X^3

Sparse Coding
Dictionary Learning
Compressed Sensing

William Edward Hahn
Elan Barenholtz
Michael Teti
Stephanie Lewkowitz

NICE Portland 2018

Outline

Sparse Coding

Dictionary Learning

Compressed Sensing

The perceptron program is not primarily concerned with the invention of devices for "artificial intelligence", but rather with investigating the physical structures and neurodynamic principles which underlie "**natural intelligence**". A perceptron is first and foremost a **brain model**, not an invention for pattern recognition. As a brain model, its utility is in enabling us to determine the physical conditions for the emergence of various psychological properties.

Frank Rosenblatt, 1962

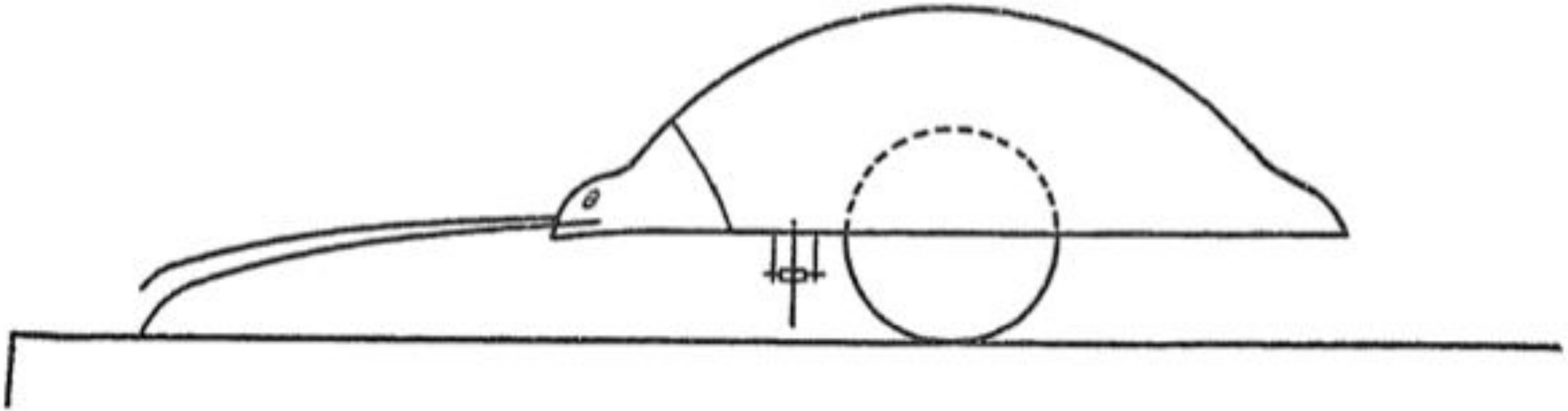


FIG. 69. MECHANICAL WALKING BEETLE, EXHIBITING THE SEVERAL CHARACTERISTIC ELEMENTS OF THE CORRELATING APPARATUS

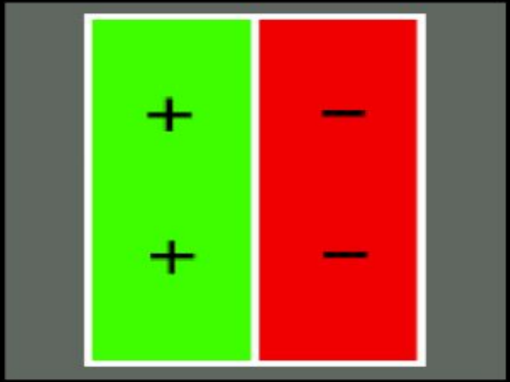
Elements of Physical Biology 1925

Robots as Model Organisms

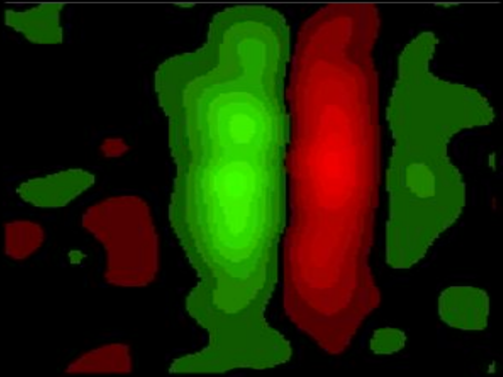
represent the most relevant visual information with the fewest physical and metabolic resources

First stage of visual processing in brain: V1

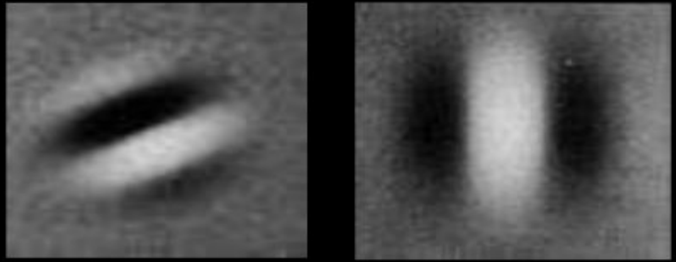
The first stage of visual processing in the brain (V1) does "edge detection."



Schematic of simple cell



Actual simple cell



"Gabor functions."

[Images from DeAngelis, Ohzawa & Freeman, 1995]

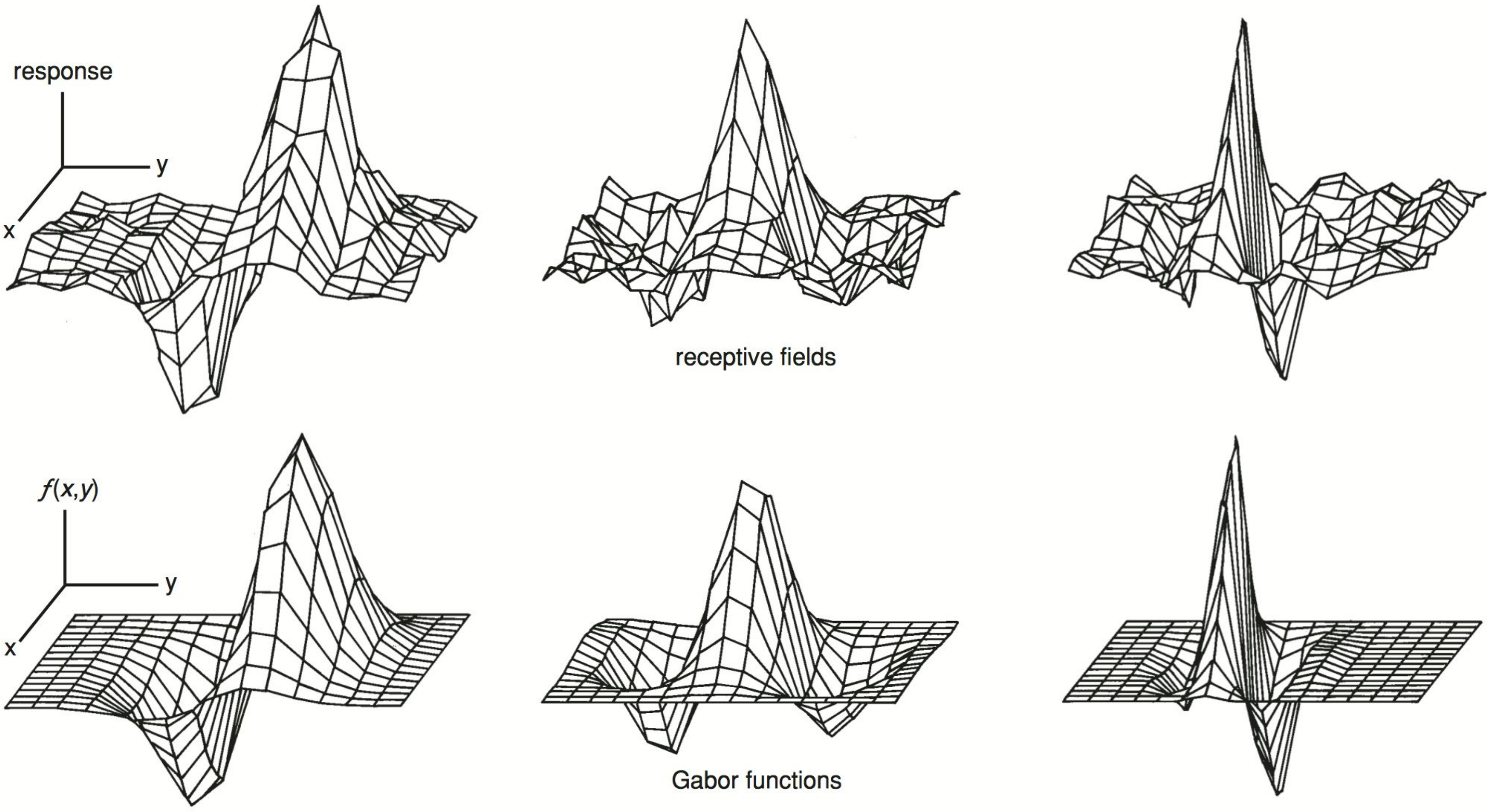


Figure 7. Receptive fields of neurons in the visual cortex of cats *(top)* resemble certain two-dimensional Gabor functions *(bottom)*. The neural circuitry of the visual system may adopt such forms of response because they are well suited to encode images efficiently. (After Daugman, 1989.)

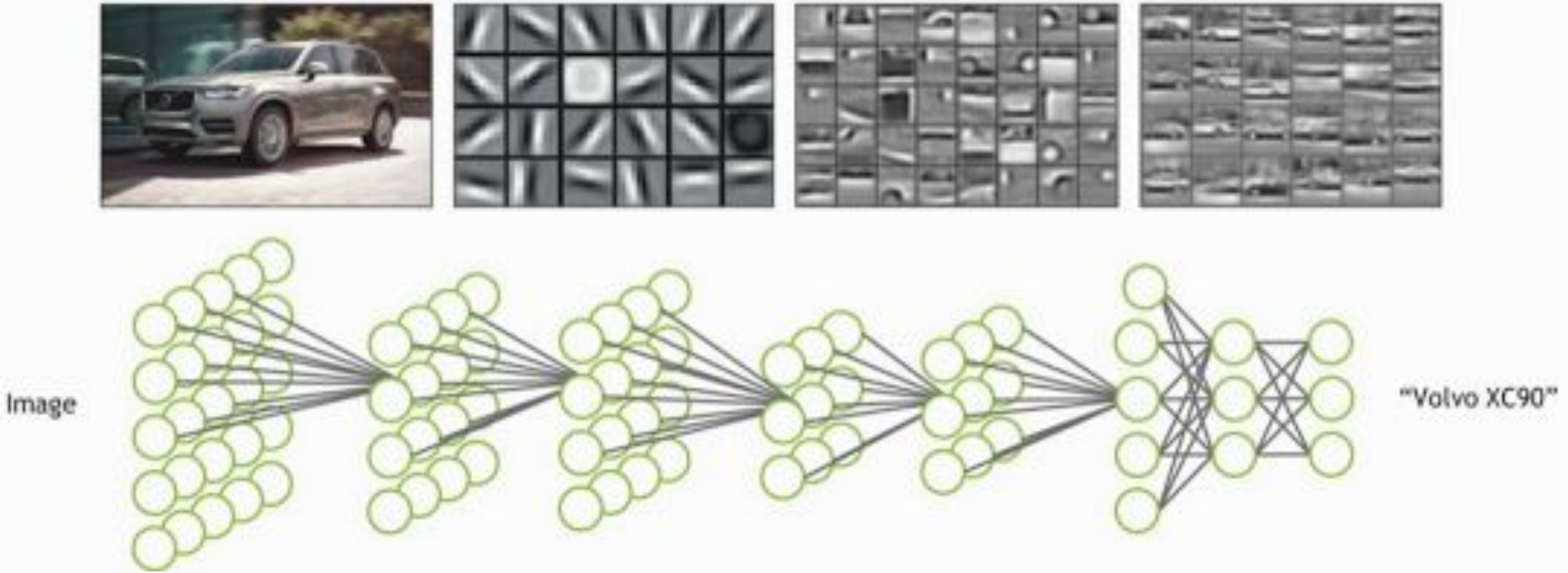
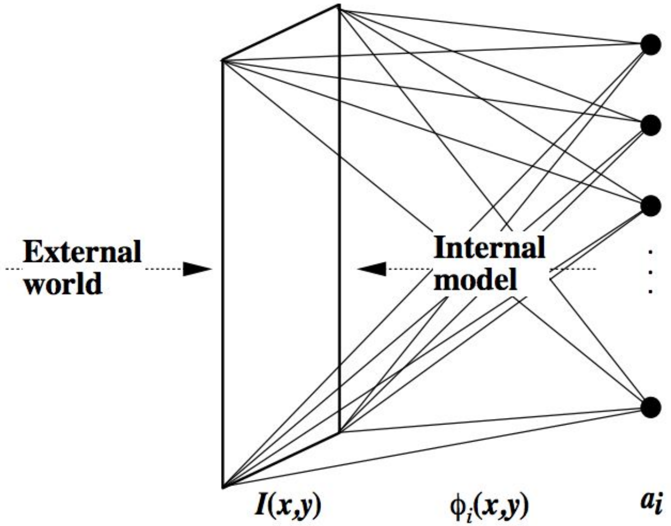


Image
"Volvo XC90"



Sparse coding image model

$$I(x,y) = \sum_i a_i \phi_i(x,y) + \epsilon(x,y)$$

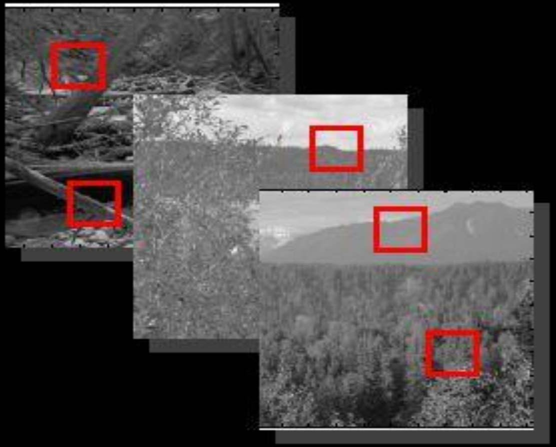
image

neural activities (sparse)

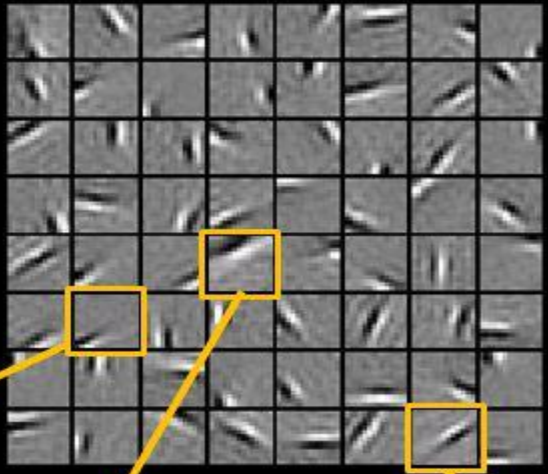
features

other stuff

Natural Images



Learned bases ($\phi_1, \ldots, \phi_{64}$): "Edges"



Test example

$$x \approx 0.8 * \phi_{36} + 0.3 * \phi_{42} + 0.5 * \phi_{63}$$

[0, 0, …, 0, **0.8**, 0, …, 0, **0.3**, 0, …, 0, **0.5**, …]

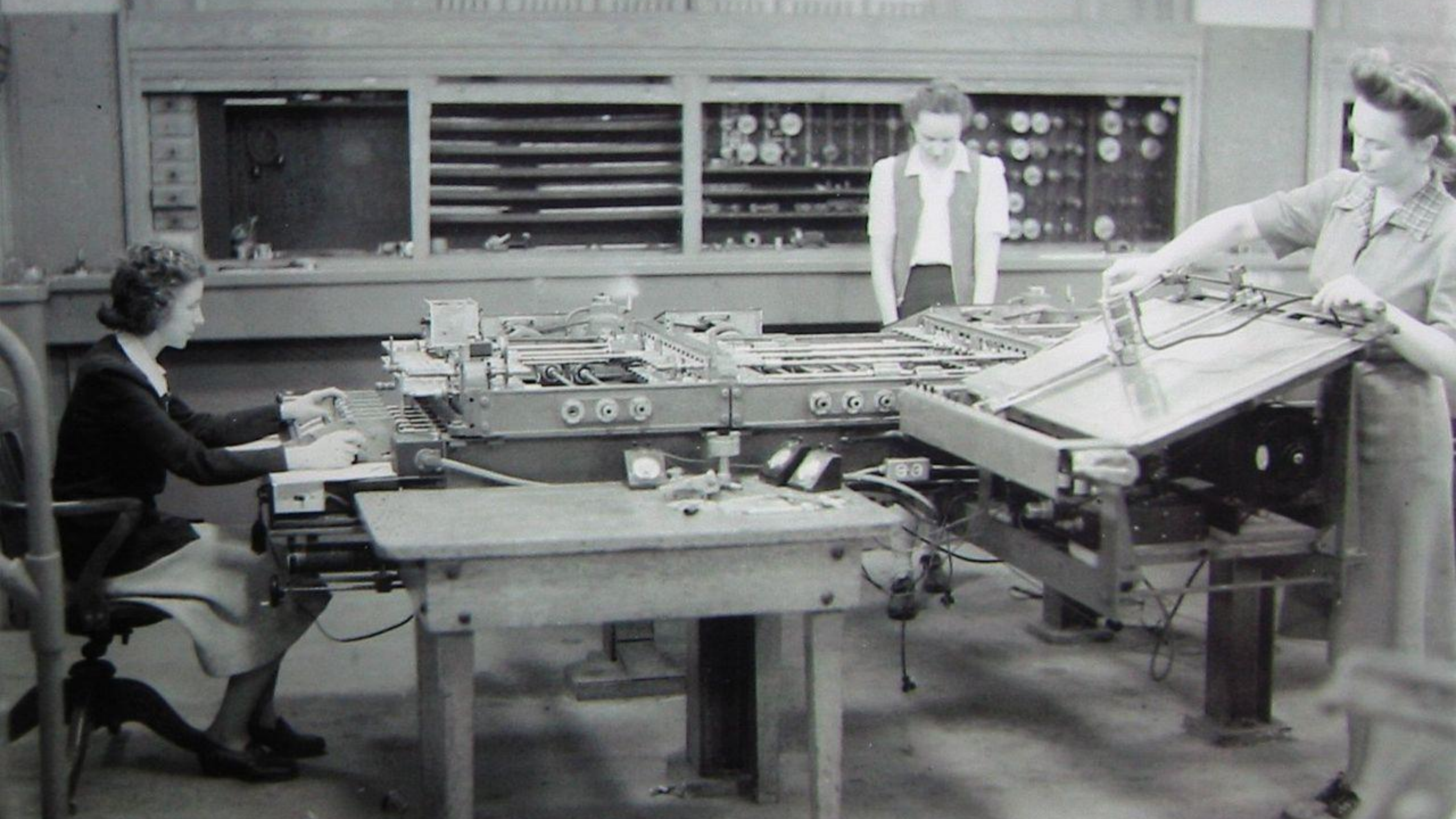
= [a_1, …, a_{64}] (feature representation)

Compact & easily interpretable

Dynamical Framework for Sparse Recovery

$$|| \mathbf{Wz} - \mathbf{x} ||_2 + \lambda || \mathbf{z} ||_0$$

"Captures a good chunk of the **computer vision** and **theoretical neuroscience** being done in the last decade" - Garrett Kenyon



There are simple systems of nonlinear differential equations that settle to the solution of

$$\min_x \ \lambda \|x\|_1 + \frac{1}{2}\|\Phi x - y\|_2^2$$

or more generally

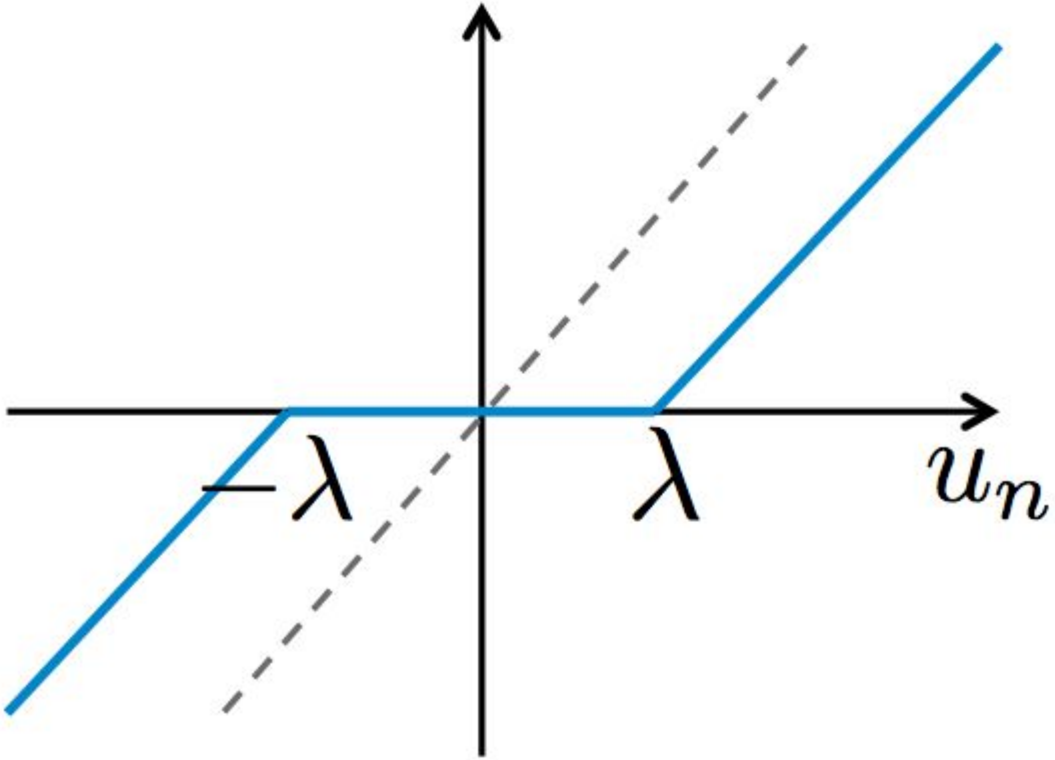
$$\min_x \ \lambda \sum_{n=1}^{N} C(x[n]) + \frac{1}{2}\|\Phi x - y\|_2^2$$

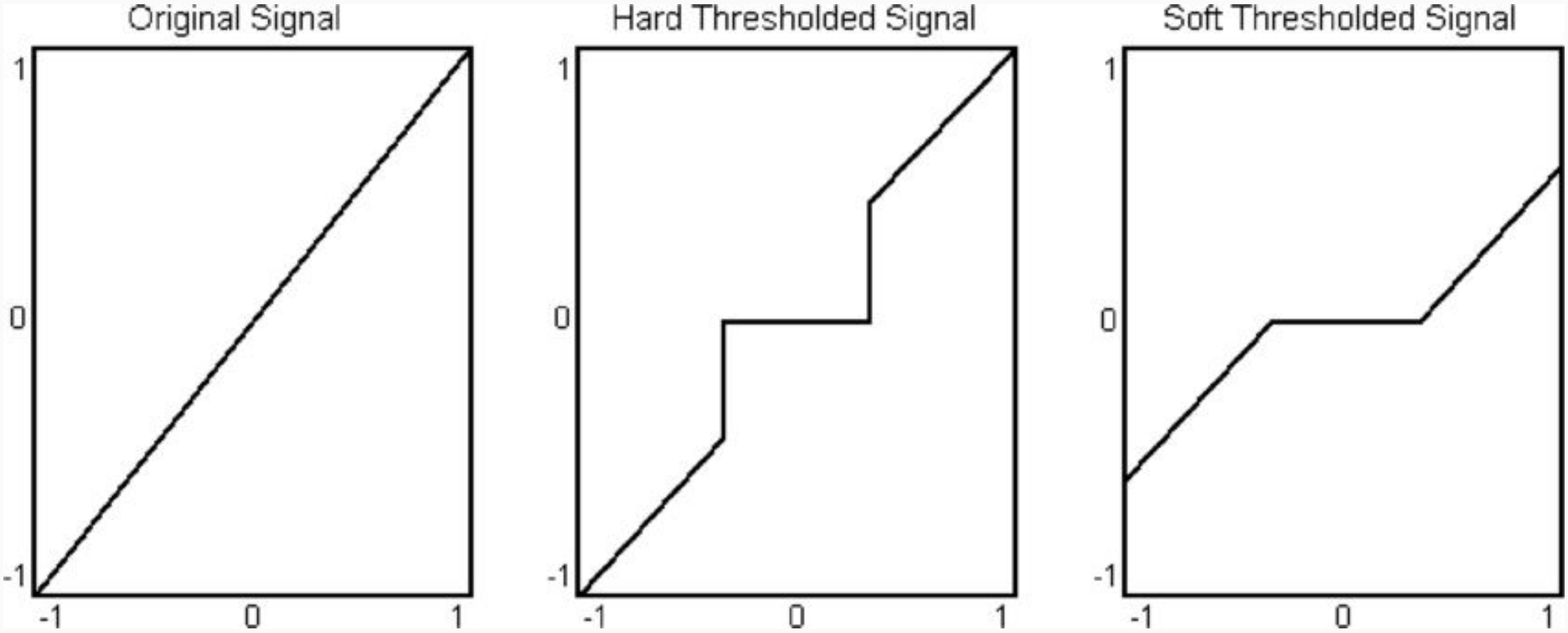
The Locally Competitive Algorithm (LCA):

$$\begin{aligned} \tau \dot{u}(t) &= -u(t) - (\Phi^T \Phi - I)x(t) + \Phi^T y \\ x(t) &= T_\lambda(u(t)) \end{aligned}$$

is a neurologically-inspired (Rozell et al 08) system which settles to the solutions of the above

$$x_n = T_\lambda(u_n)$$

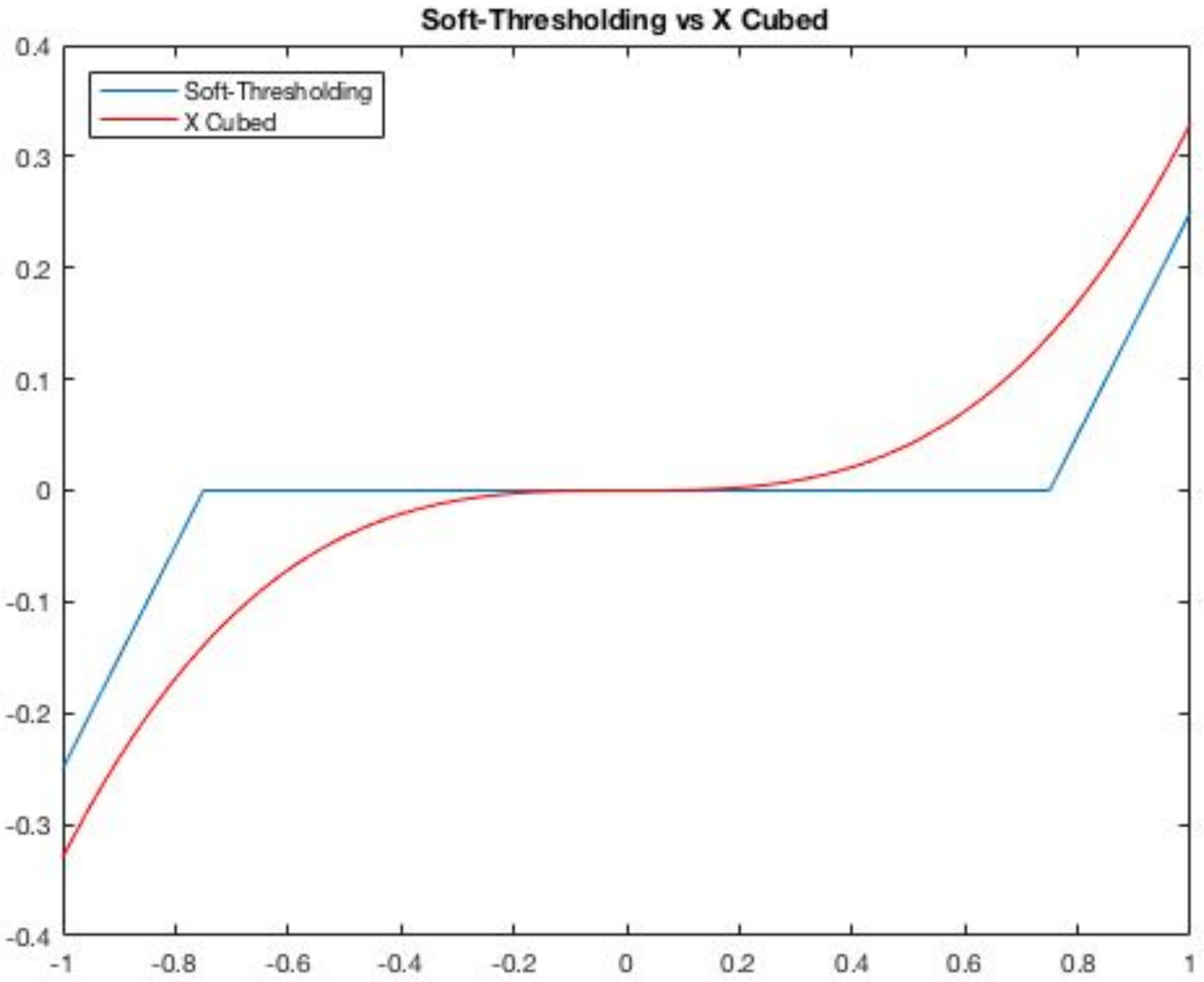




Original Signal
Hard Thresholded Signal
Soft Thresholded Signal
1
0
-1
-1
0
1

“A **third**-order power law best describes the transformation of *membrane potential to firing rate*”

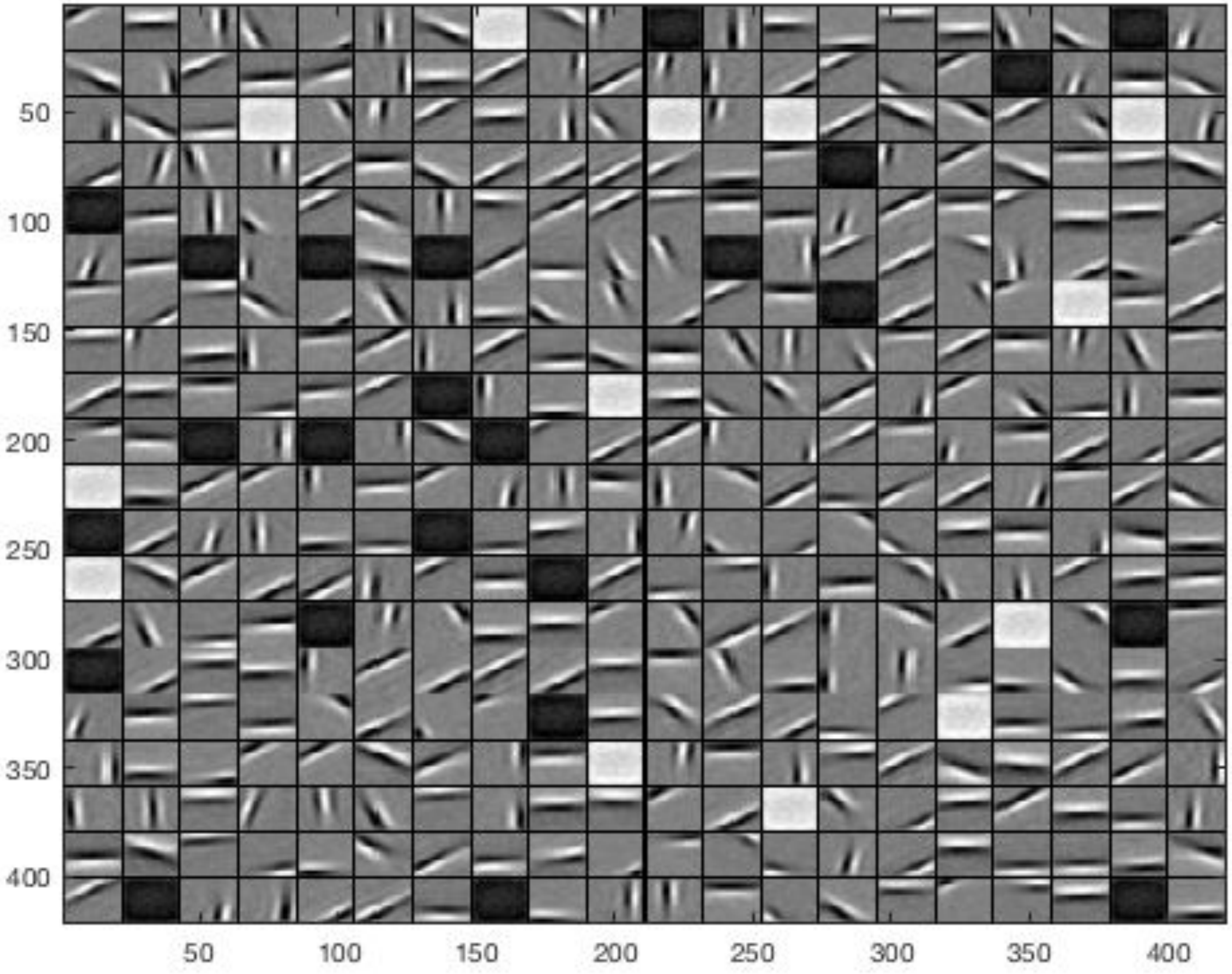
\- Christoph Koch


Soft-Thresholding vs X Cubed
Soft-Thresholding
X Cubed
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

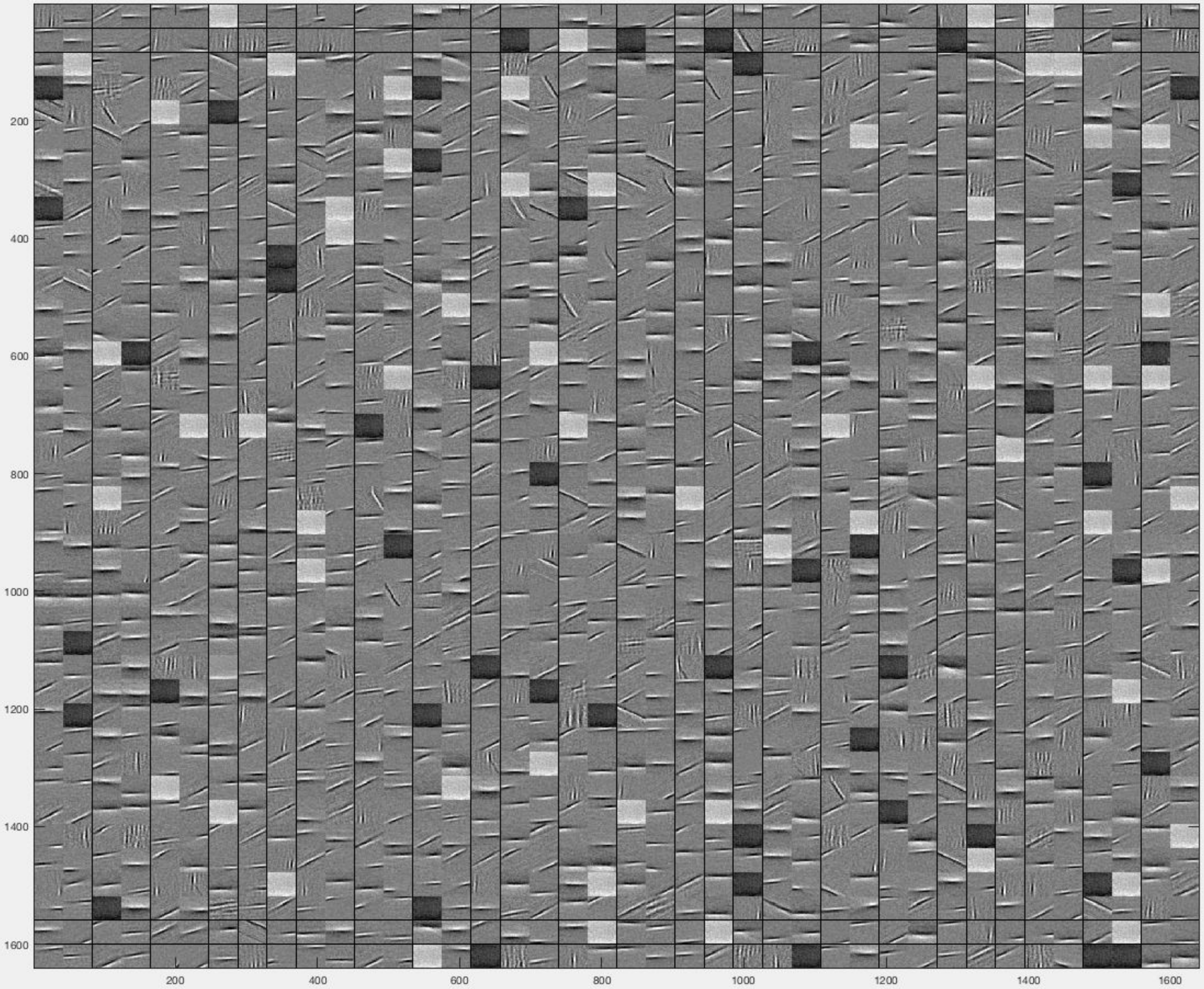
Hebbian Rule for Dictionary Learning

$$\Delta W = (x - Wz)z'$$

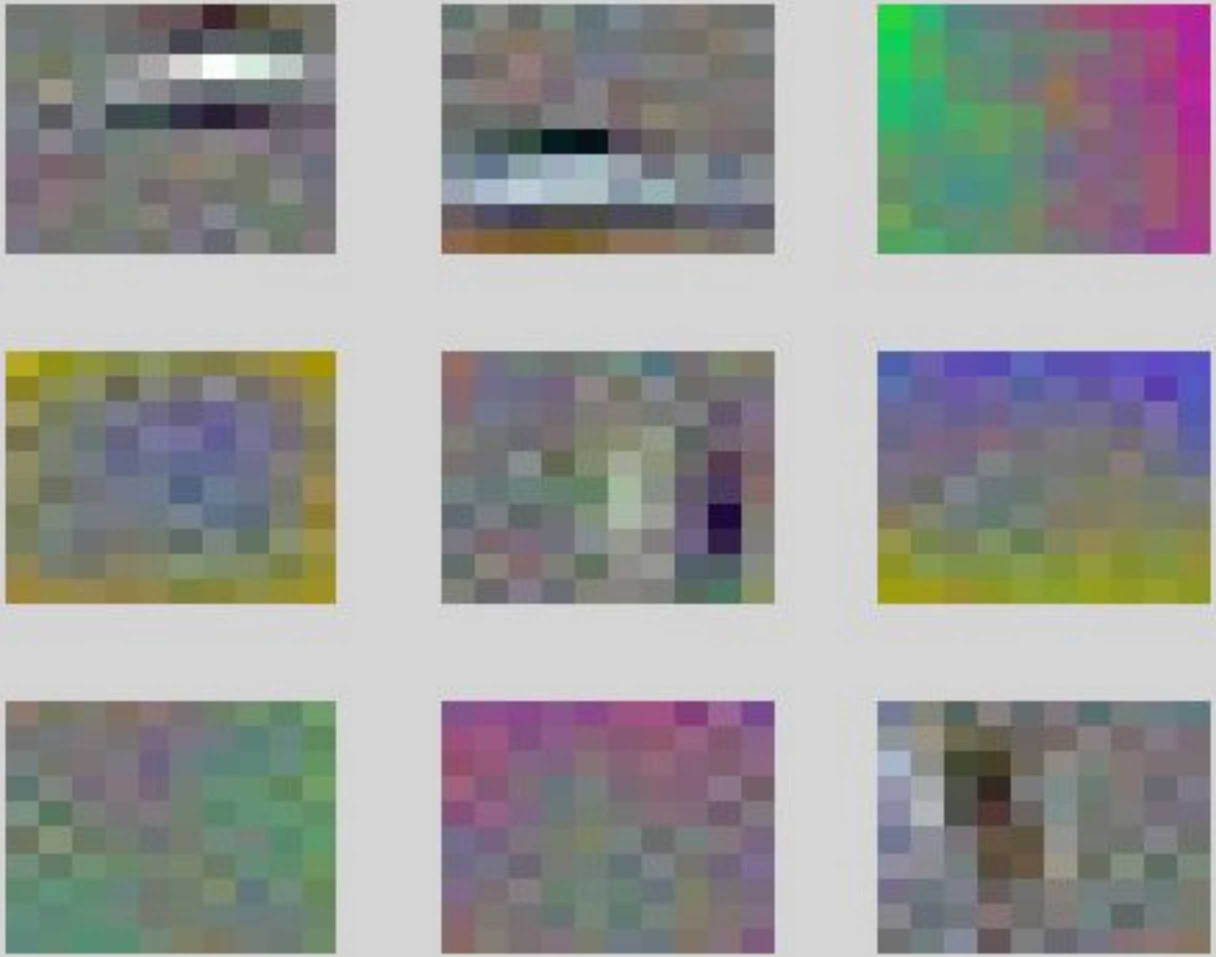
```
W = NC(W); %Normalize Columns

a = W'*X; %Feature Activations

a = NC(a); %Normalize Columns

a=0.5*a.^3;   %Cubic function acts as threshold

W = W + ((X-W*a)*a'); %Update Dictionary
```



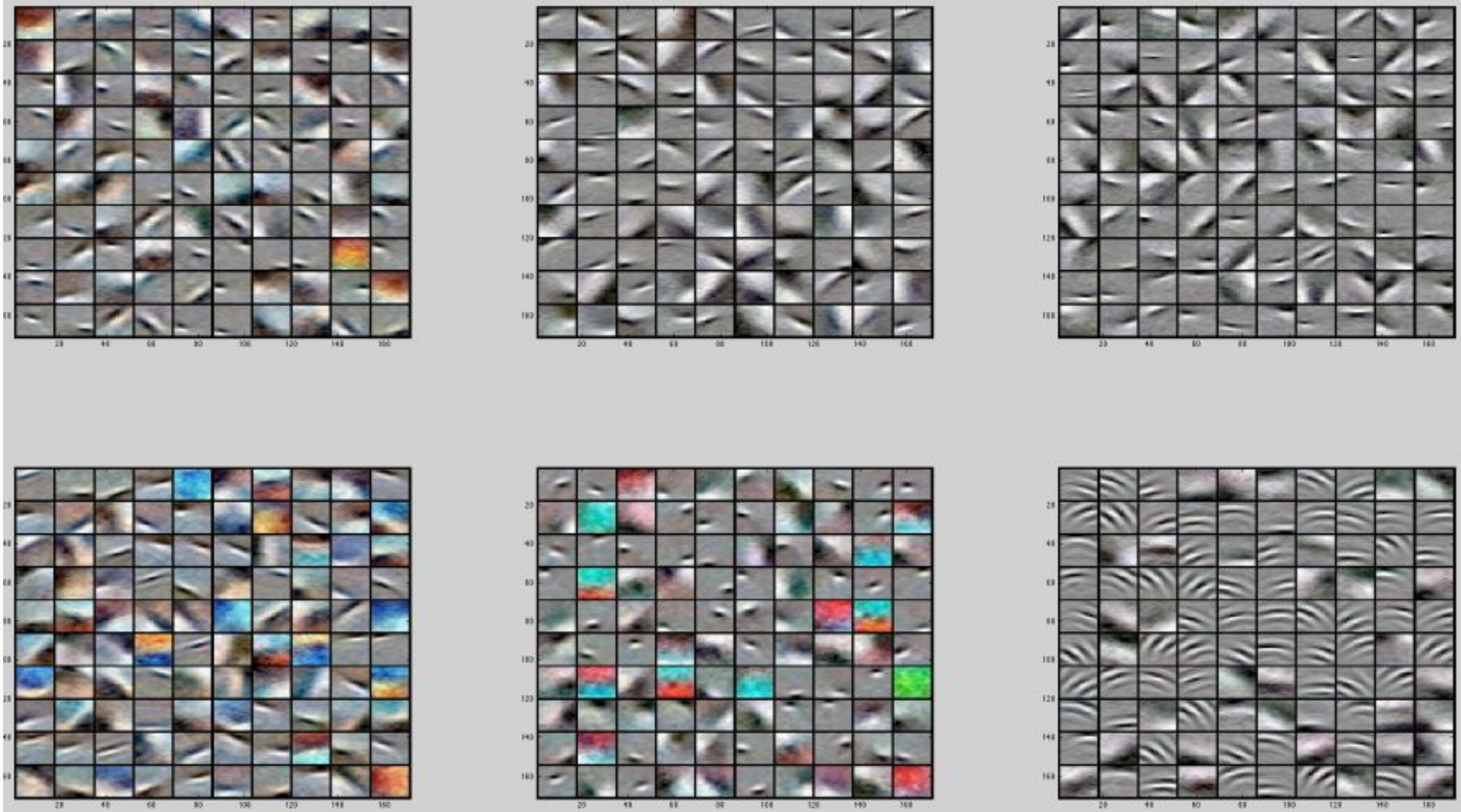
50
100
150
200
250
300
350
400
50
100
150
200
250
300
350
400

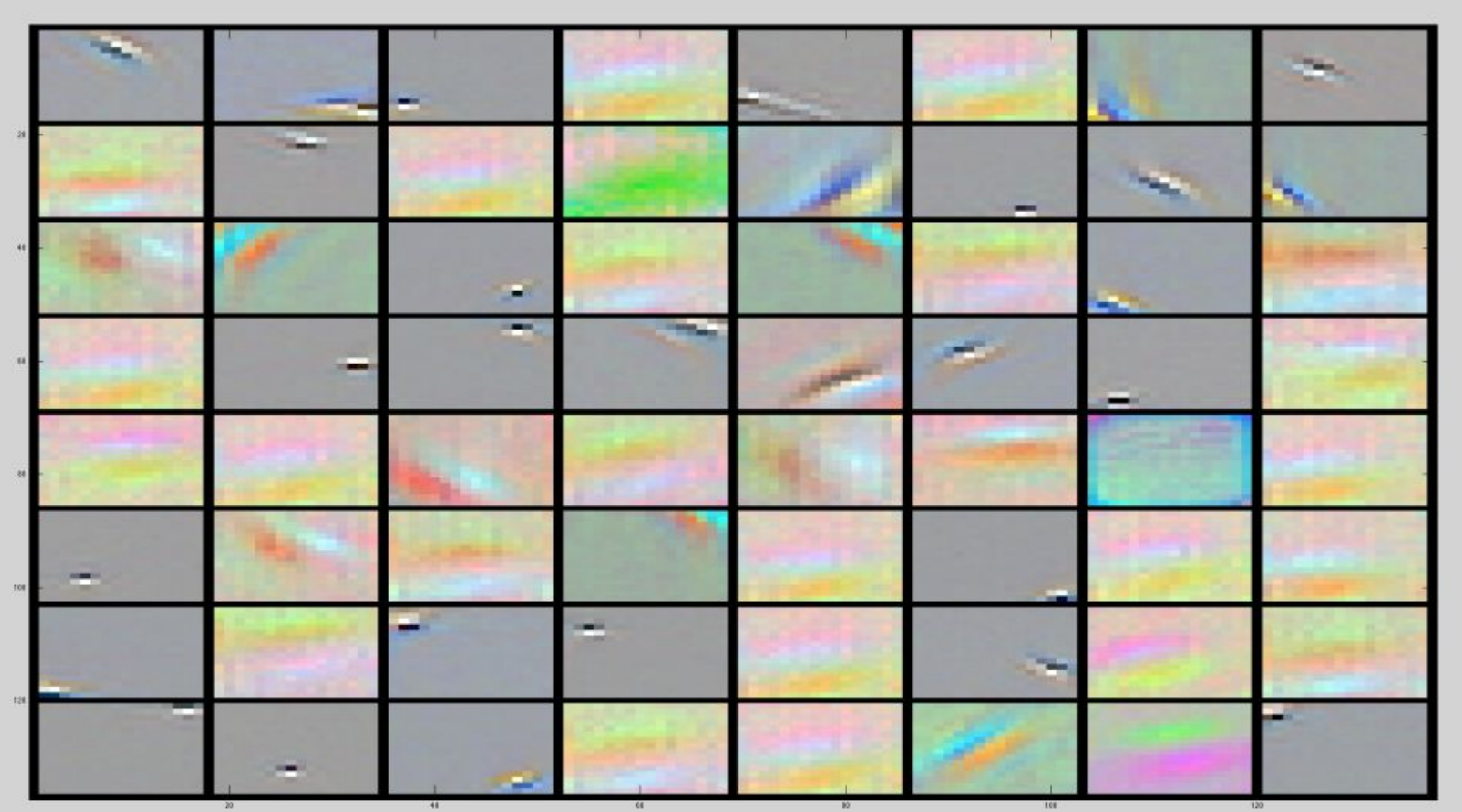


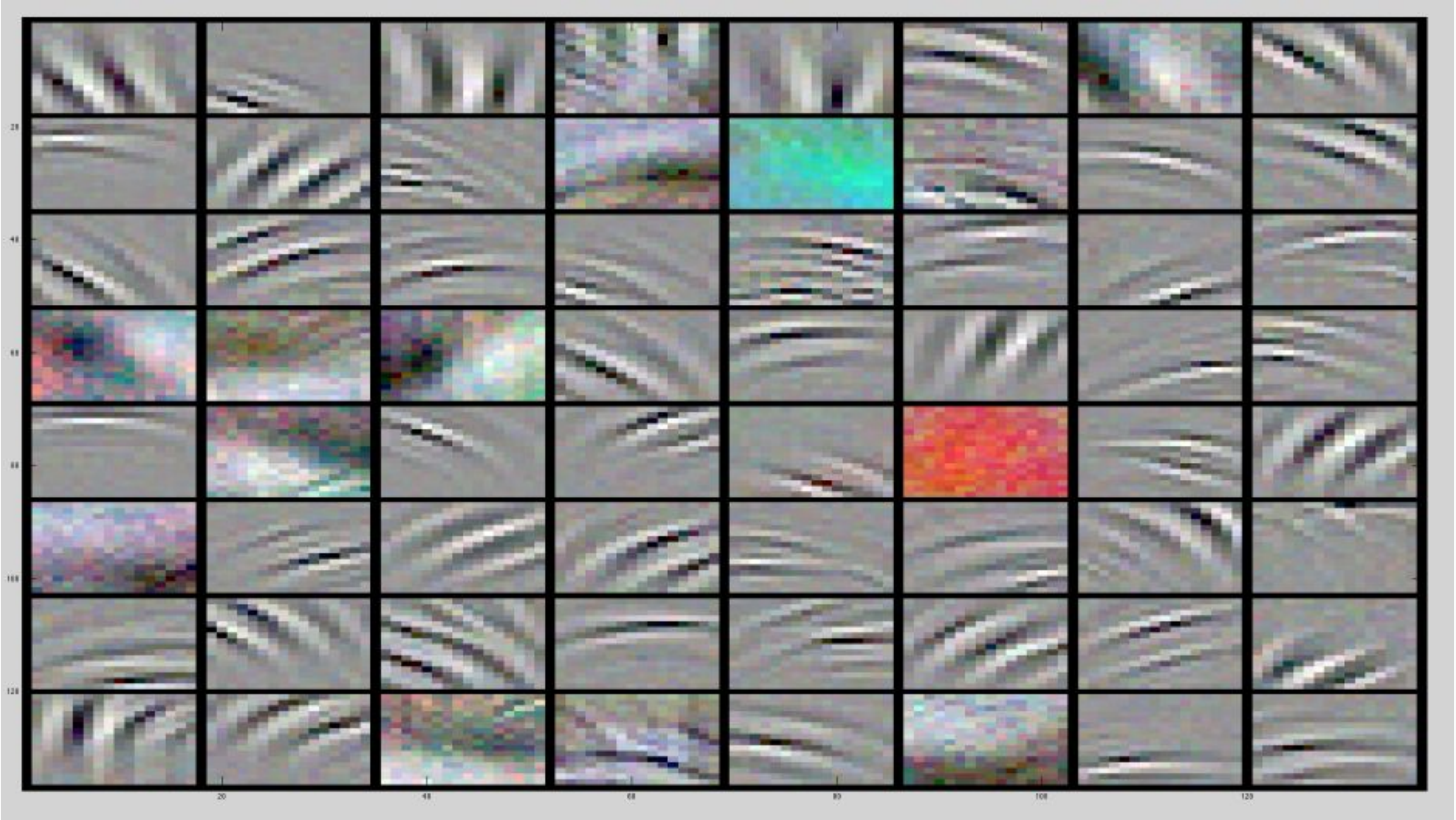
200
400
600
800
1000
1200
1400
1600
200
400
600
800
1000
1200
1400
1600



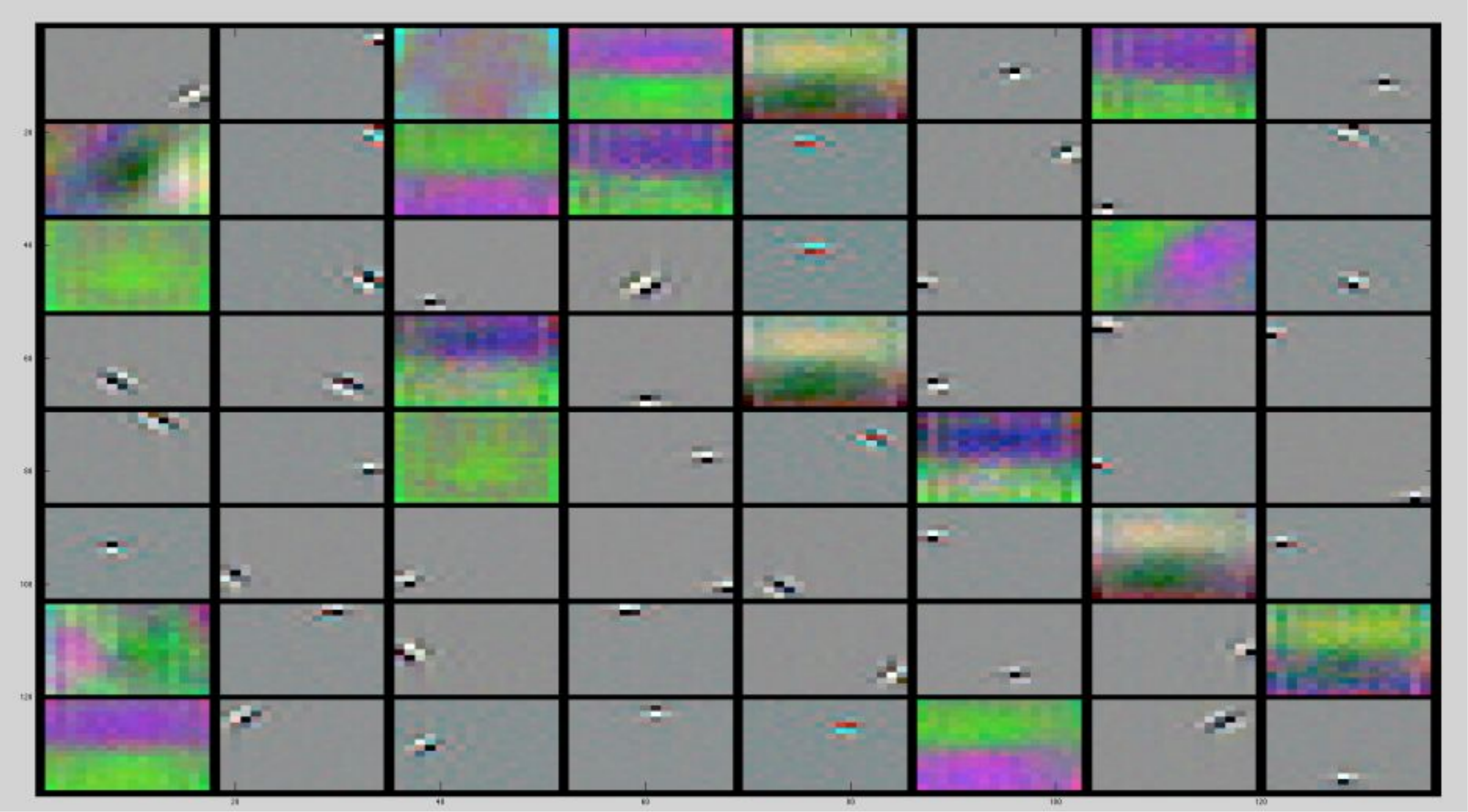

© Egil Michaelsen

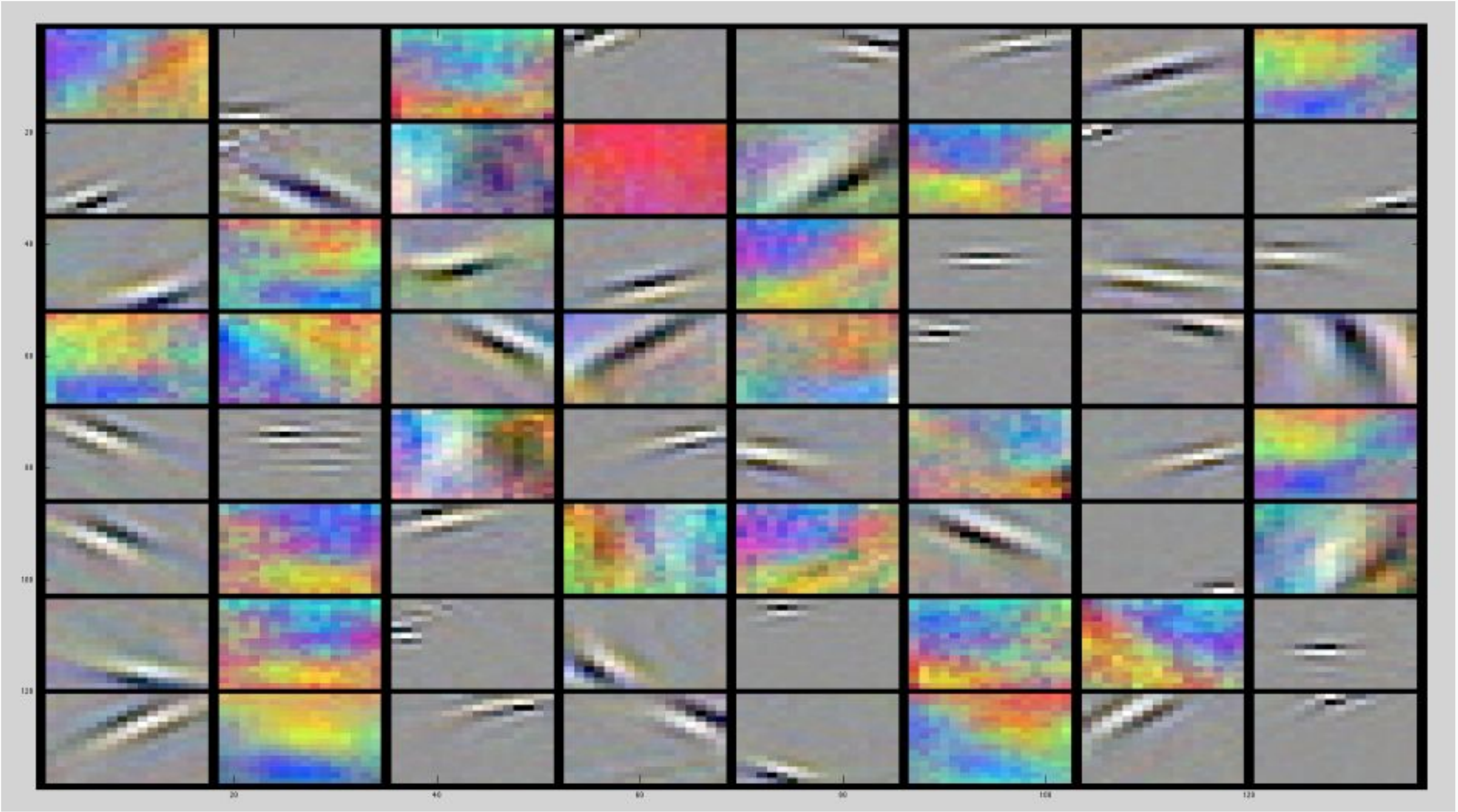



20
40
60
80
100
120

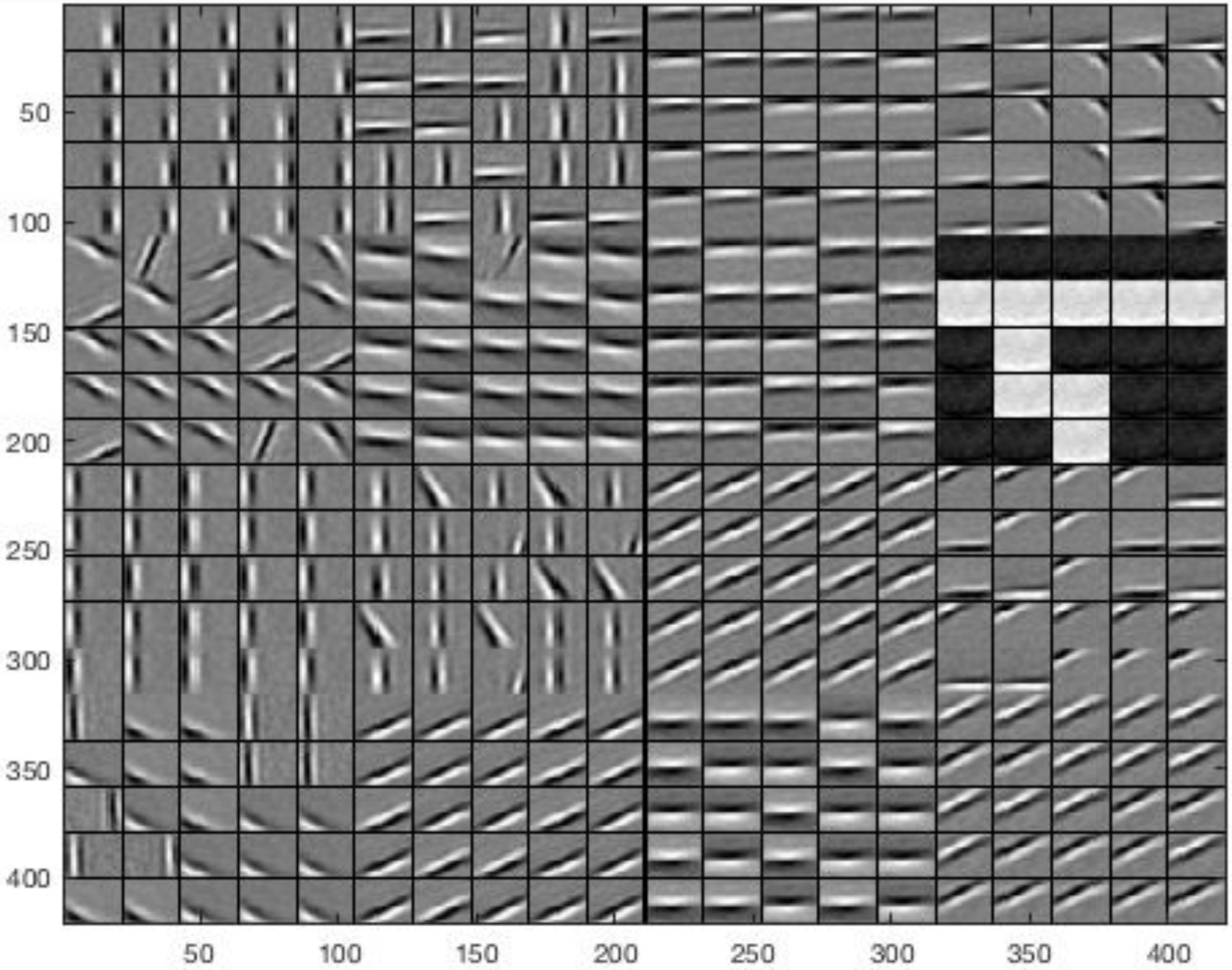


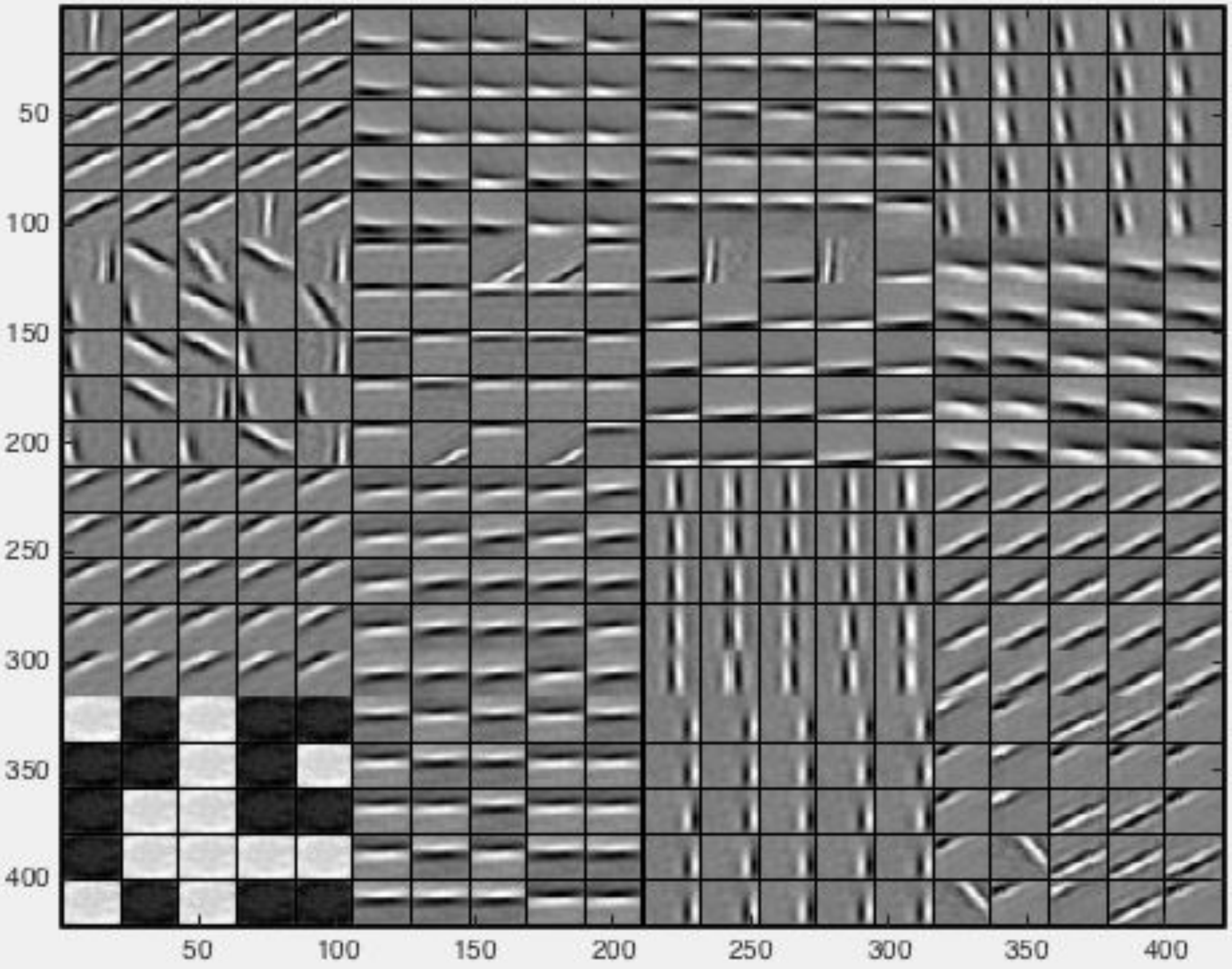
20
40
60
80
100
120
20
40
60
80
100
120


20
40
60
80
100
120

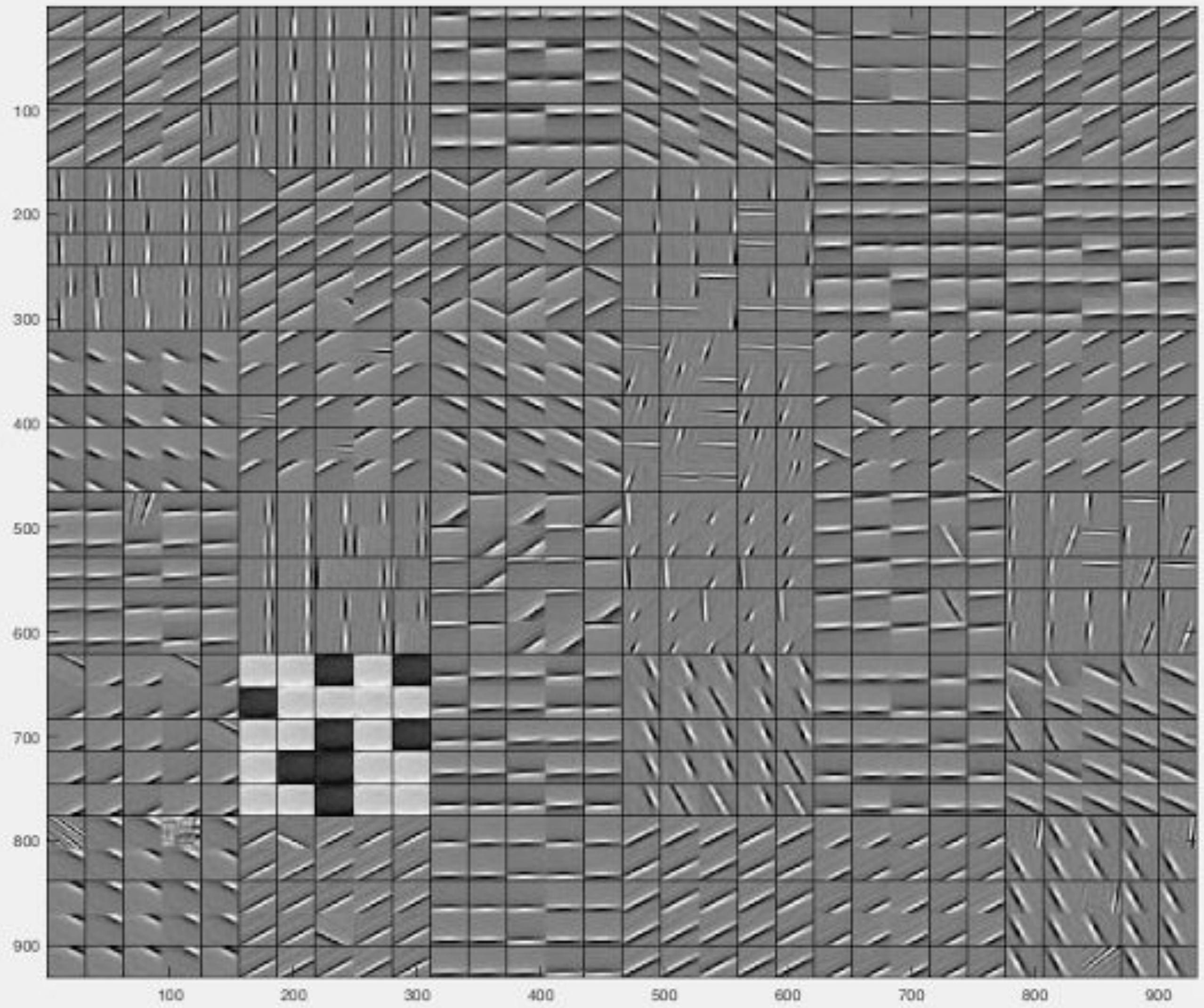


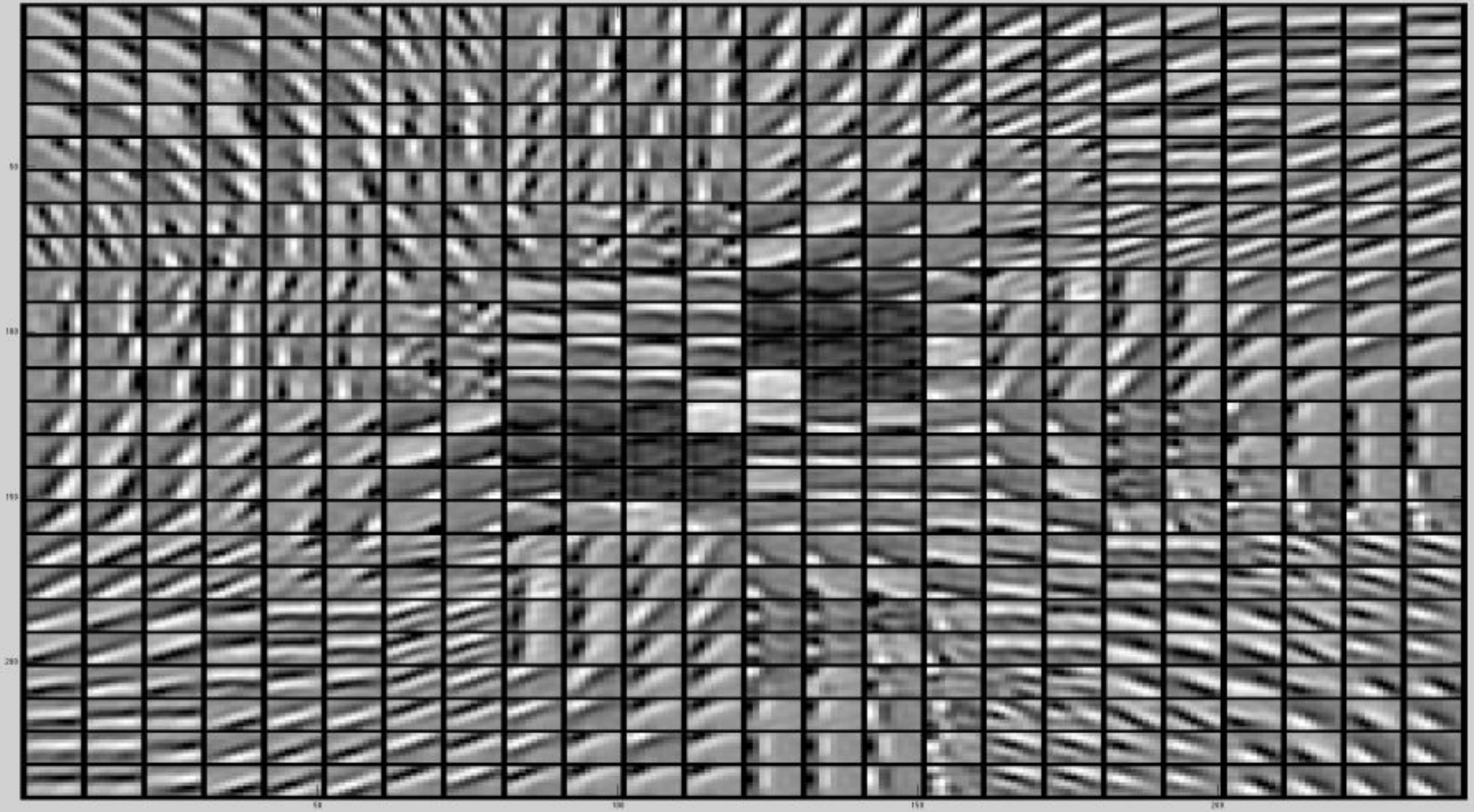
20
40
60
80
100
120
20
40
60
80
100
120

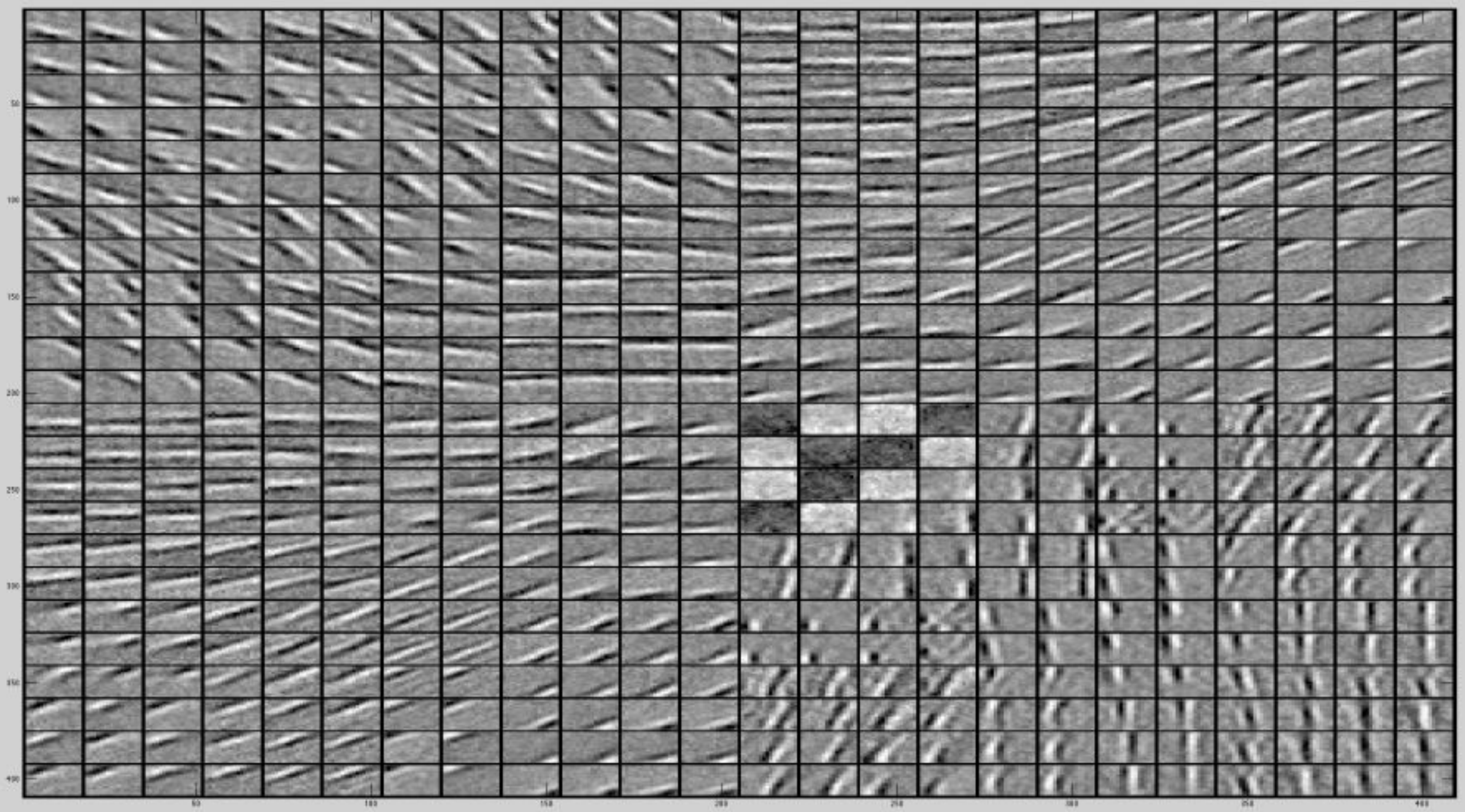

50
100
150
200
250
300
350
400
50
100
150
200
250
300
350
400

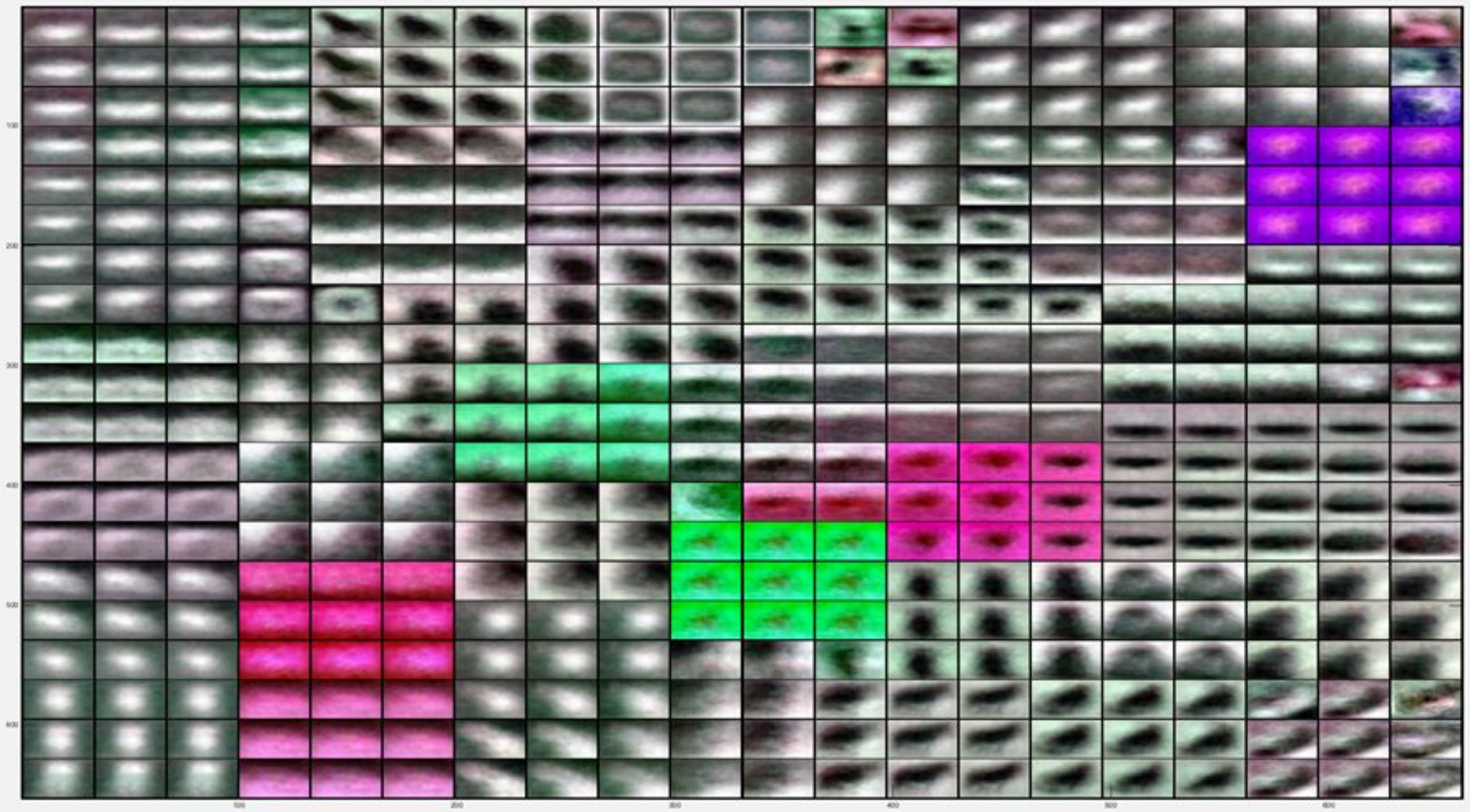


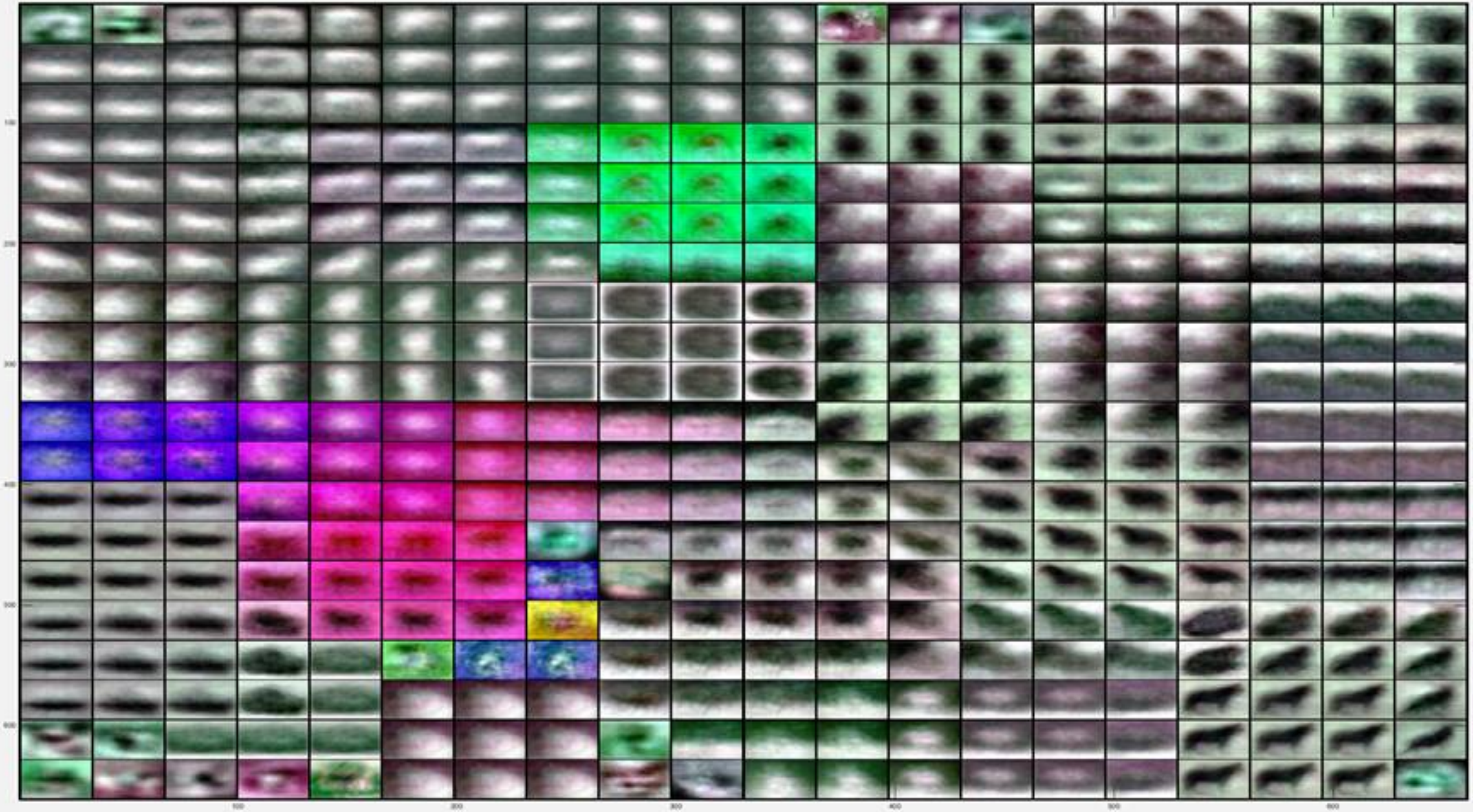
50
100
150
200
250
300
350
400
50
100
150
200
250
300
350
400

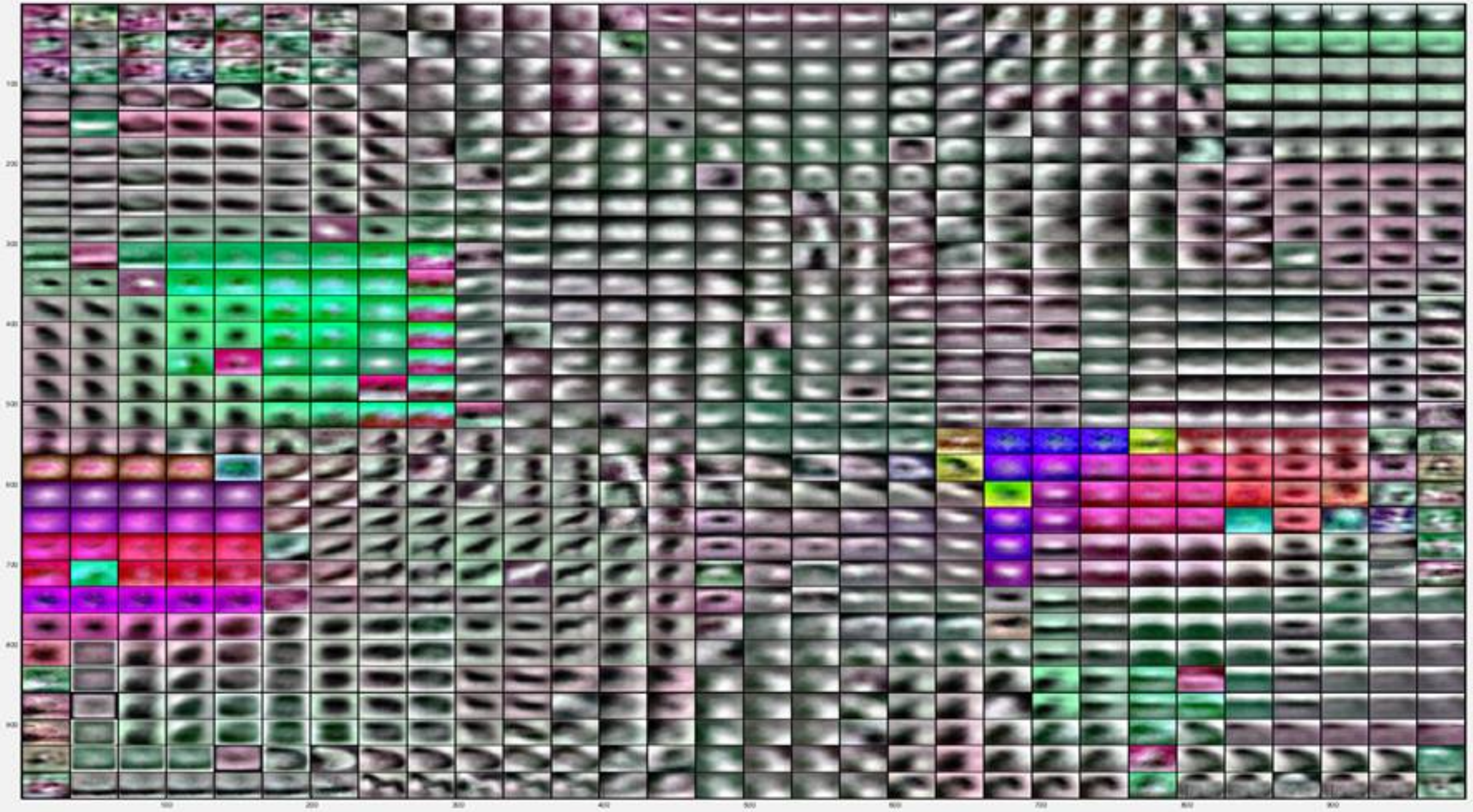

100
200
300
400
500
600
700
800
900
100
200
300
400
500
600
700
800
900

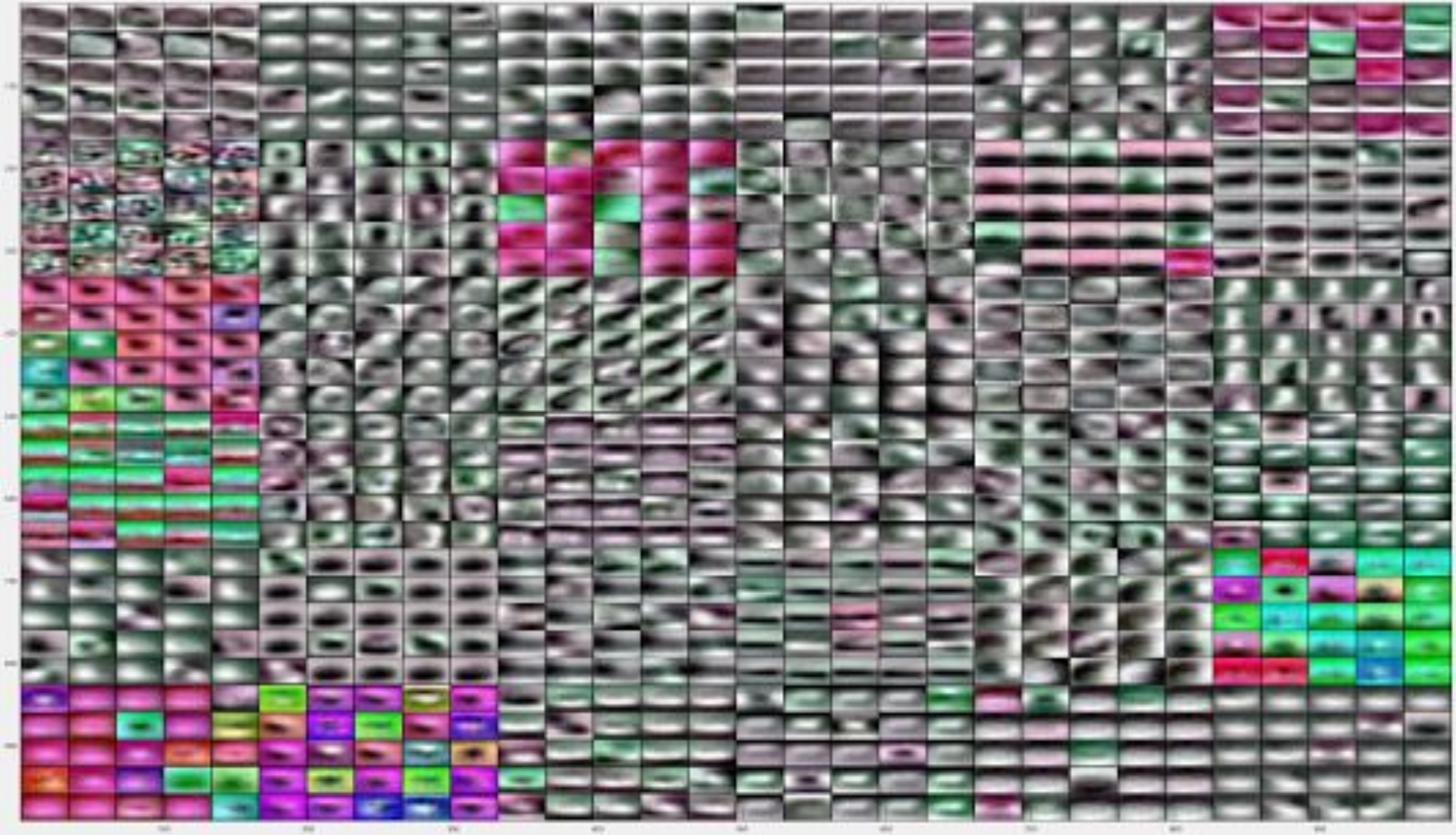


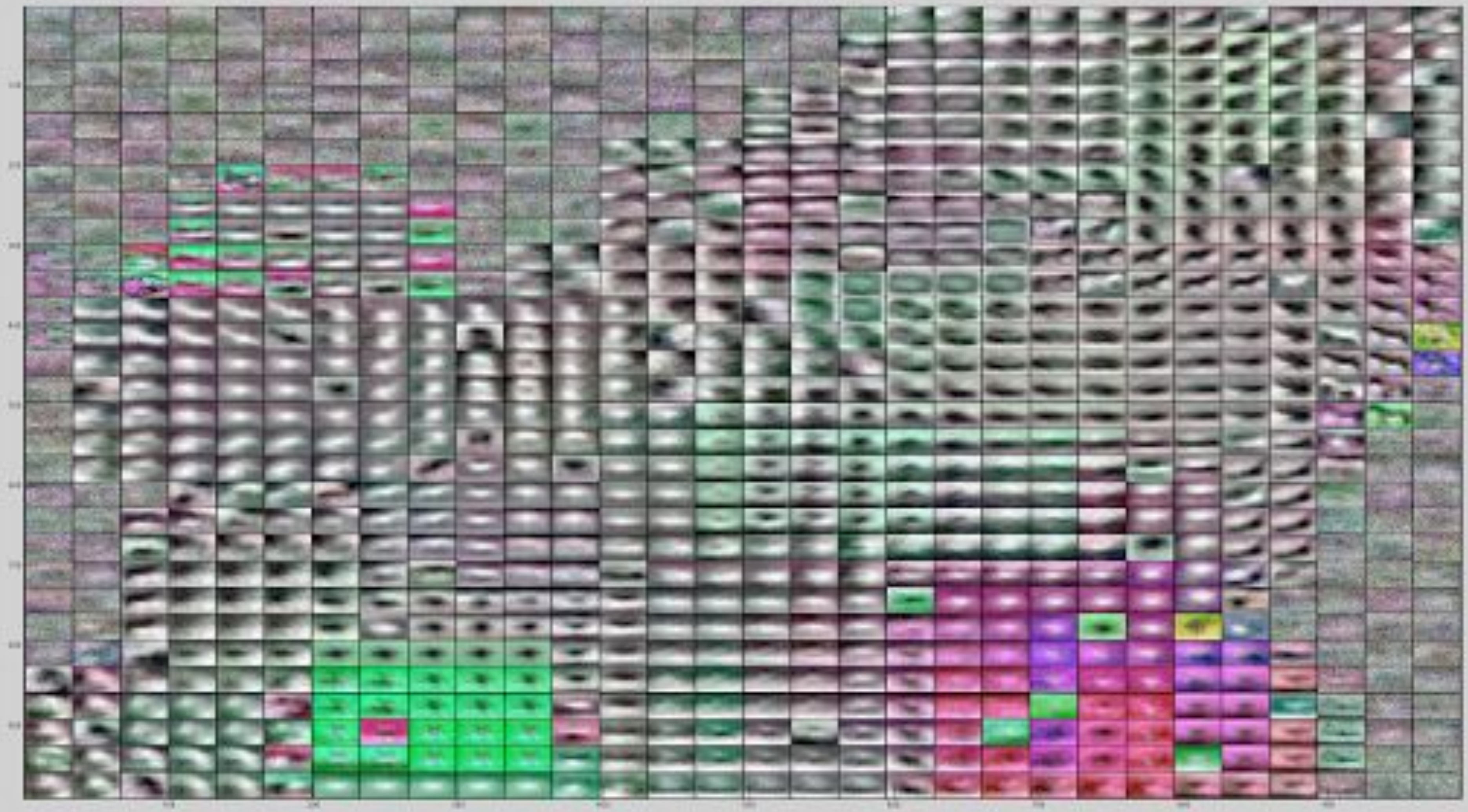

50
100
150
200
250
300
350
400

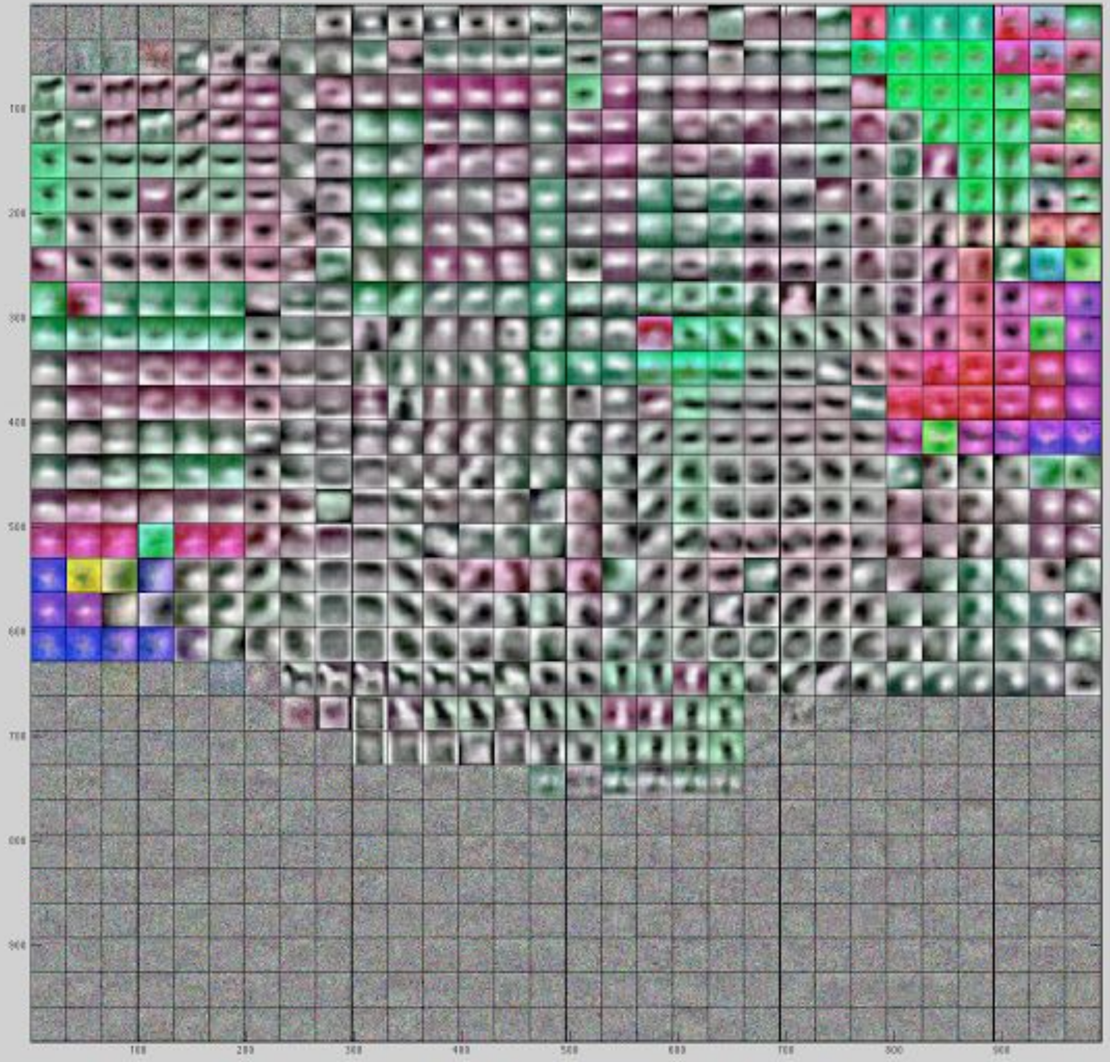


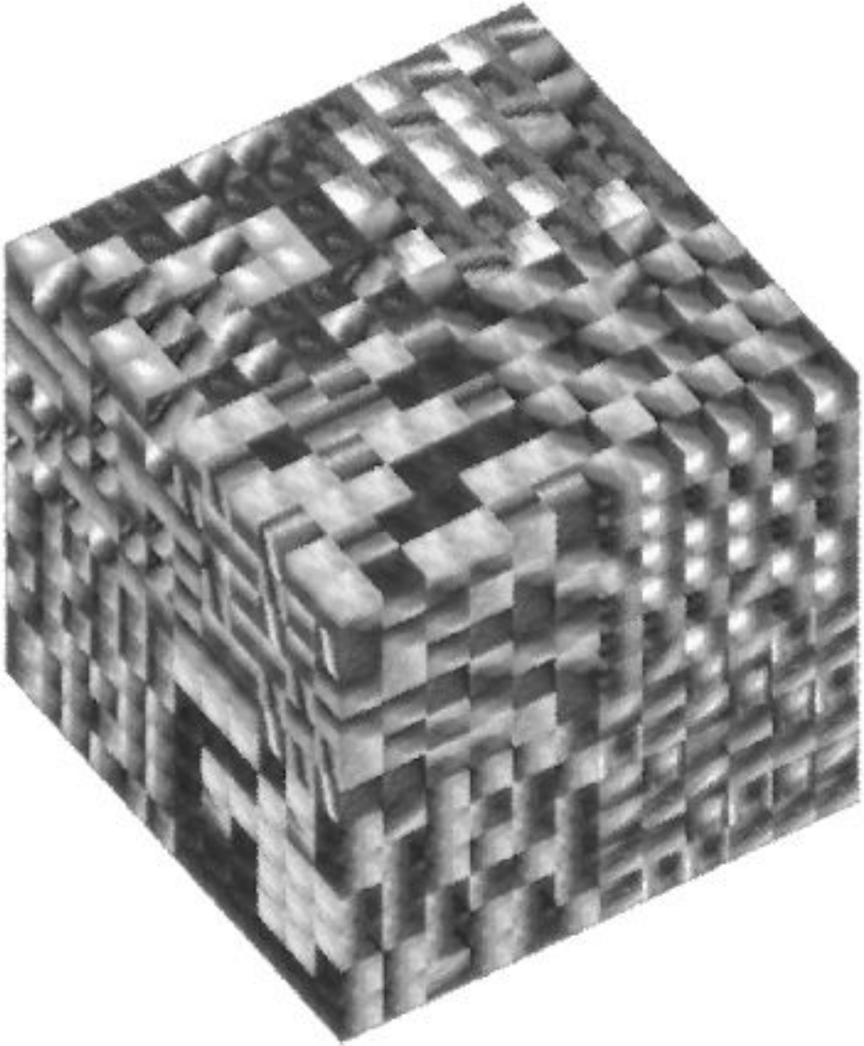


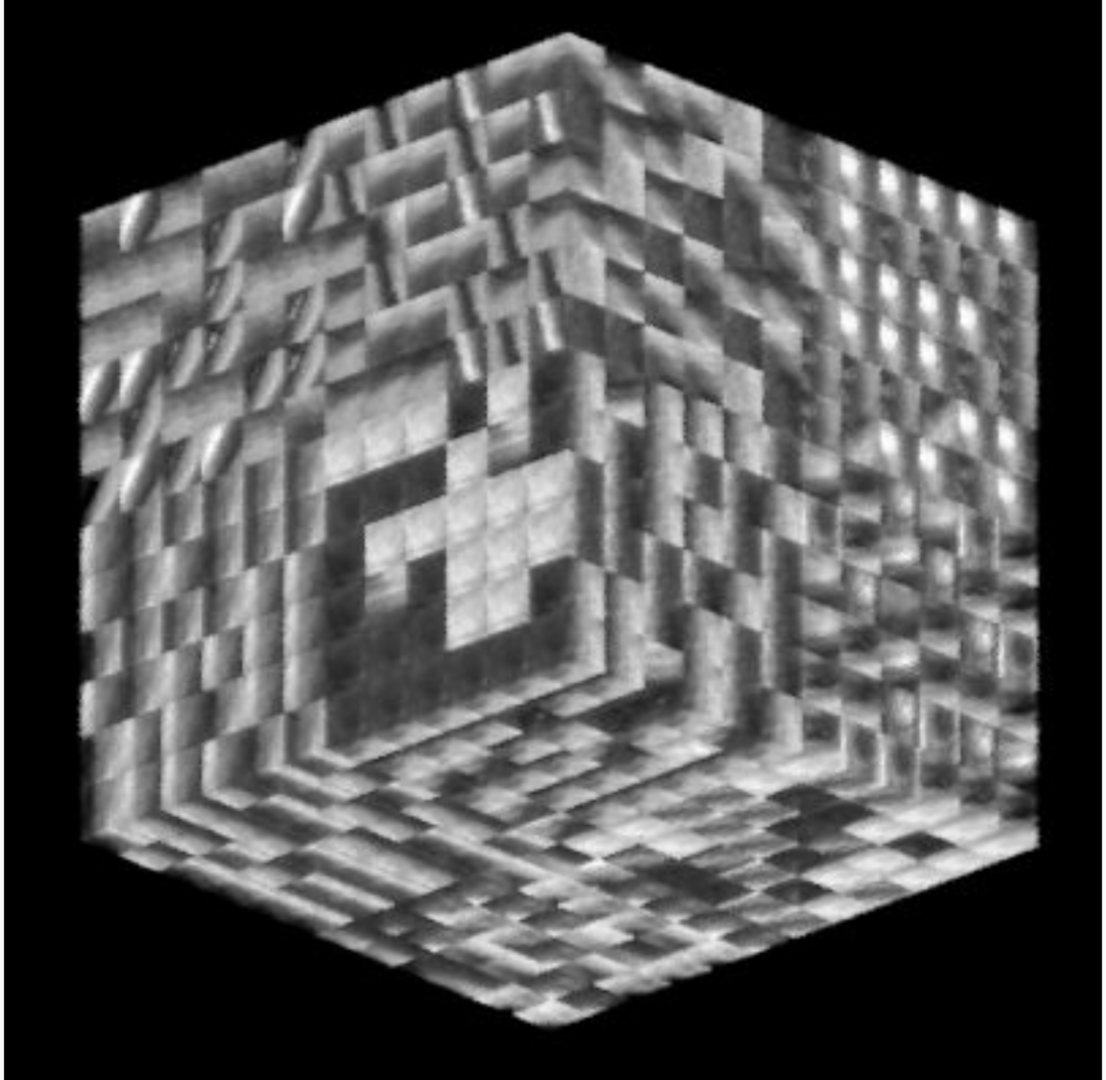












Binghamton R. R. Co.

TRANSFER.

Good only for this current trip, from point of transfer, on next car over the line punched after time canceled.

— 40 + TRADE-MARK, Subject to Rules of Company.

PICTURE PUNCHED INDICATES TYPE OF PASSENGER.

Issued by Conductor No. **47**

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31

MANUFACTURED BY GLOBE TICKET COMPANY PHILADELPHIA, PA.

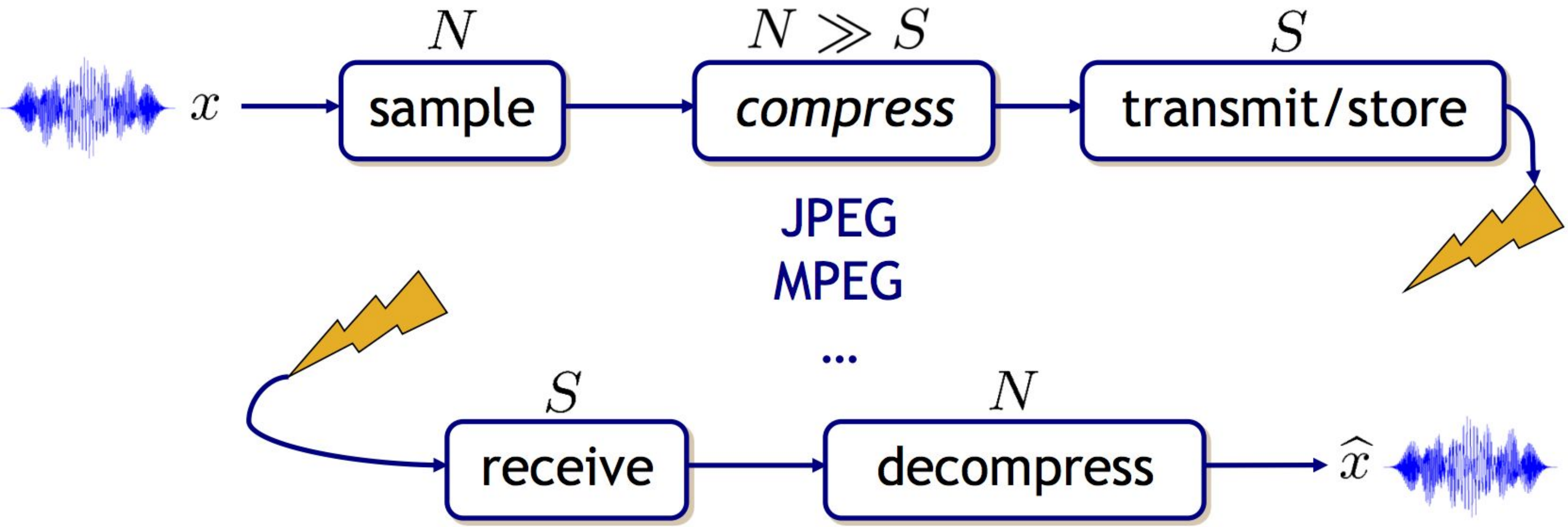
X	X	ROSS	VILLE	LESTER	SHIRE	ASY	LUM
O	HOME	GLEN	WOOD	P. DICK	INSON	DEP	OTS
C	C	WEST	MAIN	NO. CHE	NANGO	S. FO	REST
		BROAD	AVE	LEROY	STREET		

IF BLACK PORTION PUNCHED TIME IS P. M.

1	1	2	3	4	5
2	1	2	3	4	5
3	1	2	3	4	5
4	1	2	3	4	5
5	1	2	3	4	5
6	1	2	3	4	5
7	1	2	3	4	5
8	1	2	3	4	5
9	1	2	3	4	5
10	1	2	3	4	5
11	1	2	3	4	5
12	1	2	3	4	5

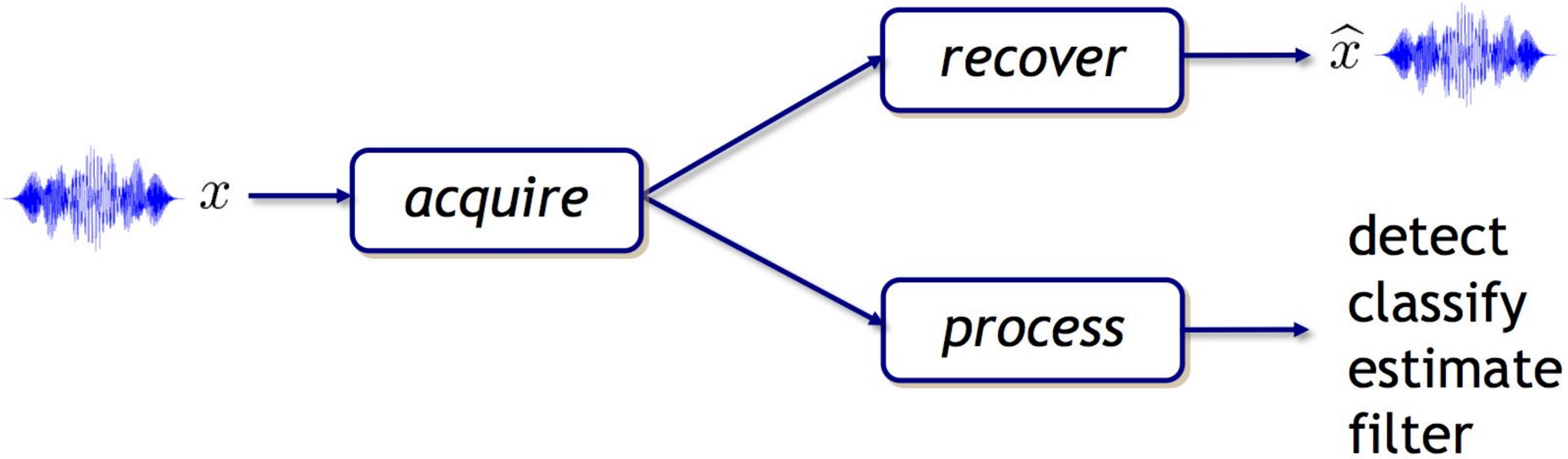
8MP
1.2MP front

16MP
2MP front



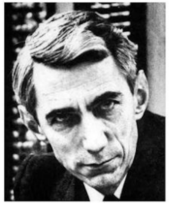
- Sample-and-compress paradigm is *wasteful*
  - samples cost $$$ and/or time

We would like to operate at the *intrinsic dimension* at all stages of the information-processing pipeline



Compressive Sensing (CS)

- Recall Shannon/Nyquist theorem
  - Shannon was a *pessimist*
  - 2x oversampling Nyquist rate is a worst-case bound for *any* bandlimited data
  - sparsity/compressibility irrelevant
  - Shannon sampling is a linear process while compression is a nonlinear process



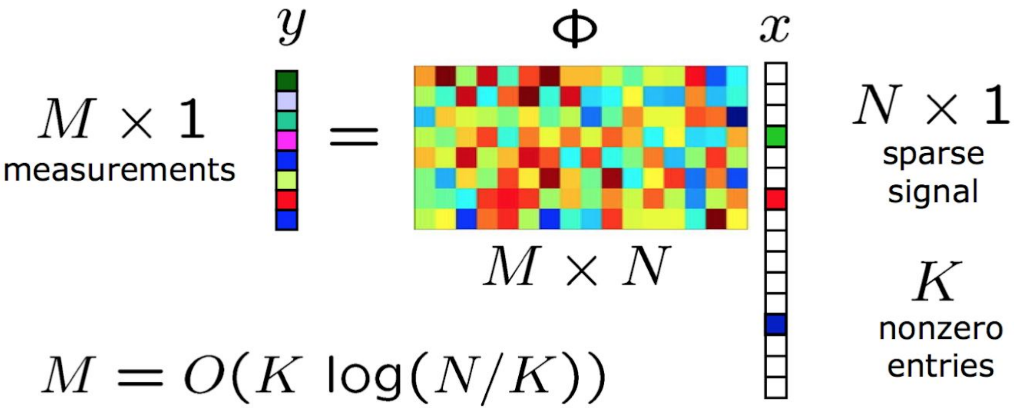
- **Compressive sensing**
  - new sampling theory that *leverages compressibility*
  - based on new *uncertainty principles*
  - *randomness* plays a key role

Compressive Data Acquisition

- When data is sparse/compressible, can directly acquire a ***condensed representation*** with no/little information loss

$$y = \Phi x$$

- ***Random projection*** will work



$$M = O(K \log(N/K))$$

[Candes-Romberg-Tao, Donoho, 2004]

Compressed Sensing

z is **Data**

x is **Code**

W is **Universal**

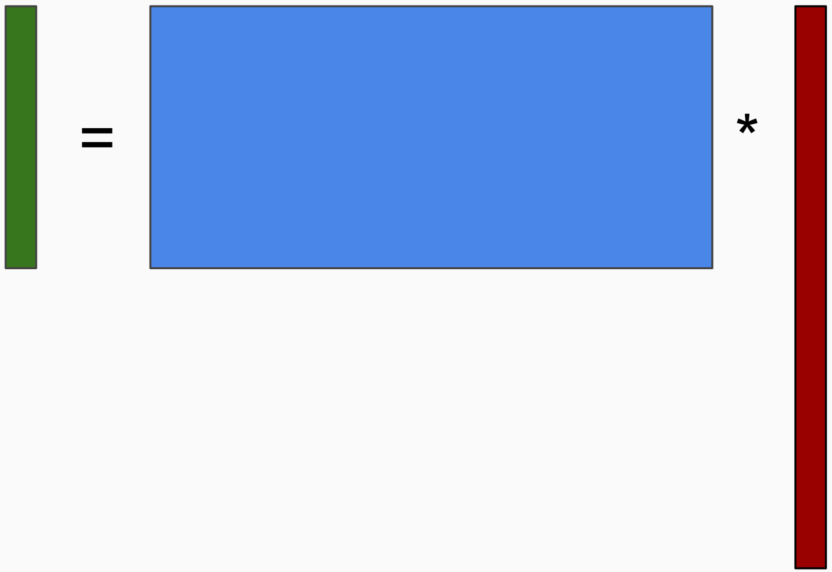
$$\| Wz - x \|_2 + \lambda \| z \|_0$$

Sparse Coding

z is **Code**

x is **Data**

W is **Adapted**



$$|| \mathbf{Wz} - \mathbf{x} ||_2 + \lambda || \mathbf{z} ||_0$$

Sparse Coding

z is **Code**

x is **Data**

W is **Adapted**

$$|| \mathbf{Wz} - \mathbf{x} ||_2 + \lambda || \mathbf{z} ||_0$$

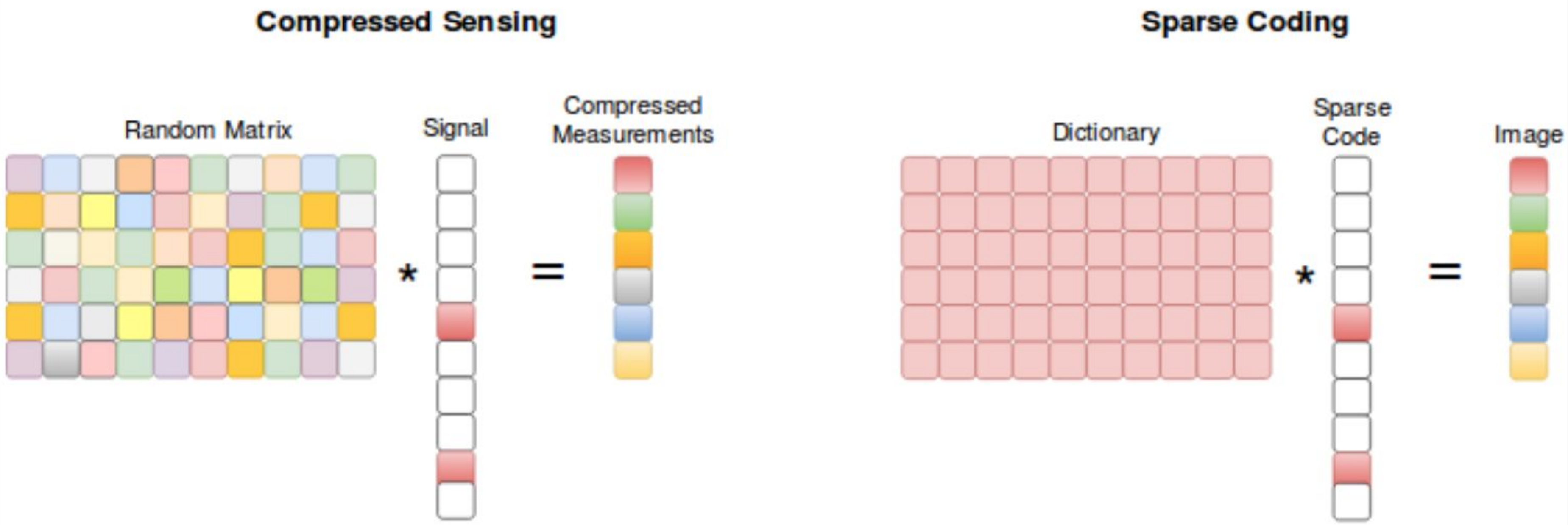
Compressed Sensing

z is **Data**

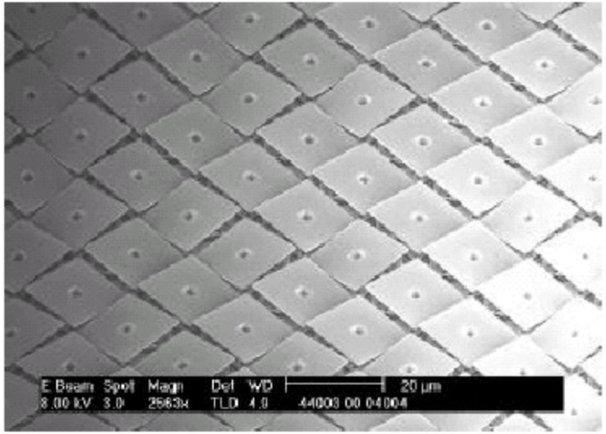
x is **Code**

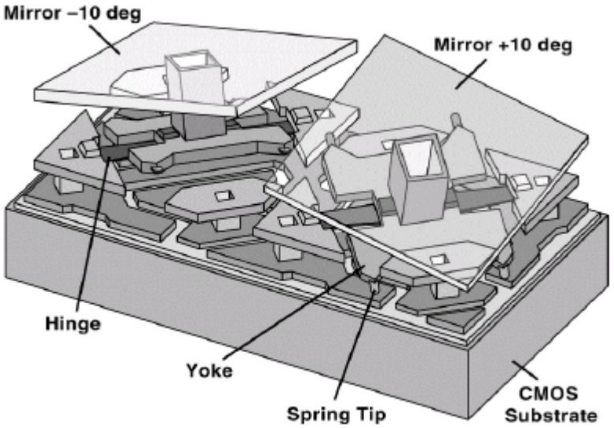
W is **Universal**

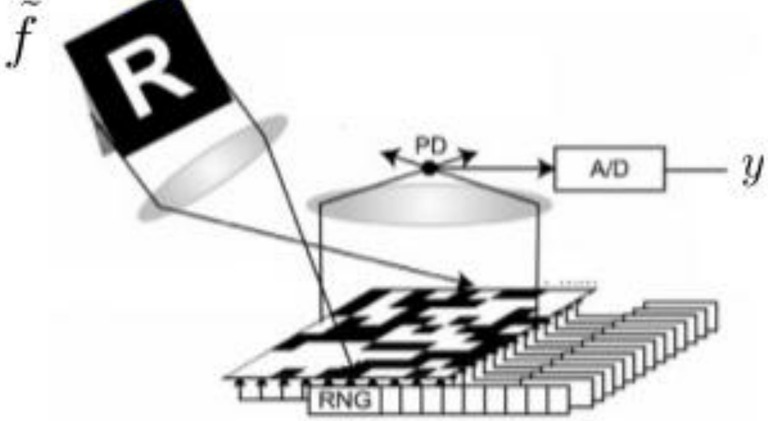
$$|| \mathbf{Wz} - \mathbf{x} ||_2 + \lambda || \mathbf{z} ||_0$$

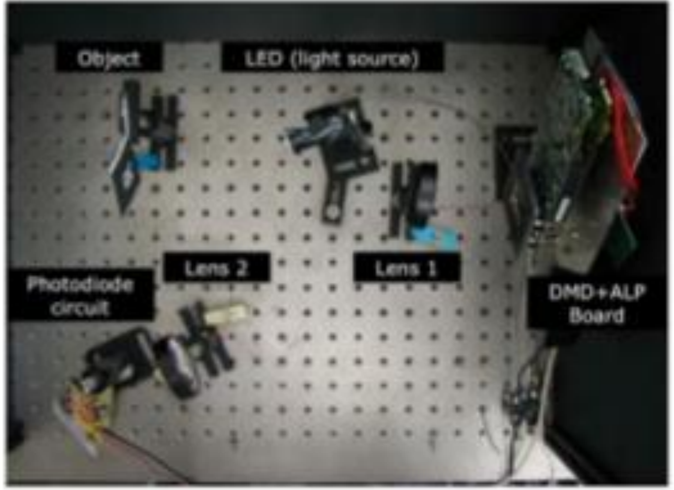

Compressed Sensing
Random Matrix
Signal
Compressed Measurements
*
=
Sparse Coding
Dictionary
Sparse Code
Image
*
=

TI Digital Micromirror Device

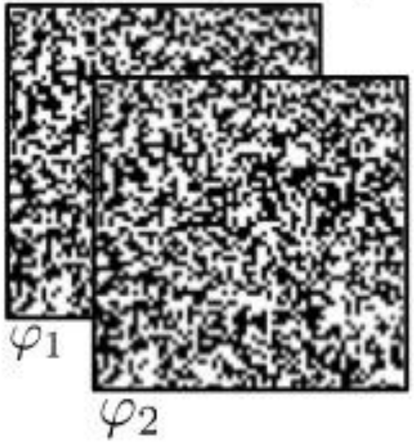






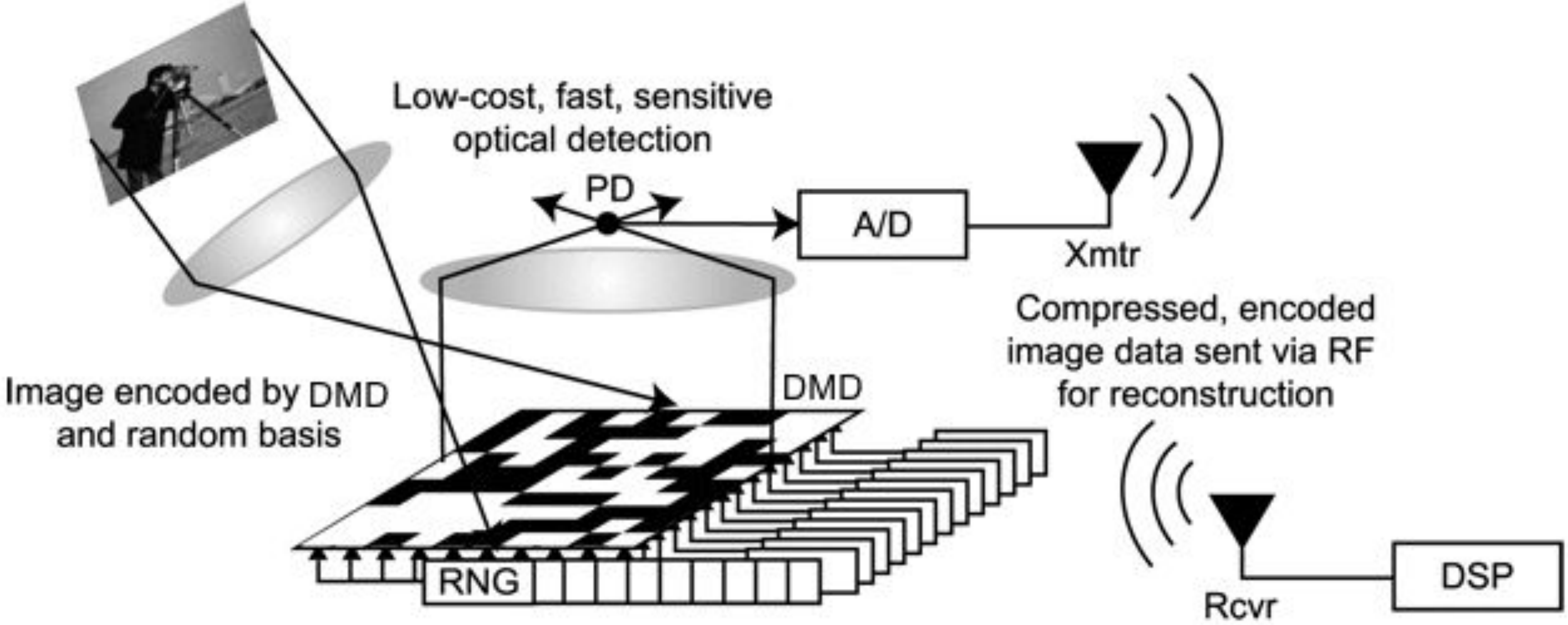


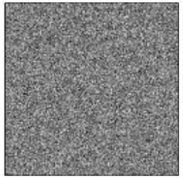
$$y[i] = \langle f, \varphi_i \rangle$$



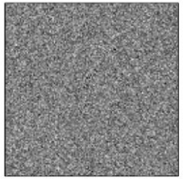
P measures $\ll N$ micro-mirrors

$P/N = 1$ $\quad$ $P/N = 0.16$ $\quad$ $P/N = 0.02$

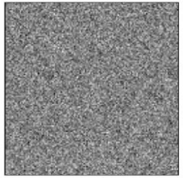

Low-cost, fast, sensitive
optical detection
PD
A/D
Xmtr
Compressed, encoded
image data sent via RF
for reconstruction
Image encoded by DMD
and random basis
DMD
RNG
Rcvr
DSP

$y_1 = \Big\langle$  ,  $\Big\rangle$

$y_2 = \Big\langle$  ,  $\Big\rangle$

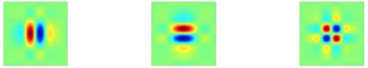
$y_3 = \Big\langle$  ,  $\Big\rangle$

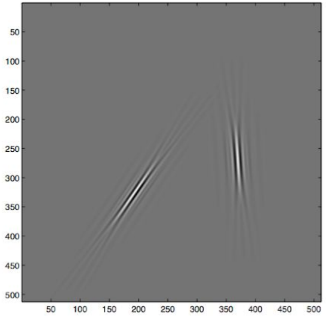
$\vdots$

$y_M = \Big\langle$  ,  $\Big\rangle$

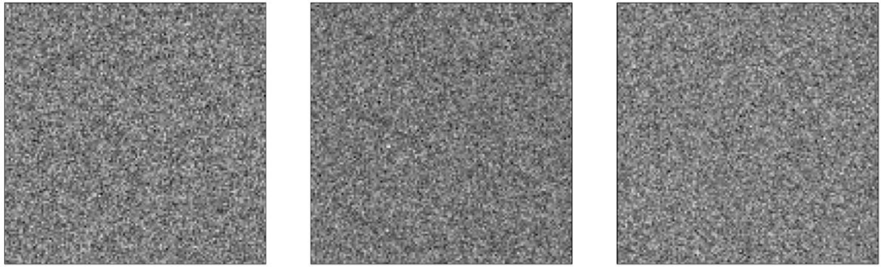
Representation vs. Measurements

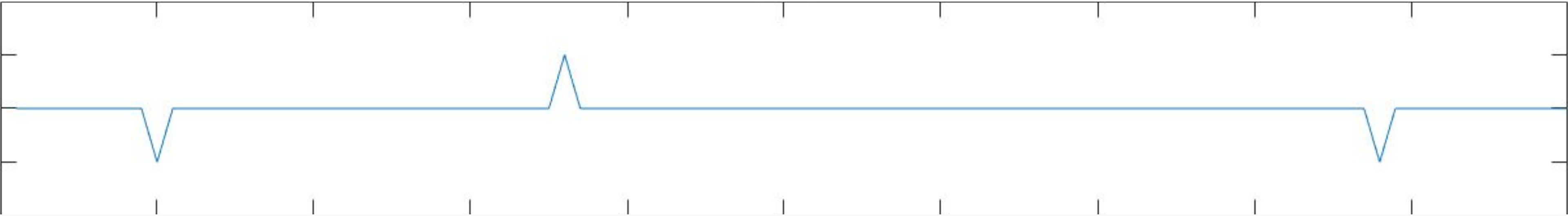
- Image structure: *local, coherent*
  Good basis functions:

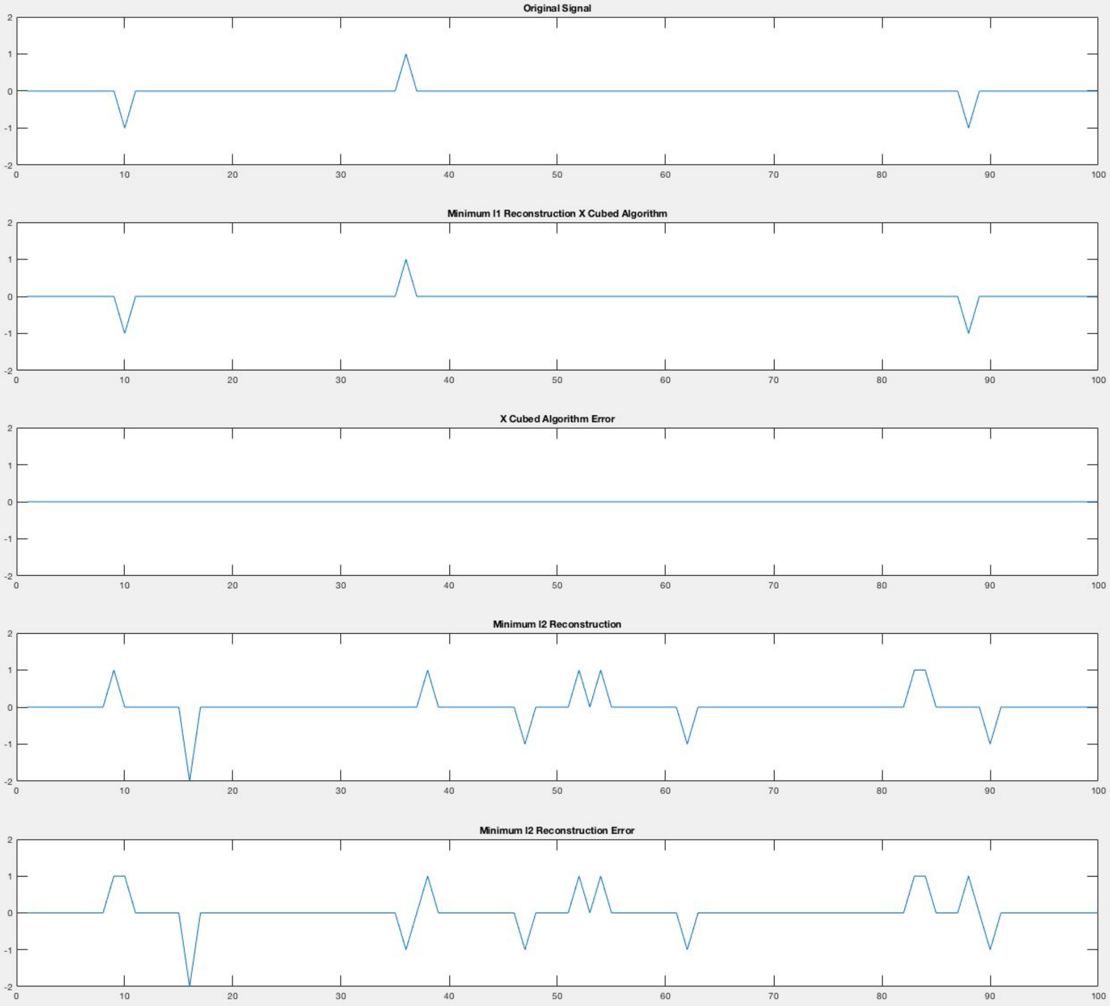


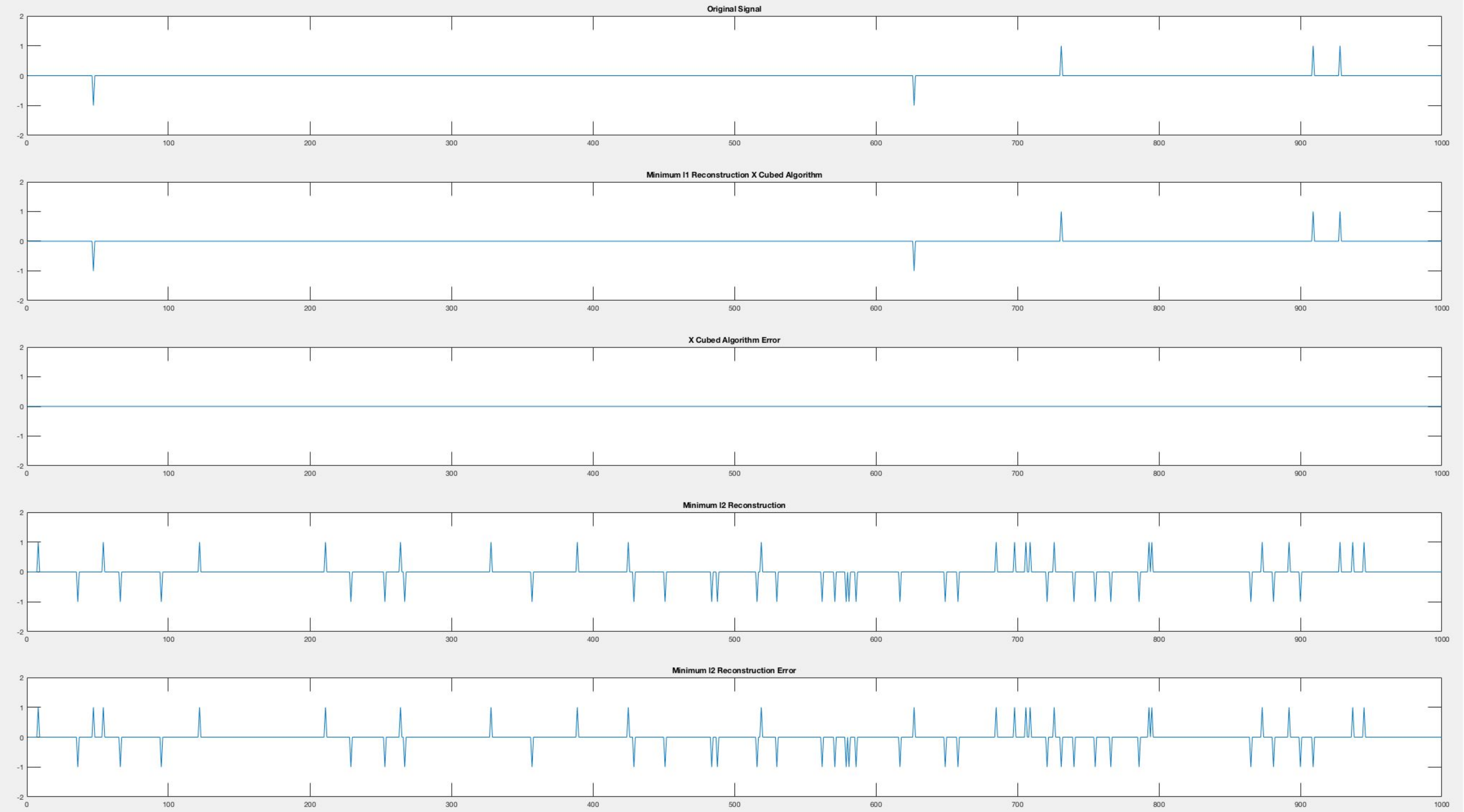


- Measurements: *global, incoherent*
  Good test functions:

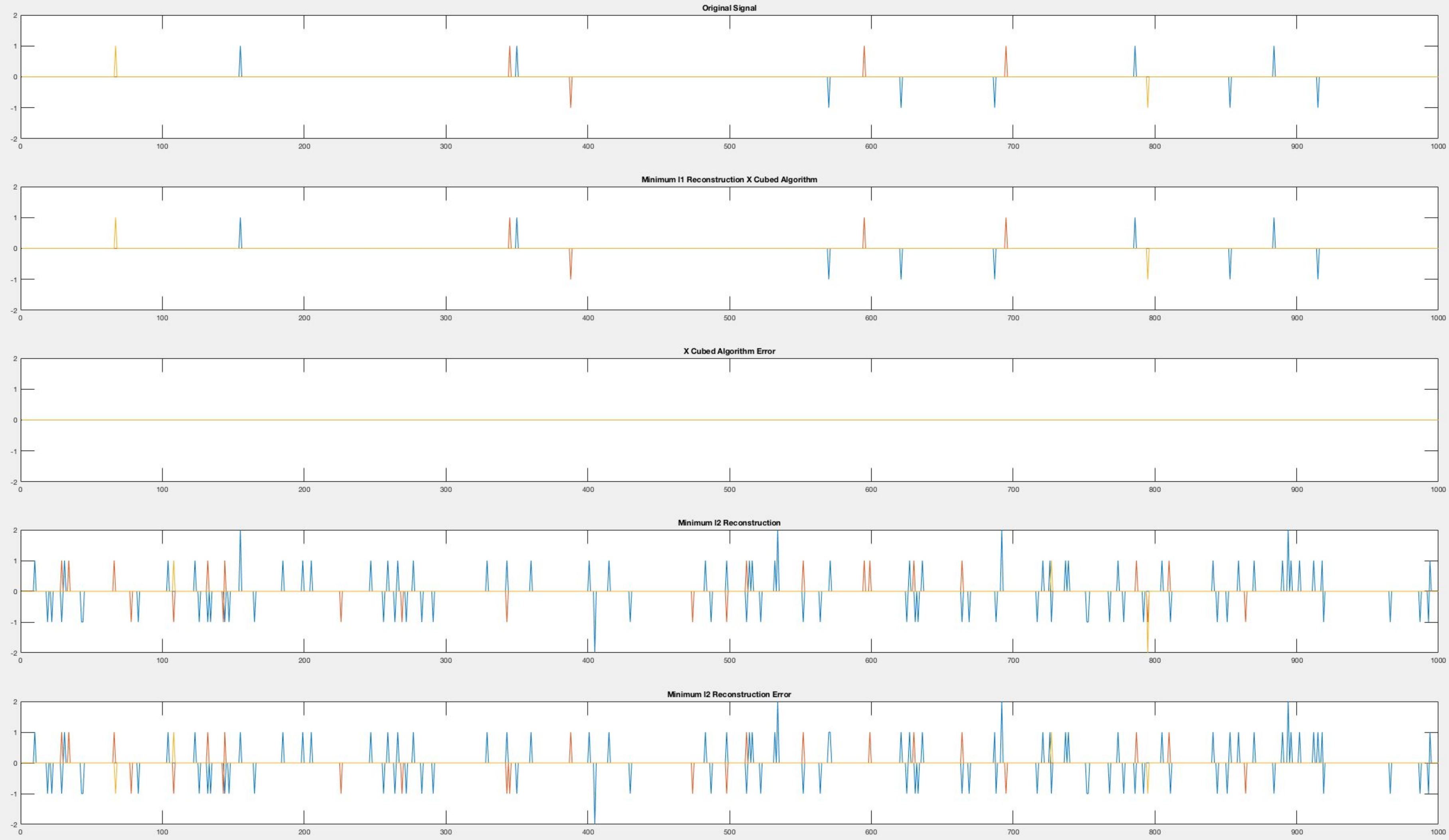




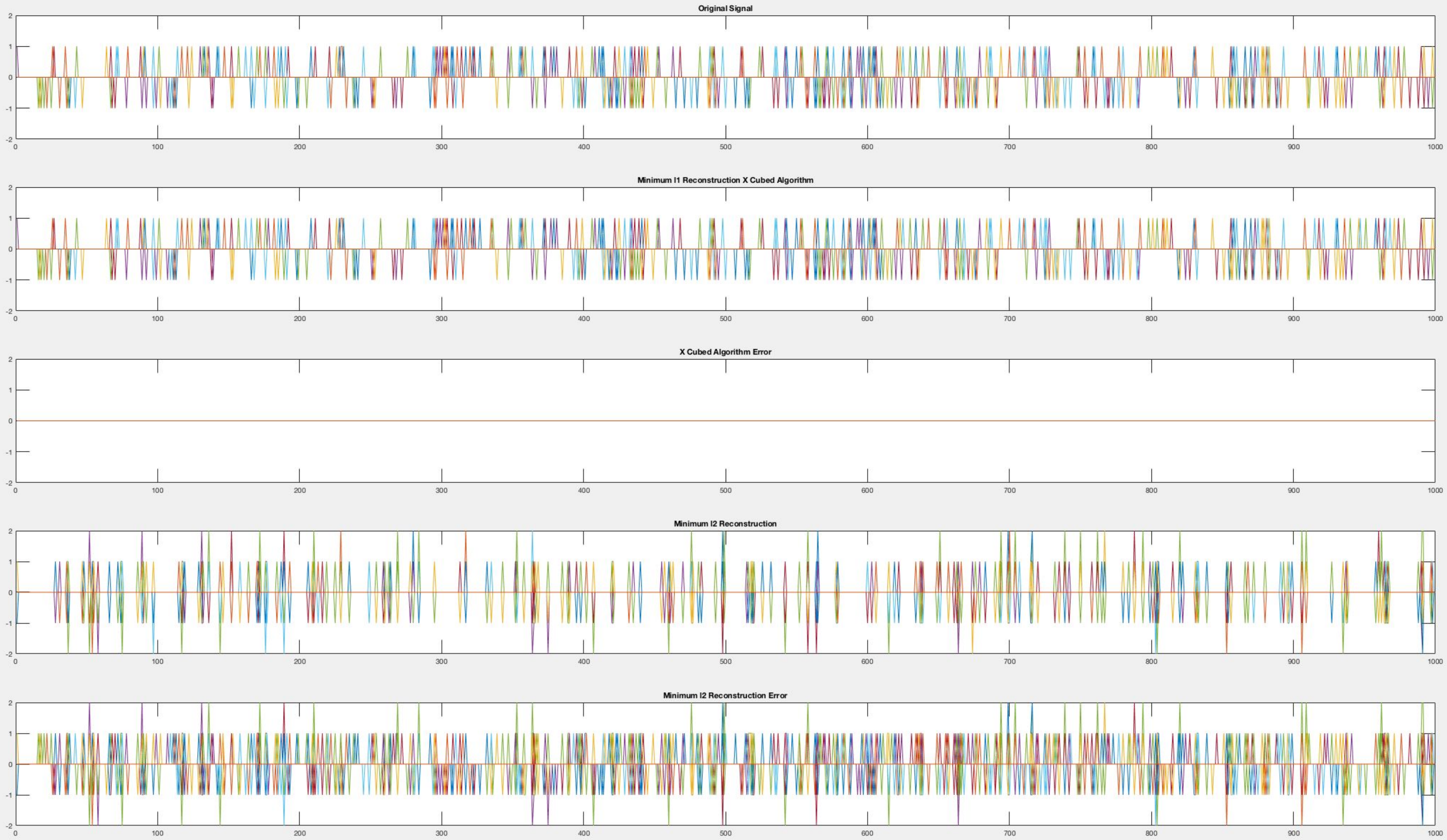

Original Signal
Minimum l1 Reconstruction X Cubed Algorithm
X Cubed Algorithm Error
Minimum l2 Reconstruction
Minimum l2 Reconstruction Error
2
1
0
-1
-2
0
10
20
30
40
50
60
70
80
90
100



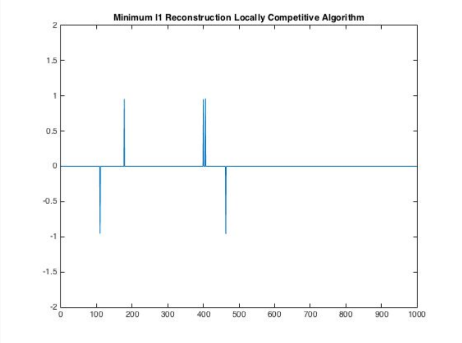
Original Signal
Minimum l1 Reconstruction X Cubed Algorithm
X Cubed Algorithm Error
Minimum l2 Reconstruction
Minimum l2 Reconstruction Error
2
1
0
-1
-2
0
100
200
300
400
500
600
700
800
900
1000

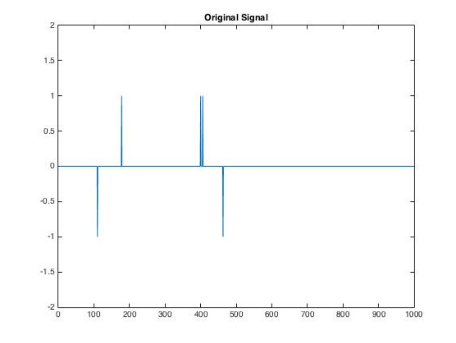


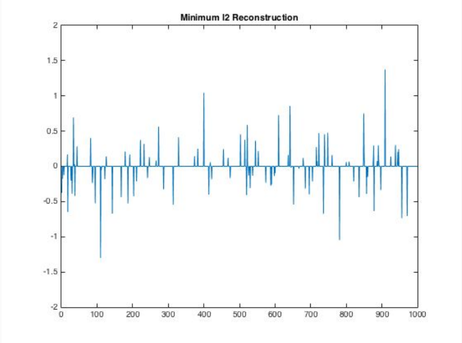
Original Signal
Minimum l1 Reconstruction X Cubed Algorithm
X Cubed Algorithm Error
Minimum l2 Reconstruction
Minimum l2 Reconstruction Error
2
1
0
-1
-2
0
100
200
300
400
500
600
700
800
900
1000

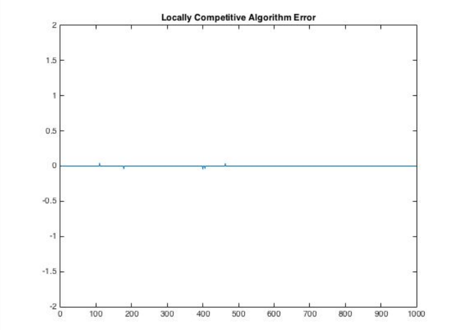


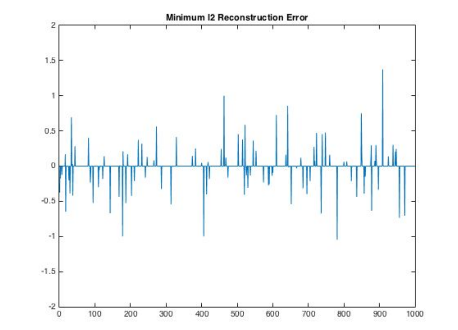
Original Signal
Minimum l1 Reconstruction X Cubed Algorithm
X Cubed Algorithm Error
Minimum l2 Reconstruction
Minimum l2 Reconstruction Error
2
1
0
-1
-2
0
100
200
300
400
500
600
700
800
900
1000

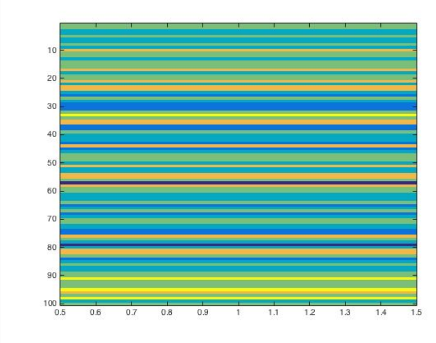

Minimum l1 Reconstruction Locally Competitive Algorithm

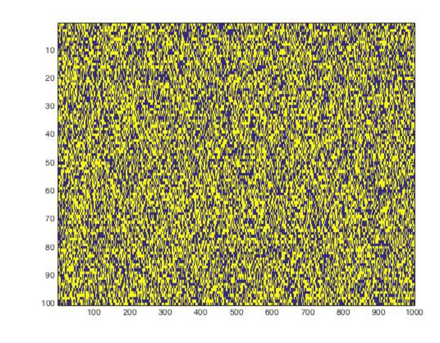

Original Signal

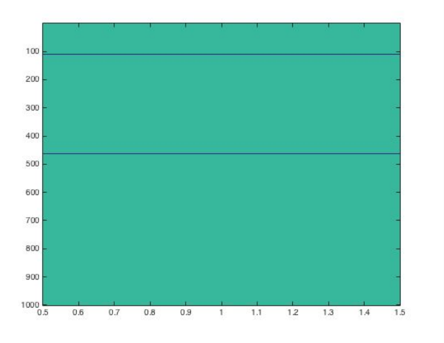

Minimum l2 Reconstruction

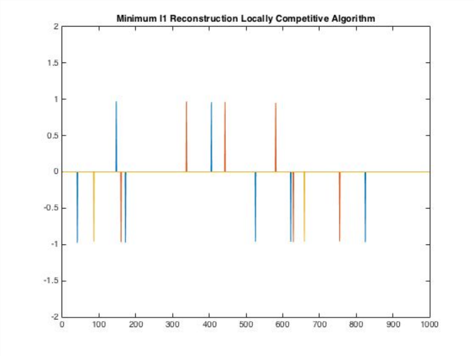

Locally Competitive Algorithm Error

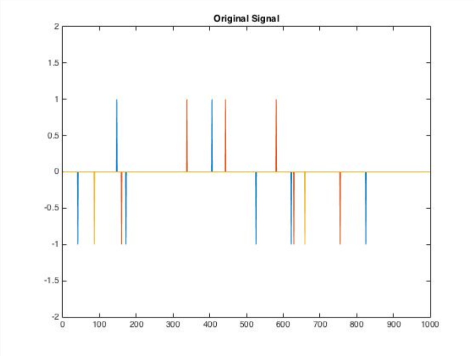

Minimum l2 Reconstruction Error

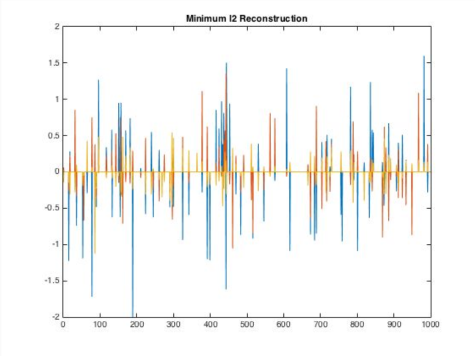


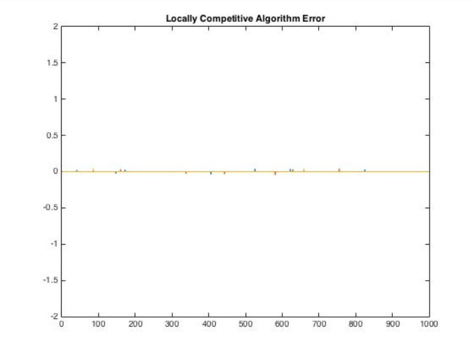


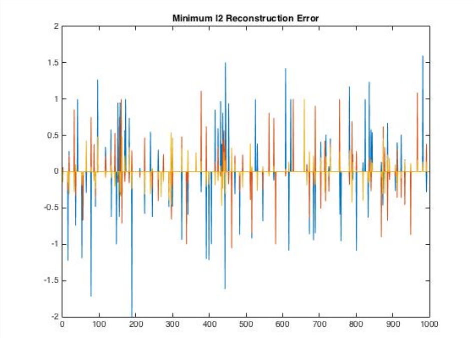


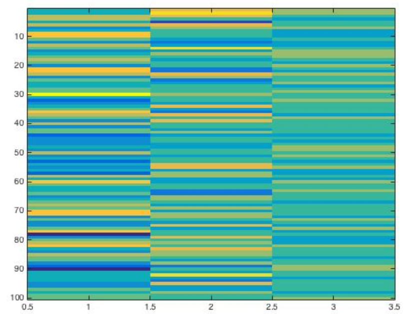

Minimum l1 Reconstruction Locally Competitive Algorithm
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0 100 200 300 400 500 600 700 800 900 1000

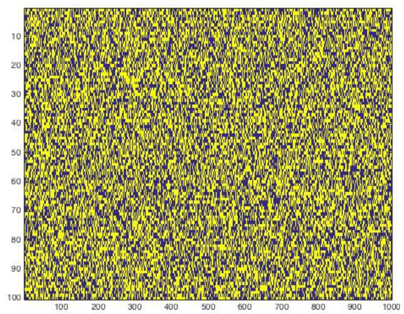

Original Signal
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0 100 200 300 400 500 600 700 800 900 1000

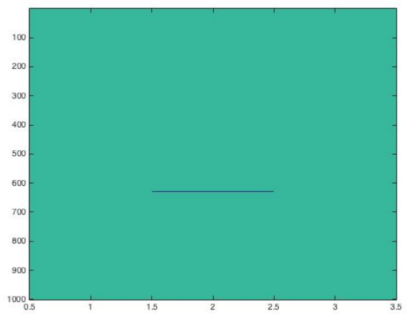

Minimum l2 Reconstruction
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0 100 200 300 400 500 600 700 800 900 1000

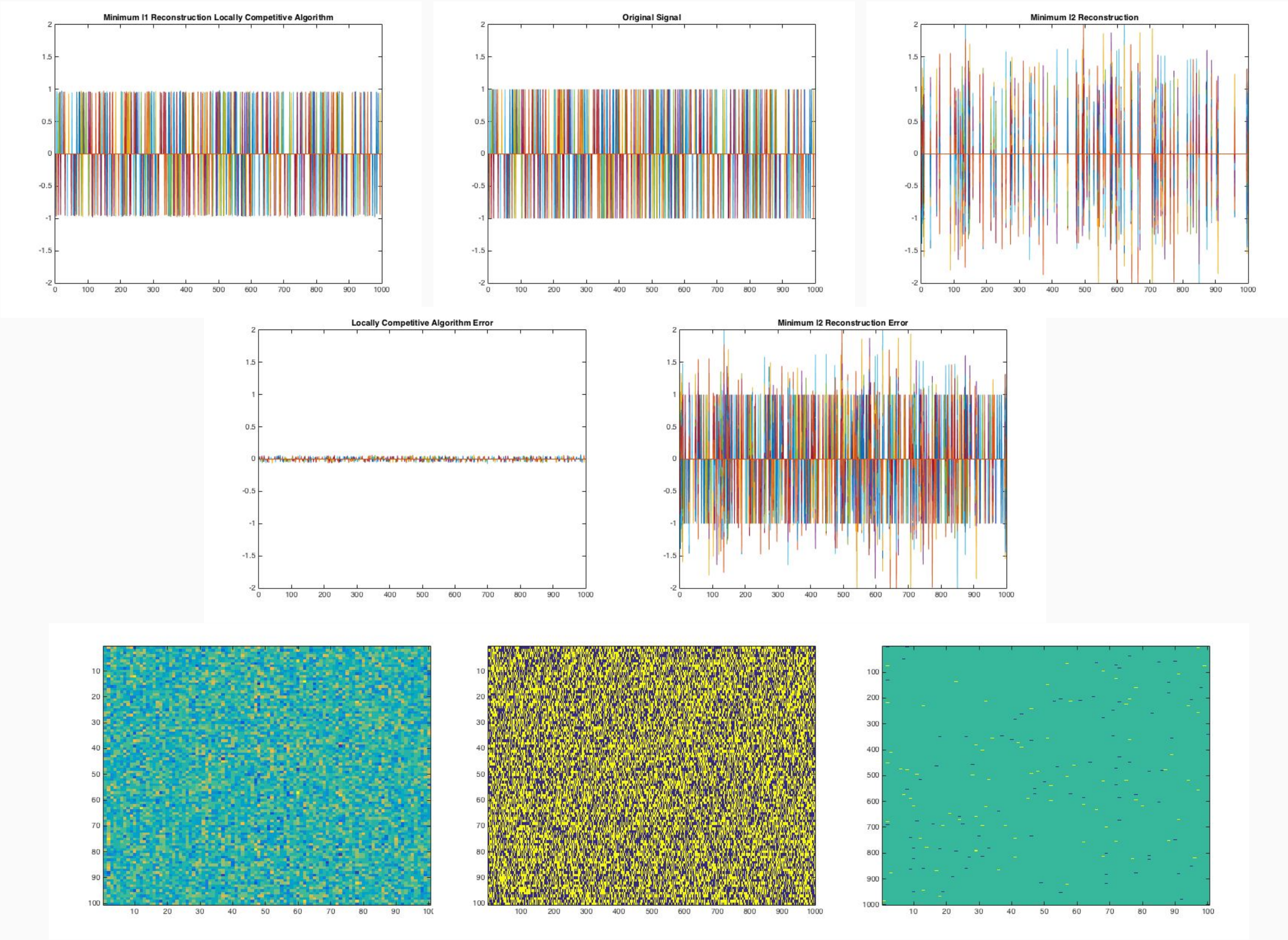

Locally Competitive Algorithm Error
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0 100 200 300 400 500 600 700 800 900 1000


Minimum l2 Reconstruction Error
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0 100 200 300 400 500 600 700 800 900 1000


10 20 30 40 50 60 70 80 90 100
0.5 1 1.5 2 2.5 3 3.5


10 20 30 40 50 60 70 80 90 100
100 200 300 400 500 600 700 800 900 1000

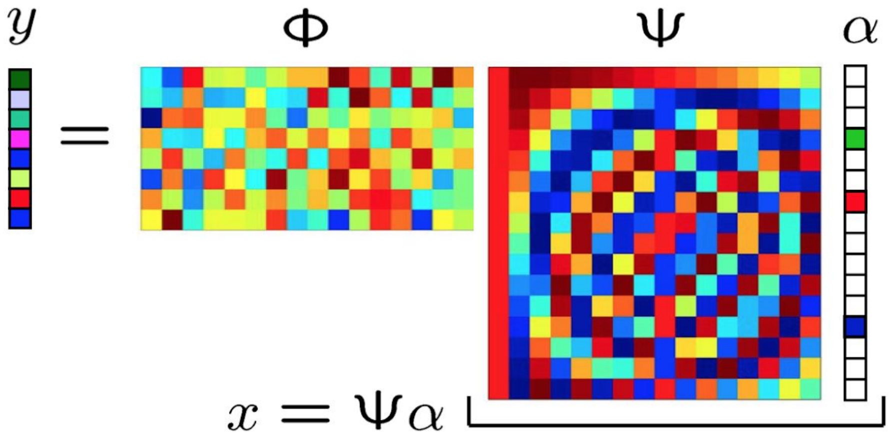

100 200 300 400 500 600 700 800 900 1000
0.5 1 1.5 2 2.5 3 3.5

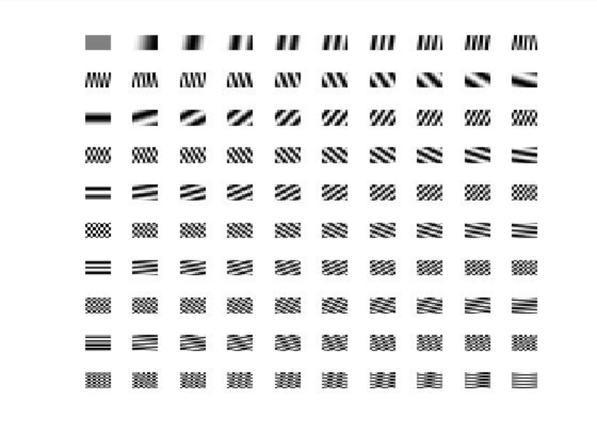

Minimum l1 Reconstruction Locally Competitive Algorithm
Original Signal
Minimum l2 Reconstruction
Locally Competitive Algorithm Error
Minimum l2 Reconstruction Error

Universality

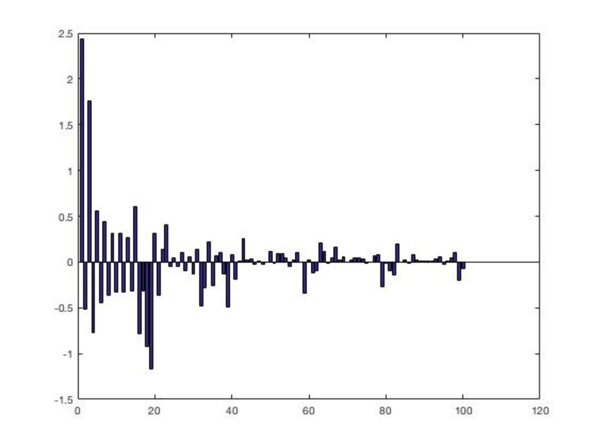
- Random measurements can be used for signals sparse in *any* basis

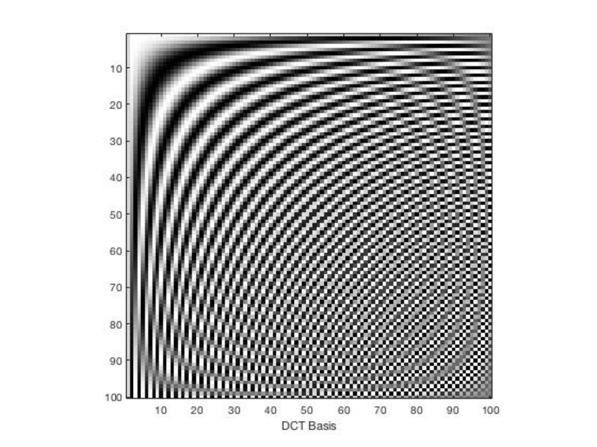
$$y = \Phi x = \Phi \Psi \alpha$$

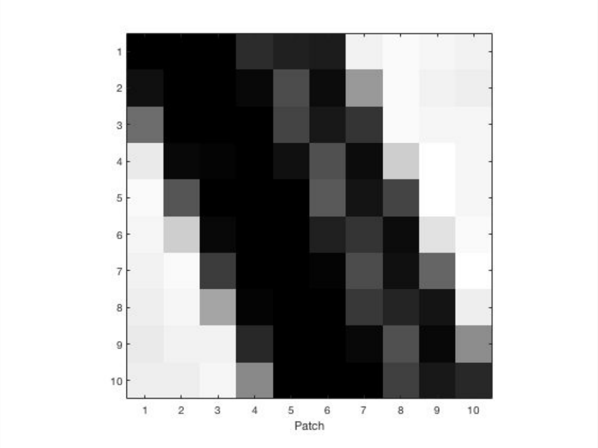


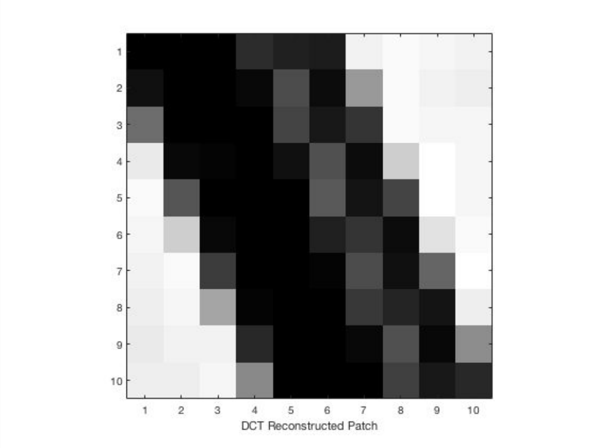


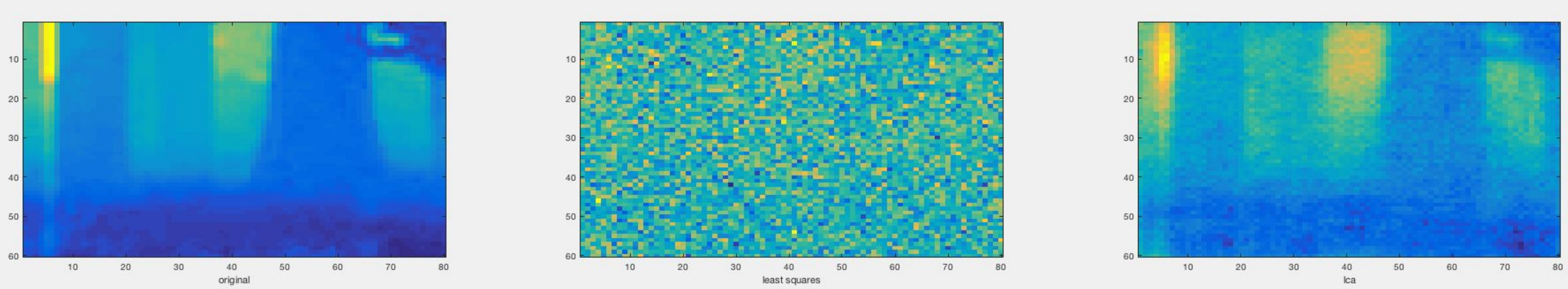

Full Image



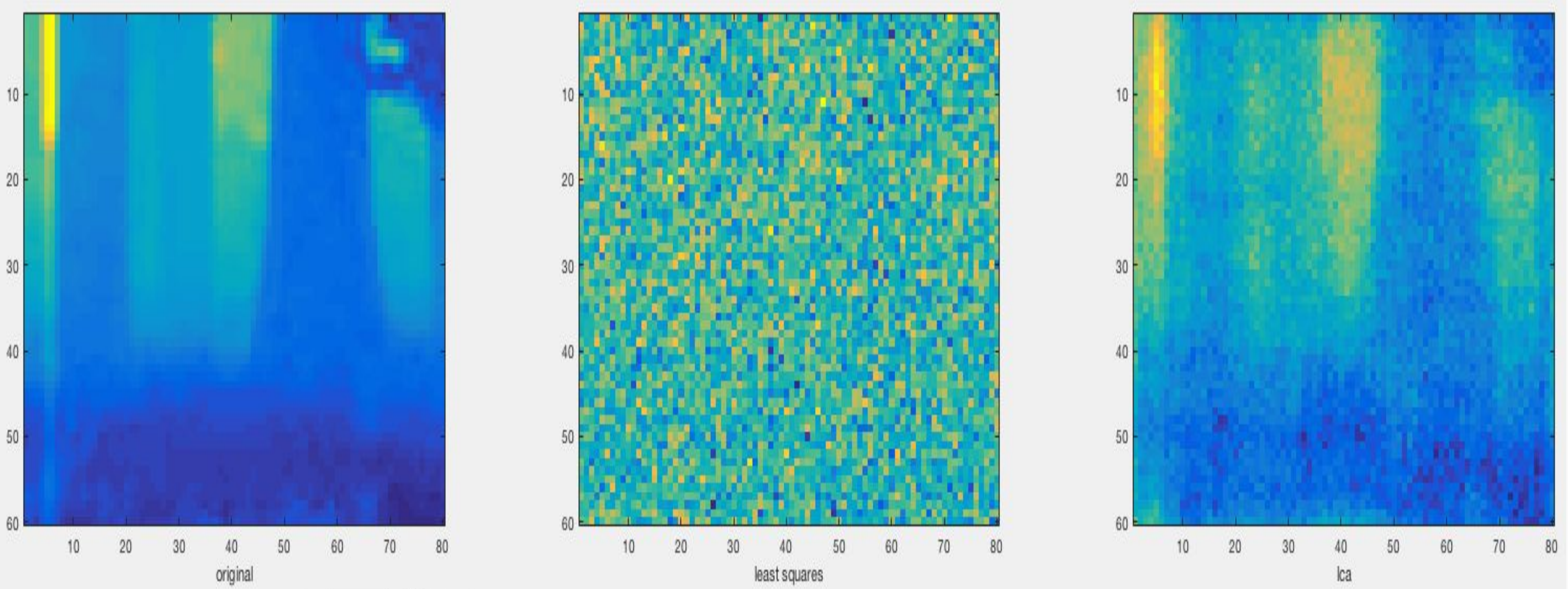

DCT Basis


Patch


DCT Reconstructed Patch

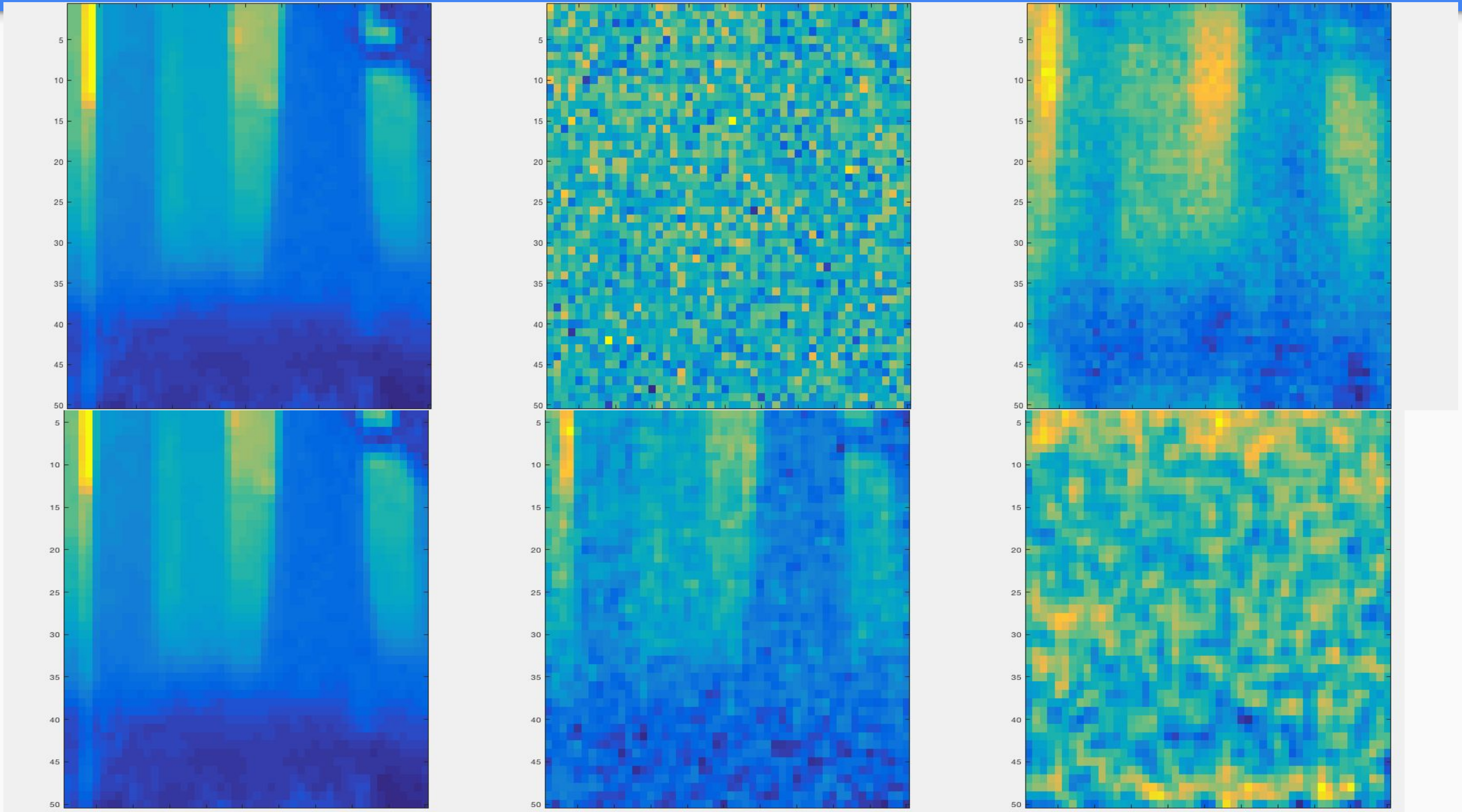


$$M = N/2$$



$M = N/3$

X^3 Dictionary for Compressed Sensing



Von Neumann wondered how

“an imperfect (biological) neural network, **containing many random connections**, can be made to perform reliably those functions which might be represented by idealized wiring diagrams.”

Information Scalability

Many applications involve signal

Inference and not **Reconstruction**

Detection < **Classification** < **Estimation** < **Reconstruction**

Information Scalability

Learning + Inference

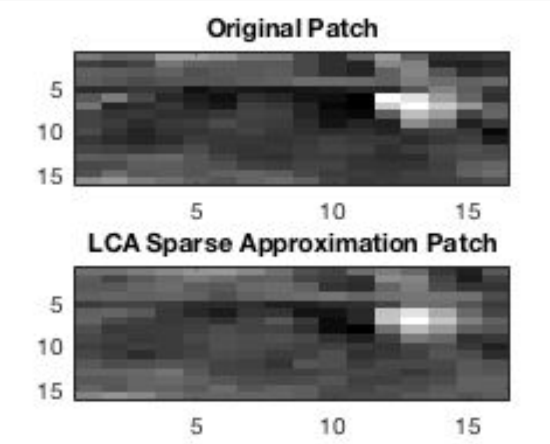
Processing directly on compressed measurements:

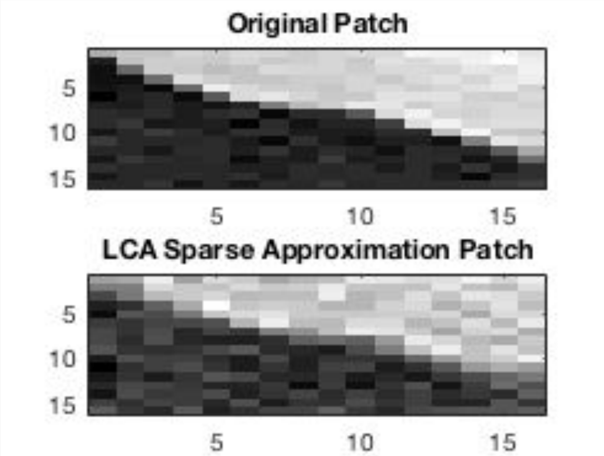
Random projections ~ sufficient statistics

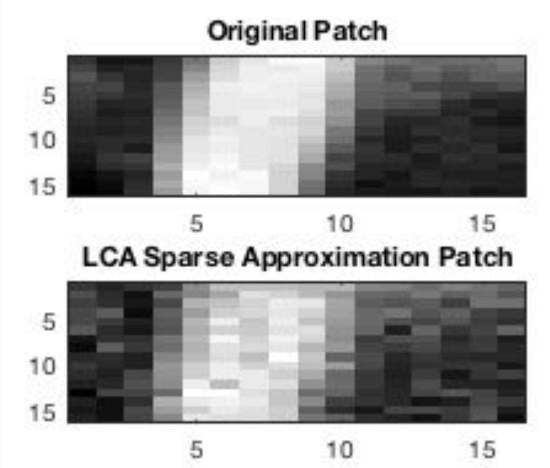
Questions and Comments

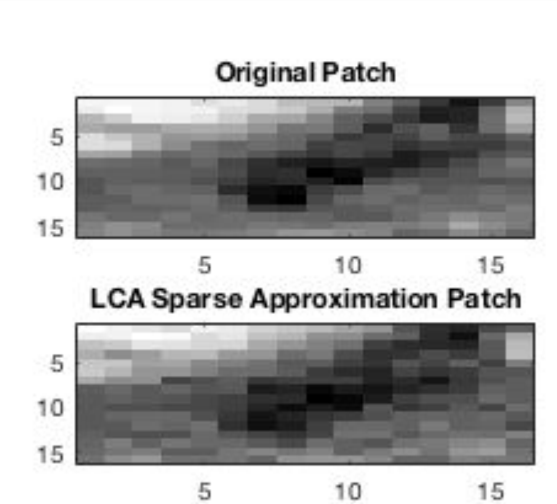
Thank You

"When Leibniz was first thinking about computation at the end of the 1600s, the thing he wanted to do was to build a machine that would effectively answer...questions." - Stephen Wolfram

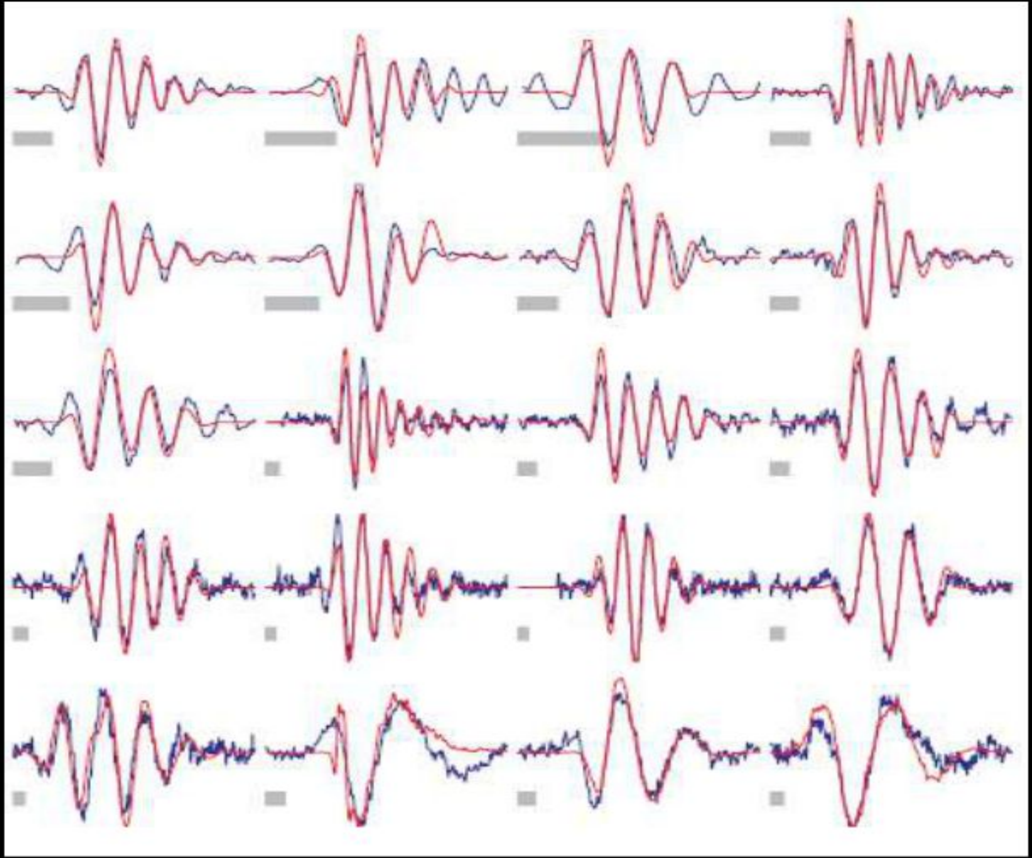

Original Patch
5
10
15
5
10
15
LCA Sparse Approximation Patch
5
10
15
5
10
15


Original Patch
5
10
15
5
10
15
LCA Sparse Approximation Patch
5
10
15
5
10
15


Original Patch
5
10
15
5
10
15
LCA Sparse Approximation Patch
5
10
15
5
10
15

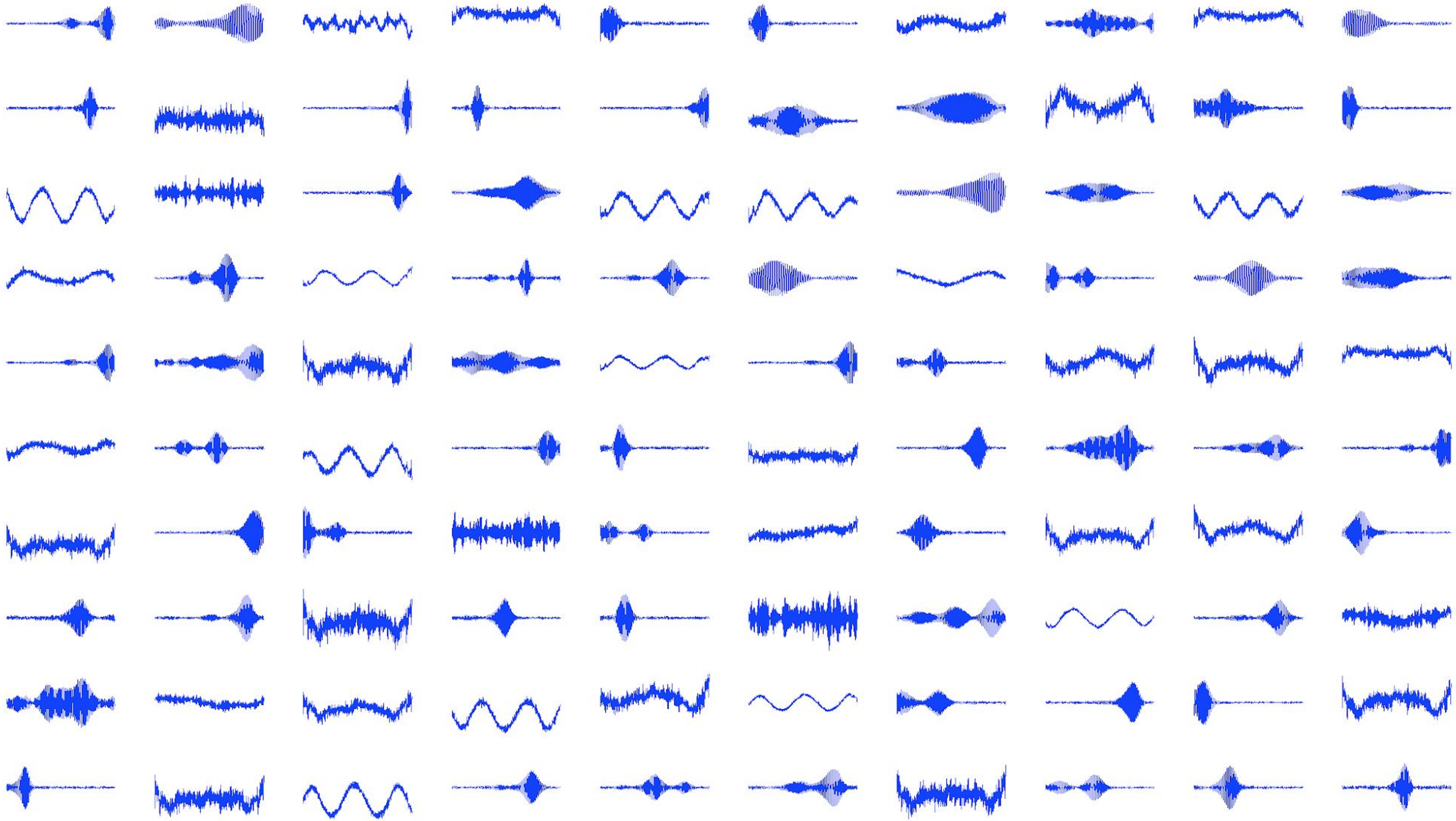

Original Patch
5
10
15
5
10
15
LCA Sparse Approximation Patch
5
10
15
5
10
15

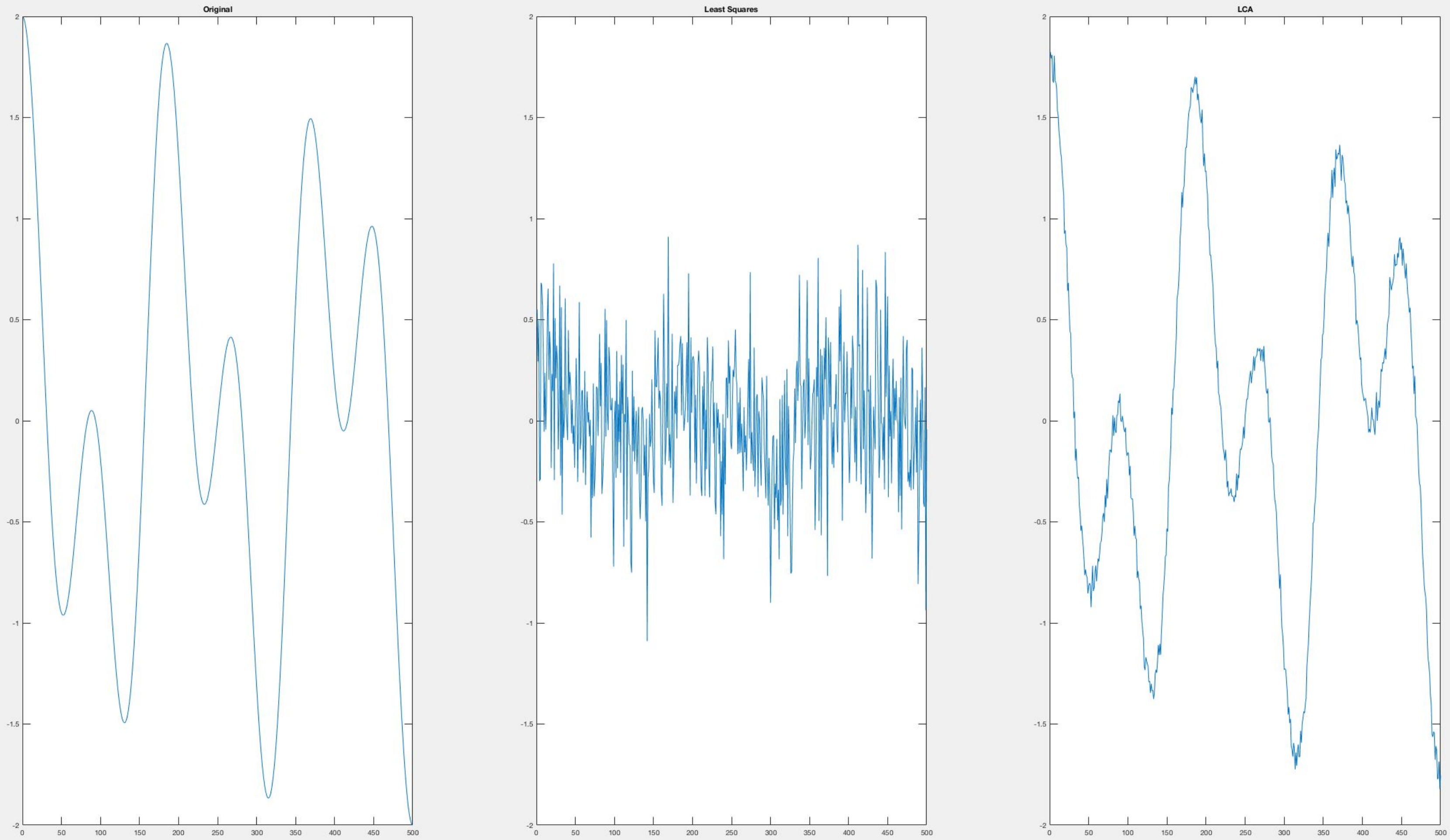
Digression: Sparse coding applied to audio



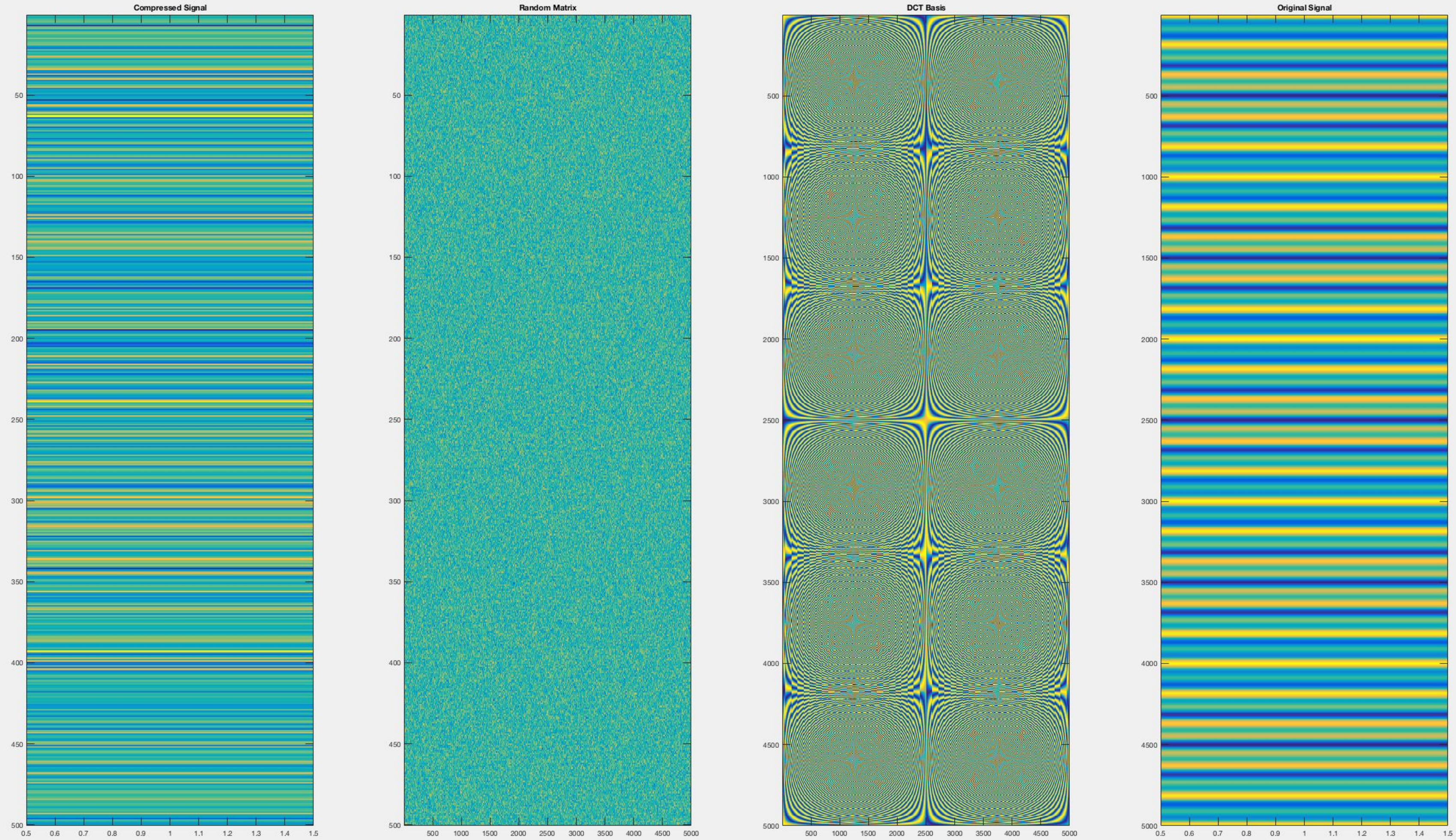
[Evan Smith & Mike Lewicki, 2006]

Andrew Ng





Original
Least Squares
LCA
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0
50
100
150
200
250
300
350
400
450
500


Compressed Signal
Random Matrix
DCT Basis
Original Signal
50
100
150
200
250
300
350
400
450
500
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
1.4
1.5
1000
1500
2000
2500
3000
3500
4000
4500
5000