DATA WITHOUT BORDERS

GAME-THEORETIC CHALLENGES IN DEMOCRATIZING DATA

MIDWEST MACHINE LEARNING SYMPOSIUM

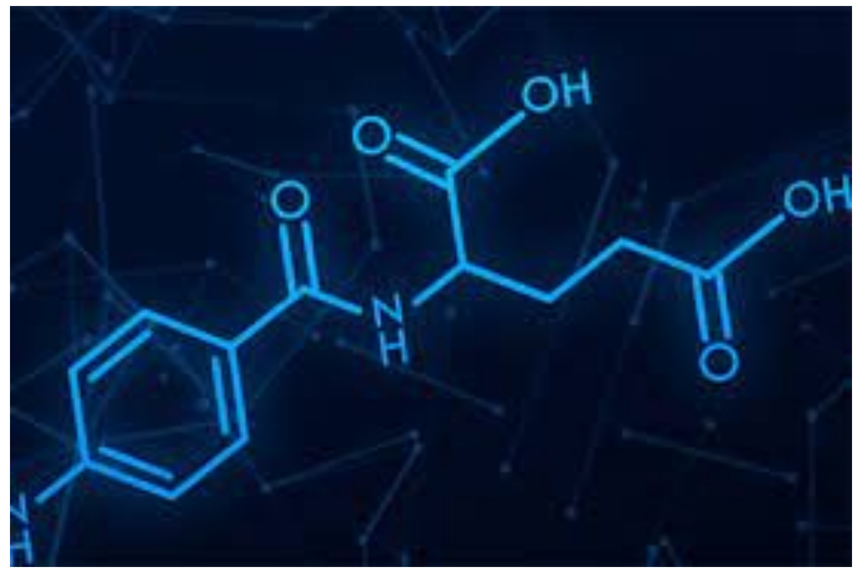
MAY 20, 2024

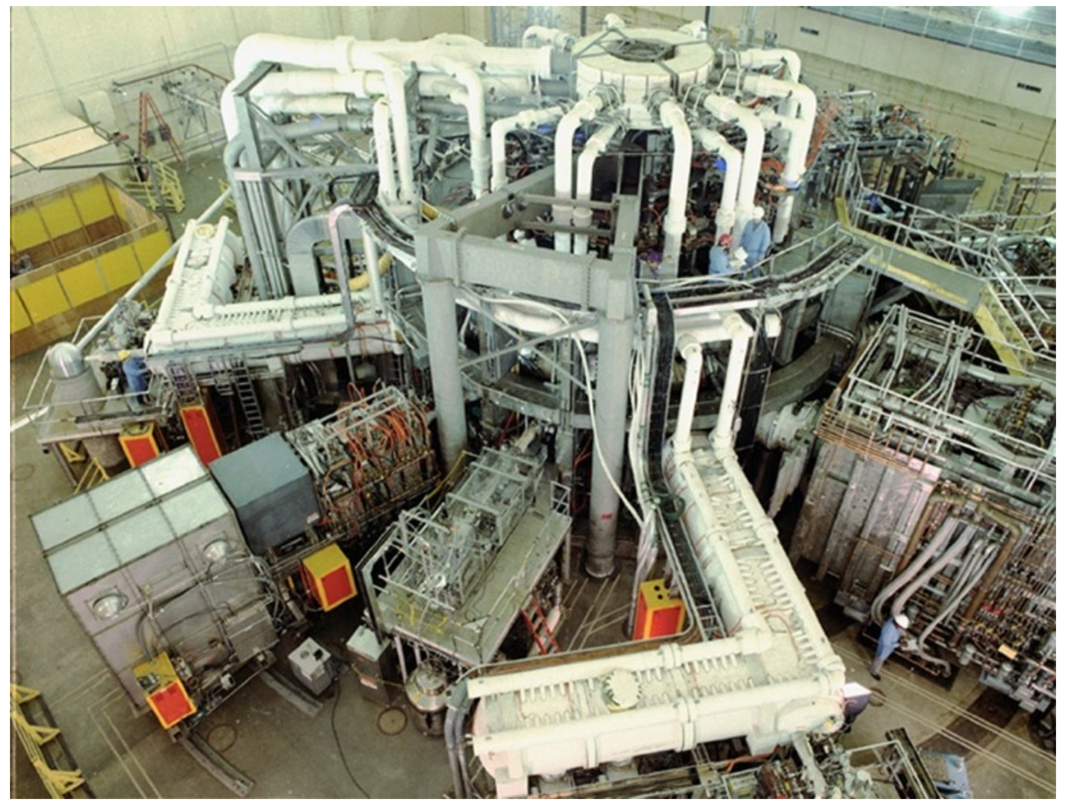
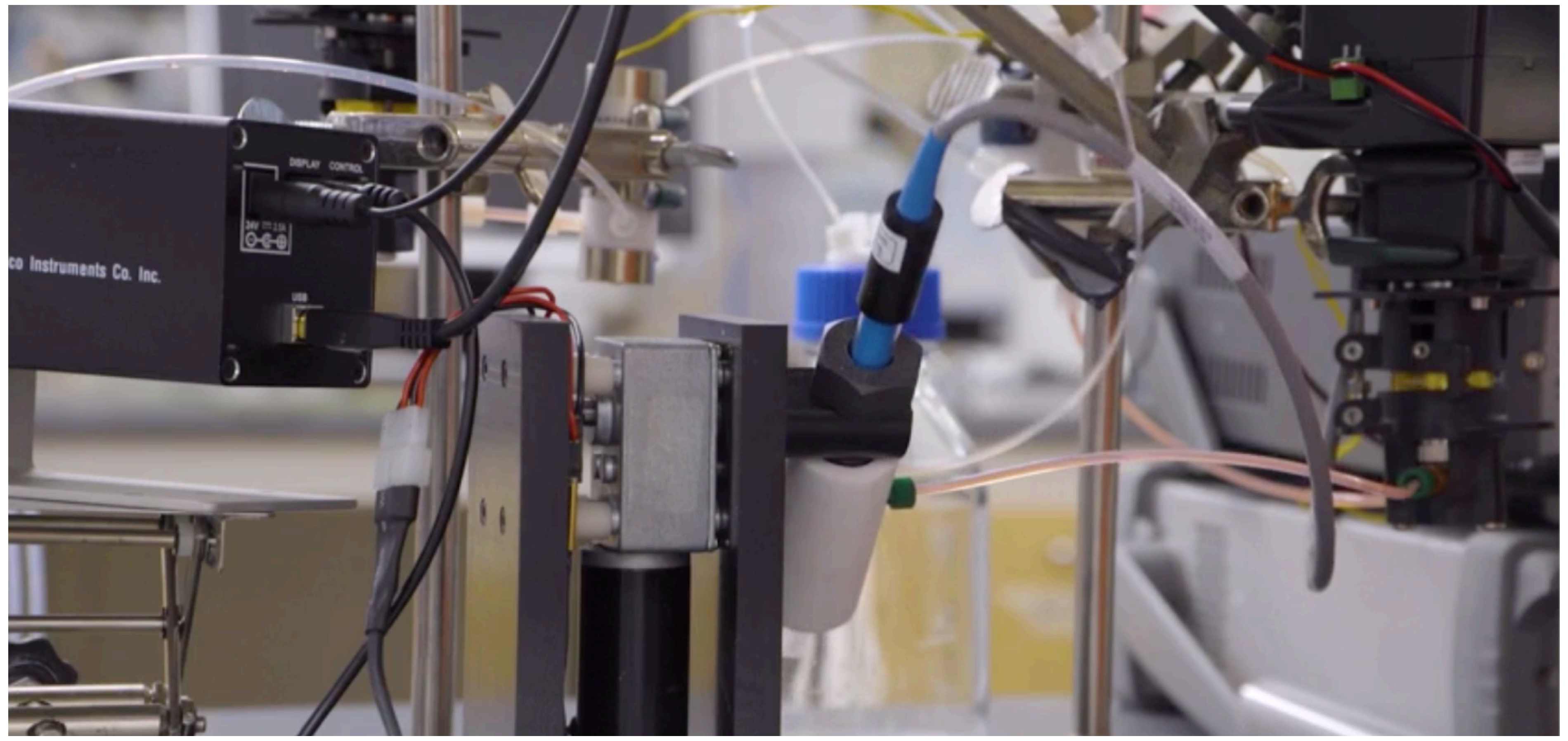
KIRTHEVASAN KANDASAMY

UNIVERSITY OF WISCONSIN-MADISON

MACHINE LEARNING IS UBIQUITOUS

- Consumer facing businesses
- Scientific research
- Industrial processes
- Transport/logistics

DATA IS AN INVALUABLE RESOURCE

DATA IS AN INVALUABLE RESOURCE

- Data is the *new oil*.
- Data is the *new gold*.

The Economist, NY Times, Forbes, Wired, Deloitte, EY, Boston Consulting Group, and several more …

DATA IS AN INVALUABLE RESOURCE

- Data is the *new oil*.
- Data is the *new gold*.

The Economist, NY Times, Forbes, Wired, Deloitte, EY, Boston Consulting Group, and several more …

- But data is different to other types of resources
  - Data is ***costly*** to produce, but ***free*** to replicate.

A UTOPIAN GOAL

Everyone collects data, everyone shares their data with others.

- Cost incurred by one organization to produce data can benefit others.
- Better for the organizations, better for society at large.

SHARING WILL DEMOCRATIZE DATA

SHARING WILL DEMOCRATIZE DATA

Small organizations with little data:

A B C D E F

SHARING WILL DEMOCRATIZE DATA

Small organizations with little data:

A B C D E F

Large organization with lots of data:

SHARING WILL DEMOCRATIZE DATA

Small organizations with little data:

A B C D E F

Large organization with lots of data:

SHARING WILL DEMOCRATIZE DATA

Small organizations with little data:

A B C D E F

Large organization with lots of data:

By sharing data with each other, small organizations can compete with larger organizations.

OBSTACLES TO DATA SHARING

OBSTACLES TO DATA SHARING

Ethical/Legal

Privacy

Ownership of data

OBSTACLES TO DATA SHARING

Ethical/Legal

Privacy

Ownership of data

Security

Data breaches

Adversarial attacks

OBSTACLES TO DATA SHARING

Ethical/Legal	Security	Logistical
Privacy	Data breaches	Inter-operability
Ownership of data	Adversarial attacks	Communication costs

OBSTACLES TO DATA SHARING

Ethical/Legal

Privacy

Ownership of data

Security

Data breaches

Adversarial attacks

Logistical

Inter-operability

Communication costs

Incentives

Free-riding

Competition

Data monetization

Data valuation

OBSTACLES TO DATA SHARING

Ethical/Legal

Privacy

Ownership of data

Security

Data breaches

Adversarial attacks

Logistical

Inter-operability

Communication costs

Incentives

Free-riding

Competition

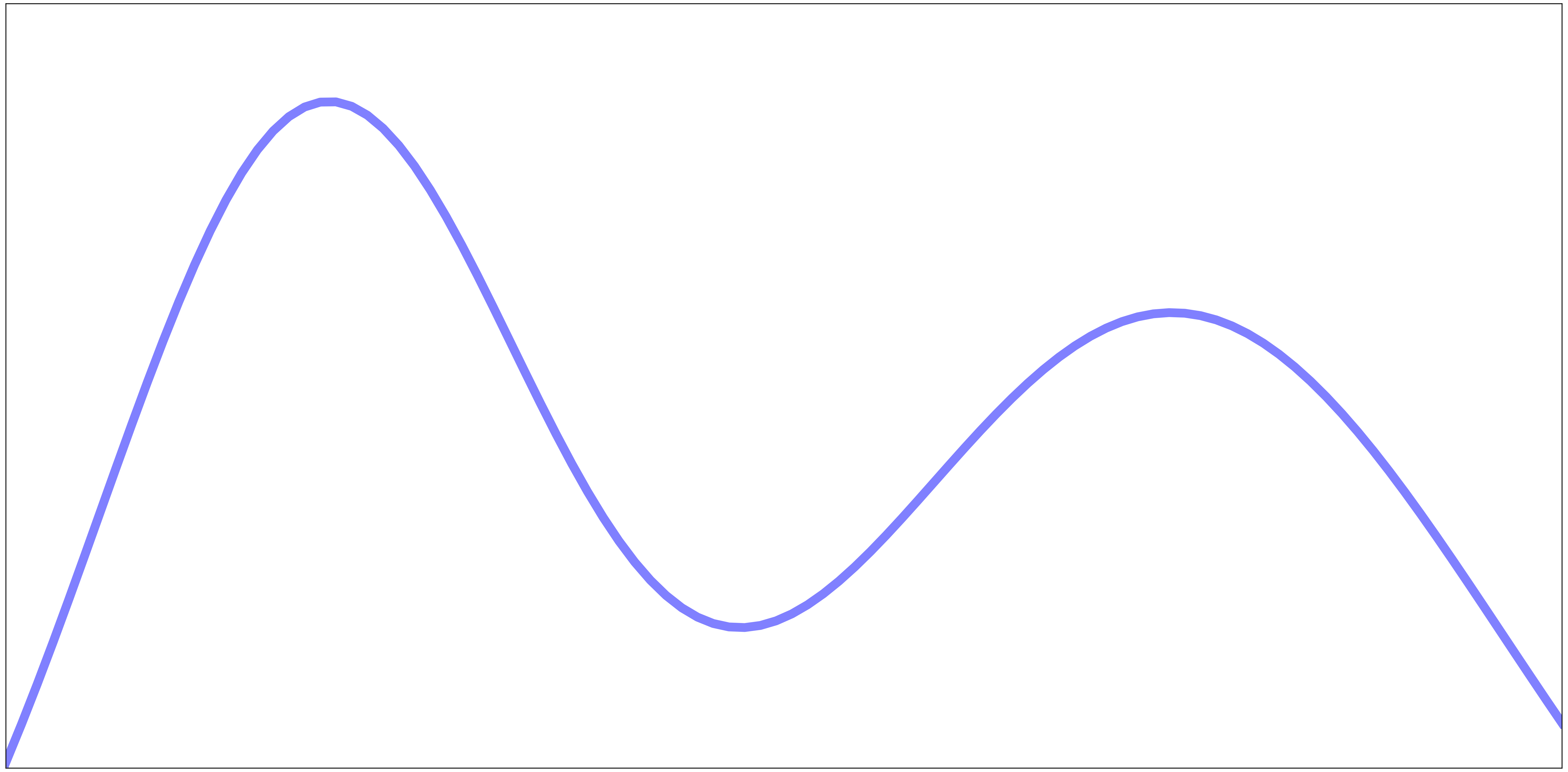
Data monetization

Data valuation

This talk

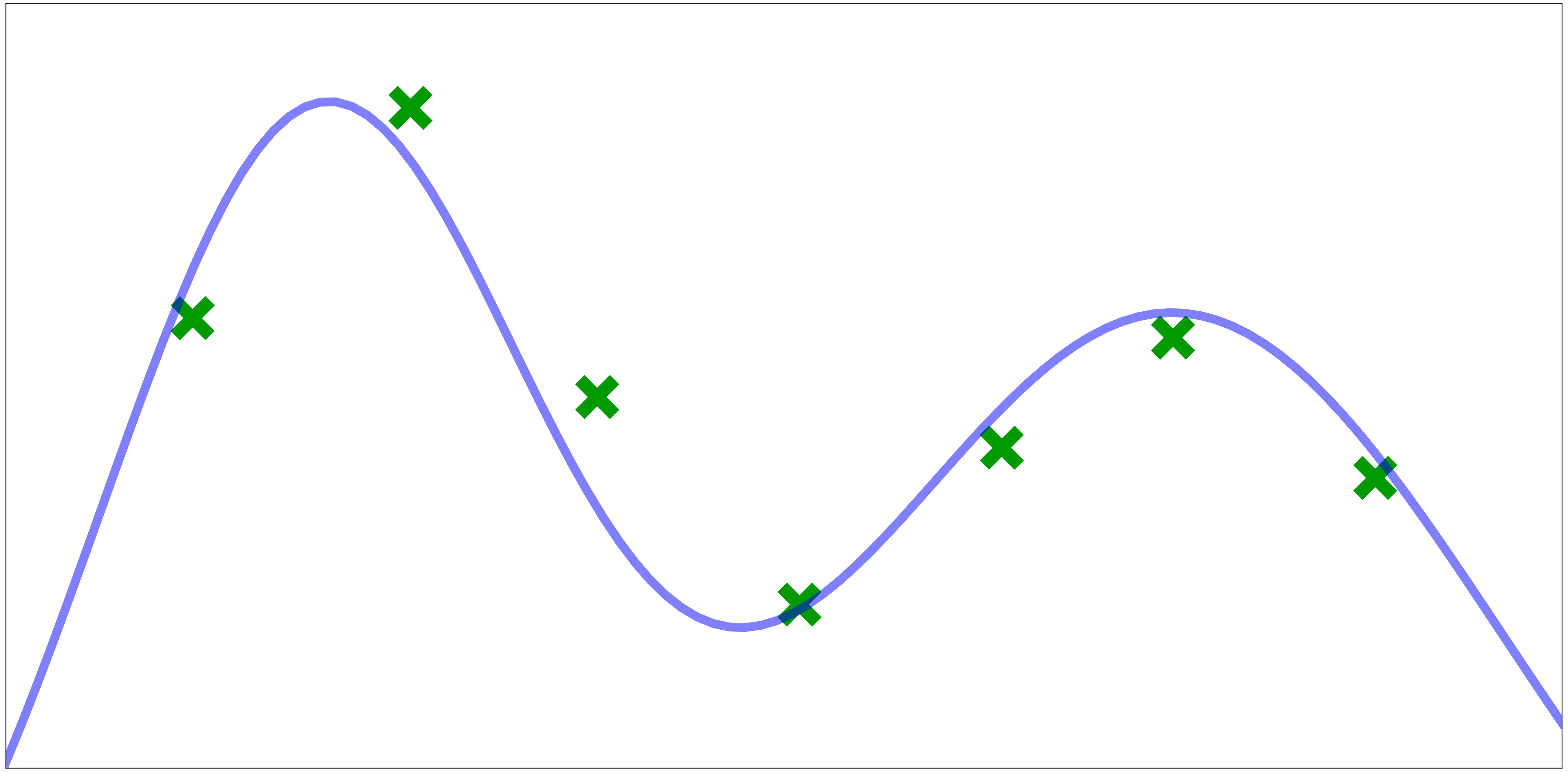
FREE-RIDING (#1/3)

agent's penalty = estimation error + cost of data collection



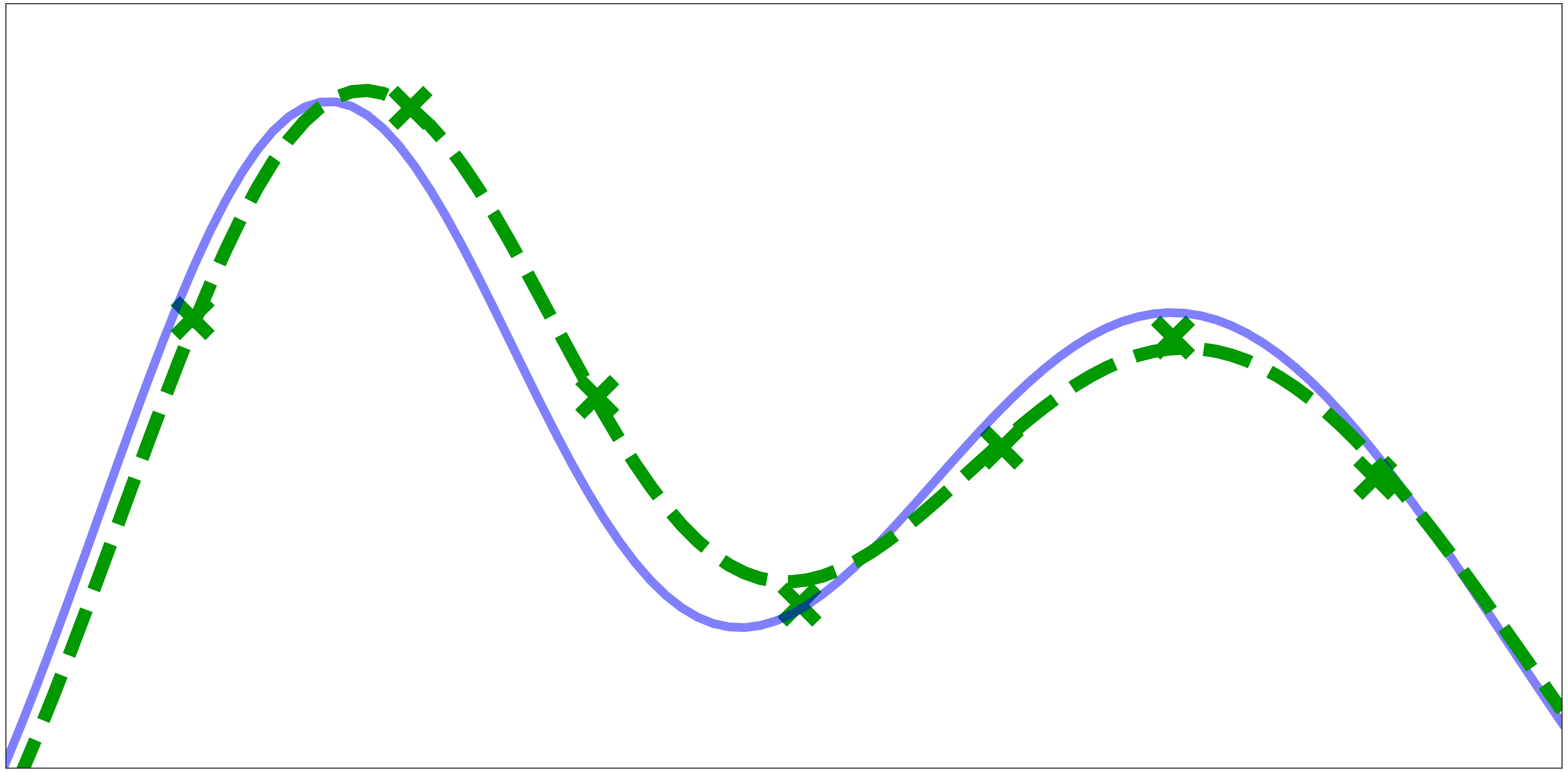
FREE-RIDING (#1/3)

agent's penalty = estimation error + cost of data collection



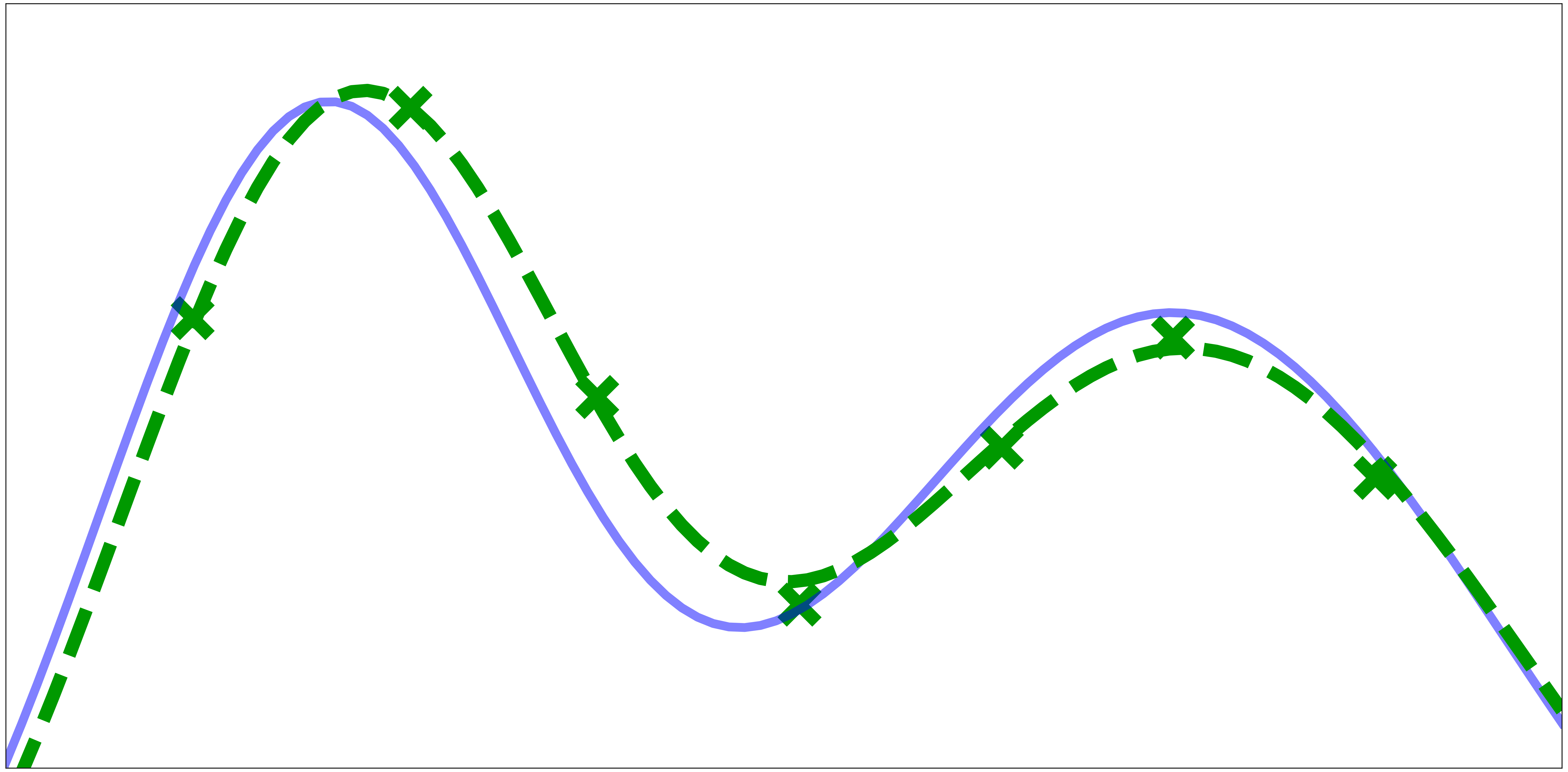
FREE-RIDING (#1/3)

agent's penalty = estimation error + cost of data collection



FREE-RIDING (#1/3)

agent's penalty = estimation error + cost of data collection

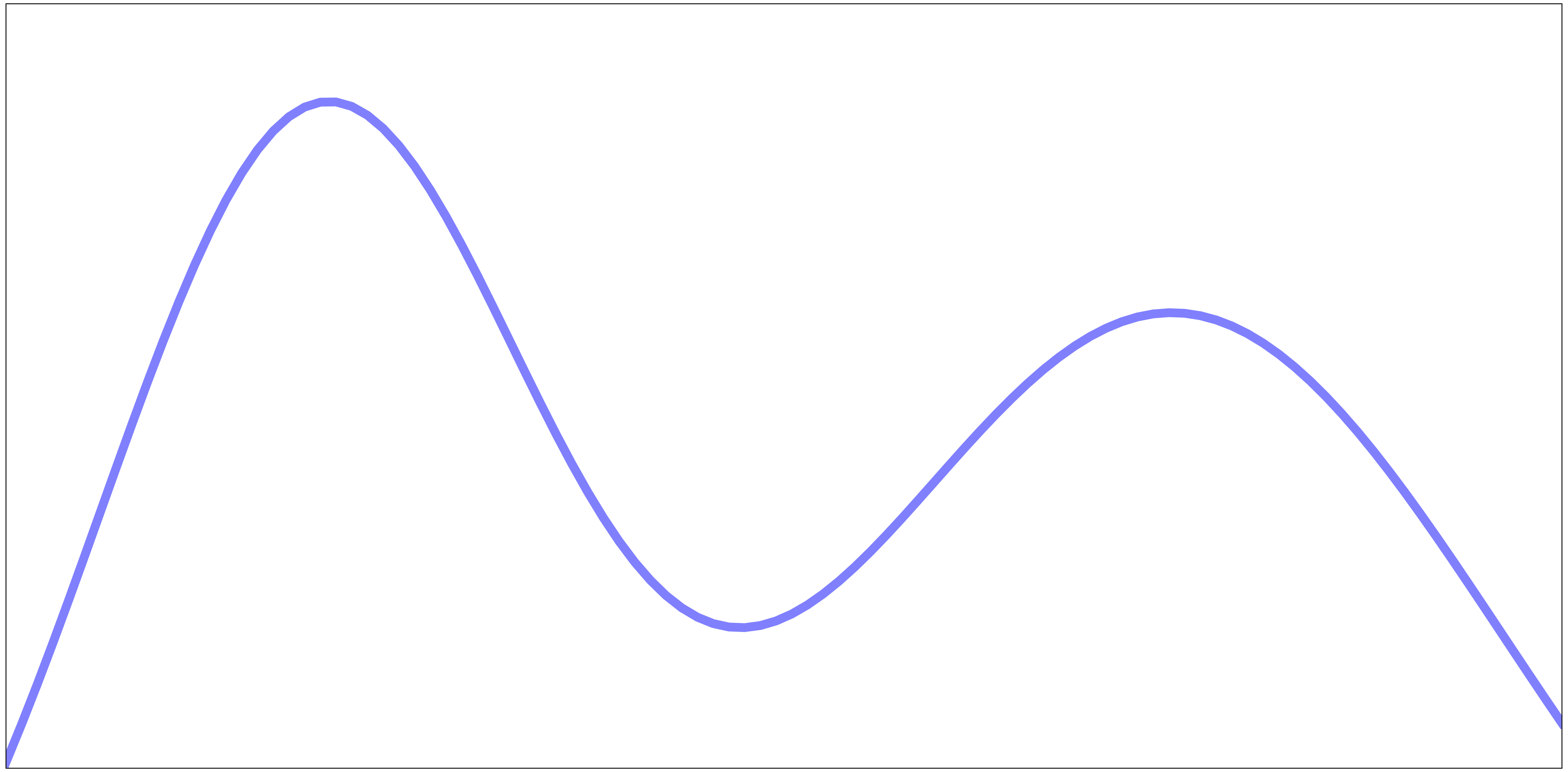


When **working on her own**, an agent will collect enough data until the cost offsets the (diminishing) increase in value from data.

FREE-RIDING (#2/3)

Multiple agents share data via a *naive* pool-and-share protocol:

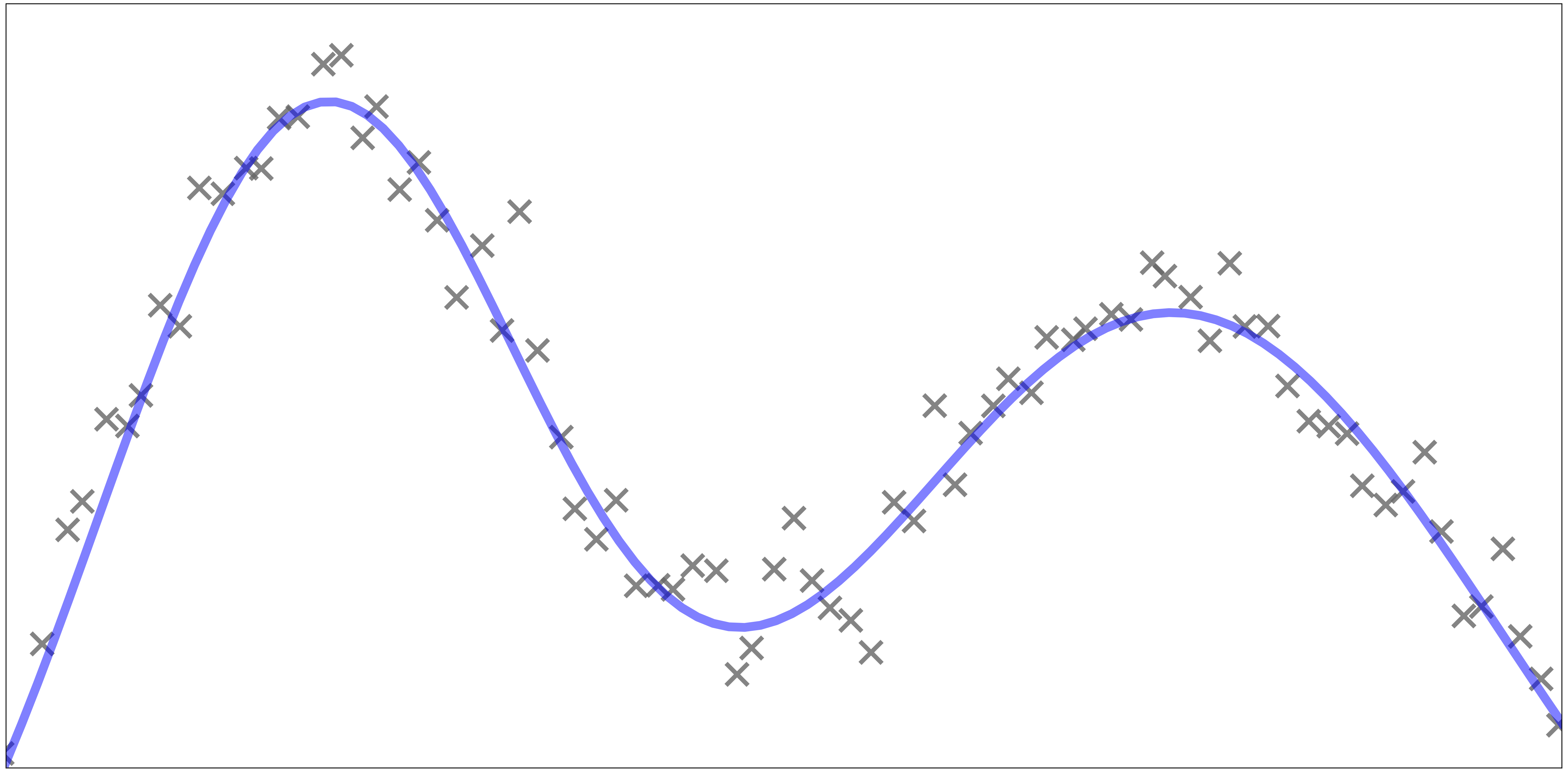
- Everyone collects data, everyone gets a copy of the others' data.



FREE-RIDING (#2/3)

Multiple agents share data via a *naive* pool-and-share protocol:

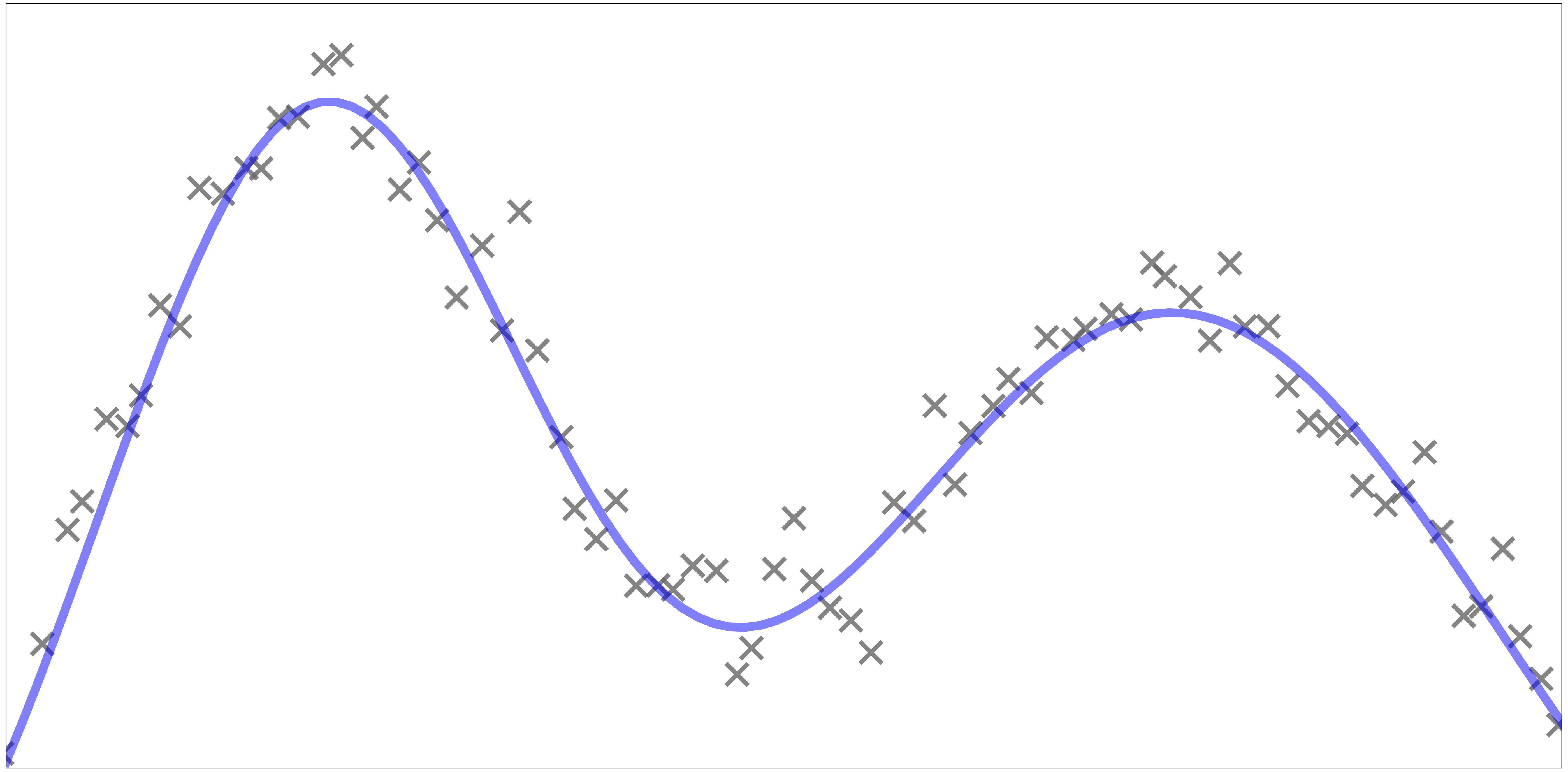
- Everyone collects data, everyone gets a copy of the others' data.



FREE-RIDING (#2/3)

Multiple agents share data via a *naive* pool-and-share protocol:

- Everyone collects data, everyone gets a copy of the others' data.

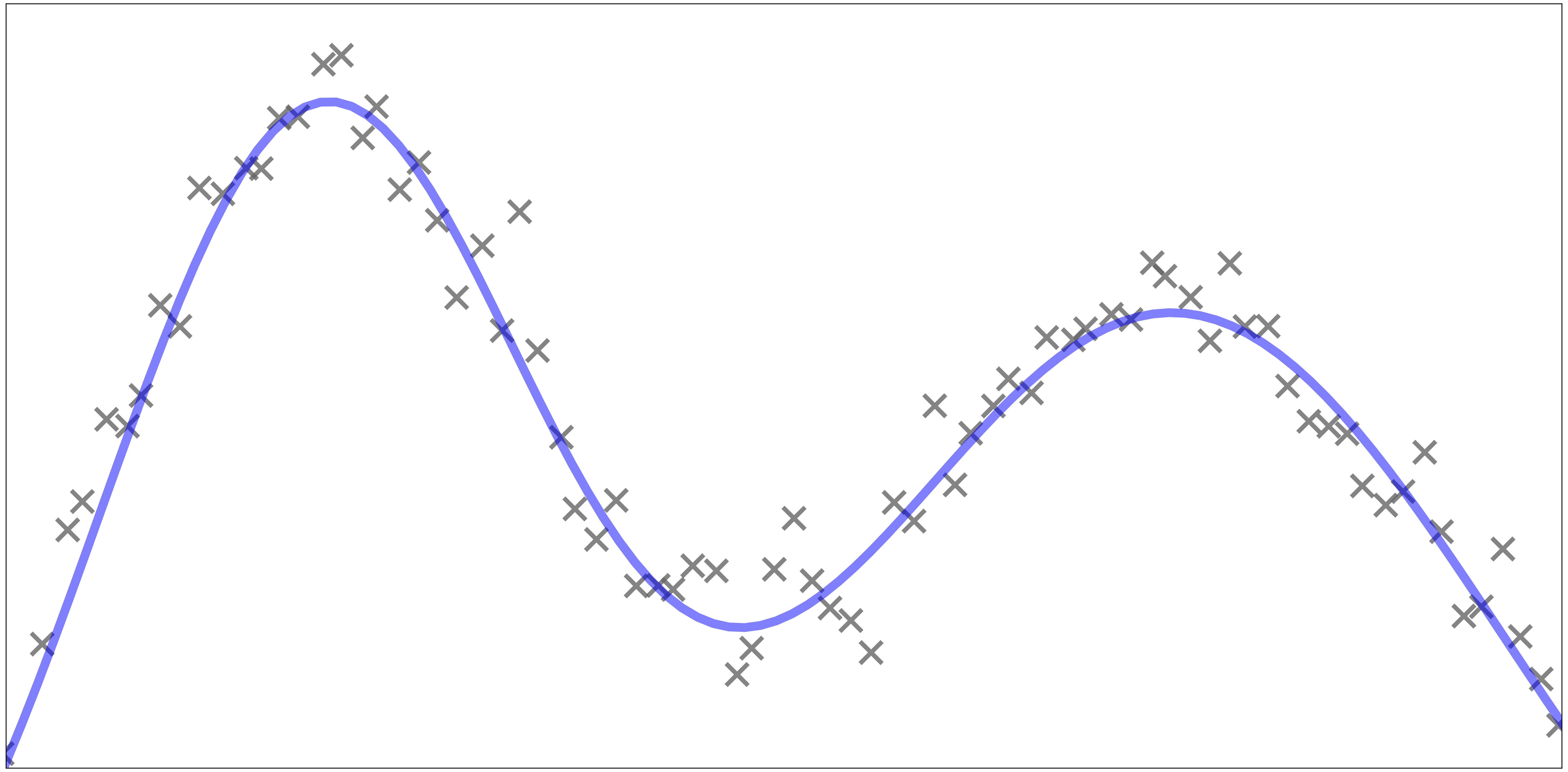


If others are already contributing large amounts of data, an agent has no incentive to collect/contribute data of her own.

FREE-RIDING (#3/3)

A seemingly plausible work-around (but does not work):

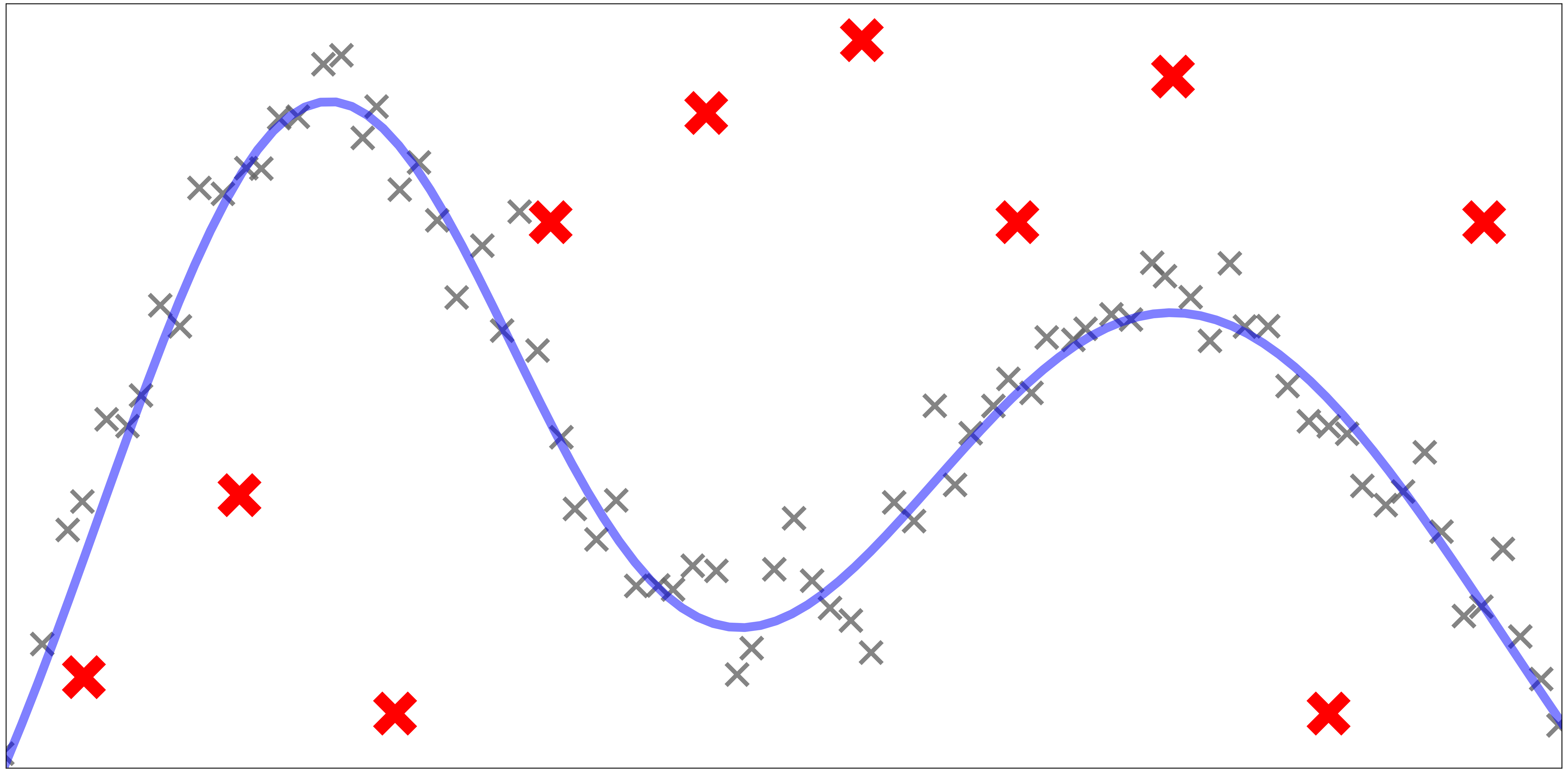
Pool-and-share but only if the agent contributes sufficient data



FREE-RIDING (#3/3)

A seemingly plausible work-around (but does not work):

Pool-and-share but only if the agent contributes sufficient data

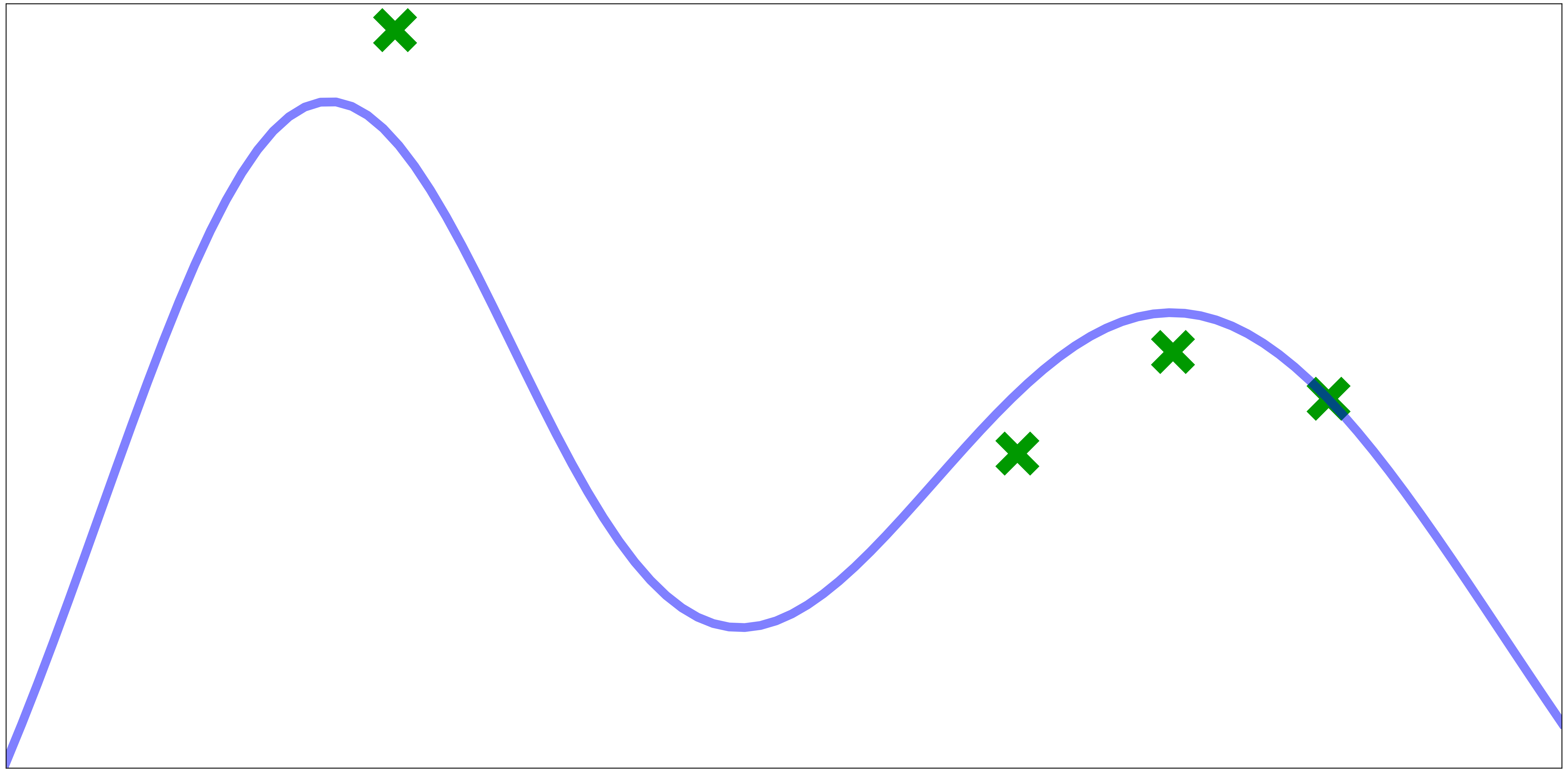


- Agent can submit fabricated data and then discard it when learning.

FREE-RIDING (#3/3)

A seemingly plausible work-around (but does not work):

Pool-and-share but only if the agent contributes sufficient data

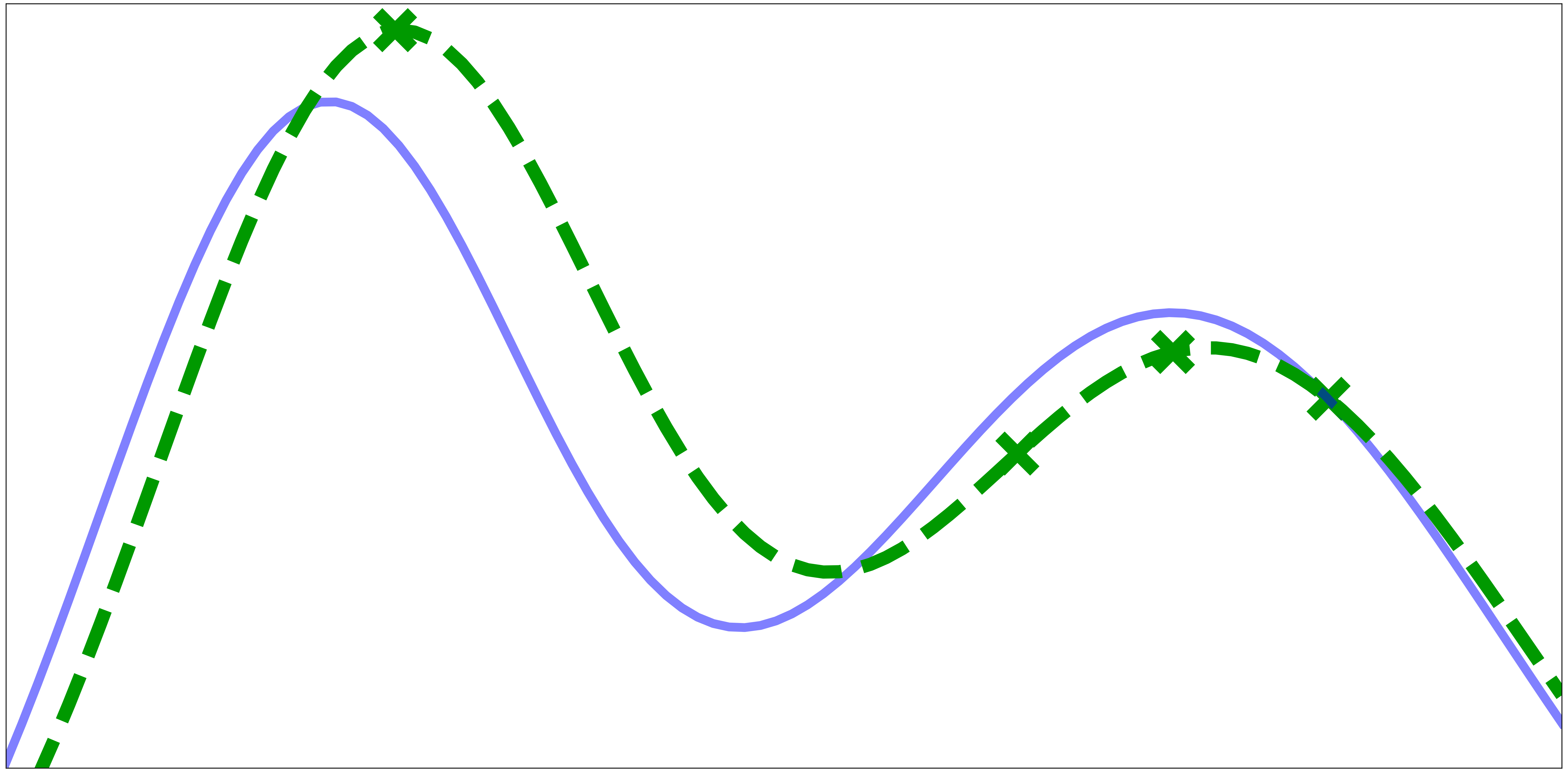


- Agent can submit fabricated data and then discard it when learning.

FREE-RIDING (#3/3)

A seemingly plausible work-around (but does not work):

Pool-and-share but only if the agent contributes sufficient data

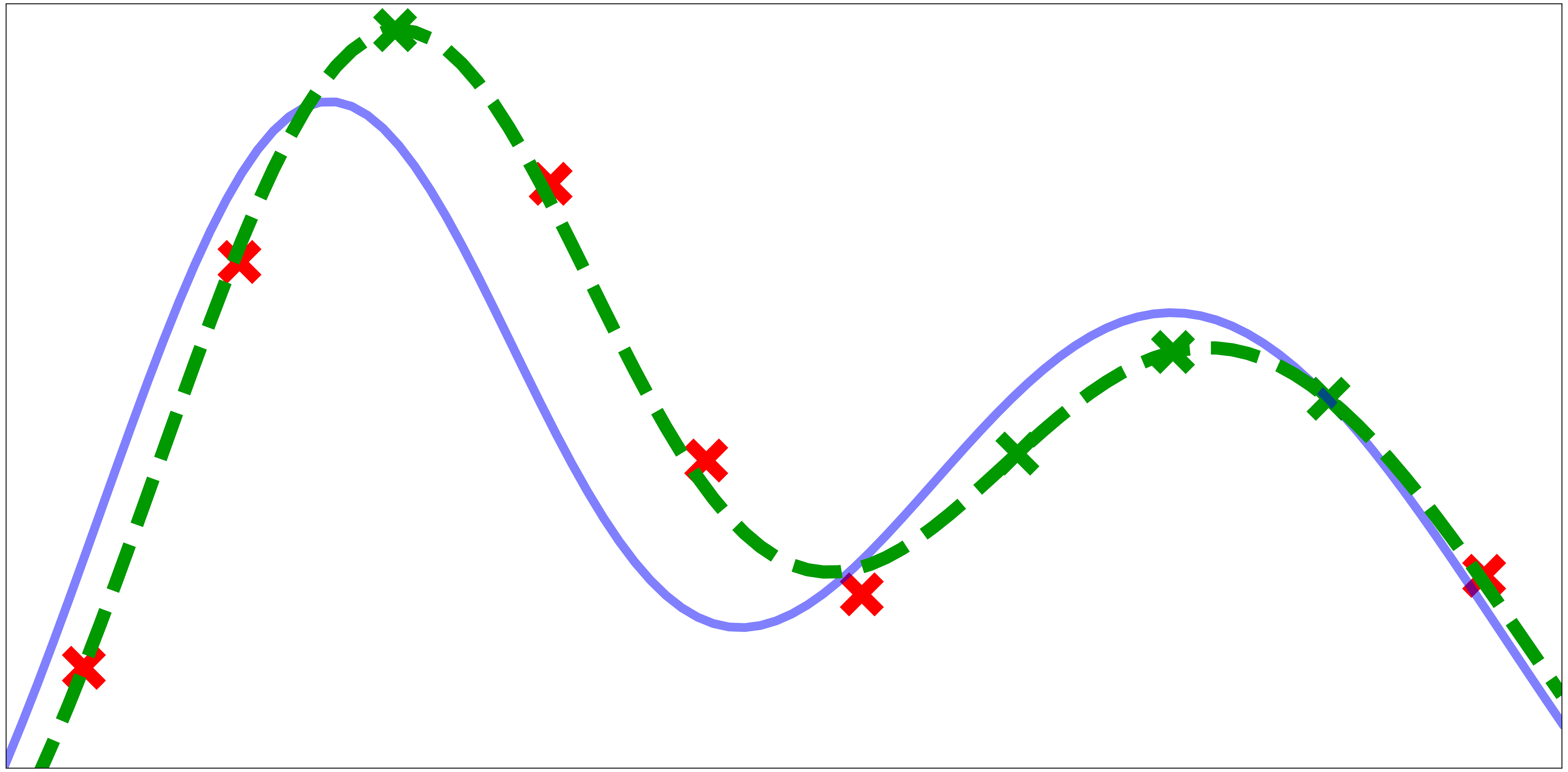


- Agent can submit fabricated data and then discard it when learning.

FREE-RIDING (#3/3)

A seemingly plausible work-around (but does not work):

Pool-and-share but only if the agent contributes sufficient data

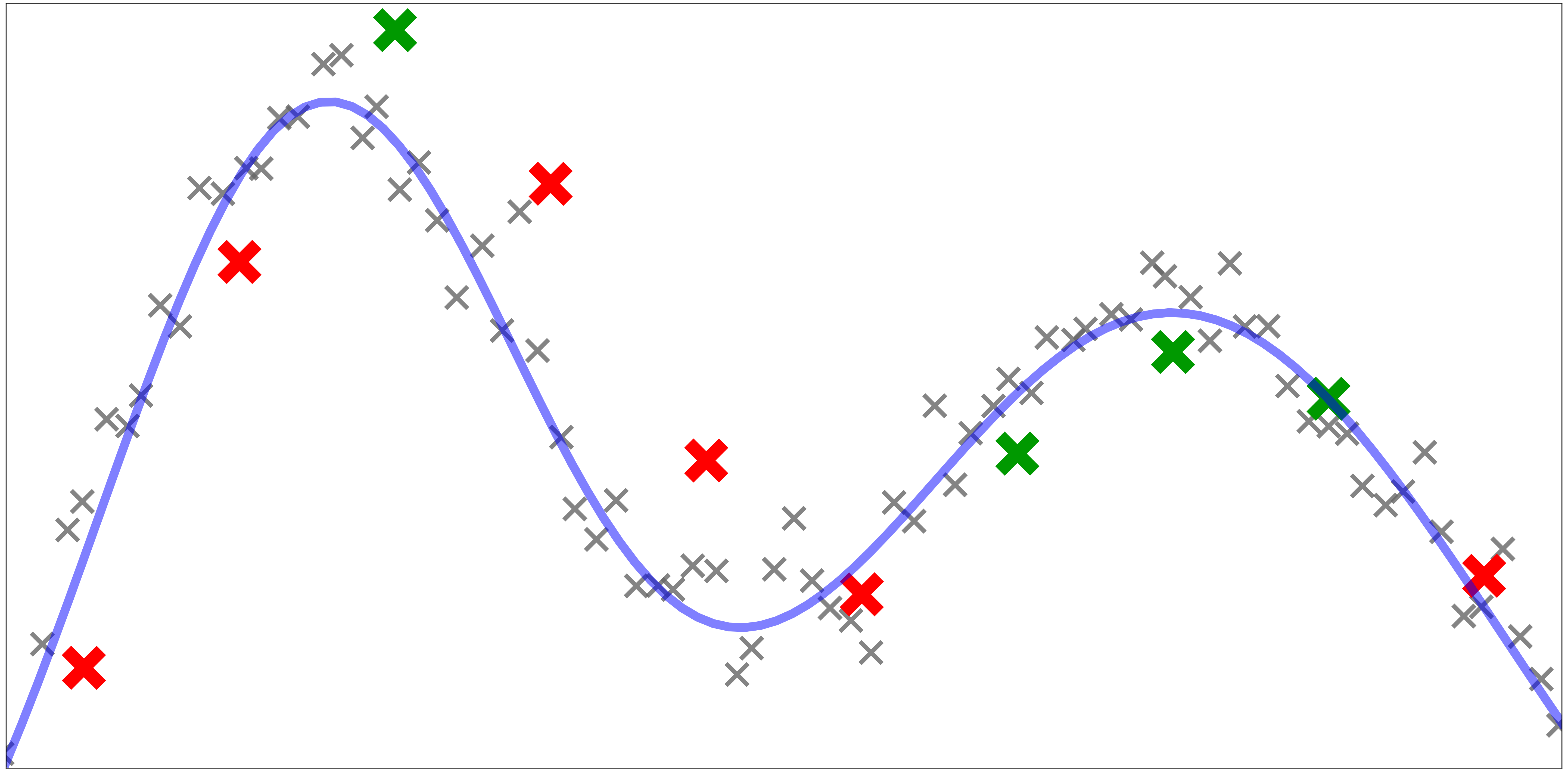


- Agent can submit fabricated data and then discard it when learning.
- Agent may fabricate based on a small sample she has collected, so it may not always be easy to detect.

FREE-RIDING (#3/3)

A seemingly plausible work-around (but does not work):

Pool-and-share but only if the agent contributes sufficient data



- Agent can submit fabricated data and then discard it when learning.
- Agent may fabricate based on a small sample she has collected, so it may not always be easy to detect.

OBSTACLES TO DATA SHARING

Ethical/Legal

Privacy

Ownership of data

Security

Data breaches

Adversarial attacks

Logistical

Inter-operability

Communication costs

Incentives

Free-riding

Competition

Data monetization

Data valuation

BUT THERE IS A DEMAND FOR DATA SHARING IN THE REAL WORLD

Data sharing platforms/consortia

PubChem

PubMed®

ChEMBL

DELTA SHARING

An open standard for secure data sharing

Marketplaces for data and ML models

aws AWS Data Exchange

Ads Data Hub

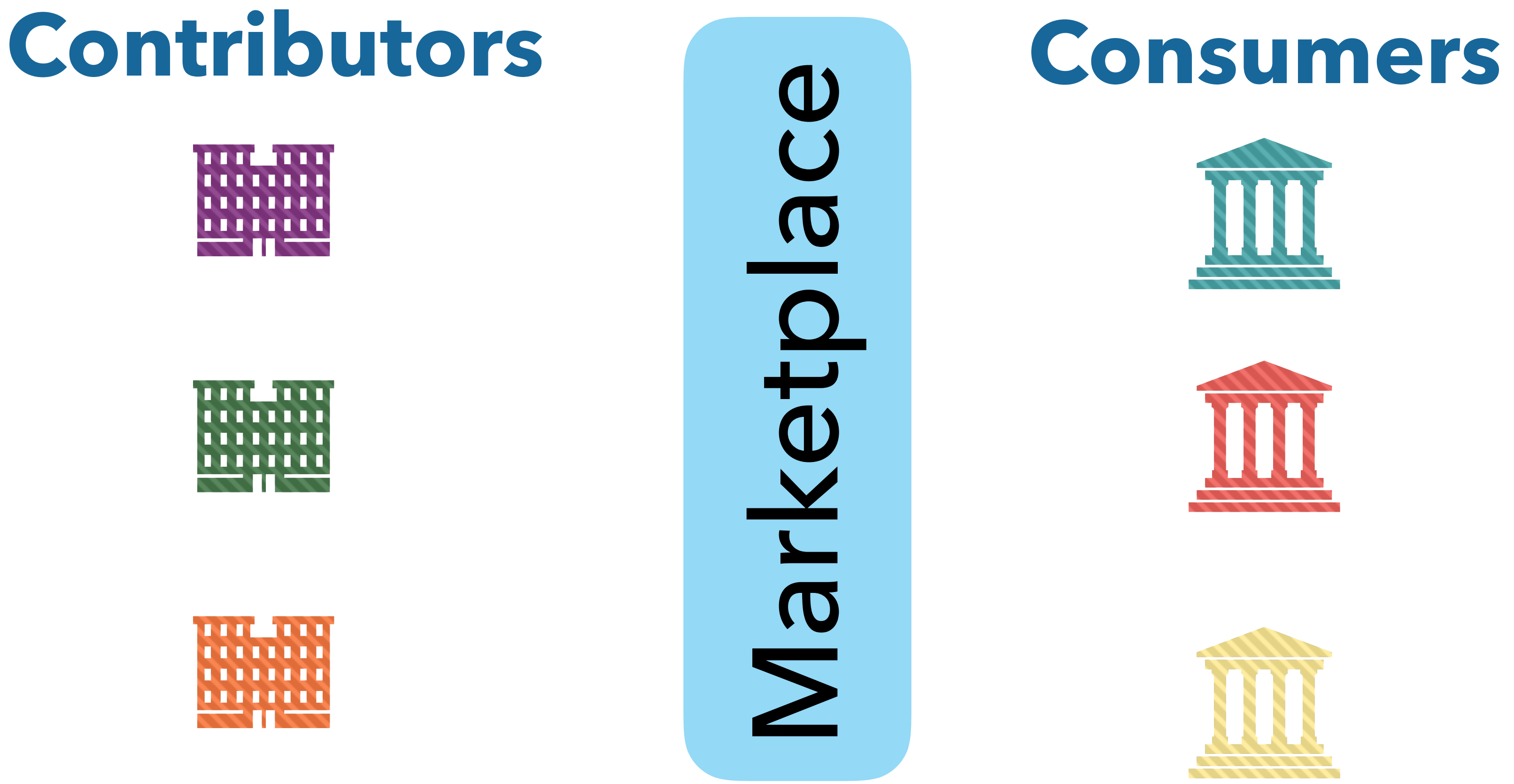
CITRINE INFORMATICS

Azure Data Share

OUR APPROACHES

Mechanisms for data sharing and federated learning

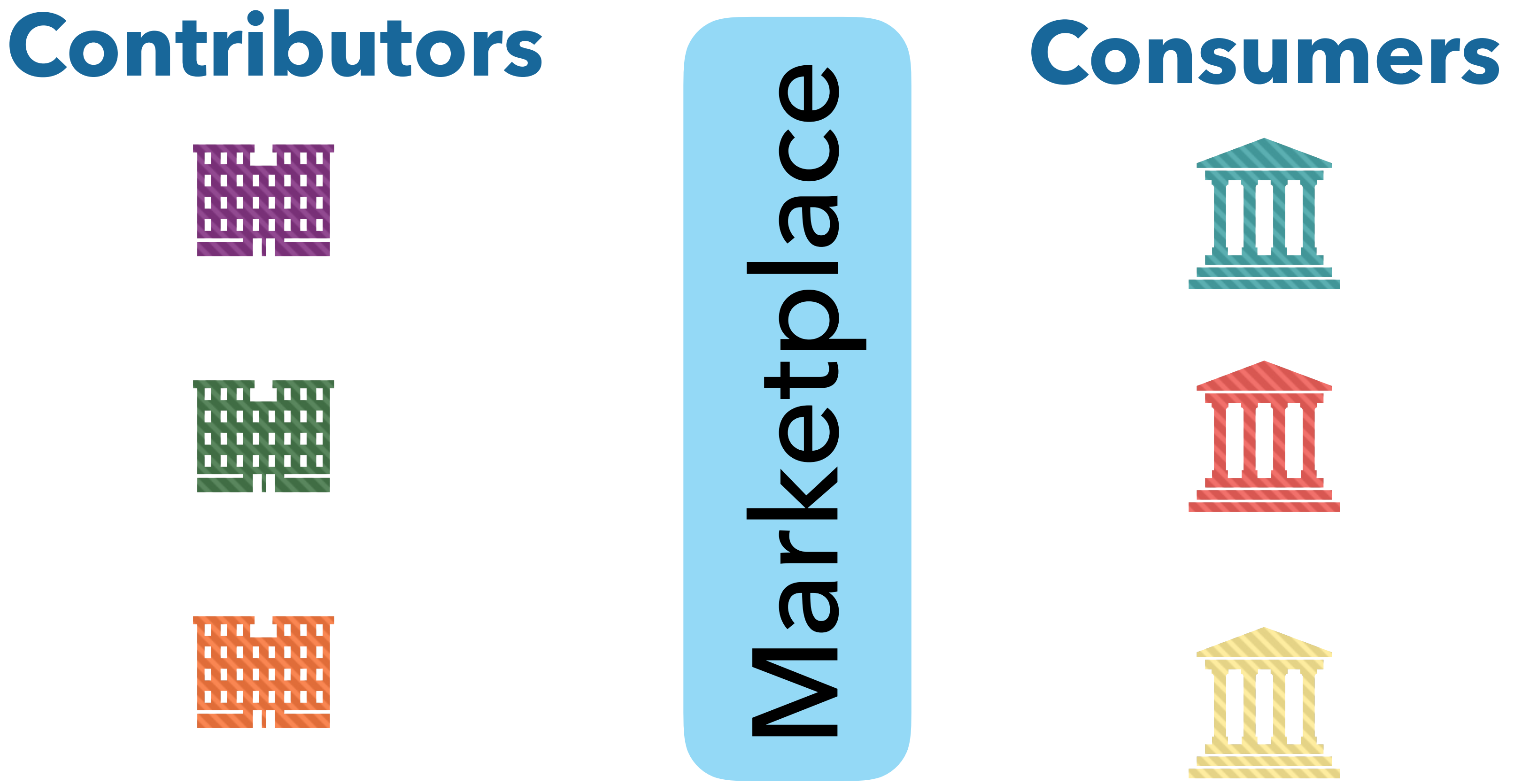
Data marketplaces



Mechanisms for data sharing and federated learning

Goal: Incentivize agents to collect as much data and share it honestly.

Data marketplaces

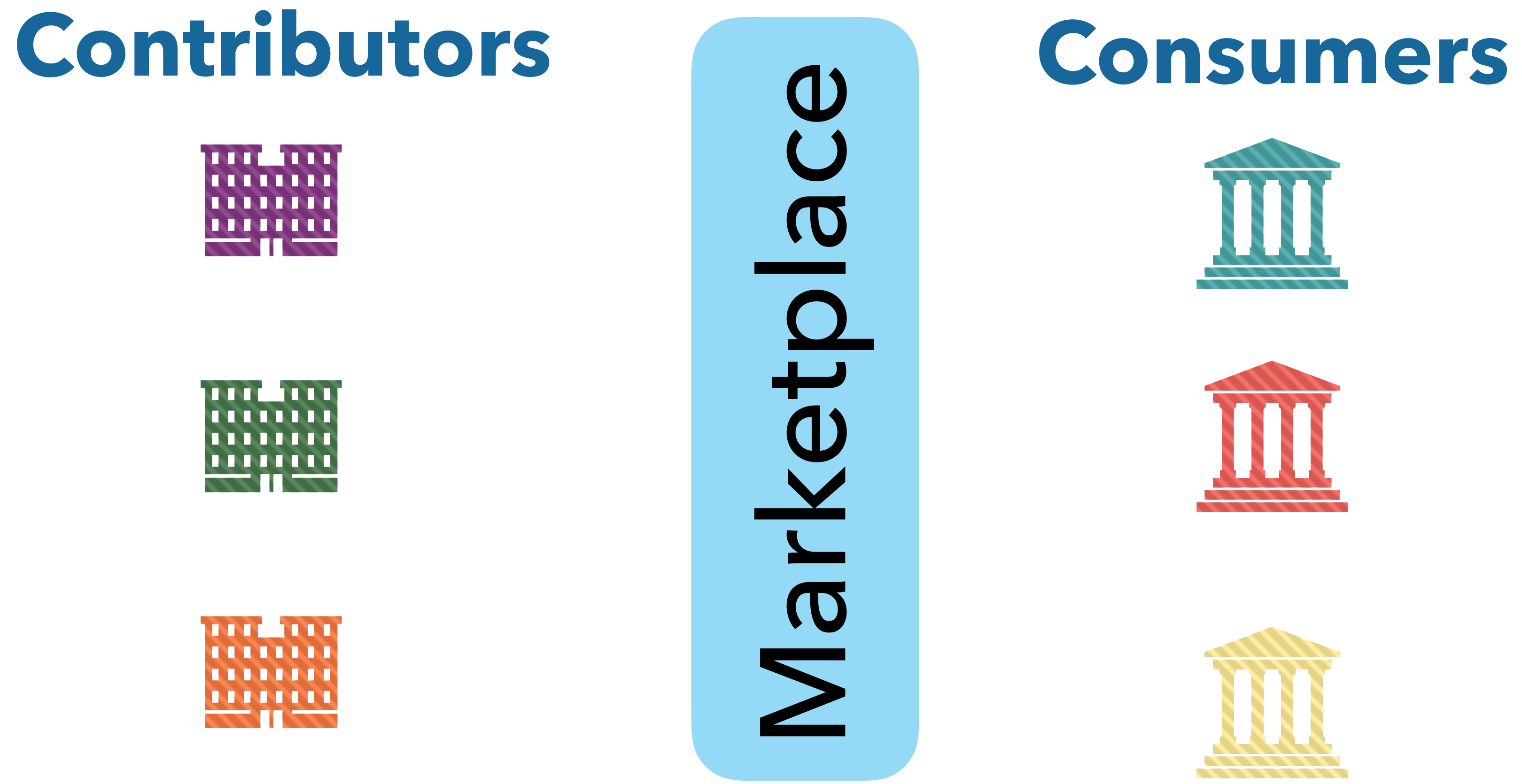


Mechanisms for data sharing and federated learning

Goal: Incentivize agents to collect as much data and share it honestly.

- Do not simply pool and share data!
- Cross-check for quality of the data contributed.

Data marketplaces

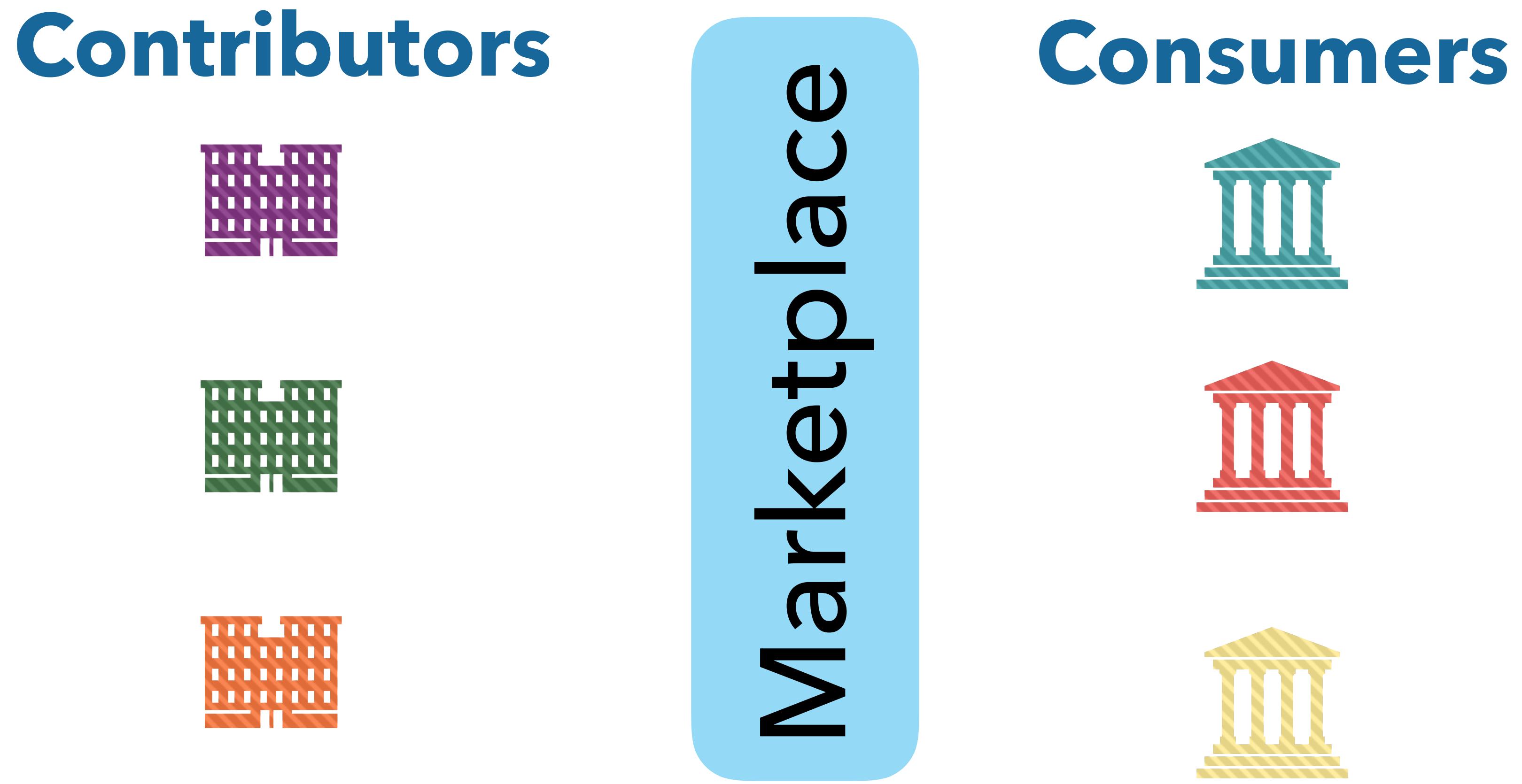


Mechanisms for data sharing and federated learning

Goal: Incentivize agents to collect as much data and share it honestly.

- Do not simply pool and share data!
- Cross-check for quality of the data contributed.
- More/better data contributed $\Longrightarrow$ more/better data received.

Data marketplaces

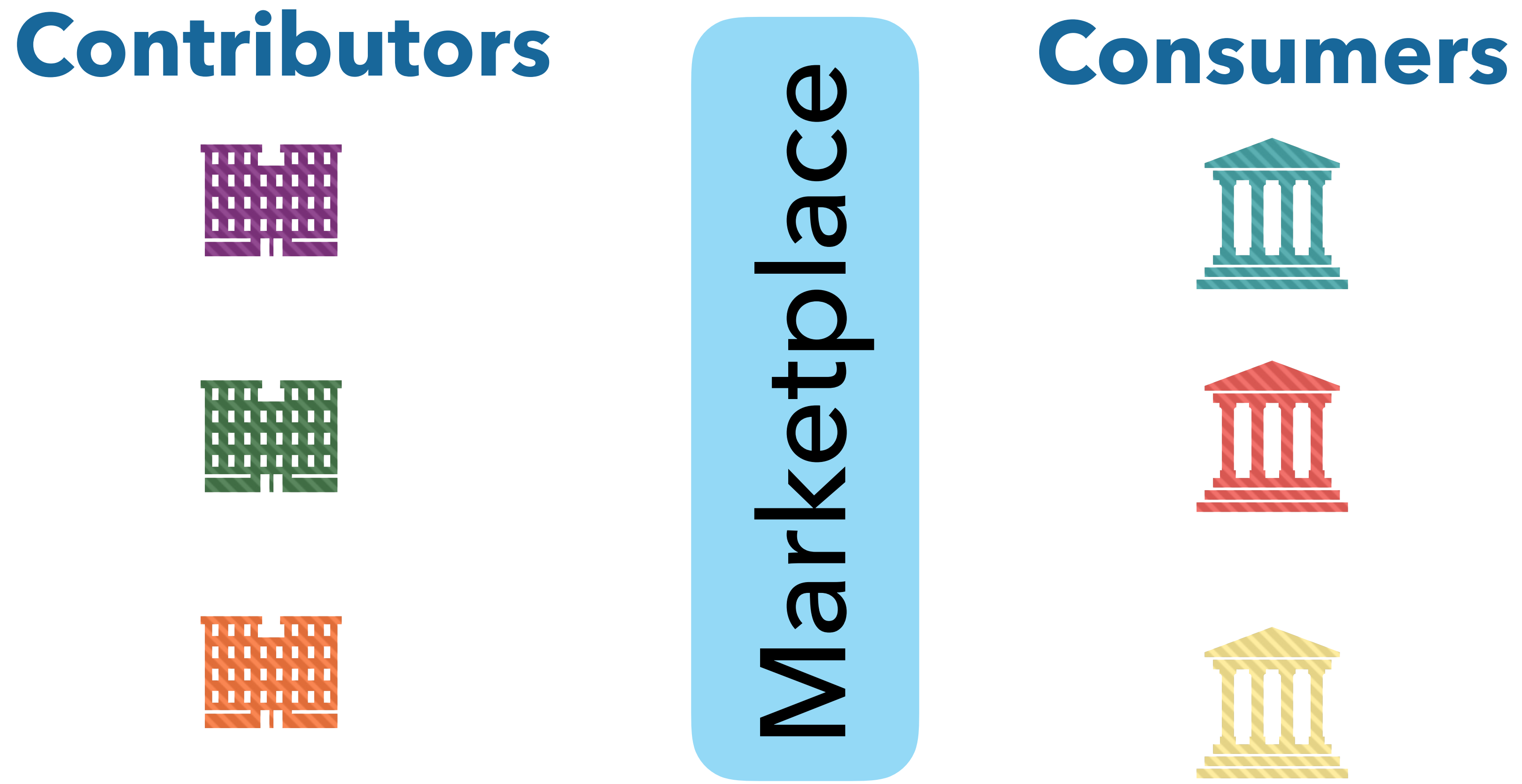


Mechanisms for data sharing and federated learning

Goal: Incentivize agents to collect as much data and share it honestly.

- Do not simply pool and share data!
- Cross-check for quality of the data contributed.
- More/better data contributed $\Longrightarrow$ more/better data received.

Data marketplaces



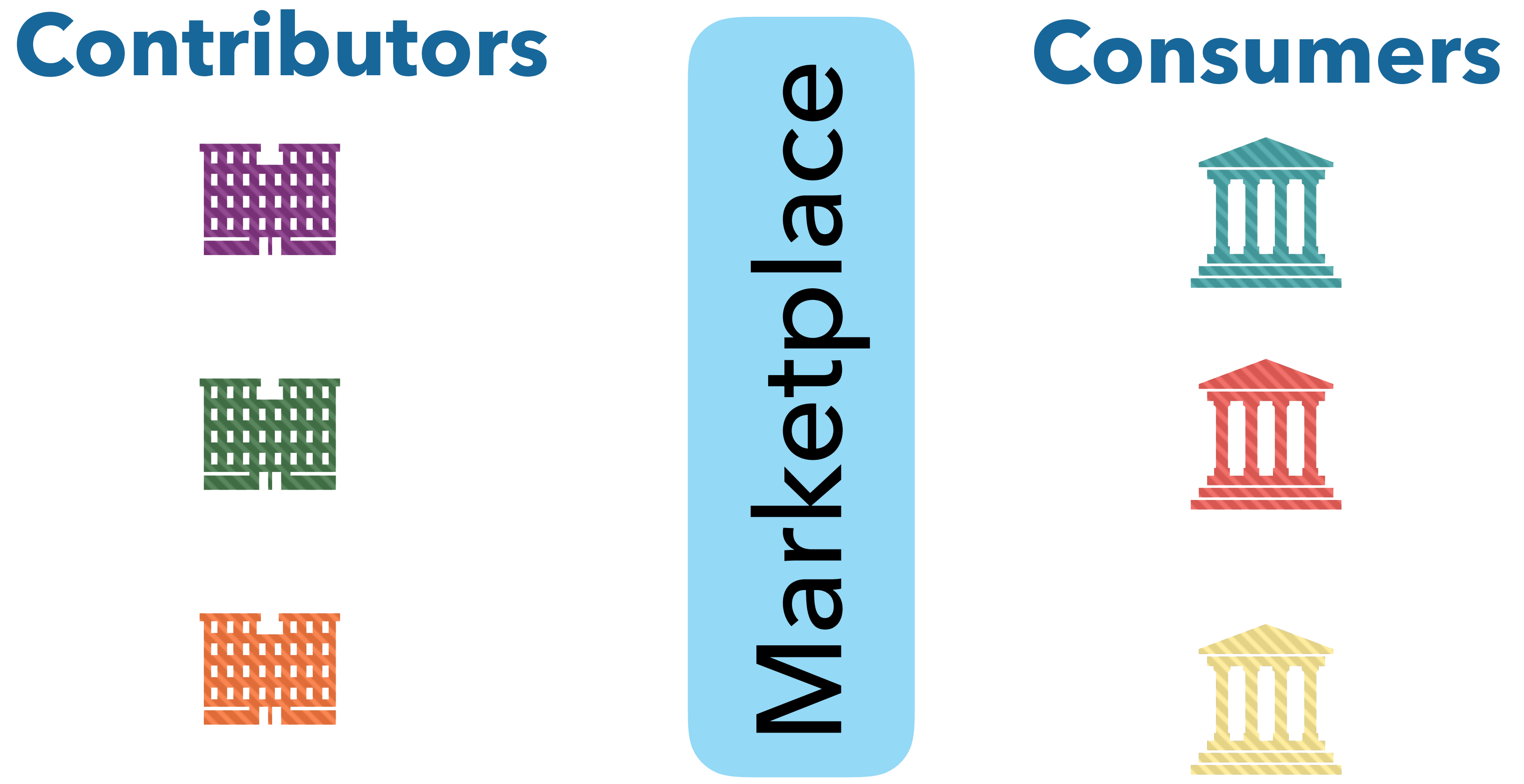
Goal: Incentivize contributors to honestly contribute lots of data. Fairly reward them for effort via payments from consumers.

Mechanisms for data sharing and federated learning

Goal: Incentivize agents to collect as much data and share it honestly.

- Do not simply pool and share data!
- Cross-check for quality of the data contributed.
- More/better data contributed $\Longrightarrow$ more/better data received.

Data marketplaces



Goal: Incentivize contributors to honestly contribute lots of data. Fairly reward them for effort via payments from consumers.

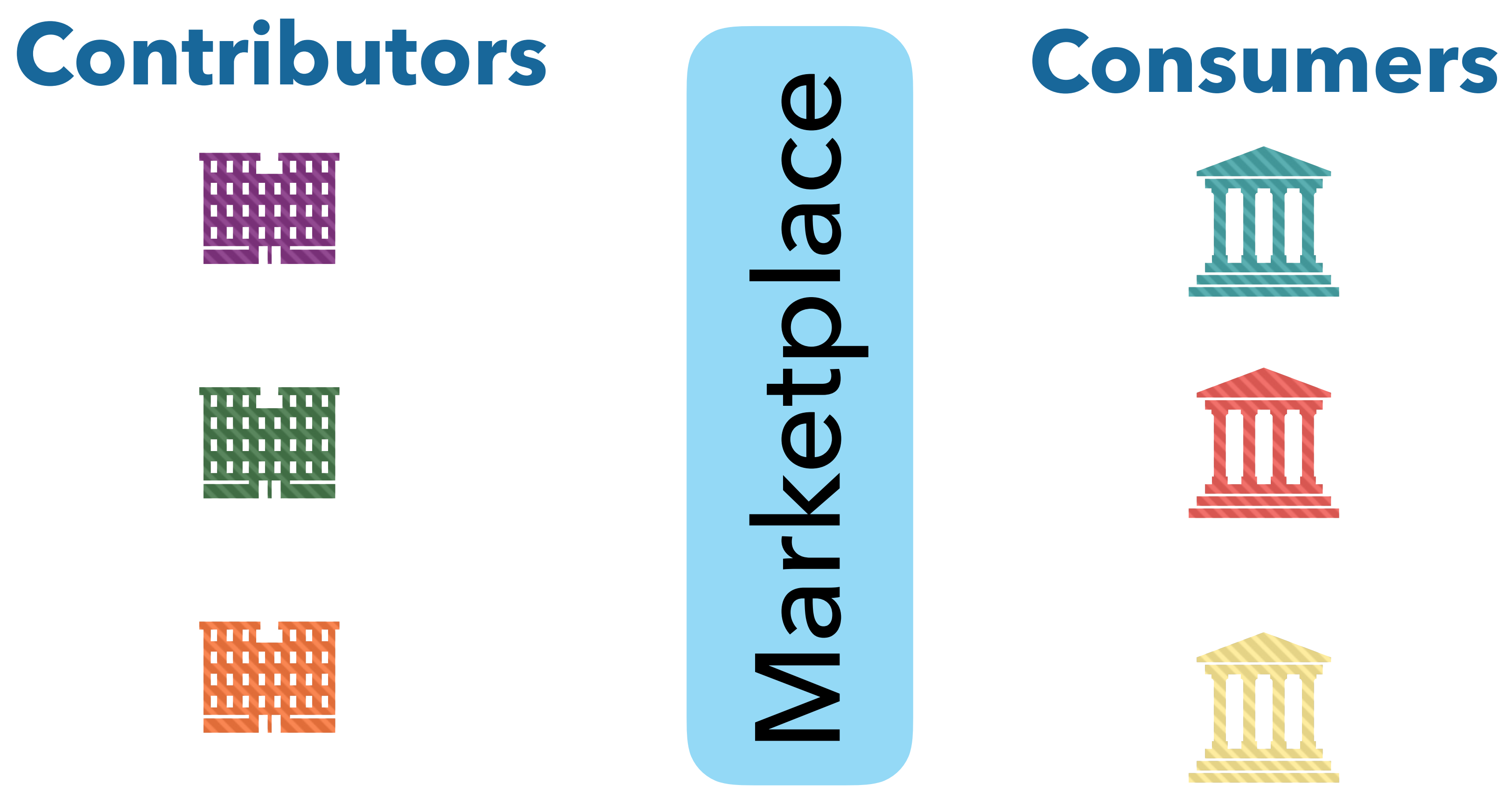
- A mediator checks for the quality of the data from contributors.

Mechanisms for data sharing and federated learning

Goal: Incentivize agents to collect as much data and share it honestly.

- Do not simply pool and share data!
- Cross-check for quality of the data contributed.
- More/better data contributed $\Longrightarrow$ more/better data received.

Data marketplaces



Goal: Incentivize contributors to honestly contribute lots of data. Fairly reward them for effort via payments from consumers.

- A mediator checks for the quality of the data from contributors.
- Higher quality data $\Longrightarrow$ higher revenue for data contributors.

PRIOR WORK

Mechanisms for data sharing and federated learning

Sim, Zhang, Chan, Low 2020
Xu, Lyu, Ma et al 2021
Blum, Haghtalab, Phillips, Shao 2021
Karimireddy, Guo, Jordan 2022
Fraboni, Vidal, Lorenzi 2021
Lin, Du, Liu 2019
Ding, Fang, Huang 2020
Liu, Tian, Chen et al 2022

Data marketplaces

Cai, Daskalakis, Papadimitriou 2015
Agarwal, Dahleh, Sarkar, 2019
Agarwal, Dahleh, Horel, Rui, 2020
Jia, Dao, Wang et al, 2019
Wang, Rausch, Zhang et al 2020

Key difference:

- All these works assume agents will always truthfully submit the data they have, i.e without fabrication/alteration.

OUTLINE

1. **Mechanism design for collaborative normal mean estimation**
   (Y. Chen, Zhu, Kandasamy, *NeurIPS 2023*)

2. **High-dimensional mean estimation with varied collection costs**
   (Clinton, Y. Chen, Zhu, Kandasamy, *Ongoing work*)

3. **Learning to price data in data marketplaces**
   (K. Chen, Huh, Kandasamy, *Ongoing work*)

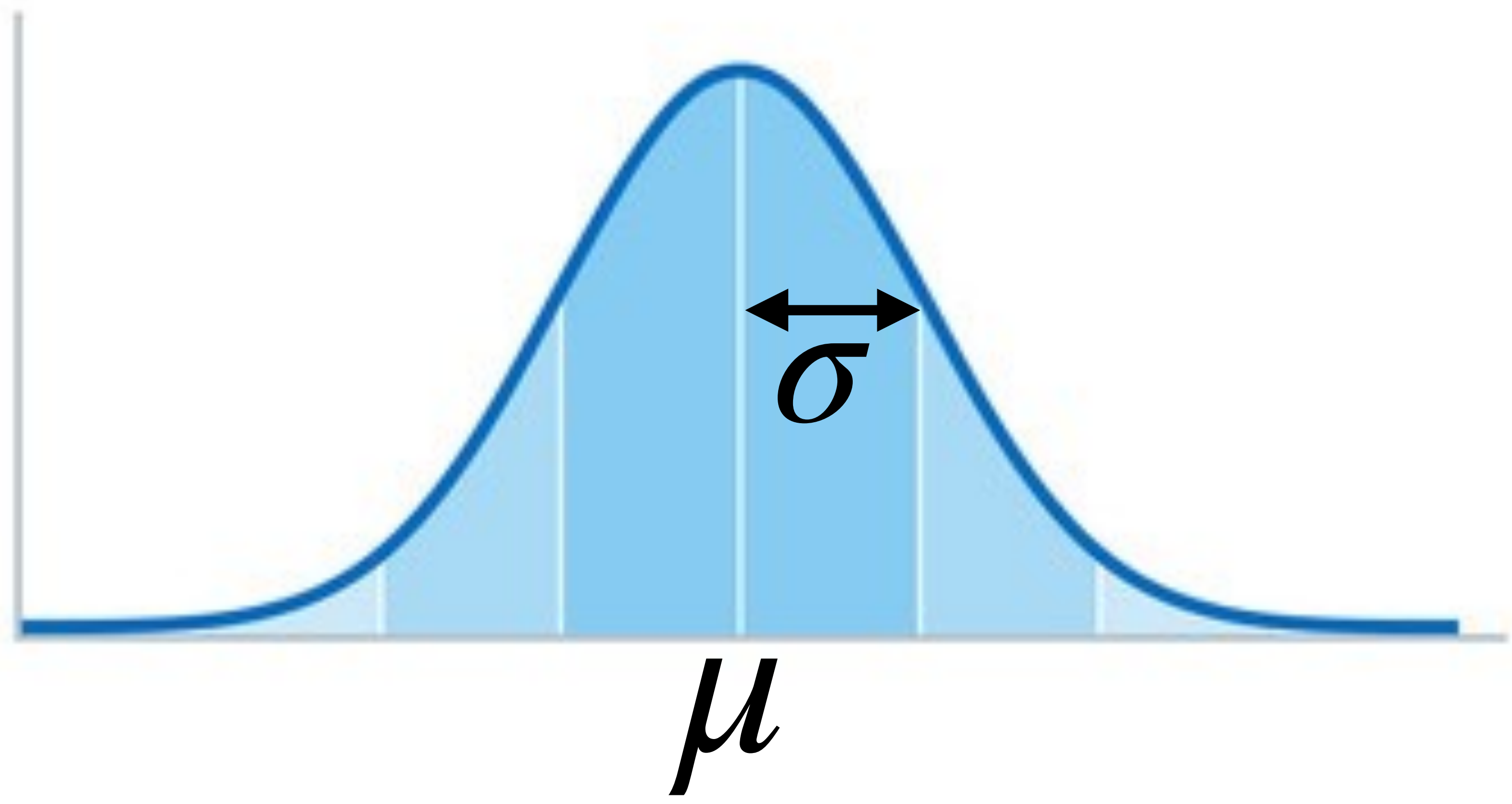
OUTLINE

1. **Mechanism design for collaborative normal mean estimation**
   (Y. Chen, Zhu, Kandasamy, *NeurIPS 2023*)

2. **High-dimensional mean estimation with varied collection costs**
   (Clinton, Y. Chen, Zhu, Kandasamy, *Ongoing work*)

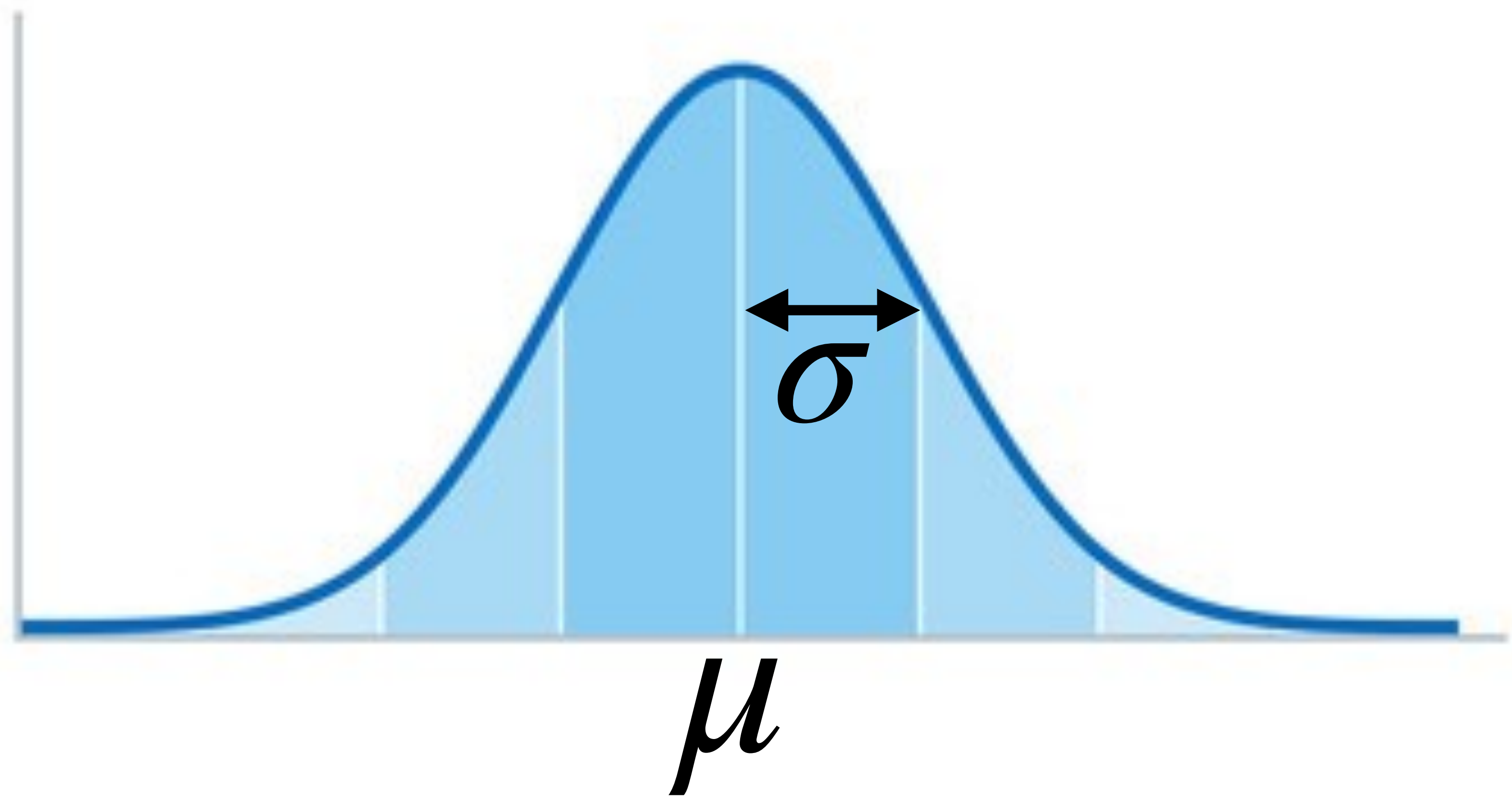
3. **Learning to price data in data marketplaces**
   (K. Chen, Huh, Kandasamy, *Ongoing work*)

NORMAL MEAN ESTIMATION



NORMAL MEAN ESTIMATION

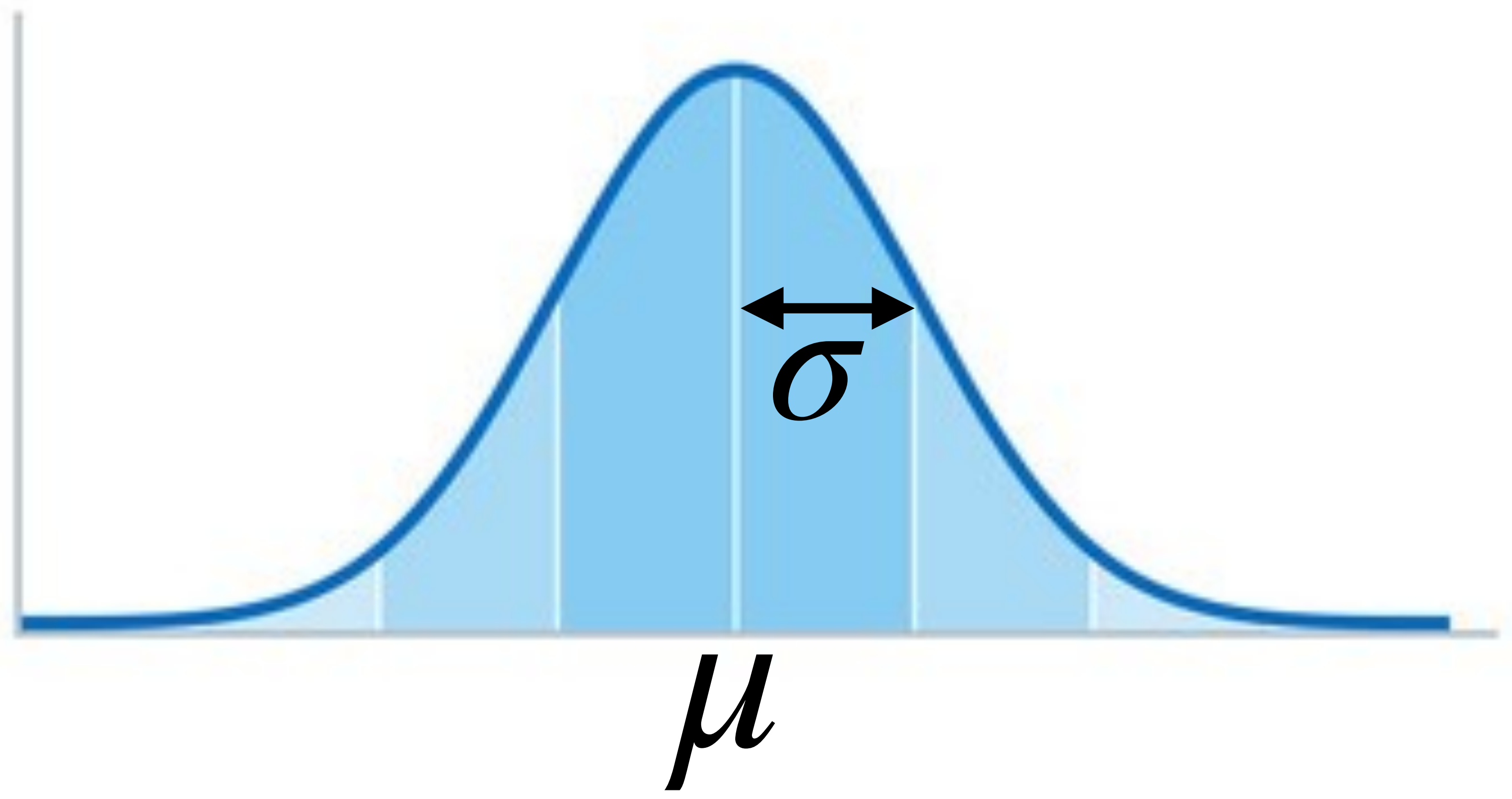
- Estimate the mean μ of a normal distribution with *known* variance σ^2.



NORMAL MEAN ESTIMATION

- Estimate the mean μ of a normal distribution with *known* variance σ^2.

- An agent can collect samples at *known* unit cost c.

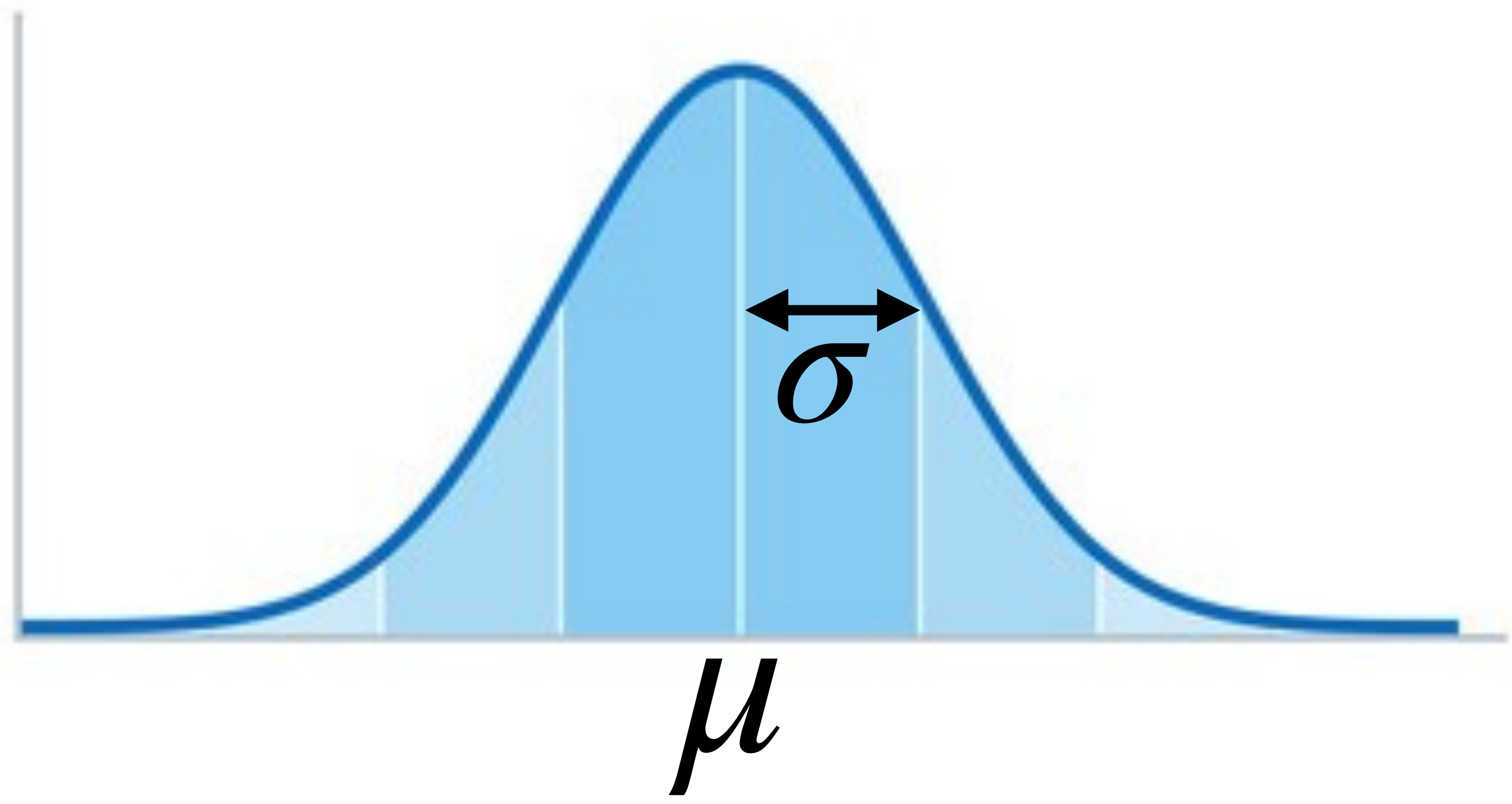


NORMAL MEAN ESTIMATION

- Estimate the mean μ of a normal distribution with *known* variance σ^2.

- An agent can collect samples at *known* unit cost c.

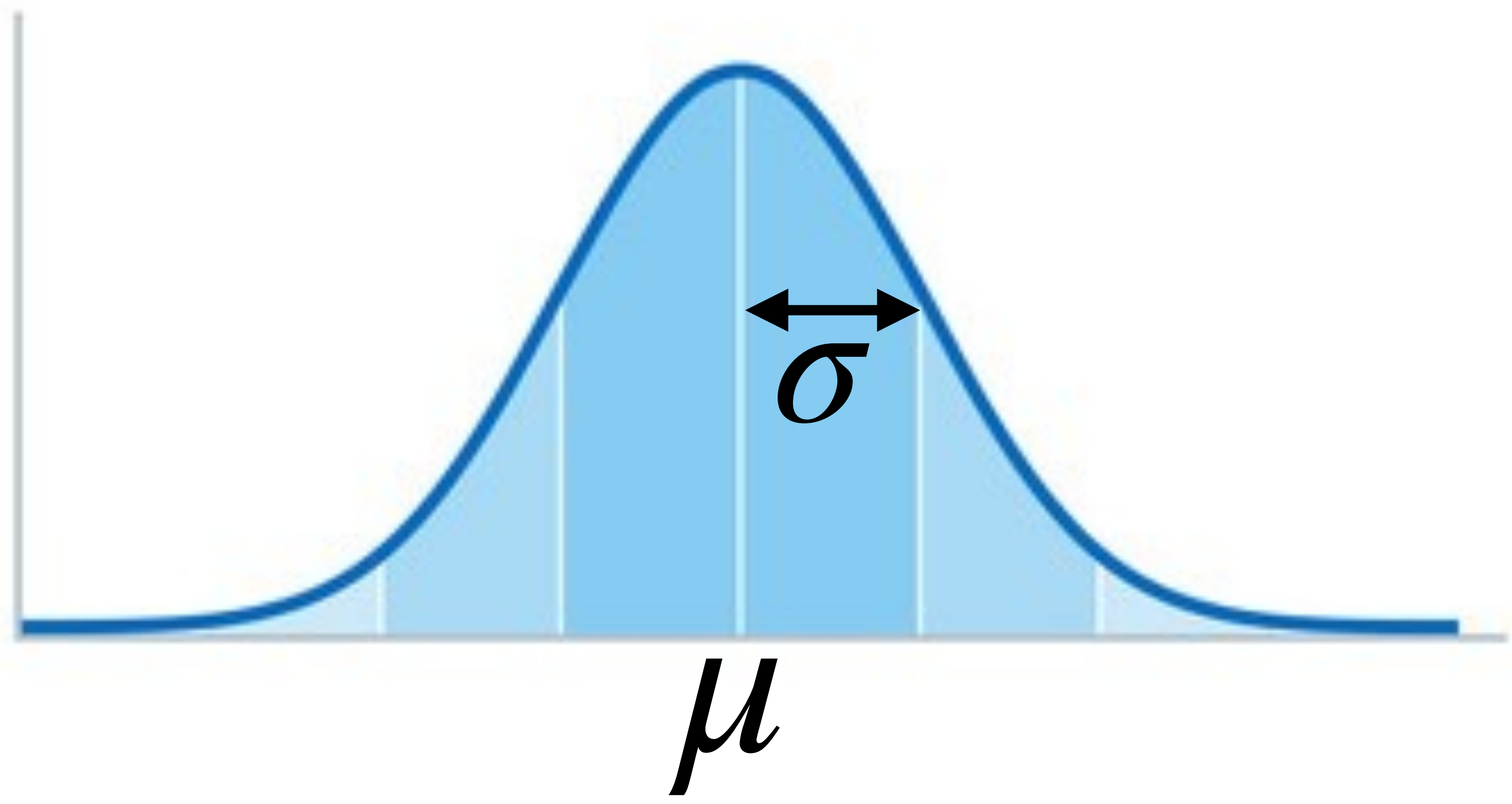
- Each agent wishes to minimize



$$\text{penalty} = \text{estimation error} + \text{data collection cost}$$

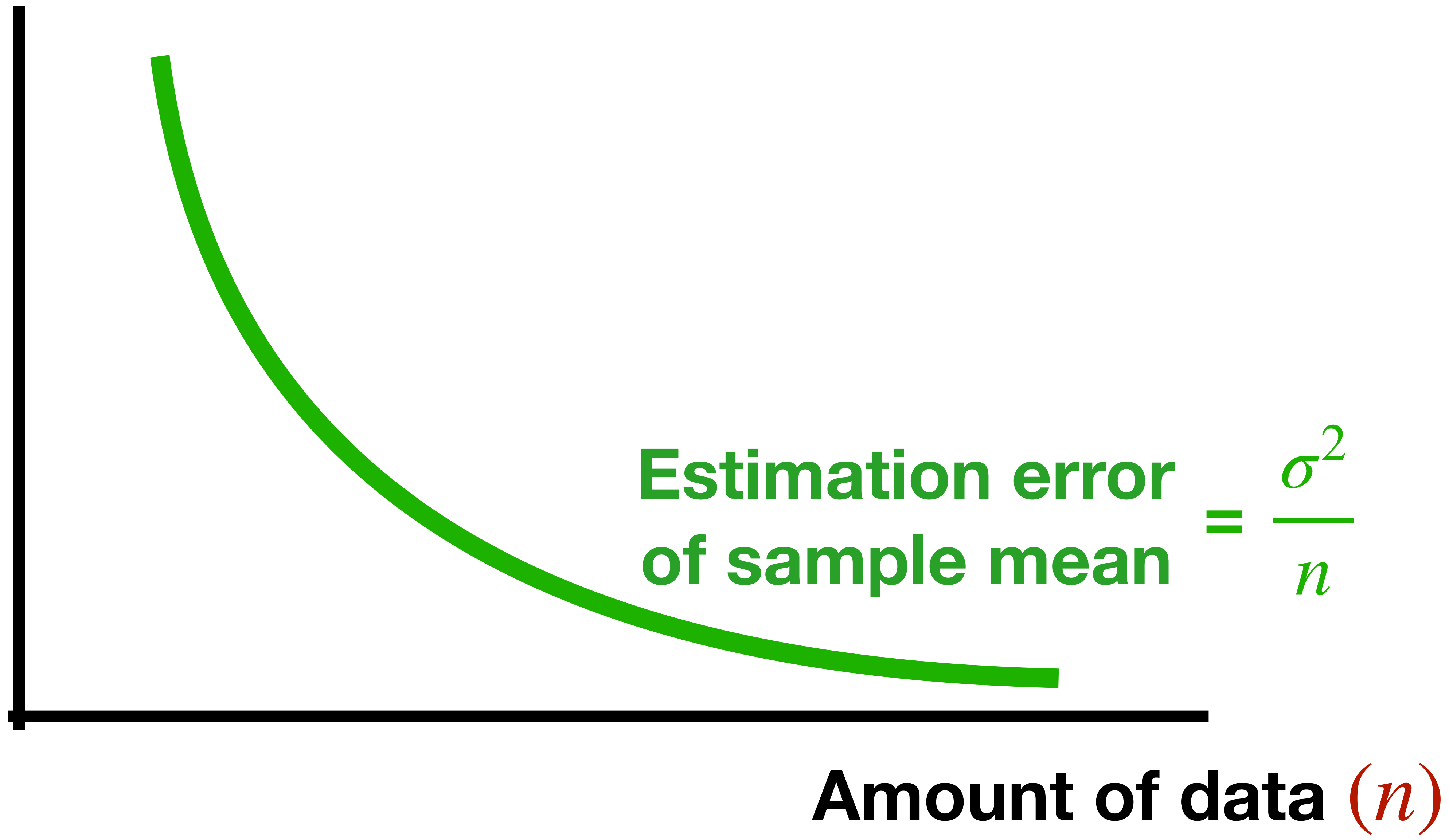
NORMAL MEAN ESTIMATION

- Estimate the mean μ of a normal distribution with *known* variance σ^2.



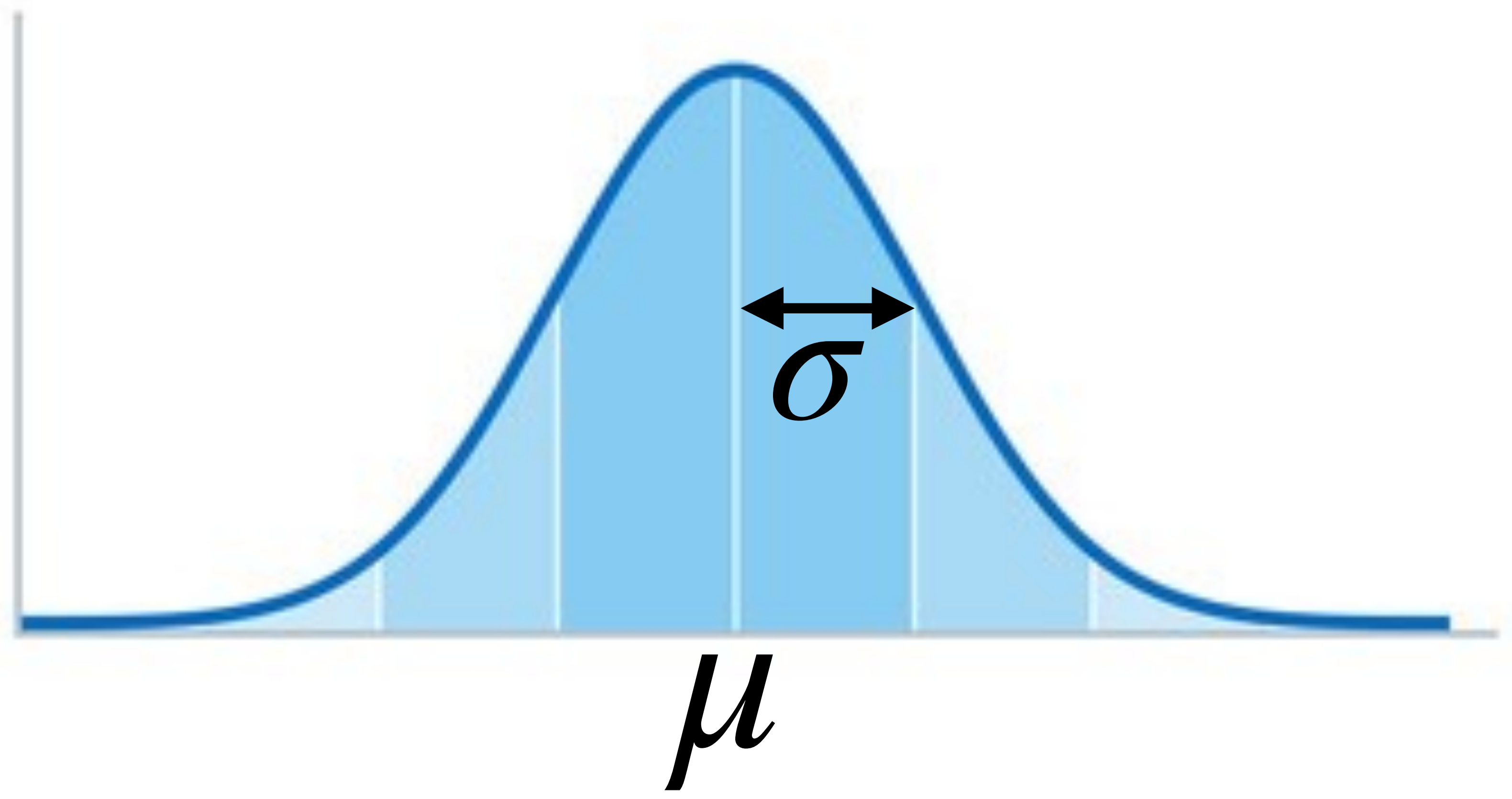
- An agent can collect samples at *known* unit cost c.
- Each agent wishes to minimize

$$\text{penalty} = \text{estimation error} + \text{data collection cost}$$
$$= \frac{\sigma^2}{n} + cn$$



NORMAL MEAN ESTIMATION

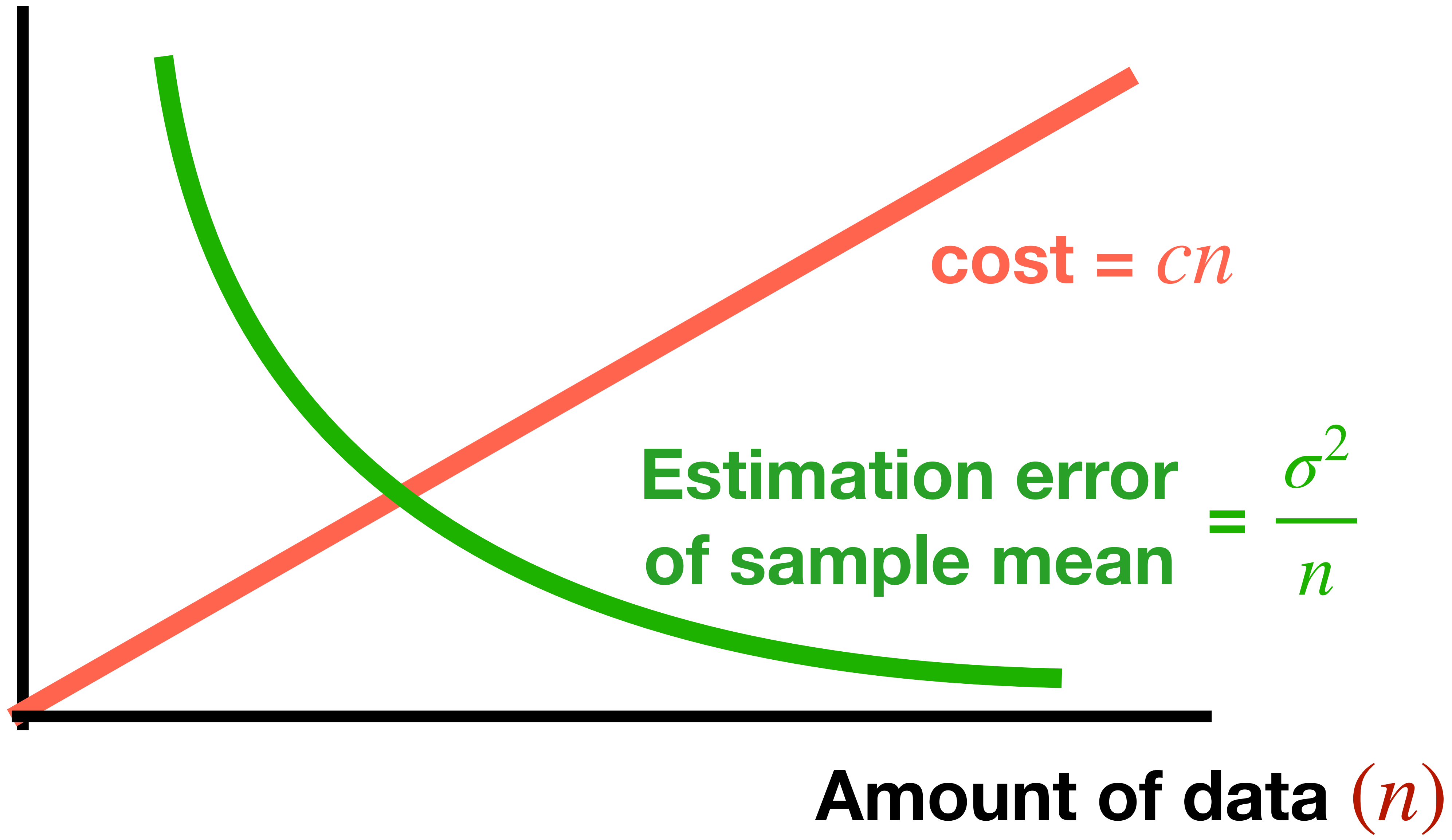
- Estimate the mean μ of a normal distribution with *known* variance σ^2.



- An agent can collect samples at *known* unit cost c.
- Each agent wishes to minimize

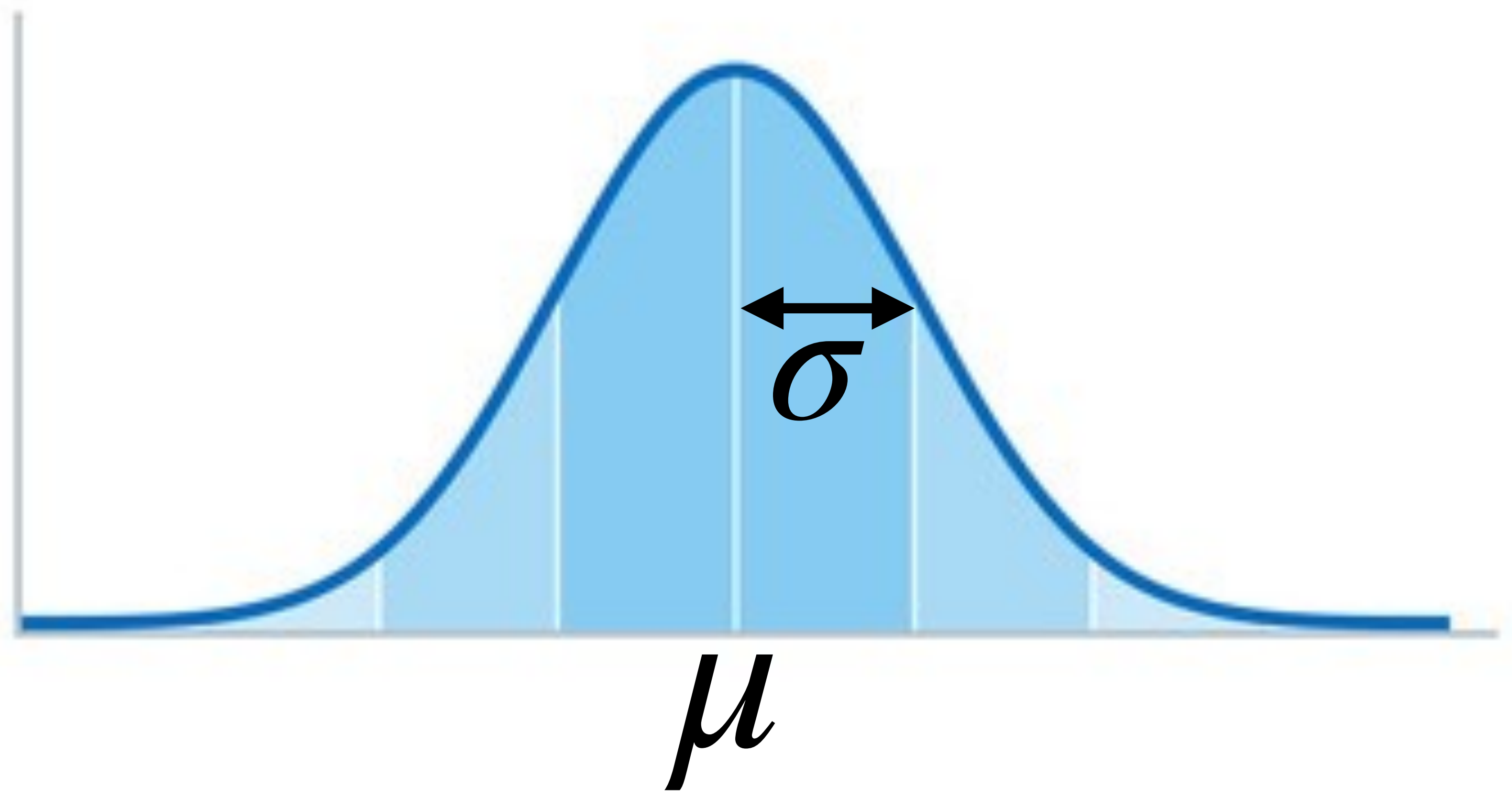
$$\text{penalty} = \text{estimation error} + \text{data collection cost}$$

$$= \frac{\sigma^2}{n} + cn$$



NORMAL MEAN ESTIMATION

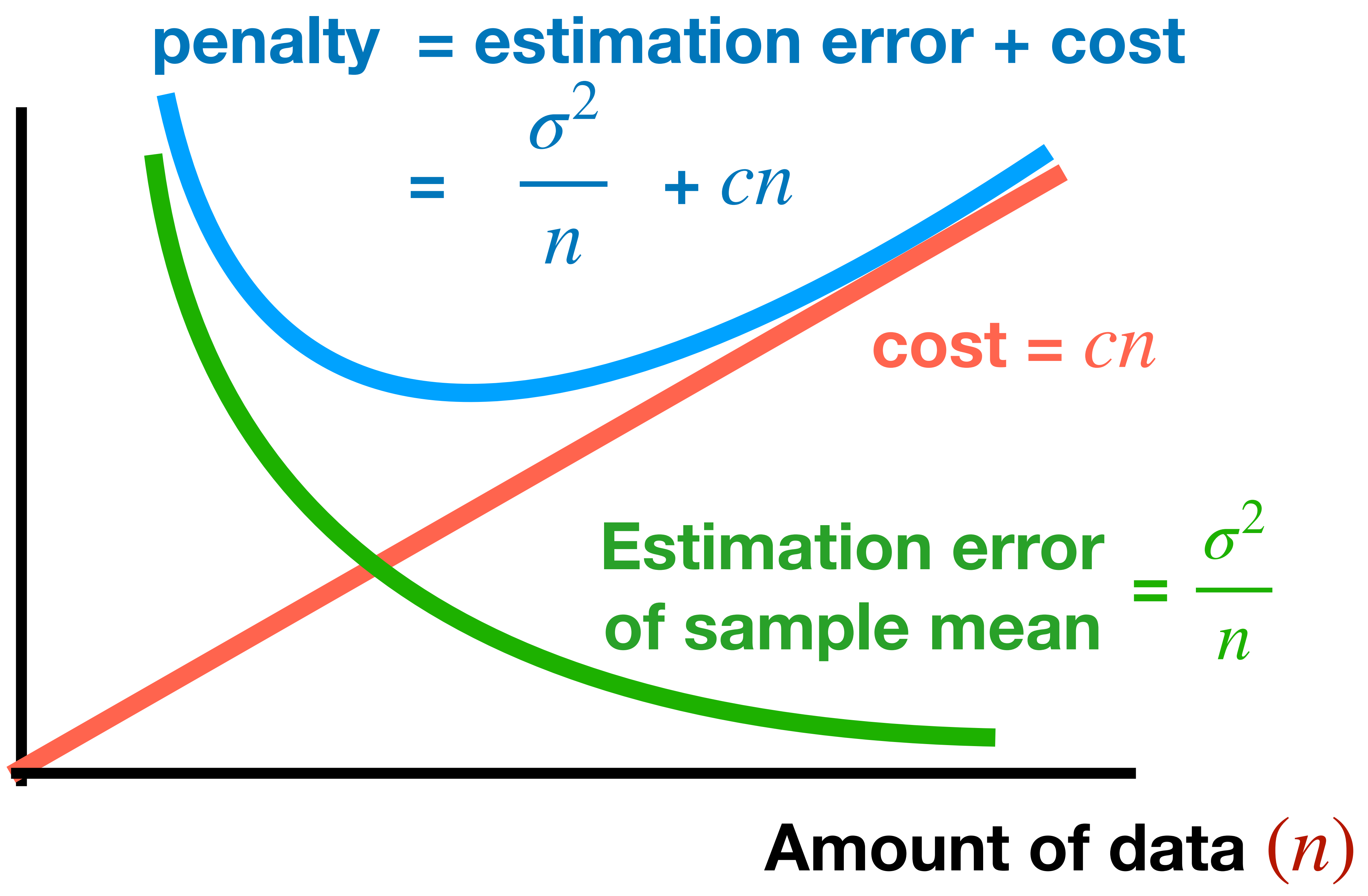
- Estimate the mean μ of a normal distribution with *known* variance σ^2.



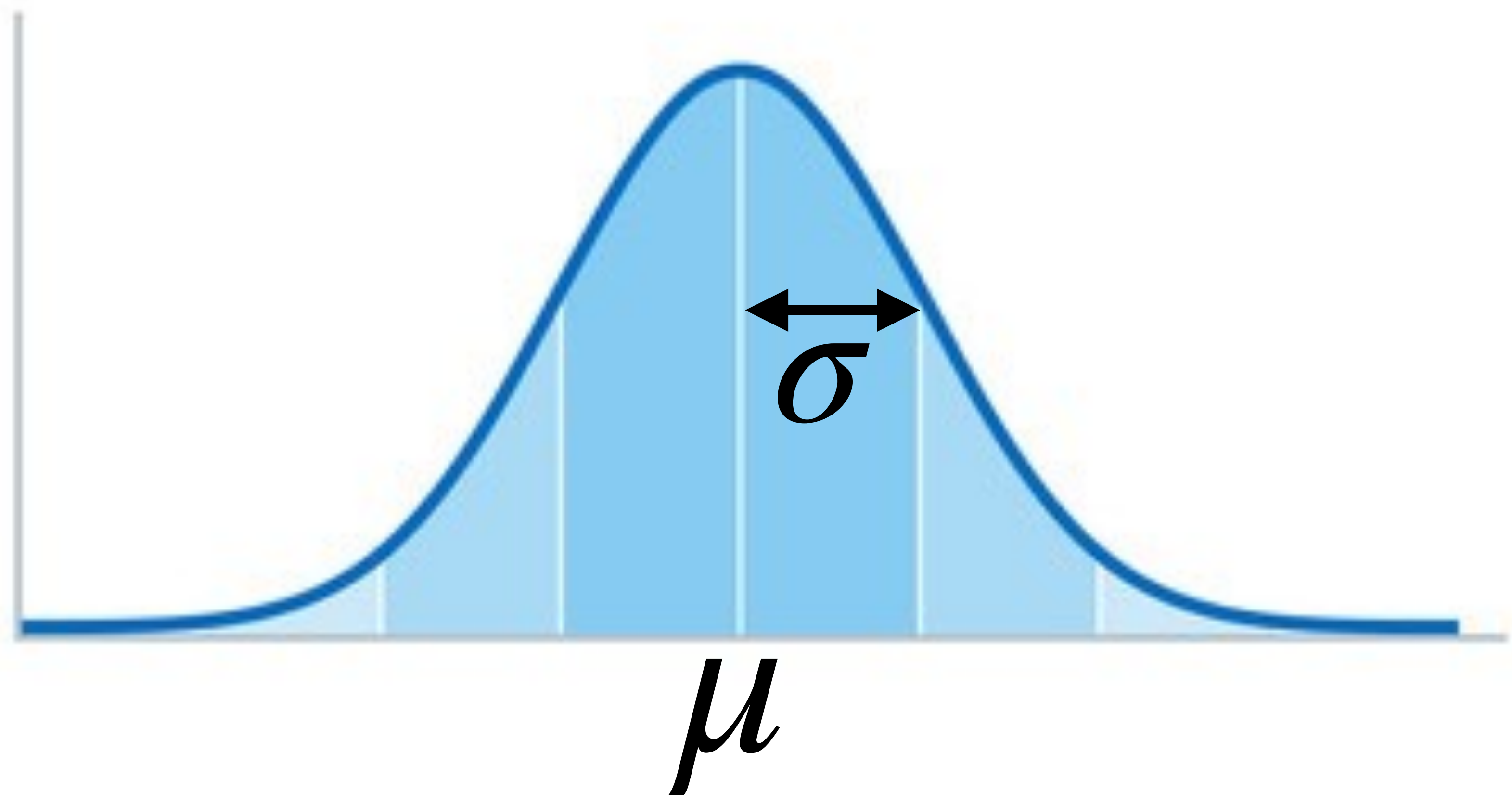
- An agent can collect samples at *known* unit cost c.
- Each agent wishes to minimize

$$\text{penalty} = \text{estimation error} + \text{data collection cost}$$

$$= \frac{\sigma^2}{n} + cn$$



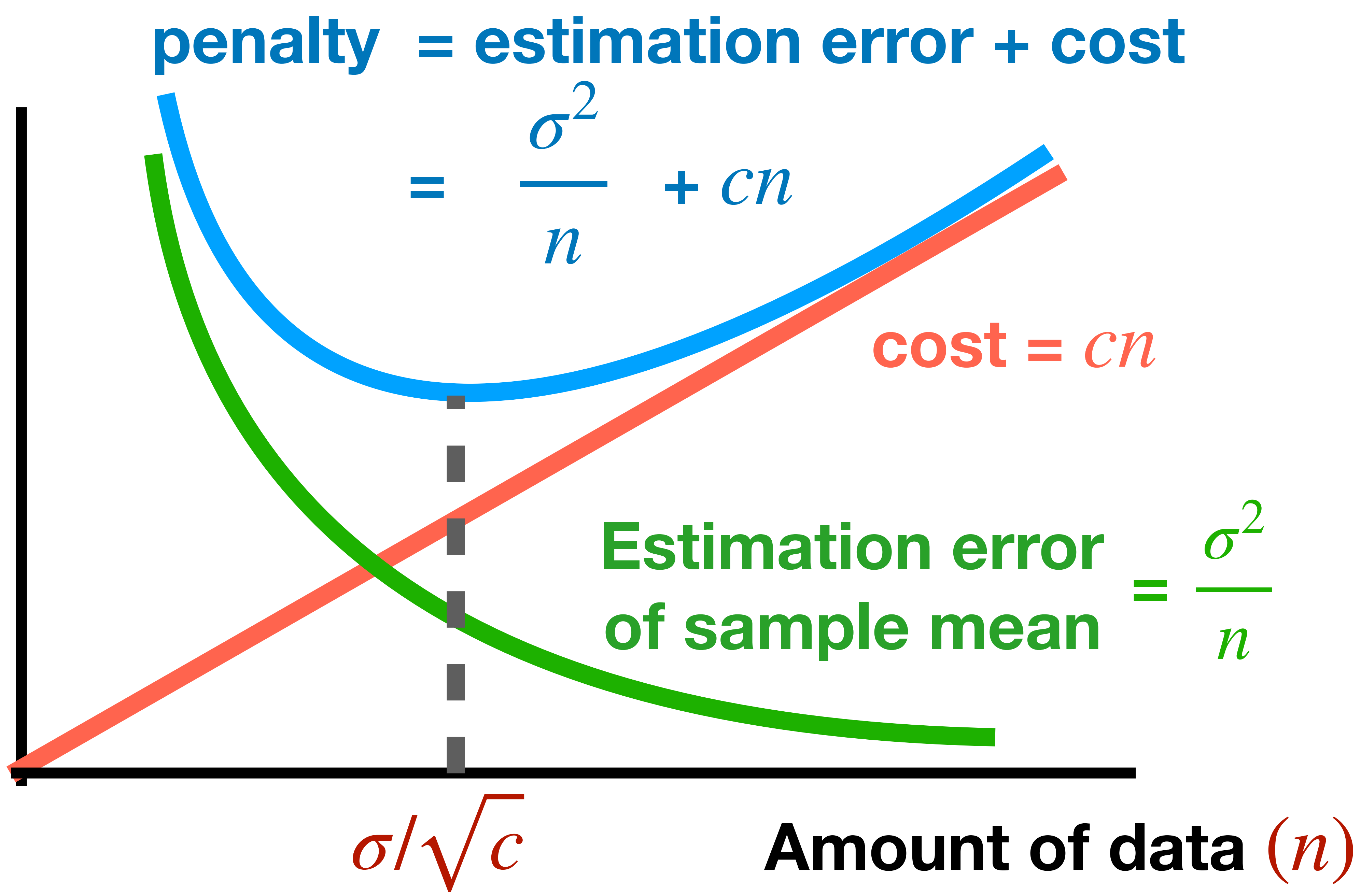
- Estimate the mean μ of a normal distribution with *known* variance σ^2.



- An agent can collect samples at *known* unit cost c.
- Each agent wishes to minimize

$$\text{penalty} = \text{estimation error} + \text{data collection cost}$$
$$= \frac{\sigma^2}{n} + cn$$

- When *working on her own*, agent will collect $\sigma/\sqrt{c}$ points to minimize penalty.



SEVERAL AGENTS WORKING TOGETHER

SEVERAL AGENTS WORKING TOGETHER

- Now consider m agents collecting and sharing their data.

SEVERAL AGENTS WORKING TOGETHER

- Now consider m agents collecting and sharing their data.

- *Social penalty* of all m agents if they collectively collect n_{tot} points.

$$\text{social penalty} = \text{ estimation error of all agents } + \text{data collection cost} = m \times \frac{\sigma^2}{n_{\text{tot}}} + c n_{\text{tot}}$$

SEVERAL AGENTS WORKING TOGETHER

- Now consider m agents collecting and sharing their data.
- *Social penalty* of all m agents if they collectively collect n_{tot} points.

$$\text{social penalty} = \text{ estimation error of all agents } + \text{data collection cost} = m \times \frac{\sigma^2}{n_{\text{tot}}} + c n_{\text{tot}}$$

- To minimize social penalty, they should collect $n^\star_{\text{tot}} = \frac{\sigma\sqrt{m}}{\sqrt{c}}$ points.

SEVERAL AGENTS WORKING TOGETHER

- Now consider m agents collecting and sharing their data.
- *Social penalty* of all m agents if they collectively collect n_{tot} points.

$$\text{social penalty} = \text{ estimation error of all agents } + \text{data collection cost} = m \times \frac{\sigma^2}{n_{\text{tot}}} + c n_{\text{tot}}$$

- To minimize social penalty, they should collect $n_{\text{tot}}^{\star} = \frac{\sigma\sqrt{m}}{\sqrt{c}}$ points.
  - Each agent needs to collect only $n^{\star} = \frac{\sigma}{\sqrt{mc}}$ points

    Only $\times\ 1/\sqrt{m}$ when compared to working on her own ($\sigma/\sqrt{c}$ points).

SEVERAL AGENTS WORKING TOGETHER

- Now consider m agents collecting and sharing their data.

- *Social penalty* of all m agents if they collectively collect n_{tot} points.

$$\text{social penalty} = \text{ estimation error of all agents } + \text{data collection cost} = m \times \frac{\sigma^2}{n_{\text{tot}}} + c n_{\text{tot}}$$

- To minimize social penalty, they should collect $n^\star_{\text{tot}} = \frac{\sigma\sqrt{m}}{\sqrt{c}}$ points.

  - Each agent needs to collect only $n^\star = \frac{\sigma}{\sqrt{mc}}$ points

    Only $\times\, 1/\sqrt{m}$ when compared to working on her own ($\sigma/\sqrt{c}$ points).

  - But she has $\times \sqrt{m}$ data.

SEVERAL AGENTS WORKING TOGETHER

	Amount of data she needs to collect (n_i)	Amount of data available to her (n_{tot})	Penalty $\frac{\sigma^2}{n_{\text{tot}}} + cn_i$
Working on her own			
Working together			

SEVERAL AGENTS WORKING TOGETHER

	Amount of data she needs to collect (n_i)	Amount of data available to her (n_{tot})	Penalty $\dfrac{\sigma^2}{n_{\text{tot}}} + cn_i$
Working on her own	$\dfrac{\sigma}{\sqrt{c}}$		
Working together	$\dfrac{\sigma}{\sqrt{cm}}$		

SEVERAL AGENTS WORKING TOGETHER

	Amount of data she needs to collect (n_i)	Amount of data available to her (n_{tot})	Penalty $\frac{\sigma^2}{n_{\text{tot}}} + cn_i$
Working on her own	$\frac{\sigma}{\sqrt{c}}$	$\frac{\sigma}{\sqrt{c}}$	
Working together	$\frac{\sigma}{\sqrt{cm}}$	$\frac{\sigma\sqrt{m}}{\sqrt{c}}$	

SEVERAL AGENTS WORKING TOGETHER

	Amount of data she needs to collect (n_i)	Amount of data available to her (n_{tot})	Penalty $\frac{\sigma^2}{n_{\text{tot}}} + cn_i$
Working on her own	$\frac{\sigma}{\sqrt{c}}$	$\frac{\sigma}{\sqrt{c}}$	$2\sigma\sqrt{c}$
Working together	$\frac{\sigma}{\sqrt{cm}}$	$\frac{\sigma\sqrt{m}}{\sqrt{c}}$	$\frac{2\sigma\sqrt{c}}{\sqrt{m}}$

SEVERAL AGENTS WORKING TOGETHER

	Amount of data she needs to collect (n_i)	Amount of data available to her (n_{tot})	Penalty $\frac{\sigma^2}{n_{\text{tot}}} + cn_i$
Working on her own	$\frac{\sigma}{\sqrt{c}}$	$\frac{\sigma}{\sqrt{c}}$	$2\sigma\sqrt{c}$
Working together	$\frac{\sigma}{\sqrt{cm}}$	$\frac{\sigma\sqrt{m}}{\sqrt{c}}$	$\frac{2\sigma\sqrt{c}}{\sqrt{m}}$

Agents can reduce data collection costs, and improve estimation error by sharing data with others.

DESIGNING A MECHANISM FOR COLLABORATIVE NORMAL MEAN ESTIMATION

DESIGNING A MECHANISM FOR COLLABORATIVE NORMAL MEAN ESTIMATION

- Naive mechanism 1: “pool and share”

DESIGNING A MECHANISM FOR COLLABORATIVE NORMAL MEAN ESTIMATION

- Naive mechanism 1: "pool and share"
  - Selfish agents will *free-ride:* not contributing any data herself, but using data that the others have contributed.

- Naive mechanism 1: "pool and share"
  - Selfish agents will *free-ride:* not contributing any data herself, but using data that the others have contributed.

$$\text{penalty when free-riding} \approx \frac{1}{2} \times \text{penalty if well-behaved}$$

DESIGNING A MECHANISM FOR COLLABORATIVE NORMAL MEAN ESTIMATION

- Naive mechanism 1: “pool and share”
  - Selfish agents will *free-ride:* not contributing any data herself, but using data that the others have contributed.

$$\text{penalty when free-riding} \approx \frac{1}{2} \times \text{penalty if well-behaved}$$

- Naive mechanism 2: “pool and share, but only if you contribute enough data”

DESIGNING A MECHANISM FOR COLLABORATIVE NORMAL MEAN ESTIMATION

- Naive mechanism 1: "pool and share"
  - Selfish agents will *free-ride:* not contributing any data herself, but using data that the others have contributed.

$$\text{penalty when free-riding} \approx \frac{1}{2} \times \text{penalty if well-behaved}$$

- Naive mechanism 2: "pool and share, but only if you contribute enough data"
  - Agents can fabricate and then discard after receiving others' data.

A MECHANISM FOR COLLABORATIVE NORMAL MEAN ESTIMATION

A MECHANISM FOR COLLABORATIVE NORMAL MEAN ESTIMATION

Each agent i will

A MECHANISM FOR COLLABORATIVE NORMAL MEAN ESTIMATION

Each agent i will

- Collect n_i points $X_i = \{x_{i,1}, \ldots, x_{i,n_i}\}$ and submit $Y_i = \{y_{i,1}, \ldots, y_{i,n_i'}\} = f_i(X_i)$.

Each agent i will

- Collect n_i points $X_i = \{x_{i,1}, \ldots, x_{i,n_i}\}$ and submit $Y_i = \{y_{i,1}, \ldots, y_{i,n_i'}\} = f_i(X_i)$.

Mechanism

A MECHANISM FOR COLLABORATIVE NORMAL MEAN ESTIMATION

Each agent i will

- Collect n_i points $X_i = \{x_{i,1}, \ldots, x_{i,n_i}\}$ and submit $Y_i = \{y_{i,1}, \ldots, y_{i,n_i'}\} = f_i(X_i)$.

Mechanism

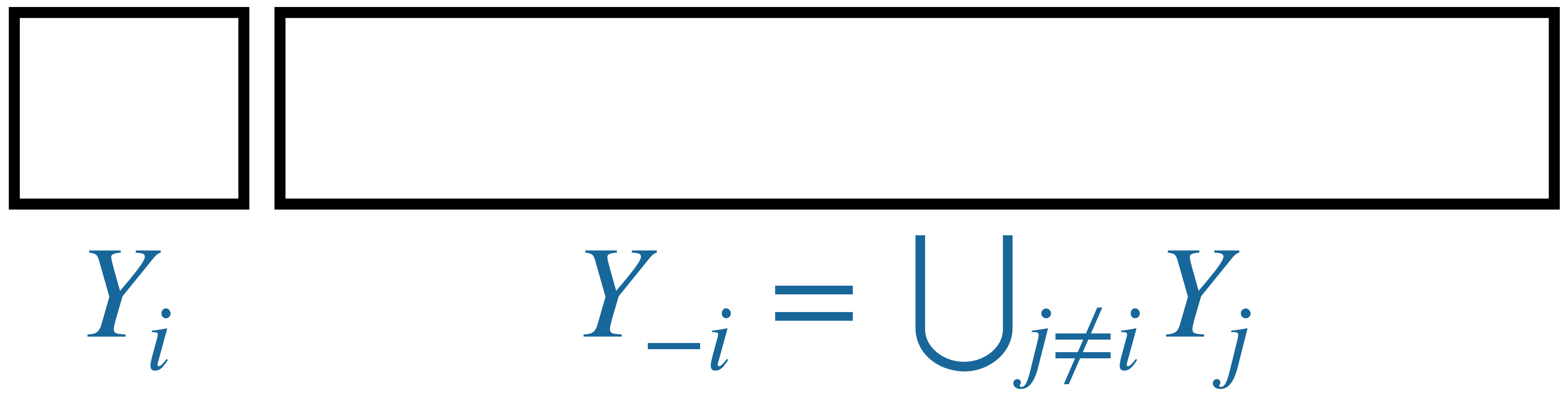
- For each agent i:

Each agent i will

- Collect n_i points $X_i = \{x_{i,1}, \ldots, x_{i,n_i}\}$ and submit $Y_i = \{y_{i,1}, \ldots, y_{i,n_i'}\} = f_i(X_i)$.

Mechanism

- For each agent i:



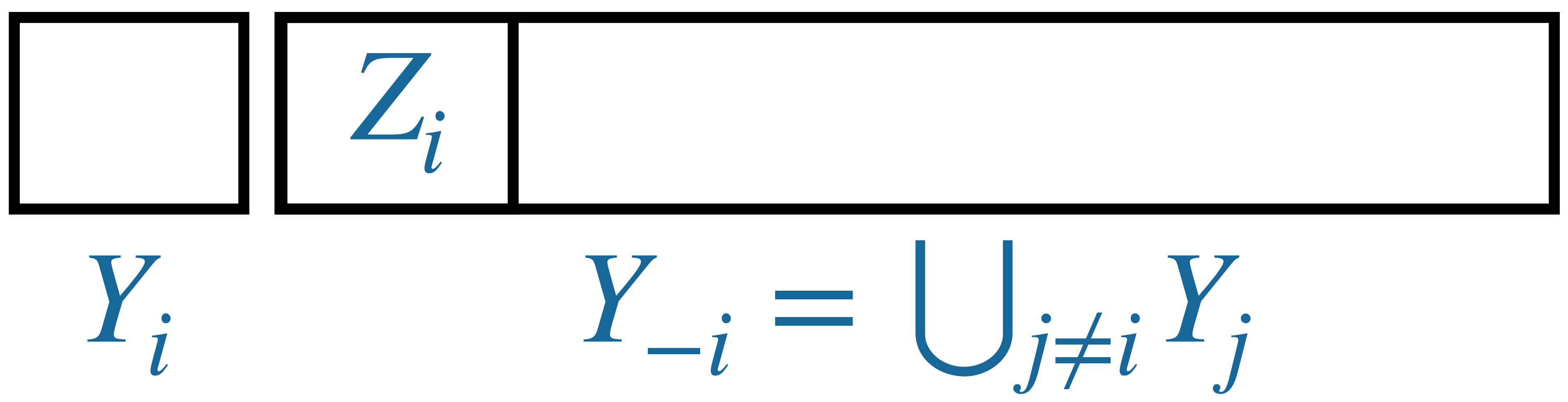
A MECHANISM FOR COLLABORATIVE NORMAL MEAN ESTIMATION

Each agent i will

- Collect n_i points $X_i = \{x_{i,1}, \ldots, x_{i,n_i}\}$ and submit $Y_i = \{y_{i,1}, \ldots, y_{i,n_i'}\} = f_i(X_i)$.

Mechanism

- For each agent i:
  - $Z_i \leftarrow$ randomly sample $\sigma/\sqrt{cm}$ points from others' submissions Y_{-i}.



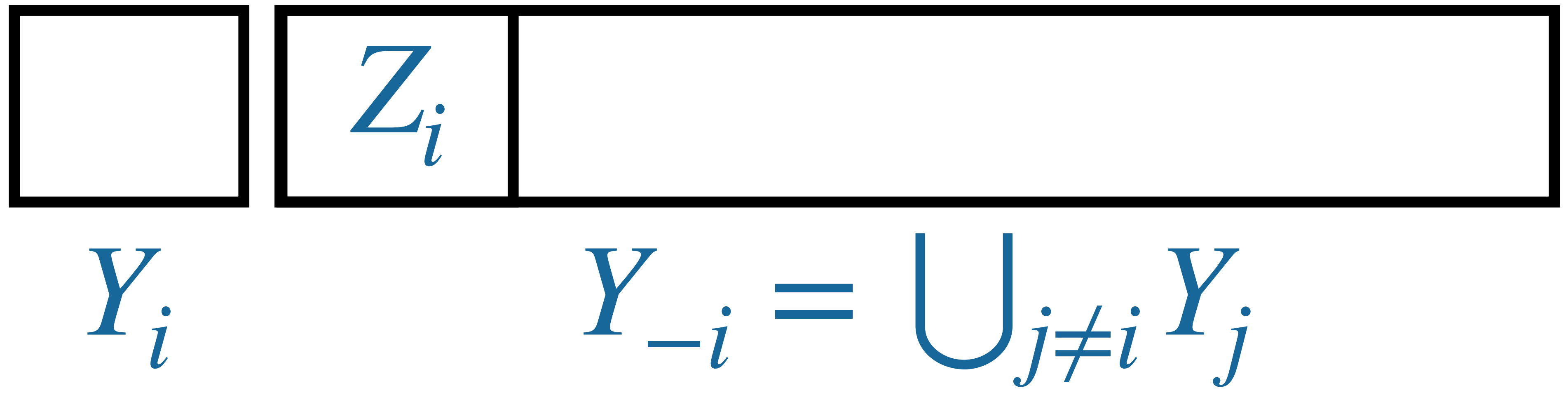
A MECHANISM FOR COLLABORATIVE NORMAL MEAN ESTIMATION

Each agent i will

- Collect n_i points $X_i = \{x_{i,1}, \ldots, x_{i,n_i}\}$ and submit $Y_i = \{y_{i,1}, \ldots, y_{i,n_i'}\} = f_i(X_i)$.

Mechanism

- For each agent i:
  - $Z_i \leftarrow$ randomly sample $\sigma/\sqrt{cm}$ points from others' submissions Y_{-i}.
  - Set noise variance $\eta_i^2 \propto \left(\text{mean}(Y_i) - \text{mean}(Z_i)\right)^2$ # Variance proportional to difference



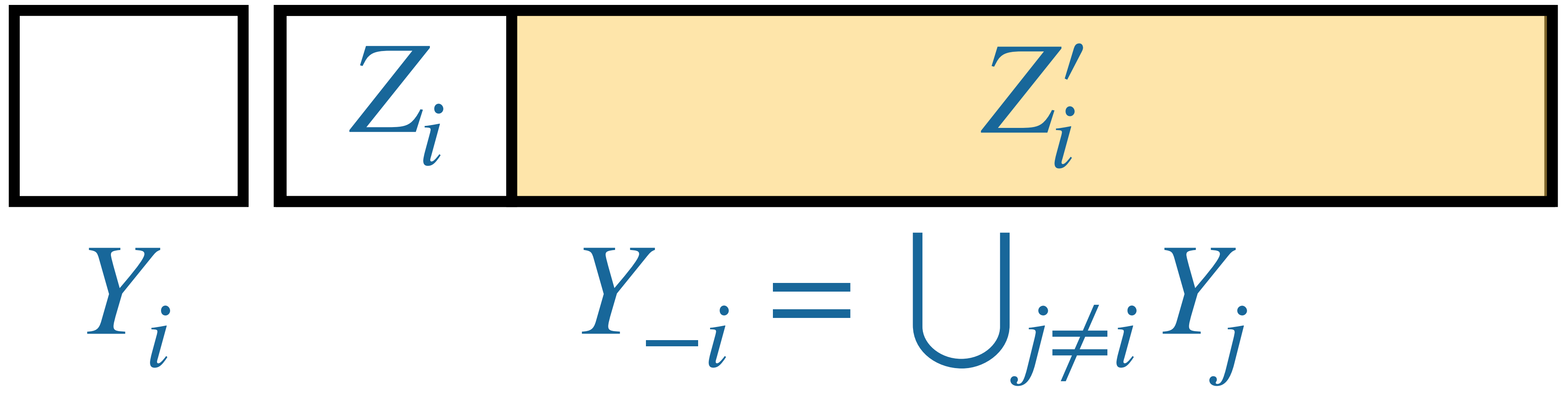
A MECHANISM FOR COLLABORATIVE NORMAL MEAN ESTIMATION

Each agent i will

- Collect n_i points $X_i = \{x_{i,1}, \ldots, x_{i,n_i}\}$ and submit $Y_i = \{y_{i,1}, \ldots, y_{i,n_i'}\} = f_i(X_i)$.

Mechanism

- For each agent i:
  - $Z_i \leftarrow$ randomly sample $\sigma/\sqrt{cm}$ points from others' submissions Y_{-i}.
  - Set noise variance $\eta_i^2 \propto \left(\text{mean}(Y_i) - \text{mean}(Z_i)\right)^2$ # Variance proportional to difference
  - $Z_i' \leftarrow \left\{ z + \epsilon_z, \quad \text{for all } z \in Y_{-i} \backslash Z_i, \quad \text{where } \epsilon_z \sim \mathcal{N}(0, \eta_i^2) \right\}$.



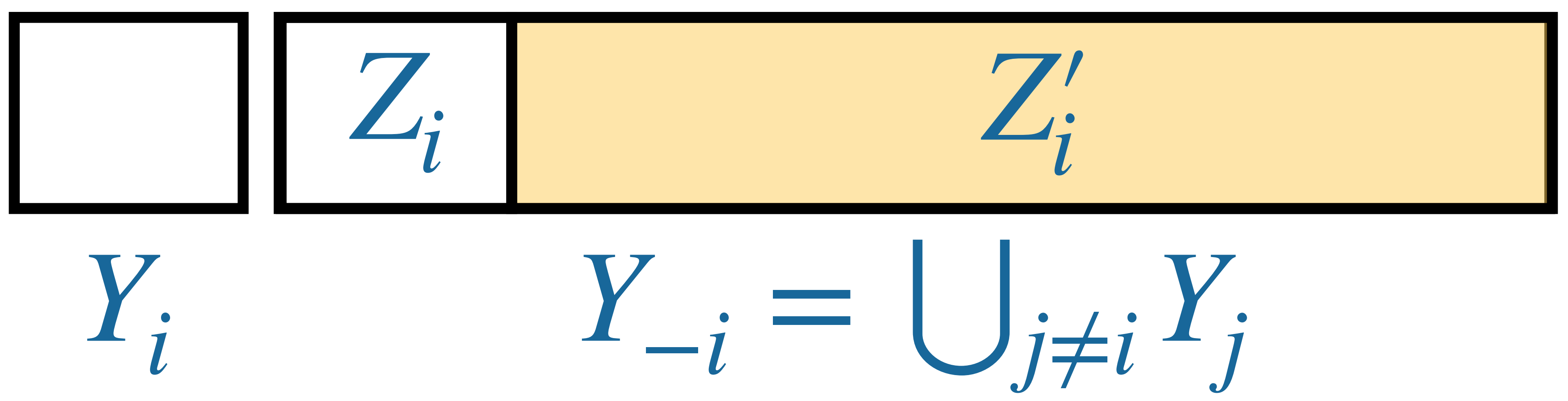
A MECHANISM FOR COLLABORATIVE NORMAL MEAN ESTIMATION

Each agent i will

- Collect n_i points $X_i = \{x_{i,1}, \ldots, x_{i,n_i}\}$ and submit $Y_i = \{y_{i,1}, \ldots, y_{i,n_i'}\} = f_i(X_i)$.

Mechanism

- For each agent i:
  - $Z_i \leftarrow$ randomly sample $\sigma/\sqrt{cm}$ points from others' submissions Y_{-i}.
  - Set noise variance $\eta_i^2 \propto \left(\text{mean}(Y_i) - \text{mean}(Z_i)\right)^2$ # Variance proportional to difference
  - $Z_i' \leftarrow \left\{ z + \epsilon_z, \quad \text{for all } z \in Y_{-i} \backslash Z_i, \quad \text{where } \epsilon_z \sim \mathcal{N}(0, \eta_i^2) \right\}$.
  - Return $A_i \leftarrow (Z_i, Z_i', \eta_i^2)$ to each agent.



A MECHANISM FOR COLLABORATIVE NORMAL MEAN ESTIMATION

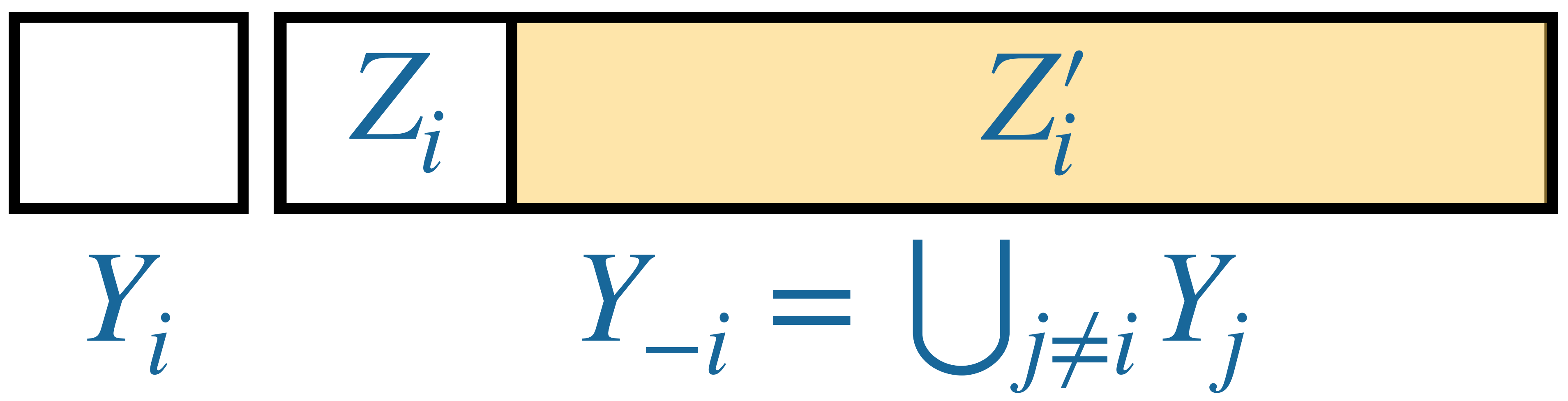
Each agent i will

- Collect n_i points $X_i = \{x_{i,1}, \ldots, x_{i,n_i}\}$ and submit $Y_i = \{y_{i,1}, \ldots, y_{i,n_i'}\} = f_i(X_i)$.

Mechanism

- For each agent i:
  - $Z_i \leftarrow$ randomly sample $\sigma/\sqrt{cm}$ points from others' submissions Y_{-i}.
  - Set noise variance $\eta_i^2 \propto \left(\text{mean}(Y_i) - \text{mean}(Z_i)\right)^2$ # Variance proportional to difference
  - $Z_i' \leftarrow \left\{ z + \epsilon_z, \quad \text{for all } z \in Y_{-i} \backslash Z_i, \quad \text{where } \epsilon_z \sim \mathcal{N}(0, \eta_i^2) \right\}$.
  - Return $A_i \leftarrow (Z_i, Z_i', \eta_i^2)$ to each agent.

Each agent i will



A MECHANISM FOR COLLABORATIVE NORMAL MEAN ESTIMATION

Each agent i will

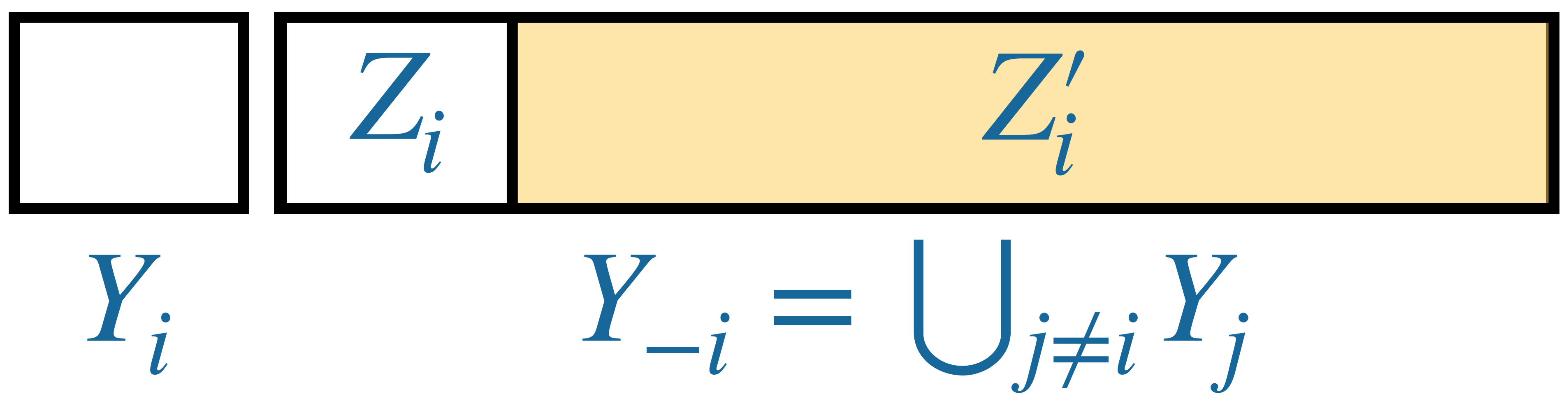
- Collect n_i points $X_i = \{x_{i,1}, \ldots, x_{i,n_i}\}$ and submit $Y_i = \{y_{i,1}, \ldots, y_{i,n_i'}\} = f_i(X_i)$.

Mechanism

- For each agent i:
  - $Z_i \leftarrow$ randomly sample $\sigma/\sqrt{cm}$ points from others' submissions Y_{-i}.
  - Set noise variance $\eta_i^2 \propto \left(\text{mean}(Y_i) - \text{mean}(Z_i)\right)^2$ # Variance proportional to difference
  - $Z_i' \leftarrow \left\{ z + \epsilon_z, \quad \text{for all } z \in Y_{-i} \backslash Z_i, \quad \text{where } \epsilon_z \sim \mathcal{N}(0, \eta_i^2) \right\}$.
  - Return $A_i \leftarrow (Z_i, Z_i', \eta_i^2)$ to each agent.

Each agent i will

- Compute her estimate $h_i(X_i, Y_i, A_i)$



Mechanisms recommends that agents follow $s_i^\star = (n_i^\star, f_i^\star, h_i^\star)$,

$$n_i^\star = \frac{\sigma}{\sqrt{cm}},$$

$$f_i^\star = \text{identity},$$

$$h_i^\star\Big(X_i, Y_i, \underbrace{(Z_i, Z_i', \eta_i^2)}_{A_i}\Big) = \frac{\sum_{u \in X_i \cup Z_i} u + \frac{1}{1+\eta_i^2/\sigma^2} \sum_{u \in Z_i'} u}{|X_i \cup Z_i| + \frac{1}{1+\eta_i^2/\sigma^2} |Z_i'|}$$

Mechanisms recommends that agents follow $s_i^\star = (n_i^\star, f_i^\star, h_i^\star)$,

$$n_i^\star = \frac{\sigma}{\sqrt{cm}},$$

$$f_i^\star = \text{identity},$$

$$h_i^\star\Big(X_i, Y_i, \underbrace{(Z_i, Z_i', \eta_i^2)}_{A_i}\Big) = \frac{\sum_{u \in X_i \cup Z_i} u + \frac{1}{1 + \eta_i^2/\sigma^2} \sum_{u \in Z_i'} u}{|X_i \cup Z_i| + \frac{1}{1 + \eta_i^2/\sigma^2} |Z_i'|}$$

That is collect a sufficient amount of data $n_i^\star$, submit it truthfully $f_i^\star$, and use a weighted average estimator $h_i^\star$.

Mechanisms recommends that agents follow $s_i^\star = (n_i^\star, f_i^\star, h_i^\star)$,

$$n_i^\star = \frac{\sigma}{\sqrt{cm}},$$

$$f_i^\star = \text{identity},$$

$$h_i^\star\Big(X_i, Y_i, \underbrace{(Z_i, Z_i', \eta_i^2)}_{A_i}\Big) = \frac{\sum_{u \in X_i \cup Z_i} u + \frac{1}{1+\eta_i^2/\sigma^2} \sum_{u \in Z_i'} u}{|X_i \cup Z_i| + \frac{1}{1+\eta_i^2/\sigma^2} |Z_i'|}$$

That is collect a sufficient amount of data $n_i^\star$, submit it truthfully $f_i^\star$, and use a weighted average estimator $h_i^\star$.

- $h^\star$ is minimax-optimal for the corrupted dataset.

THEOREM

This mechanism is

THEOREM

This mechanism is

- **Nash incentive-compatible:** Provided that other agents are well-behaved, the best strategy for an agent is to,

This mechanism is

- **Nash incentive-compatible:** Provided that other agents are well-behaved, the best strategy for an agent is to,
    - Collect a sufficient amount ($n^\star = \sigma/\sqrt{mc}$) of data.

THEOREM

This mechanism is

- **Nash incentive-compatible:** Provided that other agents are well-behaved, the best strategy for an agent is to,
    - Collect a sufficient amount ($n^\star = \sigma/\sqrt{mc}$) of data.
    - Submit it truthfully.

This mechanism is

- **Nash incentive-compatible:** Provided that other agents are well-behaved, the best strategy for an agent is to,
    - Collect a sufficient amount ($n^\star = \sigma/\sqrt{mc}$) of data.
    - Submit it truthfully.
    - Use the recommended minimax-optimal estimator.

THEOREM

This mechanism is

- **Nash incentive-compatible:** Provided that other agents are well-behaved, the best strategy for an agent is to,
  - Collect a sufficient amount ($n^\star = \sigma/\sqrt{mc}$) of data.
  - Submit it truthfully.
  - Use the recommended minimax-optimal estimator.
- **Individually rational:** Provided that others are well-behaved, an agent does not do worse than the best she could do on her own.

THEOREM

This mechanism is

- **Nash incentive-compatible:** Provided that other agents are well-behaved, the best strategy for an agent is to,
  - Collect a sufficient amount ($n^\star = \sigma/\sqrt{mc}$) of data.
  - Submit it truthfully.
  - Use the recommended minimax-optimal estimator.
- **Individually rational:** Provided that others are well-behaved, an agent does not do worse than the best she could do on her own.
- **Approximately efficient:** Social penalty at the Nash equilibrium strategies is at most a factor 2 of the global minimum.

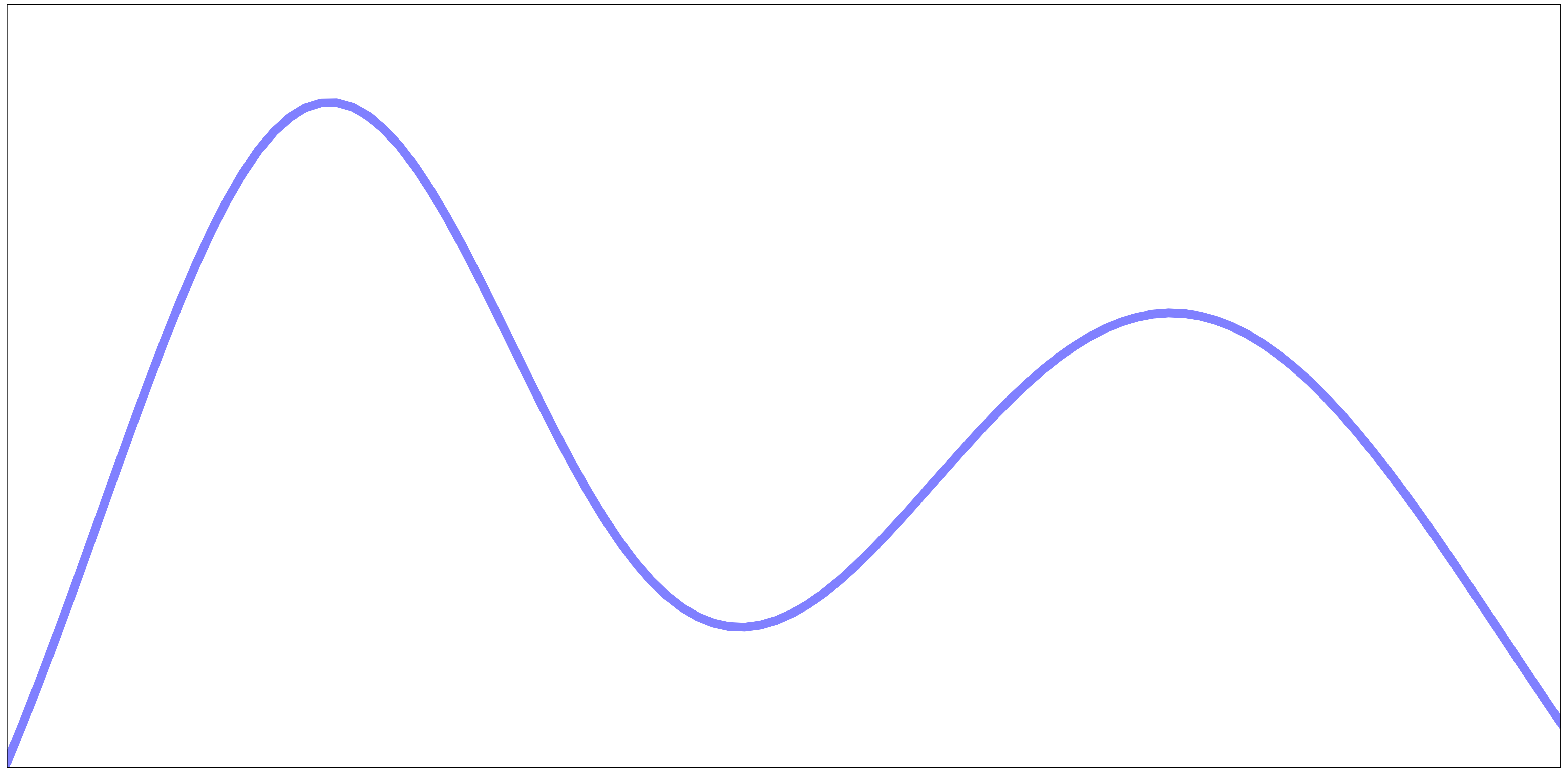
OUTLINE

1. **Mechanism design for collaborative normal mean estimation**
   (Y. Chen, Zhu, Kandasamy, *NeurIPS 2023*)

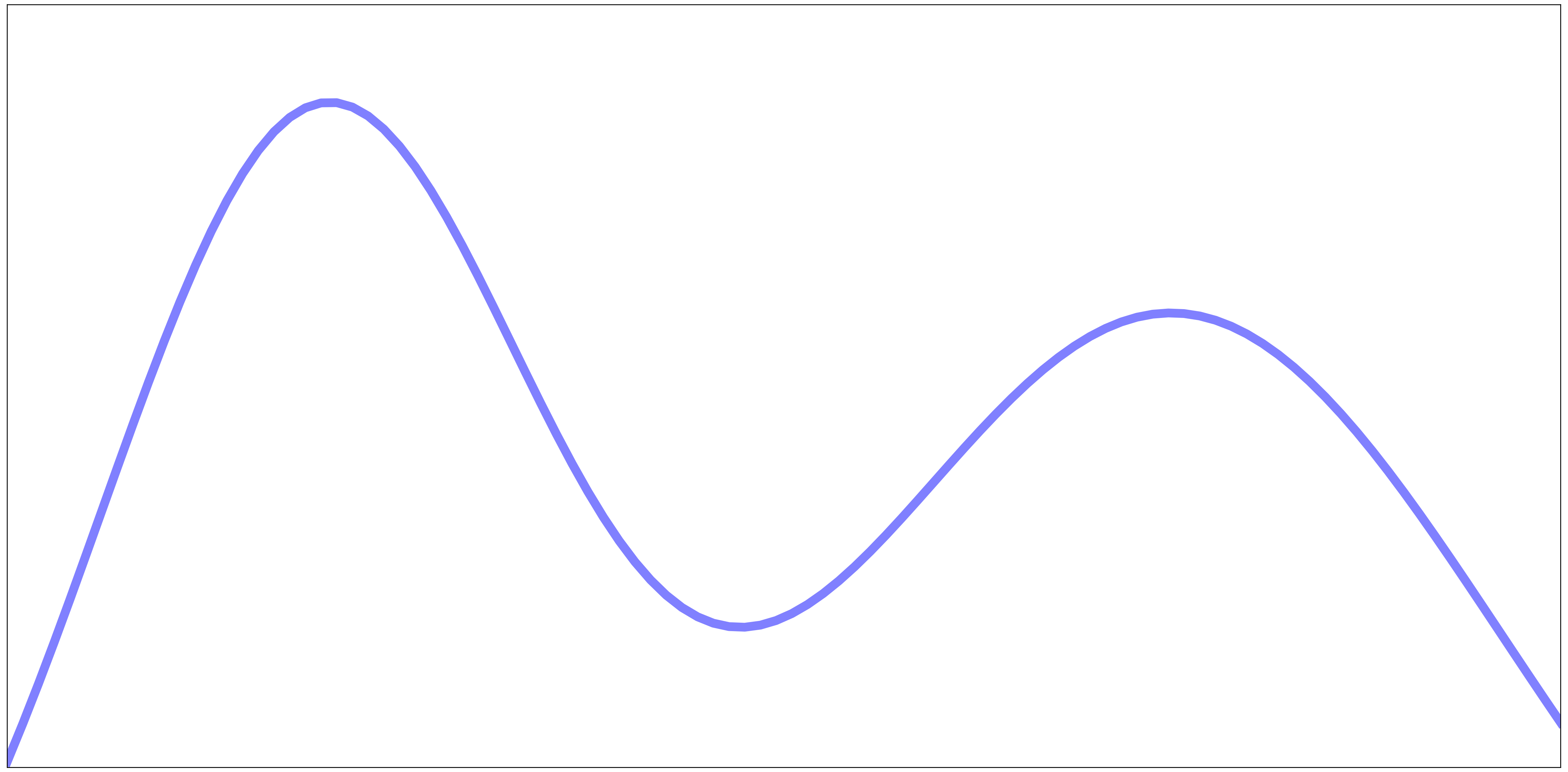
2. **High-dimensional mean estimation with varied collection costs**
   (Clinton, Y. Chen, Zhu, Kandasamy, *Ongoing work*)

3. **Learning to price data in data marketplaces**
   (K. Chen, Huh, Kandasamy, *Ongoing work*)

VARIED DATA COLLECTION COSTS

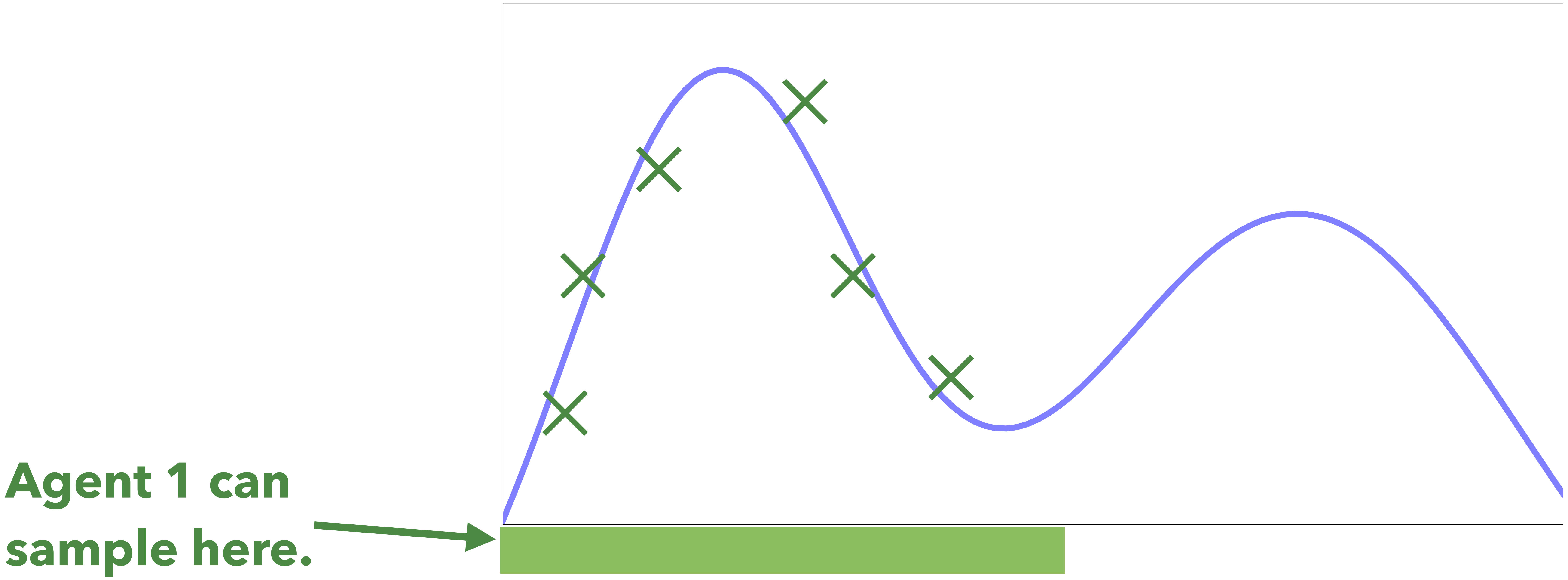


VARIED DATA COLLECTION COSTS



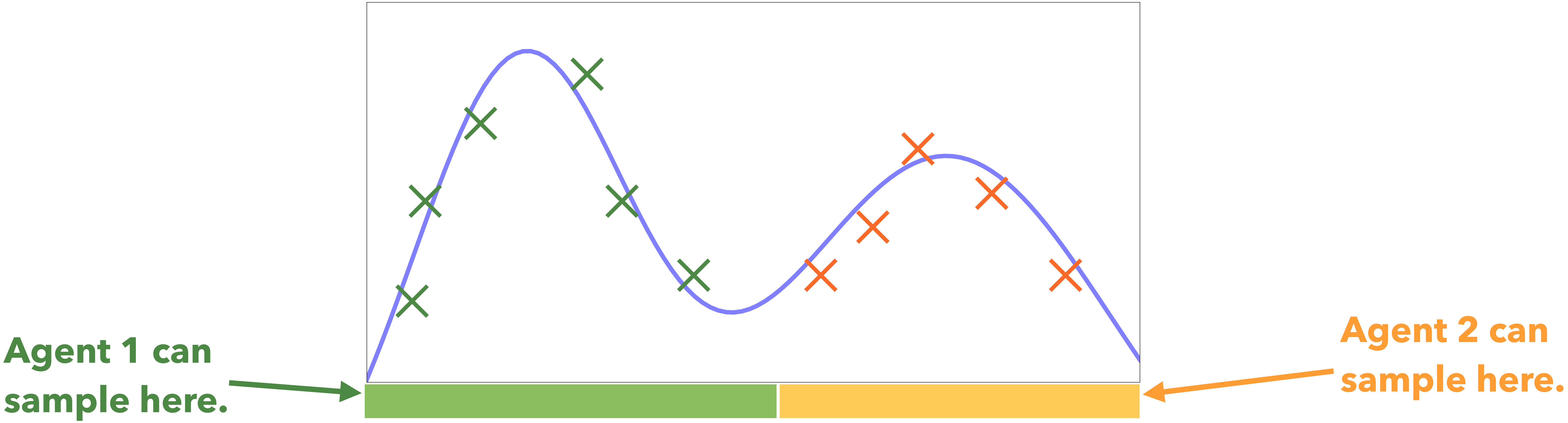
Data sharing when there is asymmetric data collection capabilities.

VARIED DATA COLLECTION COSTS



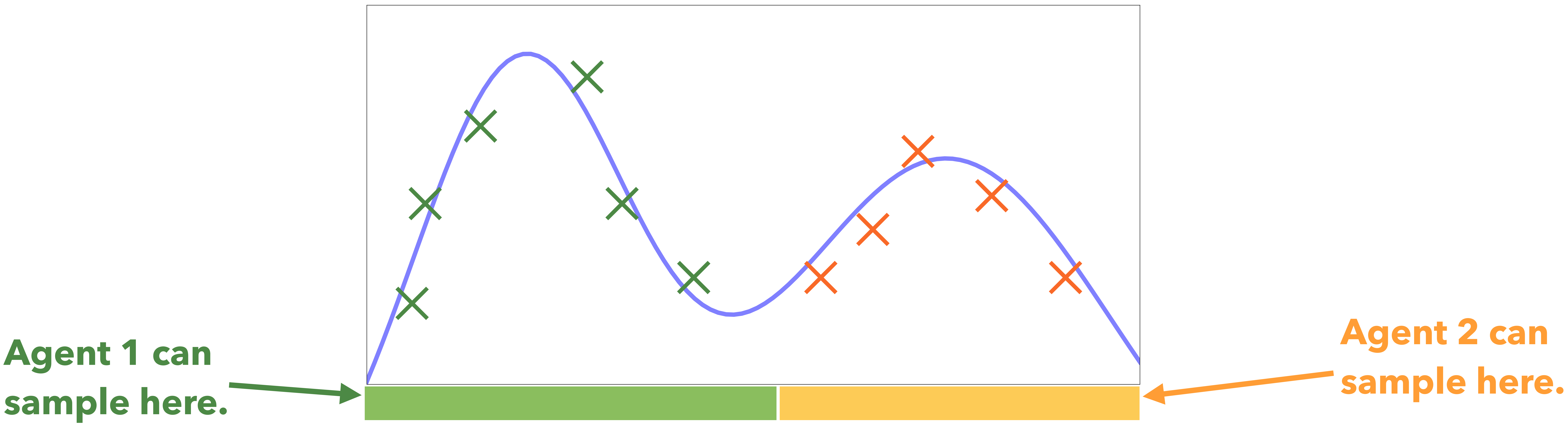
Data sharing when there is asymmetric data collection capabilities.

VARIED DATA COLLECTION COSTS



Data sharing when there is asymmetric data collection capabilities.

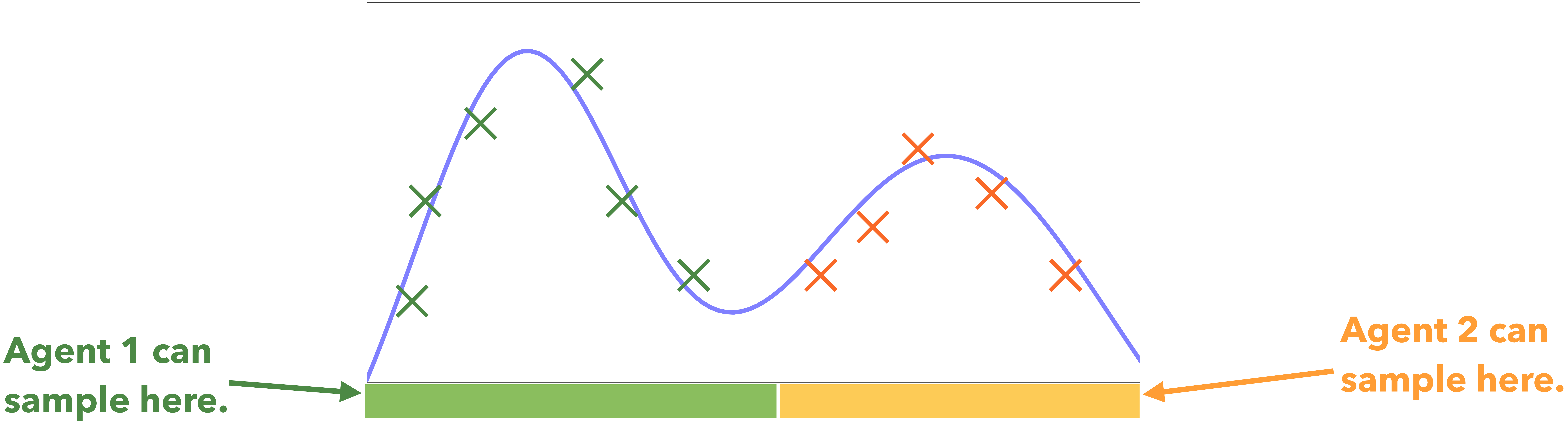
VARIED DATA COLLECTION COSTS



Data sharing when there is asymmetric data collection capabilities.

E.g: hospitals in different locations, researchers with different experimental equipment etc.

VARIED DATA COLLECTION COSTS

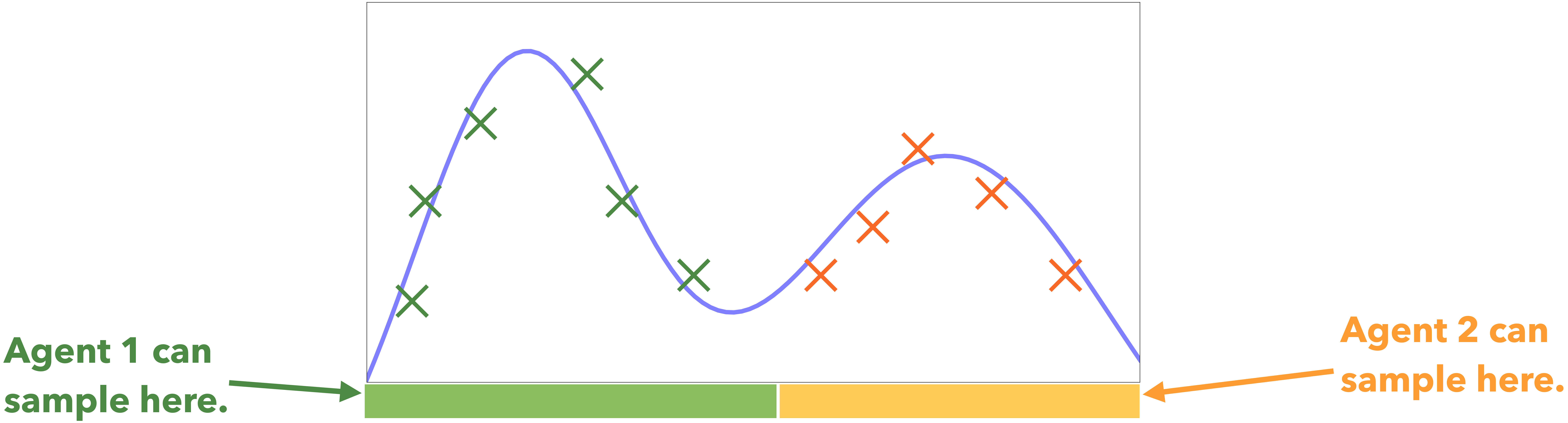


Data sharing when there is asymmetric data collection capabilities.

E.g: hospitals in different locations, researchers with different experimental equipment etc.

+ Agents will be more willing to collaborate due to complementarity of data.

VARIED DATA COLLECTION COSTS



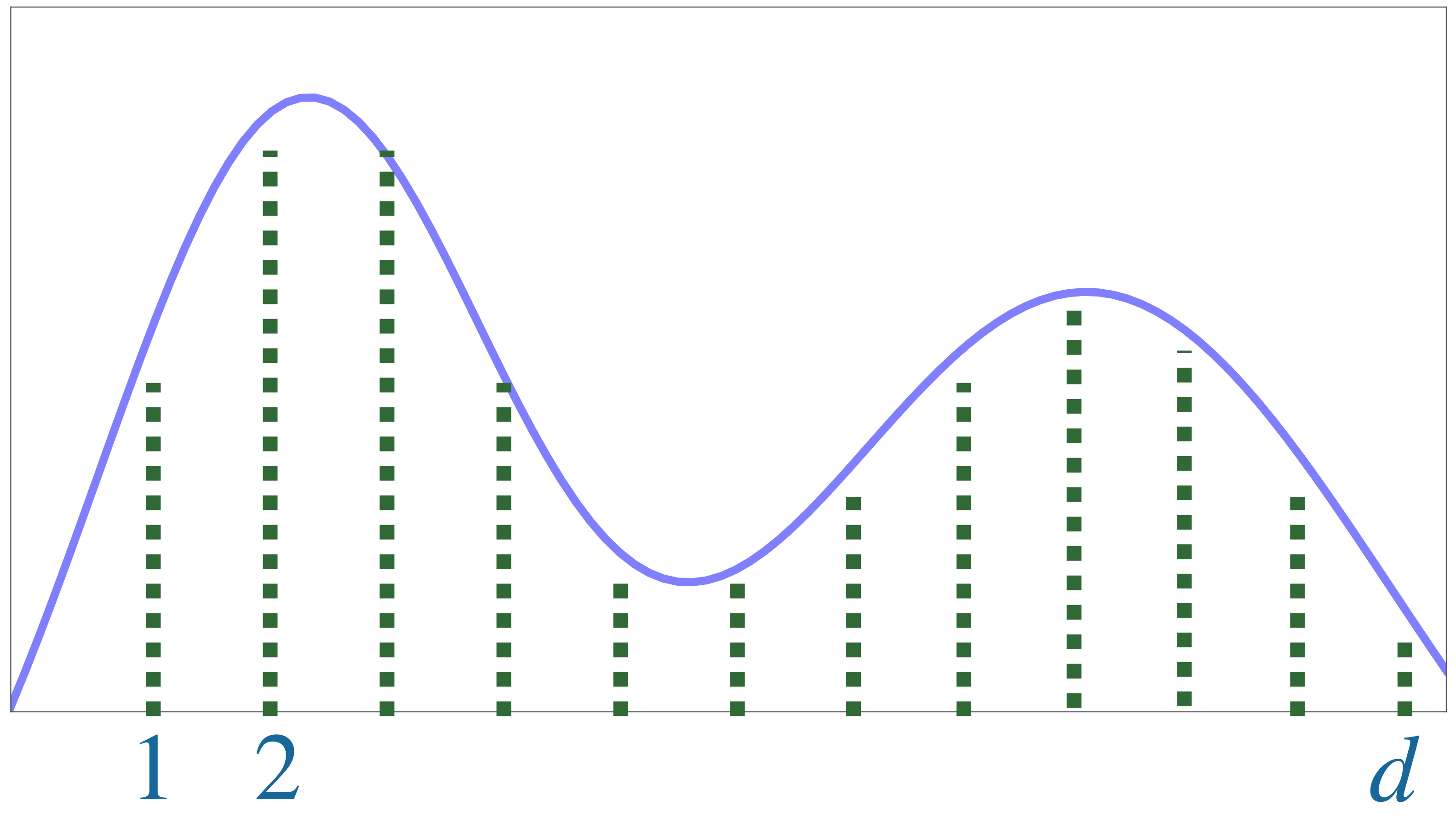
Data sharing when there is asymmetric data collection capabilities.

E.g: hospitals in different locations, researchers with different experimental equipment etc.

+ Agents will be more willing to collaborate due to complementarity of data.

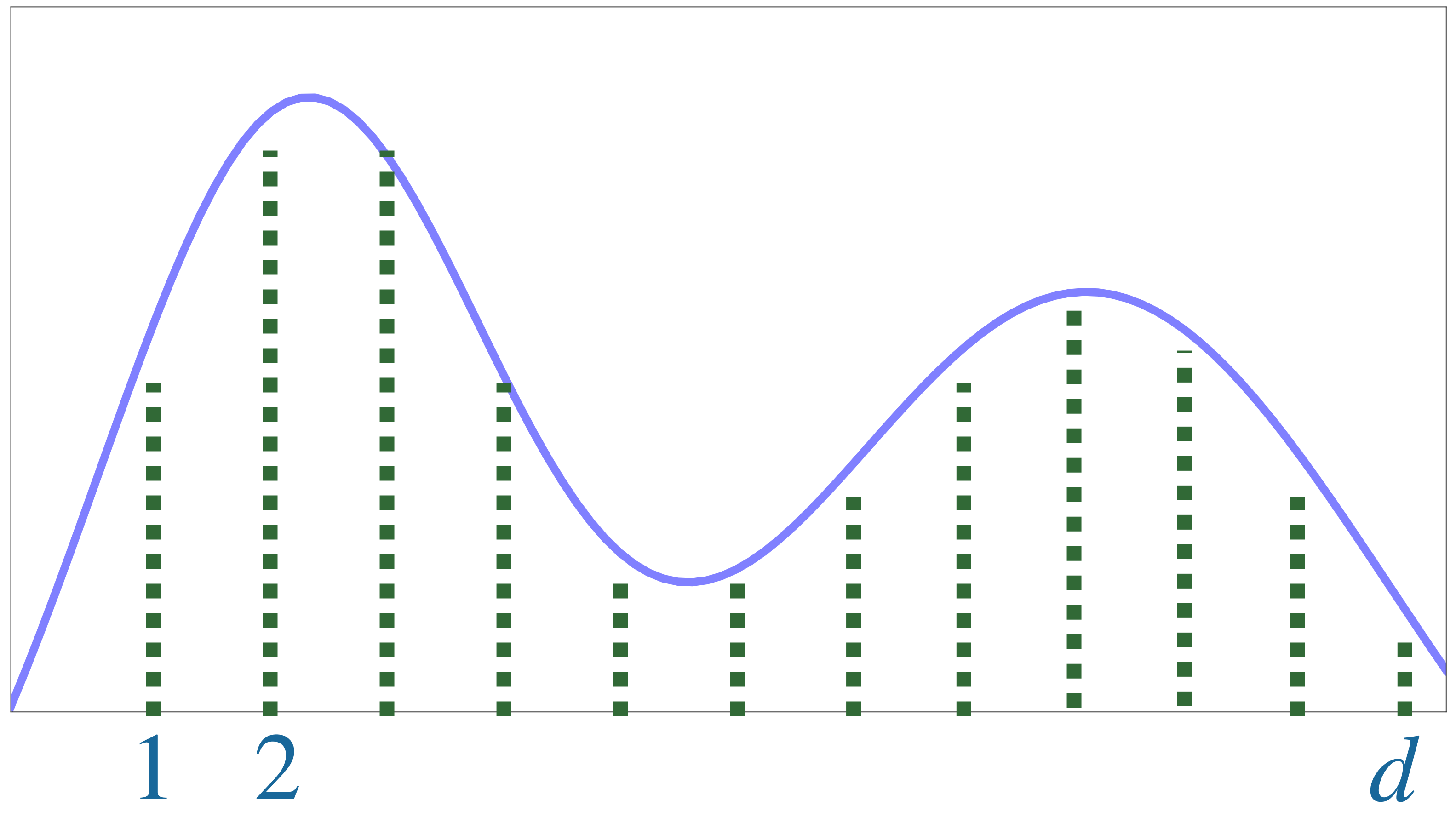
– No way to validate an agent's data with other similar data.

Consider estimating d distributions (e.g discretizing the domain)



COLLABORATIVELY LEARNING MULTIPLE DISTRIBUTIONS

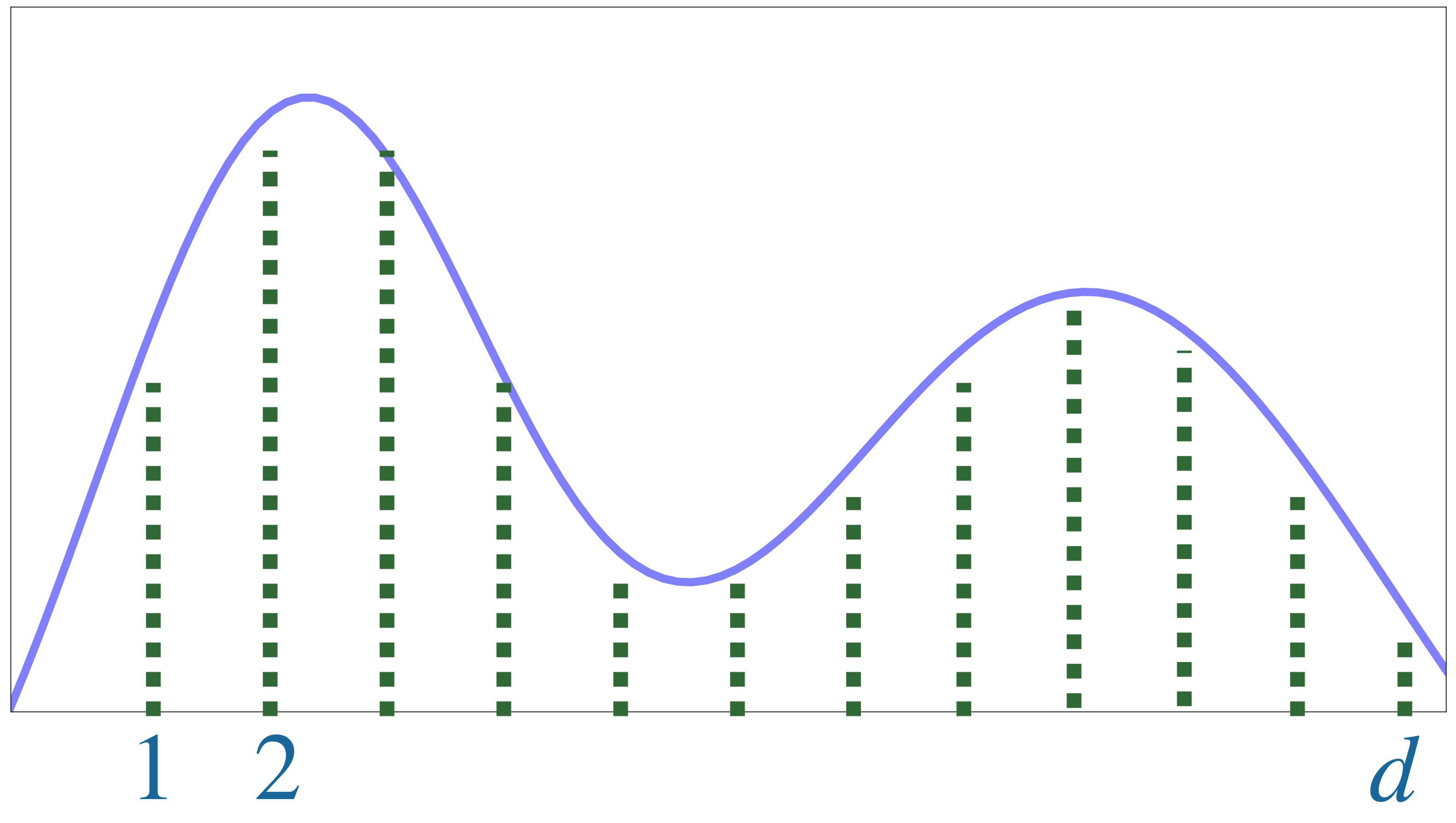
Consider estimating d distributions (e.g discretizing the domain)



Agent i can sample from distribution k at cost $c_{i,k}$.

Penalty, $p_i = \sum_{k=1}^{d} \text{est-err}_k + \sum_{k=1}^{d} c_{i,k} n_{i,k}$

Consider estimating d distributions (e.g discretizing the domain)



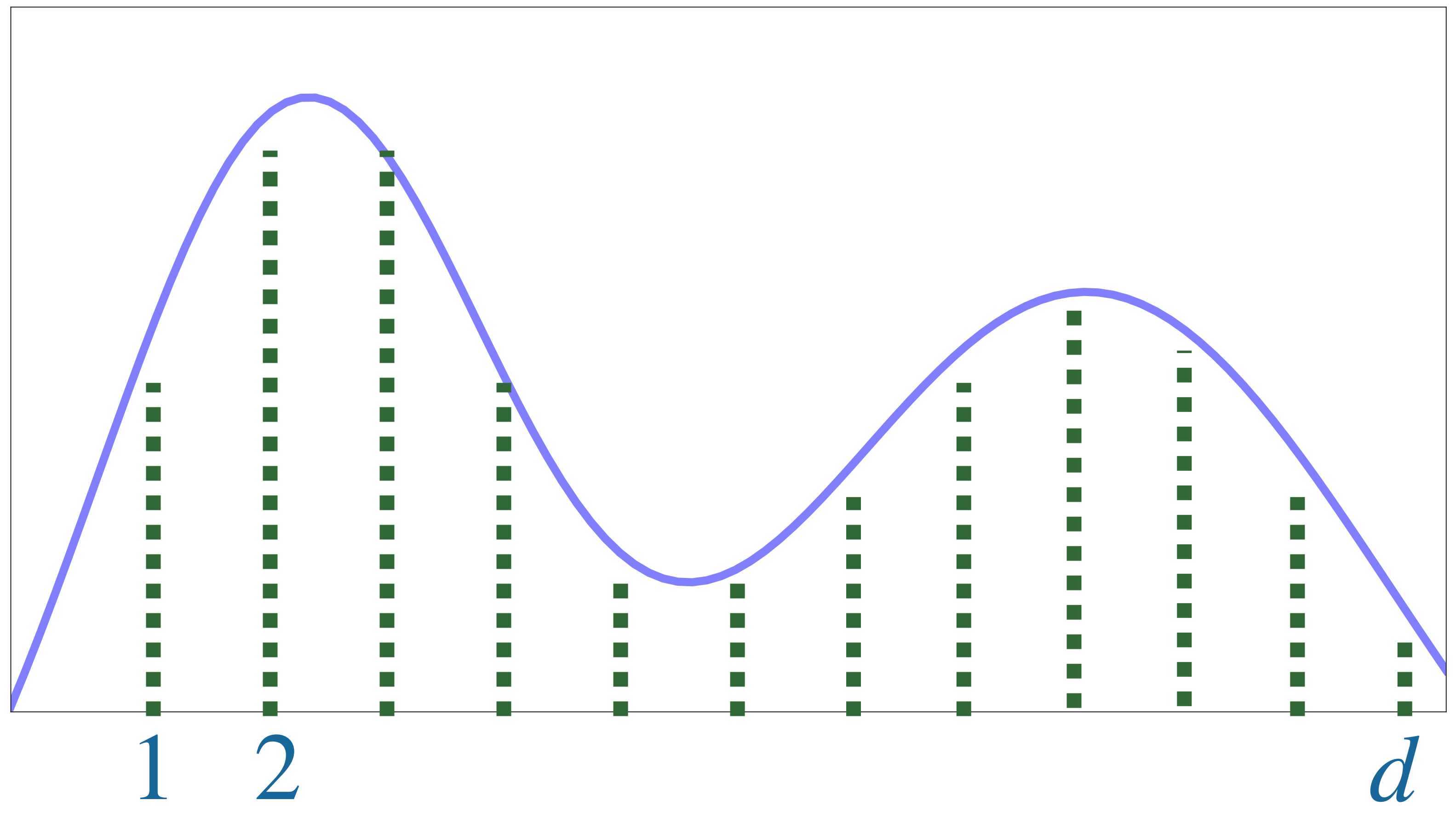
Agent i can sample from distribution k at cost $c_{i,k}$.

Penalty, $p_i = \sum_{k=1}^{d} \text{est-err}_k + \sum_{k=1}^{d} c_{i,k} n_{i,k}$

Overview of our solution:

- Uses axiomatic bargaining to define *collaboration baselines* assuming agents will always report truthfully.

Consider estimating d distributions (e.g discretizing the domain)



Agent i can sample from distribution k at cost $c_{i,k}$.

Penalty, $p_i = \sum_{k=1}^{d} \text{est-err}_k + \sum_{k=1}^{d} c_{i,k} n_{i,k}$

Overview of our solution:

- Uses axiomatic bargaining to define *collaboration baselines* assuming agents will always report truthfully.
- Enforces truthful behaviour, via corruption and other techniques.

RESULTS

Theorem: There exists a NIC and IR mechanism for which,

$$P(M, s^\star) \leq 8\sqrt{m} \cdot \inf_{M,s} P(M, s)$$

m: number of agents

RESULTS

Theorem: There exists a NIC and IR mechanism for which,

$$P(M, s^\star) \le 8\sqrt{m} \cdot \inf_{M,s} P(M, s)$$

m: number of agents

Theorem (hardness): There exists a set of costs $\{c_{i,k}\}_{i,k}$ such that for any mechanism M and any Nash equilibrium $s^\star$ of this mechanism, we have

$$P(M, s^\star) \ge \Omega\left(\sqrt{m}\right) \cdot \inf_{M,s} P(M, s)$$

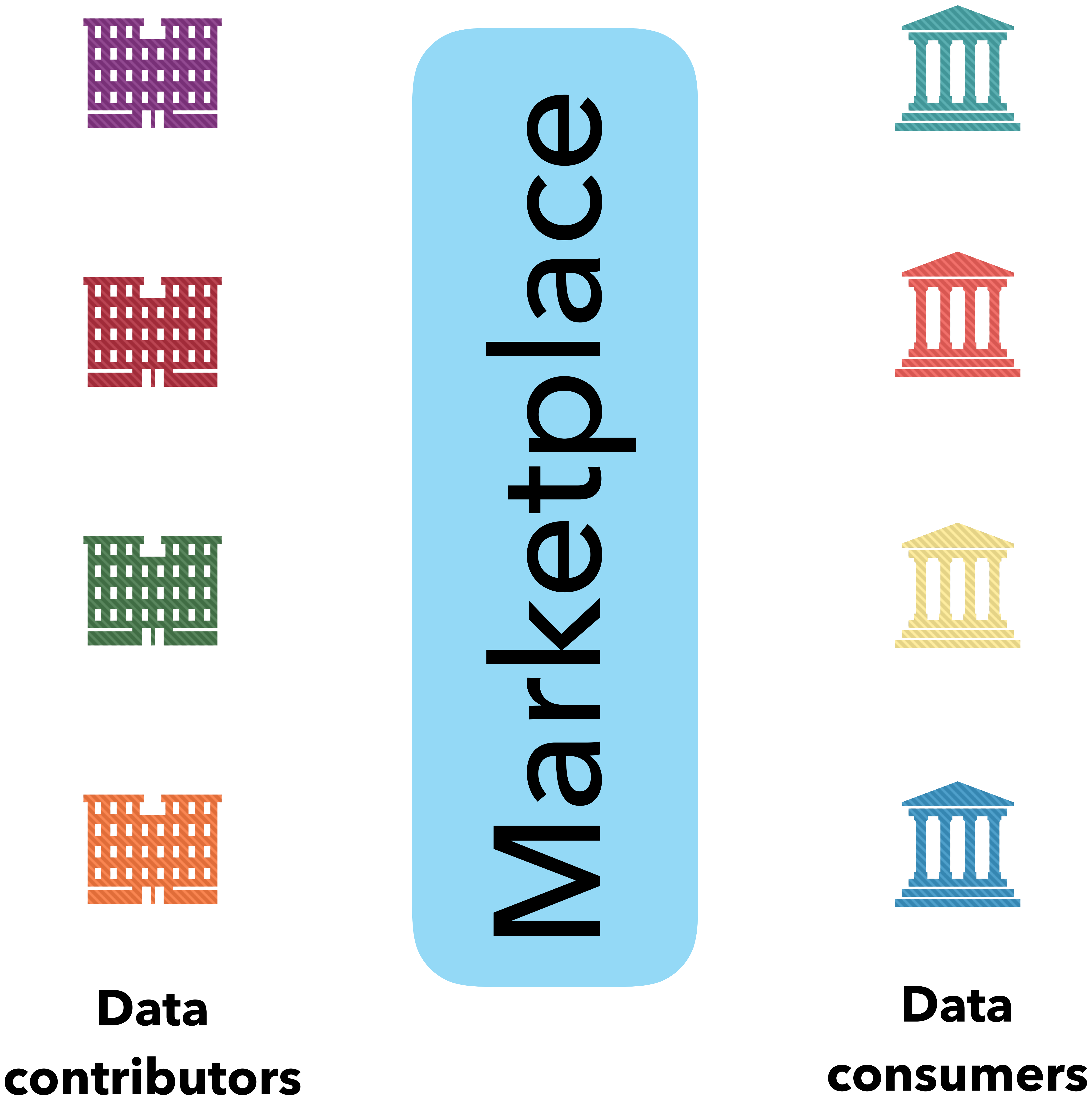
OUTLINE

1. **Mechanism design for collaborative normal mean estimation**
   (Y. Chen, Zhu, Kandasamy, *NeurIPS 2023*)

2. **High-dimensional mean estimation with varied collection costs**
   (Clinton, Y. Chen, Zhu, Kandasamy, *Ongoing work*)

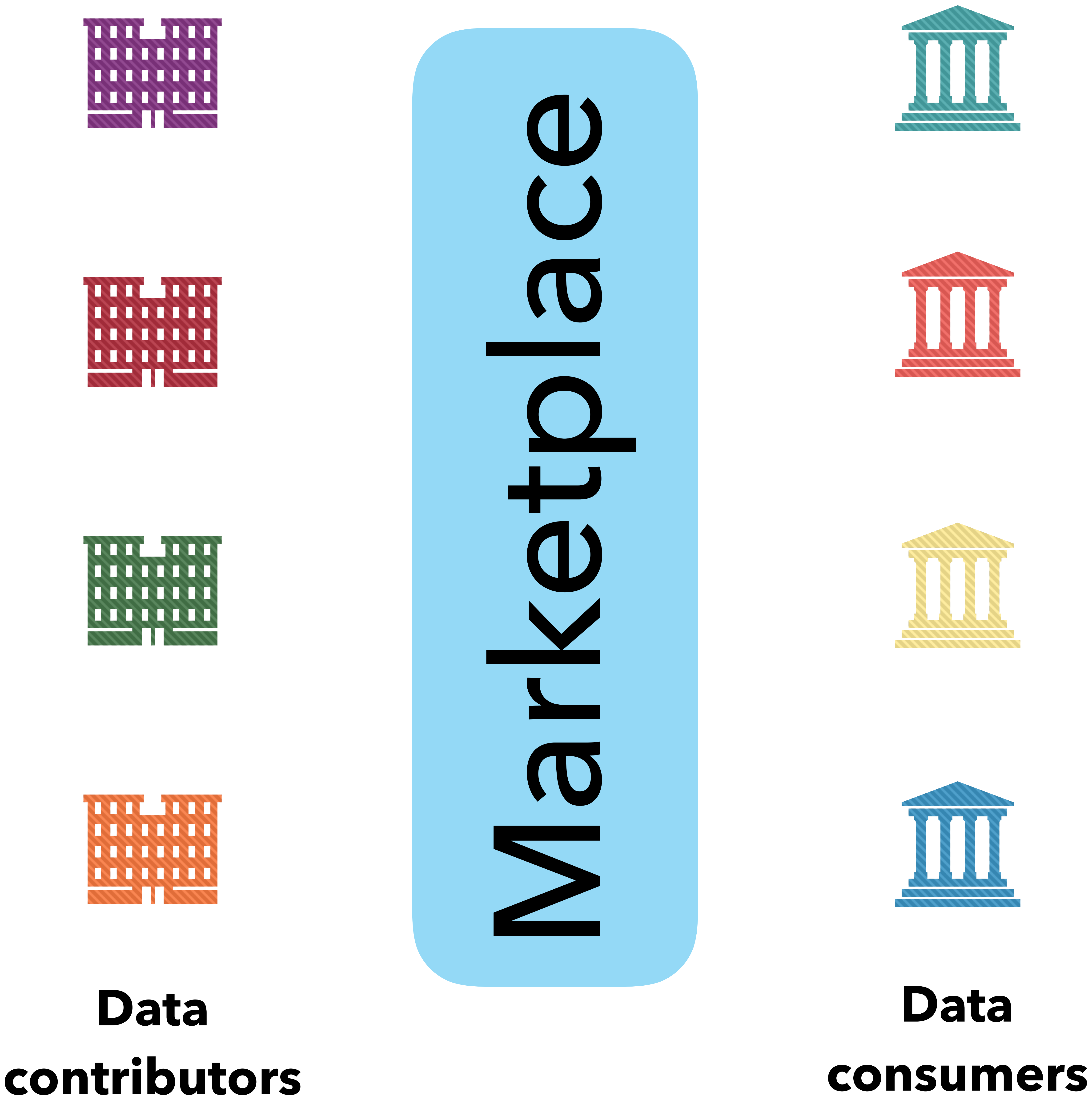
3. **Learning to price data in data marketplaces**
   (K. Chen, Huh, Kandasamy, *Ongoing work*)

DATA MARKETPLACES

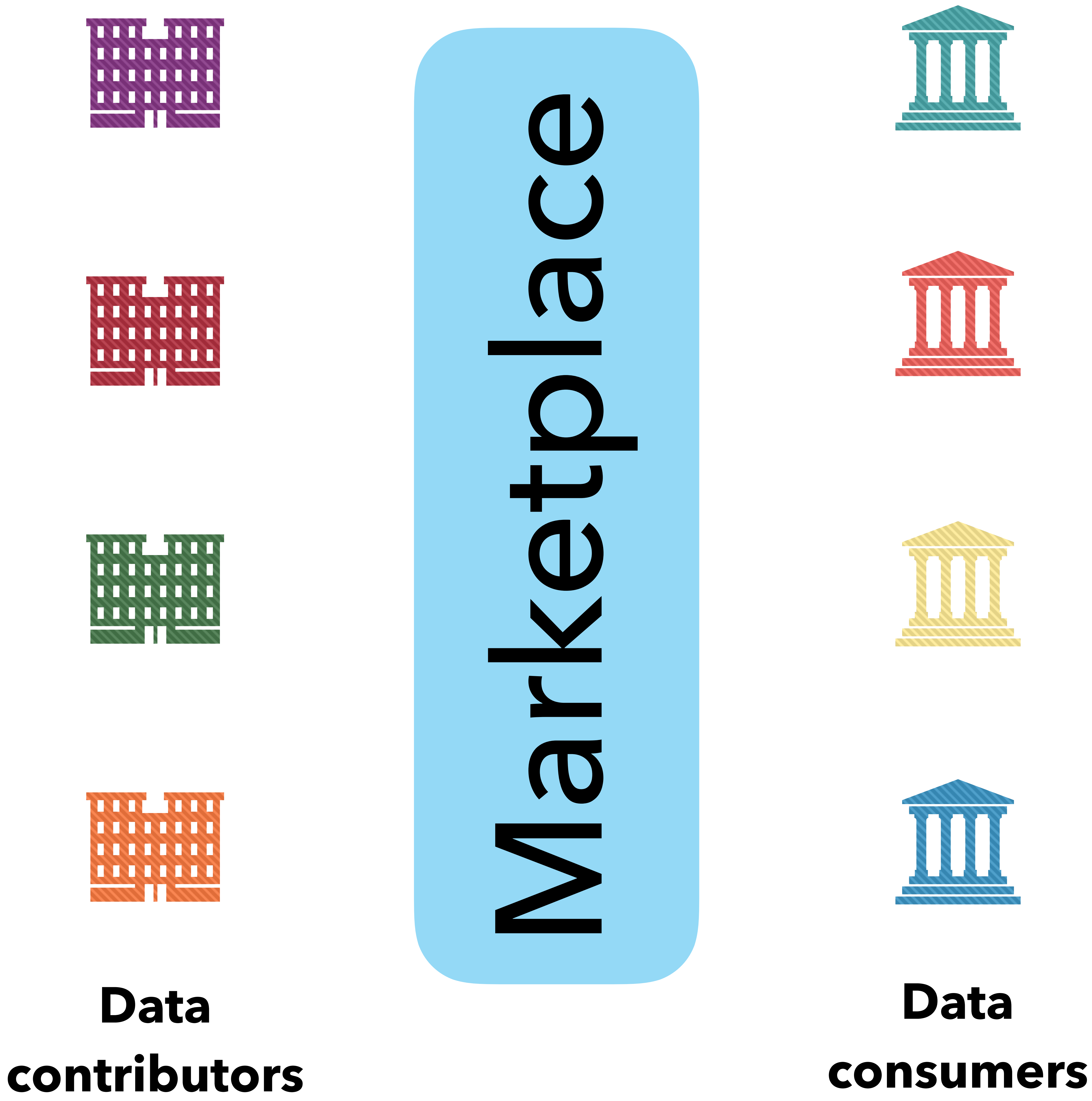


DATA MARKETPLACES

Consumers purchase data from contributors via a marketplace:



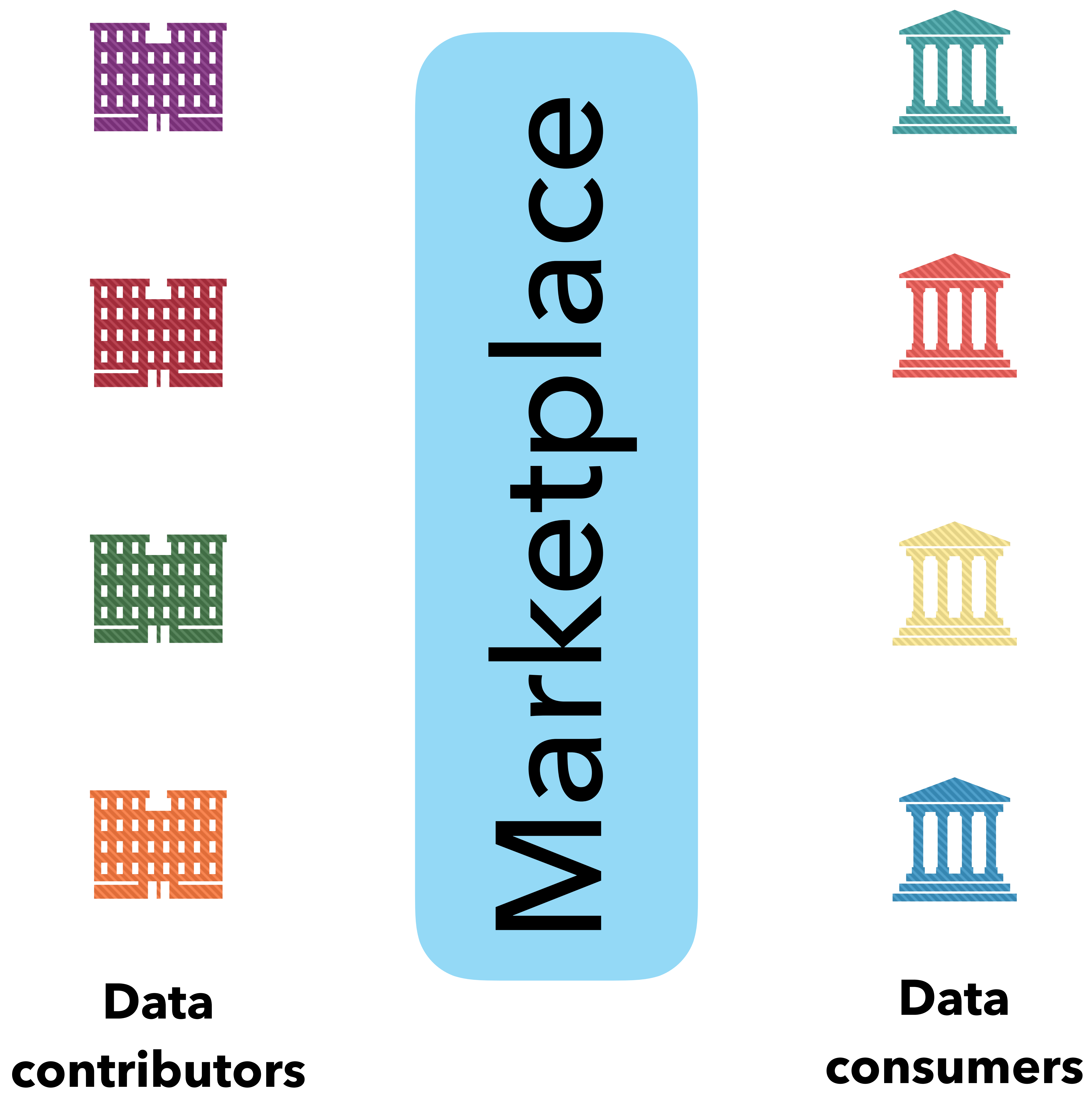
DATA MARKETPLACES



Consumers purchase data from contributors via a marketplace:

- Ensure contributors do not fabricate/poison data.

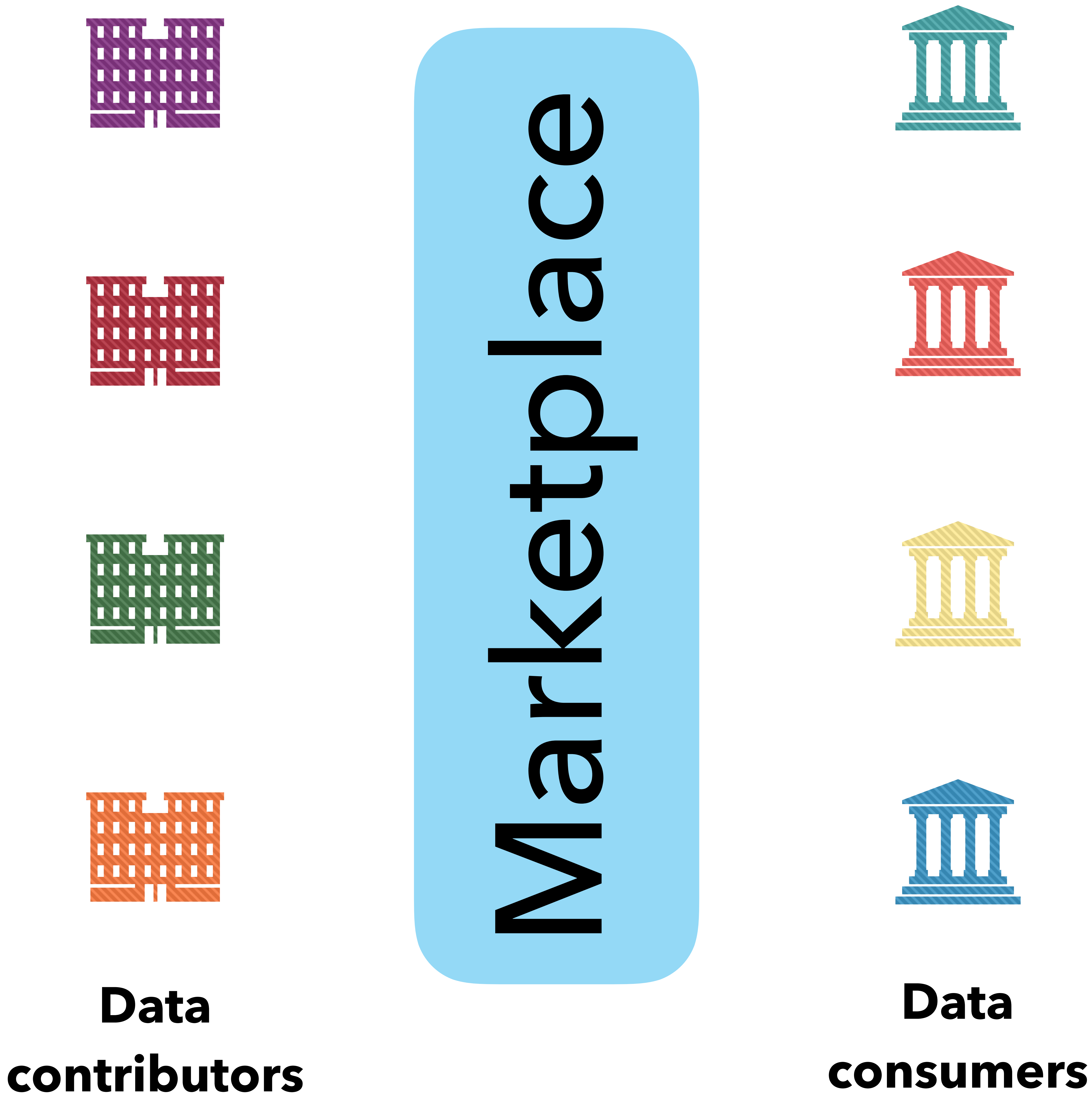
DATA MARKETPLACES



Consumers purchase data from contributors via a marketplace:

- Ensure contributors do not fabricate/ poison data.
- Pricing data that is being sold to consumers.

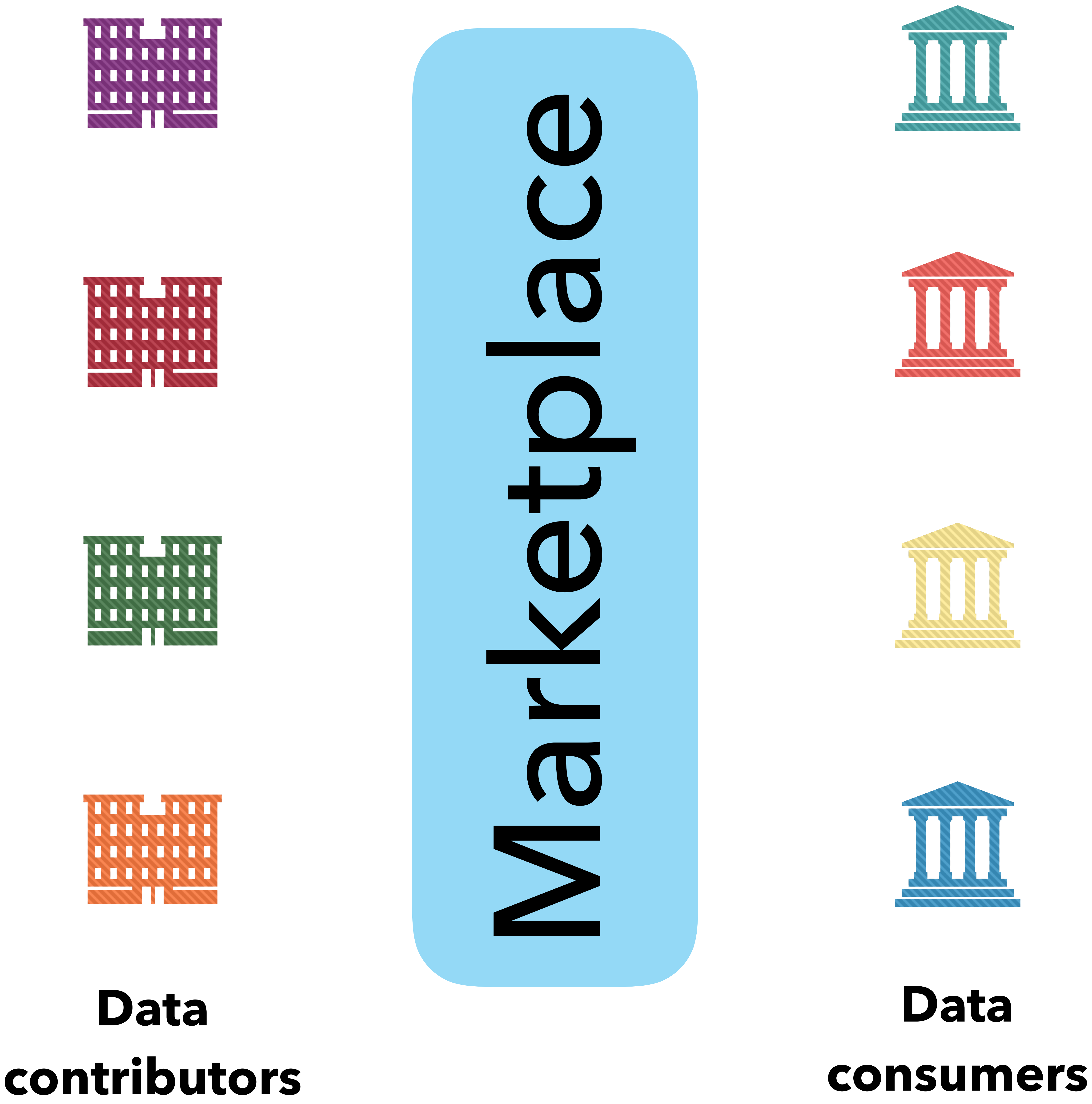
DATA MARKETPLACES



Consumers purchase data from contributors via a marketplace:

- Ensure contributors do not fabricate/ poison data.
- Pricing data that is being sold to consumers.
- Re-distributing the revenue back to the contributors.

DATA MARKETPLACES



Consumers purchase data from contributors via a marketplace:

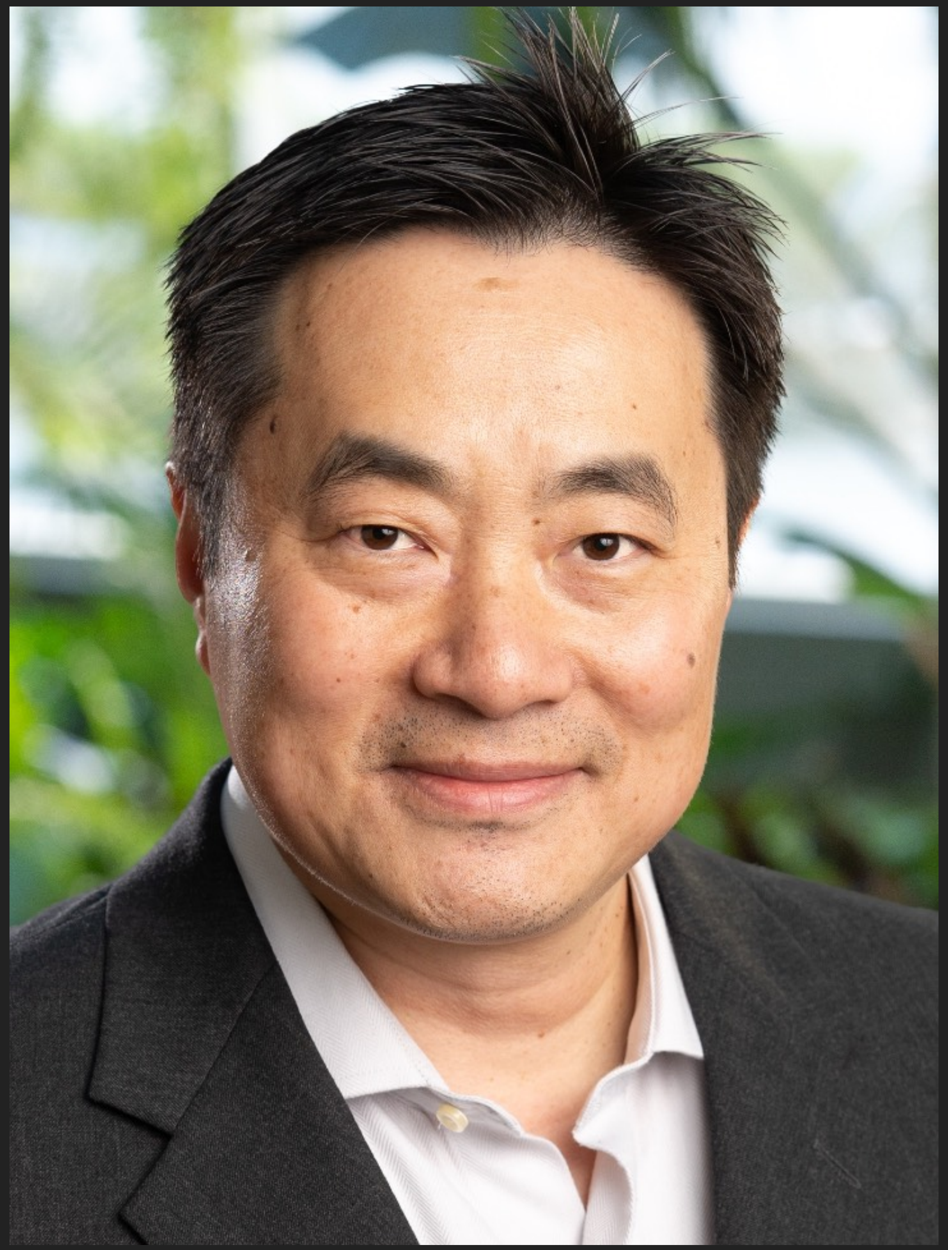
- Ensure contributors do not fabricate/ poison data.
- Pricing data that is being sold to consumers.
- Re-distributing the revenue back to the contributors.
- Learn consumer valuation of data via online feedback.

Keran Chen

Yiding Chen

Alex Clinton

Joon Suk Huh



Jerry Zhu

THANK YOU!

kandasamy@cs.wisc.edu