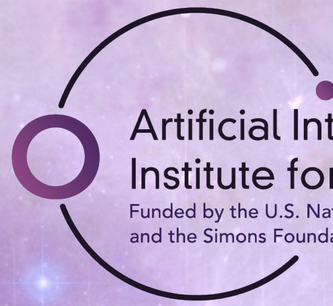
Robust-by-Design Distributional Learning from Contaminated Samples

Nevena Gligić, Arya Farahi

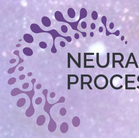
The University of Texas at Austin, NSF–Simons AI Institute for Cosmic Origins



The University of Texas at Austin
Department of Statistics and Data Sciences
College of Natural Sciences

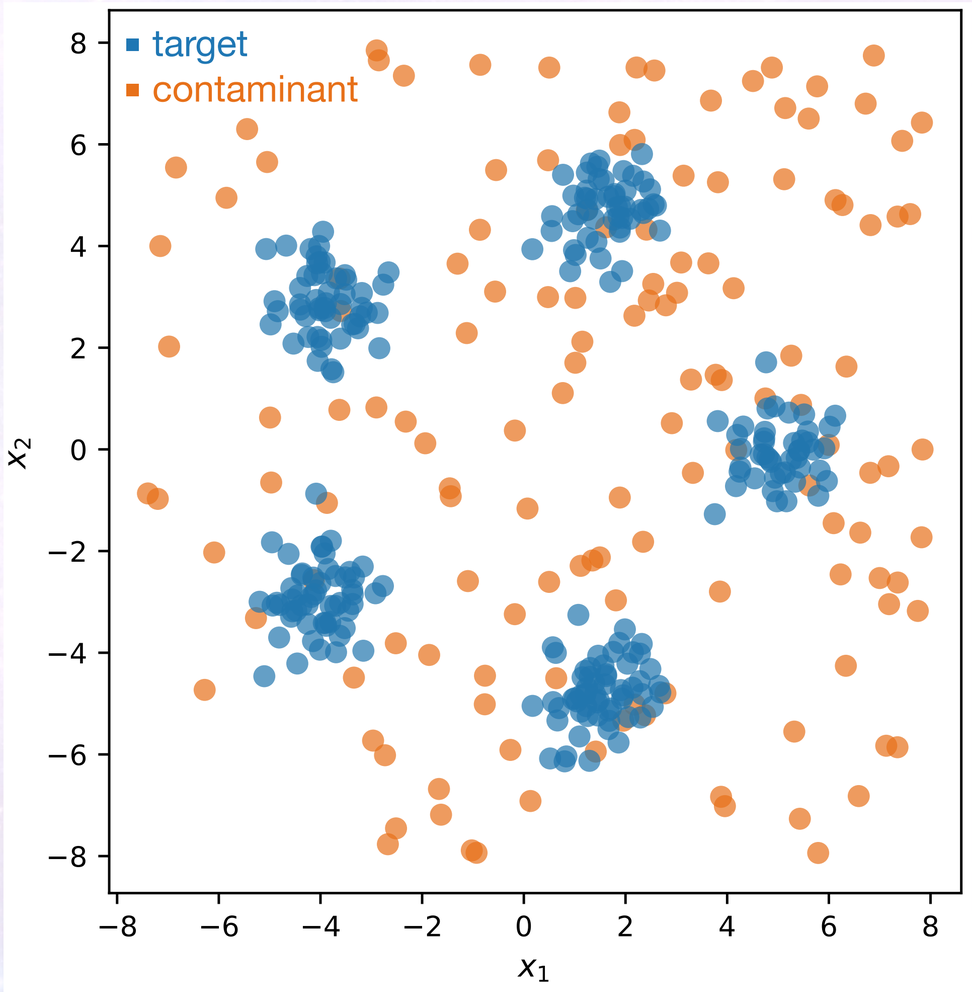


Artificial Intelligence Institute for Cosmic Origins
Funded by the U.S. National Science Foundation and the Simons Foundation



NEURAL INFORMATION PROCESSING SYSTEMS

Problem





■ MNIST digits (target) ■ Kuzushiji (contaminant)

Ancient DNA

modern human DNA
in ancient samples

Galaxy catalogs

stars and interlopers
among the galaxies

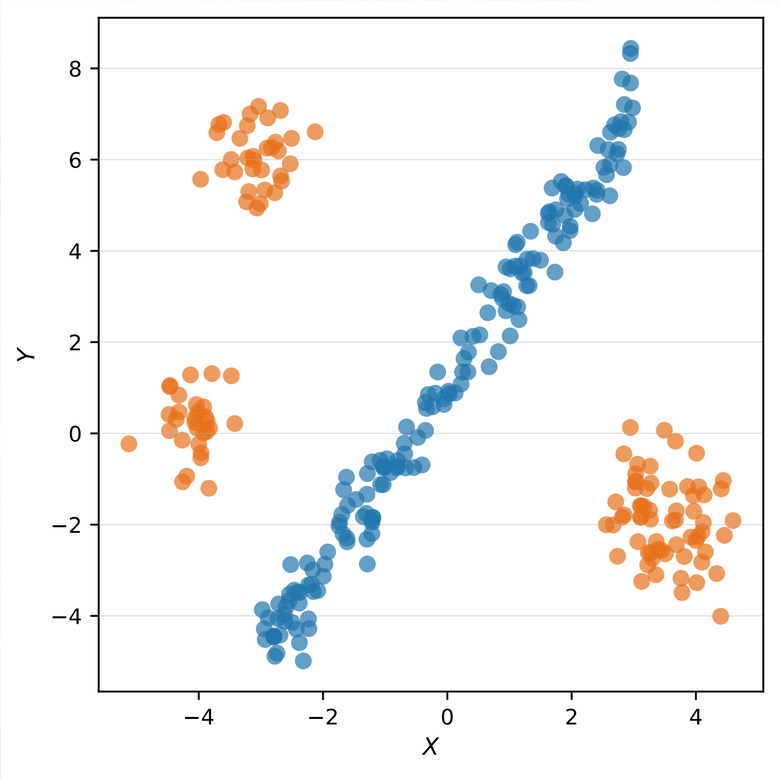
CosmicAI

Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE signal processing magazine, 29(6):141–142, 2012.

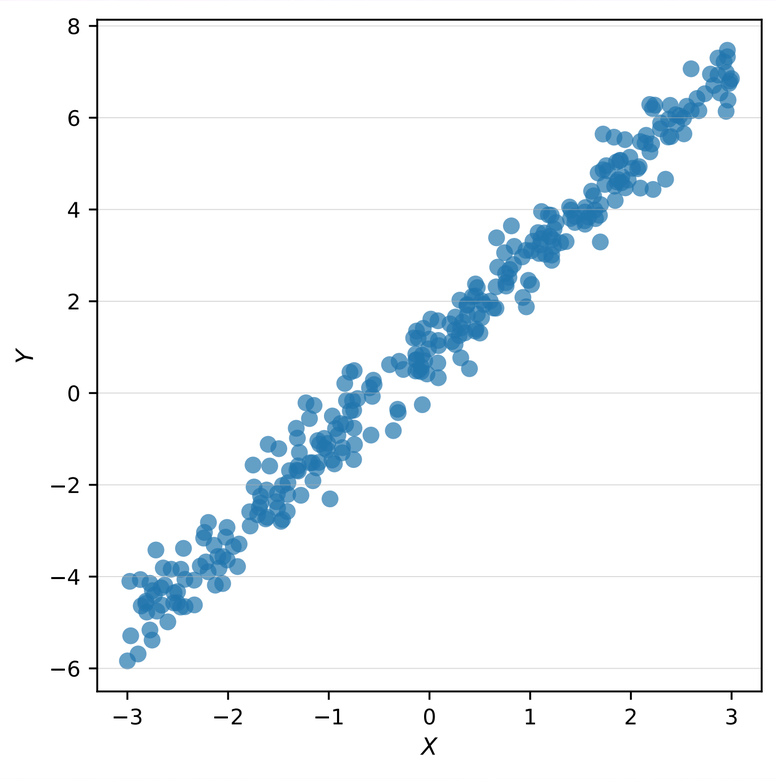
Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. arXiv preprint arXiv:1812.01718, 2018.

The setting

Contaminated dataset



Clean dataset



🎯 $\mathrm{MMD}^2(p, q)$

Maximum Mean Discrepancy (MMD), a kernel based distance measure between two distributions

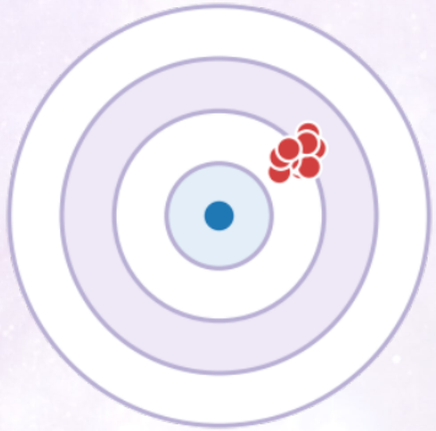
$$W_1, \ldots, W_n \overset{iid}{\sim} r$$

$$r = (1-\epsilon)q + \epsilon z \quad \epsilon \in [0,1)$$

$$X_1, \ldots, X_n \overset{iid}{\sim} p$$

CosmicAI

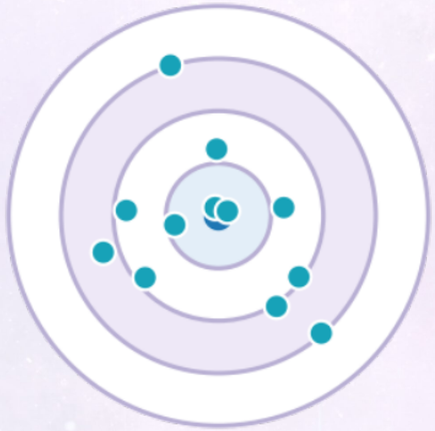
Common approaches



Whole contaminated pool (naive)

✓ no data waste

✗ biased: $\mathrm{MMD}^2(p, r)$



Audit only

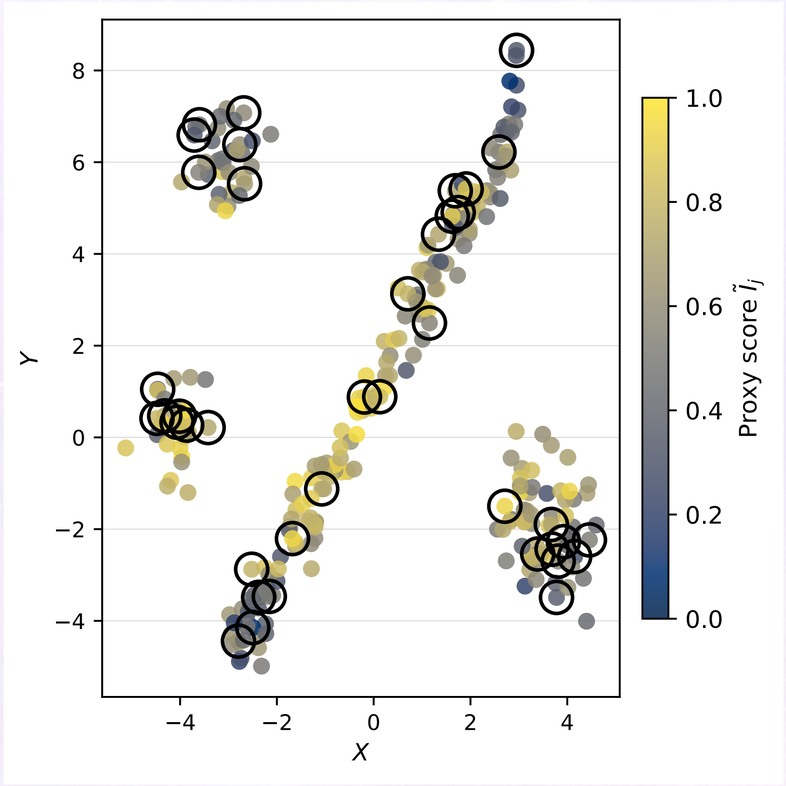
✓ unbiased: $\mathrm{MMD}^2(p, q)$

✗ data waste

CC-MMD



Audit subset $I_{j\in L} \in \{0,1\}$ + Proxy scores $\tilde{I}_j \in [0,1]$ → **CC-MMD**



✓ $\mathrm{MMD}^2(p, q)$

✓ no data waste

$L \subset \{1,\ldots,N\}$ with $|L| = m \ll N$

$I_{j\in L} \in \{0,1\}, \quad \tilde{I}_j \in [0,1]$

CosmicAI

Theoretical guarantees

Asymptotic normality

The gap between CC-MMD and the oracle ratio estimator is asymptotically Gaussian and shrinks at the $\sqrt{m}$ scale.

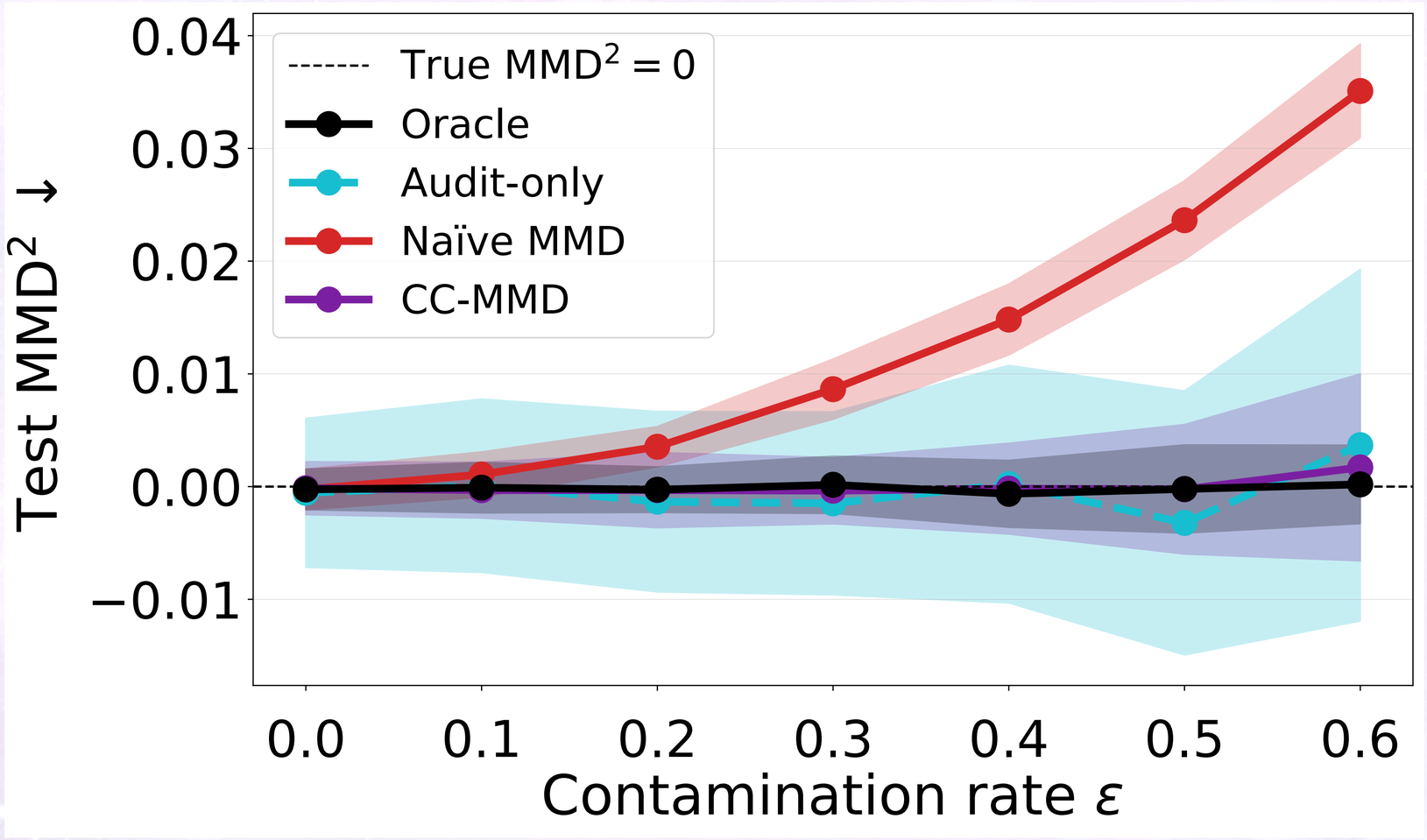
Theorem 3.2, 3.3

Consistency

$$\widehat{\mathrm{MMD}}^2_{CC} \xrightarrow{\mathbb{P}} \mathrm{MMD}^2(p, q)$$

Theorem 3.4

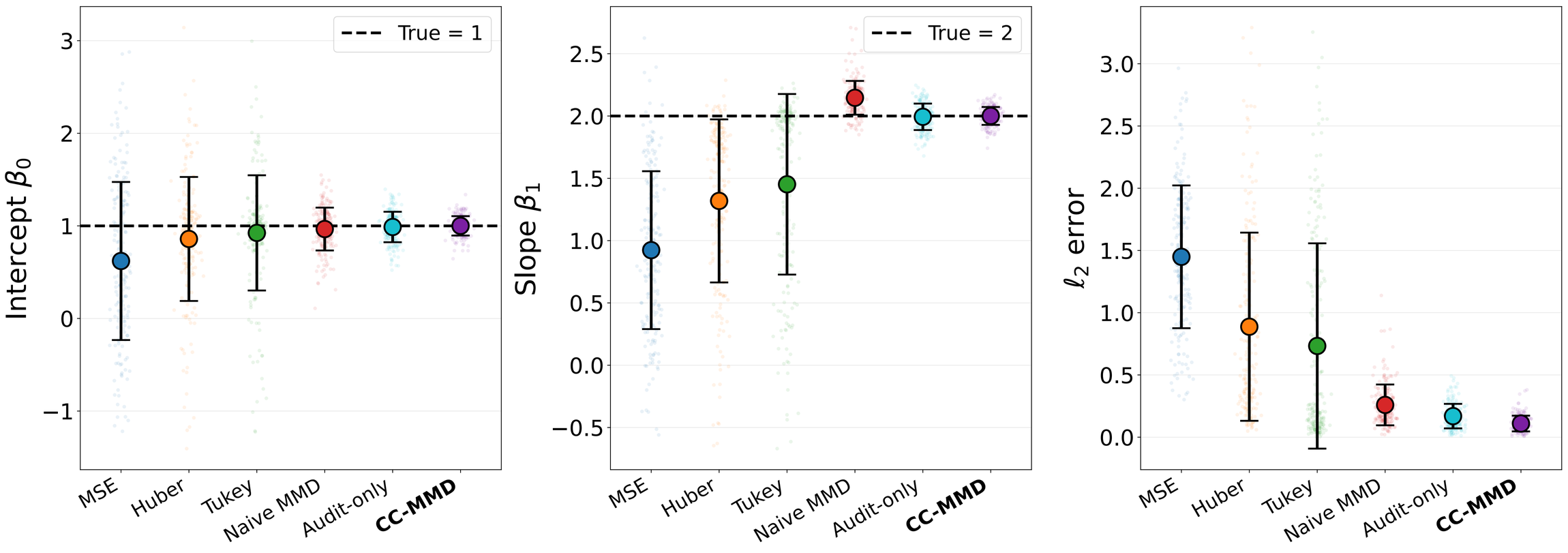
Results: MMD estimation



MMD estimation under increasing contamination.

CC-MMD follows the oracle with less spread than audit-only.

Results: regression



Parameter recovery in a contaminated linear regression problem over 200 independent realizations.

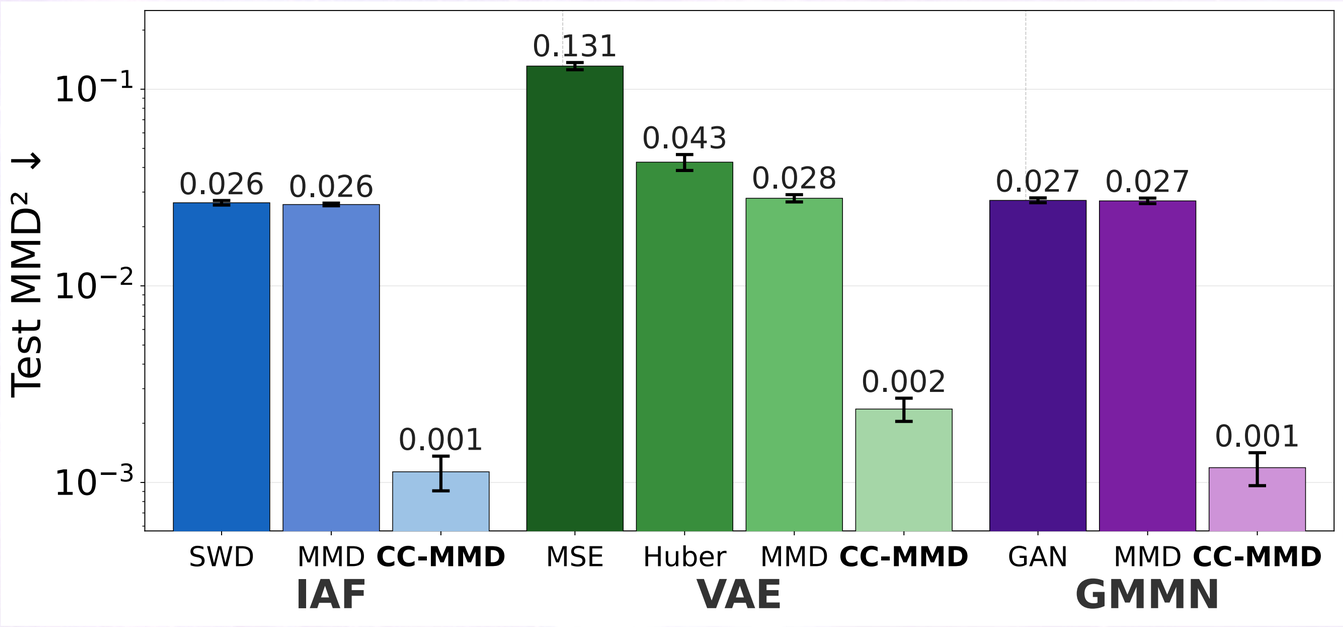
The smallest spread.

Centered on the true parameters.

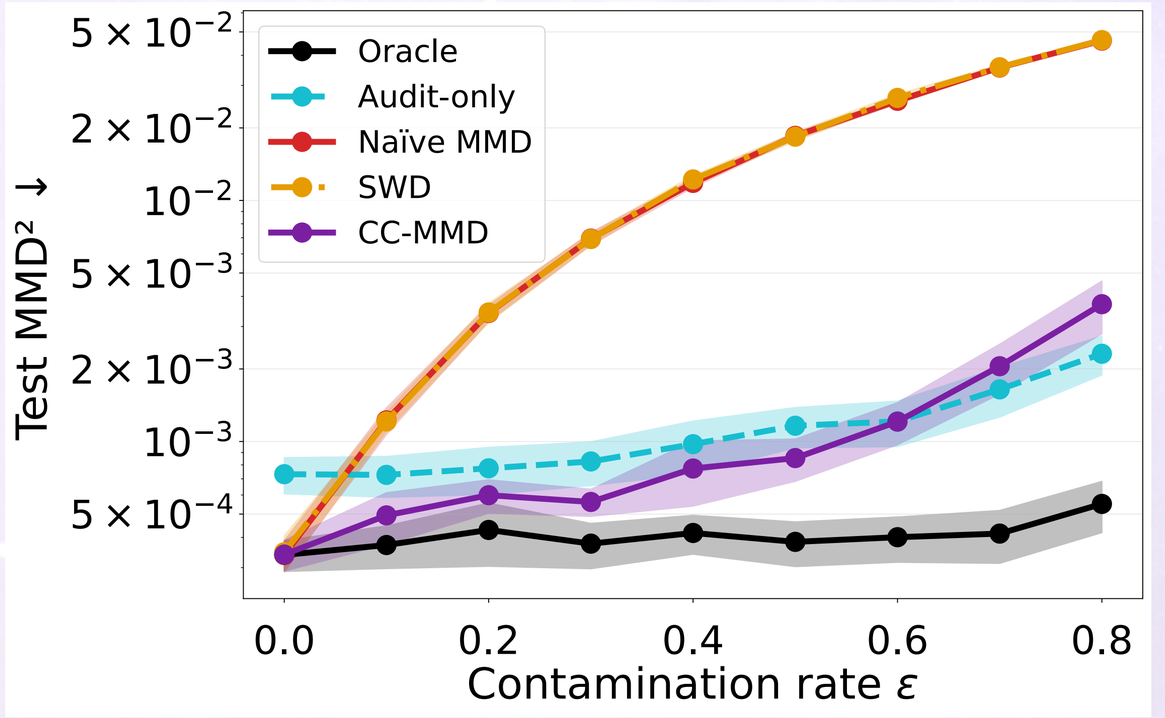
$> 35\,\%$ lower ℓ_2.

CosmicAI

Results: image generation



Generative modeling task on MNIST contaminated with K-MNIST.
Test MMD^2 of the generated images against clean MNIST images.



Generative quality vs. contamination rate.

$> 11\times$
lower test MMD^2.

Robust to increasing contamination.

CosmicAI

Takeaways

- Contaminated data ≠ unusable data
- Audit labels + proxy scores $\Longrightarrow$ CC-MMD
- Robust by design
- Works with various data types and applications

Open questions: audit size vs dimensions, poor audit subset quality, multiple contamination sources

Thank you!

https://github.com/NevenaGligic/CC-MMD

Supported by the National Science Foundation (Cooperative Agreement No. 2421782) and from the Simons Foundation (Award No. MPS-AI-00010515), through the NSF–Simons AI Institute for Cosmic Origins. Compute by the Texas Advanced Computing Center at The University of Texas at Austin.

The University of Texas at Austin
Department of Statistics and Data Sciences
College of Natural Sciences

Artificial Intelligence Institute for Cosmic Origins
Funded by the U.S. National Science Foundation and the Simons Foundation

NEURAL INFORMATION PROCESSING SYSTEMS