoring activity to understand what causes peaks in sign-up activity (notably whether sign-ups before important and popular sports matches increase beyond normal rates) in order to predict and manage the problem of efficiently scaling up server capacity to appropriate levels to meet demand. Similarly, Netflix analyses viewing data to identify and predict the most popular content, and to cache this on servers closer to users. This helps to deliver a better streaming experience while minimising the cost of transporting the same content many times from its core data centres.

OSP can also use data from their operations to help partners improve their services, thus benefitting them indirectly. Facebook has developed a service called ‘Network Insights’, which creates performance maps using data on user location, type of connection and network performance.⁴¹ These performance maps allow mobile operators to “*monitor performance, compare network quality to other operators, and identify trends that can help improve a customer’s experience on the network*”, which in turn can benefit Facebook through increased engagement.

Data on users and their behaviour enables OSPs to market their own services more effectively

Data is also used by OSPs to support the marketing of their own services (as distinct from serving third-party adverts to their users, discussed below). OSPs can more effectively target potential new users both online and offline by using data from their own services and from third parties. Data on the demographic profiles, habits and preferences of existing users can be used to target users with similar characteristics or tastes, or to determine which type of users are not being targeted optimally. Spotify has run campaigns using data on playlists, songs and artists listened to by its users, displaying localised advertising in several of its core markets.⁴² Netflix uses data to optimise its market campaigns using a technique it calls ‘quasi-experiments’, in which it might select two cities

³⁹ See: <https://www.nytimes.com/2018/10/25/technology/apple-news-humans-algorithms.html>

⁴⁰ See: <https://www.bloomberg.com/news/articles/2018-09-04/why-trump-is-suddenly-attacking-google-as-biased-quicktake>

⁴¹ See: <https://code.fb.com/connectivity/announcing-tools-to-help-partners-improve-connectivity/>

⁴² See: <https://www.smartinsights.com/traffic-building-strategy/campaign-of-the-week-how-spotify-showed-the-power-of-data-analytics-in-their-marketing-campaign/>

that appear to be broadly similar along historical demographic and behavioural data and run a marketing campaign using billboards in one city, while using only digital ads in the other. It then compares the results by looking at data on sign-ups or viewing figures from the two cities, which allows it to gauge the effectiveness of each approach.⁴³

5.2 Data on users enables OSPs to increase the value of their advertising inventory

Data can be used to improve advertising revenue opportunity, where OSPs are able to facilitate the targeting of adverts towards users with specific demographic characteristics, interests or behaviours seen as relevant by advertisers. As explained by interviewees in the advertising sector, the value of advertising inventory on publishers' sites increases if more detailed and granular information about the audience is attached to it, as this allows for more targeted advertising. If an OSP can collect more detailed data on a user, such as demographic profile and preferences, as well as information on how that user moves across the web, this data will allow advertisers to produce more targeted adverts and therefore command higher rates on the part of the OSP.

OSP with an advertising-funded revenue model are more likely to be concerned with data 'about the user', rather than simply how they use the services. For example, while Channel 4 could use the predictive power of preference-based data it primarily uses demographic data to market audience segments to advertisers, who tend to develop campaigns based on demographic and socio-economic data.⁴⁴

Wide-ranging data collection across several products and services can provide OSPs with detailed information on user behaviour and broader market trends

Google, Facebook and Amazon are present across multiple services, which allows them to collect and aggregate much more data on a user than a single-service OSP. Beyond the services discussed in the report, these players may derive significant benefits from their presence in areas such as search, mobile software, e-commerce and hardware. Combining the data gathered from these services provides a much richer picture of a user than can be gained from individual services.

Taking the example of Google, the popularity of its Android operating system, search tool and Chrome browser provides it with a wealth of data on users' online behaviour, preferences, location and network characteristics, even when they are not directly using Google apps. Analysis of a mobile handset running Android and an active Chrome browsing session revealed that the phone communicated local information 14 times per hour on average, even when the phone was stationary.⁴⁵ These data-collection capabilities, along with Google's presence across the advertising value chain – through products such as Google Analytics and Google Ad Manager – increase Google's opportunity to track users across the web and record their behaviour. Many OSPs can only gather data from their own services when users engage directly. These products, in addition to

⁴³ See: <https://medium.com/netflix-techblog/quasi-experimentation-at-netflix-566b57d2e362>

⁴⁴ See: <https://www.adweek.com/tv-video/netflix-thrives-by-programming-to-taste-communities-not-demographics/>

⁴⁵ See: <https://digitalcontentnext.org/blog/2018/08/21/google-data-collection-research/>

Google Search, give Google the ability to monitor consumer behaviour and searches, and to promote its services to an existing user base.

OSPs that gather and combine data on users across multiple services can increase their attractiveness as an advertising platform and maximise advertising revenue per user

OSPs that can collect data on users across services are potentially able to enhance the value of the data they keep on each individual user, while also becoming more attractive to advertisers and marketers due to the increased reach of their advertising platform. For example, Google has strengthened its position in the online advertising sector by connecting the data it holds from Google accounts with YouTube accounts. Advertisers can serve highly targeted ads to YouTube users, based on segmentation characteristics derived from interactions with other Google services such as search.

Instagram was able to use Facebook's advertising infrastructure and to put in place measurement tools that allowed advertisers to track the effectiveness of their ads.⁴⁶ Learning from Facebook allowed Instagram to grow rapidly from 200 000 to 500 000 advertisers on its platform between February and September 2016. By comparison, Twitter at that time had little more than 100 000 advertisers despite having spent more time building its advertising platform.

The larger the user base and the more uniquely identifiable pieces of data (e.g. email addresses or mobile phone numbers) are held against each user, the easier it becomes for advertisers to either target existing customers or to find customers with similar demographic or interest-based characteristics. Advertisers can use data they hold on users from previous campaigns and match this against an OSP's user base, or have an OSP find users that are similar in characteristics to existing customers. Facebook, for example, provides a service called 'Lookalike Audiences', which allows advertisers to upload lists of customers, which Facebook then matches to its own users enabling it to then serve adverts to a much wider audience of Facebook users that exhibit similar characteristics.⁴⁷ Interviewees stated that Facebook's ease of use, and its large user base, are attractive to advertisers', whose customer lists can be matched accurately by Facebook to its broad user base.

Companies in the ad tech space, beyond the larger platforms, have launched several initiatives to homogenise systems, lower transaction costs, and increase match rates and audience reach. For example, The Trade Desk, an ad tech player, has launched its Unified ID Solution product, which aims to improve audience matching among ad tech players by offering its global cookie footprint for free to its partners. Early results show that it is achieving match rates of 99%, which will allow media buyers to better reach their audiences on platforms other than those offered by Google and Facebook.⁴⁸ Based on our interviews, further co-operation and consolidation in the ad tech space appears likely, as companies strive to compete with the scale and reach of Google and Facebook.

⁴⁶ See: <http://fortune.com/2016/09/22/instagram-advertising-growth/>

⁴⁷ See: <https://www.facebook.com/business/help/164749007013531>

⁴⁸ See: <https://www.adweek.com/programmatic/the-trade-desk-and-index-exchange-boast-match-rates-of-99-percent/>

5.3 Data is an important input for AI/ML

AI/ML is being increasingly applied by OSPs in ways that cut across all of the data use cases described above. Interviews with market participants consistently illustrated that AI/ML techniques are seen as valuable tools in harnessing the benefits of large-scale data collection, but also that AI/ML can be costly to implement and place significant demands on infrastructure and data engineering capabilities. The term ‘artificial intelligence’ is used here in the broadest sense to refer to simulated intelligence performed by computers and machines, which includes the performance of cognitive functions such as perceiving, reasoning, learning and problem solving. Machine learning is a sub-set of AI and much of the recent interest and innovation in AI is driven by advances in machine-learning techniques.⁴⁹

Below we provide a high-level overview of relevant aspects of AI/ML, under the following headings:

- machine learning comprises supervised, unsupervised and reinforcement learning
- data needs significant preparation before it can be used for AI/ML.

Machine learning comprises supervised, unsupervised and reinforcement learning

Machine learning comprises a variety of methods by which algorithms ‘learn’ to perform functions, with diverse use cases for OSPs ranging from facial recognition or generating email replies to suggesting films for users to watch. Commonly, a distinction is made between supervised, unsupervised and reinforcement learning models.

Supervised learning models rely on training datasets, which contain labelled data and provide an example of a solution to the problem that the algorithm can use to evaluate its own accuracy.⁵⁰ The algorithm gradually improves as it compares its results with the correct labelled outcomes, detects errors and modifies the model accordingly.⁵¹ The trained algorithm is subsequently deployed on unlabelled data.

In contrast, unsupervised learning uses unlabelled data, leaving the algorithm to find key characteristics and distinctions within the dataset and to categorise the data into sets of classes that are not pre-defined. Google fed an unsupervised machine-learning algorithm to 10 million randomly selected YouTube video thumbnails, from which it developed the ability to recognise pictures of cats as a common category within the dataset, despite receiving no information on the distinguishing features of a cat.⁵² Other types of data that lend themselves to unsupervised learning techniques include payment data, where volumes of transactions are so large as to make classification by humans difficult. Unsupervised learning algorithms can find patterns and categorise consumers by certain shared behaviours that would previously have been unknown.

⁴⁹ See: <https://medium.com/mmc-writes/the-fourth-industrial-revolution-a-primer-on-artificial-intelligence-ai-ff5e7ffcae1>

⁵⁰ See: <https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/>

⁵¹ See: <https://www.digitalocean.com/community/tutorials/an-introduction-to-machine-learning>

⁵² See: <https://www.wired.com/2012/06/google-x-neural-network/>

Reinforcement learning is a more general approach than supervised or unsupervised learning, as it is based on learning from trial and error when interacting with the environment to achieve a defined goal and is used when decisions need to be made in an uncertain environment. Reinforcement learning algorithms allow a machine or software to determine the best behaviour in a situation based on feedback from the environment. As such, reinforcement learning allows for dynamic improvement without having to go through repeated training cycles. Reinforcement learning can be seen in the example of a robot learning to walk, with each attempt providing the system with a data point to be incorporated into the reinforcement system. An action such as a large step that leads to a fall will be incorporated by the algorithm and cause it to take a smaller step in the next attempt in order to correct for the previous mistake.⁵³

Data needs significant preparation before it can be used for AI/ML

The challenge of implementing sophisticated AI/ML solutions in a business environment can be illustrated using what is known as the ‘data science hierarchy of needs’ (Figure 5.2).⁵⁴ Before data scientists can get to work on a dataset, the data must be gathered, which requires logging of events, collecting sensor data or gathering UGC. Subsequently, the collected data has to be moved and stored, which requires reliable infrastructure, data pipelines and data storage capacities. Making data available has been highlighted as a key challenge in using data effectively within organisations, with many companies struggling to build the necessary data architecture.⁵⁵

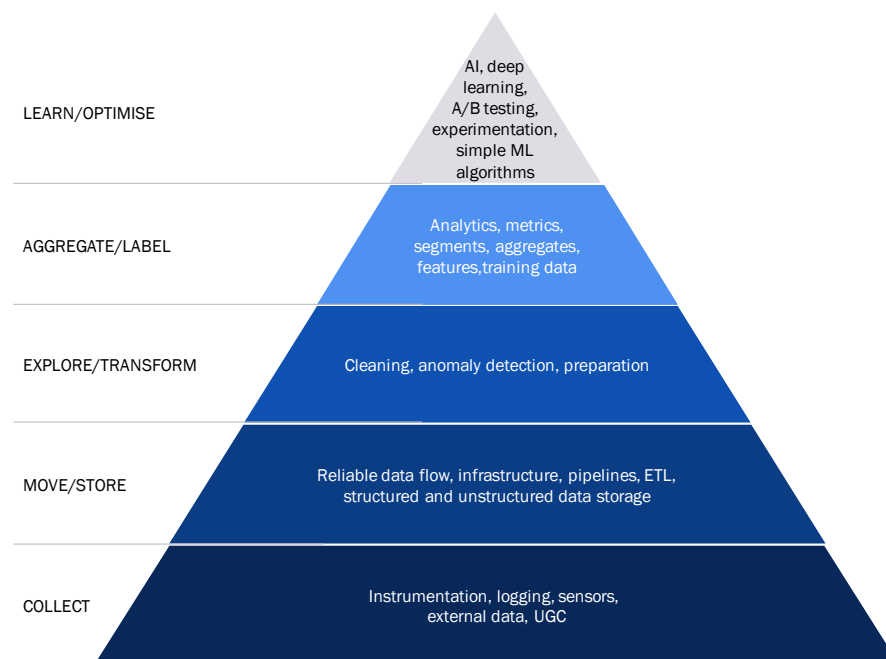


Figure 5.2: The data science hierarchy of needs [Source: Monica Rogati, 2017]

⁵³ See: <https://www.forbes.com/sites/bernardmarr/2018/10/22/artificial-intelligence-whats-the-difference-between-deep-learning-and-reinforcement-learning/#33f9d489271e>

⁵⁴ See: <https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007>

⁵⁵ See: <https://engineering.linkedin.com/distributed-systems/log-what-every-software-engineer-should-know-about-real-time-datas-unifying>

Once the data is readily available and accessible, it has to be ‘cleaned’ (removing erroneous data and ‘noise’) and prepared for further analysis. In a survey of 16 000 data scientists, nearly 50% responded saying that ‘dirty’ data requiring cleaning posed the biggest barrier at work, which large companies with greater resources are better equipped to address.⁵⁶

Following this process, the data can be used for AI/ML. Supervised machine-learning algorithms are the most commonly used, as they are simpler and require less computing power and technical know-how than unsupervised learning. However, they are typically still costly to implement as they require large amounts of labelled data with which to train the algorithm. The labelling of data for supervised learning often must be performed manually, which is a very costly and time-intensive process. Large companies such as Google and Facebook can, in some cases, rely on their user bases to perform tasks such as image labelling (e.g. through tagging of photos or hashtags), which are then used as training data. Smaller companies may not be able to generate training data in this way and have to resort to solutions such as hiring human workers or relying on third parties to label data.⁵⁷

Once data has been cleaned, aggregated and labelled, AI/ML algorithms can be put into place and tested on the data, followed by more complex techniques such as unsupervised learning. The hierarchy of needs describes the large number of pre-requisites that have to be in place before a company is in a position to perform advanced AI/ML techniques.

OSPs with multiple services can spread the costs of investment in data and related capabilities more widely, particularly in the context of AI and ML

The development of AI capabilities is expensive, while successful products can take a long time to be developed and commercialised. Large OSPs have an advantage in that they can invest in costly technology and recover the costs by implementing marginal improvements across a number of their services. For example, in 2014 Google acquired DeepMind for an estimated USD600 million, an AI lab working on the problem of developing general-purpose AI tools.⁵⁸ Google has since been able to use software developed by DeepMind to launch a text-to-speech product called WaveNet, which generates speech mimicking human voice.⁵⁹ Developers can access this functionality via Google Cloud as part of Google’s offering of AI services. DeepMind has also allowed Google to reduce the cost of its data centres by optimising the cooling units.⁶⁰

⁵⁶ See: <https://www.kaggle.com/surveys/2017?utm=cade>

⁵⁷ See: <https://scale.ai/>

⁵⁸ See: <https://www.forbes.com/sites/samshead/2018/04/20/googles-complex-relationship-with-deepmind-gets-exposed/#509a661217d6>

⁵⁹ See: <https://deepmind.com/blog/wavenet-generative-model-raw-audio/>

⁶⁰ See: <http://uk.businessinsider.com/google-is-using-deepminds-ai-to-slash-its-enormous-electricity-bill-2016-7>

6 Conclusions

Online services are becoming increasingly central to how UK consumers meet their needs in relation to communication, information and entertainment. As people spend more time engaging with online services, the level of data available to OSPs is increasing exponentially.

From the research and interviews undertaken two key messages have emerged:

- OSPs place high value on observed data relating to how their services are used
- Data about users' individual characteristics is valuable to OSPs supporting targeted advertising; such OSPs often use data from third parties to supplement what they can collect or infer.

OSPs place high value on observed data relating to how their services are used

Observed data recorded from users' interactions with individual services is seen as central to service improvement and optimisation. Most of this data is collected automatically, in particular via mobile platforms which are strongly linked to individuals, enabling better service personalisation. The value placed on observed data reflects a general preference towards revealed rather than stated preferences from users, as the former is seen as being more accurate.

By supporting incremental product improvements, observed data plays an important role, enabling OSPs to increase user engagement, retain and attract new users, and exploit network effects. Observed data is also at the heart of tracking and improving operational metrics, with many OSPs deploying AI/ML techniques to help improve efficiency.

Data about users' individual characteristics is valuable to OSPs supporting targeted advertising; such OSPs often use data from third parties to supplement what they can collect or infer

Online advertising has a cross-cutting influence, with most service categories containing OSPs that rely on advertising as their primary source of revenue.⁶¹ These OSPs place a significant emphasis (especially compared to those with subscription- or transaction-based revenue models) on developing detailed profiles of users' demographic characteristics and interests, which can be used to support targeted advertising and increase the value of their inventory.

The largest OSPs often have rich user profiles which are built up from a combination of user-provided data, observed data and inferences generated through analytics. Most OSPs, however, choose to supplement their internal data with data from third-party sources, with the vast majority of data sharing between companies being used to support online advertising.

⁶¹ The exception to this is news aggregation, where the largest players (Google News and Apple News) do not make any direct advertising revenue from the services.