JBES @ JSM Discussion
Bayesian optimization and ML Causal inference

Matt Taddy – http://taddylab.com

Microsoft Research and Chicago Booth

Contextual Bandits + Bayesian Optimization

Both optimize unknown f(x) while minimizing evaluations at sub-optimal x.

BO replaces expensive calls to f(x) with realizations from its posterior.

e.g., Schonlau+ 1998 EGO search to maximize $EI[f(x_{new})]$ *where* $I(f) = \max(f - f_{best}, 0)$.

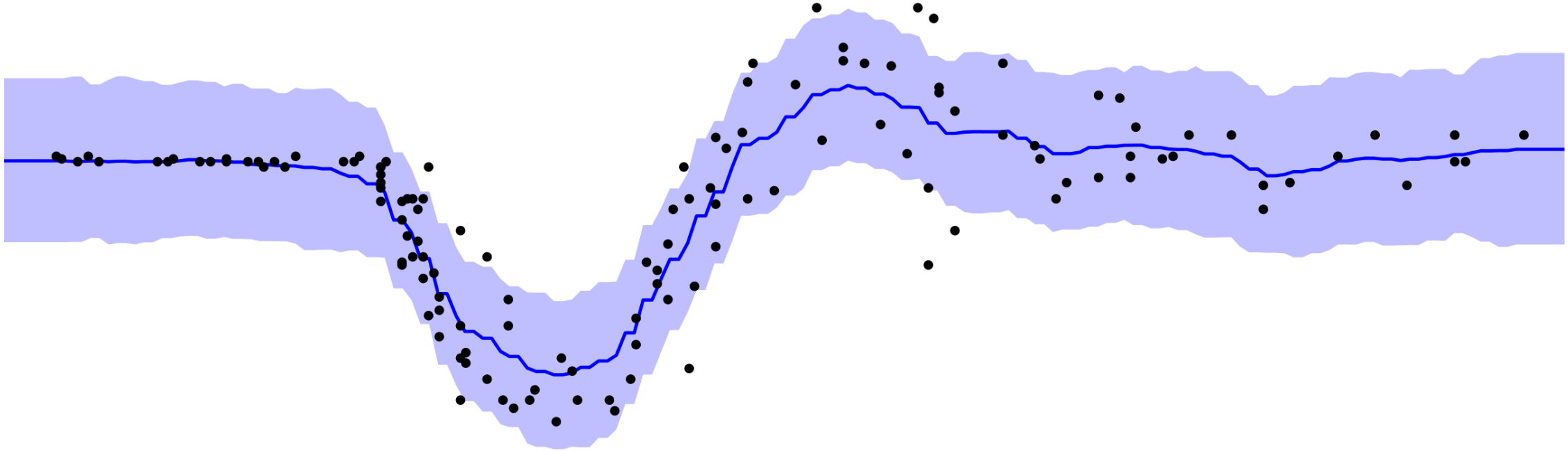
Traditional model for f is a GP, but it's just regression so bring your favorite.

T+ 2009, T+Gramacy+ 2011 and Gramacy+T+Wild 2013 all use Treed Gaussian Processes,
Snoek…Adams 2015 bring Deep Nets, Frazier GPs and the knowledge gradient at Yelp and Uber.

You can also do well combining with greedier search.

BART is [not not] parametric

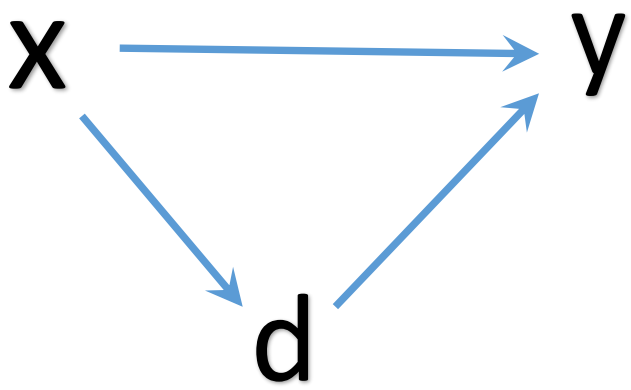
The mean is very flexible, but the additive error structure is not.



This may or may not hurt you in prediction problems (T+ 2015 ICML).

But in BO or Bandits, the differences in posterior uncertainty are key.

Backdoors and causation



If you know $y = g(x, d)$ then
$\bar{g}(d) = \int g(x, d)\partial Px$ is the *causal effect* of d on y.

To an engineer, it's the *Main Effect* of d in system g.

Esp. in sensitivity analysis; see Saltelli primer and Gramacy + T 2012 JSS.

Measurement error on g is mostly ignored.

Getting it right inside the box

Validity of the adjustment formula requires accurate $g(x, d)$.

All of the internal partial dependencies need to be correct.

Say $d = x\varphi + \nu$ and $y = d\gamma + x\beta + \varepsilon$.
We need to ensure that $\hat{\gamma} \approx \gamma$ not $\hat{\gamma} \approx \gamma + \beta\varphi$.

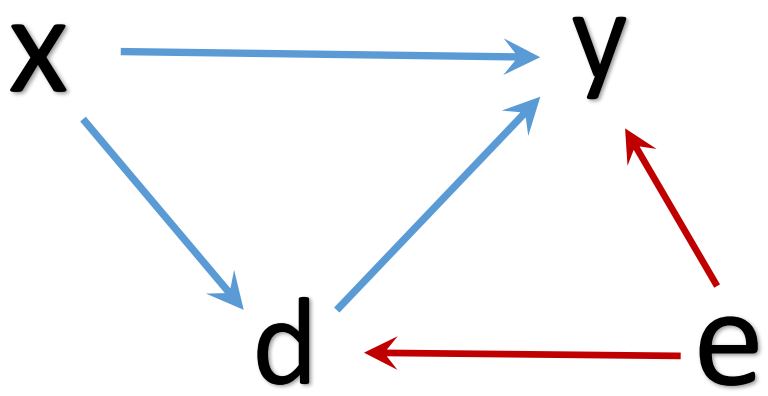
Simple OLS does a good job for low-D x, but its tough in HD

Great recent work from Econ this: Belloni, Chernozhukov, Hansen, ++.

Getting this right in $\hat{g}$ is the only thing that makes backdoor adjustment better than, say, simply running the naïve marginal regression $y \sim d$.

Causal Epsilon

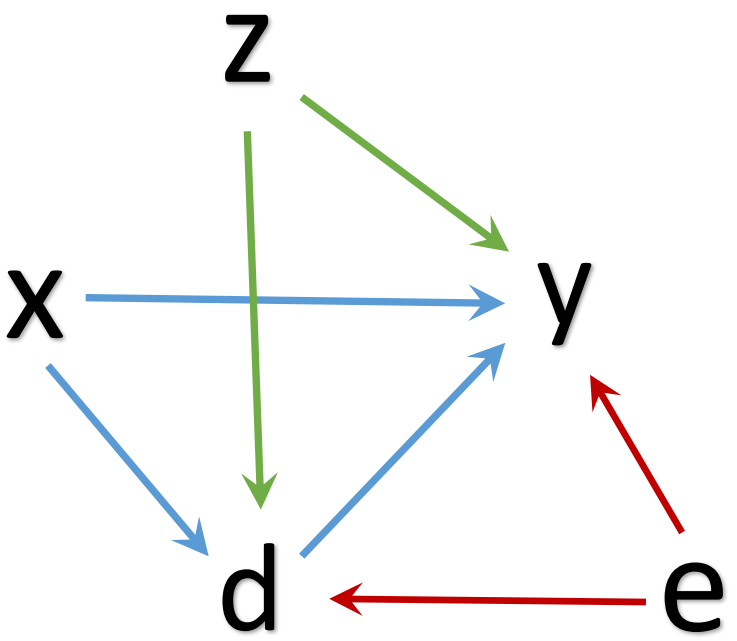
We also worry about left-over causes beyond observables,
and that these might also influence the treatment.
e.g., being downwind of downtown in NOX example.



The model is actually $y = g(x, d) + e$,
and we have an non-adjustable backdoor

Life is tough without observables! You can't model the counterfactual.
Only a few special designs (e.g., IV or RD) lead to a workable problem.

Instruments



In IV we have a special $z \perp e$.

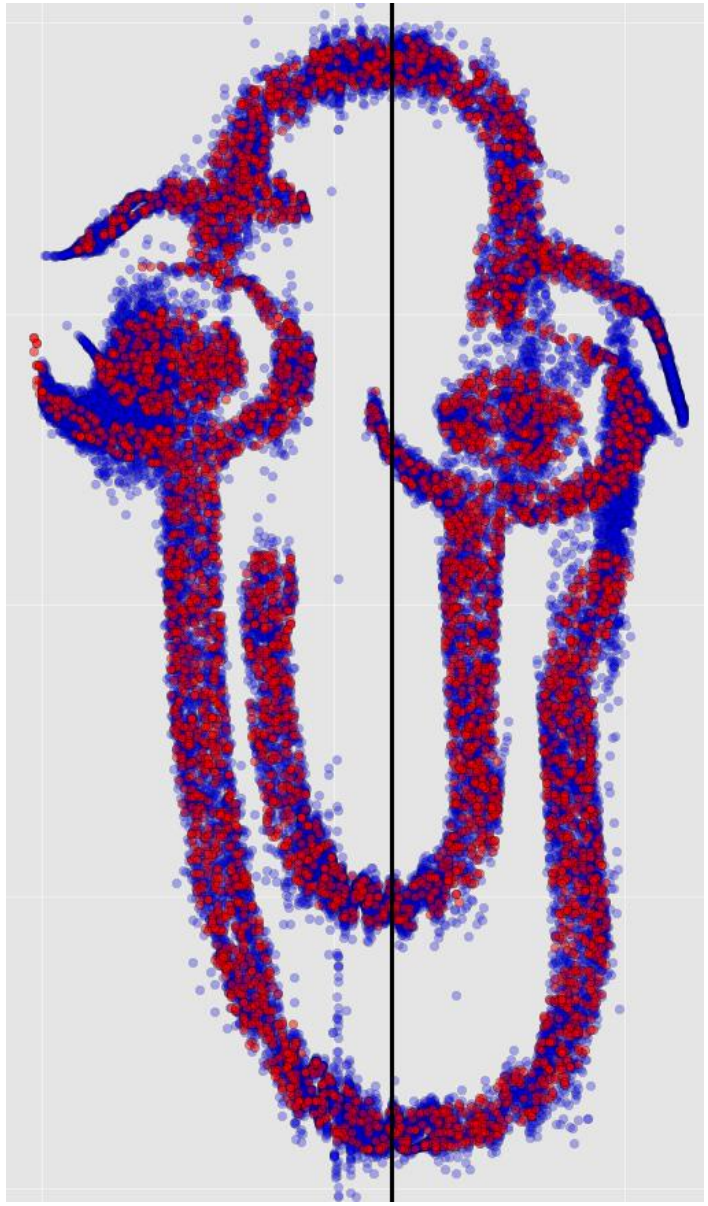
Then $E[y|x,z] = \int g(x,d)\partial P(d|x,z)$.

We have $\{d_i\} \sim P(d|x_i,z_i,e_i)$.

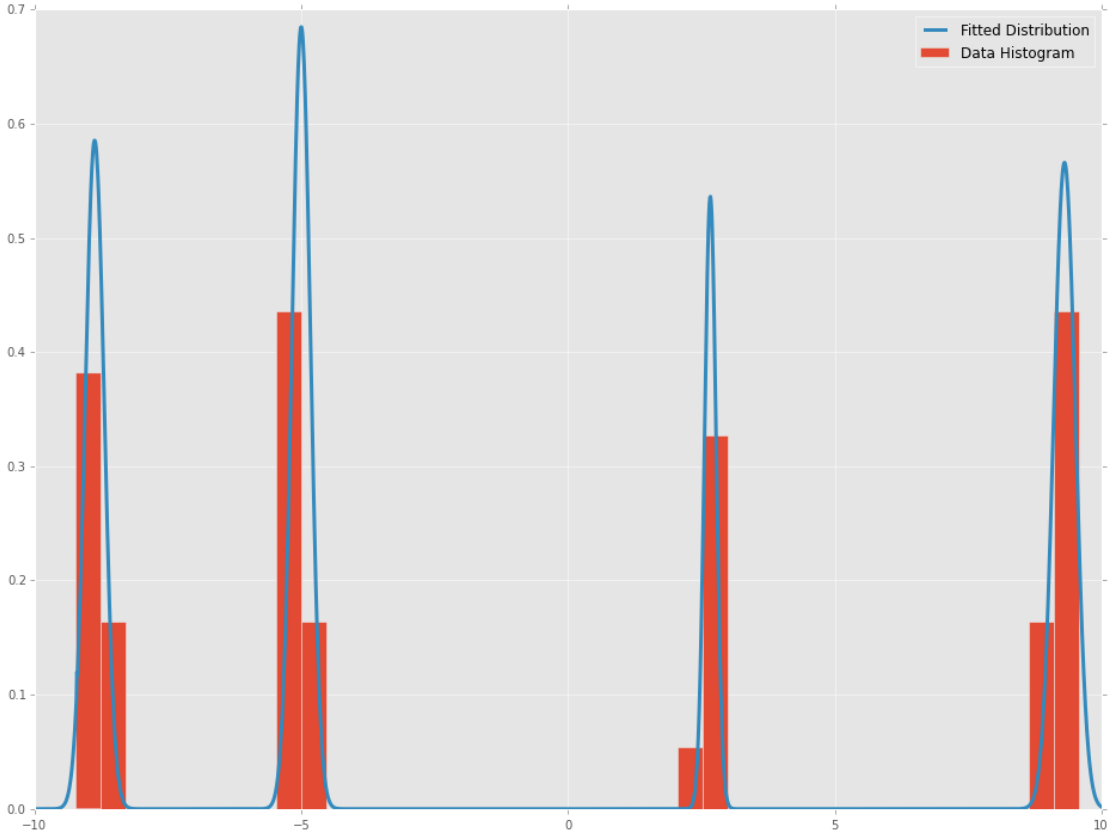
If we can instead get $\{\hat{d}_i\} \sim P(d|x_i,z_i)$, then 'black box'

machine learning for $y \sim x,\hat{d}$ will recover the correct $\hat{g}$.

You need to direct your ML first at learning $P(d|x,z)$.

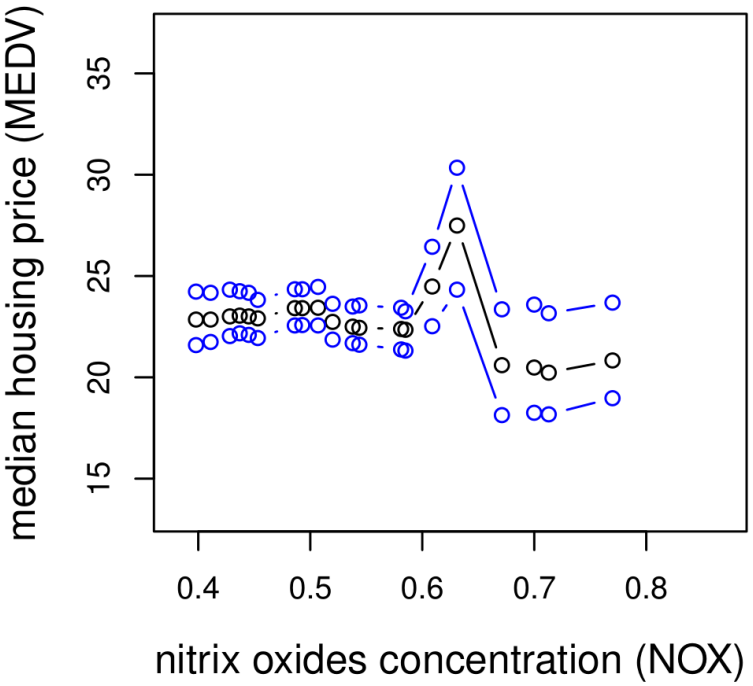


Luckily, with contemporary ML we can learn all sorts of crazy conditional densities.



with Jason Hartford (UBC + MSR) and Greg Lewis (MSR).

NOX Experiment



Always anchor yourself by imagining an experiment where you get to change the treatment independent of all other influences.