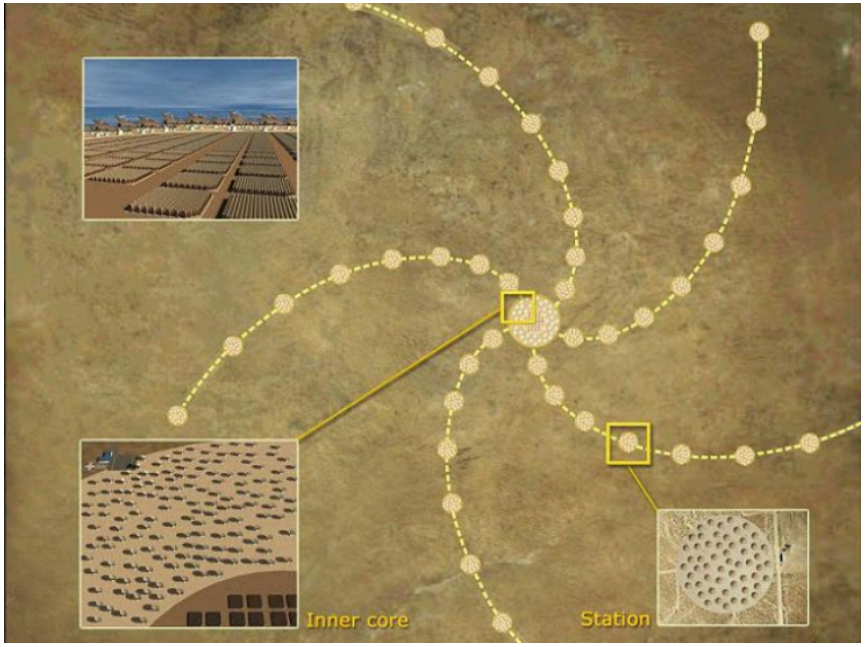
Information-theoretic astro-imaging and data analysis

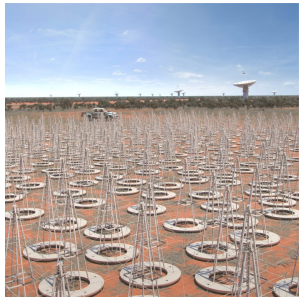
Slava Voloshynovskiy, Olga Taran and Daniel Schaerer

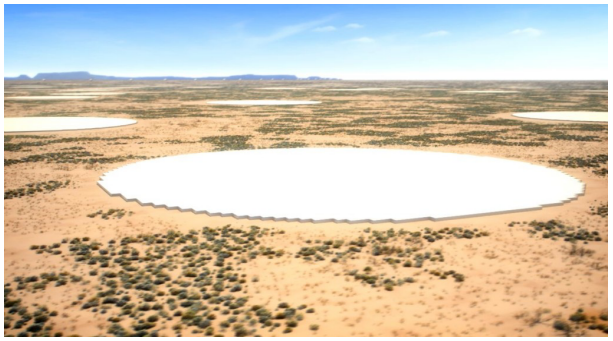
2021, Swiss SKA Days

UNIVERSITÉ DE GENÈVE

Square Kilometer Array



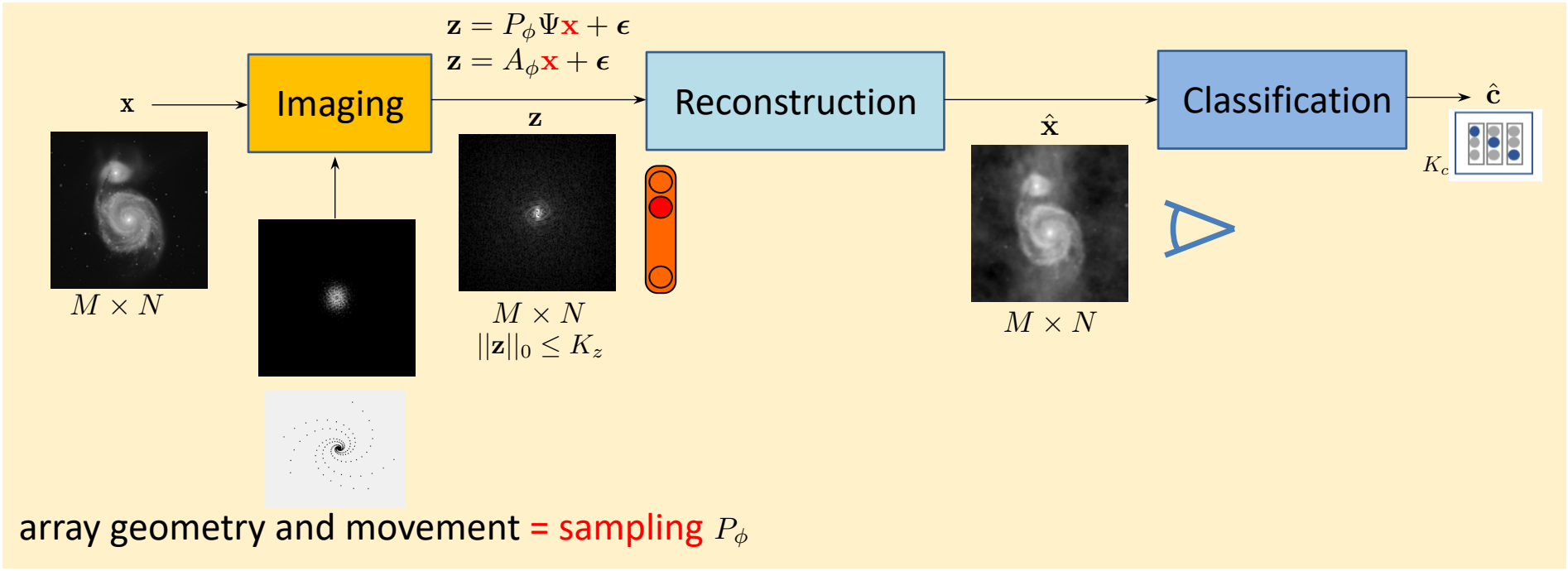




Square Kilometer Array (SKA) – imaging tool of the 21st century:

- Huge amount of data is expected
- Problems with Big Data:
  - Data communication, storage
  - Image reconstruction
  - Image analysis: classification, prediction, estimation, etc.

Traditional imaging-classification pipeline



Traditional pipeline

The image $\mathbf{x} \in \mathcal{X}^{M \times N}$ is reconstructed from the uv-observation $\mathbf{z} \in \mathcal{Z}^{K_z}$ to $\hat{\mathbf{x}} \in \mathcal{X}^{M \times N}$ and then the classification is performed based on $\hat{\mathbf{x}} \in \mathcal{X}^{M \times N}$ to produce $\hat{\mathbf{c}} \in \{0, 1, \cdots, K_c\}$

Note

The entropy of image is significsantly higher than the entropy of parameters to be estimated or labels, i.e., $H(\mathbf{X}) \geq H(\mathbf{C})$

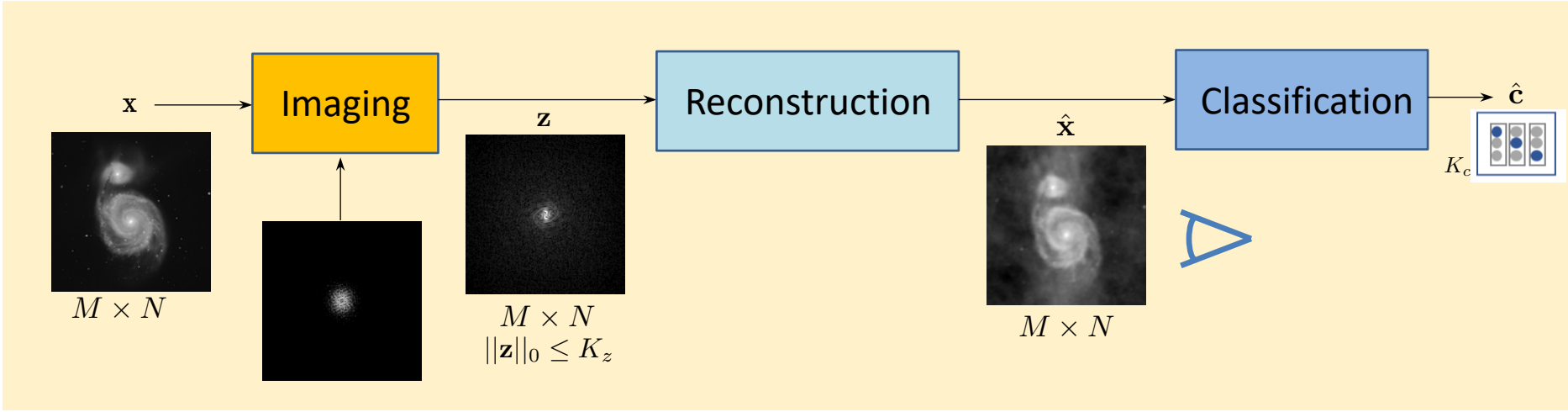
Traditional imaging-classification pipeline

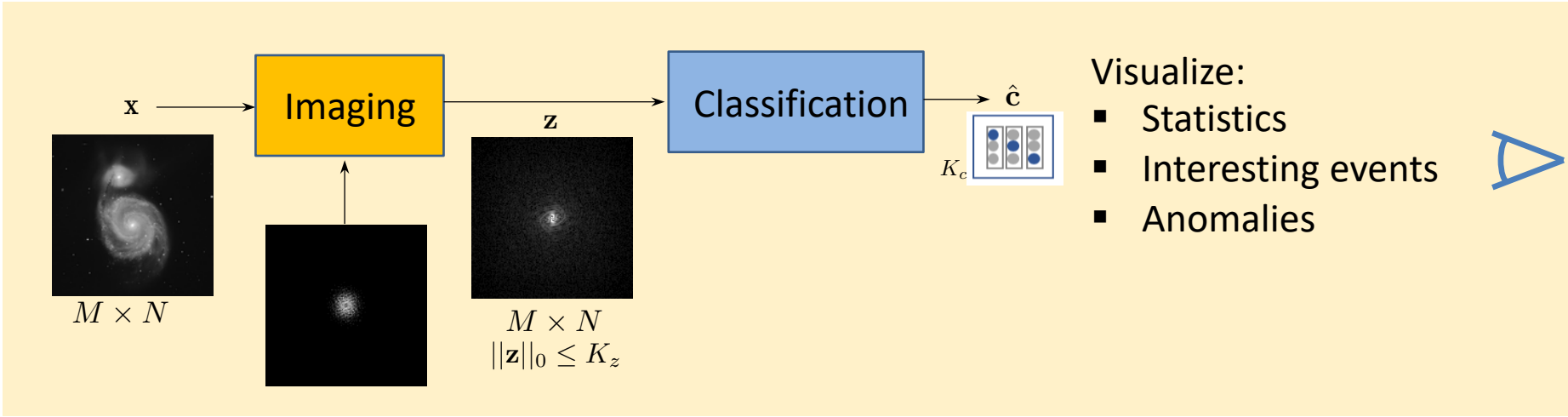
- $M \times N$ might be large and generally $K_z \ll M \times N$
- The inverse problem $\mathbf{z} \longrightarrow \hat{\mathbf{x}}$ is ill-posed and requires either "hand-crafted" or data-dependent (ML) priors and regularizers
- The overall problem complexity is very high
- The reconstructed data require huge storage and communication

Q: do we really need to reconstruct $\hat{\mathbf{x}}$ in all cases ?

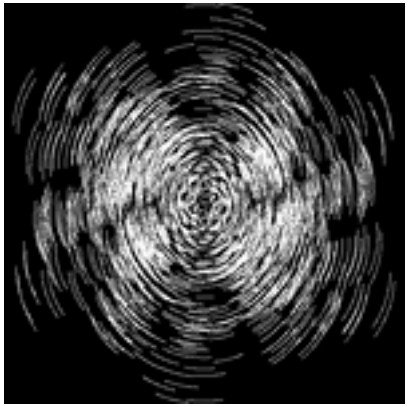
Proposed solution

- Skip the reconstruction stage (if fast decision or statistics are needed for the massive amount of data) by **proceeding with the classification task directly**

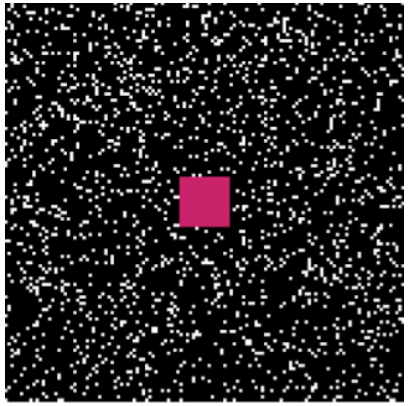




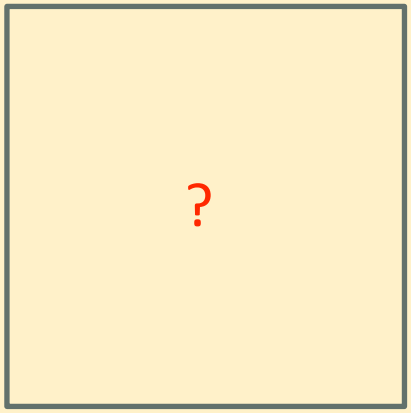
Optimal sensor-sampling



Sparse array with Earth rotation



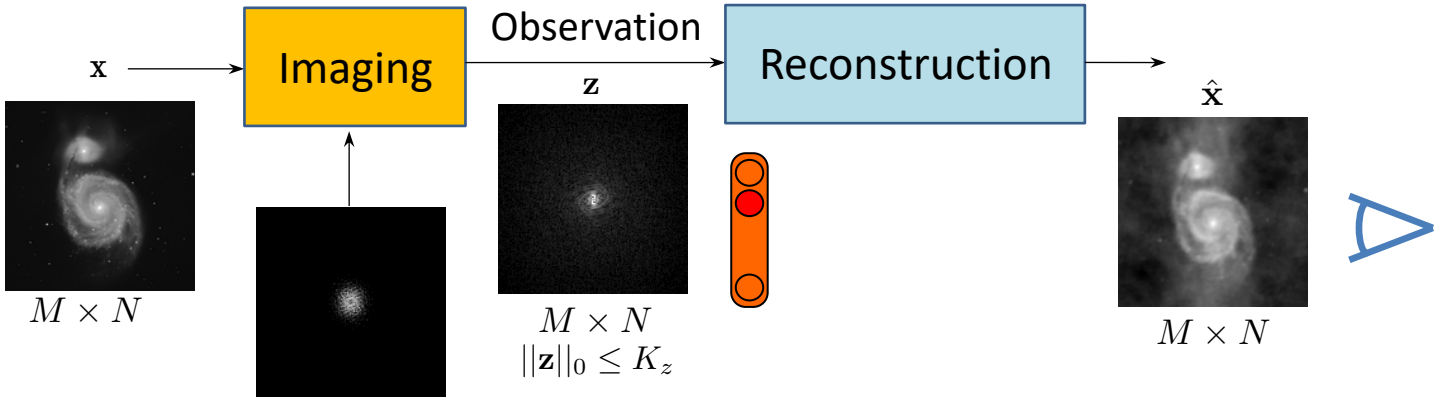
Compressive Sensing (CS) sampling

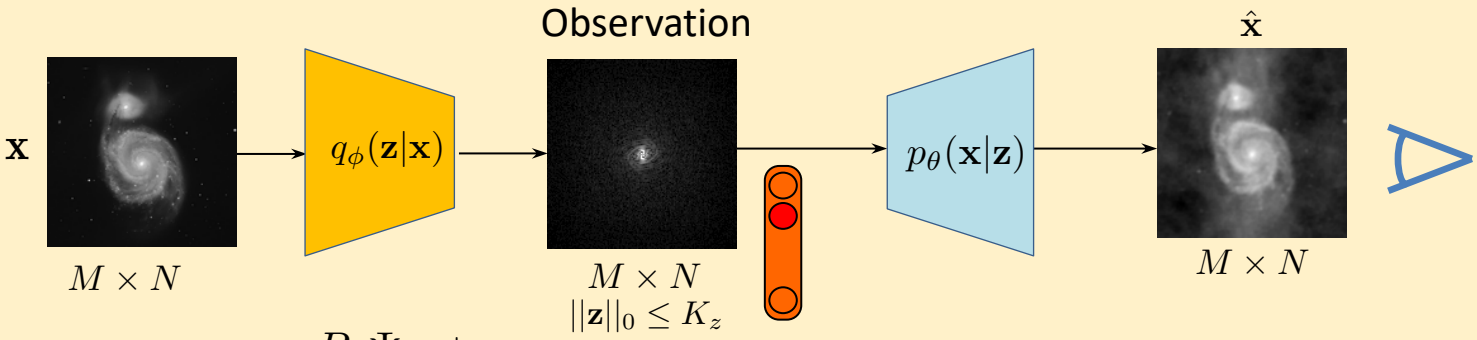


Optimal sensing for:
(a) Reconstruction
(b) Classification

Proposed solution

- Consider imaging/classification from the information-theoretic point of view

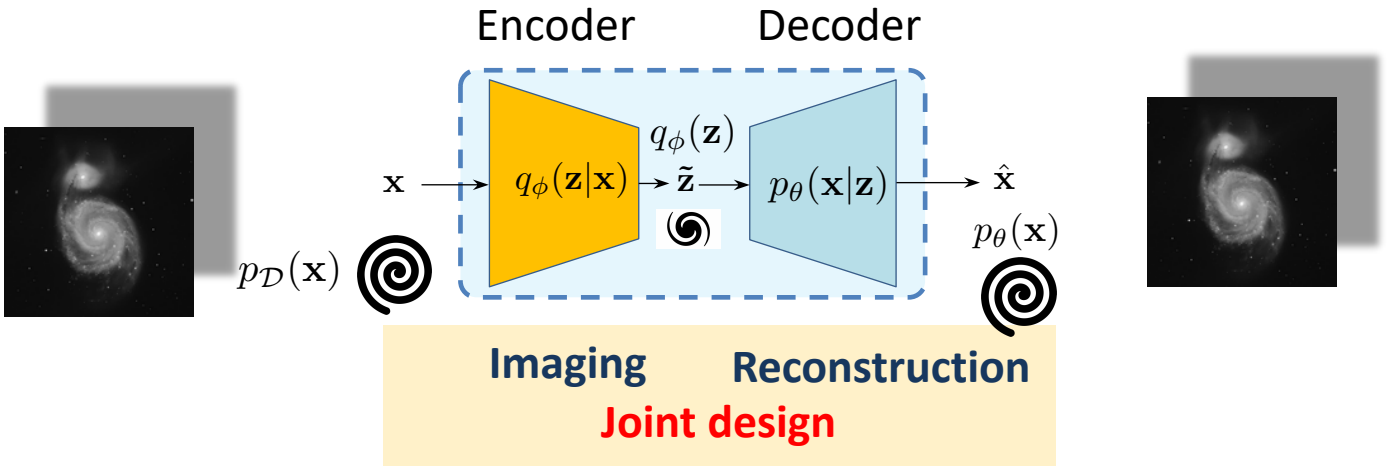




$$\mathbf{z} = P_\phi \Psi \mathbf{x} + \boldsymbol{\epsilon}$$
$$\mathbf{z} = A_\phi \mathbf{x} + \boldsymbol{\epsilon}$$

Latent space

Information-theoretic formulation: information bottleneck



Lagrangian formulation

$$(\hat{\phi}, \hat{\theta}) = \arg\min_{\phi,\theta} \mathcal{L}_{\text{BIB-AE}}(\phi, \theta)$$

$$\mathcal{L}_{\text{BIB-AE}}(\phi, \theta) = I_\phi(\mathbf{X};\mathbf{Z}) - \beta I_{\phi,\theta}(\mathbf{Z};\mathbf{X})$$

$$I_\phi(\mathbf{X};\mathbf{Z}) = \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z})}\right] \qquad I_{\theta,\phi}(\mathbf{Z};\mathbf{X}) = \mathbb{E}_{p_\mathcal{D}(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x}|\mathbf{z})}{p_\mathcal{D}(\mathbf{x})}\right]\right]$$

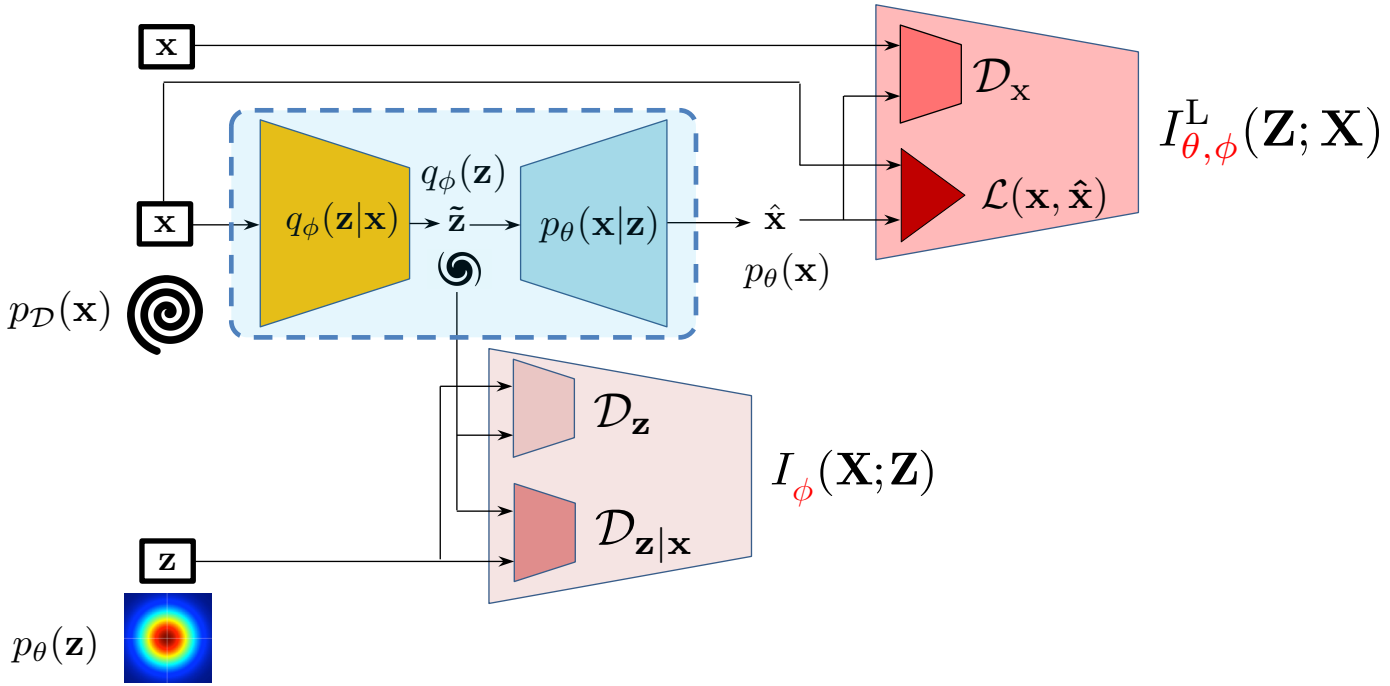
Voloshynovskiy, Kondah, Rezaeifar, Taran, Hotolyak, Rezende, Information bottleneck through variational glasses, 2019.

Bounded Information Bottleneck (BIB) Autoencoder [BIB-AE]

$$\mathcal{L}_{\text{BIB-AE}}(\phi,\theta) = I_{\phi}(\mathbf{X};\mathbf{Z}) - \beta I_{\phi,\theta}(\mathbf{Z};\mathbf{X})$$

$$I_{\phi}(\mathbf{X};\mathbf{Z}) = \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z})}\frac{p_\theta(\mathbf{z})}{p_\theta(\mathbf{z})}\right] = \underbrace{\mathbb{E}_{p_\mathcal{D}(\mathbf{x})}\left[D_{\text{KL}}\left(q_\phi(\mathbf{z}\,|\,\mathbf{X}=\mathbf{x}) \parallel p_\theta(\mathbf{z})\right)\right]}_{\mathcal{D}_{\mathbf{z}|\mathbf{x}}} - \underbrace{D_{\text{KL}}\left(q_\phi(\mathbf{z}) \parallel p_\theta(\mathbf{z})\right)}_{\mathcal{D}_{\mathbf{z}}}$$

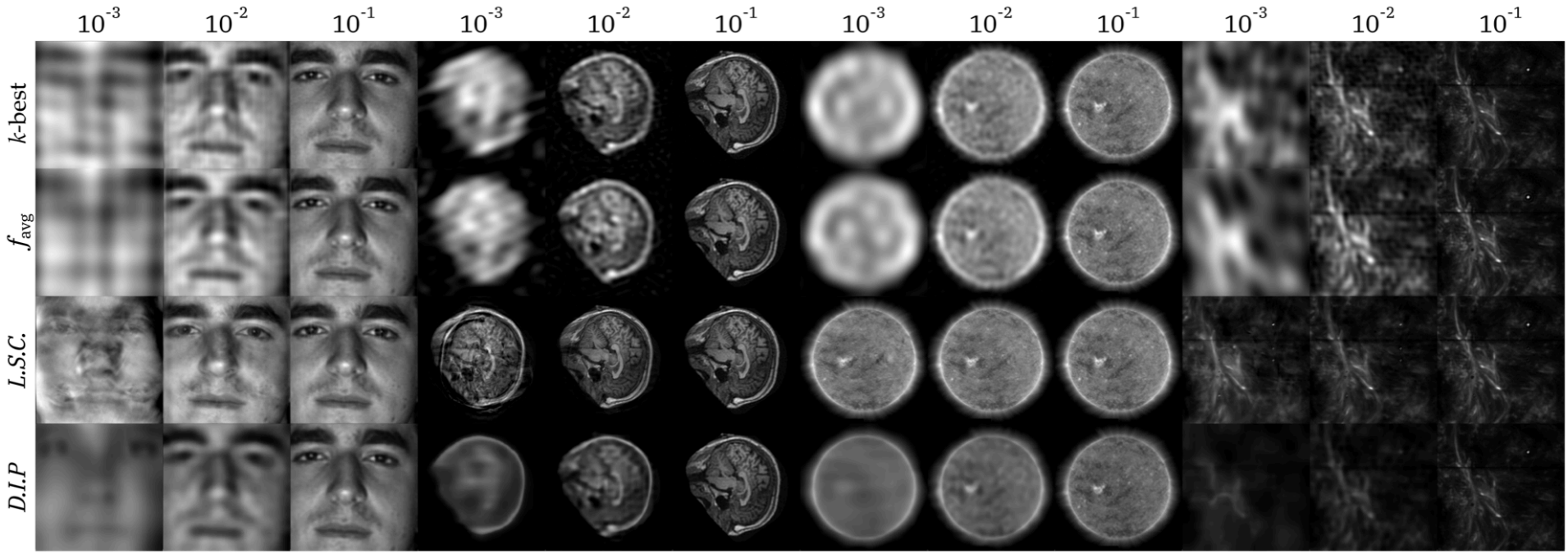
$$I^{\text{L}}_{\theta,\phi}(\mathbf{Z};\mathbf{X}) \triangleq \underbrace{-H_{\theta,\phi}(\mathbf{X}|\mathbf{Z})}_{\mathcal{L}(\mathbf{x},\hat{\mathbf{x}})} - \underbrace{D_{\text{KL}}\left(p_\mathcal{D}(\mathbf{x})\|p_\theta(\mathbf{x})\right)}_{\mathcal{D}_{\mathbf{x}}}$$



Voloshynovskiy, Kondah, Rezaeifar, Taran, Hotolyak, Rezende, Information bottleneck through variational glasses, 2019.

Challenge 2: Optimal sensing for reconstruction $\mathbf{x} \mapsto \hat{\mathbf{x}}$

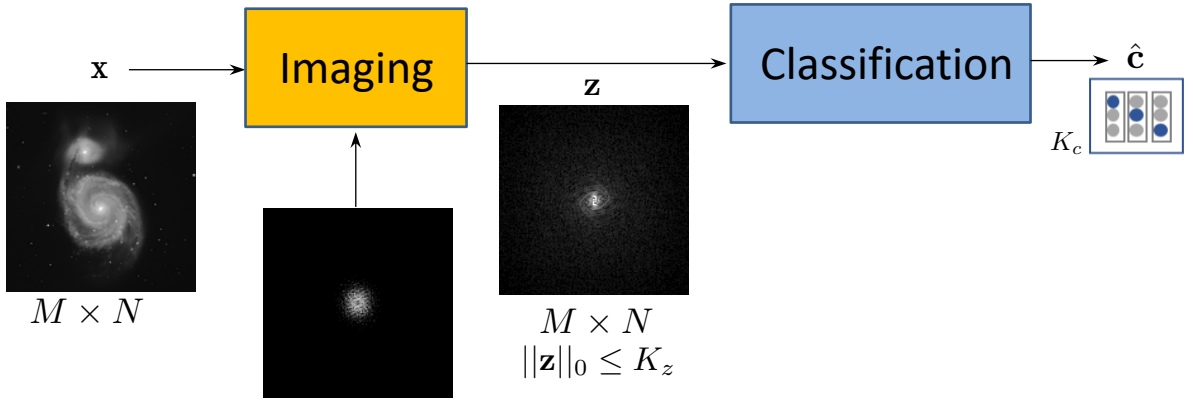
Reconstruction from undersampled data



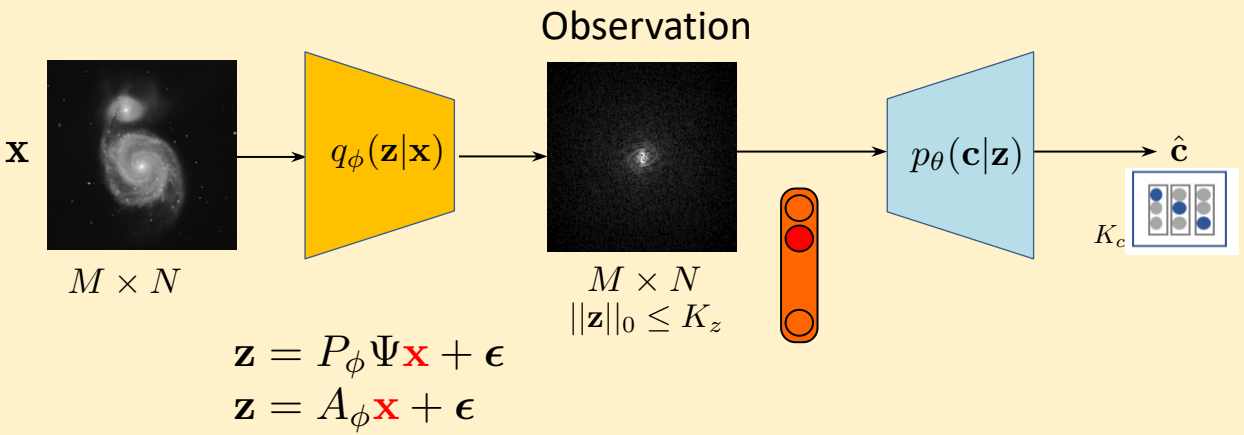
M. Ferrari, O. Taran, T. Holotyak, K. Egiazarian, and S. Voloshynovskiy, "Injecting Image Priors into Learnable Compressive Subsampling," in *Proc. 26th European Signal Processing Conference (EUSIPCO)*, Rome, Italy, 2018.

Proposed solution

- Consider imaging/classification from the information-theoretic point of view

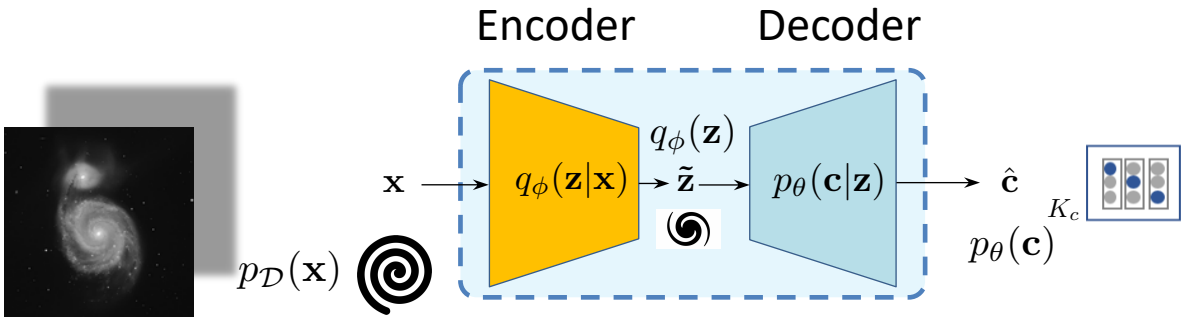


Visualize:
- Statistics
- Interesting events
- Anomalies



Visualize:
- Statistics
- Interesting events
- Anomalies

Information-theoretic formulation: information bottleneck



Imaging **Classification**

Joint designn

Lagrangian formulation

$$(\hat{\phi}, \hat{\theta}) = \arg\min_{\phi,\theta} \mathcal{L}_{\text{BIB-C}}(\phi, \theta)$$

$$\mathcal{L}_{\text{BIB-C}}(\phi, \theta) = I_\phi(\mathbf{X};\mathbf{Z}) - \beta I_{\phi,\theta}(\mathbf{Z};\mathbf{C})$$

$$I_\phi(\mathbf{X};\mathbf{Z}) = \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z})}\right] \qquad I_{\theta,\phi}(\mathbf{Z};\mathbf{C}) = \mathbb{E}_{p(\mathbf{c},\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{c}|\mathbf{z})}{p_\mathbf{c}(\mathbf{c})}\right]\right]$$

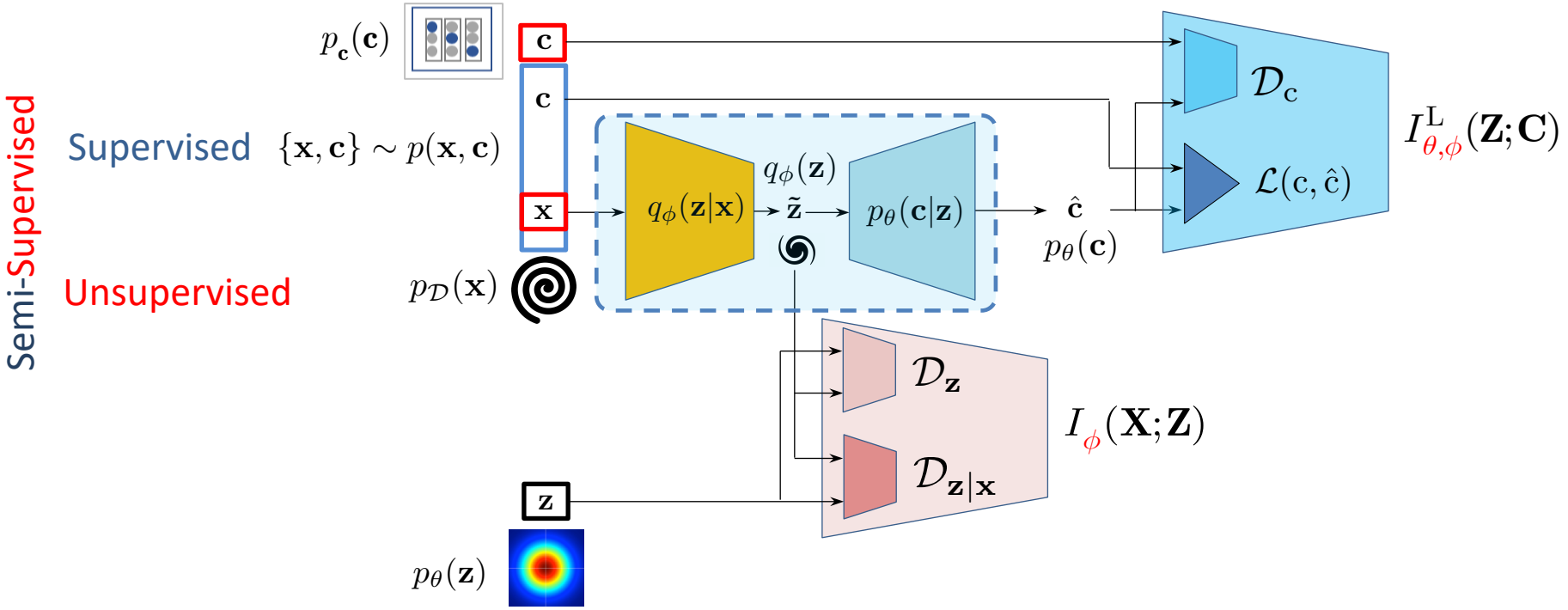
Voloshynovskiy, Kondah, Rezaeifar, Taran, Hotolyak, Rezende, Information bottleneck through variational glasses, 2019.

Bounded Information Bottleneck (BIB) Classification [BIB-C]

$$\mathcal{L}_{\text{BIB-C}}(\phi,\theta) = I_{\phi}(\mathbf{X};\mathbf{Z}) - \beta I^{\text{L}}_{\theta,\phi}(\mathbf{Z};\mathbf{C})$$

$$I_{\phi}(\mathbf{X};\mathbf{Z}) = \underbrace{\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[D_{\text{KL}}\left(q_{\phi}(\mathbf{z} \mid \mathbf{X}=\mathbf{x}) \parallel p_{\theta}(\mathbf{z}) \right)\right]}_{\mathcal{D}_{\mathbf{z}|\mathbf{x}}} - \underbrace{D_{\text{KL}}\left(q_{\phi}(\mathbf{z}) \parallel p_{\theta}(\mathbf{z}) \right)}_{\mathcal{D}_{\mathbf{z}}}$$

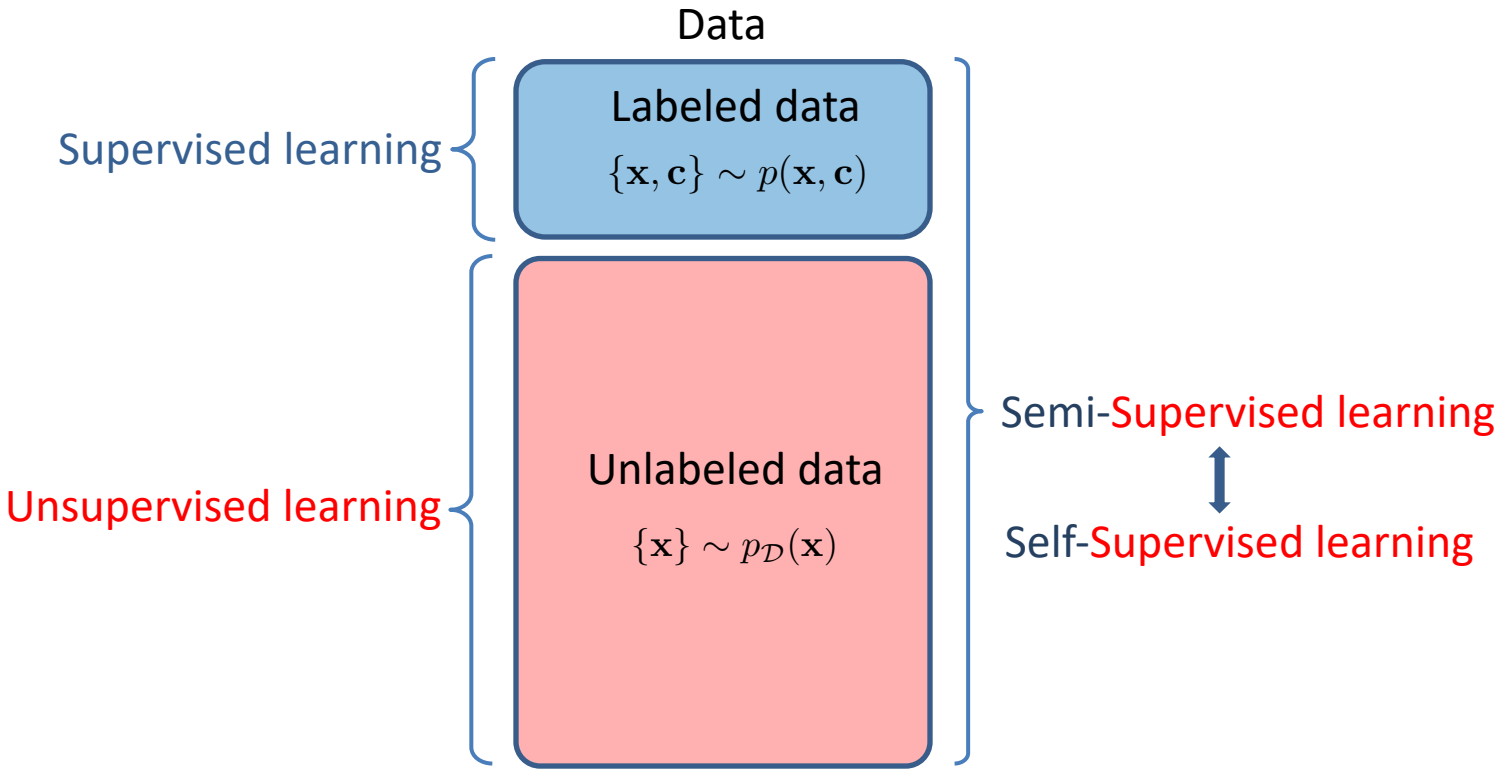
$$I^{\text{L}}_{\theta,\phi}(\mathbf{Z};\mathbf{C}) \triangleq \underbrace{-H_{\theta,\phi}(\mathbf{C} \mid \mathbf{Z})}_{\mathcal{L}(\text{c},\hat{\text{c}})} \underbrace{- D_{\text{KL}}\left(p_{\mathbf{c}}(\mathbf{c}) \| p_{\theta}(\mathbf{c})\right)}_{\mathcal{D}_{\text{c}}}$$



Conclusions

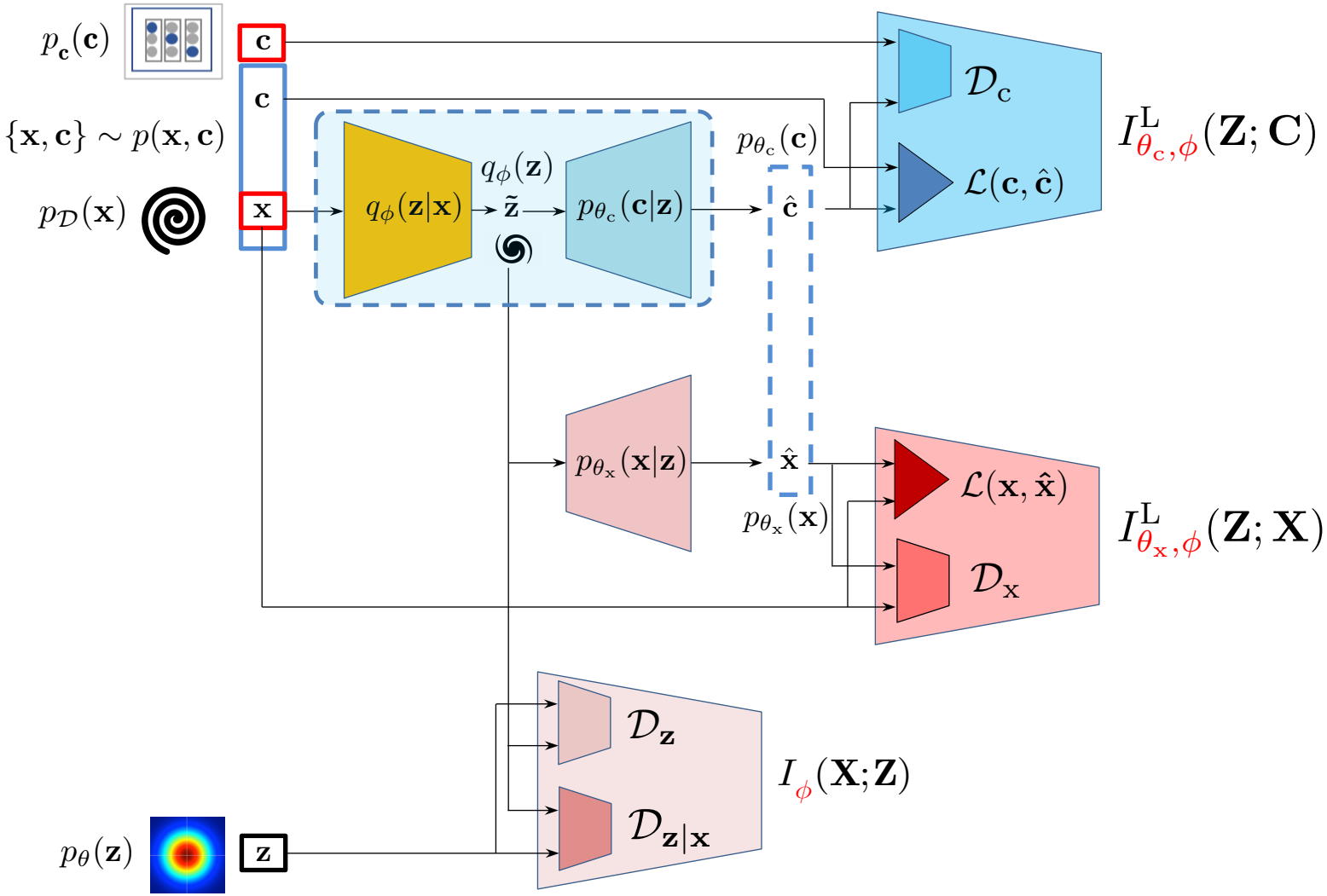
- Joint design of:
  - Sampling-reconstruction operators
  - Sampling-classification operators
- Optimal sampling for reconstruction and classification might be different
- The amount of needed samples for the classification might be significantly smaller than for the reconstruction

Traditional classification pipelines



Information Bottleneck: classification $\mathbf{x} \rightarrow \hat{\mathbf{c}}$

Learnable regularization – joint classification and reconstruction

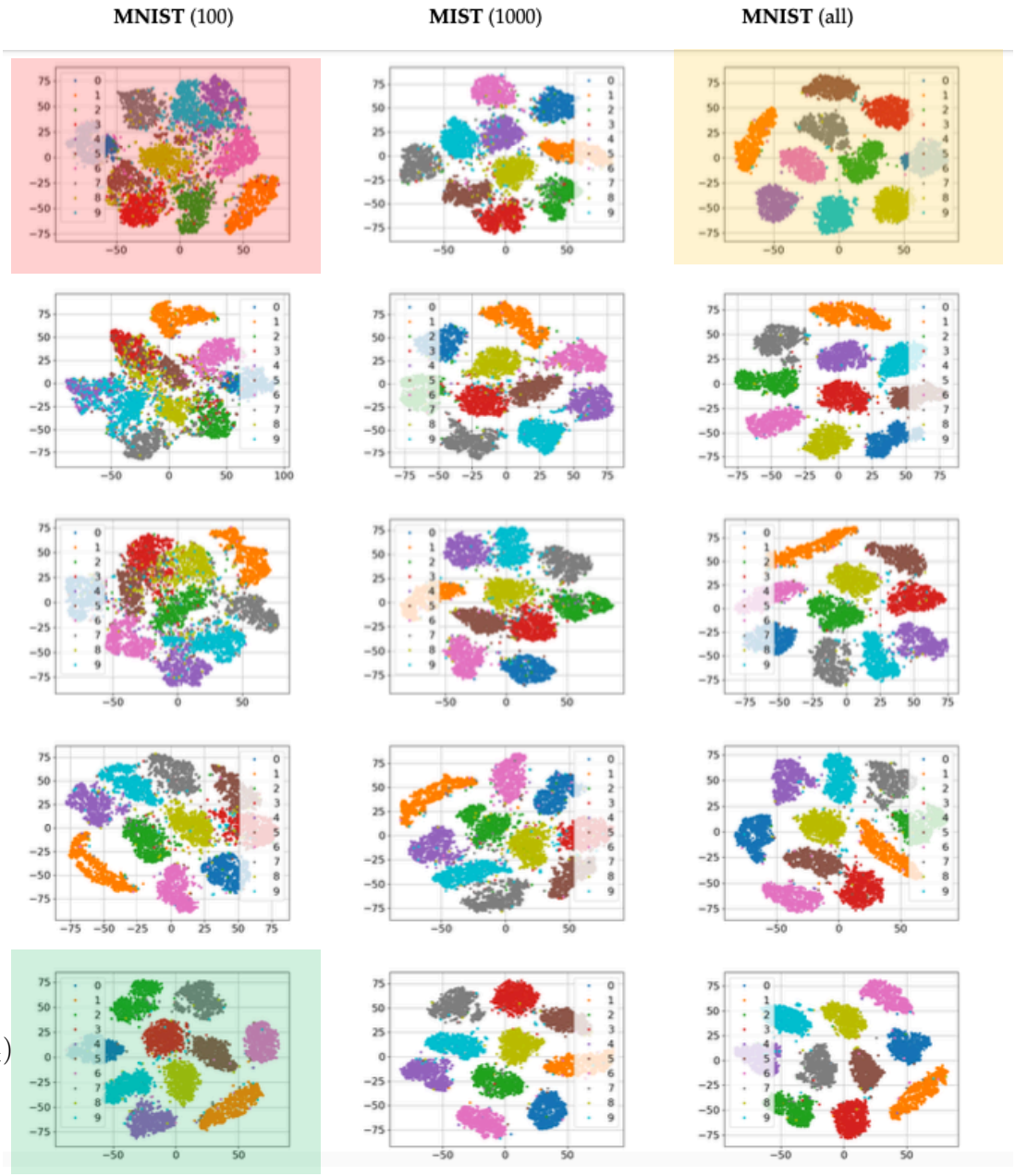


Information Bottleneck: classification $\mathbf{x} \rightarrow \hat{\mathbf{c}}$

Table 1. Semi-supervised classification performance (percentage error) for the optimal parameters (Appendices B–G) defined on the MNIST (*D*—deterministic; *S*—stochastic).

		MNIST (100)	MIST (1000)	MNIST (all)	SVHN (1000)
NN Baseline $\mathcal{L}(\mathbf{c}, \hat{\mathbf{c}})$	[*D*]	26.31 (±0.91)	7.50 (±0.19)	0.68 (±0.05)	36.16 (±0.77)
	[*S*]	26.78 (±1.66)	7.54 (±0.25)	0.70 (±0.05)	36.28 (±0.93)
InfoMax [3]	[*S*]	33.41	21.5	15.86	-
VAE [5]	[*S*]	14.26	8.71	5.02	-
MV-InfoMax [5]	[*S*]	13.22	7.39	6.07	-
IB multiview [5]	[*S*]	3.03	2.34	2.22	-
VAE (M1 + M2) [5]	[*S*]	3.33 (±0.14)	2.40 (±0.02)	0.96	36.02 (±0.10)
CatGAN	[*S*]	1.91 (±0.10)	1.73 (±0.18)	0.91	-
AAE	[*D*]	1.90 (±0.10)	1.60 (±0.08)	0.85 (±0.02)	17.70 (±0.30)
No priors on latent space					
$\mathcal{L}(\mathbf{c}, \hat{\mathbf{c}}) + \mathcal{D}_\mathrm{c}$	[*D*]	20.72 (±1.58)	4.99 (±0.28)	0.69 (±0.04)	25.78 (±0.90)
	[*S*]	19.60 (±1.37)	4.49 (±0.25)	0.67 (±0.05)	26.34 (±0.80)
Hand crafted latent space priors					
$\beta_\mathrm{c}\mathcal{L}(\mathbf{c}, \hat{\mathbf{c}}) + \mathcal{D}_\mathrm{z}$	[*D*]	27.44 (±1.40)	6.77 (±0.34)	0.91 (±0.05)	35.94 (±1.08)
	[*S*]	27.48 (±1.07)	6.91 (±0.45)	0.88 (±0.05)	35.80 (±1.21)
$\beta_\mathrm{c}(\mathcal{L}(\mathbf{c}, \hat{\mathbf{c}}) + \mathcal{D}_\mathrm{c}) + \mathcal{D}_\mathrm{z}$	[*D*]	12.04 (±4.46)	2.43 (±0.12)	0.81 (±0.05)	24.70 (±0.46)
	[*S*]	11.80 (±3.82)	2.40 (±0.10)	0.82 (±0.04)	24.62 (±0.54)
Learnable latent space priors					
$\beta_\mathrm{c}(\mathcal{L}(\mathbf{c}, \hat{\mathbf{c}}) + \mathcal{D}_\mathrm{c}) + \mathcal{D}_\mathrm{z} + \beta_\mathrm{x}\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}})$	[*D*]	1.55 (±0.21)	1.25 (±0.10)	0.74 (±0.04)	20.07 (±0.36)
	[*S*]	1.49 (±0.18)	1.43 (±0.06)	0.78 (±0.04)	20.00 (±0.31)
$\beta_\mathrm{c}(\mathcal{L}(\mathbf{c}, \hat{\mathbf{c}}) + \mathcal{D}_\mathrm{c}) + \mathcal{D}_\mathrm{z} + \beta_\mathrm{x}(\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) + \mathcal{D}_\mathrm{x})$	[*D*]	1.38 (±0.09)	1.21 (±0.10)	0.77 (±0.06)	19.75 (±0.52)
	[*S*]	1.42 (±0.10)	1.16 (±0.09)	0.79 (±0.02)	19.71 (±0.26)

Information Bottleneck: classification $\mathbf{x} \rightarrow \hat{\mathbf{c}}$



Conclusions

- We considered information-theoretic formulation of imaging and data analysis based on the classification task
- IBN is a useful tool for the analysis and generalization of existing schemes
- **Not covered problems:**
  - Data compression
  - Domain-to-domain translation
  - Prediction

Thank you!