



DATA GOVERNANCE HELPS UNLOCK THE TRUE POTENTIAL OF LOCATION DATA

Abstract

Location data is one of the most powerful data sources in an ever-evolving digital world, where data-driven innovations drive business models. While location data helps offer a multitude of personalized services and improved user experience, the associated privacy concerns are eroding users' confidence and becoming major impediment for digital transformation. Despite the despicable side of data misuse, the mounting privacy risks and the pertinent challenges of managing data, location data offers significant social and economic opportunities, particularly in the areas of healthcare, emergency response, urban planning etc. Hence, a comprehensive data strategy driven by robust data governance, user-centric considerations and privacy by design will help leverage the true potential of location data, enabling businesses and public institutions to drive economic value and solve real-life problems, whilst complying with the regulatory requirements.

CONTENTS

| | |
|--------------------------------------------------------|-----------|
| INTRODUCTION: LOCATION DATA – A 360-DEGREE VIEW | 3 |
| Sources of Location Data | 3 |
| BUSINESS USES OF LOCATION DATA | 5 |
| CHALLENGES IN UTILIZING LOCATION DATA | 7 |
| ‘The Data Explosion’ Challenge | 7 |
| Data Standardization and Hygiene | 8 |
| ‘The Privacy Problem’ | 8 |
| MITIGATING PRIVACY RISKS | 8 |
| De-identification Methods | 8 |
| Limit the Data | 10 |
| Federated Learning | 10 |
| Handle Sensitive Locations | 10 |
| CONCLUSION | 11 |
| ABOUT THE AUTHOR | 12 |

Location Data - A 360-degree View

Location data, also known as geolocation data, is the information acquired from an internet-connected device, such as a smartphone or a computer about its physical or geographical position, including latitude, longitude and altitude of the device. It may also provide contextual information, such as the activity at the location, its opening times, contiguity and proximity to other objects etc.

Location data is, most often, collected by a network or a service using Global Positioning System (GPS) satellite imagery, or WiFi access points, or mobile cell towers or a combination of them. For instance, the 'location services' in our mobile devices aggregates data from various sources such as GPS, nearby mobile cell towers and Wi-Fi networks, and bluetooth

as there are hardware sensors embedded in devices allowing them to detect a wide variety of signals.

There has been an evolution in data collection mechanisms of location data over the last decade, with an increasing prevalence of ubiquitous inter-connected devices. Most of the data is now being sourced from non-traditional sources, such as crowd-sourced data.



Figure 1: Sources of Location data

| Data Source | Definition |
|--------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  <p>Global Positioning System (GPS)</p> | <p>A technology that uses signals transmitted by the satellites orbiting the earth to determine a device's location.</p> |
|  <p>WiFi Access Points</p> | <p>A wireless technology that enables devices to emit probes to search for WiFi access points and hotspots.</p> |
|  <p>Cell Tower Triangulation</p> | <p>A technology that uses unique 'Cell Tower IDs' broadcast by the nearby Mobile cellular towers. A device is then located by the overlap of signals when it is within the communication range of three or more towers.</p> |
|  <p>Beacons</p> | <p>These are radio transmitters that transmit bluetooth signals over short distances to devices that are equipped to receive them.</p> |
|  <p>Mobile Apps</p> | <p>App services offered by mobile operating systems are used to collect location information of the device to provide location-based features. Also, third-parties provide SDK (Software Development Kit), a piece of code integrated into apps to collect location information.</p> |
|  <p>Crowd-sourced</p> | <p>Data sourced from users when they provide their address details while signing up for various services. It may also be sourced from social media apps when a user checks-in at a location or adds location while uploading a picture.</p> |
|  <p>Bidstream</p> | <p>Bidstream data is supplied by advertising networks to the advertisers through a real-time auction.</p> |



Business Uses of Location Data

Maps have, for centuries, helped humanity with explorations and navigation through the globe. In today's increasingly digital era, the once-static maps have transformed into a living digital platform – where each point on the map provides rich contextual information and deep insights about people, events, culture and businesses.

As decision-makers use data to analyze customer patterns, identify market needs, assess operational efficacy or strategize a roadmap – the element of “place” matters the most. In fact, only location data can help analyse some of the complex social and economic

trends today – unravelling the relationship of geographical location with people, transactions, events and other associated business factors. Real-time location data helps unleash the ‘power of where’ – answering questions at a very granular level, such as:



Figure 2: The ‘power of where’

The convergence of location data with interrelated enabling technologies such as Advanced Networking, Artificial Intelligence/Machine Learning (AI/ML), Augmented Reality, Internet of Things

(IoT), and 3D technologies are transforming the core of traditional businesses in ways one can only imagine. It provides digital insights and contextual information, geospatial analytics, and offers visualization

of the data, enabling policymakers and businesses to deliver location-intelligent solutions, thereby improving our everyday lives cutting across a wide range of industry sectors and public services.

“You can't use an old map to explore a new world.”
Albert Einstein





By 2025 the UK will have a coherent national location data framework. Future technologies will be underpinned by data about events occurring at a time and place. Location data will be the unifying connection between things, systems, people and the environment.



Unlocking the power of location:
The UK's geospatial strategy. Geospatial Commission, June 2020

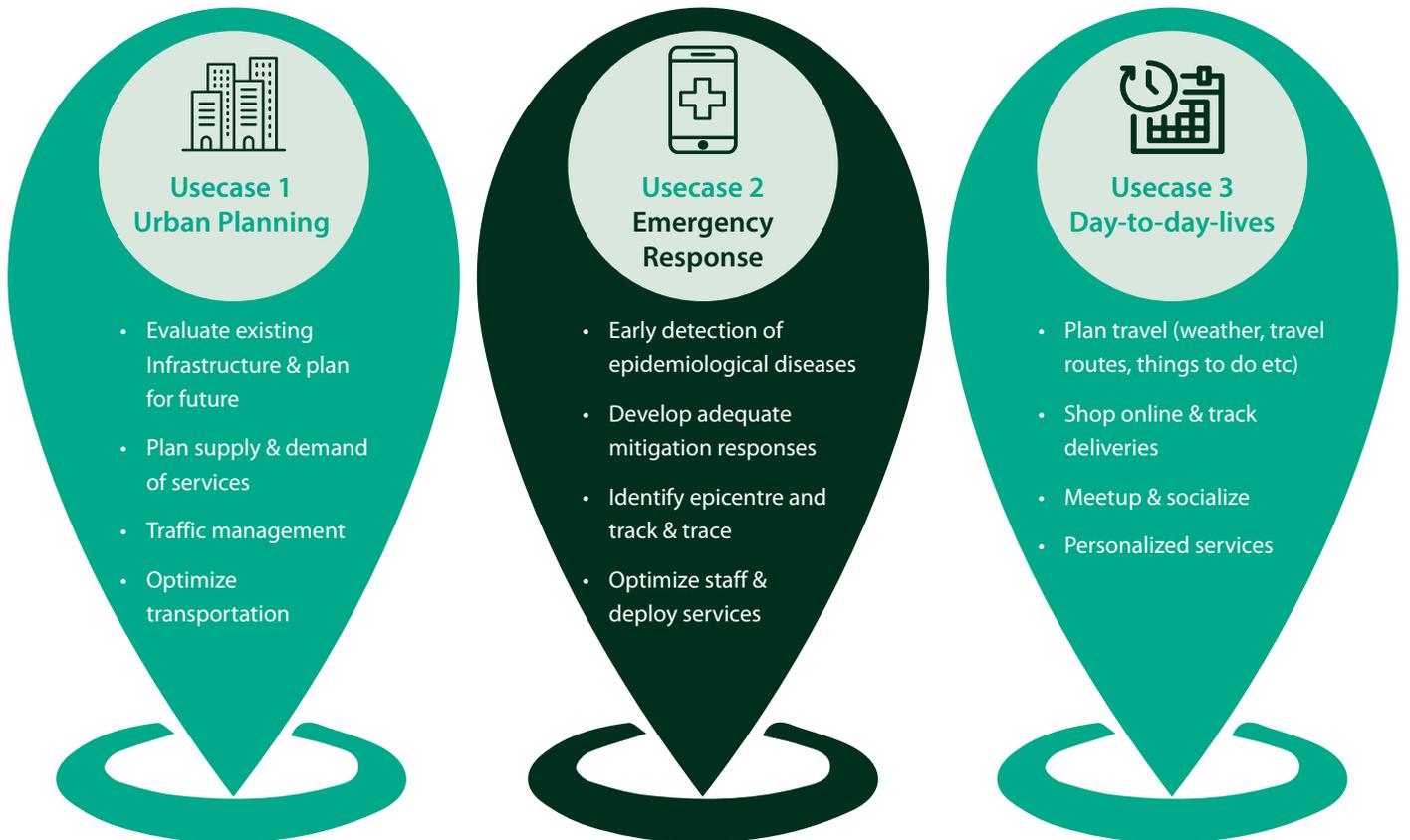


Figure 3: Usecases of Location Data

Challenges in Utilizing Location Data

While location data, in an ever-evolving digital landscape, offers enormous new potential, it comes with its share of challenges due to the increasingly complex location data ecosystem, besides the associated privacy risks. The complexities include a distributed data landscape that makes data harder to track and govern, poor data quality, multi-sourced and disparate datasets, data standardization etc. Therefore, data governance is one the

most important aspects determining the usability of location data. It is a continuous process of managing ever-increasing amounts of data, improving its quality and regulating its application, which allows for its efficient use, enabling organizations to fully leverage its true potential and minimize risks. A robust data governance strategy also ensures data privacy and regulatory compliance with data protection laws.

In this regard, the Geospatial Commission has published a study of the UK's location data, illustrating how location data offers significant social and economic opportunities. It calls for location data to be made a strategic national asset for the country's data economy.

Let's take a look at some of the critical challenges in play and how a robust, modern and unified data governance program helps overcome them.

'The Data Explosion' Challenge

A highly distributed data environment, the expanding cloud footprint, and an ever-increasing volume of data have created substantial opportunities, while bringing to fore the challenge of tracking and governing data; with diverse sources of location data and complicated data handling mechanisms, it becomes more important than ever to have a scalable, automated and centralized data governance strategy.

Centralized data governance ensures that enterprise standards for data management are followed consistently across the organization. Furthermore, the various components and technologies in the distributed data environment must be centralized and standardized, which not only helps achieve uniformity but also helps bring the licensing cost down.

The increased relevance of cloud and other enabling technologies have resulted in diverse types of data, a

growing community of data consumers, and a continuous addition of new data sources and systems, hence, necessitating the use of automation in data governance functions such as data classification, data cataloguing, data profiling, metadata management etc, as traditional approaches are proving to be unsustainable. Additionally, the access and provisioning of data has to be automated via a rules engine creating a fine balance between protecting data and making it accessible.

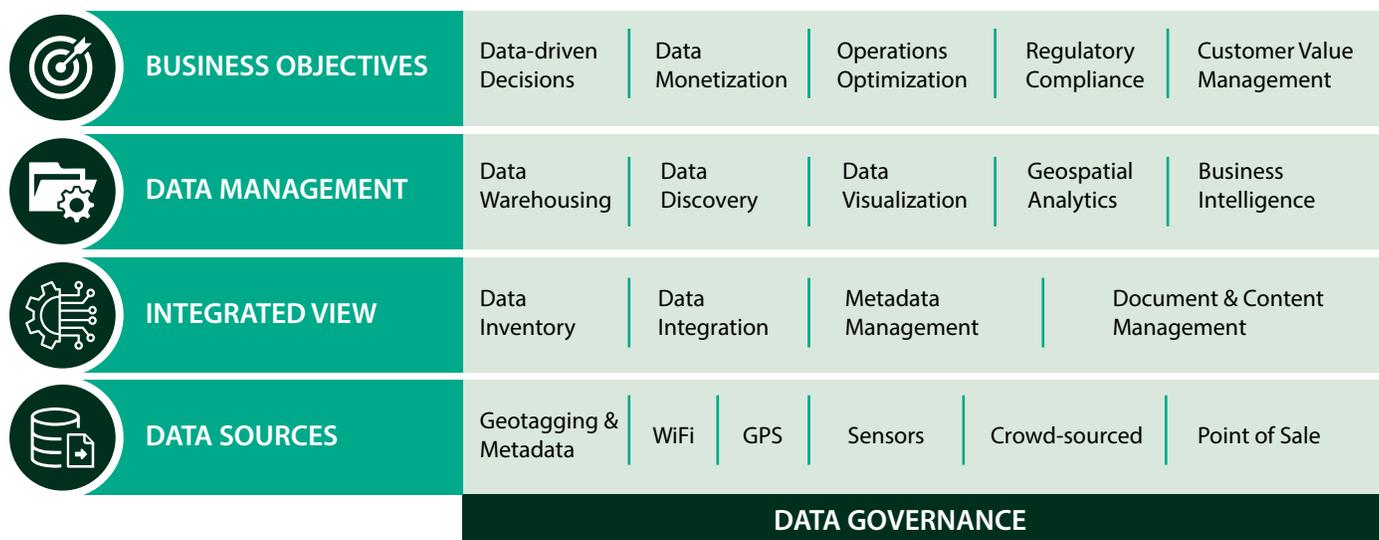


Figure 3: Location Data Strategy Map

Data Standardization and Hygiene

Huge volume of data doesn't necessarily translate into quality data and the quality of data is a key consideration while investing in data programs. Besides the inaccuracies in the location data that providers offer, the other major challenge relates to the lack of location data standards, hurting the trust in data

and adversely affecting the reliability of data-based business decisions.

For instance, a standard address format is inevitable for feeding data into location-intelligence models in order to provide efficient location-based services. However, address formats are usually marred with data entry errors, missing fields and abbreviation errors or other issues due to inaccuracies in digitizing places.

It is also important that common data formats are developed to store and process location data. Furthermore, raw location data often comprises of erroneous or incorrect data and other anomalies that necessitate data cleansing.

A robust data governance helps address this challenge as it enforces clear policies to ensure that the data remains consistent and harmonized.

'The Privacy Problem'

Location datasets largely comprise of personal and sensitive data, often the device's IP address or the precise location information collected by mobile applications, which can be used to learn the user's exact location. This data when linked with additional personal information could reveal personal behaviour of the user constituting a potential risk to the individual's privacy rights. Hence, users today are becoming increasingly wary

about how their data is being collected, who it is being shared with and what it is being used for.

As per a research, an estimated \$12 billion market is built on monetizing people's movements. There have been multiple reports of mobile applications selling location data to data brokers, thereby threatening the privacy of individuals. This becomes even more critical after the United States' Supreme Court overturned the Roe vs Wade decision, raising

apprehensions that some states may use information from period tracking apps to prosecute women travelling to other states to get an abortion. This serves as a textbook example of how data privacy is much more than a compliance issue and impacts our daily lives.

Therefore, the privacy risks and its consequences not only result in decline of users' trust and cause reputational damage, but also incur massive regulatory fines.

Mitigating Privacy Risks

Following are some of the key steps to mitigate the privacy risks around processing of location data:

1. De-identification Methods:

Data de-identification refers to the processes by which the identity of an individual can't be ascertained from the data. The purpose is to minimize the risk of connecting the data to an individual so as to protect the individual's identity from being revealed. The various de-identification techniques provide varying degrees of obscurity and should be used in combination for effectiveness. Let's discuss some of these in detail:

- **Anonymization:** It is a technique by which the personal data of an individual is irreversibly altered in such a way that it can no longer be associated with the specific individual. This is achieved by masking or removing the personal data attributes from the dataset such that utility of the data is not lost.

| ID | Name | Postcode | Age | Gender |
|-------|-----------------|----------|-----|--------|
| 10477 | John Smith | E51 A0 | 27 | M |
| 10565 | Maria Jones | N21 8D | 35 | F |
| 10461 | Robert Williams | W12 0S | 37 | M |
| 10338 | David Miller | M75 9W | 32 | M |
| 10291 | Linda Brown | SW9 2X | 22 | F |
| 10360 | Mark Taylor | M69 2D | 39 | M |

Direct Identifiers

Indirect Identifiers

Figure 4a: Original Dataset

| ID | Name | Postcode | Age | Gender |
|-------|-----------------|----------|-----|--------|
| 10477 | kcjn qvwzj | y78 x3 | 77 | X |
| 10565 | vxuwx kcneq | N18 8z | 10 | A |
| 10461 | uczeuz Wwaawxvq | W81 3q | 60 | B |
| 10338 | zxvwz vwaaeu | v77 7W | 41 | X |
| 10291 | awnzx zucwn | qW7 1X | 61 | S |
| 10360 | vxuk zxyacu | v67 1z | 11 | B |

Figure 4b: Anonymized Dataset

- **k-Anonymity:** The likelihood of re-identification of the anonymized data by linking it to other datasets through indirect identifiers exists, although low. Once direct identifiers have been masked, k-Anonymity may be used to manage the risk of re-identification. It is a technique applied on the

anonymized data by which every individual should be indistinguishable from k-1 other individuals in a dataset where there are at least k individuals who share same values for the same set of data attributes. Hence, this technique is also referred to as 'hiding in the crowd'.

| ID | Postcode | Age | Gender |
|-------|----------|-----|--------|
| 104** | E**** | 27 | M |
| 105** | N**** | 35 | F |
| 104** | W**** | 37 | M |
| 103** | M**** | 32 | M |
| 102** | S**** | 22 | F |
| 103** | M**** | 39 | M |

Figure 4c: k-Anonymous Dataset (k=2)

- **Data Aggregation:** This method reduces the granularity of the dataset by grouping the data together and provides a broader picture of the data, thereby minimizing the risk of sensitive information about individuals getting revealed. The values within a data attribute are grouped to form a single value, instead of providing raw data. For location data, the granularity of addresses could be reduced to city or county level.

| City | Gender | No. of people |
|------------|--------|---------------|
| Manchester | M | 2 |
| London | F | 2 |
| London | M | 2 |

Figure 4d: Breakdown by City & Gender

| City | Age group | No. of people |
|------------|-----------|---------------|
| Manchester | 30-39 | 2 |
| London | 20-29 | 2 |
| London | 30-39 | 2 |

Figure 4e: Breakdown by City & Age group

- **Differential Privacy:** The aggregated data is vulnerable to differencing or reconstruction attacks, wherein the attackers identify the individual by running multiple statistical programs against the aggregated data, thereby reconstructing the original data. That's where 'Differential Privacy' comes to the rescue. It is a data randomization method of adding

a controlled amount of noise to the values of attributes in an aggregated dataset such that the privacy of the individuals is protected without significantly changing the overall result. Therefore, anyone trying to identify an individual from this dataset will not be able to tell if the values of the attributes are correct as long as the noise being introduced is not predictable.

| Original Data | New Data |
|---------------|----------|
| Age | Age |
| 27 | 27 |
| 35 | 35 |
| 37 | 37 |
| 32 | 32 |
| 22 | 22 |
| 39 | 39 |
| | Noise |

Without noise:
Average age = 32

Average age:
32 + Noise = 32.5

Figure 4f: Differential privacy example



2. Limit the Data:

Simply put, the sensitive data that doesn't exist results in a risk that never materializes. Hence, the principles of 'Data Minimization' and 'Purpose Limitation' must be taken into consideration throughout the product lifecycle.

- Organizations must assess the amount of data needed for their business activity as well as the extent of processing of the data being collected.
- Organizations must ensure that data is used only for the purposes it was originally collected for.
- Context-based access controls will help restrict access to the data and provide a more granular access management based on attributes such as user identity, device, IP address, location, etc.

3. Federated Learning:

In simple words, Federated Learning is a de-centralized form of machine learning. The idea is to have training models run across individual edge devices, instead of training the models in data centers. This means that users' raw data remains local and doesn't leave their devices and only the locally computed updates are aggregated and shared with the centralized model. This technique not only helps mitigate privacy risks, but also helps optimize the costs. However, it is also vulnerable to reconstruction attacks and hence, privacy-preserving methods such as differential privacy and homomorphic encryption must be used.

4. Handle Sensitive Locations:

Tracking the precise co-ordinates of the user may not be required for providing most location-based services. Hence, precise location must be avoided, and instead approximate location or general location must be used wherever possible. Users must be given control over how they would want to share their location information. Furthermore, sensitive locations such as home or workplace must be treated with high levels of security. Most importantly, the product design must be configured with default settings that are as privacy-protective as possible.





Conclusion

With 15 billion operational ubiquitous mobile devices in the world and at least 80% of the data today having a location element, location data underpins today's digital society. It has fostered a wide range of consumer services and highly contextualized user experiences revolutionizing our daily lives.

There is an enormous potential that location data offers, and its true value can only be realized when you can address the challenges of managing and governing the data, and the ethical and privacy implications. Hence, a modern data governance strategy has emerged as a strategic business imperative.

About the Author



Mohammed Arifuddin

Data Privacy Architect with Data & Analytics unit at Infosys.

He has 14 years of experience in driving implementation of strategic initiatives in Data management and building Data privacy solutions. He has extensive knowledge in GDPR, holds a PG certificate in Cyber Law and is a OneTrust certified privacy professional.

References

<https://themarkup.org/privacy/2021/09/30/theres-a-multibillion-dollar-market-for-your-phones-location-data>

For more information, contact askus@infosys.com



© 2022 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior permission of Infosys Limited and/ or any named intellectual property rights holders under this document.