# A New Platform for High Density Data Collection in Population Health Research


Mathew Kiang[1]
@mathewkiang

Patrick Staples[2]
@ctphoenix

John Torous[3]
@JohnTorousMD

Monica Alexander[4]
@monjalexander

Jukka-Pekka "JP" Onnela[5]
@jponnela

[1]Department of Social and Behavioral Sciences, Harvard TH Chan School of Public Health
[2]Core Data Science Team, Mindstrong
[3]Department of Psychiatry, Beth Israel Deaconess Medical Center, Harvard Medical School
[4]Department of Demography, University of California Berkeley
[5]Department of Biostatistics, Harvard TH Chan School of Public Health

# INTRODUCTION

1. What is *our* problem?
2. How to approach the problem?
3. What is Beiwe?
4. What are we doing?
5. Who should use this?

# OUR PROBLEM

- Traditional data falls on a spectrum:
  - Wide and shallow (e.g., national surveys)
  - Deep but narrow (e.g., ethnographies)
  - In between but infrequent (e.g., census)

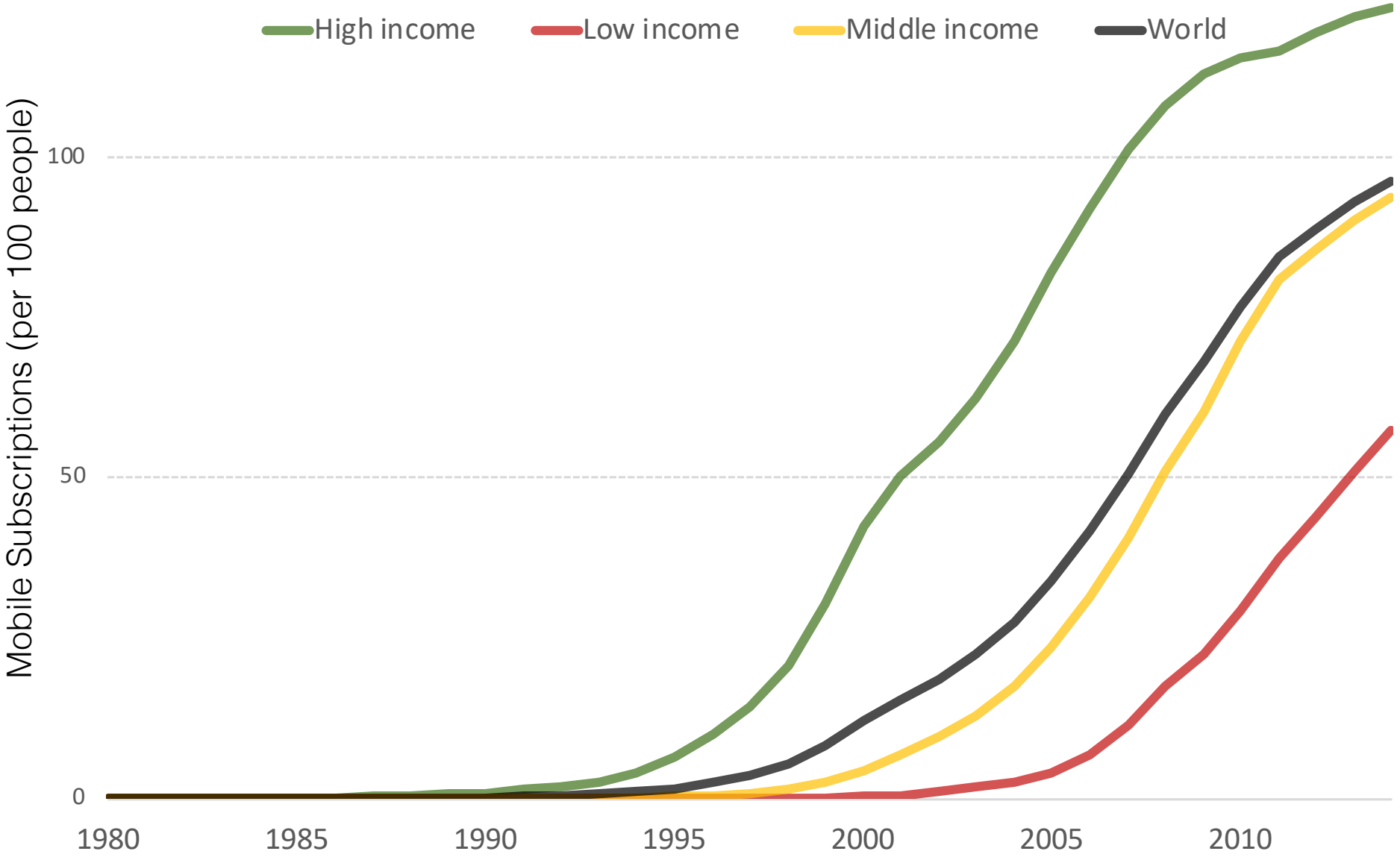
# OUR PROBLEM

- Many social science questions of interest need a lot of individual-level over a long period of time.
- (Insert “big data” joke.)

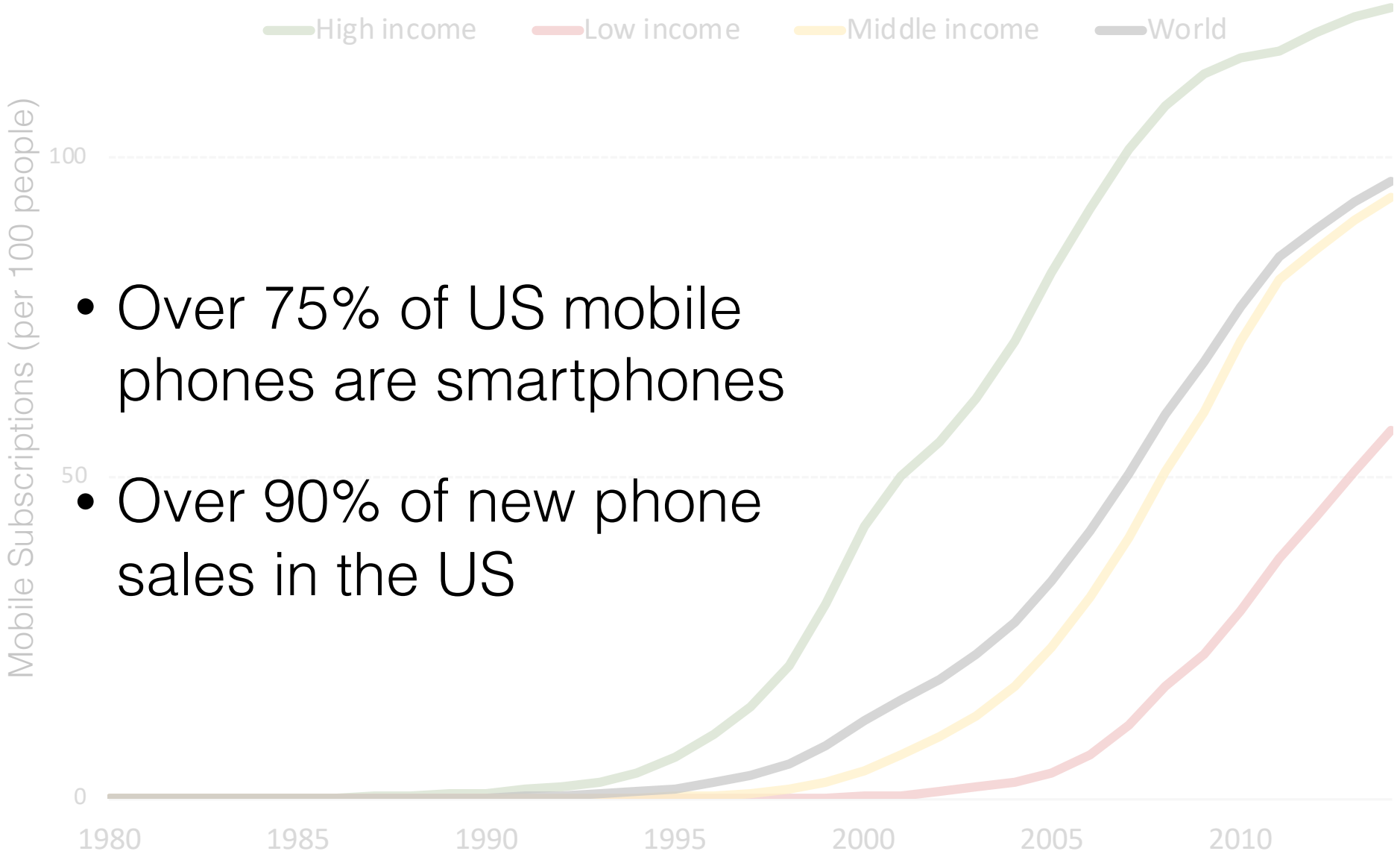
# A NEW TOOL

- More mobile phones than humans
- Fastest growing manmade phenomenon ever

# A NEW TOOL


High income
Low income
Middle income
World
Mobile Subscriptions (per 100 people)
100
50
0
1980
1985
1990
1995
2000
2005
2010

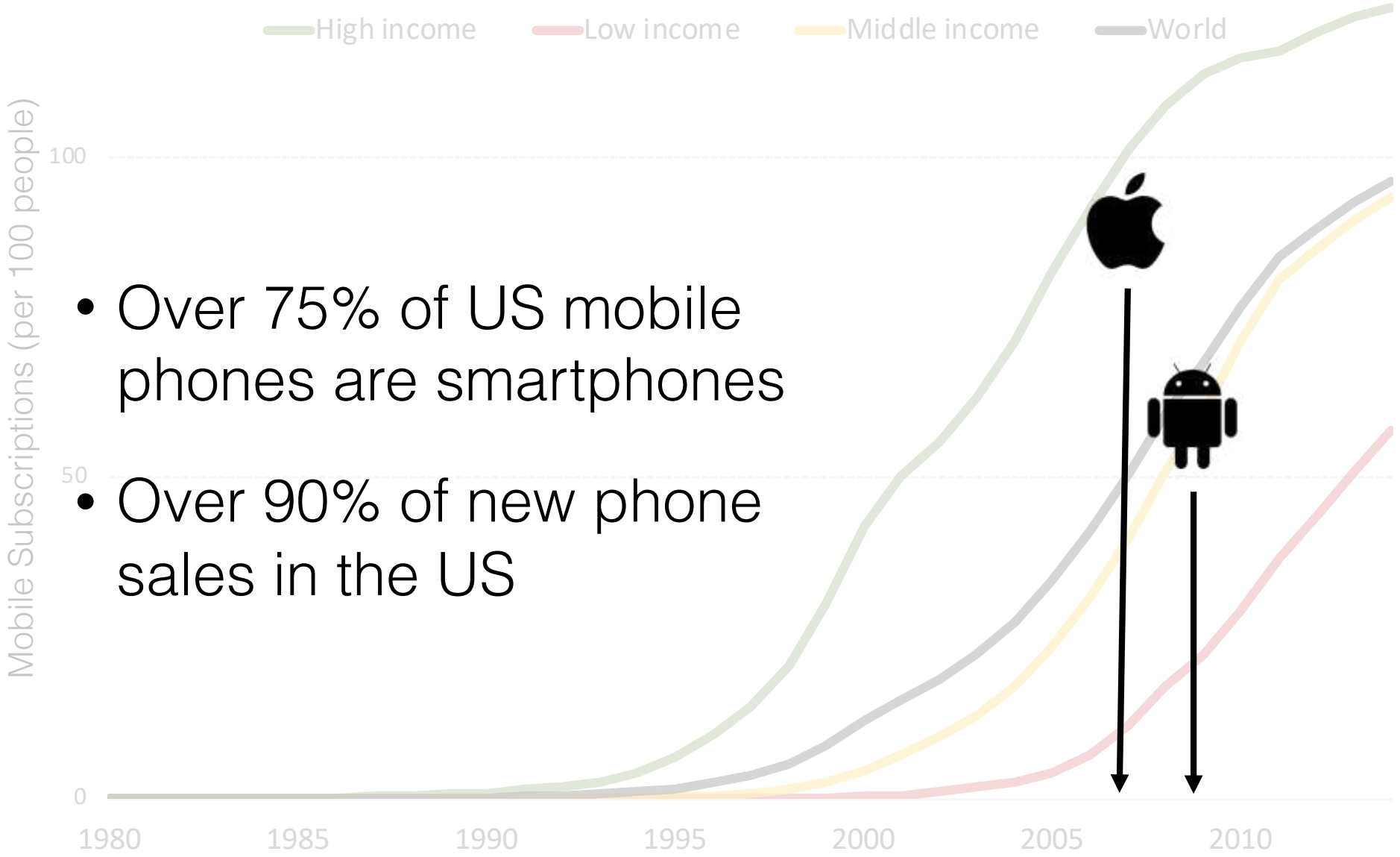
# A NEW TOOL




- Over 75% of US mobile phones are smartphones
- Over 90% of new phone sales in the US

# A NEW TOOL

- Over 75% of US mobile phones are smartphones
- Over 90% of new phone sales in the US



# A NEW TOOL

- GPS
- Accelerometer
- Gyroscope
- Barometer
- Wi-Fi
- Bluetooth
- Microphone
- Gyroscope
- Magnetometer
- Traditional surveys
- Front and rear camera
- Pressure-sensitive display

# A NEW TOOL

- GPS
- Accelerometer
- Gyroscope
- Barometer
- Wi-Fi
- Bluetooth
- Microphone
- Gyroscope
- Magnetometer
- Traditional surveys
- Front and rear camera
- Pressure-sensitive display

# WHAT IS BEIWE?

*Beiwe is an open-source, digital phenotyping research platform designed for the collection and analysis of research-grade data from smartphones*

# WHAT IS BEIWE?

*Beiwe is an open-source, digital phenotyping research platform designed for the collection and analysis of research-grade data from smartphones*

# DIGITAL PHENOTYPING

- "Moment-by-moment quantification of the individual-level human phenotype *in situ*…"[1]

- Beiwe is our attempt to use these sensors to collect data about human behavior in the real world, at scale, and over long periods of time

[1]J Torous, MV Kiang, J Lorme, & JP Onnela. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Mental Health, 3(2), 2016.

# WHAT IS BEIWE?

*Beiwe is an open-source, digital phenotyping research platform designed for the collection and analysis of research-grade data from smartphones*

# RESEARCH-GRADE

- Research-grade data collection:
  - Raw data is stored so new methods can be used on previously collected data
  - Proprietary data may change with no way to adjust
  - Data collection should be as regular and predictable as possible

# WHAT IS BEIWE?

*Beiwe is an open-source, digital phenotyping research platform designed for the collection and analysis of research-grade data from smartphones*

# RESEARCH PLATFORM

- A research platform:
  - Privacy of patient data and replicability were central design goals
  - Researchers specify which sensors, how often, how long, etc.
  - All these settings can be exported and distributed to other researchers

# RESEARCH PLATFORM

- A research platform:
  - Phone apps, server backend, analysis pipeline — all **open-source**
  - Infinitely scalable backend on AWS
  - Useful, transparent, shareable analysis pipeline

# BEIWE DATA SUMMARIES (PARTIAL LIST)

**Mobility features**

1. Time spent at home
2. Distance travelled
3. Radius of gyration
4. Maximum diameter
5. Maximum distance from home
6. Number of significant locations
7. Average flight length
8. Standard deviation of flight length
9. Average flight duration
10. Standard deviation of flight duration
11. Fraction of the day spent stationary
12. Significant location entropy
13. Minutes of GPS data missing
14. Physical circadian rhythm
15. Physical circadian rhythm stratified

**Sociability features**

1. Number of outgoing texts
2. Total outgoing text length
3. Texting out-degree
4. Number of incoming texts
5. Total incoming text length
6. Texting in-degree
7. Texting reciprocity
8. Texting responsiveness
9. Number of outgoing calls
10. Total outgoing call duration
11. Call out-degree
12. Number of incoming calls
13. Total incoming call durations
14. Call in-degree
15. Call reciprocity
16. Call responsiveness

**Surveys**

1. Survey ID
2. Survey score
3. Time to begin survey (seconds)
4. Time to complete survey once begun (seconds)
5. Survey completed?

**Activity**

1. Fraction of time accelerometer is active
2. Variation in accelerometer measurements
3. Total number of steps
4. Classification of activity

**Sleep**

1. Sleep onset time
2. Awake onset time
3. Sleep duration

**Power state**

1. Total number of screen events

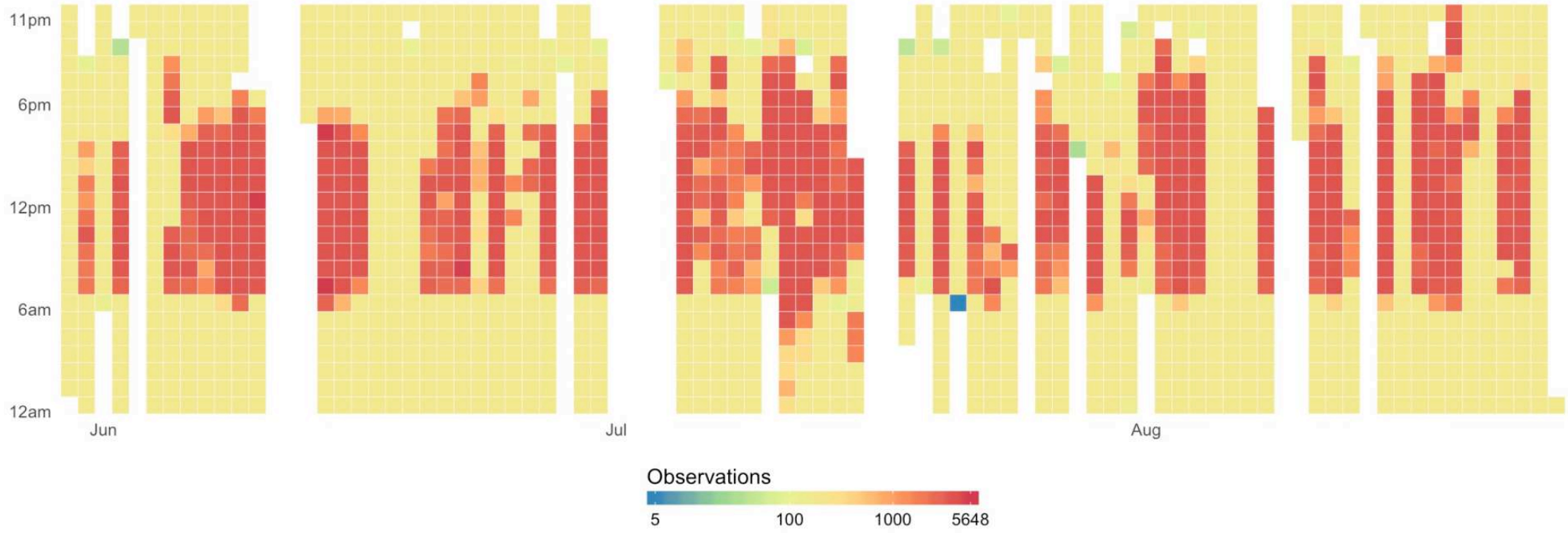
**Data coverage**

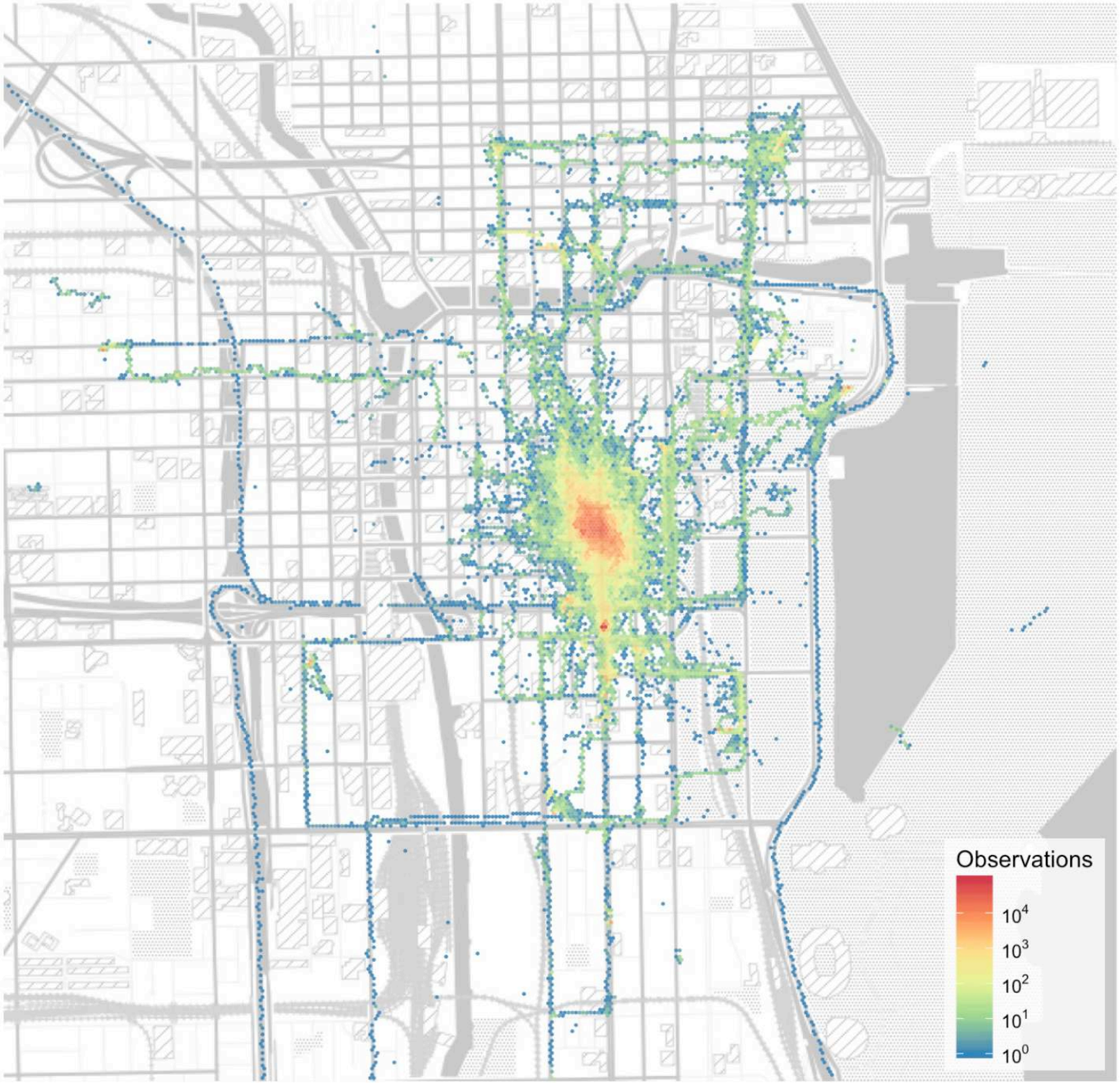
1. Bursts coverage
2. Within burst length coverage
3. Within burst frequency coverage
4. Total coverage

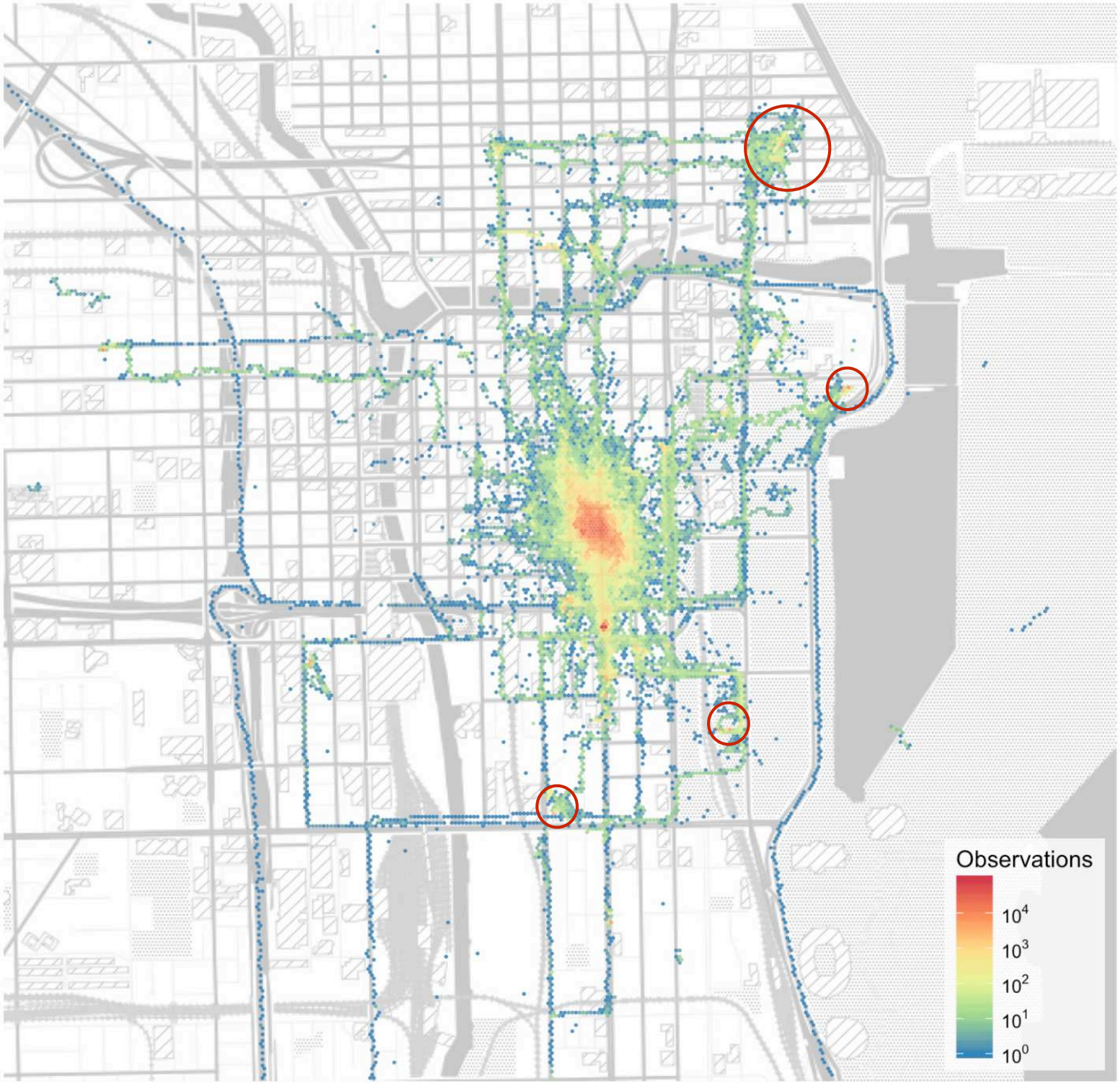
# WHAT IS BEIWE?

*Beiwe is an open-source, digital phenotyping research platform designed for the collection and analysis of research-grade data from smartphones*

# ACCELEROMETER DATA



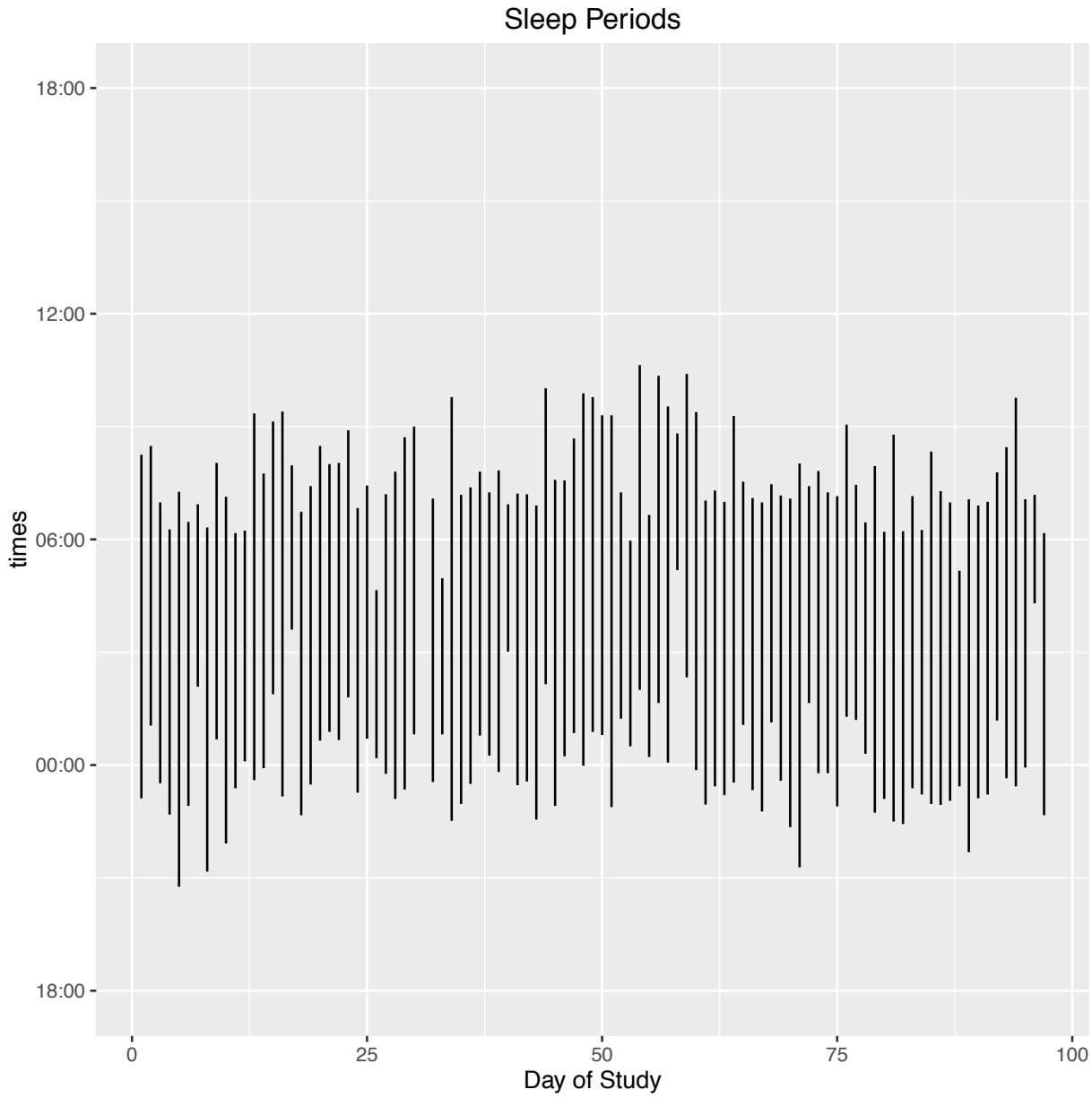

Observations
$10^4$
$10^3$
$10^2$
$10^1$
$10^0$


Observations
$10^4$
$10^3$
$10^2$
$10^1$
$10^0$

# WHAT ARE WE DOING?

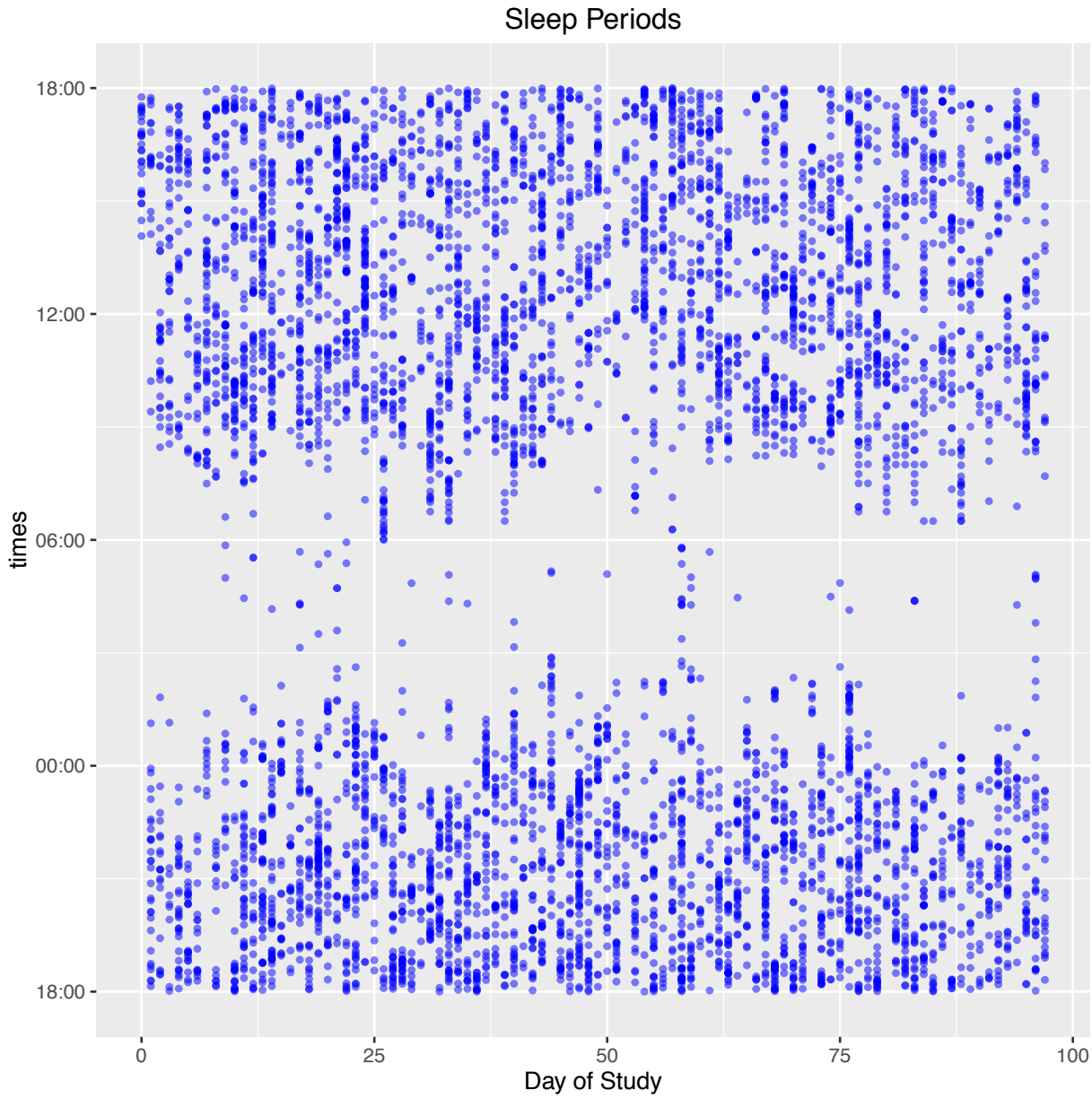
# QUANTIFYING SLEEP

- Actigraphy data from a gold standard instrument ($1,500)
- 1 patient over approximately 100 days



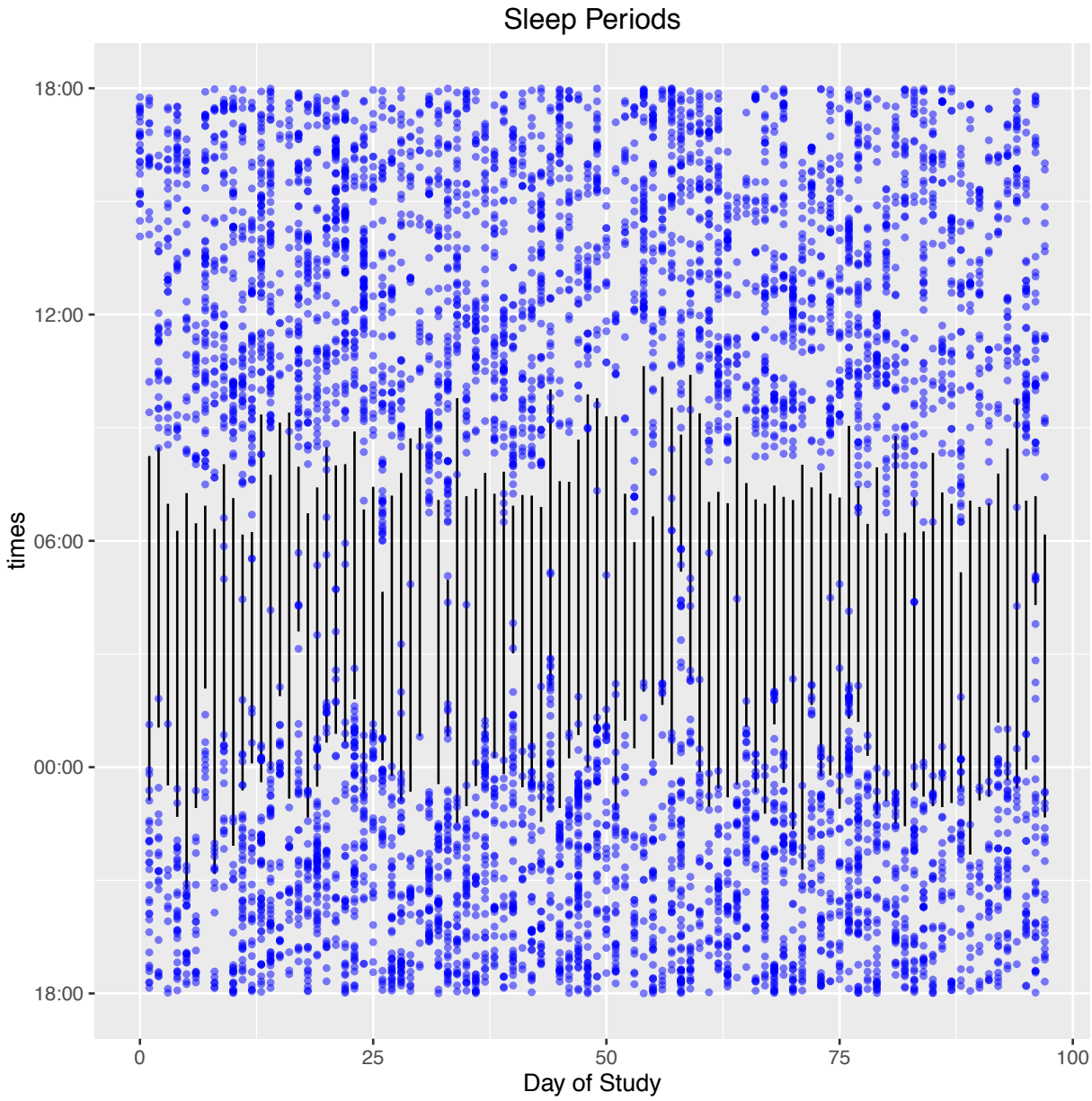
# QUANTIFYING SLEEP

- Beiwe data using only a single data stream
- Each dot is screen on/off event

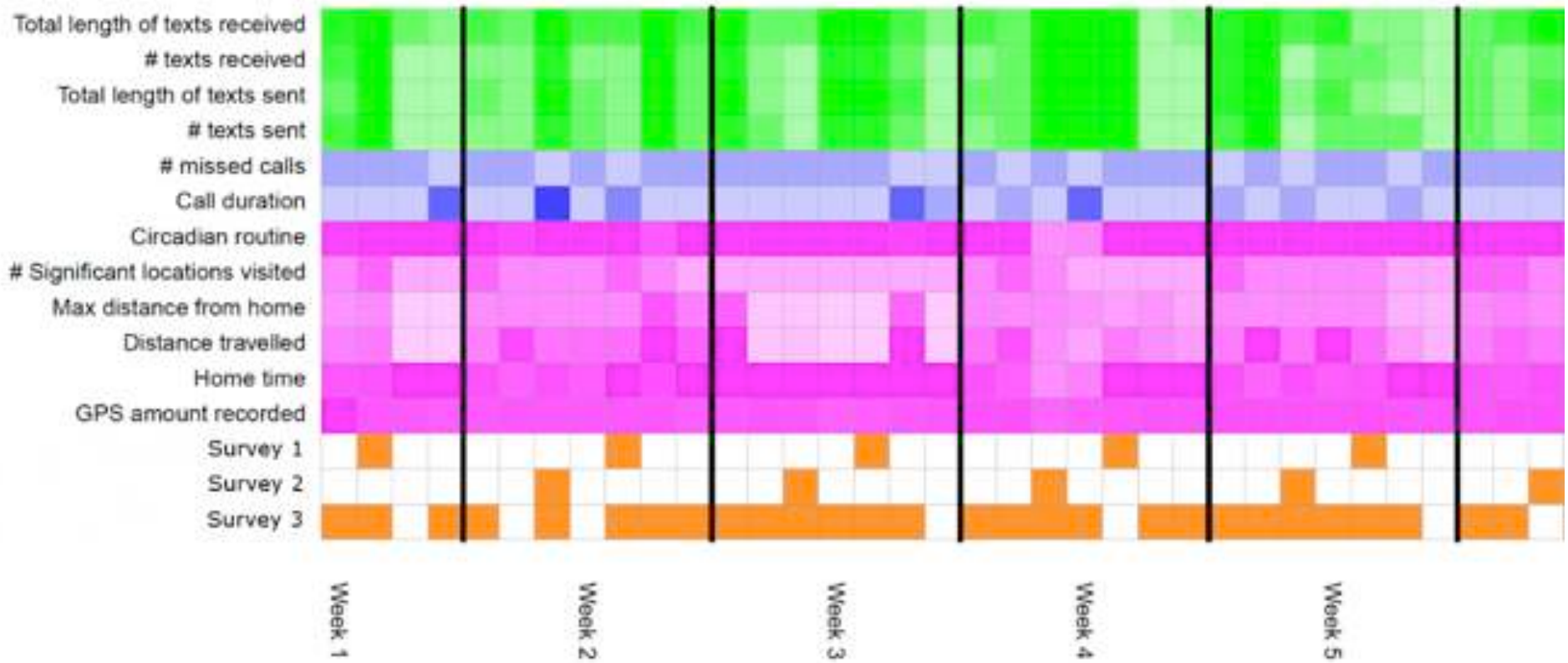


# QUANTIFYING SLEEP

- We can infer sleep patterns with reasonably high resolution
- Obviously not as high fidelity as the gold standard instrument, but significant more scalable.

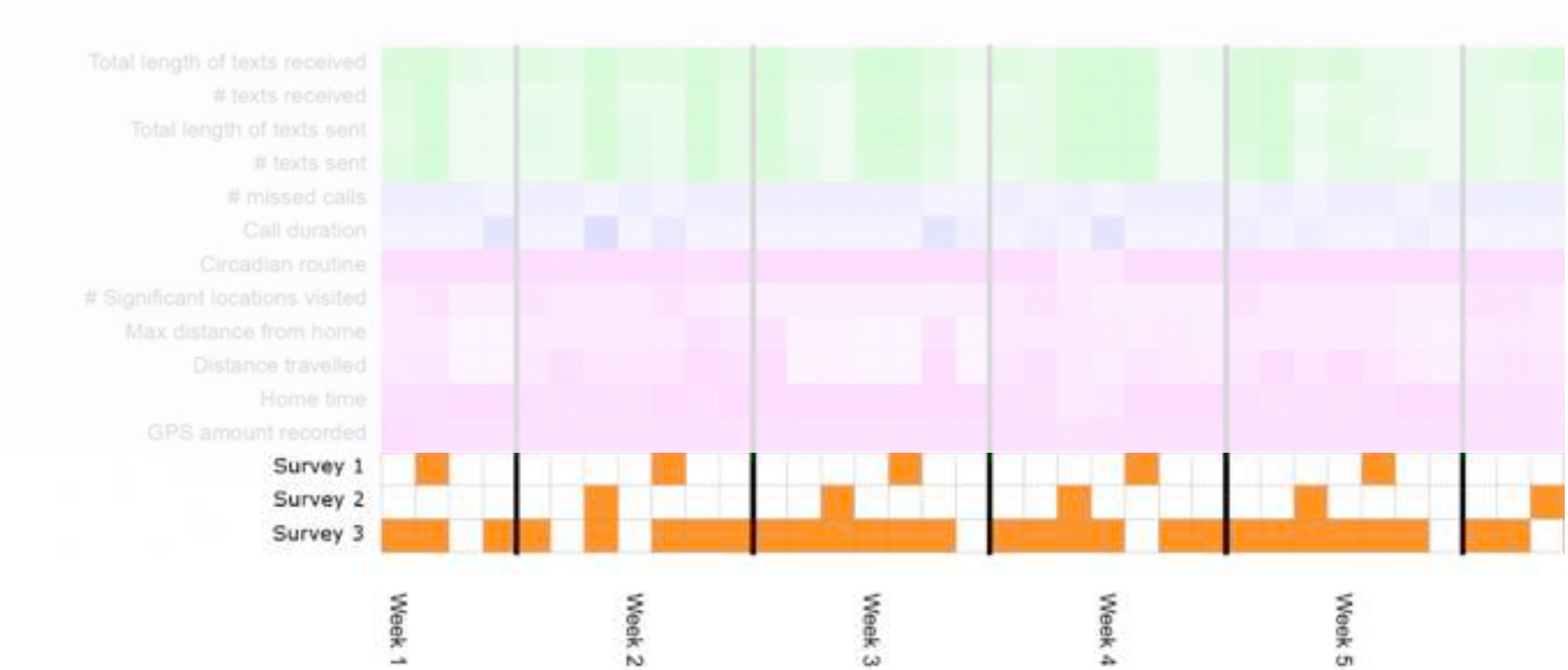


# UNDERSTANDING PAIN



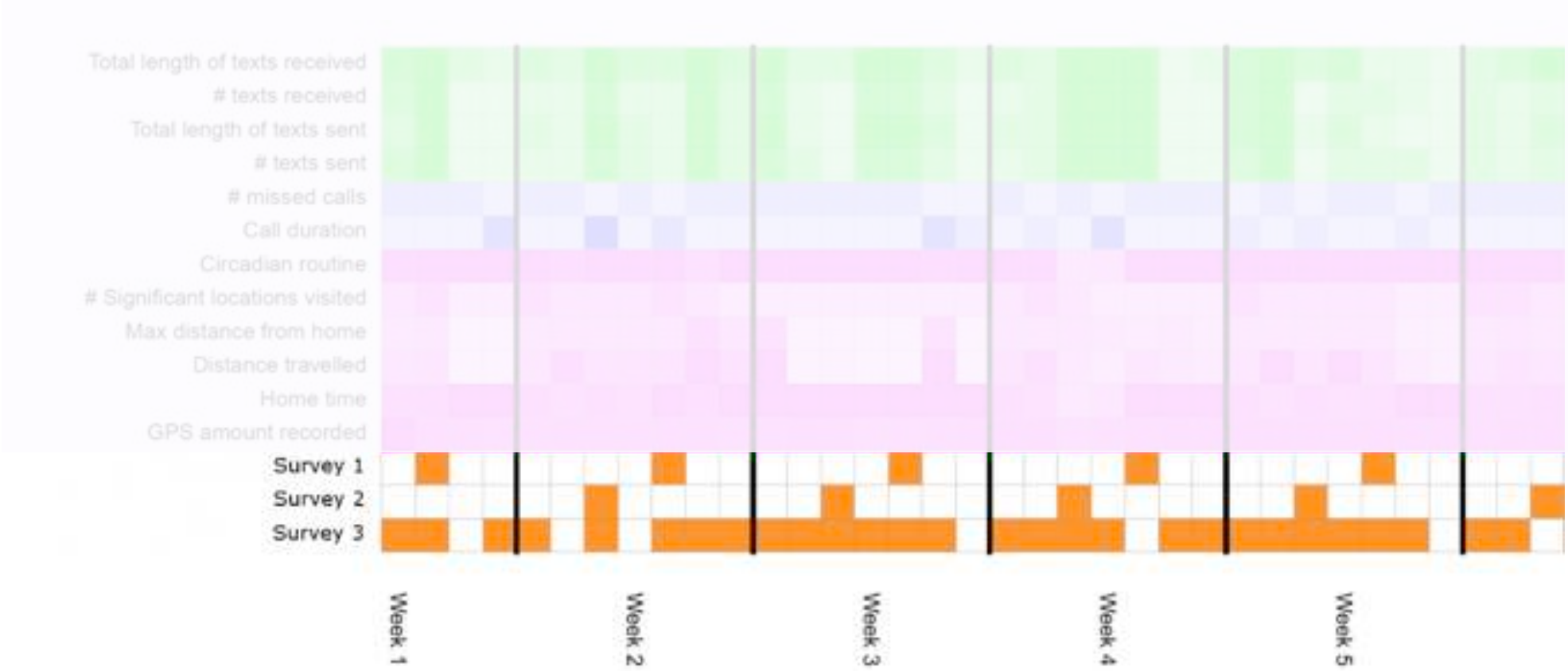

- Active (orange) and passive data streams for one patient in a spinal surgery study

# UNDERSTANDING PAIN



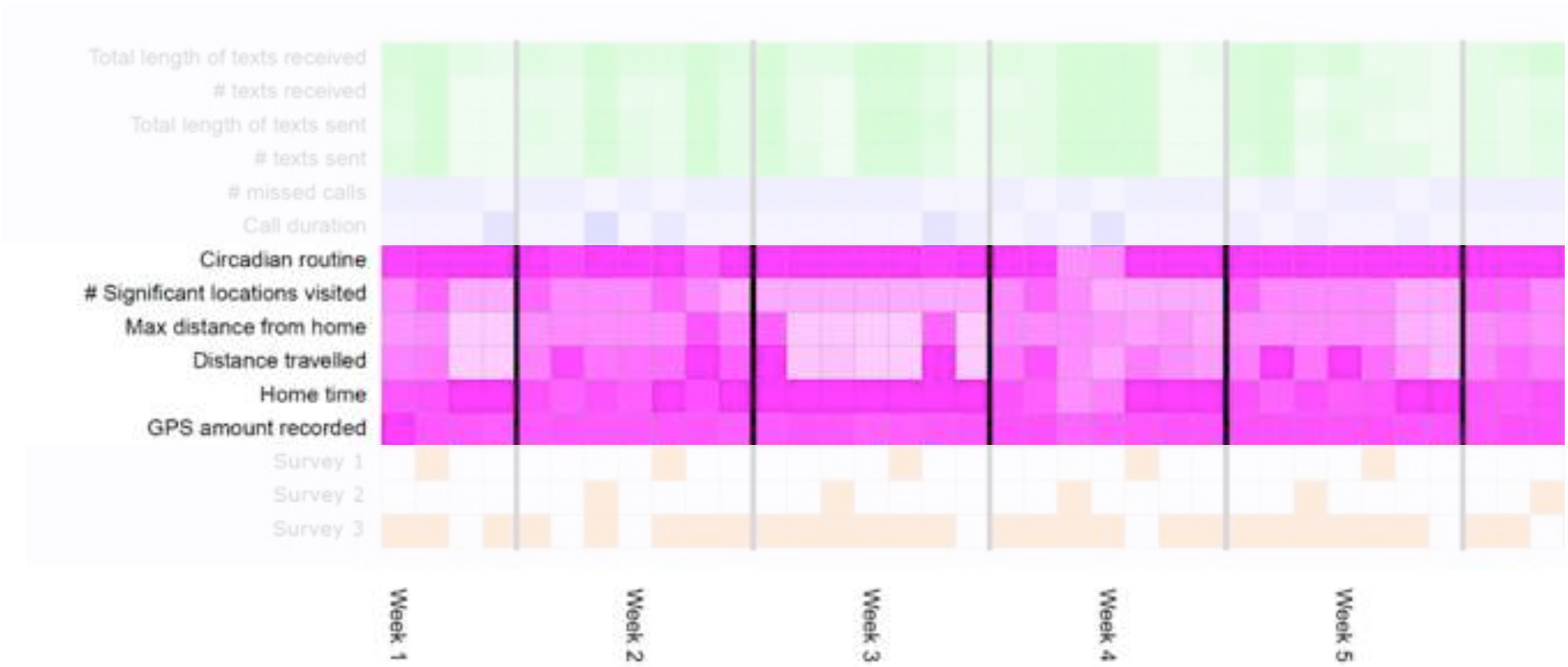

- One-question survey: Please rate your pain over the last 24 hours on a scale of 0 (no pain) to 10 (worst)?

# UNDERSTANDING PAIN



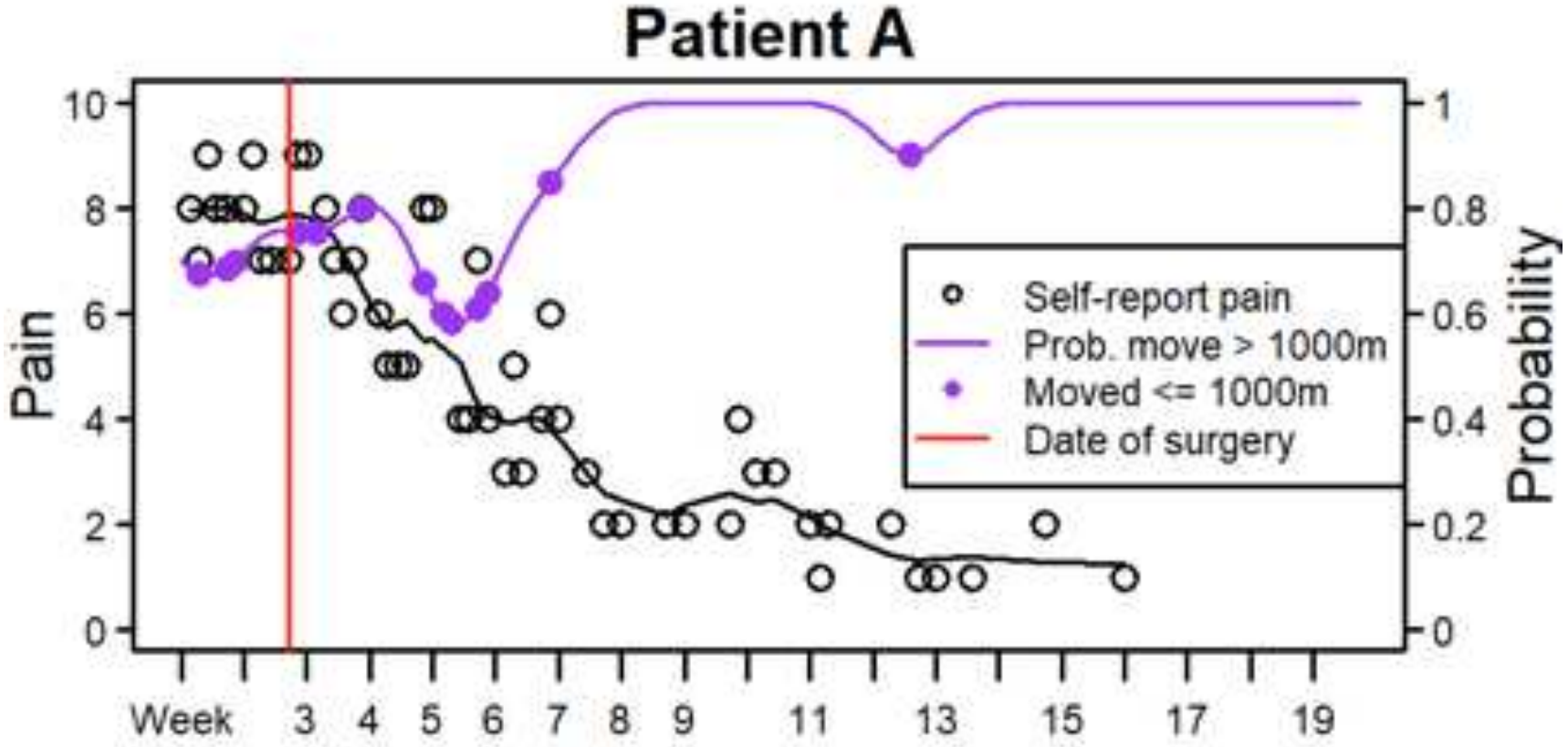

- Survey response rate on average: 43%

# UNDERSTANDING PAIN



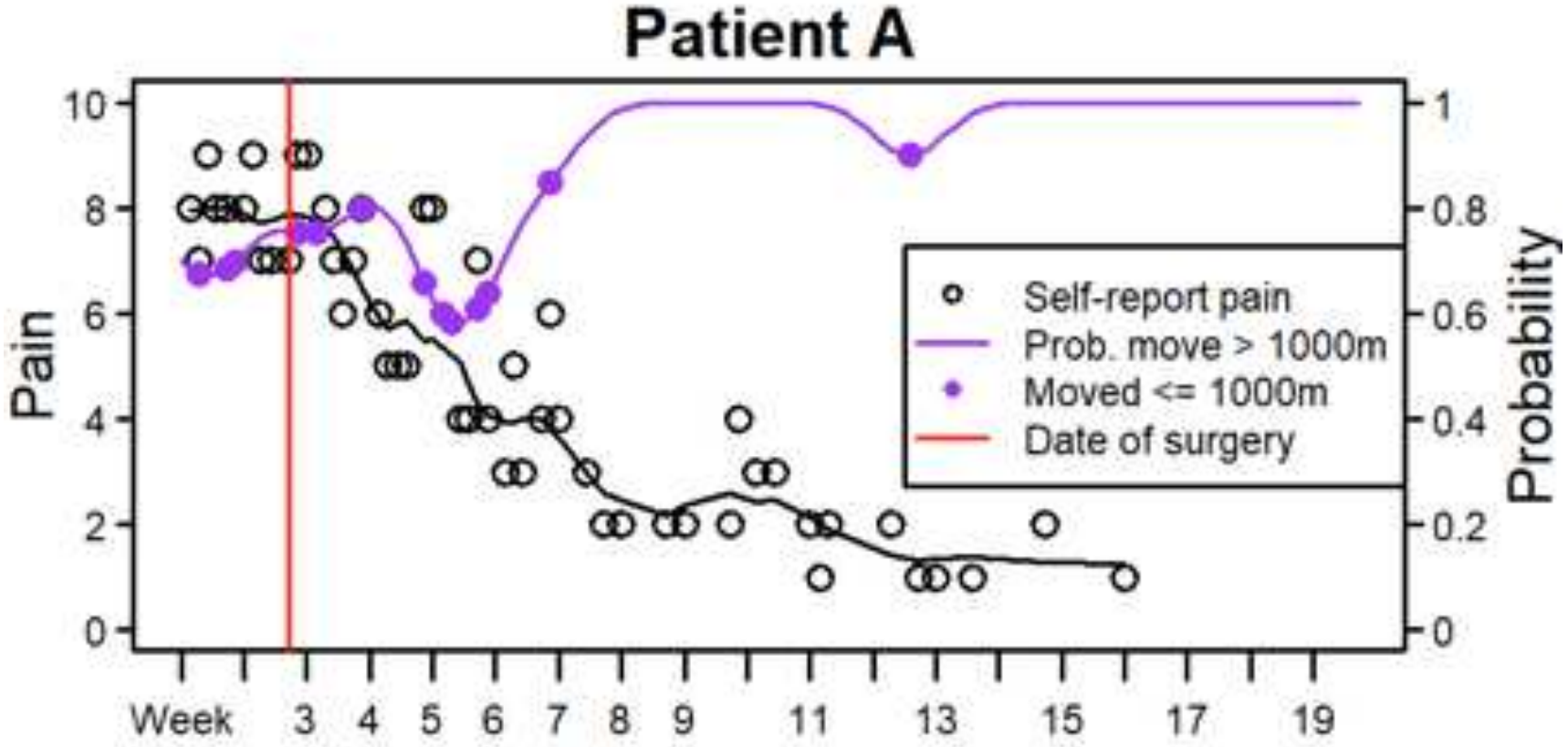

- Amount of daily GPS data available: 93%

# UNDERSTANDING PAIN




- Circles are survey responses. Purple line is the probability of moving more than 1km for that day.

# UNDERSTANDING PAIN




- GPS is inversely correlated with self-reported pain and has substantially higher data availability.

# WHO IS IT FOR?

- Researchers interested in human behavior that could potentially be collected through smartphones
- Projects designed for long-term data collection
- Research questions that need “good enough” data (vs gold-standard data)

# THANK YOU

Please see *beiwe.org* for more information

Mathew Kiang[1]
@mathewkiang

Patrick Staples[2]
@ctphoenix

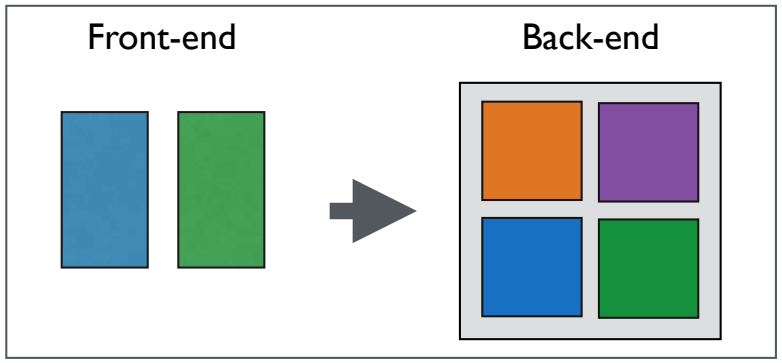
John Torous[3]
@JohnTorousMD

Monica Alexander[4]
@monjalexander

Jukka-Pekka "JP" Onnela[5]
@jponnela

# BEIWE RESEARCH PLATFORM

- **Front-end:** Android and iOS applications
- **Back-end:** Amazon Web Services (AWS) cloud computing infrastructure
  - Scalability is key: about 1 GB of **raw data** per subject-month
  - **Study portal** for managing studies and accessing data
  - **Data collection** using AWS EC2 instances and Elastic Beanstalk
  - **Data storage** using AWS S3 for medium-term and Amazon Glacier for long-term
  - **Data analysis pipeline** for data processing and computation of data summaries
- **Open source under the 3-clause BSD license (https://github.com/onnela-lab)**



# BEIWE DATA STREAMS

## Active data

- Survey responses and metadata
- Voice / audio recordings
- Cognitive tests (under development)

## Passive data

- GPS
- Accelerometer
- Gyroscope
- Phone and screen state
- WiFi routers
- Bluetooth devices
- Phone call logs
- Text message logs
- Magnetometer
- Proximity

---

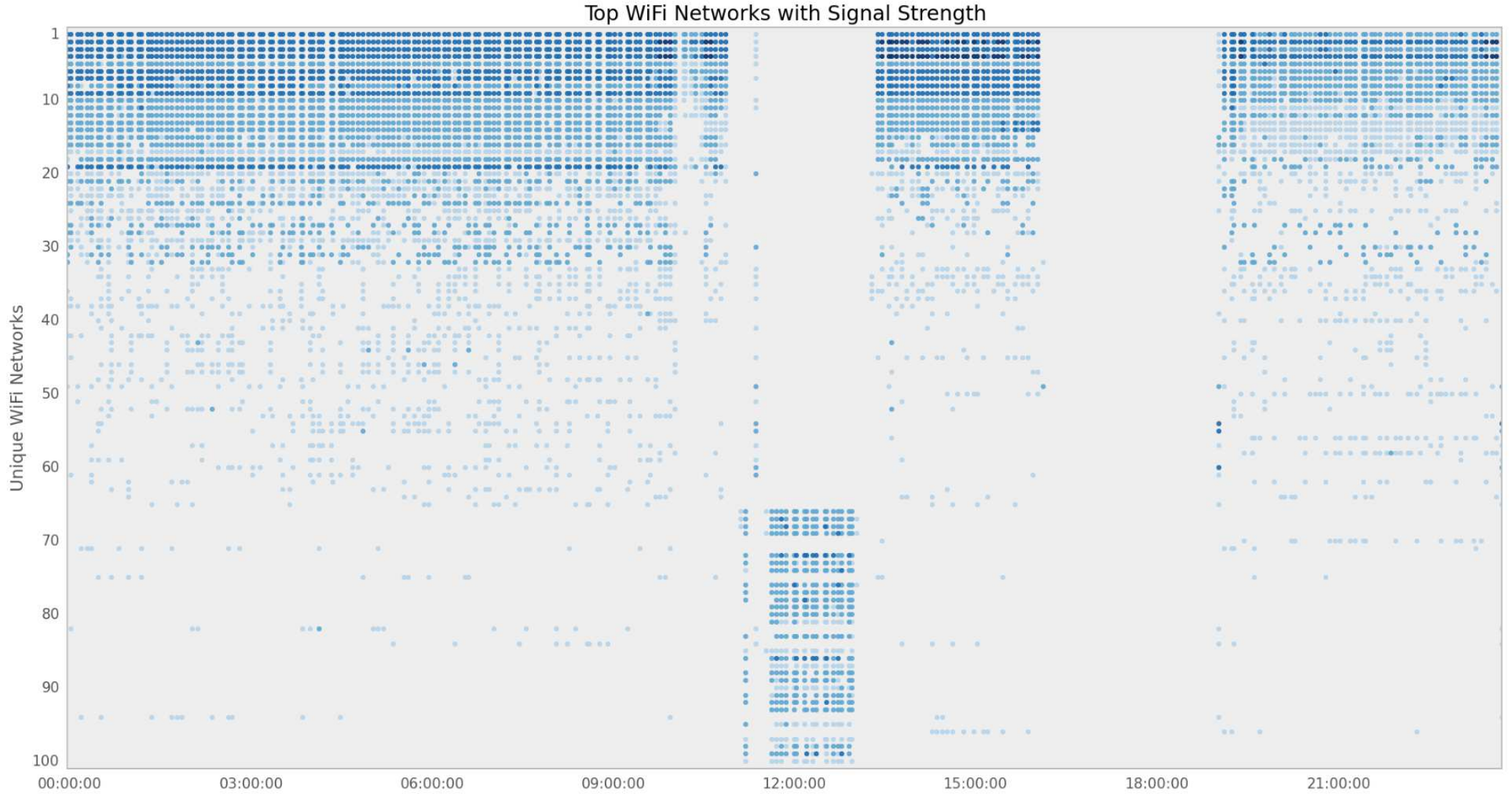
## Data flow

- Hash and encrypt, store locally, forward to server when data connection (WiFi / 4G)

# PRIVACY

- Password protection
- Data buffered on the device until next Wi-Fi network available
- Beiwe does not store any directly identifying information
- Participants are assigned randomly generated IDs (e.g., DT4TIY86)
- Potential identifiers, like phone numbers, in collected data are hashed using industry-standard SHA-256 hashing algorithm
- **All data on phones, on the server, and in-transit use industry-standard encryption techniques**
- **Beiwe uses asymmetric encryption** (cannot read its own data)
- Public 2048 bit RSA encryption key, which is then used to encrypt a symmetric AES (Advanced Encryption Standard) key for bulk encryption (generated as needed by the app)

# WIFI DATA


Top WiFi Networks with Signal Strength
Unique WiFi Networks
1
10
20
30
40
50
60
70
80
90
100
00:00:00
03:00:00
06:00:00
09:00:00
12:00:00
15:00:00
18:00:00
21:00:00