

# Approximate Inference in Industry: Two Applications at Amazon

Cédric Archambeau

[cedrica@amazon.com](mailto:cedrica@amazon.com)

Workshop on Advances in Approximate Bayesian Inference,  
Neural Information Processing Systems, 2017



Amazon.com, 2017





Today, machine learning is  
creating a paradigm shift.

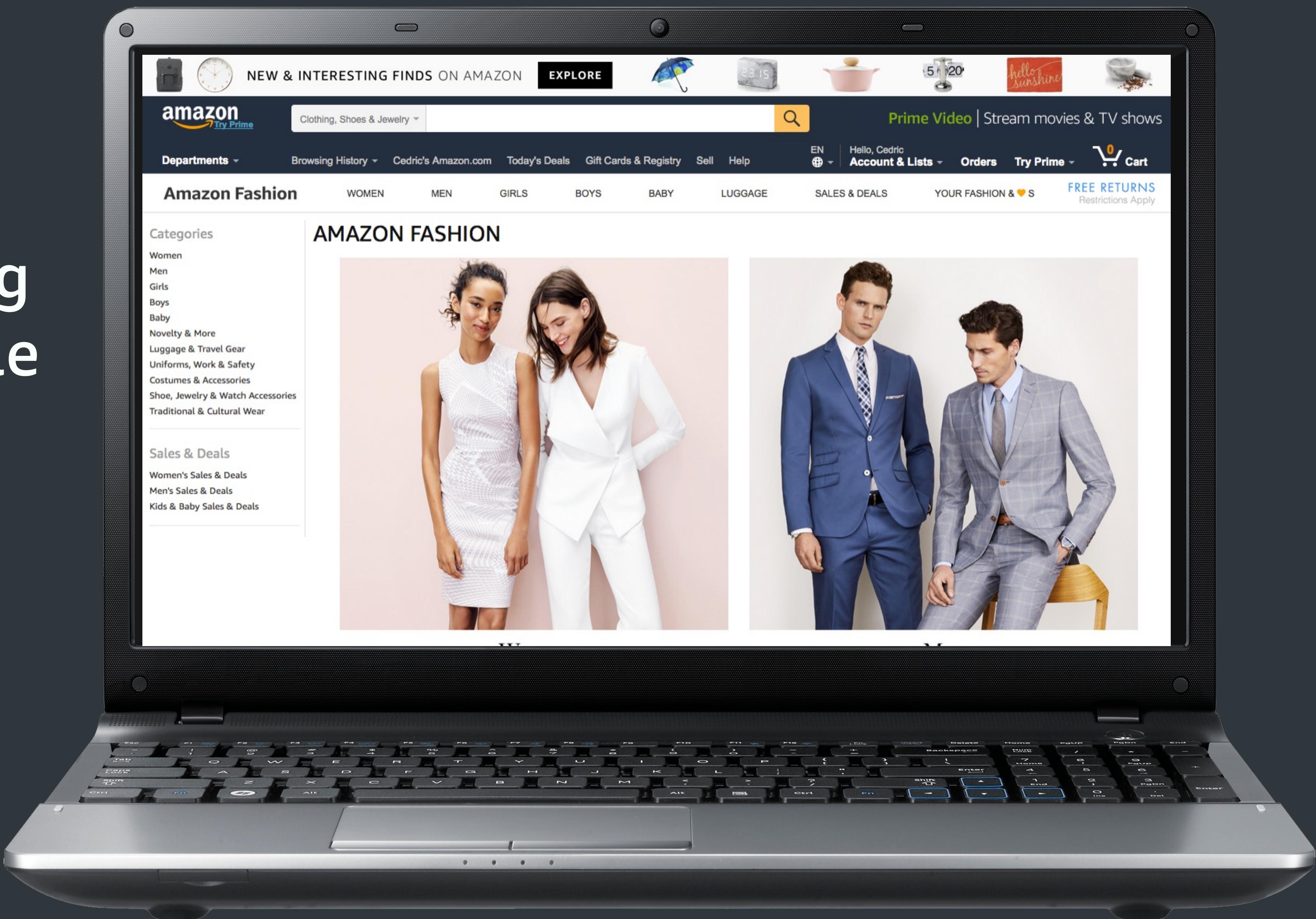
"It is a golden age. Machine learning and AI is a horizontal enabling layer. It will empower and improve every business, every government organization, every philanthropy."

Jeff Bezos in Geekwire (May 6, 2017)

$$1 + 1 =$$

$$[(27/3)/3] - 1$$

# Understanding Fashion & Style



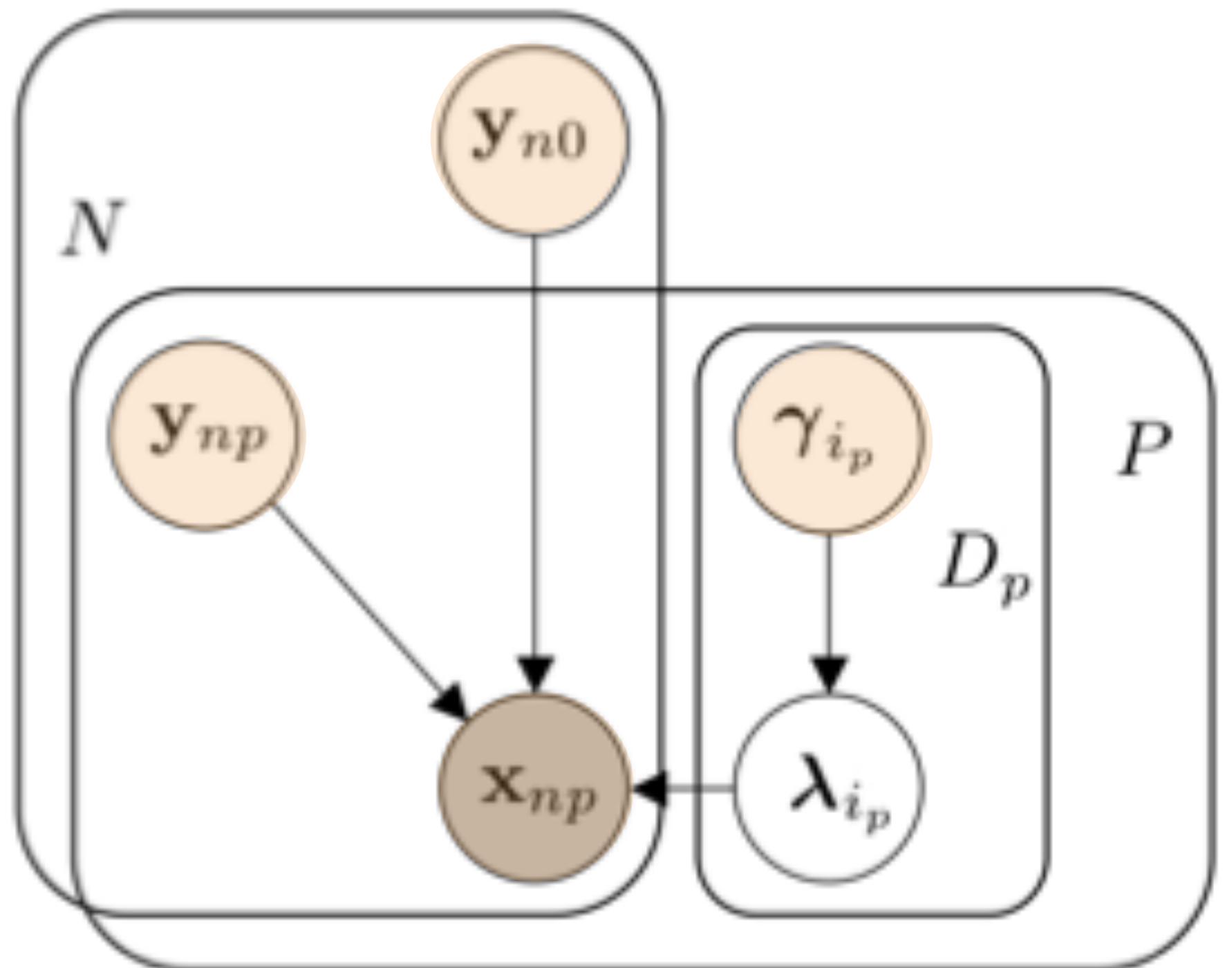
B00GIZOIDQ

null

Hair Drawing #4



# Shallow Latent Variable Model



(Archambeau and Bach, NIPS 2008)



# Variational Inference

- Using *Jensen's inequality*, we get for any distribution  $q_{\mathbf{w}}(\mathbf{Z}, \theta)$ :

$$\begin{aligned}\ln p(\mathbf{X}) &= \ln \iint p(\mathbf{X}, \mathbf{Z}, \theta) d\mathbf{Z} d\theta \\ &\geq \iint q_{\mathbf{w}}(\mathbf{Z}, \theta) \ln \frac{p(\mathbf{X}, \mathbf{Z}, \theta)}{q_{\mathbf{w}}(\mathbf{Z}, \theta)} d\mathbf{Z} d\theta \\ &= \ln p(\mathbf{X}) - \text{KL}[q_{\mathbf{w}}(\mathbf{Z}, \theta) \| p(\mathbf{Z}, \theta | \mathbf{X})] \triangleq -\mathcal{F}(\mathbf{w}).\end{aligned}$$

- A tractable solution is found by assuming  $q_{\mathbf{w}}$  factorises given the data:

$$q_{\mathbf{w}}(\mathbf{Z}, \theta) = \prod_n q(\mathbf{z}_n; \mathbf{w}_n) \times \prod_m q(\theta_m; \mathbf{w}_m).$$

# Stochastic Variational Inference

$$\mathbf{w}_n \leftarrow \arg \max_{\mathbf{w}_n} \quad \langle \ln p(\mathbf{x}_n | \mathbf{z}_n, \theta) \rangle - \text{KL}[q(\mathbf{z}_n; \mathbf{w}_n) || p(\mathbf{z}_n)],$$

$$\mathbf{w}_m \leftarrow \arg \max_{\mathbf{w}_m} \quad \sum_n \langle \ln p(\mathbf{x}_n | \mathbf{z}_n, \theta) \rangle - \text{KL}[q(\theta_m; \mathbf{w}_m) || p(\theta_m)].$$

Let  $\ell_n(\mathbf{w}) = \langle \ln p(\mathbf{x}_n | \mathbf{z}_n, \theta) \rangle$ :

$$\mathbf{w}_m \leftarrow \mathbf{w}_m + \rho_t N \frac{\partial}{\partial \mathbf{w}_m} \left( \ell_n(\mathbf{w}) - \frac{\text{KL}[q(\theta_m; \mathbf{w}_m) || p(\theta_m)]}{N} \right),$$

where  $\sum_t \rho_t = \infty$  and  $\sum_t \rho_t^2 < \infty$ .

# Incremental Variational Inference



# Demand Forecasting

amazon.co.uk Try Prime

Women's Necklaces

Shop by Department

Your Amazon.co.uk Today's Deals Gift Cards & Top Up Sell Help

Hello. Sign in Your Account

Prime

Amazon Fashion WOMEN MEN KIDS & BABY LUGGAGE BRANDS

Women > Necklaces

elements<sup>TM</sup>  
SILVER

Elements Silver 925 Ladies' Heart Tag T-Bar Sterling Silver Necklace of 46 cm

[Elements Silver and Elements Gold Jewellery Store](#)

★★★★★ 30 customer reviews

Price: £88.99 & FREE Delivery in the UK. [Delivery Details](#)

In stock.

Want it delivered by Monday, 24 Apr.? Order within 11 hrs 34 mins and choose **Priority Delivery** at checkout. [Details](#)

Dispatched from and sold by Amazon. Gift-wrap available.

Note: This item is eligible for **click and collect**. [Details](#)

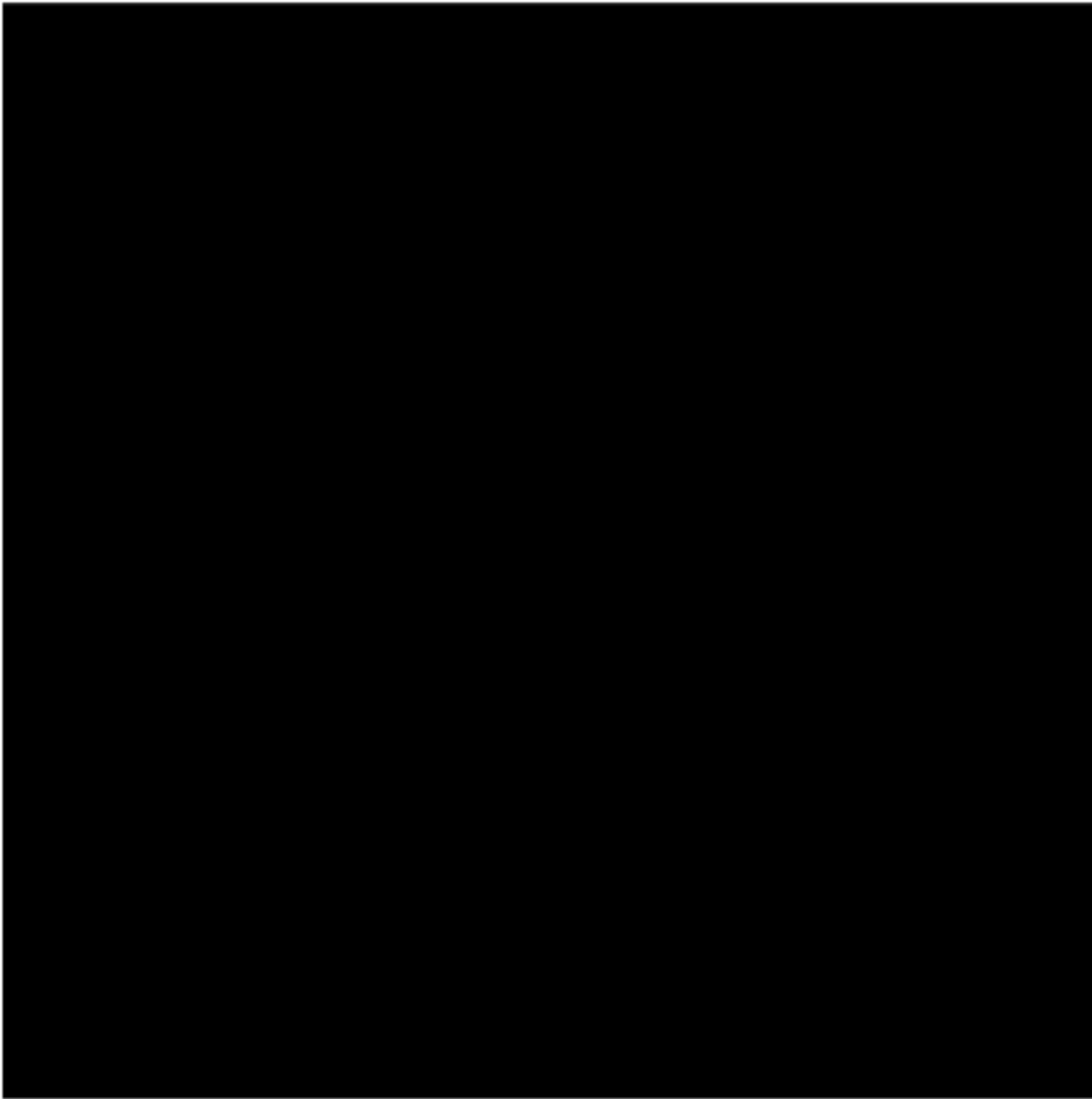
- Made with Sterling 925 silver
- Presented in a pale blue Elements gift box
- This necklace is 46 cm/18 inch in length
- Solid silver piece weighing over 40 g
- Classic T-bar necklace with delicate heart charm
- Necklace length is 46cm

amazon

# Forecasting

**Input data**

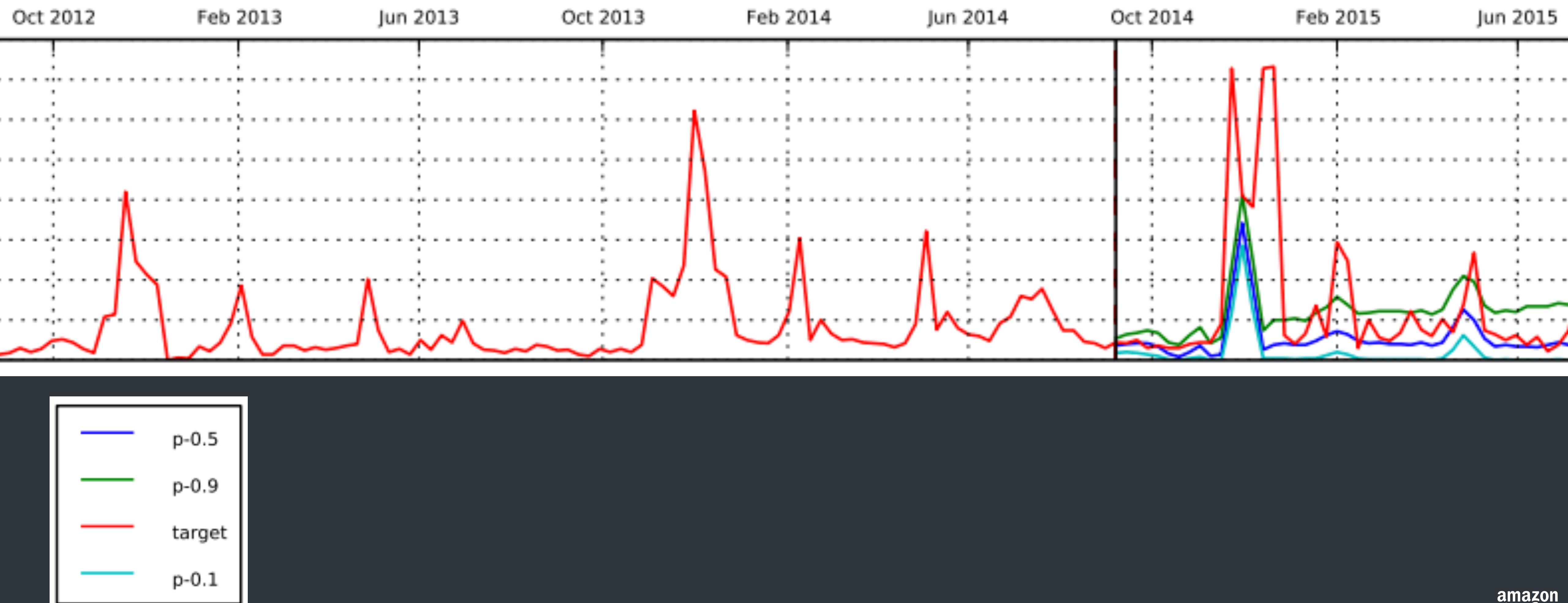
$z_1, z_2, \dots, z_{t_0-1} \longrightarrow$



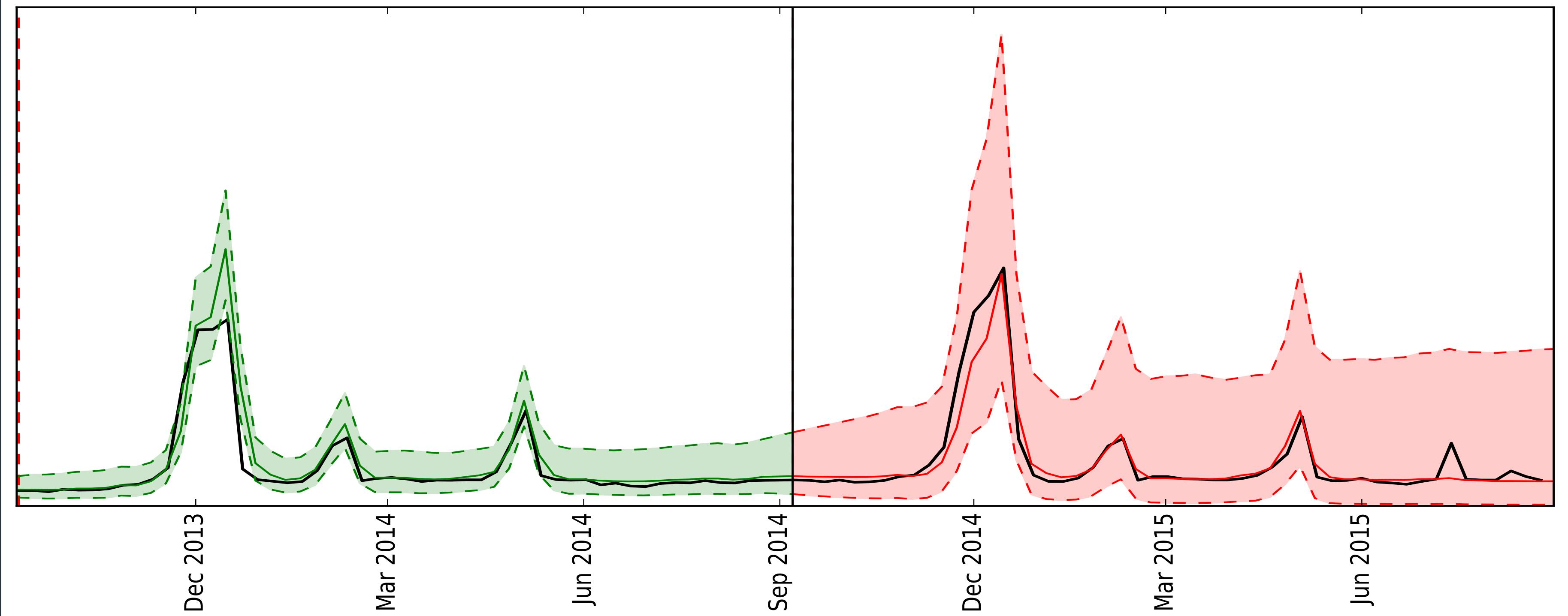
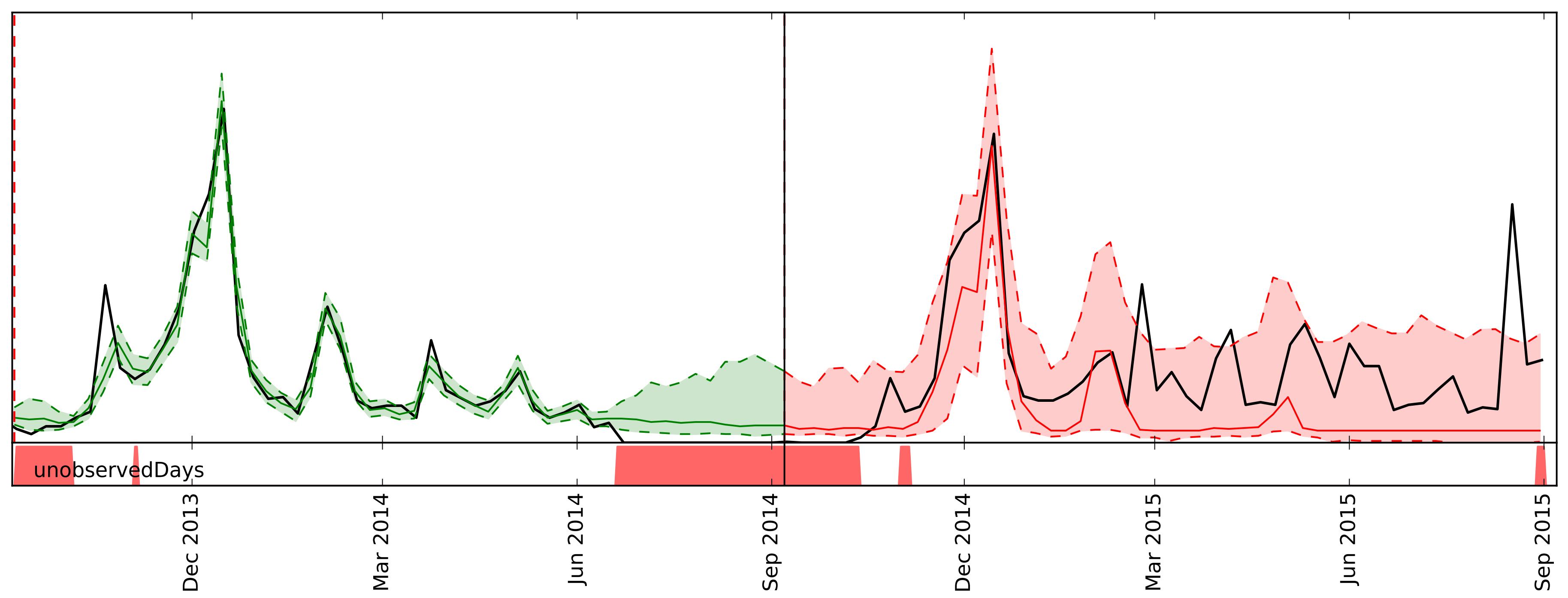
**Forecast**

$P(z_{t_0}, z_{t_0+1}, \dots, z_T)$

# Seasonality and External Events



# Probabilistic Forecasts



## Latent State Forecaster

$$z_t \sim \ell(z_t | y_t; \theta)$$

$$y_t = \mathbf{a}_t^\top \mathbf{l}_{t-1} + \mathbf{w}^\top \mathbf{x}_t$$

$$\mathbf{l}_t = \mathbf{F}\mathbf{l}_{t-1} + \mathbf{g}_t \varepsilon_t$$

$$\mathbf{l}_0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

$$\varepsilon_t \sim \mathcal{N}(0, 1)$$

$$\Theta = (\mathbf{w}, \mathbf{g}_t, \mu_0, \sigma_0^2, \theta)$$

$$\operatorname{argmax}_{\Theta} \int P(z_{1:t_0-1}, \mathbf{l}_{1:t_0-1}) d\mathbf{l}_{1:t_0-1}$$

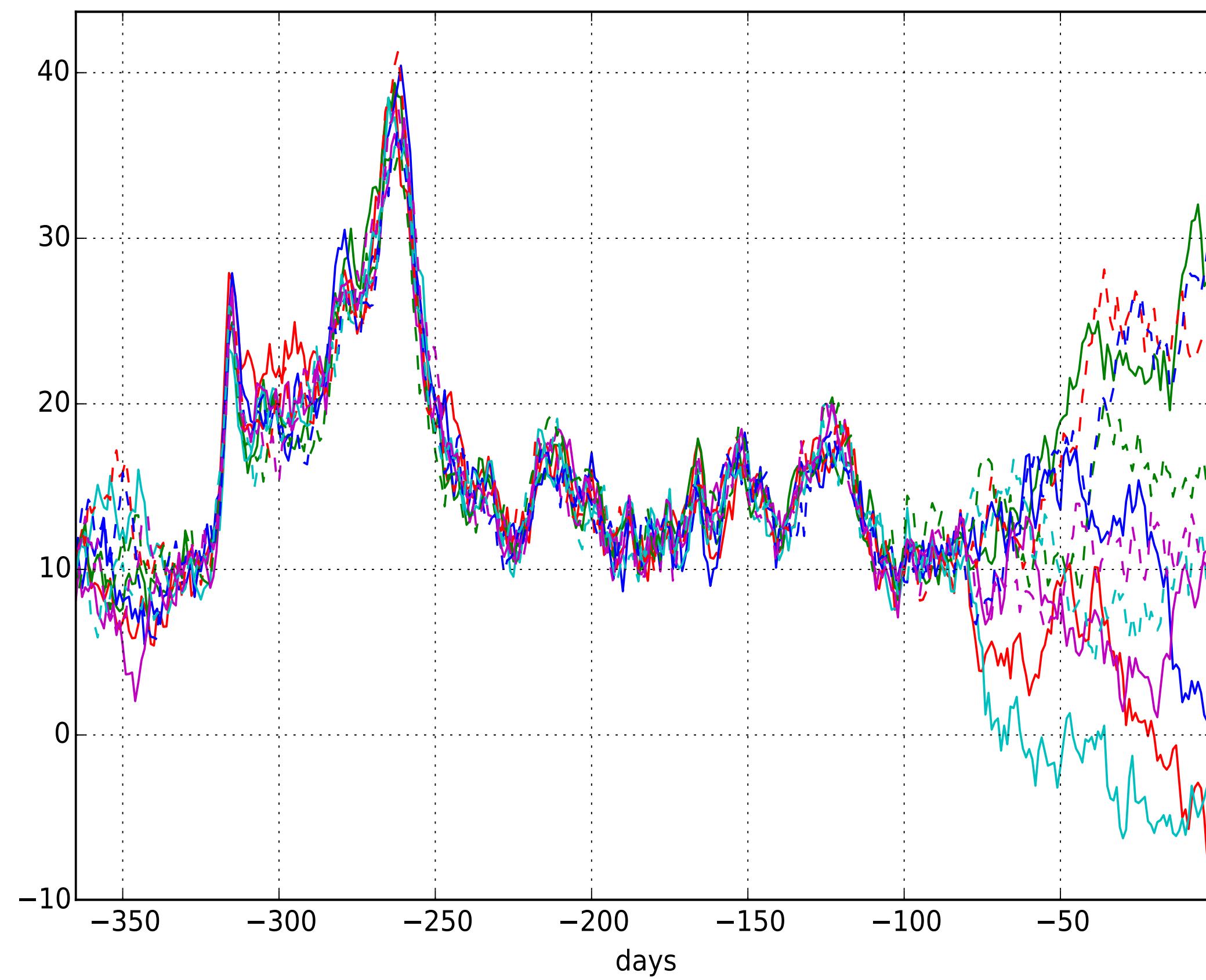
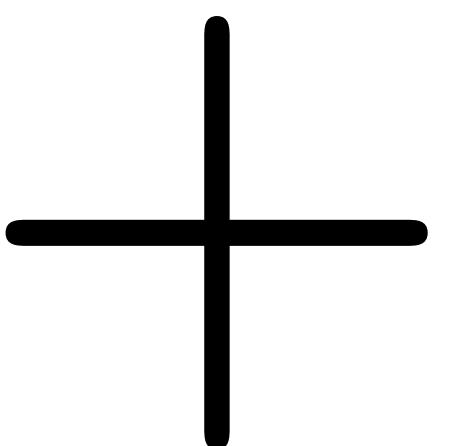
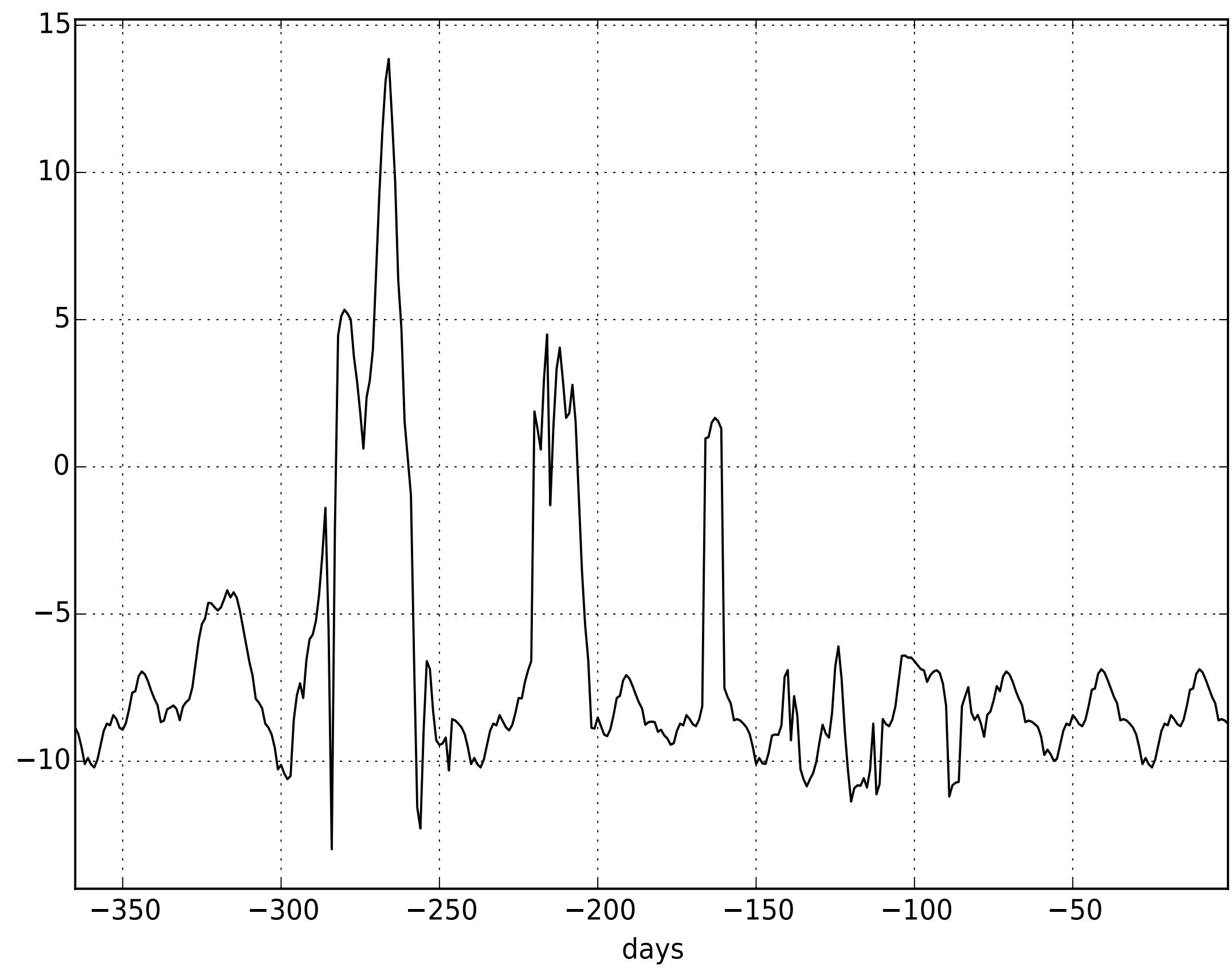
Input data

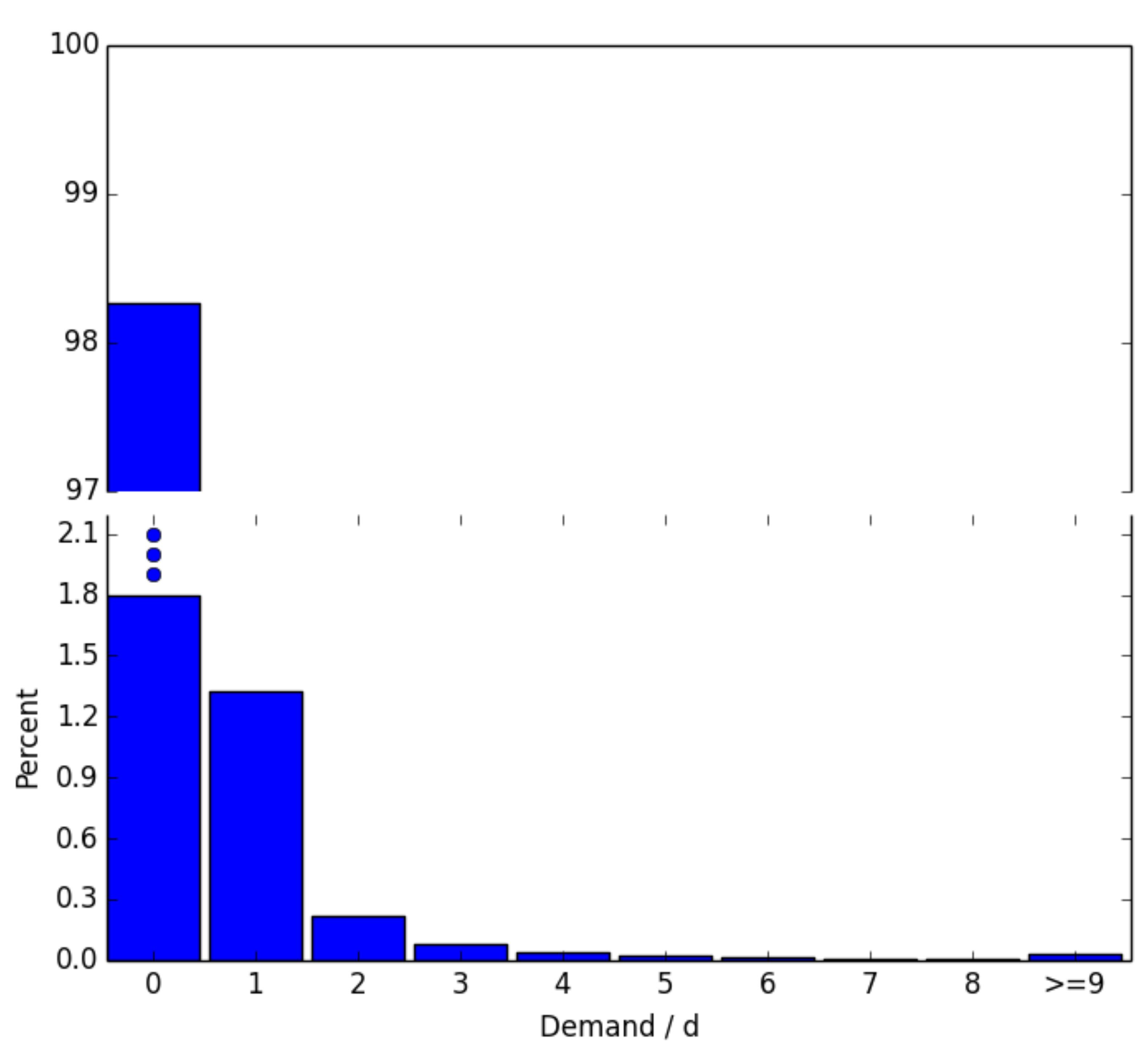
$$z_1, z_2, \dots, z_{t_0-1} \rightarrow \\ \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$$

Forecast

$$P(z_{t_0}, z_{t_0+1}, \dots, z_T)$$

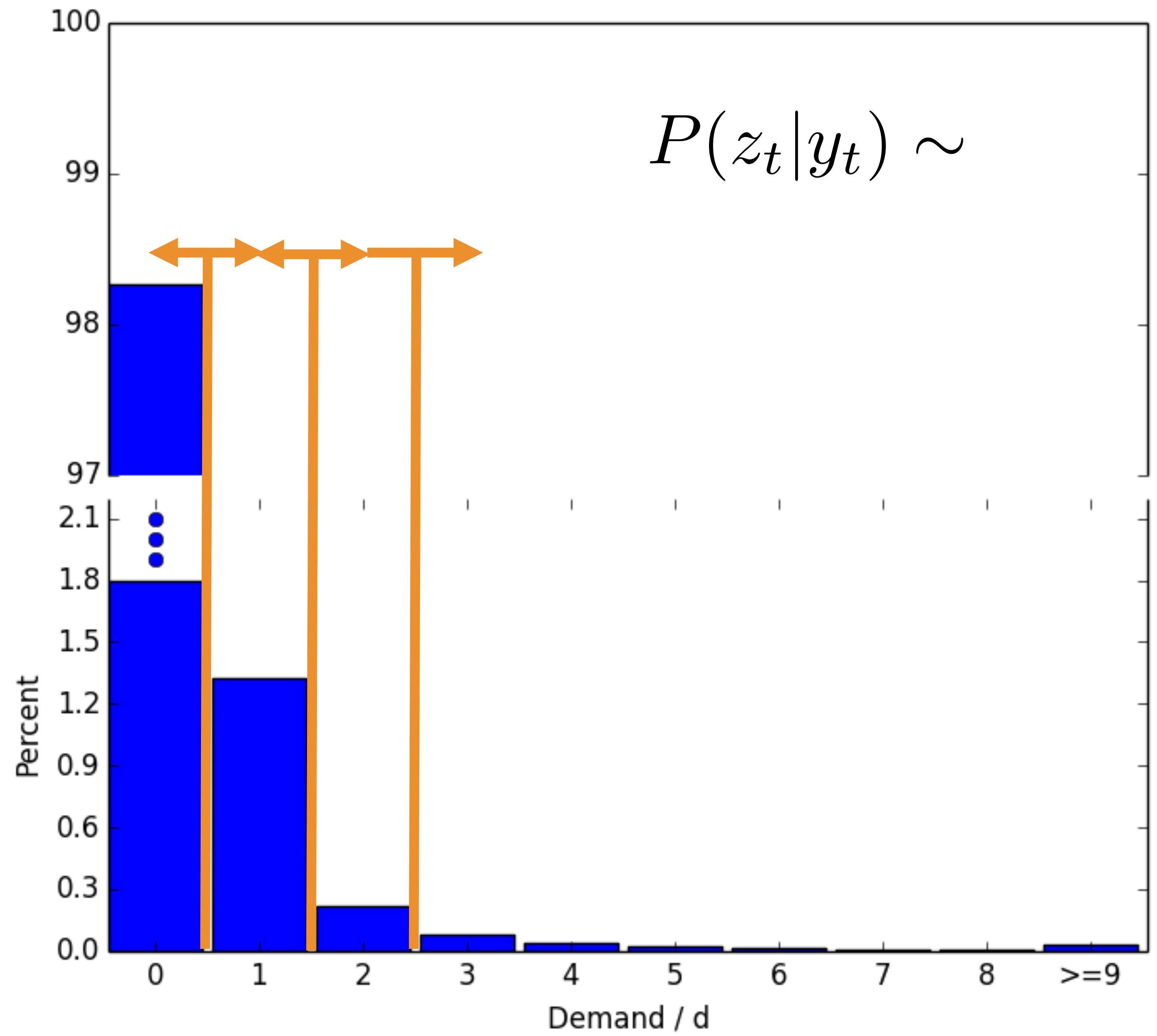
# Latent Gaussian State





Intermittent  
Demand

## Multi-stage Likelihood

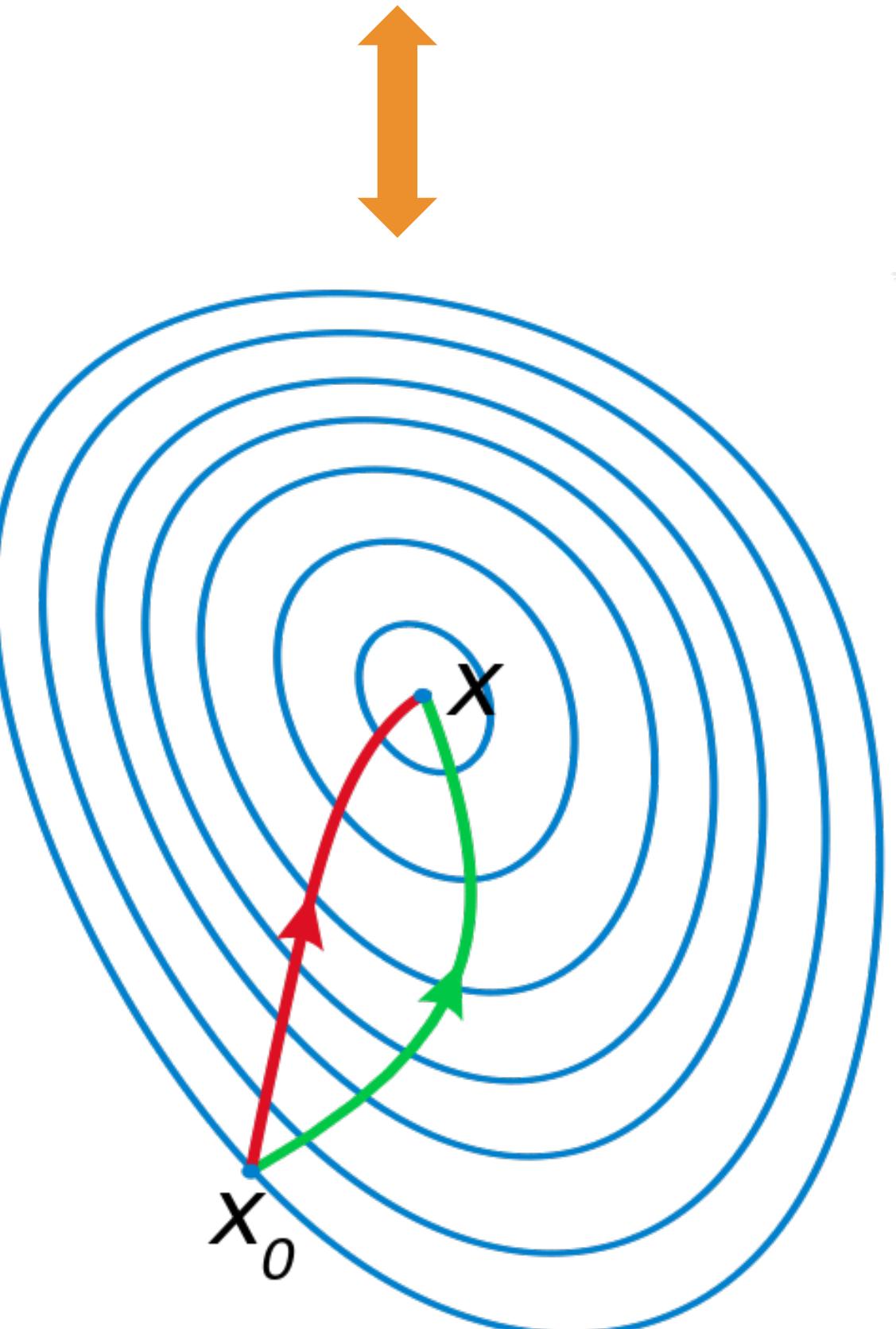


# Scalable Inference

```
while (!converged) {  
    b = kalmanSmoothing(a)  
    a = glueCode(b)  
}
```

$$\phi(\theta) = - \log \int P(\mathbf{z}_{\text{train}}, \mathbf{s} | \theta) d\mathbf{s}$$

$$[\Phi(\theta), \nabla_{\theta}\Phi]$$



|                  | Parts | EC-sub | EC-all |
|------------------|-------|--------|--------|
| # items          | 19874 | 39700  | 534884 |
| Unit $t$         | month | day    | day    |
| Median $CV^2$    | 2.4   | 5.8    | 9.7    |
| Freq. $z_t = 0$  | 54%   | 46%    | 83%    |
| In-stock ratio   | 100%  | 73%    | 71%    |
| Avg. size series | 33    | 329    | 293    |
| # item-days      | 656K  | 13M    | 157M   |

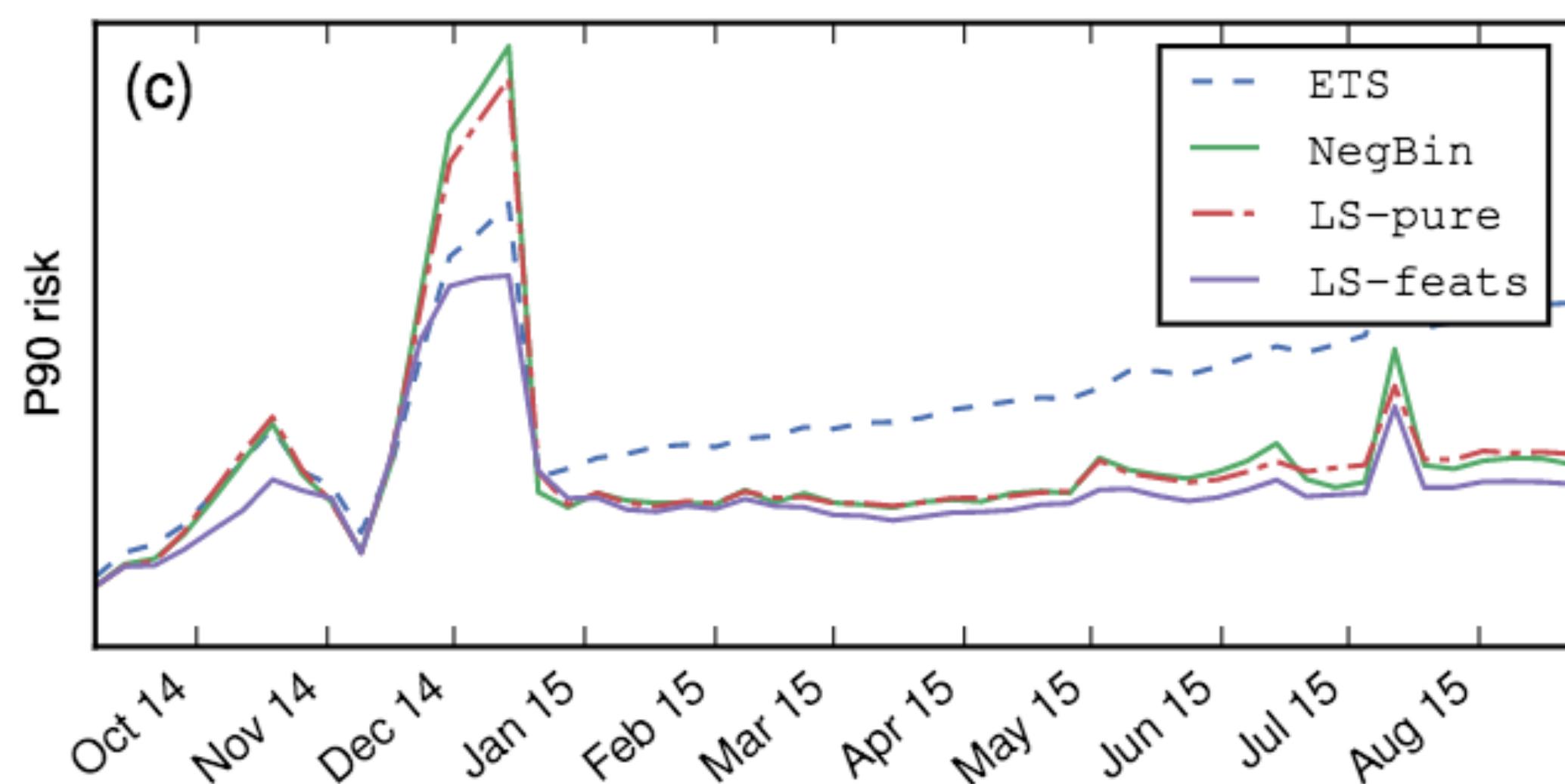
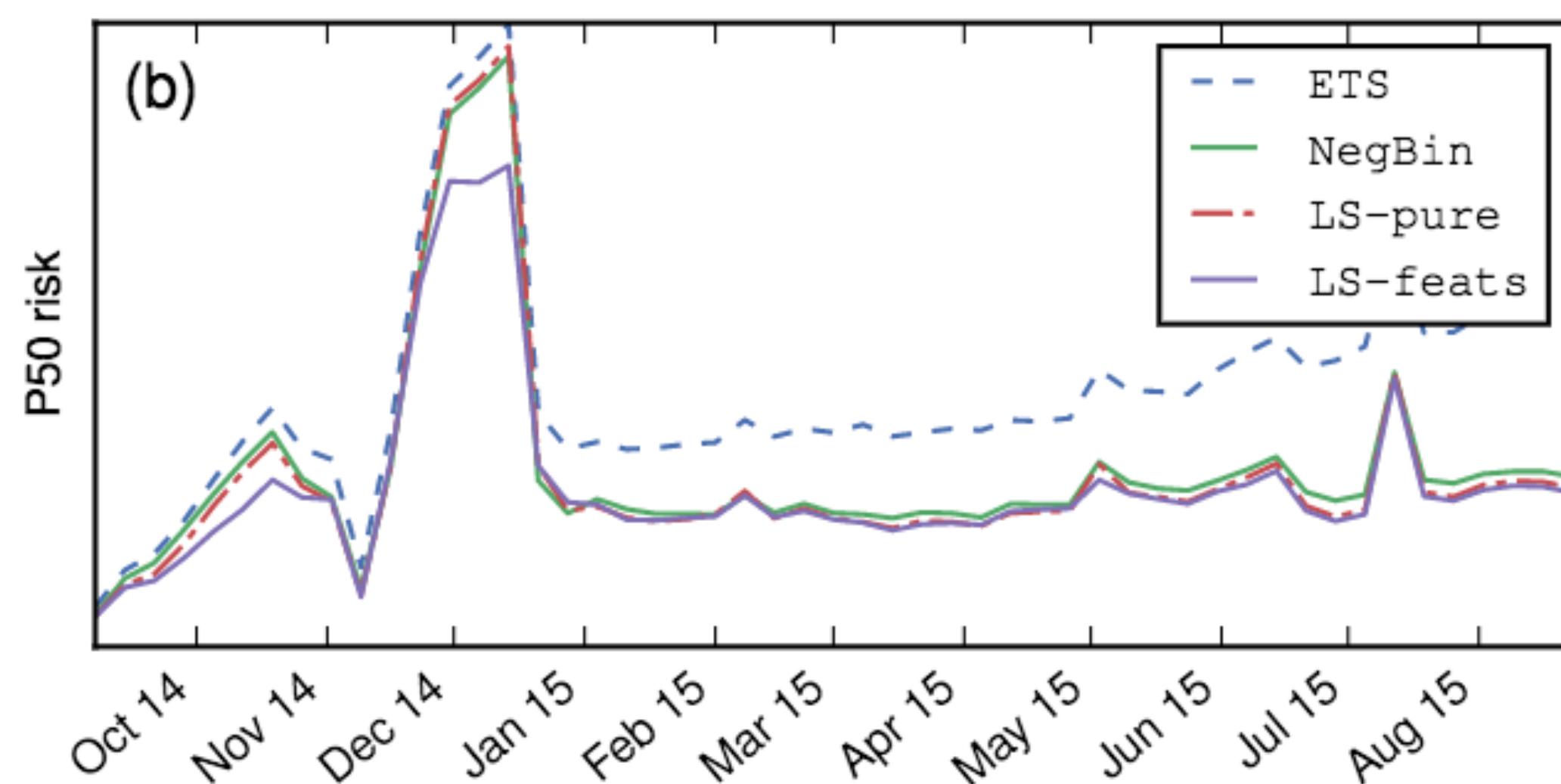
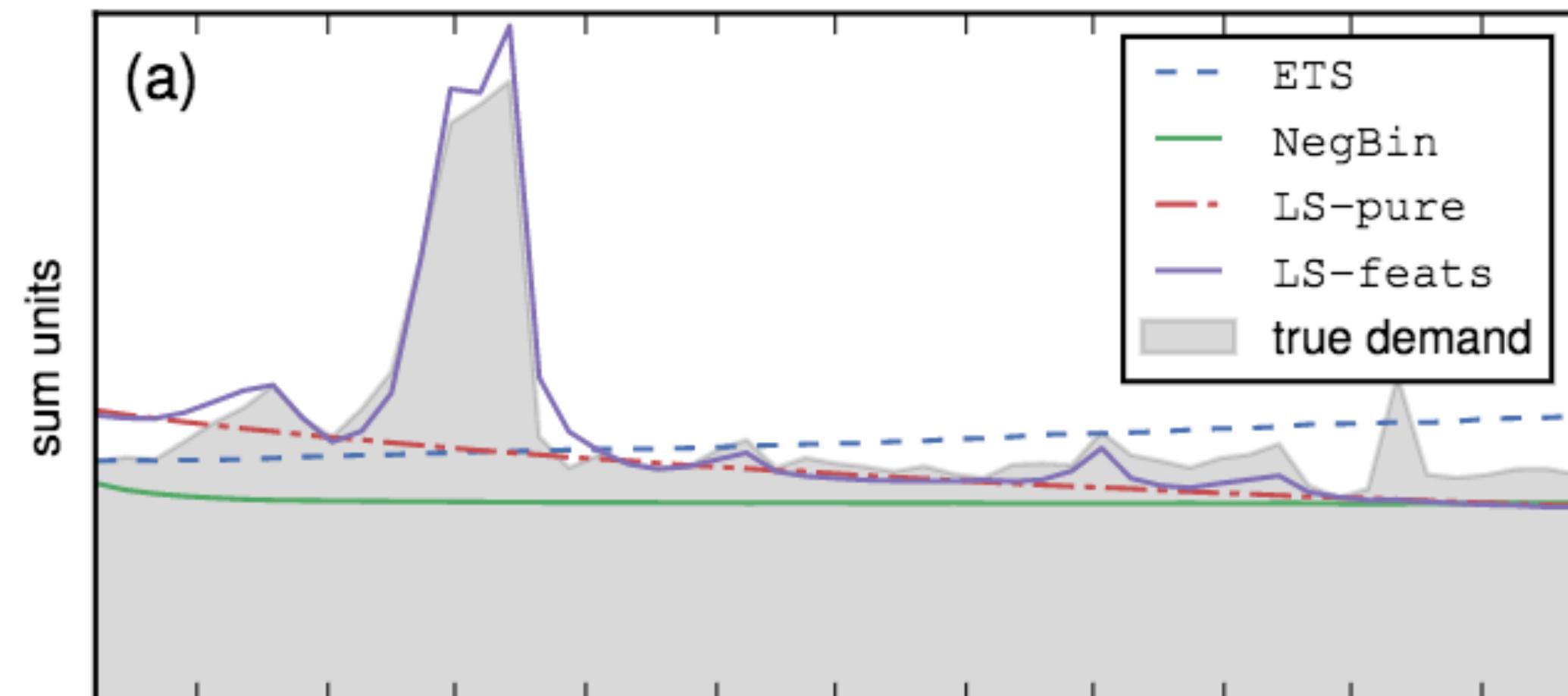


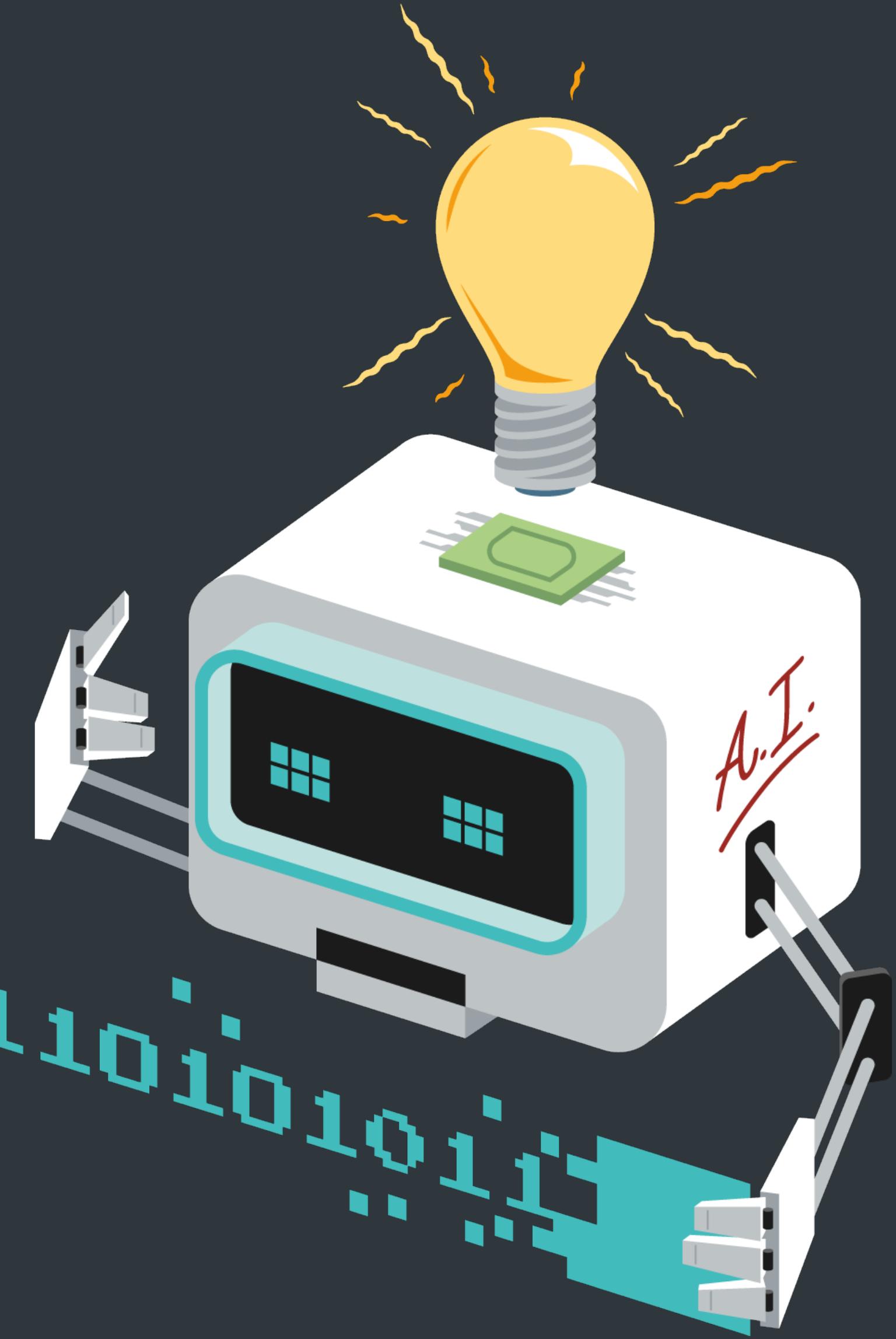
Figure 3: **Table:** Dataset properties.  $CV^2 = \text{Var}[z_t]/\text{E}[z_t]^2$  measures burstiness. **(a):** Sum of weekly P50 point (median) forecast over a one-year prediction range for the different methods (lines) as well as sum of true demand (shaded area), on dataset  $\mathcal{I} = \text{EC-sub}$ . **(b):** Weekly P50 risk  $R^{0.5}[\mathcal{I}; (7 \cdot k, 7)]$ ,  $k = 0, 1, \dots$ , for same dataset. **(c):** Same as (b) for P90 risk.

# Take away

Machine learning & AI are transformational.

Wealth of probabilistic machine learning applications.

When deploying machine learning, keep it simple (not simpler).





Beyza  
Ermis



Felix  
Biessmann



Michael  
Brückner



Matthias  
Seeger



David  
Salinas



Valentin  
Flunkert

# amazon

[cedrica@amazon.com](mailto:cedrica@amazon.com)



Pablo García  
Moreno