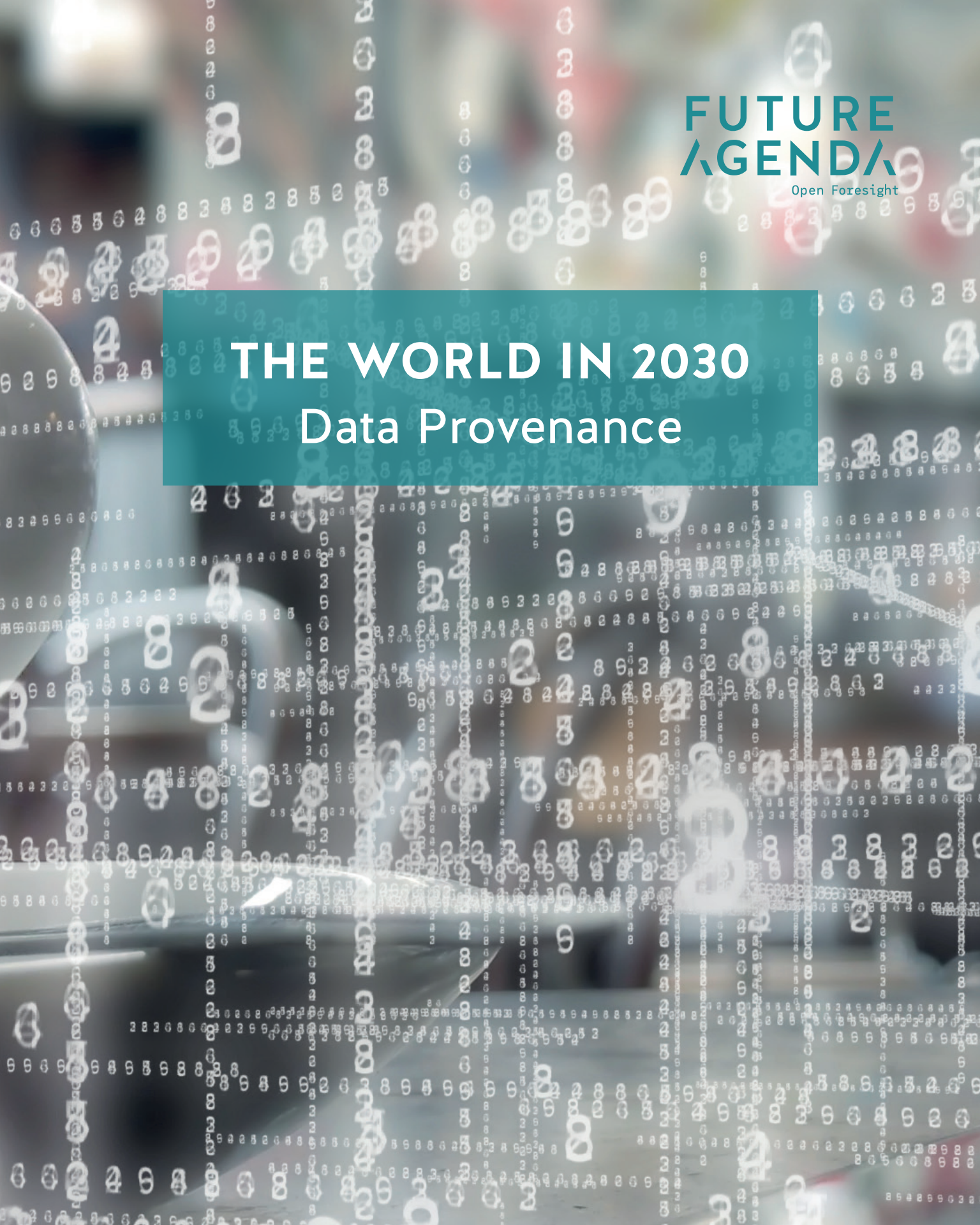
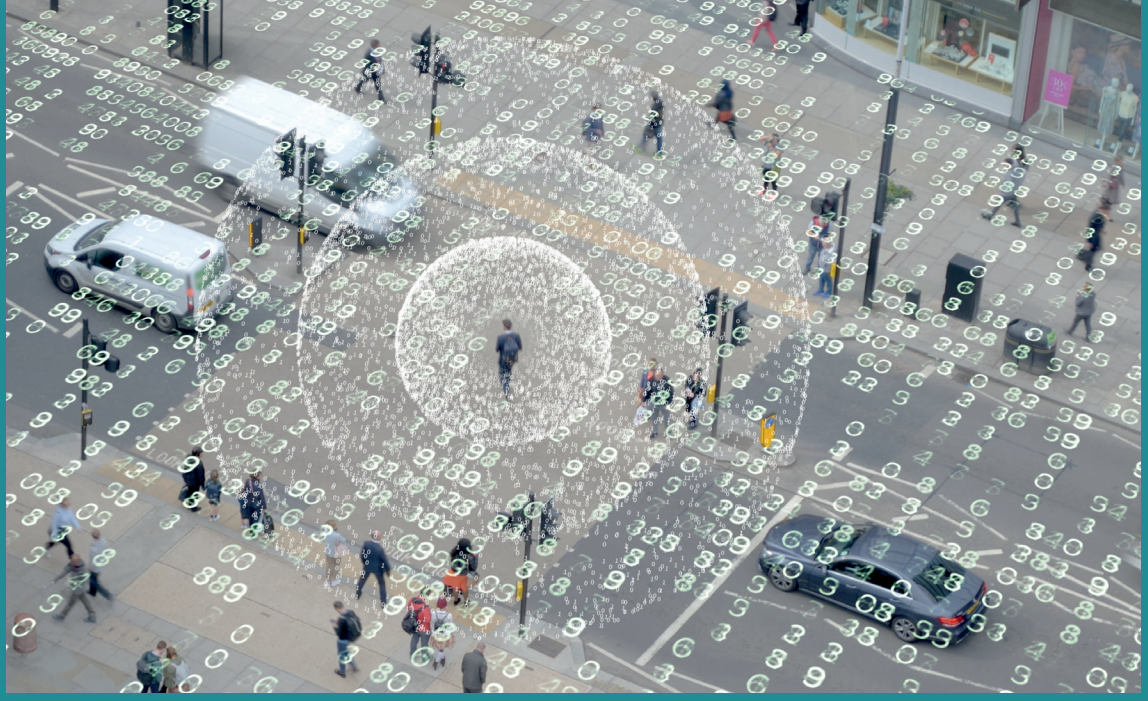


FUTURE AGENDA
Open Foresight

THE WORLD IN 2030
Data Provenance



Data Provenance

Although the ability to trust data is fundamental for economic and commercial activity, technology is exploding the scale and scope of the inauthentic. Demonstrating data provenance becomes a prerequisite to operate.

No of devices connected to the Internet by 2030 [1]	**500 billion**
% autonomous vehicles in the US by 2030 [2]	**10%**

The exponential growth of data and internet of things (IoT) devices across government, industry and in homes, from logistics and manufacturing, to oil and gas, retail and even academia is ushering in an era of data abundance. In this ever-expanding, dynamic, and complex environment, it is increasingly vital to know the origin or the source of data - and whether this data can be trusted. Verifying data is increasingly tricky and therefore expensive. Deciding who or what will help with this process is trickier still. This matters because the world has become, not just interconnected, but also deeply intertwined, with mutual dependencies often at the heart of key infrastructure. Looking ahead, as we become even more reliant on data generated not just by people, but also by things, where the information is continually interpreted by AI and machines, the stakes are going to get higher. Systems will fail if they are not trusted. This is where data provenance comes in.

Provenance is derived from the French word provenir, meaning 'to come from'. Demonstrating provenance has long been a way of validating a work of art; and likewise, in digital libraries it can document a digital object's lifecycle. Recording data provenance is important to confirm its authenticity or its origin, to enable it to be identified, reused and so maintain the integrity of the system. Data provenance shows the pedigree of the data - the record of components, inputs, systems and processes that affect collected data and provide historical context. It provides an understanding of where data comes from, how it is collected and how it can best be applied. It allows devices and systems to receive reliable updates and timely security upgrades and the algorithms used to power analytics and AI to be effective and trustworthy. In the IoT, cloud-based, just-in-time world we are busily constructing…data provenance is, well, nearly everything.

Although the principles behind the need to establish data provenance remains the same as that for physical objects, the nature of data presents specific challenges. As a 'resource' data is, for instance, different to traditional physical resources. Like knowledge and ideas, when data is 'used' it doesn't get 'used up.' This means the same piece of data, such as the location or speed of a bus, can be used for multiple different purposes by multiple different parties – passengers, other road users, smart city networks and the varied public transport bodies just as much as the vehicle manufacturers and a whole host of suppliers. And, far from being a depleting resource, data is an accumulating one. This creates potentially vastly complex lines of provenance for a single data point, let alone the trillions we create each day. The implications and ramifications of these unique characteristics of data are vast; they up-end our notions we often see as fundamental such as the idea of 'ownership'; they force us to challenge many of the assumptions that lie at the heart of economic analysis; they transform both the possibilities data creates and the dilemmas it generates; and they change relationships between stakeholders, be they individuals, communities, networks, organisations, government or wider society. This is why establishing provenance is so difficult.

Systems will fail if they are not trusted. This is where data provenance comes in.

The Benefits of Provenance

The W3C Provenance Incubator Group defines provenance as: “a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance”. As with a piece of art, data provenance concerns the lineage and derivation– its ‘custodial history’.

This record can be useful for a number of reasons.

- Knowing the origin of data can help to assess its quality, accuracy and reliability and therefore can provide a degree of trust that we (or an entity acting on our behalf, or a machine) places in it. This is significant when you consider the amount of trust we already have, for example, for voice assistants and devices such as smart meters and, going ahead, will need to be in place for the successful implementation of smart cities. These will embed responsive IoT that can be trusted into infrastructure and built environments. Alongside connected buses expect streetlights that change intensity based on the presence of humans and the prevalence of autonomous vehicles. Even with robust data provenance, with every additional connection it will become harder to establish where a vulnerability has emerged.

- Data provenance can also be relevant in debugging or detecting software errors, or even in detecting the entities or actions that produced the errors in the first place. And of course, knowing the custodial history of data can map both the connection and the responsibility. Bruce Schneier, cryptographer and author of ‘Click Here to Kill Everybody’, has recently observed that “most software is poorly written and insecure”. While for some safety critical areas, such as avionics, software has to be bug-free, for the vast majority of applications the onus placed on coders is less stringent. Such is the expense of development; much software is routinely released in beta format so that bugs can later be discovered by users. “Updates” are then regularly sent to effect necessary repairs and we accept this. Looking ahead data provenance will likely be used to apply more rigour on software development whether undertaken by man or machine.

- Data provenance will also have a growing role in the assessment of data value and ownership. Current GDP calculations do not include the full extent of digital activities or measure the further value that is generated when data is resold or reused. The first step to address this will be to ensure that we can track the different ways data is being used and quantify how it can have different values dependent on its use and who or what it is used by. An individual’s data is almost worthless in isolation and cannot be sold or traded in any serious way. Value is only derived when it is aggregated and combined with others’ data. It then has differential value to different data companies depending on the intended use. Its provenance will influence its value and as data is increasingly and more openly monetised, traded, and, again just like an object in the physical world, validated.

Current GDP calculations do not include the full extent of digital activities or measure the further value that is generated when data is resold or reused.

New Standards

A core challenge is the need for standardisation. Consider the potential prevalence of autonomous vehicles on our roads in the next decade or so. Although some see that all cars will be independent, the majority currently foresee that in most use-cases each vehicle will rely on the data provided from all the others in order not to crash. It is reasonable to assume then that each vehicle must be certain of the provenance of the data it is being provided with and is acting upon: that it is accurate, timely and comes from a certified source.

The authors of 'A survey on data provenance in IoT' suggest that any future data provenance system should follow the following eight guidelines: Completeness, Trustworthiness, Granularity, Depth, Accuracy, Efficiency, Verifiability and Scalability.[3] In addition, there are four other security needs to be met:

- Integrity: throughout the whole lifecycle of the data; including source and path integrity.
- Confidentiality: so that an adversary cannot gain any information about the provenance through analysis.
- Freshness: timeliness of provenance information should be guaranteed, and outdated provenance information cannot be used.
- Privacy: as provenance is a kind of metadata describing the data it needs legal protection.

To date, however, no single system fully meets all of these needs for an IoT environment. Other issues that should be addressed include; processing and storing large data volumes; incorporating a flexible approach for attaching provenance, privacy protections and aligning with other IoT characteristics of decentralisation, distributed networking and immense scale.

Provenance in Action

Provenance data has multiple uses. For example establishing data provenance is part of the jigsaw puzzle that can facilitate the creation of a digital identity, establishing the key credentials that can be used to identify an individual when conducting transactions, applying for a driver's licence, registering for a course or asking for a loan and so on. A data provenance system could also be used to prevent data manipulation in research by providing a complete, transparent audit trail of all data that is collected, processed, and accessed by academics. If blockchain technology was also implemented any modifications that were made to research data would require majority consensus from stakeholders and would be visible to everyone - ensuring high data quality and preventing individuals from acting dishonestly.

Establishing better provenance will mean that it is essential that data is stored in a tamperproof and replicable way. While block-chain is, as ever, frequently cited in discussions, other provenance determination methods in the mix may well be simpler and better suited to the task at hand, such as the use of well-maintained digital signature and provenance databases.

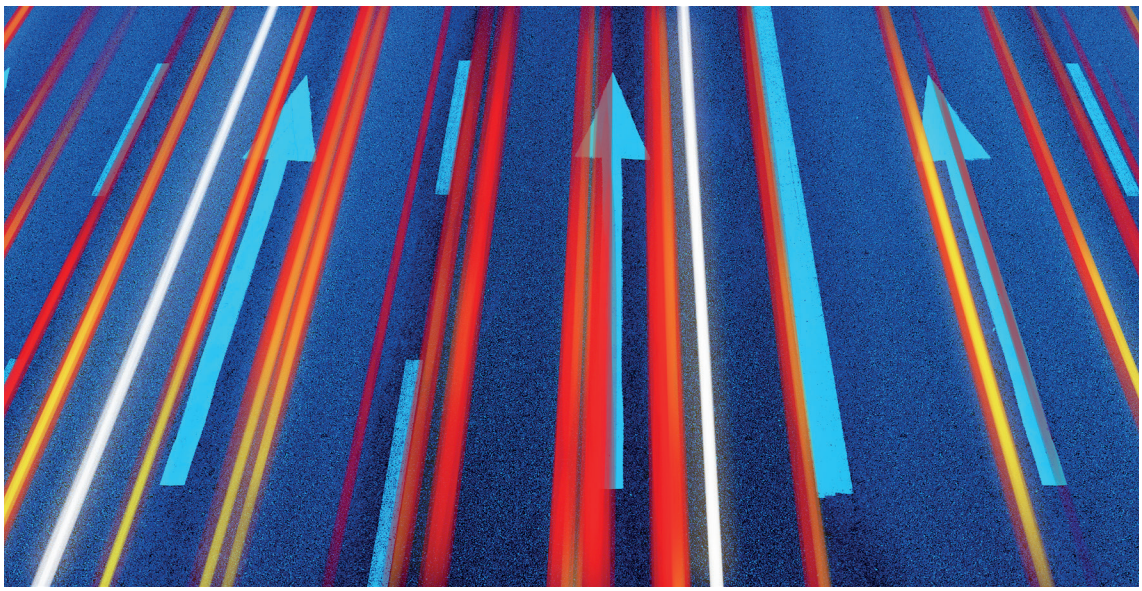
Leading in 2030

As often with nascent technological developments, regulation is lagging behind the technology. A number of organizations are working on this including the W3C that is developing standards for expressing and exchanging 3rd party verified claims, with the aim being to make these more secure and easier.

Technical issues that need to be addressed such as how to manage and store the huge amount of data which a provenance system will generate and how to create a system which will allow access to data provenance detail in a timely and flexible manner.[4] Moreover, how to ensure data provenance privacy has yet to be clarified. It may well be that new mechanisms will be enlisted to solve outstanding issues such as data fusion and filtering mechanisms perhaps.

Given the necessity for stronger data provenance systems, experts agree that stronger regulation and more ubiquitous provenance systems and mechanism will be in place in ten years' time. The real questions lie in what this means for all of us. For example, does ubiquitous provenance mean that we will never be able to do anything, even remotely privately, ever again? And will the benefits of greater cyber security be an adequate compensation for this? Does it lead inevitably to the permanent balkanisation of the internet, as different regimes of trust stop the free flow of data forever? Will it mean that we are all able to benefit financially from being able to accurately track the contributions the data we create make to data-driven processes?

The real questions lie in what this means for all of us. For example, does ubiquitous provenance mean that we will never be able to do anything, even remotely privately, ever again?



References

[1] https://www.ibm.com/thought-leadership/institute-business-value/report/auto-2030

[2] https://hbr.org/2017/12/driverless-cars-will-change-auto-insurance-heres-how-insurers-can-adapt

[3] https://link.springer.com/article/10.1007/s11280-019-00746-1#Sec6

[4] Hu, R., Yan, Z., Ding, W. et al. World Wide Web (2019). https://doi.org/10.1007/s11280-019-00746-1

FUTURE AGENDA
Open Foresight

The World in 2030

This is one of 50 global foresights from Future Agenda's World in 2030 Open Foresight programme, an initiative which gains and shares views on some of the major issues facing society over the next decade. It is based on multiple expert discussions across all continents and covers a wide range of topics. We do not presume to cover every change that will take place over the next decade however we hope to have identified the key areas of significance. Each foresight provides a comprehensive 10-year view drawn from in-depth expert discussions. All foresights are on https://www.futureagenda.org/the-world-in-2030/

Previous Global Programmes

The World in 2020 was published in 2010 and based on conversations from 50 workshops with experts from 1500 organisations undertaken in 25 countries as part of the first Future Agenda Open Foresight programme. This ground-breaking project has proven to be highly accurate in anticipating future change and the results have been used by multiple companies, universities, NGOs and governments globally. Rising obesity, access not ownership, self-driving cars, drone wars, low cost solar energy, more powerful cities and growing concerns over trust were just some of the 50 foresights generated. For more details: https://www.futureagenda.org/the-world-in-2020/

Five years on, the World in 2025 programme explored 25 topics in 120 workshops hosted by 50 different organisations across 45 locations globally. Engaging the views of over 5000 informed people, the resulting foresights have again proven to be very reliable. Declining air quality, the growing impact of Africa, the changing nature of privacy, the increasing value of data and the consequence of plastics in our oceans are some of the foresights that have already grown in prominence. For more details: https://www.futureagenda.org/the-world-in-2025/

About Future Agenda

Future Agenda is an open source think tank and advisory firm. It runs the world's leading Open Foresight programme, helping organisations to identify emerging opportunities, and make more informed decisions. Future Agenda also supports leading organisations on strategy, growth and innovation.

Please contact us via:
douglas.jones@futureagenda.org

Future Agenda
84 Brook Street
London W1K 5EH
www.futureagenda.org
@futureagenda