From Supervised to Unsupervised Computational Sensing

Ali Mousavi

Aug 12th 2019

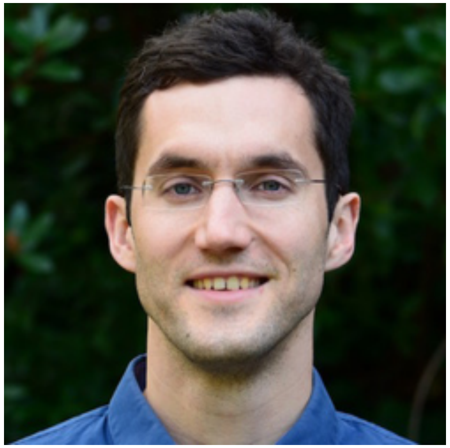
Google Brain

Collaborators

Rich Baraniuk
Rice University

Arian Maleki
Columbia University

Chris Metzler
Stanford University

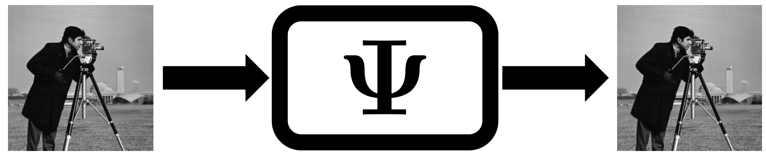


Reinhard Heckel
Rice University

Gautam Dasarathy
Arizona StateUniversity

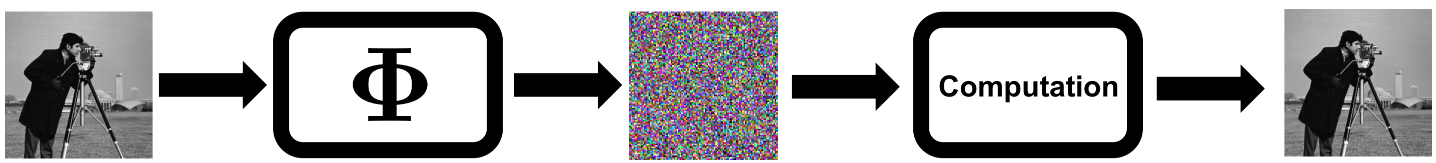
Computational Sensing

- **Conventional Sensing**

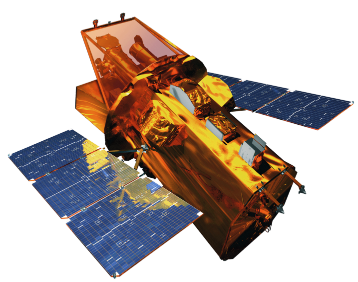


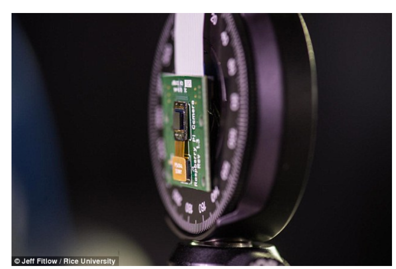
Subject   Expensive Hardware

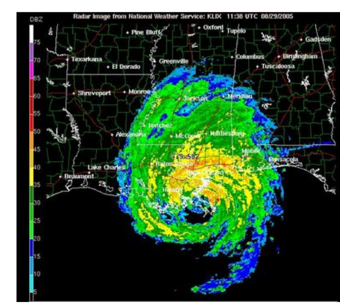
- **Computational Sensing:** **Reduce costs** in acquisition systems by replacing expensive hardware w/ **cheap hardware + computation**



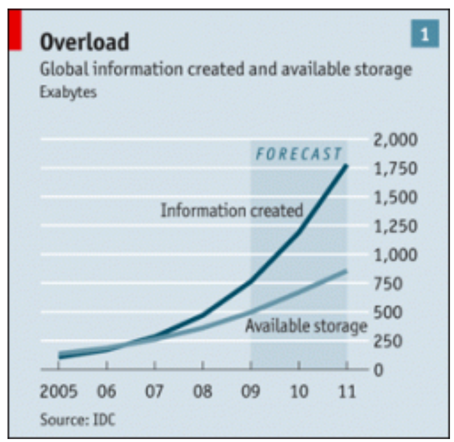
Subject   Simpler Hardware   Measurements   Software

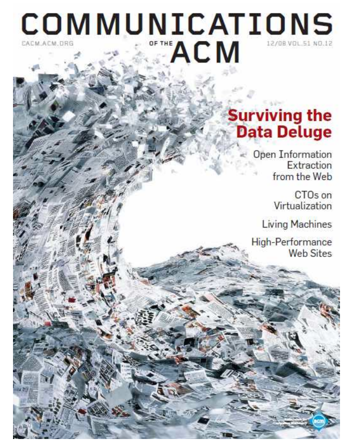




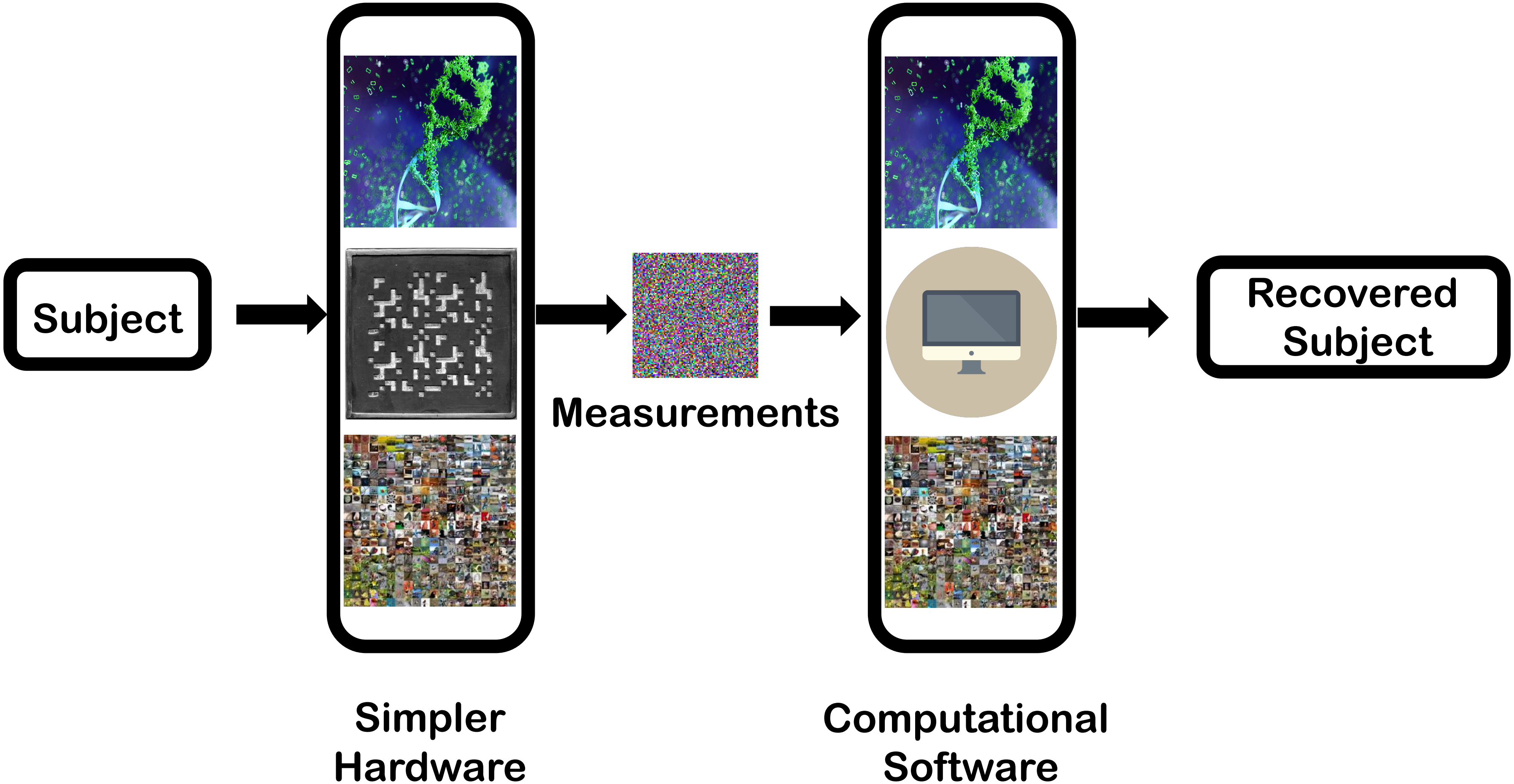


Large Scale Datasets

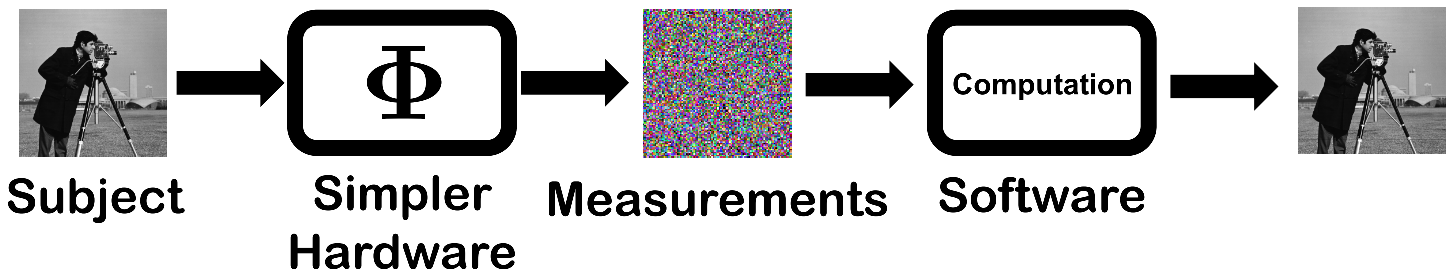




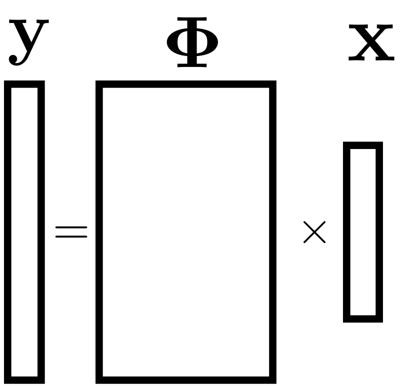
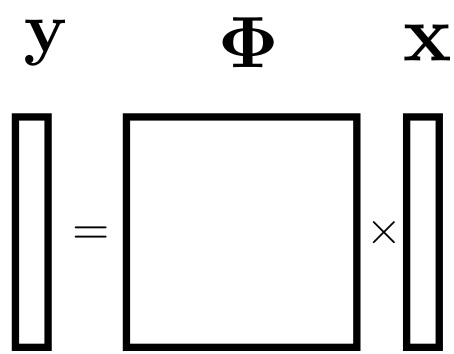
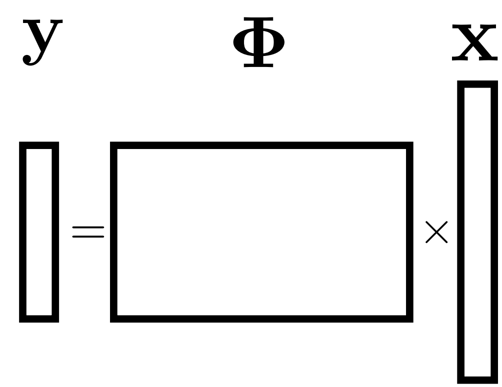
Data-Driven Computational Sensing



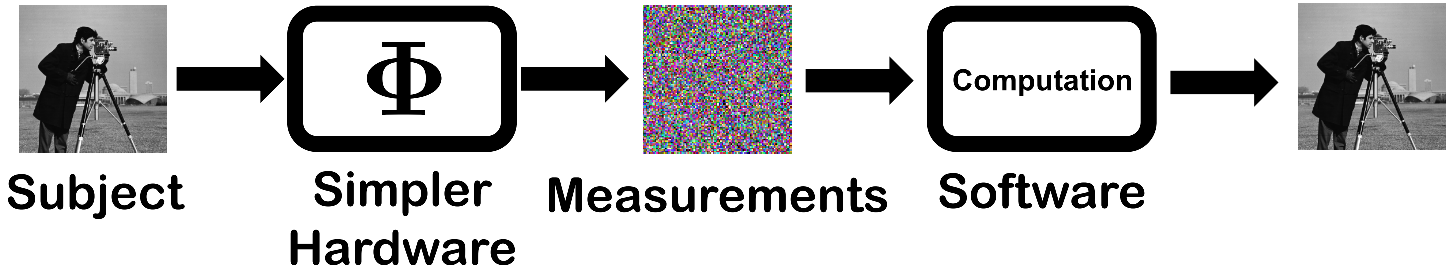
Model



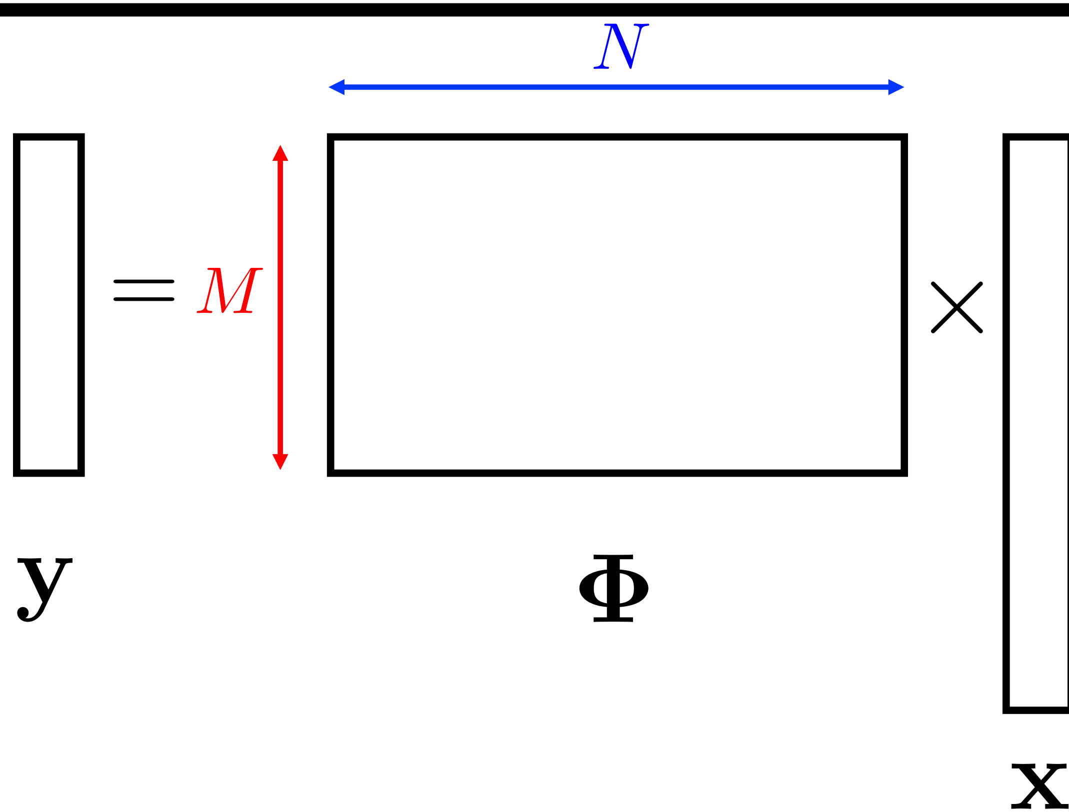
$$\mathbf{x} \in \mathbb{R}^N \xrightarrow{\mathbf{\Phi}(.)} \mathbf{y} = \mathbf{\Phi}(\mathbf{x}) \in \mathbb{R}^M \xrightarrow{\mathbf{\Phi}^{-1}(.)} \hat{\mathbf{x}} \in \mathbb{R}^N$$

Overdetermined	Determined	Underdetermined
 $M > N$	 $M = N$	 $M < N$

Model

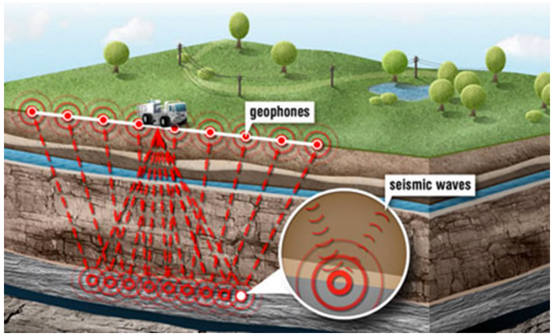


$$\mathbf{x} \in \mathbb{R}^N \xrightarrow{\boldsymbol{\Phi}(.)} \mathbf{y} = \boldsymbol{\Phi}(\mathbf{x}) \in \mathbb{R}^M \xrightarrow{\boldsymbol{\Phi}^{-1}(.)} \hat{\mathbf{x}} \in \mathbb{R}^N$$

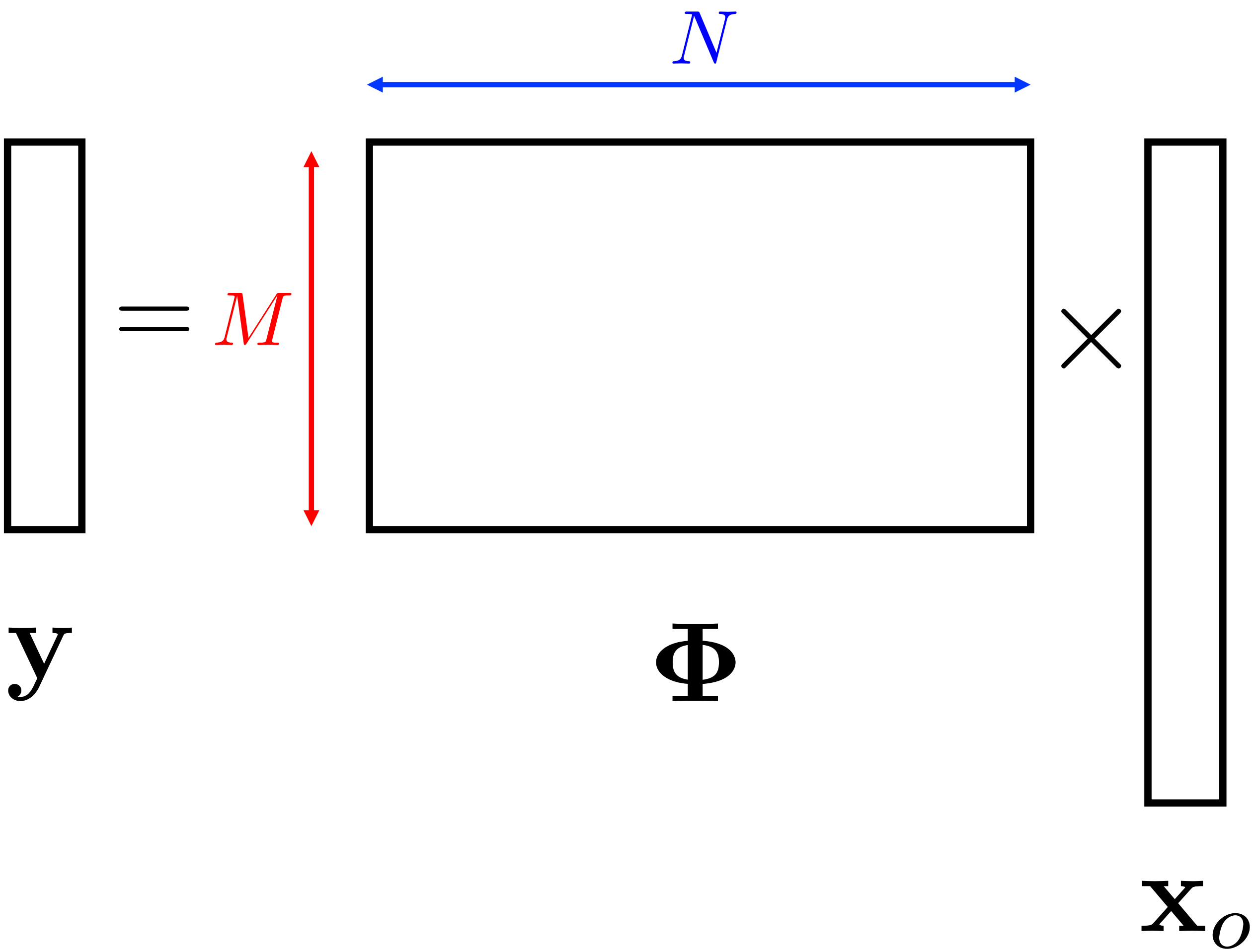


Applications



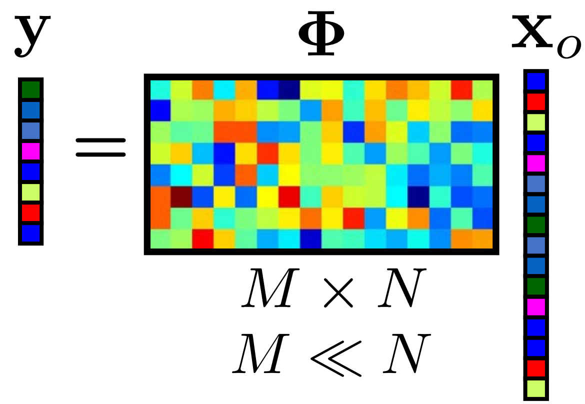



Data-Driven Computational Sensing

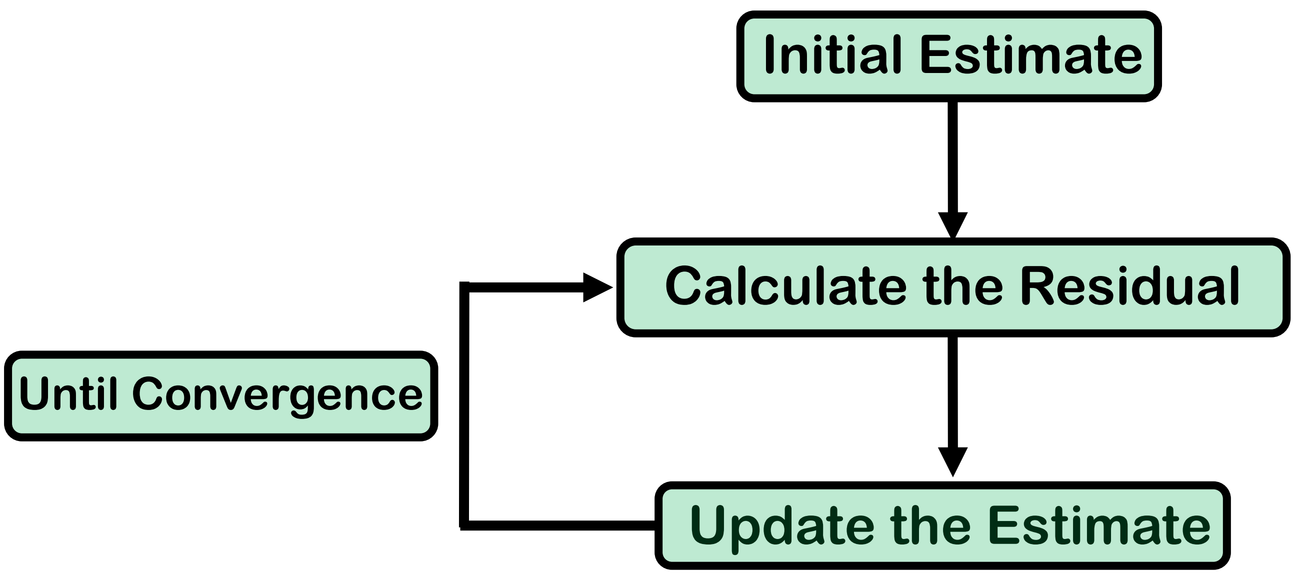


$$\min_{\mathbf{x}} \|\mathbf{y} - \boldsymbol{\Phi}\mathbf{x}\|_2^2 + \lambda \times f(\mathbf{x})$$

Iterative Algorithms

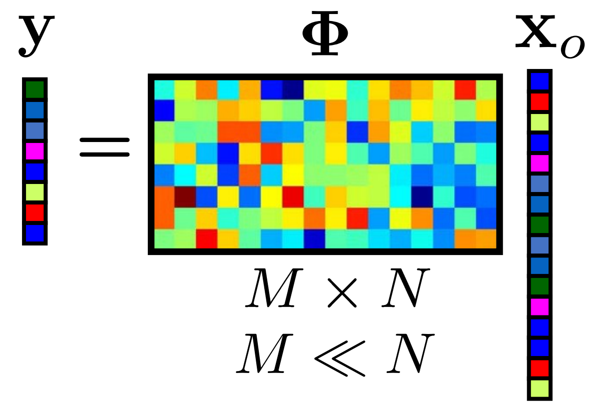


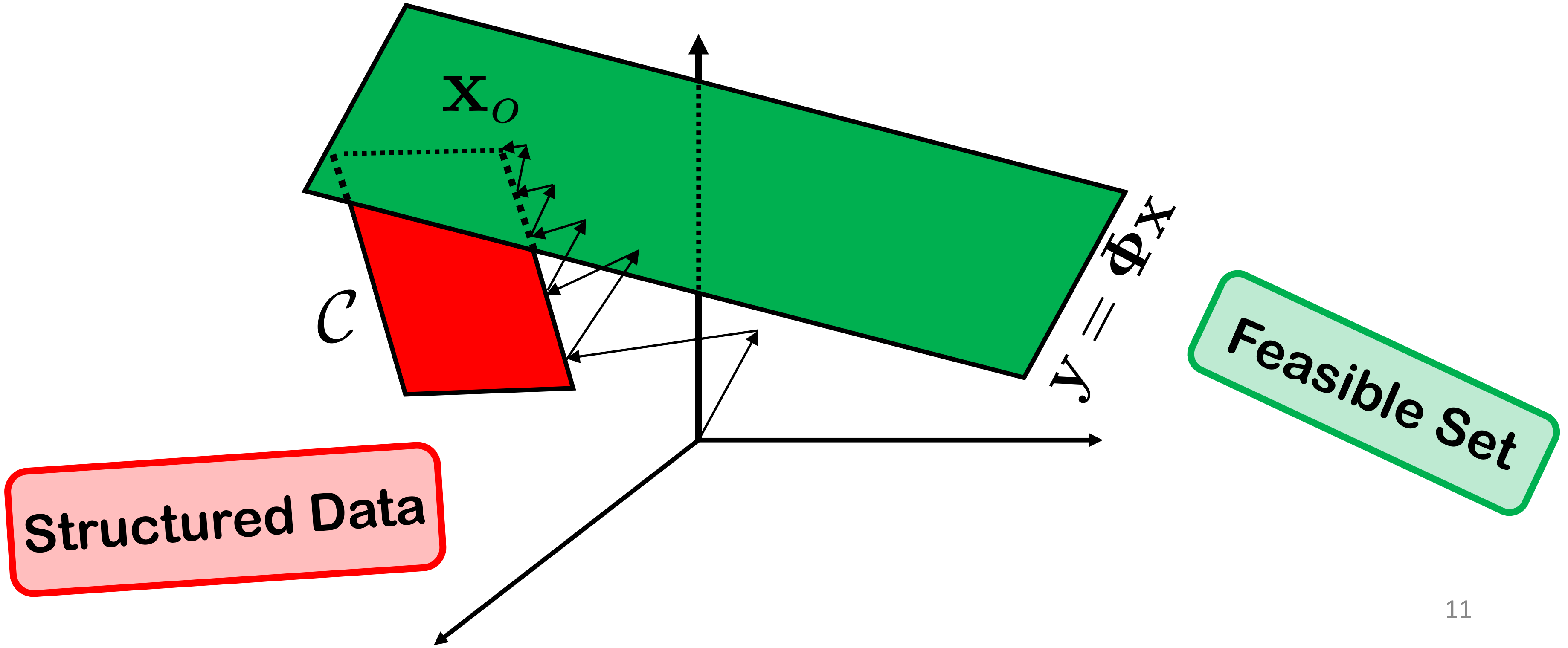
$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi x}\|_2^2 + \lambda \times f(\mathbf{x})$$



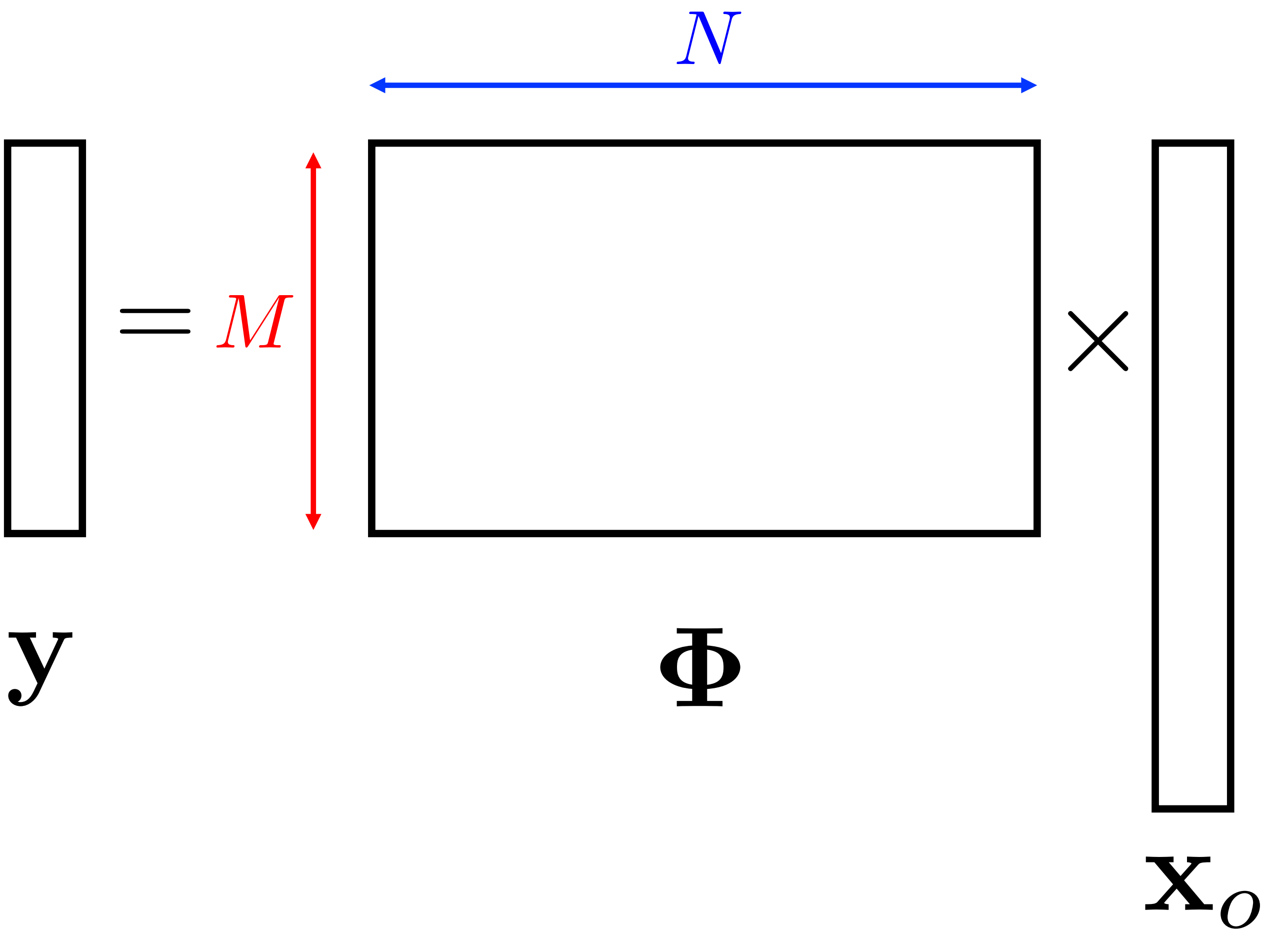
Iterative Algorithms

$$\min_{\mathbf{x}} \|\mathbf{y} - \boldsymbol{\Phi}\mathbf{x}\|_2^2 + \lambda \times f(\mathbf{x})$$



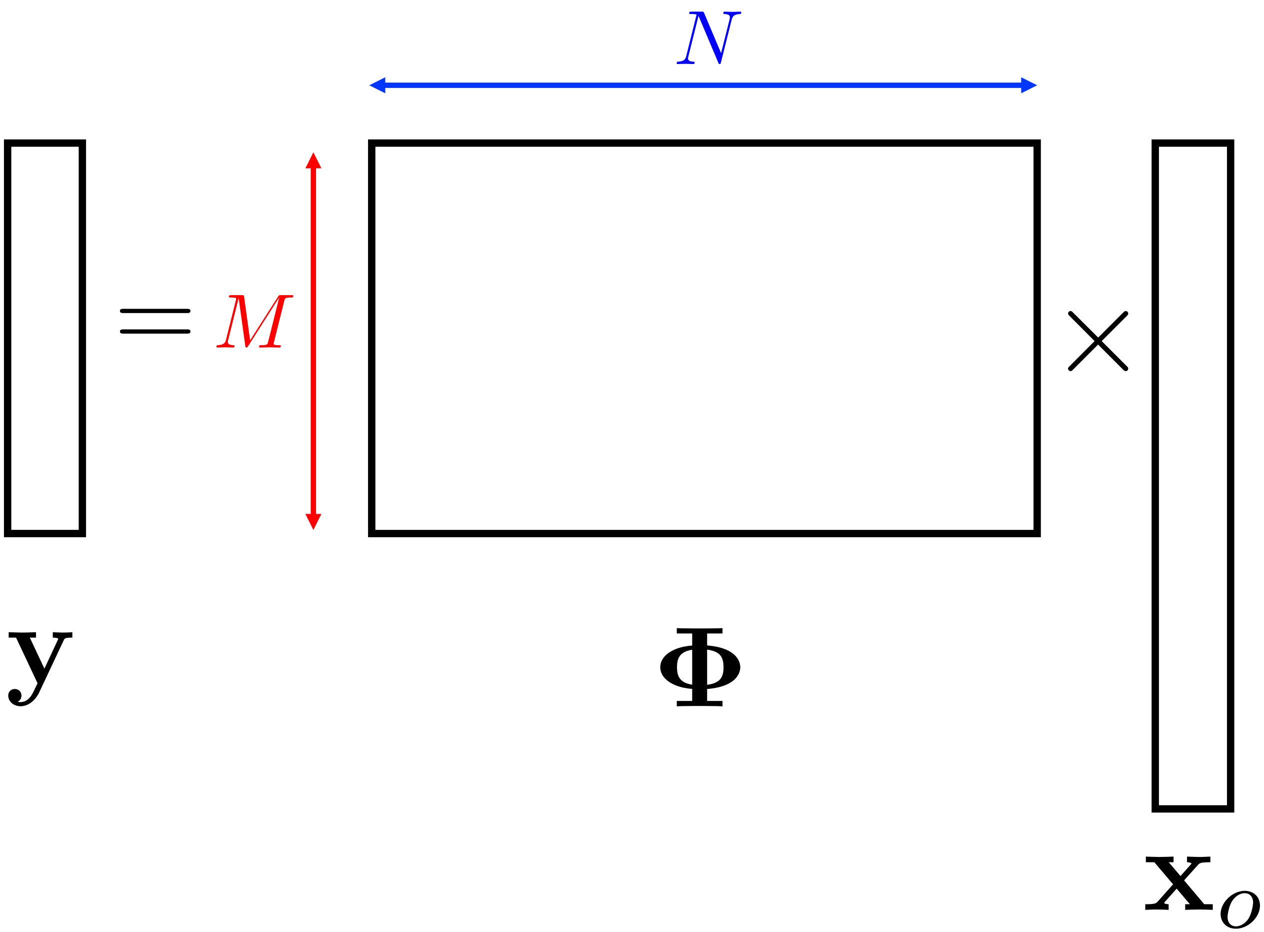


Data-Driven Computational Sensing



$$\min_{\mathbf{x}} \|\mathbf{y} - \boxed{\mathbf{\Phi}}\mathbf{x}\|_2^2 + \boxed{\lambda} \times \boxed{f(\mathbf{x})}$$

Data-Driven Computational Sensing

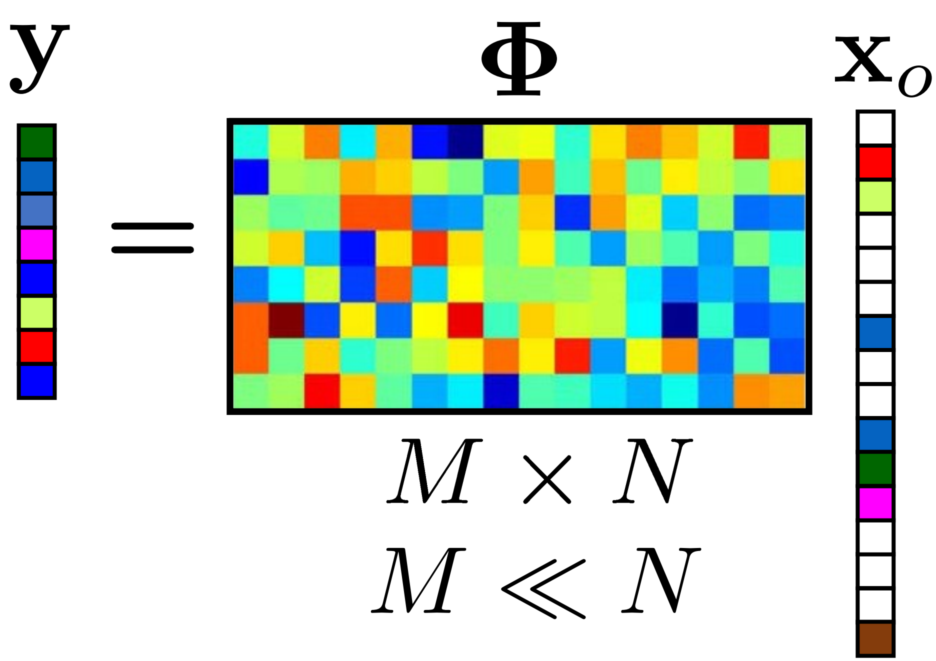


$$\min_{\mathbf{x}} \|\mathbf{y} - \boxed{\boldsymbol{\Phi}}\mathbf{x}\|_2^2 + \boxed{\lambda} \times \boxed{f(\mathbf{x})}$$

Sparse Regression

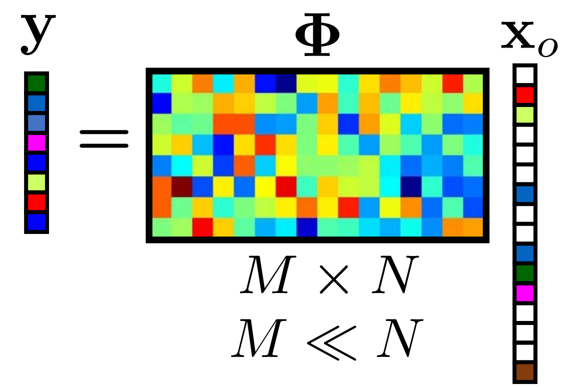
$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \lambda \times f(\mathbf{x})$$

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$$



Approximate Message Passing

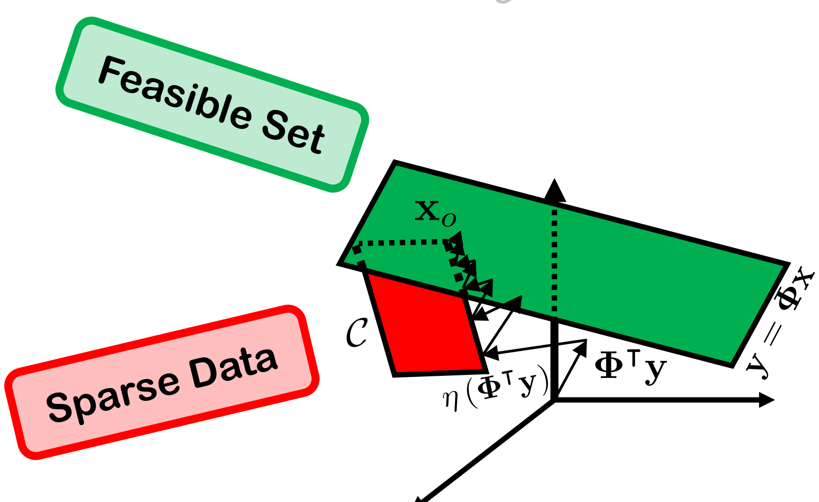
$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$$



- **Approximate Message Passing (AMP)** [Donoho, Maleki, Montanari 2009]

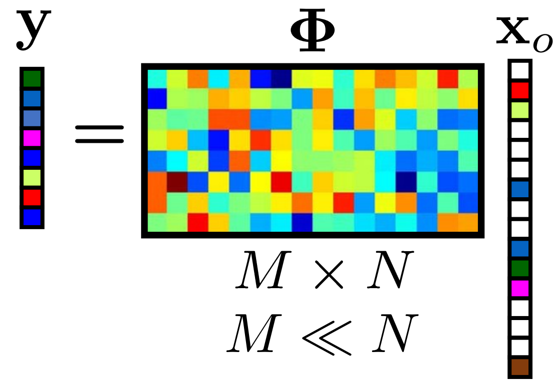
$$\mathbf{x}^{t+1} = \eta(\mathbf{x}^t + \mathbf{\Phi}^\intercal \mathbf{z}^t; \tau^t)$$

$$\mathbf{z}^t = \mathbf{y} - \mathbf{\Phi}\mathbf{x}^t + \frac{1}{\delta}\mathbf{z}^{t-1}\langle \eta'(\mathbf{x}^{t-1} + \mathbf{\Phi}^\intercal \mathbf{z}^{t-1})\rangle$$



Approximate Message Passing

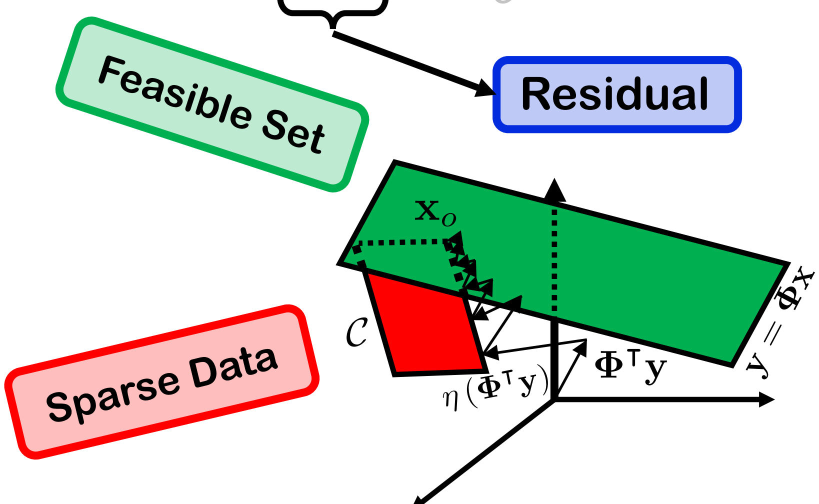
$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$$



- **Approximate Message Passing (AMP)** [Donoho, Maleki, Montanari 2009]

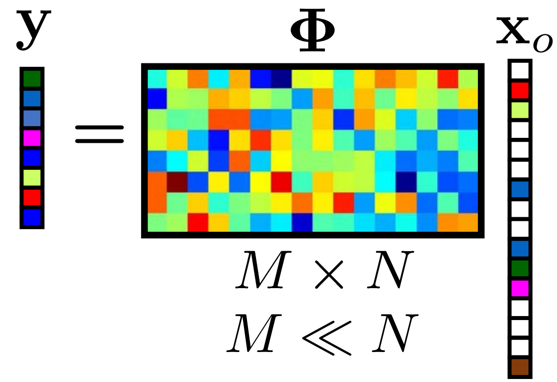
$$\mathbf{x}^{t+1} = \eta(\mathbf{x}^t + \mathbf{\Phi}^\intercal \mathbf{z}^t; \tau^t)$$

$$\underbrace{\mathbf{z}^t}_{\text{Residual}} = \mathbf{y} - \mathbf{\Phi}\mathbf{x}^t + \frac{1}{\delta}\mathbf{z}^{t-1}\left\langle \eta'(\mathbf{x}^{t-1} + \mathbf{\Phi}^\intercal \mathbf{z}^{t-1})\right\rangle$$



Approximate Message Passing

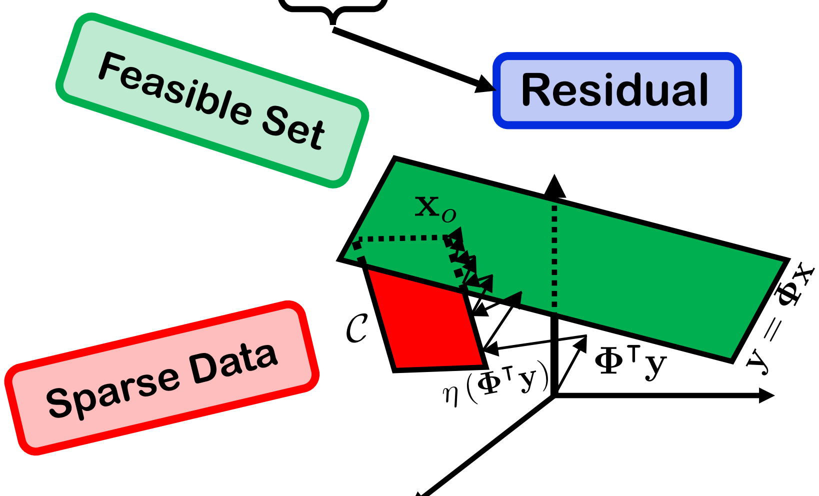
$$\min_{\mathbf{x}} \|\mathbf{y} - \boldsymbol{\Phi}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$$



- **Approximate Message Passing (AMP)** [Donoho, Maleki, Montanari 2009]

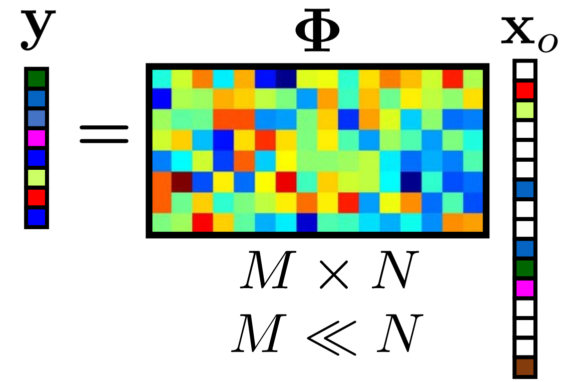
$$\mathbf{x}^{t+1} = \eta(\mathbf{x}^t + \overbrace{\boldsymbol{\Phi}^\intercal\mathbf{z}^t}^{\text{Gradient Step}}; \tau^t)$$

$$\underbrace{\mathbf{z}^t}_{\text{Residual}} = \mathbf{y} - \boldsymbol{\Phi}\mathbf{x}^t + \frac{1}{\delta}\mathbf{z}^{t-1}\left\langle \eta'(\mathbf{x}^{t-1} + \boldsymbol{\Phi}^\intercal\mathbf{z}^{t-1}) \right\rangle$$



Approximate Message Passing

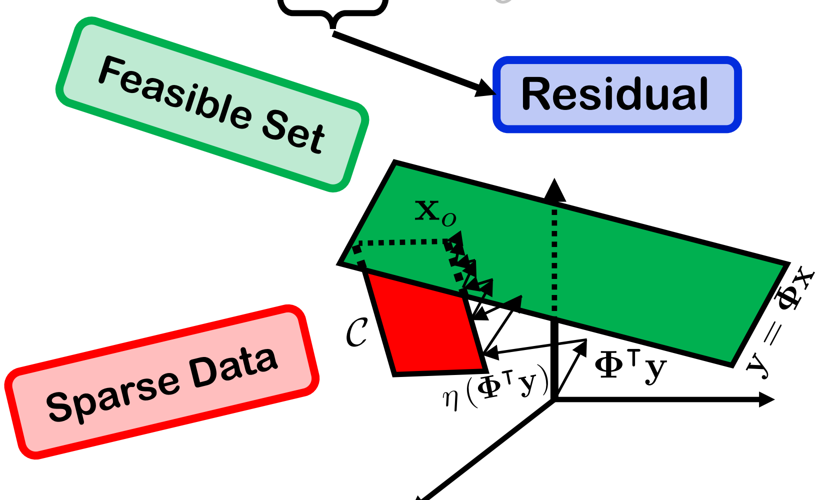
$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$$



- **Approximate Message Passing (AMP)** [Donoho, Maleki, Montanari 2009]

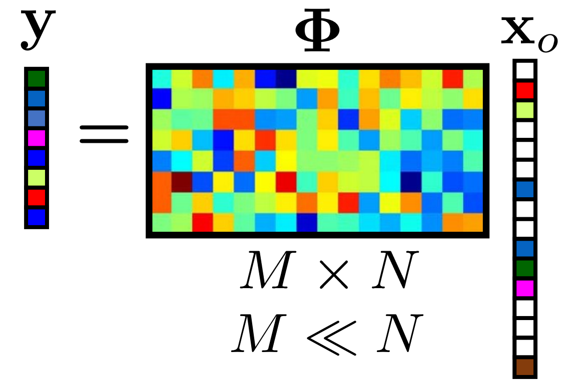
$$\mathbf{x}^{t+1} = \overbrace{\eta}^{\text{Projection Operator}}(\mathbf{x}^t + \overbrace{\mathbf{\Phi}^\intercal\mathbf{z}^t}^{\text{Gradient Step}}; \tau^t)$$

$$\underbrace{\mathbf{z}^t}_{\text{Residual}} = \mathbf{y} - \mathbf{\Phi}\mathbf{x}^t + \frac{1}{\delta}\mathbf{z}^{t-1}\left\langle \eta'(\mathbf{x}^{t-1} + \mathbf{\Phi}^\intercal\mathbf{z}^{t-1})\right\rangle$$



Approximate Message Passing

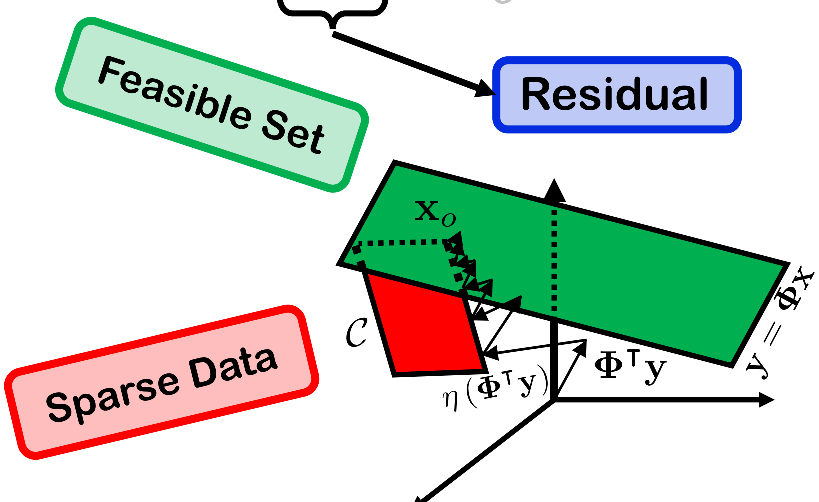
$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi x}\|_2^2 + \lambda \|\mathbf{x}\|_1$$



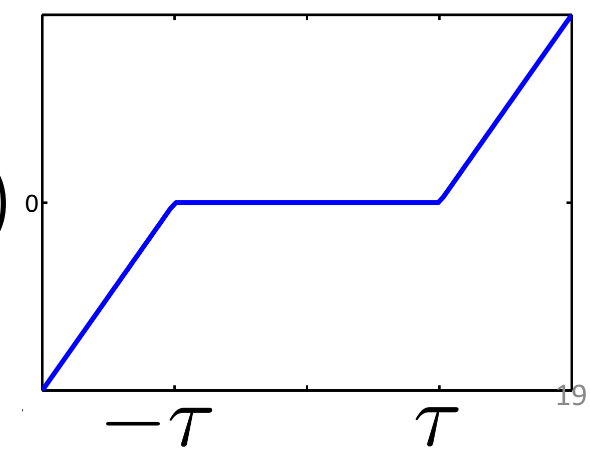
- **Approximate Message Passing (AMP)** [Donoho, Maleki, Montanari 2009]

$$\mathbf{x}^{t+1} = \overbrace{\eta}^{\text{Projection Operator}}(\mathbf{x}^t + \overbrace{\mathbf{\Phi}^\intercal \mathbf{z}^t}^{\text{Gradient Step}}; \tau^t)$$

$$\underbrace{\mathbf{z}^t}_{\text{Residual}} = \mathbf{y} - \mathbf{\Phi x}^t + \frac{1}{\delta}\mathbf{z}^{t-1} \langle \eta'(\mathbf{x}^{t-1} + \mathbf{\Phi}^\intercal \mathbf{z}^{t-1}) \rangle$$

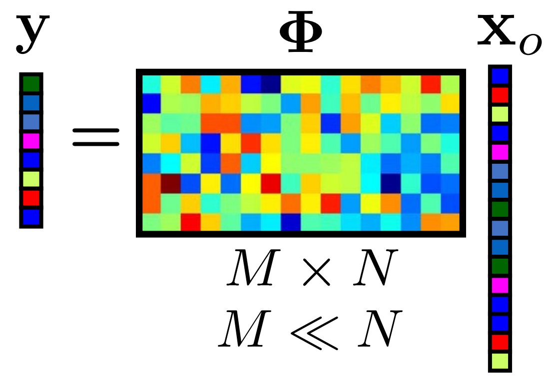


$$\eta(x, \tau)$$

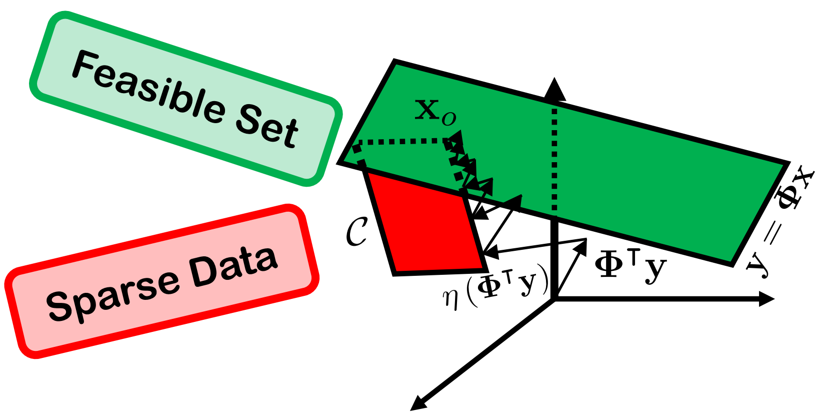


Sparse Regression

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$$



- **Approximate Message Passing (AMP)** [Donoho, Maleki, Montanari 2009]

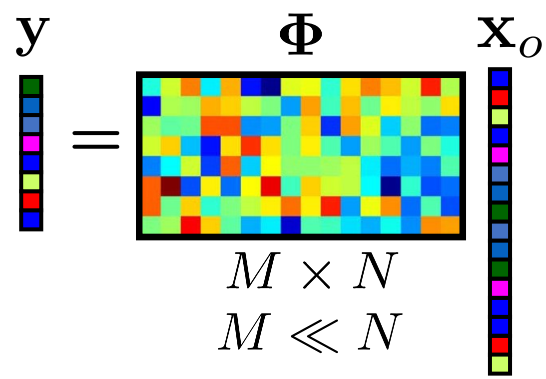


$$\mathbf{x}^{t+1} = \boxed{\eta}(\mathbf{x}^t + \mathbf{\Phi}^\intercal \mathbf{z}^t; \tau^t)$$

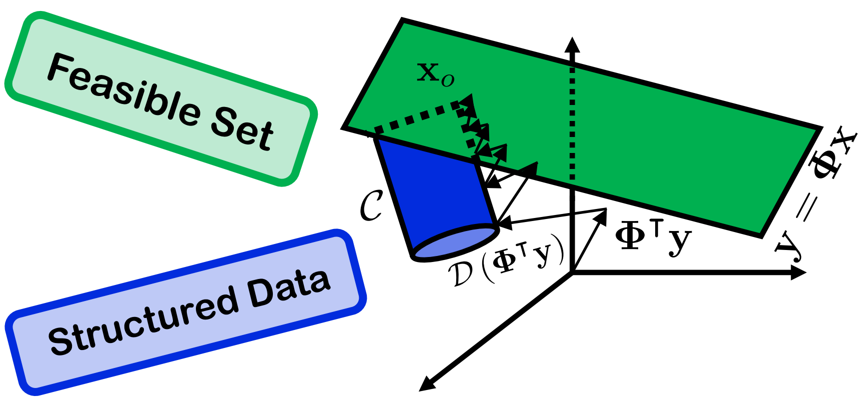
$$\mathbf{x}^t + \mathbf{\Phi}^\intercal \mathbf{z}^t = \mathbf{x}_o + \underbrace{\mathbf{v}^t}_{\text{Effective Noise}}$$

Structured Regression

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \lambda f(\mathbf{x})$$

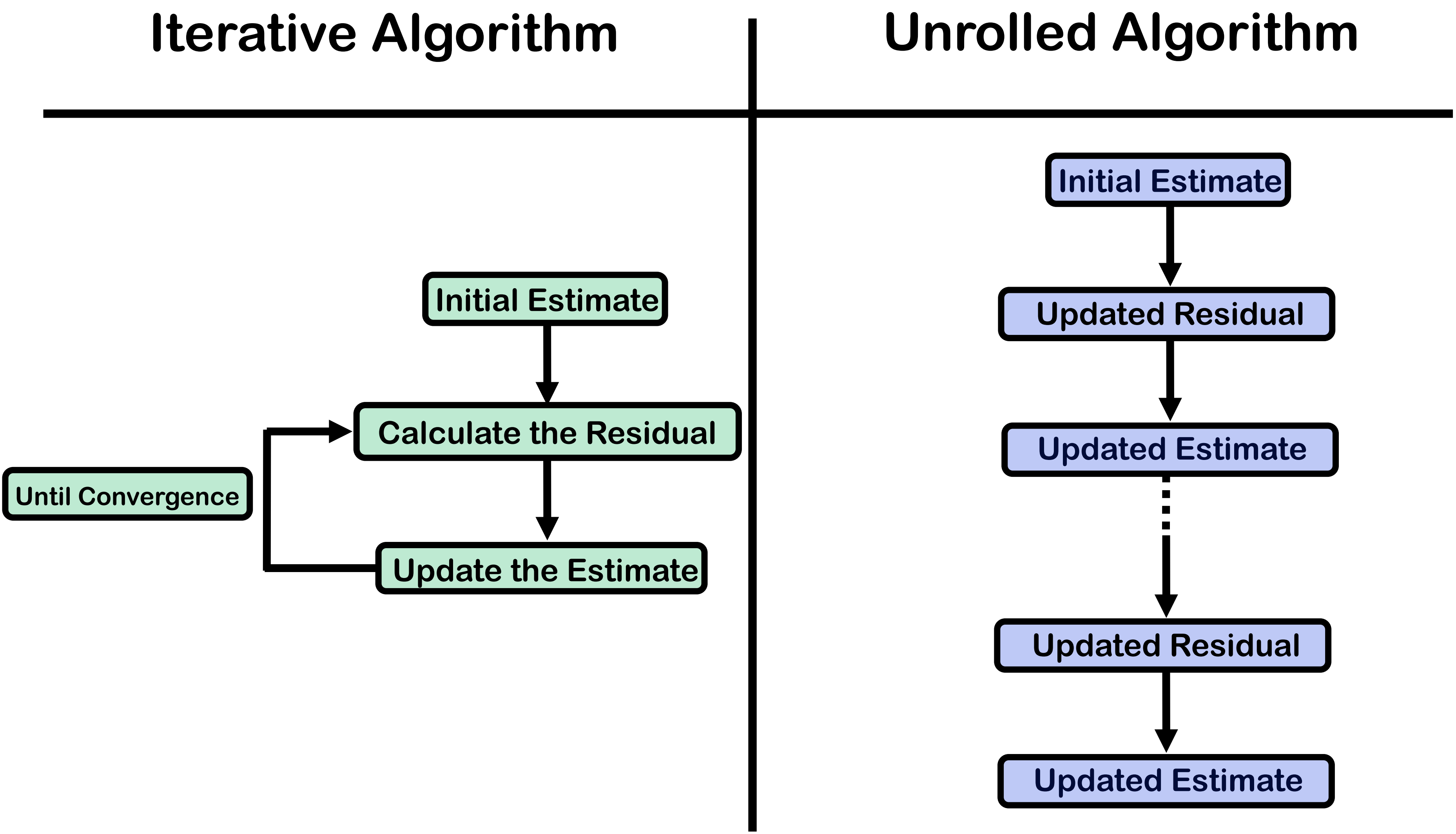


- **Denoising Approximate Message Passing (D-AMP)** [Metzler, Maleki, Baraniuk 2015]



$$\mathbf{x}^{t+1} = \boxed{\mathcal{D}^t}(\mathbf{x}^t + \mathbf{\Phi}^\intercal \mathbf{z}^t)$$

Unrolling Iterative Algorithms



[Gregor and LeCun, 2010]

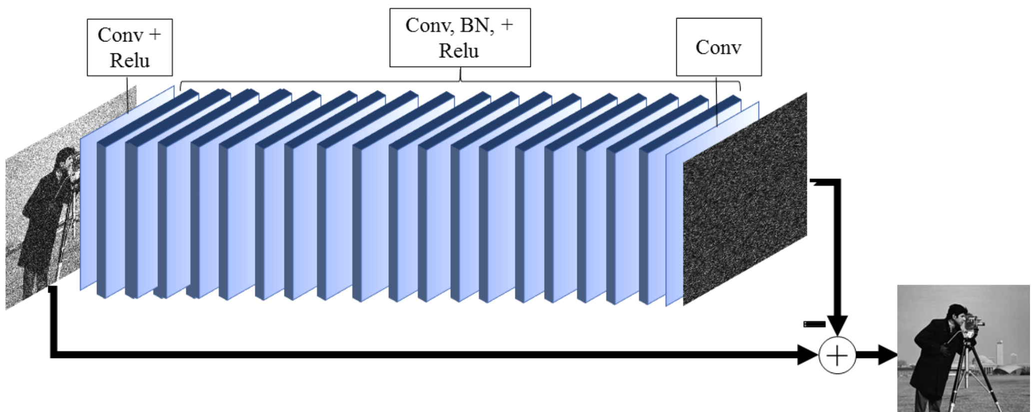
Learned-Denoising-AMP

Learned-Denoising-AMP (LDAMP) [Metzler, Mousavi, Baraniuk, *NIPS 2017*]

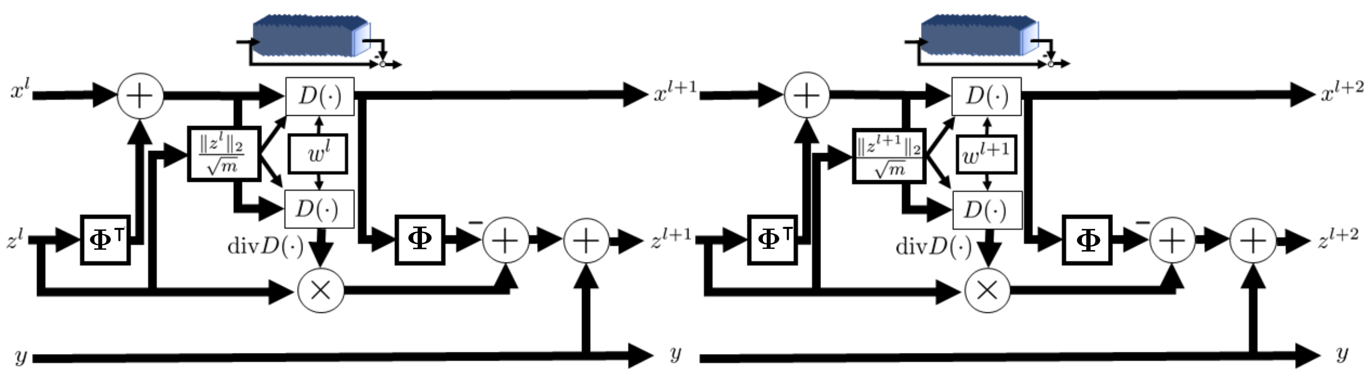
$$\mathbf{x}^{l+1} = \mathcal{D}^l(\mathbf{x}^l + \mathbf{\Phi}^\intercal \mathbf{z}^l)$$

$$\mathbf{z}^l = \mathbf{y} - \mathbf{\Phi}\mathbf{x}^l + \frac{1}{\delta}\mathbf{z}^{l-1}\langle \mathrm{div}\mathcal{D}^l(\mathbf{x}^{l-1} + \mathbf{\Phi}^\intercal \mathbf{z}^{l-1})\rangle$$

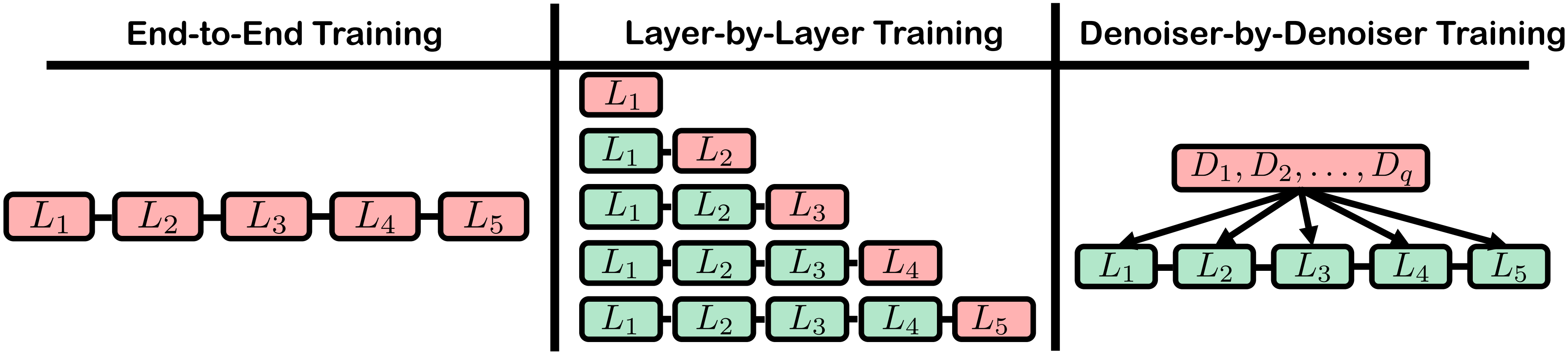
- **We use a 20-layer convolutional network as a denoiser** [Zhang et al. 2017]



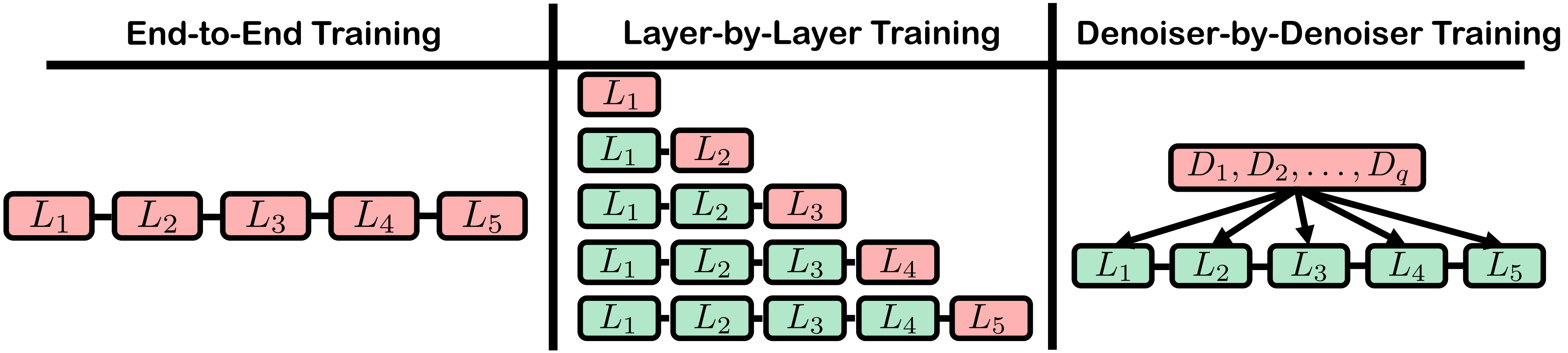
- **Two layers of the LDAMP network**



Training LDAMP and LDIT



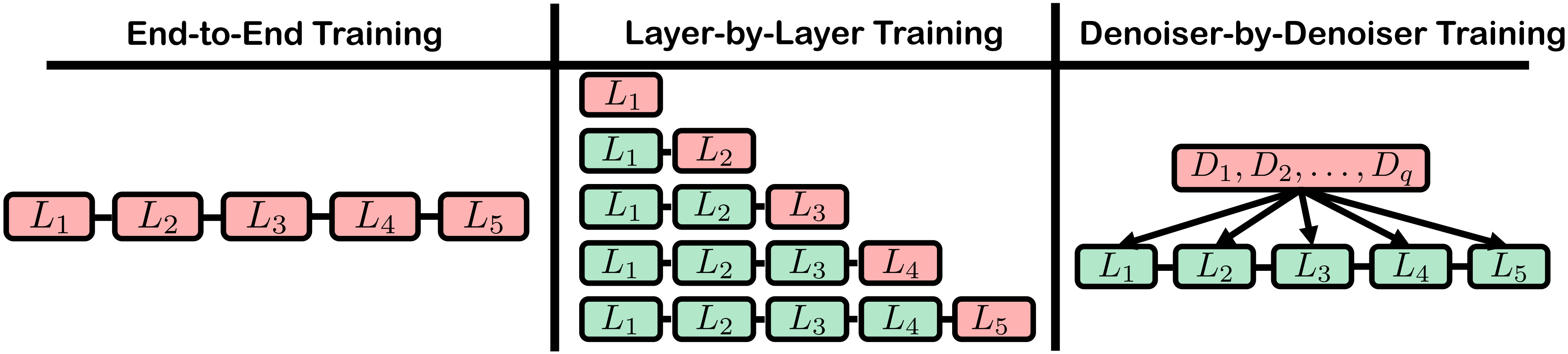
Training LDAMP



- **Lemma 1** [Metzler, Mousavi, Baraniuk, *NIPS 2017*]
  Layer-by-layer training of LDAMP is MMSE optimal.

- **Lemma 2** [Metzler, Mousavi, Baraniuk, *NIPS 2017*]
  Denoiser-by-denoiser training of LDAMP is MMSE optimal.

Training LDAMP



- **Lemma 1** [Metzler, Mousavi, Baraniuk, *NIPS 2017*]
  Layer-by-layer training of LDAMP is MMSE optimal.

- **Lemma 2** [Metzler, Mousavi, Baraniuk, *NIPS 2017*]
  Denoiser-by-denoiser training of LDAMP is MMSE optimal.

Average PSNR (dB) of one hundred 40x40 images Recovered from i.i.d Gaussian Measurements

	Training: $\frac{M}{N} = 0.2$ Testing: $\frac{M}{N} = 0.2$		Training: $\frac{M}{N} = 0.2$ Testing: $\frac{M}{N} = 0.05$	
Training Method	LDIT	LDAMP	LDIT	LDAMP
End-to-End	32.1	33.1	8.0	18.7
Layer-by-Layer	26.1	33.1	-2.6	18.7
Denoiser-by-Denoiser	28.0	31.6	22.1	25.9

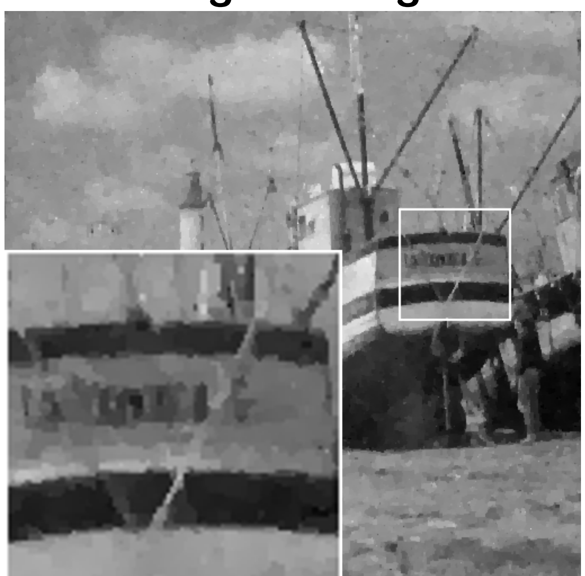
- **Noise discretization degrades the performance.**
- **Denoiser-by-denoiser is more generalizable.**

Compressive Image Recovery

512x512 images, 20x undersampling, noiseless measurements

Original Image

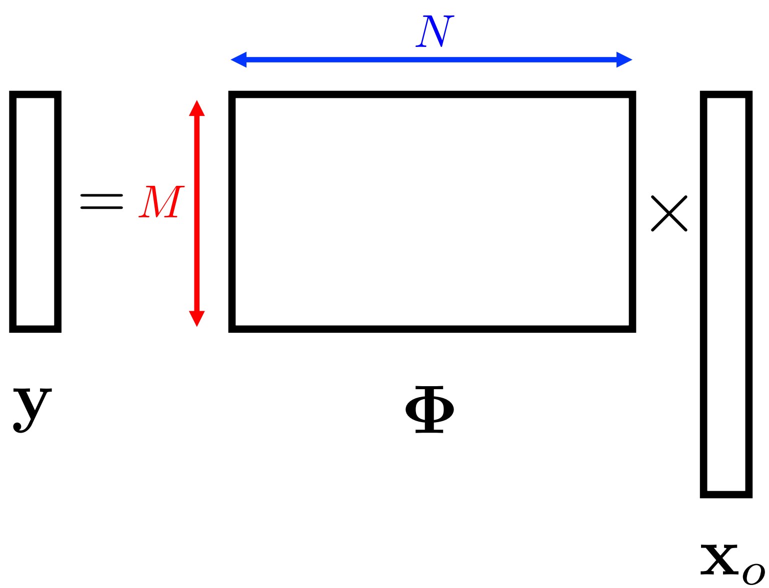
BM3D-AMP (27.2 dB, 75.04 sec)



TVAL3 (26.4 dB, 6.85 sec)

LDAMP (28.1 dB, 1.22 sec)

summary so far

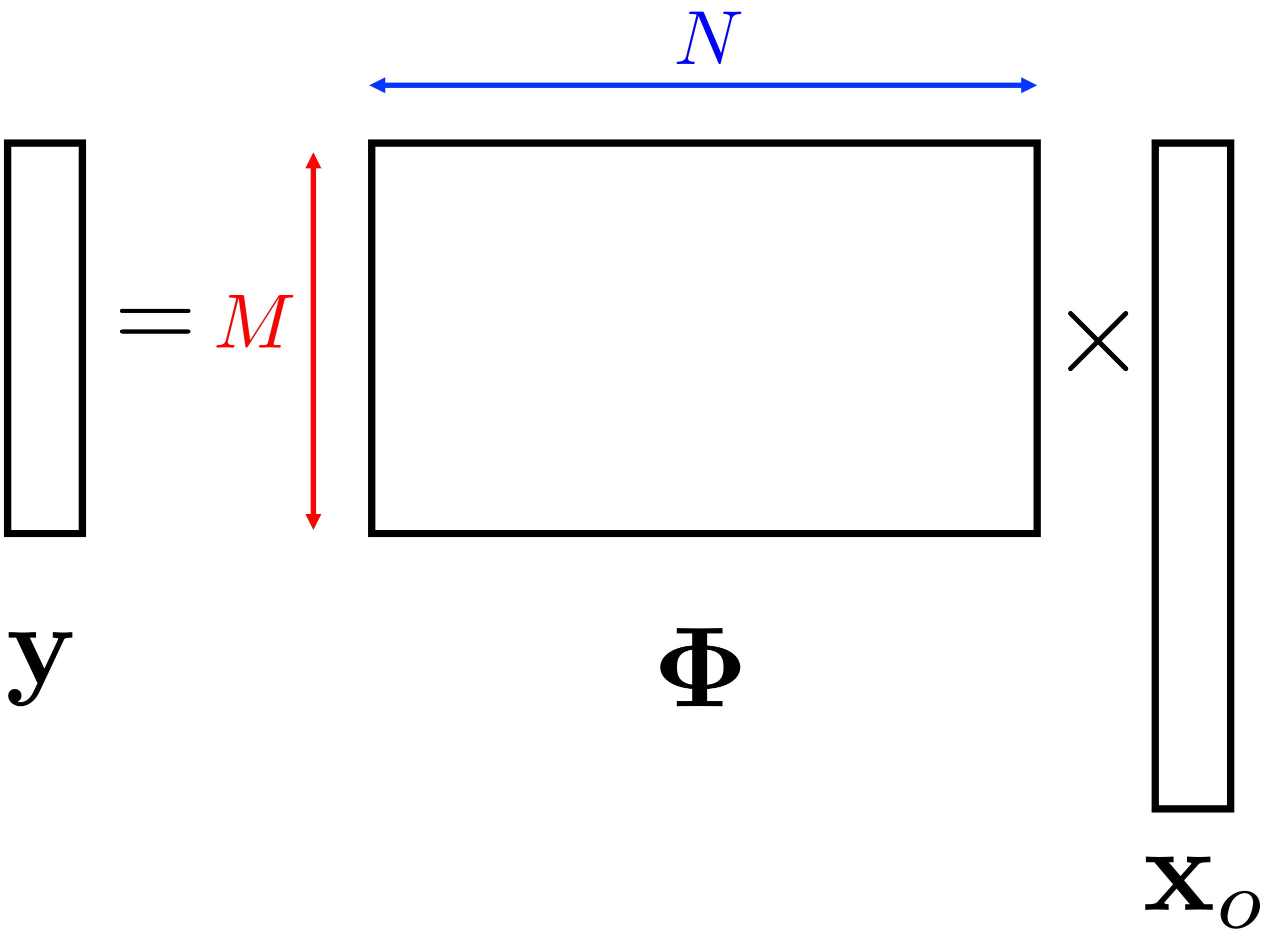


$$\arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 \text{ subject to } \mathbf{x} \in \mathcal{C}$$

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \lambda \times f(\mathbf{x})$$

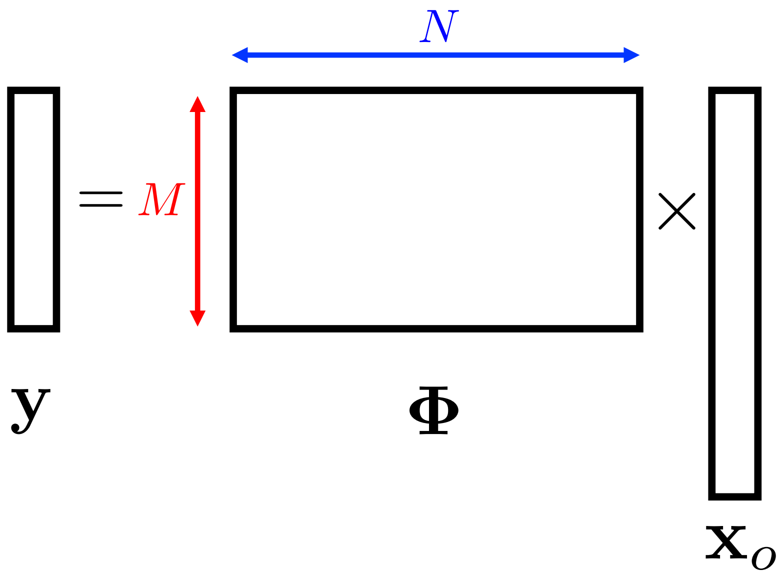
$$\mathbf{x}_1, \mathbf{x}_1, \ldots, \mathbf{x}_L$$

Data-Driven Computational Sensing



$$\min_{\mathbf{x}} \|\mathbf{y} - \boxed{\mathbf{\Phi}}\mathbf{x}\|_2^2 + \boxed{\lambda} \times \boxed{f(\mathbf{x})}$$

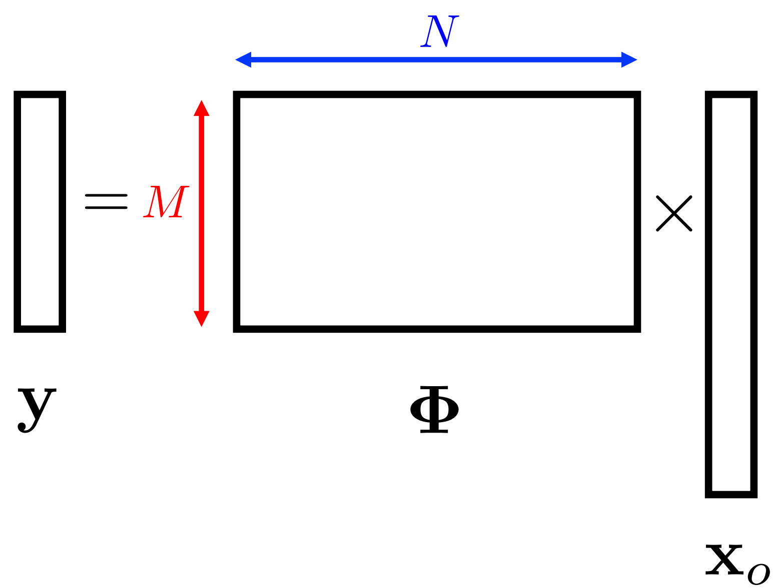
Data-Driven Computational Sensing



$$\min_{\mathbf{x}} \|\mathbf{y} - \boxed{\boldsymbol{\Phi}}\mathbf{x}\|_2^2 + \boxed{\lambda} \times \boxed{f(\mathbf{x})}$$

- Mousavi, Maleki, Baraniuk, '[Consistent Parameter Estimation]', Annals of Statistics 2017
- Mousavi, Dasarathy, Baraniuk, '[Data-Driven Sparse Representation'], ICLR 2019

Summary so far

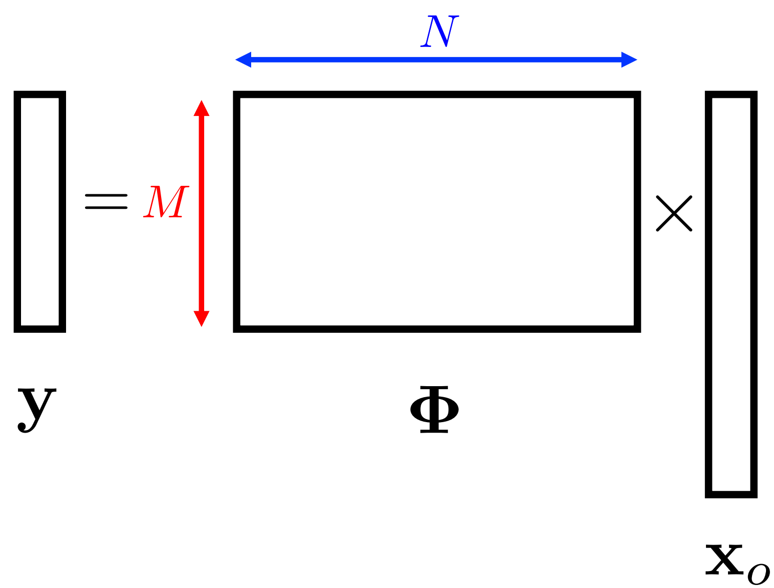


$$\arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 \text{ subject to } \mathbf{x} \in \mathcal{C}$$

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \lambda \times f(\mathbf{x})$$

Supervised $\leftarrow$ $\mathbf{x}_1, \mathbf{x}_1, \ldots, \mathbf{x}_L$

Next Step



$$\arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 \text{ subject to } \mathbf{x} \in \mathcal{C}$$

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \lambda \times f(\mathbf{x})$$

Unsupervised ⟵ $\boxed{\cancel{\mathbf{x}_1, \mathbf{x}_1, \cdots, \mathbf{x}_L}}$

Stein's Unbiased Risk Estimator (SURE) [Stein '81]

- **A statistical model selection technique**

$\mathbf{x}$ **Unknown**

$$\mathbf{y} = \mathbf{x} + \mathbf{w}$$

$$\mathbf{w} \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$$

$f_\theta(.)$ **Weakly differentiable**

$$\mathbb{E}_{\mathbf{w}}\left[\frac{1}{n}\|\mathbf{x} - f_\theta(\mathbf{y})\|^2\right] = \mathbb{E}_{\mathbf{w}}\left[\frac{1}{n}\|\mathbf{y} - f_\theta(\mathbf{y})\|^2\right] - \sigma_w^2 + \frac{2\sigma_w^2}{n}\mathrm{div}_{\mathbf{y}}(f_\theta(\mathbf{y}))$$

$$\mathrm{div}_{\mathbf{y}}(f_\theta(\mathbf{y})) = \sum_{n=1}^{N} \frac{\partial f_\theta(\mathbf{y})}{\partial y_n}$$

Monte-Carlo SURE [Ramani, Blu, Unser, 2008]

- **Challenge: Computing the divergence**

$$\mathrm{div}_{\mathbf{y}}(f_\theta(\mathbf{y})) = \sum_{n=1}^{N} \frac{\partial f_\theta(\mathbf{y})}{\partial y_n}$$

- **For bounded functions:**

$$\mathbf{b} \sim \mathcal{N}(0, \mathbb{I})$$

$$\mathrm{div}_{\mathbf{y}}(f_\theta(\mathbf{y})) = \lim_{\epsilon \to 0} \mathbb{E}_{\mathbf{b}} \left\{ \mathbf{b}^t \left(\frac{f_\theta(\mathbf{y} + \epsilon \mathbf{b}) - f_\theta(\mathbf{y})}{\epsilon} \right) \right\}$$

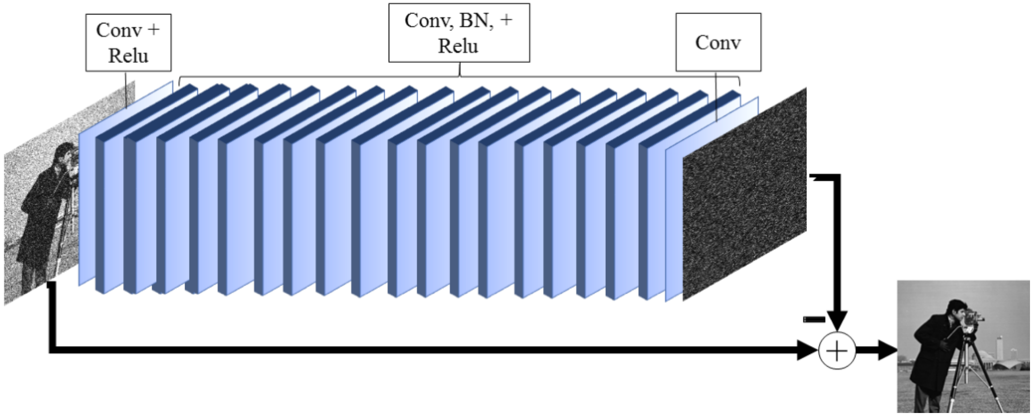
- **Approximation:**

$$\epsilon = \frac{\max(\mathbf{y})}{1000}$$

$$\mathrm{div}_{\mathbf{y}}(f_\theta(\mathbf{y})) \approx \mathbf{b}^t \left(\frac{f_\theta(\mathbf{y} + \epsilon \mathbf{b}) - f_\theta(\mathbf{y})}{\epsilon} \right)$$

Denoising with Noisy Data

- **DnCNN Denoiser:**



- **Training Data:**

$$\cancel{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L}$$

$$\mathbf{y} = \mathbf{x} + \mathbf{w}$$
$$\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_L$$

- **Loss Function:**

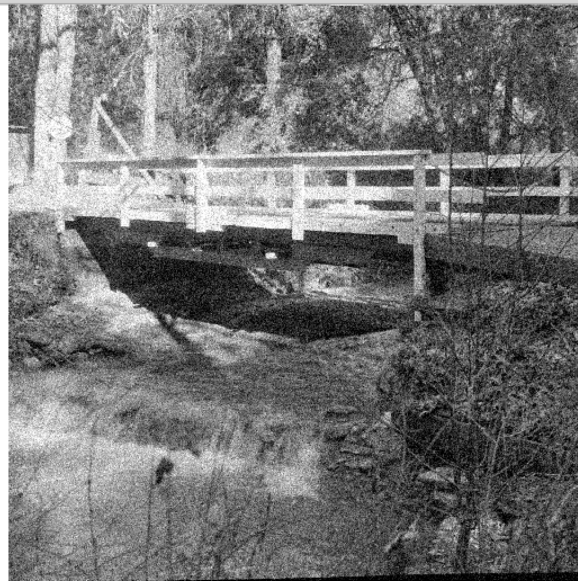
MSE

$$\sum_{\ell=1}^{L} \frac{1}{n} \|\mathbf{x}_\ell - f_\theta(\mathbf{y}_\ell)\|^2$$

SURE

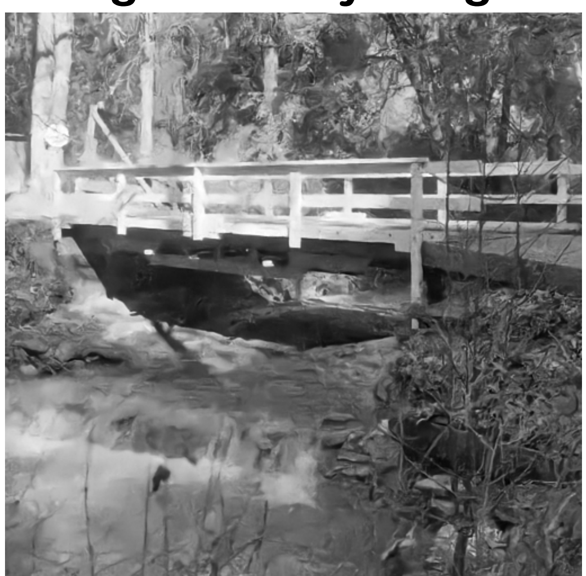
$$\sum_{\ell=1}^{L} \frac{1}{n} \|\mathbf{y}_\ell - f_\theta(\mathbf{y}_\ell)\|^2 - \sigma_w^2 + \frac{2\sigma_w^2}{n} \mathrm{div}_{\mathbf{y}_\ell}\{(f_\theta(\mathbf{y}_\ell))\}$$

Denoising with Noisy Data Results



Original Noisy Image

BM3D (26.0 dB, 4.01 sec.)



DnCNN SURE (26.5 dB, 0.04 sec.)

DnCNN MSE (26.7 dB, 0.04 sec.)

Compressive Image Recovery w/ Noisy Data

- **Problem Formulation:**

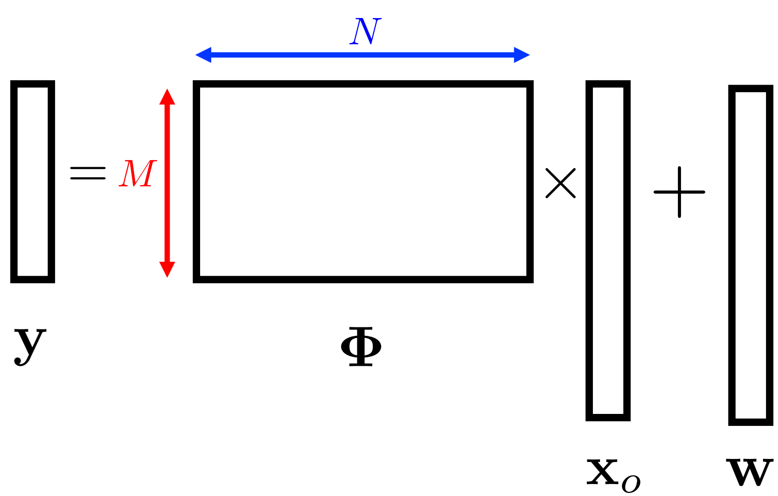
Image: $\mathbf{x} \in \mathbb{R}^N$

Measurements: $\mathbf{y} = \mathbf{\Phi x} + \mathbf{w}, \quad \mathbf{y} \in \mathbb{R}^M$

Measurement Operator: $\mathbf{\Phi} \in \mathbb{R}^{M \times N}$

Noise: $\mathbf{w} \in \mathbb{R}^M$

Setting: $M \ll N$



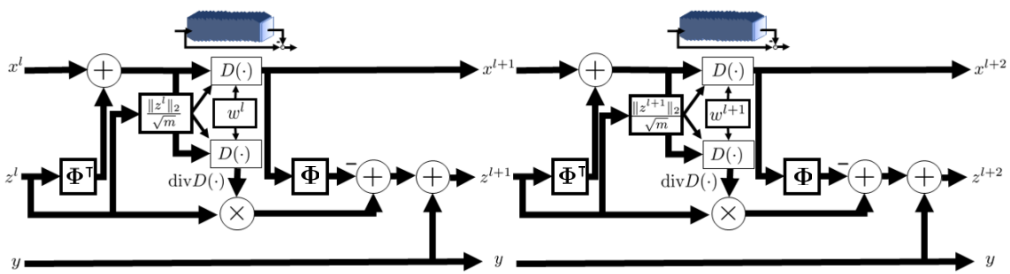
Recovery Algorithm

- **Learning Denoising-based AMP (LDAMP) Neural Network (for k=1,…,K):**

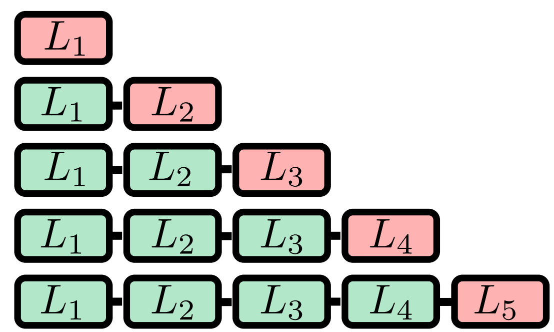
$$\mathbf{z}^k = \mathbf{y} - \mathbf{\Phi}\mathbf{x}^k + \frac{1}{m}\mathbf{z}^{k-1}\mathrm{div}D^{k-1}_{\theta^{k-1}}(\mathbf{x}^{k-1} + \mathbf{\Phi}^*\mathbf{z}^{k-1})$$

$$\sigma^k = \frac{\|\mathbf{z}^k\|_2}{\sqrt{m}}$$

$$\mathbf{x}^{k+1} = D^k_{\theta^k}(\mathbf{x}^k + \mathbf{\Phi}^*\mathbf{z}^k)$$



Layer by Layer Training



- **Decouples image recovery into a series of denoising problems:**

$$\mathbf{x}^k + \mathbf{\Phi}^*\mathbf{z}^k = \mathbf{x}_o + \sigma\mathbf{v} \qquad \mathbf{v} \sim \mathcal{N}(0, \mathbb{I})$$

[Donoho et al. 2009, 2011]
[Bayati and Montanari, 2011]

- **Layerwise Training of the LDAMP Network:**

MSE

$$\theta_{\mathrm{MSE}} = \arg\min_\theta \sum_{\ell=1}^{L} \frac{1}{n}\|\mathbf{x}_\ell - f_\theta(\mathbf{y}_\ell)\|^2$$

SURE

$$\theta_{\mathrm{SURE}} = \arg\min_\theta \sum_{\ell=1}^{L} \frac{1}{n}\|\mathbf{y}_\ell - f_\theta(\mathbf{y}_\ell)\|^2 - {\sigma^k}^2 + \frac{2(\sigma^k)^2}{n}\mathrm{div}_{\mathbf{y}_\ell}(f_\theta(\mathbf{y}_\ell))$$

Compressive Image Recovery

5x undersampling

Original Image

BM3D-AMP (31.3 dB, 13.2 sec.)

LDAMP MSE (34.6 dB, 0.4 sec.)

LDAMP SURE (31.9 dB, 0.4 sec.)

Take away Messages!

- There are three major paradigms for signal acquisition.

- Each paradigm puts resources on one of the sampling, modeling, or reconstruction tasks.

- There seems to be a preservation of computation between different paradigms.