# CIMPA 2013 Course on

# High Dimensional Classification with Deep Scattering Networks

*Stéphane Mallat*

**École Normale Supérieure**

www.di.ens.fr/data/scattering

# Big Data Analysis

• Tremendous increase of data acquistion: audio, images, video medical/biological data, industrial processes, social networks...

• Automatic analysis becomes critical for industries, science medecine, Internet search, new services.

• Memory and computational capacity now allow to sovle large size complex classification problems.

• Effective but complex non-linear algorithms are developped such as deep neural networks.

• Lack of Mathematics.

# Big Data Analysis

- Classification problems:

  find the label $y(x) \in \{1, ..., K\}$ for a data vector $x \in \mathbb{R}^d$

  given a training set of examples: $\left\{(x_i, y_i)\right\}_i$

  Data dimension $d \geq 10^6$, Number of classes $2 \leq K \leq 10^4$

  Number of training examples per class from 10 to 1000

- An interpolation problem:

  find a good approximation $\tilde{y}(x)$ of $y(x)$, with $\left\{\tilde{y}(x_i) = y_i\right\}_i$

- Piecewise constant interpolation: nearsest neighbor classifier

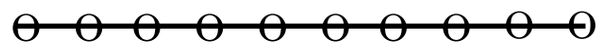
  $\tilde{y}(x) = y_j$ if $x_j = \arg\min_{x_i} \|x - x_i\|$ Failure

# Curse of Dimensionality

- Failure of standard analysis in high dimension: here $d \geq 10^6$.

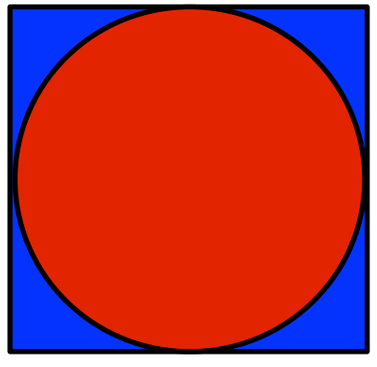
- Points are far away in high dimensions $d$:

  - 10 points cover $[0,1]$ at a distance $10^{-1}$

  - $10^d$ points cover $[0,1]^d$ at a distance $10^{-1}$.

$$\lim_{d \to \infty} \frac{\text{volume sphere of radius r}}{\text{volume } [0,r]^d} = 0$$



: nearly all points are in the corners!

$\Rightarrow$ there is typically no close data point in high dimension.

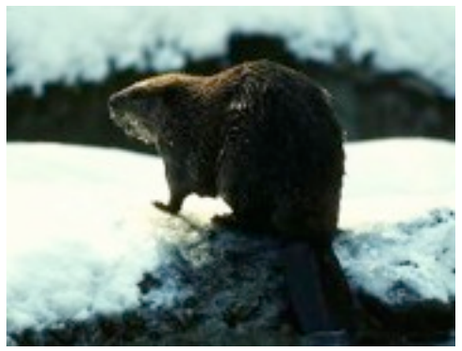
# High Dimensional Classification

*CalTech 101*

Anchor

Joshua Tree

Beaver



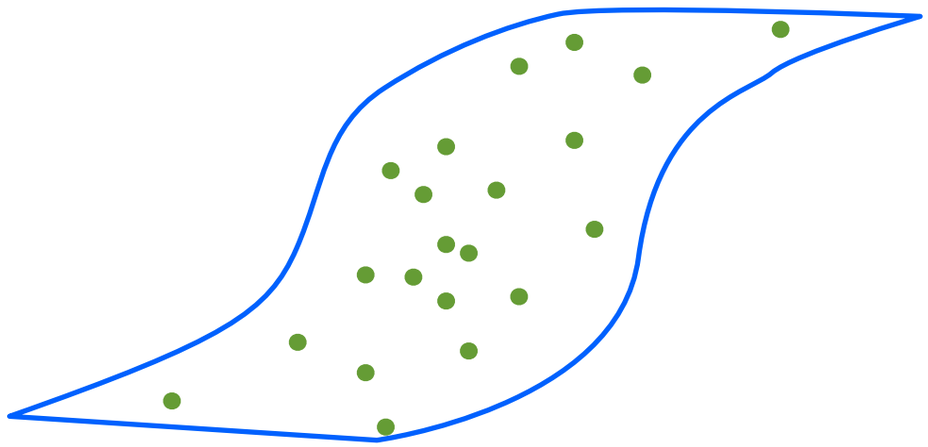
Lotus

Water Lily

- Considerable variability in each class.
- Euclidean distances are meaningless
- Need to find **discriminative invariants**.

# Low-Dimensional Manifold

- The curse of dimensionality is not a problem if signals belong to low-dimensional manifolds:



$\Rightarrow$ Euclidean distances provide local similarity measures

- Manifold technics: find intrinsic coordinates

for example by diagonlizing the Laplace-Beltrami operator.

- Applies to output of low-dimensional dynamical systems but not valid for complex signals such as music, speech, images, geophysical data, medical signals, financial series...

# Stochastic Models

- Each $x \in \mathbb{R}^d$ of class $k$ is a realization of a process $X_k$ of probability distribution $p_k(x)$

  $\Rightarrow$ classification is about discriminating random processes.

- Not enough samples to estimate $p_k(x)$ in high dimension, but can discriminate different $p_k(x)$ from their projections:

$$\left\{ E(F_m(X_k)) = \int_{\mathbb{R}^d} F_m(x)\, p_k(x)\, dx \right\}_m$$

- For classification: $E(F_m(X_k))$ must be estimated from one realization $x$ $\Rightarrow$ need ergodicity property

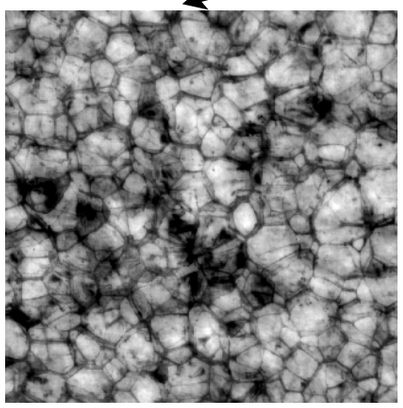
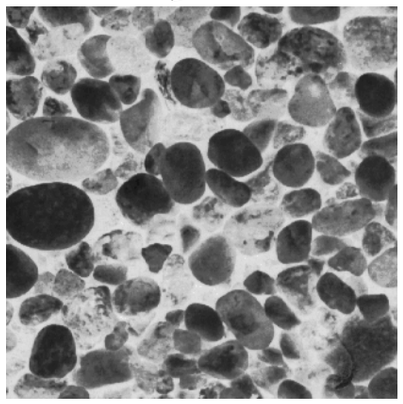
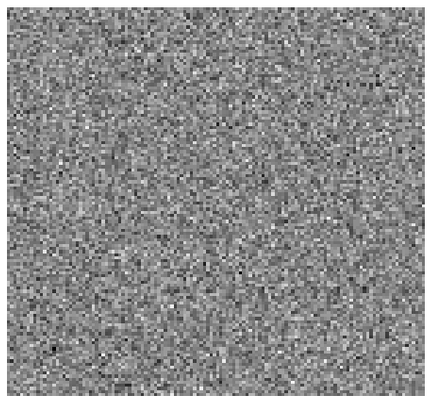
  $\widetilde{F}_m(x) \approx E(F_m(X_k)) = \mu_k$ **invariant** in the class $k$.

# Texture Discrimination

• Textures are realizations of high-dimensional stationary processes, which are typically not Gaussian or Markovian.

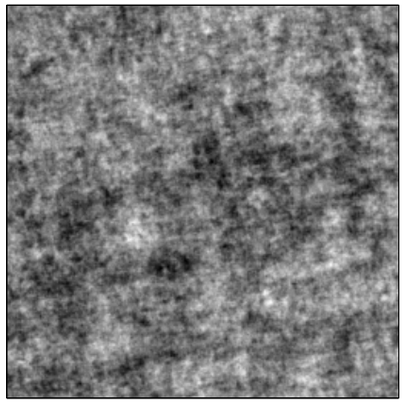
• Second order moment projections: $E(X(t)\,X(t-m)) = R(m)$ estimated with weak ergodicity conditions: power spectrum.

**same second order moments**



**same second order moments:** not discriminative.

• Use higher order moments ?
Estimators have a large variance
$\Rightarrow$ not sufficiently invariant.

# Audio Textures

J. McDermott textures

same second order moments

- Natural Sounds (1s) Original Gaussian model
  - Hammer
  - Water
  - Applause

# Learning Invariants

- Classifications can be reduced to multiple binary classifications

- Two classes $\mathcal{C}_k$ and $\mathcal{C}_l$ can be discriminated by finding an invariant operator $F_{k,l}$ with:

$$\forall x \in \mathcal{C}_k \quad F_{k,l}(x) \approx \mu_k \in \mathbb{R}$$

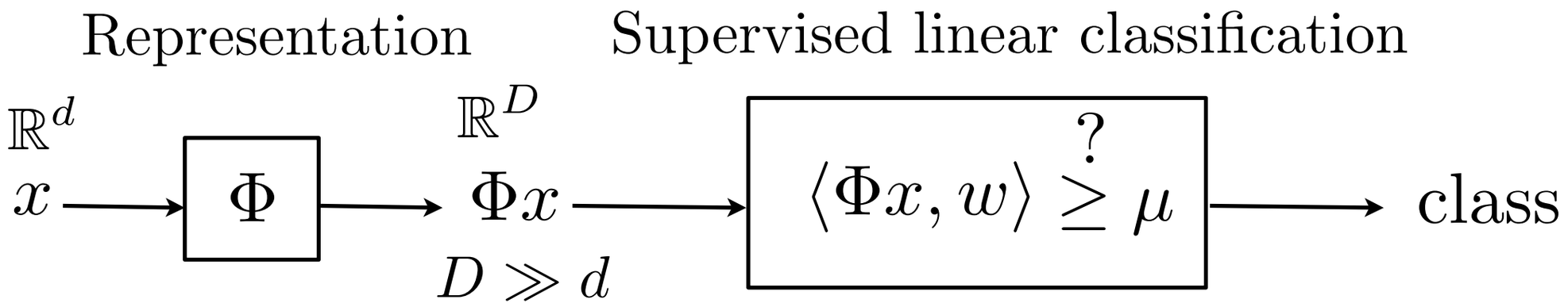
$$\forall x \in \mathcal{C}_l \quad F_{k,l}(x) \approx \mu_l \neq \mu_k$$

- Linear classifier compute $F_{k,l}(x)$ from a $\Phi(x) = \{\Phi_n(x)\}_{n \leq D}$

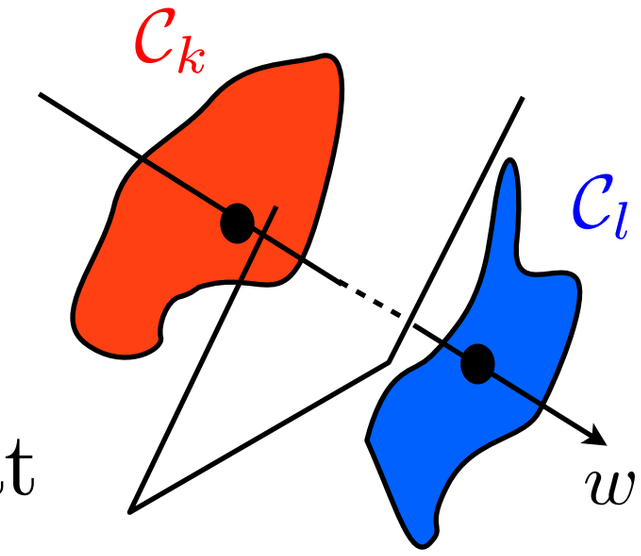
$$F_{k,l}(x) = \sum_{n \leq D} w_n \, \Phi_n x \ .$$

Strong classifier aggregating many "weak features".

# Hyperplane Separation




• Support Vector Machines optimize the choice of hyperplane: $(w, \mu)$ from examples.



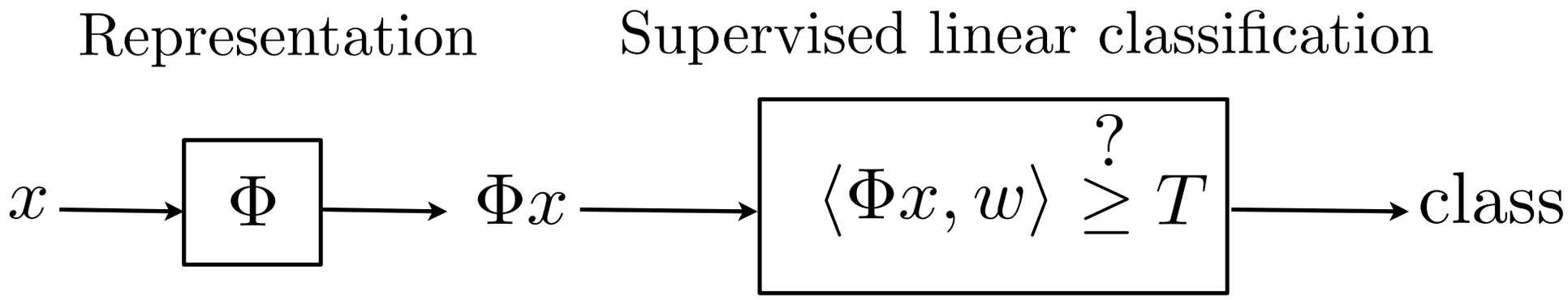

For any two classes $\mathcal{C}_k$ and $\mathcal{C}_l$ finds $w$ so that

$$\langle \Phi x, w \rangle = \sum_n w_n \, \Phi_n x$$

is nearly invariant and different in $\mathcal{C}_k$ and $\mathcal{C}_l$.

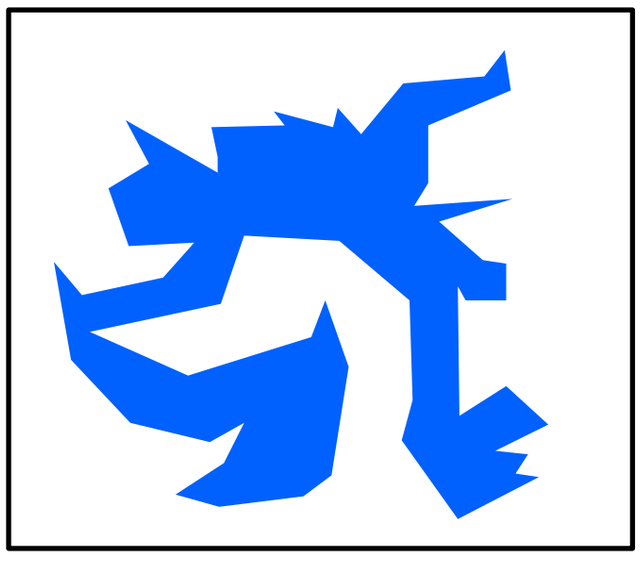
• How to define $\Phi$ to get linear discriminative invariants ?

# Classification with Invariants



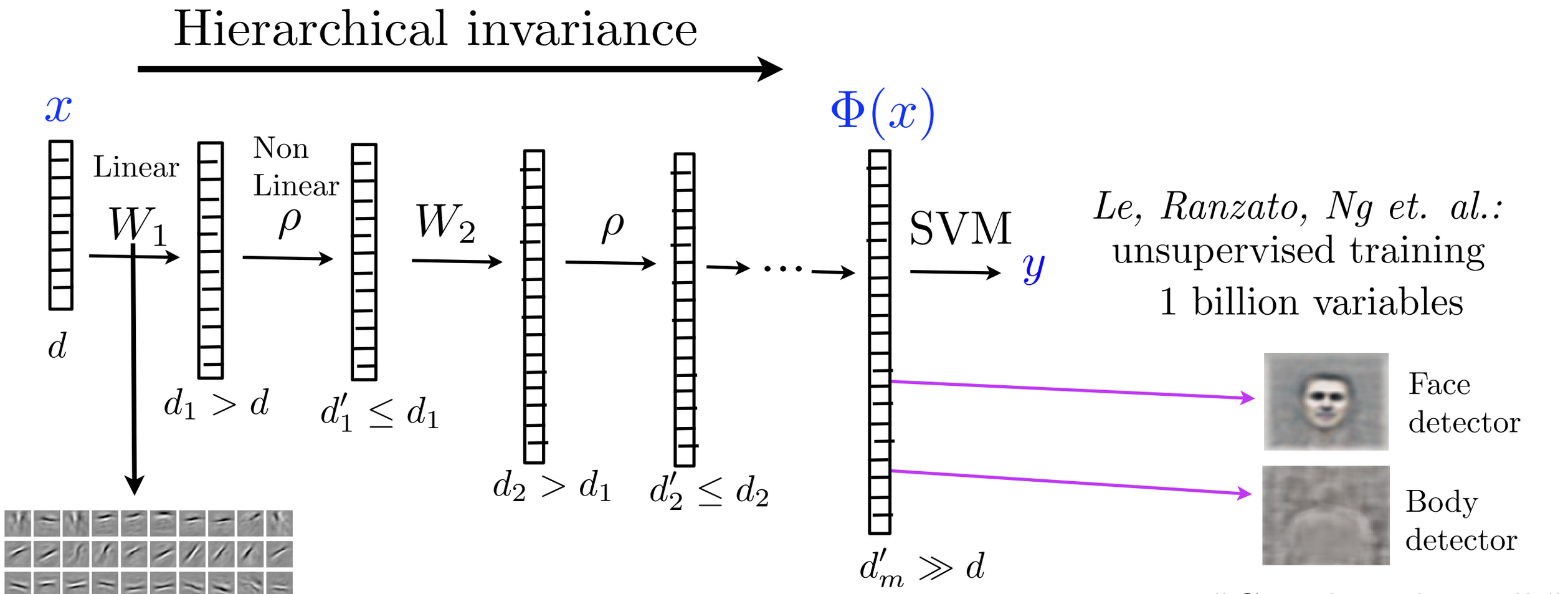

- $\Phi$ may be defined from prior knowledge on data.

- **Unsupervised learning** of $\Phi$ from unlabeled examples $\{x_i\}$: requires to model a very high dimension distribution.



# Deep Neural Networks

*J. Hinton, Y. LeCun* "State of the art results"



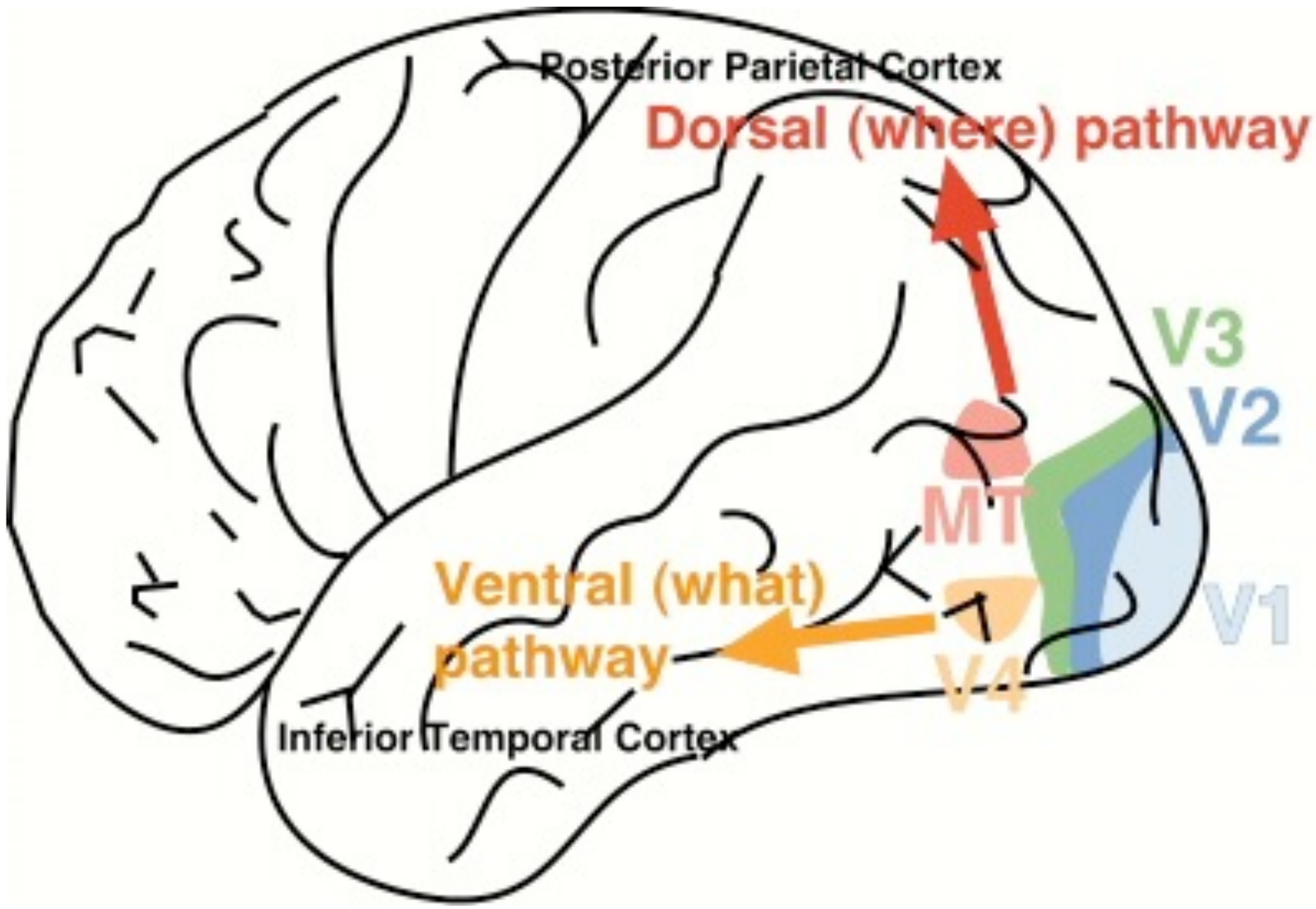

*Hinton, Bengio, Ranzato et. al.:*
usupervised learning with sparse auto-encoders

Why and how does it work ?

# The Best Image Classifier

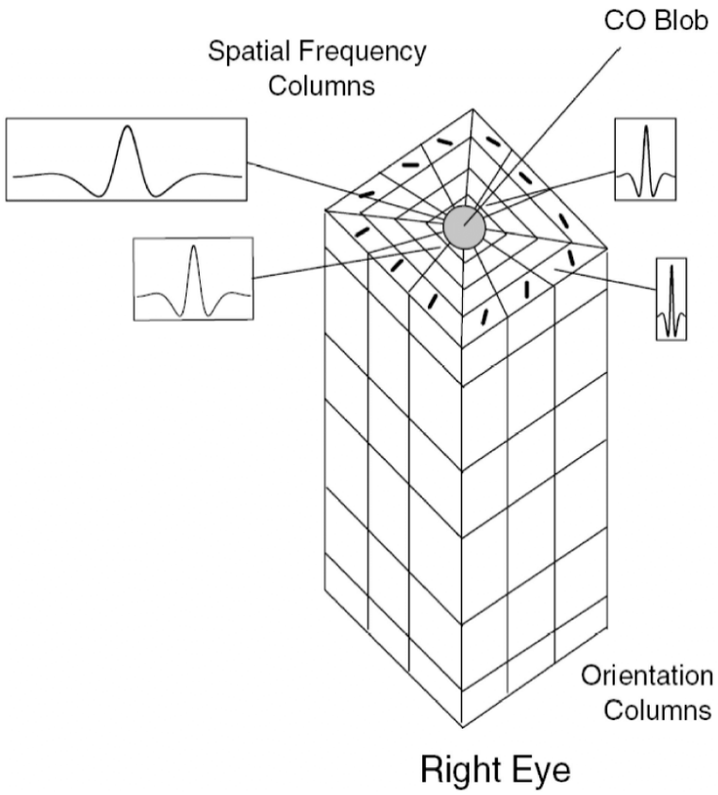
Over 30% of the brain for vision

Huge amount of memory

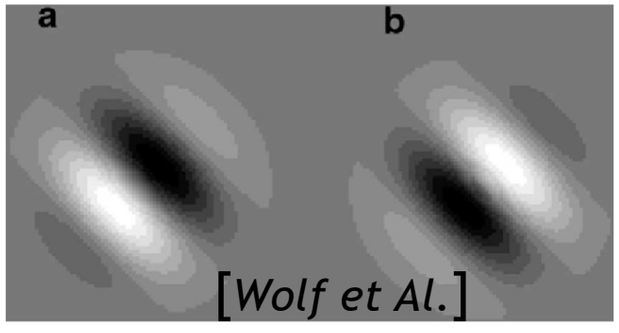


# Psychophysics of Vision

## **Hypercolumns in V1:**
directional wavelets



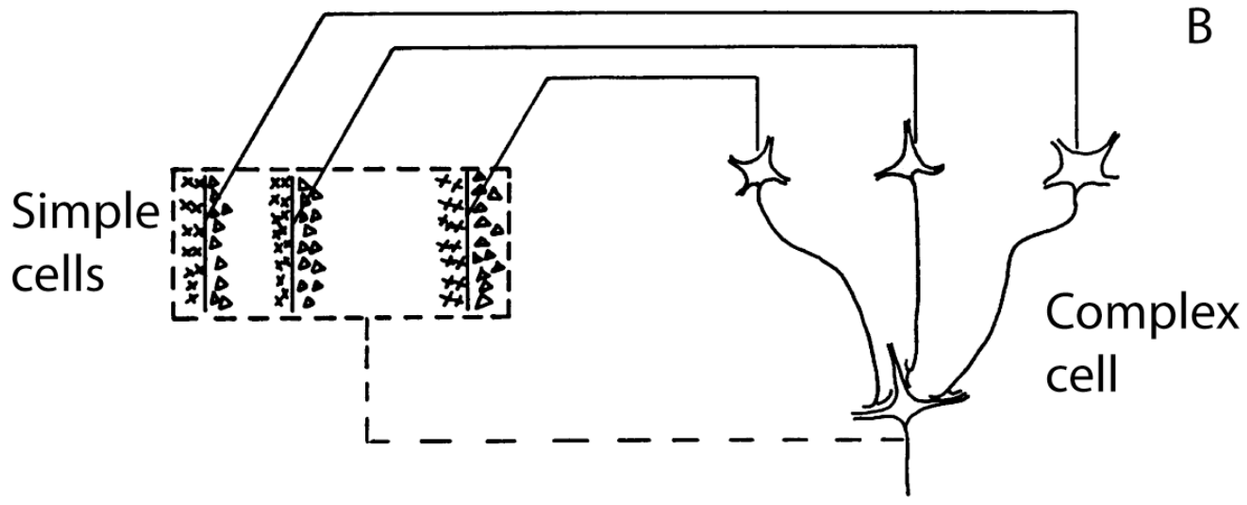

Simple cells Gabor linear models




$$\psi(x) = \theta(x)e^{i\xi x}$$

## **Complex Cells**

- Non-linear
- Large receptive fields
- Some forms of invariance

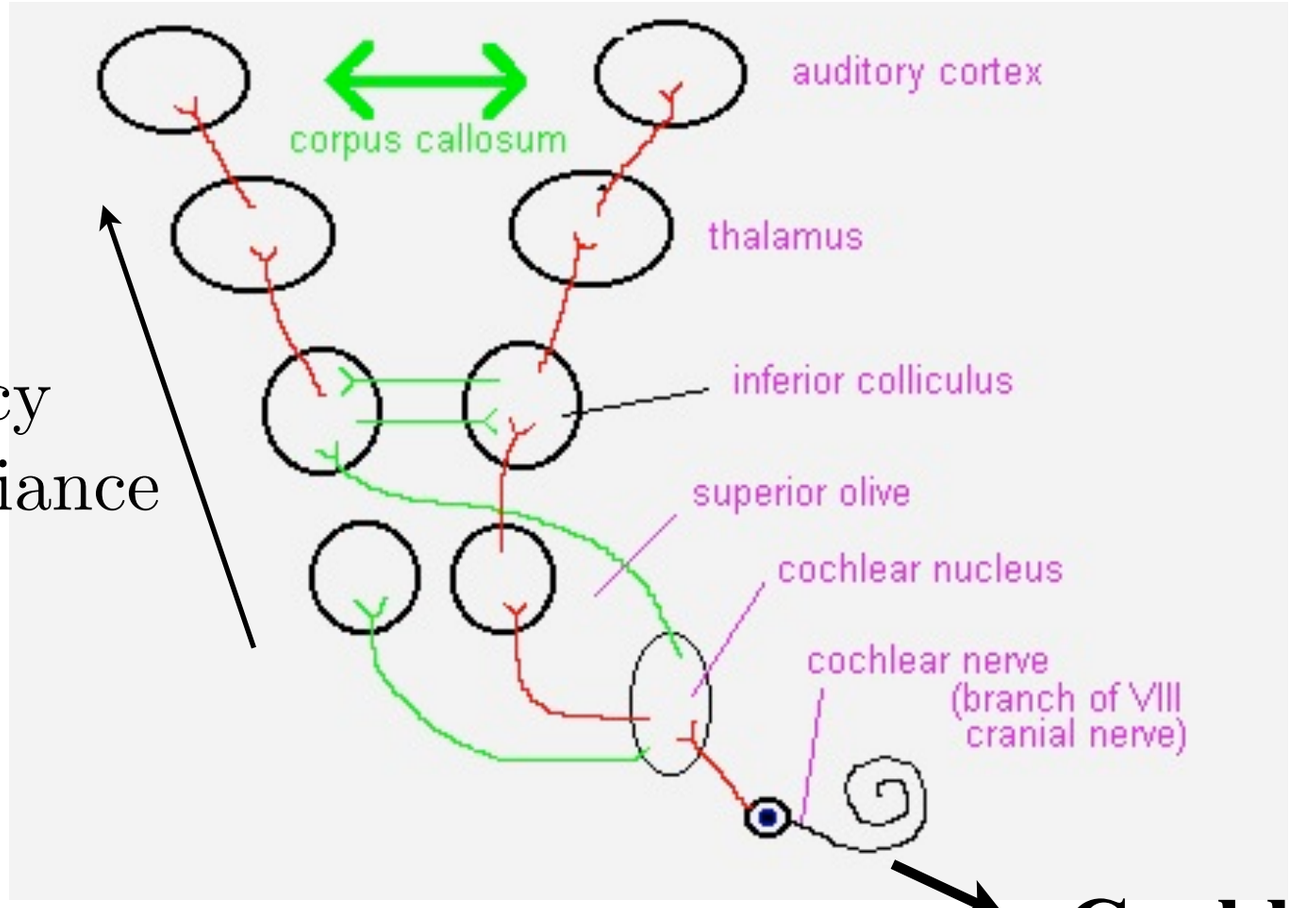



## **«What» Pathway towards V4:**

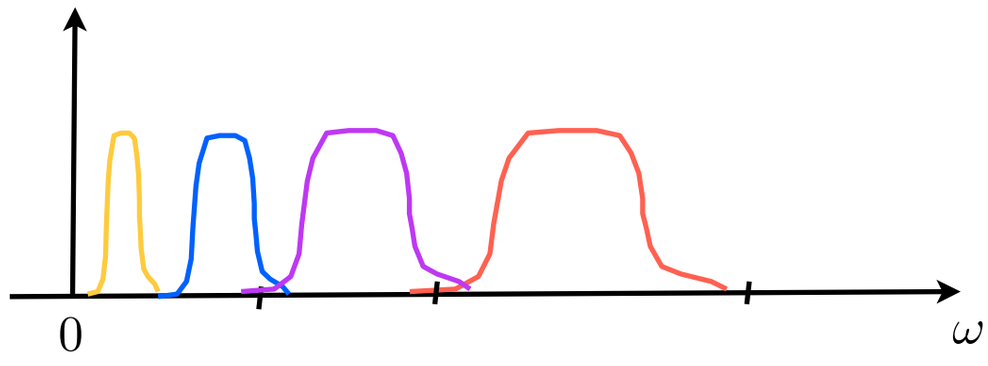
- More specialized invariance
- «Grand mother cells»

# Audio Psychophysics

Time and Frequency
"translation" invariance




**Cochlea:**
dilated wavelet filters




• Wavelets appear at early stages of vision and audition.
WHY ?

# Overview

- **Part I:**

  Invariants and stability to diffeomorphisms

  Scattering and deep neural networks

- **Part II:**

  Limit scattering transform

  Expected scattering of stationary processes

- **Part III:**
- Texture discrimination and synthesis

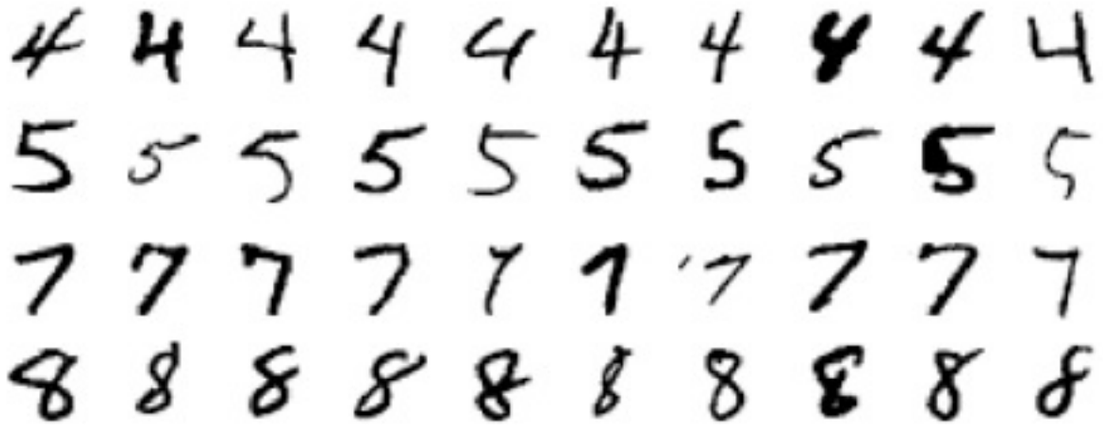
  Multifractal analysis

  Scattering on Lie Groups

  Unsupervised learning of representations

# Translations and Deformations

- Patterns are translated and deformed (class dependent)

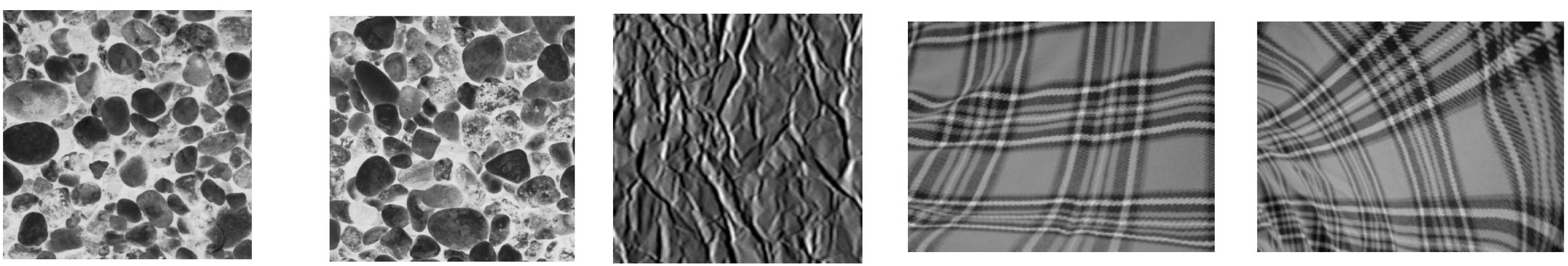


Group: $\mathbb{R}^2 \times \mathrm{Diff}(\mathbb{R}^2)$

two dimensions — infinite dimensions

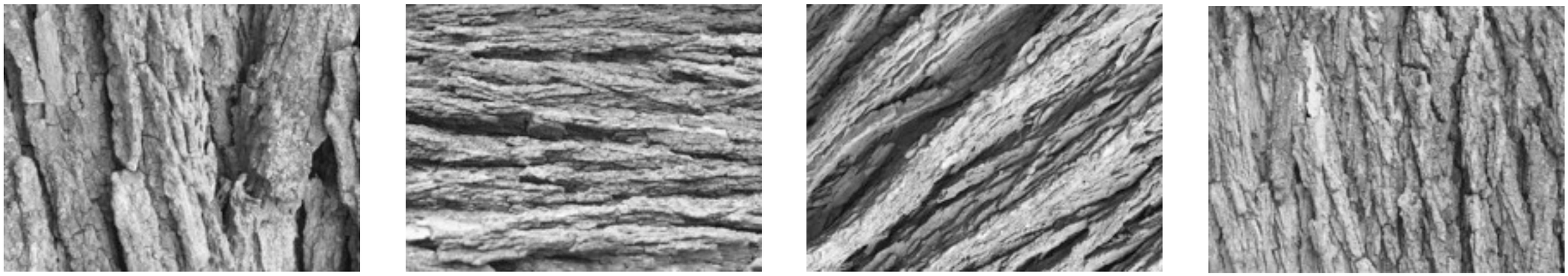
- Textures are stationary (translation invariant) processes

with deformations



# Rotation and Scaling Variability

- Rotation and deformations

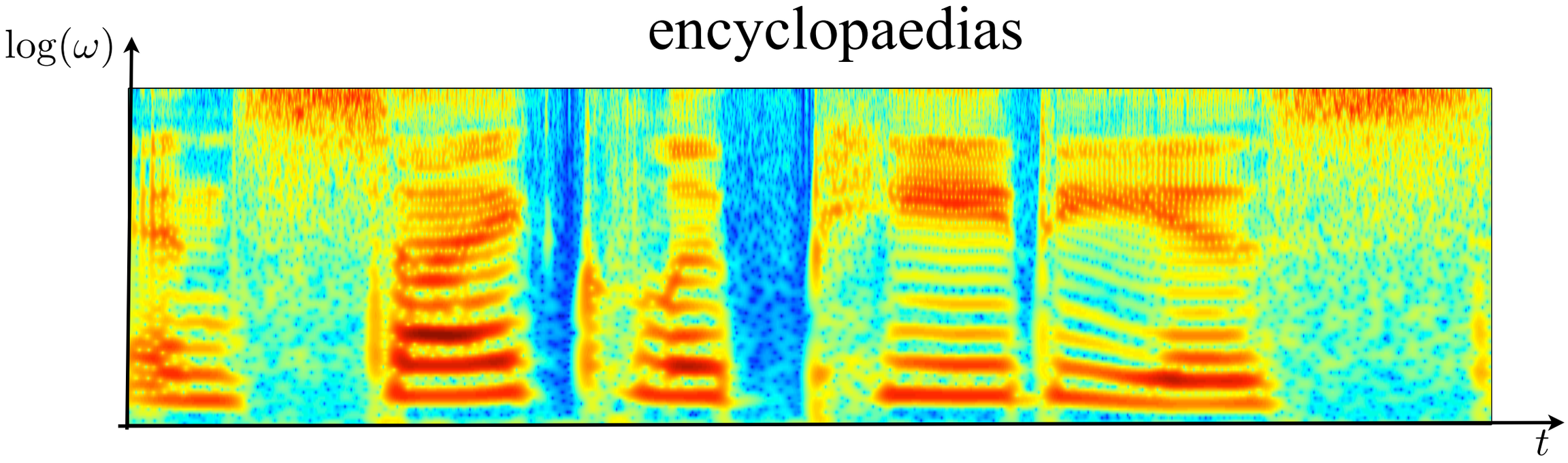


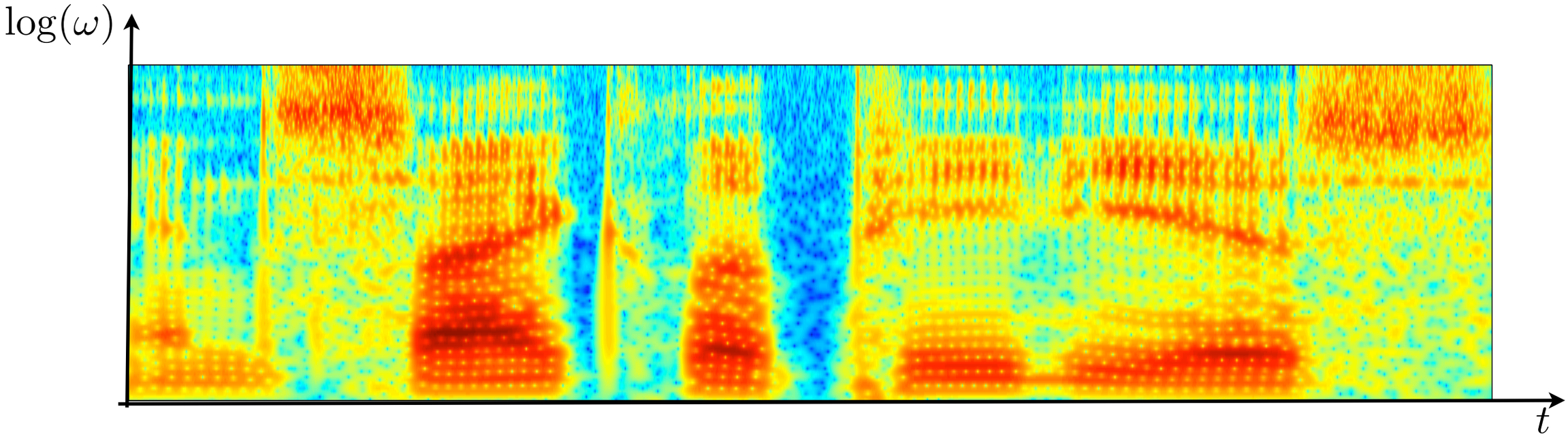
Group: $SO(2) \times \text{Diff}(SO(2))$

- Scaling and deformations

Group: $\mathbb{R} \times \text{Diff}(\mathbb{R})$

# Frequency Transpositions

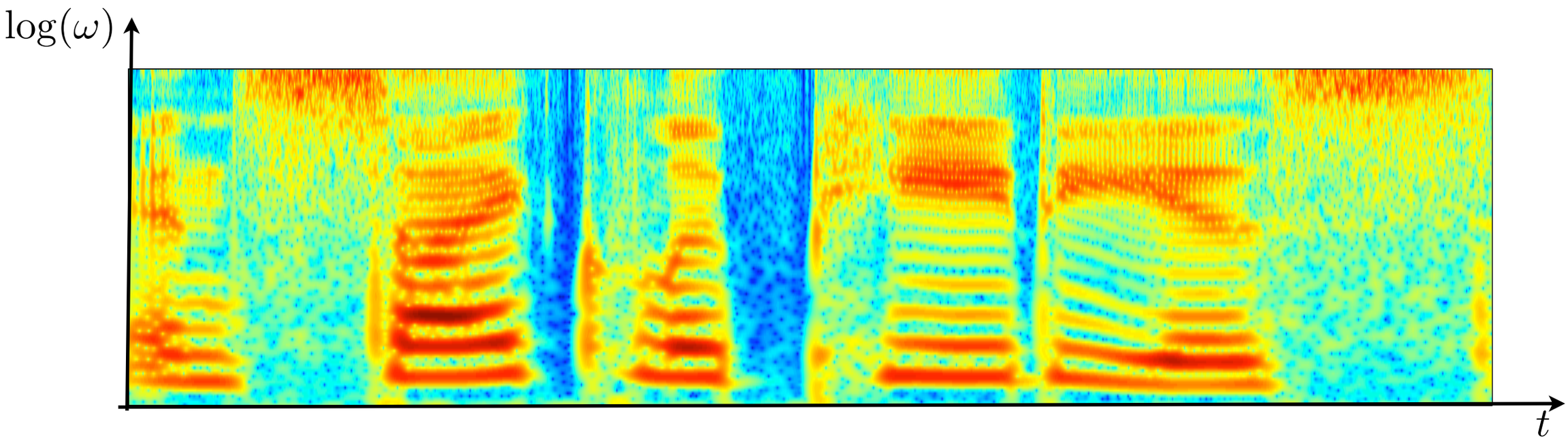



H : Heisenberg group of "time-frequency" translations

# Frequency Transpositions

Time and frequency translations and deformations:

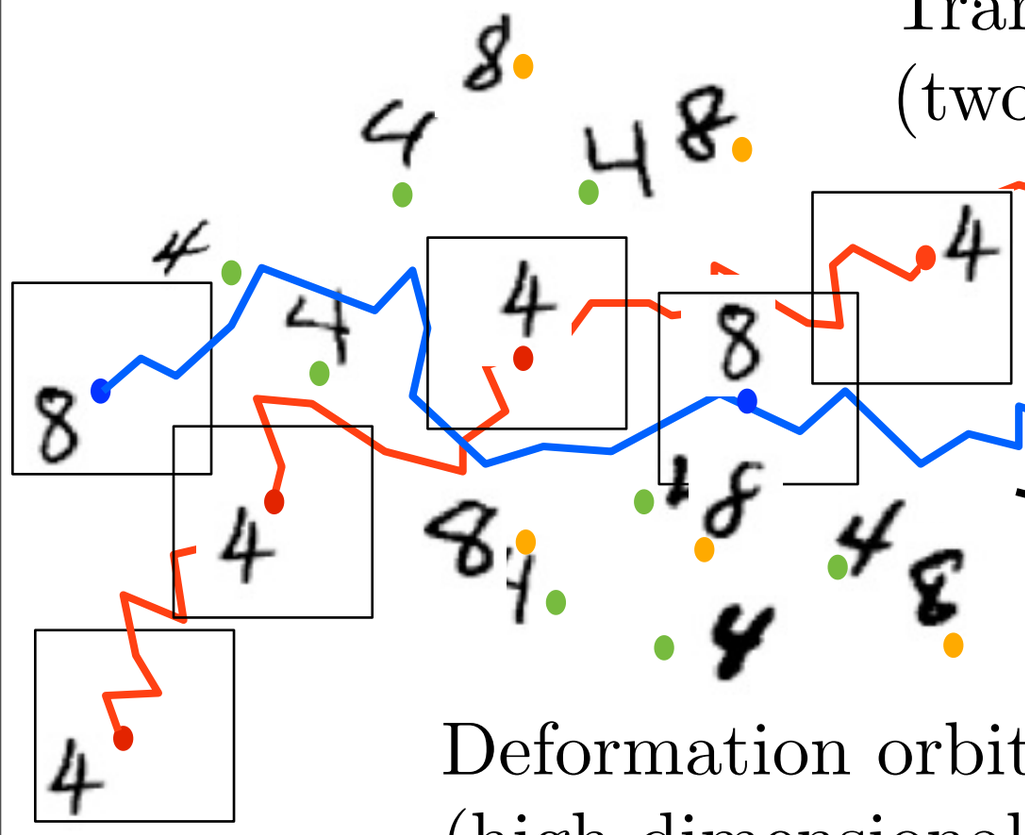



- Learning frequency transposition invariance:

  for speech recognition not for locutor recognition.

# Translation Invariance

- Specific deformation invariance must be learned.



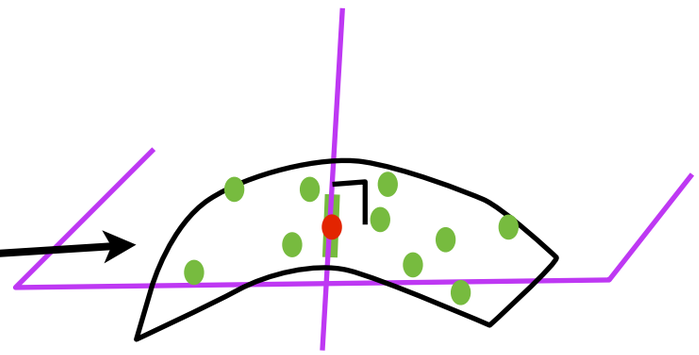
Translation orbits
(two-dimensional)

Deformation orbits
(high dimensional)

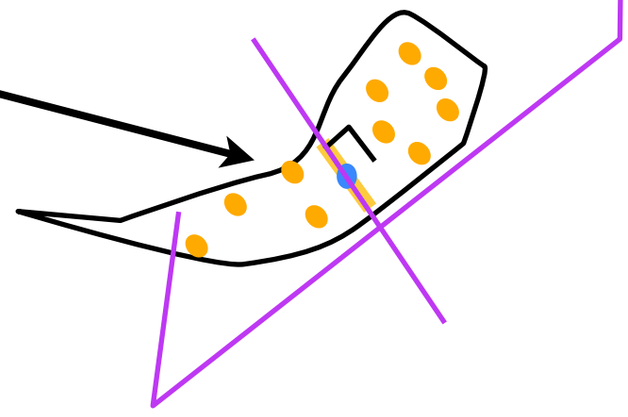
$\Phi$

**Invariant** to translations
**"Linearizes"** deformations
**Discriminant**

Supervised learning:



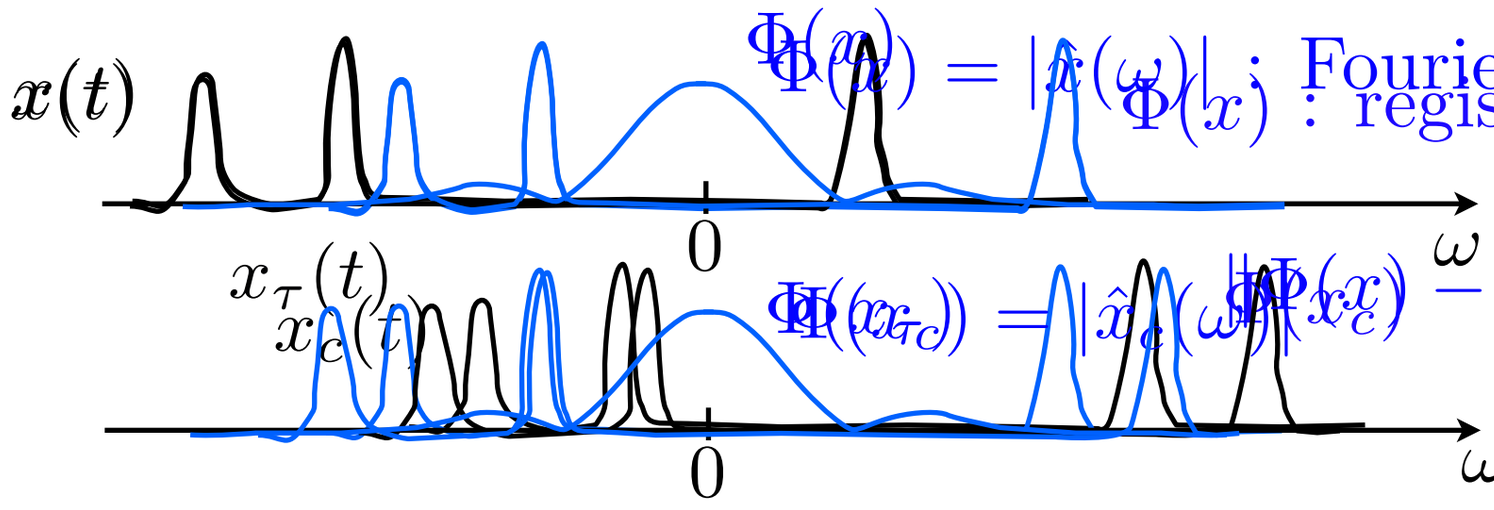
$\mathbf{P}_{\mathbf{V}_k^\perp}$
nearly invariant
to deformations



# Stable Translation Invariants

- **Invariance** to translations $x_c(t) = x(t-c)$

$$\forall c \in \mathbf{R} \quad , \quad \Phi(x_c) = \Phi(x) \ .$$




Not stable

Fourier invariants are not stable either.

- **Lipschitz stable** to diffeomorphisms $x_\tau(t) = x(t-\tau(t))$

small deformations of $x \implies$ small modifications of $\Phi(x)$

$$\forall \tau \quad , \quad \|\Phi(x_\tau) - \Phi(x)\| \leq C \underbrace{\sup_t |\nabla \tau(t)|}_{\text{diffeomorphism metric}} \|x\| \ .$$

# Fourier Translation Invariance

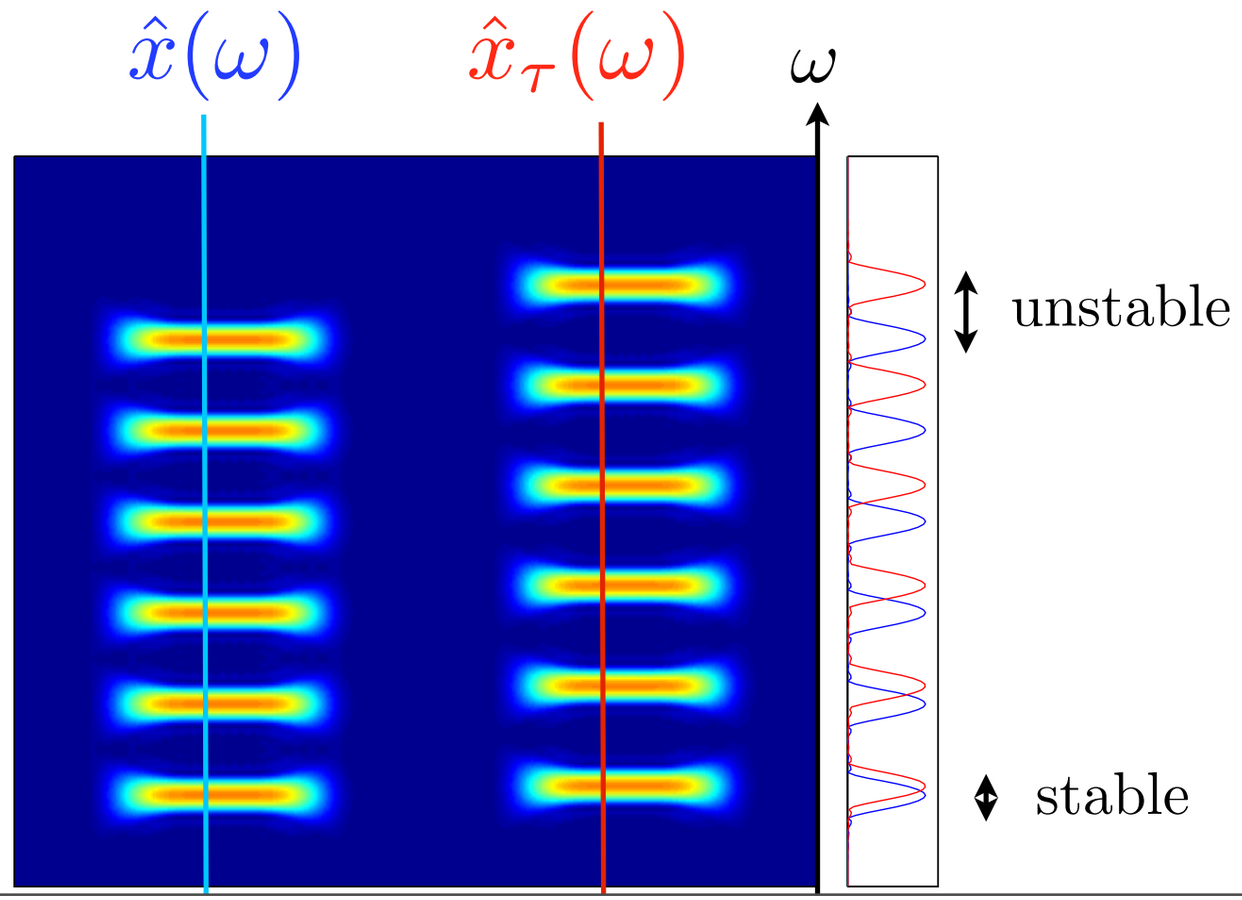
- Fourier transform $\hat{x}(\omega) = \int x(t)\, e^{-i\omega t}\, dt$ invariance:

$$\text{if } x_c(t) = x(t-c) \text{ then } |\widehat{x_c}(\omega)| = |\hat{x}(\omega)|$$

- Instabilites to small deformations $x_\tau(t) = x(t - \tau(t))$ :

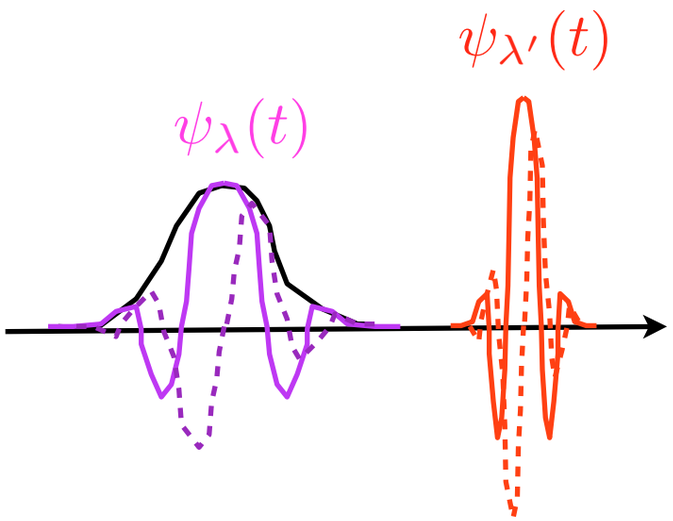
$$\big|\, |\hat{x}_\tau(\omega)| - |\,\hat{x}(\omega)|\,\big| \quad \text{is big at high frequencies}$$

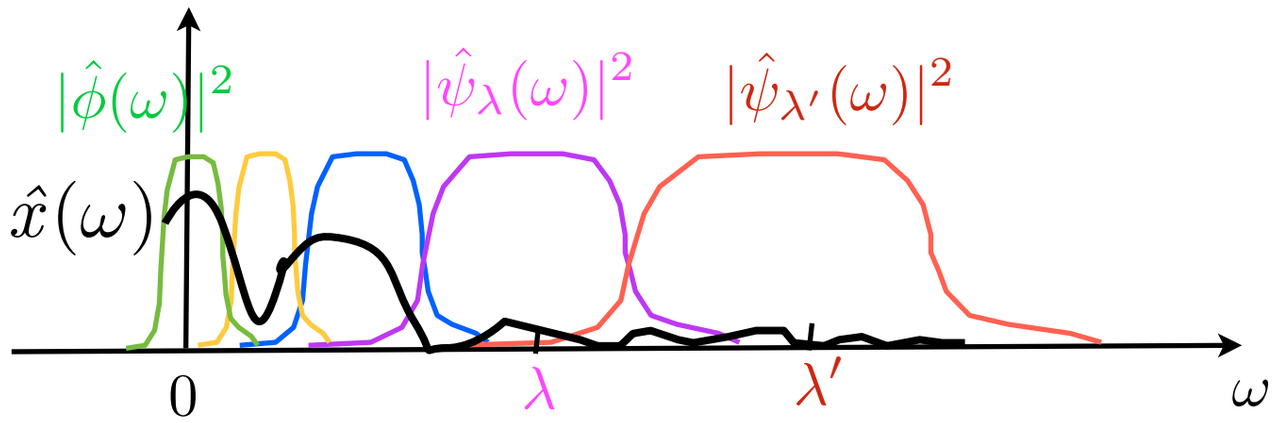
$$\tau(t) = \epsilon\, t$$



# Wavelet Transform

- Complex analytic wavelet: $\psi(t) = \psi^a(t) + i\,\psi^b(t)$
- Dilated: $\psi_\lambda(t) = \alpha^{-j}\,\psi(\alpha^{-j}t)$ with $\lambda = \alpha^{-j}$ .




- Wavelet transform: $x \star \psi_\lambda(t) = \int x(u)\,\psi_\lambda(t-u)\,du$

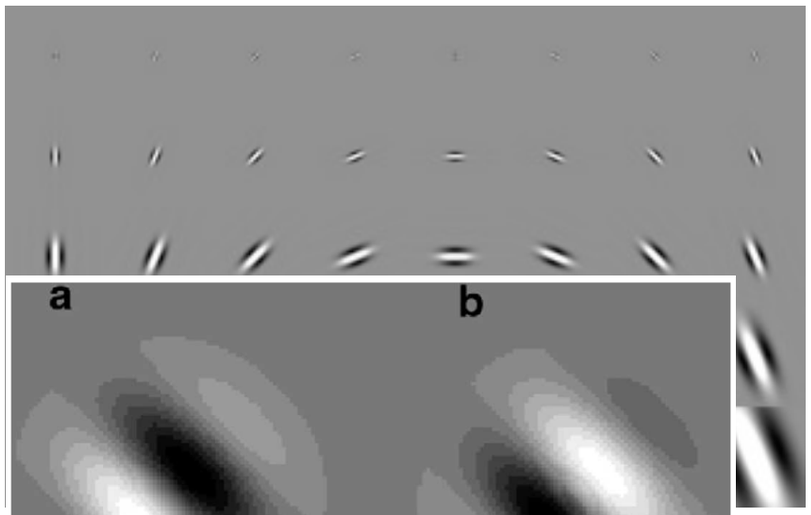
$$Wx = \begin{pmatrix} x \star \phi(t) \\ x \star \psi_\lambda(t) \end{pmatrix}_{t,\lambda}$$

# Image Wavelet Transform

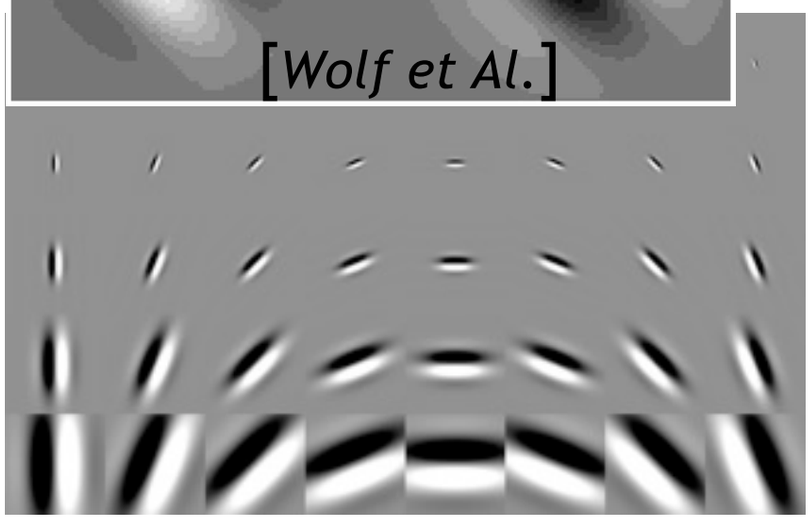
- Complex wavelet: $\psi(t) = \psi^a(t) + i\,\psi^b(t) \quad , \quad t = (t_1, t_2)$

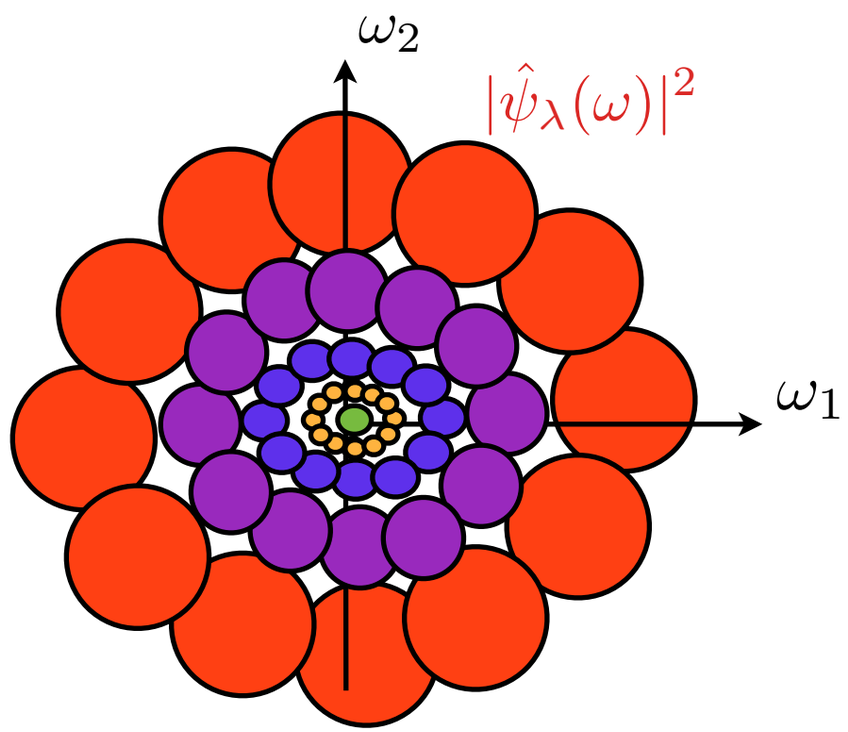
  rotated and dilated: $\psi_\lambda(t) = 2^{-j}\,\psi(2^{-j} r t)$ with $\lambda = (2^j, r)$

real parts




imaginary parts






- Wavelet transform: $Wx = \begin{pmatrix} x \star \phi(t) \\ x \star \psi_\lambda(t) \end{pmatrix}_{t,\lambda}$

# Unitary Wavelet Transforms

$$Wx = \begin{pmatrix} x \star \phi(t) \\ x \star \psi_\lambda(t) \end{pmatrix}_{t,\lambda}$$

Denote $\|x\|^2 = \int |x(t)|^2 \, dt$

**Proposition:** *(Littlewood-Paley)*

The wavelet transform is unitary for $x(t) \in \mathbb{R}$

$$\|Wx\|^2 = \|x \star \phi\|^2 + \sum_\lambda \|x \star \phi_\lambda\|^2 = \|x\|^2$$

if and only if for almost all $\omega$.

$$|\hat{\phi}(\omega)|^2 + \frac{1}{2} \sum_\lambda \left( |\hat{\psi}_\lambda(\omega)|^2 + |\hat{\psi}_\lambda(-\omega)|^2 \right) = 1$$

# Why Wavelets ?

- The wavelet dictionary $\{\psi_\lambda(t-u)\}_{t,\lambda}$ is translation invariant.

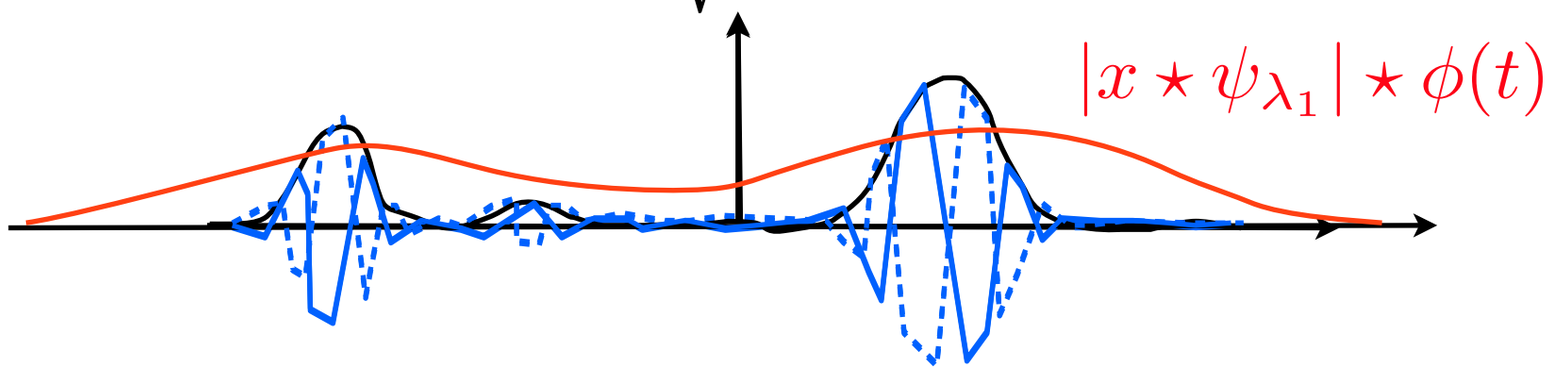
- Wavelets are uniformly stable to deformations:

  if $\psi_{\lambda,\tau}(t) = \psi_\lambda(t-\tau(t))$ then

$$\|\psi_\lambda - \psi_{\lambda,\tau}\| \leq C \sup_t |\nabla\tau(t)| \ .$$

# Wavelet Translation Invariance

$$|x \star \psi_{\lambda_1}| \star \phi(t) \quad \text{pooling}$$




- The modulus $|x \star \psi_{\lambda_1}|$ is a regular envelop

- The average $|x \star \psi_{\lambda_1}| \star \phi(t)$ is invariant to small translations relatively to the support of $\phi$.

- Full translation invariance at the limit:

$$\lim_{\phi \to 1} |x \star \psi_{\lambda_1}| \star \phi(t) = \int |x \star \psi_{\lambda_1}(u)| \, du = \|x \star \psi_{\lambda_1}\|_1$$

but few invariants.

# Wavelet Stabilization

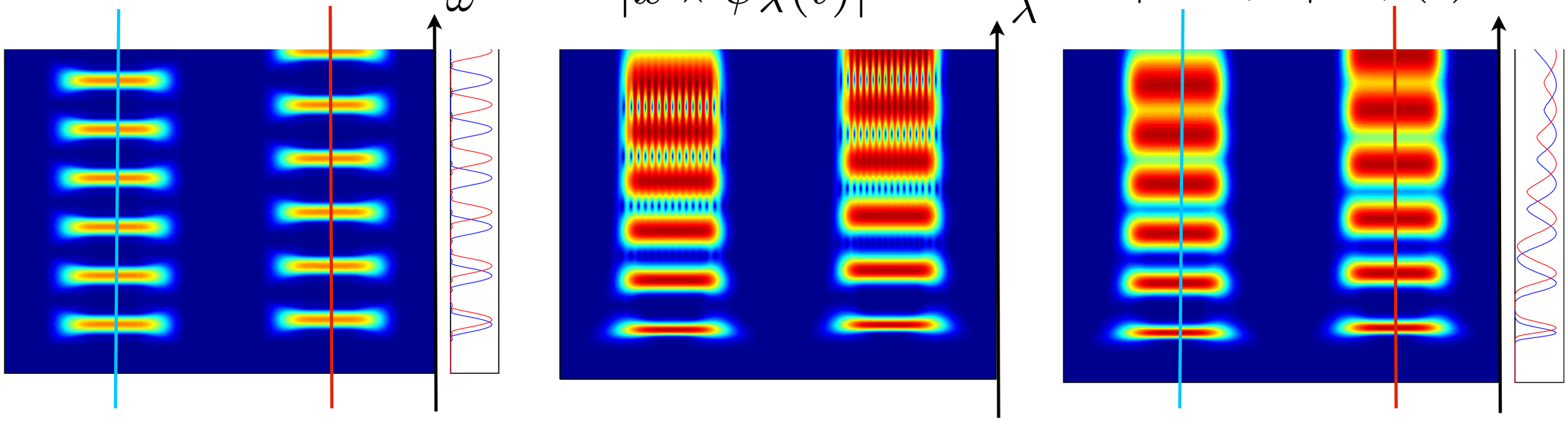
Window Fourier

Wavelet time-frequency

$|x \star \psi_\lambda(t)|$

Time averaging

$|x \star \psi_\lambda| \star \phi(t)$



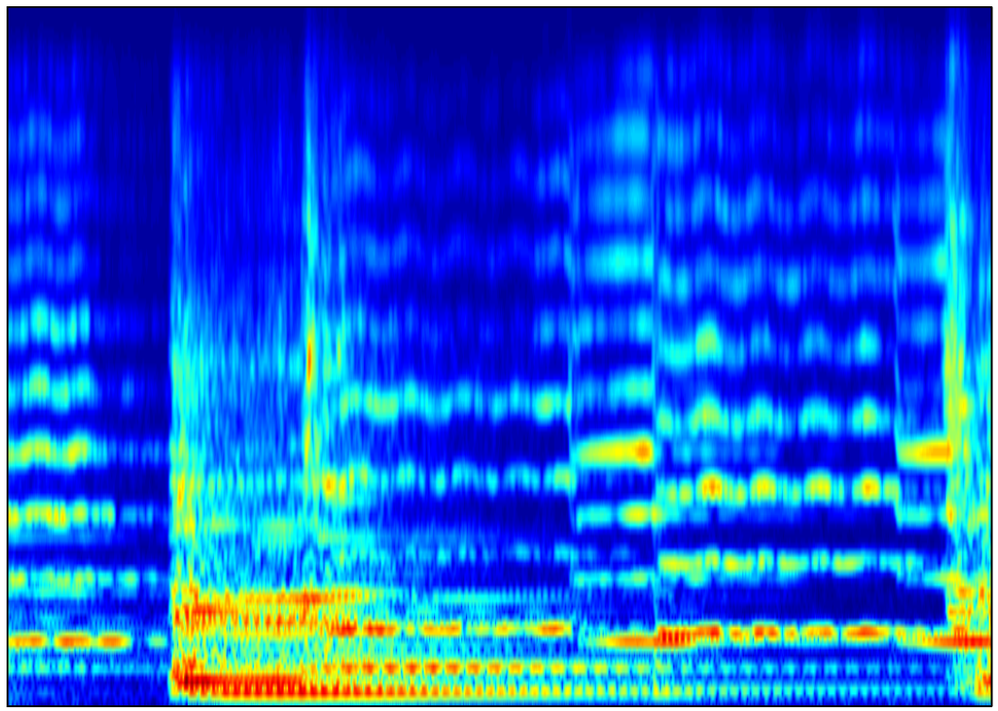

MFSC (audio) on **25ms**

Locally invariant to translations and stable to deformations

# Wavelet Stabilization

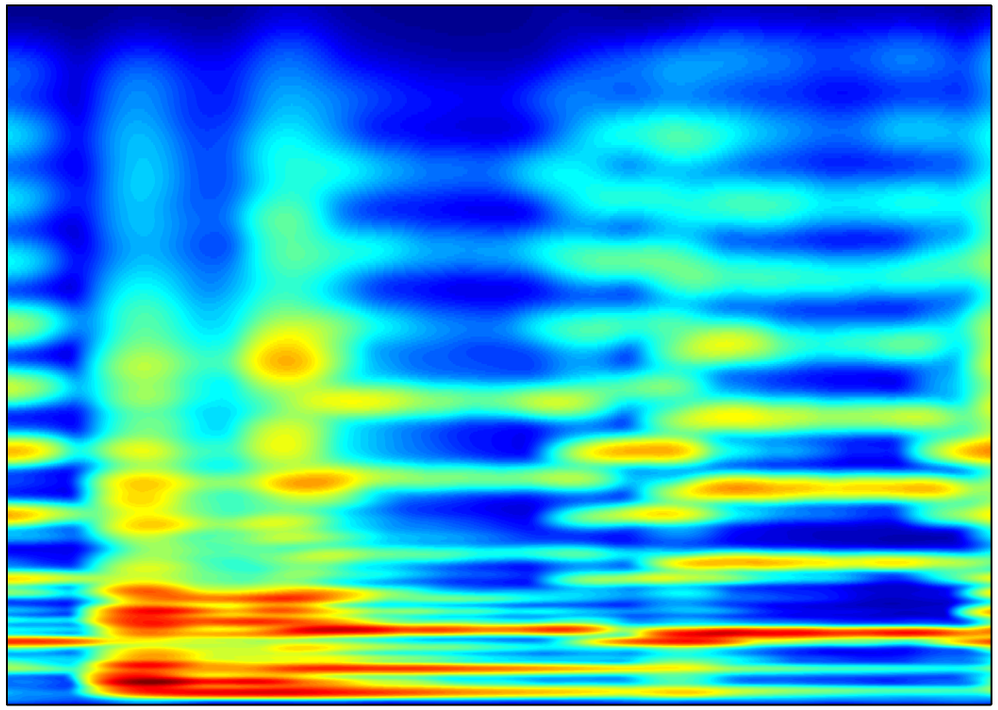
Wavelet time-frequency

$$|x \star \psi_\lambda(t)|$$



Time averaging on **370ms**

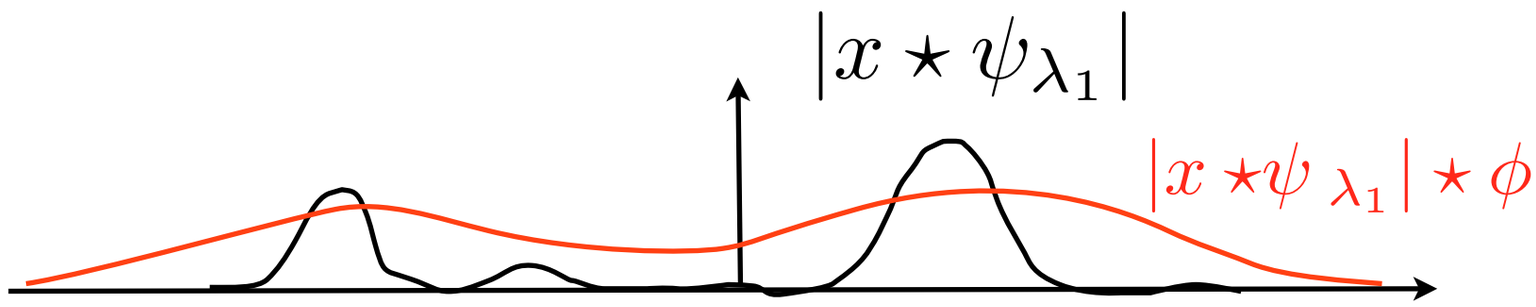
$$|x \star \psi_\lambda| \star \phi(t)$$



Locally invariant to translations and stable to deformations

But loss of information $\Rightarrow$ MFSC (audio) on **25ms** : too small.

# Recovering Lost Information



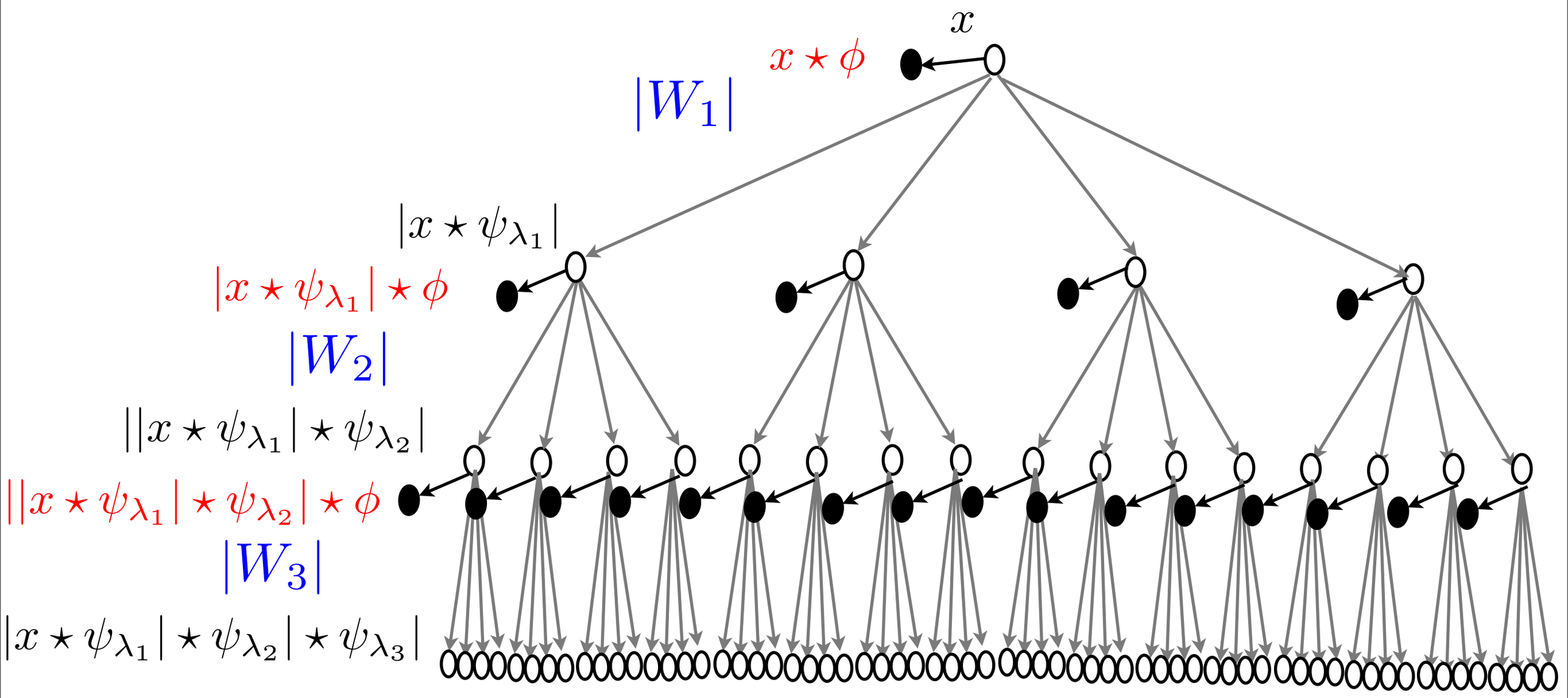

- The high frequencies of $|x \star \psi_{\lambda_1}|$ are in wavelet coefficients:

$$W|x \star \psi_{\lambda_1}| = \begin{pmatrix} |x \star \psi_{\lambda_1}| \star \phi(t) \\ |x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}(t) \end{pmatrix}_{t,\lambda_2}$$

- Translation invariance by time averaging the amplitude:

$$\forall \lambda_1 , \lambda_2 , \quad ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(t)$$

# Deep Convolution Network



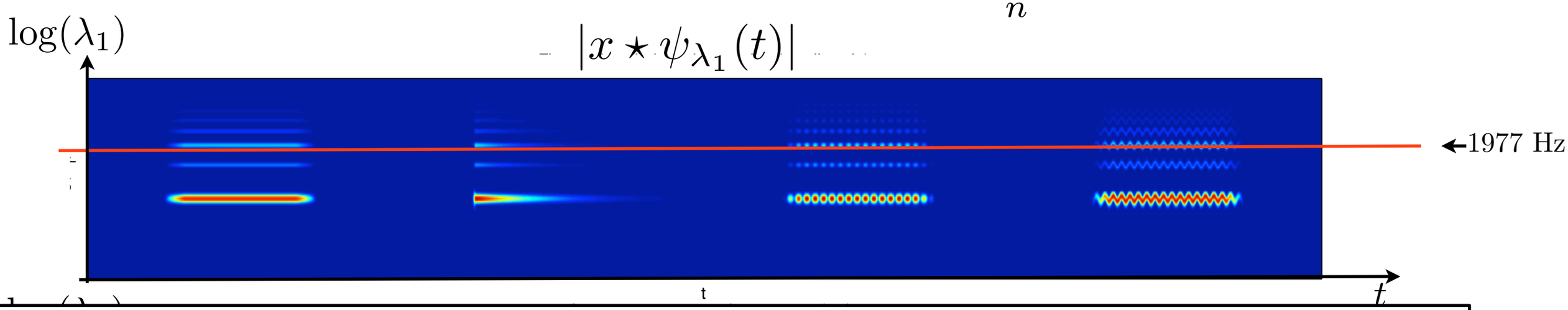
# Scattering Vector

Network ouptut:

$$Sx = \begin{pmatrix} x \star\phi\left(u\right) \\ |x \star \psi_{\lambda_1}| \star \phi(u) \\ ||x \star\psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(u) \\ |||x \star\psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi(u) \\ \cdots \end{pmatrix}_{u,\lambda_1,\lambda_2,\lambda_3,\ldots}$$

# Amplitude Modulation

$$x_i(t) = a_i(t)\Big(c \star h(t)\Big) \quad \text{with} \quad c(t) = \sum_n \delta(t - nT) \ .$$



# Second Order gives Intervals

Music chord :


Minor 3rd

$\omega_3 - \omega_2$

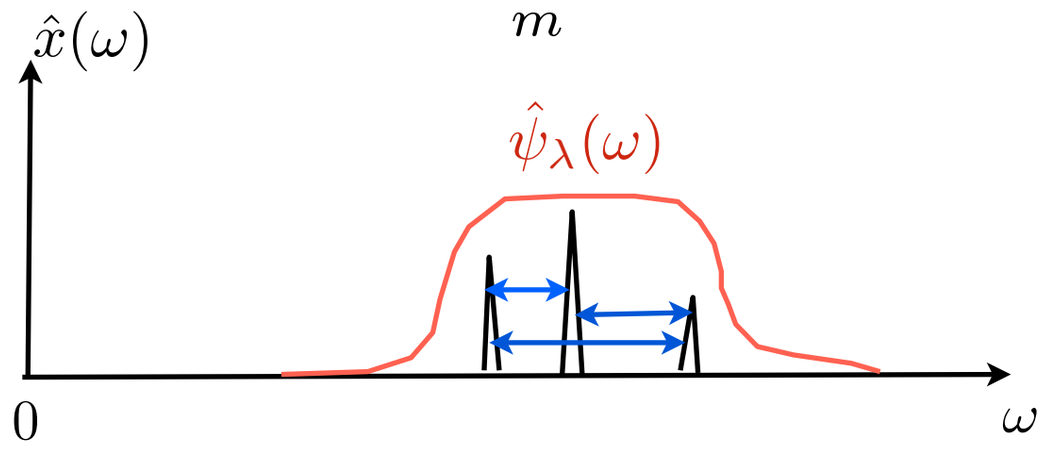
$\omega_2 - \omega_1$

Major 3rd

Perfect 5th

$\omega_3 - \omega_1$

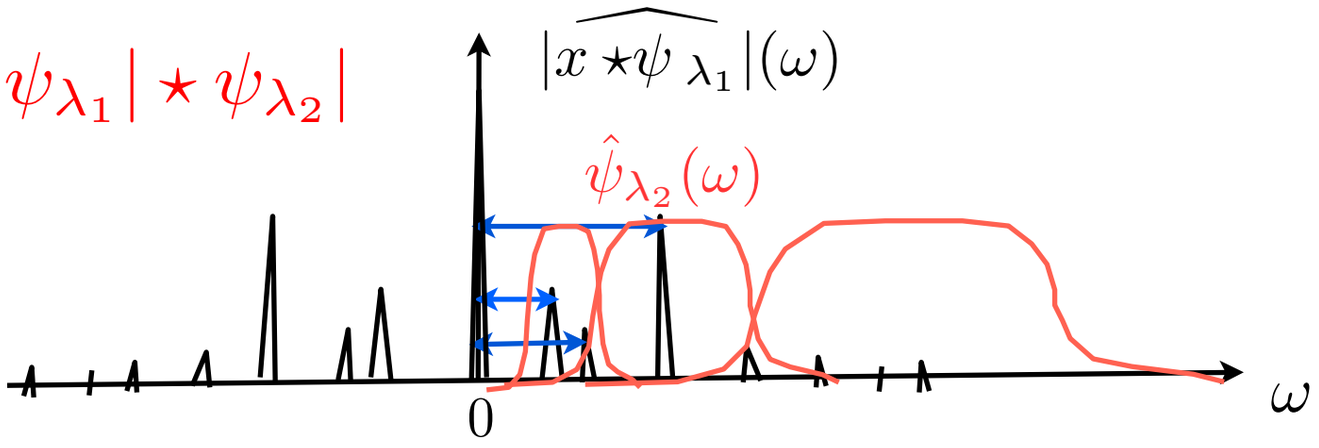
$$x(t) = \sum_m a_m \cos(\omega_m t)$$



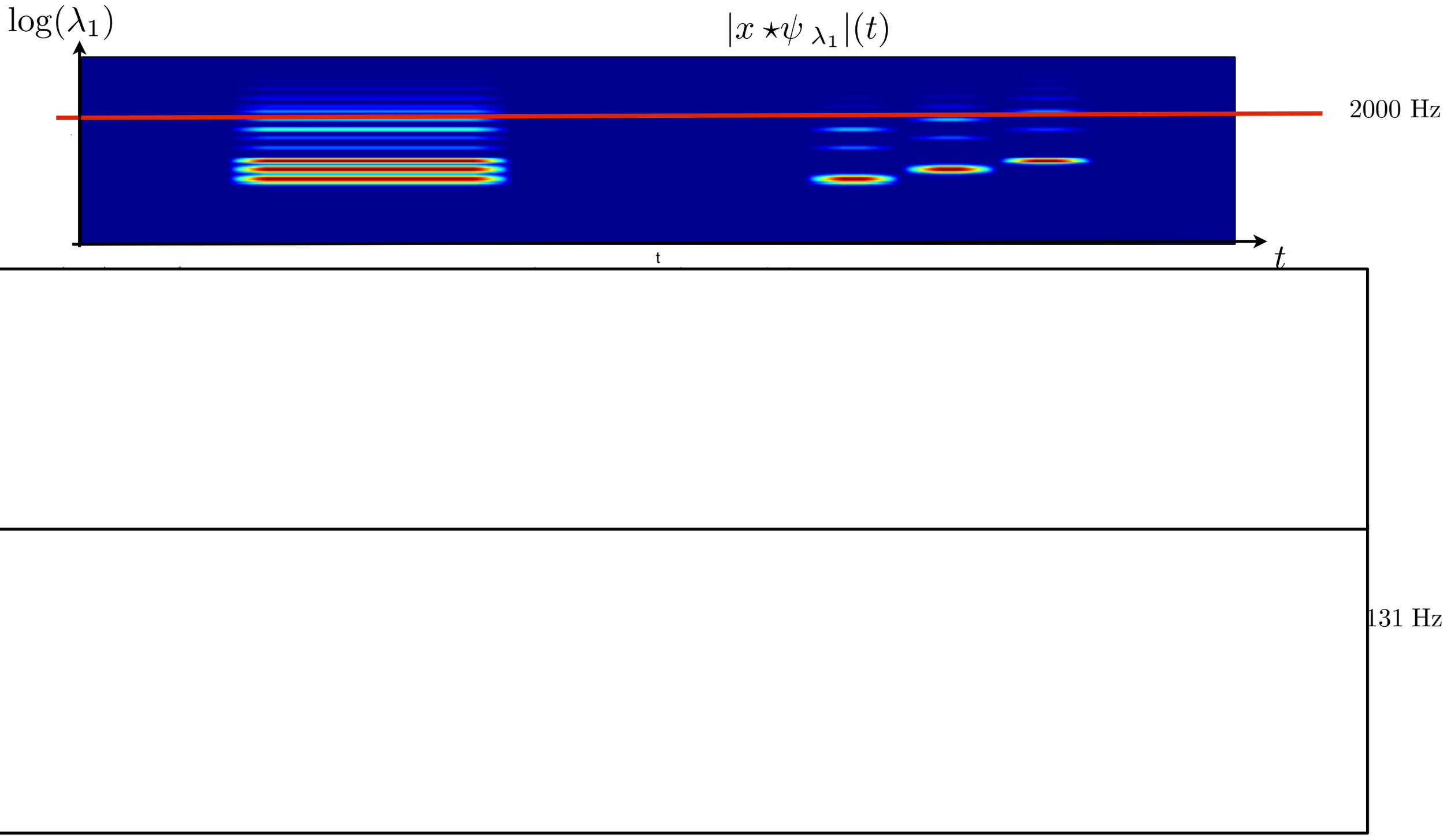

## *Interferences :*

$$|x \star \psi_\lambda(t)|^2 = e_\lambda^2 + \sum_{m' \neq m} c_{m,m'} \cos(\omega_m - \omega_{m'})t$$

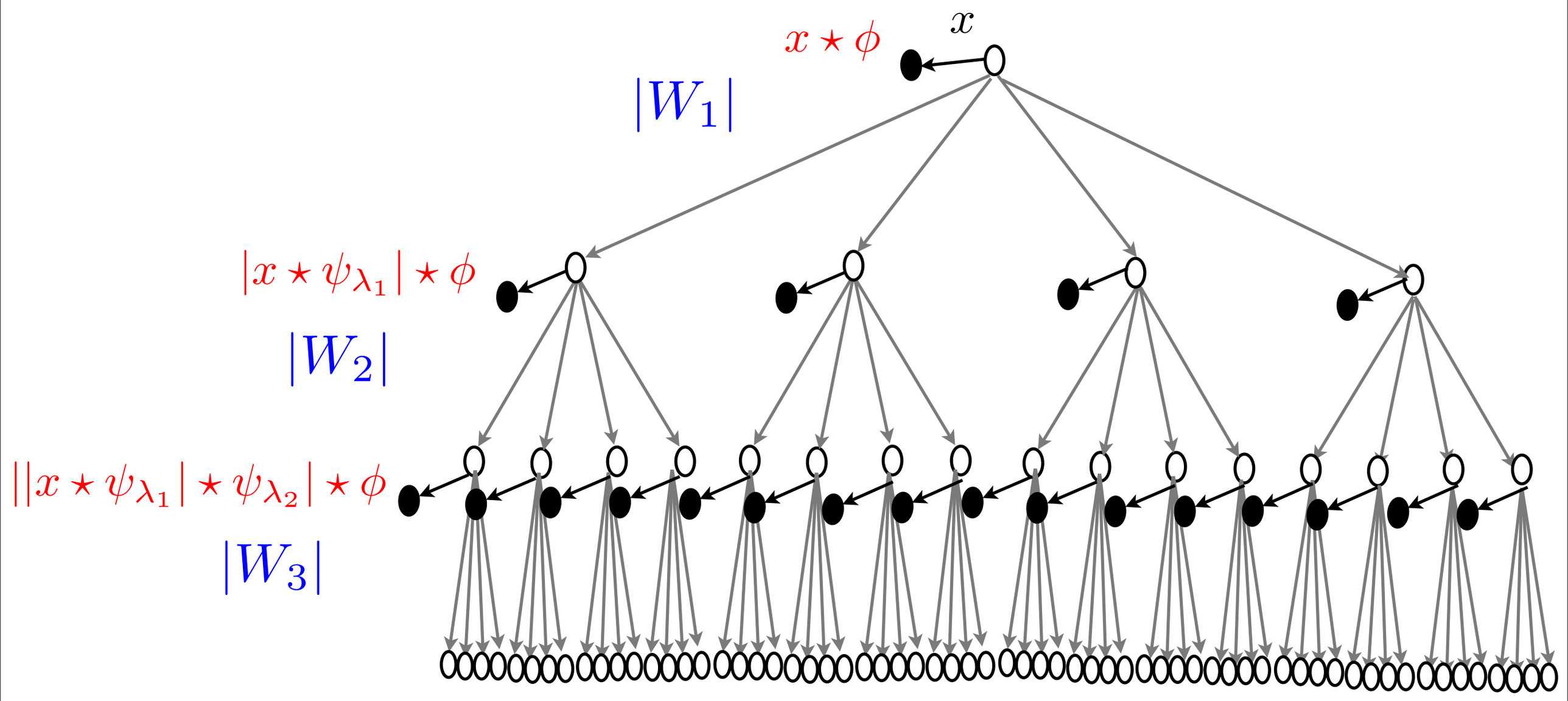
Second order coefficients: $||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|$



# Arpeggio



# Deep Convolution Network



# Scattering Properties

$$Sx = \begin{pmatrix} x \star \phi\,(u) \\ |x \star \psi_{\lambda_1}| \star \phi(u) \\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(u) \\ |||x \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi(u) \\ \dots \end{pmatrix}_{u,\lambda_1,\lambda_2,\lambda_3,\dots}$$

$$\|Sx\|^2 = \sum_{m=0}^{\infty} \sum_{\lambda_1,\dots,\lambda_m} \left\| |||x \star \psi_{\lambda_1}| \star \dots| \star \psi_{\lambda_m}| \star \phi \right\|^2$$

**Theorem**: *For appropriate wavelets, a scattering is*

*contractive* $\|Sx - Sy\| \leq \|x - y\|$

*preserves norms* $\|Sx\| = \|x\|$

*stable to deformations* $x_\tau(t) = x(t - \tau(t))$

$$\|Sx - Sx_\tau\| \leq C \sup_t |\nabla \tau(t)|\, \|x\|$$

# Contraction

$$Wx = \begin{pmatrix} x \star \phi(t) \\ x \star \psi_\lambda(t) \end{pmatrix}_{t,\lambda} \quad \text{is linear and } \|Wx\| = \|x\|$$

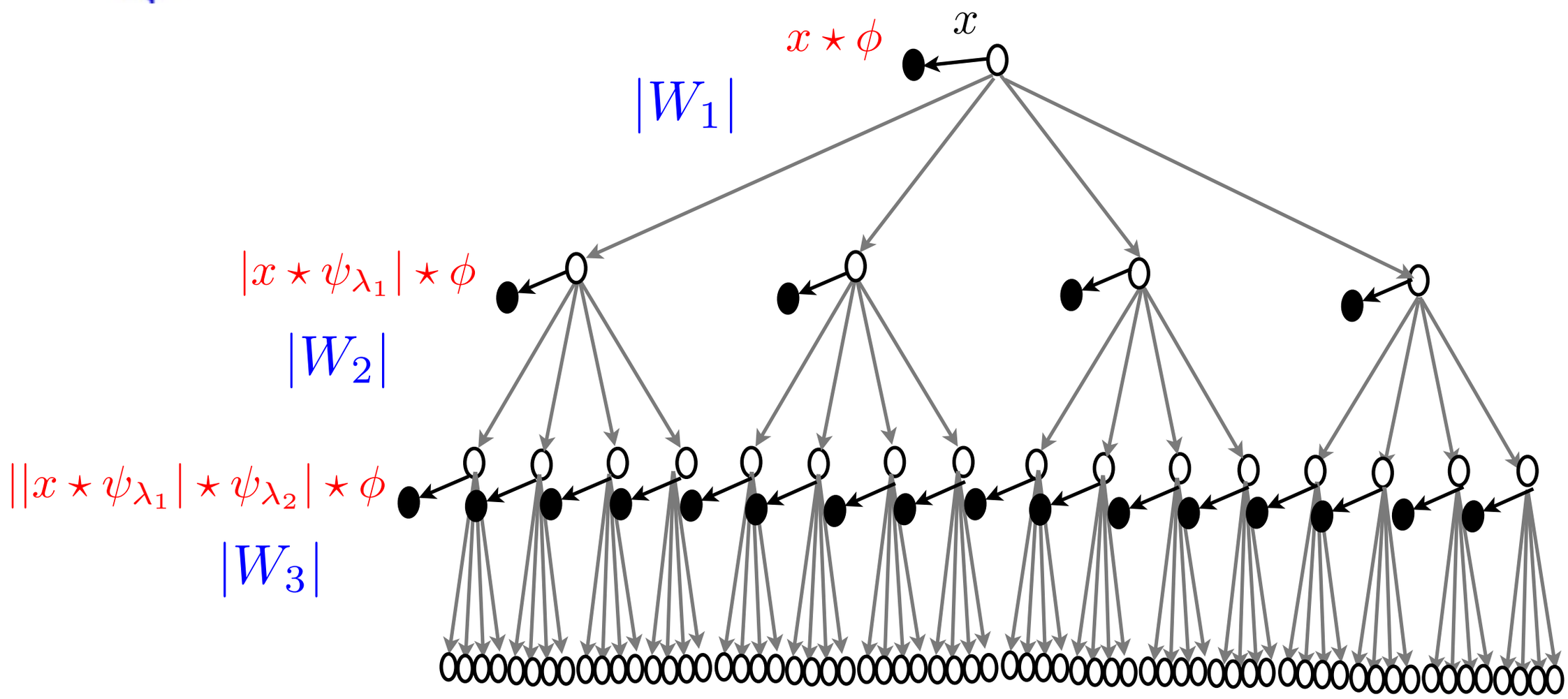
$$|W|x = \begin{pmatrix} x \star \phi(t) \\ |x \star \psi_\lambda(t)| \end{pmatrix}_{t,\lambda} \quad \text{is non-linear}$$

- it preserves the norm $\||W|x\| = \|x\|$

- it is contractive $\||W|x - |W|y\| \leq \|x - y\|$

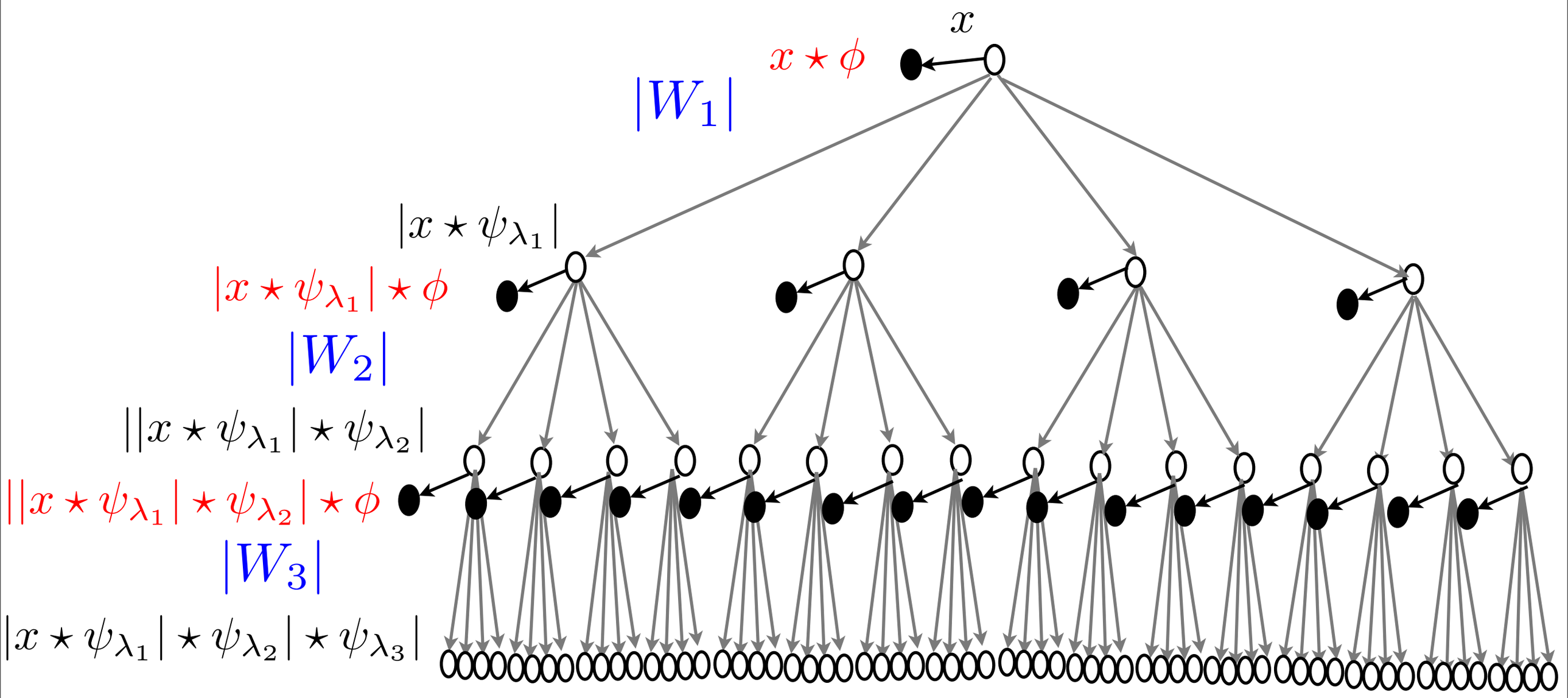
because for $(a, b) \in \mathbb{C}^2$ $||a| - |b|| \leq |a - b|$

# Scattering Contraction




- $S$ is contractive because product of contractive operators.

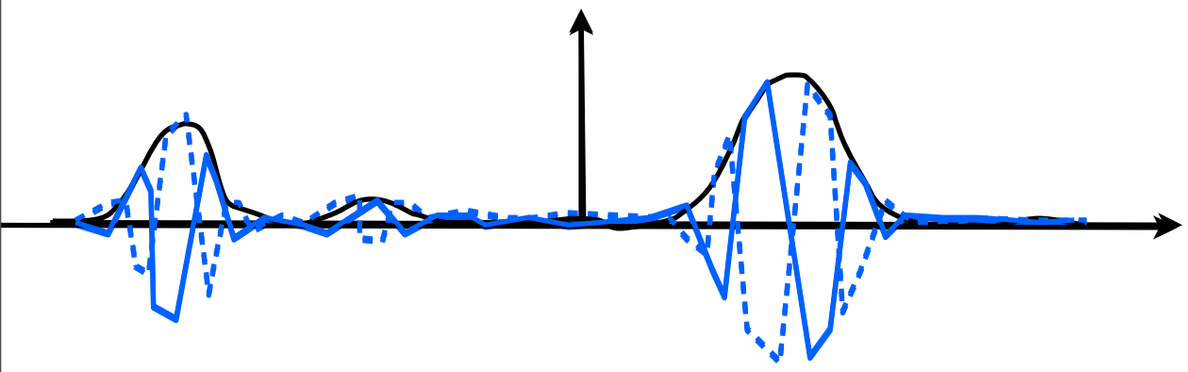
# Scattering Energy Conservation

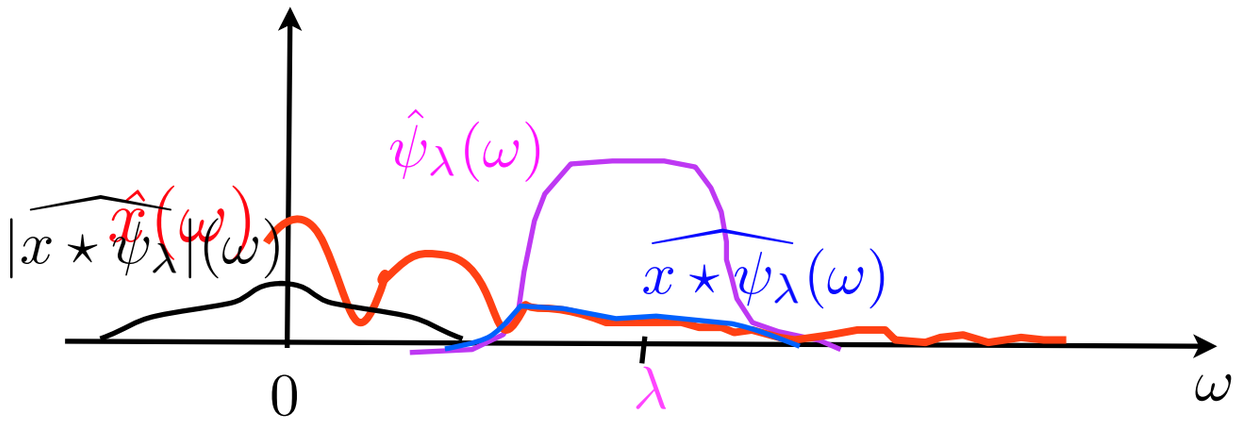



- $S$ preserves the norm because inner layer energy converge to zero as the depth increases.

# Modulus «Demodulation»

$$x \star \psi_{\lambda_1}(t) = x \star \psi^a_{\lambda_1}(t) + i\, x \star \psi^b_{\lambda_1}(t)$$






- The modulus $|x \star \psi_{\lambda_1}|$ is a regular lower frequency envelop

Modulus shift wavelet coefficient energy to low frequencies.

# Lipschitz Stability to Deformations

Wavelet transforms "nearly commute" with deformations:

$$D_\tau x(t) = x(t - \tau(t))$$

Commutator operator:

$$[W, D_\tau] = W\, D_\tau - D_\tau\, W$$

**Lemma :**

$$\| [W, D_\tau] \| \leq C \sup_t |\nabla \tau(t)| \ .$$

and $\| [|W|, D_\tau] \| \leq \| [W, D_\tau] \|$

because modulus commutes with diffeomorphisms.

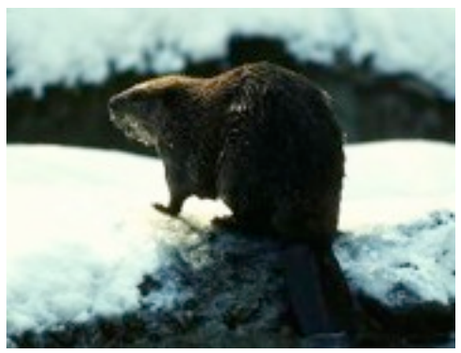
# Part II: High Dimensional Classification

*CalTech 101*

Anchor

Joshua Tree

Beaver



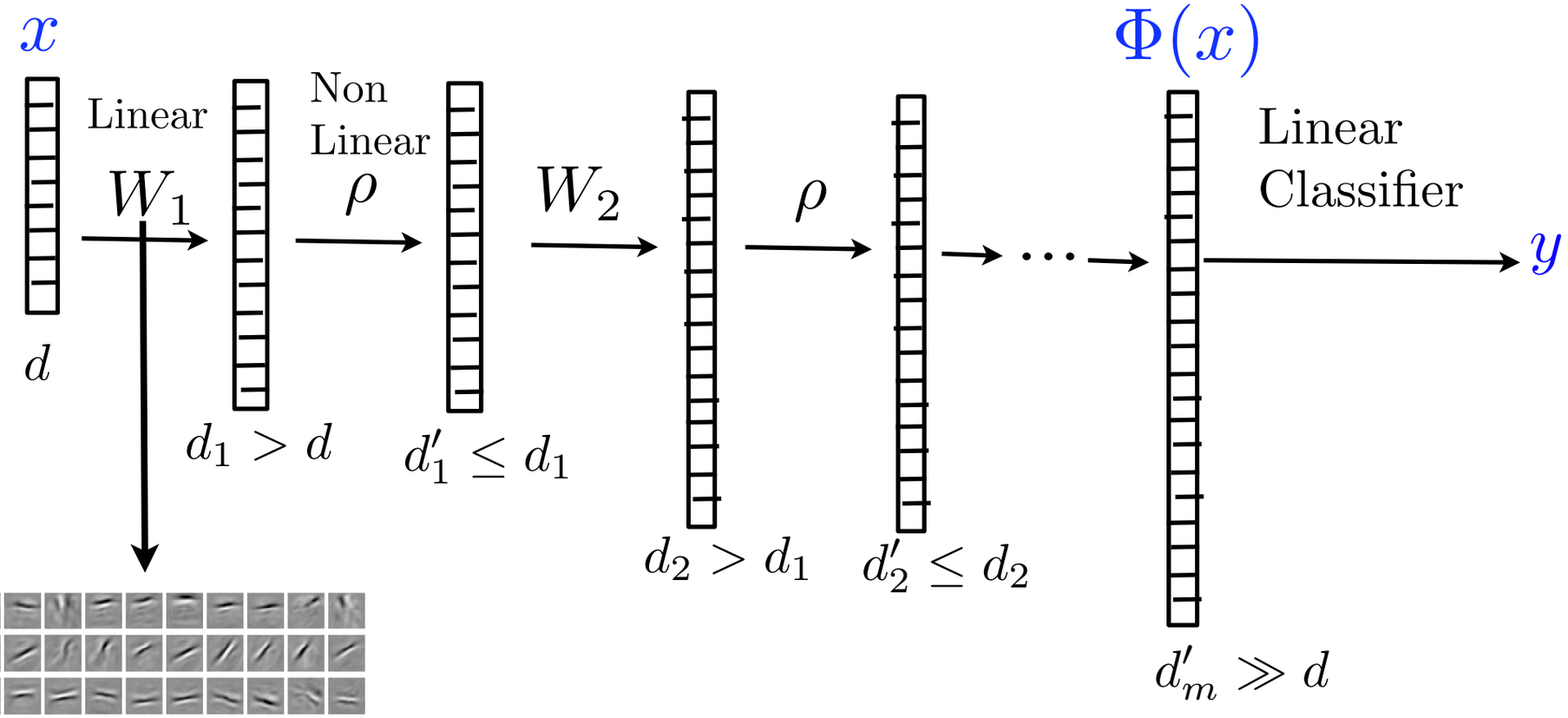
Lotus

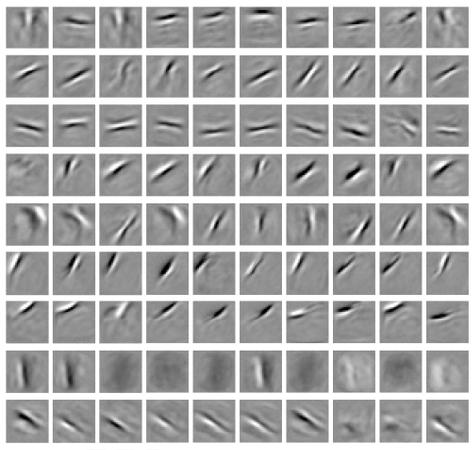
Water Lily

- Considerable variability in each class.
- Euclidean distances are meaningless
- Need to find **discriminative invariants**.

# Deep Neural Networks

*J. Hinton, Y. LeCun* "State of the art results"



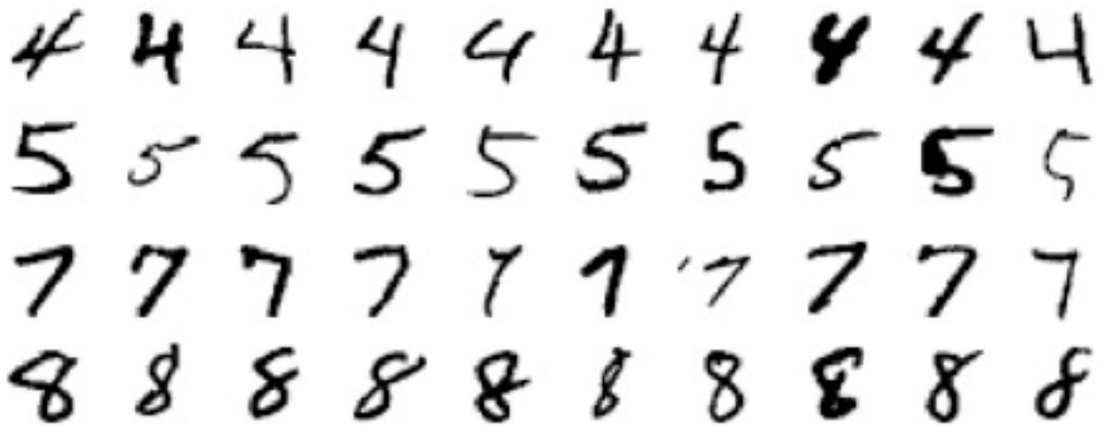
Wavelets

The $W_k$ are learned with a sparsity criteria

Why and how does it work ?

# Translations and Deformations

- Patterns are translated and deformed (class dependent)



Group: $\mathbb{R}^2 \times \mathrm{Diff}(\mathbb{R}^2)$

- **Invariance** to translations $x_c(t) = x(t-c)$

$$\forall c \in \mathbf{R} \quad , \quad \Phi(x_c) = \Phi(x) \ .$$

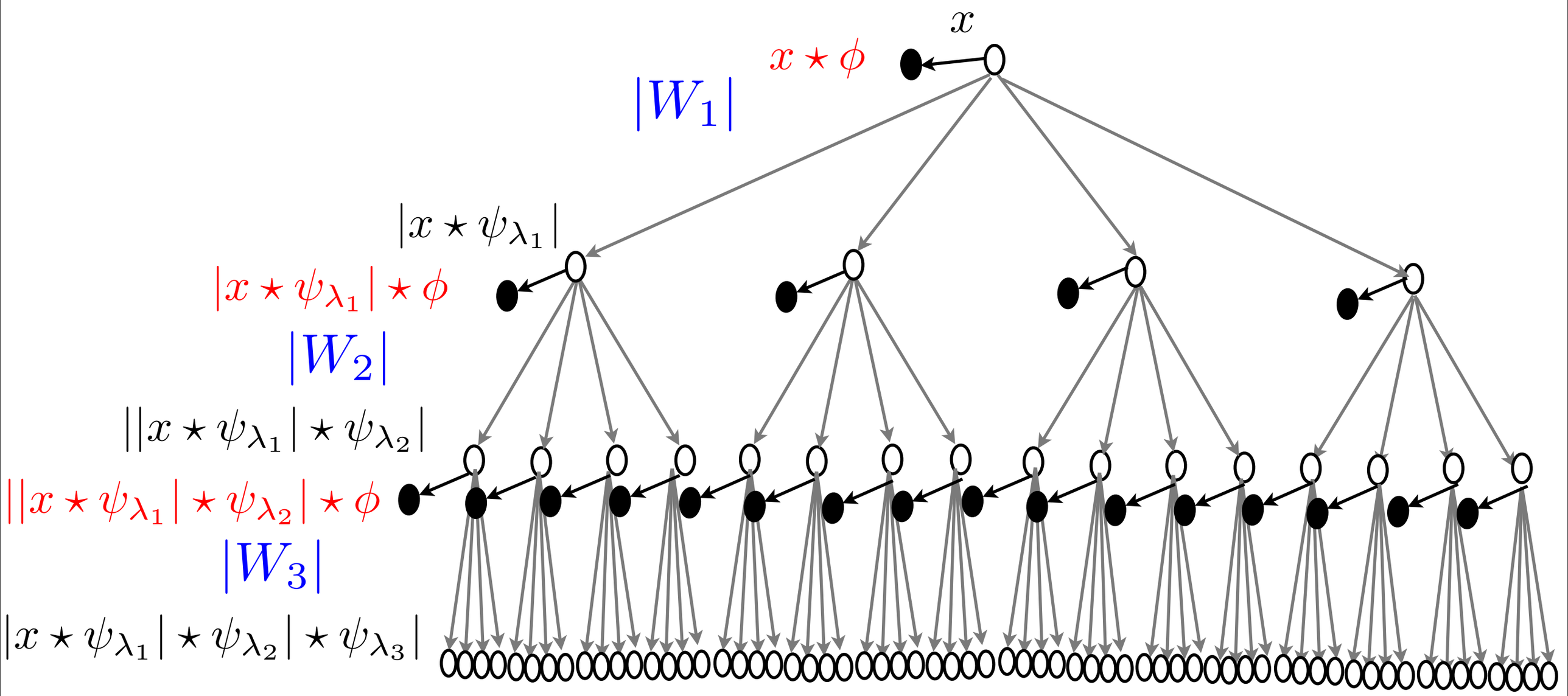
- **Lipschitz stable** to diffeomorphisms $x_\tau(t) = x(t - \tau(t))$

$$\forall \tau \quad , \quad \|\Phi(x_\tau) - \Phi(x)\| \leq C \sup_t |\nabla \tau(t)| \, \|x\| \ .$$

Fourier Failure

diffeomorphism metric

# Local Scattering Transform



# Scattering Properties

$$Sx = \begin{pmatrix} x \star \phi\,(u) \\ |x \star \psi_{\lambda_1}| \star \phi(u) \\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(u) \\ |||x \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi(u) \\ \dots \end{pmatrix}_{u,\lambda_1,\lambda_2,\lambda_3,\dots}$$

$$\|Sx\|^2 = \sum_{m=0}^{\infty} \sum_{\lambda_1,\dots,\lambda_m} \left\| |||x \star \psi_{\lambda_1}| \star \dots| \star \psi_{\lambda_m}| \star \phi \right\|^2$$

**Theorem**: *For appropriate wavelets, a scattering is*

*contractive* $\|Sx - Sy\| \leq \|x - y\|$

*preserves norms* $\|Sx\| = \|x\|$

*stable to deformations* $x_\tau(t) = x(t - \tau(t))$

$$\|Sx - Sx_\tau\| \leq C \sup_t |\nabla \tau(t)|\, \|x\|$$

# Part II : Overview

- ***Deterministic Scattering transform***

  Limit scattering integral

  Inversion

  Image classification application

- ***High-dimensional stochastic models***

  Scattering models of stationary processes

# Fourier versus Scattering

$e^{i\xi t}\, x(t)$

Frequencies $\omega = m\xi$

$$e^{im\xi t}x(t) = e^{i\xi t} \ldots e^{i\xi t} e^{i\xi t} x(t)$$

Countable frequency set

Local Fourier:

$$\int e^{im\xi u}\, x(u)\, \phi(t-u) du$$

$\hat{\phi}(\omega)$ in $[-\xi, \xi]$

Fourier transform:

$$\hat{x}(\omega) = \int e^{i\omega u}\, x(u)\, du$$

$$\hat{\delta}(\omega) = 1$$

Frequency set: $\mathbb{R}$

$|x \star \psi_\lambda(t)|$ with $\lambda = 2^j \geq \xi$

Paths $p = (\lambda_1, \lambda_2 ..., \lambda_m)$

$$|\, |x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \ldots | \star \psi_{\lambda_m}(t)|$$

Countable path set

Local scattering:

$$\int |\, |x \star \psi_{\lambda_1}| \ldots | \star \psi_{\lambda_m}(u)|\, \phi(t-u)\, du\ .$$

$\phi \longrightarrow 1$

$\xi \longrightarrow 0$

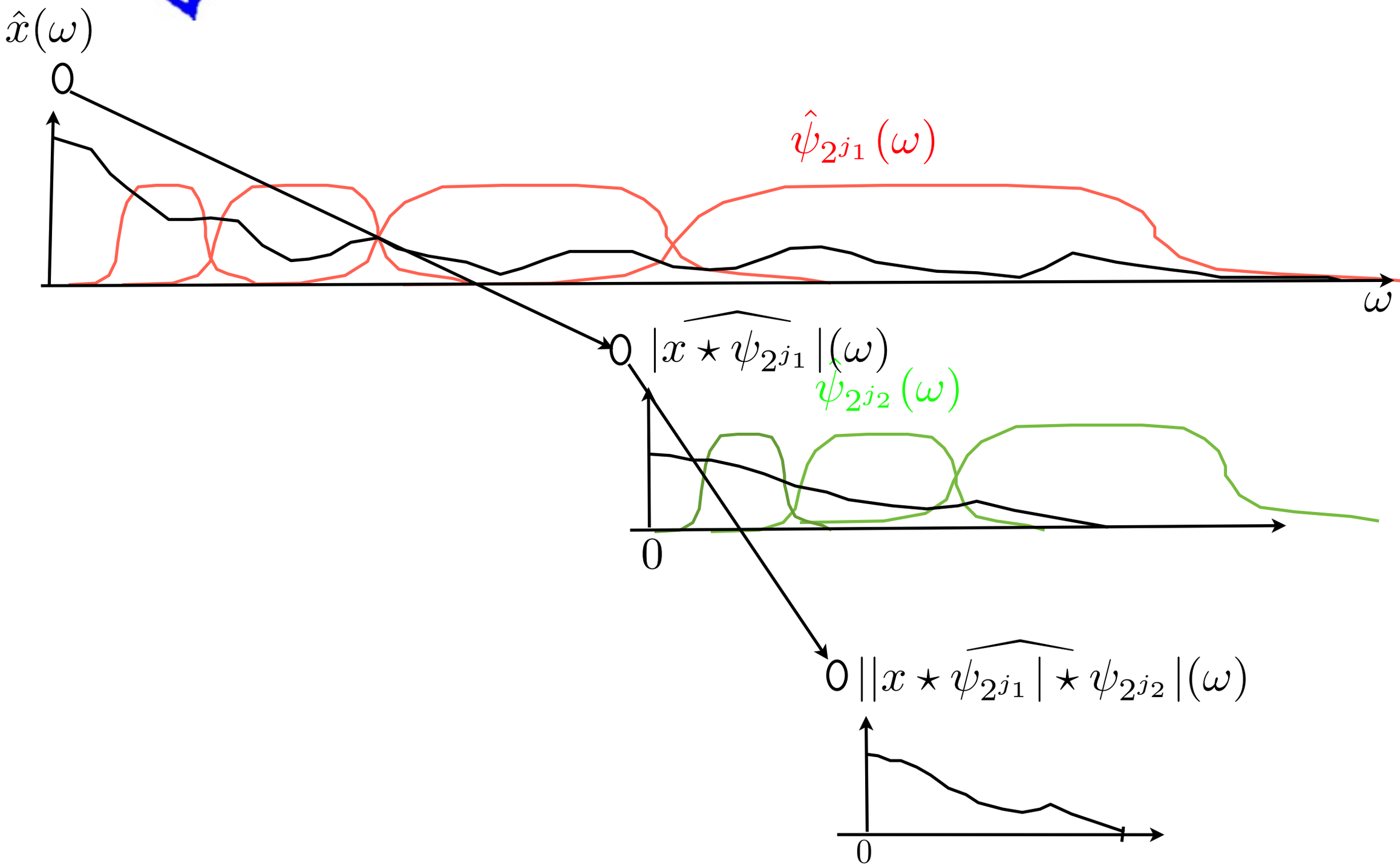
Scattering transform:

$$\overline{S}x(p) = \mu_p^{-1} \int |\, |x \star \psi_{\lambda_1}| \ldots | \star \psi_{\lambda_m}(u)|\, du$$

$$\overline{S}\delta(p) = 1$$

Path set $\mathcal{P} \sim \mathbb{Z}^{\mathbb{N}} \sim \mathbb{R}$

# Frequency Subdivision




$p = (2^{j_1}, 2^{j_2}, 2^{j_3}, ...)$ yields a non-linear frequency subdivision.

# Limit Scattering Transform

- Give a meaning to $\lim_{\phi \to 1} S = \overline{S}$.

- $S$ is defined on finite paths: $p = \left(2^{j_k}\right)_{k \leq m}$ with $2^{j_k} \geq \xi$.

  Countable path set.

  $\hat{\phi}(\omega)$ in $[-\xi, \xi]$

- $\bar{S}$ is defined on inifinite paths: $p = \left(2^{j_k}\right)_{k \in \mathbb{N}}$ with $2^{j_k} \geq 0$.

  Path set $\mathcal{P} = \mathbb{N}^{\mathbb{Z}} \sim \mathbb{R}$.

- Must define a measure $d\mu(p)$ on $\mathcal{P}$ hence a $\sigma-$algebra.

  finite path p = cylinder set of infinite path beginning by p.

  $d\mu(p)$: scattering mass of a Dirac on a cylinder set.

# Scattering Transform

**Theorem** *$S$ converges weakly to $\overline{S}$ when $\phi$ goes to 1*

*There exists a measure $d\mu$ on $\mathcal{P}$ such that*

$$\forall x \in \mathbf{L^2(R)} \quad , \quad \overline{S}x(p) \in \mathbf{L^2}(\mathcal{P}, d\mu)$$

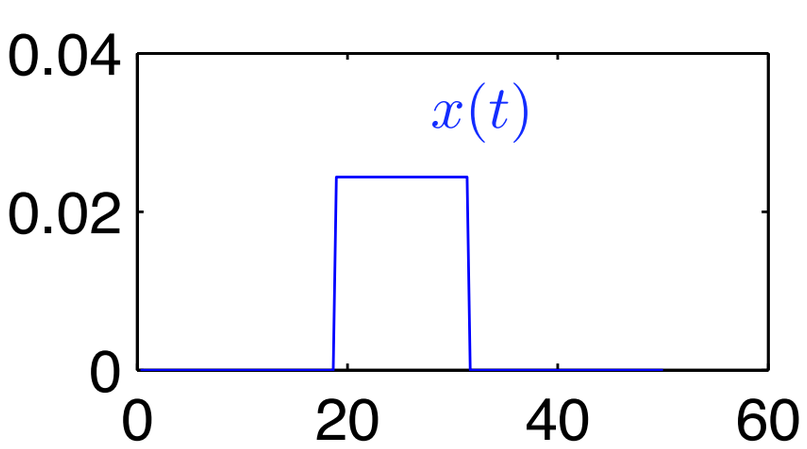
$$\int_{\mathcal{P}} |\overline{S}x(p)|^2 \, d\mu(p) < \infty \ .$$

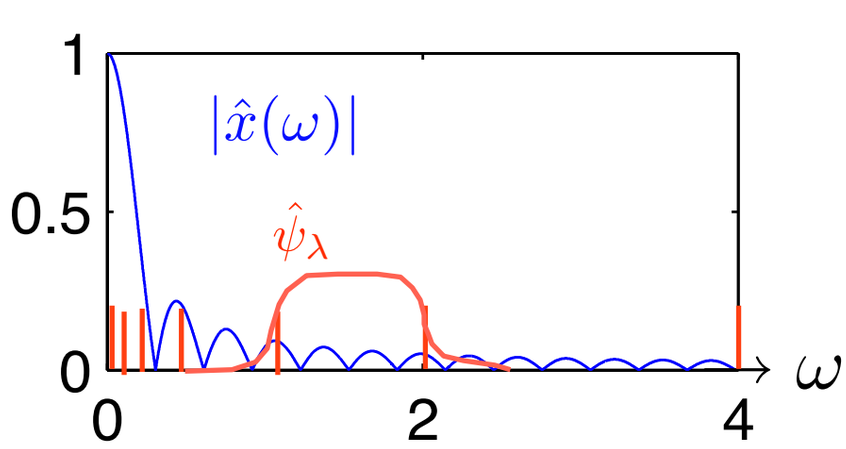
We know that $\|Sx\|^2 = \|x\|^2$ and $\lim_{\phi \to 1} S = \overline{S}$

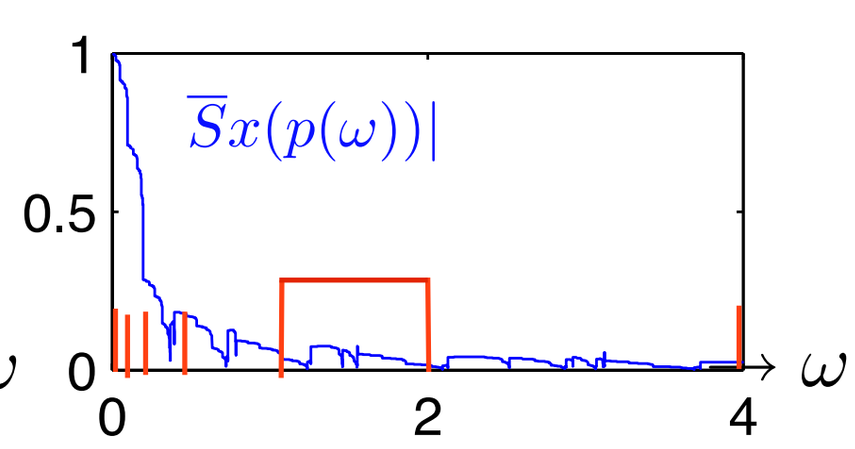
**Conjecture:** $\displaystyle\int_{\mathcal{P}} |\overline{S}x(p)|^2 \, d\mu(p) = \|x\|^2 \ .$

# Frequency to Paths Mapping

$$\begin{aligned}
\mathbb{R} &\longrightarrow \mathcal{P} \\
\omega &\longrightarrow p(\omega) \quad \text{with } d\mu(p(\omega)) = d\omega \\
\omega &\longrightarrow \overline{S}x(p(\omega))
\end{aligned}$$



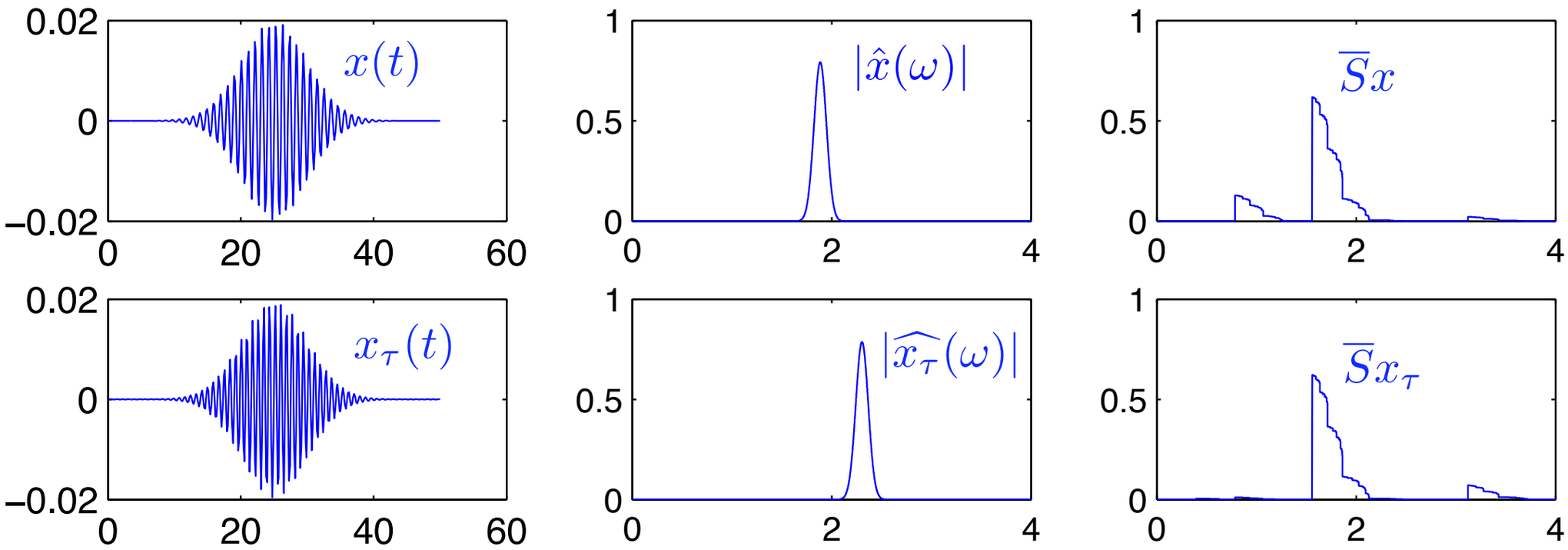




$$\int |\hat{x}(\omega)|^2 \, |\hat{\psi}(2^j\omega)|^2 \, d\omega = \int_{2^j\pi}^{2^{j+1}\pi} |\overline{S}x(p(\omega))|^2 \, d\omega$$

# Scattering Integral Examples

$$x_\tau(t) = x(t - \tau(t)) \quad \text{with} \quad \tau(t) = \epsilon\, t \ .$$




$$\frac{\| \, |\hat{x}| \ast \hat{\psi} \, |\hat{x}_\tau| \, \|}{\|x\| \, \|\tau'\|_\infty} = 13 \qquad \frac{\|\overline{S}x - \overline{S}x_\tau\|_{\mathcal{P}}}{\|x\| \, \|\tau'\|_\infty} = 1.4$$

Fourier transforms maps regularity and decay and vice-versa.

What notion of regularity defined by the scattering decay ?

Depends on the sparsity/geometry of wavelet coefficients.

# Image Scattering Transforms

Image

$x(t)$

$t = (t_1, t_2)$

Fourier Modulus

$|\hat{x}(\omega)|$

$\omega = (\omega_1, \omega_2)$

Scattering $\phi(t) = 1$

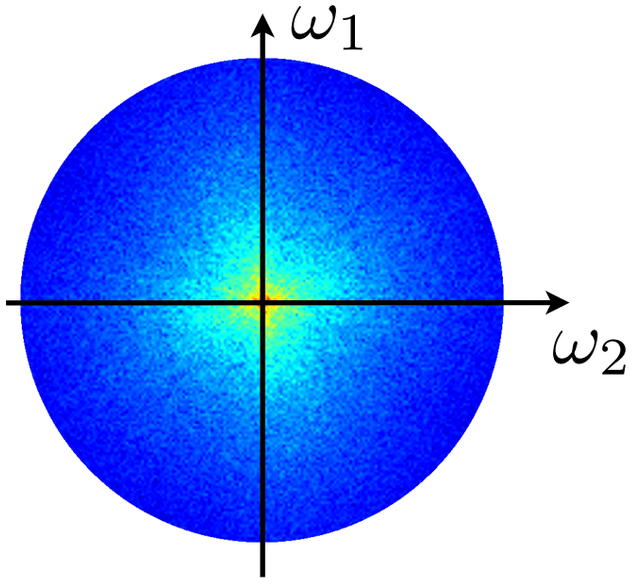
$|x \star \psi_{\lambda_1}| \star \phi$ $\quad$ $||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi$

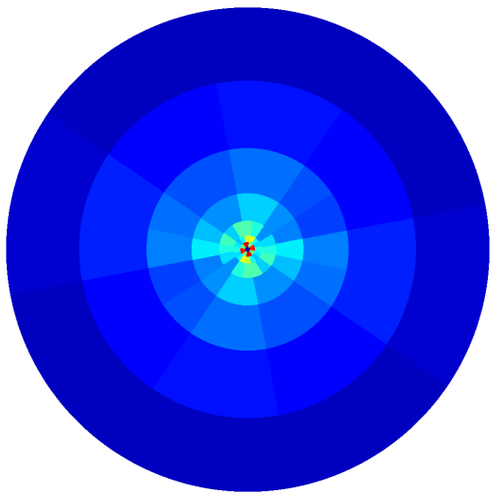
$\|x \star \psi_{\lambda_1}\|_1$ $\quad$ $\||x \star \psi_{\lambda_1}| \star \psi_{2^{j_2}}\|_1$

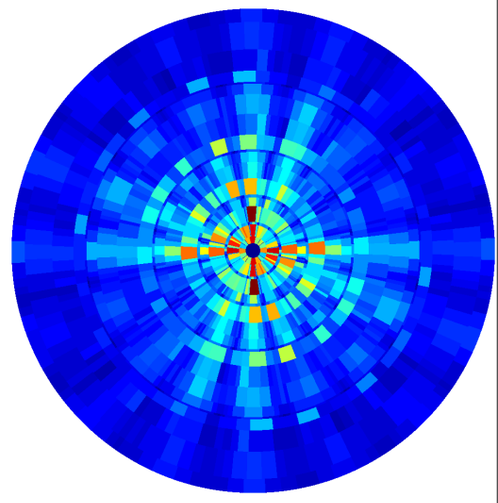
$\lambda_1 = 2^{j_1} r_{\theta_1}$

$\lambda_1 = 2^{j_1} r_{\theta_1}$

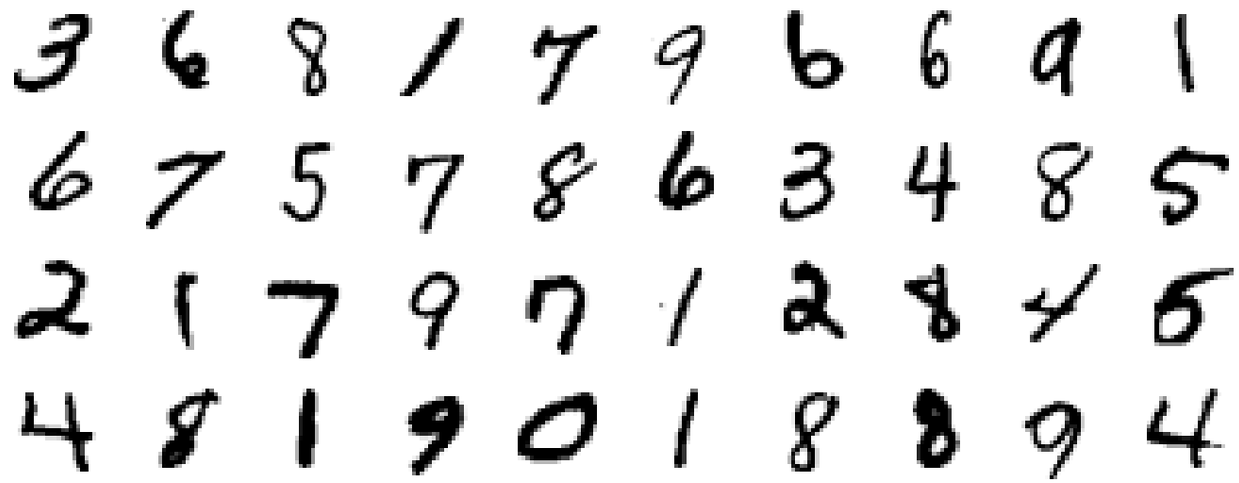
$\lambda_2 = 2^{j_2} r_{\theta_2}$



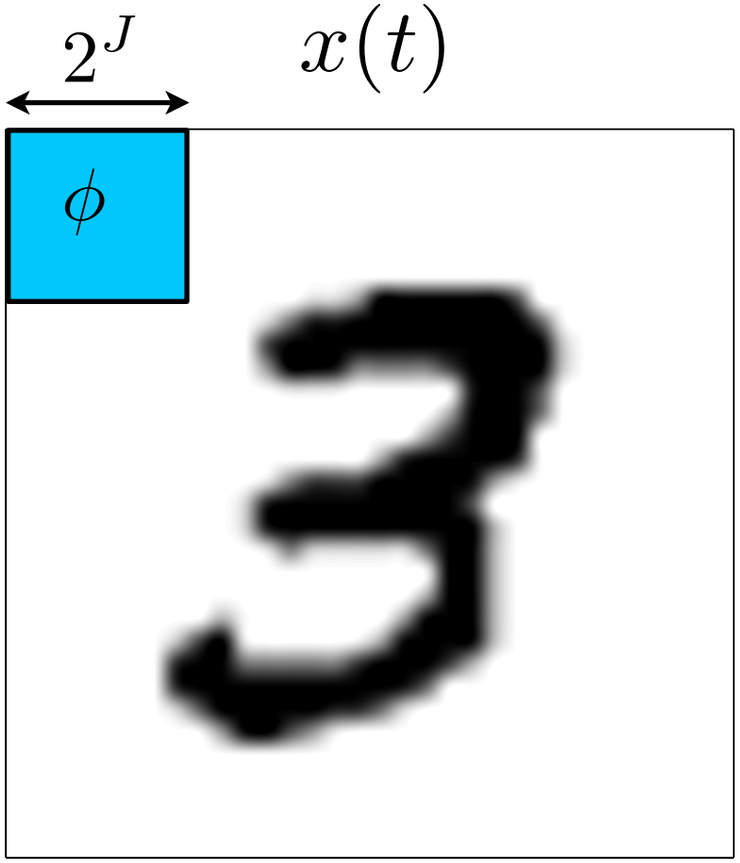


# Digit Classification: MNIST

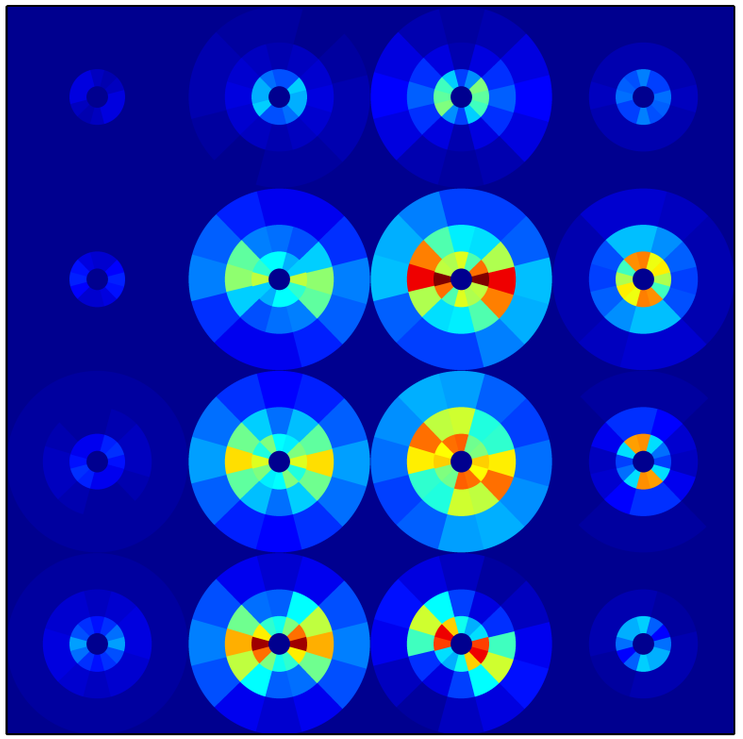


# Digit Classification: MNIST

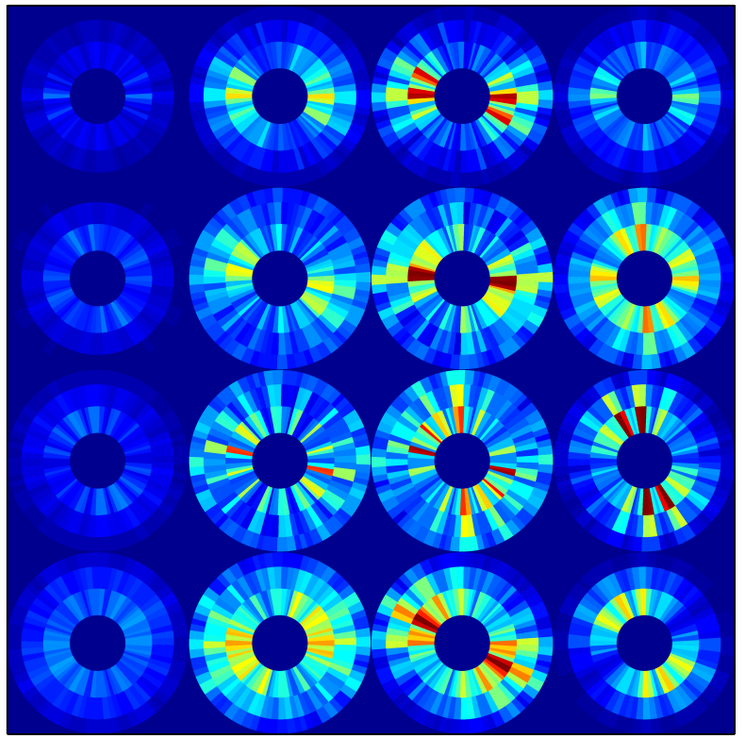
Second order Scattering $Sx$:




$|x \star \psi_{\lambda_1}| \star \phi(2^J n)$



$||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(2^J n)$



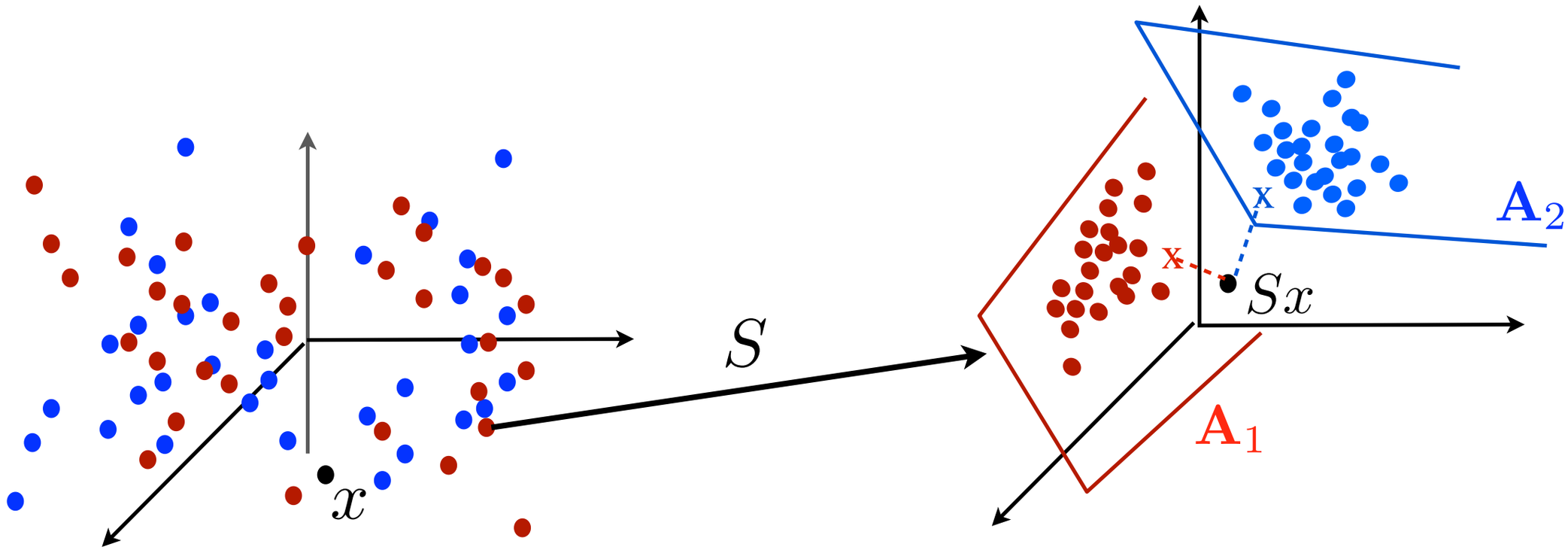
# Affine Space Classification

*Joan Bruna*

Each image is represented by its scattering tranfsorm

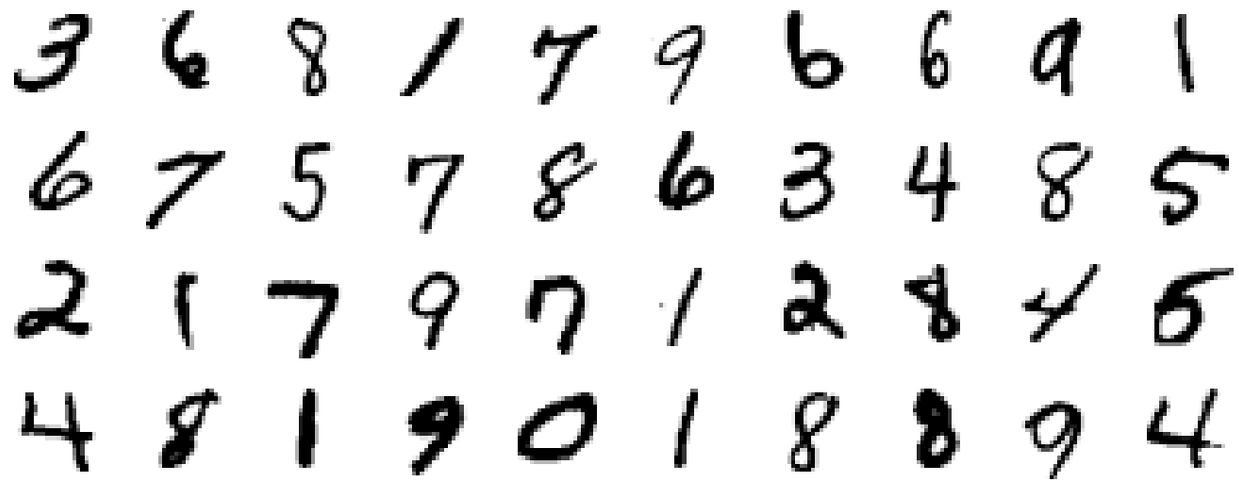
Each class $\mathcal{C}_k$ is approximated by an affine space $\mathbf{A}_k$

computed from examples with a Principal Component Analysis



# Digit Classification: MNIST

*Joan Bruna*



Classification Errors

| Training size | Conv. Net. | Scattering |
|---|---|---|
| 300 | 7.2% | **4.4%** |
| 5000 | 1.5% | **1.0%** |
| 20000 | 0.8% | **0.6%** |
| 60000 | 0.5% | **0.4%** |

LeCun et. al.

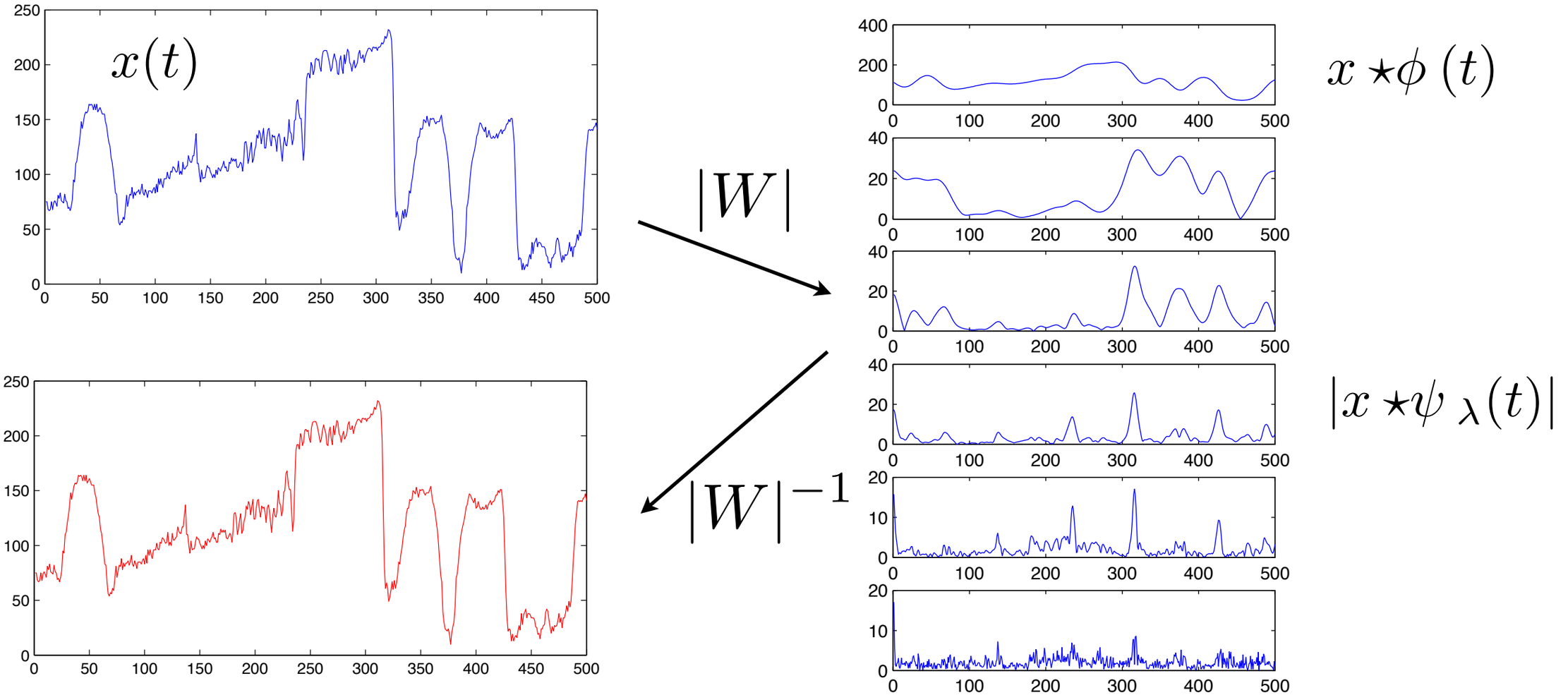
# Scattering Inversion: Phase Recovery

$$Wx = \left\{ x \star \phi \,,\; x \star \psi_\lambda \right\}_\lambda \quad \text{is linear and unitary.}$$

**Theorem** For appropriate wavelets *I. Waldspurger*

$$|W|x = \left\{ x \star \phi \,,\; |x \star \psi_\lambda| \right\}_\lambda$$

is invertible and the inverse is continuous.



# Scattering Inversion: Phase Recovery

*I. Waldspurger*

**Theorem** For appropriate wavelets

$$|W|x = \left\{ x \star \phi \,,\ |x \star \psi_\lambda| \right\}_\lambda$$

is invertible and the inverse is continuous.

Inverse scattering:

$$x$$
$$\uparrow |W|^{-1}$$
$$\left\{ x \star \phi \,,\ |x \star \psi_{\lambda_1}| \right\}_{\lambda_1}$$
$$\uparrow |W|^{-1}$$
$$\left\{ |x \star \psi_{\lambda_1}| \star \phi \,,\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \right\}_{\lambda_1,\lambda_2}$$
$$\uparrow |W|^{-1}$$
$$\left\{ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi \,,\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \right\}_{\lambda_1,\lambda_2,\lambda_3}$$
$$\uparrow |W|^{-1}$$
$$\left\{ |||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi \,,\ ||||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \psi_{\lambda_4}| \right\}_{\lambda_1,\lambda_2,\lambda_3,\ldots}$$

Propagation of errors

# Audio Reconstruction

*J. Anden*

Original audio signal $x$

Reconstruction from $Sx$ for an averaging window $\phi$ of 1 s

from 1st layer coefficients $|x \star \psi_{\lambda_1}| \star \phi$

adding 2nd layer coefficients $||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi$

# Expected Scattering Transform

- If $X(t)$ is a stationary process then
  $||X \star \psi_{\lambda_1}| \star ...| \star \psi_{\lambda_m}(t)|$ is also stationary.

Scattering :

$$SX(t) = \begin{pmatrix} X \star \phi(t) \\ |X \star \psi_{\lambda_1}| \star \phi(t) \\ ||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(t) \\ |||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi(t) \\ ... \end{pmatrix}_{\lambda_1,\lambda_2,\lambda_3,...}$$

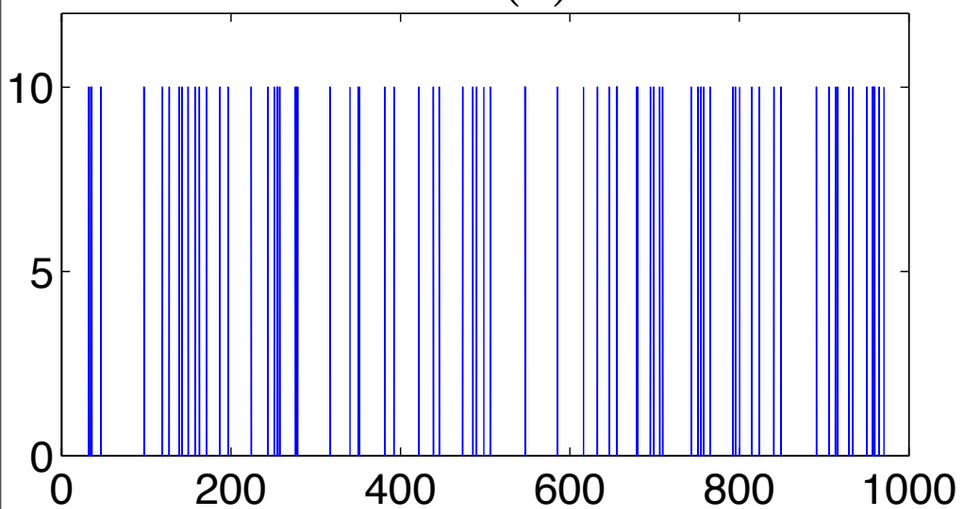
- When $\phi \to 1$ with "appropriate" ergodicity conditions"
  $SX(t)$ may converge to the expected scattering transform:

$$\overline{S}X = \begin{pmatrix} E(X) \\ E(|X \star \psi_{\lambda_1}|) \\ E(||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|) \\ E(|||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}|) \\ ... \end{pmatrix}_{\lambda_1,\lambda_2,\lambda_3,...}$$

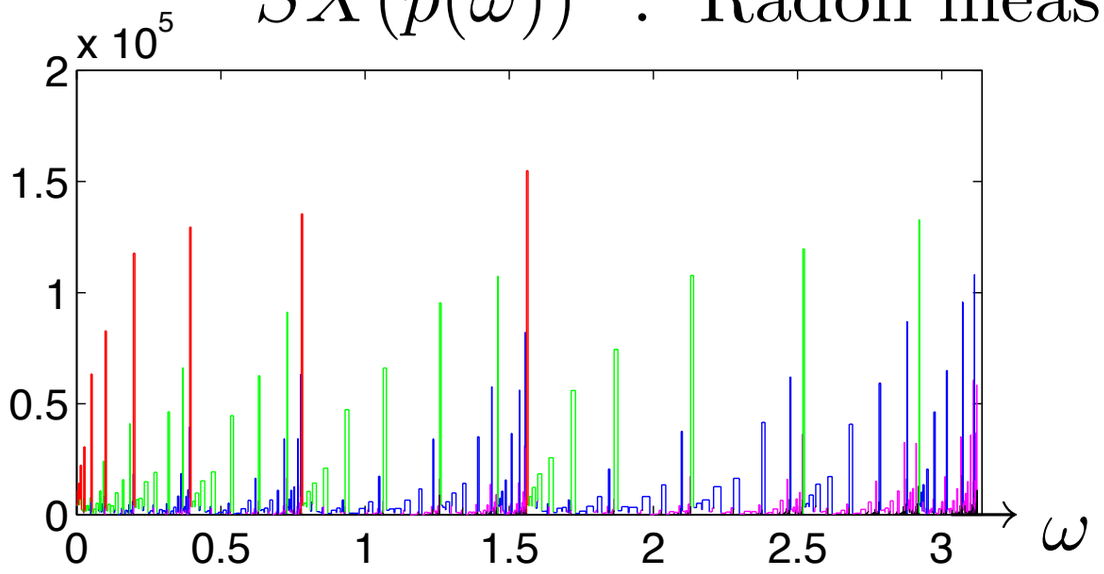
# Scattering White Noises

Constant Fourier power spectrum: $\hat{R}_X(\omega) = \sigma^2$.

Bernoulli $X(t)$

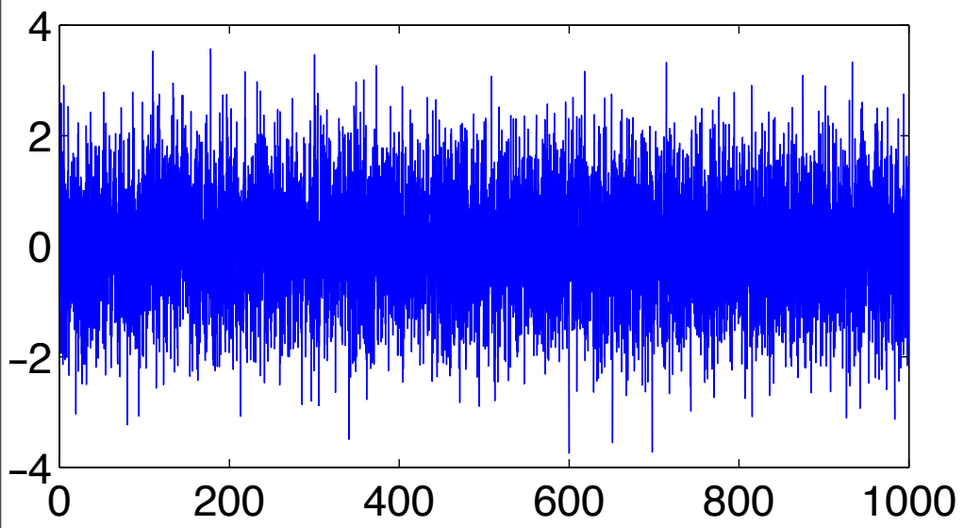



$\overline{S}X(p(\omega))^2$ : Radon measure

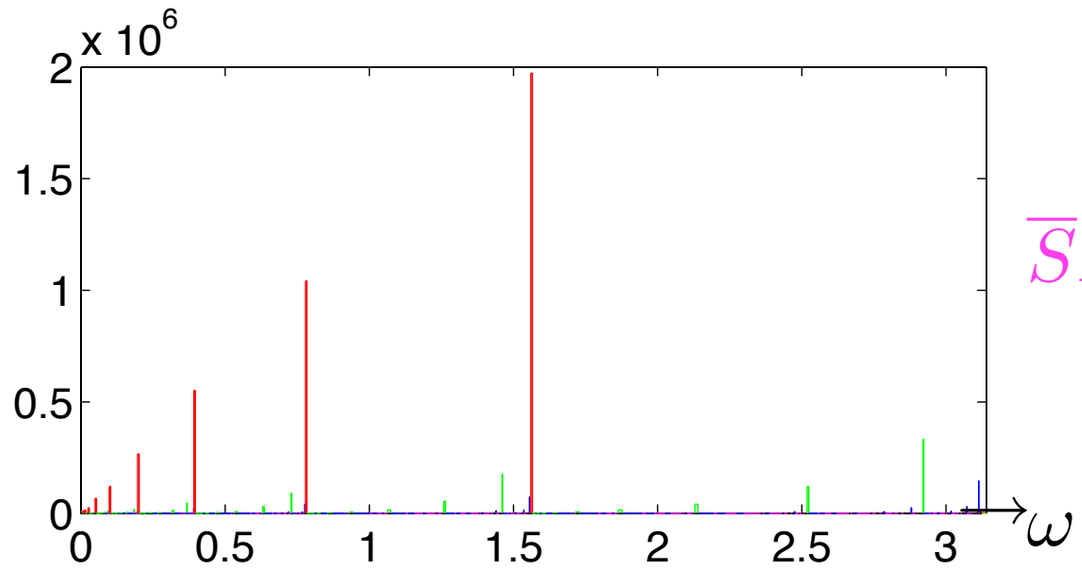



$\overline{S}X(\lambda_1)$

$\overline{S}X(\lambda_1, \lambda_2)$




Gaussian White




$\overline{S}X(\lambda_1, \lambda_2, \lambda_3)$

$\overline{S}X(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$

$$\int \hat{R}_X(\omega)\, |\hat{\psi}_{2^j}(\omega)|^2\, d\omega = \int_{2^j}^{2^{j+1}} \overline{S}X(p(\omega))^2\, d\omega \ .$$

# Expected Scattering Transform

$X(t)$ stationary process:

$$\overline{S}X = \begin{pmatrix} E(X) \\ E(|X \star \psi_{\lambda_1}|) \\ E(||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|) \\ E(|||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}|) \\ \dots \end{pmatrix}_{\lambda_1,\lambda_2,\lambda_3,\dots}$$

**Theorem**: *A scattering is*

*contractive* $\|\overline{S}X - \overline{S}Y\|^2 \leq E(|X-Y|^2)$

*preserves norms* $\|\overline{S}X\|^2 = E(|X|^2)$ *(for finite random vectors)*

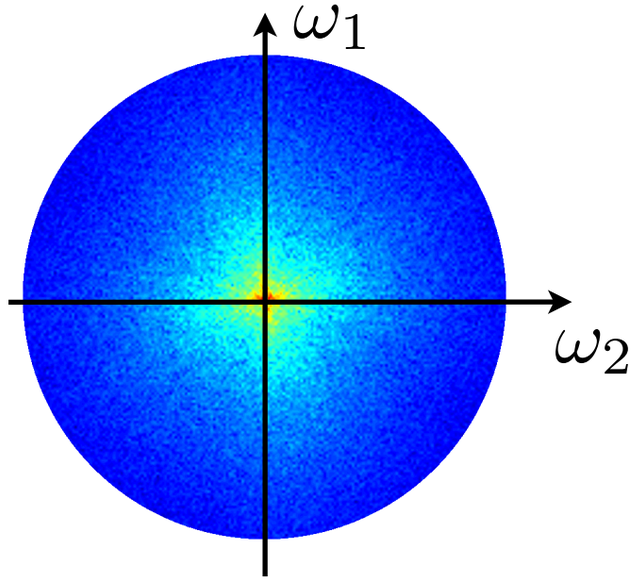
*stable to stationary deformations* $X_\tau(t) = X(t-\tau(t))$

$$\|\overline{S}X - \overline{S}X_\tau\| \leq C \sup_t |\nabla\tau(t)|\, E(|X|^2)^{1/2} .$$

# Textures with Same Spectrum

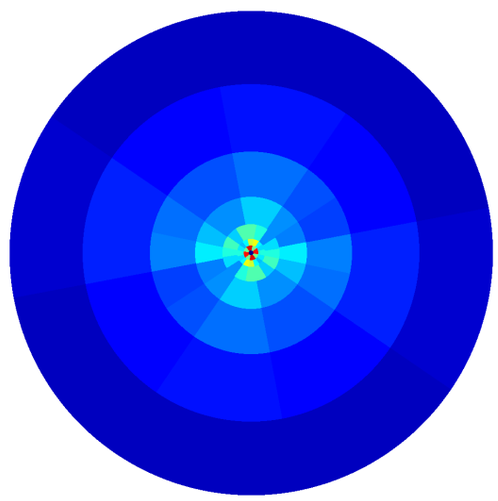
$X(t)$
stationary process

$\hat{R}_X(\omega)$
Power Spectrum

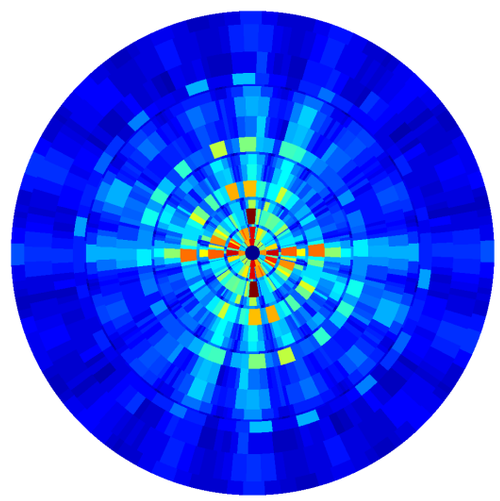


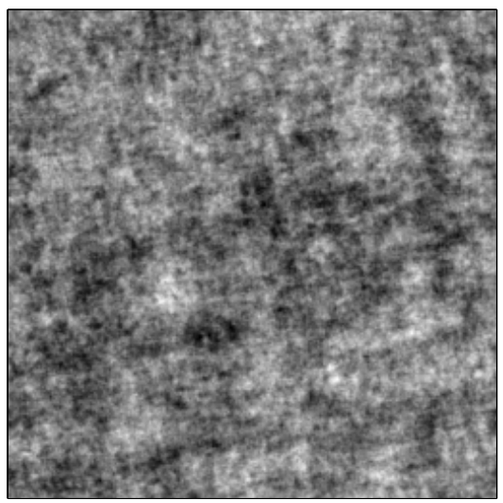

Expected Scattering

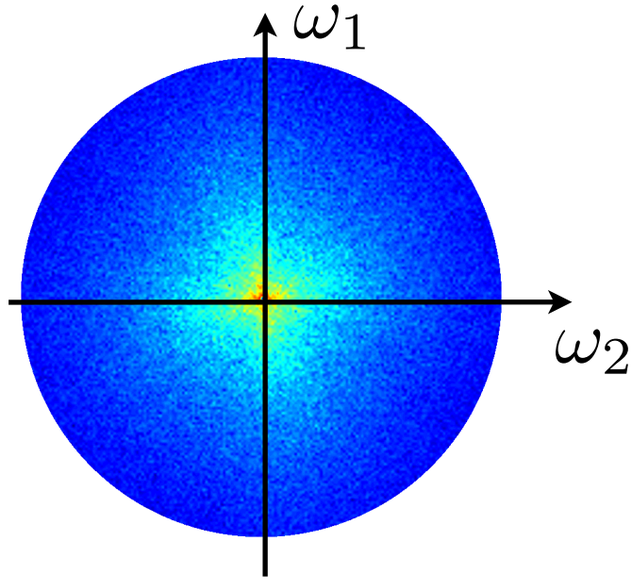
$E(|X \star \psi_{\lambda_1}|)$

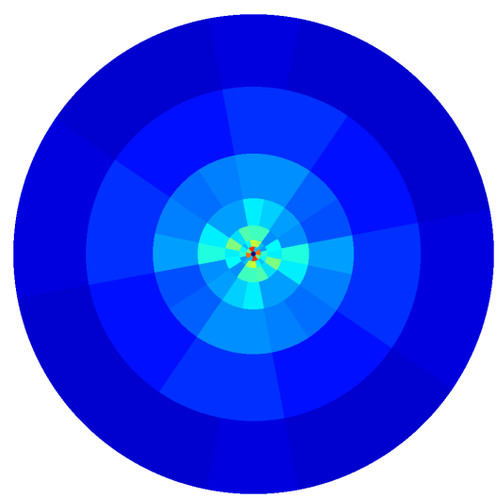


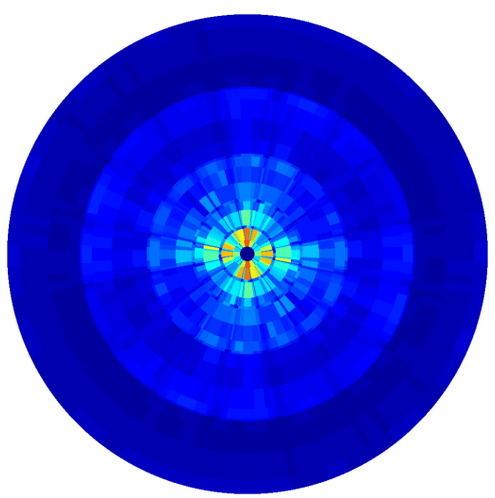
$E(||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|)$









# Sounds with Same Spectrum

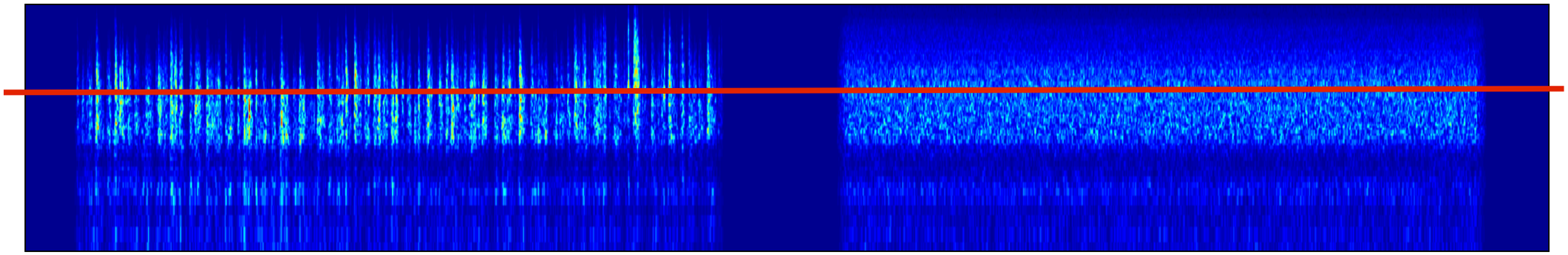
$X$: stationary process

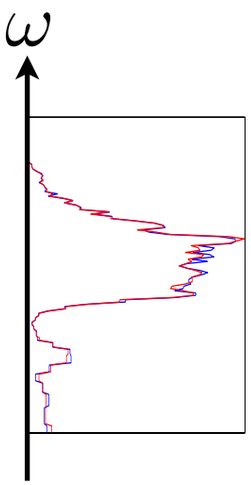
Fourier Spectrum

$\log(\lambda_1)$

J. McDermott

$|x \star \psi_{\lambda_1}|(t)$

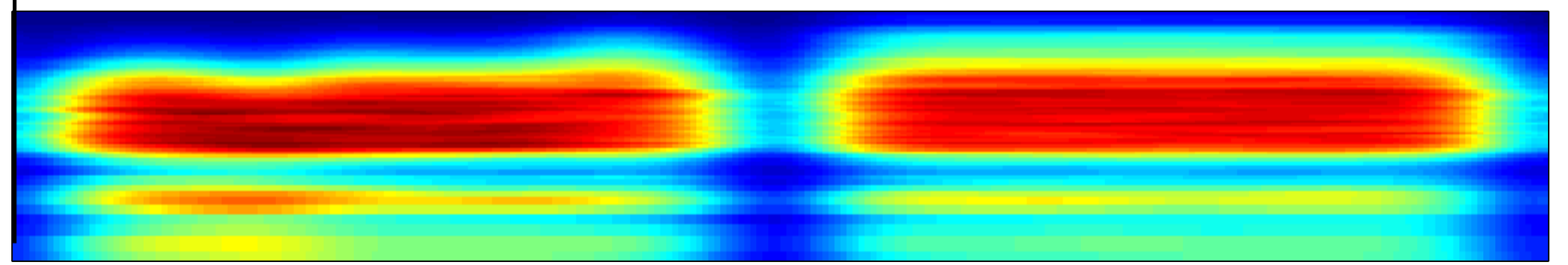





$\log(\lambda_1)$

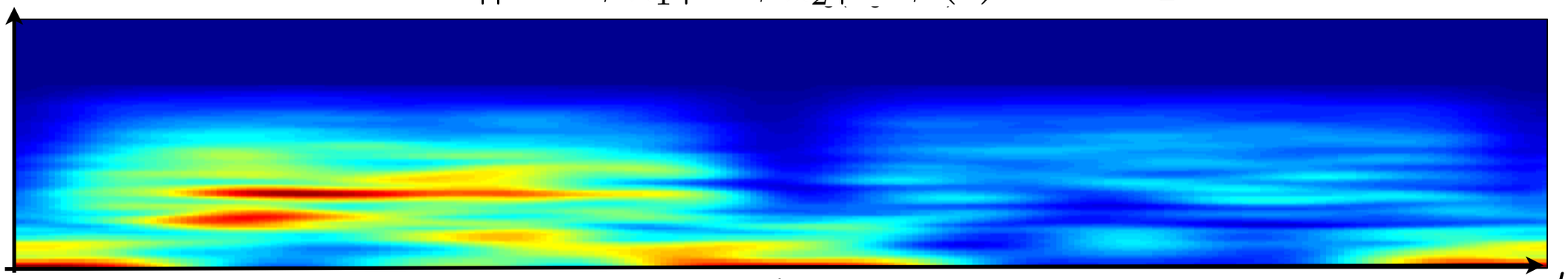
2s window

$|x \star \psi_{\lambda_1}| \star \phi(t)$



$\log(\lambda_2)$

$||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(t)$ for $\lambda_1 = 2000$



$t$

# Representation of Random Processes

- An expected scattering is a non-complete representation

$$\overline{S}X = \begin{pmatrix} E(X) & = & E(U_0 X) \\ E(|X \star \psi_{\lambda_1}|) & = & E(U_1 X) \\ E(||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|) & = & E(U_2 X) \\ E(|||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}|) & = & E(U_3 X) \\ \ldots & & \end{pmatrix}_{\lambda_1, \lambda_2, \lambda_3, \ldots}$$

**Theorem** *(Boltzmann)* The distribution $p(x)$ which satisfies

$$\int_{\mathbb{R}^N} U_m x \; p(x)\, dx = E(U_m X)$$

and maximizes the entropy $-\int p(x) \, \log p(x)\, dx$

can be written: $$p(x) = \frac{1}{Z} \exp\left( \sum_{m=1}^{\infty} \lambda_m \, . \, U_m x \right)$$

# Synthesis from Second Order

J. McDermott textures

*Joakim Anden  Joan Bruna*

- Maximum entropy estimation of $X(t)$ :
  - Gaussian model from 2nd order moments
  ($N$ power spectrum coefficients)
  - Scattering model 1st & 2nd orders ($(\log_2 N)^2$ coefficients)

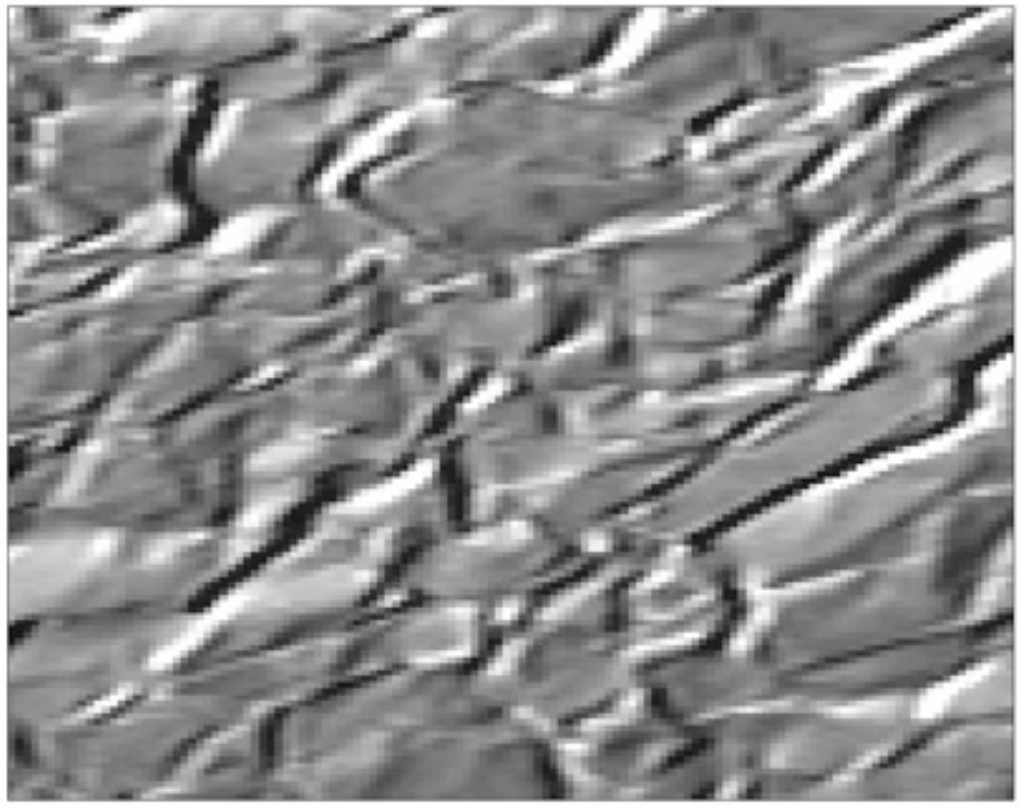
    - Original jackhammer
    - Gaussian model
    - Scattering model

    - Original water
    - Gaussian model
    - Scattering model

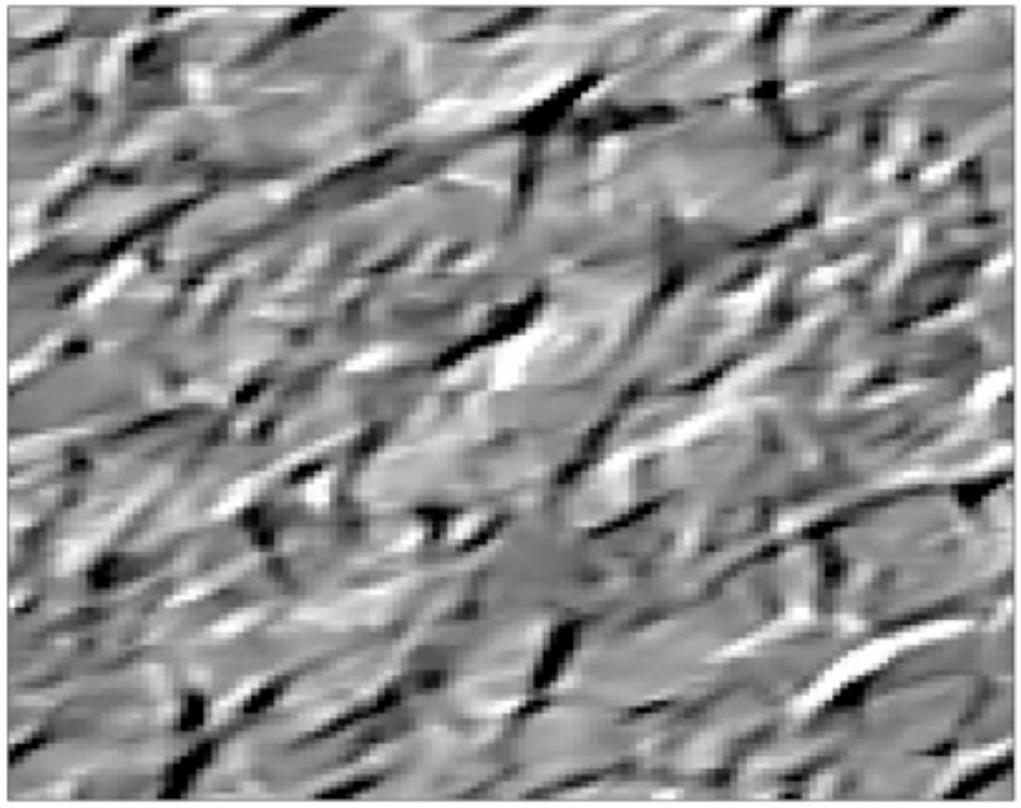
    - Original applause
    - Gaussian model
    - Scattering model

# Image Reconstruction

Original



Reconstructed



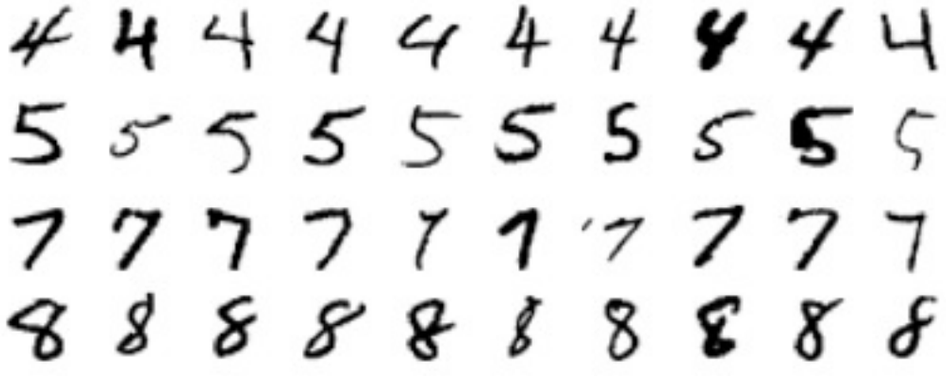
# Part III: High Dimensional Classification

- How to represent high-dimensional data $x \in \mathbb{R}^d$ for classification ?

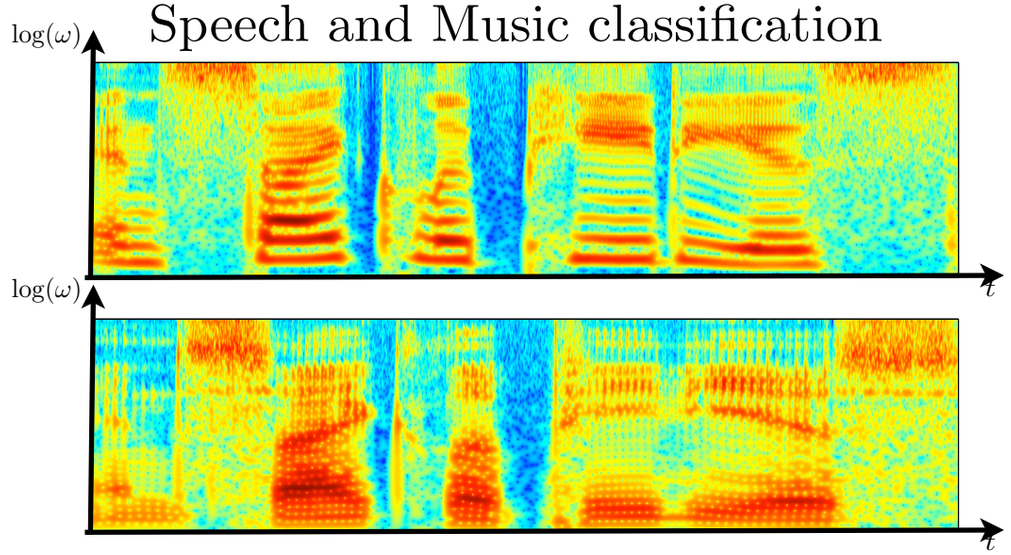
$$d \geq 10^6$$

- Need to compute **discriminative invariants**.

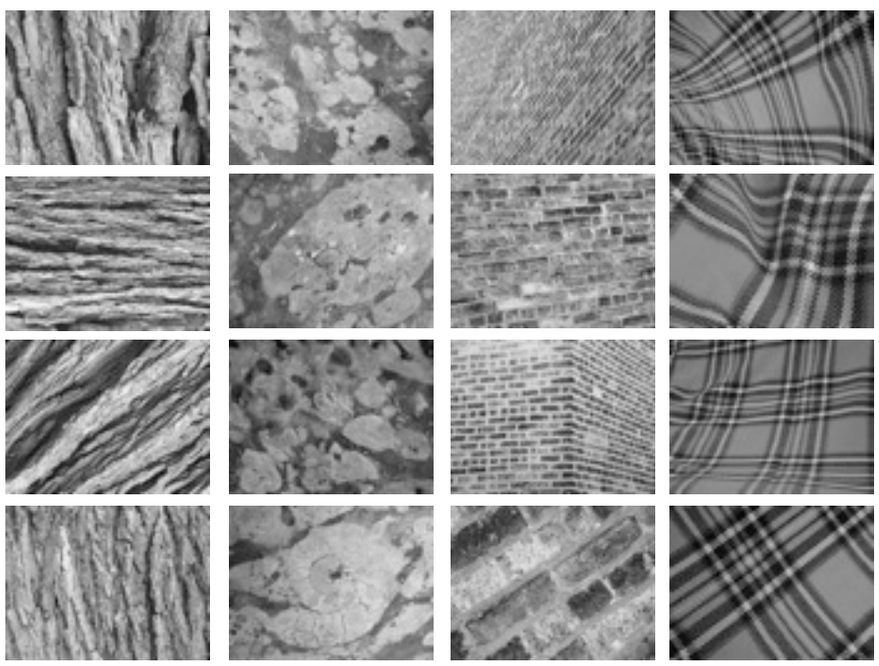
MNIST digit classification



Speech and Music classification

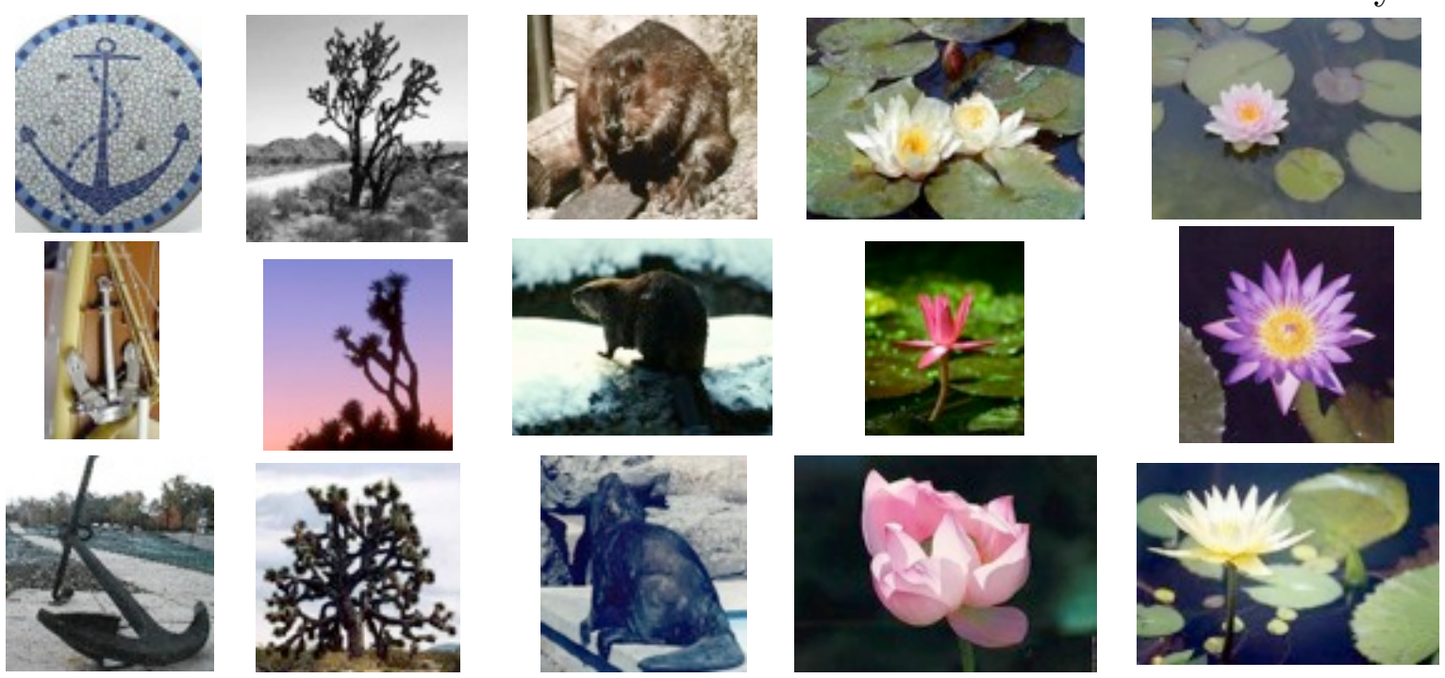



Texture classification



*CalTech 101*

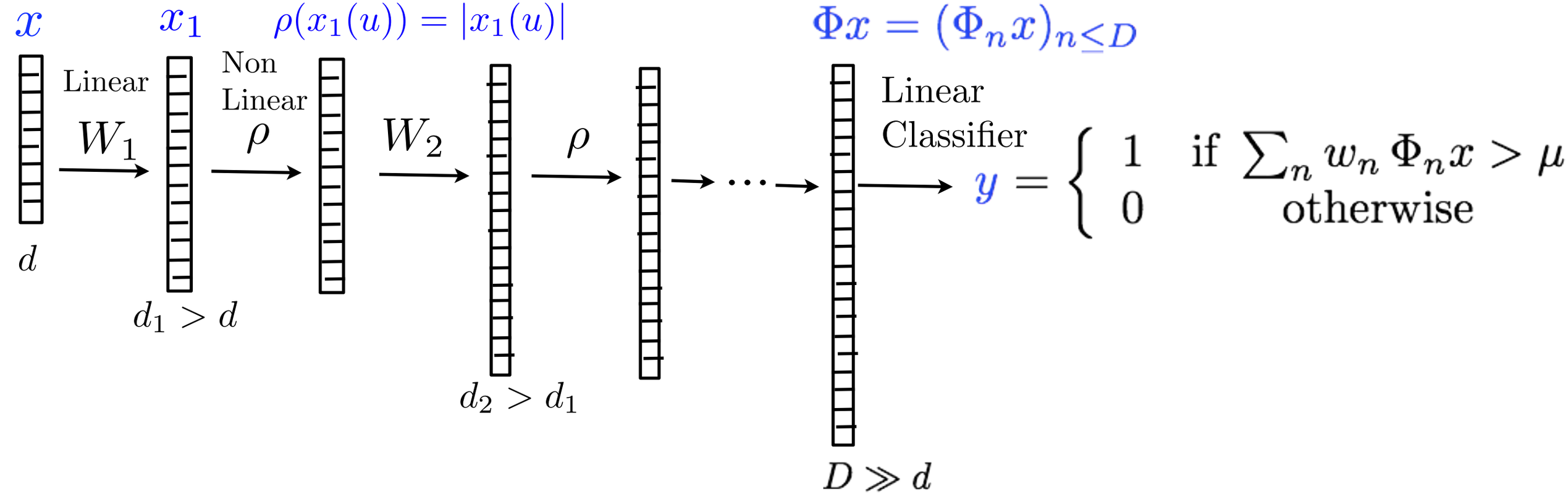
Anchor | Joshua Tree | Beaver | Lotus | Water Lily



# Deep Neural Network Classifiers

*J. Hinton, Y. LeCun* "State of the art results"

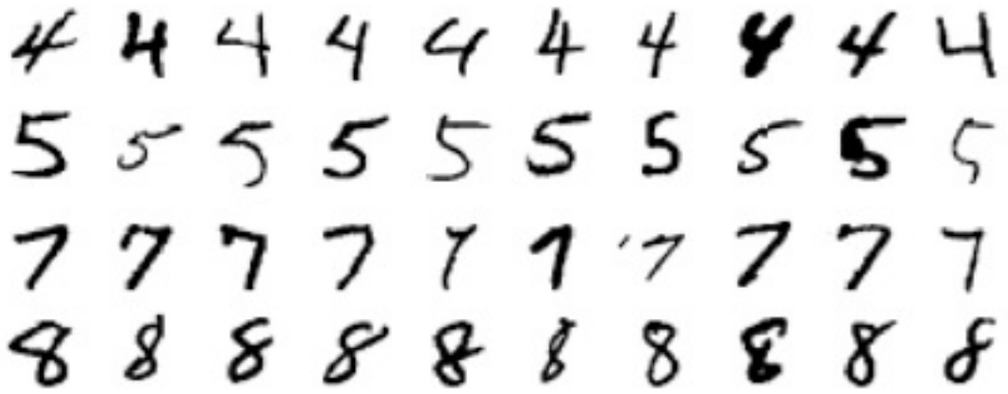
Hierarchical invariance




- Deep network algorithms learn the $W_k$ with sparsity.

Why does it work ?

# Translations and Deformations



- **Invariance** to translations $x_c(t) = x(t - c)$

$$\forall c \in \mathbf{R} \quad , \quad \Phi(x_c) = \Phi(x) \ .$$

Fourier invariant: $\Phi(x) = |\hat{x}(\omega)| = \left| \int x(t)\, e^{-it\omega}\, d\omega \right|$

- **Lipschitz stable** to diffeomorphisms $x_\tau(t) = x(t - \tau(t))$

$$\forall \tau \quad , \quad \|\Phi(x_\tau) - \Phi(x)\| \leq C \sup_t |\nabla \tau(t)| \, \|x\| \ .$$

Fourier Failure

$\sup_t |\nabla \tau(t)|$: diffeomorphism metric

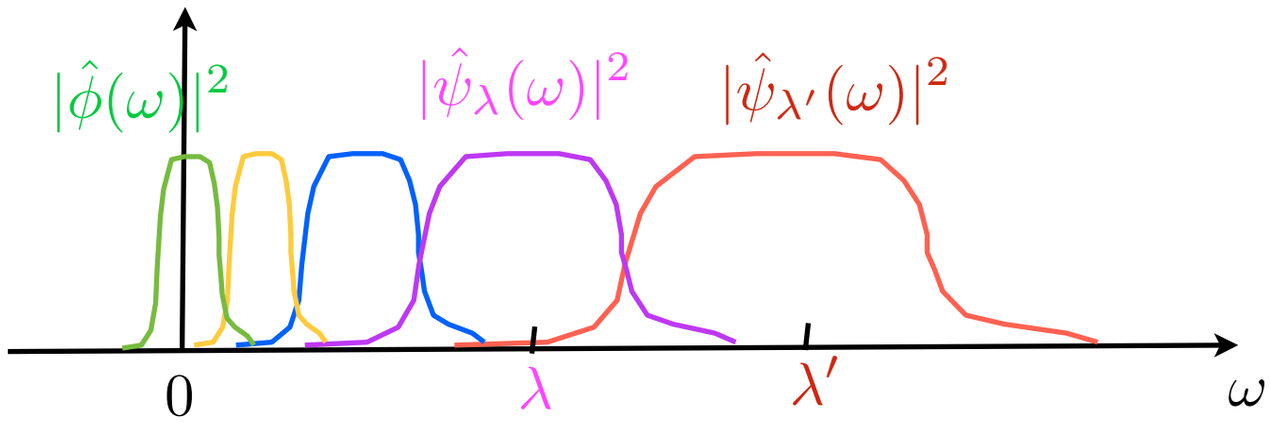
# Wavelet Transform

- Complex analytic wavelet: $\psi(t) = \psi^a(t) + i\,\psi^b(t)$

- For $t \in \mathbb{R}$, dilated

$$\psi_\lambda(t) = 2^j \psi(2^j t)$$

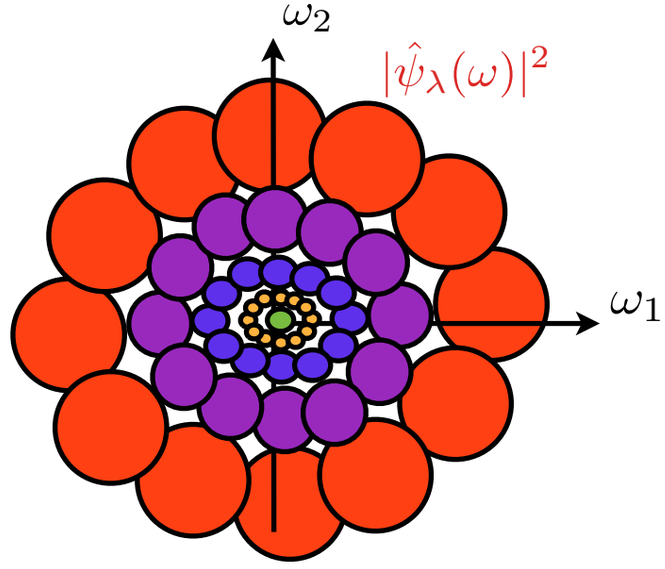
$$\text{with } \lambda = 2^j$$




- For $t \in \mathbb{R}^2$, dilated and rotated

$$\psi_\lambda(t) = 2^j\, \psi(2^j r t)$$

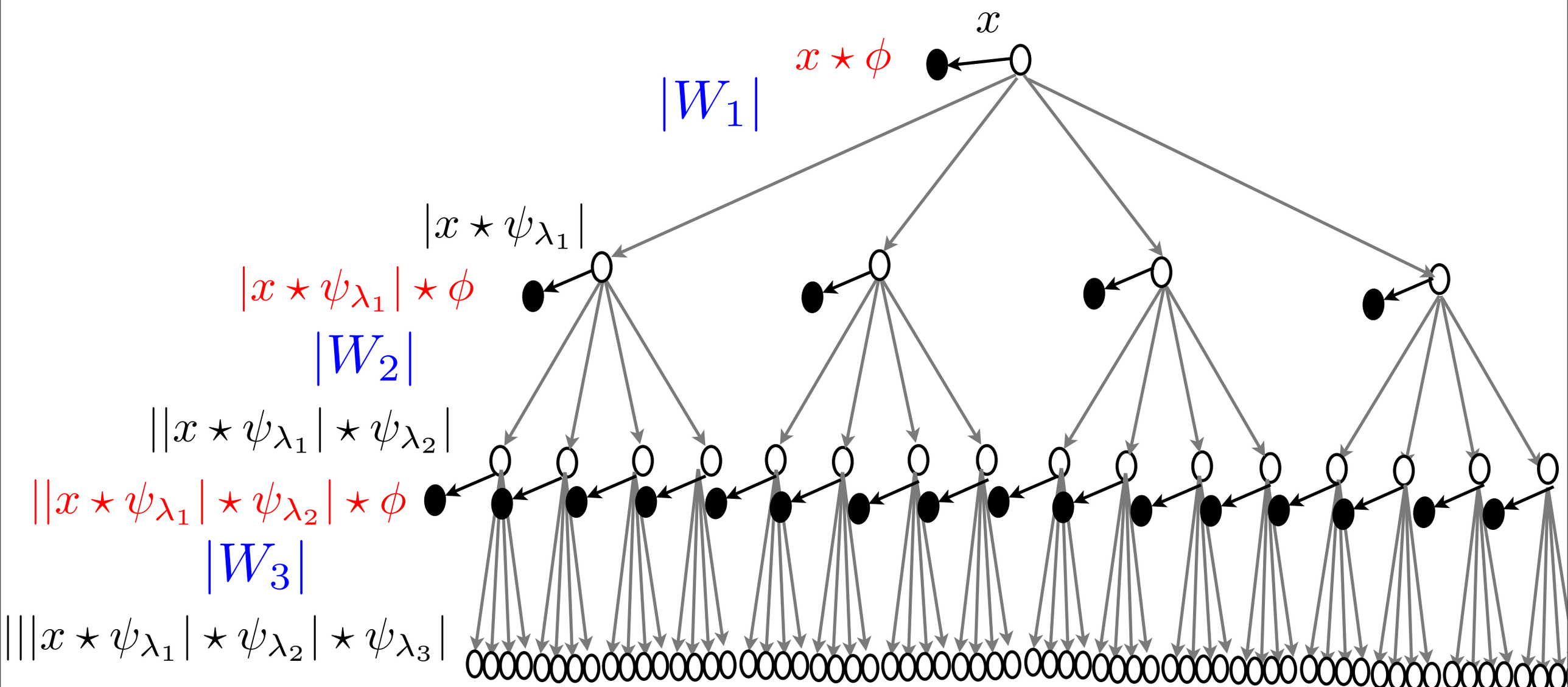
$$\text{with } \lambda = (2^j, r)$$




$$\text{Wavelet transform } Wx = \Big(x \star \phi(t)\,,\; x \star \psi_\lambda(t)\Big)_\lambda$$

$$\text{Tight frame } \|Wx\|^2 = \|x \star \phi\|^2 + \sum_\lambda \|x \star \psi_\lambda\|^2 = \|x\|^2$$

# Wavelet Scattering Transform



# Scattering Transform in $\mathbf{L^2}(\mathbb{R}^n)$

- Wavelet scattering of $x(t)$:

$$Sx = \begin{pmatrix} x \star \phi(t) \\ |x \star \psi_{\lambda_1}| \star \phi(t) \\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(t) \\ |||x \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi(t) \\ \dots \end{pmatrix}_{\lambda_1,\lambda_2,\lambda_3,\dots}$$

**Theorem**

$$\|Sx - Sy\| \leq \|x - y\|$$

$$\|Sx\| = \|x\|$$

Lipschitz continuous to diffeomorphism

- If $x \in \mathbf{L^2}(\mathbb{R}^n)$ then full translation invariance with

$$\lim_{\phi \to 1} Sx \to \begin{pmatrix} \int x(t)\, dt \\ \|x \star \psi_{\lambda_1}\|_1 \\ \||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}\|_1 \\ \|||x \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}\|_1 \\ \dots \end{pmatrix}_{\lambda_1,\lambda_2,\lambda_3,\dots}$$

similarities with Fourier transform

$$\to \overline{S}x \in \mathbf{L^2}(\mathcal{P})$$

# Scattering Random Processes

- Wavelet scattering of $x(t)$:

$$Sx = \left( \begin{array}{c} x \star \phi(t) \\ |x \star \psi_{\lambda_1}| \star \phi(t) \\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi(t) \\ |||x \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi(t) \\ \ldots \end{array} \right)_{\lambda_1, \lambda_2, \lambda_3, \ldots}$$

- If $x(t)$ is a realization of a sationary process $X(t)$ then convergence to an expected scattering:

$$\lim_{\phi \to 1} Sx = \overline{S}X = \left( \begin{array}{c} E(X) \\ E(|X \star \psi_{\lambda_1}|) \\ E(||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|) \\ E(|||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}|) \\ \ldots \end{array} \right)_{\lambda_1, \lambda_2, \lambda_3, \ldots}$$

contractive representation of a random proces

# Part III : Overview

- ***Scattering of Stationary Processes***

  Texture discrimination

  Multifractal analysis and applications

- ***Invariance to Frequency Transposition, Rotation and Scaling***

  Scattering wavelets on Lie Groups

- ***Unsupervised Learning of Deep Networks***

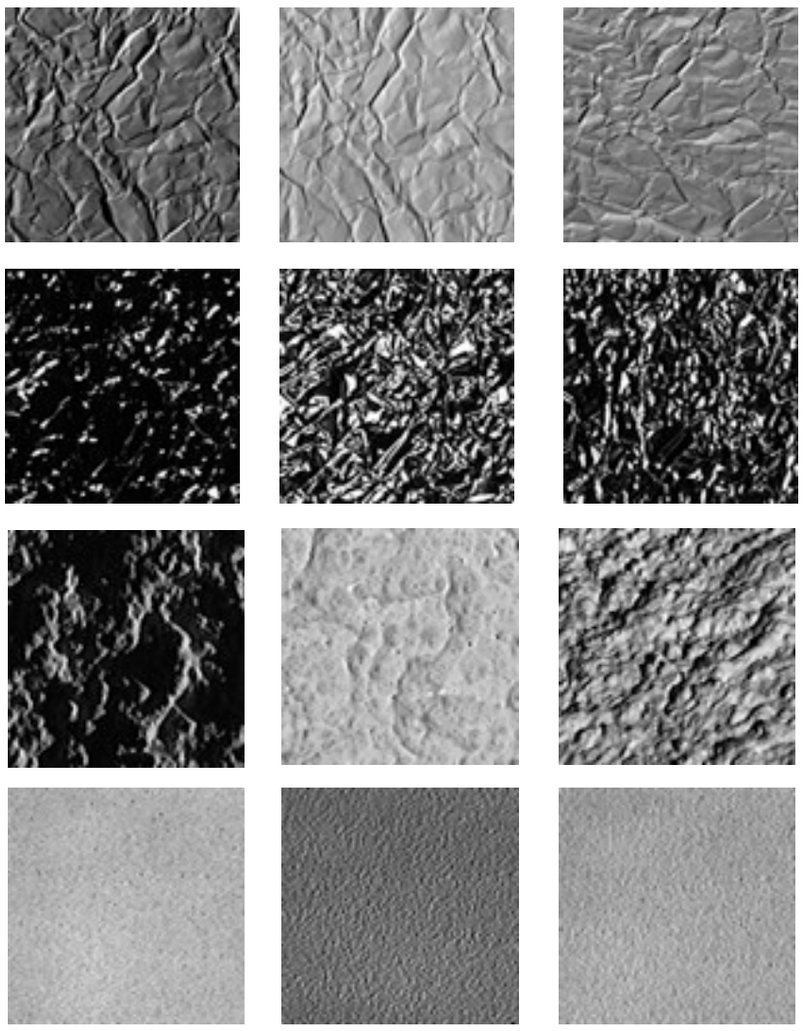
  Scattering with frames

  Sparsity and contraction.

# Classification of Textures

CUREt database

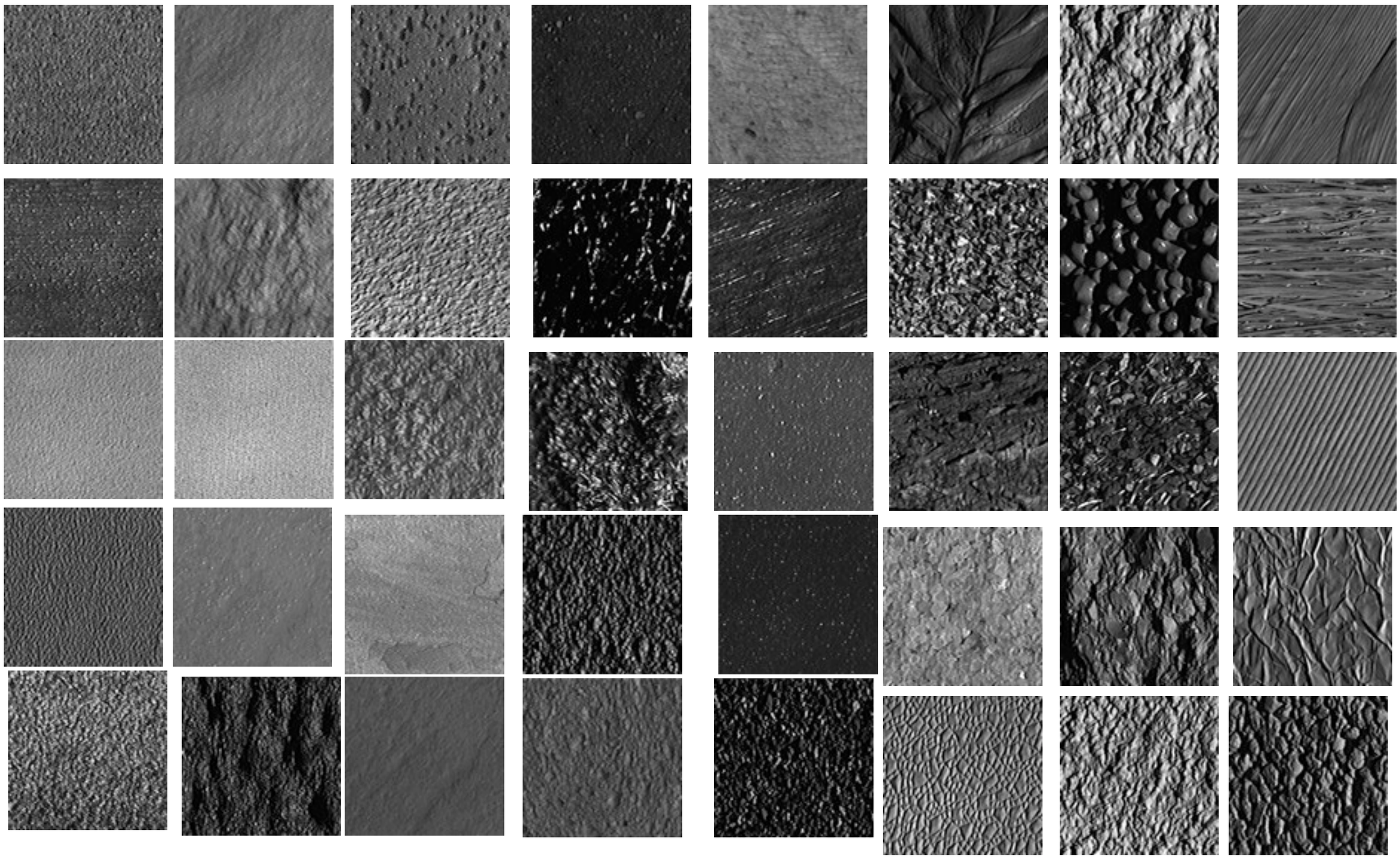
61 classes



Rotations and illumination variations.

# Classification of Textures

40 classes of CureT



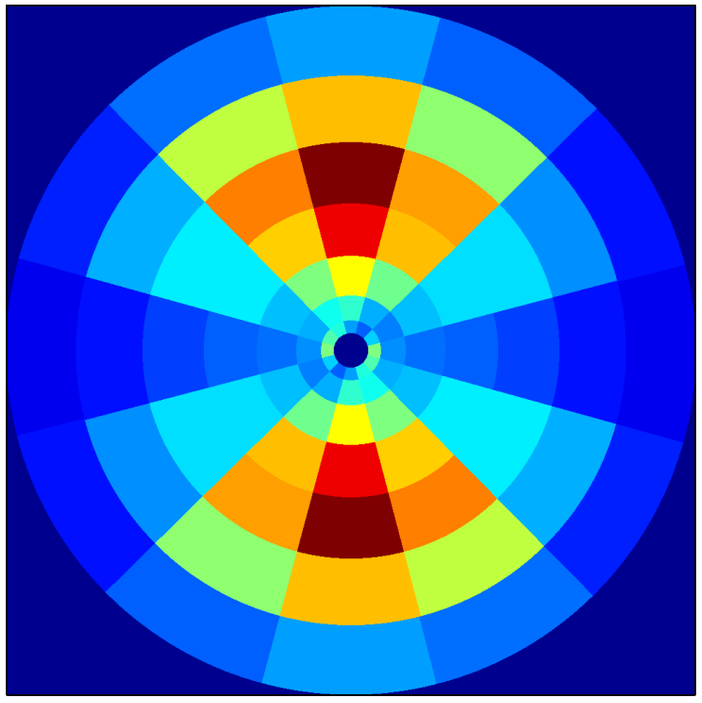
# Classification of Textures

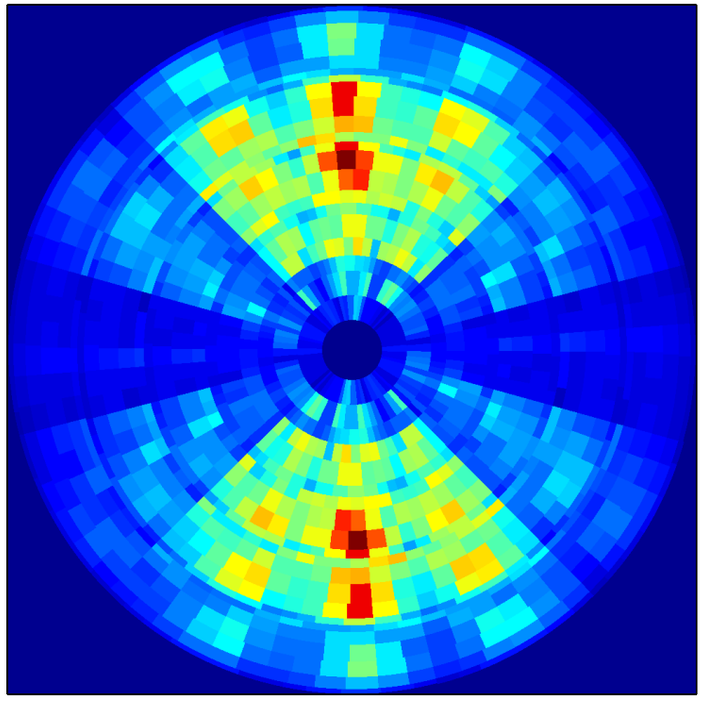
Expected Scattering
estimated with $\phi = 1$

$X$

$|X \star \psi_{\lambda_1}| \star \phi$

$||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi$

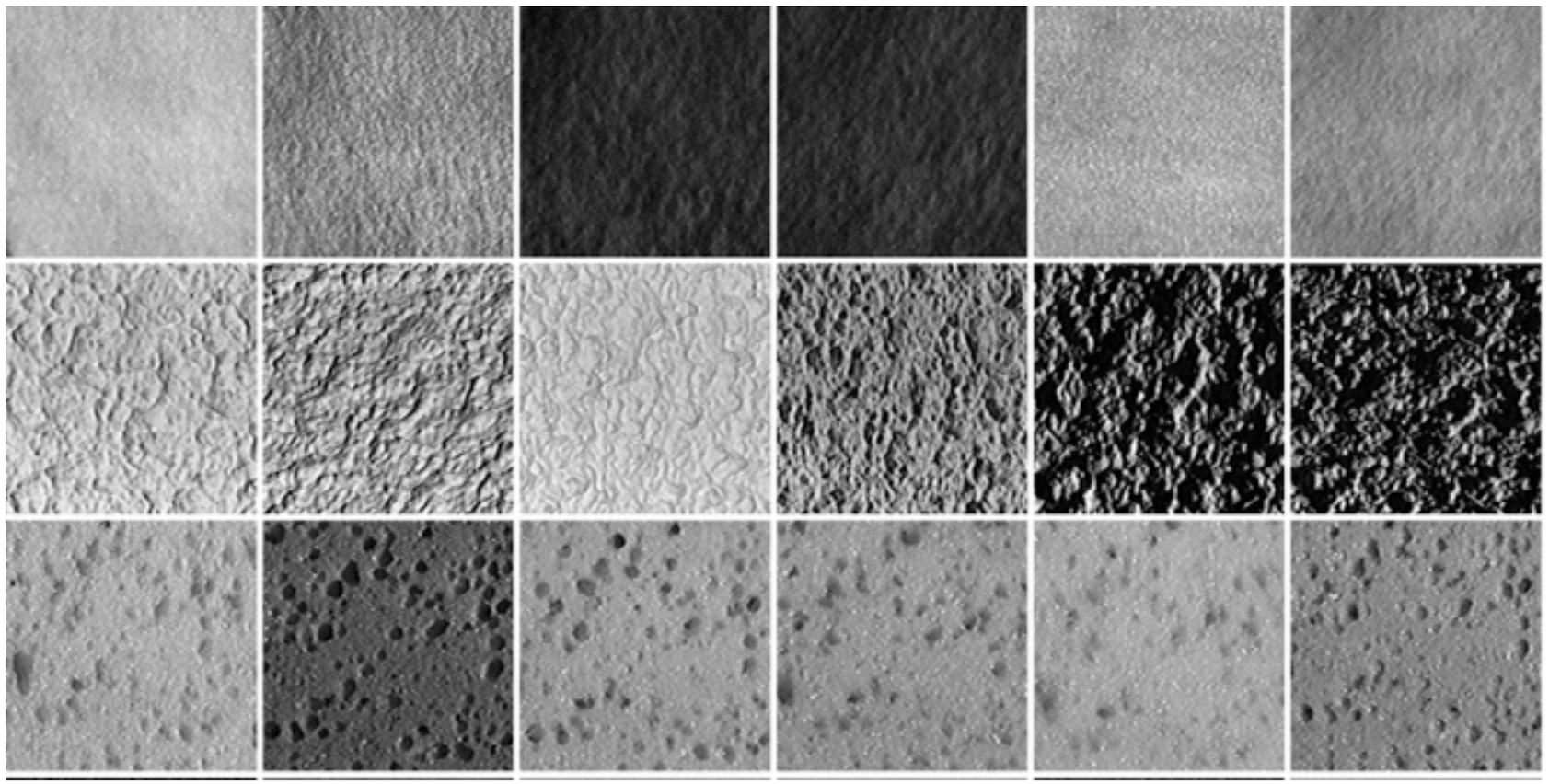


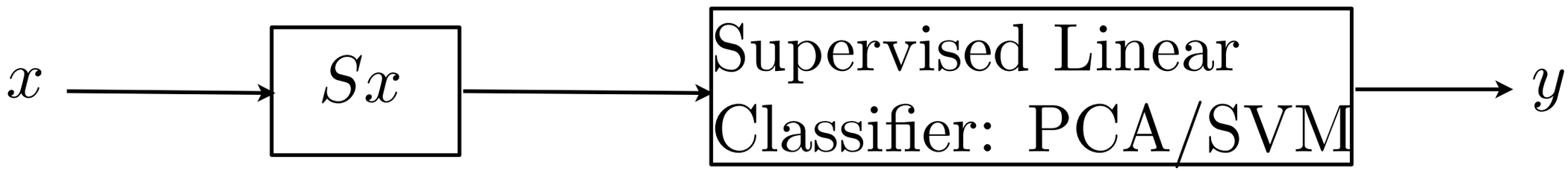


# Classification of Textures

*J. Bruna*

CUREt database
61 classes






| Training per class | Fourier Spectr. | Histogr. Features | Scattering |
|---|---|---|---|
| 46 | 1% | 1% | **0.2** % |

# Self-Similar Multifractals

- Stochastic self-similarity: $X(st) \equiv A_s\, X(t)$

  where $A_s$ is a random variable independant of $X$ and

$$E(|A_s|^q) \sim s^{\zeta(q)}$$

  and $X(t)$ has stationary increments.

- $A_s$ is constant for fractional Brownians and Levy Stable:

$$\Rightarrow \quad \zeta(q) = \zeta(1)\, q \ .$$

- $A_s$ is a log-normal random variable for Mandelbrot cascades.

# Scattering Multifractals

$J.Bruna, E.Bacry, J.F.Muzy$

$$X(st) \equiv A_s\, X(t) \quad \text{with} \quad E(|A_s|^q) \sim s^{\zeta(q)} \ .$$

- First order scattering coefficients

$$\overline{S}X(2^{j_1}) = E(|X \star \psi_{j_1}|) \sim 2^{\zeta(1) j_1}$$

Not sufficient to discriminate different selsimilar processes.

Avoid high order moments: numerical instabilities.

- Normalized second order scattering

$$\widetilde{S}X(2^{j_1}, 2^{j_2}) = \frac{E(||X \star \psi_{2^{j_1}}| \star \psi_{2^{j_2}}|)}{E(|X \star \psi_{2^{j_1}}|)}$$

**Proposition** If $X$ has stationary increments and self-similar:

$$\widetilde{S}X(2^{j_1}, 2^{j_2}) = \widetilde{S}X(2^{j_1 - j_2}) \ .$$

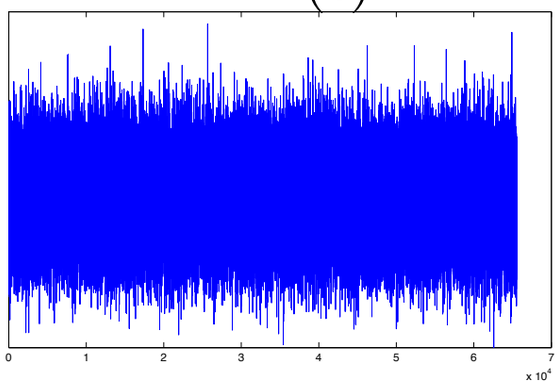
# Fractional Brownian Scattering

**Proposition:** For fractional Brownian motion and noise

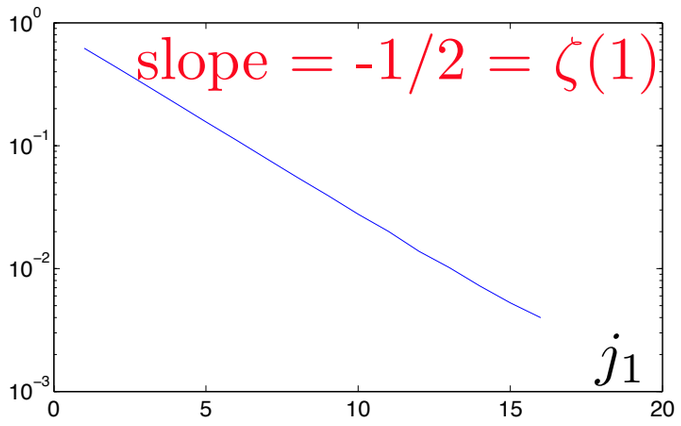
$$\widetilde{S}X(2^{j_1}, 2^{j_2}) = \frac{E(||X \star \psi_{2^{j_1}}| \star \psi_{2^{j_2}}|)}{E(|X \star \psi_{2^{j_1}}|)} \sim 2^{-(j_2 - j_1)/2}$$

Gaussian white noise

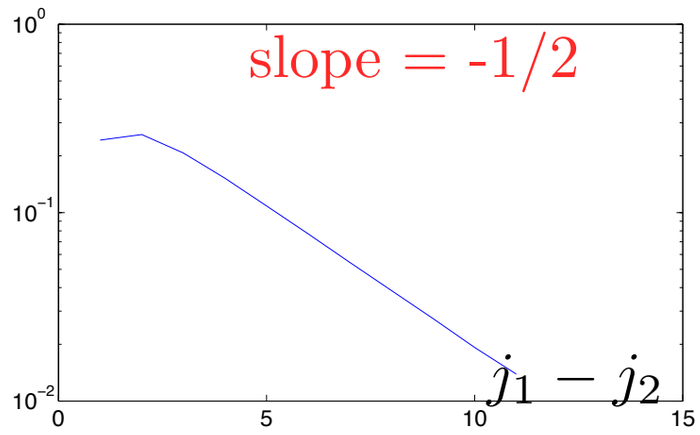
X(t)



$\log \overline{S}X(2^{j_1}) \sim \zeta(1)\, j_1$

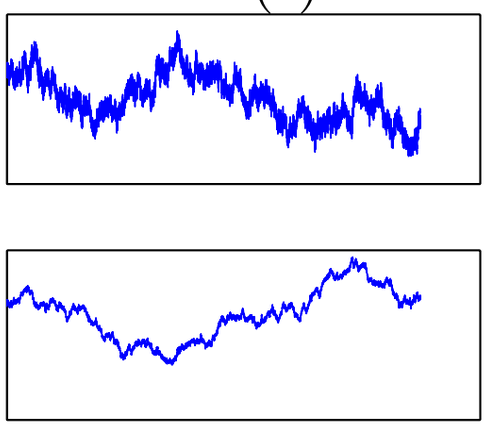



$\log \overline{S}X(2^{j_1}, 2^{j_1}) \sim -(j_2 - j_1)/2$

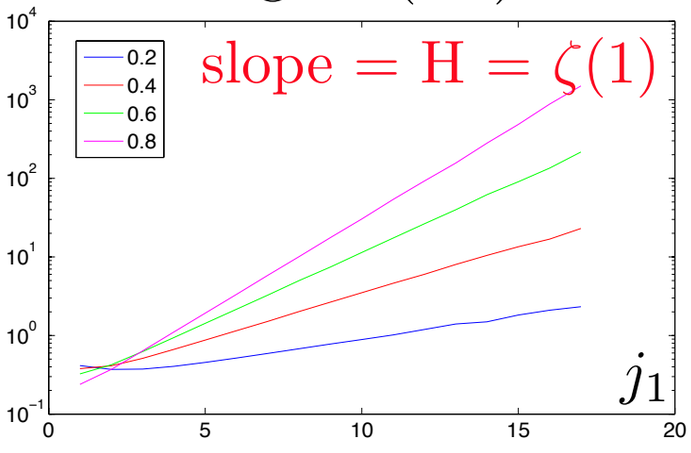



Fractional Brownian

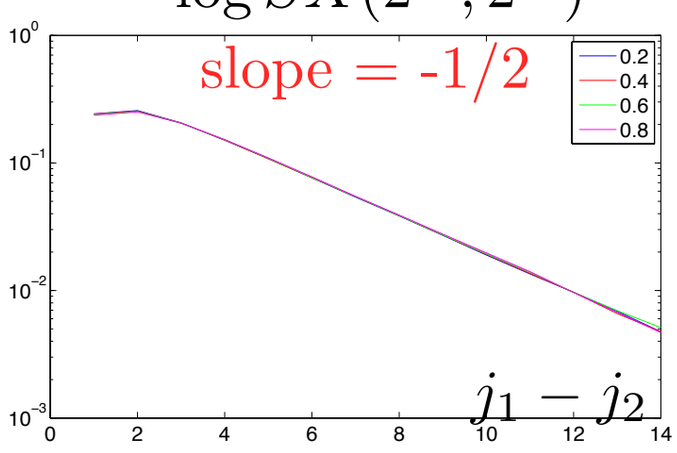
X(t)



$\log \overline{S}X(2^{j_1})$



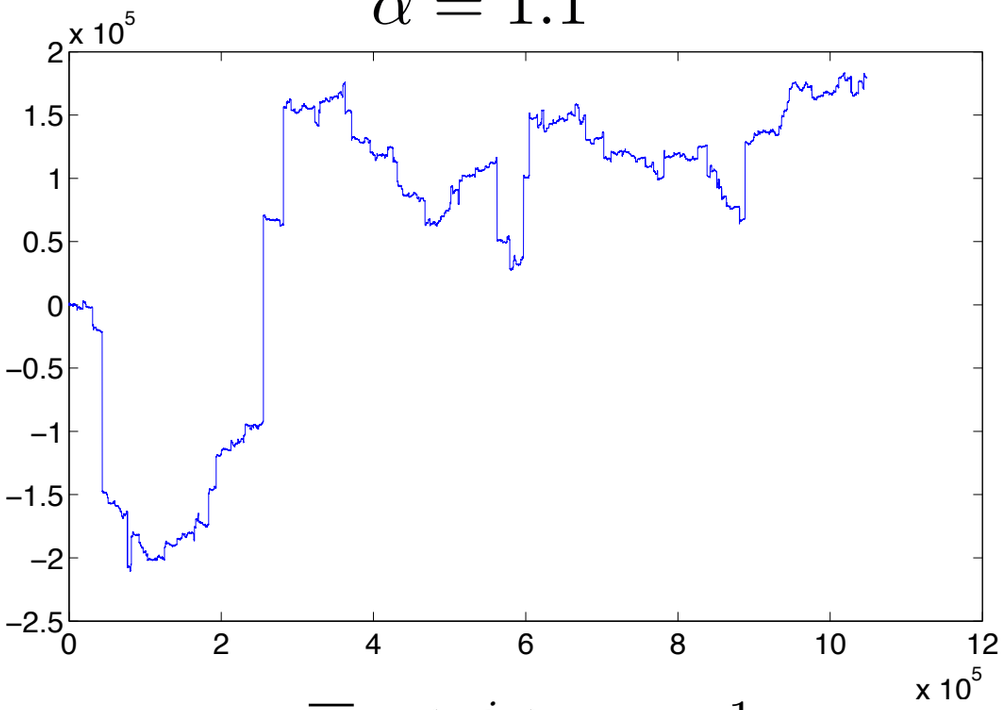

$\log \widetilde{S}X(2^{j_1}, 2^{j_2})$



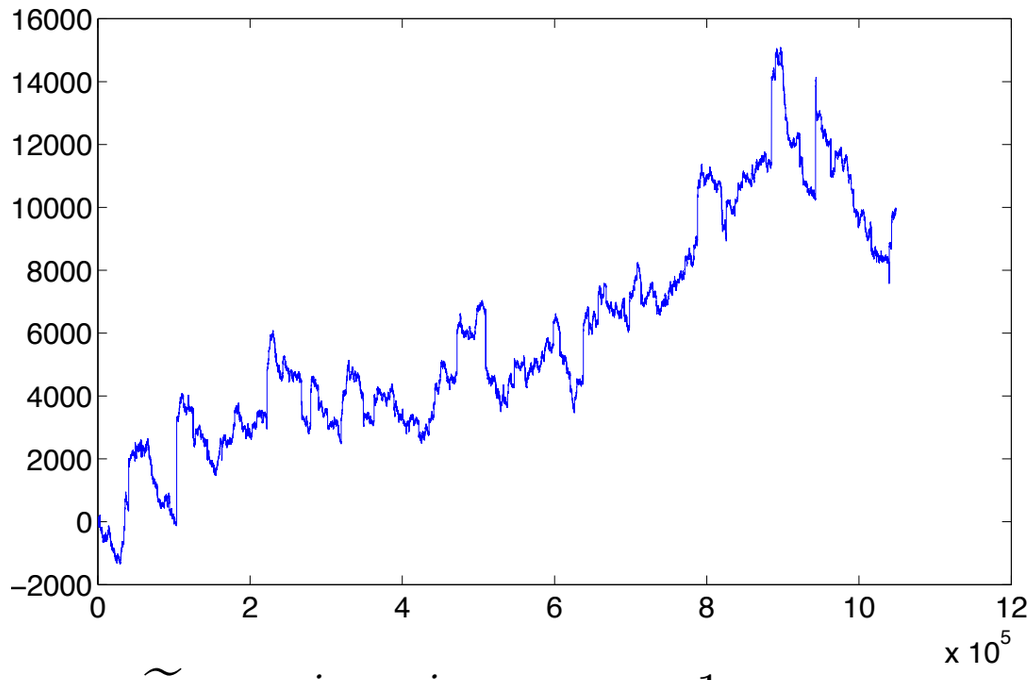
# Scattering Stable Levy Measures

$X(t)$

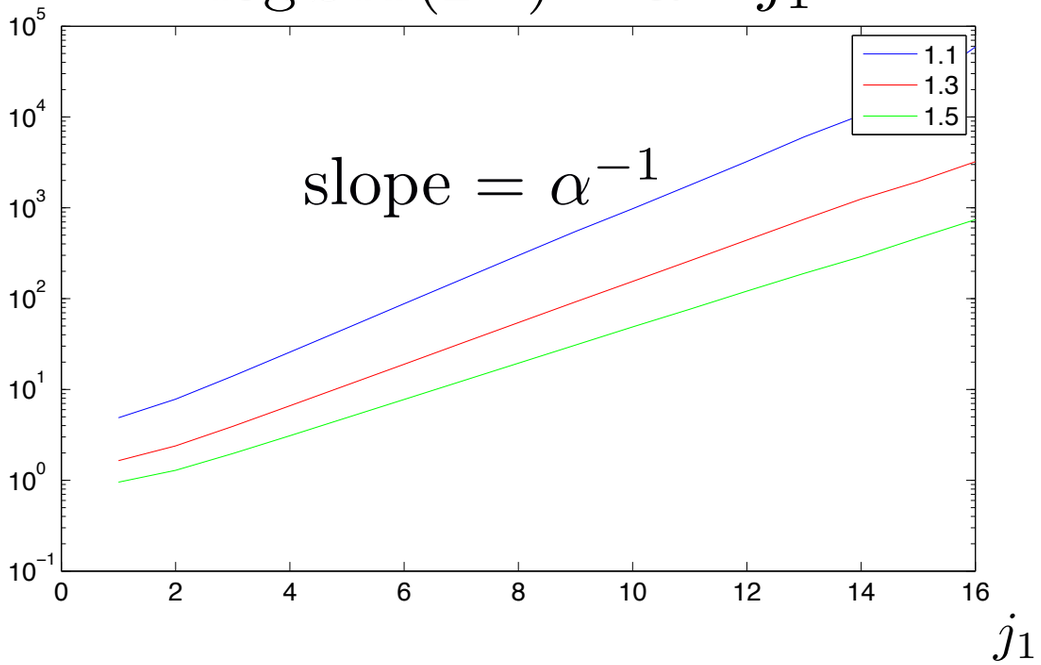
$\alpha = 1.1$



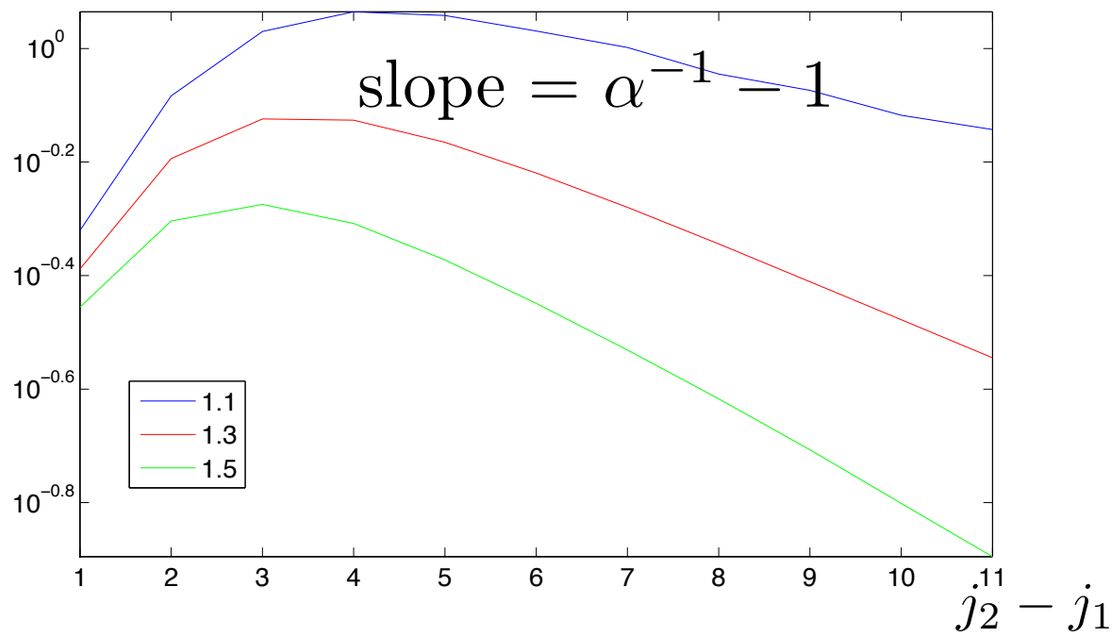

$\alpha = 1.5$




$\log \overline{S}X(2^{j_1}) \sim \alpha^{-1} j_1$




$\log \widetilde{S}X(2^{j_1}, 2^{j_2}) \sim (\alpha^{-1} - 1)(j_2 - j_1)$




$\alpha = 2$ : Brownian motion.

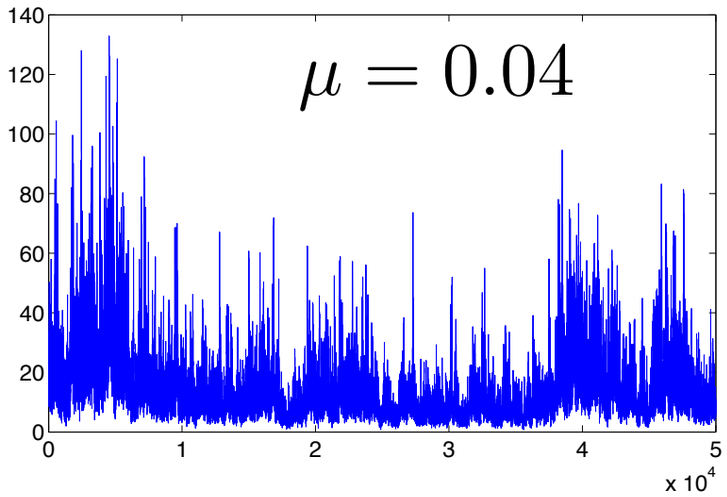
# Mandelbrot Cascades

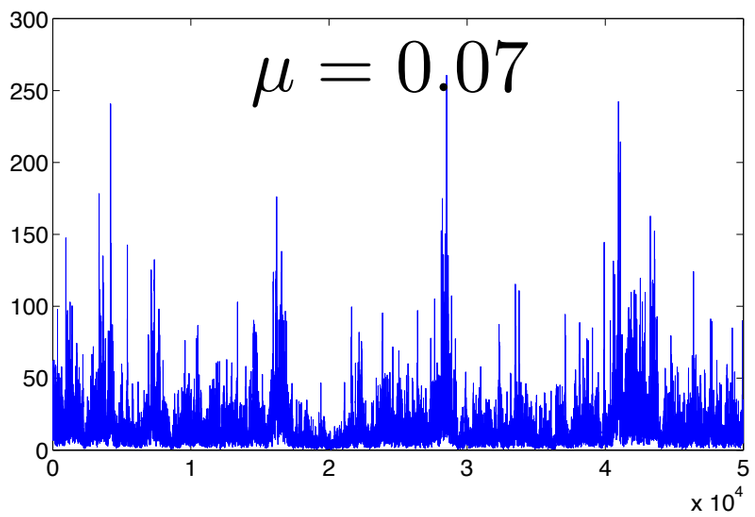
*Barral, Mandelbrot*

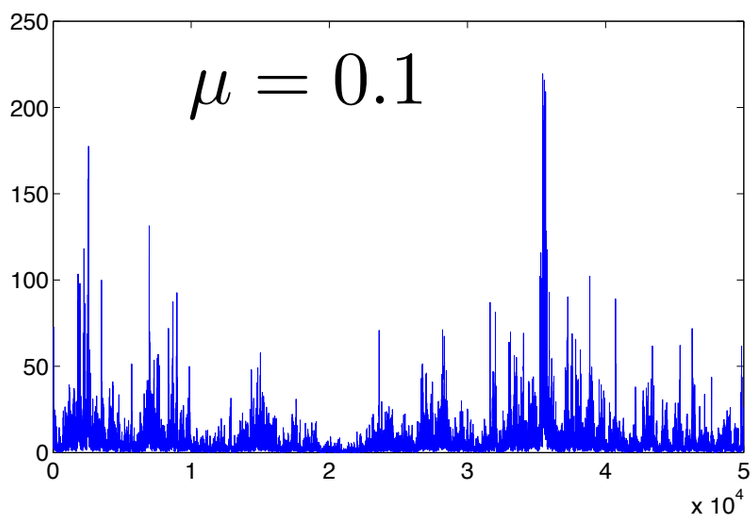
- Stationary log normal random measure $dX(t)$ obtained as multiscale products of log-normal random variables.

$$\zeta(q) = \left(1 + \frac{\mu}{2}\right) q - \frac{\mu}{2} q^2$$

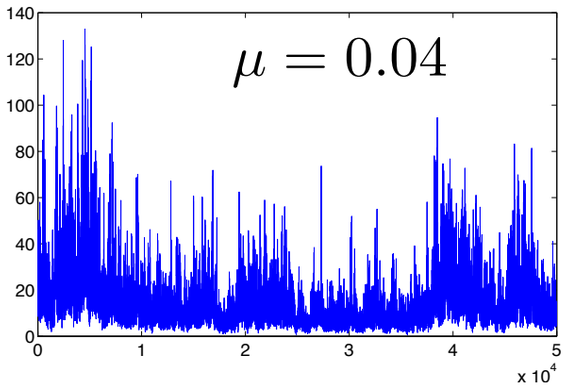
$\mu$ is an "intermittency" factor.

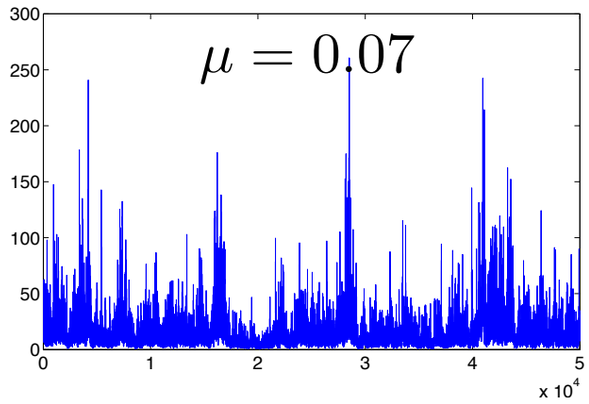


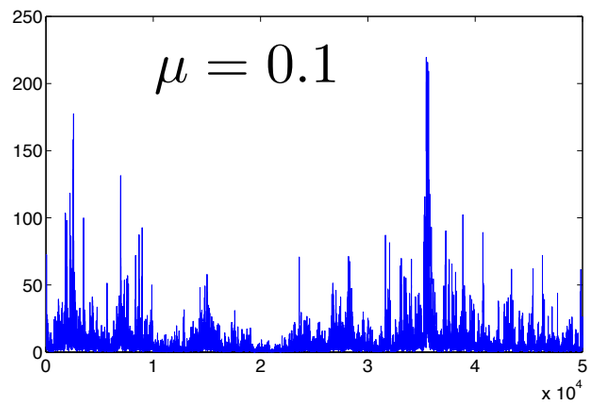



# Scattering Mandelbrot Cascades



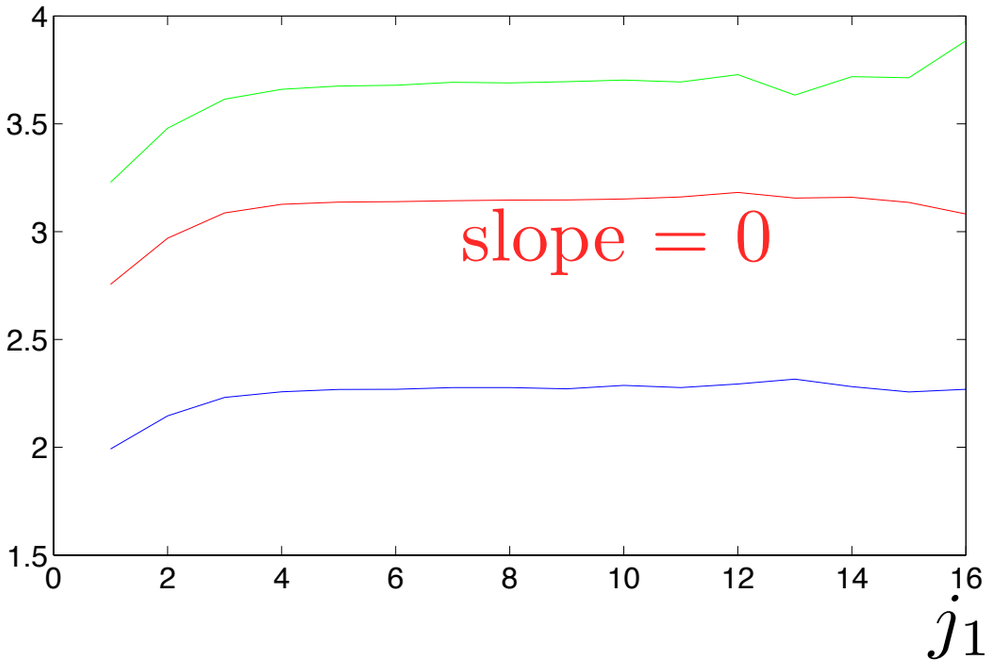




*J.Bruna, E.Bacry, J.F.Muzy*

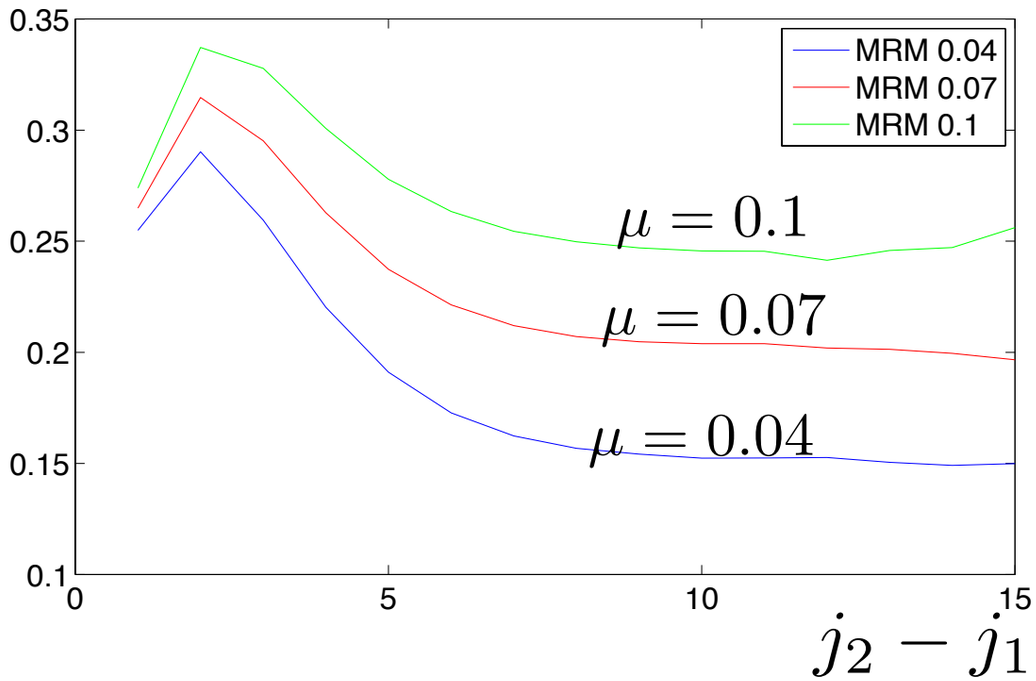
**Theorem:** Mandelbort Random Measures $dX$ satisfy:

$$\lim_{j_2 - j_1 \to \infty} \widetilde{S} dX(2^{j_1}, 2^{j_2}) = C_2 \, \mu \ .$$

$$\log \overline{S} dX(2^{j_1}) \sim C_1$$

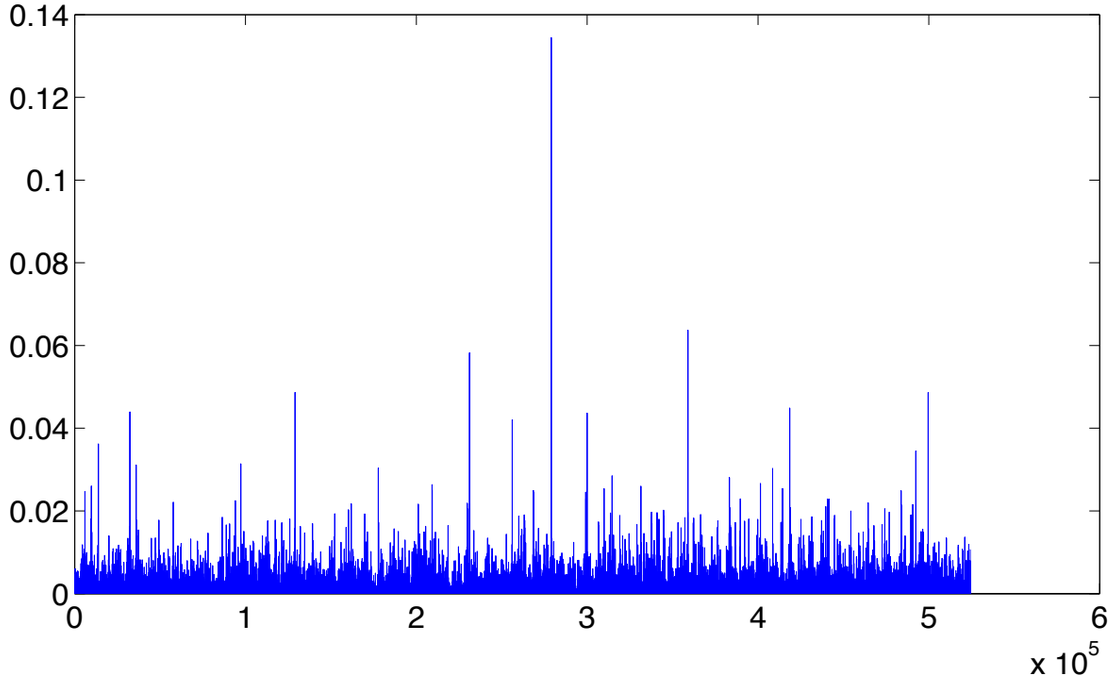



$$\log \widetilde{S} dX(2^{j_1}, 2^{j_2})$$

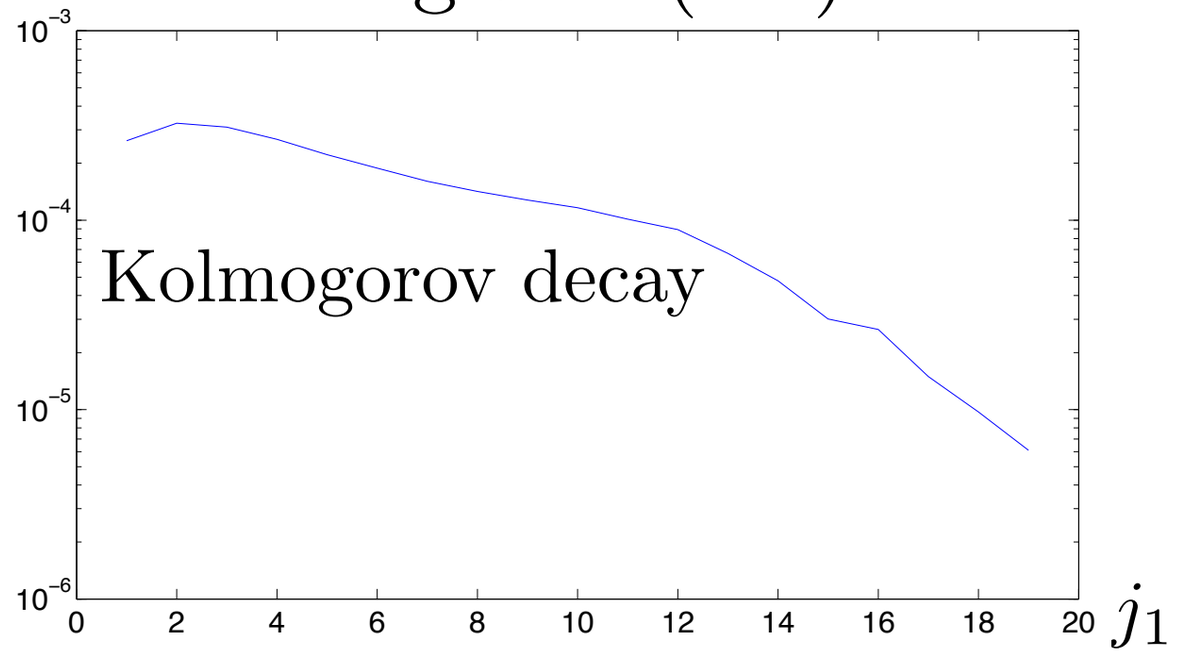


# Scattering Turbulence

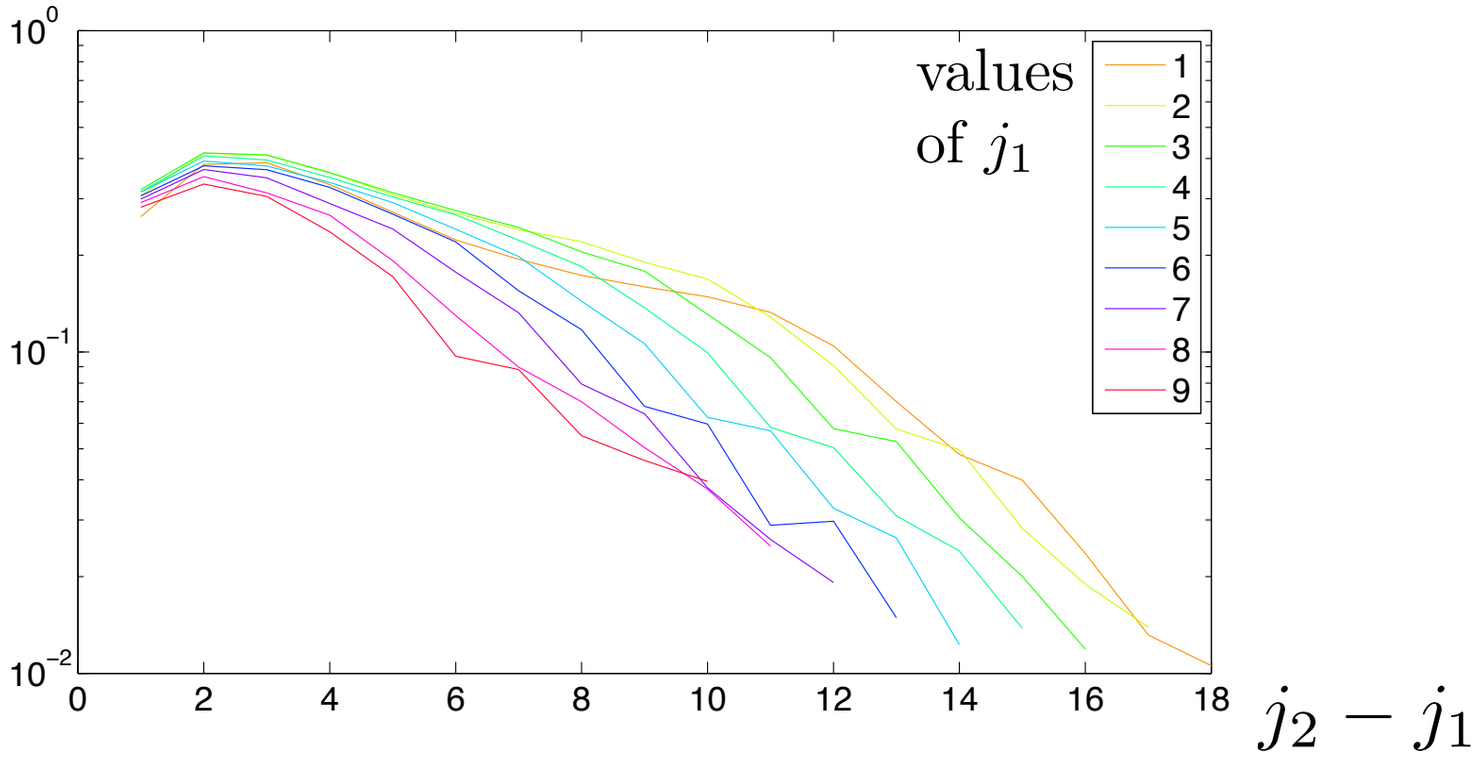
$dX(t)$ : dissipation energy




$\log \overline{S} dX(2^{j_1})$



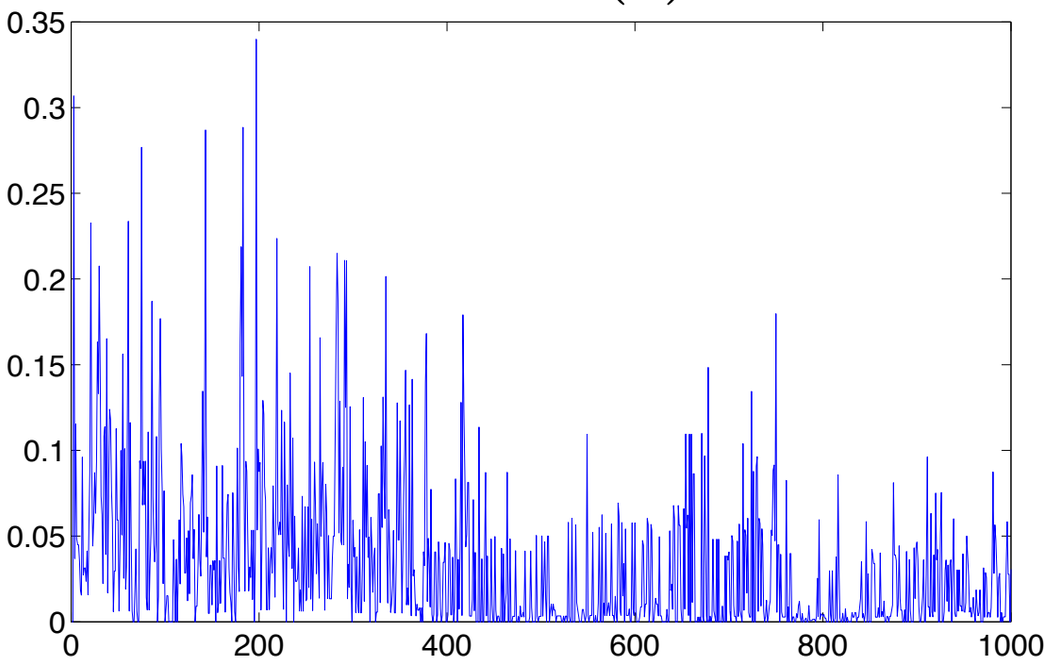

$\log \widetilde{S}(2^{j_1}, 2^{j_2}) \neq F(j_2 - j_1)$ : not self-similar.



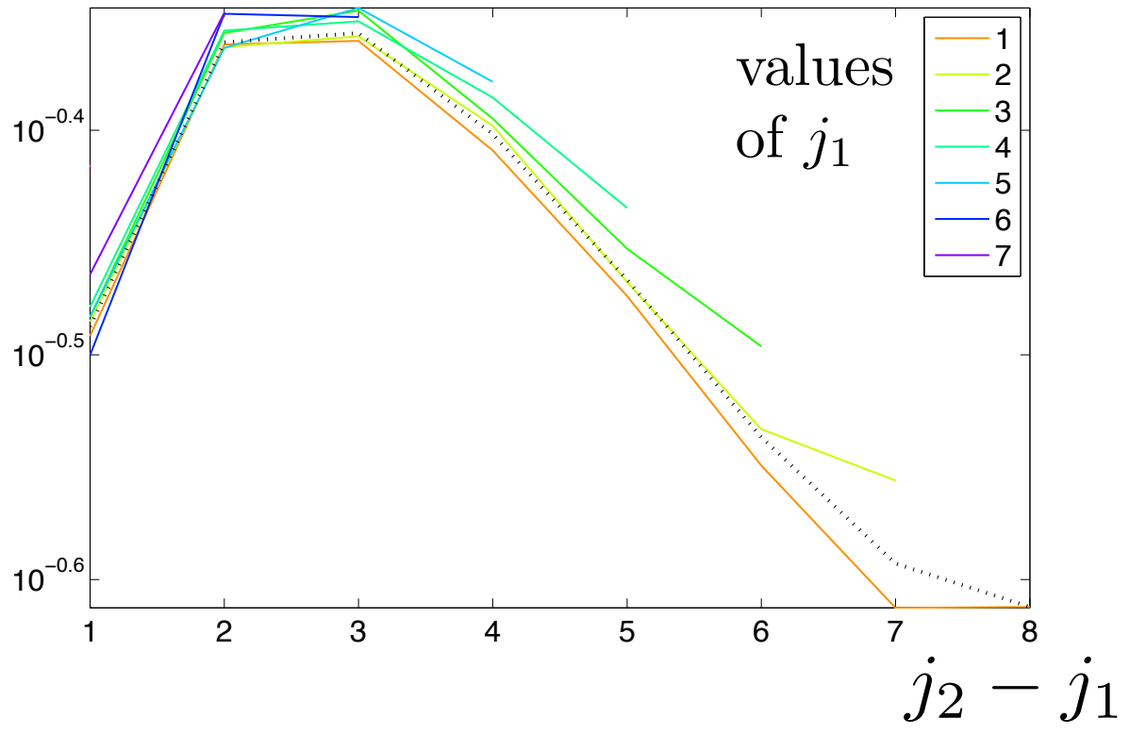
# Financial Time Series

1 Trading day of German Bund.

$dX(t)$




$\log \widetilde{S}(2^{j_1}, 2^{j_2}) \approx F(j_2 - j_1)$



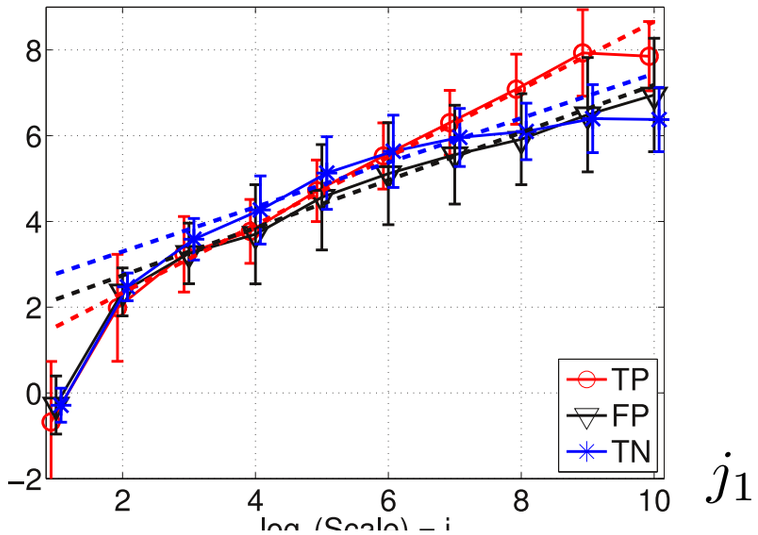
# Fetal Heart Rate Variability

*P. Abry, J. Anden, V. Chudacek, M. Doret*

Fetal heart rate monitoring gives information on the stress level of babies before delivery.

"Fractal behavior"

$\log \overline{S}X(2^{j_1})$

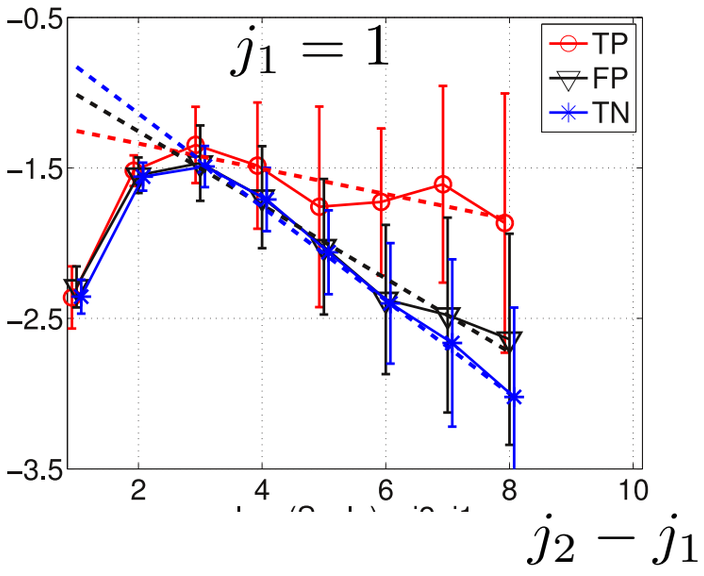


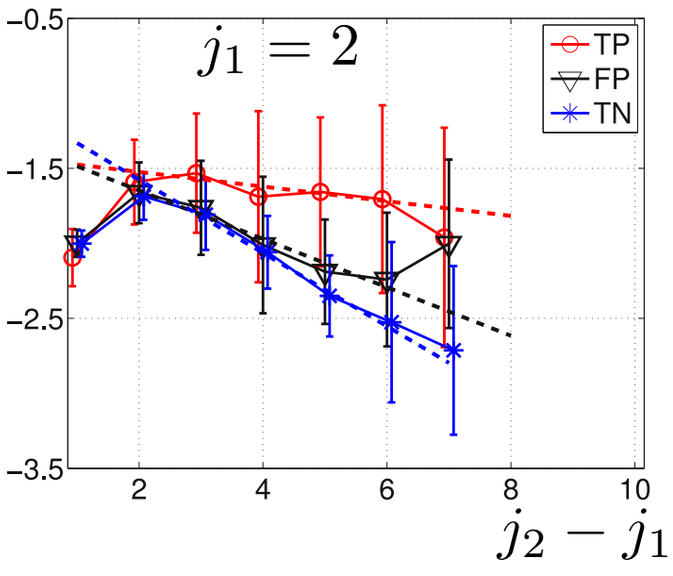

True Unhealthy

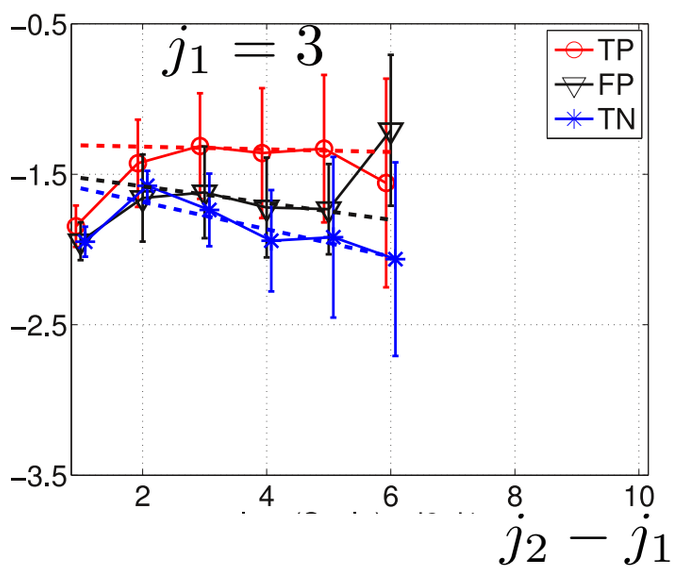
False Healthy

True Healty

$\log \tilde{S}X(2^{j_1}, 2^{j_2}) \neq F(j_2 - j_1) \Rightarrow$ Not self-similar



# Invariance to Lie Group Actions

- Invariance to a Lie group action and stability to diffeomorphisms
  - Translation and frequency transpositions
  - Translations and rotations
  - Tranvariance to translation-rotations and scaling

# Transposition Invariance

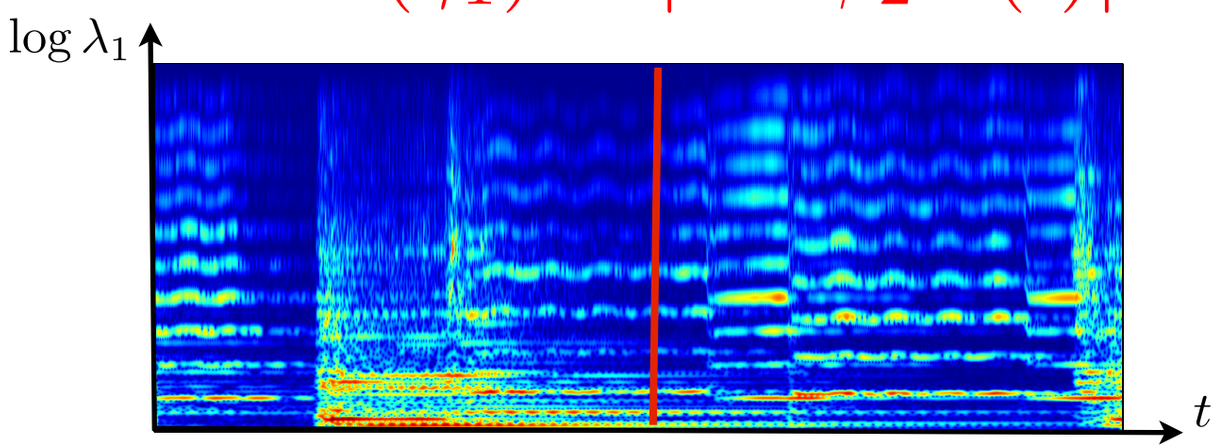
*J.Anden*

- Frequency transposition is a common source of variability
- Transposition $\Leftrightarrow$ translation and deformations in $\log \lambda_1$
- Invariance with a "frequency scattering" along $\log \lambda_1$

$$x(t) \longrightarrow \boxed{\text{Scattering along } t} \longrightarrow \boxed{\text{Scattering along } \log \lambda_1} \longrightarrow \Phi x$$

Scattering along log frequency $\gamma_1 = \log_2 \lambda_1$:

$$z(\gamma_1) = |x \star \psi_{2^{\gamma_1}}(t)|$$



# Genre Classification (GTZAN)

*J.Anden*

- GTZAN: music genre classification (jazz, rock, classical, ...) 10 classes and 30 seconds tracks.
- Each frame is classified using a Gaussian kernel SVM.

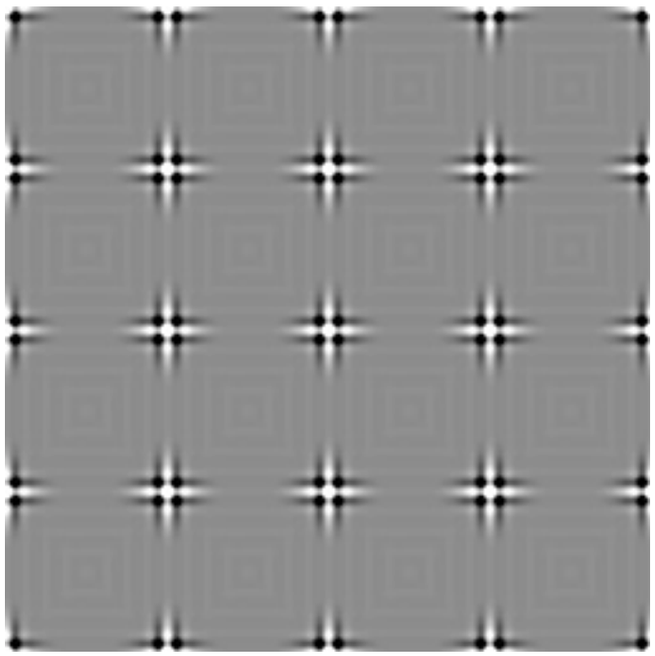
$$T = 370 \text{ ms}$$

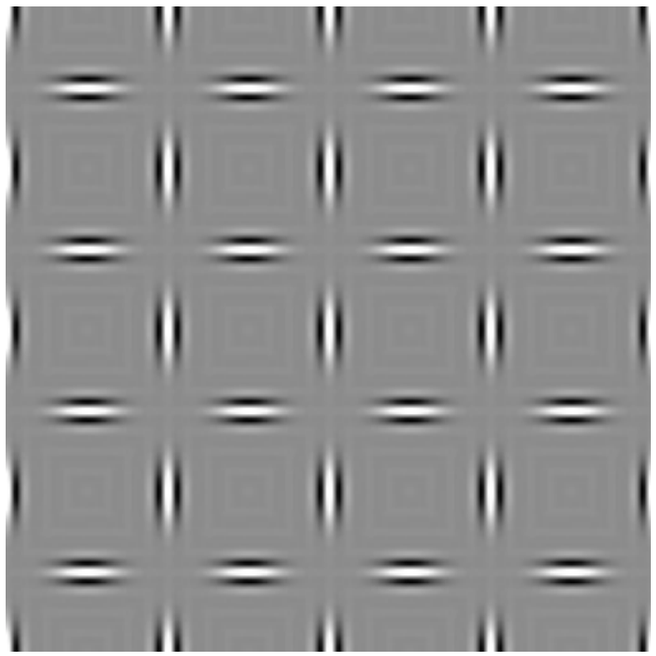
| Feature Set | Error (%) |
|---|---|
| Δ-MFCC (32 ms) | 19.3 |
| Time Scat., m = 1 | 17.9 |
| Time Scat., m = 2 | 12.3 |
| Time & Frequency Scat., m=2 | 10.3 |

# Joint versus Separable Invariants

- Separable cascade of invariants loose joint distributions.

- Separable rotation and translation invariants can not discriminate:





$\Rightarrow$ need to build invariant on the joint roto-translation group.

# Roto-Translation Group

- Roto-translation group $G = \{g = (r, t) \in SO(2) \times \mathbb{R}^2\}$

$$(r, t) \,.\, x(u) = x(r^{-1}(u - t))$$

- Group multiplication:

$$(r', t') \,.\, (r, t) = (r'r \,,\, r't + t') \;\; : \;\; \text{not commutative.}$$

- An averaging invariant is convolution on $\mathbf{L^2}(G)$: $x(g) = x(r, t)$

for translations: $\phi \star x(t) = \int_{\mathbb{R}^2} \overline{\phi}(t' - t)\, x(t')\, dt'$

for roto-translations: $\phi \star x(g) = \int_G \overline{\phi}(g'^{-1} g)\, x(g')\, dg'$

- Roto-translation Haar measure : $dg = dt\, d\theta$ (rotation angle $\theta$)

# Scattering on a Lie Group

*L. Sifre*

- One can define separable complex wavelets $\overline{\psi}_{\lambda_2}(r,t) \in \mathbf{L^2}(G)$

$$W_2 x = \begin{pmatrix} x \circledast \overline{\phi}(r,t) \\ x \circledast \overline{\psi}_{\lambda_2}(r,t) \end{pmatrix}_{\lambda_2,r,t} \quad \text{is unitary over } \mathbf{L^2}(G).$$

- A roto-translation scattering applies

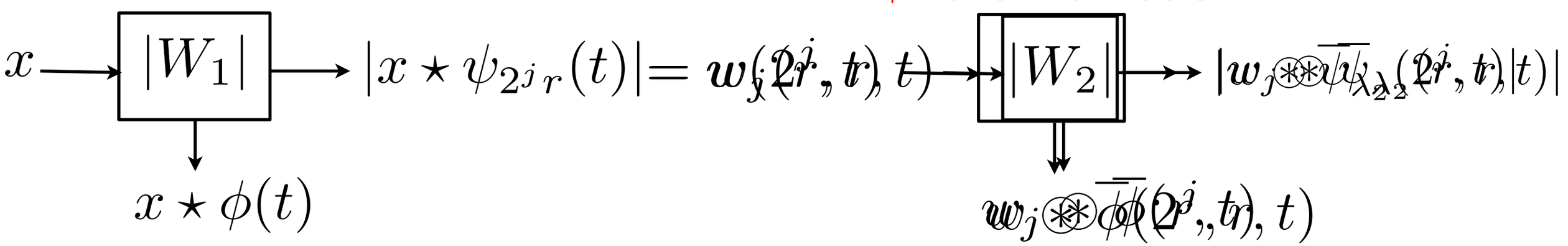
$$|W_2|x = \begin{pmatrix} x \circledast \overline{\phi}(r,t) \\ |x \circledast \overline{\psi}_{\lambda_2}(r,t)| \end{pmatrix}_{\lambda_2,r,t} \quad \text{and } |W_m| = |W_2| \text{ for } m \geq 2.$$

translation

scalo-roto-translation

roto-translation

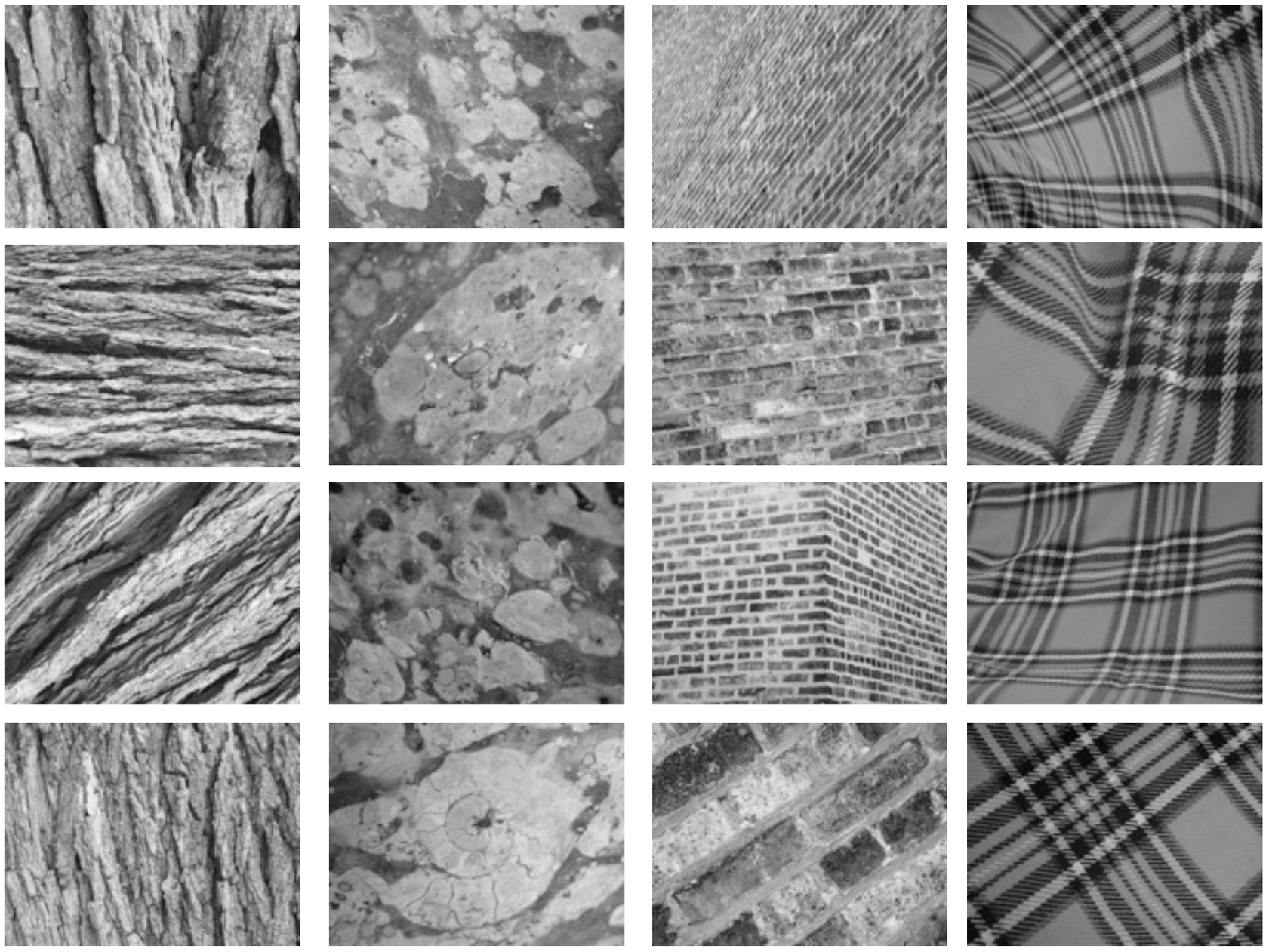
+ renormalization



# Rotation and Scaling Invariance

*Laurent Sifre*

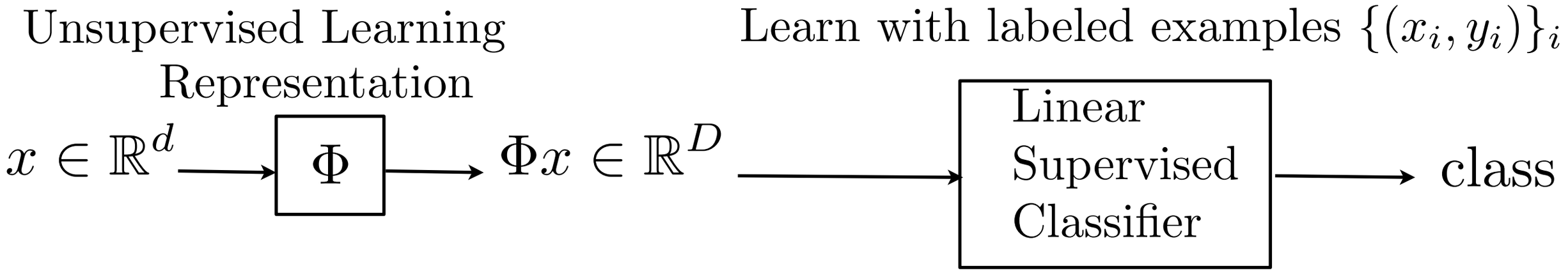
UIUC database:
25 classes



Scattering classification errors

| Training | Translation | Supervised Linear Transl + Rotation Classifier: PCA/SVM | + Scaling |
|---|---|---|---|
| 20 | 20 % | 2% | **0.6**% |

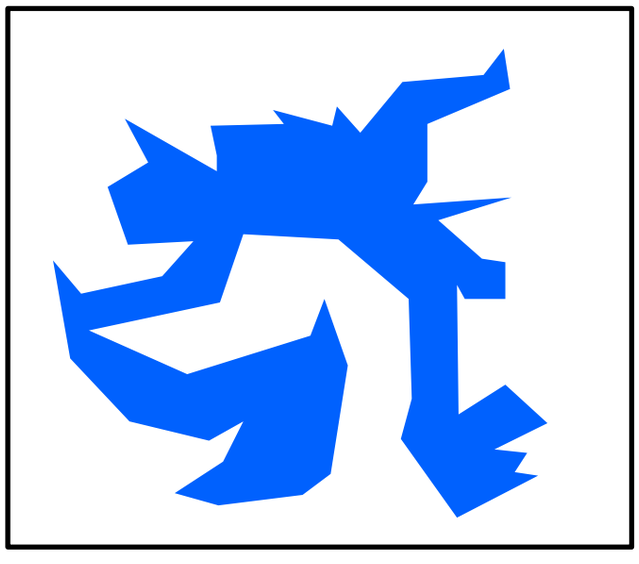
# Learning Representations




- **Unsupervised learning** of $\Phi$ from unlabeled examples $\{x_i\}$:

  - model the $\{x_i\}_i$ as realization of a random vector $X \in \mathbb{R}^d$

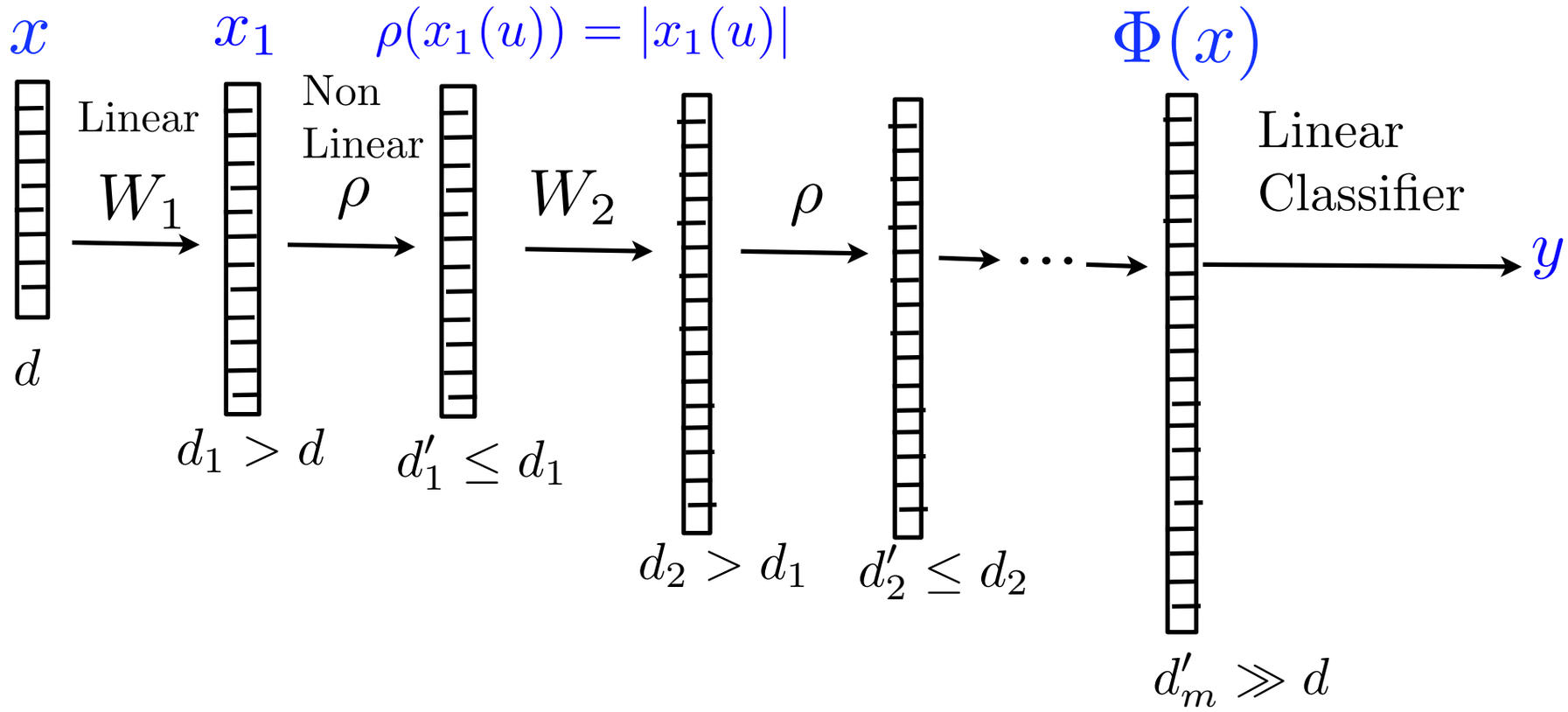
  - adapt $\Phi$ to the high-dimensional distribution $p(x)$ of $X$



but we can not estimate $p(x)$...

# Scattering Generalization

- Towards general deep networks:



# Revisit Expected Scattering

- Expected wavelet scattering transform:

$$\overline{S}X = \begin{pmatrix} E(X) = E(X_0) \\ E(|X \star \psi_{\lambda_1}|) = E(X_1) \\ E(||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|) = E(X_2) \\ E(|||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}|) = E(X_3) \\ \dots \end{pmatrix}_{\lambda_1,\lambda_2,\lambda_3,\dots}$$

Initialize $X_0 = X$

For $W_m Z = \left( \sum_n Z(n) \ , \ Z \star \psi_\lambda(n) \right)_\lambda$ iteratively compute

$$X_1 = |W_1(X_0 - E(X_0))|$$
$$X_2 = |W_2(X_1 - E(X_1))|$$
$$X_3 = |W_3(X_2 - E(X_2))|$$
$$\dots$$

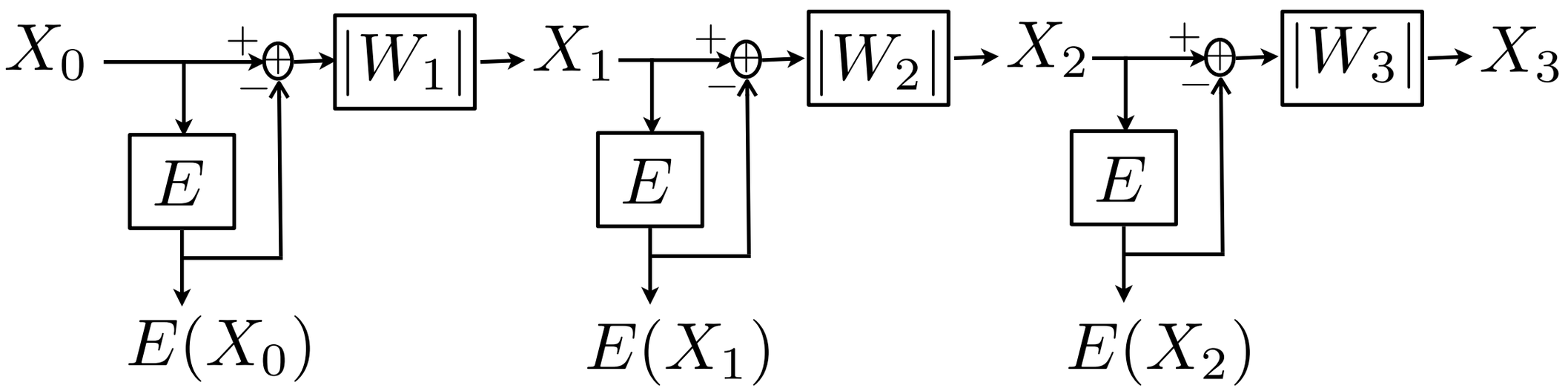
- Expected scattering: $\overline{S}X = \left( E(X_m) \right)_{m \in \mathbb{N}}$

# Generalized Scattering

- Define $W_m x = \Big( \langle x, \theta_n \rangle \Big)_{n \leq N_{m+1}}$ from $\mathbb{R}^{N_m}$ to $\mathbb{C}^{N_{m+1}}$: HOW ?

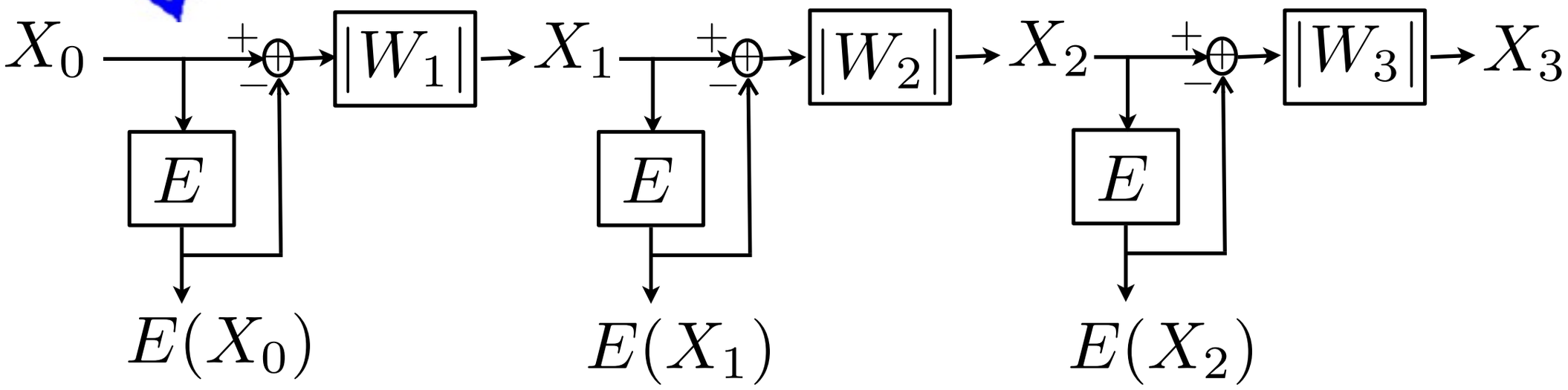
Tight frame: $\sum_n |\langle x, \theta_n \rangle|^2 = \|x\|^2 \Leftrightarrow W_m^* W_m = Id$

$$\begin{aligned} X_m &= |W_m(X_{m-1} - E(X_{m-1})| \\ &= \Big( |\langle X_{m-1} - E(X_{m-1}), \theta_n \rangle| \Big)_n \end{aligned}$$




- Expected scattering transform: $\overline{S}X = \{E(X_m)\}_{m \in \mathbb{N}}$

# Scattering Properties




$$\overline{S}X = \Big(E(X_m)\Big)_{m\in\mathbb{N}} \quad \text{and} \quad \|\overline{S}X\|^2 = \sum_{m\in\mathbb{N}} |E(X_m)|^2$$

- Since $W_m$ is a tight frame operator $\|W_m x\| = \|x\|$ and

$$\||W_m|x - |W_m y|\| \leq \|x - y\|$$

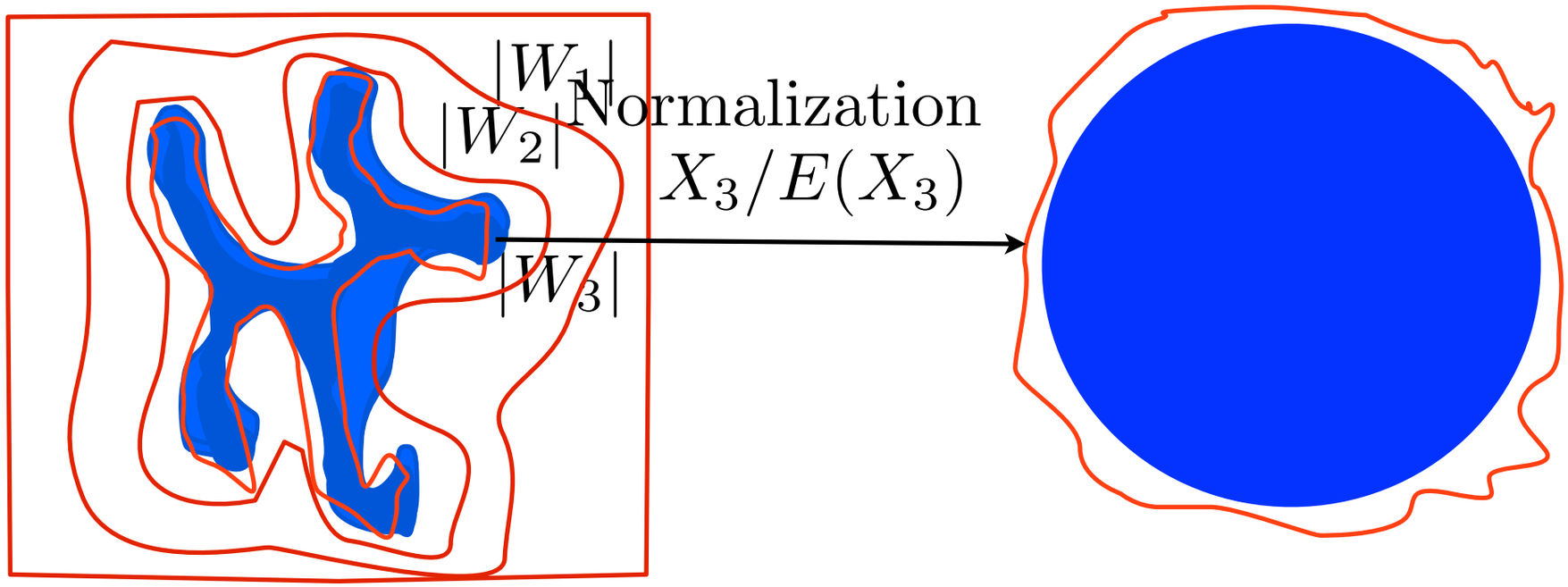
**Theorem:** *I. Waldspurger*

$$\|\overline{S}X - \overline{S}Y\| \leq E(\|X - Y\|^2)$$

$$\|\overline{S}X\| = E(\|X\|^2)$$

# Optimized Space Contraction

- A generalized scattering progressively contracts the space
- For classification, we need to squeeze the space while minimizing the data volume reduction




**Proposition:** The data volume reduction at layer $m$ is

$$E(\|X_{m-1} - E(X_{m-1})\|^2) - E(\|X_m - E(X_m)\|^2) = \|E(X_m)\|^2$$

$$\Rightarrow \text{ for all } m \text{ minimize } \|E(X_m)\| \ .$$

# Sparse Layerwise Learning

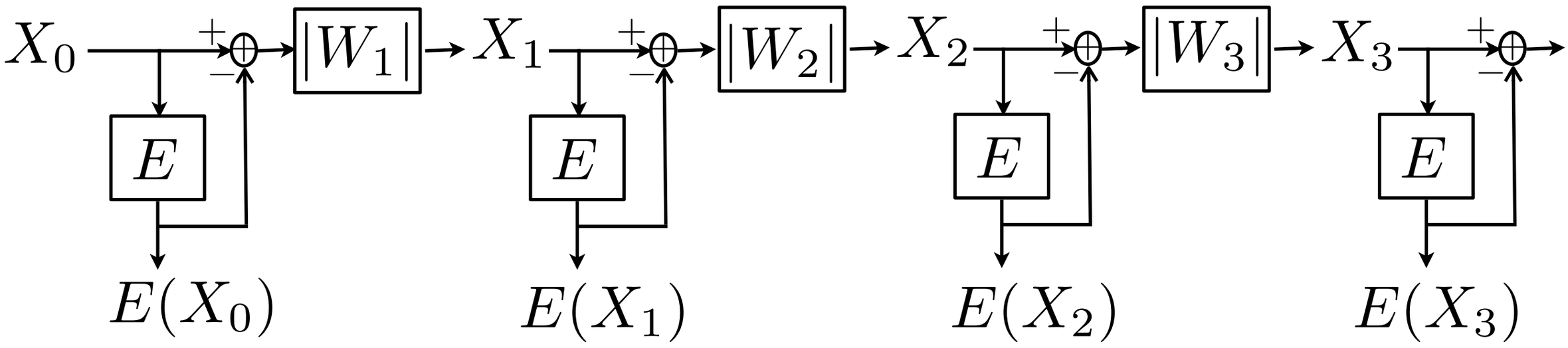
$$X_m = |W_m(X_{m-1} - E(X_{m-1})| \quad \text{with } W_m^* W_m = Id.$$

- Given $X_{m-1} - E(X_{m-1})$ we compute $W_m$ by minimizing

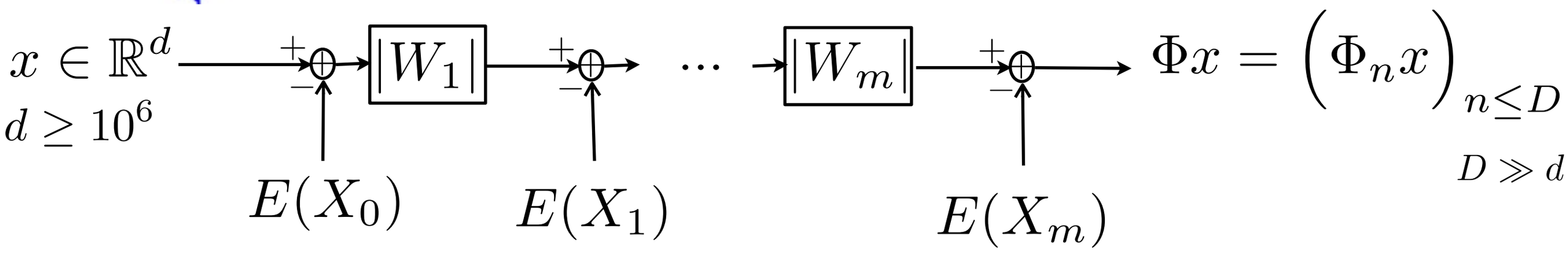
$$\|E(X_m)\| = \Big\| \underbrace{E\Big(|W_m(X_{m-1} - E(X_{m-1})|\Big)}_{\mathbf{l^1} \text{ norm across realizations}} \Big\|$$

$\Rightarrow W_m$ defines a sparse representation of $X_{m-1} - E(X_{m-1})$

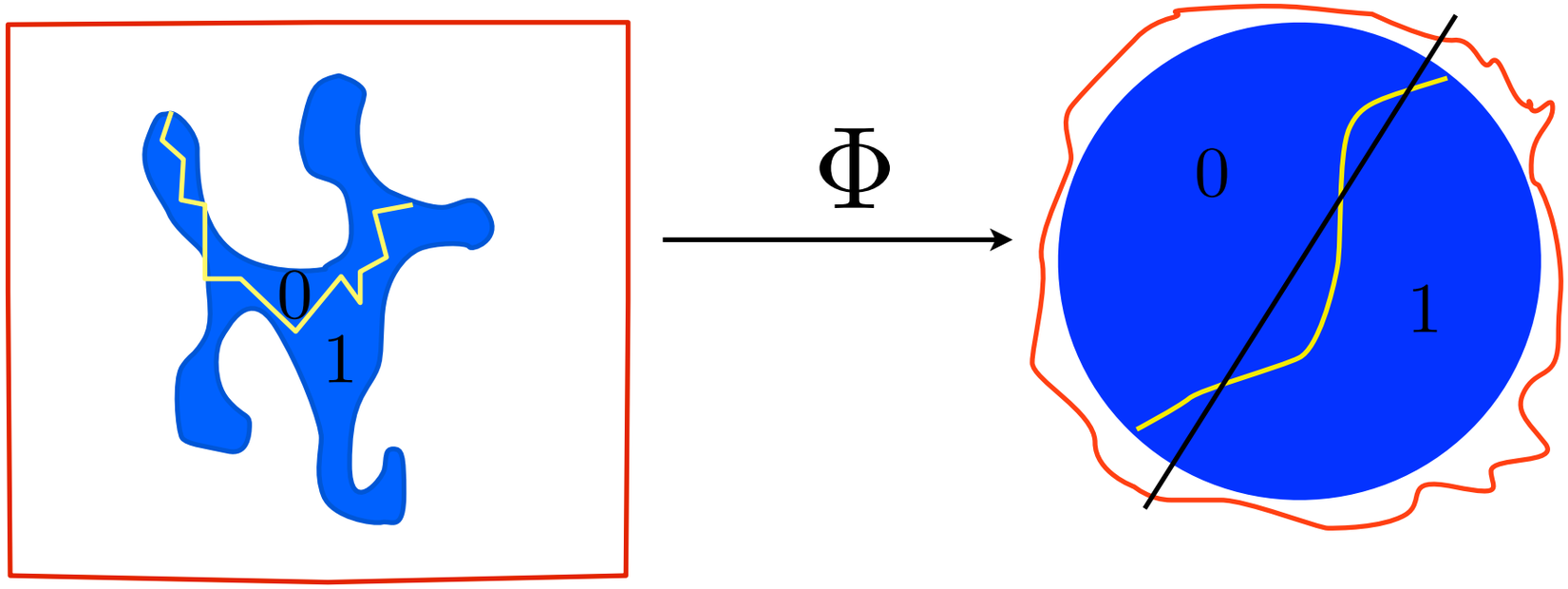
Sparse dictionary learning problem.



# Supervised Linear Classifiers




Binary classification
$y(x) = 0$ or $1$




- A linear classifier approximates the frontier of $y(x)$ by

$$\sum_n w_n \, \Phi_n x$$

- Problems of functional approximations in high dimension...

# Conclusion

- High dimensional classification algorithms have considerably improved in the last few years with many applications.

- Beautiful problems but lack of mathematics and mathematicians working in this area.

- **Papers and Softwares:** www.di.ens.fr/data/scattering