

Sparse Modeling of Data

and its Relation to

Deep Learning

Michael Elad

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

27-28 June 2019

ETH zürich

Department of Mathematics

FIM - Institute for Mathematical Research

Series of Lectures on Waves and Imaging (III)

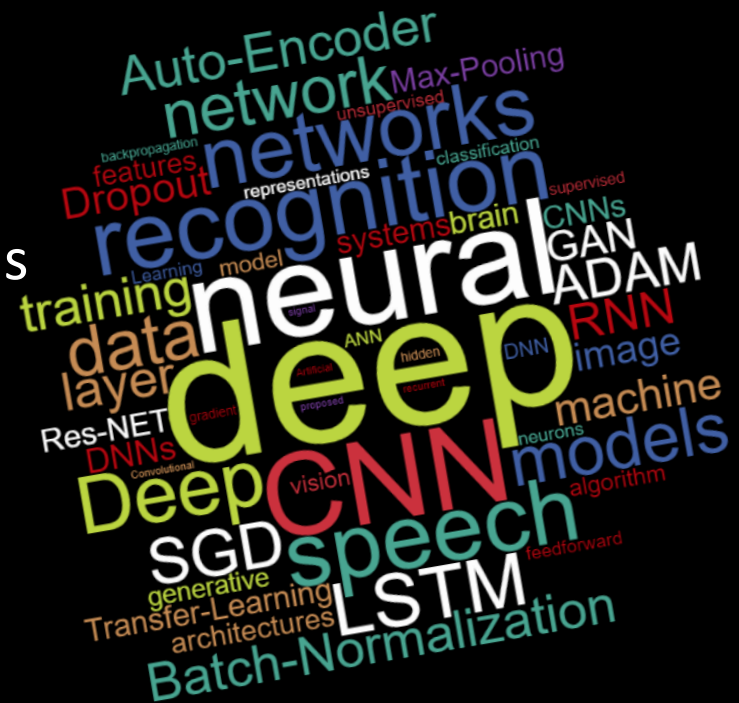


This Lecture is About ...

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



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



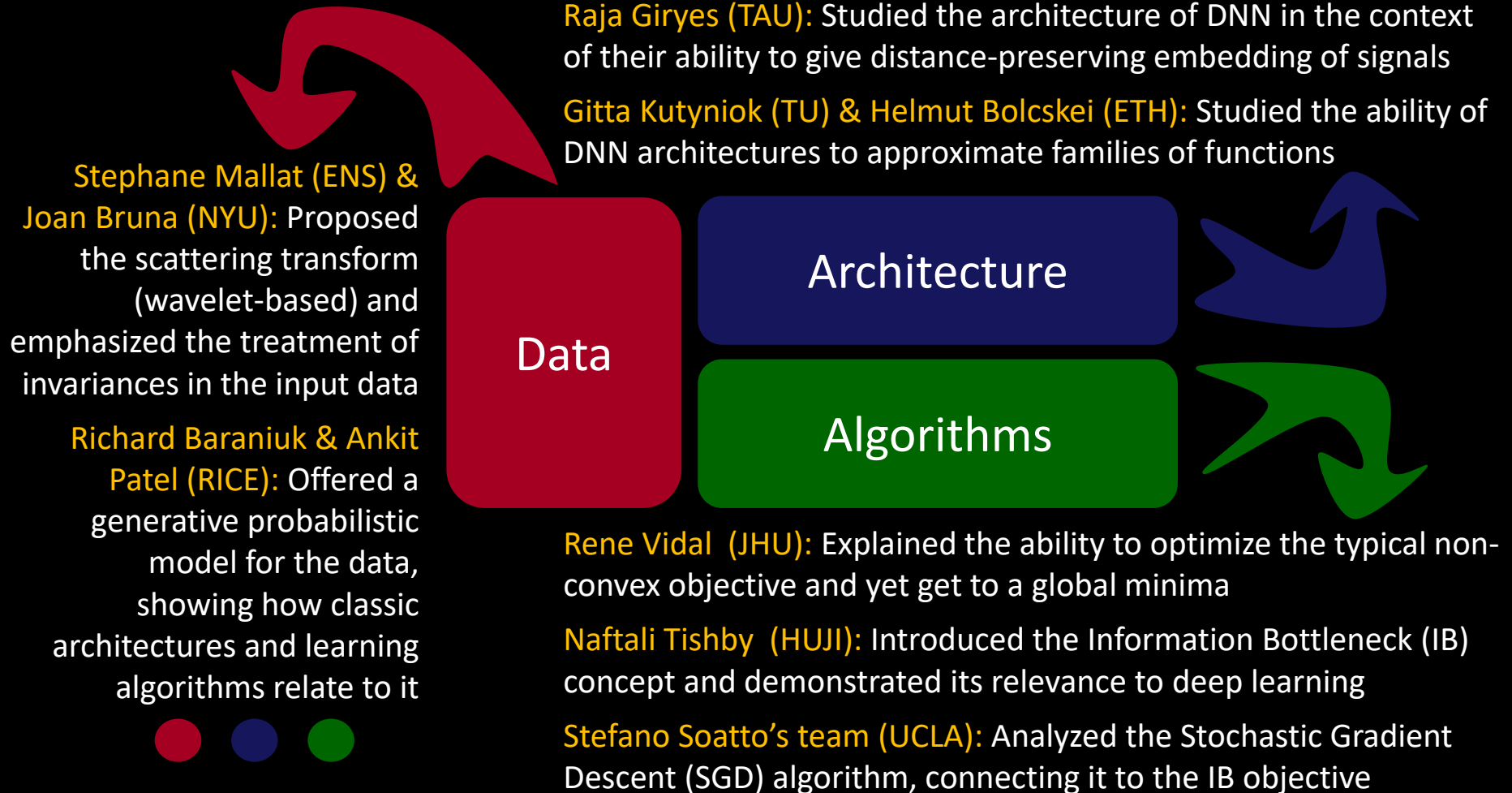
Understanding is a good thing ... but another goal is inventing methods. In the history of science and technology, engineering

preceded theoretical understanding:

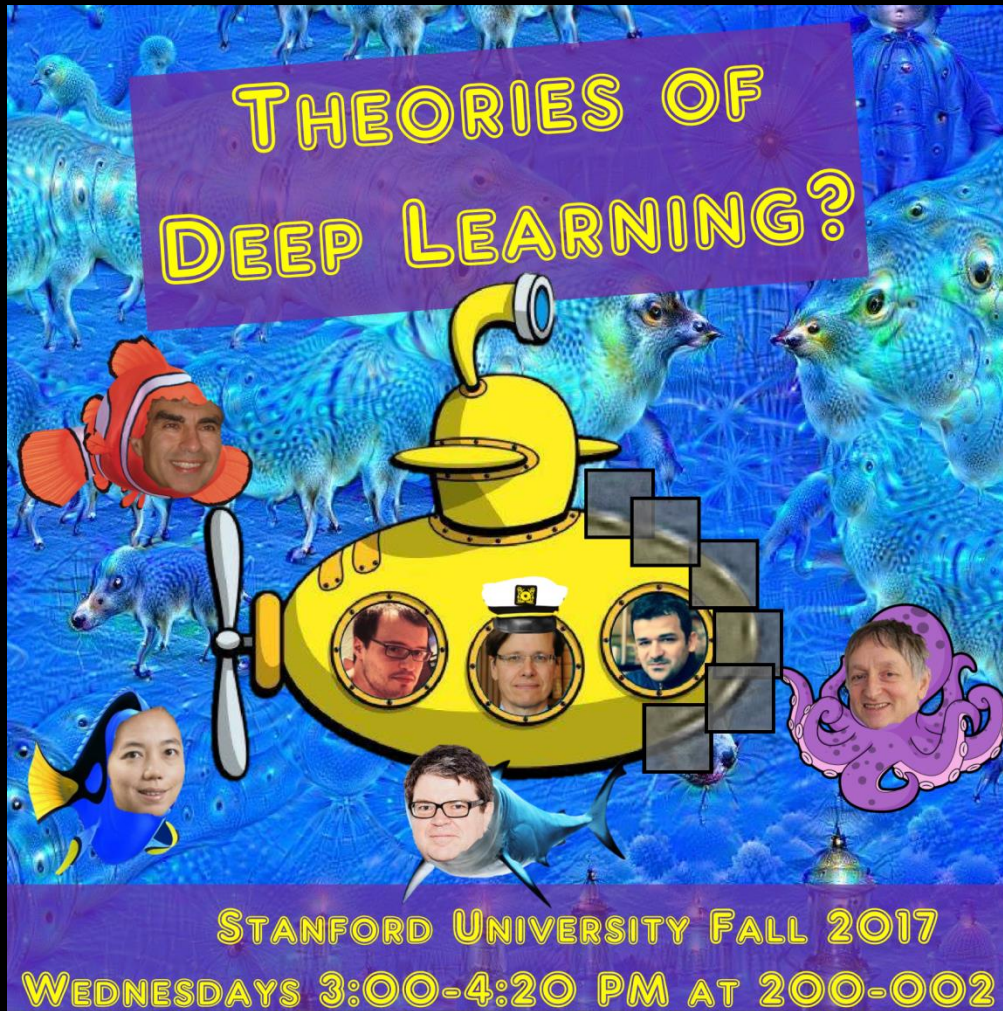
- Lens & telescope → Optics
- Steam engine → Thermodynamics
- Airplane → Aerodynamics
- Radio & Comm. → Info. Theory
- Computer → Computer Science



A Theory for DL ?



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 ..."*



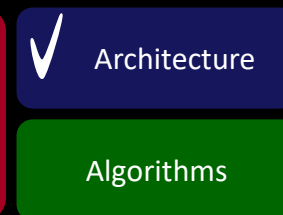
David Donoho: *"... these mirrors are taken from Cinderella's story, telling each that he is the **most beautiful**"*



- Today's talk is on our proposed theory:



Yaniv Romano Vardan Papyan Jeremias Sulam Aviad Aberdam



... and our theory is the best 🤪

This Lecture: More Specifically



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



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

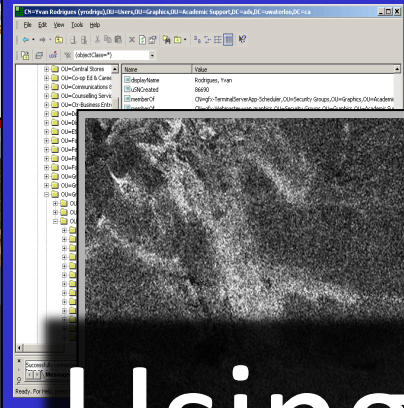


Our Data is Structured

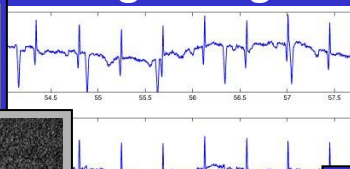
Stock Market



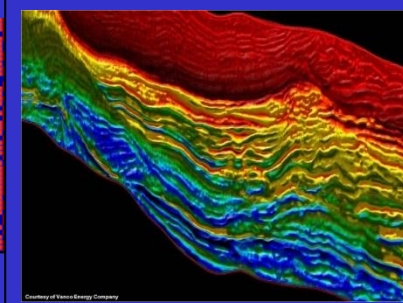
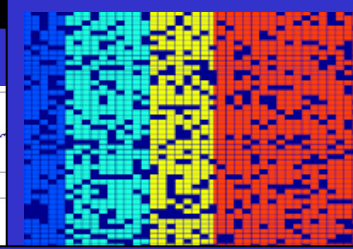
Text Documents



Biological Signals



Matrix Data

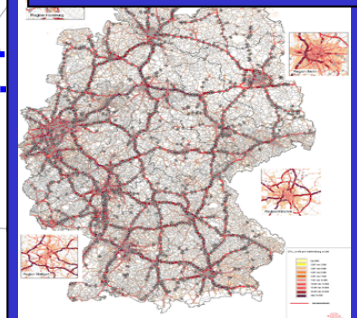
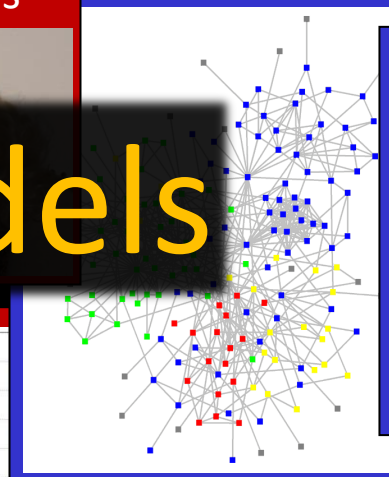


Seismic Data

Still Images

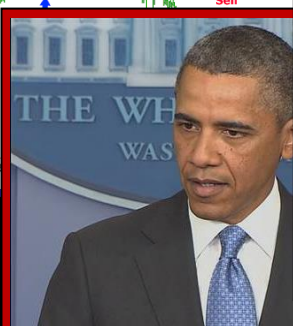


Social Networks



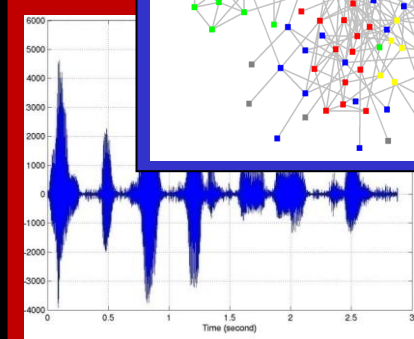
Traffic info

Videos

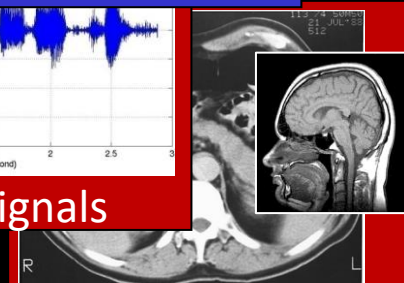


Using models

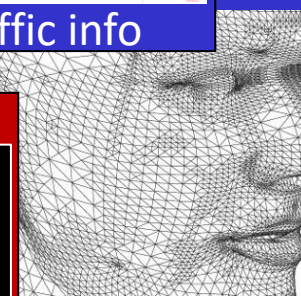
Voice Signals



Medical Imaging



3D Objects

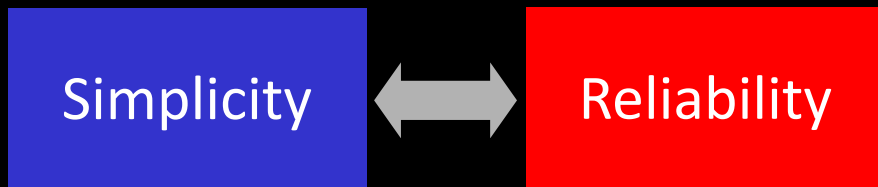


- We are surrounded by various diverse sources of massive information
- Each of these sources have an internal structure, which can be exploited
- This structure, when identified, is the engine behind the ability to process data
- How to identify structure?



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



- 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

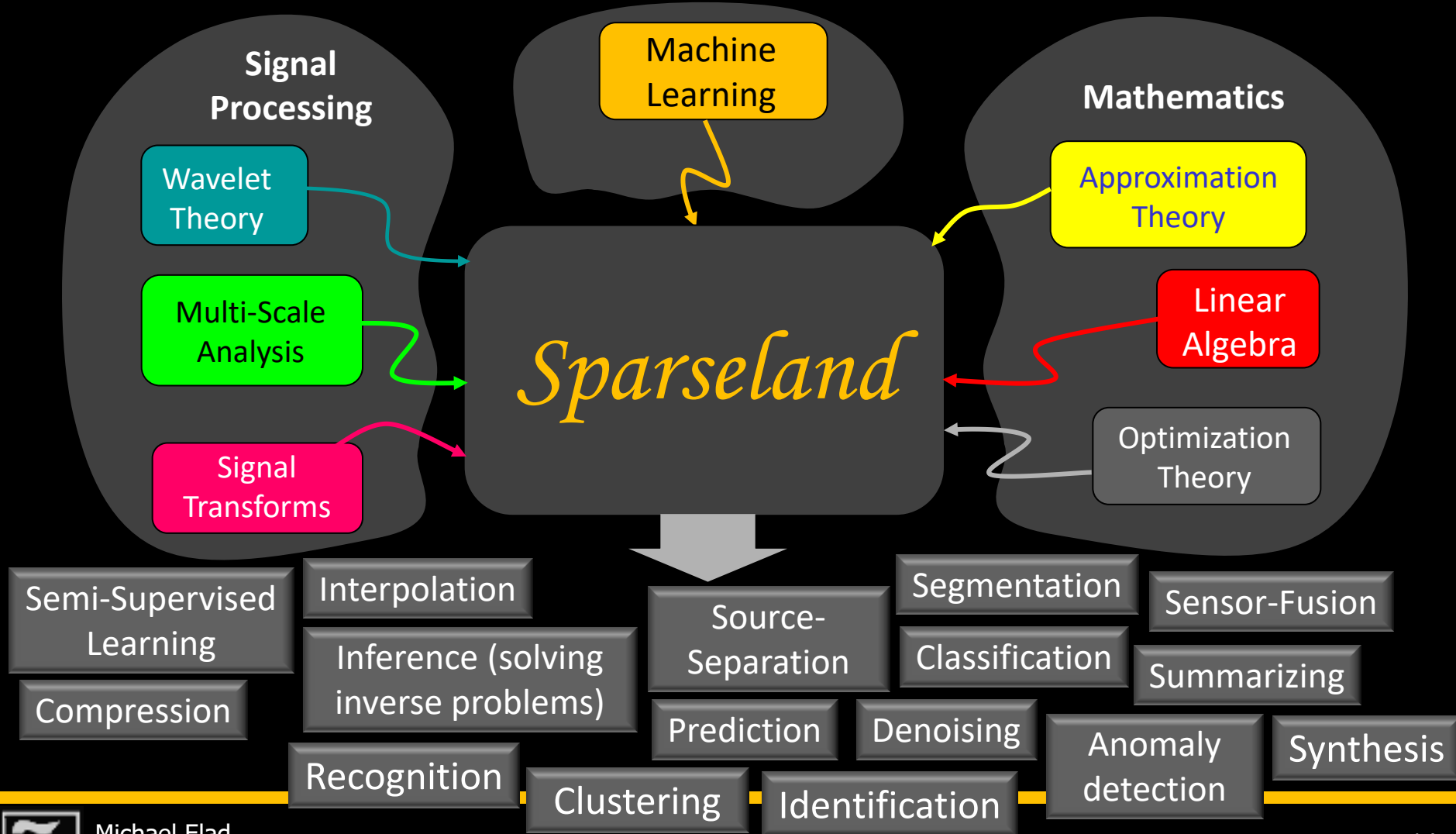
- We shall describe this and descendant versions of it that lead all the way to ... **deep-learning**



Multi-Layered Convolutional Sparse Modeling

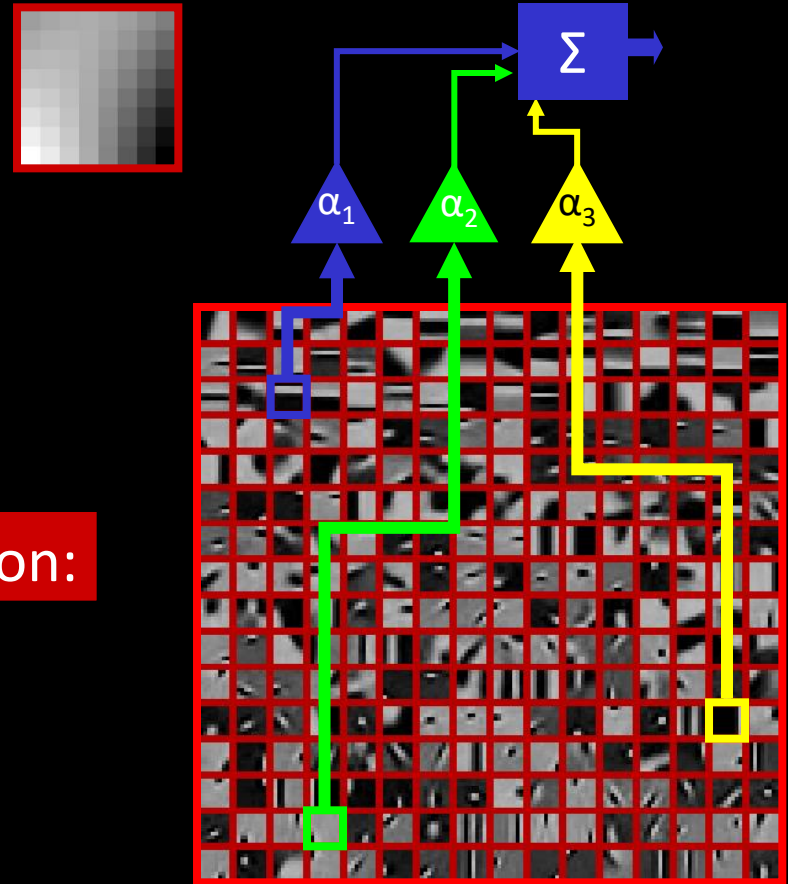


A New Emerging Model



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

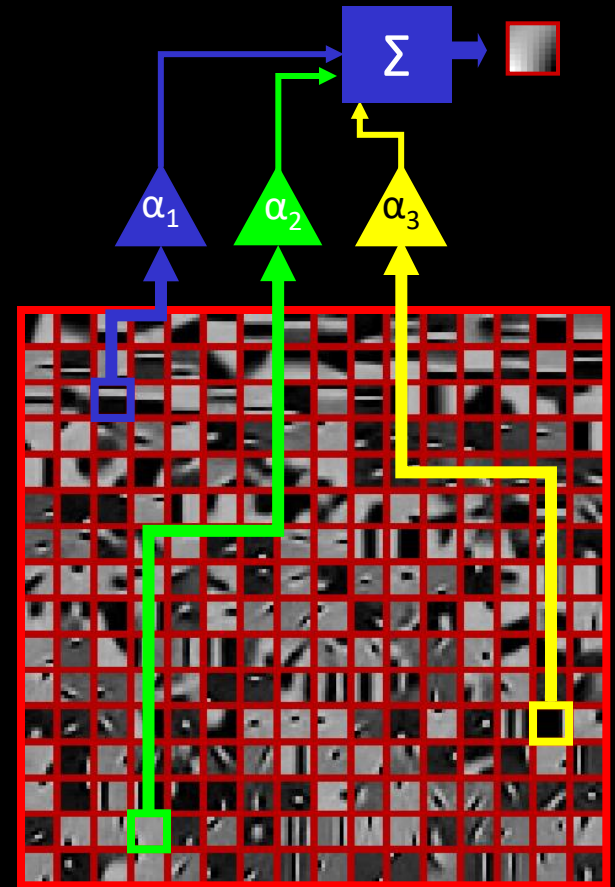


The *Sparseland* Model

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)

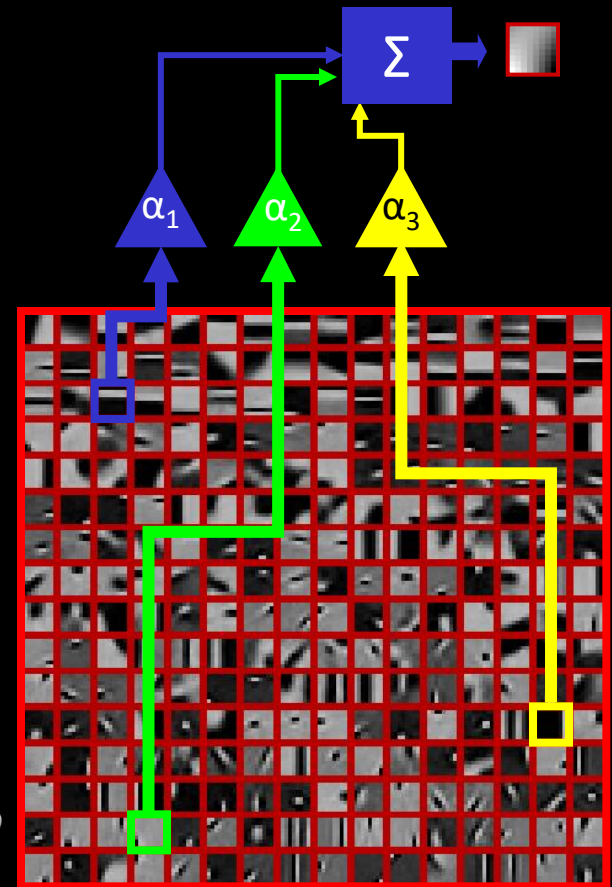
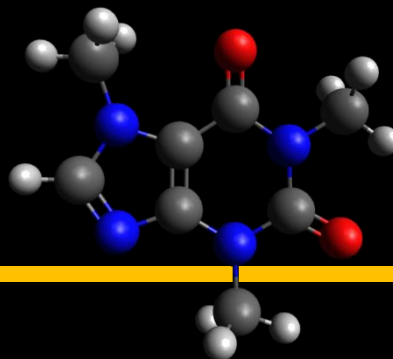


Chemistry of Data

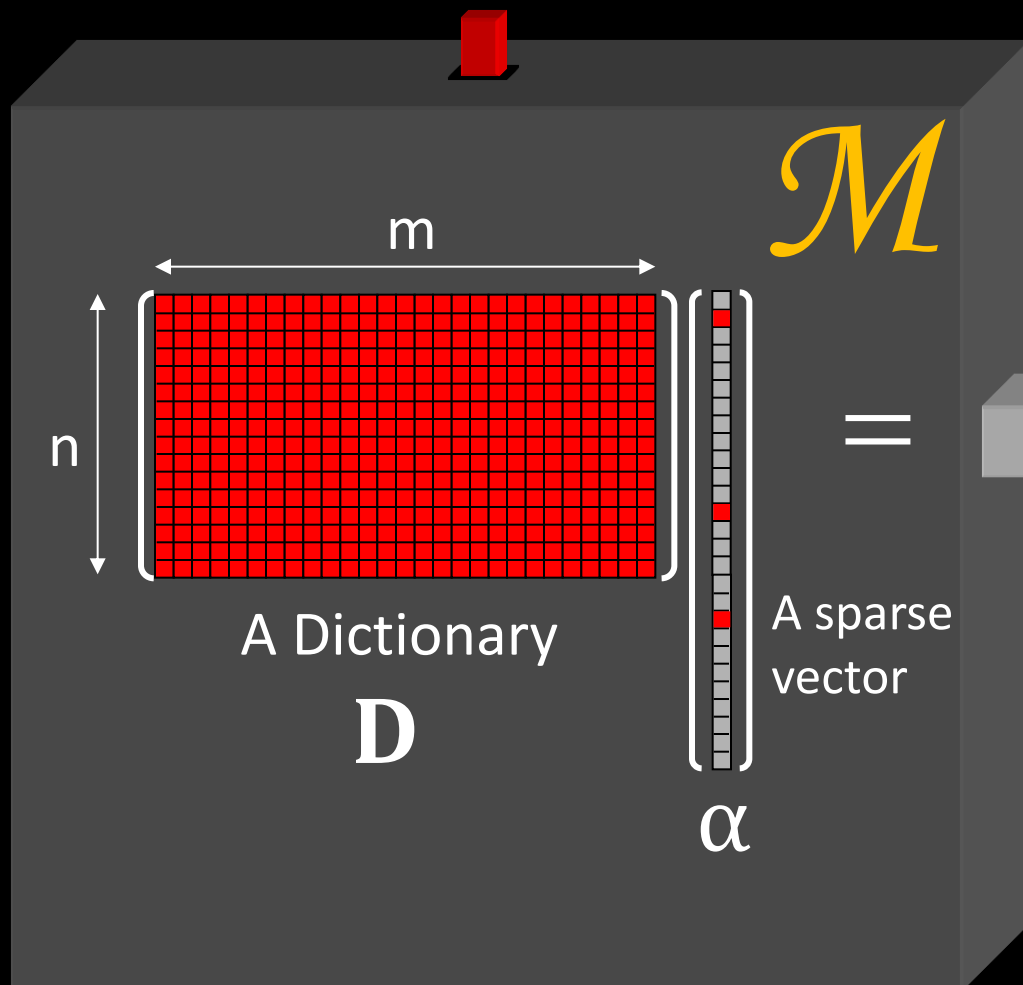
We could refer to the *Sparseland* model as the **chemistry** of information:

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

1	H																	2	He																
3	Li	4	Be																	5	B	6	C	7	N	8	O	9	F	10	Ne				
11	Na	12	Mg																	13	Al	14	Si	15	P	16	S	17	Cl	18	Ar				
19	K	20	Ca	21	Sc	22	Ti	23	V	24	Cr	25	Mn	26	Fe	27	Co	28	Ni	29	Cu	30	Zn	31	Ga	32	Ge	33	As	34	Se	35	Br	36	Kr
37	Rb	38	Sr	39	Y	40	Zr	41	Nb	42	Mo	43	Tc	44	Ru	45	Rh	46	Pd	47	Ag	48	Cd	49	In	50	Sn	51	Sb	52	Te	53	I	54	Xe
55	Cs	56	Ba	*	72	Hf	73	Ta	74	W	75	Re	76	Os	77	Ir	78	Pt	79	Au	80	Hg	81	Tl	82	Pb	83	Bi	84	Po	85	At	86	Rn	
87	Fr	88	Ra	**	104	Rf	105	Db	106	Sg	107	Bh	108	Hs	109	Mt	110	Ds	111	Rg	112	Cn	113	Uut	114	Fl	115	Uup	116	Lv	117	Uus	118	Uuo	
				57	La	58	Ce	59	Pr	60	Nd	61	Pm	62	Sm	63	Eu	64	Gd	65	Tb	66	Dy	67	Ho	68	Er	69	Tm	70	Yb	71	Lu		
				89	Ac	90	Th	91	Pa	92	U	93	Np	94	Pu	95	Am	96	Cm	97	Bk	98	Cf	99	Es	100	Fm	101	Md	102	No	103	Lr		



Sparseland: A Formal Description



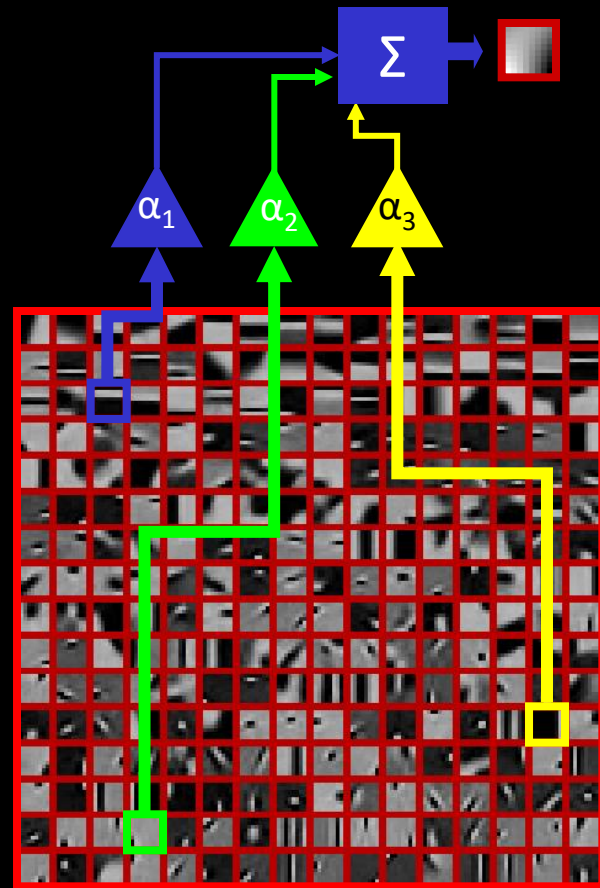
- Every column in \mathbf{D} (**dictionary**) is a prototype signal (**atom**)
- The vector $\underline{\alpha}$ is generated with few non-zeros at arbitrary locations and values
- This is a generative model that describes how (**we believe**) signals are created

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

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

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



Atom Decomposition Made Formal

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



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

$$\begin{matrix} n \\ \left[\begin{array}{c} \text{Red Grid } D \end{array} \right] \\ m \end{matrix} \alpha = x$$

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

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

Approximation Algorithms

Basis Pursuit

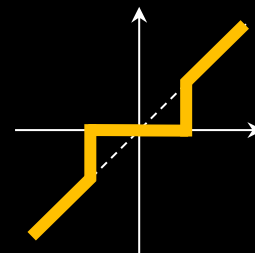
Change the L_0 into L_1
and then the problem
becomes convex and
manageable

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

Thresholding

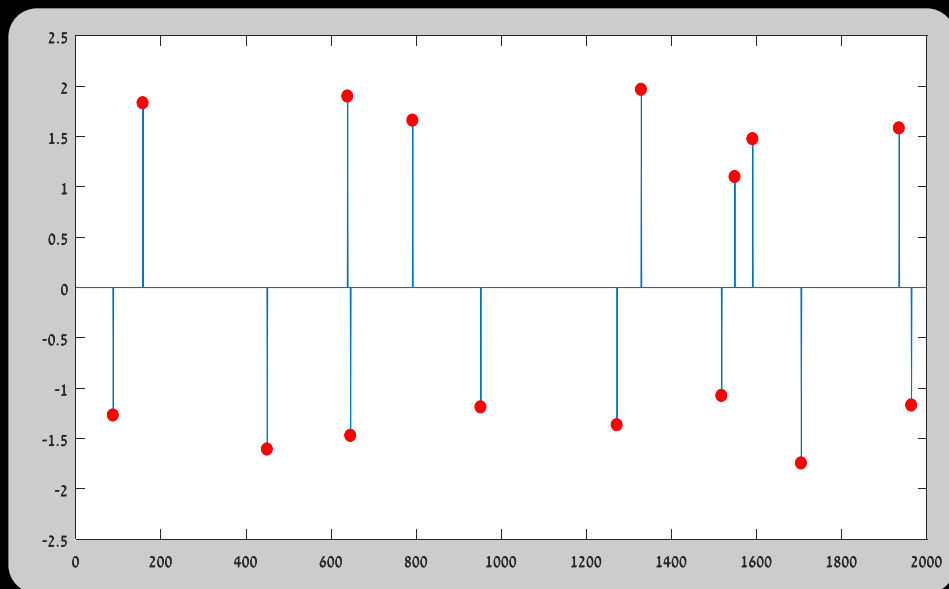
Multiply y by \mathbf{D}^T
and apply shrinkage:

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

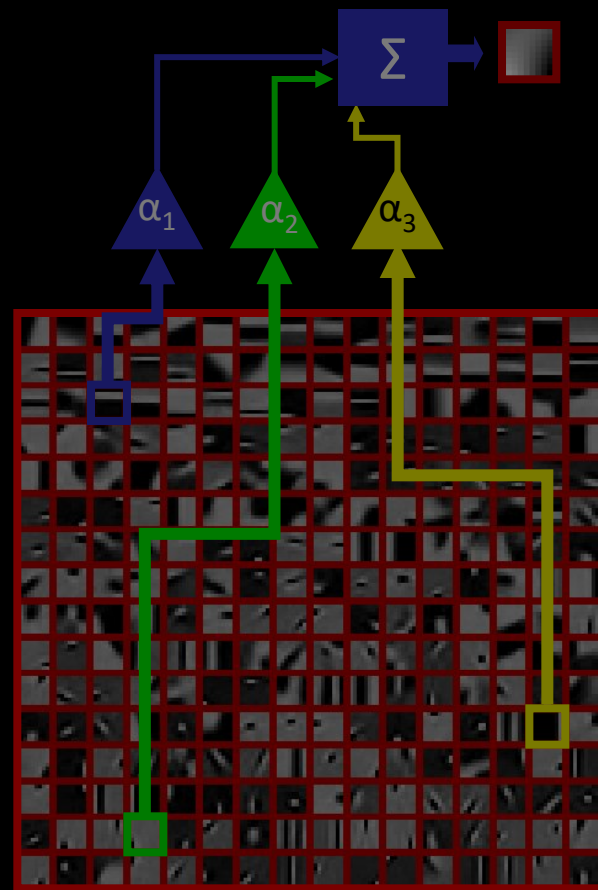


Difficulties with *Sparseland*

- 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

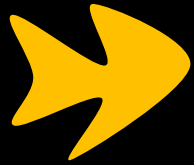


The Mutual Coherence

- Compute $\begin{bmatrix} \mathbf{D}^T \end{bmatrix} \begin{bmatrix} \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{D}^T \mathbf{D} \end{bmatrix}$
Assume normalized columns
- 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



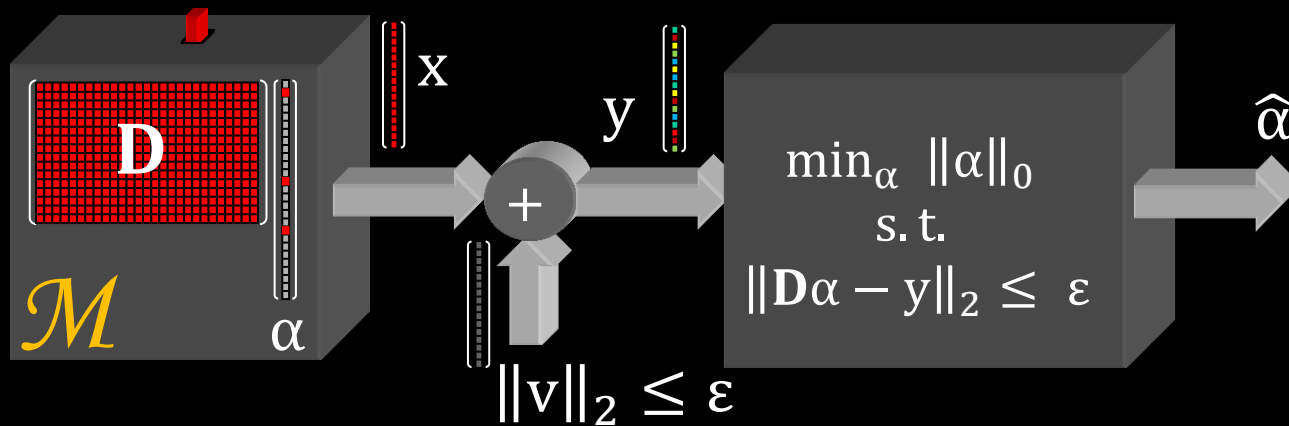
Theorem: **Given** a noisy signal $y = \mathbf{D}\alpha + v$ where $\|v\|_2 \leq \varepsilon$ and α is sufficiently sparse,

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

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

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

Donoho, Elad & Temlyakov ('06)



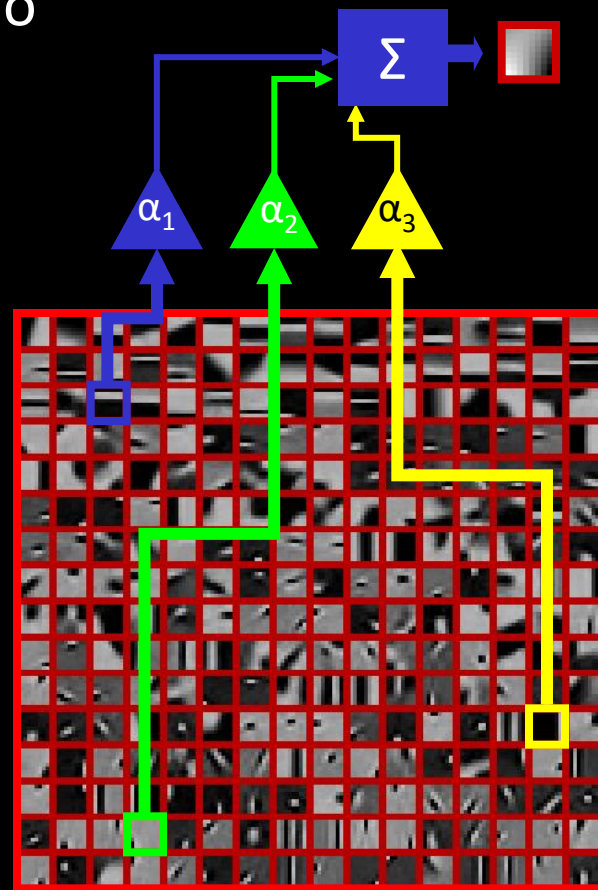
Comments:

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



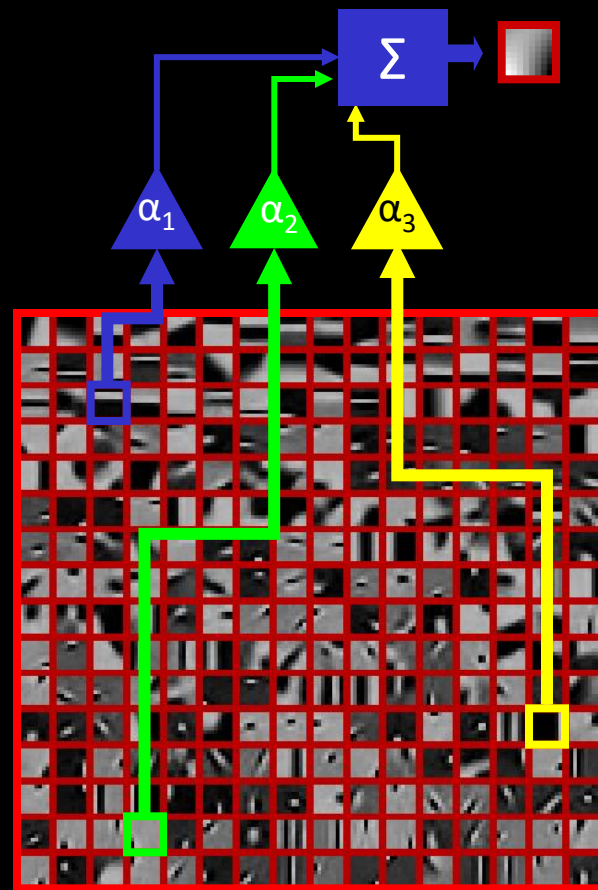
Difficulties with *Sparseland*

- 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



Difficulties with *Sparseland*

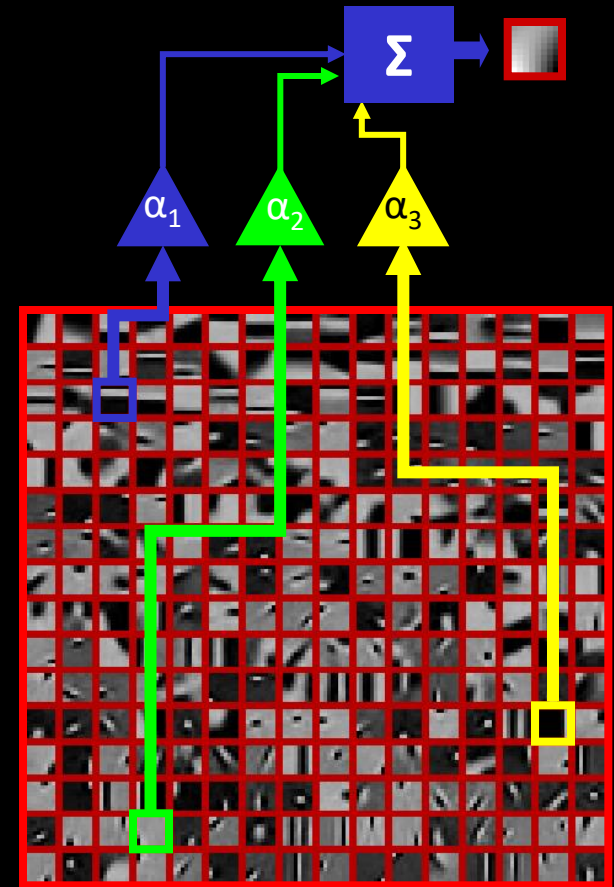
- Problem 3: Why is this model suitable to describe various sources? e.g., Is it good for images? Audio? Stocks? ...
- General answer: Yes, this model is extremely effective in representing various sources
 - **Theoretical answer:** Clear connection to other models
 - **Empirical answer:** In a large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



Difficulties with *Sparseland*?

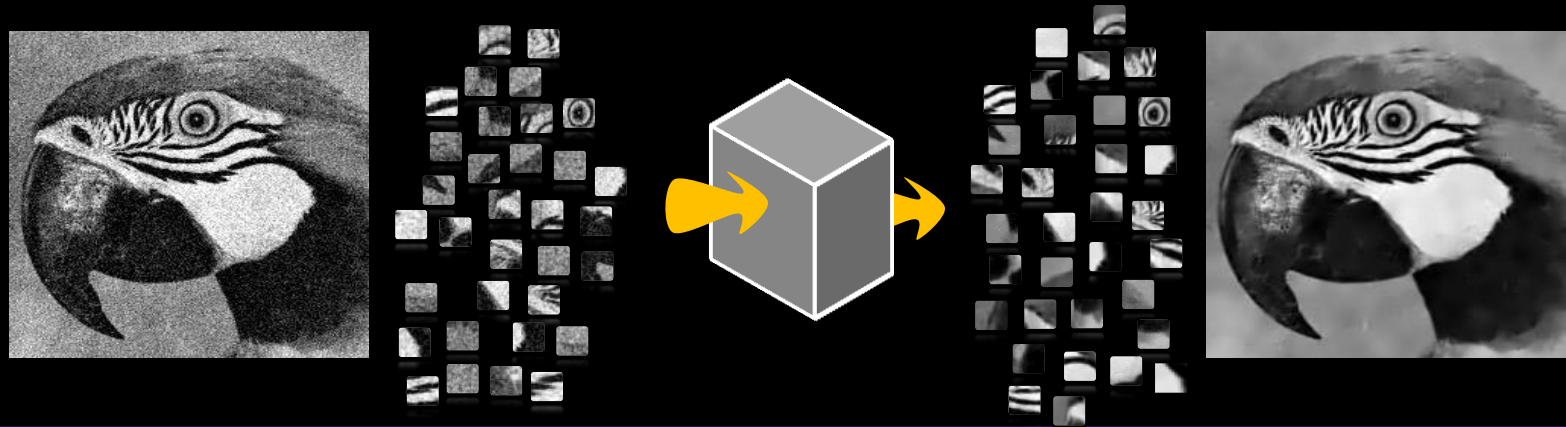
- Problem 1: Given an image patch, how can we find its atom decomposition?
- Problem 2: Given a family of signals, how do we find the dictionary to represent it well?
- Problem 3: Is this model flexible enough to describe various sources? E.g., Is it good for images? audio? ...

**ALL ANSWERED
POSITIVELY AND
CONSTRUCTIVELY**



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

1. V. Pappas, J. Sulam, and M. Elad, *Working Locally Thinking Globally: Theoretical Guarantees for Convolutional Sparse Coding*, IEEE Trans. on Signal Processing, Vol. 65, No. 21, Pages 5687-5701, November 2017.

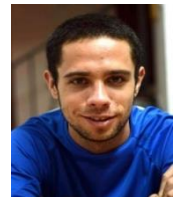
Joint work with



Yaniv Romano



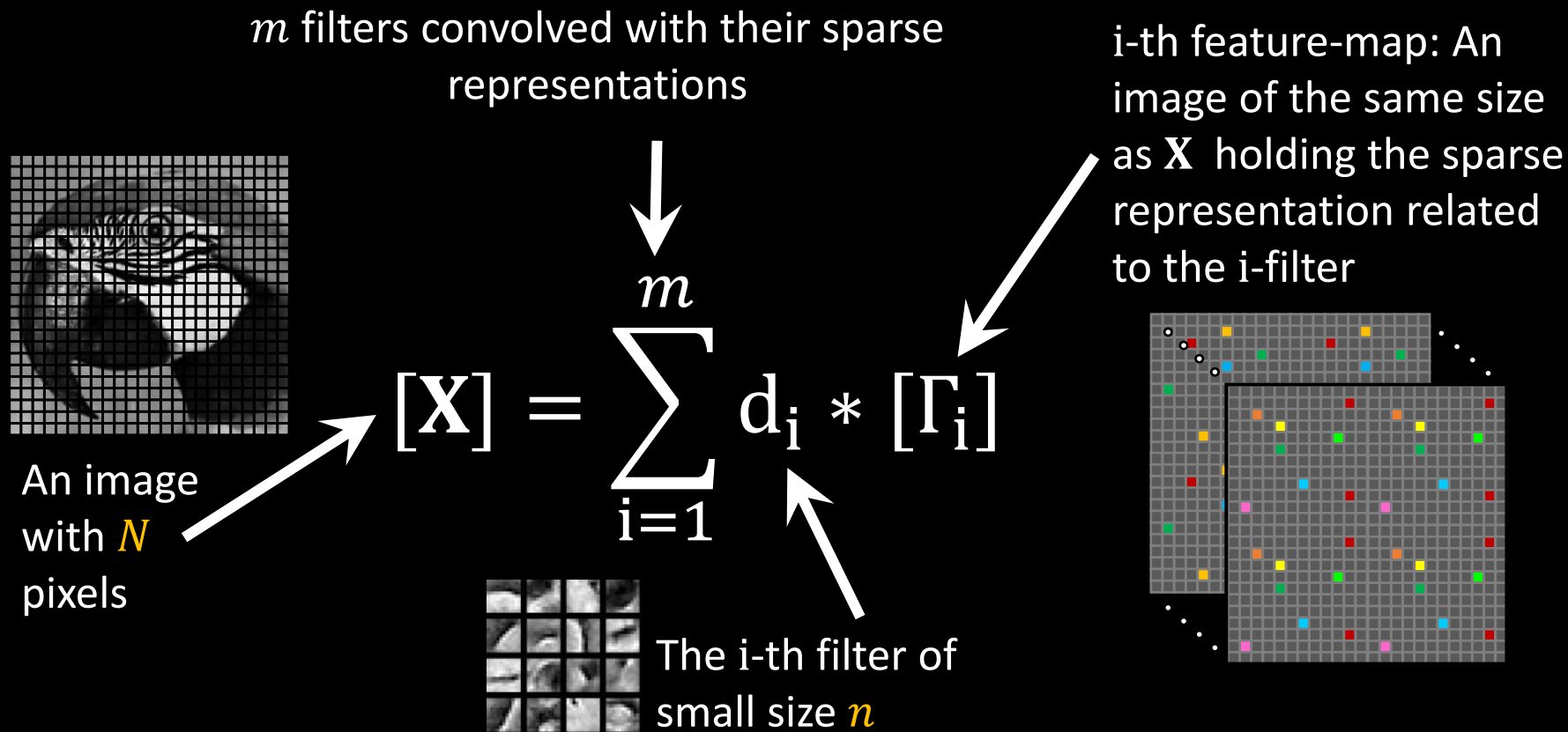
Vardan Pappas



Jeremias Sulam



Convolutional Sparse Coding (CSC)



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



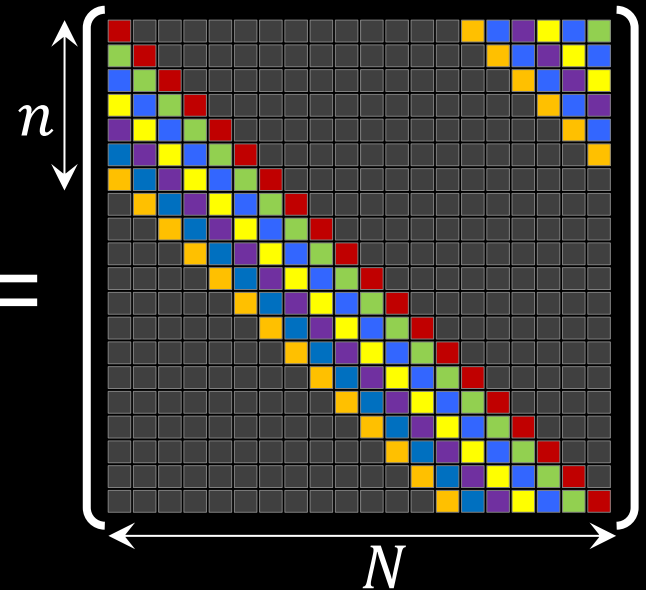
CSC in Matrix Form

- Here is an alternative global sparsity-based model formulation

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

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

$$\begin{bmatrix} \text{red} \\ \text{green} \\ \text{blue} \\ \text{yellow} \\ \text{purple} \\ \text{orange} \end{bmatrix} \Rightarrow \mathbf{C}^i =$$



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

The CSC Dictionary

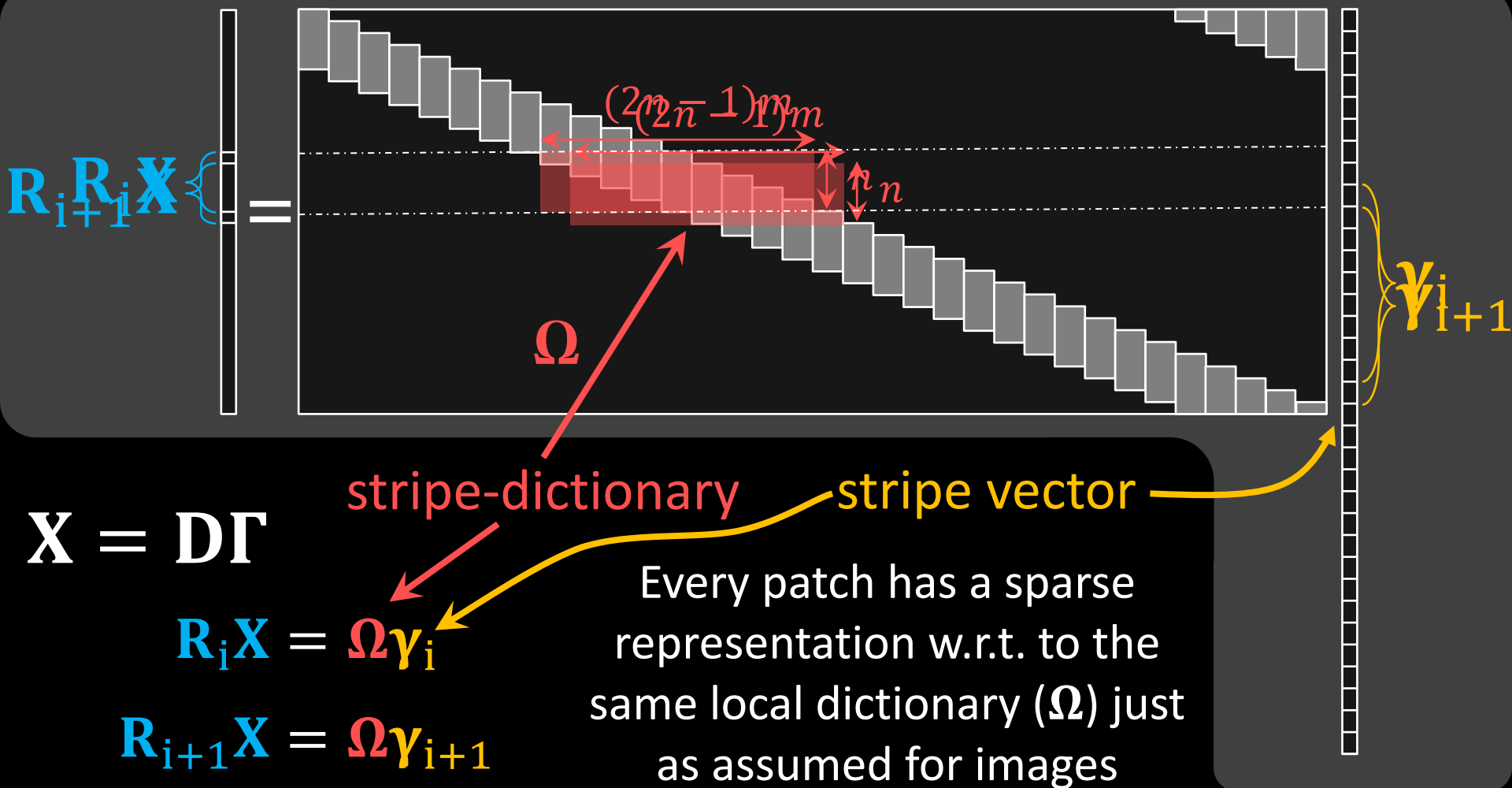
$$[\mathbf{C}^1 \ \mathbf{C}^2 \ \mathbf{C}^3] = \left[\begin{array}{ccc} \text{Grid 1} & \text{Grid 2} & \text{Grid 3} \end{array} \right]$$

$\mathbf{D}_L \swarrow$

$$\mathbf{D} = \left[\begin{array}{c} \text{Large Grid} \end{array} \right]$$

The diagram illustrates the structure of the CSC Dictionary. It shows three small grids ($\mathbf{C}^1, \mathbf{C}^2, \mathbf{C}^3$) and a large grid (\mathbf{D}). The large grid \mathbf{D} is composed of multiple blocks, each of size $m \times n$, arranged in a sparse pattern. The label \mathbf{D}_L points to the first block in the first row of \mathbf{D} .

Why CSC?



Classical Sparse Theory for CSC ?

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

Theorem: BP is guaranteed to “succeed” if $\|\Gamma\|_0 < \frac{1}{4} \left(1 + \frac{1}{\mu}\right)$

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

$$\mu \geq 0.063$$

- Success of pursuits is

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

- On the other hand, μ is GLOBALLY are allowed!!! This is a very pessimistic result!



Moving to Local Sparsity: **Stripes**

$\ell_{0,\infty}$ Norm: $\|\Gamma\|_{0,\infty}^s = \max_i \|\gamma_i\|_0$

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

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

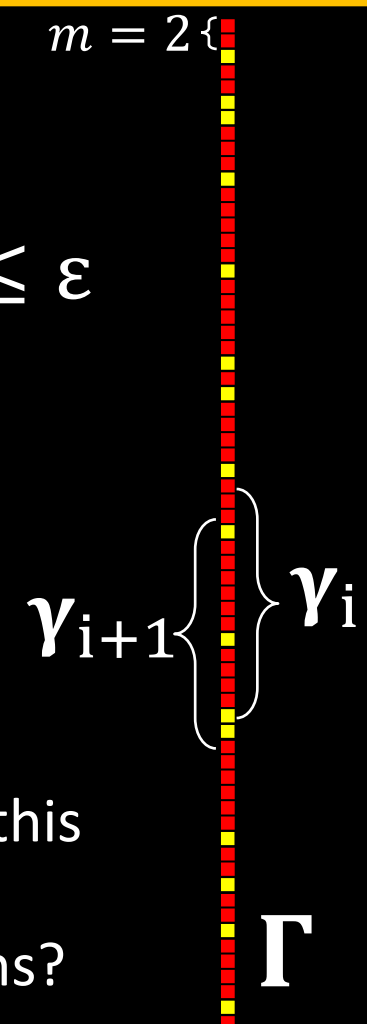
The main question we aim to address is this:

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?

$m = 2\{$

$\gamma_{i+1} \left\{ \right. \gamma_i$


Γ



Success of the Basis Pursuit

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

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


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

then Basis Pursuit performs very-well:

1. The support of Γ_{BP} is contained in that of Γ
2. $\|\Gamma_{\text{BP}} - \Gamma\|_{\infty} \leq 7.5\|E\|_{2,\infty}^p$
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

Papayan, Sulam
& Elad ('17)



Multi-Layered Convolutional Sparse Modeling



Yaniv Romano



Vardan Papyan



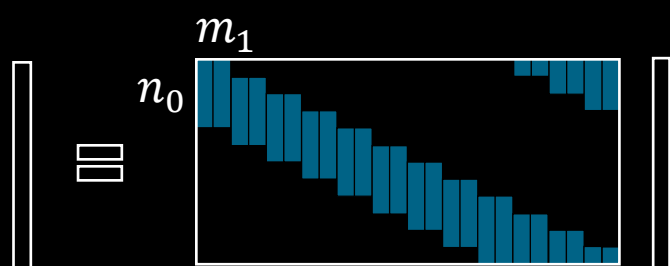
Jeremias Sulam

2. V. Papyan, Y. Romano, and M. Elad, Convolutional Neural Networks Analyzed via Convolutional Sparse Coding, *Journal of Machine Learning Research*, Vol. 18, Pages 1-52, July 2017.
3. V. Papyan, Y. Romano, J. Sulam, and M. Elad, *Theoretical Foundations of Deep Learning via Sparse Representations*, *IEEE Signal Processing Magazine*, Vol. 35, No. 4, Pages 72-89, June 2018.



From CSC to Multi-Layered CSC

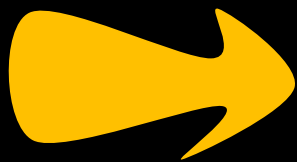
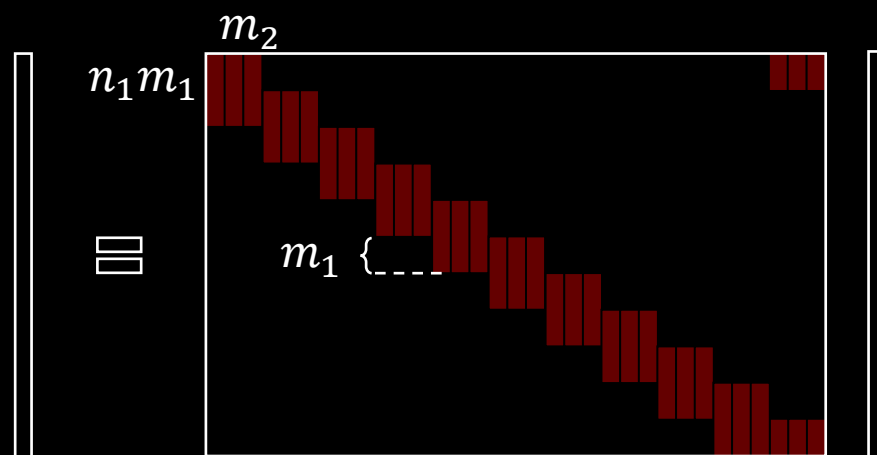
$$\mathbf{X} \in \mathbb{R}^N \quad \mathbf{D}_1 \in \mathbb{R}^{N \times Nm_1} \quad \mathbf{\Gamma}_1 \in \mathbb{R}^{Nm_1}$$



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

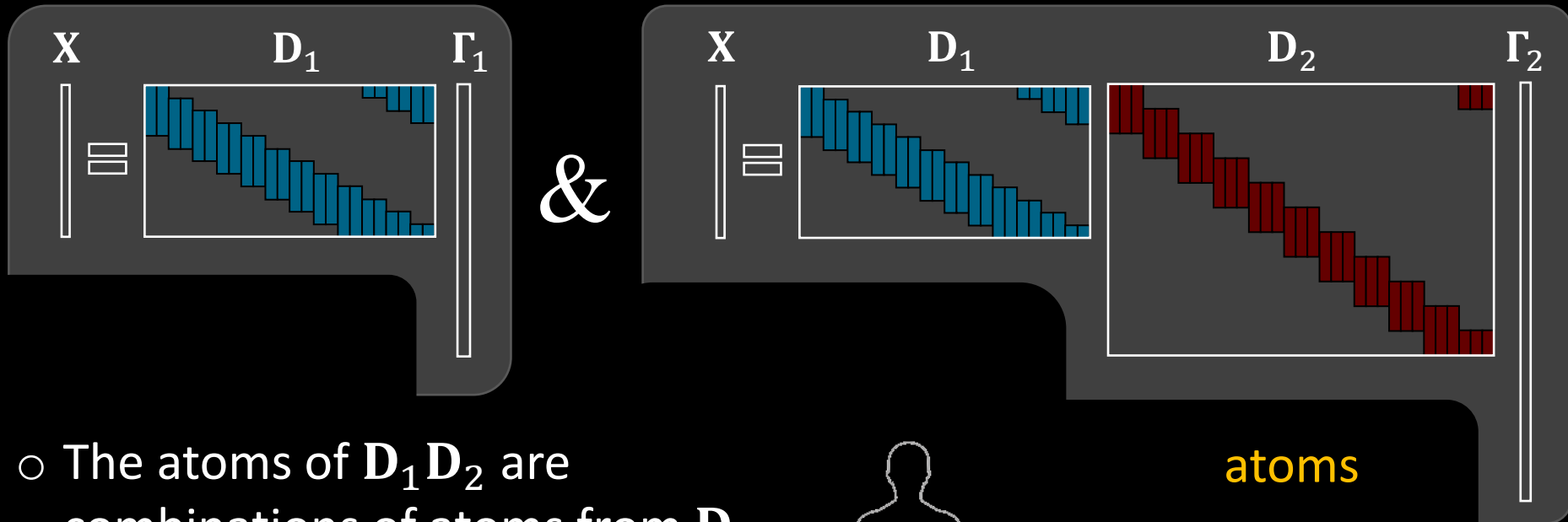
We propose to impose the same structure on the representations **themselves**

$$\mathbf{\Gamma}_1 \in \mathbb{R}^{Nm_1} \quad \mathbf{D}_2 \in \mathbb{R}^{Nm_1 \times Nm_2} \quad \mathbf{\Gamma}_2 \in \mathbb{R}^{Nm_2}$$

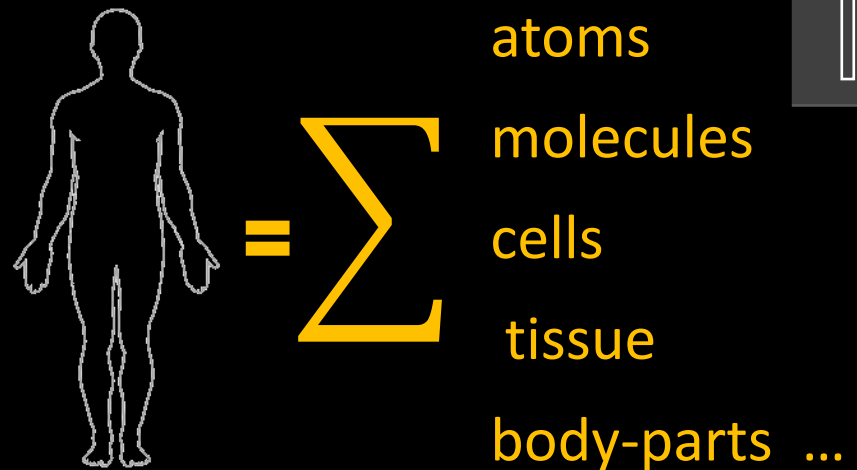


Multi-Layer CSC (ML-CSC)

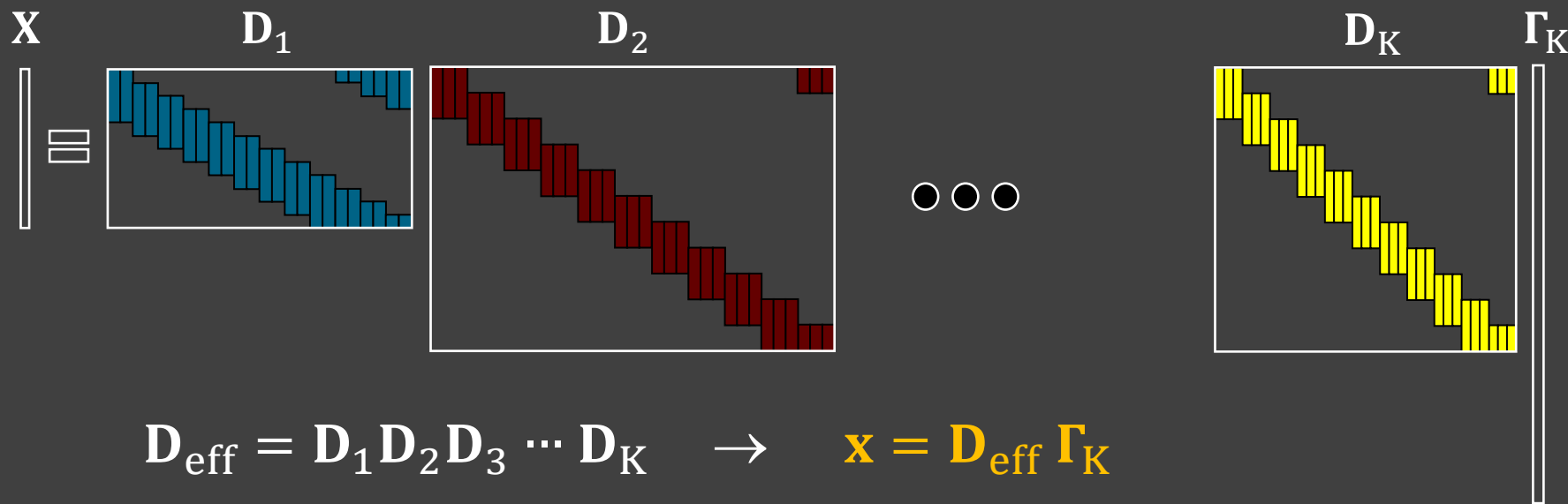
Intuition: From Atoms to Molecules



- The atoms of $D_1 D_2$ are combinations of atoms from D_1 - these are now **molecules**
- Thus, this model offers different **levels of abstraction** in describing X



Intuition: From Atoms to Molecules



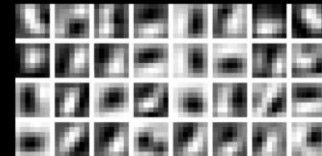
- This is a special *SparseLand* (indeed, a CSC) model
- However: A key property in our model: the intermediate representations are required to be sparse as well

A Small Taste: Model Training (MNIST)

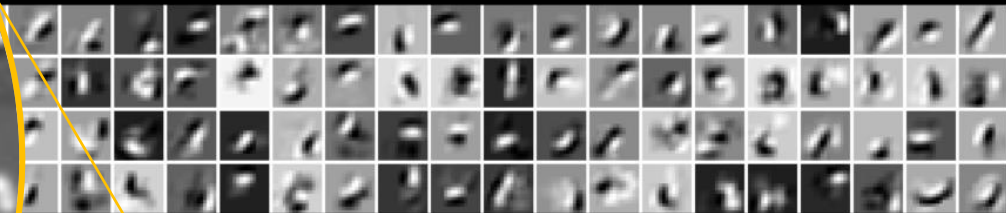
MNIST Dictionary:

- D_1 : 32 filters of size 2 (dense)
- D_2 : 128 filters of size 1 - 99.09 % sparse
- D_3 : 1024 filters of size 1 - 99.99 % sparse

D_1 (7×7)



$D_1 D_2$ (15×15)



$D_1 D_2 D_3$ (28×28)



ML-CSC: Pursuit

- Deep-Coding Problem (**DCP_λ**) (dictionaries are known):

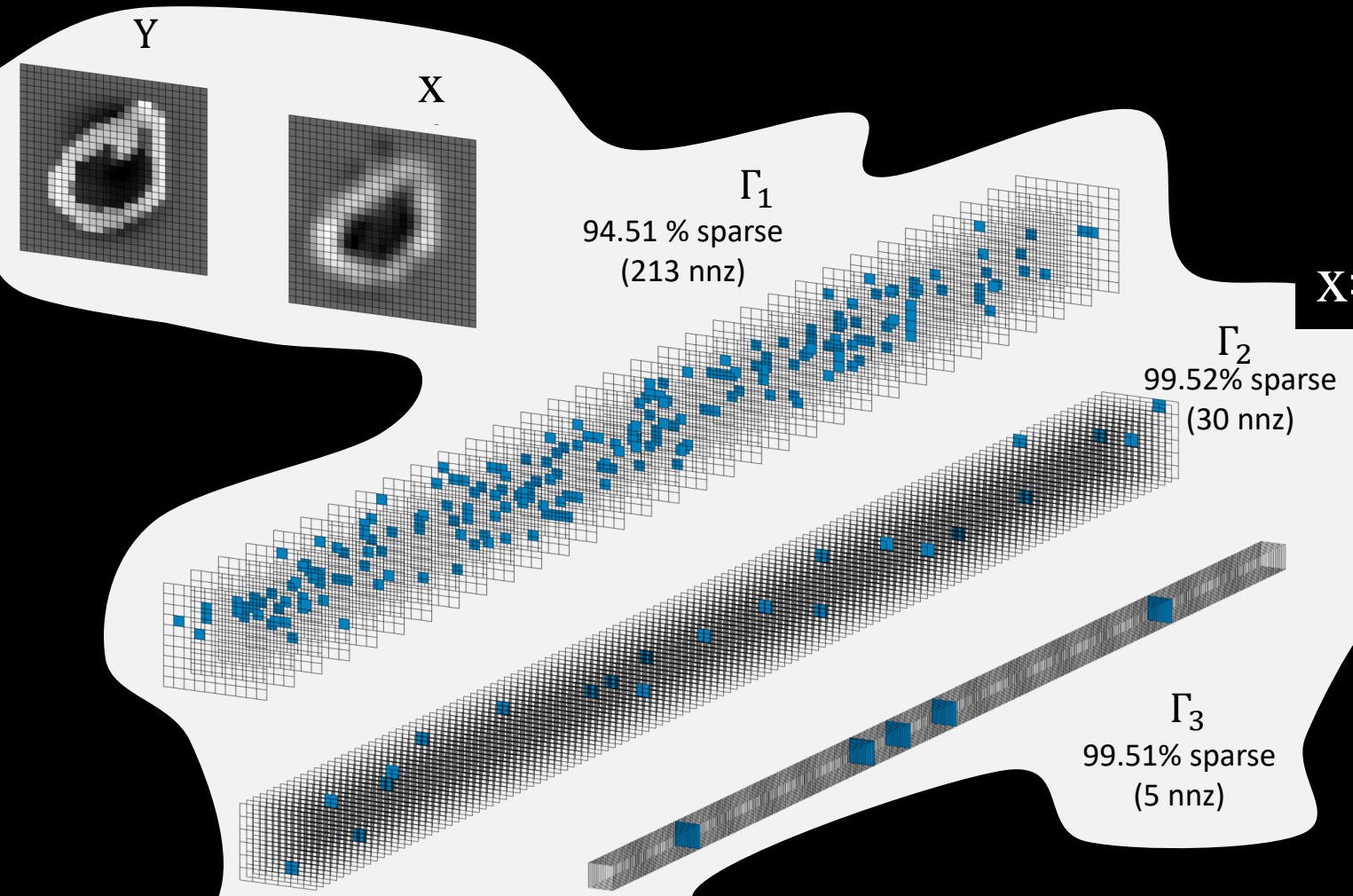
$$\left\{ \begin{array}{ll} \mathbf{X} = \mathbf{D}_1 \mathbf{\Gamma}_1 & \|\mathbf{\Gamma}_1\|_{0,\infty}^s \leq \lambda_1 \\ \mathbf{\Gamma}_1 = \mathbf{D}_2 \mathbf{\Gamma}_2 & \|\mathbf{\Gamma}_2\|_{0,\infty}^s \leq \lambda_2 \\ \vdots & \vdots \\ \mathbf{\Gamma}_{K-1} = \mathbf{D}_K \mathbf{\Gamma}_K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

- Or, more realistically for noisy signals,

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



A Small Taste: Pursuit



ML-CSC: The Simplest Pursuit



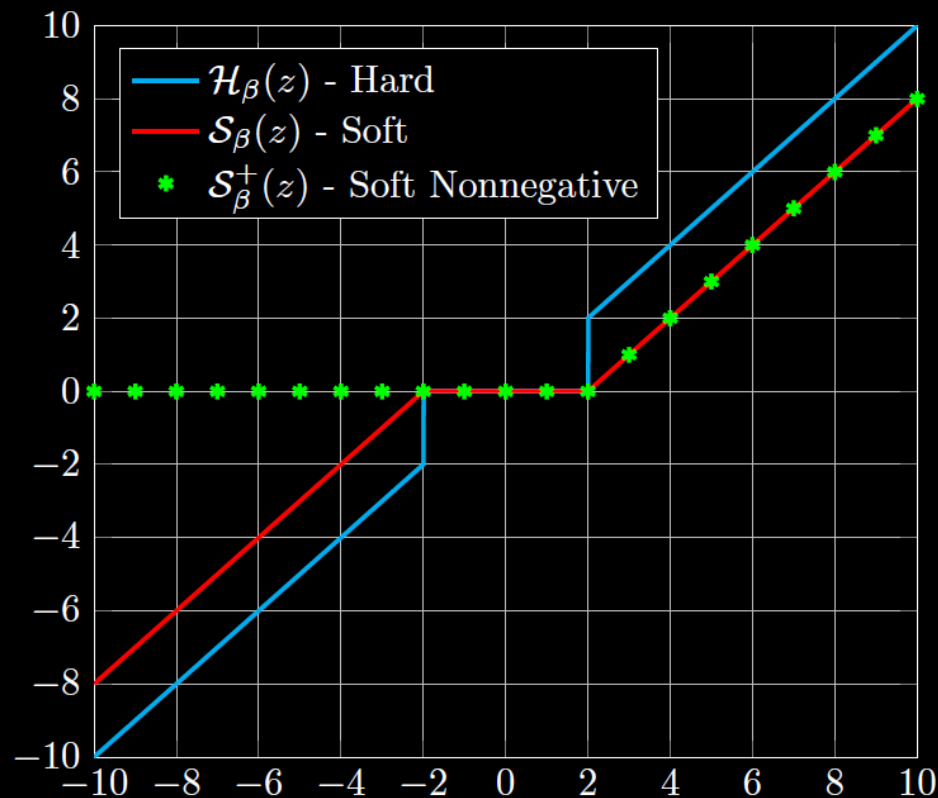
The simplest pursuit algorithm (single-layer case) is the THR algorithm, which operates on a given input signal \mathbf{Y} by:

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

and $\mathbf{\Gamma}$ is sparse



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



Consider this for Solving the DCP

- Layered Thresholding (LT):

Estimate Γ_1 via the THR algorithm

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

Estimate Γ_2 via the THR algorithm

$$(\mathbf{DCP}_{\lambda}^{\varepsilon}): \text{Find } \{\Gamma_j\}_{j=1}^K \text{ s.t. } \left\{ \begin{array}{ll} \|\mathbf{Y} - \mathbf{D}_1 \Gamma_1\|_2 \leq \varepsilon & \|\Gamma_1\|_{0,\infty}^s \leq \lambda_1 \\ \Gamma_1 = \mathbf{D}_2 \Gamma_2 & \|\Gamma_2\|_{0,\infty}^s \leq \lambda_2 \\ \vdots & \vdots \\ \Gamma_{K-1} = \mathbf{D}_K \Gamma_K & \|\Gamma_K\|_{0,\infty}^s \leq \lambda_K \end{array} \right\}$$

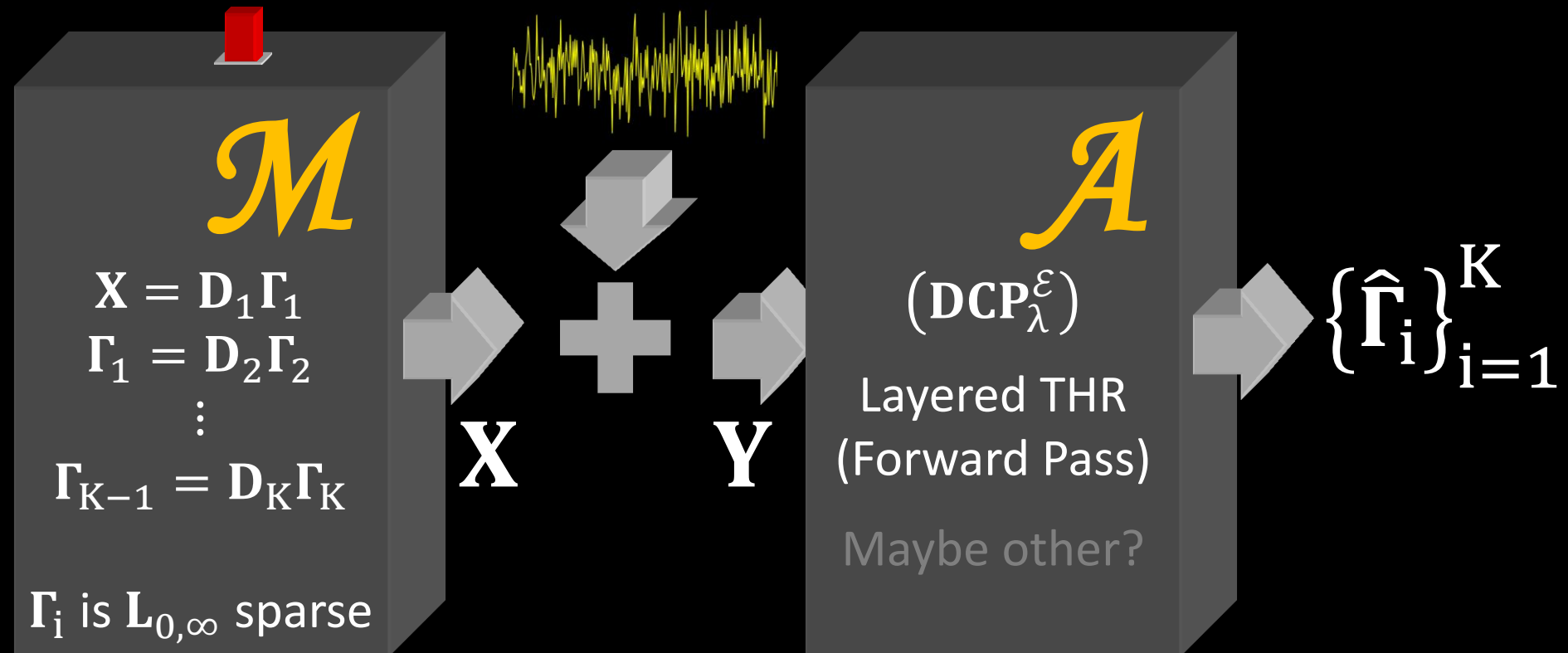
- Now let's take a look at how Conv. Neural Network operates:

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

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 $\|\Gamma_i\|_{0,\infty}^s < \frac{1}{2} \left(1 + \frac{1}{\mu(\mathbf{D}_i)} \cdot \frac{|\Gamma_i^{\min}|}{|\Gamma_i^{\max}|} \right) - \frac{1}{\mu(\mathbf{D}_i)} \cdot \frac{\varepsilon_L^{i-1}}{|\Gamma_i^{\max}|}$
then the **Layered Hard THR** (with the proper thresholds)
finds the correct supports and $\|\Gamma_i^{LT} - \Gamma_i\|_{2,\infty}^p \leq \varepsilon_L^i$, where
we have defined $\varepsilon_L^0 = \|\mathbf{E}\|_{2,\infty}^p$ and

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

Papayan, 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 ...

$(\mathbf{DCP}_{\lambda}^{\varepsilon})$: Find $\{\mathbf{\Gamma}_j\}_{j=1}^K$ s. t.

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

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



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



⋮


Deconvolutional networks

[Zeiler, Krishnan, Taylor & Fergus '10]



Success of the Layered BP

Theorem: Assuming that $\|\Gamma_i\|_{0,\infty}^s < \frac{1}{3} \left(1 + \frac{1}{\mu(\mathbf{D}_i)}\right)$
then the Layered Basis Pursuit performs very well:

- 
1. The support of Γ_i^{LBP} is contained in that of Γ_i
 2. The error is bounded: $\|\Gamma_i^{\text{LBP}} - \Gamma_i\|_{2,\infty}^p \leq \varepsilon_L^i$, where

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

3. Every entry in Γ_i greater than

$$\varepsilon_L^i / \sqrt{\|\Gamma_i\|_{0,\infty}^p} \text{ will be found}$$

Problems:

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

Papayan, Romano & Elad ('17)



Layered Iterative Thresholding

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



Layered Iterative Soft-Thresholding Algorithm (ISTA):

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

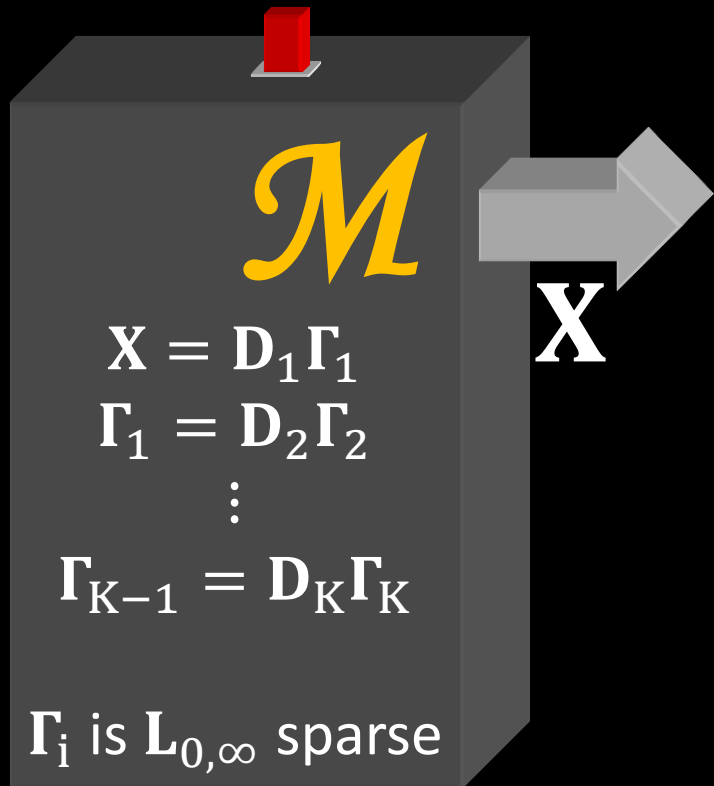
Note that our suggestion implies that groups of layers share the same dictionaries

Can be seen as a very deep residual neural network

[He, Zhang, Ren, & Sun '15]



Where are the Labels?

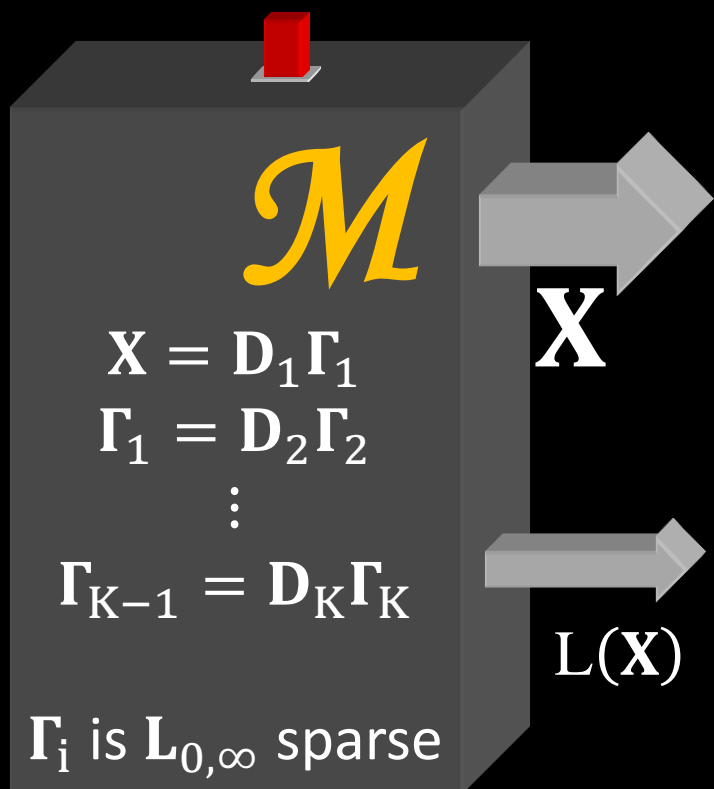


Answer 1:

- 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

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

Where are the Labels?



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

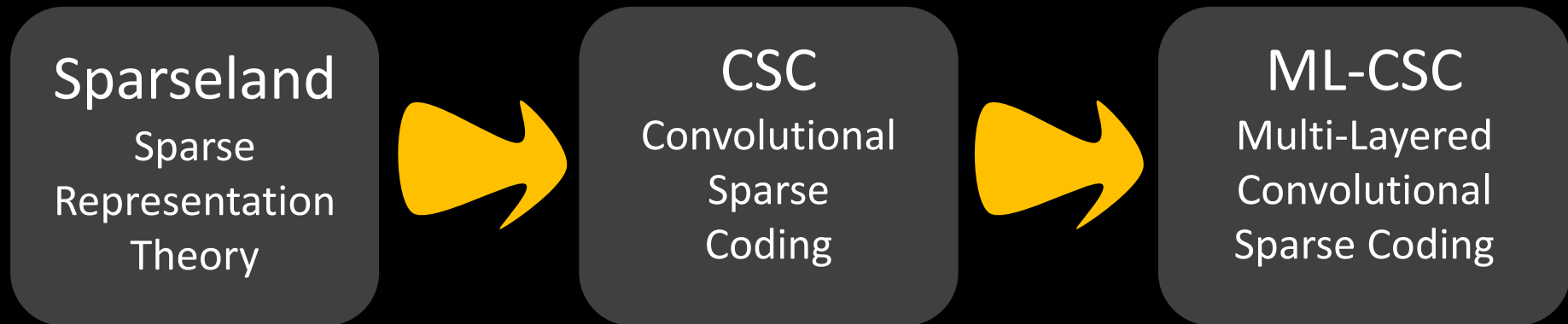
Answer 2:

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

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

- This assumes that knowing the representations suffices for classification → **supervised** mode
- Thus, a successful pursuit algorithm can lead to an accurate recognition if the network is augmented by a FC classification layer
- In fact, we can analyze theoretically the classification accuracy and the sensitivity to adversarial noise – see later

What About Learning?



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: Two algorithms were proposed – unsupervised and supervised

Time to Conclude

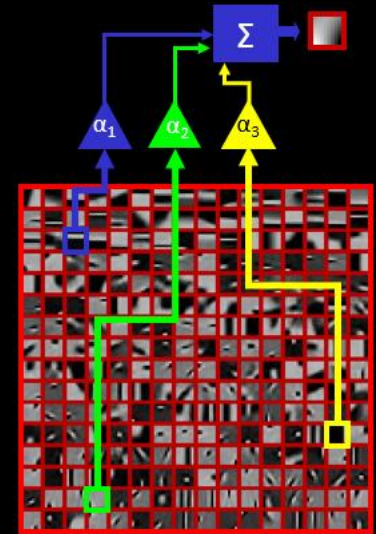


This Talk

Sparseland



The desire to
model data



We spoke about the importance of models in signal/image processing and described *Sparseland* in details



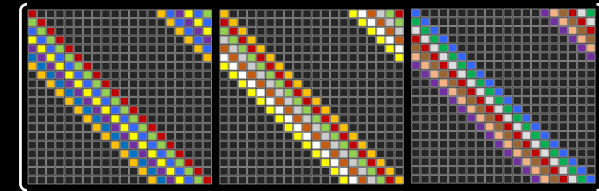
This Talk

Sparseland

The desire to
model data



Novel View of
Convolutional
Sparse Coding



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



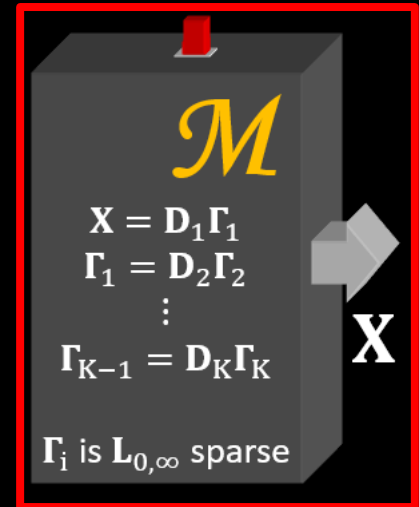
This Talk

Sparseland

The desire to
model data

Novel View of
Convolutional
Sparse Coding

Multi-Layer
Convolutional
Sparse Coding



We propose a multi-layer extension of
CSC, shown to be tightly connected to CNN

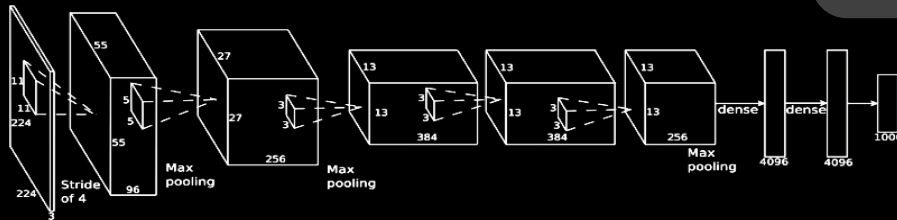


This Talk

Sparseland

The desire to
model data

Novel View of
Convolutional
Sparse Coding



A novel interpretation
and theoretical
understanding of CNN

Multi-Layer
Convolutional
Sparse Coding

The ML-CSC was shown to enable a theoretical
study of CNN, along with new insights



This Talk

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

A novel interpretation
and theoretical
understanding of CNN

Sparseland

The desire to
model data

Novel View of
Convolutional
Sparse Coding

Multi-Layer
Convolutional
Sparse Coding



This Talk

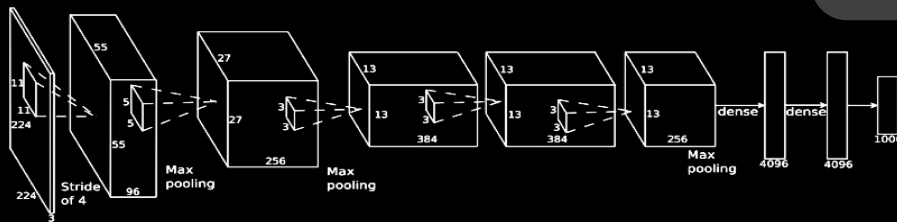
Sparseland

The desire to
model data

Novel View of
Convolutional
Sparse Coding

Multi-Layer
Convolutional
Sparse Coding

A novel interpretation
and theoretical
understanding of CNN

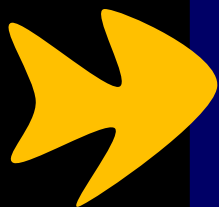


Take Home Message 2:
The Multi-Layer
Convolutional Sparse
Coding model could be
a new platform for
understanding and
developing deep-
learning solutions



On a Personal Note ...

Disclaimer: True, I am biased, so take what I say with a grain of salt ...



Conjecture: **Sparse modeling** of data is at the heart of the Deep-Learning ideas, and as such, it **is the main avenue for** developing **theoretical foundations** for this field.

Elad ('19)

My research activity in the foreseen future will be dedicated to establishing this connection, addressing various aspects of it



Fresh from the Oven

My team's work proceeds along the above-described line of thinking:

4. J. Sulam, V. Pappyan, Y. Romano, and M. Elad, Multi-Layer Convolutional Sparse Modeling: **Pursuit and Dictionary Learning**, IEEE Trans. on Signal Proc., Vol. 66, No. 15, Pages 4090-4104, August 2018.
5. A. Aberdam, J. Sulam, and M. Elad, Multi Layer Sparse Coding: **the Holistic Way**, SIAM Journal on Mathematics of Data Science (SIMODS), Vol. 1, No. 1, Pages 46-77.
6. J. Sulam, A. Aberdam, A. Beck, and M. Elad, **On Multi-Layer Basis Pursuit**, Efficient Algorithms and Convolutional Neural Networks, to appear in IEEE T-PAMI.
7. Y. Romano, A. Aberdam, J. Sulam, and M. Elad, **Adversarial Noise Attacks** of Deep Learning Architectures – Stability Analysis via Sparse Modeled Signals, to appear in JMIV.
8. Ev Zisselman, Jeremias Sulam, and Michael Elad, A Local Block Coordinate Descent Algorithm **for the CSC Model**, CVPR 2019.
9. I. Rey-Otero, J. Sulam, and M. Elad, **Variations on the CSC model**, submitted to IEEE Transactions on Signal Processing.
10. D. Simon and M. Elad, **Rethinking the CSC model** for Natural Images, submitted to NIPS.



A New Massive Open Online Course



Courses ▾ Programs ▾ Schools & Partners About ▾

Search:



Sign In

Register



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



Michael Elad
The Computer-Science Department
The Technion

