Production and Operations Management Society
Applied Research Challenge
May 6th, 2016

From Predictive to Prescriptive Analytics

by D. Bertsimas (MIT) & N. Kallus (Cornell)

Presenters:

Nathan Kallus — Asst. Professor, Cornell
Amjad Hussain — CEO, Silkroute

Applied ML in Data Science

Data	Prediction Problem	Prescription Problem
Web Search	Predict video game demand (Goel et al. ’10)	Inventory management for video game titles
Twitter	Predict box-office gross (Asur & Huberman ‘10)	Assign capacities (cinemas)
Blogs	Predict amazon book sales (Gruhl et al. ‘05)	Facility location, shipment planning
Twitter & News	Predict civil unrest (Kallus ’14)	Supply chain management

… … …

A general problem

- Data $y^1, \ldots, y^N$ on quantity(ies) of interest Y
  E.g. demands at locations/of products, % returns
- Data $x^1, \ldots, x^N$ on associated covariates X
  E.g. recent sales figures, search engine attention
- Decision $z \in \mathcal{Z}$ to minimize *uncertain* costs $c(z; Y)$
  after observing $X = x$

The predictive prescription problem

- Problem of interest:

$$z^*(x) \in \arg\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z;Y) \middle| X = x\right]$$

- Hypothetical full-information optimum
  - Uses knowledge of $\mu_{X,Y}$ to leverage $X = x$ to greatest possible extent in reducing costs

- **Our task**:
  use data $S_N = \{(x^1, y^1), \ldots, (x^N, y^N)\}$ to construct a data-driven predictive prescription

$$\hat{z}_N(x) : \mathcal{X} \to \mathcal{Z}$$

Standard Data-Driven Optimization

- Data $y^1, \ldots, y^N$ on quantity(ies) of interest Y
- Decision $z \in \mathcal{Z}$ to minimize *uncertain* costs $c(z; Y)$
- Problem of interest is $\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z; Y)\right]$
- Standard data-driven solution is sample average approximation (SAA)

$$\hat{z}_N^{\text{SAA}} \in \arg\min_{z \in \mathcal{Z}} \frac{1}{N} \sum_{i=1}^{N} c(z; y^i)$$

  – Also: SA (Robins '51), Robust SAA (Bertsimas, Gupta, Kallus '14), Data-Driven RO (Bertsimas, Gupta, Kallus '13), Data-Driven DRO (Delage & Ye '10, Calafiore & El Gahoui '06)

- In our problem, standard data-driven optimization **accounts for uncertainty but not for auxiliary data**

Standard Supervised Learning in ML

- Data $y^1, \dots, y^N$ on quantity(ies) of interest Y
- Data $x^1, \dots, x^N$ on associated covariates X
- Problem of interest is prediction, i.e., $\mathbb{E}\left[Y \middle| X = x\right]$
- Standard approaches: linear regression, random forest
- Standard use in decision making (as taught in 15.060):
  - Fit a predictive model $\hat{m}_N(x) \approx \mathbb{E}\left[Y \middle| X = x\right]$ to data (e.g. a random forest) and optimize deterministically

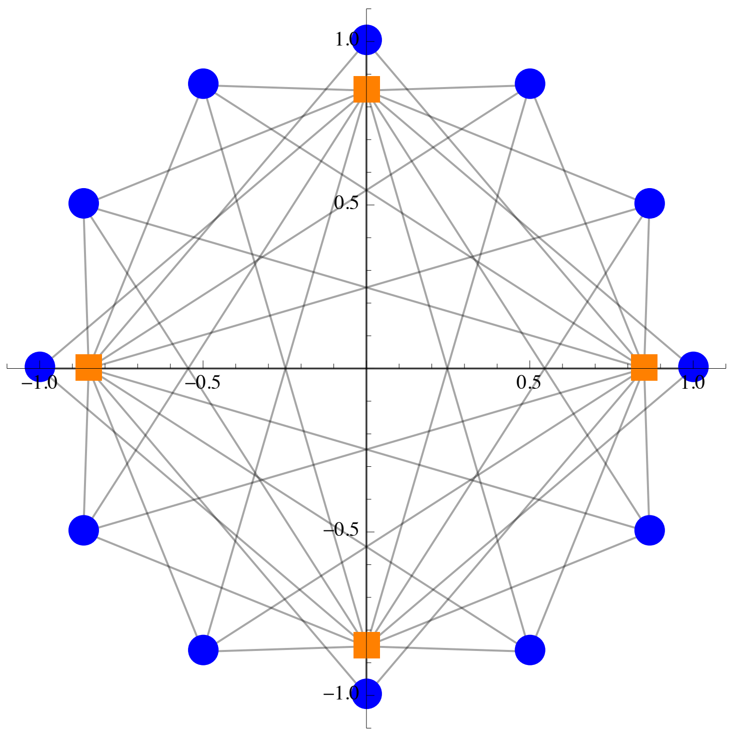
$$\hat{z}_N^{\text{point-pred}}(x) \in \arg\min_{z \in \mathcal{Z}} c(z; \hat{m}_N(x))$$

- In our problem, ML point-prediction-driven decisions **account for auxiliary data but not for uncertainty**

Shipment planning example

- Stock 4 warehouses to fulfill demand in 12 locations
- Observe predictive features X about demand in a week

$$c(z;y) = p_1 \sum_{i=1}^{d_z} z_i \quad + \quad \min \quad \left(p_2 \sum_{i=1}^{d_z} t_i + \sum_{i=1}^{d_z} \sum_{j=1}^{d_y} c_{ij} s_{ij} \right)$$



$$\text{s.t. } t_i \geq 0 \qquad \forall i$$

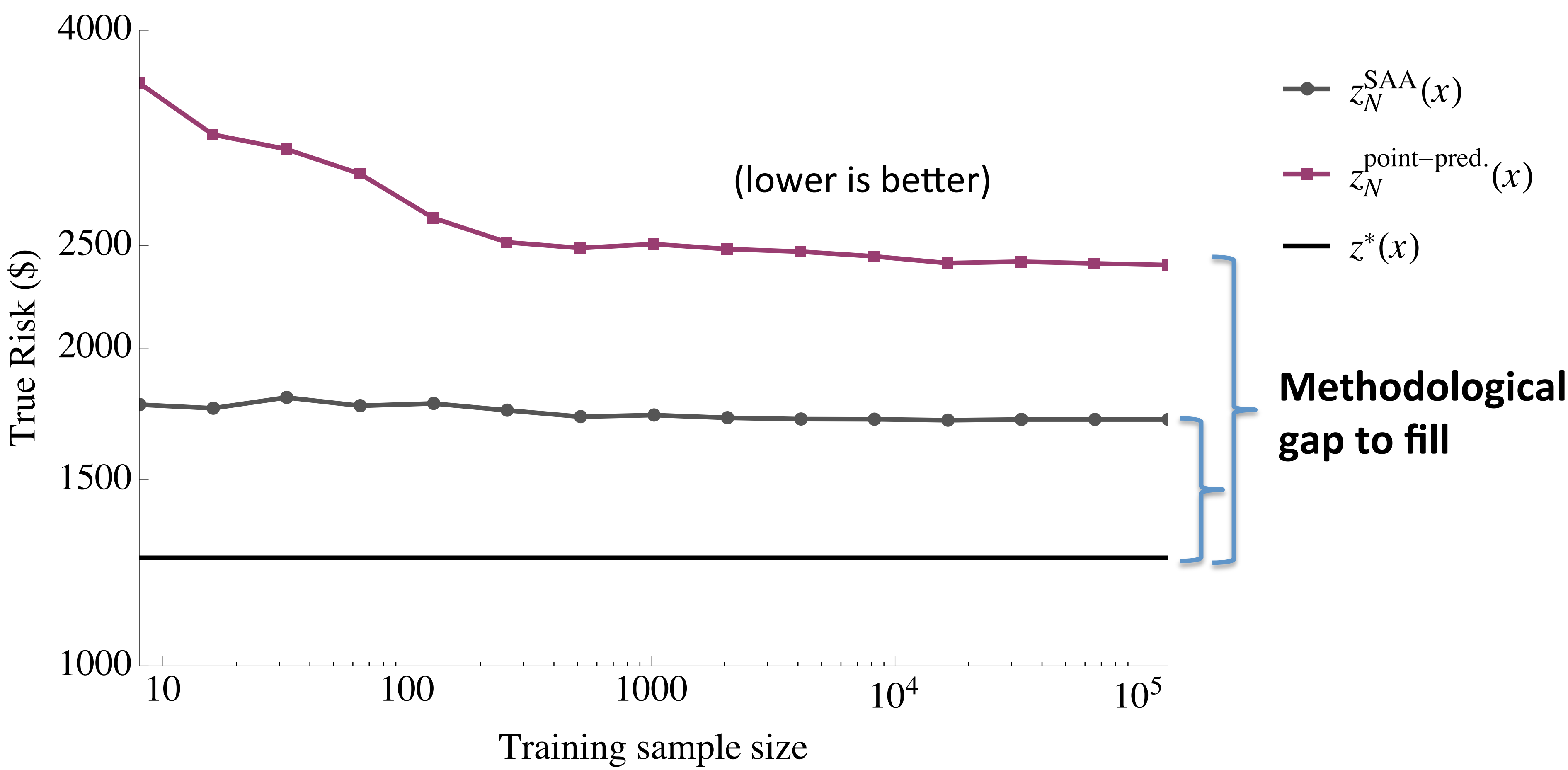
$$s_{ij} \geq 0 \qquad \forall i, j$$

$$\sum_{i=1}^{d_z} s_{ij} \geq y_j \qquad \forall j$$

$$\sum_{j=1}^{d_y} s_{ij} \leq z_i + t_i \qquad \forall i$$

Shipment planning example

- Stock 4 warehouses to fulfill demand in 12 locations
- Observe predictive features X about demand in a week



Contributions

- **A new framework**
  - General purpose
  - Coefficient of prescriptiveness
- **Theory**
  - Computational tractability
  - Asymptotic optimality
- **Practice**
  - Case study of huge media distributor
  - In collaboration with Silkroute
  - Study ***prescriptive*** power of large-scale data

Our approach

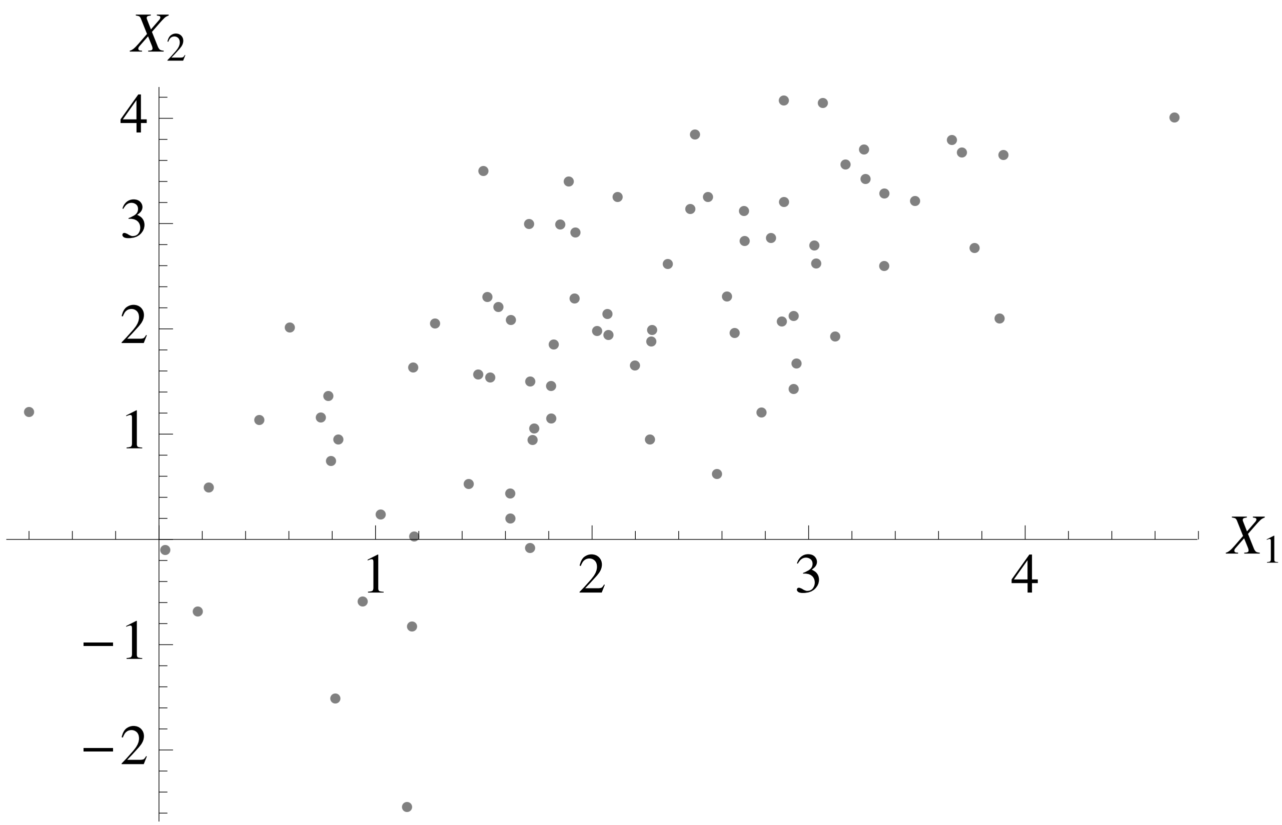
- A local learning approach to prescription
- Re-weight Y data using data-driven weights
  - Emphasize data that is similar to new observation (Analogy breaks down in general)
- Construct predictive prescriptions of the form

$$\hat{z}_N(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^{N} w_N^i(x) c(z; y^i)$$

- Draws on ideas from non-parametric predictive statistics (Stone '77) and extends to *optimization*

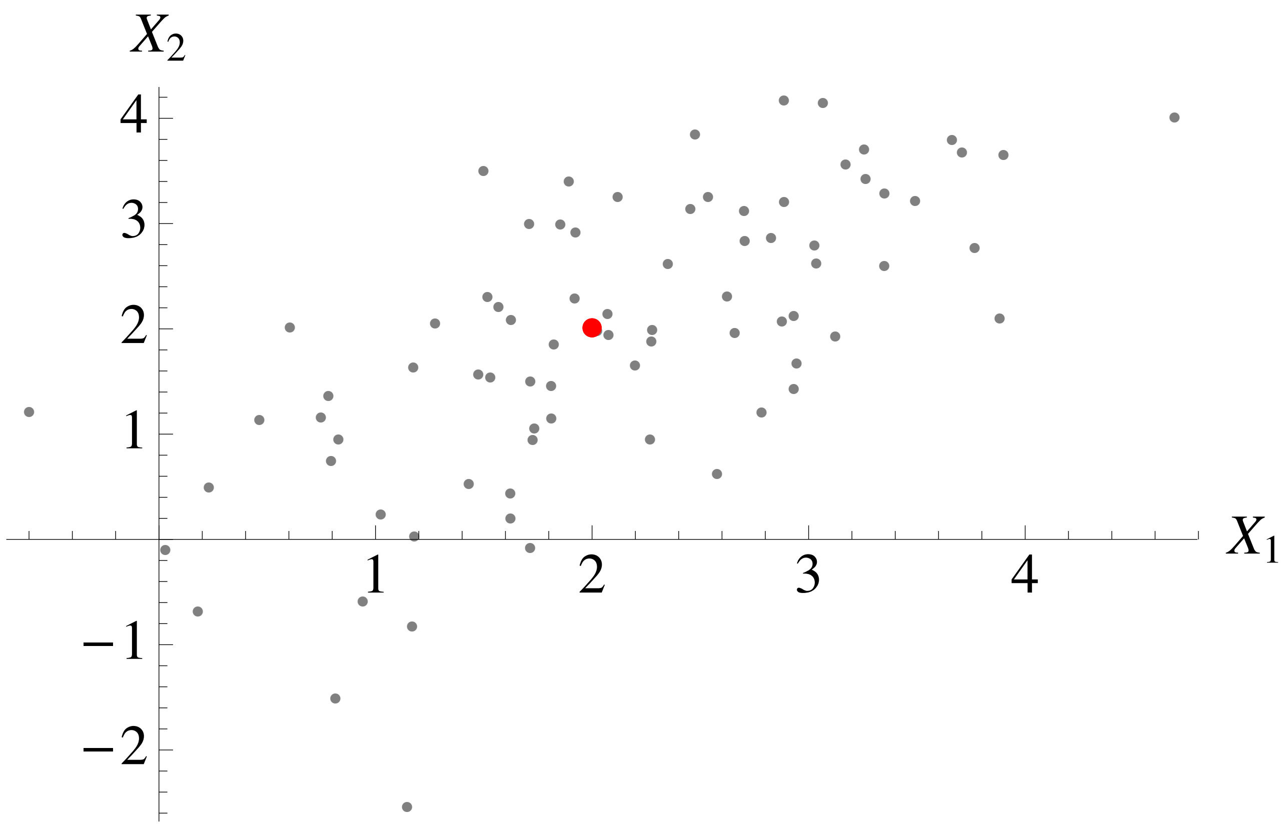
Weights using nearest neighbors

$$\hat{z}_N^{k\mathrm{NN}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{x^i \text{ is } k\mathrm{NN} \text{ of } x} c(z; y^i)$$



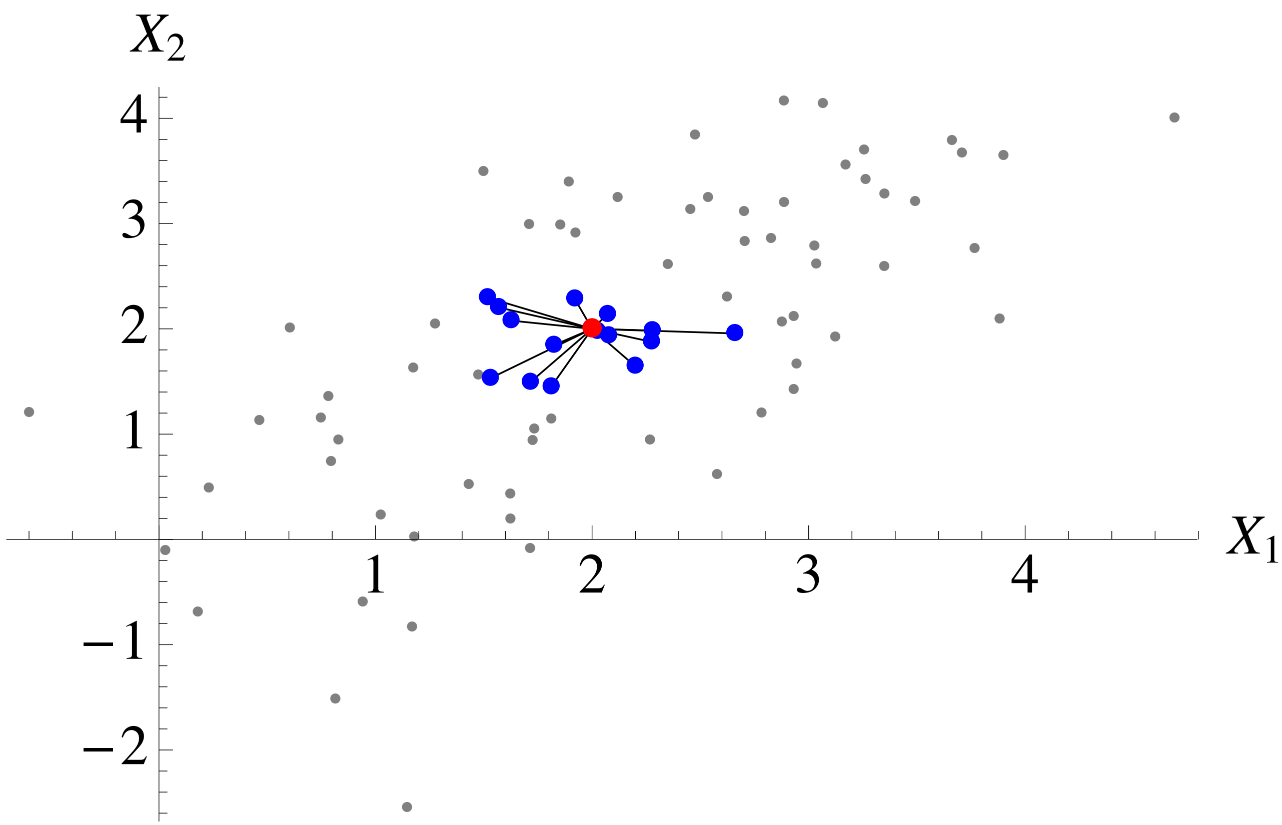
Weights using nearest neighbors

$$\hat{z}_N^{k\mathrm{NN}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{x^i \text{ is } k\mathrm{NN} \text{ of } x} c(z; y^i)$$



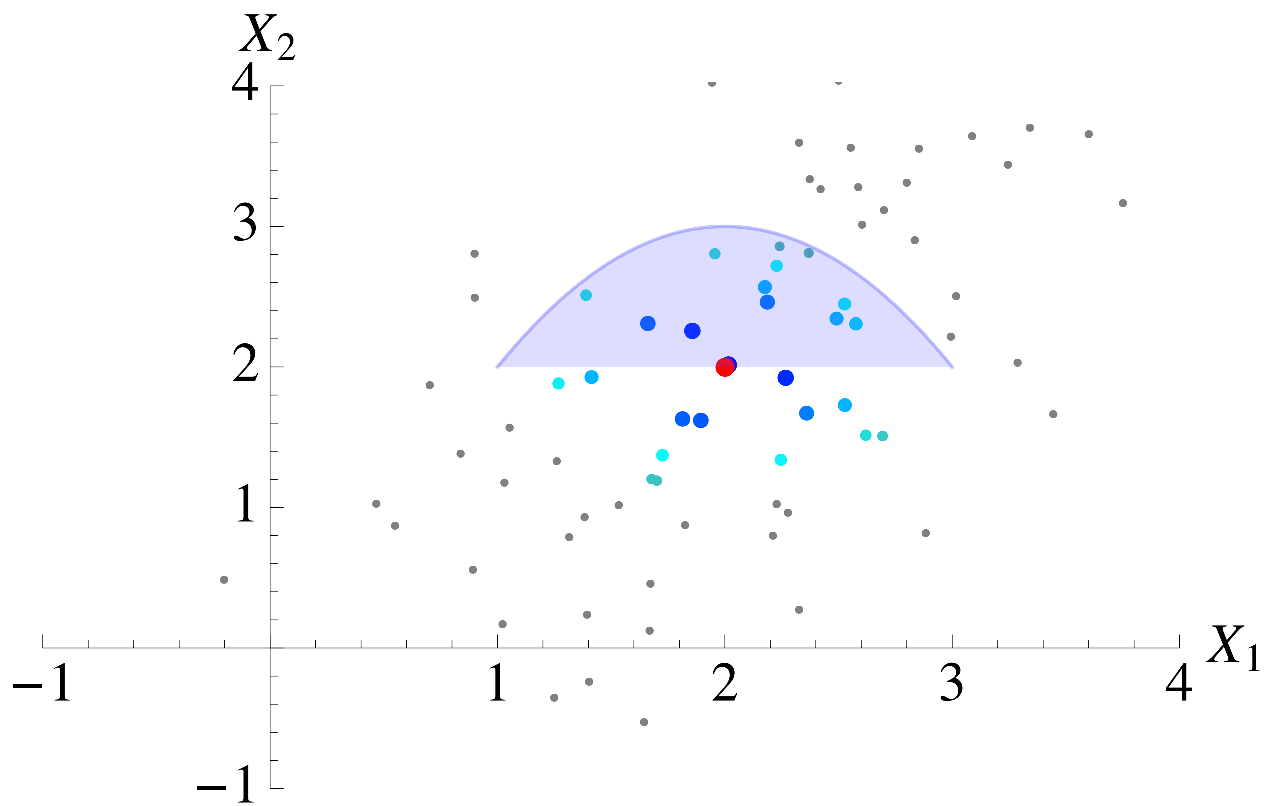
Weights using nearest neighbors

$$\hat{z}_N^{k\text{NN}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{x^i \text{ is } k\text{NN of } x} c(z; y^i)$$



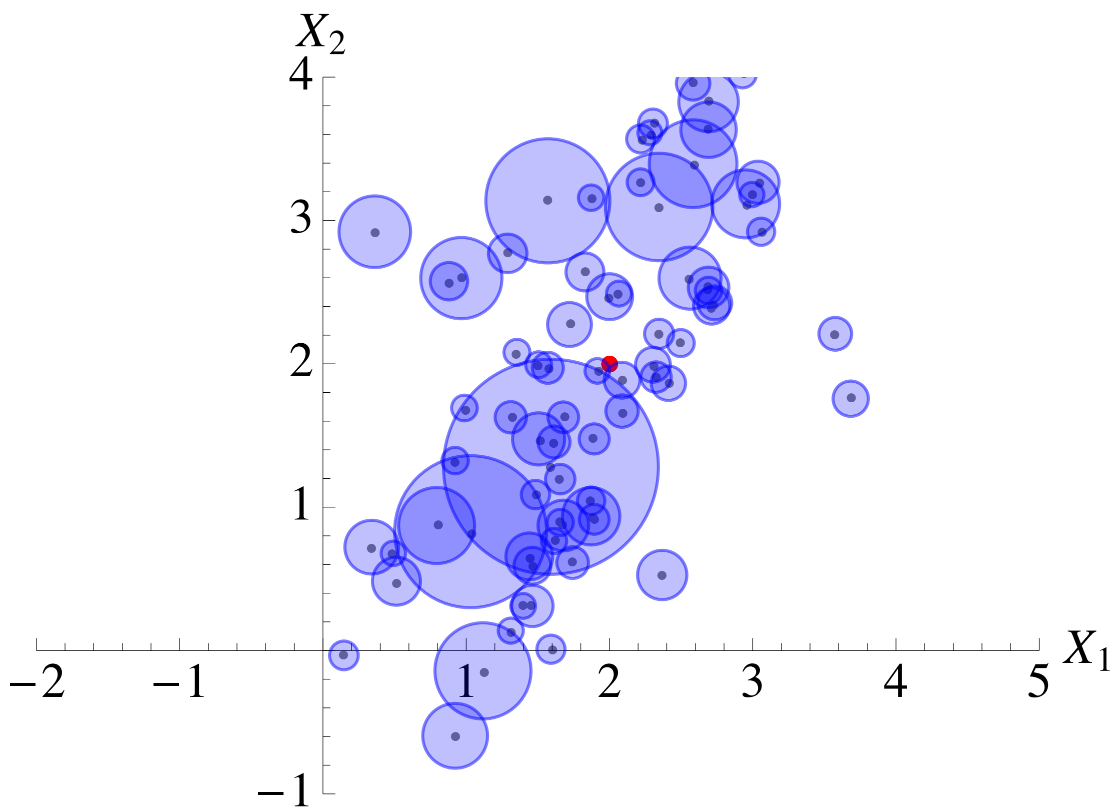
Weights using Parzen windows

$$\hat{z}_N^{\mathrm{KR}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^{N} K((x^i - x)/h_N)c(z; y^i)$$



Weights using recursive Parzen

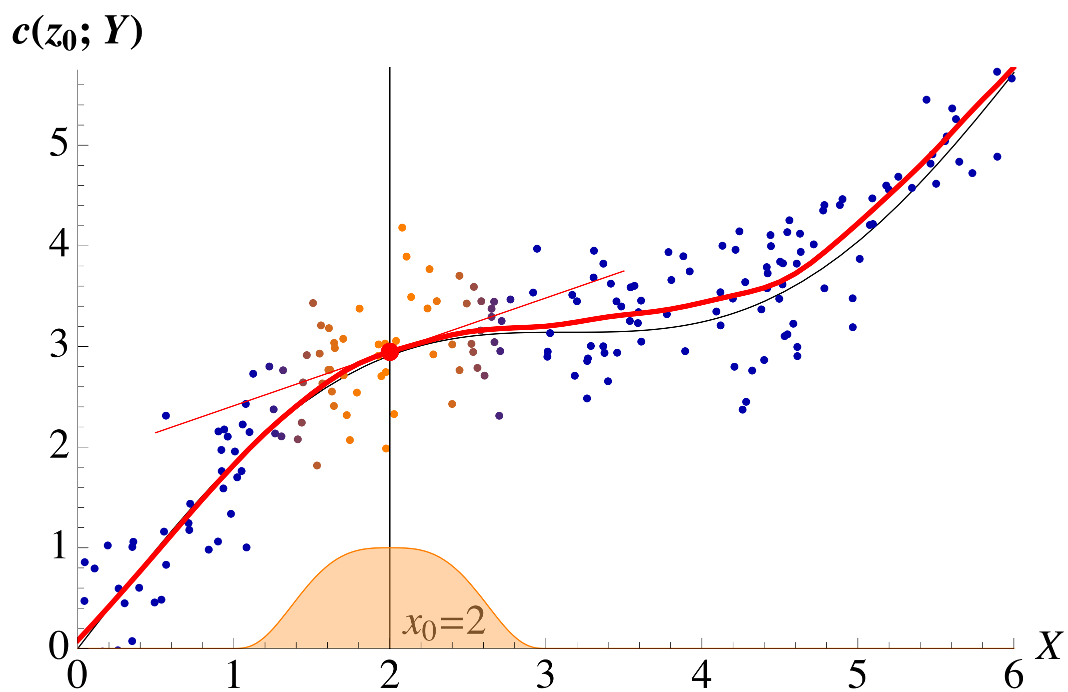
$$\hat{z}_N^{\text{Rec-KR}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^{N} K((x^i - x)/h_{\boldsymbol{i}})c(z; y^i)$$



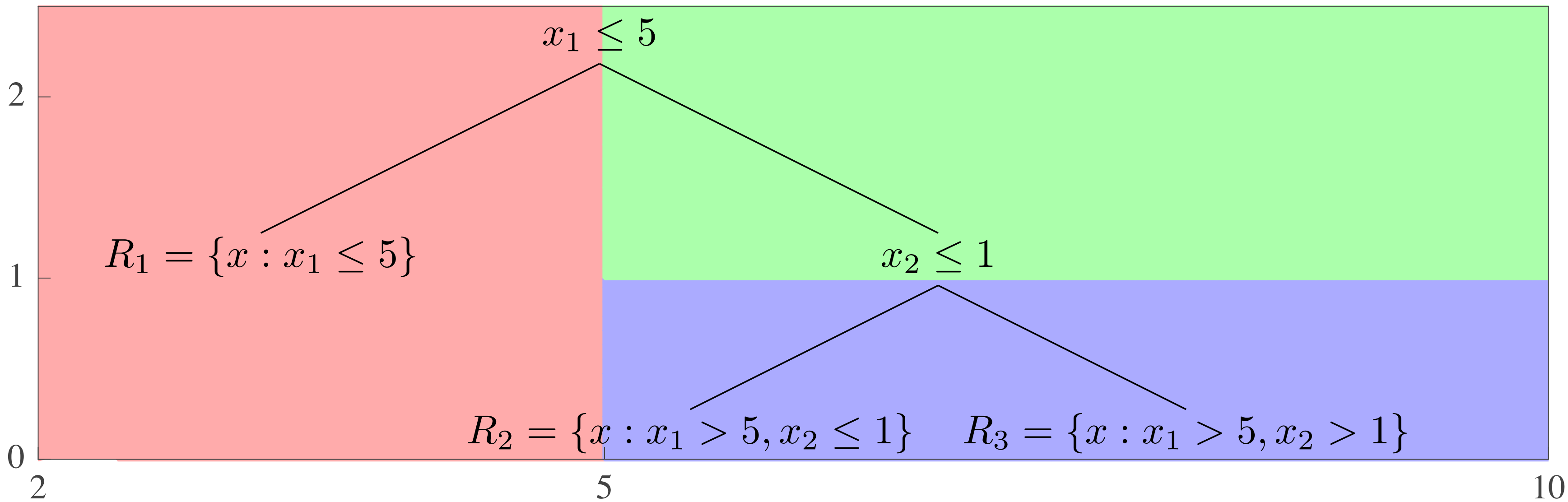
Weights using LOESS

$$\hat{z}_N^{\mathrm{LOESS}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^{N} k_i(x) \left(1 - \sum_{j=1}^{n} k_j(x)(x^j - x)^T \Xi(x)^{-1}(x^i - x) \right) c(z; y^i)$$

$$\Xi(x) = \sum_{i=1}^{n} k_i(x)(x^i - x)(x^i - x)^T \quad k_i(x) = \left(1 - \left(\left|\left|x^i - x\right|\right| / h_N\right)^3\right)^3 \mathbb{I}\left[\left|\left|x^i - x\right|\right| \leq h_N\right]$$



Weights using recursive partitions



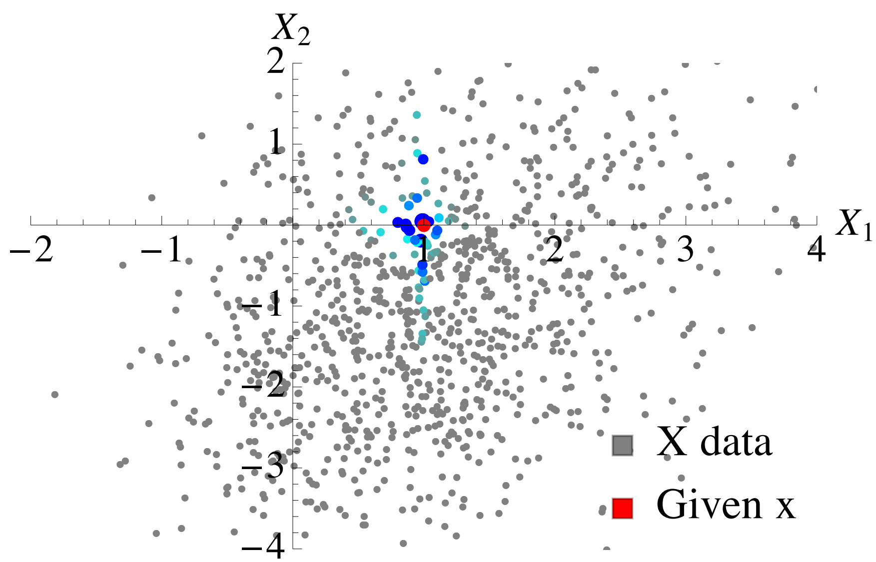
Implied binning rule $R(x) = (j \text{ s.t. } x \in R_j)$

$$\hat{z}_N^{\text{CART}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{\mathcal{R}(x^i) = \mathcal{R}(x)} c(z; y^i)$$

Weights using bagging

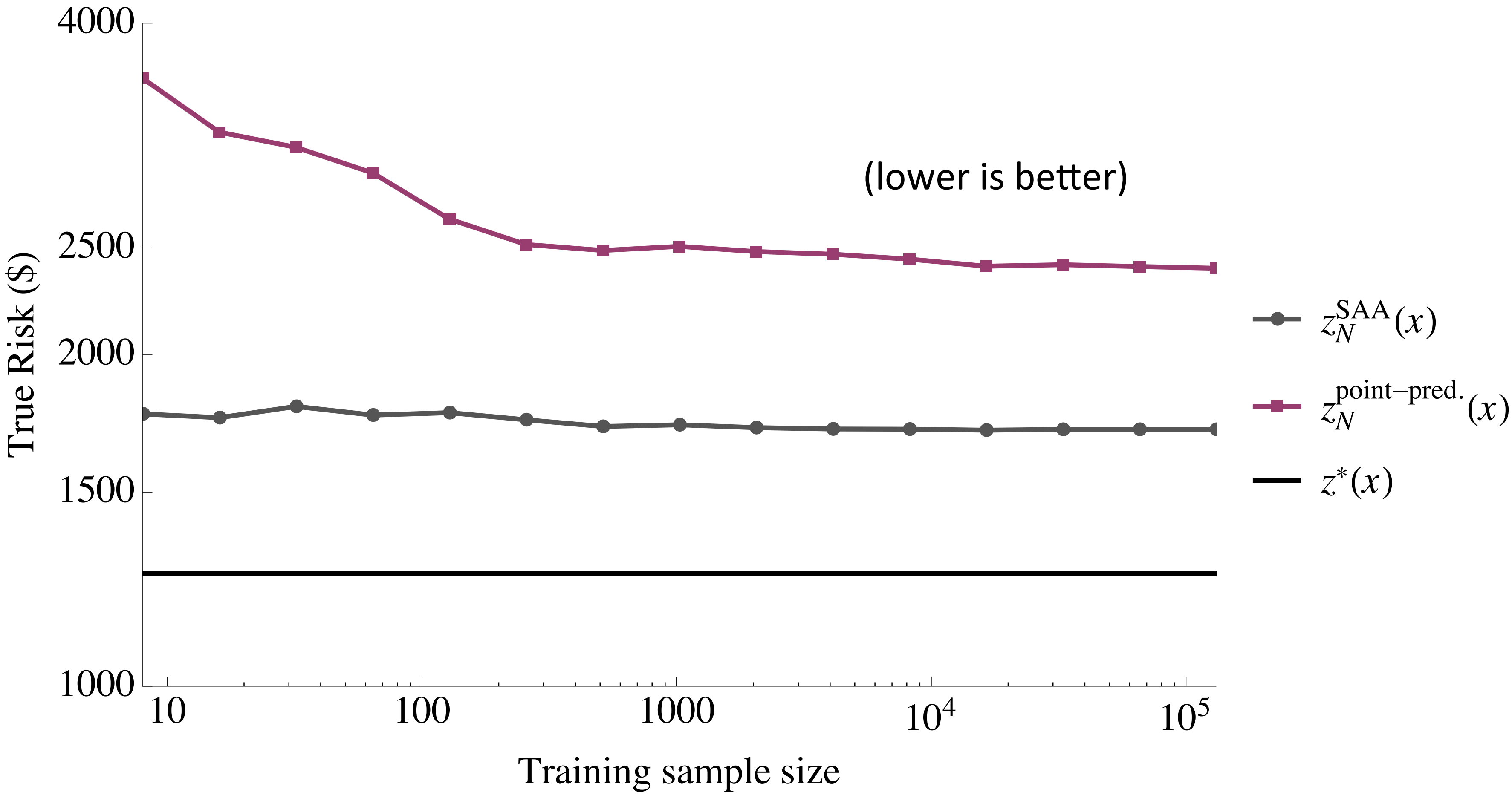
- Train T tree partitions on bootstrapped samples and random feature subsets
- Get T binning rules $R^t(x) = \left(j \text{ s.t. } x \in R_j^t\right)$

$$\hat{z}_N^{\mathrm{RF}}(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{t=1}^{T} \frac{1}{|\{j : R^t(x^j) = R^t(x)\}|} \sum_{R^t(x^i) = R^t(x)} c(z; y^i)$$



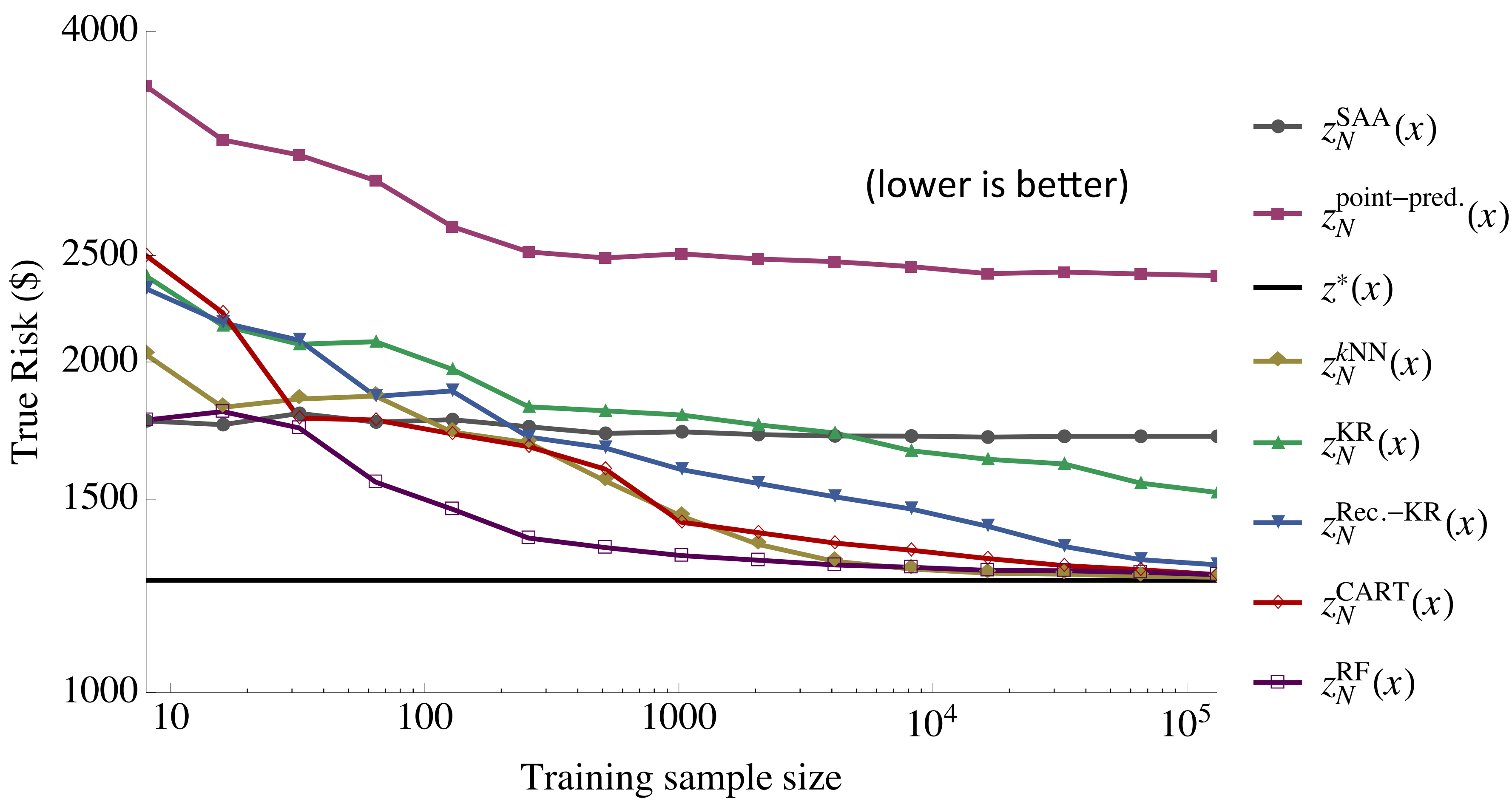
Shipment planning example

- Stock 4 warehouses to fulfill demand in 12 locations
- Observe predictive features X about demand in a week



Shipment planning example

- Stock 4 warehouses to fulfill demand in 12 locations
- Observe predictive features X about demand in a week



Coefficient of Prescriptiveness

Data-poor prescription

Our prescription

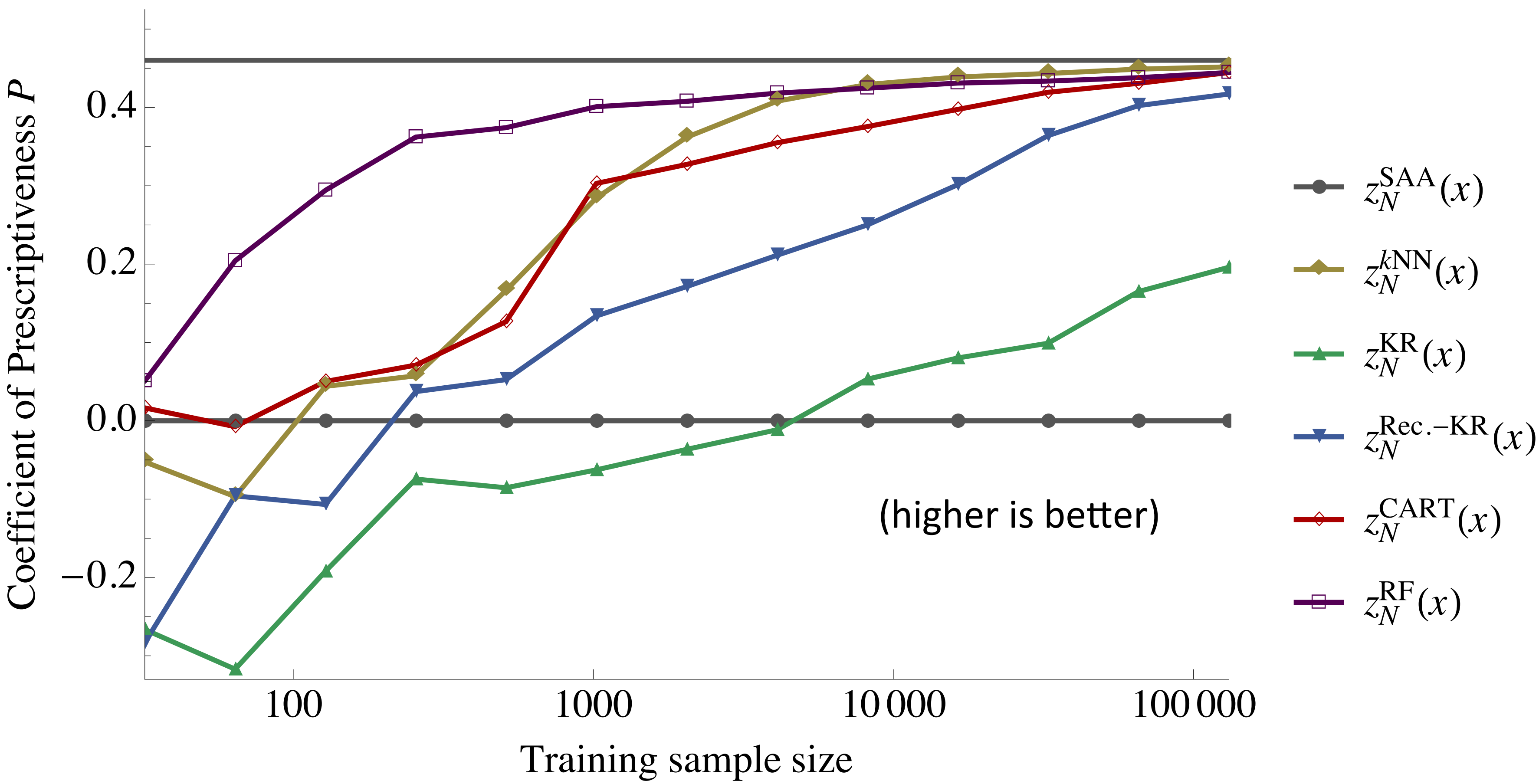
$$P = \frac{\min_{z \in \mathcal{Z}} \sum_{i=1}^{N} c(z; y^i) - \sum_{i=1}^{N} c(\hat{z}_N(x^i); y^i)}{\min_{z \in \mathcal{Z}} \sum_{i=1}^{N} c(z; y^i) - \sum_{i=1}^{N} \min_{z \in \mathcal{Z}} c(z; y^i)} \leq 1 \\ \to [0, 1]$$

Perfect foresight (deterministic)

- Measures the prescriptive value of X and of the of the prescription trained
- To be measured out of sample

Shipment planning example

- X can get us **43%** of the way from no data to perfect foresight
  – less if prescription is not well trained / insuff. data



Asymptotic Optimality

- Want

Def: predictive prescription $\hat{z}_N(x)$ is ***asymptotically optimal*** if, with probability 1, for almost everywhere x, as $N \to \infty$

$$\lim_{N\to\infty} \mathbb{E}\left[c(\hat{z}_N(x);Y)\middle|X=x\right] = \min_{z\in\mathcal{Z}} \mathbb{E}\left[c(z;Y)\middle|X=x\right]$$

$$L\left(\{\hat{z}_N(x) : N \in \mathbb{N}\}\right) \subset \arg\min_{z\in\mathcal{Z}} \mathbb{E}\left[c(z;Y)\middle|X=x\right]$$

- Need

Assumption 1: The full-info problem is well defined, i.e.,

$$\mathbb{E}\left[|c(z;Y)|\right] < \infty$$

Assumption 2: $c(z;y)$ is equicontinuous in z .

Assumption 3: $\mathcal{Z}$ is closed and either (a) also bounded, (b) $c(z;y)$ is coercive, or (c) $c(z;y)$ is convex.

Asymptotic Optimality

Weights using Nearest Neighbors:

Thm: If Assumptions 1-3 hold and $k = \min\{\lceil CN^\delta \rceil, N-1\}$, then $\hat{z}_N^{k\mathrm{NN}}(x)$ is asymptotically optimal.

Weights using Parzen windows:

Thm: If Assumptions 1-3 hold, $h_N = CN^{-\delta}$, and costs satisfy $\mathbb{E}\left[|c(z;Y)|\left(\log|c(z;Y)|\right)_+\right] < \infty$, then $\hat{z}_N^{\mathrm{KR}}(x)$ is asymptotically optimal.

Weights using Recursive Parzen windows:

Thm: If Assumptions 1-3 hold and $h_i = Ci^{-\delta}$, then $\hat{z}_N^{\text{Rec-KR}}(x)$ is asymptotically optimal.

Weights using LOESS:

Thm: If Assumptions 1-3 hold, μ_X is abs. cts., costs dominated, and $h_N = CN^{-\delta}$. Then $\hat{z}_N^{\mathrm{LOESS}}(x)$ is asymptotically optimal.

Computational tractability

- Construct predictive prescriptions of the form

$$\hat{z}_N(x) \in \arg\min_{z \in \mathcal{Z}} \sum_{i=1}^{N} w_N^i(x) c(z; y^i)$$

Thm: if $c(z;y)$ is convex + subgrad oracle, $\mathcal{Z}$ is convex and separation oracle is given, then we can compute $\hat{z}_N(x)$ in polynomial time and oracle calls.

Case Study: Distribution Arm of International Media Conglomerate

- silkroute provides analytics solutions for manufacturers, distributors and retailers
- Client is Fortune Global 100 company
  - 100+ million units of entertainment media shipped per year
  - Sells 1/2 million different titles on CD/DVD/Bluray at over 40,000 retailers worldwide
  - Need: SaaS solution for **Vendor-Managed Inventory** with **Scan-Based Trading**

- Our target: Maximize number media sold

Case Study: Distribution Arm of International Media Conglomerate

- Want to maximize number of items sold.
- Focus on video media, Europe

$$\max \quad \mathbb{E}\left[\sum_{j=1}^{d} \min\{Y_j,\, z_{trj}\} \,\middle|\, X = x_{tr}\right]$$

$$\text{s.t.} \quad \sum_{j=1}^{d} z_{trj} \leq K_r$$

$$z_{trj} \geq 0 \qquad \forall j = 1, \ldots, d$$

Case Study: Distribution Arm of International Media Conglomerate

- Key issues:
  - Limited shelf space at retail locations
  - Huge array of potential titles
  - Highly uncertain demand for new releases


PHARMACY
UR ONE STOP
R FLU SHOTS
4.75
sale
HANGOVER PART III
BOURNE LEGA
RUSH
LORAX
BRIDES MAIDS
GRINCH

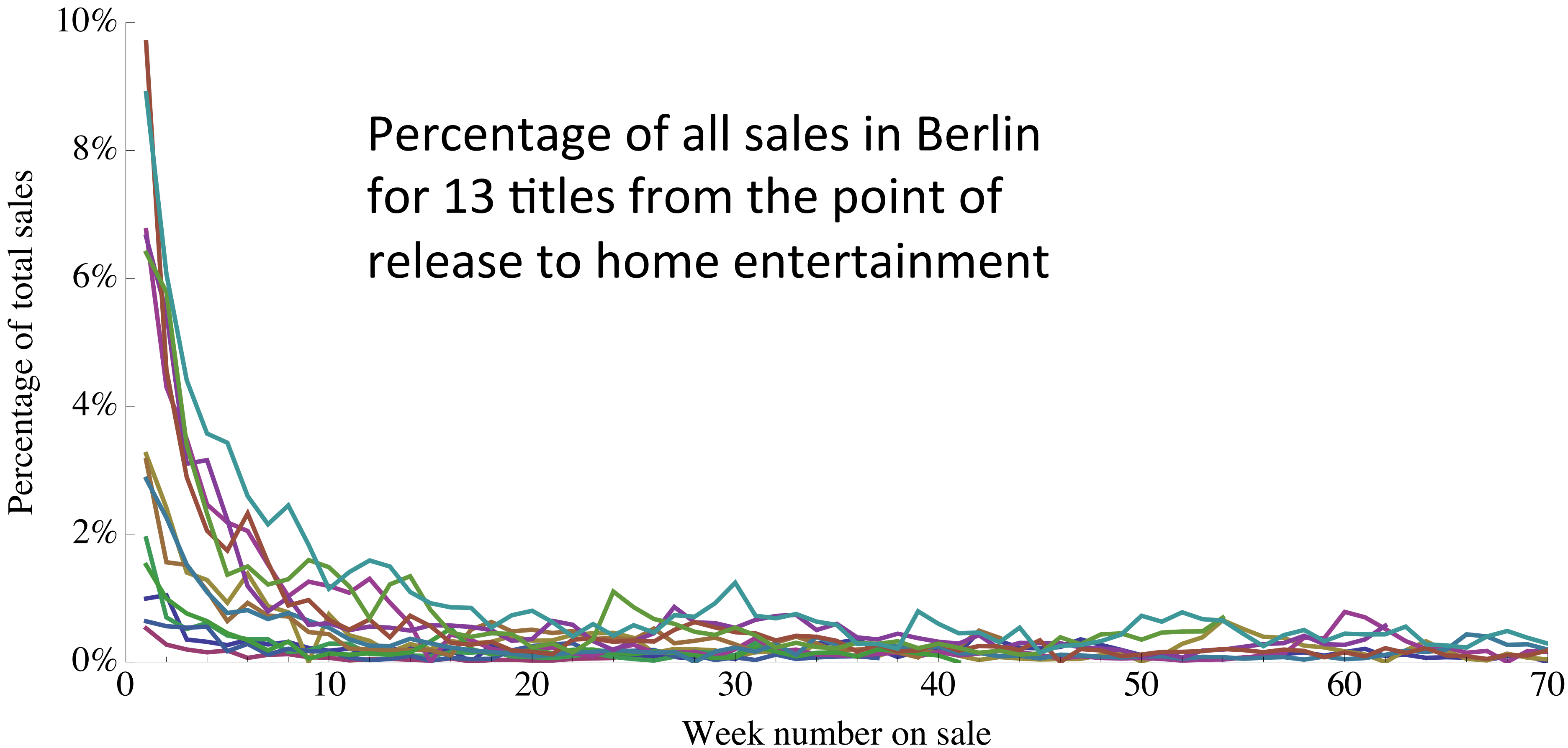
Case Study: Distribution Arm of International Media Conglomerate

- Key issues:
  - Limited shelf space at retail locations
  - Huge array of potential titles
  - Highly uncertain demand for new releases

Release date: 5/24/16

Internal Company Data

- Sales by item/location, 2010 to present
- ~50GB *after* aggregating transaction records by week



Dealing with Censored Data

- Observe sales, not demand (quantity of interest Y)

$$U = \min\{Y,\, V\}$$

- Adjust weights for right-censored data

$$\tilde{w}_{N,(i)}(x) = \begin{cases} \left(\frac{w_{N,(i)}(x)}{\sum_{\ell=i}^{N} w_{N,(\ell)}(x)}\right) \prod_{k\le i-1\,:\,u^{(k)}<v^{(k)}} \left(\frac{\sum_{\ell=k+1}^{N} w_{N,(\ell)}(x)}{\sum_{\ell=k}^{N} w_{N,(\ell)}(x)}\right) & \text{if } u^{(i)} < v^{(i)}, \\ 0 & \text{otherwise.} \end{cases}$$

Thm: Under same assumptions as before and if in addition (a) Y and V conditionally independent given X, (b) Y and V share no atoms, and (c) upper support of V greater than that of Y given $X = x$, then $\hat{z}_N(x)$ is ***asymptotically optimal***.

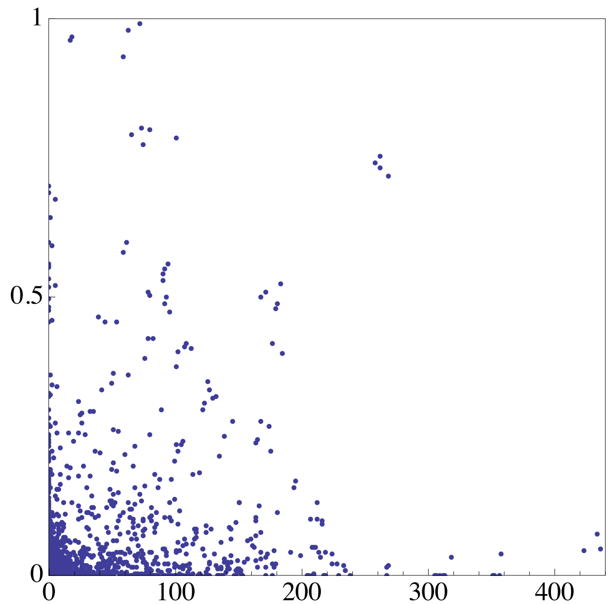
Internal Company Data

- Sales by item/location, 2010 to present
- ~50GB *after* aggregating transaction records by week
- Location info:
  - Address
    - Google Geocoding API
- Item info:
  - Medium (DVD/BLU)
  - Obfuscated title
    - Disambiguation

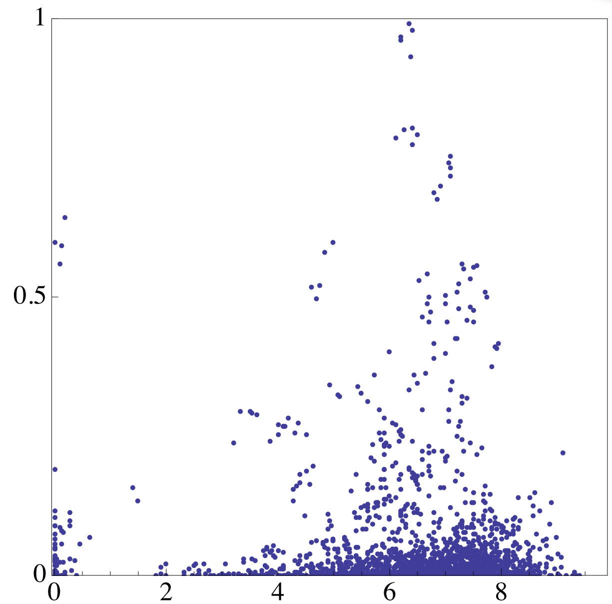
Beyond internal company data: Harvesting public data (more X)

- **Movie/series**
- **Actors** (find actor communities; Blondel et al 2008)
- **Plot summary** (cosine similarities, hierarchically clustered)
- **Box office gross, US**
- **Oscar wins and nominations** and other awards
- **Professional (meta-)ratings, user ratings**
- **Num of user ratings**
- **Genre** (can be multiple)
- **MPAA rating**

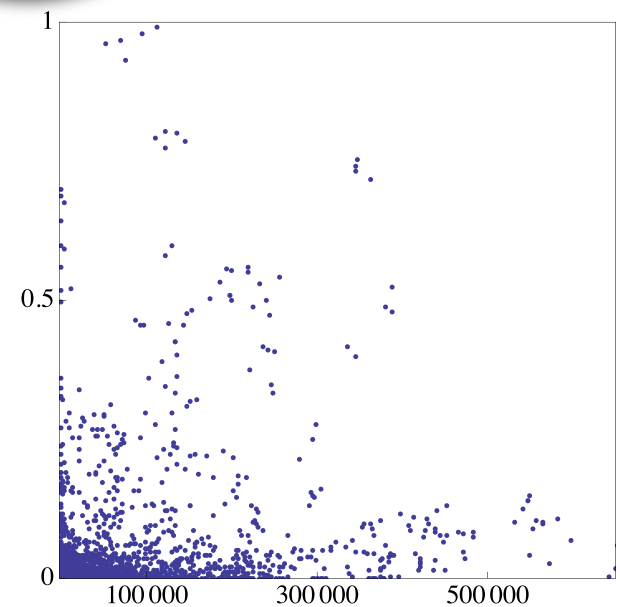
Beyond internal company data: Harvesting public data (more X)



Box office gross
$\rho = 0.32$

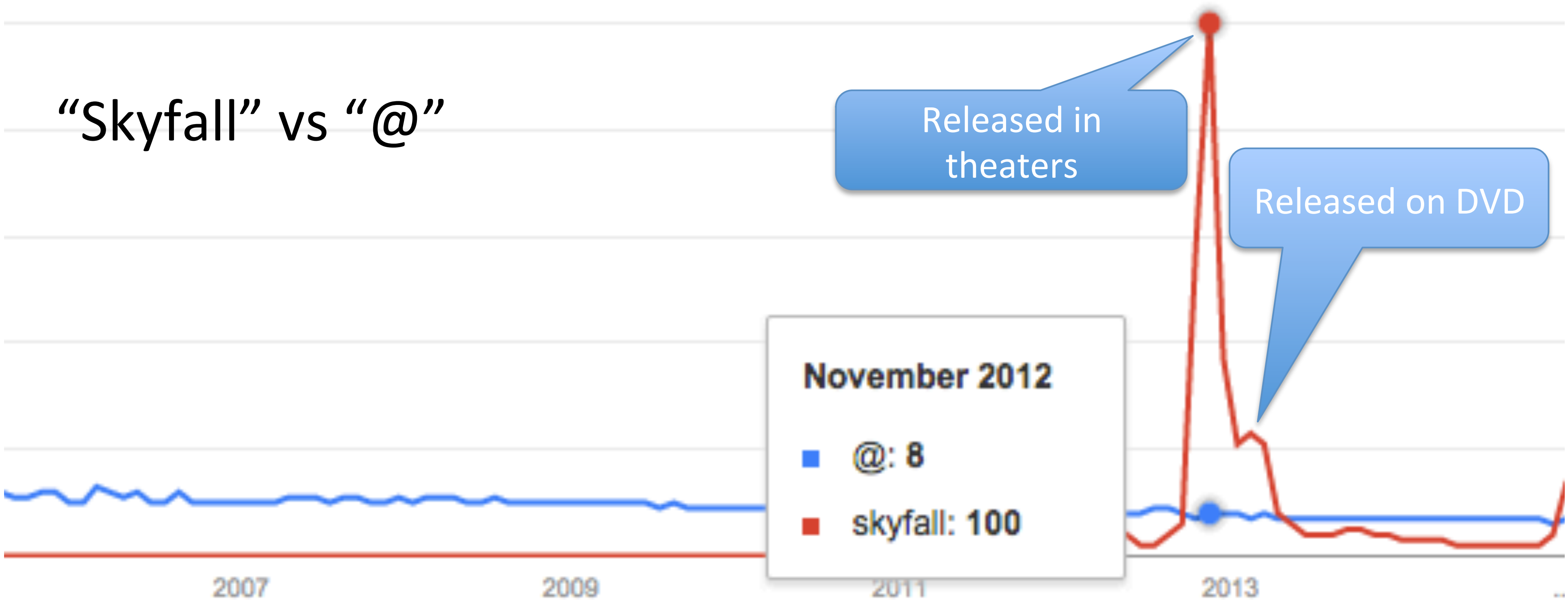


IMDb rating
$\rho = 0.02$

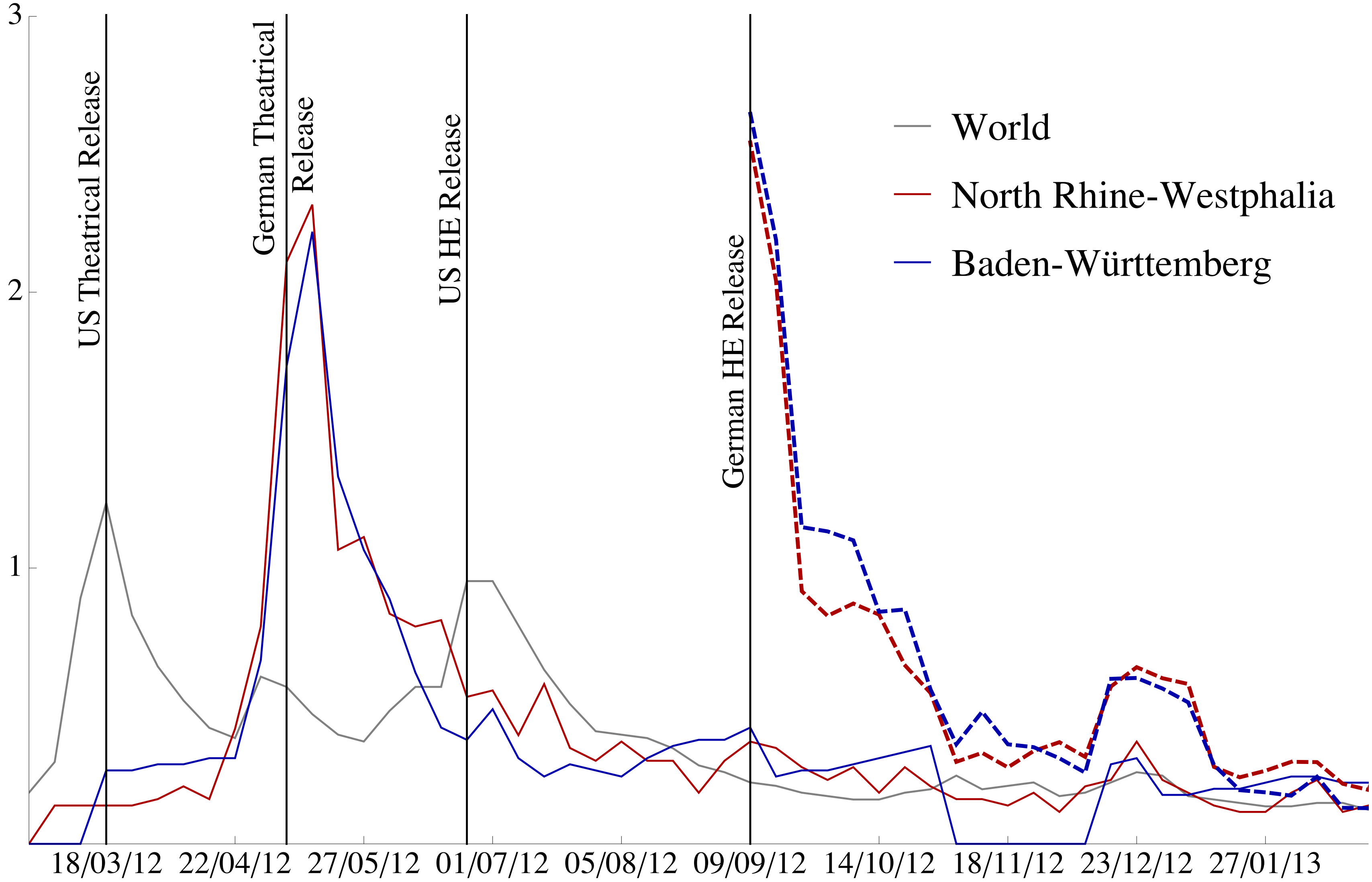


Number user votes
$\rho = 0.25$

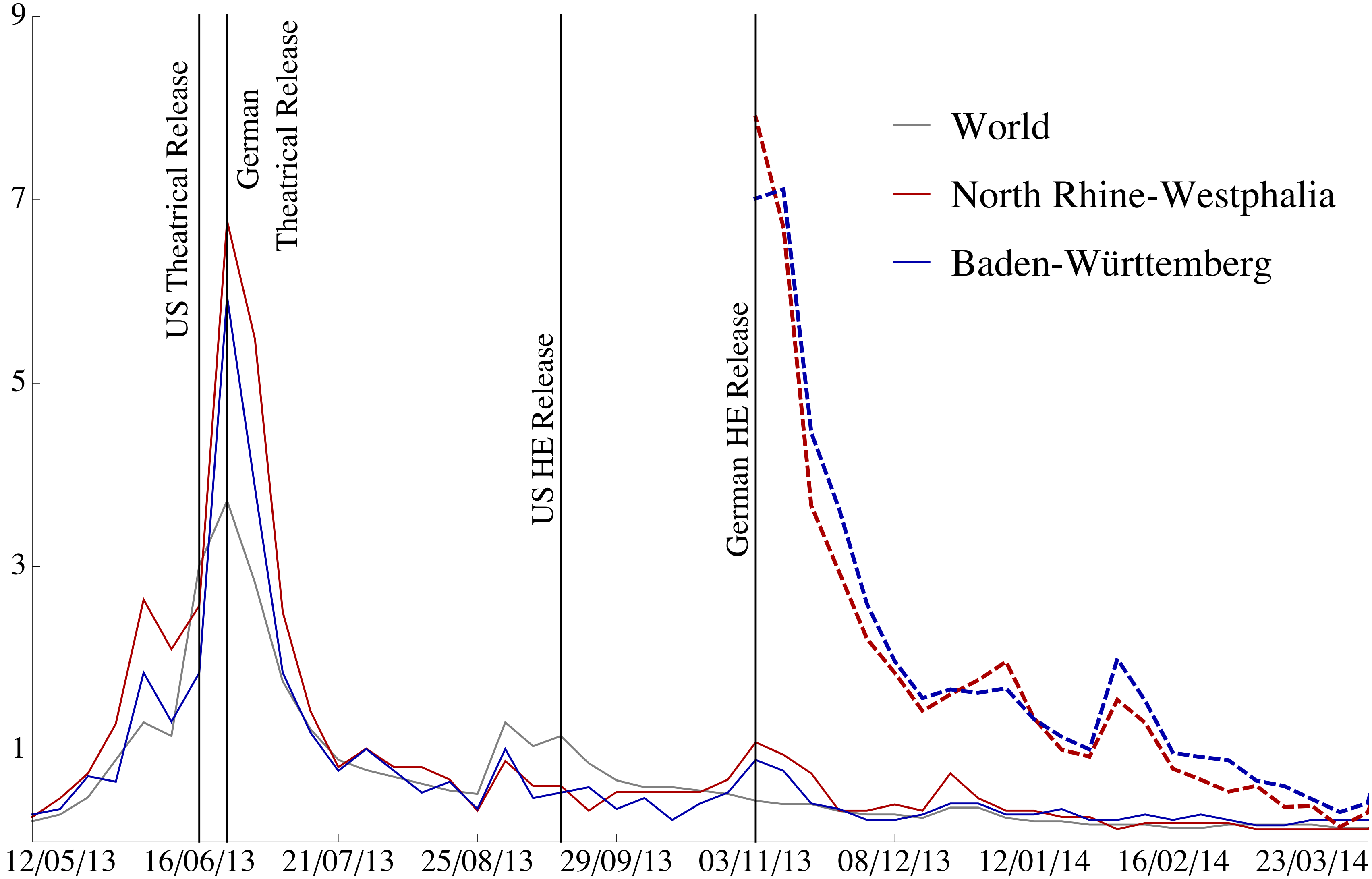
Beyond internal company data: Harvesting public data (more X)



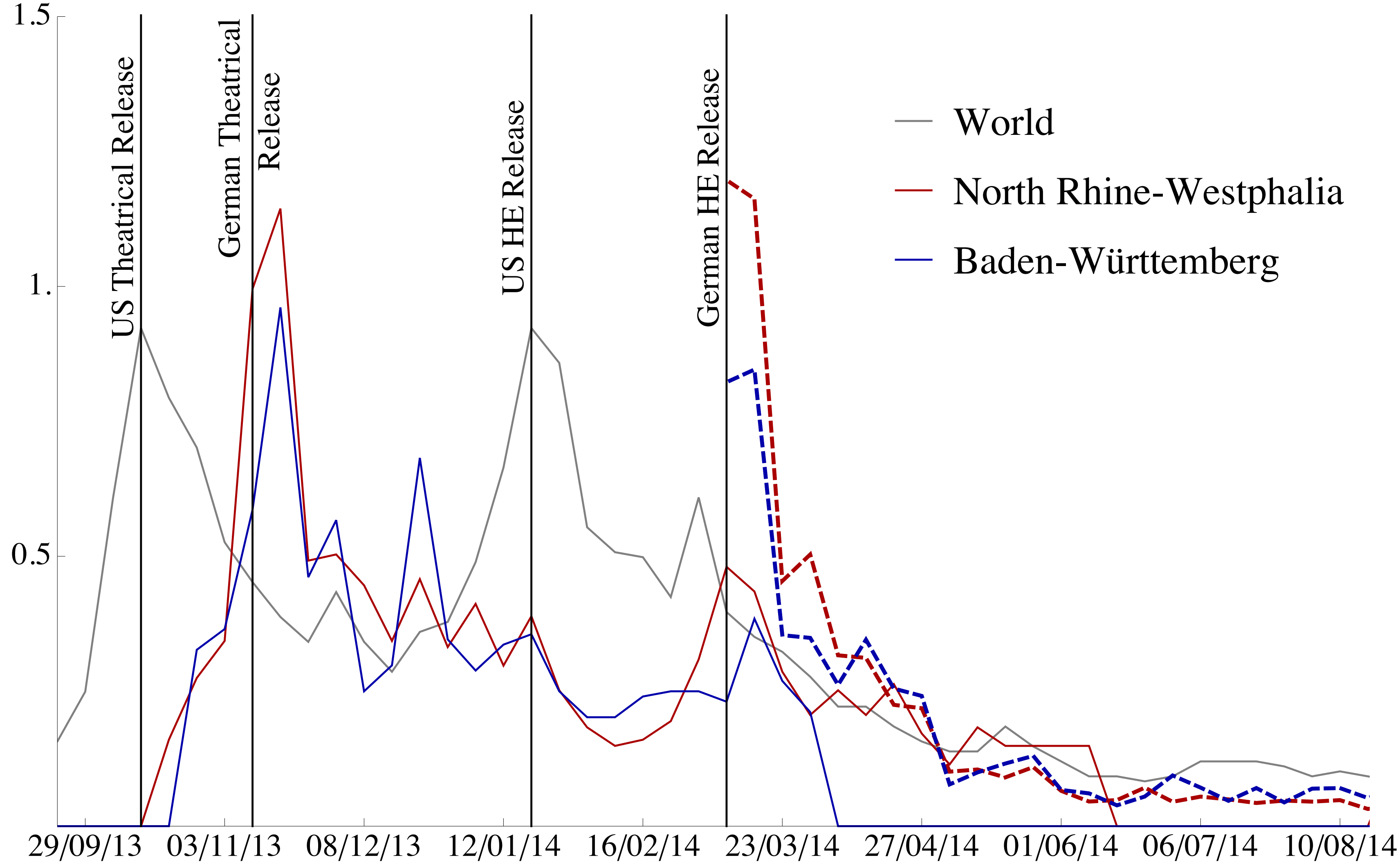
Beyond internal company data: Harvesting public data (more X)



Beyond internal company data: Harvesting public data (more X)



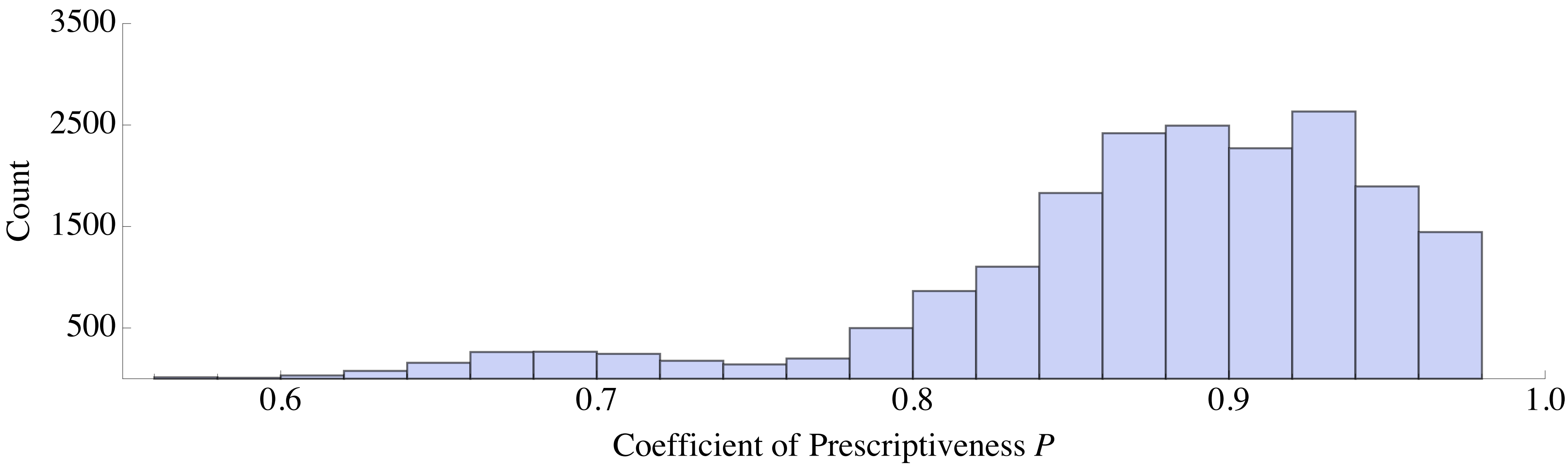
Beyond internal company data: Harvesting public data (more X)



Prescribing Order Quantities

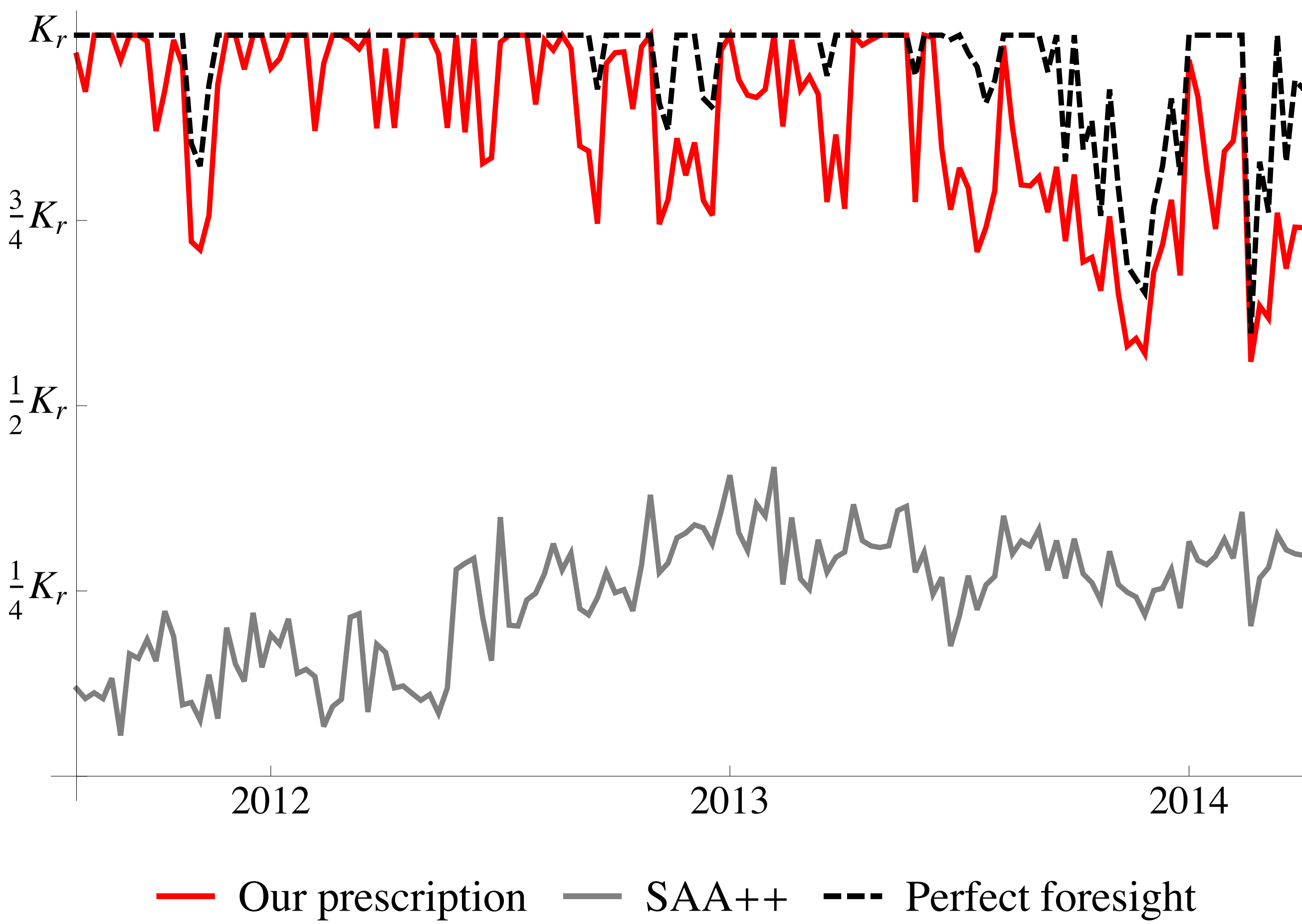
- Using our bagged prescription trees and all our data…

- ***Out-of-sample***
  P = 0.88

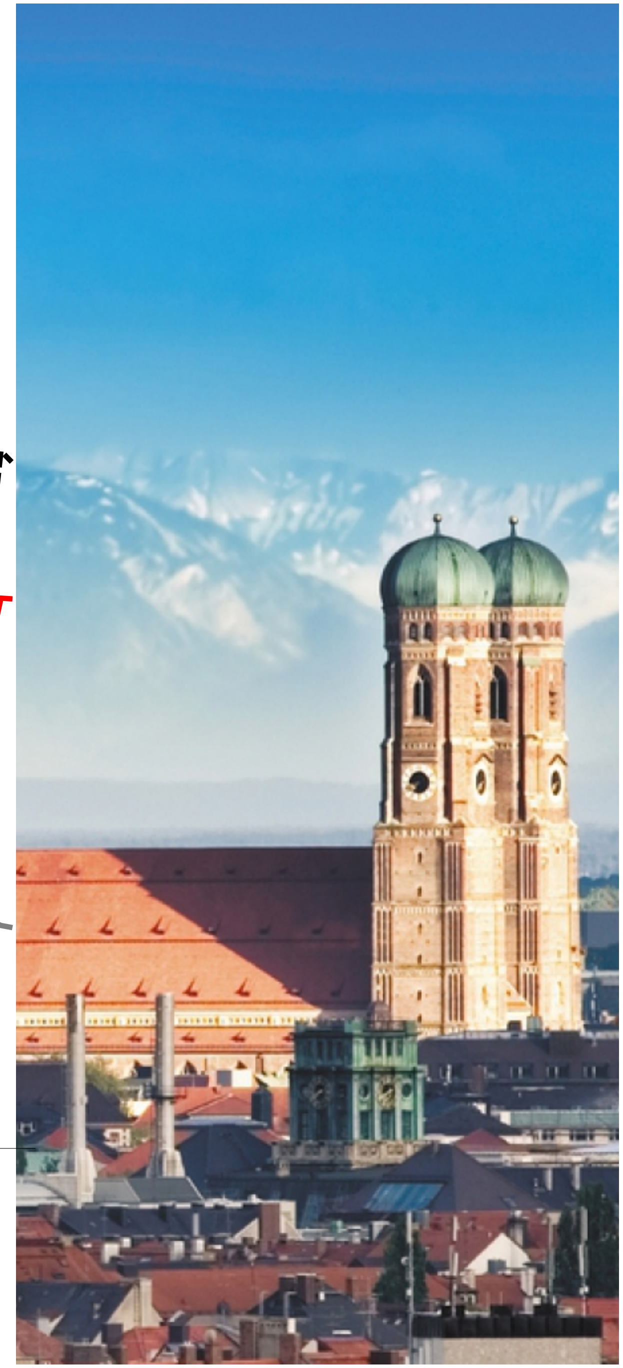


Munich

$$P = 0.89$$

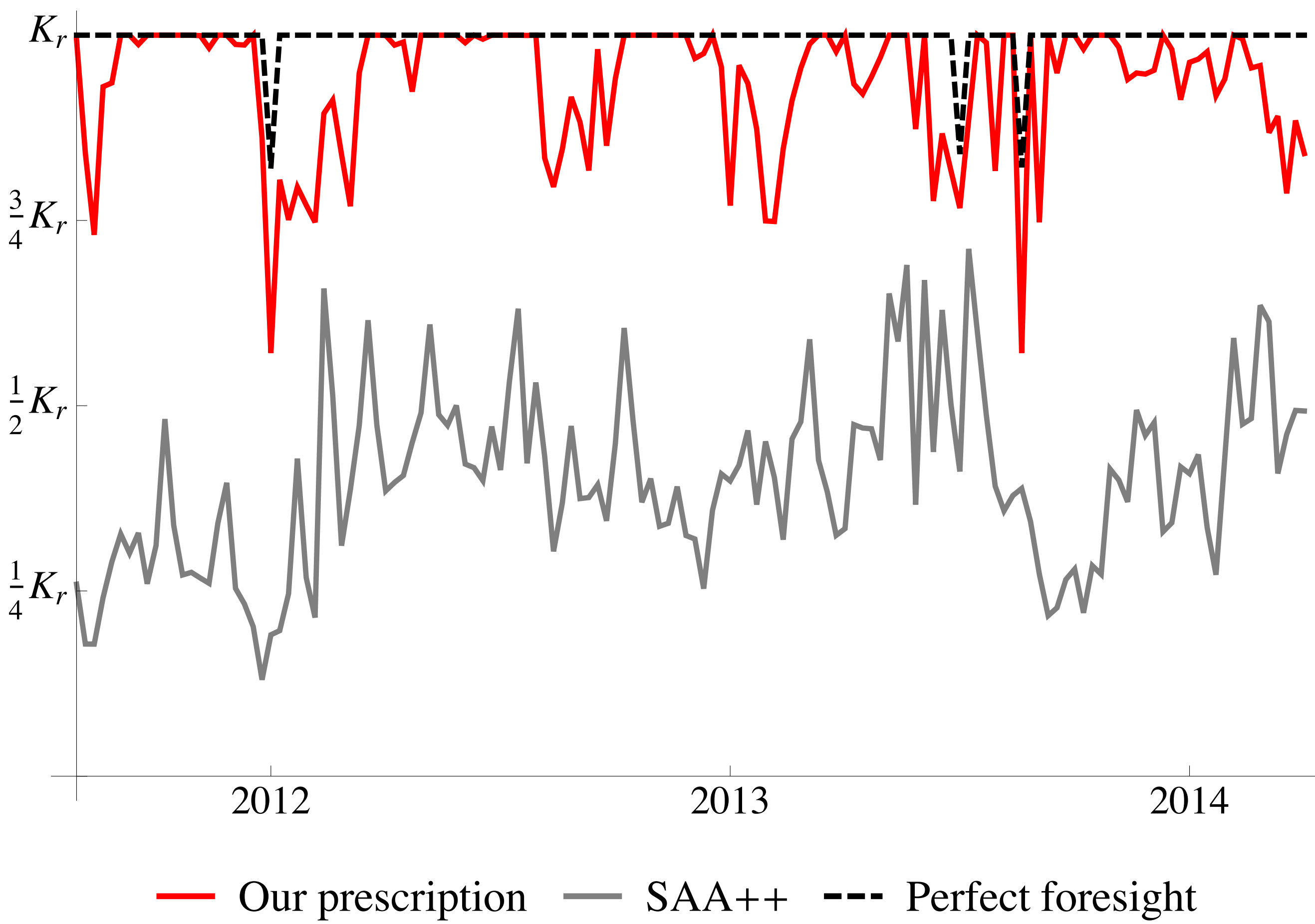


K_r
$\frac{3}{4}K_r$
$\frac{1}{2}K_r$
$\frac{1}{4}K_r$
2012
2013
2014
Our prescription
SAA++
Perfect foresight

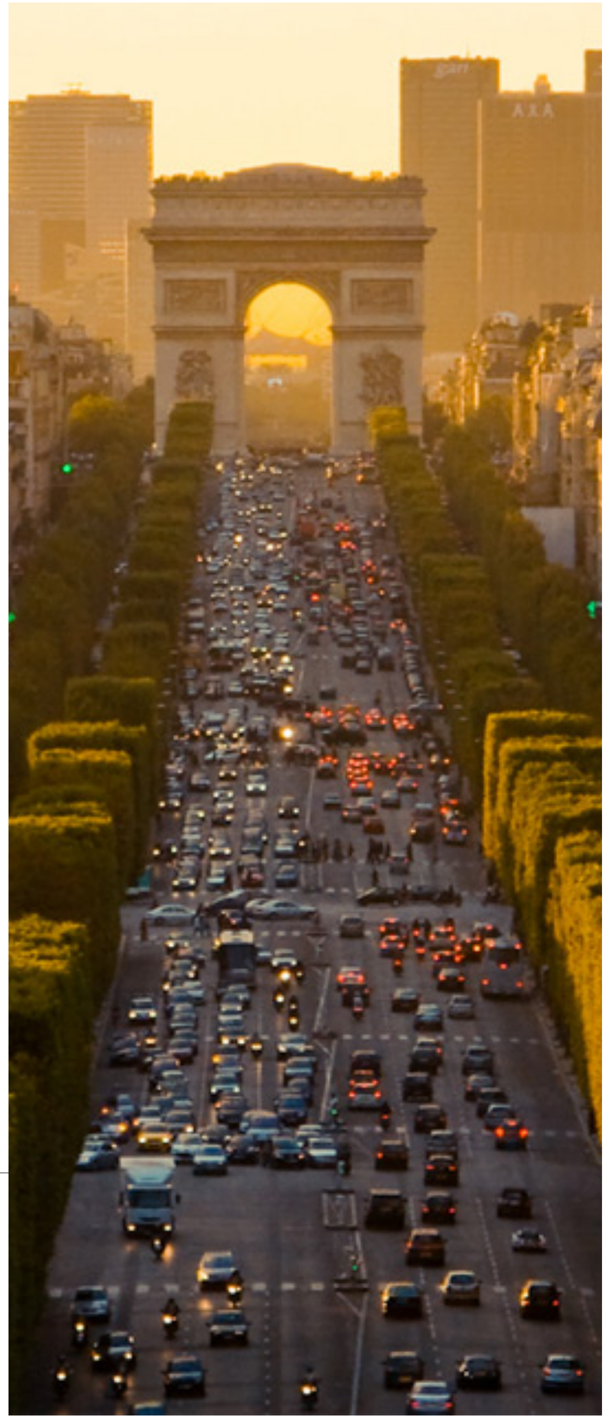


Paris

$$P = 0.90$$



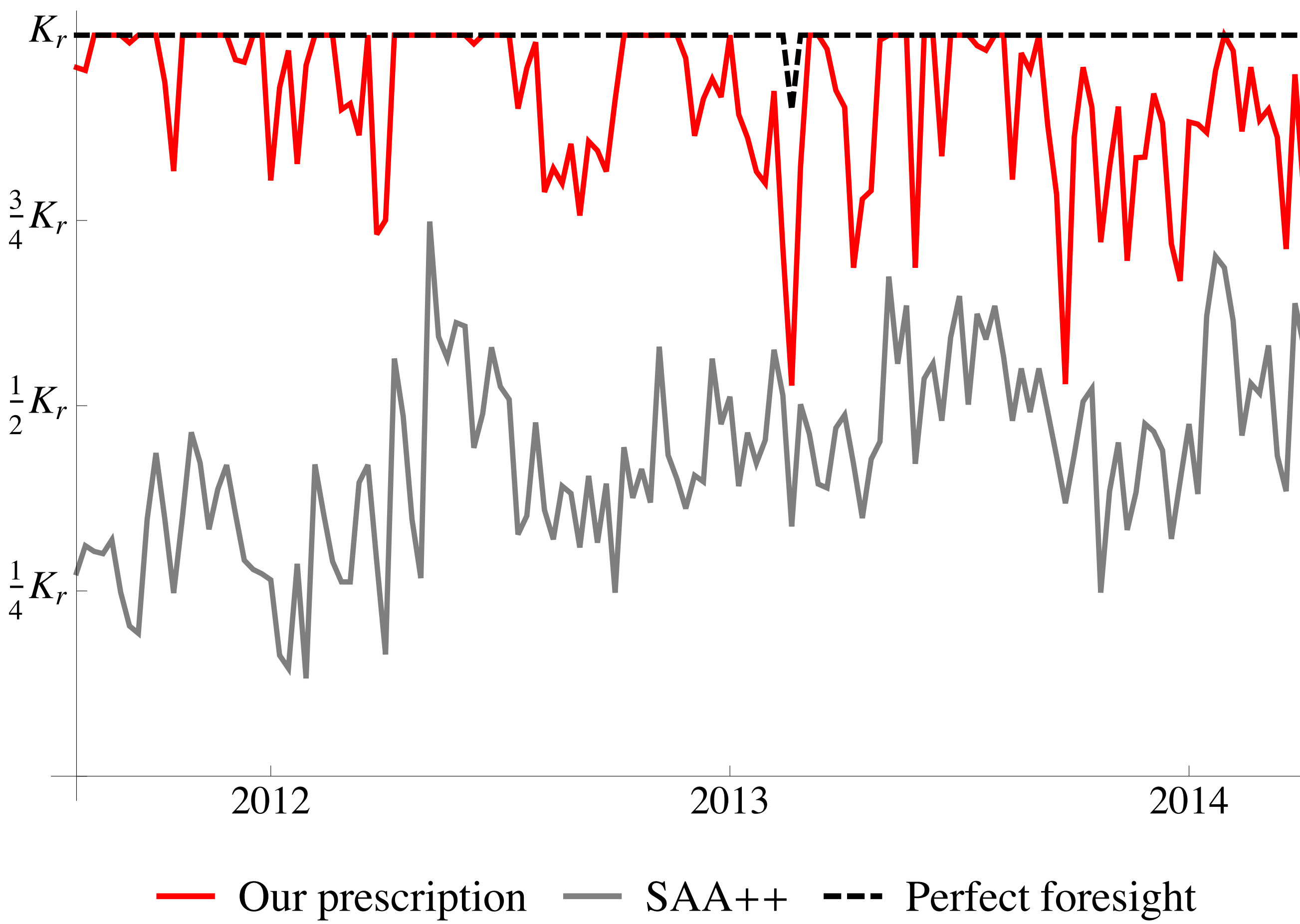
K_r
$\frac{3}{4}K_r$
$\frac{1}{2}K_r$
$\frac{1}{4}K_r$
2012
2013
2014
Our prescription
SAA++
Perfect foresight



AXA

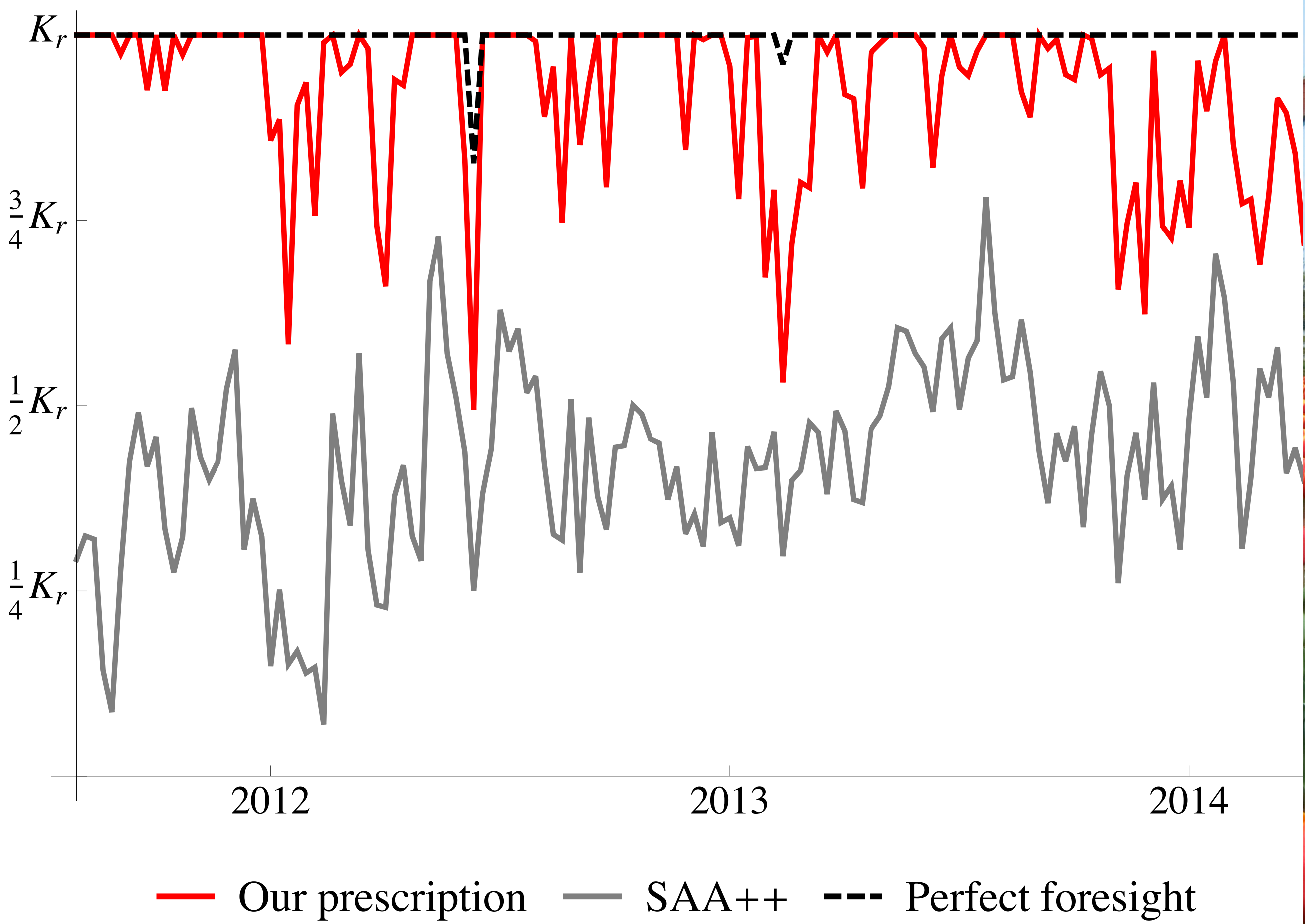
Waterloo

$$P = 0.85$$

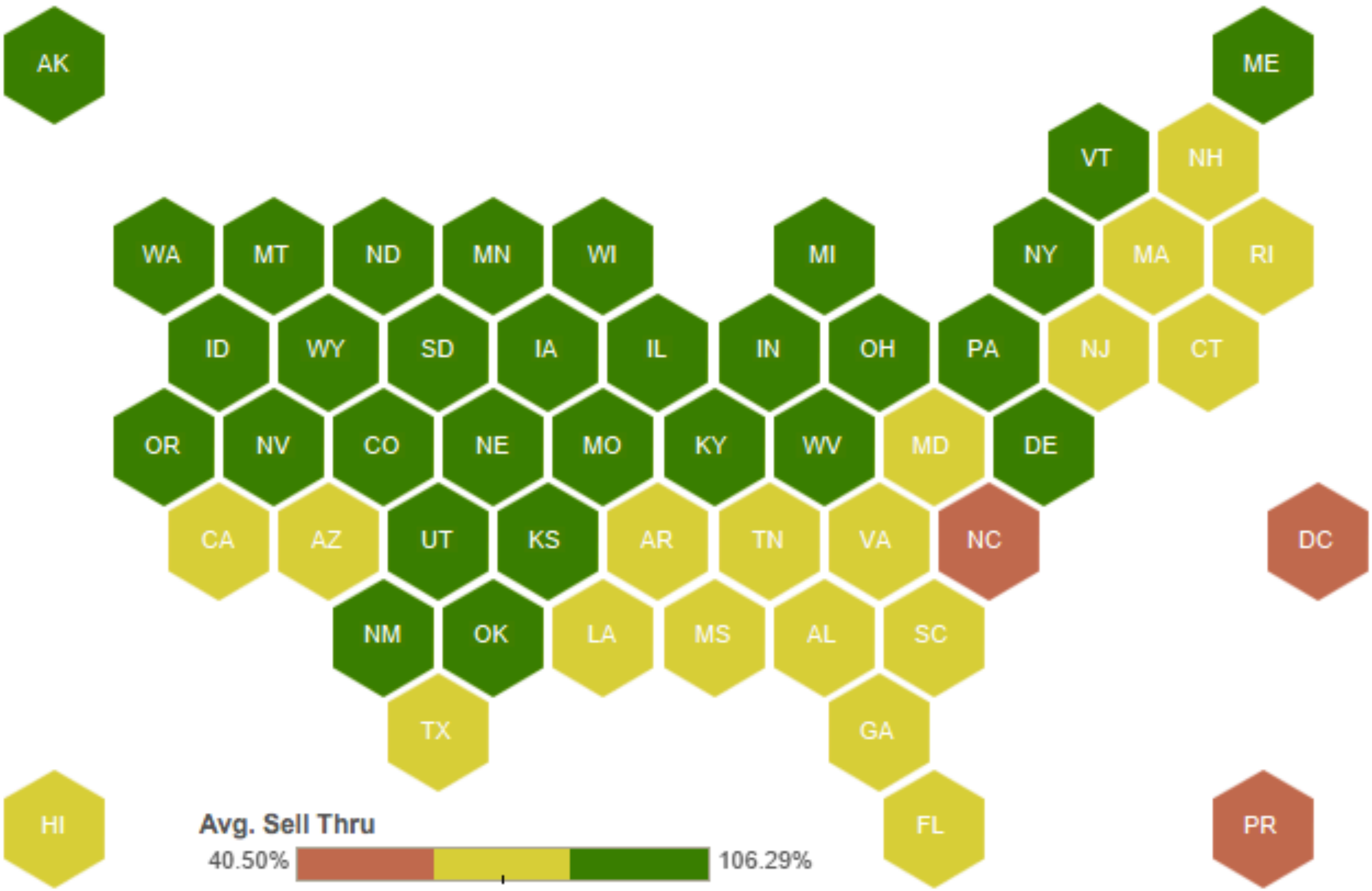


The Hague

$$P = 0.86$$



Initial Shipments	Shipped Units	Sold Units	Returned Units	Current On Hand Qty	Avg. Unit Retail	Current WOS	Sell Thru	Day 1 Sell Thru	Week 1 Sell Thru	4 Week Sell Thru	Returns %	Allocation Accuracy
415,182	531,897	433,206	103,108	6.81	$21.02	22.46	81.45%	11.17%	53.28%	87.57%	19.38%	90.97%



Contributions

- **A new framework**
  - General purpose
  - Coefficient of prescriptiveness
- **Theory**
  - Computational tractability
  - Asymptotic optimality
- **Practice**
  - Case study of huge media distributor
  - In collaboration with Silkroute
  - Study ***prescriptive*** power of large-scale data

Thank you!