METHODS FOR SPARSE ANALYSIS OF HIGH-DIMENSIONAL DATA, I

Rebecca Willett

DUKE
ENGINEERING
EDMUND T. PRATT, JR.
SCHOOL OF ENGINEERING

HIGH-DIMENSIONAL DATA

CONSUMER PREFERENCES

A company records how much you like each of N products in its database, and wants to predict what else you'll like.

ACTUARIAL SCIENCE

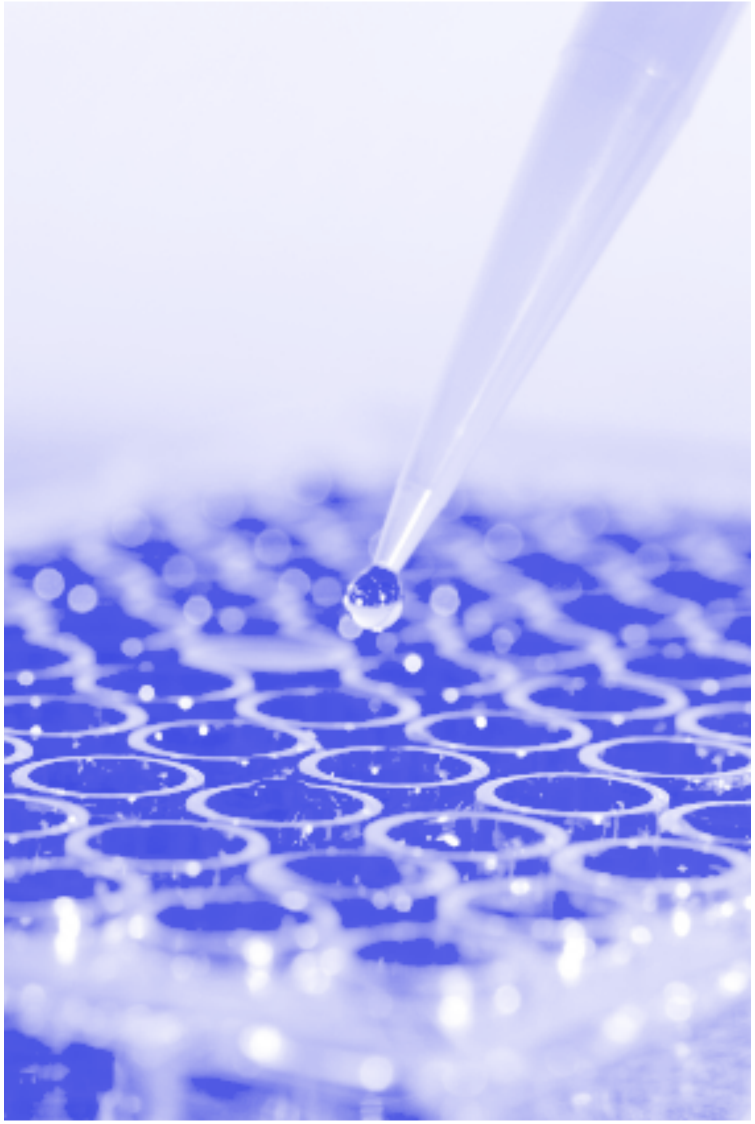
Your insurance company asks you N questions about yourself and family. Based on your responses and history, they want to predict how much you'll cost the insurance company.

"He's 104. He always comes back on his birthday - I worked out he'd be gone at 69."

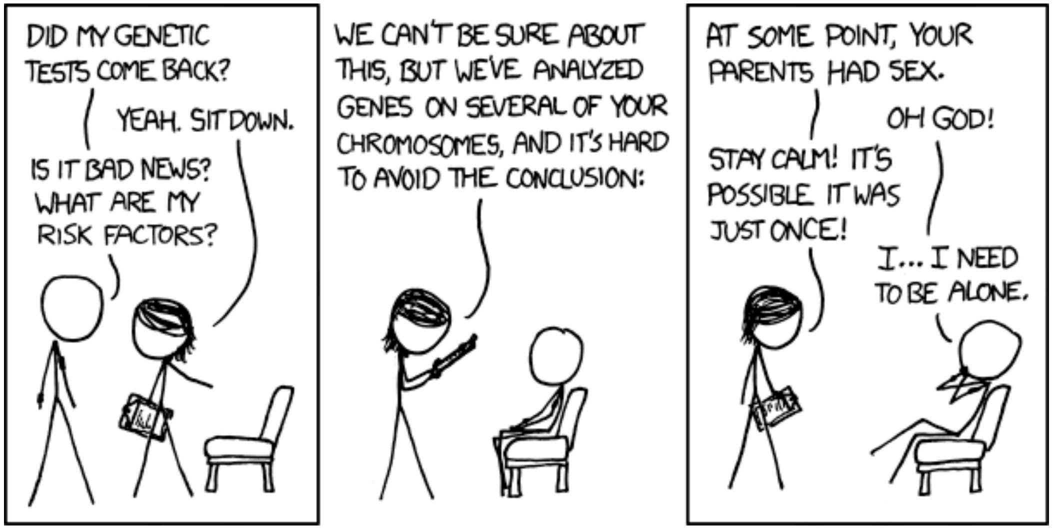
Image Processing and Analysis

An N-pixel image is a single point in $\mathbb{R}^N$.

Genetic Analysis



We record N genes for each person in a population. Only a few people have a given genetic disease.



The Curse of Dimensionality

- In many such settings, we have a small number of points in $\mathbb{R}^N$, where N can be very high.
- We also have a prediction task (e.g. regression/function estimation, classification, approximation, clustering, optimization)
- If we want to perform that task with accuracy ε, then we need

$$O[(1/\varepsilon)^N]$$

data points or observations, which are unavailable in real-world settings or create massive run times.

Modeling and approximation are mathematically and computationally FORMIDABLE.

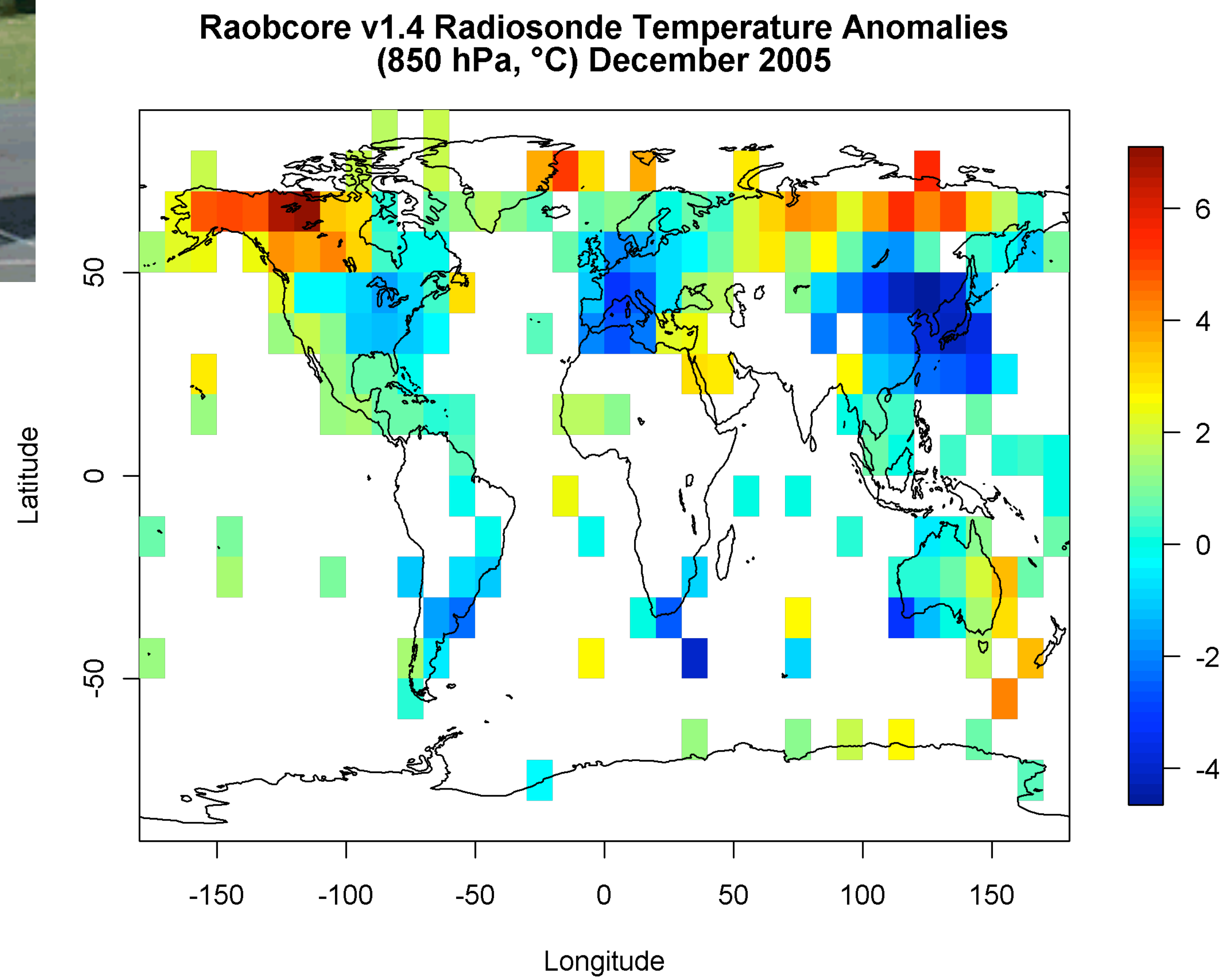
Significant Data-Processing Challenges

Experiments and measurements are noisy, corrupted, or unreliable.

Information processing and decision making must be robust to uncertainty.

We need to learn and analyze the structure of networks

We can't observe/sense everything all the time; incomplete, missing, or indirect data are the norm



Signals can require significant storage space (111 kB)

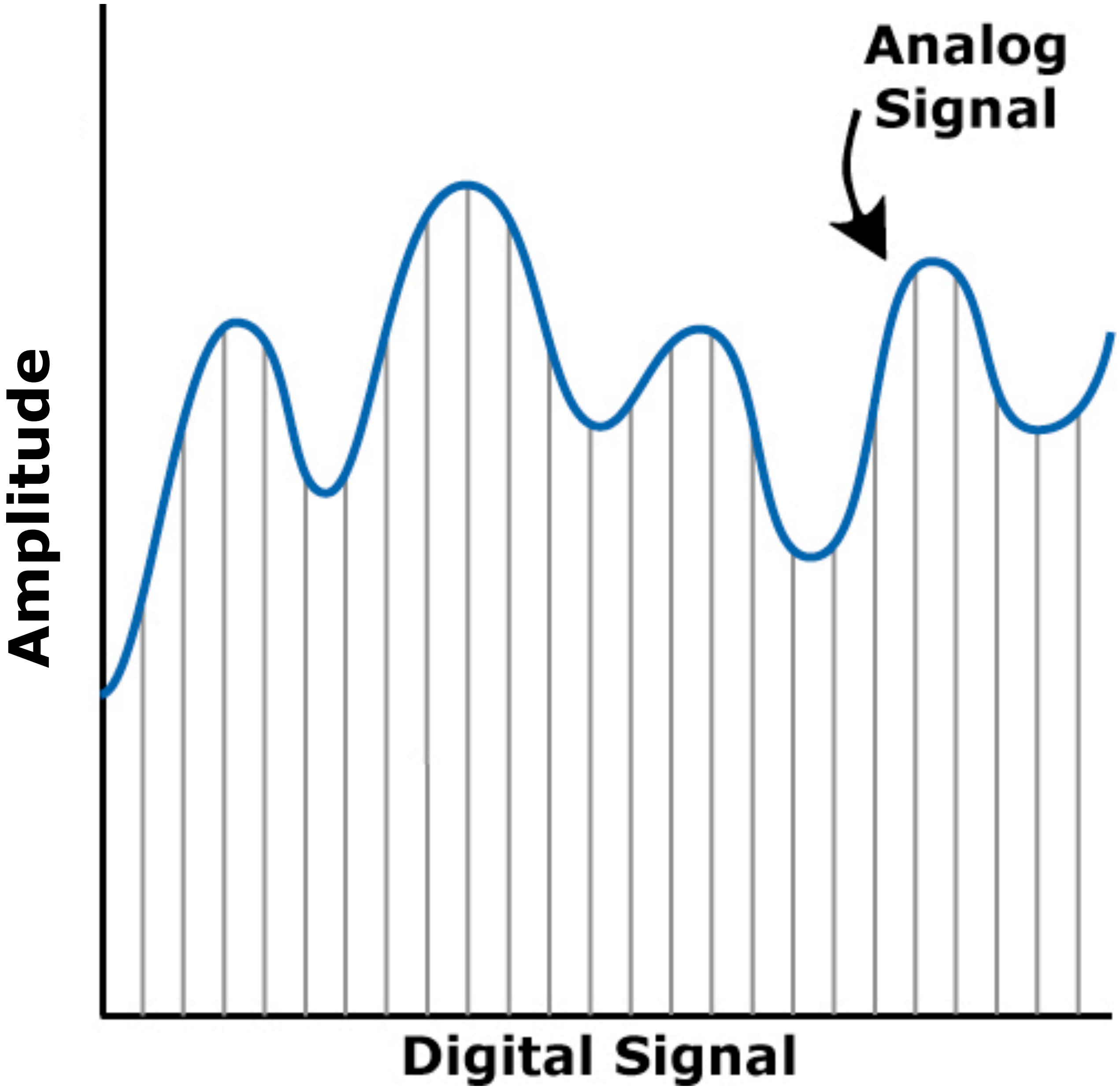
We need to minimize storage space utilized (12 kB)


We need to process huge amounts of streaming data

FAST COMPUTATIONS ON STREAMING DATA

- Imaging computing the Fourier transform of a length-N signal, where N is HUGE.
- If we use naïve matrix multiplication, this takes
$$O(N^2)$$
operations.
- If we use the Fast Fourier Transform, this takes
$$O(N \log N)$$
operations.
- Can we do even better?

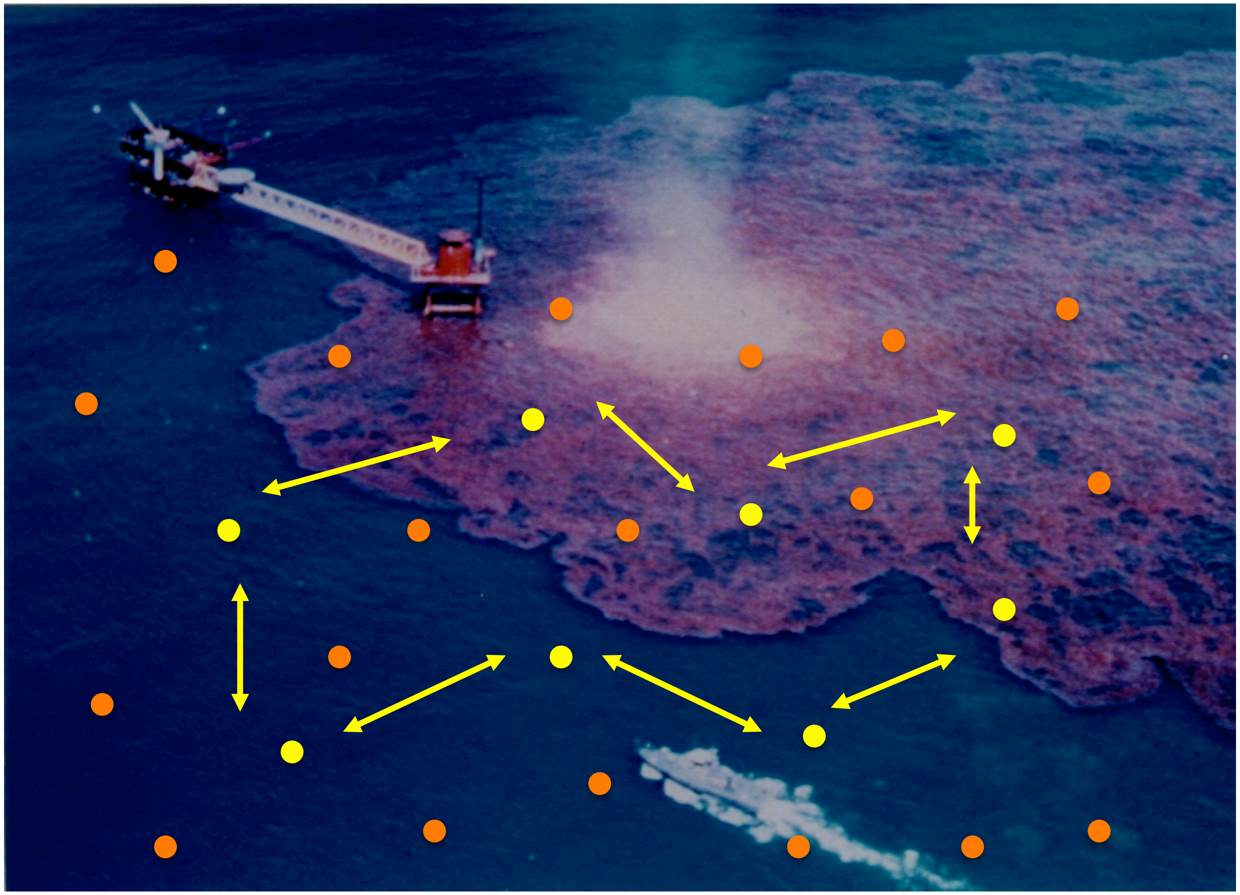
We need to improve Analog-to-Digital converters



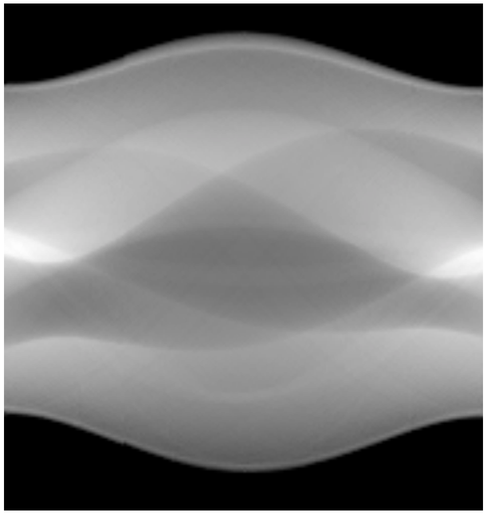
We need to improve Analog-to-Digital converters

This is what you would get with a typical low-resolution camera. Can we do better?

We need to reduce power consumption in sensor networks



We need to solve inverse problems



Sinogram data

$$y = Af + \epsilon$$

data image noise

Tomographic projections

Brain slice

$$\hat{f}(y)$$

Reconstruction from data

Diversity: data come from disparate sources; we must integrate info from different sensors, experiments, people, etc.

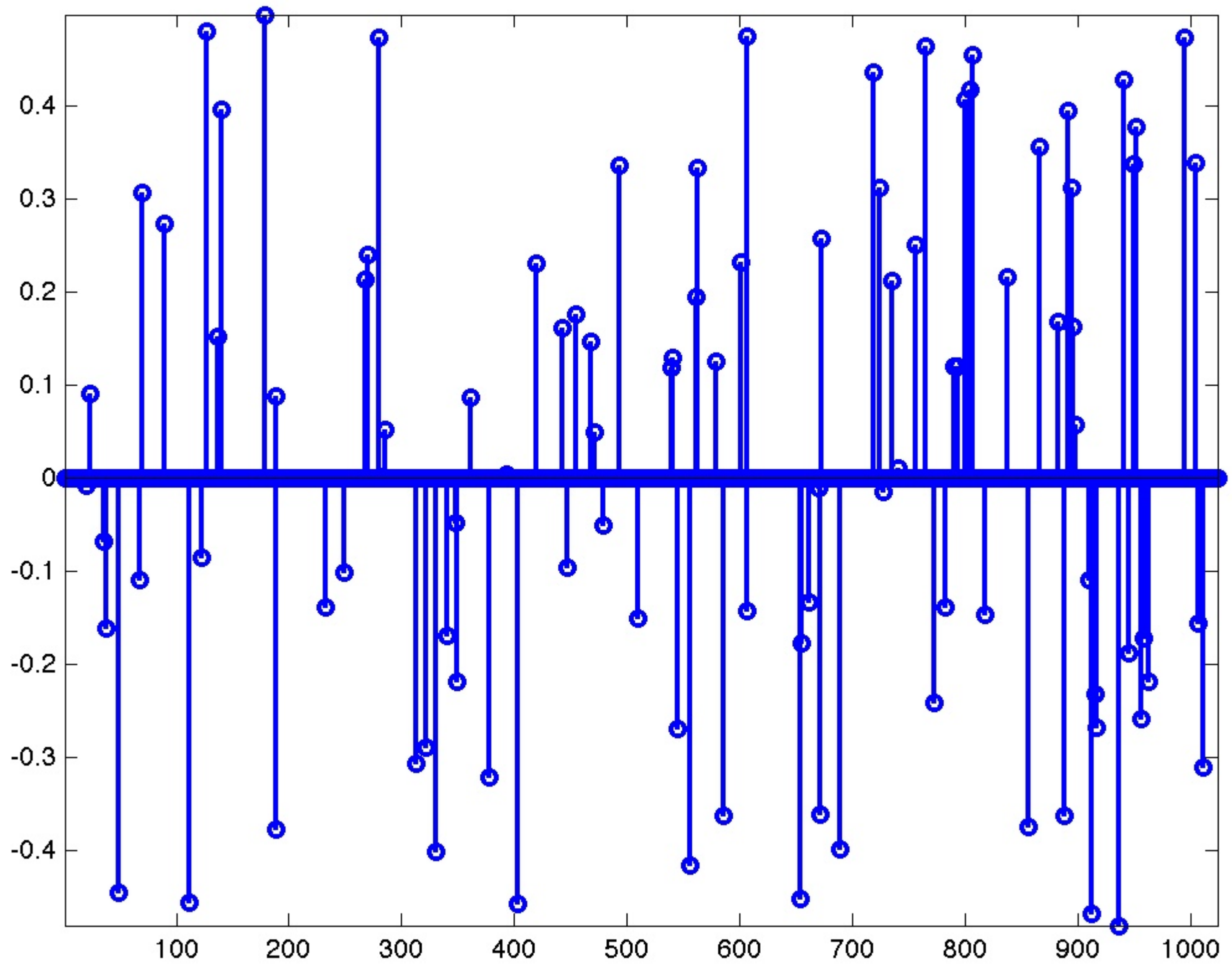

Sparsity
Compu-
tation

SPARSITY:
BASES, DICTIONARIES, AND APPROXIMATION

Rebecca Willett

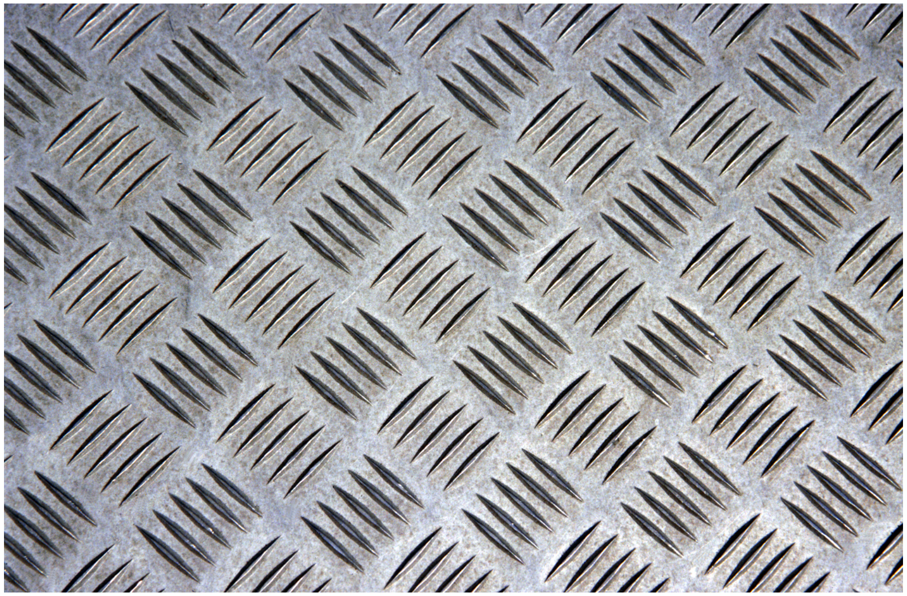
DUKE
EDMUND T. PRATT, JR.
SCHOOL OF ENGINEERING

What is sparsity?



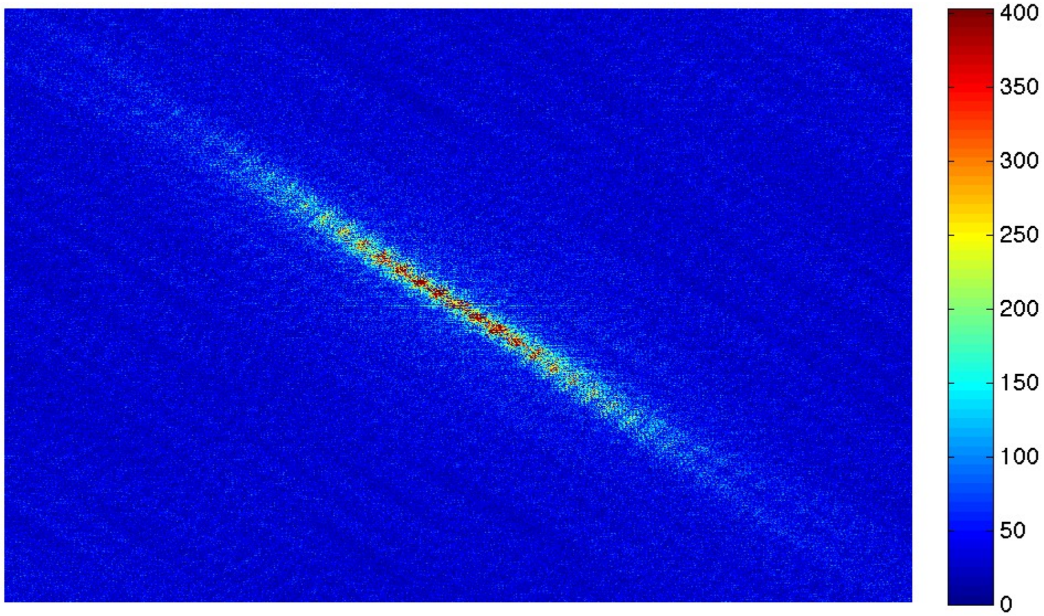
Most elements are zero-valued; only ~10% are non-zero.


Most pixels are zero-valued (black)



This image is not sparse…

but it's Fourier Transform is.



The collection of links/edges connecting people is sparse

WHY SHOULD SPARSITY HELP?

First some initial insight…

Pill weighing problem

- On the shelf you have 10 identical bottles of identical pills (let's say there's one pill in each bottle). However, one of those 10 bottles contains a cheap knockoff pill. (sparsity)
- The only way to differentiate fake pills from real pills is the weight - while real pills weigh 1 g each, the knockoff pills are only 0.9 g.
- You have one scale that shows the exact weight (down to the mg) of whatever is weighed.
- How can you tell which bottle contains fake pills with as few weighings as possible?

PILL WEIGHING PROBLEM

- On the shelf you have 10 identical bottles of identical pills (let’s say there’s one hundred pills in each bottle). However, one of those 10 bottles contains cheap knockoff pills. (sparsity)
- The only way to differentiate fake pills from real pills is the weight - while real pills weigh 1 g each, the knockoff pills are only 0.9 g.
- You have one scale that shows the exact weight (down to the mg) of whatever is weighed.
- How can you tell which bottle contains fake pills with just **1** weighing?

PARAMETRIC SIGNALS

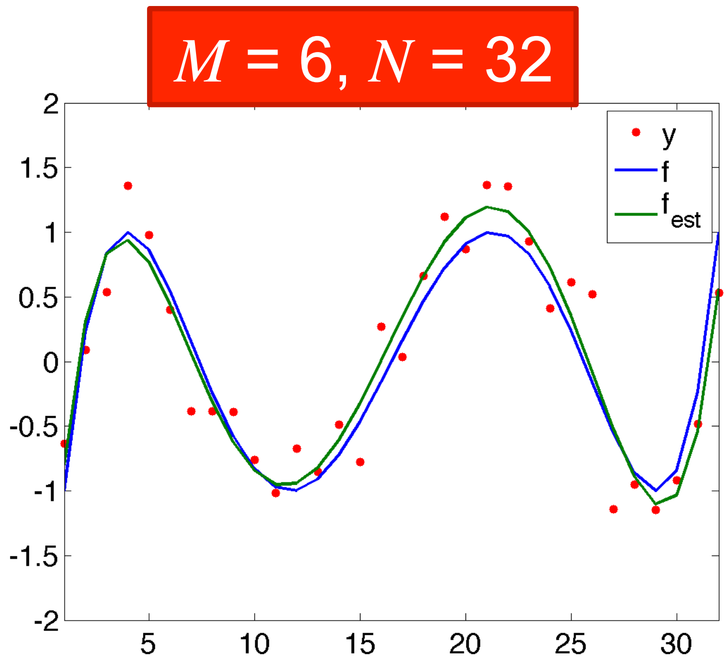
Say we make noisy measurements of a parametric signal:

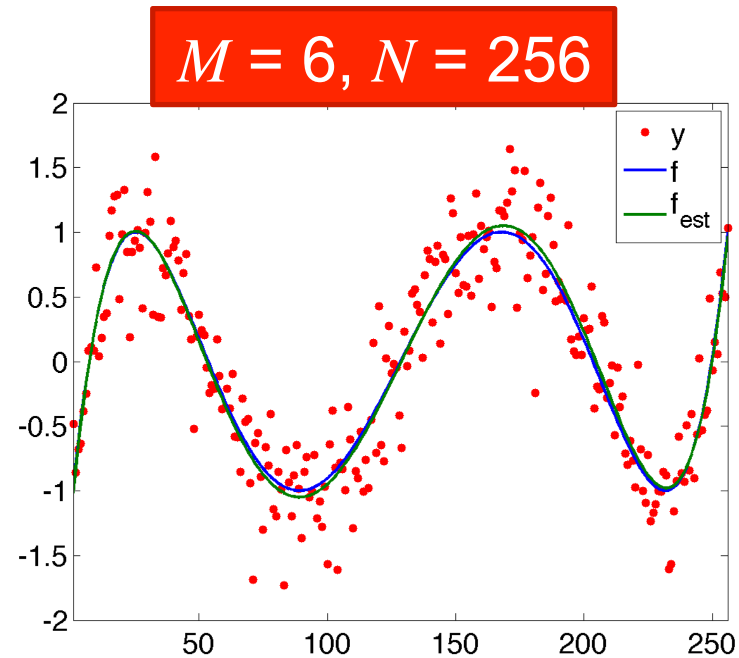
$$y_n = f_n + \epsilon_n, \quad n = 1, \ldots, N,$$

where, for instance,

$$f_n = a_0 + a_1 n + a_2 n^2 + \cdots + a_{M-1} n^{M-1}.$$

We want to estimate $f \triangleq [f_1, \ldots, f_N]$ from $y \triangleq [y_1, \ldots, y_N]$.

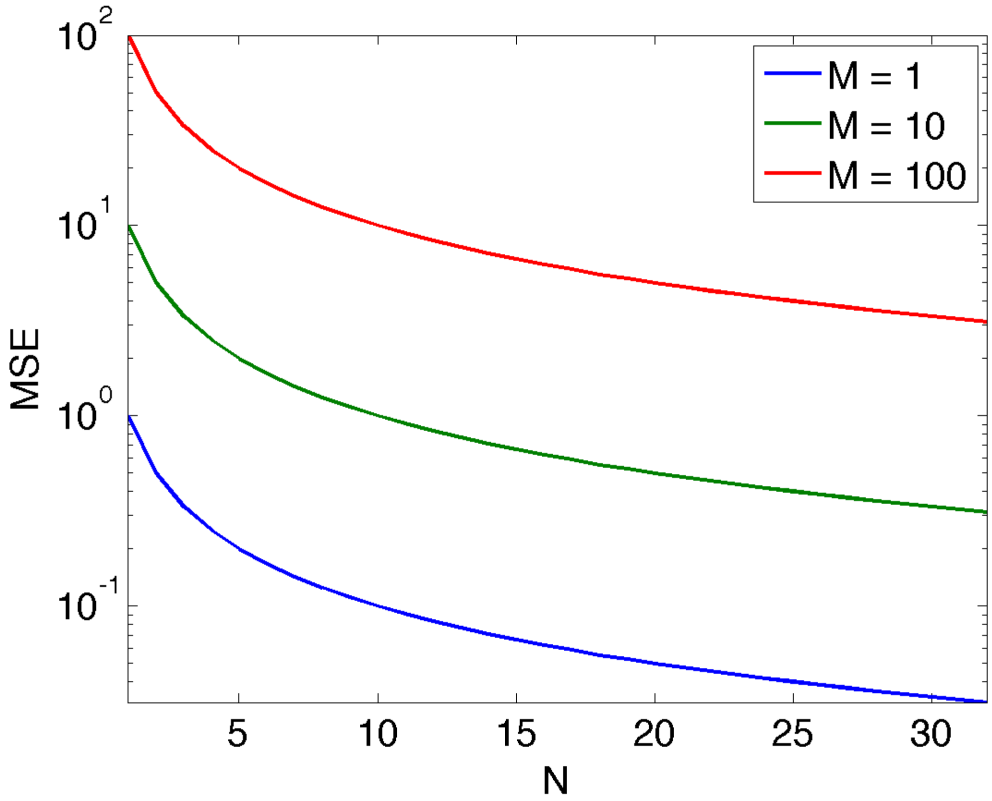




PARAMETRIC SIGNALS

In general, the best possible mean squared error (MSE) decays as we collect more data (i.e. as N increases) like

$$\text{MSE} \triangleq \frac{\|f - \widehat{f}\|_2^2}{N} = \frac{1}{N}\sum_{n=1}^{N}(f_n - \widehat{f}_n)^2 \preceq \frac{M}{N}.$$



Non-parametric signals

With parametric signals, we have M degrees of freedom – M different parameters to estimate. However, in many real-world problems we don't have access to a good parametric model.

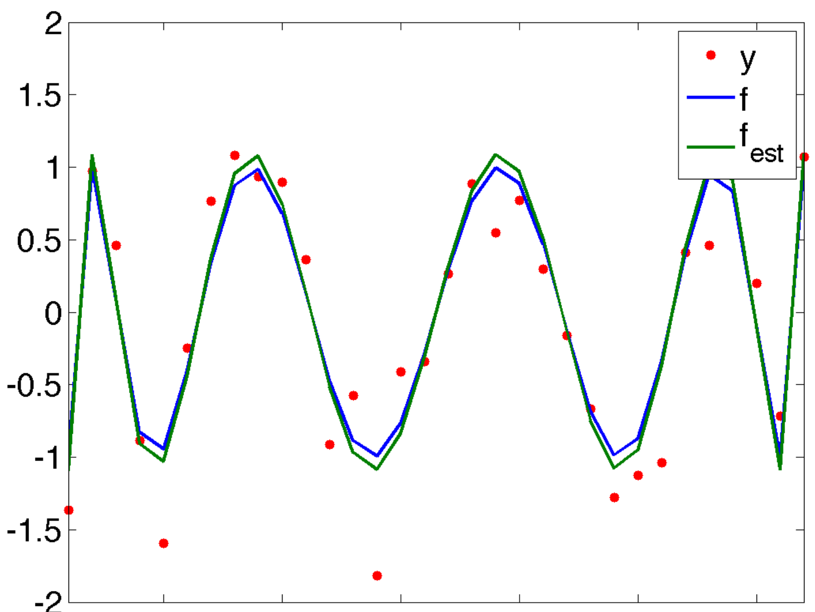
Without a parametric model, we have $M \approx N$ degrees of freedom, and without additional assumptions our MSE is $O(1)$ – i.e. our error does not go down as we collect more data.

SPARSE SIGNALS

With sparse signals, we assume that only K of the N possible degrees of freedom are significant or non-zero.

Most techniques which exploit sparsity have two components:
(a) determining which K-sparse model is best, and
(b) using that best sparse model as a parametric model.

The amazing part is that, with the right tools, we can often do almost as well as if we knew a parametric model in advance (e.g. MSE $= O(K/N)$).



M = 9, N = 32, K = 1

This is a high degree polynomial, but *sparse* in the Chebyshev polynomial basis.

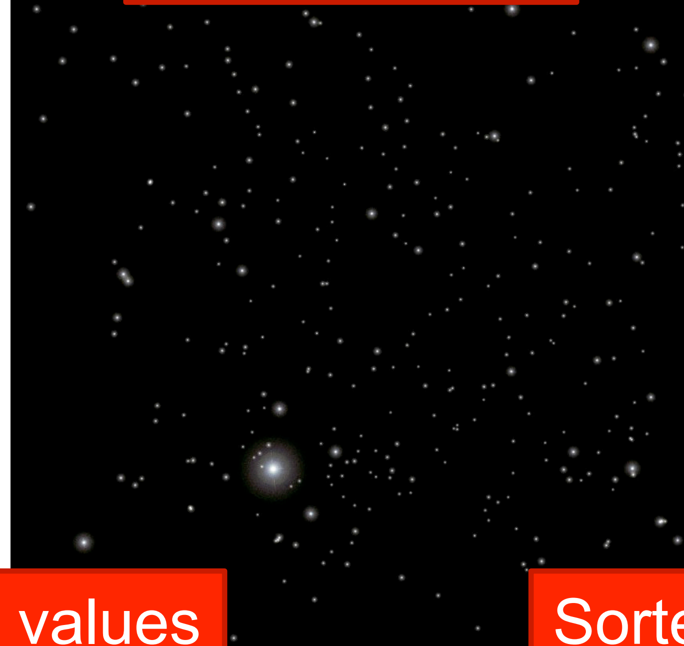
Sparsity and Compressibility

Definition: *A signal f is K-sparse if K or fewer elements of f are non-zero.*

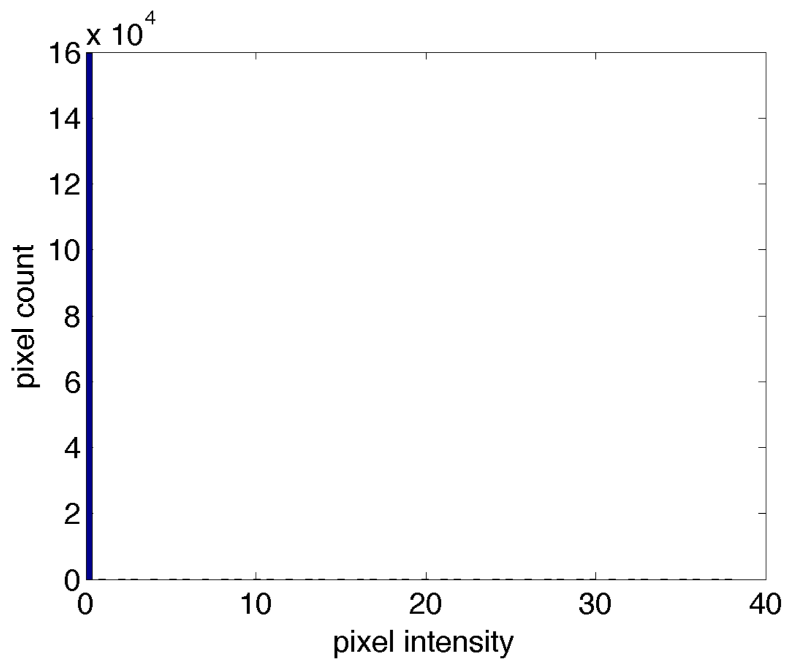
$$K \triangleq \#\{n : f_n \neq 0, n = 1, \ldots, N\}$$

This image has $N = 400^2$ pixels and is 344-sparse.

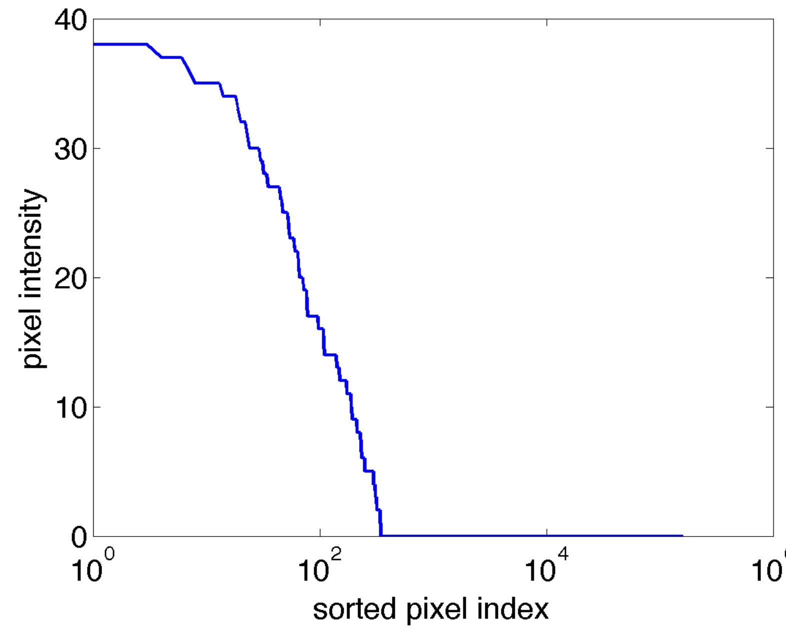
Original image



Histogram of pixel values



Sorted pixel intensities



Compressible Signals

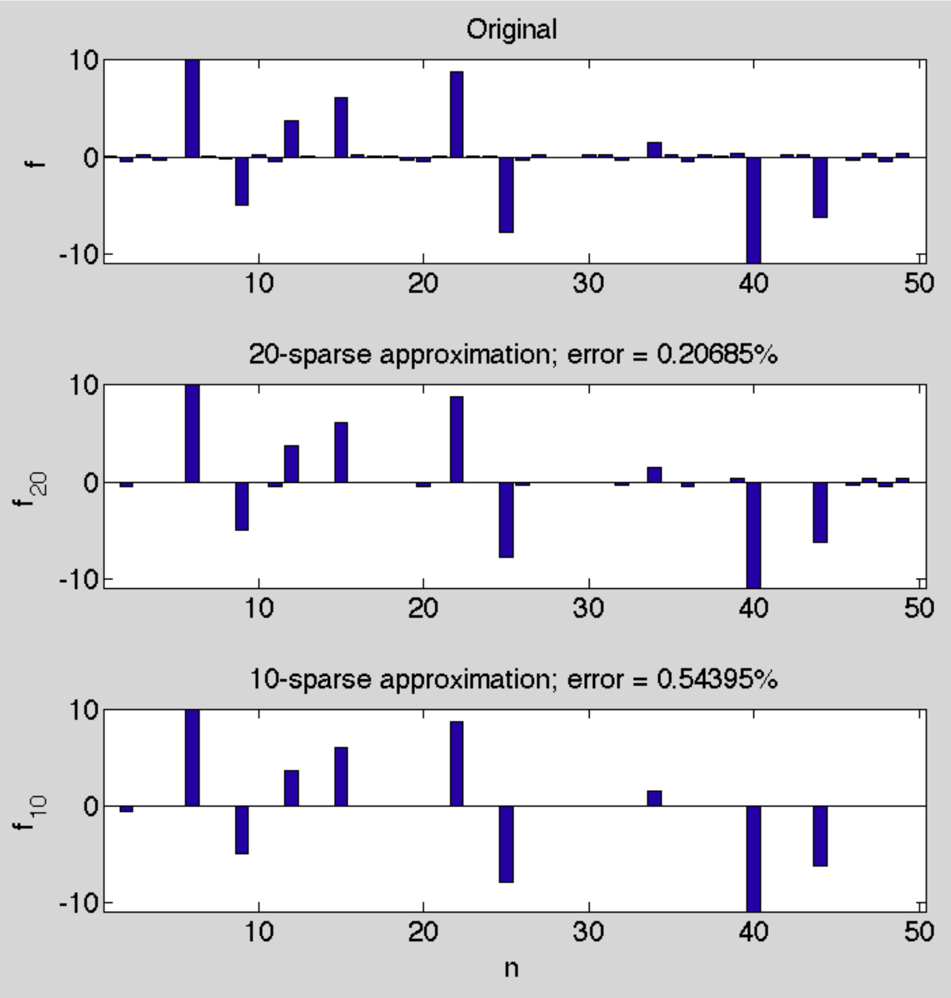
In some cases, our signal is not exactly K-sparse.

However, it may have a K-sparse approximation which is very accurate. We then say the signal is compressible.

Specifically, we can define the K-sparse approximation as follows. Let σ_K be the value of the K^{th} largest (in magnitude) element of f, and set

$$f_{K,i} \triangleq \begin{cases} f_i & |f_i| \geq \sigma_K \\ 0 & \text{otherwise} \end{cases}$$

$$f_K \triangleq [f_{K,1}, \ldots, f_{K,N}]$$

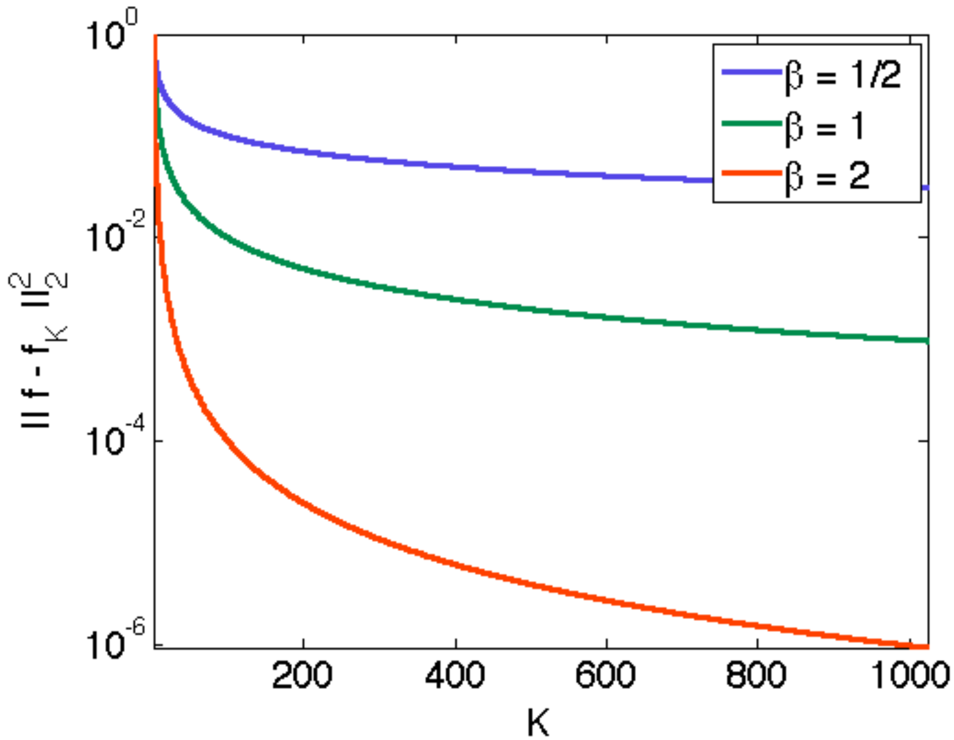


Approximation Error Decay Rate

Ideally, the approximation f_K obeys

$$\frac{\|f - f_K\|_2^2}{N} \equiv \frac{1}{N}\sum_{i=1}^{N}(f_i - f_{K,i})^2 \preceq K^{-\beta}$$

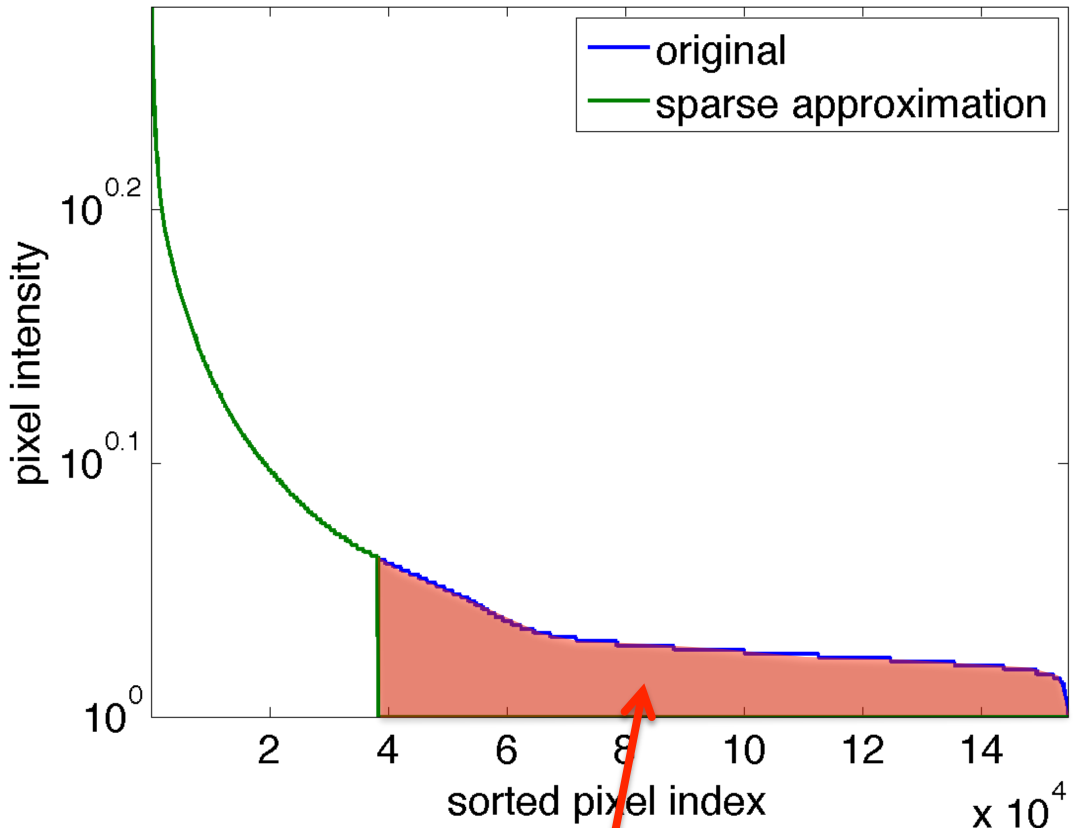
for some $\beta > 0$. This bound tells us how well the sparse signal f_K approximates the original signal f. Bigger β suggests we can get a highly accurate representation of with a very sparse approximation.



APPROXIMATION EXAMPLE

Original

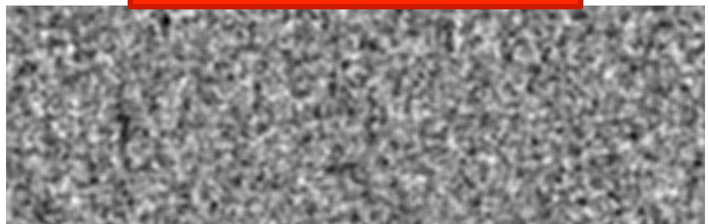
Sparse approximation



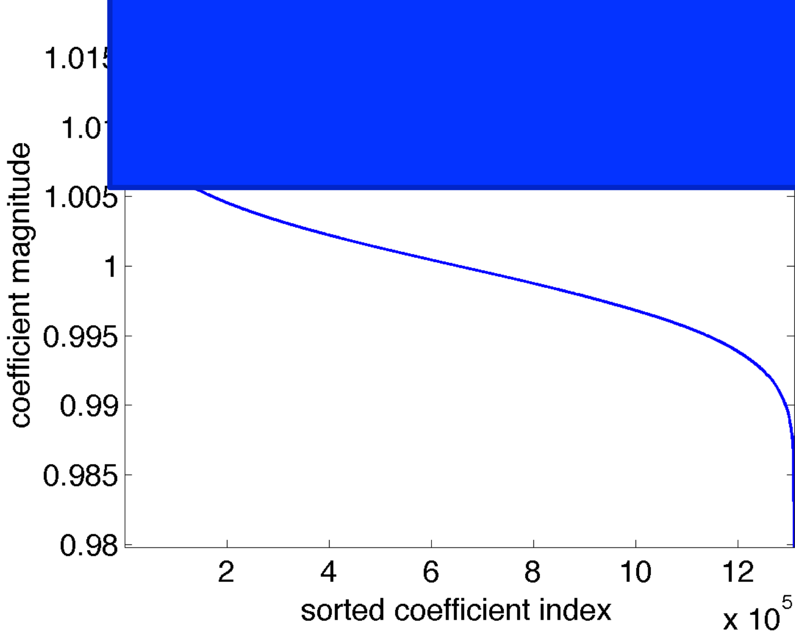
Approximation error

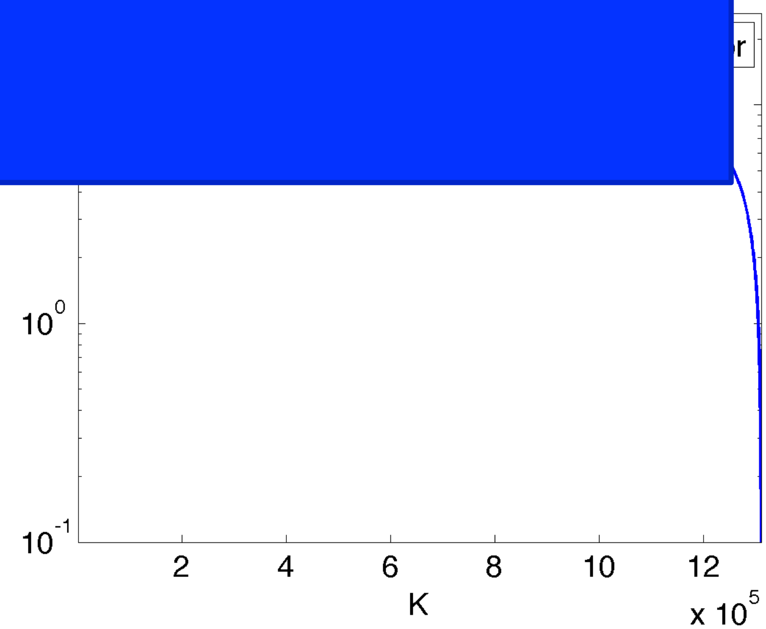
OF COURSE, IN THE REAL WORLD MOST SIGNALS AREN'T *IMMEDIATELY* SPARSE OR COMPRESSIBLE

Original image



Fourier transform



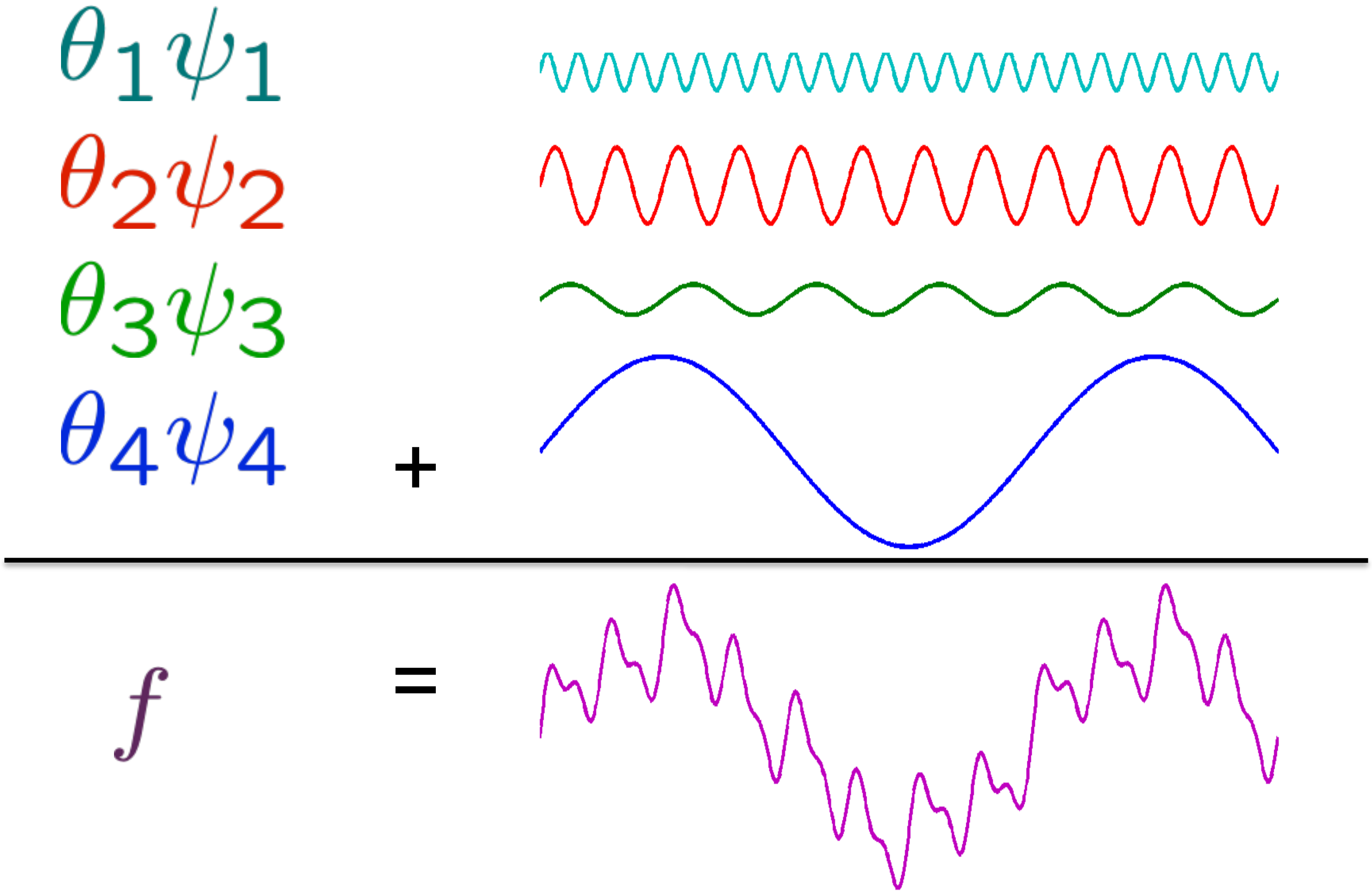


Fourier Transform

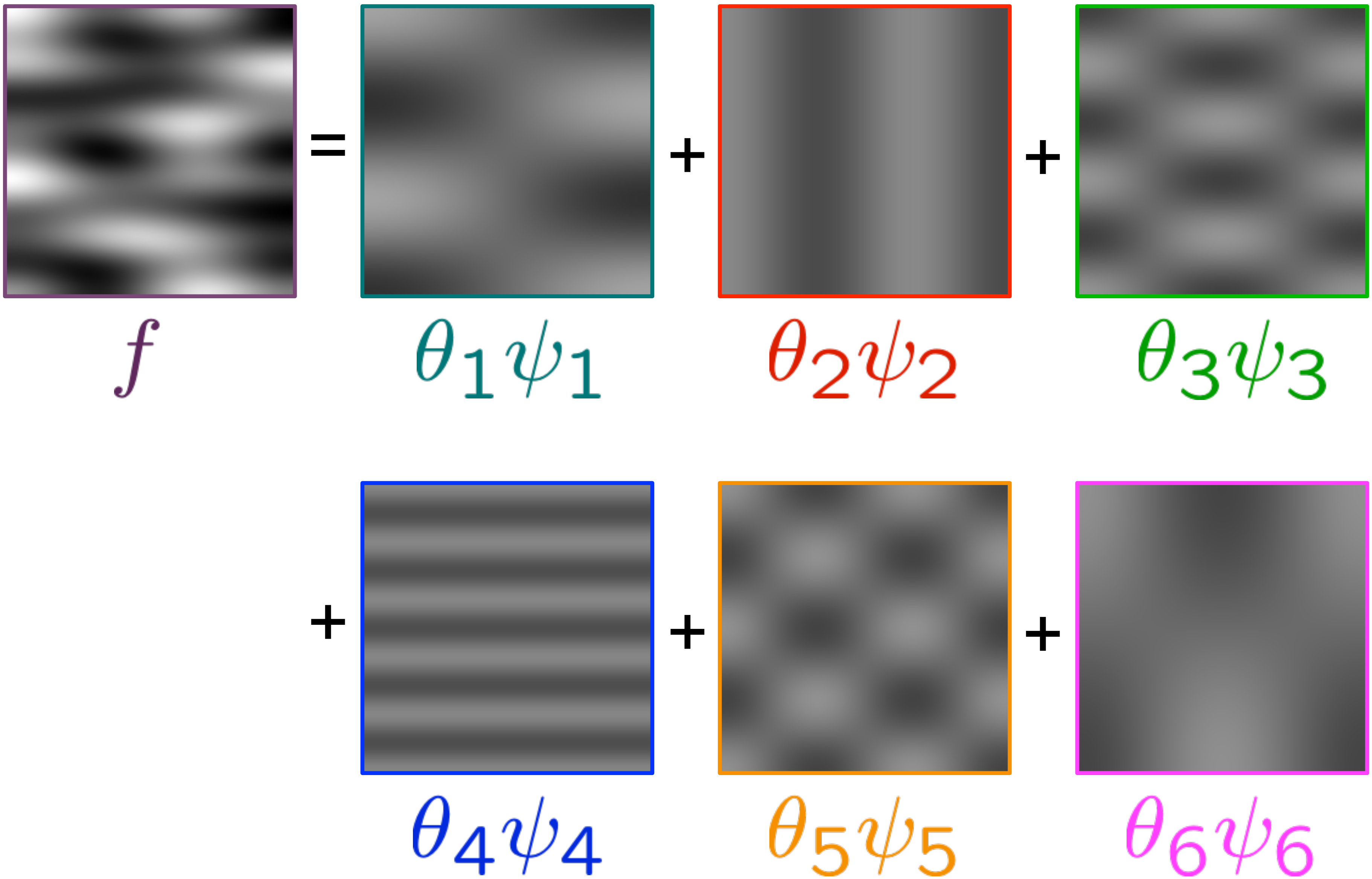
$$f = \Psi\theta = \sum_{i=1}^{N} \theta_i \psi_i$$

- f: Signal
- Ψ: Basis matrix
- θ: Basis coefficients
- θ_i: Weights
- ψ_i: Sinusoidal basis function

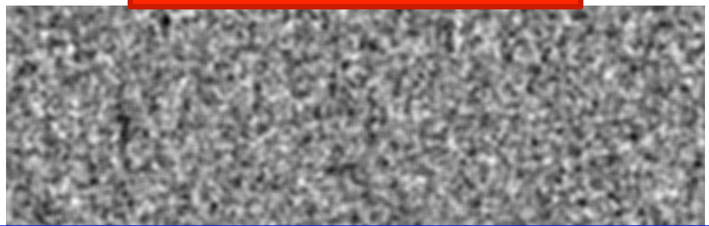
Example: 1-D



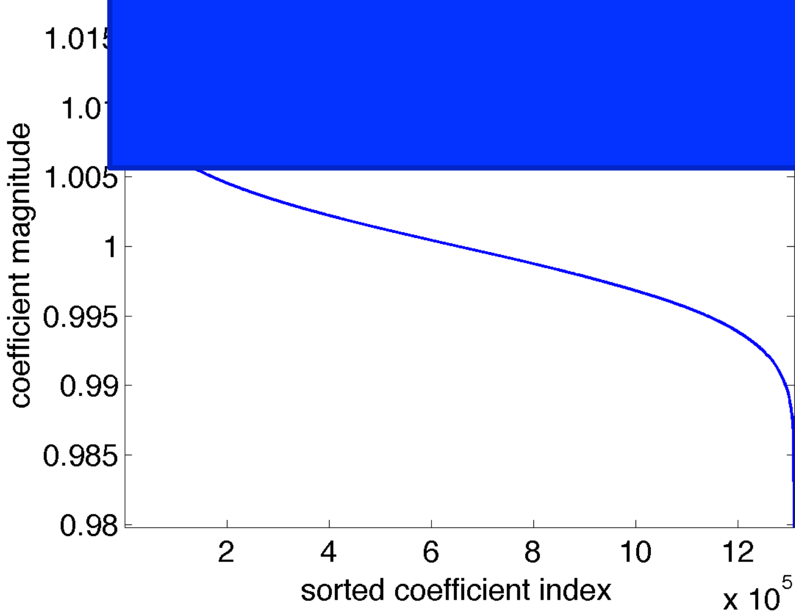
Example: 2-D

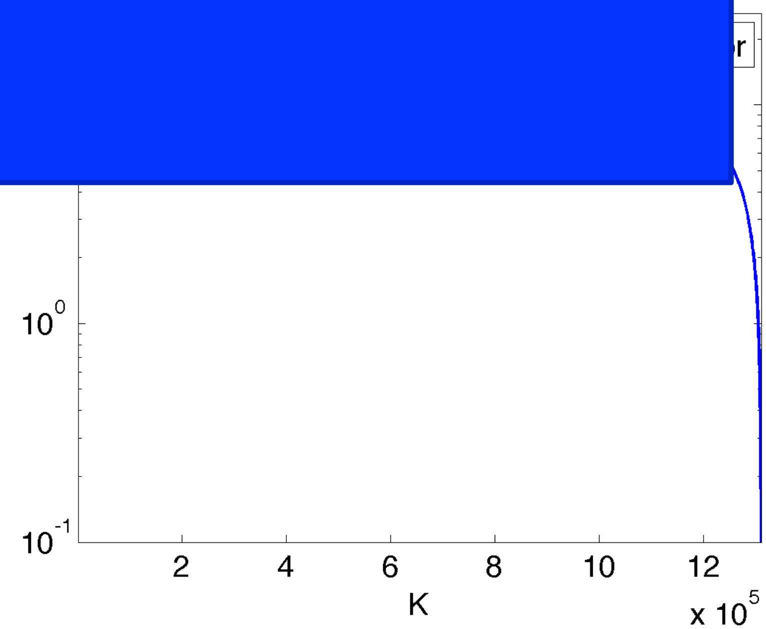

f
=
$\theta_1\psi_1$
+
$\theta_2\psi_2$
+
$\theta_3\psi_3$
+
$\theta_4\psi_4$
+
$\theta_5\psi_5$
+
$\theta_6\psi_6$

Original image

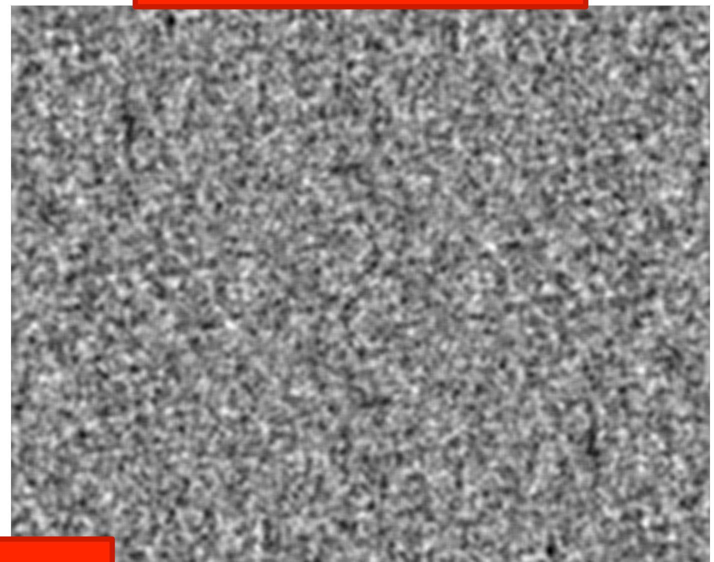


Fourier transform

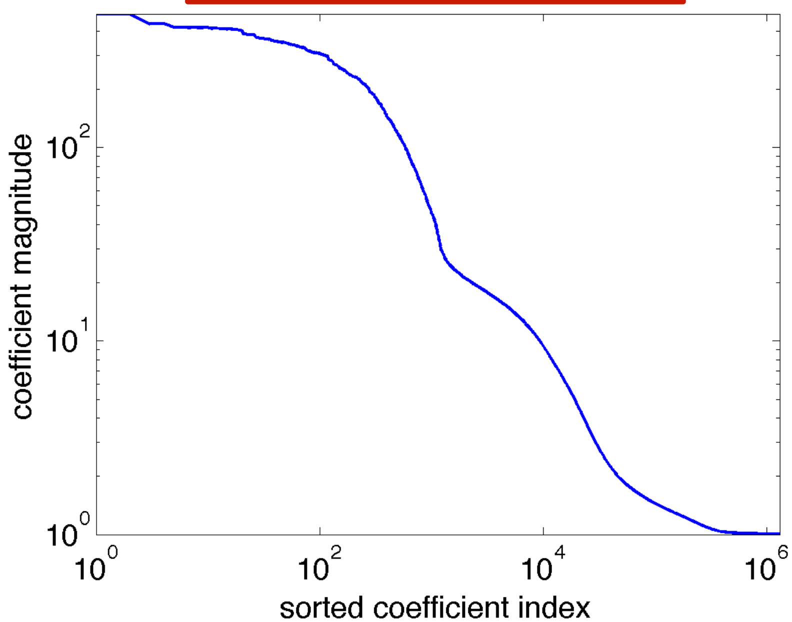




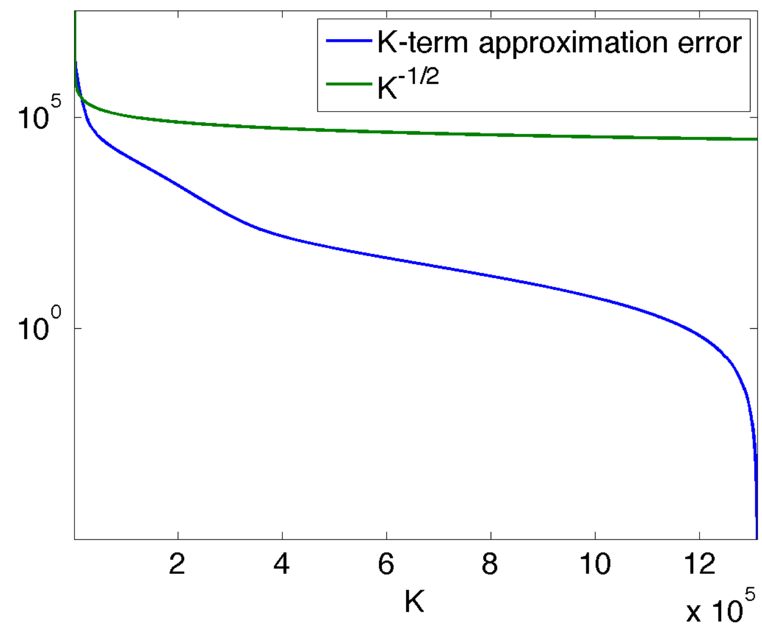
Original image



Sorted Fourier coeff. intensities



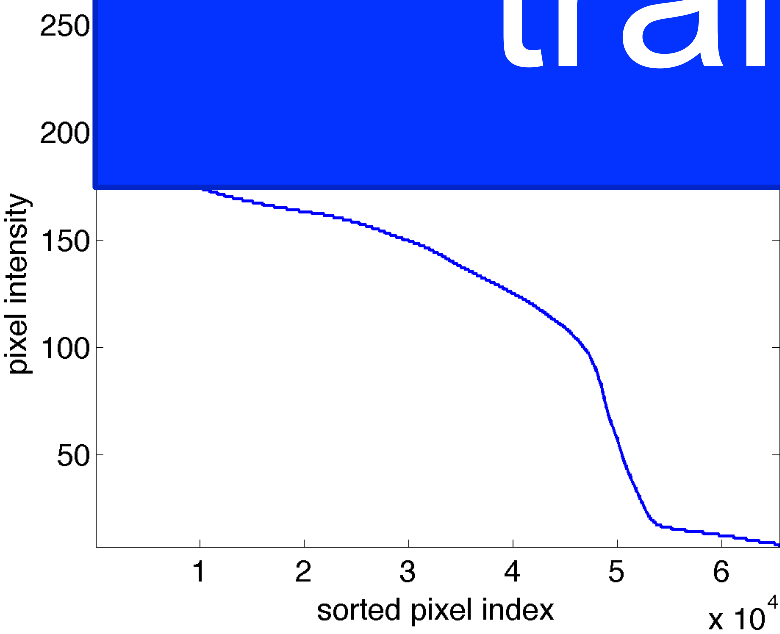
Approximation error decay

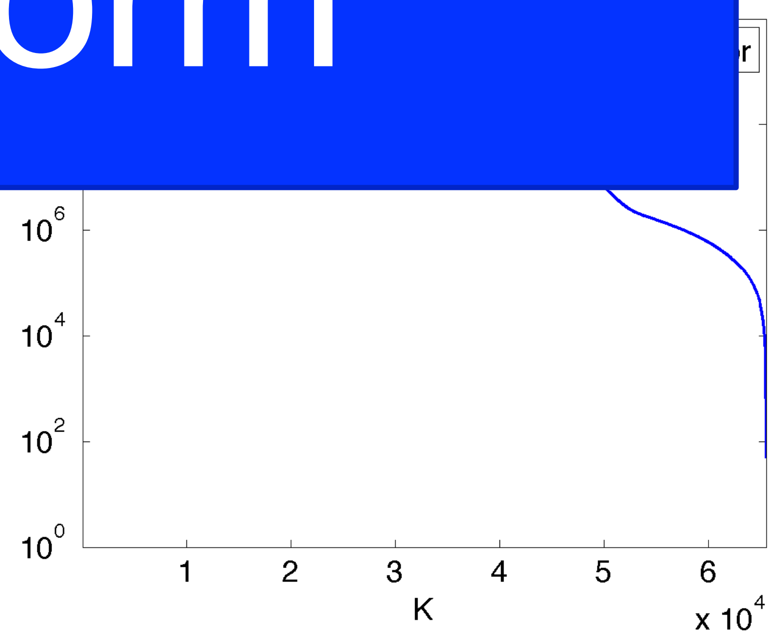


Original image

Wavelet transform

ay





WAVELET TRANSFORM

$$f = \Psi\theta = \sum_{i=1}^{N} \theta_i \psi_i$$

Signal

Basis matrix

Basis coefficients

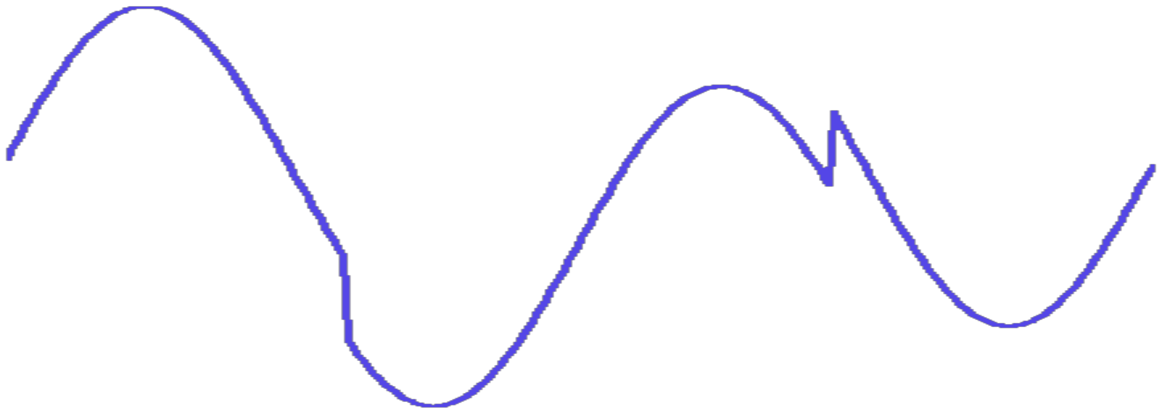
Weights

Wavelet basis function

Just like the Fourier transform, but with different basis functions

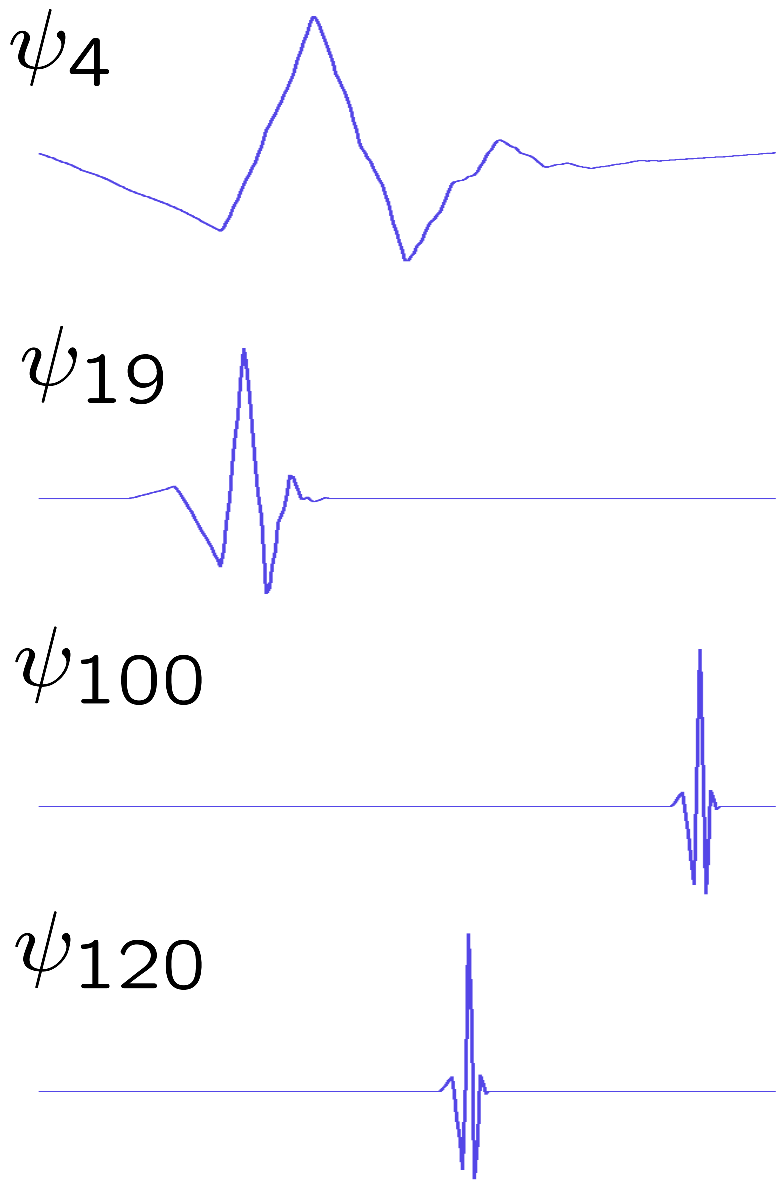
WAVELETS

- The Dirac or canonical basis is restrictive; only a small fraction of signals of interest are sparse here, and it is difficult to model scene structure.
- The Fourier basis if good for smooth signals, but as soon as a single discontinuity is introduced, the signal is no longer sparse in the Fourier basis.
- A wavelet basis gives a sparse representation of piecewise-smooth signals.



Wavelets

- Wavelet basis functions correspond to a single "mother wavelet" at various scales and shifts.
- They form an orthonormal basis.
- They decompose signals into an initial course approximation followed by successive levels of refinement.



Example: 1-D

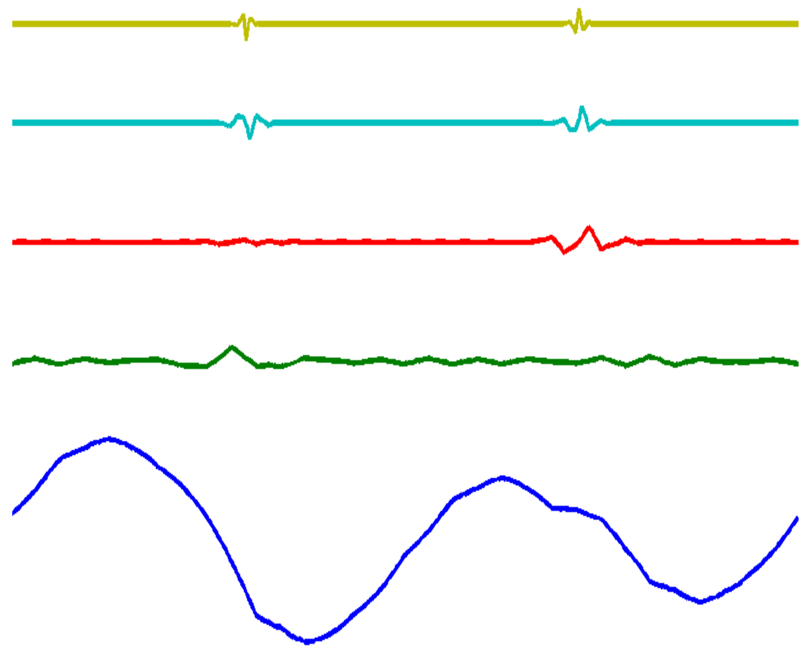
$\sum_{i \in \text{scale}_1} \theta_i \psi_i$

$\sum_{i \in \text{scale}_2} \theta_i \psi_i$

$\sum_{i \in \text{scale}_3} \theta_i \psi_i$

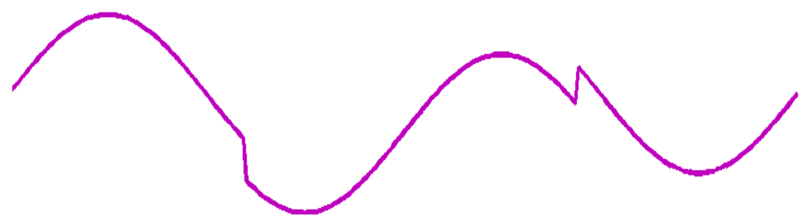
$\sum_{i \in \text{scale}_4} \theta_i \psi_i$

$\sum_{i \in \text{scale}_5} \theta_i \psi_i$ +

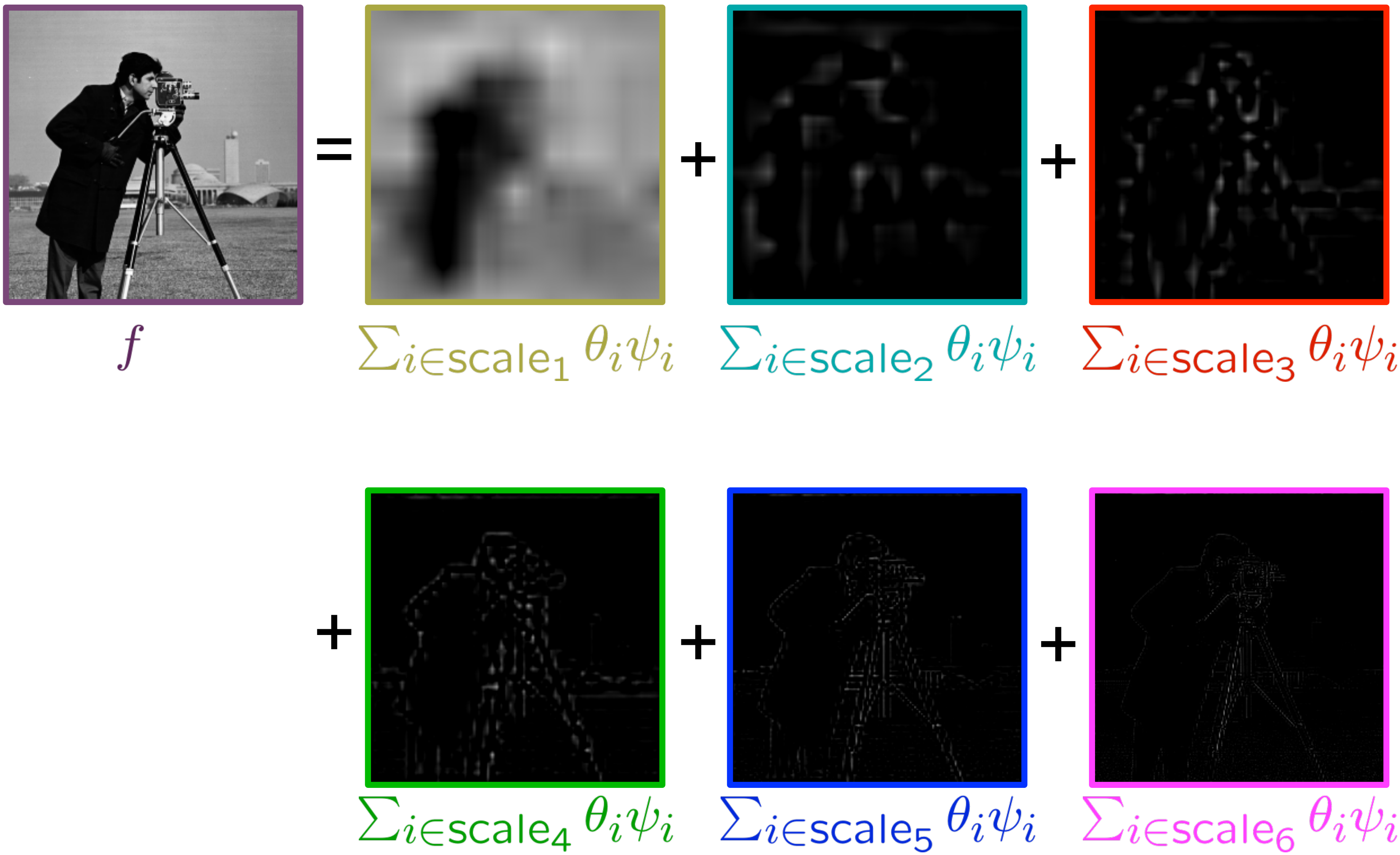


---

f =



EXAMPLE: 2-D

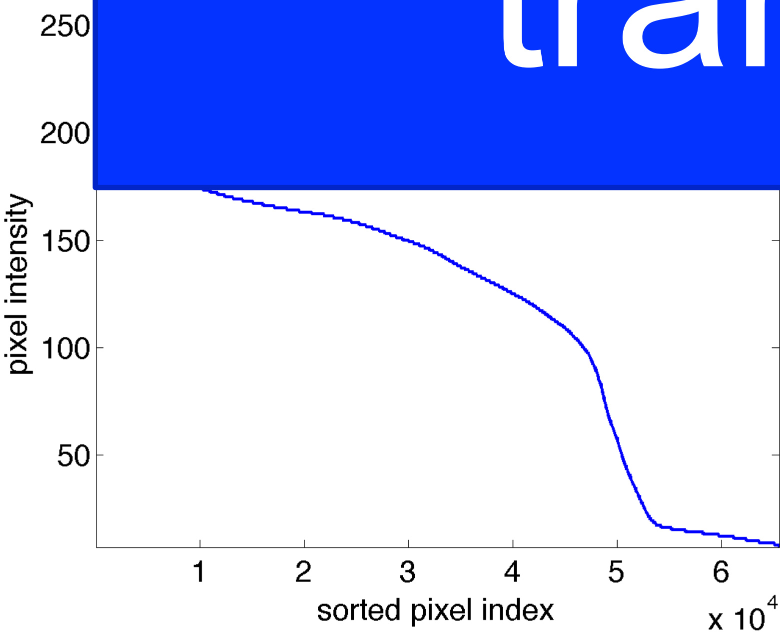


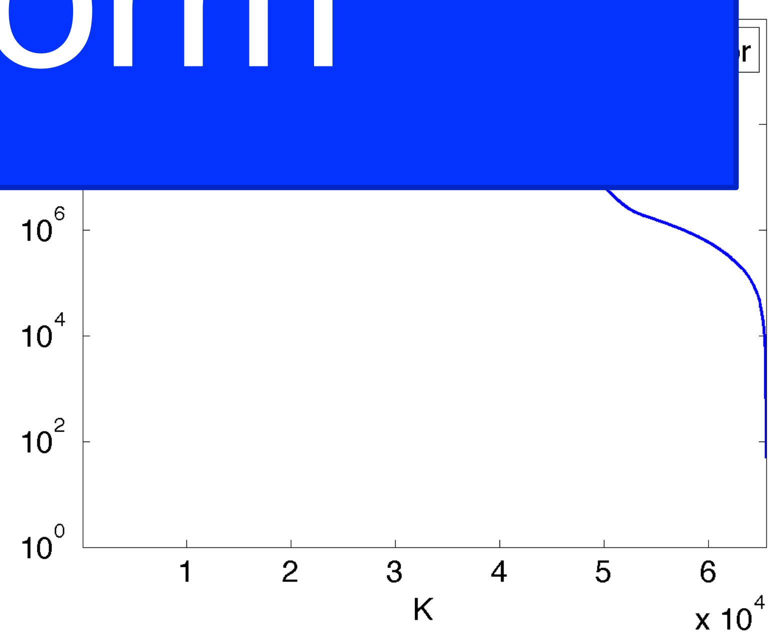
f
=
$\sum_{i \in \text{scale}_1} \theta_i \psi_i$
+
$\sum_{i \in \text{scale}_2} \theta_i \psi_i$
+
$\sum_{i \in \text{scale}_3} \theta_i \psi_i$
+
$\sum_{i \in \text{scale}_4} \theta_i \psi_i$
+
$\sum_{i \in \text{scale}_5} \theta_i \psi_i$
+
$\sum_{i \in \text{scale}_6} \theta_i \psi_i$

Original image

Wavelet transform

ay



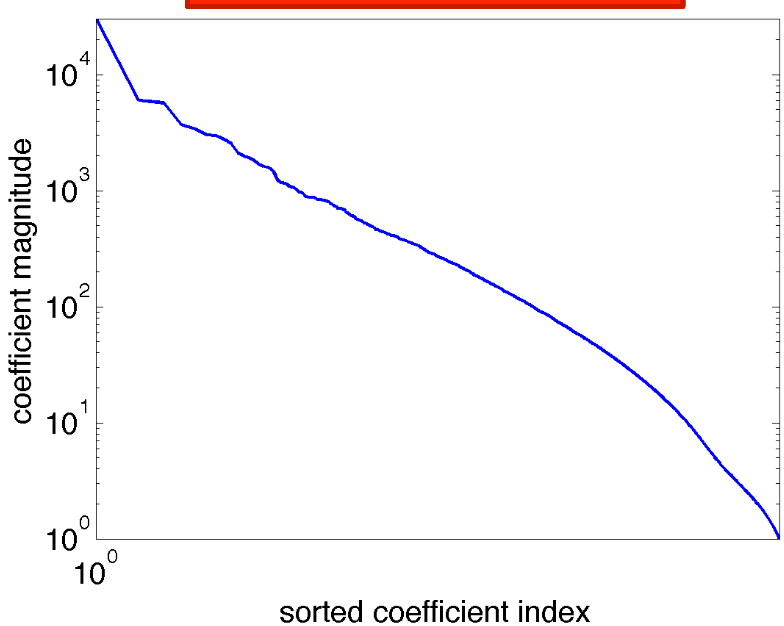


Original image

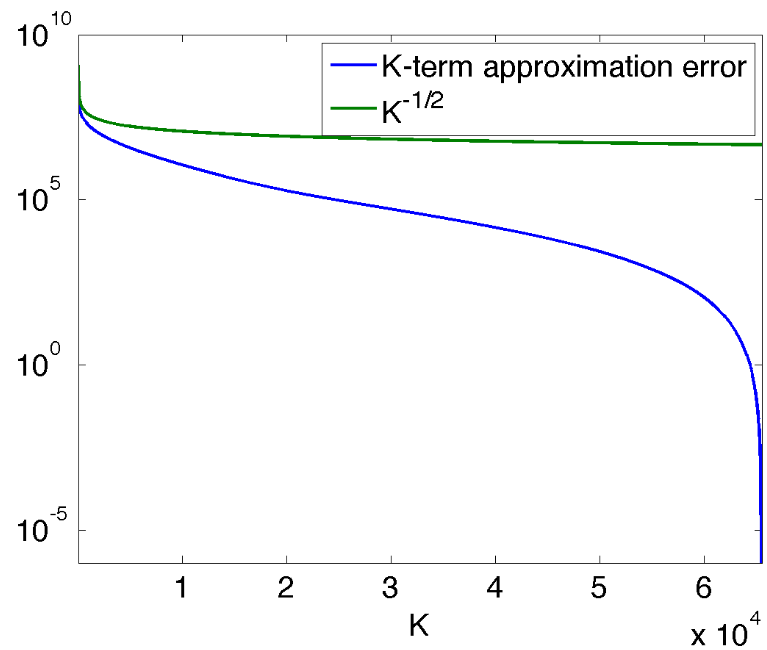
Wavelet coefficients

Original image

Sorted wavelet coeff. intensities



Approximation error decay



Wavelets yield sparse approximations of broad classes of signals and images.

For instance, all functions in a "Besov space" have a sparse wavelet approximation.

COMING NEXT…

- Now that we can represent signals using sparse approximations, how can we use this to solve real-world problems?
- Can we use sparsity to reduce the amount of data we need to collect?
- Can we get better sparse approximations than what we see with Fourier or wavelet bases?