Dictionary Learning for Graph Signals

236862 – Introduction to Sparse and Redundant Representations

joint work with

Prof. Michael Elad

Yael Yankelevsky

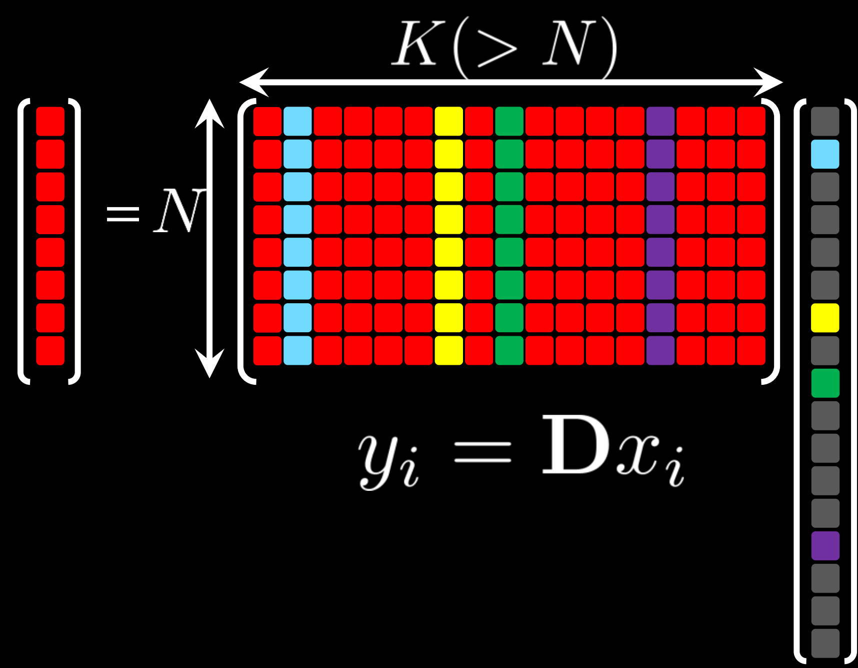
22.12.2019

Technion
Israel Institute of Technology

The *Sparseland* Model

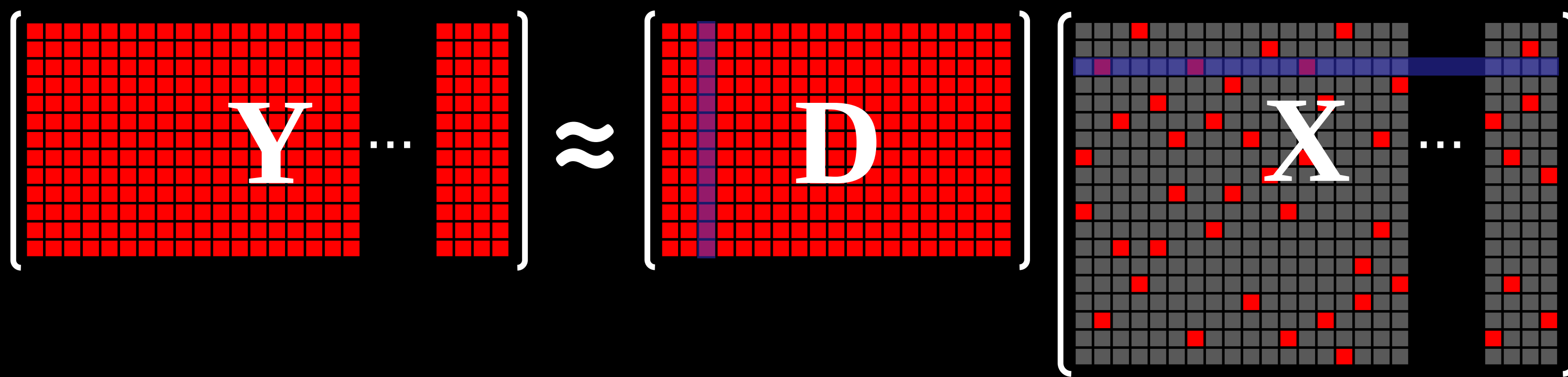
Dictionary Learning:

$$\boxed{\begin{aligned}&\arg\min_{\mathbf{D},\mathbf{X}} \|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2 \\ &\text{s.t.} \quad \|x_i\|_0 \le T \quad \forall i\end{aligned}}$$



Model assumption: All data vectors are linear combinations of FEW ($T \ll N$) atoms from $\mathbf{D}$

K-SVD Algorithm Overview [Aharon et al. '06]

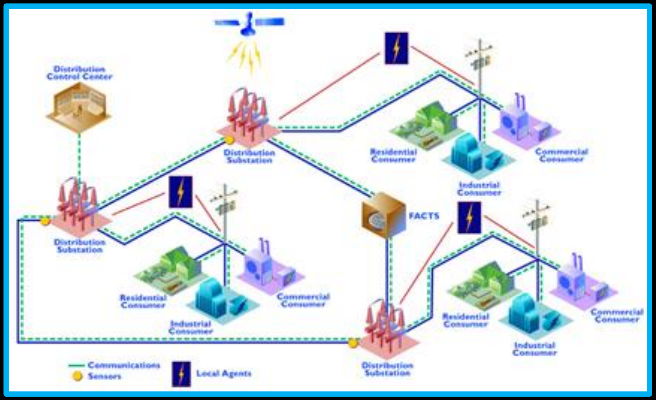


Initialize Dictionary → Sparse Coding (Using OMP) → Dictionary Update (Atom-by-atom + coeffs.)

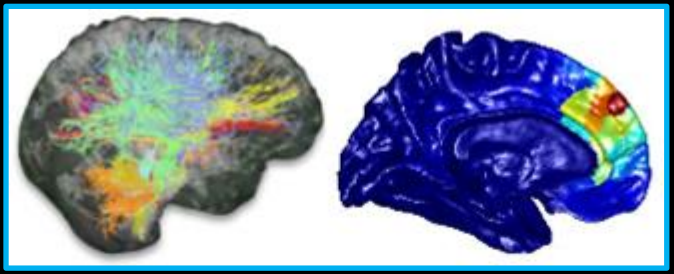
For the j-th atom:

$$\begin{cases} d_j = \mathbf{E}_j \mathbf{P}_j x_j^R / \|x_j^R\|_2^2 \\ x_j^R = \mathbf{P}_j^T \mathbf{E}_j^T d_j / \|d_j\|_2^2 \end{cases}$$

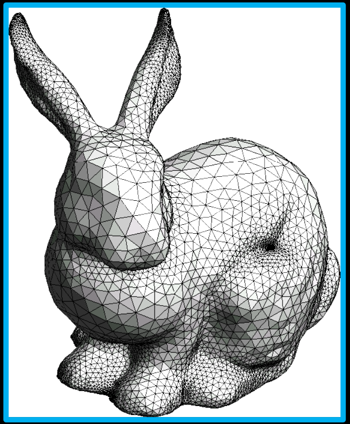
Data is often structured…



Energy Networks



Biological Networks



Meshes & Point Clouds

Transportation Networks

Social Networks

What happens for non-conventionally structured signals?

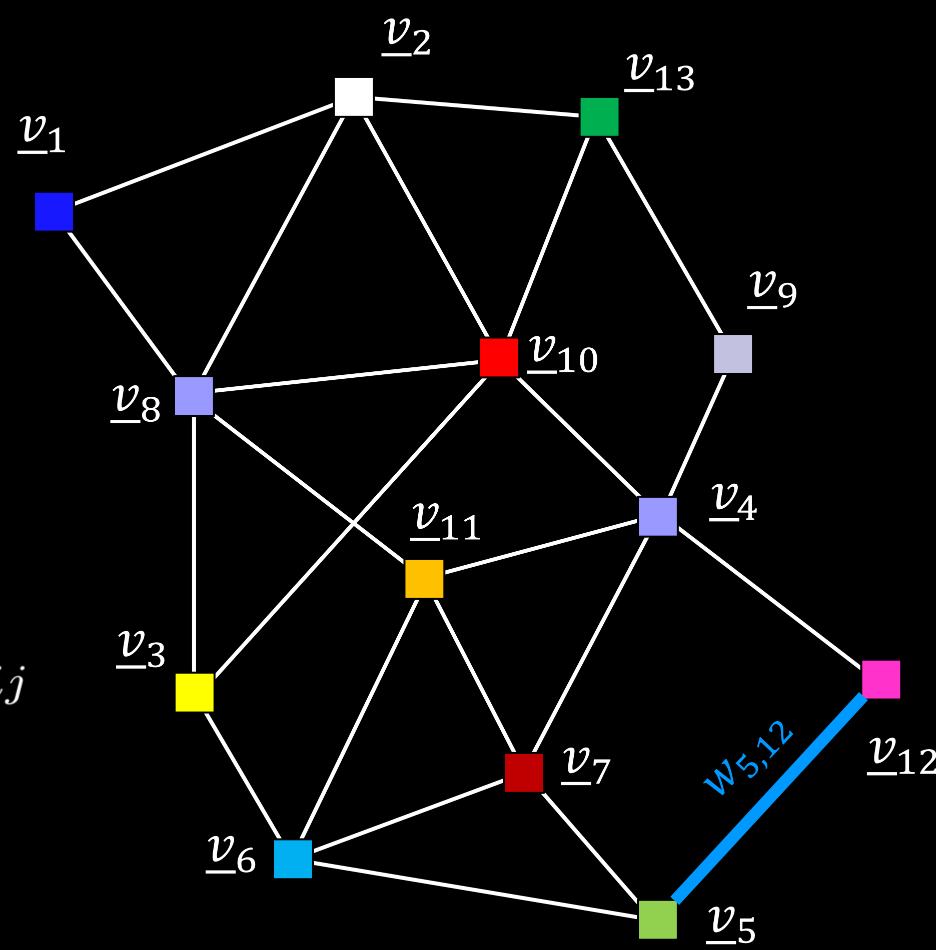
Can dictionary learning work well for such signals as well?

The general idea:
Model the underlying structure as a graph and incorporate it in the dictionary learning algorithm

Basic Notations

We are given a graph: $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$

- The i^{th} node is characterized by a feature vector $\underline{v}_i$
- The edge between the i^{th} and j^{th} nodes carries a weight $w_{ij} \propto d(\underline{v}_i, \underline{v}_j)^{-1}$
- The degree matrix: $D_{ii} = \sum_j w_{ij}$
- Graph Laplacian: $L = D - W$



Basic Notations

- The i^{th} node has a value f_i

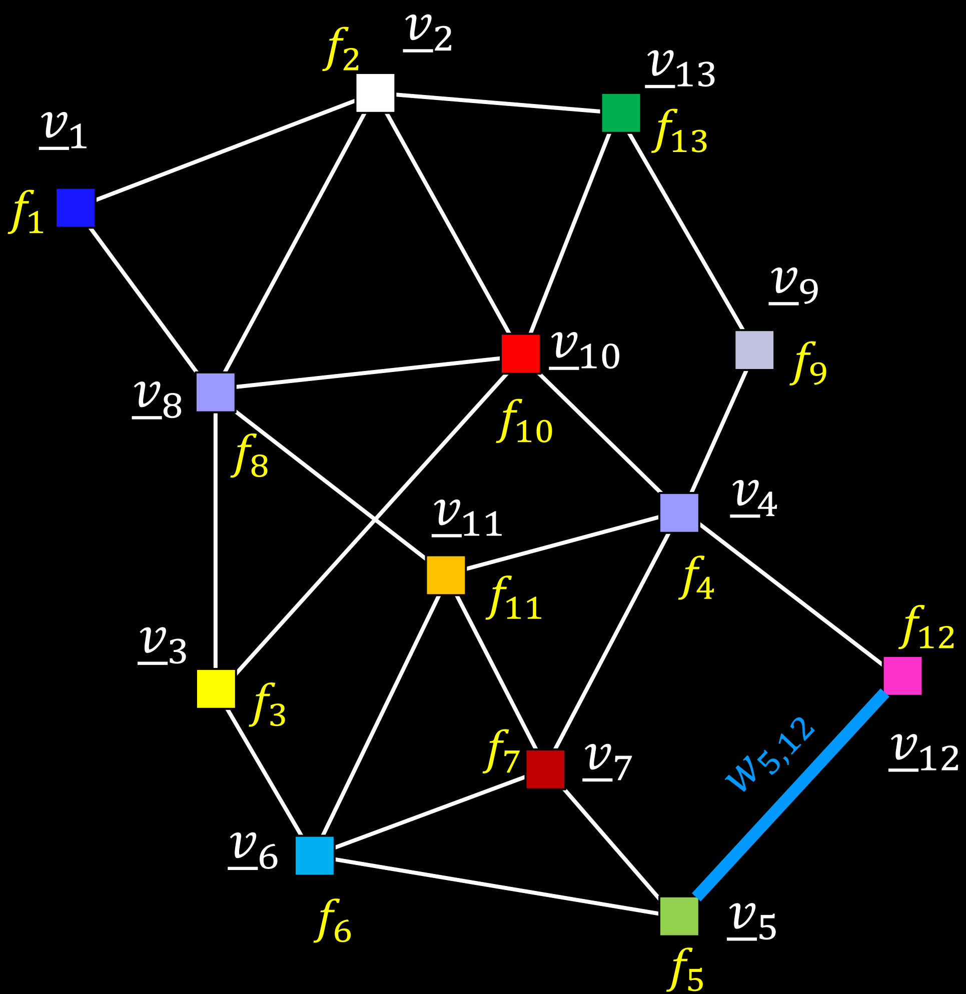
  $\underline{f}$ = graph signal

- The combinatorial Laplacian is a differential operator:

$$(Lf)_i = \sum_j w_{ij}(f_i - f_j)$$

- Defines global regularity on the graph (Dirichlet energy):

$$\boxed{f^T L f = \frac{1}{2}\sum_{i,j} w_{ij}(f_i - f_j)^2}$$



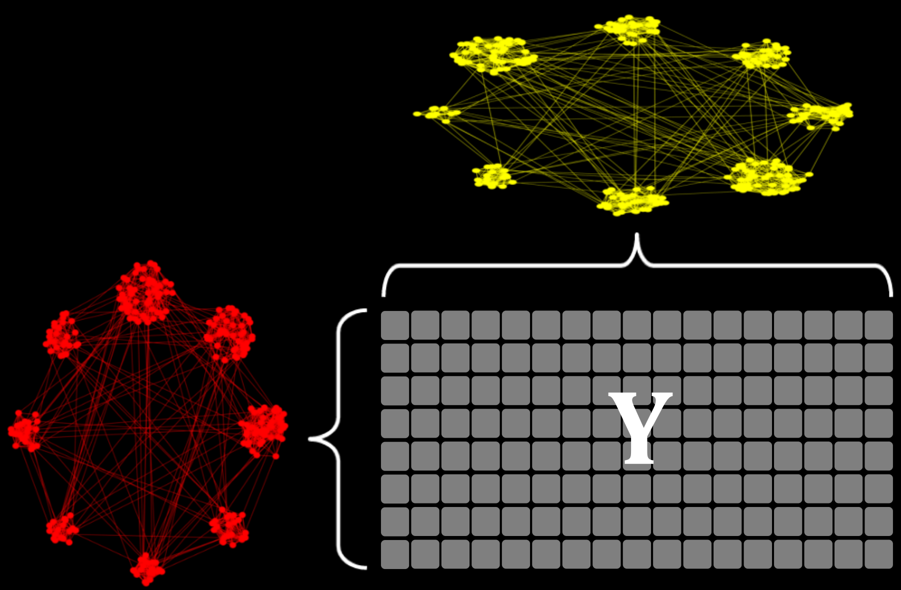
Related Work: Dictionaries for Graph Signals

- Ignore structure (MOD, K-SVD) [Engan et al. ’99],[Aharon et al. ’06]
- Analytic transforms
  - Graph Fourier transform [Sandryhaila & Moura ’13]
  - Windowed Graph Fourier transform [Shuman et al. ’12]
  - Graph Wavelets [Coifman & Maggioni ’06],[Gavish et al. ’10],[Hammond et al. ’11],[Ram et al. ’12],[Shuman et al. ’16],…
- Structured learned dictionaries [Zhang et al. ’12],[Thanou et al. ’14]

Our solution: Graph Regularized Dictionary Learning

The Basic Concept

Construct 2 graphs capturing the feature dependencies and the data manifold structure



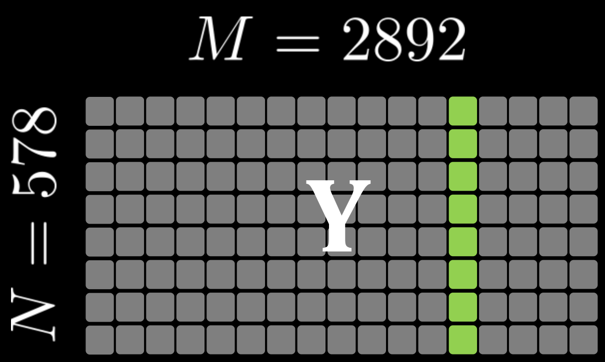
$$w_{ij}^{\mathcal{M}} = \exp\left(-\frac{\|y_i - y_j\|_2^2}{\varepsilon_{\mathcal{M}}}\right)$$

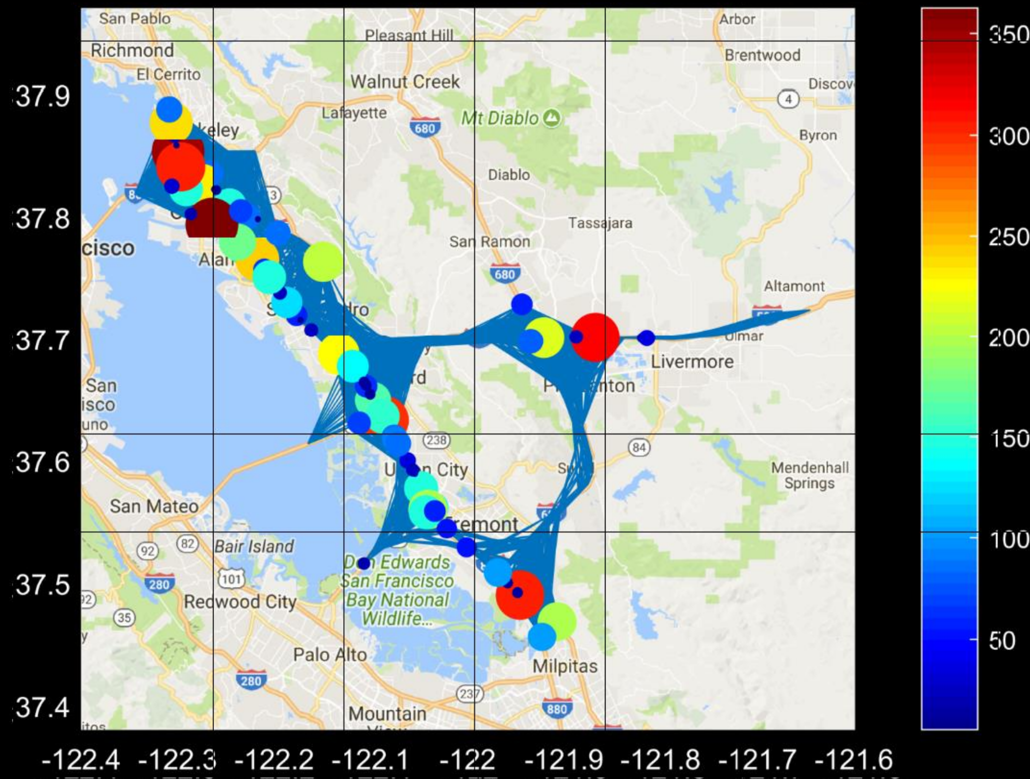
$$\Downarrow$$

$$\mathrm{L_c}$$

$$w_{ij}^{\mathcal{G}} = \exp\left(-\frac{\|Y(i,:) - Y(j,:)\|_2^2}{\varepsilon_{\mathcal{G}}}\right) \Rightarrow \mathrm{L}$$

Example: Traffic Dataset





Dual Graph Regularized Dictionary Learning

Introduce graph regularization terms that preserve these structures

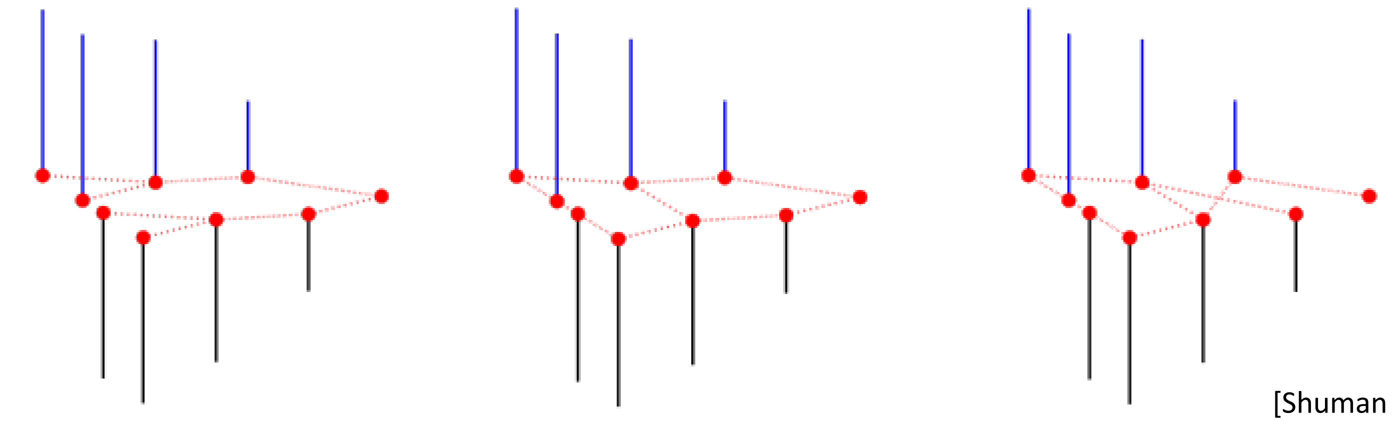
$$\underset{\mathbf{D},\mathbf{X}}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \alpha Tr(\mathbf{D}^T \mathbf{L} \mathbf{D}) + \beta Tr(\mathbf{X} \mathbf{L_c} \mathbf{X}^T)$$

$$\text{s.t.} \quad \|x_i\|_0 \leq T \quad \forall i$$

Imposed smoothness (graph Dirichlet energy):

$$Tr(\mathbf{X} \mathbf{L_c} \mathbf{X}^T) = \frac{1}{2} \sum_{i,j} w_{ij}^{\mathcal{M}} \|x_i - x_j\|_2^2$$

The Importance of the Underlying Graph



[Shuman et al. ’13]

A good estimation of $\mathbf{L}$ is crucial!

We can learn $\mathbf{L}$ and adapt it to promote the desired smoothness [Hu et al. ’13], [Dong et al. ‘15], [Kalofolias ’16], [Segarra et al. ‘17],…

Dual Graph Regularized Dictionary Learning

$$\arg\min_{\mathbf{D},\mathbf{X},\mathbf{L}} \|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2 + \alpha Tr(\mathbf{D}^T\mathbf{L}\mathbf{D}) + \beta Tr(\mathbf{X}\mathbf{L_c}\mathbf{X}^T) + \mu\|\mathbf{L}\|_F^2$$

$$\text{s.t.} \quad \|x_i\|_0 \leq T \quad \forall i$$

$$\mathbf{L}_{ij} = \mathbf{L}_{ji} \leq 0 \quad (i \neq j)$$

$$\mathbf{L} \cdot \mathbf{1} = \mathbf{0}$$

$$Tr(\mathbf{L}) = \gamma N$$

Key idea:

- Dictionary atoms preserve feature similarities
- Similar signals have similar sparse representations
- The graph is adapted to promote the desired smoothness

The DGRDL Algorithm

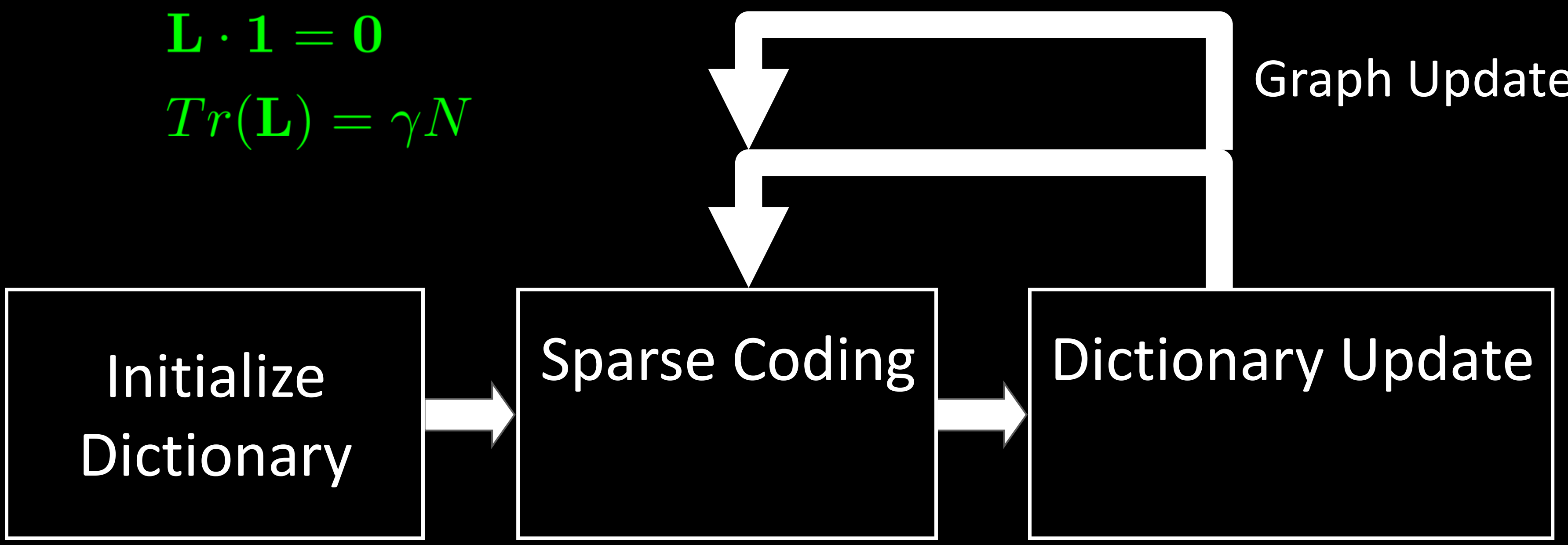
$$\underset{\mathbf{D},\mathbf{X},\mathbf{L}}{\arg\min} \ \|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2+\alpha Tr(\mathbf{D}^T\mathbf{L}\mathbf{D})+\beta Tr(\mathbf{X}\mathbf{L_c}\mathbf{X}^T)+\mu\|\mathbf{L}\|_F^2$$

$$\text{s.t.} \quad \|x_i\|_0 \leq T \quad \forall i$$

$$\mathbf{L}_{ij}=\mathbf{L}_{ji}\leq 0 \quad (i\neq j)$$

$$\mathbf{L}\cdot\mathbf{1}=\mathbf{0}$$

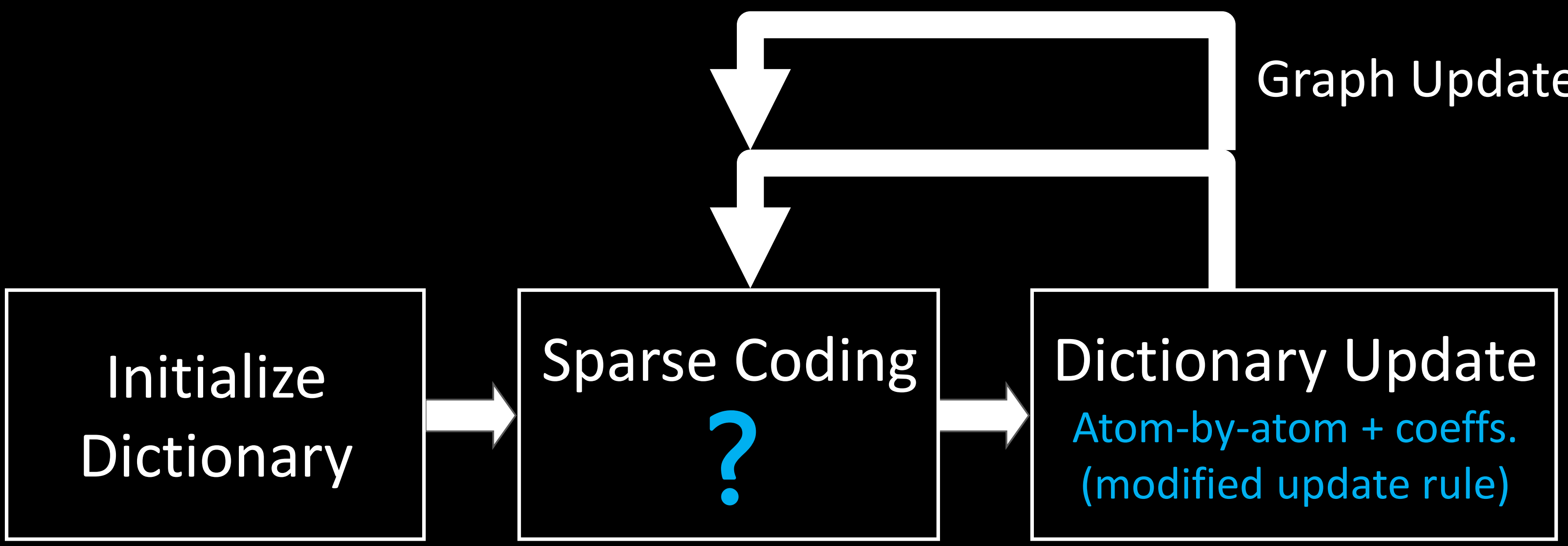
$$Tr(\mathbf{L})=\gamma N$$



The DGRDL Algorithm

For the j-th atom:

$$\begin{cases} d_j = (\|x_j^R\|_2^2\mathbf{I} + \alpha\mathbf{L})^{-1}\mathbf{E}_j\mathbf{P}_j x_j^R \\ x_j^R = (\|d_j\|_2^2\mathbf{I} + \beta\mathbf{P}_j^T\mathbf{L_c}\mathbf{P}_j)^{-1}\mathbf{P}_j^T\mathbf{E}_j^T d_j \end{cases}$$



Graph Regularized Pursuit

$$\arg\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \beta Tr(\mathbf{X}\mathbf{L_c}\mathbf{X}^T)$$

$$\text{s.t.} \quad \|x_i\|_0 \leq T \quad \forall i$$

Graph Regularized Pursuit

$$\arg\min_{\mathbf{X},\mathbf{Z}} \|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2+\beta Tr(\mathbf{X}\mathbf{L_c}\mathbf{X}^T)$$

$$\text{s.t.} \quad \|z_i\|_0 \le T \quad \forall i,$$

$$\mathbf{X}=\mathbf{Z}$$

ADMM: [Boyd et al. '11]

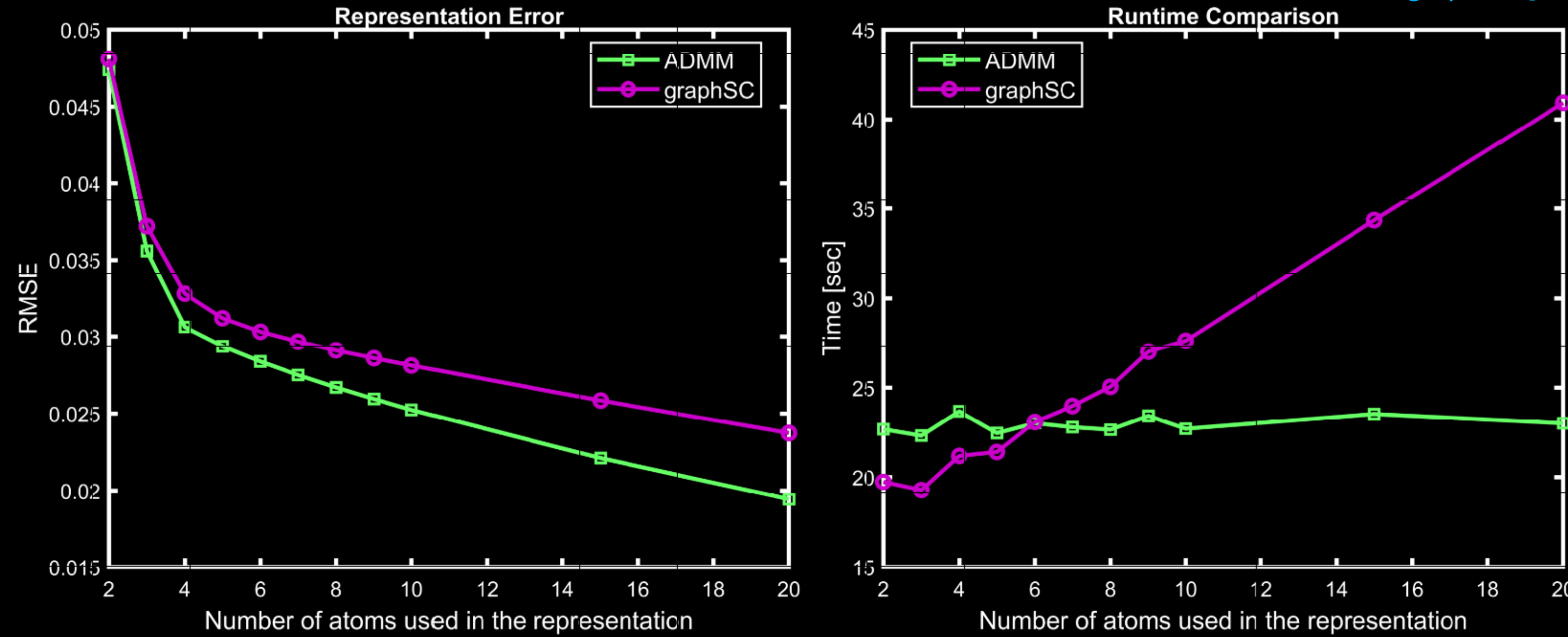
$$\mathbf{X}^{(k)} \leftarrow (\mathbf{D}^T\mathbf{D}+\rho\mathbf{I})\mathbf{X}+\beta\mathbf{X}\mathbf{L_c}=\mathbf{D}^T\mathbf{Y}+\rho\left(\mathbf{Z}^{(k-1)}-\mathbf{U}^{(k-1)}\right)$$

$$\mathbf{Z}^{(k)} \leftarrow \mathcal{P}_T\left(\mathbf{X}^{(k)}+\mathbf{U}^{(k-1)}\right)$$

$$\mathbf{U}^{(k)} \leftarrow \mathbf{U}^{(k-1)}+\mathbf{X}^{(k)}-\mathbf{Z}^{(k)}$$

Graph Regularized Pursuit

graph SC [Zheng et al. ’11]



$$\mathbf{X}^{(k)} \leftarrow (\mathbf{D}^T\mathbf{D} + \rho\mathbf{I})\mathbf{X} + \beta\mathbf{X}\mathbf{L_c} = \mathbf{D}^T\mathbf{Y} + \rho\left(\mathbf{Z}^{(k-1)} - \mathbf{U}^{(k-1)}\right)$$

$$\mathbf{Z}^{(k)} \leftarrow \mathcal{P}_T\left(\mathbf{X}^{(k)} + \mathbf{U}^{(k-1)}\right)$$

$$\mathbf{U}^{(k)} \leftarrow \mathbf{U}^{(k-1)} + \mathbf{X}^{(k)} - \mathbf{Z}^{(k)}$$

Theoretical Guarantees

Classical sparse theory:

$$(P_0^\epsilon) \qquad \arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \|\mathbf{y}-\mathbf{D}\mathbf{x}\|_2^2 \le \epsilon^2$$

Theorem: If the true representation $\mathbf{x}$ satisfies

$$\|\mathbf{x}\|_0 = s < \frac{1}{2}\left(1+\frac{1}{\mu(\mathbf{D})}\right)$$

then a solution $\hat{\mathbf{x}}$ for $(\mathbf{P}_0^\epsilon)$ must be close to it

$$\|\hat{\mathbf{x}}-\mathbf{x}\|_2^2 \le \frac{4\epsilon^2}{1-\delta_{2s}} \le \frac{4\epsilon^2}{1-(2s-1)\mu(\mathbf{D})}$$

Theoretical Guarantees

Graph sparse coding:

$$(P_{0,\infty}^{\epsilon}) \qquad \arg\min_{\mathbf{X}} \|\mathbf{X}\|_{0,\infty} \quad \text{s.t.} \quad \|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2 + \beta Tr(\mathbf{X}\mathbf{L_c}\mathbf{X}^T) \le \epsilon^2$$

Theorem: If the true representation $\mathbf{X}$ satisfies

$$\|\mathbf{X}\|_{0,\infty} = s < \frac{1}{2}\left(1 + \frac{1 + f(\beta, \mathbf{L_c})}{\mu(\mathbf{D})}\right)$$

then a solution $\widehat{\mathbf{X}}$ for $(\mathbf{P}_{\mathbf{0},\infty}^{\epsilon})$ must be close to it

$$\left\|\widehat{\mathbf{X}} - \mathbf{X}\right\|_F^2 \le \frac{4\epsilon^2}{1-\delta_{2s}} \le \frac{4\epsilon^2}{1-(2s-1)\mu(\mathbf{D}) + f(\beta, \mathbf{L_c})}$$

$$\ge 0$$

Back to DGRDL...

$$\arg\min_{\mathbf{D},\mathbf{X},\mathbf{L}} \|\mathbf{Y}-\mathbf{DX}\|_F^2 + \alpha Tr(\mathbf{D}^T\mathbf{L}\mathbf{D}) + \beta Tr(\mathbf{X}\mathbf{L_c}\mathbf{X}^T) + \mu\|\mathbf{L}\|_F^2$$

$$\text{s.t.} \quad \|x_i\|_0 \le T \quad \forall i$$

$$\mathbf{L}_{ij} = \mathbf{L}_{ji} \le 0 \quad (i \ne j)$$

$$\mathbf{L}\cdot\mathbf{1} = \mathbf{0}$$

$$Tr(\mathbf{L}) = \gamma N$$

dictionary atoms are smooth graph signals

similar signals have similar sparse codes

graph is adapted to promote smoothness

Graph Update

Initialize Dictionary

Sparse Code
Using ADMM pursuit

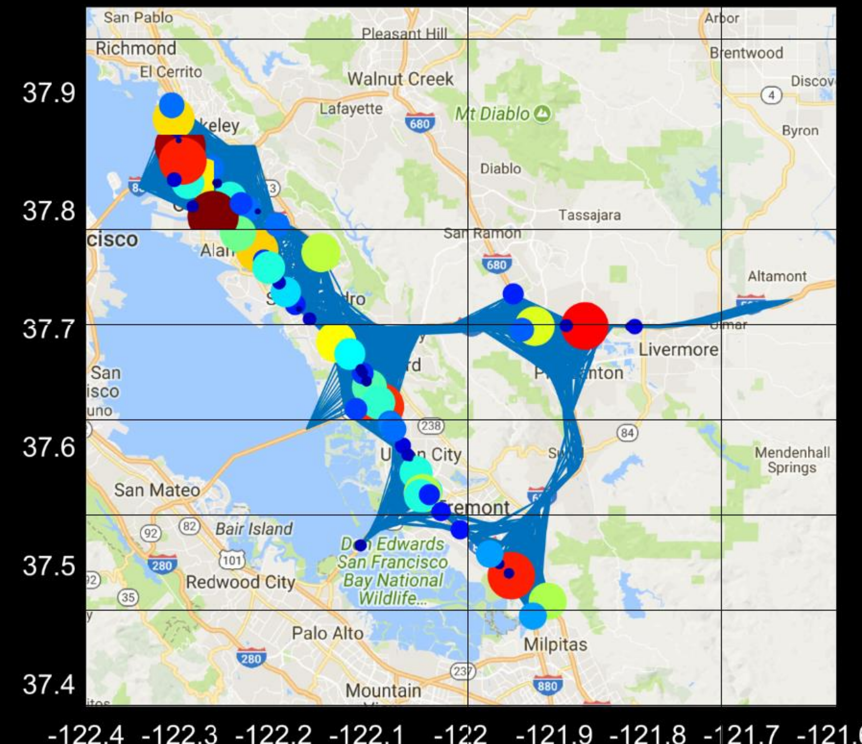
Dictionary Update
Atom-by-atom + coeffs. (modified update rule)

Results:
Network Data Recovery

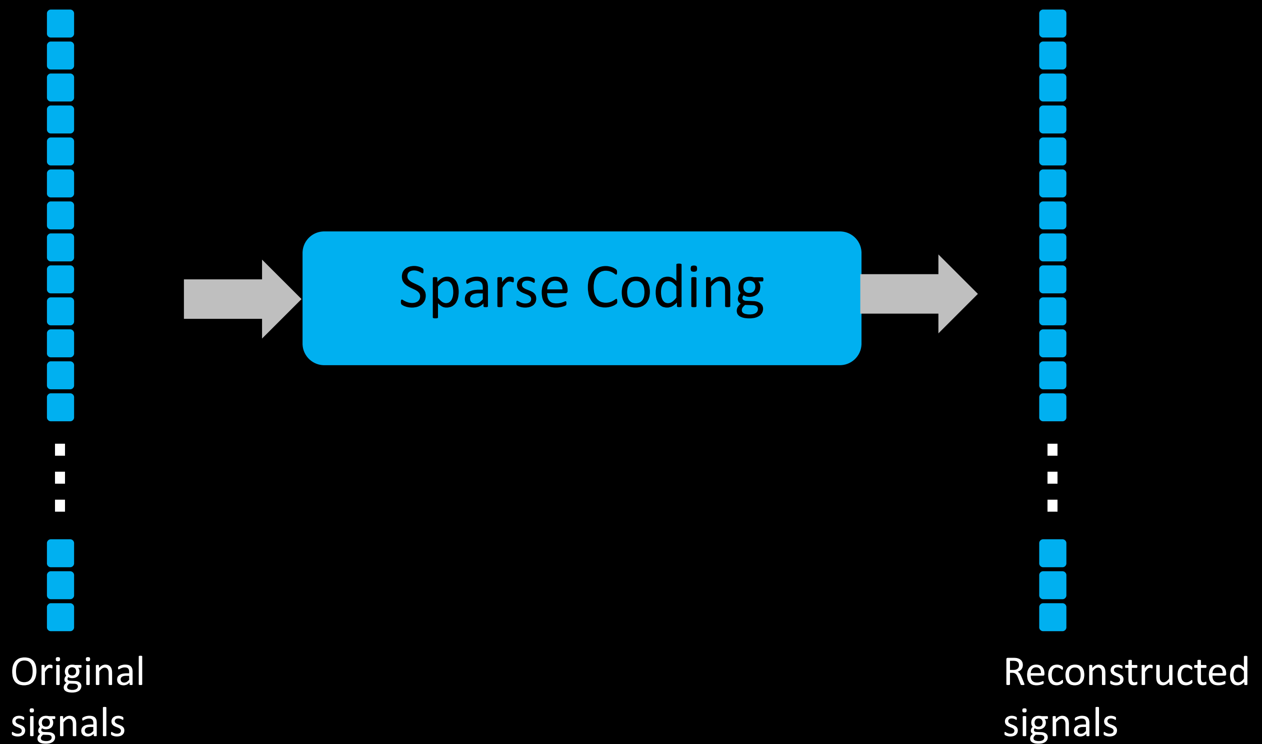
Traffic Dataset

Settings:

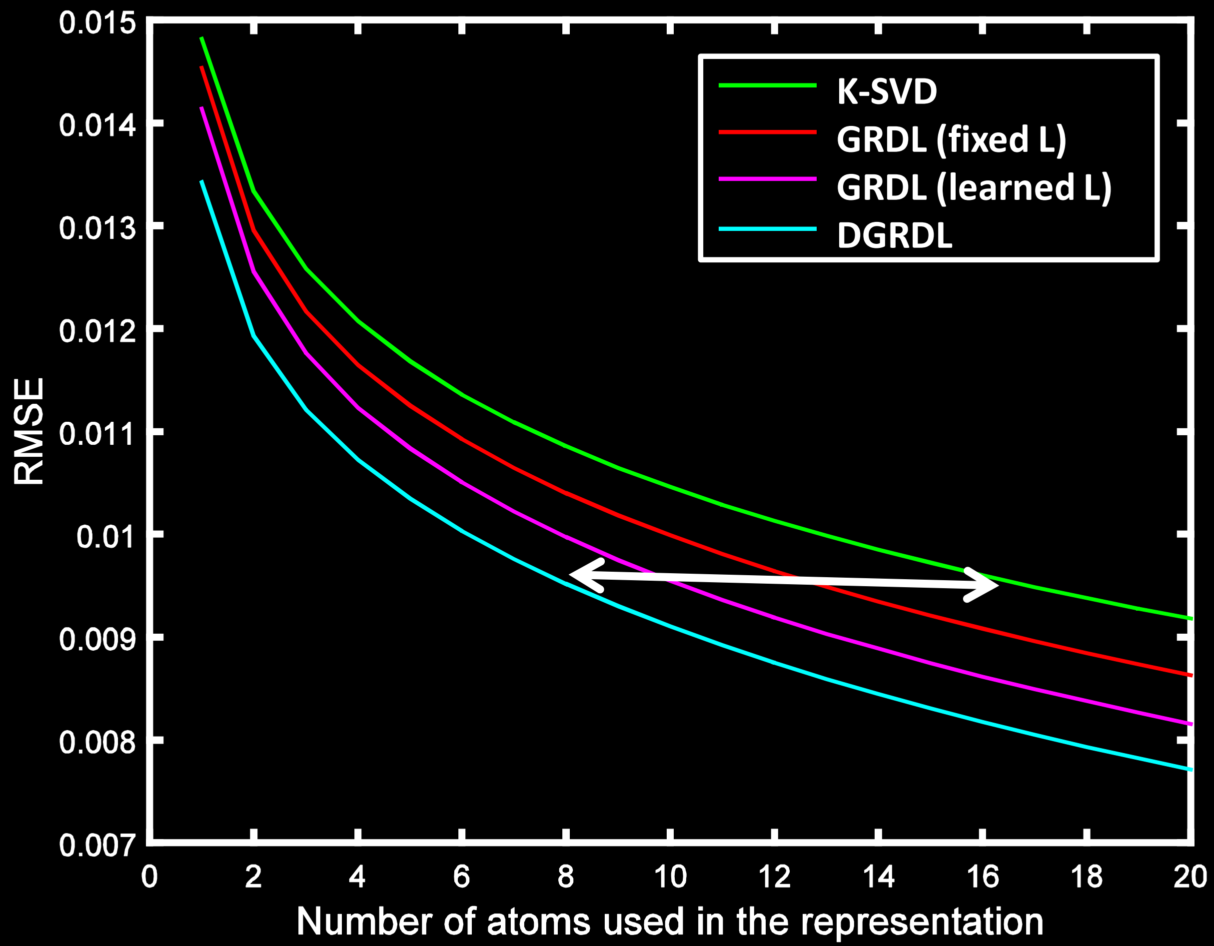
- N=578 sensors
- M=2892 signals
  - 1500 for training
  - 1392 for testing
- Graph signal = daily avg. bottleneck (min.) measured at each station



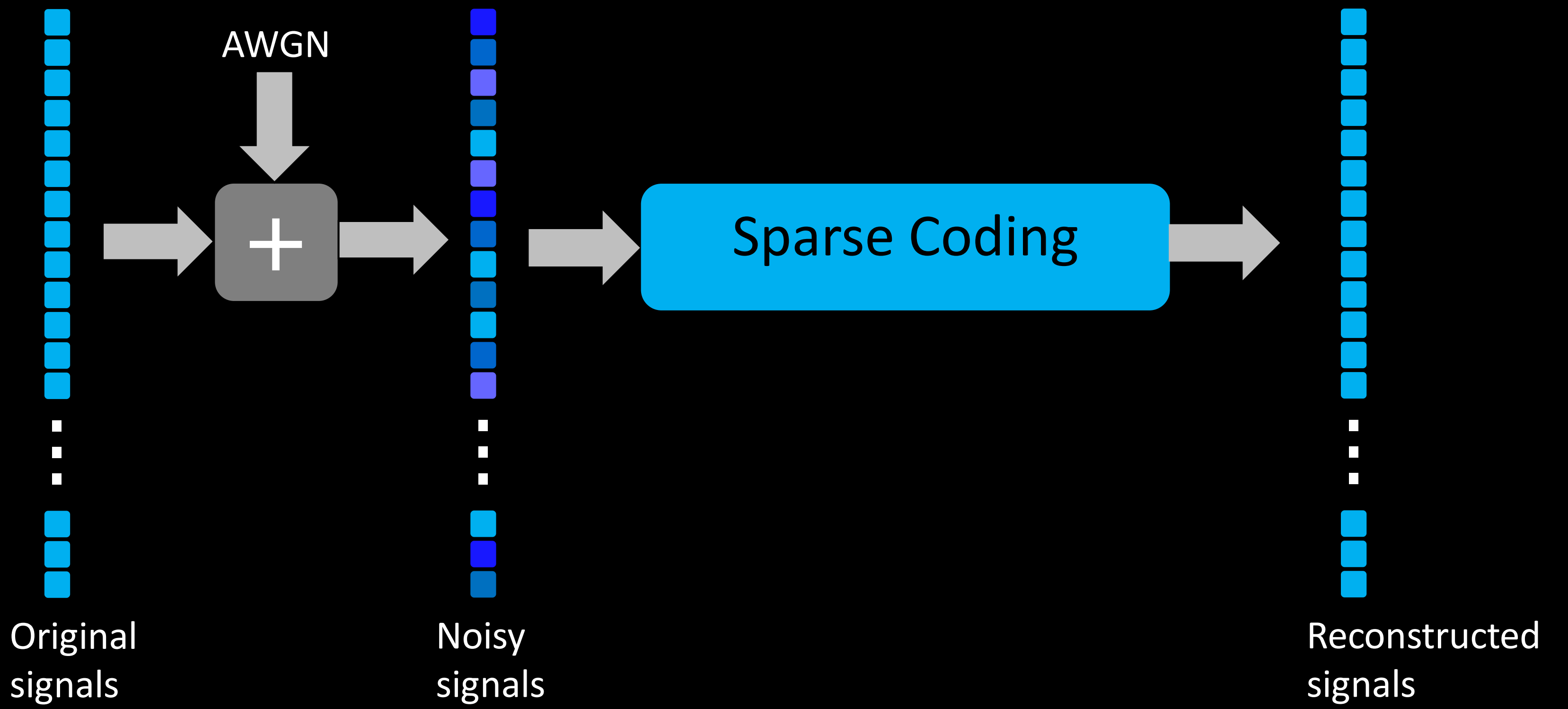
Representation



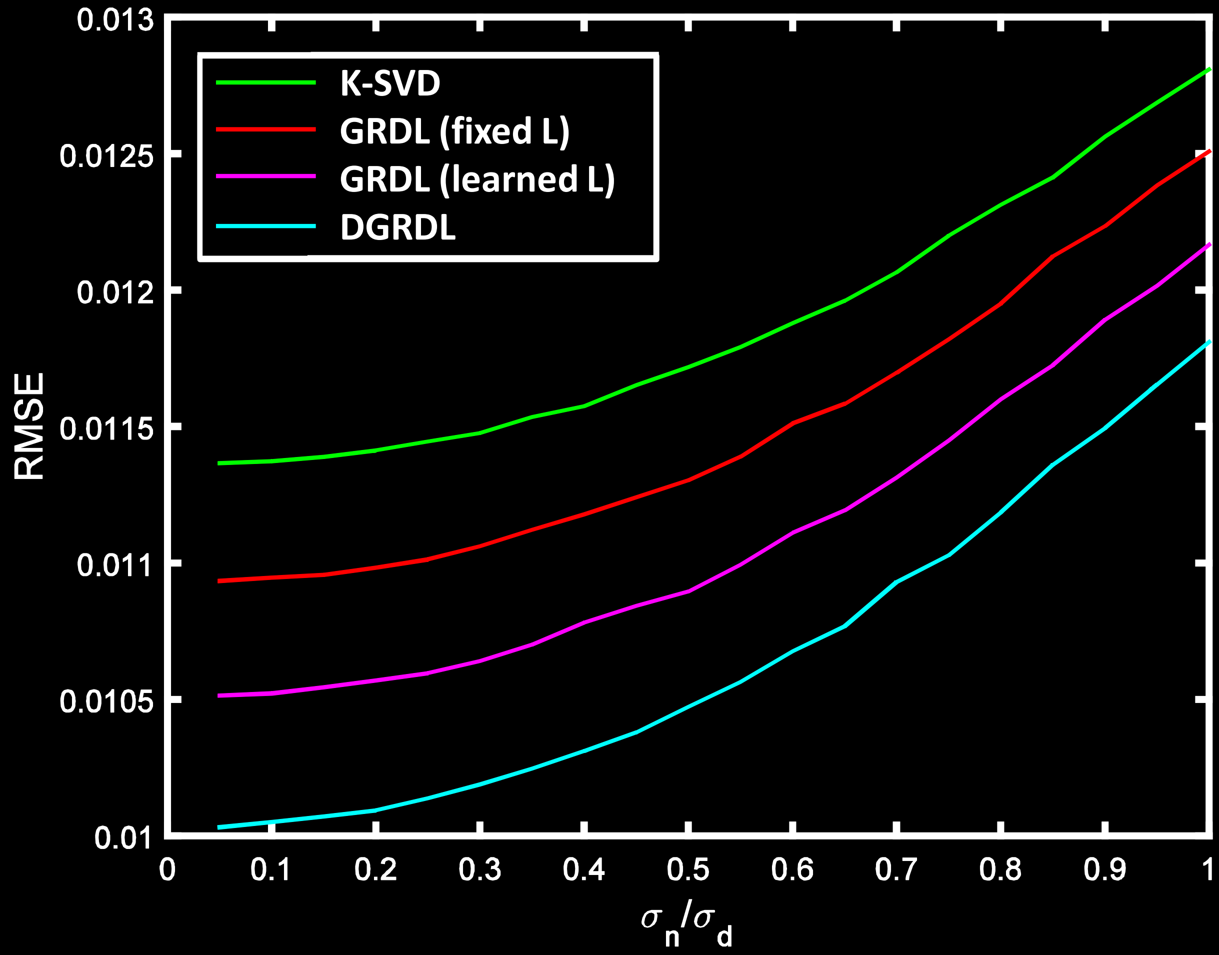
Representation



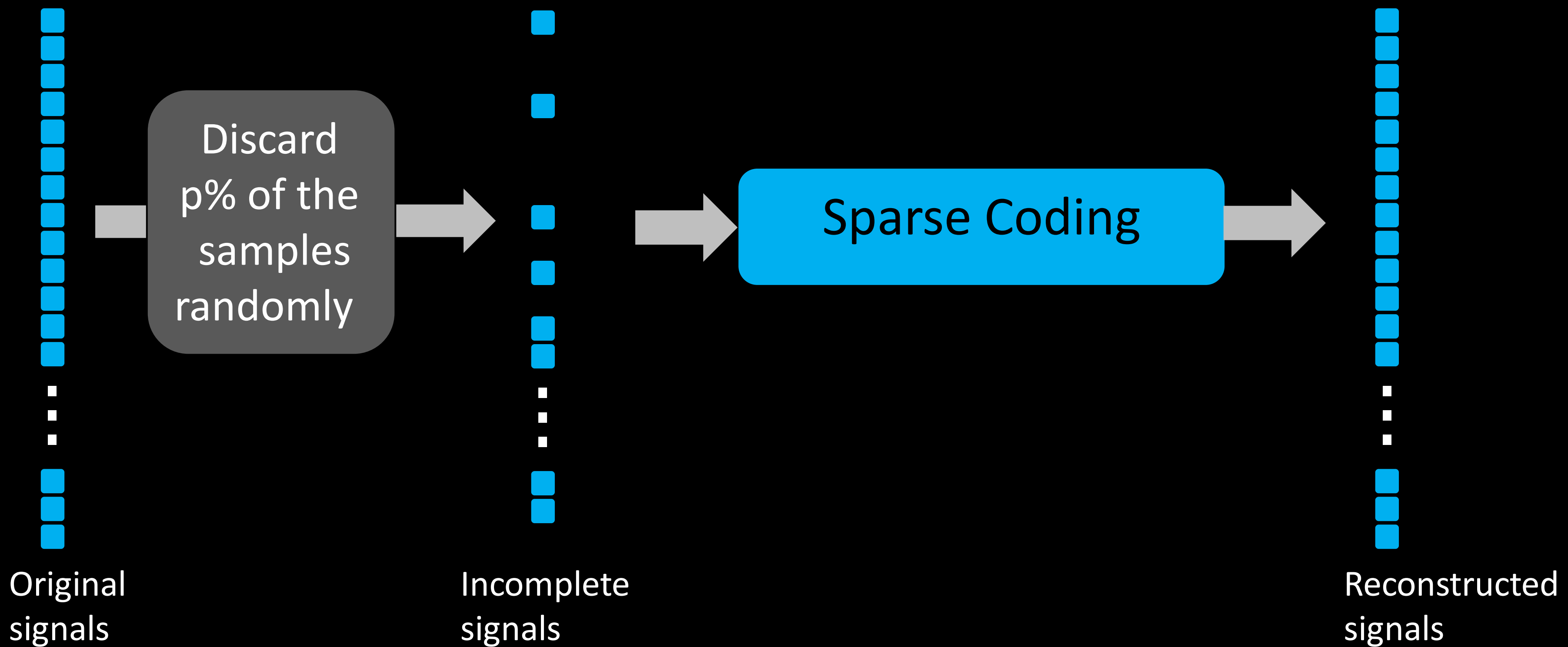
Denoising


AWGN
+
Sparse Coding
Original signals
Noisy signals
Reconstructed signals

Denoising

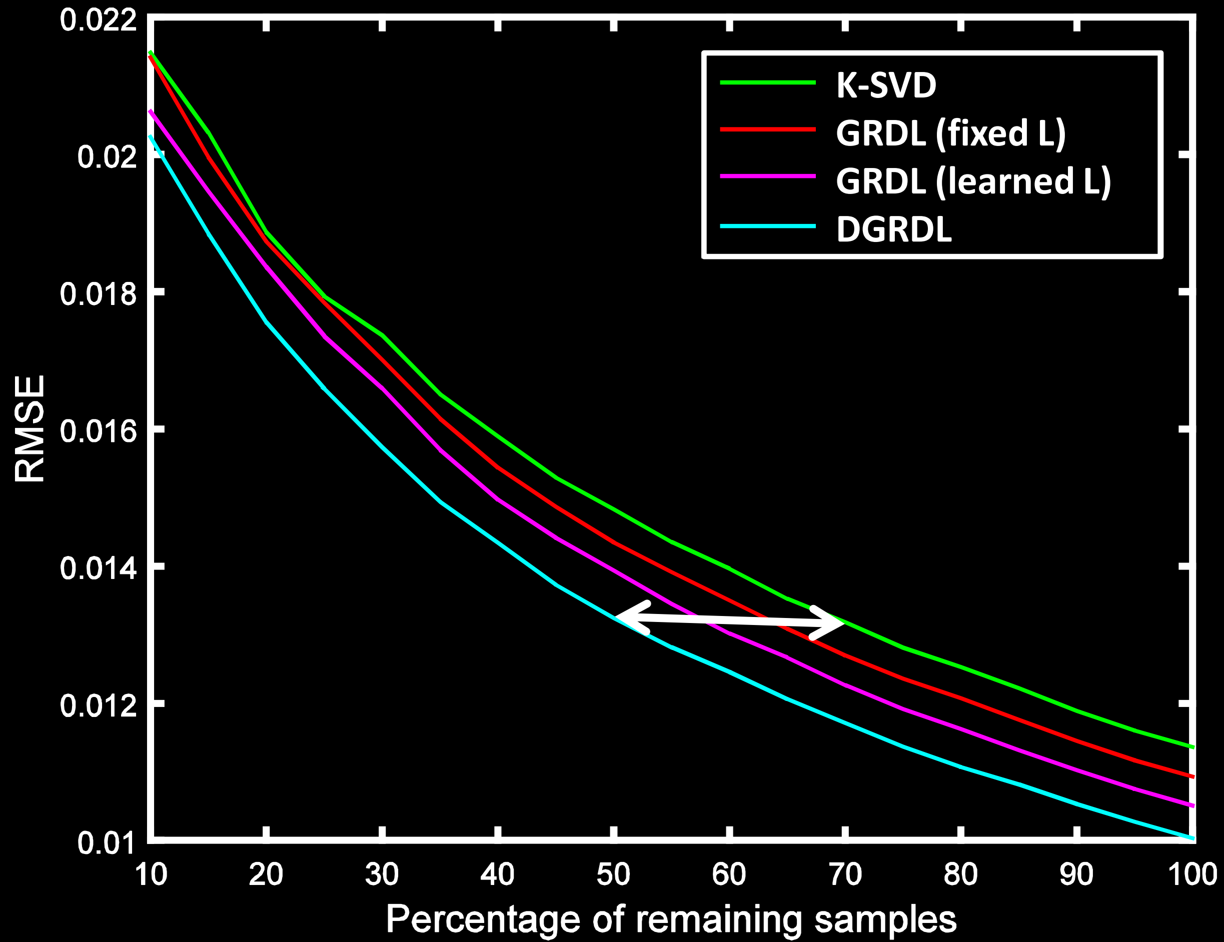

K-SVD
GRDL (fixed L)
GRDL (learned L)
DGRDL
RMSE
0.013
0.0125
0.012
0.0115
0.011
0.0105
0.01
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
σ_n/σ_d

Inpainting

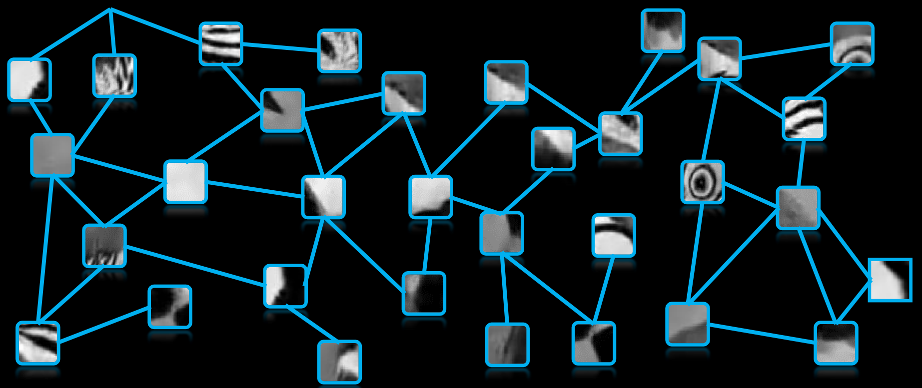


Discard p% of the samples randomly
Sparse Coding
Original signals
Incomplete signals
Reconstructed signals

Inpainting

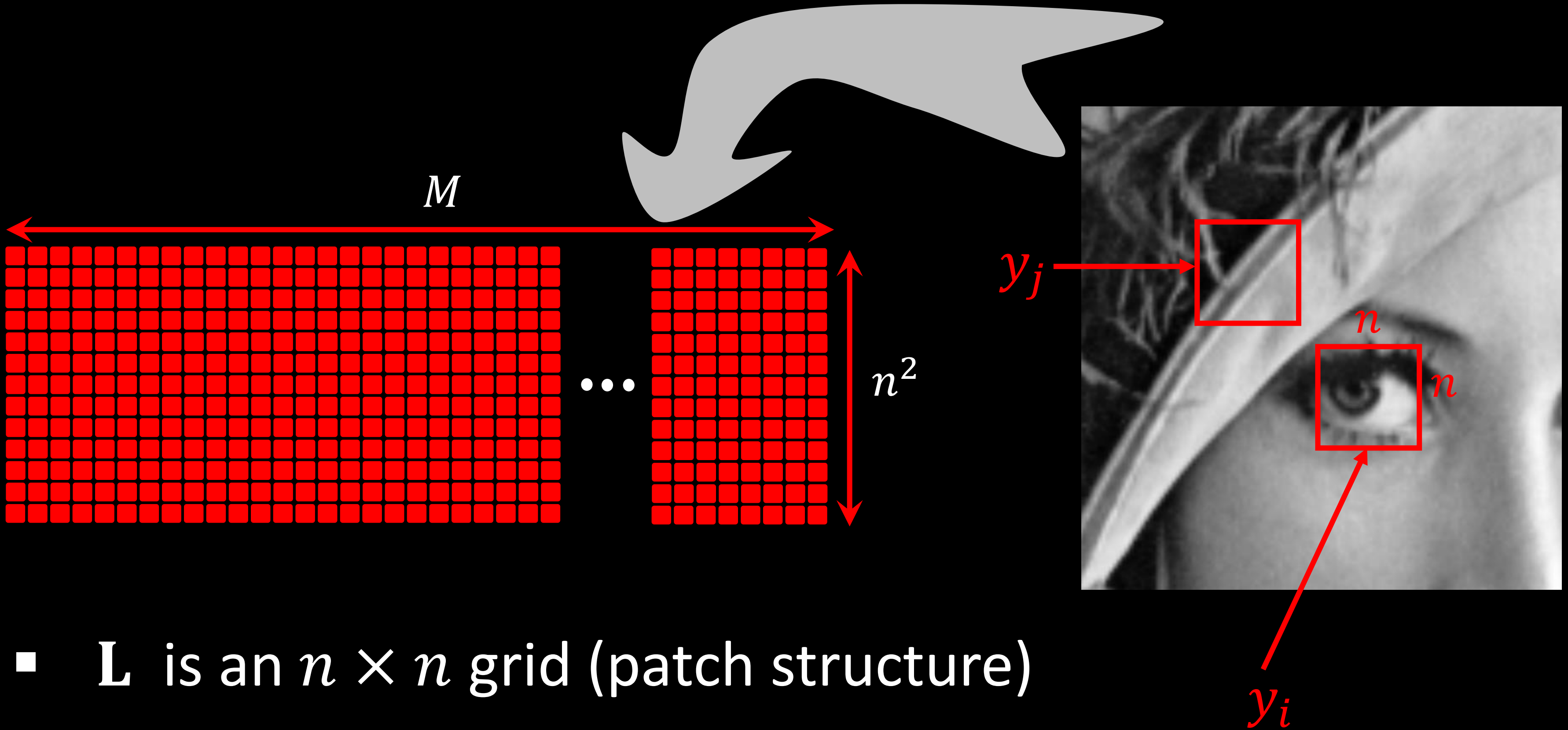


K-SVD
GRDL (fixed L)
GRDL (learned L)
DGRDL
RMSE
0.022
0.02
0.018
0.016
0.014
0.012
0.01
10
20
30
40
50
60
70
80
90
100
Percentage of remaining samples



Results:
Image Denoising Revisited

A Glimpse at Image Processing



- $\mathbf{L}$ is an $n \times n$ grid (patch structure)
- $\mathbf{D}$ is learned from only 1000 patches

Image Denoising (σ=25)

Original **Noisy (20.18dB)** **K-SVD (28.35dB)** **DGRDL (28.50dB)**

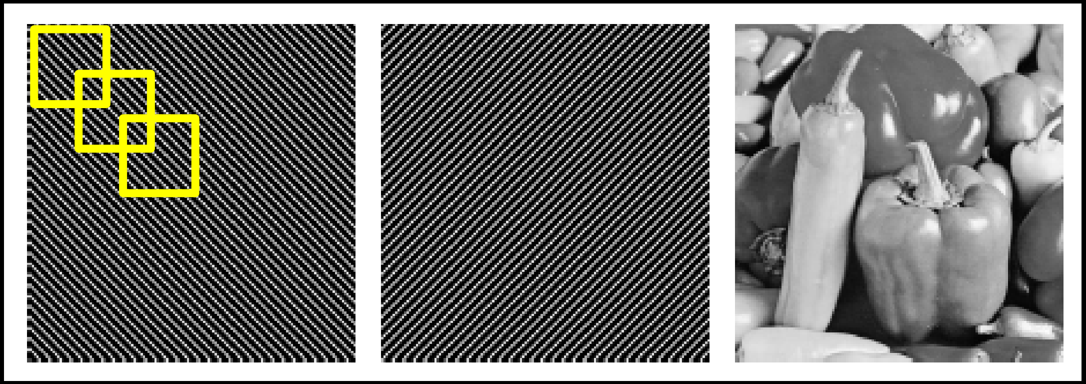
Original **Noisy (20.18dB)** **K-SVD (30.56dB)** **DGRDL (30.71dB)**

$$+0.15[dB]$$

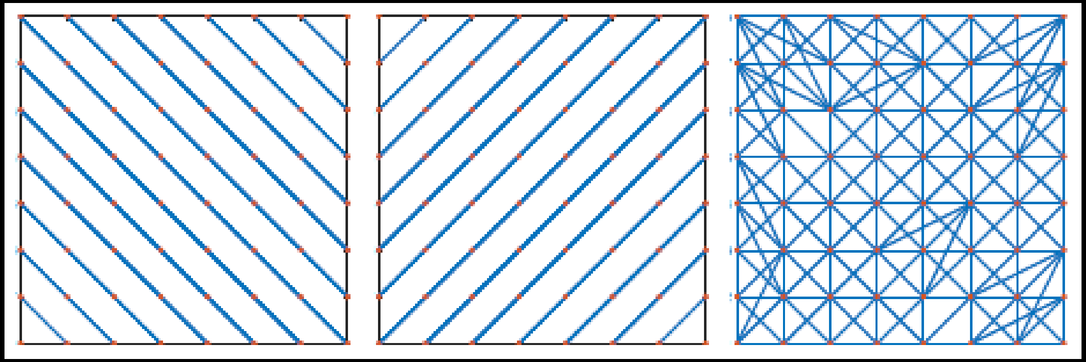
Structure Inference

Learn the underlying patch structure (pixel dependencies) from the data

input image



learned $\mathbf{L}$



Time to Conclude…

Processing data is enabled by an appropriate modeling that exposes its inner structure

We have shown how sparsity-based models become applicable also for graph structured data

We developed an efficient algorithm for joint learning of the dictionary and the graph

We demonstrated how various applications can benefit from the new model

Extensions include supervised dictionary learning and supporting high dimensions

Thank You