# Executive Summaries of Selected Projects

## Project 11

## Cybersecurity – Detect Command and Control Channel in DNS Logs

Project image placeholder

**BLUF:** Designed a real-time, streaming analytics system to detect covert malware **command-and-control (C&C)** communications in DNS traffic and alert defenders immediately.

### Introduction

* Monitoring DNS logs can uncover distributed malicious communications hidden as ordinary queries[[1]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=Clear%20Aim%20Monitor%20DNS%20logs,they%20can%20take%20required%20actions)
* Attackers use DNS as a covert C&C channel, making threats difficult to spot without specialized analysis[[2]](file://file_000000002ee861f586ba20612cc9ac3b#:~:text=,Detects%20anomalies%20for%20codeword)
* Massive, continuous DNS data streams overwhelm manual monitoring, necessitating automated detection[[3]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=Target%20Application%20Design%20a%20dashboard,control%20channel%20in%20DNS%20logs)
* Undetected DNS-based **C&C** traffic can persist and escalate into breaches if not caught early (drives need for this solution)[[4]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=Problem%20Statement%20Detect%20command%20and,control%20channel%20in%20DNS%20logs)

### Key Questions Addressed

* How to ingest and process high-volume DNS logs continuously for timely threat analysis?[[5]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=1,DNS%20logs%2C%20metadata%20for%20enrichment)
* How to detect both fast and stealthy (low-and-slow) malicious DNS communication patterns in real time?[[6]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=2.%20Real,cleaning%2C%20and%20enrichment%20of%20DNS)[[7]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=3.%20Large)
* How to leverage threat intelligence feeds and historical data to improve detection accuracy?[[8]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=o%20Combine%20historical%20data%2C%20community,identify%20slow%2C%20stealthy%20malicious%20behavior)
* How to promptly alert stakeholders and visualize DNS threats as they emerge?[[9]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=5,for)

### The Problem

* Malware can abuse DNS queries as a covert C&C channel, allowing attackers to bypass standard defenses[[2]](file://file_000000002ee861f586ba20612cc9ac3b#:~:text=,Detects%20anomalies%20for%20codeword)
* Organizations face a flood of DNS log data, and manual analysis cannot keep up[[3]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=Target%20Application%20Design%20a%20dashboard,control%20channel%20in%20DNS%20logs)
* Slow, stealthy DNS tunnels blend into normal traffic, evading simplistic or batch detection methods[[7]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=3.%20Large)
* Without quick detection, malicious domains remain active, posing ongoing risks to data and systems[[4]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=Problem%20Statement%20Detect%20command%20and,control%20channel%20in%20DNS%20logs)

### The Importance

* Early detection of DNS-based threats enables faster incident response, cutting off attacks before damage occurs[[4]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=Problem%20Statement%20Detect%20command%20and,control%20channel%20in%20DNS%20logs)
* Adds a critical security layer focused on an often-overlooked attack vector (DNS traffic) for comprehensive defense[[1]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=Clear%20Aim%20Monitor%20DNS%20logs,they%20can%20take%20required%20actions)
* Real-time alerts empower security teams to act immediately, reducing attacker dwell time in networks[[10]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=Monitor%20DNS%20logs%20round,the%20stakeholders%20in%20real%20time)
* Shares newly discovered malicious indicators with the community, strengthening collective cyber defenses[[11]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=o%20Ensure%20only%20authorized%20users,Contribution%20Back%20to%20Community)

### The Solution

* Built a microservices-based pipeline to ingest, enrich (GeoIP, WHOIS), and analyze DNS logs in real time[[5]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=1,DNS%20logs%2C%20metadata%20for%20enrichment)[[12]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=We%20clustered%20requirements%20into%20microservices%3B,Ingestion%20%26%20Preprocessing%20Service)
* Short-window anomaly detection flags immediate threats, while long-window analytics finds stealthy, slow attacks[[6]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=2.%20Real,cleaning%2C%20and%20enrichment%20of%20DNS)[[13]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=o%20Detects%20anomalies%20in%20short,confidence%20events)
* Integrated threat intelligence feeds (blacklists/whitelists) to continuously update known malicious domains[[14]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=4)[[15]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=2,append%20source%20and%20threat%20hierarchies)
* Provided an alerting module and central dashboard for real-time visualization of suspicious domains and events[[9]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=5,for)[[16]](file://file_000000002ee861f586ba20612cc9ac3b#:~:text=,based%20access%20%26%20query%20interfaces)
* Shared confirmed Indicators of Compromise via a knowledge base, contributing new domain intel back to the community[[11]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=o%20Ensure%20only%20authorized%20users,Contribution%20Back%20to%20Community)[[17]](file://file_000000002ee861f586ba20612cc9ac3b#:~:text=)

### Architecture Overview

* **Modular microservices design:** Independent services for ingestion, threat intel aggregation, detection, analytics, alerting, and UI[[12]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=We%20clustered%20requirements%20into%20microservices%3B,Ingestion%20%26%20Preprocessing%20Service)[[18]](file://file_000000002ee861f586ba20612cc9ac3b#:~:text=Overview)
* **Streaming data pipeline:** A message broker (e.g. Kafka) glues services together, ensuring scalable, real-time data flow[[19]](file://file_000000002ee861f586ba20612cc9ac3b#:~:text=Kafka%20)
* **Multi-tier storage:** In-memory fast tier for real-time processing and a bulk storage tier for historical analysis[[20]](file://file_000000002ee861f586ba20612cc9ac3b#:~:text=In)[[21]](file://file_000000002ee861f586ba20612cc9ac3b#:~:text=Storage%20)
* **Tiered deployment:** Services grouped by tier (data, application, storage) for optimized performance and scalability[[21]](file://file_000000002ee861f586ba20612cc9ac3b#:~:text=Storage%20)[[22]](file://file_000000002ee861f586ba20612cc9ac3b#:~:text=With%20this%20microservice,ingest%20continuous%20DNS%20logs%2C%20run)
* **Flexibility & resilience:** The microservice architecture allows independent scaling, easy updates, and robust fault isolation[[22]](file://file_000000002ee861f586ba20612cc9ac3b#:~:text=With%20this%20microservice,ingest%20continuous%20DNS%20logs%2C%20run)

### Results and Impacts

* Continuous DNS monitoring with anomaly detection can yield near-immediate alerts for emerging threats[[22]](file://file_000000002ee861f586ba20612cc9ac3b#:~:text=With%20this%20microservice,ingest%20continuous%20DNS%20logs%2C%20run)
* Combining historical analysis and dynamic threat intel improves detection accuracy, catching both abrupt and subtle attacks[[23]](file://file_000000002ee861f586ba20612cc9ac3b#:~:text=system%20can%20ingest%20continuous%20DNS,logs%2C%20run)
* Actionable alerts and a live dashboard give security teams situational awareness to quickly contain incidents[[24]](file://file_000000002ee861f586ba20612cc9ac3b#:~:text=real,window%20analysis)
* Knowledge-sharing components foster broader cybersecurity defense by contributing new threat insights to community feeds[[25]](file://file_000000002ee861f586ba20612cc9ac3b#:~:text=integrate%20dynamic%20threat%20intelligence%2C%20generate,and%20provide%20a%20central%20dashboard)

### Skills and Tools Used

* **Streaming Data Processing:** Utilized message brokers (e.g. Apache Kafka) for high-throughput DNS log ingestion[[19]](file://file_000000002ee861f586ba20612cc9ac3b#:~:text=Kafka%20)
* **Microservices & Containerization:** Designed a containerized microservice architecture for scalable, modular deployment[[26]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=Step%202,handled%20by%20a%20different%20team)[[12]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=We%20clustered%20requirements%20into%20microservices%3B,Ingestion%20%26%20Preprocessing%20Service)
* **Data Enrichment:** Integrated GeoIP and WHOIS lookup services to add contextual metadata to DNS queries[[27]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=1,DNS%20logs%2C%20metadata%20for%20enrichment)
* **Threat Intelligence Integration:** Aggregated external threat intel feeds (blacklists/whitelists) via APIs and databases[[14]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=4)
* **Dashboard & Alerting:** Built web interfaces for interactive monitoring and automated multi-channel notifications to stakeholders[[28]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=6)

### Cross-project Capabilities

* **Real-time Analytics:** Streaming anomaly detection techniques honed here apply to live data in other domains (e.g. social media feeds)[[6]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=2.%20Real,cleaning%2C%20and%20enrichment%20of%20DNS)[[29]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=stream%20of%20tweets%20containing%20at,Figure%201)
* **Multi-source Data Fusion:** Experience combining internal logs with external intel mirrors integration of diverse data (social media + demographics in health projects)[[14]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=4)[[30]](file://file_000000005a7461f59b15548059f89c06#:~:text=improve%20surveillance%20and%20reporting%2C%20two,and%20Chicago%20were%20successful%20in)
* **Dashboard Development:** Skills in creating intuitive dashboards for analysts translate across cybersecurity and epidemiology use-cases[[28]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=6)[[31]](file://file_000000005a7461f59b15548059f89c06#:~:text=FoodBorne%20Chicago%2C%20the%20Dashboard%20monitors,The%20Dashboard%20improves%20surveillance)
* **Machine Learning & NLP:** Applied machine learning for pattern detection in network logs, akin to text classification in health surveillance projects[[32]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=4.%20Historical%20Analytics%20%26%20Large,threats%20or%20advanced%20persistent%20threats)[[33]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=automated%20detection%20of%20foodborne%20illness,We%20developed%20a%20support%20vector)
* **Scalable Big Data Handling:** Competence in managing and analyzing high-volume streaming data is leveraged in projects from DNS security to global mobility[[5]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=1,DNS%20logs%2C%20metadata%20for%20enrichment)[[34]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=Cells%20with%20sufficient%20data%20to,2c%29%2C%20while%20data)

### Published Papers/Tools

* **DNS Tunnel Detection – Design Blueprint (2025):** Internal system design presentation by Gaurav Tuli (Feb 2025)[[35]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=DNS)
* *(This project was delivered as an internal architecture study; no external publication available)*

## Project 12

## Foodborne-1 – Foodborne Illness Surveillance Using Twitter

Project image placeholder

**BLUF:** Created a Twitter-monitoring dashboard for **foodborne illness** that significantly increased citizen reporting of food poisoning cases and enabled faster, targeted restaurant inspections.

### Introduction

* Foodborne illness is common (affecting ~1 in 4 Americans annually) but vastly underreported to health authorities[[36]](file://file_000000005a7461f59b15548059f89c06#:~:text=A,capture%20a%20small%20fraction%20of)
* Traditional surveillance captures only a small fraction of cases, making it hard to prevent outbreaks[[37]](file://file_000000005a7461f59b15548059f89c06#:~:text=become%20sick%20seek%20medical%20care,improving%20disease%20surveillance%20and%20reporting)[[38]](file://file_000000005a7461f59b15548059f89c06#:~:text=systems%20only%20capture%20a%20small,improving%20disease%20surveillance%20and%20reporting)
* Local health departments rely on consumer complaints to trigger inspections, but very few people report illness[[39]](file://file_000000005a7461f59b15548059f89c06#:~:text=Inspections%20prompted%20by%20consumers%20identify,2%2C4)[[40]](file://file_000000005a7461f59b15548059f89c06#:~:text=leading%20to%20a%20higher%20proportion,2%2C4)
* New media platforms (Twitter, etc.) offer a promising data source to improve detection and response for food safety[[41]](file://file_000000005a7461f59b15548059f89c06#:~:text=challenging,improving%20disease%20surveillance%20and%20reporting)[[30]](file://file_000000005a7461f59b15548059f89c06#:~:text=improve%20surveillance%20and%20reporting%2C%20two,and%20Chicago%20were%20successful%20in)

### Key Questions Addressed

* How can social media (Twitter) be used to identify local incidents of food poisoning in real time?[[42]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=uses%20digital%20data%20sources%20such,easy%20to%20use%20and%20designed)
* Will engaging with illness-related tweets encourage more individuals to submit official food poisoning reports?[[43]](file://file_000000005a7461f59b15548059f89c06#:~:text=FoodBorne%20Chicago%2C%20the%20Dashboard%20monitors,detail%20on%20restaurants%2C%20types%20of)
* Do tweet-identified complaints lead to findings of serious restaurant violations comparable to traditional reports?[[44]](file://file_00000000bba861f58ef1975149dab1ba#:~:text=tweets%20resulted%20in%20more%20filed,of%20current%20data%2C%20allowing%20direct)
* Can a Twitter-based surveillance tool be easily implemented by different city health departments?[[45]](file://file_000000005a7461f59b15548059f89c06#:~:text=restaurants%2C%20and%20areas%20with%20higher,3%29%20low%20cost)

### The Problem

* Few sick individuals (~3%) seek medical care or report their illness, so most foodborne cases go unreported[[36]](file://file_000000005a7461f59b15548059f89c06#:~:text=A,capture%20a%20small%20fraction%20of)
* Key violations and outbreaks are missed because health departments receive too little timely information from the public[[39]](file://file_000000005a7461f59b15548059f89c06#:~:text=Inspections%20prompted%20by%20consumers%20identify,2%2C4)
* Existing reporting methods (phone hotlines, fax, long web forms) are inconvenient and slow, deterring public participation[[46]](file://file_000000005a7461f59b15548059f89c06#:~:text=In%20St,screen%20web%20form%20in%20the)
* Without citizen reports, contaminated restaurants remain unchecked longer, prolonging the risk to the community[[39]](file://file_000000005a7461f59b15548059f89c06#:~:text=Inspections%20prompted%20by%20consumers%20identify,2%2C4)[[40]](file://file_000000005a7461f59b15548059f89c06#:~:text=leading%20to%20a%20higher%20proportion,2%2C4)

### The Importance

* Quick identification of foodborne outbreaks allows officials to intervene sooner, preventing further illnesses[[47]](file://file_000000005a7461f59b15548059f89c06#:~:text=Significance%20The%20project%20is%20significant,in%20local%20health%20department%20settings)
* Using real-time social media data enables time-sensitive public health education and engagement with affected individuals[[48]](file://file_00000000bba861f58ef1975149dab1ba#:~:text=other%20mechanisms,spread%20of%20Zika%20virus%20infection)
* Demonstrates how modern **digital surveillance** can augment traditional epidemiology, leading to improved reporting and response[[48]](file://file_00000000bba861f58ef1975149dab1ba#:~:text=other%20mechanisms,spread%20of%20Zika%20virus%20infection)[[49]](file://file_00000000bba861f58ef1975149dab1ba#:~:text=provide%20time,spread%20of%20Zika%20virus%20infection)
* Success in this domain provides a model that can be extended to other public health issues (e.g. tracking disease outbreaks, public safety)[[49]](file://file_00000000bba861f58ef1975149dab1ba#:~:text=provide%20time,spread%20of%20Zika%20virus%20infection)

### The Solution

* Implemented the **HealthMap Foodborne Dashboard**, a web-based tool that monitors Twitter for tweets about food poisoning in a given region[[50]](file://file_000000005a7461f59b15548059f89c06#:~:text=Harvard%20Medical%20School%20developed%20the,reporting%20by%20facilitating%20replies%20to)
* Developed keyword filters and a **machine learning classifier** (SVM) to automatically flag tweets indicating probable foodborne illness[[33]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=automated%20detection%20of%20foodborne%20illness,We%20developed%20a%20support%20vector)[[51]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=poisoning%20after%20eating%20at%20a,feature%20selection%20and%20parameter%20configuration)
* Integrated the dashboard with local health department workflows, allowing inspectors to quickly follow up on credible complaints[[52]](file://file_000000005a7461f59b15548059f89c06#:~:text=Aim%201%3A%20Implement%20the%20HealthMap,processes%20for%20reporting%20foodborne%20illness)
* Enabled staff to respond to detected tweets in real time with a link to an official reporting form, converting tweets into formal reports[[53]](file://file_000000005a7461f59b15548059f89c06#:~:text=can%20be%20replied%20to%2C%20discarded%2C,use%20for%20researchers%20and)
* Dashboard provided contextual details (restaurant name, location clustering) to help target inspections and identify hotspots[[31]](file://file_000000005a7461f59b15548059f89c06#:~:text=FoodBorne%20Chicago%2C%20the%20Dashboard%20monitors,The%20Dashboard%20improves%20surveillance)[[54]](file://file_000000005a7461f59b15548059f89c06#:~:text=improves%20reporting%20by%20facilitating%20replies,3%29%20low%20cost)

### Architecture Overview

* **Data pipeline:** Streamed tweets via Twitter API; enriched data by inferring geolocation for tweets lacking GPS (using profile location & Google Maps API)[[55]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=Our%20database%20currently%20consist%20of,5%20million%20foodservice%20reviews)[[56]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=and%20680%20million%20tweets,A%20report)
* **Automated filtering:** Used a curated keyword list (symptoms, “food poisoning” etc.) to catch relevant tweets, then applied an SVM classifier to eliminate false positives[[33]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=automated%20detection%20of%20foodborne%20illness,We%20developed%20a%20support%20vector)
* **Geo-spatial tagging:** Mapped tweets to the city/county level, enabling visualization of outbreak clusters and affected establishments[[55]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=Our%20database%20currently%20consist%20of,5%20million%20foodservice%20reviews)[[31]](file://file_000000005a7461f59b15548059f89c06#:~:text=FoodBorne%20Chicago%2C%20the%20Dashboard%20monitors,The%20Dashboard%20improves%20surveillance)
* **Real-time dashboard:** Flagged tweets appeared instantly on an interface where officials could review details, reply with one click, or mark false alarms[[31]](file://file_000000005a7461f59b15548059f89c06#:~:text=FoodBorne%20Chicago%2C%20the%20Dashboard%20monitors,The%20Dashboard%20improves%20surveillance)[[53]](file://file_000000005a7461f59b15548059f89c06#:~:text=can%20be%20replied%20to%2C%20discarded%2C,use%20for%20researchers%20and)
* **Deployability:** The system was built to be low-cost and easy to implement in any region, requiring minimal technical setup by health departments[[45]](file://file_000000005a7461f59b15548059f89c06#:~:text=restaurants%2C%20and%20areas%20with%20higher,3%29%20low%20cost)[[57]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=to%20ensure%20food%20safety,easy%20implementation%20for%20any%20region)

### Results and Impacts

* Captured 193 relevant food-poisoning tweets in the first 7 months of deployment (Oct 2015–May 2016)[[58]](file://file_00000000bba861f58ef1975149dab1ba#:~:text=Results%3A%20In%20its%20first%207,following%20reports%20through%20other%20mechanisms)
* Replies to tweets via the dashboard led to more official illness reports than all existing reporting channels in the same period[[58]](file://file_00000000bba861f58ef1975149dab1ba#:~:text=Results%3A%20In%20its%20first%207,following%20reports%20through%20other%20mechanisms)
* Restaurants investigated from tweet-based complaints had rates of serious violations comparable to those found via traditional complaints (no loss in quality of detection)[[59]](file://file_00000000bba861f58ef1975149dab1ba#:~:text=during%20the%20same%20time%20frame,sensitive%20education%20and%20mobilizing)
* The real-time approach proved effective and scalable – the dashboard model was later adopted by health departments in **six U.S. states**[[60]](file://file_0000000078a8622f820b1ed13735a100#:~:text=Interested%20in%20collaborating%20or%20using,departments%20in%20six%20US%20states)

### Skills and Tools Used

* **Twitter API & Streaming:** Leveraged Twitter’s streaming API to continuously collect tweets mentioning food poisoning symptoms[[55]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=Our%20database%20currently%20consist%20of,5%20million%20foodservice%20reviews)
* **Natural Language Processing:** Employed text parsing and keyword matching to filter tweets, and an SVM machine learning model to classify relevant reports[[33]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=automated%20detection%20of%20foodborne%20illness,We%20developed%20a%20support%20vector)
* **Geo-coding:** Used location data (GPS coordinates or user profile locations) and the Google Maps API to determine where incidents occurred[[56]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=and%20680%20million%20tweets,A%20report)
* **Web Development:** Built an interactive web dashboard (UI/UX) for real-time monitoring and one-click response to incidents[[31]](file://file_000000005a7461f59b15548059f89c06#:~:text=FoodBorne%20Chicago%2C%20the%20Dashboard%20monitors,The%20Dashboard%20improves%20surveillance)
* **Public Health Integration:** Collaborated with local health officials to integrate the tool into existing reporting systems and train users[[52]](file://file_000000005a7461f59b15548059f89c06#:~:text=Aim%201%3A%20Implement%20the%20HealthMap,processes%20for%20reporting%20foodborne%20illness)

### Cross-project Capabilities

* **Digital Epidemiology:** Pioneered use of user-generated online data (tweets) for disease surveillance – a technique further applied using review sites in the Yelp project[[61]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=of%20foodborne%20illness%20on%20social,responding%20to%20illness%20reports%20and)[[30]](file://file_000000005a7461f59b15548059f89c06#:~:text=improve%20surveillance%20and%20reporting%2C%20two,and%20Chicago%20were%20successful%20in)
* **Geo-spatial Analysis:** Experience mapping health data geographically (city hotspots) is transferable to global-scale mobility mapping (5 km grids in mobility project)[[55]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=Our%20database%20currently%20consist%20of,5%20million%20foodservice%20reviews)[[62]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=vals%20regardless%20of%20cell%20phone,of%20cells%20as%20well%20as)
* **Real-time Alerting:** Expertise in real-time data pipelines and alert generation (tweets to notifications) aligns with real-time anomaly alerts in cybersecurity projects[[29]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=stream%20of%20tweets%20containing%20at,Figure%201)[[63]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=2.%20Real,cleaning%2C%20and%20enrichment%20of%20DNS)
* **User-Centric Design:** Built tools for non-technical public health users (dashboard ease-of-use), reflecting a broader strength in creating accessible tech solutions[[31]](file://file_000000005a7461f59b15548059f89c06#:~:text=FoodBorne%20Chicago%2C%20the%20Dashboard%20monitors,The%20Dashboard%20improves%20surveillance)[[64]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=6,query%20interfaces%2C%20and%20summary%20views)
* **Machine Learning Application:** Applied ML classifiers to unconventional data (social media text), similar to using ML for pattern detection in network security and large-scale data analysis[[33]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=automated%20detection%20of%20foodborne%20illness,We%20developed%20a%20support%20vector)[[65]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=3.%20Real,confidence%20events)

### Published Papers/Tools

* **Harris et al., 2017 – *Using Twitter to Identify and Respond to Food Poisoning*** (Journal of Public Health Management & Practice)[[66]](file://file_00000000bba861f58ef1975149dab1ba#:~:text=Using%20Twitter%20to%20Identify%20and,Brownstein%2C%20PhD)[[58]](file://file_00000000bba861f58ef1975149dab1ba#:~:text=Results%3A%20In%20its%20first%207,following%20reports%20through%20other%20mechanisms)
* **HealthMap Foodborne Dashboard:** Real-time foodborne illness surveillance platform (deployed at St. Louis DOH and adopted by multiple health departments)[[60]](file://file_0000000078a8622f820b1ed13735a100#:~:text=Interested%20in%20collaborating%20or%20using,departments%20in%20six%20US%20states)

## Project 13

## Foodborne-2 – Foodborne Illness Reporting Disparity Using Yelp

Project image placeholder

**BLUF:** Analyzed 1.5 million Yelp reviews to reveal how climate and socioeconomics affect online food poisoning reports, showing that affluent communities report more – a **digital disparity** in public health surveillance.

### Introduction

* Food safety is a major public health concern worldwide, causing hundreds of millions of illnesses and ~420,000 deaths annually[[67]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=Food%20safety%20is%20%E2%80%9Cthe%20assurance,and%20three%20thousand%20deaths%20an)
* In the U.S., an estimated 48 million people get sick from foodborne diseases each year[[68]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=%28World%20Health%20Organization%20%28WHO%29%2C%202015%29,2011a)
* New technologies (e.g. DNA sequencing, informatics tools) and digital data sources are enhancing our ability to detect and monitor foodborne outbreaks[[69]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=nc,crowdsourced%20reports%20of%20suspected%20foodborne)
* Social media and online review platforms have demonstrated potential for identifying restaurant-related illness clusters that traditional systems miss[[70]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=health%20have%20shown%20that%20crowdsourced,on%20the%20business%20review%20site)

### Key Questions Addressed

* How do **seasonal trends** (climate, time of year) influence the volume of foodborne illness reports posted online?[[71]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=with%20climate%2C%20socio,were%20posi)
* What **demographic and socioeconomic factors** (income, education, etc.) are associated with higher rates of digital illness reporting?[[72]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=factors%20typically%20associated%20with%20affluence,study%2C%20it%20is%20an%20impor)
* Do patterns of online reporting reflect known offline health disparities between communities?[[73]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=capita%20and%20education%20were%20the,the%20online%20public%20health%20environment)
* Which local factors (e.g. restaurant density, income, education) emerge as the strongest predictors of online food poisoning report rates?[[72]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=factors%20typically%20associated%20with%20affluence,study%2C%20it%20is%20an%20impor)

### The Problem

* Although health officials use digital disease reports, little is known about what drives people in different areas to report illness online[[74]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=Although%20digital%20reports%20of%20disease,2014%2C%20and%20use%20census%20data)
* Environmental and community characteristics could bias who reports illness on platforms like Yelp, risking blind spots in surveillance[[75]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=sion%20making%2C%20little%20is%20known,reporting%20of%20disease%20is%20associated)
* If online reporting is uneven across socioeconomic groups, certain outbreaks or communities might be underrepresented in the data[[73]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=capita%20and%20education%20were%20the,the%20online%20public%20health%20environment)
* Without understanding these biases, public health cannot fully trust or calibrate digital surveillance inputs for decision-making[[76]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=results%20suggest%20thatwell,the%20online%20public%20health%20environment)

### The Importance

* Uncovering biases in online reporting is crucial for **equitable** public health surveillance – ensuring all populations are represented[[73]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=capita%20and%20education%20were%20the,the%20online%20public%20health%20environment)
* Knowing seasonal peaks in reporting can help align public health resources with times of higher outbreak risk[[77]](file://file_0000000078a8622f820b1ed13735a100#:~:text=A%20fixed,associated%20with%20higher%20socioeconomic%20status)
* Identifying socio-economic gaps allows health agencies to adjust outreach and not over-rely on data from more vocal (affluent) communities[[72]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=factors%20typically%20associated%20with%20affluence,study%2C%20it%20is%20an%20impor)
* This study provides insight on interpreting digital illness data, making surveillance more reliable as an adjunct to traditional systems[[76]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=results%20suggest%20thatwell,the%20online%20public%20health%20environment)

### The Solution

* Collected ~1.5 million Yelp restaurant reviews (2004–2014) and identified those mentioning foodborne illness using machine learning and keyword filters[[78]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=between%20climate%2C%20demographic%20and%20socio,digital%20reports%20of%20foodborne%20ill)
* Augmented the review dataset with external data: local climate information and county-level demographics (income, education, etc.) from the U.S. Census[[78]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=between%20climate%2C%20demographic%20and%20socio,digital%20reports%20of%20foodborne%20ill)
* Employed statistical models (including a fixed-effects panel regression) to quantify how various factors correlate with the frequency of reported illnesses in reviews[[77]](file://file_0000000078a8622f820b1ed13735a100#:~:text=A%20fixed,associated%20with%20higher%20socioeconomic%20status)[[72]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=factors%20typically%20associated%20with%20affluence,study%2C%20it%20is%20an%20impor)
* Analyzed seasonal patterns by examining trends in review-based reports across different months and climates, validating against known outbreak seasonality[[77]](file://file_0000000078a8622f820b1ed13735a100#:~:text=A%20fixed,associated%20with%20higher%20socioeconomic%20status)
* Assessed the impact of socioeconomics by comparing reporting rates in communities with different income levels, education attainment, and restaurant densities[[72]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=factors%20typically%20associated%20with%20affluence,study%2C%20it%20is%20an%20impor)

### Architecture Overview

* **Data compilation:** Built a decade-long panel dataset of online foodborne illness reports at the county level, linking each Yelp review with its restaurant’s county and timeline[[78]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=between%20climate%2C%20demographic%20and%20socio,digital%20reports%20of%20foodborne%20ill)
* **Review filtering:** Used keyword searches for illness symptoms and context in reviews; likely applied an ML classifier to ensure mentions were true illness reports (excluding irrelevant mentions)[[79]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=reference%20approximately%201.5mil,summer%2C%20when%20most%20foodborne%20out)
* **Data integration:** Merged review data with county attributes (weather patterns, median income, education rates, population) to enable multivariate analysis[[78]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=between%20climate%2C%20demographic%20and%20socio,digital%20reports%20of%20foodborne%20ill)
* **Regression analysis:** Implemented a fixed-effects regression model controlling for unobserved county differences, to isolate the effect of each factor on reporting rates[[77]](file://file_0000000078a8622f820b1ed13735a100#:~:text=A%20fixed,associated%20with%20higher%20socioeconomic%20status)
* **Significance testing:** Evaluated which factors (e.g. “restaurants per capita”, income, education) significantly influence the frequency of reports, identifying the strongest predictors[[72]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=factors%20typically%20associated%20with%20affluence,study%2C%20it%20is%20an%20impor)

### Results and Impacts

* **Seasonality:** Online reports of foodborne illness showed clear seasonal variation – peaking in the summer months (when foodborne outbreaks are most common) and with a smaller rise in winter[[77]](file://file_0000000078a8622f820b1ed13735a100#:~:text=A%20fixed,associated%20with%20higher%20socioeconomic%20status)
* **Socioeconomic correlation:** Wealthier and more educated communities had higher rates of online illness reporting – areas with higher median income and a greater fraction of college graduates reported more incidents[[72]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=factors%20typically%20associated%20with%20affluence,study%2C%20it%20is%20an%20impor)
* **Key predictors:** The density of restaurants and the education level of the population were the most significant predictors of a county’s volume of reports[[72]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=factors%20typically%20associated%20with%20affluence,study%2C%20it%20is%20an%20impor)
* **Digital disparity:** The findings suggest that well-known health disparities are mirrored online – affluent areas tend to generate more reports, indicating a digital divide in participatory surveillance[[73]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=capita%20and%20education%20were%20the,the%20online%20public%20health%20environment)
* This study underscores the need for public health to account for such biases and possibly complement digital data with targeted outreach in underreporting communities[[73]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=capita%20and%20education%20were%20the,the%20online%20public%20health%20environment)

### Skills and Tools Used

* **Data Mining:** Processed a vast corpus of Yelp reviews (1.5M) to extract relevant health-related reports[[79]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=reference%20approximately%201.5mil,summer%2C%20when%20most%20foodborne%20out)
* **Natural Language Processing:** Used text analysis (keywords and ML classification) to identify reviews describing food poisoning incidents from free-text feedback
* **Statistical Analysis:** Applied advanced statistics (panel regression, correlation analysis, significance testing) to epidemiological data[[77]](file://file_0000000078a8622f820b1ed13735a100#:~:text=A%20fixed,associated%20with%20higher%20socioeconomic%20status)
* **Data Integration:** Combined heterogeneous data sources (crowdsourced reviews, climate records, census data) into a unified analytical dataset[[78]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=between%20climate%2C%20demographic%20and%20socio,digital%20reports%20of%20foodborne%20ill)
* **Computing Tools:** Utilized programming languages and libraries (e.g. R’s statistical packages) for data manipulation, modeling, and visualization[[80]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=Code%20availability%20Standard%20R,No%20custom%20code%20was%20developed)

### Cross-project Capabilities

* **Data-Driven Health Insights:** Expertise in transforming non-traditional big data (consumer reviews) into public health evidence, paralleling the Twitter project’s use of social media and the mobility project’s use of telecom data[[70]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=health%20have%20shown%20that%20crowdsourced,on%20the%20business%20review%20site)[[81]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=we%20describe%20global%20human%20mobility,data%20and%20find%20that%20human)
* **Multivariate Modeling:** Skilled in applying and interpreting complex statistical models – a competency also crucial in global mobility analysis (e.g. validating mobility metrics against socio-demographic indicators)[[77]](file://file_0000000078a8622f820b1ed13735a100#:~:text=A%20fixed,associated%20with%20higher%20socioeconomic%20status)[[82]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=was%20aggregated%20to%20the%20country,fit%20of%20human%20mobility%20metrics)
* **Domain Expertise Integration:** Ability to merge epidemiological context with data science techniques, seen here and in other projects (e.g. integrating threat intel in cybersecurity, or health metrics in mobility)[[83]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=lance%20of%20foodborne%20diseases,on%20the%20business%20review%20site)[[14]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=4)
* **Temporal Trend Analysis:** Experienced in analyzing time-series and seasonal patterns in data (illness reporting peaks, mobility periodicity), useful across projects dealing with time-dependent phenomena[[77]](file://file_0000000078a8622f820b1ed13735a100#:~:text=A%20fixed,associated%20with%20higher%20socioeconomic%20status)[[84]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=Travel%20synchrony%20and%20periodicity%20across,Surprisingly%2C%20however%2C%20Northern%20and)
* **Scalable Analytics:** Proven capacity to handle and analyze millions of records (reviews, tweets, mobility points) using efficient pipelines, a cross-cutting skill in all projects[[78]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=between%20climate%2C%20demographic%20and%20socio,digital%20reports%20of%20foodborne%20ill)[[55]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=Our%20database%20currently%20consist%20of,5%20million%20foodservice%20reviews)

### Published Papers/Tools

* **Henly et al., 2017 – *Disparities in digital reporting of illness: A demographic and socioeconomic assessment*** (Preventive Medicine)[[85]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=j%20ourna%20l%20homepage%3A%20www,Hospital%2C%20Boston%2C%20MA%2C%20United%20States)[[72]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=factors%20typically%20associated%20with%20affluence,study%2C%20it%20is%20an%20impor)
* *(This project yielded a research publication; no separate software tool was developed for public use apart from analytical scripts)*

## Project 14

## Mapping Global Variation in Human Mobility (Google)

Project image placeholder

**BLUF:** Mapped human mobility worldwide by analyzing anonymized Google Location data from 300 million users – revealing new insights into global movement patterns, disparities between regions, and providing an unprecedented open dataset for epidemiological modeling.

### Introduction

* Human mobility underpins many societal dynamics – it affects economic development, traffic, disease spread, disaster response, and more[[86]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=Human%20mobility%20is%20an%20essential,conflict5%2C%20natural%20disasters6%2C7%2C%20infectious%20disease)
* A global understanding of mobility is crucial for sustainability and public health (e.g., reducing carbon footprint, improving access to healthcare)[[87]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=transmission8%2C9%2C%20access%20to%20healthcare10%2C11%2C%20and,and%20sociocultural%20back)
* Until recently, studies of human movement were limited by data – often confined to single-country surveys or sparse mobile phone records[[88]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=Previous%20work%20and%20data%20on,for%20only%20a%20single%20country25)
* Basic, comprehensive measures of movement across all countries were lacking, hindering our ability to quickly assess global events like pandemics or natural disasters[[89]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=been%20limited%20by%20the%20geographical,income%20regions13%E2%80%9315)

### Key Questions Addressed

* How can we obtain a **globally consistent** measure of human mobility covering nearly all countries?[[89]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=been%20limited%20by%20the%20geographical,income%20regions13%E2%80%9315)
* How do mobility patterns vary between regions with different sociodemographic and environmental contexts (e.g. high-income vs low-income countries)?[[90]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=causes%20of%20these%20differences%20may,Further%2C%20statistical%20descriptions%20of)[[91]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=The%20geographic%20variation%20of%20human,international%20movement%20patterns%20across%20national)
* What universal “laws” or scaling patterns govern human travel distances and frequencies around the world, and how do these differ by context?[[91]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=The%20geographic%20variation%20of%20human,international%20movement%20patterns%20across%20national)
* Can large-scale mobile phone location data be used in a privacy-safe way to fill gaps in our understanding of global movement?[[92]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=prehensive%20global%20dataset%20comprising%20de,extensive%20information%20on%20human%20move)[[93]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=sis%2C%20locations%20were%20abstracted%20and,recorded%20human%20movement%20data%20were)

### The Problem

* **Data scarcity:** There was no accurate, high-resolution data covering human movement for the majority of countries[[89]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=been%20limited%20by%20the%20geographical,income%20regions13%E2%80%9315)
* Prior mobility studies were fragmented – limited to specific regions or requiring active user input (e.g., travel surveys, call records), resulting in sparse or biased datasets[[88]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=Previous%20work%20and%20data%20on,for%20only%20a%20single%20country25)
* Without a global mobility dataset, we couldn’t robustly compare movement patterns across rich and poor regions, or urban vs rural areas[[94]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=paramount,lacking%2C%20thereby%20prohibiting%20accurate%20and)[[88]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=Previous%20work%20and%20data%20on,for%20only%20a%20single%20country25)
* This lack of data hampered our ability to model and respond to global phenomena (e.g., predict disease spread internationally or plan infrastructure in low-mobility areas)[[89]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=been%20limited%20by%20the%20geographical,income%20regions13%E2%80%9315)

### The Importance

* Knowing how people move on a global scale improves epidemiological models for infectious diseases (by providing realistic data on travel and mixing)[[95]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=Human%20mobility%20is%20an%20essential,conflict5%2C%20natural%20disasters6%2C7%2C%20infectious%20disease)
* It aids in planning transport and urban development, and in understanding economic connectivity between regions[[86]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=Human%20mobility%20is%20an%20essential,conflict5%2C%20natural%20disasters6%2C7%2C%20infectious%20disease)[[96]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=global%20scales,amount%20and%20quality%20of%20data)
* Highlighting differences in mobility (e.g., much shorter travel in low-income regions) points to inequalities in access and can inform targeted improvements in infrastructure and health access[[94]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=paramount,lacking%2C%20thereby%20prohibiting%20accurate%20and)
* A global mobility typology establishes a baseline for research – enabling countries to **benchmark** against each other and track changes (e.g., due to interventions or crises)[[89]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=been%20limited%20by%20the%20geographical,income%20regions13%E2%80%9315)

### The Solution

* Partnered with **Google** to use de-identified, aggregated data from Google Location History, representing movements of over 300 million smartphone users in 2016[[97]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=A%20global%20human%20mobility%20map,The%20data%20were%20initially%20col)
* Processed the raw location pings into trips between places: identified places where users spend time and aggregated movement flows between ~5×5 km grid cells worldwide[[92]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=prehensive%20global%20dataset%20comprising%20de,extensive%20information%20on%20human%20move)[[62]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=vals%20regardless%20of%20cell%20phone,of%20cells%20as%20well%20as)
* Applied **differential privacy** techniques – only analyzing flows that meet anonymity thresholds – to protect individual privacy while using granular data[[93]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=sis%2C%20locations%20were%20abstracted%20and,recorded%20human%20movement%20data%20were)
* Developed a typology of human movement by using statistical and machine learning methods to characterize distance and travel frequency distributions for each region[[91]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=The%20geographic%20variation%20of%20human,international%20movement%20patterns%20across%20national)
* Validated the data by fitting models (e.g., power-law distance decay) and comparing to known indicators (population, income), and released the mobility metrics as an open dataset for researchers[[98]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=terns%20vary%20across%20sociodemographic%20and,demographic%2C%20economic%20and%20environmental%20changes)[[99]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=borders,demographic%2C%20economic%20and%20environmental%20changes)

### Architecture Overview

* **Data pipeline:** Ingested billions of location points from Google’s opt-in users globally; identified “stops” (significant locations) and computed **origin-destination** flows between stops for all user trips[[92]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=prehensive%20global%20dataset%20comprising%20de,extensive%20information%20on%20human%20move)[[62]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=vals%20regardless%20of%20cell%20phone,of%20cells%20as%20well%20as)
* **Unified spatial grid:** Mapped all coordinates to a standard S2 geometry grid (~5 km cells), enabling uniform analysis across countries and seamless aggregation of movement data[[62]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=vals%20regardless%20of%20cell%20phone,of%20cells%20as%20well%20as)
* **Privacy filtering:** Implemented strict criteria (differential privacy) – only included flows with sufficient users and aggregated all data to ensure no individual could be identified[[93]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=sis%2C%20locations%20were%20abstracted%20and,recorded%20human%20movement%20data%20were)
* **Analytical framework:** Used statistical models to derive key mobility metrics (e.g. average distances traveled, frequency of travel above certain distances) and to test how these correlate with external factors (e.g. income level)[[91]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=The%20geographic%20variation%20of%20human,international%20movement%20patterns%20across%20national)[[82]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=was%20aggregated%20to%20the%20country,fit%20of%20human%20mobility%20metrics)
* **Collaboration & tools:** Multi-institution effort (Google and academic partners); employed standard data science tools (R statistical packages) to analyze the data efficiently at scale[[80]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=Code%20availability%20Standard%20R,No%20custom%20code%20was%20developed)

### Results and Impacts

* Created the first comprehensive **global human mobility map**, covering nearly all countries and 65% of Earth’s populated surface (data from regions home to ~2.9 billion people)[[100]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=The%20geographic%20variation%20of%20human,and%20environmental%20contexts%20and%20present)[[34]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=Cells%20with%20sufficient%20data%20to,2c%29%2C%20while%20data)
* Quantified stark differences in movement: in low-income settings, people travel much shorter distances on average and movement drops off 40% faster with distance, compared to high-income settings[[98]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=terns%20vary%20across%20sociodemographic%20and,demographic%2C%20economic%20and%20environmental%20changes)
* Validated that human movement follows certain scaling laws universally, but with different parameters – for example, “movement laws” that applied at city scales also hold globally but at ~10× shorter characteristic distances in poorer regions[[98]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=terns%20vary%20across%20sociodemographic%20and,demographic%2C%20economic%20and%20environmental%20changes)
* Revealed globally synchronized mobility patterns: identified seasonal peaks in mid-year (July–August) that align across hemispheres, and observed expected dips during winter and major holidays[[84]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=Travel%20synchrony%20and%20periodicity%20across,Surprisingly%2C%20however%2C%20Northern%20and)
* **Open data impact:** The resulting mobility dataset and findings were made publicly available, providing a valuable resource for epidemic modeling, urban planning, and for comparing mobility in future studies[[99]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=borders,demographic%2C%20economic%20and%20environmental%20changes)

### Skills and Tools Used

* **Big Data Processing:** Managed and analyzed an extremely large dataset (billions of location records) with cloud-based data processing pipelines
* **Geospatial Analysis:** Applied spatial algorithms (grid indexing, distance calculations) to derive movement flows and visualize mobility patterns[[62]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=vals%20regardless%20of%20cell%20phone,of%20cells%20as%20well%20as)
* **Privacy Engineering:** Implemented differential privacy and anonymity checks, ensuring ethical use of sensitive location data[[93]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=sis%2C%20locations%20were%20abstracted%20and,recorded%20human%20movement%20data%20were)
* **Statistical Computing:** Utilized R and Python for heavy data analysis, pattern recognition, and model fitting on large-scale movement data[[80]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=Code%20availability%20Standard%20R,No%20custom%20code%20was%20developed)
* **Interdisciplinary Collaboration:** Coordinated between tech industry and academic teams, blending data science with epidemiological insight to interpret mobility data[[97]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=A%20global%20human%20mobility%20map,The%20data%20were%20initially%20col)

### Cross-project Capabilities

* **Large-Scale Analytics:** Demonstrated ability to derive insights from huge datasets (global mobility) – a skill mirrored in handling multi-million record datasets in other projects (tweets, reviews)[[34]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=Cells%20with%20sufficient%20data%20to,2c%29%2C%20while%20data)[[55]](file://file_000000009db461f5b7ed2b239df171d5#:~:text=Our%20database%20currently%20consist%20of,5%20million%20foodservice%20reviews)
* **Privacy & Data Governance:** Experience applying privacy safeguards (differential privacy for mobility data) is relevant to managing sensitive data in health and security projects (e.g., compliance in DNS analytics)[[93]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=sis%2C%20locations%20were%20abstracted%20and,recorded%20human%20movement%20data%20were)[[101]](file://file_000000004058620c96cf547f8d3b01a4#:~:text=o%20Provide%20historical%20and%20trending,Governance%20%26%20Compliance)
* **Global vs Local Analysis:** Versatility in analyzing data at different scales – from granular city-level outbreaks (Foodborne projects) to worldwide patterns – showing adaptability of methods across domains[[91]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=The%20geographic%20variation%20of%20human,international%20movement%20patterns%20across%20national)[[52]](file://file_000000005a7461f59b15548059f89c06#:~:text=Aim%201%3A%20Implement%20the%20HealthMap,processes%20for%20reporting%20foodborne%20illness)
* **Advanced Modeling:** Fitting and validating complex models (power-law mobility decay) reflects an advanced analytical skill set, also employed in other projects (e.g. regression in digital illness reporting)[[98]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=terns%20vary%20across%20sociodemographic%20and,demographic%2C%20economic%20and%20environmental%20changes)[[72]](file://file_00000000dc5061f599fcea8c5bb2d46d#:~:text=factors%20typically%20associated%20with%20affluence,study%2C%20it%20is%20an%20impor)
* **Cross-sector Collaboration:** Proven ability to work with corporate data providers and cross-disciplinary teams, similar to engaging health departments in community surveillance projects[[97]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=A%20global%20human%20mobility%20map,The%20data%20were%20initially%20col)[[52]](file://file_000000005a7461f59b15548059f89c06#:~:text=Aim%201%3A%20Implement%20the%20HealthMap,processes%20for%20reporting%20foodborne%20illness)

### Published Papers/Tools

* **Kraemer et al., 2020 – *Mapping global variation in human mobility*** (Nature Human Behaviour)[[102]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=The%20geographic%20variation%20of%20human,quantify%20how%20human%20movement%20pat)[[98]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=terns%20vary%20across%20sociodemographic%20and,demographic%2C%20economic%20and%20environmental%20changes)
* **Global Mobility Dataset (2016):** Anonymized worldwide human movement data (aggregated flows) released with the above study for public use[[99]](file://file_0000000057b461f593f231f1d9e1f6a2#:~:text=borders,demographic%2C%20economic%20and%20environmental%20changes)