Sparse Modeling
in
Image Processing and Deep Learning

Michael Elad

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

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

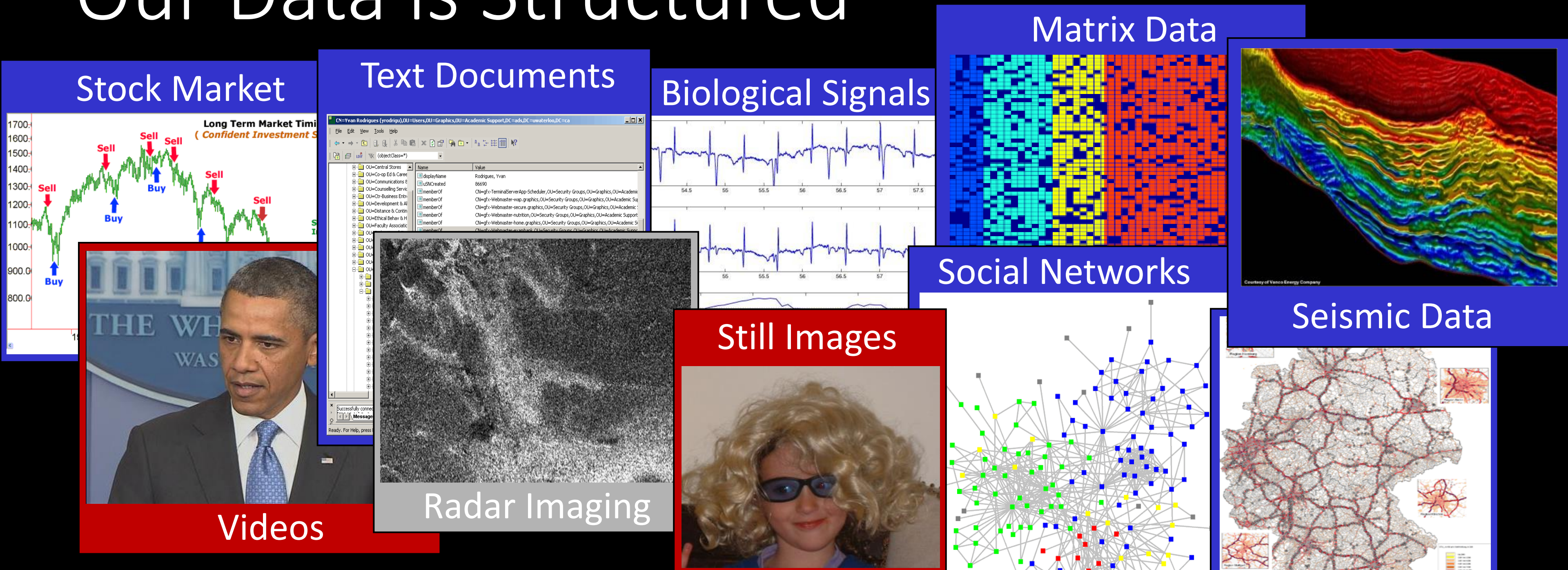
Generative modeling of data sources enables

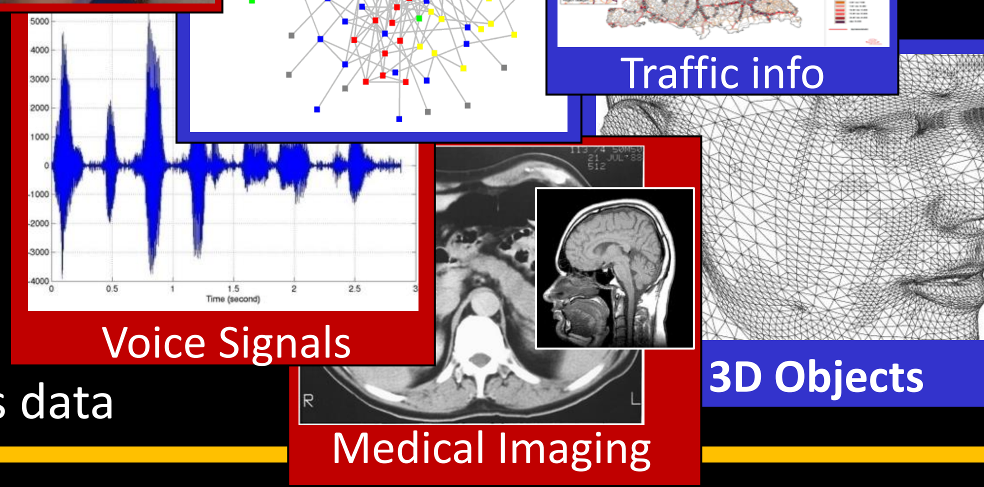
- A systematic algorithm development, &
- A theoretical analysis of their performance

Michael Elad
The Computer-Science Department
The Technion

Multi-Layered Convolutional Sparse Modeling

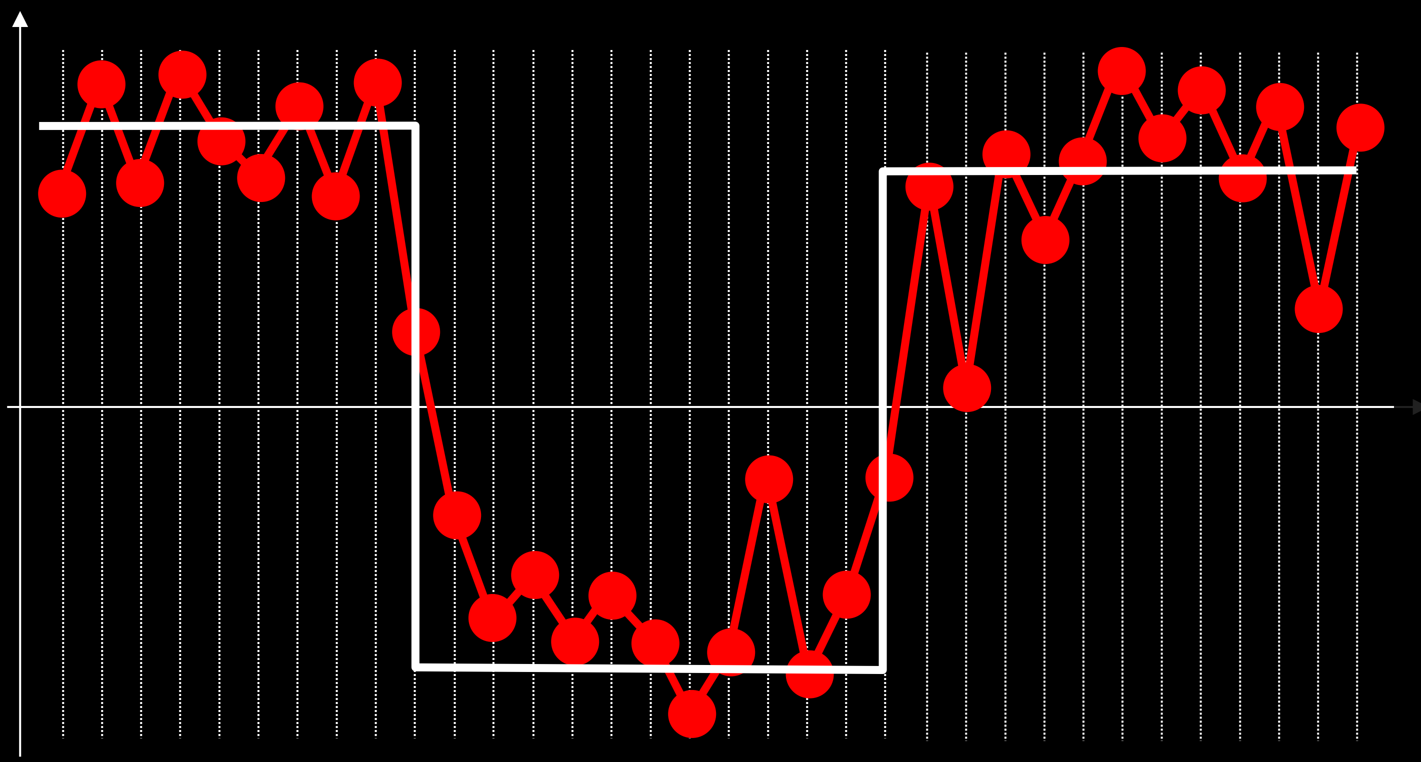
Our Data is Structured





- 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

Model?



Fact 1:
This signal contains AWGN $\mathbb{N}(0,1)$

Fact 2:
The clean signal is believed to be PWC

Effective removal of noise (and many other tasks) relies on an proper modeling of the signal

Michael Elad
The Computer-Science Department
The Technion

Which Model to Choose?

- 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

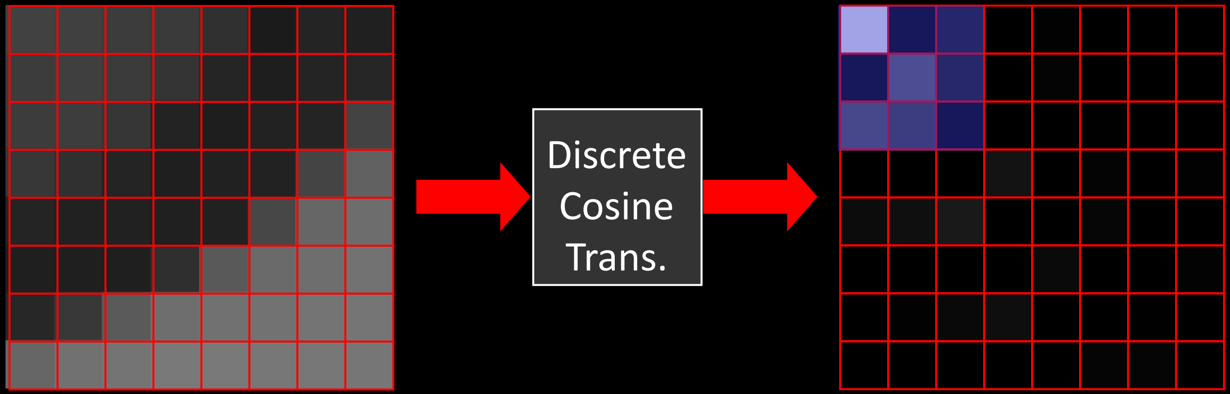
Besov-Spaces

Total-Variation

Beltrami-Flow

An Example: JPEG and DCT

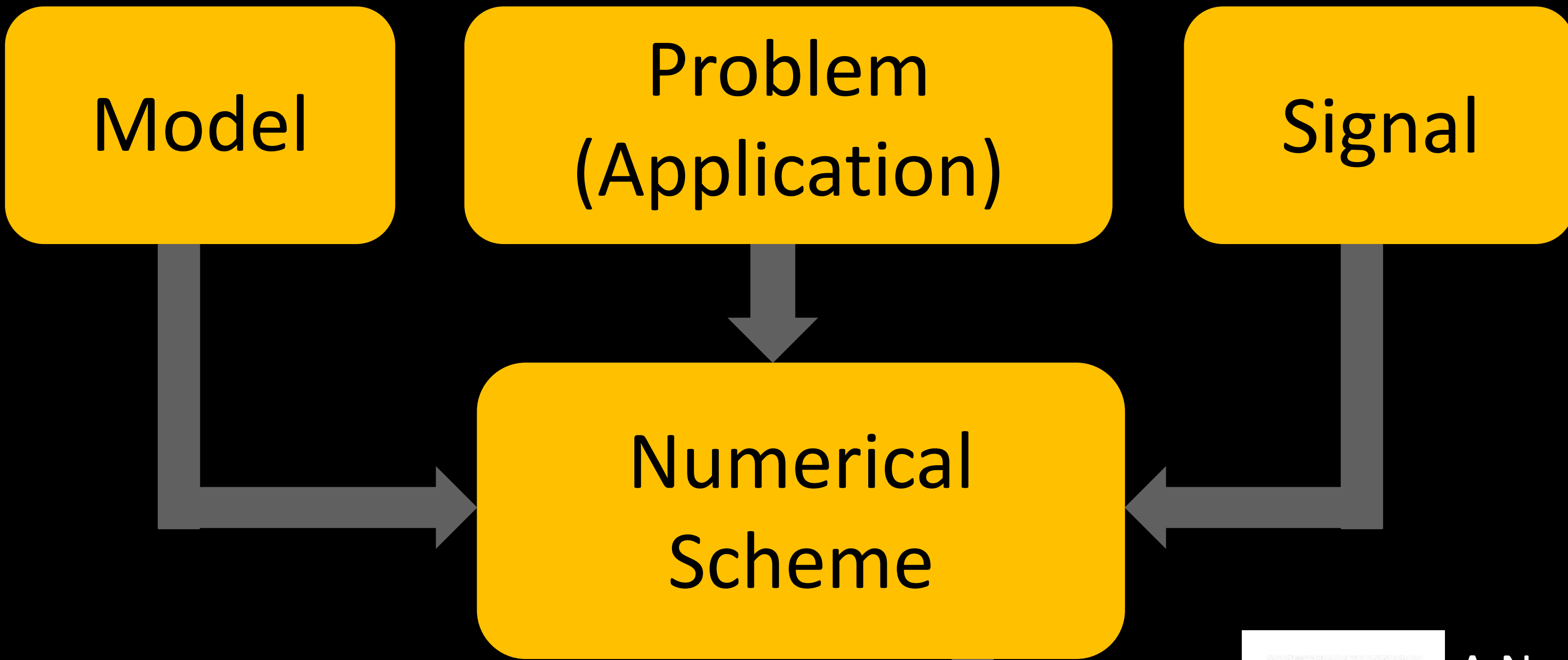
How & why does it works?

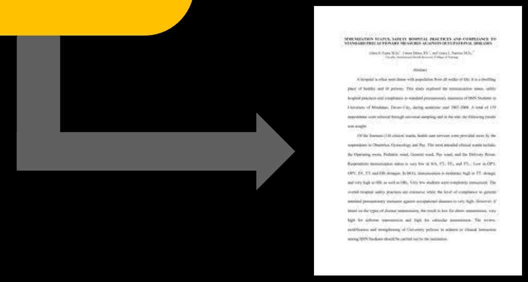


The model assumption: after DCT, the top left coefficients to be dominant and the rest zeros

Michael Elad
The Computer-Science Department
The Technion

Research in Signal/Image Processing





A New Research Work (and Paper) is Born

The fields of signal & image processing are essentially built of an evolution of models and ways to use them for various tasks

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

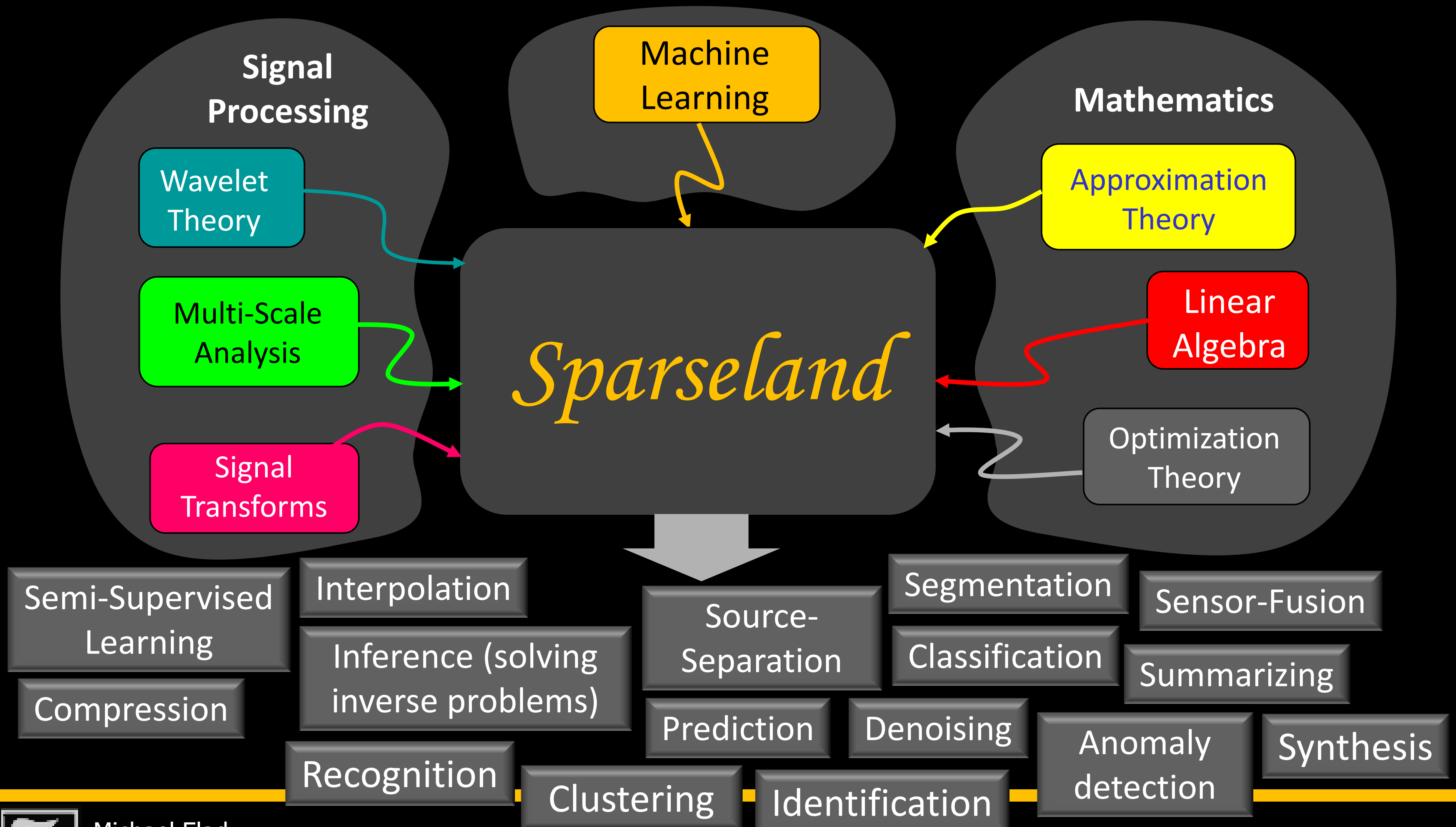
Sparseland

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

Multi-Layered Convolutional

Sparse Modeling

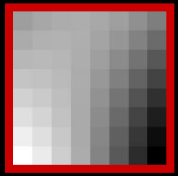
A New Emerging Model



Michael Elad
The Computer-Science Department
The Technion

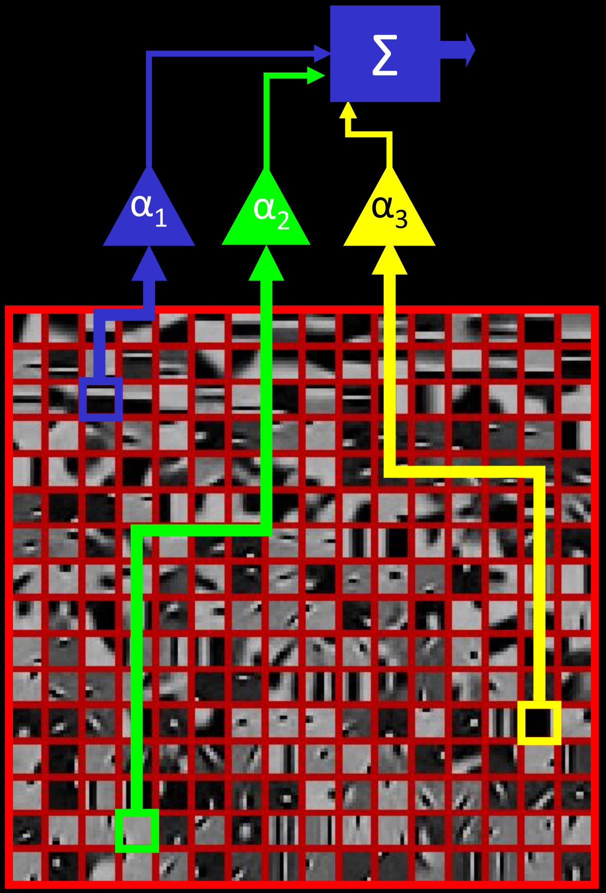
The *Sparseland* Model

- 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



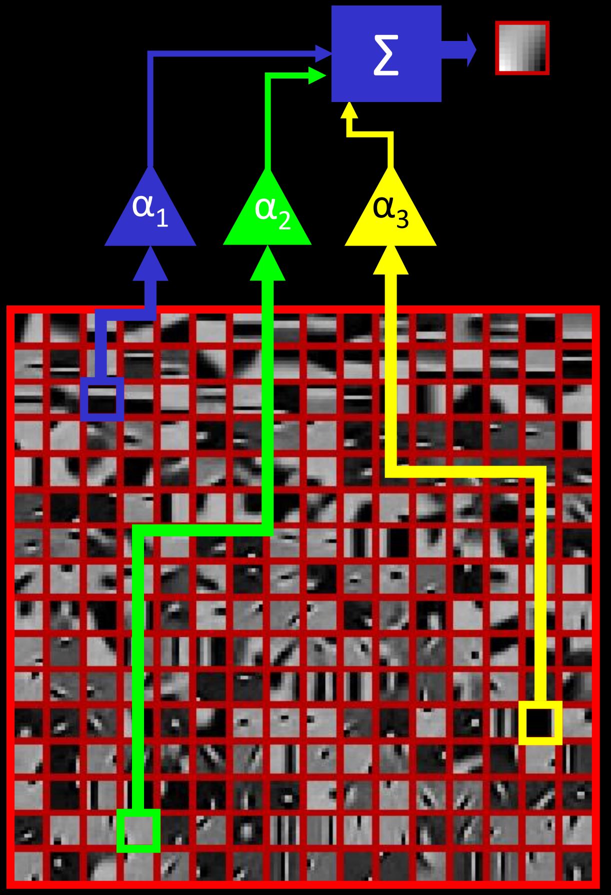
Michael Elad
The Computer-Science Department
The Technion

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)

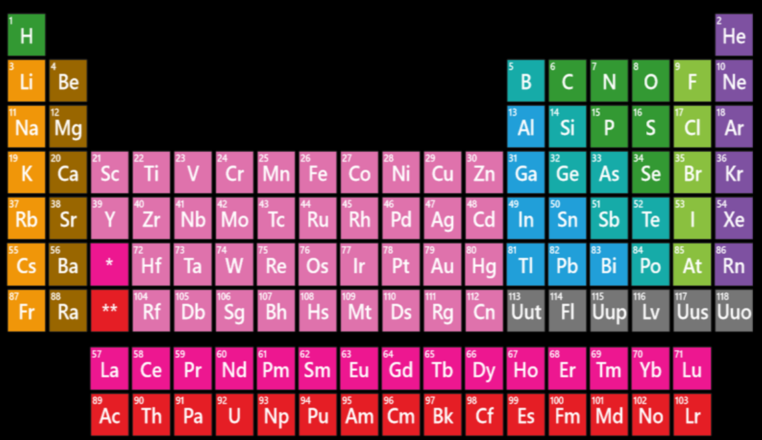


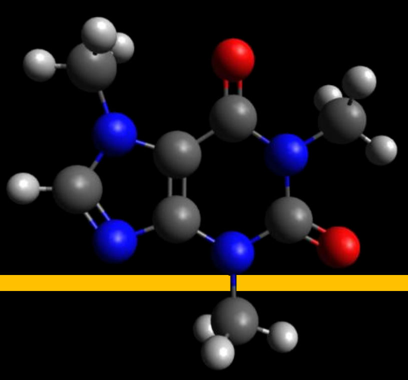
Michael Elad
The Computer-Science Department
The Technion

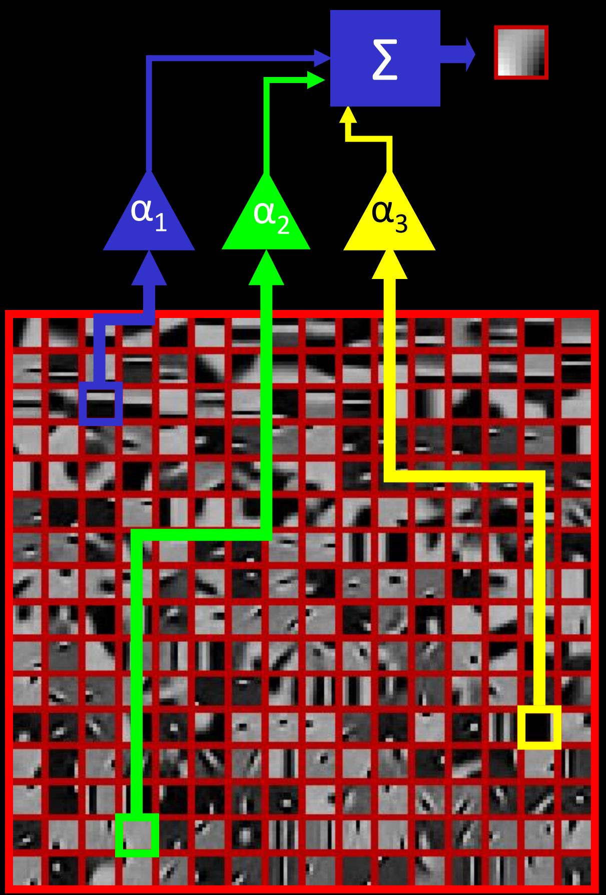
Chemistry of Data

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

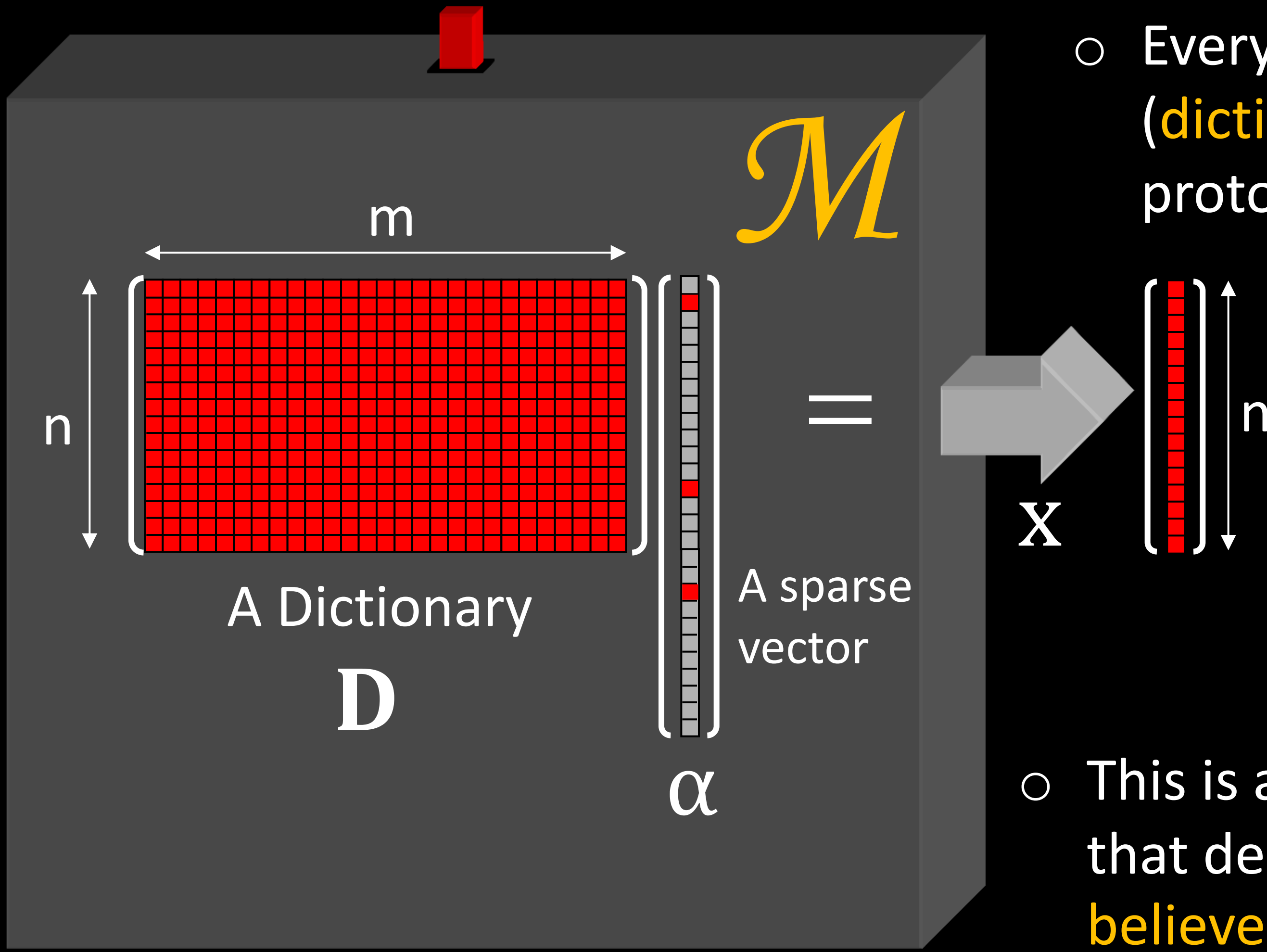
- Our dictionary stands for the **Periodic Table** containing all the elements
- Our model follows a similar rationale: Every molecule is built of **few** elements







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

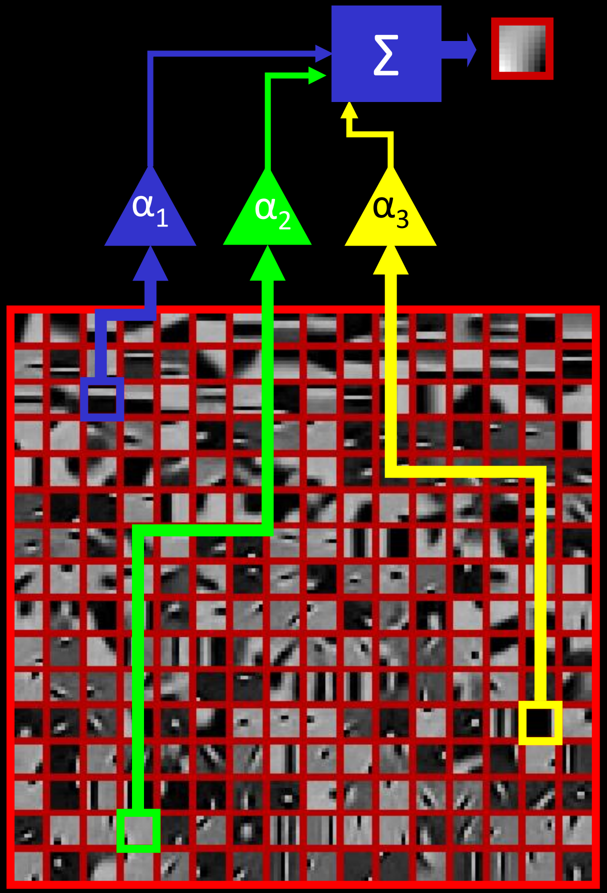
Michael Elad
The Computer-Science Department
The Technion

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

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

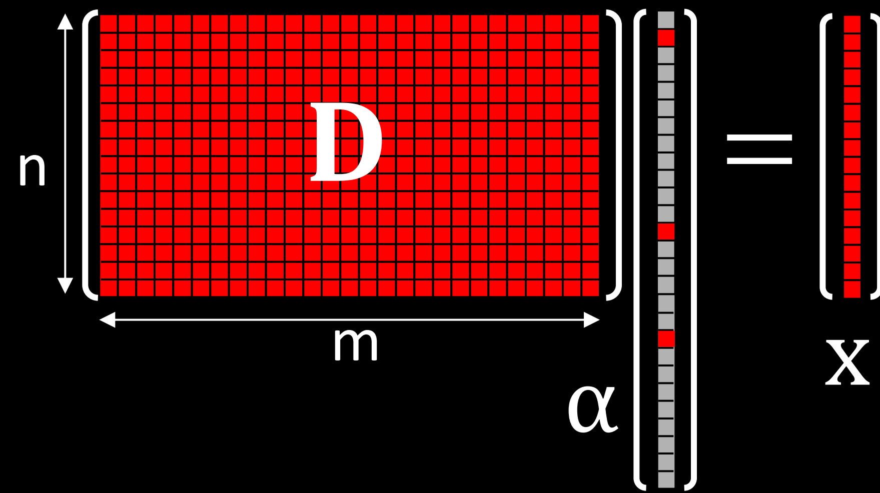
  - If each of these takes 1nano-sec to test, will take ~7.5e20 years to finish !!!!!!
- So, are we stuck?



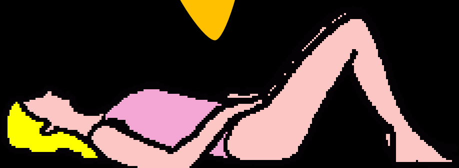
Atom Decomposition Made Formal

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

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



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

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

Approximation Algorithms

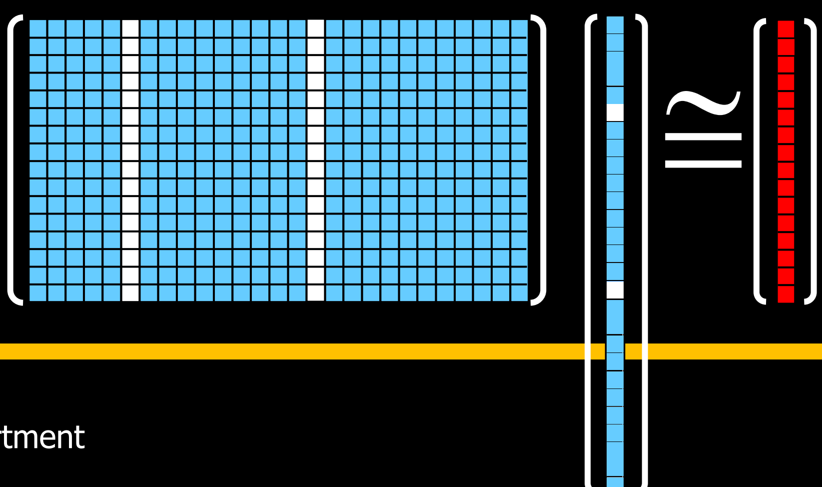
Basis Pursuit

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

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

Matching Pursuit

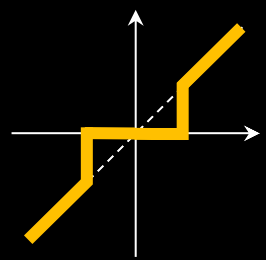
Find the support greedily, one element at a time



Thresholding

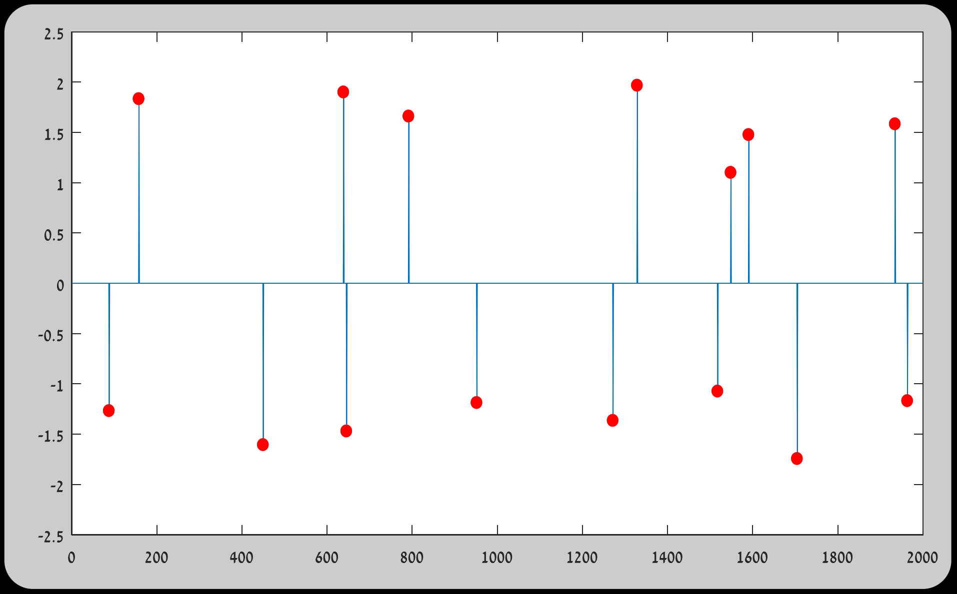
Multiply y by $\mathbf{D}^{\mathbf{T}}$ and apply shrinkage:

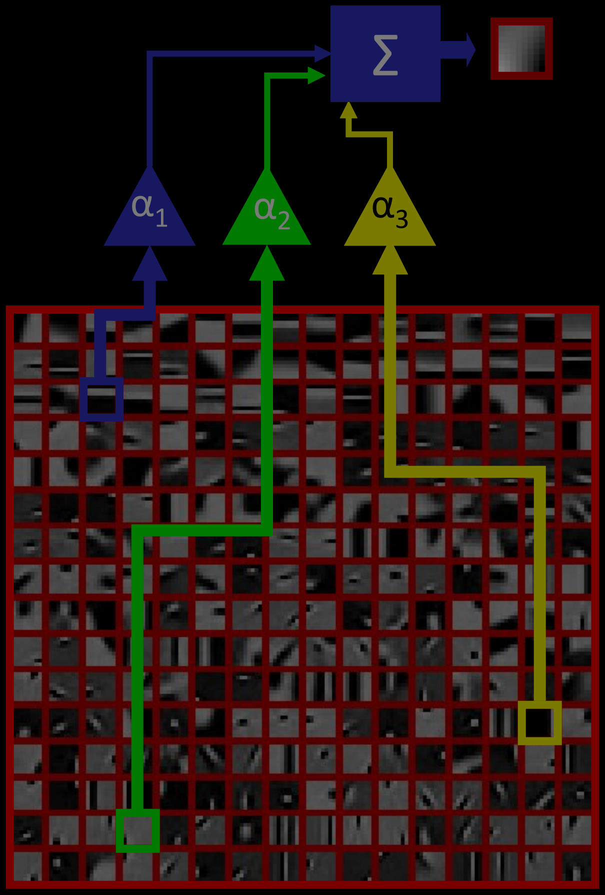
$$\hat{\alpha} = \mathcal{P}_{\beta}\{\mathbf{D}^{\mathbf{T}}y\}$$



Difficulties with *Sparseland*

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

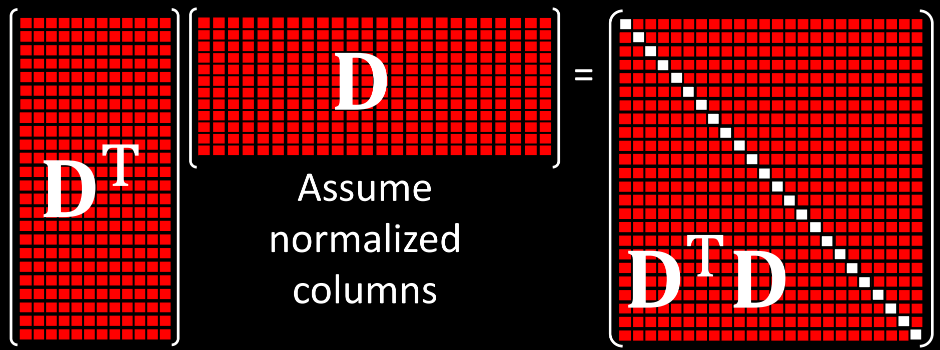




- Surprising fact: Many of these algorithms are often accompanied by theoretical guarantees for their success, if the unknown is sparse enough

The Mutual Coherence

- Compute



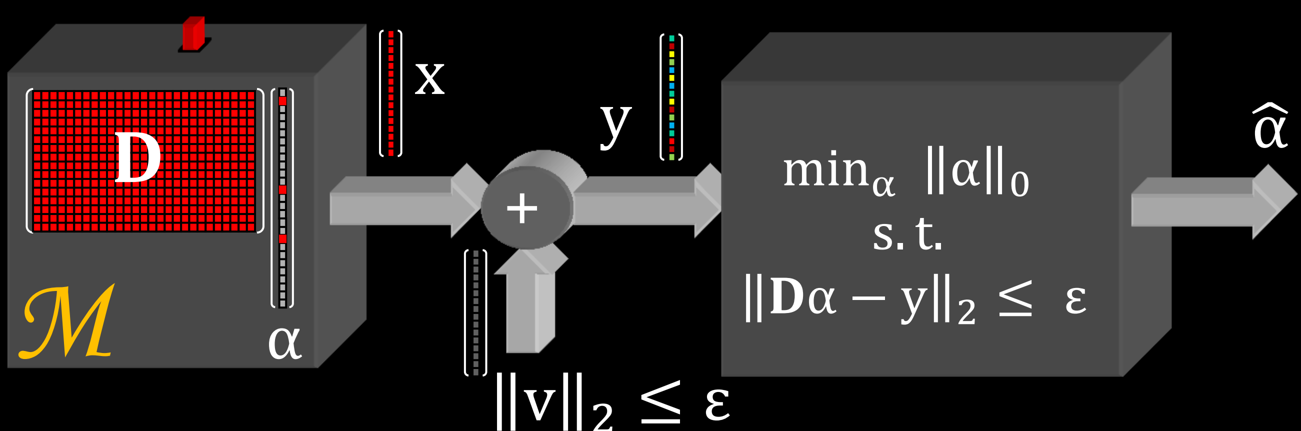
- The Mutual Coherence $\mu(\mathbf{D})$ is the largest off-diagonal entry in absolute value
- We will pose all the theoretical results in this talk using this property, due to its simplicity
- You may have heard of other ways to characterize the dictionary (Restricted Isometry Property - RIP, Exact 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: $\|\hat{\alpha} - \alpha\|_2^2 \le \dfrac{4\varepsilon^2}{1-\mu(4\|\alpha\|_0-1)}$

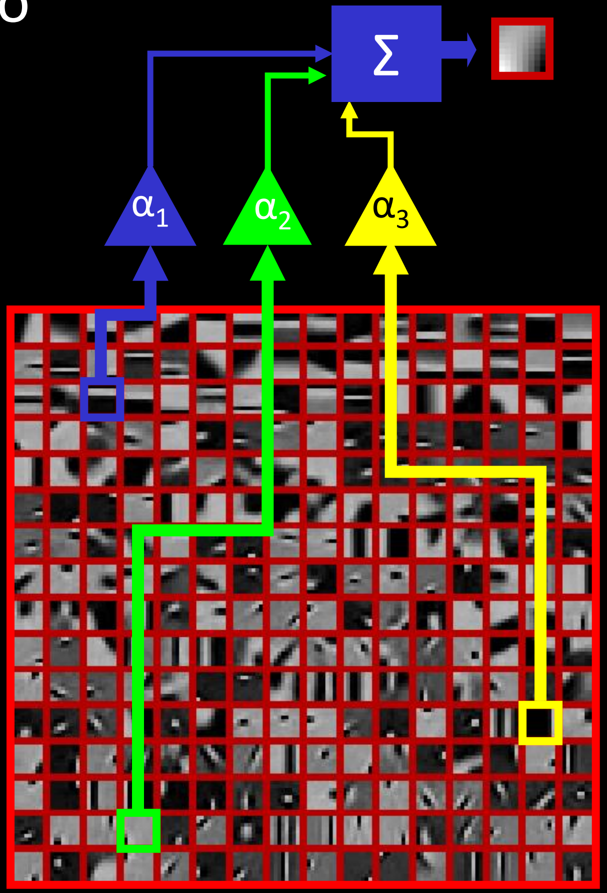
Donoho, Elad & Temlyakov ('06)



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

Difficulties with *Sparseland*

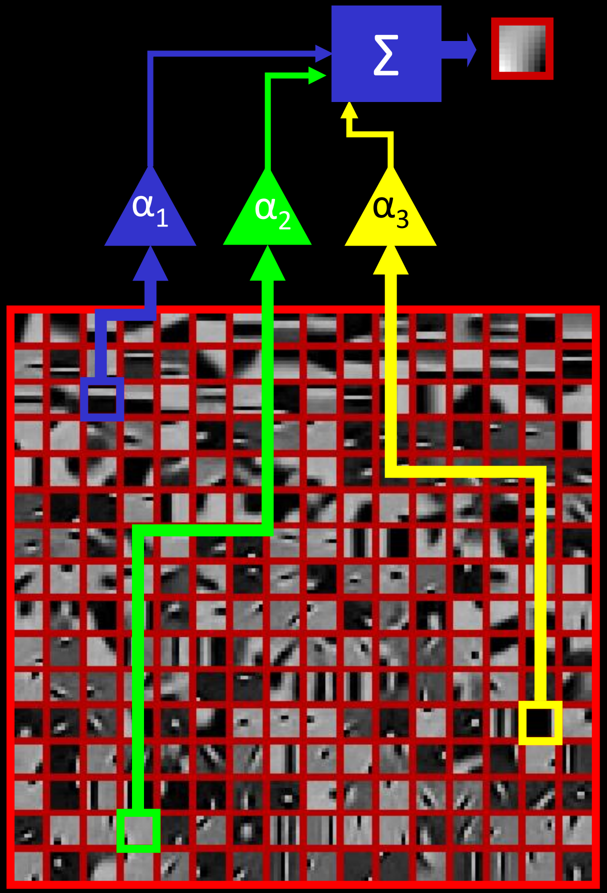
- Problem 2: Given a family of signals, how do we find the dictionary to represent it well?
- Solution: Learn! Gather a large set of signals (many thousands), and find the dictionary that sparsifies them
- Such algorithms were developed in the past 10 years (e.g., K-SVD), and their performance is surprisingly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



Michael Elad
The Computer-Science Department
The Technion

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

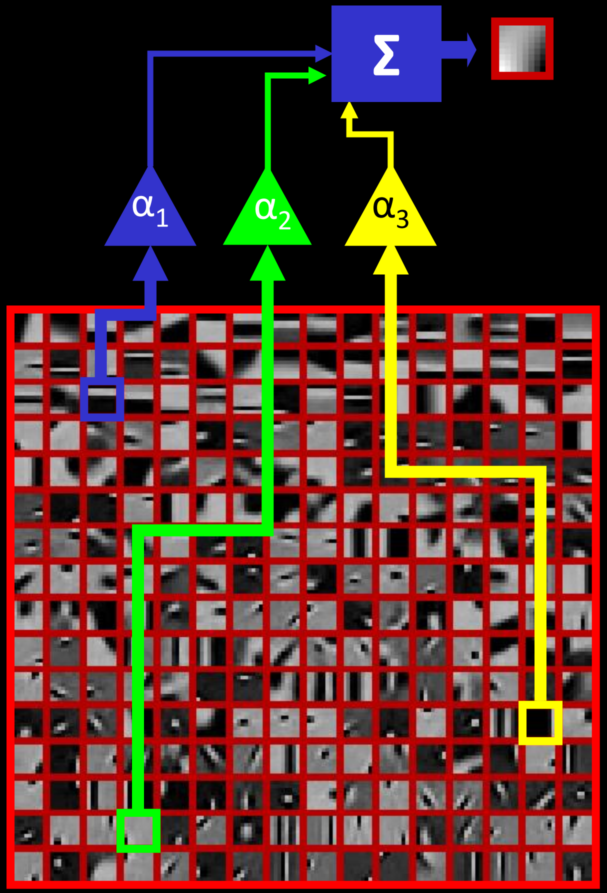


Michael Elad
The Computer-Science Department
The Technion

Difficulties with *Sparseland* ?

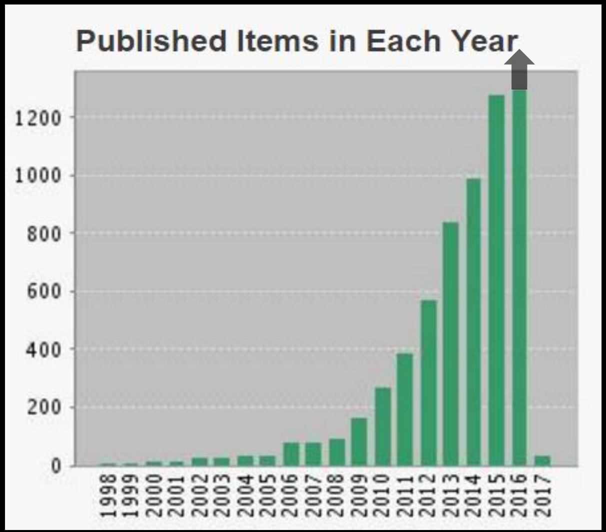
- Problem 1: Given an image patch, how can we find its atom decomposition ?
- Problem 2: Given a family of signals, how do we find the dictionary to represent it well?
- Problem 3: Is this model flexible enough to describe various sources? E.g., Is it good for images? audio? …

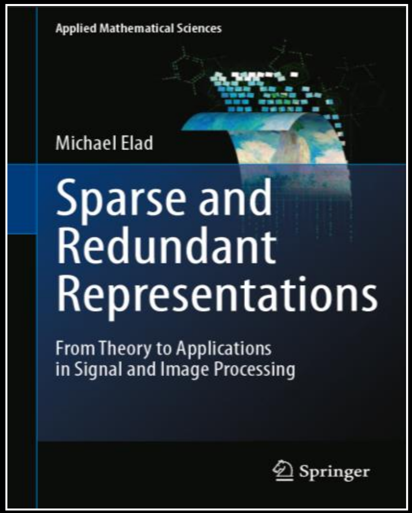
ALL ANSWERED POSITIVELY AND CONSTRUCTIVELY

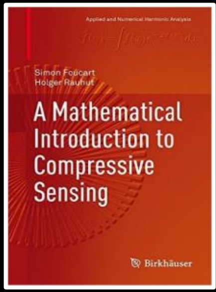


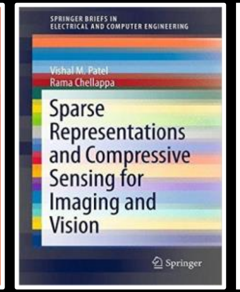
This Field has been rapidly GROWING …

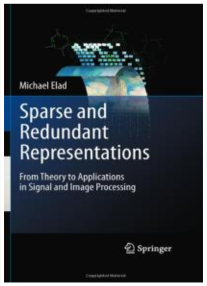
- *Sparseland* has a great success in signal & image processing and machine learning tasks
- In the past 8-9 years, many books were published on this and closely related fields

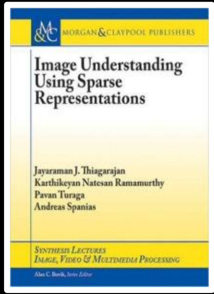


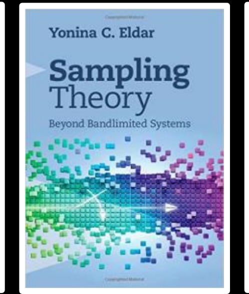


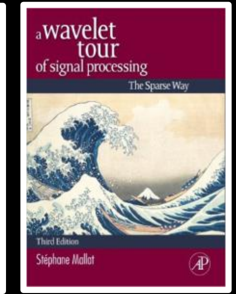


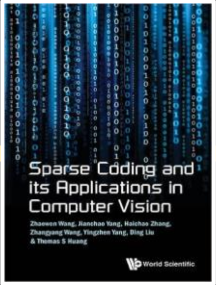


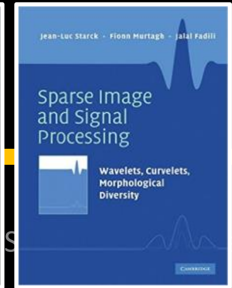


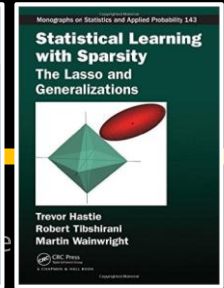


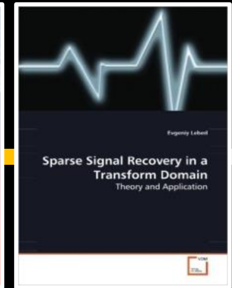


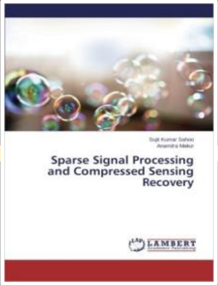


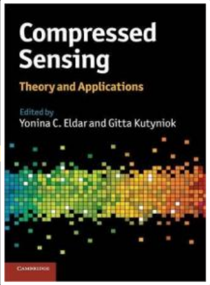


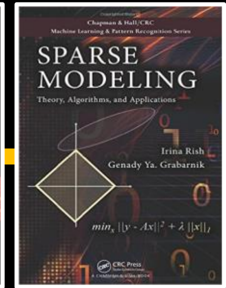












Coming Up: A Massive Open Online Course

edX Courses Programs Schools & Partners About Search: Sign In Register

Israel X
The National Online Education Initiative

Sparse Representations in Signal and Image Processing

Learn the theory, tools and algorithms of sparse representations and their impact on signal and image processing.

Start the Professional Certificate Program

Courses in the Professional Certificate Program

Sparse Representations in Signal and Image Processing: Fundamentals
Learn about the field of sparse representations by understanding its fundamental theoretical and algorithmic foundations.
Learn more

Starts on October 25, 2017
Enroll Now
I would like to receive email from IsraelX and learn about other offerings related to Sparse Representations in Signal and Image Processing: Fundamentals.

Sparse Representations in Image Processing: From Theory to Practice
Learn about the deployment of the sparse representation model to signal and image processing.
Learn more

Starts on February 28, 2018
Enroll Now
I would like to receive email from IsraelX and learn about other offerings related to Sparse Representations in Image Processing: From Theory to Practice.

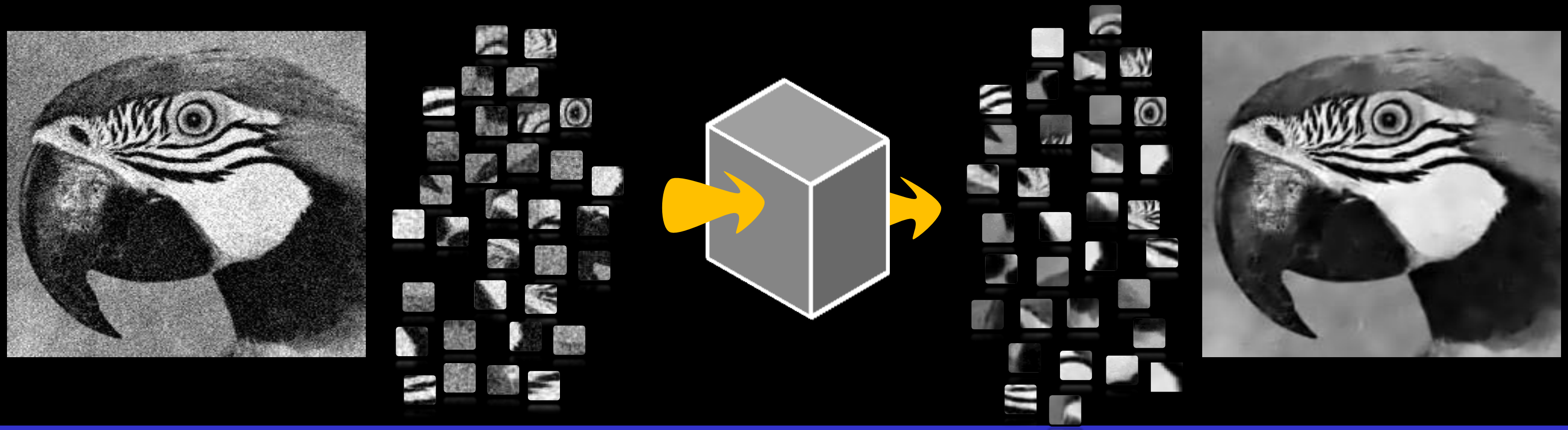
Instructors

Yaniv Romano

Michael Elad

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

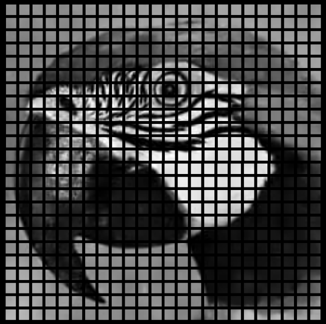

Vardan Papyan


Jeremias Sulam

Convolutional Sparse Coding (CSC)

m filters convolved with their sparse representations

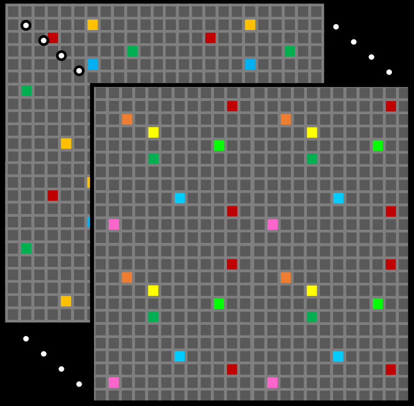
i-th feature-map: An image of the same size as $\mathbf{X}$ holding the sparse representation related to the i-filter



An image with N pixels

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

The i-th filter of small size n



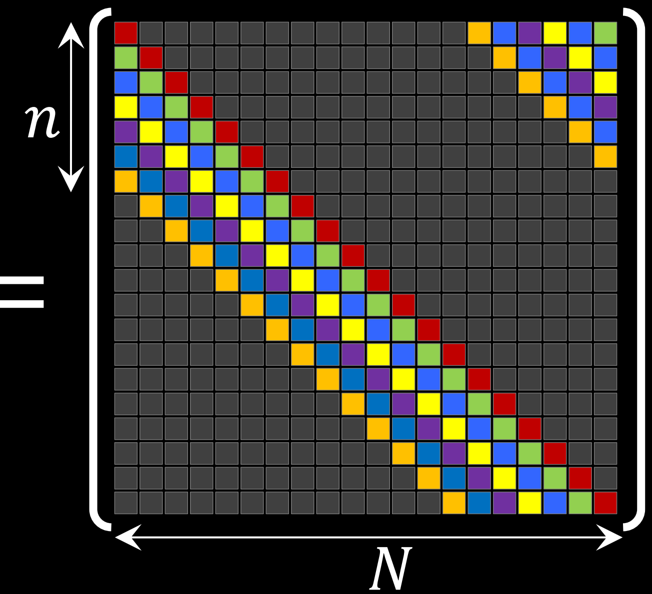
CSC in Matrix Form

- Here is an alternative global sparsity-based model formulation

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

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

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

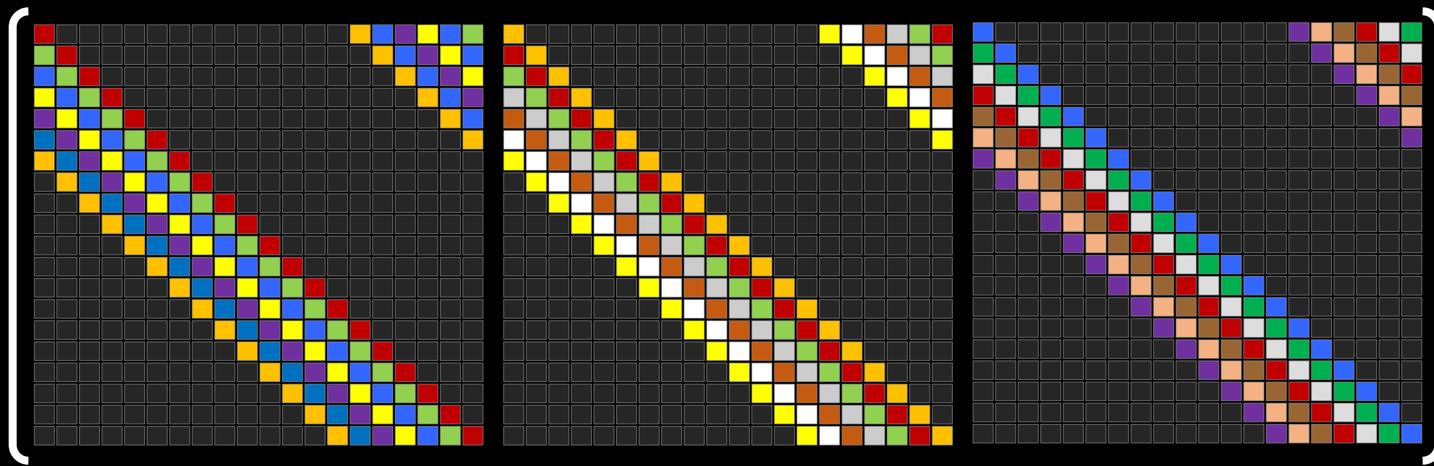


- $\mathbf{\Gamma}^{\mathrm{i}} \in \mathbb{R}^{N}$ are the corresponding coefficients ordered as column vectors

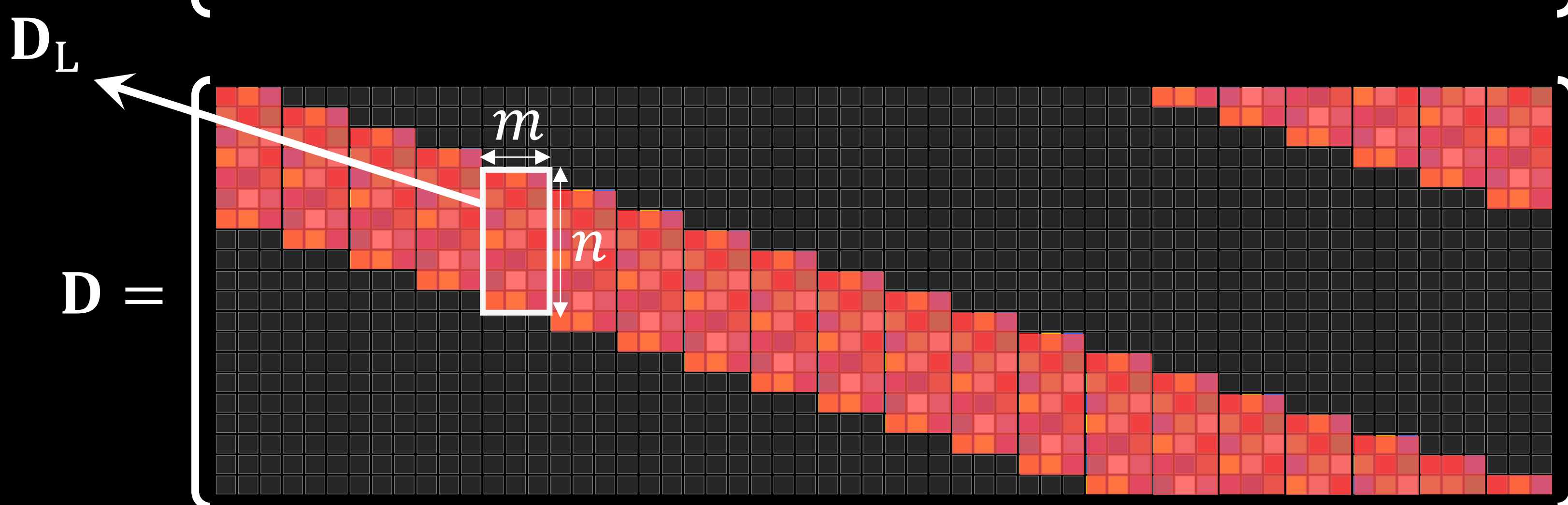
Michael Elad
The Computer-Science Department
The Technion

The CSC Dictionary

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

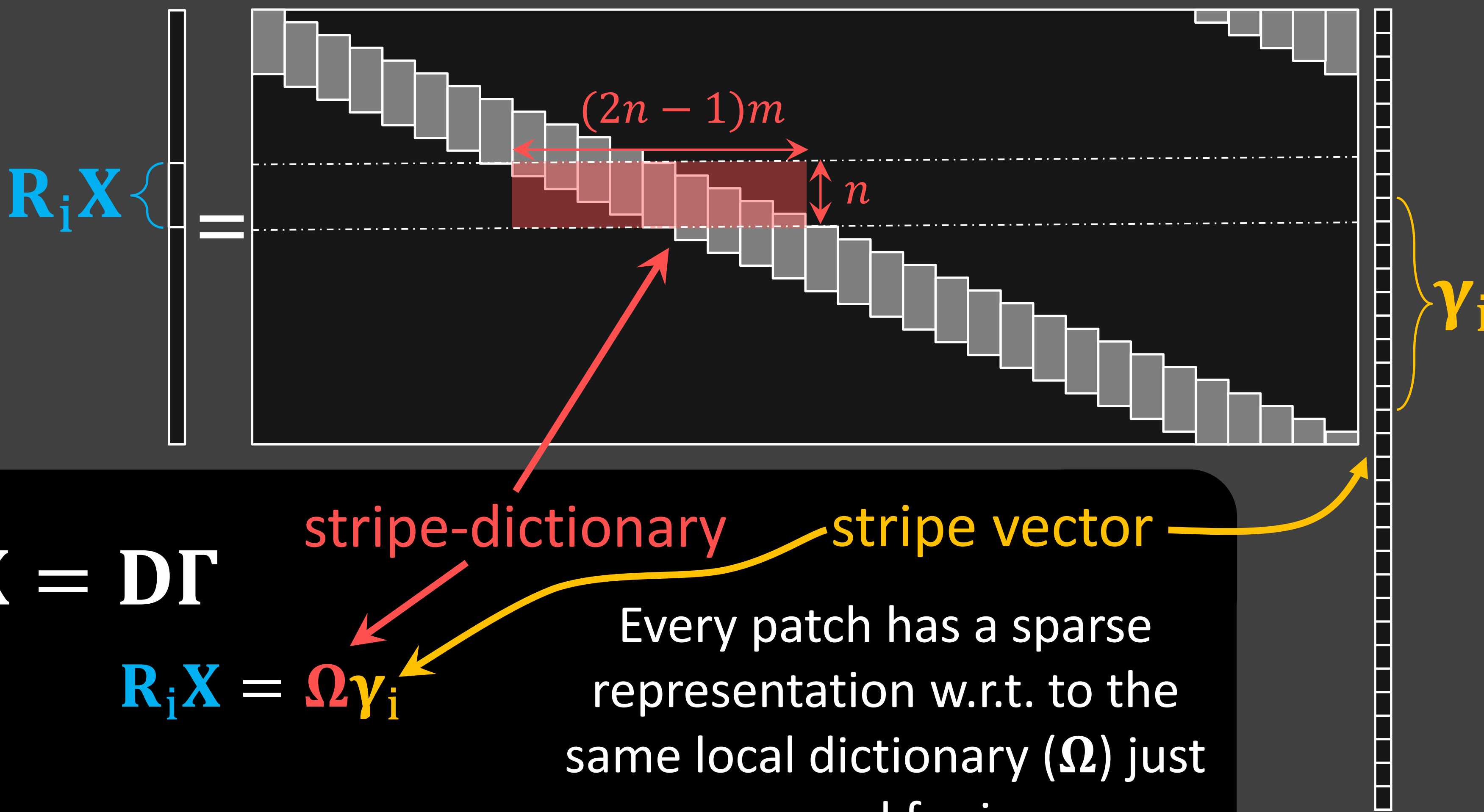


$\mathbf{D} =$



Michael Elad
The Computer-Science Department
The Technion

Why CSC?

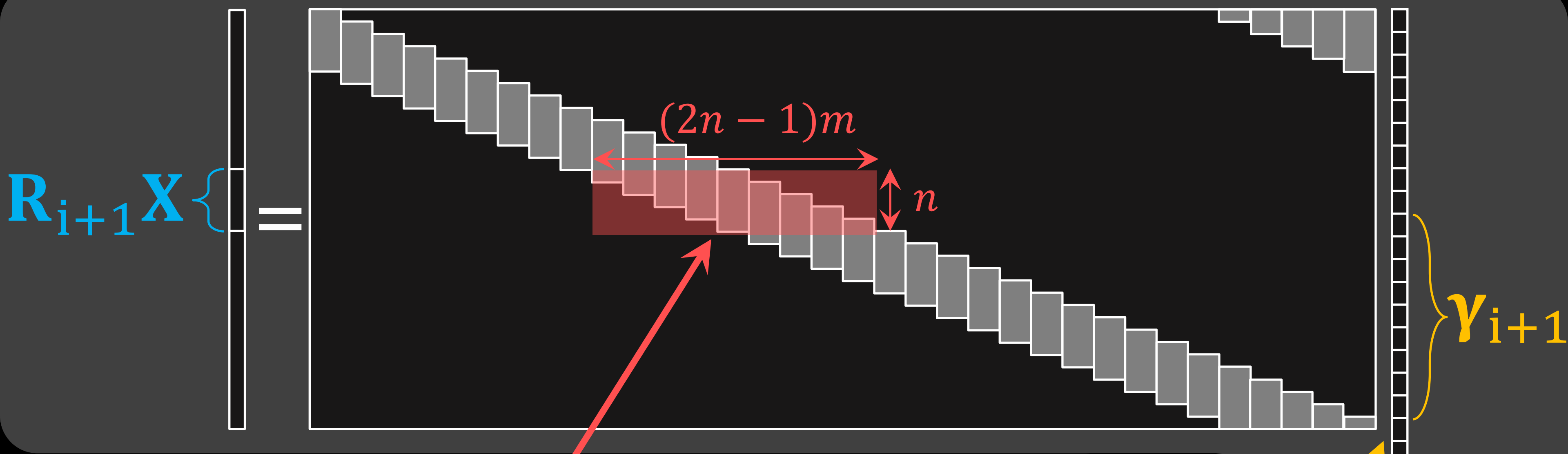


$$\mathbf{X} = \mathbf{D\Gamma}$$

$$\mathbf{R_i X} = \mathbf{\Omega}\boldsymbol{\gamma}_\mathrm{i}$$

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

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

Classical Sparse Theory for CSC ?

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

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

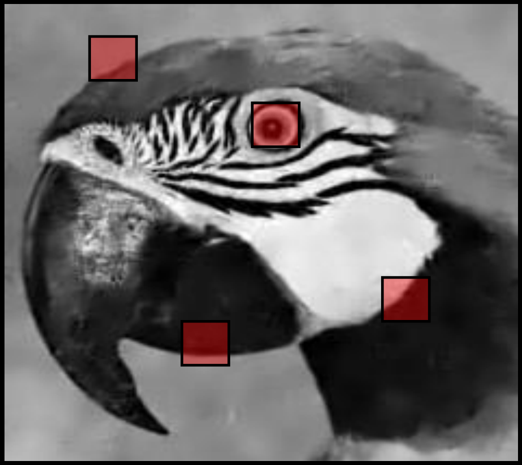
- Assuming that $m = 2$ and $n = 64$ we have that [Welch, '74]

$$\mu \ge 0.063$$

- Success of pursuits is guaranteed as long as

$$\|\mathbf{\Gamma}\|_0 < \frac{1}{4}\left(1+\frac{1}{\mu(\mathbf{D})}\right) \le \frac{1}{2}\left(1+\frac{1}{0.063}\right) \approx 4.2$$

- Only few (4) non-zeros GLOBALLY are allowed!!! This is a very pessimistic result!



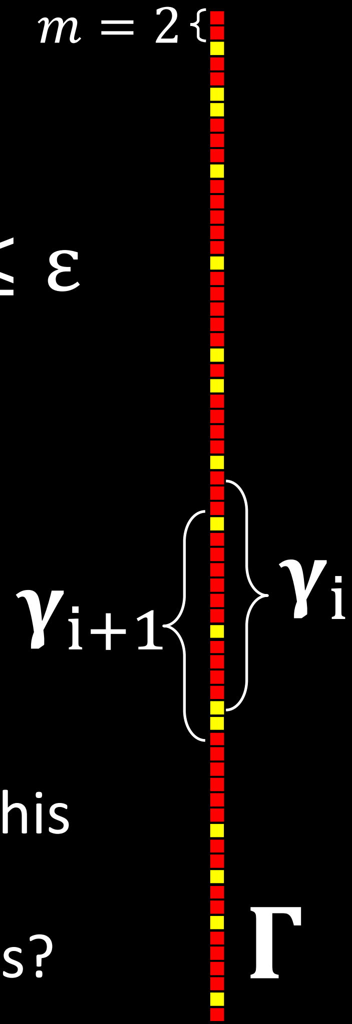
- The classic *Sparseland* Theory does not cover well the CSC model

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 OMP

Local noise (per patch)

Theorem: If $\mathbf{Y} = \mathbf{D\Gamma} + \mathbf{E}$ where

$$\|\mathbf{\Gamma}\|_{0,\infty}^{s} < \frac{1}{2}\left(1 + \frac{1}{\mu}\right) - \frac{1}{\mu} \cdot \frac{\|\mathbf{E}\|_{2,\infty}^{p}}{|\mathbf{\Gamma}_{\min}|}$$

then **OMP** run for $\|\mathbf{\Gamma}\|_0$ iterations

1. **Finds the correct support**
2. $\|\mathbf{\Gamma}_{\mathrm{OMP}} - \mathbf{\Gamma}\|_2^2 \leq \frac{\|\mathbf{E}\|_2^2}{1-\left(\|\mathbf{\Gamma}\|_{0,\infty}^{s}-1\right)\mu}$

Papyan, Sulam & Elad ('17)

This is a much better result – it allows few non-zeros locally in each stripe, implying a permitted $O(N)$ non-zeros globally

Success of the Basis Pursuit

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

Recent works tackling the convolutional sparse coding problem via BP

[Bristow, Eriksson & Lucey ‘13]
[Wohlberg ‘14]
[Kong & Fowlkes ‘14]
[Bristow & Lucey ‘14]
[Heide, Heidrich & Wetzstein ‘15]
[Šorel & Šroubek ‘16]

Success of the Basis Pursuit

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

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

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

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)

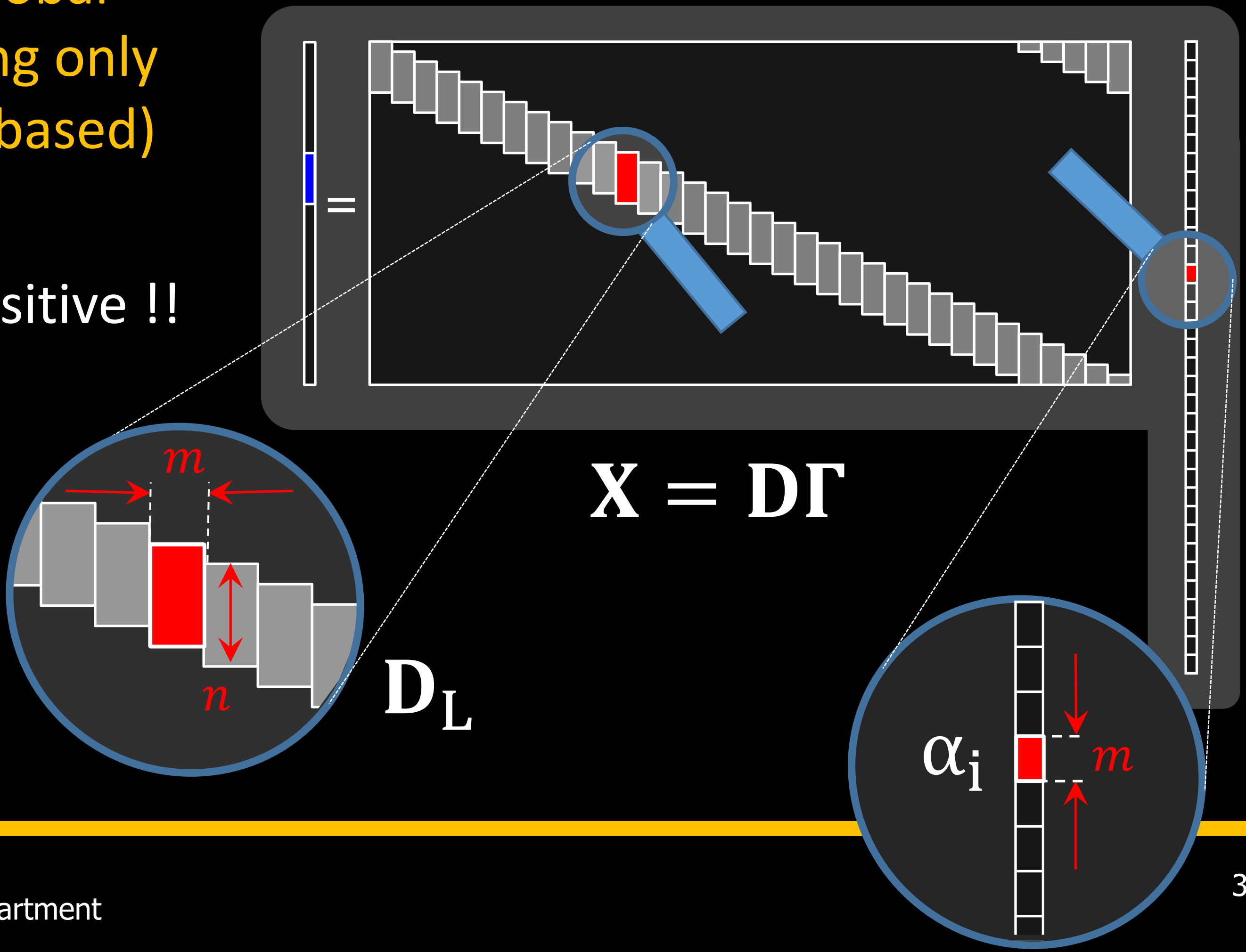
SKIP ?

Global Pursuit via Local Processing

- Could we suggest a solution of the global Basis Pursuit using only local (e.g. patch-based) operations ?
- The answer is positive !!
- We define image slices :

$$s_i \equiv \mathbf{D}_L \alpha_i$$

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



Global Pursuit via Local Processing

$$(\mathbf{P}_1^{\epsilon}): \quad \mathbf{\Gamma}_{\mathrm{BP}} = \min_{\mathbf{\Gamma}} \ \frac{1}{2}\|\mathbf{Y} - \mathbf{D\Gamma}\|_2^2 + \lambda\|\mathbf{\Gamma}\|_1$$

These two are convex & equivalent

Redefine this problem using s_i and α_i

$$\min_{\boldsymbol{\alpha}_i, \mathbf{s}_i} \ \frac{1}{2}\left\|\mathbf{Y} - \sum_i \mathbf{R}_i^T \mathbf{s}_i\right\|_2^2 + \lambda \sum_i \|\boldsymbol{\alpha}_i\|_1 \quad \text{s.t.} \quad \{\mathbf{s}_i = \mathbf{D}_L \boldsymbol{\alpha}_i\}_i$$

Update the $\boldsymbol{\alpha}_i$ by a local BP

Update the slices $\mathbf{s}_i$ by a simple LS & patch-averaging

If you apply the above two steps only once, you get a known patch-based denoising algorithm

Global Pursuit via Local Processing

$$(P_1^{\epsilon}):\quad \Gamma_{BP} = \min_{\Gamma} \frac{1}{2}\|\mathbf{Y} - \mathbf{D}\Gamma\|_2^2 + \lambda\|\Gamma\|_1$$

These two are convex & equivalent

Redefine this problem using s_i and α_i

$$\min_{\alpha_i, s_i} \frac{1}{2}\left\|\mathbf{Y} - \sum_i \mathbf{R}_i^T \mathbf{s}_i\right\|_2^2 + \lambda \sum_i \|\boldsymbol{\alpha}_i\|_1 \quad \text{s.t.} \quad \{\mathbf{s}_i = \mathbf{D}_L \boldsymbol{\alpha}_i\}_i$$

Update the α_i by a lo…

Upd… lices s_i … LS & … ging

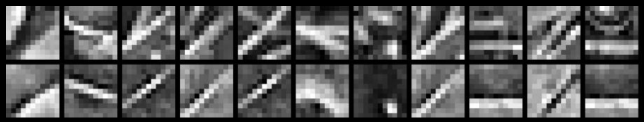
If you ap… steps only once, you get a kn… tch-based denoising algorithm

This algorithm operates locally while guaranteeing to solve the global problem

Two Comments About this Scheme

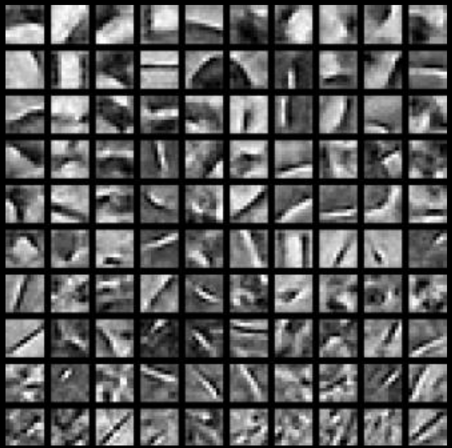
We work with Slices and not Patches

Patches extracted from natural images, and their corresponding slices. Observe how the slices are far simpler, and contained by their corresponding patches

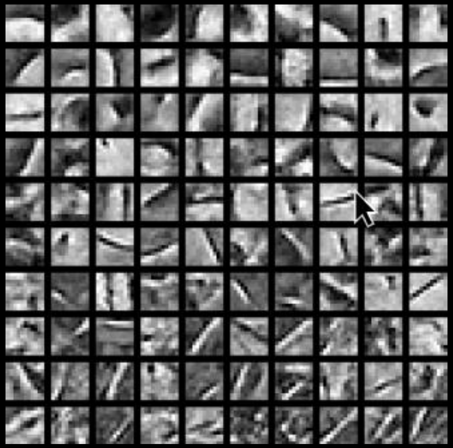


The Proposed Scheme can be used for Dictionary ($\mathrm{D_L}$) Learning

Slice-based DL algorithm using standard patch-based tools, leading to a faster and simpler method, compared to existing methods



[Wohlberg, 2016]



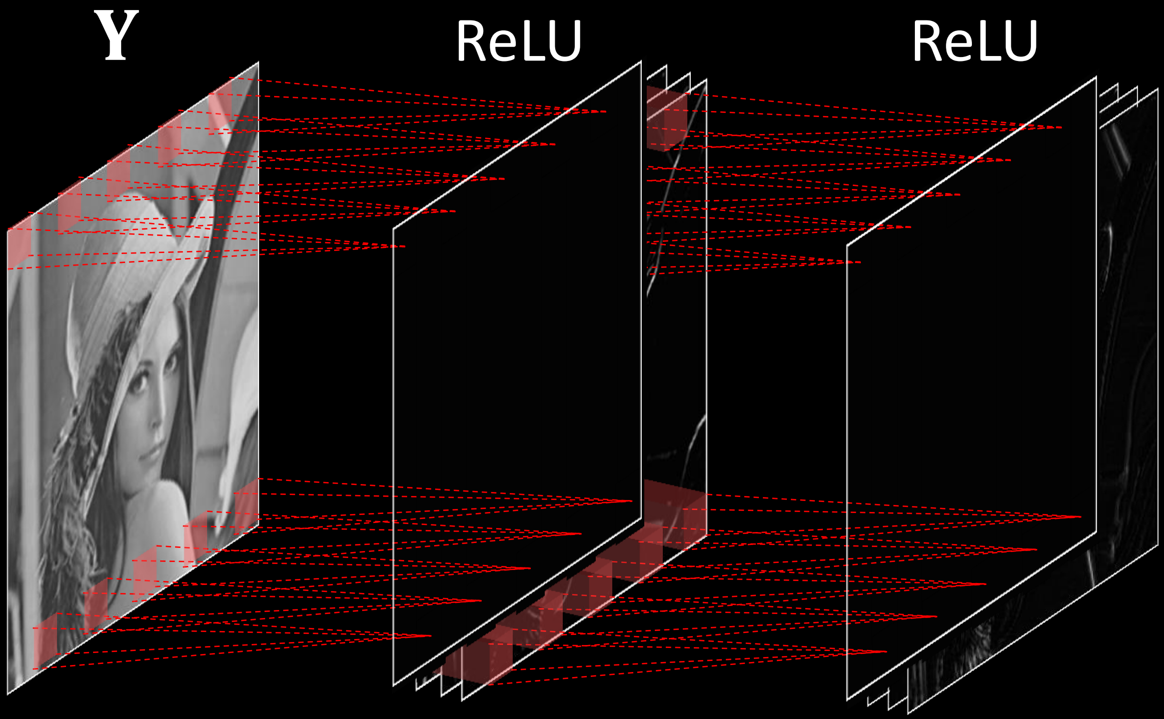
Ours

Multi-Layered Convolutional Sparse Modeling

CSC and CNN

- There is a rough analogy between CSC and CNN:
  - Convolutional structure
  - Data driven models
  - ReLU is a sparsifying operator
- We shall now propose a principled way to analyze CNN
- But first, a brief review of CNN…

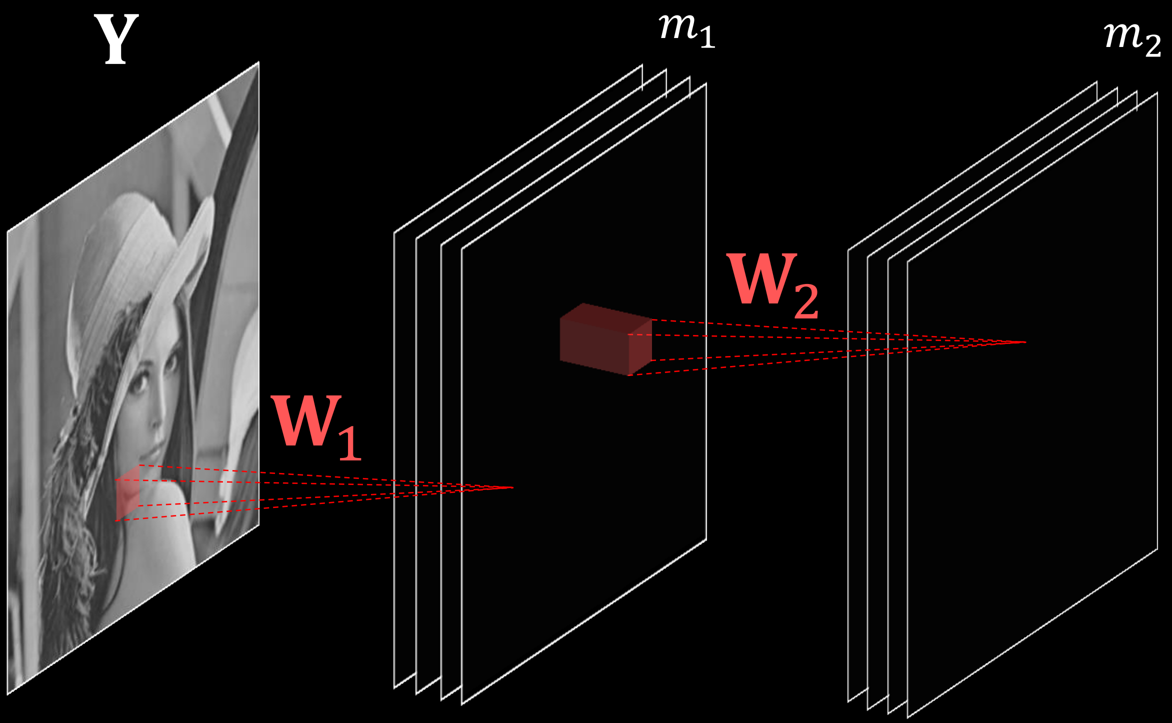
CNN



[LeCun, Bottou, Bengio and Haffner '98]
[Krizhevsky, Sutskever & Hinton '12]
[Simonyan & Zisserman '14]
[He, Zhang, Ren & Sun '15]

$$\mathrm{ReLU}(z) = \max(\mathrm{Thr}, z)$$

CNN

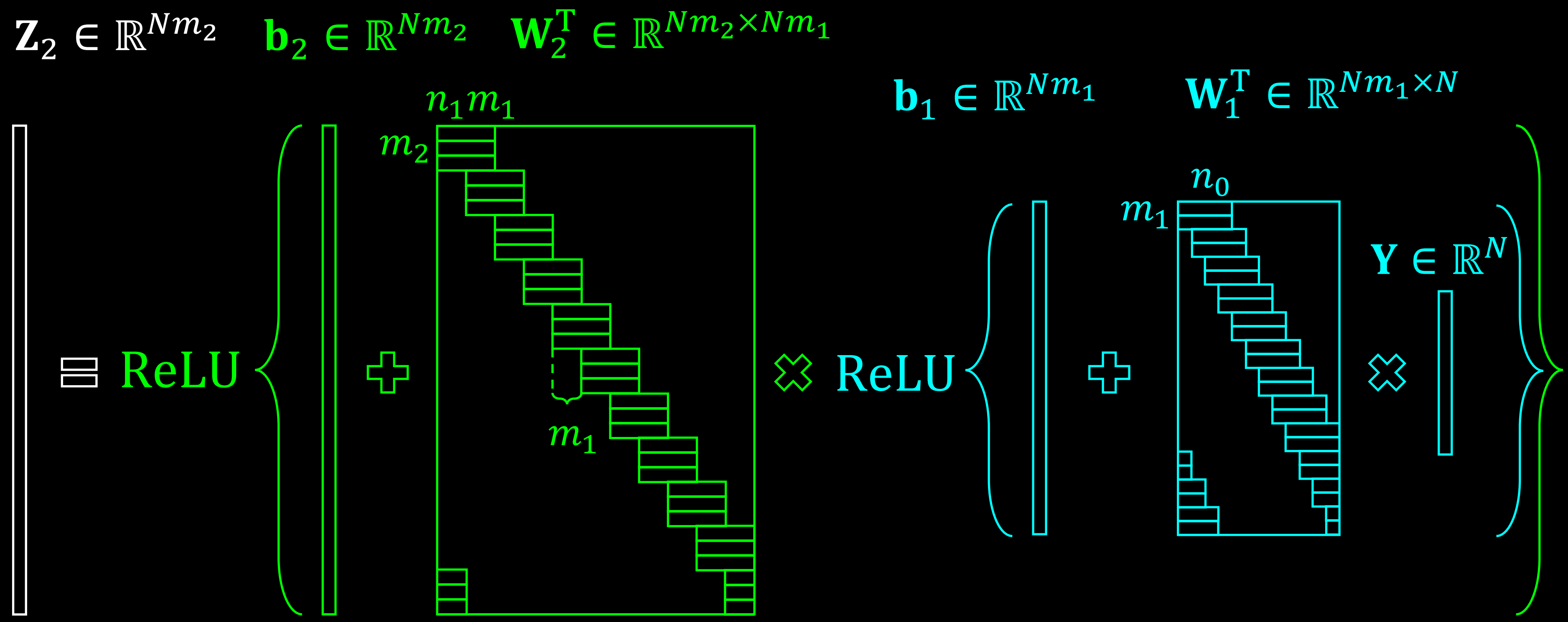


[LeCun, Bottou, Bengio and Haffner '98]
[Krizhevsky, Sutskever & Hinton '12]
[Simonyan & Zisserman '14]
[He, Zhang, Ren & Sun '15]

$$\mathrm{ReLU}(z) = \max(\mathrm{Thr}, z)$$

Mathematically...

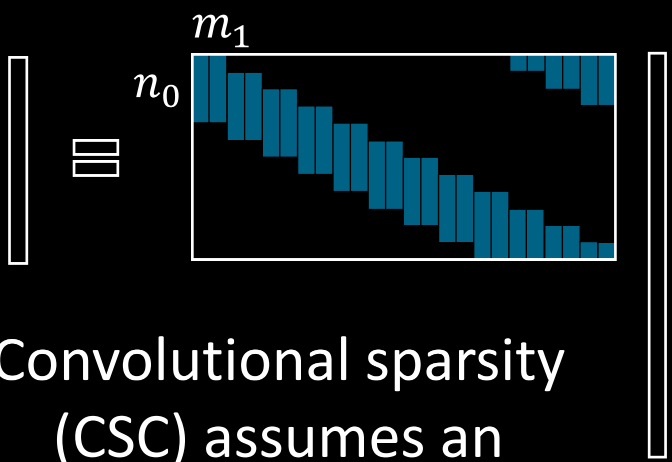
$$f(\mathbf{Y}) = \text{ReLU}\left(\mathbf{b}_2 + \mathbf{W}_2^{\text{T}}\,\text{ReLU}\left(\mathbf{b}_1 + \mathbf{W}_1^{\text{T}}\mathbf{Y}\right)\right)$$



Michael Elad
The Computer-Science Department
The Technion

From CSC to Multi-Layered CSC

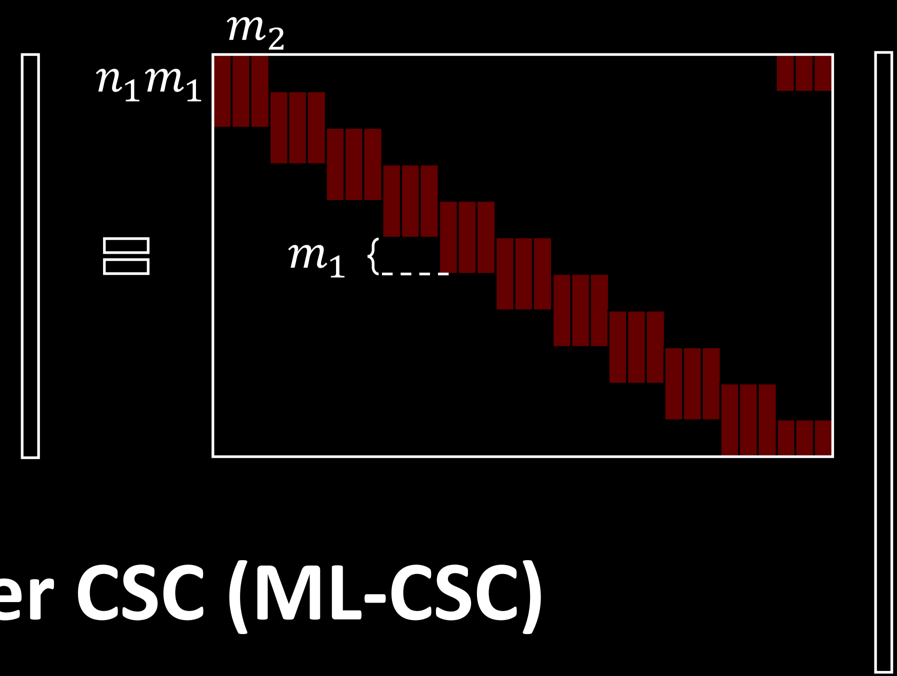
$\mathbf{X} \in \mathbb{R}^{N}$ $\mathbf{D}_1 \in \mathbb{R}^{N \times N m_1}$ $\mathbf{\Gamma}_1 \in \mathbb{R}^{N m_1}$



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

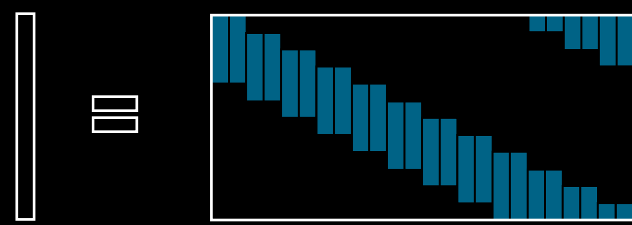
We propose to impose the same structure on the representations themselves

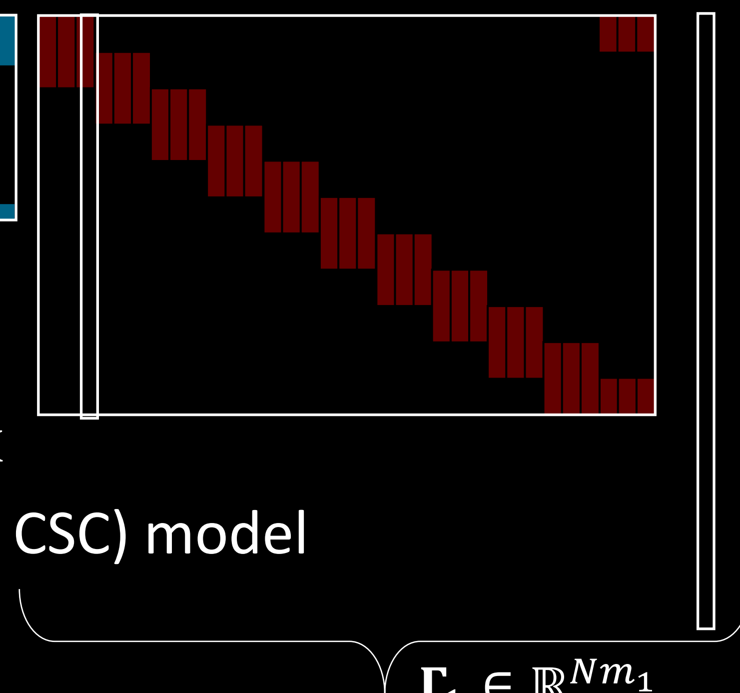
$\mathbf{\Gamma}_1 \in \mathbb{R}^{N m_1}$ $\mathbf{D}_2 \in \mathbb{R}^{N m_1 \times N m_2}$ $\mathbf{\Gamma}_2 \in \mathbb{R}^{N m_2}$



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 → molecules → cells → tissue → body-parts …

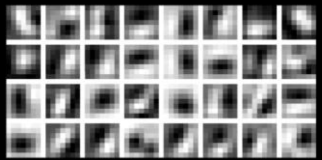
Michael Elad
The Computer-Science Department
The Technion

A Small Taste: Model Training (MNIST)

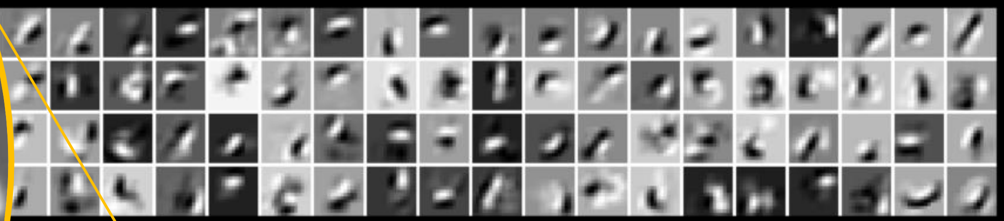
MNIST Dictionary:

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

$\mathbf{D}_1$ (7×7)

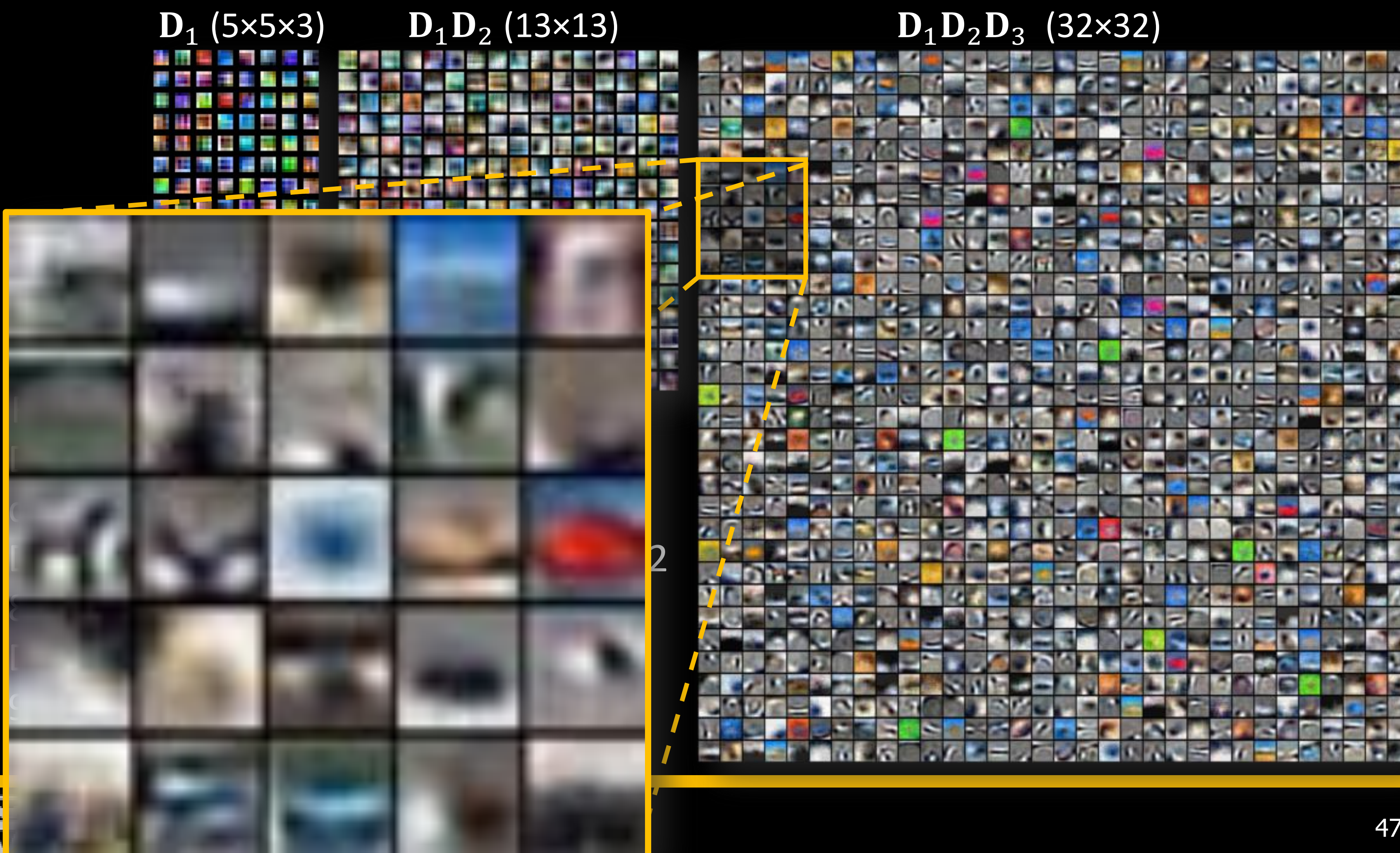


$\mathbf{D}_1\mathbf{D}_2$ (15×15)



$\mathbf{D}_1\mathbf{D}_2\mathbf{D}_3$ (28×28)

A Small Taste: Model Training (CiFAR)



ML-CSC: Pursuit

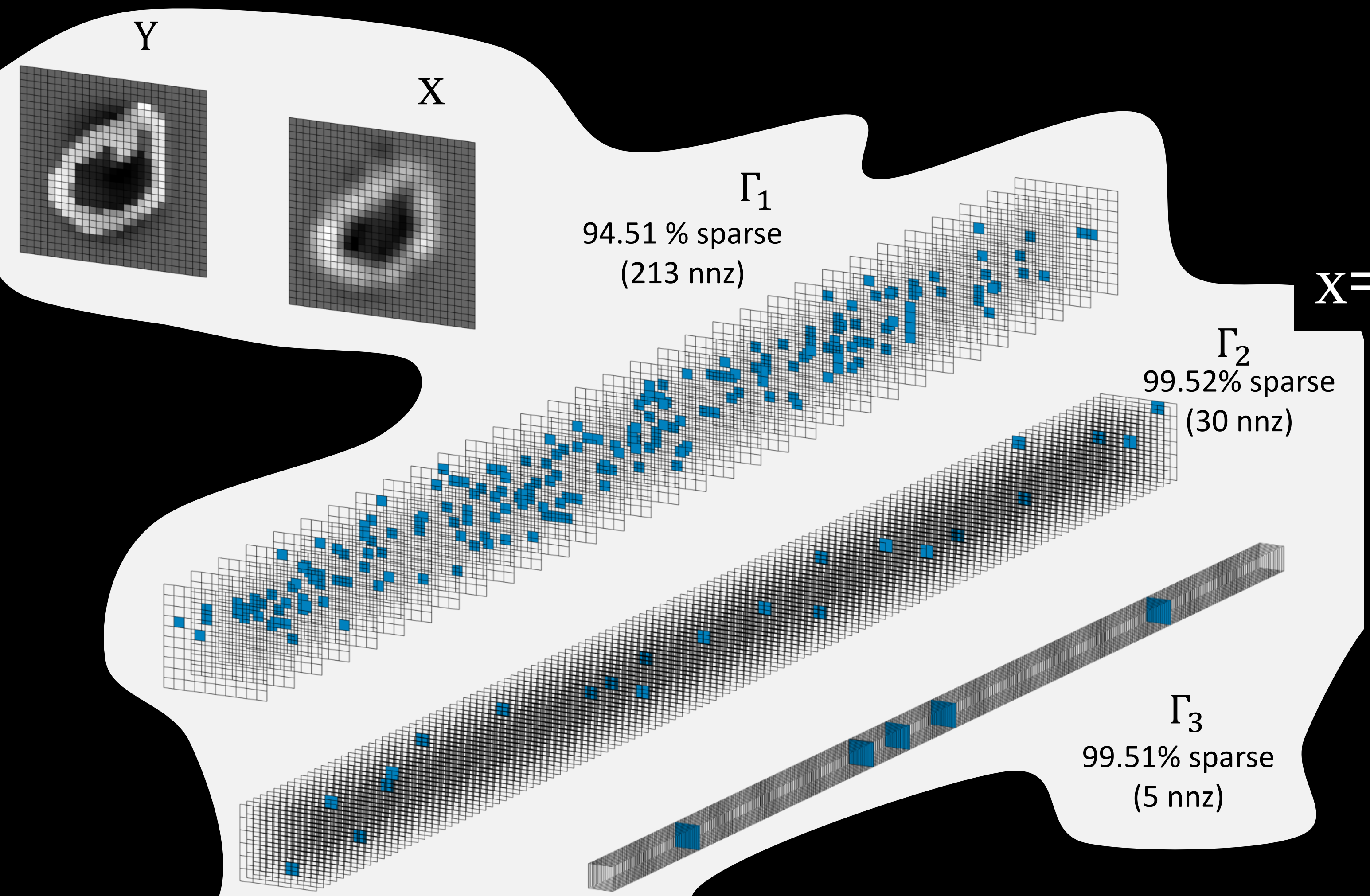
- Deep–Coding Problem $(\mathbf{DCP_\lambda})$ (dictionaries are known):

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

- Or, more realistically for noisy signals,

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

A Small Taste: Pursuit



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

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

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

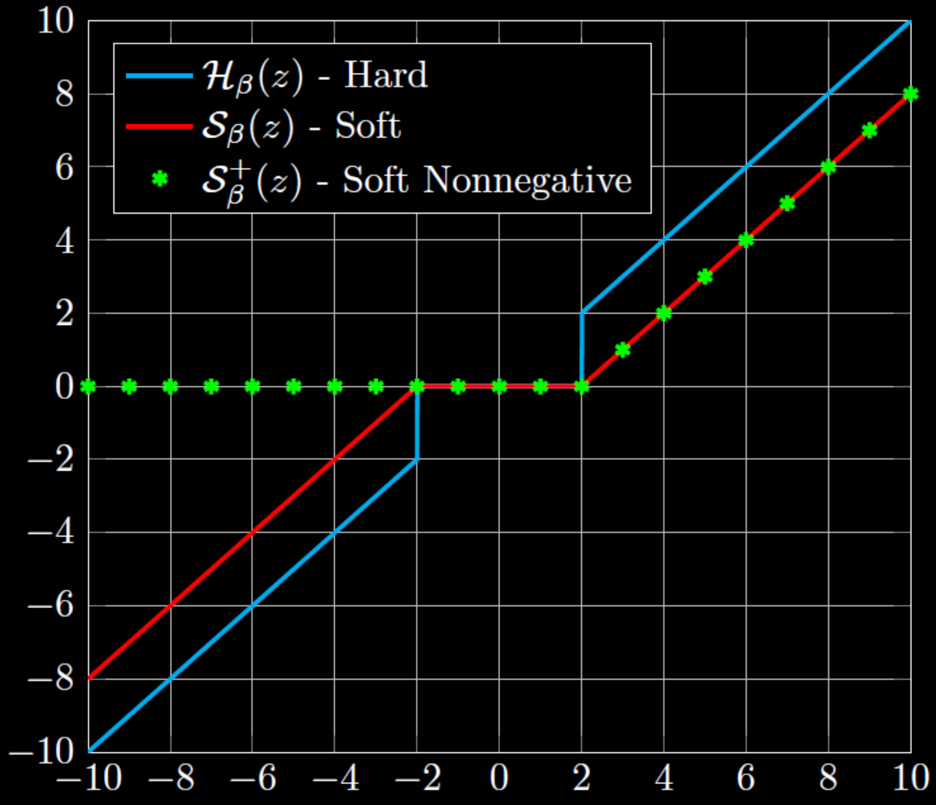
ML-CSC: The Simplest Pursuit

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

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



Consider this for Solving the DCP

- Layered thresholding (LT):

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

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

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

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

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

- 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 questions

Theoretical Path: Possible Questions

- Having established the importance of the ML-CSC model and its associated pursuit, the DCP problem, we now turn to its analysis

- The main questions we aim to address:

I. Stability of the solution obtained via the hard layered THR algorithm (forward pass) ?

II. Limitations of this (very simple) algorithm and alternative pursuit?

… and here are questions we will not touch today:

III. Algorithms for training the dictionaries $\{\mathbf{D}_i\}_{i=1}^{K}$ vs. CNN ?

IV. New insights on how to operate on signals via CNN ?

Success of the Layered-THR

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

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

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

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 …

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

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

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

$$\downarrow$$

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

$$\downarrow$$

$$\vdots$$

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

Success of the Layered BP

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

then the Basis Pursuit performs very well:

1. The support of $\boldsymbol{\Gamma}_{\mathrm{i}}^{\mathrm{LBP}}$ is contained in that of $\boldsymbol{\Gamma}_{\mathrm{i}}$
2. The error is bounded: $\|\boldsymbol{\Gamma}_{\mathrm{i}}^{\mathrm{LBP}} - \boldsymbol{\Gamma}_{\mathrm{i}}\|_{2,\infty}^{\mathrm{p}} \leq \varepsilon_{\mathrm{L}}^{\mathrm{i}}$, where

   $\varepsilon_{\mathrm{L}}^{\mathrm{i}} = 7.5^{\mathrm{i}}\|\mathbf{E}\|_{2,\infty}^{\mathrm{p}} \prod_{\mathrm{j}=1}^{\mathrm{i}} \sqrt{\|\boldsymbol{\Gamma}_{\mathrm{j}}\|_{0,\infty}^{\mathrm{p}}}$
3. Every entry in $\boldsymbol{\Gamma}_{\mathrm{i}}$ greater than

   $\varepsilon_{\mathrm{L}}^{\mathrm{i}} / \sqrt{\|\boldsymbol{\Gamma}_{\mathrm{i}}\|_{0,\infty}^{\mathrm{p}}}$ will be found

Problems:

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

Papyan, Romano & Elad ('17)

Layered Iterative Thresholding

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

Layered Iterative Soft-Thresholding:

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

Time to Conclude

This Talk

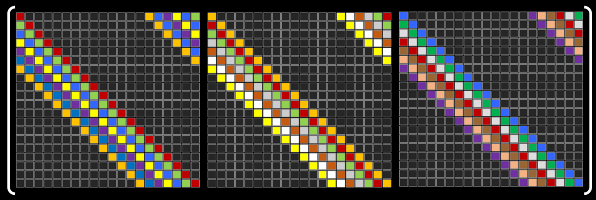
Sparseland

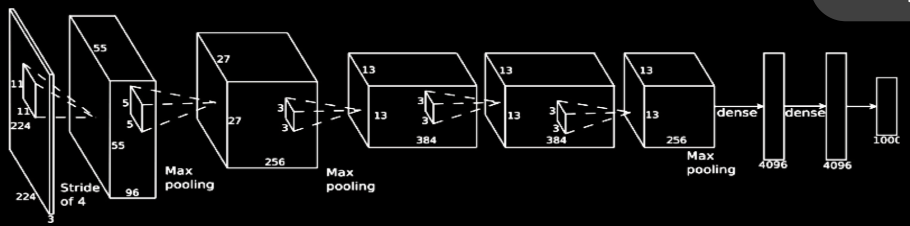
The desire to model data

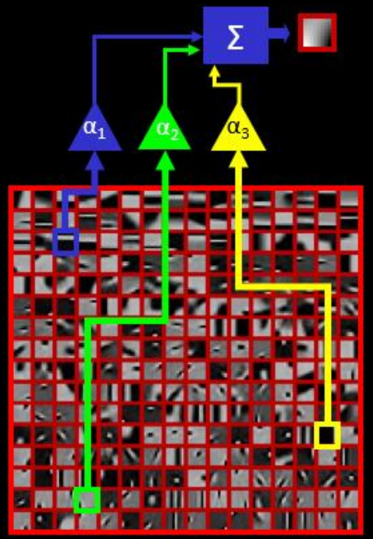
Novel View of Convolutional Sparse Coding

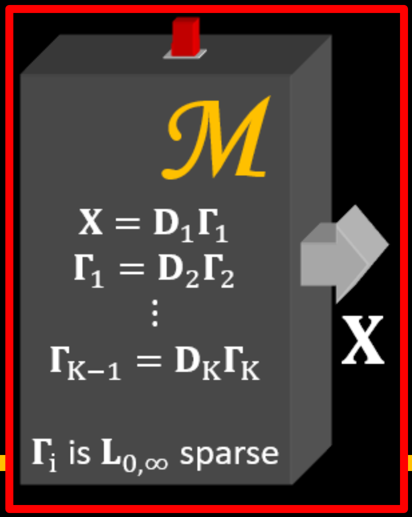
Multi-Layer Convolutional Sparse Coding

A novel interpretation and theoretical understanding of CNN









Michael Elad
The Computer-Science Department
The Technion

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:

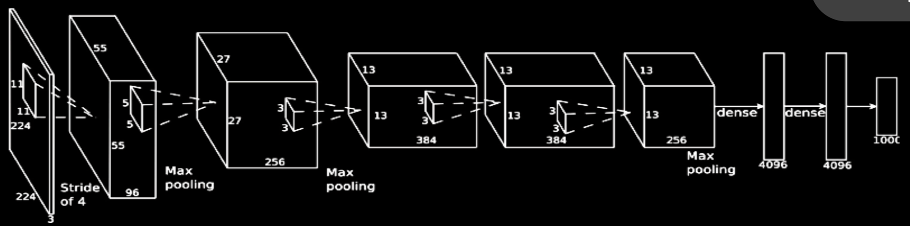
Generative modeling of data sources enables algorithm development along with theoretically analyzing algorithms' performance

This Talk

Sparseland

The desire to model data

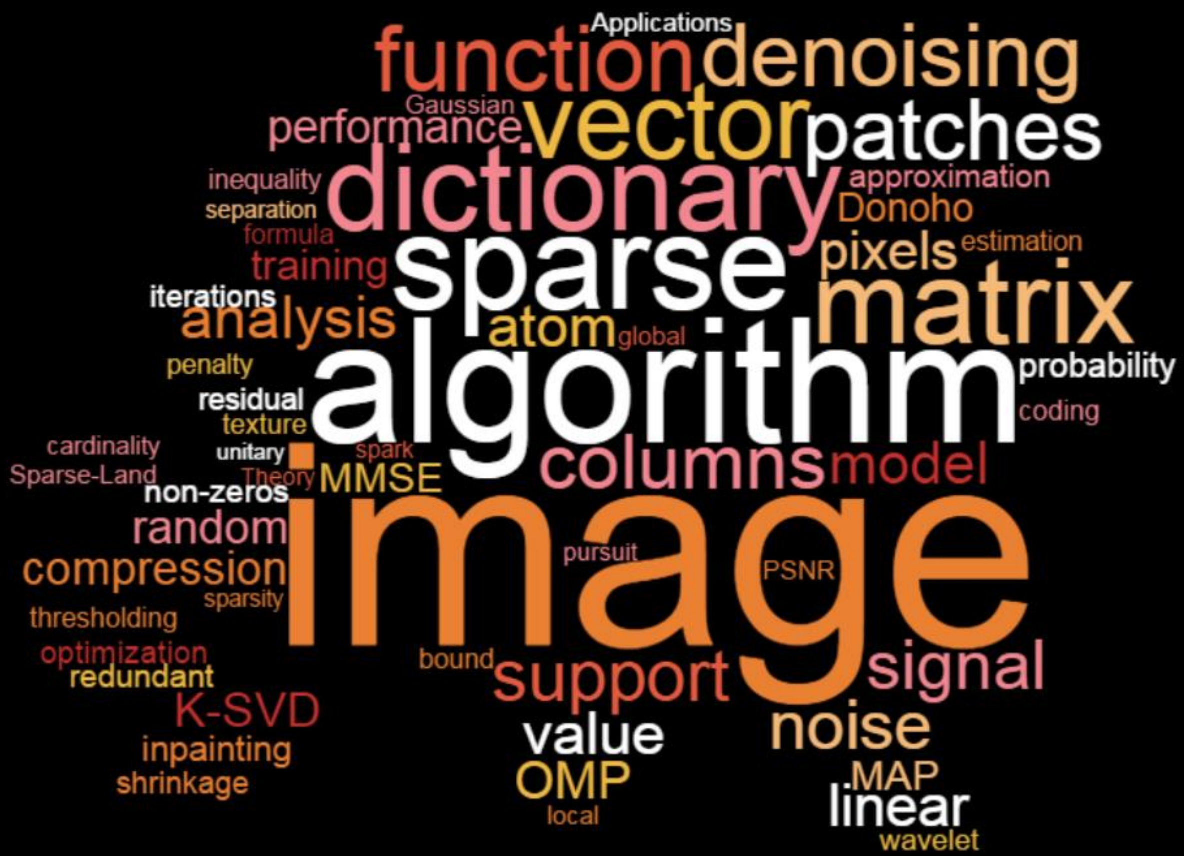
Novel View of Convolutional Sparse Coding



Multi-Layer Convolutional Sparse Coding

A novel interpretation and theoretical understanding of CNN

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



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