Sparsity in Data Analysis and Computation

Ingrid Daubechies

Abel Symposium, Oslo, August 2012

Sparsity

Sparsity

- emerging role over last $\sim$ 4 decades

Sparsity

- emerging role over last $\sim$ 4 decades
- powerful tool

Sparsity

- emerging role over last $\sim$ 4 decades
- powerful tool
  - for data analysis

Sparsity

- emerging role over last $\sim$ 4 decades
- powerful tool
  - for data analysis
  - for computation

Sparsity

- emerging role over last $\sim$ 4 decades
- powerful tool
  - for data analysis
  - for computation
- better understanding will have enormous impact

Data analysis

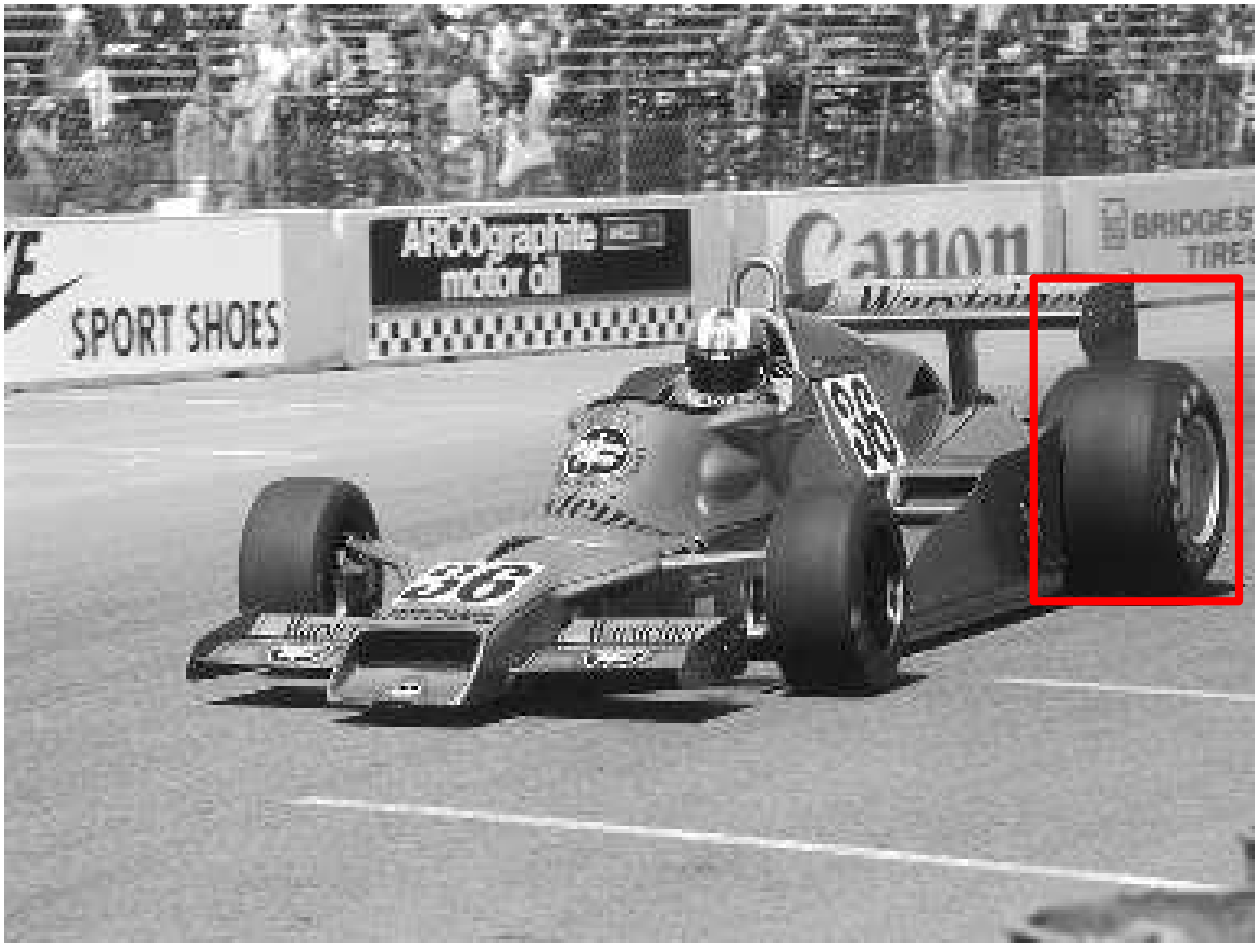
Data analysis: Images

Data analysis: Images

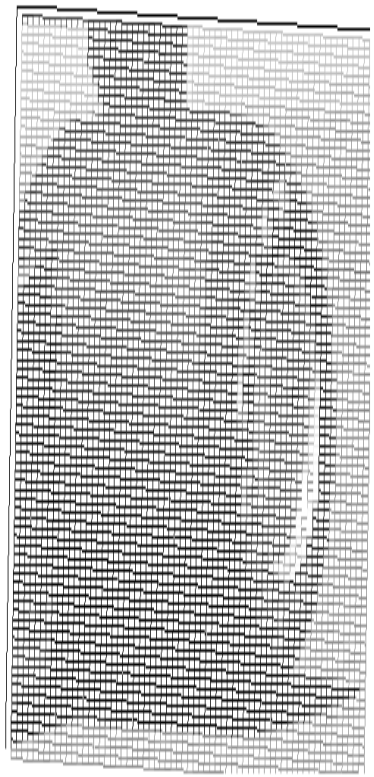
What is an image?

Data analysis: Images

Data analysis: Images

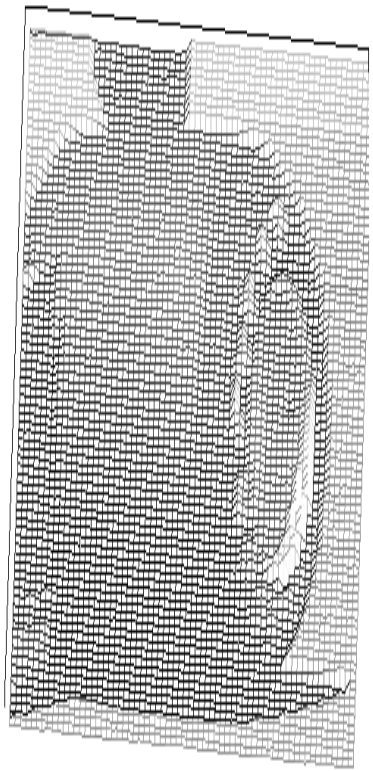


Data analysis: Images



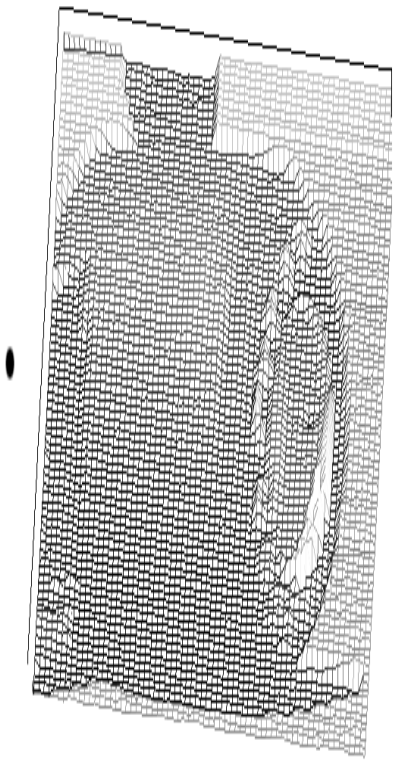
$$f : [a,b] \times [c,d] \longrightarrow \mathbb{R}_+$$

Data analysis: Images



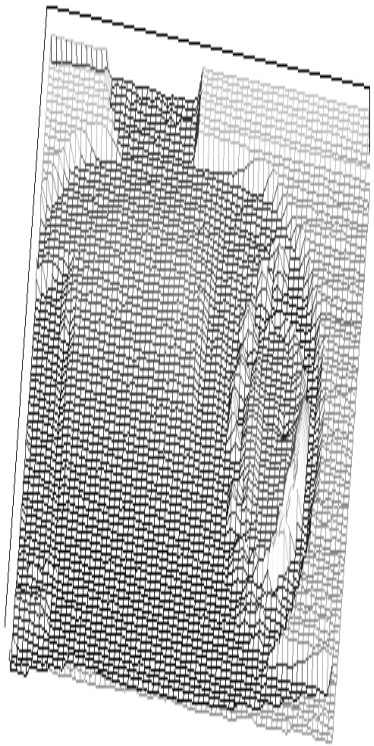
$$f : [a,b] \times [c,d] \longrightarrow \mathbb{R}_+$$

Data analysis: Images



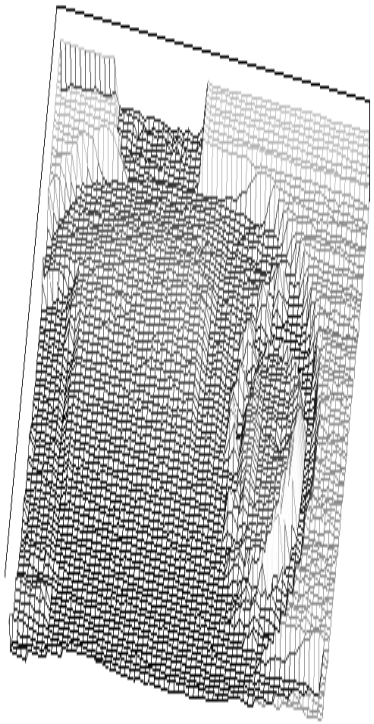
$$f : [a,b] \times [c,d] \longrightarrow \mathbb{R}_+$$

Data analysis: Images



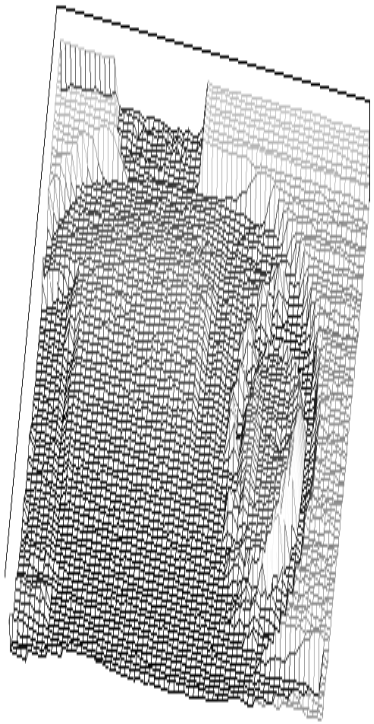
$$f : [a,b] \times [c,d] \longrightarrow \mathbb{R}_+$$

Data analysis: Images



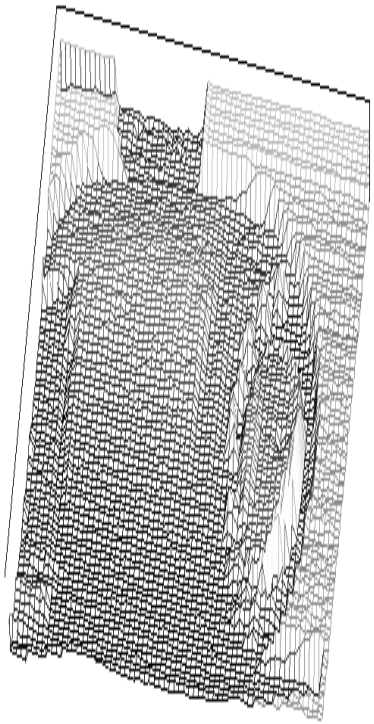
$$f : [a,b] \times [c,d] \longrightarrow \mathbb{R}_+$$

Data analysis: Images



$$f : [a,b] \times [c,d] \longrightarrow [0,1]$$

Data analysis: Images



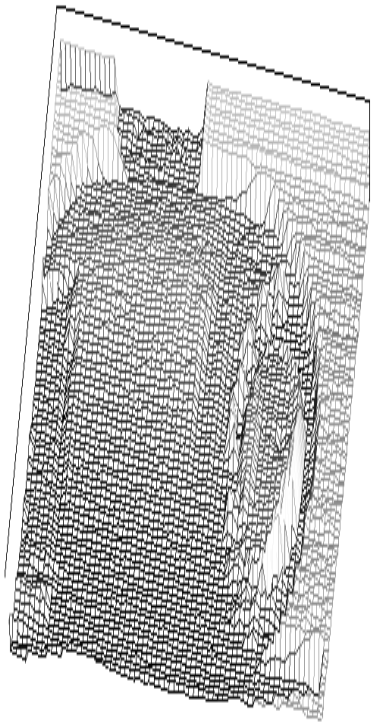
$$f : [a,b] \times [c,d] \longrightarrow [0,1]$$

sampled: $f\left(a + \frac{b-a}{M} m,\ c + \frac{d-c}{N} n\right)$

$m = 0, 1, \ldots, M$

$n = 0, 1, \ldots, N$

Data analysis: Images



$$f : [a,b] \times [c,d] \longrightarrow [0,1]$$

sampled: $f\left(a + \frac{b-a}{M} m,\ c + \frac{d-c}{N} n\right)$

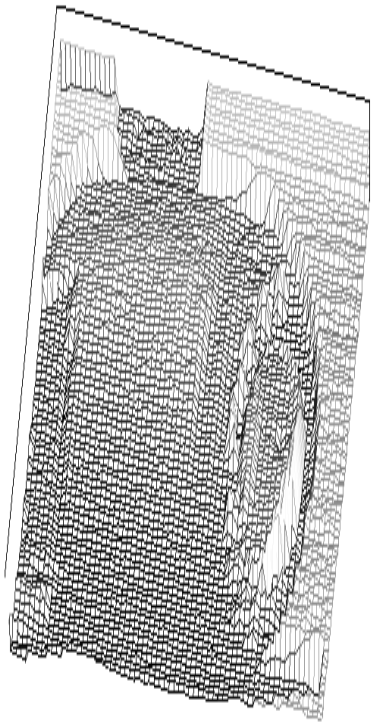
$m = 0, 1, \ldots, M$

$n = 0, 1, \ldots, N$

or averaged:

$$\int_{|s - \frac{b-a}{M} m| < \frac{b-a}{2M}} \int_{|t - \frac{d-c}{N}| < \frac{d-c}{2N}} f(s,t)\, ds\, dt$$

Data analysis: Images



$$f : [a,b] \times [c,d] \longrightarrow [0,1]$$

$$F : \{0,\dots,M\} \times \{0,\dots N\} \longrightarrow [0,1]$$

sampled: $f\left(a + \frac{b-a}{M} m,\ c + \frac{d-c}{N} n\right)$

$m = 0, 1, \dots, M$

$n = 0, 1, \dots, N$

or averaged:

$$\int_{|s - \frac{b-a}{M} m| < \frac{b-a}{2M}} \int_{|t - \frac{d-c}{N}| < \frac{d-c}{2N}} f(s,t)\, ds\, dt$$

Data analysis: Images

In any case : object in high-dimensional space.

Data analysis: Images

In any case : object in high-dimensional space.

For practical purposes : need compression

Data analysis: Images

In any case : object in high-dimensional space.

For practical purposes : need compression

storage

Data analysis: Images

In any case : object in high-dimensional space.

For practical purposes : need compression

storage , transmission

Data analysis: Images

In any case : object in high-dimensional space.

For practical purposes : need compression

storage, transmission, analysis

Data analysis: Images

In any case : object in high-dimensional space.

For practical purposes : need compression

How ?

Data analysis: Images

In any case : object in high-dimensional space.

For practical purposes : need compression

How ?

→ exploit mathematical properties of the class

Data analysis: Images

In any case : object in high-dimensional space.

For practical purposes : need compression

How ?

↳ exploit mathematical properties of the class

Translation invariance !

Data analysis: Images

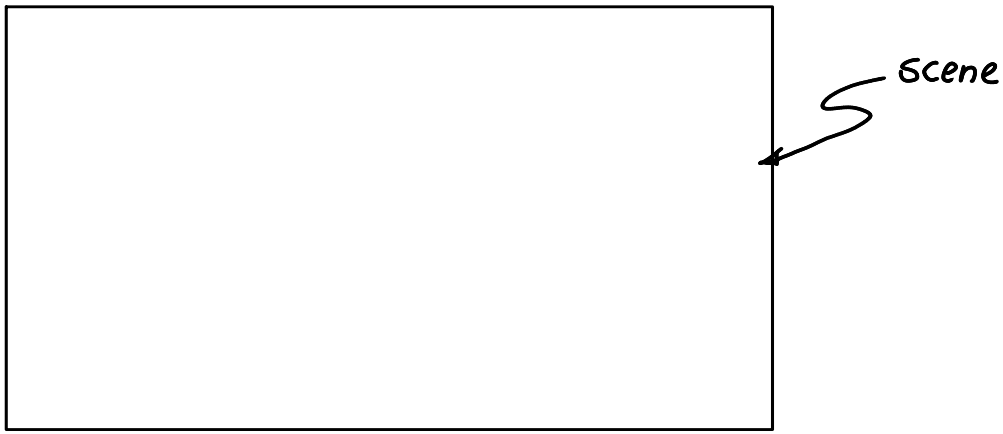
In any case : object in high-dimensional space.

For practical purposes : need compression

How ?

→ exploit mathematical properties of the class

Translation invariance !



Data analysis: Images

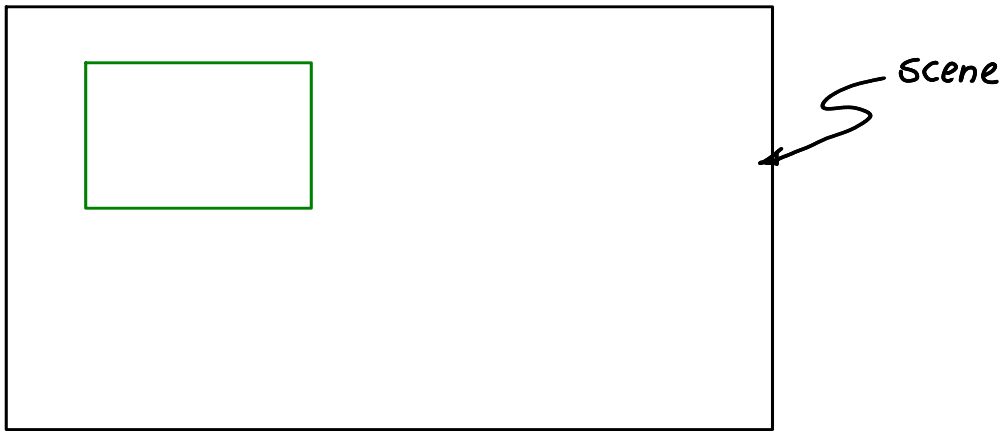
In any case : object in high-dimensional space.

For practical purposes : need compression

How ?

→ exploit mathematical properties of the class

Translation invariance !



Data analysis: Images

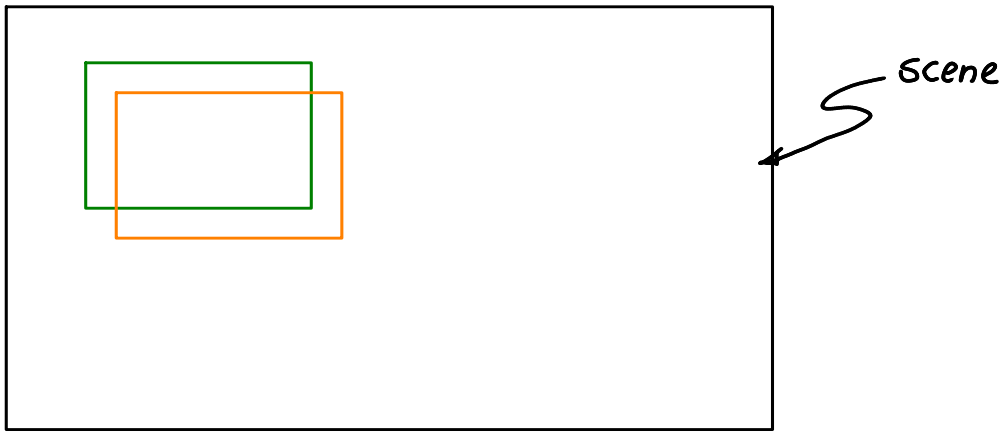
In any case : object in high-dimensional space.

For practical purposes : need compression

How ?

→ exploit mathematical properties of the class

Translation invariance !



Data analysis: Images

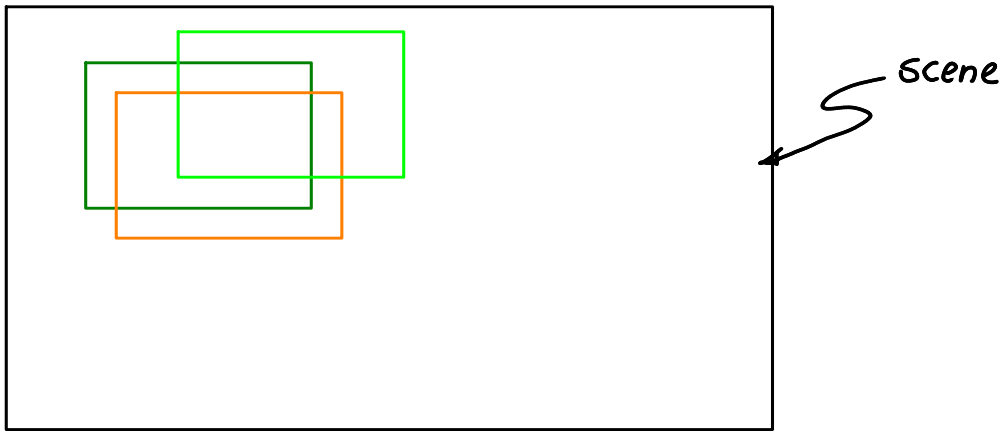
In any case : object in high-dimensional space.

For practical purposes : need compression

How ?

→ exploit mathematical properties of the class

Translation invariance !



Data analysis: Images

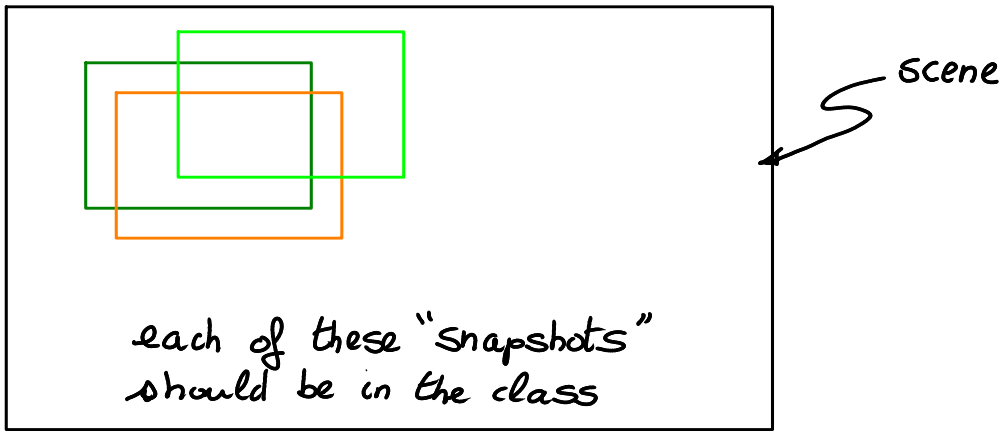
In any case : object in high-dimensional space.

For practical purposes : need compression

How ?

→ exploit mathematical properties of the class

Translation invariance !



Data analysis: Images

Invariance under Translation Group

Data analysis: Images

Invariance under Translation Group

$\Longrightarrow$ use irreducible representations to decompose the class

Data analysis: Images

Invariance under Translation Group

$\Longrightarrow$ use irreducible representations to decompose the class

$\Longrightarrow$ Fourier analysis

Data analysis: Images

Invariance under Translation Group

$\Longrightarrow$ use irreducible representations to decompose the class

$\Longrightarrow$ Fourier analysis

Indeed:

Data analysis: Images

Invariance under Translation Group

$\Longrightarrow$ use irreducible representations to decompose the class

$\Longrightarrow$ Fourier analysis

Indeed: JPEG standard for image compression uses DCT (discrete cosine transform)

Data analysis: Images

Invariance under Translation Group

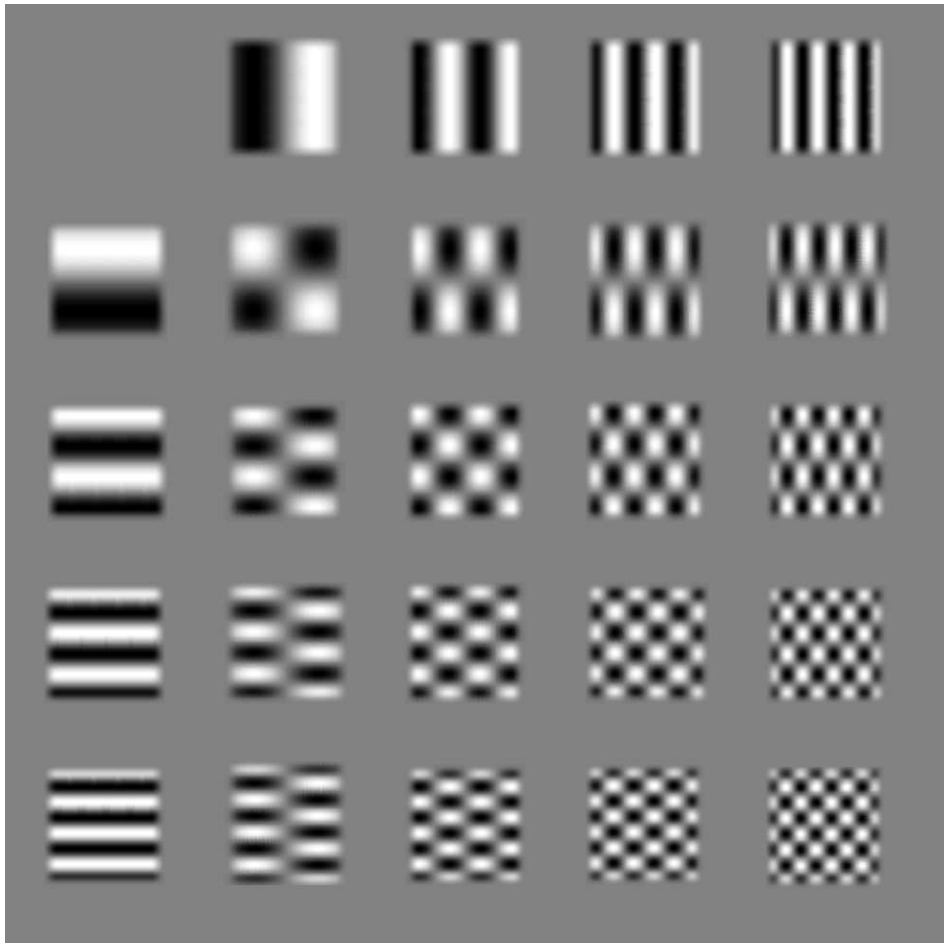
$\Longrightarrow$ use irreducible representations to decompose the class

$\Longrightarrow$ Fourier analysis!

Indeed: JPEG standard for image compression uses DCT

(discrete cosine transform)

Data analysis: Images



Data analysis: Images

JPEG standard :

uses DCT on 8×8 blocks

Data analysis: Images

JPEG standard :

uses DCT on 8 x 8 blocks

technical reasons

Data analysis: Images

JPEG standard :

uses DCT on 8 x 8 blocks

technical reasons

in early 80s: → 16x16

expected to go to even larger...

Data analysis: Images

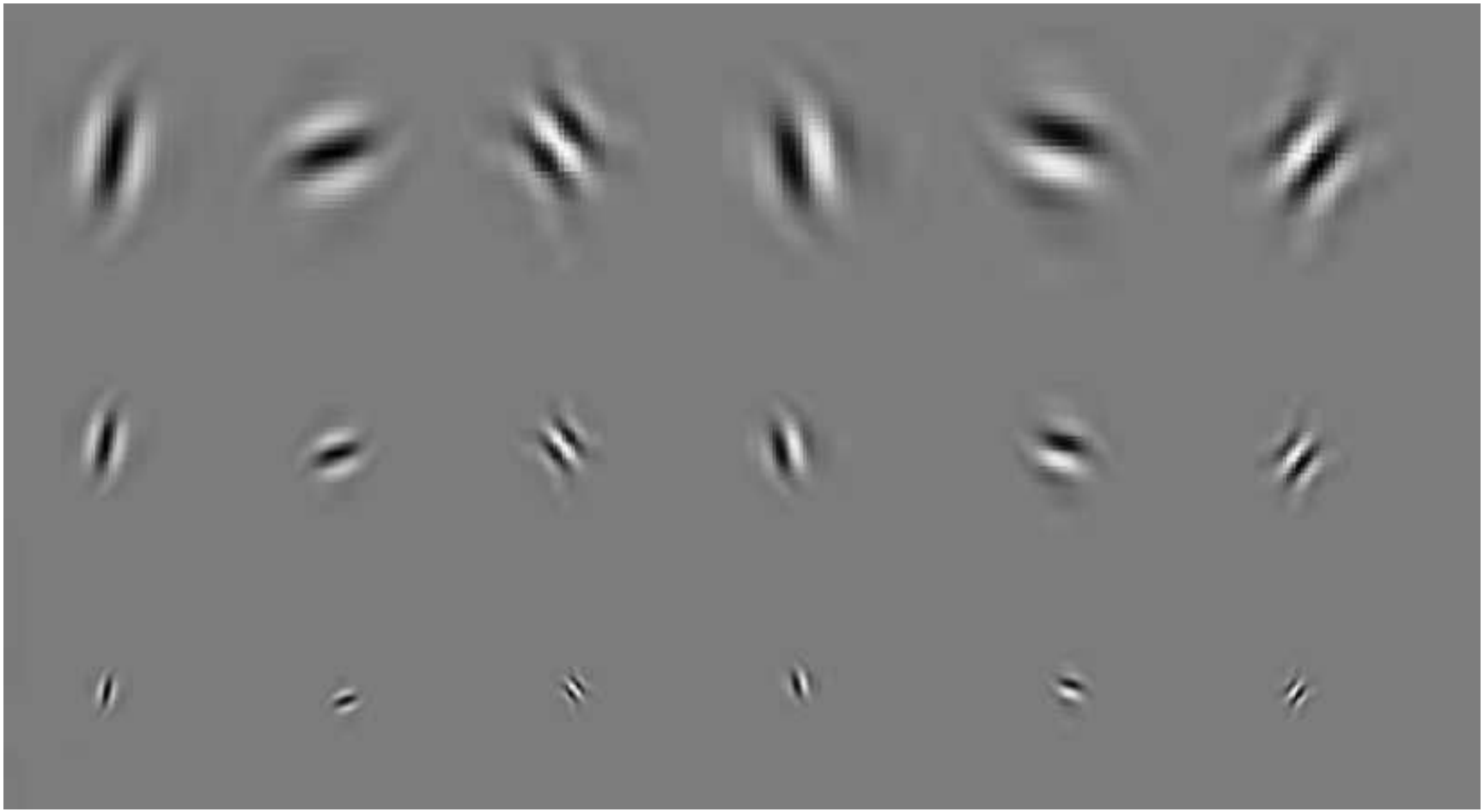
In 1980s : start of use of wavelet transform for images.

Data analysis: Images

In 1980s : start of use of wavelet transform for images.

$\longrightarrow$ decomposition of images into different type of building blocks.

Data analysis: Images



Data analysis: Images

Wavelets

Data analysis: Images

Wavelets

. high frequency wavelets much more "narrow" than low frequency wavelets

Data analysis: Images

Wavelets

- high frequency wavelets much more "narrow" than low frequency wavelets

⇒ need many more fine scale wavelets to cover the image domain than coarse scale wavelets

Data analysis: Images

Wavelets

. high frequency wavelets much more "narrow" than low frequency wavelets

⇒ need many more fine scale wavelets to cover the image domain than coarse scale wavelets

⇒ traditional representation of wavelet decompositions of an image.

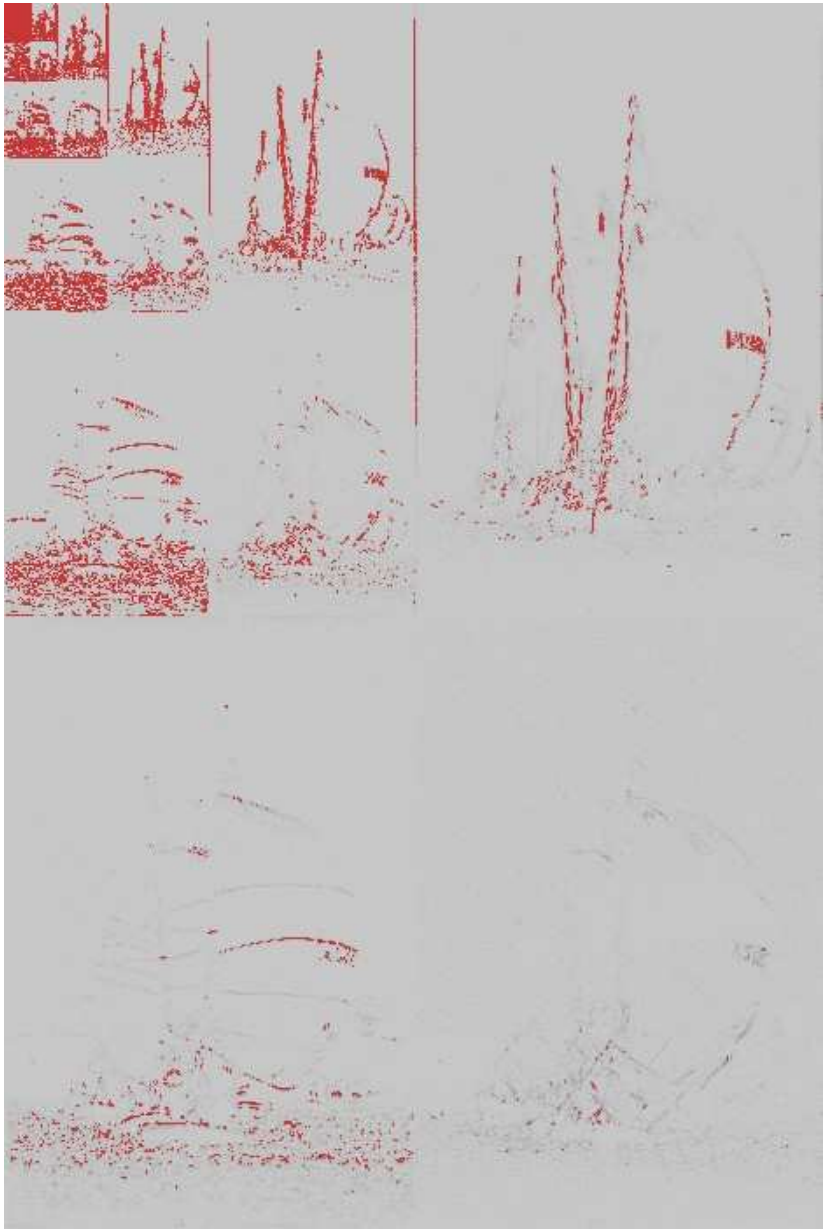

14255


14255

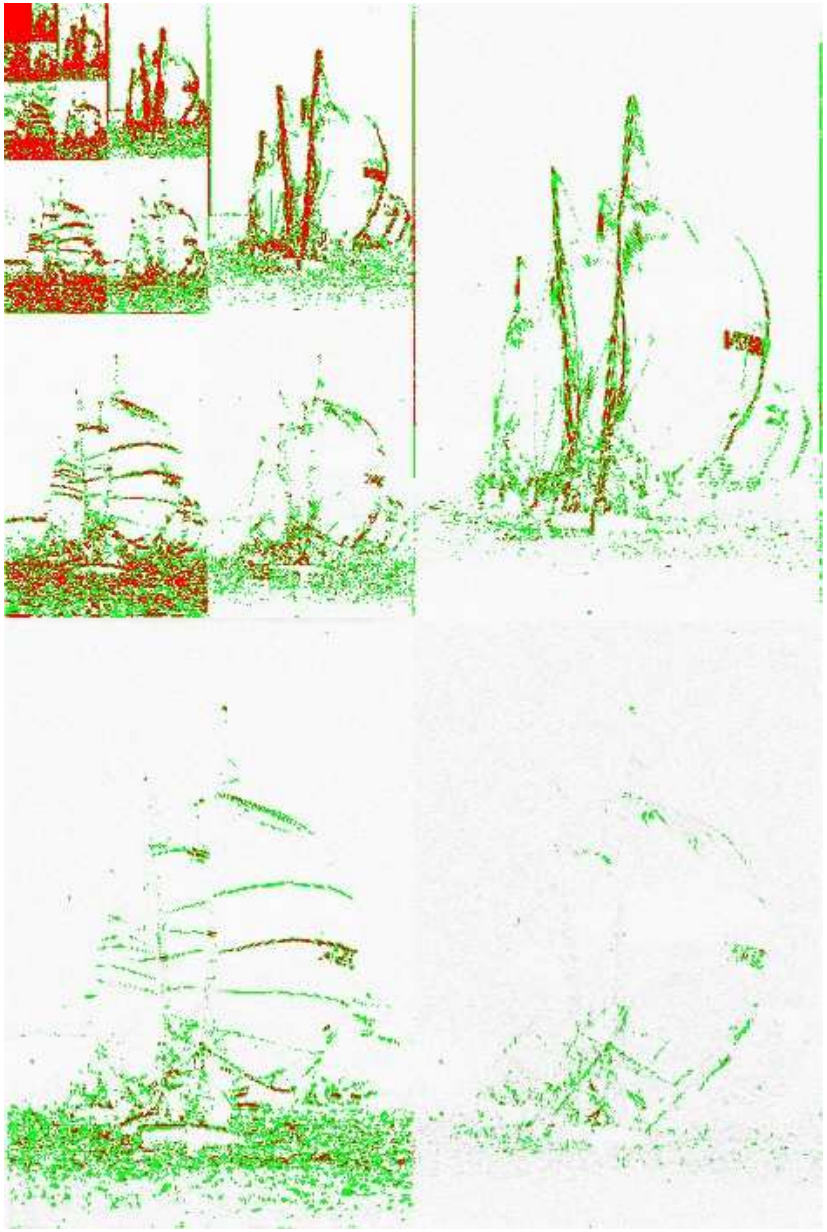

14255

Compression


14255



Compression Ratio: 3.3%



Compression Ratio: 10%

Data analysis: Images

In JPEG-2000 image standard:

wavelets instead of DCT.

Data analysis: Images

In JPEG-2000 image standard:

wavelets instead of DCT.

major reasons:

Data analysis: Images

In JPEG-2000 image standard:

wavelets instead of DCT.

major reasons: . graceful degradation as rate drops

Data analysis: Images

In JPEG-2000 image standard:

wavelets instead of DCT.

major reasons:
- graceful degradation as rate drops
- ease of implementing lossy/lossless compr.

Data analysis: Images

In JPEG-2000 image standard:

wavelets instead of DCT.

major reasons:
- graceful degradation as rate drops
- ease of implementing lossy/lossless compr.

impact:

Data analysis: Images

In JPEG-2000 image standard:

wavelets instead of DCT.

major reasons: . graceful degradation as rate drops
. ease of implementing lossy/lossless compr.

impact: . none really on consumer products

Data analysis: Images

In JPEG-2000 image standard:

wavelets instead of DCT.

major reasons: . graceful degradation as rate drops
. ease of implementing lossy/lossless compr.

impact: . none really on consumer products
. digital movies, sports reporting

Data analysis: Images

Why are wavelets a good idea for images?

What was "wrong" with the Fourier analysis argument?

Really the difference between

Linear and Non-linear

approximation.

Data analysis: Images

Consider a simple class of functions on $\mathbb{T}$

$\mathcal{C}$ $\quad f \in \mathcal{C}$

$f : \mathbb{T} \longrightarrow \mathbb{C}$ "nice"

on $\mathcal{C}$: probability measure

invariant under translations

Then one can prove that the "best" basis in which the $f \in \mathcal{C}$ can be decomposed is the Fourier basis

Namely:

If one wants to find the basis $\varphi_1, \varphi_2, \dots \varphi_n, \dots$ of functions such that

$$\mathbb{E}\left(\int_{\mathbb{T}} \left| f(t) - \langle f, \varphi_1 \rangle \varphi_1(t) \right|^2 dt \right)$$

$$\vdots$$

$$\mathbb{E}\left(\int_{\mathbb{T}} \left| f(t) - \sum_{n=1}^{N} \langle f, \varphi_n \rangle \varphi_n(t) \right|^2 dt \right)$$

$$\vdots$$

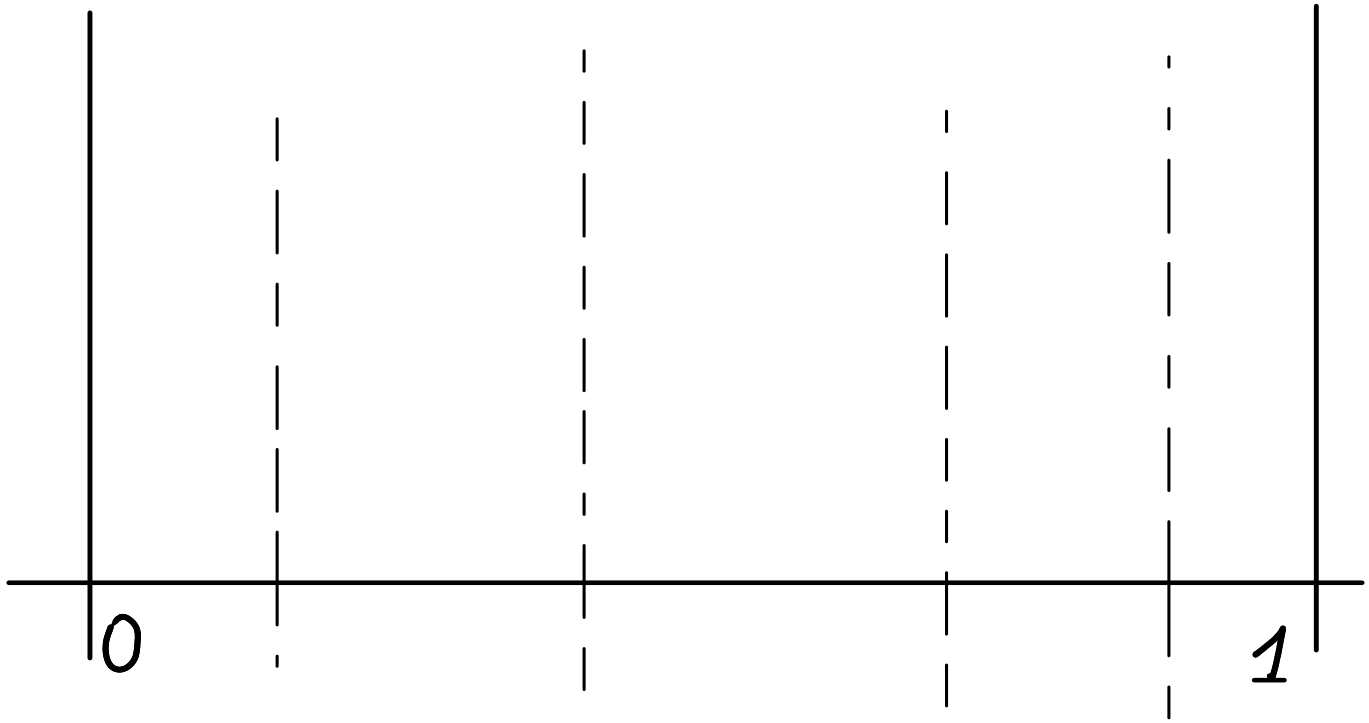
are minimal, then these must be the Fourier exponentials $e^{2\pi i n t}$

Data analysis: Images

However, consider the following example:

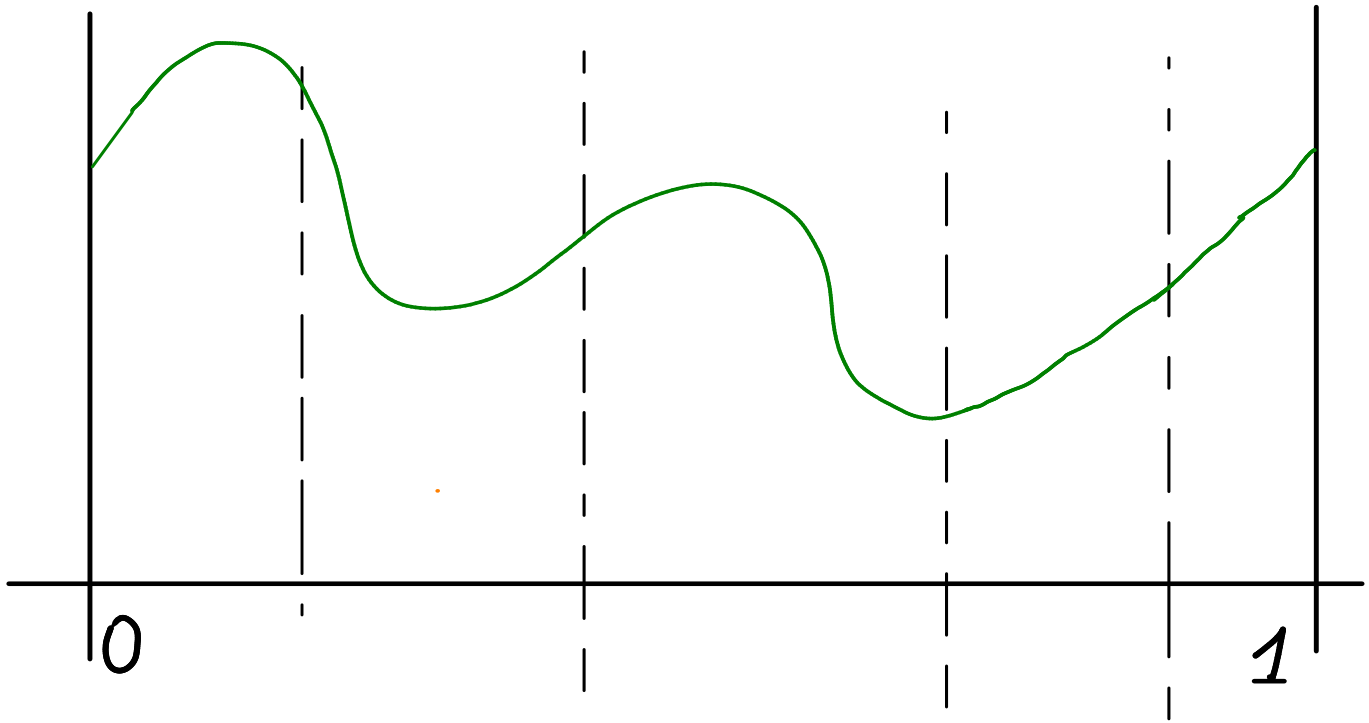
Data analysis: Images

However, consider the following example:



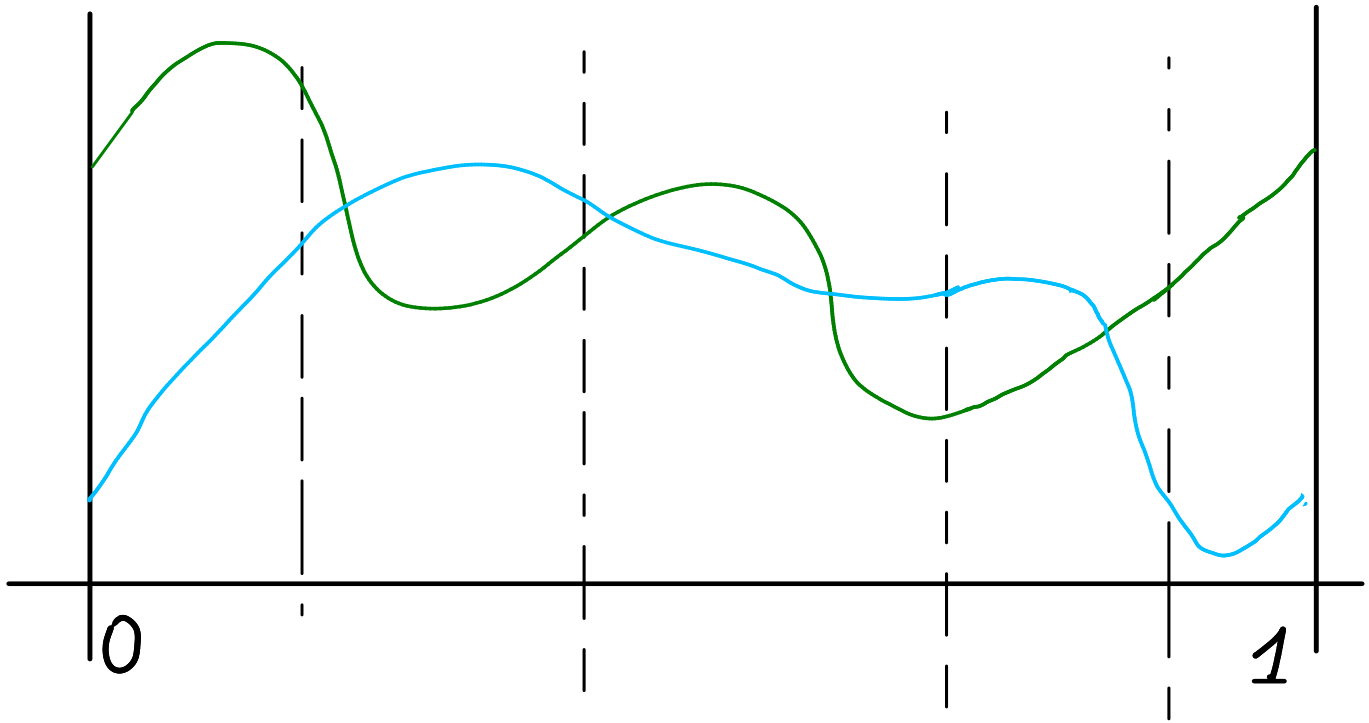
Data analysis: Images

However, consider the following example:



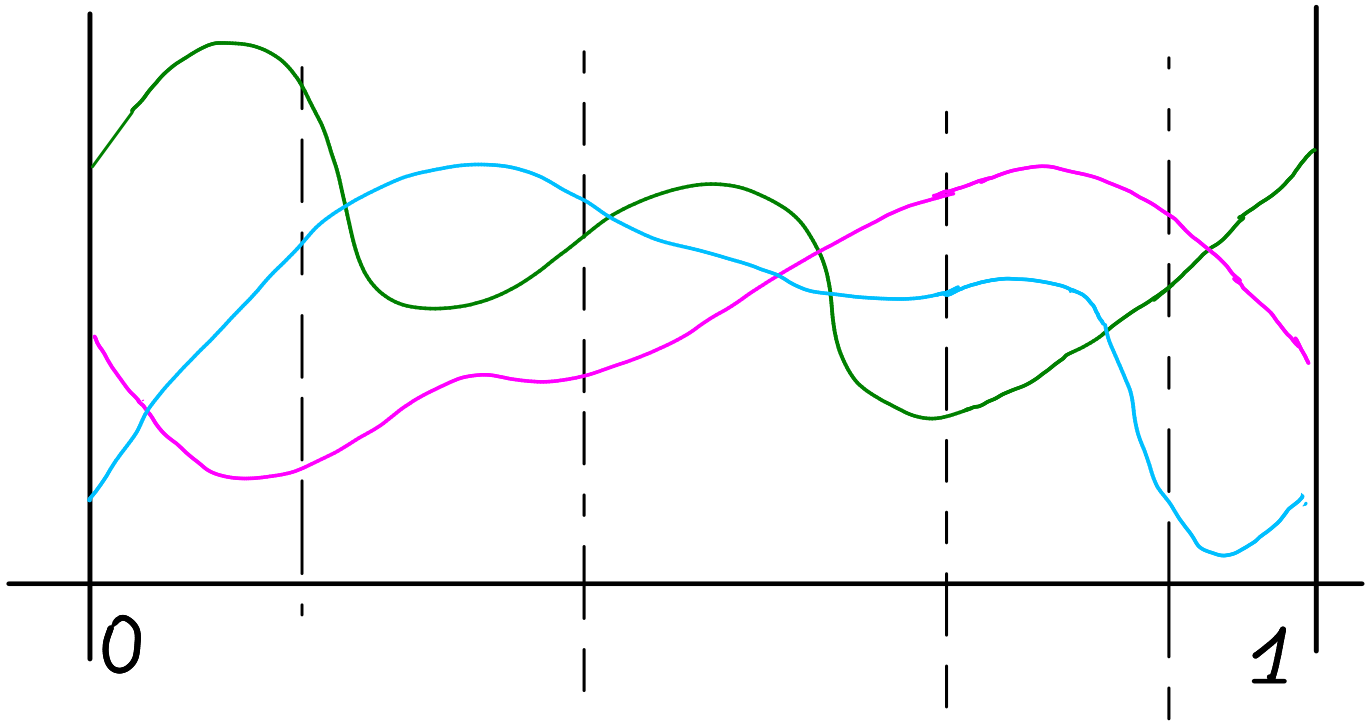
Data analysis: Images

However, consider the following example:



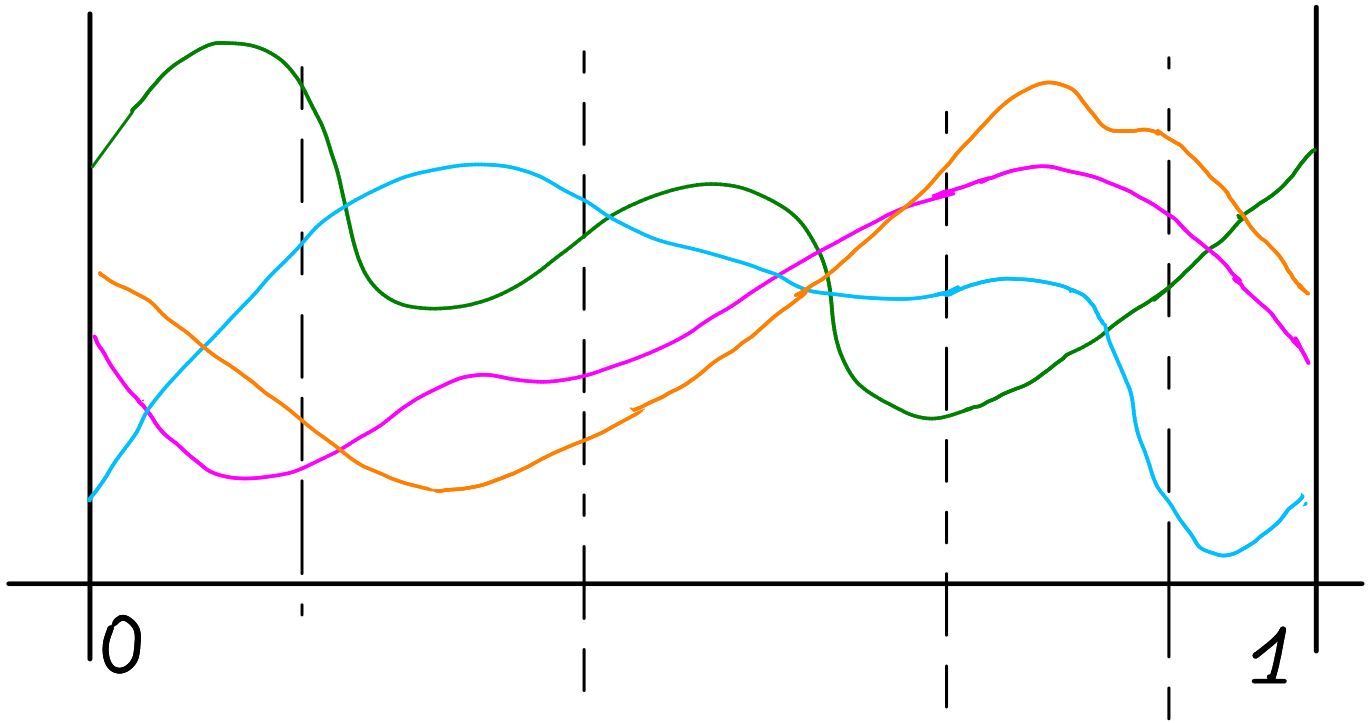
Data analysis: Images

However, consider the following example:

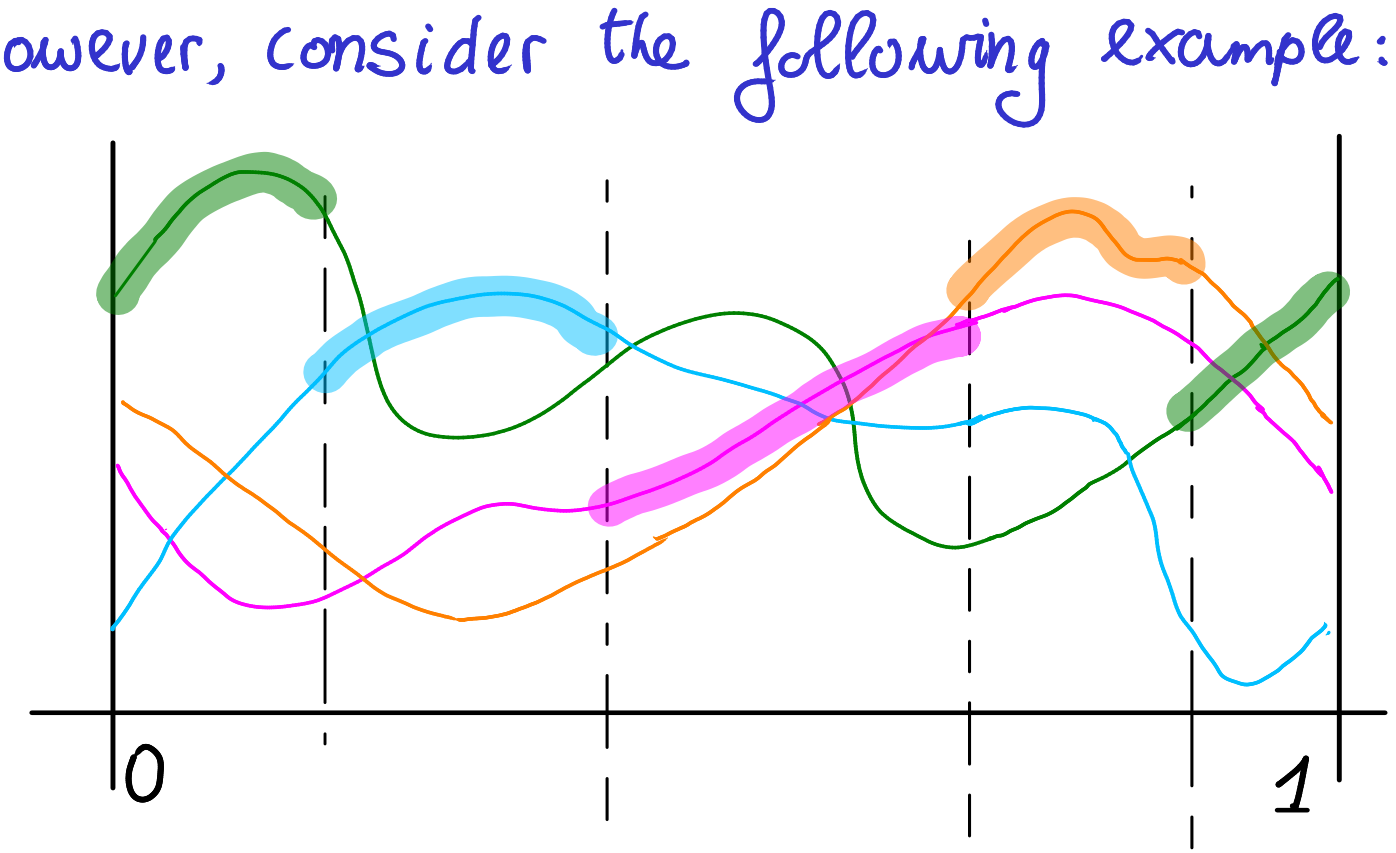


Data analysis: Images

However, consider the following example:

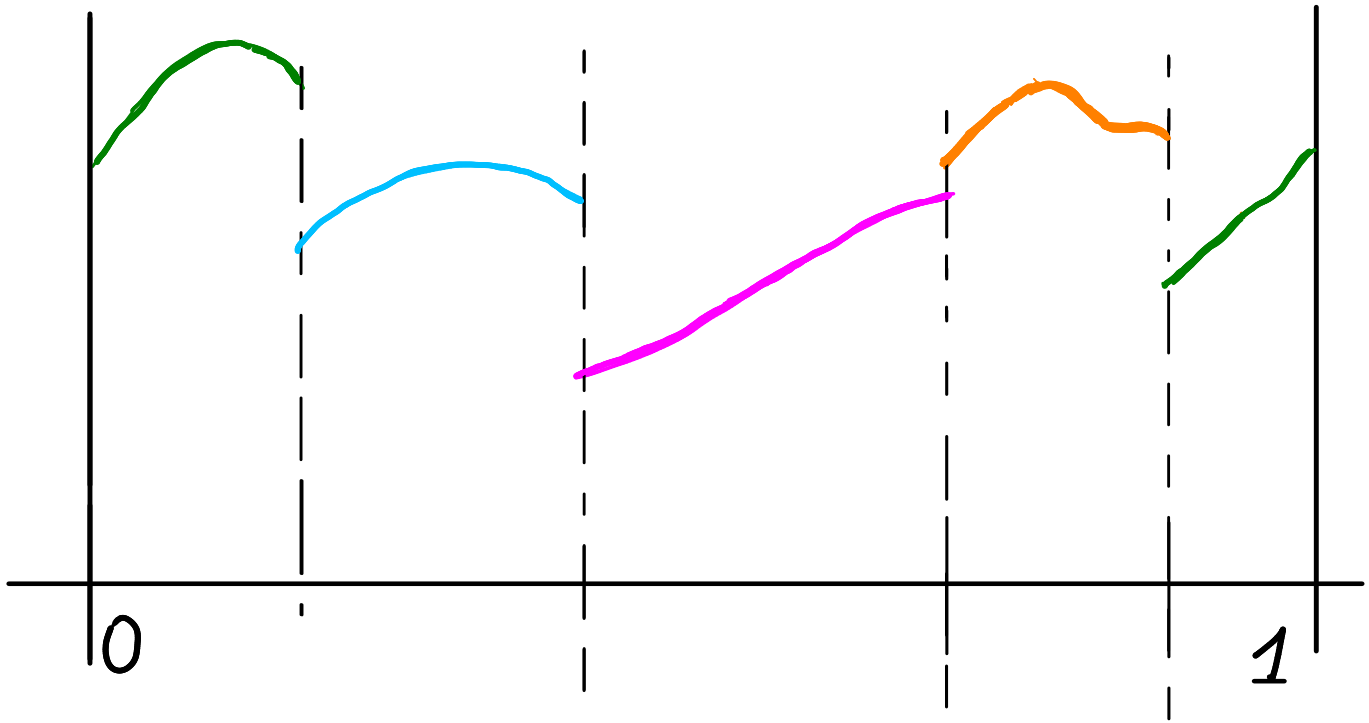


Data analysis: Images



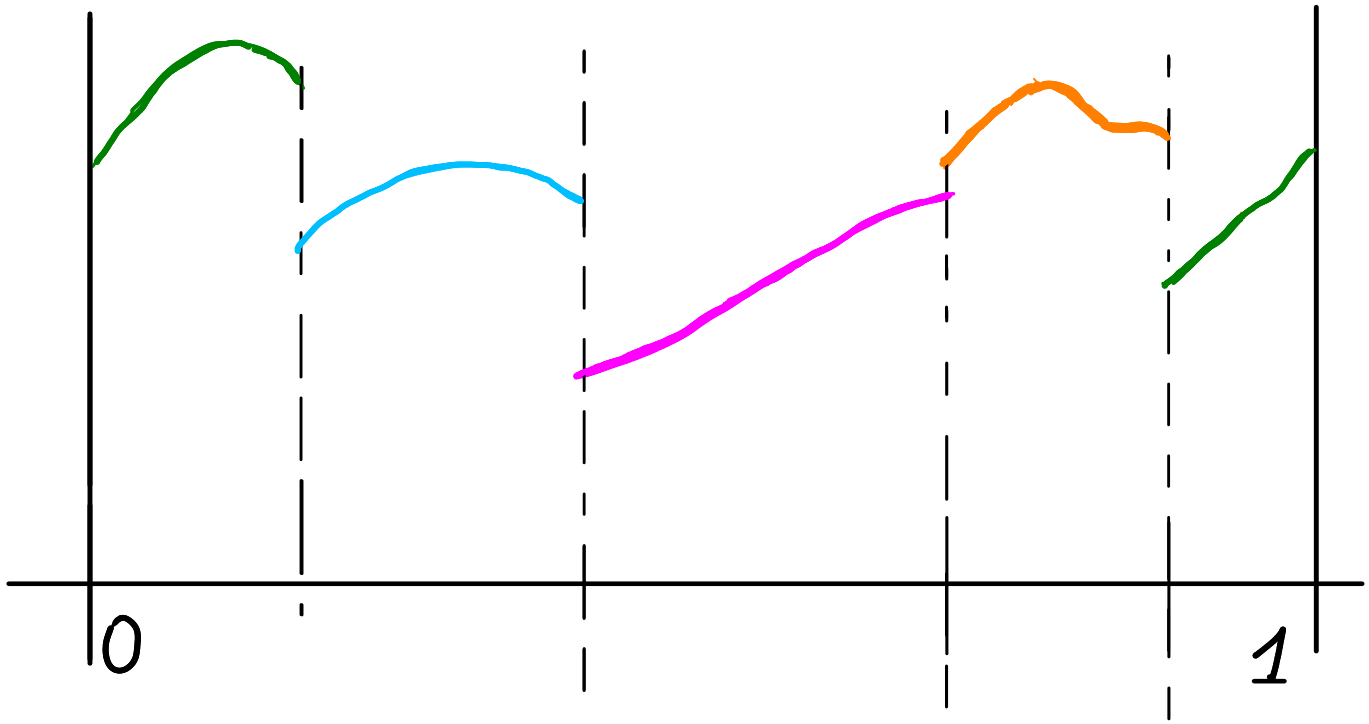
Data analysis: Images

However, consider the following example:



Data analysis: Images

However, consider the following example:



Clearly translation invariant process..

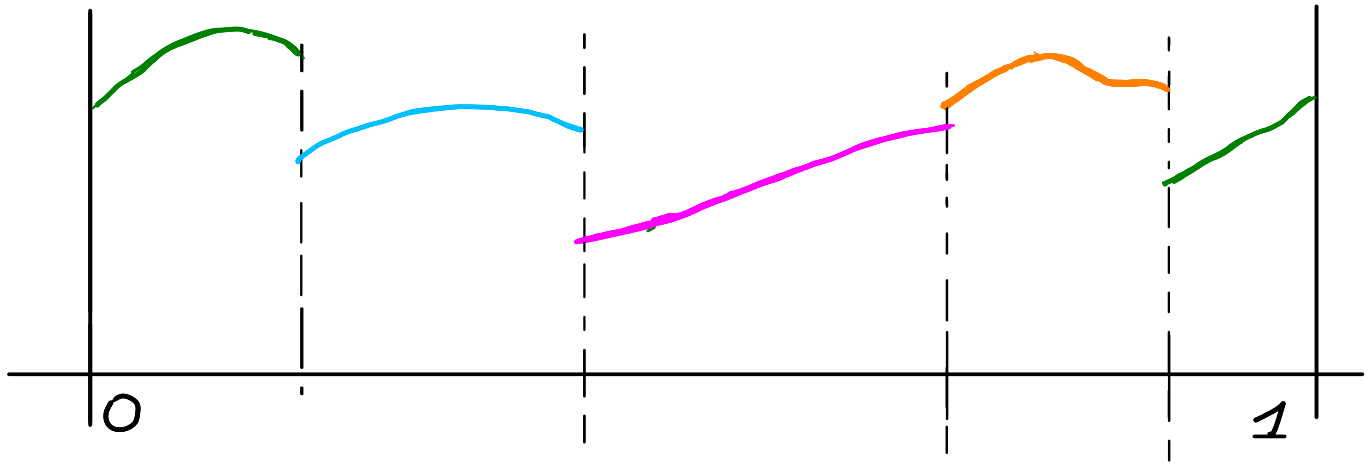
Data analysis: Images

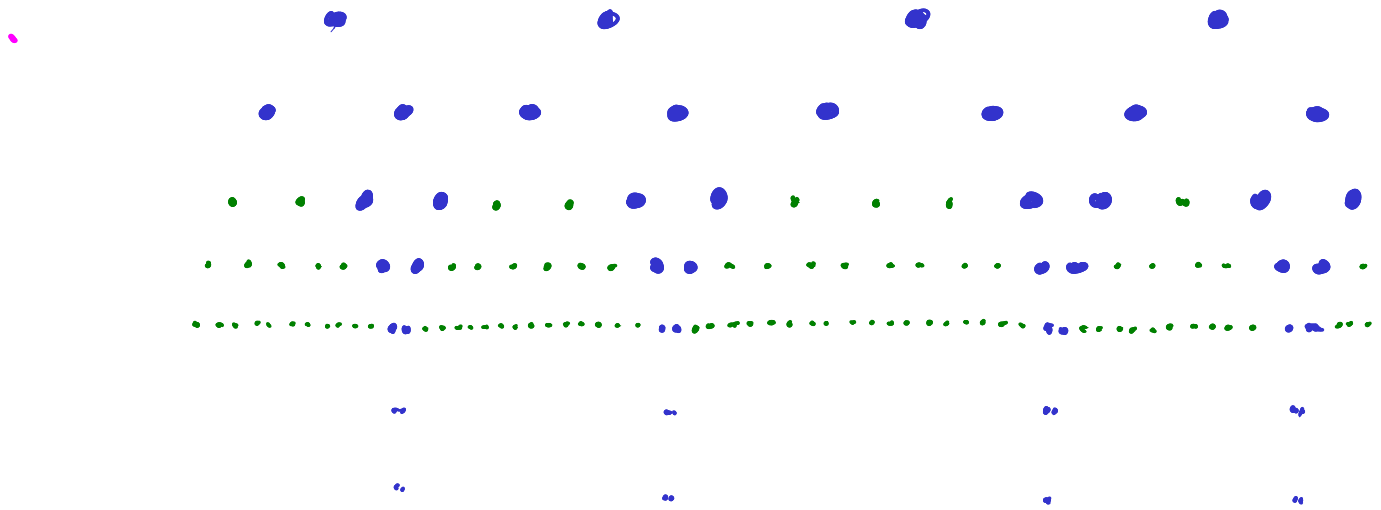
Yet, one can prove that

$$\mathbb{E}\left(\int_{\mathbb{T}} \left| f(t) - \sum_{|n| \leq N} \langle f, e_n \rangle e_n(t) \right|^2 dt \right) \geqslant C \frac{1}{N}$$

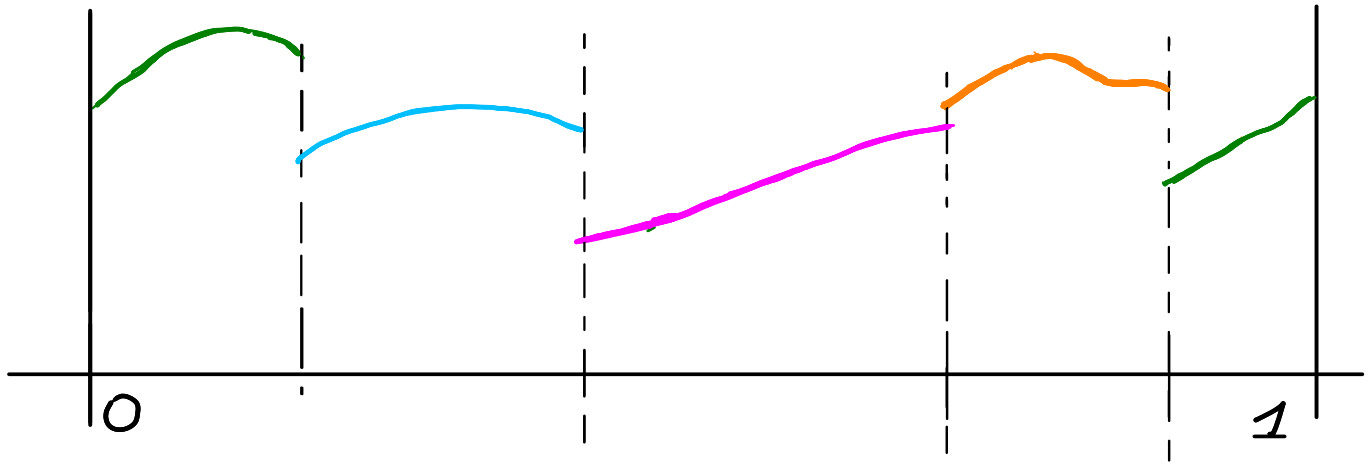
But with a wavelet expansion it is very simple to find a strategy that does better ...

Data analysis: Images





Data analysis: Images



One easily proves that with this strategy,

$$\mathbb{E}\left(\int_{\mathbb{T}} |f(t) - A_{2N+1}(f)(t)|^2 \, dt \right) \leq C N^{-2}$$

where $A_{2N+1}(f)$ is an approximation to f that uses only $2N+1$ coeffs.

Data analysis: Images

But there is an enormous difference.

In 1 case

$$\mathbb{E}\left(\| f - \sum_{\ell=1}^{L} \langle f, \varphi_\ell \rangle \varphi_\ell \|^2 \right)$$

is minimal

In the other case,

$$\mathbb{E}\left(\| f - \sum_{\ell \in \Lambda_L(f)} \langle f, \psi_\ell \rangle \psi_\ell \|^2 \right)$$

is considered, with $\# \Lambda_L(f) = L$

In both cases, L coeffs allowed, but in 2nd case their choice can depend on f.

Data analysis: Images

Linear approximation:

$$\varphi_1, \varphi_2 \dots \varphi_n, \dots \quad \Rightarrow \quad \mathrm{Span}(\varphi_1, \dots \varphi_n) = V_n$$

and study $\mathrm{dist}(f, V_n)$

Nonlinear approximation:

$$\varphi_1, \varphi_2, \dots \varphi_n, \dots$$

$$\Sigma_n = \left\{ \sum_{\ell \in \mathbb{N}} c_\ell \varphi_\ell \; ; \; \#\{\ell; c_\ell \neq 0\} \leq n \right\}$$

now study $\mathrm{dist}(f, \Sigma_n)$.

Data analysis: Images

Wavelets are a good basis for nonlinear approximation of images, because images have sparse wavelet expansions.

Data analysis: Images

Wavelets are a good basis for nonlinear approximation of images, because images have sparse wavelet expansions.

With hindsight : first example of benefit of sparse expansions.

Why do wavelets have this property ?

Wavelets are connected with beautiful and strong theorems in harmonic analysis

Calderón - Zygmund theory

In fact, wavelets are not even the best basis for 2D - images

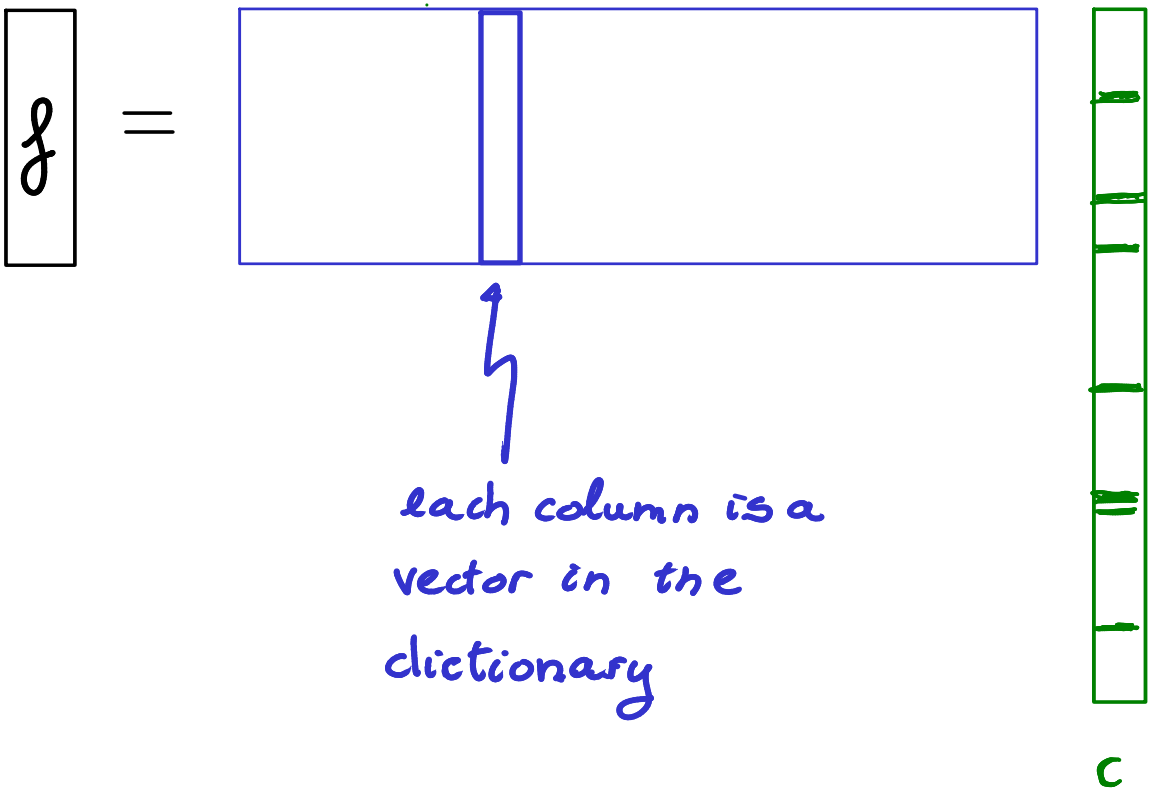
Images really need curvelets (or shearlets)

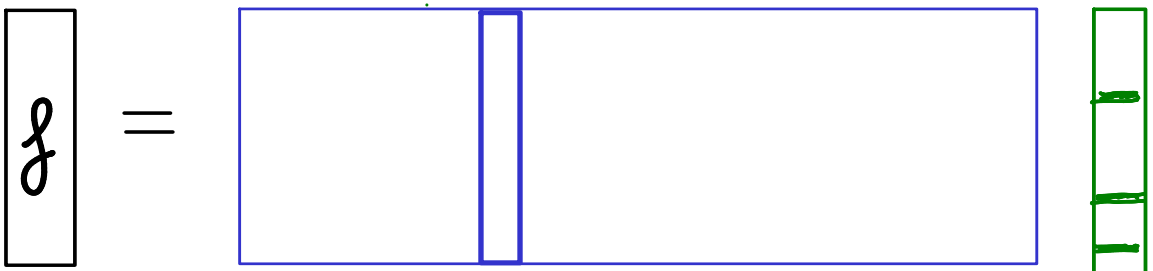
For wavelets, we were lucky : we "guessed" a good basis.

Can we search for a good basis for sparse expansions ?

Find good basis for sparse expansions?

- Search within "dictionaries"
  ↓
  union of many bases.
- Nonlinear (adaptive) singular value decompositions


each column is a
vector in the
dictionary
c



each column is a vector in the dictionary

c

Same kind of situation as in compressed sensing !

Compressed sensing.

Back to images, for a moment.

Images are sparse when expressed as a combination of wavelets.

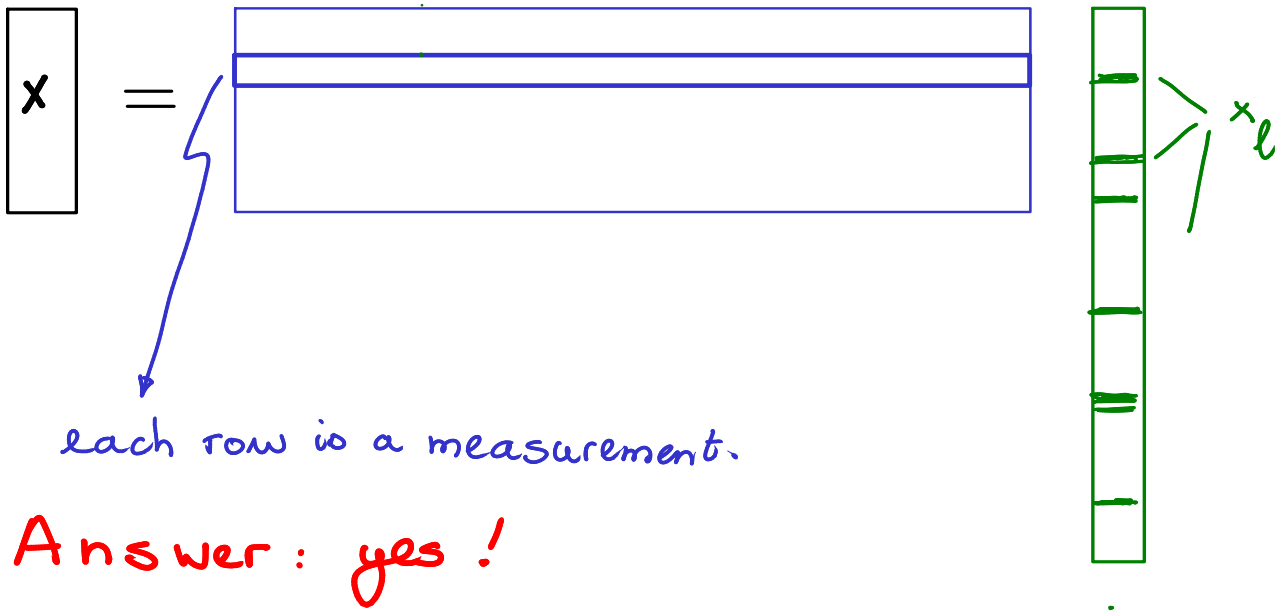
For compression applications:

- use fast transform to decompose into wavelets
- retain only the significant coeffs. (identity depends on image)

Why bother first getting all these coeffs? Why not "acquire" image sparsely?

In other words, if we know $x \in \mathbb{R}^N$ is a sparse vector, i.e. $\#\{\ell\,;\, x_\ell \neq 0\} \leq K << N$, can we then determine x by making fewer than N measurements?

In other words, if we know $x \in \mathbb{R}^N$ is a sparse vector, i.e. $\#\{\ell\,;\ x_\ell \neq 0\} \leq K << N$, can we then determine x by making fewer than N measurements?



Answer: yes!

Compressive sensing is related to results in theoretical computer science

use Johnson-Lindenstrauss Lemma.

$x_1, \dots x_L$ vectors in V

high-dim. space

$\dim V = D$

Consider projections of x_ℓ on randomly picked d-dim. subspace of V.

Compare $\langle P x_i , P x_j \rangle \frac{D}{d}$ with $\langle x_i, x_j \rangle$

How large should d be for these 2 matrices to be close with high probability?

Basically: $\log L$

This result in CS has had a tremendous impact

- verify that proofs are correct with high probability by "random sampling"
- fast computation algorithms (with small probability of failure).

Fast computations: example.

$f \in \mathbb{C}^N$ N huge

There exists $x \in \mathbb{C}^N$, with only $K << N$ non-zero entries, that is close to f.

$\Rightarrow$ To get a good approximation to f, one needs to take only (non-adaptively!) $O(K \log N)$ random samples of f

and algorithm runs in $O(K \log N)$ time as well.

Finding good ways to represent data.

↳ Knowing (or "believing") that there is a sparse expansion can be exploited to reconstruct from seemingly very insufficient data.

Search in a dictionary
↔ compressed sensing.

Find the dictionary if given a class of objects?



Johnson - Lindenstrauss
Compressed sensing
} dimension reduction.

One last salvo about computation made feasible by "dimension reduction"

↳ comparing surfaces, with applications to biology.

$$d_p(\mathcal{L}, \mathcal{L}') = \inf_{\substack{m : \mathcal{L} \to \mathcal{L}' \\ \text{matching}}} \left[\min_{R \in \text{Euclidean gp.}} \sum_{p \in \mathcal{L}} \| m(p) - Rp \|^2 \right]^{1/2}$$

$$d_p(\mathcal{L}, \mathcal{L}') = \inf_{\substack{m : \mathcal{L} \to \mathcal{L}' \\ \text{matching}}} \left[\min_{R \in \text{Euclidean gp.}} \sum_{p \in \mathcal{L}} \| m(p) - Rp \|^2 \right]^{1/2}$$

$$\mathcal{D}_p(S, S') = \inf_{C : S \to S' \text{ area-preserving}} \left[\min_{R \in \text{Euclidean gp.}} \int_S \| C(x) - Rx \|^2 \, dA_S \right]^{1/}$$

If $\mathcal{D}_p(S, S')$ is small,

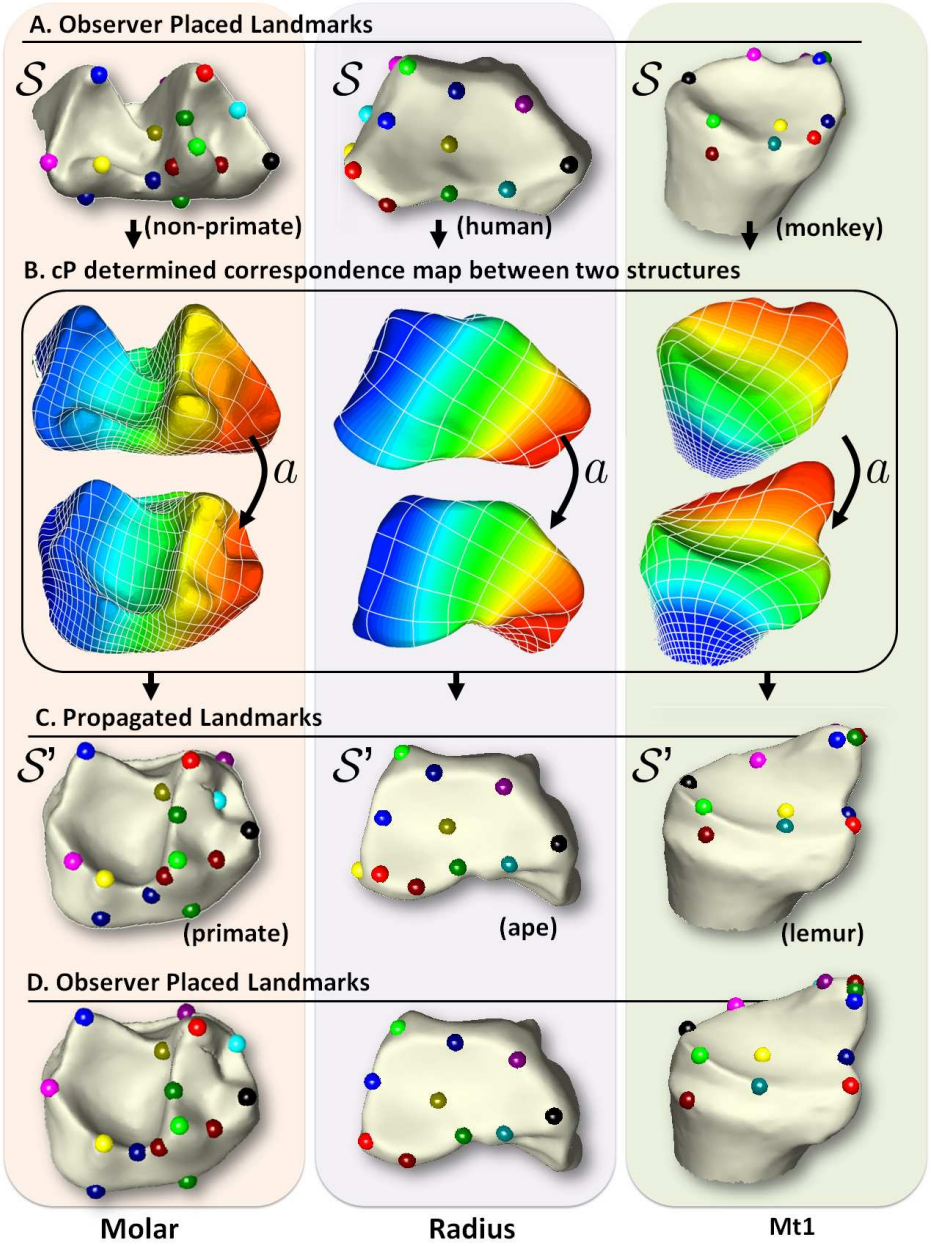
then $\exists$ conformal map $m : S \to S'$

so that

$$\min_R \int_S \| m(x) - Rx \|^2 \, dA_S \leq C \, \mathcal{D}_p(S, S')^{1/2} .$$

↪ use this to compute approx. to $\mathcal{D}_p(S, S')$

by searching "deformations of conformal maps"


A. Observer Placed Landmarks
$\mathcal{S}$
(non-primate)
$\mathcal{S}$
(human)
$\mathcal{S}$
(monkey)
B. cP determined correspondence map between two structures
a
a
a
C. Propagated Landmarks
$\mathcal{S}'$
(primate)
$\mathcal{S}'$
(ape)
$\mathcal{S}'$
(lemur)
D. Observer Placed Landmarks
Molar
Radius
Mt1

With apologies to

A. Cohen
J.P d'Ales
Y. Meyer
R. De Vore
R. Coifman
E. Candès
D. Donoho
J. Romberg
T. Tao
W. Dahmen

L. Carin
A. Gilbert
M. Strauss
R. Calderbank
Y. Lipman
O. Yilmaz
D. Boyer
J. Jernvall

and many, many more ...