From Predictive to Prescriptive Analytics

Dimitris Bertsimas

Joint work with Nathan Kallus

Massachusetts Institute of Technology

OPERATIONS RESEARCH CENTER

Philosophy/Motivation

- Operations Research/Management Science typically starts with models and aims to make optimal decisions. Data is often an afterthought.
- Machine Learning/Statistics typically starts with data and aims to make predictions. Decisions are typically not considered.
- Models, in my opinion, exist in our imagination.
- Yet, we typically teach models to our students.

Opportunity

- Availability of data (often big data) in electronic form.
- Can we develop a theory that unifies OR/MS and

  ML/S that goes:

  From data to predictions to prescriptions?
- What are the implications in Education? Should we change our courses?

Contributions

- A theory that unifies OR/MS and ML/S that goes

  From data to predictive to prescriptive analytics.

- Asymptotic optimality, as the data increases the decisions are optimal.
- Computational Tractability
- Coefficient of Prescriptiveness P that measures how much auxiliary data improves a baseline, generalizing R^2 for predictions.
- Real world example that shows that this theory makes a material improvement in a Global Fortune 100 multimedia media company.

A Real World Problem

- A Global Fortune 100 multimedia company.
- 1 billion units of entertainment media shipped per year
- Sells 1/2 million different titles on CD/DVD/Bluray at over 50,000 retailers worldwide

Key Issues

- Limited shelf space at retail locations
- Huge array of potential titles
- Highly uncertain demand for new releases
- Which titles to order and in what quantities?
- Maximize number media sold

Data

- 4 years of sales data across a network of 50,000 retailers
- Data harvested from public online sources
- How to leverage all this data?

The general problem

- Data $y^1, \dots, y^N$ on quantities of interest Y
  E.g. demands at locations/of products,
- Data $x^1, \dots, x^N$ 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 $X = x$

Outline

- From Predictive to Prescriptive Analytics
  - A gap in decision making
  - Our approach
  - Asymptotic optimality
  - Coefficient of Prescriptiveness
  - Real world problem

Standard Data-Driven Optimization

- Data $y^1, \ldots, y^N$ on quantities of interest Y
- Decision $z \in \mathcal{Z}$ to minimize *uncertain* costs $c(z; Y)$
- Problem of interest is

$$\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z; Y)\right]$$

- But we only have data
- Distributions only exist in our imagination

Standard Data-Driven Optimization

- Problem of interest is

$$\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z;Y)\right]$$

- Standard data-driven solution is sample average approximation

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

- Other approaches: Stochastic Approximation (Robins 1951),
  Robust SAA (Bertsimas, Gupta, Kallus 2014)

- General theorem: solutions from these approaches converge to the full-info problem (cf. Shapiro et al. 2009)

- In our problem data-driven approaches like SAA account for uncertainty but not for auxiliary data

Standard Supervised Learning in ML

- Data $y^1, \ldots, y^N$ on quantities of interest Y
- Data $x^1, \ldots, x^N$ on associated covariates X
- Obtain a prediction $\hat{m}_N(x)$ for the future value of Y after observing $X = x$ so that the squared difference between our best prediction and the true value is small.
- For example, a random forest!

Standard Supervised Learning in ML

- Problem of interest is

$$\mathbb{E}\left[Y \middle| X = x\right]$$

- How to use for decision-making?

- Fit a ML predictive model $\hat{m}_N(x) \approx \mathbb{E}\left[Y \middle| X = x\right]$ (e.g. a random forest) and solve a deterministic problem

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

- In our problem, this point-prediction-driven decision accounts for auxiliary data but not for uncertainty

The predictive prescription problem

- Problem of interest:

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

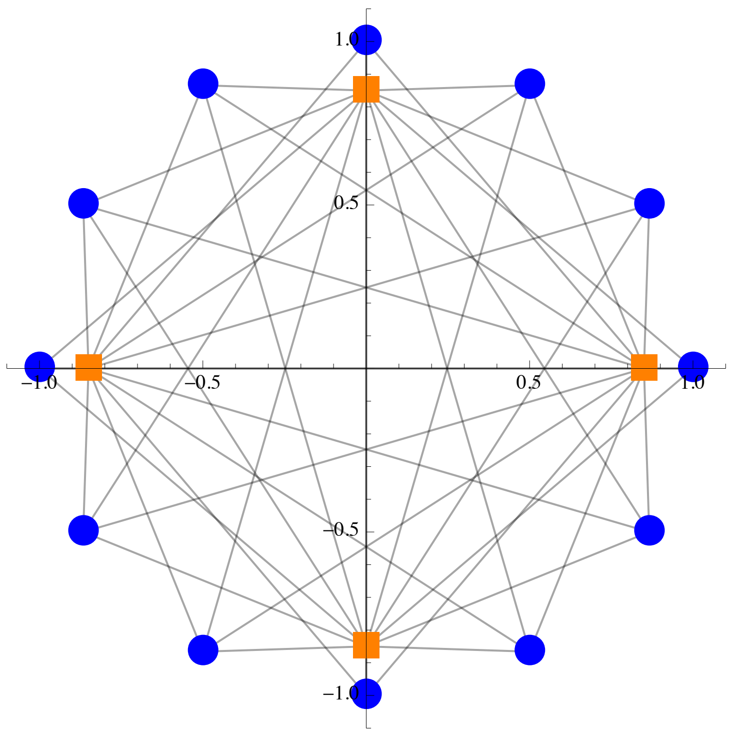
- **Task**:
  use data $S_N = \{(x^1, y^1), \ldots, (x^N, y^N)\}$ to construct a data-driven predictive prescription

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

Shipment planning example

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

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



$$\begin{aligned} \text{s.t. } t_i &\geq 0 && \forall i \\ s_{ij} &\geq 0 && \forall i,\ j \\ \sum_{i=1}^{d_z} s_{ij} &\geq y_j && \forall j \\ \sum_{j=1}^{d_y} s_{ij} &\leq z_i + t_i && \forall i \end{aligned}$$

Shipment planning example

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

X: Sales, weather forecasts, volume of Google searches

Portfolio example

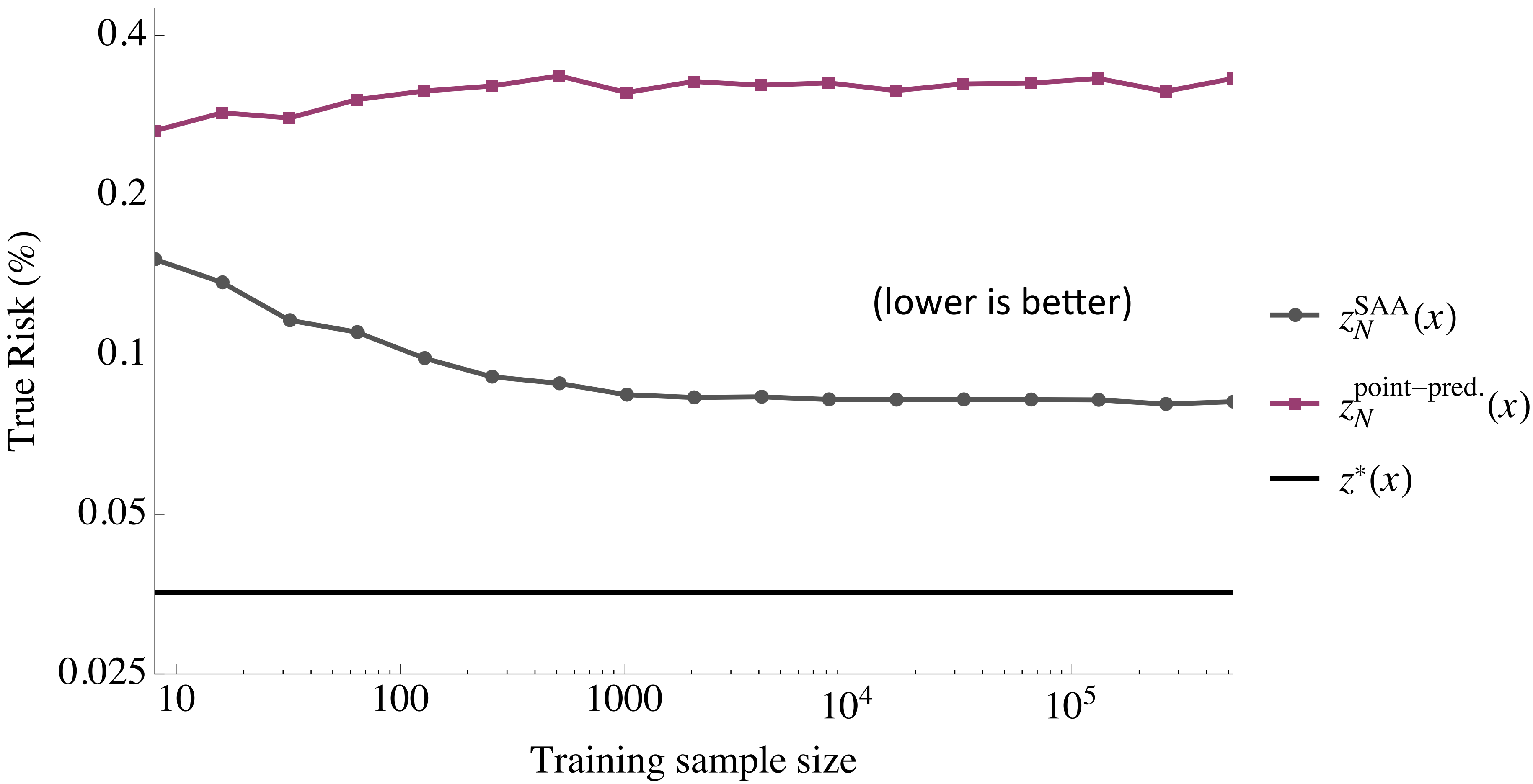
- Mean-CVaR$_{15\%}$ portfolio allocation with 12 securities
- Observe market factors X correlated with future returns

$$c((z,\beta);y) = \beta + \frac{1}{\epsilon}\max\left\{-z^T y - \beta, 0\right\} - \lambda z^T y$$

$$\mathcal{Z} = \{\beta \in \mathbb{R},\ z \geq 0,\ \sum_{i=1}^{d_z} z_i = 1\}$$

Portfolio example

- Mean-CVaR$_{15\%}$ portfolio allocation with 12 securities
- Observe market factors X correlated with future returns



Outline

- From Predictive to Prescriptive Analytics
  - A gap in decision making
  - Our approach
  - Asymptotic optimality
  - Coefficient of Prescriptiveness
  - Real world problem

Our approach

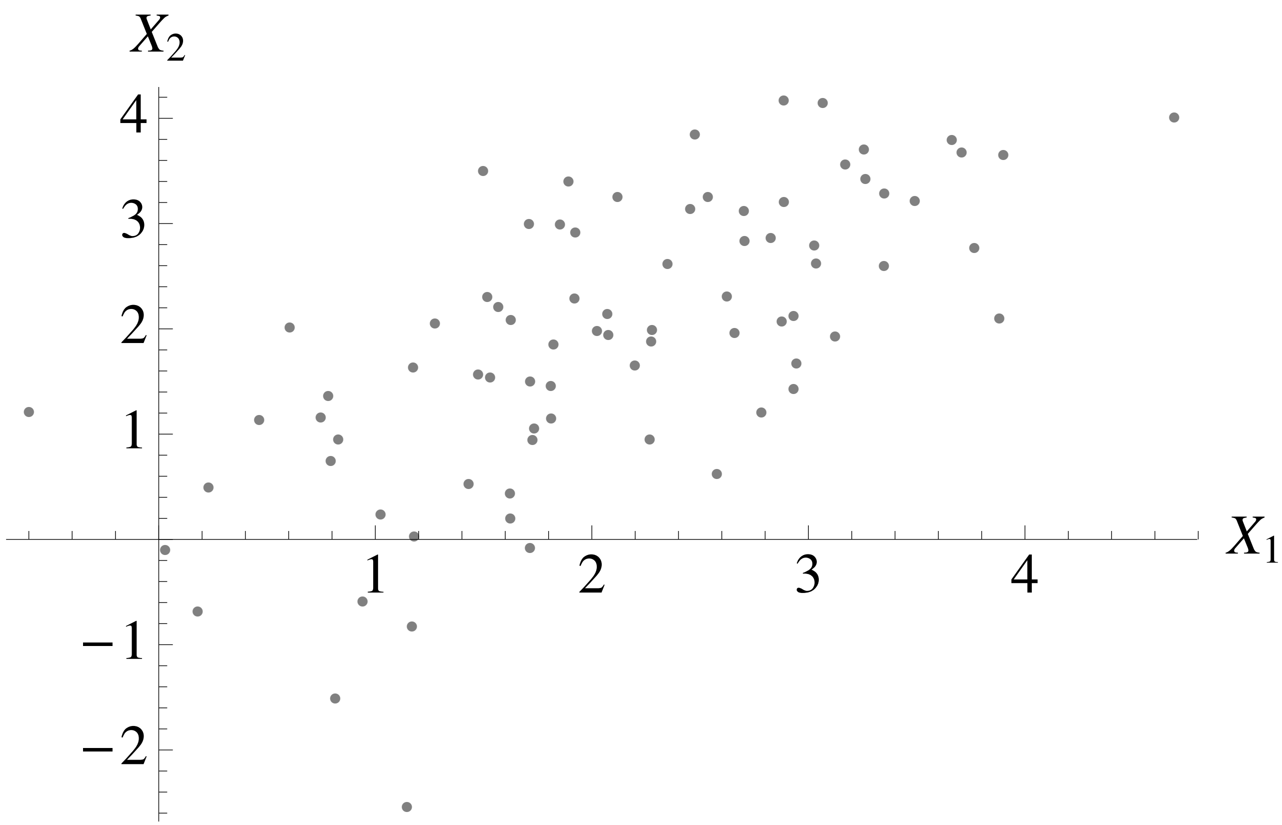
- Construct predictive prescriptions of the form

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

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

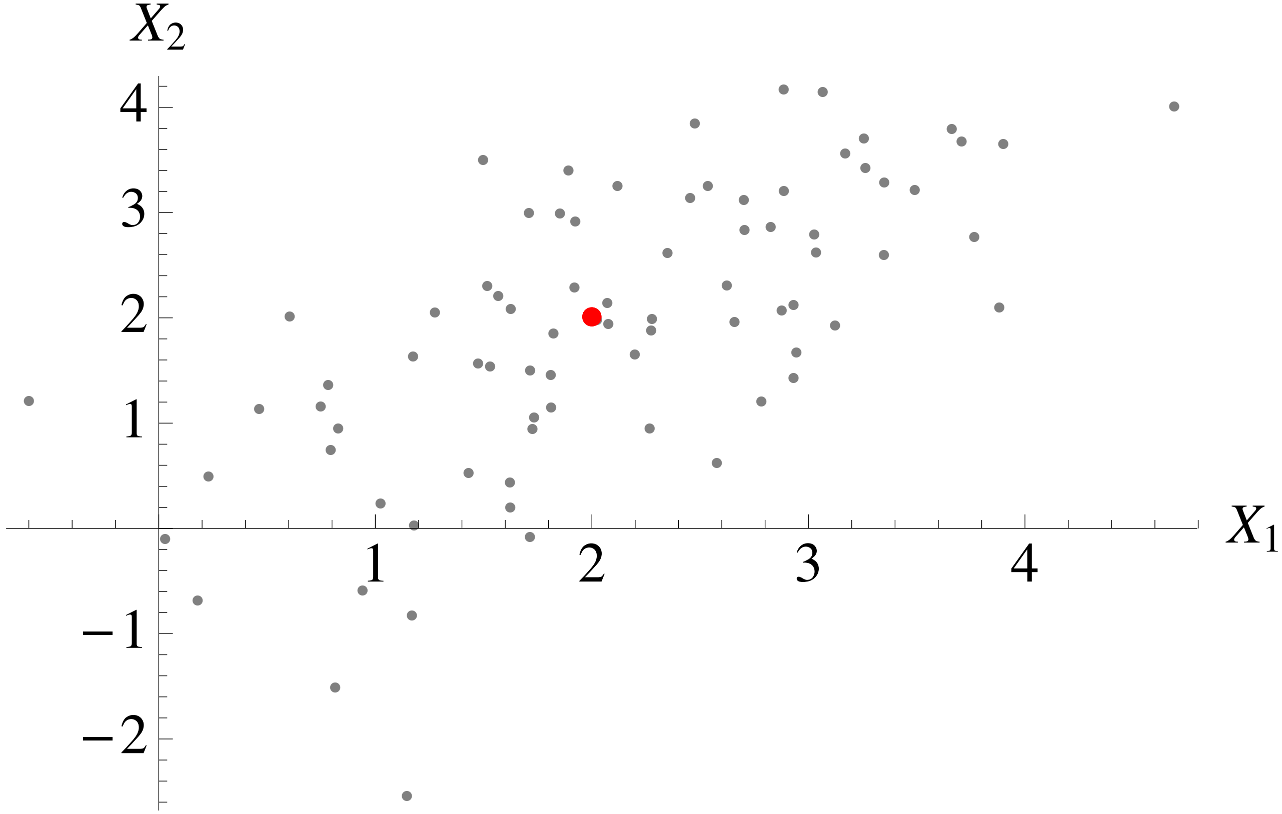
*k*NN

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



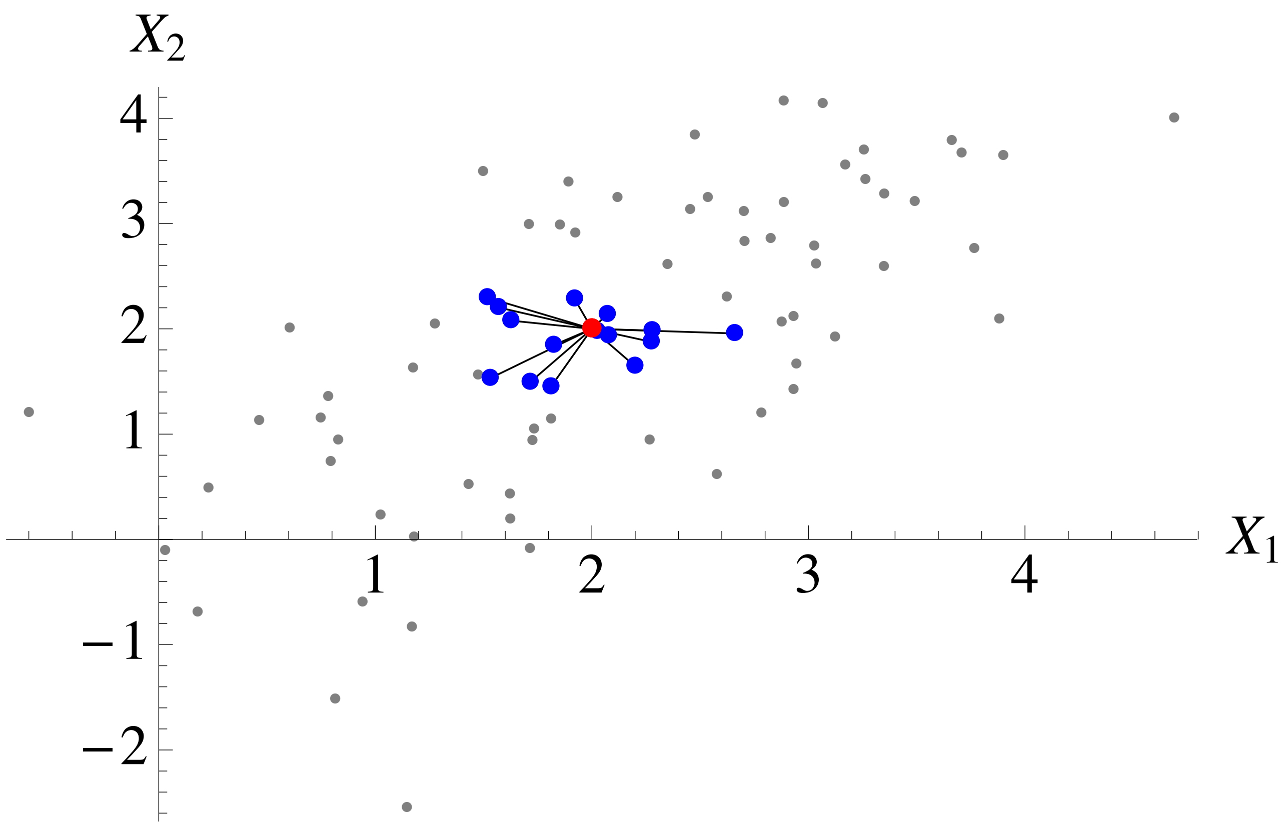
kNN

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



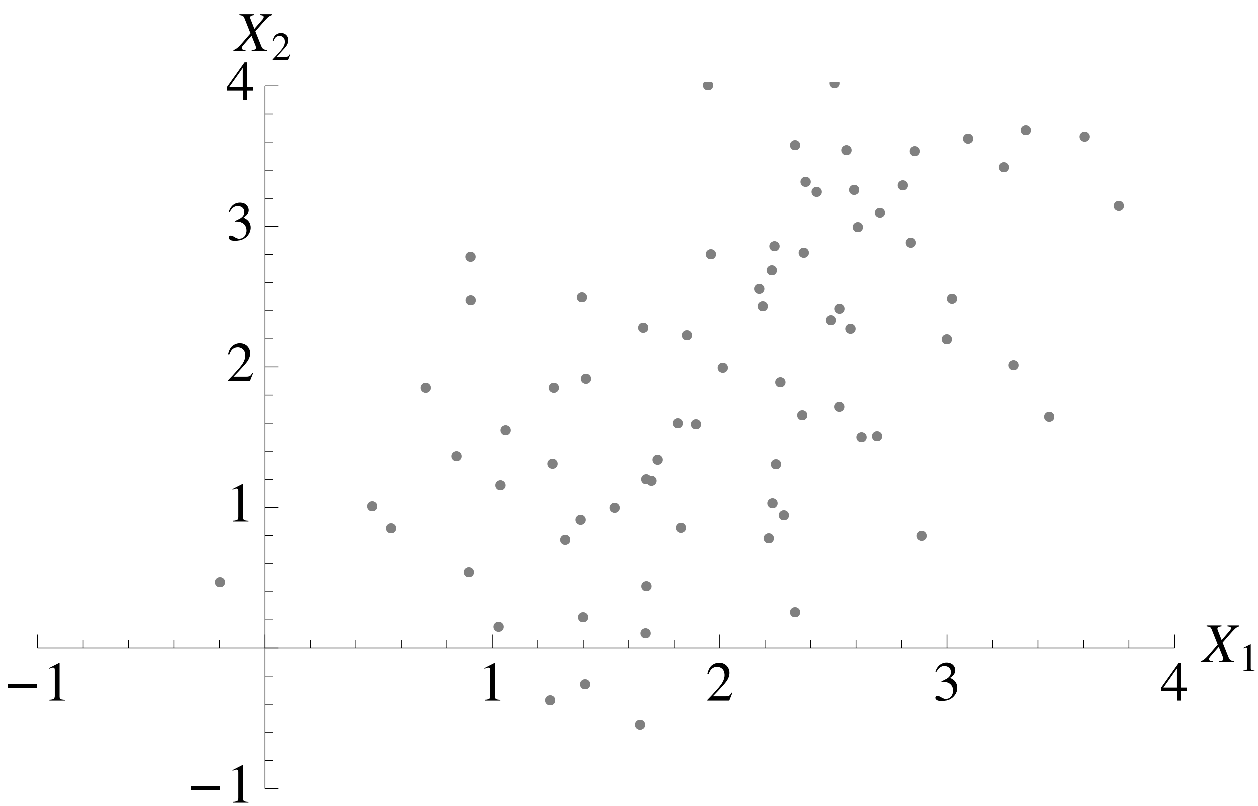
kNN

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



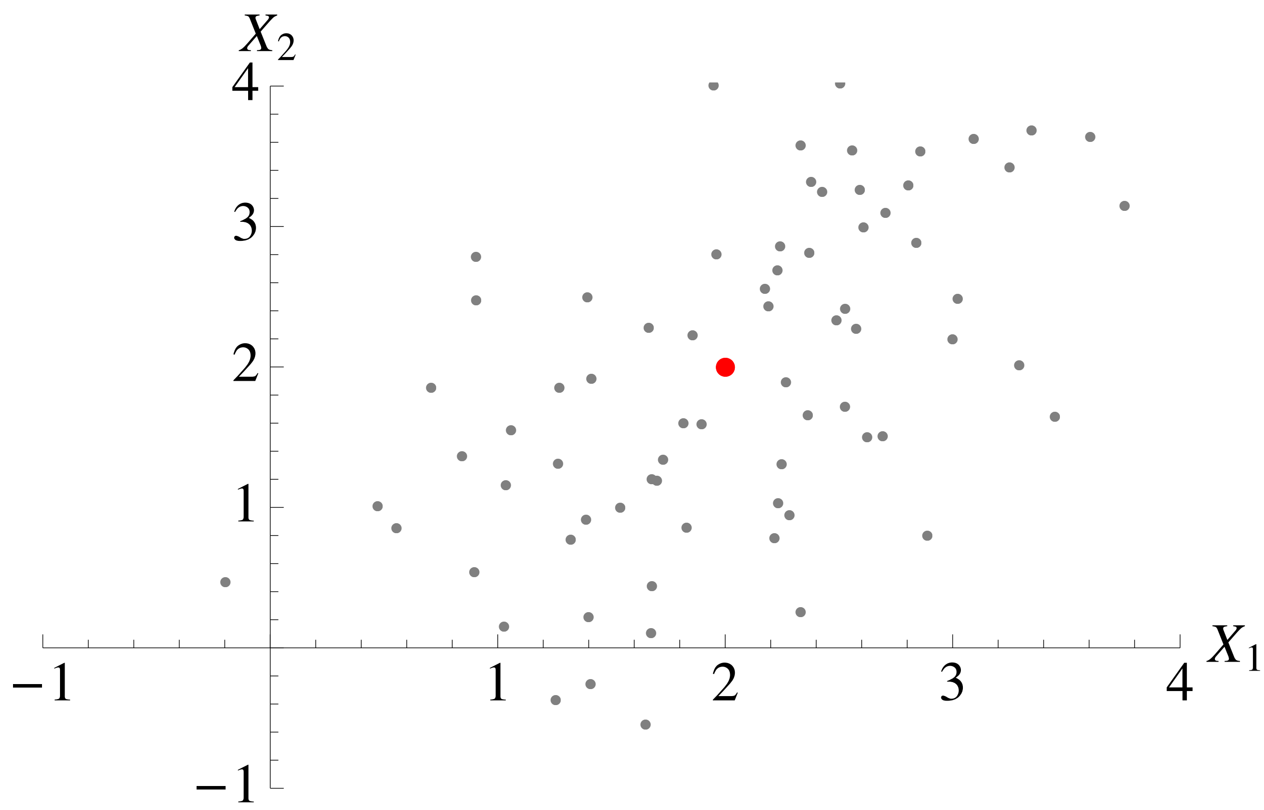
Parzen windows

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



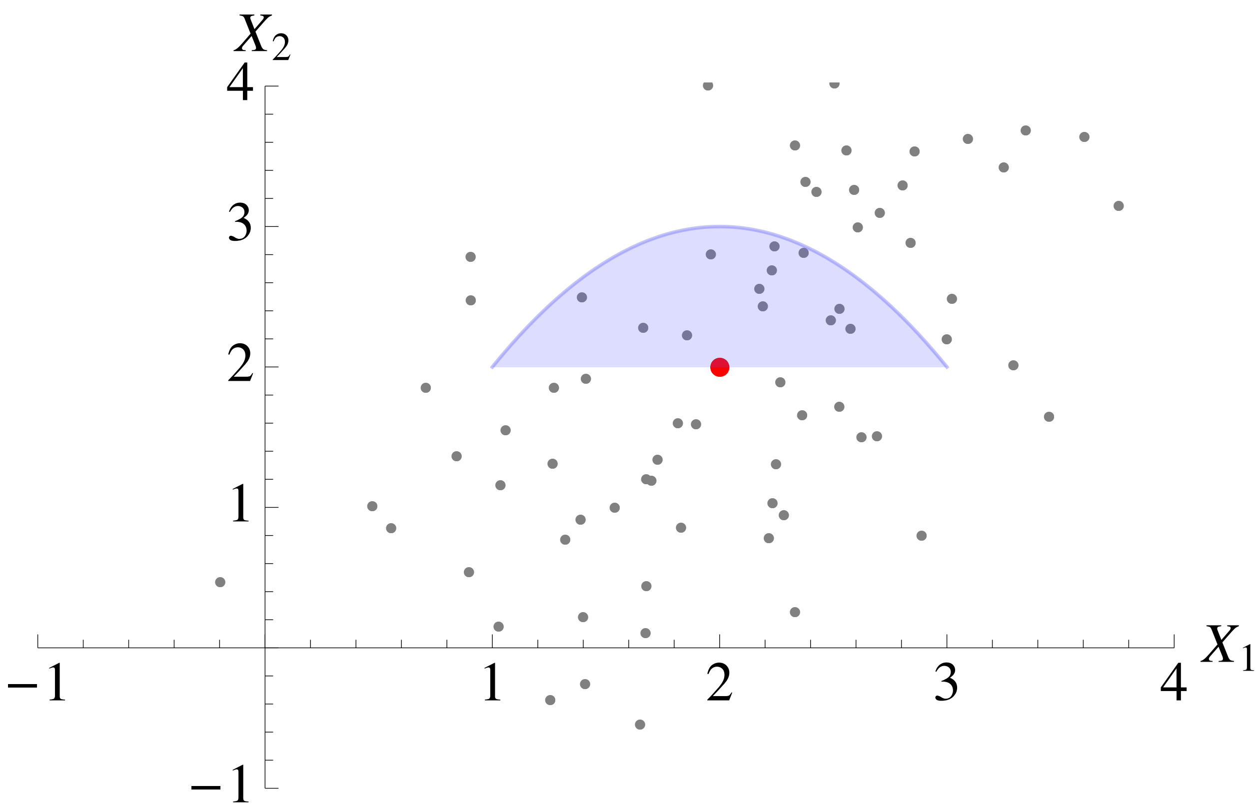
Parzen windows

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



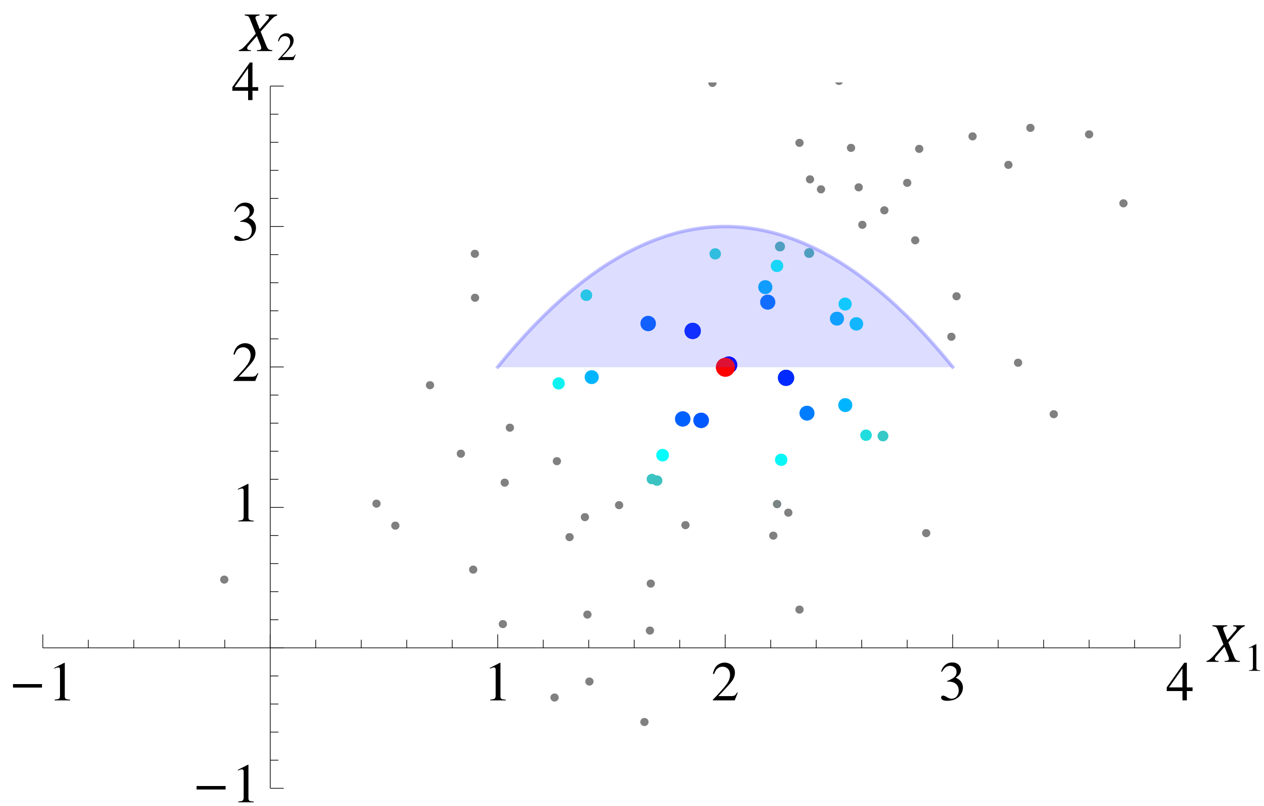
Parzen windows

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



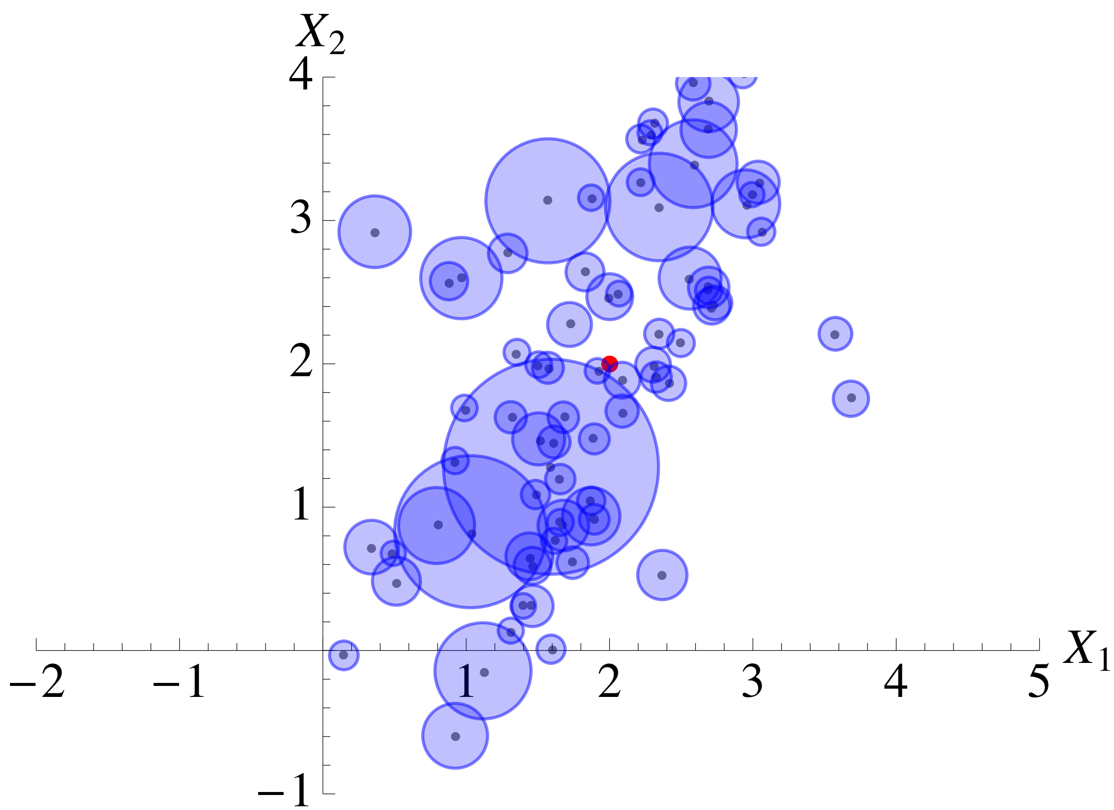
Parzen windows

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



Recursive Parzen windows

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



Local linear regression

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

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

CART

$$\left\{(x^1,\, y^1), (x^2,\, y^2), (x^2,\, y^2), (x^3,\, y^3), (x^4,\, y^4), (x^5,\, y^5), (x^6,\, y^6), (x^7,\, y^7), (x^8,\, y^8), (x^9,\, y^9), (x^{10},\, y^{10})\right\}$$

$$\hat{m}(x) = \tfrac{1}{10}\left(y^1 + y^2 + y^3 + y^4 + y^5 + y^6 + y^7 + y^8 + y^9 + y^{10}\right)$$

CART

$$x_1 \leq 5$$

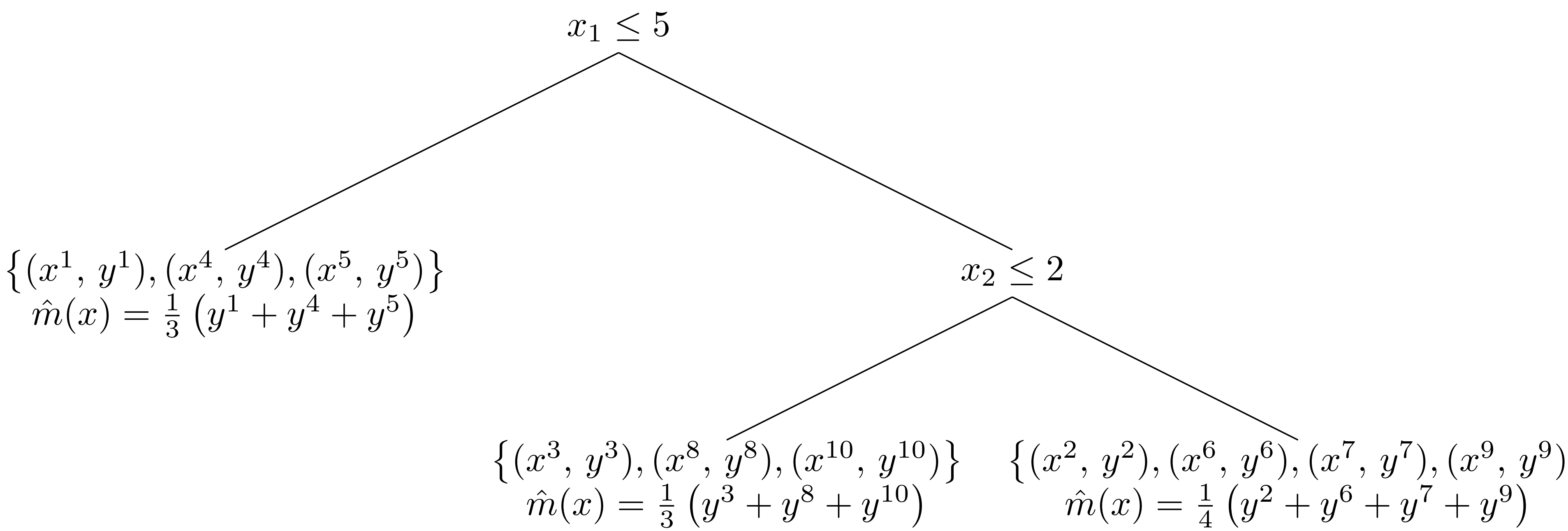
$$\left\{(x^1,\, y^1), (x^4,\, y^4), (x^5,\, y^5)\right\}$$

$$\hat{m}(x) = \frac{1}{3}\left(y^1 + y^4 + y^5\right)$$

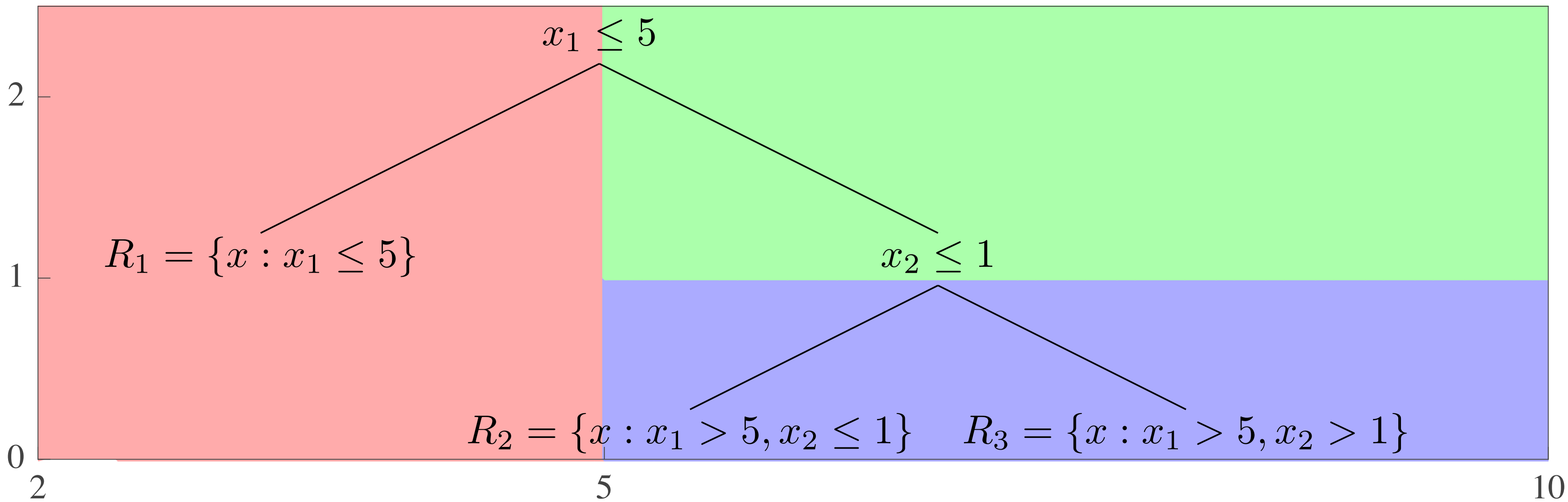
$$\left\{(x^2,\, y^2), (x^3,\, y^3), (x^6,\, y^6), \ldots, (x^{10},\, y^{10})\right\}$$

$$\hat{m}(x) = \frac{1}{7}\left(y^2 + y^3 + y^6 + \cdots + y^{10}\right)$$

CART


$x_1 \leq 5$
$\{(x^1, y^1), (x^4, y^4), (x^5, y^5)\}$
$\hat{m}(x) = \frac{1}{3}\left(y^1 + y^4 + y^5\right)$
$x_2 \leq 2$
$\{(x^3, y^3), (x^8, y^8), (x^{10}, y^{10})\}$
$\hat{m}(x) = \frac{1}{3}\left(y^3 + y^8 + y^{10}\right)$
$\{(x^2, y^2), (x^6, y^6), (x^7, y^7), (x^9, y^9)$
$\hat{m}(x) = \frac{1}{4}\left(y^2 + y^6 + y^7 + y^9\right)$

CART



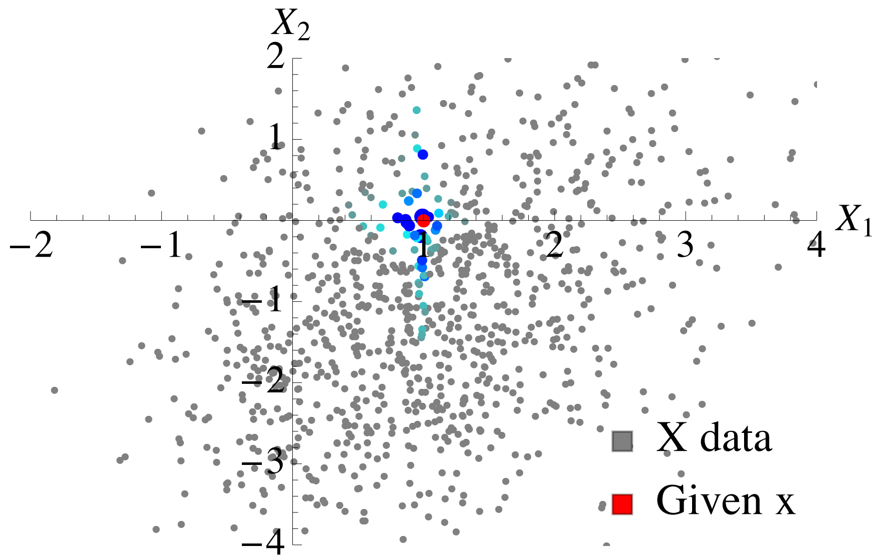
Implied binning rule $R(x) = (j \text{ s.t. } x \in R_j)$

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

Random Forest

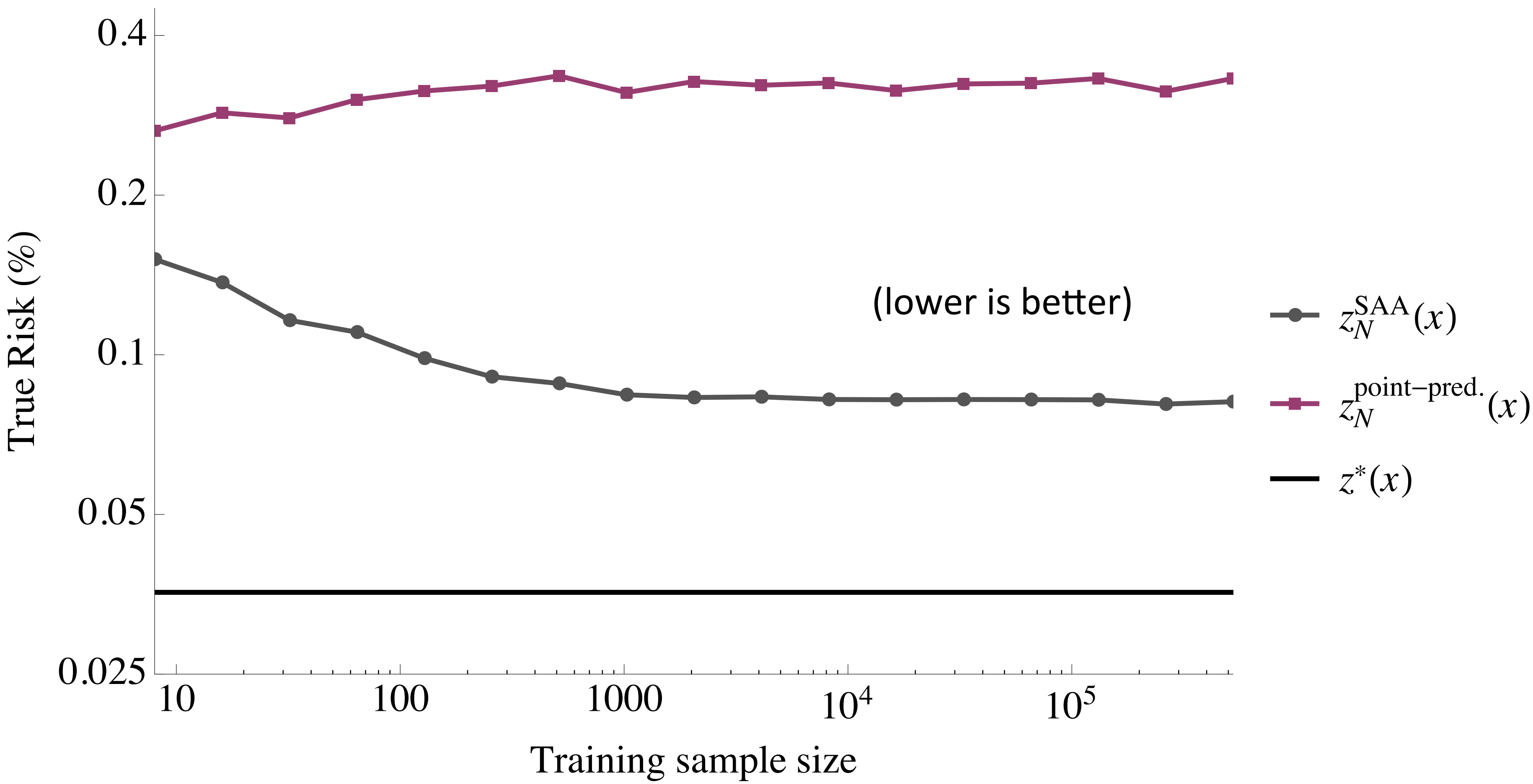
- Train T trees on bootstrapped samples and randomly selected feature subsets
- Get T binning rules $R^t(x) = \left(j \text{ s.t. } x \in R_j^t\right)$

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



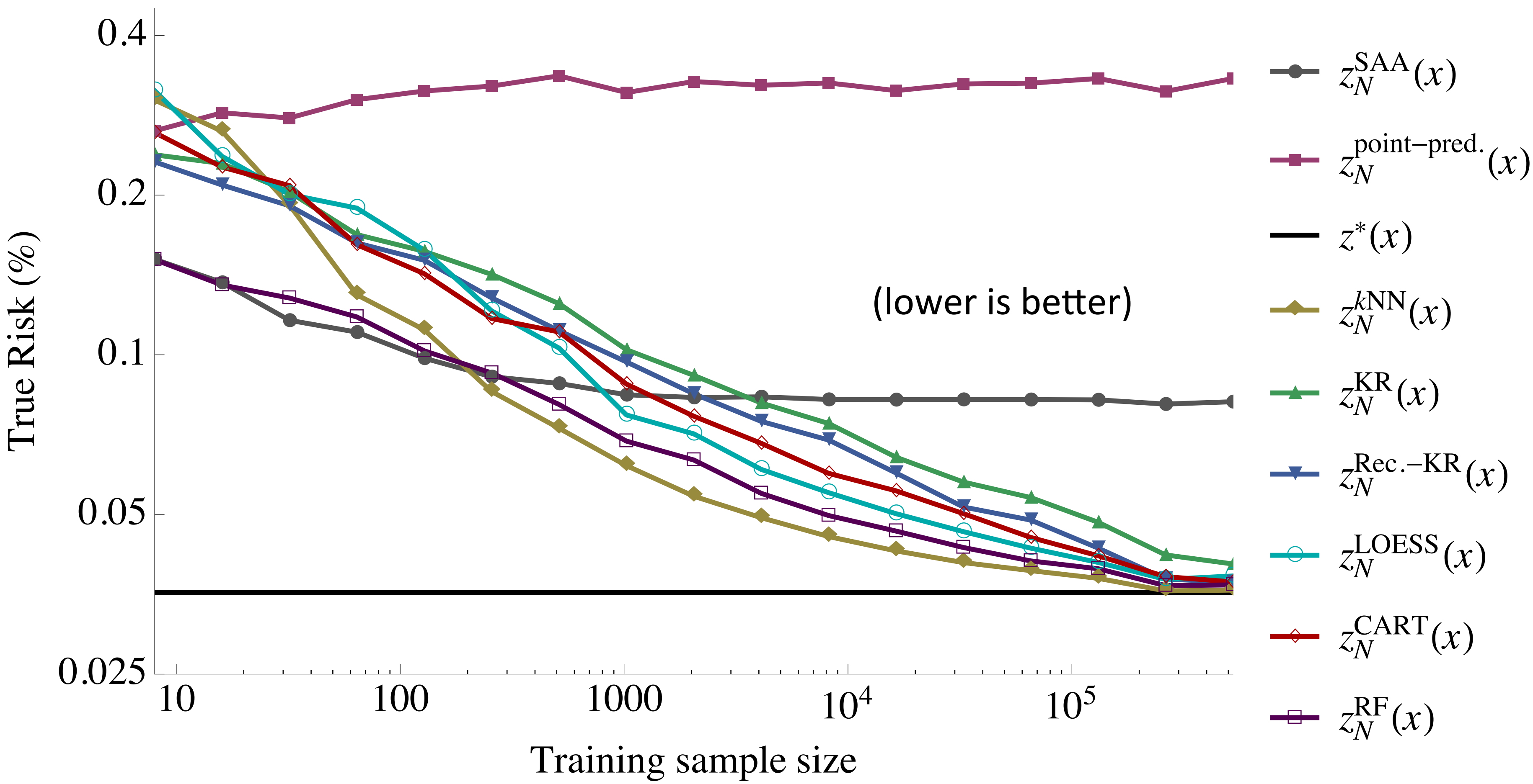
Portfolio example

- Mean-CVaR$_{15\%}$ portfolio allocation with 12 securities
- Observe market factors X correlated with future returns



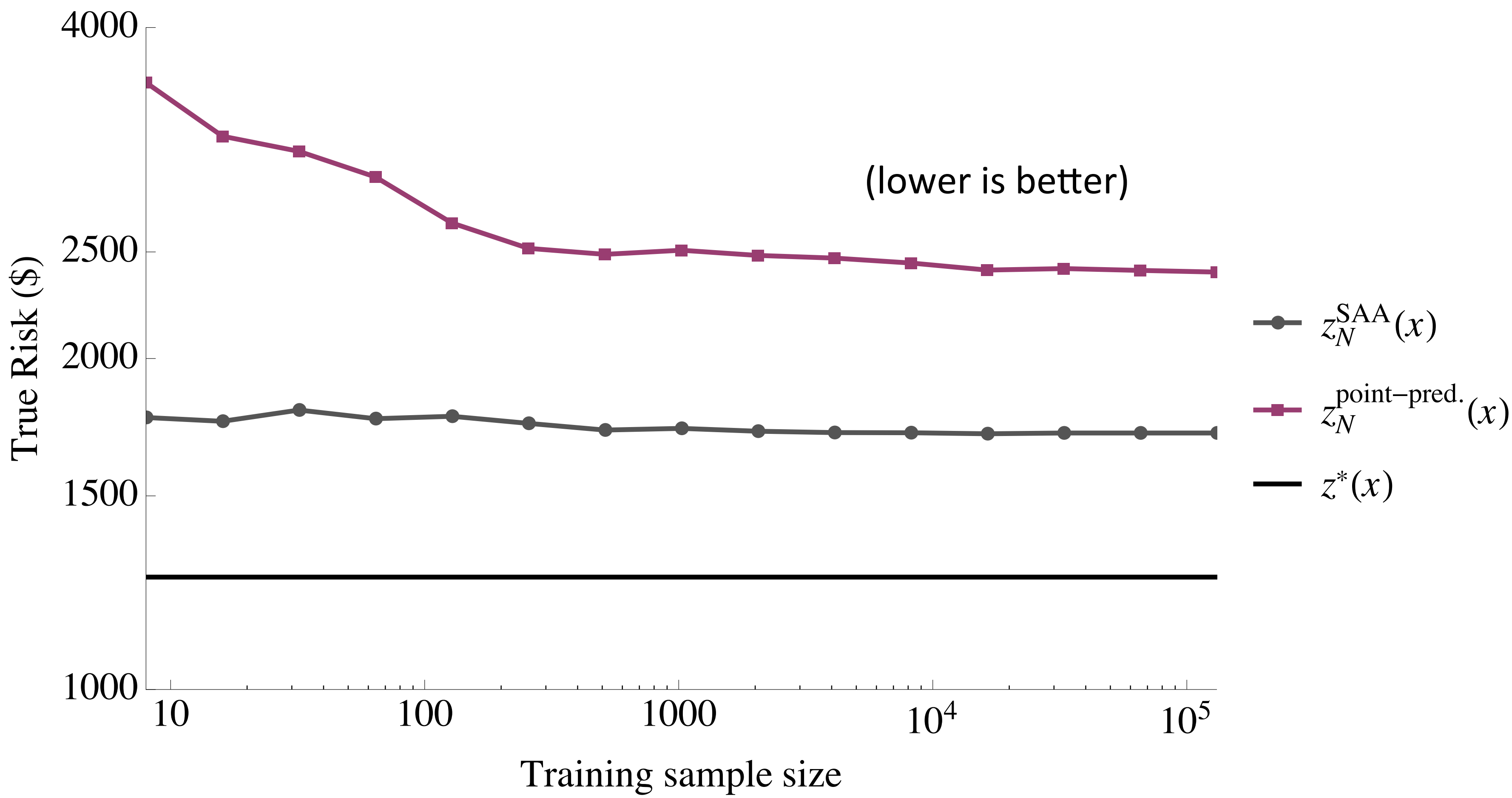
Portfolio example

- Mean-CVaR$_{15\%}$ portfolio allocation with 12 securities
- Observe market factors X correlated with future returns



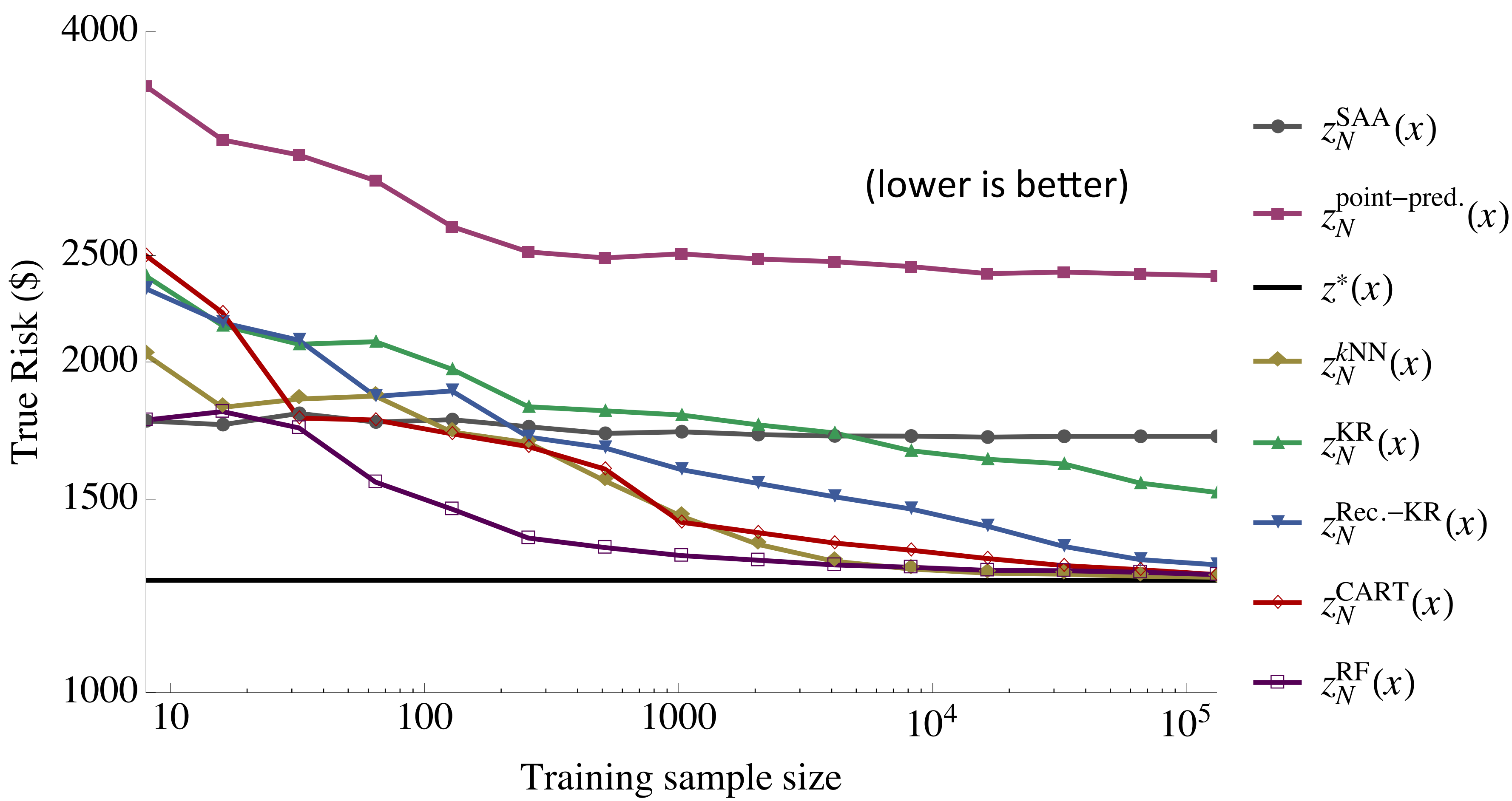
Shipment planning example

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



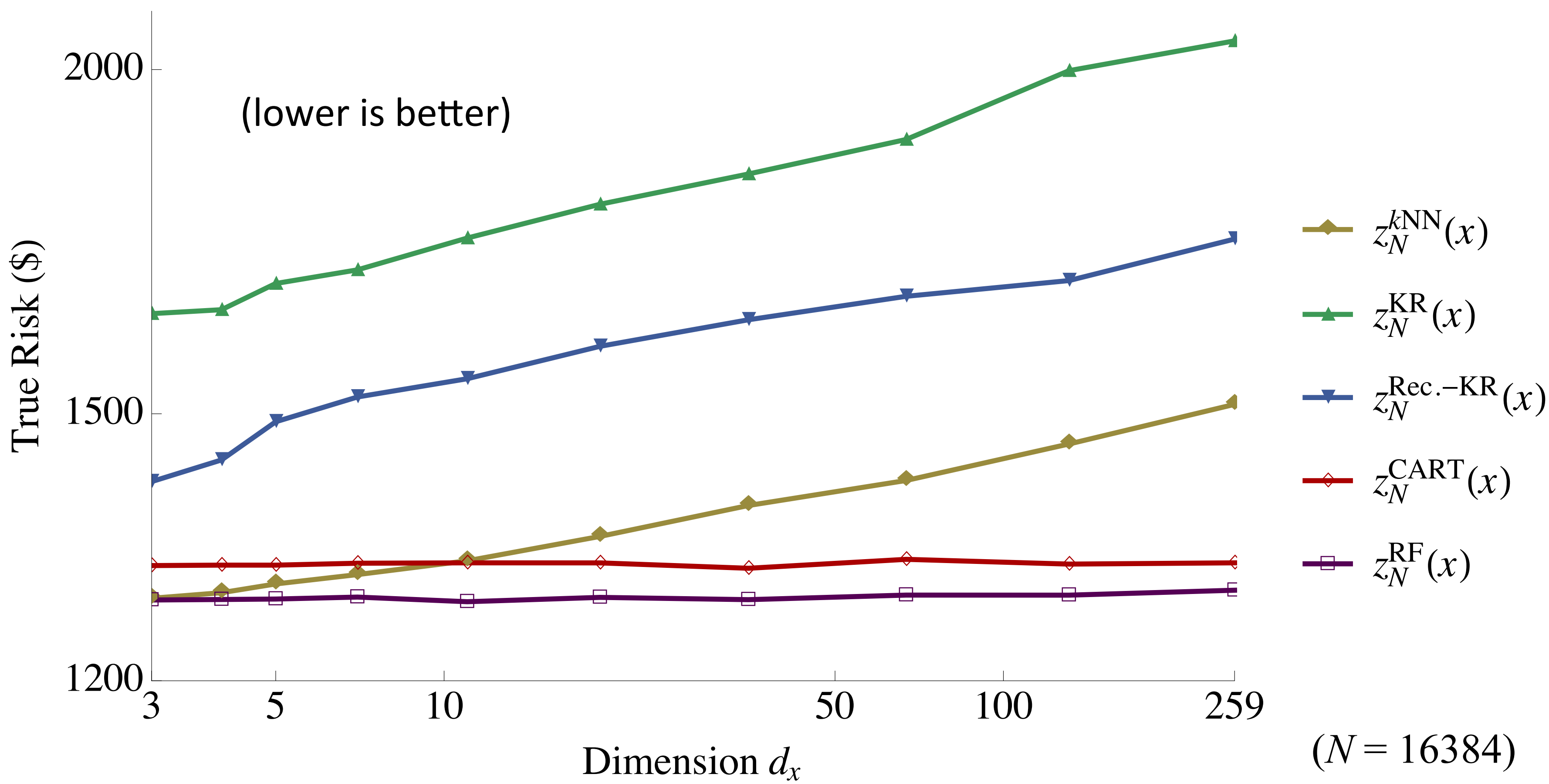
Shipment planning example

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



Shipment planning example

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



Outline

- From Predictive to Prescriptive Analytics
  - A gap in decision making
  - Our approach
  - Asymptotic optimality
  - Coefficient of Prescriptiveness
  - Real world problem

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 bounded, and $c(z;y)$ is convex.

Data collection as a mixing process

- Instead of IID consider a data collection process
$$(x_1,\ y_1), (x_2,\ y_2), \ldots$$
that is a stationary mixing process

- i.e., as the lag ℓ gets bigger,
$$(x_1,\ y_1), \ldots, (x_t,\ y_t)$$
and
$$(x_{t+\ell},\ y_{t+\ell}), (x_{t+\ell+1},\ y_{t+\ell+1}), \ldots$$
are more and more independent.

- Encompasses ARMA, GARCH, Markov processes.
- Can represent more realistic data collection from interdependent weekly demands, stock returns, volume of Google searches on a topic, ...

Asymptotic Optimality: *k*NN

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

Thm: Suppose Assumptions 1, 2, & 3 hold, data collection is IID, and $k = \min\left\{\lceil CN^{\delta} \rceil, N-1\right\}$ with $0 < \delta < 1$.
Then $\hat{z}_N^{k\mathrm{NN}}(x)$ is asymptotically optimal.

Asymptotic Optimality: Parzen Windows

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

Thm: Suppose Assumptions 1, 2, & 3 hold, data collection is mixing, costs satisfy $\mathbb{E}\left[|c(z;Y)|\left(\log|c(z;Y)|\right)_+\right] < \infty$, K is one of given kernels, and $h_N = CN^{-\delta}$, $0 < \delta < 1/d_x$. Then $\hat{z}_N^{\mathrm{KR}}(x)$ is asymptotically optimal.

Asymptotic Optimality: Recursive Parzen Windows

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

Thm: Suppose Assumptions 1, 2, & 3 hold, data collection is mixing, K is one of given kernels,
and $h_i = Ci^{-\delta}$, $0 < \delta < 1/(2d_x)$.
Then $\hat{z}_N^{\text{Rec-KR}}(x)$ is asymptotically optimal.

Asymptotic Optimality: LOESS

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

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

Thm: Suppose Assumptions 1, 2, & 3 hold, data collection is mixing, μ_X abs. continuous, costs bounded $|c(z;y)| \leq g(z)$, and $h_N = CN^{-\delta}$, $0 < \delta < 1/d_x$.

Then $\hat{z}_N^{\text{LOESS}}(x)$ is asymptotically optimal.

Outline

- From Predictive to Prescriptive Analytics
  - A gap in decision making
  - Our approach
  - Asymptotic optimality
  - Coefficient of Prescriptiveness
  - Real world problem

Value of a Prescription

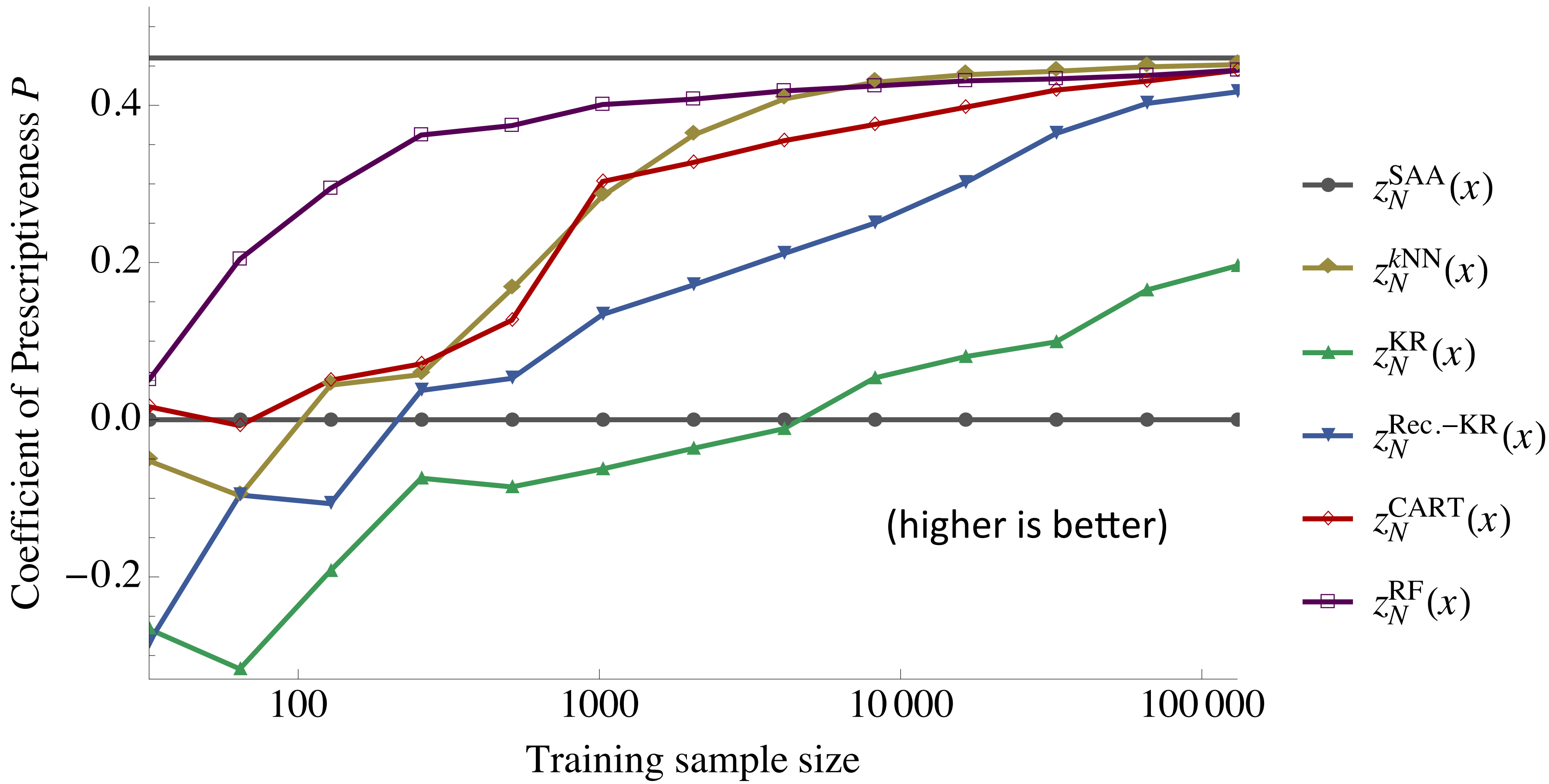
- *Coefficient of Prescriptiveness*

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

- Measures the prescriptive value of X and of the of the prescription trained
- Contrast with R^2.

Shipment planning example

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



Outline

- From Predictive to Prescriptive Analytics
  - A gap in decision making
  - Our approach
  - Asymptotic optimality
  - Coefficient of Prescriptiveness
  - Real world problem

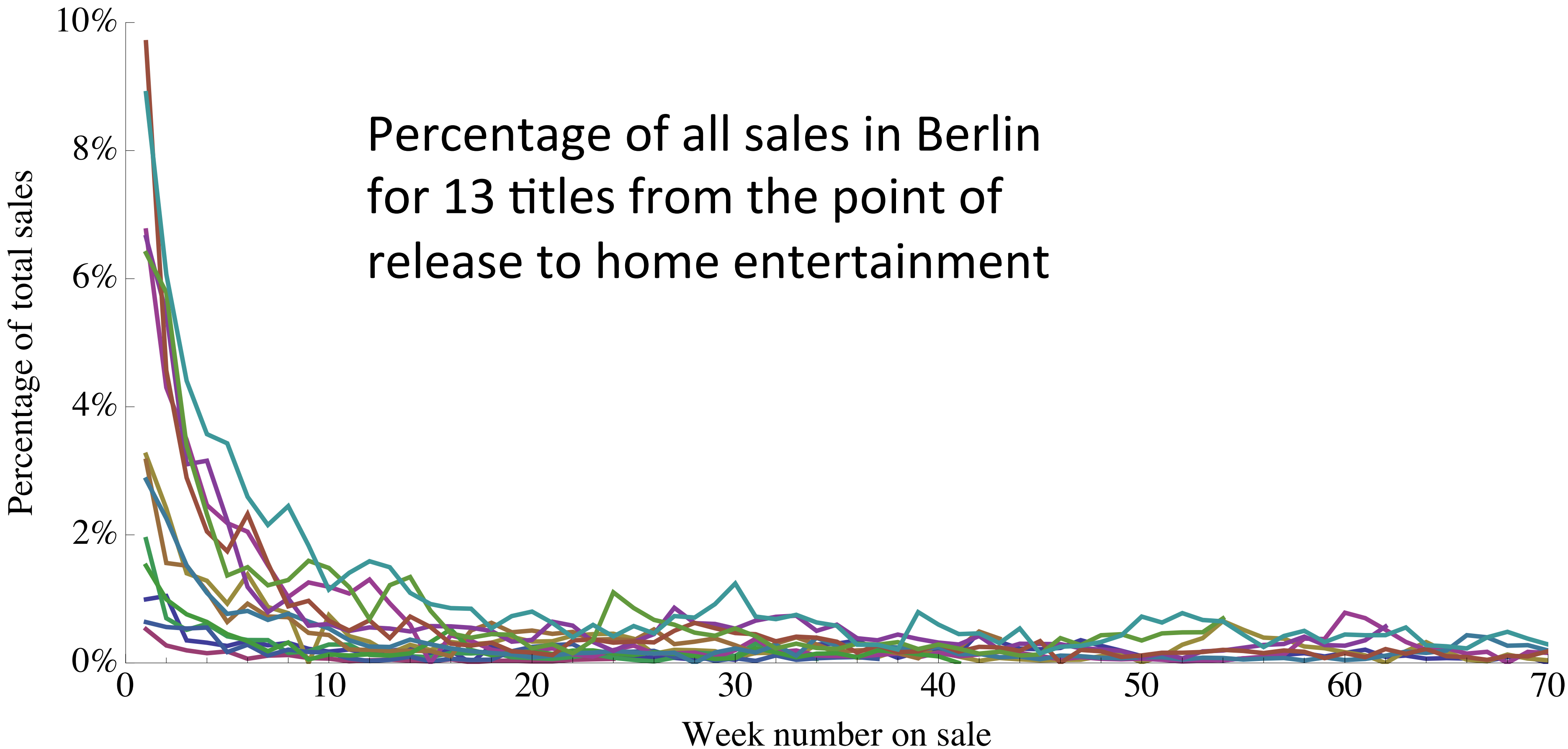
Back to our media distribution application

- Recall: want to maximize number of items sold.
- Focus on video media, Europe
- r index locations, t index periods, j index products.
- Y_j demand for j, z_{trj} order, x_{tr} auxiliary data.

$$
\begin{aligned}
\max \quad & \mathbb{E}\left[\sum_{j=1}^{d} \min\{Y_j,\, z_{trj}\} \,\middle|\, X = x_{tr}\right] \\
\text{s.t.} \quad & \sum_{j=1}^{d} z_{trj} \le K_r \\
& z_{trj} \ge 0 \qquad\qquad \forall j = 1, \ldots, d
\end{aligned}
$$

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***.

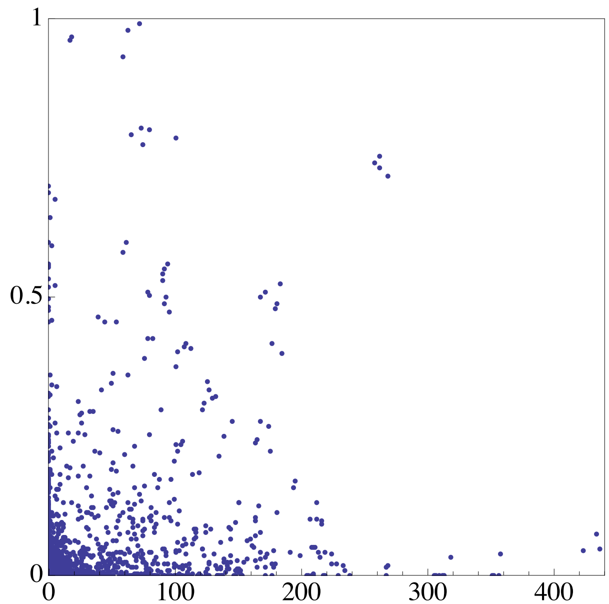
Internal Company Data

- 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

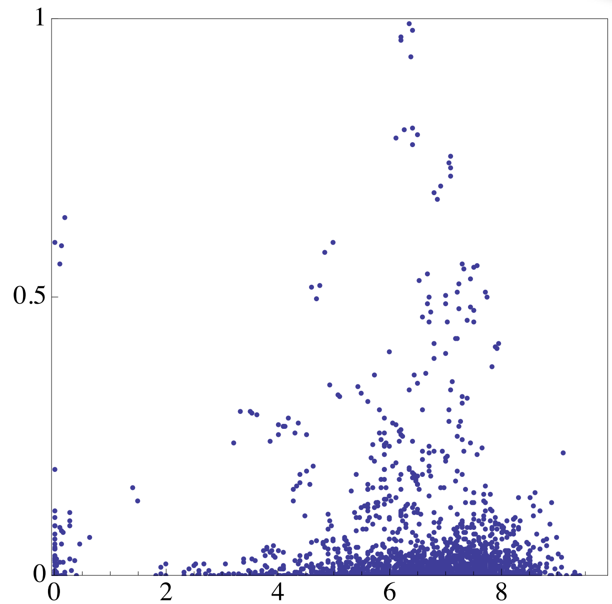
Beyond internal company data: Harvesting public data (more X)

- **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**

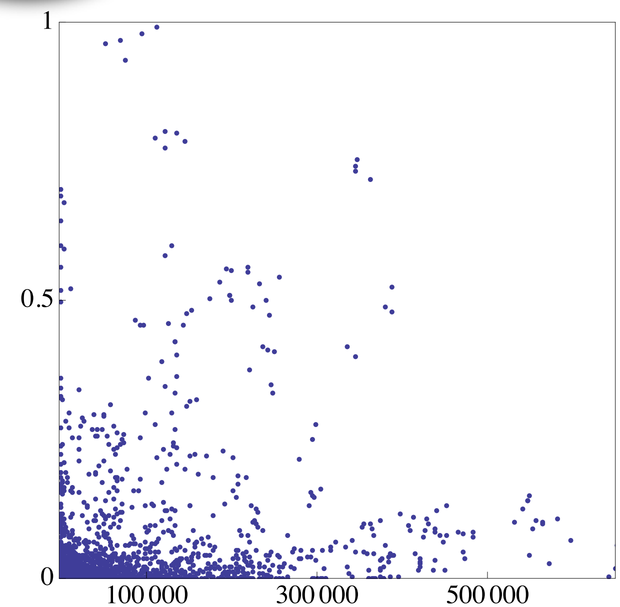
Beyond internal company data: Harvesting public data (more X)



Box office gross
ρ = 0.32

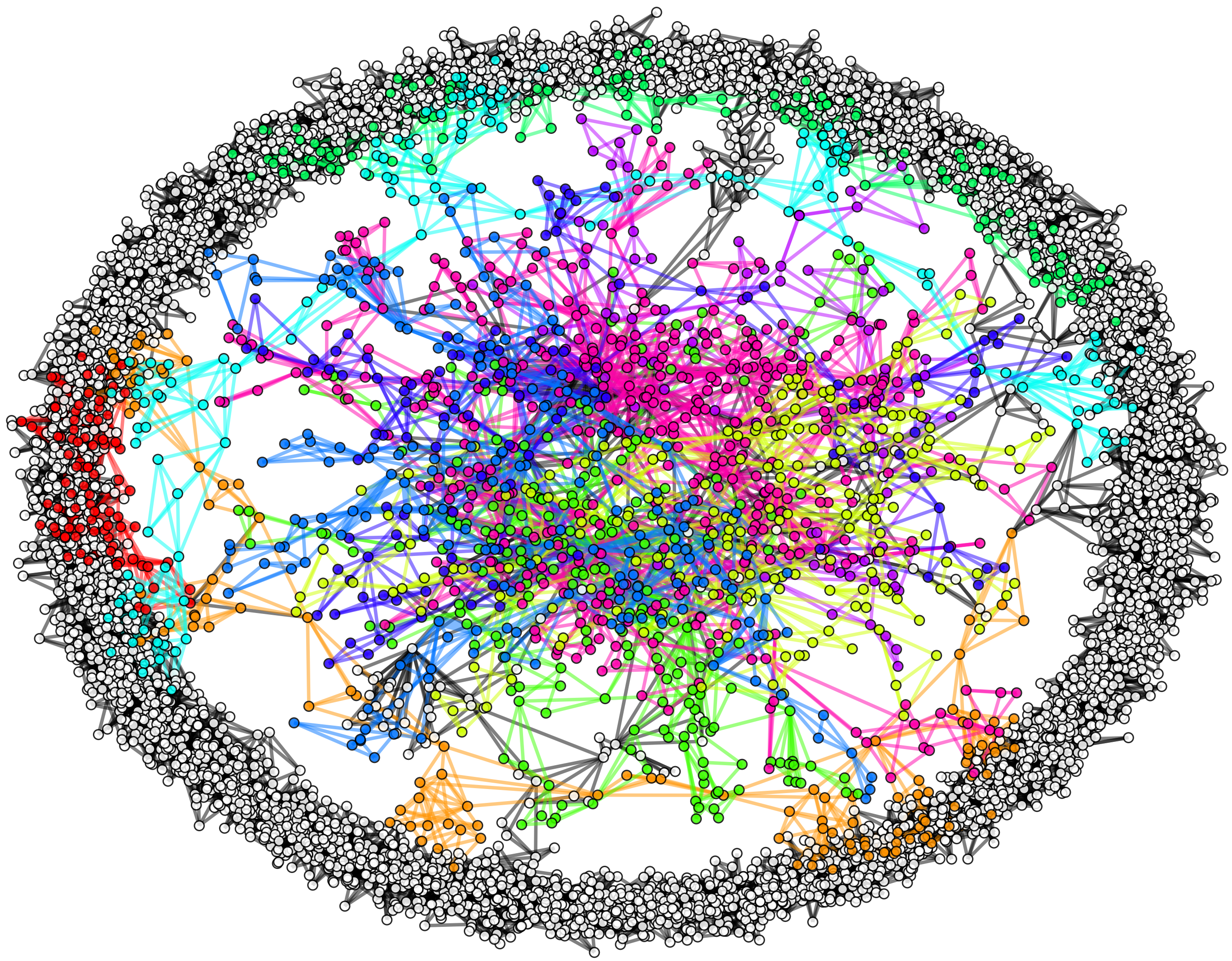


IMDb rating
ρ = 0.02

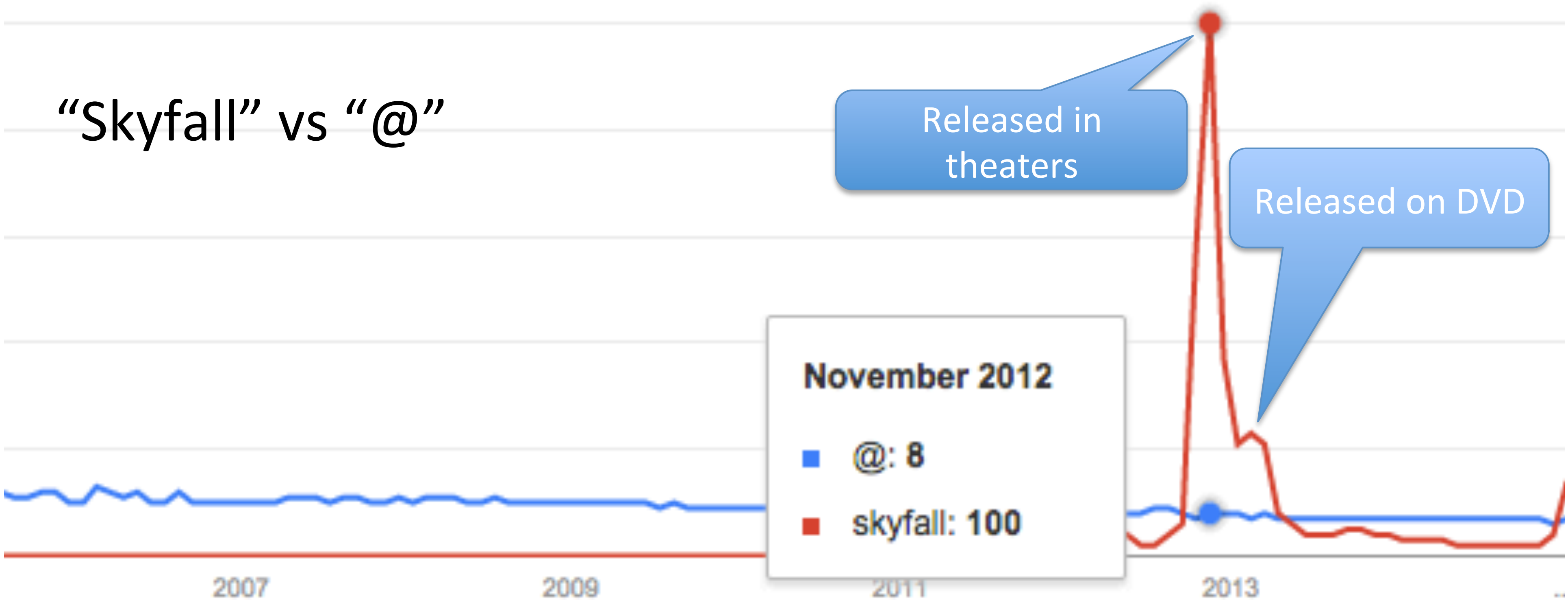


Number user votes
ρ = 0.25

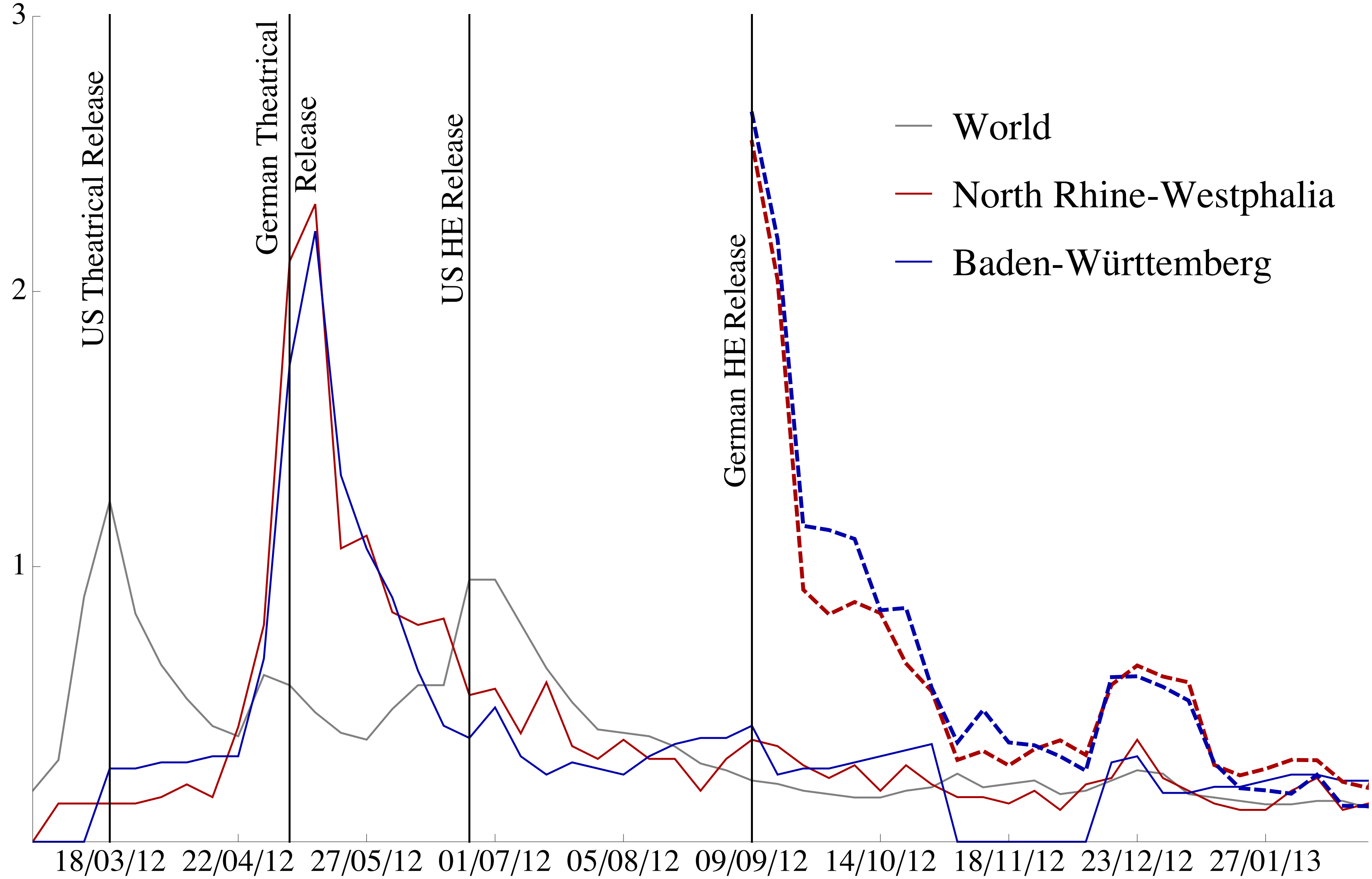
Beyond internal company data: Harvesting public data (more X)



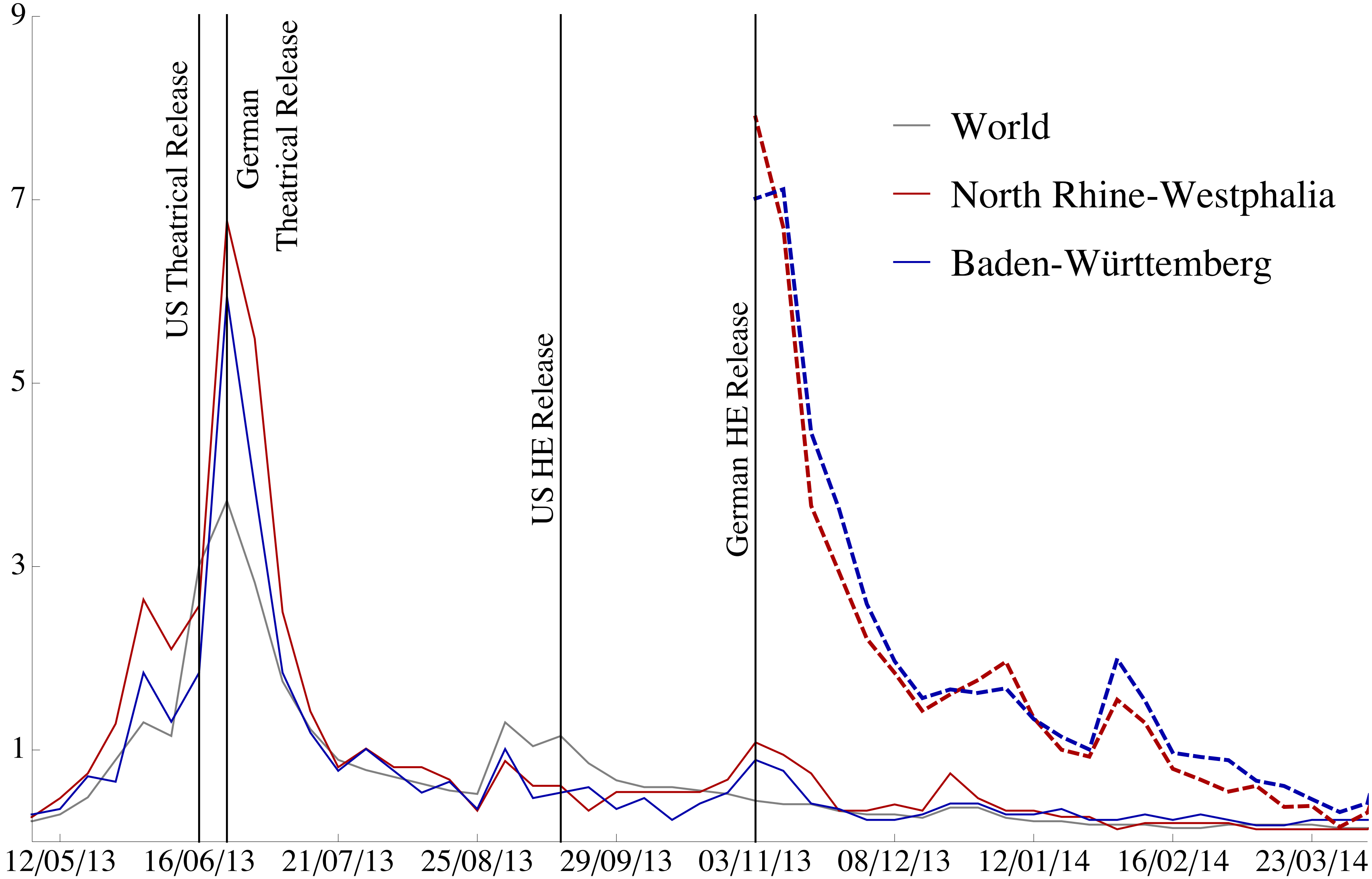
Beyond internal company data: Harvesting public data (more X)



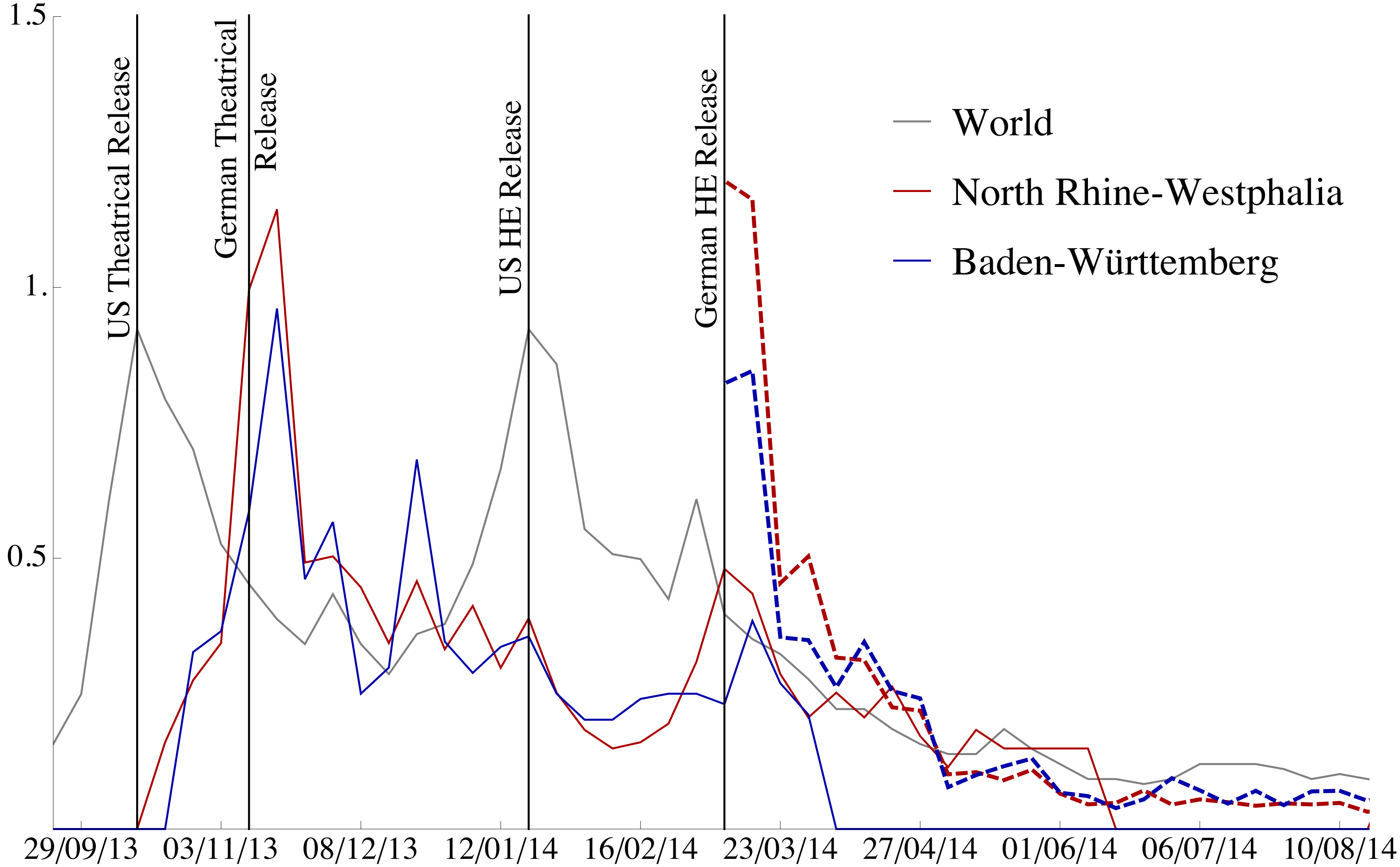
Beyond internal company data: Harvesting public data (more X)



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

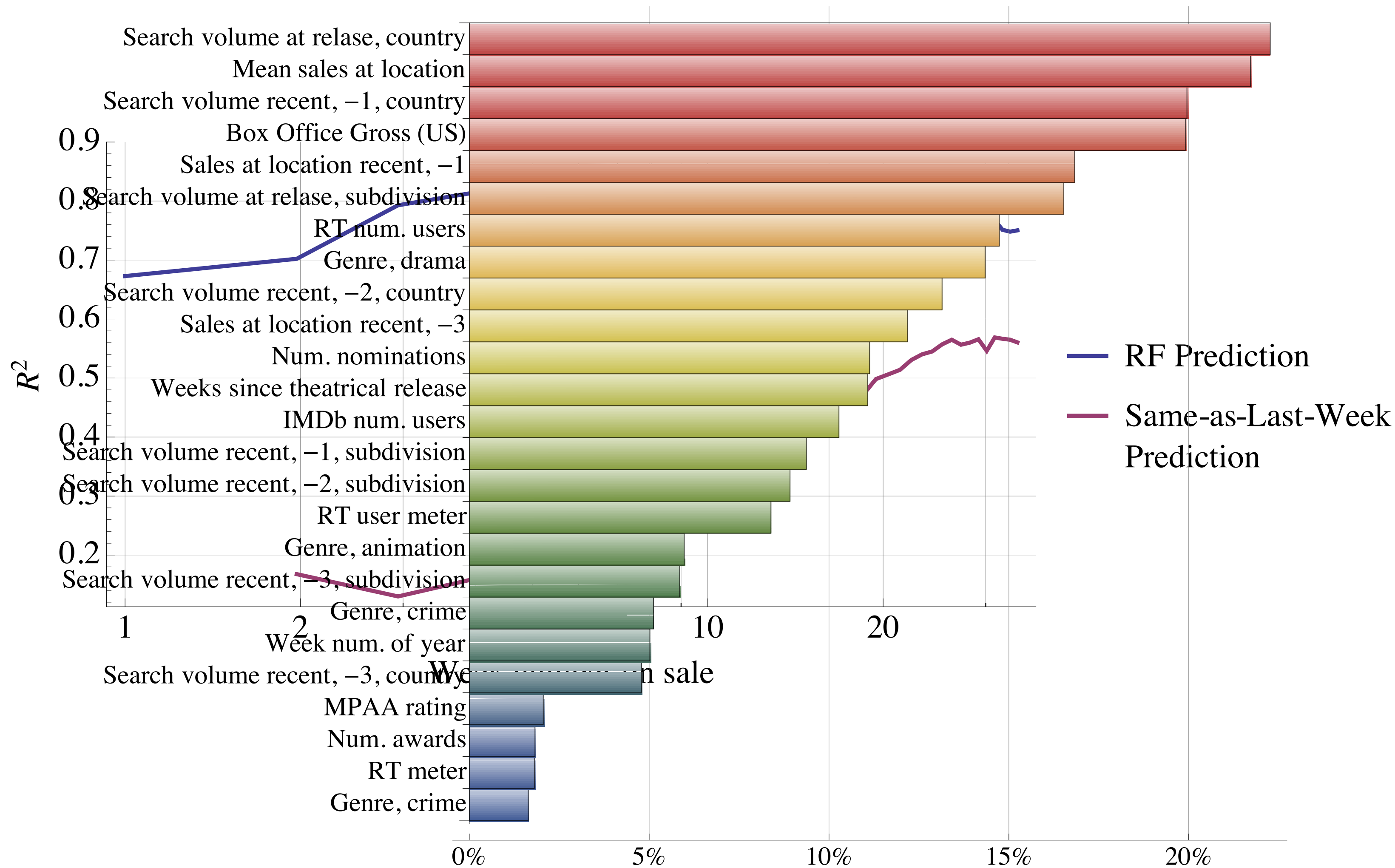


Beyond internal company data: Harvesting public data (more X)



Predicting Demand

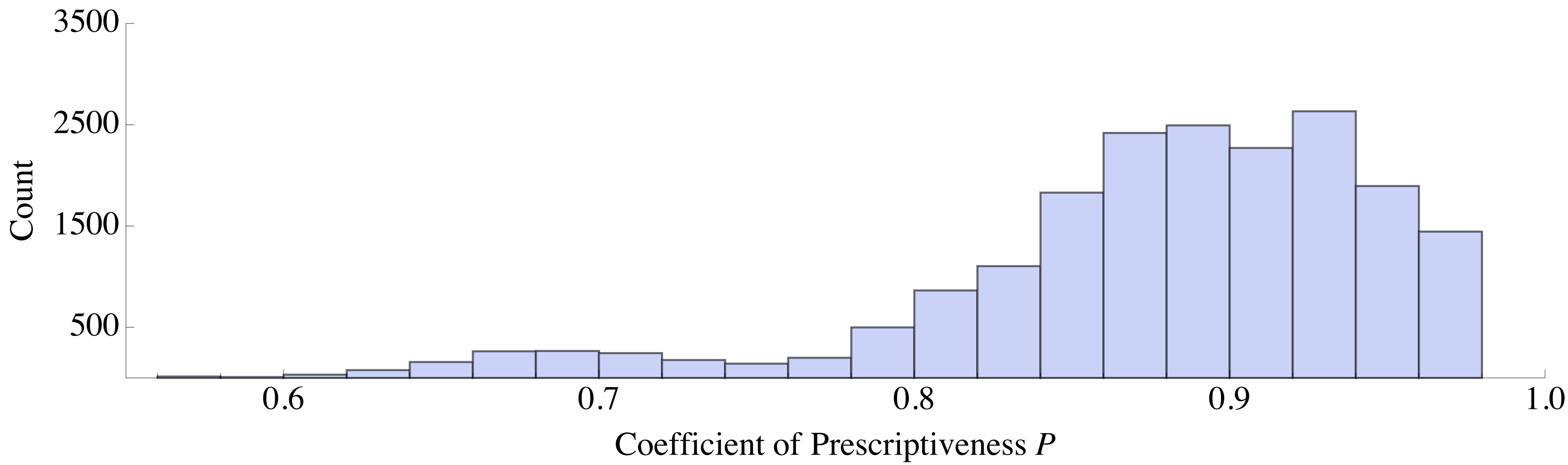
- Random forest regressor
- ***New*** titles: out-of-sample R^2 = 0.67



Prescribing Order Quantities

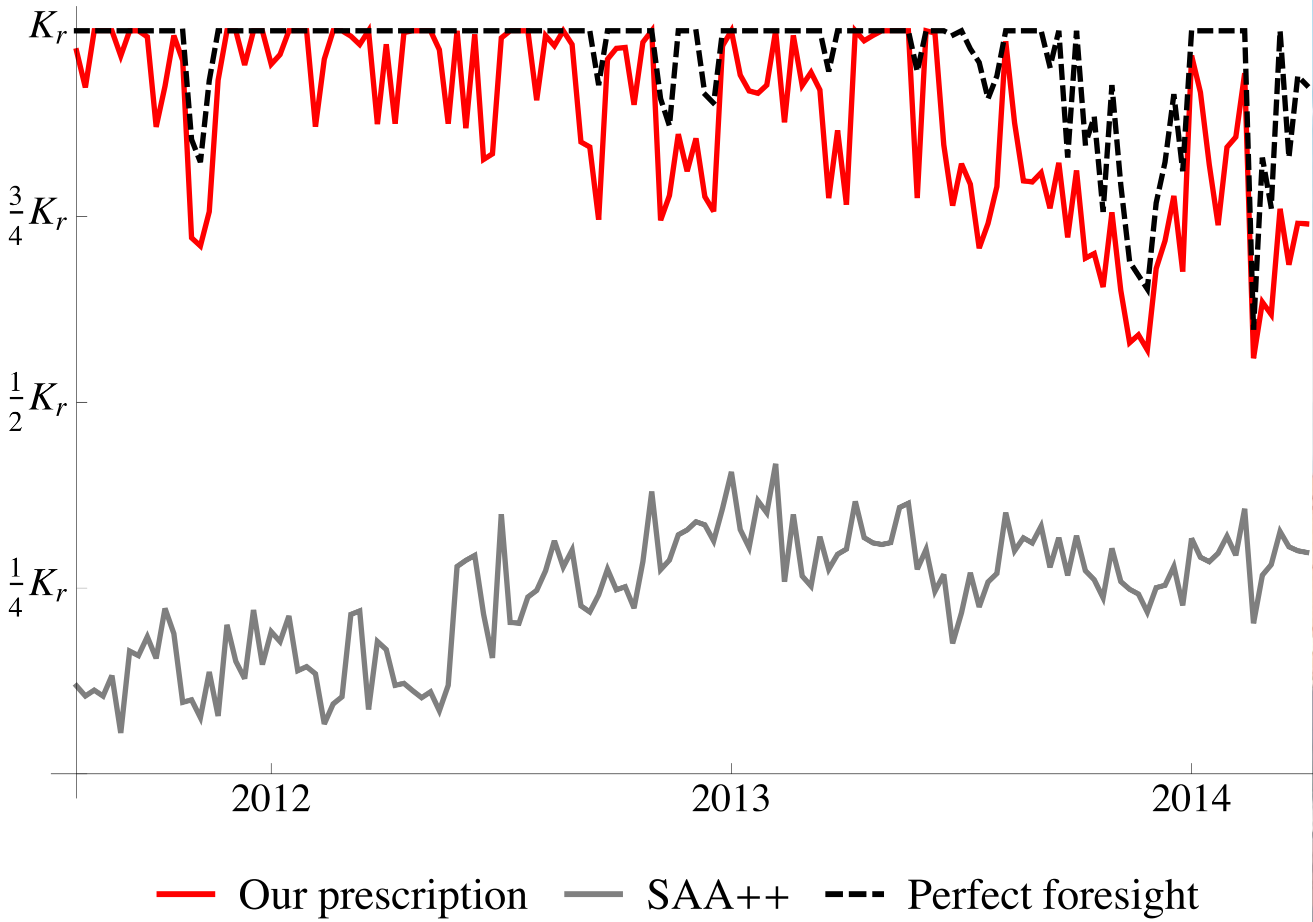
- Construct a predictive prescription based on our random forest…

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

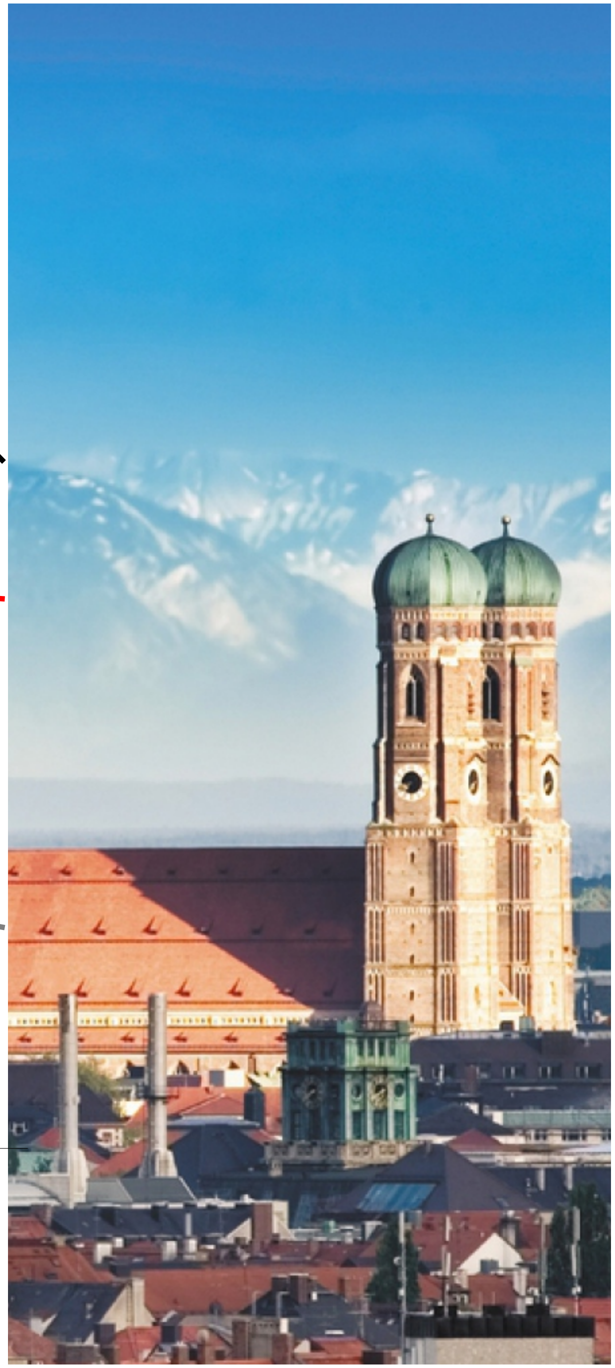


Munich

$$P = 0.89$$

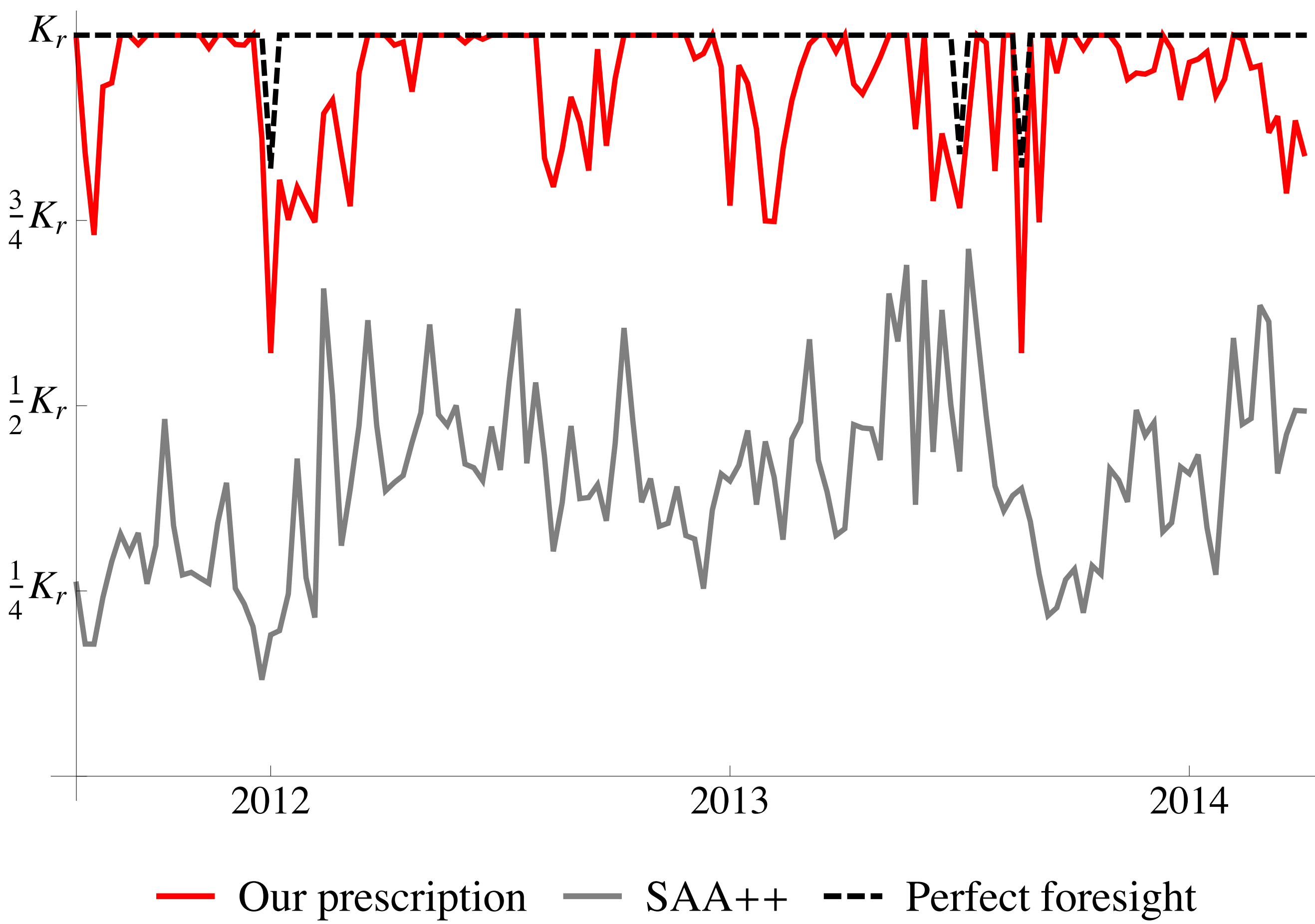


K_r
$\frac{3}{4}K_r$
$\frac{1}{2}K_r$
$\frac{1}{4}K_r$
2012
2013
2014
Our prescription
SAA++
Perfect foresight

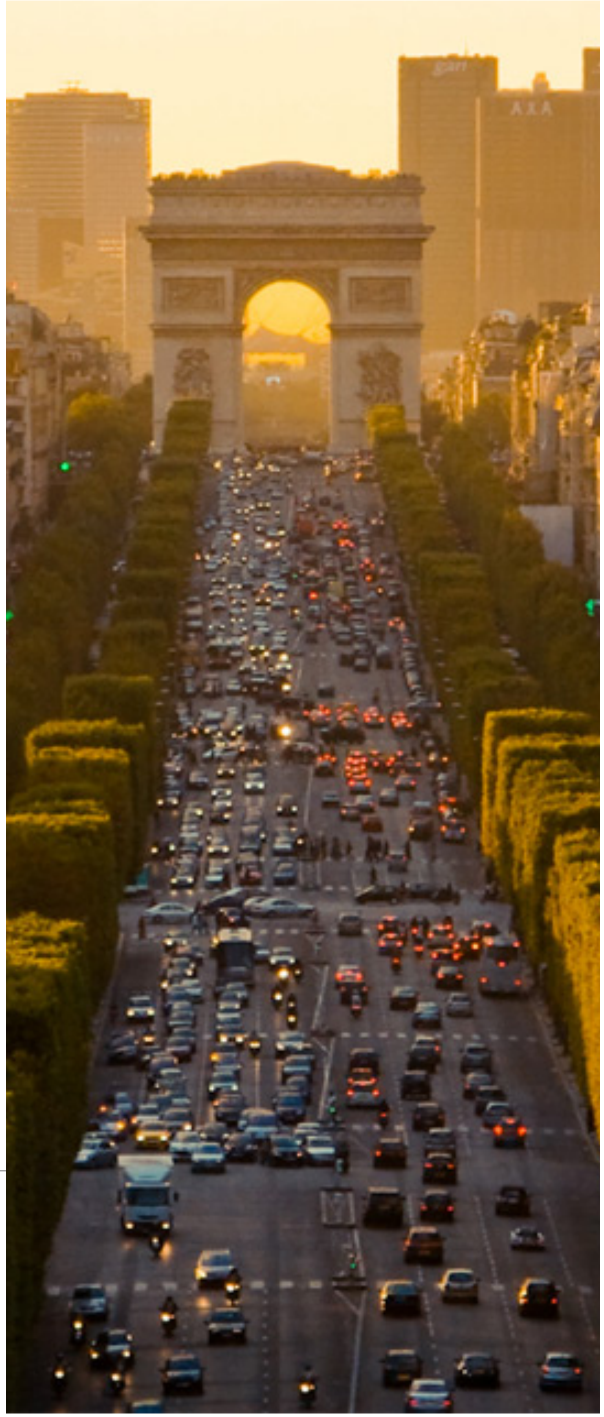


Paris

$$P = 0.90$$

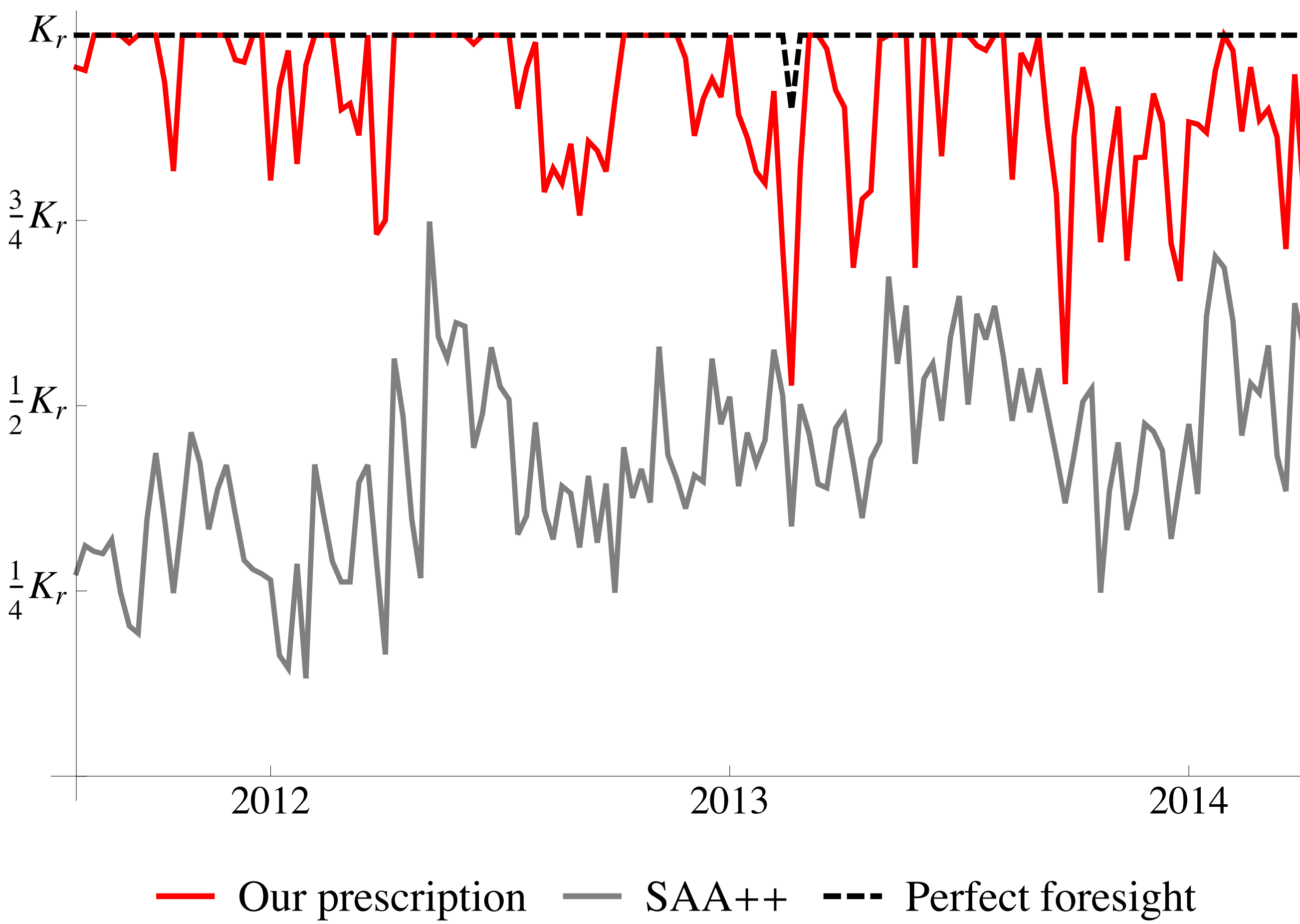


K_r
$\frac{3}{4}K_r$
$\frac{1}{2}K_r$
$\frac{1}{4}K_r$
2012
2013
2014
Our prescription
SAA++
Perfect foresight



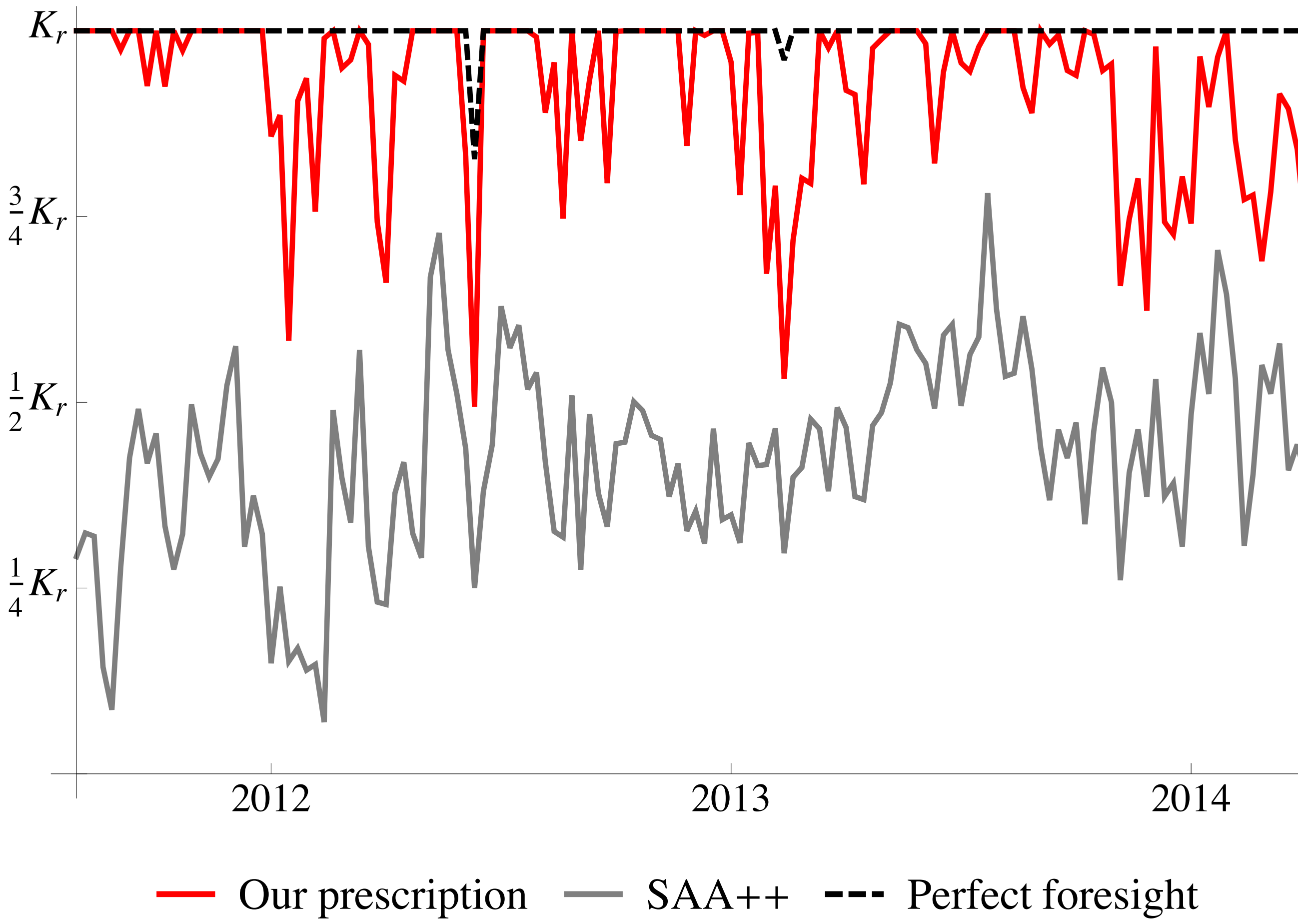
Waterloo

$$P = 0.85$$



The Hague

$$P = 0.86$$



Conclusions

- **A new framework**
  - Unifies ML and OR/MS
  - General purpose
- **Theory**
  - Computational tractability
  - Asymptotic optimality
- **Performance metric**
  - Coefficient of prescriptiveness
- **Practice**
  - Material Improvement for A Global Fortune 100 company.

Education

- A new class at the Master level that starts with Data: Analytics Edge

- A new class at the PhD level that aspires to go
  From Data to predictions to prescriptions.