When anonymized data is anything but: protecting citizens' privacy in the age of urban mobility

Table of contents

Want to talk? We do, too. Get in touch here.

“Collected over time, people’s movements from place to place reveal a good deal about them: where they work, where they play, where they worship, their political leanings, and even personal and familial relationships.”

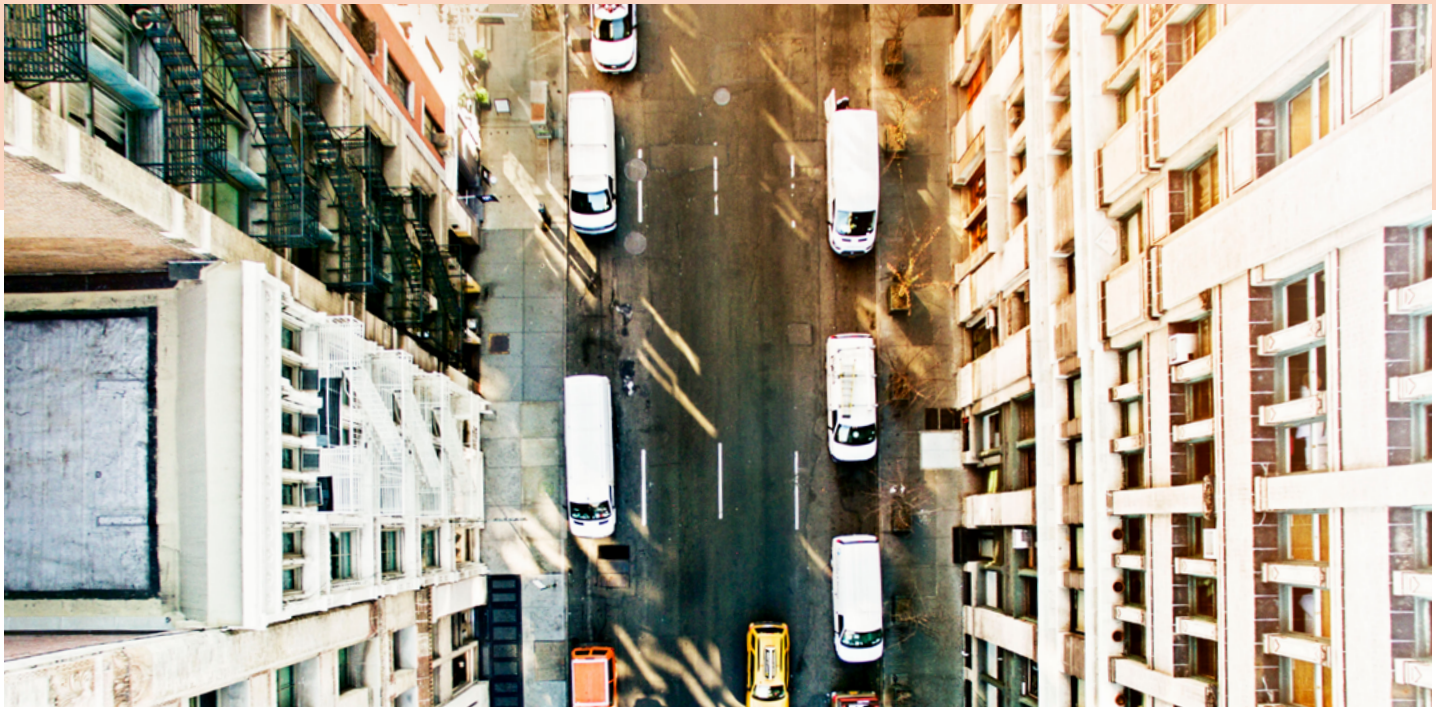
Nathan Sheard, Electronic Frontier Foundation, April 2019

“Our results suggest that even heavily sampled anonymized datasets are unlikely to satisfy the modern standards for anonymization set forth by GDPR and seriously challenge the technical and legal adequacy of the de-identification release-and-forget model.”

Researchers Luc Rocher, Julien M. Hendrickx and Yves-Alexandre de Montjoye, “Estimating the success of re-identifications in incomplete datasets using generative models”, Nature Communications, July 2019.

Want to talk? We do, too. Get in touch here.

Introduction



Whenever our data is ‘anonymized’, it has been manipulated in such a way that it can no longer be linked back to us, right?

Wrong. In the digital era, we do not just blend into the crowd, especially when our location data is part of the equation. This is because our mobility habits are unique to us. Our route to the office, from door to door, is different from anyone else’s, and our traces are highly identifiable with only a few location data points.

Technologies like anonymization offer service providers a wealth of ways to protect their users’ privacy while retaining enough data to serve some utility. However, the challenge is finding that balance. As regulations tighten and people become more privacy-wary following a spate of privacy and data breach scandals, both the private and public sectors must get smarter about how they handle data.

In the following pages, we take a closer look at anonymization and discuss some of the ways public and private organizations can help protect users’ privacy.

About HERE Technologies

HERE Technologies is the world’s leading mapping and location data platform. Organizations use our location data, tools and services to power better real-world outcomes. Journeys become faster, more efficient and safer; fleets optimize deliveries; supply chains become more predictable; and services become location intelligent.

At HERE, we strive to go beyond mere regulatory compliance. We put privacy at the heart of all our products and make privacy an integral part of our corporate culture.

Want to talk? We do, too. Get in touch here.

The truth behind location data anonymity

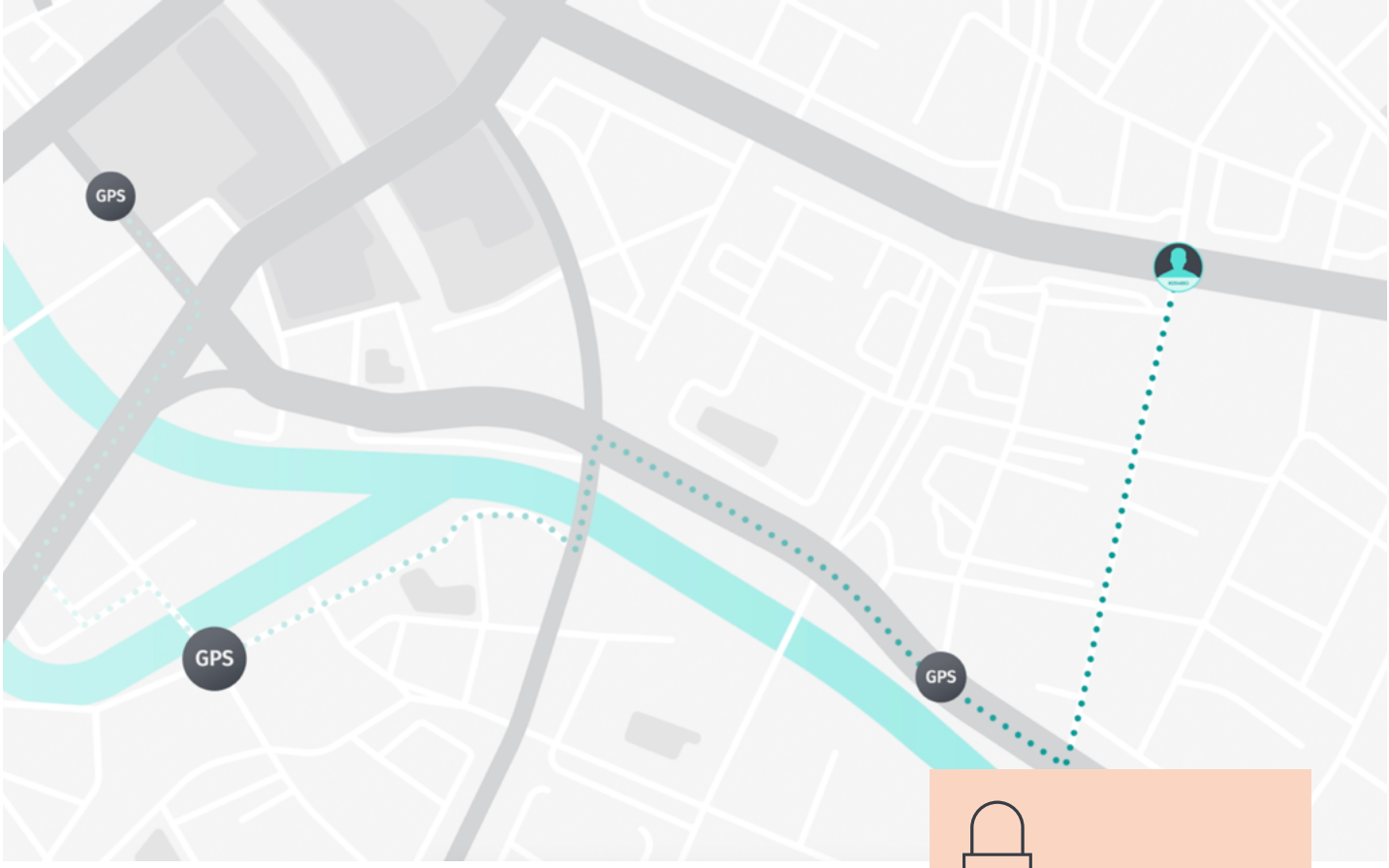
In the world of location data privacy, you may think that removing your personal data secures your anonymity. Unfortunately, that is not always the case.

Your car, your cellular carrier, or an app on your smartphone are the most common things that are likely tracking your location. Understandably, mentioning this might prompt some nervous feelings. Before we explore this any further, please take comfort in two facts. First, most organizations are using tracking data responsibly in order to create safer and smarter data-driven services. Second, most organizations (HERE Technologies included) sincerely want to protect your privacy. In any case, laws enforcing enhanced privacy protections are being enacted around the world and the penalties for breaking them can be severe.

What we are examining in this paper is a particularly challenging aspect of information privacy: location data privacy.

As you move through the world, your devices are not necessarily limited to generating single points of separated data. Rather, they can create a linked set of data points that are more than the sum of their parts. Travelling from place to place produces a whole sequence of locations and timestamps that come together to resemble a path on a map. That sequence, which we call a trajectory, can be particularly revealing and is what can make this location data more challenging than other types of data from a privacy perspective.

Want to talk? We do, too. Get in touch here.



Human mobility patterns are unique and predictable. **With 4 randomly chosen points** in a trajectory you can re-identify **95% of people** with a pseudonym.

The consequences of privacy breaches go beyond reputation damage – it can affect people's safety.

We sometimes hear people say they were "in the right place at the right time," which to the average person means something fortunate happened to someone in a specific place and time. To a statistical researcher or a data scientist, it means a low-probability spatio-temporal event occurred. What is key to understand here is that we are combining the probability of an event with a very specific space and a very specific moment.

This is why location data privacy is so tricky. A company which is tracking you can remove your personal data from the data points and trajectories which they would later make public. However, anyone (especially parties other than the company publishing the data) can potentially add their own insights or other related publicly available data to those published trajectories and use that combination of data to identify an individual. This is possible because humans are creatures of habit. We repeat patterns of behavior, such as the routes we take to work, day after day. In fact, MIT researchers have shown that it is possible to uniquely identify 95% of all individuals using only four randomly chosen location data points.[1]

This is not theoretical – it happens a lot. In a well-documented case study[2], the New York City Taxi and Limousine Commission released the full dataset of every taxi ride in the city in 2013. The information included the taxis' locations, pickup and drop-off times, fares, and tips to the drivers. None of the drivers' names, license numbers, or other personally identifiable information was published. Each car's information was replaced with a numeric identifier, or a hash.

Want to talk? We do, too. Get in touch here.

This is where the external, publicly available related information plays its part. A photojournalist noticed they had stock of pictures of celebrities getting out of cabs in front of buildings. The time and place of those photos was documented, and the numbers on the taxis were visible in many of the pictures.

All the journalist needed to do was combine the information from any single photo with the data points collected at that same time and place in the NYC Taxi database – and there it was. They discovered the full sequence of data points from that specific taxi for the entire day. That in turn revealed where the taxi had picked up the celebrity, or where they were dropped off after the photo was taken.

This example was a journalist taking a dataset and applying some common sense. When data scientists and researchers took the same information, they were able to uncover the names of the taxi drivers, their home addresses, their incomes, and their detailed driving patterns. Moreover, they accomplished this using only publicly available information.

There are a variety of other examples that demonstrate the revealing potential of linked location data. A student was able to locate military bases in the Middle East through anonymized data from a fitness application.[3] Meanwhile, MIT researchers have shown how easy it is to identify individuals if they know only the data and location of just four of their credit card transactions.[4]

De-identification does not always mean anonymity

What we have established is that removing your personal information from location data does not automatically make you anonymous. It is little wonder then that privacy advocates worry that similar events will occur again.

Back in the world of urban mobility, organizations including the Electronic Frontier Foundation (EFF) have raised privacy concerns about the Mobility Data Specification (MDS), a set of data specifications and data-sharing requirements developed for the Los Angeles Department of Transportation (LADOT), but now also used by several other U.S. cities.[5] Launched in 2018, MDS requires companies that operate dockless scooters and bicycles to provide certain usage data back to the city.

The rationale behind MDS makes a lot of sense: it helps cities more easily verify scooter numbers, understand popular routes, and assess whether fleets are

Want to talk? We do, too. Get in touch here.

deployed equitably across neighborhoods. The data also informs future policy and investment into, say, new parking zones or dedicated pathways.

However, the roll-out of MDS has stoked privacy fears, as pointed out by the EFF and other observers. Chief among these is that the data required from the mobility providers is too revealing. While users' names are stripped out before it is shared, the data includes time-stamped information for complete routes taken by the vehicles, including start and end points of a trip. This kind of raw, untruncated data makes it potentially easy to re-identify individuals, especially regular users who might take the same route to work.

For its part, LADOT says it has designated the data as confidential and is using it only for internal regulatory and planning purposes, with no intention to publicly release it. The broader point, however, is that these kinds of data privacy practices can put cities, transport agencies and data providers in a tough spot if the data becomes publicly available (and even if not, if requested by an authority to provide data), especially as privacy regulations tighten in California and elsewhere.

How to identify someone from 'anonymized' data

Probe data is often 'anonymized' by splitting trajectories into pseudonymous sub-trajectories – in other words, breaking it up into multiple pieces.

In practice, there are risks that such datasets could be de-anonymized by a malicious actor; that is, those sub-trajectories could be reconstructed back into their original trajectories, potentially exposing information related to individuals.

It is a task not unlike completing an unknown number of jigsaw puzzles after being given a stack of pieces. You find the next piece that fits with the current partial reconstruction, and then determine when one single puzzle is complete. While the first problem requires estimations of how likely two pieces fit together, the second problem assesses how a single piece fits in the big picture. Arduous work, but modern computing makes it easier.

Playing the role of attacker, HERE researchers have used a novel mix of machine learning algorithms and graph theoretical methods to "stitch" back together – with a high level of reliability – location trajectory data that had been anonymized using segmentation and suppression methods. The reconstructed trajectories could then be used to discover information related to individuals.

The goal is, of course, not to run roughshod over people's privacy. Instead, such a reconstruction attack framework could be used by organizations to assess how robust their anonymization techniques may be and ensure they make more informed choices about their anonymization parameters.

Want to talk? We do, too. Get in touch here.

Four ways cities can safeguard citizens' location data privacy

Many privacy incidents are, of course, unintentional. Any entity that provides consumer data to outside parties, including HERE, is capable of inadvertently providing information, which may enable the identification of the people from whom it originates.

Some cities and businesses may rest easier if data is confined internally to improve their own services or guide decision-making, such as where to place the next cycle path. However, many will also share data with third-party organizations or open it up to the public, and that calls for more careful consideration of the privacy payloads involved.

We recommend that public sector organizations keep in mind the following privacy recommendations:

Four ways cities can safeguard citizens' location data privacy

#1: Remember that location data is often personal

Our first recommendation is about mindset rather than technology. People are sensitive when it comes to their location data, and understandably so. Our location footprint tells the world a lot about ourselves – arguably more so than what we search for online.

HERE's research has shown that there is a lot of mistrust, concern and uncertainty over how companies and service providers collect and use location data.

In our 2018 survey of more than 8,000 people across eight countries[6], we found that individuals feel their trust is abused when there are insufficient controls for the management of personal data, coupled with a lack of transparency on the part of data collectors. Two in five people discovered they share location data with more apps than they thought. Only one in five felt they have full control over their location data, while even fewer felt that their data will be stored and used properly.

Our more recent research[7] suggests that the trust issue remains, with 75% still concerned about sharing personal information digitally.

This lack of trust is a cause for concern today. But it also raises further questions about how current click-to-consent privacy practices can stand up in the future world of autonomous transportation, checkout-free grocery stores, and drone-delivered packages. If your privacy is inherently about control over information relating to you, how will you control your exposure to others in urban environments where communications are machine-to-machine and time-sensitive?

70% of consumers are willing to share their location data sometimes, usually, very often or always.

Only **26%** of consumers trust that services collecting their location data will handle their data as they should.

Sharing their location data makes **32%** of consumers feel vulnerable or stressed; **38%** are nervous about burglaries, stalkers or digital/physical harm when sharing their location data.

Only **33%** of consumers feel they are aware of what happens with their personal information after they share it with a data collector.

Only **29%** of consumers feel they have full control over their location data.

Source: "The Privacy Paradox Reloaded: Changes in Consumer Behavior and Attitudes since 2018", HERE Technologies, 11 September 2019. Research Survey of 10,176 people in 10 countries (Australia, Brazil, China, France, Germany, India, Japan, the Netherlands, UK and U.S.)[7]

Want to talk? We do, too. Get in touch here.

Four ways cities can safeguard citizens' location data privacy

#2: Build in privacy from the get-go

Privacy by design is a sound approach in the development of any product or service. It can also be applied to standards, such as MDS. In our experience, it is always much easier to incorporate relevant privacy considerations into any standardization effort from the outset. While standards do evolve, if privacy is an after-thought things can become problematic.

We have seen this in the development of the Cooperative Intelligent Transport Systems (C-ITS) platform in the EU. Andrea Jelinek, the Chair of The European Data Protection Board, for example, has expressed that some issues relating to privacy remain unresolved, including the application of data protection by design and by default in C-ITS services.[8]

Foreseeable data uses and privacy protection methods should therefore be considered early on. This is sometimes easier said than done but can be achieved when involving right stakeholders.

Four ways cities can safeguard citizens' location data privacy

#3: Reduce data collection to only what you need

One of the simplest ways of preserving the location data privacy of users is to tailor data collection for the intended use case and minimize what you collect. HERE has sought to do this with respect to certain services.

For example, to build a live picture of the road environment for our road hazard alerts service, we crowdsource data captured by the sensors of millions of vehicles. However, all the data is anonymized before it comes into our platform. This reduces the likelihood that a vehicle or its owner could be identified.

Similarly, for our smartphone positioning service, which is used by leading smartphone brands to help calculate the 'blue dot' that represents your location on a map, we limit data collection to selected attributes like cell and Wi-Fi traces and do not collect additional parameters such as the IMEI of devices that could be used to identify an end-user.

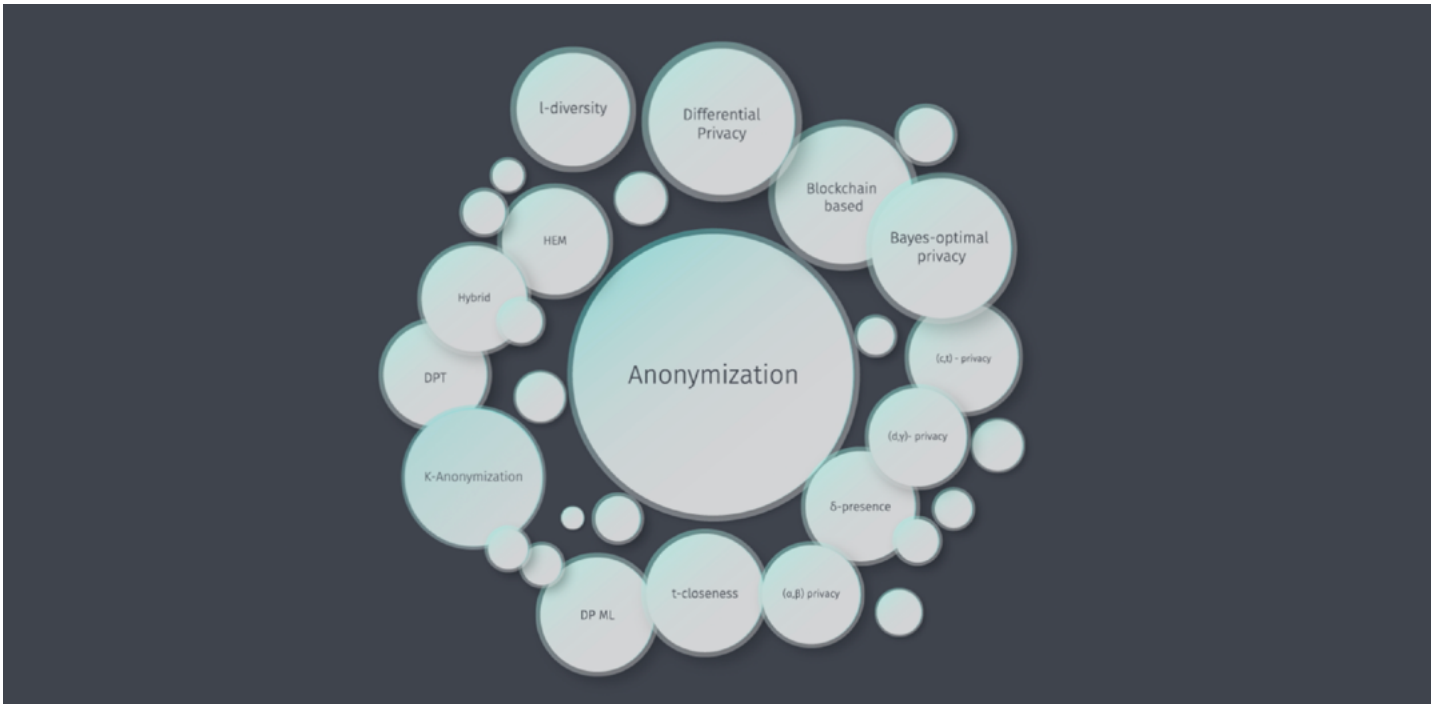
Four ways cities can safeguard citizens' location data privacy

#4: Consider new and emerging privacy-enhancing technologies

Use case specification also helps us understand the opportunities for data anonymization and other privacy-enhancing methods. At HERE, we use various pseudonymization techniques to reduce the possibility of data being connected to individuals while retaining a required level of utility. For example, for our traffic service, we apply automated random re-assignment and rotation of ID numbers associated with vehicles and devices supplying traffic probe data.

In the near future, the anonymization and privacy toolset will expand significantly. Novel technologies, such as differential privacy and federated machine learning, offer powerful new ways for cities to enhance their citizens' privacy, while simultaneously preserving and maximizing the utility of their data.

There is no one-size-fits-all when it comes to privacy, which is why organizations are likely to employ a mix of methods to meet their and their users' needs.



 Want to talk? We do, too. Get in touch here.

Geospatial differential privacy

Already widely used by various technology companies, this technique involves adding random noise to an original dataset to make it less precise, while not changing the aggregate meaning of that data. While the user's exact location is masked, the data is still useful for certain services such as location-based advertising. This approach is also useful in machine learning, enabling algorithms to access more data for training purposes.

Federated machine learning

Federated machine learning is one potential avenue for learning from user, enterprise or city data while still giving those entities a degree of control and privacy. This is because federated learning makes it possible that end-users and businesses never need to transfer their data.

With this approach, the model is transferred to the user, enterprise or city to learn from their data locally. When the data is aggregated and modeled locally, only the derived knowledge, or model parameters, is shared back to the cloud, without personal information.

Privacy diagnostic measurements

This is not an anonymization tool in itself, but rather a way of testing how well your anonymized dataset stands up to a reconstruction attack. A machine learning-driven model attempts to de-anonymize your data and provides a privacy risk rating.

Such an approach could form the basis of an analytical framework for enterprises to assess their own privacy practices, helping them weigh up which methods of anonymization provide the best privacy protection while simultaneously preserving data utility.

Want to talk? We do, too. Get in touch here.

Privacy as an opportunity

Removing or manipulating data, yet leaving behind just enough for a specified purpose, is a fine line to walk. In any event, anonymization is something that should never go unquestioned. We cannot just say “oh, we’ve anonymized that data and we’re good.” Broadly speaking, organizations can no longer just pay lip service to location data privacy. Mere regulatory compliance will not be sufficient to ensure success in the new business environment.

But nor should privacy be seen as a burden. Under the umbrella of anonymization, there are different methods and technologies that can help turn privacy into an opportunity.

Want to talk? We do, too. Get in touch here.

Sources

1. Yves-Alexandre de Montjoye, Cesar A. Hidalgo, Michel Verleysen & Vincent D. Blondel: "Unique in the Crowd: The privacy bounds of human mobility," Scientific Reports, 2013.

2. "NYC Taxi Data Blunder Reveals Which Celebs Don't Tip–And Who Frequents Strip Clubs," Fast Company, 2 October 2019.

3. "Fitness app Strava lights up staff at military bases", BBC News, 29 January 2018.

4. Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh and Alex "Sandy" Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata," Science, 30 Jan 2015.

5. "The Los Angeles Department of Transportation's Ride Tracking Pilot is Out of Control," EFF, 9 April 2019.

6. "Privacy and Location Data: Global Consumer Study," HERE Technologies, 5 March 2018; survey of 8,073 people across eight countries (Australia, Brazil, France, Germany, Japan, the Netherlands, UK and U.S.)

7. "The Privacy Paradox Reloaded: Changes in Consumer Behavior and Attitudes since 2018," HERE Technologies, 11 September 2019. Survey of 10,176 people across 10 countries (Australia, Brazil, China, France, Germany, India, Japan, the Netherlands, UK and U.S.)

8. Letter from Andrea Jelinek to Henrik Hololei, Director-General of the Directorate-General for Mobility and Transport at the European Commission, 5 June 2019.

Want to talk? We do, too. Get in touch here.

About HERE Technologies
HERE, a location data and technology platform, moves people, businesses and cities forward by harnessing the power of location. By leveraging our open platform, we empower our customers to achieve better outcomes - from helping a city manage its infrastructure or a business optimize its assets to guiding drivers to their destination safely. To learn more about HERE, please visit **360.here.com** and **here.com**