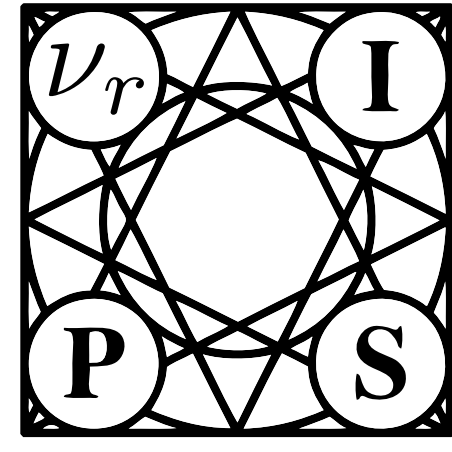


Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift



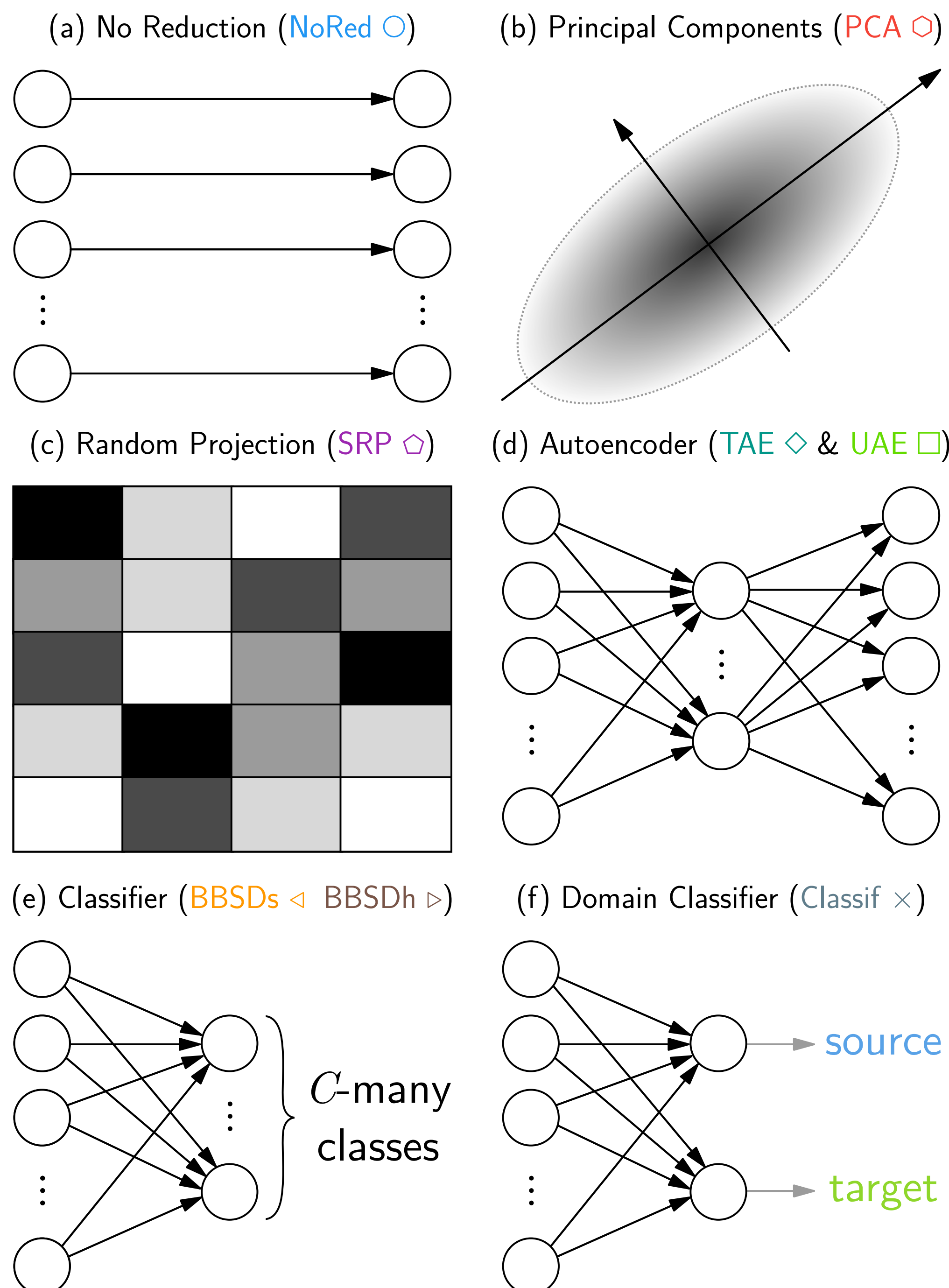
Stephan Rabanser¹ Stephan Günnemann² Zachary C. Lipton³
¹AWS AI Labs ²Technical University of Munich ³Carnegie Mellon University



Motivation

- The reliable functioning of software depends crucially on tests (unit tests, input validation).
- Despite their power, many machine learning models are sensitive to shifts in the data distribution.
- In practice, ML pipelines rarely inspect incoming data for any signs of distributional shift.
- Best practices for detecting shift in high-dimensional and real-life data have not yet been established.
- Existing solutions to addressing pure shifts like covariate shift or label shift often rely on strict preconditions, producing wrong results whenever these conditions are not met.

Dimensionality Reduction



Our Framework

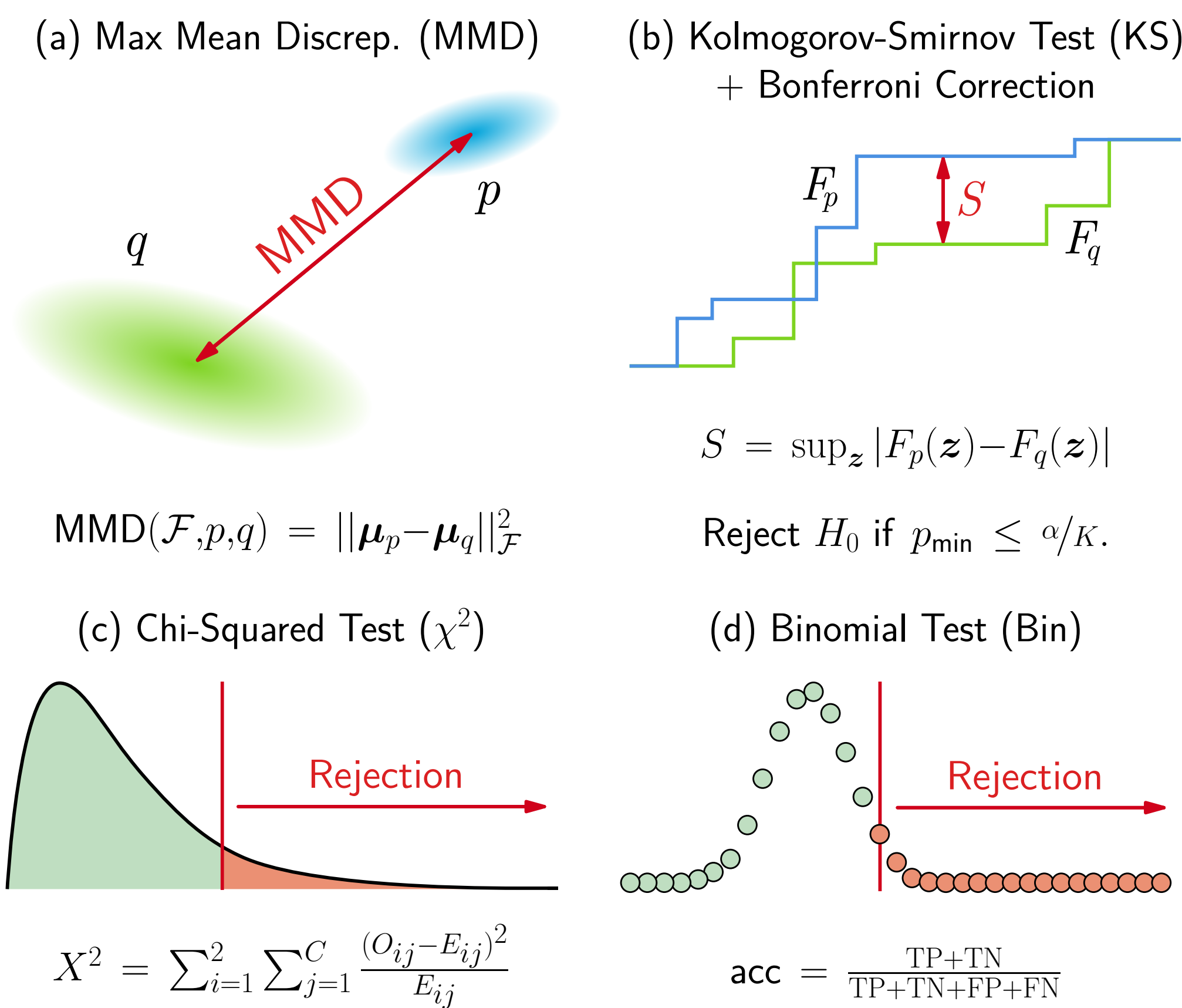
Given labeled data $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\} \sim p$ and unlabeled data $\{\mathbf{x}'_1, \dots, \mathbf{x}'_m\} \sim q$, our task is to determine whether $p(\mathbf{x})$ equals $q(\mathbf{x}')$:

$$H_0 : p(\mathbf{x}) = q(\mathbf{x}') \quad \text{vs} \quad H_A : p(\mathbf{x}) \neq q(\mathbf{x}').$$

We explore the following design considerations:

- what **representation** to run the test on;
- which **two-sample test** to run;
- when the representation is multidimensional; whether to run a **single multivariate test** or **multiple univariate two-sample tests**; and
- how to combine** their results.

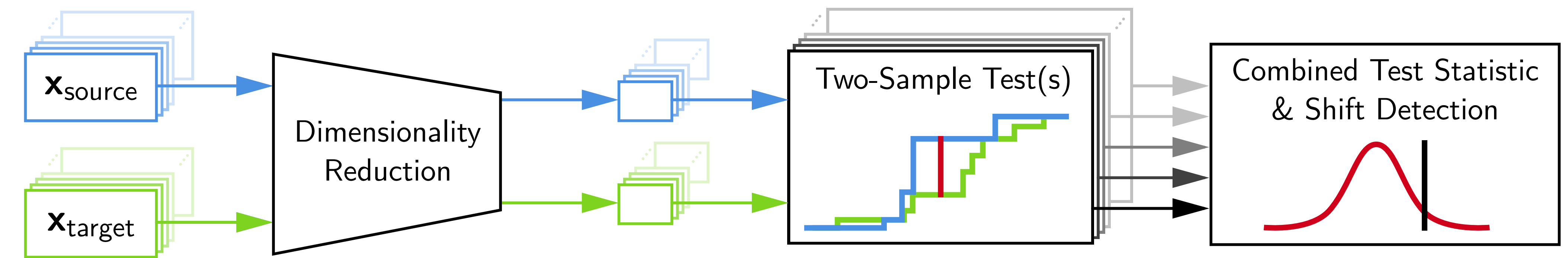
Statistical Two-Sample Testing



Key Observations

- Multiple univariate tests and multivariate kernel tests offer comparable detection performance.
- BBSDs** (univariate) and **UAE** (multivariate) are the best-performing shift detectors, respectively.
- Top different samples from **Classif** are helpful in characterizing a shift's nature and malignancy.
- MNIST original split is not i.i.d., but harmless.

Shift Detection Pipeline Overview



Key Results

(a) Detection accuracy of different dimensionality reduction techniques across all simulated shifts on MNIST and CIFAR-10. **Green bold** entries indicate the best DR method at a given sample size, **red italic** the worst.

Test	DR	Number of samples from test							
		10	20	50	100	200	500	1,000	10,000
Univ. tests	NoRed	0.03	0.15	0.26	0.36	0.41	0.47	0.54	0.72
	PCA	0.11	0.15	0.30	0.36	0.41	0.46	0.54	0.63
	SRP	0.15	0.15	0.23	0.27	0.34	0.42	0.55	0.68
	UAE	0.12	0.16	0.27	0.33	0.41	0.49	0.56	0.77
	TAE	0.18	0.23	0.31	0.38	0.43	0.47	0.55	0.69
	BBSDs	0.19	0.28	0.47	0.47	0.51	0.65	0.70	0.79
	Classif	0.01	0.03	0.11	0.21	0.28	0.42	0.51	0.67
Multiv. tests	NoRed	0.14	0.15	0.22	0.28	0.32	0.44	0.55	-
	PCA	0.15	0.18	0.33	0.38	0.40	0.46	0.55	-
	SRP	0.12	0.18	0.23	0.31	0.31	0.44	0.54	-
	UAE	0.20	0.27	0.40	0.43	0.45	0.53	0.61	-
	TAE	0.18	0.26	0.37	0.38	0.45	0.52	0.59	-
	BBSDs	0.16	0.20	0.25	0.35	0.35	0.47	0.50	-
	Classif	0.01	0.03	0.11	0.21	0.28	0.42	0.51	0.67

(b) Detection accuracy of different shifts on MNIST and CIFAR-10 using the best DR techniques. **Green bold** shifts are identified as harmless, **red italic** shifts as harmful.

Test	Shift	Number of samples from test							
		10	20	50	100	200	500	1,000	10,000
Univariate BBSDs	s_gn	0.00	0.00	0.03	0.03	0.07	0.10	0.10	0.10
	m_gn	0.00	0.00	0.10	0.13	0.13	0.13	0.23	0.37
	l_gn	0.17	0.27	0.53	0.63	0.67	0.83	0.87	1.00
	s_img	0.00	0.00	0.23	0.30	0.40	0.63	0.70	0.93
	m_img	0.30	0.37	0.60	0.67	0.70	0.80	0.90	1.00
	l_img	0.30	0.50	0.70	0.70	0.77	0.87	0.97	1.00
	adv	0.13	0.27	0.40	0.43	0.53	0.77	0.83	0.90
	ko	0.00	0.00	0.07	0.07	0.07	0.33	0.40	0.70
	m_img+ko	0.13	0.40	0.87	0.93	0.90	1.00	1.00	1.00
	oz+m_img	0.67	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Multivariate UAE	s_gn	0.03	0.03	0.03	0.03	0.03	0.07	0.07	-
	m_gn	0.03	0.03	0.03	0.03	0.17	0.27	0.30	-
	l_gn	0.50	0.57	0.67	0.70	0.80	0.90	1.00	-
	s_img	0.17	0.20	0.27	0.30	0.40	0.47	0.63	-
	m_img	0.23	0.33	0.37	0.40	0.47	0.60	0.70	-
	l_img	0.30	0.30	0.37	0.47	0.60	0.77	0.87	-
	adv	0.03	0.20	0.27	0.27	0.33	0.40	0.40	-
	ko	0.10	0.13	0.13	0.13	0.17	0.17	0.30	-
	m_img+ko	0.20	0.30	0.37	0.53	0.54	0.63	0.87	-
	oz+m_img	0.27	0.63	0.77	1.00	1.00	1.00	1.00	-

