Deep Learning for Econometrics

Greg Lewis (MSR + NBER) Matt Taddy (MSR + Chicago)
Jason Hartford (UBC) Kevin Leyton-Brown (UBC)
Kui Tang (Columbia) Dave Blei (Columbia)
Matt Goldman (MSFT) Justin Rao (MSR) Di Wang (MSFT)
James Zou (Stanford) Mengting Wan (UCSD) Richard Li (UW)

What do economists do?

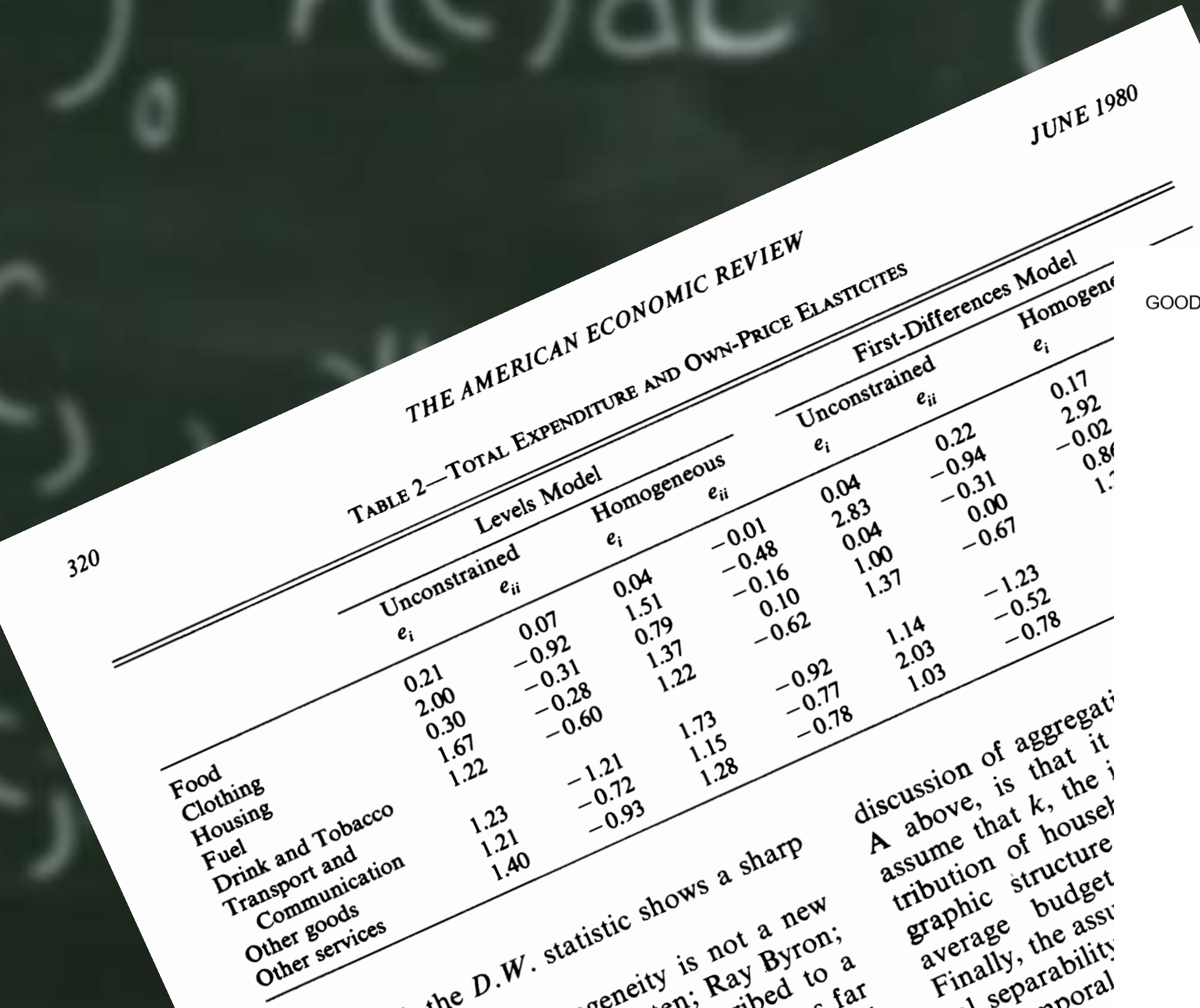
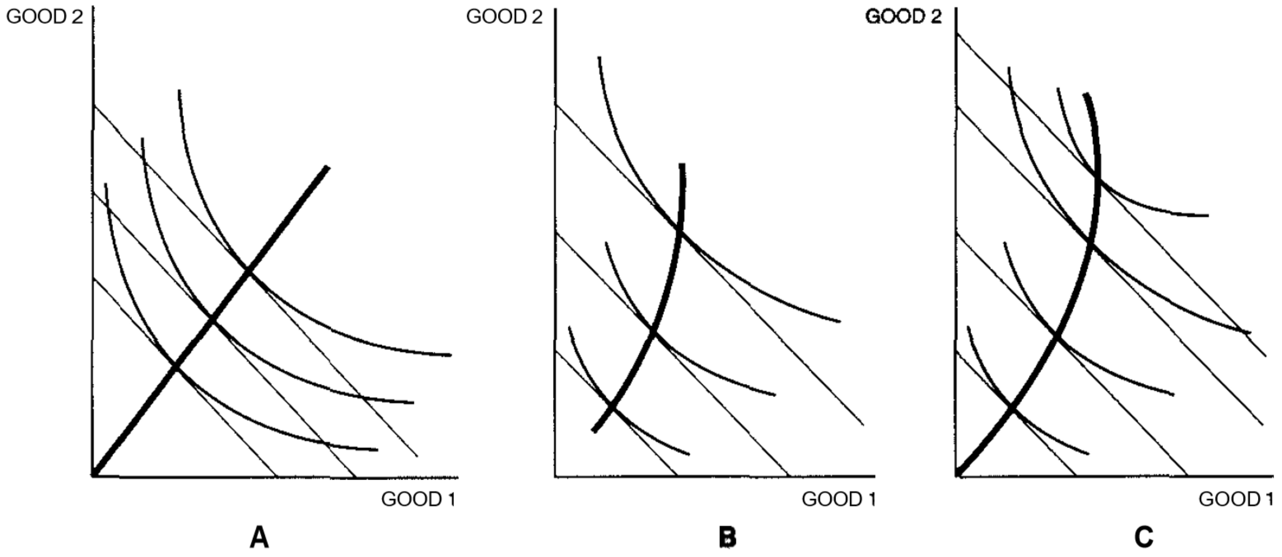


TABLE 2—TOTAL EXPENDITURE AND OWN-PRICE ELASTICITES

	Levels Model				First-Differences Model		
	Unconstrained		Homogeneous		Unconstrained		Homogene[ous]
	e_i	e_{ii}	e_i	e_{ii}	e_i	e_{ii}	e_i
Food	0.21	0.07	0.04	−0.01	0.04	0.22	0.17
Clothing	2.00	−0.92	1.51	−0.48	2.83	−0.94	2.92
Housing	0.30	−0.31	0.79	−0.16	0.04	−0.31	−0.02
Fuel	1.67	−0.28	1.37	0.10	1.00	0.00	0.8[illegible]
Drink and Tobacco	1.22	−0.60	1.22	−0.62	1.37	−0.67	1.[illegible]
Transport and Communication	1.23	−1.21	1.73	−0.92	1.14	−1.23	
Other goods	1.21	−0.72	1.15	−0.77	2.03	−0.52	
Other services	1.40	−0.93	1.28	−0.78	1.03	−0.78	

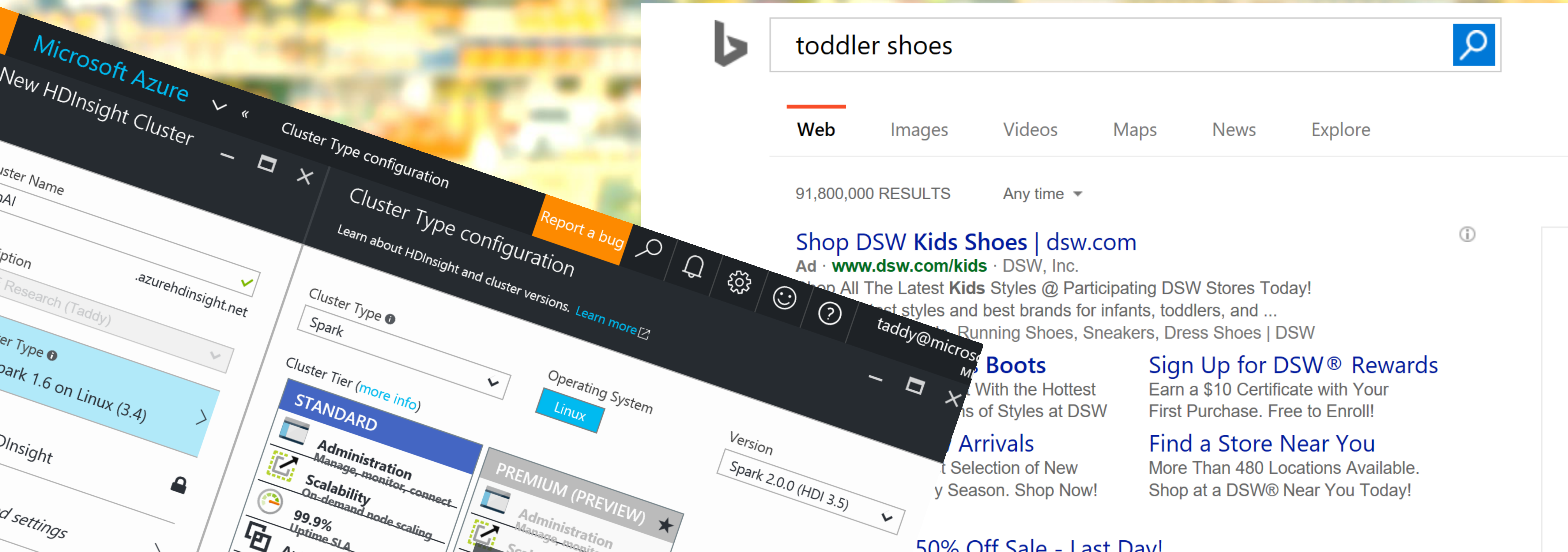
... the *D.W.* statistic shows a sharp
...geneity is not a new
...ten; Ray Byron;
...ribed to a
...f far

discussion of aggregati[on]
A above, is that it
assume that *k*, the i[...]
tribution of househ[...]
graphic structure
average budget
Finally, the assu[...]
...l separability
...mporal



Income expansion paths. Panel **A** depicts unit elastic demands, in panel B good 2 is a luxury good, and in panel C, good 1 is an inferior good.

What do they need to do today?



What can we (AI) do to help?

The dimension of the economist’s problem space has exploded

We can develop ML to navigate this space: stay safe and automate

We can also build new econometrics via deep structure

Example: Demand System

Suppose that you have transactions ‘t’ on products ‘j’.

Write the quantity bought ‘q’ as

$$q_{tj} \;=\; \alpha_j(\boldsymbol{d}_t) \;+\; \gamma_j \log p_{tj} + e_{tj}$$

a function of utility we can ($\alpha_j(\boldsymbol{d}_t)$) and can’t ($e_{tj}$) see, plus price p_{tj}.

You need to have a model like this to target customers or set prices.

But it’s a system!

For example: There many different products

Demand for j depends on substitutes and complements

Or: where does price come from?

$$\log p_{tj} \;=\; \varphi_t(\boldsymbol{c}_j) \;+\; \psi_j q^{\star}_{tj} \;+\; \nu_{tj}$$

and the *demand system* is in equilibrium when $q^{\star}_{tj} = q_{tj}$

This equilibrium introduces `price endogeneity’: $\mathbb{E}[p_{tj} e_{tj}] \neq 0$

Sometimes it's just regression

If we treat ε_{tj} as independent this is a prediction problem
e.g., model store transactions with covariates $\boldsymbol{x}_{tj}$ as

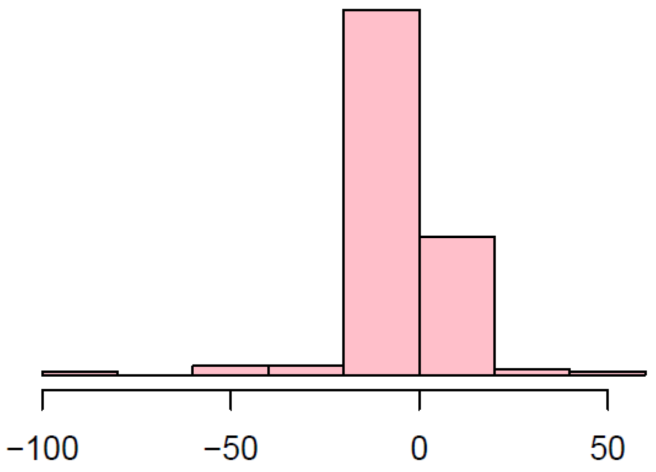
$$\mathbb{E}\log q_{tj} = \boldsymbol{x}_{tj}'\boldsymbol{\beta} + \log p_{tj}\, \boldsymbol{x}_{tj}'\boldsymbol{\gamma}$$

Elasticities:

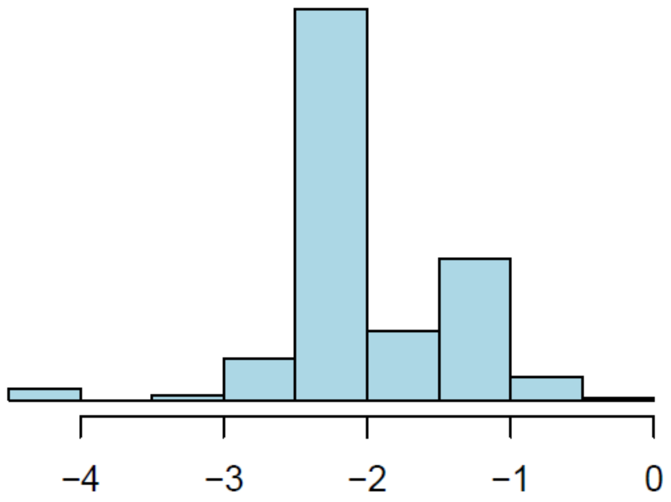
one shared: $x_{tj} = 1$

$$\frac{dq}{dp}\frac{p}{q} = -0.23$$

brand-specific: $x_{tjk} = \mathbb{1}_{[k=j]}$



$\boldsymbol{x}_{tj}$= *featurized description*



Moving inside a demand system (AIDS)

It's *almost* ideal:

$$\boldsymbol{s}_t = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\log(\boldsymbol{p}_t) + \boldsymbol{\beta}\log\frac{e_t}{\phi_t} + \boldsymbol{\varepsilon}_t$$

s_{tj} is the budget share for product j in basket t and e_t is the budget

($e_t = \sum_j \$_{tj}$ and $s_{tj} = \$_{tj}/e_t$)

ϕ_t is the translog price index $\sum_j \log p_{tj}\left[\alpha_j + \sum_k \gamma^{\star}_{jk} \log p_{tk}\right]$

(which we will replace with a plug-in for estimation)

This is meaningful after aggregation, and we can actually estimate it

Factorizing Γ

The price terms are key to finding complements and substitutes

$$\mathbb{E}s_{tj} = \alpha_j + \sum_k \gamma_{jk} \log p_{tj} + \beta_j \frac{e_t}{\phi_t}$$

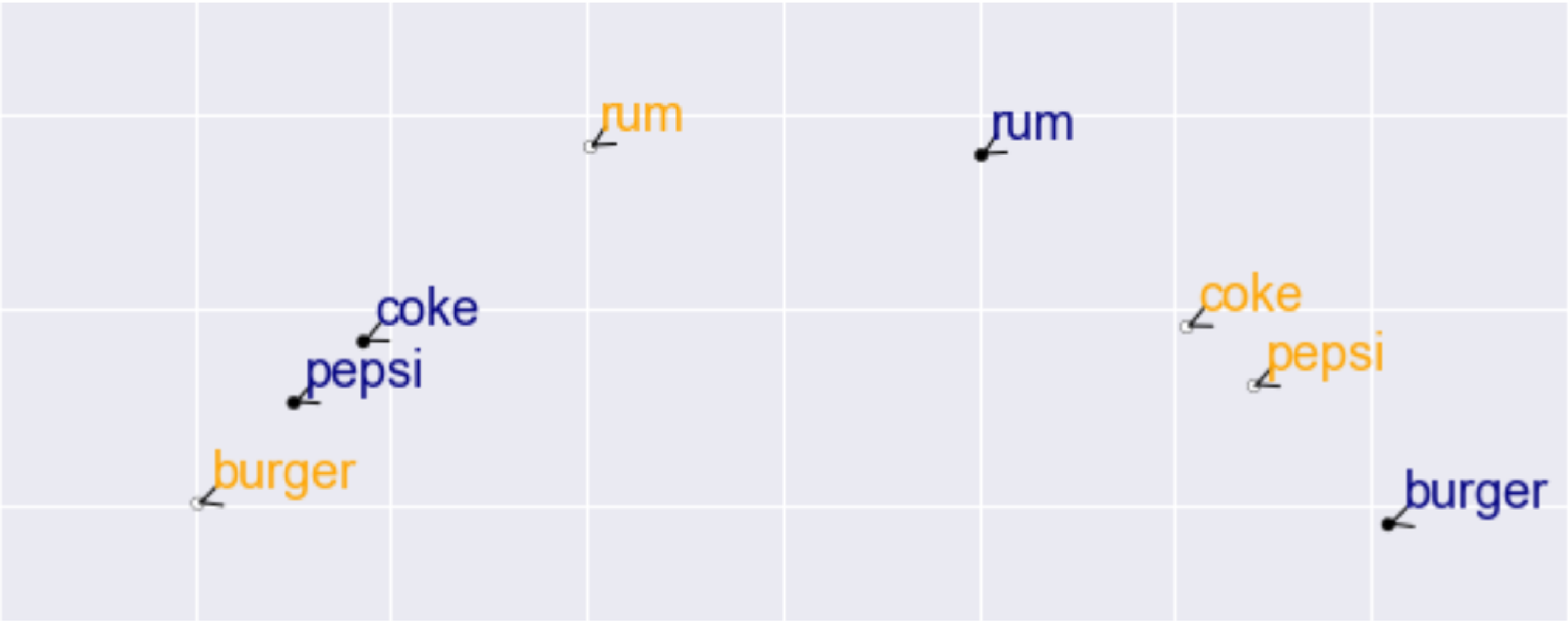
$\boldsymbol{\Gamma}$ is $J \times J$, so we need to reduce dimension if J is going to go big

One option: square matrix factorizations from word/prod embedding

$\boldsymbol{\Gamma} = \boldsymbol{UV}' + \boldsymbol{VU}' + \boldsymbol{D}$ where $\boldsymbol{u}_j, \boldsymbol{v}_j$ are S-vectors and D is J-diagonal

(AIDS implies restrictions: $\gamma_{jk} = \gamma_{kj}$, $\quad \sum_j \gamma_{jk} = \sum_j \gamma_{kj} = \sum_j \beta_j = 0$*)*

Product Embeddings



substitutes (synonyms) are close in the same vector space
complements (topical words) are close across vector spaces

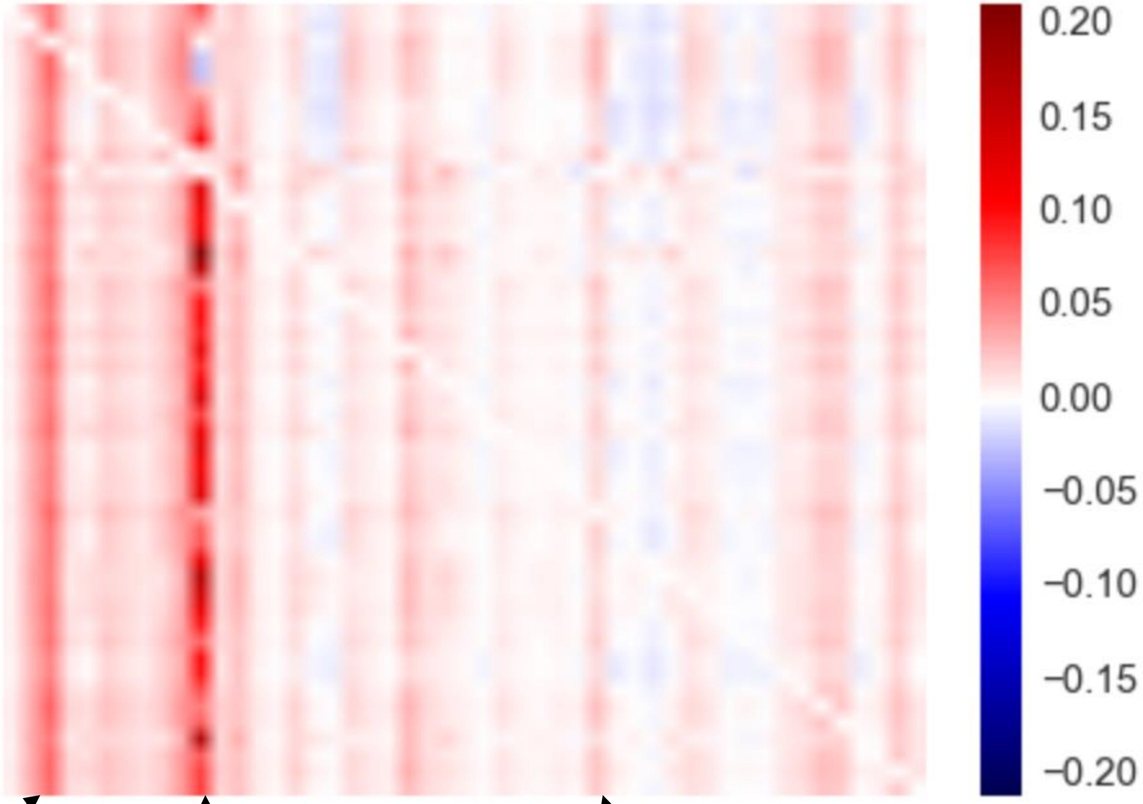
Beer

We fit on store-week totals.

Translate the γ_{jk} values into [compensated] elasticities as

$$\frac{\gamma_{jk}}{\bar{s_j}} - \bar{s_k} - \mathbb{1}_{[k=j]}$$

Elasticity matrix (omitting diagonal)



But wait… it's still a system

$$\boldsymbol{s}_t = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\log(\boldsymbol{p}_t) + \boldsymbol{\beta}\log\frac{e_t}{\phi_t} + \boldsymbol{e}_t$$

Recall: where does price come from?

$$\log p_{tj} = \varphi_t(\boldsymbol{c}_j) + \psi_j q_{tj}^{\star} + \nu_{tj}$$

and the *demand system* is in equilibrium when $q_{tj}^{\star} = q_{tj}$

This equilibrium introduces `price endogeneity': $\mathbb{E}[p_{tj}e_{tj}] \neq 0$

Endogenous Errors

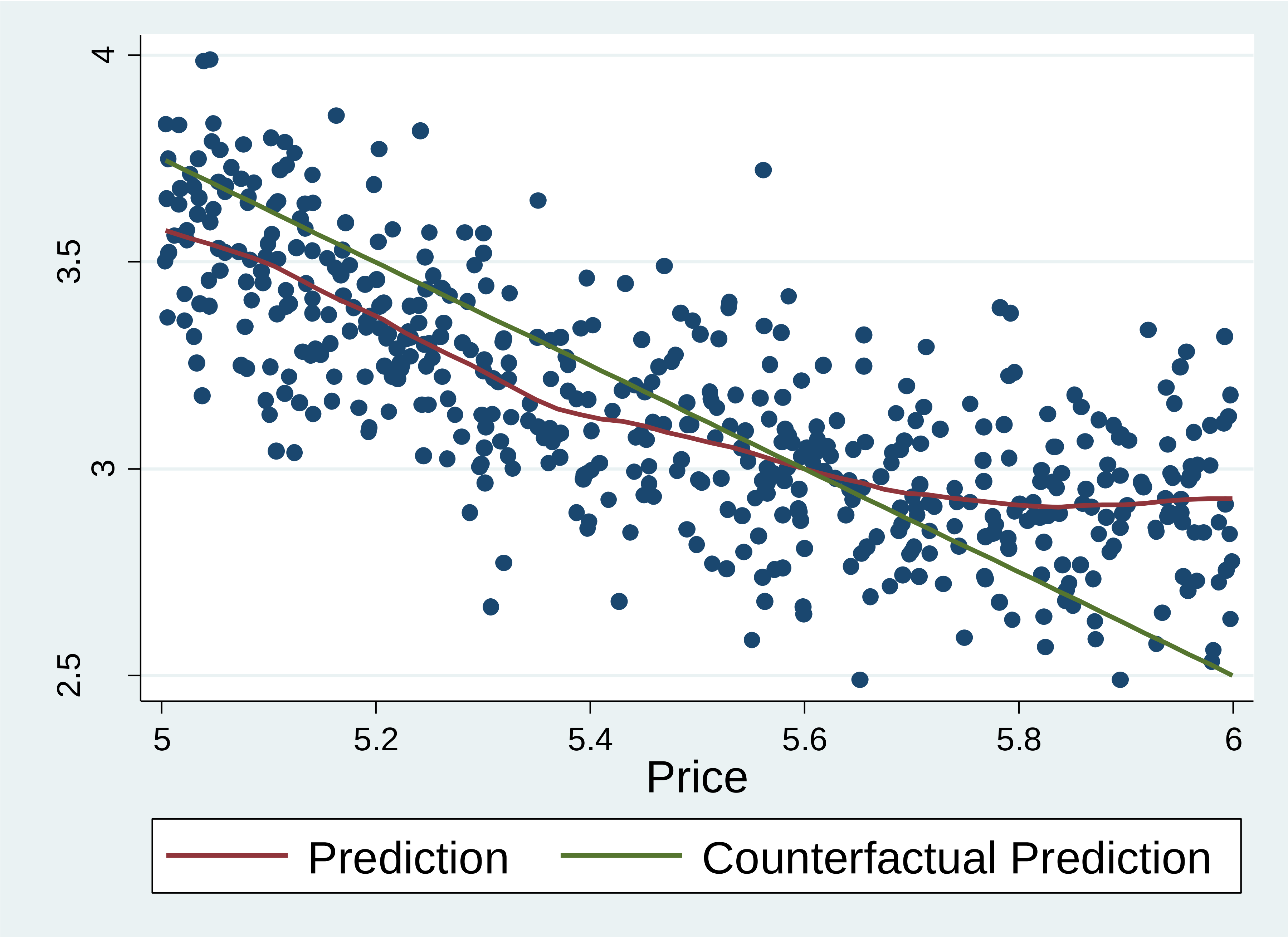
$$y = g(p, \boldsymbol{x}) + e \text{ and } \mathbb{E}[\, p\, e\,] \neq 0$$

If you estimate this using naïve ML, you'll get

$$E[y|p, \boldsymbol{x}] = E_{e|p}[g(p, \boldsymbol{x}) + e] = g(p, \boldsymbol{x}) + E[e|t, \boldsymbol{x}]$$

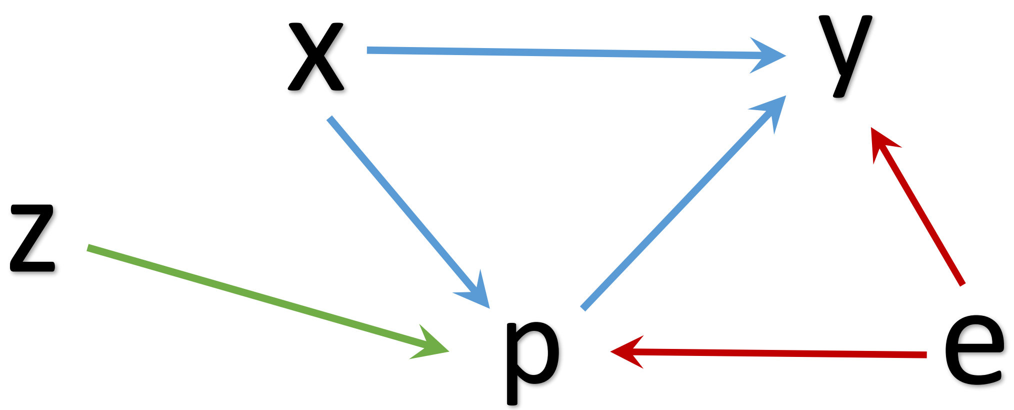
This works for prediction. It doesn't work for counterfactual inference:

What happens if I change p independent of e ?



4
3.5
3
2.5
5
5.2
5.4
5.6
5.8
6
Price
Prediction
Counterfactual Prediction

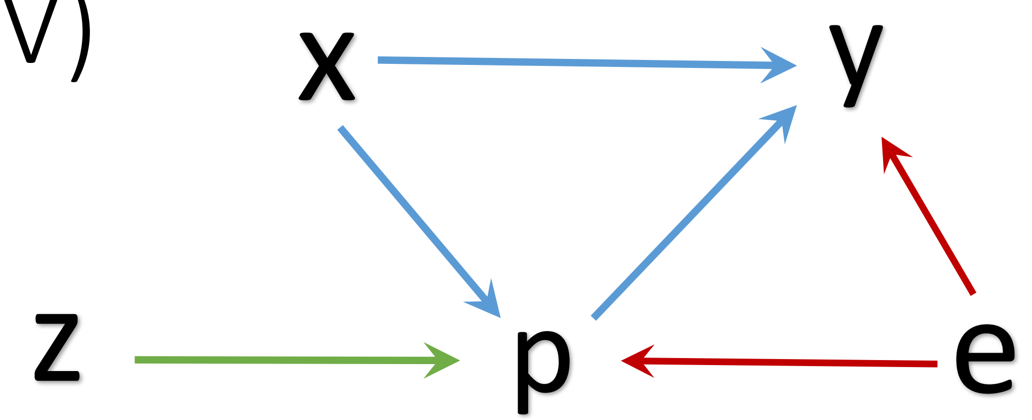
Instrumental Variables (IV)



In IV we have a special $z \perp e$ that influences policy p but not response y.

- Supplier costs that move price independent of demand (e.g., fish, oil)
- Any source of treatment randomization (intent to treat, AB tests, lottery)

Instrumental Variables (IV)



The *exclusion structure* implies

$$E[y|x,z] = E[g(p,x) + e|x,z] = \int g(p,x)d\mathrm{P}(p|x,z)$$

So to solve for *structural* $g(p,x)$ we have a new learning problem

$$\min_{g \in G} \sum \left(y_i - \int g(p,x_i)d\mathrm{P}(p|x_i,z_i) \right)^2$$

$$\min_{g \in G} \sum \left(y_i - \int g(p, x_i) d\mathrm{P}(p | x_i, z_i) \right)^2$$

You might have seen 2SLS:

$p = \beta z + \nu$ and $g(p) = \tau p$ so that $\int g(p) d\mathrm{P}(p|z) = \tau \hat{p} = \tau \hat{\beta} z$

So you first regress p on z then regress y on $\hat{p}$ to recover $\hat{\tau}$.

This requires strict assumptions and homogeneous treatment effects.

$$\min_{g \in G} \sum \left(y_i - \int g(p, x_i) d\mathrm{P}(p | x_i, z_i) \right)^2$$

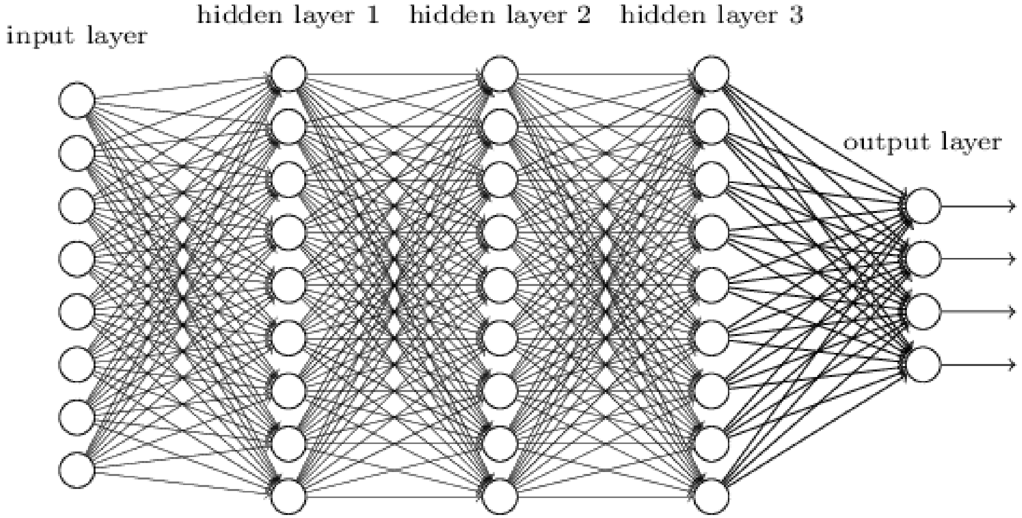
We can target this integral loss function directly with flexible g and P.

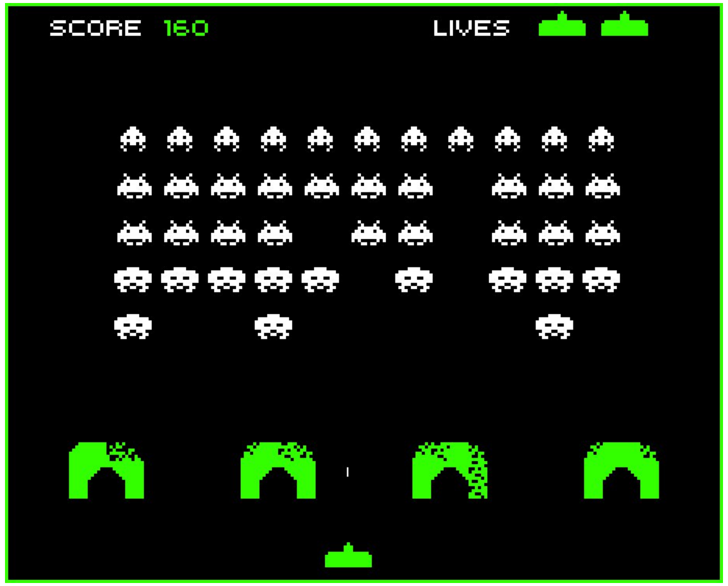
Brute force version

- Fit conditional distributions $\hat{\mathrm{P}}(p | x_i, z_i)$.
- Generate $\{\hat{p}_{ib}\}_{b=1}^{B} \sim \hat{\mathrm{P}}(p | x_i, z_i)$ for each i.
- Train $\hat{g}$ to minimize $[\, \mathrm{y_i} - B^{-1} \sum_b g(\hat{p}_{ib}, x_i) \,]^2$.

Turns IV into two ML tasks: we can use DNNs for both $\hat{\mathrm{P}}$ and $\hat{g}$.

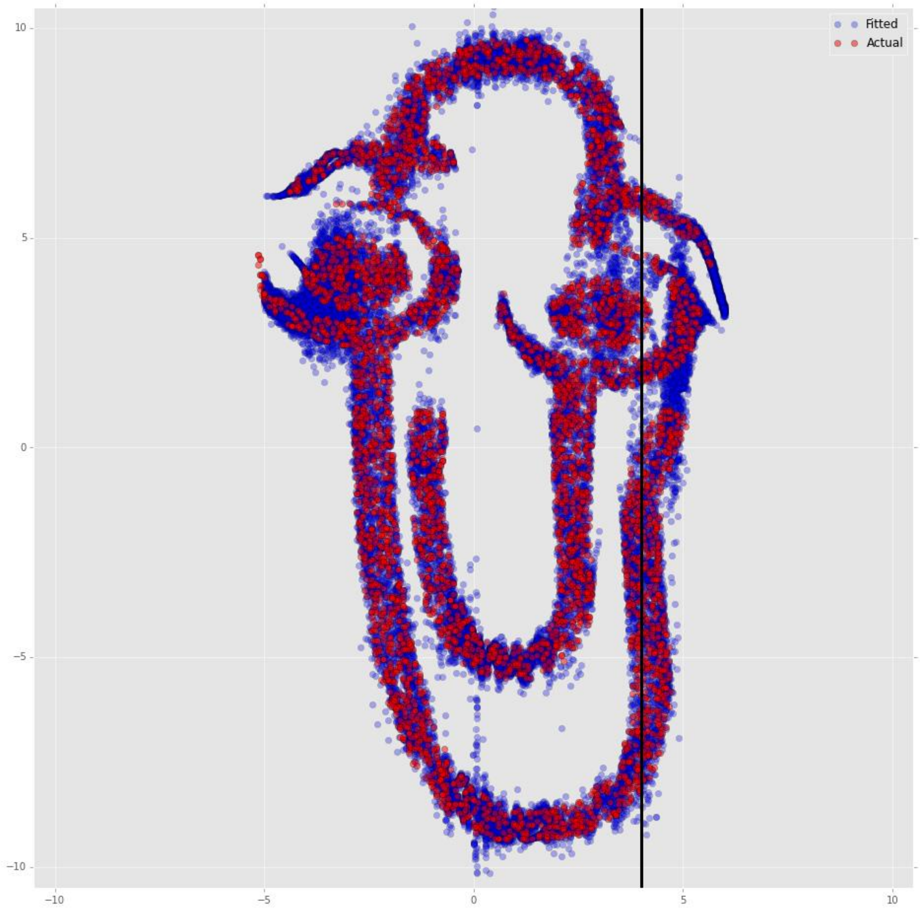
Learning to love Deep Nets

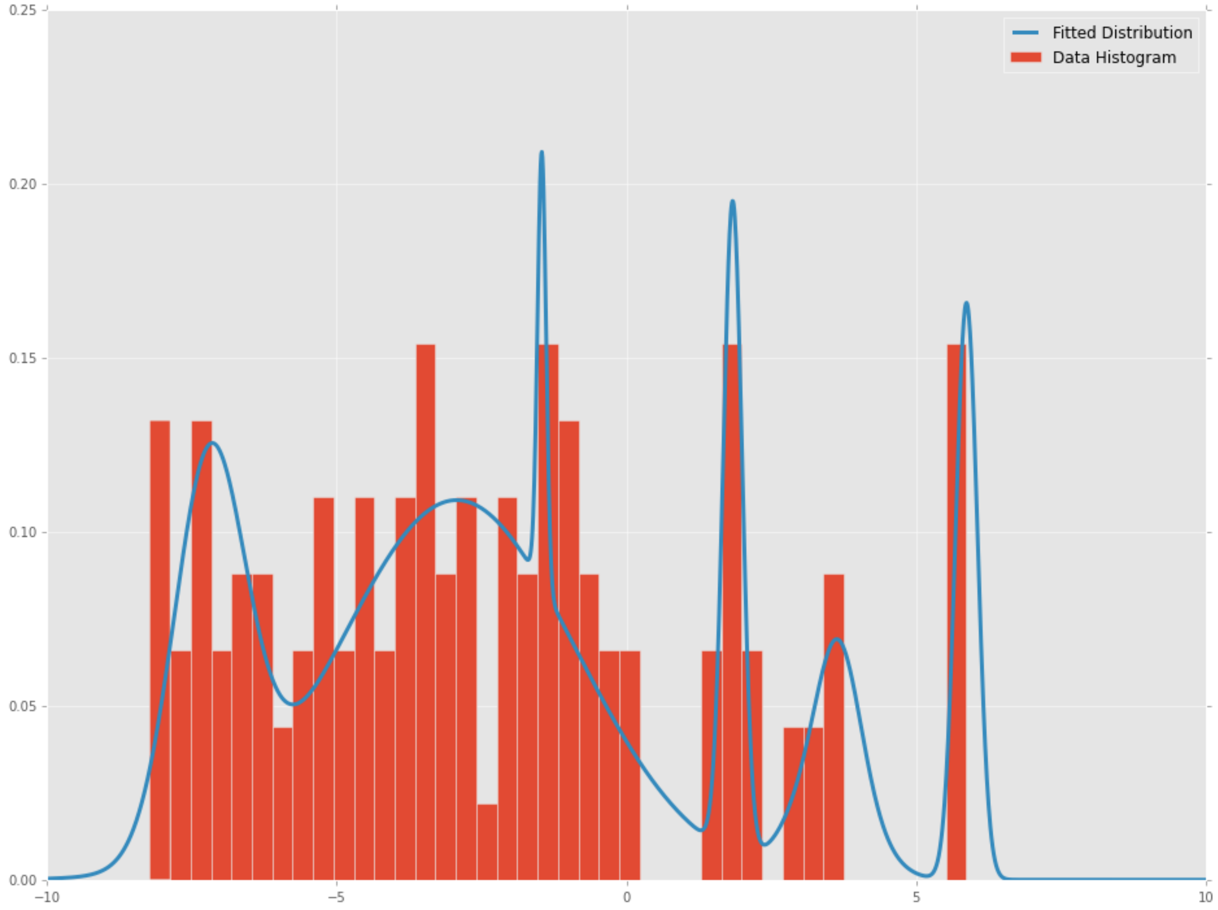




First Stage is out-of-the-box ML: learn $\mathrm{P}(p|x_i, z_i)$

e.g., DNN fits distribution to maximize likelihood for a mixture of Gaussians.





The second stage involves an integral loss function

Brute force just samples from $\hat{\mathrm{P}}(p|x_i, z_i)$ to evaluate

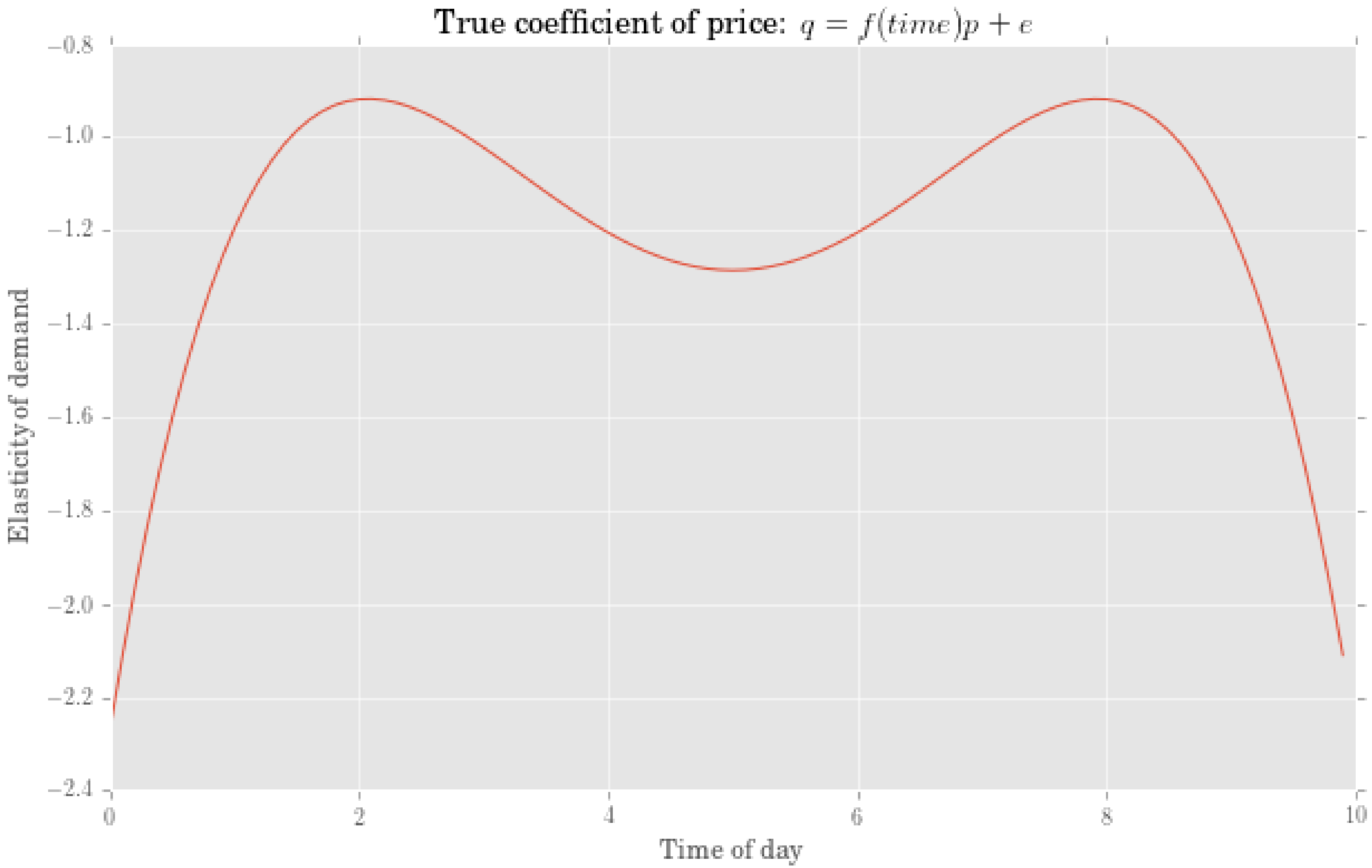
$$\min_{\theta\in\Theta} \frac{1}{N}\sum_i \left(y_i - \frac{1}{B}\sum_b g(\hat{p}_{ib}, x_i; \theta)\right)^2, \quad \hat{p}_{ib} \sim \hat{\mathrm{P}}(p|x_i, z_i)$$

Instead, *Stochastic Gradient Descent*: optimize via *unbiased* gradient estimates based upon mini-batch sample of the full dataset.

We can do SGD by pairing each observation with *two* treatment draws

$$\nabla g(\theta) \approx (y_i - g(\hat{p}_{i1}, x_i; \theta))\, g'(\hat{p}_{i2}, x_i; \theta)$$

Linear Demand, Heterogeneous Effects



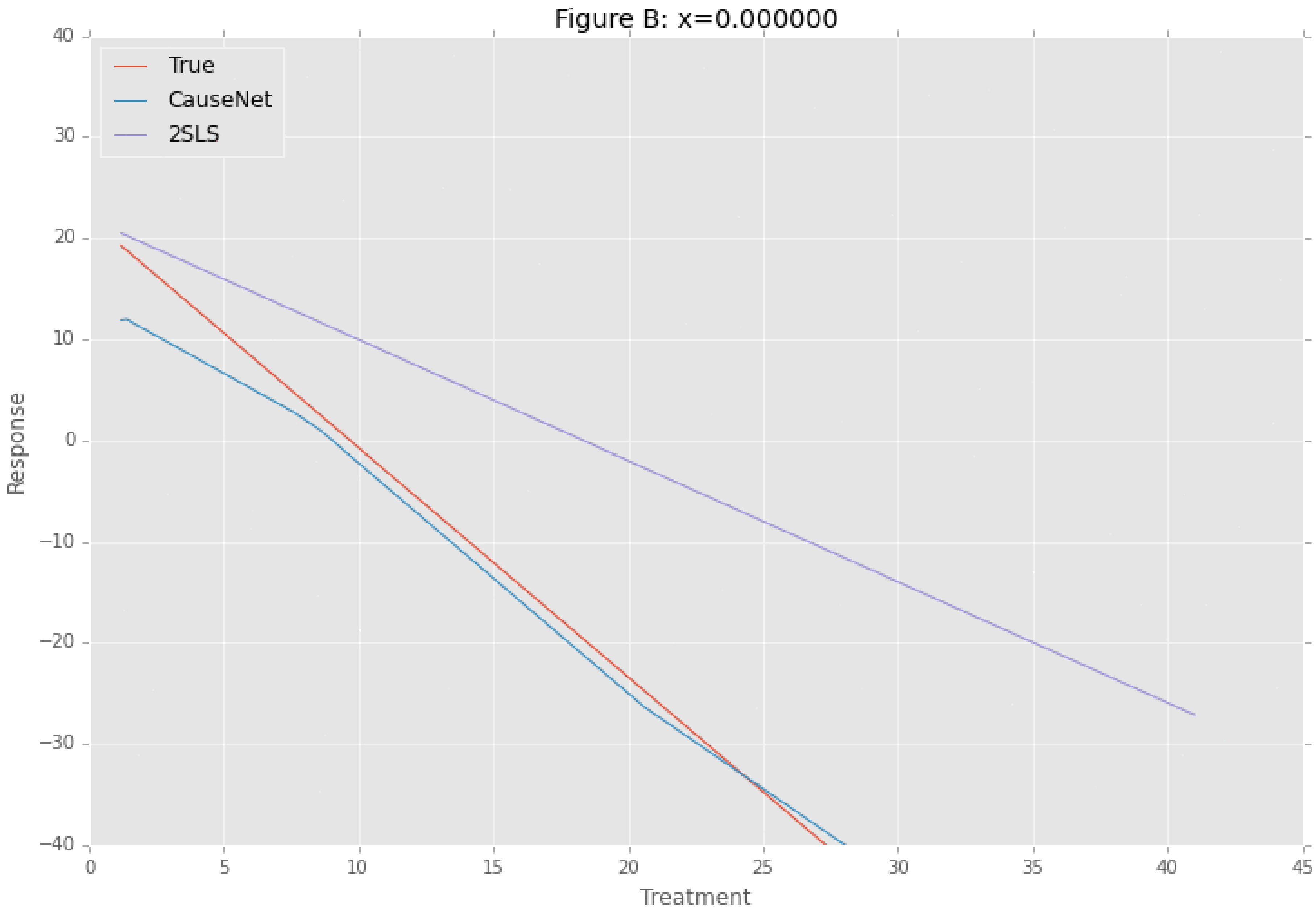
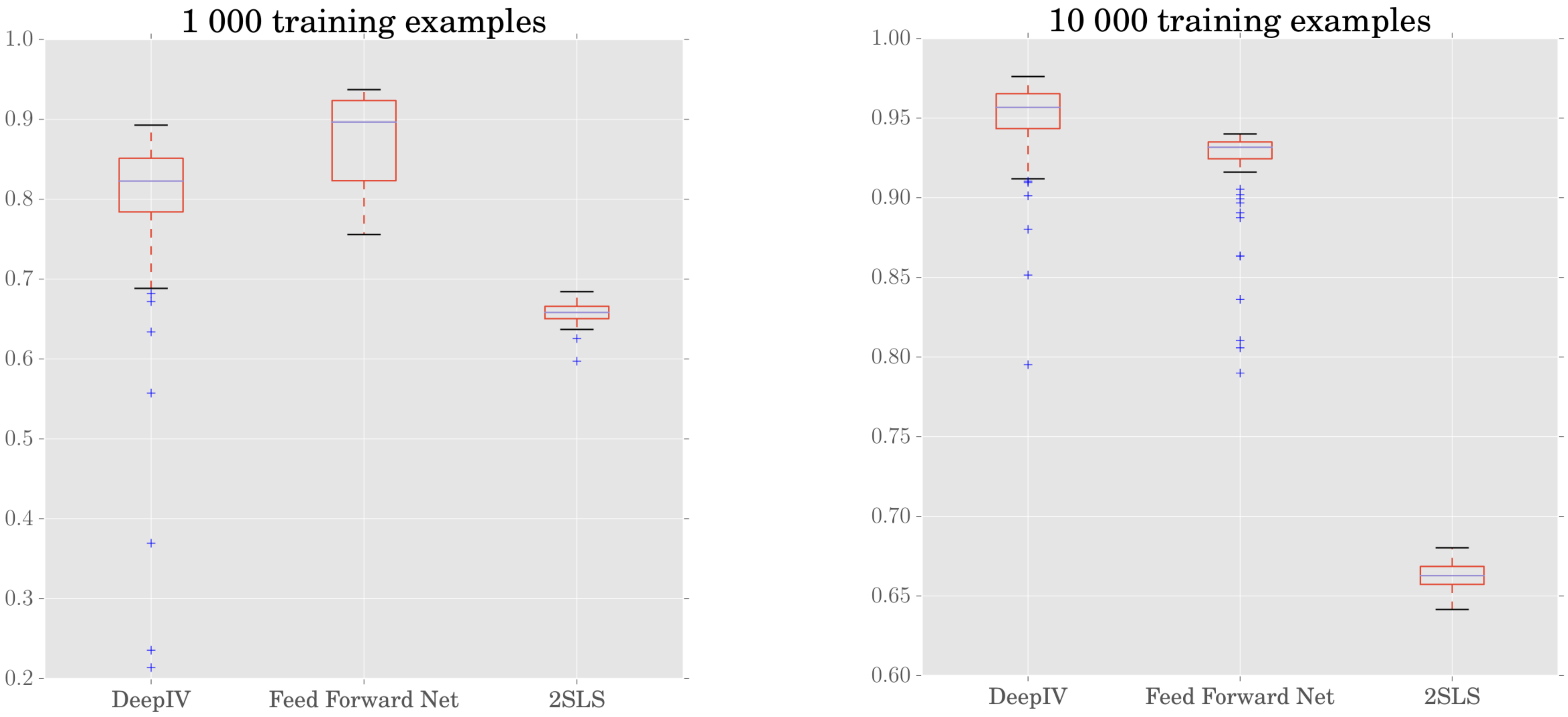


Figure B: x=0.000000
True
CauseNet
2SLS
Response
Treatment
40
30
20
10
0
−10
−20
−30
−40
0
5
10
15
20
25
30
35
40
45

Store Example: R2 Values



Ads Application

Taken from Goldman and Rao (2014)

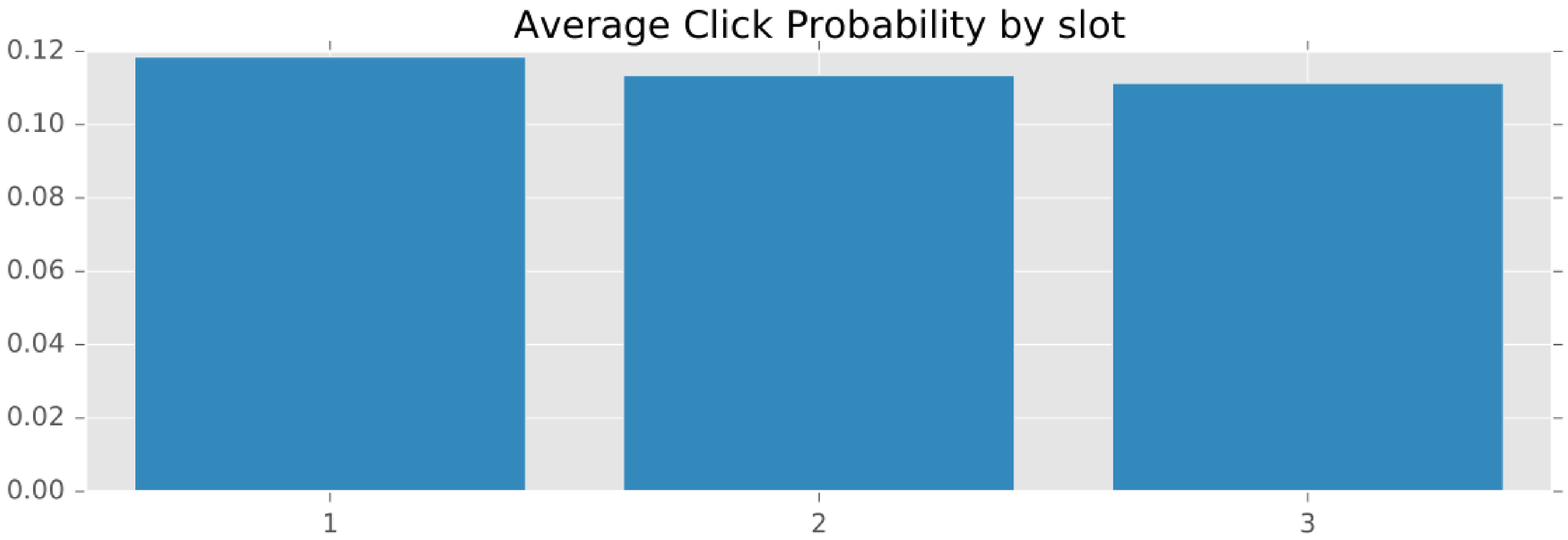
We have 74 mil click-rates over 4 hour increments for 10k search terms

Treatment: ad position 1-3

Instrument: background AB testing

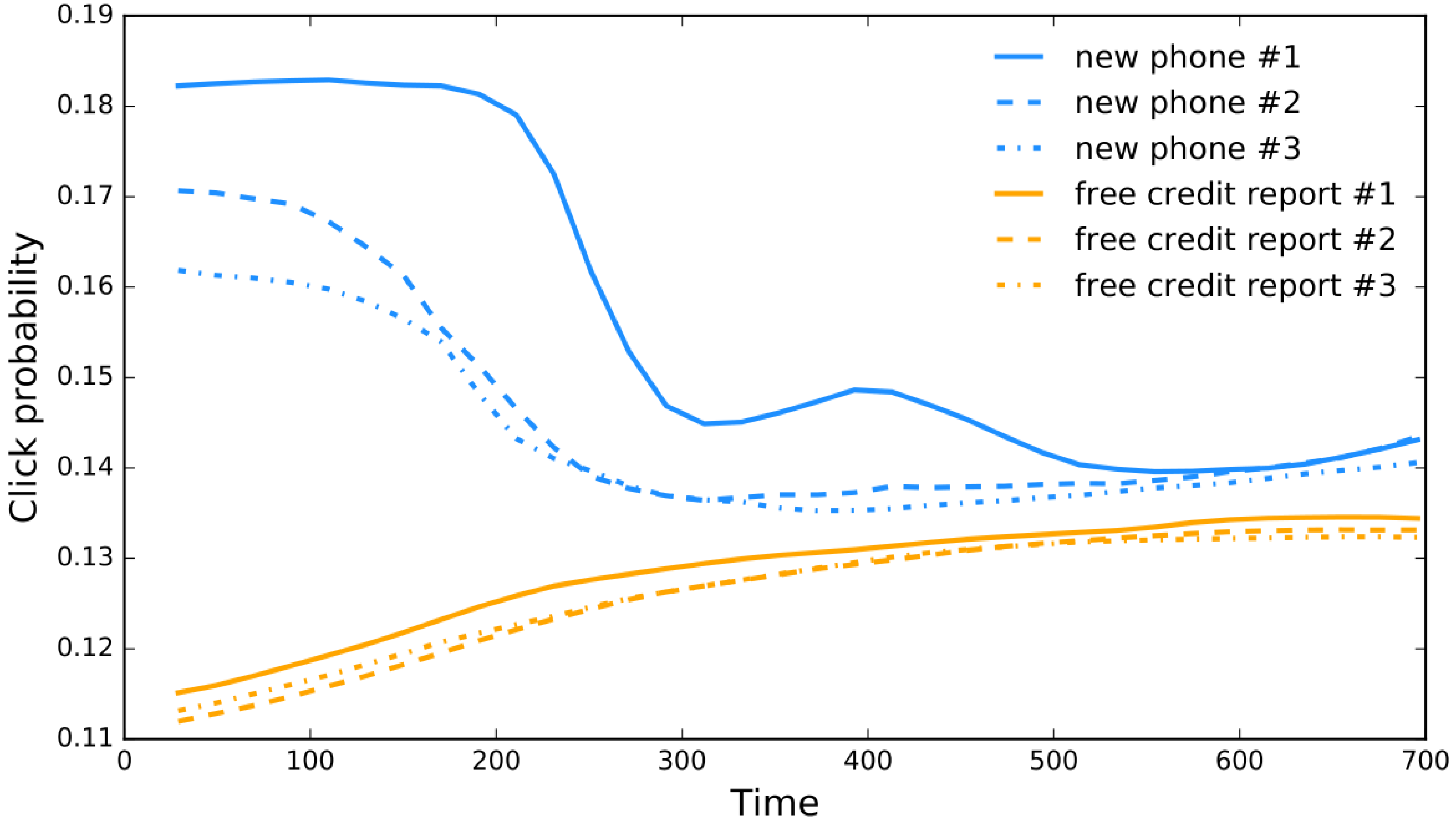
Covariates: search text and time

Average Treatment Effects



These compare to observed click probabilities of 0.33, 0.1, and 0.05.

Heterogeneous Treatment Effects



Economics and Artificial Intelligence

We have a track record pointing ML at questions of science + causation.
We're going to replicate this success at scale on unstructured data

We use economic theory to build systems of tasks that can be addressed with Deep nets and other state-of-the-art ML.

This is the construction of systems for Artificial *Economic* Intelligence.