Why do travel surveys matter in the Age of Big Data?

Patricia L. Mokhtarian
Georgia Institute of Technology
patmokh@gatech.edu

2018 NHTS Workshop
Washington, DC August 8-9, 2018

About the speaker…

Father: peripatetic Army helicopter pilot

+

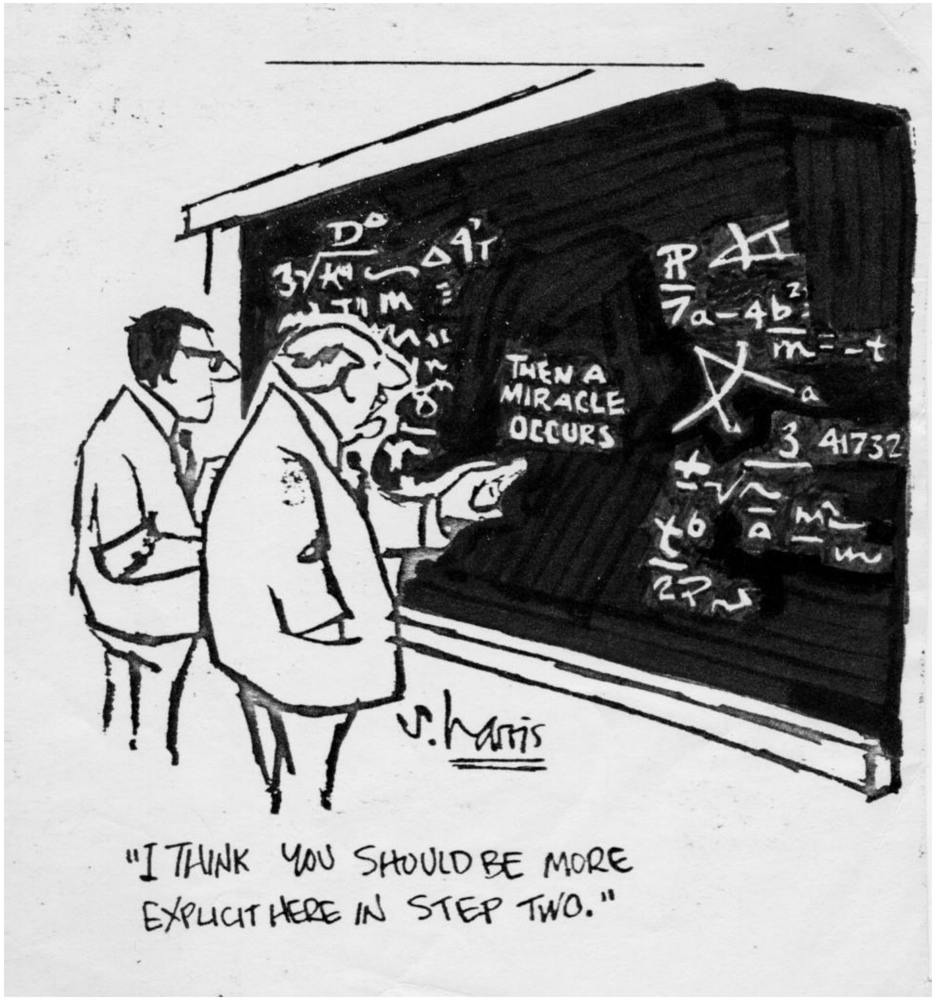
Mother: marriage and family therapist

=

Daughter: specialist in travel behavior analysis

About the speaker… (2)

- Math major



About the speaker… (2)

- Math major
- Entire career (since 1970s) devoted to the design, adminis-tration, and analy-sis of surveys measuring travel-related attitudes and behavior

About the speaker… (2)

- Math major
- Entire career (since 1970s) devoted to the design, administration, and analysis of surveys measuring travel-related attitudes and behavior
- Have been teaching a course on survey-based research methods for most of my 28-year faculty career
- Compulsive about survey design details

About the speaker… (3)

- Have I peaked???

https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=imgres&cd=&cad=rja&uact=8&ved=2ahUKEwj2gLmCpdncAhXKnuAKHY8FAi4QjRx6BAgBEAU&url=https%3A%2F%2Fnsmb.com%2Farticles%2F2017-kona-hei-hei%2F&psig=AOvVaw2nyHd5c62UmA2qwhQT5Da8&ust=1533674539304990

About the speaker… (3)

- Have I peaked???
- Even now, it's rare to find a course on survey methods in graduate *transportation* programs
  - Common in *psychology or sociology*
  - In *transportation*, you may find a course on *"data acquisition methods"*
- Will survey design have disappeared entirely from such courses within 5 years?
  - Replaced by "*Using Machine Learning Methods to Analyze Big Data*"?

Get with the times!

- In an era of
  - GPS traces
  - Transit smart cards
  - Clickstreams
  - RFID chips and scanner data
  - Twitter feeds and other social media posts
  - Remote sensing
  - Targeted marketing and credit reporting data
  - and more
- … who needs old-fashioned surveys???

Why do we still need surveys?

Three reasons:

- There's not always a Big Data source for what we need to know
- The Big Datasets we do have are incomplete
- Big Data is *even more valuable* when used in conjunction with survey data!

1. There's not always a Big Data source for what we need to know

a. Qualitative research

- ***Interviews*** about
  - Procurement
  - Priority-setting
  - Intra-household decision-making
  - Activity rescheduling
  - Other decision processes
- ***Focus groups*** on
  - Unmet needs, latent demand
  - Prospective policy impacts
  - Product/service design
- ***Charrettes*** on
  - Land use/transportation system changes

1. There's not a Big Data source for … (2)

b. Reliably identifying and measuring small/specialized populations

- Infrastructure performance managers at State DOTs
- Municipal traffic engineers
- Recent immigrants
- Single parents

1. There's not a Big Data source for… (3)

c. Hypothetical choices

- Impacts of *currently unavailable technologies* on travel, residential/job location
- Behavioral impacts of *proposed new policies*
- Removal of *constraints*
- Behavioral *intentions*

2. Even when we have Big Data sources…

- The data are far from perfect
  - Take GPS traces (please!*):
    - Broken trips
    - Urban canyons:
      - Signal blockage
      - Multi-path interference
    - Poor within-building performance
    - Dead batteries
    - Forgotten phones
    - etc…

*With a nod to Henny Youngman

2. Even when we have Big Data sources… (2)

- The data are far from perfect
- Vital context is missing:
  - Often even standard demographics are unknown
  - Want to apply aggregate statistics for the associated geographical unit?
  - Beware the ***ecological fallacy****!

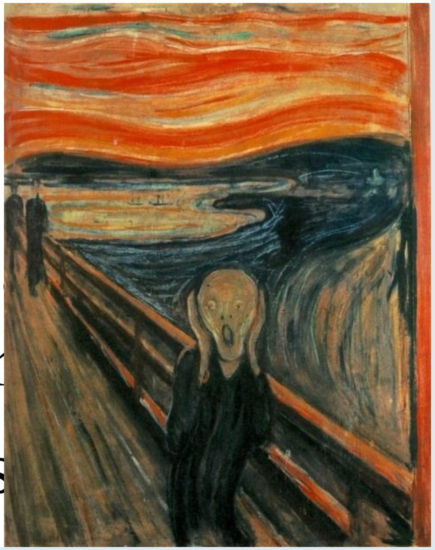
* Relationships at the ***aggregate*** level can be very different than – even the reverse of – those at the ***disaggregate*** level.

2. Even when we have Big Data sources… (3)

- The data are far from perfect
- Vital context is missing:
  - Often even standard demographics are unknown
  - Understanding the "why" of human behavior generally requires ***measuring the unobservable***, including
    - Constraints
    - Motivations (values)
    - Intentions
    - Personality
    - Lifestyle
    - Attitudes

2. Even when we have Big Data sources… (4)

- The data are far from perfect
- Vital context is missing
- Representativeness is (more) dubious
  - ~~“Since our sample is so large, representativeness is not a concern”~~
  - 1936 *Literary Digest* poll predicted over Roosevelt landslide: N = 2.3 m
  - Some exclusions are obvious, others



2. Even when we have Big Data sources… (5)

- The data are far from perfect
- Vital context is missing
- Representativeness is (more) dubious
- Correlation doesn’t equal causality

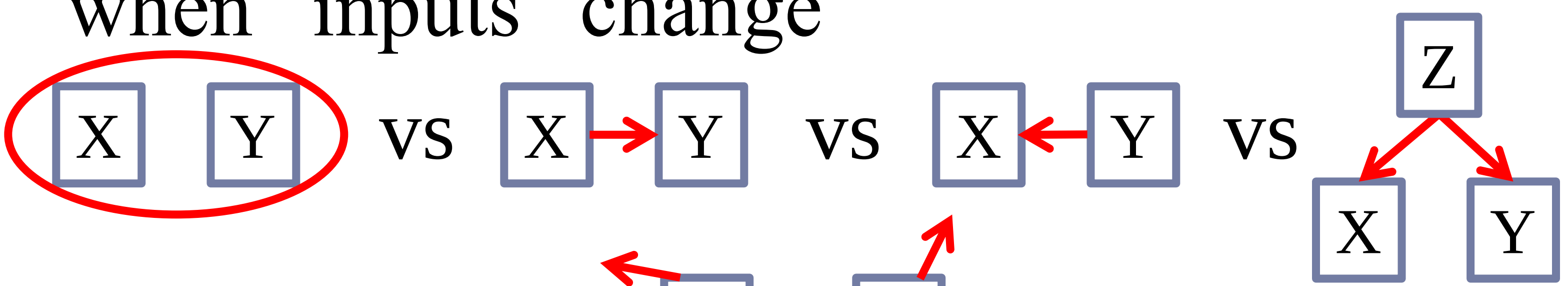
Wait – isn’t causality passé?

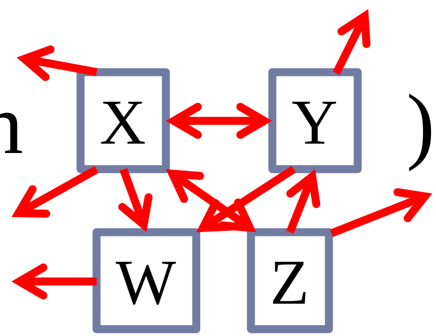
- “But today, data is so readily available and computers are so fast and powerful that **experts … have stopped trying to figure out why something – say, crime – happens**. Instead they look at crimes and notice what events or behaviors seem to precede them… [T]he tricky work of turning information into knowledge has **shifted from causation to correlation**.”

– Fareed Zakaria, *Time*, 7/8/2013

So why *does* causality matter?

- We're hardwired to ask, "Why?"
- It's our best hope of predicting "outcomes" when "inputs" change



(not to mention 

)

- because knowing the "why?" improves our understanding of the "what will happen if?"

A tale of two causal models

- "Whenever the dog tries to attack you, you give him a treat to get him to stop?" "Yes, and it works every time!"
- Human's model:
  - give treat → attack stops
- Poodle's model:
  - attack human → receive treat

"A client came in with her poodle and warned us that the dog would bite. She said that it would often corner her in a room at home, too, and snarl and sometimes bite. I asked how she handled it, and she said, 'Well, I started throwing food to get him away from me, and it worked. So now I keep snacks in every room just in case.' 'So ...,' I asked incredulously, 'whenever he tries to attack you, you give him a treat?' 'Yes,' she answered, 'and it works every time!'"

Dennis Leon, DVM

Reader's Digest, May 2012, p. 188

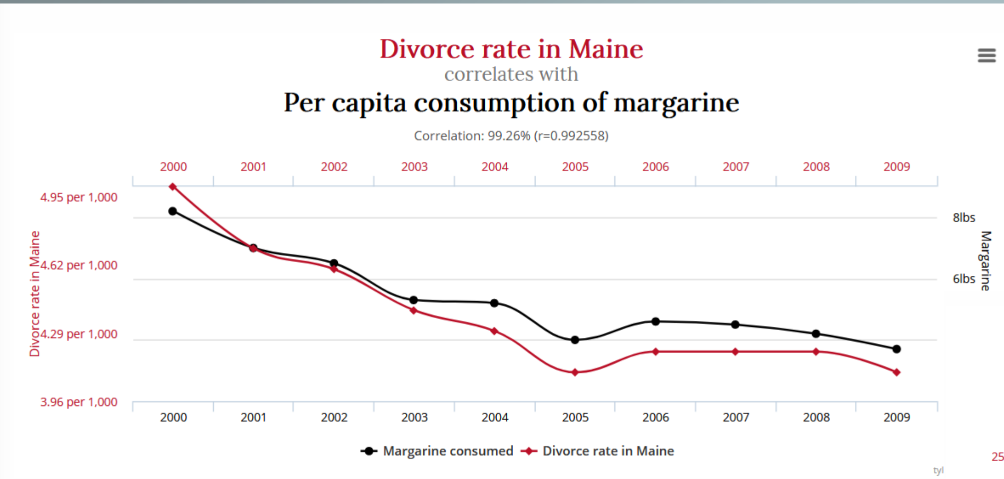
A tale of two causal models (2)

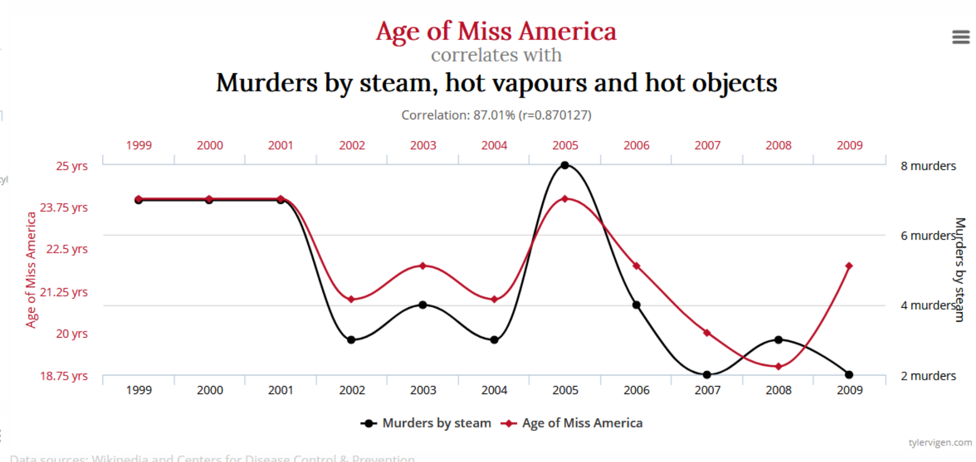
- Human
  - ***Model:*** give treat → attack stops
  - ***Policy implication:*** keep giving treats

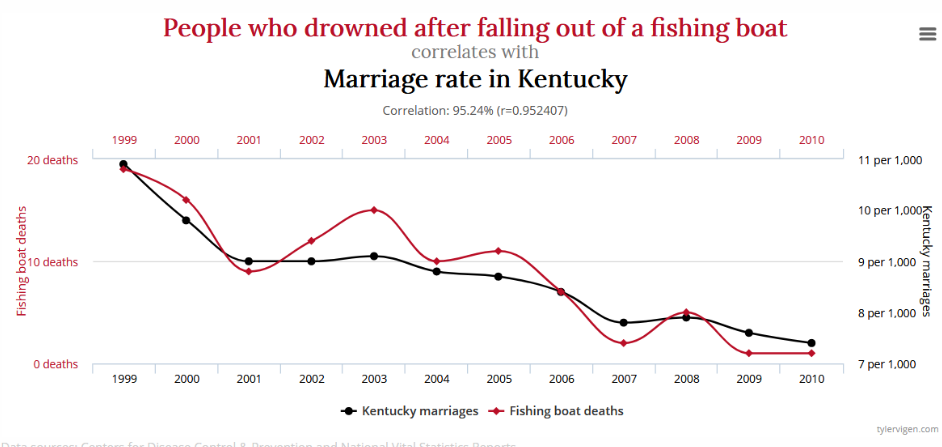
- Poodle
  - ***Model:*** attack human → receive treat
  - ***Policy implication:*** keep attacking

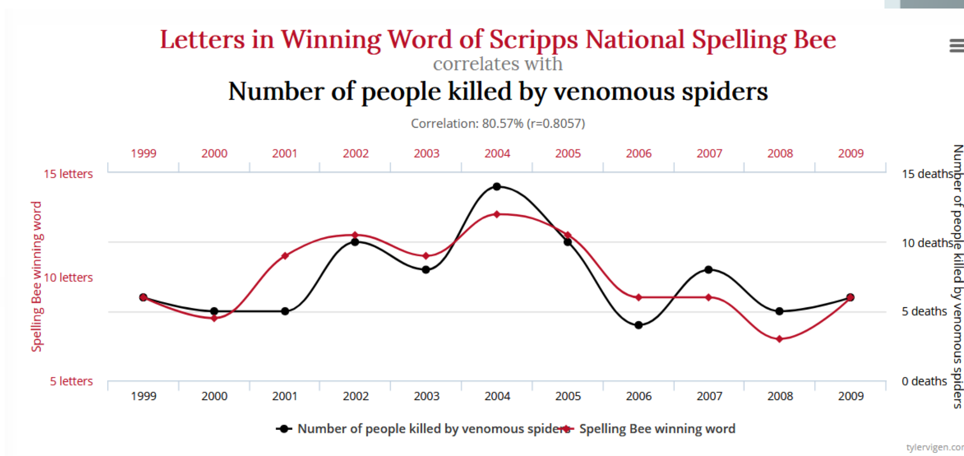
- Neither view of reality achieves the socially-optimal outcome…
- There's no substitute for domain knowledge…

If correlation is all you look at…









http://tylervigen.com/spurious-correlations

3. Big Data can enrich survey data

- We've previously considered some advantages afforded by ***survey data***
  - Measurement of specialized populations
  - Measurement of important, but unobserved, variables (constraints, motivations, etc.)
  - Greater representativeness
  - Greater illumination of "why?"
- Let's now consider some advantages offered by ***Big Data***

Also see Varian (2014)

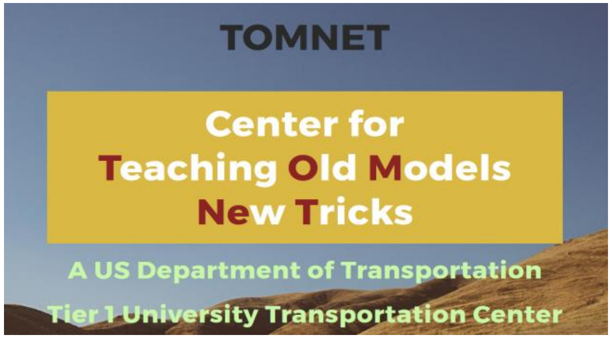
3. Big Data can enrich survey data (2)

- Some Big Data advantages *for causal models*
  - Improved matching
  - More cases around a regression discontinuity
  - Ability to analyze population segments
    - Assuming you can identify those segments, you're likely to have a lot of cases in them
  - Ability to "experiment" on a large scale, in "ecologically valid settings"
  - Ability to track dynamics

Grimmer 2015; also see Varian 2014

TOMNET (a Tier 1 UTC): Teaching Old Models (and Modelers) New Tricks

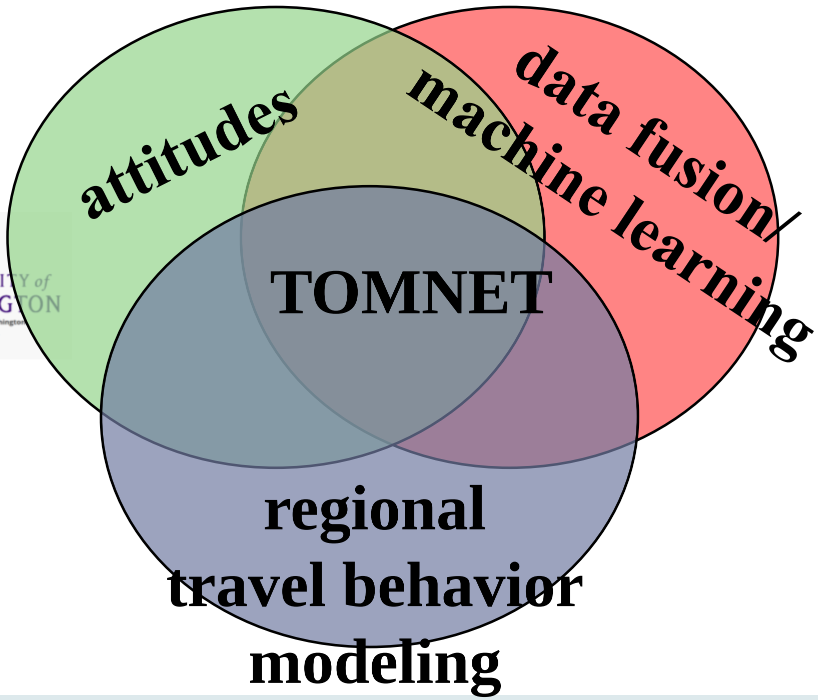
Arizona State University



USF UNIVERSITY OF SOUTH FLORIDA
University of South Florida
Tampa, FL

UNIVERSITY of WASHINGTON
University of Washington
Seattle, WA

Georgia Tech
Georgia Institute of Technology
Atlanta, GA

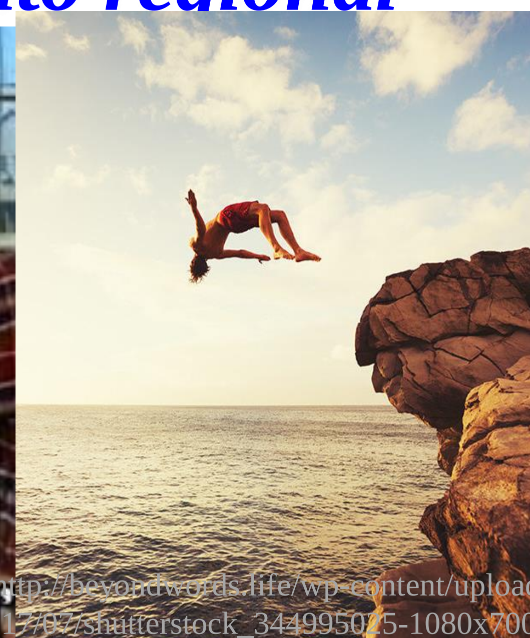


TOMNET (2)

- If you'd told me a few years ago that I'd be embracing ***machine learning***, and using it to pursue a decades-long dream of ***bringing attitudinal information into regional models***…

I'd have said …

But just look at me no



http://beyondwords.life/wp-content/uploads/2017/07/shutterstock_344995025-1080x700.jpg

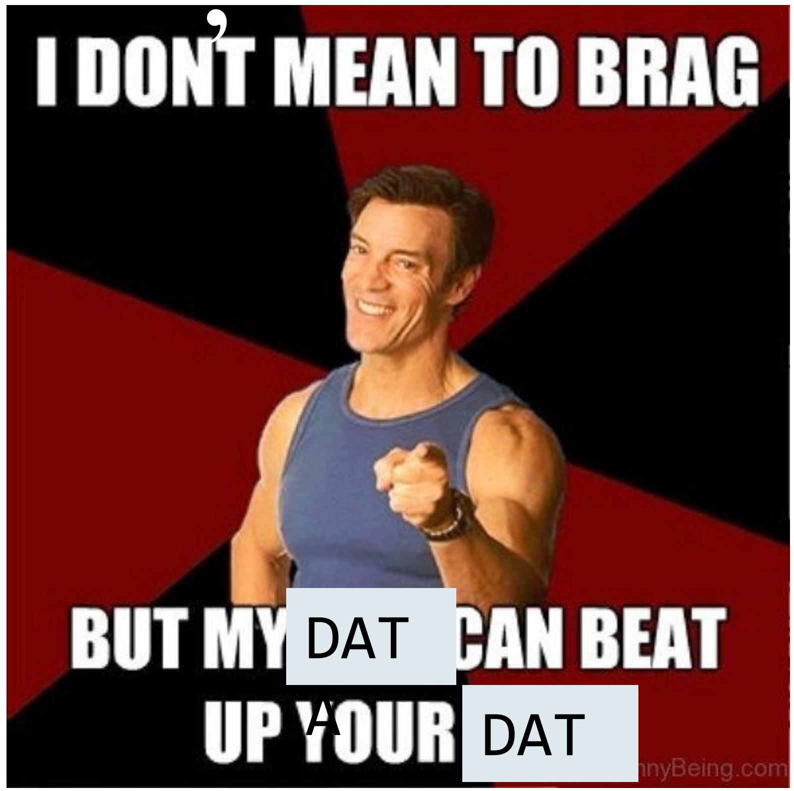
Here's what we're working on

GDOT survey data	NHTS (GA subsample)	Virtual survey
Donor (source)	*Recipient (target)*	*Fused data*
N = 3,000	N = 8,000	N = 8,000

Adapted from Fig. 1 (p. 2) of van der Putten et al. (2002). See Malokin et al. (2018).

In conclusion,
may I suggest …

Enough with the data strut and swagger, already!



http://www.funnybeing.com/wp-content/uploads/2017/03/I-Dont-Mean-To-Brag-600x600.jpg

Enough already! (2)

- We’ve seen that each of these approaches can
  - do things the other one can’t; and
  - make the other one better
- Discarding either one deprives planning/ policymaking of the insights made possible by the “other” kind of data, and by *both* kinds of data working in harmony

Enough already! (3)

- So instead of arguing about why we don't need this kind, or how the other kind can't be trusted, let's
  - have them *both* in our arsenal, using each singly – and both together – as appropriate
  - consider both perspectives, and how each can improve the other, e.g.,
    - Consider causality, representativeness when using Big Data
    - Integrate "Big Data methods" into survey data analysis
    - Combine survey data and passive data collection – like NextGen NHTS…

“Looking at things in multiple ways creates a richer and more true understanding of the world”

– Susan Handy (2013)

(speaking on the power of combining qualitative and quantitative methods)

Enough already! (4)

- Yes, that means I advise transportation students nowadays to take courses in machine learning…
- … while not forsaking the classics of survey design and causal modeling
- IT'S NOT AN EITHER-OR PROPOSITION…

Enough already! (4)

- Yes, that r… on students nowadays… e learning…
- … while r… f survey design and…
- IT'S NOT… OSITION…

- Let's make this the start of a beautiful friendship!

Selected references

- **Athey, Susan and Guido W. Imbens** (2015) Machine Learning Methods for Estimating Heterogeneous Causal Effects. arXiv:1504.01132v1.
- **Athey, Susan** (2015) Machine Learning and Causal Inference for Policy Evaluation. http://dx.doi.org/10.1145/2783258.2785466.
- **Breiman, Leo** (2001) Statistical modeling: The two cultures. *Statistical Science* **16(3)**, 199-231.
- **Grimmer, Justin** (2015) We are all social scientists now: how Big Data, machine learning, and causal inference work together. American Political Science Association, doi:10.1017/S1049096514001784.
- **Handy, Susan** (2013) The power of mixed methods: Examples from the field of travel behavior research. Guest lecture, TTP 200 (Transportation Survey Methods), University of California, Davis.
- **Malokin, Aliaksandr, Patricia L. Mokhtarian, and Giovanni Circella** (2018) An investigation of methods to enrich National Household Travel Survey data with attitudinal variables. Paper presented at the 2018 Annual Meeting of the Transportation Research Board, available from the authors.

Selected references (2)

- **Rose, Sherri, Richard J.C.M. Starmans, and Mark J. van der Laan** (2012) Targeted Learning for Causality and Statistical Analysis in Medical Research. In *Statistics: Discovering Your Future Power*. China Statistics Press. https://biostats.bepress.com/ucbbiostat/paper297/
- **Sliva, Amy, Scott Neal Reilly, David Blumstein, Steve Hookway, and John Chamberlain** (2017) Modeling Causal Relationships in Sociocultural Systems Using Ensemble Methods. In S. Schatz and M. Hoffman (eds.), *Advances in Cross-Cultural Decision Making*, Advances in Intelligent Systems and Computing 480.
- **Van Der Puttan, Peter, Joost N. Kok, and Amar Gupta** (2002) Data fusion through statistical matching. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=297501, accessed Aug. 6, 2018.
- **Varian, Hal** (2014) Machine Learning and Econometrics. https://web.stanford.edu/class/ee380/Abstracts/140129-slides-Machine-Learning-and-Econometrics.pdf, accessed Aug. 6, 2018.
- **Wu, Xindong and Vipin Kumar** (2009) *The Top Ten Algorithms in Data Mining*. Boca Raton, FL: Chapman & Hall/CRC.

Thank you!

Questions?

patmokh@gatech.edu

http://ce.gatech.edu/people/faculty/6251/overview