SAPTrees: Using Conditional Inference Trees to Characterize Heterogeneity in Human Activity Patterns

Roland Brown

Division of Biostatistics, University of Minnesota

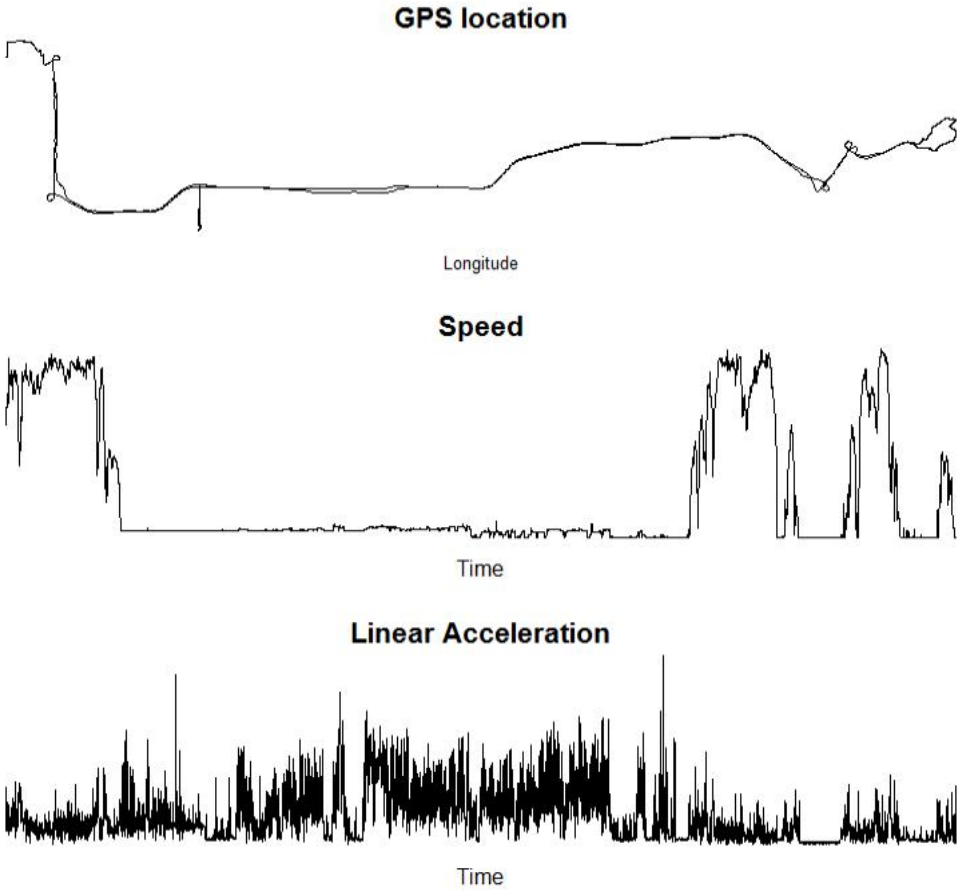
International Conference on Health Policy Statistics

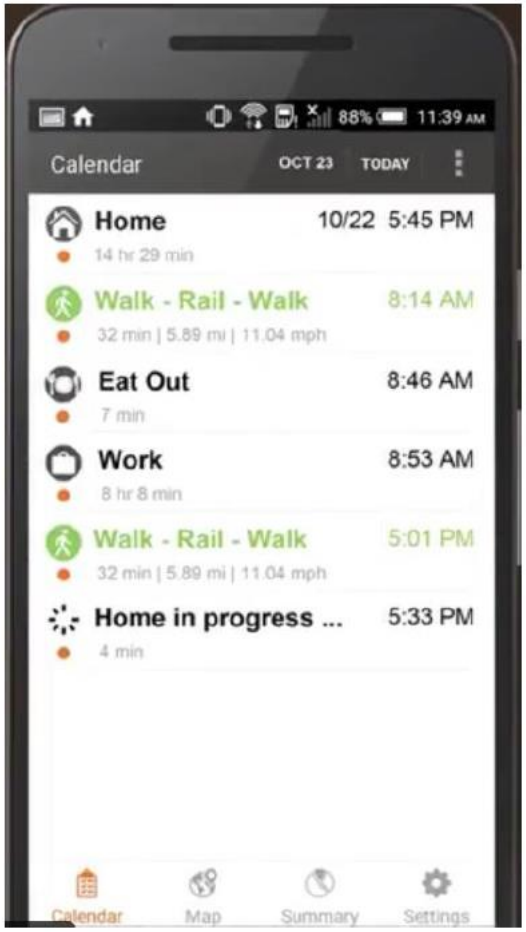
January 8, 2020

Talk Outline

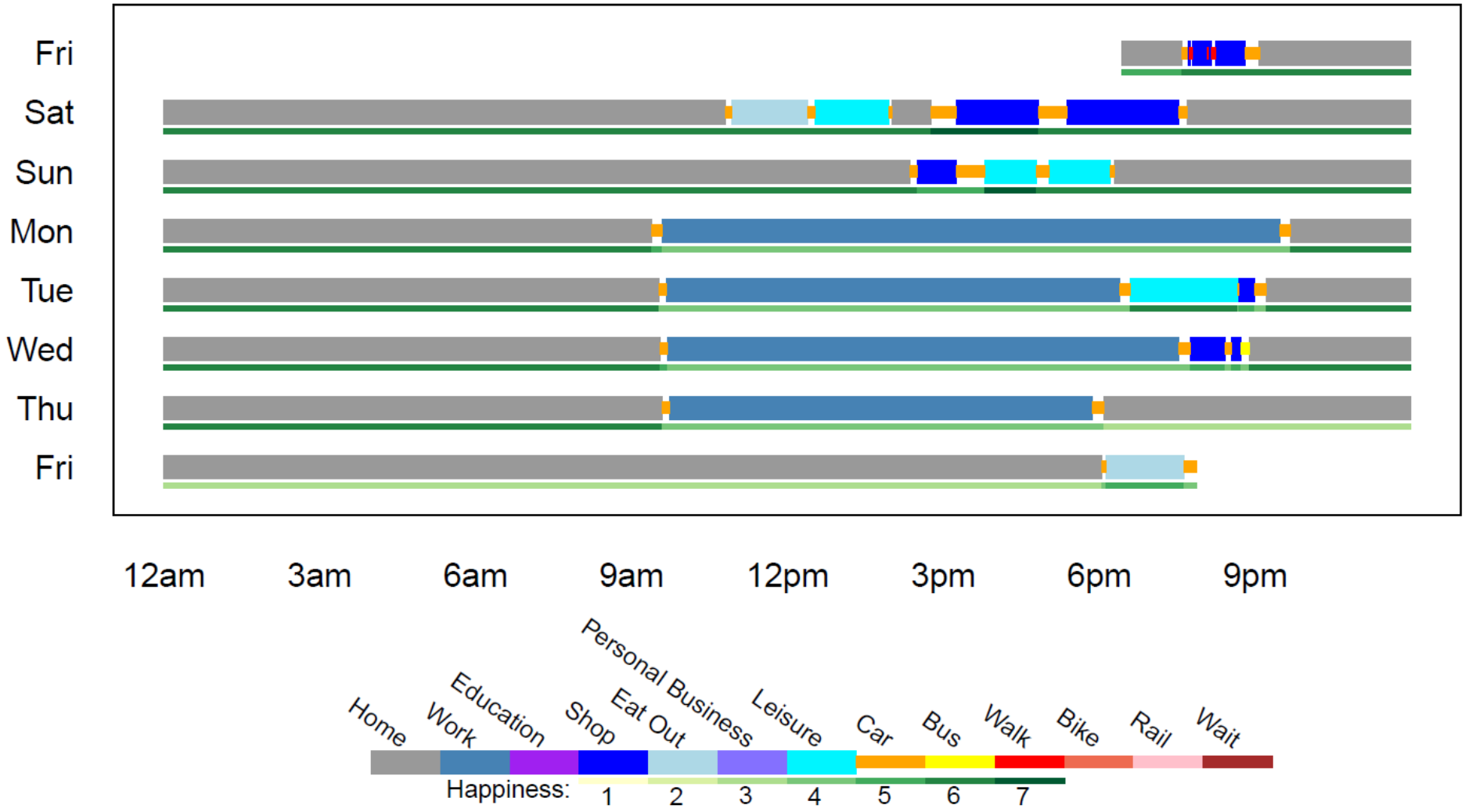
- Smartphone sensor data and Daynamica
- Smartphone data as a character sequence
- Sequential Activity Pattern Trees (SAPTrees)
  - Defining sequence distances
  - Multivariate distance matrix regression (MDMR)
  - Conditional inference trees (CTrees)
  - SAPTree Algorithm
- Application to smartphone data

Daynamica


GPS location
Longitude
Speed
Time
Linear Acceleration
Time


88% 11:39 AM
Calendar
OCT 23
TODAY
Home
10/22 5:45 PM
14 hr 29 min
Walk - Rail - Walk
8:14 AM
32 min | 5.89 mi | 11.04 mph
Eat Out
8:46 AM
7 min
Work
8:53 AM
8 hr 8 min
Walk - Rail - Walk
5:01 PM
32 min | 5.89 mi | 11.04 mph
Home in progress ...
5:33 PM
4 min
Calendar
Map
Summary
Settings

Daynamica Data

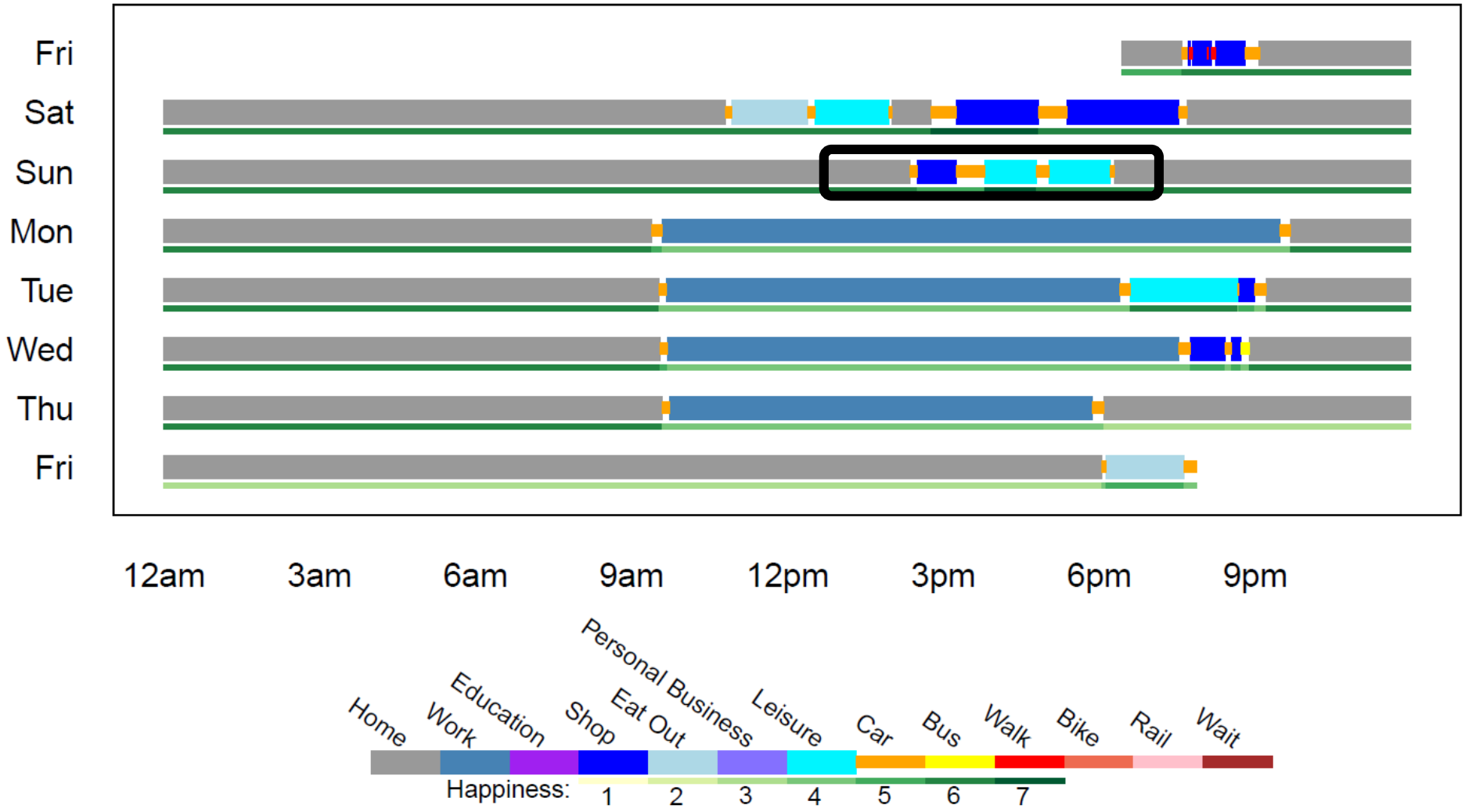


Daynamica Data

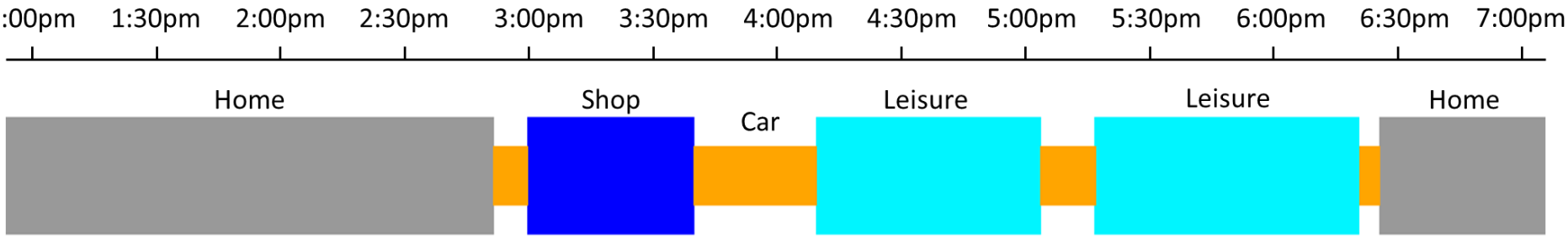
Day	Start Time	End Time	Event
1	00:00:00	08:34:35	HOME
1	08:34:35	08:41:29	CAR
1	08:41:29	17:00:05	WORK
1	17:00:05	17:10:48	CAR
1	17:10:48	17:49:15	PERSONAL_BUSINESS
1	17:49:15	18:04:17	CAR
1	18:04:17	18:37:22	EAT_OUT
1	18:37:22	18:55:22	CAR
1	18:55:22	23:59:59	HOME
2	00:00:00	09:16:32	HOME
2	09:16:32	09:23:37	CAR
2	09:23:37	17:12:08	WORK

Goal: Formally characterize heterogeneity in user's activity patterns

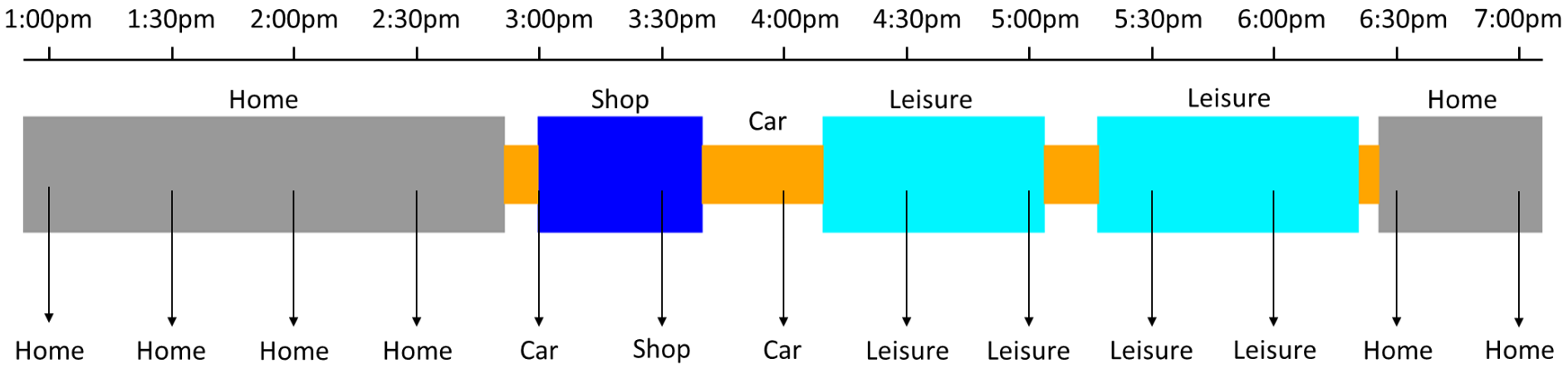
Daynamica Data



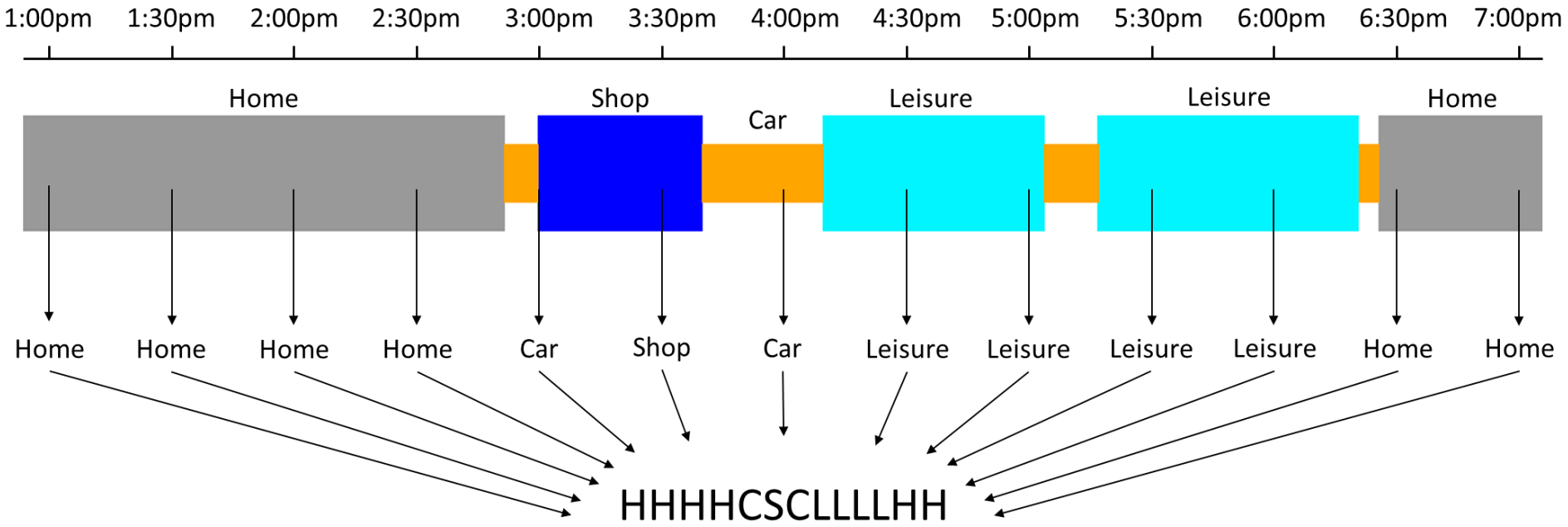
Sequences from Daynamica Data



Sequences from Daynamica Data



Sequences from Daynamica Data



Sequence-Based Edit Distances

- Common in bioinformatics, computer science
- Minimum total cost of changing one sequence to another via series of operations:
  - Substitution
  - Insertion
  - Deletion
- Specified via a substitution cost matrix

Sequence-Based Edit Distances

$$\Gamma = \begin{array}{c} \\ a \\ b \\ 1 \\ 2 \\ \phi \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & 1 & 2 & \phi \end{array} \\ \begin{bmatrix} \gamma_{aa} & \gamma_{ab} & \gamma_{a1} & \gamma_{a2} & \gamma_{a\phi} \\ \gamma_{ba} & \gamma_{bb} & \gamma_{b1} & \gamma_{b2} & \gamma_{b\phi} \\ \gamma_{1a} & \gamma_{1b} & \gamma_{11} & \gamma_{12} & \gamma_{1\phi} \\ \gamma_{2a} & \gamma_{2b} & \gamma_{21} & \gamma_{22} & \gamma_{2\phi} \\ \gamma_{\phi a} & \gamma_{\phi b} & \gamma_{\phi 1} & \gamma_{\phi 2} & \gamma_{\phi\phi} \end{bmatrix} \end{array}$$

- Cost matrix for sequences with states (a,b,1,2)
- ϕ is “null” state, responsible for indels

Sequence-Based Edit Distances

$$\boldsymbol{\Gamma} = \begin{array}{c} \\ a \\ b \\ 1 \\ 2 \\ \phi \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & 1 & 2 & \phi \end{array} \\ \begin{bmatrix} \gamma_{aa} & \gamma_{ab} & \gamma_{a1} & \gamma_{a2} & \gamma_{a\phi} \\ \gamma_{ba} & \gamma_{bb} & \gamma_{b1} & \gamma_{b2} & \gamma_{b\phi} \\ \gamma_{1a} & \gamma_{1b} & \gamma_{11} & \gamma_{12} & \gamma_{1\phi} \\ \gamma_{2a} & \gamma_{2b} & \gamma_{21} & \gamma_{22} & \gamma_{2\phi} \\ \gamma_{\phi a} & \gamma_{\phi b} & \gamma_{\phi 1} & \gamma_{\phi 2} & \gamma_{\phi\phi} \end{bmatrix} \end{array}$$

- Levenshtein edit distance: all diagonal entries 0, all other entries 1
- Distance between ‘b12a’ and ‘a22’ is 3:
  1. Delete ‘a’ → ‘b12’
  2. Substitute ‘1’ with ‘2’ → ‘b22’
  3. Substitute ‘b’ with ‘a’ → ‘a22’

Pairwise Sequence Distance Matrix

$$\boldsymbol{D} = \{d_{ij}\} = \begin{bmatrix} 0 & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} \\ & 0 & d_{23} & d_{24} & d_{25} & d_{26} \\ & & 0 & d_{34} & d_{35} & d_{36} \\ & & & 0 & d_{45} & d_{46} \\ & & & & 0 & d_{56} \\ & & & & & 0 \end{bmatrix}$$

6x6 distance matrix summarizing dissimilarity between 6 sequences

Distance-Based Methods

- Heatmaps
- Dendrograms
- Hierarchical and other clustering methods
- Distance-based regression (distances as predictors)

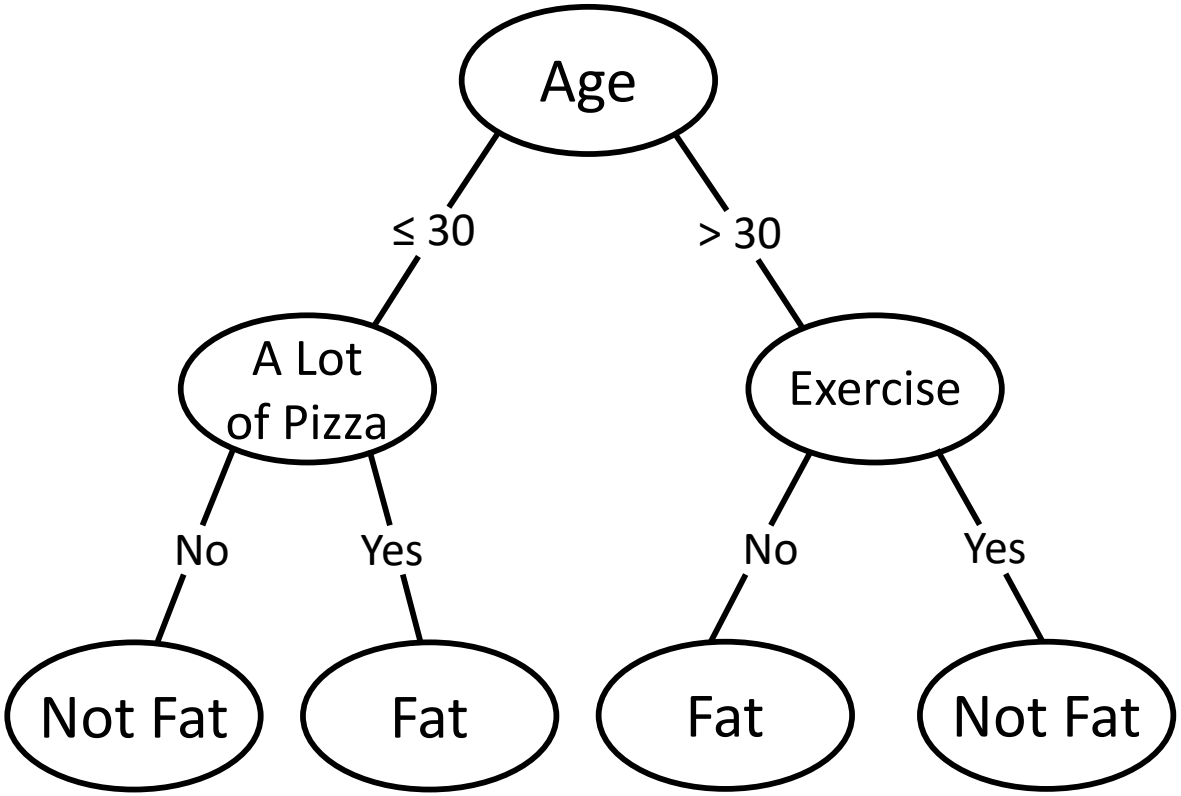
Distance Matrix as Response

- Relate set of covariates were to sequence distance matrix outcome
- A regression method with a distance matrix as the “response variable”
- Multivariate Distance Matrix Regression (MDMR)

MDMR: Overview

- Regress an $n \times n$ distance matrix $\boldsymbol{D}$ (outcome) on covariates $\boldsymbol{X}$
- Similar to OLS, partition sum of squared distances (SSD) instead
- MDMR p-values assess strength of evidence for "association" between covariate(s) and distance matrix
- No measure of effect size
- Interpretability of a "significant" association is lacking

Decision Tree



Conditional Inference Trees (CTrees)

Alternative to CART, primary differences:

- Variable **selection** and **splitting** distinct sequential processes
  - Variable **selection**: which **variable** to split on
  - Variable **splitting**: which **value** to split on

- Variable **selection** done in hypothesis testing framework

- Variable **splitting** determines optimal threshold for chosen variable

Sequence Activity Pattern Trees (SAPTrees)

SAPTrees use the CTree algorithm with:

- Sequences/resulting distance matrix as the outcomes
- MDMR used for the **selection** step and **splitting** step

Sequence Activity Pattern Trees (SAPTrees)

1. **Compute distance matrix from sequences.**

2. **Variable selection step.**
   - Single covariate MDMR model for each candidate covariate
   - Covariate w/ lowest p-value (also lower than threshold) chosen
   - If none lower, terminate (controls Type I error)

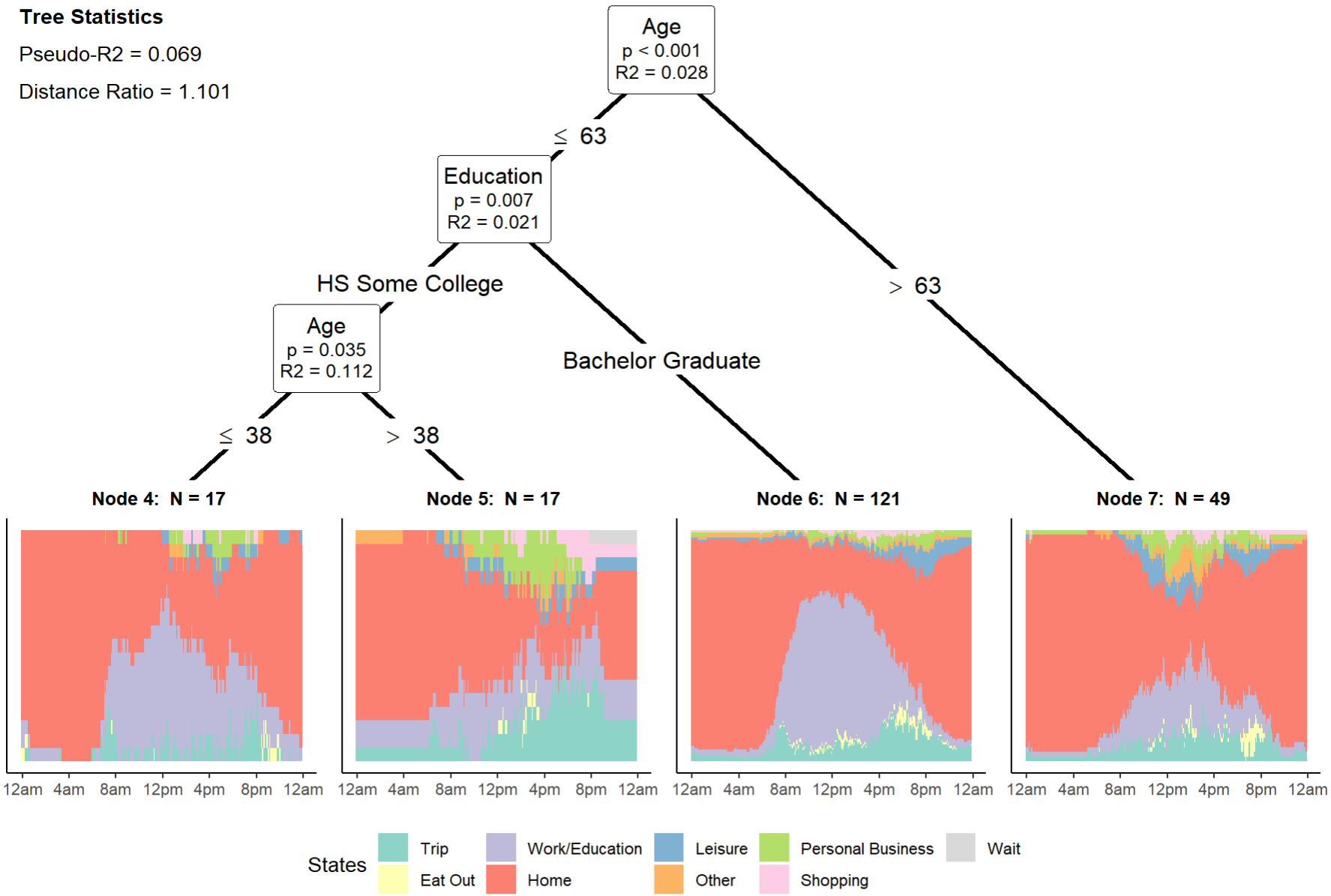
3. **Variable splitting step.**
   - MDMR at each split point used to determine optimal split value

4. **Recursion.**
   - Repeat steps (2) and (3) recursively in each node until no more splits

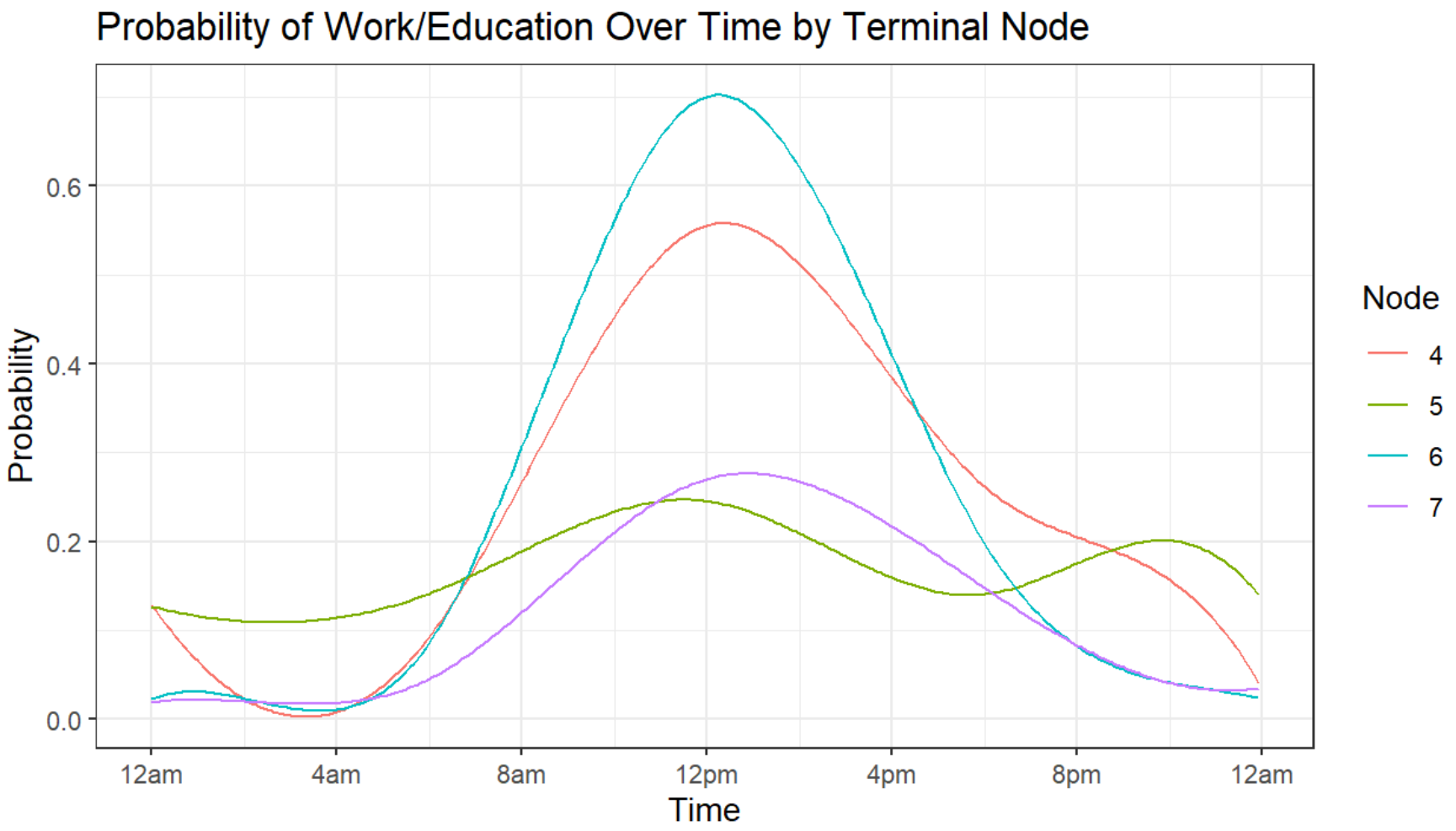
SAPTree Application

- Minneapolis-area study. ~250 users with multiple days of data.
- SAPTree fit to 24-hour Wednesday sequence data for each individual
- 5-minute resolution: 288-character sequences
- ***D*** calculated using Levenshtein distance
- Covariates:
  - Gender (male vs. female)
  - Age (quantitative)
  - Education (HS, associate/some college, bachelor, graduate)
  - Race (white vs. other)
  - Income (<25k, 25-50k, 50-75k, 75-100k, >100k)
  - Parental Status (children vs. no children)
- Multiplicity adjustment: Bonferroni
- Nominal Type I error rate at each split point: $\alpha=0.05$

SAPTree Visualization

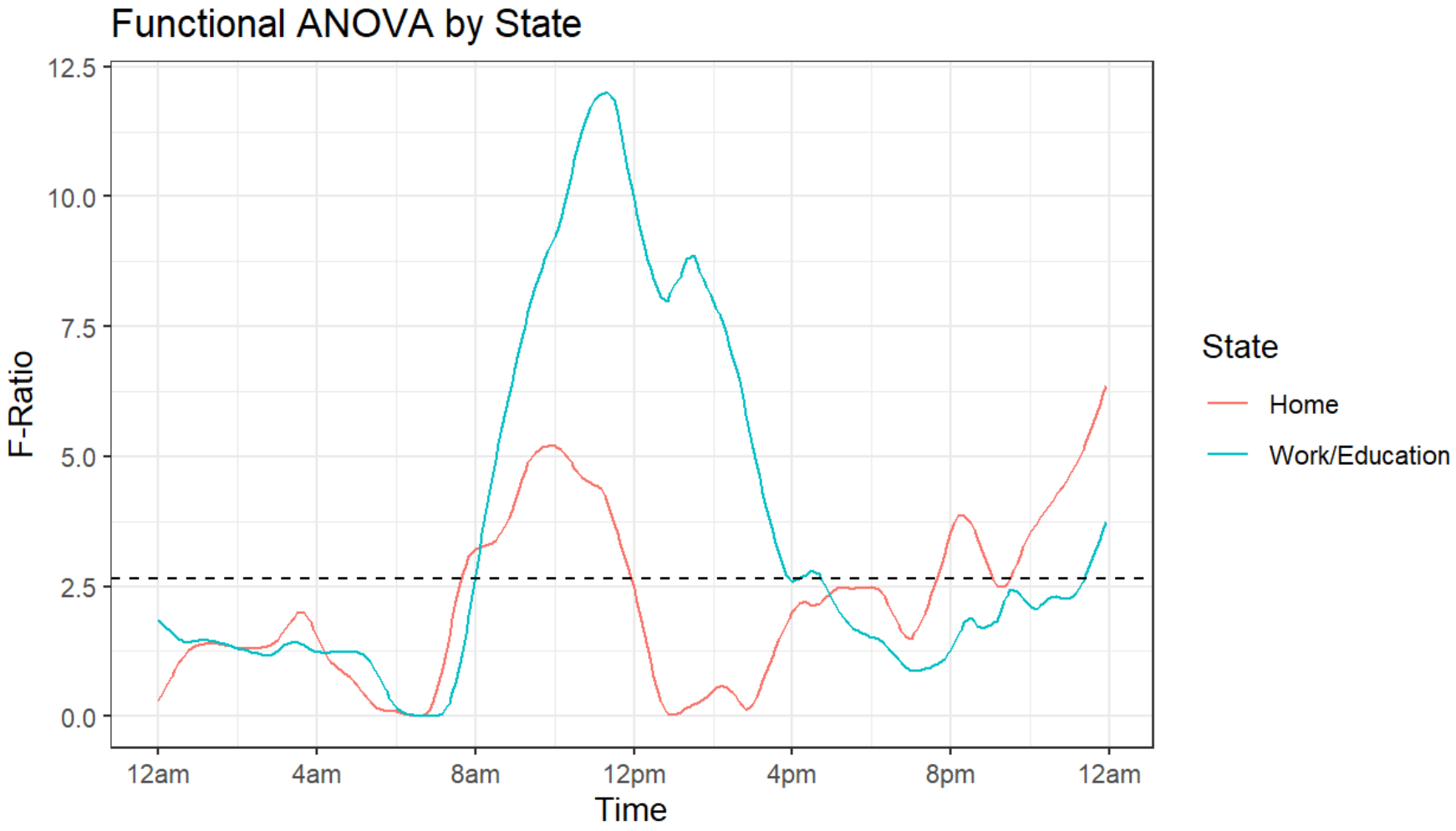


Functional Data Summaries



Node-specific probabilities of **Work** over time using function-on-scalar regression

Functional Data Summaries



Pointwise ANOVA to assess at what time points do terminal nodes significantly differ with respect to a particular state

Potential SAPTree Uses

"Personalized medicine": identifying activity-based subgroups for targeted interventions

- ID'ing patients in exercise rehab programs who differentially complete prescribed outpatient exercise regimens

Healthcare cost savings

- SAPTree analysis of sensor data collected from mobile care teams could be used to streamline their time use

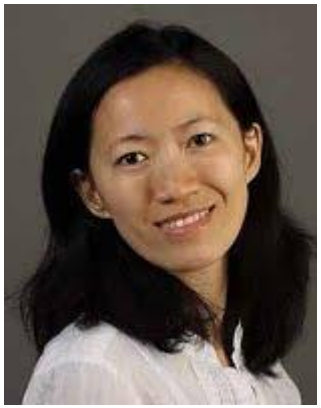
Outside of healthcare

- Transportation planning: understanding how different population subgroups utilize public transit network

Conclusions

- A method for characterizing covariate-based heterogeneity in activity sequence patterns
- Interpretable results and visualizations
- Sensitive to variety of activity pattern differences undetectable using traditional outcomes
- Many potential application areas
- Future Work:
  - Effect of sampling resolution, distance metric choice
  - Multiple sequences (e.g., adding accelerometer data)

Acknowledgements



Yingling Fan, PhD
Professor
Urban and Regional Planning
Humphrey School of Public Affairs

Julian Wolfson, PhD
Associate Professor
Biostatistics

Chen-Fu Liao, MS, PhD
Senior Research Associate
Mechanical Engineering
App Developer

Kirti Das
PhD Student, Public Affairs
Project Manager &
Research Assistant

Andy Becker
PhD Student &
Research Assistant
Biostatistics

Yin Song, PhD
Assistant Professor
Geography

Siyang Ren
MS Student &
Research Assistant
Biostatistics

Yaxuan Zhang
MS Student &
Research Assistant
Geography

Past Members

Gediminas Adomavicius, PhD
Professor
Information/Decision Sciences
Carlson School of Management

Akshay Kulkarni
M.S., Data Science
Data Scientist
(as of 05/17: Google)

Feng Liu
M.S., Computer Science
App Programmer
(as of 05/17: Amazon)

Yash Khandelwal
M.S., Computer Science
App Programmer
(as of 9/15: Epic Systems)

Jie Kang
M.S., Computer Science
App Programmer
(as of 9/16: Facebook)

Wei Ni
M.S., Computer Science
App Programmer

Questions?