Data Ethics

Uijeong “UJ” Hwang

What is Data Ethics?

*"In the simplest terms, **data** refer to factual information, such as measurements or statistics, used as a basis for reasoning, discussion, or calculation. **Data Ethics** are the norms of behavior that promote appropriate judgments and accountability when acquiring, managing, or using data, with the goals of protecting civil liberties, minimizing risks to individuals and society, and maximizing the public good."*

(Federal Data Strategy, 2020)

https://resources.data.gov/assets/documents/fds-data-ethics-framework.pdf
https://imstat.org/2018/12/14/hand-writing-right-legitimate-and-proper-the-new-world-of-data-ethics/

What is Data Ethics?

*"In the simplest terms, **data** refer to factual information, such as measurements or statistics, used as a basis for reasoning, discussion, or calculation. **Data Ethics** are the norms of behavior that promote appropriate judgments and accountability when acquiring, managing, or using data, with the goals of protecting civil liberties, minimizing risks to individuals and society, and maximizing the public good."*

(Federal Data Strategy, 2020)

*"…more recently, recognition of the need for such ethical oversight has grown, mainly because of raised awareness of the **potential and pervasiveness of big data, data science, and artificial intelligence**. Attention has shifted, from rather specialised concerns for informed consent in clinical trials, the preservation of anonymity in survey work, avoiding prohibited variables in insurance decisions, and so on, to much more "in-your-face" issues. These are matters such as selection bias leading to racist decisions, chatbots being gratuitously offensive, and questions of who is responsible when a driverless car crashes or data theft leads to fraud."*

(Hand, 2018)

https://resources.data.gov/assets/documents/fds-data-ethics-framework.pdf
https://imstat.org/2018/12/14/hand-writing-right-legitimate-and-proper-the-new-world-of-data-ethics/

Frameworks and Guidelines

- Data Ethics Framework (Federal Data Strategy)
  - Data Ethics Tenets (page 4): https://resources.data.gov/assets/documents/fds-data-ethics-framework.pdf
- General Data Protection Regulation (EU)
  - Article 5. Principles relating to processing of personal data: https://gdpr-info.eu/art-5-gdpr/
- GIS Code of Ethics (GIS Certification Institute)
  - Pages 2-4: https://www.gisci.org/Ethics/Code-of-Ethics

Key Ethical Principles

- Privacy and Informed Consent
  - Purpose Limitation
  - Storage limitation
- Security and Confidentiality
- Transparency
- Data Quality and Integrity
- Fairness and Non-discrimination
- Accountability
- Beneficence

Key Ethical Principles

- Privacy and Informed Consent
  - Purpose Limitation
  - Storage limitation
- Security and Confidentiality
- Transparency
- Data Quality and Integrity
- Fairness and Non-discrimination
- Accountability
- Beneficence

Individuals must control their personal data, which can only be collected and used with explicit, informed consent.

Data should only be used for the specific, agreed-upon purposes.

Data must be retained only as long as necessary for its purpose and then securely deleted or anonymized.

Key Ethical Principles

- Privacy and Informed Consent
  - Purpose Limitation
  - Storage limitation
- Security and Confidentiality
- Transparency
- Data Quality and Integrity
- Fairness and Non-discrimination
- Accountability
- Beneficence

Ensure data is protected from unauthorized access, breaches, and misuse, while sensitive information is shared only with authorized entities.

Clearly communicate how data is collected, used, stored, and shared to foster trust.

Key Ethical Principles

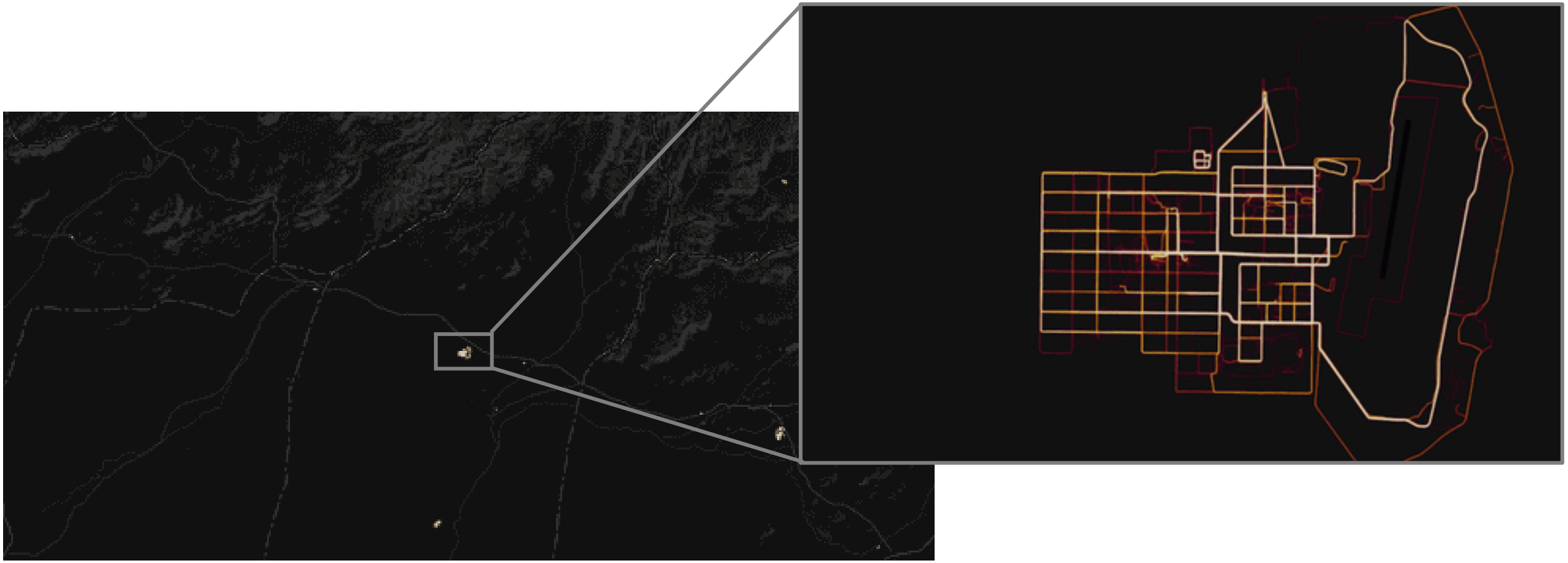
- Privacy and Informed Consent
  - Purpose Limitation
  - Storage limitation
- Security and Confidentiality
- Transparency
- Data Quality and Integrity — Maintain accurate, complete, and reliable data for ethical and effective use.
- Fairness and Non-discrimination — Data practices must avoid bias and ensure equitable treatment of all individuals or groups.
- Accountability
- Beneficence

Key Ethical Principles

- Privacy and Informed Consent
  - Purpose Limitation
  - Storage limitation
- Security and Confidentiality
- Transparency
- Data Quality and Integrity
- Fairness and Non-discrimination
- Accountability — Organizations must take responsibility for complying with ethical principles and demonstrate adherence.
- Beneficence — Ensure data practices maximize benefits while minimizing potential harm to individuals and society.

Real-World Examples

Strava Heatmap Scandal



https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases

Target Predicts Pregnancy of Customer

https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

Apple Card Gender Bias Scandal

- Apple's credit card algorithm allegedly offered significantly lower credit limits to women, even with similar financial backgrounds to men.

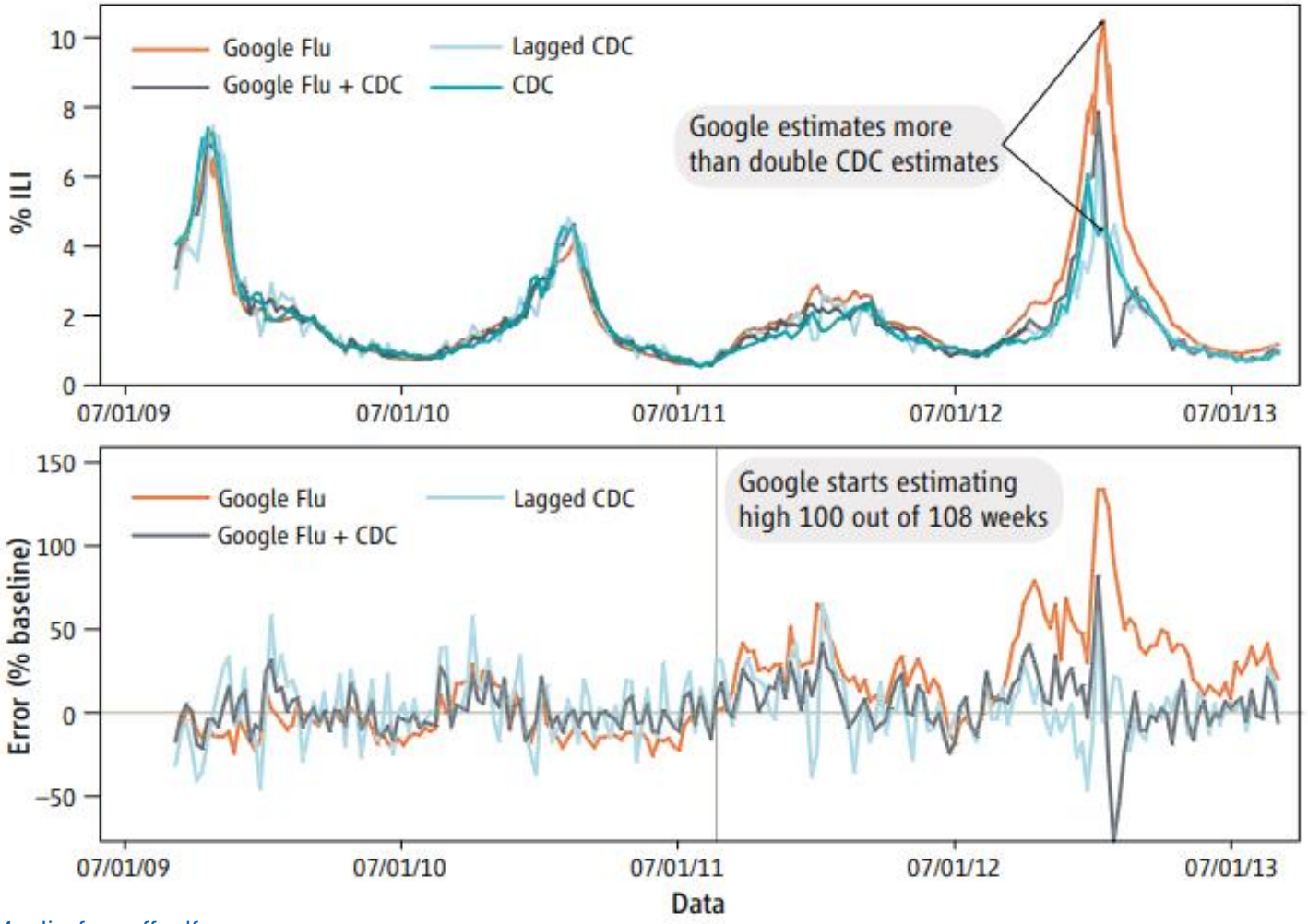
https://www.youtube.com/watch?v=rD0hpPk92LY&ab_channel=CBSMornings

Facebook-Cambridge Analytica Data Scandal

https://youtu.be/mrnXv-g4yKU?si=c2I9ksbwMzvcnO_O

Failure of Google Flu Trends

- Google Flu Trends overestimated flu outbreaks due to reliance on search data that reflected panic rather than actual cases.



https://gking.harvard.edu/files/gking/files/0314policyforumff.pdf

NYC Taxi Dataset Controversy

Poorly anonymized logs reveal NYC cab drivers' detailed whereabouts

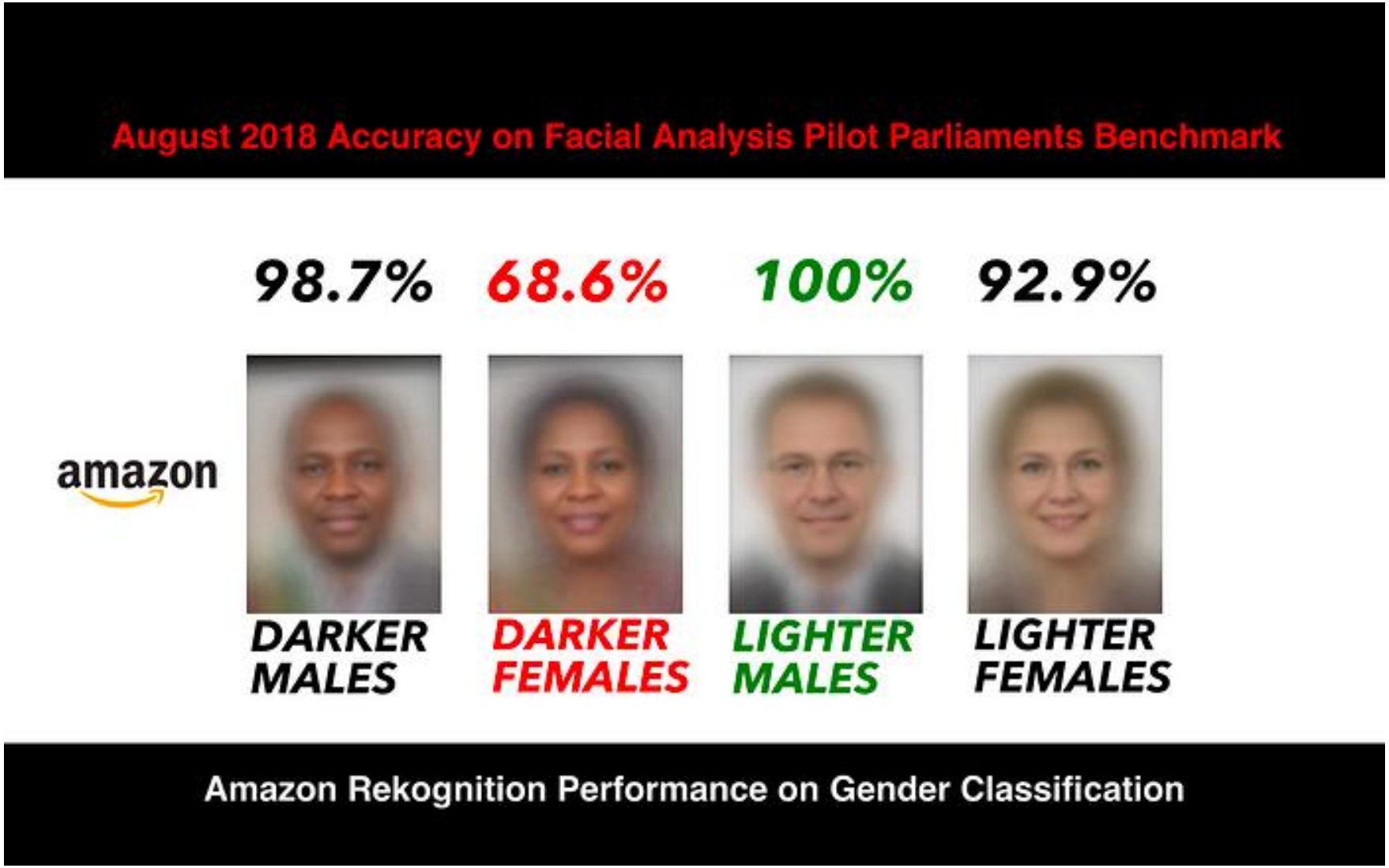
Botched attempt to scrub data reveals driver details for 173 million taxi trips.

DAN GOODIN – JUN 23, 2014 1:25 PM | 31

Credit: David R. Tribble

https://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn

Biases in Big Data and AI



https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced
https://youtu.be/PWCtoVt1CJM?si=dJkhaAZrncN0jLkh

Personally Identifiable Information

Personally Identifiable Information (PII)

According to US Department of Labor,

- *PII is defined as any representation of information that permits the identity of an individual to whom the information applies to be reasonably inferred by either direct or indirect means.*

https://www.dol.gov/general/ppii

Personally Identifiable Information (PII)

According to US Department of Labor,

- *PII is defined as any representation of information that permits the identity of an individual to whom the information applies to be reasonably inferred by either direct or indirect means.*
- *PII is information:*
  - *(i) that **directly identifies** an individual (e.g., name, address, social security number or other identifying number or code, telephone number, email address, etc.) or*
  - *(ii) by which an agency intends to identify specific individuals in conjunction with other data elements, i.e., **indirect identification**. (These data elements may include a combination of gender, race, birth date, geographic indicator, and other descriptors).*
  - *Additionally, information permitting the physical or online contacting of a specific individual is the same as personally identifiable information. This information can be maintained in either paper, electronic or other media.*

https://www.dol.gov/general/ppii

PII in Household Travel Survey Data

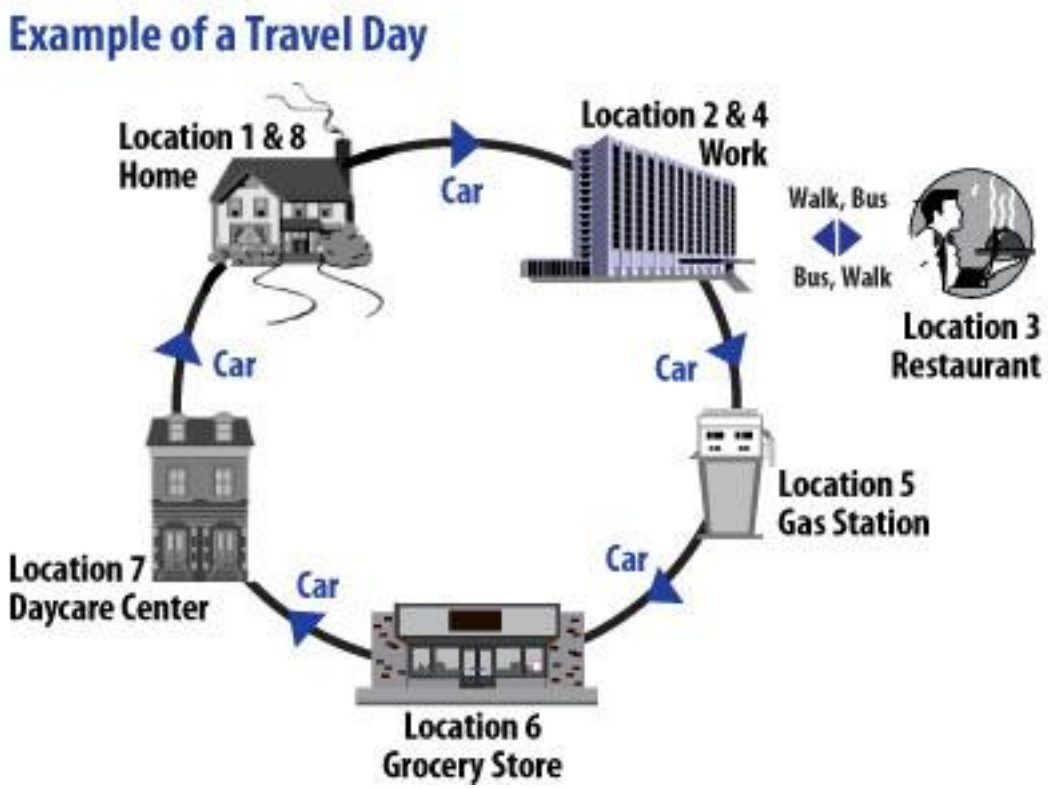
TABLE 3: SELECT/EXAMPLE DATA ELEMENTS BY DATA TABLE

DATA LEVEL		SELECT/EXAMPLE DATA ELEMENTS	
Demographics/Household Composition	Household	• Carshare program use • Home location (primary & previous) • Home ownership • Household income • Household size	• Household vehicle count • Reasons for home relocation • Residence type and duration • Seasonal residency
	Vehicle	• Disability pass status • Fuel type	• Make, model, and model year • Purchase year
	Person	• Age category • Autonomous vehicle use (concerns, interest) • Commute benefits • Commute mode (typical) • Commute frequency (typical) • Driver's license type • Educational attainment • Employment status • Factors to influence more bike travel • Factors to influence more transit travel	• Gender • Job count • Job location (current & previous) • Parking at work • Race/ethnicity • Relationship to primary householder • School type, location, travel mode • Student status • Transit payment methods • Vehicle driven most often • Walk/bike/transit/ride hailing frequency
Travel/Location Information	Day	• Diary reporter (i.e., self-reported or assisted) • Date/day of week • Day trip count • Home deliveries on travel day	• If used toll roads on travel day • If paid for parking on travel day • Online shop time on travel day • Reasons for not traveling (if applicable) • Telecommute time on travel day
	Trip	• Carpool start and end location • Date/day of week • Parking costs • Parking location type • Persons in travel party • Taxi/bus/ferry/air fares	• Toll road fares • Transit lines used • Travel modes • Travel times/trip durations • Trip purpose • Vehicle driver

https://www.psrc.org/media/3248

PII in Household Travel Survey Data

- Travel data (especially geolocation) can reveal sensitive personal information, such as home address, work location, social activities, and routines.
- Even anonymized data can sometimes be re-identified through sophisticated data matching techniques.



https://metrocouncil.org/Transportation/Performance/Travel-Behavior-Inventory/Data/2022-Household-Survey-How-We-Move/2010-Travel-Behavior-Inventory-(TBI)-Survey-Data-R.aspx

Discussions

Scenario 1: Data Sharing for Disaster Response

You are a data scientist working for a government agency coordinating disaster response after a major event (e.g., hurricane). Your agency collects various types of real-time data—satellite images, social media feeds, smartphone data, and IoT sensor data—to assist in emergency operations. Your agency plans to share this data with external partners, such as humanitarian organizations, local governments, and emergency responders, to improve relief efforts. However, the data might expose individuals' private information, which raises concerns about privacy, consent, and ethical use.

- How can we protect privacy while keeping the data useful?
- What steps can ensure transparency and prevent data misuse by external partners?

The 2011 Japan Earthquake Case

- Following the 2011 Great East Japan Earthquake, the Japanese government allowed the University of Tokyo Disaster Information Center to conduct research and surveys involving survivors of the earthquake under strict protocols to minimize privacy risks and ensure ethical research practices.

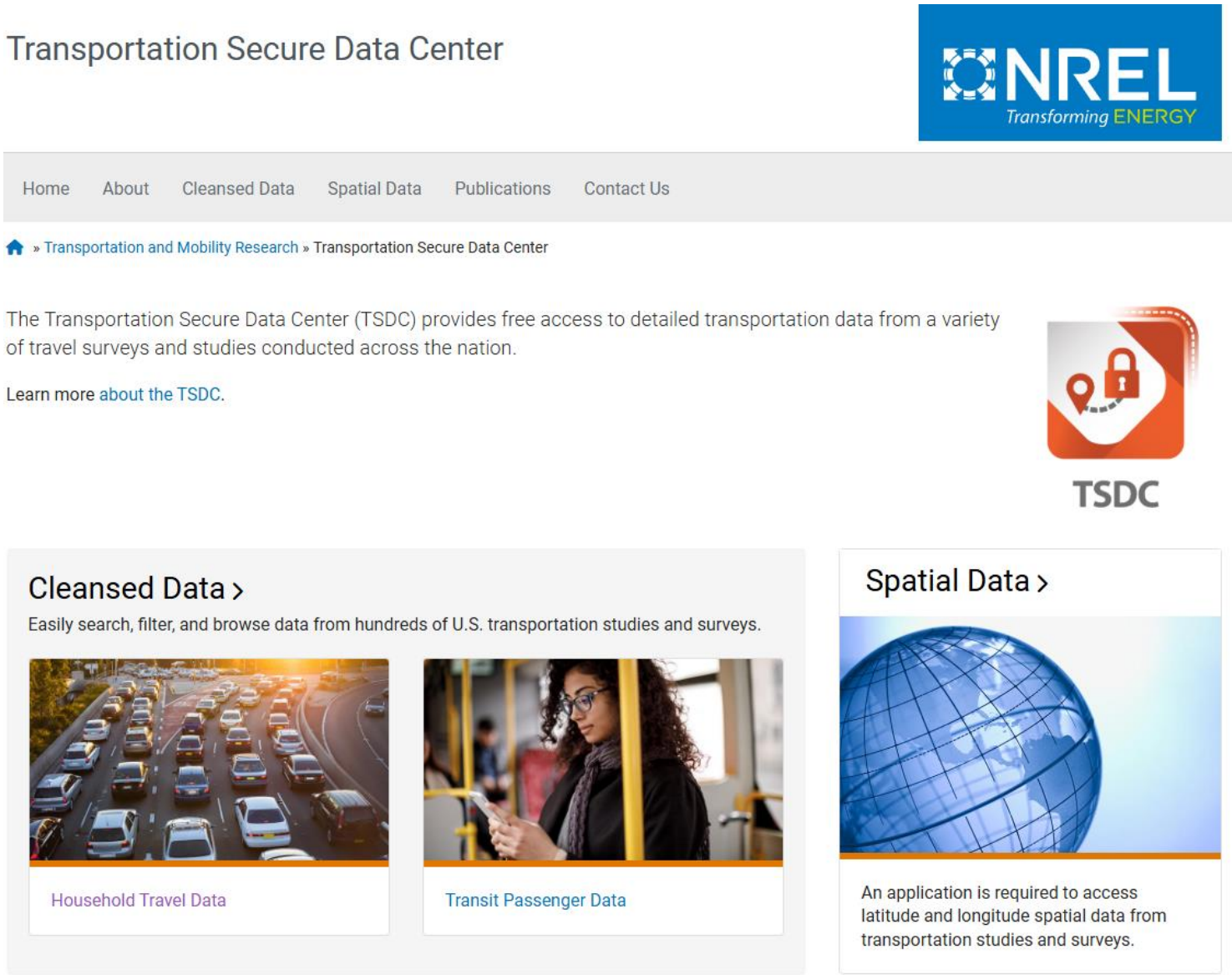
https://www.britannica.com/event/Japan-earthquake-and-tsunami-of-2011/Relief-and-rebuilding-efforts

Scenario 2: Publishing Household Travel Survey Data

*You are a city planner in charge of conducting a **household travel survey** to collect data on transportation patterns in your city. After gathering valuable insights, you want to publish the data publicly to support urban planning and policy-making. However, the data includes sensitive information, such as travel routes, GPS coordinates, and sociodemographic details, which could pose privacy risks if exposed.*

- What types of data could lead to ethical issues if made public?
- What strategies would you use to secure the data?
- How would you balance privacy protection with other values?
    - Such as transparency, data democracy, innovation, and social progress

Transportation Secure Data Center

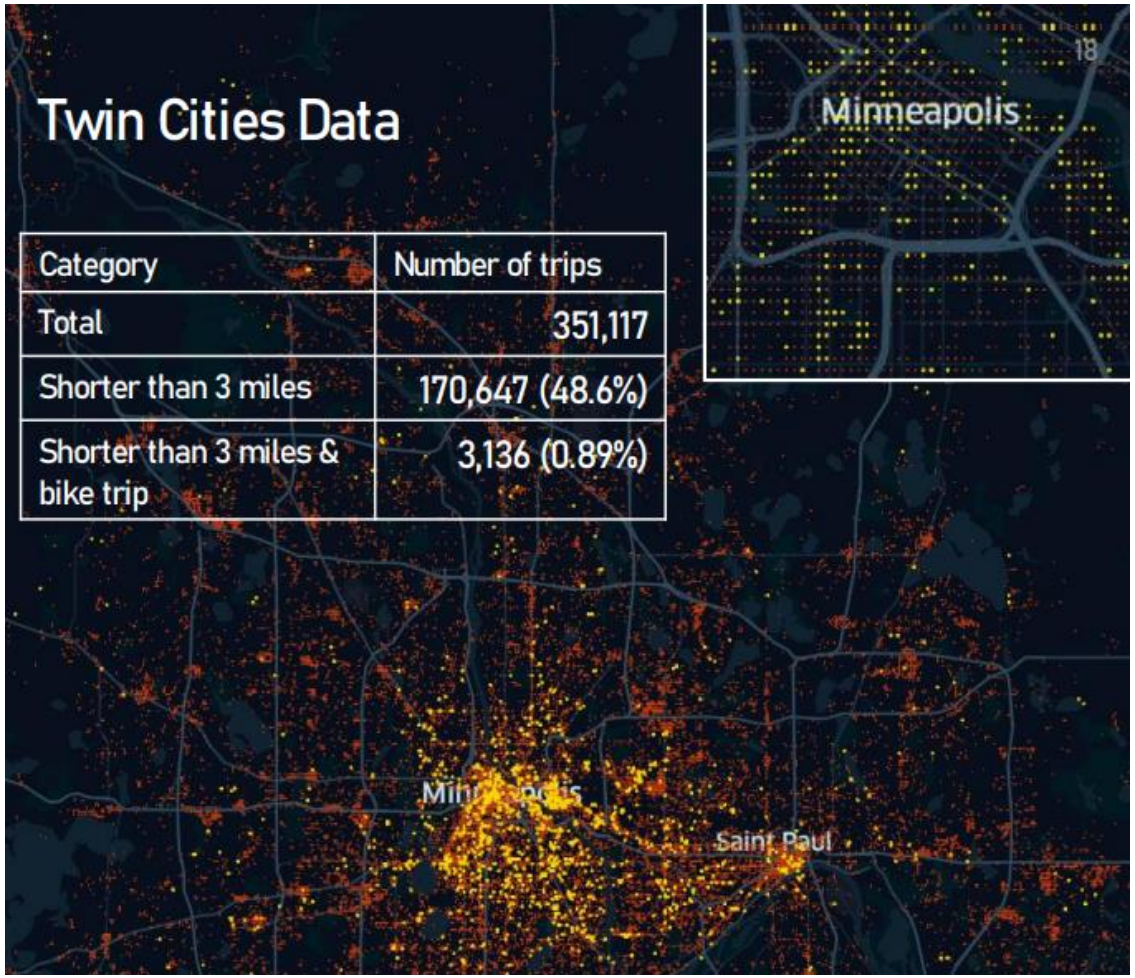


https://www.nrel.gov/transportation/secure-transportation-data/

Handling Geolocation Data Ethically

- Aggregation
- Respect for sensitive locations
- Adjusting spatial/temporal resolution
- Adding noise (perturbation)

A Key Consideration:
A delicate balance between leveraging its value for social well-being while protecting individuals' privacy.



Category	Number of trips
Total	351,117
Shorter than 3 miles	170,647 (48.6%)
Shorter than 3 miles & bike trip	3,136 (0.89%)

Best Practices

- Conduct an ethical review before starting projects.
- Consider the broader impact and unintended consequences.
- Ensure participants know exactly what data is being collected and why.
- Anonymize and aggregate data.
- Communicate data practices transparently with stakeholders.