Sparse Modeling of Data

and its Relation to

Deep Learning

Michael Elad

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

Tuesday, November 27th 2018

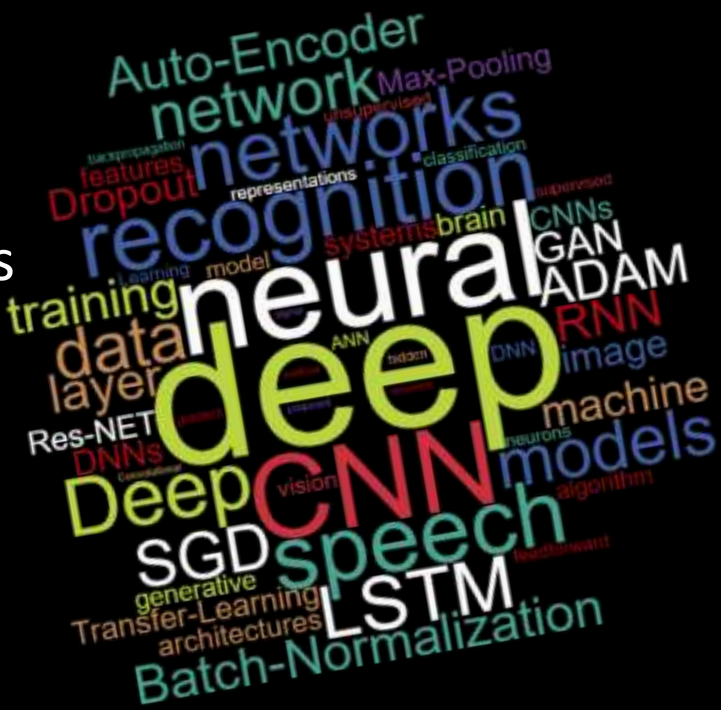
The research leading to these results has been received funding from the European union's Seventh Framework Program (FP/2007-2013) ERC grant Agreement ERC-SPARSE- 320649

This Lecture is About …

A Proposed Theory for Deep-Learning (DL)

Explanation:

- DL has been extremely successful in solving a variety of learning problems
- DL is an empirical field, with numerous tricks and know-how, but almost no theoretical foundations
- A theory for DL has become the holy-grail of current research in Machine-Learning and related fields



Michael Elad
The Computer-Science Department
The Technion

Who Needs Theory ?

We All Do !!

… because … A theory

- … could bring the next rounds of ideas to this field, breaking existing barriers and opening new opportunities
- … could map clearly the limitations of existing DL solutions, and point to key features that control their performance
- … could remove the feeling with many of us that DL is a "dark magic", turning it into a solid scientific discipline

Ali Rahimi: NIPS 2017 Test-of-Time Award

"Machine learning has become alchemy"

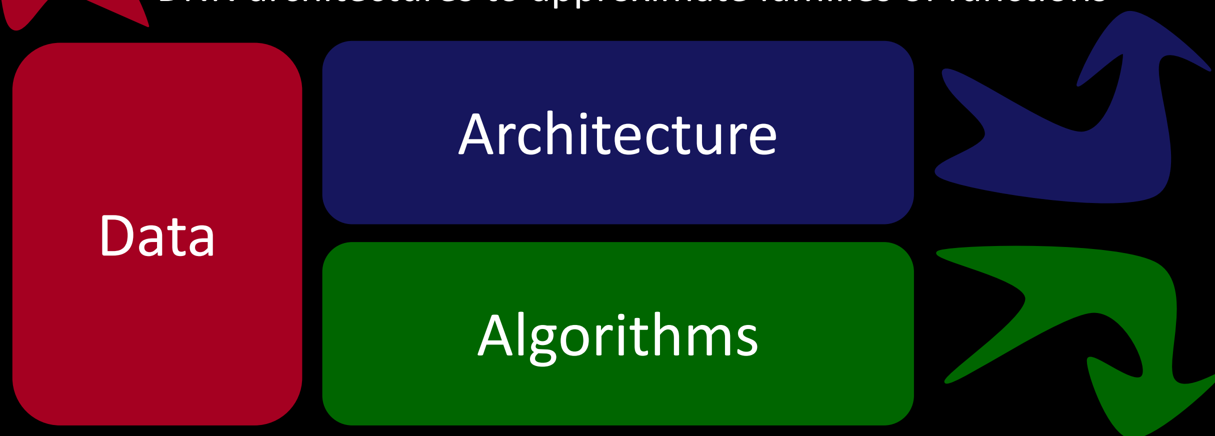
Yan LeCun

Understanding is a good thing … but another goal is inventing methods. In the history of science and technology, engineering preceded theoretical understanding:

- Lens & telescope → Optics
- Steam engine → Thermodynamics
- Airplane → Aerodynamics
- Radio & Comm. → Info. Theory
- Computer → Computer Science

Michael Elad
The Computer-Science Department
The Technion

A Theory for DL ?



Raja Giryes (TAU): Studied the architecture of DNN in the context of their ability to give distance-preserving embedding of signals

Gitta Kutyniok (TU) & Helmut Bolcskei (ETH): Studied the ability of DNN architectures to approximate families of functions

Stephane Mallat (ENS) & Joan Bruna (NYU): Proposed the scattering transform (wavelet-based) and emphasized the treatment of invariances in the input data

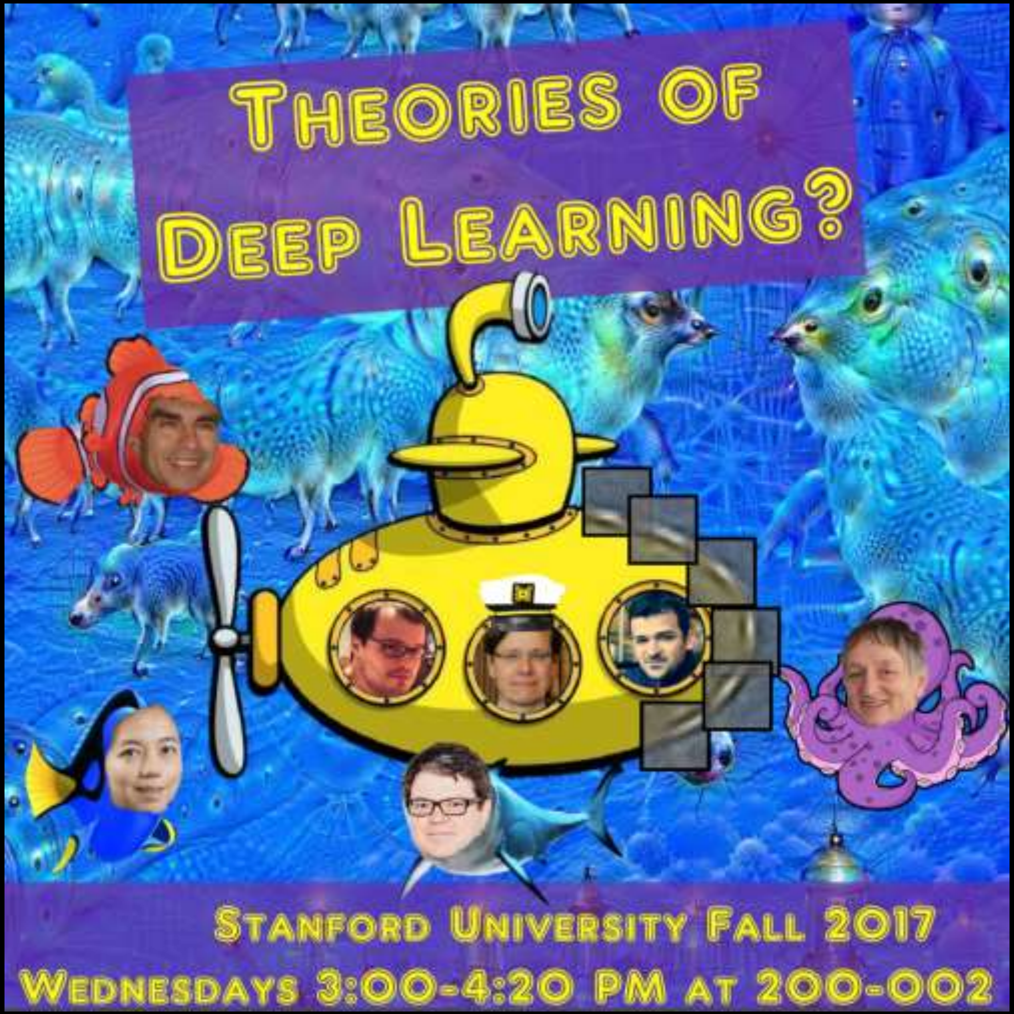
Richard Baraniuk & Ankit Patel (RICE): Offered a generative probabilistic model for the data, showing how classic architectures and learning algorithms relate to it

Rene Vidal (JHU): Explained the ability to optimize the typical non-convex objective and yet get to a global minima

Naftali Tishby (HUJI): Introduced the Information Bottleneck (IB) concept and demonstrated its relevance to deep learning

Stefano Soatto's team (UCLA): Analyzed the Stochastic Gradient Descent (SGD) algorithm, connecting it to the IB objective

So, is there a Theory for DL ?



The answer is tricky:

There are already various such attempts, and some of them are truly impressive

… but …

none of them is complete

Interesting Observations

- Theory origins: Signal Proc., Control Theory, Info. Theory, Harmonic Analysis, Sparse Represen., Quantum Physics, PDE, Machine learning …

Ron Kimmel: "*DL is a dark monster covered with mirrors. Everyone sees his reflection in it …*"

David Donoho: "*… these mirrors are taken from Cinderella's story, telling each that he is the most beautiful*"

√ Data

√ Architecture

Algorithms

- Today's talk is on our proposed theory:

Yaniv Romano

Vardan Papyan

Jeremias Sulam

… and yes, our theory is the best

This Lecture: More Specifically

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

Sparsity-Inspired Models ⟶ Deep-Learning

Another underlying idea that accompanies us

Generative modeling of data sources enables

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

Disclaimer: Being a lecture on the theory of DL, this lecture is … theoretical … and mathematically oriented

Michael Elad
The Computer-Science Department
The Technion

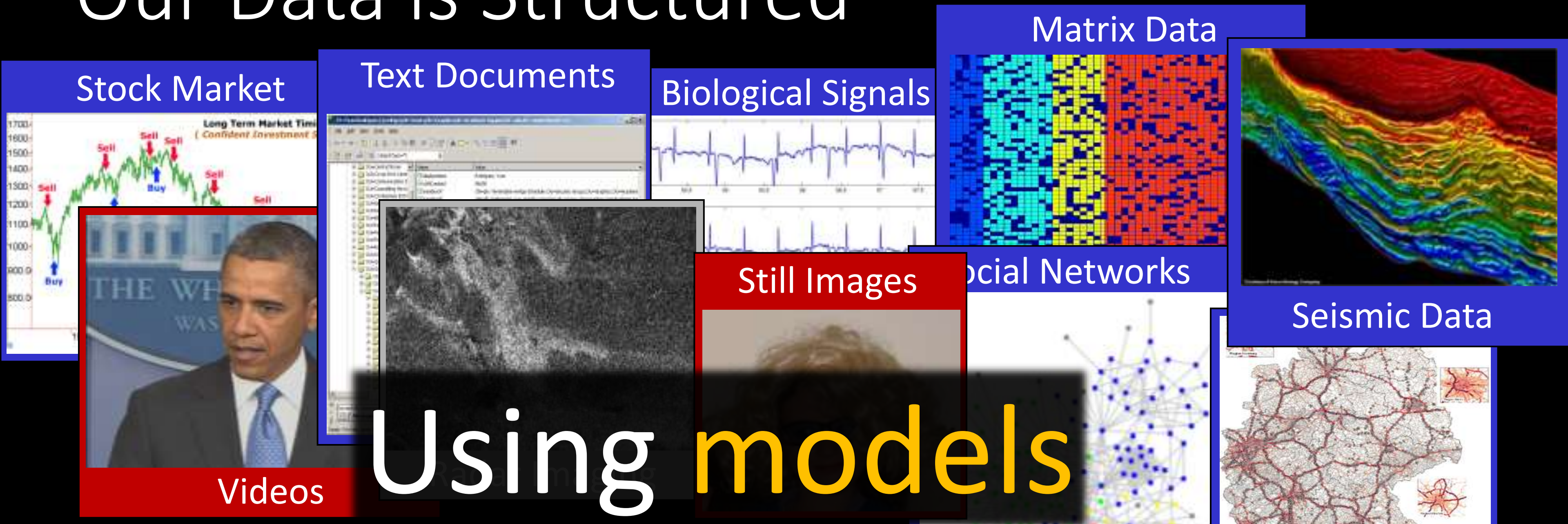
Our eventual goal in today's talk is to present the …

Multi-Layered Convolutional Sparse Modeling

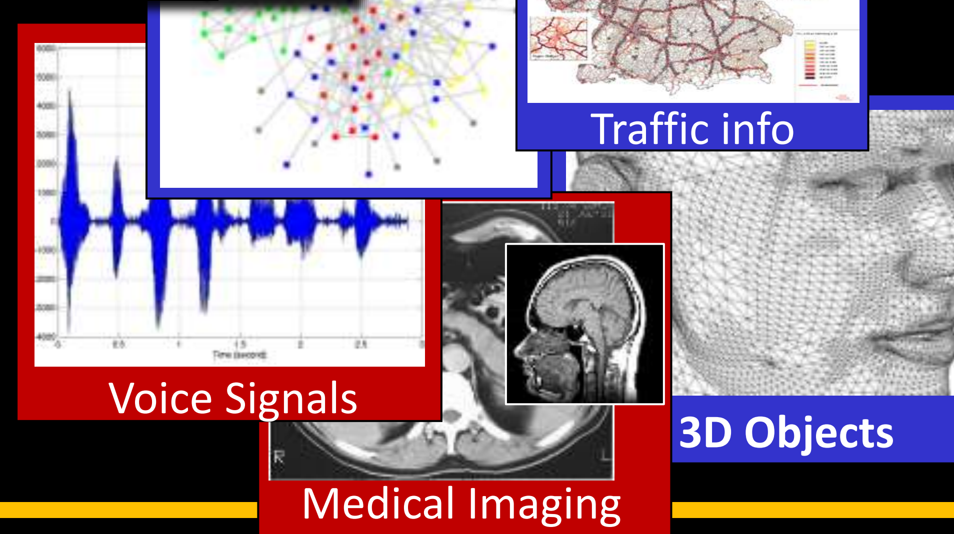
So, lets use this as our running title,
parse it into words,
and explain each of them

Multi-Layered Convolutional Sparse Modeling

Our Data is Structured



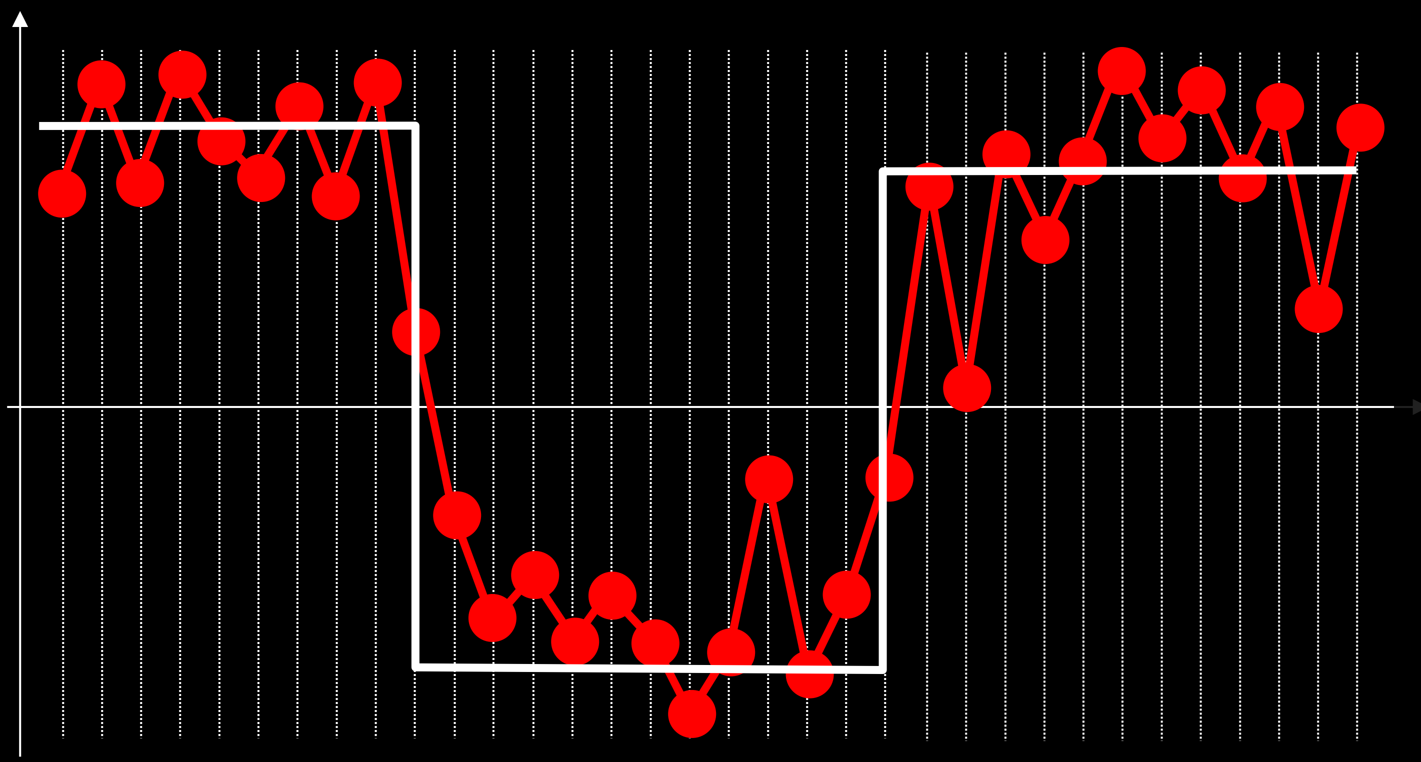
Using models



- 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 the ability to process data
- How to identify structure?

Michael Elad
The Computer-Science Department
The Technion

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

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

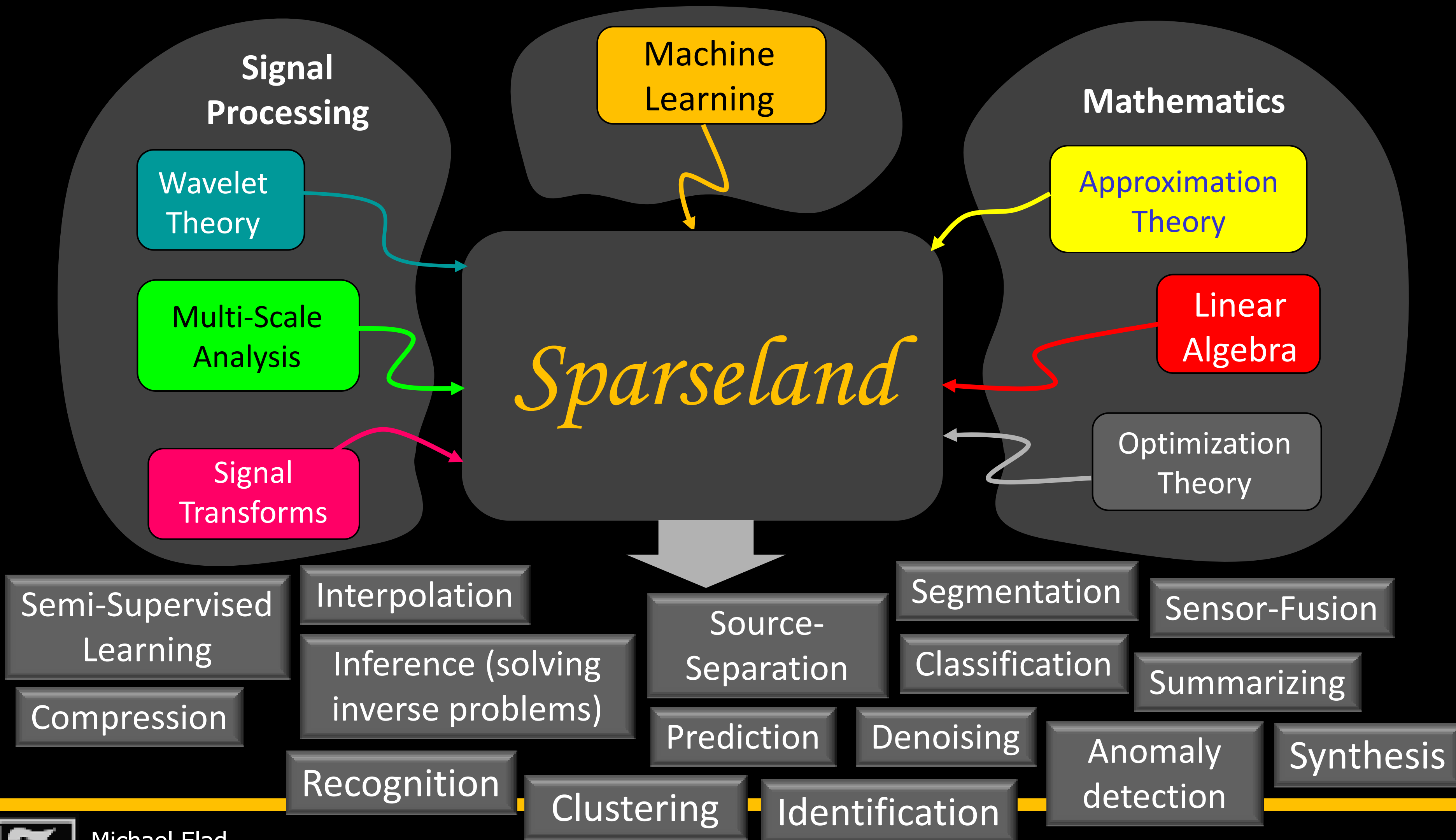
Sparseland

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

Multi-Layered Convolutional

Sparse Modeling

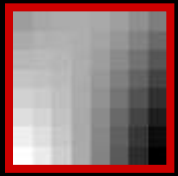
A New Emerging Model



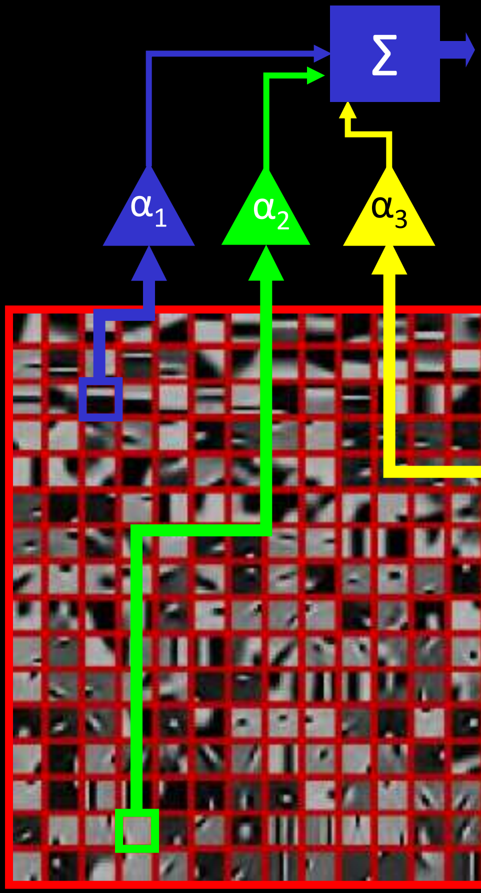
Michael Elad
The Computer-Science Department
The Technion

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

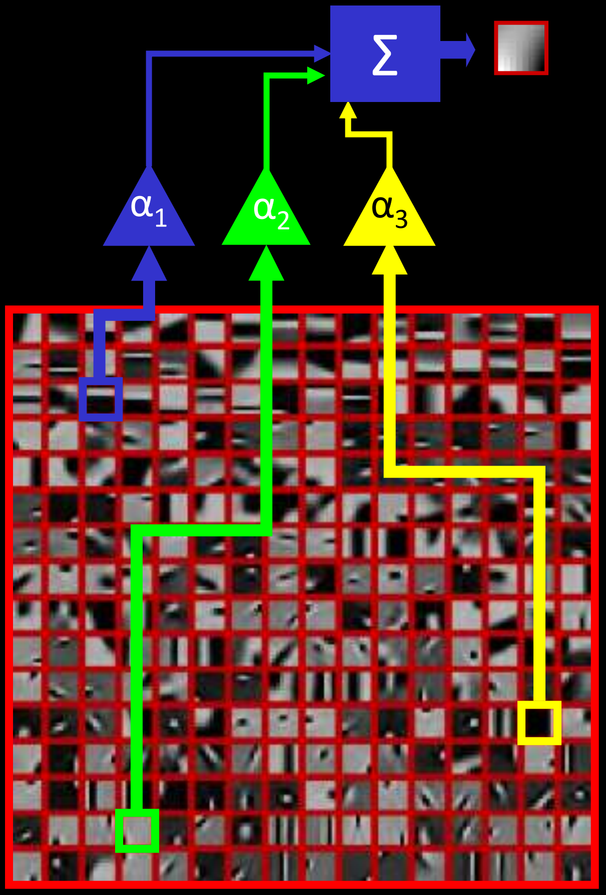


The *Sparseland* Model

Properties of this model:

Sparsity and Redundancy

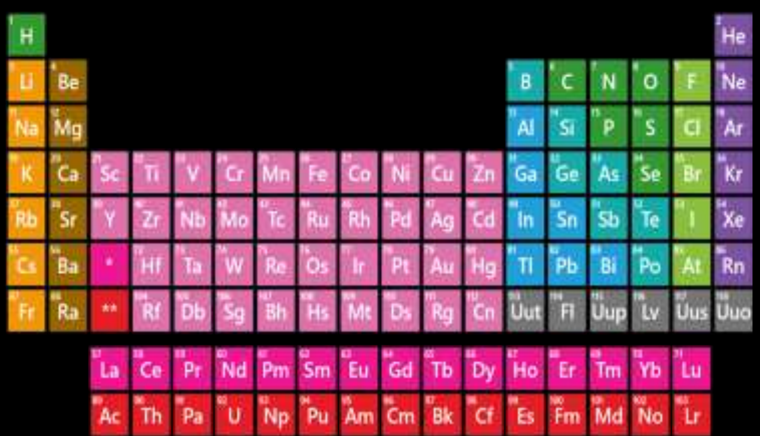
- We start with a 8-by-8 pixels patch and represent it using 256 numbers
  - – This is a redundant representation
- However, out of those 256 elements in the representation, only 3 are non-zeros
  - – This is a sparse representation
- Bottom line in this case: 64 numbers representing the patch are replaced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

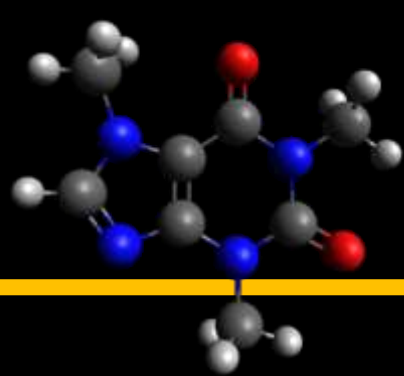


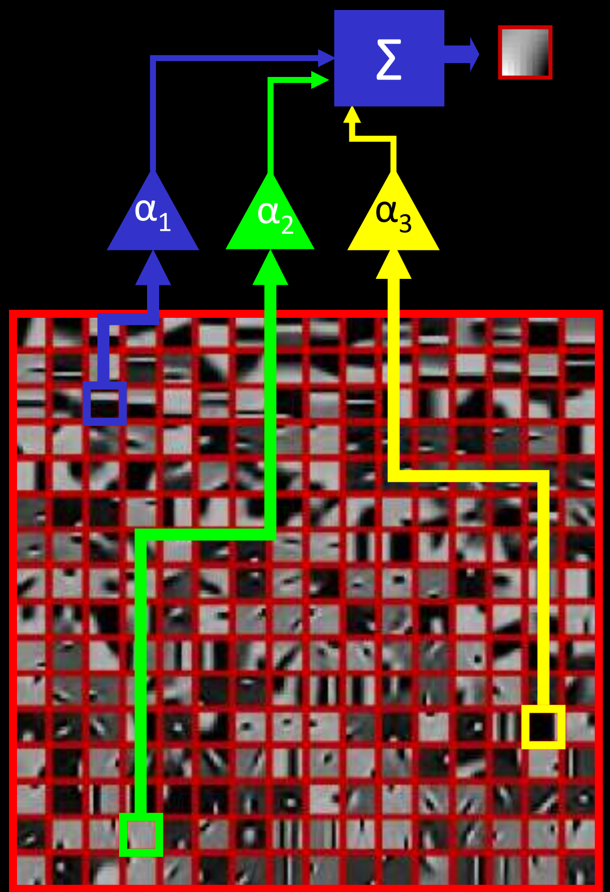
Chemistry of Data

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

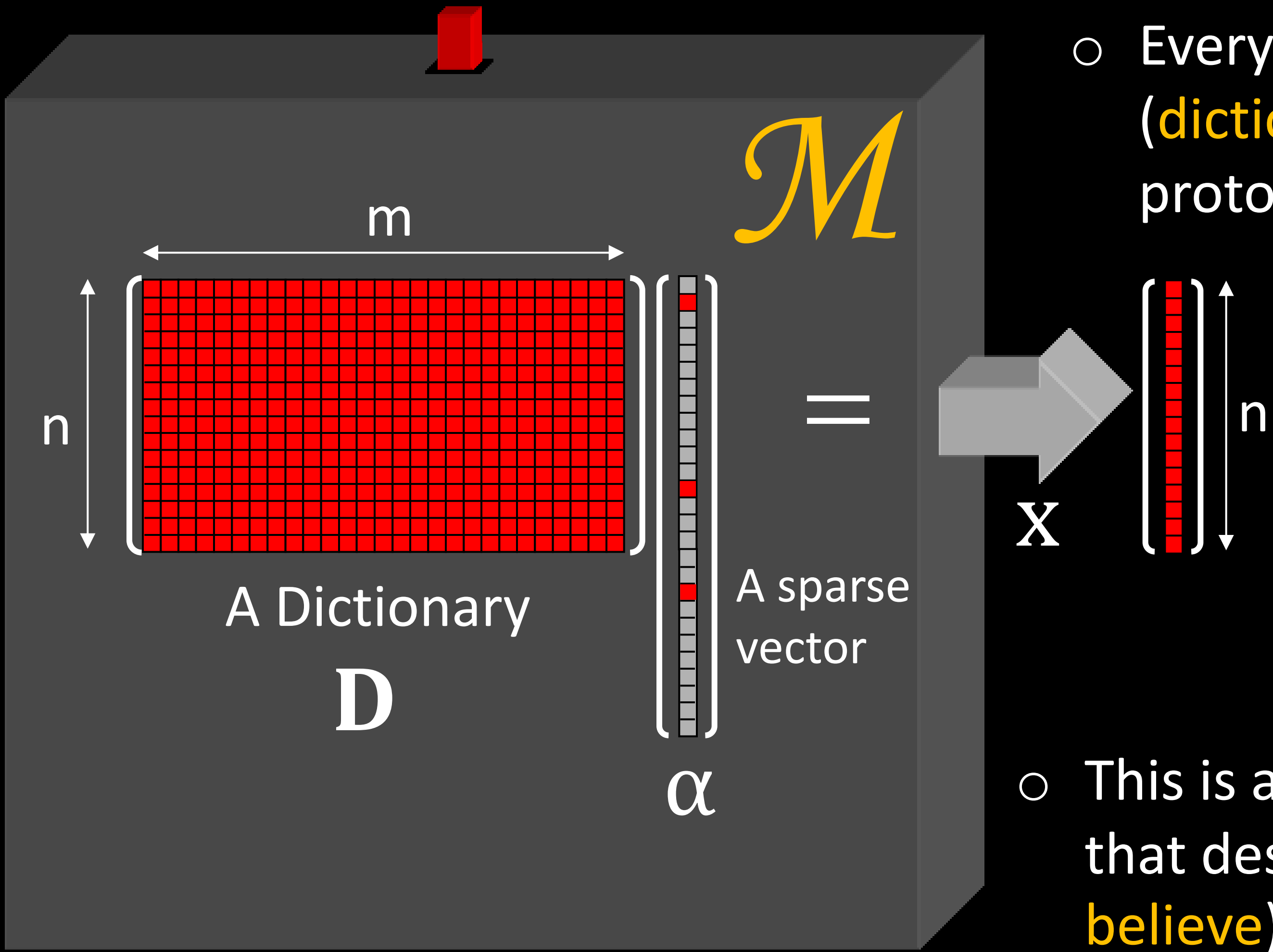






Michael Elad
The Computer-Science Department
The Technion

Sparseland : A Formal Description



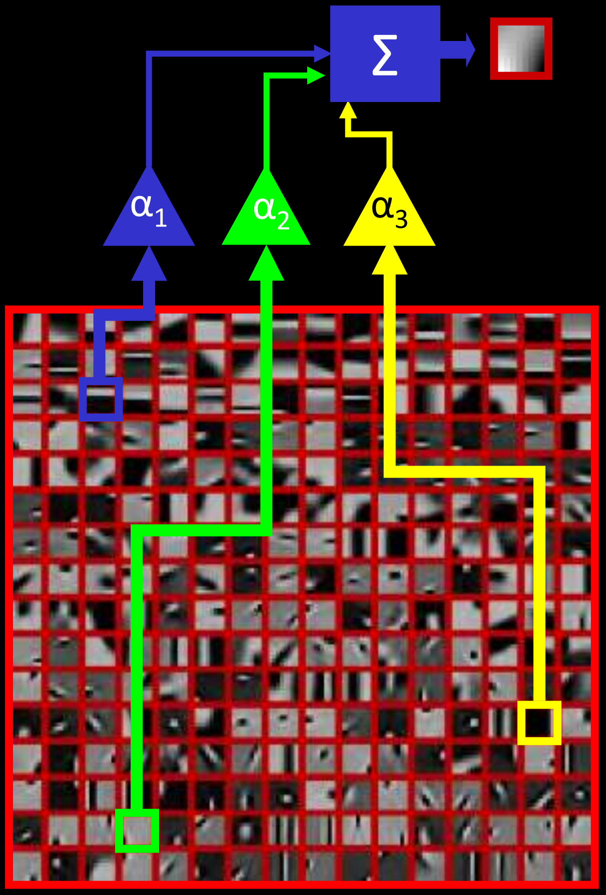
- Every column in **D** (dictionary) is a prototype signal (atom)
  - The vector $\underline{\alpha}$ is generated with few non-zeros at arbitrary locations and values
- This is a generative model that describes how (we believe) signals are created

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

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

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

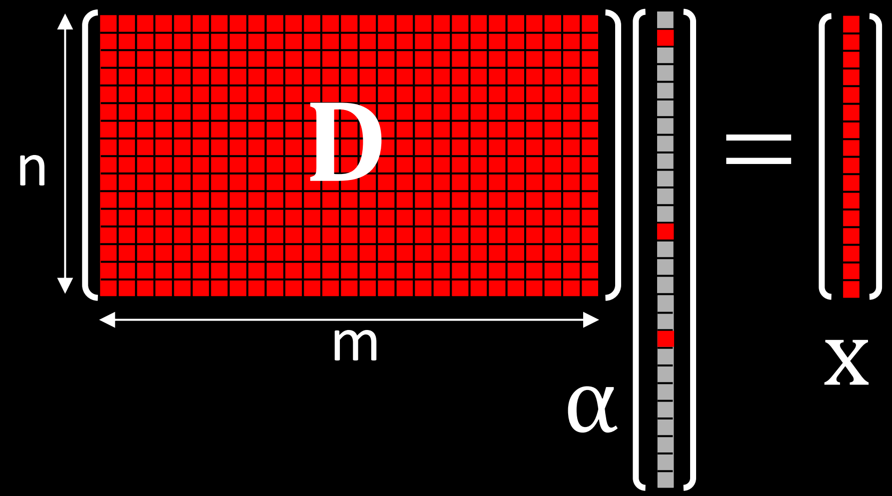


Michael Elad
The Computer-Science Department
The Technion

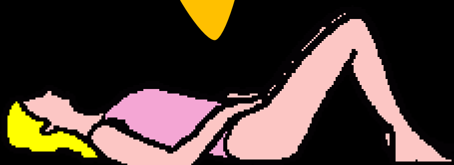
Atom Decomposition Made Formal

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

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



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

Michael Elad
The Computer-Science Department
The Technion

Pursuit Algorithms

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

Approximation Algorithms

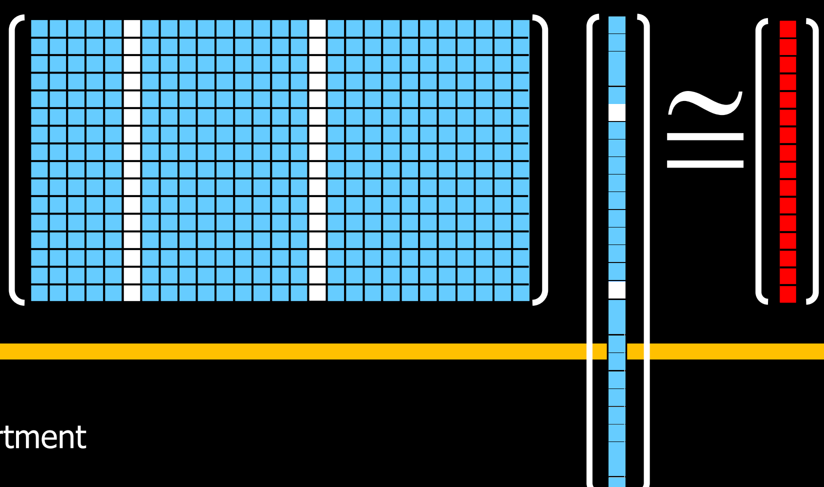
Basis Pursuit

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

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

Matching Pursuit

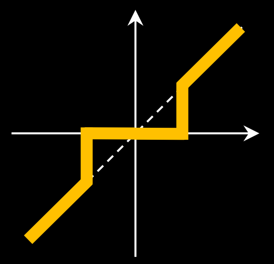
Find the support greedily, one element at a time



Thresholding

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

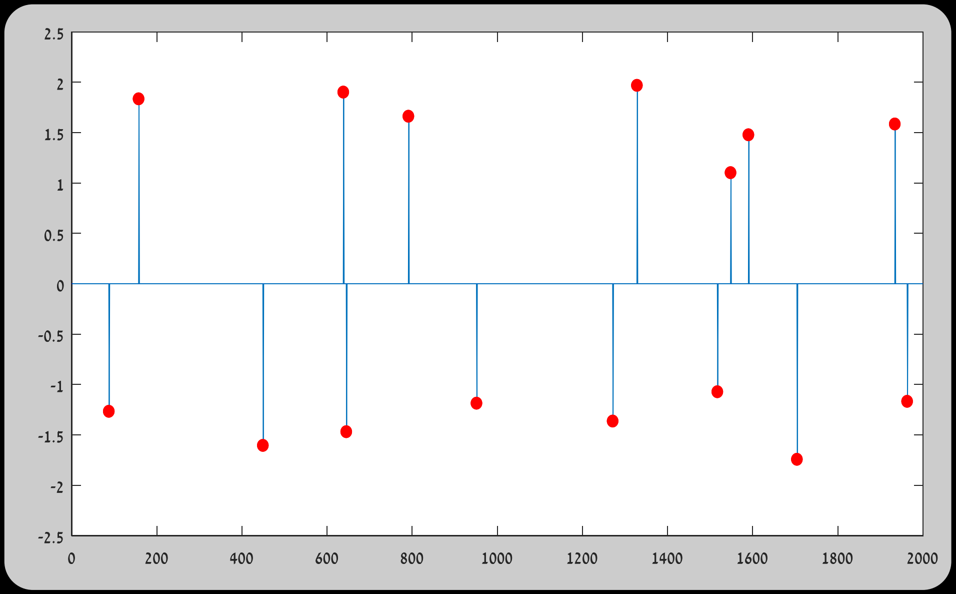
$$\hat{\alpha} = \mathcal{P}_{\beta}\{\mathbf{D}^{\mathbf{T}}\mathrm{y}\}$$

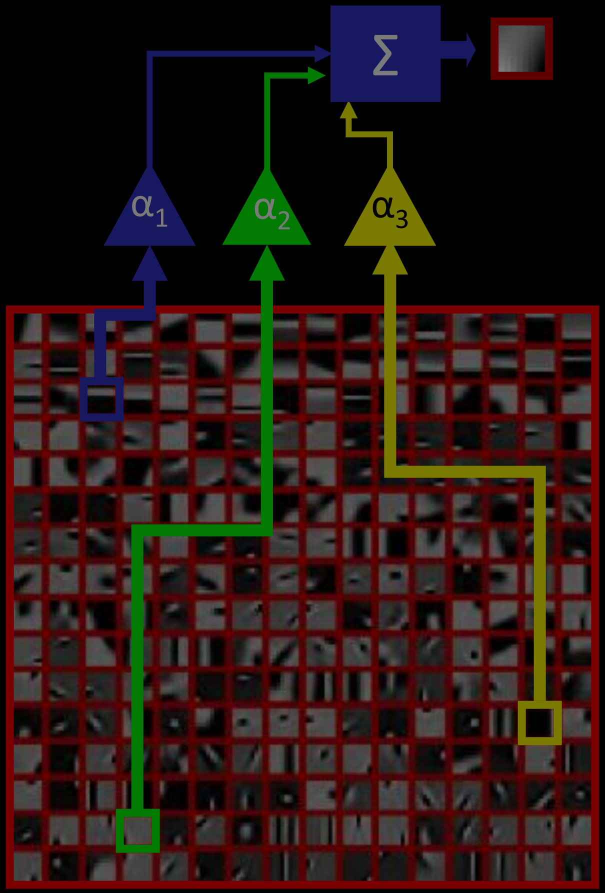


Michael Elad
The Computer-Science Department
The Technion

Difficulties with *Sparseland*

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



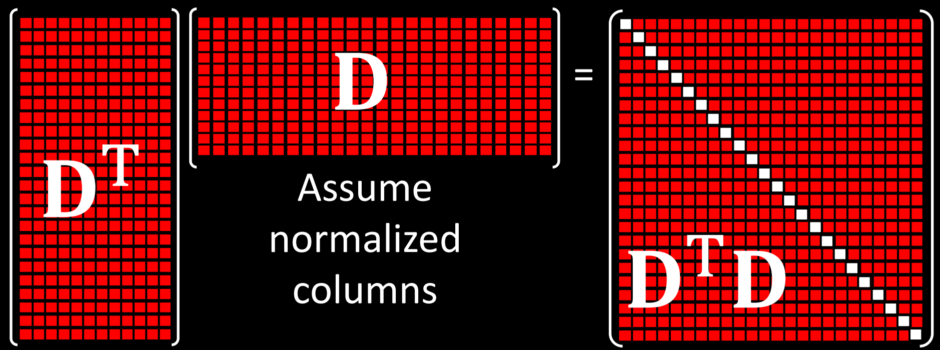


- Surprising fact: Many of these algorithms are often accompanied by theoretical guarantees for their success, if the unknown is sparse enough

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, …)

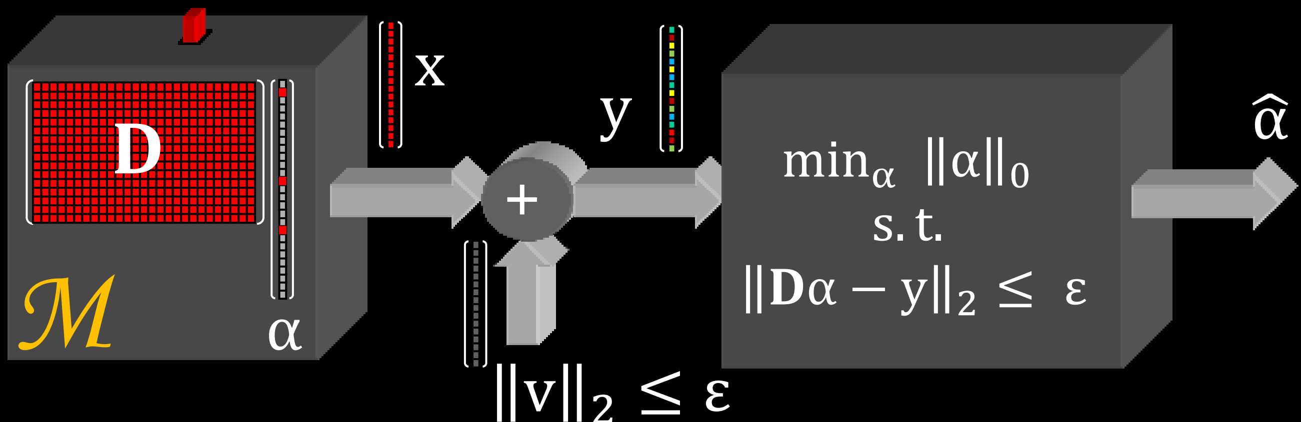
Basis-Pursuit Success

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: $\|\hat{\alpha} - \alpha\|_2^2 \le \frac{4\varepsilon^2}{1-\mu(4\|\alpha\|_0-1)}$

Donoho, Elad & Temlyakov ('06)

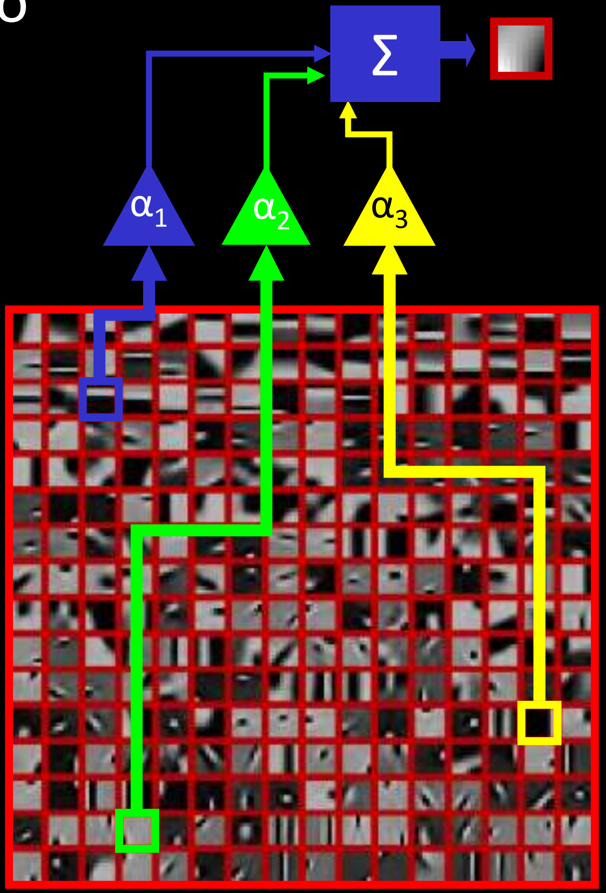


Comments:

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

Difficulties with *Sparseland*

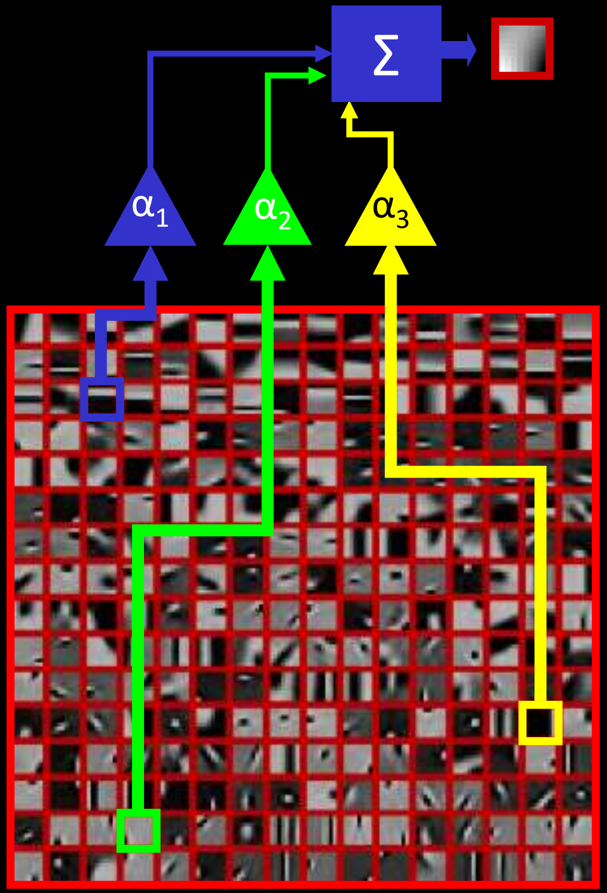
- Problem 2: Given a family of signals, how do we find the dictionary to represent it well?
- Solution: Learn! Gather a large set of signals (many thousands), and find the dictionary that sparsifies them
- Such algorithms were developed in the past 10 years (e.g., K-SVD), and their performance is surprisingly 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

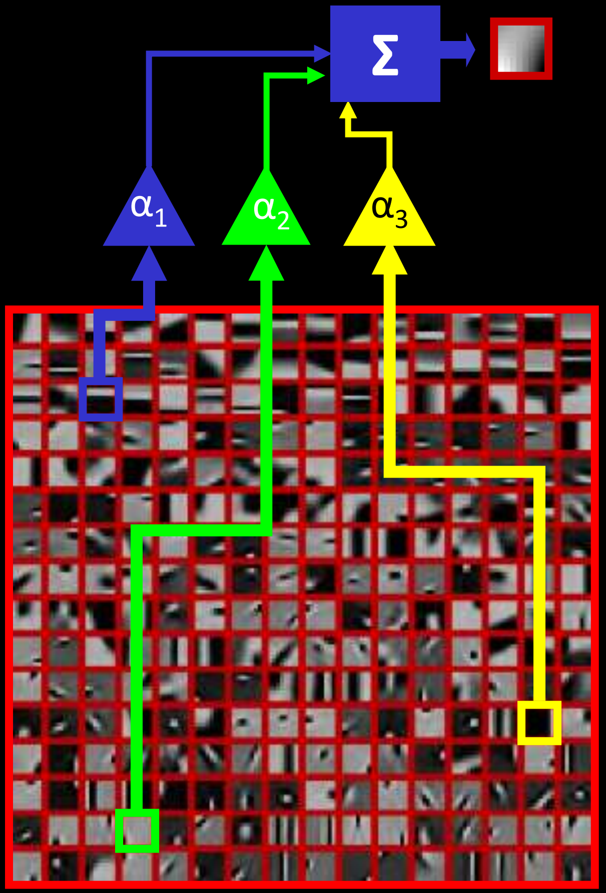


Michael Elad
The Computer-Science Department
The Technion

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? …

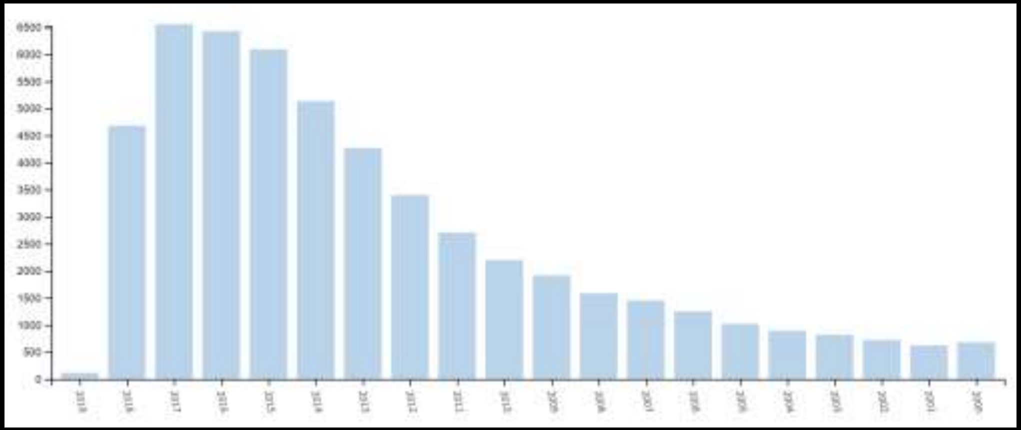
ALL ANSWERED POSITIVELY AND CONSTRUCTIVELY

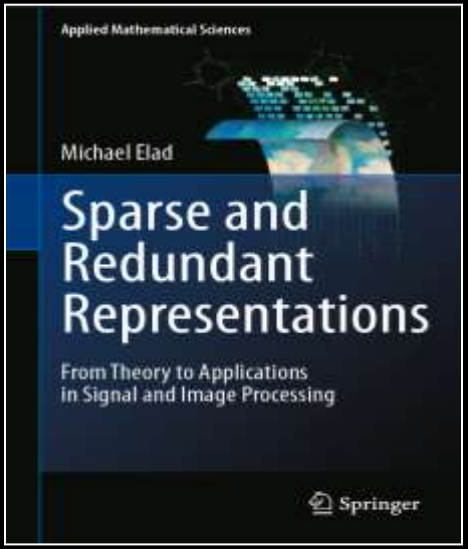


Michael Elad
The Computer-Science Department
The Technion

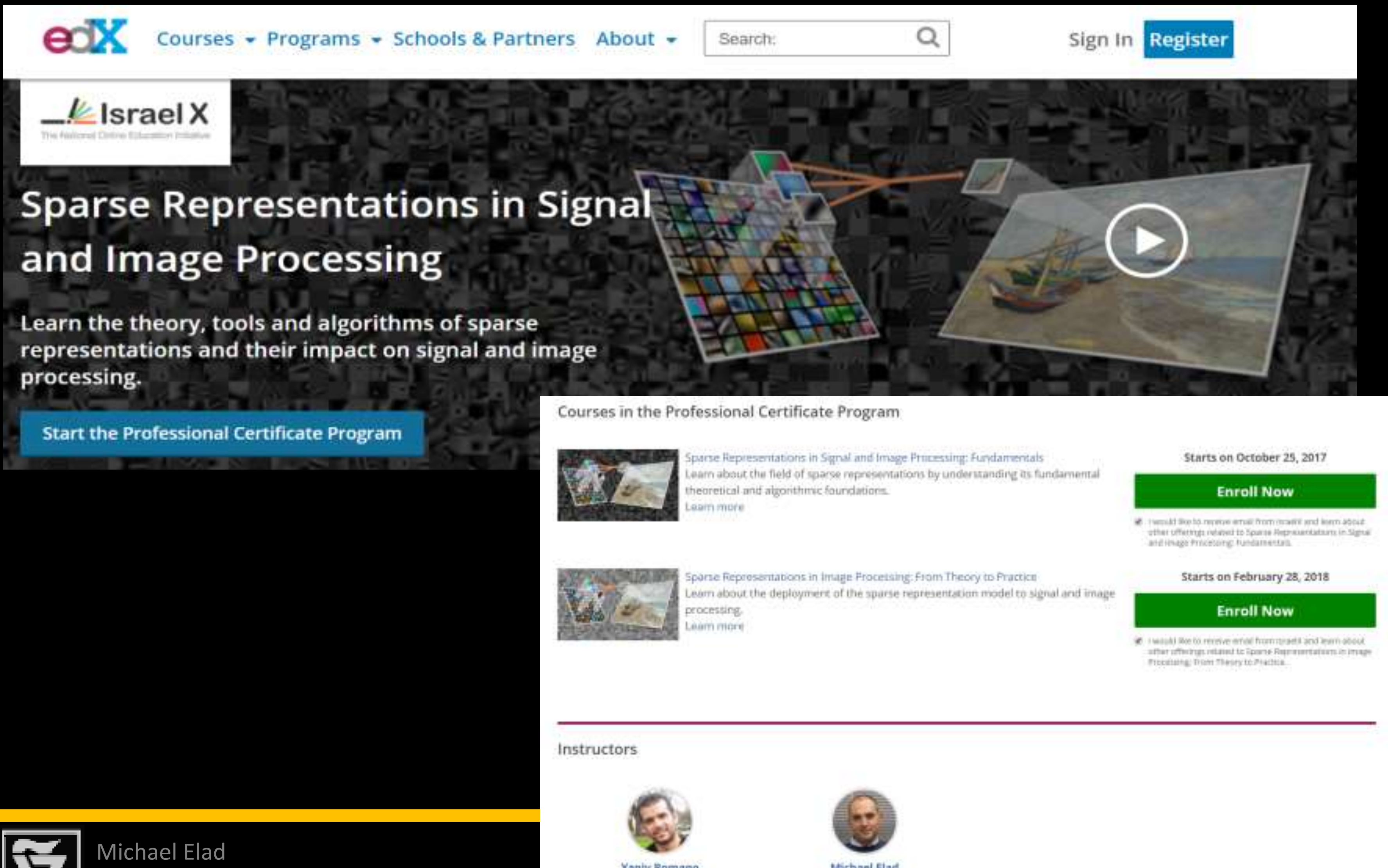
This Field has been rapidly GROWING …

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



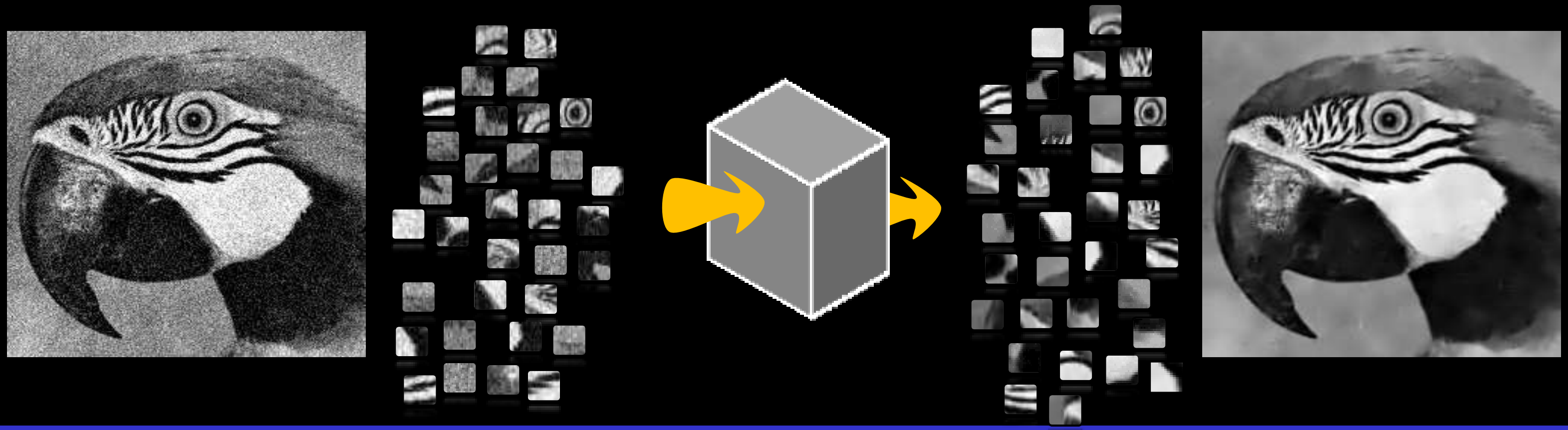


A New Massive Open Online Course



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)

Multi-Layered Convolutional Sparse Modeling

Joint work with


Yaniv Romano

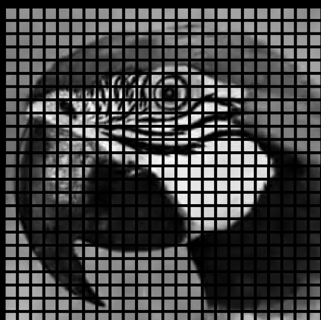

Vardan Papyan


Jeremias Sulam

Convolutional Sparse Coding (CSC)

m filters convolved with their sparse representations

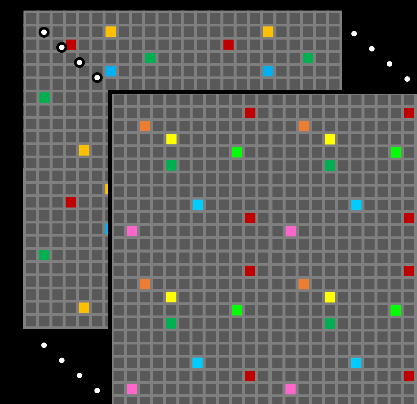
i-th feature-map: An image of the same size as $\mathbf{X}$ holding the sparse representation related to the i-filter



An image with N pixels

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

The i-th filter of small size n



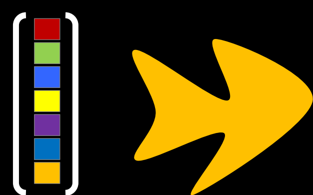
This model emerged in 2005-2010, developed and advocated by Yan LeCun and others. It serves as the foundation of Convolutional Neural Networks

CSC in Matrix Form

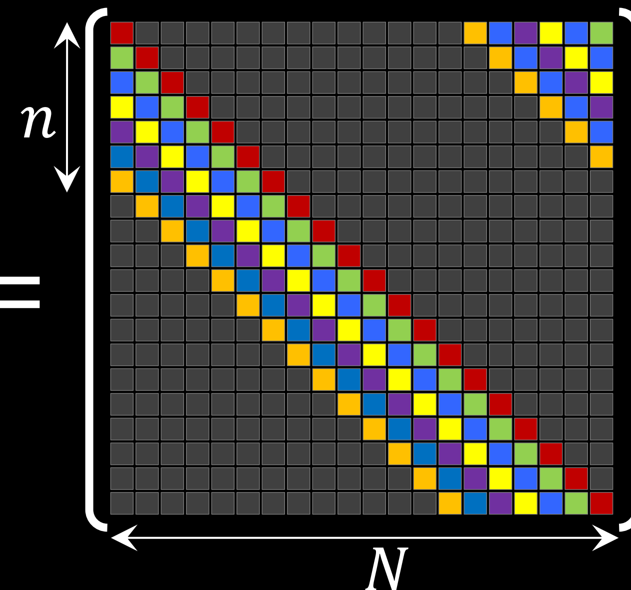
- Here is an alternative global sparsity-based model formulation

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

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



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

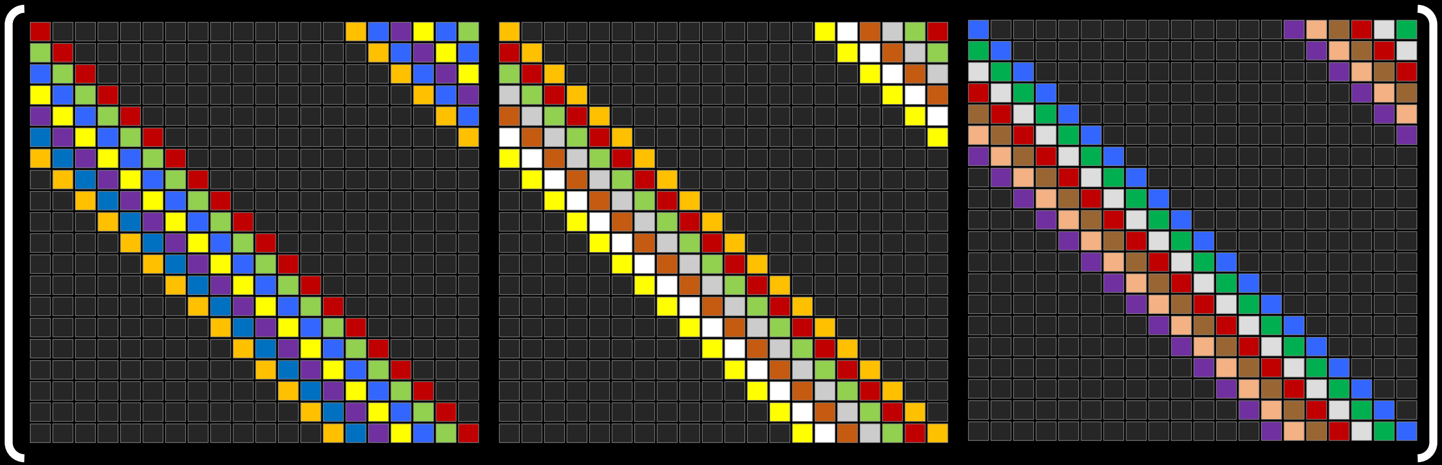


- $\boldsymbol{\Gamma}^i \in \mathbb{R}^{N}$ are the corresponding coefficients ordered as column vectors

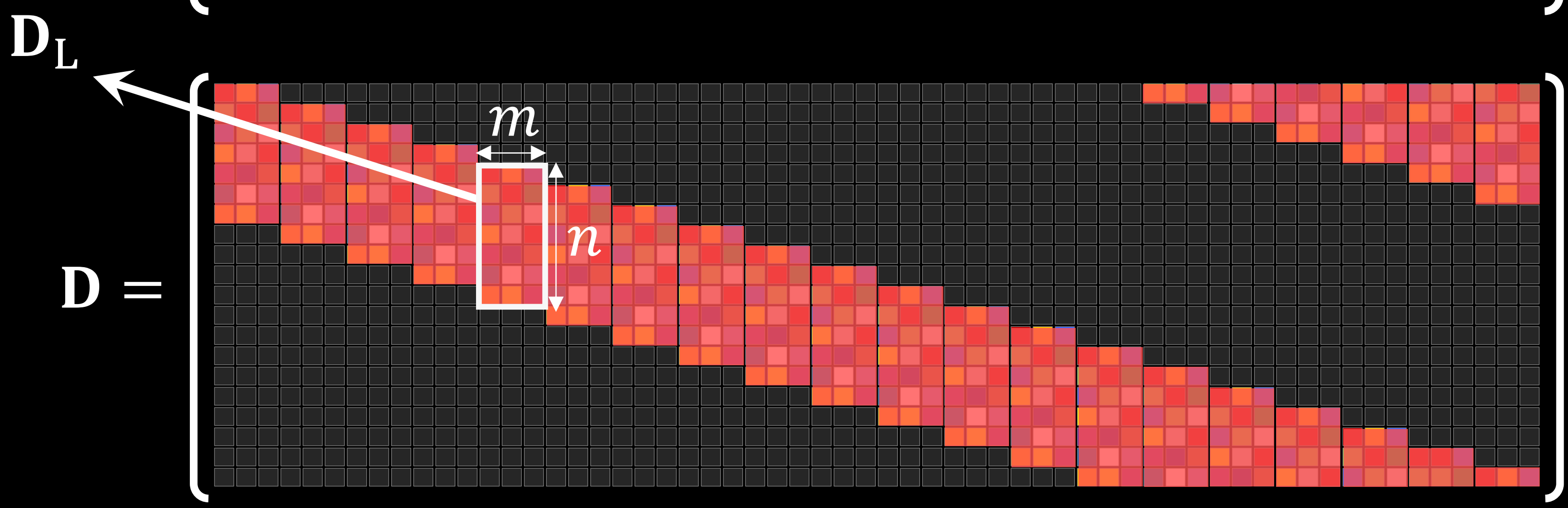
Michael Elad
The Computer-Science Department
The Technion

The CSC Dictionary

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



$\mathbf{D} =$



Classical Sparse Theory for CSC ?

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

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

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

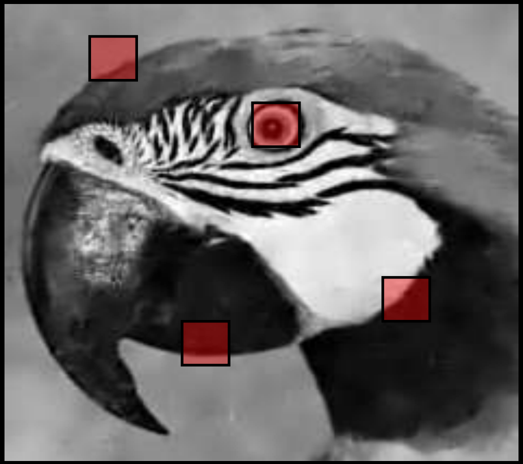
$$\mu \geq 0.063$$

- Success of pursuits i…

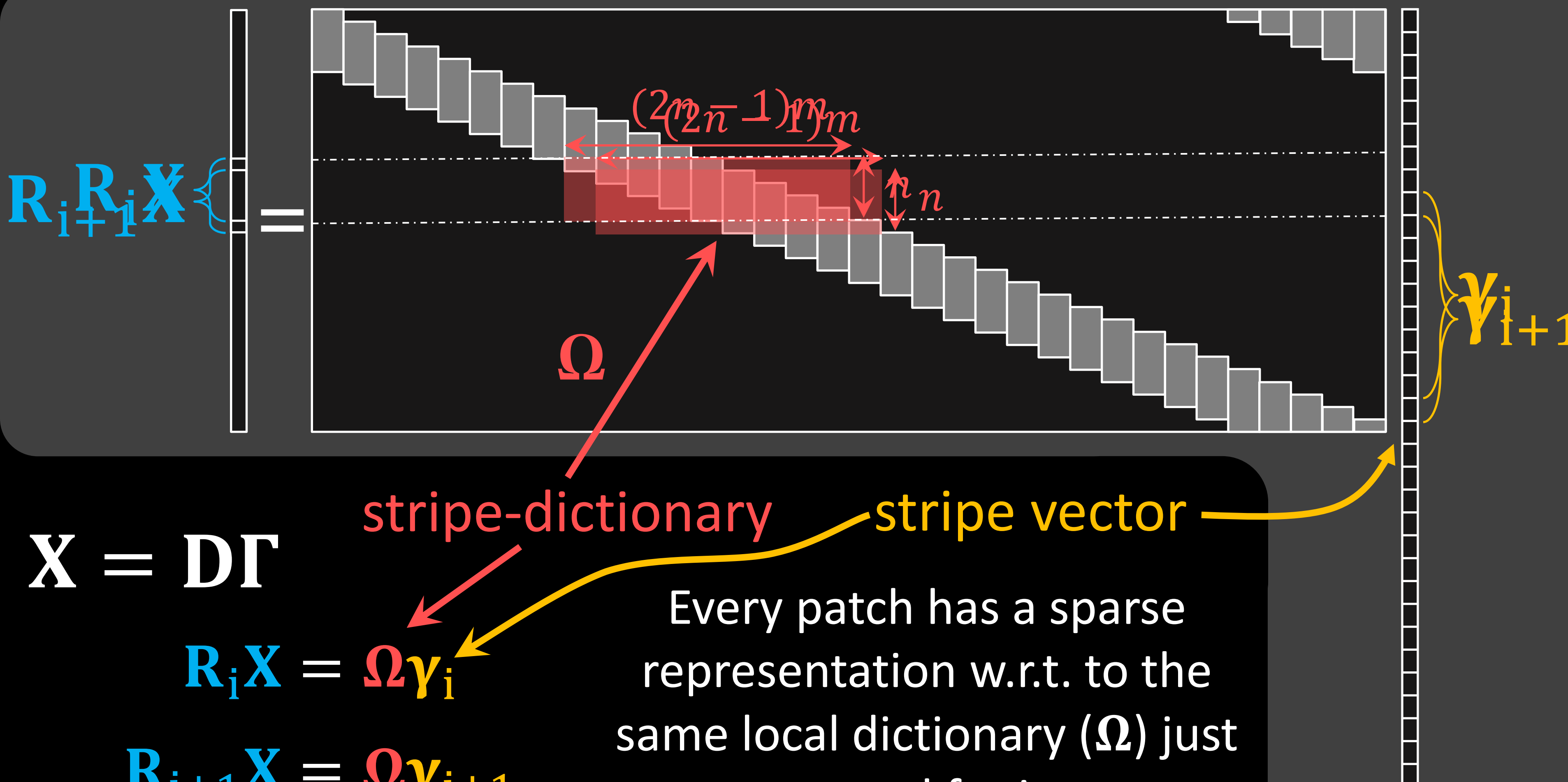
2

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

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



Why CSC?



stripe-dictionary

stripe vector

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

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

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

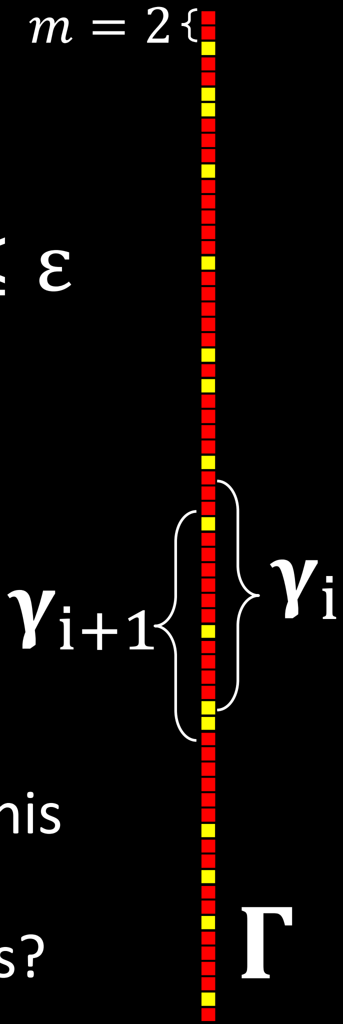
Every patch has a sparse representation w.r.t. to the same local dictionary ($\mathbf{\Omega}$) just as assumed for images

Moving to Local Sparsity: Stripes

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

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

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



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

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

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

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

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

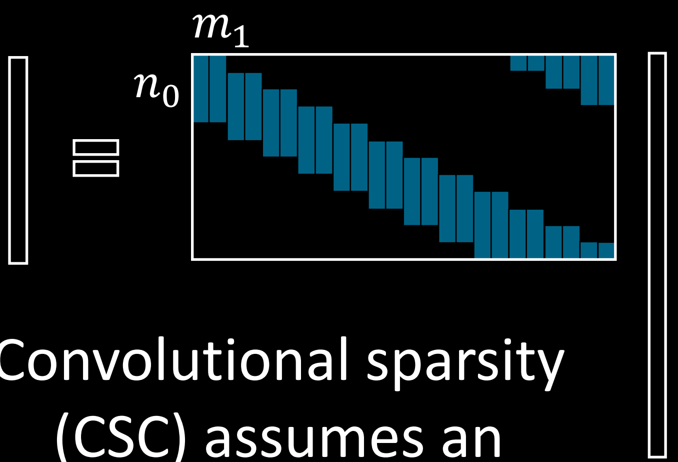
This is a much better result – it allows few non-zeros locally in each stripe, implying a permitted $O(N)$ non-zeros globally

Papyan, Sulam & Elad ('17)

Multi-Layered Convolutional Sparse Modeling

From CSC to Multi-Layered CSC

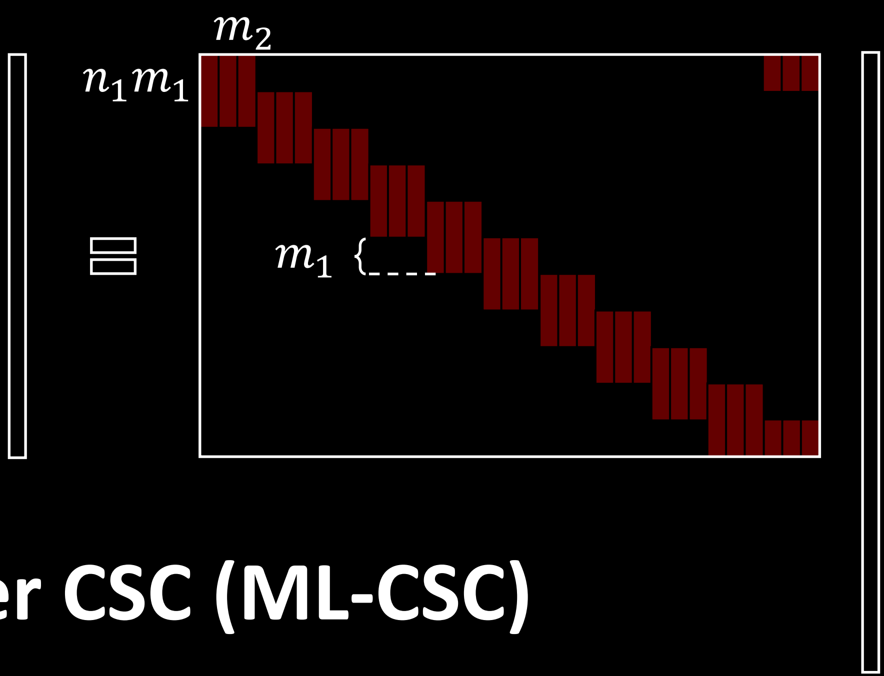
$\mathbf{X} \in \mathbb{R}^N$ $\mathbf{D}_1 \in \mathbb{R}^{N \times N m_1}$ $\mathbf{\Gamma}_1 \in \mathbb{R}^{N m_1}$



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

We propose to impose the same structure on the representations themselves

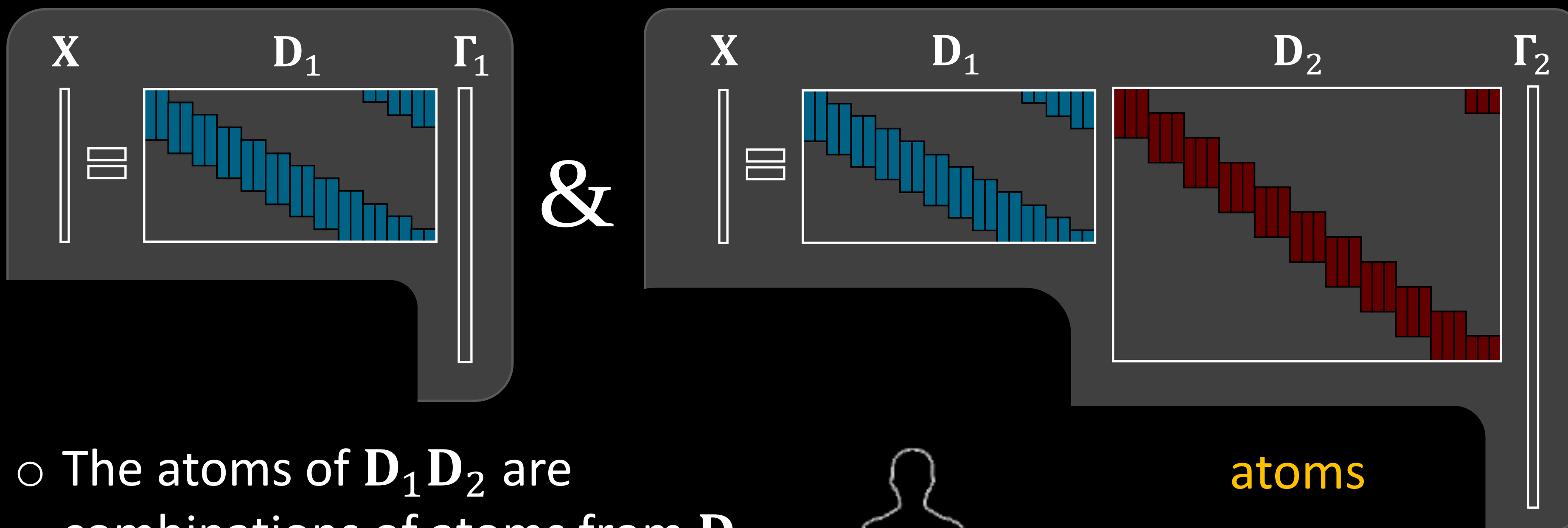
$\mathbf{\Gamma}_1 \in \mathbb{R}^{N m_1}$ $\mathbf{D}_2 \in \mathbb{R}^{N m_1 \times N m_2}$ $\mathbf{\Gamma}_2 \in \mathbb{R}^{N m_2}$



Multi-Layer CSC (ML-CSC)

Michael Elad
The Computer-Science Department
The Technion

Intuition: From Atoms to Molecules

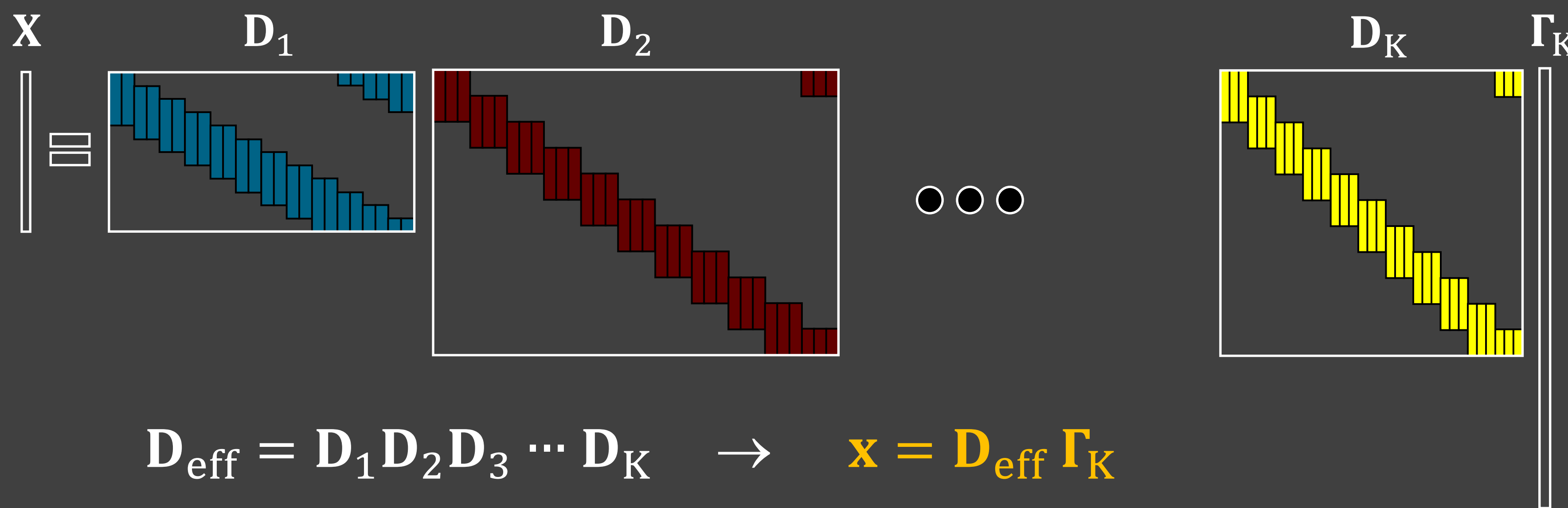


- The atoms of $\mathbf{D}_1\mathbf{D}_2$ are combinations of atoms from $\mathbf{D}_1$ - these are now molecules
- Thus, this model offers different levels of abstraction in describing X

= ∑ atoms, molecules, cells, tissue, body-parts …

Michael Elad
The Computer-Science Department
The Technion

Intuition: From Atoms to Molecules



$$\mathbf{D}_{\text{eff}} = \mathbf{D}_1 \mathbf{D}_2 \mathbf{D}_3 \cdots \mathbf{D}_K \quad \rightarrow \quad \mathbf{x} = \mathbf{D}_{\text{eff}} \, \mathbf{\Gamma}_K$$

- This is a special *Sparseland* (indeed, a CSC) model
- However: A key property in our model: the intermediate representations are required to be sparse as well

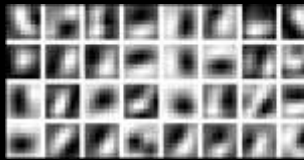
Michael Elad
The Computer-Science Department
The Technion

A Small Taste: Model Training (MNIST)

MNIST Dictionary:

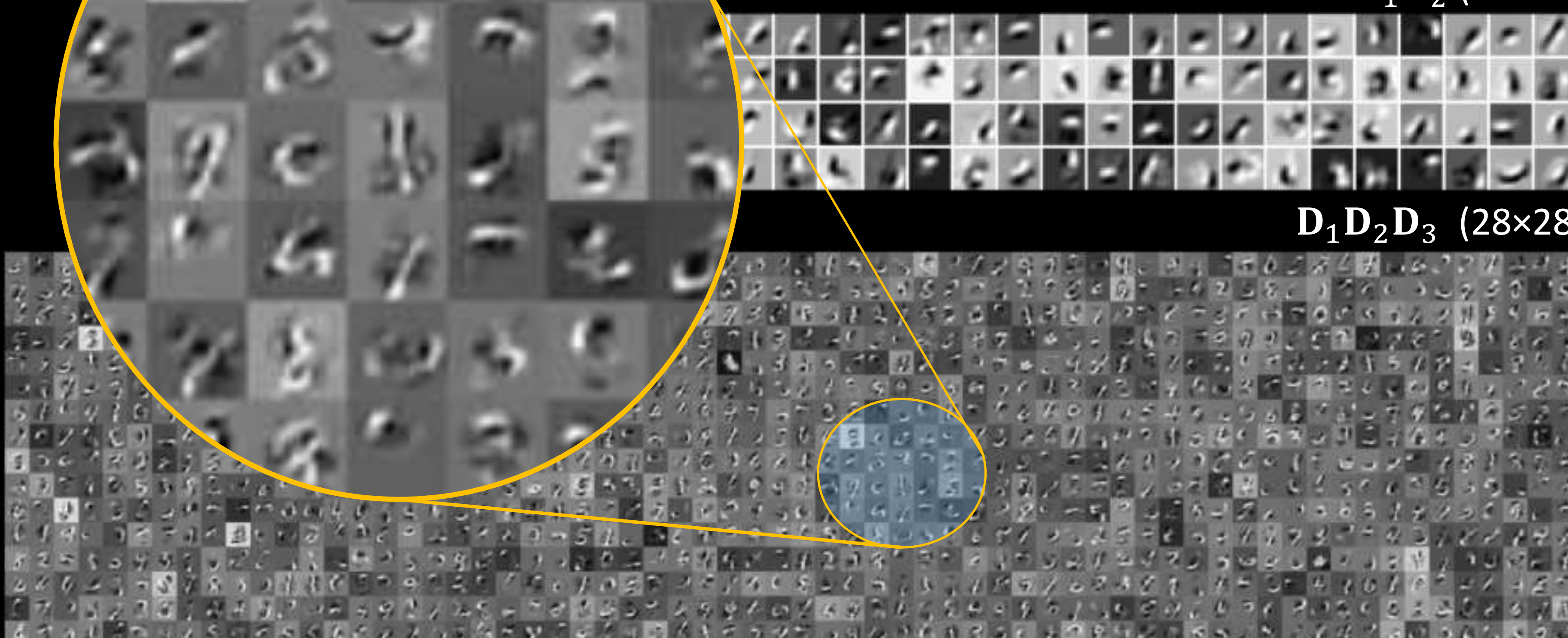
- $\mathbf{D}_1$: 32 filters o … f 2 (dense)
- $\mathbf{D}_2$: 128 fil … f 1 - 99.09 % sparse
- $\mathbf{D}_3$: 102 … arse

$\mathbf{D}_1$ (7×7)



$\mathbf{D}_1\mathbf{D}_2$ (15×15)

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



Michael Elad
The Computer-Science Department
The Technion

ML-CSC: Pursuit

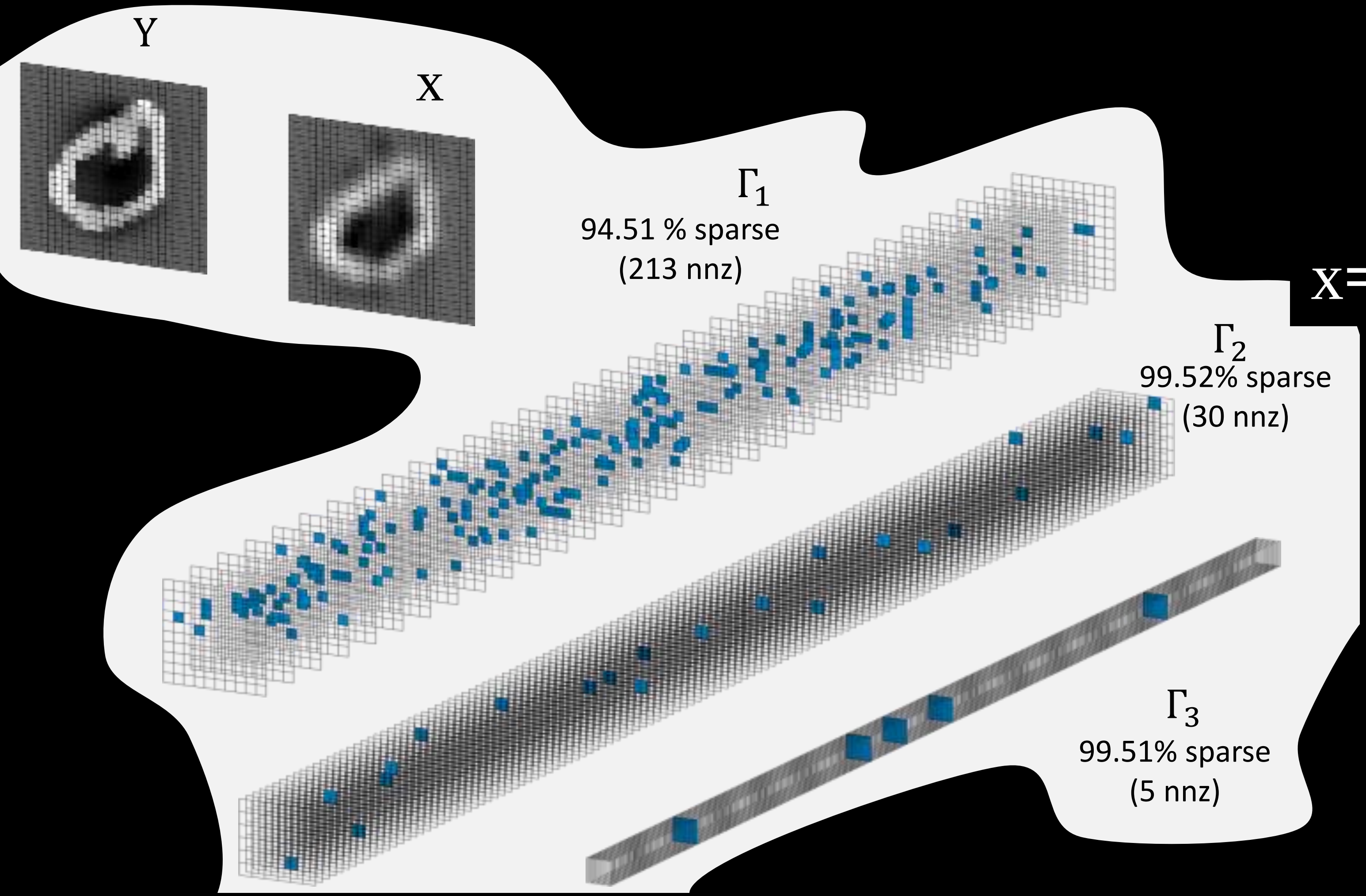
- Deep–Coding Problem $(\mathbf{DCP}_\lambda)$ (dictionaries are known):

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

- Or, more realistically for noisy signals,

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

A Small Taste: Pursuit



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

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

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

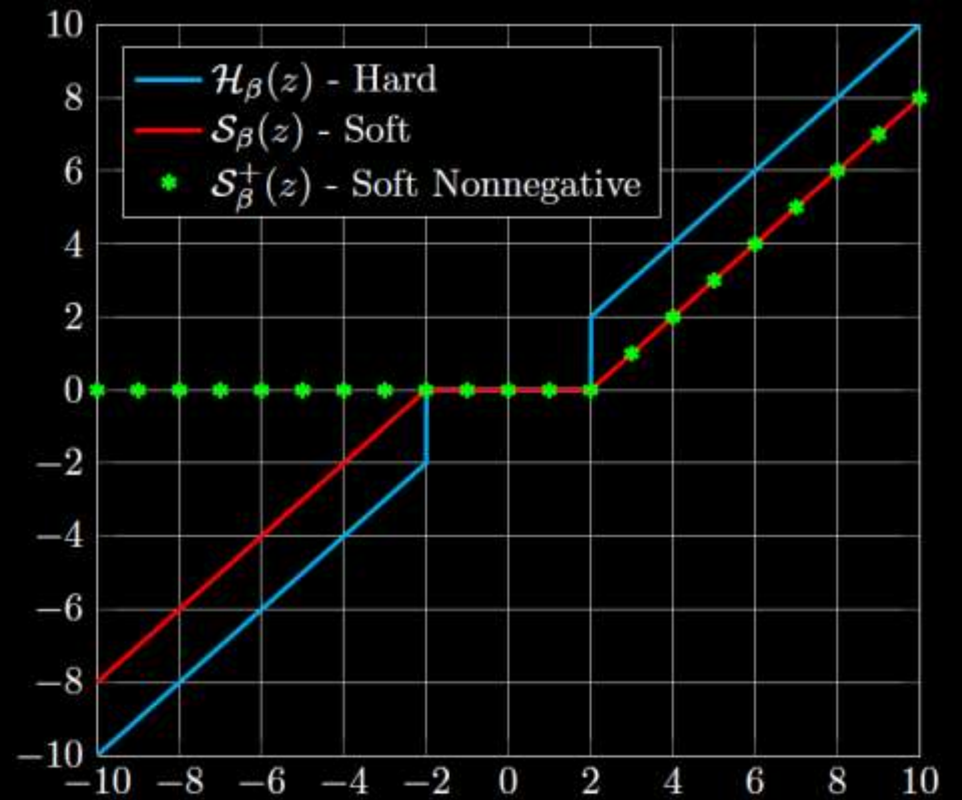
ML-CSC: The Simplest Pursuit

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

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

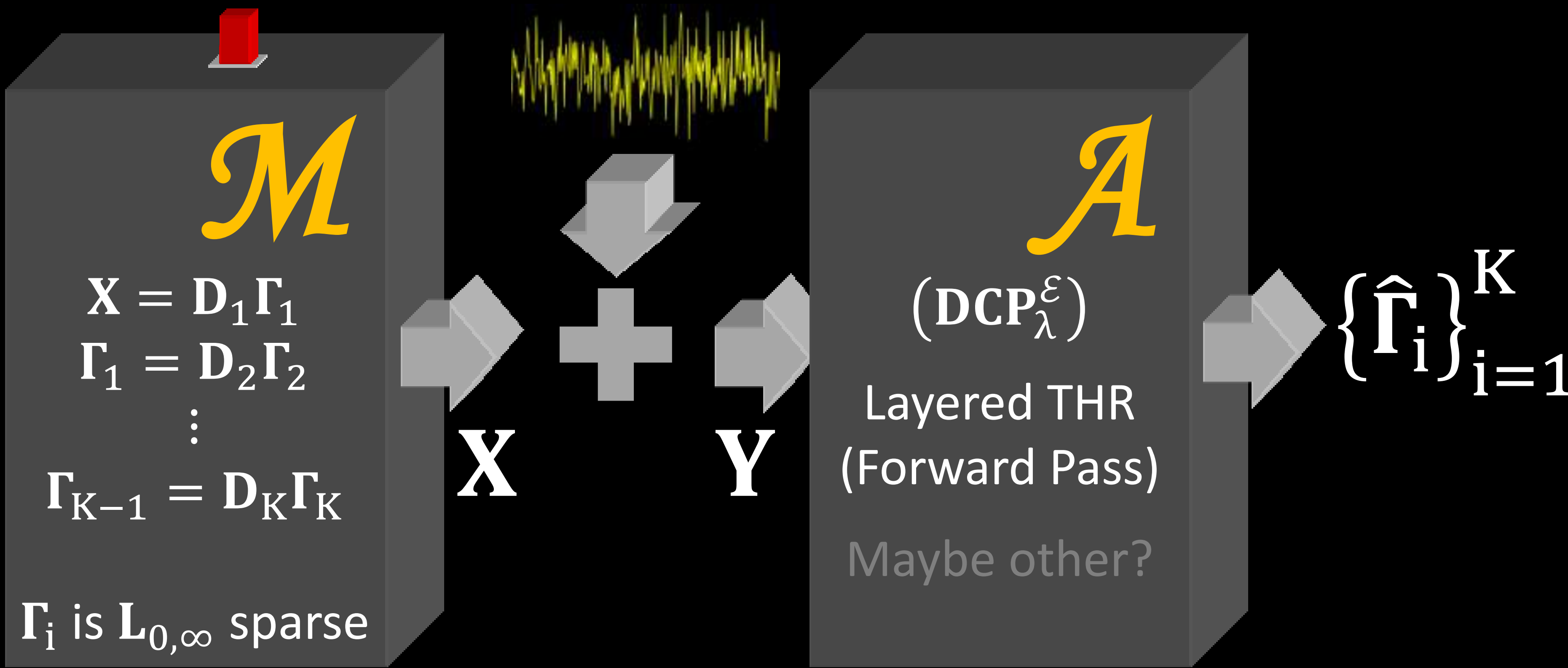
$$\left\{\begin{array}{ll} \|\mathbf{Y} - \mathbf{D}_1\mathbf{\Gamma}_1\|_2 \le \mathcal{E} & \|\mathbf{\Gamma}_1\|_{0,\infty}^{s} \le \lambda_1 \\ \mathbf{\Gamma}_1 = \mathbf{D}_2\mathbf{\Gamma}_2 & \|\mathbf{\Gamma}_2\|_{0,\infty}^{s} \le \lambda_2 \\ \vdots & \vdots \\ \mathbf{\Gamma}_{K-1} = \mathbf{D}_K\mathbf{\Gamma}_K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

- 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

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 $\left\|\boldsymbol{\Gamma}_i^{LT} - \boldsymbol{\Gamma}_i\right\|_{2,\infty}^{p} \leq \varepsilon_L^i$, where we have defined $\varepsilon_L^0 = \|\mathbf{E}\|_{2,\infty}^{p}$ and

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

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 …

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

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

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

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

⋮

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

Michael Elad
The Computer-Science Department
The Technion

Success of the Layered BP

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

then the Layered 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: $\left\|\boldsymbol{\Gamma}_\mathrm{i}^\mathrm{LBP} - \boldsymbol{\Gamma}_\mathrm{i}\right\|_{2,\infty}^\mathrm{p} \leq \varepsilon_\mathrm{L}^\mathrm{i}$, 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

Papyan, Romano & Elad ('17)

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

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 Algorithm (ISTA):

t

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

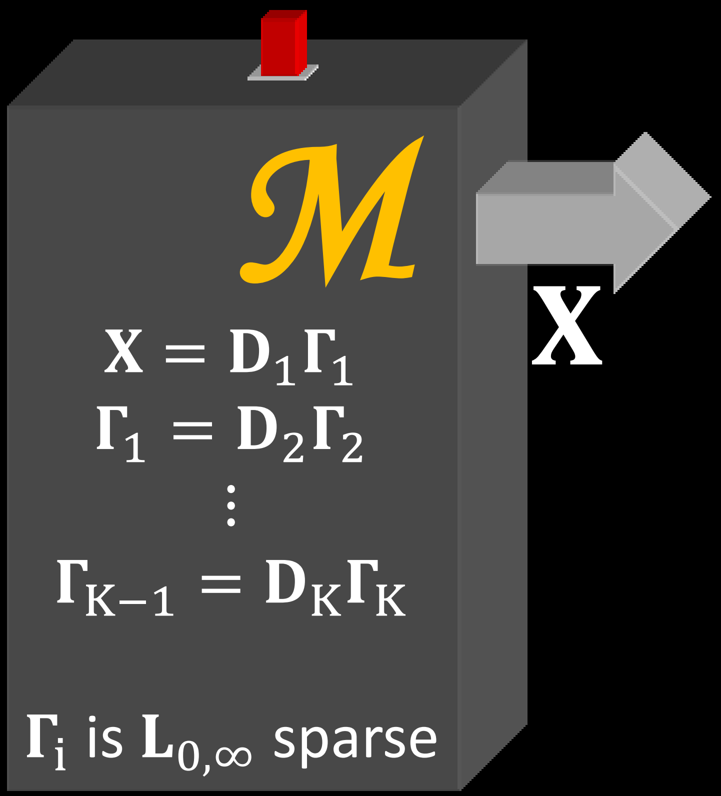
j

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]

Michael Elad
The Computer-Science Department
The Technion

Where are the Labels?

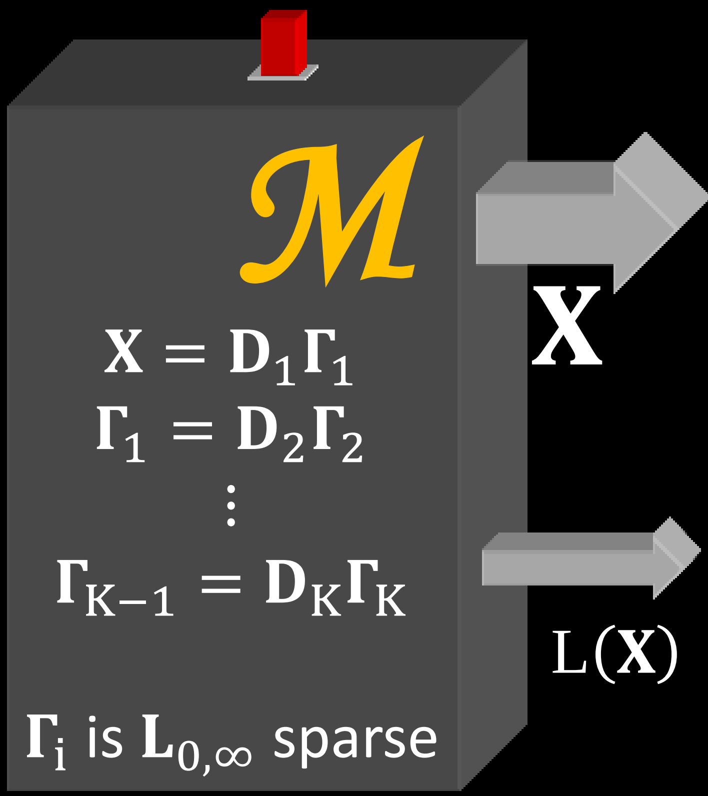


We presented the ML-CSC as a machine that produces signals $\mathbf{X}$

Answer 1:

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

Where are the Labels?



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

Answer 2:

- 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 suffice for identifying the label
- A successful pursuit algorithm can lead to an accurate recognition if the network is augmented by a FC classification layer
- See our recent paper (on ArXiv), analyzing bounds on adversarial noise permitted and the influence of the pursuit algorithm

Michael Elad
The Computer-Science Department
The Technion

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: We proposed two such algorithms – see ArXiv (handling both unsupervised and supervised learning)

Time to Conclude

This Talk

Sparseland

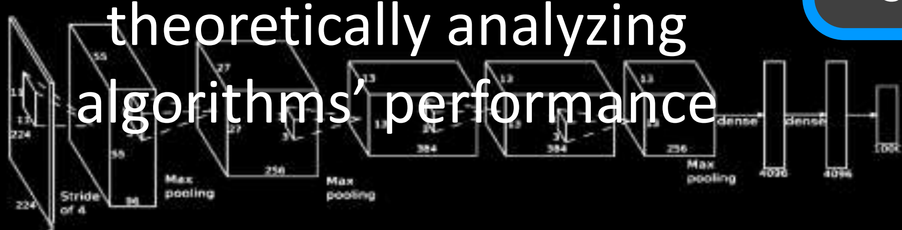
The desire to model data

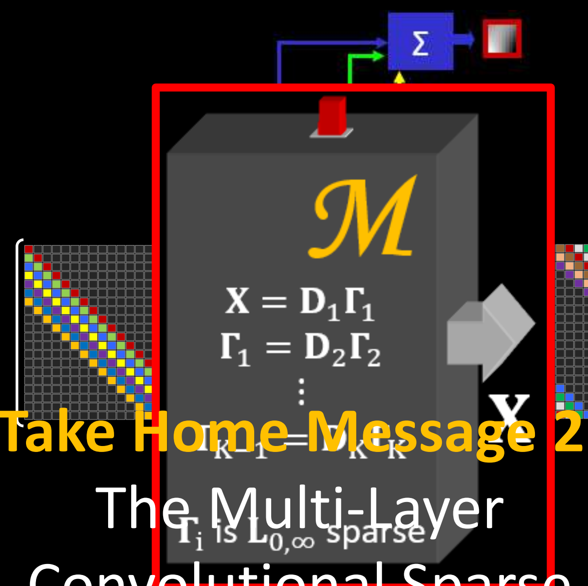
Novel View of Convolutional Sparse Coding

Multi-Layer Convolutional Sparse Coding

A novel interpretation and theoretical understanding of CNN

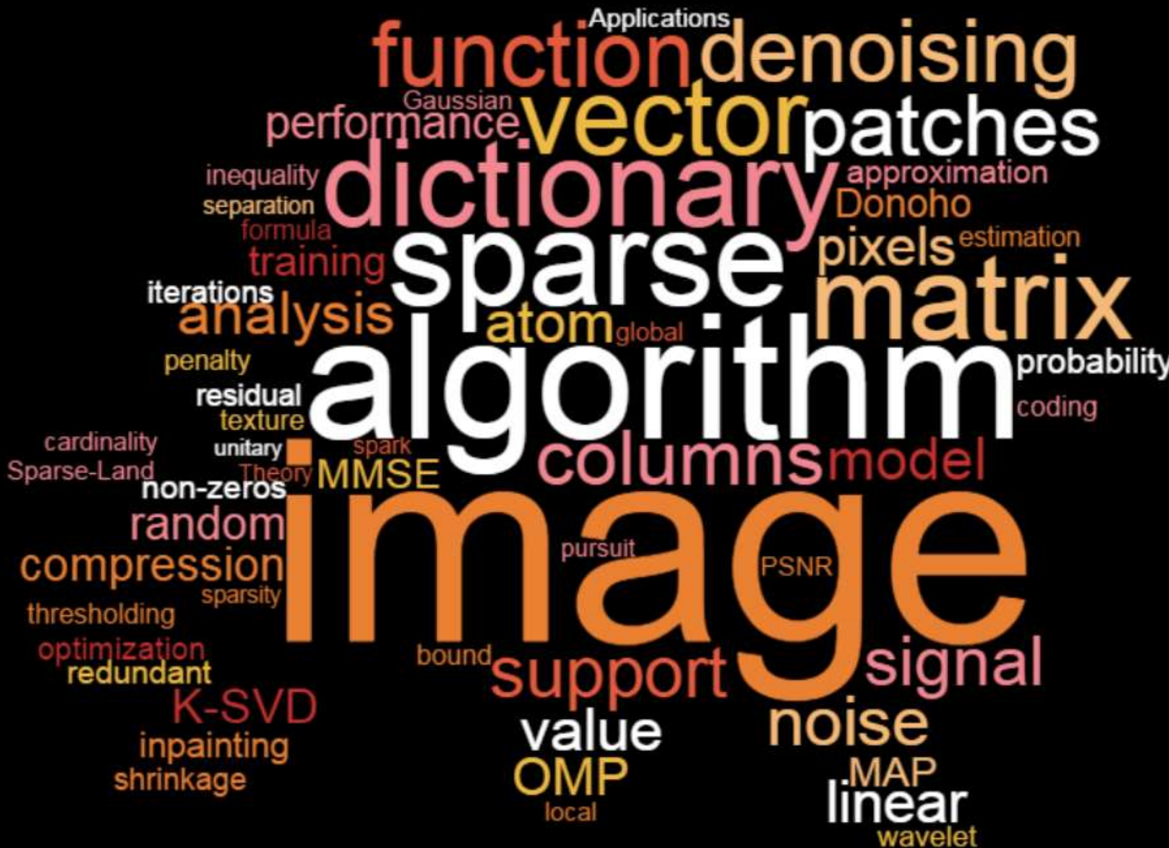
Take Home Message 1:
Generative modeling of data sources enables algorithm development along with theoretically analyzing algorithms' performance





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