Data Reliability Scoring

Yiling Chen, Shi Feng, Paul Kattuman, Fang-Yi Yu

Harvard John A. Paulson School of Engineering and Applied Sciences

GEORGE MASON UNIVERSITY

Outline

- What is data reliability score?
- How can we define data reliability?
- Gram Determinant score and main result
- Experiments

Data reliability

POLITICS | POLICY

Trump Orders Firing of Bureau of Labor Statistics Chief

The president's move throws the quality of America's statistical apparatus into question

By *Justin Lahart* Follow, *Alex Leary* Follow and *Matt Grossman* Follow

Updated Aug. 1, 2025 6:33 pm ET

Data reliability

Econ Lett. 2022 Apr; 213: 110348.

Published online 2022 Feb 7. doi: 10.1016/j.econlet.2022.11

Data misreporting during the COVID19 crisis

Antonis Adam[a,b] and Sofia Tsarsitalidou[c,d,*]

POLITICS | POLICY

Trump Orders Firing of Bureau of Labor Statistics Chief

The president's move throws the quality of America's statistical apparatus into question

By Justin Lahart Follow, Alex Leary Follow and Matt Grossman Follow

Updated Aug. 1, 2025 6:33 pm ET

Data reliability

Econ Lett. 2022 Apr; 213: 110348.

Published online 2022 Feb 7. doi: 10.1016/j.econlet.2022.11

Data misreporting during the COVID19 crisis

Antonis Adam[a,b] and Sofia Tsarsitalidou[c,d,*]

POLITICS | POLICY

Trump Orders Firing of Bureau of Labor Statistics Chief

The president's move throws the quality of America's statistical apparatus into question

By Justin Lahart Follow, Alex Leary Follow and Matt Grossman Follow

Updated Aug. 1, 2025 6:33 pm ET

Perspect Health Inf Manag. 2021 Oct 1;18(4):1f.

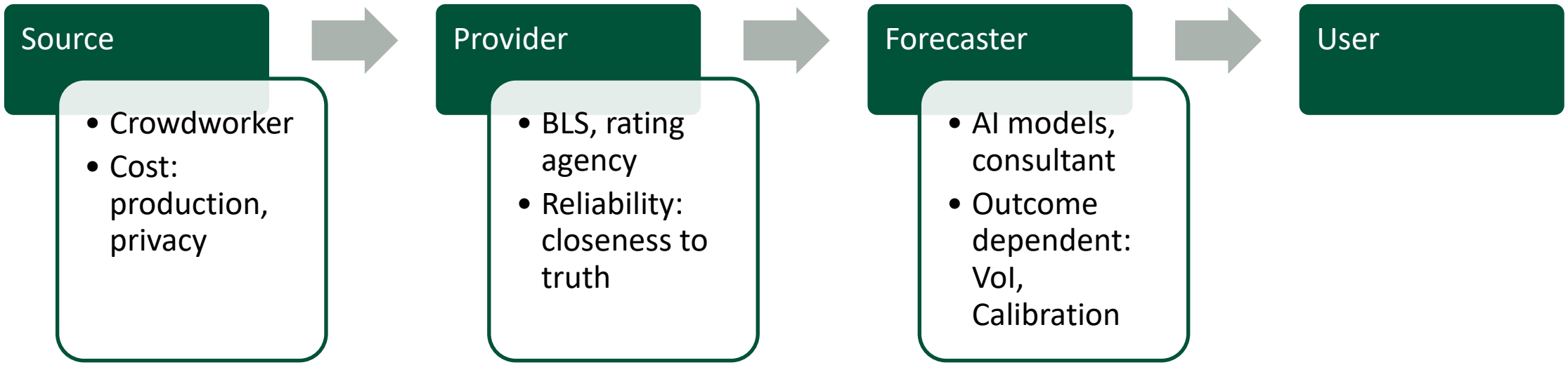
UPCODING MEDICARE: IS HEALTHCARE FRAUD AND ABUSE INCREASING?

Alberto Coustasse

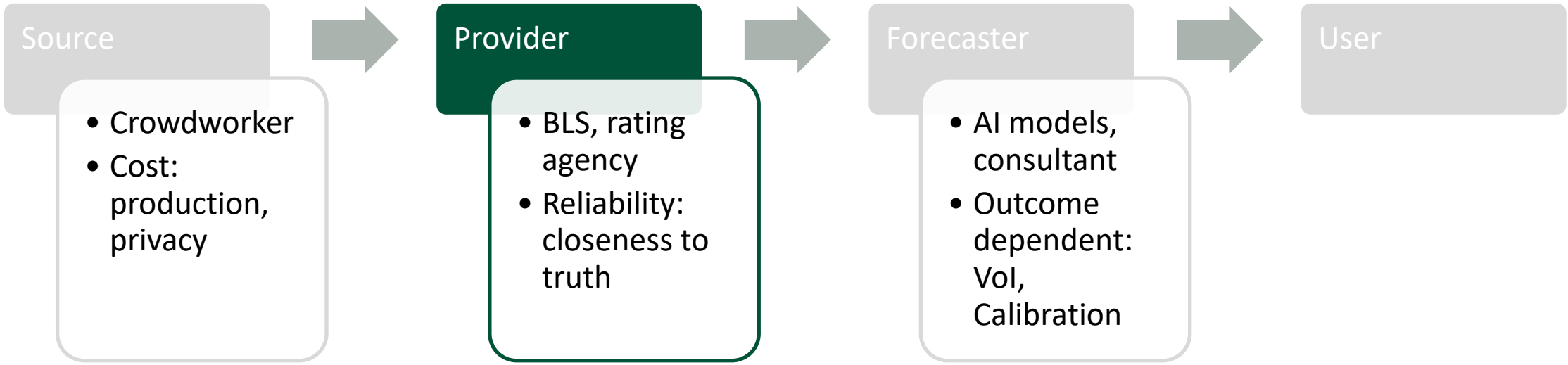
Article notes Copyright and License information

PMCID: PMC8649706 PMID: 34975355

Scores for Data



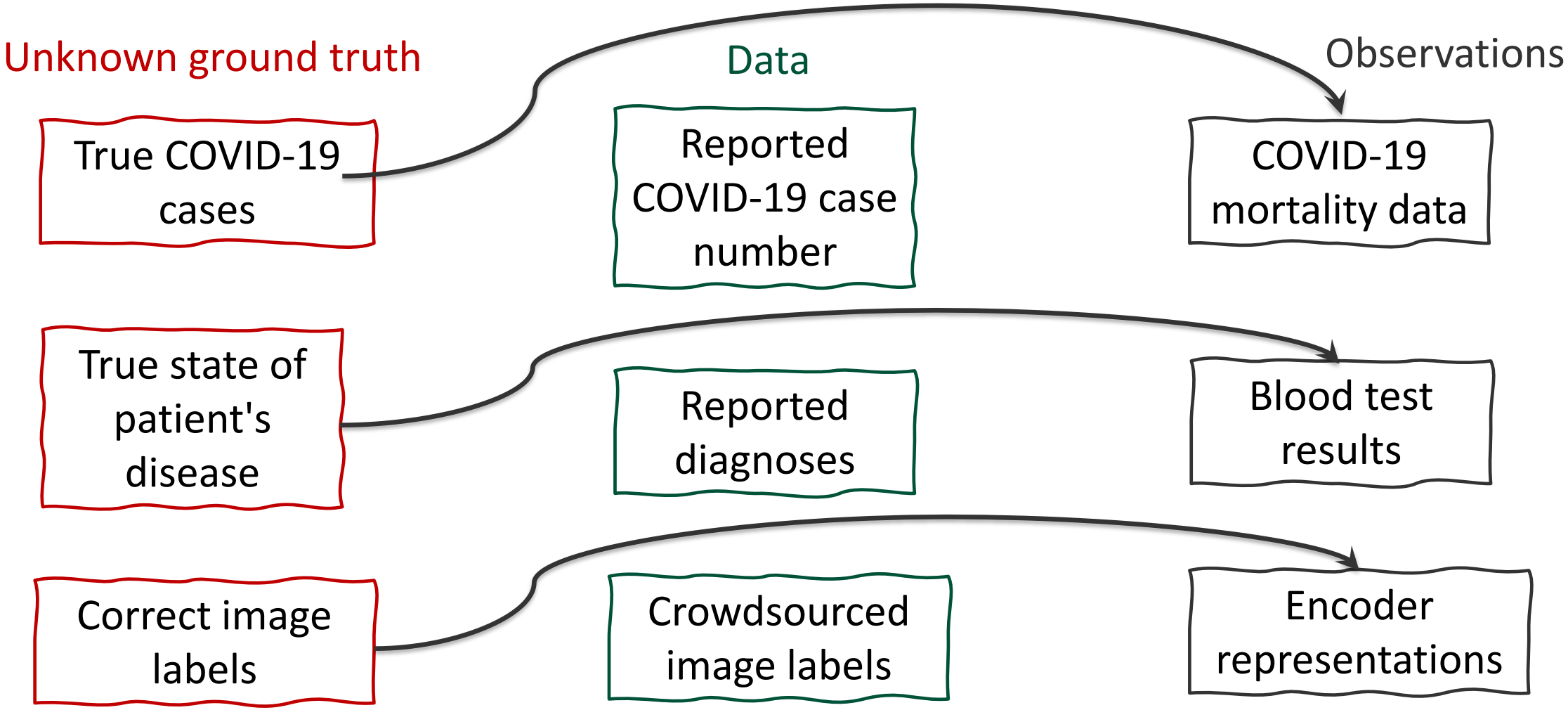
Data reliability score



Why difficult?

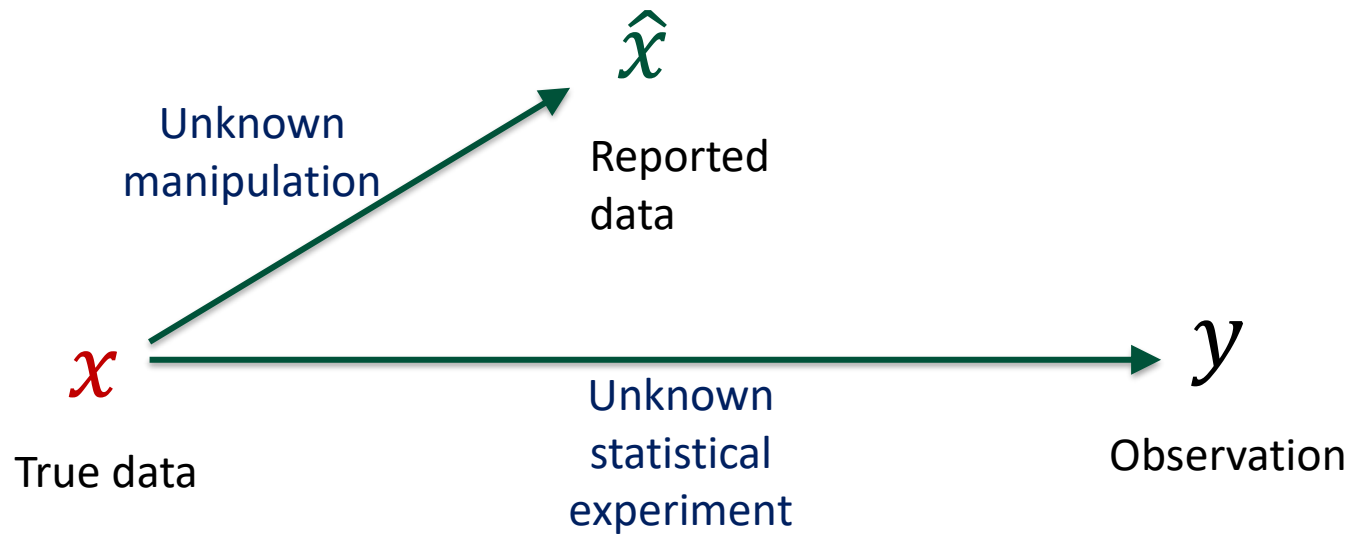
Unknown ground truth	Data
True COVID-19 cases	Reported COVID-19 case number
True state of patient's disease	Reported diagnoses
Correct image labels	Crowdsourced image labels

Why promising?



Research Question

- Can we design a reliability score $S(\hat{x}, y)$ that assigns a higher expected score to datasets that are closer to the ground truth?

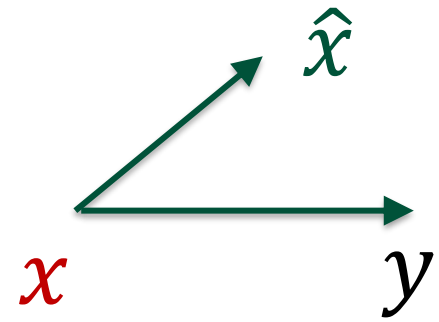


Outline

- What is data reliability score?
- How can we define data reliability?
- Gram determinant score and main result
- Experiments

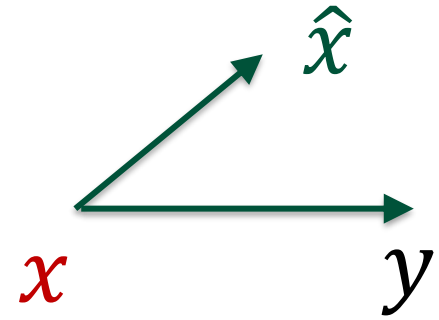
Reliability Ordering of Datasets

- To compare $\hat{x}$ and $\hat{x}'$ w.r.t. true data x, we propose partial orderings $>^{\cdot}$
  - Exact match ordering: $\hat{x} >^{x}_{EXACT} \hat{x}'$ if $\hat{x} = x$ but $\hat{x}' \neq x$



Reliability Ordering of Datasets

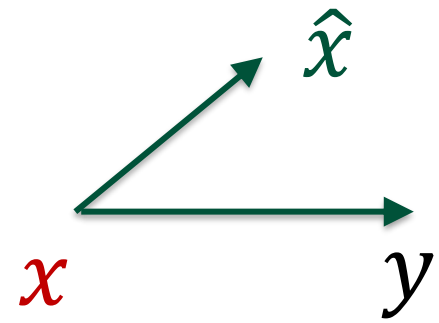
- To compare $\hat{x}$ and $\hat{x}'$ w.r.t. true data x, we propose partial orderings $>^{\cdot}$
  - Exact match ordering: $\hat{x} >^{x}_{EXACT} \hat{x}'$ if $\hat{x} = x$ but $\hat{x}' \neq x$
  - Blackwell dominant ordering: one is a garbling of the other
  - Hamming ordering: $\hat{x}$ is closer to x than $\hat{x}'$ in Hamming distance



Scoring to Preserve Reliability Ordering

- To compare $\hat{x}$ and $\hat{x}'$ w.r.t. true data x, we propose partial orderings $>^{\cdot}$
  - Exact match ordering
  - Blackwell dominant
  - Hamming ordering
- A reliability score S preserves $>^{\cdot}$ if

$$\mathbb{E}[S(\hat{x}, y)] > \mathbb{E}[S(\hat{x}', y)] \text{ for all } \hat{x} >^{x} \hat{x}'$$



Problem Formulation

Design reliability score S preserves $>^{\cdot}$ so that

$$\mathbb{E}[S(\hat{x}, y)] > \mathbb{E}[S(\hat{x}', y)] \text{for all } \hat{x} >^{x} \hat{x}'$$

- Three datasets
  - (Hidden) true data $x = (x_1, \ldots, x_N) \in [d]^N$
  - Reports $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_N) \in [d]^N$
  - Observation $y = (y_1, \ldots, y_N) \in \mathcal{Y}^N$

Problem Formulation

Design reliability score S preserves $>^{\cdot}$ so that

$$\mathbb{E}[S(\hat{x}, y)] > \mathbb{E}[S(\hat{x}', y)] \text{for all } \hat{x} >^{x} \hat{x}'$$

- Three datasets
  - (Hidden) true data $x = (x_1, \ldots, x_N) \in [d]^N$
  - Reports $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_N) \in [d]^N$
  - Observation $y = (y_1, \ldots, y_N) \in \mathcal{Y}^N$
- Assumption: $y_n \sim P(\cdot, x_n)$ independently for all i

Outline

- What is data reliability score?
- How can we define data reliability?
- Gram Determinant score and main result
- Experiments

Main Result

We introduce **Gram Determinant score (GDS)**

$$\det(\hat{G})$$

where $\hat{G}$ is the Gram matrix of report and observation

Main Result

We introduce **Gram Determinant score (GDS)**

$$\det(\hat{G})$$

which preserves

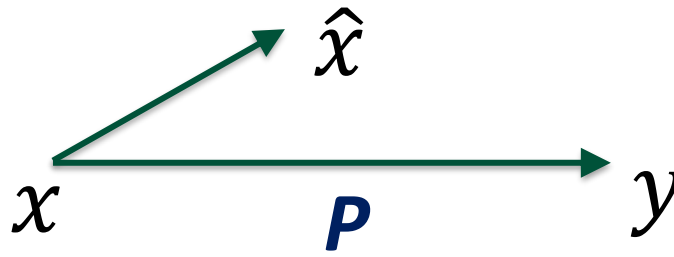
- Exact match ordering under $\mathcal{P}_{indep}$ and $\mathcal{Q}_{nonperm}$
- Blackwell ordering under $\mathcal{P}_{indep}$ and $\mathcal{Q}_{reg}$
- $\frac{1}{4L}$ −Hamming ordering under $\mathcal{P}_{indep}$ and $\mathcal{Q}_{L,K}$

Outline

- What is data reliability?
- How can we define data reliability?
- Gram Determinant score and main result
  Proof idea: GDS preserves Exact match
- Experiments

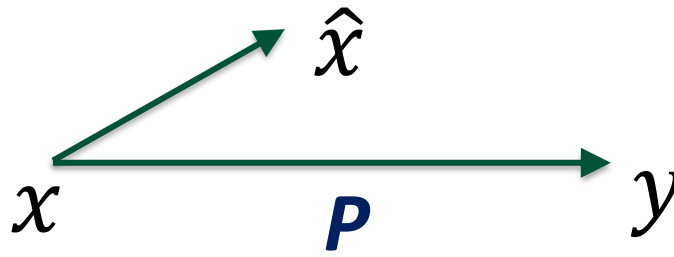
GDS=volume of $\hat{x}$, y dist.

- Given three datasets x, $\hat{x}$, y,
  - Experiment $P \in \mathbb{R}^{y \times d}$ independently for all i
  - Misreport matrix $Q \in \mathbb{R}^{d \times d}$: $Q(i,j) = \frac{\#[x_n = i, \hat{x}_n = j]}{N}$
  - Joint distribution and $\hat{x}$ and y: $PQ = PQ_x Q_{\hat{x}|x}^T$
  - Gram matrix of $\hat{x}$ and y: $\hat{G} = (PQ)^T PQ$



GDS=volume of $\hat{x}$, y dist.

- Given three datasets x, $\hat{x}$, y,
  - Experiment $P \in \mathbb{R}^{y \times d}$ independently for all i
  - Misreport matrix $Q \in \mathbb{R}^{d \times d}$: $Q(i,j) = \frac{\#[x_n = i, \hat{x}_n = j]}{N}$
  - Joint distribution and $\hat{x}$ and y: $PQ = PQ_x Q_{\hat{x}|x}^T$
  - Gram matrix of $\hat{x}$ and y: $\hat{G} = (PQ)^T PQ$



- GDS $\det(\hat{G}) = vol(PQ)$ volume spanned by the joint distribution

GDS preserves Exact match

Claim: $\mathbb{E}[S(\hat{x}', y)] < \mathbb{E}[S(x, y)]$for all $\hat{x} \neq x$

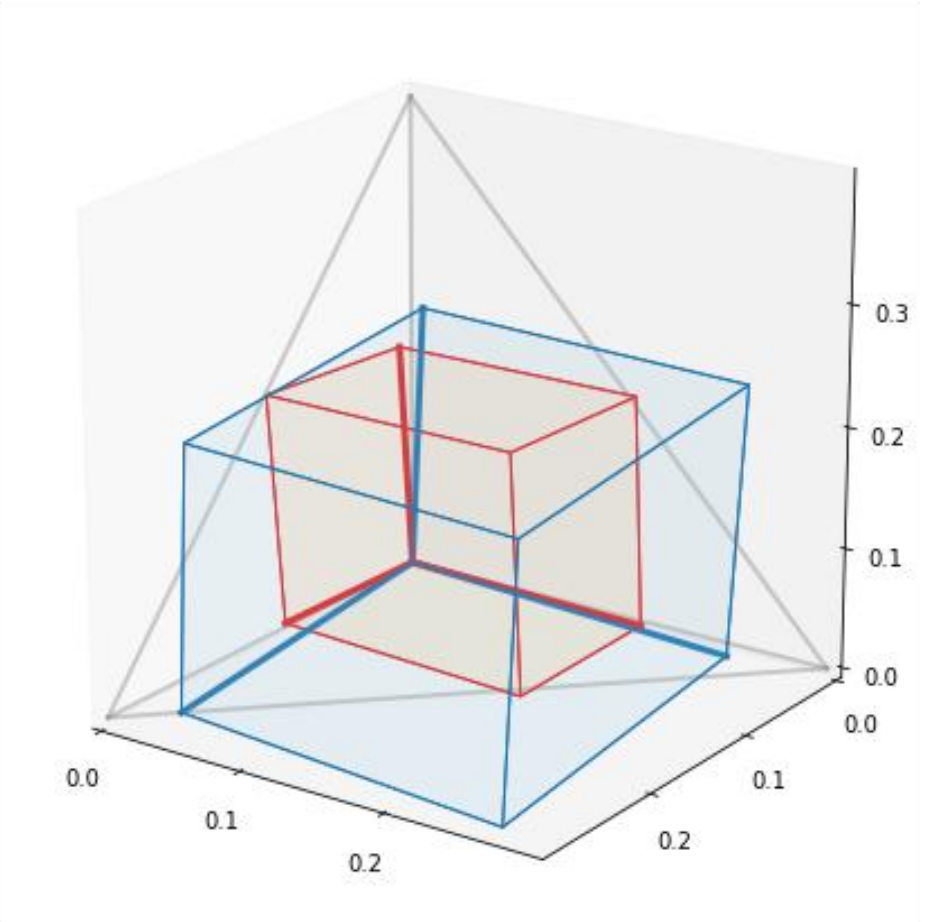
- Three facts for $vol(A) \coloneqq \det A^T A$
  1. $\det AB = \det A \det B$ and $\det B = \det B^T$
  2. $\det A = \prod eign_i(A)$
  3. eigenvalue of stochastic matrix≤ 1

GDS preserves Exact match

Claim: $\mathbb{E}[S(\hat{x}', y)] < \mathbb{E}[S(x, y)]$ for all $\hat{x} \neq x$

- Three facts for $vol(A) := \det A^T A$
  1. $\det AB = \det A \det B$ and $\det B = \det B^T$
  2. $\det A = \prod eign_i(A)$
  3. eigenvalue of stochastic matrix ≤ 1
- Proof:

$$\mathbb{E}[S(\hat{x}, y)] = vol\left(PQ_x Q_{\hat{x}|x}^T\right)$$
$$= vol(PQ_x)vol\left(Q_{\hat{x}|x}\right)$$
$$\leq vol(PQ_x) = \mathbb{E}[S(x, y)]$$

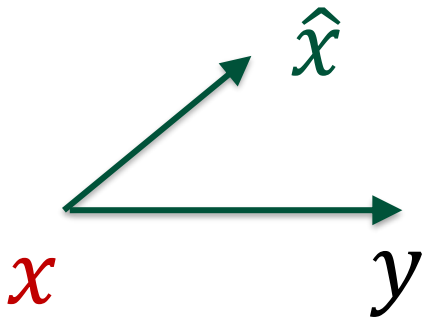


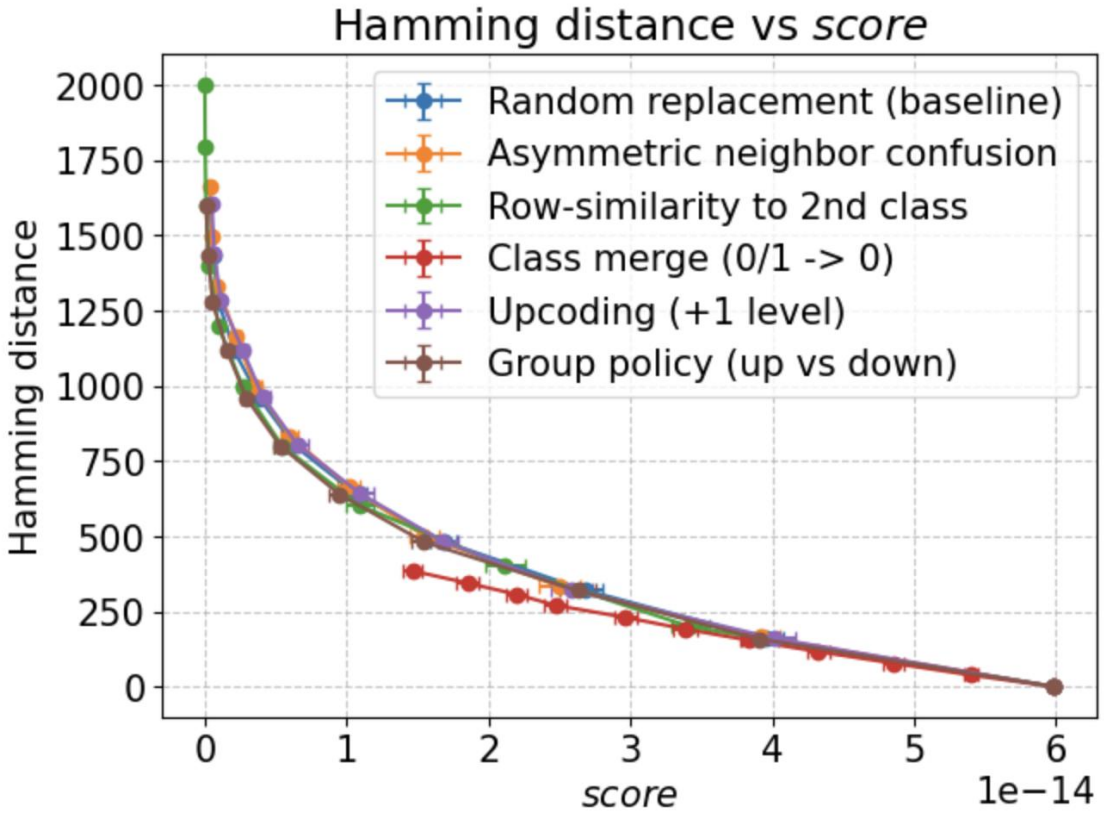
Outline

- What is data reliability score?
- How can we define data reliability?
- Gram Determinant score and main result
- Experiments

Empirical Results – Synthetic Data

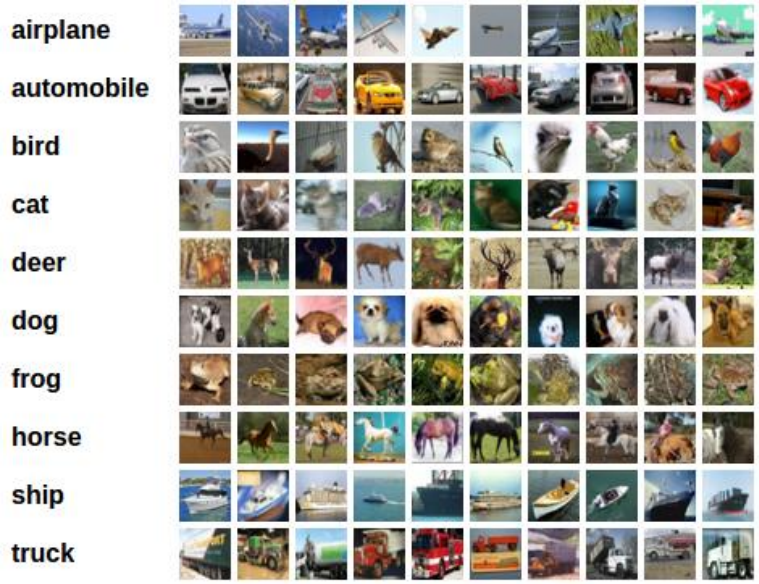
- $x \sim \{0,1,2,3,4\}^N$ with $N = 4000$
- $\hat{x}$: noisy label
- y: random experiment $P \in [0,1]^{5\times5}$

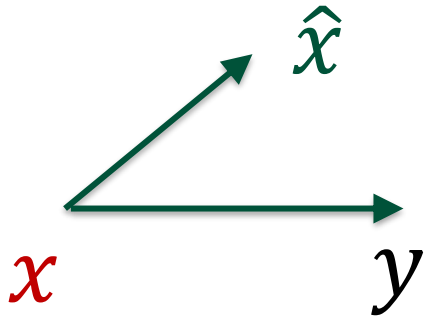


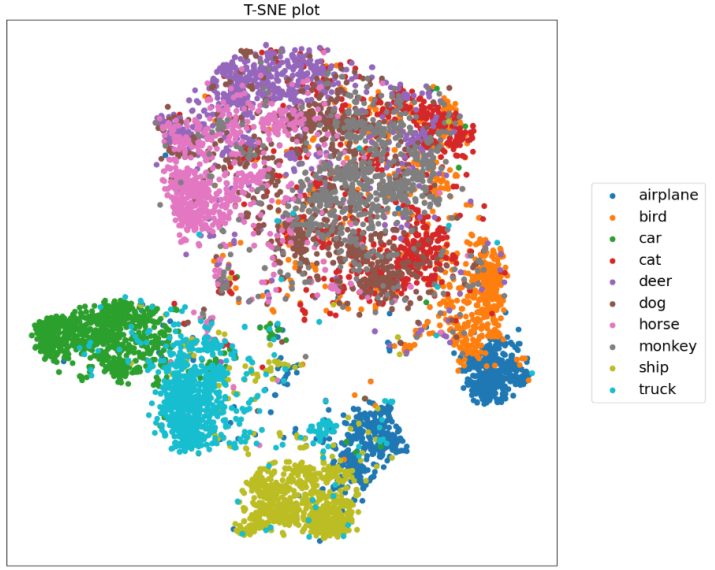


Empirical Results – CIFAR-10 data

- x: real label of image
- $\hat{x}$: noisy label
- y: feature embedding from contrastive learning

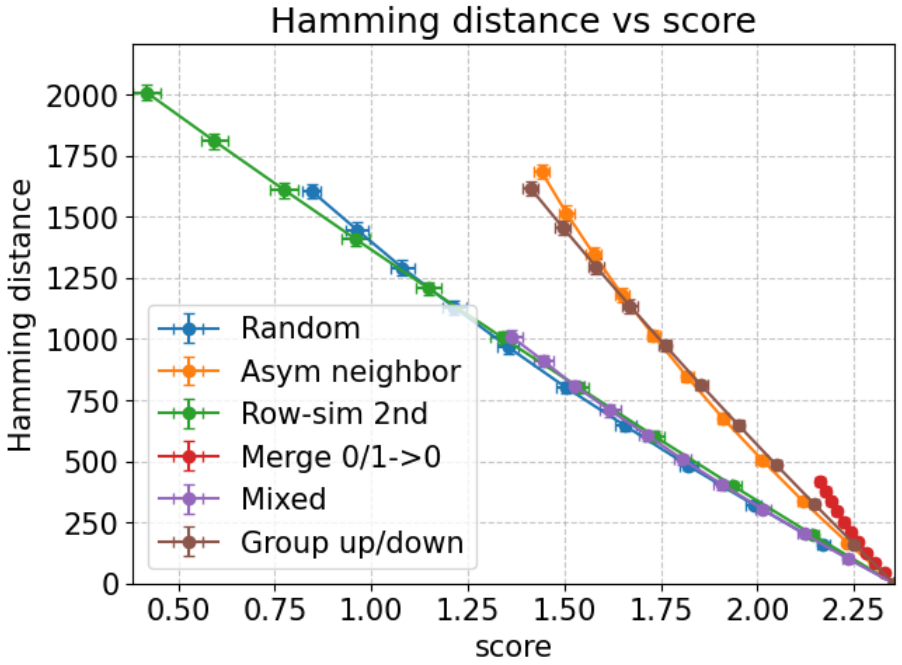






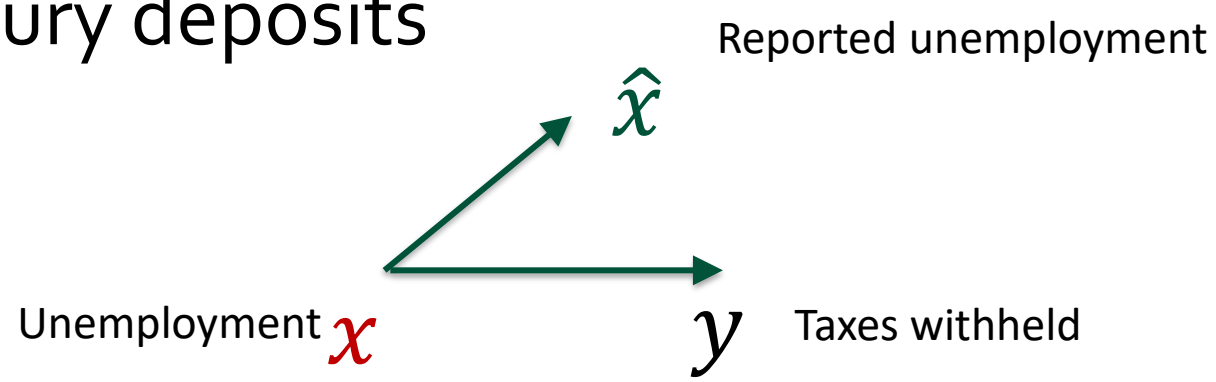
Empirical Results – CIFAR-10 data

- x: real label of image
- $\hat{x}$: noisy label
- y: feature embedding from contrastive learning



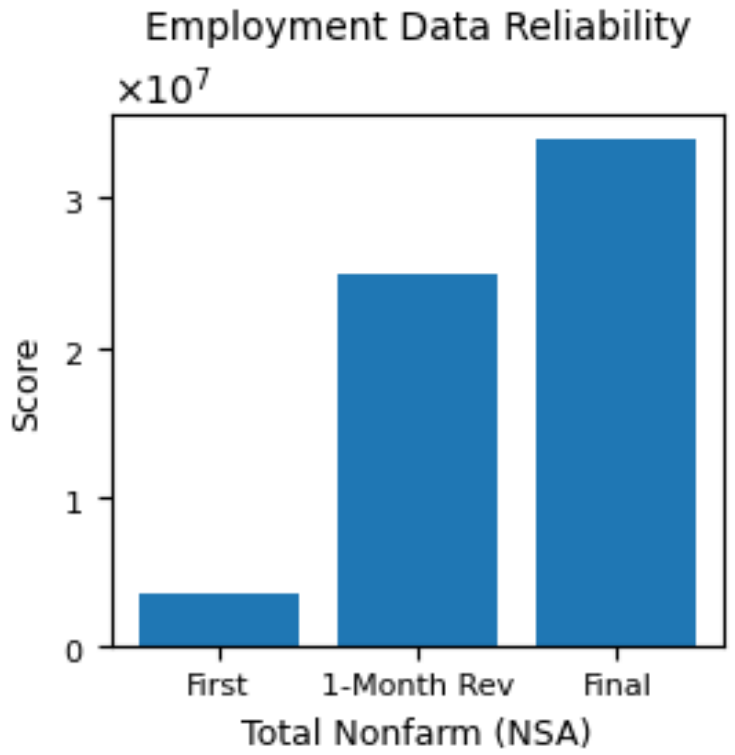
Unemployment data

- x: real unemployment number
- $\hat{x}$: reports
  - First report,
  - One month revision, and
  - Final revision
- y: Withheld Taxes from Treasury deposits



Unemployment data

- x: real unemployment number
- $\hat{x}$: reports
  - First report,
  - One month revision, and
  - Final revision
- y: Withheld Taxes from Treasury deposits



Discussion

- Summary
  - Reliability ordering
  - GDS: respects several orderings and agonistic to experiment
  - Kernel treat for high dimensional Y
- Future directions:
  - How to score data without verification?
    - Time series data, or large signal space X
    - ordering vs testing
  - Empirical results