Sparse Modeling
in
Image Processing and Deep Learning

Michael Elad

Computer Science Department
The Technion - Israel Institute of Technology
Haifa 32000, Israel

New Deep Learning Techniques
February 5-9, 2018

The research leading to these results has been received funding from the European union's Seventh Framework Program (FP/2007-2013) ERC grant Agreement ERC-SPARSE- 320649

This Lecture

Sparseland
Sparse Representation Theory

→

CSC
Convolutional Sparse Coding

→

ML-CSC
Multi-Layered Convolutional Sparse Coding

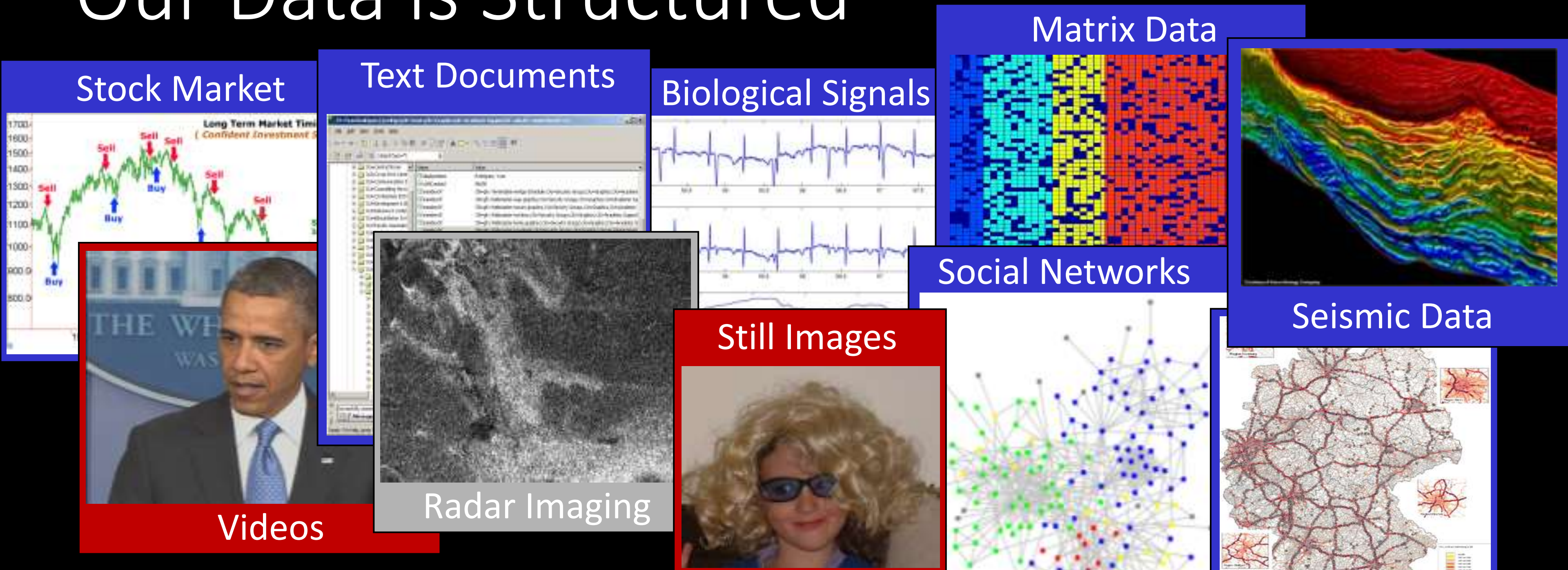
Sparsity-Inspired Models → Deep-Learning

Another underlying idea that will accompany us

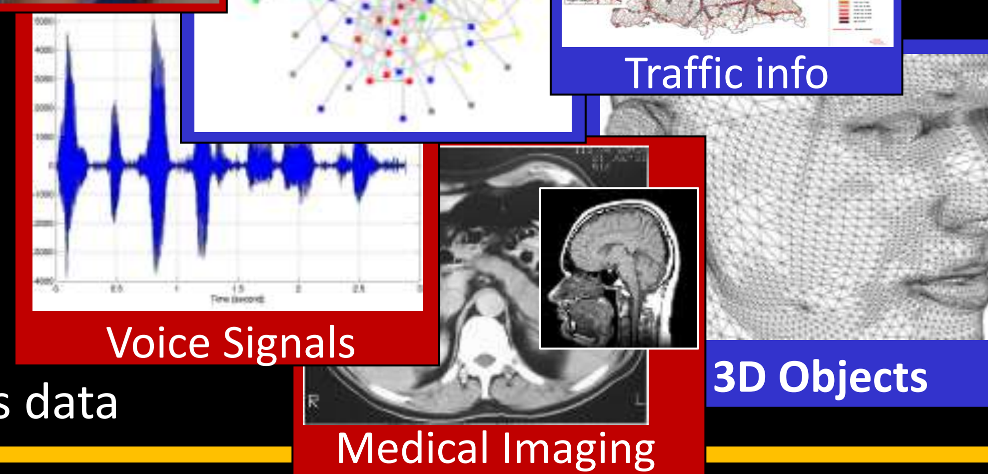
Generative modeling of data sources enables
- A systematic algorithm development, &
- A theoretical analysis of their performance

Multi-Layered Convolutional Sparse Modeling

Our Data is Structured



- We are surrounded by various diverse sources of massive information
- Each of these sources have an internal structure, which can be exploited
- This structure, when identified, is the engine behind our ability to process this data



Models

- A model: a mathematical description of the underlying signal of interest, describing our beliefs regarding its structure
- The following is a partial list of commonly used models for images
- Good models should be simple while matching the signals

Simplicity ⟷ Reliability

- Models are almost always imperfect

Principal-Component-Analysis
Gaussian-Mixture
Markov Random Field
Laplacian Smoothness
DCT concentration
Wavelet Sparsity
Piece-Wise-Smoothness
C2-smoothness
Besov-Spaces
Total-Variation
Beltrami-Flow

What This Talk is all About?

Data Models and Their Use

- Almost any task in data processing requires a model – true for denoising, deblurring, super-resolution, inpainting, compression, anomaly-detection, sampling, recognition, separation, and more
- Sparse and Redundant Representations offer a new and highly effective model – we call it

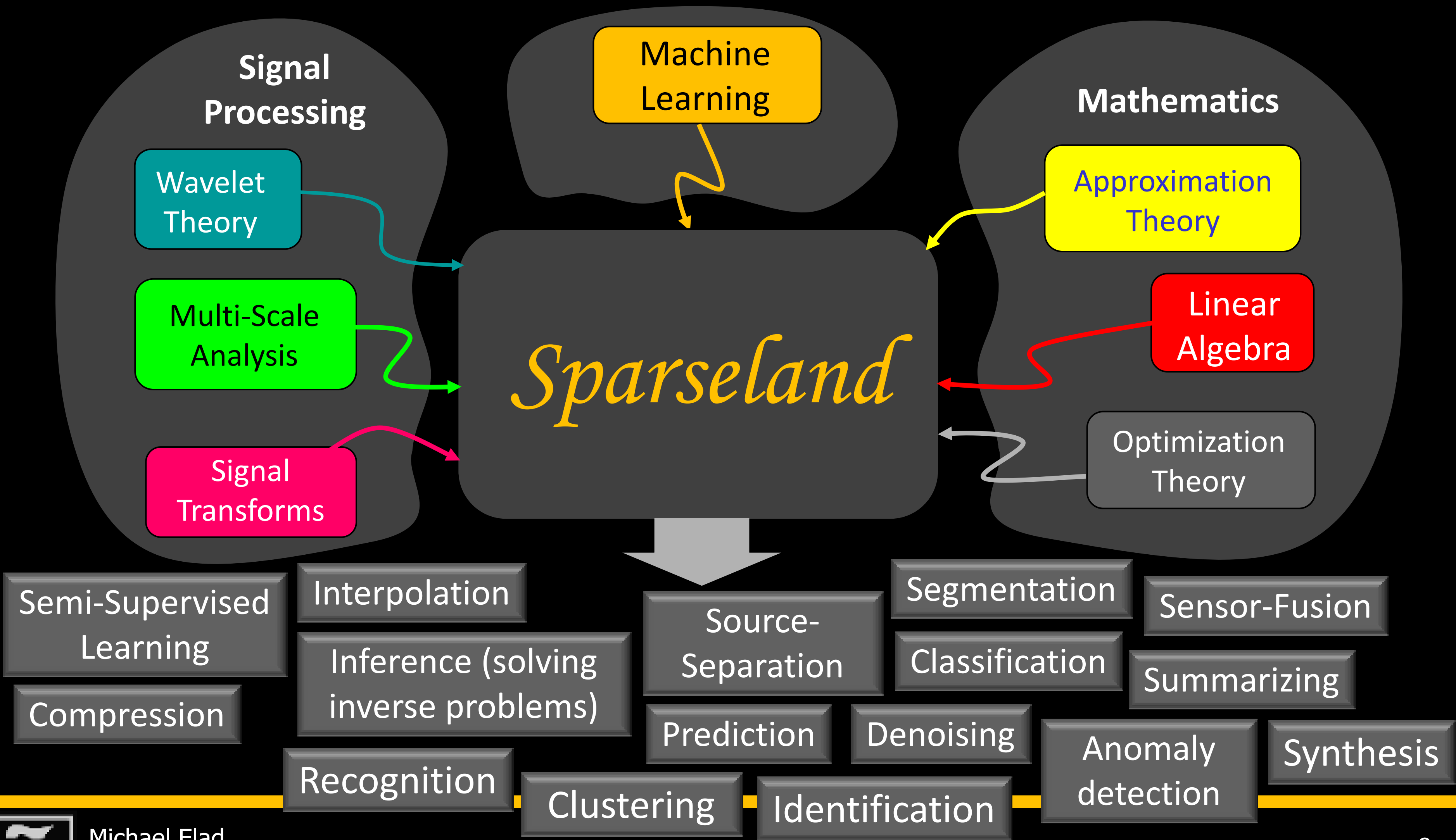
Sparseland

- We shall describe this and descendant versions of it that lead all the way to … deep-learning

Multi-Layered Convolutional

Sparse Modeling

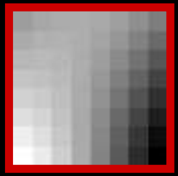
A New Emerging Model



Michael Elad
The Computer-Science Department
The Technion

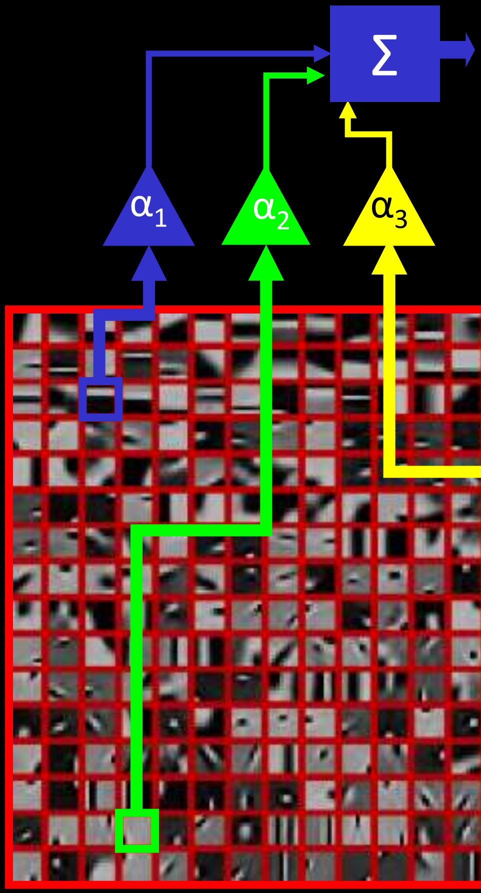
The *Sparseland* Model

- Task: model image patches of size 8×8 pixels



- We assume that a **dictionary** of such image patches is given, containing 256 **atom** images

- The *Sparseland* model assumption: every image patch can be described as a linear combination of **few** atoms



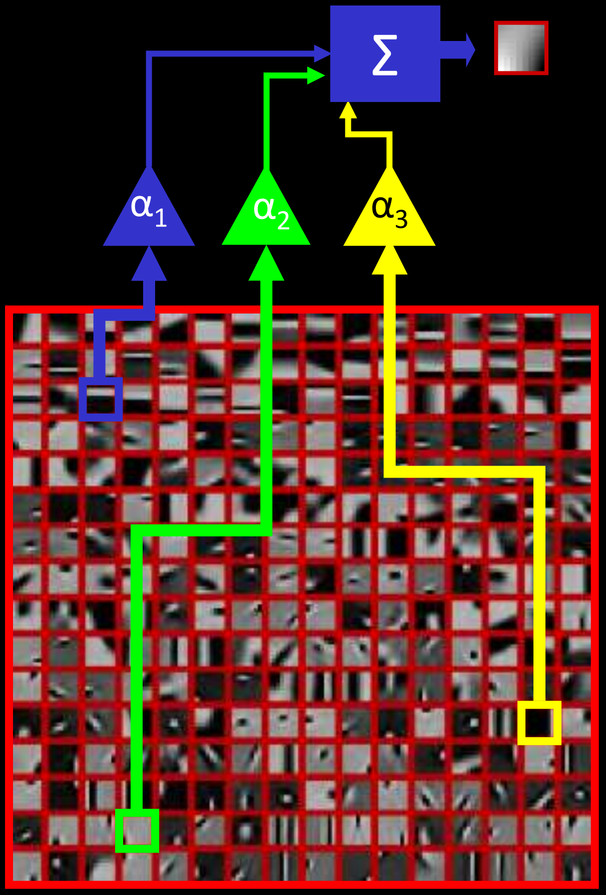
Michael Elad
The Computer-Science Department
The Technion

The *Sparseland* Model

Properties of this model:

Sparsity and Redundancy

- We start with a 8-by-8 pixels patch and represent it using 256 numbers
  – This is a redundant representation
- However, out of those 256 elements in the representation, only 3 are non-zeros
  – This is a sparse representation
- Bottom line in this case: 64 numbers representing the patch are replaced by 6 (3 for the indices of the non-zeros, and 3 for their entries)

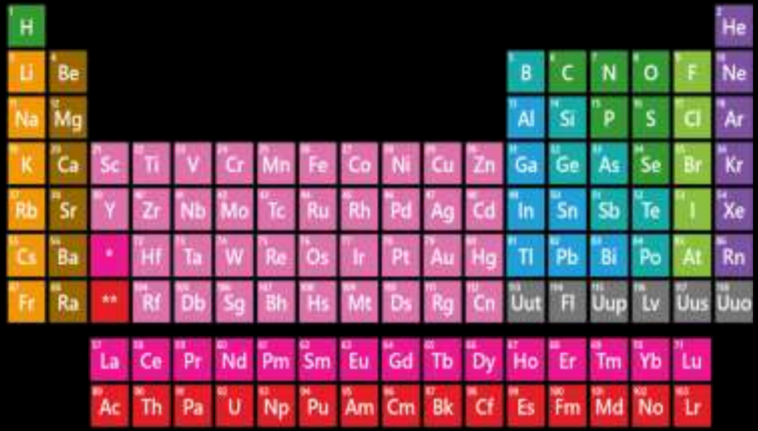


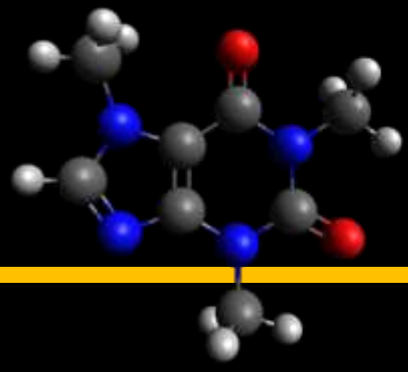
Michael Elad
The Computer-Science Department
The Technion

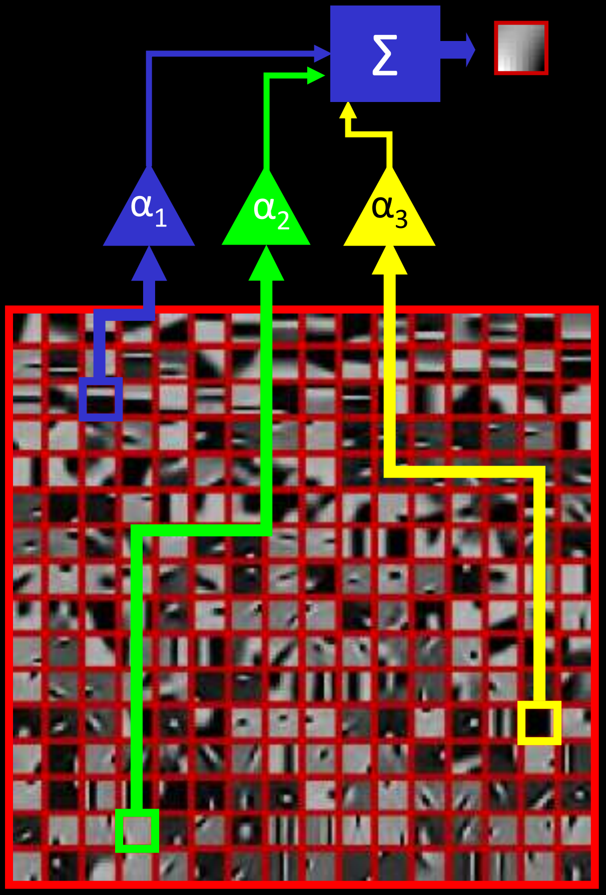
Chemistry of Data

We could refer to the *Sparseland* model as the chemistry of information:

- Our dictionary stands for the Periodic Table containing all the elements
- Our model follows a similar rationale: Every molecule is built of **few** elements

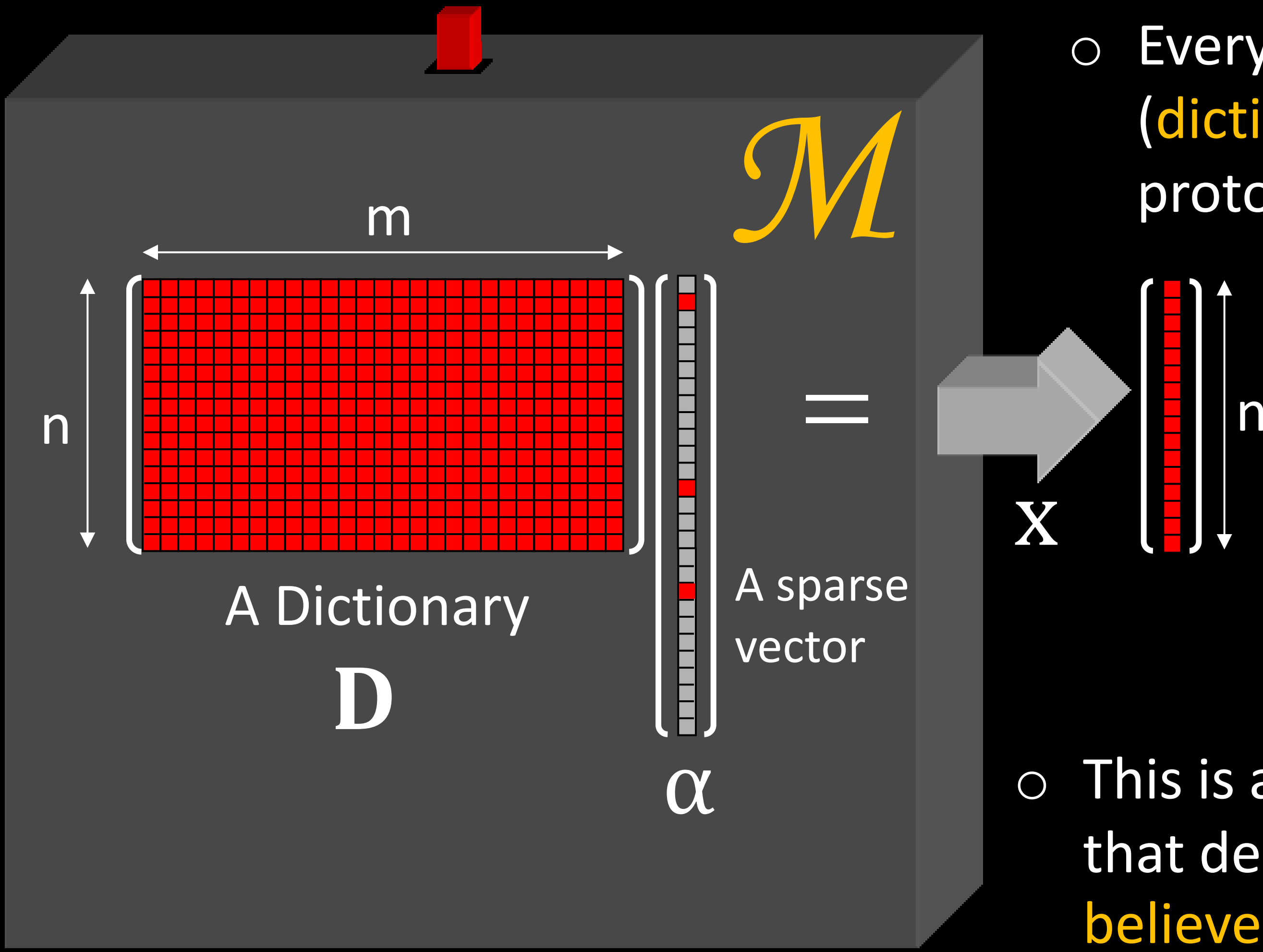






Michael Elad
The Computer-Science Department
The Technion

Sparseland : A Formal Description



- Every column in **D** (dictionary) is a prototype signal (atom)
  - The vector $\underline{\alpha}$ is generated with few non-zeros at arbitrary locations and values
- This is a generative model that describes how (we believe) signals are created

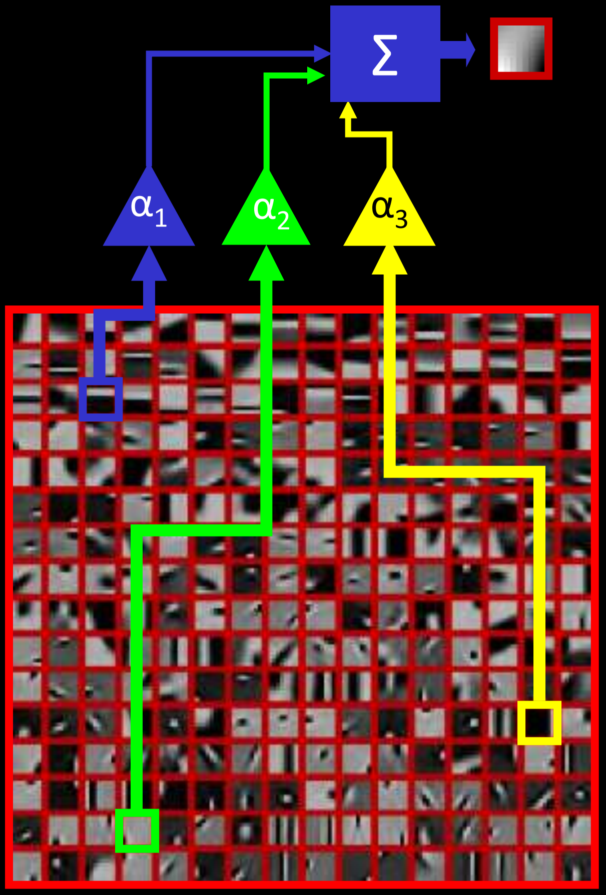
Michael Elad
The Computer-Science Department
The Technion

Difficulties with *Sparseland*

- Problem 1: Given a signal, how can we find its atom decomposition?
- A simple example:
  - There are 2000 atoms in the dictionary
  - The signal is known to be built of 15 atoms

$$\Rightarrow \binom{2000}{15} \approx 2.4e+37 \text{ possibilities}$$

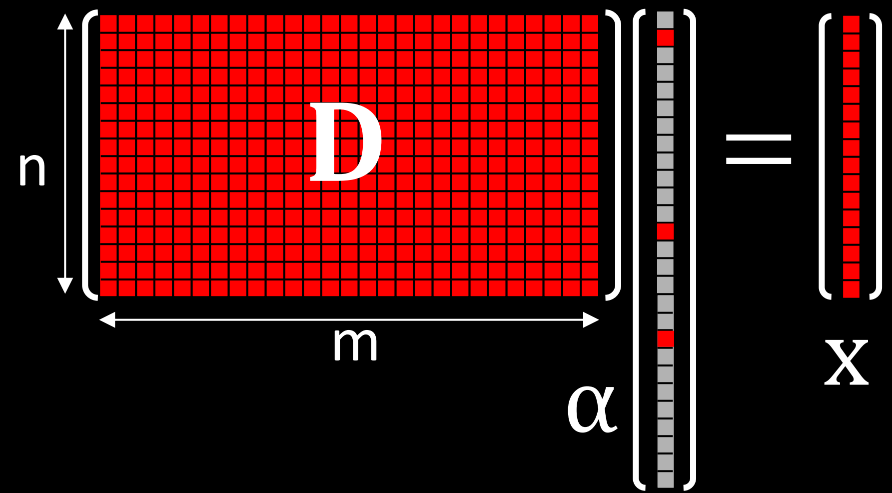
  - If each of these takes 1nano-sec to test, will take ~7.5e20 years to finish !!!!!!
- So, are we stuck?



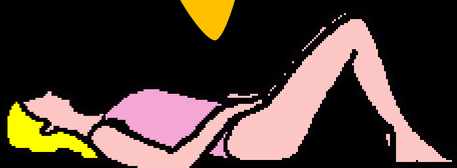
Atom Decomposition Made Formal

$$\min_\alpha \|\alpha\|_0 \;\; \text{s.t.} \;\; x = \mathbf{D}\alpha$$

$$\min_\alpha \|\alpha\|_0 \;\; \text{s.t.} \;\; \|\mathbf{D}\alpha - y\|_2 \le \varepsilon$$



Approximation Algorithms



Relaxation methods

Basis-Pursuit

Greedy methods

Thresholding/OMP

- L_0 – counting number of non-zeros in the vector
- This is a projection onto the *Sparseland* model
- These problems are known to be NP-Hard problem

Pursuit Algorithms

$$\min_{\alpha} \|\alpha\|_0 \;\; \text{s.t.} \;\; \|\mathbf{D}\alpha - y\|_2 \le \varepsilon$$

Approximation Algorithms

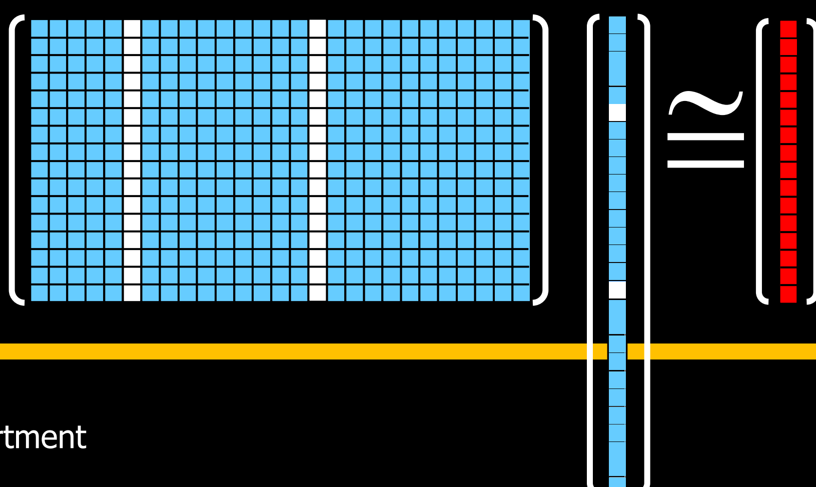
Basis Pursuit

Change the L_0 into L_1 and then the problem becomes convex and manageable

$$\min_{\alpha} \|\alpha\|_1 \;\; \text{s.t.} \;\; \|\mathbf{D}\alpha - y\|_2 \le \varepsilon$$

Matching Pursuit

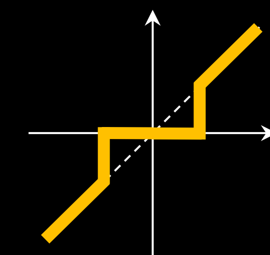
Find the support greedily, one element at a time



Thresholding

Multiply y by $\mathbf{D}^T$ and apply shrinkage:

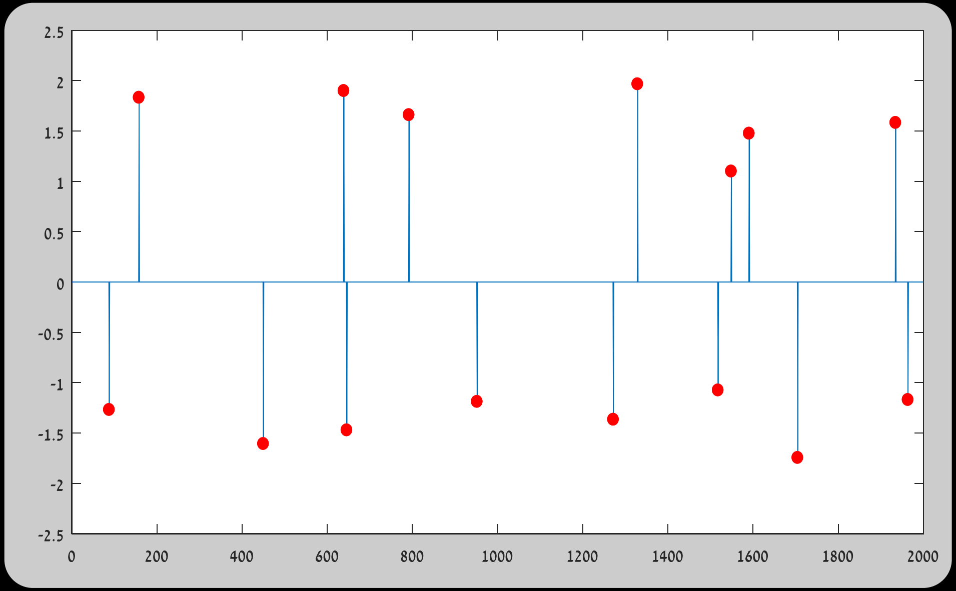
$$\hat{\alpha} = \mathcal{P}_{\beta}\{\mathbf{D}^T y\}$$



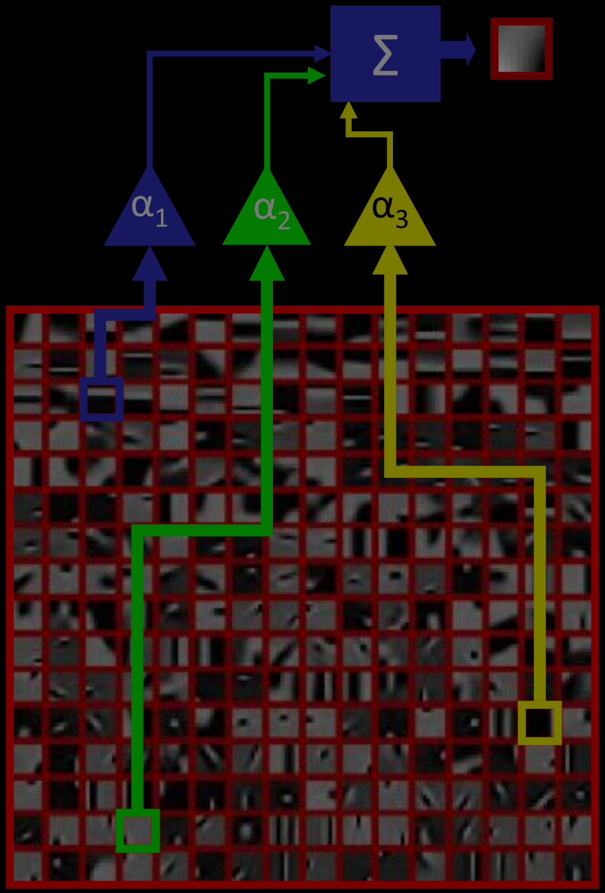
Michael Elad
The Computer-Science Department
The Technion

Difficulties with *Sparseland*

- There are various pursuit algorithms
- Here is an example using the Basis Pursuit (L_1):



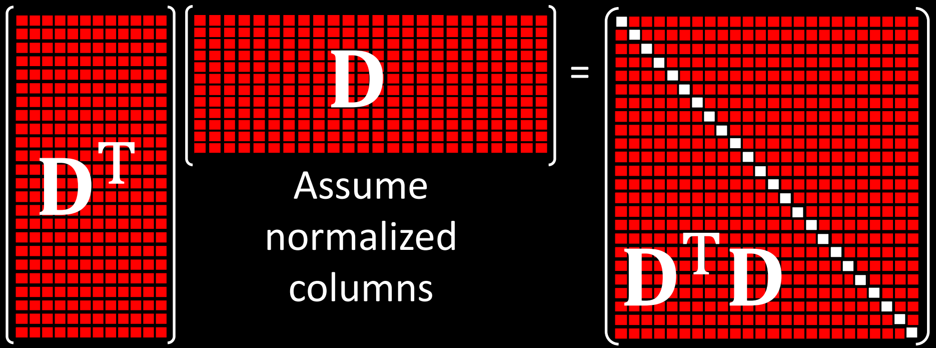
- Surprising fact: Many of these algorithms are often accompanied by theoretical guarantees for their success, if the unknown is sparse enough



Michael Elad
The Computer-Science Department
The Technion

The Mutual Coherence

- Compute



- The Mutual Coherence $\mu(\mathbf{D})$ is the largest off-diagonal entry in absolute value
- We will pose all the theoretical results in this talk using this property, due to its simplicity
- You may have heard of other ways to characterize the dictionary (Restricted Isometry Property - RIP, Exact Recovery Condition - ERC, Babel function, Spark, …)

Basis-Pursuit Success

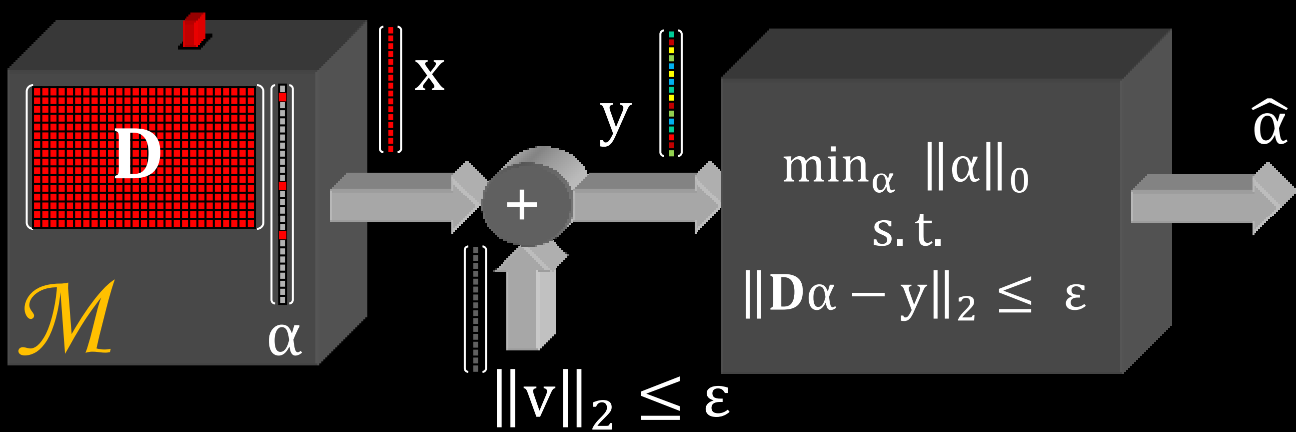
Theorem: Given a noisy signal $y = \mathbf{D}\alpha + v$ where $\|v\|_2 \le \varepsilon$ and α is sufficiently sparse,

$$\|\alpha\|_0 < \frac{1}{4}\left(1 + \frac{1}{\mu}\right)$$

then Basis-Pursuit: $\min_\alpha \ \|\alpha\|_1 \ \text{ s.t. } \ \|\mathbf{D}\alpha - y\|_2 \le \varepsilon$

leads to a stable result: $\|\hat{\alpha} - \alpha\|_2^2 \le \dfrac{4\varepsilon^2}{1-\mu(4\|\alpha\|_0 - 1)}$

Donoho, Elad & Temlyakov (‘06)



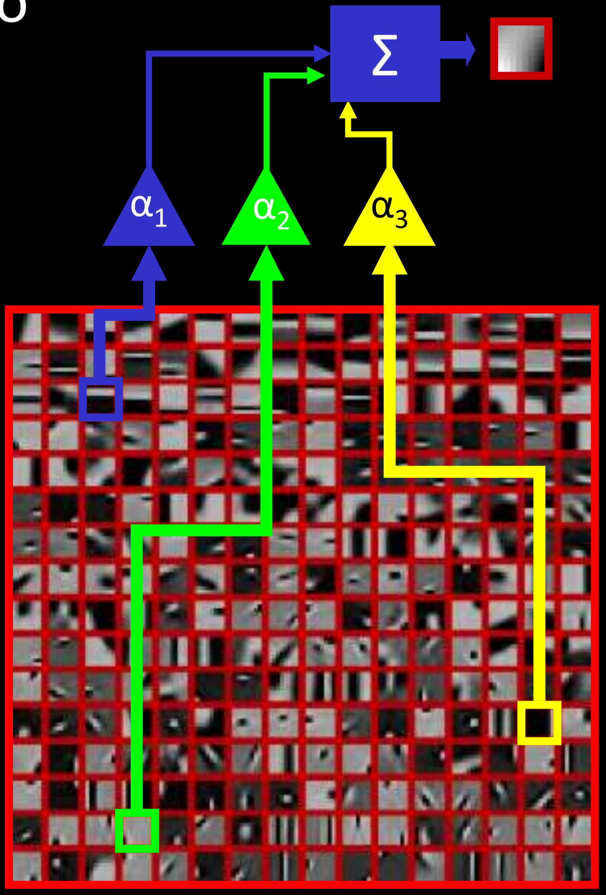
Comments:

- If $\varepsilon=0 \rightarrow \hat{\alpha} = \alpha$
- This is a worst-case analysis – better bounds exist
- Similar theorems exist for many other pursuit algorithms

Michael Elad
The Computer-Science Department
The Technion

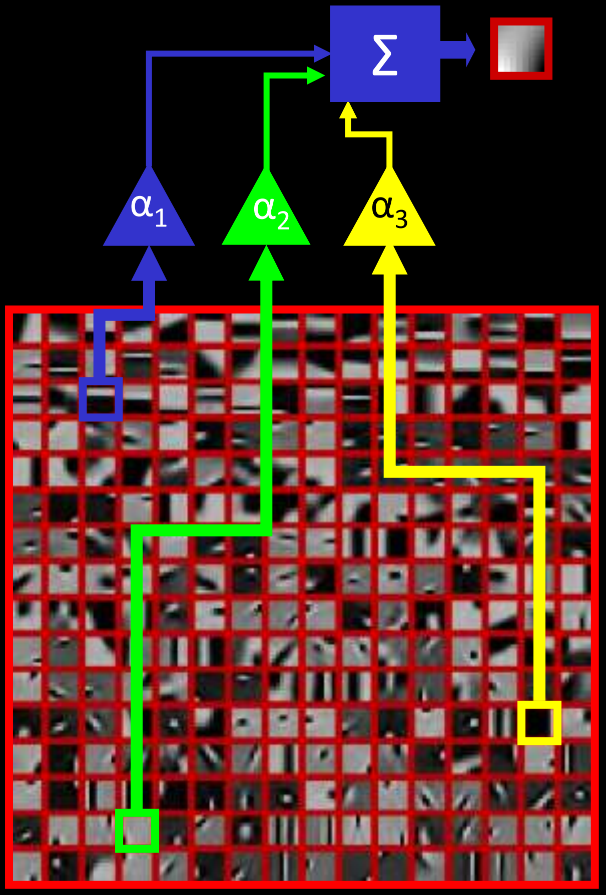
Difficulties with *Sparseland*

- Problem 2: Given a family of signals, how do we find the dictionary to represent it well?
- Solution: Learn! Gather a large set of signals (many thousands), and find the dictionary that sparsifies them
- Such algorithms were developed in the past 10 years (e.g., K-SVD), and their performance is surprisingly good
- We will not discuss this matter further in this talk due to lack of time



Difficulties with *Sparseland*

- Problem 3: Why is this model suitable to describe various sources? e.g., Is it good for images? Audio? Stocks? …
- General answer: Yes, this model is extremely effective in representing various sources
  - Theoretical answer: Clear connection to other models
  - Empirical answer: In a large variety of signal and image processing (and later machine learning), this model has been shown to lead to state-of-the-art results

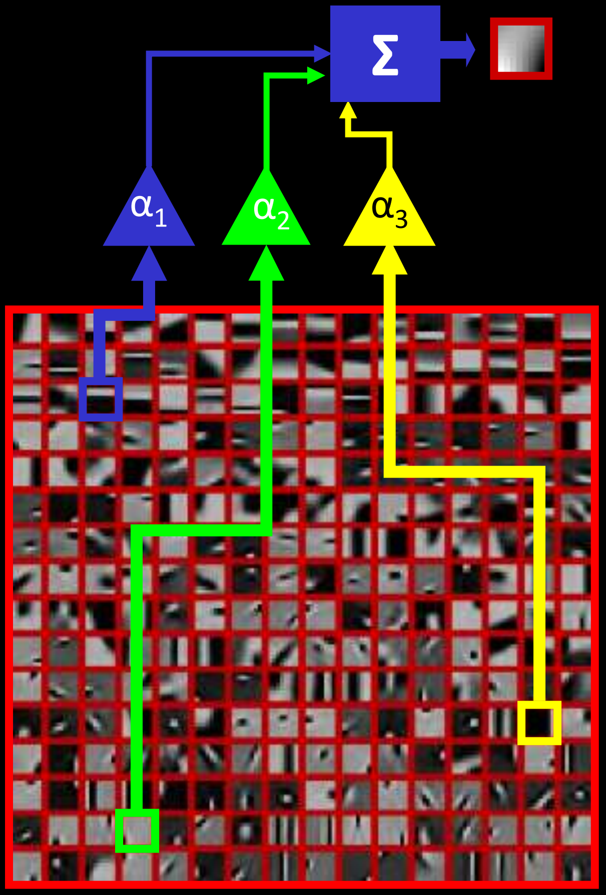


Michael Elad
The Computer-Science Department
The Technion

Difficulties with *Sparseland* ?

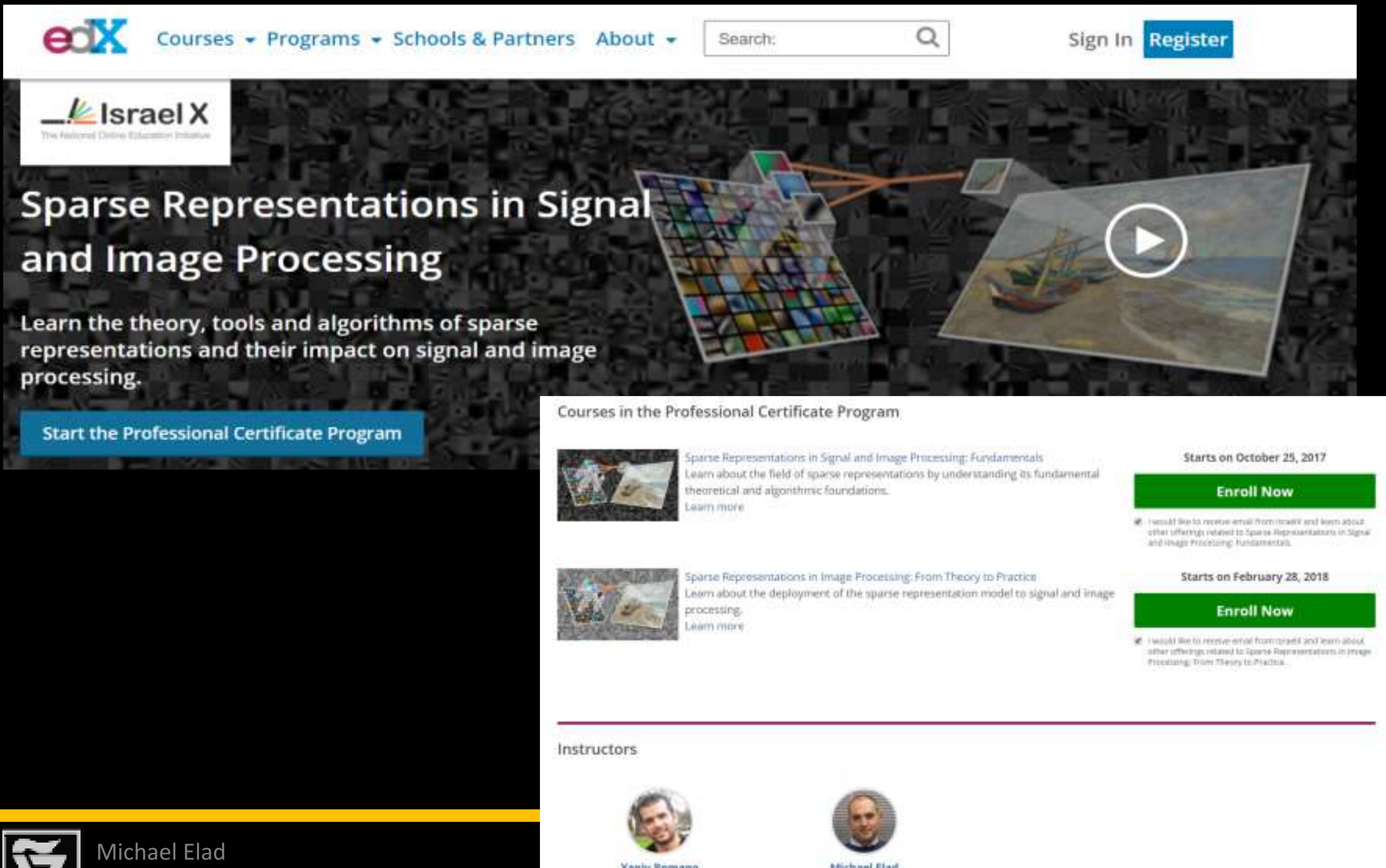
- Problem 1: Given an image patch, how can we find its atom decomposition ?
- Problem 2: Given a family of signals, how do we find the dictionary to represent it well?
- Problem 3: Is this model flexible enough to describe various sources? E.g., Is it good for images? audio? …

ALL ANSWERED POSITIVELY AND CONSTRUCTIVELY



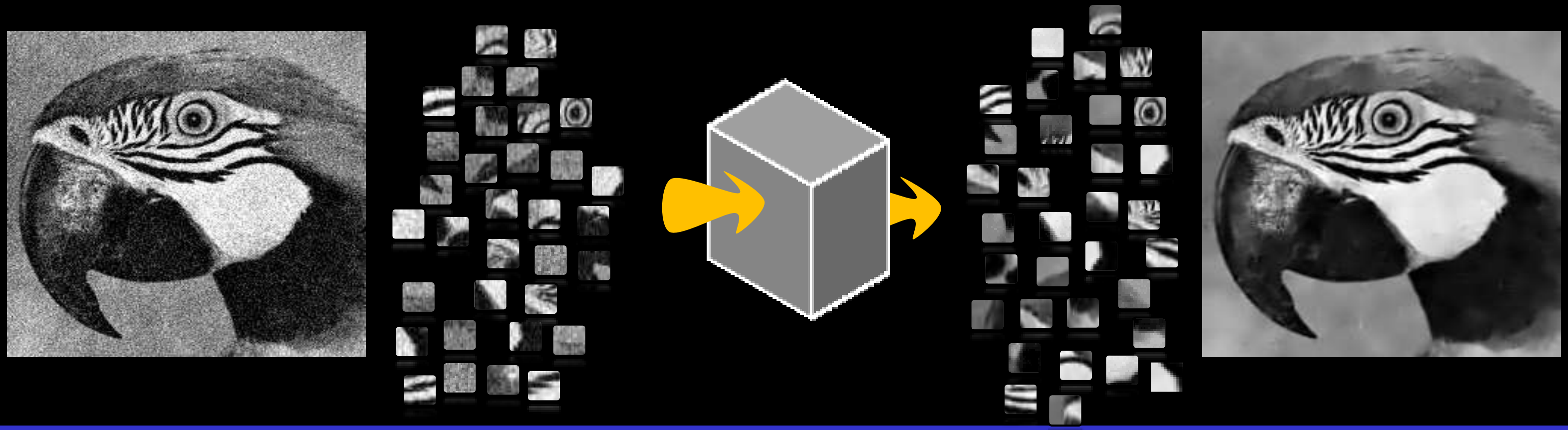
Michael Elad
The Computer-Science Department
The Technion

A New Massive Open Online Course



Sparseland for Image Processing

- When handling images, *Sparseland* is typically deployed on small overlapping patches due to the desire to train the model to fit the data better



- The model assumption is: each patch in the image is believed to have a sparse representation w.r.t. a common local dictionary

- What is the corresponding global model? This brings us to … the Convolutional Sparse Coding (CSC)

Michael Elad
The Computer-Science Department
The Technion

Multi-Layered Convolutional Sparse Modeling

Joint work with


Yaniv Romano


Vardan Papyan


Jeremias Sulam

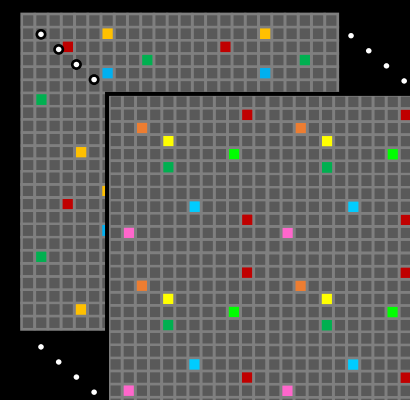
Convolutional Sparse Coding (CSC)

m filters convolved with their sparse representations

i-th feature-map: An image of the same size as $\mathbf{X}$ holding the sparse representation related to the i-filter

An image with N pixels

$$[\mathbf{X}] = \sum_{i=1}^{m} \mathrm{d}_i * [\Gamma_i]$$



The i-th filter of small size n

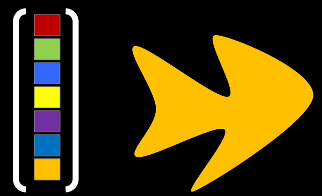
Michael Elad
The Computer-Science Department
The Technion

CSC in Matrix Form

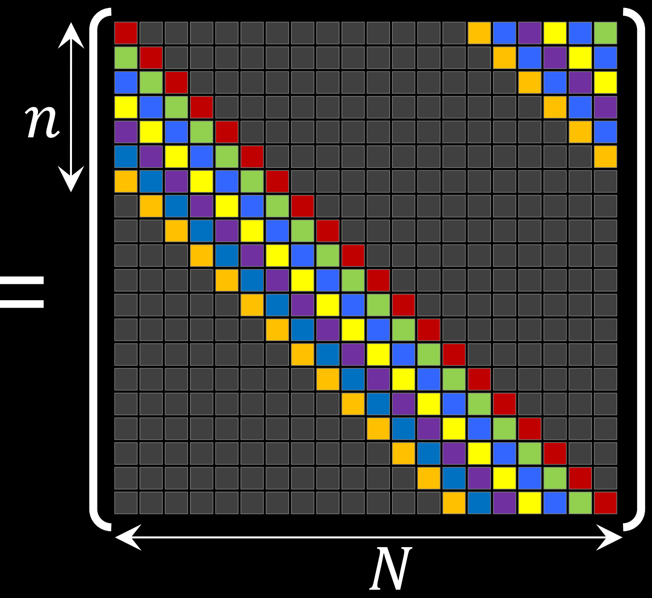
- Here is an alternative global sparsity-based model formulation

$$\mathbf{X} = \sum_{\mathrm{i}=1}^{m} \mathbf{C}^{\mathrm{i}}\boldsymbol{\Gamma}^{\mathrm{i}} = [\mathbf{C}^1 \cdots \mathbf{C}^m] \begin{bmatrix} \boldsymbol{\Gamma}^1 \\ \vdots \\ \boldsymbol{\Gamma}^m \end{bmatrix} = \mathbf{D}\boldsymbol{\Gamma}$$

- $\mathbf{C}^{\mathrm{i}} \in \mathbb{R}^{N \times N}$ is a banded and Circulant matrix containing a single atom with all of its shifts



$$\mathbf{C}^{\mathrm{i}} =$$

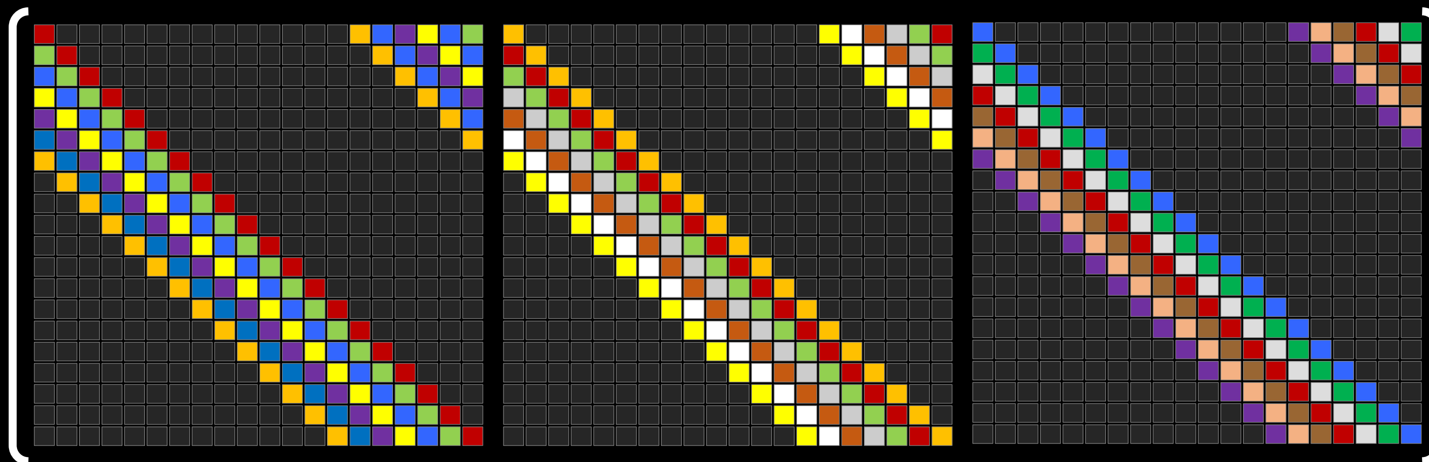


- $\boldsymbol{\Gamma}^{\mathrm{i}} \in \mathbb{R}^{N}$ are the corresponding coefficients ordered as column vectors

Michael Elad
The Computer-Science Department
The Technion

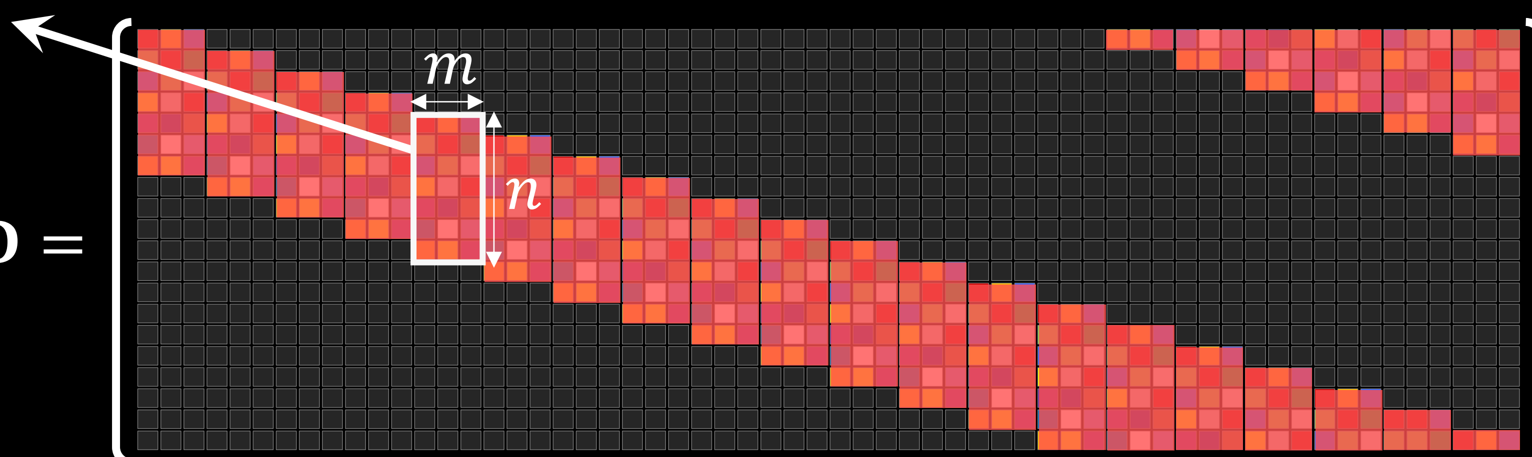
The CSC Dictionary

$[\mathbf{C}^1 \ \mathbf{C}^2 \ \mathbf{C}^3] =$



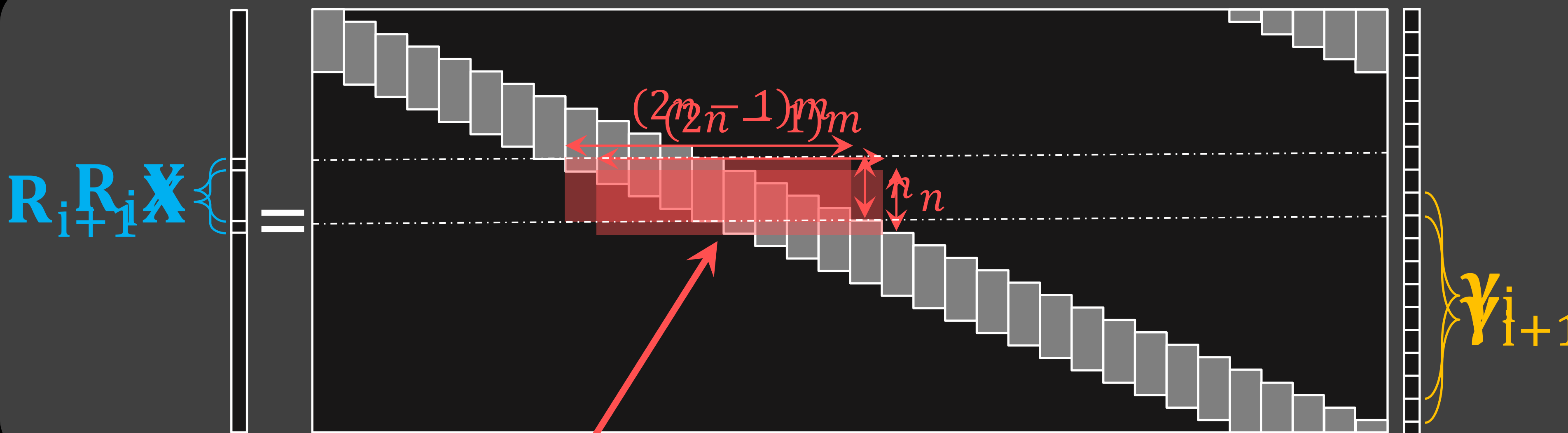
$\mathbf{D}_\mathrm{L}$

$\mathbf{D} =$



Michael Elad
The Computer-Science Department
The Technion

Why CSC?



stripe-dictionary

stripe vector

$$\mathbf{X} = \mathbf{D\Gamma}$$

$$\mathbf{R}_i\mathbf{X} = \boldsymbol{\Omega}\boldsymbol{\gamma}_i$$

$$\mathbf{R}_{i+1}\mathbf{X} = \boldsymbol{\Omega}\boldsymbol{\gamma}_{i+1}$$

Every patch has a sparse representation w.r.t. to the same local dictionary ($\boldsymbol{\Omega}$) just as assumed for images

Michael Elad
The Computer-Science Department
The Technion

Classical Sparse Theory for CSC ?

$$\min_{\mathbf{\Gamma}} \ \|\mathbf{\Gamma}\|_0 \ \ \text{s.t.} \ \|\mathbf{Y}-\mathbf{D\Gamma}\|_2 \leq \varepsilon$$

Theorem: **BP is guaranteed to "succeed" …. if** $\|\mathbf{\Gamma}\|_0 < \frac{1}{4}\left(1+\frac{1}{\mu}\right)$

- Assuming that $m = 2$ and $n = 64$ we have that [Welch, '74]

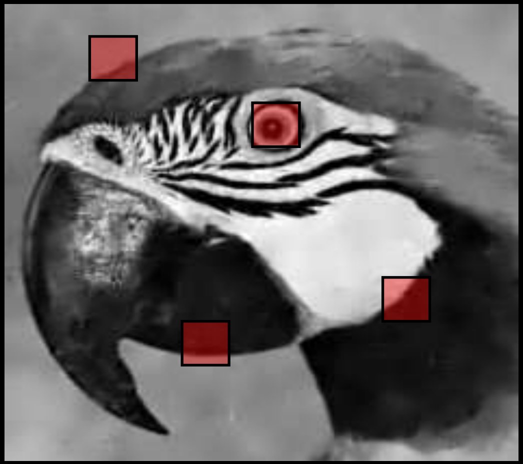
$$\mu \geq 0.063$$

- Success of pursuits is …

2

- O… zeros GLOBALLY are allowed!!! This is a very pessimistic result!

The classic *Sparseland* Theory does not provide good explanations for the CSC model

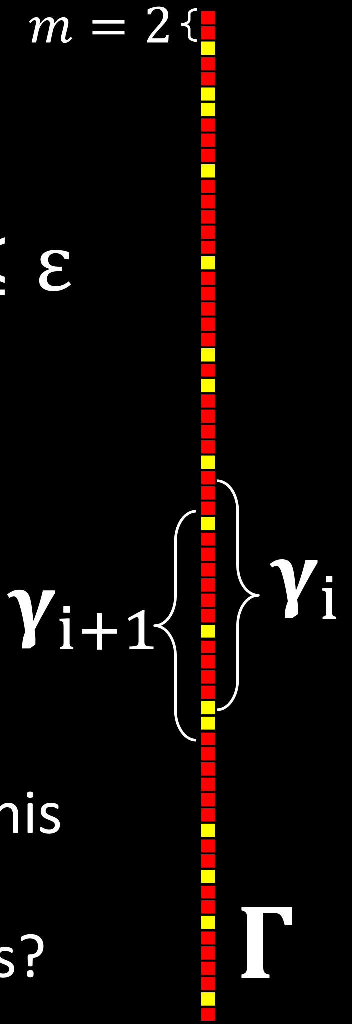


Moving to Local Sparsity: Stripes

$\ell_{0,\infty}$ Norm: $\|\boldsymbol{\Gamma}\|_{0,\infty}^{s} = \max_{i} \|\boldsymbol{\gamma}_i\|_0$

$$\min_{\boldsymbol{\Gamma}} \|\boldsymbol{\Gamma}\|_{0,\infty}^{s} \text{ s.t. } \|\mathbf{Y} - \mathbf{D}\boldsymbol{\Gamma}\|_2 \leq \varepsilon$$

$\|\boldsymbol{\Gamma}\|_{0,\infty}^{s}$ is low $\rightarrow$ all $\boldsymbol{\gamma}_i$ are sparse $\rightarrow$ every patch has a sparse representation over $\boldsymbol{\Omega}$



The main question we aim to address is this:

Can we generalize the vast theory of *Sparseland* to this new notion of local sparsity? For example, could we provide guarantees for success for pursuit algorithms?

Michael Elad
The Computer-Science Department
The Technion

Success of the Basis Pursuit

$$\Gamma_{BP} = \min_{\Gamma} \frac{1}{2}\|Y - D\Gamma\|_2^2 + \lambda\|\Gamma\|_1$$

Local noise (per patch)

Theorem: For $Y = D\Gamma + E$, if $\lambda = 4\|E\|_{2,\infty}^{p}$, **if**

$$\|\boldsymbol{\Gamma}\|_{0,\infty}^{s} < \frac{1}{3}\left(1 + \frac{1}{\mu(D)}\right)$$

then Basis Pursuit performs very-well:

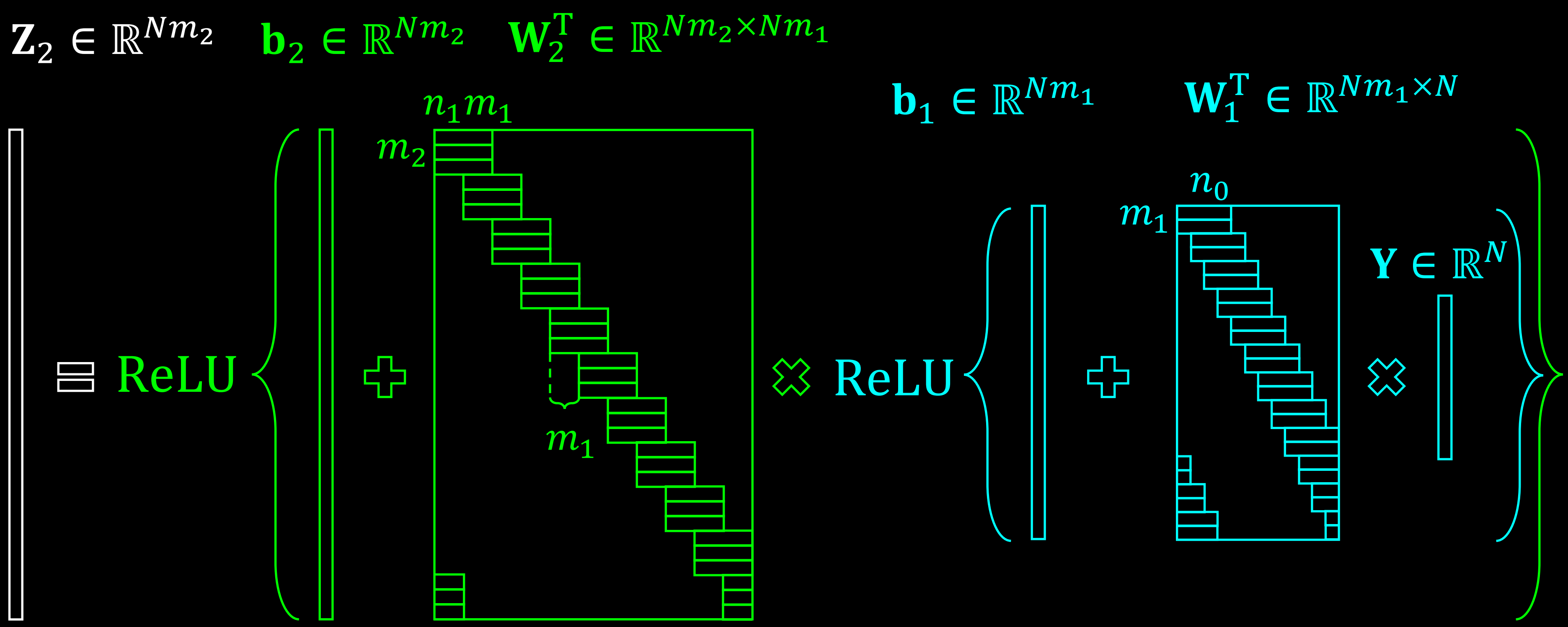
1. The support of Γ_{BP} is contained in that of Γ
2. $\|\Gamma_{BP} - \Gamma\|_{\infty} \leq 7.5\|E\|_{2,\infty}^{p}$
3. Every entry greater than $7.5\|E\|_{2,\infty}^{p}$ is found
4. Γ_{BP} is unique

Papyan, Sulam & Elad ('17)

Multi-Layered Convolutional Sparse Modeling

Quick Recall: The Forward Pass

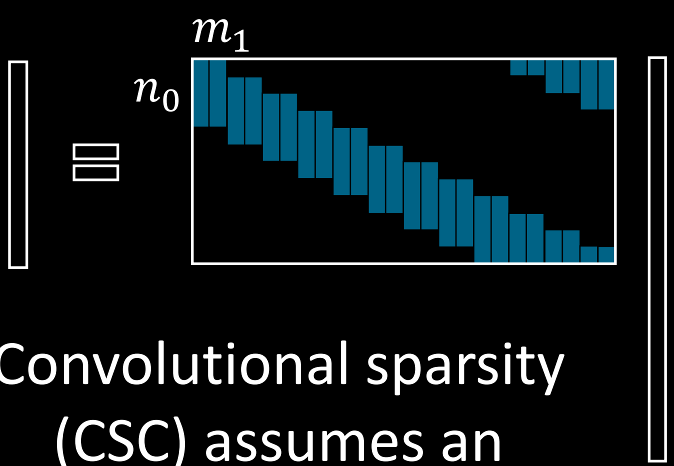
$$f(\mathbf{Y}) = \text{ReLU}\left(\mathbf{b}_2 + \mathbf{W}_2^{\mathrm{T}}\,\text{ReLU}\left(\mathbf{b}_1 + \mathbf{W}_1^{\mathrm{T}}\mathbf{Y}\right)\right)$$



Michael Elad
The Computer-Science Department
The Technion

From CSC to Multi-Layered CSC

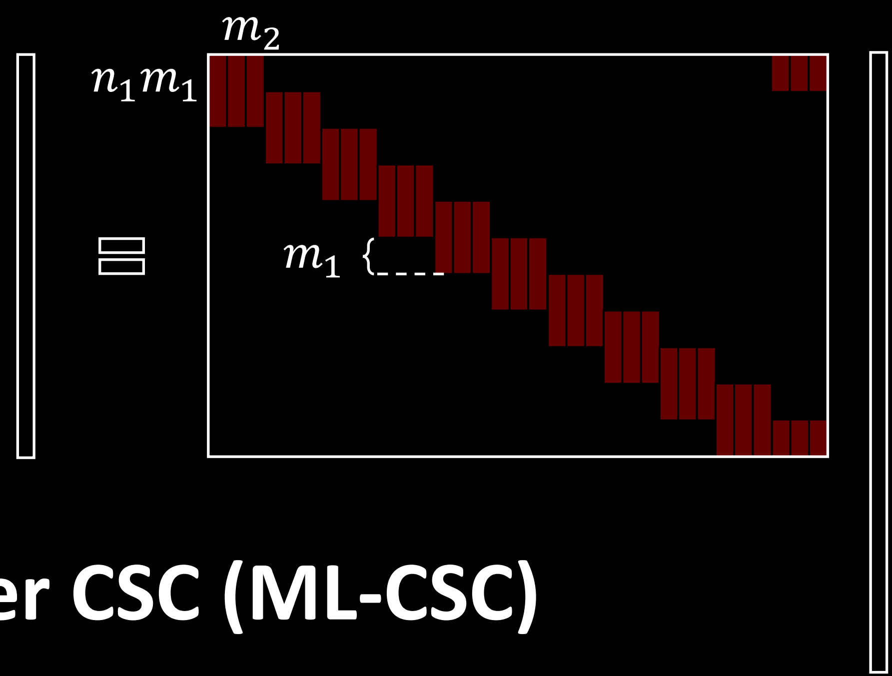
$\mathbf{X} \in \mathbb{R}^{N}$ $\mathbf{D}_1 \in \mathbb{R}^{N \times N m_1}$ $\mathbf{\Gamma}_1 \in \mathbb{R}^{N m_1}$



Convolutional sparsity (CSC) assumes an inherent structure is present in natural signals

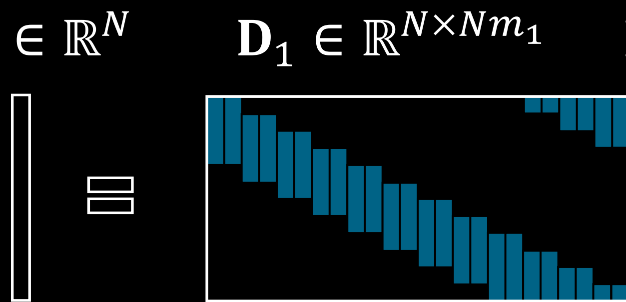
We propose to impose the same structure on the representations themselves

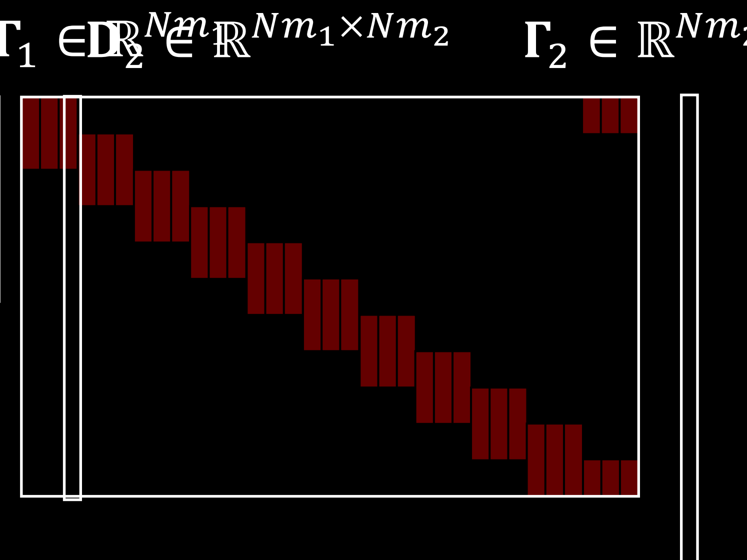
$\mathbf{\Gamma}_1 \in \mathbb{R}^{N m_1}$ $\mathbf{D}_2 \in \mathbb{R}^{N m_1 \times N m_2}$ $\mathbf{\Gamma}_2 \in \mathbb{R}^{N m_2}$



Multi-Layer CSC (ML-CSC)

Intuition: From Atoms to Molecules





- We can chain the all the dictionaries into one effective dictionary $\mathbf{D}_{\text{eff}} = \mathbf{D}_1 \mathbf{D}_2 \mathbf{D}_3 \cdots \mathbf{D}_{\text{K}} \rightarrow \mathbf{x} = \mathbf{D}_{\text{eff}} \mathbf{\Gamma}_{\text{K}}$
- This is a special *Sparseland* (indeed, a CSC) model

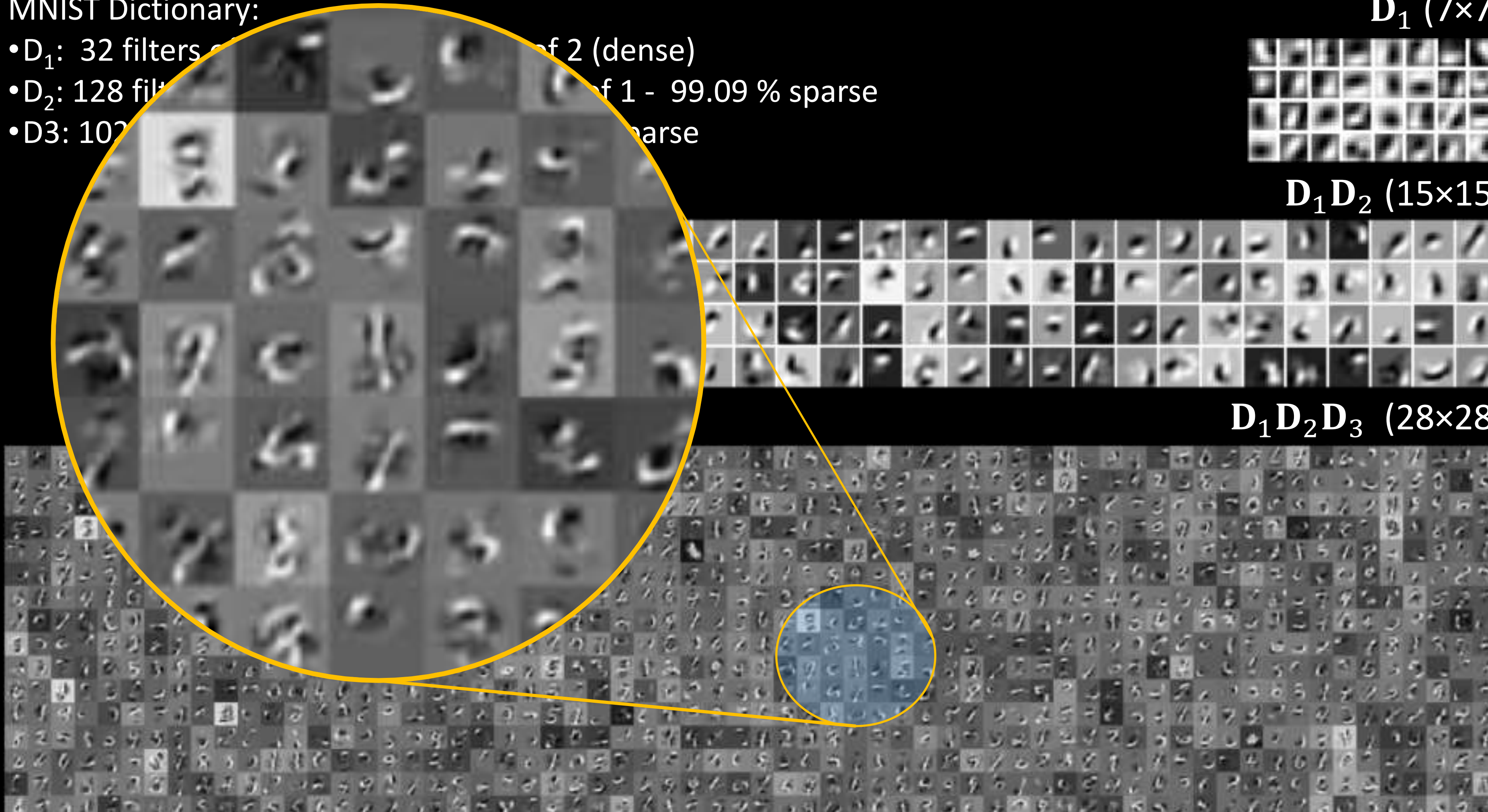
- However:
  - A key property in this model: sparsity of the intermediate representations
  - The effective atoms: atoms

Michael Elad
The Computer-Science Department
The Technion

A Small Taste: Model Training (MNIST)

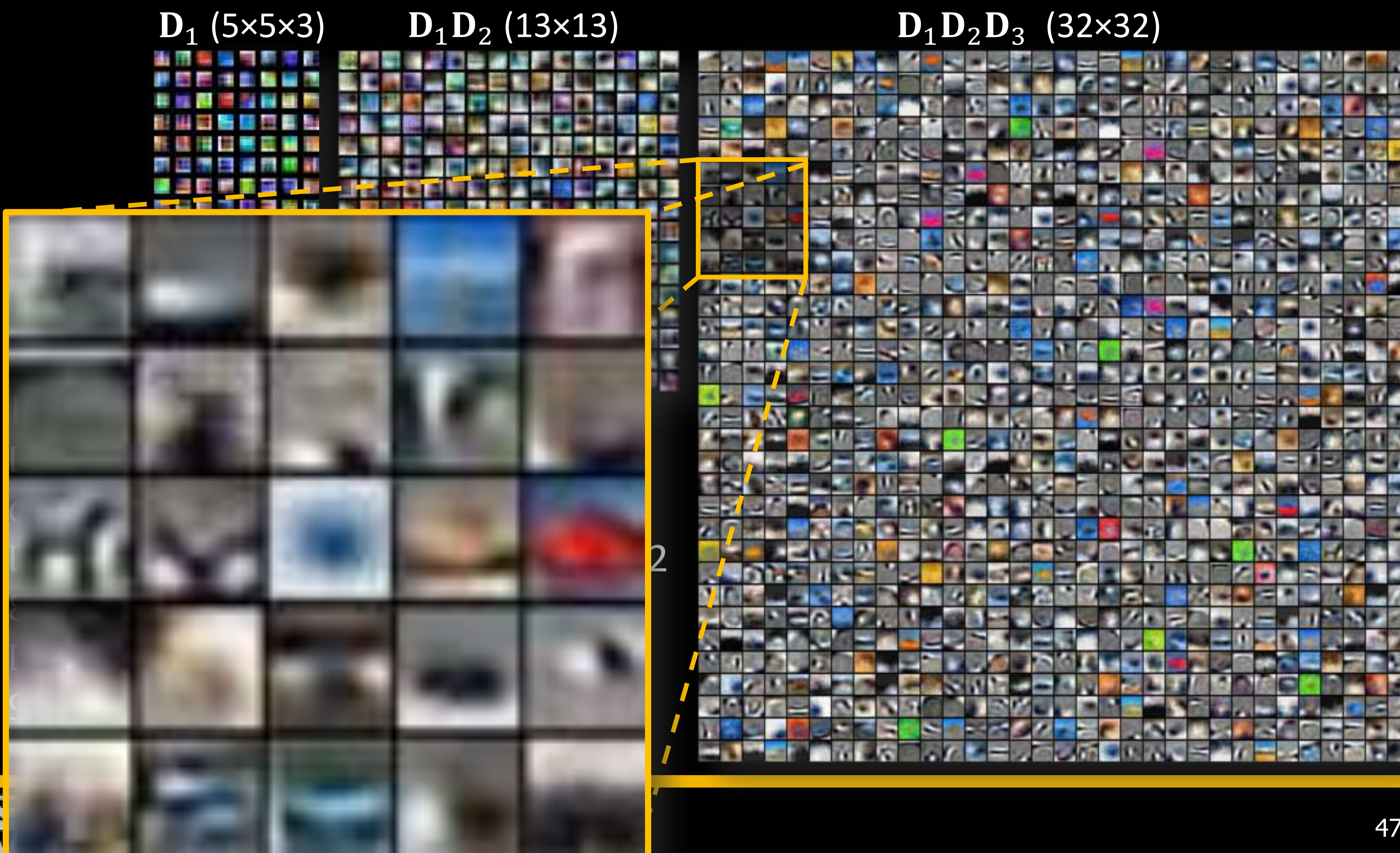
MNIST Dictionary:

- D_1: 32 filters ... f 2 (dense)
- D_2: 128 fil ... f 1 - 99.09 % sparse
- D3: 102 ... arse



Michael Elad
The Computer-Science Department
The Technion

A Small Taste: Model Training (CiFAR)



ML-CSC: Pursuit

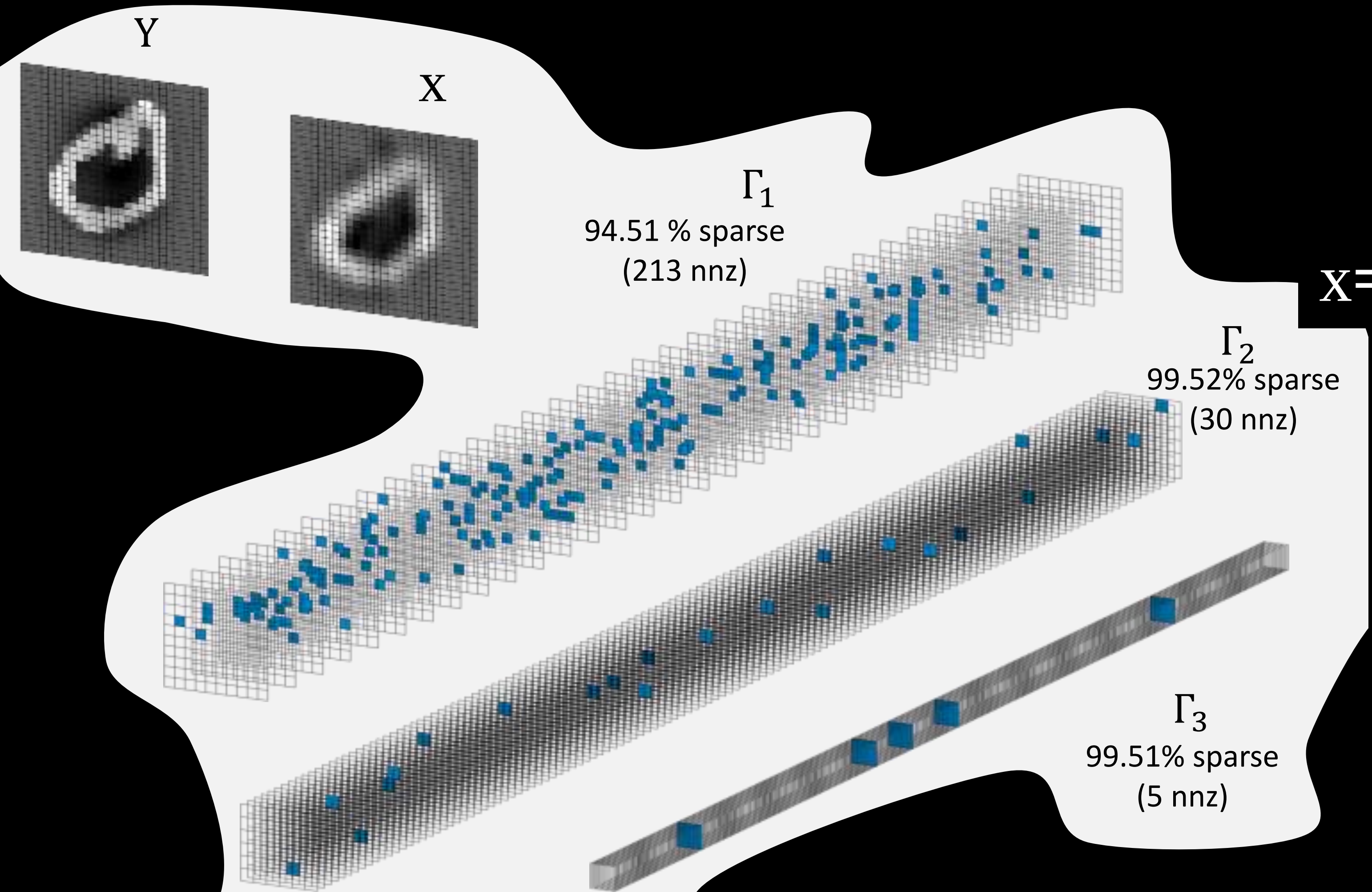
- Deep–Coding Problem $(\mathbf{DCP}_\lambda)$ (dictionaries are known):

$$\begin{Bmatrix} \mathbf{X} = \mathbf{D}_1\mathbf{\Gamma}_1 & \|\mathbf{\Gamma}_1\|_{0,\infty}^{s} \le \lambda_1 \\ \mathbf{\Gamma}_1 = \mathbf{D}_2\mathbf{\Gamma}_2 & \|\mathbf{\Gamma}_2\|_{0,\infty}^{s} \le \lambda_2 \\ \vdots & \vdots \\ \mathbf{\Gamma}_{K-1} = \mathbf{D}_K\mathbf{\Gamma}_K & \|\mathbf{\Gamma}_K\|_{0,\infty}^{s} \le \lambda_K \end{Bmatrix}$$

- Or, more realistically for noisy signals,

$$\text{Find } \{\mathbf{\Gamma}_j\}_{j=1}^{K} \quad s.t. \begin{Bmatrix} \|\mathbf{Y} - \mathbf{D}_1\mathbf{\Gamma}_1\|_2 \le \mathcal{E} & \|\mathbf{\Gamma}_1\|_{0,\infty}^{s} \le \lambda_1 \\ \mathbf{\Gamma}_1 = \mathbf{D}_2\mathbf{\Gamma}_2 & \|\mathbf{\Gamma}_2\|_{0,\infty}^{s} \le \lambda_2 \\ \vdots & \vdots \\ \mathbf{\Gamma}_{K-1} = \mathbf{D}_K\mathbf{\Gamma}_K & \|\mathbf{\Gamma}_K\|_{0,\infty}^{s} \le \lambda_K \end{Bmatrix}$$

A Small Taste: Pursuit



$$x = \mathbf{D}_1 \Gamma_1$$

$$x = \mathbf{D}_1 \mathbf{D}_2 \Gamma_2$$

$$x = \mathbf{D}_1 \mathbf{D}_2 \mathbf{D}_3 \Gamma_3$$

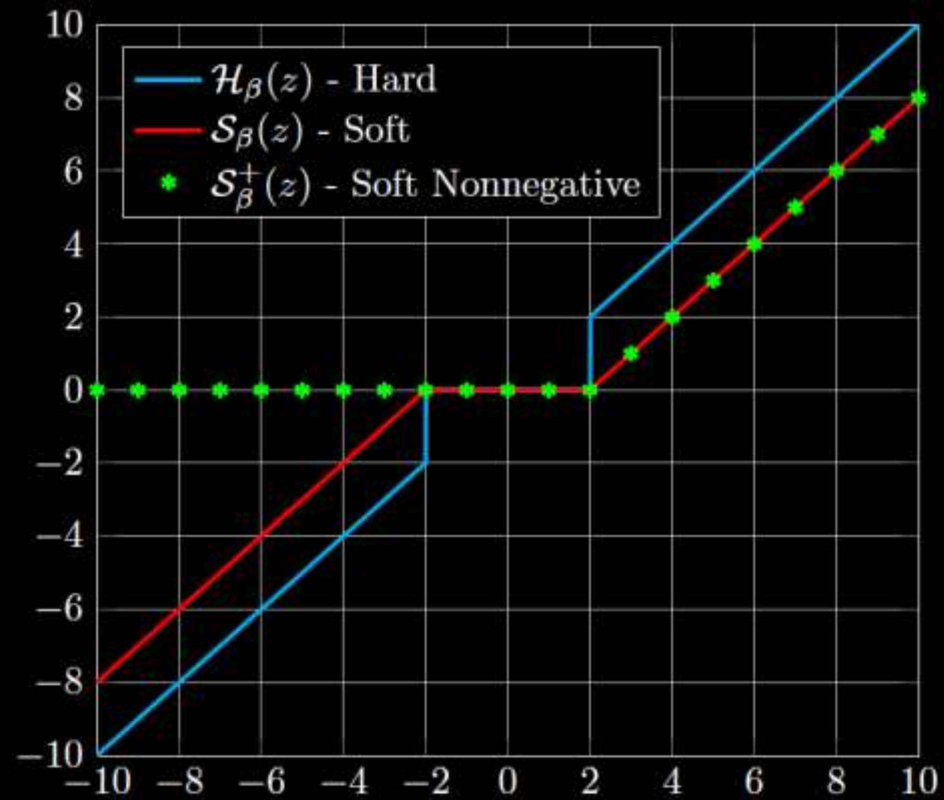
ML-CSC: The Simplest Pursuit

The simplest pursuit algorithm (single-layer case) is the THR algorithm, which operates on a given input signal $\mathbf{Y}$ by:

$$\mathbf{Y} = \mathbf{D}\boldsymbol{\Gamma} + \mathbf{E}$$

and $\boldsymbol{\Gamma}$ is sparse

$$\hat{\boldsymbol{\Gamma}} = \mathcal{P}_\beta\left(\mathbf{D}^{\mathrm{T}}\mathbf{Y}\right)$$



Michael Elad
The Computer-Science Department
The Technion

Consider this for Solving the DCP

- Layered thresholding (LT):

Estimate $\mathbf{\Gamma}_1$ via the THR algorithm

$$\hat{\mathbf{\Gamma}}_2 = \mathcal{P}_{\beta_2}\left(\mathbf{D}_2^T \mathcal{P}_{\beta_1}\left(\mathbf{D}_1^T \mathbf{Y}\right)\right)$$

Estimate $\mathbf{\Gamma}_2$ via the THR algorithm

$$\left(\mathbf{DCP}_\lambda^\mathcal{E}\right): \text{ Find } \left\{\mathbf{\Gamma}_j\right\}_{j=1}^K \quad s.t.$$

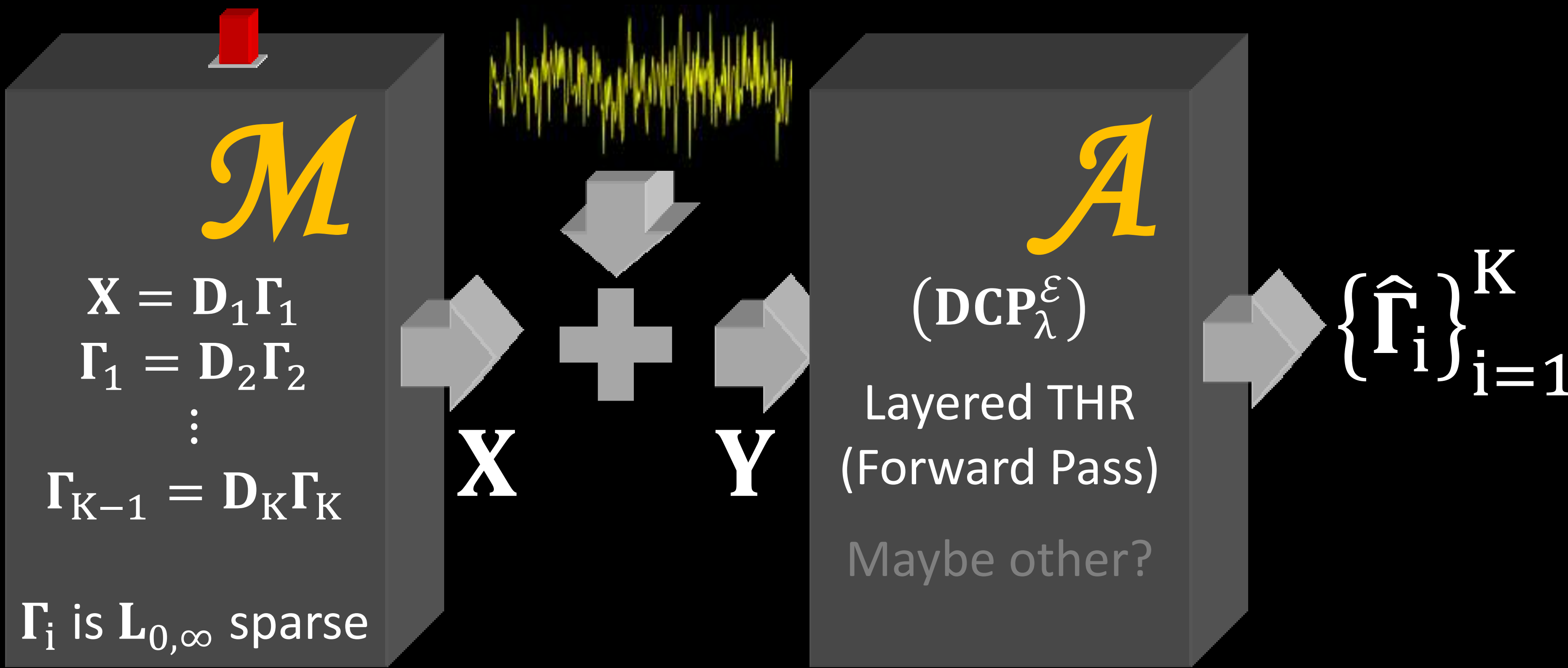
$$\begin{cases} \|\mathbf{Y} - \mathbf{D}_1\mathbf{\Gamma}_1\|_2 \le \mathcal{E} & \|\mathbf{\Gamma}_1\|_{0,\infty}^s \le \lambda_1 \\ \mathbf{\Gamma}_1 = \mathbf{D}_2\mathbf{\Gamma}_2 & \|\mathbf{\Gamma}_2\|_{0,\infty}^s \le \lambda_2 \\ \vdots & \vdots \\ \mathbf{\Gamma}_{K-1} = \mathbf{D}_K\mathbf{\Gamma}_K & \|\mathbf{\Gamma}_K\|_{0,\infty}^s \le \lambda_K \end{cases}$$

- Now let's take a look at how Conv. Neural Network operates:

$$f(\mathbf{Y}) = \text{ReLU}\left(\mathbf{b}_2 + \mathbf{W}_2^T \text{ ReLU}\left(\mathbf{b}_1 + \mathbf{W}_1^T \mathbf{Y}\right)\right)$$

The layered (soft nonnegative) thresholding and the CNN forward pass algorithm are the very same thing !!!

Theoretical Path



Armed with this view of a generative source model, we may ask new and daring theoretical questions

Michael Elad
The Computer-Science Department
The Technion

Success of the Layered-THR

Theorem: If $\|\boldsymbol{\Gamma}_i\|_{0,\infty}^{s} < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D}_i)} \cdot \frac{|\boldsymbol{\Gamma}_i^{min}|}{|\boldsymbol{\Gamma}_i^{max}|}\right) - \frac{1}{\mu(\mathbf{D}_i)} \cdot \frac{\varepsilon_L^{i-1}}{|\boldsymbol{\Gamma}_i^{max}|}$

then the **Layered Hard THR** (with the proper thresholds) **finds the correct supports** and $\|\boldsymbol{\Gamma}_i^{LT} - \boldsymbol{\Gamma}_i\|_{2,\infty}^{p} \leq \varepsilon_L^i$, where we have defined $\varepsilon_L^0 = \|\mathbf{E}\|_{2,\infty}^{p}$ and

$$\varepsilon_L^i = \sqrt{\|\boldsymbol{\Gamma}_i\|_{0,\infty}^{p}} \cdot \left(\varepsilon_L^{i-1} + \mu(\mathbf{D}_i)\left(\|\boldsymbol{\Gamma}_i\|_{0,\infty}^{s} - 1\right)|\boldsymbol{\Gamma}_i^{max}|\right)$$

Papyan, Romano & Elad ('17)

The stability of the forward pass is guaranteed if the underlying representations are **locally** sparse and the noise is **locally** bounded

Problems:
1. Contrast
2. Error growth
3. Error even if no noise

Layered Basis Pursuit (BP)

- We chose the Thresholding algorithm due to its simplicity, but we do know that there are better pursuit methods – how about using them?
- Lets use the Basis Pursuit instead …

$$\left(\mathbf{DCP}_{\lambda}^{\mathcal{E}}\right): \text{ Find } \left\{\boldsymbol{\Gamma}_j\right\}_{j=1}^{K} \quad s.t.$$

$$\begin{Bmatrix} \|\mathbf{Y}-\mathbf{D}_1\boldsymbol{\Gamma}_1\|_2 \le \mathcal{E} & \|\boldsymbol{\Gamma}_1\|_{0,\infty}^{s} \le \lambda_1 \\ \boldsymbol{\Gamma}_1 = \mathbf{D}_2\boldsymbol{\Gamma}_2 & \|\boldsymbol{\Gamma}_2\|_{0,\infty}^{s} \le \lambda_2 \\ \vdots & \vdots \\ \boldsymbol{\Gamma}_{K-1} = \mathbf{D}_K\boldsymbol{\Gamma}_K & \|\boldsymbol{\Gamma}_K\|_{0,\infty}^{s} \le \lambda_K \end{Bmatrix}$$

$$\boldsymbol{\Gamma}_1^{\mathrm{LBP}} = \min_{\boldsymbol{\Gamma}_1} \frac{1}{2}\|\mathbf{Y}-\mathbf{D}_1\boldsymbol{\Gamma}_1\|_2^2 + \lambda_1\|\boldsymbol{\Gamma}_1\|_1$$

$$\boldsymbol{\Gamma}_2^{\mathrm{LBP}} = \min_{\boldsymbol{\Gamma}_2} \frac{1}{2}\left\|\boldsymbol{\Gamma}_1^{\mathrm{LBP}}-\mathbf{D}_2\boldsymbol{\Gamma}_2\right\|_2^2 + \lambda_2\|\boldsymbol{\Gamma}_2\|_1$$

⋮

Deconvolutional networks
[Zeiler, Krishnan, Taylor & Fergus '10]

Success of the Layered BP

Theorem: **Assuming that** $\|\mathbf{\Gamma_i}\|_{0,\infty}^{s} < \frac{1}{3}\left(1 + \frac{1}{\mu(\mathbf{D_i})}\right)$

then the Layered Basis Pursuit performs very well:

1. The support of $\mathbf{\Gamma}_i^{LBP}$ is contained in that of $\mathbf{\Gamma}_i$
2. The error is bounded: $\|\mathbf{\Gamma}_i^{LBP} - \mathbf{\Gamma}_i\|_{2,\infty}^{p} \le \varepsilon_L^i$, where

$$\varepsilon_L^i = 7.5^i \|\mathbf{E}\|_{2,\infty}^{p} \prod_{j=1}^{i} \sqrt{\|\mathbf{\Gamma}_j\|_{0,\infty}^{p}}$$

3. Every entry in $\mathbf{\Gamma}_i$ greater than $\varepsilon_L^i / \sqrt{\|\mathbf{\Gamma}_i\|_{0,\infty}^{p}}$ will be found

Papyan, Romano & Elad ('17)

Problems:

1. ~~Contrast~~
2. Error growth
3. ~~Error even if no noise~~

Layered Iterative Thresholding

Layered BP: $\Gamma_j^{LBP} = \min_{\Gamma_j} \frac{1}{2} \left\| \Gamma_{j-1}^{LBP} - \mathbf{D}_j \Gamma_j \right\|_2^2 + \xi_j \left\| \Gamma_j \right\|_1$ j

Layered Iterative Soft-Thresholding:

t $\Gamma_j^t = \mathcal{S}_{\xi_j / c_j}\left(\Gamma_j^{t-1} + \mathbf{D}_j^T \left(\hat{\Gamma}_{j-1} - \mathbf{D}_j \Gamma_j^{t-1} \right) \right)$ j

Note that our suggestion implies that groups of layers share the same dictionaries

Can be seen as a very deep recurrent neural network
[Gregor & LeCun '10]

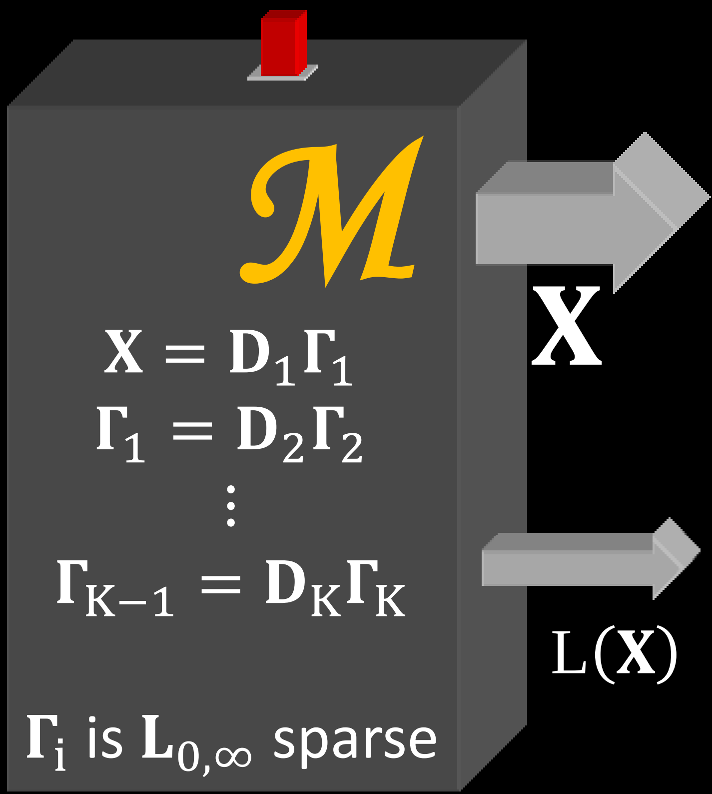
What About Learning?

Sparseland — Sparse Representation Theory → **CSC** — Convolutional Sparse Coding → **ML-CSC** — Multi-Layered Convolutional Sparse Coding

- All these models rely on proper Dictionary Learning Algorithms to fulfil their mission:
  - Sparseland: We have unsupervised and supervised such algorithms, and a beginning of theory to explain how these work
  - CSC: We have few and only unsupervised methods, and even these are not fully stable/clear
  - ML-CSC: One algorithm has been proposed (unsupervised) – see ArxiV

Michael Elad
The Computer-Science Department
The Technion

Where are the Labels?



We presented the ML-CSC as a machine that produces signals $\mathbf{X}$

Answer 1:

- We do not need labels because everything we show refers to the unsupervised case, in which we operate on signals, not necessarily in the context of recognition

Answer 2:

- In fact, this model could be augmented by a synthesis of the corresponding label by:

$$\mathrm{L}(\mathbf{X}) = sign\{c + \sum_{j=1}^{K} w_j^T \Gamma_j\}$$

- This assumes that knowing the representations (or maybe their supports?) suffice for identifying the label
- Thus, a successful pursuit algorithm can lead to an accurate recognition if the network is augmented by a FC classification layer

Time to Conclude

This Talk

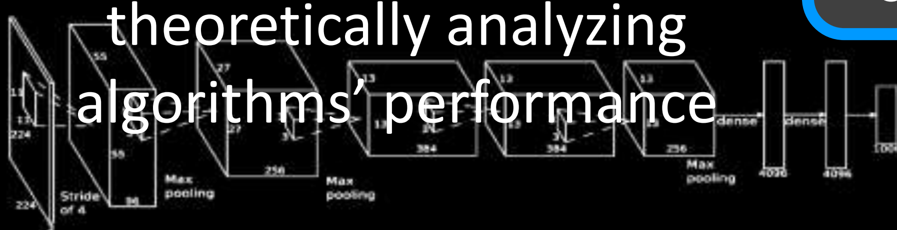
Sparseland

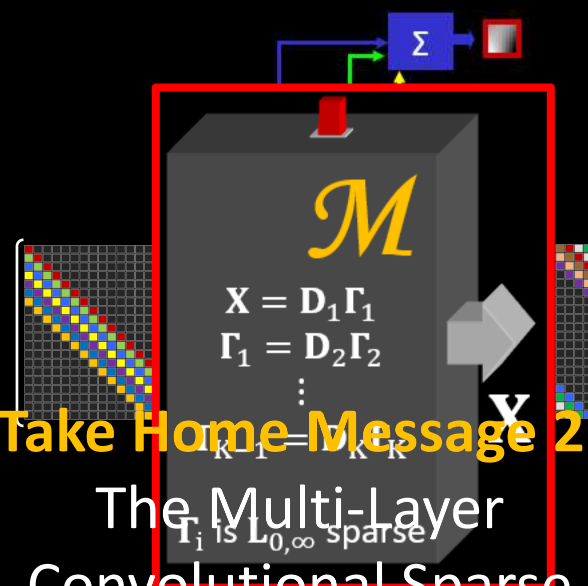
The desire to model data

Novel View of Convolutional Sparse Coding

Multi-Layer Convolutional Sparse Coding

A novel interpretation and theoretical understanding of CNN

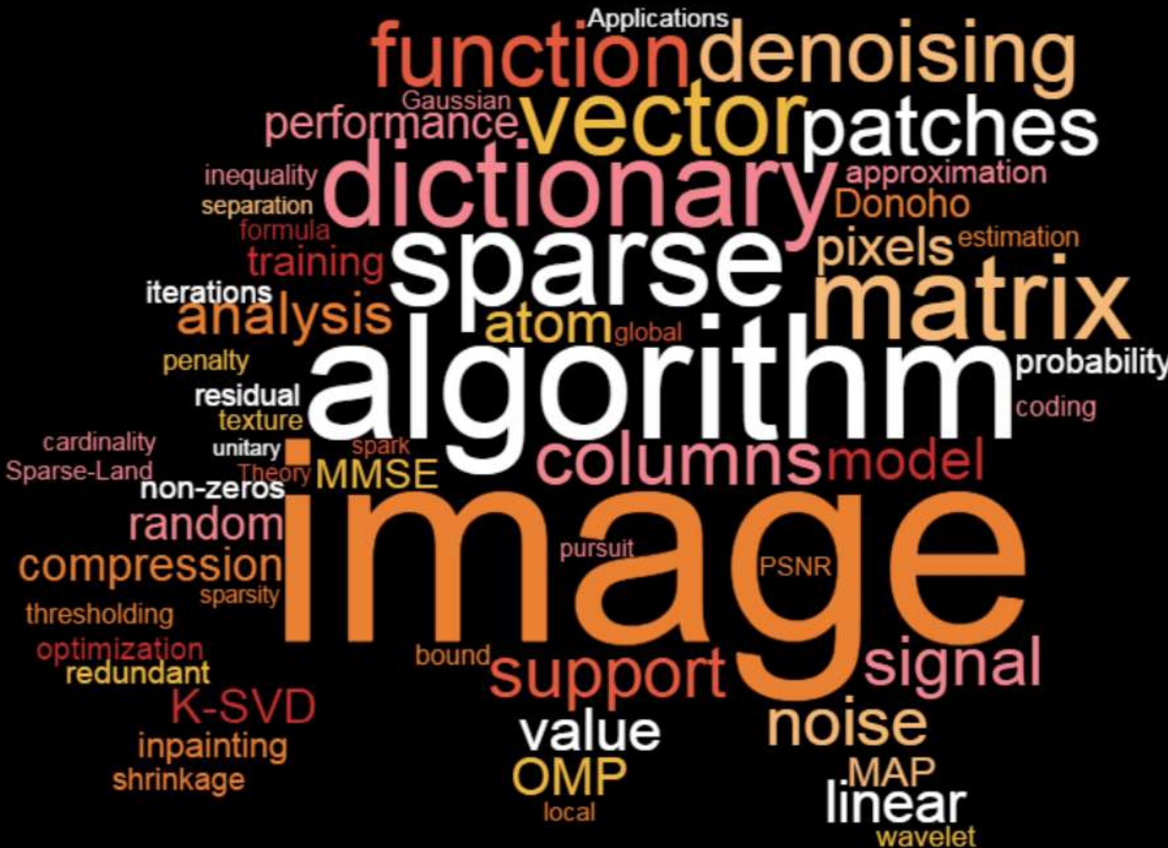




Take Home Message 1:
Generative modeling of data sources enables algorithm development along with theoretically analyzing algorithms' performance

Take Home Message 2:
The Multi-Layer Convolutional Sparse Coding model could be a new platform for understanding and [illegible]ing deep-learning solutions

We presented a theoretical study of the CSC model and how to operate locally while getting global optimality



More on these (including these slides and the relevant papers) can be found in http://www.cs.technion.ac.il/~elad

Michael Elad
The Computer-Science Department
The Technion