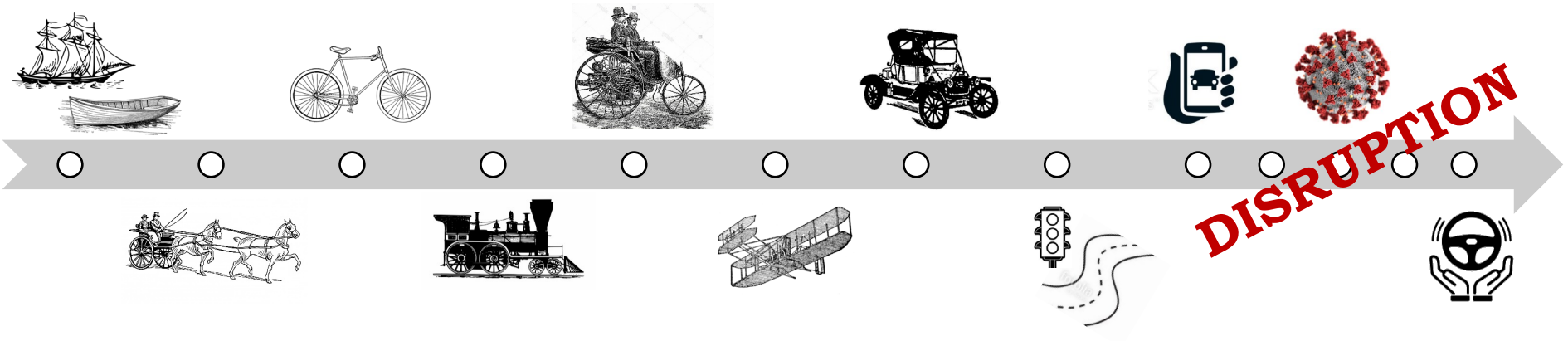
Enriching Transportation Survey Datasets Using Big Data and Machine Learning

With an Application for Transferring Attitudinal Variables across Transport Surveys

F. Atiyya Shaw & Patricia L. Mokhtarian
Georgia Institute of Technology
Applied Urban Modelling 2020: Modelling the New Urban World
2 November 2020

Unprecedented changes occurring rapidly…

across multiple dimensions



Technologies

- Uber, Lyft, car/bike-sharing, etc.
- Long-range electric vehicles
- Autonomous vehicles, drones
- Virtual reality “travel” experiences

Societal shifts

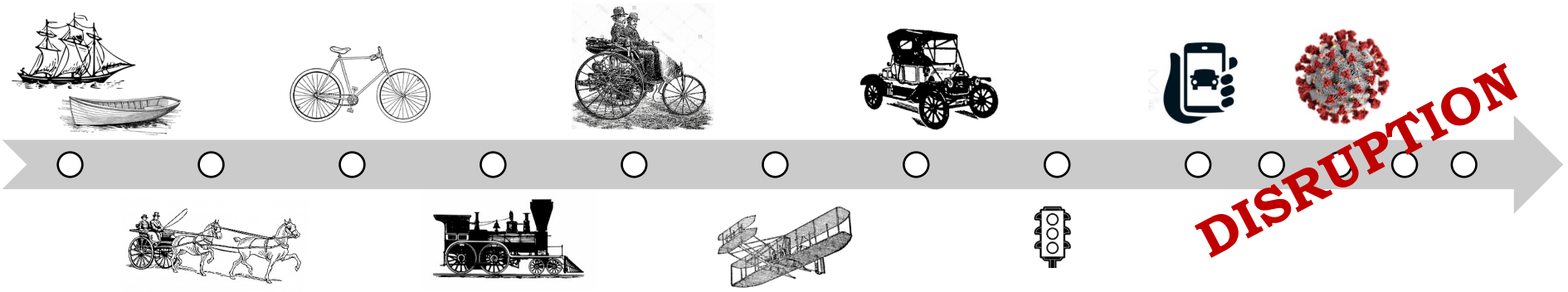
- Delayed or deferred marriage and childbearing
- Greater education
- Increasing ethnic diversity
- Shifting values

Policy instruments

- Denser, more diverse land uses
- Changing revenue base away from fuel taxes
- Yet-to-be-determined AV policies

Improving behavioral forecasting

Two approaches: functions or data



Behavioral decision-making

$$\begin{cases} Mobility\ Patterns_{today} = f_{today}(x_{today}) \\ Mobility\ Patterns_{future} = f_{today}(x'_{future}) \end{cases}$$

functions:

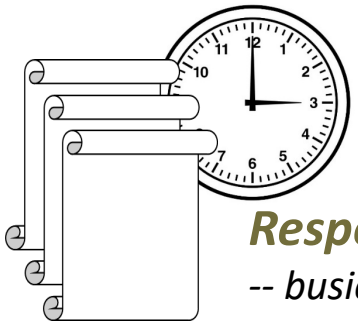
- *machine learning*
- *latent variable models*
- *advanced discrete choice models*

data:

- *passive data streams*
  - *big data: mobile phone location data, etc.*
- *active data streams*
  - *survey data*

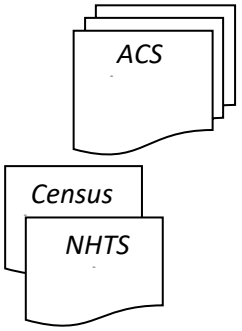
Obtaining good survey data is getting harder…

…and all evidence indicates this will continue

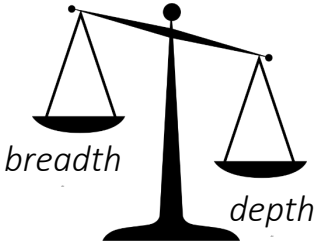


Respondents have gotten

-- busier
-- more "jaded"
-- more distracted

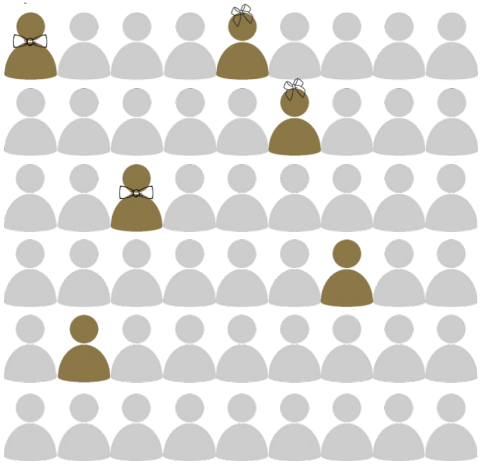


Finding a balance in content



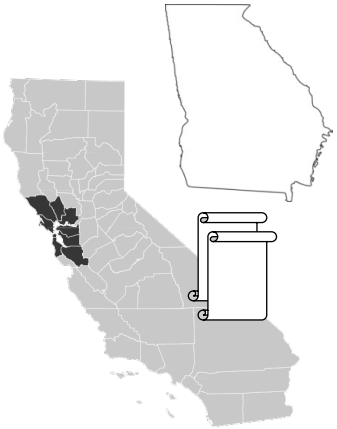
Longer surveys

→ lower response rates,
→ increased survey bias



Large-scale surveys like the National Household Travel Survey (NHTS)

-- socio-economic characteristics
-- observed travel behavior attributes
-- across a nationwide sample



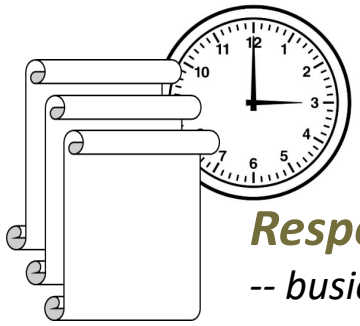
Smaller studies

-- broader set of variables
-- smaller samples
-- limited geographic areas

Large-scale household travel surveys

How can we get rich variables like attitudes into them?

Option 1: ask some attitudinal q's on the survey itself - easier said than done



Respondents have gotten

-- busier

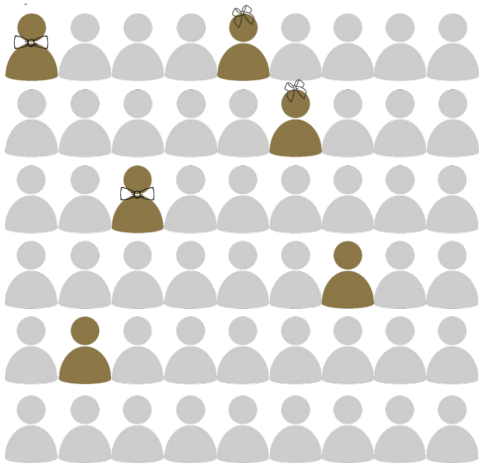
-- more "jaded"

-- more distracted

Longer surveys

→ lower response rates,

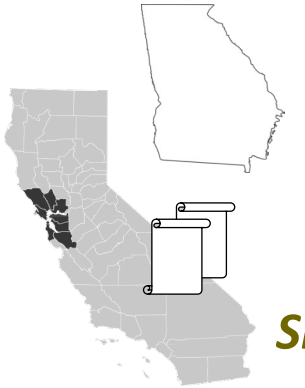
→ increased survey bias

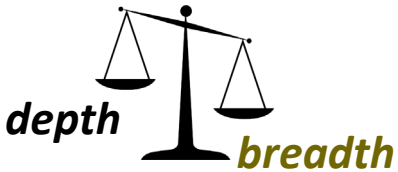


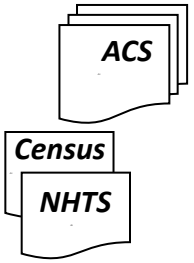
Large-scale household travel surveys

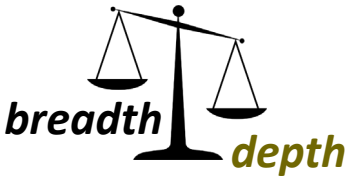
How can we get rich variables like attitudes into them?

Option 2: *"transfer" the info from other surveys*









Smaller studies

-- broader set of variables

-- smaller samples

-- limited geographic areas

National/regional surveys like NHTS

-- socio-economic characteristics

-- observed travel behavior attributes

-- sample drawn across a large area

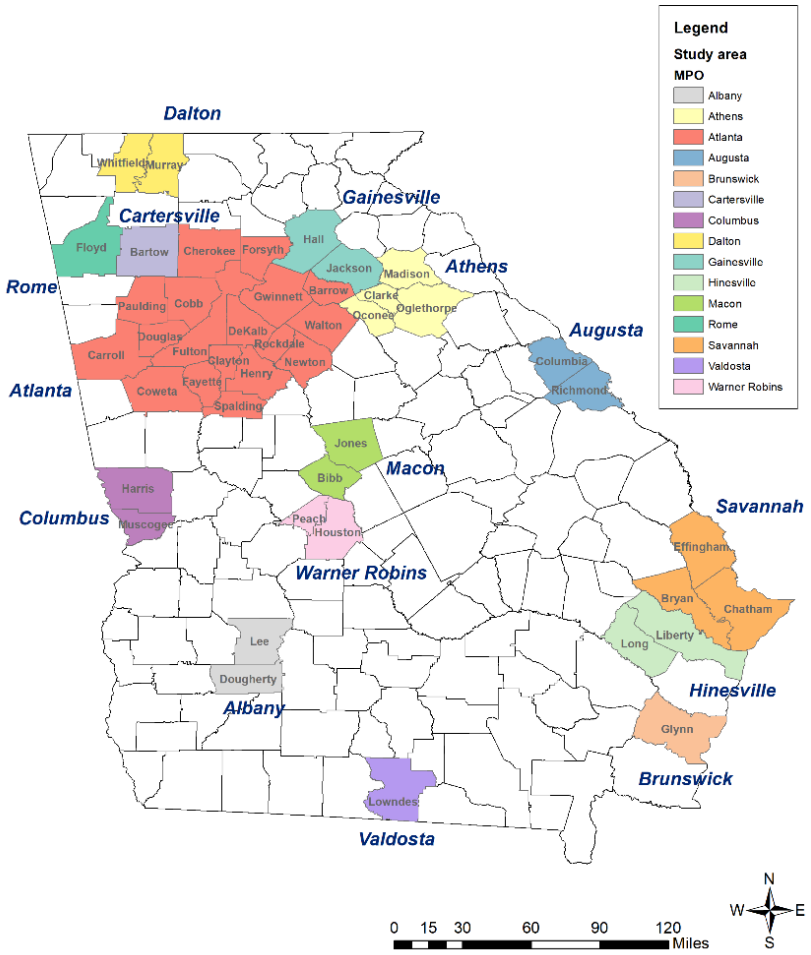
*Taking information from richer studies could be useful for lots of difficult-to-measure variable types (e.g. other **psychometric variables**)*

GDOT emerging technologies survey

The smaller, variable-rich survey – the "donor"

- **Purpose**:
  - To understand the impacts of emerging technologies and trends on travel behavior in Georgia (GA)
- **Details:**
  - Conducted Fall 2017
  - Invited sample: 15 MPO areas
  - ***Current*** N ~ 3300
- **Contents:**
  - ***A : Attitudes and personality***
  - *B : Technology usage*
  - *C : Key aspects of lifestyle*
  - *D : How you travel*
  - ***E : Evolving transportation services***
  - *F : Desires for future travel*
  - *G : Autonomous vehicles*
  - ***H : Sociodemographic traits***

Study area (15 MPOs, 45 counties)

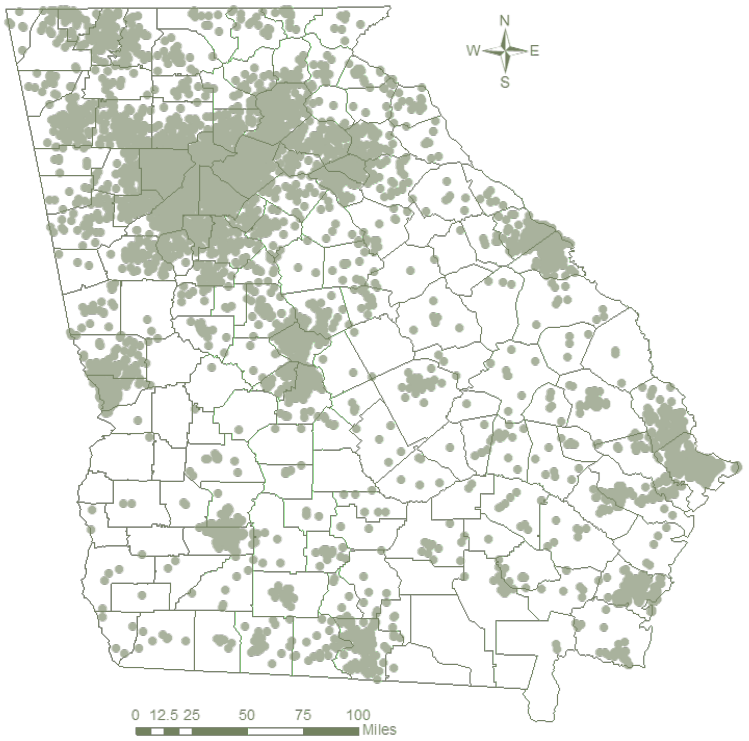


National Household Travel Survey (NHTS)

The national large-scale survey – the "recipient"

- **Purpose:**
  - To support travel demand modeling & long-range transportation planning across U.S.
- **Details:**
  - Repeated cross-sectional travel behavior survey
  - Georgia subsample used in this study
  - Wave: April 2016 to May 2017
  - ***Original*** N ~ 8632 (GA respondents)
- **Contents:**
  - ***Household data module***
  - *Long distance module*
  - *Vehicle data module*
  - ***Person level module***
  - *Person trips module*
  - ***Person health module***
  - *Person drive module*

Georgia subsample of 2016-17 NHTS respondents

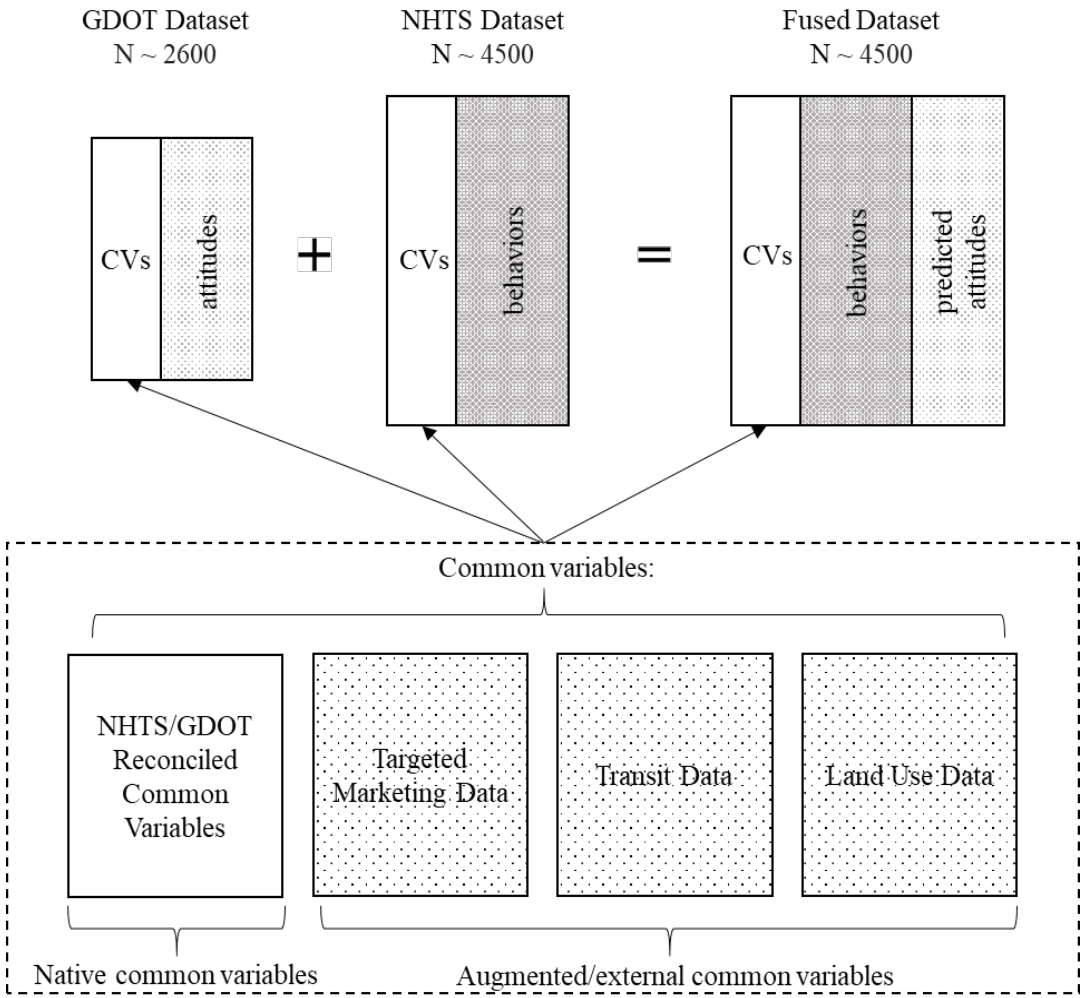


Expanding transportation survey datasets

Using common variables to transfer attitudes

$$Attitudes_{GDOT} = f_{GDOT}(CV_{GDOT}, augCV_{GDOT}) + \varepsilon_{GDOT}$$

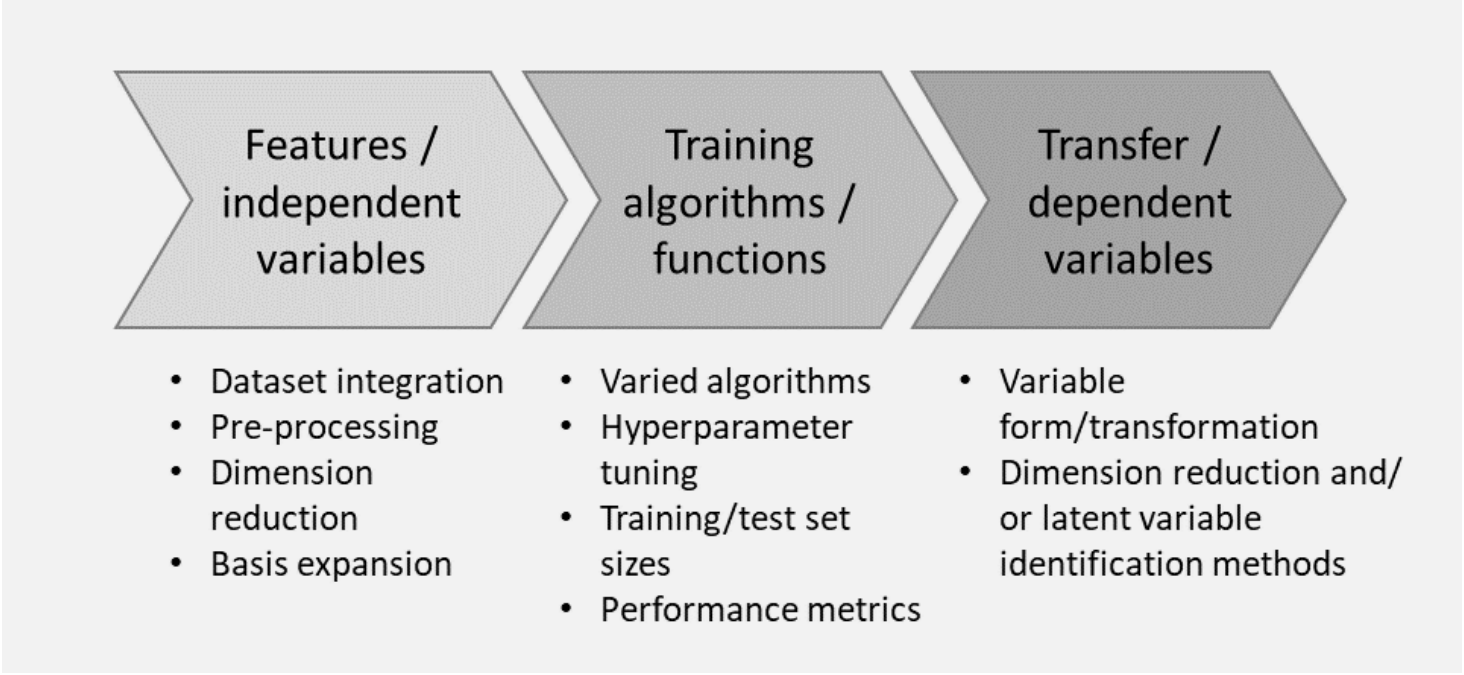
$$\widehat{Attitudes}_{Fused} = f_{GDOT}(CV_{NHTS}, augCV_{NHTS})$$



Expanding transportation survey datasets

Components of transfer process

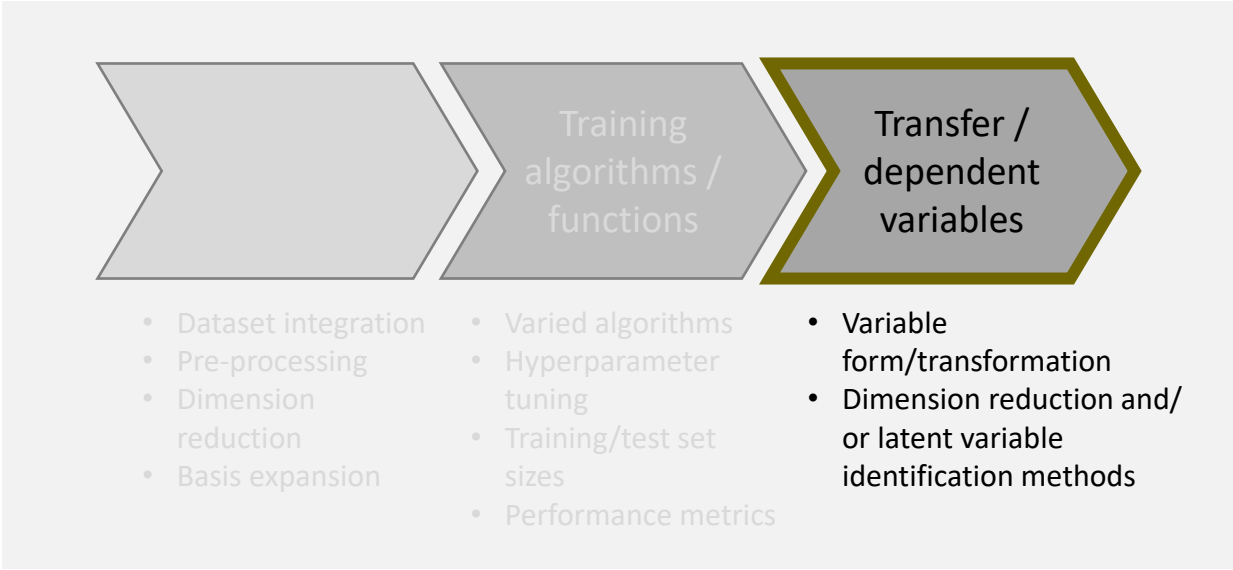
- **Transfer variables**: variables of interest being transferred across datasets
- **Features**: inputs to the training algorithm that are used to model/predict the transfer variables
- **Training algorithms**: used to transfer the variables across datasets



Components of the transfer process

Transfer variables

- **Transfer variables**: variables of interest being transferred across datasets
- **Features**: inputs to the training algorithm that are used to model/predict the transfer variables
- **Training algorithms**: used to transfer the variables across datasets



Transfer variables

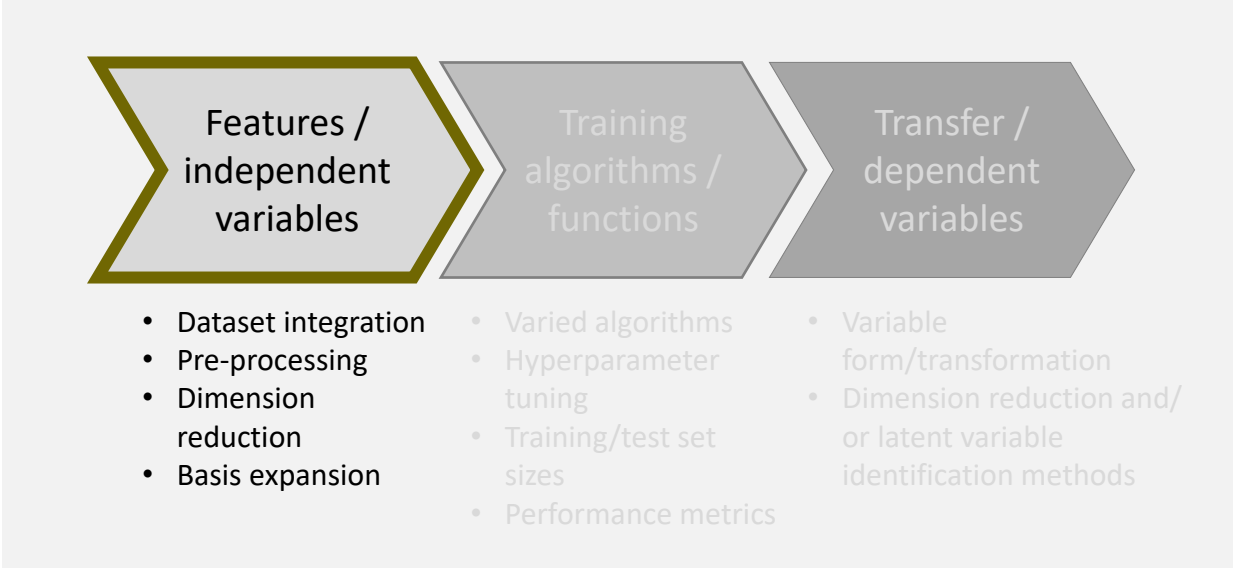
Introducing the attitudinal variables for transfer

	Name	Example statement
Lifestyle	Tech Savvy	Learning how to use new technologies is often frustrating for me (-)
	Work-oriented	… having fun is more important to me than working hard (-)
	Pro-exercise	I am committed to exercising regularly
	Materialistic	I would/do enjoy having a lot of luxury things
	Family-oriented	Family/friends play a big role in how I schedule my time
Land use	Pro-suburban	I prefer to live in a spacious home, even if it's farther away ...
	Urbanite	I like … having stores, … mixed among the homes in my n'hood
Travel	Non Car Mode	I like the idea of walking [bicycling, PT] as a means of travel for me
	Commute Benefit	My commute is a useful transition between home and work
	Travel Liking	I generally enjoy the act of traveling itself
	Car-owning	I definitely want to own a car
Personality	Polychronic	I prefer to do one thing at a time (-)
	Wait Tolerant	Having to wait is an annoying waste of time (-)
	Environmental	Cost or convenience takes priority over environmental impacts … (-)
	Sociable	I consider myself to be a sociable person

Components of the transfer process

Features

- **Transfer variables**: variables of interest being transferred across datasets
- **Features**: inputs to the training algorithm that are used to model/predict the transfer variables
- **Training algorithms**: used to transfer the variables across datasets



Features

Two types of features available for use

Survey Data Streams

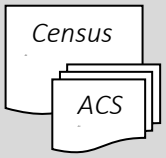
GDOT Emerging Technologies Survey
N ~ 3000

National Household Travel Survey Georgia Subsample
N ~ 8000

External Data Streams

Targeted Marketing Data
N ~ 6000

All Transit Data



Land Use Data

Native common variables – active data sources

- Exist initially in both datasets
- Tend to be SED variables
- Often must be adjusted and recoded across sources

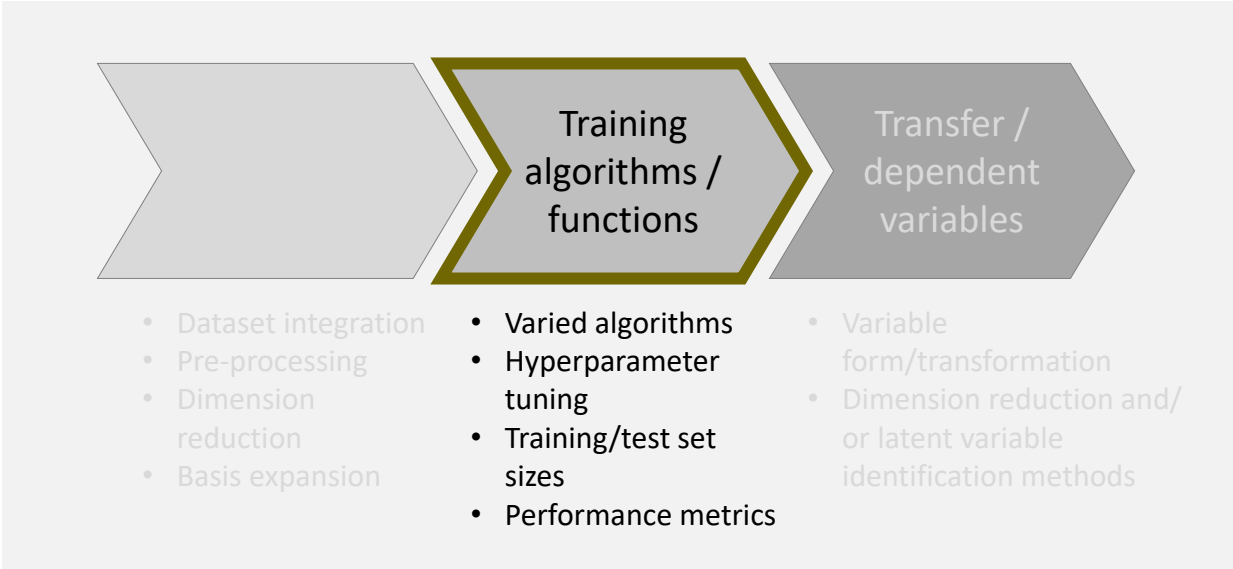
Augmented common variables – passive/active data sources

- Obtained from external active or passive datasets
- Must be appended to both the donor and recipient datasets at either:
  - Household level
  - Individual level
  - Geographic level

Components of the transfer process

Algorithms

- **Transfer variables**: variables of interest being transferred across datasets
- **Features**: inputs to the training algorithm that are used to model/predict the transfer variables
- **Training algorithms**: used to transfer the variables across datasets



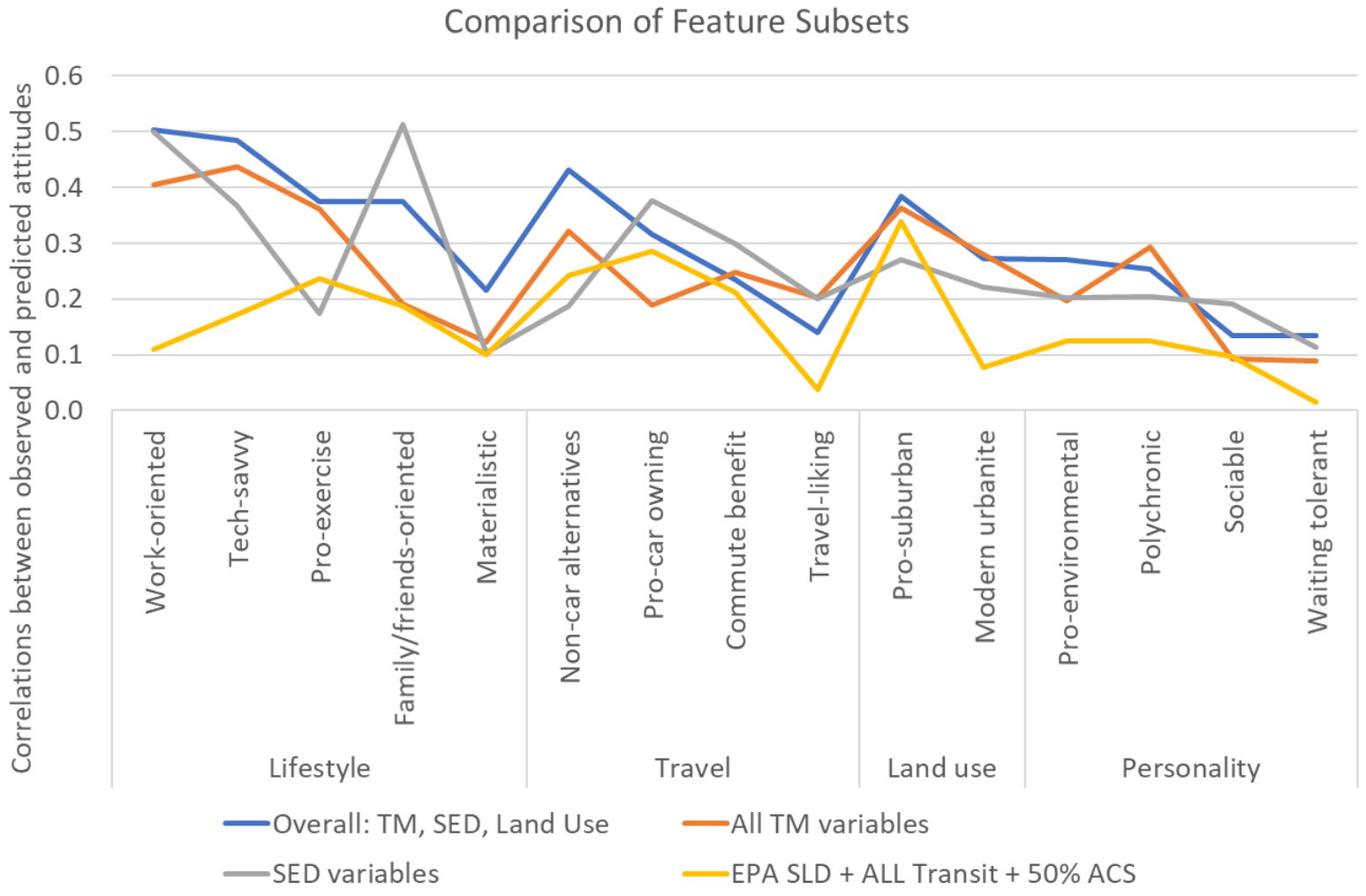
Algorithms

Possible variations within algorithm selection process

- **Algorithm selection**
  - Linear regression, random forest, support vector machine, **elastic net regression**, lasso regression, ridge regression, random forest, extreme gradient boosting
- **Algorithm tuning**
  - Training/test sample split: 80/20
  - Hyperparameter tuning using k-fold cross validation
  - Final metrics on test/hold-out sample
- **Algorithm performance**
  - Possible metrics: R-squared, **correlations (between observed and predicted),** mean squared error, misclassification error, etc.

How well are we transferring variables?

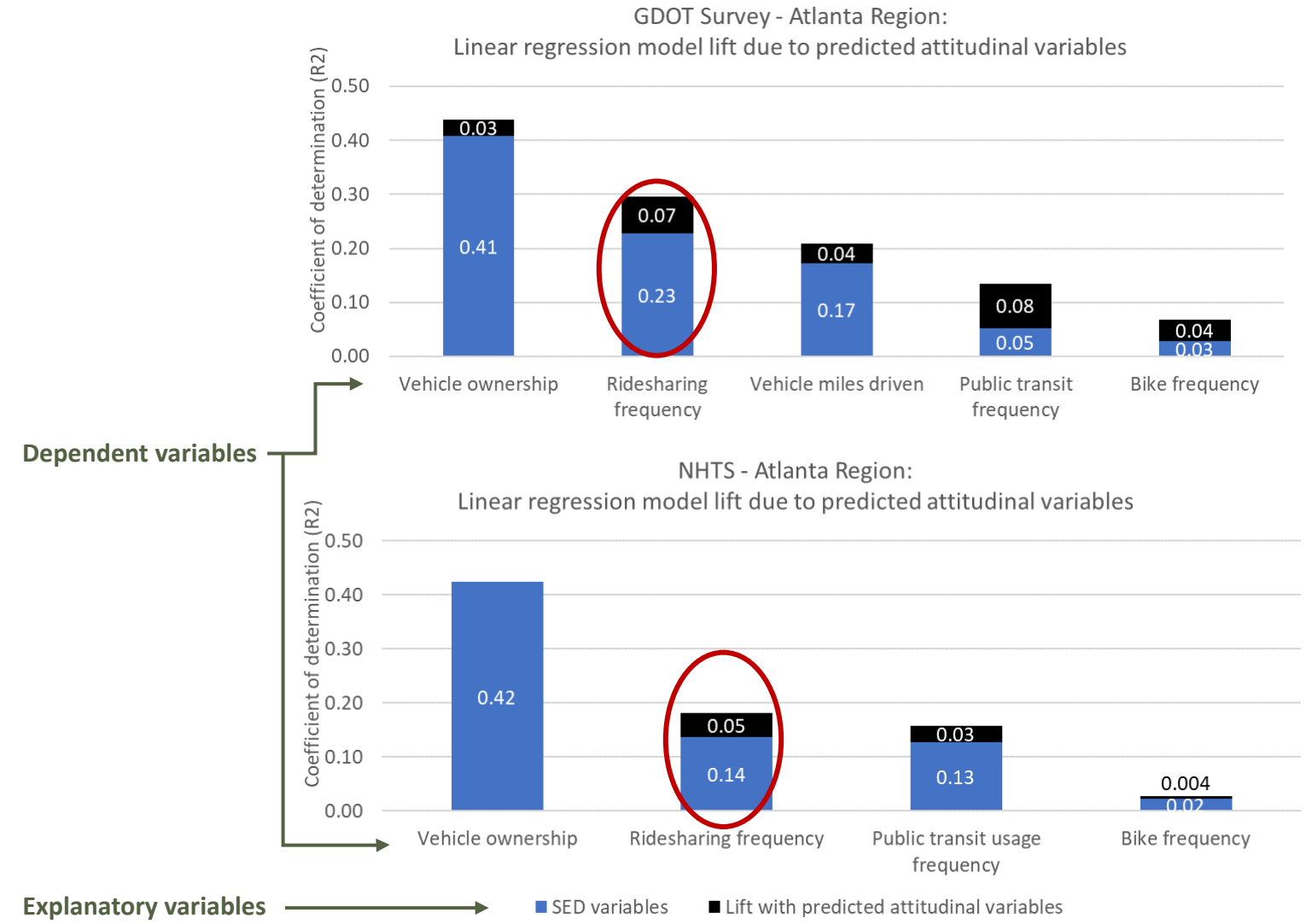
Internal validation



Are the transferred variables any good?

External validation:

How much are they helping in travel behavior models?



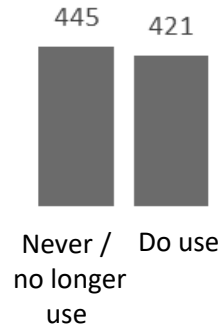
What will we be modeling for external validation?

Atlanta region: Ridehailing frequency

GDOT Survey

4. Please indicate *how often* you typically use each of the following transportation services.

	Never used / No longer use	*Less than once a month*	*1-3 times a month*	*1-2 times a week*	*3 or more times a week*
a. Carsharing	☐1	☐2	☐3	☐4	☐5
b. On-demand ride service	☐1	☐2	☐3	☐4	☐5
c. Shared on-demand ride service	☐1	☐2	☐3	☐4	☐5
d. Traditional taxi service	☐1	☐2	☐3	☐4	☐5

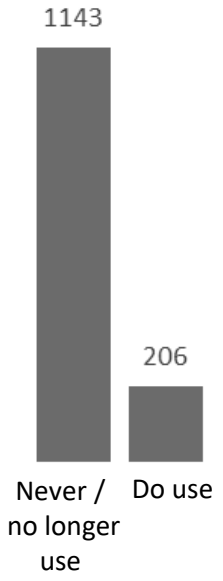


NHTS

RIDESHARE
Range: 0 - 99
ProgrammerNote: Asked if subject is at least 16 years of age

In the past 30 days, how many times [$HAVE_YOU] purchased a ride with a smartphone rideshare app (e.g. Uber, Lyft, Sidecar)?

WEB ATEXT	CATI ATEXT	AVALUE
ENTER NUMBER	ENTER NUMBER	
I don't know	DON'T KNOW	-8
I prefer not to answer	REFUSED	-7



How much are the transferred variables helping?

External validation: linear regression, R^2 values

GDOT Survey

explanatory variables

NHTS

explanatory variables

How much are the transferred variables helping?

Ext· validation: prediction accuracies for choice models

GDOT Survey

NHTS

LCCM: Latent Class Choice Model

Latent class choice model of ridehailing adoption

A preliminary look at some insights!

GDOT model w/ observed attitudes: three latent classes

- Travel liking class
- Family-oriented & car-owning
- Urbanite & tech-savvy

} *Emergent classes*

Choice model

Predictors	Travel-liking	Family-oriented	Urbanite
Education	--	(+) *	(+) **
Age	(-)***	(-) *	(-) **
Household income	(+) ***	--	--
Household size	--	(-) *	(-) **

Benefits:

- Improved predictive accuracy
- Nuanced interpretation
- Slight increase in model fit

Will the predicted attitudes yield similar benefits for GDOT and NHTS?

Study overview

The process in a nutshell…

1 DATA ACQUISITION

- Acquire **source** domain data
- Acquire **target** domain data
- Acquire **targeted marketing data (TMD)** for source and target domain

2 ESTABLISH COMMON VARIABLE SPACE

- Determine **common variables** original to source and target domains
- Determine **TMD augmented** common variables for source and target domains
- Determine **land use augmented** common variables for source and target domains

3 DATA PROCESSING

Prepare the following datasets for analysis:

- **Source** domain data
- **Target** domain data
- **TMD** data
- **Land-use** data

4 MACHINE LEARNING ALGORITHMS

- Determine ML algorithms for testing
- Train, optimize, and cross validate ML algorithms
- Evaluate ML algorithms relative to each other and benchmark measures

5 EXTERNAL VALIDATION

- Select dependent variable(s) that are possible candidates for external validation
- Evaluate model improvement due to imputed attitudinal variables

Study overview

Takeaways

- We can **impute attitudes** reasonably well!
  - …and we're getting better
- **Attitudes improve**:
  - Model fit
  - Model interpretability
  - Predictive accuracy (small increases)
- And **interpretations improve**:
  - Latent class choice models of travel behavior in demand forecasting models!
    - Policy implications of more nuanced segments

Study overview

Looking to the future

- In the **short-term**: let's use what we have!
  - Use trained algorithm to predict attitudes into Atlanta Household Travel Survey
  - How does it affect the final outcomes?
    - Is it moving the outcome in the same direction as the adjustment factors?
- In the **medium-term**: let's get better at predicting attitudes
  - Continued refinement of machine learning models
  - Additional common variables purchased and integrated into attitudinal prediction functions
- In the **long-term**: let's talk attitudinal variables!
  - Attitudinal marker statements that could be included on future household travel surveys
    - Can help us obtain improved predictions of attitudes
    - Can be used directly to improve our travel demand models
    - Will not be highly correlated with other explanatory variables

Thank you!

Atiyya Shaw
atiyya@gatech.edu
Patricia Mokhtarian
patmokh@gatech.edu

Acknowledgements

STRIDE | Southeastern Transportation Research, Innovation, Development and Education Center

TOMNET Transportation Center
Teaching Old Models New Tricks

GDOT Georgia Department of Transportation

NSF