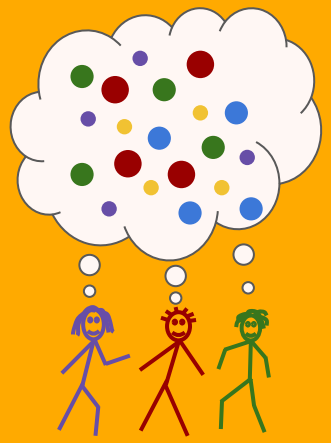
Acquiring and Aggregating Information from Strategic Sources



Bo Waggoner

PhD Defense, Harvard Computer Science
advised by Yiling Chen
May 2016

This PhD made possible by…

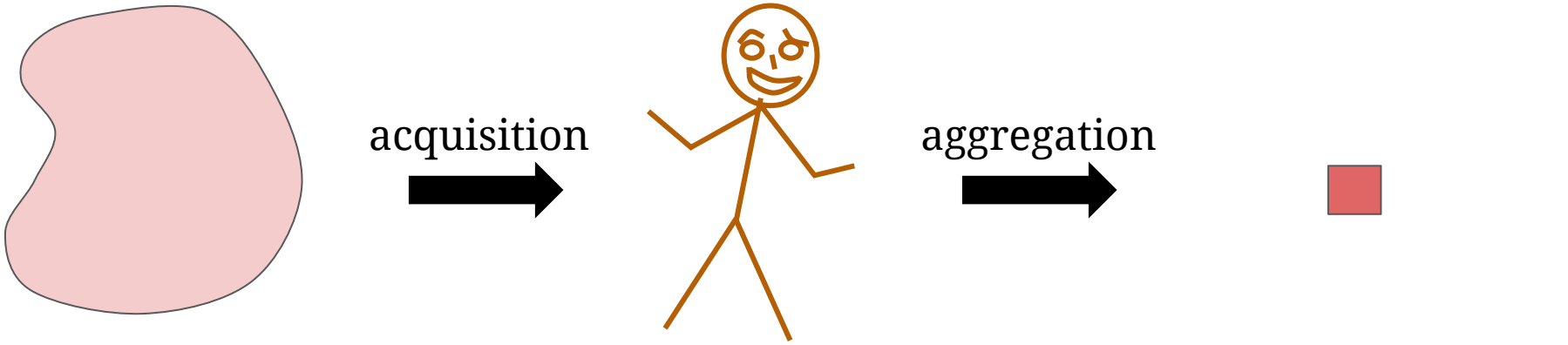
Advisor 

and committee:

Collaborators and mentors:

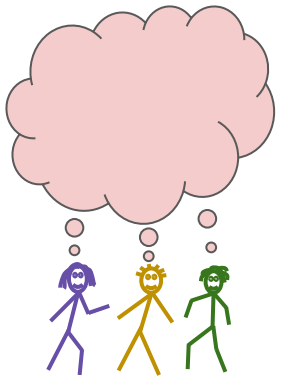
Friends, family, Priya, mentors, colleagues, coaches, Duke, Harvard, Google, Microsoft, Siebel Foundation, taxpayers, chocolate, electricity, mitochondria, …

A common pattern in theory and practice



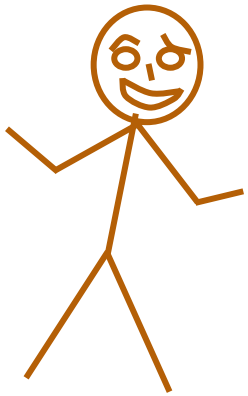
*drawing not to scale

This thesis: info is held by strategic agents



information

acquisition

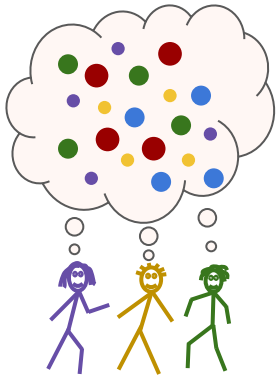


designer

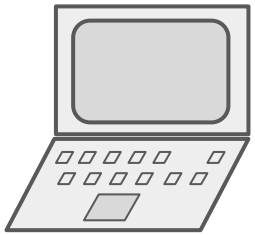
aggregation

useful summary

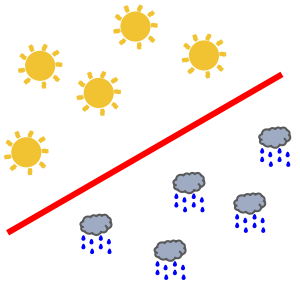
Case #1: data and hypotheses



acquisition



aggregation



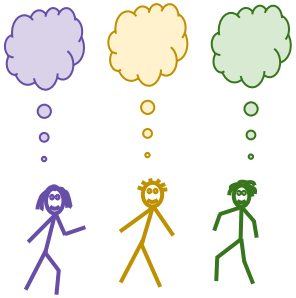
information
= data

algorithm

useful summary
= a hypothesis

How to A&A data controlled by strategic agents into a machine-learning hypothesis?

Case #2: expert beliefs and prediction



acquisition

aggregation

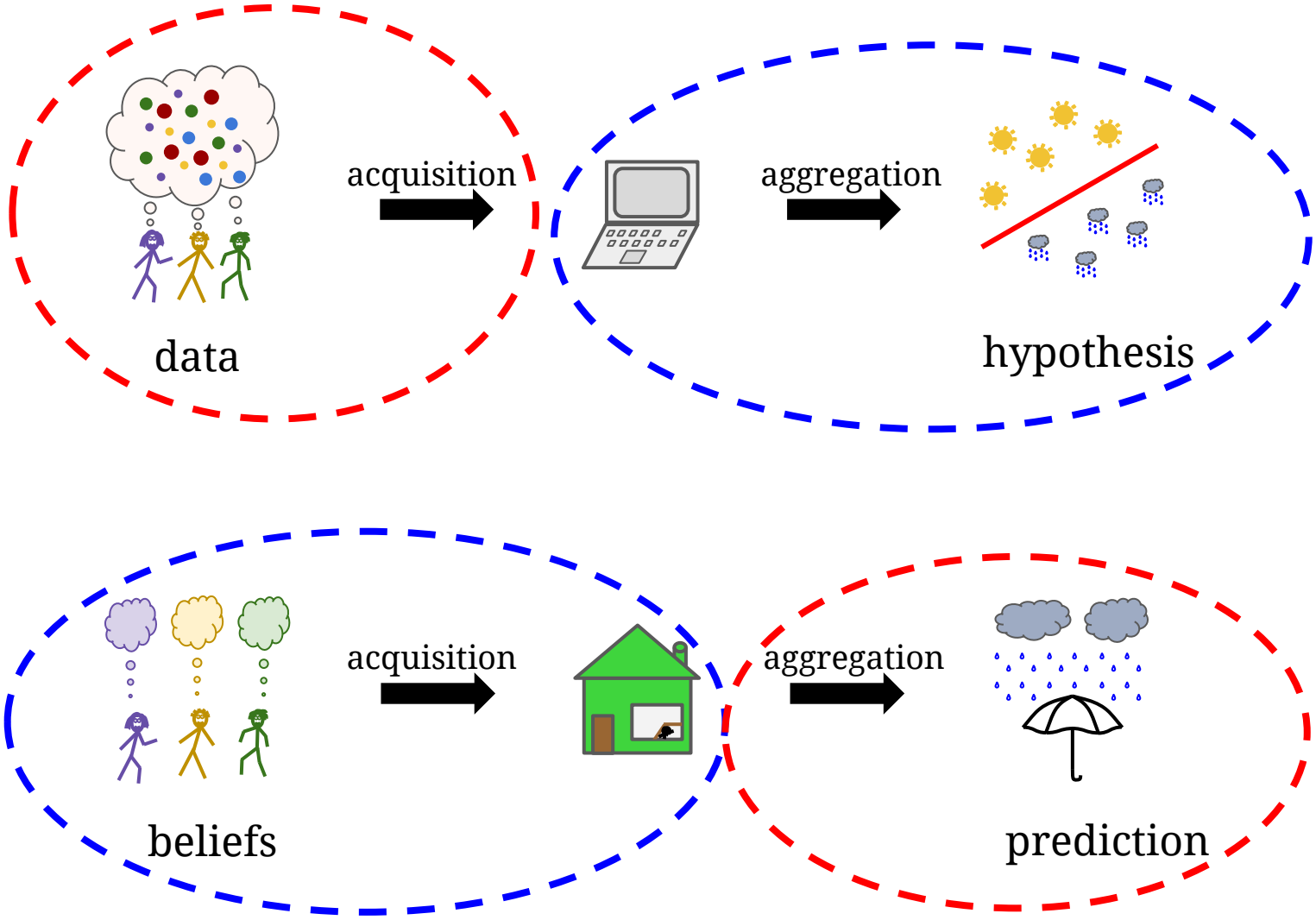
information
= expert beliefs

mechanism

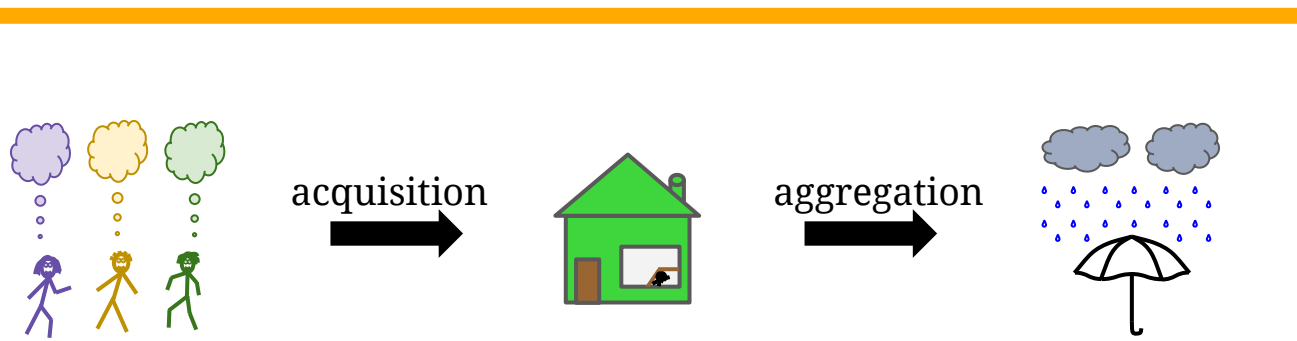
useful summary
= prediction

How to A&A beliefs controlled by strategic agents into a prediction?

The pieces are well-studied…



…but piece-wise approaches do not suffice!



Outline

Case #1: data and hypotheses

- a model for A&A of data
- “actively procuring data”

Case #2: beliefs and predictions

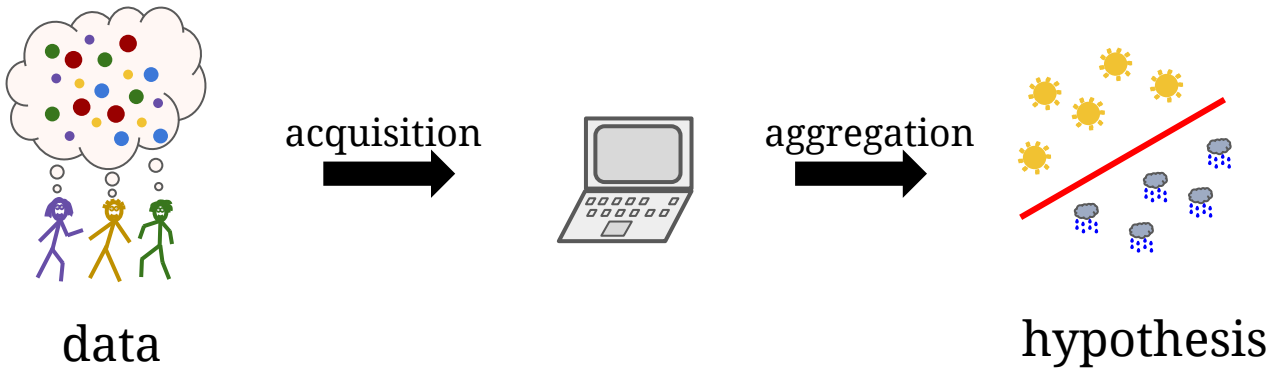
- “substitutes and complements” of information
- analyzing mechanisms for A&A of beliefs

Bringing the cases together

- mechanisms for both data and beliefs

Case #1: data and hypotheses

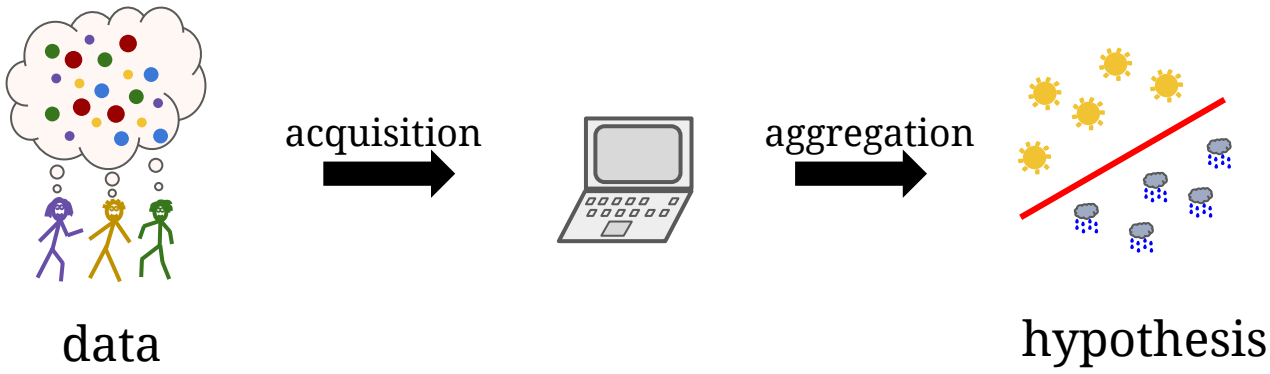
How to A&A data controlled by strategic agents into a machine-learning hypothesis?



Case #1: data and hypotheses

How to A&A data controlled by strategic agents into a machine-learning hypothesis?

Challenge: the acquisition process can **bias** the data.

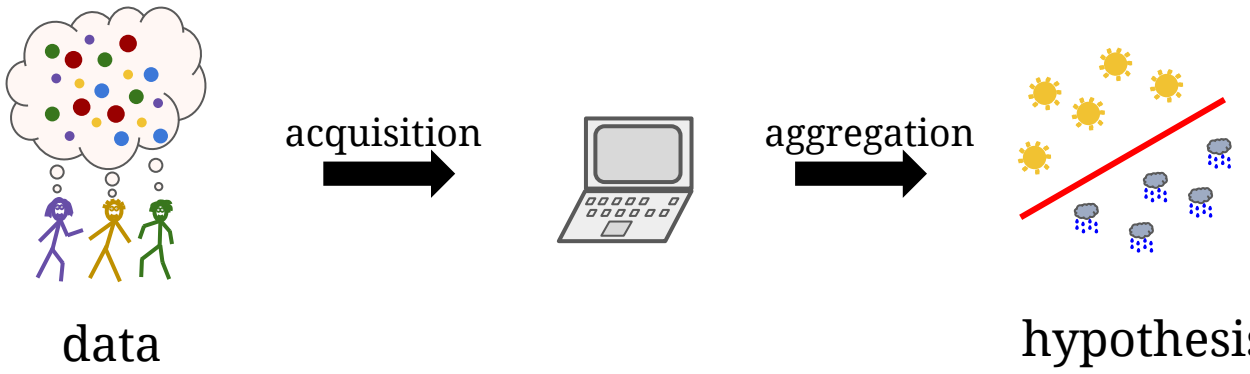


Case #1: data and hypotheses

How to A&A data controlled by strategic agents into a machine-learning hypothesis?

Challenge: the acquisition process can **bias** the data.

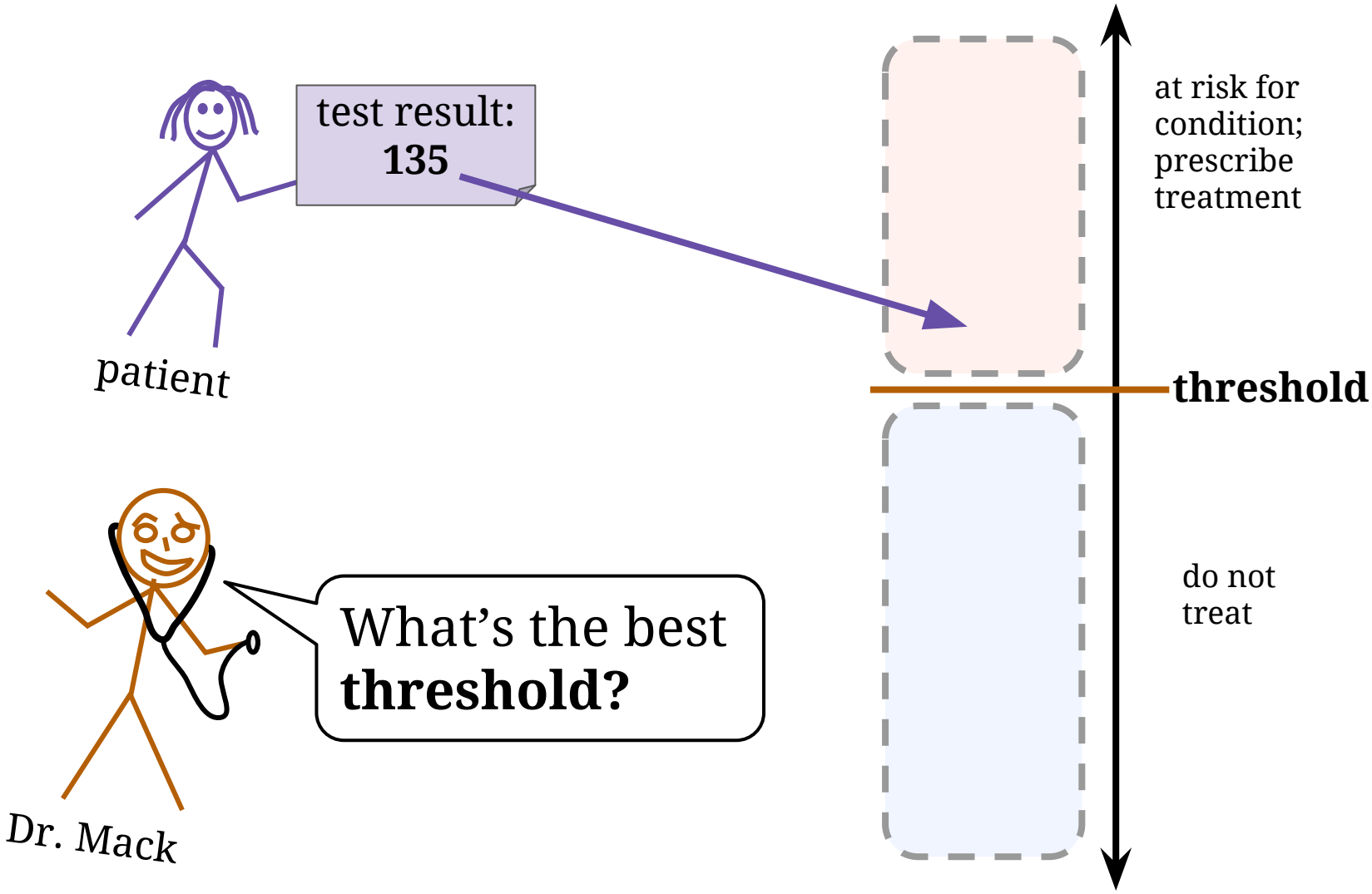
Challenge: we want to focus on acquiring **useful** data.



Outline for case #1

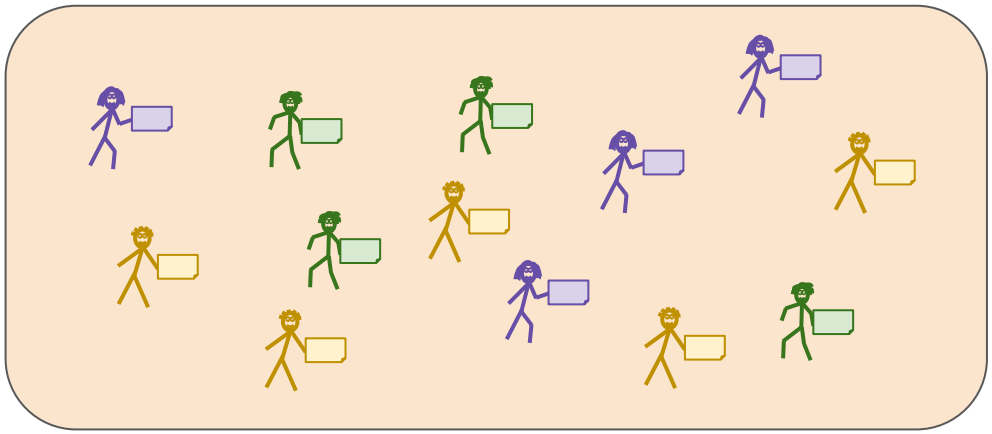
- **Introducing Dr. Mack**
- **A simple model and solution for Dr. Mack**
- **More complex problems**

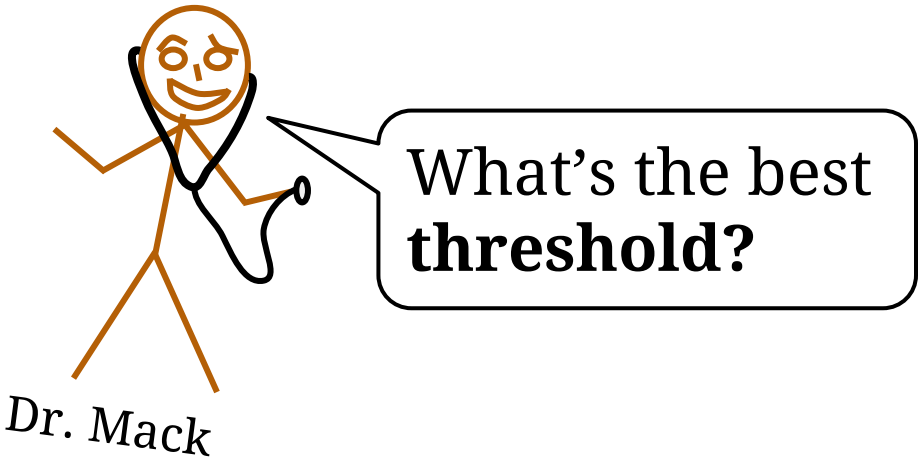
An example from Dr. Mack

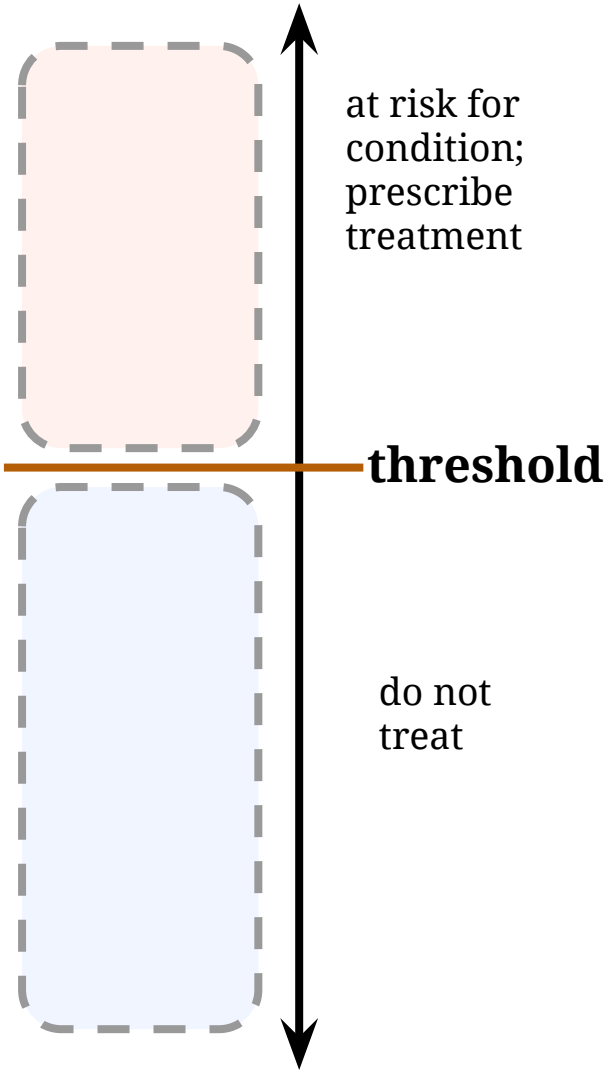


Goal: acquire and aggregate past data

Set of past patients with:
(test result, eventually-had-condition?)





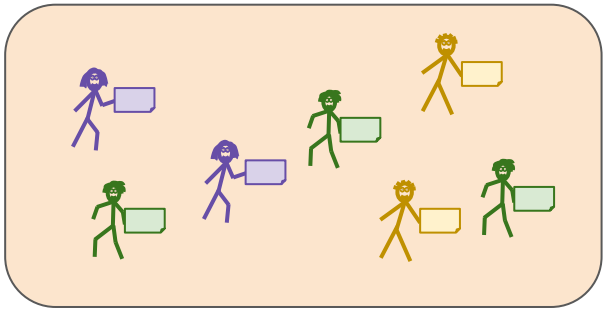


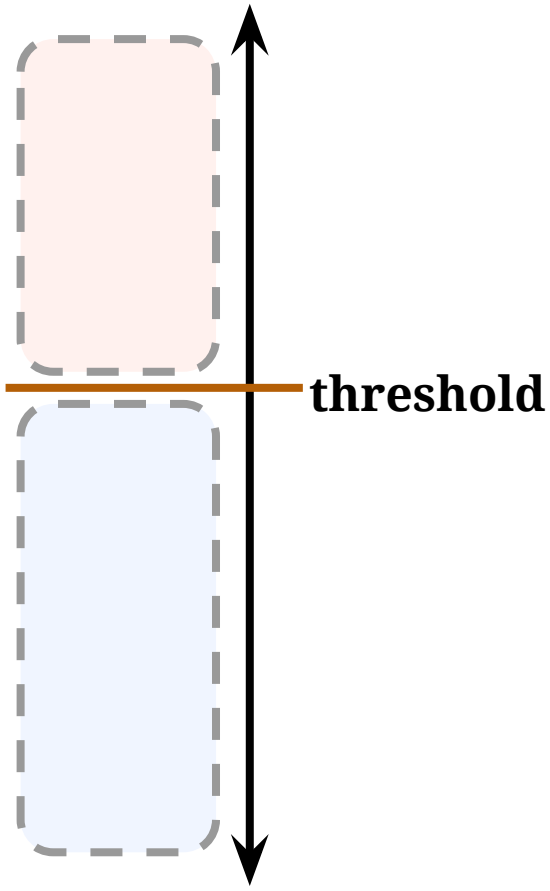
If Dr. Mack already had the data…

… he could use *e.g.* Rosenblatt's "perceptron" (1958):

1. Start with some threshold **h**

2. If **h** is wrong on data point, move toward it



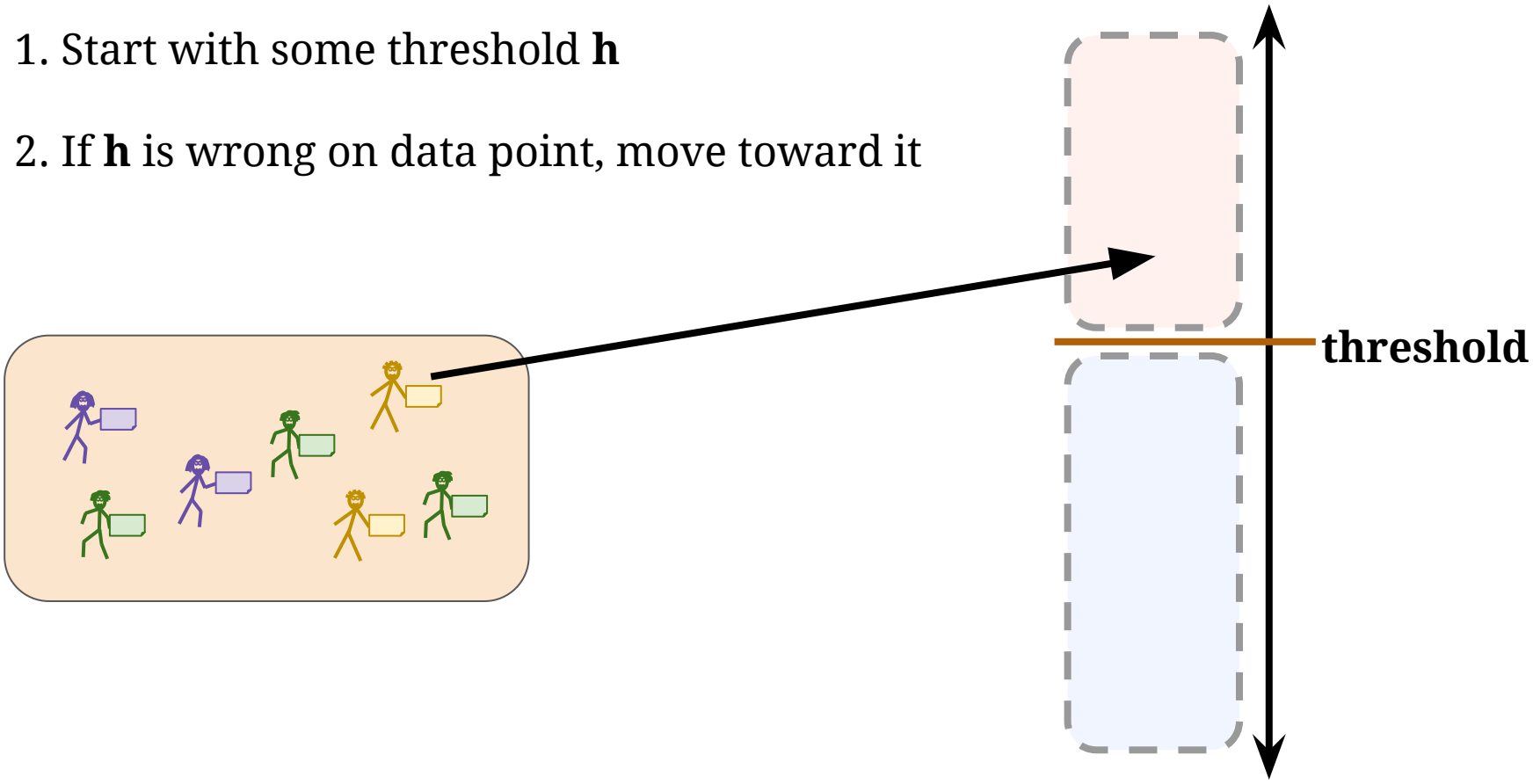


If Dr. Mack already had the data…

… he could use *e.g.* Rosenblatt's "perceptron" (1958):

1. Start with some threshold **h**

2. If **h** is wrong on data point, move toward it

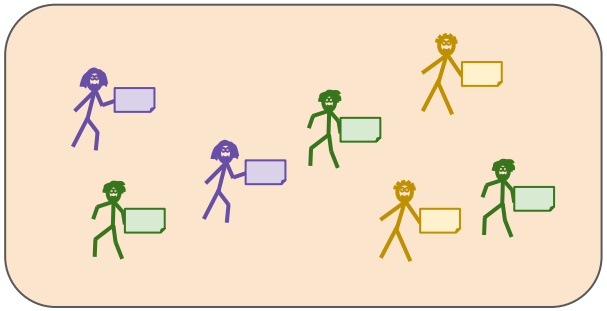


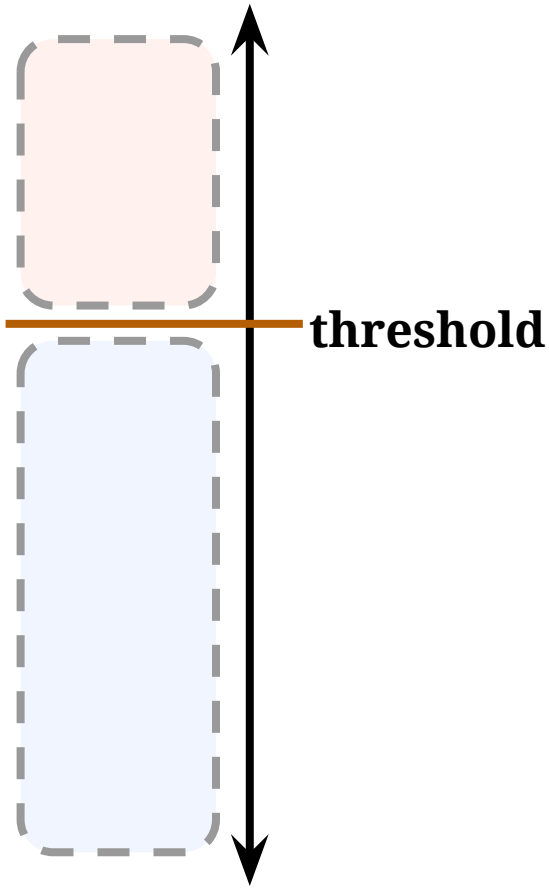
If Dr. Mack already had the data…

… he could use *e.g.* Rosenblatt's "perceptron" (1958):

1. Start with some threshold **h**

2. If **h** is wrong on data point, move toward it





If Dr. Mack already had the data…

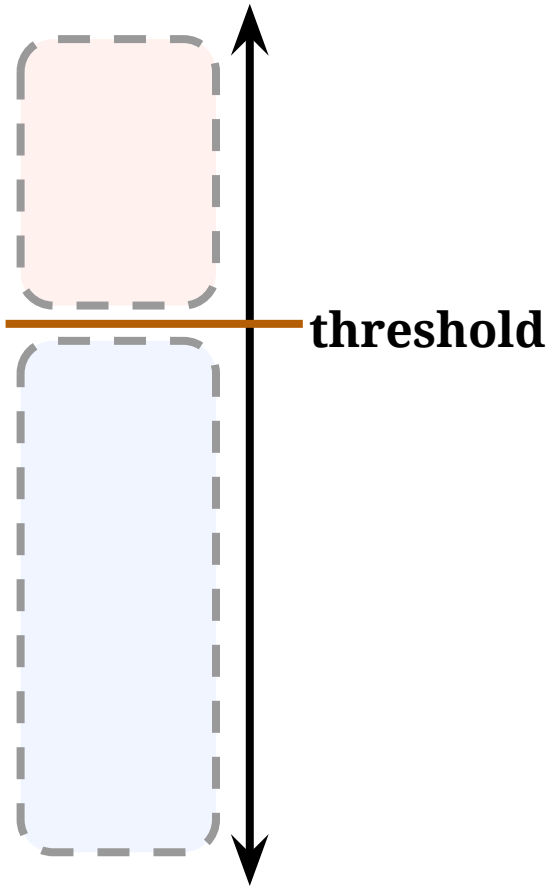
… he could use *e.g.* Rosenblatt's "perceptron" (1958):

1. Start with some threshold **h**

2. If **h** is wrong on data point, move toward it:

   $\mathbf{h} \leftarrow \mathbf{h} + \eta\,(x - \mathbf{h})$
   where x = patient's test result

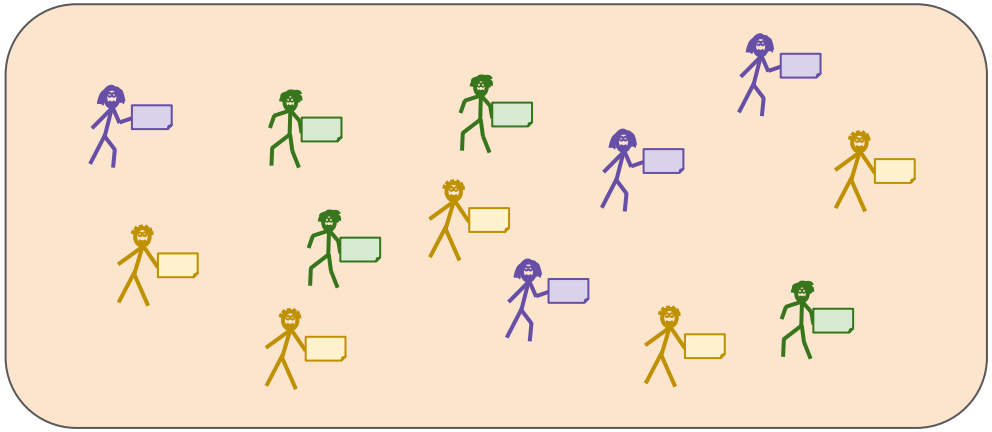
3. Repeat for all data points



But: data is controlled by the agents

Proposed model:

- Each agent holds a data point…
- … and agrees to disclose only if offered $100



Strategies for Dr. Mack

Keep buying data at $100 per until budget is exhausted.

Pro: Works seamlessly with previous algorithm.

Con: not a good strategy.

Strategies for Dr. Mack

Keep buying data at $100 per until budget is exhausted.

Pro: Works seamlessly with previous algorithm.

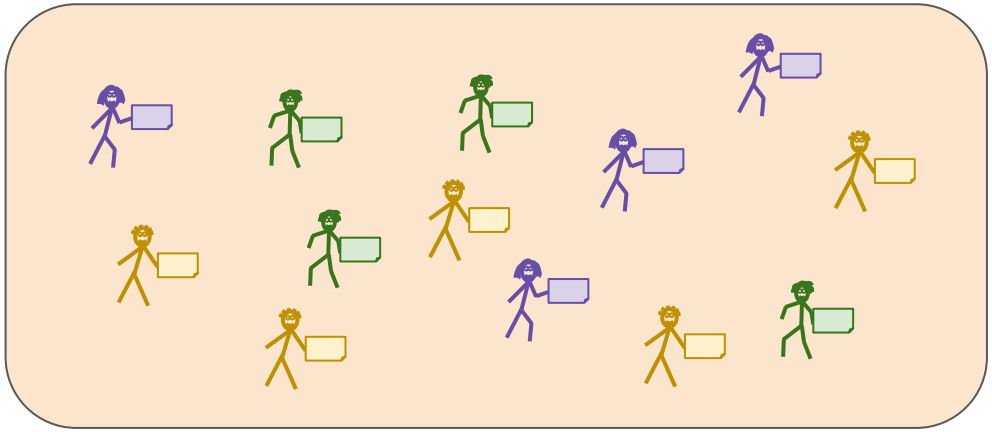
Pro: is a good strategy.

Only offer to buy data on which current algorithm is **wrong.**

A more sophisticated model

Updated model:

- Each agent holds a data point and **cost** $\leq$ \$100…
- … and agrees to disclose only if offered a higher price for the data point.



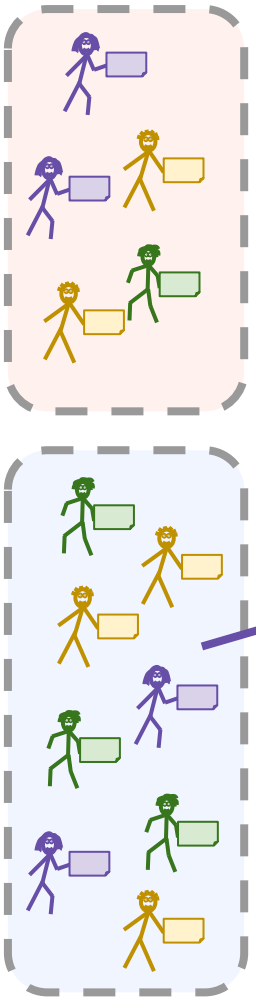
A more sophisticated strategy

Offer **randomly chosen prices** (only for data on which the current algorithm is wrong).

Pro: spends less budget.

Con: obtains biased data.

De-biasing the data from random prices



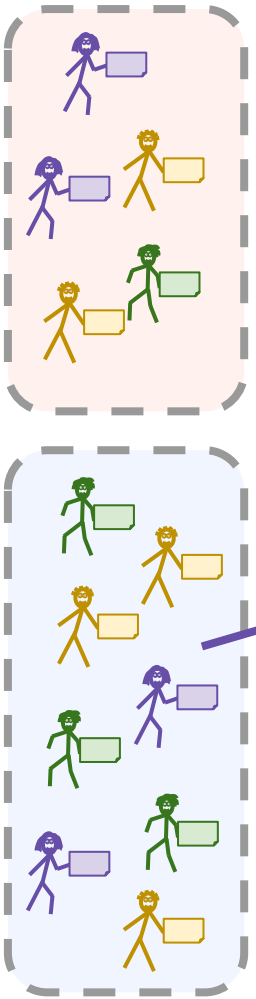
Example: **h** = 130.

test result:
135

Baseline algorithm (η=0.1):

update $h \leftarrow h + \eta(x - h)$
$= h + 0.5$

De-biasing the data from random prices



Example: **h** = 130.

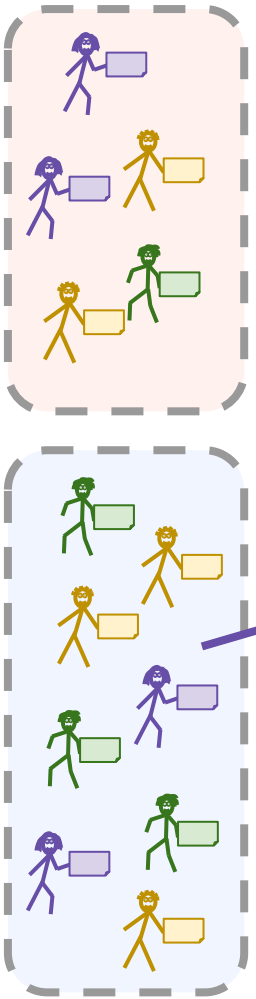
test result:
135

Baseline algorithm (η=0.1):

update $h \leftarrow h + \eta(x - h)$
$= h + 0.5$

Idea: draw random price,
make update of size 0.5 “on average”

De-biasing the data from random prices



Example: **h** = 130.
Example: price drawn uniform [$0, $100].

test result:
135

cost: $50

Dr. Mack's algorithm:

if agent agrees to price,
with cost = $50:
update h ← h + 1.0

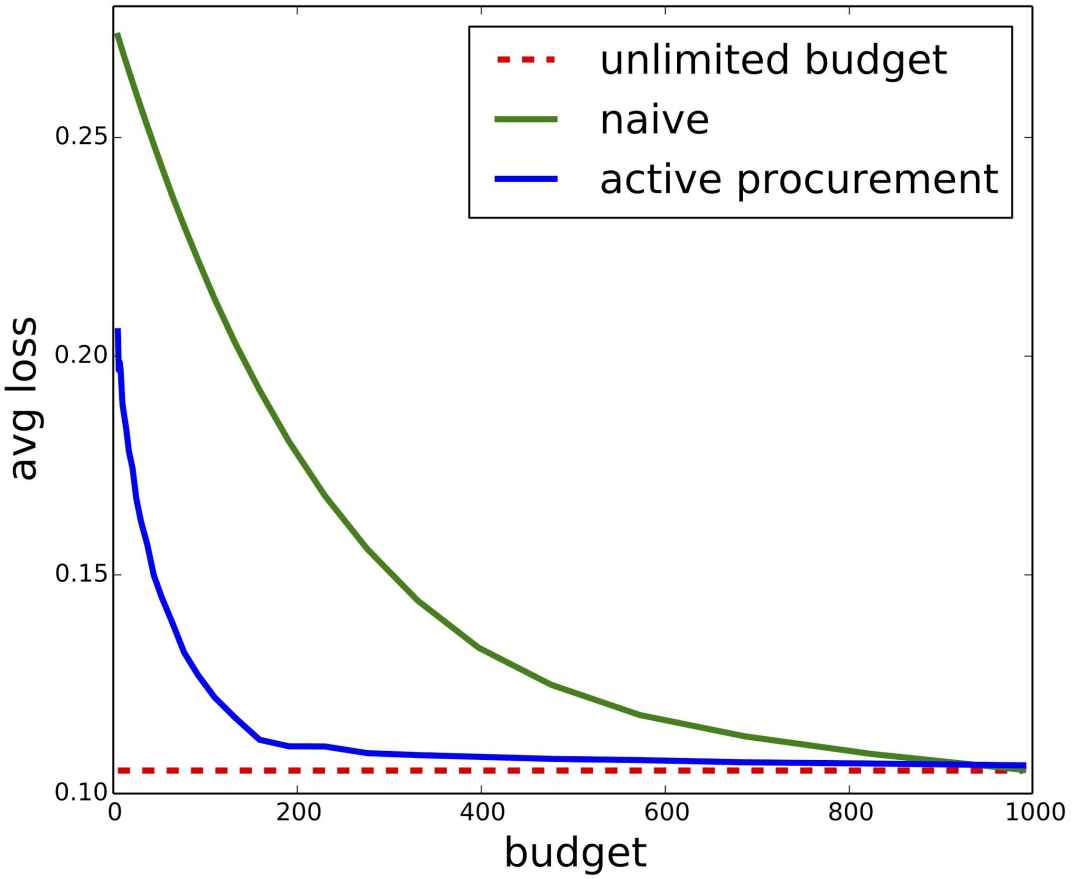
because we only get their data "half the time"

and so on.

Example plot

1000 patients, costs in [0,1].
(note: main results are theoretical…this is just for illustration!)

loss measures the performance of the final hypothesis



More generality in the thesis

- hypothesis is a vector in R^d; some convex loss function.
- proves bounds on “regret” and “risk”.
- more sophisticated measure for “value” of data.

Takeaways

The main ideas:

- actively procuring the most useful data
- can prove learning bounds with monetary resources
- algorithms → mechanisms

Takeaways

The main ideas:

- actively procuring the most useful data
- can prove learning bounds with monetary resources
- algorithms → mechanisms

Outline

Case #1: data and hypotheses

- a model for A&A of data
- “actively procuring data”

Case #2: beliefs and predictions

- “substitutes and complements” of information
- analyzing mechanisms for A&A of beliefs

Bringing the cases together

- mechanisms for both data and beliefs

Case #2: beliefs and predictions

How to A&A beliefs controlled by strategic agents into a prediction?

Challenge: Agents can lie, bluff, etc.

Challenge: how do different agents' beliefs **interact**?

Outline for case #2

- **Introducing Dr. Martha**
- **Prediction markets as a model for A&A**
- **Substitutes and complements**

Helping out Dr. Martha

Dr. Martha needs to predict the chance of rain tomorrow.
Alice and Bob have beliefs based on private information.

Helping out Dr. Martha

Dr. Martha needs to predict the chance of rain tomorrow.
Alice and Bob have beliefs based on private information.

Tool for acquisition: proper scoring rules.
1. Alice reports probability p of rain.
2. Martha pays S(p, 1) if it rains and S(p, 0) otherwise.

Helping out Dr. Martha

Dr. Martha needs to predict the chance of rain tomorrow.
Alice and Bob have beliefs based on private information.

Tool for acquisition: proper scoring rules.
1. Alice reports probability p of rain.
2. Martha pays S(p, 1) if it rains and S(p, 0) otherwise.

Example 1: $S(p,z) = -(p-z)^2$

Example 2: $S(p,1) = \log(p)$, $S(p,0) = \log(1-p)$.

Proper scoring rules are not enough

Problems:

- Dr. Martha may pay extra for redundant information
- How should Dr. Martha **aggregate** these reports?

Proper scoring rules are not enough

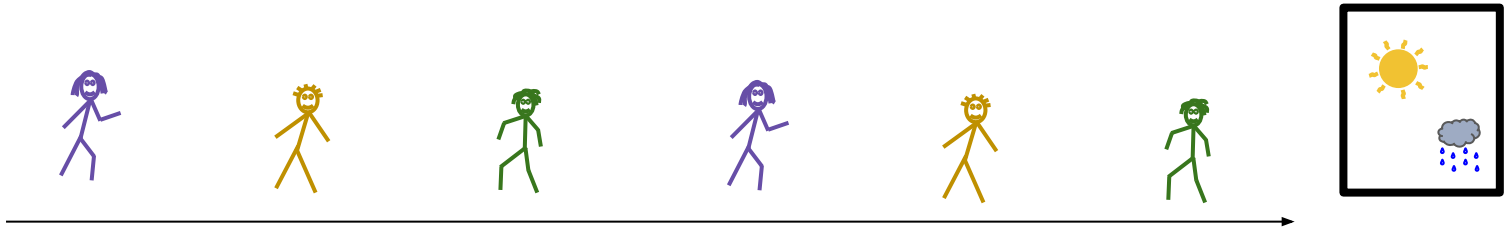
Problems:

- Dr. Martha may pay extra for redundant information
- How should Dr. Martha **aggregate** these reports?

A solution (Hanson 2003):

1. Alice sets initial prediction $p^{(1)}$
2. Bob updates prediction to $p^{(2)}$
3. Event is observed:
   Dr. Martha pays Alice $S(p^{(1)}, z)$
   Dr. Martha pays Bob $S(p^{(2)}, z) - S(p^{(1)}, z)$

Prediction market model



traders iteratively update market prediction

event is revealed,
payments assigned

Payment for changing prediction from p to p' is
S(p', z) - S(p, z).

An unsolved question!

Suppose Alice participates first. 

Then Bob. 

Then Alice again. 

In “equilibrium”, what do they do?

To see our solution, an analogy

Imagine Dr. Martha wants to buy **items** rather than **information**.

To see our solution, an analogy

Imagine Dr. Martha wants to buy **items** rather than **information**.

At each time, she will pay the **her marginal value** for a set of items:

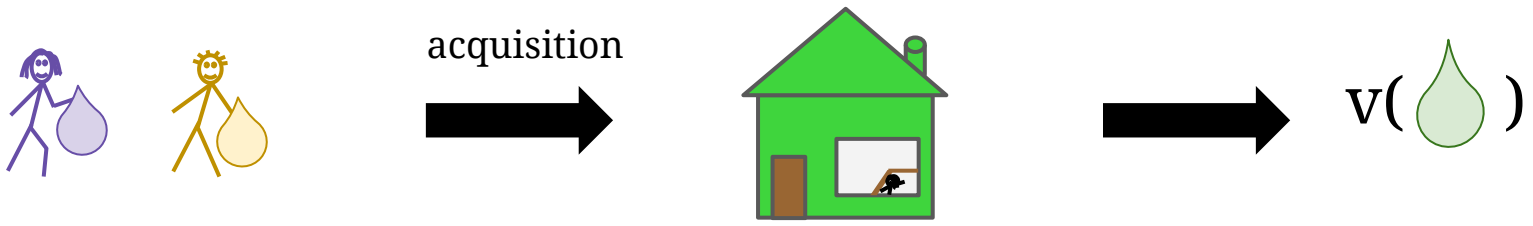
v(old items & new items) - v(old items).

Continuing the analogy

Consider the **Alice - Bob - Alice** market.

What if Alice has a left shoe and Bob has a right shoe?

What if Alice has chocolate ice cream and Bob has vanilla?



Stretching the analogy…

If Alice and Bob each have a **set** of items,
does Alice sells all items in the beginning?

Does she sell them all at the end?

Stretching the analogy…

If Alice and Bob each have a **set** of items,
does Alice sells all items in the beginning?

Does she sell them all at the end?

A: Yes if items are **substitutes** (resp., **complements**).

(Formally, corresponds to sub- and super-modular v.)

S&C for information

Our idea: make a general definition of **substitutes** and **complements** for pieces of information.

S&C for information

Our idea: make a general definition of **substitutes** and **complements** for pieces of information.

1. Martha has some **utility function**.
   u(d, z) = utility for taking decision d when event is z
   e.g. u(☂, ☼).

2. This leads to a **value for information**.

3. Now S&C can be defined analogously to items.
   diminishing marginal value = substitutes
   increasing marginal value = complements

Back to the unsolved question!

Suppose Alice participates first. 

Then Bob. 

Then Alice again. 

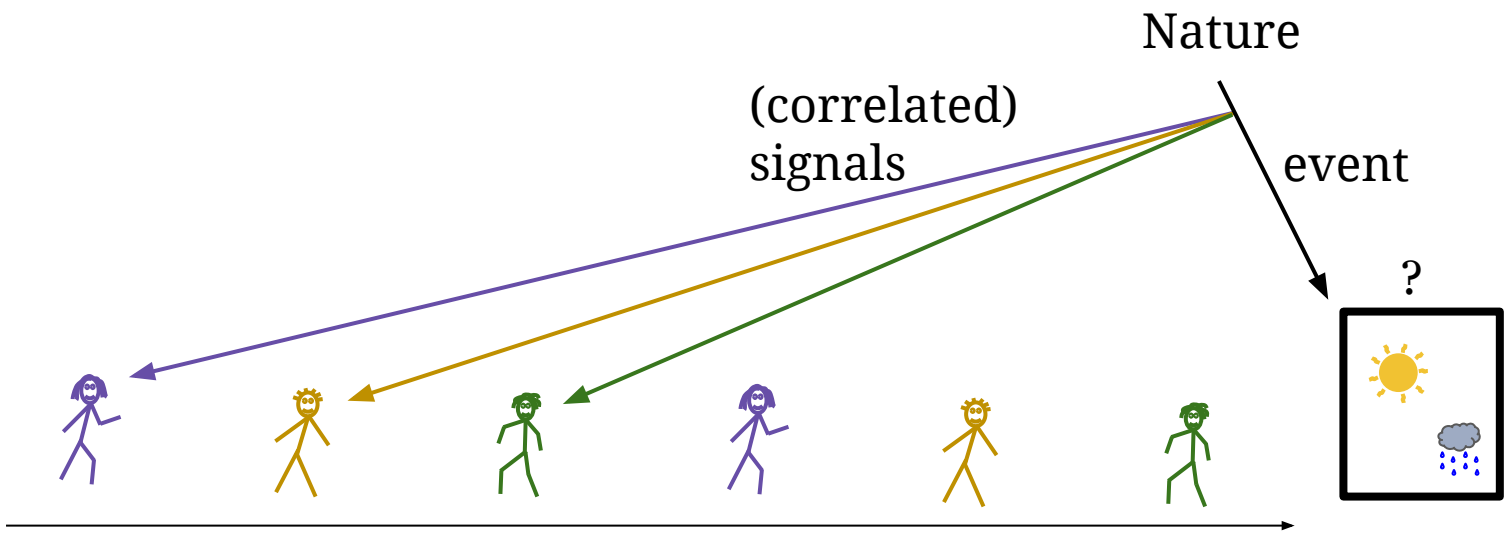
In “equilibrium”, what do they do?

Answer:

informational substitutes = rush to report
informational complements = delay

A bigger unsolved question

In general prediction markets, when do participants **rush to truthfully report and aggregate?**



A bigger unsolved question

In general prediction markets, when do participants **rush to truthfully report and aggregate?**

Answer: if and only if their signals are **substitutes.**

And: they fully delay if and only if **complements.**

A bigger unsolved question

In general prediction markets, when do participants **rush to truthfully report and aggregate?**

Answer: if and only if their signals are **substitutes.**

And: they fully delay if and only if **complements.**

Similar results apply for some crowdsourcing contests and question-and-answer forums.

Some big picture takeaways

- Information + incentives is hard!
- Analogies between items and information are useful…
  …up to a point.
- **Structure** and **context** both matter in determining value of information, S&C.

Outline

Case #1: data and hypotheses

- a model for A&A of data
- “actively procuring data”

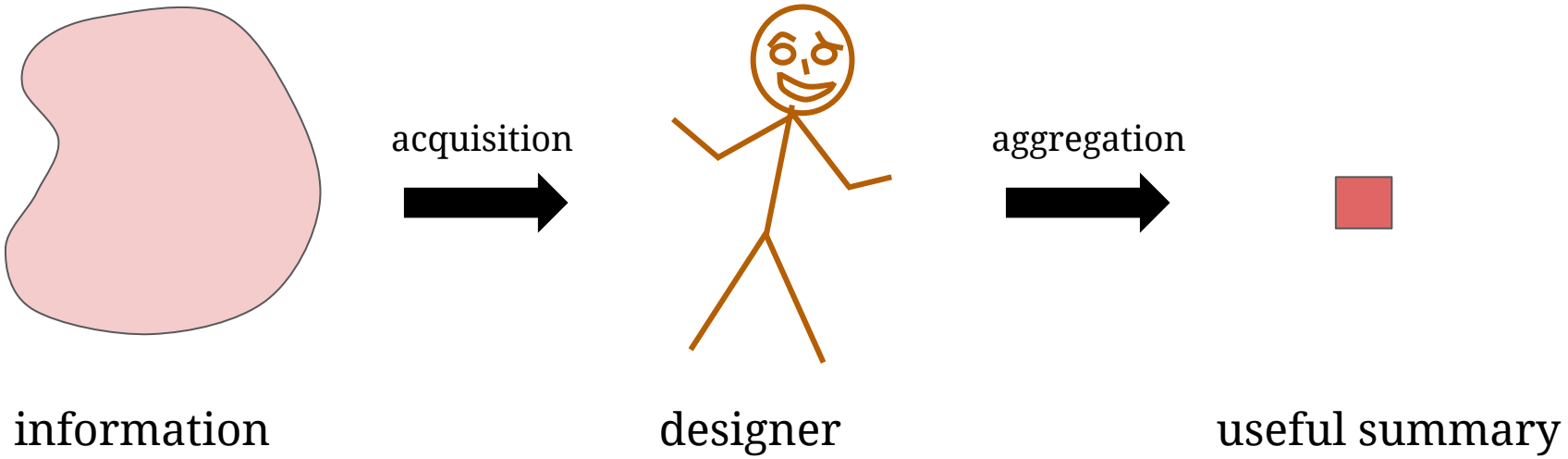
Case #2: beliefs and predictions

- “substitutes and complements” of information
- analyzing mechanisms for A&A of beliefs

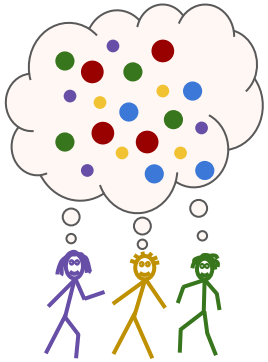
Bringing the cases together

- mechanisms for both data and beliefs

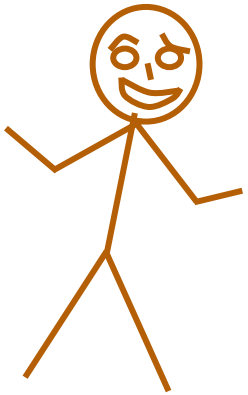
Recall the problem, and two approaches

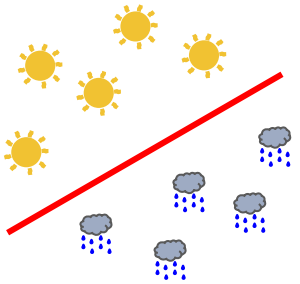


Recall the problem, and two approaches

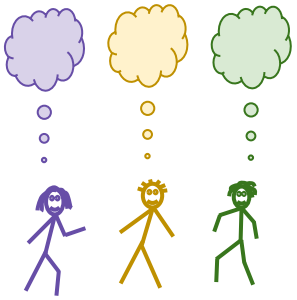


data





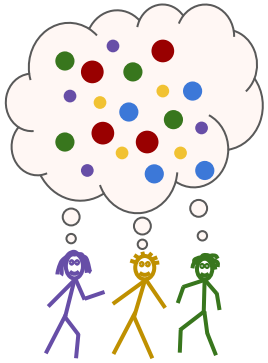
hypothesis

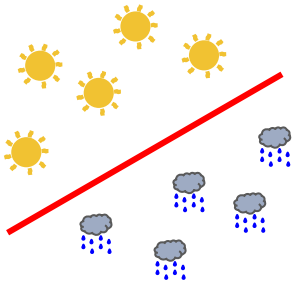


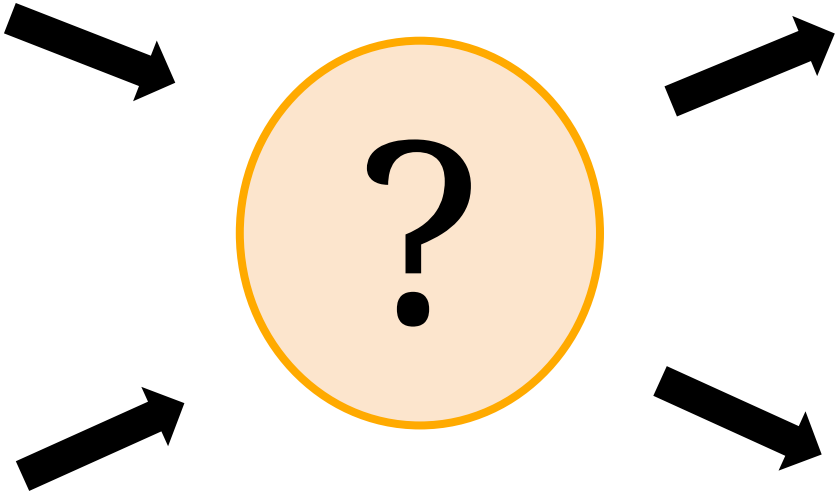
(expert) opinions
and beliefs

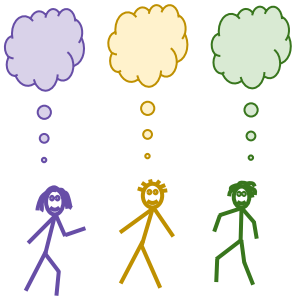
prediction or
decision

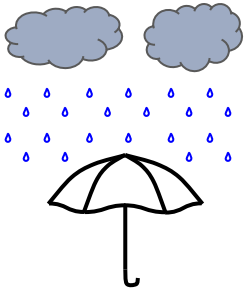
Challenge going forward











What can these approaches teach each other?

An illustrative mechanism

Goal: pick a good **threshold** for Dr. Mack.

Market Framework:

1. Designer chooses initial threshold **h**.

2. Traders arrive, iteratively update to new threshold.

3. Designer draws a test data point from the population.
   Each trader's update gets paid
   *loss(*new **h**, test data*)* - *loss(*old **h**, test data*)*.

An illustrative mechanism

Goal: pick a good **threshold** for Dr. Mack.

Market Framework:

1. Designer chooses initial threshold **h**.

2. Traders arrive, iteratively update to new threshold.

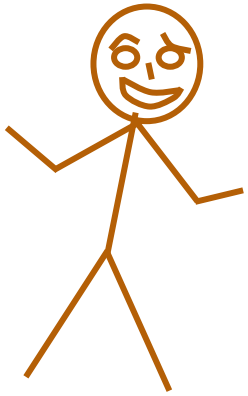
3. Designer draws a test data point from the population.
   Each trader's update gets paid
   *loss(*new **h**, test data*)* - *loss(*old **h**, test data*)*.

A trader with beliefs can update **h** to reflect those beliefs.
A trader with data can submit that data; a learning algorithm uses it to update the hypothesis.

Our results

Can use tools from both worlds for this model:

- solve machine-learning problems with data (achieve low “risk” or predictive error)
- good incentive properties: truthful reporting of beliefs, rushing if substitutes, ….



Some final thoughts

- Moving toward a world where **people are in control of their own data**

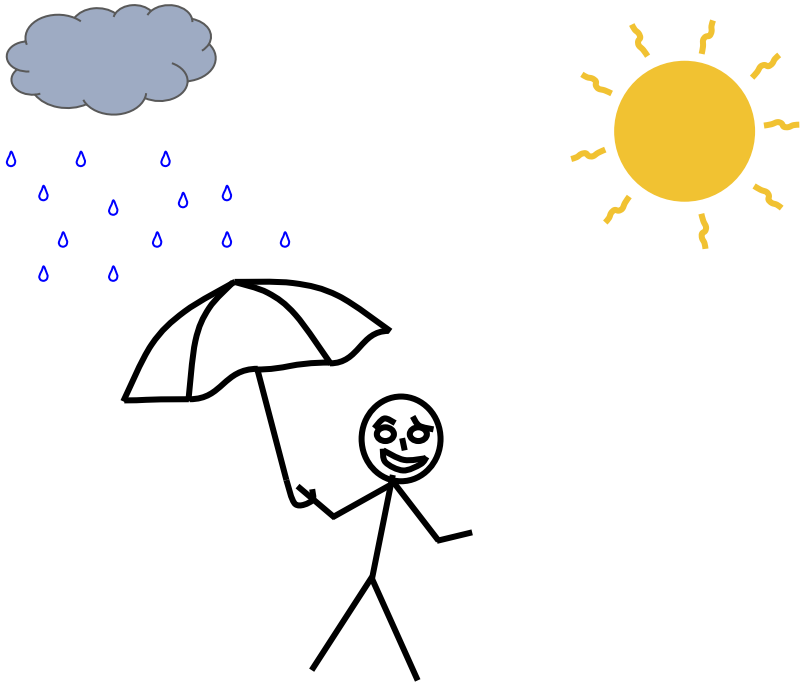
Some final thoughts

- Moving toward a world where **people are in control of their own data**
- The (relative) value of information derives from both **structure** and **context**

Some final thoughts

- Moving toward a world where **people are in control of their own data**
- The (relative) value of information derives from both **structure** and **context**
- We can do a lot of things with information, but there is a huge amount left to **understand…**

That’s it!



Thanks!

Tiger got to hunt,
Bird got to fly;
Man got to sit and wonder, “Why, why, why?”

Tiger got to sleep,
Bird got to land;
Man got to tell himself he understand.

The Books of Bokonon