Agenda

1 **Introductions**
Who are you?

2 **Syllabus Walkthrough**
What will we be doing all quarter?

3 **Papers**
Why are we reading these?

4 **Weekly Structure**
What can we expect each week?

5 **Topic Motivation**
What are travel demand models and why is any of this interesting?

Introductions

Deloitte FutureScape Data Science Team

Steve Hardy, Ph.D.
Managing Director

Dan Baeder
Data Scientist / Specialist Master

George Panteras, Ph.D.
Data Scientist / Specialist Master

Corey Ducharme, Ph.D.
Data Scientist / Specialist Senior

We use data science to build digital replicas of entire cities

Modeling and simulation is a common technique to analyze hypothetical situations or new technologies. Historically, there is a tradeoff between detail, scale, and speed. Our team builds algorithms and software to remove that limitation.

MASSIVE SCALE AGENT BASED SIMULATION

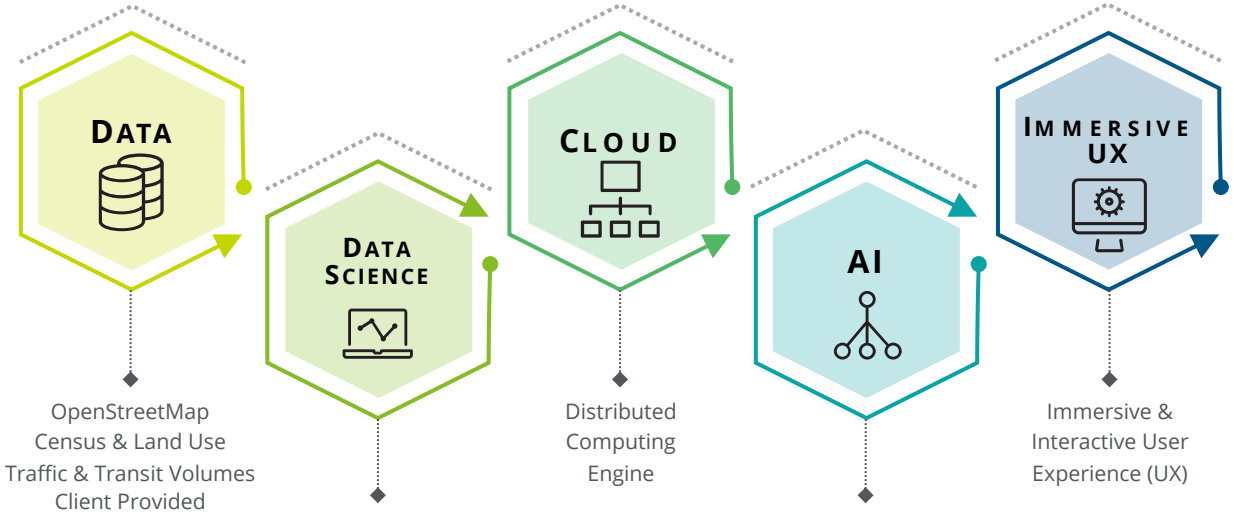
We develop and apply a cloud-distributed simulation platform that allows researchers to:

- **Model an entire region at once:** Models can cover thousands of square miles, millions of people, and multiple network layers.
- **Model at high fidelity:** Simulate every person, every road, and every mode of mobility.
- **Run quickly and simultaneously:** Leverage elastic cloud computing to scale simulations horizontally (many nodes on one model) and vertically (many models at once).

Generating realistic data & what-if analyses at massive scale

Agent-based simulations, tuned to real-world data and behaviors, generate synthetic anonymized data and answer "what if" questions. With FutureScape, simulations are not limited by scale or complexity.

BUILDING BLOCKS



DATA
OpenStreetMap
Census & Land Use
Traffic & Transit Volumes
Client Provided

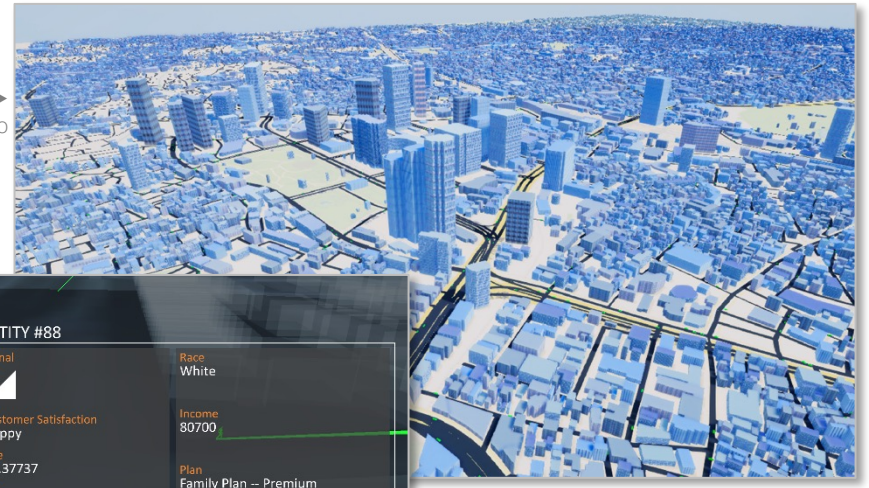
DATA SCIENCE
Synthetic Population
Trip List
Mode choice
Activity Modeling

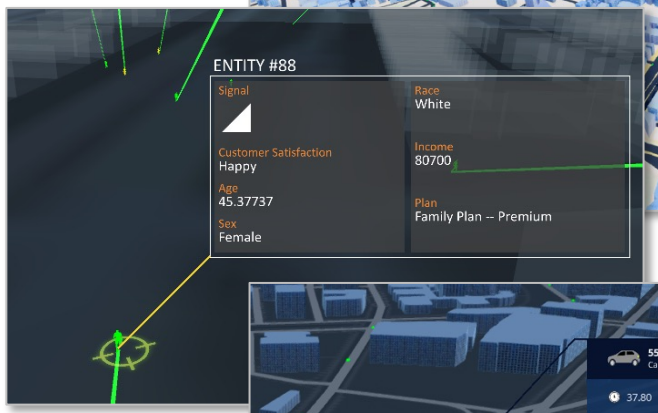
CLOUD
Distributed
Computing
Engine

AI
Custom AI-Algorithms
Driving Entity Behaviors
and Choices

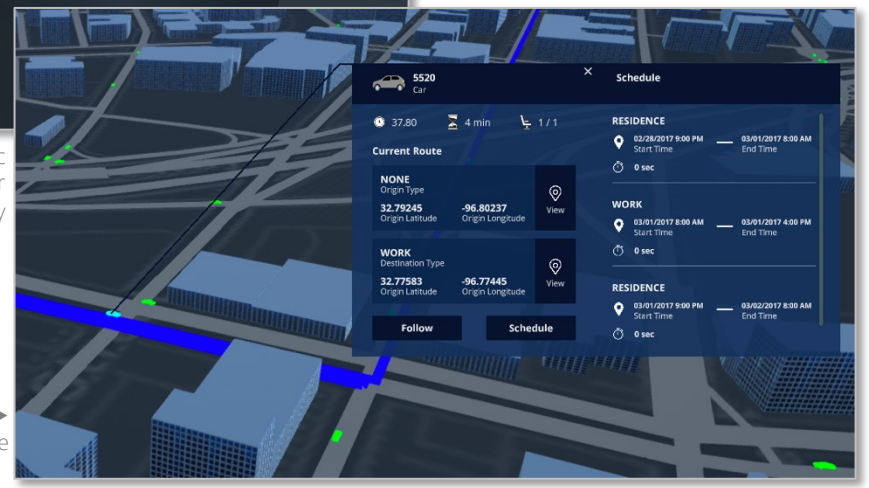
IMMERSIVE UX
Immersive &
Interactive User
Experience (UX)

Full scale sim ▶
of Tokyo





▲Agent socioeconomic
info and tower
connectivity



Agent details, route ▶
and activity schedule

Introductions

WHO ARE YOU?

- Name
- Interest in this topic
- Goals for the course
- Career aspirations

Syllabus

Weeks 1 – 5 (September 29 – October 27)

1 Intro to Travel Demand Models

Initial conversation about travel demand models to motivate the course. Discussion on the current state of the art.

2 Mode Choice Modeling

Discussion of current methods for modeling mode choice within travel demand models

3 Paper Intro / Data Preparation

Introduction to *Budholiya et al*, the paper we'll focus on for learning techniques and replicating the approach. Introduction to the Cleveland heart dataset used by the authors, along with some data cleaning approaches.

4 XGBoost

Introduction to extreme gradient boosted trees (XGBoost) including advantages of the technique and the math behind the method.

CHECKPOINT 1 DUE: Clean and prepare the heart dataset for modeling

5 XGBoost Classifier Construction

Overview of how to construct an XGBoost classifier using the xgboost Python package and the methods described in *Budholiya et al.*

Syllabus

Weeks 6 – 10 (November 3 – December 1)

6 Feature Selection

Learn how to use XGBoost model output to understand feature importance and craft a simplified model with a reduced feature set.

CHECKPOINT 2 DUE: Discuss how you will reduce the number of features in your classification model

7 Analyzing Model Results

Following the approach in *Budholiya et al*, learn how to assess the quality of your classification model

8 Hyperparameter Optimization

Overview of Bayesian optimization for hyperparameter tuning of your XGBoost model, in order to improve model performance.

9 Ethics of Mobility Data Collection

Discussion on the ethics of collecting and using mobility data to inform better travel demand models. Even when anonymized, this data can be revealing as *de Montjoye et al* show.

10 Final Presentations

Culmination of Q1 – a tuned XGBoost classification model using the Cleveland heart dataset.

Massive-Scale Models of Urban Infrastructure and Populations

Baeder et al.

Abstract

As the world becomes more dense, connected, and complex, it is increasingly difficult to answer "what-if" questions about our cities and populations. Most modeling and simulation tools struggle with scale and connectivity. We present a new method for creating digital twin simulations of city infrastructure and populations from open source and commercial data.

Relevance

The next advancement in travel demand modeling may be in the form of agent-based models that use more precise or granular input datasets. This paper showcases one approach for doing just that.

Massive-Scale Models of Urban Infrastructure and Populations

Daniel Baeder, Eric Christensen, Anhvinh Doanvo, Andrew Han, Ben F. M. Intoy, Steven Hardy, Zachary Humayun, Melissa Kain, Kevin Liberman, Adrian Myers, Meera Patel, William J. Porter III, Lenny Ramos, Michelle Shen, Lance Sparks, Allan Toriel, and Benjamin Wu

Deloitte Consulting LLP, Arlington, Virginia, USA
bintoy@deloitte.com

Abstract. As the world becomes more dense, connected, and complex, it is increasingly difficult to answer "what-if" questions about our cities and populations. Most modeling and simulation tools struggle with scale and connectivity. We present a new method for creating digital twin simulations of city infrastructure and populations from open source and commercial data. We transform cellular location data into activity patterns for synthetic agents and use geospatial data to create the infrastructure and world in which these agents interact. We then leverage technologies and techniques intended for massive online gaming to create 1:1 scale simulations to answer these "what-if" questions about the future.

Keywords: Simulation · Urban mobility · Pattern mining

1 Introduction

There is tremendous value to studying complex systems, but these systems cannot be understood simply by analyzing their individual components [2]. If a complex system cannot be studied analytically, an alternative option is to explore the system experimentally. However, these experiments may be difficult to perform due to cost or infeasibility. An easier, more rapid, and more scalable solution would be to instead model and then simulate the system.

R. Thomson et al. (Eds.): SBP-BRiMS 2019, LNCS 11549, pp. 113–122, 2019.
https://doi.org/10.1007/978-3-030-21741-9_12

An optimized XGBoost-based diagnostic system for effective prediction of heart disease

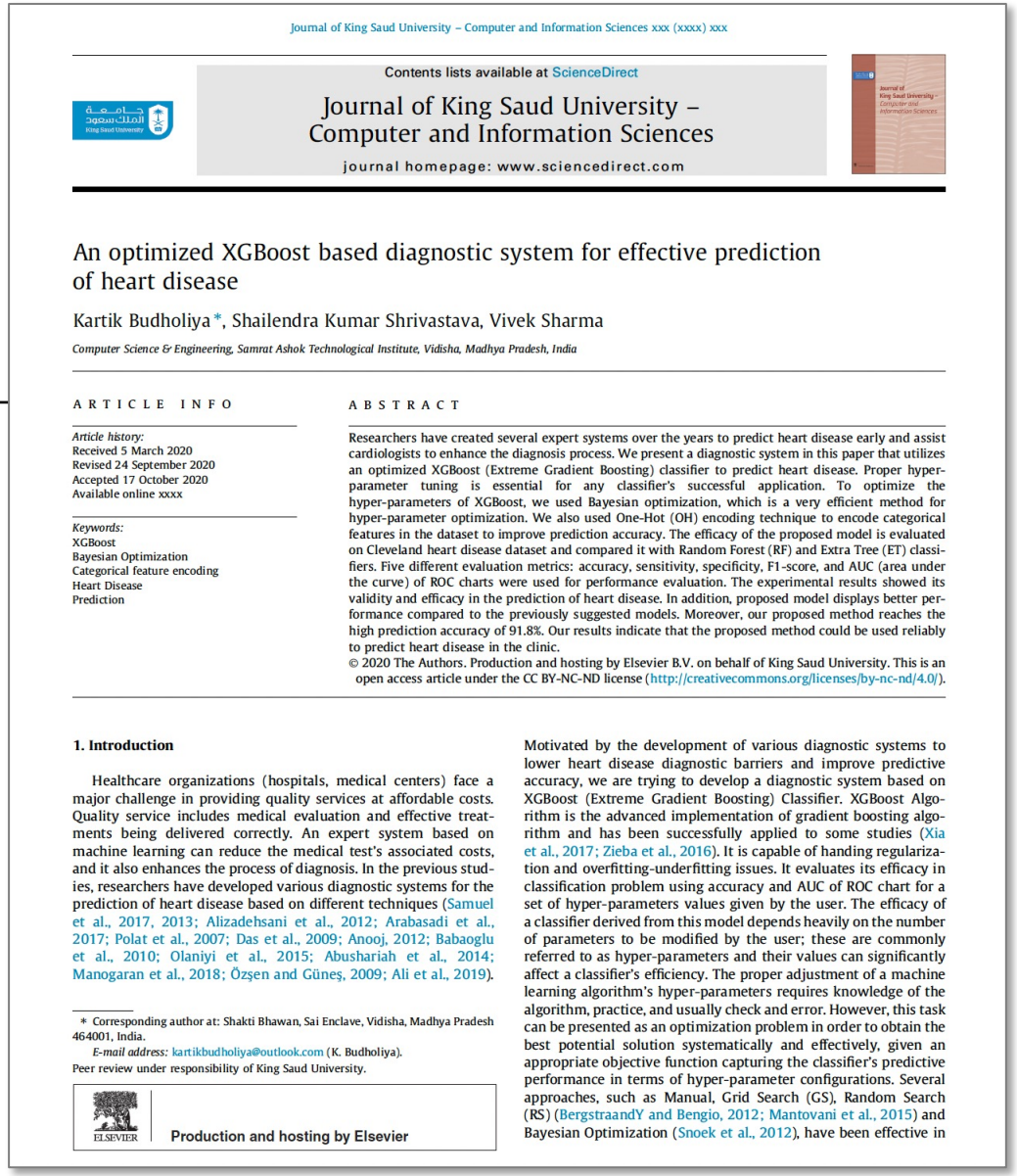
Budholiya et al.

Abstract

Researchers have created several expert systems over the years to predict heart disease early and assist cardiologists to enhance the diagnosis process. We present a diagnostic system in this paper that utilizes an optimized XGBoost (Extreme Gradient Boosting) classifier to predict heart disease.

Relevance

This paper sets forth a clear methodology for building an XGBoost classifier using a readily available dataset.
We will use this paper as our primary reference for learning techniques during the first quarter.



Journal of King Saud University – Computer and Information Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

An optimized XGBoost based diagnostic system for effective prediction of heart disease

Kartik Budholiya *, Shailendra Kumar Shrivastava, Vivek Sharma

Computer Science & Engineering, Samrat Ashok Technological Institute, Vidisha, Madhya Pradesh, India

ARTICLE INFO

Article history:
Received 5 March 2020
Revised 24 September 2020
Accepted 17 October 2020
Available online xxxx

Keywords:
XGBoost
Bayesian Optimization
Categorical feature encoding
Heart Disease
Prediction

ABSTRACT

Researchers have created several expert systems over the years to predict heart disease early and assist cardiologists to enhance the diagnosis process. We present a diagnostic system in this paper that utilizes an optimized XGBoost (Extreme Gradient Boosting) classifier to predict heart disease. Proper hyper-parameter tuning is essential for any classifier's successful application. To optimize the hyper-parameters of XGBoost, we used Bayesian optimization, which is a very efficient method for hyper-parameter optimization. We also used One-Hot (OH) encoding technique to encode categorical features in the dataset to improve prediction accuracy. The efficacy of the proposed model is evaluated on Cleveland heart disease dataset and compared it with Random Forest (RF) and Extra Tree (ET) classifiers. Five different evaluation metrics: accuracy, sensitivity, specificity, F1-score, and AUC (area under the curve) of ROC charts were used for performance evaluation. The experimental results showed its validity and efficacy in the prediction of heart disease. In addition, proposed model displays better performance compared to the previously suggested models. Moreover, our proposed method reaches the high prediction accuracy of 91.8%. Our results indicate that the proposed method could be used reliably to predict heart disease in the clinic.

1. Introduction

Healthcare organizations (hospitals, medical centers) face a major challenge in providing quality services at affordable costs. Quality service includes medical evaluation and effective treatments being delivered correctly. An expert system based on machine learning can reduce the medical test's associated costs, and it also enhances the process of diagnosis. In the previous studies, researchers have developed various diagnostic systems for the prediction of heart disease based on different techniques (Samuel et al., 2017, 2013; Alizadehsani et al., 2012; Arabasadi et al., 2017; Polat et al., 2007; Das et al., 2009; Anooj, 2012; Babaoglu et al., 2010; Olaniyi et al., 2015; Abushariah et al., 2014; Manogaran et al., 2018; Özşen and Güneş, 2009; Ali et al., 2019).

Motivated by the development of various diagnostic systems to lower heart disease diagnostic barriers and improve predictive accuracy, we are trying to develop a diagnostic system based on XGBoost (Extreme Gradient Boosting) Classifier. XGBoost Algorithm is the advanced implementation of gradient boosting algorithm and has been successfully applied to some studies (Xia et al., 2017; Zieba et al., 2016). It is capable of handing regularization and overfitting-underfitting issues. It evaluates its efficacy in classification problem using accuracy and AUC of ROC chart for a set of hyper-parameters values given by the user. The efficacy of a classifier derived from this model depends heavily on the number of parameters to be modified by the user; these are commonly referred to as hyper-parameters and their values can significantly affect a classifier's efficiency. The proper adjustment of a machine learning algorithm's hyper-parameters requires knowledge of the algorithm, practice, and usually check and error. However, this task can be presented as an optimization problem in order to obtain the best potential solution systematically and effectively, given an appropriate objective function capturing the classifier's predictive performance in terms of hyper-parameter configurations. Several approaches, such as Manual, Grid Search (GS), Random Search (RS) (BergstraandY and Bengio, 2012; Mantovani et al., 2015) and Bayesian Optimization (Snoek et al., 2012), have been effective in

* Corresponding author at: Shakti Bhawan, Sai Enclave, Vidisha, Madhya Pradesh 464001, India.
E-mail address: kartikbudholiya@outlook.com (K. Budholiya).

Peer review under responsibility of King Saud University.

ELSEVIER Production and hosting by Elsevier

Unique in the Crowd: The privacy bounds of human mobility

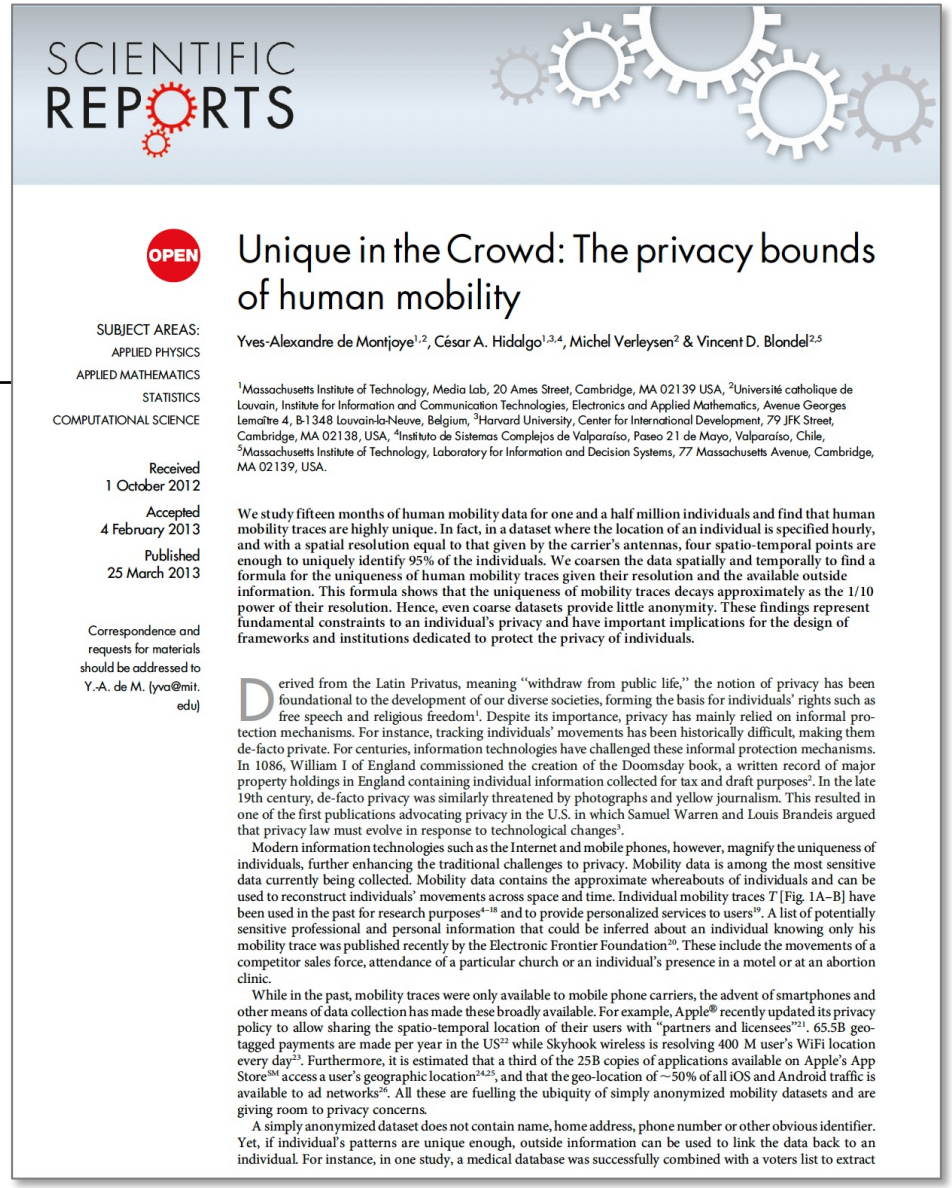
de Montjoye et al.

Abstract

We study fifteen months of human mobility data for one and a half million individuals and find that human mobility traces are highly unique. In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals.

Relevance

Mobility data can serve as a critical input to travel demand models in the future, increasing the realism of the models and providing better insights into the impact of infrastructure or policy changes. There are concerns though: most of this data is collected passively and anonymization may not really be possible.



SCIENTIFIC REPORTS

OPEN

Unique in the Crowd: The privacy bounds of human mobility

Yves-Alexandre de Montjoye[1,2], César A. Hidalgo[1,3,4], Michel Verleysen[2] & Vincent D. Blondel[2,5]

[1]Massachusetts Institute of Technology, Media Lab, 20 Ames Street, Cambridge, MA 02139 USA, [2]Université catholique de Louvain, Institute for Information and Communication Technologies, Electronics and Applied Mathematics, Avenue Georges Lemaître 4, B-1348 Louvain-la-Neuve, Belgium, [3]Harvard University, Center for International Development, 79 JFK Street, Cambridge, MA 02138, USA, [4]Instituto de Sistemas Complejos de Valparaíso, Paseo 21 de Mayo, Valparaíso, Chile, [5]Massachusetts Institute of Technology, Laboratory for Information and Decision Systems, 77 Massachusetts Avenue, Cambridge, MA 02139, USA.

SUBJECT AREAS:
APPLIED PHYSICS
APPLIED MATHEMATICS
STATISTICS
COMPUTATIONAL SCIENCE

Received
1 October 2012

Accepted
4 February 2013

Published
25 March 2013

Correspondence and requests for materials should be addressed to Y.-A. de M. (yva@mit.edu)

We study fifteen months of human mobility data for one and a half million individuals and find that human mobility traces are highly unique. In fact, in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. We coarsen the data spatially and temporally to find a formula for the uniqueness of human mobility traces given their resolution and the available outside information. This formula shows that the uniqueness of mobility traces decays approximately as the 1/10 power of their resolution. Hence, even coarse datasets provide little anonymity. These findings represent fundamental constraints to an individual's privacy and have important implications for the design of frameworks and institutions dedicated to protect the privacy of individuals.

Derived from the Latin Privatus, meaning "withdraw from public life," the notion of privacy has been foundational to the development of our diverse societies, forming the basis for individuals' rights such as free speech and religious freedom[1]. Despite its importance, privacy has mainly relied on informal protection mechanisms. For instance, tracking individuals' movements has been historically difficult, making them de-facto private. For centuries, information technologies have challenged these informal protection mechanisms. In 1086, William I of England commissioned the creation of the Doomsday book, a written record of major property holdings in England containing individual information collected for tax and draft purposes[2]. In the late 19th century, de-facto privacy was similarly threatened by photographs and yellow journalism. This resulted in one of the first publications advocating privacy in the U.S. in which Samuel Warren and Louis Brandeis argued that privacy law must evolve in response to technological changes[3].

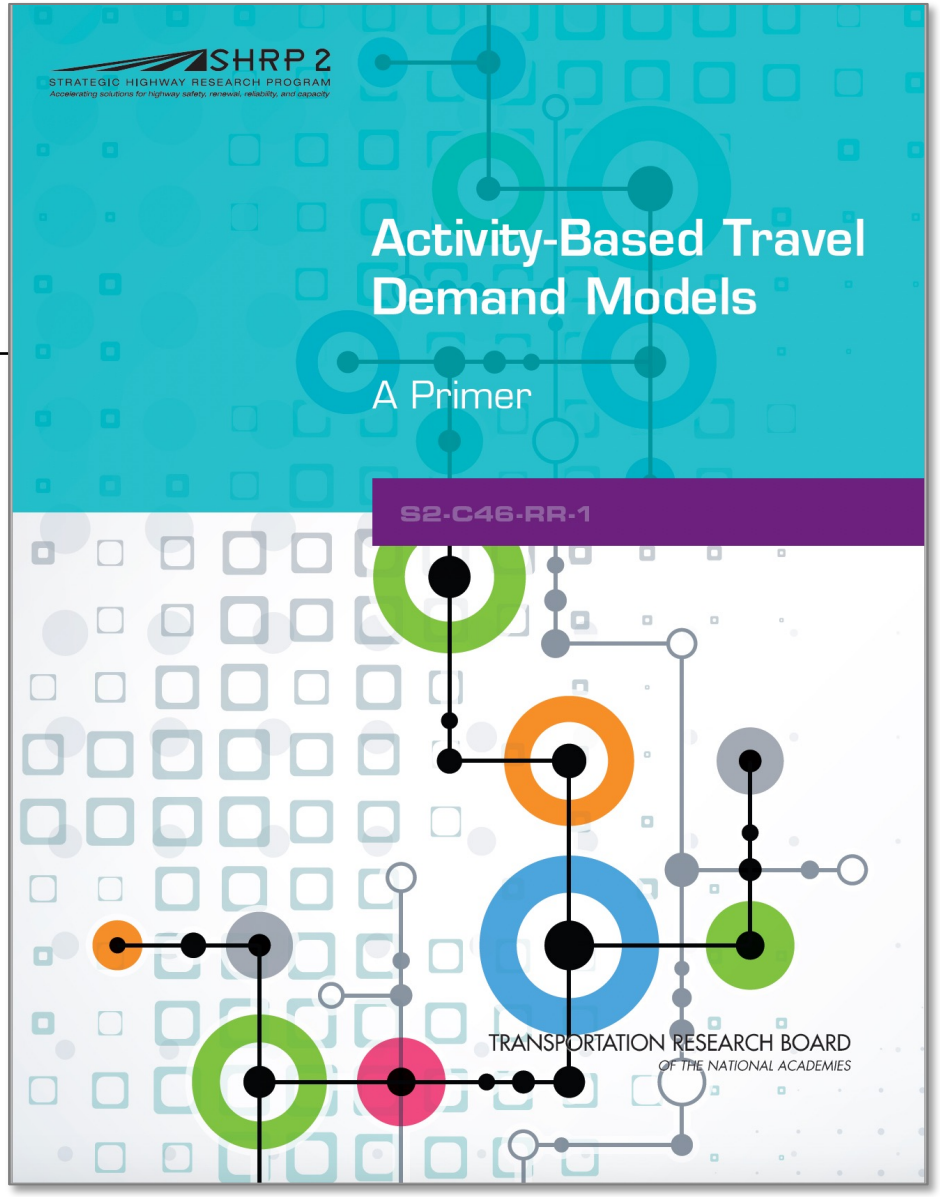
Modern information technologies such as the Internet and mobile phones, however, magnify the uniqueness of individuals, further enhancing the traditional challenges to privacy. Mobility data is among the most sensitive data currently being collected. Mobility data contains the approximate whereabouts of individuals and can be used to reconstruct individuals' movements across space and time. Individual mobility traces T [Fig. 1A–B] have been used in the past for research purposes[4–18] and to provide personalized services to users[19]. A list of potentially sensitive professional and personal information that could be inferred about an individual knowing only his mobility trace was published recently by the Electronic Frontier Foundation[20]. These include the movements of a competitor sales force, attendance of a particular church or an individual's presence in a motel or at an abortion clinic.

While in the past, mobility traces were only available to mobile phone carriers, the advent of smartphones and other means of data collection has made these broadly available. For example, Apple® recently updated its privacy policy to allow sharing the spatio-temporal location of their users with "partners and licensees"[21]. 65.5B geo-tagged payments are made per year in the US[22] while Skyhook wireless is resolving 400 M user's WiFi location every day[23]. Furthermore, it is estimated that a third of the 25B copies of applications available on Apple's App Store℠ access a user's geographic location[24,25], and that the geo-location of ~50% of all iOS and Android traffic is available to ad networks[26]. All these are fuelling the ubiquity of simply anonymized mobility datasets and are giving room to privacy concerns.

A simply anonymized dataset does not contain name, home address, phone number or other obvious identifier. Yet, if individual's patterns are unique enough, outside information can be used to link the data back to an individual. For instance, in one study, a medical database was successfully combined with a voters list to extract

Activity-Based Travel Demand Models: A Primer

Transportation Research Board of the National Academies



Abstract

Transportation decision makers confront difficult questions and must make informed choices. How will the national, regional, or even local transportation system perform 30 years into the future? A travel model is an analysis tool that provides a systematic framework for representing how travel demand changes in response to different input assumptions.

Relevance

We do a lot of work with transportation modeling on FutureScape, and our hypothesis is that mode choice selection can be simplified by identifying a reduced set of important features. We'll get to that in Q2 but understanding travel demand models is essential foundation for that later work.

Weekly Class Structure

Course delivery primarily via Zoom

Hopefully we'll get out to San Diego at least once per quarter though!

Discussion

Each class will start with a discussion of the assigned reading. Of interest are your reactions, concerns, and questions about the material. Noteworthy concepts will be called out, although we'll keep slides to a minimum.

Questions

We'll reserve time to specifically address any questions you have, whether about the reading or about code development.

Follow-Up

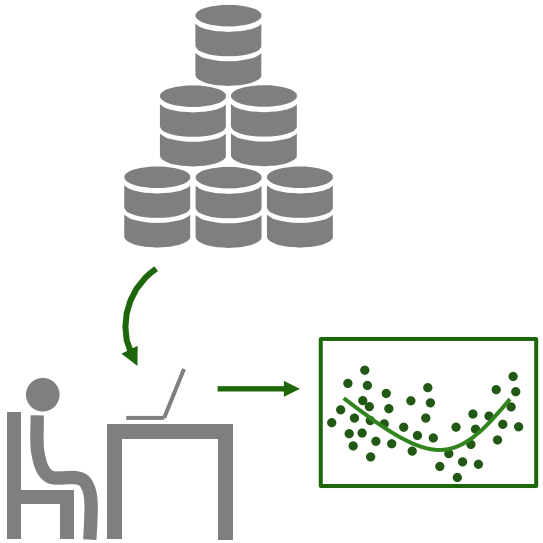
Each week we'll have office hours on **Friday from 10am – 11am.** While you can pose questions at any time, this time is set aside for direct communication.

Introduction to Travel Demand Modeling

Little data exists on things that **haven't happened yet**

Metropolitan planning organizations, departments of transportation, and other public agencies are responsible for crafting long-term plans that stretch years or decades into the future.

WITH DATA:

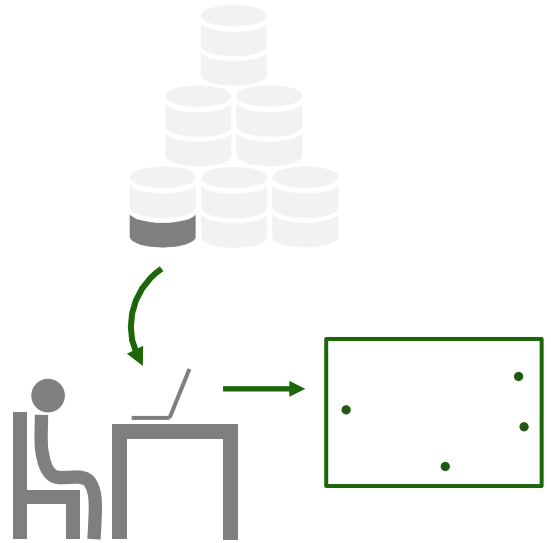


HOW CAN WE UNDERSTAND THE SIGNAL IN ALL OF THIS DATA?

BUT:

DECISIONS OFTEN HAVE TO BE MADE ON PROBLEMS FOR WHICH THERE IS LITTLE OR NO DATA

LITTLE TO NO DATA:



HOW CAN WE SOLVE PROBLEMS WITHOUT DATA TO ANALYZE?

Planners and modelers are faced with a choice

DO I…

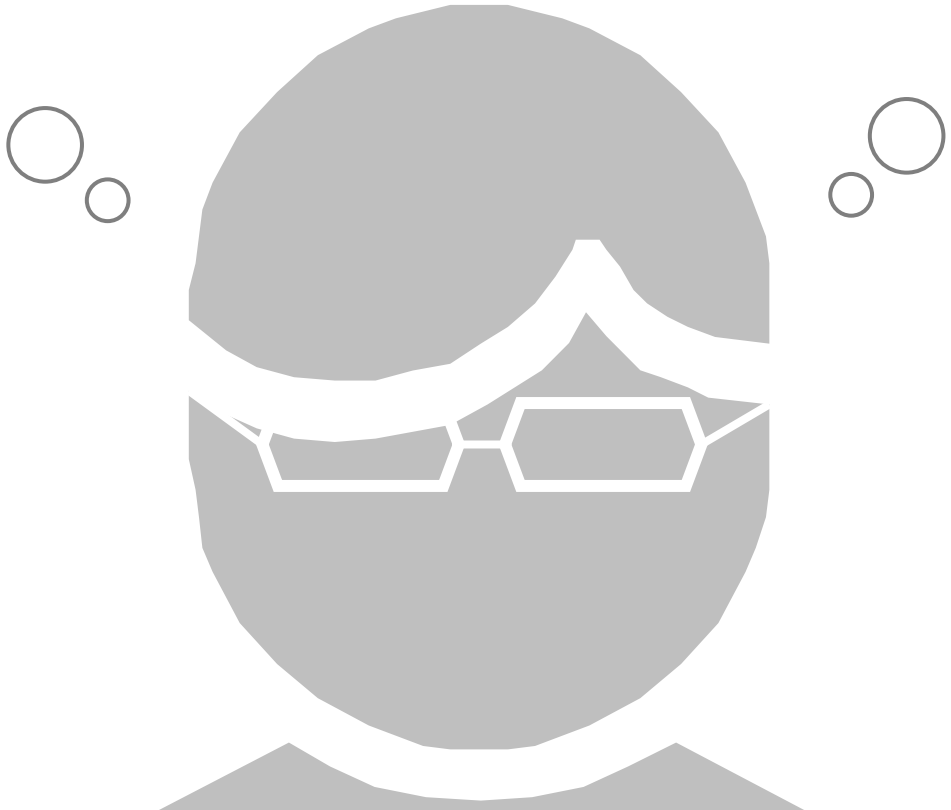
BUILD / PROTOTYPE?

Pros:

- Easy to assess impact

Cons:

- Expensive
- Time-consuming
- High risk



MODEL / SIMULATE?

Pros:

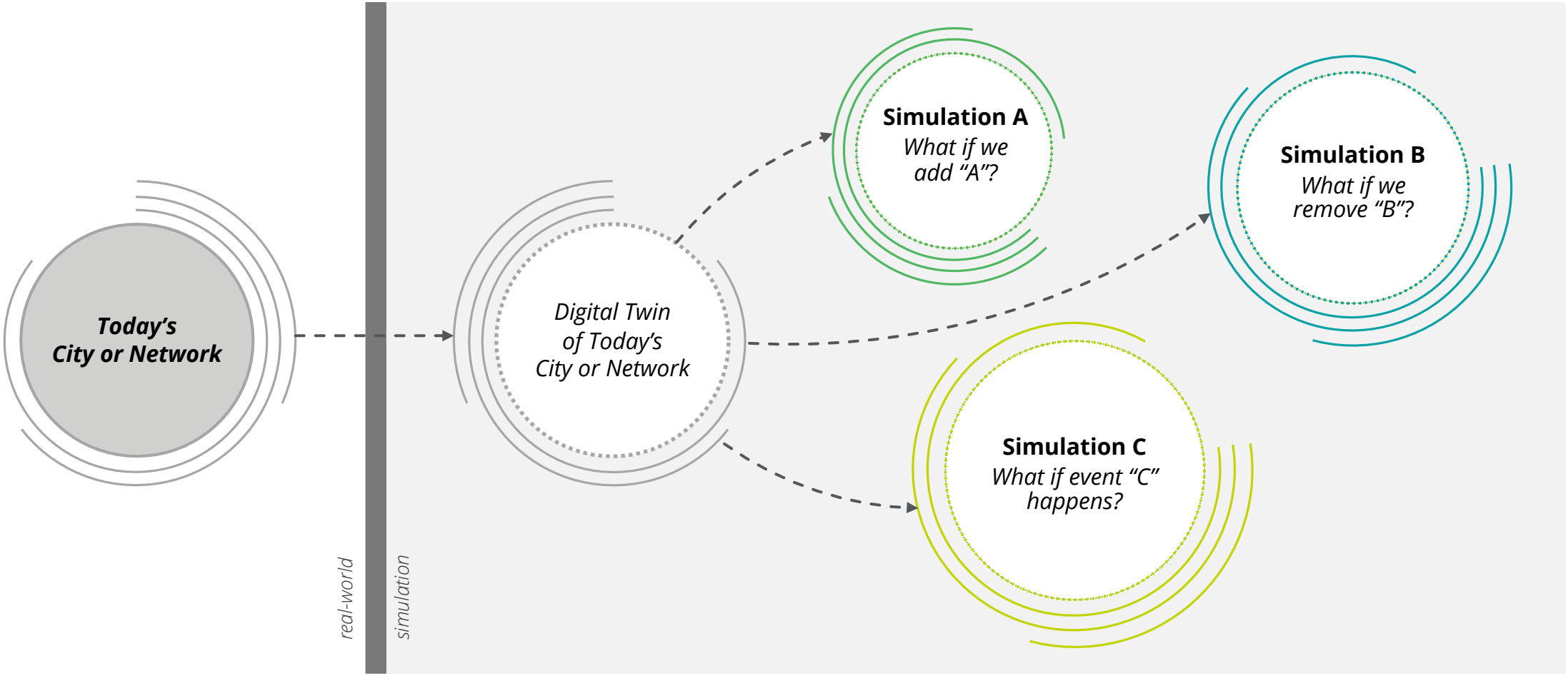
- Cost-effective
- Faster
- More flexible
- Limited risk

Cons:

- Need data
- Need to validate model

Organizations can use **simulation** to test improvements

Organizations can use simulation to understand complex, real-world systems (e.g., city, transportation systems) and test ways to improve them. A simulation is a faster, less risky way to answer “what if” questions.

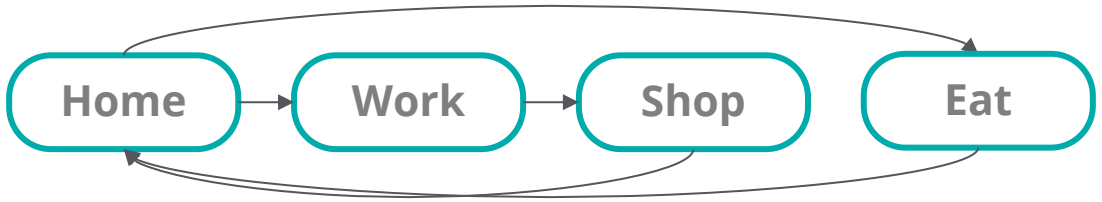


Activity-based models

As the transportation modeling field has evolved, the activity-based model has become the current state of the art

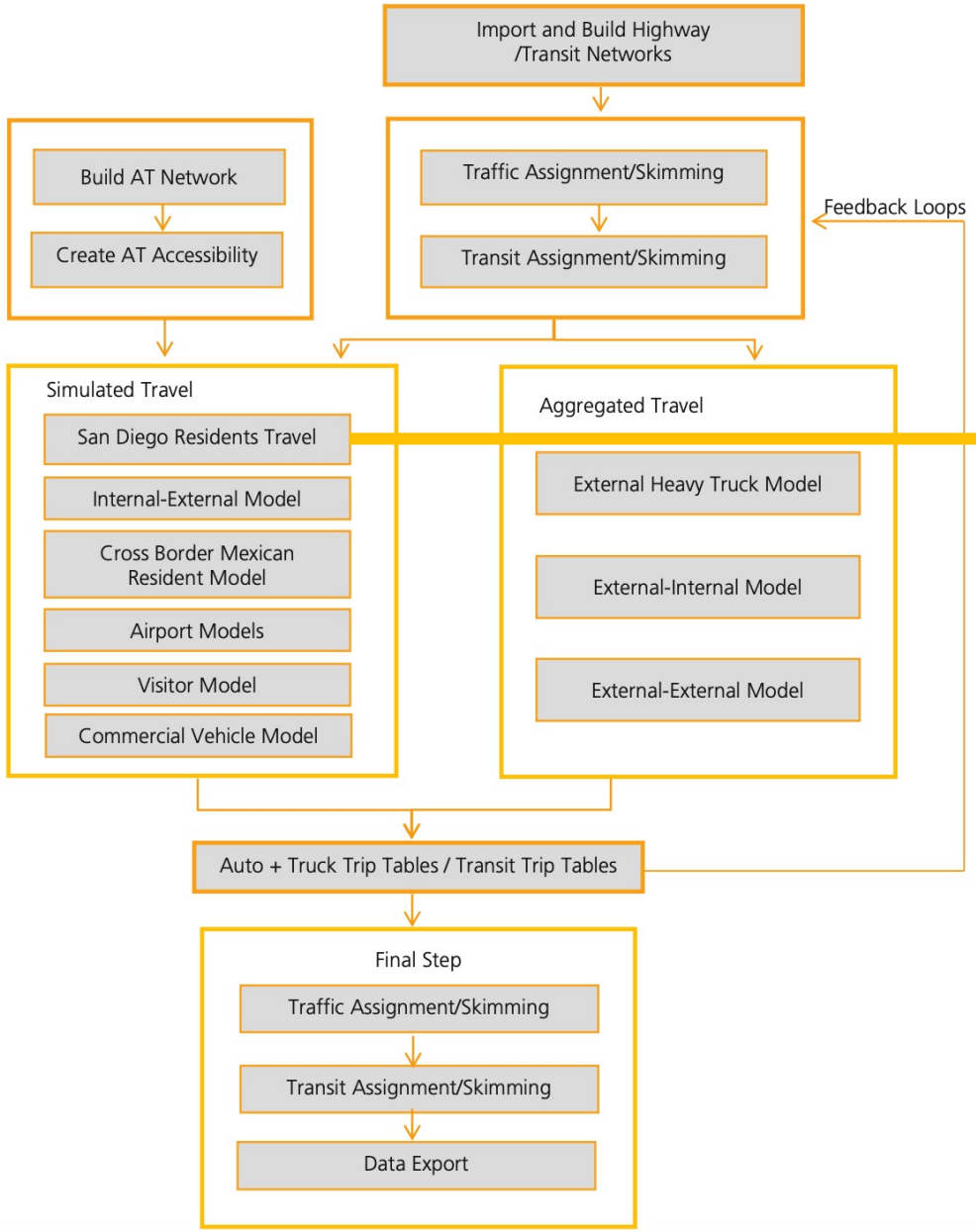
TRIP-BASED (4-STEP MODEL)

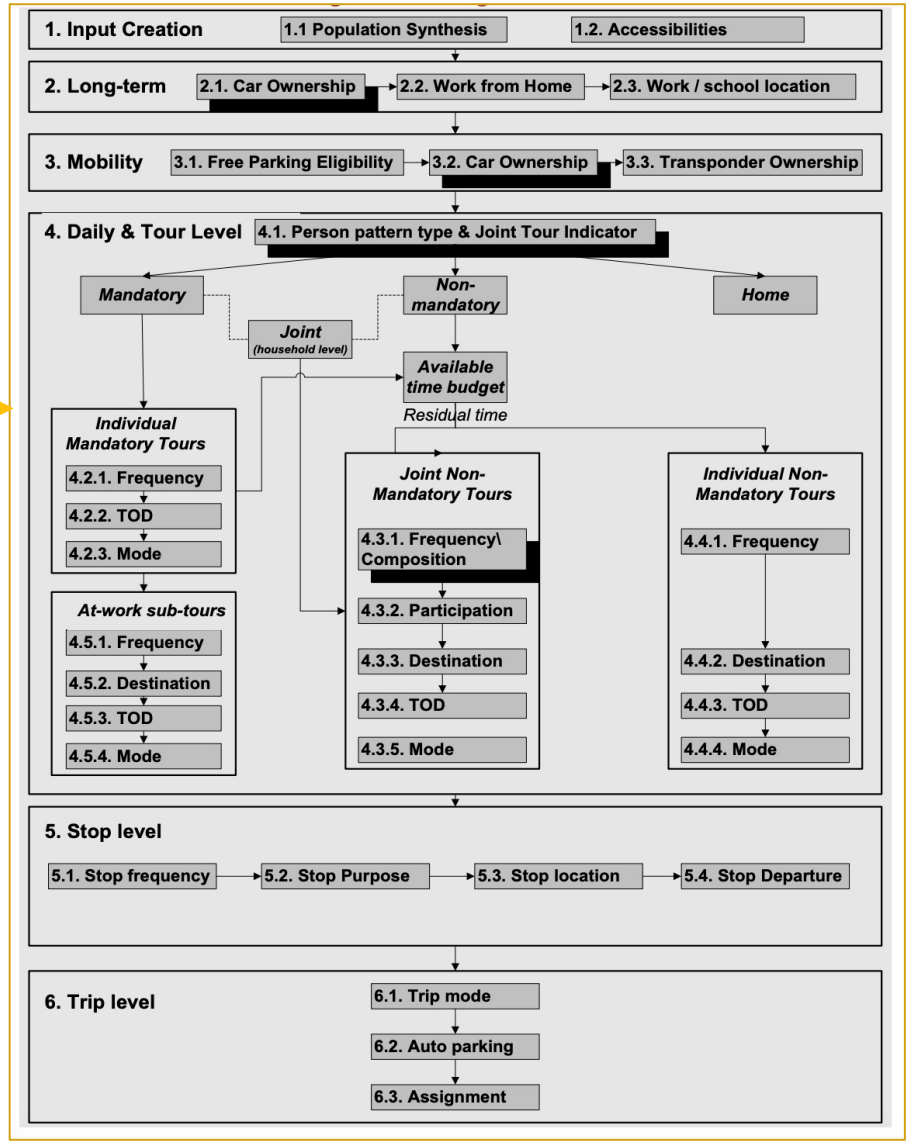
ACTIVITY-BASED



"Activity-based travel demand models **predict the long-term choices** (such as work location and automobile ownership) and the daily activity patterns of a given synthetic population, **including activity purposes, locations, timing, and modes of access.** These estimates of travel demand can be used to help **evaluate alternative transportation, land use, and other scenarios**."

SANDAG ABM2 Flow Chart





SANDAG Travel Demand Model and Forecasting Documentation

Where is the field heading?

Our point of view is that there is more data, better techniques, and enough computational power available today to advance the transportation modeling field in a number of ways

	ACTIVITY-BASED MODELS	AGENT-BASED MODELS
Granularity	Detail at the level of individual tours and trips	Precise detail at the level of individual agents
Population	Synthetic population generated from sampling Census data	Synthetic population generated using deep generative models
Activity Pattern	Activity patterns modeled based on survey results	Activity patterns modeled using mobility data as input
Mode Choice	Utility-based, nested logit mode choice models	Simplified and generalized mode choice model using tree-based methods

Tying it all together and looking ahead

Today

- Kicked off the course by setting the stage for who we all are and what we plan to cover
- Discussed transit demand modeling

Next week

- We'd like you to read chapter 3.1 in *Activity-Based Travel Demand Models: A Primer* and come ready to discuss mode choice modeling
- We'll discuss the current state of mode choice modeling and where we think there are opportunities to improve

Rest of Q1

- We'll use the approach in *Budholiya et al* as a framework for learning about classifier construction with XGBoost.
- Our data is a heart disease dataset with over 50 features. You'll need to examine that data, build a model, reduce the feature set (if necessary), and tune your final classifier.

Q2

- We have a large dataset from the mode choice selection process that occurs within FutureScape.
- Can you build a multiclass classification model to accurately select trip mode using a concise set of features?

