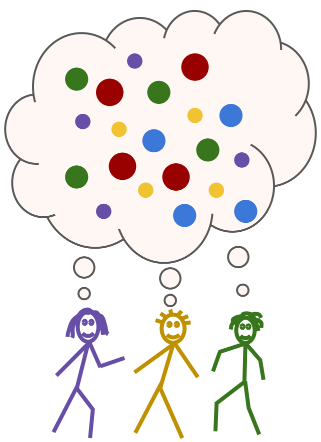
Some approaches for

Acquiring and Aggregating Information

Bo Waggoner

Harvard

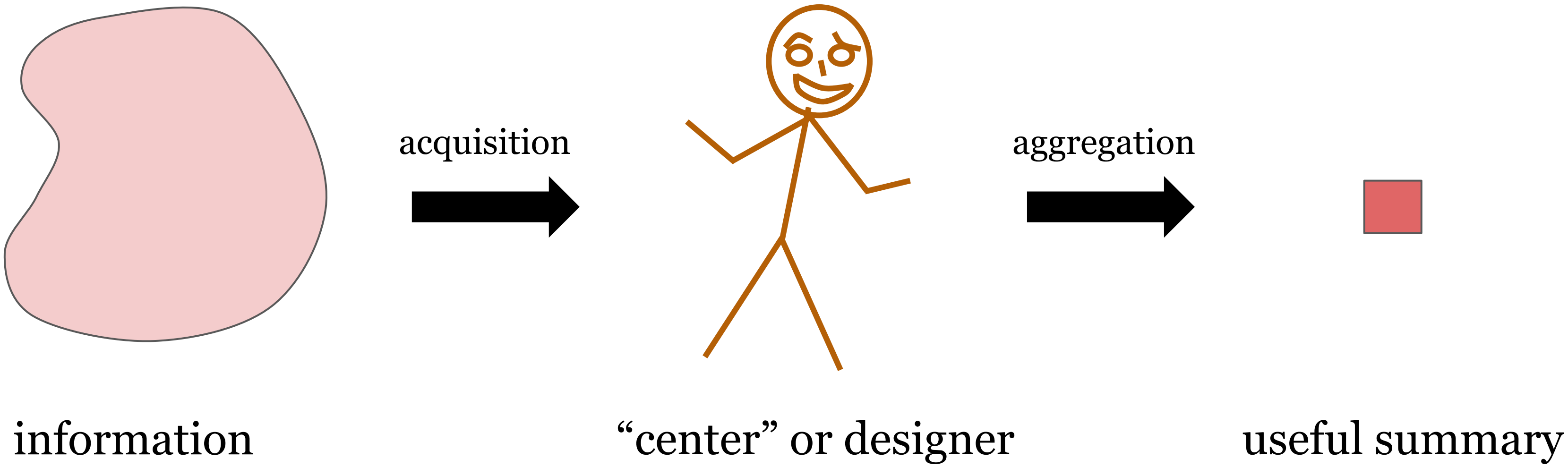


Based on joint works with:

- Jacob Abernethy
- Yiling Chen
- Rafael Frongillo
- Chien-Ju Ho

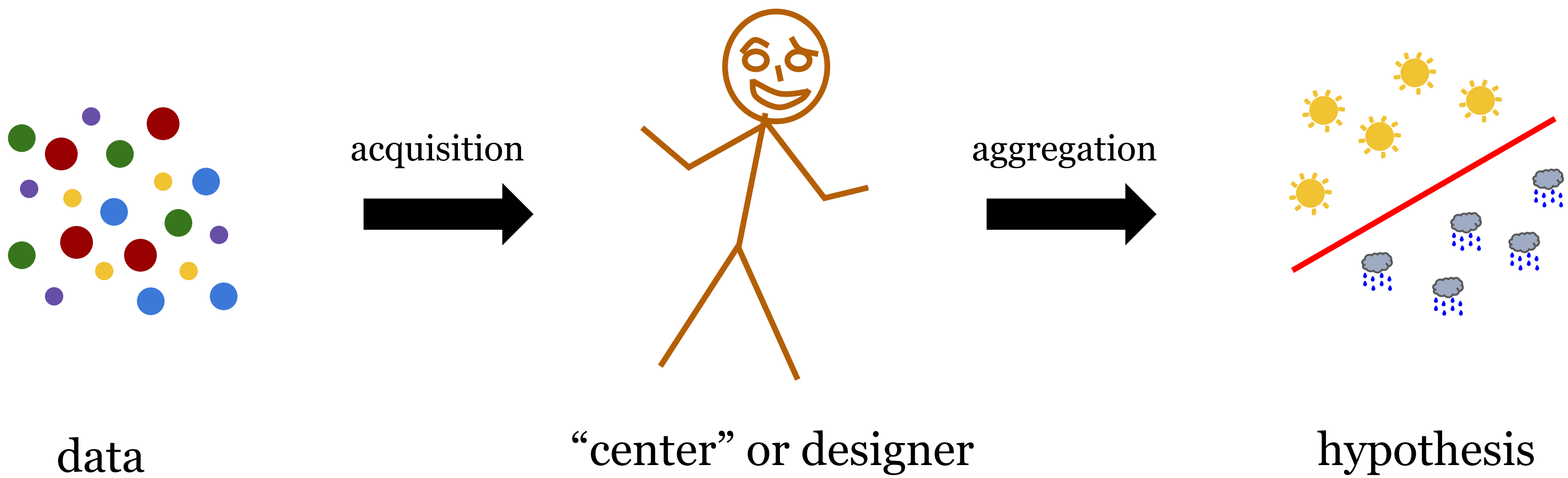
March 2016

A common pattern



*drawing not to scale

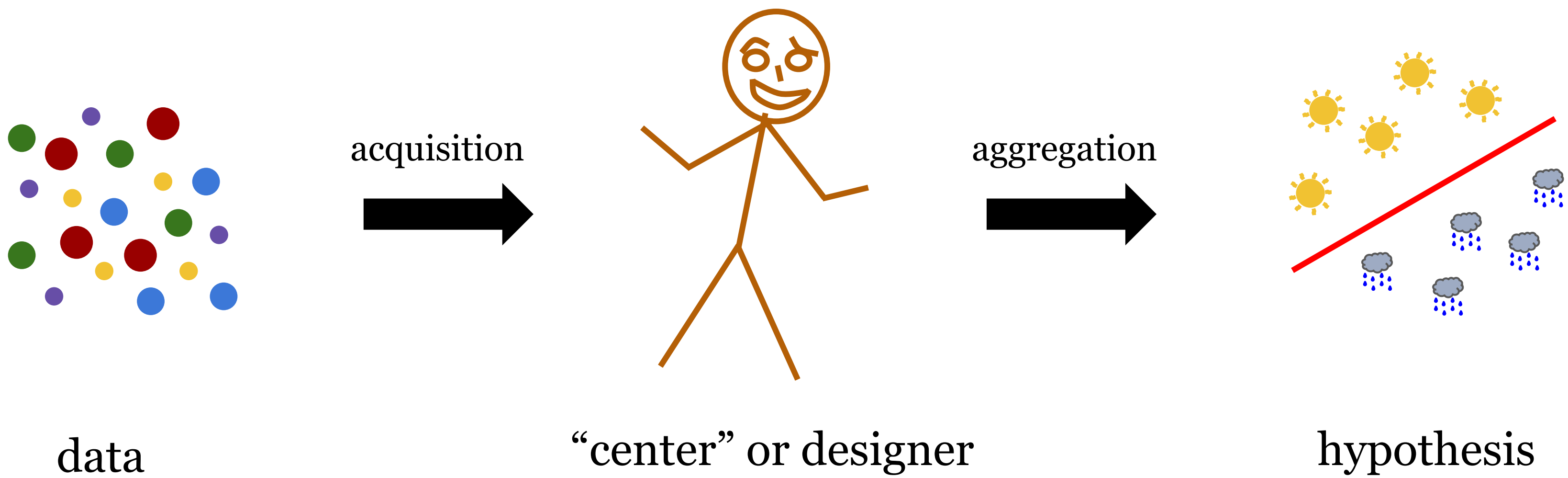
An important instance



*drawing not to scale

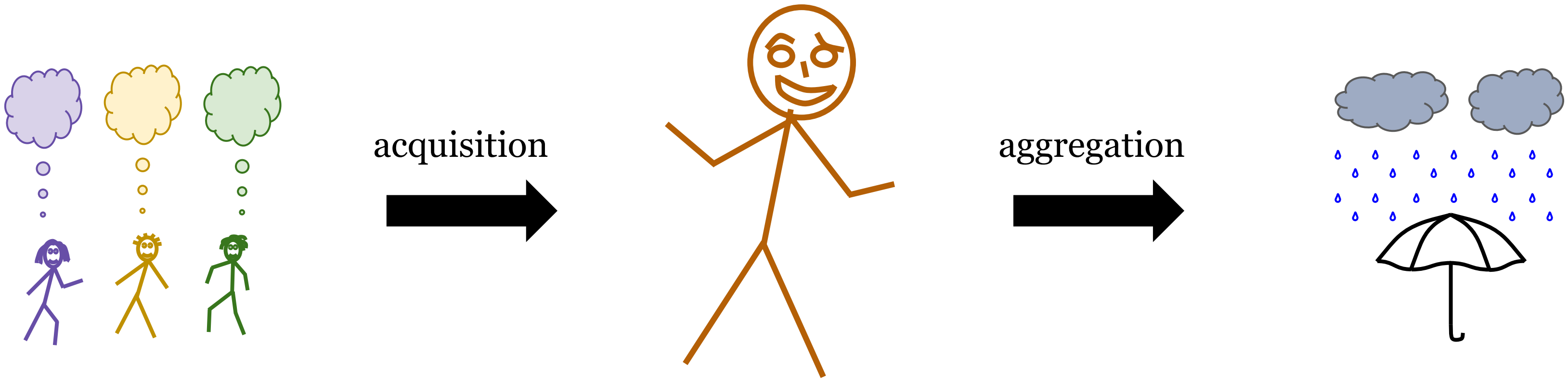
An important instance

Example: individuals' medical data, for predicting disease from features



*drawing not to scale

Another important instance



(expert) opinions and beliefs

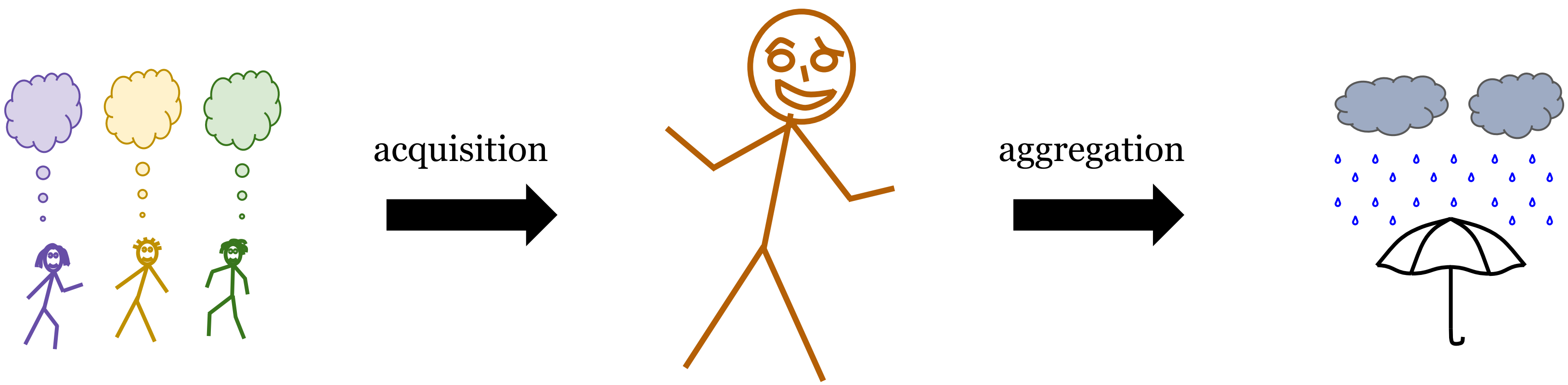
"center" or designer

prediction or decision

*drawing not to scale

Another important instance

Example: a prediction market for predicting whether a study on medical data will be replicated successfully.



(expert) opinions and beliefs

“center” or designer

prediction or decision

*drawing not to scale

Outline

1. **Approach #1:** Purchasing data for learning (main part of today's talk)

2. **Approach #2:** strategic aggregation of beliefs

3. **Discussion** and future directions

Outline

1. **Approach #1:** Purchasing data for learning (main part of today's talk)

2. **Approach #2:** strategic aggregation of beliefs

3. **Discussion** and future directions

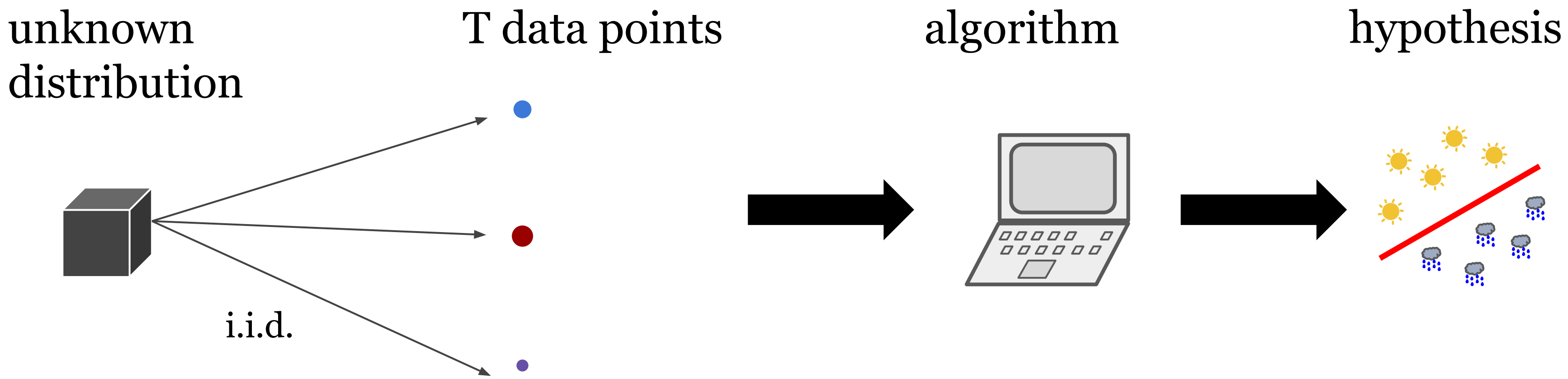
J Abernethy, Y Chen, C Ho, B Waggoner. **Low-Cost Learning via Active Data Procurement.** EC 2015.

Outline for "purchasing data"

1. Motivation, goal, and obstacles
2. Model, result, and approach
3. Discussion

J Abernethy, Y Chen, C Ho, B Waggoner. **Low-Cost Learning via Active Data Procurement.** EC 2015.

The machine-learning approach



Given: hypothesis class H, loss function *loss(h, z)* on hypothesis *h* and data point

Goal: minimize **“excess risk” (ER)**

ER := (expected loss of alg’s hypothesis) - (expected loss of optimal h)

(expectation over a new data point from that distribution)

The machine-learning approach

ER := (expected loss of alg's hypothesis) - (expected loss of optimal h)

Example result:
For binary classification, *loss(h, (x,y))* = 1 if h(x) = y and 0 otherwise,

$$\mathrm{ER} \leq O\left(\sqrt{\frac{\mathrm{VCdim(H)}}{T}}\right)$$

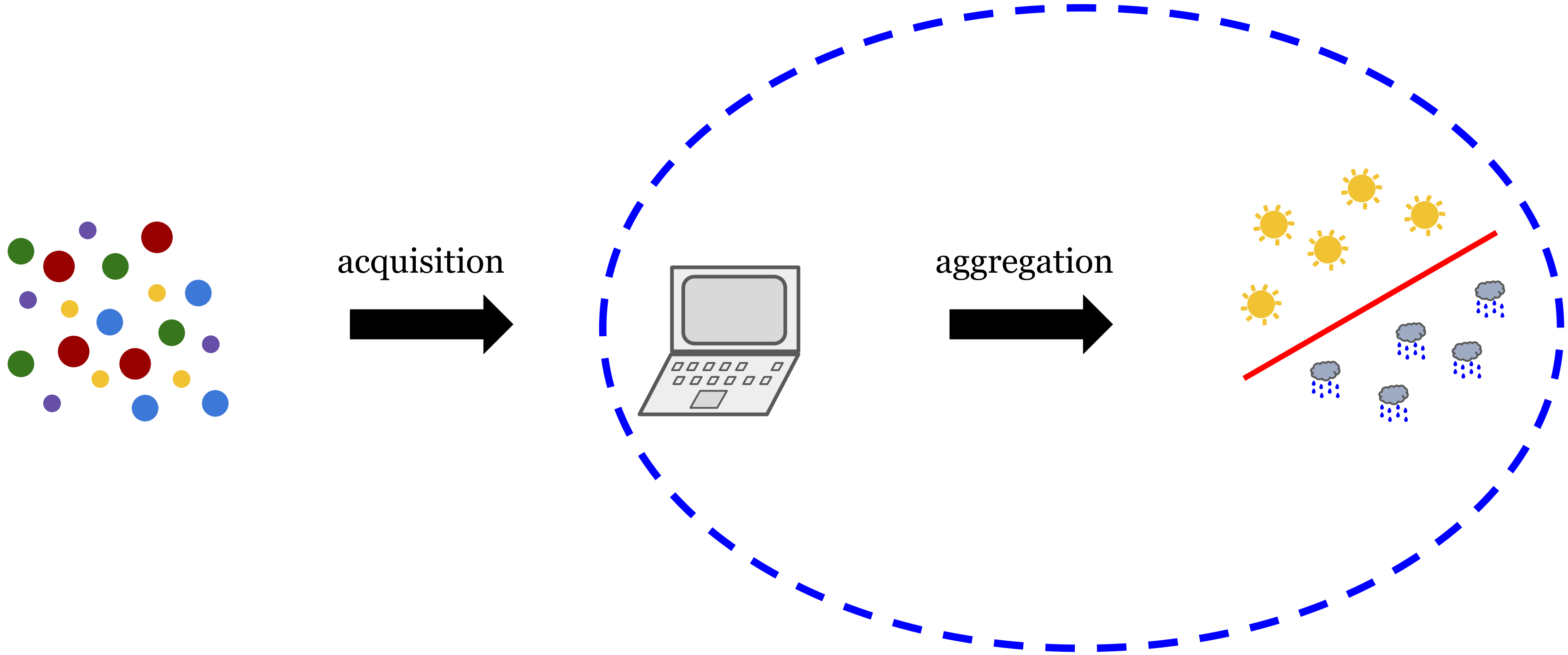
measure of complexity

quantity of resources

Some strengths of ML:
- very general and effective algorithms
- GE bounds capturing relationship of success to *complexity* and *resources*

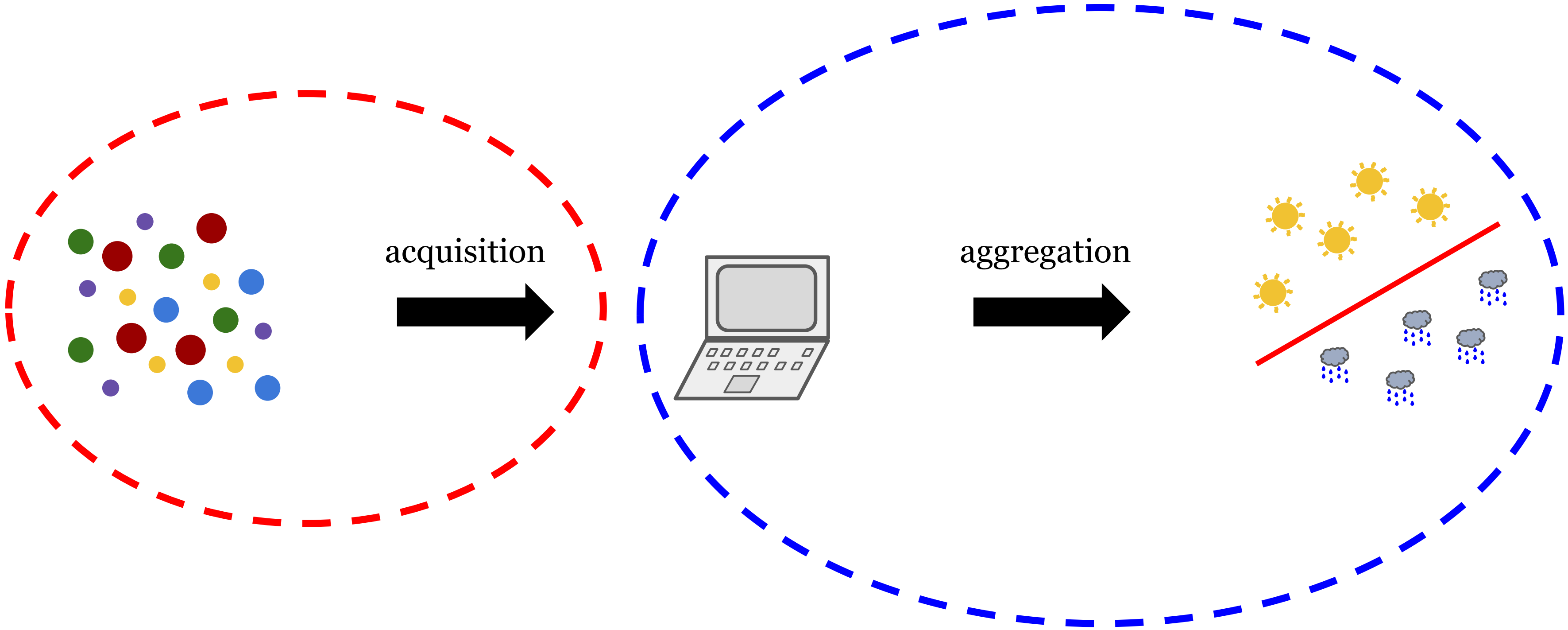
The gap in theory...

Here: machine learning theory is excellent



The gap in theory...

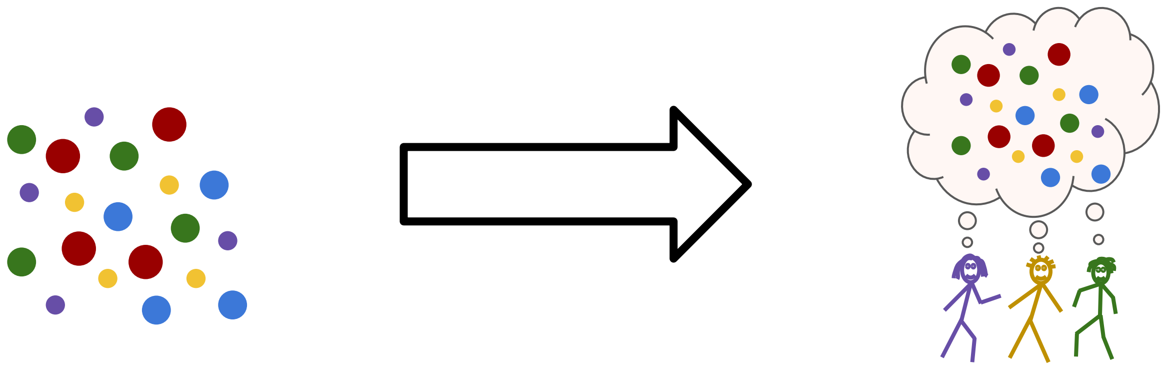
Here: machine learning theory is excellent



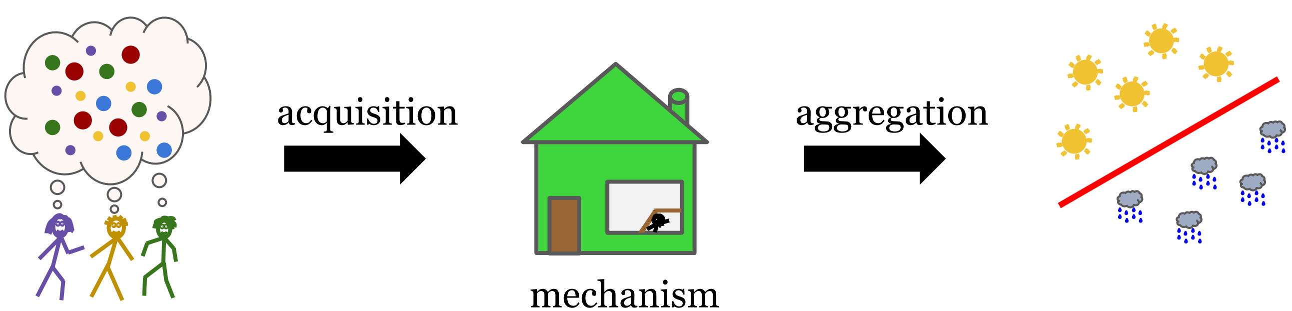
Here: extremely lacking!

Why is this a problem?

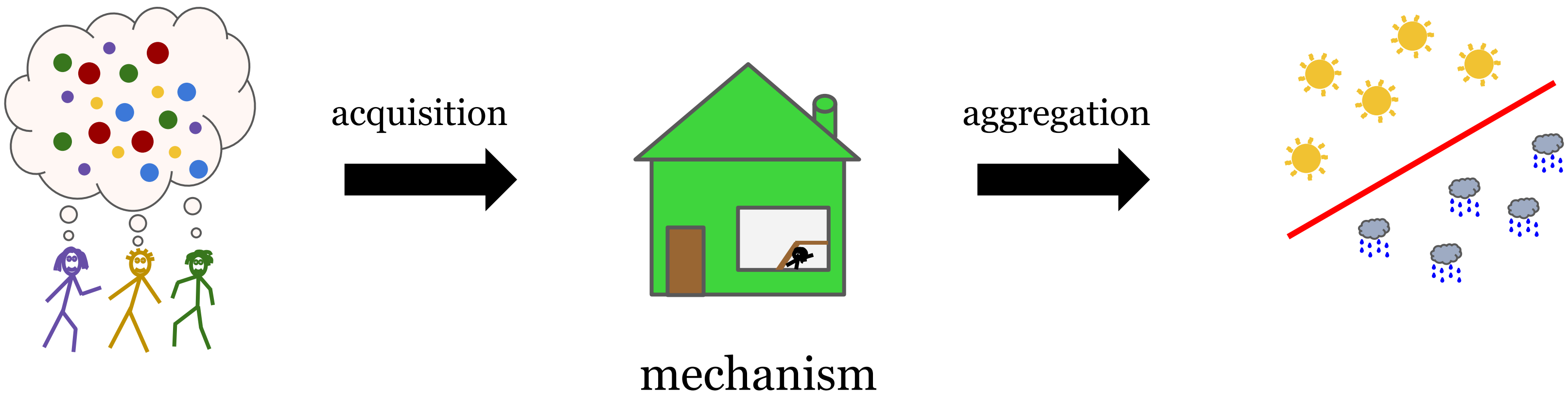
Often, data comes from *strategic agents*.



Challenge: design *mechanisms* to acquire and aggregate data.



What has been done?



Very exciting and active area!
Varied models and objectives: preserve privacy, principal-agent “effort” models, data may be falsifiable / not verifiable, ….

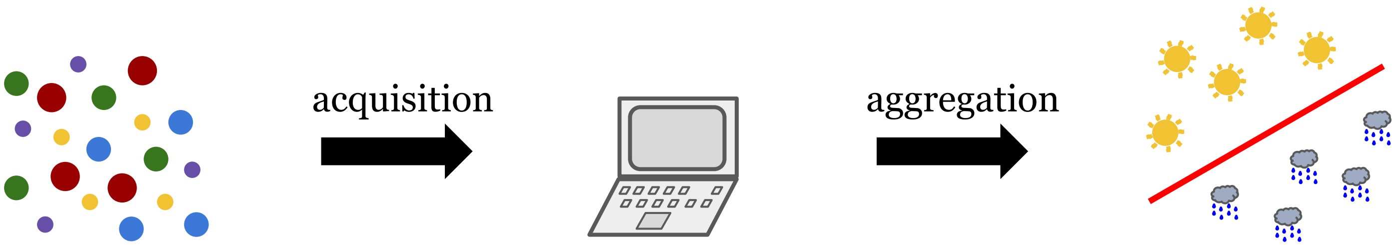
But the literature generally does not:

- offer solutions for generic loss functions
- leverage existing ML algorithms
- give bounds relating success, *complexity*, and *resources*

Roth, Schoenebeck EC 2012
Horel, Ioannidis, Muthukrishnan LATIN 2014
Ghosh, Roth EC 2011
Ligett, Roth WINE 2012
Cummings, Ligett, Roth, Wu, Ziani ITCS 2015
Cai, Daskalakis, Papadimitriou COLT 2015
Cummings, Ioannidis, Ligett COLT 2015
…

Two key goals for this field of research

(1) Given ML algorithm with ER bound "K"...



...produce *mechanism* with ER bound "f(K)".

(2) Understand properties of this new bound (in terms of *complexity* and *resources*).

Two key goals for this field of research

(1) Given ML algorithm with ER bound “K”…

Sneak peek: We’ll achieve these for one class of algorithms and an incomplete understanding of complexity.

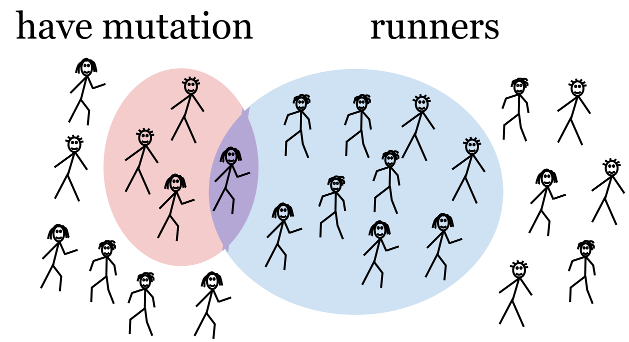
…produce *mechanism* with ER bound “f(K)”.

(2) Understand properties of this new bound (in terms of *complexity* and *resources*).

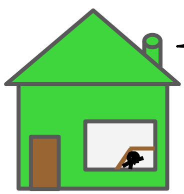
Obstacles / challenges

1. Relatively few data are **useful**

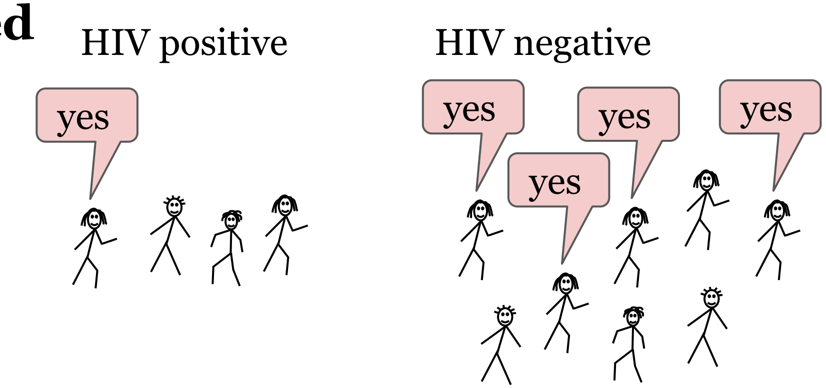
studying ACTN-3
mutation and running



2. Data and cost to reveal it may be **correlated**



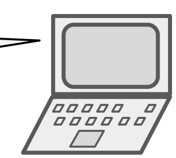
Want $10 to participate
in a study on HIV?



3. Usefulness of data (ML) and price paid (econ) live in **different worlds**

auctions, budgets,
reserve prices, value
distributions....

gradients, entropies, loss
functions, divergences...



Outline for “purchasing data”

1. Motivation, goal, and obstacles
2. Model, result, and approach
3. Discussion

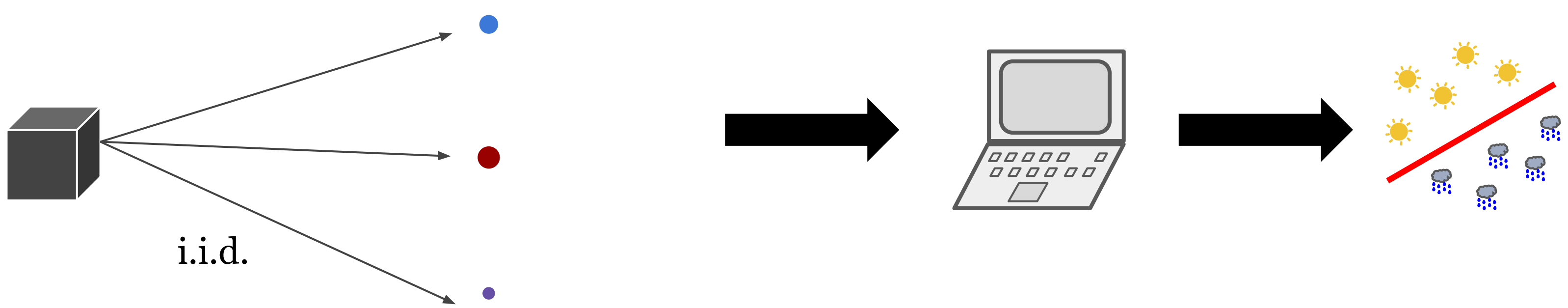
The classic statistical learning model

unknown distribution

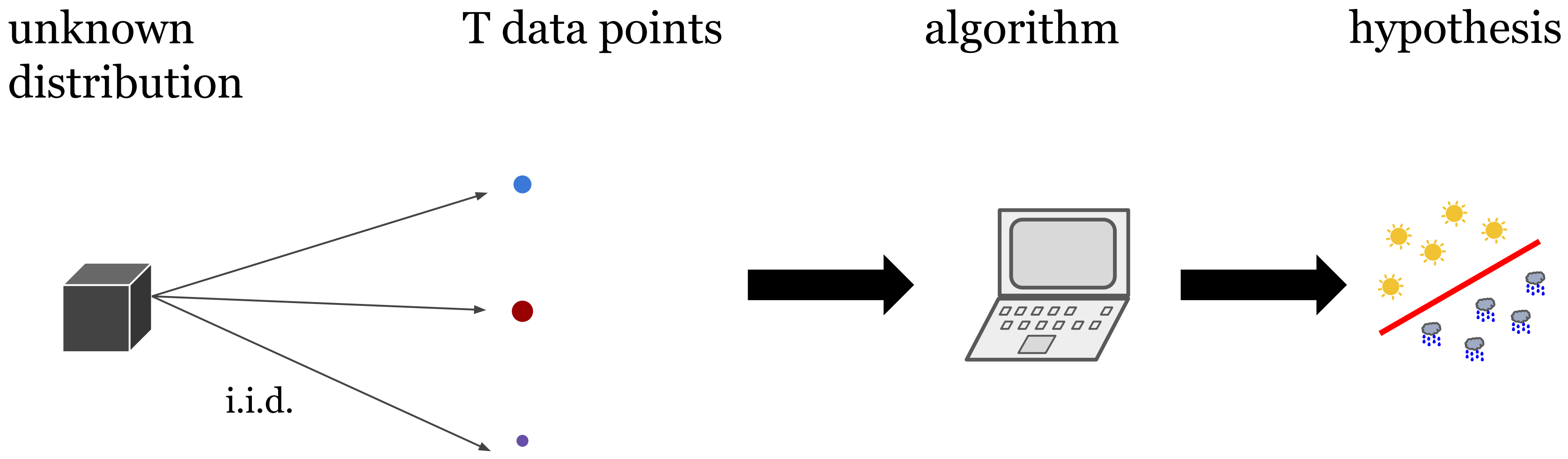
T data points

algorithm

hypothesis



The classic statistical learning model



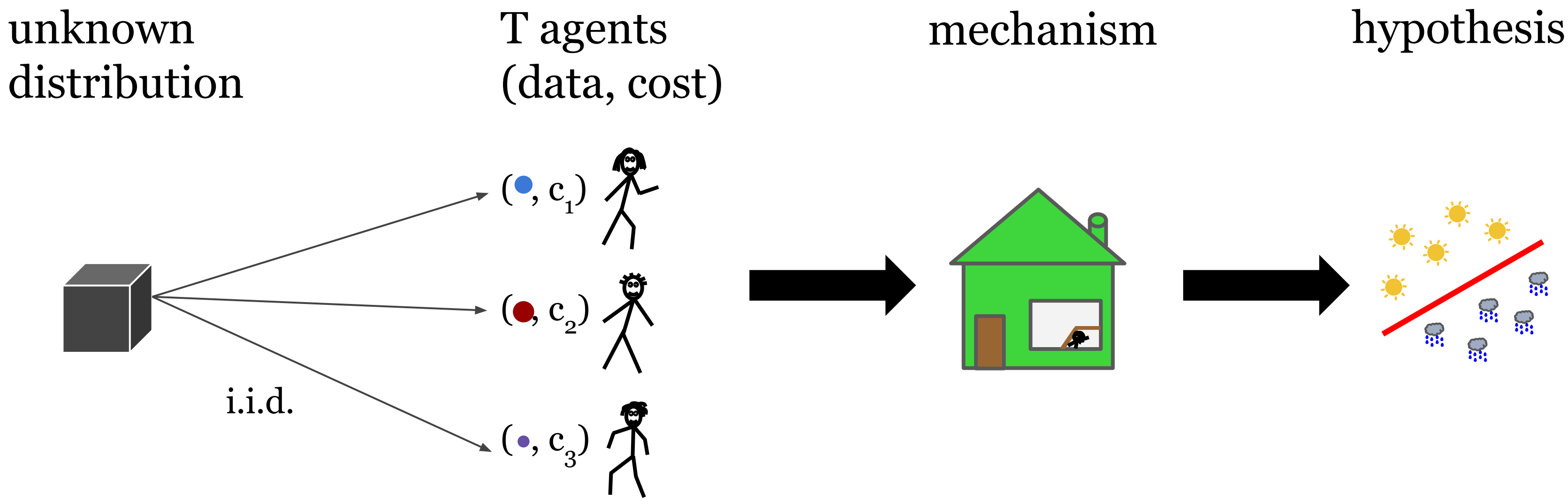
Follow-the-Regularized Leader (FTRL):

- hypothesis class is a Hilbert space (*e.g.* hyperplanes)
- loss function is Lipschitz and convex in h (*e.g.* hinge loss)
- processes data points online, outputting a hypothesis at each step

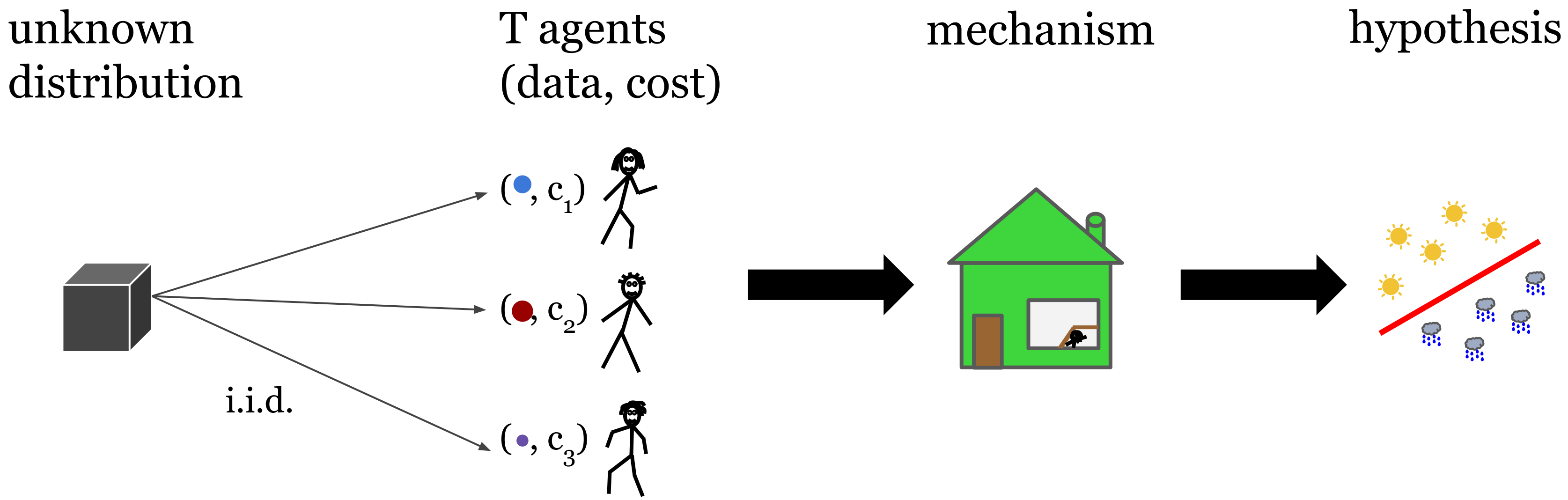
Regret: (total loss of these on arriving data) - (loss of optimal h in hindsight)

Classic FTRL result: "regret" $\leq O\left(\sqrt{T}\right)$, even if data is chosen adversarially.
Online-to-batch conversion $\Rightarrow$ ER $\leq O\left(1/\sqrt{T}\right)$.

A model that adds incentives



A model that adds incentives



In our model:

- agents arrive online
- costs may depend on the data arbitrarily (even chosen by an adversary)
- costs bounded in [0,1]
- model of cost: threshold "take-it-or-leave-it price" for which agent reveals data
- data cannot be fabricated or falsified

Our main result

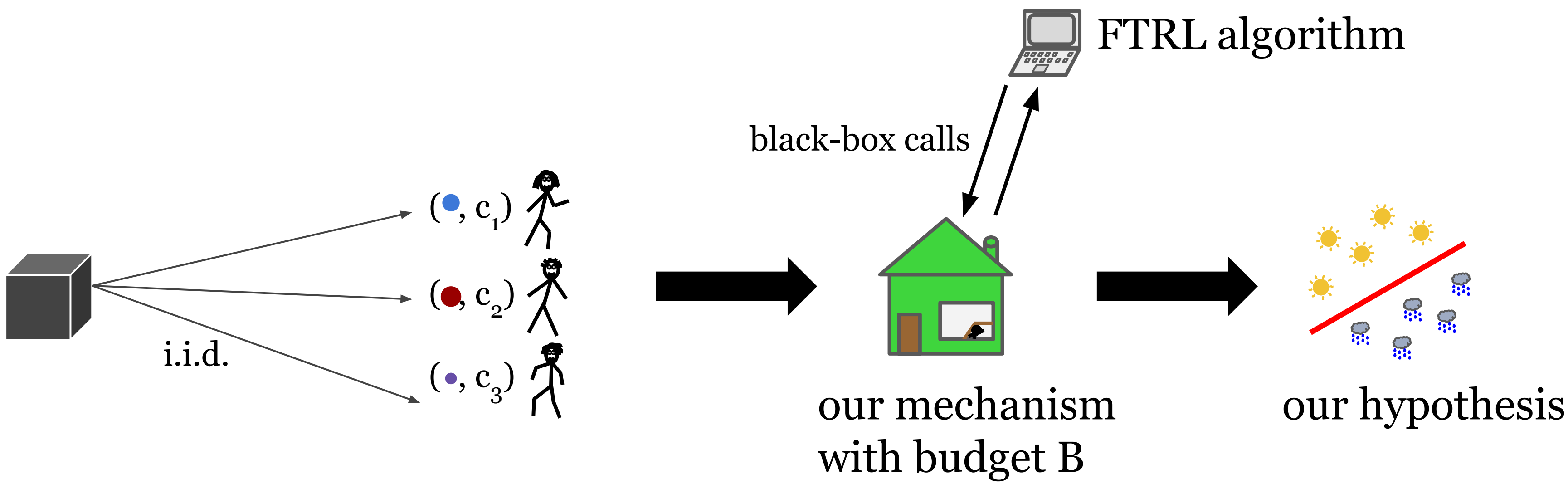
Given a Hilbert space of hypotheses, a Lipschitz convex loss function, and budget constraint B, our mechanism achieves excess risk

$$\mathrm{ER} \le O\left(\sqrt{\frac{\gamma}{B}}\right)$$

measure of complexity

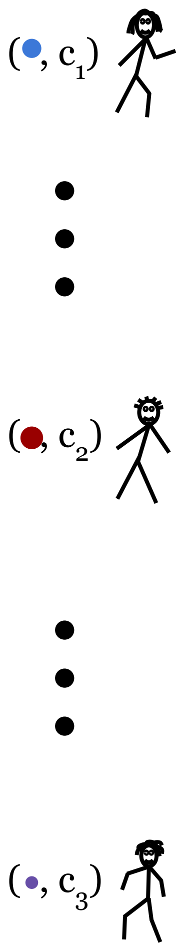
quantity of resources

where gamma in [0,1], to be discussed later.



How does the mechanism work?

agents arrive online

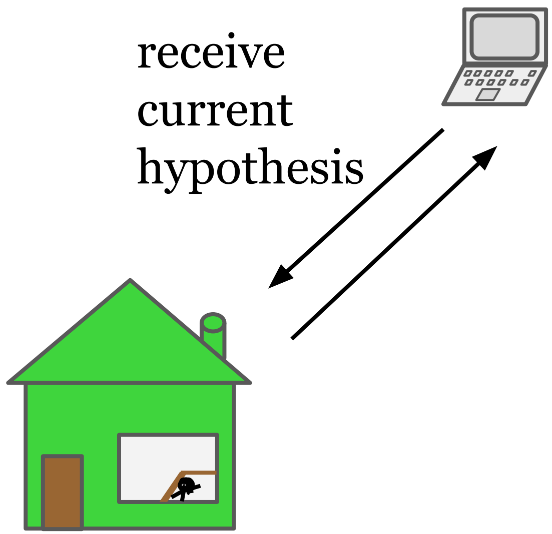


before each arrival,
post a take-it-or-leave-it
menu of prices

data	price
● (blue)	\$0.71
● (red)	\$0.38
	…

implicitly specified
by an algorithm

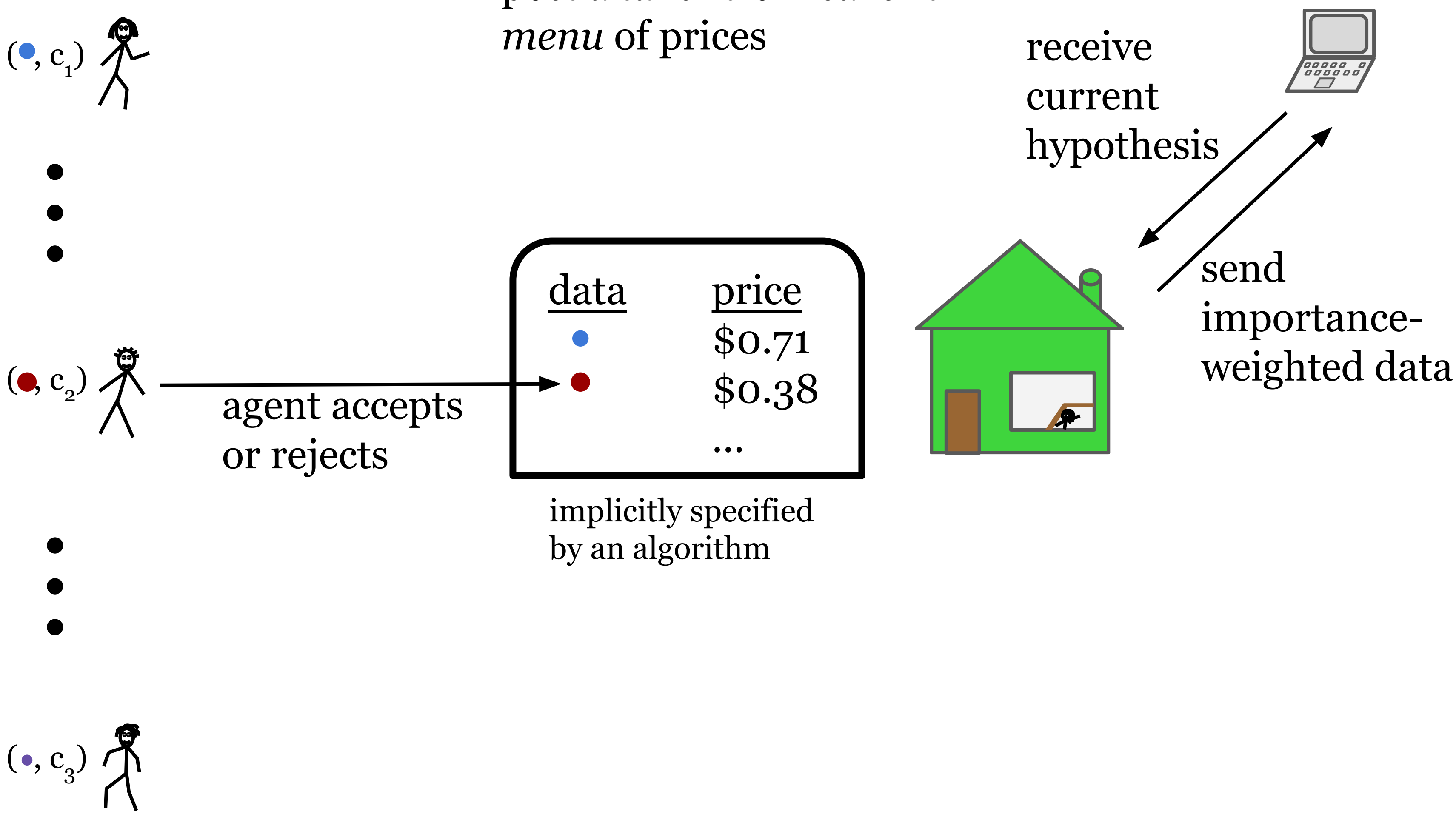
receive
current
hypothesis



How does the mechanism work?

agents arrive online

before each arrival,
post a take-it-or-leave-it
menu of prices



How does the mechanism work?

agents arrive online

(●, c_1)

⋮

(●, c_2)

agent accepts or rejects

⋮

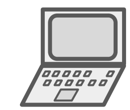
(•, c_3)

before each arrival,
post a take-it-or-leave-it
menu of prices

data	price
●	\$0.71
●	\$0.38
	…

implicitly specified
by an algorithm

receive
current
hypothesis



send
importance-
weighted data

Goal: show that these hypotheses have low regret on the data sequence. (online-to-batch ⇒ low excess risk)

How does the mechanism work?

agents arrive online

(•, c_1) 

•
•
•

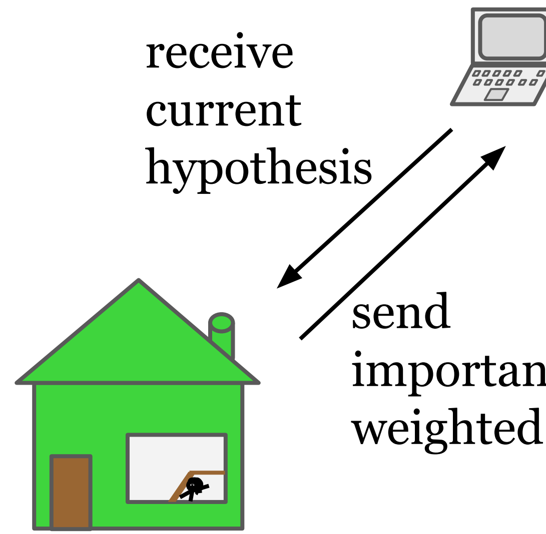
(•, c_2) 

agent accepts or rejects

before each arrival,
post a take-it-or-leave-it
menu of prices

data	price
•	$0.71
•	$0.38
	…

implicitly specified
by an algorithm



•
•
•

(•, c_3) 

How to choose the prices to post?

Roadmap: deriving the pricing strategy

1. Start from FTRL analysis for low regret.

2. Consider simple setting where all costs are 1.
   Prove regret guarantee.
   (Have matching lower bound.)

3. Consider simple setting where agents report costs truthfully to mechanism.
   Derive "optimal" price-posting strategy and prove regret guarantee.
   (Have matching lower bound.)

4. Leverage previous solution to get a regret guarantee for the general setting.
   (Gap to known lower bound -- *price of strategic behavior*!)

First step: the analysis of FTRL

FTRL: At time t, pick $h_t = \arg\min_h \sum_{s<t} \text{loss}(h, z_s) + \frac{G(h)}{\eta}$

where:

- z_s is the data point arriving at time s
- G is a strongly-convex function (called the "regularizer")
- η is a parameter to be chosen later

First step: the analysis of FTRL

FTRL: At time *t*, pick $h_t = \arg\min_h \sum_{s<t} \mathrm{loss}(h, z_s) + \frac{G(h)}{\eta}$

where:

- z_s is the data point arriving at time *s*
- *G* is a strongly-convex function (called the "regularizer")
- η is a parameter to be chosen later

Key regret lemma: show that regret $\leq O(1)/\eta + 2\eta \sum_t \Delta_t^2$
where $\Delta_t = \|\nabla \mathrm{loss}(h_t, z_t)\|$

Using the lemma: By assumption, Δ_t in [0,1].
Choose $\eta = 1/\sqrt{T}$ to get regret $\leq O\left(\sqrt{T}\right)$

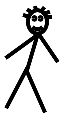
Can do better (sometimes): Imagine we knew in advance $g = \frac{1}{T} \sum_t \Delta_t^2$
Can choose $\eta = 1/\sqrt{\sum_t \Delta_t^2}$ to get regret $O\left(\sqrt{gT}\right)$

Second step: all costs are 1

Just make a yes/no decision on each data point.

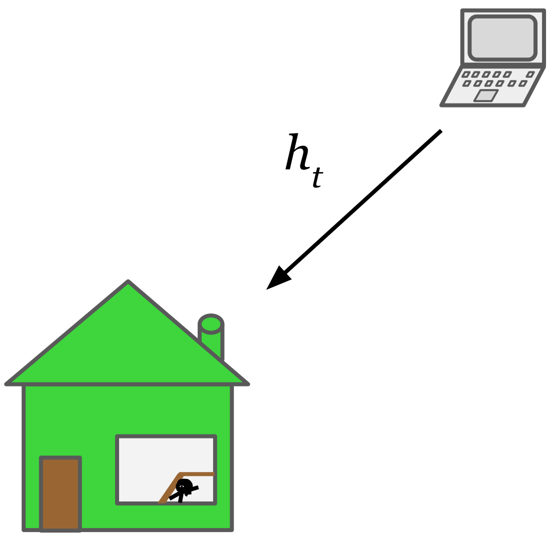
arriving agent

(●, 1)



today's menu

data	price
●	\$1
●	\$0
	...



Key idea: must decide randomly! (to defeat adversary)

data	Pr[samp]
●	0.34
●	0.15
	...

Second step: all costs are 1

Recall FTRL regret lemma: Regret $\leq O(1)/\eta + 2\eta \sum_t \Delta_t^2$
where $\Delta_t = \|\nabla \mathrm{loss}(h_t, z_t)\|$.

Challenge: not enough budget to purchase every data point.
(Must randomly subsample.)

Importance-weighted loss: given data point z when Pr[samp] = p,
send "importance-weighted" loss function $h \mapsto \frac{\mathrm{loss}(h,z)}{p}$.

"Importance-weighted" regret lemma:
Let q_t = Pr[sample arrival t]. Then for any choices of q_t,
by feeding FTRL "importance-weighted losses",

$$\text{regret} \leq O(1)/\eta + 2\eta \sum_t \frac{\Delta_t^2}{q_t}.$$

Second step: all costs are 1

Recall importance-weighted regret lemma:
by feeding FTRL importance-weighted losses (when data is obtained) and zeroes (otherwise), regret $\leq O(1)/\eta + 2\eta \sum_t \frac{\Delta_t^2}{q_t}$.

Result: Setting every $q_t = B/T$ and choosing $\eta = \sqrt{B}/T$ yields regret $\leq O\left(T/\sqrt{B}\right)$.
Lower bound: regret $\geq T/\sqrt{B}$ (identifying a slightly biased coin).

Imagine we could solve the following problem…

$$\min_q \sum_t \frac{\Delta_t^2}{q_t}$$
$$\text{s.t. } \sum_t q_t \leq B.$$

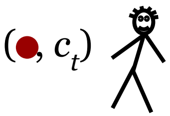
Actually, with a tiny bit of prior knowledge, we can! Choose $q_t \propto \Delta_t$.

Better result: With advance knowledge of $g' = \mathbb{E}\frac{1}{T}\sum_t \Delta_t$,
can achieve regret $\leq O\left(g'T/\sqrt{B}\right)$.

Third step: “at-cost”

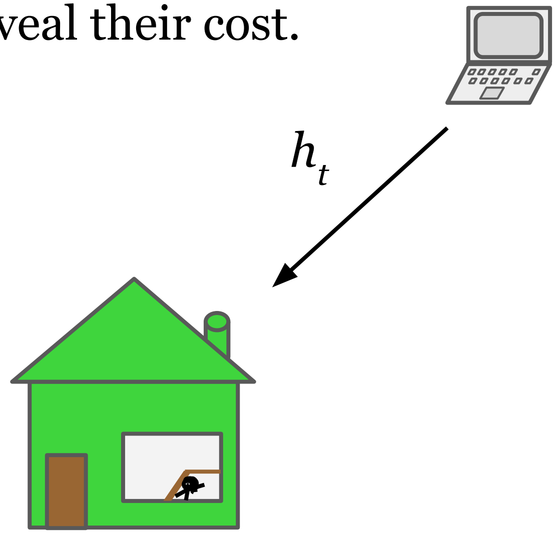
Suppose that: agents, when they arrive, truthfully reveal their cost.
(for purposes of analysis only)

arriving agent



today’s menu

data	Pr[post c_t]
● (blue)	0.55
● (red)	0.08
	...



Key idea: almost identical approach as when all costs were 1!

Result: With advance knowledge of $\gamma = \mathbb{E}\frac{1}{T}\sum_t \sqrt{c_t \Delta_t^2}$, by picking $q_t \propto \frac{\Delta_t}{\sqrt{c_t}}$ can achieve regret $\le O\left(\gamma T/\sqrt{B}\right)$.

Result: matching lower bound (see paper for details on what this means).

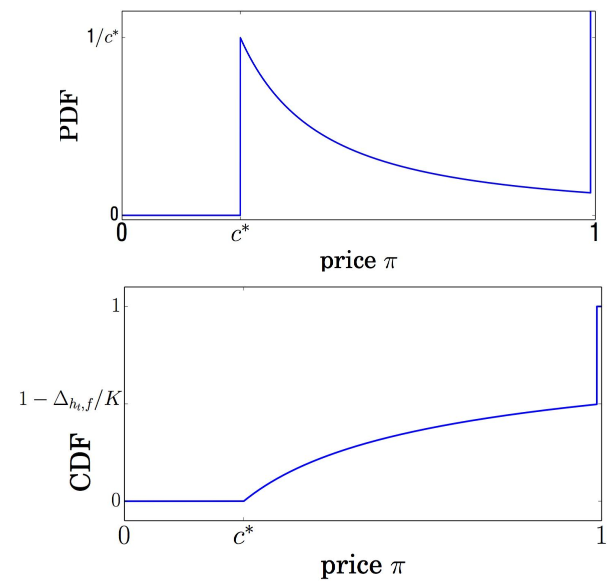
Final step: the price-posting distribution

What we'd like to do: obtain the data point with probability $q_t = \frac{\Delta_t}{K\sqrt{c_t}}$
Problem: the data and cost may be adversarially chosen.

Unfairly tricky-yet-simple insight: Draw a price according to cdf $F(x) = 1 - \frac{\Delta_t}{K\sqrt{x}}$

Why?? For *every* c_t, …

Result: With advance knowledge of $\gamma = \mathbb{E}\frac{1}{T}\sum_t \sqrt{c_t \Delta_t^2}$,
get regret $\leq O\left(\sqrt{\gamma}T/\sqrt{B}\right)$.



Note the loss versus the previous result:
cost due to strategic behavior!

(This loss is the gap between our upper and lower bounds…)

Outline for “purchasing data”

1. Motivation, goal, and obstacles
2. Model, result, and approach
3. Discussion

Revisiting the main result, discussion

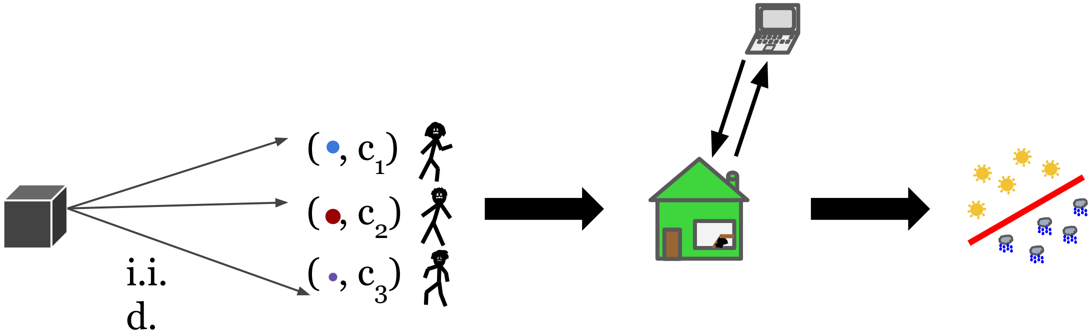
Given a Hilbert space of hypotheses, a Lipschitz convex loss function, budget constraint B, *and advance knowledge of gamma*, our mechanism achieves

$$\mathrm{ER} \leq O\left(\sqrt{\frac{\gamma}{B}}\right)$$

where $\gamma = \mathbb{E}\frac{1}{T}\sum_t \sqrt{c_t \Delta_t^2}$.

Feasibility of knowing gamma?

- Just a single scalar (compare to *e.g*. knowing marginal distribution of costs)
- In practice (and our simulations), gamma can be learned online
- Can replace gamma with any upper bound that is known, and get a corresponding ER guarantee. Example: gamma ≤ sqrt(average cost).



Discussion on meaning of result

Given a Hilbert space of hypotheses, a Lipschitz convex loss function, budget constraint B, *and advance knowledge of gamma*, our mechanism achieves

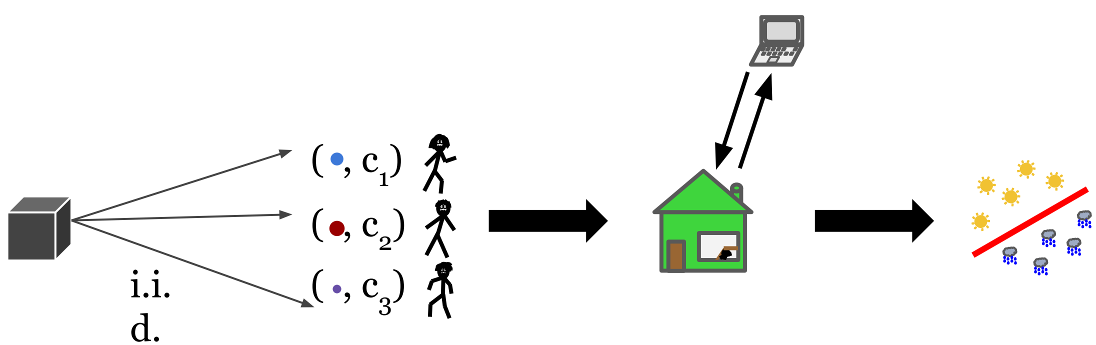
$$\mathrm{ER} \leq O\left(\sqrt{\frac{\gamma}{B}}\right)$$

where $\gamma = \mathbb{E}\frac{1}{T}\sum_t \sqrt{c_t \Delta_t^2}$

Recall: the FTRL algorithm that sees all T data points could "at best" guarantee ER $\leq O\left(\sqrt{\frac{g}{T}}\right)$ where $g = \frac{1}{T}\sum_t \Delta_t^2$.

Implications:

- gamma ≤ sqrt(average cost).
- gamma ≤ sqrt(average "difficulty").
- Can take advantage of beneficial correlations!

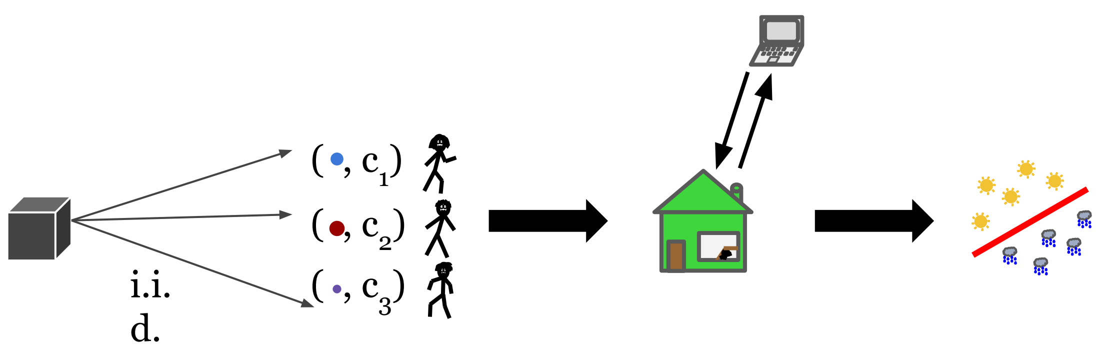


Recap: the key points

- Proposed a **model** of strategic data-holders grounded in statistical learning.
- Proposed mechanism **utilizing existing** FTRL learning algorithms.
- Proved **regret and ER bounds** as function of *“complexity”* and *budget*.
- We also saw:
  - a way to trade off algorithmic and monetary “value” of a data point
  - a “price of strategic behavior”: gap in bounds when agents maximize profit

Future directions:

- More models of strategic data holders
- Interface with more ML algorithms
- Better measures and understanding of “problem complexity”



Outline

1. **Approach #1:** Purchasing data for learning (main part of today's talk)
2. **Approach #2:** strategic aggregation of beliefs
3. **Discussion** and future directions

Y Chen, B Waggoner. **Informational Substitutes for Prediction and Play.** Working paper, 2016.

Motivation: strategizing in aggregation

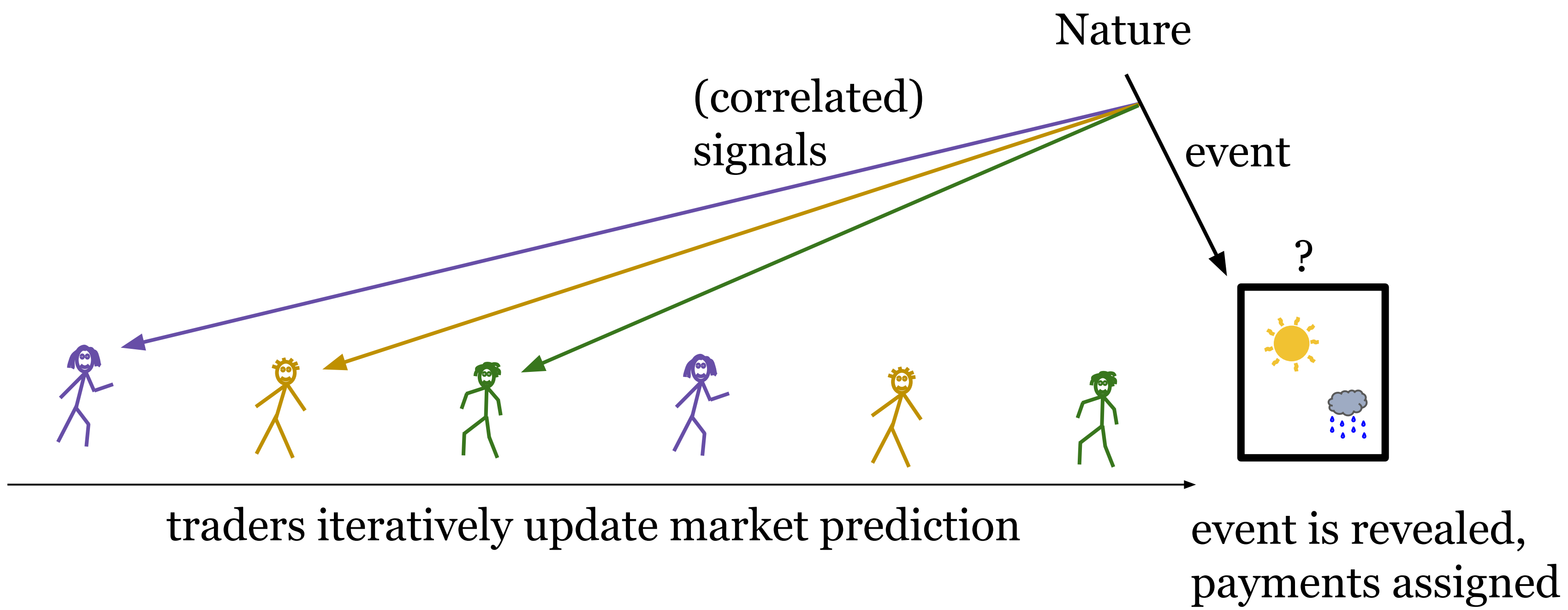
We don't understand how agents strategically reveal and aggregate information (even in relatively simple settings).

agents mechanism prediction

In my opinion:

- prediction markets are the simplest/cleanest model for studying this problem
- we know almost nothing about information aggregation in prediction markets!

Prediction market model



Payment for changing prediction from p to p' with outcome is $S(p', \ ) - S(p, \ )$, where S is any proper scoring rule.

Ex: the popular “log” scoring rule is $S(p, \ ) = \log p(\ )$.

Prior work on aggregation in markets

- Chen, Reeves, Pennock, Hanson, Fortnow, Gonen, WINE 2007:
  For the log scoring rule, if signals are conditionally independent, information is "immediately" aggregated.

- Dimitrov, Sami, EC 2008:
  For the log scoring rule, information is not always immediately aggregated.

- Gao, Zhang, Chen, EC 2013:
  For the log scoring rule, if signals are independent, information is aggregated "as late as possible".

Our results

We propose a definition of informational substitutes and complements.
For *every* scoring rule and information structure,

- information is "immediately" aggregated if and only if signals are **substitutes**.
- information is aggregated "as late as possible" if and only if signals are **complements**.

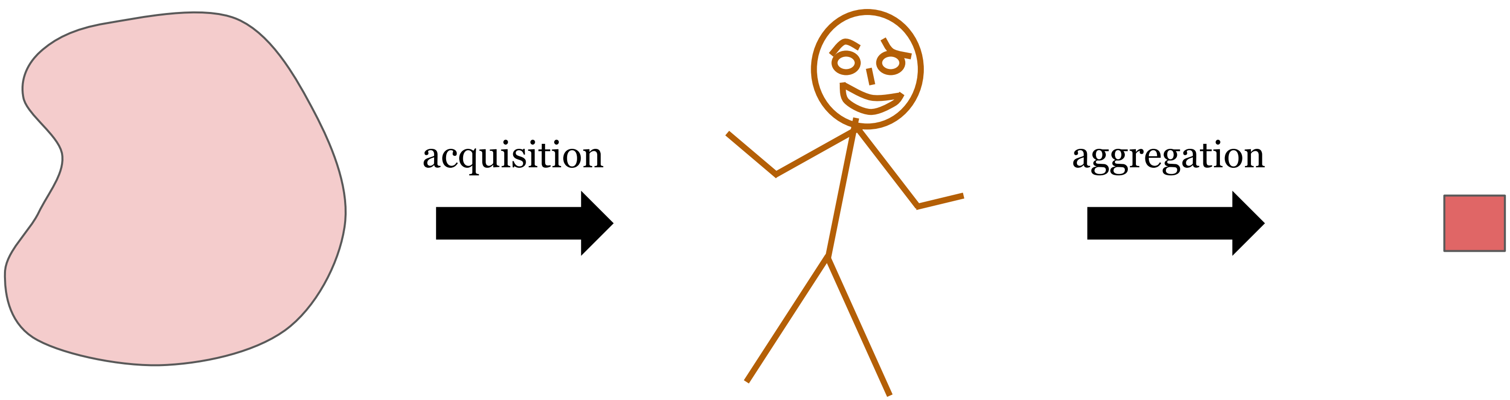
Prior results are special cases for the log scoring rule (easy to show).

Sidenote: definitions have natural characterizations, algorithmic applications....

Outline

1. **Approach #1:** Purchasing data for learning (main part of today's talk)
2. **Approach #2:** strategic aggregation of beliefs
3. **Discussion** and future directions

Recall the problem, and two approaches

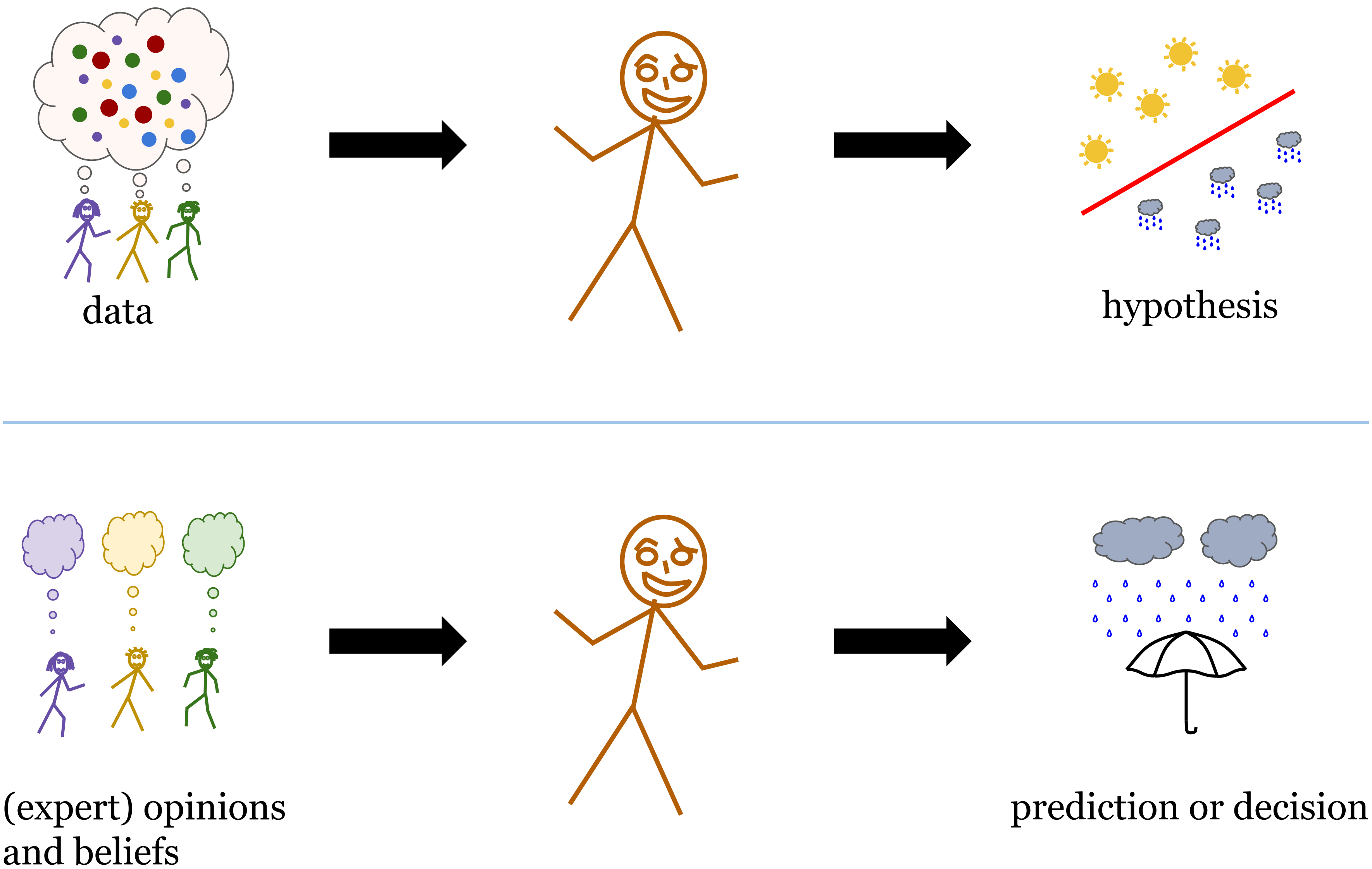


information

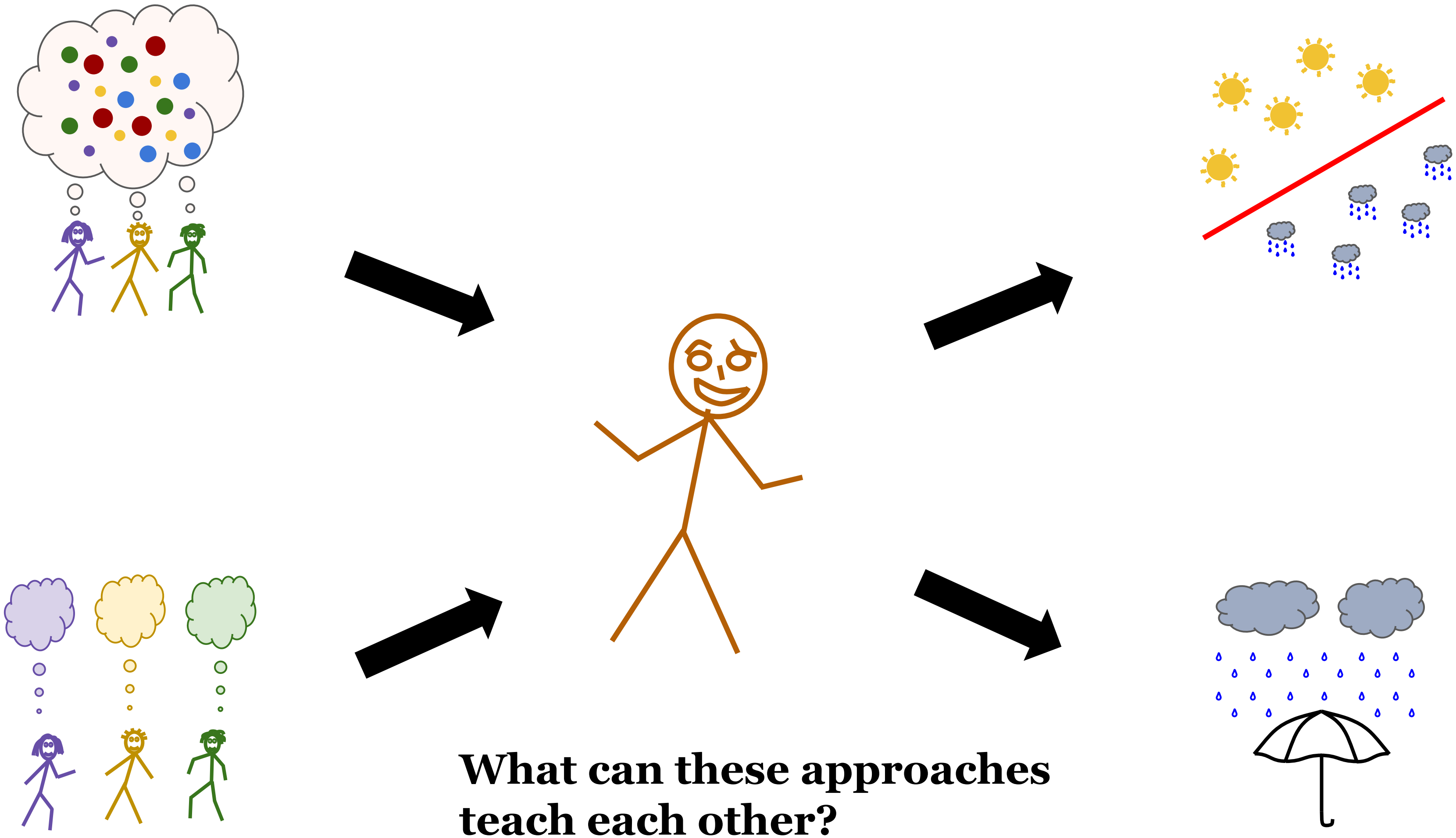
"center" or designer

useful summary

Recall the problem, and two approaches



Challenge going forward



What can these approaches teach each other?

An illustrative mechanism

Example: linear regression.
Goal: accurately predict a test data point using $y = ax + b$.

Market Framework:

1. Designer chooses initial parameters a,b.

2. Traders arrive, iteratively update parameters to a', b'.

3. Designer draws a test data point (x,y).
   Each update gets paid $loss(a,b, x,y) - loss(a',b', x,y)$, where
   $loss(a,b, x,y) = (y - (ax+b))^2$.

An illustrative mechanism

Example: linear regression.
Goal: accurately predict a test data point using $y = ax + b$.

Market Framework:

1. Designer chooses initial parameters a,b.

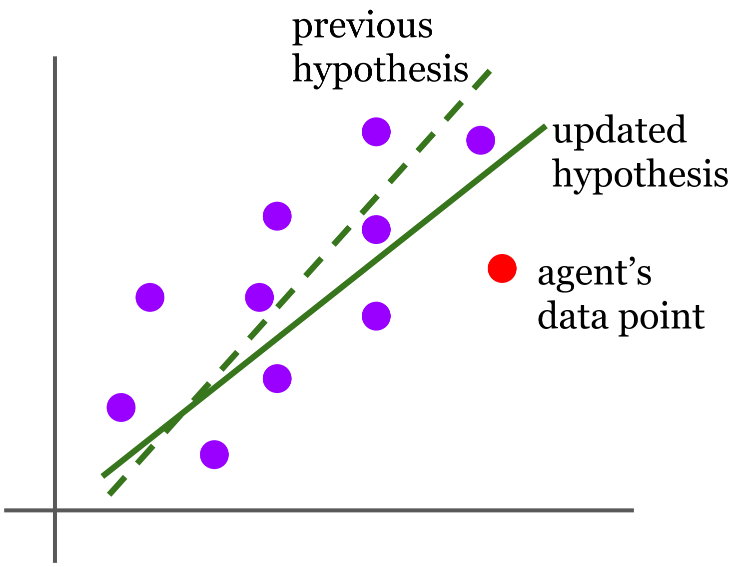
2. Traders arrive, iteratively update parameters to a', b'.

3. Designer draws a test data point (x,y).
   Each update gets paid $loss(a,b, x,y) - loss(a',b', x,y)$, where $loss(a,b, x,y) = (y - (ax+b))^2$.

Note: First proposed in Abernethy-Frongillo NIPS 2011.
Updated to add differential privacy for traders, other features in Waggoner-Frongillo-Abernethy NIPS 2015.

An illustrative mechanism

- What if traders just have **data** rather than **beliefs**?
- **Easy!** Run one iteration of a learning algorithm on their data point(s). Use its output as the updated market hypothesis.
- If data point was drawn i.i.d. from the underlying distribution, trader can *a priori* expect to make a profit.



Market framework:

1. Designer picks (a,b)
2. Traders update to (a',b')
   (repeat)
3. Designer draws test data, pays by improvement in loss

Raises questions pointing at future work

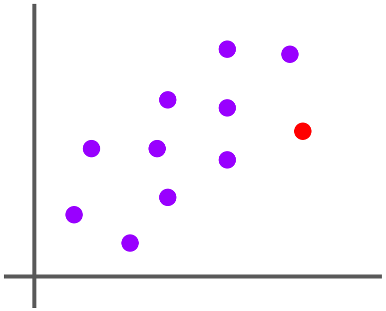
This mechanism accepts both kinds of inputs -- data and beliefs.
But it raises more questions than it answers ...

Q: What does "truthfulness" mean for this mechanism? Is it achieved?

Q: Where is the line between data and beliefs in this setting?

Q: To what extent is this a learning algorithm iteratively updating versus a mechanism relying on agents to aggregate?

→ Each of these questions points at a direction for future work!

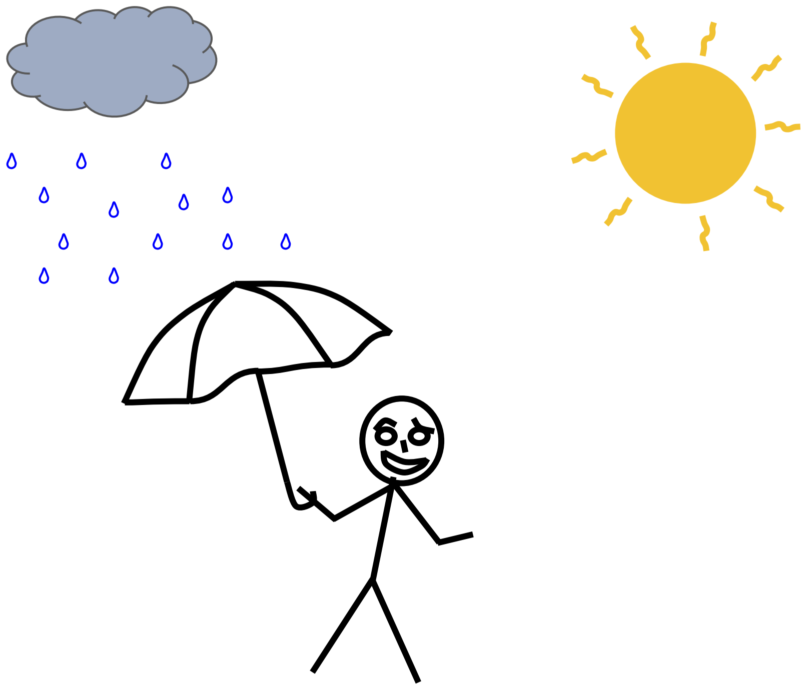


Conclusion: toward the future

- Machine learning *must* deal with strategic data.
  Not just to guarantee good learning bounds, but due to privacy, user control, efficient use of financial resources, ...

- Mechanisms for belief aggregation must deal with structure of information.
  Hopefully structure such as substitutes allows us to leverage algorithms to help.

- Mechanisms of the future should draw on the strengths of both approaches.

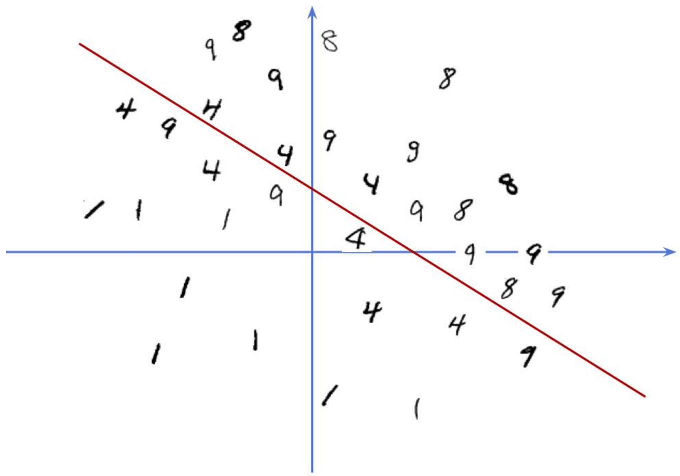


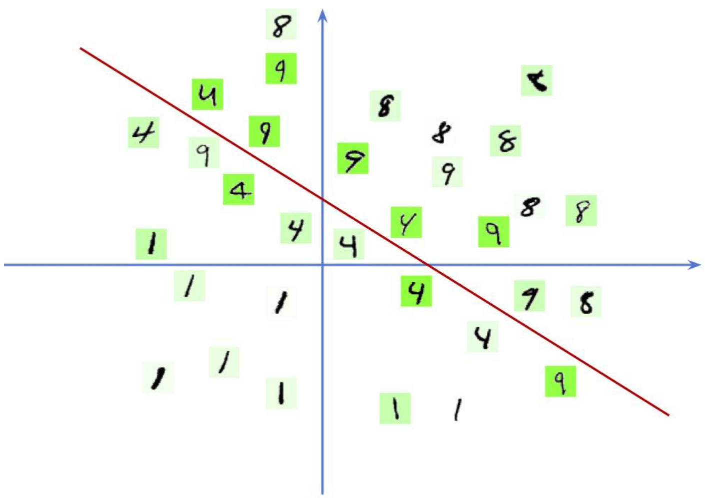
Thanks!

Additional slides

Simulation results

MNIST dataset -- handwritten digit classification



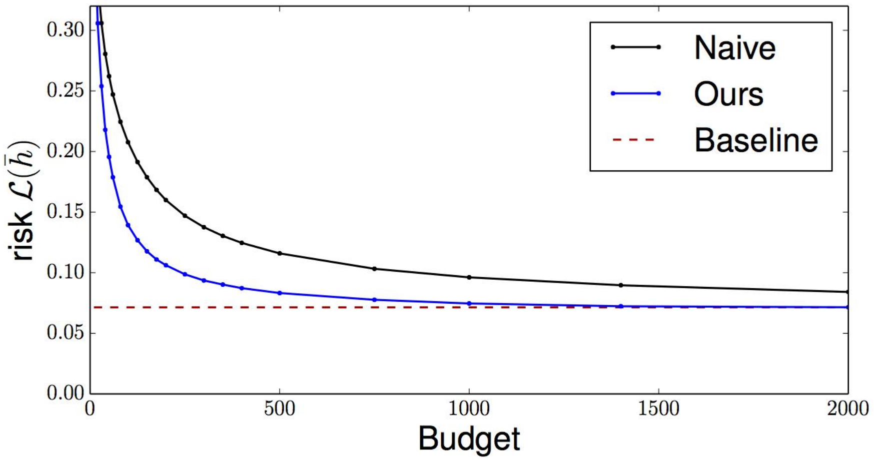


Toy problem: classify (1 or 4) vs (9 or 8)

Brighter green = higher cost

Simulation results

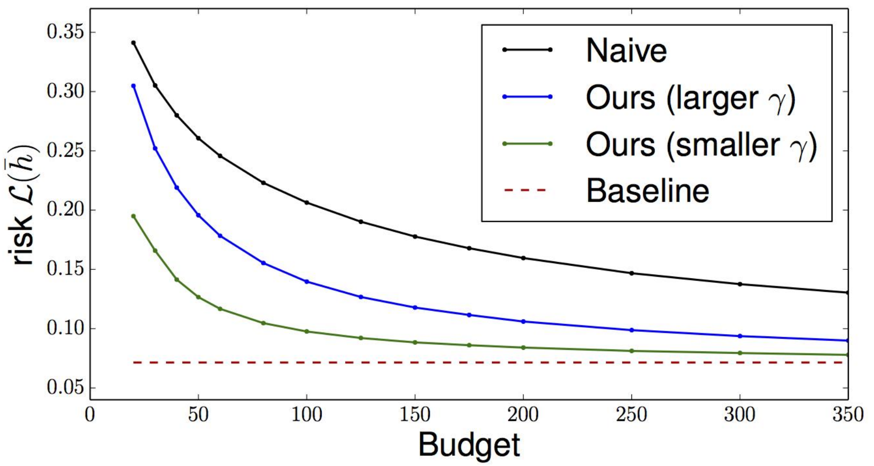
- T = 8503
- train on half,
  test on half
- Alg: Online Gradient Descent



Naive: pay 1 until budget is exhausted, then run alg

Baseline: run alg on all data points (no budget)

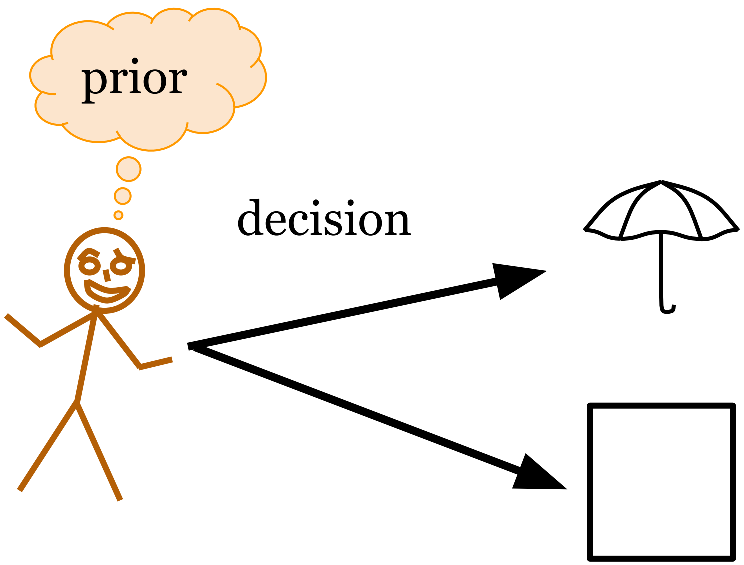
Large γ: bad correlations
Small γ: independent cost/data



Defining informational substitutes

(Much harder to define than substitutable goods!)

Question: What is the “value” of information in the first place?
A: given a *decision problem*, the expected utility to observe that signal before acting.

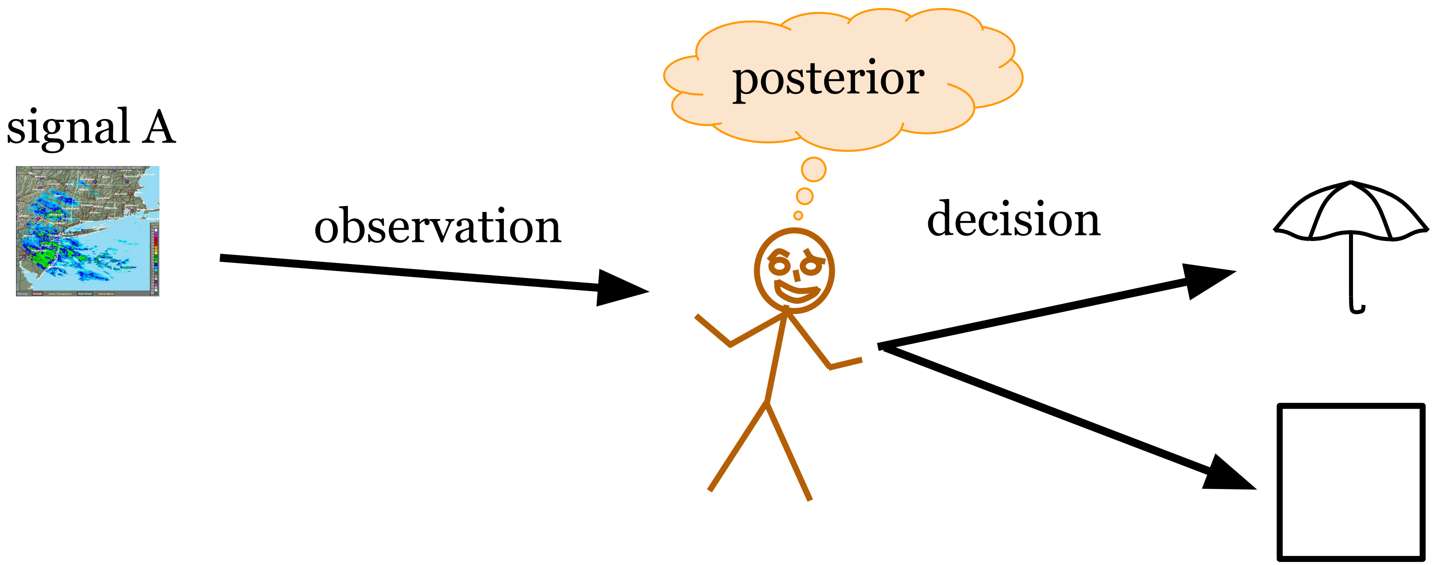


V(∅) = expected utility when observing no signals before deciding

Defining informational substitutes

(Much harder to define than substitutable goods!)

Question: What is the “value” of information in the first place?
A: given a *decision problem*, the expected utility to observe that signal before acting.

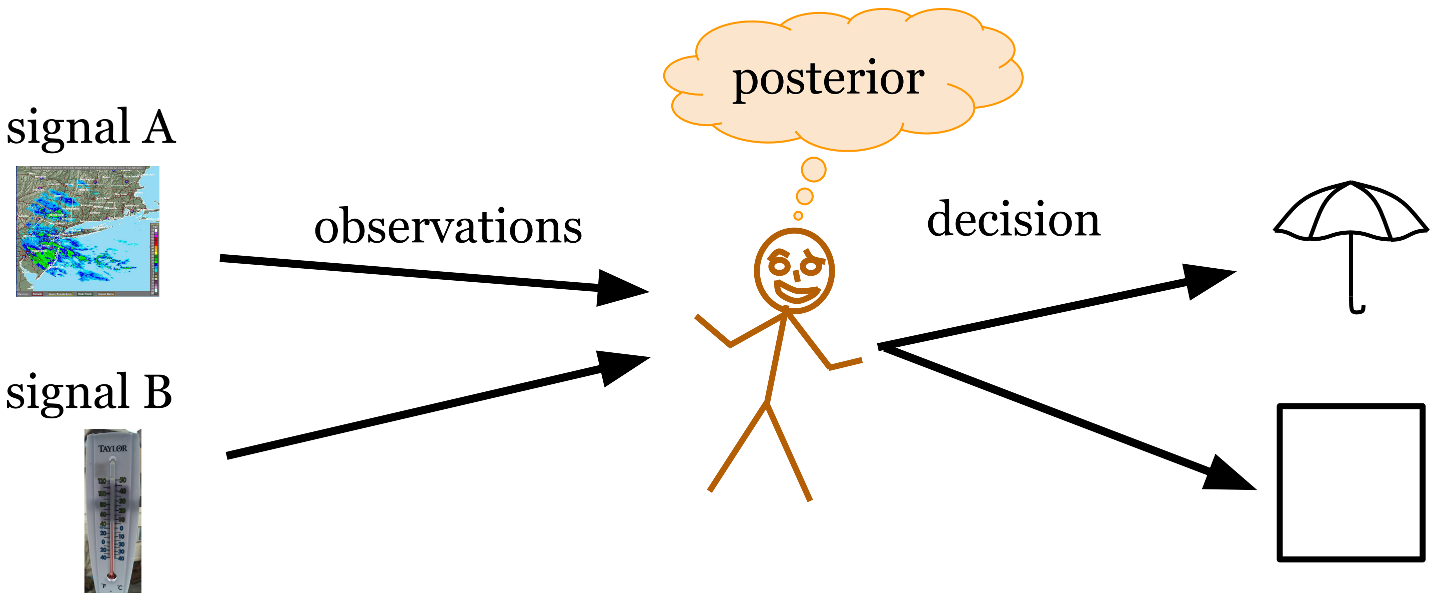


V(A) = expected utility for observing A, then deciding
V(A) - V(∅) = marginal value of A

Defining informational substitutes

(Much harder to define than substitutable goods!)

Question: What is the "value" of information in the first place?
A: given a *decision problem*, the expected utility to observe that signal before acting.



V(A,B) = expected utility for observing A and B, then deciding
V(A,B) - V(A) = marginal value of B if already observing A

Defining informational substitutes

Definition: Signals A and B are **substitutes** with respect to a particular decision problem if the *marginal value* of B *diminishes* with knowledge of A:

$$V(A,B) - V(A) \leq V(B) - V(\varnothing) .$$

and analogously with roles reversed.

Example: Say I only choose umbrella if [rainy and cold] or [sunny and warm]. Then radar map and thermometer reading are complements.

But: When choosing clothes for a run, these two signals are substitutes!

Some nice facts about substitutes

- A set of signals are substitutes iff expected utility is a **submodular** function on a (continuous) lattice defined over the signals.

- Consider the amount of **"bits" of information** a signal reveals about an event. A and B are substitutes iff the amount revealed by B *diminishes* given A.

- Consider the **"distance" moved by Bayesian updating** a distribution on B. A and B are substitutes iff this distance diminishes given A.

- **Algorithmic application**: how to choose what signals to purchase under constraints? (1-1/e)-approximation for substitutes; hard in general.

Algorithmic application of substitutes

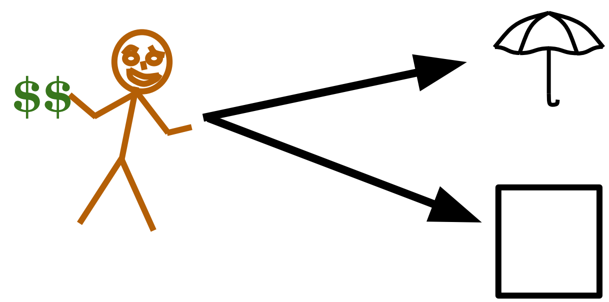
Input:

- decision problem
- set of signals A,B,... with prices π_A, π_B, ...
- Budget constraint

Output:
set of signals to purchase
maximizing utility,
subject to budget constraint

$0.31 

$0.09 



Our result: Substitutes ⇒ efficient (1-1/e)-approximation algorithm.

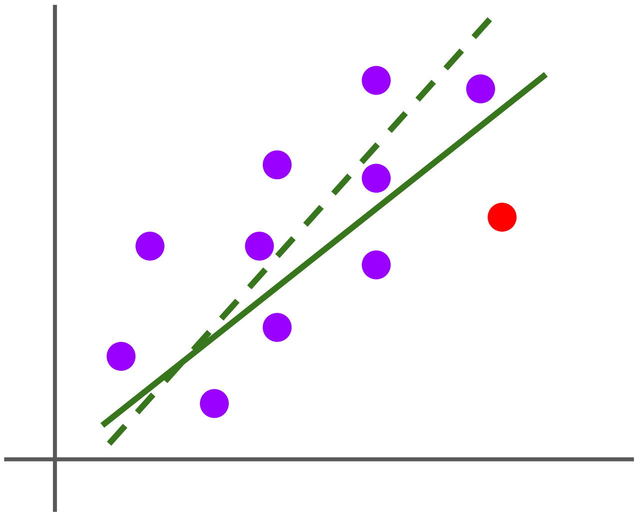
(Generalizes approach/results of Guestrin, Krause, Singh, ICML 2005 and related literature.)

Some further notes about WFA-NIPS'15

- Allows market to minimize any divergence-based loss function.
  Extends to nonparametric hypotheses via sample-based scoring rules of Zawadzki and Lahaie, AAAI 2015.

- (beautiful connections to exponential-family distributions as in above paper)

- Can ensure differential privacy for traders' data / updates if of bounded size, via adaptation of "continual observation". (Works for nonparametric hypotheses when combined with Hall, Rinaldo, Wasserman, JMLR 2013.)



Market framework:

1. Designer picks (a,b)
2. Traders update to (a',b')
   (repeat)
3. Designer draws test data,
   pays by improvement in loss