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

Class 2 of Digital Data Flows Masterclass: *Emerging Technologies*

27 November, 2018 | Brussels
# DIGITAL DATA FLOWS MASTERCLASS:
EMERGING TECHNOLOGIES

<table>
<thead>
<tr>
<th>Curriculum</th>
<th>Date*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Session 1:</strong> 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</td>
<td>25 October, 2018 - side event, ICDPPC (Brussels) (with remote participation)</td>
</tr>
<tr>
<td><strong>Session 2:</strong> Location Data: GPS, Wi-Fi, and Spatial Analytics</td>
<td>November 2018 - Brussels (with remote participation)</td>
</tr>
<tr>
<td><strong>Session 3:</strong> Advertising Technologies: Online Data Flows, Behavioral Targeting, and Cross-Device Tracking</td>
<td>January 2019 - Brussels (with remote participation)</td>
</tr>
<tr>
<td><strong>Session 4:</strong> Mobile Apps: Operating Systems, Software Development Kits (SDKs), and User Controls</td>
<td>March 2019 - Virtual</td>
</tr>
<tr>
<td><strong>Session 5:</strong> Transportation and Mobility: Video Analytics, Sensors, and Connected Infrastructure</td>
<td>April 2019 - Virtual</td>
</tr>
<tr>
<td><strong>Session 6:</strong> Biometric Data: Facial Recognition, Voice, and Digital Fingerprints</td>
<td>June 2019 - Virtual</td>
</tr>
<tr>
<td><strong>Session 7:</strong> Tracking in Physical Spaces: Retail Technologies, Smart Homes, and the “Internet of Things”</td>
<td>July 2019 - Virtual</td>
</tr>
<tr>
<td><strong>Session 8:</strong> De-Identification: Multi-party Computing, Differential Privacy, and Homomorphic Encryption</td>
<td>Sept. 2019 - Virtual</td>
</tr>
</tbody>
</table>

*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](http://www.fpf.org/classes)
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
I. Introduction
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

<table>
<thead>
<tr>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>52.50333486</td>
<td>13.33955726</td>
</tr>
<tr>
<td>52.50333535</td>
<td>13.33955777</td>
</tr>
<tr>
<td>52.50333633</td>
<td>13.33955888</td>
</tr>
<tr>
<td>52.5033732</td>
<td>13.33955984</td>
</tr>
<tr>
<td>52.5033383</td>
<td>13.33956087</td>
</tr>
<tr>
<td>52.5033929</td>
<td>13.3395619</td>
</tr>
<tr>
<td>52.5033982</td>
<td>13.33956245</td>
</tr>
<tr>
<td>52.5034027</td>
<td>13.33956293</td>
</tr>
<tr>
<td>52.5034126</td>
<td>13.33956397</td>
</tr>
<tr>
<td>52.5034225</td>
<td>13.3395656</td>
</tr>
<tr>
<td>52.5034323</td>
<td>13.33956603</td>
</tr>
</tbody>
</table>
II. Sources of Data
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

- 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

iPhone 7
source: ifixit.com, Creative Commons
Mobile sensors generate input for OS to provide standardized latitude-longitude

Accelerometer
Bluetooth
Cell Towers
Compass
GPS
Gyroscope
Wi-Fi Networks

“Location Services”

back to OS to improve services

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

ID=NULL
A closer look at…

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

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

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

Challenges:
- 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.

<table>
<thead>
<tr>
<th>Cell Tower Database</th>
<th>Unique Cell Towers</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenCellID</td>
<td>&gt; 6 million</td>
<td>Public</td>
</tr>
<tr>
<td>Combain</td>
<td>&gt; 72 million</td>
<td>Private</td>
</tr>
<tr>
<td>LocationAPI.org</td>
<td>&gt; 72 million</td>
<td>Private</td>
</tr>
<tr>
<td>Mozilla</td>
<td>&gt; 26 million</td>
<td>Public</td>
</tr>
<tr>
<td>Navizon</td>
<td>&gt; 71 million</td>
<td>Private</td>
</tr>
<tr>
<td>Mylnikov GEO</td>
<td>&gt; 15 million</td>
<td>Public</td>
</tr>
<tr>
<td>WiGLE</td>
<td>&gt; 6 million</td>
<td>Private</td>
</tr>
</tbody>
</table>
3. Nearby Wi-Fi Networks

Source: https://wigle.net/
3. Nearby Wi-Fi Networks

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

Source: https://wigle.net/
4. Bluetooth (Beacons)

Beacons are inexpensive radio transmitters that send one-way signals to devices equipped to receive them.
5. Proximity Alternatives – LED, Audio

Indoor Location-Based Services Using LED Lighting

How it Works

1. ByteLight-enabled GE LED fixtures “communicate” a unique light pattern using Visible Light Communication and Bluetooth Low Energy

2. Connected shoppers opt-in to “listen” with retailer’s app on any smartphone and tablet with a camera and/or Bluetooth Smart

3. Camera detects unique light pattern and Bluetooth signal emitted by GE Lumanation™ LED Luminaires; application notifies ByteLight platform of shopper’s position and direction with sub-meter accuracy

4. Platform ties to retailer’s digital marketing systems to deliver location-based services and personalized content to each shopper

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
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**

<table>
<thead>
<tr>
<th></th>
<th>2006</th>
<th>2012</th>
<th>Change</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vehicles per Person</strong></td>
<td>0.79</td>
<td>0.74</td>
<td>-6.3%</td>
</tr>
<tr>
<td><strong>Vehicles per Household</strong></td>
<td>2.05</td>
<td>1.93</td>
<td>-5.9%</td>
</tr>
</tbody>
</table>

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**

- **Bridges**: $12.8B per Year vs. $7.7B per Year
- **Roads**: $918B per Year vs. $79B per Year

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.

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 techniques² (and that doesn’t include the externalities!)
Location-Based Services Data Location
Circle radii vary: they accurately reflect the spatial precision of each unique data point

Navigation-GPS Data Location
Circle enlarged for visibility
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.

<table>
<thead>
<tr>
<th>TAZ ID</th>
<th>Avg Trip Duration (sec)</th>
<th>Avg Trip Speed (mph)</th>
<th>Sum under 1 mile</th>
<th>Sum under 3 mile</th>
</tr>
</thead>
<tbody>
<tr>
<td>851</td>
<td>1186</td>
<td>27</td>
<td>5%</td>
<td>30%</td>
</tr>
<tr>
<td>850</td>
<td>1433</td>
<td>27</td>
<td>6%</td>
<td>25%</td>
</tr>
<tr>
<td>849</td>
<td>1427</td>
<td>30</td>
<td>4%</td>
<td>21%</td>
</tr>
<tr>
<td>848</td>
<td>916</td>
<td>23</td>
<td>5%</td>
<td>47%</td>
</tr>
<tr>
<td>847</td>
<td>1420</td>
<td>27</td>
<td>9%</td>
<td>39%</td>
</tr>
<tr>
<td>846</td>
<td>1275</td>
<td>29</td>
<td>4%</td>
<td>28%</td>
</tr>
<tr>
<td>845</td>
<td>1180</td>
<td>23</td>
<td>6%</td>
<td>38%</td>
</tr>
<tr>
<td>844</td>
<td>1129</td>
<td>26</td>
<td>7%</td>
<td>37%</td>
</tr>
<tr>
<td>843</td>
<td>1504</td>
<td>27</td>
<td>5%</td>
<td>25%</td>
</tr>
<tr>
<td>842</td>
<td>1485</td>
<td>30</td>
<td>4%</td>
<td>27%</td>
</tr>
<tr>
<td>841</td>
<td>1460</td>
<td>26</td>
<td>7%</td>
<td>31%</td>
</tr>
<tr>
<td>840</td>
<td>1403</td>
<td>26</td>
<td>3%</td>
<td>24%</td>
</tr>
<tr>
<td>839</td>
<td>1177</td>
<td>25</td>
<td>4%</td>
<td>37%</td>
</tr>
<tr>
<td>838</td>
<td>1359</td>
<td>26</td>
<td>6%</td>
<td>34%</td>
</tr>
<tr>
<td>837</td>
<td>1272</td>
<td>28</td>
<td>3%</td>
<td>30%</td>
</tr>
<tr>
<td>836</td>
<td>1397</td>
<td>28</td>
<td>8%</td>
<td>45%</td>
</tr>
<tr>
<td>835</td>
<td>1732</td>
<td>33</td>
<td>6%</td>
<td>36%</td>
</tr>
</tbody>
</table>
City of Lafayette, California: Pinpointing the Cause of Congestion Downtown

**Downtown Congestion Study**

**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
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

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.

<table>
<thead>
<tr>
<th>DEGREES OF IDENTIFIABILITY</th>
<th>INFORMATION CONTAINING DIRECT AND INDIRECT IDENTIFIERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicitly Personal</td>
<td>Directly identifiable data contains personal identifiers</td>
</tr>
<tr>
<td>Potentially Identifiable</td>
<td>Data can be linked to personal identifiers if sufficiently linked</td>
</tr>
<tr>
<td>Not Readily Identifiable</td>
<td>Data cannot be linked to personal identifiers without additional information</td>
</tr>
</tbody>
</table>

**PSEUDONYMOUS DATA**

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

**DIRECT IDENTIFIERS**

Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SSN).

**INDIRECT IDENTIFIERS**

Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be identified (e.g., DOB, gender).

**SAFE GUARDS AND CONTROLS**

Technical, organizational, and legal controls preventing employees, researchers or other third parties from re-identifying individuals.

**SELECTED EXAMPLES**

- Name, address, phone number, SSN, government-issued ID (e.g., Jane Smith, 123 Main Street, 015-555-3313)
- Unique device ID, license plate, medical record number, cookie, IP address (e.g., MVC address 8646 6603 66 96 0306)
- Clinical or research datasets where only casual research keys exist (e.g., pseudo-IMPA identifiers and legal representations)
- Unique, artificial pseudonyms replace direct identifiers (e.g., HINSA Limited Dataset, Jane Doe = 15.77.436.972) Use these keywords to search the database.

**ANONYMOUS DATA**

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

**FUTURE OF PRIVACY FORUM**

Produced by Future of Privacy Forum

**BRUSSELS PRIVACY HUB**

Compiled by 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.
Questions?