Location Data: GPS, Wi-Fi, and Spatial Analytics

Class 2 of Digital Data Flows Masterclass:
Emerging Technologies

27 November, 2018 | Brussels

FUTURE OF PRIVACY FORUM
VUB
BRUSSELS PRIVACY HUB

DIGITAL DATA FLOWS MASTERCLASS: EMERGING TECHNOLOGIES

Curriculum	Date*
Session 1: Artificial Intelligence and Machine Learning - featuring Dr. Swati Gupta, Assistant Professor in the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Tech; and Dr. Oliver Grau, Chair of ACM's Europe Technology Policy Committee, Intel Automated Driving Group, and University of Surrey	25 October, 2018 - side event, ICDPPC (Brussels) (with remote participation)
Session 2: Location Data: GPS, Wi-Fi, and Spatial Analytics	November 2018 - Brussels (with remote participation)
Session 3: Advertising Technologies: Online Data Flows, Behavioral Targeting, and Cross-Device Tracking	January 2019 - Brussels (with remote participation)
Session 4: Mobile Apps: Operating Systems, Software Development Kits (SDKs), and User Controls	March 2019 - Virtual
Session 5: Transportation and Mobility: Video Analytics, Sensors, and Connected Infrastructure	April 2019 - Virtual
Session 6: Biometric Data: Facial Recognition, Voice, and Digital Fingerprints	June 2019 - Virtual
Session 7: Tracking in Physical Spaces: Retail Technologies, Smart Homes, and the "Internet of Things"	July 2019 - Virtual
Session 8: De-Identification: Multi-party Computing, Differential Privacy, and Homomorphic Encryption	Sept. 2019 - Virtual

*dates may change.

All sessions are free and will support remote participation.
Priority registration may be held for government staff.
Enroll and receive updates on the full course at: www.fpf.org/classes

FUTURE OF PRIVACY FORUM

AGENDA

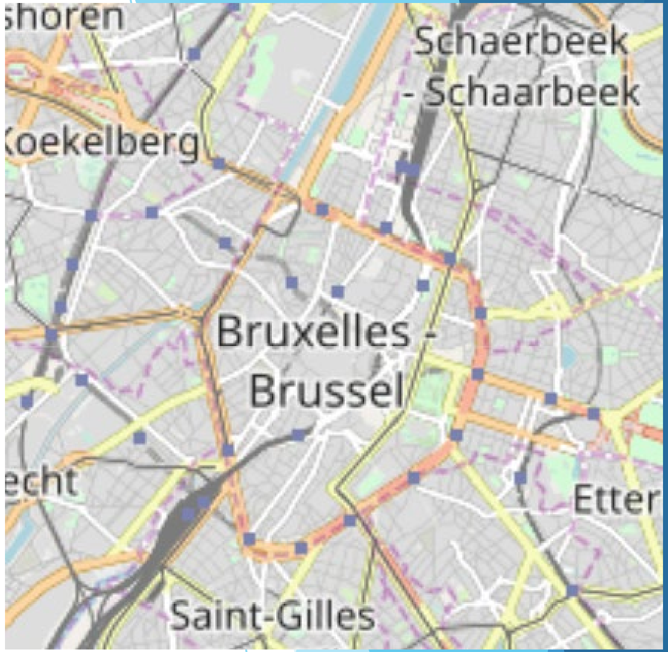
I. Introduction to Geo-Location Data
II. Sources of Data: *Mobile Sensors, Wi-Fi Analytics*
III. Data Flows & Case Studies
IV. Current De-identification Methods

FUTURE OF PRIVACY FORUM
VUB
BRUSSELS PRIVACY HUB

I. Introduction

FUTURE OF PRIVACY FORUM VUB BRUSSELS PRIVACY HUB

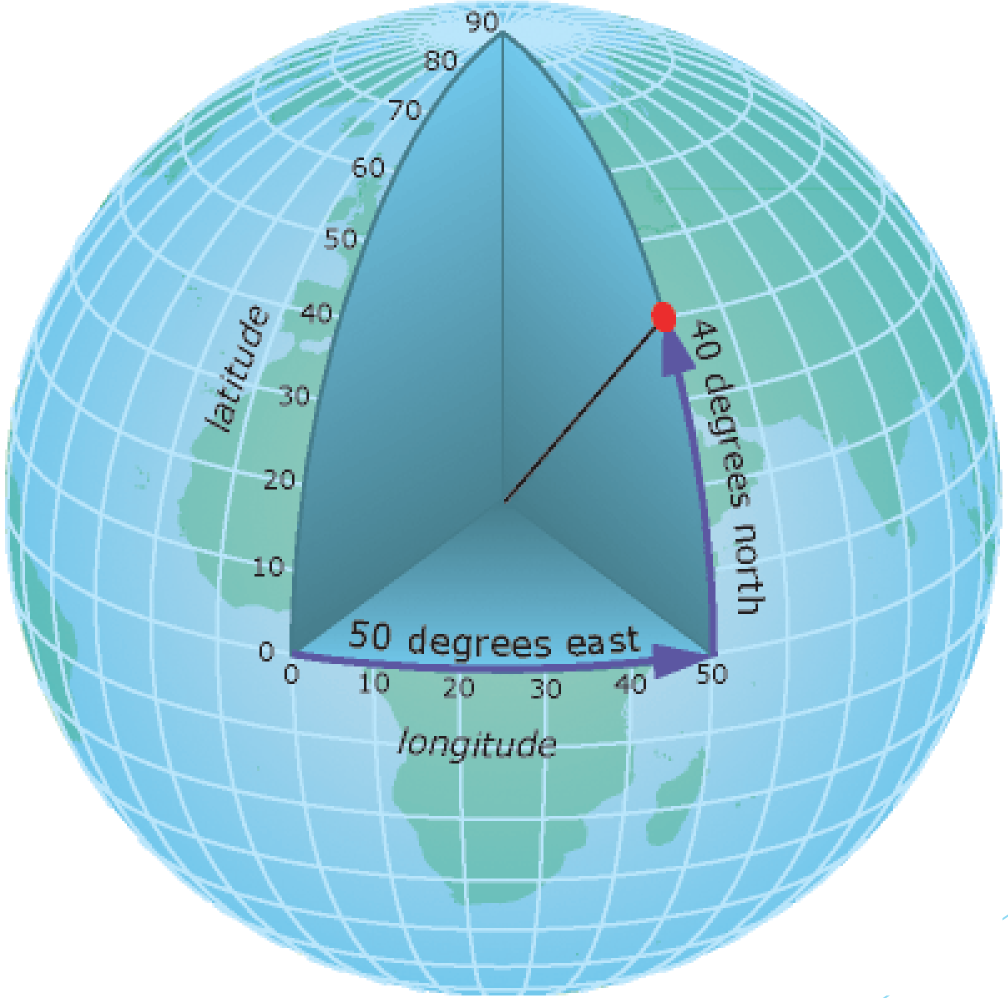
Digital maps & Geographic Information Systems (GIS)



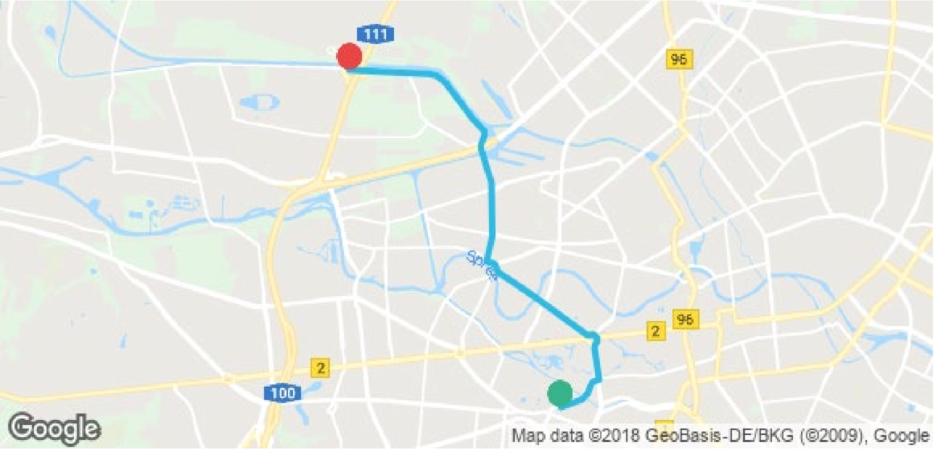
- A digital representation of the real world – a “**geobase**”
- Loaded with layers of additional information, static & dynamic
- Queried for e.g. *“what are geocoordinates of object XYZ?”* or *“at geocoordinates x,y what objects exist there?”*
- Visualized using a map projection
- Created and maintained using surveying, crowd sourcing and lots of computing & labor

6 dimensions of location data

- Latitude
- Longitude
- Altitude
- Time
- Frequency
- Precision

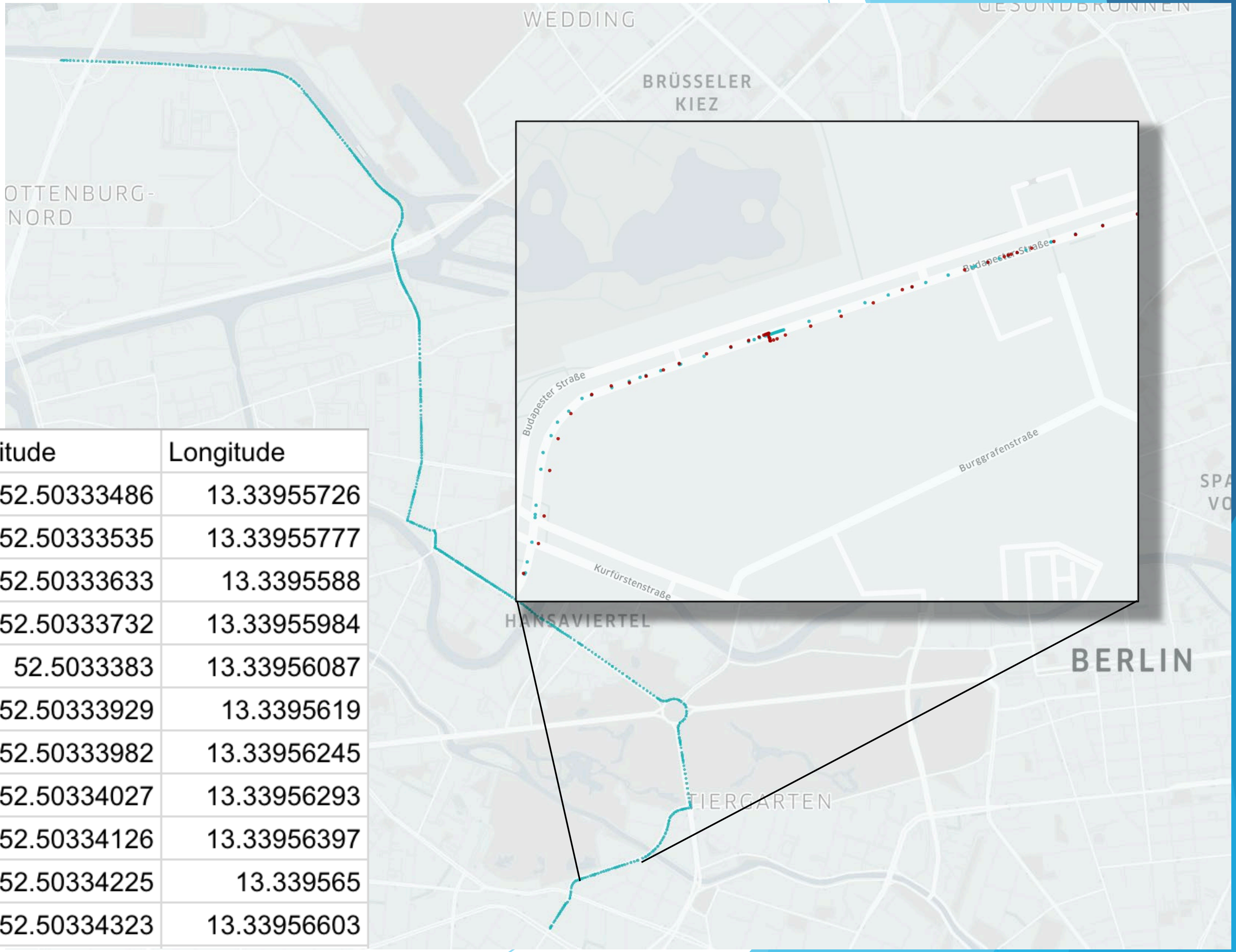


Example: Uber ride in Berlin



- Smartphone based location data
- Collection every 2 seconds
- Map matched to reduce inaccuracy

Latitude	Longitude
52.50333486	13.33955726
52.50333535	13.33955777
52.50333633	13.3395588
52.50333732	13.33955984
52.5033383	13.33956087
52.50333929	13.3395619
52.50333982	13.33956245
52.50334027	13.33956293
52.50334126	13.33956397
52.50334225	13.339565
52.50334323	13.33956603



FUTURE OF PRIVACY FORUM
VUB
BRUSSELS PRIVACY HUB

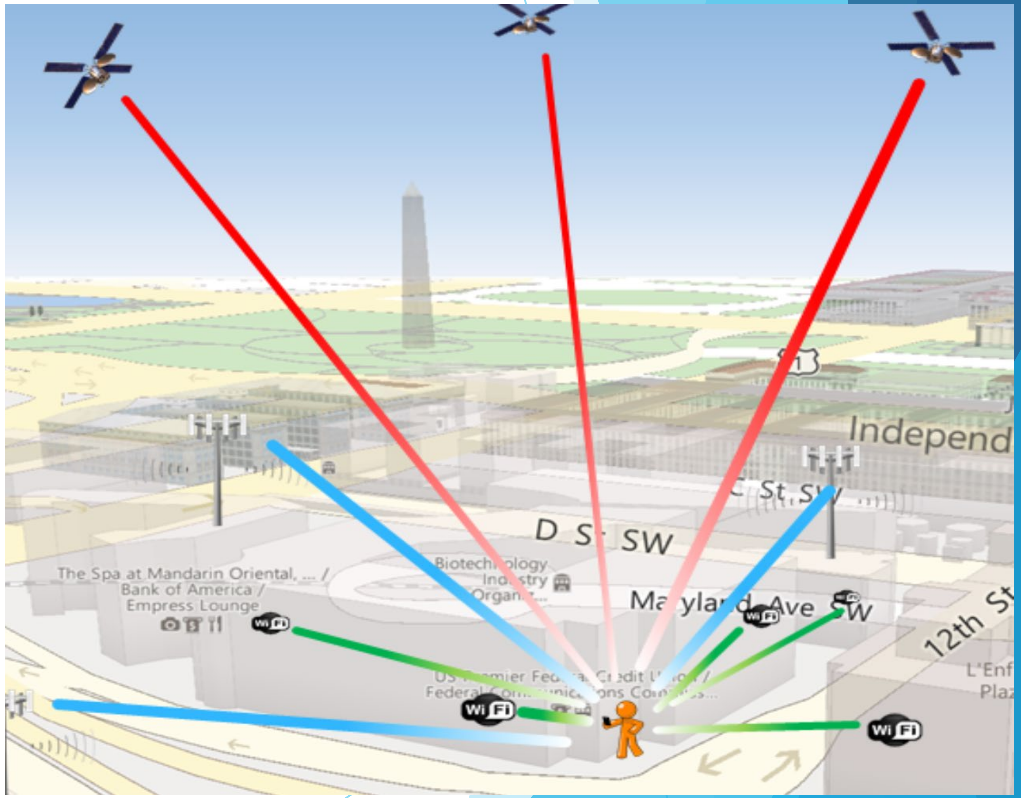
II. Sources of Data

FUTURE OF PRIVACY FORUM VUB BRUSSELS PRIVACY HUB

Sources of Location Data

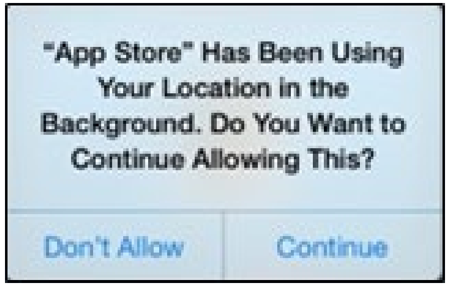
- **“Location Services” and Platform Controls**
- **Hardware Sensors:**
  - GNSS/GPS
  - Nearby Cell Towers
  - Nearby Wi-Fi Networks
  - Beacons and Proximity
  - Emerging Alternatives: LED, Audio

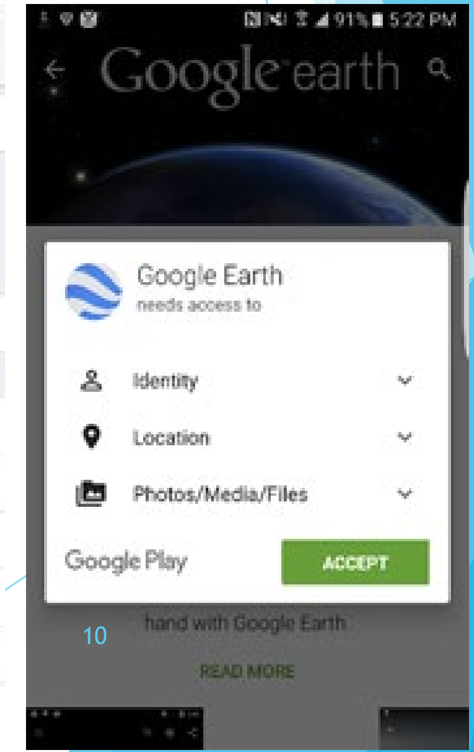
- **Connectivity Information,** e.g. CSLI
- **Wi-Fi Analytics (Tracking in Public Spaces)**



Smartphone "Location Services"

- Operating system (e.g. iOS or Android) controls access to a device's geo-location
- Apps/websites must usually get user permission
- ***Location Services*** aggregates data from many different sources—including satellites, nearby cell towers, nearby Wi-Fi networks, and Bluetooth





Hardware Sensors

iPhone 7
source: ifixit.com, Creative Commons

Accelerometer
Ambient light
Barometer
Bluetooth Radio
Cameras
Cellular Radio
Compass (Magnetometer)
Face ID (iPhone)
GPS Receiver
Gyroscope
Microphones
Moisture sensor
Touch ID
Wi-Fi Radio

FUTURE OF PRIVACY FORUM
VUB
BRUSSELS PRIVACY HUB

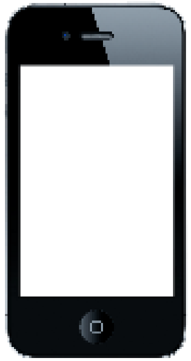
Mobile sensors generate input for OS to provide standardized latitude-longitude

ID=NULL

back to OS to improve services

"Location Services"

Accelerometer
Bluetooth
Cell Towers
Compass
GPS
Gyroscope
Wi-Fi Networks



38.9072° N, 77.0369° W

+ Identifier
e.g. the device's mobile Ad ID:
6D92078A-8246-4BA4-AE5B-76104861E7DC

Apps + Third Parties using **SDKs** in Apps

A closer look at…

1. *GPS*
2. *Cell Tower IDs*
3. *Wi-Fi Networks*
4. *Bluetooth Beacons*
5. *Alternatives – Audio and LED*

FUTURE OF PRIVACY FORUM VUB BRUSSELS PRIVACY HUB

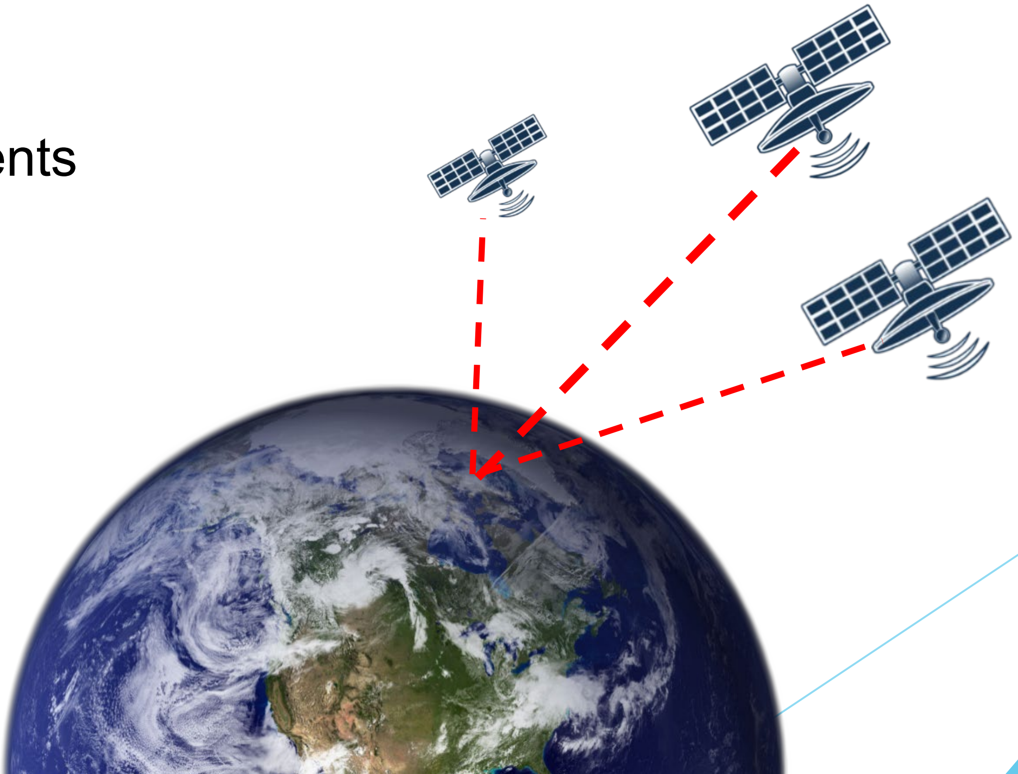
1. Global Navigation Satellite Systems

E.g. Global Positioning System (GPS) (U.S.)

Allows devices to determine their location (latitude-longitude) using time signals transmitted by satellites.

<u>Challenges</u>:

- Weather
- Buildings / urban environments
- Indoor positioning



2. Cell Tower IDs

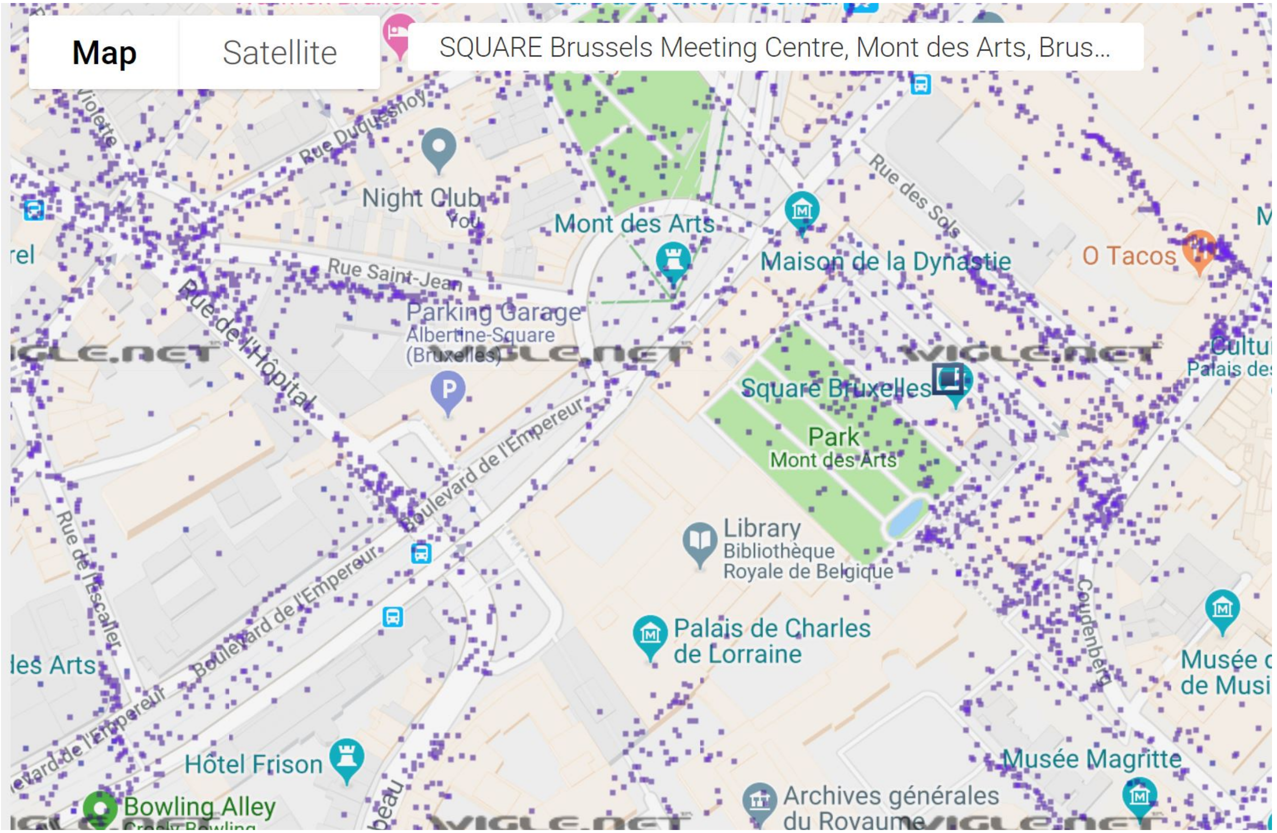
Cell towers broadcast unique **Cell IDs**, which are compiled in both private and publicly available databases. Privately owned databases are often larger, some containing over 72 million unique cell towers.

Approximate location can be inferred by comparing detected Cell IDs and signal strengths with the known locations of cell towers.

Cell Tower Database	*Unique Cell Towers*	*Availability*
OpenCellID	*> 6 million*	*Public*
Combain	*> 72 million*	*Private*
LocationAPI.org	*> 72 million*	*Private*
Mozilla	*> 26 million*	*Public*
Navizon	*> 71 million*	*Private*
Mylnikov GEO	*> 15 million*	*Public*
WiGLE	*> 6 million*	*Private*

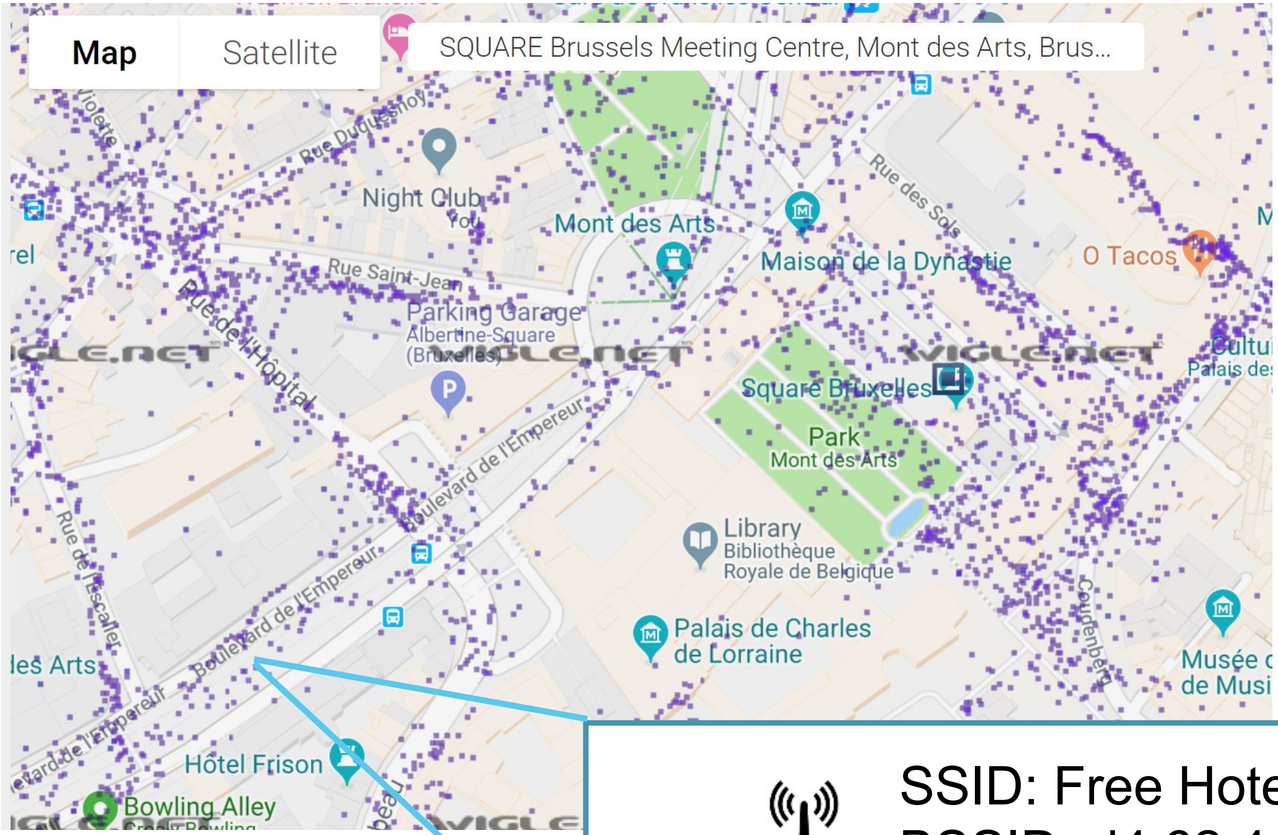
FUTURE OF PRIVACY FORUM VUB BRUSSELS PRIVACY HUB

3. Nearby Wi-Fi Networks



Source: https://wigle.net/

3. Nearby Wi-Fi Networks

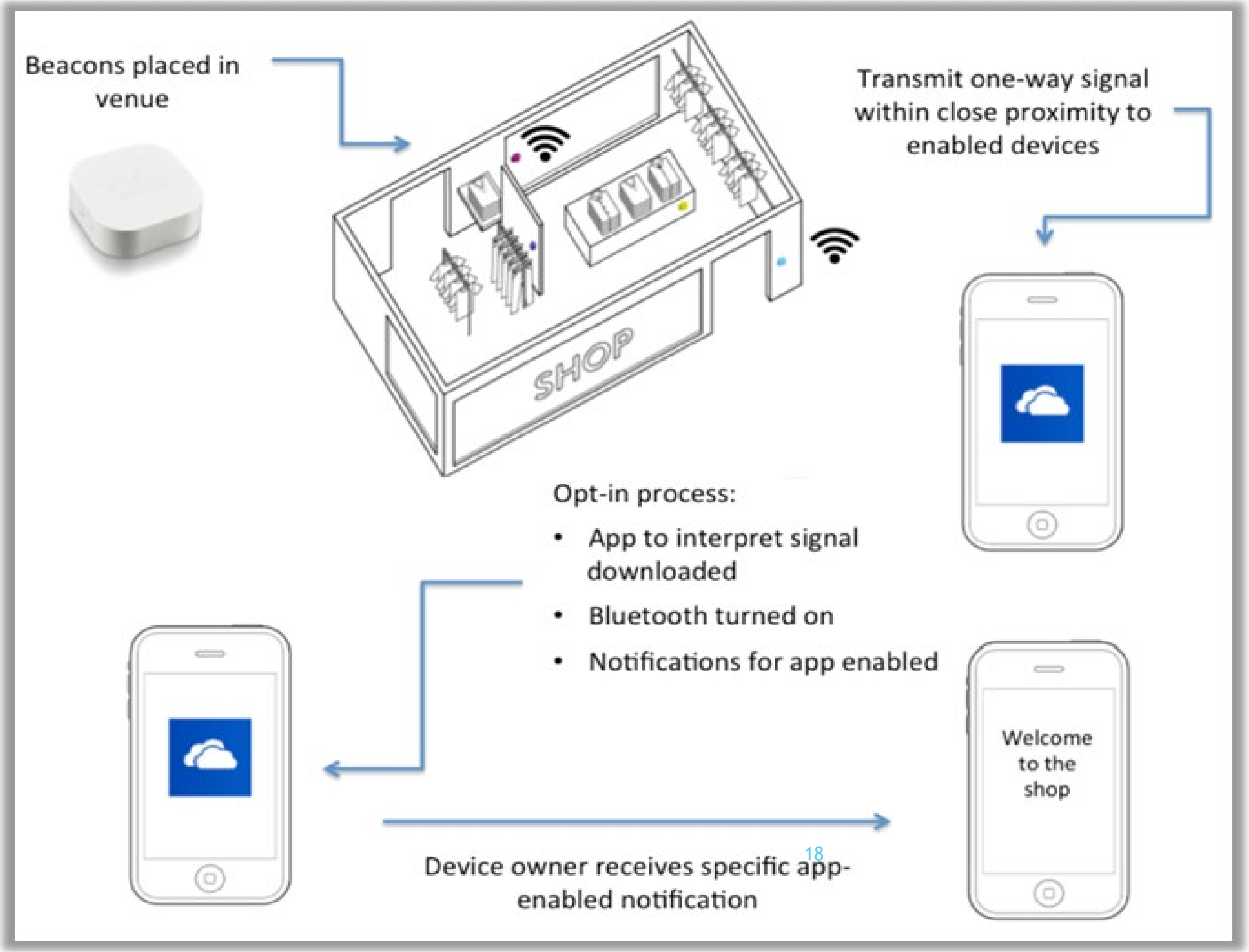


Source: https://wigle.net/

SSID: Free Hotel Wifi
BSSID: d4:62:4d:2c:c8:ec
Vendor: Ruckus Wireless

4. Bluetooth (Beacons)

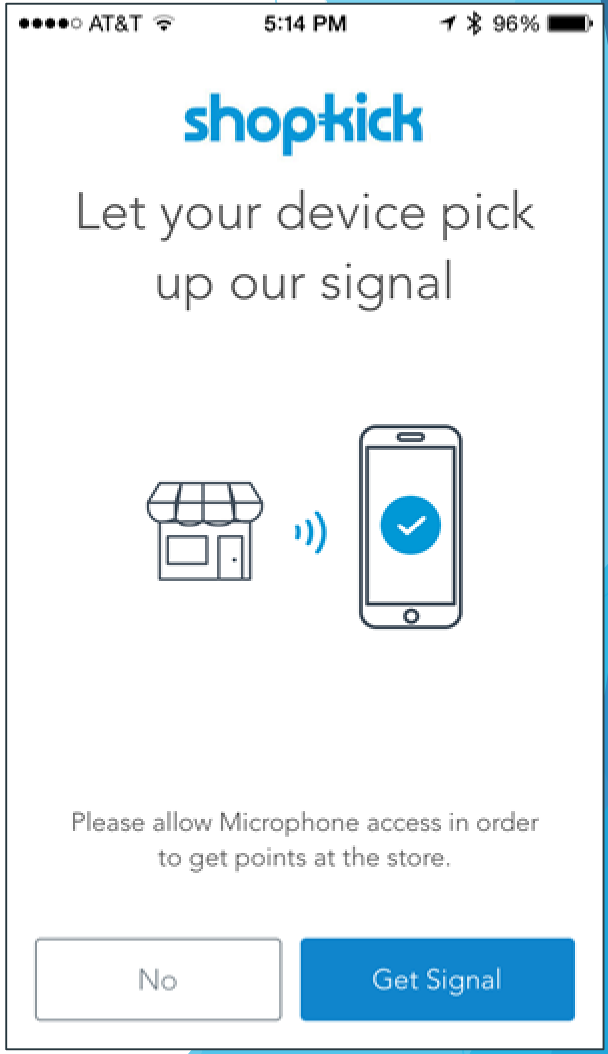
- Beacons are inexpensive radio transmitters that send one-way signals to devices equipped to receive them



FUTURE OF PRIVACY FORUM VUB BRUSSELS PRIVACY HUB

5. Proximity Alternatives – LED, Audio

Requires an **app** with **camera** access permission or **microphone** access permission



Mobile Location Analytics (MLA)

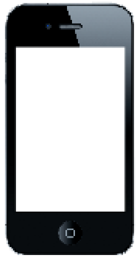
Devices with WiFi or Bluetooth capabilities broadcast their WiFi MAC Address and/or Bluetooth MAC address. Venues use MLA technology to detect how devices are moving within a space or to identify repeat visitors.

- Looks like **68:A8:6D:E5:65:03.**
- Since different device manufacturers have been assigned groups of MAC addresses, your MAC indicates if your device is made by Apple, Samsung or another company.
- Most smartphones now randomize MAC addresses for privacy reasons.

III. Data Flows & Case Studies

FUTURE OF PRIVACY FORUM VUB BRUSSELS PRIVACY HUB

Location Data Ecosystem



First Parties:

- App or website that requests location
- Service providers (e.g. bikeshare company, mobile carrier’s “cell site” location information)

Third Parties:

- Provider of an “SDK” (software development kit) integrated into an app to collect location information, e.g. for advertising or location analytics

Location Data Ecosystem

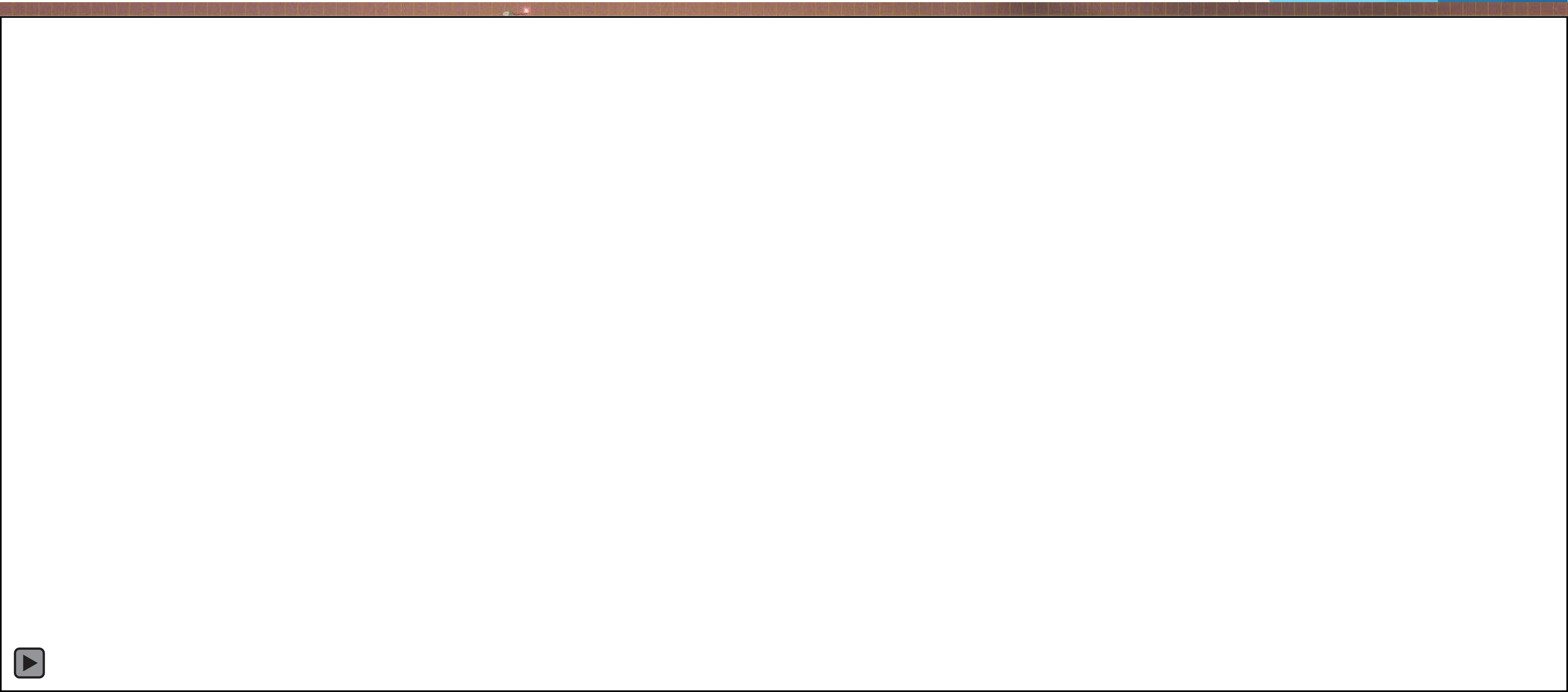
First Party Uses:

- **Raw data** – *e.g. to analyze trends, user behavior, detect security threats, improve a geo-aware service*
- **Geo-fencing** – *e.g. to alert users of local promotions, events, or messages (e.g. Amber alerts)*

Secondary Uses:

- **Marketing profiles** across publishers or brands – e.g. coffee shop fan, frequent traveler
- **Measurement** of ad effectiveness (offline <-> online)
- **Data analysis:**
  - transportation analysis
  - city planning and Smart Cities

Case Studies: Data Creation in Smart Communities Today



What Questions Can Location Data Answer for Transportation Planners?

What types of trips cause **congestion** on a particular roadway?

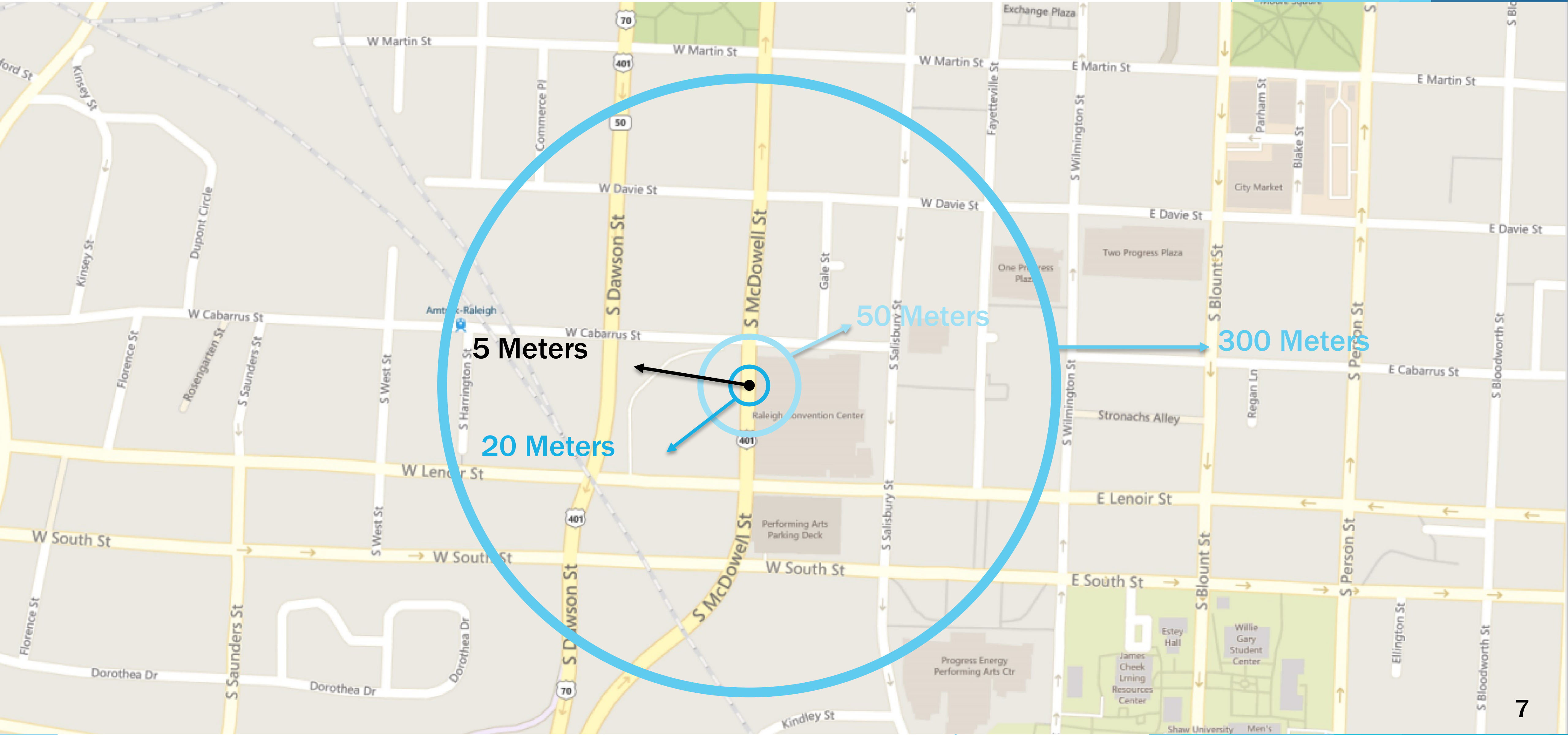
What are the origins and destinations of **travelers** on a particular roadway?

How **do travel patterns vary** during different types and types of day?

What are the **demographic characteristics** of travelers?

Where do **commuters** live, and where do residents work?

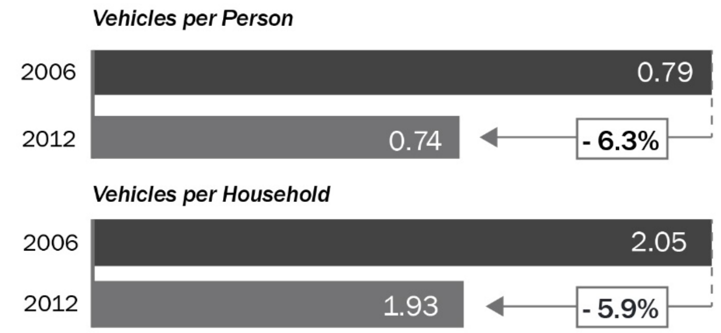
When Selecting Data Sources, Spatial Precision is an Important Factor for Planners



Transportation Behavior Is Changing – But Infrastructure and Budgets Have Not Kept Pace in U.S.

Transportation Behavior is Changing

Vehicle Ownership Trends: 2006 - 2012



Source: Michael Sivak, *Has motorization in the U.S. peaked?* UMTRI, January 2014

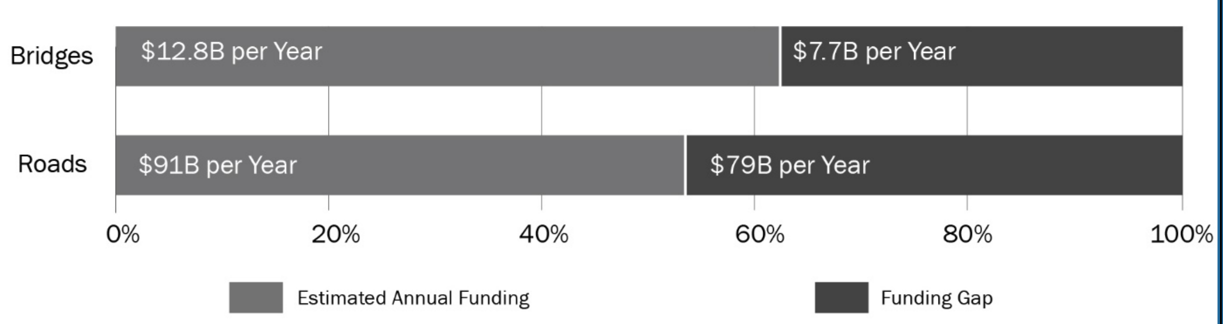
By 2045, It Will Change Even More

- **32%** increase in urban population
- **30%** decrease in rural population
- **Up to 27%** more Vehicle Miles Traveled
- **44%** increase in trucks' freight volume

Source: US DOT, *Beyond Traffic* Final Report, January 2015

Infrastructure Budgets Have Not Kept Up

The Transportation Infrastructure Funding Gap: 2008 – 2028



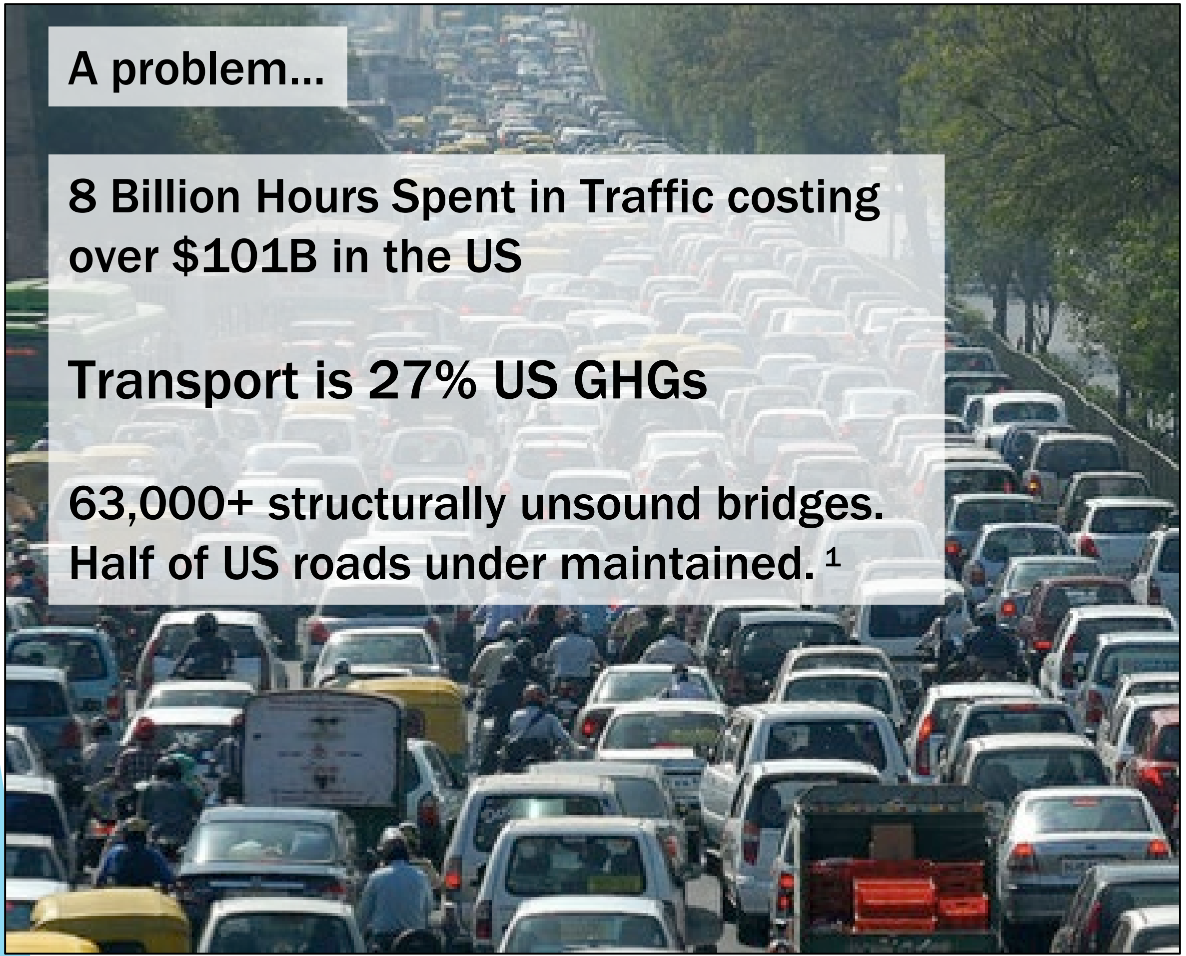
Source: American Society of Civil Engineers, *2013 Report Card for America's Infrastructure*, March 2013

But according to the McKinsey Global Institute, 22% ($400B) per year could be saved globally by using data to optimize expenditures.

Source: McKinsey & Company, *Big Data vs. Congestion: Using Information to Improve Transport*, July 2015

FUTURE OF PRIVACY FORUM VUB BRUSSELS PRIVACY HUB

Transportation is an Expensive, Dangerous Mystery



A problem…

8 Billion Hours Spent in Traffic costing over $101B in the US

Transport is 27% US GHGs

63,000+ structurally unsound bridges. Half of US roads under maintained. [1]

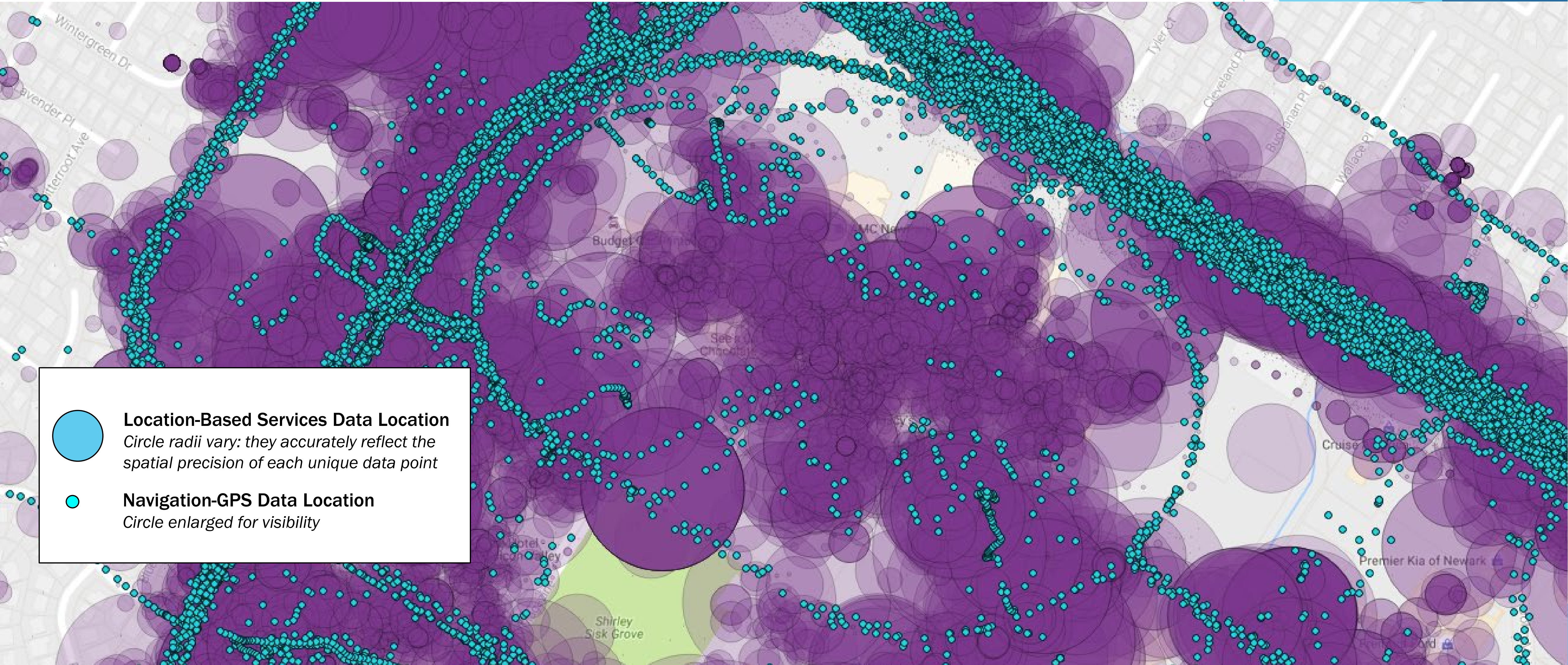
An opportunity

$130B/year US recommended transportation infrastructure spend.[1]

$3T/year global transport infrastructure spend expected.[2]

22% infrastructure expenditures could be saved with data-driven techniques[2]
(and that doesn't include the externalities!)

Example of Mobile Data – Fremont, California



Northern Virginia: Identifying and Prioritizing TDM Projects

Transportation Demand Management

Scanning for Opportunities

Need: Evaluate and prioritize solutions to traffic when highway expansion is not an option due to widespread residential and commercial development

Question to Answer: Where are the highest volume of short trips between O-D pairs that could be converted to other modes?

Challenge: Northern Virginia had to scan hundreds of miles of roads to identify and prioritize the best TDM opportunities, which was not possible to do cost-effectively with conventional data sources

TAZ ID	Avg Trip Duration (sec)	Avg Trip Speed (mph)	Sum under 1 mile	Sum under 3 mile
851	1186	27	5%	30%
850	1433	27	6%	25%
849	1427	30	4%	21%
848	916	23	5%	47%
847	1420	27	9%	39%
846	1275	29	4%	28%
845	1180	23	6%	38%
844	1129	26	7%	37%
843	1504	27	5%	25%
842	1485	30	4%	27%
841	1460	26	7%	31%
840	1403	26	3%	24%
839	1177	25	4%	37%
838	1359	26	6%	34%
837	1272	28	3%	30%
836	1397	28	8%	45%
835	1732	33	6%	36%

City of Lafayette, California: Pinpointing the Cause of Congestion Downtown

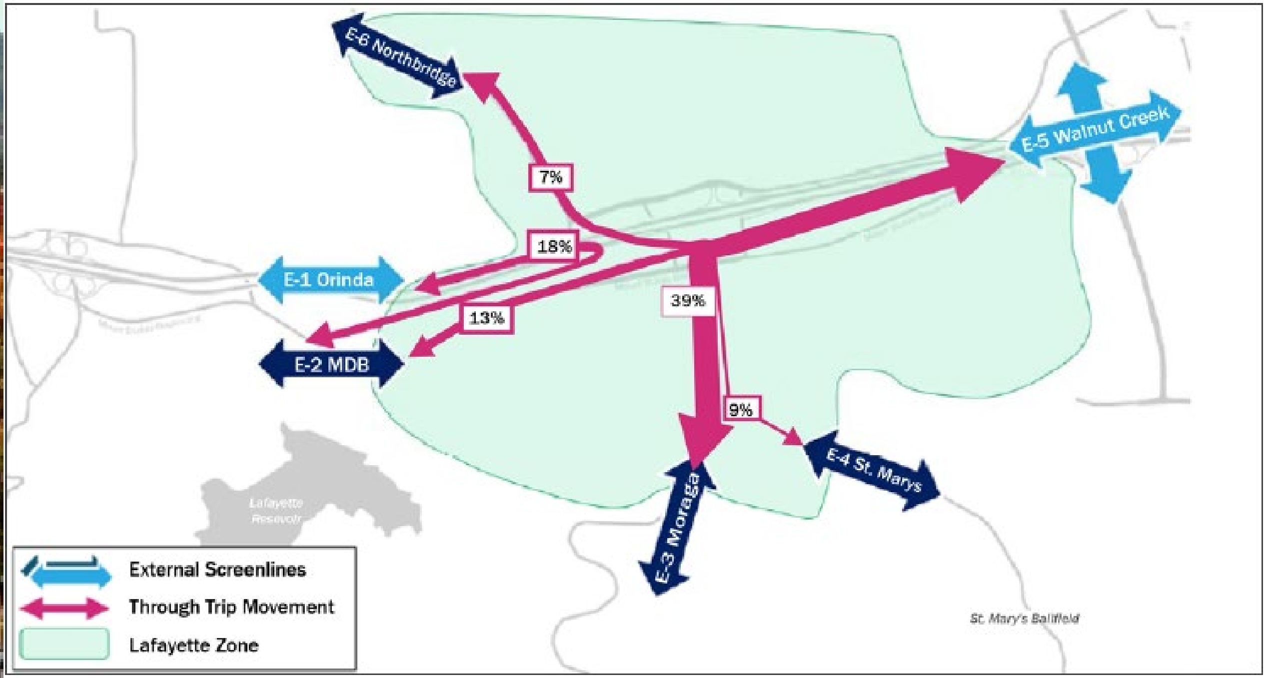
Downtown Congestion Study

O-D for Select Link

Need: Evaluate and prioritize solutions to congestion in downtown corridor

Question to Answer: Understand what which type of trip causes congestion: School drop-offs, commuters to downtown, or "first/last mile" commuters to transit stop

Challenge: Studies were not providing satisfactory answers. The city had counts, but they didn't show origins and destinations, and surveys were inconclusive.



Charlotte, North Carolina: Calibrating a Travel Demand Model

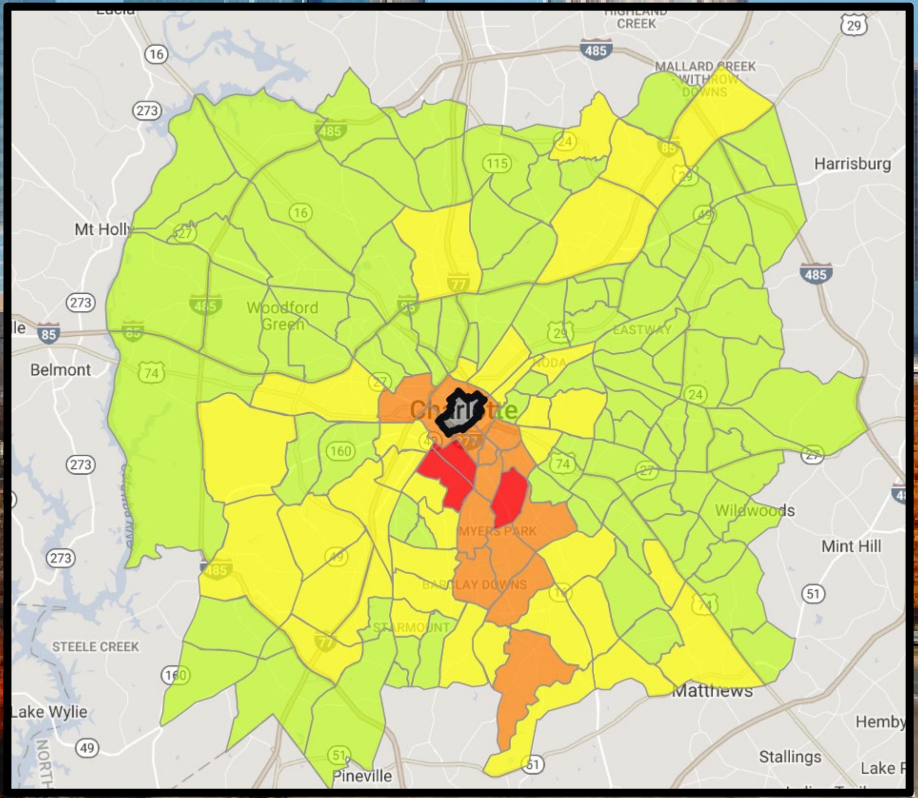
Hypothetical
Transport. Demand Modeling

Origin-Destination for North Carolina MPO

Need: Accurate O/D for calibration or transportation demand model without expensive/time consuming survey for personal and medium/heavy duty commercial trips.

Question: How do travel patterns vary by demographic group and time of day?

Challenge: Planners need to understand how all groups travel, but MPO survey respondents were disproportionately higher income, making it difficult to determine the impact of plans on lower income travelers.



IV. De-Identification: Current Methods

FUTURE OF PRIVACY FORUM VUB BRUSSELS PRIVACY HUB

two or more objects can *not* be
at the *same* place at the *same* time

“Identity” and “identification” according to Wikipedia

- Identity (philosophy), also called sameness, is whatever makes an entity definable and recognizable
- Identity (social science), individuality, personal identity, social identity, and cultural identity in psychology, sociology, and philosophy
- Identity (mathematics), an equality that holds regardless of the values of its variables
- Identification (information), the capability to find, retrieve, report, change, or delete specific data without ambiguity

De-identifying location data adding ambiguity

Various methods exist, such as:

- **Replacing** identifiers with pseudo-identifiers, eg through hashing or lookup tables
- **Stripping identifiers**: (numeric) values that are relatable to individuals
- **Removing sections of data** that combined with other data could allow for identification e.g. begin/end of trip
- **Adding inaccuracy** in time and/or space
- **Aggregating** into “buckets” of time and space

A VISUAL GUIDE TO PRACTICAL DATA DE-IDENTIFICATION

Produced by FUTURE OF PRIVACY FORUM
FPF.org

What do scientists, regulators and lawyers mean when they talk about de-identification? How does anonymous data differ from pseudonymous or de-identified information? Data identifiability is not binary. Data lies on a spectrum with multiple shades of identifiability.

This is a primer on how to distinguish different categories of data.

DEGREES OF IDENTIFIABILITY

Information containing direct and indirect identifiers.

PSEUDONYMOUS DATA

Information from which direct identifiers have been eliminated or transformed, but indirect identifiers remain intact.

DE-IDENTIFIED DATA

Direct and known indirect identifiers have been removed or manipulated to break the linkage to real world identities.

ANONYMOUS DATA

Direct and indirect identifiers have been removed or manipulated together with mathematical and technical guarantees to prevent re-identification.

	EXPLICITLY PERSONAL	POTENTIALLY IDENTIFIABLE	NOT READILY IDENTIFIABLE	KEY CODED	PSEUDONYMOUS	PROTECTED PSEUDONYMOUS	DE-IDENTIFIED	PROTECTED DE-IDENTIFIED	ANONYMOUS	AGGREGATED ANONYMOUS
DIRECT IDENTIFIERS Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SSN)	INTACT	PARTIALLY MASKED	PARTIALLY MASKED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
INDIRECT IDENTIFIERS Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be singled out (e.g., DOB, gender)	INTACT	INTACT	INTACT	INTACT	INTACT	INTACT	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
SAFEGUARDS and CONTROLS Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals	NOT RELEVANT due to nature of data	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	NOT RELEVANT due to nature of data	NOT RELEVANT due to high degree of data aggregation
SELECTED EXAMPLES	Name, address, phone number, SSN, government-issued ID (e.g., Jane Smith, 123 Main Street, 555-555-5555)	Unique device ID, license plate, medical record number, cookie, IP address (e.g., MAC address 68:A8:6D:35:65:03)	Same as Potentially Identifiable except data are also protected by safeguards and controls (e.g., hashed MAC addresses & legal representations)	Clinical or research datasets where only curator retains key (e.g., Jane Smith, diabetes, HgB 15.1 g/dl = Csrk123)	Unique, artificial pseudonyms replace direct identifiers (e.g., HIPAA Limited Datasets, John Doe = 5L7T LX619Z) (unique sequence not used anywhere else)	Same as Pseudonymous, except data are also protected by safeguards and controls	Data are suppressed, generalized, perturbed, swapped, etc. (e.g., GPA: 3.2 = 3.0-3.5, gender: female = gender: male)	Same as De-Identified, except data are also protected by safeguards and controls	For example, noise is calibrated to a data set to hide whether an individual is present or not (differential privacy)	Very highly aggregated data (e.g., statistical data, census data, or population data that 52.6% of Washington, DC residents are women)

FUTURE OF PRIVACY FORUM · VUB · BRUSSELS PRIVACY HUB

Can location data be anonymous?

Yes. But it is very hard to achieve.

Taking an ongoing risk based approach is key.

Technical, organization and contractual measures can provide a “tripod” of assurance.

FUTURE OF PRIVACY FORUM

VUB

BRUSSELS PRIVACY HUB

Questions?

FUTURE OF PRIVACY FORUM VUB BRUSSELS PRIVACY HUB