MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS

# Performative Prediction

## An overview

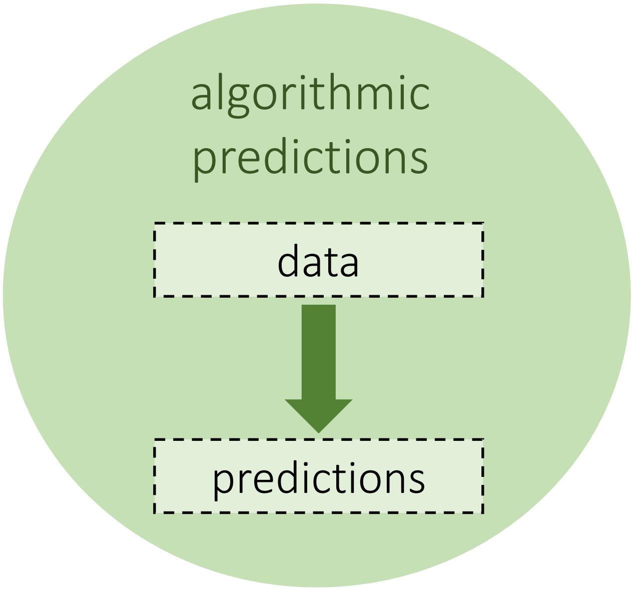

Celestine Mendler-Dünner

*Max Planck Institute for Intelligent Systems*

*Tübingen, Germany*

# Traditional view on machine learning

Supervised learning:
*static, isolated*



# Traditional view on machine learning

kaggle

**Competitions**

criteo

**Display Advertising Challenge**
Predict click-through rates on display ads
Research · 717 Teams · 9y ago

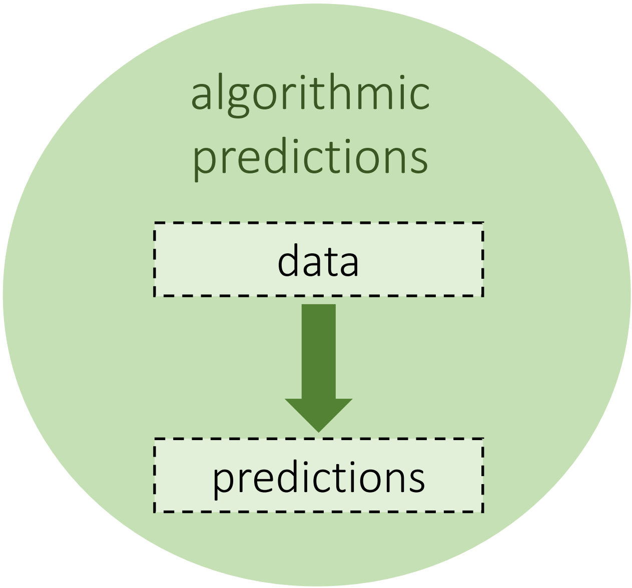
Zillow

**Zillow Prize: Zillow's Home Value Prediction (Zestimate)**
Can you improve the algorithm that changed the world of real estate?
Featured · 3770 Teams · 5 years ago

**Home Credit Default Risk**
Can you predict how capable each applicant is of repaying a loan?
Featured · 7176 Teams · 5 years ago

IDB

**Costa Rican Household Poverty Level Prediction**
Can you identify which households have the highest need for social welfare assistance?
Playground · Code Competition · 616 Teams · 5 years ago

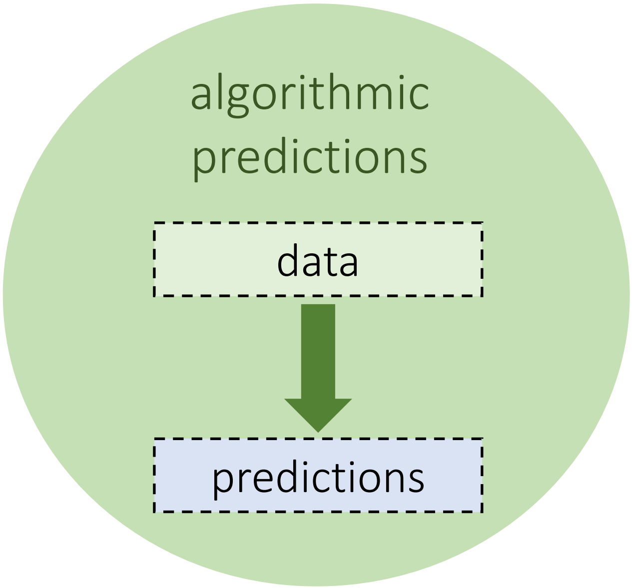



Goal: Fit patterns in static dataset

# Machine learning in societal systems

## Machine learning in practice

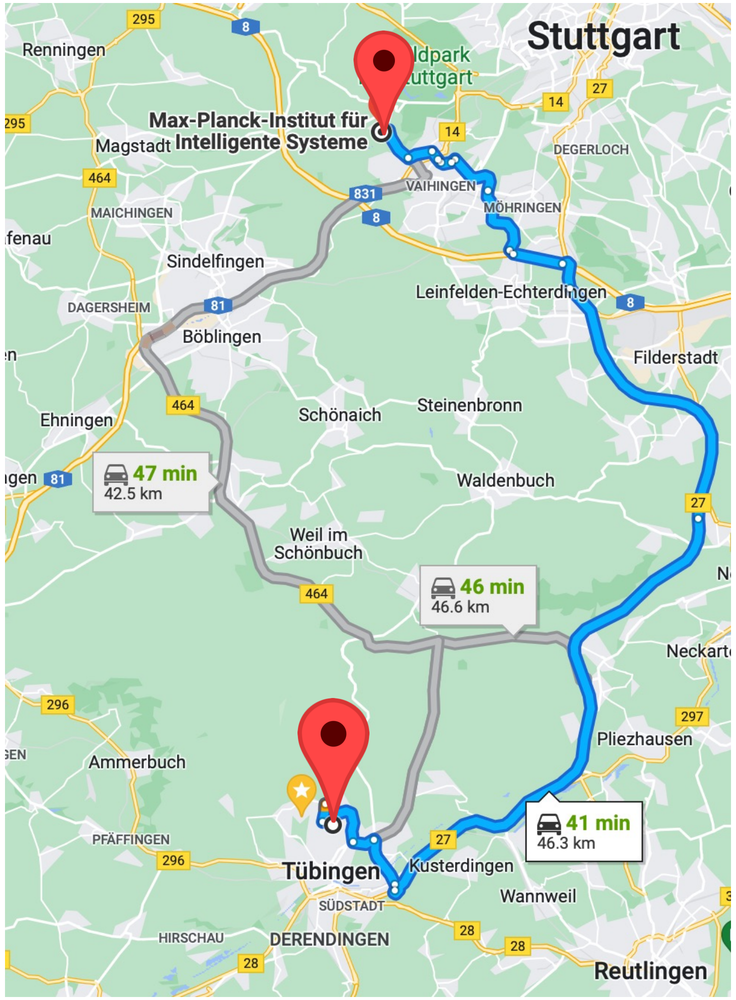
- Click through rate prediction informs targeted advertising
- Zillow’s Zestimate is released to inform buyers
- Credit risk prediction is used to determine interest rates
- Poverty index scores are used to allocate resources

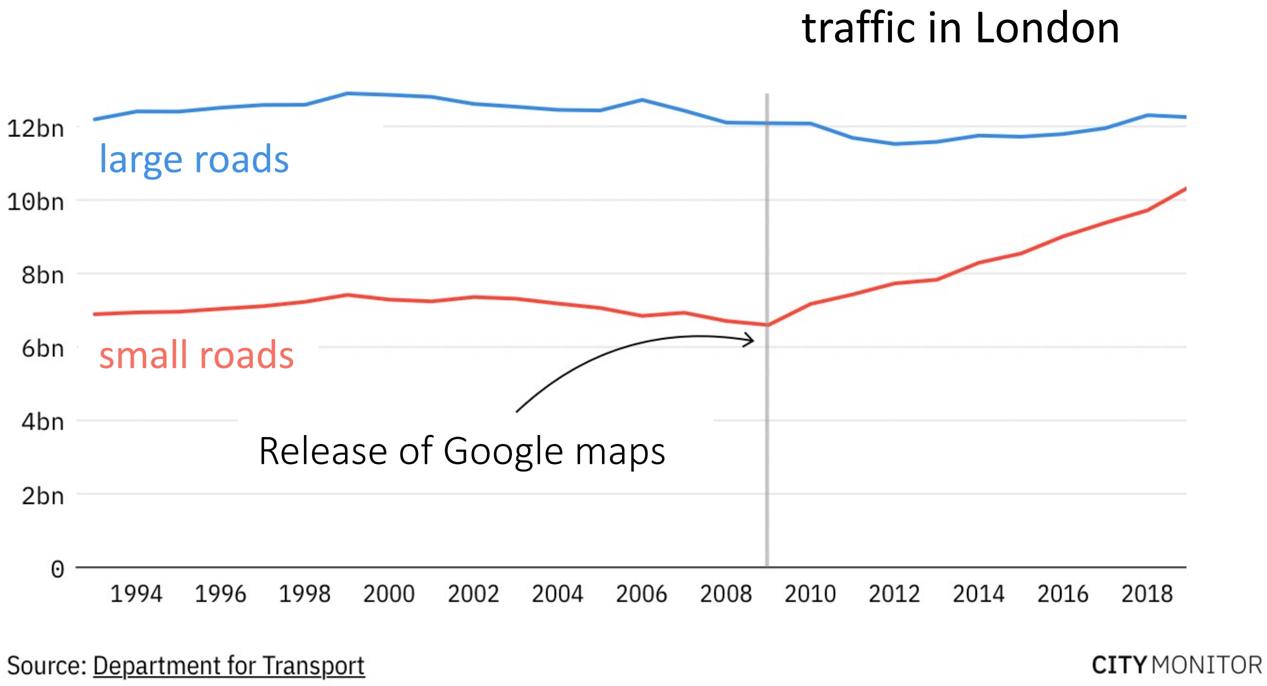


# Machine learning in societal systems


predictions impact people

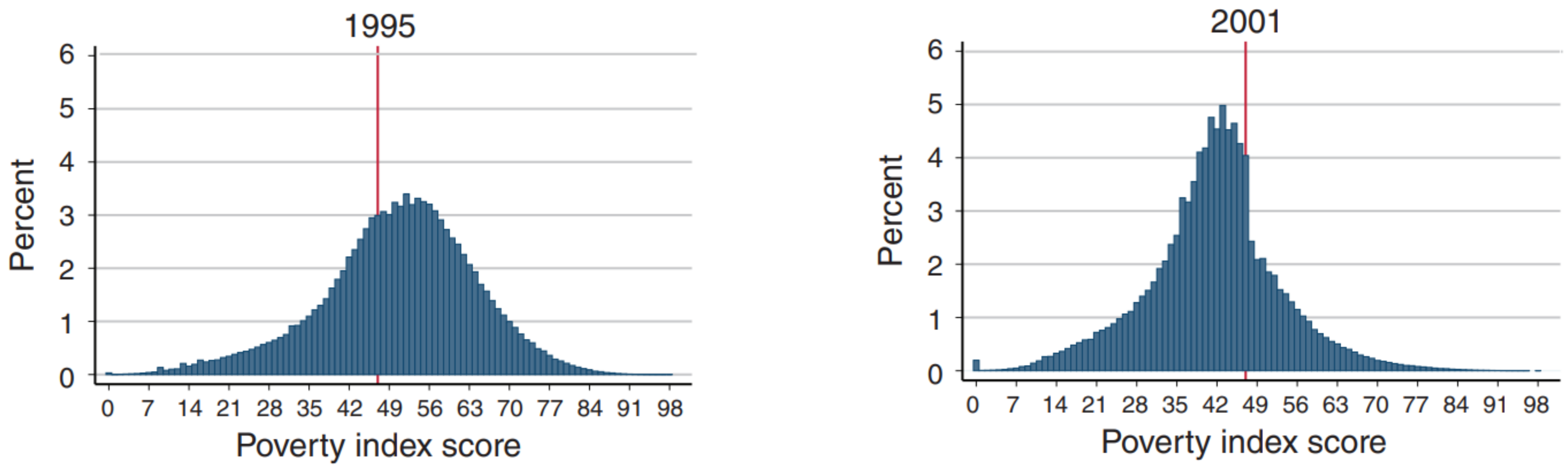
# Predictions inform decisions



# Predictions cause change in people's behavior

Poverty index score
used as targeting instrument




Eligibility for social welfare program in Colombia
*Camacho & Conover, American Economic Journal, 2011*

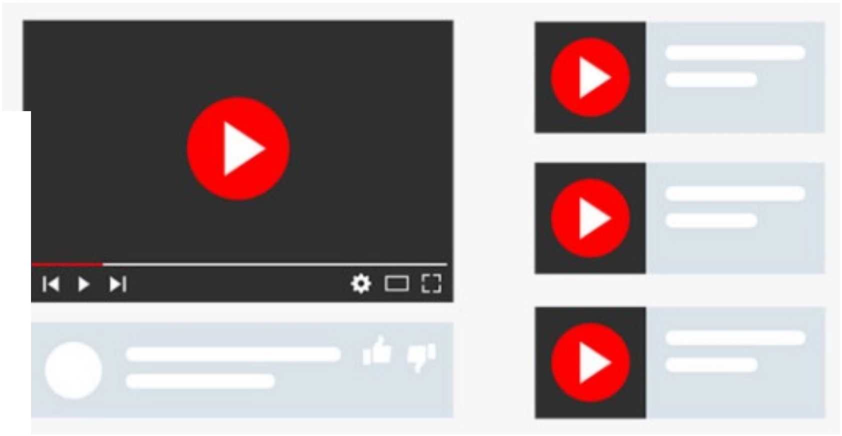
# Predictions shape markets

*“Option pricing theory—a “crown jewel” of neoclassical economics—succeeded empirically not because it discovered preexisting price patterns but because it pushed the market to conform to its predictions […].”*

MacKenzie & Millo, American Journal of Sociology, 2003

# Predictions mediate our everyday lives

… moderate public discourse
… redirect attention
… shape preferences



# Lessons from economics

## Why it is a bad idea to ignore causal effects of predictions

Grunberg, Modigliani (1954)
"The predictability of social events"
Private predictions ≠ public predictions

Goodhardt's law (1975):
"any statistical regularity will tend to collapse once pressure it put upon it for control purposes"

Lucas' critique (1976):
*Macroeconomic policy can disrupt the statistical patterns motivating the policy*

THE JOURNAL OF
POLITICAL ECONOMY

Volume LXII DECEMBER 1954 Number 6

THE PREDICTABILITY OF SOCIAL EVENTS[1]

EMILE GRUNBERG AND FRANCO MODIGLIANI[2]
Carnegie Institute of Technology

I. THE PROBLEM

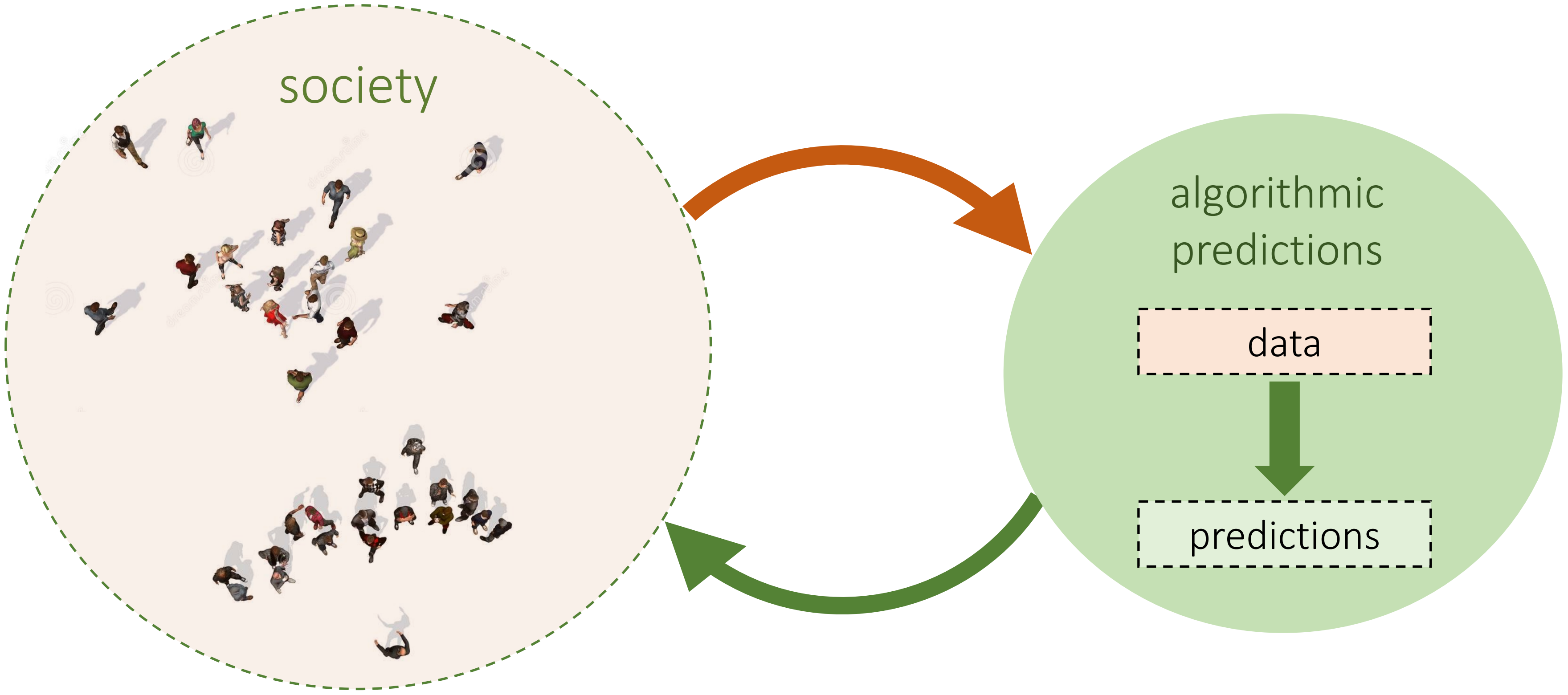
THE fact that human beings react to the expectations of future events seems to create difficulties for the social sciences unknown to the

of this paper is to investigate the validity of this claim. Since it is specifically concerned with the problem raised by the agents' reaction to a published prediction and not with the broader problem

# Predictions have causal powers

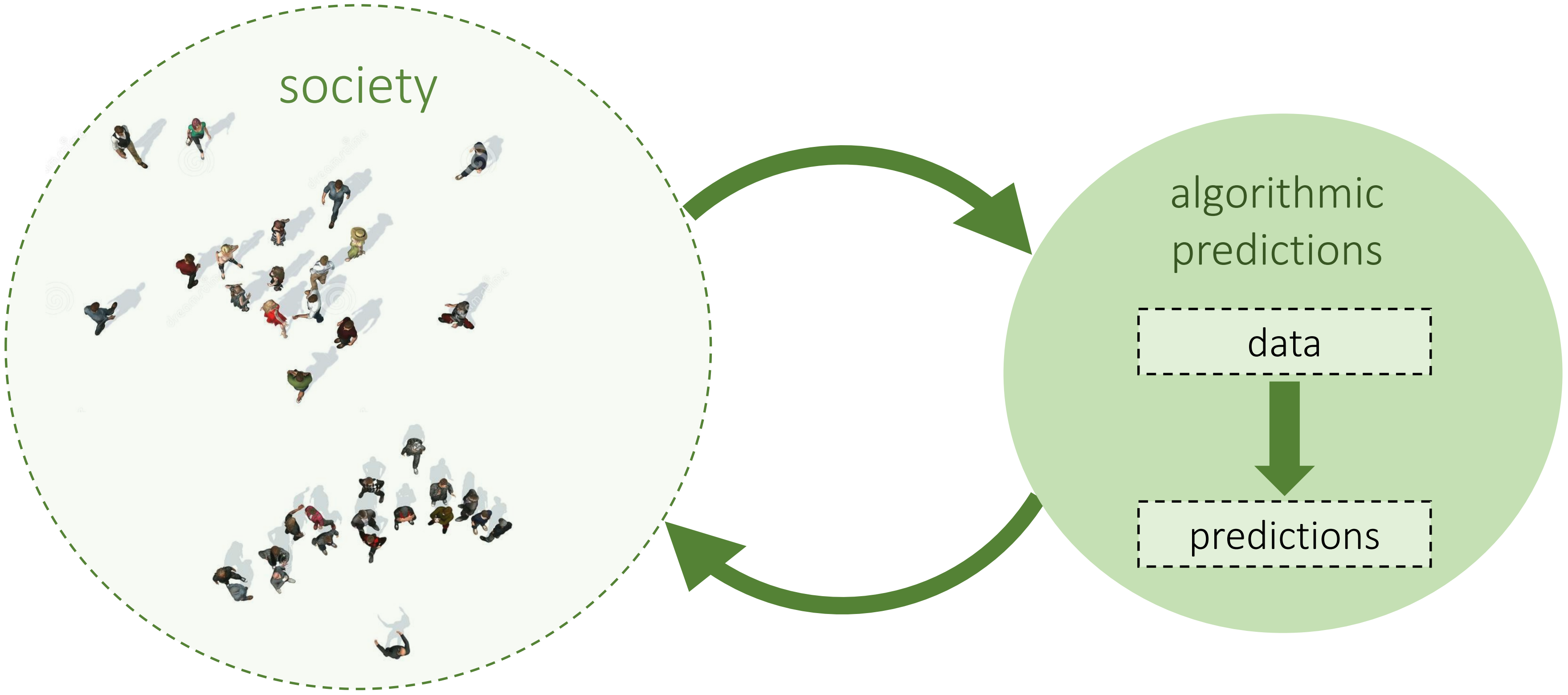

Why does it matter for machine learning?

# Predictions have causal powers




Why does it matter for machine learning?

# Predictions have causal powers




we call predictions performative when they impact the population they aim to predict

# This talk

- Performative prediction: A general conceptual framework to reason about the causal powers of predictions in machine learning.
  - Solution concepts
  - Emerging optimization results
  - Economic models
- Performative power: Measuring power in digital economies
  - On the difference between learning and steering in optimization
  - Connections to anti trust and market regulation

[PZMH20]

# Framework of Performative Prediction

$$\boxed{\mathrm{Risk}(\theta, D) = \mathrm{E}_{z \sim D}\ [\,\ell(z;\theta)\,]}$$

**Supervised learning:**

- Represent population as a fixed distribution $D$ over data instances $\mathrm{Z} = (\overset{\text{feature}}{X}, \overset{\text{label}}{Y})$
- Represent the predictive model by a parameter vector $\theta \in \Theta$
- Find model that minimizes risk

$$\mathrm{R}(\theta) = \mathrm{Risk}(\theta, D)$$

**Performativity thesis:**

The data distribution $D$ depends on the model $\theta$ that is being deployed.

[PZMH20]

# Framework of Performative Prediction

$$\mathrm{Risk}(\theta, D) = \mathrm{E}_{z\sim D}\ [\,\ell(z;\theta)\,]$$

**Distribution map:** let $D(\theta)$ denote the distribution over data instances $Z = (X, Y)$ induced by a model $\theta \in \Theta$

- 'Macro-level' description of the distribution shift

**Performative risk:** Risk of a model measured after deployment

$$\mathrm{PR}(\theta) = \mathrm{Risk}(\theta, D(\theta))$$

model-dependent distribution

[PZMH20]

# Solution Concepts in Performative Prediction

we take distribution shift into account

Performative optimality: $\theta_{\mathrm{PO}} \in \operatorname{argmin}_\theta \mathrm{PR}(\theta)$ $\quad\quad \mathrm{PR}(\theta) \coloneqq \mathrm{Risk}(\theta, D(\theta))$

Performative stability: $\theta_{\mathrm{PS}} \in \operatorname{argmin}_\theta \mathrm{Risk}(\theta, D(\theta_{\mathrm{PS}}))$

→ Exposing $D(\theta)$ shows new solution concepts for risk minimization

- Performative optimality minimizes risk after deployment
- Performative stability is a natural equilibrium notion in observation-driven optimization

[PZMH20]

# Sensitivity assumption

Definition: We say the distribution map $D(\theta)$ is $\epsilon$-sensitive if for all $\theta, \theta'$

$$W_1\left(D(\theta), D(\theta')\right) \leq \epsilon \left|\left|\theta - \theta'\right|\right|_2$$

*"Similar models lead to similar distributions"*

- Self-fulfilling prophecy
  Small changes in model parameters lead to small changes in predictions, and hence outcomes

# Sensitivity assumption

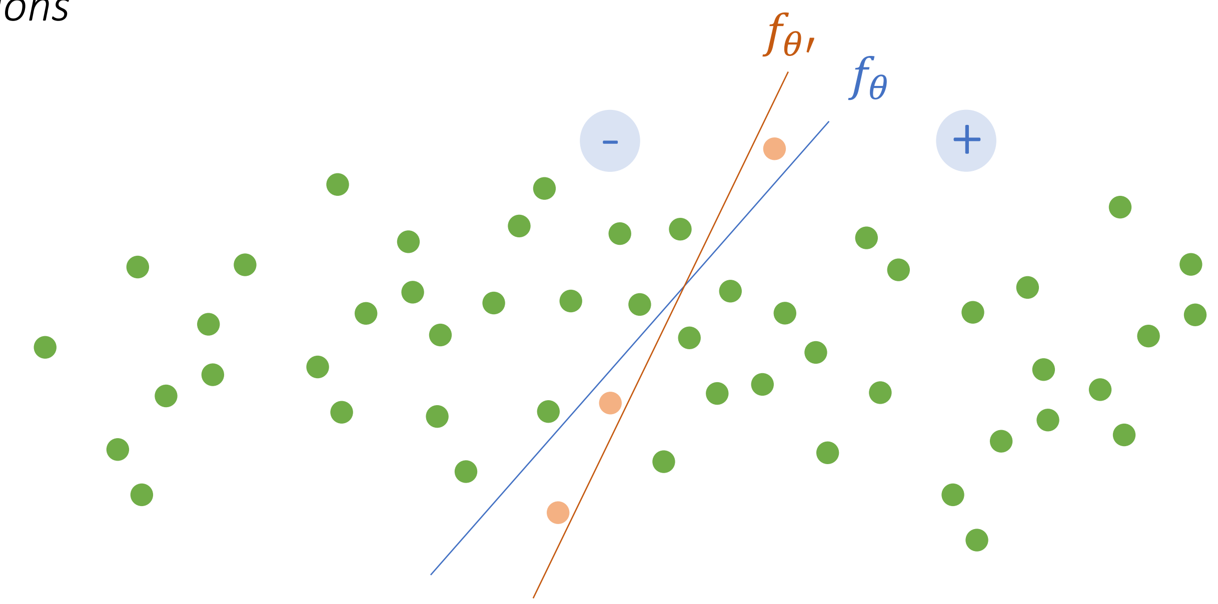
<u>Definition</u>: We say the distribution map $D(\theta)$ is $\epsilon$-sensitive if for all $\theta, \theta'$

$$W_1\big(D(\theta), D(\theta')\big) \leq \epsilon \|\theta - \theta'\|_2$$

*"Similar models lead to similar distributions"*

- **Self-fulfilling prophecy**
  Small changes in model parameters lead to small changes in predictions, and hence outcomes

- **Consequential decisions**:
  Small changes to decision boundary impact only few individuals



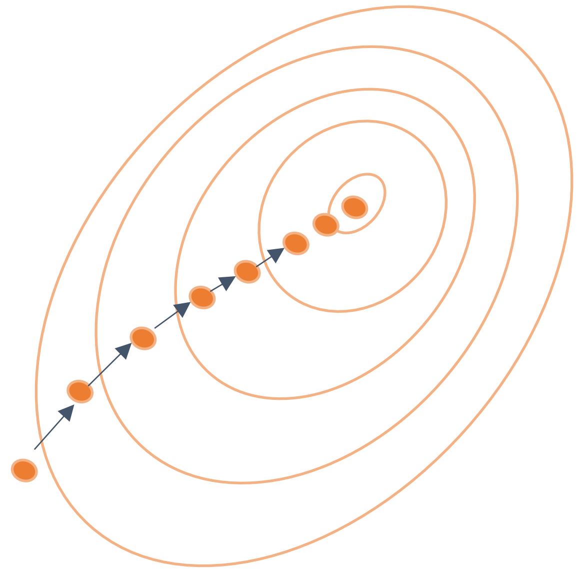
[PZMH20]

# Retraining converges to stable points

*A heuristic for dealing with distribution shifts*

Repeated risk minimization (RRM):

1. deploy the model $\theta_k$
2. observe the induced distribution $D(\theta_k)$
3. let $\theta_{k+1}$ be the risk minimizer on $D(\theta_k)$
4. repeat



<u>Theorem</u> [PZMH20]:

If the loss function is strongly convex and smooth in the data and the distribution map is not too sensitive, then retraining converges to stable points at a linear rate.

→ if any of the three conditions is violated convergence is not guaranteed!

# Beyond risk minimization

Retraining heuristics as natural fixed point dynamic under performativity

$$\theta_{k+1} \leftarrow \text{argmin}_{\theta} \text{ Risk}(\theta, D(\theta_k))$$

Empirical risk using samples of $D(\theta_k)$

gradient update $\theta_k - \eta \, \mathrm{E}_{z \sim D(\theta_k)} [\nabla \ell(z; \theta_k)]$

- ERM and repeated gradient descent [PZ**M**H20]
- Stochastic optimization [**M**PZH22]
- Proximal point methods [DX20]
- Projected gradient descent [WBD21]
- Time-dependent, stateful shifts [BSI20, BHK22,LW22, RRDF22, MTR22]
- Multi-player performative prediction [NFDFR22, PY22, LYW22]

→ small enough sensitivity and appropriate loss function convergence to stable points

# Beyond risk minimization

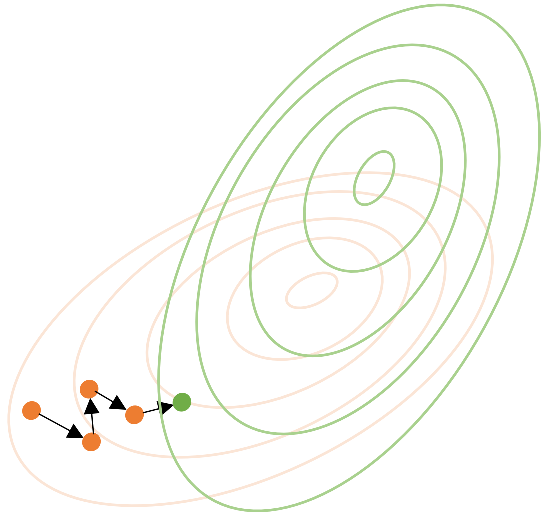
Retraining heuristics as natural fixed point dynamic under performativity

$$\theta_{k+1} \leftarrow \operatorname{argmin}_\theta \operatorname{Risk}(\theta, D(\theta_k))$$

Empirical risk using samples of $D(\theta_k)$

gradient update $\theta_k - \eta\, \mathrm{E}_{z\sim D(\theta_k)} [\nabla \ell(z;\theta_k)]$

- ERM and repeated gradient descent [PZ**M**H20]
- Stochastic optimization [**M**PZH22]



**Consideration for algorithm design:**

Tradeoff sample collection and deployment costs by deciding when to deploy

→ The more samples you collect between deployments, the more samples, but the fewer deployments you need

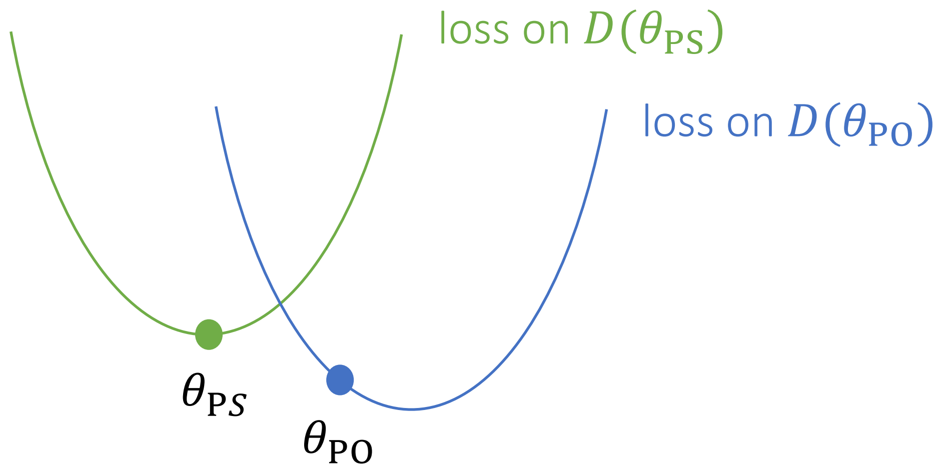
# Performative optimality

we take distribution shift into account

Performative optimality: $\theta_{\mathrm{PO}} \in \operatorname{argmin}_{\theta} \mathrm{PR}(\theta)$

$$\mathrm{PR}(\theta) \coloneqq \mathrm{Risk}(\theta, D(\theta))$$

Performatively stable points are not necessarily performatively optimal!



# Performative optimality

we take distribution shift
into account

Performative optimality: $\theta_{\mathrm{PO}} \in \operatorname{argmin}_{\theta} \mathrm{PR}(\theta)$

$$\mathrm{PR}(\theta) \coloneqq \mathrm{Risk}(\theta, D(\theta))$$

Performatively stable points are not necessarily performatively optimal!

Performative optimality is a natural solution concept under experimentation and modeling

A/B testing,
iterative policy evaluation,
black box optimization

closed form expression
for distribution map, solution
can be evaluated analytically

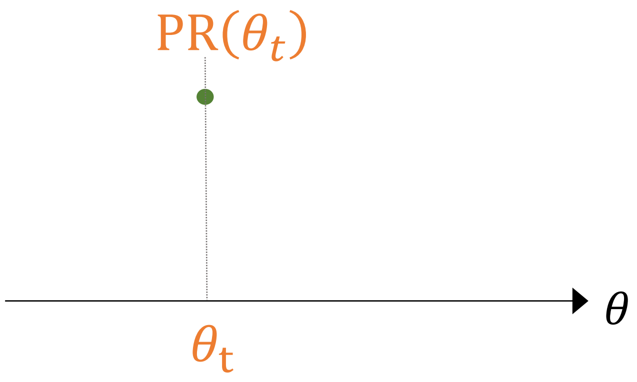
# Exploration-based approach

“live experiments”

- deploy a model $\theta_t$
- observe distribution $D(\theta_t)$
- evaluate performance on induced distribution $\rightarrow \mathrm{PR}(\theta_t)$

**Black-box approach**

Inspired by multi-armed bandits



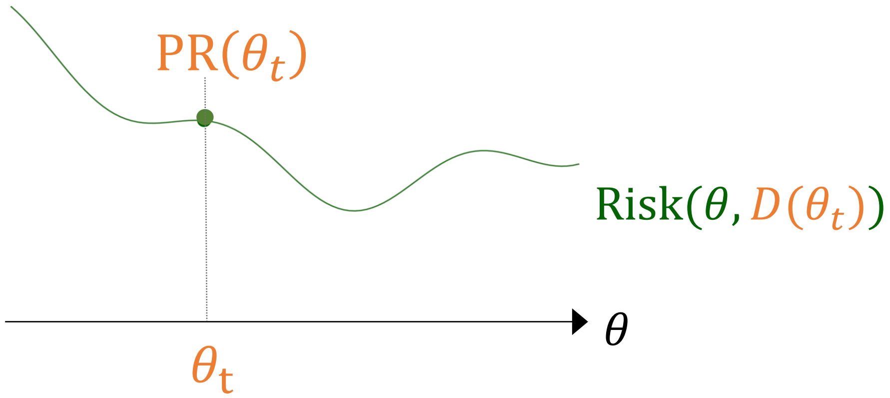
# Exploration-based approach

"live experiments"

- deploy a model $\theta_t$
- observe distribution $D(\theta_t)$
- evaluate performance on induced distribution $\rightarrow \mathrm{PR}(\theta_t)$

**Black-box approach**

Inspired by multi-armed bandits



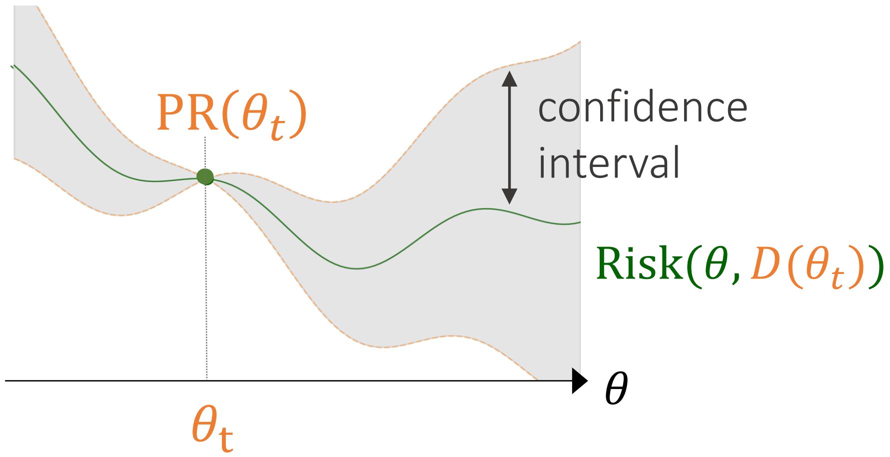
# Exploration-based approach

"live experiments"

- deploy a model $\theta_t$
- observe distribution $D(\theta_t)$
- evaluate performance on induced distribution $\rightarrow \mathrm{PR}(\theta_t)$

**Black-box approach**

Inspired by multi-armed bandits



[JZM22]

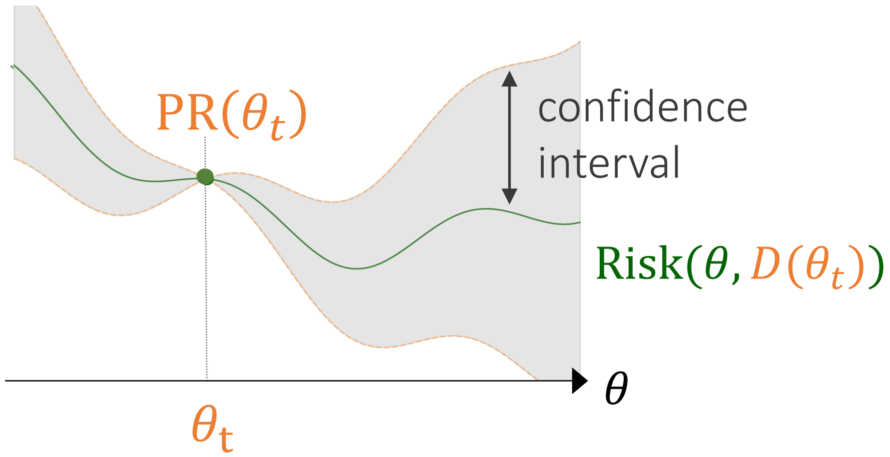
# Exploration-based approach

“live experiments”

- deploy a model $\theta_t$
- observe distribution $D(\theta_t)$
- evaluate performance on induced distribution $\rightarrow \mathrm{PR}(\theta_t)$

**Black-box approach**

Inspired by multi-armed bandits




If the distribution map is not too sensitive and the loss is Lipschitz in the data, then with targeted exploration you can find performative optima with sublinear regret.

[JZM22]

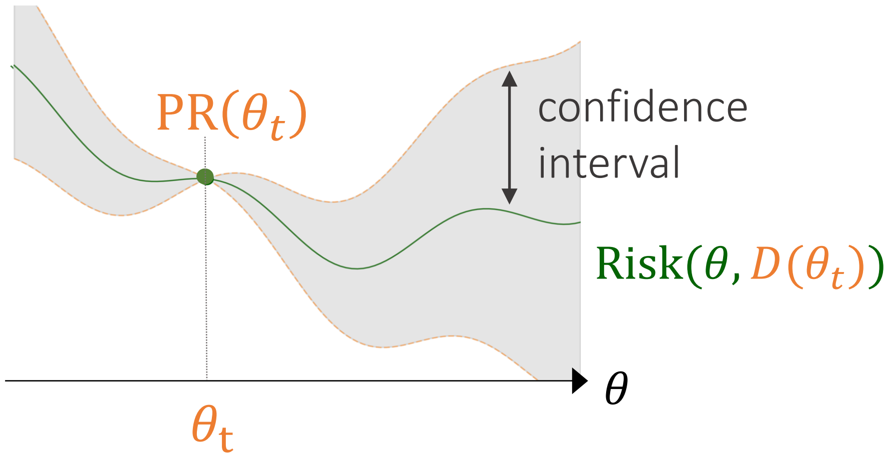
# Exploration-based approach

"live experiments"

- deploy a model $\theta_t$
- observe distribution $D(\theta_t)$
- evaluate performance on induced distribution $\rightarrow \mathrm{PR}(\theta_t)$

**Black-box approach**

Inspired by multi-armed bandits




**Limitations and open challenges:**

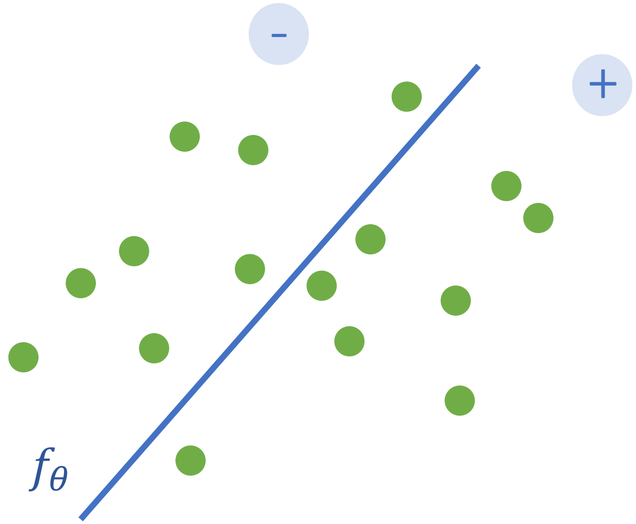
- Incorporating practical constraints on exploration
- Respecting cost and risk of deployments
- Incorporating prior knowledge about distirbution shift

If the distribution map is not too sensitive and the loss is Lipschitz in the data, then with targeted exploration you can find performative optima with sublinear regret.

# Model-based approach

Example: **Strategic classification** [HMPW16]

Distribution $D(\theta)$ comes from strategic behavior of individuals trying to adapt to decision rule




Rational agent model

$$x(\theta) = \underset{x}{\operatorname{argmax}} \; \gamma\, f_\theta(x) - \mathrm{cost}(x_0, x)$$

gain of positive clasification

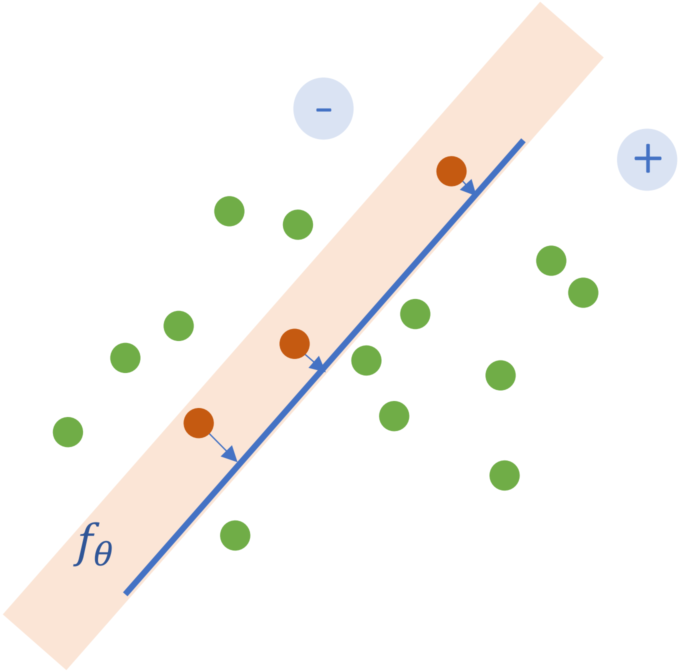
cost of feature manipulation

$D(\theta)$ is "best response map" over $(x(\theta), y)$

# Model-based approach

Example: **Strategic classification** [HMPW16]

Distribution $D(\theta)$ comes from strategic behavior of individuals trying to adapt to decision rule




Rational agent model

$$x(\theta) = \operatorname*{argmax}_x \; \gamma\, f_\theta(x) - \mathrm{cost}(x_0, x)$$

gain of positive clasification

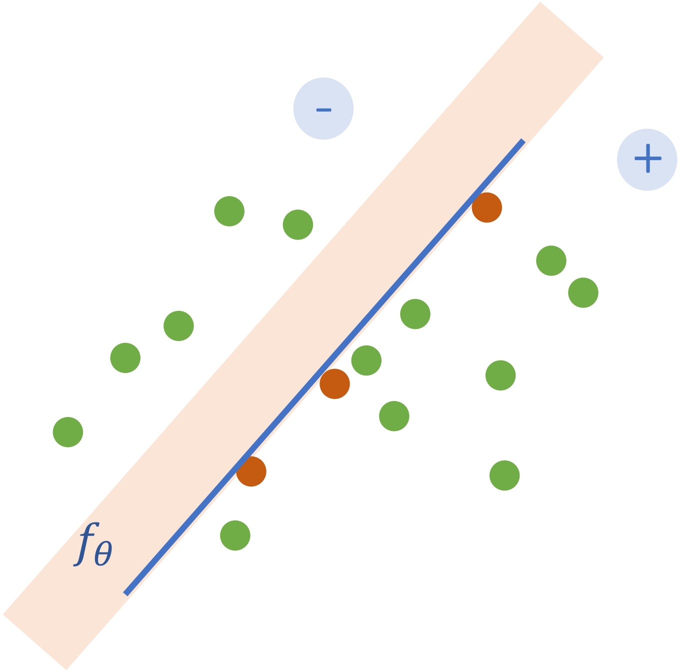
cost of feature manipulation

$D(\theta)$ is “best response map“ over $(x(\theta), y)$

# Model-based approach

Example: **Strategic classification** [HMPW16]

Distribution $D(\theta)$ comes from strategic behavior of individuals trying to adapt to decision rule




Rational agent model

$$x(\theta) = \operatorname*{argmax}_x \; \gamma\, f_\theta(x) - \mathrm{cost}(x_0, x)$$

gain of positive clasification

cost of feature manipulation

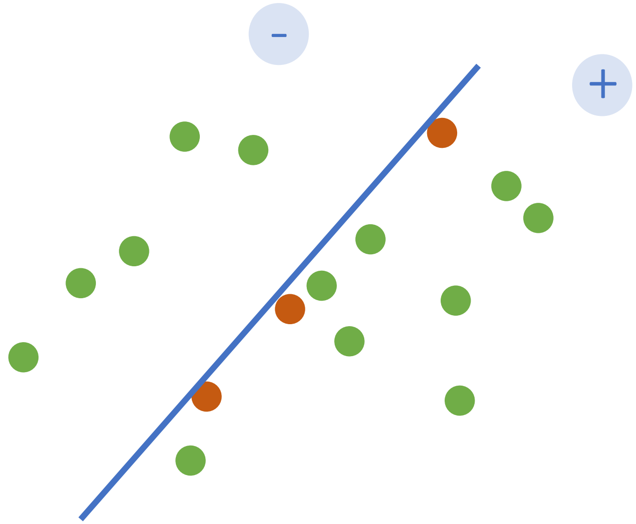
$D(\theta)$ is "best response map" over $(x(\theta), y)$

# Model-based approach

Advantage: precise understanding of distirbution shift allows for analytical solutions

Example: **Strategic classification** [HMPW16]

Distribution $D(\theta)$ comes from strategic behavior of individuals trying to adapt to decision rule




→ Performative optimum corresponds to Stackleberg equilibrium in game between learner and population

<u>Rational agent model</u>

$$x(\theta) = \underset{x}{\operatorname{argmax}} \; \gamma\, f_\theta(x) - \mathrm{cost}(x_0, x)$$

gain of positive clasification

cost of feature manipulation

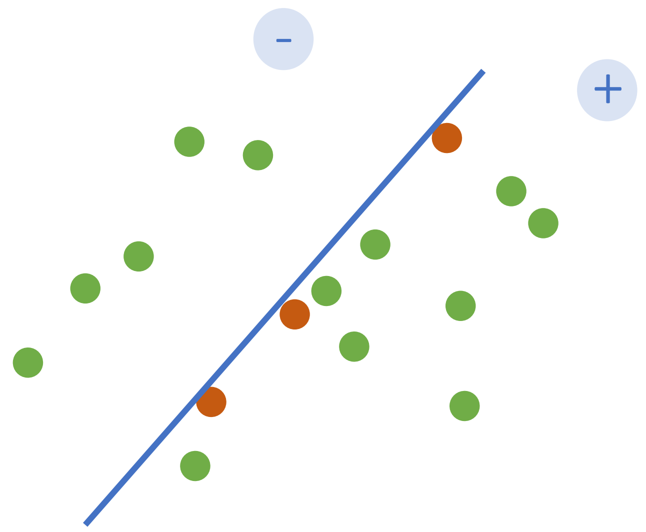
$D(\theta)$ is "best response map" over $(x(\theta), y)$

# Model-based approach

Advantage: precise understanding of distribution shift allows for analytical solutions

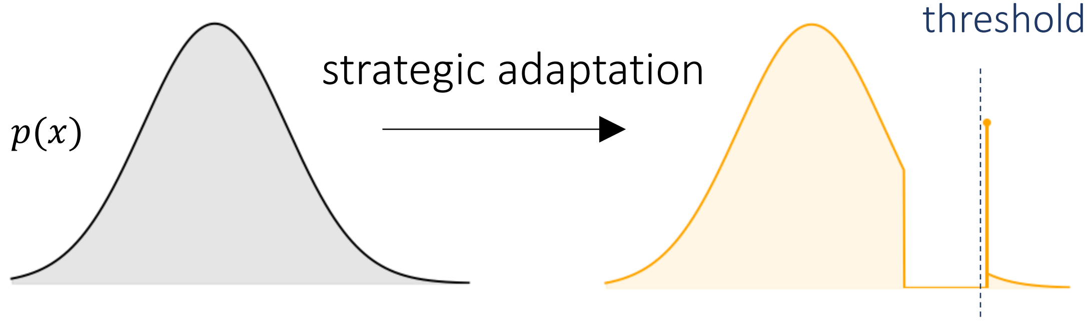
Example: **Strategic classification** [HMPW16]

Distribution $D(\theta)$ comes from strategic behavior of individuals trying to adapt to decision rule




→ Performative optimum corresponds to Stackleberg equilibrium in game between learner and population

Tesion between micro and macro




Microfoundation models can lead to degenerate aggregates and brittle conclusions about learning dynamics in performative prediction, as well as large negative externalities

→ Randomized smoothing can help [JMH21]

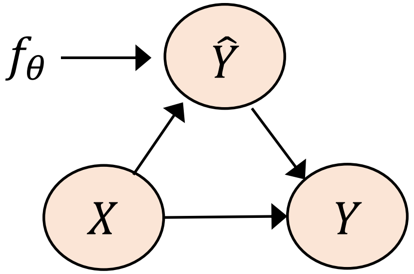
# Alternate models

Behavioral modeling

- Variations in agent costs and model families
- Partial information
- Beyond rationality (e.g., approximate best response)
- Social interactions (e.g., interference, peer effects)

Structural causal models

- Model impact of prediction on outcome (e.g., self-fulfilling prophecy)
- Model interaction dynamics between decision maker and individuals




Macro-models

- Parametric assumptions on distribution map (e.g., location-scale family [MPZ21])

Modeling assumptions permit analytical solution for performative optimality

# Recap

- Performativity is everywhere!

- Performative prediction offers a conceptual framework to reason about performativity in machine learning

- Solution concepts:
  - Performative stability as a natural equilibrium notion for retraining heuristics
  - Performative optimality as the optimal solution post intervention

- Optimization results:
  - A sensitivity assumption on the distribution shift permits interesting theory
  - Performative prediction + microfoundations = strategic classification: Modeling allows anticipating shifts and finding optima analytically

# How performative are predictions?

5M+
downloads

10B+
downloads

# How performative are predictions?

5M+
downloads

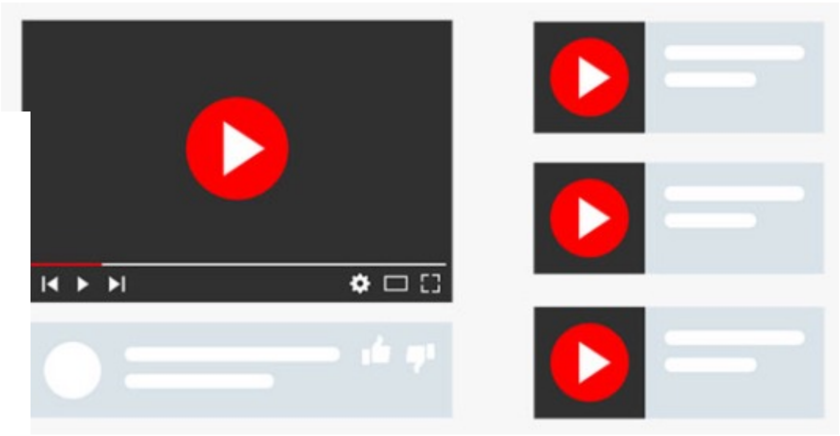
10B+
downloads

- It depends on who is making the prediction
- It depends on power

Can we use performativity to reason about power of predictive systems?

**Lexical definition of power:**
"the capacity or ability to direct or influence others or the course of events"

# How much power do digital platforms have?



# Digital platforms are tricky for market regulation

Stigler Committee on Digital platforms: Final Report 2019

*"Pinpointing the locus of competition can be challenging because markets are multisided and often ones with which economists and lawyers have little experience. This can make market definition another hurdle to effective enforcement."*

European Commission:

*"less emphasize on analysis of market definition, and more emphasis on the theory of harm and identification of anti-competitive strategies.*

# Performative Power

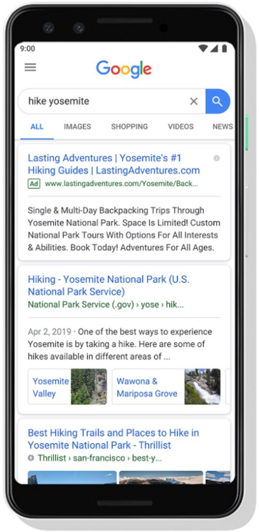
**Informal Definition** [HJM22]: Performative Power is the largest change a firm can cause to a population $U$ with respect to a set of algorithmic actions $F$ and attributes $Z$.

counterfactual data under $f$

$$\mathrm{P} := \sup_{\text{action } f \in \mathrm{F}} \frac{1}{|U|} \sum_{\mathrm{u} \in \mathrm{U}} \mathrm{E}\left[\, \mathrm{dist}(z(u), z_f(u))\right]$$

# Performative Power

**Informal Definition** [HJM22]: Performative Power is the largest change a firm can cause to a population $U$ with respect to a set of algorithmic actions $F$ and attributes $Z$.




$$P := \sup_{\text{action } f \in F} \frac{1}{|U|} \sum_{u \in U} \mathrm{E}\left[\, \mathrm{dist}(z(u), z_f(u)) \right]$$

how to display search results

population of internet users

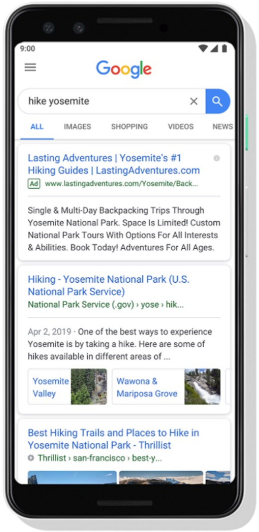
content user clicks

content user clicked under $f$

*'counterfactual'*

# Performative Power

**Informal Definition** [HJM22]: Performative Power is the largest change a firm can cause to a population $U$ with respect to a set of algorithmic actions $F$ and attributes $Z$.




average treatment effect

$$\mathrm{P} := \sup_{\text{action } f\in\mathrm{F}} \boxed{\frac{1}{|U|}\sum_{\mathrm{u}\in\mathrm{U}} \mathrm{E}\left[\,\mathrm{dist}(z(u), z_f(u))\right]}$$

how to display search results

population of internet users

content user clicks

content user clicked under $f$

*'counterfactual'*

# Performative Power

**Informal Definition** [HJM22]: Performative Power is the largest change a firm can cause to a population $U$ with respect to a set of algorithmic actions $F$ and attributes $Z$.

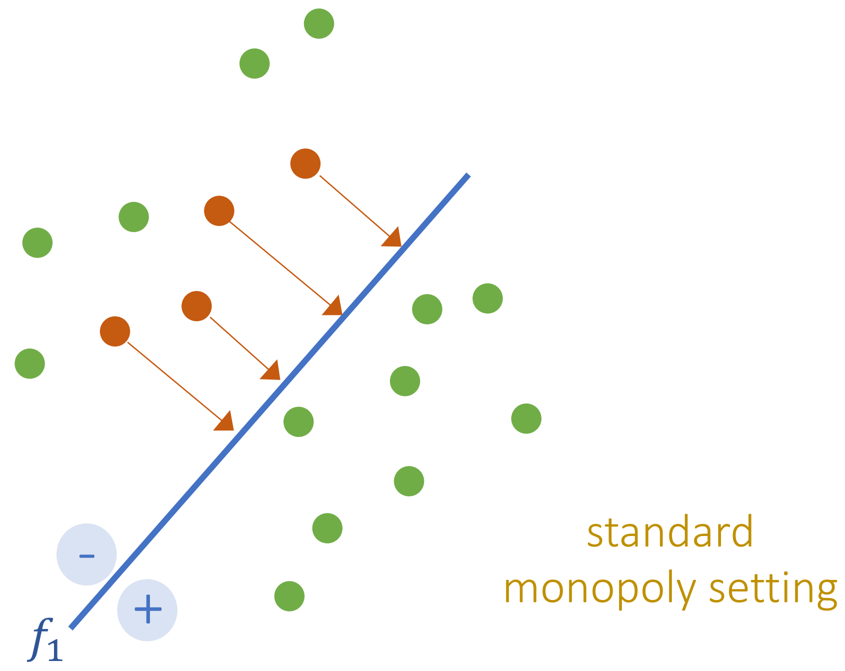
Advantages

- ✓ gives a ‘type-signature’ to power
- ✓ does not require model for competition, concept of prices, equilibria, etc.
- ✓ is a causal statistical notion that can be assessed from data

# Properties of performative power

How performative power relates to the economic study of competition

Multi-player strategic classificaton model



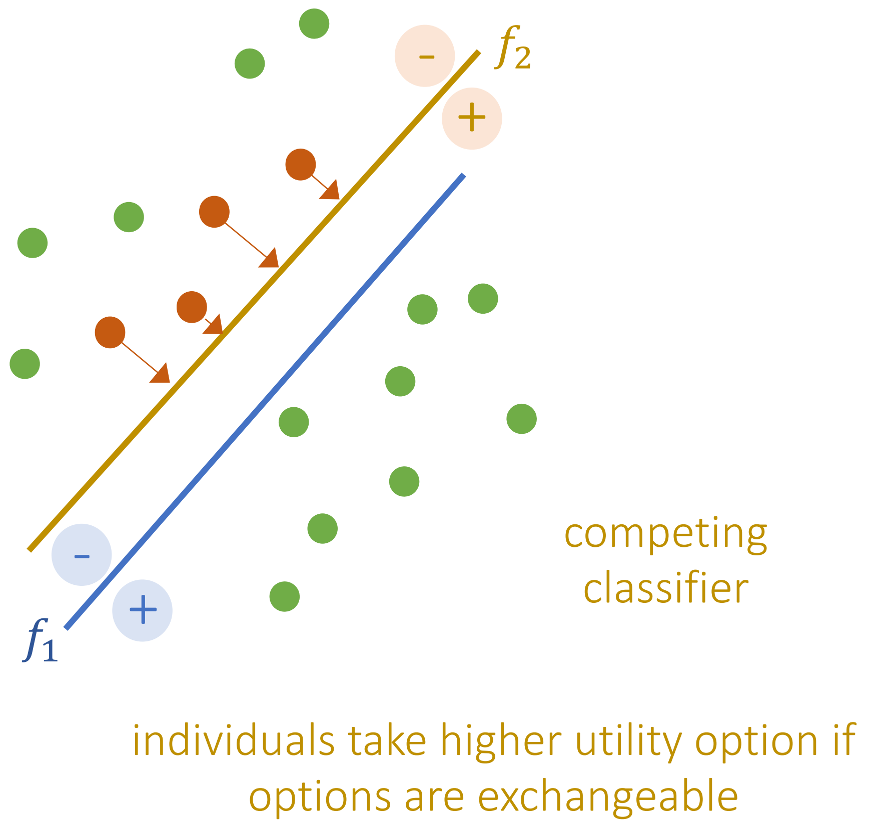

individuals invest up to full surplus utility for adaptation

# Properties of performative power

How performative power relates to the economic study of competition

Multi-player strategic classificaton model

✓ Competition decreases performative power

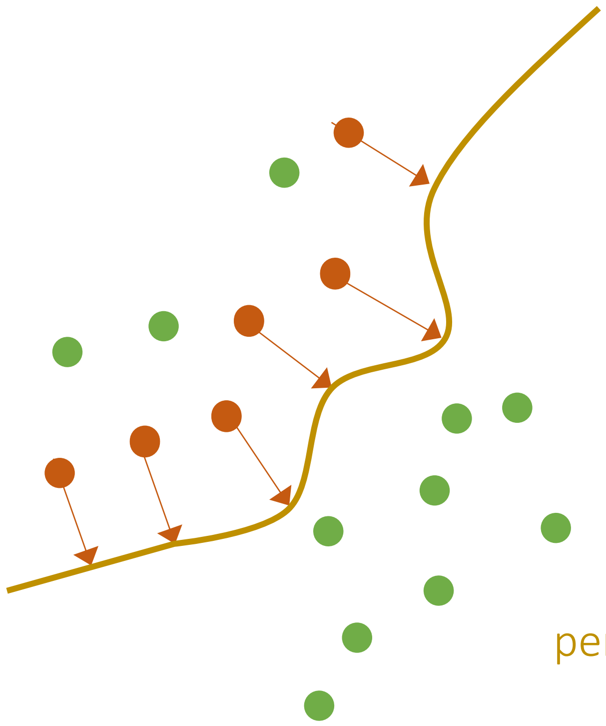


# Properties of performative power

## How performative power relates to the economic study of competition

Multi-player strategic classificaton model

- ✓ Competition decreases performative power
- ✓ The ability to personalize increases performative power



personalization

extract maximum utility simultaneously
from every individual

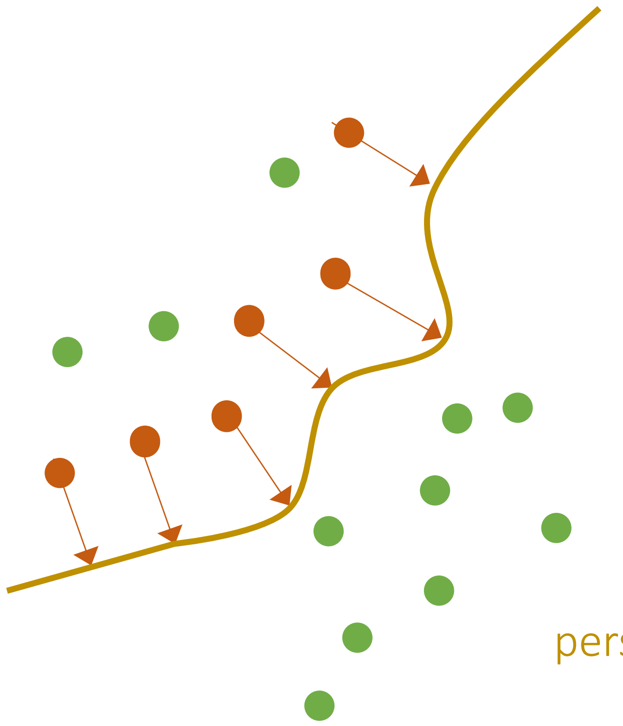
# Properties of performative power

How performative power relates to the economic study of competition

Multi-player strategic classificaton model

- ✓ Competition decreases performative power
- ✓ The ability to personalize increases performative power
- ✓ Outside options decrease performative power

**Sanity check**: Performative power exhibits qualitatively ‘right’ behavior in the presence of competition



personalization

extract maximum utility simultaneously from every individual

# The role of power in prediction

For any given data distribution $D$

$$\mathrm{PR}(\theta) \quad = \quad \mathrm{Risk}(\theta, D) \ - \ \left[\, \mathrm{Risk}(\theta, D) \ - \ \mathrm{Risk}\big(\theta, D(\theta)\big) \,\right]$$

finding performative optima = optimizing given data + steering towards favorable data

# The role of power in prediction

For any given data distribution $D$

zero performative power

$$\mathrm{PR}(\theta) \quad = \quad \mathrm{Risk}(\theta, D) \; - \; \left[ \mathrm{Risk}(\theta, D) \; - \; \mathrm{Risk}\big(\theta, D(\theta)\big) \right]$$

finding performative optima = optimizing given data + steering towards favorable data

*price-taking*

# The role of power in prediction

For any given data distribution $D$

$$\mathrm{PR}(\theta) \quad = \quad \mathrm{Risk}(\theta, D) \;-\; \left[\,\mathrm{Risk}(\theta, D) \;-\; \mathrm{Risk}\big(\theta, D(\theta)\big)\,\right]$$

finding performative optima = optimizing given data + steering towards favorable data

*price-taking* *price-making*

# The role of power in prediction

For any given data distribution $D$

the larger performative power,
the larger the potential for steering

$$\mathrm{PR}(\theta) \quad = \quad \mathrm{Risk}(\theta, D) \; - \; \left[\, \mathrm{Risk}(\theta, D) \; - \; \mathrm{Risk}\big(\theta, D(\theta)\big) \,\right]$$

finding performative optima = optimizing given data + steering towards favorable data

*price-taking* *price-making*

$$\mathrm{PR}(\theta) - \mathrm{Risk}(\theta, D(\theta')) \; \leq O(\mathrm{P}) \text{ for any } \theta'$$

given Lipschitzness

# The potential harms of steering

Stigler Committee on Digital platforms [2019]
concerns about high market power

"Strategies such as offering addictive content at moments when consumers lack self-control increase time spent on the platform and profitable ad sales even as the platform lowers the quality of content. These tactics increase the welfare costs of market power."

# Google shopping antitrust case

... ongoing since 2010

In 2017 the EU commission found that Google has infringed Article 102 TEUF by abusing its dominant position in the search for favoring its own comparison shopping service over competitors (‘*self-preferencing*’)


Home > Press corner > Antitrust: Commission fines Google €2.42 billion

Press release | 27 June 2017 | Brussels

**Antitrust: Commission fines Google €2.42 billion for abusing dominance as search engine by giving illegal advantage to own comparison shopping service**

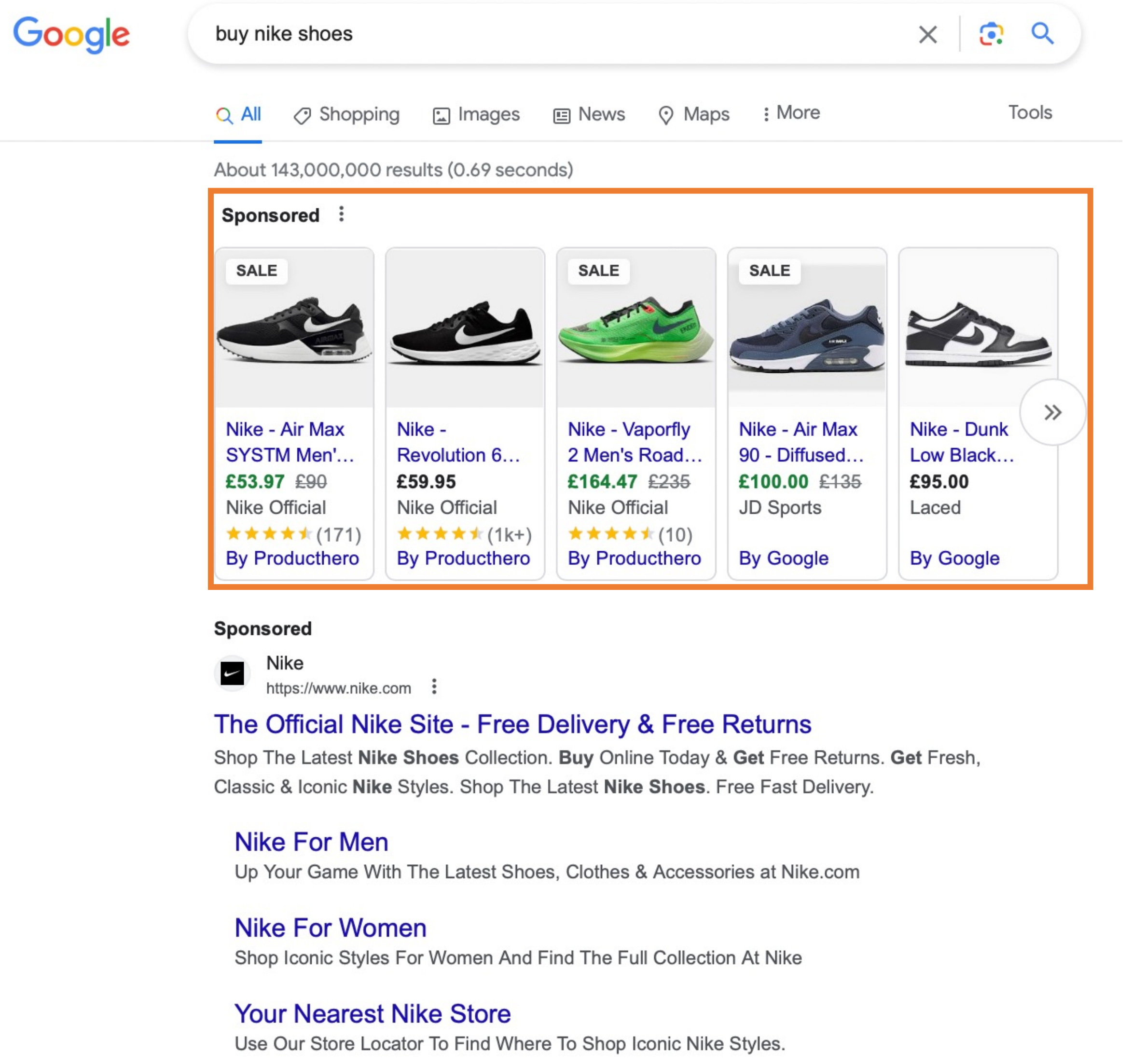
# Google shopping antitrust case

… ongoing since 2010

In 2017 the EU commission found that Google has infringed Article 102 TEUF by abusing its dominant position in the search for favoring its own comparison shopping service over competitors (‘*self-preferencing*’)

**Key technical claim**:
Arrangement of content steer traffic to Google away from competitors

Causal question of display bias at the heart of the investigation



# The role of performative power?

Trying to get at the causal question at the heart of an investigation

- Links the consequences of actions to the concept of power
- Performative power offers a framework to apply tools from causal inference for estimating power

# The role of performative power?

Trying to get at the causal question at the heart of an investigation

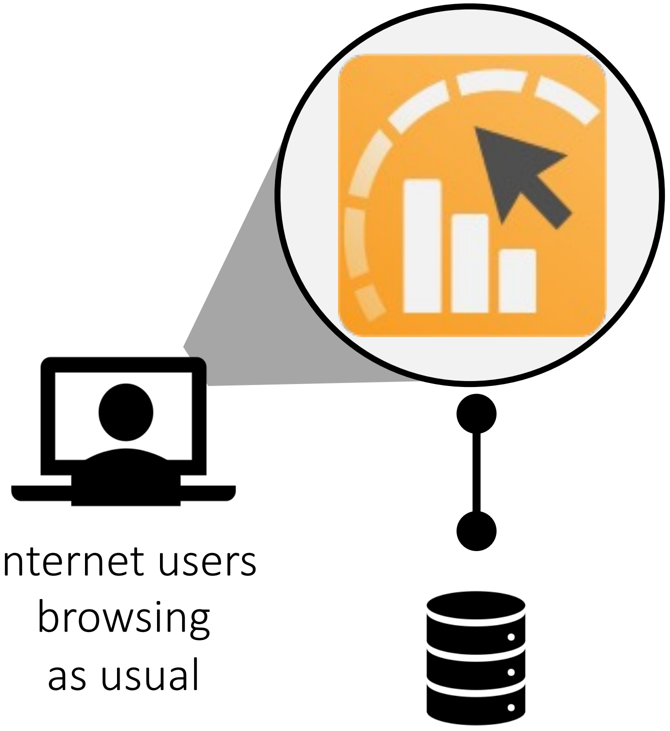
- Links the consequences of actions to the concept of power
- Performative power offers a framework to apply tools from causal inference for estimating power

a) Insights from existing studies

*Anderson, Magruder* (2012) “An extra half-star rating [on Yelp] causes restaurants to sell out 19 percentage points (49%) more frequently”
*Narayanan, Kalyanam* (2015) “Being ranked 2 instead of 1 in Google Ads reduces CTR by 21%”

b) New experimental designs

Powermeter: An ongoing research project with Gabriele Carovano and Moritz Hardt



# Summary

- Predictions have causal powers, performativity is everywhere
- Performative prediction offers a conceptual framework to reason about the causal effect of predictions in machine learning
- Simple syntactic changes to classical risk minimization allows to distinguish solution concepts, brings forth new algorithms, and articulates important optimization challenges
- Performativity allows us to articulate the difference between learning and steering
- Performativity and the causal power of predictions plays an important role in digital market investigations

Framework for brining together machine learning, causality, behavioral economics, control theory, game theory, macroeconomics and social sciences more broadly

# Thanks to my great collaborators

Moritz Hardt
MPI-IS

Meena Jagadeesan
UC Berkeley

Juan C. Perdomo
UC Berkeley
(soon Harvard)

Tijana Zrnic
UC Berkeley
(soon Stanford)

Gabriele Carovano
Uni Tübingen

# Questions, thoughts, suggestions?

cmendler@tuebingen.mpg.de

# References

[PZMH20] Performative Prediction. ICML 2020
J. Perdomo, T. Zrnic, C. Mendler-Dünner, M. Hardt

[MPZH20] Stochastic Optimization for Performative Prediction. NeurIPS 2020.
C. Mendler-Dünner, J. Perdomo, T. Zrnic, M. Hardt

[JZM22] Regret Minimization with Performative Feedback. ICML 2022.
M. Jagadeesan, T. Zrnic, C. Mendler-Dünner

[JMH21] Alternative Microfoundations for Strategic Classification. ICML 2021
M.Jagadeesan, C. Mendler-Dünner, M. Hardt

[HJM22] Performative Power. NeurIPS 2022
M.Hardt, M. Jagadeesan, C. Mendler-Dünner

[MDW22] Anticipating Performativity by Predicting from Predictions. NeurIPS 2022
C. Mendler-Dünner, F. Ding, and Yixin Wang

[CHM23] Causal Inference out of Control: Estimating the Steerability of Consumption. Arxiv 2023.
G. Cheng, M. Hardt, C. Mendler-Dünner

# References


[DX20] Stochastic optimization with decision-dependent distributions, Arxiv 2020 and MOOR 2022.
D. Drusvyatskiy, L. Xiao

[LYW22] Multi-agent Performative Prediction with Greedy Deployment and Consensus Seeking Agents. NeurIPS 2022.
Q. Li, C. Yau, H. Wai

[PY22] Multi-agent Performative Prediction: From Global Stability and Optimality to Chaos. Arxiv 2022.
G. Piliouras, F. Yu

[BHK22] Performative Prediction in a Stateful World. AISTATS 2022.
G. Brown, S. Hod, I. Kalemaj

[LW22] State Dependent Performative Prediction with Stochastic Approximation. AISTATS 2022.
Q. Li, H. Wai

# References


[RRDF22] Decision-Dependent Risk Minimization in Geometrically Decaying Dynamic Environments. AAAI 2022.
M. Ray, L. Ratliff, D. Drusvyatskiy, M. Fazel

[MTR22] Performative Reinforcement learning. Arxiv 2022.
D. Mandal, S. Triantafyllou, G. Radanovic

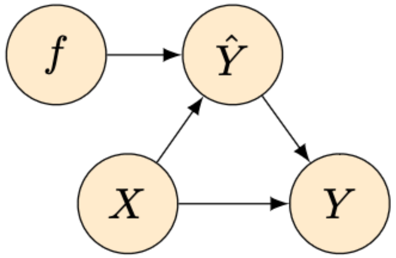
[MMG22] Performative Prediction with Neural Networks. DistShift@NeurIPS 2022.
M. Mofakhami, I. Mitliagkas, G. Gidel

# Thank you!

# Causal modeling

[MDW22]

# Performativity of Predictions




$$X = \xi_X \qquad \xi_X \sim \mathcal{D}_X \qquad (1)$$

$$\hat{Y} = f(X, \xi_f) \qquad \xi_f \sim \mathcal{D}_f \qquad (2)$$

$$Y = g(X, \hat{Y}) + \xi_Y \qquad \xi_Y \sim \mathcal{D}_Y \qquad (3)$$

Figure 1: Performative effects in the outcome mediated by the prediction for a given $f$

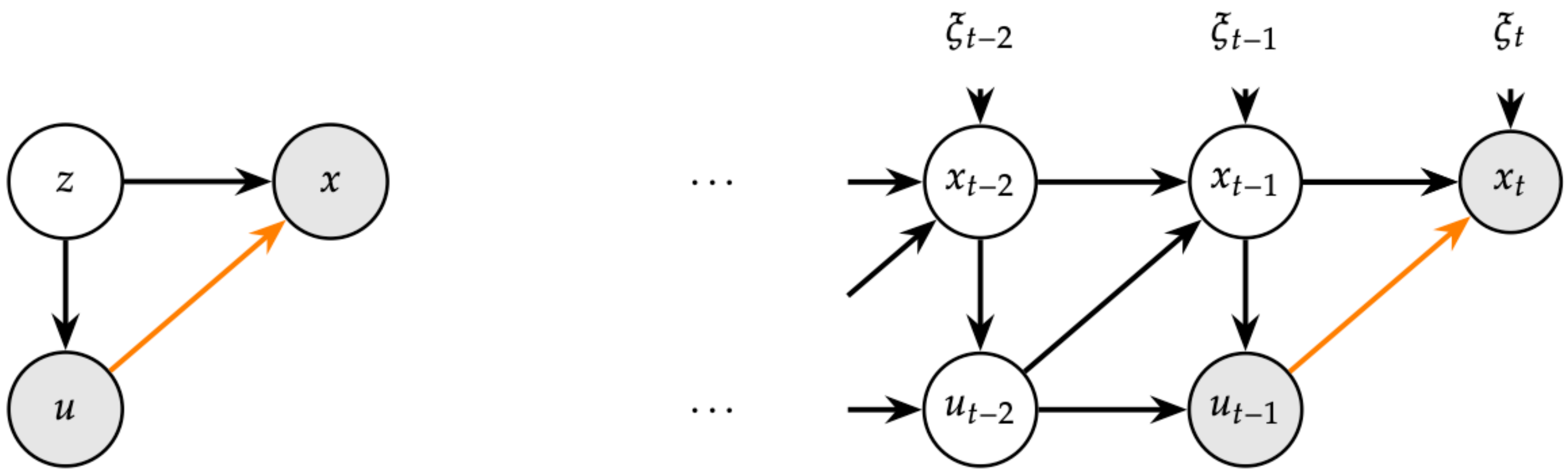
Challenge for identifiability:
correlation of prediction with outcome and deterministic nature of predictions leads to positivity violations

Sufficient conditions for identifiability:

a) Randomization in predictions (e.g. for differential privacy or fairness)
b) Incongruence in modality + separability (e.g. discrete predictions)
c) Incongruence on functional complexity + separability (e.g. overparameterized models)

[CHM23]

# Estimating Steerability of Consumption




(a) standard model

(b) modeling temporal confounding structure

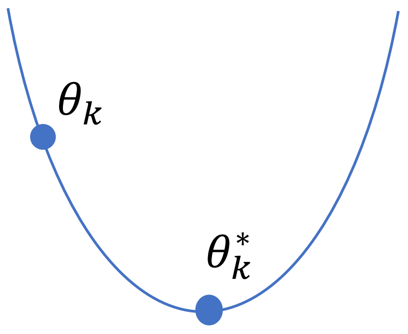
**Challenge for identifiability**:
Positivity violation in confounding variable (confounder can be long rollout over past actions and states)

**Our approach**
Explicitly model temporal dynamics
Assume platform action is sufficiently sensitive.
Consumption shocks propagate through system and allow valid observational designs.

# Proof sketch: Convergence of retraining

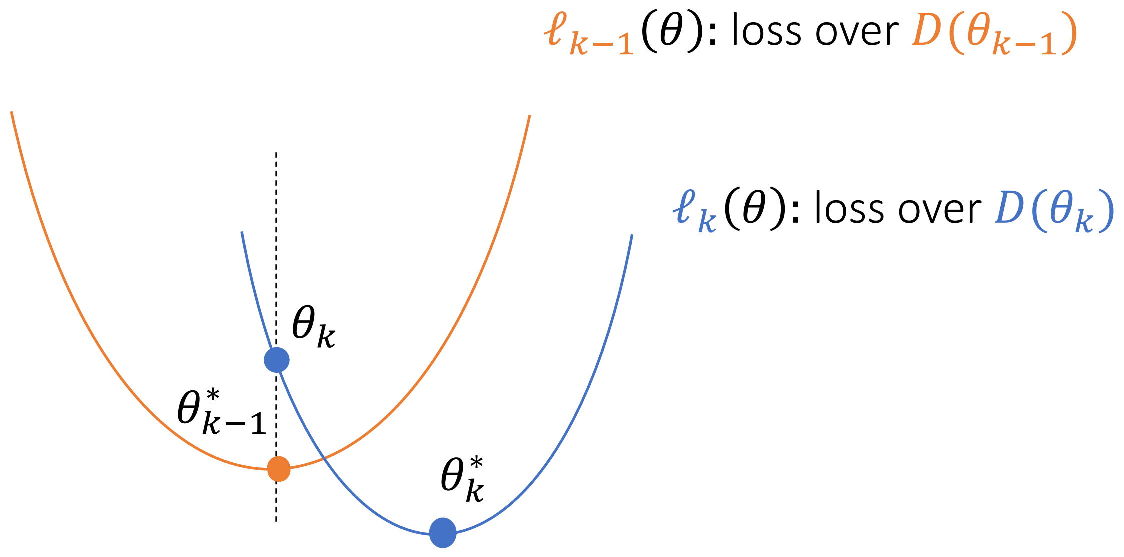
# Proof sketch

$\ell_k(\theta)$: loss over $D(\theta_k)$



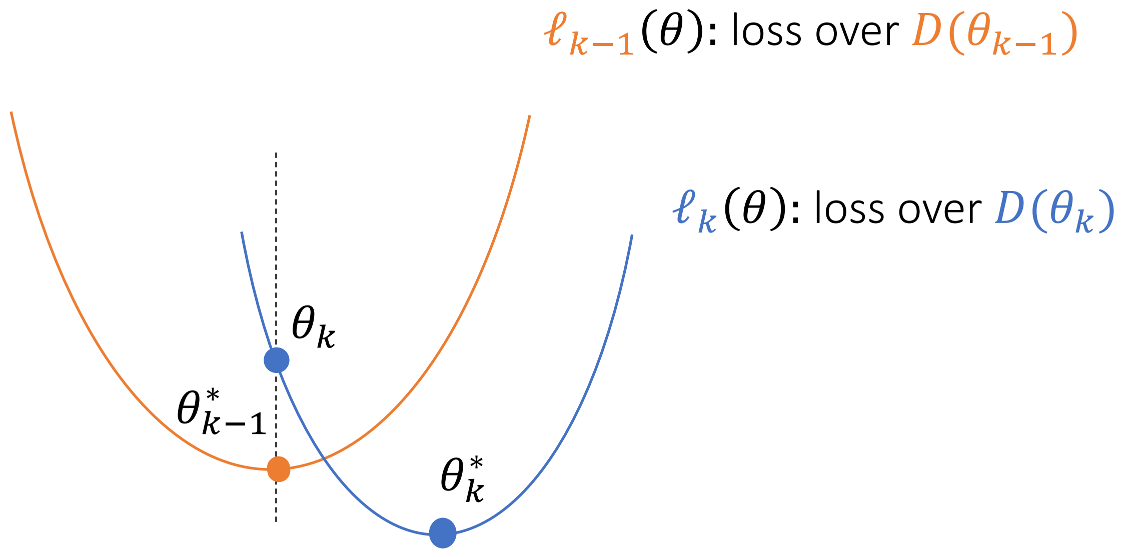

- $\gamma$-strong convexity of the loss in $\theta$: $[\nabla \ell_k(\theta_k) - \cancelto{0}{\nabla \ell_k(\theta_k^*)}]^T(\theta_k - \theta_k^*) \geq \gamma||\theta_k - \theta_k^*||^2$

# Proof sketch




- $\gamma$-strong convexity of the loss in $\theta$: $[\nabla \ell_k(\theta_k) - \cancelto{0}{\nabla \ell_k(\theta_k^*)}]^T(\theta_k - \theta_k^*) \geq \gamma||\theta_k - \theta_k^*||^2$

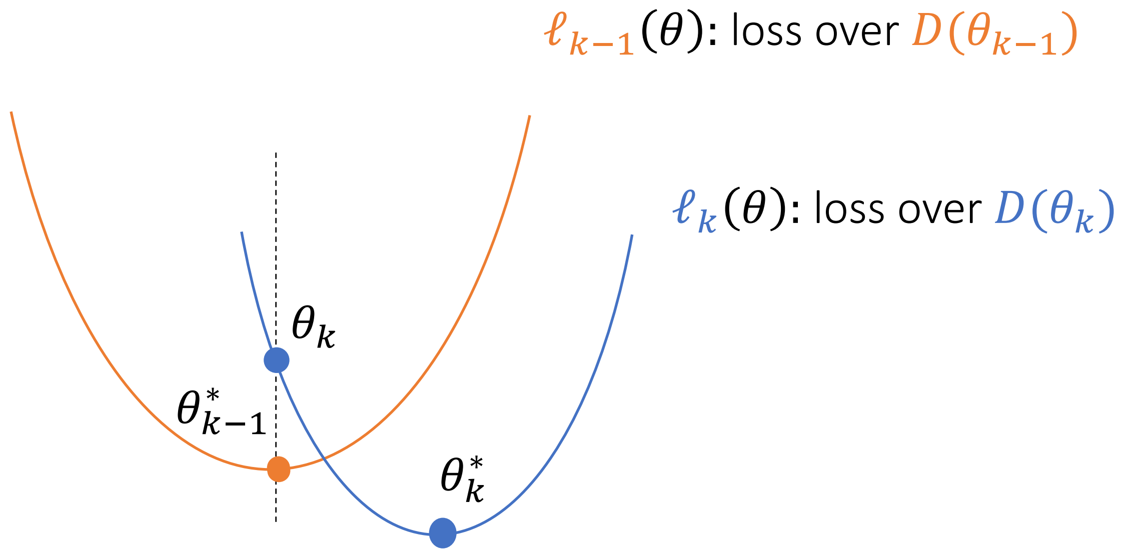
# Proof sketch




- $\gamma$-strong convexity of the loss in $\theta$: $\quad [\nabla \ell_k(\theta_k) - \cancelto{0}{\nabla \ell_k(\theta_k^*)}]^T(\theta_k - \theta_k^*) \geq \gamma||\theta_k - \theta_k^*||^2$
- $\beta$-smoothness of the loss in the data: $\quad [\nabla \ell_k(\theta_k) - \cancelto{0}{\nabla \ell_{k-1}(\theta_k)}]^T(\theta_k - \theta_k^*) \leq \beta\left||\theta_k - \theta_k^*|\right| W(D(\theta_{k-1}), D(\theta_k))$

Kantorovich Rubinstein
for L-Lipschitz functions $f$
$\mathrm{E}_{x\sim D_1} f(x) - \mathrm{E}_{x\sim D_2} f(x) \leq L\, W(D_1, D_2)$

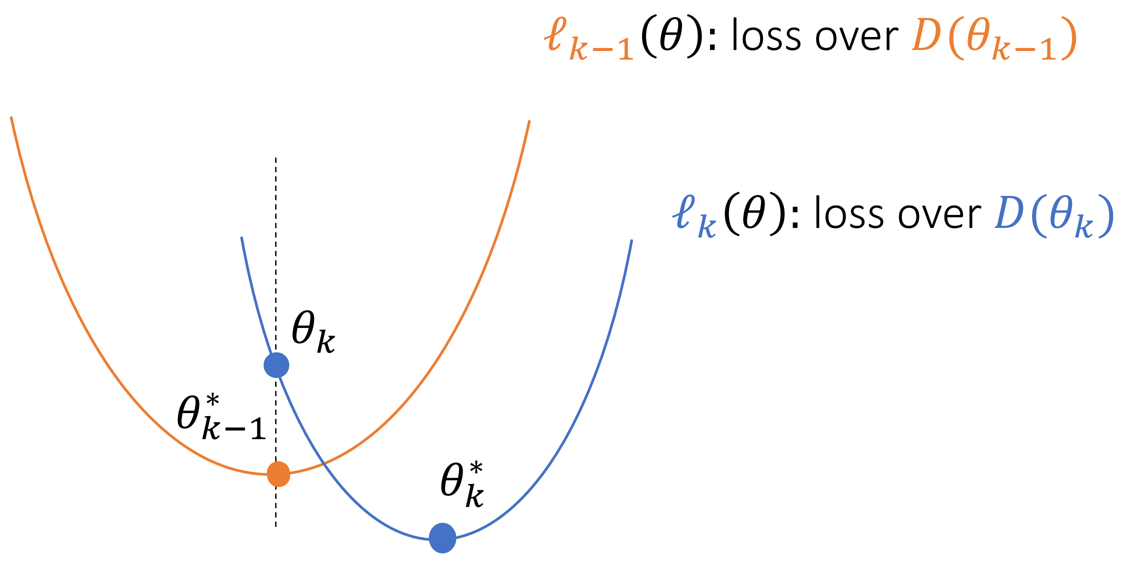
# Proof sketch




- $\gamma$-strong convexity of the loss in $\theta$: $[\nabla \ell_k(\theta_k) - \cancelto{0}{\nabla \ell_k(\theta_k^*)}]^T(\theta_k - \theta_k^*) \geq \gamma ||\theta_k - \theta_k^*||^2$
- $\beta$-smoothness of the loss in the data: $[\nabla \ell_k(\theta_k) - \cancelto{0}{\nabla \ell_{k-1}(\theta_k)}]^T(\theta_k - \theta_k^*) \leq \beta \big|| \theta_k - \theta_k^* \big|| \, W(D(\theta_{k-1}), D(\theta_k))$

$$\Rightarrow \gamma ||\theta_k - \theta_k^*|| \leq \beta \, W(D(\theta_{k-1}), D(\theta_k))$$

# Proof sketch




- $\gamma$-strong convexity of the loss in $\theta$: $\quad [\nabla \ell_k(\theta_k) - \cancelto{0}{\nabla \ell_k(\theta_k^*)}]^T(\theta_k - \theta_k^*) \geq \gamma ||\theta_k - \theta_k^*||^2$

- $\beta$-smoothness of the loss in the data: $\quad [\nabla \ell_k(\theta_k) - \cancelto{0}{\nabla \ell_{k-1}(\theta_k)}]^T(\theta_k - \theta_k^*) \leq \beta \big||\theta_k - \theta_k^*|\big| \, W(D(\theta_{k-1}), D(\theta_k))$

$$\Rightarrow \gamma ||\theta_k - \theta_k^*|| \leq \beta \, W(D(\theta_{k-1}), D(\theta_k))$$

- $\epsilon$-sensitivity of $D(\cdot)$: $\quad \leq \beta\epsilon \, \big||\theta_{k-1} - \theta_k|\big|$

$$= \beta\epsilon \, \big||\theta_{k-1} - \theta_{k-1}^*|\big| \qquad \text{contraction for } \epsilon < \gamma/\beta$$

$\square$

# Stochastic optimization in performative prediction

[MPZH20]

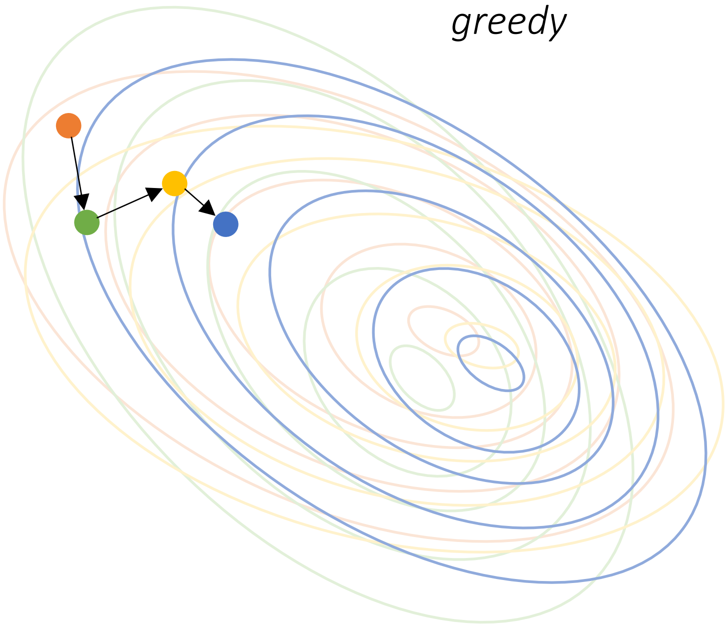
# Stochastic Optimization in Performative Prediction

Samples arrive one at a time: $\theta_{t+1} = \theta_t - \eta_t \nabla \ell(z, \theta_t)$ with $z \sim D(\theta_{deploy(t)})$

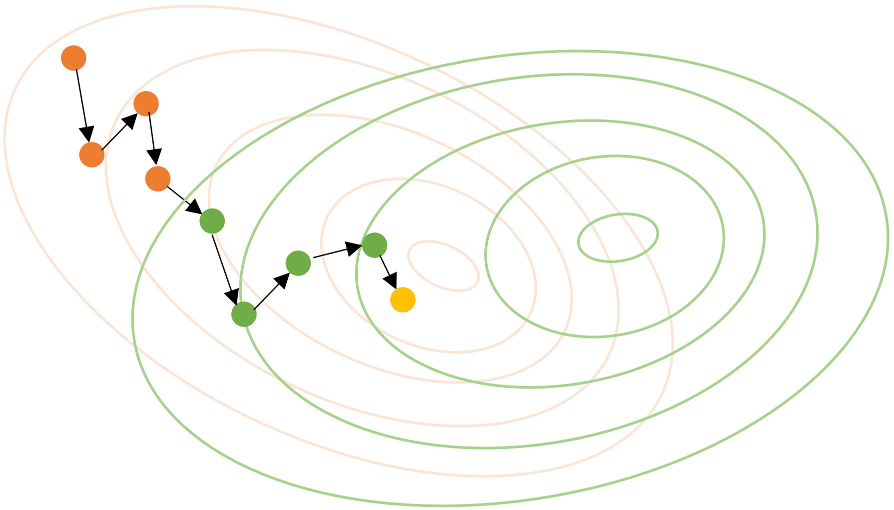
index of deployed model at time step $t$

- **Greedy deploy**: Deploy model after every single update
- **Lazy deploy:** Set $\alpha > 0$ and perform $ck^{\alpha}$ updates between deployments $k$ and $k+1$

*greedy*



*lazy*



[MPZH20]

# Stochastic Optimization in Performative Prediction

**Bounded second moment:**

$\theta_\phi^* = \mathrm{argmin}_\theta \ \mathrm{Risk}(\theta, D(\phi))$

$$\mathrm{E}_{z\sim D(\phi)}\left[||\nabla \ell(z,\theta)||_2^2\right] \leq \sigma^2 + L^2||\theta - \theta_\phi^*|| \quad \text{for any } \theta, \phi$$

In addition, assume a) $\beta$-smooth loss in z and $\theta$, b) $\gamma$-strongly convex loss in $\theta$, c) $\epsilon < \gamma/\beta$

**Proposition**: With an appropriate stepsize schedule, a solution $\theta^*$ with $||\theta^* - \theta_{PS}|| \leq \delta$ is reached after

- $O(1/\delta)$ updates and $O(1/\delta)$ deployments for greedy deploy
- $O(1/\delta^{\frac{1+\alpha}{\alpha}})$ updates and $O(1/\delta^{\frac{1}{\alpha}})$ deployments for lazy deploy

→ For $\alpha \gg 1$ lazy deploy has asympthotic sample complexity $O(1/\delta)$ with only $O(1/\delta^{\frac{1}{\alpha}})$ deployments.

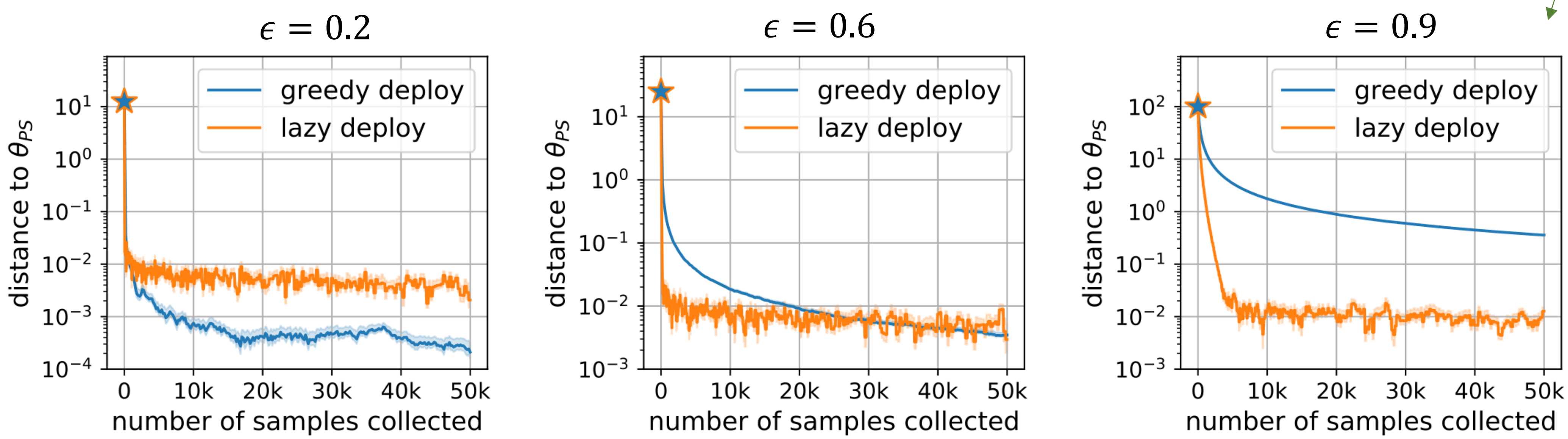
- Stepsize for greedy deploy is globally decreasing and becomes more conservative as $(\gamma - \epsilon\beta) \to 0$
- Stepsize for lazy deploy is locally decreasing between deployments and is independent of $\epsilon$

# Stochastic Optimization in Performative Prediction

Different regimes depending on strength of performativity

- Greedy deploy is better if performativity is weak.
- Lazy deploy is better at dealing with strong shifts and poor initialization.

deployments
greedy: 50K
lazy: 200




Setup: Mean estimation $z \sim N(\mu + \epsilon\theta, \sigma^2)$ using $\ell(z, \theta) = \frac{1}{2}(z - \theta)^2$

→ see paper for a semi-synthetic credit scoring example

# Confidence bound algorithm

[JZM22]

# Tighter confidence bounds intuition

*Ignore finite sample considerations for now*

- After deploying $\theta_t$ we observe $D(\theta_t)$
- What do we learn about performative risk of an unexplored $\theta_{new}$?

$$\mathrm{PR}(\theta_{\mathrm{new}}) - \mathrm{PR}(\theta_{\mathrm{t}}) = \mathrm{Risk}(\theta_{\mathrm{new}}, D(\theta_{\mathrm{new}})) - \mathrm{Risk}(\theta_{\mathrm{new}}, D(\theta_{\mathrm{t}}))$$

uncertainty due to distribution shift

$$+ \mathrm{Risk}(\theta_{\mathrm{new}}, D(\theta_{\mathrm{t}})) - \mathrm{Risk}(\theta_{\mathrm{t}}, D(\theta_{\mathrm{t}}))$$

uncertainty due to changing predictive model

- We can use feedback about $D(\theta_t)$ and knowledge of the loss to evaluate second term offline
- We only pay for uncertainty due to distribution shift

$\rightarrow$ we need Lipschitzness of $\mathrm{Risk}(\theta, D(\phi))$ in $\phi$ to control the first term

*Lipschitz loss in z*
*+ sensitivity*

→ see [JZM22] for more details

# Performative regret bound

*to deal with finite sample uncertainty we proceed in phases and progressively refine precision of risk estimate*

Assume the distribution map $D(\theta)$ is $\epsilon$-sensitive and the loss $\ell(z;\theta)$ is $L_z$-Lipschitz in $z$. Then, there exists an algorithm that after $T$ deployments achieves a regret bound of

$$\mathrm{Reg}(T) = \tilde{O}\left(\sqrt{T} + T^{\frac{d+1}{d+2}}(L_z\epsilon)^{\frac{d}{d+2}}\right)$$

where $d$ denotes the "zooming dimension" of the problem

Baseline: Lipschitz bandits [Kleinberg et al. 2008] $\mathrm{Reg}(T) = \tilde{O}\left(T^{\frac{d'+1}{d'+2}} L^{\frac{d'}{d'+2}}\right)$

$L$ Lipschitz constant PR
$d' \geq d$ zooming dimension

**Benefits of our bound:**

the complexity of the distribution shift

- regret bound primarily scales with $L_z\epsilon$ and not with $L$
- as $\epsilon \to 0$ bound scales as $\tilde{O}(\sqrt{T})$ (no dimension dependence)
- no constraint on loss as a function of $\theta$