

# The Impact of Machine Learning on Economics

Susan Athey, Stanford University

# References: My Research on ML & Causal Inference

- ▶ **ATE/Unconfoundedness**
  - ▶ Athey, S., G. Imbens, S. Wager: “Efficient Inference of Average Treatment Effects in High Dimensions via Approximate Residual Balancing,” 2016, <http://arxiv.org/abs/1604.07125>
- ▶ **Robustness and Supplementary Analysis**
  - ▶ Athey, S., & Imbens, G. (2015). “A measure of robustness to misspecification.” *The American Economic Review*, 105(5), 476-480. (Canvas: Robustness)
  - ▶ Athey, S., G. Imbens, T. Pham, and S. Wager. Estimating average treatment effects: Supplementary analyses and remaining challenges. *American Economic Review*, 107(5):278{81, 2017b.
- ▶ **Difference-in-Difference and Panel Data**
  - ▶ Athey, Bayati, Doudchenko, Imbens, Khosravi (2016), “Matrix Completion Methods for Causal Panel Data Models.”
- ▶ **Heterogeneous treatment effects-subgroup analysis with valid confidence intervals**
  - ▶ Athey, S. and G. Imbens, “Recursive Partitioning for Heterogeneous Effects,” 2016, *Proceedings of the National Academy of Science*. <http://arxiv.org/abs/1504.01132>
- ▶ **Heterogeneous treatment effects-personalized estimates for experiments, unconfounded designs, instrumental variables, and general method of moments models, with asymptotic normality and confidence intervals**
  - ▶ Wager, S. and S. Athey, “Estimation and Inference for Heterogeneous Treatment Effects Using Random Forests,” forthcoming, *Journal of the American Statistical Association*. <http://arxiv.org/abs/1510.04342>
  - ▶ Athey, S., J. Tibshirani, and S. Wager. Generalized random forests. arXiv:1610.01271, 2017d. <https://arxiv.org/abs/1610.01271>
- ▶ **Optimal Policy estimation**
  - ▶ Athey, S. and S. Wager. Efficient policy estimation. arXiv:1702.02896, 2017. <https://arxiv.org/abs/1702.02896>
- ▶ **Surrogates, combining experimental and observational data**
  - ▶ Athey, S., et al. “Estimating Treatment Effects using Multiple Surrogates: The Role of the Surrogate Score and the Surrogate Index.” arXiv:1603.09326 (2016).
- ▶ **Testing for peer effects in Network Experiments**
  - ▶ Athey, S., D. Eckles, and G. W. Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, (just-accepted), 2016b.
- ▶ **Bandits**
  - ▶ Athey, S., M. Diamakopoulou, W. Du, G. Imbens, “Contextual Bandits and Causal Inference”
- ▶ **Unsupervised learning for text analysis**
  - ▶ Athey, S., M. Mobius, and J. Pal. The impact of aggregators on internet news consumption. 2017.
- ▶ **Large-scale Bayesian structural models**
  - ▶ Athey, S., and D. Nekipelov. “Heterogeneity and Advertiser Preferences in Sponsored Search Auctions,” in progress.
  - ▶ Athey, S., D. Blei, R. Donnelly, and F. Ruiz. “Counterfactual inference for consumer choice across many product categories.” 2017a.
  - ▶ Athey, S., D. Blei, F. Ruiz, “Item Embeddings for Demand Estimation.”

# Some accessible references and surveys

## ▶ Intro

- ▶ Hal Varian, “Big Data: New Tricks for Econometrics,” *The Journal of Economic Perspectives*, 28 (2), Spring 2014, 3-27.  
<http://people.ischool.berkeley.edu/~hal/Papers/2013/ml.pdf>

## ▶ Prediction v. Estimation

- ▶ Mullainathan, Sendhil, and Jann Spiess. “Machine learning: an applied econometric approach.” *Journal of Economic Perspectives* 31.2 (2017): 87-106.

## ▶ Prediction policy

- ▶ Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. “Prediction policy problems.” *The American Economic Review* 105, no. 5 (2015): 491-495.

## ▶ Prediction v. Causal Inference

- ▶ S. Athey. Beyond prediction: Using big data for policy problems. *Science*, 355 (6324):483-485, 2017.
- ▶ A. Belloni, V. Chernozhukov, C. Hansen: “High-Dimensional Methods and Inference on Structural and Treatment Effects,” *Journal of Economic Perspectives*, 28 (2), Spring 2014, 29-50.  
<https://www.aeaweb.org/articles?id=10.1257/jep.28.2.29>

## ▶ Overview of Heterogeneous Treatment Effects

- ▶ S. Athey and G.W. Imbens. The state of applied econometrics: Causality and policy evaluation. *The Journal of Economic Perspectives*, 31 (2):3-32, 2017.

# Machine Learning

## Supervised ML

- ▶ Outcomes  $Y$
- ▶ Features  $X$
- ▶ Independent obs.
- ▶ Goal: Use  $X$  to predict  $Y$  on an independent test set

$$\hat{\mu}(x) = E[Y|X = x]$$

## Unsupervised ML

- ▶ Features  $X$
- ▶ Goals:
  - ▶ Clustering
  - ▶ Dimensionality Reduction
- ▶ “I discovered cats!”



# I discovered Town and Country!

cluster	location	city
17	palo alto high school	palo alto
17	trader joe's (el camino real)	palo alto
17	calafia cafe	palo alto
17	mayfield bakery & cafe	palo alto
17	gotts roadside	palo alto
17	whole foods market (emerson street)	palo alto
17	philz coffee (forest ave.)	palo alto
17	bloomingdale's	palo alto
17	apple retail store (university avenue)	palo alto
17	foothills tennis & swimming club	palo alto
17	lytton gardens community housing	palo alto
17	nest labs	palo alto
17	tin pot creamery (el camino real)	palo alto
17	palo alto clay and glass festival	palo alto
17	blue bottle coffee	palo alto
17	kirks steakburgers	palo alto
17	orens hummus shop (university ave)	palo alto
17	pediatric dentistry of palo alto	palo alto
17	douce france cafe & bakery	palo alto
17	cvs (el camino real)	palo alto
17	nob hill foods (grant road)	mountain view
17	peets coffee (el camino real)	palo alto



# Predictions for Economics

- ▶ Adoption of off-the-shelf ML methods for their intended tasks (prediction, classification, and clustering, e.g. for textual analysis)
- ▶ Extensions and modifications of prediction methods to account for considerations such as fairness, manipulability, and interpretability
- ▶ Development of new econometric methods based on machine learning designed to solve traditional social science estimation tasks, e.g. causal inference
- ▶ Increased emphasis on model robustness and other supplementary analysis to assess credibility of studies
- ▶ Adoption of new methods by empiricists at large scale
- ▶ Revival and new lines of research in productivity and measurement
- ▶ New methods for the design and analysis of large administrative data, including merging these sources
- ▶ Increase in interdisciplinary research
- ▶ Changes in organization, dissemination, and funding of economic research
- ▶ “Economist as engineer” engages with firms, government to design and implement policies in digital environment
- ▶ Design and implementation of digital experimentation, both one-time and as an ongoing process, in collaboration with firms and government
- ▶ Increased use of data analysis in all levels of economics teaching; increase in interdisciplinary data science programs
- ▶ Research on the impact of AI and ML on economy

# What Are Unique Features of Cross-Sectional Econometrics v. Other Branches of Statistics?

- ▶ Framework and language for causality
- ▶ Causal inference from observational data
  - ▶ Theory and PRACTICE
- ▶ Structural models to do counterfactuals for environments that have never been observed
- ▶ Emphasis on interpretable (~causal) models
- ▶ Relatively little emphasis on systematic model selection in applied micro-econometrics
  - ▶ Even in environments where theory does not motivate functional forms
- ▶ Emphasis on standard errors for a pre-specified models
  - ▶ Estimators must have established properties

# What We Say v. What We Do (Econometrics)

## ▶ What We Say

- ▶ Causal inference and counterfactuals
- ▶ God gave us the model
- ▶ We report estimated causal effects and appropriate standard errors
- ▶ Plus a few additional specifications for robustness

## ▶ What we do

- ▶ Run OLS or IV regressions
  - ▶ Try a lot of functional forms
  - ▶ Report standard errors as if we ran only one model
  - ▶ Have research assistants run hundreds of regressions and pick a few “representative” ones
- ▶ Use complex structural models
  - ▶ Make a lot of assumptions without a great way to test them



# Some Broad Generalizations About ML Versus Cross-Sectional Econometrics

- ▶ Guiding principle: prediction
  - ▶ Training, testing
  - ▶ Big concern: overfitting with small data
  - ▶ Also: underfitting with large data
- ▶ Counterfactuals: within current “regime”
  - ▶ If joint distribution among variables changes, just retrain your model
  - ▶ Many argue that predicting for a new stochastic process not justified
- ▶ Some key features
  - ▶ Quality of a predictive algorithm can be summarized in a single number per observation
  - ▶ Can assess performance in a model-free way
- ▶ Relatively small ML literature on causality
  - ▶ “graphical” representations of causal relationships (Judea Pearl)
  - ▶ Reinforcement learning & bandit problems
  - ▶ Little empirical work outside of randomized experiments, no IV or IV analog
  - ▶ If model predicts well in current regime, what more do you need?
- ▶ Relatively little emphasis on statistical properties of estimators or interpretability of models
- ▶ Not historically an empirical field—not about measurement/estimation or about the numbers

# What We Say v. What We Do (ML)

## ▶ What we say

- ▶ ML = Data Science, statistics
  - ▶ Is there anything else?
- ▶ Use language of answering questions or solving problems, e.g. advertising allocation, salesperson prioritization
- ▶ Aesthetic: human analyst does not have to make any choices
- ▶ All that matters is prediction

## ▶ What we do

- ▶ Use predictive models and ignore other considerations, e.g. causality
- ▶ Wonder/worry about interpretability/reliability/robustness/adaptability, but have little way to conceptualize or ask algos to optimize for it
- ▶ Limited conceptual framework for feedback effects, equilibrium, etc.

# Some Lessons for Econometrics: More Emphasis on Validation

- ▶ Model “validation” essential in ML but often neglected in econometrics
  - ▶ To be fair, we are asking harder counterfactual questions
  - ▶ We are using models less prone to “overfitting”
- ▶ Examples in econometrics
  - ▶ Fitting moments that weren’t used for estimation
  - ▶ Testing assumptions of structural models
  - ▶ Meta-studies of merger predictions v. outcomes
  - ▶ Athey/Levin/Seira (QJE), Athey-Coey-Levin (AEJ:Micro) on timber where we estimate on sealed-bid, unrestricted sales and predict to open ascending or small business

# Some Lessons for Econometrics: More Emphasis on Model Selection

- ▶ We don't *really* pick specifications in advance, but we don't emphasize our selection procedures
  - ▶ For larger datasets, really need systematic model selection
    - ▶ Regularized regression, etc.
  - ▶ Robustness
    - ▶ Athey and Imbens, 2015—standard deviation of estimates across models
  - ▶ Supplementary Analysis
    - ▶ See Athey and Imbens 2017 (JEP) for a review
    - ▶ Athey, Imbens, Pham and Wager (2017), etc.
- ▶ Need methods palatable and interpretable for applied research, valid standard errors

# Insights and Applications of the New ML/Causal Inference Literature

- ▶ ML will not solve identification problems, by definition
  - ▶ A parameter is “identified” if you could learn it with an infinite amount of data
  - ▶ ML is about more systematic and exhaustive model selection
- ▶ ML may help analyst be much more systematic about model selection for “predictive part” of models
- ▶ Applications
  - ▶ Better controls for confounding
  - ▶ Personalized/heterogeneous parameter estimates
  - ▶ Personalized policies
  - ▶ Dynamic experimentation (bandits)
- ▶ Example: ATE under unconfoundedness
  - ▶ Environment where treatment is as good as random conditional on a large set of weak confounders
  - ▶ The small data literature has had limited success; different methods and functional forms get very different answers
- ▶ Using ML to systematically search for specifications to control for confounders improves performance
  - ▶ But ONLY if you modify the objective!!


# The Potential Outcome Setup for Causal Inference

For a set of i.i.d. subjects  $i = 1, \dots, n$ , we observe a tuple  $(X_i, Y_i, W_i)$ , comprised of

- ▶ A **feature vector**  $X_i \in \mathbb{R}^p$ ,
  - ▶ A **response**  $Y_i \in \mathbb{R}$ , and
  - ▶ A **treatment assignment**  $W_i \in \{0, 1\}$ .
- 
- ▶ Define the **average treatment effect (ATE)**, the **average treatment effect on the treated (ATT)**

$$\tau = \tau^{\text{ATE}} = \mathbb{E} \left[ Y^{(1)} - Y^{(0)} \right]; \tau^{\text{ATT}} = \mathbb{E} \left[ Y^{(1)} - Y^{(0)} \mid W_i = 1 \right];$$

- ▶ and, the **conditional average treatment effect (CATE)**

$$\tau(x) = \mathbb{E} \left[ Y^{(1)} - Y^{(0)} \mid X = x \right].$$


# ML and Causal Inference: Average Treatment Effects Under Unconfoundedness

- ▶ Focusing on **prediction** only using off-the-shelf ML leads to bias
  - ▶ Off-the-shelf:
    - ▶ Regress  $Y$  on  $W$  and  $X$  using, e.g., LASSO
  - ▶ We know we need to control for confounders to eliminate bias
  - ▶ Focusing on prediction “zero’s out” confounders with weak effect on outcomes, even if they are confounders
  - ▶ Belloni, Chernozukov, and Hansen (series of papers)
- ▶ Use LASSO as a variable selection method
  - ▶  $Y$  on  $X$
  - ▶  $W$  on  $X$
  - ▶ OLS of  $Y$  on  $W$ , union of selected  $X$ ’s
  - ▶ Early example to show that Prediction and ML should have different objectives!
- ▶ Estimating propensity scores/assignment model neither necessary or a good idea
  - ▶ Assignment models often complex
  - ▶ Hard to estimate accurately in high dimensions
  - ▶ Focus directly on covariate balance
  - ▶ Athey, Imbens and Wager (2016) method does not rely on estimable propensity score
- ▶ Orthogonalization helps
  - ▶ Both BCH and AIW approaches rely on residualization
  - ▶ Hard to estimate high-dimensional models accurately
- ▶ Residual on Residual regression using ML – Chernozhukov et al (2017)

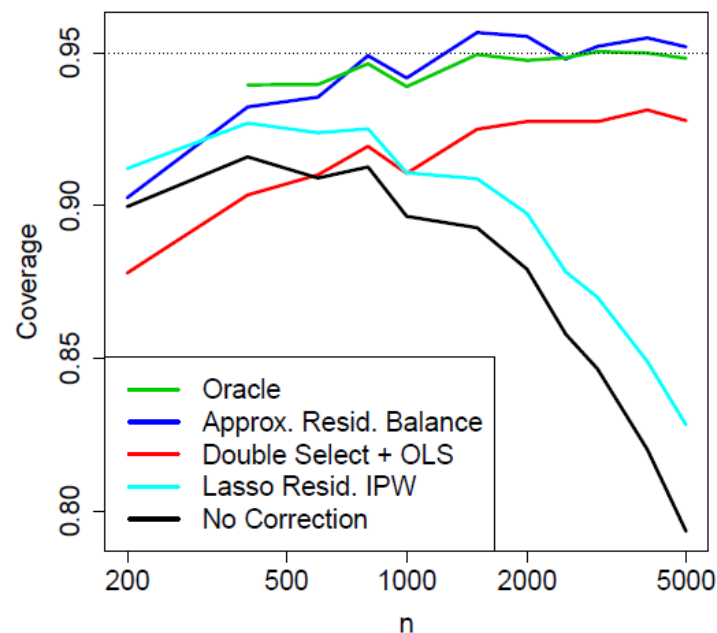
# Conclusions for ATE Under Unconfoundedness

- ▶ ML-based methods systematically improve over traditional methods in simulations and empirical examples
- ▶ Which ML-based method depends on attributes of problem
  - ▶ AIW's residual balancing works well when treatment allocation function is complex and nonlinear
  - ▶ LASSO models work well when environment is simpler and sparse
  - ▶ Double Machine Learning allows a range of ML methods that can be selected based on the applications
- ▶ Broader insight: Pay special attention to causal elements and considerations, and use ML for predictive parts



Data from the California GAIN Program, as in Hotz et al. (2006).

- ▶ Program separately randomized in: Riverside, Alameda, Los Angeles, San Diego.
- ▶ Outcome: mean earnings over next 3 years.
- ▶ We hide county information. Seek to compensate with  $p = 93$  controls.
- ▶ Full dataset has  $n = 19170$ .



# Difference in Difference, Panel Data

- ▶ **Key task in DID:**
  - ▶ *Predict* what would have happened to treatment units if they had not been treated
- ▶ **Doudchenko and Imbens (2017)**
  - ▶ Regularized regression for Synthetic Control
- ▶ **Bai: Analysis of latent factor models**
- ▶ **Athey, Bayati, Doudchenko, Imbens, Khosravi (2017)**
  - ▶ Fit a matrix to panel data with penalization for “complexity”, building and extending recent ML methods
  - ▶ Find general cross-sectional and time series patterns
  - ▶ Works with “wide” or “narrow” data
  - ▶ Observation: estimating what would have happened in the absence of the treatment is a prediction problem
  - ▶ Improves on existing methods when there is information in both cross-sectional and time series patterns

# Heterogeneous Treatment Effects: Experiments, Unconfoundedness, IV, GMM

- ▶ Estimating heterogeneity with limited complexity
  - ▶ Causal Tree (Athey and Imbens, PNAS 2016)
    - ▶ Tailored objective, std errors
    - ▶ Sample splitting
    - ▶ Many applications from health to field experiments
  - ▶ Trees with GMM/ML Models
    - ▶ Zeileis (2008)
    - ▶ Asher, Nekipelov, Novosad, Ryan (2016)
    - ▶ Athey, Tibshirani, and Wager (2016)
  - ▶ LASSO
    - ▶ “Interpretability”? Arguably harder than trees when omitted variables.
    - ▶ E.g. Imai and Ratkovic, 2013
  - ▶ “Deep IV”
    - ▶ Matt Taddy, Greg Lewis et al (2017)
- ▶ Non-parametric estimation
  - ▶  $\hat{\tau}(x) = E[\tau_i | X_i = x]$
  - ▶ This is a hard problem!
- ▶ Forest-based methods
  - ▶ Wager and Athey (2015) provide first asymptotic normality results, confidence intervals
  - ▶ Athey, Tibshirani, and Wager (2016) – any GMM model, e.g. IV, with confidence intervals
  - ▶ Use forests to generate weights
  - ▶ Forests replace kernels wherever they are used
- ▶ “Deep IV”
  - ▶ Matt Taddy, Greg Lewis et al (2017)

# Heterogeneous Treatment Effects in Medicine

- ▶ “Targeting weight loss interventions to reduce cardiovascular complications of type 2 diabetes: a machine learning-based post-hoc analysis of heterogeneous treatment effects in the Look AHEAD trial” Baum et al, *Lancet*, July 2017.
- ▶ The Action for Health in [Diabetes](#) (Look AHEAD) trial investigated whether long-term [cardiovascular](#) disease morbidity and mortality could be reduced through a weight loss intervention among people with type 2 diabetes. Despite finding no significant reduction in cardiovascular events on average, it is possible that some subpopulations might have derived benefit. In this [post-hoc analysis](#), we test the hypothesis that the overall neutral [average treatment effect](#) in the trial masked important heterogeneous treatment effects (HTEs) from intensive weight loss interventions.
- ▶ We used causal forest modelling, which identifies HTEs, using a random half of the trial data (the training set). We applied Cox proportional hazards models to test the potential HTEs on the remaining half of the data (the testing set).
- ▶ Look AHEAD participants with moderately or poorly controlled diabetes (HbA1c 6·8% or higher) and subjects with well controlled diabetes (HbA1c less than 6·8%) and good self-reported health (85% of the overall study population) averted cardiovascular events from a behavioural intervention aimed at weight loss. However, 15% of participants with well controlled diabetes and poor self-reported general health experienced negative effects that rendered the overall study outcome neutral. HbA1c and a short questionnaire on general health might identify people with type 2 diabetes likely to derive benefit from an intensive lifestyle intervention aimed at weight loss.

## Application: General Social Survey

The General Social Survey is an extensive survey, collected since 1972, that seeks to measure demographics, political views, social attitudes, etc. of the U.S. population.

Of particular interest to us is a **randomized experiment**, for which we have data between 1986 and 2010.

- ▶ **Question A:** Are we spending too much, too little, or about the right amount on **welfare**?
- ▶ **Question B:** Are we spending too much, too little, or about the right amount on **assistance to the poor**?

**Treatment effect:** how much less likely are people to answer **too much** to question B than to question A.

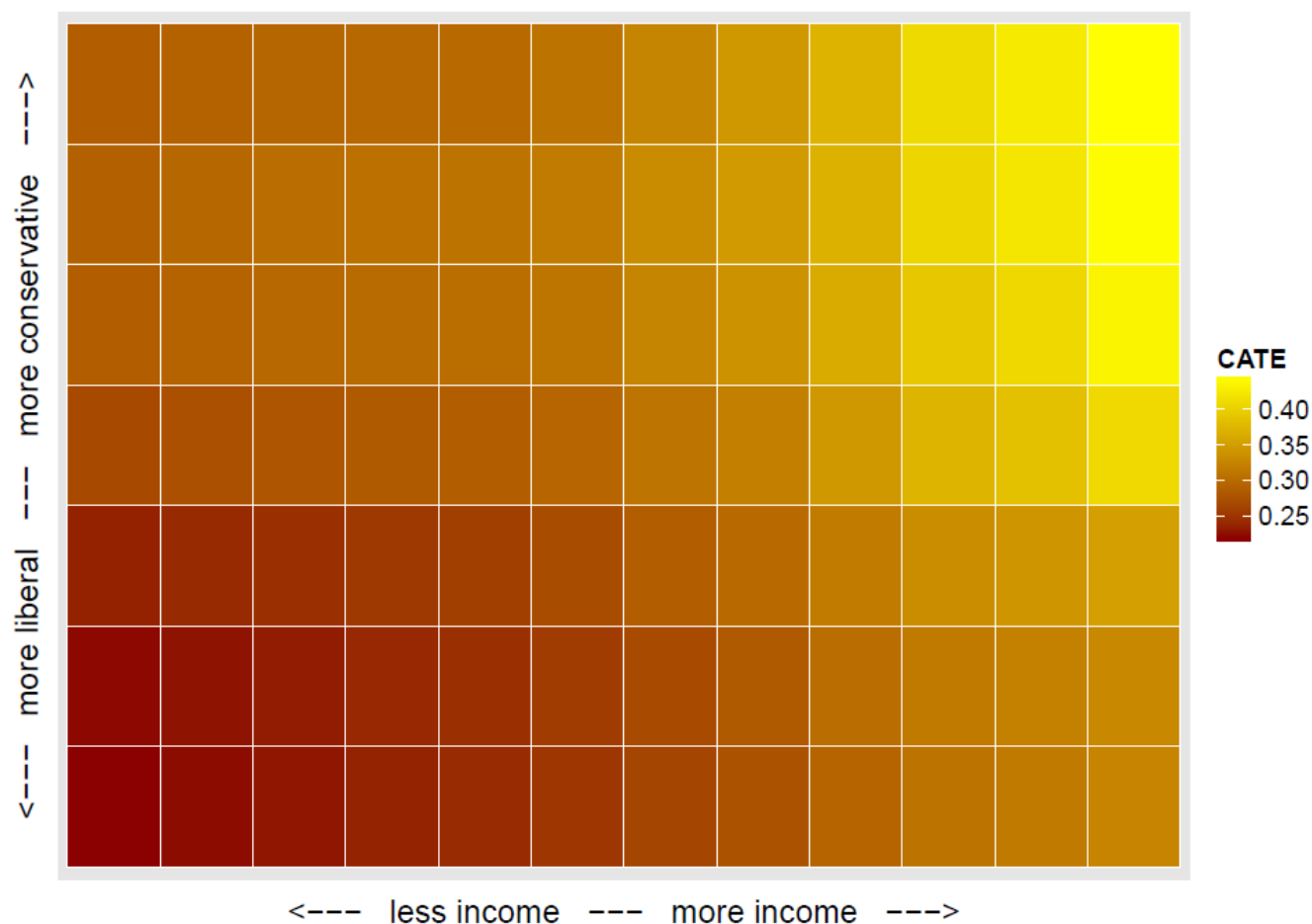
- ▶ We want to understand how the treatment effect depends on **covariates**: political views, income, age, hours worked, ...

**NB:** This dataset has also been analyzed by Green and Kern (2012) using Bayesian additive regression trees (Chipman, George, and McCulloch, 2010).



# Application: General Social Survey

A causal forest analysis uncovers **strong treatment heterogeneity** ( $n = 28,686$ ,  $p = 12$ ).



## Empirical Application: Family Size

Angrist and Evans (1998) study the effect of family size on women's labor market outcomes. Understanding heterogeneity can guide policy.

- ▶ Outcomes: participation, female income, hours worked, etc.
- ▶ Treatment: more than two kids
- ▶ Instrument: first two kids same sex
- ▶ First stage effect of same sex on more than two kids: .06
- ▶ Reduced form effect of same sex on probability of work, income: .008, \$132
- ▶ LATE estimates of effect of kids on probability of work, income: .133, \$2200



## Empirical Application: Family Size

Angrist and Evans (1998) study the effect of family size on women's labor market outcomes. Understanding heterogeneity can guide policy.

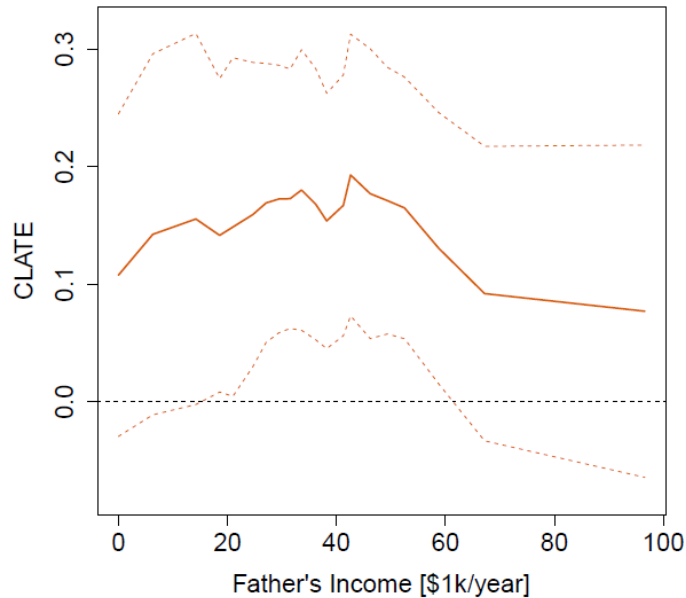
- ▶ Outcomes: participation, female income, hours worked, etc.
- ▶ Treatment: more than two kids
- ▶ Instrument: first two kids same sex
- ▶ First stage effect of same sex on more than two kids: .06
- ▶ Reduced form effect of same sex on probability of work, income: .008, \$132
- ▶ LATE estimates of effect of kids on probability of work, income: .133, \$2200



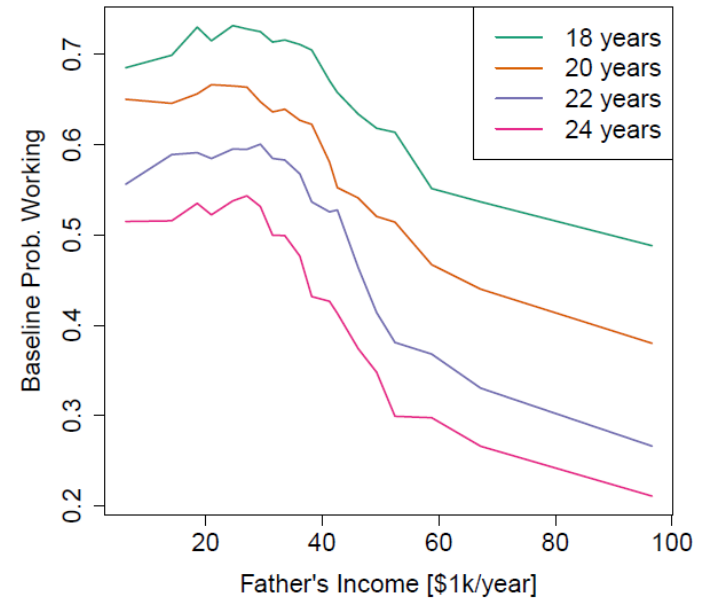


# Treatment Effects: Magnitude of Decline

## Effect on Participation



## Baseline Probability of Working



# Optimal Policy Estimation

- ▶ E.g. personalized medicine
  - ▶ Estimate policy mapping from covariates to treatment.  $\pi: X \rightarrow W$
- ▶ A variety of approaches from ML literature
  - ▶ Imports ideas from causal inference literature such as propensity score weighting
  - ▶ Little attention to econometric efficiency
- ▶ Kitagawa and Tetenov (forthcoming, EMA)
- ▶ Athey and Wager (2017)
  - ▶ Improve the performance bringing in orthogonalization and ideas from econometric efficiency
- ▶ Bandits & Contextual Bandits
  - ▶ Steve Scott (Google)
  - ▶ John Langford team (MSR)
  - ▶ Eytan Bakshy team (Facebook)
  - ▶ Athey et al (methods & applications in progress... stay tuned)

# Some Lessons for Econometrics: Large Scale Bayesian Models

- ▶ ML & Econometrics closest when we do Bayesian statistics
- ▶ ML has well-developed literature on large scale
- ▶ Athey-Nekipelov (2014) – advertisers with heterogeneous preferences in search
- ▶ David Blei et al techniques
- ▶ Use matrix factorization for consumer demand systems with aggregated (Taddy et al 2017) or individual discrete choice (Athey et al (2017))