Rethinking Data Technical Briefing Phase One: landscaping

What this outline contains

This document briefs on emerging directions in data collection and processing from a technical point of view. Its purpose is to give a flavour of possible areas to explore in a fuller, detailed report.

The document comprises a series of short comments on potential topics that might be of interest to the group. Specifically, we give an intuition as to the nature of particular technologies and trends, and point to common narratives, promises, and some factors relevant for applications.

Again, the aim is to provide a brief primer, rather than a detailed elaboration of technologies and their considerations. This is to initiate a discussion regarding the areas to consider in a detailed, follow-up report[1].

How we discuss broader themes and trends

The first chapter provides a brief perspective on broader data trends, separate from any particular technologies. This to indicate how a follow on report can work to integrate the discussion on particular topics.

How we classify and discuss specific technologies and approaches

We then present our set of briefs across three areas:

---

[1] In focusing on data processing concerns, here we do not consider specific application areas (e.g. autonomous vehicles, emotion recognition, AR/VR, deepfakes, etc.), though we would be happy to if this is deemed of interest.

1. The **sources and contexts** of data collection. This covers emerging types of data, and methods of generating and collecting data.
2. **Models of computation**. This is a high-level view of data processing arrangements. These are general approaches or architectures used in practice, rather than specific technologies.
3. **Building-block technologies**. These represent specific technologies for data processing which are expressly designed in a way that is sensitive or aware of data considerations. In contrast to the general approaches described as "models" above, these are tools and building blocks for constructing and managing systems.

Within these, we list a number of technologies and trends, noting the following:

Summary: a lay description of the technology or trend in question.
Maturity: how close to widespread or otherwise notable use is this technology, and what preconditions must be met for this to occur.
Emerging trends: which of the key themes and trends discussed by the report this technology or setting contributes to, or otherwise interfaces with.
Hype check: a quick critical rundown of the actual benefits and drawbacks of this technology, and how it compares to claims or projections.
The bottom line: potential implications and impact of the technology, and what to look for.

For *sources and contexts*, we will also include:

Novelty: does this setting introduce a significantly new type of data?
Risk impact: how the emergence of this setting mitigates or creates new risks

And for *models of computation* and *building-block technologies*:

Motivation/Context: the problems and historical context from which the technology or model arises.

Themes and Trends

In this chapter we will put forward themes and trends in three areas: the collection of data, its subsequent processing, and power dynamics. These provide the broader context for the specific briefs provided in the report.

Directions in collection

Here we give an outline of how the practice of data collection has shifted as a whole—from deliberate, small-scale, and easily understood to ambient, expansive, and almost intractable in scope.

Pervasive/ambient collection

Sensing technology has proliferated to an extraordinary degree. Not only is potentially every interaction with a computerised system logged, but other ('real-world') phenomena increasingly leave a digital trace. In these cases, the subject of collection is passive, potentially having no direct involvement in or even awareness of the collection. Sensor technology is embedded in people's devices (smartphones, wearables), appliances and machinery (domestic or commercial), and increasingly in the built environment. Indeed, there is a trend of commercial actors being motivated to deploy as many sensors (and connectivity) into their products as possible. Information can be collected on the interactions between individuals or groups, about the physical environment, and about the operation of systems.

Behavioural data

Much of the data gathered on individuals is now records of activity (rather than identifiers or properties—name, favourite colour, etc.), from which properties might be inferred. For example, rather than collecting directly from an individual their preferred genre of film, this insight is generated from the log of all films they have watched.

The pervasiveness of behavioural data seems to be consistently underestimated. A substantial proportion of 'useful' information is now derived, rather than explicitly collected. This development shifts how we should set expectations about the data held on an individual.

Speculative accumulation

The data collected—about an individual or about any other system or phenomenon—has become significantly more granular. Traditionally data collection has been thought of as obtaining information in the form of specific, well-labeled, and intuitively understandable data types—traditionally termed "microdata", e.g. email address, name, etc. However, many collection efforts today focus on obtaining many data points as possible, even those that carry no immediately apparent utility or information in isolation. From a technical perspective, this becomes viable due to the pervasiveness of sensing technologies, low costs of storage, and growing processing capabilities.

The utility of this data emerges from its potential to be linked with other data (see the processing trend ***Linkage*** for more). That is, we have witnessed a shift whereby the value gained from data is not always determined by its quantity or specific informativeness, but by the degree to which it can be linked to other data[2].

Derived data

Derived data refers to data that is not directly collected, from individuals or elsewhere, but produced from summarising, aggregating, or otherwise processing that data. For reasons ranging from data paucity to regulatory friction, such datasets are becoming more attractive for analysis. As this trend progresses, we see the processes by which these derived datasets are produced becoming more sophisticated, and often more opaque.

A key factor to watch for as this trend progresses is the degree to which derived products retain information from the original data (and how well we can even measure this)—particularly given the proliferation of machine learning technologies.

Directions in processing

The infrastructure status quo

As a backdrop for the trends discussed here, note that the current state of data processing is largely cloud-based, with infrastructure provided (as a service) by large companies. Many of these leverage centralised control over geographically distributed datacentres. Generally there is significant consolidation across all levels of the technical and data processing infrastructure, with major firms dominating.

---

[2]Recall Metcalfe's Law about networks, which tells us that the value of a network is proportional to the square of the number of participants. Similarly, the value of a firm's collection of datasets is proportional to how many linkages can be constructed between them.

Linkage

Data doesn't exist in a vacuum. Other data(sets) can be used to form assumptions about what that data is supposed to look like, the statistical distributions from which it may be drawn, and how it may be dependent on or correlated with other information. As data has grown ever more granular, we have become better at uncovering these relationships —even between independently collected datasets—and leveraging them to learn more. The drawing of these relationships is termed *linkage*.

Linkage is used to provide more information, through the fusing of datasets, or to improve the interpretation of data—using accessory data as "supporting evidence" to increase confidence in statements or inferences being made.

Data-driven exploration

Again, data is often collected without a clear prior expectation of what its utility or informativeness night be. This is partly because processing techniques such as machine learning now allow automated (or semi-automated) 'insight' generation—interdependencies and correlations can be drawn from data; i.e. to indicate to an analyst where they might look and what they might find.

Such an approach can raise concerns, where sensitive information can be derived from seemingly innocuous data (even data that was supposedly 'scrubbed', but underlying dependencies remain), where biases in the insight process compound into biased analyses and outcomes, and so forth.

Interoperability

This is a topic gaining prominence, as a potential way of tackling the anti-competitive practice and monopolies of the tech industry. Interoperability can take many forms; the most commonly discussed are the portability of data, and mandating that particular services allow direct interaction with others. Note, however, that at a technical level realising such interoperability can be complex and challenging, and is an area requiring further discussion and consideration.

Power dynamics

Inverting the effort to disclose

As shown by the trends we highlighted earlier, data collection has moved beyond individuals disclosing particular information, to the passive collection of (sometimes

inscrutable) data. The opacity and complexity of technical and organisational processes often render it difficult for individuals to know what data has been collected, or could be collected, and how this can be quantified—let alone prevented. Where the subject of data collection once had to exert effort to share that data, the pressure has entirely inverted, and they must now exert effort if they wish to withhold it!

Platforms and ecosystems

It is well known that information technologies exhibit strong lock-in and network effects, which allows established players to accrue greater competitive power. More than ever we are seeing the concentration of power around platforms—products and services (or collections thereof) that create an entire ecosystem around them. Importantly we see this occurring at *all levels* of technical infrastructure, not just amongst the more visible, 'user facing' services. Firms create platforms by providing an easy surface for other parties to build atop, usually through service offerings which comprise another firm's 'technical stack'. The platforms then build up sufficient material advantages so unassailable that they become the *de facto* arbiters of market dynamics. This gives platform owners tremendous power to influence which technologies are developed—deciding which ones find traction and how they are used—and creates other concerns regarding monopolistic behaviour (gatekeeping, issues regarding dominating adjacent markets, and so on).

Locus of computation: centralisation vs. decentralisation and other tensions

Throughout the history of computing, the locus of computation has often ebbed and flowed between centralised infrastructure and the edge—devices directly operated by the user, or otherwise out 'in the wild'. These dynamics are driven (in part) by the alternating commodification of user and server hardware (think the shift from mainframes to personal computers, and later from personal computers to the cloud). And as the locus of computation moves, so too does data and where parties place trust.

Where we are next headed on this centralisation/decentralisation tradeoff will (partly) depend on the viability of models and tech discussed in this report. The increase of computational power at the edge—smartphones, IoT, etc.—suggests that we are headed towards more decentralised models of computation. On the other hand, this is countered by the non-linear accrual of market and analytic power that firms have recently enjoyed by centralising data.

Capability is the new data

The relative power of two firms is no longer a matter of how much data they accumulate, but what they can do with it. A dataset is more valuable in the hands of a party with the most sophisticated algorithms for extracting insights from it. Parties that can bring more computational power and expertise to bear (think companies with large datacentres, access to top talent, and other resources) have an advantage.

This trend is reinforced by the increased importance of linkage to data processing. As data is more available than ever before, and its value is realised through linkage to other datasets, the parties with the expertise or technological capacity to construct more links between datasets and aggregate them at scale have a distinct advantage. Firms like Google and Facebook have vast datasets, but these datasets are not valuable simply because of their scale, but also how well they are linked, understood, and can be manipulated.

Sources and Contexts

In this chapter we will present a collection of emerging contexts regarding data sources and collection—starting with contexts with more pervasive data capture, narrowing to those where capture is more targeted and (potentially) intrusive. We also highlight the existence of data markets and directions in synthetic data.

The Internet of Things (IoT)

We here will discuss a range of settings that are often discussed under the umbrella term "Internet of Things". The term IoT largely refers to networked physical objects—both objects that were already present in an environment, such as "smart" appliances, and new physical sensors and actuators embedded into the environment. These new technologies exist in somewhat varied settings, and so we split our outline of the Internet of Things into three parts: the built environment, industrial systems, and the smart home.

The capabilities of networked physical objects fall into the two broad categories of *sensing* (which perceive aspects of the physical environment) and *actuation* (which materially interact with the environment). IoT devices range from simple, single-function devices to complex devices that perform their own processing and interface with other devices.

The risks introduced by turning more and more systems into "connected" systems accordingly fall into two categories: pervasive systems systems that collect fine-grained data creating privacy and surveillance risks, and harms resulting from the devices' actuation capabilities (i.e. material effects).

Security note: increasing the number of networked objects precipitates a step change in the number of devices connected to the internet globally, and where manufacturers' security practices have been lax IoT devices have been a major target of cybercriminals.

The built environment

Summary: here we discuss networked technology in public or shared spaces (beyond the home). This is a setting currently dominated by sensing, where data is collected either on an individual or an aggregate level by entities—either commercial or public—who in some way manage the physical space.

In the commercial space, one common IoT deployment is in retail settings, enabling the tracking of customers (or potential customers) en masse or person-by-person, and analysing patterns of movement to appraise the performance of shop layouts or gauge individual interest in products. As well as traditional image recognition using video surveillance systems, identification of individuals is performed through facial or gait recognition, or through wireless passive tracking of smartphones. People usually are not aware of the signals they created by moving through and interacting with public environments.

Public authorities already deploy embedded sensor systems at large for a number of purposes, performing city-scale data fusion to gain insights around the management of resources—such as co-ordinating traffic or train service—as well as for law enforcement. The sensor technologies used include those used by commercial entities, but public authorities also have greater opportunities to embed sensors in the built environment such as car-to-infrastructure communication systems.

The smart office falls somewhere between the commercial and public settings, where companies wish to surveil employees for a number of reasons. These might be similar to the public setting where the interest is in resource allocation, or could be more focused on individuals' behaviour, to evaluate a worker's performance.

Maturity: use of pervasive sensing for population-level data collection is growing, as is individual tracking, empowering local authorities—which previously engaged in traditional surveillance methods—to build sophisticated automated data fusion platforms. As data processing capabilities improve and new sensors are introduced, we will certainly see them incorporated into public data collection. Moving forward, we might expect greater degrees of actuation capabilities, where more aspects of the environment 'responds' to occurrences.

Novelty: instrumented physical environments are not particularly new. However, the barriers to integrating technical components into built environments has fallen, and the demand for data (and data-driven insights) means IoT in built environments is increasing in prevalence.

Risk impact: public and commercial institutions' appetite for data, as mentioned above, is ever growing. However many data collection activities in public settings are unknown to the subject, let alone the outcomes of that collection. And as sensor technologies improve and the built environment becomes more embedded with those sensors, the data collected will become increasingly invasive. This is clearly poses the risk of broad-brush surveillance—levels of monitoring previously only feasible for individuals can now be scaled to larger populations.

Hype check: data collection in public is becoming more generalised—instead of simply CCTV footage many sorts of small data points are collected en masse. This reduces overhead for surveillance or tracking, as instead of actively following an individual, information can be mined at scale from linked datasets.

The bottom line: granular data collection in public is pervasive and increasingly available, and the definition of surveillance must shift accordingly. As the scale of datasets and the degree to which linkage can be performed increases, even small actions that leave small traces can add up to detailed pictures of populations or individuals.

Industrial systems

Summary: Industrial Control Systems (ICS) are a family of technologies that are becoming increasingly networked, and so have ended up discussed under the umbrella of the Internet of Things. These systems range from low complexity, such as sensors in a factory line, to complex systems that perform significant processing and perform complex tasks, such as nuclear fuel processing controllers. There are significant efficiency gains from establishing more comprehensive real-time data collection in industrial processes, as well as automating the management of those systems.

Maturity: the world of ICS can be slow-moving and is littered with legacy systems, due to their large number of interconnected parts with embedded computer systems that are difficult and expensive to get to, let alone update. This means that developments in security can take a long time to fully take hold in ICS networks. Furthermore, there is significant engineering effort required for any changes, due to requirements for extensive testing. This means that systems, while they are now usually networked, are slow to take up new governance practices, for security and otherwise.

Novelty: connected industrial systems are becoming very popular, driven by a significantly increased appetite for data collection, as well as the ability for remote control and management of industrial processes. We're seeing a proliferation of new wireless communications protocols, as industries push to digitally connect every part of their operations.

Risk impact: there are many risks in ICS. These systems usually consist of expensive machinery, can have significant potential for personal harm, or may be involved in key operations to such a degree that compromise would leak trade secrets. Industrial networks are usually heavily policed and separated ("airgapped") from the wider internet. While true airgapping has always been something of a myth, given the need for software maintenance and equipment upgrades, it has become harder to maintain as more

connected devices mean more potential points of entry into a system that was intended to be isolated.

Hype check: The term IIoT (Industrial Internet of Things) has gained momentum. This, as with any marketing term, means no small part of the hype is driven by exaggeration and over-promising; but the increase in devices running complex networked software is real.

The bottom line: industrial control systems are converging with more general (including consumer) IoT products, and this means that the vaunted engineering rigour of industrial systems is being eroded. As companies chase the business benefits of data-driven insights into their operations, they are poking a thousand tiny holes into their closely guarded ecosystems.

Smart Homes

Summary: it is now common for homes to have "smart" appliances—appliances with a networked component, used to connect to the public internet or other devices within the home. These range from existing appliances with pre-existing use-cases (e.g. light bulbs) to new categories of home device whose primary function fundamentally relies on networking (e.g. smart assistants).

As platforms have arisen and smart home devices become more interoperable, the key perceived benefit of smart home devices is light-touch, "ambient intelligence" experiences, usually through smart assistants. In contrast, consumer demand for more high-involvement data collection-oriented products has not proven to be particularly high, perhaps due to surveillance concerns, as well as perceptions that their functionality is unnecessary.

Maturity: we are past the first wave of "smartening" appliances—consumers have mostly settled on what existing devices they want networked. The companies that control the interface to the smart home, usually via smart assistants, i.e. Apple, Amazon, and Google, have outcompeted the device manufacturers to become the de facto platforms of the smart home. These large players have recently formed a consortium—Project Connected Home over IP—seeking to fully consolidate protocols.

Novelty: the smart home has been a dream for decades. Early systems were highly proprietary and cumbersome—one company would fit out an entire building, and that system was self-contained. The step forward due to the Internet of Things is in heterogeneity and interoperability. The systems are now highly modular, as devices can be low-power and networked with pre-existing technology (Wi-Fi), and protocols are for the most part open to anyone.

Risk impact: significant. Consumer products are easily bought and integrated into a home, while manufacturers have a poor track record for shipping secure systems with robust standards for interoperability and authentication.

Actuators pose an obvious tangible risk, as they directly manipulate the physical environment. Potential harms from unauthorised access are obvious—devices could be caused to operate in unsafe ways, malfunction, or be disabled entirely.

Sensors pose a privacy risk. Their function is to surveil, collecting information about the physical environment and private activity patterns. Beyond the security risks of contributing data about household occupancy, behavioural data in the home can be highly specific to individuals—conferring insights about specific activities, or more intangible factors such as interpersonal relationships.

Due to this pervasive sensing, we are concerned with the exfiltration of this data. The degree to which this happens varies across devices and platforms, with some performing almost all computation on devices inside the home and some reporting back to the cloud frequently. Even Apple, who markets themselves as more privacy-conscious, uploads voice recordings to the cloud for the purposes of improving its machine learning models. Security systems such as Amazon's Ring share data with law enforcement, apparently without the input or sometimes knowledge of the resident, which is a significant privacy and civil liberties concern—made no less alarming by the fact that there is no evidence that "smart home security" products are actually effective in protecting against crime.

Efforts to limit data exfiltration from the smart home is confounded by the heterogeneity of devices in the ecosystem; a smart home may consist of devices from many manufacturers, each with their own software stack powering the device's functionality—auditing all of these is difficult.

Hype check: concern about the capabilities and risks of smart homes are warranted, though so far the security vulnerabilities of IoT devices have not appeared to cause widespread harms to users themselves (though have been used in other attacks). However, as the richness of data collected in the home increases, the individual privacy risks will necessarily increase. A troubling trend is the overhyping of the capabilities or benefits of smart home systems as a way to entice uptake, selling users connected systems that provide questionable security benefit either to fuel platform lock-in or to collect monetisable data. This trend warrants more attention than it has attracted to date.

The bottom line: the smart home has mostly matured as the initial market share land-grabs and platform wars have started to settle into a status quo. The next phase to look out for is whether companies see the smart home as a valuable data source, and how they decide to monetise that data, and how individuals (and the market) respond.

Biometrics

Summary: the collection and analysis of data that is specific to an individual's body. The first and most well known use for biometrics is in identity and authentication—verifying identity and controlling access to systems—originally using fingerprints and more recently with iris, face, subdermal blood vessel patterns, and gait. These physiological characteristics are highly specific to a person and are persist for long periods of time, in many cases for the subject's lifetime.

The other main use of biometrics is generating data on the state of a person's physiology, usually for health applications. Of course, medical devices could be considered biometric systems, and biometric sensors previously only found in medical tech have crossed over into consumer electronics—fitness trackers, smart watches, and smartphones have arrays of sensors and ample processing power to analyse that data to produce detailed physiological measurements. Commodity wearable hardware can now monitor heart rate, blood oxygen, blood glucose, electric conductivity of the skin (galvanic skin response), and more. These can add up to profiles that can be monitored over time, providing insights for both medical professionals and for personal fitness.

Sensors do not always need to be worn to provide biometric measurements. For example, advances in computer vision and modelling techniques mean that only a camera is needed to gain insights about mobility and physical performance. A prime example of this is gait recognition, which can be performed entirely passively. Generally, as the number of sensors in the built environment increases—visible or IR cameras, air quality sensors—and signal processing becomes more sophisticated and efficient, it becomes more feasible to collect detailed biometric signals passively. Some consider this the next frontier of surveillance technology.

Maturity: in wide use. Biometric authentication is already ubiquitous, e.g. used in passports, on smartphones, etc. Serious consideration is currently being given to passive biometrics for commercial or public purposes, such as tracking individuals around shopping centres or other public spaces using gait. In fact, in addition to voluntary non-contact temperature checks in shops and other public places due to the SARS-CoV-2 pandemic, the monitoring of body temperatures en masse using CCTV-like thermographic cameras has been widely deployed.

Novelty: while biometrics have had a long history, the capabilities of biometric devices and the richness of data now available is surprising to many, and newly feasible collection-at-scale is effecting new applications.

Risk impact: as the number of wearable and environment—embedded sensors increases, the collection of biometric signals will become progressively easier at scale. As a person's

biometric characteristics are long-lived, any breaches or uncontrolled disclosure of biometric data could harm a user's privacy for long periods of time with little opportunity for mitigation.

Hype check: public mass collection of biometric signals has the potential for novel social and technological applications—while many aspects of biometric technologies aren't necessarily new (biometric tech is already widely used), increased availability enables many potential applications in emerging systems. The level of sophistication of discourse around the implications for personal privacy has been somewhat lacking, perhaps because the risk lifetimes dwarf the expected lifetime for any particular tech. Biometric signals are highly detailed and specific to a person for long periods of time—management and disclosure of biometric data requires close attention.

The bottom line: public and population-level biometrics are an up-and-coming surveillance tool, and a lot of biometric data is long-lived and highly specific, so difficult to rescind or repudiate. Think of a raw biometric signal as a Pandora's box—your best chance for controlling risks is actively at the point of collection, before you let it out into the world.

Brain-Computer Interfaces

Summary: currently BCI works to monitor activity levels in a few areas of the brain. Certain areas are known to be active during certain tasks, and when they show activity we can attempt to compare these patterns to previously observed ones. There have been some minor advances in recent years in non-invasive and portable brain sensing technologies, but they are still somewhat primitive and are a focus of research, with few proven useful applications. The data collected is a very coarse-grained observation, and far from actually understanding of the processes that produce that activity. To use an analogy, we can see the shadows on the wall, not what's casting them.

Note that we're not talking about therapeutic implants for conditions such as epilepsy, which fall more under the umbrella of wearable technology or body augmentation.

Maturity: still a nascent technology, currently at the low-level hardware stage of development; i.e. producing safe and reliable sensors that are minimally invasive. They currently don't yet yield very sophisticated signals, and the potential utility (if any) of those signals is a matter of debate.

Novelty: if we are able to really isolate signals within the brain, this has the potential to provide very personal data—unconscious or subconscious neural reactions to various stimuli would be a new frontier of data intimacy. There may also be sufficient uniqueness in those signals that a "fingerprint" of brain activity could be constructed. As a research area, BCI still is dominated by conjecture.

Risk impact: a number of concerns are raised by the concept. There are a number of large question marks, given the technology is nascent—the specifics of how it develops will determine the potential risks and harms.

Hype check: despite what some billionaires would have you believe, this tech is still in the very early stages: basic sensor development and signal processing. Research is primarily at the stage of exploring "if", rather than "how".

The bottom line: beyond some basic accessibility use-cases, such as moving a mouse pointer around with your mind by putting electrodes on your head, specific data capture at scale appears some time off.

Data markets

Summary: as the collection of data has become more pervasive, it has also become commodified. This marketisation of data has become big business—while adtech may be the obvious and most visible example of collectors monetising their datasets, the promises of "Big Data" across industries has incentivised entities to collect data and sell it on where appropriate; most every data point can be contributed to some profile or larger dataset, adding value for a third party's analyses. This has led to entire ecosystems and supply chains for data—networks of data brokers who aggregate, link, and sell on datasets.

Maturity: widespread, but difficult to quantify as these markets are hard to map, and the parties within are resistant to being surveyed.

Novelty: the barrier to entry when analysing a new dataset is substantially lowered, as accessory data that could be used in analysis or linkage is now readily available.

Risk impact: opaque markets trading in vast datasets make it difficult to appraise risks for any given dataset, as it is intractable to gauge the availability of accessory datasets against which it could be linked.

Hype check: attention to data selling is mostly paid to consumer-facing companies, particularly in the shape of minor scandals about how they monetise data contrary to the assumptions of their users. But this narrative belies a sophisticated industry that creates extensive data flows across companies, silos, and borders.

The bottom line: data markets have fundamentally changed the availability of data, and accelerated trends towards linkage-based data processing. Data flows more freely than ever, posing questions about validity, quality, and provenance.

N.B. *an important subset of data markets are real-time bidding (RTB) marketplaces, prominent in programmatic advertising. If data markets are analogous to money markets, RTB is high-frequency trading. There is substantial attention, including active legal complaints, being paid to this space at present.*

Synthetic data

Summary: usually refers to producing representative ("lookalike") datasets—i.e. datasets with similar properties to the real data—to enable certain processing. For example, synthetic data is attractive when you don't want to use real individuals' data, or when you have insufficient data to perform a computation. Synthetic data can be produced by a number of technologies, from generative models to full system simulation. Differential privacy could be considered a synthetic data generation technology, as it produces an anonymised dataset from a raw dataset.

Maturity: still quite young. Synthetic datasets are currently a hot research topic in medical data, and as machine learning techniques become more able to capture nuances in datasets, generative models such as GANs are being keenly watched for use in applications where data is scarce or too sensitive to use raw.

Novelty: arguably an iteration of well-worn statistical techniques for producing aggregate reports. But now with machine learning techniques, it is possible to capture even more statistical features of a dataset and produce even more realistic lookalikes. Use cases are new as well, no longer simply data release but for training AI/ML systems.

Risk impact: difficulties arise regarding the certainty that any sensitive patterns or biases from the real, source data have been removed. This is especially difficult as biases often don't become apparent until well after the data has entered use. There are also risks around how representative the data is. In many cases, you don't know what you're looking for in data until you start poking around—if the process of making lookalike data doesn't pick up the needle-in-the-haystack cases, it may not prove useful at all.

Hype check: a promising alternative when one is dealing with large datasets full of granular, well-defined data. Some particular limitations, such as capturing unexpected features in the source data, and limited applicability to non-numerical data—text fields, for example. Also, if you're trying to preserve privacy, every time you perform a new synthesis from the same source data, you leak information about that source (see differential privacy).

The bottom line: important to watch this space, particularly in its first big implementations. It's likely this will be a gold standard for datasets where you're looking for aggregate or demographic data, where the statistical properties you want to preserve are known. However, it's not yet clear that in cases like health data we are able to preserve all the interdependencies of the original dataset. In other words, it may not prove useful when you don't yet know what you're looking for in the data. Those most set to benefit are incumbent firms who already have access to huge amounts of data from which they can synthesise new data.

Models of computation

Next we will present our cohort of notable and growing models of computation. This refers to high-level system architectures, design patterns or philosophies, and general technological approaches, rather than specific technologies.

Edge computing

Summary: the term edge computing generally refers to performing computation at the edge (i.e. closer to the user and/or data source). The idea is to leverage the processing power at—or closer to—endpoints (e.g. user devices), and avoid the overheads in terms of bandwidth and storage associated with more centralised processing. Edge might involve performing particular tasks to serve that locale, or might be part of some broader distributed/collective computation. As edge is a generic concept, the actual shape of these systems will depend on the technologies that become popular.

Edge computing holds clear benefits even for the most entrenched players. Edge would enable processing over data that would otherwise be unavailable, either because of privacy concerns or limited resources. And even in many cases where processing is done today in centralised architectures, it may well be that processing can be pushed to the edge with little to no loss in results. In fact, robust approaches to edge computing could be a competitive advantage for established firms, as they have the scale, resources, and market power to push their own approaches. These approaches would likely be leveraged into platform advantages.

However, firms do have some competing incentives when it comes to adopting edge computing. There are certain cases and practices that are not amenable to edge, and in these contexts centralised processing will continue to be used. Also, the industry status quo is widespread collection and linkage, where companies tend to want maximum data visibility, centralising and hoarding as much data as possible to extract value. While we can debate the extent to which data hoarding is a successful or rational practice for these firms, an entire industry of data brokerage has been created, and is likely to be a source of inertia or pushback.

Context: there are three key drivers behind the popularisation of edge computing. First, the growing scale and richness of data being produced becomes impractical for latency, bandwidth and other constraints on the aggregation and central processing of data;

second, an increase in resistance, liabilities, and other issues regarding data sharing which can be avoided by not centralising data in the first place; and third, the significant amount of computing power available at the edge (i.e. on people's devices, neighbourhood hubs, etc).

Maturity: the latest rebirth of a paradigm that comes and goes—the locus of computing has historically moved multiple times between centralisation and decentralisation: from mainframes to personal computers, to the cloud, to smartphones (to oversimplify). It is not yet clear if this trend will take hold and edge computing will becomes the status quo; the trend is currently in a stage where there are a number of technologies under its umbrella, and it remains to be seen which will gain traction and become foundational.

Hype check: edge computing is likely to become a significant part of the data processing landscape, but the extent to which it becoems the dominant paradigm will depend on market forces, regulatory influence, and if new technologies for multiparty computation and federated processing reach maturity and become mainstream. These factors do appear to be moving in that direction—prominent platforms and infrastructures are already being rolled out to support edge.

The bottom line: edge computing is in line with current moves towards more decentralised processing. There are strong drivers for its uptake, as well as disincentives and inertia, and it looks likely (at least in the nearer term) to support and complement rather than entirely supplant centralised forms of processing.

Distributed Ledger Technologies (e.g. blockchain)

Summary: blockchain is an example of distributed ledger technology (DLT), a type of database whose maintenance is distributed across multiple parties. The structure of DLTs is broadly the same—each new entry in a ledger contains a cryptographically derived representation of the previous entry, which works to 'link' entries together. The cryptographic function that produces this representation is publicly known, which enables integrity checks—any party can recompute the cryptographic representation for an entry to ensure that it is correct (i.e. hasn't been tampered with).

The process of adding an entry to the ledger is by *consensus*—where a set of parties must agree (according to a set protocol) whether the next entry can be appended to the database. Some consensus processes are (deliberately) computationally very expensive, requiring significant energy.

In short, the technology is akin to a traditional ledger, but with multiple eyes watching it. It is important to note that blockchains are only one type of DLT, and that other DLTs exist, with different definitions of consensus and means of achieving it.

Context: distributed ledgers that preserve integrity have been around for quite some time. The novelty in modern DLT is the methods for achieving consensus between all parties, as opposed to a 'clearing agent' (i.e. a trusted central authority).

Maturity: there are many DLT implementations out there, most prominently blockchains which are usually used as cryptocurrencies (Bitcoin, Ethereum). Numerous research projects and in-the-wild experiments have shown that blockchains are somewhat inflexible, offering very strong but narrow security properties with significant compromises.

Hype check: blockchain is a technology that has been significantly hyped (particularly given the ideological appeal of cryptocurrencies) and much critiqued. But beyond speculative 'currency' investments there haven't been any real transformative applications (or 'killer apps') of blockchain—mostly just 'niche' examples of its potential use. Some major tech firms offer services that claim to leverage blockchain, but usually in 'closed' contexts, with well-defined business actors, rather than low-trust scenarios. (And it is unclear how much this is a marketing gimmick.)

Other DLTs and related consensus or conflict-resolving technologies, on the other hand, are probably underhyped—there is a growing pool of building block technologies for decentralised data storage, which could fuel future technologies that do not rely as much on central authorities as the status quo.

The bottom line: in most cases where blockchain-based solutions are proposed, an existing (potentially distributed) database technology will do. There may be some cases in which the specific properties of a DLT are required, but these appear few and far between.

AI-as-a-Service (AIaaS)

Summary: service (cloud) providers are increasingly featuring AI-as-a-Service offerings. These tend to fall into three categories:

- the infrastructure and support enabling customers to perform their machine learning;
- pre-defined types of models but trained on customer data or tuned to their use case;
- ready-made, pre-trained models that can simply be applied to the client's problem (e.g. audio transcription, object recognition, facial recognition etc).

The latter two categories provide access to AI models in a 'plug-and-play' fashion, often marketed as needing no machine learning expertise to incorporate, and are often available on demand, at a few clicks. Note that some services use customers' data to help better refine their service offering.

Context: the latest extension of the modularisation of systems. Building machine learning models requires expertise, data, and resources. AIaaS is seen as a means to meet high levels of demand for AI capabilities. Note that dominant firms have a clear advantage in this space, having the expertise, access to data and technical capabilities and resources to make their offerings 'state of the art'.

Maturity: most major providers already offer AIaaS.

Hype check: already exist and growing in popularity, given they make available the power of machine learning models for significantly lower entry cost. Will be interesting to see how specialised the services become. Problematic AI services have already manifested, e.g. facial recognition-as-a-service has recently attracted much public attention.

The bottom line: AIaaS is here to stay. Consider platform power and monoculture effects, e.g. regarding gatekeeping, problem proliferation (e.g. a model's bias propagating across customers' systems), etc.

Code-to-data systems

Summary: code-to-data systems are another example of moving the locus of computation away from 'untrusted' third parties, and more under the control of the dataset owner. In this sense, they are closely related to (and share building blocks with) edge computing and Personal Data Stores. In particular, code-to-data systems are meant to be alternatives to dataset publishing—cases where just one actor holds a sensitive dataset, usually on the behalf of many subjects.

The key function of these systems is to limit information disclosure as strongly as possible by never providing access to raw data to any third parties. Instead, they receive third parties' analysis code and run it in a closely monitored and trusted environment, and send back only the results. By controlling the locus of computation, the dataset holder is able to apply many other data protection mechanisms, such as applying mathematical transformations to the data before processing, restricting the operations available to incoming code, and manual audit of code and results.

Maturity: this is a growing model of interest for data publishing. There are many systems used by governments or public authorities, such as healthcare providers, that already operate on this model. Research into these systems and attempts to describe their security and privacy properties are a nascent academic field.

Context: this model has much in common with personal data stores, but instead of a subject-by-subject approach, it aggregates data and trusts a central party to manage it. The key difference between code-to-data systems and technologies like federated learning and secure multiparty computation is less reliance on strong mathematical techniques, and greater leverage of "soft" risk mitigation techniques like auditing, adding friction to the processing experience, and the imposition of contractual constraints.

Hype check: the code-to-data systems most people are familiar with are hosted solutions focused on enabling data science for non-technical users. Because such systems have been independently "invented" from scratch by institutions in niche settings, it has not yet emerged as one clear model in the public consciousness of data scientists. It is possible that similar systems will emerge soon as a result of businesses' growing interest in internal "data platforms".

The bottom line: look out for more models of computation where the dataset holder does not need to actually hand data to third parties to extract value from it. Code-to-data systems may be seen as a more practical solution than wholesale adoption of distributed processing, due to the practical benefits afforded by having a central trusted party. Expect more systems that try to strike new tradeoffs between enabling data availability and the loss of control from releasing data.

Personal data stores

Summary: personal data stores envisage an individual being given a computational data 'container' (perhaps involving a physical device). This container empowers the user, either directly or through setting a policy, to mediate and control the transfer or processing of that data.

Maturity: the idea has been around for a long time. There are several academic initiatives, and some commercial—Tim Berners-Lee's Solid project is one that is of particular attention at the time of writing—however none have yet gained significant traction, nor demonstrated a viable business model. PDS ideas cross-pollinate with edge computing, code-to-data systems, federated and distributed computation, etc.

Context: the technology generally aims to directly challenge and offer an alternative to the current centralised approach whereby data is aggregated by a third party for processing, upon which the individual loses control. It is seen by some as an avatar for the ideal of data 'ownership' (or at least 'empowerment' and 'control').

Hype check: PDS has long been the goal of decentralisation partisans, but it remains to be seen whether there is a niche in the actual world of data processing where it can fit. Though no concrete PDS projects have gained widespread traction to date, the field in general has served as a nursery for various technologies in decentralised or edge computing. It is also debatable whether it is appropriate or realistic to rely on individual agency to effectively manage such concerns. Furthermore, as data collection becomes ambient and highly granular, it is not clear how a system could capture and withhold all data generated about a subject.

The bottom line: PDS might be considered more a design experiment—an extra-strict access control system, research into which provides a staging ground for distributed processing technologies—rather than a viable potential disruptor of the current data processing landscape. Personal data stores in forms discussed appear unlikely to materialise, given that we seem past the point where all the data generated by or about us is even enumerable, let alone manageable through a user-oriented control mechanism.

Simulation

Summary: simulation techniques promise to reduce reliance on individuals' data and enable investigation of problems that would otherwise be impractical.

Agent-based simulation can help the generation of synthetic data, but also enables more —it is the approach of constructing agents that exhibit behaviour that mimics a real-world actor, and building an environment with rules and constraints for that agent. Then, the agent can be allowed to interact with the environment and its behaviour monitored.

Agents can be defined manually and painstakingly, with a set of rules that govern how they respond to different situations or stimuli. Recently, however, machine learning advances have removed the need for this painstaking process. Now, agents can be created from data captured about a real-world subject, which allows us to capture nuances of behaviour that a human observer perhaps would not have caught, or could not describe well in logical, rule-based terms.

Maturity: accessible high-performance computing has made simulation techniques that were previously prohibitively expensive viable, and all sorts of simulation techniques have become more widespread.

Rules-based agents in particular were one of the earliest (and most computationally expensive) AI technologies, and have decades of research behind them. Today, as computation at scale is cheap, they are coming back into vogue. The new breed of ML-trained agents is a novel development and also contributes to this resurgence.

Context: agent-based simulation has been a key area of AI research since its early days, but often involved laborious design. Many advances in the space have come from gaming, and then carried over to other industries.

Hype check: now that we have the power, look out for agent-based simulation as a popular modelling technique—there was a significant amount of hype earlier this year for using simulation techniques for modelling spread of the SARS-CoV-2 pandemic. But remember that while machine learning can alleviate much of the burden of designing agents, the setting in which they interact still requires thought and expertise to construct.

The bottom line: using detailed and repeated simulation of scenarios is rapidly becoming a popular way of analysing systems without some of the risks or costs that come with real-world experimentation.

Building-block technologies

Finally, we present an overview of some data processing technologies. As this is a rapidly evolving space, we will limit our discussion to a group of indicative technologies and families thereof, whose uptake is either underway or sufficiently past the point of theory.

Federated learning

Summary: federated learning is a subfield of machine learning which separates the training of models across multiple parties. This enables a model to be trained on a number of datasets from different sources without the need to centralise the data in one place.

One driver for this is confidentiality—you do not need to disclose any of your raw data to other parties, whom you may not trust. Instead, derived mathematical parameters are exchanged between the parties to enable broader computation to occur. Complete confidentiality is not achieved, as those parameters (returned and exchanged) are still derived from the raw data and so can leak some amount of information.

Another driver (as with other distributed models) is the potential for alleviating some of the management responsibilities for centralised data—costs of storage, engineering for transfer at scale, etc.

Maturity: this is a nascent but rapidly emerging area. How much of the machine learning we currently do by centralising data can be moved to a federated approach is unclear, and will also depend on other trends, such as moves towards edge computing, personal data stores, and so forth. A significant factor in its uptake will be how quickly it advances relative to the wider machine learning field—for a while it is likely to be less efficient, and so the benefits of decentralisation may insufficient to warrant sustained attention and development.

Motivation: the wider field of distributed computing has been around for a long time, as such techniques allow efficient use of resources, or processing to be *parallelised*-i.e. data and tasks to be decomposed and computed separately (in parallel) for efficiency gains.

Previous distributed learning techniques relied upon stronger assumptions of homogeneity across parties, both in terms of statistical similarity of individual datasets and the participation of parties in processing.

The benefit of federated learning techniques is the relaxation of these assumptions. This allows different parties to hold very different data (crucial for most multiparty learning), and more gracefully handles participants who may contribute at different times, or have differing levels of processing power. In other words, it enables new computation architectures with stronger privacy properties.

Hype check: not all machine learning can be done in a federated fashion, and even where it can, it is important to note that confidentiality is not necessarily achieved. While only mathematical parameters generated from the data are transferred, those results can still be revealing. If your data is significantly different from another participant's, the parameters you generate will accordingly be different from theirs, perhaps in a way that leaks information about your data. Thus the key benefit of federated learning systems, the ability to incorporate heterogeneous data from participants, also underlies their limitations regarding confidentiality.

The bottom line: federated learning is a step towards decentralised processing, though note that it does not fully mitigate data risks. The field is yet young, and we may see that a significant proportion of machine learning models may be computed in a federated fashion. Look out for how future systems navigate the privacy tradeoff between information leakage and learning from new, outlying data.

Differential privacy

Summary: the core idea of differential privacy is that, for a statistical computation performed on a dataset made up of data from a number of individuals, the result should not (overly) depend on any one individual's data. In practice, this means that if you were to perform the computation on the database before and after removing an individual's data, the difference between the results would be indistinguishable from random noise. As a result, nothing can be inferred about any one individual from a differentially private release of the data. This is usually achieved by adding some random noise to each record in the database—the more records in the database, the less noise needed to achieve differential privacy.

Motivation: differential privacy is a strong information-theoretic definition of privacy, with an intuitive definition and a notion of 'privacy budget' (ε) which can be set and whose depletion can be monitored. Notably, the definition is so strong that while this privacy budget has not been exhausted, differentially private data is impervious to any reidentification attacks through linkage

Maturity: differential privacy has found widespread uptake, from the US Census Bureau, where microdata published as a result of the 2020 Census will not be raw data but a differentially private transformation, to telemetry data collected from mobile devices. It is important to note that differential privacy is not a silver bullet for all kinds of data and use cases—in fact, differential privacy is often critiqued for being narrowly applicable, and there is widespread misunderstanding of what it can and cannot do.

Hype check: differential privacy is a significant step forward for data disclosure, but care must be taken around its applicability; not all data types can be usefully transformed into a differentially private form. For example, noise cannot be applied to text collected from a questionnaire, and even in cases where noise could be applied it could hinder the utility of the data, e.g. such as in DNA bases from genome sequencing, where a few changes could result in an entirely different protein expression. Further still, differential privacy was initially designed for dataset publishing, and while it has been extended to other contexts, modification for use in new contexts requires significant work, and oversights can easily lose you its guarantees.

The bottom line: differential privacy is a prominent method for protecting large datasets, either at collection or in publishing or querying—though some modes of data use are less amenable to this technique. For example, frequently updated databases are difficult to make differentially private, as every new release depletes the privacy budget.

Secure Multiparty Computation (SMC)

Summary: this is a family of methods dealing with computation that multiple parties must work together to accomplish, and which relies on disparate data that each holds, but where no party wants to divulge their data to the others. In these cases, there is no trusted third party that can see all the data that needs to be input to the computation—instead specially crafted calculations are performed on cryptographic transformations of each party's data. An example is comparing two contact lists to see if you have mutual friends with a stranger, where your goal is to find the contacts that appear in both of your lists without showing any of your other contacts to then stranger.

Motivation: SMC solutions for common computations would be a boon for decentralisation efforts. These methods usually require each party to place little to no trust in any of the other parties, allowing distributed tasks to be performed without the disclosure of secrets. Federated learning systems, discussed earlier, are often built using SMC algorithms to mitigate limitations of confidentiality and information leakage.

Maturity: there are a number of well-explored problems in this field. The example above belongs to a subfamily of problems called Private Set Intersection, for which multiple methods exist. SMC is usually much more computationally costly than performing the analogous computation where the data were simply shared.

Hype check: it takes a significant investment of research to produce a SMC solution to a problem that can otherwise be centrally managed, and that solution is often notably less energy and time efficient. However, for common problems the investment may be worthwhile, especially where it would enable an entirely different model of computation to be used.

The bottom line: not every problem can be solved efficiently with SMC, but significant progress has been made on some common problems, where even a costly solution could bring huge benefits by removing the need to trust a central party.

Trusted execution environments (secure enclaves)

Summary: a trusted execution environment (TEE), or secure enclave, aims at securing data and computation through hardware-based mechanisms with various security properties. In essence, TEEs use hardware-backed cryptographic means to provide an computational environment that isolates particular data and code from the rest of system (and outside world), and gives certain guarantees over that data and processing—including around system, code, and data integrity, and means for verifying that (only) certain computation occurred.

Modern smartphones often use a TEE to handle authentication, storing and comparing biometric identifiers (such as thumbprint or Apple's FaceID scans)—this works to protect the sensitive biometric information. To extract data from a TEE in an unauthorised way usually requires laborious and invasive techniques, such as the use of electron microscope, and the hardware is often designed so that such attempts to violate its physical integrity would destroy the data onboard.

Motivation: there are many points of failure in a modern system. There are many applications where you need to be able to trust at least one part of the system—know that it is running the software you think it's running, that its secrets have not been exfiltrated, and that it will behave as you expect. TEEs are an attempt to provide an environment where you can have certain guarantees about the confidentiality and integrity of certain code and data, as a basis for protecting certain functionality, and as a 'root of trust' for building on, verifying and validating broader system behaviour.

There is considerable interest by cloud technology firms in enabling 'confidential computing' which aims to remove a provider from the 'trust loop'. Cloud providers seek to minimise the risk of handling unencrypted data as much as possible. To do this they usually aim to make as much data transfer as possible 'end-to-end' encrypted, meaning that the data can only ever be decrypted at the intended recipient. In addition, these providers encrypt data 'at rest' (when it is stored). These measures are intended to minimise the possibilities for the unencrypted data to ever be disclosed to an untrusted party. However, the service provider still needs to decrypt the data to process it.

In a 'confidential computing' system that leverages TEEs, users can upload encrypted data to a provider, which is only decrypted within the TEE. The provider performs verifiable computation over data for the user within the TEE—with guarantees that certain computation occurred, and without the provider ever having access to the legible (unencrypted) data (or code).

Maturity: TEEs are already commonplace, included in chips by all major manufacturers (e.g. Intel SGX), and have gained traction for application in low-trust environments (such as shared servers), or for secure remote computation. Current usage focuses mostly on protecting small, specific tasks—e.g. encryption key generation and storage, authentication, system integrity (configuration) and code/data (tampering) checks. There is interest in using TEEs to encapsulate more complete software or data processing operations, though this is an area of ongoing work—the larger and more complex the code in an TEE, the less concretely you can describe the guaranteed security.

Hype check: TEEs are useful where hardware attacks are common and supply chains can be of questionable integrity. They have also been shown to be useful for giving confidentiality and integrity guarantees to protect a few, specific, mission critical operations (such as authentication, decryption, etc). Cloud providers themselves are adopting TEEs at scale and advocating for their wider use. This is because it would significantly improve the security properties of building systems on their platforms, while also minimising risk exposure.

Currently in the field of secure software engineering there is a push to put more and more code in the trusted area, but TEEs are a complex target to write code for, and poorly architected software executing in a TEE environment can undermine the security properties the technology aims to provide. Therefore care must be undertaken to design software to best leverage TEE capabilities—this requires specialist expertise which is in short supply.

The bottom line: TEEs are already widely deployed, but are becoming more commonplace, fuelled by significant industry interest from infrastructure providers. Expect TEEs to form the bedrock of (some) new system architectures that either separate out into more modular and verifiable blocks, or that push more and more processing into trusted hardware.

Homomorphic encryption

Summary: to conceal information from a third party, data can be encrypted. This transforms the data in a way that it becomes unintelligible unless one holds a specific key, which tells you how the transformation is performed and allows you to reverse it. Usually, to perform any computation on data, it must be unencrypted. However, in fully homomorphic encryption, computation can be performed on encrypted data. This means that an untrusted party could perform a computation on the encrypted, unintelligible data, without ever seeing the original (decrypted) data. The result of this computation could then be decrypted by the owner.

Motivation: like TEEs, homomorphic encryption would be a dream scenario for cloud service providers, or any multiparty system where one party performs computation on another's behalf, as it enables them to remove themselves from the 'trust loop'. Like restricting decryption to a TEE, homomorphic encryption also ensures that the unencrypted data would never be in the service provider's hands, absolving the client of the need to trust the provider.

Maturity: while fully homomorphic encryption (FHE), which allows all computation without divulging the underlying information appears very relevant for a range of data concerns, currently it is incredibly inefficient and few actual systems exist. Partially homomorphic encryption (PHE) allows a more limited set of operations, and is more practicable (but still computationally expensive). In PHE, instead of preserving all of the information, you choose some properties of the underlying data that are maintained during computation on the encrypted form.

Hype check: homomorphic encryption is an active research topic, and there is skepticism that fully homomorphic encryption will ever be achieved in a practical sense. However, partially homomorphic encryption methods exist and may yield some narrower practical solutions.

The bottom line: FHE shows promise, though may still be some time off. On the other hand, PHE mechanisms have already been deployed, and some of partially homomorphic encryption forms the basis for SMC solutions.

Data provenance

Summary: in a data context, provenance concerns mapping out data processing pipelines. This is by capturing information that describes data—recording the data's lineage, potentially including where it came from, where it moves to, and the associated dependencies, contexts (environmental and computational), and processing steps undertaken.

Generally, provenance systems aim to provide some combination of (a) a 'language' defining, describing or otherwise encapsulating what happens; (b) capturing what actually happens; and/or (c) describing what should happen and alerting to violations.

In essence, it can be thought of as a form of logging describing what happens to data.

Motivation: currently what happens to data is largely invisible or opaque—provenance attempts to establish traces over data and its processing. It tends to align well with the modelling of business workflows, to enable monitoring, audit and possible recreation.

A key driver for provenance work was research reproducibility, by capturing data on scientific computing workflows in order to provide visibility over the process, enabling the research to be reproduced. More recently provenance has seen some traction in systems security contexts, where logs of data flow/processing within systems can indicate potential security issues (i.e. indicating spurious behaviour).

Maturity: provenance has been around for some time. Provenance 'languages' (ontologies) are well-established and exist for many domains. Generally provenance methods and associated tools tend to be application or domain specific, tailored to a particular scenario or workflow. General tooling to provide records for enabling data audit are largely still in the research domain.

Hype check: there appears to be a role for provenance in dealing with data/algorithmic accountability concerns, though this usage has yet to see traction. There are technical barriers for making such methods work at scale. There is some discussion of provenance by blockchain (DLT) communities—primarily because DLTs are ledger-based—though it is important to note that the area of provenance is far broader, and that DLT technology would only suit a few provenance use-cases.

The bottom line: provenance systems show potential for tracing what happens to data, particularly in complex data sharing environments, but significantly more research is required.

Prepared for the Ada Lovelace Institute

Authors:

Jovan Powar
jsp50@cl.cam.ac.uk

Jat Singh
jatinder.singh@cl.cam.ac.uk

University of Cambridge