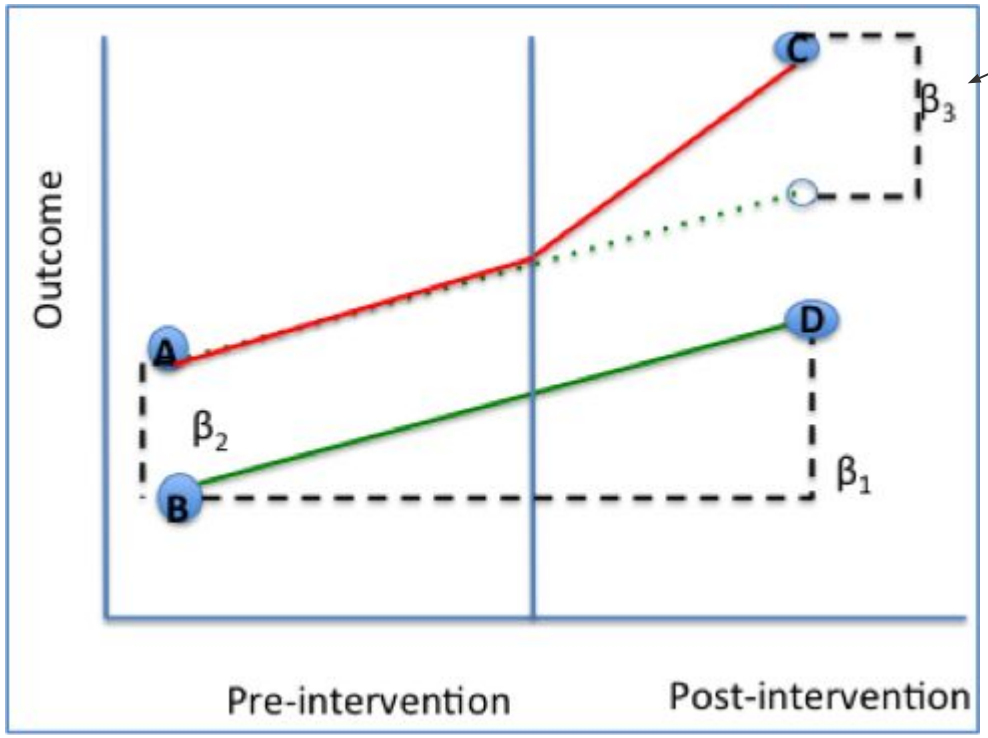

**#ml-papers May 2019**

**Inferring Causal Impact Using Bayesian Structural Time-Series Models**

# Intro: Measuring Causal Impact



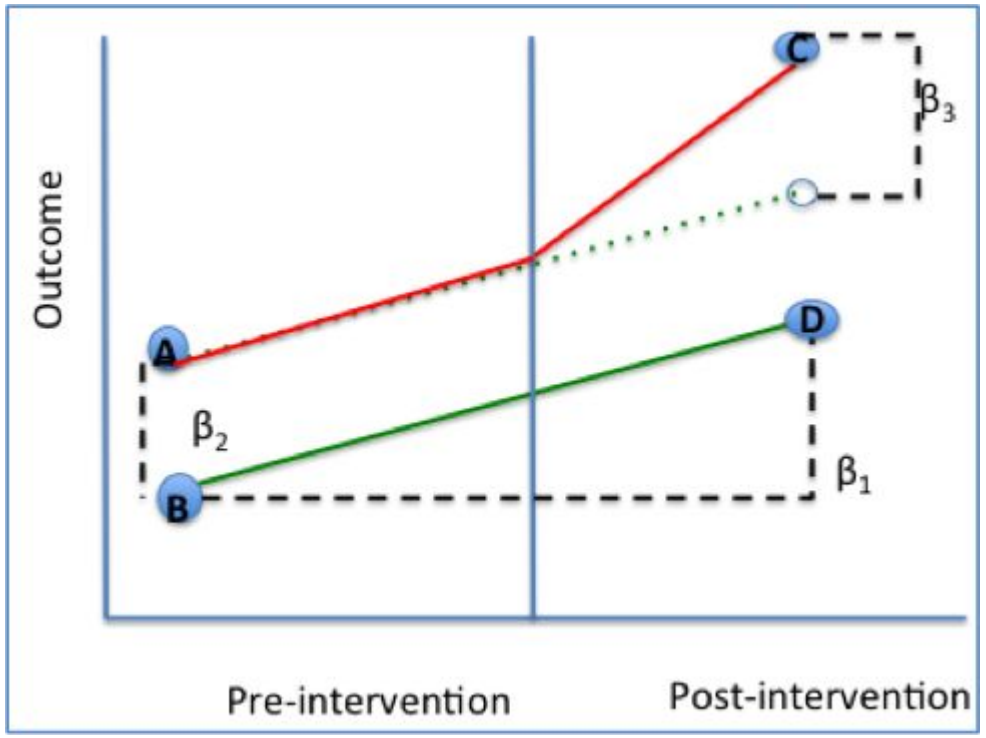

Diff in diffs: how we typically measure causal impact. **Requires a control group.**

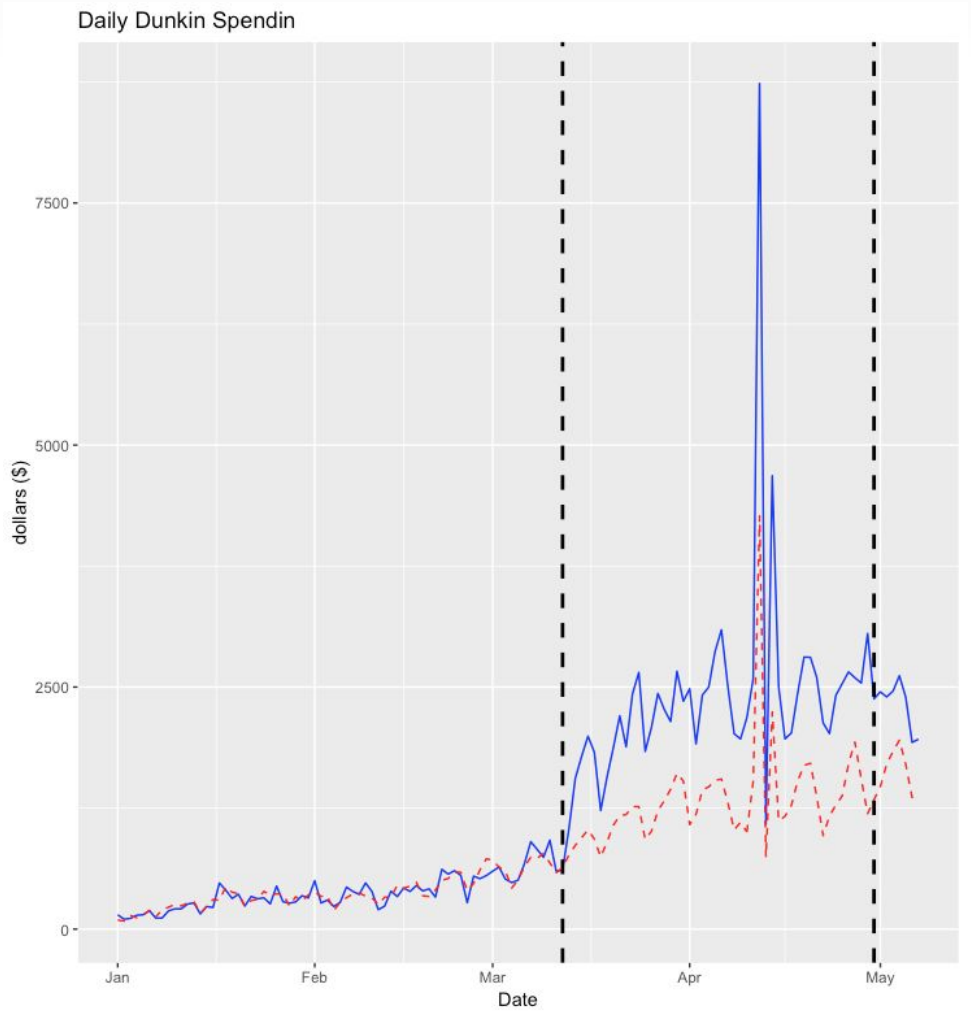
Paper cites 3 common problems with this approach:

1) Assumes i.i.d despite temporal nature
2) Typically considers 2 time periods, before and after intervention. Ignores how the effect evolves over time.
3) Previous time series based DD techniques restrict how we construct controls

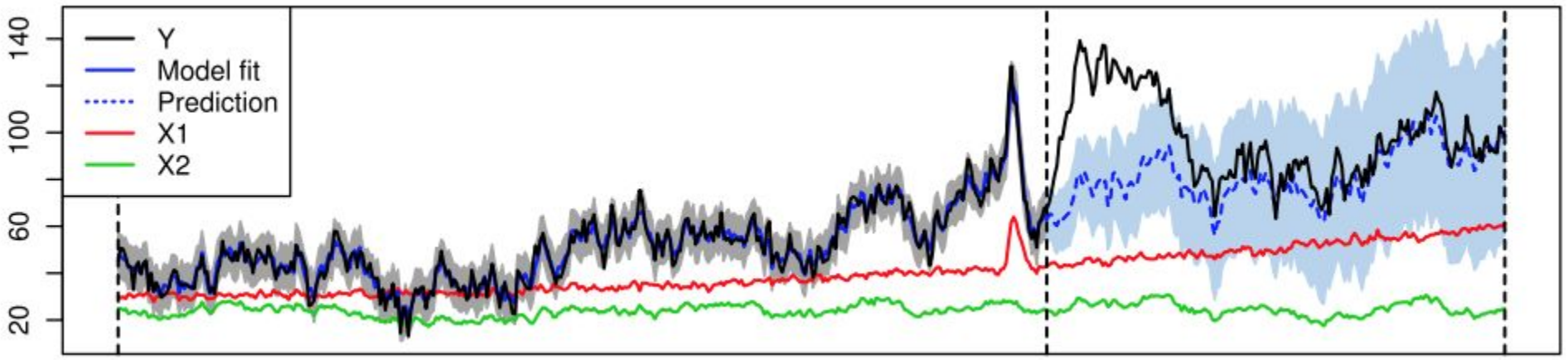
**but the real problem is that we can't always choose to apply treatment to part of the population**

# Intro: Estimating the Counterfactual



# Intro: What do we use to ~the CF




We can think of our **outcome** time-series as a combination of local trends, seasonal effects and "**contemporaneous covariates" (X1, X2..)**. This is where the "structural" word comes from, we're constructing ts from other ts.

Covariates should exhibit similar patterns as Y, but their relationship with Y needs to be unaffected by the treatment.

Covariate example for debit spend? Covariate example for FB bidding?

# BSTS Models: what is a state-space model?

- States change through time steps
- The next state (alpha t+1) is a function of the previous state and a learned **transition matrix (T)**
- State models have no memory of past states, any info based on past has to be encoded into the current state (Markov Property).. parallels with RNNs which feed time step gradients to the next iteration
- Parallels with RL, but no branching factor because we're not making decisions

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t,$$

# BSTS Models: what is part of "state"?

The "state" is a vector with 3 key components:

1) Local Linear trend -> local changes in Y & momentum

$$\mu_{t+1} = \mu_t + \delta_t + \eta_{\mu,t},$$

$$\delta_{t+1} = \delta_t + \eta_{\delta,t},$$

2) Seasonality -> parameterized seasonality equations lets us try apply different seasonal effects as we like

$$\gamma_{t+1} = -\sum_{s=0}^{S-2} \gamma_{t-s} + \eta_{\gamma,t},$$

3) Covariates -> these help control our Y to external influences we can't measure

$$\mathbf{x}_t^{\mathrm{T}} \beta_t = \sum_{j=1}^{J} x_{j,t} \beta_{j,t},$$

$$\beta_{j,t+1} = \beta_{j,t} + \eta_{\beta,j,t},$$

# BSTS Models: what is a state-space model?

Now that you know the state is made up of a bunch of components:

**Rt** controls the error term according to the covariances of the state components.

Intuition: if components errors covary the total error could potentially be much higher than if they were conjugate.

Also.. error estimates grow as we move further into the post-treatment period because of cumulative effect.. you're making estimates based on previous time step estimates so potential error gets bigger with every step

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t,$$

# BSTS Models: what is a state-space model?

Once we've constructed our "state" we can iterate through time and learn our **T**, but predicting the next state is only part of the model..

We have to use the state to predict the Yt. Zt is learned, e is error.

$$\alpha_{t+1} = T_t \alpha_t + R_t \eta_t,$$

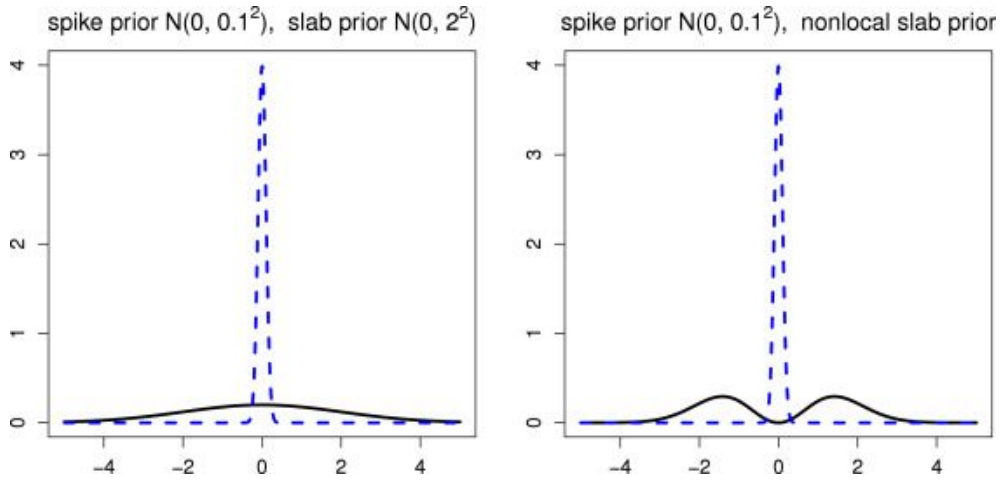
$$y_t = Z_t^{\mathrm{T}} \alpha_t + \varepsilon_t,$$

# BSTS Models: benefits

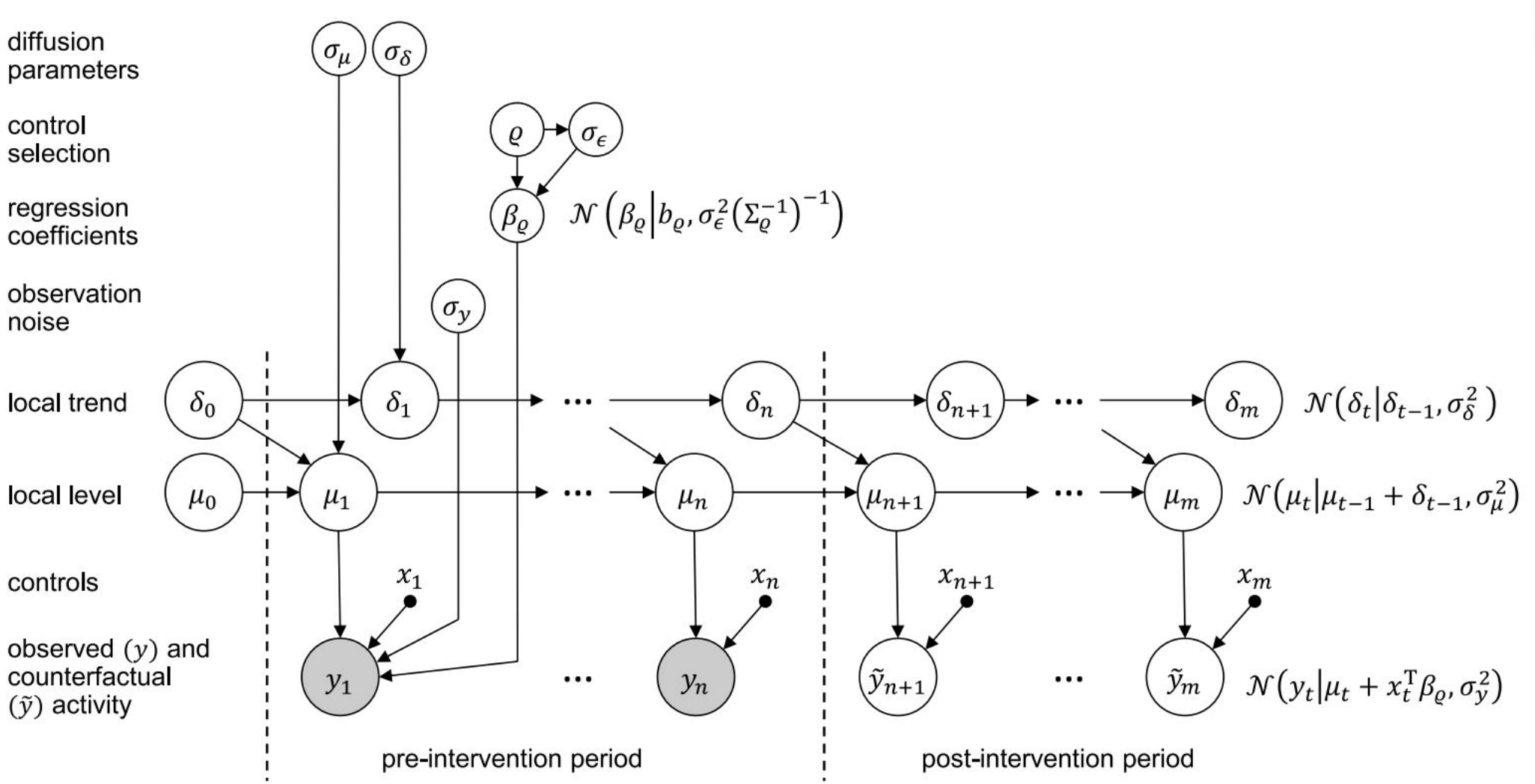
Structural time-series models are useful in practice because they are flexible and modular. They are flexible in the sense that a very large class of models, including all ARIMA models, can be written in the state-space form given by (2.1) and (2.2). They are modular in the sense that the latent state as well as the associated model matrices $Z_t$, $T_t$, $R_t$, and $Q_t$ can be assembled from a library of component sub-models to capture important features of the data. There are several widely used state-component models for capturing the trend, seasonality or effects of holidays.

# BSTS Models: final notes on Bayesian stuff

- **Priors**: to initialize this model we have to make some prior assumption on the model weights (theta) and a posterior on the initial state p(alpha[t=0] | theta).. paper uses "diffuse" priors, also called "uninformative" priors which are weak assumptions that will be corrected for by the time we get to the treatment period. They specifically advocate using a "spike-and-slab" prior, which is a Bayesian feature selection method that allows us to incorporate a lot of different X controls into the state and let the model learn which ones are useful.
- **MCMC**: since we're using Bayesian framework it's important to remember that as we go through state-space we're storing **state and parameter distributions** (not actual values) and **sampling** from these distributions using monte carlo sampling

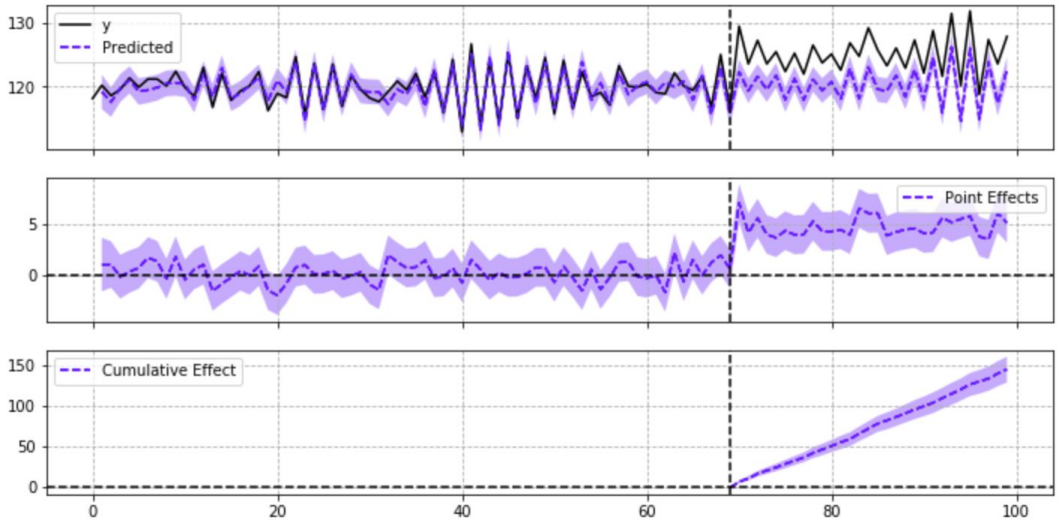


# BSTS Models: one picture



# BSTS Models: usage

```
import numpy as np
import pandas as pd
from statsmodels.tsa.arima_process import ArmaProcess
from causalimpact import CausalImpact
np.random.seed(12345)
ar = np.r_[1, 0.9]
ma = np.array([1])
arma_process = ArmaProcess(ar, ma)
X = 100 + arma_process.generate_sample(nsample=100)
y = 1.2 * X + np.random.normal(size=100)
y[70:] += 5
data = pd.DataFrame({'y': y, 'X': X}, columns=['y', 'X'])
pre_period = [0, 69]
post_period = [70, 99]
ci = CausalImpact(data, pre_period, post_period)
print(ci.summary())
ci.plot()
```




Read this!

# Application to Simulated Data

★ Simulated data:
- Dynamic regression component w. 2 contemporaneous covariates

$$y_t = \beta_{t,1} z_{t,1} + \beta_{t,2} z_{t,2} + \mu_t + \varepsilon_t$$

$\mu_t \sim \mathcal{N}(\mu_{t-1}, 0.1^2)$ : Local level; evolves with a random walk

$\beta_t \sim \mathcal{N}(\beta_{t-1}, 0.01^2)$ : Covariate Coefficient; evolves with a random walk

- Post intervention simulation: *(1+e)*Y_t*

# Application to Simulated Data

- ★ Causal effect:
  - central 95% posterior prob of cumulative effect excludes 0

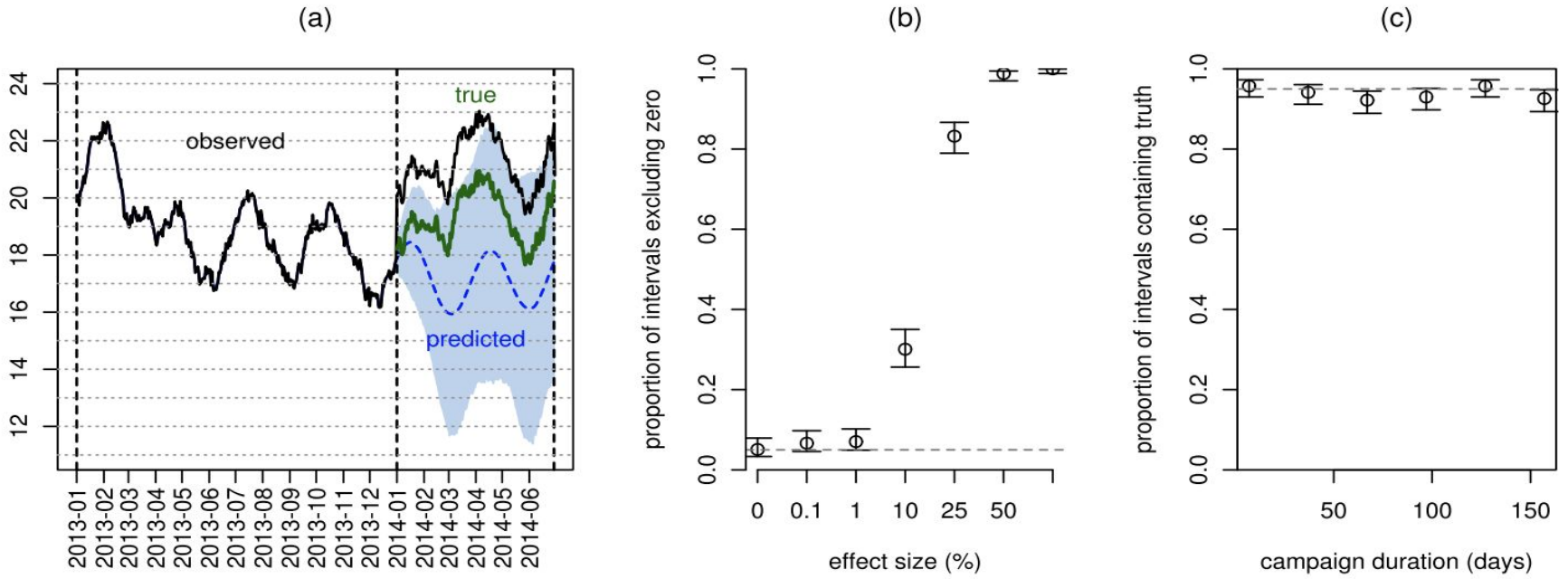
- ★ Estimation accuracy:
  - absolute percentage estimation error:

$$a_{i,t} := \frac{|\hat{\phi}_{i,t} - \phi_t|}{\phi_t}.$$

  - Forecast period +, accuracy -
    - under local linear trend model
    - standard deviation of random walk of local level tripled

# Application to Simulated Data

- ★ Sensitivity & specificity
  - b: #causal/#all: Effect size +, correct detection +
  - c: #(95% interval w. causal detected)/#all: indep. from duration



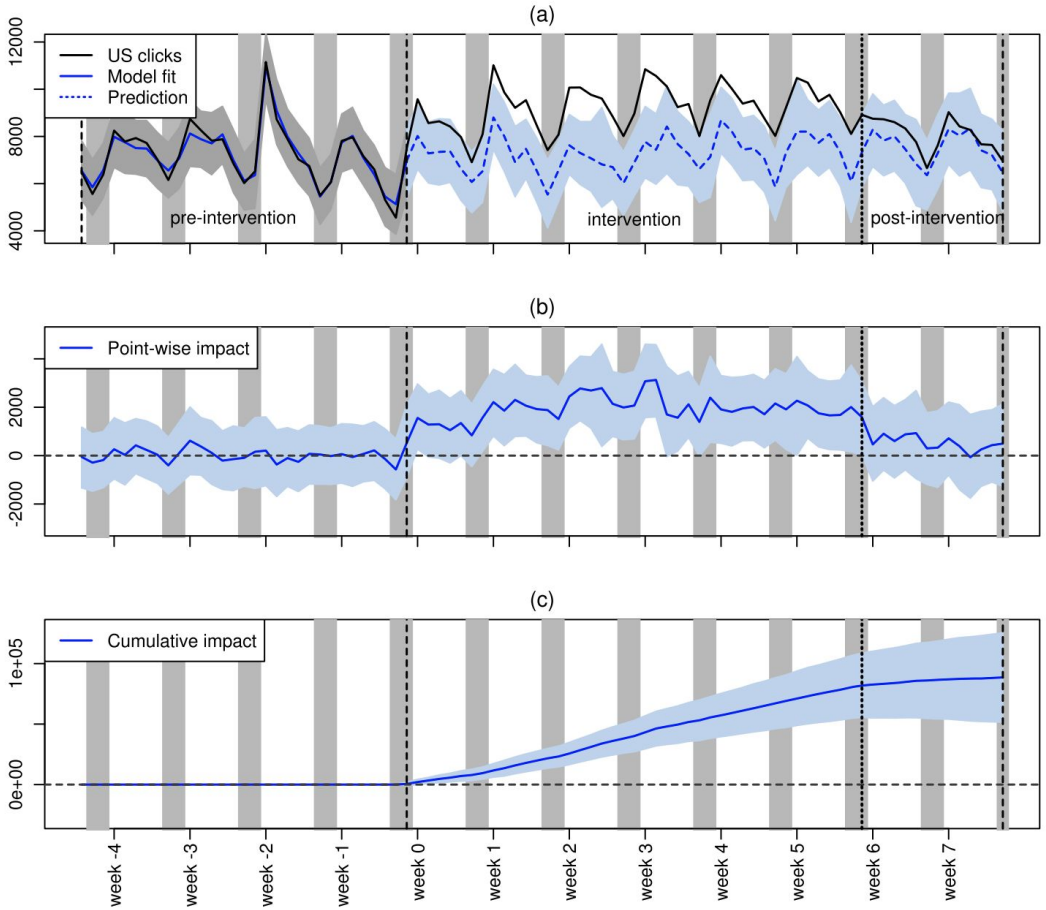
# Application to Empirical Data

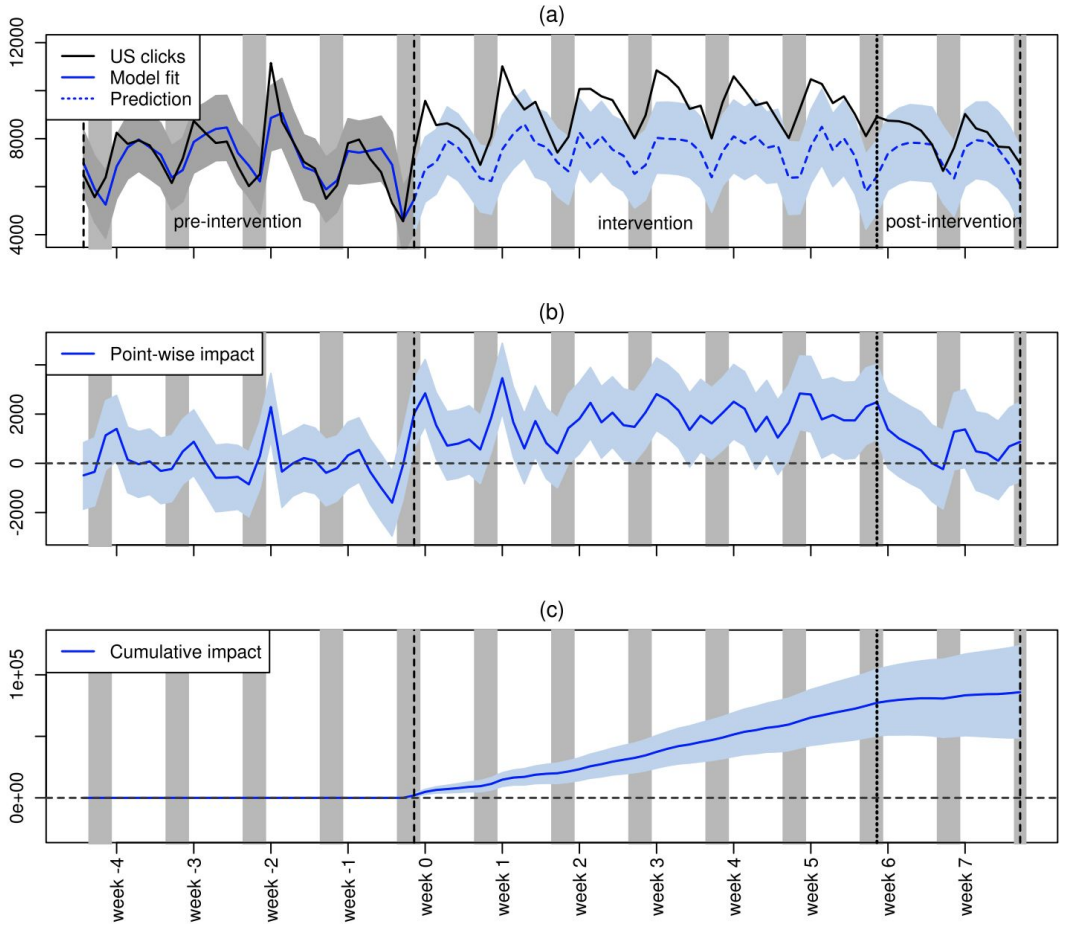
- ★ Google ads campaign
  - Control/experiment randomly split;
    then aggregate experiment regions
  - Outcome: total clicks (organic clicks + paid clicks)
  - Components:
    - 1. local level: inverse-Gamma prior
    - 2. static regression: spike-and-slab prior

# Application to Empirical Data

★ Effect on the treated

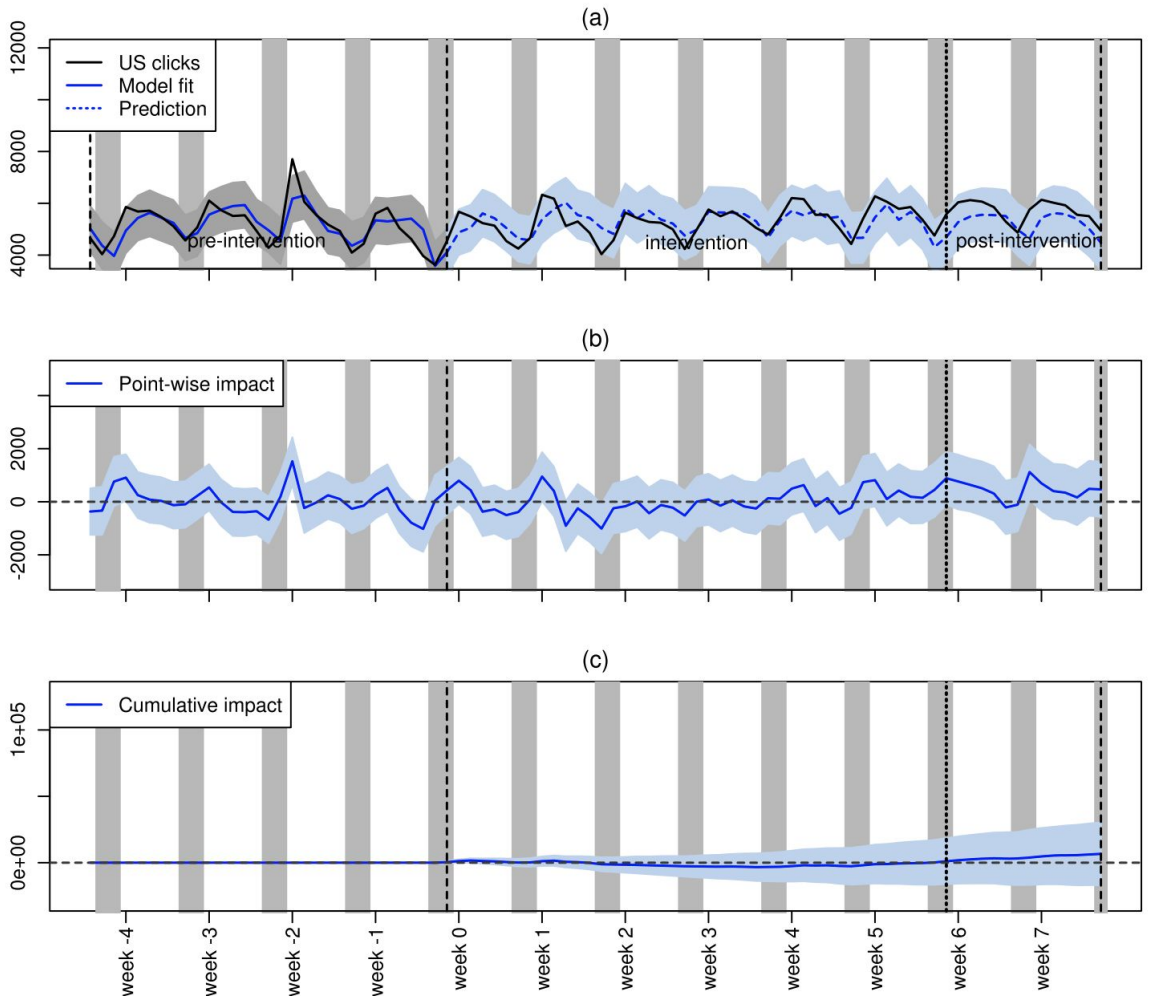
- Randomized control (control regions) vs. Observational control (discard control regions, use keyword search results)



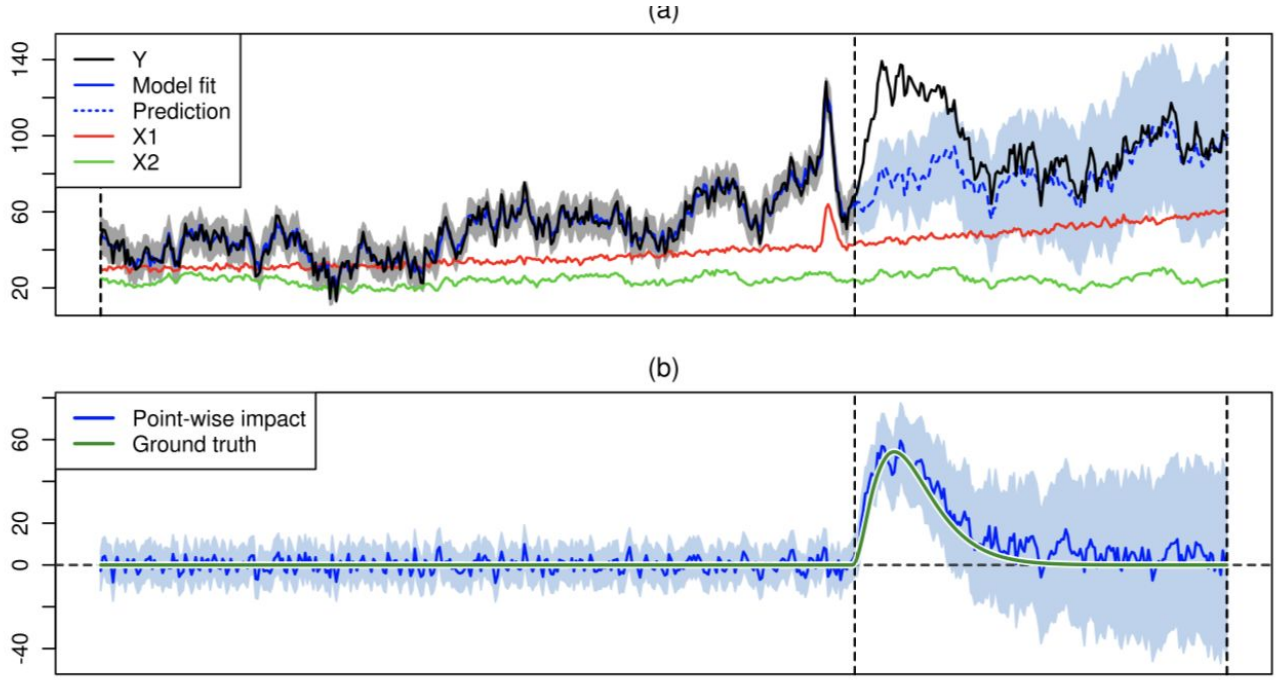
# Application to Empirical Data

★ Absence effect on the control



# Discussion

- The paper introduces a Bayesian structural time series model to estimate the effect of market interventions in time series data, generalizing the classical differences-in-differences approach.
- The main idea is to build a Bayesian time series model that predicts the evolution of the observed outcome, taking into account various predictors, such as seasonality, outcomes in control groups, and external factors.
- Once a model is fit, it can be used to predict the future outcome if no treatment was applied.
- We then subtract the observed future outcome (after treatment) from the predicted outcome obtained by the model. The difference is the estimated effect.



# Discussion (cont.)

- In contrast to many other methods, the approach introduced is fully Bayesian, meaning all the parameters are estimated via sampling. There is no closed form solution, so the recommended approach to estimate the model is via Markov Chain Monte Carlo (MCMC).
- Because the model is calculated using a large number of Markov Chain samples, uncertainty estimates can also be calculated.
- While the main application of this method is to determine the effect of an intervention, it can also be used for power-analysis. Even if no intervention occurred, we can pretend one occurred at a point in time and estimate the uncertainty in a hypothetical effect.
- One potential risk with the proposed method is it can lead to incorrect results if the treatment effect correlates with the likelihood of being treated ("activity bias").
- The authors created an R package CasaulImpact, making it fairly easy to use the proposed method. They state that all their experiments ran in < 30 seconds.