SUSAN ATHEY

THE ECONOMICS OF TECHNOLOGY PROFESSOR, STANFORD GSB

The Impact of Machine Learning on Economics and the Economy

Two-day course on machine learning and causal inference with videos and scripts: https://www.aeaweb.org/conference/cont-ed/2018-webcasts

Survey paper: https://www.nber.org/chapters/c14009.pdf

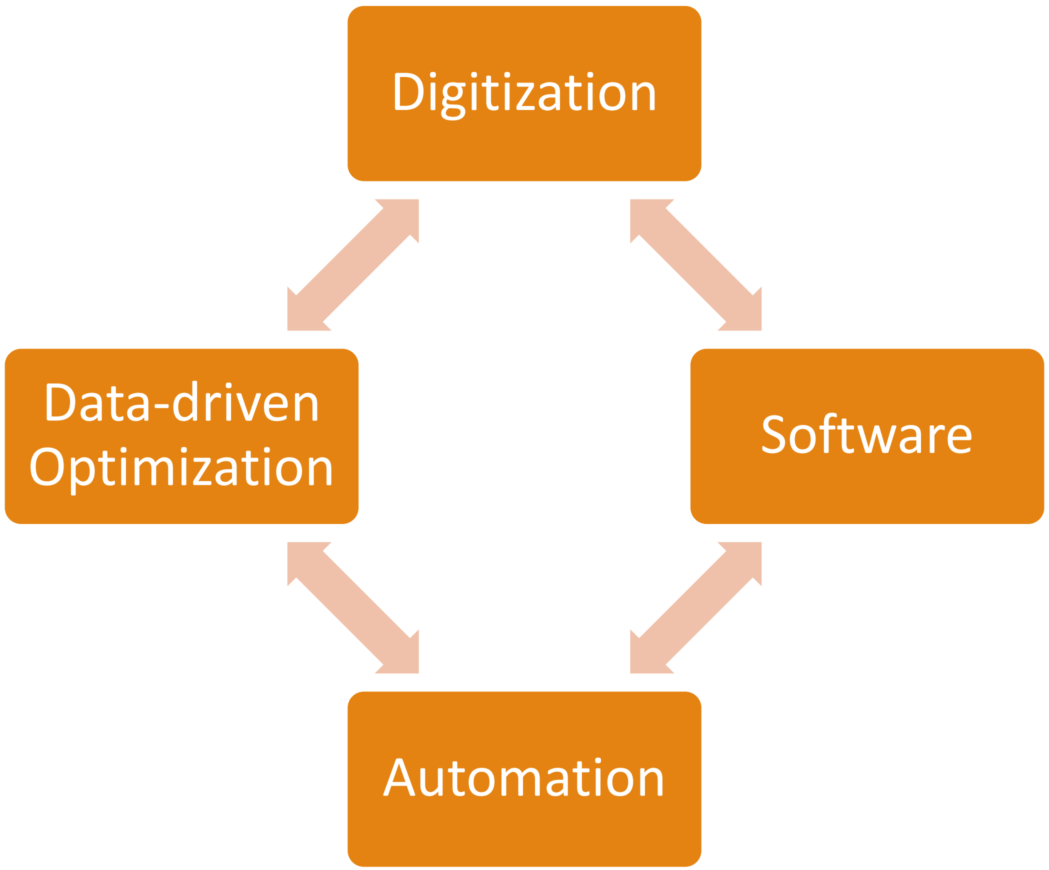
Links to papers: https://athey.people.stanford.edu/research

Software is Eating The World
Every Company is a Tech Company

"My own theory is that we are in the middle of a dramatic and broad technological and economic shift in which software companies are poised to take over large swathes of the economy.

More and more major businesses and industries are being run on software and delivered as online services—from movies to agriculture to national defense."

-Marc Andreessen (2013)



Machine Learning

Advances in Supervised ML dramatically improve quality of image classification

Supervised Machine Learning

Labelled data (X,Y)

Objective: use X to predict Y in a test set

Used to classify images without using any structure or prior knowledge

X_i

$$\Pr(Y_i = CAT|X_i) = .95$$
$$\Pr(Y_i = DOG|X_i) = .05$$

What's New About ML?

Flexible, rich, data-driven **algorithms** select from a family of models to optimize **goodness of fit**

Computational tricks/engineering

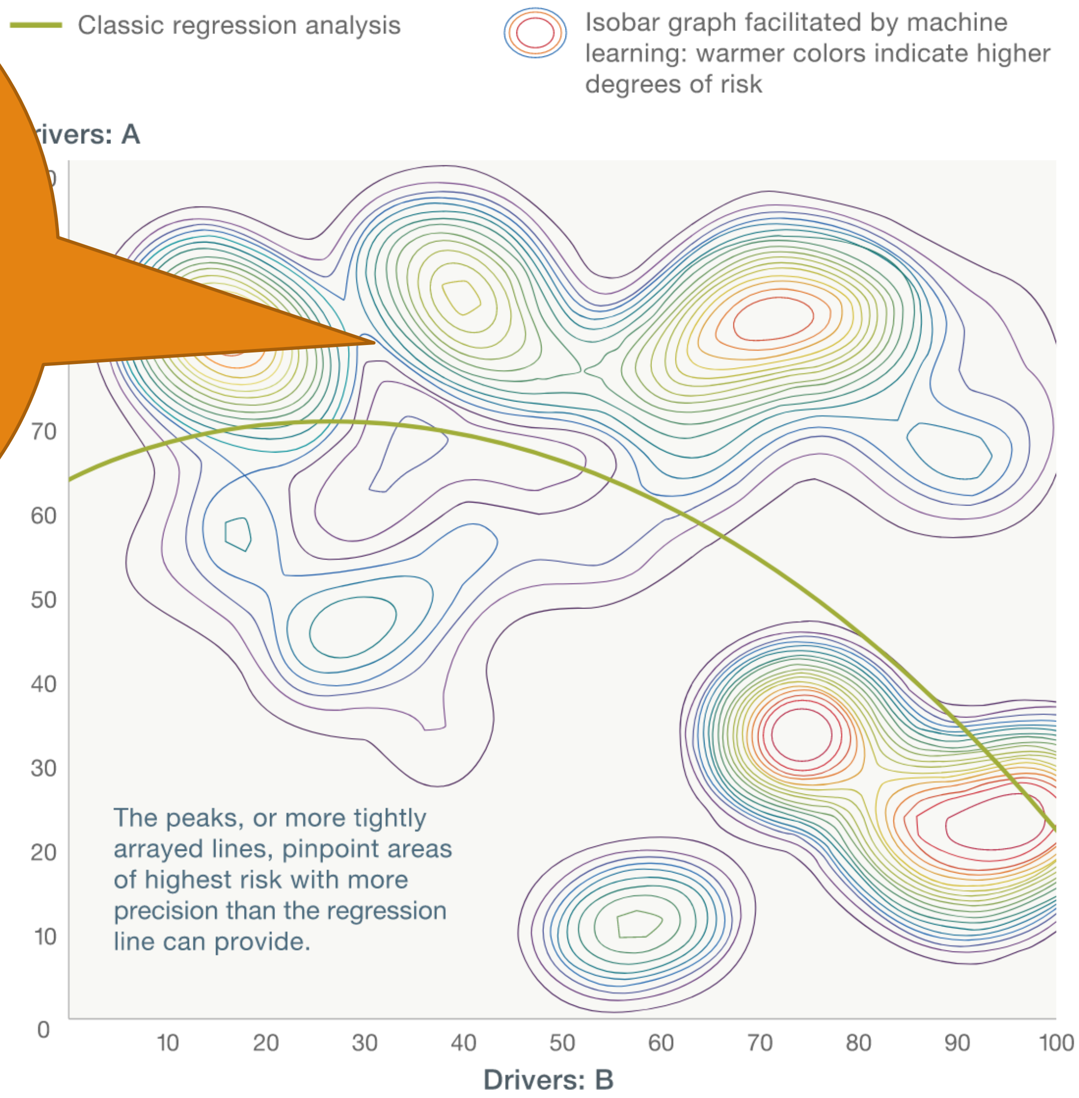
Methods (e.g. cross-validation) to avoid over-fitting

Increase in personalization and precision

Do we really think this relationship is plausible? Stable, robust, causal?

The contrast between routine statistical analysis and data generated by machine learning can be quite stark.

Value at risk from customer churn, telecom example

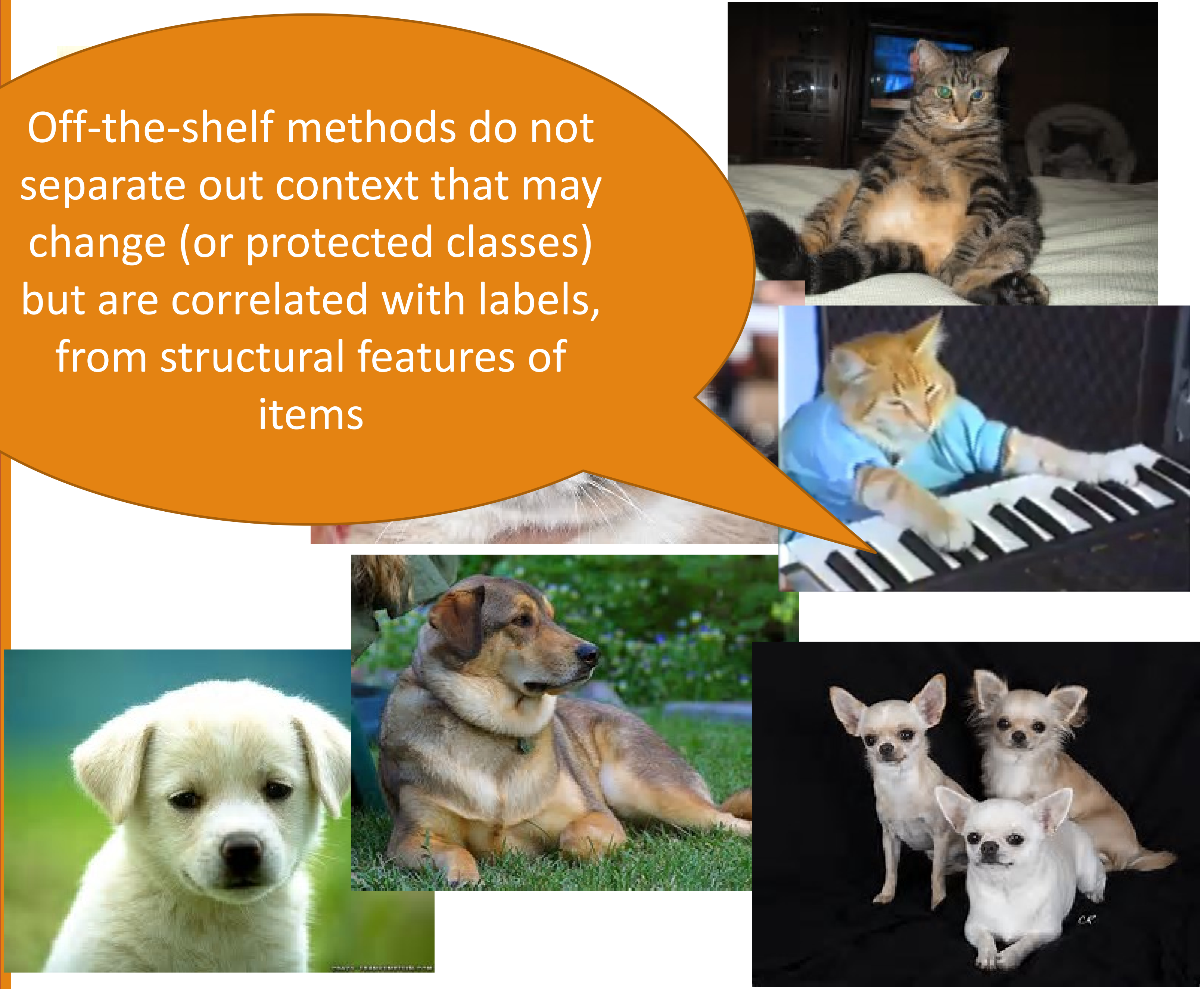


McKinsey&Company

Machine Learning and AI

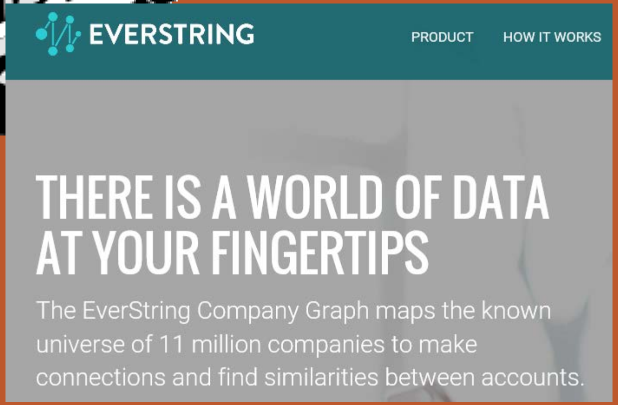
Advances in ML dramatically improve quality of image classification

Off-the-shelf methods do not separate out context that may change (or protected classes) but are correlated with labels, from structural features of items



Applications of Prediction Across Industries

NYC Firecast



Text and image recognition as input to other processes

Risk scoring/decision support

Threat detection/content moderation

Prioritization of resources

- Sales calls
- Advertising
- Auditing
- City inspections
- Restaurant hygiene

Monitoring workers

- Video/voice
- Mobile phones

Identifying or reducing discrimination

- Hiring
- Justice

Application: Monitoring and Incentives

Marketplaces need to provide incentives and screen for quality

Ratings are noisy, often missing and biased, uncomfortable and time consuming for customers

Alternative: direct monitoring and feedback to sellers

Approaches

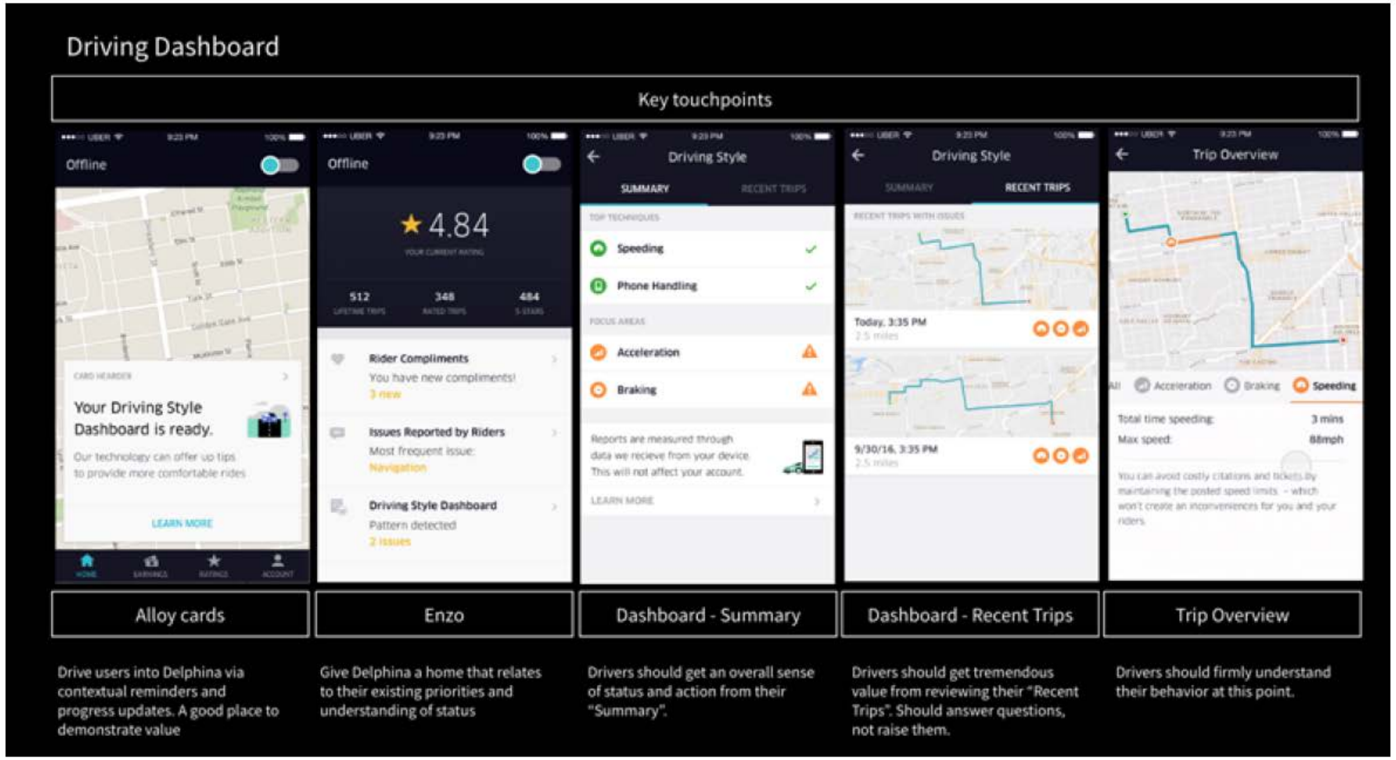
Gather data passively

Gather customer satisfaction data from a sample, or passively from customer behavior

Train a model to estimate quality of service

Provide feedback and coaching to seller, require training, explicit incentives

Nudging Drivers to Better Performance

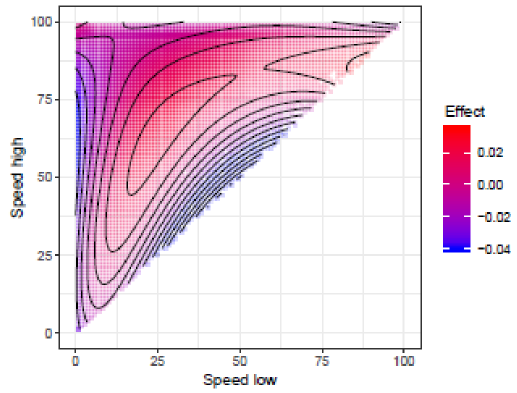


Experiment:

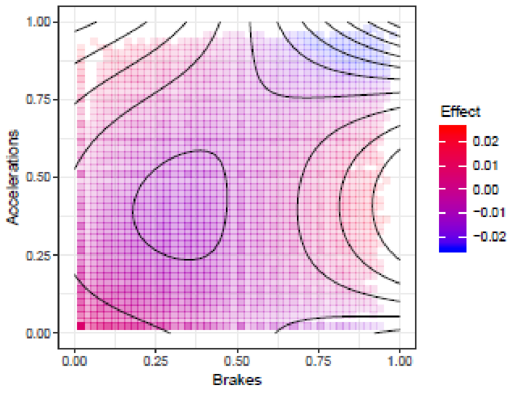
- Randomly select drivers have access to app
- Small effect improving driver safety on average
- Much larger effect for drivers whose performance was poor prior to experiment

Monitoring Workers or Service Providers for Quality: UberX drivers provide higher quality than taxi's

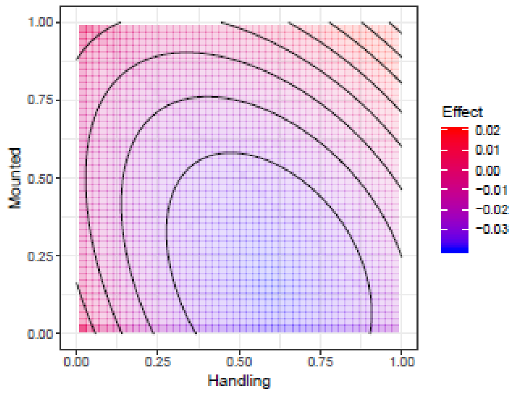
Predicted Star Ratings as a Function of Telematics



(a) Speed metrics



(b) Acceleration and brake metrics



(c) Cell phone use metrics

Experimental Estimates of Informational Nudges

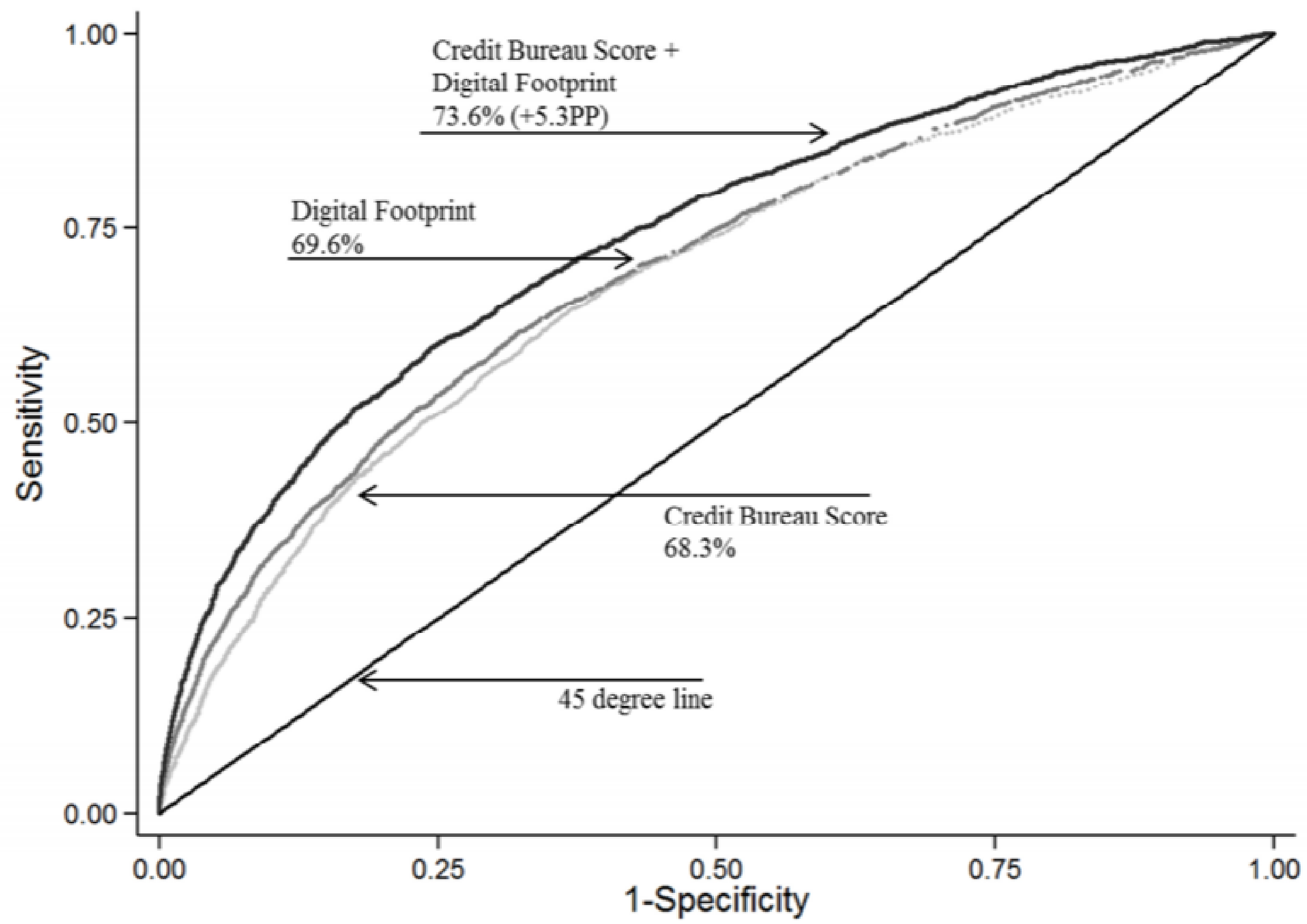
	Dependent variable:		
	Score F	Score S	Score NS
	(1)	(2)	(3)
	Panel A: Intent to treat estimator		
Bottom 10th Perc. Before	−0.0271***	−0.0082***	−0.0255***
	(0.0005)	(0.0004)	(0.0004)
Treatment x Not Bottom 10th Perc.	0.0001	0.0001	0.00004
	(0.0002)	(0.0001)	(0.0001)
Treat x Bottom 10th Perc.	0.0015**	0.0006	0.0014**
	(0.0006)	(0.0005)	(0.0005)
Observations	4,254,109	4,254,109	4,254,109
	Panel B: 2SLS estimator		
Bottom 10th Perc. Before	−0.0008*	−0.0005**	0.0002
	(0.0005)	(0.0002)	(0.0004)
App Int. x Not Bottom 10th Perc.	0.0003	0.0001	0.0002
	(0.0002)	(0.0001)	(0.0002)
App Int. x Bottom 10th Perc.	0.0028***	0.0007	0.0027***
	(0.0010)	(0.0005)	(0.0009)
Observations	4,254,109	4,254,109	4,254,109

Using digital footprints for credit scoring

"On the Rise of FinTechs – Credit Scoring Using Digital Footprints," Berg, Burg, Gombovic, Puri, forthcoming

Figure 3: AUC (Area Under Curve) for scorable customers for various model specifications

This figure illustrates the discriminatory power of three different model specifications by providing the rec operating characteristics curve (ROC-curve) and the area under curve (AUC). The ROC-curves are estimated usi logit regression of the default dummy on the credit bureau score (light gray), the digital footprint (gray), both c bureau score and digital footprint (dark gray). The sample only includes customers with credit bureau scores. sample period is from October 19, 2015 to December 2016. For variable definitions see Appendix Table 1.



Using digital footprints for credit scoring

- Manipulability
- Stability

DRIVERS' E-FAIL Admiral hikes insurance costs for drivers using Hotmail email addresses

It follows our story yesterday on how insurers charge drivers called Mohammed more

EXCLUSIVE

Katie Hodge | Ben Leo

23 Jan 2018, 0:01 | Updated: 23 Jan 2018, 20:25

2 COMMENTS

CAR insurer Admiral last night admitted hiking premiums for drivers applying via Hotmail.

Challenges for Management/ Regulation of ML in Financial Services

Algorithms have demonstrable errors

Engineers build black-box algorithms, but are not trained to evaluate

Need “best practices” to analyze the black box

Credit Scoring Example

- **Instability** of joint distribution of outcomes, novel features
- **Poor performance** when extrapolating
- **Manipulation** of novel features
- **Discrimination** and **Fairness**
- Ever-changing **adverse selection** problem as competing firms change models, marketing strategies
- When are results more or less **reliable**?

Equilibrium effects

- Agents using ML interact
- Collusion (airline prices)
- Instability (financial market crashes, correlated mistakes across firms)
- Google maps examples

Need models of individual behavior and eqm selection to study eqm changes

- Why existing AI/ML is a long way from solving “harder” problems

Policy and Productivity of the Financial Sector

Financial services have great potential for application of ML/AI

Need regulatory policy that seeks efficiency

Fraud and cybersecurity
- Great application of ML/AI
- Cat and mouse game
- Economics of attacks and prevention: public good

Regulating processes v. regulating outcomes
- Black box makes process regulation obsolete
- Discrimination measured by outcomes not inputs
- Need value judgements, cost-benefit analysis

Exposure for firms to document risks, processes

Labor displacement and retraining

The Value of Data, Productivity, and Industry Structure for AI/ML

AI/ML performs better with more data

How much depends on circumstances

Can Europe be a full participant in the AI revolution?

International differences

- Population/market size (China > U.S. > Europe)
- Privacy policies
- Industrial policy

Scale economies in AI

- At the firm level or the market level?
- Cloud computing and shared services in principle bring to market level

General purpose technology

- Technology is fairly straightforward, open; innovations diffuse quickly
- Domain-specific know-how, data, active users vary

Artificial Intelligence/Machine Learning Desired Properties for Applications

DESIRED PROPERTIES

Interpretability

Stability/Robustness

Transferability

Fairness/Non-discrimination

"Human-like" decision-making

- Reasonable decisions in never-experienced situations

CAUSAL INFERENCE FRAMEWORK

Goal: learn model of how the world works

- Impact of interventions can be context-specific
- Model maps contexts and interventions to outcomes
- Formal language to separate out correlates and causes

Ideal causal model is by definition stable, interpretable

Transferability: straightforward for new context dist'n

- If you estimate treatment effect heterogeneity

Fairness: Many aspects of algorithmic discrimination relate to correlation v. causation

- Gender and race may be correlated with factors that shift distributions of characteristics like test scores or credit scores, relatively limited direct causal effects

ML and Econometrics

Causal inference vs. Supervised ML

Supervised learning:

- Can evaluate in test set in model-free way

$$\text{MSE: } \sum (Y_i - \hat{\mu}(X_i))^2$$

Causal inference

- Objective: unbiased/consistent parameter estimation
- Parameters of interest not observed in test set
- Can estimate objective (MSE of parameter), but requires maintained assumptions, often not model-free

$$\textbf{Infeasible MSE: } \sum \left(\theta_i - \hat{\theta}(X_i)\right)^2$$

- Tune for counterfactuals: distinct from tuning for fit, also different counterfactuals select different models
- Theoretical assumptions, domain knowledge
- Sampling variation matters even in large data sets
  - Statistical theory and inference play important roles

Causal Inference Approaches

"Program evaluation", "Treatment effect estimation"

For each

Estimand X Design

New **ML-based method,** theory, confidence intervals

Goal: **estimate the causal impact** of interventions or treatment assignment policies

- Low dimensional intervention
- Desire **confidence intervals**

Estimands

- Average effect
- Heterogeneous effects
- Optimal policy

Designs that enable identification and estimation of these effects

- Randomized experiments
- Unconfoundedness
- "Natural" experiments (IV)
- Regression discontinuity
- Difference-in-difference
- Longitudinal data
- Randomized and natural experiments in social network/settings w/ interference

My own work on ML/Causal Inference

Pitfalls of Pure Prediction

- “Beyond Prediction: Using Big Data for Policy Problems,” *Science*, 2017
- “The Impact of Machine Learning on Economics,” *The Economics of Artificial Intelligence*

Stable/robust prediction and estimation

- “Stable Prediction across Unknown Environments,” (with Kun Kuang, Ruoxuan Xiong, Peng Cui, Bo Li), *Knowledge Discovery & Data Mining*, 2018.
- “Estimating Average Treatment Effects: Supplementary Analyses and Remaining Challenges,” (with Guido Imbens, Thai Pham, and Stefan Wager), *American Economic Review*, May 2017
- “A Measure of Robustness to Misspecification” (with Guido Imbens), *American Economic Review*, May 2015, 105 (5), 476-480

Surrogates

- “Estimating Treatment Effects using Multiple Surrogates: The Role of the Surrogate Score and the Surrogate Index” (with Raj Chetty, Guido Imbens, Hyunseung Kang), 2016

Combining ML and Structural Models of Consumer Behavior

- “Estimating Heterogeneous Consumer Preferences for Restaurants and Travel Time Using Mobile Location Data,” (with David Blei, Robert Donnelly, Francisco Ruiz, and Tobias Schmidt), *American Economic Review Papers and Proceedings*, May, 2018
- “SHOPPER: A Probabilistic Model of Consumer Choice with Substitutes and Complements,” 2017, (with Francisco Ruiz and David Blei).
- “Counterfactual Inference for Consumer Choice Across Many Product Categories” (with David Blei, Rob Donnelly, Francisco Ruiz)

Causal Panel Data Models

- Athey, Bayati, Duodechenko, Khosravi, Imbens: “Matrix Completion Methods for Causal Panel Data Models” 2018
- Arkhangelsky, Athey, Hirschberg, Imbens, Wager: “Synthetic Difference in Differences” 2018
- Johannemann, Hadad, Athey, Wager: “Sufficient Representations for Categorical Variables”

Treatment Effects, Assignment Policies

- “Recursive Partitioning for Heterogeneous Causal Effects” (with Guido Imbens), *PNAS* 2016
- “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests” (with Stefan Wager), *Journal of the American Statistical Association*, 2018.
- “Generalized Random Forests,” with Julie Tibshirani and Stefan Wager, *Annals of Statistics,* 2019.
- “Efficient Policy Learning,” with Stefan Wager, 2017.
- “Offline Multi-Action Policy Learning: Generalization and Optimization,” (with Zhengyuan Zhou and Stefan Wager)
- “Local Linear Forests,” (with Rina Friedberg, Julie Tibshirani, and Stefan Wager), 2018.

Contextual Bandits

- “Balanced Linear Contextual Bandits,” with Maria Dimakopoulou, Zhengyuan Zhou, and Guido Imbens, *Association for the Advancement of Artificial Intelligence (AAAI),* forthcoming.

Generative Adversarial Networks

- “Using Wasserstein Generative Adversial Networks for the Design of Monte Carlo Simulations” with Guido Imbens, Jonas Metzger, Evan Munro

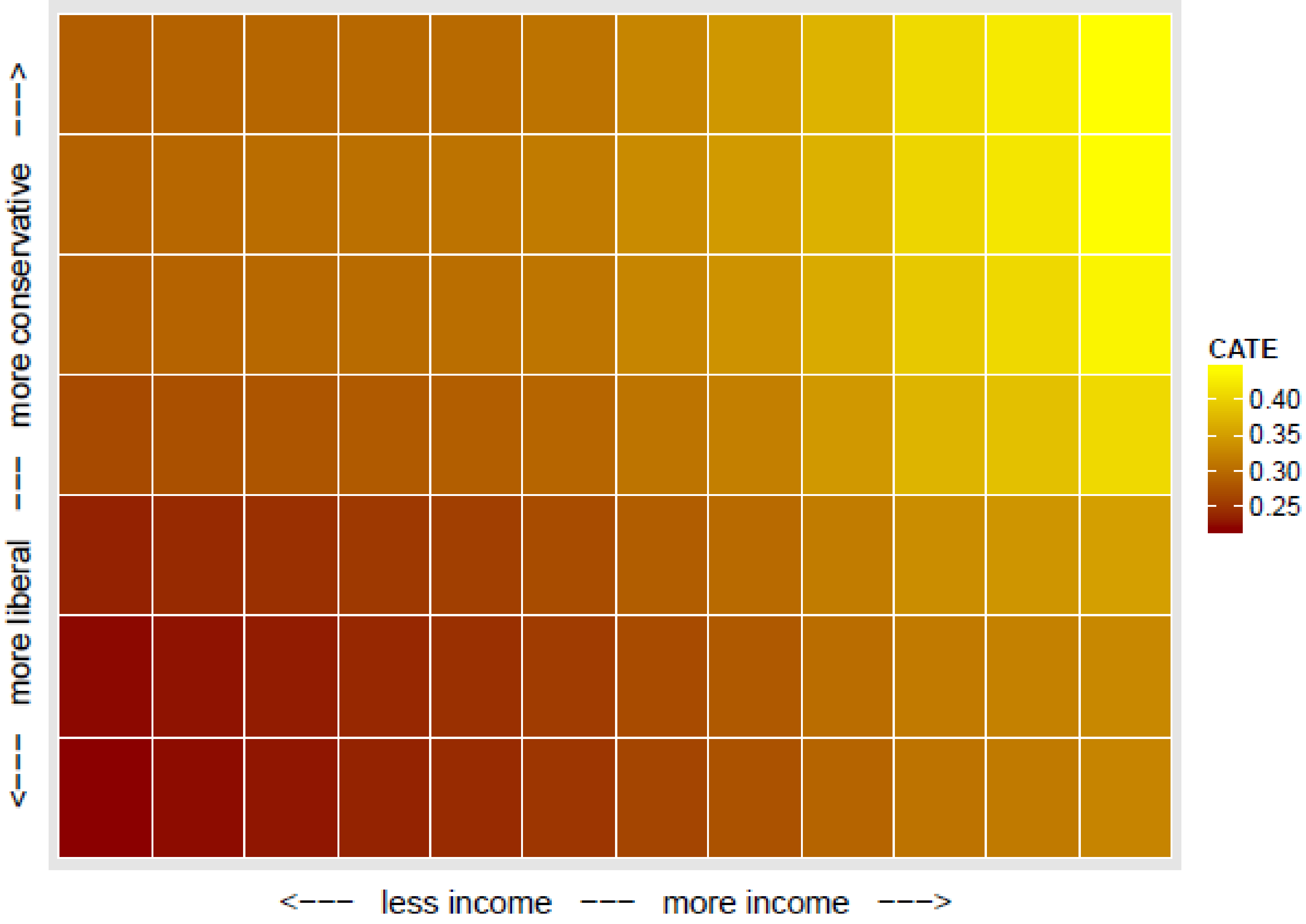
General Social Survey Experiment

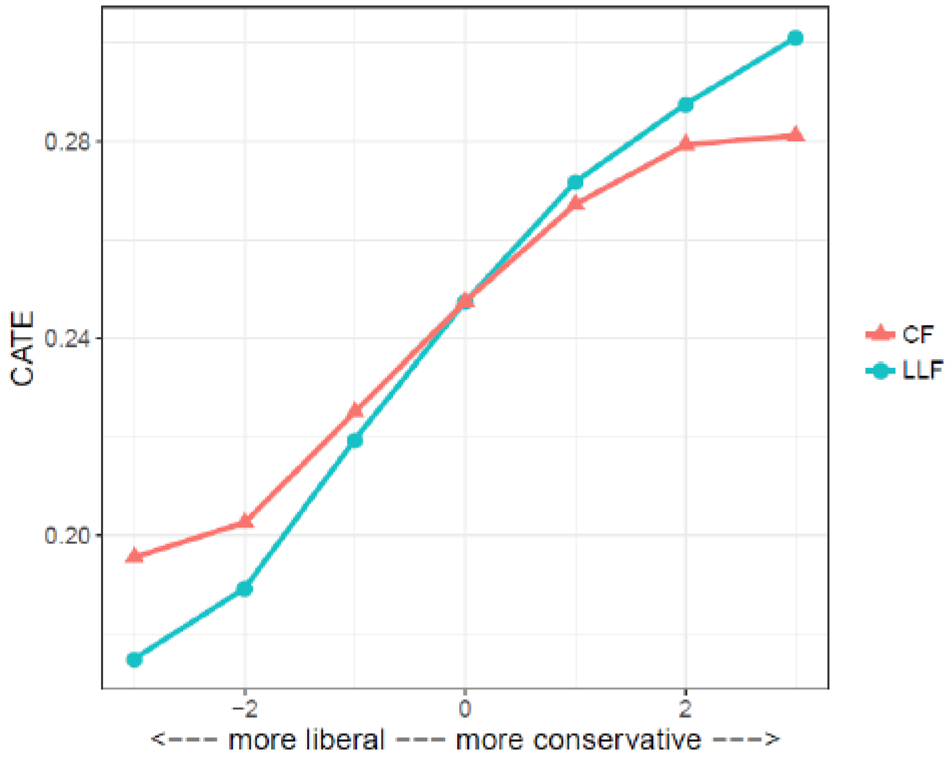
Are you in favor of

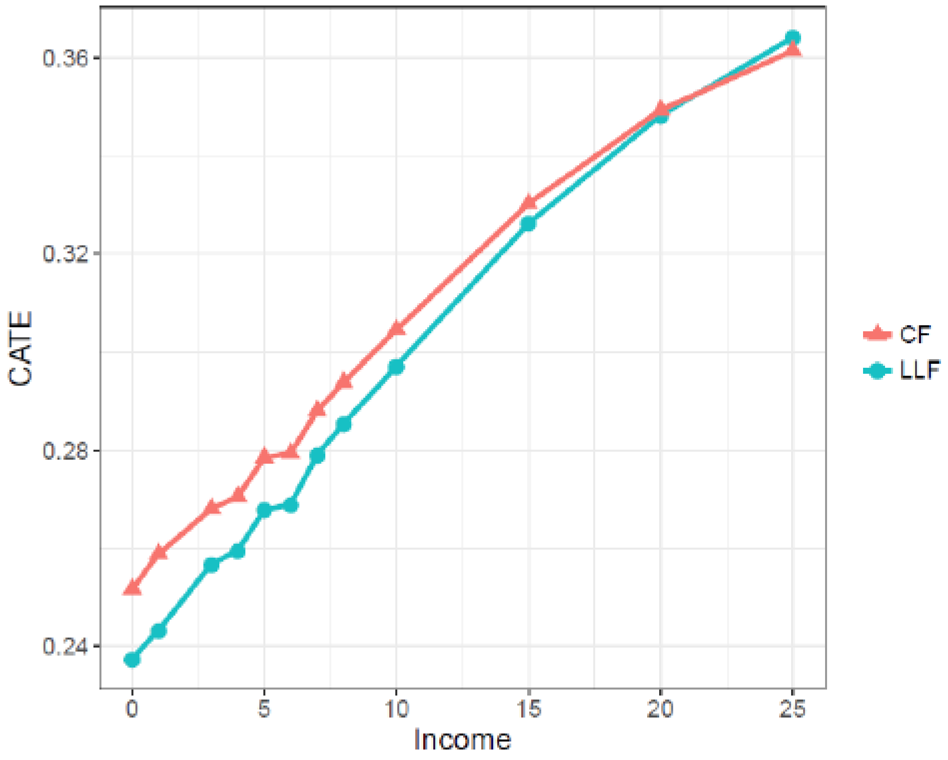
> “**Assistance to the poor**” vs. “**Welfare**”

Data-driven search for heterogeneity; confidence intervals

Methods: Causal forest (Wager and Athey (JASA 2018), Athey, Tibshirani, and Wager (AOS forthcoming))







Causal forest v. Local linear forest (Friedberg, Athey, Tibshirani and Wager (2018))

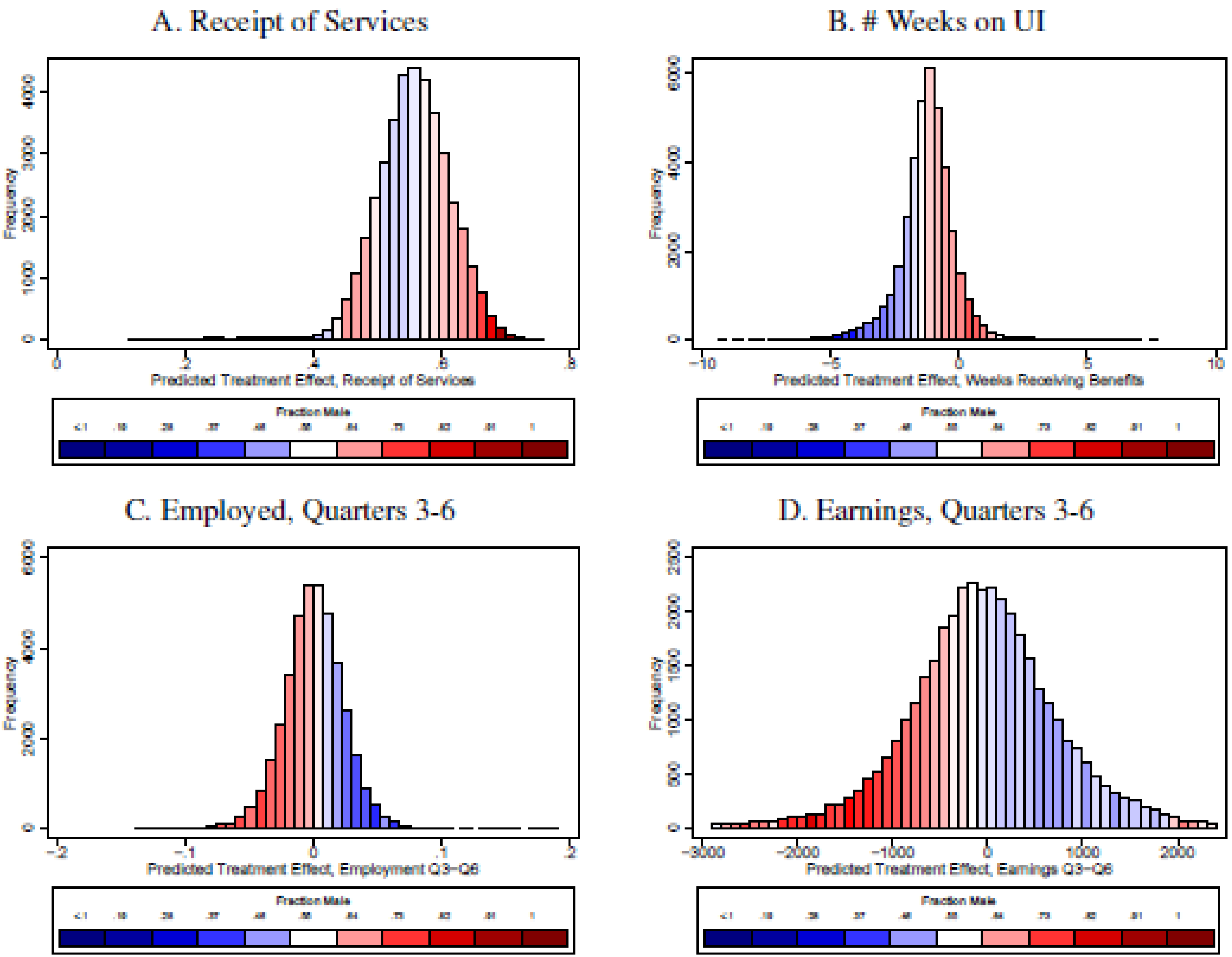
Improve ML methods bringing in ideas from stats/econ (bias correction at boundaries) and allow modeling mixed structure (linear effects and more complex interactions)

Machine Learning Examples

Using “causal forests” (Wager and Athey, 2018; Athey, Tibshirani and Wager, 2018) to estimate heterogeneous treatment effects from training program

Athey, Campbell, Chyn, Hastings and White (in progress) using data from RIPL

Figure 2: Distribution of Predicted Treatment Effects with Gender

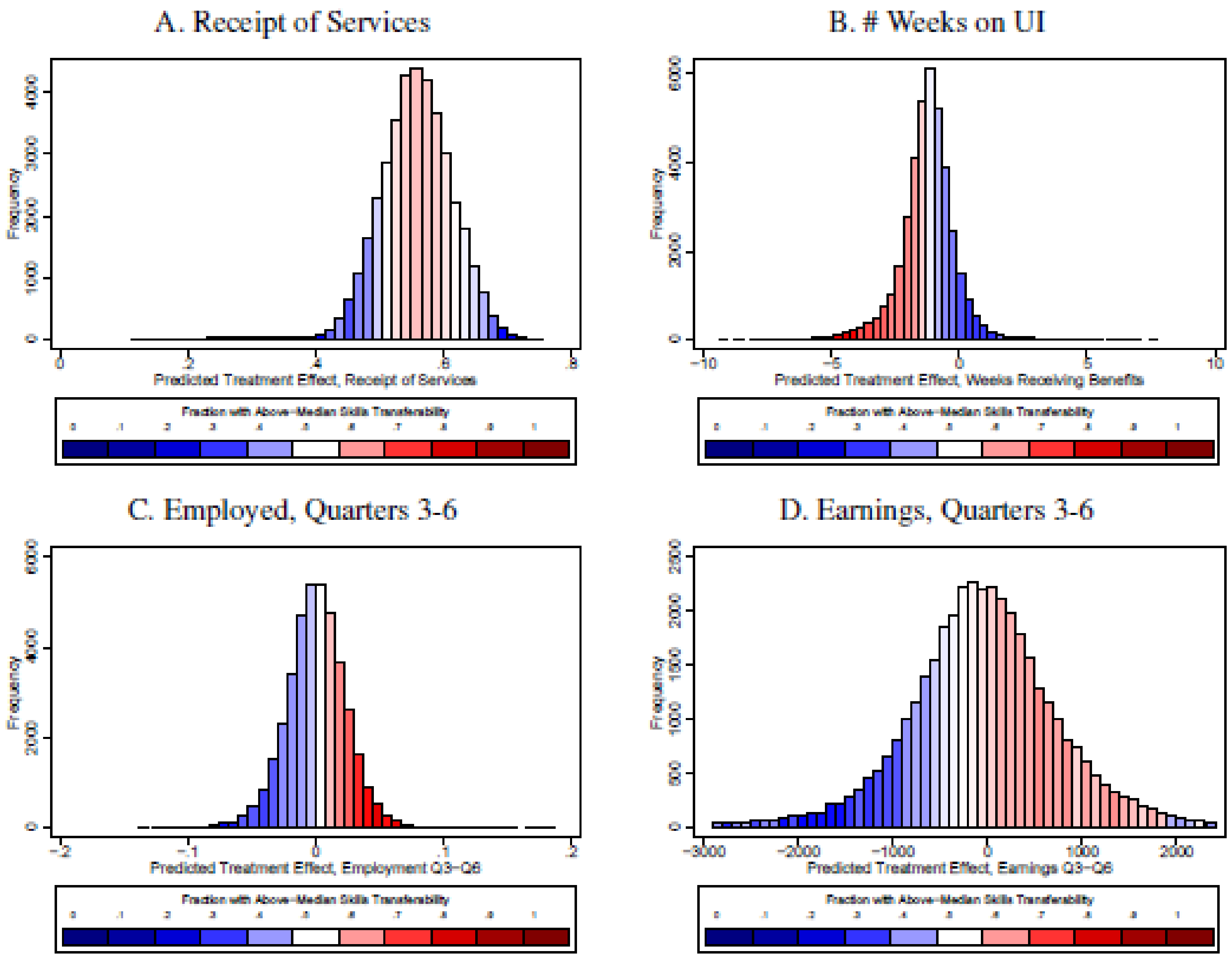


Machine Learning Examples

Using “causal forests” (Wager and Athey, 2018; Athey, Tibshirani and Wager, 2018) to estimate heterogeneous treatment effects from training program

Athey, Campbell, Chyn, Hastings and White (in progress) using data from RIPL

Figure 5: Distribution of Predicted Treatment Effects with Occupational Skills Transferability



Machine Learning Examples

ESTIMATING HETEROGENEOUS TREATMENT EFFECTS OF THE EARLY RETIREMENT REFORM

Susan Athey, Rina Friedberg, Nicolaj Mühlbach, Henrike Steimer & Stefan Wager

- In Denmark, the Retirement Reform increased the early retirement age (ERA) gradually by 1/2 years annually from 2014 for cohorts born after 1954

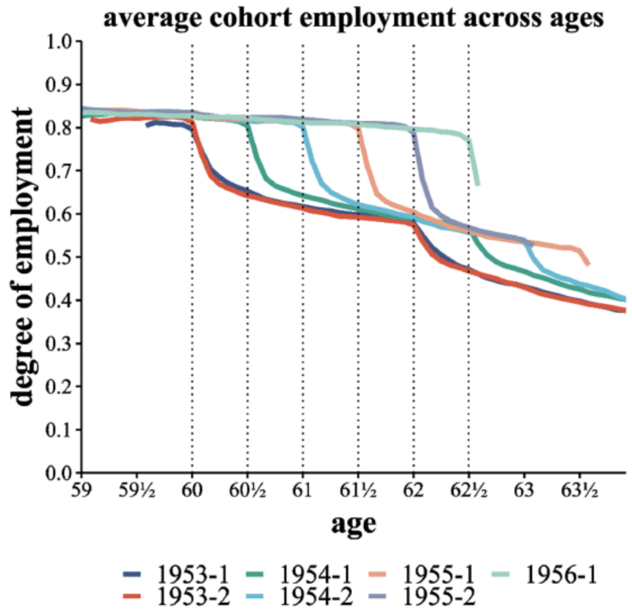


Figure: Average cohort employment for different ages

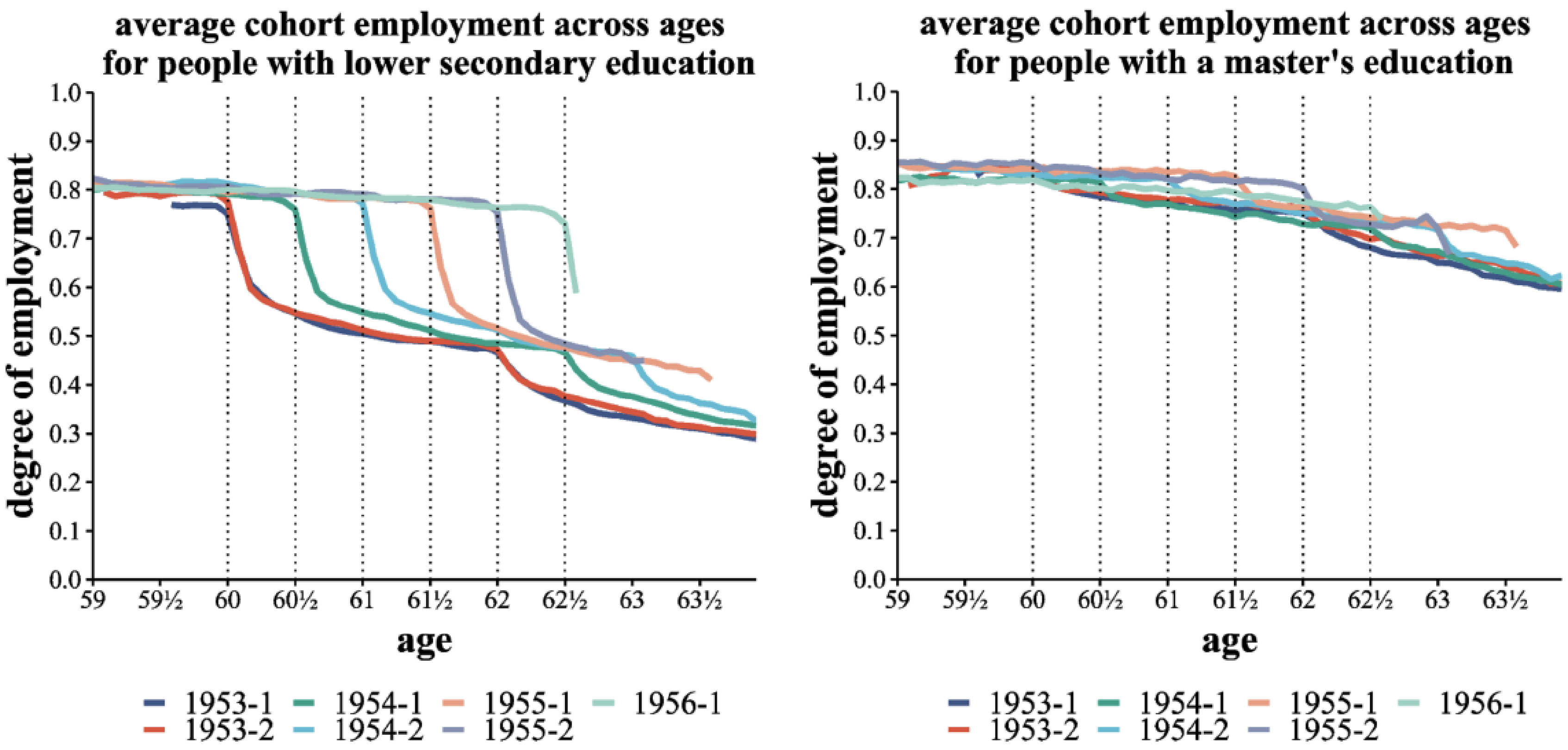


Figure: Average cohort employment for different ages by education level

Machine Learning methods get a better fit to the sign of treatment effect heterogeneity

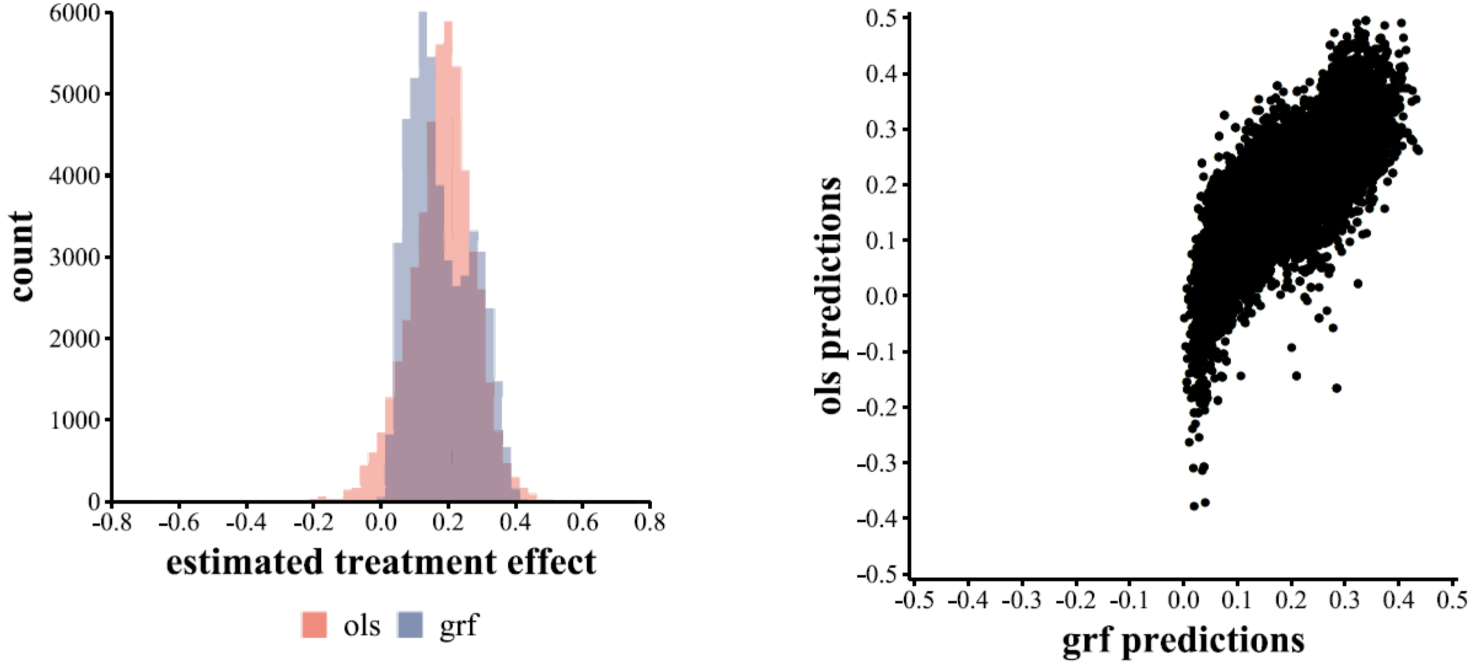


Figure: Distribution of predicted treatment effects

Causal forest discovers significant treatment effect heterogeneity as evaluated “out of bag”

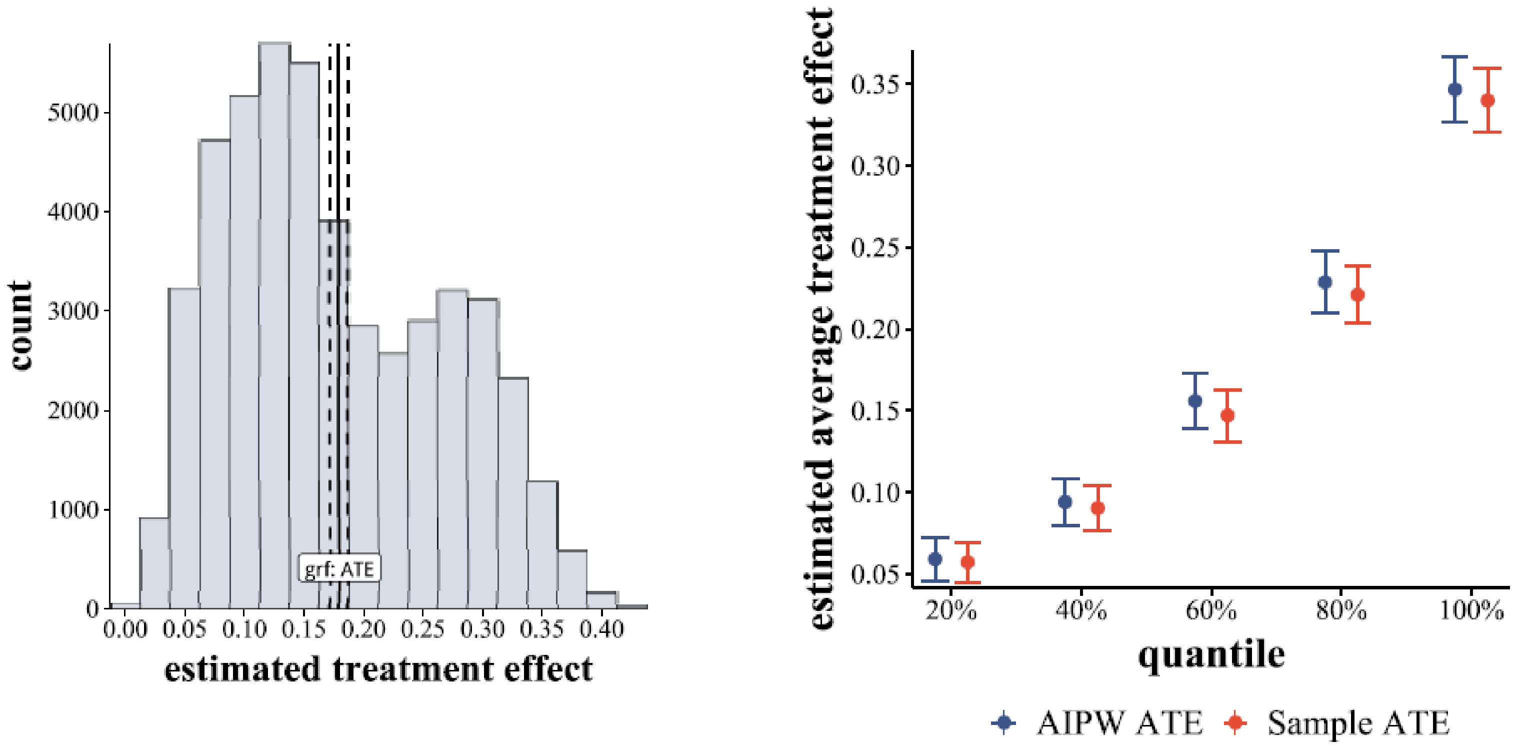
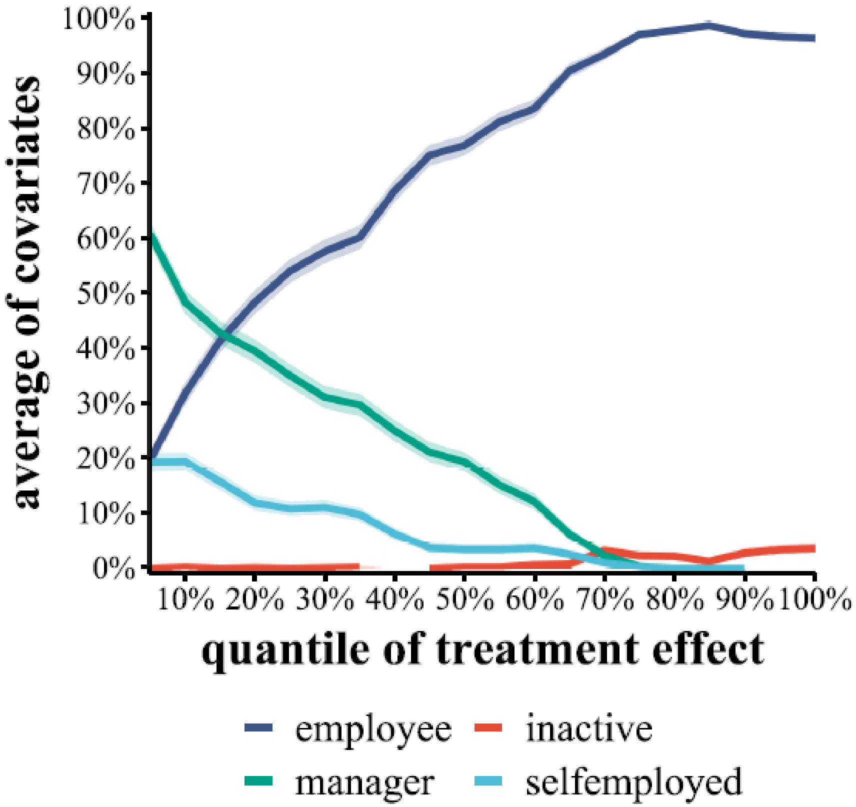
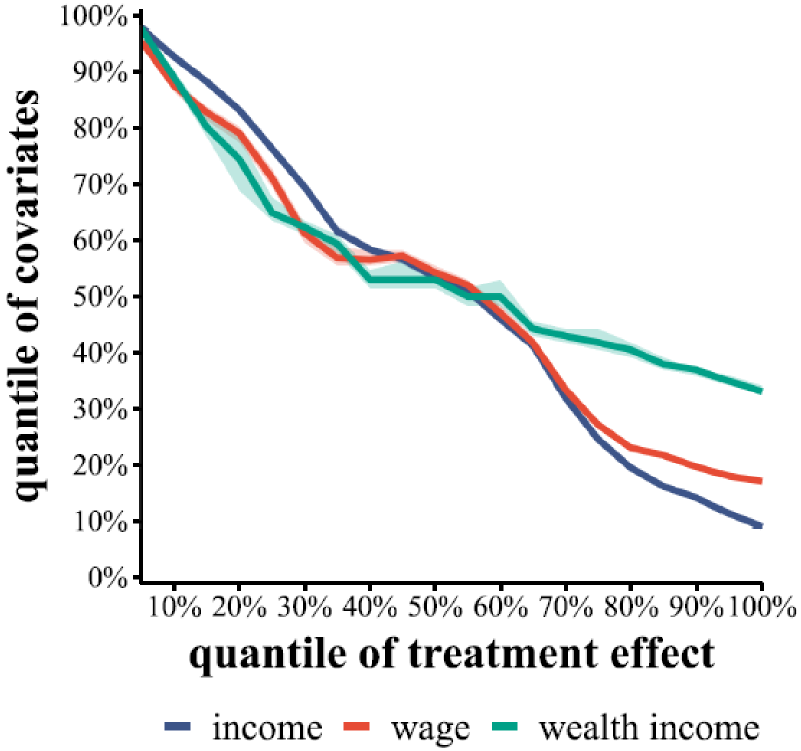


Figure: Distribution of estimated out-of-bag treatment effects

Average values of covariates for different quantiles of estimated treatment effects





ML and Structural Models: Shopping Application

Combine structural model with matrix factorization techniques and computational methods from ML

Scanner data from supermarket
- Product hierarchy (category, class, subclass, UPC)
- Prices change Tuesday evening
- Study 123 high-frequency categories with 1263 UPCs
  - Multiple UPCs per category
  - Typically purchase only one UPC per trip in categroy
  - Independent price changes
  - Not too much seasonality
  - 333,000 shopping trips for ~2000 consumers over 20 months

Economic Goals:
- Optimal pricing
- Benefits of personalization versus simpler segmentation

Methodological Goals:
- Contrast off-the-shelf ML, off-the-shelf econometrics with combined models
- Tune and test models for counterfactual performance

Joint work with Rob Donnelly, David Blei, Fran Ruiz

Structural Model

Mixed logit

- User u, product i, time t

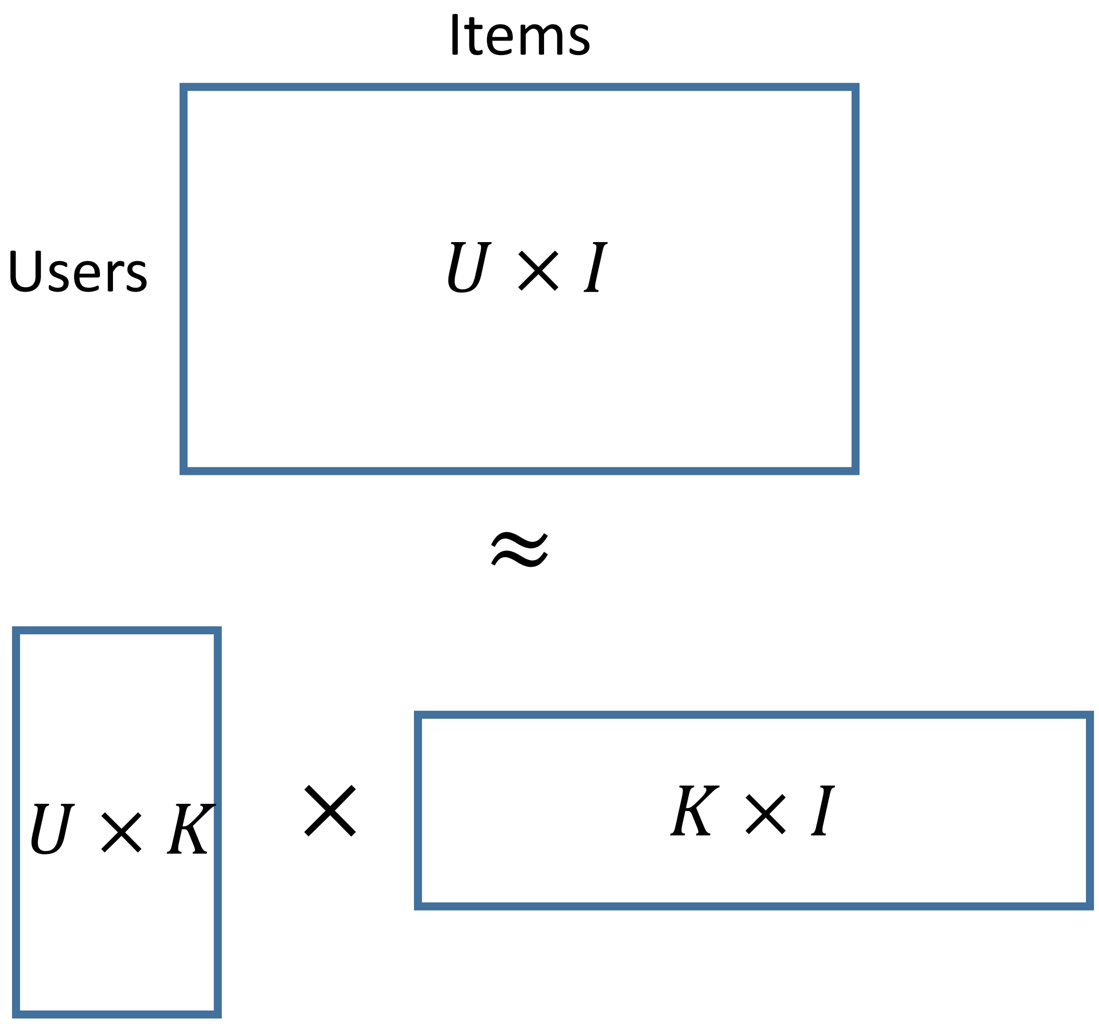
$$\mu_{uit} = \nu_{ui} + \beta X_i - \alpha_u p_{it}$$
$$U_{uit} = \mu_{uit} + \epsilon_{uit}$$

- If ϵ_{uit} i.i.d. Type I EV, then

$$\Pr(Y_{uit} = i) = \frac{\exp(\mu_{uit})}{\sum_j \exp(\mu_{ujt})}$$

- Counterfactuals
  - Out of stock
  - Price changes

Matrix Factorization



Structural Model

Mixed logit

- User *u*, product *i*, time *t*

$$\mu_{uit} = \nu_{ui} + \kappa X_i - \alpha_u p_{it}$$
$$U_{uit} = \mu_{uit} + \epsilon_{uit}$$

- If ϵ_{uit} i.i.d. Type I EV, then

$$\Pr(Y_{uit} = i) = \frac{\exp(\mu_{uit})}{\sum_j \exp(\mu_{ujt})}$$

- Counterfactuals
  - Out of stock
  - Price changes

+ Factorization

Mixed logit + factors

- User *u*, product *i*, time *t*

$$\mu_{uit} = \beta_u \theta_i + \kappa_u X_i - \rho_u \alpha_i p_{it}$$

- Add in nesting for outside good
  - Implement as two-stage estimation with inclusive value (McFadden)
  - Also factorization of outside good

Model Comparisons

Nested Factorization
- All categories estimated in single model
- Items substitutes within category, independent across
- Tuned on held-out validation set

Hierarchical Poisson Factorization (HPF)
- All items in single model, each item **independent** of others
- A form of matrix factorization allowing for covariates
- Ignores prices
- **Scales easily**

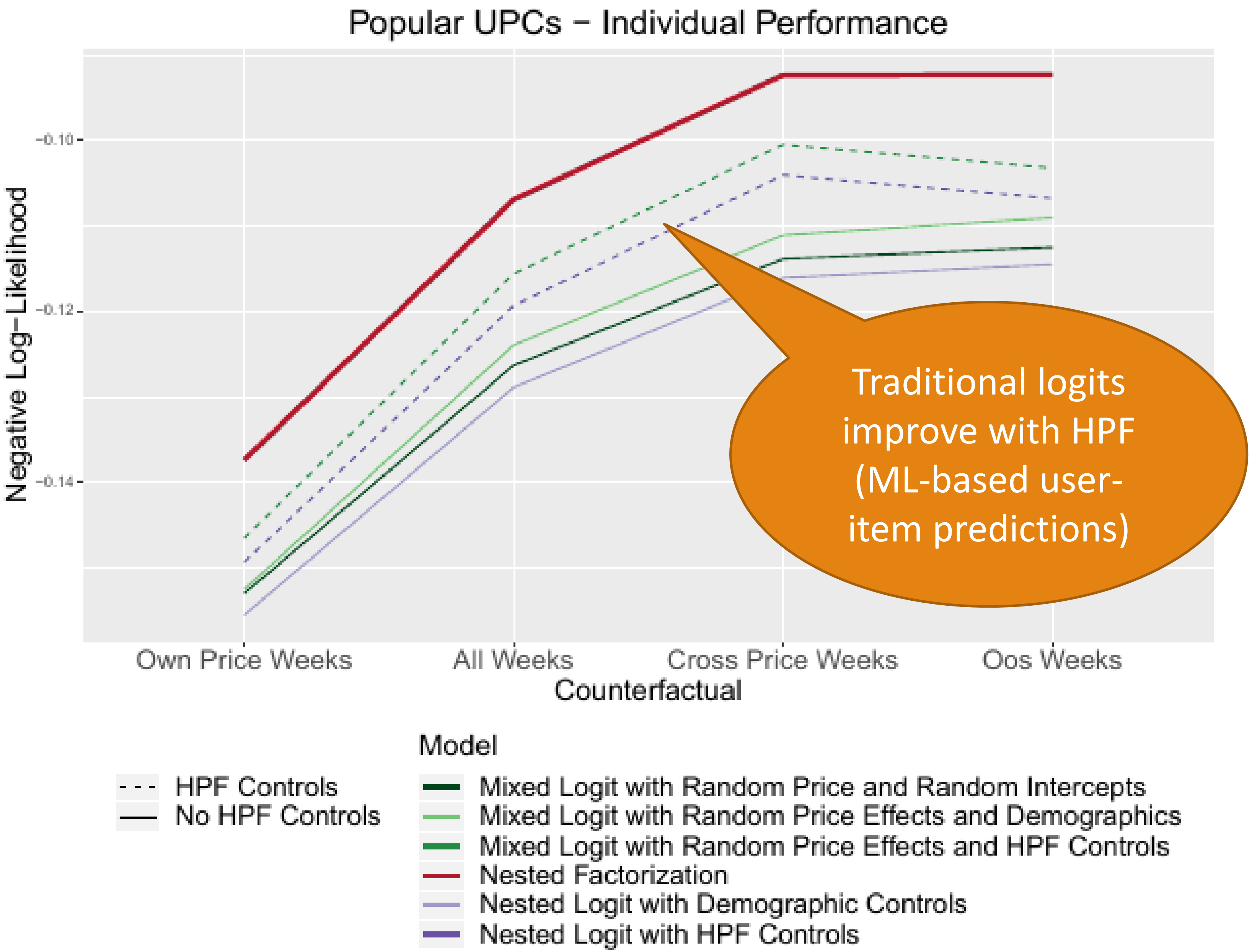
Category by category logits
- Mixed logit (random coefficients)
- Nested Logit
- With various controls (demographic, etc.)

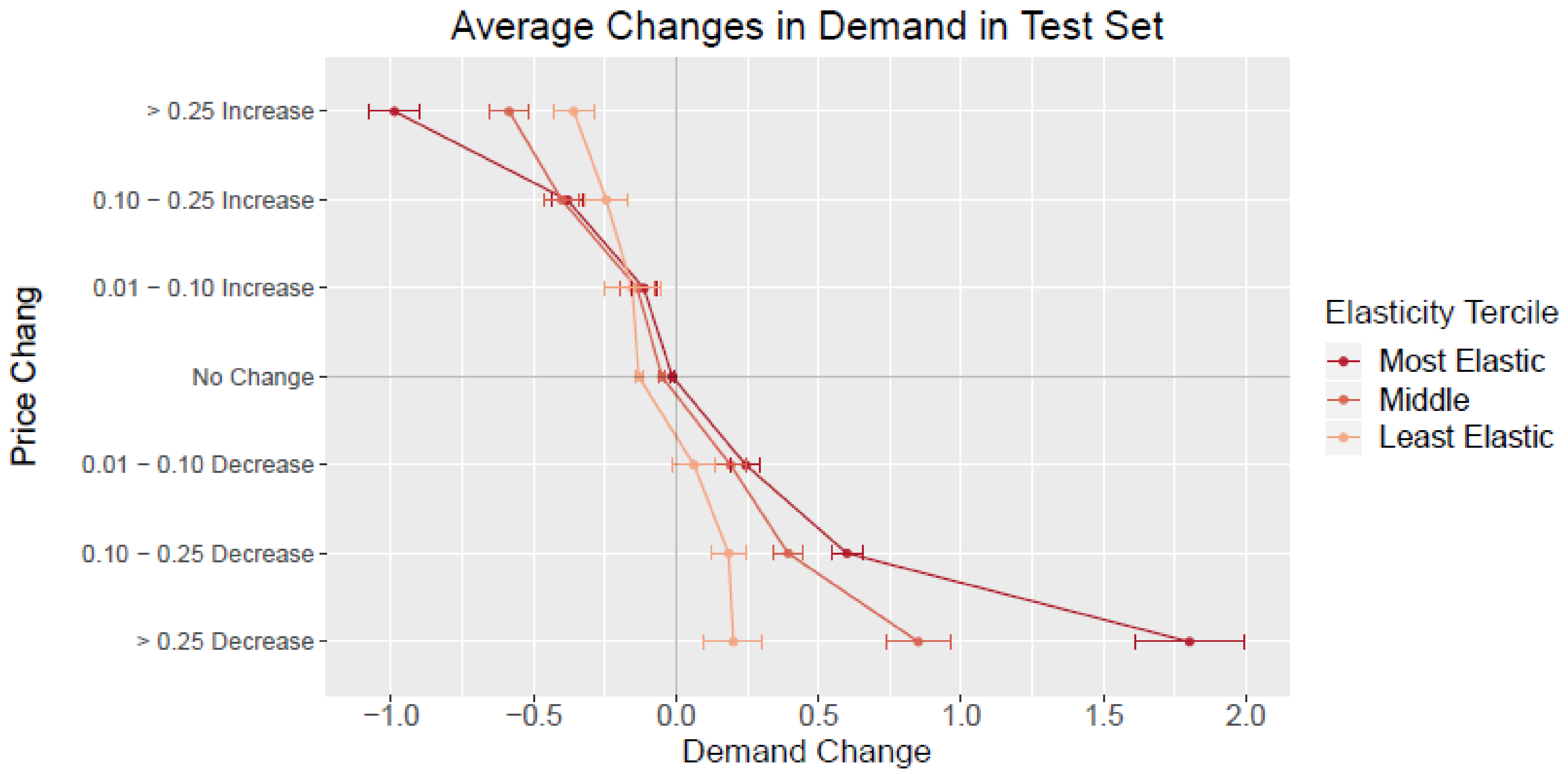
Logits with HPF Factors
- Include user-item prediction from HPF model

Performance by Scenario (Counterfactual)

Evaluate log-likelihood only in weeks where an item falls into specified scenarios:

- Price changed for the item this week
- Price changed for another item in the same category this week
- Another item in the same category is out of stock at least one day this week





Validation of Structural Parameter Estimates

Compare Tues-Wed change in price to Tues-Wed change in demand, in test set

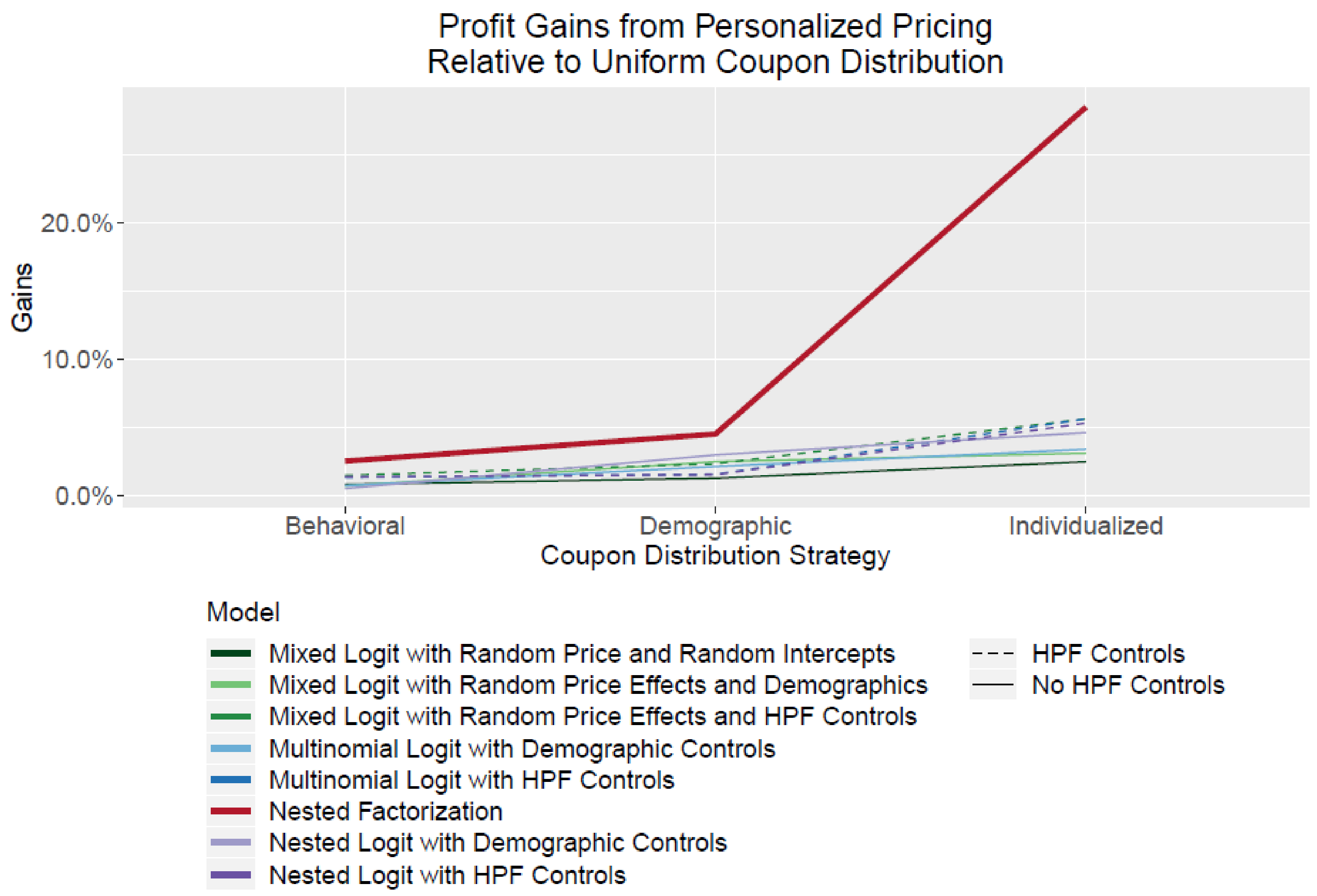
Break out results by how price-sensitive (elastic) we have estimated consumers to be

ML Approach Improves Ability to Profit from Customer Targeting

How much profit can be made by giving a 30% off coupon for a single product to a targeted selection of 30% of the shoppers in the store?

Compare:

Random allocation, demographic targeting, or individual targeting



What recent advances in AI can directly help solve economic, business and social problems?

Stanford Initiatives:

Shared Prosperity and Innovation

Human-Centered Artificial Intelligence

Active learning can be very useful in environments where analyst can intervene

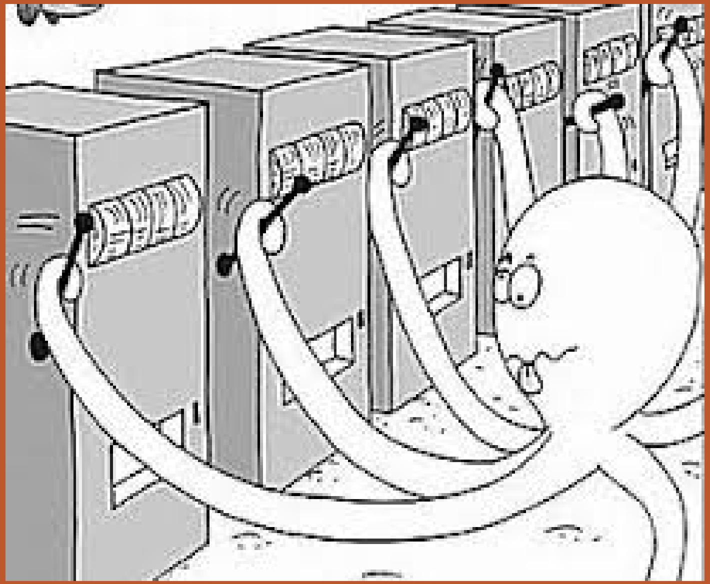
- Incremental improvement is key to tech firm success
- Digital interaction with ability to use dynamic experimentation
- RCT's 3.0: iteratively optimize across many alternatives, with targeting and customization

Examples/Applications

- Nudges for financial health (Ideas42)
- Targeted application of training programs (e.g. RIPL)
- Digital tutors/training (e.g. 17Zuoye)
- Decision-making applications
  - Information for first-generation college students (Ideas42)
  - Contraception selection in developing countries (World Bank)
- Charitable giving
  - Contextual bandits to learn best prompt and charity (IPA/Gates/PayPal)
- Advice/nudge app for newly released prisoners (Ideas42)
- Worker relocation, job search (Facebook)

Active Learning

System interacts with its environment, taking actions or assigning treatments



Bandits:

- Balance **exploration** (learning) and **exploitation** (getting the best outcome for each subject)
- Heuristics such as Thompson Sampling
  - Assign treatment in proportion to probability it is optimal

Contextual bandits:

- Learn a (time-varying) targeted treatment assignment policy mapping from individual characteristics to treatments

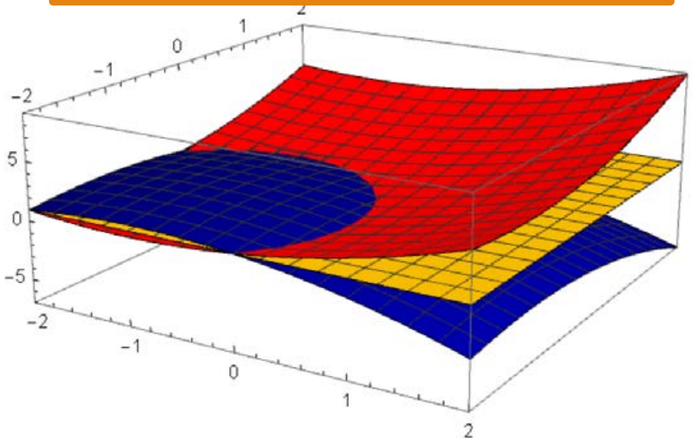
$$\pi_t : \mathbb{X} \rightarrow \mathbb{W}$$

- Consider subjects in batches
- After each batch, estimate model $\hat{\mu}_t(x, w)$
- Apply bandit heuristics
- Modifications in my work: consider **scientific discovery** as goal, methods for valid **hypothesis testing**, incorporate **econometric insights** in algos

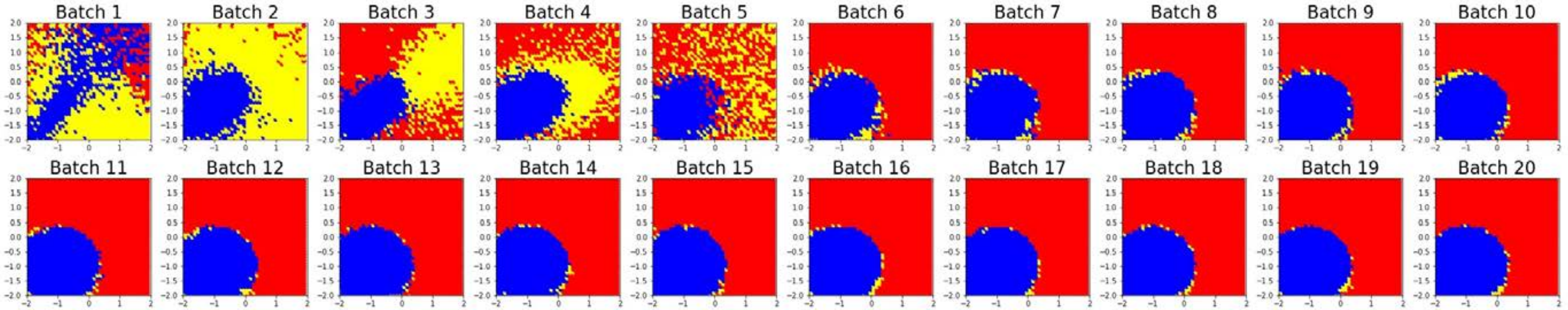
Reinforcement learning:

- Treatment/action affects state
- Context includes state
- E.g. dynamic educational apps

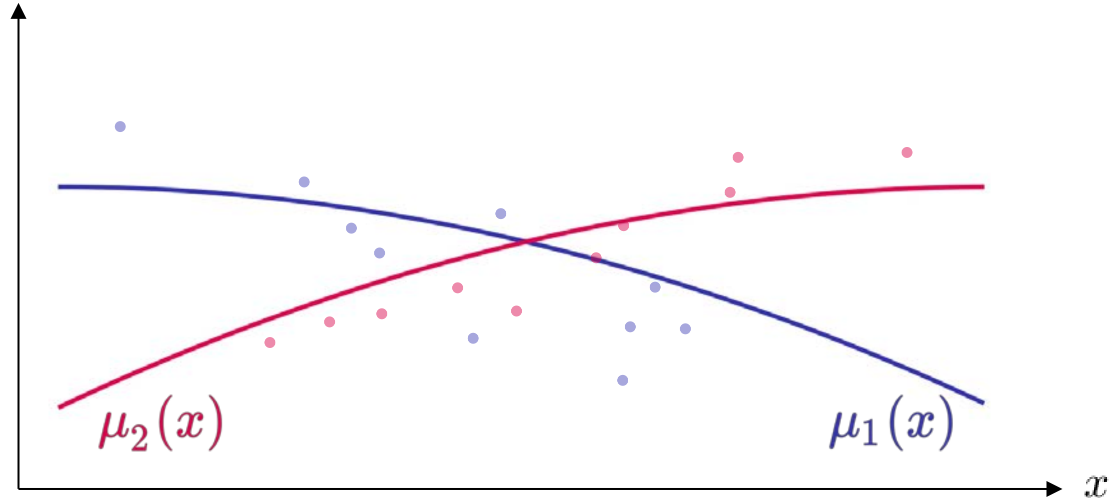
Outcomes for different arms depend on contexts

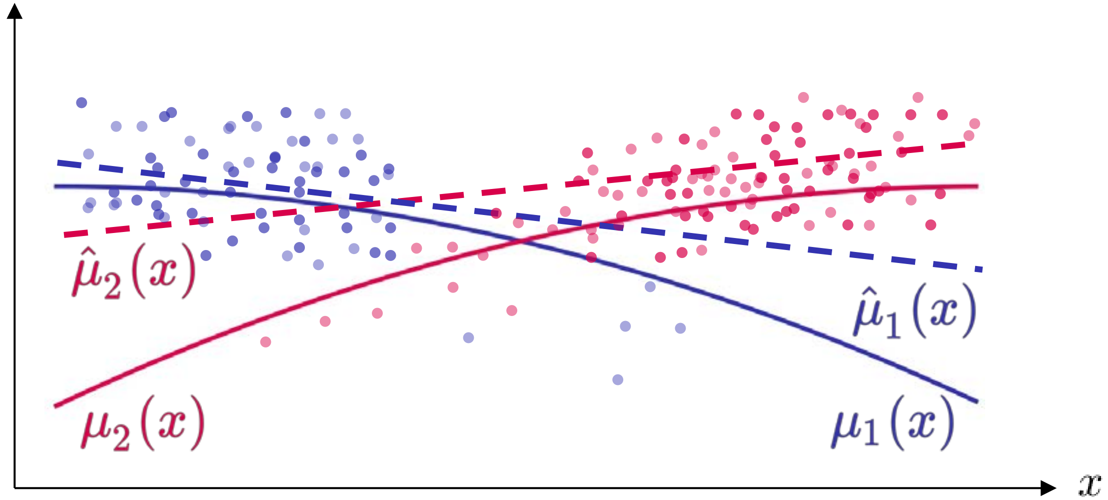


Doubly robust contextual bandit learns the optimal treatment assignment policy



Estimation along the path plagued by adaptivity of assignment process; weighting creates variance as assignment probabilities converge

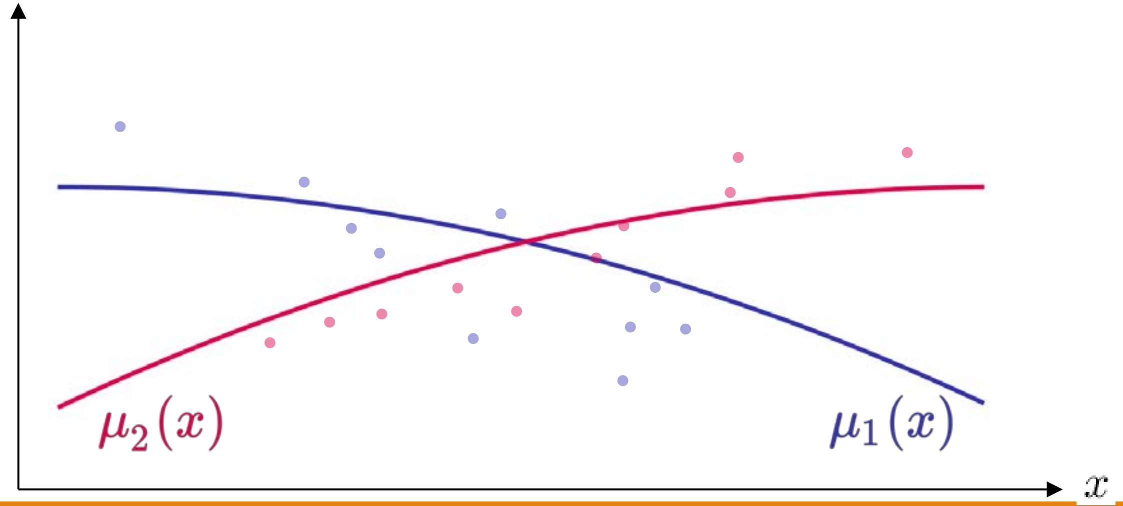


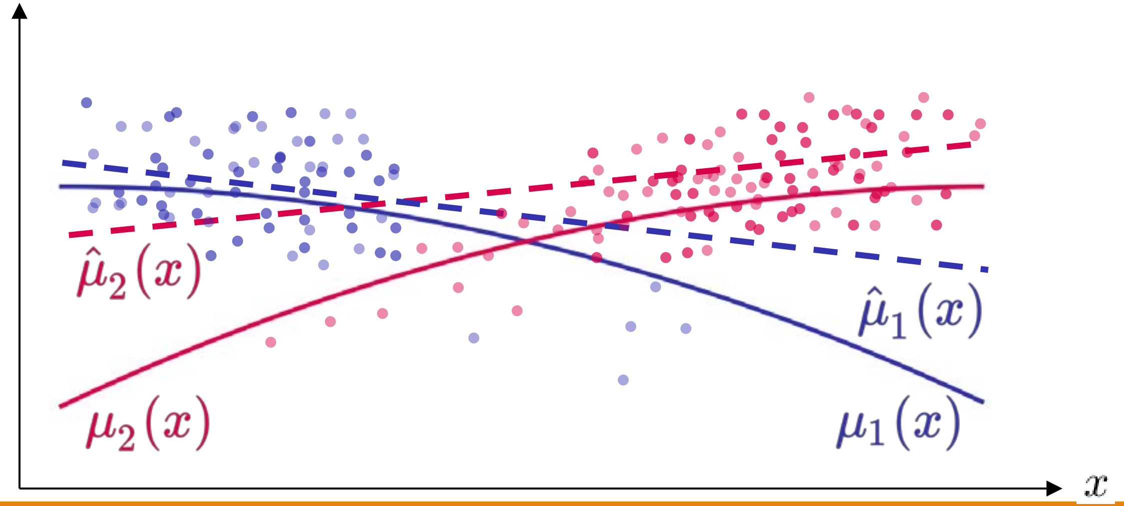


Appendix

Estimation is challenging: Contextual Bandit example

- Inherent bias in estimation due to adaptive assignment of contexts to arms.
  - context more likely assigned to high-performing arm
  - creates systematically unbalanced data
- Algorithmic selection (on observables) similar to selection biases from agent optimization
- See Diamakopoulou, Zhou, Athey and Imbens (2019, AAAI) who provide regret bounds for doubly robust approaches





Economists as Engineers: A New Chapter

Services, education, training, advice delivered digitally by firms, governments, and philanthropy

AI and econometric theory needs work but not the main constraint

Instead, success will depend on:

- Understanding broader context
  - Social science to identify opportunities to intervene
- Defining measures of success that are measurable in the short term and related to long term outcomes
  - Non-manipulable
  - Don't let the AI "teach to the test"
- Reaching target audience
  - Finding partners with access to individual time and attention
  - Distributing digital services
  - Making engaging and effective content (treatments)
- Social scientists key contributors to multi-disc. teams
  - Evaluation is embedded in system and not separable from system design

Conclusions

Causal inference is key to using machine learning and artificial intelligence to make decisions

- This is a tautological statement: but not fully appreciated

Black box algorithms come with risks and challenges

AI/ML in causal framework has desirable properties (stability, fairness, robustness, transfer, ….)

Enormous literature on theory and applications of causal inference in variety of design settings

- Conceptual framework for both static and dynamic settings
- Structural models enable counterfactuals for never-seen worlds

ML can greatly improve practical performance, scalability

- With careful modifications, attention to objective functions, cross-fitting/sample splitting

Challenges: data sufficiency, finding sufficient/useful variation in historical data

- Recent advances in computational methods in ML don't help with this
- But tech firms conducting lots of experiments, running bandits, and interacting with humans at large scale can greatly expand ability to learn about causal effects and solve societal problems

References

Selected Overview Articles: Econometrics and ML

Survey

- S. Athey, “The Impact of Machine Learning on Economics.”

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