Welcome to *Sparseland*

Sparse & Redundant Representations and their Applications in Signal and Image Processing

Michael Elad

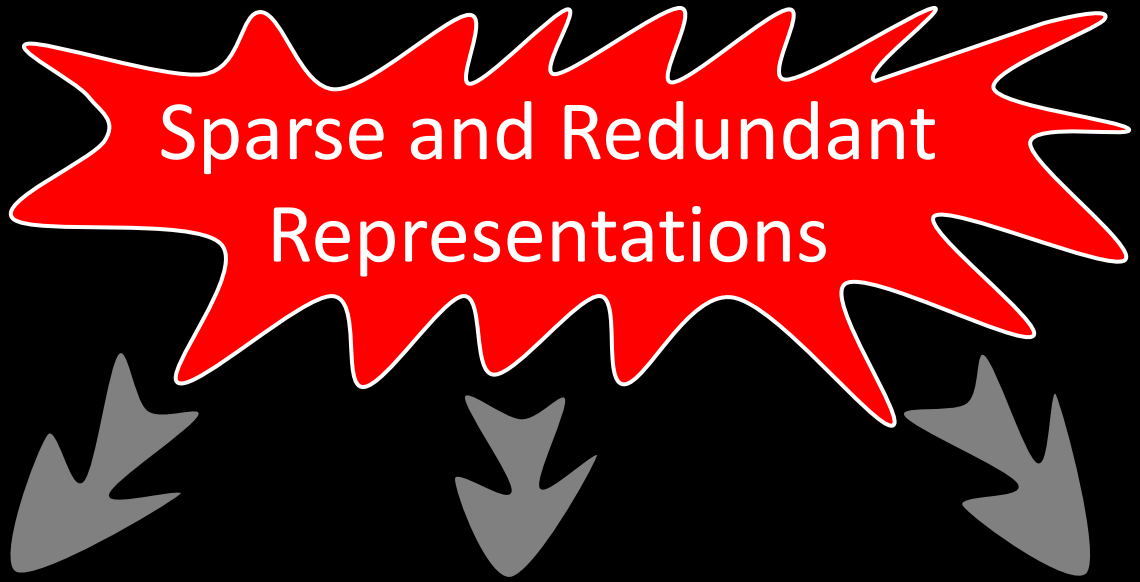
The Computer Science Department
The Technion – Israel Institute of Technology
Haifa 32000, Israel

2017 Summer School on Signal Processing Meets Deep Learning
September 4-8, 2017

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

… will review nearly 20 years of tremendous progress in the field of



- Theory
- Numerical Problems
- Applications (image processing and Machine Learning)

This Lecture

… is Built of Three Parts:

Part 1: A General Introduction to *Sparseland*:
Presenting this Model and its Importance

Part 2: Diving In: Theory and Applications
in Sparse Representations

Part 3: Relation to Deep-Learning:
A Tale of three models: *Sparseland* $\rightarrow$CSC $\rightarrow$CNN

Welcome to *Sparseland*

Part 1: A General Introduction to *Sparseland*

Michael Elad

The Computer Science Department

The Technion – Israel Institute of Technology

Haifa 32000, Israel

2017 Summer School on Signal Processing Meets Deep Learning

September 4-8, 2017

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

What This Field is All About ?

Michael Elad
The Computer-Science Department
The Technion

The Answer is Not Trivial

Depends whom you ask, as the researchers in this field come from various disciplines:

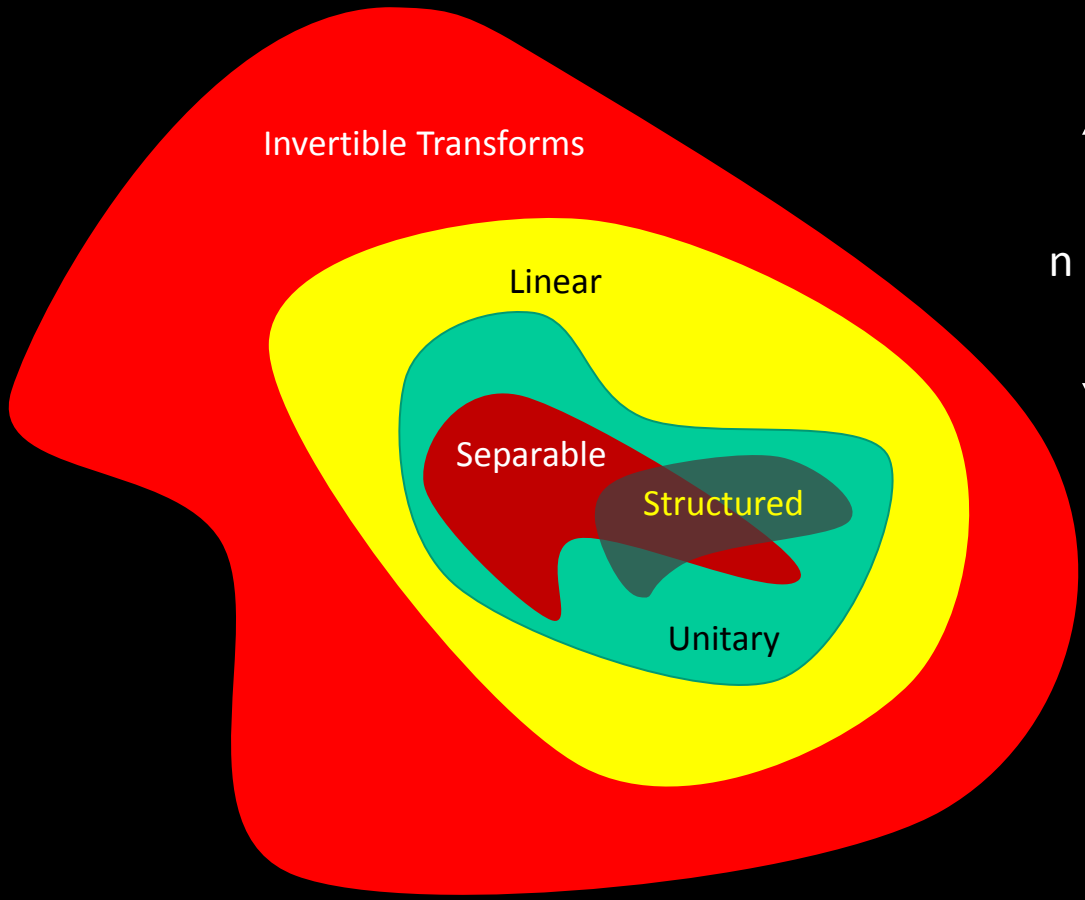
- Mathematics
- Applied Mathematics
- Statistics
- Signal & Image Processing: CS, EE, Bio-medical, …
- Computer-Science Theory
- Machine-Learning
- Physics (optics)
- Geo-Physics
- Physics - Astronomy
- Psychology (neuroscience)
- …

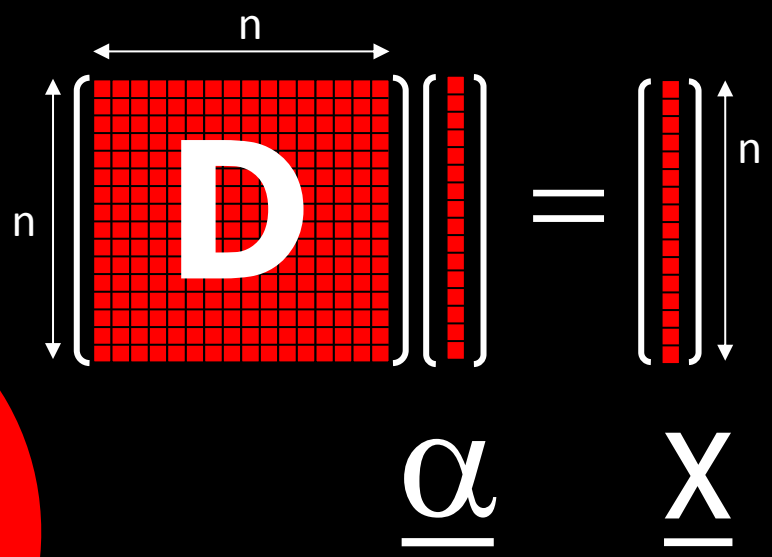
My Answer (For Now)

A New Transform for Signals

- We are all well-aware of the idea of transforming a signal and changing its representation
- We apply a transform to gain something – efficiency, simplicity of the subsequent processing, speed, …
- There are many known transforms: Fourier, DCT, Hadamard, Wavelets and its descendants, and more
- Our message: There is a new transform in town, based on sparse and redundant representations

Transforms – The General Picture

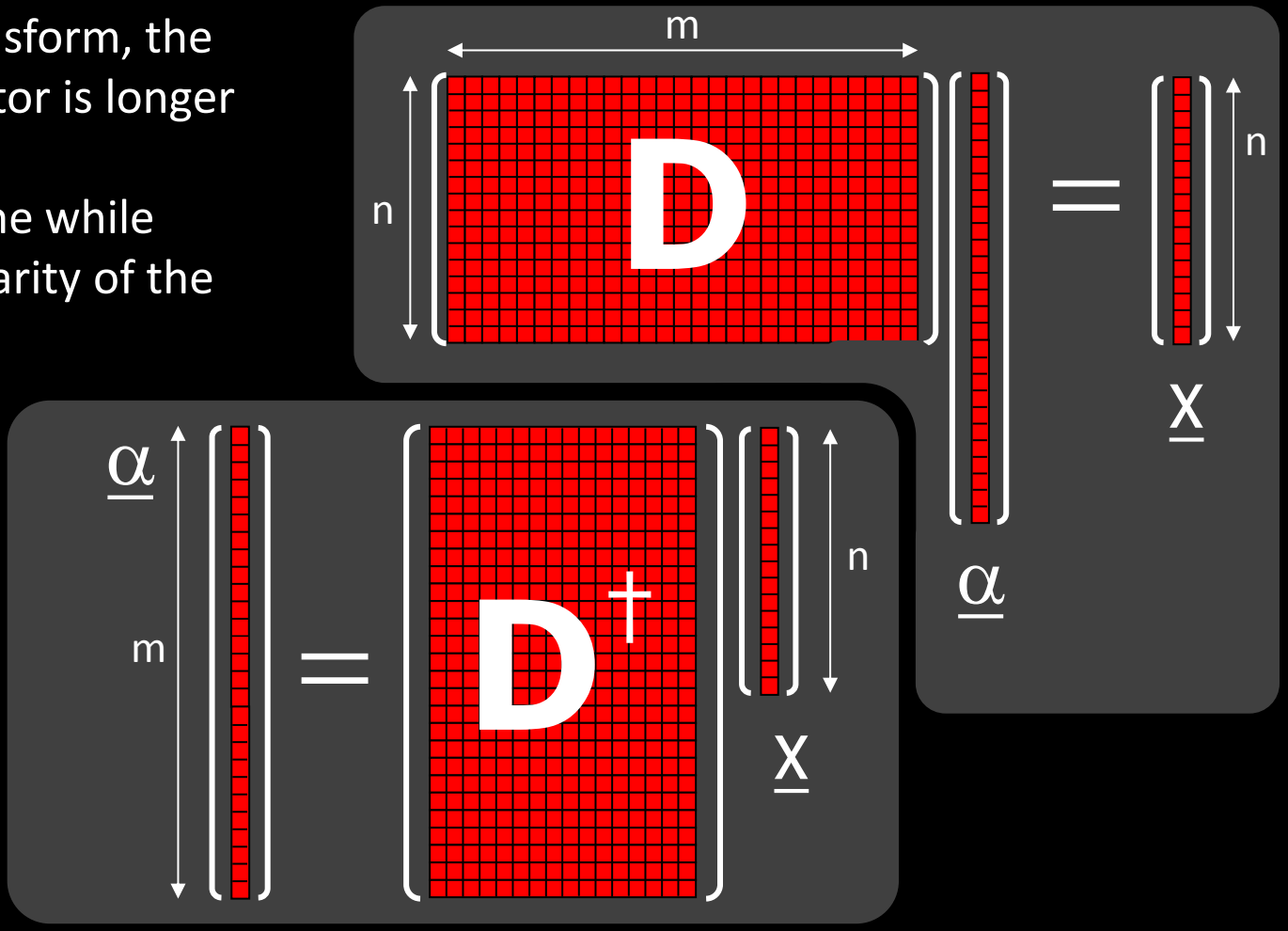




Redundancy?

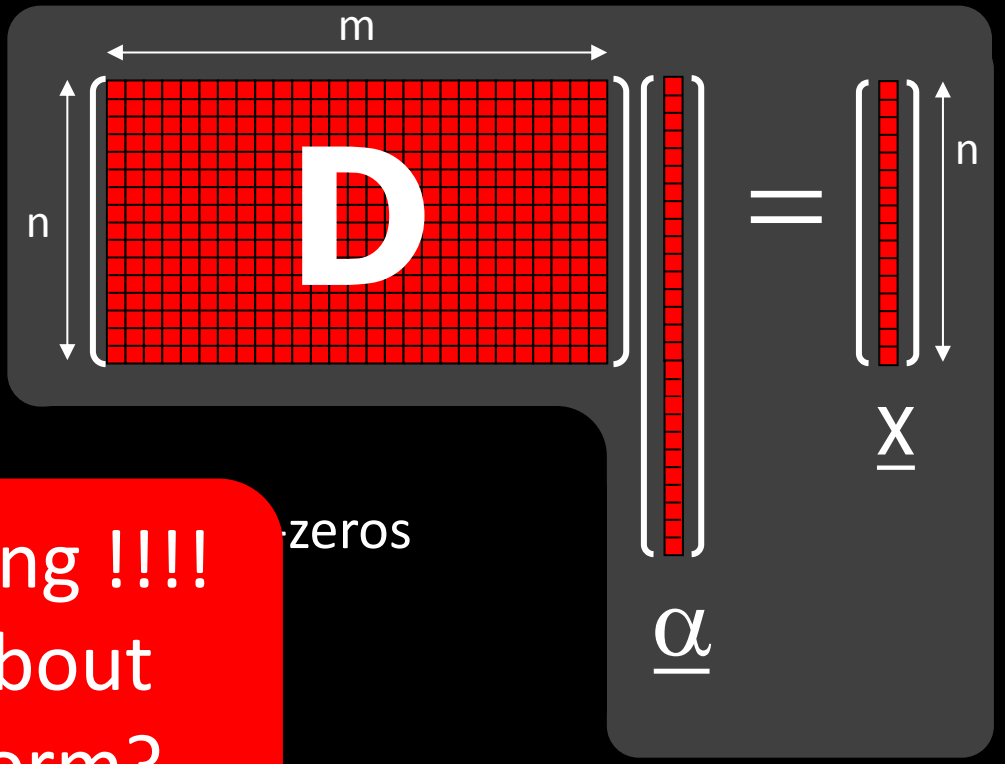
- In a redundant transform, the representation vector is longer (m>n)
- This can still be done while preserving the linearity of the transform:

$$\underline{x} = \mathbf{D}\underline{\alpha}$$
$$= \mathbf{D}\mathbf{D}^{\dagger}\underline{x}$$
$$= \mathbf{I}$$
$$= \underline{x}$$



Sparse & Redundant Representation

- We shall keep the linearity of the inverse-transform
- As for the forward (computing $\underline{\alpha}$ from $\underline{x}$), there are infinitely many possible solutions
- We shall seek the sparsest of all soluti… …zeros
- This mal… non-line…
- The field… …cations is all about defining clearly this transform, solving various theoretical and numerical issues related to it, and showing how to use it in practice



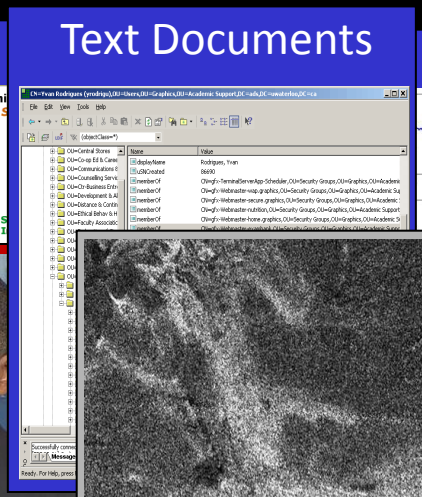
Sounds … Boring !!!!
Who cares about a new transform?

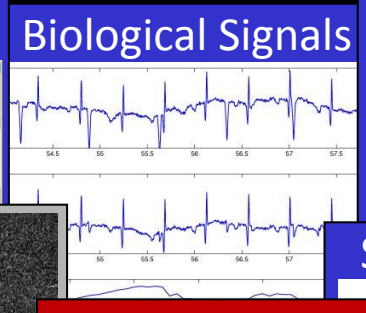
What This Field is All About ?

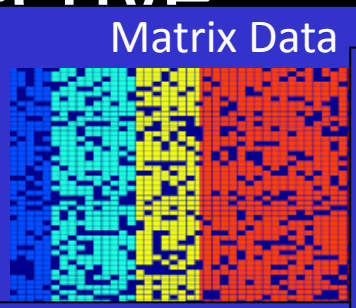
Take 2

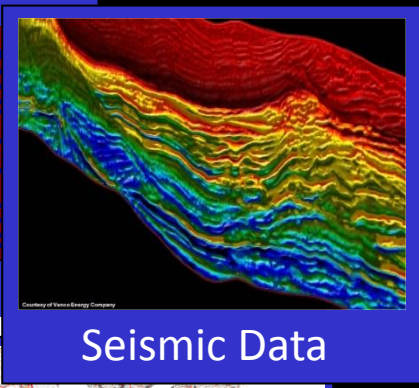
Michael Elad
The Computer-Science Department
The Technion

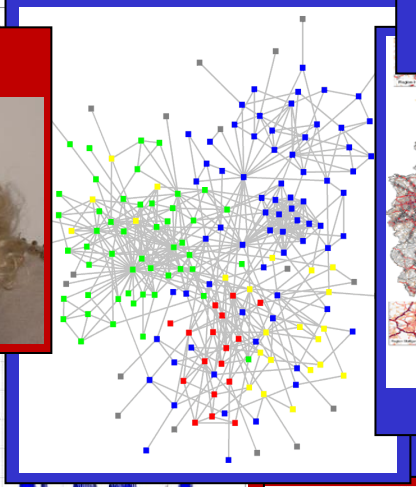
Lets Take a Wider Perspective

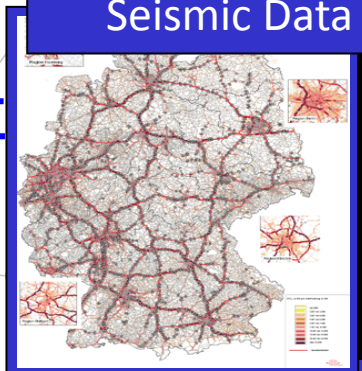


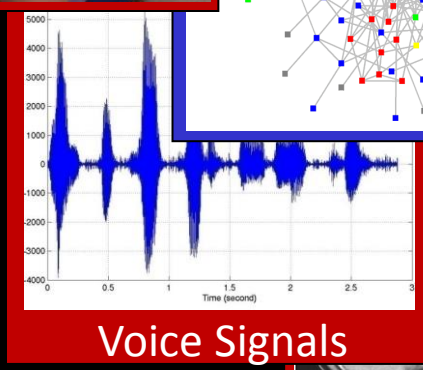


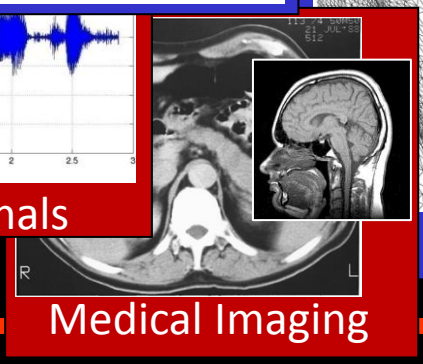


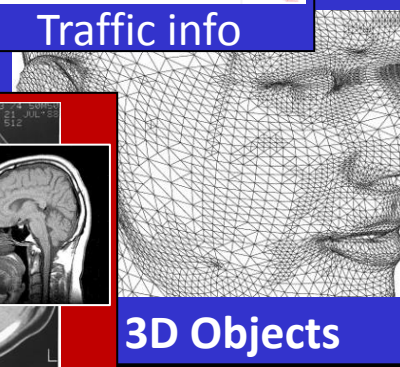






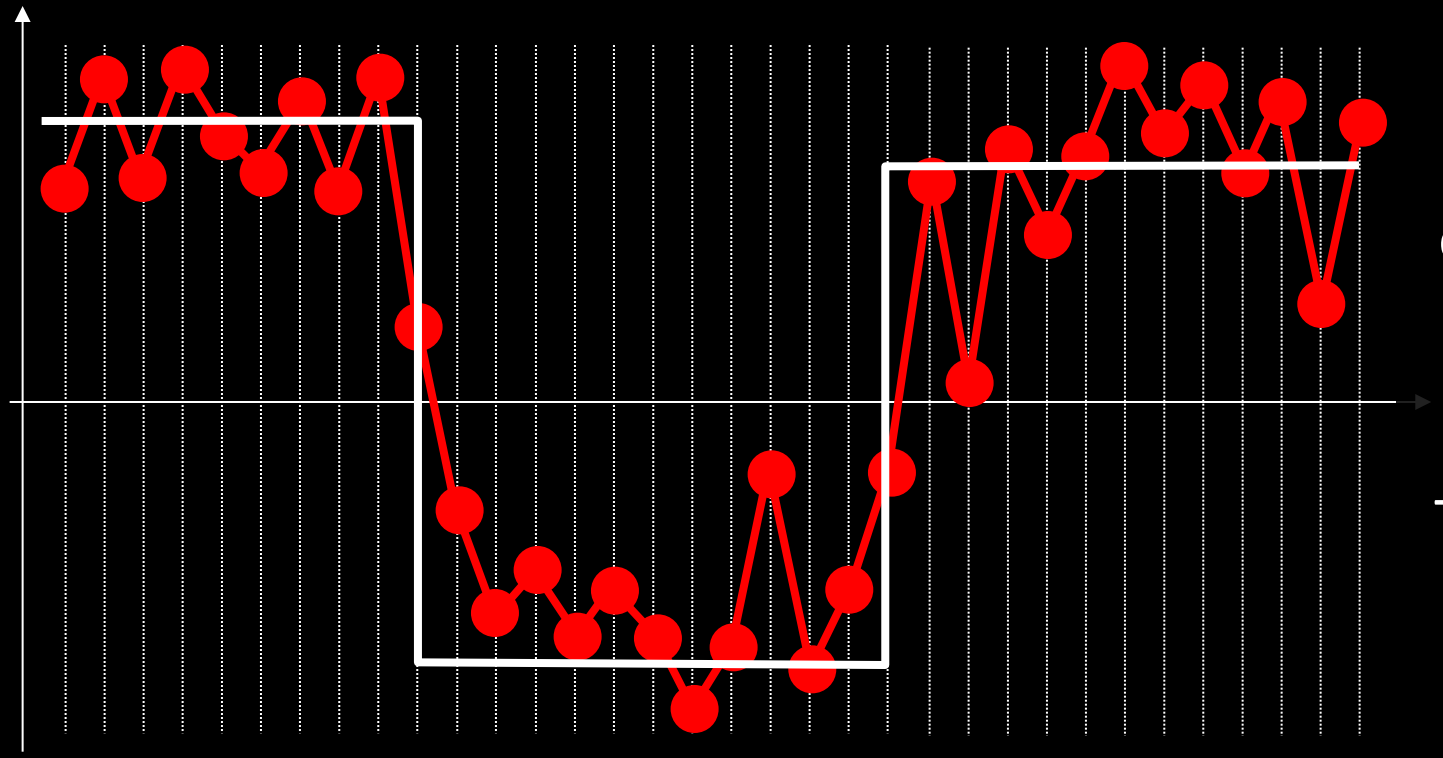






- We are surrounded by various sources of massive information
- All these sources have some internal structure, which can be exploited
- This structure, when identified, is the engine behind our ability to process this data

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

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

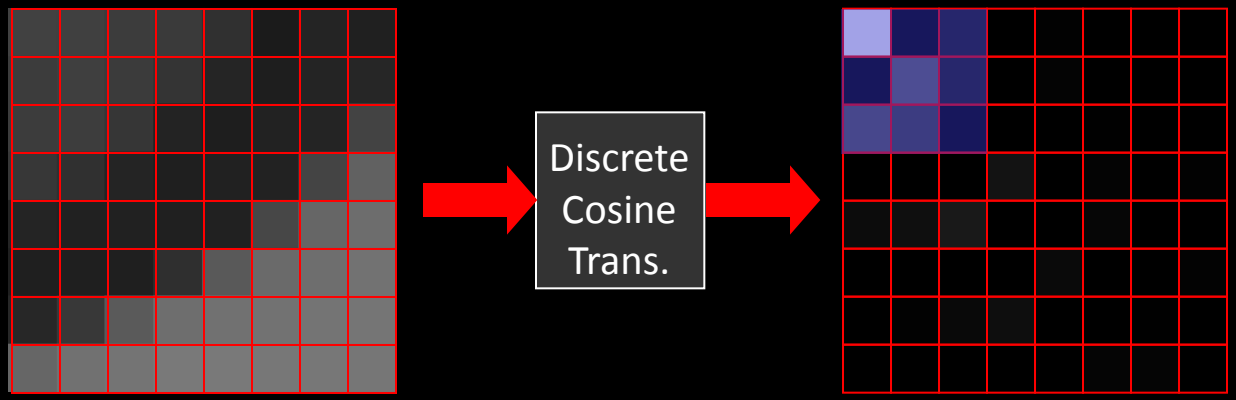
Besov-Spaces

Total-Variation

Beltrami-Flow

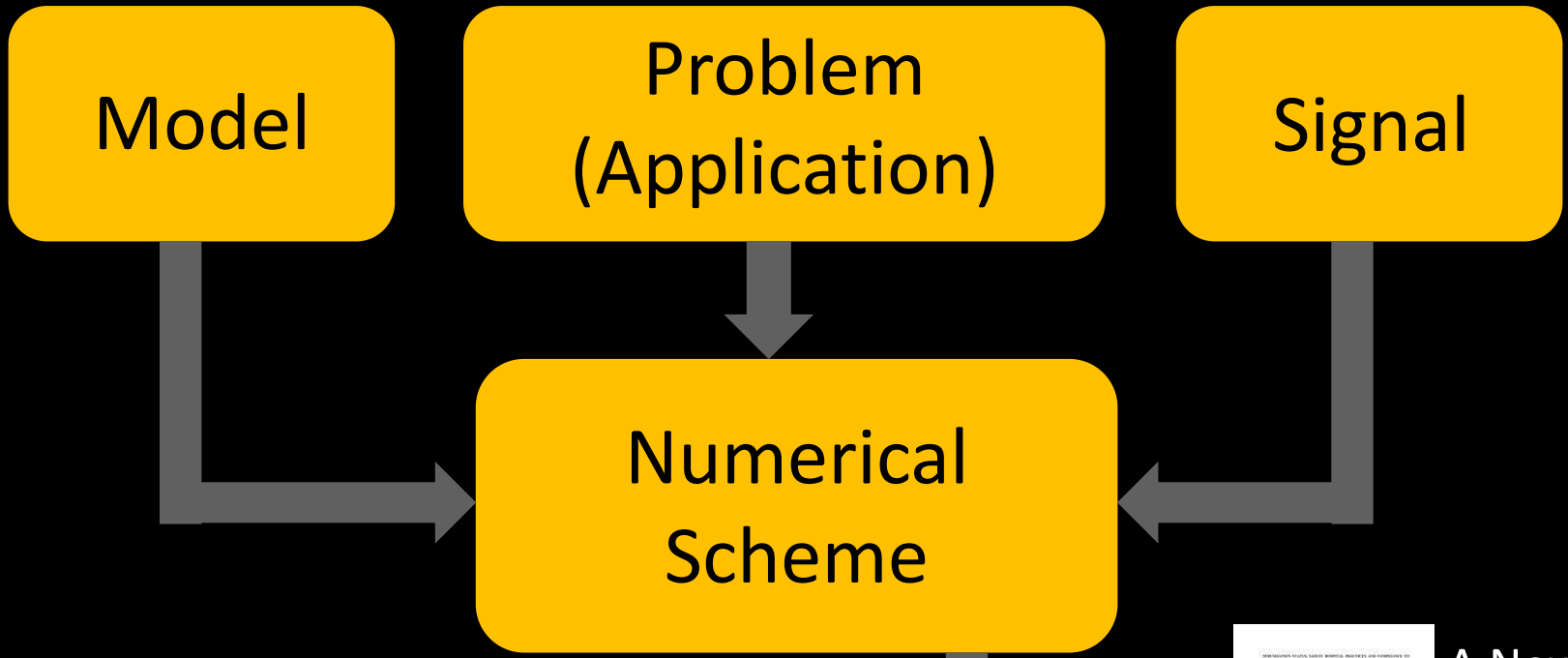
An Example: JPEG and DCT

How & why does it works?

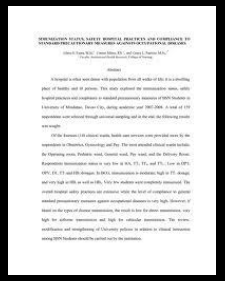


The model assumption: after DCT, the top left coefficients to be dominant and the rest zeros

Research in Signal/Image Processing



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



A New Research Work (and Paper) is Born

Again: What This Field is all About?

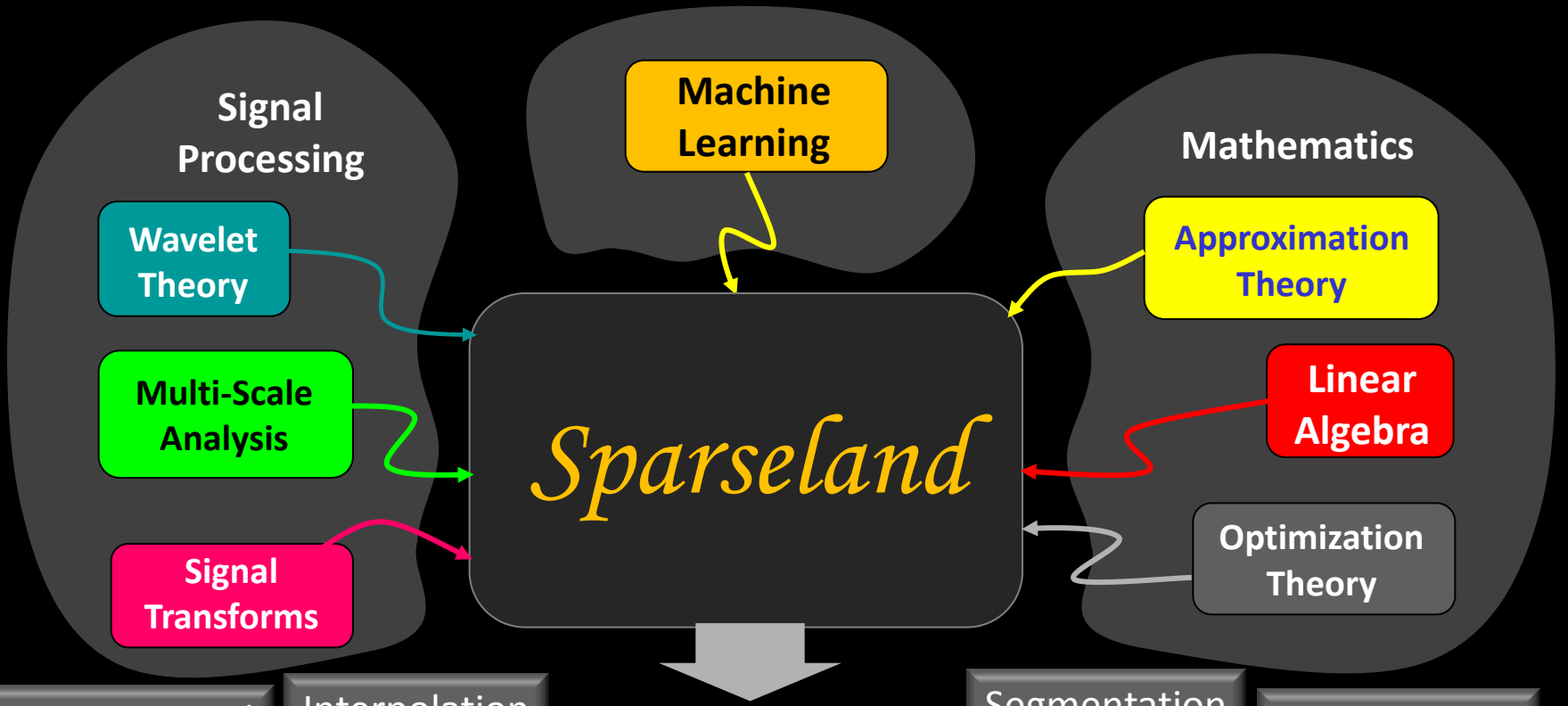
A Data Model and its Use

- Almost any task in data processing requires a model – true for denoising, deblurring, super-resolution, inpainting, compression, anomaly-detection, sampling, recognition, …
- There is a new model in town – sparse and redundant representation – we will call it

Sparseland

- This is a **flexible model** that can adjust to the signal

A New Emerging Model



A Closer Look at the *Sparseland* Model

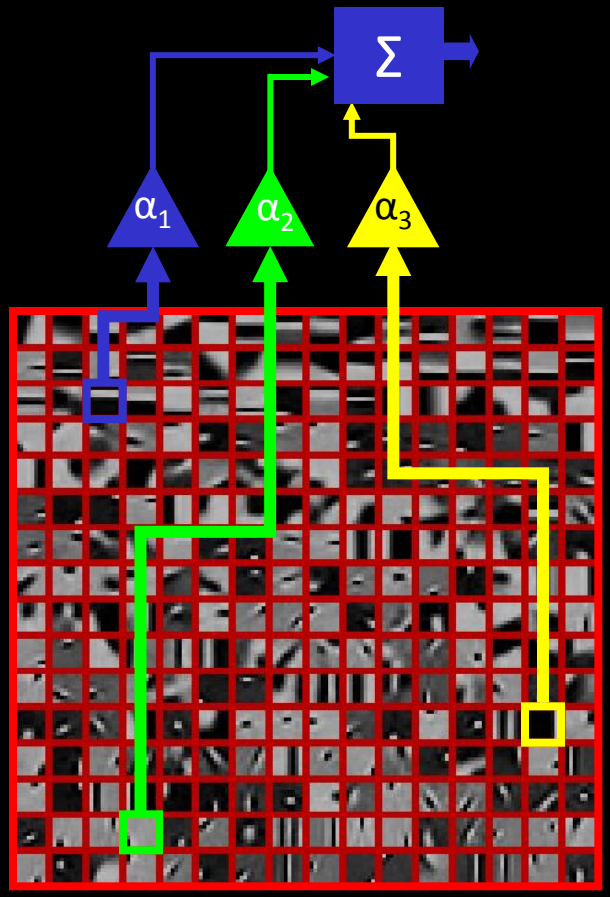
Michael Elad
The Computer-Science Department
The Technion

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

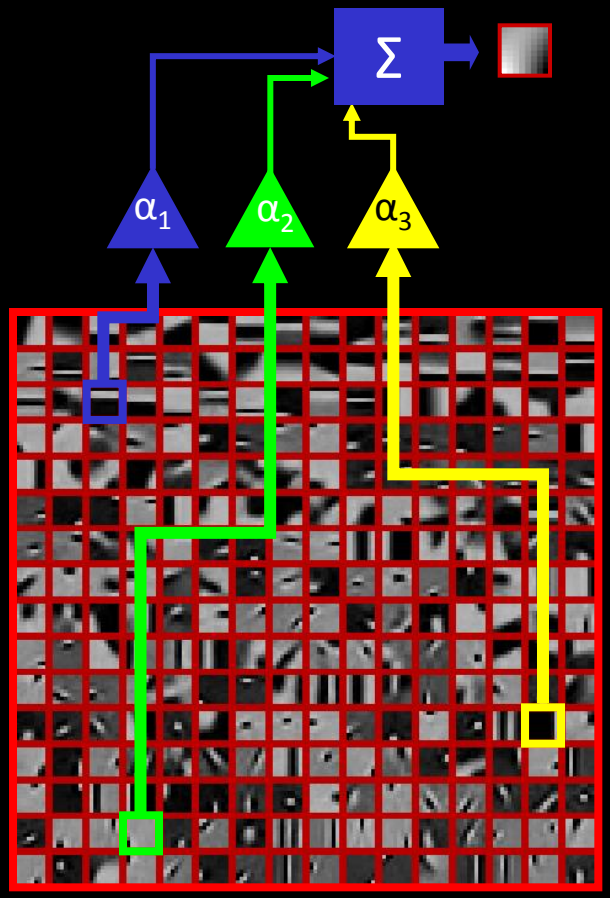


The *Sparseland* Model

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)

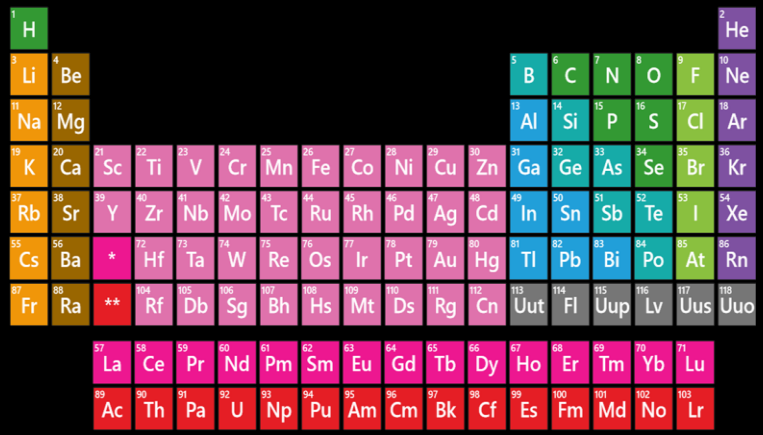


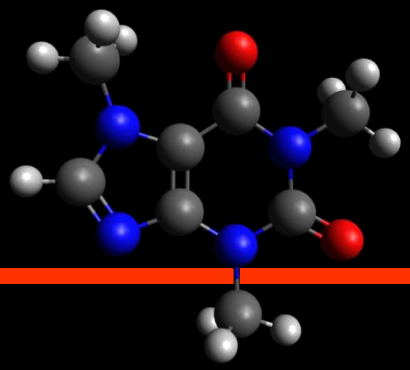
Chemistry of Data

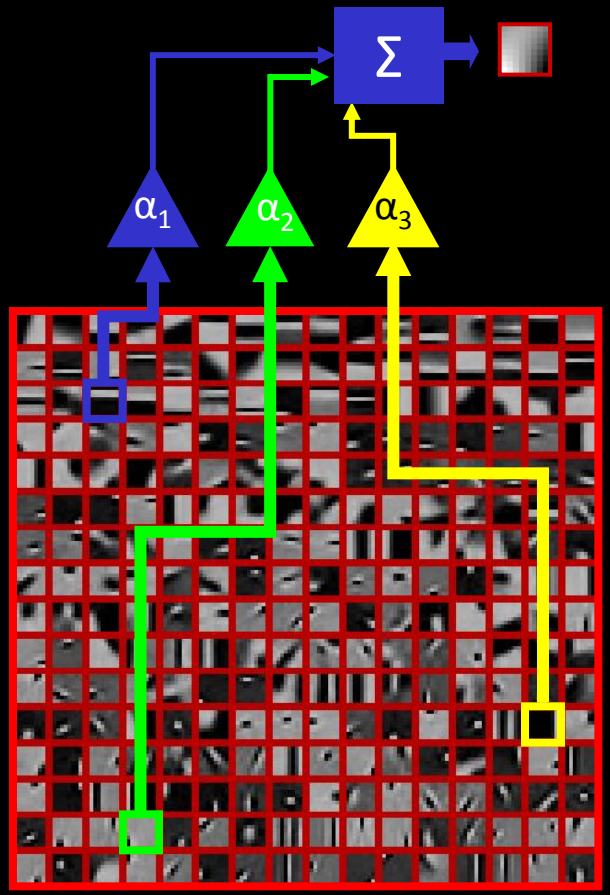
We could refer to the *Sparseland* model as the chemistry of information:

- Our dictionary stands for the **Periodic Table** containing all the elements
- Our model follows a similar rationale:

  Every molecule is built of **few** elements

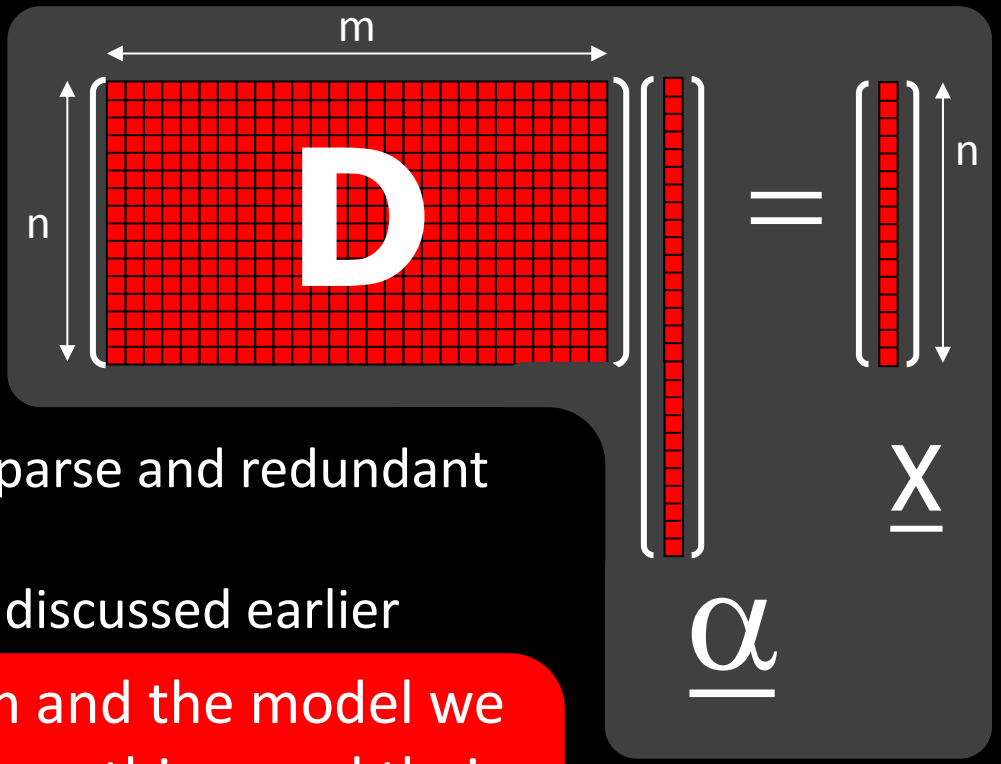






Model vs. Transform ?

- The relation between the signal $\underline{x}$ and its representation $\underline{\alpha}$ is the following linear system, just as described earlier
- We shall be interested in seeking sparse solutions to this system when deploying the sparse and redundant representation model
- This is EXACTLY the transform we discussed earlier



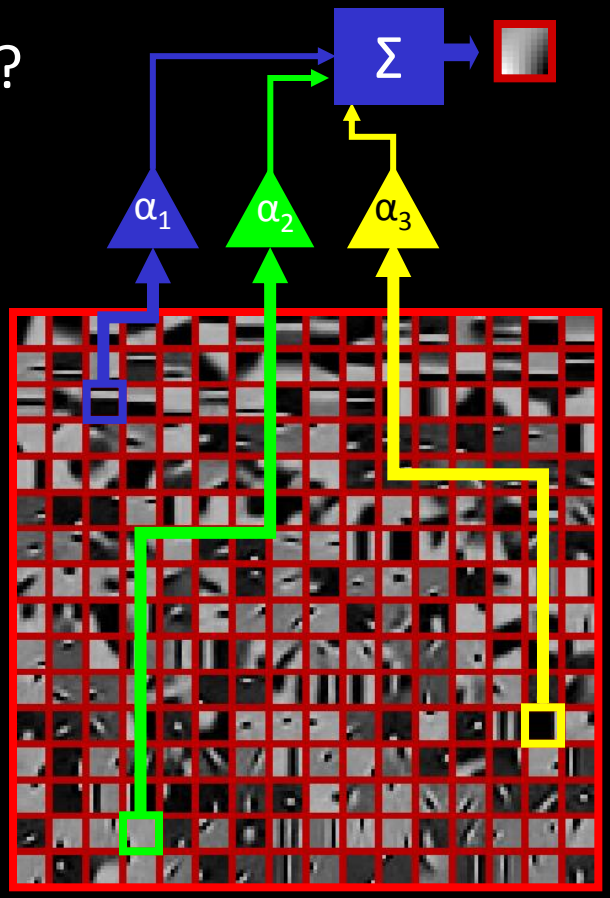
Bottom Line: The transform and the model we described above are the same thing, and their impact on signal/image processing is profound and worth studying

Difficulties with *Sparseland*

- Problem 1: Given an image patch, 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

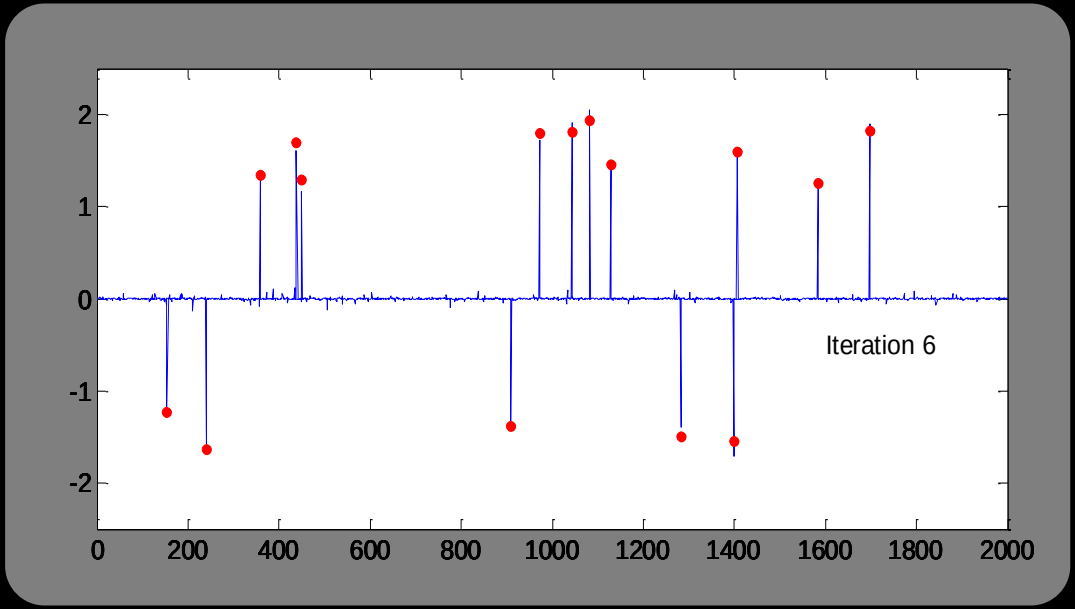
$$\binom{2000}{15} \approx 2.4e+37$$ possibilities

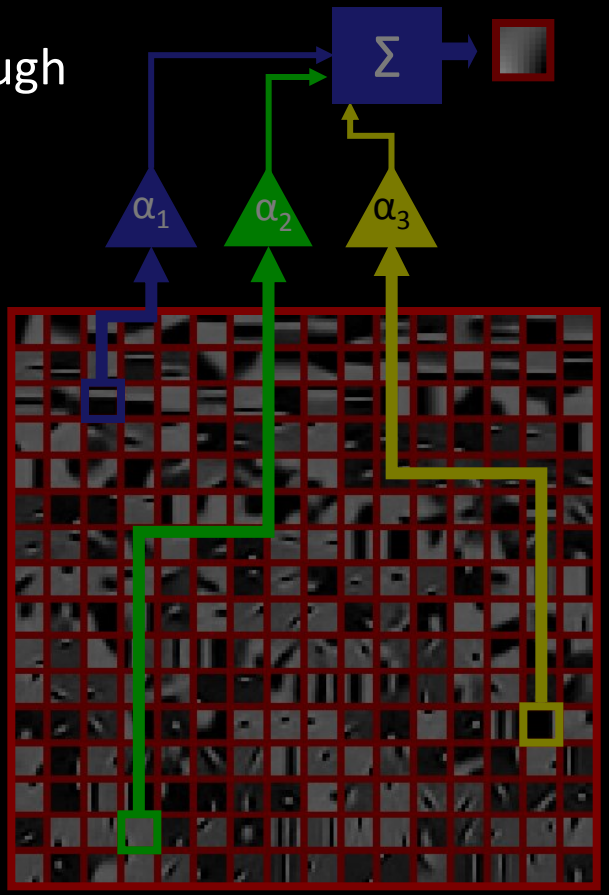
  - If each of these takes 1nano-sec to test, will take ~7.5e20 years to finish !!!!!!
- Solution: Approximation algorithms



Difficulties with *Sparseland*

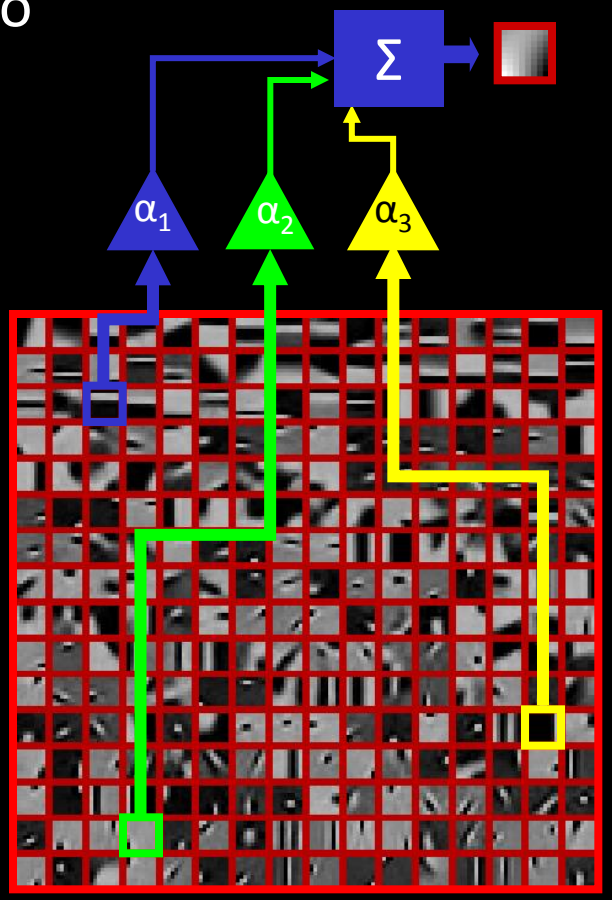
- Various algorithms exist. Their theoretical analysis guarantees their success if the solution is sparse enough
- Here is an example – the Iterative Reweighted LS:





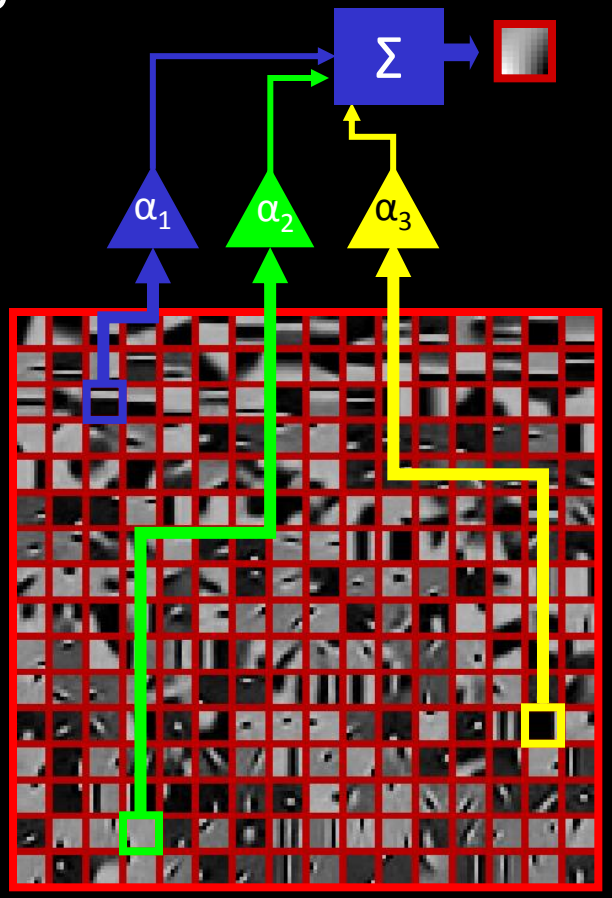
Difficulties with *Sparseland*

- 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
- Many such algorithms were developed in the past 10 years (e.g., K-SVD), and their performance is surprisingly good
- This is only the beginning of a new era in signal processing …



Difficulties with *Sparseland*

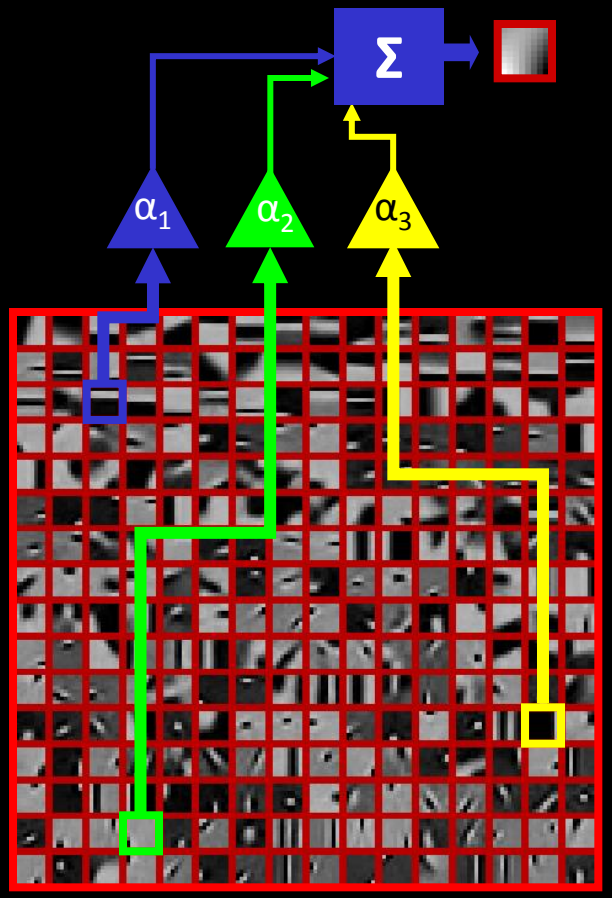
- Problem 3: Is this model flexible enough to describe various sources? e.g., Is it good for images? Audio? Stocks? …
- General answer: Yes, this model is extremely effective in representing various sources
  - Theoretical answer: Relation to other known models
  - Empirical answer: we will see in this course, several image processing applications, where this model leads to the best known results (benchmark tests)



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



Who Works on This ?

Michael Elad
The Computer-Science Department
The Technion

Who is Working in This Field ?


Donoho, Candes – Stanford


Tropp – CalTech


Baraniuk, W. Yin – Rice Texas


Gilbert, Vershynin, Plan– U-Michigan


Gribonval, Fuchs – INRIA France


Starck – CEA – France


Vandergheynst – EPFL Swiss


Rao, Delgado – UC San-Diego


Do, Bresler – U-Illinois


Davies – Edinburgh UK


Elad, Zibulevsky, Bruckstein, Eldar, Segev, Mendelson – Technion


Goyal – MIT


Mallat – ENS Paris


Nowak, Willet – Wisconsin


Coifman – Yale


Romberg – GaTech


Lustig, Wainwright – Berkeley


Sapiro, Daubachies – Duke


Friedlander – UBC Canada


Tarokh – Harvard


Cohen, Combettes – Paris VI

David L. Donoho

- An extremely talented mathematician and statistician from the Stanford Statistics Department
- He is among the few who founded this field sparse and redundant representations and its spin-off topic of compressed sensing
- In 2013 he won the Shaw prize ("the Nobel of the east")

THE SHAW PRIZE 邵逸夫獎

Announcement and Citation

Announcement

The Shaw Prize in Mathematical Sciences 2013

is awarded to

David L Donoho

for his profound contributions to modern mathematical statistics and in particular the development of optimal algorithms for statistical estimation in the presence of noise and of efficient techniques for sparse representation and recovery in large data-sets.

28 May 2013 Hong Kong

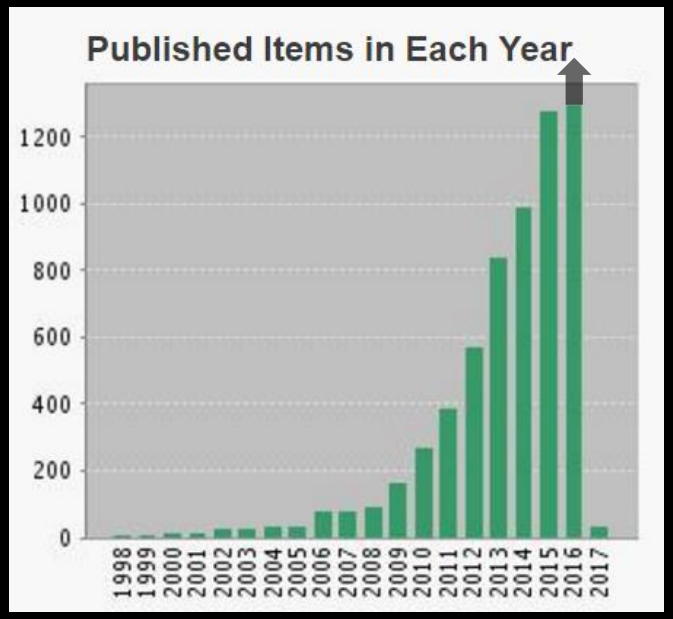
This Field is rapidly GROWING …

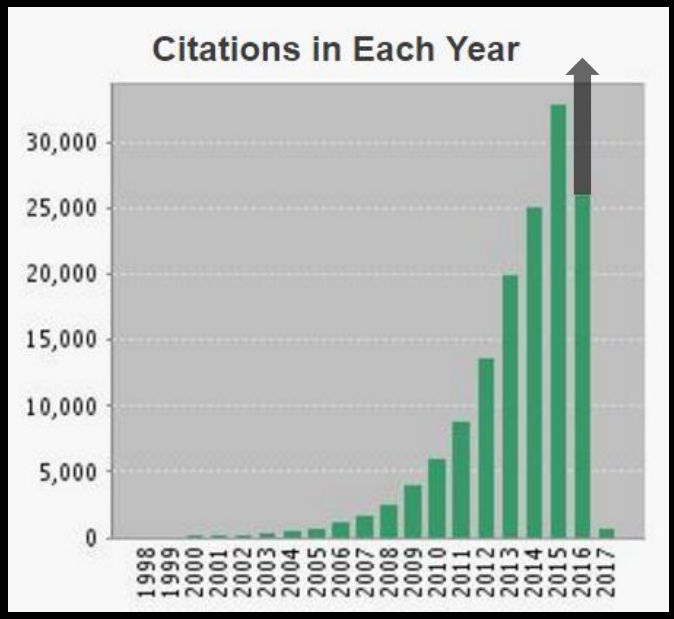
- Searching ISI-Web-of-Science (December 26th 2016):

  Topic= ((spars* and (represent* or approx* or solution or estimation) and (dictionary or pursuit or convex)) or (compres* and sens* and spars*))

  led to 6354 papers (it was ~4000 papers a year ago)

- Here is how they spread over time (with ~146178 citations):





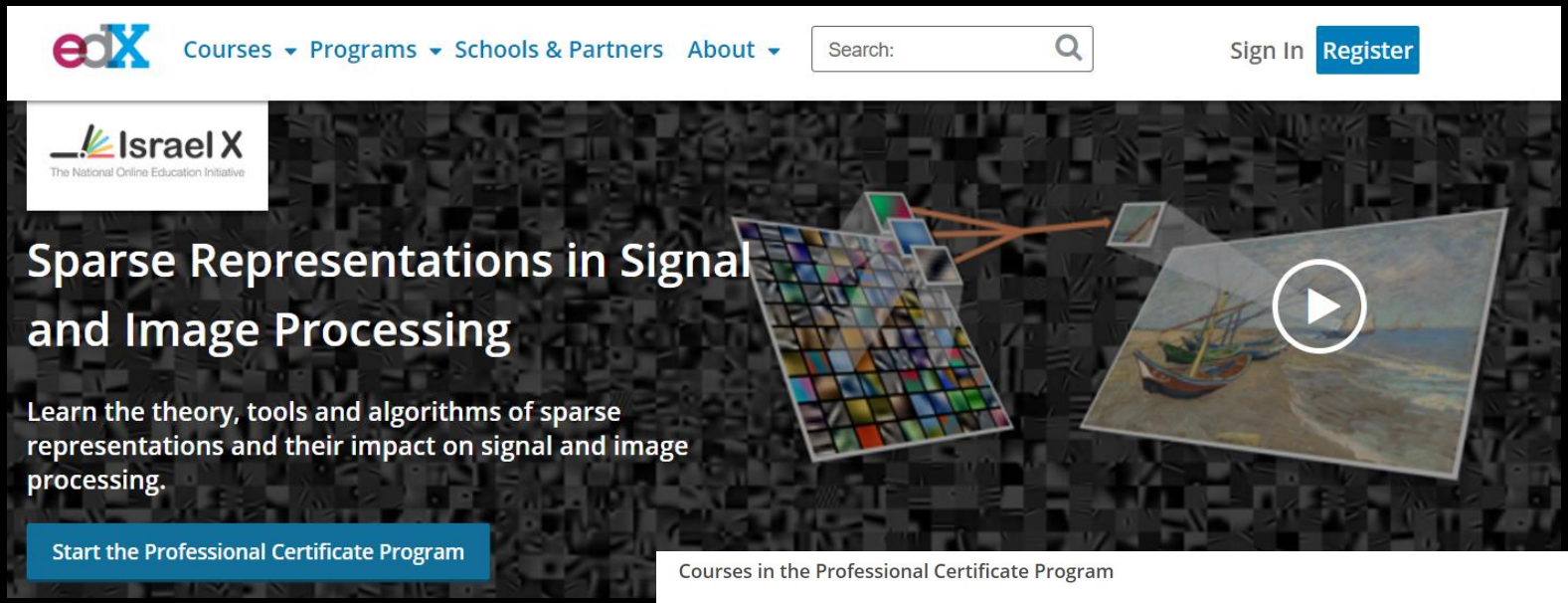
Which Countries?

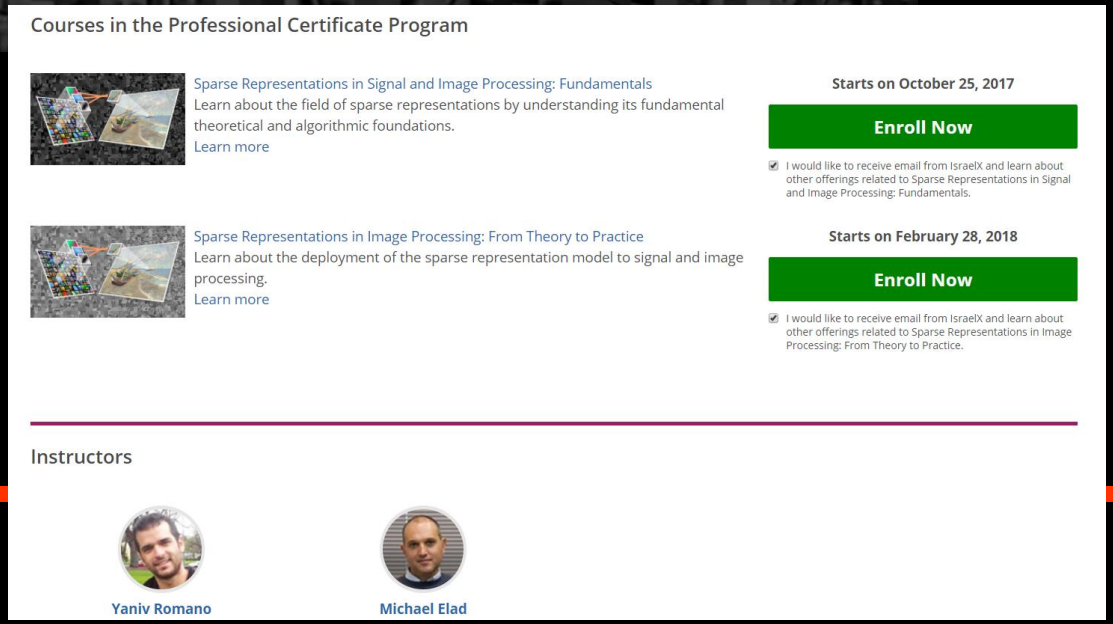
Field: Countries/Territories	Record Count	% of 6354	Bar Chart
PEOPLES R CHINA	2573	40.494 %	
USA	2171	34.167 %	
FRANCE	416	6.547 %	
ENGLAND	286	4.501 %	
CANADA	272	4.281 %	
GERMANY	257	4.045 %	
ISRAEL	189	2.975 %	
AUSTRALIA	170	2.675 %	
SOUTH KOREA	160	2.518 %	
ITALY	154	2.424 %	
SWITZERLAND	144	2.266 %	
JAPAN	141	2.219 %	
SINGAPORE	137	2.156 %	
INDIA	115	1.810 %	
IRAN	115	1.810 %	
SPAIN	84	1.322 %	
BELGIUM	73	1.149 %	
SCOTLAND	72	1.133 %	
TURKEY	72	1.133 %	
TAIWAN	71	1.117 %	
SWEDEN	64	1.007 %	
AUSTRIA	58	0.913 %	
DENMARK	55	0.866 %	
NETHERLANDS	55	0.866 %	
FINLAND	51	0.803 %	

Books in this field

- The following book was published in 2010, and it has served as the textbook for my advanced course
- Since then, it has been adopted by other courses worldwide (Stanford, Duke, Oxford, IISc, UCSD, …)
- In the past 8-9 years, many other books were published in this and closely related fields

A Massive Open Online Course : Coming Up



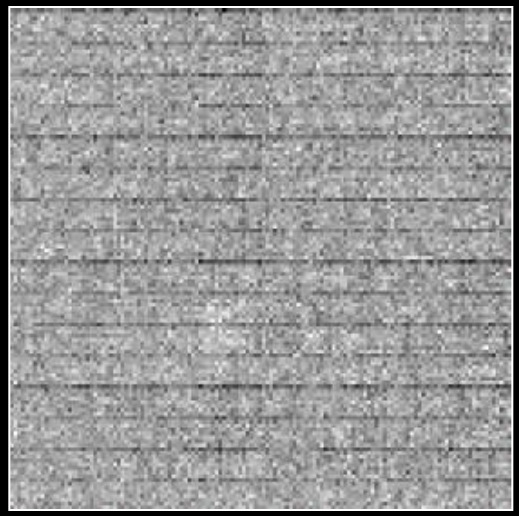


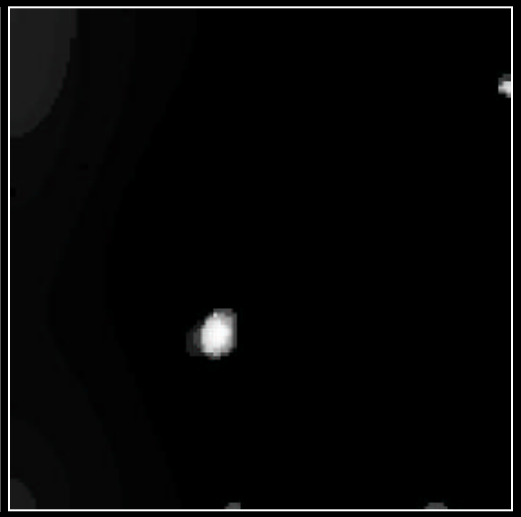
Several Examples: Applications Leveraging this Model

Michael Elad
The Computer-Science Department
The Technion

Image Separation [Starck, Elad, & Donoho (`04)]

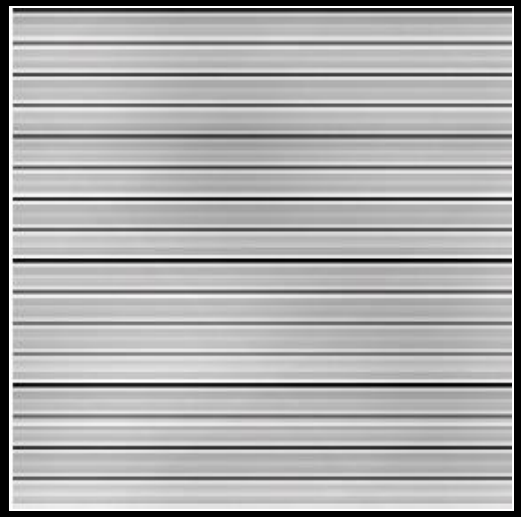
The original image - Galaxy SBS 0335-052 as photographed by Gemini

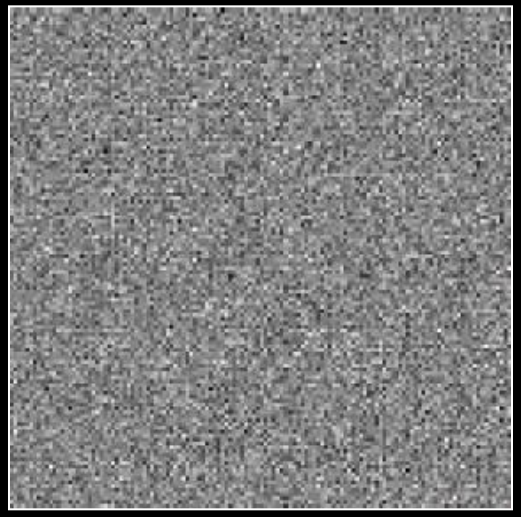




The Cartoon part spanned by wavelets

The texture part spanned by global DCT

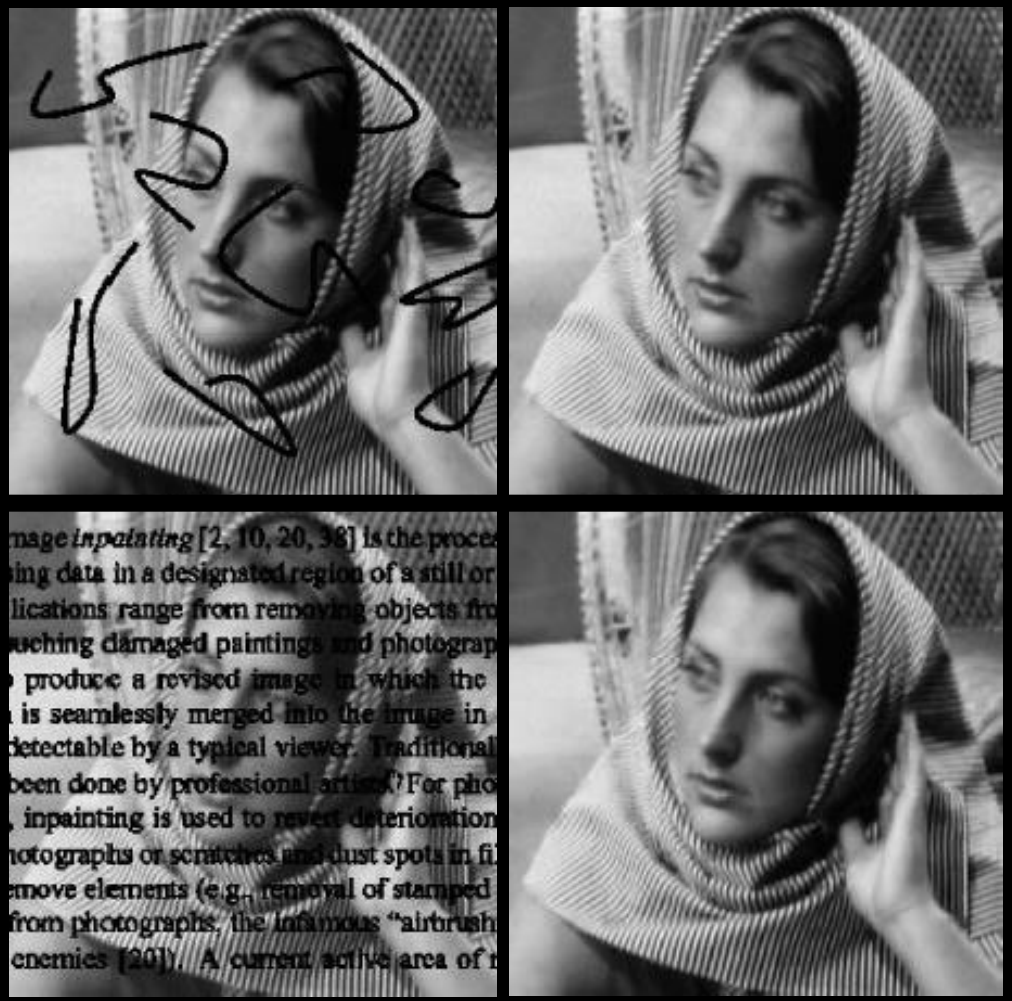




The residual being additive noise

Inpainting [Starck, Elad, and Donoho ('05)]

Source



Outcome

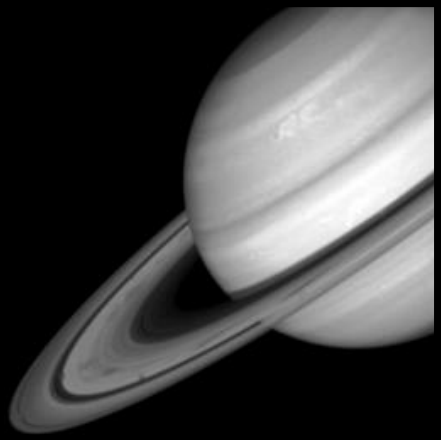
Denoising [Mairal, Elad & Sapiro, ('06)]

Original

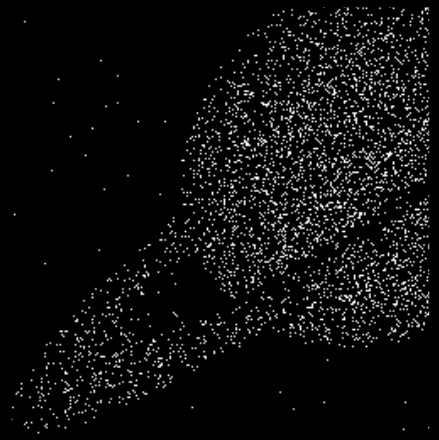
Noisy (20.43dB)

Result (30.75dB)

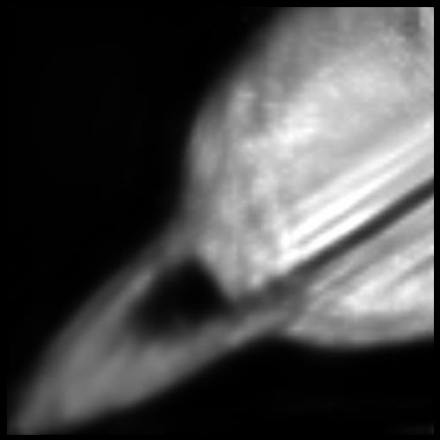
Poisson Denoising [Giryes & Elad ('14)]



Original



Noisy (peak=0.2)



Result (PSNR=24.16dB)

Original

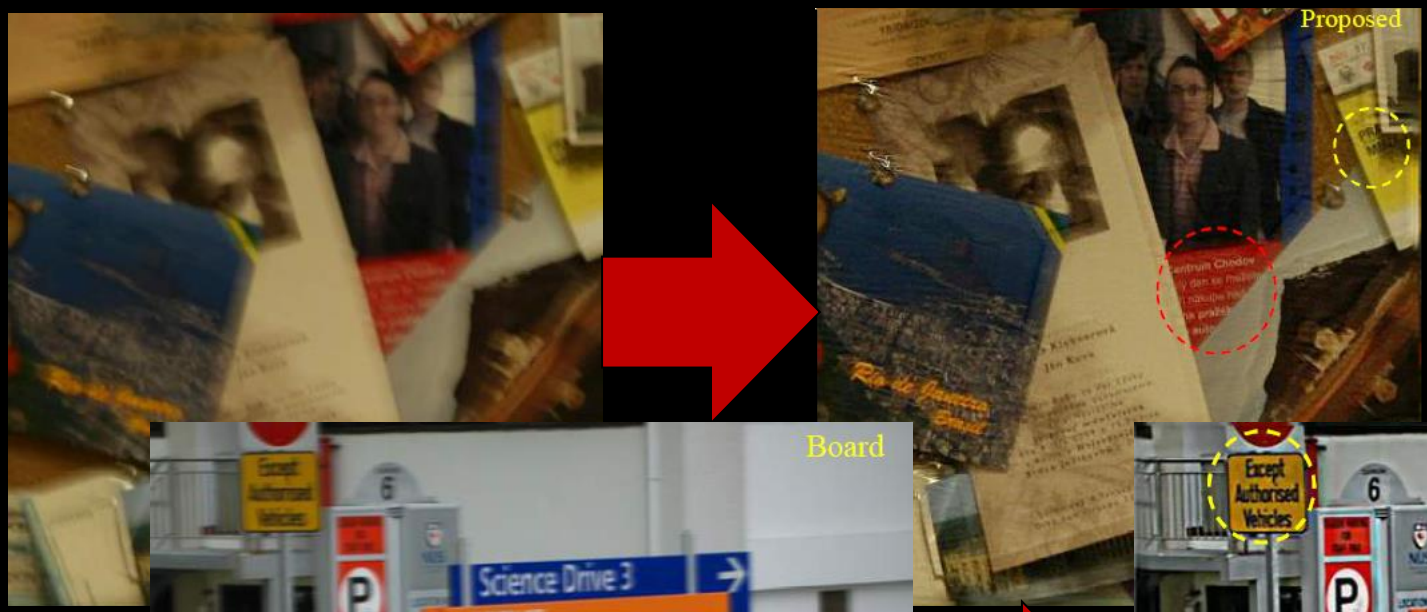
Noisy (peak=2)

Result (PSNR=24.76dB)

Deblurring [Elad, Zibulevsky and Matalon, ('07)]

original Measured Restored (right): ISNR=7.0322 dB

Blind Deblurring [Shao and Elad ('14)]



Inpainting (Again!) [Mairal, Elad & Sapiro, (‘06)]

To Summarize

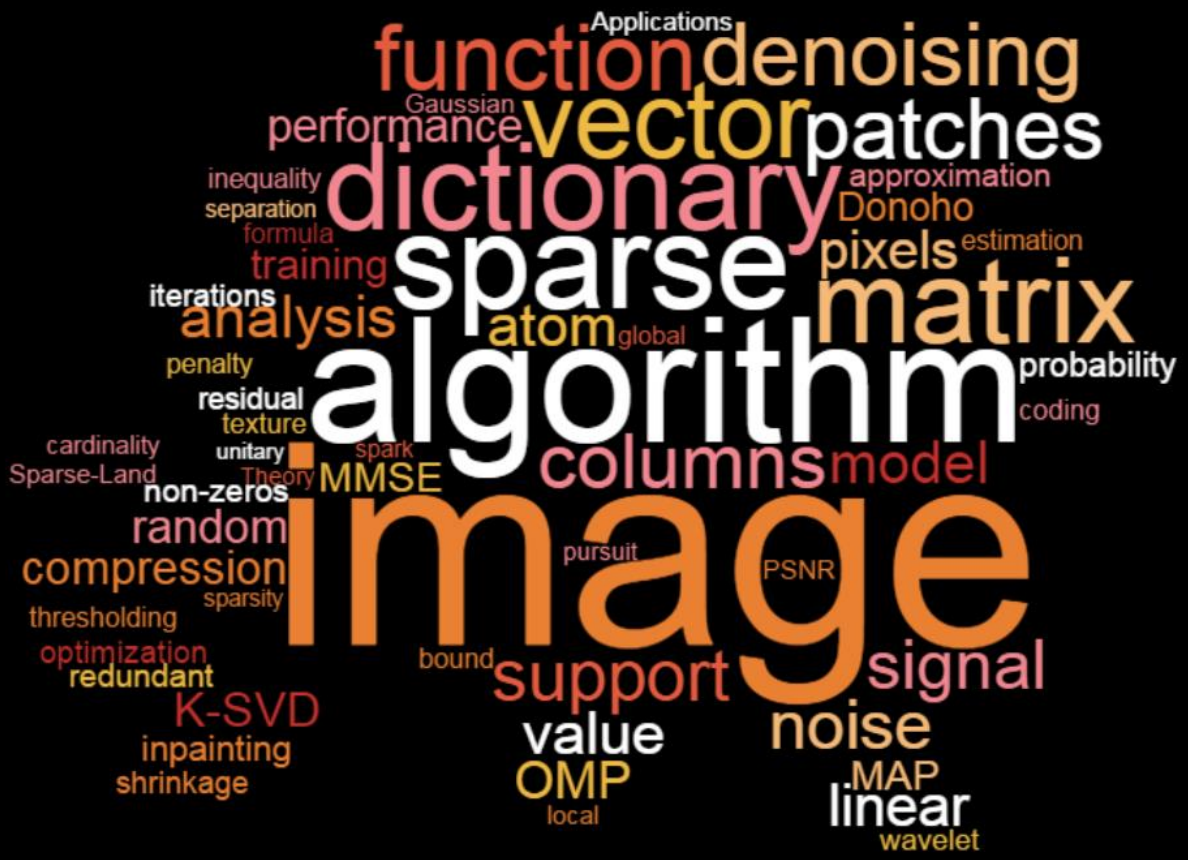
An effective (yet simple) model for signals/images is key in getting better algorithms for various applications

Which model to choose?

Sparse and redundant representations & trained dictionaries are drawing a considerable attention in recent years, due to the elegant theory and the impressive applications

So, what next?

Lets take a closer look at this model, clarify what we mean by the theory behind it, and show several of the algorithms it leads to for handling image processing tasks



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

Welcome to *Sparseland*

Part 2: Diving In

Michael Elad

The Computer Science Department

The Technion – Israel Institute of Technology

Haifa 32000, Israel

2017 Summer School on Signal Processing Meets Deep Learning

September 4-8, 2017

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

Agenda

Part I – Denoising by Sparse & Redundant Representations

Part II – Theoretical & Numerical Foundations

Part III – Dictionary Learning & The K-SVD Algorithm

Part IV – Back to Denoising … and Beyond – handling stills and video denoising & inpainting, demosaicing, super-res., and compression

Part V – Summary & Conclusions

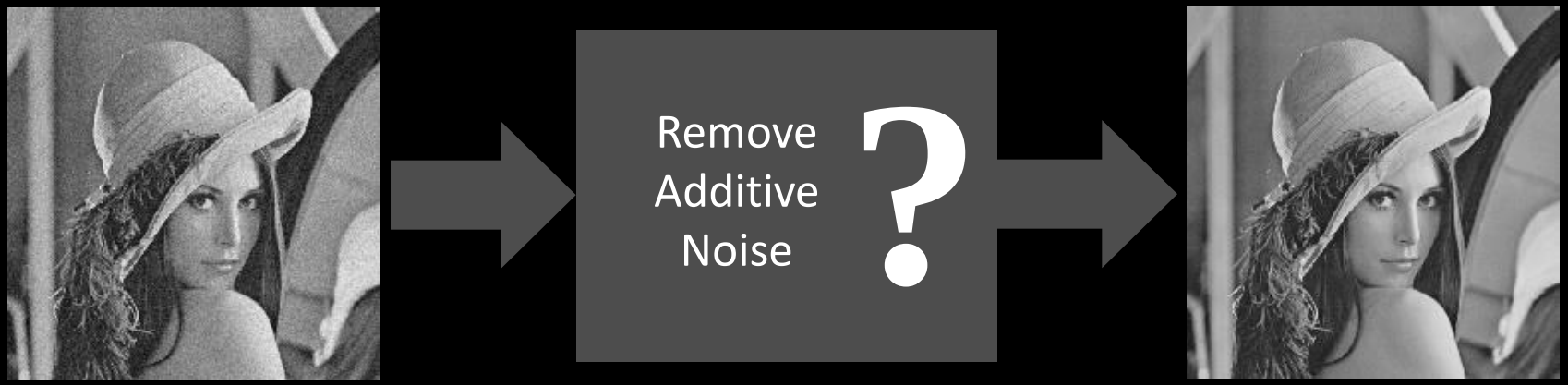
In this part we will show that

- Sparsity and Redundancy are valuable and well-founded tools for modeling data.
- When used in image processing, they lead to state-of-the-art results.

Denoising by Sparse & Redundant Representations

Noise Removal?

Our story begins with image denoising …



- Important: (i) Practical application; (ii) A convenient platform (being the simplest inverse problem) for testing basic ideas in image processing, and (iii) If you can denoise, you can do much more – see Peyman's talk
- Many Considered Directions: Partial differential equations, Statistical estimators, Adaptive filters, Inverse problems & regularization, Wavelets, Example-based techniques, Sparse representations, …

Denoising by Energy Minimization

Many of the proposed image denoising algorithms are related to the minimization of an energy function of the form

$$f(\underline{x}) = \frac{1}{2}\|\underline{x} - \underline{y}\|_2^2 + G(\underline{x})$$

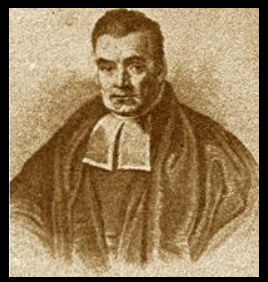
Relation to measurements

Prior or regularization

$\underline{y}$: Given measurements

$\underline{x}$: Unknown to be recovered

- This is in-fact a Bayesian point of view, adopting the Maximum-A-posteriori Probability (MAP) estimation.
- Clearly, the wisdom in such an approach is within the choice of the prior – modeling the images of interest.



Thomas Bayes
1702 - 1761

The Evolution of G(x)

During the past several decades we have made all sort of guesses about the prior G(x) for images:

$$G(\underline{x}) = \lambda \|\underline{x}\|_2^2$$

Energy

$$G(\underline{x}) = \lambda \|\mathbf{L}\underline{x}\|_2^2$$

Smoothness

$$G(\underline{x}) = \lambda \|\mathbf{L}\underline{x}\|_{\mathbf{W}}^2$$

Adapt+ Smooth

$$G(\underline{x}) = \lambda \rho\{\mathbf{L}\underline{x}\}$$

Robust Statistics

$$G(\underline{x}) = \lambda \|\nabla \underline{x}\|_1$$

Total-Variation

$$G(\underline{x}) = \lambda \|\mathbf{W}\underline{x}\|_1$$

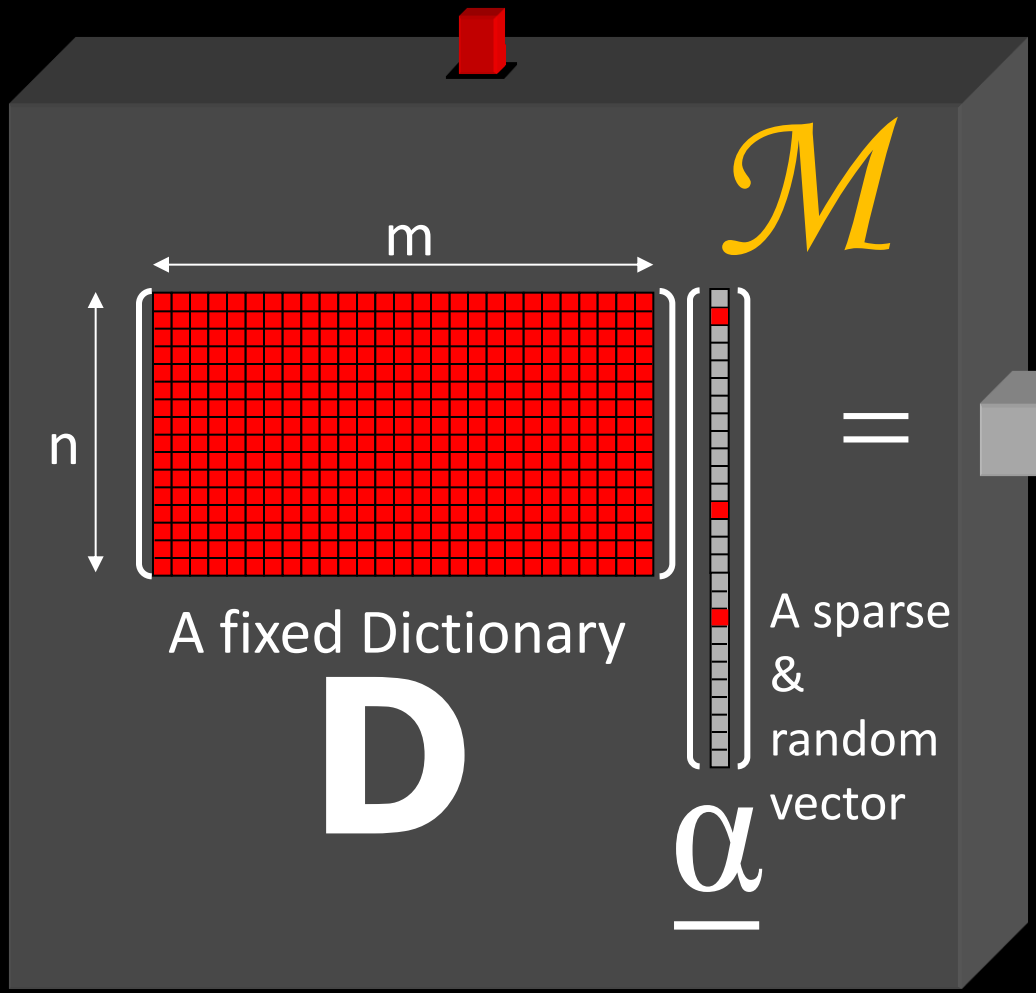
Wavelet Sparsity

$$G(\underline{x}) = \lambda \|\underline{\alpha}\|_0 \quad \text{for} \quad \underline{x} = \mathbf{D}\underline{\alpha}$$

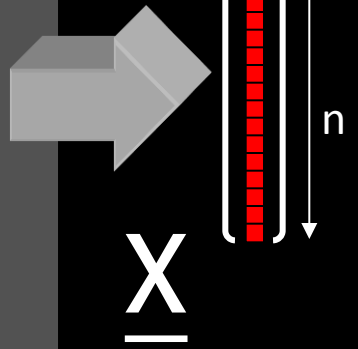
Sparse & Redundant

- Hidden Markov Models,
- Compression algorithms as priors,
- Fields of Experts
- …

Sparse Modeling of Signals

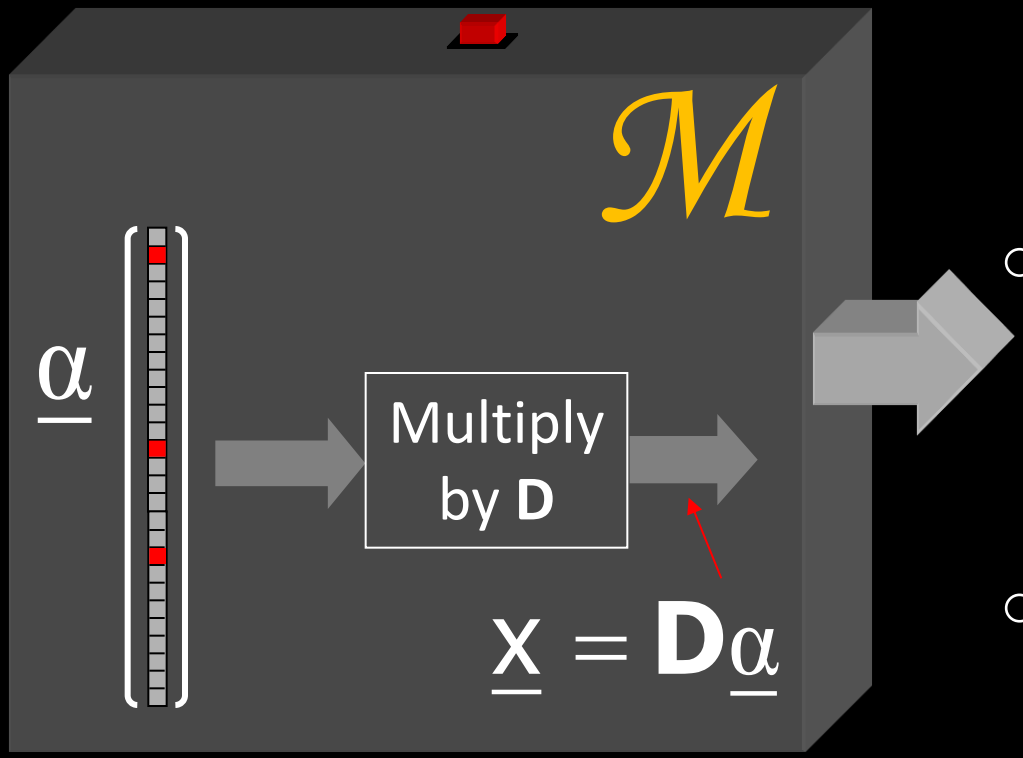


- Every column in **D** (dictionary) is a prototype signal (atom)



- The vector $\underline{\alpha}$ is generated randomly with few (say L) non-zeros at random locations and with random values
- We shall refer to this model as *Sparseland*

Sparseland Signals are Special



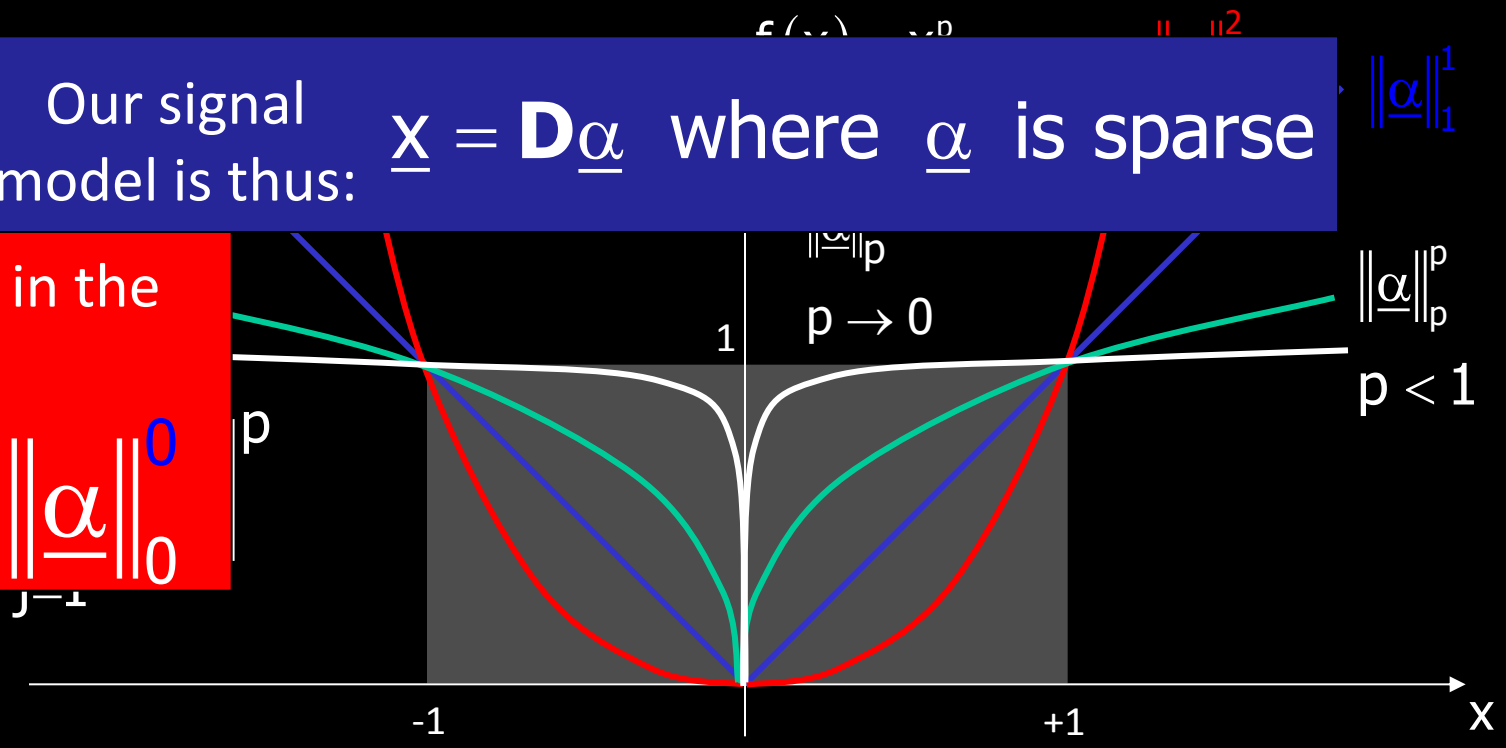
- Simple: Every generated signal is built as a linear combination of few atoms from our dictionary **D**
- Rich: A general model: the obtained signals are a union of many low-dimensional Gaussians
- Familiar: We have been using this model in other context for a while now (wavelet, JPEG2000, …)

Sparse & Redundant Rep. Modeling?

As $p \to 0$ a count of non-zeros in the vector

$$\Rightarrow \|\underline{\alpha}\|_0^0$$

Our signal model is thus: $\underline{x} = \mathbf{D}\underline{\alpha}$ where $\underline{\alpha}$ is sparse



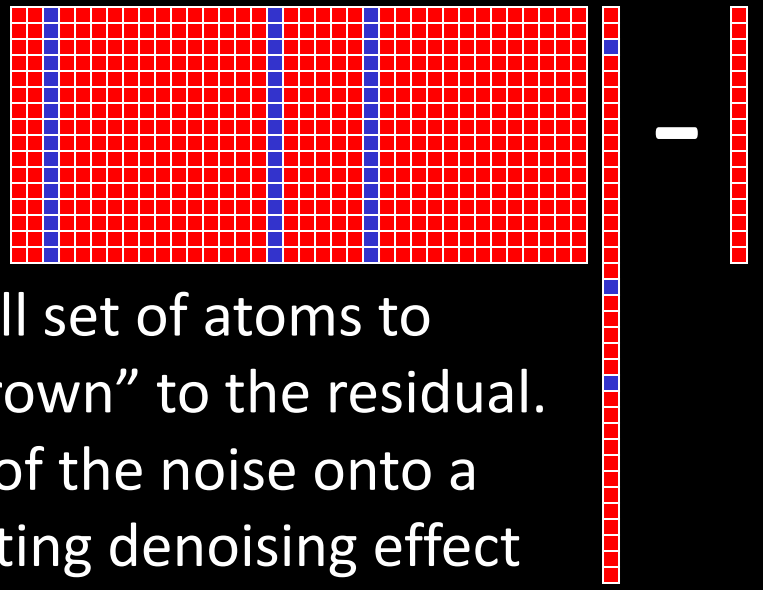
$$\underline{x} = \mathbf{D}\underline{\alpha} \quad \text{where} \quad \|\underline{\alpha}\|_0 \le L$$

Back to the MAP Energy Function

- L_0 norm is effectively counting the number of non-zeros in $\underline{\alpha}$
- The vector $\underline{\alpha}$ is the representation (sparse/redundant) of the desired signal $\underline{x}$

$$\frac{1}{2}\left\| \underline{x} - \underline{y} \right\|_2^2$$

$$\mathbf{D}\underline{\alpha} - \underline{y} =$$



- The core idea: One cannot find a small set of atoms to represent the noise and thus it is "thrown" to the residual.
  $\rightarrow$ We obtain an effective projection of the noise onto a very low-dimensional space, thus getting denoising effect

Wait! There are Some Issues

- Numerical Problems: How should we solve or approximate the solution of the problem

$$\min_{\underline{\alpha}} \left\|\mathbf{D}\underline{\alpha}-\underline{y}\right\|_2^2 \quad \text{s.t.} \quad \left\|\underline{\alpha}\right\|_0 \le L \qquad \min_{\underline{\alpha}} \left\|\underline{\alpha}\right\|_0 \quad \text{s.t.} \quad \left\|\mathbf{D}\underline{\alpha}-\underline{y}\right\|_2^2 \le \varepsilon^2$$

$$\min_{\underline{\alpha}} \lambda\left\|\underline{\alpha}\right\|_0 + \left\|\mathbf{D}\underline{\alpha}-\underline{y}\right\|_2^2 \; ?$$

- Theoretical Problems: Is there a unique sparse representation? If we are to approximate the solution somehow, how close will we get?
- Practical Problems: What dictionary **D** should we use, such that all this leads to effective denoising? Will all this work in applications?

To Summarize So Far …

Image denoising (and many other problems in image processing) requires a model for the desired image

What do we do?

We proposed a model for signals/images based on sparse and redundant representations

Great! No?

There are some issues:

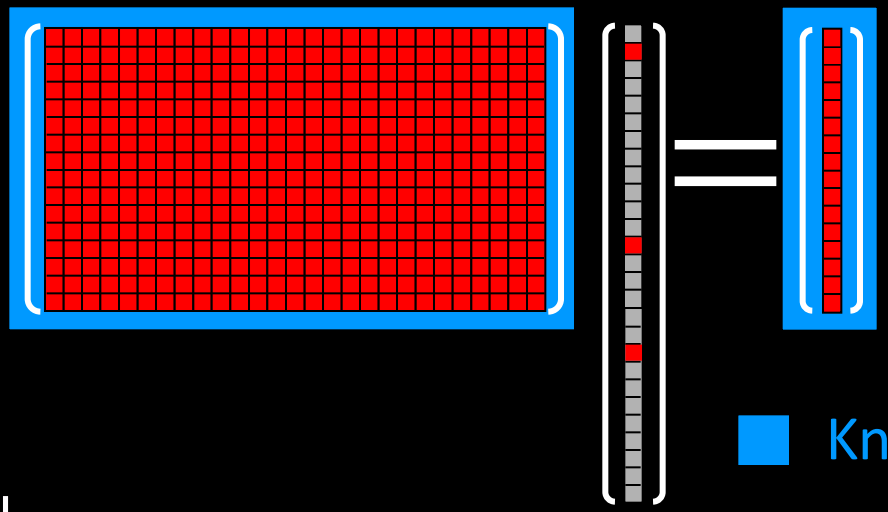
1. Theoretical
2. How to approximate?
3. What about **D**?

Theoretical & Numerical Foundations

Lets Start with the Noiseless Problem

Suppose we build a signal by the relation

$$\mathbf{D}\underline{\alpha} = \underline{x}$$



We aim to find the signal's representation:

$$\underline{\hat{\alpha}} = \underset{\underline{\alpha}}{\text{Arg Min}} \left\|\underline{\alpha}\right\|_0 \quad \text{s.t.} \quad \underline{x} = \mathbf{D}\underline{\alpha}$$

Uniqueness

Why should we necessarily get $\underline{\hat{\alpha}} = \underline{\alpha}$?

It might happen that eventually $\left\|\underline{\hat{\alpha}}\right\|_0 < \left\|\underline{\alpha}\right\|_0$.

Matrix "Spark"

Definition: Given a matrix **D**, σ=Spark{**D**} is the smallest number of columns that are linearly dependent *

Donoho & E. ('02)

Example:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Rank = 4

Spark = 3

* In tensor decomposition, Kruskal defined something similar already in 1989

Uniqueness Rule

Suppose this problem has been solved somehow

$$\hat{\underline{\alpha}} = \text{Arg}\underset{\underline{\alpha}}{\text{Min}} \|\underline{\alpha}\|_0 \quad \text{s.t.} \quad \underline{x} = \mathbf{D}\underline{\alpha}$$

Uniqueness

Donoho & E. ('02)

If we found a representation that satisfy

$$\|\hat{\underline{\alpha}}\|_0 < \frac{\sigma}{2}$$

Then necessarily it is unique (the sparsest).

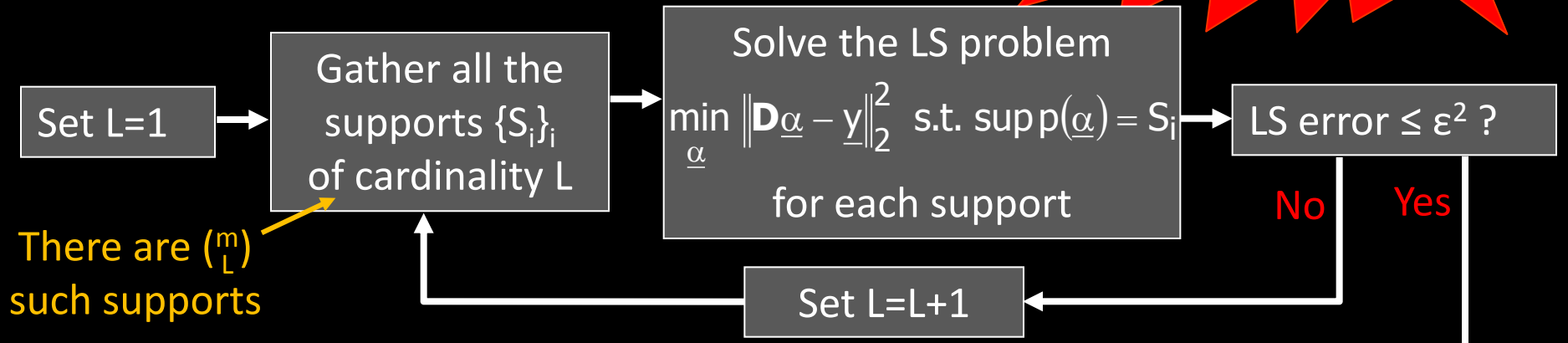
This result implies that if $\mathcal{M}$ generates signals using "sparse enough" $\underline{\alpha}$, the solution of the above will find it exactly

Our Goal

$$\min_{\underline{\alpha}} \|\underline{\alpha}\|_0 \quad \text{s.t.} \quad \|\mathbf{D}\underline{\alpha} - \underline{y}\|_2^2 \le \varepsilon^2$$

This is a combinatorial problem, proven to be NP-Hard!

Here is a recipe for solving this problem:



Assume: m=2000, L=15 (known!), 1 nano-sec per each LS

➡ We shall need ~7.5e20 years to solve this problem !!!!!

Lets Approximate

$$\min_{\underline{\alpha}} \|\underline{\alpha}\|_0 \quad \text{s.t.} \quad \|\mathbf{D}\underline{\alpha} - \underline{y}\|_2^2 \le \varepsilon^2$$

Relaxation methods

Smooth the L_0 and use continuous optimization techniques

Greedy methods

Build the solution one non-zero element at a time

Relaxation – The Basis Pursuit (BP)

Instead of solving

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

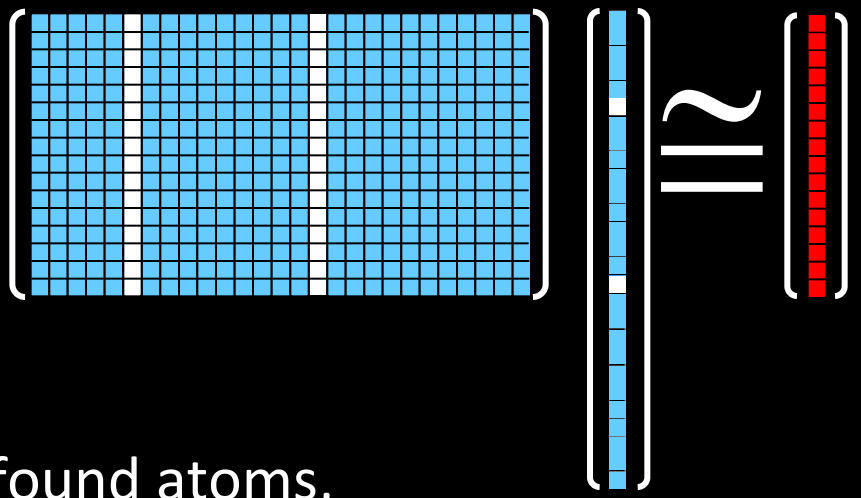
Solve Instead

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

- This is known as the Basis-Pursuit (BP) [Chen, Donoho & Saunders ('95)]
- The newly defined problem is convex (quad. programming)
- Very efficient solvers can be deployed:
  - Interior point methods [Chen, Donoho, & Saunders ('95)] [Kim, Koh, Lustig, Boyd, & D. Gorinevsky (`07)]
  - Sequential shrinkage for union of ortho-bases [Bruce et.al. ('98)]
  - Iterative shrinkage [Figuerido & Nowak ('03)] [Daubechies, Defrise, & De-Mole ('04)] [E. ('05)] [E., Matalon, & Zibulevsky ('06)] [Beck & Teboulle (`09)] …

Go Greedy: Matching Pursuit (MP)

- The MP is one of the greedy algorithms that finds one atom at a time [Mallat & Zhang ('93)]
- Step 1: find the one atom that best matches the signal



- Next steps: given the previously found atoms, find the next one to best fit the residual
- The algorithm stops when the error $\|\mathbf{D}\underline{\alpha} - \underline{y}\|_2$ is below the destination threshold.
- The Orthogonal MP (OMP) is an improved version that re-evaluates the coefficients by Least-Squares after each round.

Pursuit Algorithms

$$\min_{\underline{\alpha}} \|\underline{\alpha}\|_0 \quad \text{s.t.} \quad \|\mathbf{D}\underline{\alpha} - \underline{y}\|_2^2 \le \varepsilon^2$$

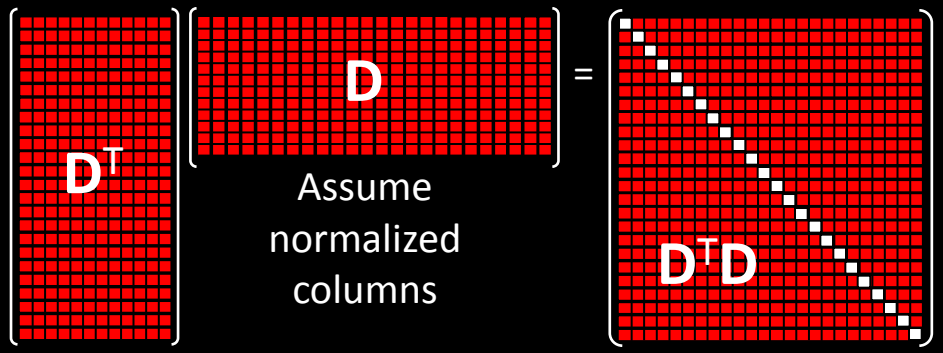
There are various algorithms designed for approximating the solution of this problem:

- Greedy Algorithms: Matching Pursuit, Orthogonal Matching Pursuit (OMP), Least-Squares-OMP, Weak Matching Pursuit, Block Matching Pursuit [1993-today].
- Relaxation Algorithms: Basis Pursuit (a.k.a. LASSO), Dnatzig Selector & numerical ways to handle them [1995-today].
- Hybrid Algorithms: StOMP, CoSaMP, Subspace Pursuit, Iterative Hard-Thresholding [2007-today].
- ...

Why should we trust these methods ?

The Mutual-Coherence

- Compute



- The Mutual-Coherence μ is the largest off-diagonal entry in absolute value

- The Mutual-Coherence is a property of the dictionary (just like the "Spark"). In fact, the following relation can be shown:

$$\sigma \geq 1 + \frac{1}{\mu}$$

BP and MP Equivalence (No Noise)

Equivalence

Donoho & E. (‘02)
Gribonval & Nielsen (‘03)
Tropp (‘03)
Temlyakov (‘03)

Given a signal $\underline{x}$ with a representation $\underline{x} = \mathbf{D}\underline{\alpha}$, assuming that $\|\underline{\alpha}\|_0 < 0.5\left(1 + 1/\mu\right)$, BP and MP are guaranteed to find the sparsest solution.

$$\hat{\underline{\alpha}} = \text{Arg}\,\text{Min}_{\underline{\alpha}} \|\underline{\alpha}\|_0 \quad \text{s.t.} \quad \underline{x} = \mathbf{D}\underline{\alpha}$$

- MP and BP are different in general (hard to say which is better)
- The above result corresponds to the worst-case, and as such, it is too pessimistic
- Average performance results are available too, showing much better bounds [Donoho (`04)] [Candes et.al. (‘04)] [Tanner et.al. (‘05)] [E. (‘06)] [Tropp et.al. (‘06)] … [Candes et. al. (‘09)]

BP Stability for the Noisy Case

Stability

Given a signal $\underline{y} = \mathbf{D}\underline{\alpha} + \underline{v}$ with a representation satisfying $\|\underline{\alpha}\|_0 < 1/3\mu$ and a white Gaussian Noise $\underline{v} \sim N(0, \sigma^2\mathbf{I})$

Ben-Haim, Eldar & E. ('09)

* With very high probability

- For σ=0 we get a wea
- This result is the orac

$$\min_{\underline{\alpha}} \lambda \|\underline{\alpha}\|_1 + \|\mathbf{D}\underline{\alpha} - \underline{y}\|_2^2$$

- Similar results exist for other pursuit algorithms (Dantzig Selector, Orthogonal Matching Pursuit, CoSaMP, Subspace Pursuit, …)

To Summarize So Far ...

Image denoising (and many other problems in image processing) requires a model for the desired image

What do we do?

We proposed a model for signals/images based on sparse and redundant representations

Problems?

We have seen that there are approximation methods to find the sparsest solution, and there are theoretical results that guarantee their success.

What next?

The Dictionary **D** should be found somehow !!!

Dictionary Learning: The K-SVD Algorithm

What Should **D** Be?

$$\hat{\underline{\alpha}} = \arg\min_{\underline{\alpha}} \|\underline{\alpha}\|_0 \quad \text{s.t.} \quad \frac{1}{2}\|\mathbf{D}\underline{\alpha} - \underline{y}\|_2^2 \le \varepsilon^2 \quad \Rightarrow \quad \hat{\underline{x}} = \mathbf{D}\hat{\underline{\alpha}}$$

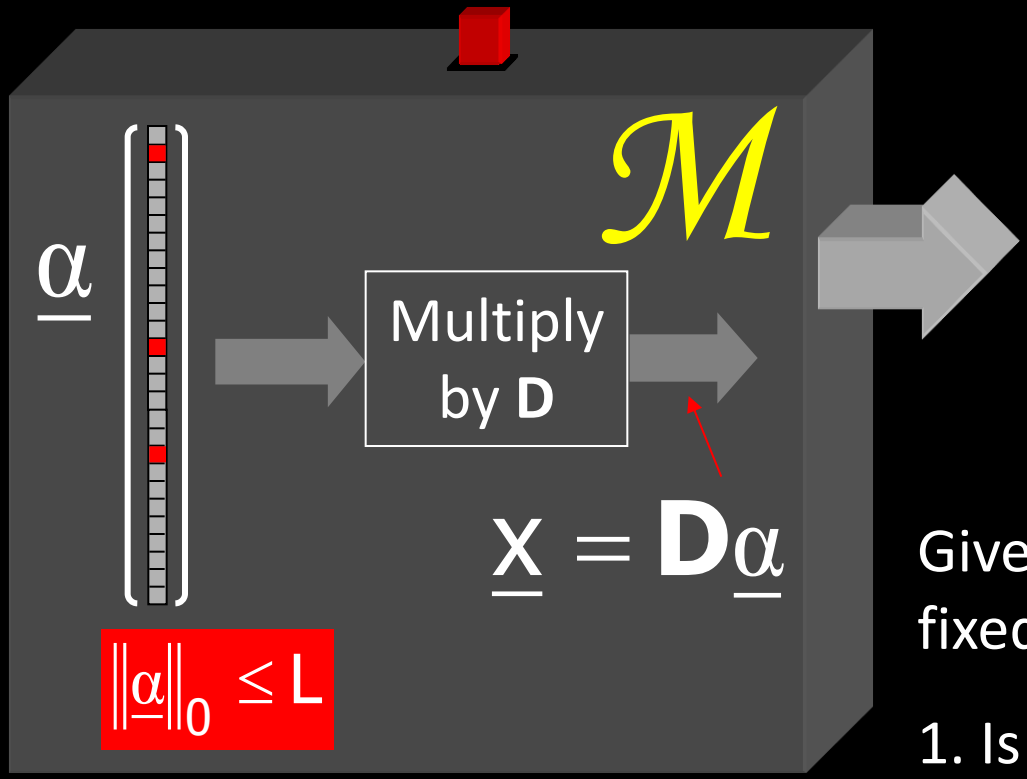
Our Assumption: Good-behaved Images have a sparse representation

D should be chosen such that it sparsifies the representations

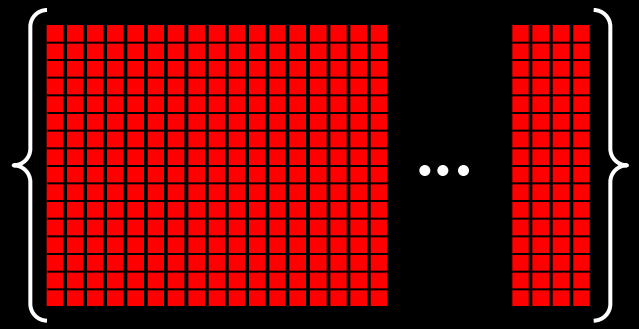
One approach to choose **D** is from a known set of transforms (Steerable wavelet, Curvelet, Contourlets, Bandlets, Shearlets …)

The approach we will take for building **D** is training it, based on Learning from Image Examples

Dictionary Learning: Problem Setting



$$\left\{ \underline{X}_j \right\}_{j=1}^{P}$$

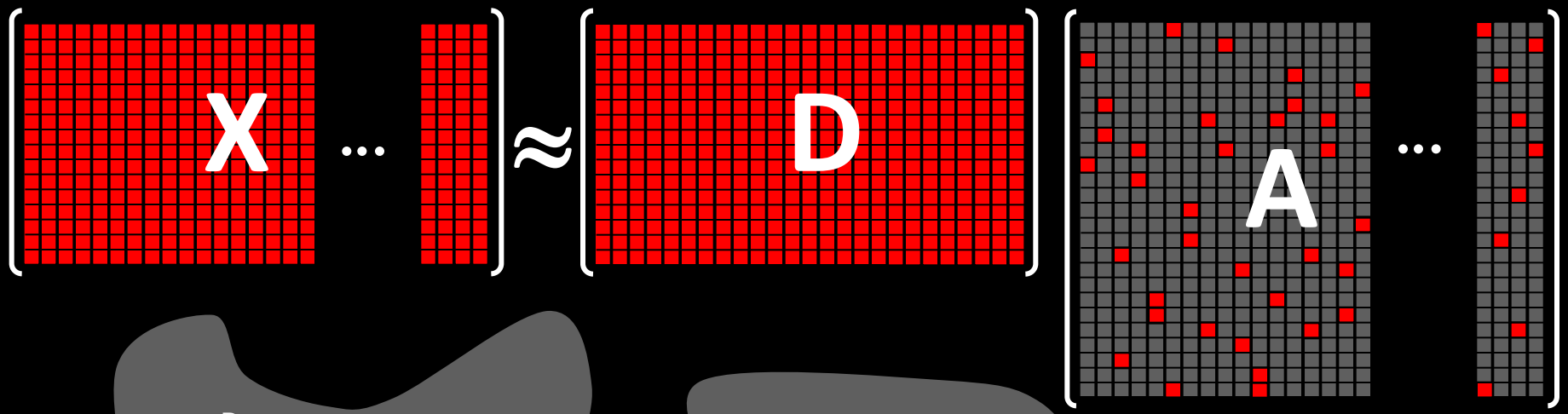


Given these P examples and a fixed size [n×m] dictionary **D**:

1. Is **D** unique?

2. How would we find **D**?

Measure of Quality for **D**



$$\underset{\mathbf{D},\mathbf{A}}{\text{Min}} \sum_{j=1}^{P} \left\| \mathbf{D}\underline{\alpha}_j - \underline{x}_j \right\|_2^2 \quad \text{s.t.} \quad \forall j, \ \left\| \underline{\alpha}_j \right\|_0 \le L$$

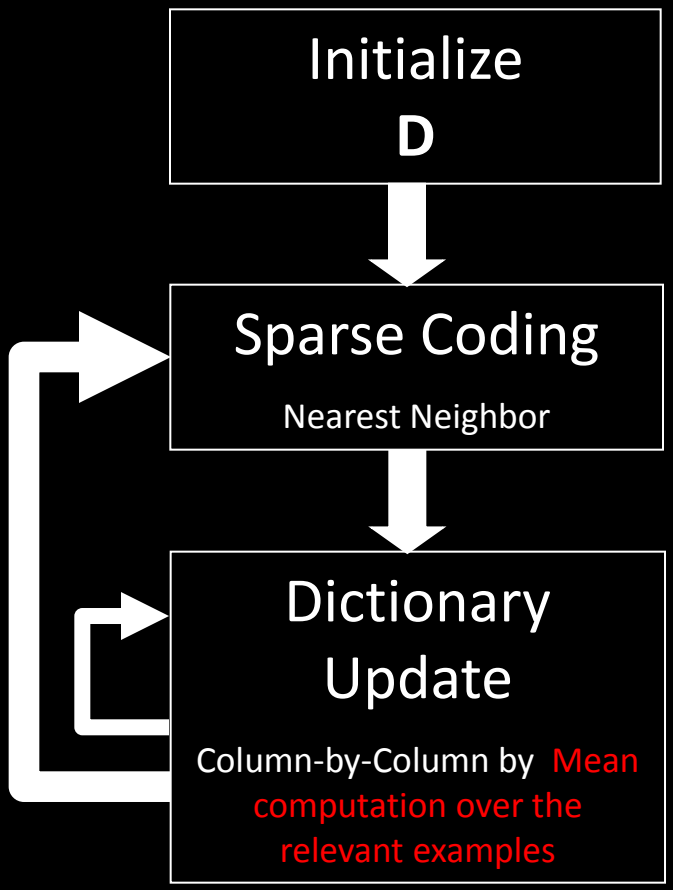
Each example is a linear combination of atoms from **D**

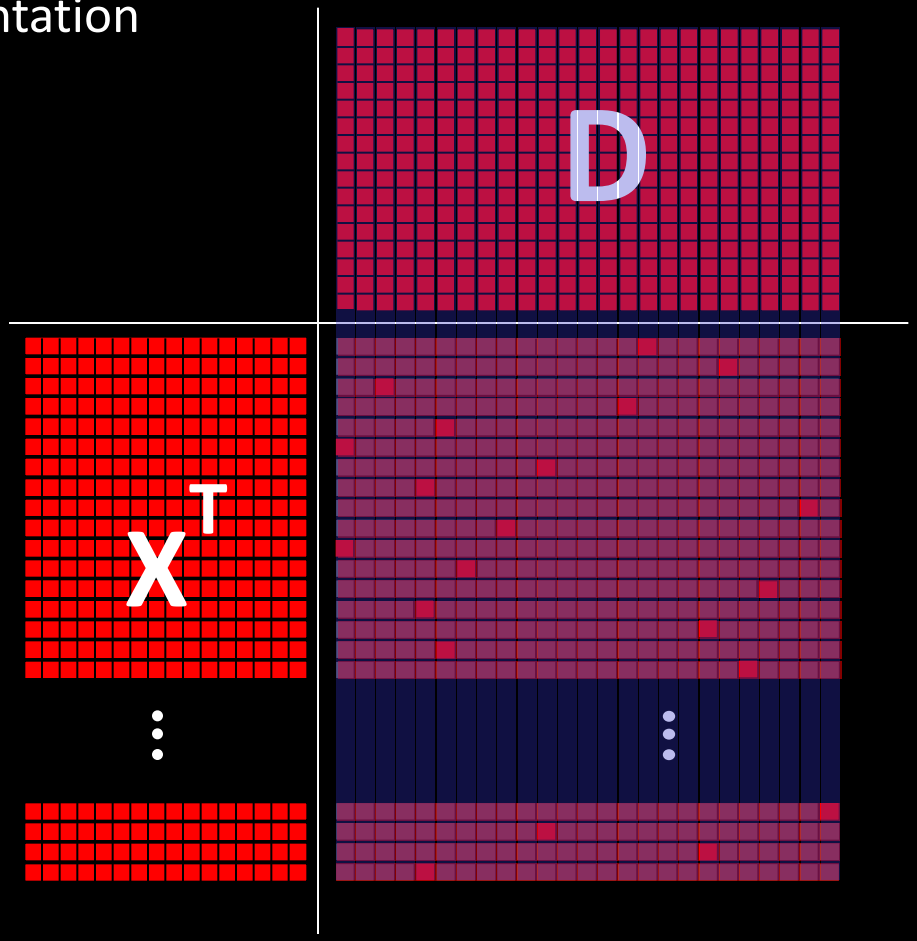
Each example has a sparse representation with no more than L atoms

[Field & Olshausen ('96)]
[Engan et. al. ('99)]
[Lewicki & Sejnowski ('00)]
[Cotter et. al. ('03)]
[Gribonval et. al. ('04)]
[Aharon, E. & Bruckstein ('04)]
[Aharon, E. & Bruckstein ('05)]

K–Means For Clustering

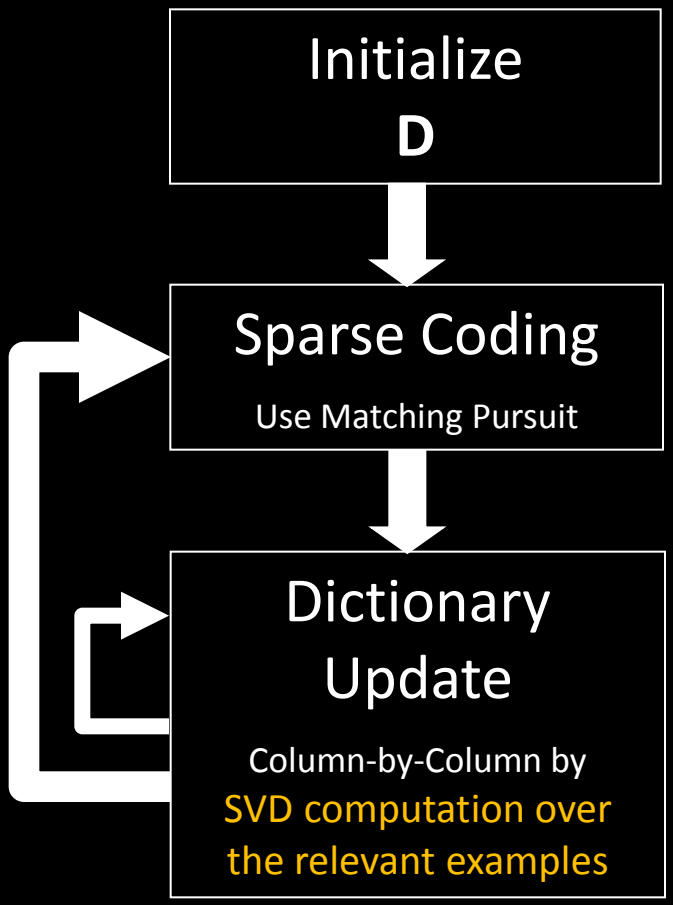
Clustering: An extreme sparse representation

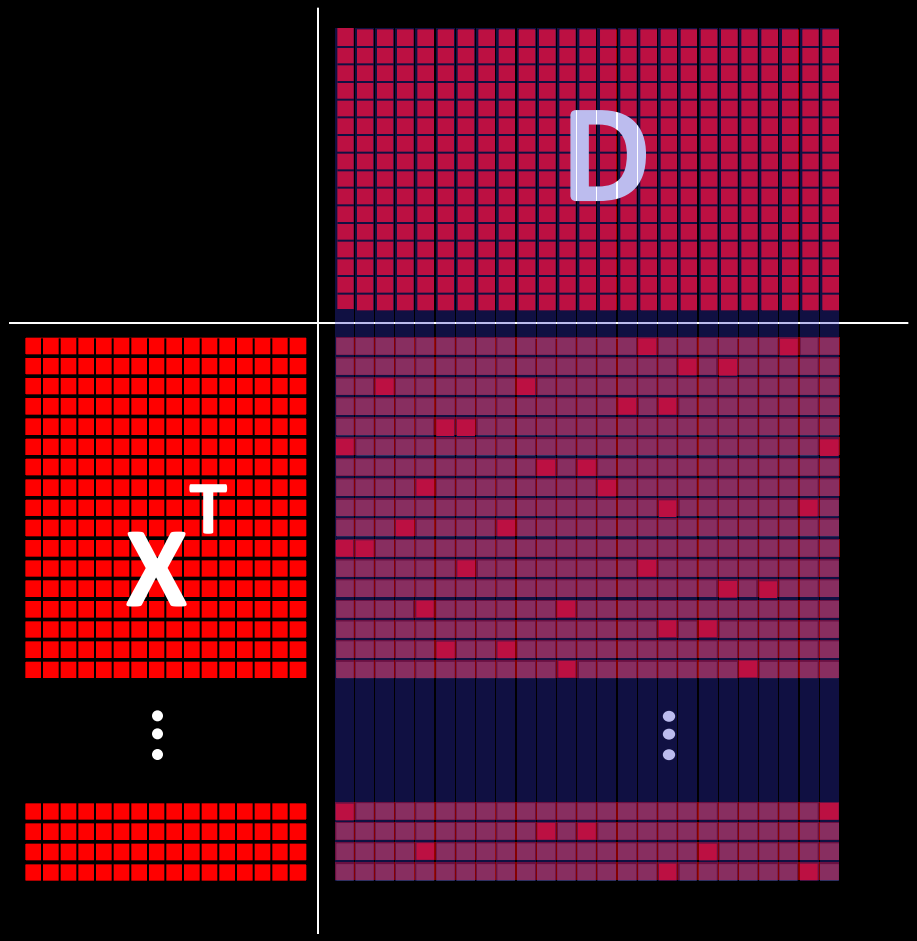




The K–SVD Algorithm – General

[Aharon, E. & Bruckstein ('04,'05)]





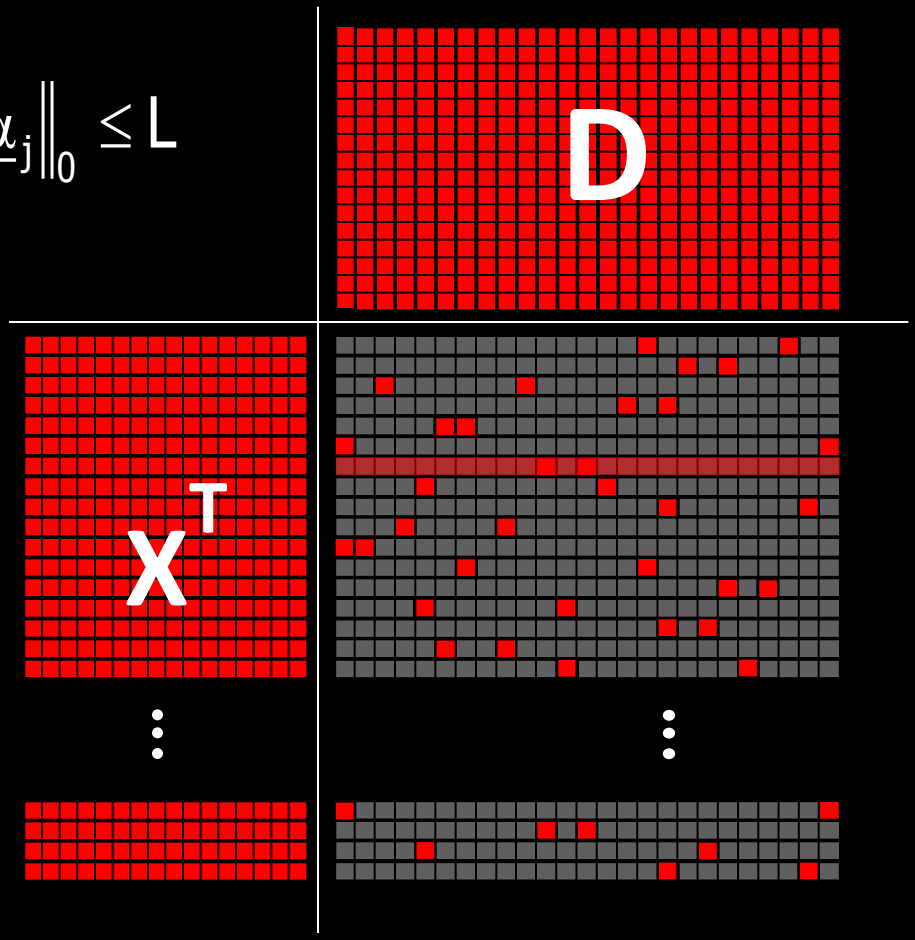
K–SVD: Sparse Coding Stage

$$\min_{\mathbf{A}} \sum_{j=1}^{P} \left\| \mathbf{D}\underline{\alpha}_j - \underline{x}_j \right\|_2^2 \quad \text{s.t.} \quad \forall j, \ \left\| \underline{\alpha}_j \right\|_0 \le L$$

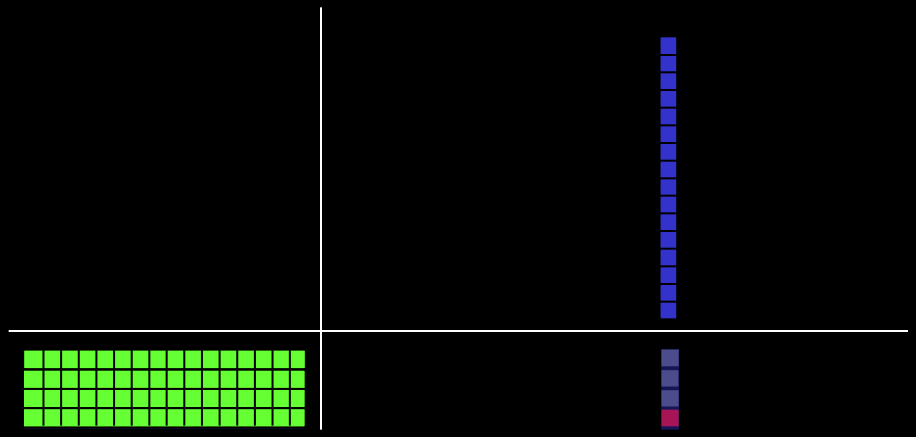
D is known! For the j^{th} item we solve

$$\min_{\underline{\alpha}} \left\| \mathbf{D}\underline{\alpha} - \underline{x}_j \right\|_2^2 \quad \text{s.t.} \quad \left\| \underline{\alpha} \right\|_0 \le L$$

Solved by
A Pursuit Algorithm



K–SVD: Dictionary Update Stage



We refer only to the examples that use the column $\underline{d}_k$

We should solve:

$$\underset{\underline{d}_k,\alpha_k}{\text{Min}} \left\| \alpha_k \underline{d}_k^T - \mathbf{E} \right\|_F^2$$

SVD

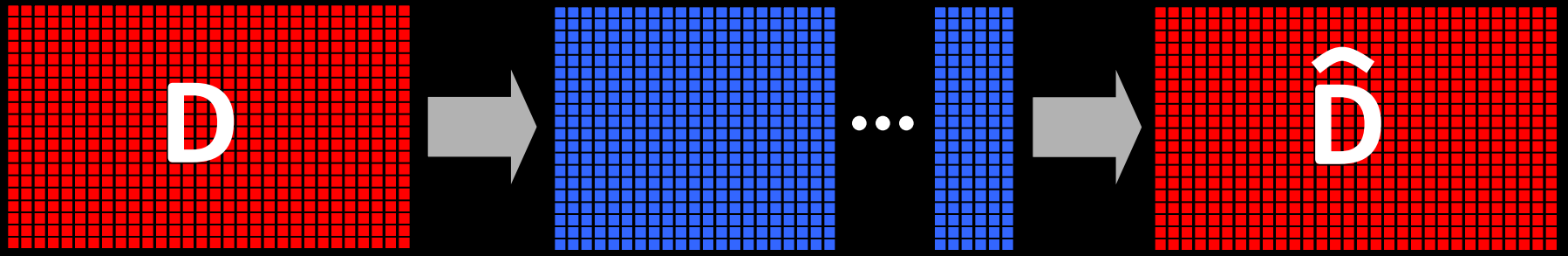
Fixing all **A** and **D** apart from the k^{th} column, and seek both $\underline{d}_k$ and the k^{th} column in A to better fit the residual!

A Synthetic Experiment

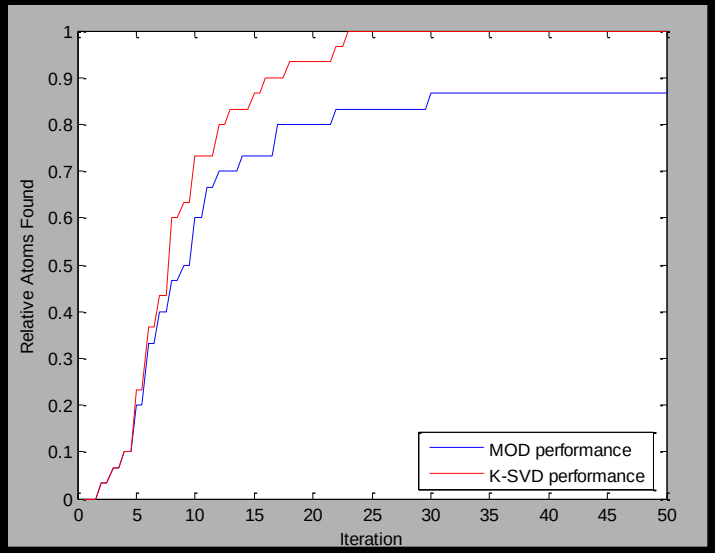
Create A 20×30 random dictionary with normalized columns

Generate 2000 signal examples with 3 atoms per each and add noise

Train a dictionary using the KSVD and MOD and compare



Results



Improved Dictionary Learning

$$\underset{\mathbf{D},\mathbf{A}}{\text{Min}} \sum_{j=1}^{P} \left\| \mathbf{D}\underline{\alpha}_j - \underline{x}_j \right\|_2^2 \quad \text{s.t. } \forall j, \left\| \underline{\alpha}_j \right\|_0 \le L$$

MOD Algorithm

Fix **D** and update **A**

Fix **A** and update **D**

K-SVD Algorithm

Fix **D** and update **A**

for j=1:1:m

- Fix **A** & **D** apart from the j-th atom its coefficients
- Update $\underline{d}_j$ and its coef. in **A**

end

Improved Dictionary Learning

$$\underset{\mathbf{D},\mathbf{A}}{\text{Min}} \sum_{j=1}^{P} \left\| \mathbf{D}\underline{\alpha}_j - \underline{x}_j \right\|_2^2 \quad \text{s.t. } \forall j, \left\| \underline{\alpha}_j \right\|_0 \le L$$

Improved Algorithm
[Smith & E. 2013]

Fix **D** and update **A**

Fix **A** and update **D**

Zeros

MOD and K-SVD can be considered as crude approximation of this method

$\mathbf{M} = \mathbf{0}$

This can be done in two ways:

1. Apply several rounds of the atoms' update in the K-SVD, or
2. Extend the MOD to update the non-zero elements in **A**

To Summarize So Far …

Image denoising (and many other problems in image processing) requires a model for the desired image

What do we do?

We proposed a model for signals/images based on sparse and redundant representations

Problems?

We have seen approximation methods that find the sparsest solution, and theoretical results that guarantee their success. We also saw a way to learn **D**

What next?

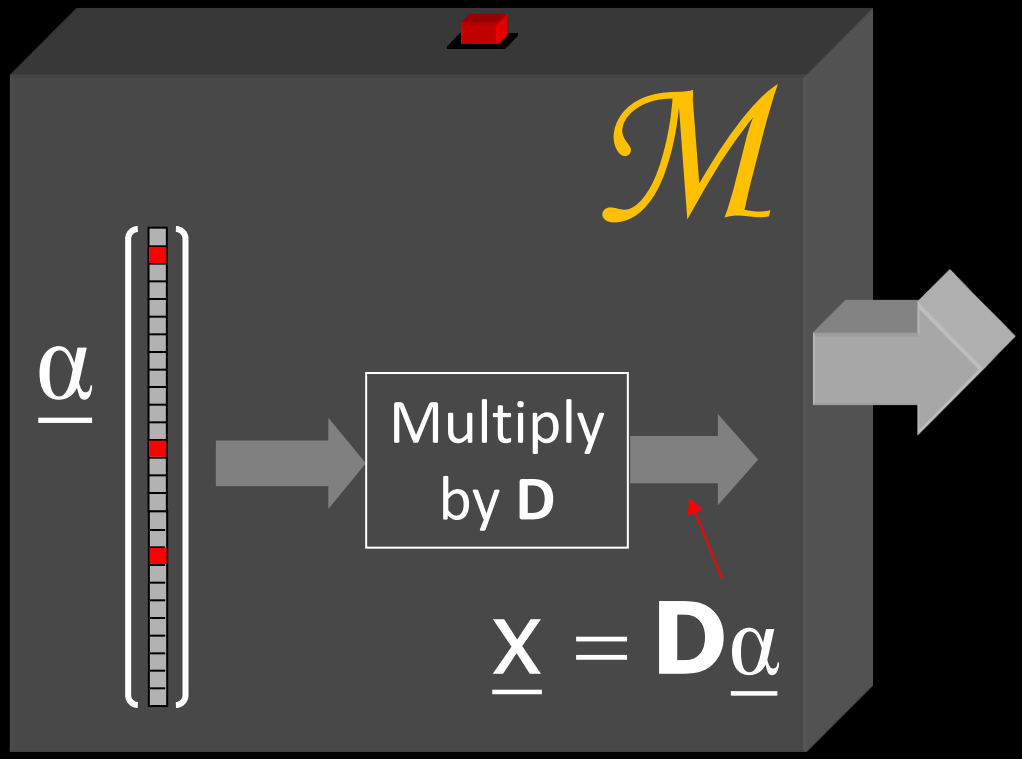
Will it all work in applications?

BREAK

Back to Denoising …
and Beyond – Combining it All

Bringing Sparseland to Applications

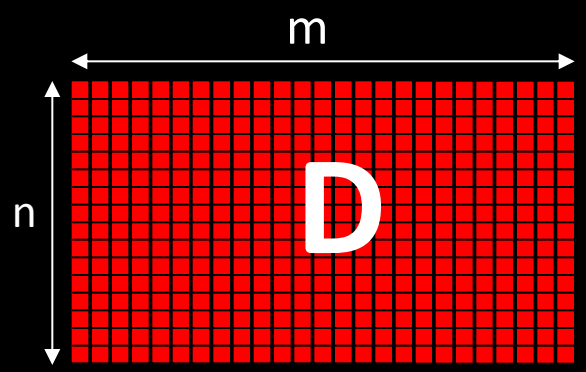
- While the Sparseland model is clear and well-defined, there are various ways to bring it into an actual algorithm in applications



- The bad news: It is not obvious how to turn this model into a successful algorithm
- The good news: As we are about to see, there is a lot of room for ingenuity & originality in designing algorithms in image processing

From Local to Global Treatment

- The K-SVD is reasonable for low-dimension signals (n in the range 10-400). As n grows, the complexity and the memory requirements of the K-SVD are prohibitive



- So, how should large images be handled?
- The solution: Force shift-invariant sparsity – operate on patches of size n-by-n (n=8) in the image, including overlaps

$$\hat{\underline{x}} = \underset{\underline{x},\{\underline{\alpha}_{ij}\}_{ij}}{\text{ArgMin}} \frac{1}{2}\left\|\underline{x}-\underline{y}\right\|_2^2 + \mu \sum_{ij} \left\|\mathbf{R}_{ij}\underline{x} - \mathbf{D}\underline{\alpha}_{ij}\right\|_2^2 \quad \text{s.t.} \quad \left\|\underline{\alpha}_{ij}\right\|_0 \le L$$

Extracts a patch in the (i,j) location

Our prior

What Data to Train On?

Option 1:

- Use a database of images
- We tried that, and it works fine (~0.5-1dB below the state-of-the-art)

Option 2:

- Use the corrupted image itself !!
- Simply sweep through all patches of size $\sqrt{n}$-by-$\sqrt{n}$ (overlapping blocks)
- Image of size 1000^2 pixels ➡ ~10^6 examples to use – more than enough
- This works much better!

K-SVD Image Denoising

$$\hat{\underline{x}} = \underset{\underline{x},\{\underline{\alpha}_{ij}\}_{ij}, \mathbf{D}?}{\text{ArgMin}} \frac{1}{2}\left\|\underline{x}-\underline{y}\right\|_2^2 + \mu\sum_{ij}\left\|\mathbf{R}_{ij}\underline{x}-\mathbf{D}\underline{\alpha}_{ij}\right\|_2^2 \quad \text{s.t.} \quad \left\|\underline{\alpha}_{ij}\right\|_0 \le L$$

$\underline{x}=\underline{y}$ and **D** known

$\underline{x}$ and α_{ij} known

D and α_{ij} known

Compute α_{ij} per patch

$$\underline{\alpha}_{ij} = \underset{\underline{\alpha}}{\text{Min}}\left\|\mathbf{R}_{ij}\underline{x}-\mathbf{D}\underline{\alpha}\right\|_2^2$$

$$\text{s.t.} \left\|\underline{\alpha}\right\|_0 \le L$$

using the matching pursuit

Compute **D** to minimize

$$\underset{\underline{\alpha}}{\text{Min}}\sum_{ij}\left\|\mathbf{R}_{ij}\underline{x}-\mathbf{D}\underline{\alpha}\right\|_2^2$$

using SVD, updating one column at a time

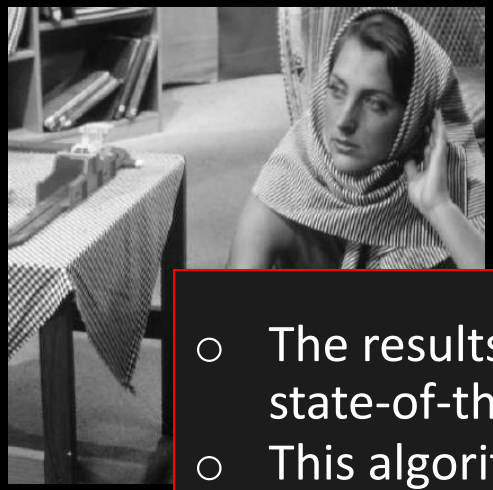
K-SVD

Compute $\underline{x}$ by

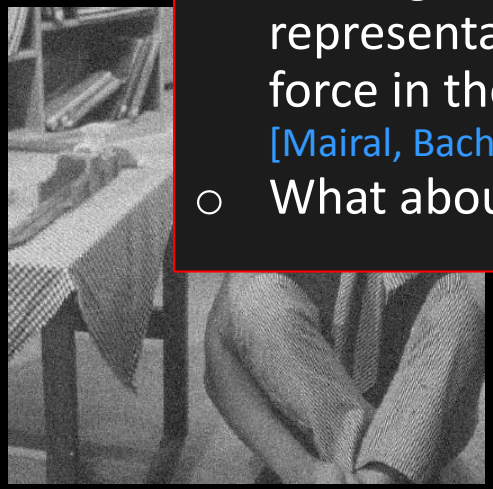
$$\underline{x} = \left[\mathrm{I}+\mu\sum_{ij}\mathbf{R}_{ij}^{\mathrm{T}}\mathbf{R}_{ij}\right]^{-1}\left[\underline{y}+\mu\sum_{ij}\mathbf{R}_{ij}^{\mathrm{T}}\mathbf{D}\underline{\alpha}_{ij}\right]$$

which is a simple averaging of shifted patches

Image Denoising (Gray) [E. & Aharon ('06)]



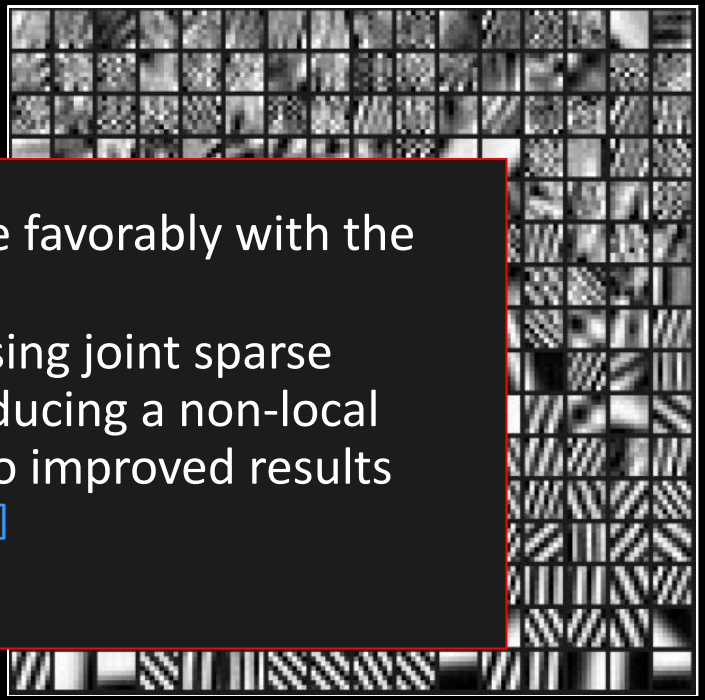
Source



Noisy image

$\sigma = 20$

- The results of this algorithm compete favorably with the state-of-the-art
- This algorithm can be extended by using joint sparse representation on the patches, introducing a non-local force in the denoising, thus leading to improved results [Mairal, Bach, Ponce, Sapiro & Zisserman ('09)]
- What about EPLL ? …

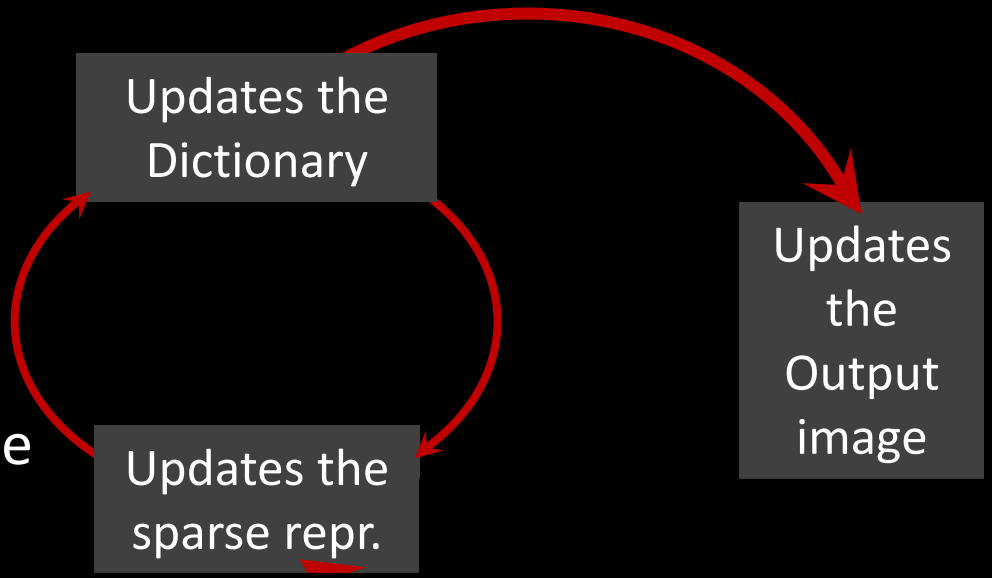


The obtained dictionary after 10 iterations

EPLL Improvement [Sulam and E. ('15)]

$$\hat{\underline{x}} = \underset{\underline{x},\{\underline{\alpha}_{ij}\}_{ij},\mathbf{D}}{\text{ArgMin}} \frac{1}{2}\left\|\underline{x}-\underline{y}\right\|_2^2 + \mu\sum_{ij}\left\|\mathbf{R}_{ij}\underline{x}-\mathbf{D}\underline{\alpha}_{ij}\right\|_2^2 \text{ s.t. } \left\|\underline{\alpha}_{ij}\right\|_0 \le L$$

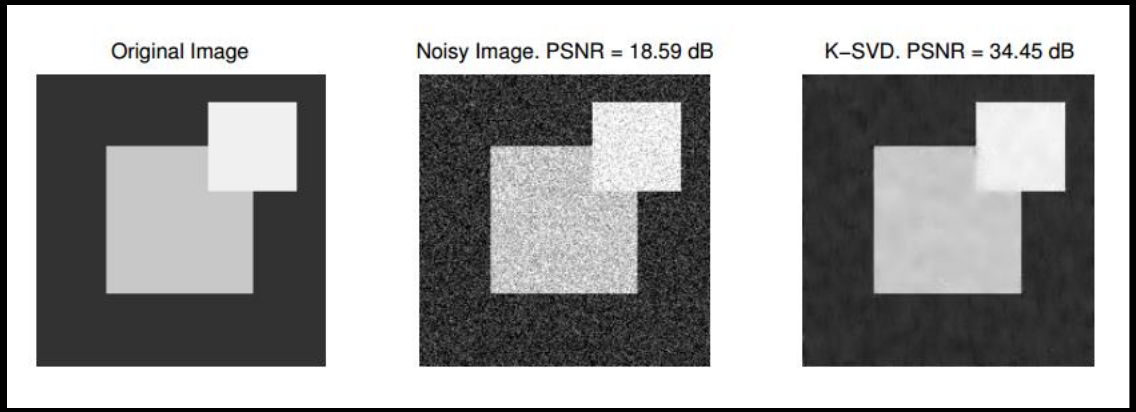
- The algorithm we proposed updates $\underline{x}$ only once at the end
- Why not repeat the whole process several times?
- The rationale: The sparse representation model should be imposed on the patches of the FINAL image. After averaging, this is ruined



EPLL Improvement [Sulam and E. ('15)]

- Expected Patch Log Likelihood (EPLL) is an algorithm that came to fix this problem [Zoran and Weiss, ('11)] , originally in the context of a GMM prior
- An extension of EPLL to *Sparsland* is proposed in [Sulam and E. ('15)]. The core idea is:
  - After the image has been computed, we proceed the iterative process, and apply several such overall rounds of updates
  - Sparse coding must be done with a new threshold, based on the remaining noise in the image. This is done by evaluating the noise level based on the linear projections (disregarding the support detection by the OMP)
  - This algorithm leads to state-of-the-art results, with 0.5-1dB improvement over the regular K-SVD algorithm shown before

EPLL Improvement [Sulam and E. ('15)]



Noisy image has σ=25

KSVD PSNR 31.42 dB

EPLL PSNR 31.83 dB

Denoising (Color) [Mairal, E. & Sapiro ('08)]

- When turning to handle color images, the main

Original

Noisy (20.43dB)

Result (30.75dB)

Denoising (Color) [Mairal, E. & Sapiro ('08)]

Our experiments lead to state-of-the-art denoising results, giving ~1dB better results compared to [Mcauley et. al. ('06)] which implements a learned MRF model (Field-of-Experts)

Original

Noisy (12.77dB)

Result (29.87dB)

Video Denoising [Protter & E. (‘09)]

Our experiments lead to state-of-the-art video denoising results, giving ~0.5dB better results on average compared to [Boades, Coll & Morel (‘05)] and comparable to [Rusanovskyy, Dabov, & Egiazarian (‘06)]

2)

[Buades, Col, and Morel (06)]

Original | Noisy (σ=15) | Denoised (PSNR=29.98)

Low-Dosage Tomography [Shtok, Zibulevsky & E. ('10)]

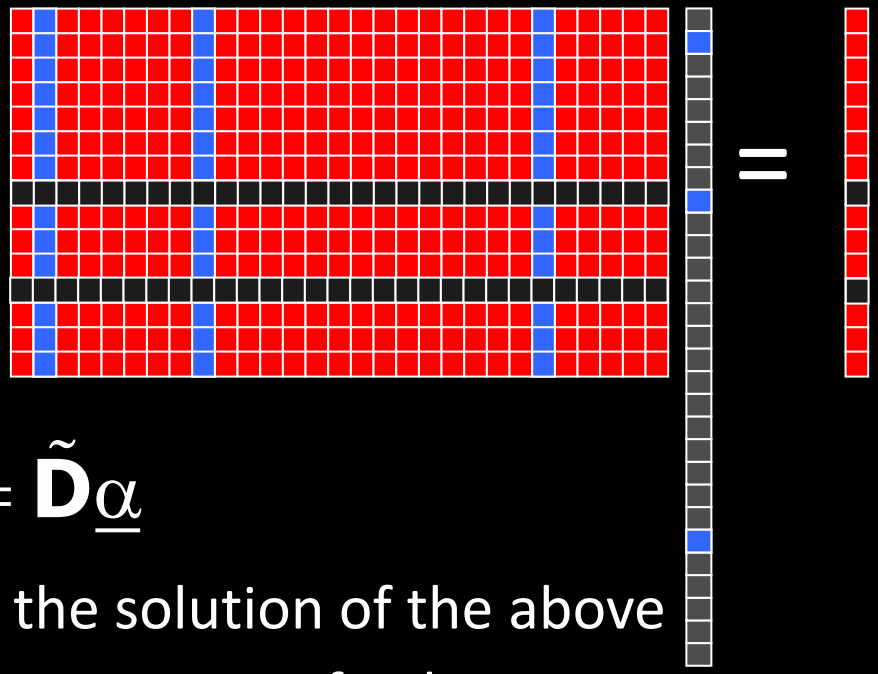
- In Computer-Tomography (CT) reconstruction, an image is recovered from a set of its projections
- In medicine, CT projections are obtained by X-ray, and it typically requires a high dosage of radiation in order to obtain a good quality reconstruction
- A lower-dosage projection implies a stronger noise (Poisson distributed) in data to work with
- Armed with sparse and redundant representation modeling, we can denoise the data and the final reconstruction … enabling CT with lower dosage

Image Inpainting – The Basics

- Assume: the signal $\underline{x}$ has been created by $\underline{x}=\mathbf{D}\underline{\alpha}_0$ with very sparse $\underline{\alpha}_0$

$$\mathbf{D}\,\underline{\alpha}_0 = \underline{x}$$

- Missing values in $\underline{x}$ imply missing rows in this linear system



- By removing these rows, we get

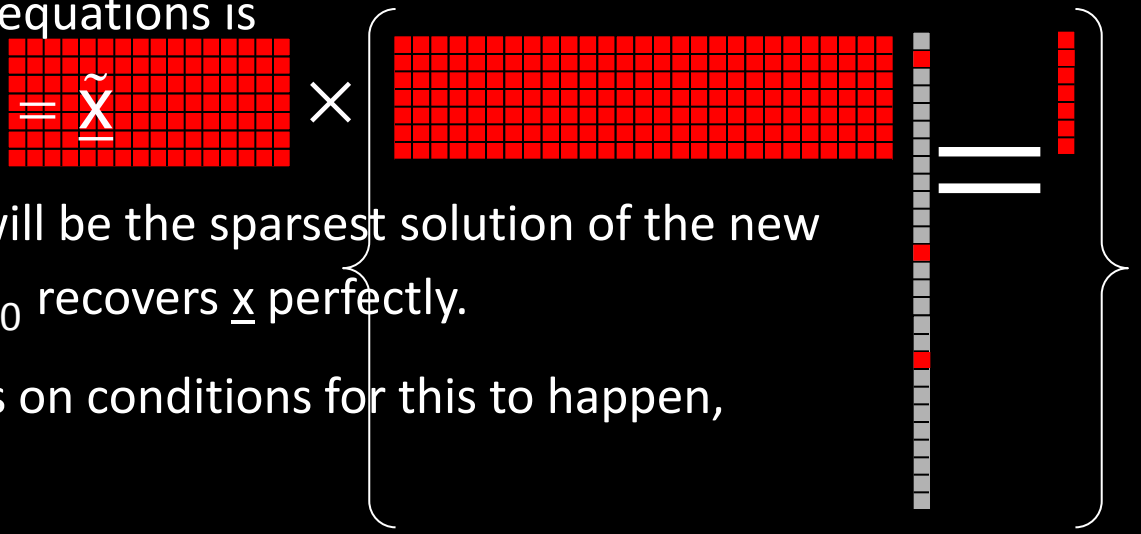
$$\tilde{\mathbf{D}}\underline{\alpha} = \underline{\tilde{x}}$$

- Now solve $\underset{\underline{\alpha}}{\text{Min}} \left\|\underline{\alpha}\right\|_0 \;\; \text{s.t.} \;\; \underline{\tilde{x}} = \tilde{\mathbf{D}}\underline{\alpha}$
- If $\underline{\alpha}_0$ was sparse enough, it will be the solution of the above problem! Thus, computing $\mathbf{D}\underline{\alpha}_0$ recovers $\underline{x}$ perfectly

Side Note: Compressed-Sensing

- Compressed Sensing is leaning on the very same principal, leading to alternative sampling theorems.
- Assume: the signal $\underline{x}$ has been created by $\underline{x}=\mathbf{D}\underline{\alpha}_0$ with very sparse $\underline{\alpha}_0$.
- Multiply this set of equations by the matrix **Q** which reduces the number of rows.
- The new, smaller, system of equations is

$$\mathbf{QD}\underline{\alpha} = \mathbf{Q}\underline{x} \quad \Rightarrow \quad \tilde{\mathbf{D}}\underline{\alpha} = \underline{\tilde{x}}$$

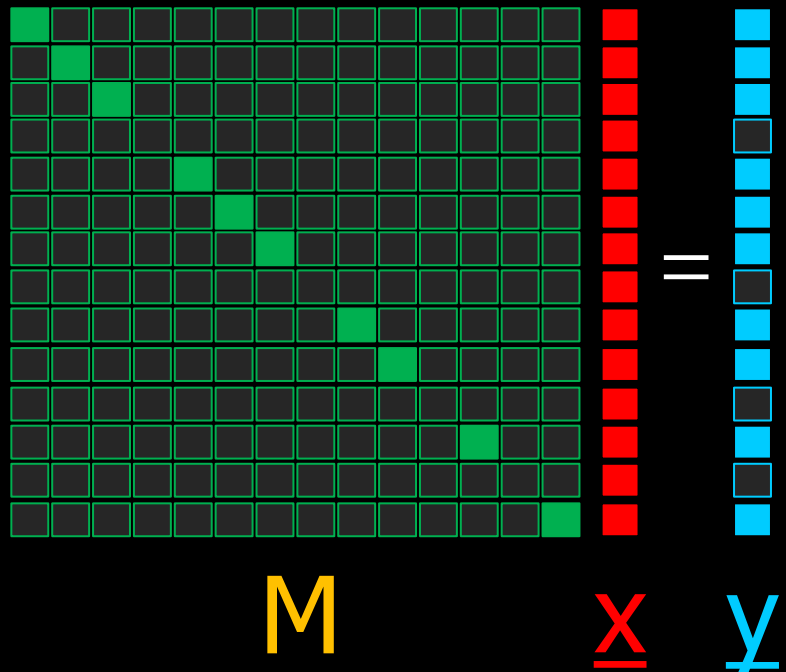


- If $\underline{\alpha}_0$ was sparse enough, it will be the sparsest solution of the new system, thus, computing $\mathbf{D}\underline{\alpha}_0$ recovers $\underline{x}$ perfectly.
- Compressed sensing focuses on conditions for this to happen, guaranteeing such recovery.

Inpainting Formulation [Mairal, E. & Sapiro ('08)]

$$\hat{\underline{x}} = \underset{\underline{x},\{\underline{\alpha}_{ij}\}_{ij},\mathbf{D}}{\text{ArgMin}} \frac{1}{2}\left\|\mathbf{M}\underline{x} - \underline{y}\right\|_2^2 + \mu\sum_{ij}\left\|\mathbf{R}_{ij}\underline{x} - \mathbf{D}\underline{\alpha}_{ij}\right\|_2^2 \text{ s.t. } \left\|\underline{\alpha}_{ij}\right\|_0 \le L$$

The matrix **M** is a mask matrix, obtained by the identity matrix with some of its rows omitted, corresponding to the missing samples



Inpainting Formulation [Mairal, E. & Sapiro ('08)]

$$\hat{\underline{x}} = \underset{\underline{x},\{\underline{\alpha}_{ij}\}_{ij},\mathbf{D}}{\text{ArgMin}} \frac{1}{2}\left\|\mathbf{M}\underline{x} - \underline{y}\right\|_2^2 + \mu\sum_{ij}\left\|\mathbf{R}_{ij}\underline{x} - \mathbf{D}\underline{\alpha}_{ij}\right\|_2^2 \quad \text{s.t.} \quad \left\|\underline{\alpha}_{ij}\right\|_0 \le L$$

$\underline{x}=\underline{y}$ and **D** known

Compute α_{ij} per patch

$$\underline{\alpha}_{ij} = \underset{\underline{\alpha}}{\text{Min}}\left\|\mathbf{M}_{ij}\left(\mathbf{R}_{ij}\underline{x} - \mathbf{D}\underline{\alpha}\right)\right\|_2^2 \quad \text{s.t.} \quad \left\|\underline{\alpha}\right\|_0 \le L$$

using the matching pursuit

$\underline{x}$ and α_{ij} known

Compute **D** to minimize

$$\underset{\underline{\alpha}}{\text{Min}}\sum_{ij}\left\|\mathbf{M}_{ij}\left(\mathbf{R}_{ij}\underline{x} - \mathbf{D}\underline{\alpha}\right)\right\|_2^2$$

using SVD, updating one column at a time

~K-SVD

D and α_{ij} known

Compute $\underline{x}$ by

$$\underline{x} = \left[\mathbf{M}^T\mathbf{M} + \mu\sum_{ij}\mathbf{R}_{ij}^T\mathbf{R}_{ij}\right]^{-1} \cdot \left[\mathbf{M}^T\underline{y} + \mu\sum_{ij}\mathbf{R}_{ij}^T\mathbf{D}\underline{\alpha}_{ij}\right]$$

which is a again a simple averaging of patches

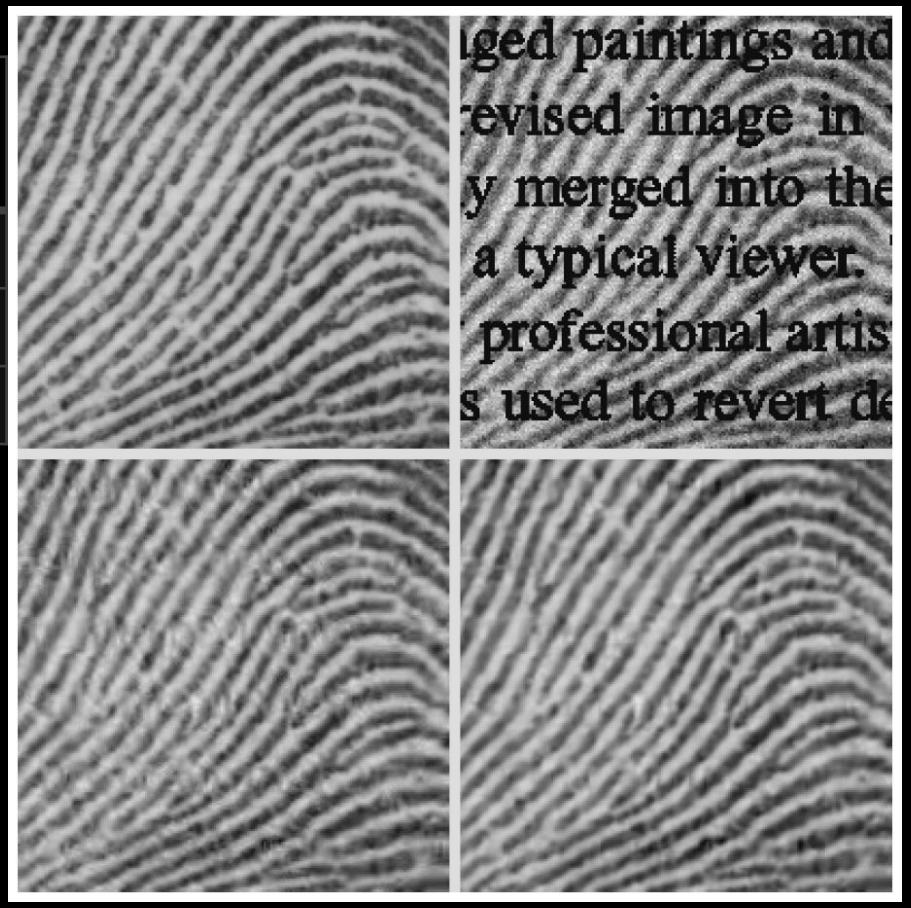
Inpainting [Mairal, E. & Sapiro ('08)]

For the `Peppers` image

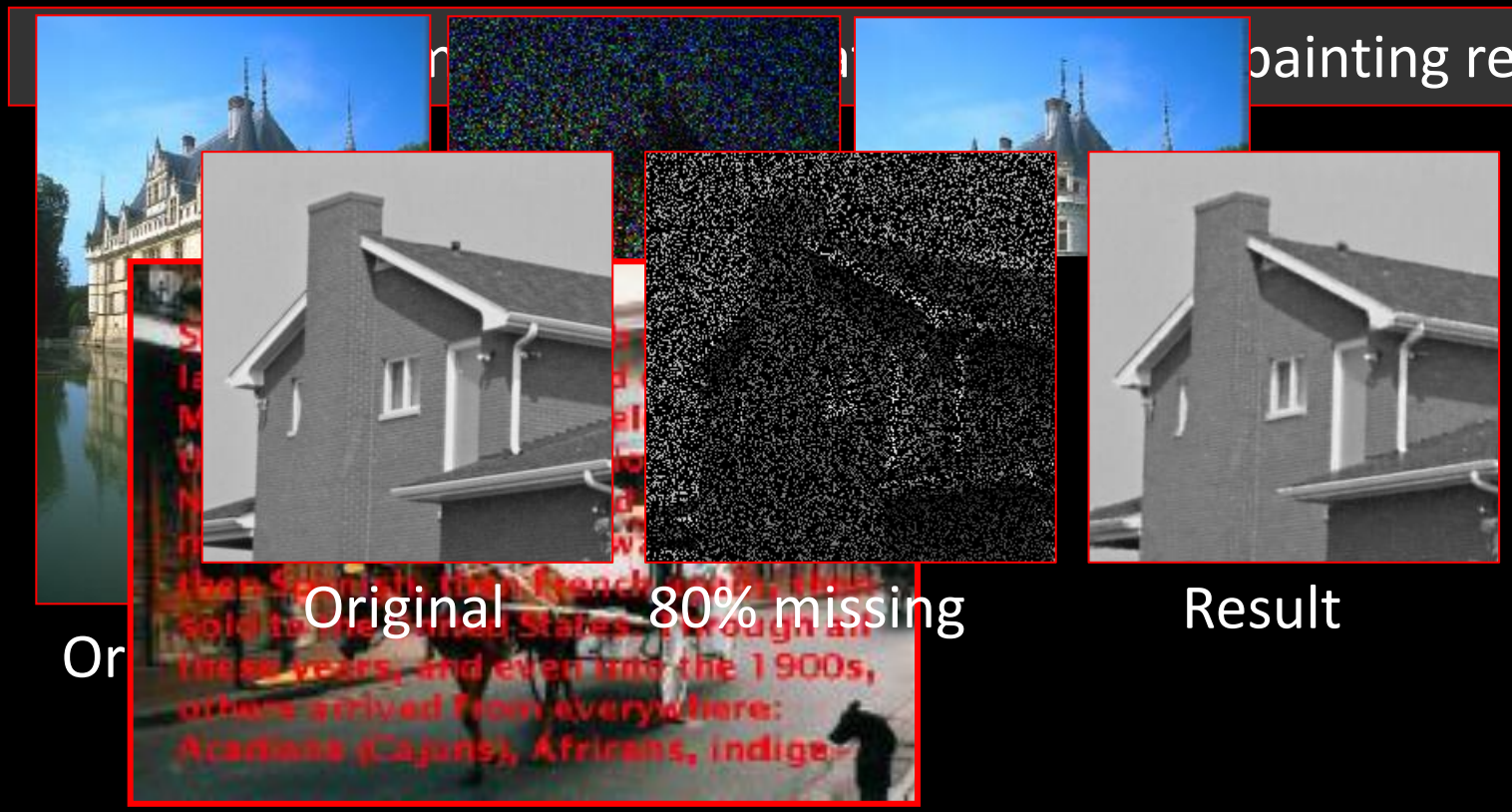
Alg.	RMSE for 25% missing	RMSE for 50% missing
DCT: No-overlap	14.55	19.61
DCT: Overlap	9.00	11.55
K-SVD	8.1	10.05

This is a more challenging case, where the DCT is not a suitable dictionary.

- For Redundant DCT we get RMSE=16.13, and
- For K-SVD (15 iterations) we get RMSE=12.74

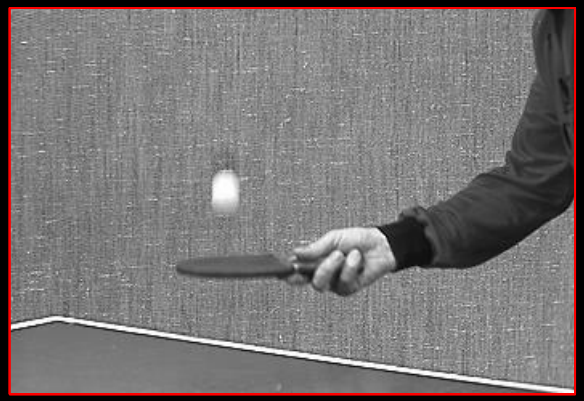


Inpainting [Mairal, E. & Sapiro ('08)]

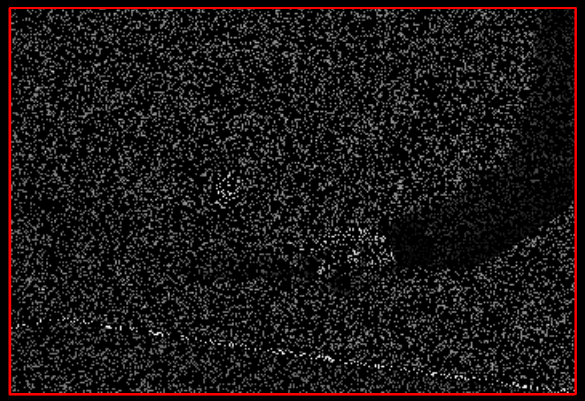


Inpainting [Mairal, E. & Sapiro ('08)]

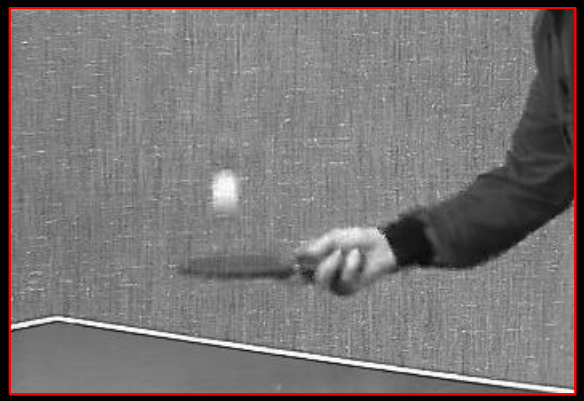
The same can be done for video, very much like the denoising treatment: (i) 3D patches, (ii) no need to compute the dictionary from scratch for each frame, and (iii) no need for explicit motion estimation



Original



80% missing

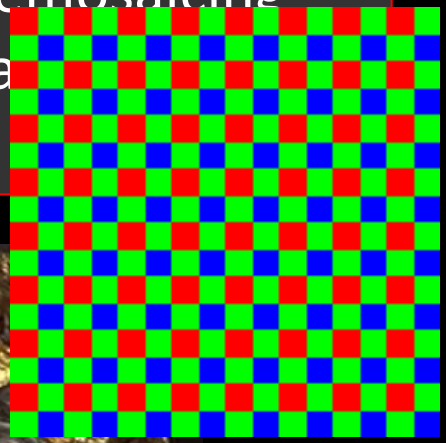


Result

Demosaicing [Mairal, E. & Sapiro ('08)]

Our experiments lead to state-of-the-art demosaicing results, giving ~0.2dB better results on a compared to [Chang & Chan ('06)]

- Today's cameras are sensing only one color per pixel, leaving the rest for interpolated
- Generalizing the inpainting scheme to handle demosaicing is tricky because of the possibility to learn the mosaic pattern within the dictionary
- In order to avoid "over-fitting", we handle the demosaicing problem while forcing strong sparsity and applying only few iterations



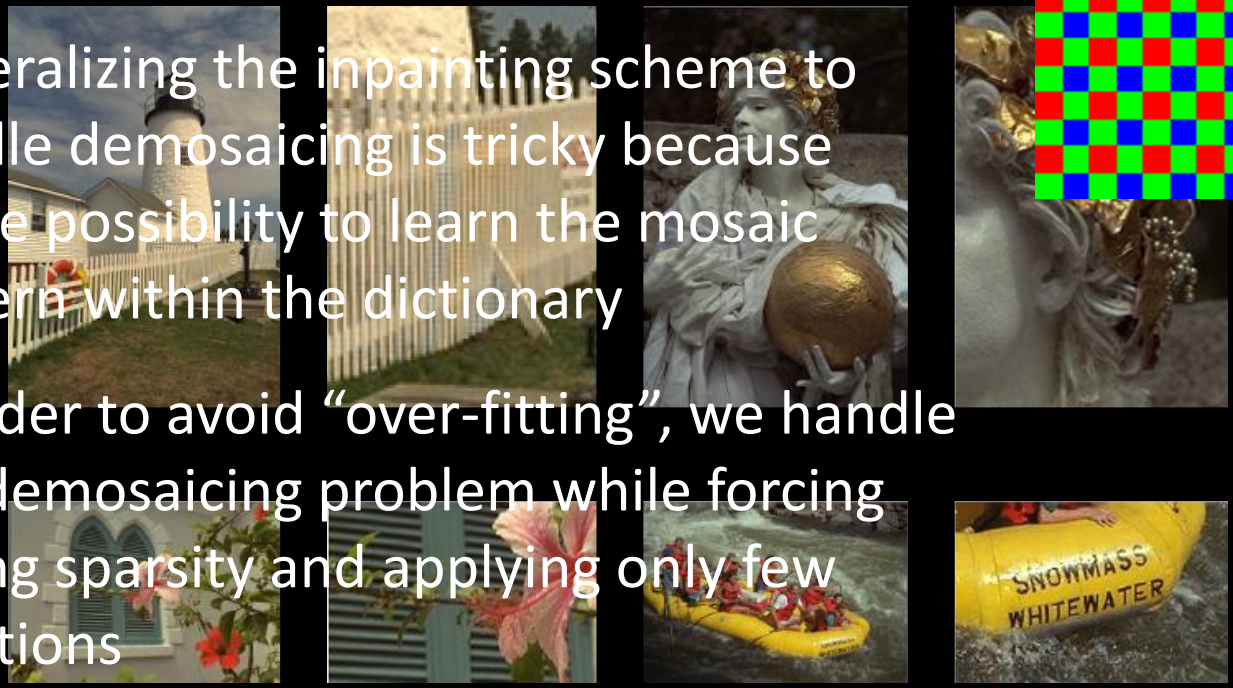


Image Compression [Bryt and E. ('08)]

- The problem: Compressing photo-ID images
- General purpose methods (JPEG, JPEG2000) do not take into account the specific family
- By adapting to the image-content (PCA/K-SVD), better results could be obtained
- For these techniques to operate well, train locally (per patch) using a training set of images is required
- In PCA, only the (quantized) coefficients are stored, the K-SVD requires storage of the indices
- Geometric alignment of the image is very helpful should be done [Goldenberg, Kimmel, & E. ('05)]

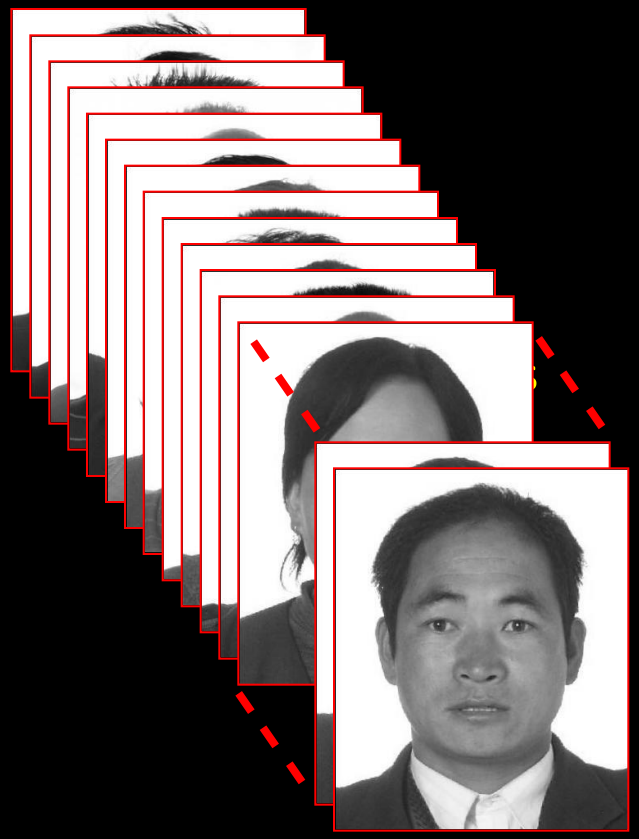


Image Compression

Detect main features and warp the images to a common reference (20 parameters)

↓

Divide the image into disjoint 15-by-15 patches. For each compute mean and dictionary

↓

Per each patch find the operating parameters (number of atoms L, quantization Q)

↓

Warp, remove the mean from each patch, sparse code using L atoms, apply Q, and dewarp

On the training set

On the test image

Training set (2500 images)

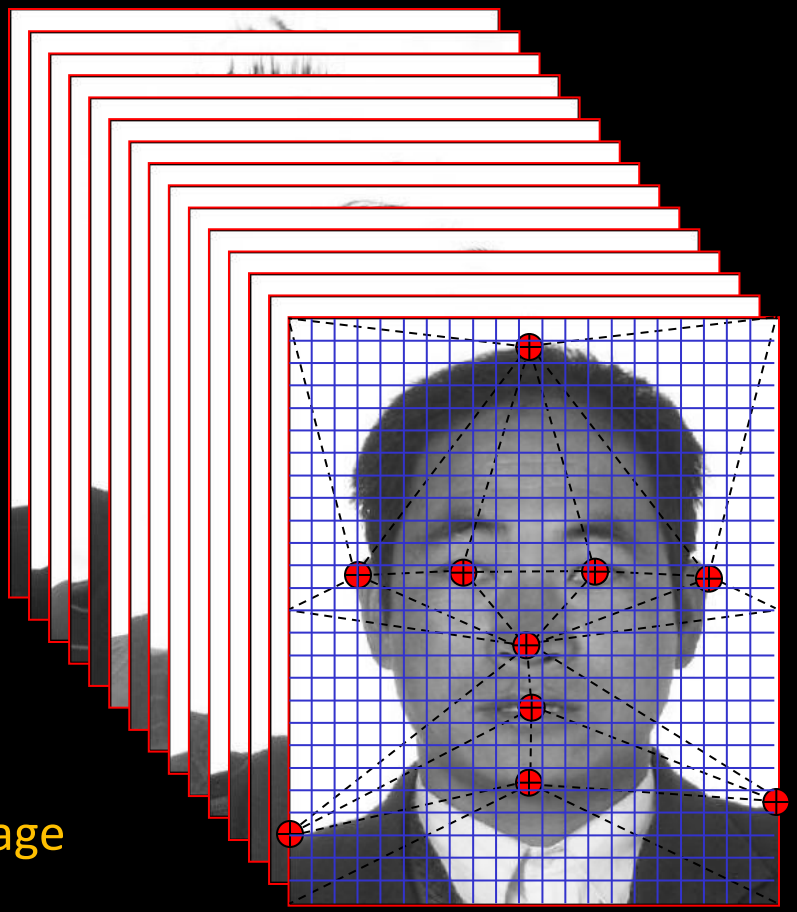


Image Compression Results

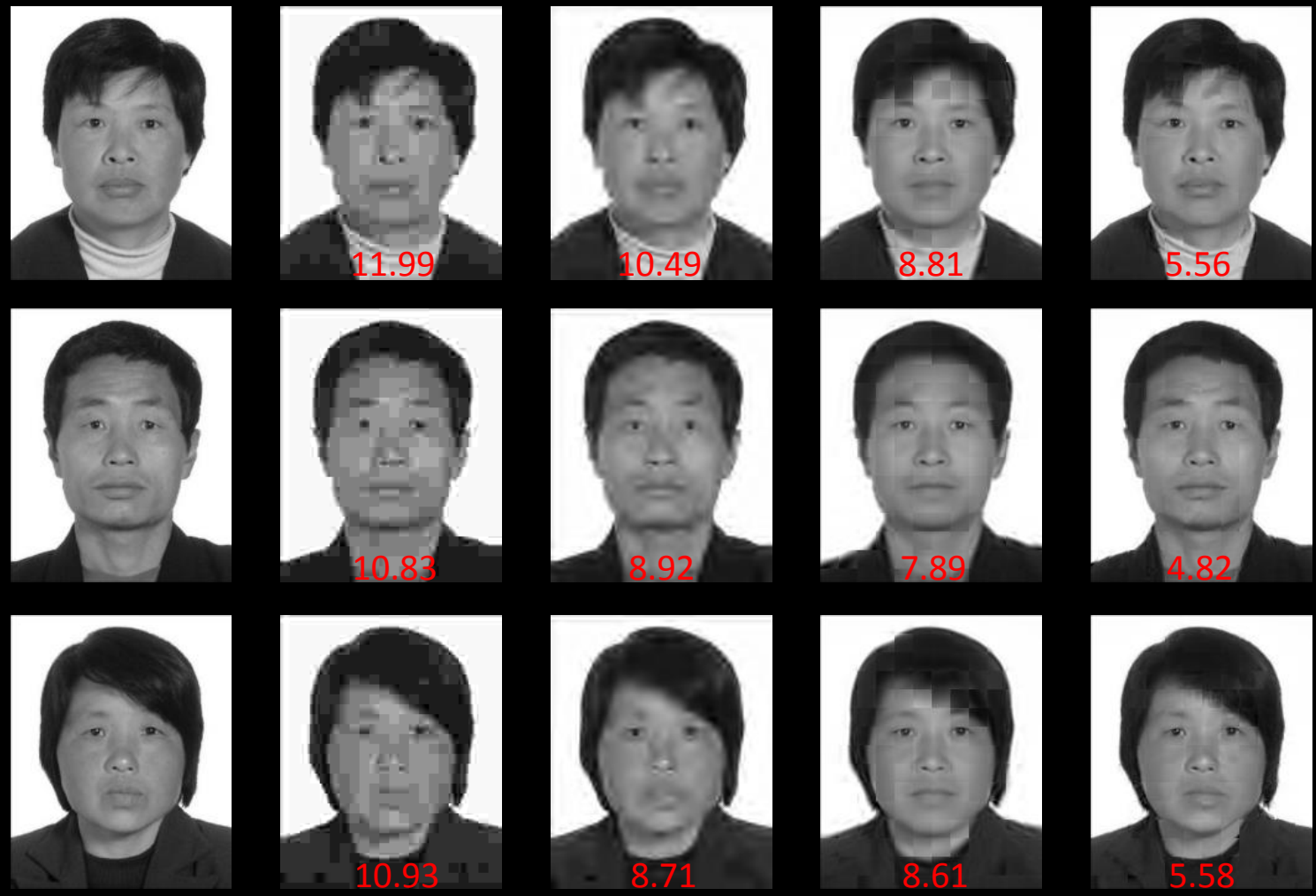


Image Compression Results

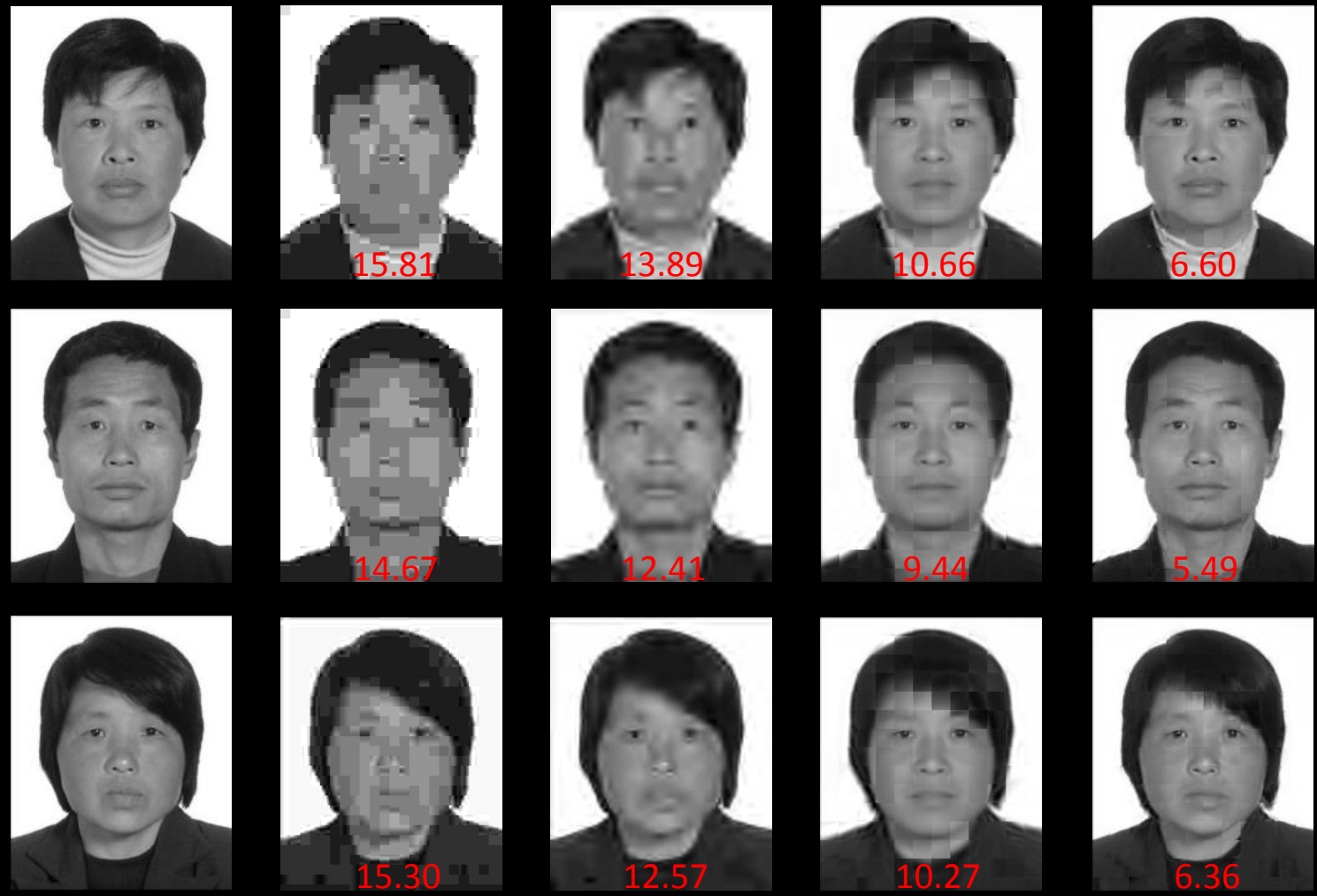


Image Compression Results

Deblocking the Results [Bryt and E. (`09)]

550 bytes K-SVD results with and without deblocking

K-SVD (6.60)

K-SVD (5.49)

K-SVD (6.45)

K-SVD (11.67)

Deblock (6.24)

Deblock (5.27)

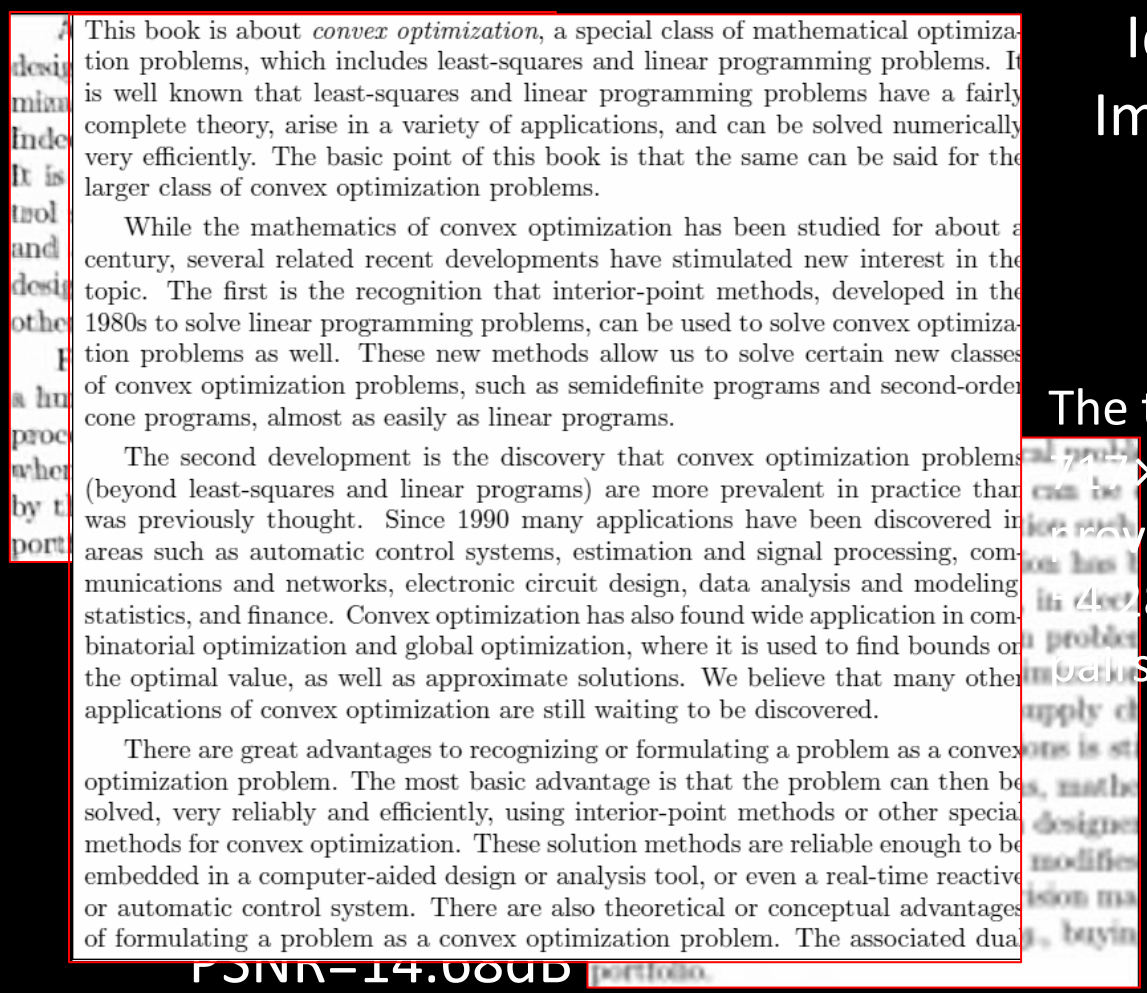
Deblock (6.03)

Deblock (11.32)

Super-Resolution [Zeyde, Protter, & E. ('11)]

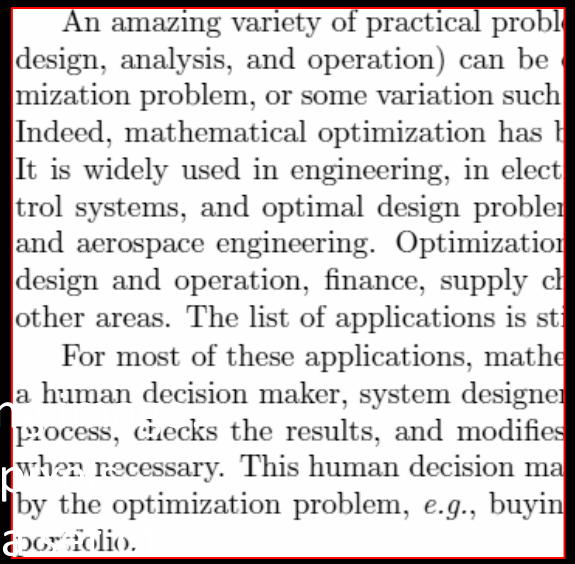
- Given a low-resolution image, we desire to enlarge it while producing a sharp looking result. This problem is referred to as "Single-Image Super-Resolution"
- Image scale-up using bicubic interpolation is far from being satisfactory for this task
- A brilliant and very different sparse and redundant representation technique was proposed [Yang, Wright, Huang, and Ma ('08)] for solving this problem, by training a coupled-dictionaries for the low- and high res. images
- We extended and improved their algorithms and results

Super-Resolution – Results (1)

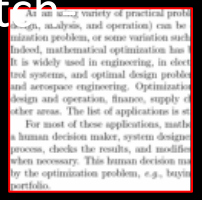


PSNR=14.68dB

Ideal Image



The trainin
×717 p
roviding a
89 training patch
s.



Given Image

Super-Resolution – Results (2)

Given image

Scaled-Up (factor 2:1) using the proposed algorithm,
PSNR=29.32dB (3.32dB improvement over bicubic)

Super-Resolution – Results (2)

The Original

Bicubic Interpolation

SR result

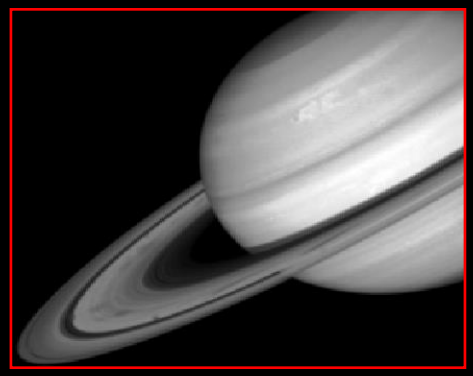
Super-Resolution – Results (2)

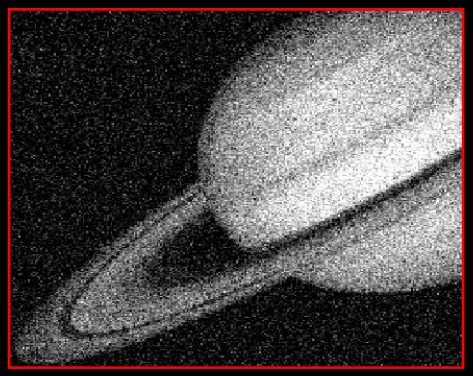
The Original

Bicubic Interpolation

SR result

Poisson Denoising


+

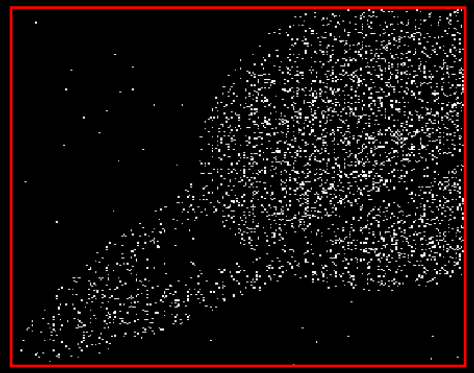
=


$$\underline{Y} = \underline{X} + \underline{V}$$

$$\underline{V} \sim \mathbb{N}\left(\underline{0}, \sigma^2 \mathbf{I}\right)$$

peak $= 100$

peak $= 0.1$



$$P(y \mid x) = \frac{x^y}{y!} e^{-x}$$

$$\text{peak} \triangleq \max_{i,j} \left\{ x_{i,j} \right\}$$

Poisson Denoising [Salmon et. al., 2011] [Giryes et. al., 2013]

- Anscombe transform converts Poisson distributed noise into an approximately Gaussian one, with variance 1 using the following formula [Anscombe, 1948]:

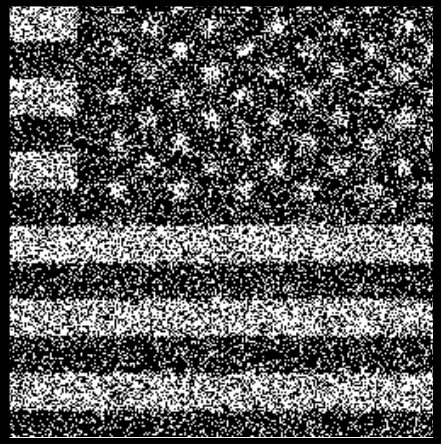
$$f_{\text{Anscombe}}\left(y\right)=2\sqrt{y+\frac{3}{8}}$$

- However, this is of reasonable accuracy only if peak>4.
- For lower peaks (poor illumination), we use the patch-based approach with dictionary learning, BUT … in the exponent domain:

$$\left\{\begin{array}{c}\underline{x}=\mathbf{D}\underline{\alpha}\\ \text{where}\ \ \left\|\underline{\alpha}\right\|_0\le L\end{array}\right\}\Rightarrow\left\{\begin{array}{c}\underline{x}=\exp\left\{\mathbf{D}\underline{\alpha}\right\}\\ \text{where}\ \ \left\|\underline{\alpha}\right\|_0\le L\end{array}\right\}$$

Poisson Denoising – Results (1)

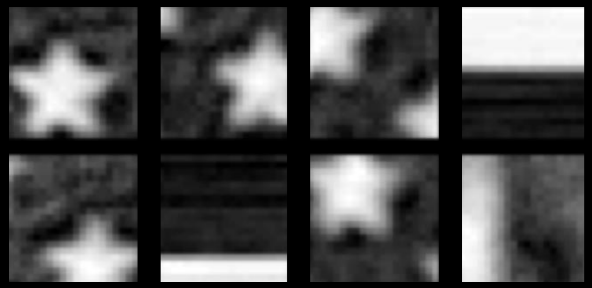
Original



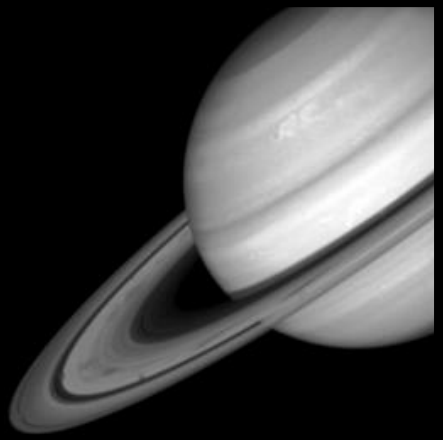
Noisy (peak=1)

Result (PSNR=22.59dB)

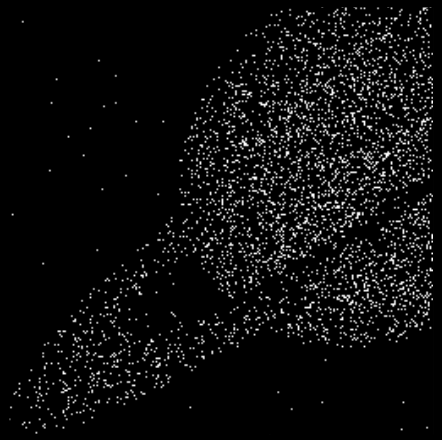
Dictionary learned atoms:



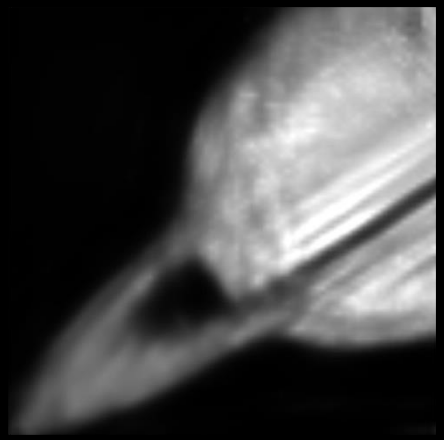
Poisson Denoising – Results (2)



Original



Noisy (peak=0.2)



Result (PSNR=24.16dB)

Original

Noisy (peak=2)

Result (PSNR=24.76dB)

Other Applications?

- Poisson Denoising & Inpainting
- Blind deblurring
- Audio inpainting
- Dynamic MRI reconstruction
- Clutter reduction in ultrasound
- Single image interpolation
- Anomaly detection
- …

Summary and Conclusion

In this Part we Have Seen that …

Sparsity, Redundancy, and the use of examples are important ideas that can be used in designing better tools in signal/image processing

What do we do?

In our work on we cover theoretical, numerical, and applicative issues related to this model and its use in practice

What next?

We keep working on:

- Improving the model
- Improve the dictionaries
- Demonstrate on other applications
- …

Next we focus on …

Many of the results we got focused on patch-based methods – it is time to understand better this choice and its limitations, with the hope to lead to new insights

Thank You

All this Work is Made Possible Due to

my teachers and mentors


A.M. Bruckstein


D.L. Donoho

colleagues & friends collaborating with me


G. Sapiro


J.L. Starck


I. Yavneh


M. Zibulevsky

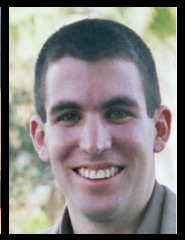

P. Milanfar

and my students


M. Aharon


O. Bryt


J. Mairal

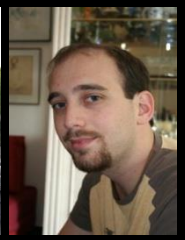

M. Protter


R. Rubinstein

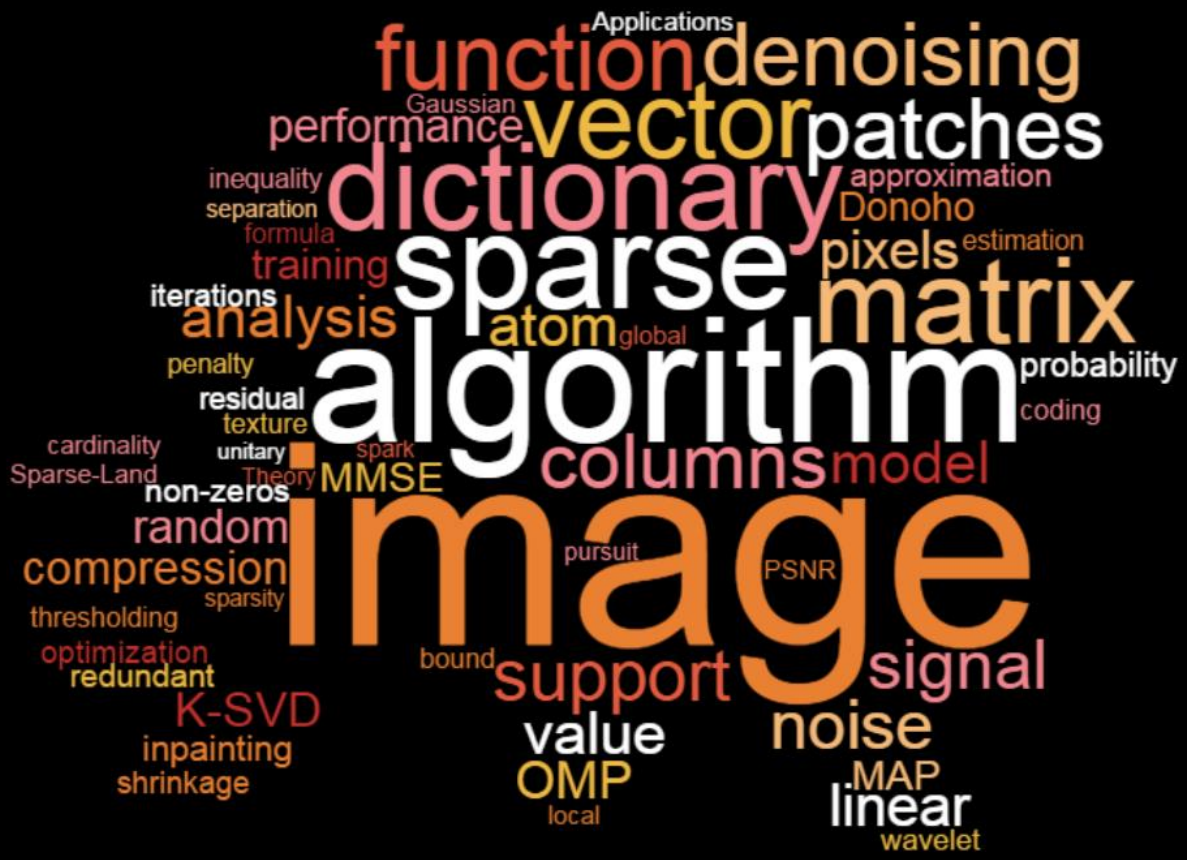

J. Shtok


R. Giryes


Z. Ben-Haim


J. Turek


R. Zeyde



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

Welcome to *Sparseland*

Part 3: A Tale of Three Models
Sparseland→CSC→CNN

Michael Elad

The 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

In This Talk

The Underlying Idea

Generative Modeling

of data sources enable

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

* Only CNN? What about other architectures ?

The Results Presented

… are the fruit of a joint work with

Yaniv Romano

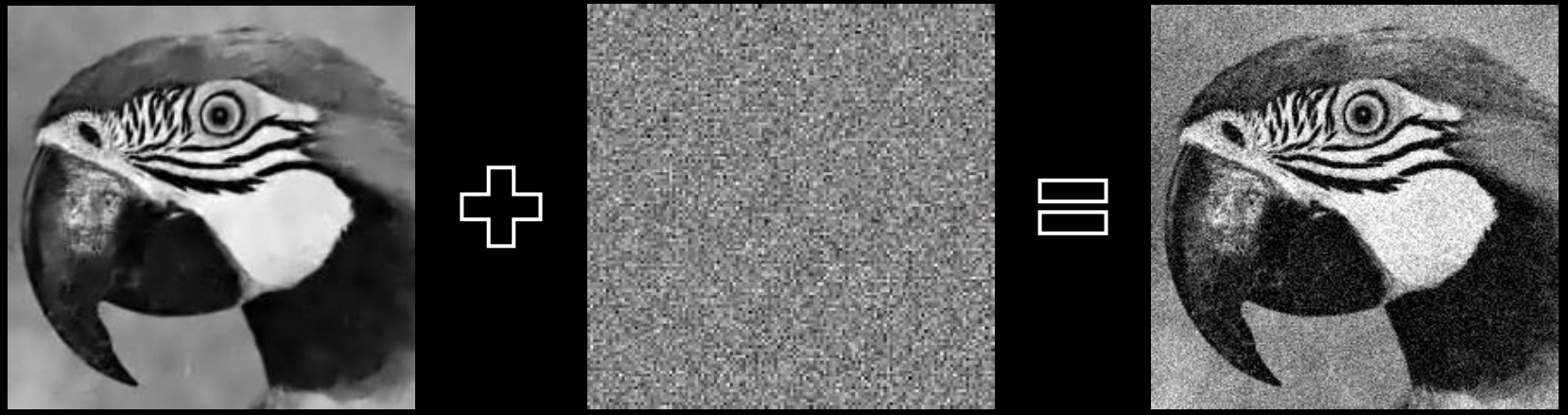
Vardan Papyan

Jeremias Sulam

Part I
Motivation and Background

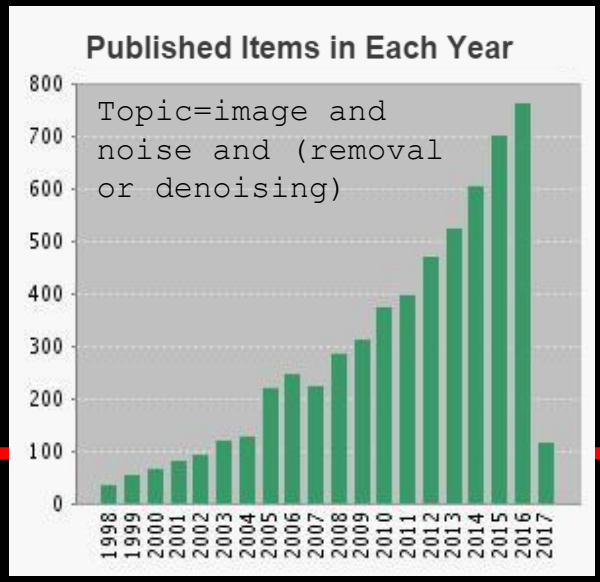
Our Starting Point: Image Denoising

Original Image $\mathbf{X}$ + White Gaussian Noise $\mathbf{E}$ = Noisy Image $\mathbf{Y}$



Many (thousands) image denoising algorithms have been proposed over the years, some of which are extremely effective

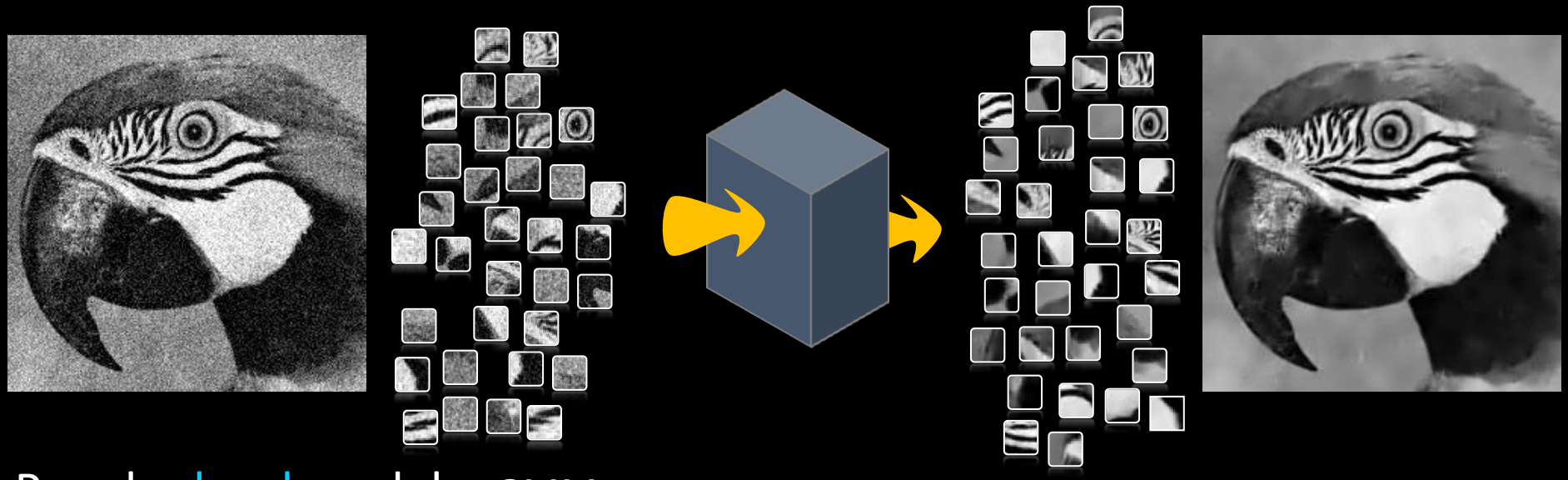
$\mathbf{Y}$ → Denoising Algorithm → $\hat{\mathbf{X}}$



Michael Elad
The Computer-Science Department
The Technion

Leading Image Denoising Methods…

are built upon powerful patch-based local models:

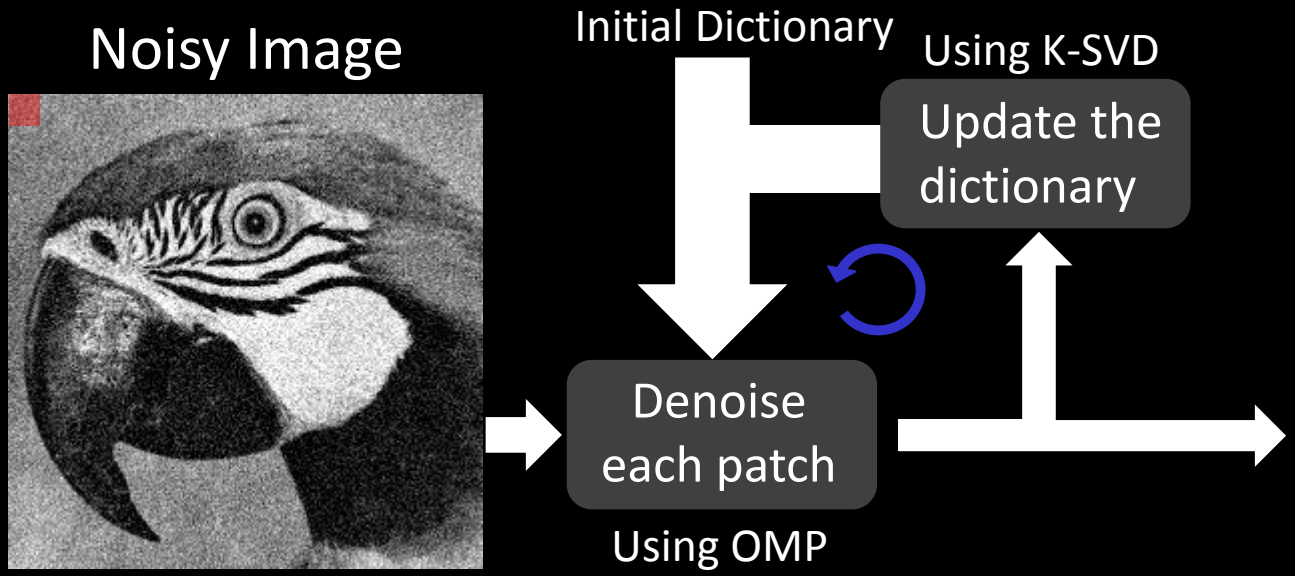


Popular local models:
- GMM
- Sparse-Representation
- Example-based
- Low-rank
- Field-of-Experts &
- Neural networks

Patch-Based Image Denoising

- K-SVD: sparse representation modeling of image patches [Elad & Aharon, '06]
- BM3D: combines sparsity and self-similarity [Dabov, Foi, Katkovnik & Egiazarian '07]
- EPLL: uses GMM of the image patches [Zoran & Weiss '11]
- MLP: multi-layer perceptron [Burger, Schuler & Harmeling '12]
- NCSR: non-local sparsity with centralized coefficients [Dong, Zhang, Shi & Li '13]
- WNNM: weighted nuclear norm of image patches [Gu, Zhang, Zuo & Feng '14]
- SSC–GSM: nonlocal sparsity with a GSM coefficient model [Dong, Shi, Ma & Li '15]

Recall K-SVD Denoising [Elad & Aharon, '06]



- Despite its simplicity, this is a very well-performing algorithm
- Its origins can be traced back to Guleryuz's local DCT recovery
- A small modification of this method leads to state-of-the-art results [Mairal, Bach, Ponce, Spairo, Zisserman, `09]

What is Missing?

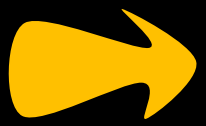
- Over the years, many kept revisiting this algorithm and its line of thinking, with a clear feeling that key features are still lacking
- What is missing? Here is what **WE** thought of…
  - A multi-scale treatment [Ophir, Lustig & Elad '11] [Sulam, Ophir & Elad '14] [Papyan & Elad '15]
  - Exploiting self-similarities [Ram & Elad '13] [Romano, Protter & Elad '14]
  - Pushing to better agreement on the overlaps [Romano & Elad '13] [Romano & Elad '15]
  - Enforcing the local model on the final patches (EPLL) [Sulam & Elad '15]
- Eventually, we realized that the key part that is missing is

A Theoretical Backbone

Missing Theoretical Backbone?

- The core global-local model assumption on $\mathbf{X} \in \mathbb{R}^N$:

$$\forall \mathrm{i} \quad \mathbf{R}_\mathrm{i}\mathbf{X} = \boldsymbol{\Omega}\boldsymbol{\gamma}_\mathrm{i} \quad \text{where} \quad \|\boldsymbol{\gamma}_\mathrm{i}\|_0 \leq \mathrm{k}$$



Every patch in the unknown signal is expected to have a sparse representation w.r.t. the same dictionary $\boldsymbol{\Omega}$

- Questions to consider:
  - Who are the signals belonging to this model? Do they exist?
  - How should we project a signal on this model (pursuit)?
  - Could we offer theoretical guarantees for this model/algorithms?
  - Could we offer a global pursuit algorithm that operates locally?
  - How should we learn $\boldsymbol{\Omega}$ if this is indeed the model?

- As we will see, all these questions are very relevant to recent developments in signal processing and machine learning

Coming Up

Limitations of patch averaging

Convolutional Sparse Coding (CSC) model

Theoretical study of CSC

Multi-Layer Convolutional Sparse Coding (ML-CSC)

Convolutional neural networks (CNN)

Fresh view of CNN through the eyes of sparsity

Part II
Convolutional Sparse Coding

(IEEE-TSP)

Working Locally Thinking Globally:

Theoretical Guarantees for Convolutional Sparse Coding

Vardan Papyan, Jeremias Sulam and Michael Elad

(ICCV 2017)

Convolutional Dictionary Learning via Local Processing

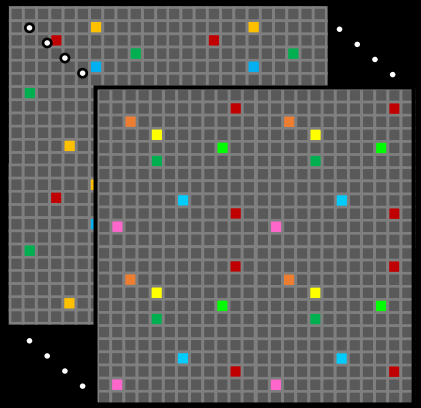
Vardan Papyan, Yaniv Romano, Jeremias Sulam, and Michael Elad

Convolutional Sparse Coding (CSC)

m filters convolved with their sparse representations

$$\mathbf{x} = \sum_{i=1}^{m} \mathbf{d}_i * \mathbf{z}_i$$

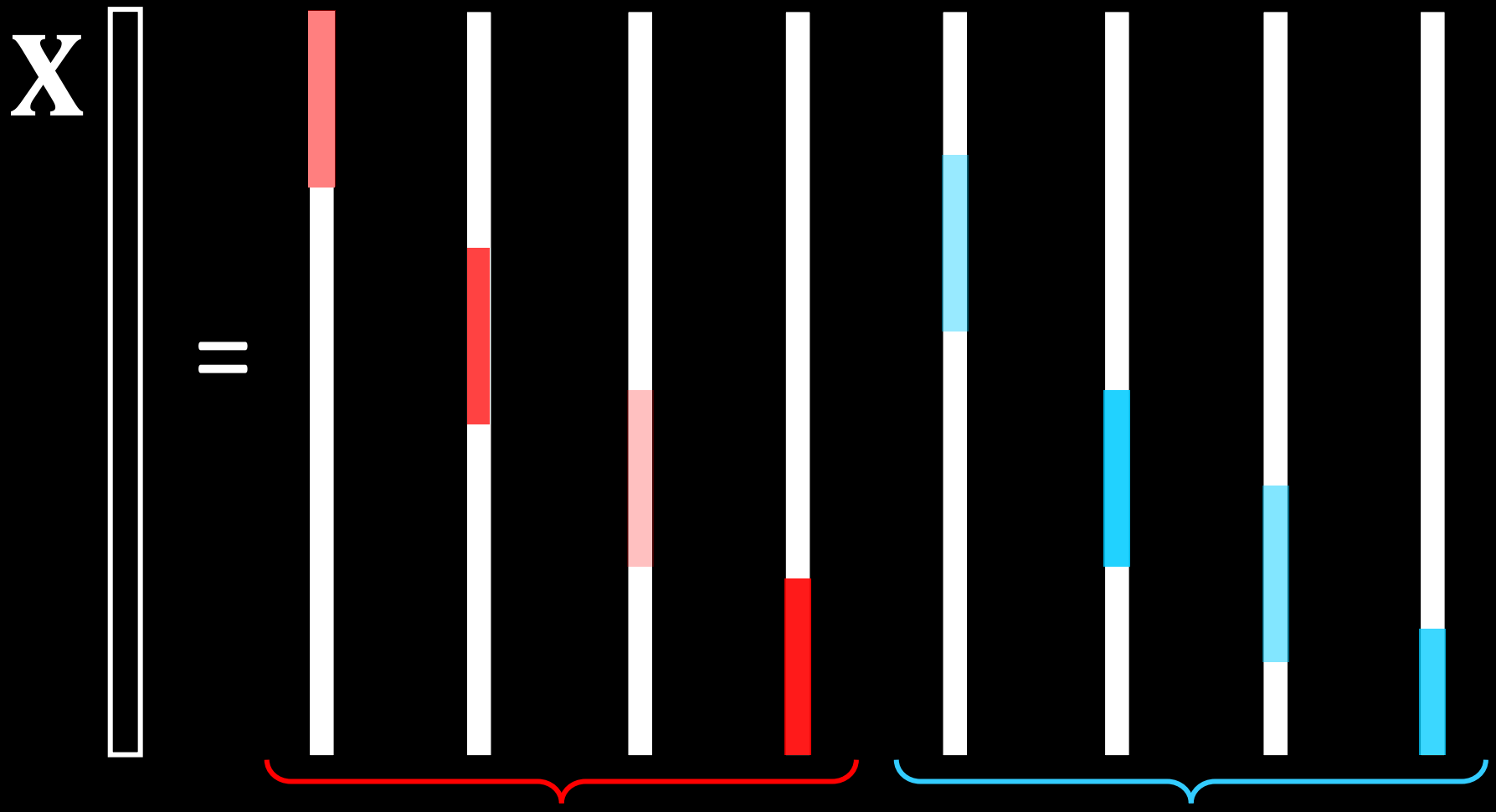
i-th feature-map: An image of the same size as $\mathbf{X}$ holding the sparse representation related to the i-filter



An image held as a column vector of length N

The j-th filter of small support n

Intuitively …

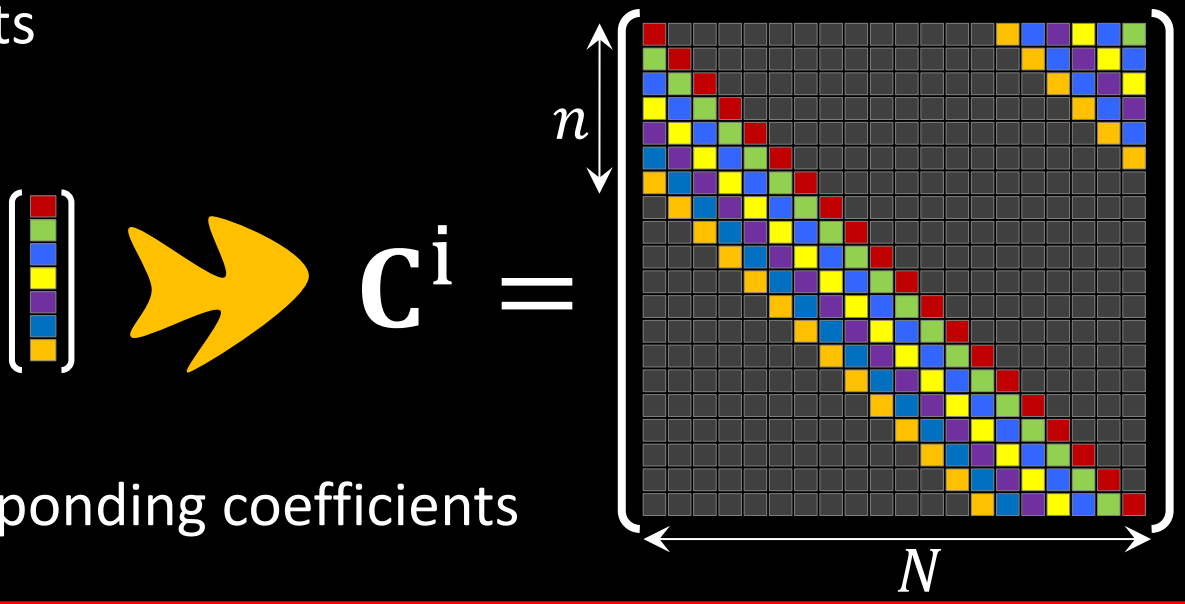


CSC in Matrix Form

- Here is an alternative global sparsity-based model formulation

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

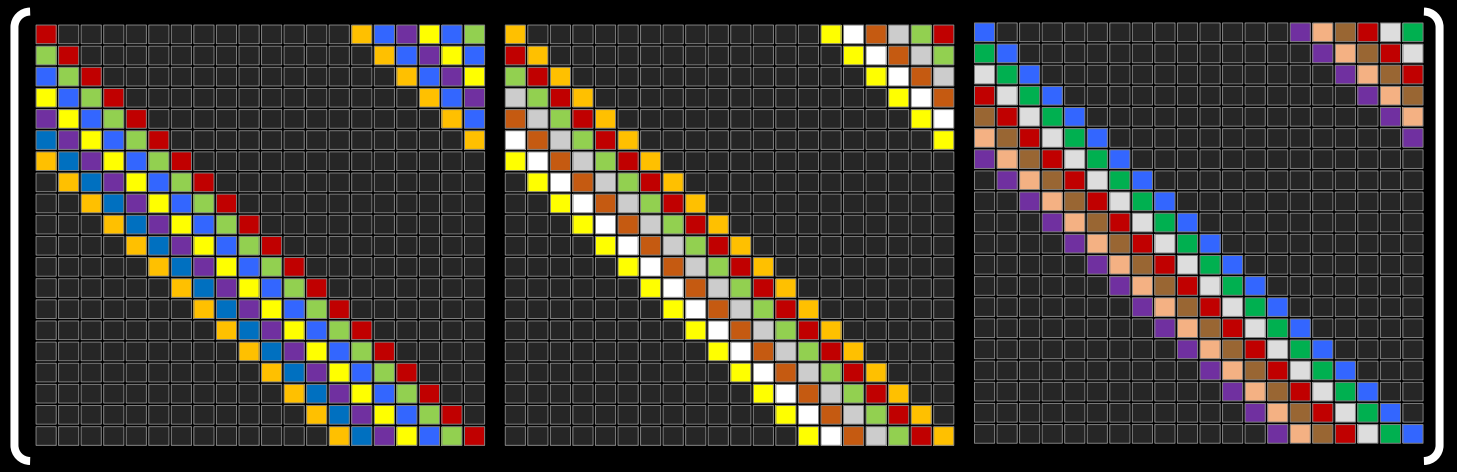
- $\mathbf{C}^i \in \mathbb{R}^{N \times N}$ is a banded and Circulant matrix containing a single atom with all of its shifts



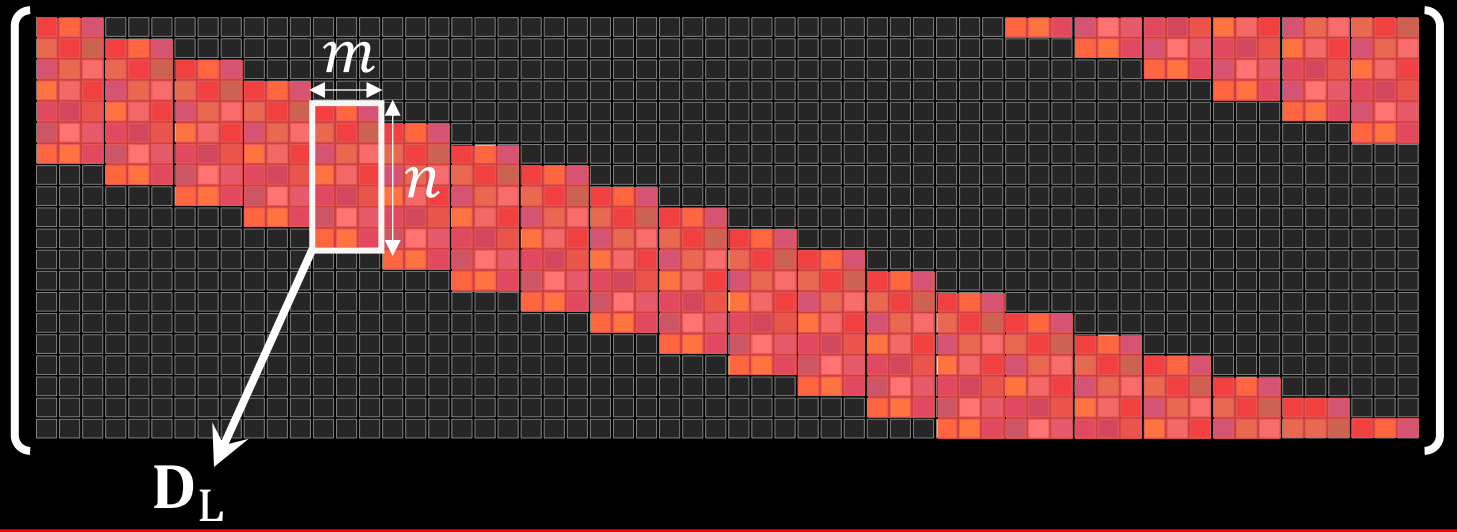
- $\boldsymbol{\Gamma}^i \in \mathbb{R}^N$ are the corresponding coefficients

Two Interpretations

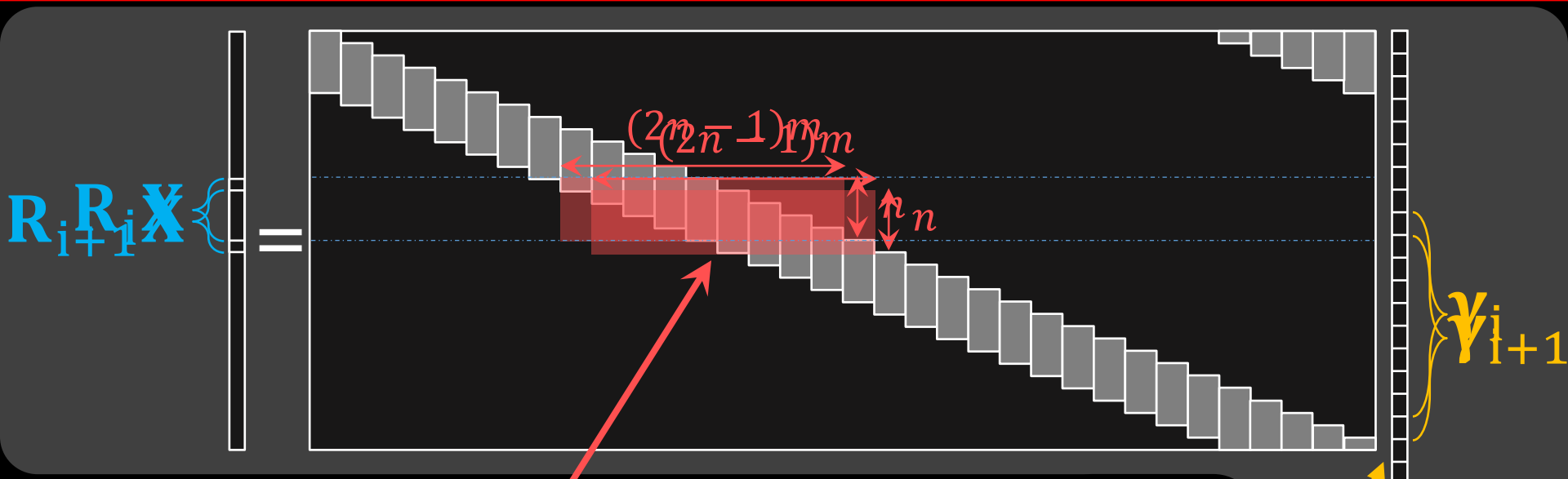
$[\mathbf{c}^1\ \mathbf{c}^2\ \mathbf{c}^3] =$



$\mathbf{D} =$



Why CSC?



stripe-dictionary

stripe vector

$$X = D\Gamma$$

$$R_i X = \Omega \gamma_i$$

Every patch has a sparse representation w.r.t. to the same local dictionary Ω, just as we have assumed

CSC Relation to Our Story

- A clear global model: every patch has a sparse representation w.r.t. to the same local dictionary $\boldsymbol{\Omega}$, just as we have assumed
- No notion of disagreement on the patch overlaps
- Related to the current common practice of patch averaging ($\mathbf{R}_i^T$ - put the patch $\boldsymbol{\Omega}\boldsymbol{\gamma}_i$ back in the i-th location of the global vector)

$$\mathbf{X} = \mathbf{D}\boldsymbol{\Gamma} = \frac{1}{n}\sum_i \mathbf{R}_i^T \boldsymbol{\Omega}\boldsymbol{\gamma}_i$$

- What about the Pursuit?
  - "Patch averaging": independent sparse coding for each patch
  - CSC: should seek all the representations together
- Is there a bridge between the two? We'll come back to this later …

PREVIOUS WORK

- This model has been used in the past [Lewicki & Sejnowski '99] [Hashimoto & Kurata, '00]
- Most works have focused on solving *efficiently* its associated pursuit, called **convolutional sparse coding**, using the BP algorithm

$$(\mathbf{P}_1^{\epsilon}): \quad \min_{\boldsymbol{\Gamma}} \quad \|\boldsymbol{\Gamma}\|_1 + \lambda\|\mathbf{Y} - \mathbf{D}\boldsymbol{\Gamma}\|_2^2$$

Convolutional dictionary

- Several applications were demonstrated:
  - Pattern detection in images and the analysis of instruments in music signals [Mørup, Schmidt & Hansen '08]
  - Inpainting [Heide, Heidrich & Wetzstein '15]
  - Super-resolution [Gu, Zuo, Xie, Meng, Feng & Zhang '15]
- However, little is known regrading its theoretical aspects. Why? Perhaps because the regular SparsLand theory is sufficient?

Classical Sparse Theory (Noiseless)

$$(\mathbf{P}_0): \quad \min_{\boldsymbol{\Gamma}} \quad \|\boldsymbol{\Gamma}\|_0 \quad \text{s.t.} \quad \mathbf{X} = \mathbf{D}\boldsymbol{\Gamma}$$

Definition: Mutual-Coherence: $\mu(\mathbf{D}) = \max_{i \neq j} |d_i^T d_j|$ [Donoho & Elad '03]

Theorem: For a signal $\mathbf{X} = \mathbf{D}\boldsymbol{\Gamma}$, if $\|\boldsymbol{\Gamma}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right)$ then this solution is necessarily the sparsest

[Donoho & Elad '03]

Theorem: The OMP and BP are guaranteed to recover the true sparse code assuming that $\|\boldsymbol{\Gamma}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right)$

[Tropp '04], [Donoho & Elad '03]

The Need for a Theoretical Study

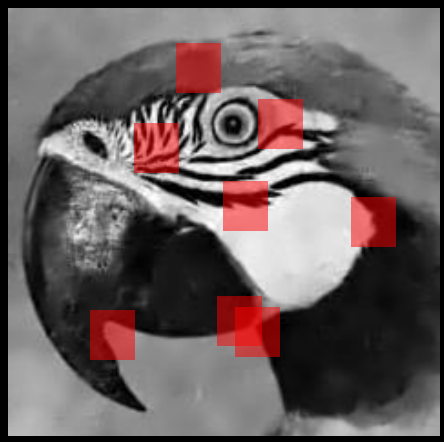
- Assuming that $m = 2$ and $n = 64$ we have that [Welch, '74]

$$\mu(\mathbf{D}) \geq 0.063$$

- As a result, uniqueness and success of pursuits is guaranteed as long as

$$\|\boldsymbol{\Gamma}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right) \leq \frac{1}{2}\left(1 + \frac{1}{0.063}\right) \approx 8$$

- Less than 8 non-zeros GLOBALLY are allowed!!! This is a very pessimistic result!
- Repeating the above for the noisy case leads to even worse performance predictions
- Bottom line: Classic SparseLand Theory cannot provide good explanations for the CSC model

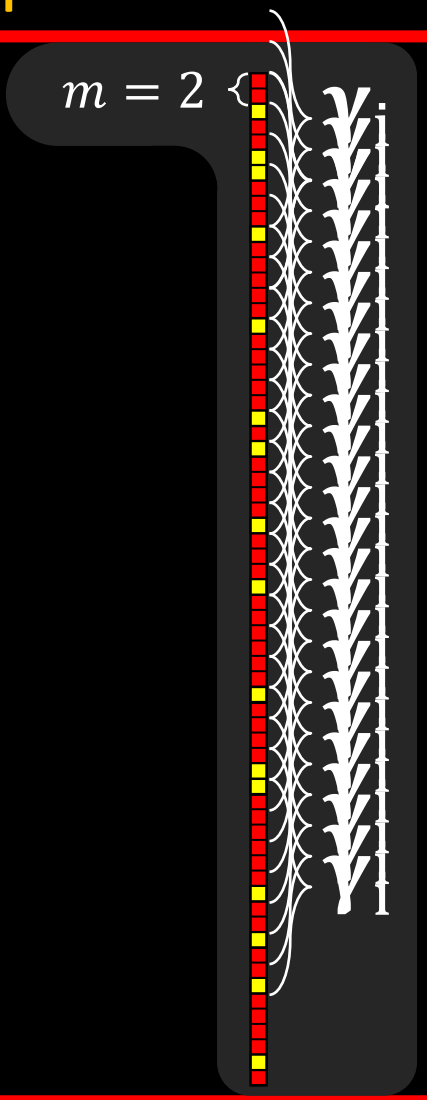


Moving to Local Sparsity: Stripes

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

$$\left(\mathbf{P}_{0,\infty}\right): \quad \min_{\boldsymbol{\Gamma}} \|\boldsymbol{\Gamma}\|_{0,\infty}^{s} \text{ s.t. } \mathbf{X} = \mathbf{D}\boldsymbol{\Gamma}$$

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



The Main Questions we Aim to Address:

I. Is the solution to this problem unique ?

II. Can we recover the solution via a global OMP/BP ?

Stripe-Spark and Uniqueness

$$\left(\mathbf{P}_{0,\infty}\right): \quad \min_{\boldsymbol{\Gamma}} \ \|\boldsymbol{\Gamma}\|_{0,\infty}^{s} \ \text{s.t.} \ \mathbf{X} = \mathbf{D}\boldsymbol{\Gamma}$$

Definition: Stripe Spark $\eta_{\infty}(\mathbf{D}) = \min_{\boldsymbol{\Delta}} \ \|\boldsymbol{\Delta}\|_{0,\infty}^{s} \ \text{s.t.} \ \begin{Bmatrix} \mathbf{D}\boldsymbol{\Delta} = 0 \\ \boldsymbol{\Delta} \neq 0 \end{Bmatrix}$

Theorem: If a solution $\boldsymbol{\Gamma}$ is found for $\left(\mathbf{P}_{0,\infty}\right)$ such that:

$$\|\boldsymbol{\Gamma}\|_{0,\infty}^{s} < \frac{1}{2}\eta_{\infty}$$

then it is necessarily the optimal solution to this problem

Theorem: The relation between the Stripe-Spark and the Mutual Coherence is:

$$\eta_{\infty}(\mathbf{D}) \geq 1 + \frac{1}{\mu(\mathbf{D})}$$

Uniqueness via Mutual Coherence

$$\left(\mathbf{P}_{0,\infty}\right): \quad \min_{\boldsymbol{\Gamma}} \ \|\boldsymbol{\Gamma}\|_{0,\infty}^{s} \ \text{s.t.} \ \mathbf{X} = \mathbf{D}\boldsymbol{\Gamma}$$

Theorem: If a solution $\boldsymbol{\Gamma}$ is found for $\left(\mathbf{P}_{0,\infty}\right)$ such that:

$$\|\boldsymbol{\Gamma}\|_{0,\infty}^{s} < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right)$$

then this is necessarily the unique optimal solution to this problem

This result is exciting: This and later results pose a **local** constraint for a **global** guarantee, and as such, they are far more optimistic compared to the **global** guarantees

For k non-zeros per stripe, and filters of length n, we get

$$\|\boldsymbol{\Gamma}\|_0 \cong \frac{k}{2n-1} \cdot N$$

non-zeros globally

Recovery Guarantees

$$\left(\mathbf{P}_{0,\infty}\right): \quad \min_{\mathbf{\Gamma}} \ \|\mathbf{\Gamma}\|_{0,\infty}^{s} \ \text{s.t.} \ \mathbf{X} = \mathbf{D}\mathbf{\Gamma}$$

Lets solve this problem via OMP or BP*, applied globally

Theorem: If a solution $\mathbf{\Gamma}$ of $\left(\mathbf{P}_{0,\infty}\right)$ satisfies:

$$\|\mathbf{\Gamma}\|_{0,\infty}^{s} < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right)$$

then **global** OMP and BP are guaranteed to find it

Both OMP and BP do not assume **local** sparsity but still guaranteed to succeed. One could propose algorithms that rely on this assumption

* How about variants that would exploit the local sparsity?

From Ideal to Noisy Signals

- So far, we have assumed an ideal signal $\mathbf{X} = \mathbf{D\Gamma}$
- However, in practice we usually have $\mathbf{Y} = \mathbf{D\Gamma} + \mathbf{E}$ where $\mathbf{E}$ is due to noise or model deviations
- To handle this, we redefine our problem as:

$$\left(\mathbf{P}_{0,\infty}^{\epsilon}\right): \quad \min_{\mathbf{\Gamma}} \quad \|\mathbf{\Gamma}\|_{0,\infty}^{s} \quad \text{s.t.} \quad \|\mathbf{Y} - \mathbf{D\Gamma}\|_2 \leq \epsilon$$

- **The Main Questions We Aim to Address:**
  I. Stability of the solution to this problem ?
  II. Stability of the solution obtained via global OMP/BP ?
  III. Could the same recovery be done via local (patch) operations ?

Stability of via Stripe-RIP

$$\left(\mathbf{P}_{0,\infty}^{\epsilon}\right): \quad \min_{\boldsymbol{\Gamma}} \ \|\boldsymbol{\Gamma}\|_{0,\infty}^{s} \ \text{s.t.} \ \|\mathbf{Y}-\mathbf{D}\boldsymbol{\Gamma}\|_2 \le \epsilon \quad \Rightarrow \quad \widehat{\boldsymbol{\Gamma}}$$

Theorem: If the true representation $\boldsymbol{\Gamma}$ satisfies

$$\|\boldsymbol{\Gamma}\|_{0,\infty}^{s} = k < \frac{1}{2}\left(1+\frac{1}{\mu(\mathbf{D})}\right)$$

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

$$\left\|\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}\right\|_2^2 \le \frac{4\epsilon^2}{1-(2k-1)\mu(\mathbf{D})}$$

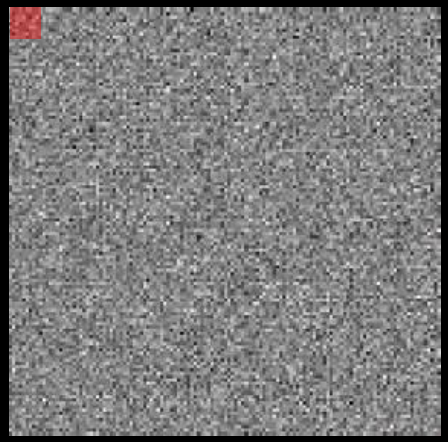
If you carefully review this result, you should be disappointed, as we see a noise magnification !! Is this true ?

Answer: No!!! This is a worst-case (with adversarial noise) analysis

Local Noise Assumption

- Thus far, our analysis relied on the local sparsity of the underlying solution $\boldsymbol{\Gamma}$, which was enforced through the $\ell_{0,\infty}$ norm
- In what follows, we present stability guarantees for both OMP and BP that will also depend on the local energy in the noise vector E
- This will be enforced via the $\ell_{2,\infty}$ norm, defined as:

$$\|\mathbf{E}\|_{2,\infty}^{\mathrm{p}} = \max_{\mathrm{i}} \|\mathbf{R}_{\mathrm{i}}\mathbf{E}\|_2$$



Stability of OMP

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

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

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

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

Stability of Lagrangian BP

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

Theorem: For $\mathbf{Y} = \mathbf{D}\boldsymbol{\Gamma} + \mathbf{E}$, if $\lambda = 4\|\mathbf{E}\|^{\mathrm{p}}$ and

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

Then we are guaranteed that

1. The support of $\boldsymbol{\Gamma}_{\mathrm{BP}}$ is contain
2. $\|\boldsymbol{\Gamma}_{\mathrm{BP}} - \boldsymbol{\Gamma}\|_{\infty} \leq 7.5\|\mathbf{E}\|_{2,\infty}^{\mathrm{p}}$
3. Every entry greater than $7.5\|$
4. $\boldsymbol{\Gamma}_{\mathrm{BP}}$ is unique

Theoretical foundation for 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]

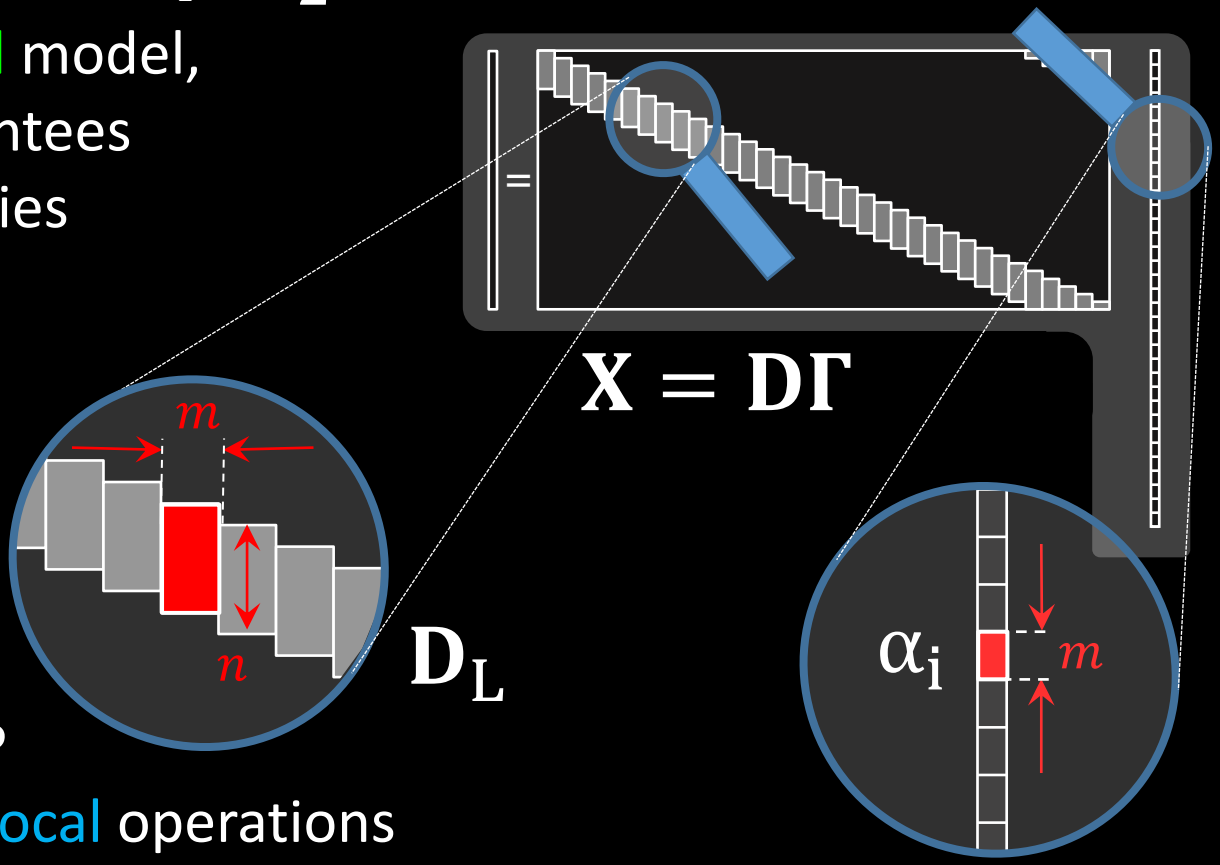
Proof relies on the work of [Tropp '06]

Global Pursuit via Local Processing

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

- While CSC is a global model, its theoretical guarantees rely on local properties

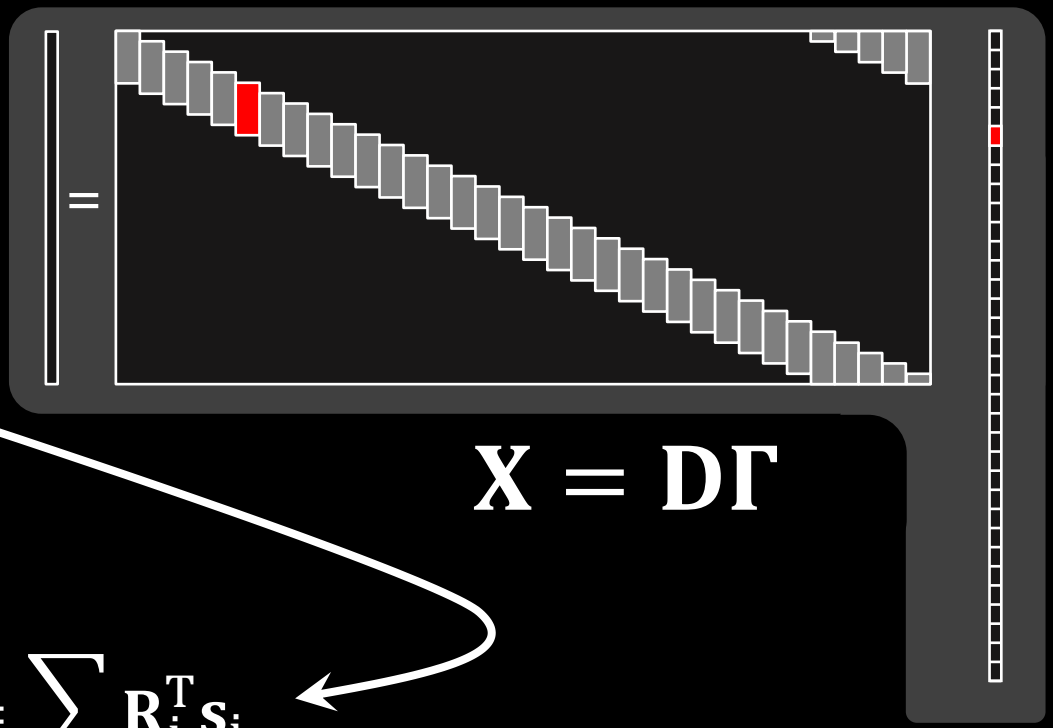
- We aim to show that this global-local relation can also be exploited for solving the global BP problem using only local operations



Global Pursuit via Local Processing (2)

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

- $\mathbf{s}_i$ are slices – local patches that overlap to form the full image



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

$$\mathbf{X} = \mathbf{D}\boldsymbol{\Gamma} = \sum_i \mathbf{R}_i^T \mathbf{D}_L \boldsymbol{\alpha}_i = \sum_i \mathbf{R}_i^T \mathbf{s}_i$$

Global Pursuit via Local Processing (2)

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

Turning to the local form and using the Augmented Lagrangian

$$\min_{\boldsymbol{\alpha}_i,\mathbf{s}_i} \ \frac{1}{2}\left\|\mathbf{Y}-\sum_i \mathbf{R}_i^T\mathbf{s}_i\right\|_2^2 + \sum_i\left(\lambda\|\boldsymbol{\alpha}_i\|_1 + \frac{\rho}{2}\|\mathbf{s}_i-\mathbf{D}_L\boldsymbol{\alpha}_i+u_i\|_2^2\right)$$

- These two problems are equivalent, and convex w.r.t their variables
- The new formulation targets the local slices, and their sparse representations
- The vectors u_i are the Lagrange multipliers for the constraints $\mathbf{s}_i=\mathbf{D}_L\boldsymbol{\alpha}_i$

Global Pursuit via Local Processing (2)

$$\min_{\boldsymbol{\alpha}_i,\mathbf{s}_i} \frac{1}{2}\left\|\mathbf{Y}-\sum_i \mathbf{R}_i^T\mathbf{s}_i\right\|_2^2 + \sum_i\left(\lambda\|\boldsymbol{\alpha}_i\|_1 + \frac{\rho}{2}\|\mathbf{s}_i-\mathbf{D}_L\boldsymbol{\alpha}_i+\mathbf{u}_i\|_2^2\right)$$

ADMM

- Slice-update: $\min_{\mathbf{s}_i} \frac{1}{2}\left\|\mathbf{Y}-\sum_i \mathbf{R}_i^T\mathbf{s}_i\right\|^2 + \sum_i \frac{\rho}{2}\|\mathbf{s}_i-\mathbf{D}_L\boldsymbol{\alpha}_i+\mathbf{u}_i\|_2^2$

  nd averaging

- Sparse-Upda $\sum_i\left(\lambda\|\boldsymbol{\alpha}_i\|_1 + \frac{\rho}{2}\|\mathbf{s}_i-\mathbf{D}_L\boldsymbol{\alpha}_i+\mathbf{u}_i\|_2^2\right)$

  Separable and local LARS problems

This algorithm operates locally while guaranteeing to solve the global problem

Comment: One iteration of this procedure amounts to … the very same patch-averaging algorithm we started with

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 ($\mathbf{D}_L$) Learning

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

[Wohlberg, 2016] Ours

Partial Summary of CSC

- What we have seen so far is a new way to analyze the global CSC model using local sparsity constraints. We proved:

Uniqueness of the solution for the noiseless problem

Stability of the solution for the noisy problem

Guarantee of success and stability of both OMP and BP

We obtained guarantees and algorithms that operate locally while claiming global optimality

We mentioned briefly the mater of learning the model (i.e. dictionary learning for CSC), and presented our competitive approach which is based on simple local steps

Part III
Going Deeper

(JMLR 2017)

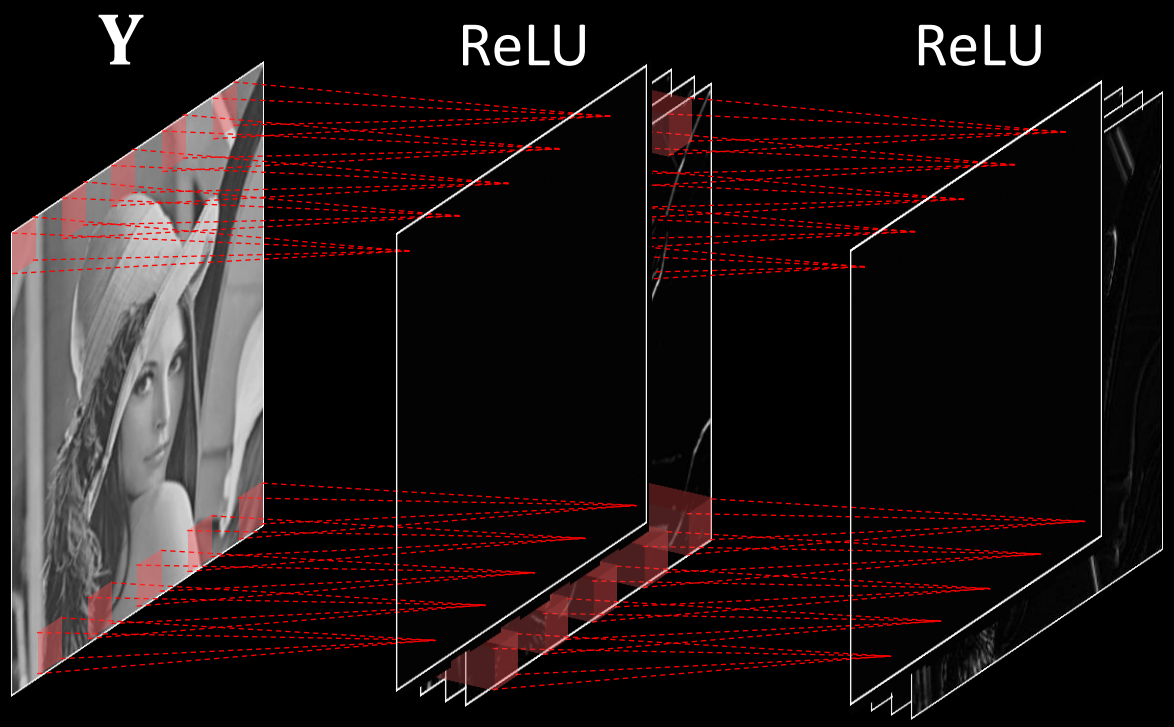
Convolutional Neural Networks Analyzed via Convolutional Sparse Coding

Vardan Papyan, Yaniv Romano and Michael Elad

CSC and CNN

- There seems to be a connection between CSC and CNN:
  - Convolutional structure
  - Data driven models
  - ReLU is a sparsifying operator
- We propose a principled way to analyze CNN
- But first, a short review of CNN…

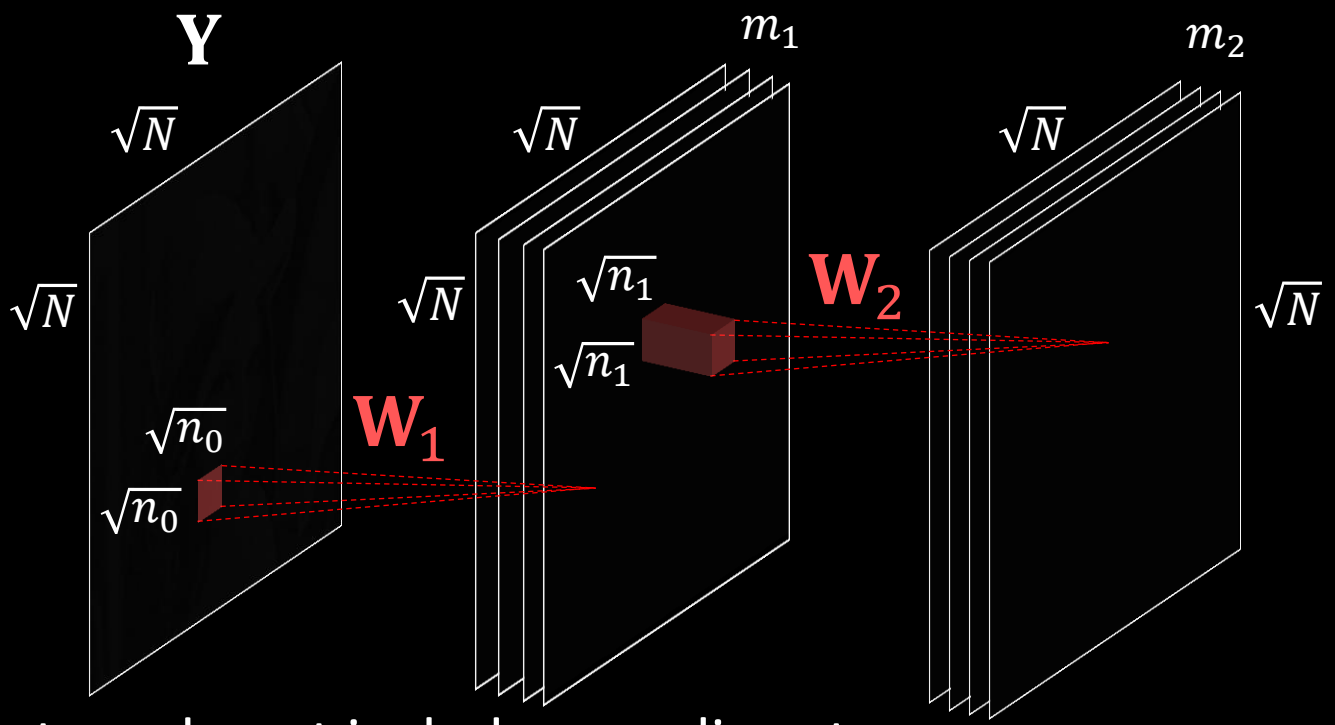
CNN



[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



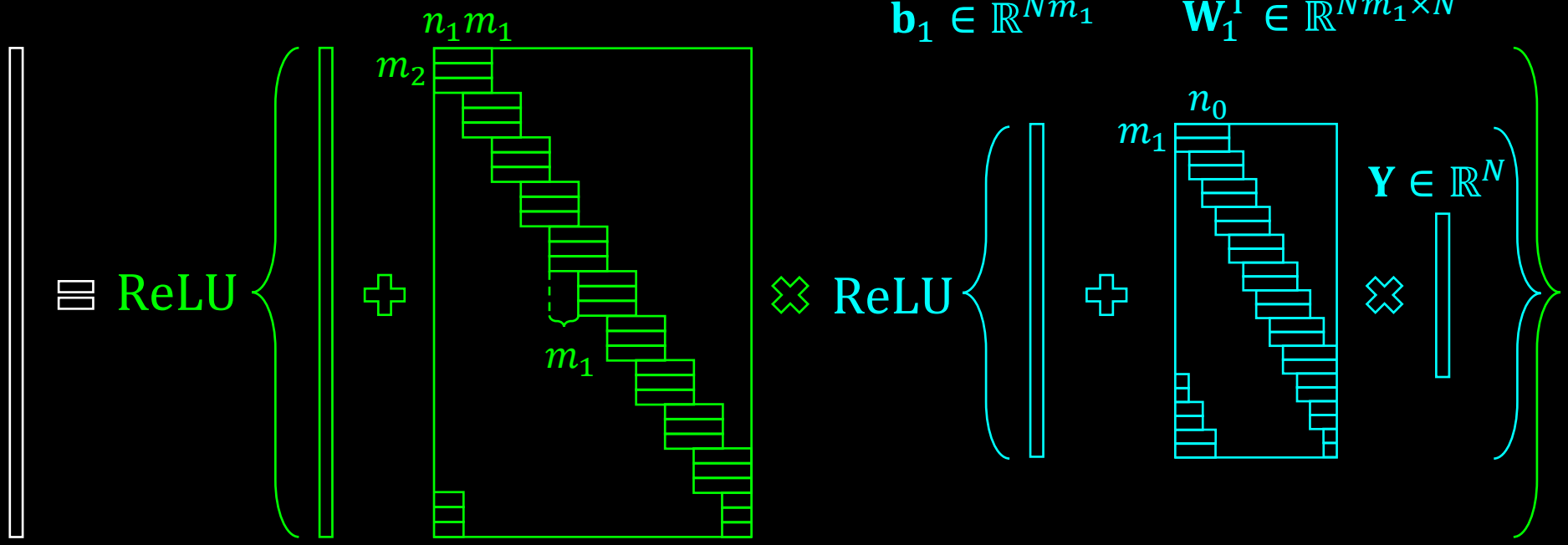
Notice that we do not include a pooling stage:

- Can be replaced by a convolutional layer with increased stride without loss in performance [Springenberg, Dosovitskiy, Brox & Riedmiller '14]
- The current state-of-the-art in image recognition does not use it [He, Zhang, Ren & Sun '15]

Mathematically…

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

$\mathbf{Z}_2 \in \mathbb{R}^{Nm_2}$ $\mathbf{b}_2 \in \mathbb{R}^{Nm_2}$ $\mathbf{W}_2^T \in \mathbb{R}^{Nm_2 \times Nm_1}$ $\mathbf{b}_1 \in \mathbb{R}^{Nm_1}$ $\mathbf{W}_1^T \in \mathbb{R}^{Nm_1 \times N}$ $\mathbf{Y} \in \mathbb{R}^{N}$



Training Stage of CNN

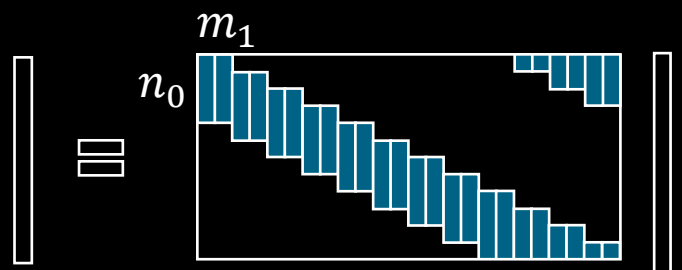
- Consider the task of classification
- Given a set of signals $\{\mathbf{Y}_j\}_j$ and their corresponding labels $\{h(\mathbf{Y}_j)\}_j$, the CNN learns an end-to-end mapping

$$\min_{\{\mathbf{W}_i\},\{\mathbf{b}_i\},\mathbf{U}} \sum_j \ell\left(h(\mathbf{Y}_j), \mathbf{U}, f\left(\mathbf{Y}_j, \{\mathbf{W}_i\}, \{\mathbf{b}_i\}\right)\right)$$

True label Classifier Output of last layer

Back to CSC

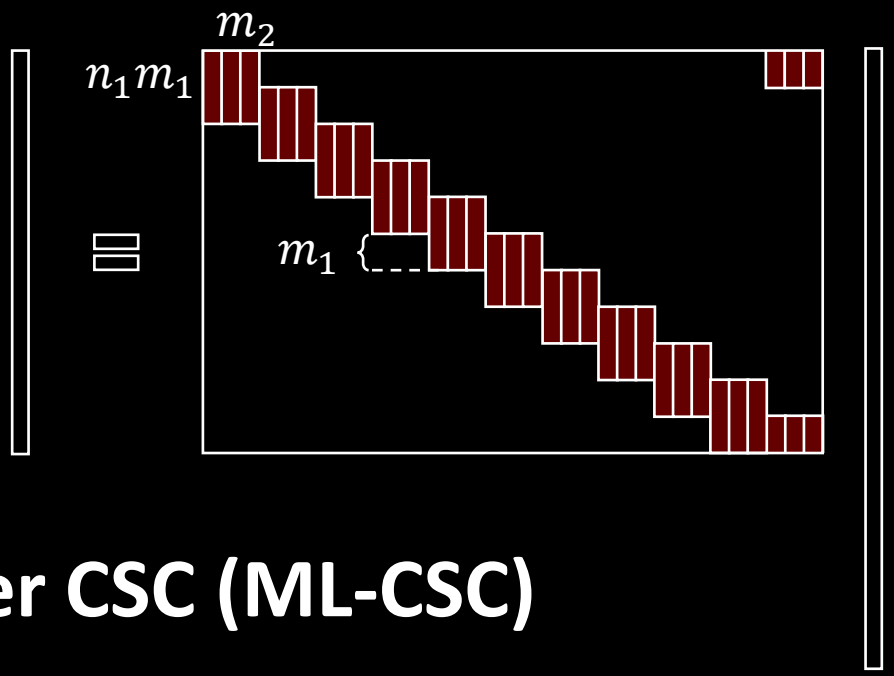
$\mathbf{X} \in \mathbb{R}^{N}$ $\quad \mathbf{D}_1 \in \mathbb{R}^{N \times Nm_1}$ $\quad \boldsymbol{\Gamma}_1 \in \mathbb{R}^{Nm_1}$



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

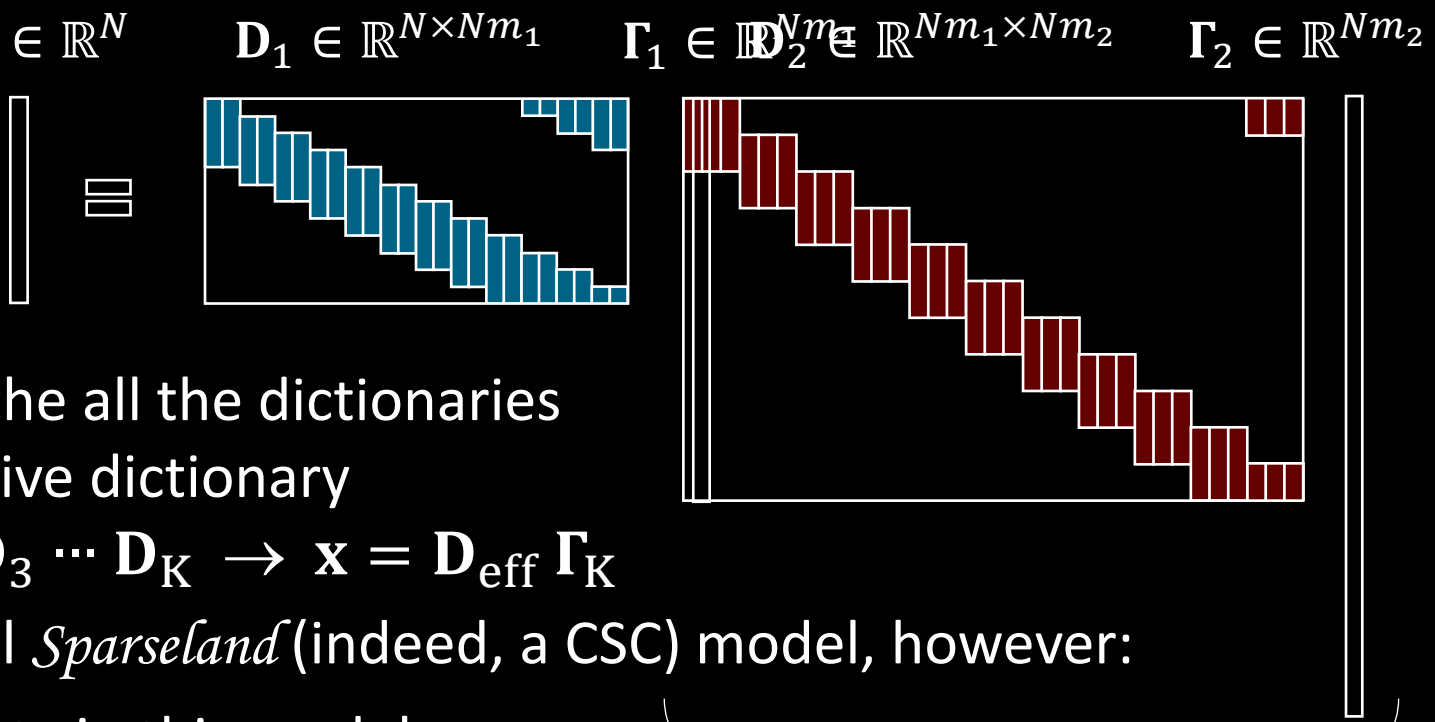
We propose to impose the same structure on the representations themselves

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



Multi-Layer CSC (ML-CSC)

Intuition: From Atoms to Molecules

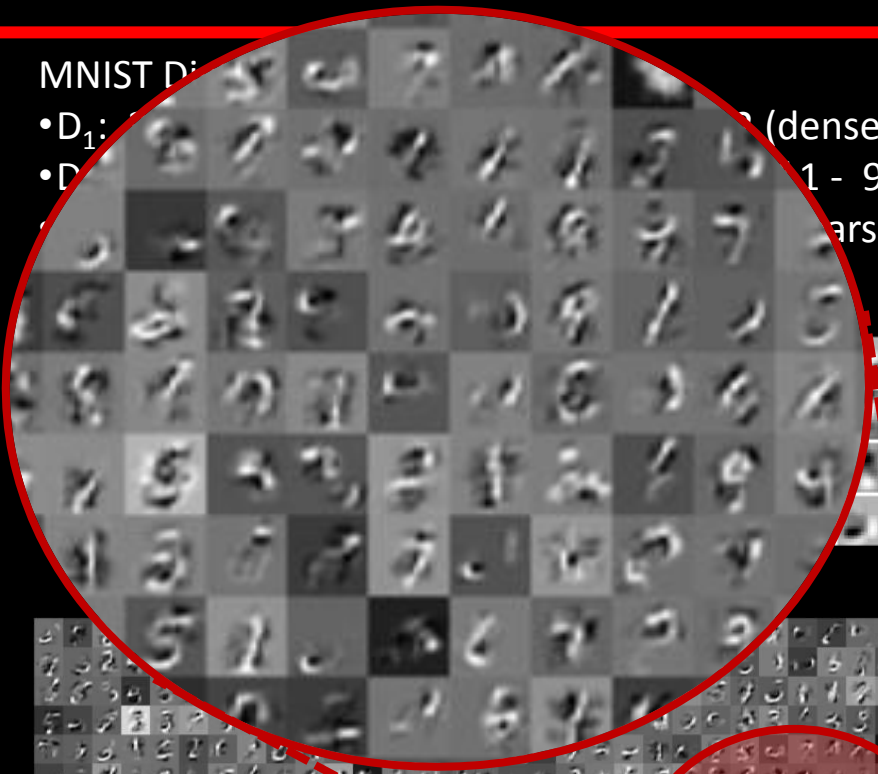


- We can chain the all the dictionaries into one effective dictionary
  $\mathbf{D}_{\text{eff}} = \mathbf{D}_1\mathbf{D}_2\mathbf{D}_3 \cdots \mathbf{D}_{\text{K}} \rightarrow \mathbf{x} = \mathbf{D}_{\text{eff}}\,\mathbf{\Gamma}_{\text{K}}$
- This is a special *Sparseland* (indeed, a CSC) model, however:
  - A key property in this model: sparsity of intermediate representations
  - The effective atoms are combinations of the original atoms → molecules → cell → tissue → body-part …

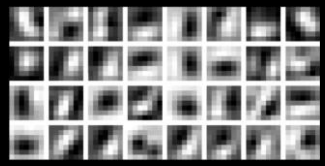
A Small Taste: Model Training (MNIST)

MNIST D:

- D_1: (dense)
- D: 1 - 99.09 % sparse
- arse



$\mathbf{D}_1$ (7×7)



$\mathbf{D}_1\mathbf{D}_2$ (15×15)

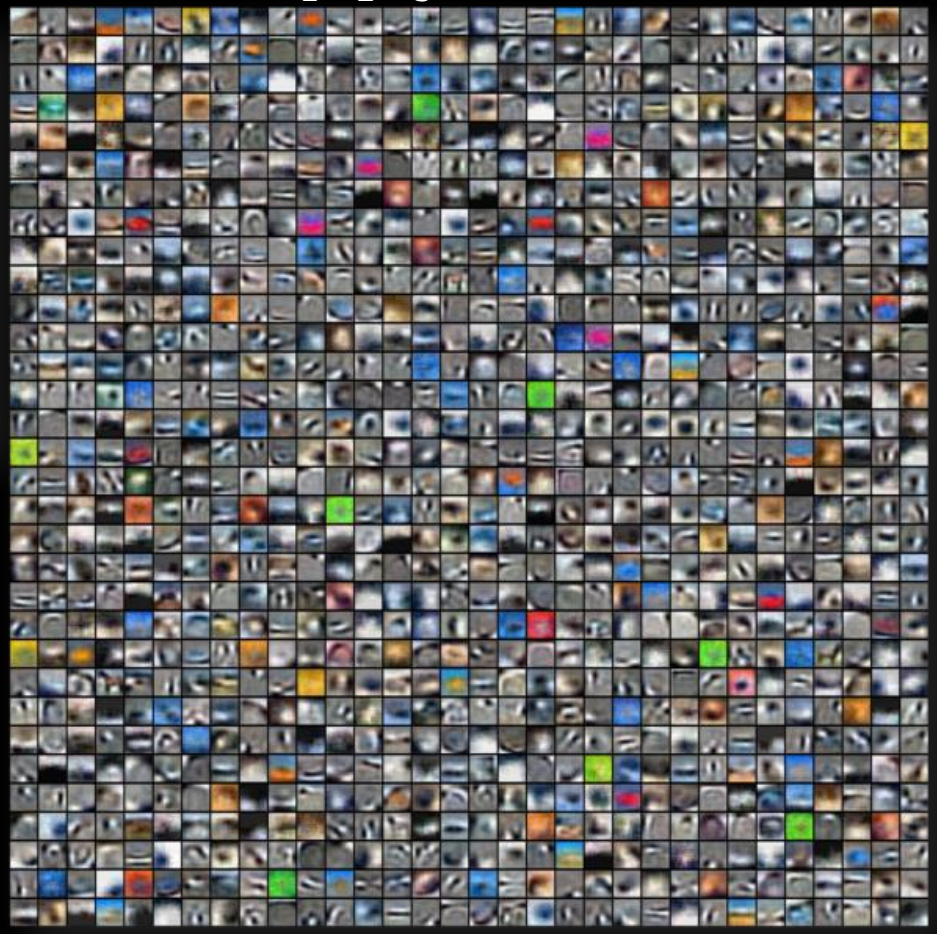
$\mathbf{D}_1\mathbf{D}_2\mathbf{D}_3$ (28×28)

A Small Taste: Model Training (CFAR)

$\mathbf{D}_1$ (5×5×3)

$\mathbf{D}_1\mathbf{D}_2$ (13×13)

$\mathbf{D}_1\mathbf{D}_2\mathbf{D}_3$ (32×32)



CIFAR Dictionary:

- D_1: 64 filters of size 5x5x3, stride of 2
  dense
- D_2: 256 filters of size 5x5x64, stride of 2
  82.99 % sparse
- D_3: 1024 filters of size 5x5x256
  90.66 % sparse

ML-CSC: Pursuit

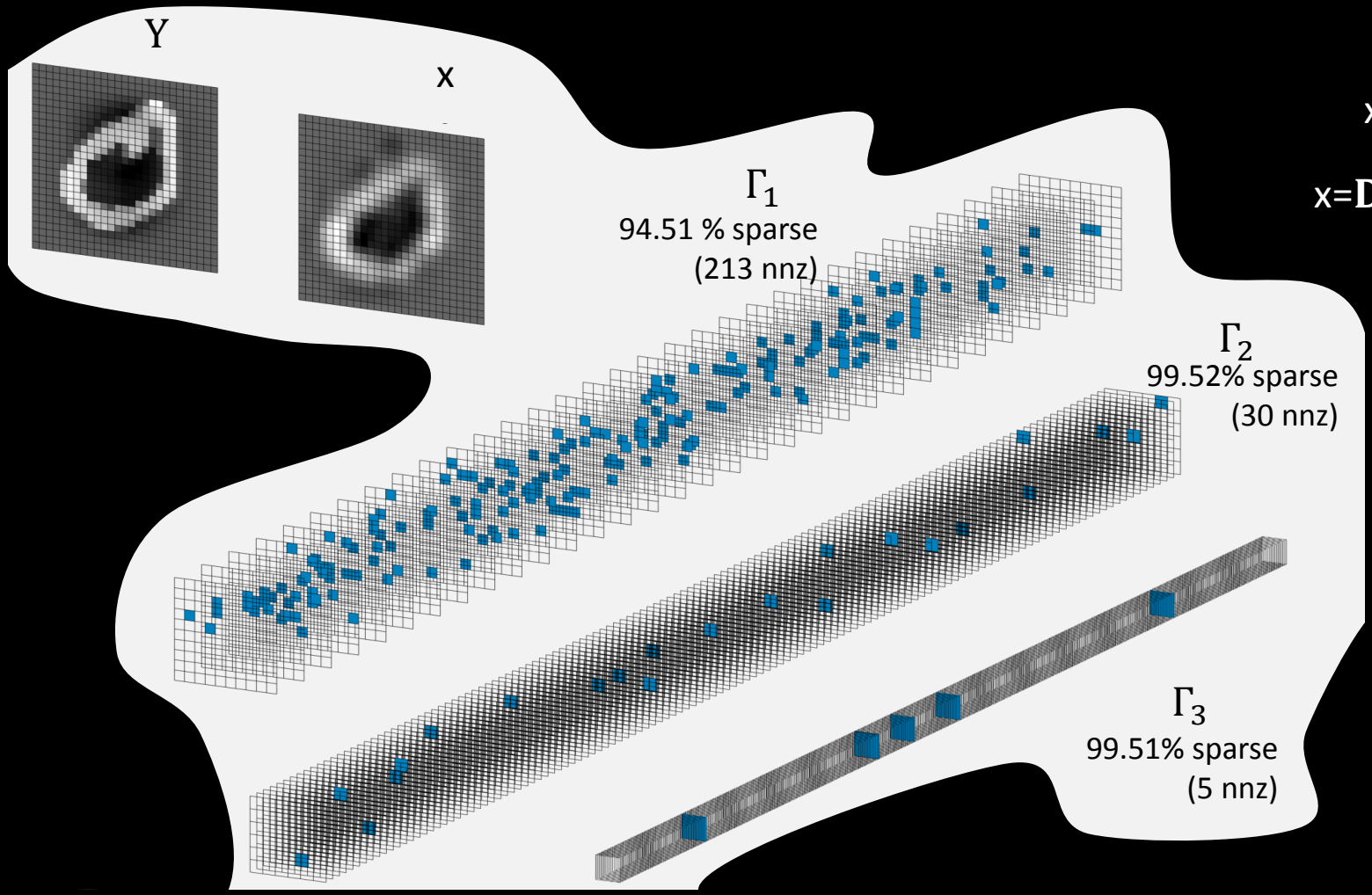
- Deep–Coding Problem ($\mathbf{DCP}_\lambda$) (dictionaries are known):

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

- Or, more realistically for noisy signals,

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

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: Dictionary Learning

- Deep-Learning Problem $(\mathbf{DLP}_\lambda)$:

$$\text{Find} \quad \{\mathbf{D}_i\}_{i=1}^{K} \quad s.t. \quad \left\{ \begin{array}{ll} \left\|\mathbf{Y}_j - \mathbf{D}_1\boldsymbol{\Gamma}_1^j\right\|_2 \le \varepsilon & \left\|\boldsymbol{\Gamma}_1^j\right\|_{0,\infty}^{s} \le \lambda_1 \\ \boldsymbol{\Gamma}_2^j = \mathbf{D}_2\boldsymbol{\Gamma}_1^2 & \left\|\boldsymbol{\Gamma}_2^j\right\|_{0,\infty}^{s} \le \lambda_2 \\ \vdots & \vdots \\ \boldsymbol{\Gamma}_K^j = \mathbf{D}_K\boldsymbol{\Gamma}_K^j & \left\|\boldsymbol{\Gamma}_K^j\right\|_{0,\infty}^{s} \le \lambda_K \end{array} \right\}_{j=1}^{J}$$

- While the above is an unsupervised DL, a supervised version can be envisioned

  [Mairal, Bach & Ponce '12]

$$\min_{\{\mathbf{D}_i\}_{i=1}^{K},\mathbf{U}} \sum_j \ell\left(\mathrm{h}\left(\mathbf{Y}_j\right), \mathbf{U}, \mathbf{DCP}^\star\left(\mathbf{Y}_j, \{\mathbf{D}_i\}\right)\right)$$

The deepest representation $\boldsymbol{\Gamma}_K$ obtained by solving the DCP

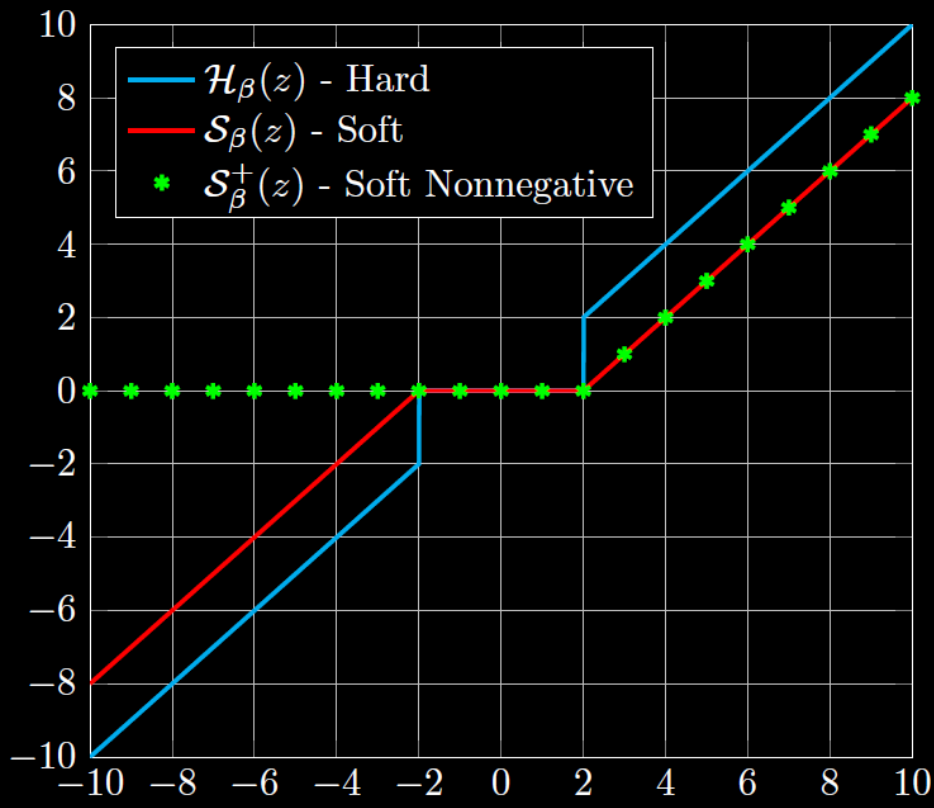
ML-CSC: The Simplest Pursuit

- 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

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

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

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

- Forward pass of CNN:

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

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

Consider this for Solving the DLP

- DLP (supervised*):

$$\min_{\{\mathbf{D}_i\}_{i=1}^{K},\mathbf{U}} \sum_j \ell\left(\mathrm{h}(\mathbf{Y}_j), \mathbf{U}, \underbrace{\mathbf{DCP}^{\star}\left(\mathbf{Y}_j, \{\mathbf{D}_i\}\right)}\right)$$

The thresholds for the DCP should also learned

Estimate via the layered THR algorithm

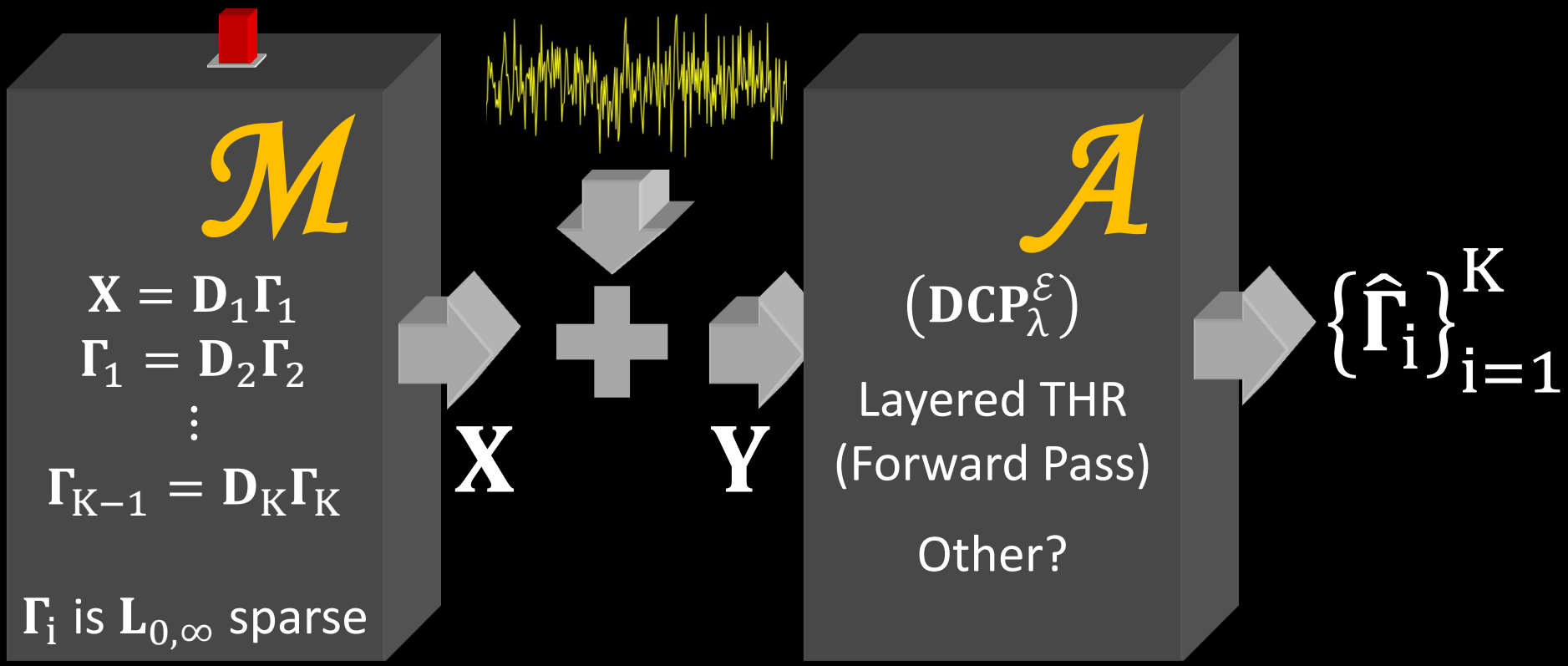
- CNN training:

$$\min_{\{\mathbf{W}_i\},\{\mathbf{b}_i\},\mathrm{U}} \sum_j \ell\left(\mathrm{h}(\mathbf{Y}_j), \mathbf{U}, f(\mathbf{Y}, \{\mathbf{W}_i\}, \{\mathbf{b}_i\})\right)$$

The problem solved by the training stage of CNN and the DLP are equivalent as well, assuming that the DCP is approximated via the layered thresholding algorithm

* Recall that for the ML-CSC, there exists an unsupervised avenue for training the dictionaries that has no simple parallel in CNN

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. Uniqueness of the solution (set of representations) to the $(\mathbf{DCP}_\lambda)$?
II. Stability of the solution to the $\left(\mathbf{DCP}_\lambda^\mathcal{E}\right)$ problem ?

III. Stability of the solution obtained via the hard and soft layered THR algorithms (forward pass) ?
IV. Limitations of this (very simple) algorithm and alternative pursuit?

V. Algorithms for training the dictionaries $\{\mathbf{D}_i\}_{i=1}^K$ vs. CNN ?
VI. New insights on how to operate on signals via CNN ?

Uniqueness of $(\mathbf{DCP}_\lambda)$

$(\mathbf{DCP}_\lambda)$: Find a set of representations satisfying

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

Is this set unique?

Theorem: If a set of solutions $\{\boldsymbol{\Gamma}_i\}_{i=1}^{K}$ is found for $(\mathbf{DCP}_\lambda)$ such that:

$$\|\boldsymbol{\Gamma}_i\|_{0,\infty}^{s} \le \lambda_i < \frac{1}{2}\left(1+\frac{1}{\mu(\mathbf{D}_i)}\right)$$

then these are necessarily the unique solution to the DCP problem

The feature maps CNN aims to recover are unique

Stability of $\left(\mathbf{DCP}_{\lambda}^{\mathcal{E}}\right)$

Is this set stable?

- The problem we aim to solve is this

$\left(\mathbf{DCP}_{\lambda}^{\mathcal{E}}\right)$: Find a set of representations satisfying

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

- Suppose that we manage to solve the $\left(\mathbf{DCP}_{\lambda}^{\mathcal{E}}\right)$ and find a feasible set of representations satisfying all the conditions

$$\left\{\hat{\boldsymbol{\Gamma}}_i\right\}_{i=1}^{K}$$

- The question we pose is How close is $\hat{\boldsymbol{\Gamma}}_i$ to $\boldsymbol{\Gamma}_i$?

Stability of $\left(\mathbf{DCP}_{\lambda}^{\mathcal{E}}\right)$

Theorem: If the true representations $\{\boldsymbol{\Gamma}_i\}_{i=1}^K$ satisfy

$$\|\boldsymbol{\Gamma}_i\|_{0,\infty}^s \leq \lambda_i < \frac{1}{2}\left(1+\frac{1}{\mu(\mathbf{D}_i)}\right)$$

then the set of solutions $\{\hat{\boldsymbol{\Gamma}}_i\}_{i=1}^K$ obtained by solving this problem (somehow) must obey

$$\left\|\hat{\boldsymbol{\Gamma}}_i - \boldsymbol{\Gamma}_i\right\|_2^2 \leq \mathcal{E}_i^2 \quad \text{for}$$

$$\mathcal{E}_0^2 = 4\mathcal{E}^2, \qquad \mathcal{E}_i^2 = \frac{\mathcal{E}_{i-1}^2}{1-(2\lambda_i-1)\mu(\mathbf{D}_i)}$$

The problem CNN aims to solve is stable under certain conditions

Observe this annoying effect of error magnification as we dive into the model

Stability of 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) will find 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

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

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

Better Pursuit ?

- $(\mathbf{DCP}_\lambda)$ Noiseless: Find a set of representations satisfying

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

- So far we proposed the Layered THR:

$$\widehat{\boldsymbol{\Gamma}}_K = \mathcal{P}_{\beta_K}\left(\mathbf{D}_K^T \ldots \mathcal{P}_{\beta_2}\left(\mathbf{D}_2^T\,\mathcal{P}_{\beta_1}\left(\mathbf{D}_1^T\mathbf{X}\right)\right)\right)$$

- The motivation is clear – getting close to what CNN use
- However, this is the simplest and weakest pursuit known in the field of sparsity – Can we offer something better?

Layered Basis Pursuit (Noiseless)

- Our Goal: $(\mathbf{DCP}_\lambda)$: Find a set of representations satisfying

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

- We can propose a Layered Basis Pursuit Algorithm:

$$\boldsymbol{\Gamma}_1^{\mathrm{LBP}} = \min_{\boldsymbol{\Gamma}_1} \|\boldsymbol{\Gamma}_1\|_1 \quad \text{s.t.} \quad \mathbf{X}=\mathbf{D}_1\boldsymbol{\Gamma}_1$$

$$\boldsymbol{\Gamma}_2^{\mathrm{LBP}} = \min_{\boldsymbol{\Gamma}_2} \|\boldsymbol{\Gamma}_2\|_1 \quad \text{s.t.} \quad \boldsymbol{\Gamma}_1^{\mathrm{LBP}} = \mathbf{D}_2\boldsymbol{\Gamma}_2$$

$\bullet\bullet\bullet$

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

Guarantee for Success of Layered BP

- As opposed to prior work in CNN, we can do far more than just proposing an algorithm – we can analyze its terms for success:

Theorem: If a set of representations $\{\mathbf{\Gamma}_i\}_{i=1}^K$ of the Multi-Layered CSC model satisfy

$$\|\mathbf{\Gamma}_i\|_{0,\infty}^s \leq \lambda_i < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D}_i)}\right)$$

then the Layered BP is guaranteed to find them

- Consequences:
  - The layered BP can retrieve the underlying representations in the noiseless case, a task in which the forward pass fails to provide
  - The Layered-BP's success does not depend on the ratio $\left|\mathbf{\Gamma}_i^{min}\right|/|\mathbf{\Gamma}_i^{max}|$

Layered Basis Pursuit (Noisy)

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

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

We can invoke a result we have seen already, referring to the BP for the CSC model:

RECALL

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

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

then we are guaranteed that

$$\|\boldsymbol{\Delta}\|_{2,\infty}^{p} \leq 7.5\,\varepsilon_L^0\sqrt{\|\boldsymbol{\Gamma}\|_{0,\infty}^{p}}$$

Stability of Layered BP

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

then for correctly chosen $\{\lambda_i\}_{i=1}^{K}$ we are guaranteed that

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

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

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

Problems:

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

Layered Iterative Thresholding

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

Layered Iterative Soft-Thresholding:

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

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

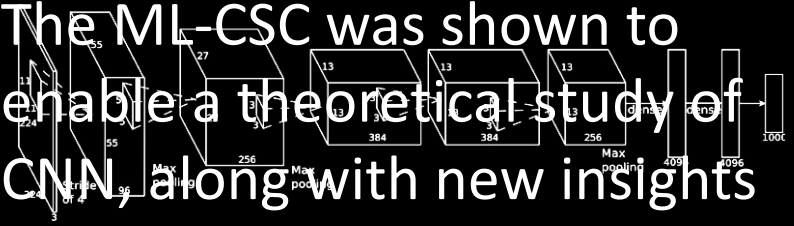
Can be seen as a recurrent neural network
[Gregor & LeCun '10]

* $c_i > 0.5\,\lambda_{max}(\mathbf{D}_i^T\mathbf{D}_i)$

Time to Conclude

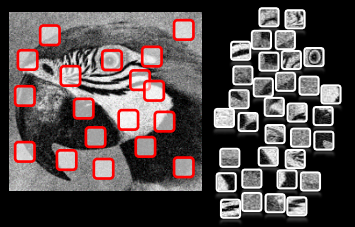
This Talk

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



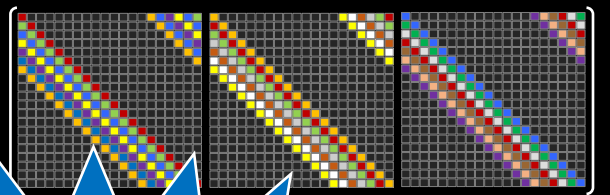
Independent patch-processing

Local Sparsity



Convolutional Neural Networks

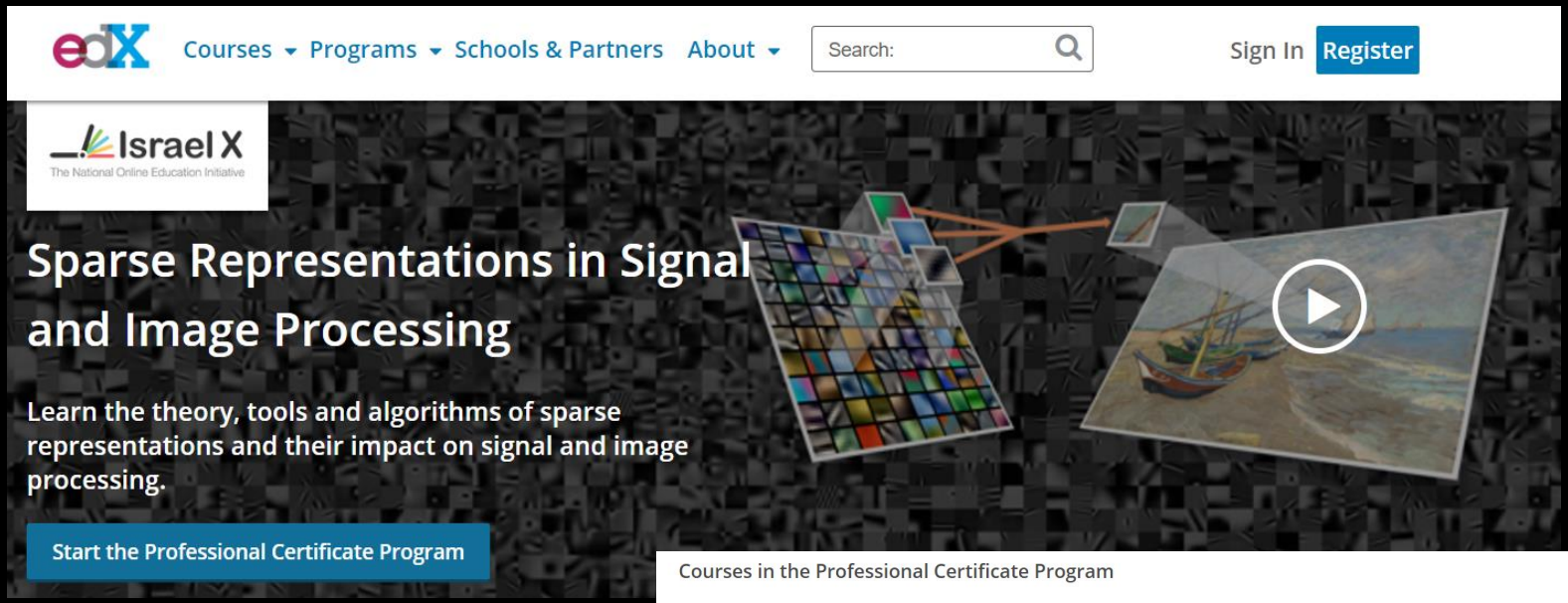
Novel View of Convolutional Sparse Coding

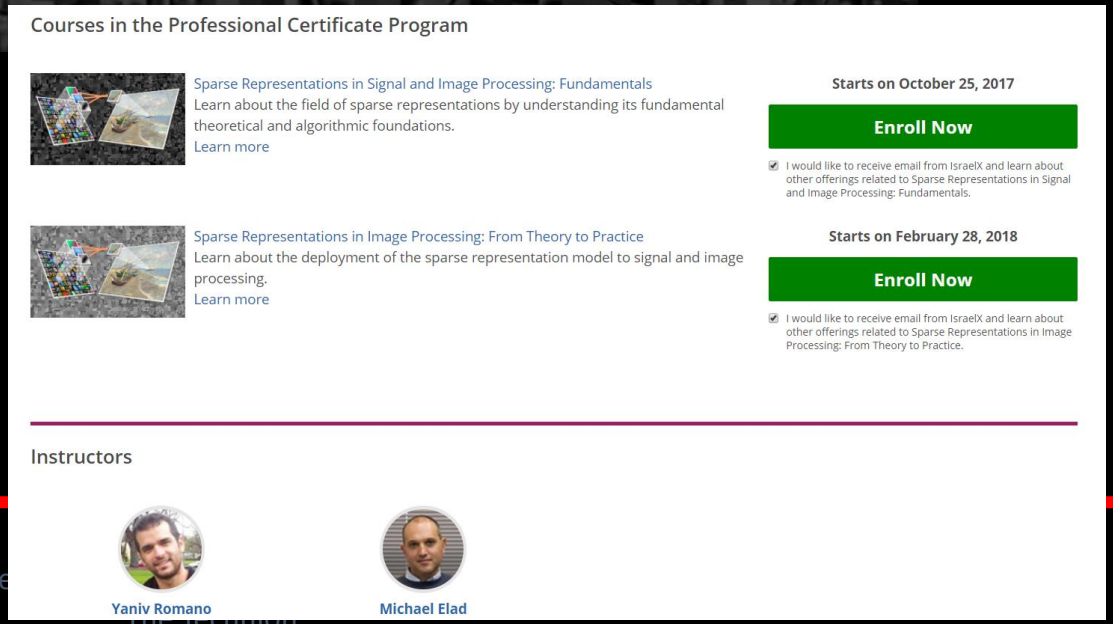


Multi-Layer Convolutional Sparse Coding

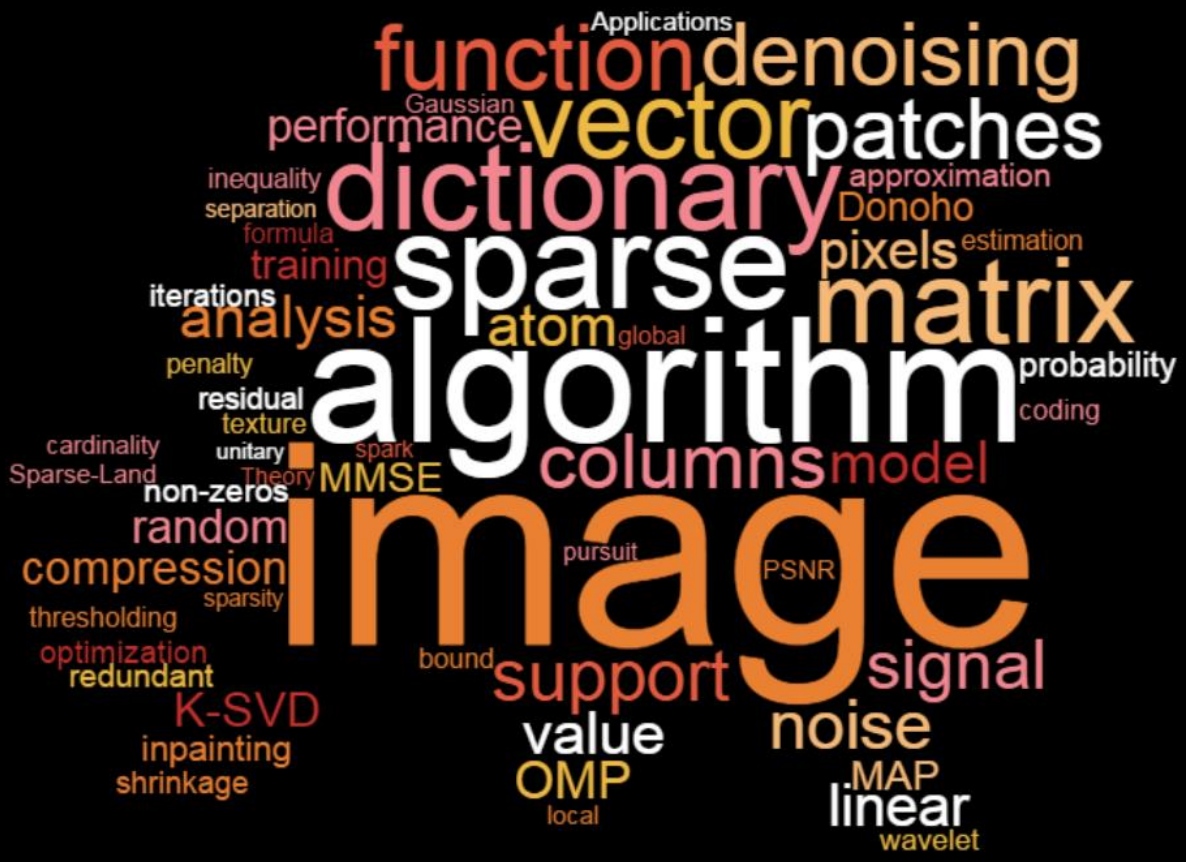
The underlying idea:
Modeling the data source in order to be able to theoretically analyze algorithms' performance

A Massive Open Online Course : Coming Up





Michael Elad
The Computer-Science Department
The Technion



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

Michael Elad
The Computer-Science Department
The Technion