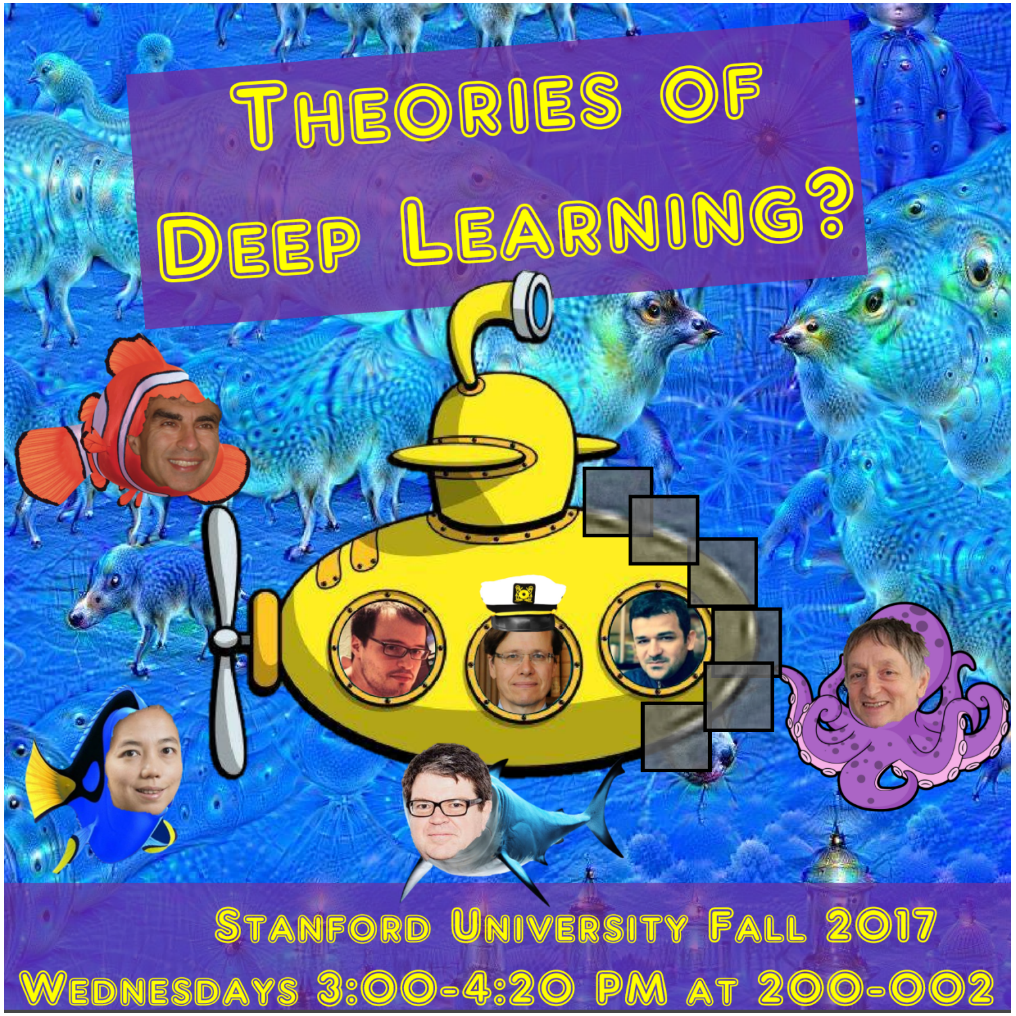
Sparsity in Convolutional Neural Networks

Speaker: Qingyun Sun
Math PhD @ Stanford

Acknowledgement:
Stats 385 @ Stanford
https://stats385.github.io/



The talk is based on:

Convolutional Neural Networks in View of Sparse Coding, Vardan Papyan @ Stats 385, Stanford

Based on work of:
Vardan Papyan, Jeremias Sulam, Yaniv Romano, Michael Elad

Sparsity: Central idea in Stats

Compressive Sensing:

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

$$\hat{\boldsymbol{\Gamma}} = \arg\min_{\boldsymbol{\Gamma}} \ \|\boldsymbol{\Gamma}\|_1 \ \text{ s.t. } \ \mathbf{X} = \mathbf{D}\boldsymbol{\Gamma}$$

Sparsity: Central idea in Stats

Lasso:

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

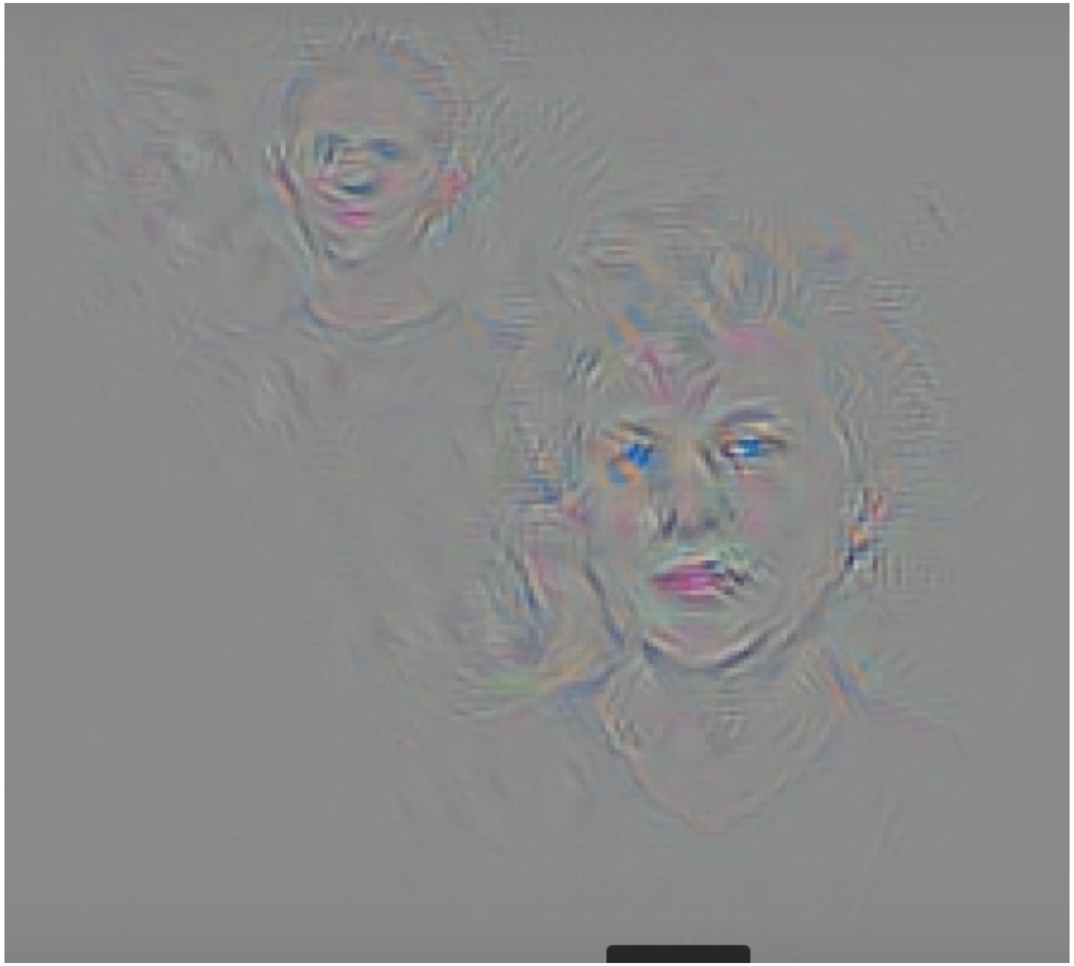
$$\hat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} \ \frac{1}{2}\|\mathbf{Y} - \mathbf{D}\mathbf{\Gamma}\|_2^2 + \lambda\|\mathbf{\Gamma}\|_1$$

After Deep Revolution, is sparsity still important?

Sparsity observed in CNN

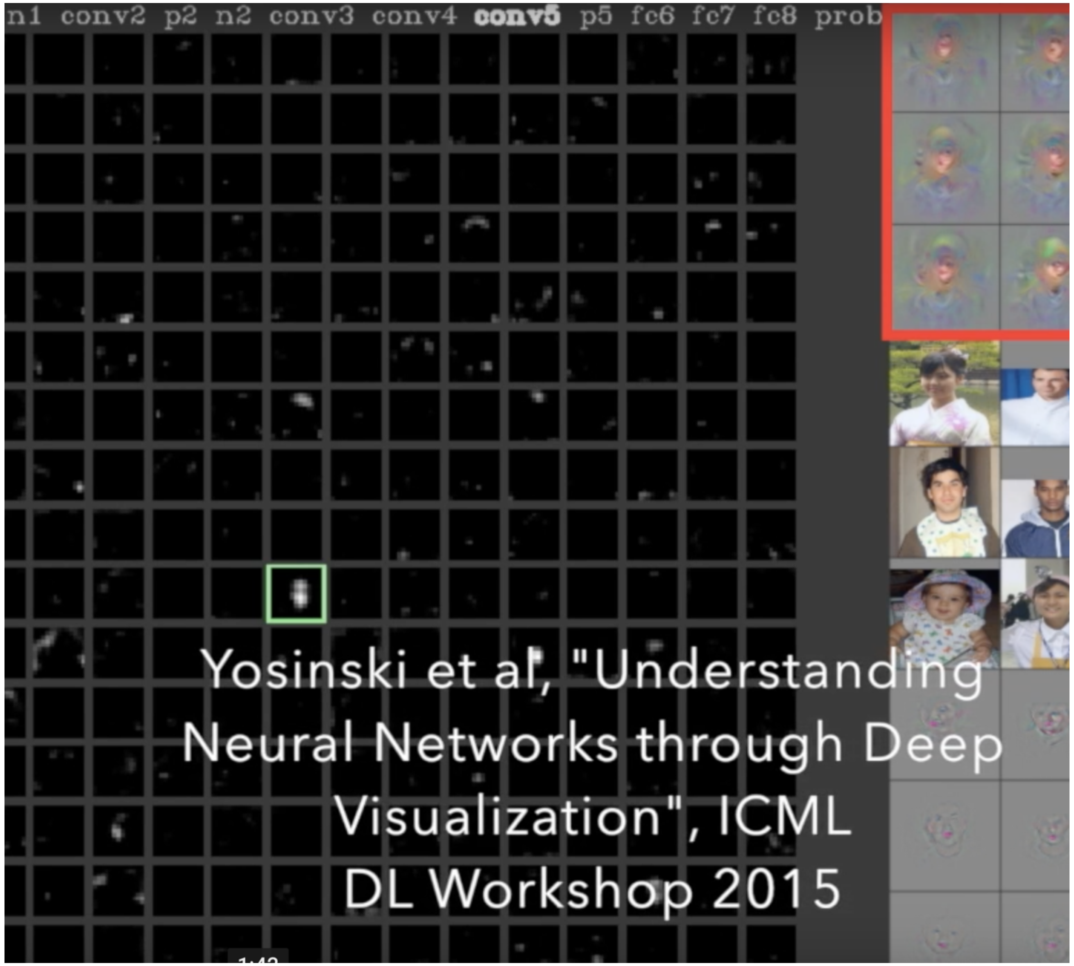
Sparsity in Practice

The activation of RELU layer is sparse.

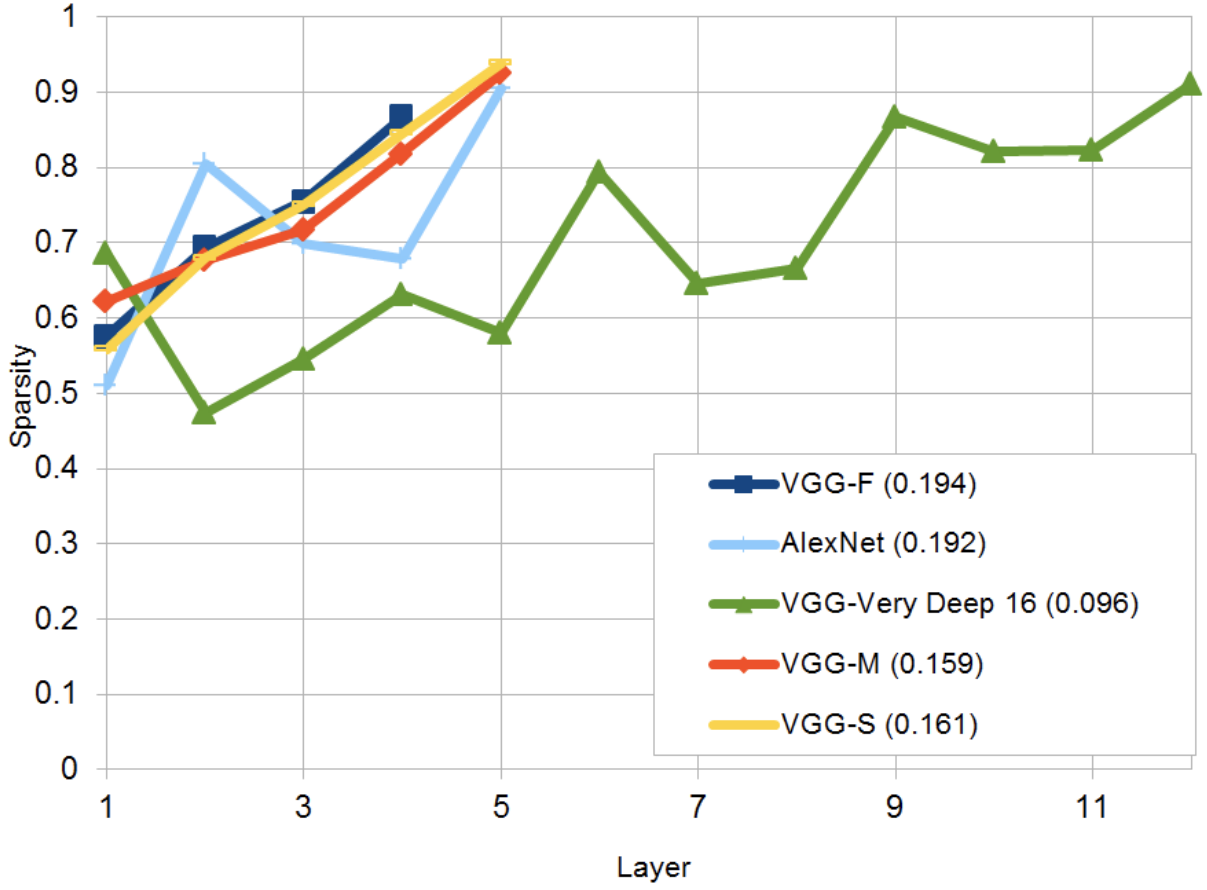


Sparsity in Practice

The activation of RELU layer is sparse.

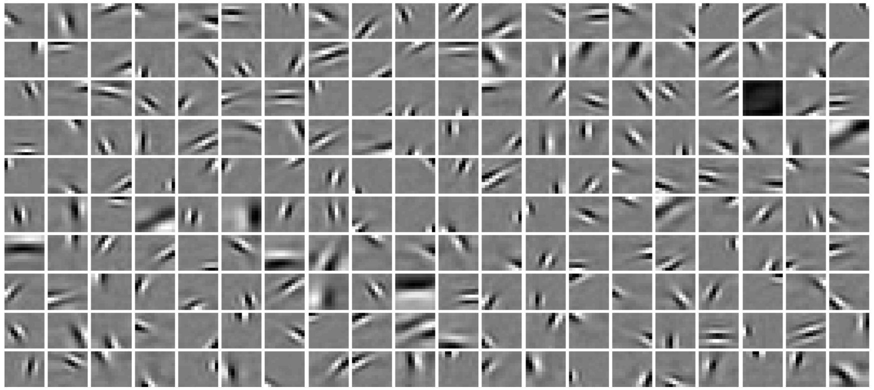


Sparsity in Practice



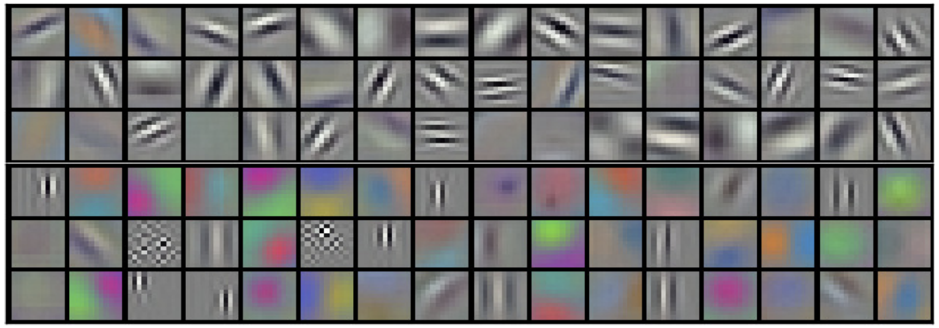
Olshausen & Field and AlexNet

Olshausen & Field



explicit sparsity

AlexNet



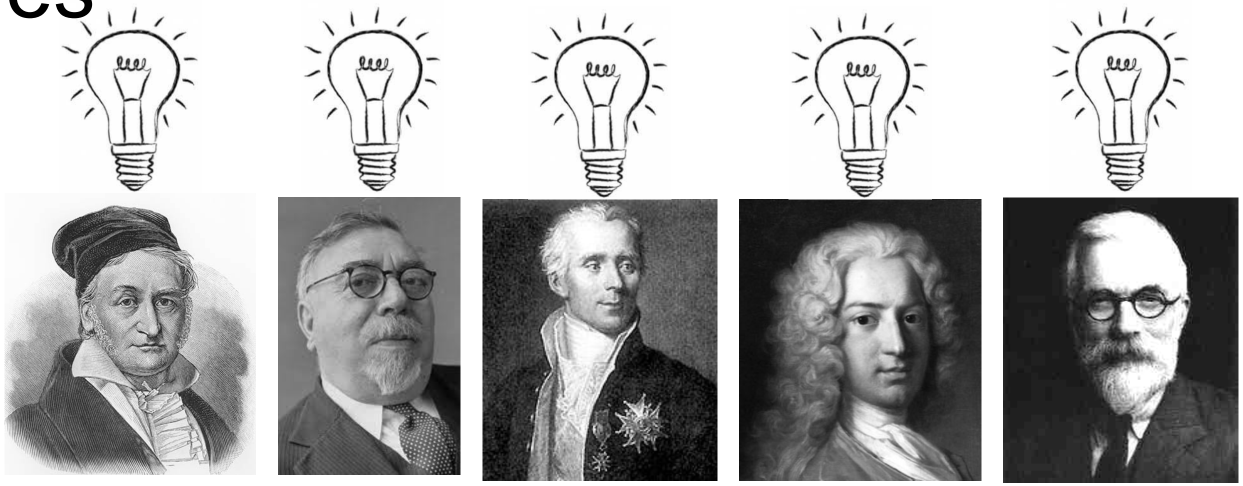
implicit sparsity

Credit to: Vardan, Stats385@Stanford

Theory of sparsity in CNN?

Breiman's "Two Cultures"

Generative modeling



Gauss Wiener Laplace Bernoulli Fisher

Predictive modeling

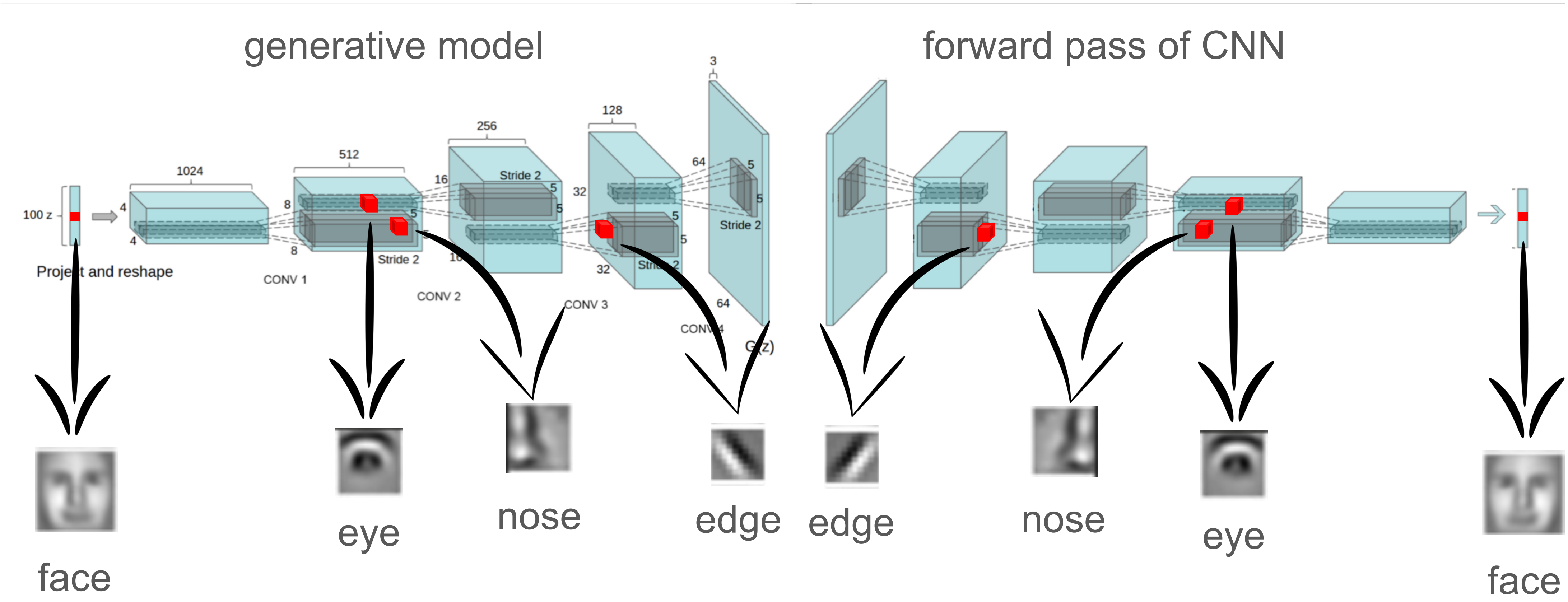
Generative modeling

Seeks to develop stochastic models which fit the data, and then make inferences about the data-generating mechanism based on the structure of those models.

Predictive modeling

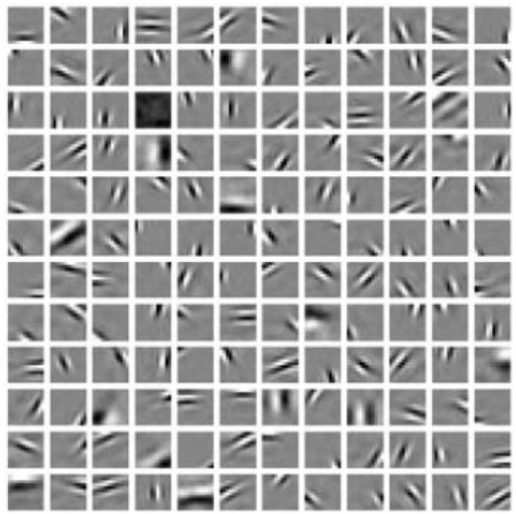
Predictive modeling is effectively silent about the underlying mechanism generating the data, and allows for many different predictive algorithms, preferring to discuss only accuracy of prediction made by different algorithm on various datasets.

Generative Modeling

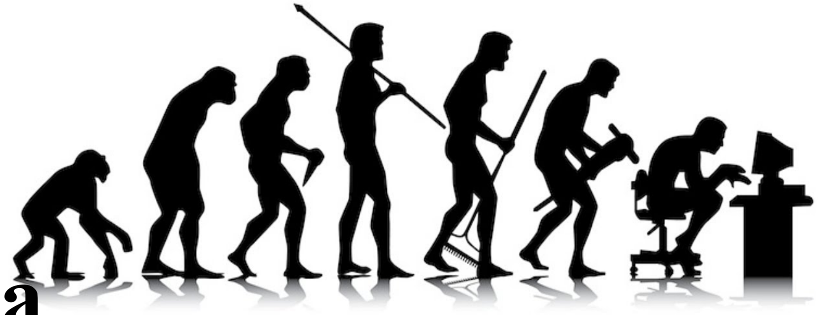


Sparse Representation Generative Model

- Receptive fields in visual cortex are spatially localized, oriented and bandpass
- Coding natural images while promoting sparse solutions results in a set of filters satisfying these properties [Olshausen and Field 1996]
- Two decades later…
  - vast theoretical study
  - different inference algorithms
  - different ways to train the model



Evolution of Models



Multi-Layered
Convolutional
Neural
Network

First Layer of a
Convolutional
Neural Network

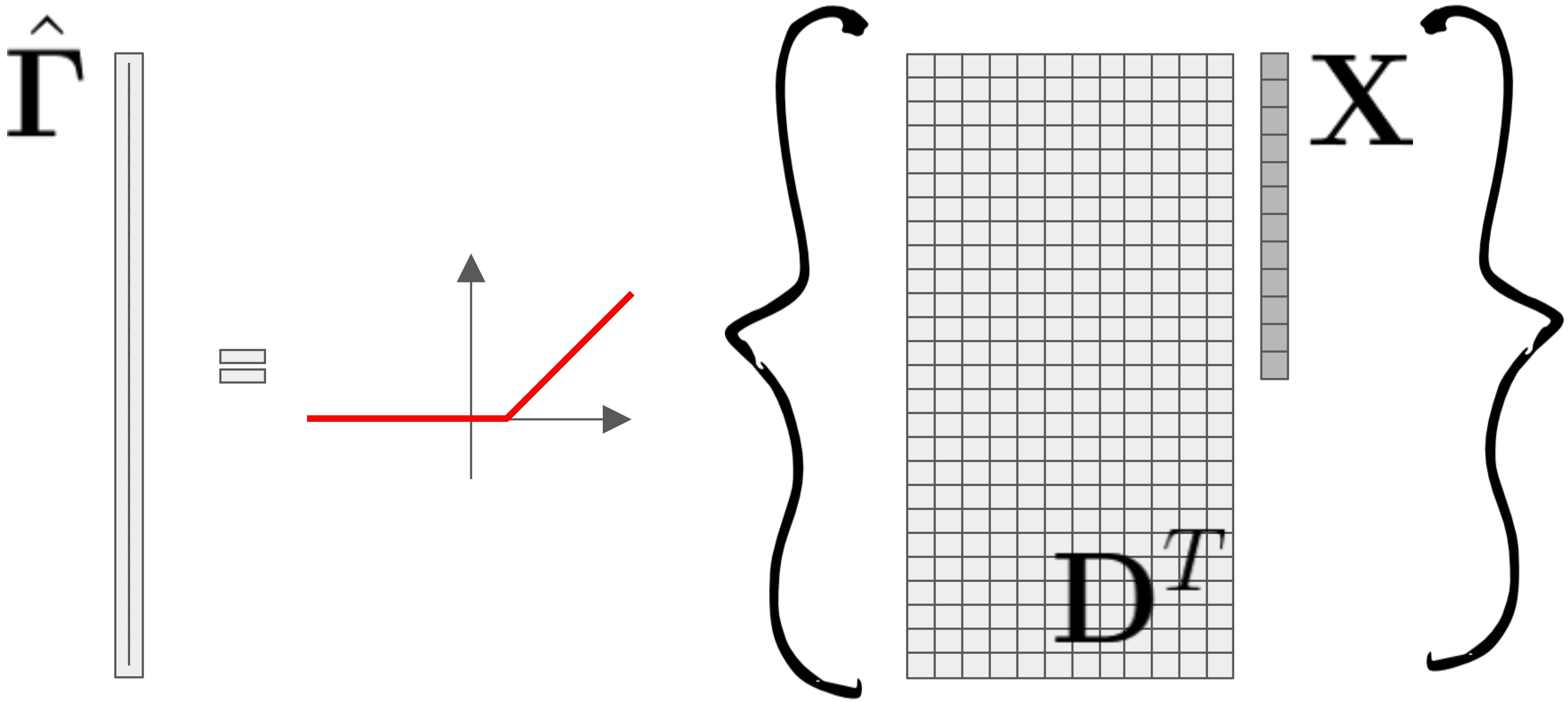
First Layer of a
Neural Network

Multi-Layered
Convolutional
Sparse
Representatio
n

Convolutional
sparse
representation

Sparse
representations

First Layer of a Neural Network



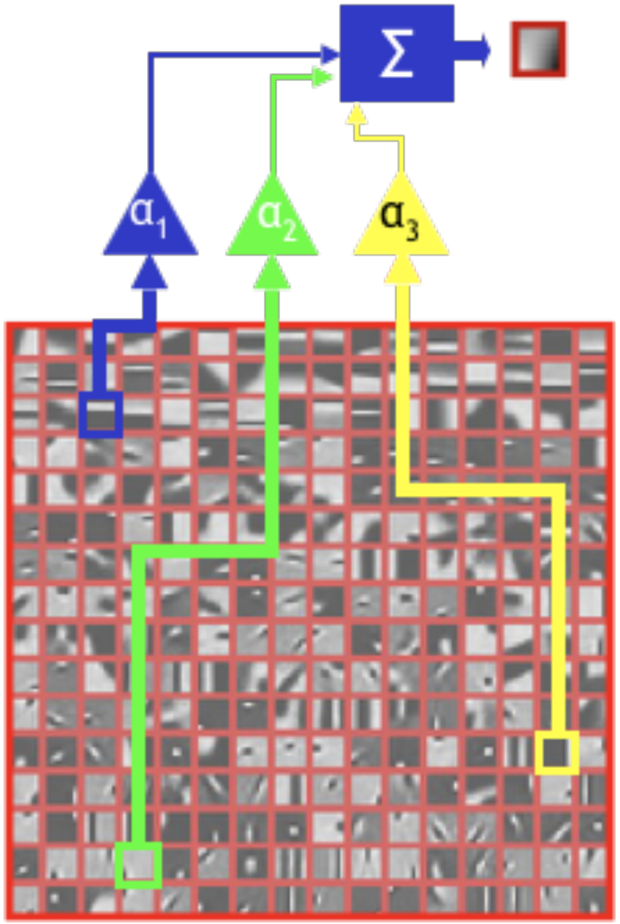
Sparse Modeling

Task: model image patches of size 8x8 pixels 

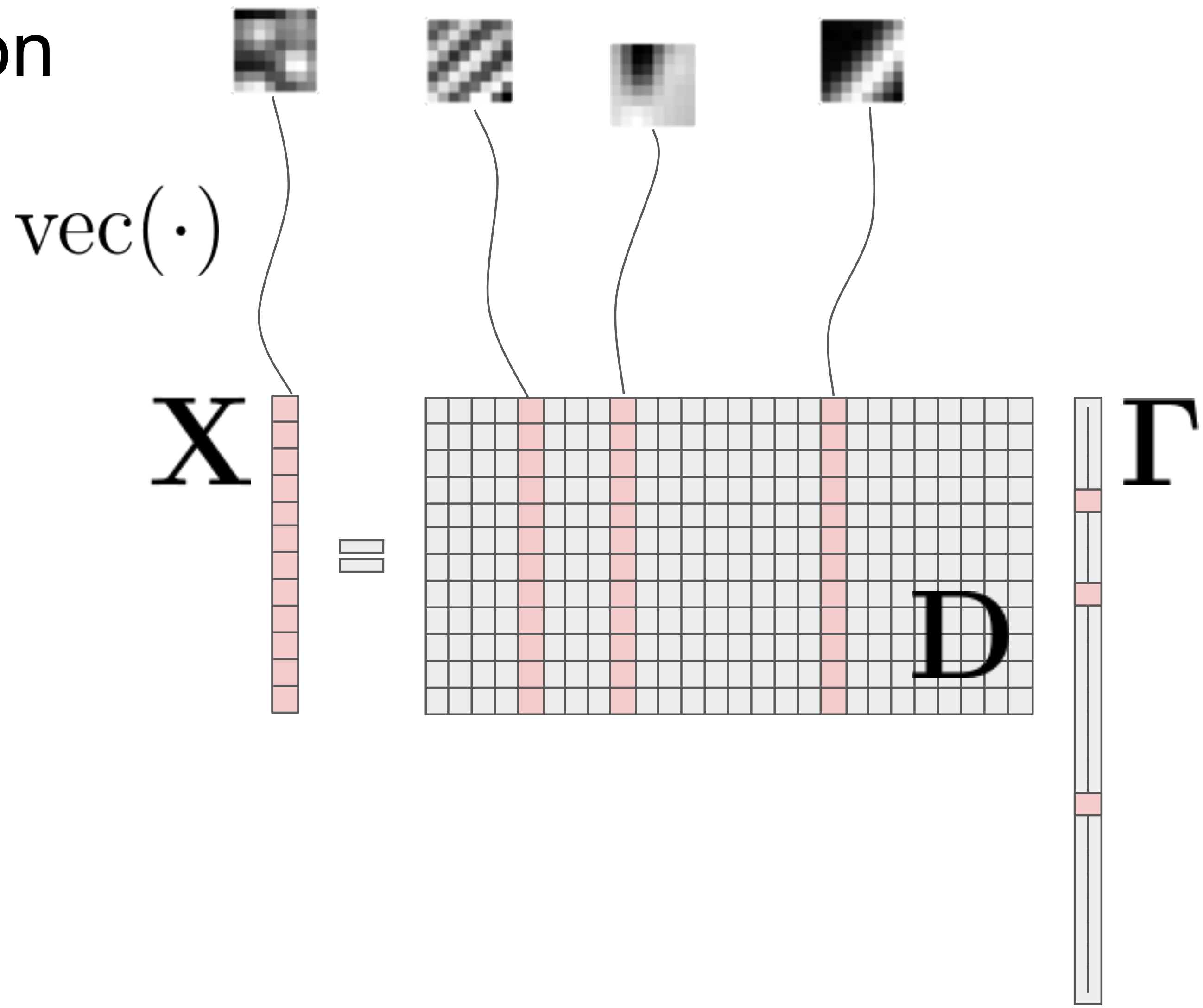
We assume a dictionary of such image patches is given, containing 256 atoms

Assumption: every patch can be described as a linear combination of a few atoms

Key properties: sparsity and redundancy



Matrix Notation



Sparse Coding

Given a signal, we would like to find its sparse representation

$$\min_{\mathbf{\Gamma}} \ \|\mathbf{\Gamma}\|_0 \ \text{ s.t. } \ \mathbf{X} = \mathbf{D}\mathbf{\Gamma}$$

Convexify

$$\min_{\mathbf{\Gamma}} \ \|\mathbf{\Gamma}\|_1 \ \text{ s.t. } \ \mathbf{X} = \mathbf{D}\mathbf{\Gamma}$$

Sparse Coding

Given a signal, we would like to find its sparse representation

$$\min_{\mathbf{\Gamma}} \ \|\mathbf{\Gamma}\|_0 \ \text{ s.t. } \ \mathbf{X} = \mathbf{D}\mathbf{\Gamma}$$

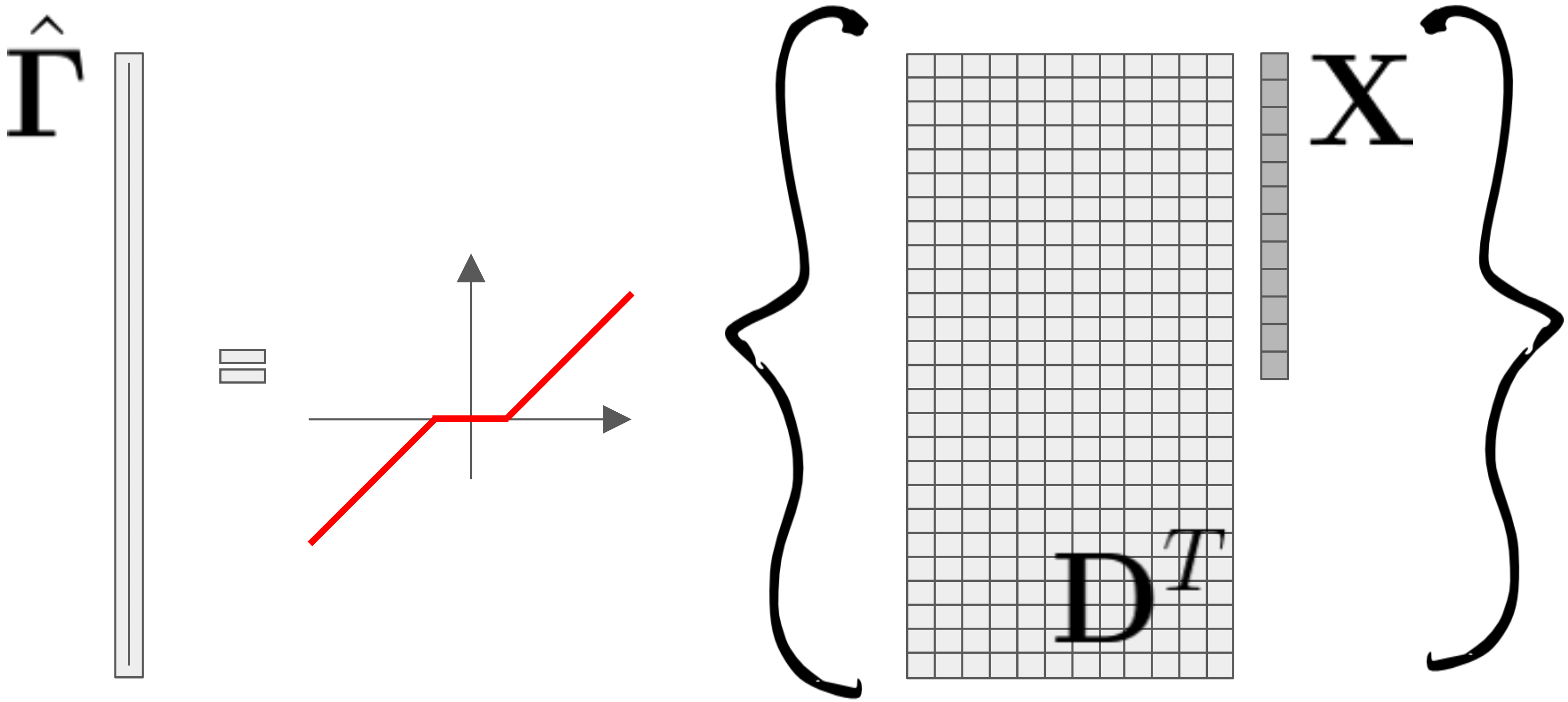
Convexify

$$\min_{\mathbf{\Gamma}} \ \|\mathbf{\Gamma}\|_1 \ \text{ s.t. } \ \mathbf{X} = \mathbf{D}\mathbf{\Gamma}$$

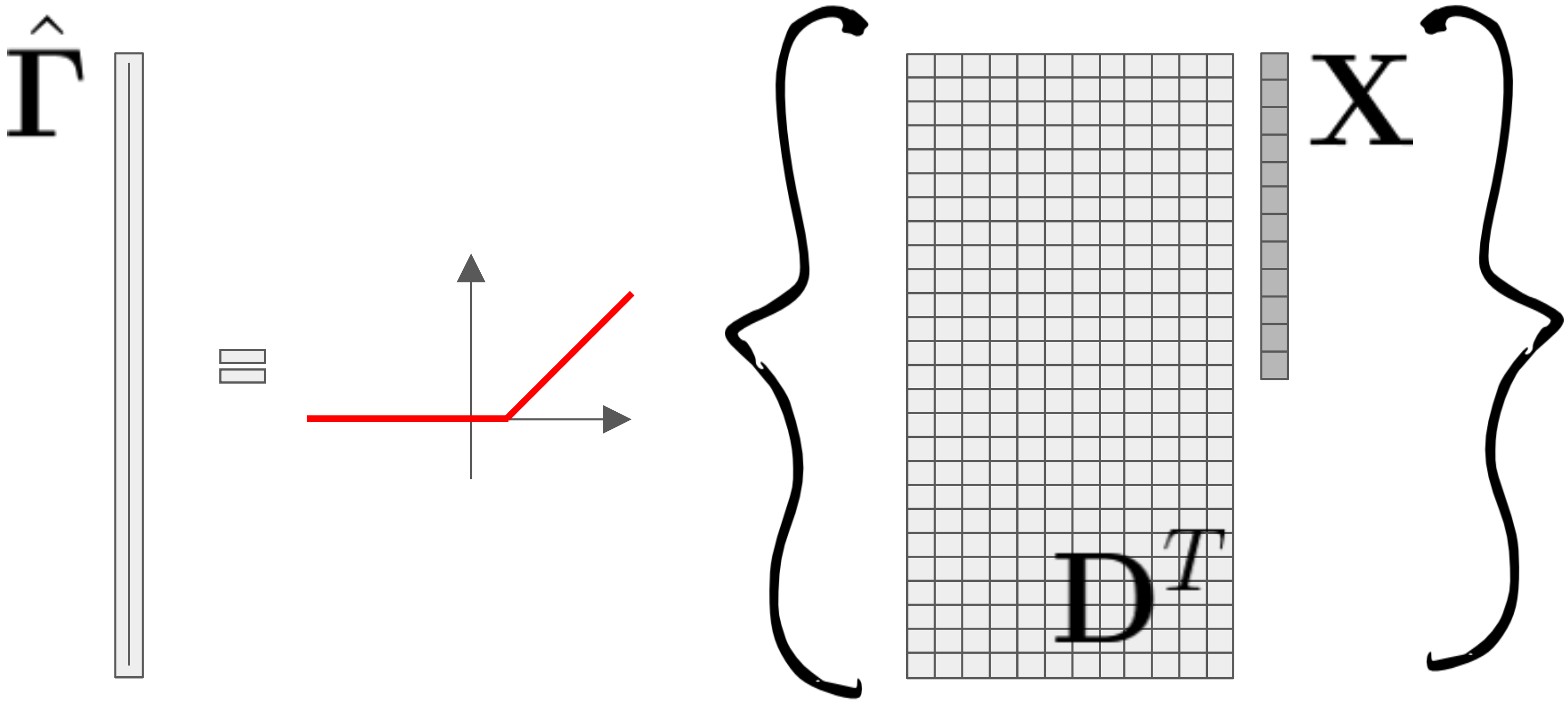
Crude approximation

$$\mathcal{S}_\beta\{\mathbf{D}^T\mathbf{X}\}$$

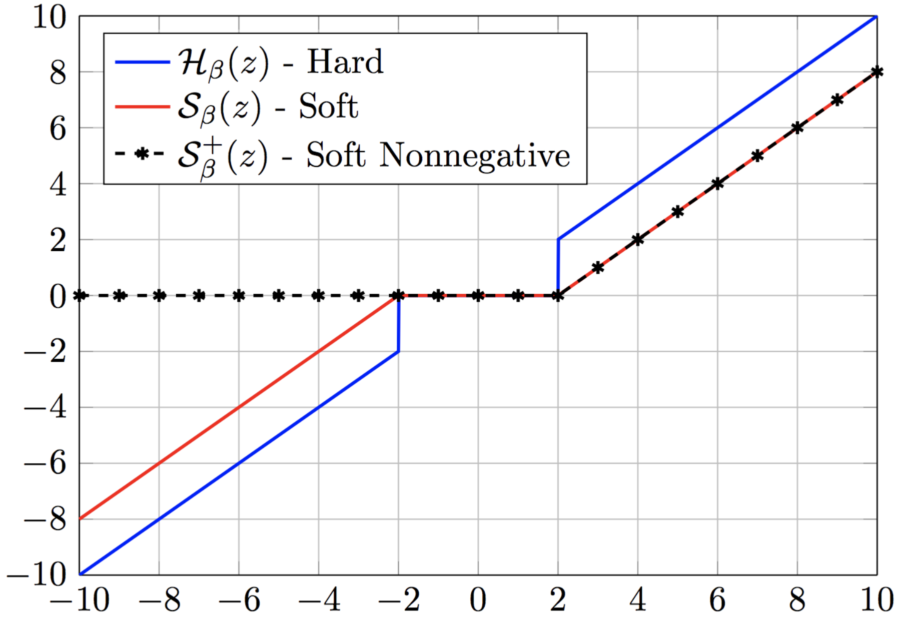
Thresholding Algorithm



First Layer of a Neural Network

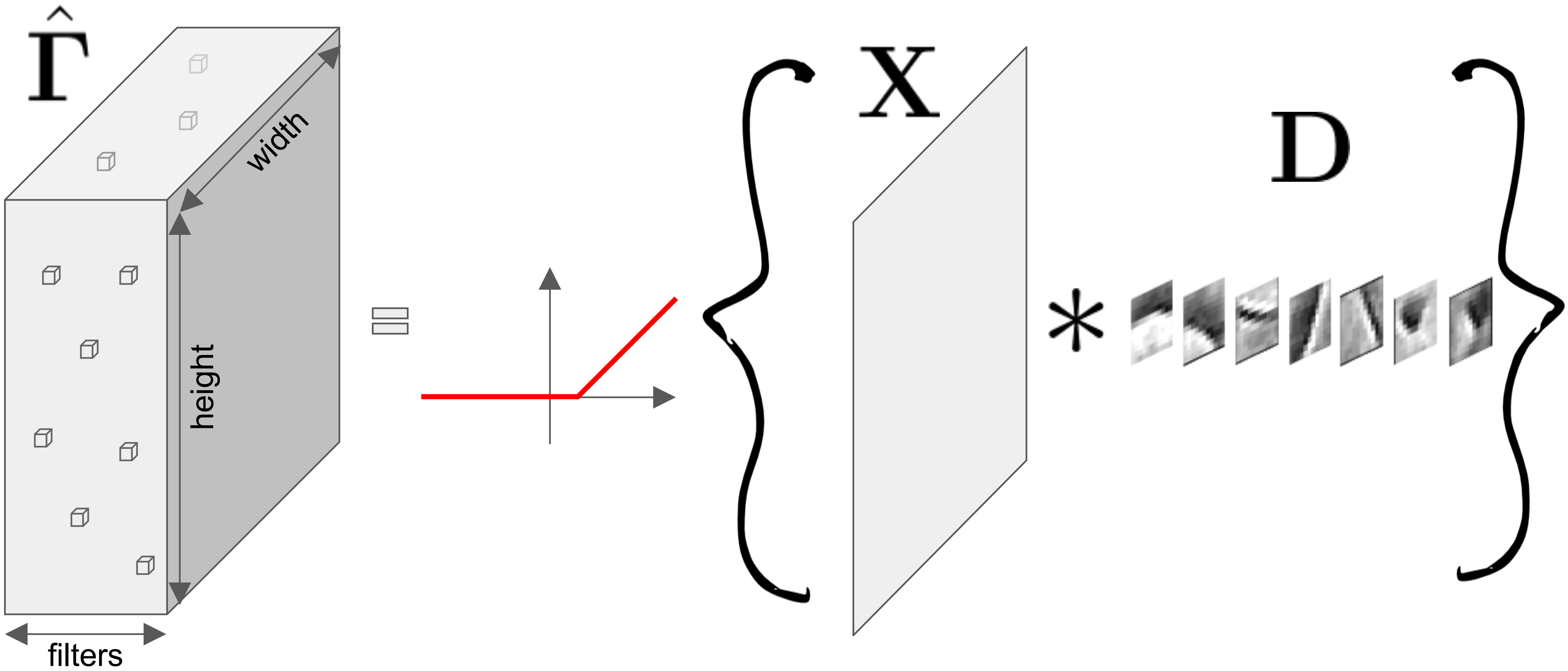


ReLU = Soft Nonnegative Thresholding

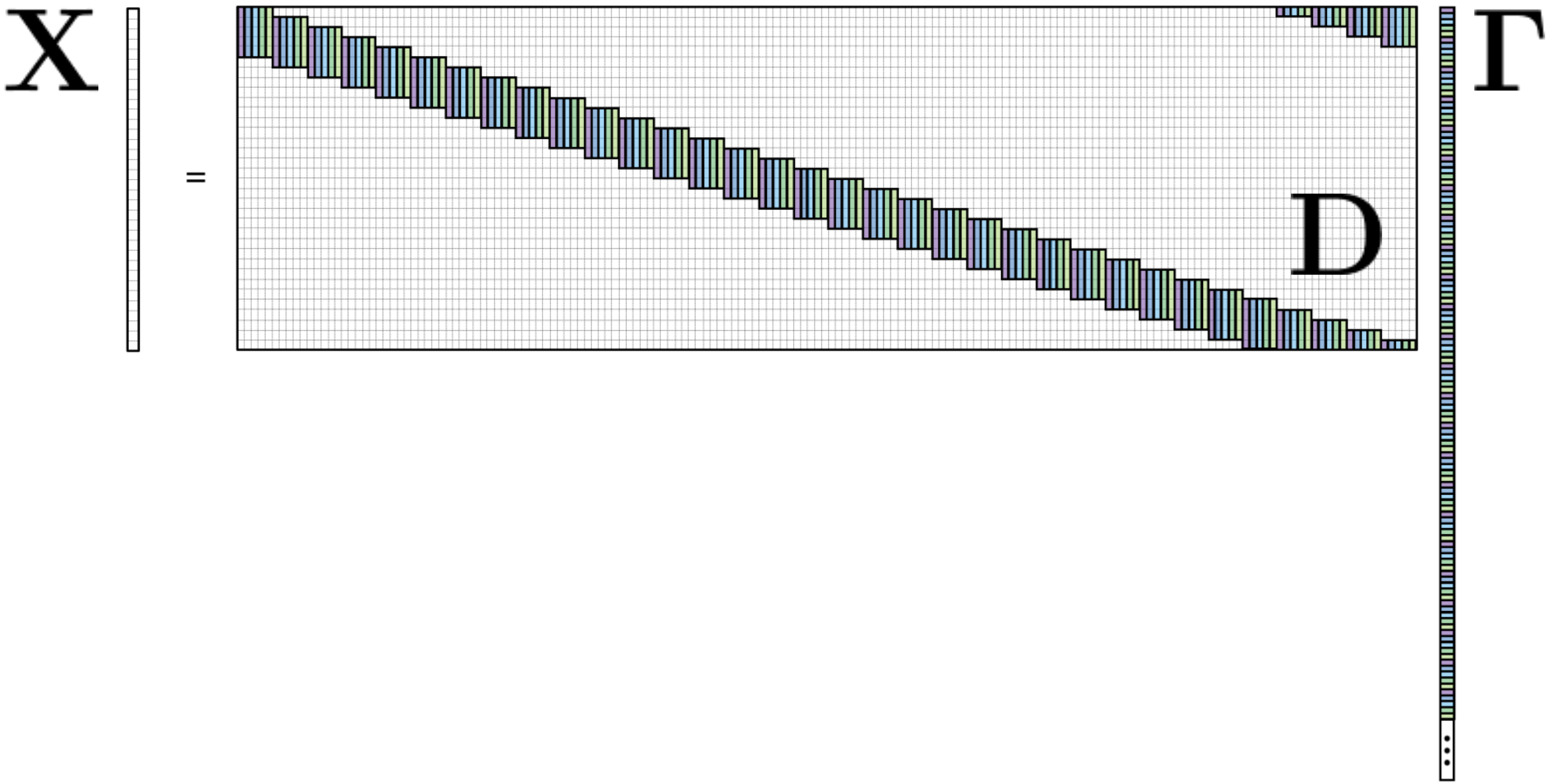


ReLU is equivalent to soft nonnegative thresholding

First layer of a **Convolutional** Neural Network

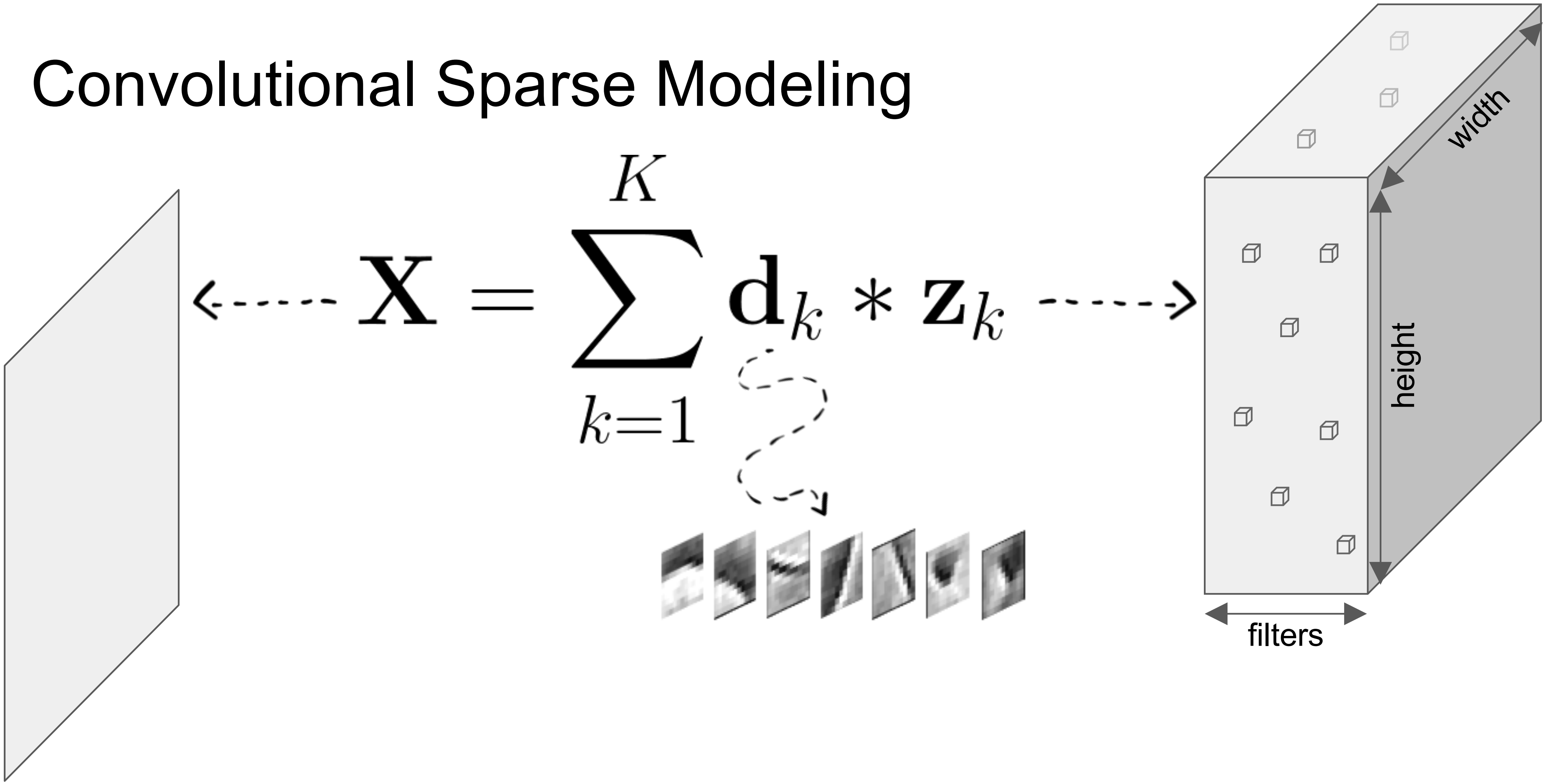


Convolutional Sparse Modeling

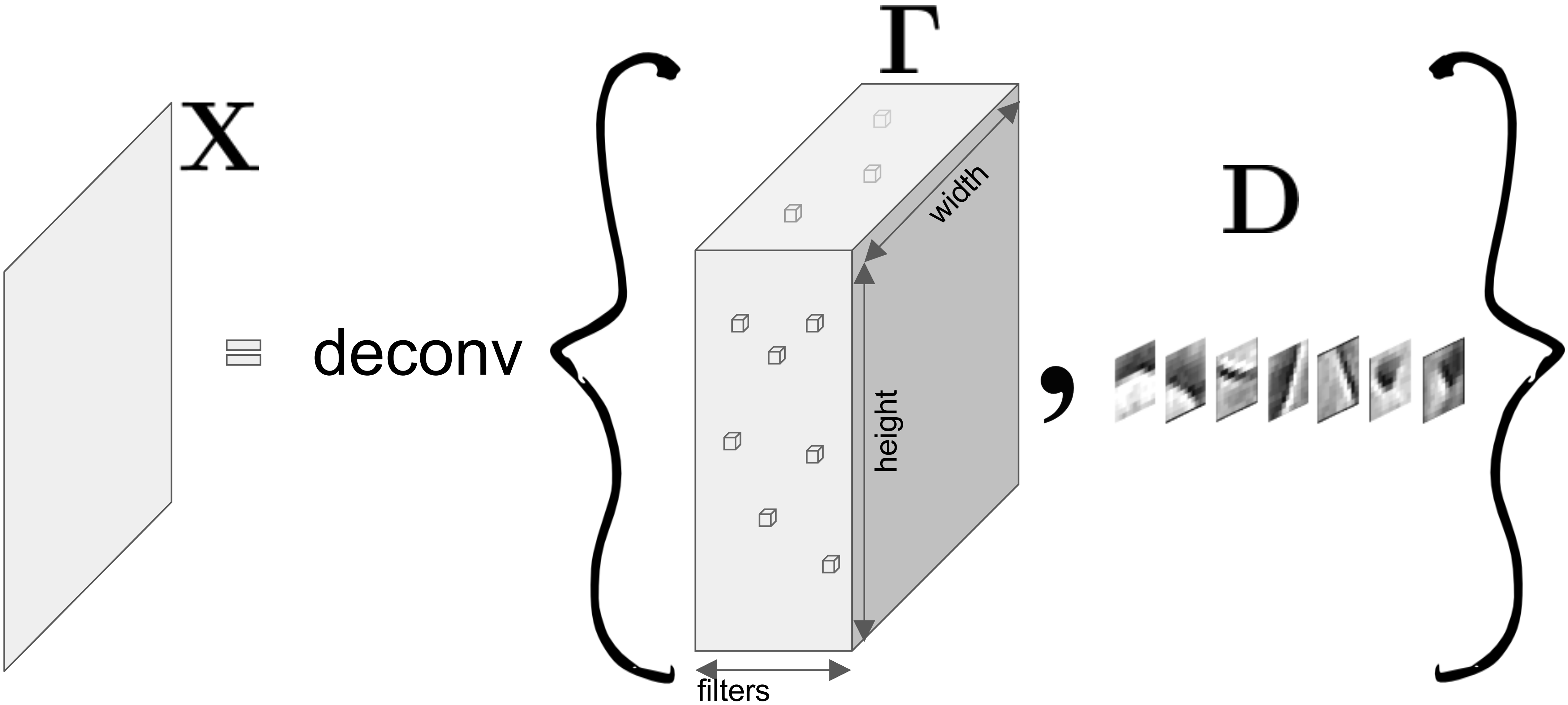


Convolutional Sparse Modeling

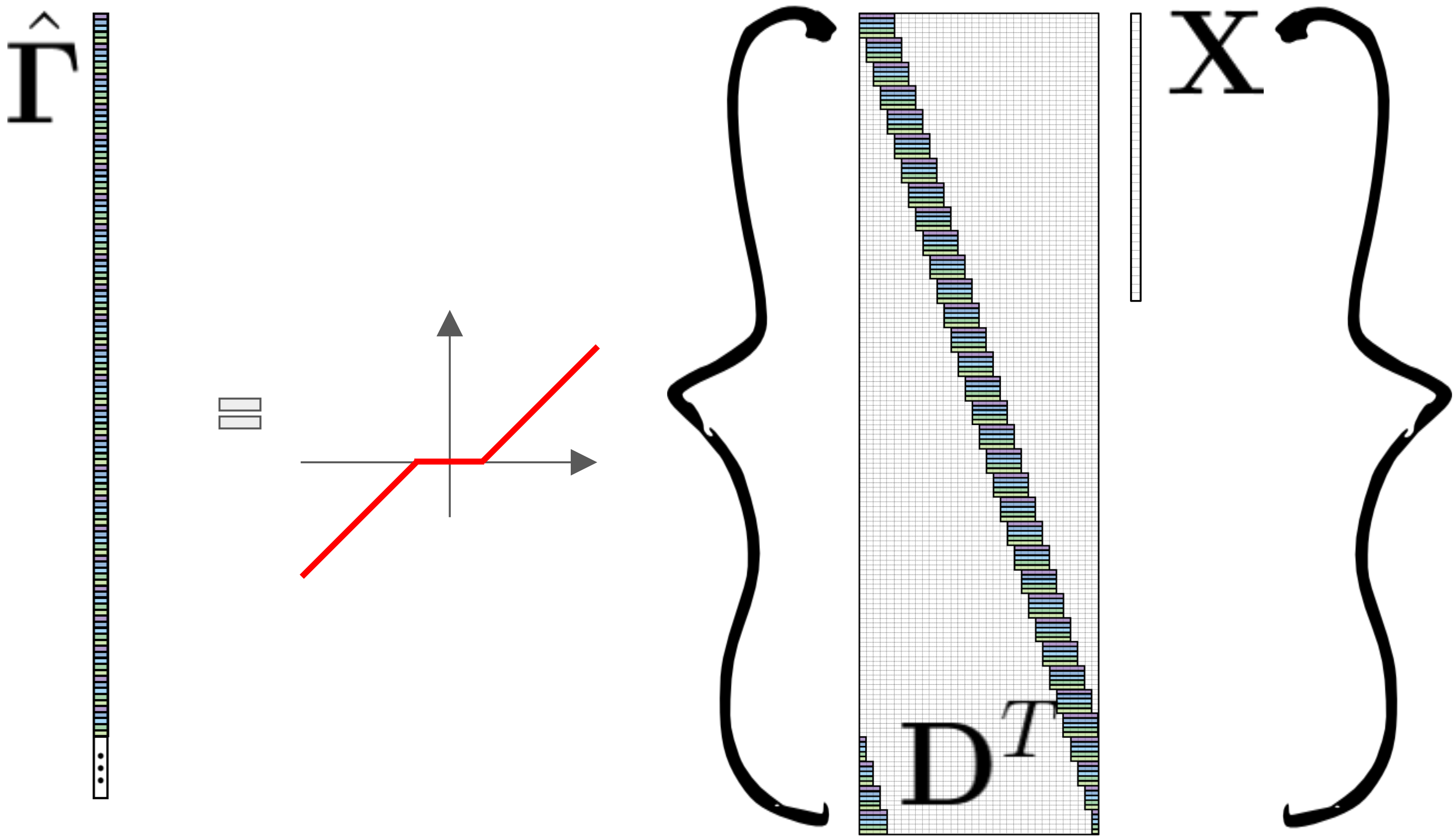
$$\mathbf{X} = \sum_{k=1}^{K} \mathbf{d}_k * \mathbf{z}_k$$



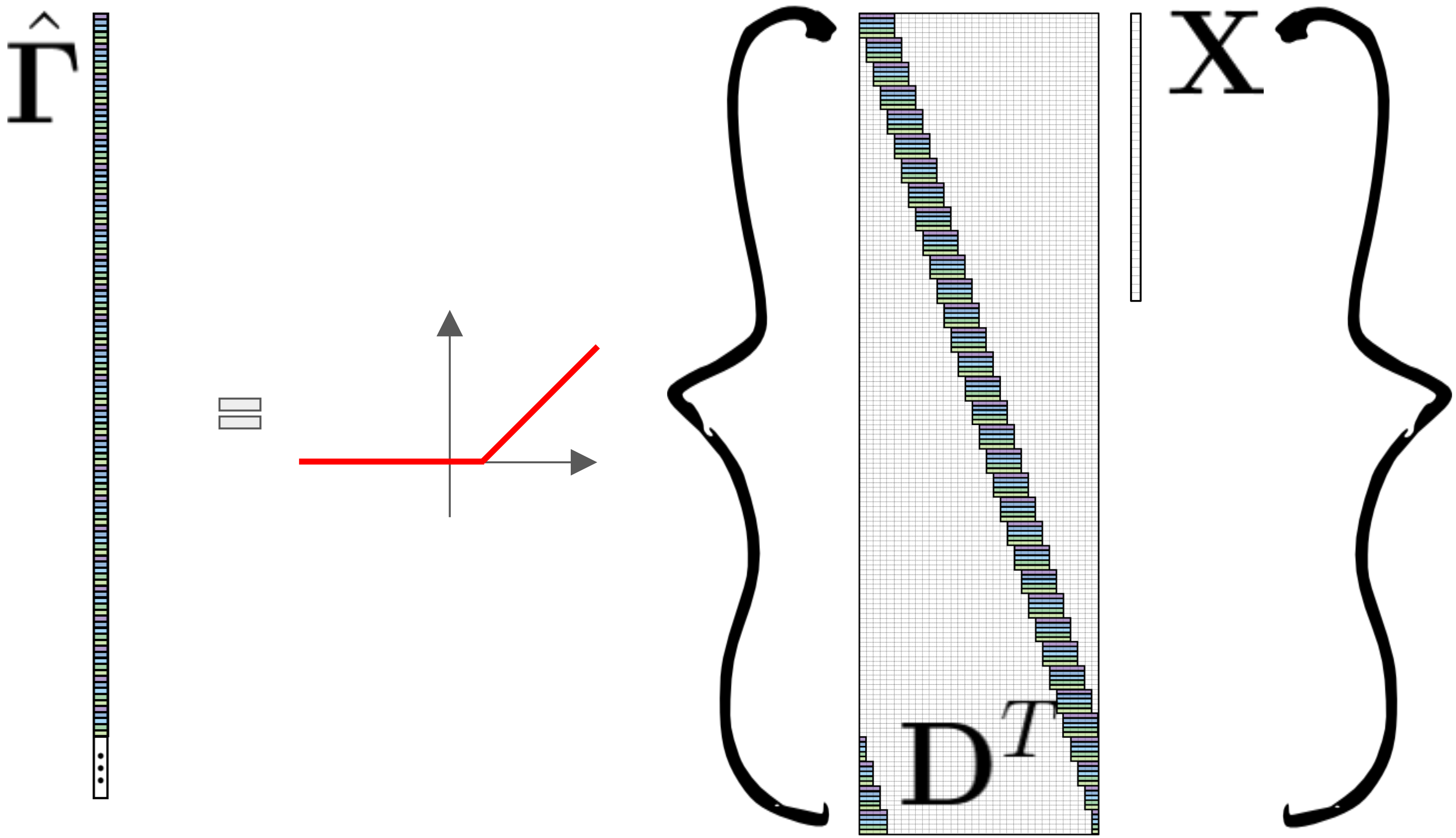
Convolutional Sparse Modeling



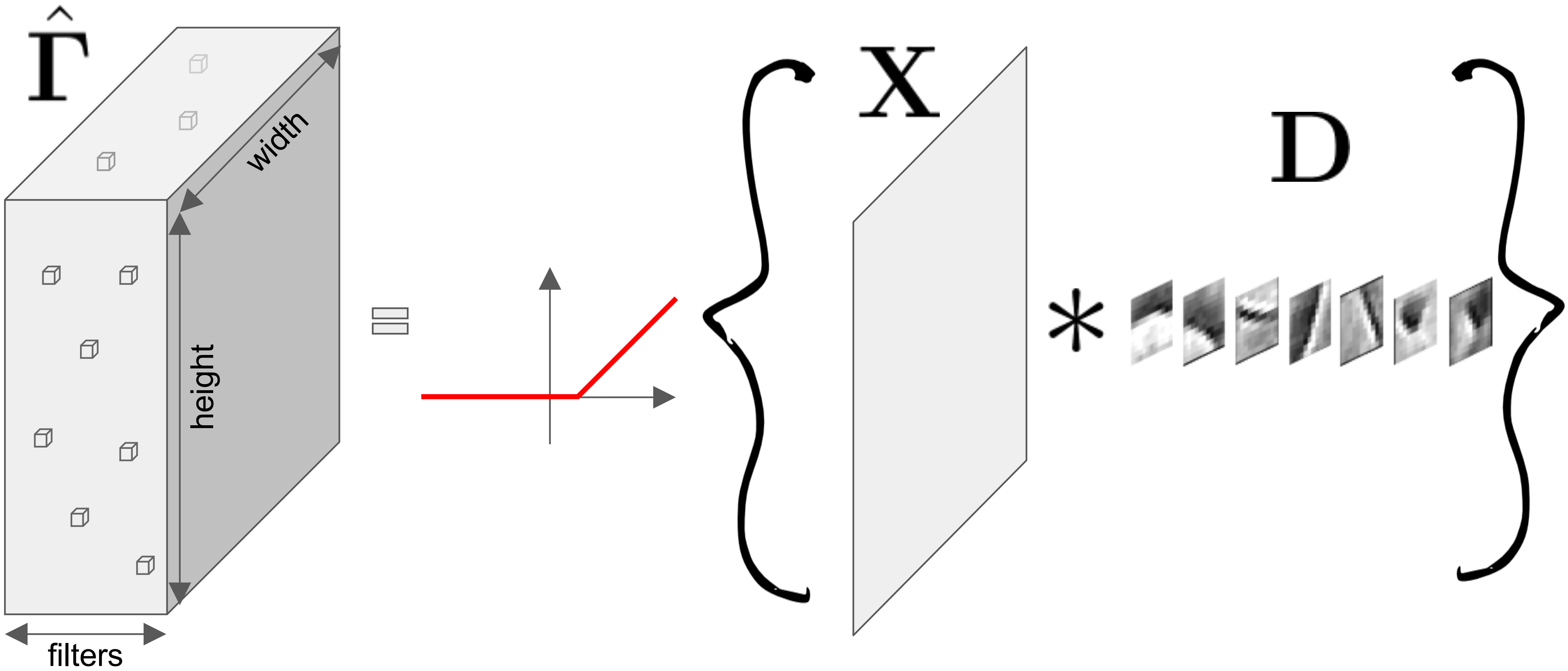
Thresholding Algorithm

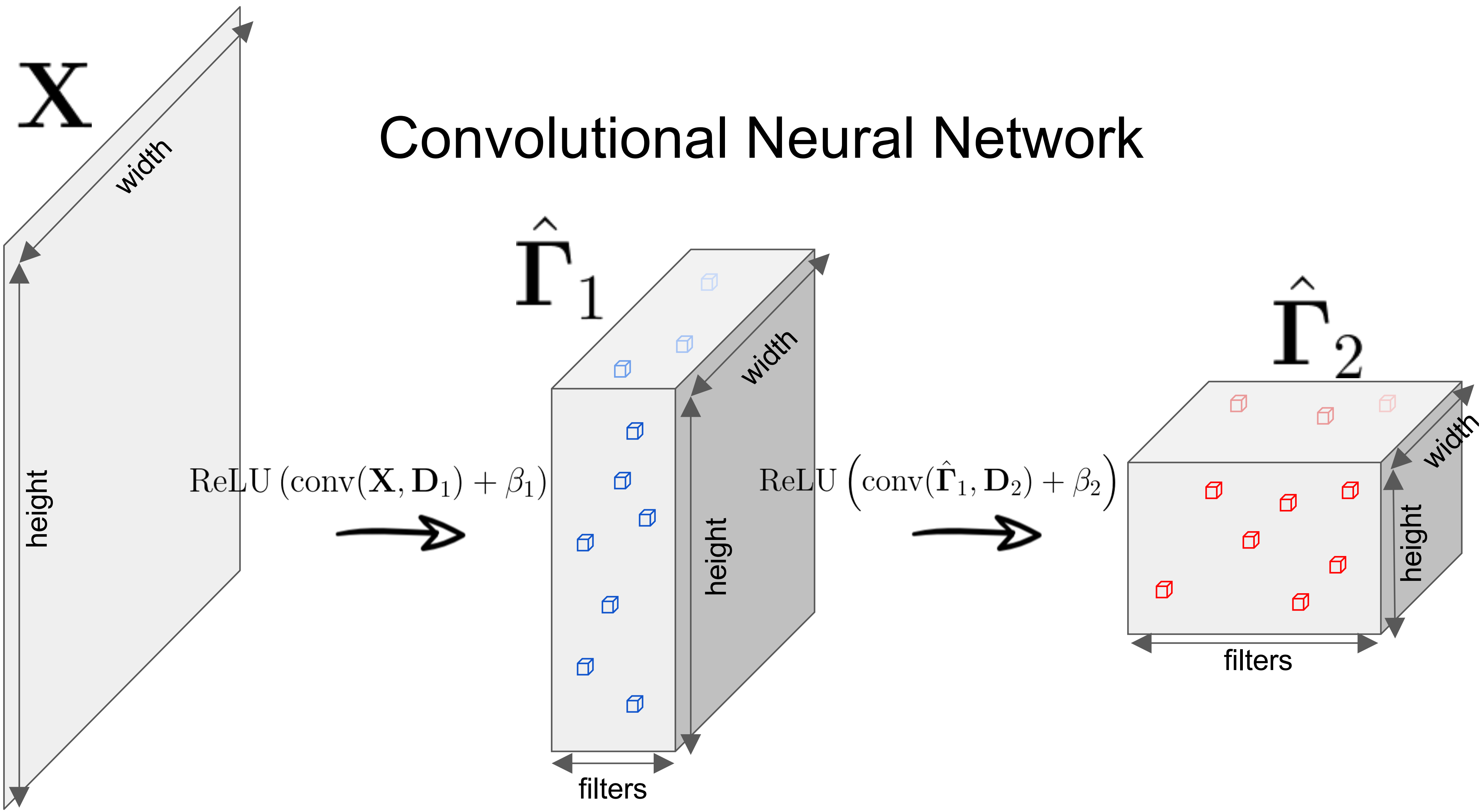


First layer of a Convolutional Neural Network

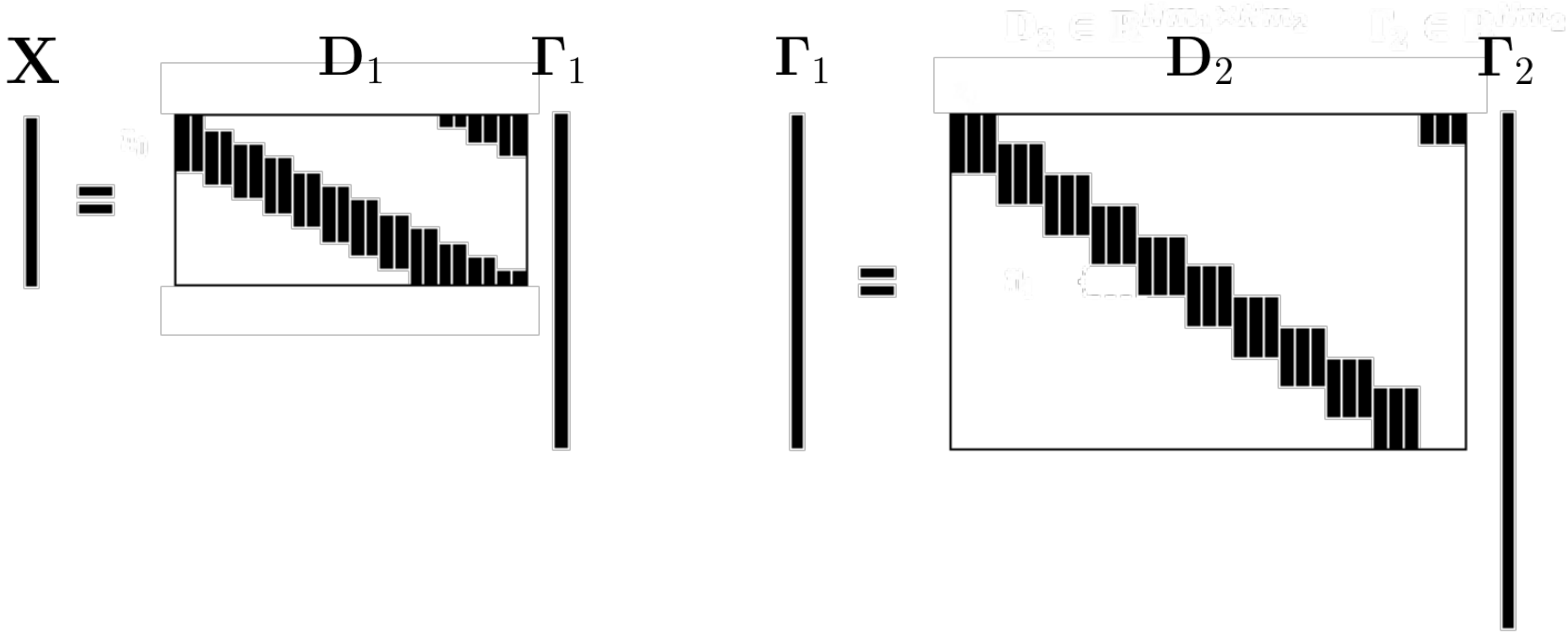


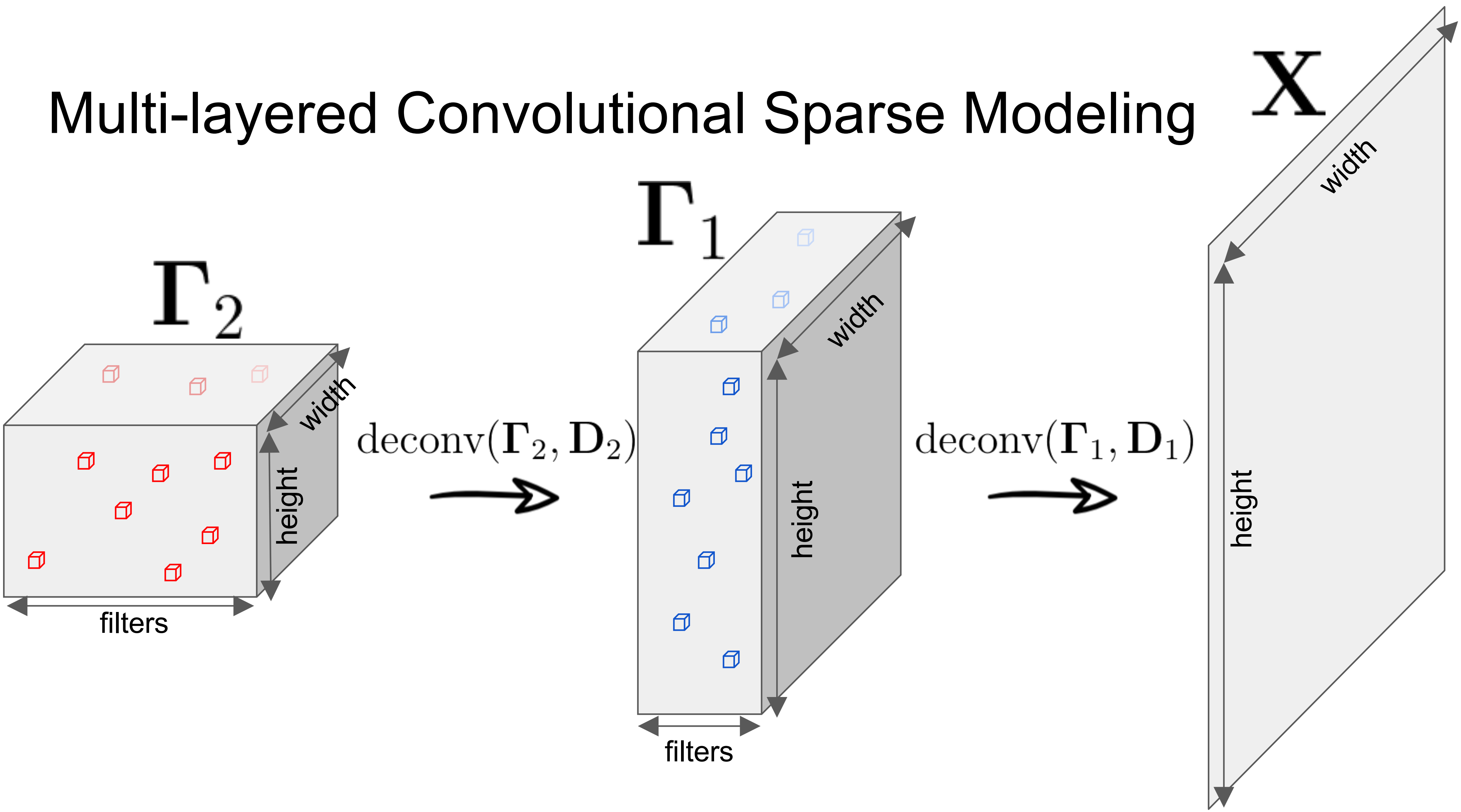
First layer of a Convolutional Neural Network




Convolutional Neural Network
$\mathbf{X}$
width
height
$\mathrm{ReLU}\left(\mathrm{conv}(\mathbf{X}, \mathbf{D}_1) + \beta_1\right)$
$\hat{\Gamma}_1$
width
height
filters
$\mathrm{ReLU}\left(\mathrm{conv}(\hat{\mathbf{\Gamma}}_1, \mathbf{D}_2) + \beta_2\right)$
$\hat{\Gamma}_2$
width
height
filters

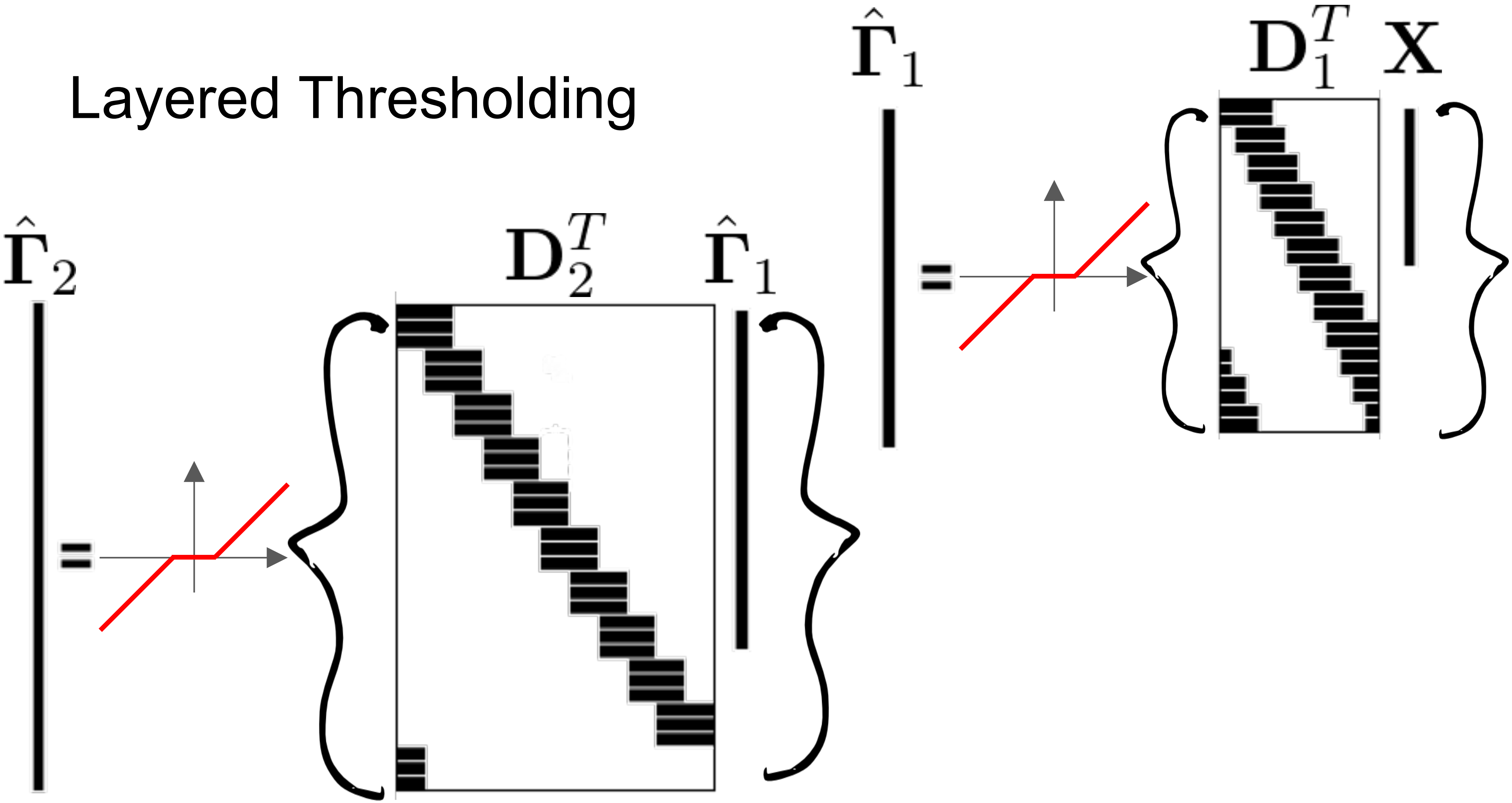
Multi-layered Convolutional Sparse Modeling

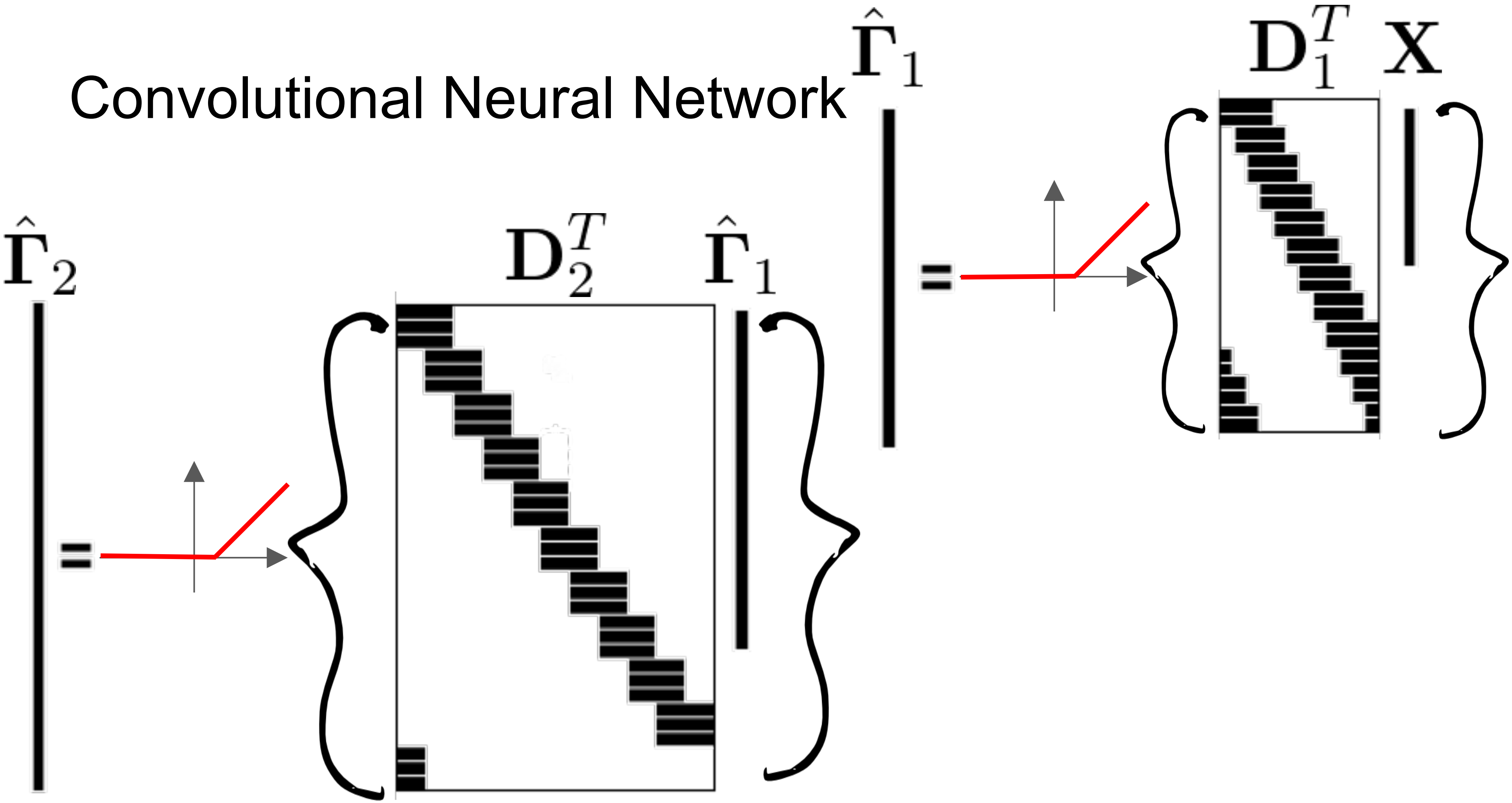


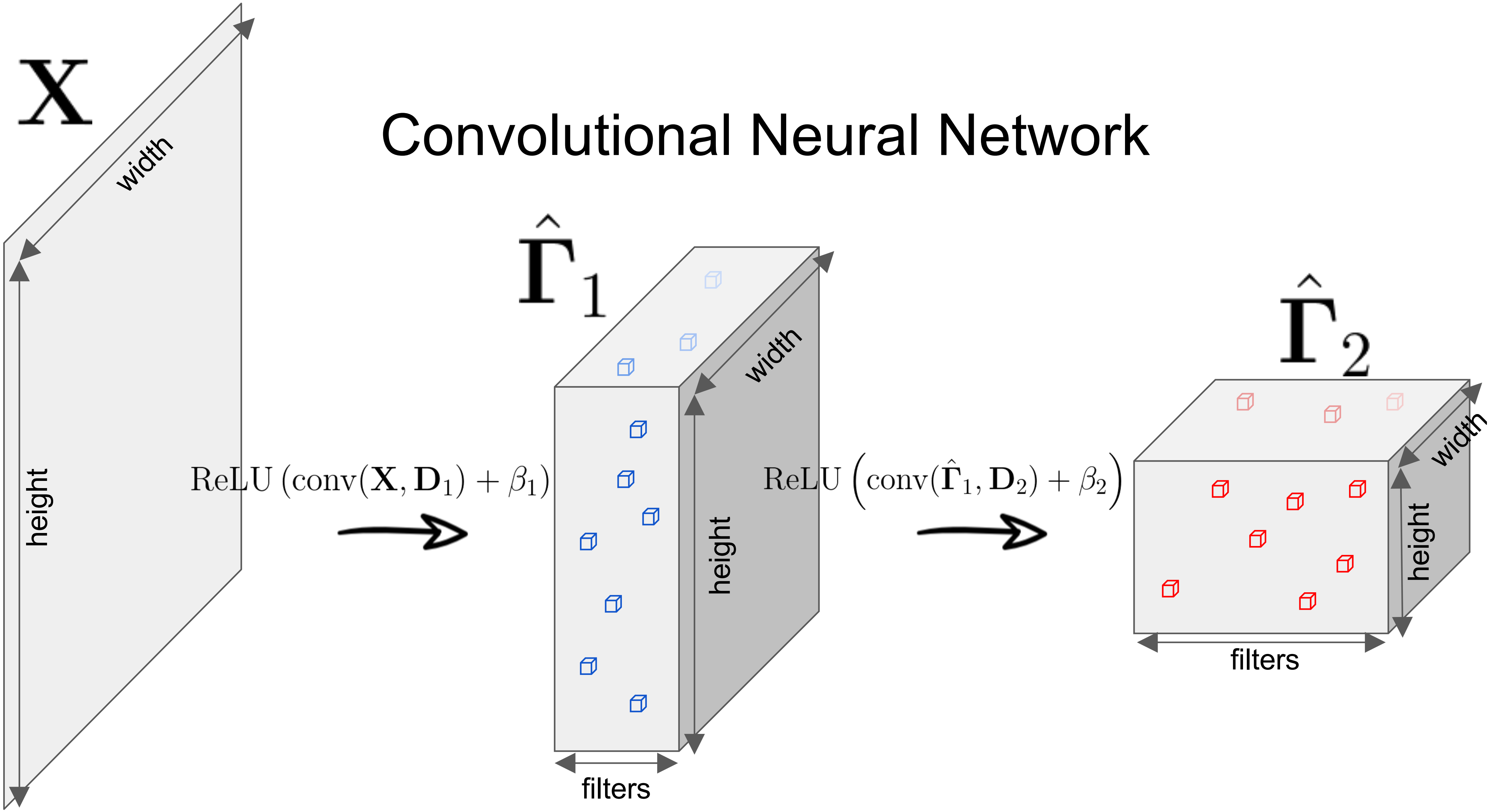


Multi-layered Convolutional Sparse Modeling
$\mathbf{X}$
Γ_2
Γ_1
width
height
filters
$\text{deconv}(\mathbf{\Gamma}_2, \mathbf{D}_2)$
$\text{deconv}(\mathbf{\Gamma}_1, \mathbf{D}_1)$
width
height
filters
width
height

Layered Thresholding



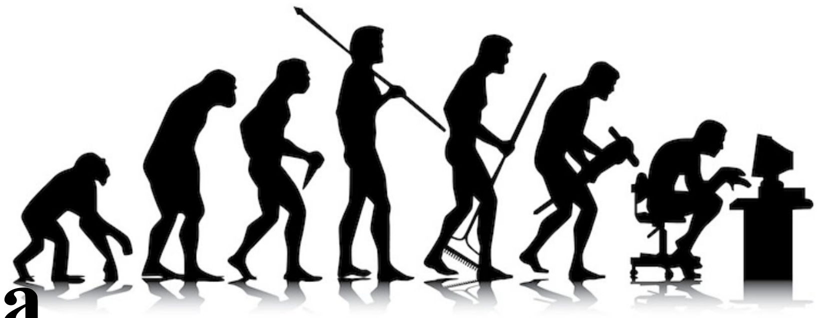

Convolutional Neural Network
$\hat{\mathbf{\Gamma}}_1$
$\mathbf{D}_1^T$
$\mathbf{X}$
$\hat{\mathbf{\Gamma}}_2$
$\mathbf{D}_2^T$
$\hat{\mathbf{\Gamma}}_1$



Convolutional Neural Network
$\mathbf{X}$
width
height
$\hat{\Gamma}_1$
$\mathrm{ReLU}\left(\mathrm{conv}(\mathbf{X}, \mathbf{D}_1) + \beta_1\right)$
width
height
filters
$\hat{\Gamma}_2$
$\mathrm{ReLU}\left(\mathrm{conv}(\hat{\mathbf{\Gamma}}_1, \mathbf{D}_2) + \beta_2\right)$
width
height
filters

Theories of Deep Learning

Evolution of Models



Multi-Layered Convolutional Neural Network

First Layer of a Convolutional Neural Network

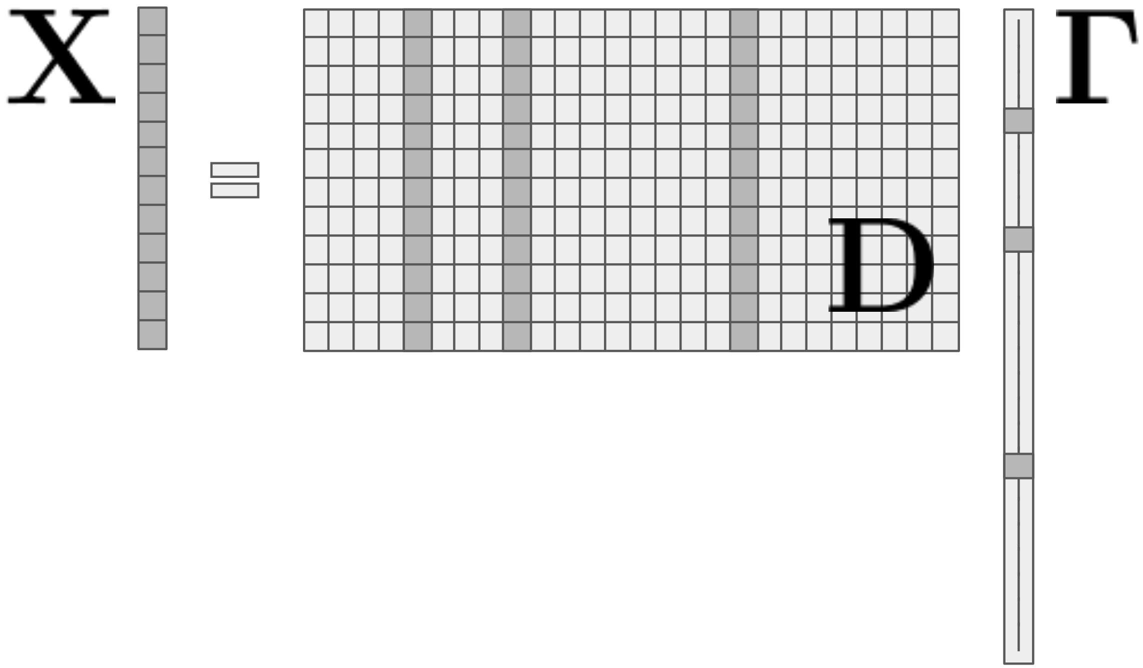
First Layer of a Neural Network

Sparse representations

Convolutional sparse representation

Multi-Layered Convolutional Sparse Representatio
n

Sparse Modeling



Classic Sparse Theory

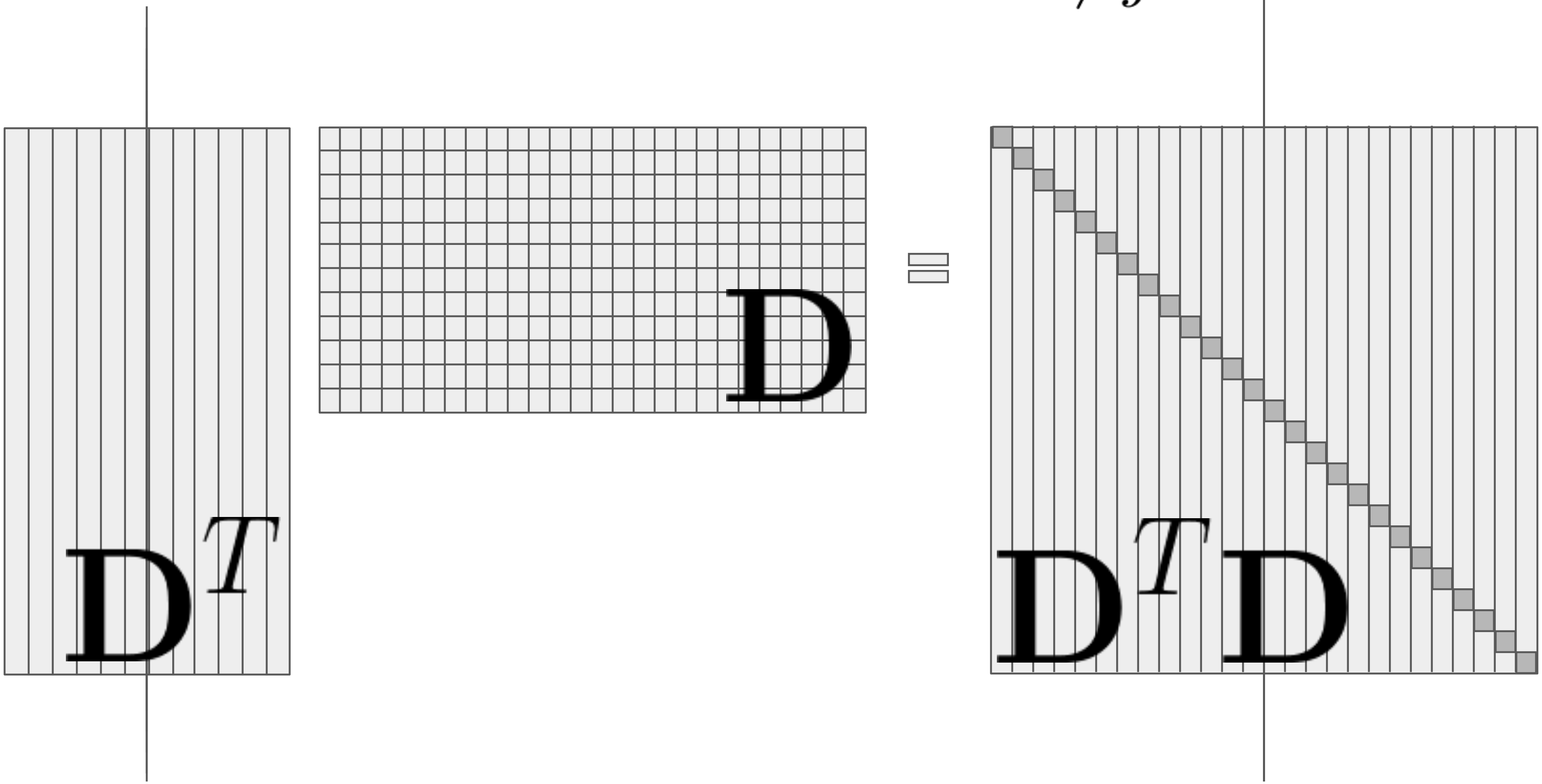
$$\mathbf{X} = \mathbf{D}\mathbf{\Gamma}$$

$$\hat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} \ \|\mathbf{\Gamma}\|_1 \ \text{ s.t. } \ \mathbf{X} = \mathbf{D}\mathbf{\Gamma}$$

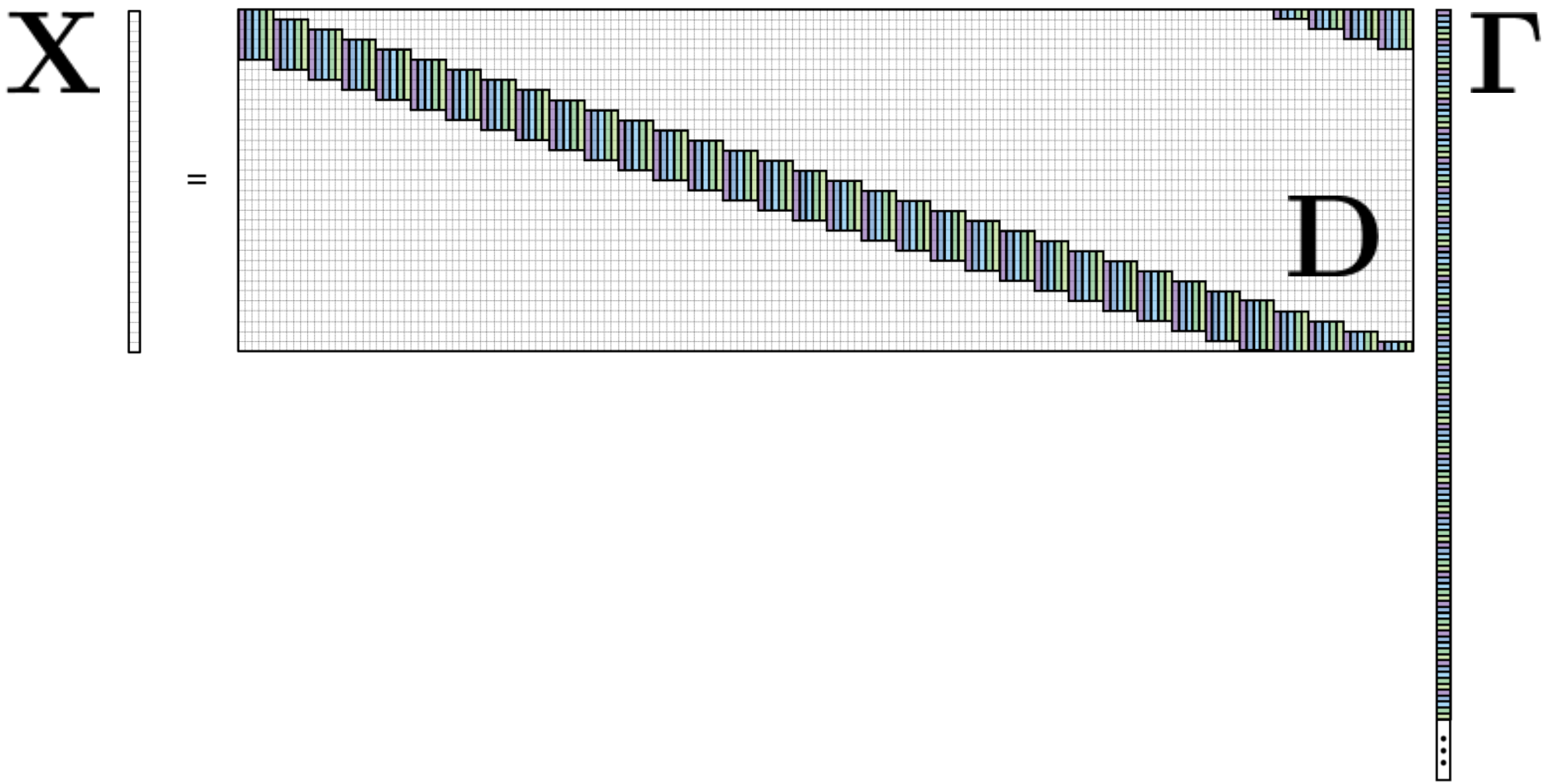
Theorem: [Donoho and Elad, 2003]
Basis pursuit is guaranteed to recover the true sparse vector assuming that

$$\|\mathbf{\Gamma}\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right)$$

Mutual Coherence: $\mu(\mathbf{D}) = \max_{i \neq j} \left| \left(\mathbf{D}^T \mathbf{D}\right)_{i,j} \right|$



Convolutional Sparse Modeling



Classic Sparse Theory for Convolutional Case

Theorem: [Donoho and Elad, 2003]

Basis pursuit is guaranteed to recover the true sparse vector assuming that

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

Assuming 2 atoms of length 64 $\mu(\mathbf{D}) \geq 0.063$ [Welch, 1974]

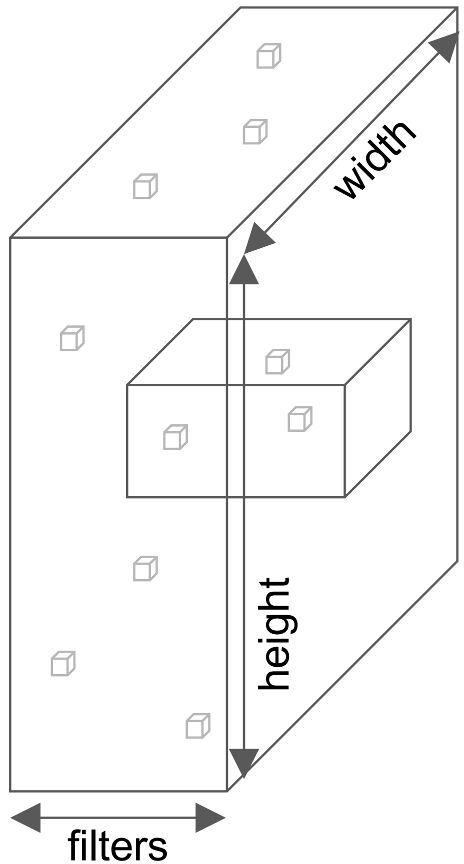
Success guaranteed when $\|\boldsymbol{\Gamma}\|_0 < 8.43$

Very pessimistic!

Local Sparsity

$\|\boldsymbol{\Gamma}\|_{0,\infty}$ maximal number of non-zeroes in a local neighborhood

$$\min_{\boldsymbol{\Gamma}} \ \|\boldsymbol{\Gamma}\|_{0,\infty} \quad \text{s.t.} \quad \mathbf{X} = \mathbf{D}\boldsymbol{\Gamma}$$



Success of Basis Pursuit

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

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

Theorem: [Papyan, Sulam and Elad, 2016]

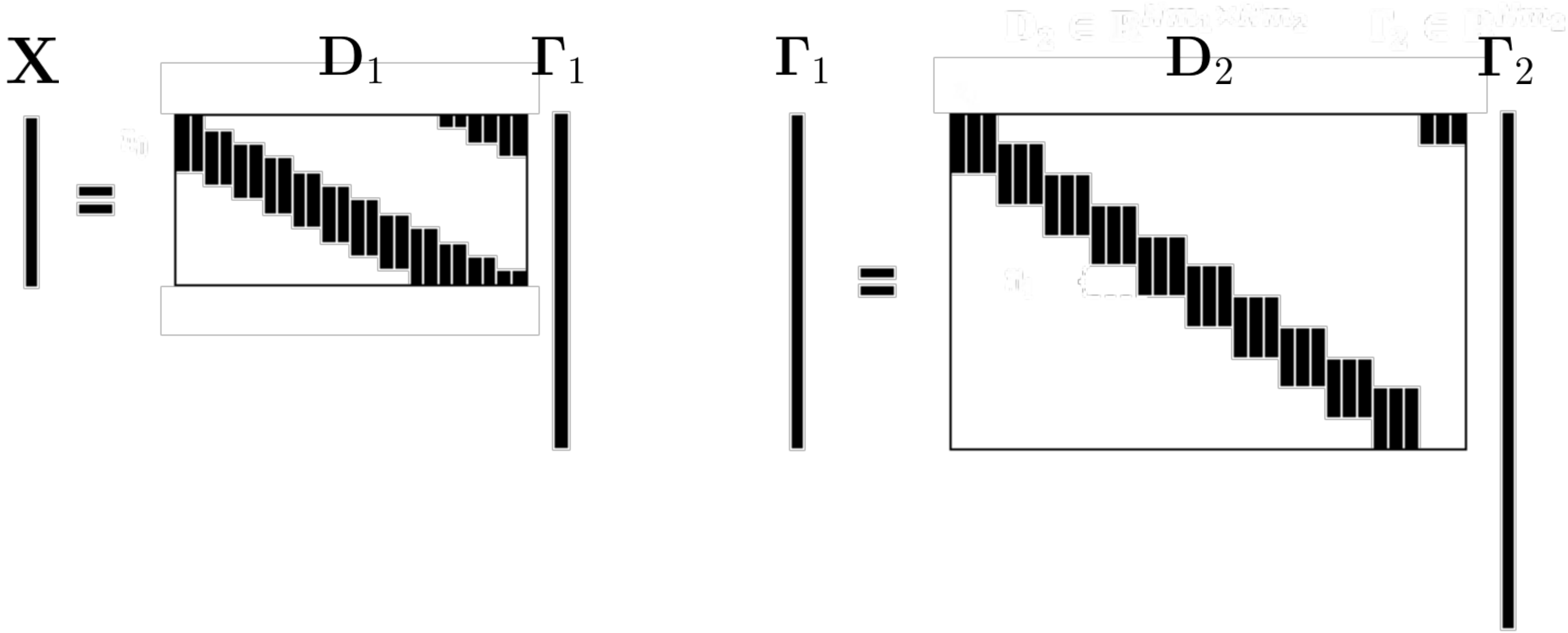
Assume: $\|\mathbf{\Gamma}\|_{0,\infty} < \frac{1}{3}\left(1 + \frac{1}{\mu(\mathbf{D})}\right)$

Then: $\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\infty} \leq 7.5\|\mathbf{E}\|_{2,\infty}$

Theoretical guarantee for:

- [Zeiler et. al 2010]
- [Wohlberg 2013]
- [Bristow et. al 2013]
- [Fowlkes and Kong 2014]
- [Zhou et. al 2014]
- [Kong and Fowlkes 2014]
- [Zhu and Lucey 2015]
- [Heide et. al 2015]
- [Gu et. al 2015]
- [Wohlberg 2016]
- [Šorel and Šroubek 2016]
- [Serrano et. al 2016]
- [Papyan et. al 2017]
- [Garcia-Cardona and Wohlberg 2017]
- [Wohlberg and Rodriguez 2017]
- ...

Multi-layered Convolutional Sparse Modeling



Deep Coding Problem

Given $\mathbf{X}$, find a set of representations satisfying:

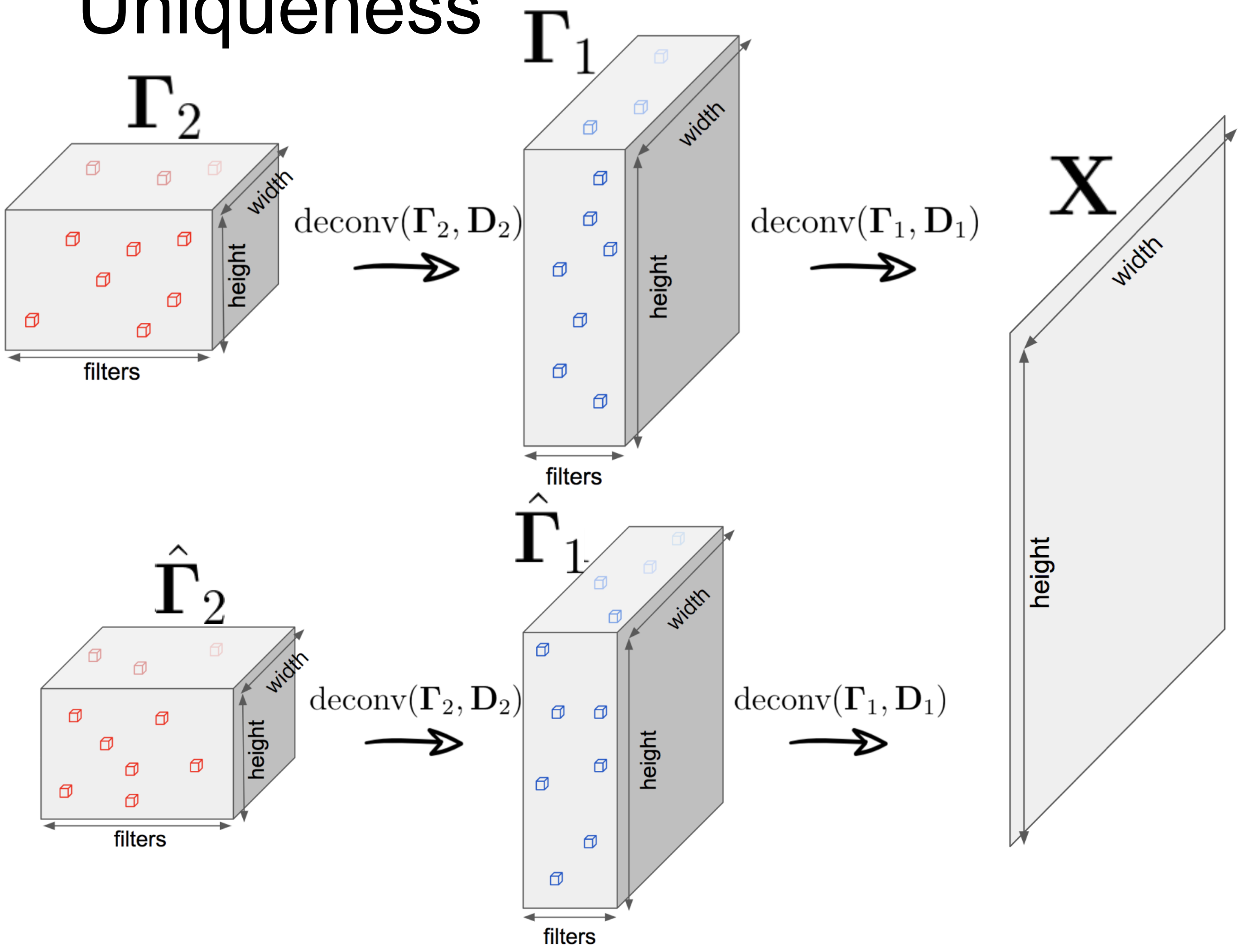
$$
\begin{aligned}
\mathbf{X} &= \mathbf{D}_1\mathbf{\Gamma}_1, & \|\mathbf{\Gamma}_1\|_{0,\infty} &\leq \lambda_1 \\
\mathbf{\Gamma}_1 &= \mathbf{D}_2\mathbf{\Gamma}_2, & \|\mathbf{\Gamma}_2\|_{0,\infty} &\leq \lambda_2 \\
&\vdots \\
\mathbf{\Gamma}_{L-1} &= \mathbf{D}_L\mathbf{\Gamma}_L, & \|\mathbf{\Gamma}_L\|_{0,\infty} &\leq \lambda_L
\end{aligned}
$$

Deep Coding Problem

Given $\mathbf{Y}$, find a set of representations satisfying:

$$
\begin{aligned}
\|\mathbf{Y} - \mathbf{D}_1\mathbf{\Gamma}_1\|_2 \leq \epsilon, &\quad \|\mathbf{\Gamma}_1\|_{0,\infty} \leq \lambda_1 \\
\mathbf{\Gamma}_1 = \mathbf{D}_2\mathbf{\Gamma}_2, &\quad \|\mathbf{\Gamma}_2\|_{0,\infty} \leq \lambda_2 \\
&\vdots \\
\mathbf{\Gamma}_{L-1} = \mathbf{D}_L\mathbf{\Gamma}_L, &\quad \|\mathbf{\Gamma}_L\|_{0,\infty} \leq \lambda_L
\end{aligned}
$$

Uniqueness



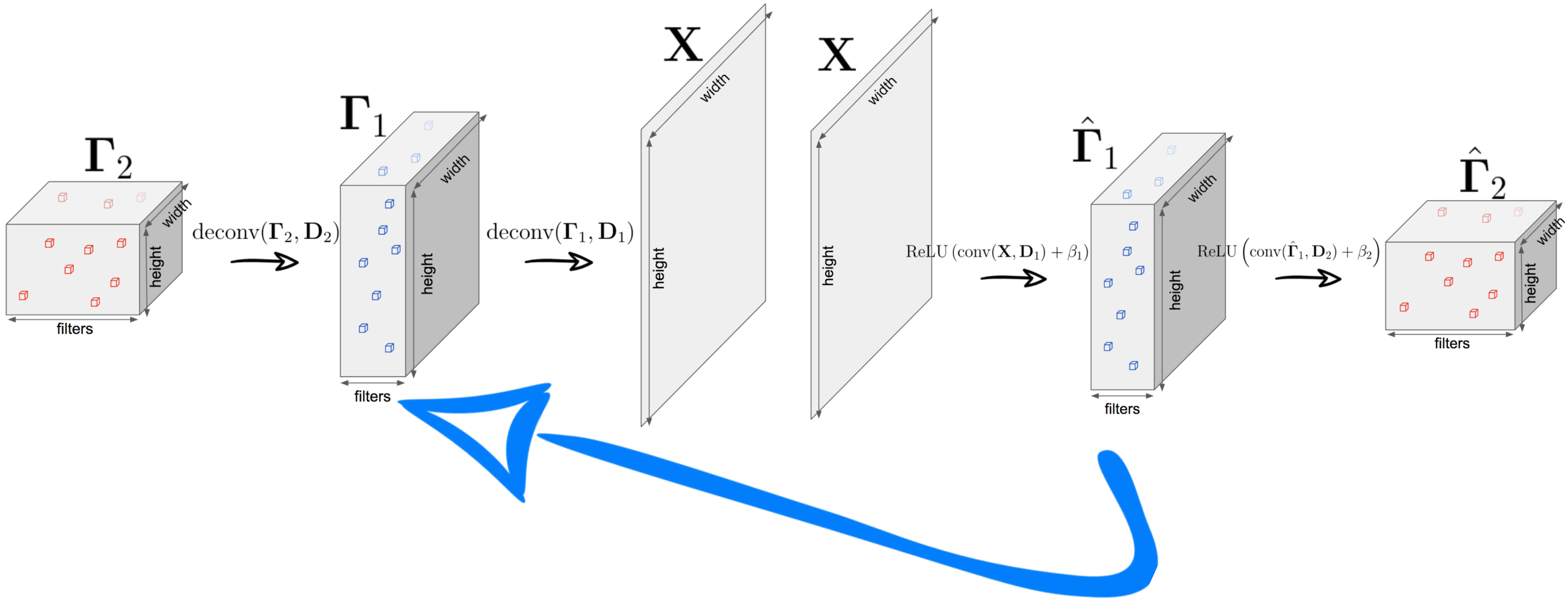
$\mathbf{\Gamma}_2$
width
height
filters
$\text{deconv}(\mathbf{\Gamma}_2, \mathbf{D}_2)$
$\mathbf{\Gamma}_1$
width
height
filters
$\text{deconv}(\mathbf{\Gamma}_1, \mathbf{D}_1)$
$\mathbf{X}$
width
height
$\hat{\mathbf{\Gamma}}_2$
width
height
filters
$\text{deconv}(\mathbf{\Gamma}_2, \mathbf{D}_2)$
$\hat{\mathbf{\Gamma}}_1$
width
height
filters
$\text{deconv}(\mathbf{\Gamma}_1, \mathbf{D}_1)$

Uniqueness Theorem

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

$$\{\mathbf{\Gamma}_l\}_{l=1}^{L} \text{ are the unique feature maps of } \mathbf{X}$$

Success of Forward Pass



Success of Forward Pass Theorem

$$\|\mathbf{\Gamma}_l\|_{0,\infty} < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D}_l)}\frac{|\Gamma_l^{\min}|}{|\Gamma_l^{\max}|}\right) - \frac{1}{\mu(\mathbf{D}_l)}\frac{\epsilon_{l-1}}{|\Gamma_l^{\max}|}$$

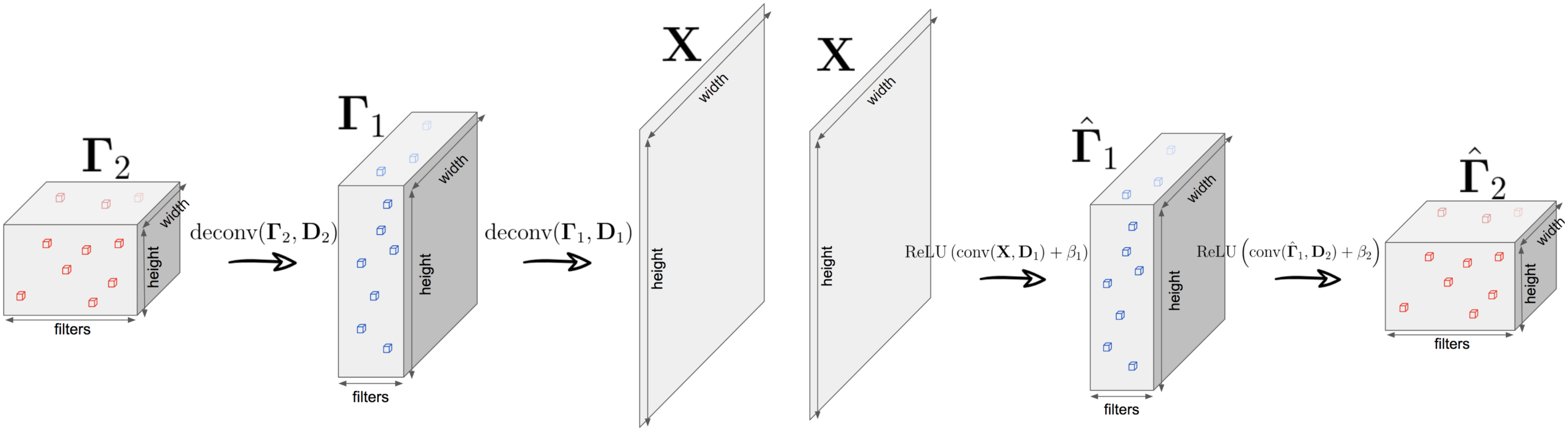
Layered thresholding guaranteed:

1. Find correct places of nonzeros

$$\|\hat{\mathbf{\Gamma}}_l - \mathbf{\Gamma}_l\|_{2,\infty} \leq \epsilon_l$$

- Forward pass always fails at recovering representations exactly
- Success depends on ratio
- Distance increases with layer

Generative Model and Crude Inference



Layered Lasso

\# StatsDepartment

$$\hat{\mathbf{\Gamma}}_1 = \arg\min_{\mathbf{\Gamma}_1} \ \frac{1}{2}\|\mathbf{Y} - \mathbf{D}_1\mathbf{\Gamma}_1\|_2^2 + \alpha_1\|\mathbf{\Gamma}_1\|_1$$

$$\hat{\mathbf{\Gamma}}_2 = \arg\min_{\mathbf{\Gamma}_2} \ \frac{1}{2}\|\hat{\mathbf{\Gamma}}_1 - \mathbf{D}_2\mathbf{\Gamma}_2\|_2^2 + \alpha_2\|\mathbf{\Gamma}_2\|_1$$

Success of Layered Lasso

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

Layered Lasso guaranteed:

1. Find only correct places of nonzeros
2. Find all coefficients that are big enough

$$\|\hat{\mathbf{\Gamma}}_l - \mathbf{\Gamma}_l\|_{2,\infty} \le \epsilon_l$$

~~Forward pass always fails at recovering representations exactly~~

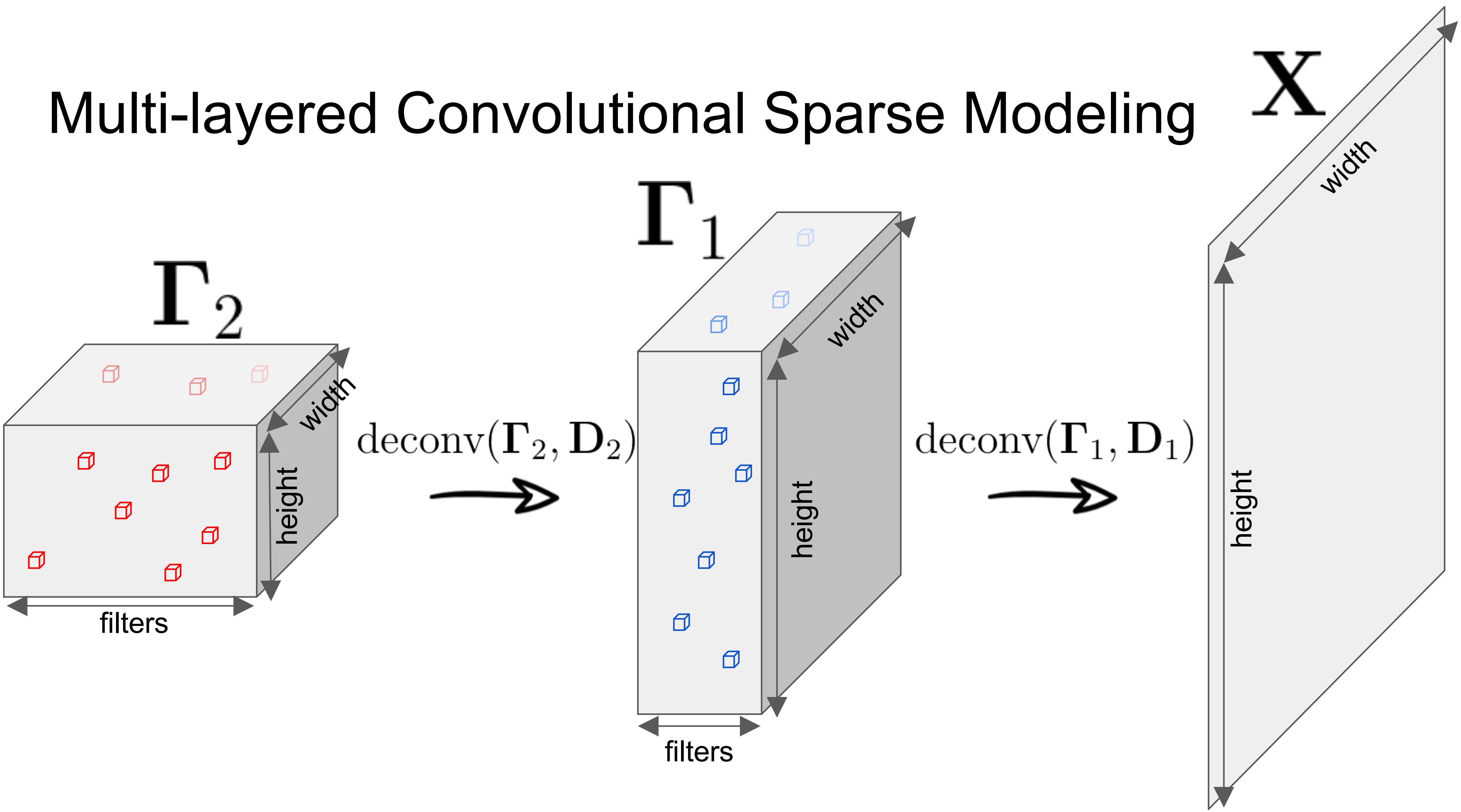
~~Success depends on ratio~~

Distance increases with layer

Layered Iterative Thresholding

$$\mathbf{\Gamma}_1^t = \mathcal{S}_{\alpha_1}\left(\mathbf{D}_1^T\mathbf{Y} + \left(\mathbf{I} - \mathbf{D}_1^T\mathbf{D}_1\right)\mathbf{\Gamma}_1^{t-1}\right)$$

$$\mathbf{\Gamma}_2^t = \mathcal{S}_{\alpha_2}\left(\mathbf{D}_2^T\hat{\mathbf{\Gamma}}_1 + \left(\mathbf{I} - \mathbf{D}_2^T\mathbf{D}_2\right)\mathbf{\Gamma}_2^{t-1}\right)$$


Multi-layered Convolutional Sparse Modeling
$\mathbf{X}$
$\mathbf{\Gamma}_1$
$\mathbf{\Gamma}_2$
width
height
filters
$\mathrm{deconv}(\mathbf{\Gamma}_2, \mathbf{D}_2)$
width
height
filters
$\mathrm{deconv}(\mathbf{\Gamma}_1, \mathbf{D}_1)$
width
height

Summary

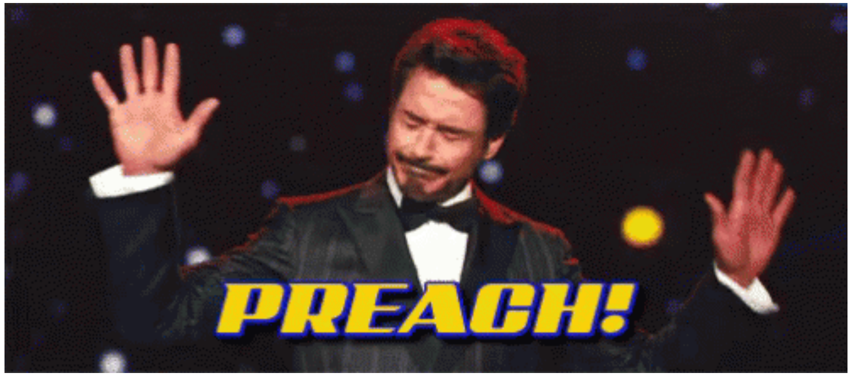
1  Sparsity well established theoretically

2  Sparsity is covertly exploited in practice: ReLU, dropout, stride, dilation, ...

3  Sparsity is the secret sauce behind CNN

4  Need to bring sparsity to the surface to better understand CNNs

5  Andrej Karpathy agrees


PREACH!