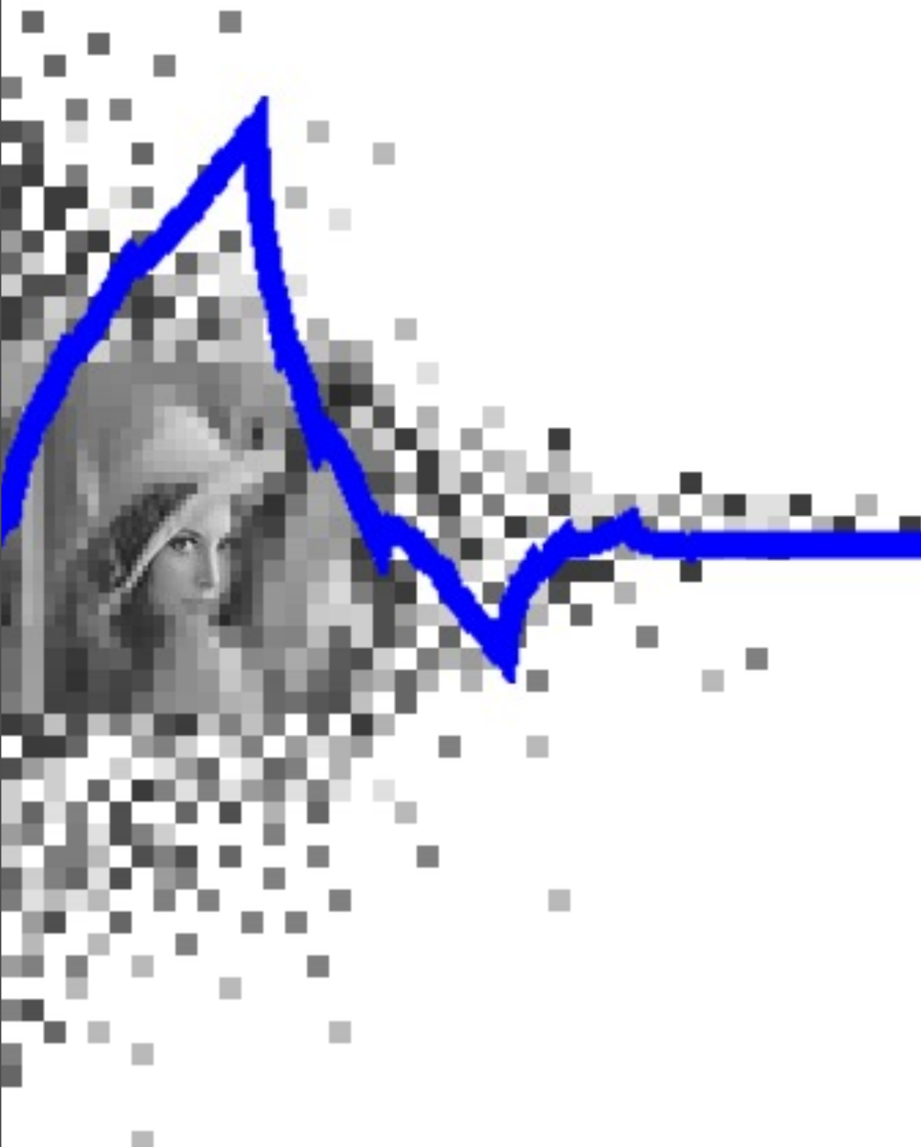
Course on High Dimensional Signal Analysis

Peyresq 2014



Stéphane Mallat

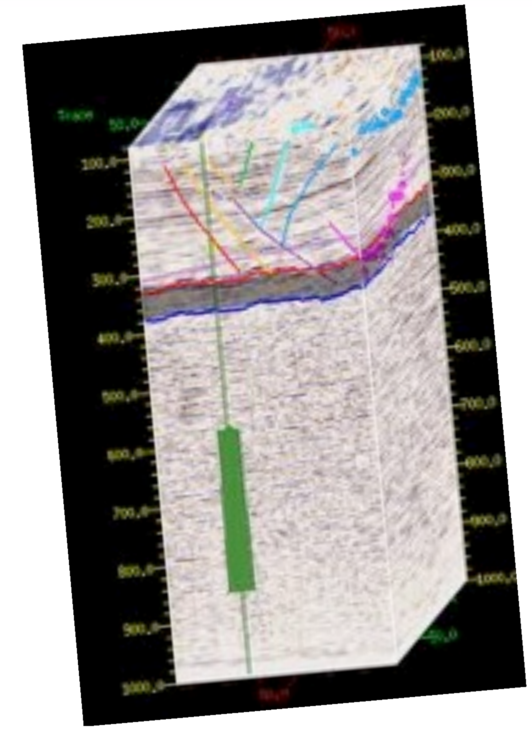
École Normale Supérieure

www.di.ens.fr/data/scattering

Deluge de Signaux

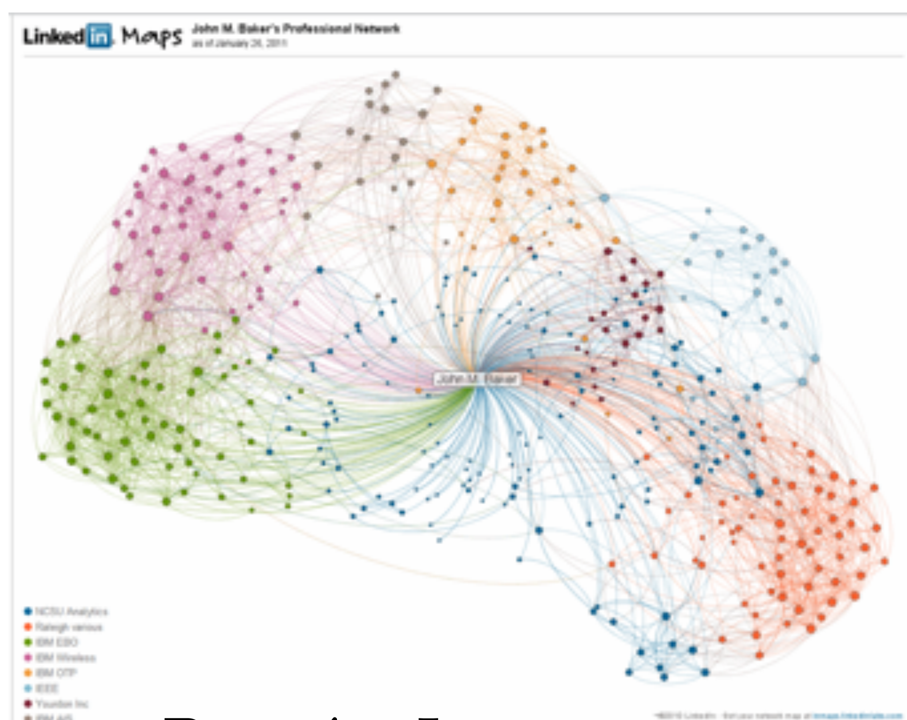
Audio

Video

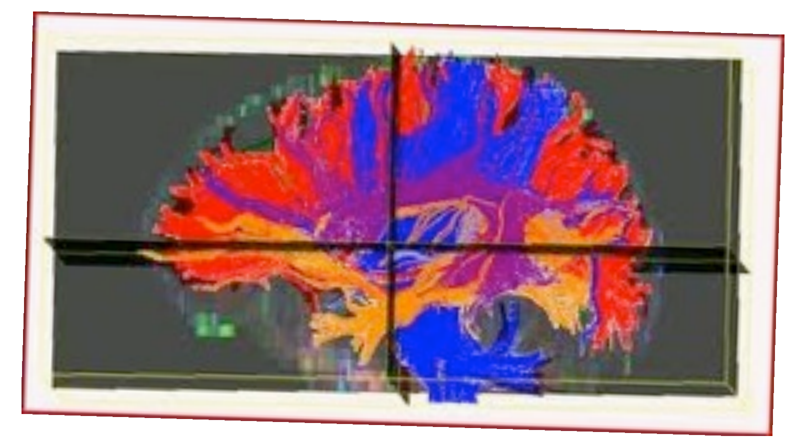


Seismic data

Satellite images



Données Internet



Medical data

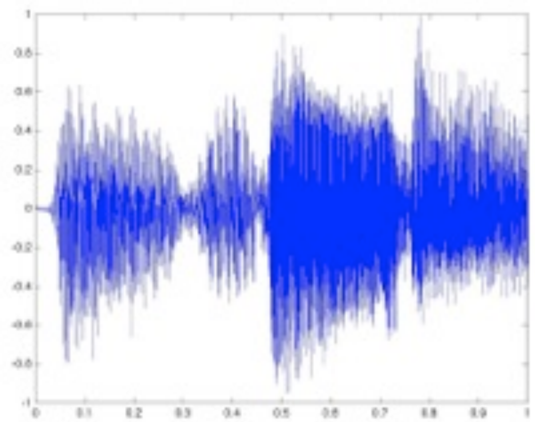
- Comment analyser automatiquement ces informations ?

High Dimensional Analysis

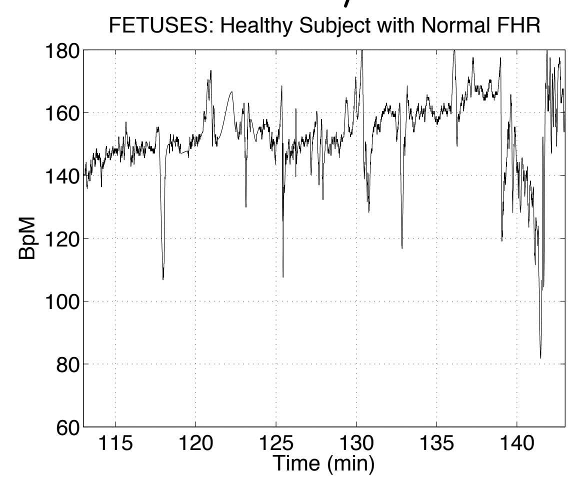
- High-dimensional signals $x = (x(1), ..., x(d))$:

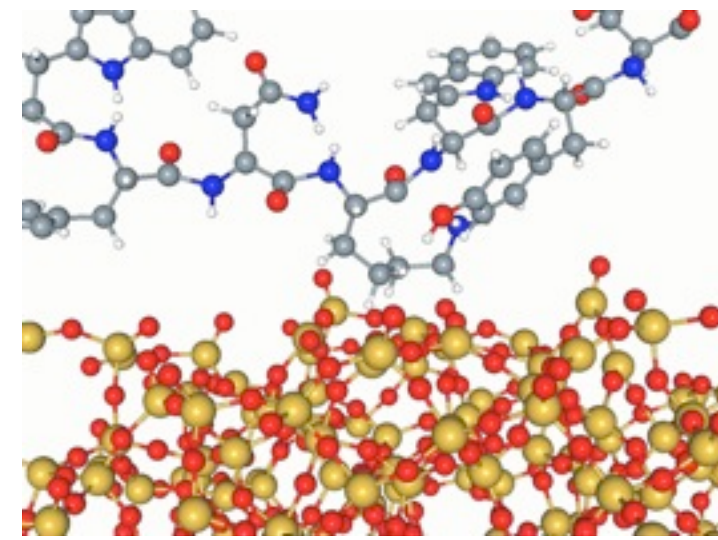
$d = 10^6$

$d = 10^4/s$



$d = 10^4/hour$





- *Supervised learning*: learn a function $f(x)$ (class label) given n sample values $\{x_i\, ,\, y_i = f(x_i)\}_{i \leq n}$

Different Class of Problems

- Unsupervised problems: we know examples $\{x_i\}_i$ and want to estimate a probability density $p(x)$ or clusters.

- Supervised problems: we know examples $\{x_i\}_i$ and their values $f(x_i)$ and want to estimate $f(x)$ for all $x \in \Omega$.

Different Level of Complexity:

- Low dimensional problems: $\dim(\Omega)$ is small.

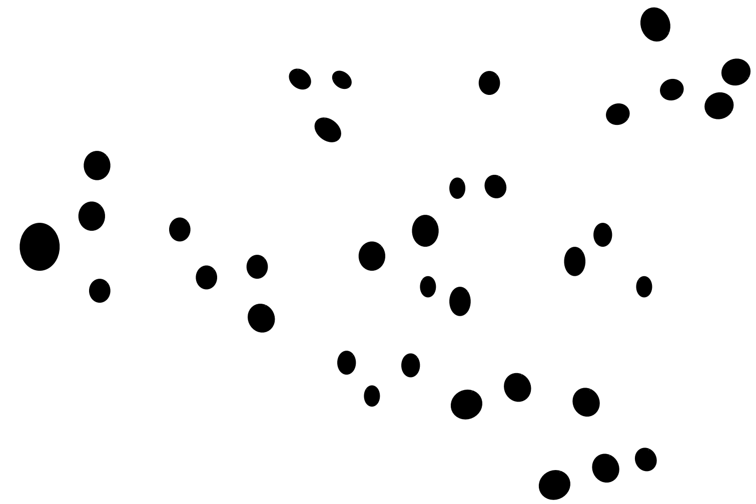
- High-dimensional but separable: $f(x) = \prod_{k=1}^{d} f_k(w_k.x)$

with $1D$ functions $f_k(u)$ for $u \in \mathbb{R}$ $\Rightarrow$ indépendant components.

- High-dimensional Ω with many interactions: OUR PROBLEM.

Many Body Interactions

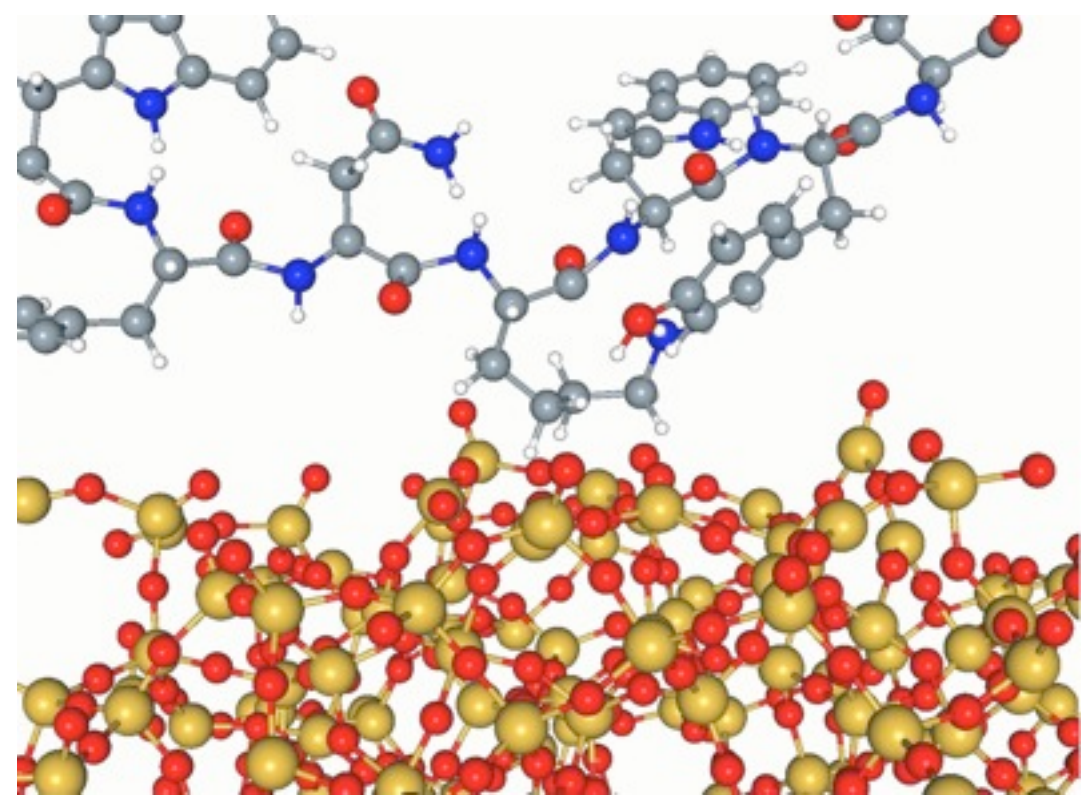
Long range interactions:
each body interacts
with the d others



Interaction energy $f(x)$ of a system $x = \Big\{ \text{positions, values} \Big\}$

Astronomy Masses

Quantum Chemistry Charges



Découvrir la Physique

- Recherche des lois de la physique: synthèses d'observations.

  Intelligences exceptionnelles: Newton, Maxwell, Einstein...

- Vraiment ?

Electric Eel
500 V

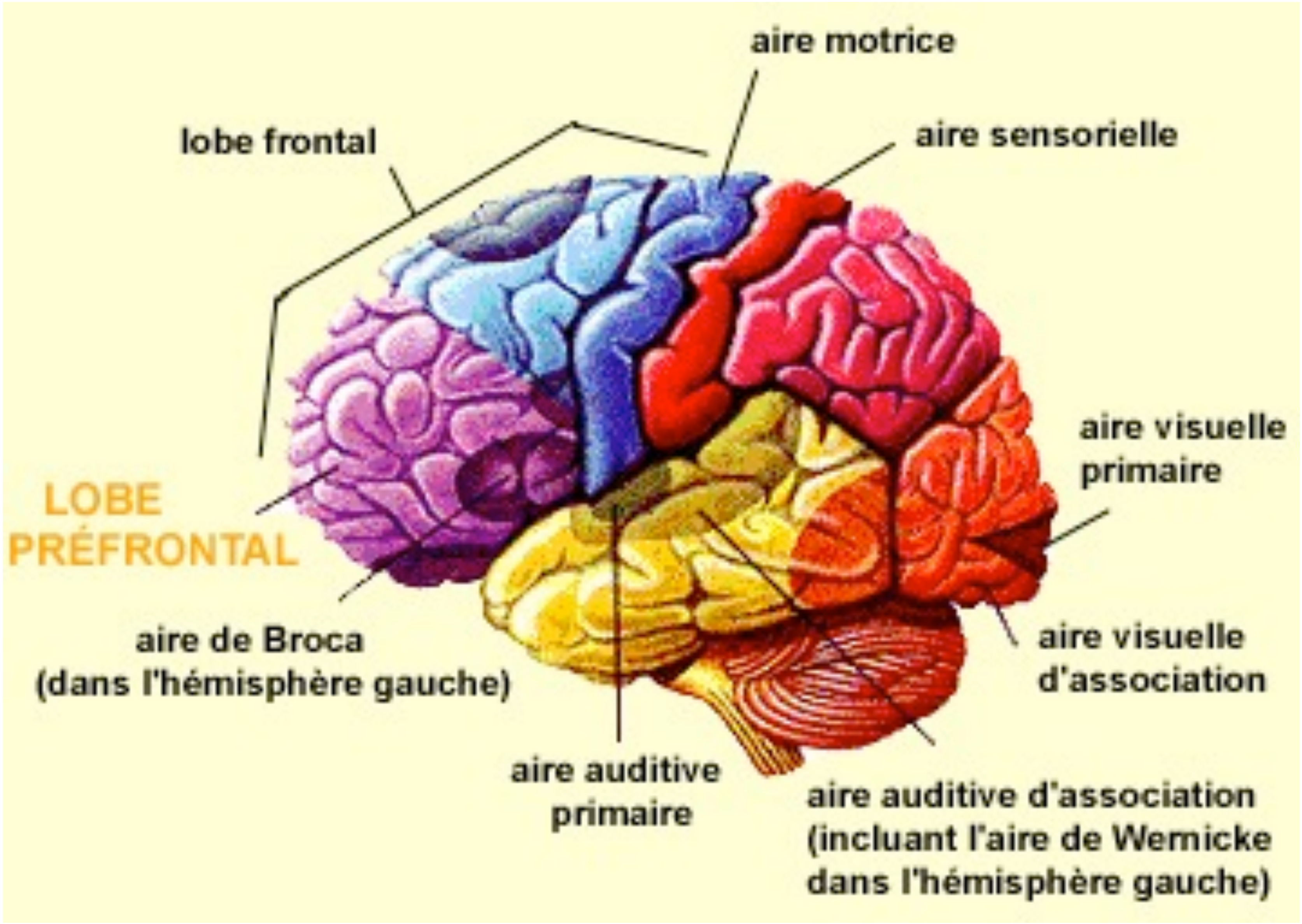
Blackghost Knifefish
1 mV

Un poisson peut résoudre les équations de l'électromagnétique bien mieux que nous.

- On peut apprendre directement à partir des données,

  mais il en faut **beaucoup, beaucoup**...

Une Architecture de Traitement de Données Massives



- Comment et quoi apprendre ?
- Pourquoi faut il beaucoup de données et de mémoire ?

Overview

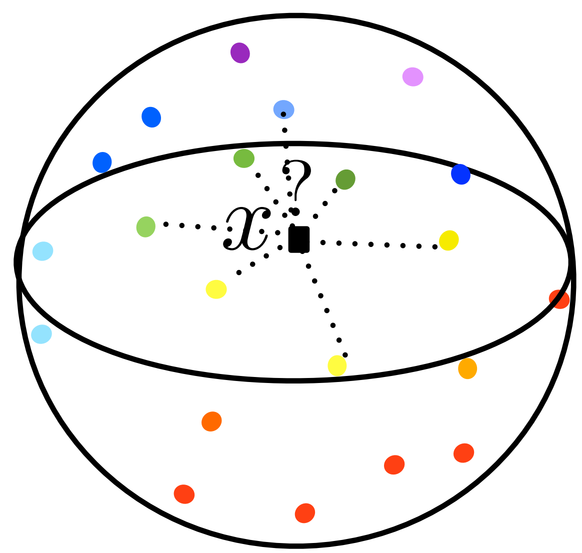
- I- Curse of Dimensionality and Kernel Classifiers
  - Support vector machines
  - Kernels and metrics

- II- Deep Neural Networks and Wavelet Scattering Transforms
  - Geometric invariants
  - Iterated wavelet transforms
  - Image and audio classification, and physics learning
  - Stationary Processes: beyond Gaussian processes

- III- Unsupervised Kerenel Learning with Deep Neural Networks

Nearest Neighbor Approximations

- $f(x)$ can be approximated from examples $\{x_i\,,\ f(x_i)\}_i$ by local interpolation if f is regular and there are close examples:

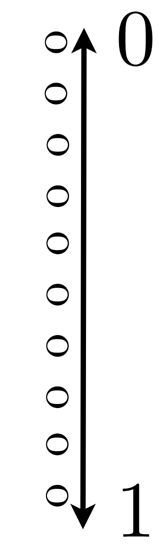


$$\text{Regression: } \tilde{f}(x) = \sum_i \alpha_i\, K(x, x_i)$$

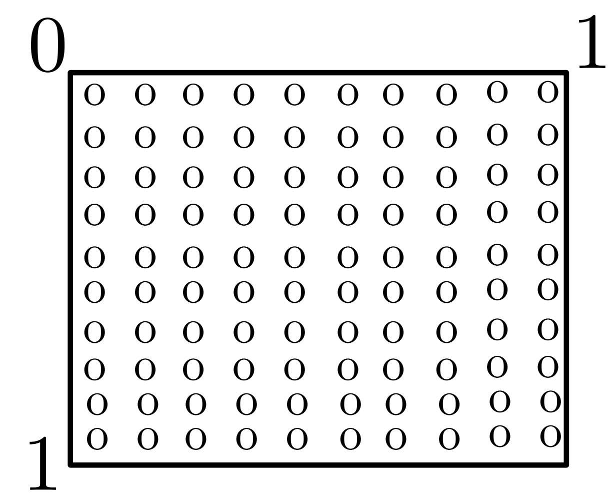
$$\text{Classification binaire: } \tilde{f}(x) = \text{sign}\Big(\sum_i \alpha_i\, K(x, x_i) \Big)$$

La Malédiction de la Dimensionalité

Problème des plus proche voisins: ils sont trés loin en grande dimension.



- il faut 10 points pour couvrir $[0,1]$ à intervalles 10^{-1}



- il en faut 100 pour $[0,1]^2$

- il en faut 10^d pour $[0,1]^d$: impossible si $d \geq 100$

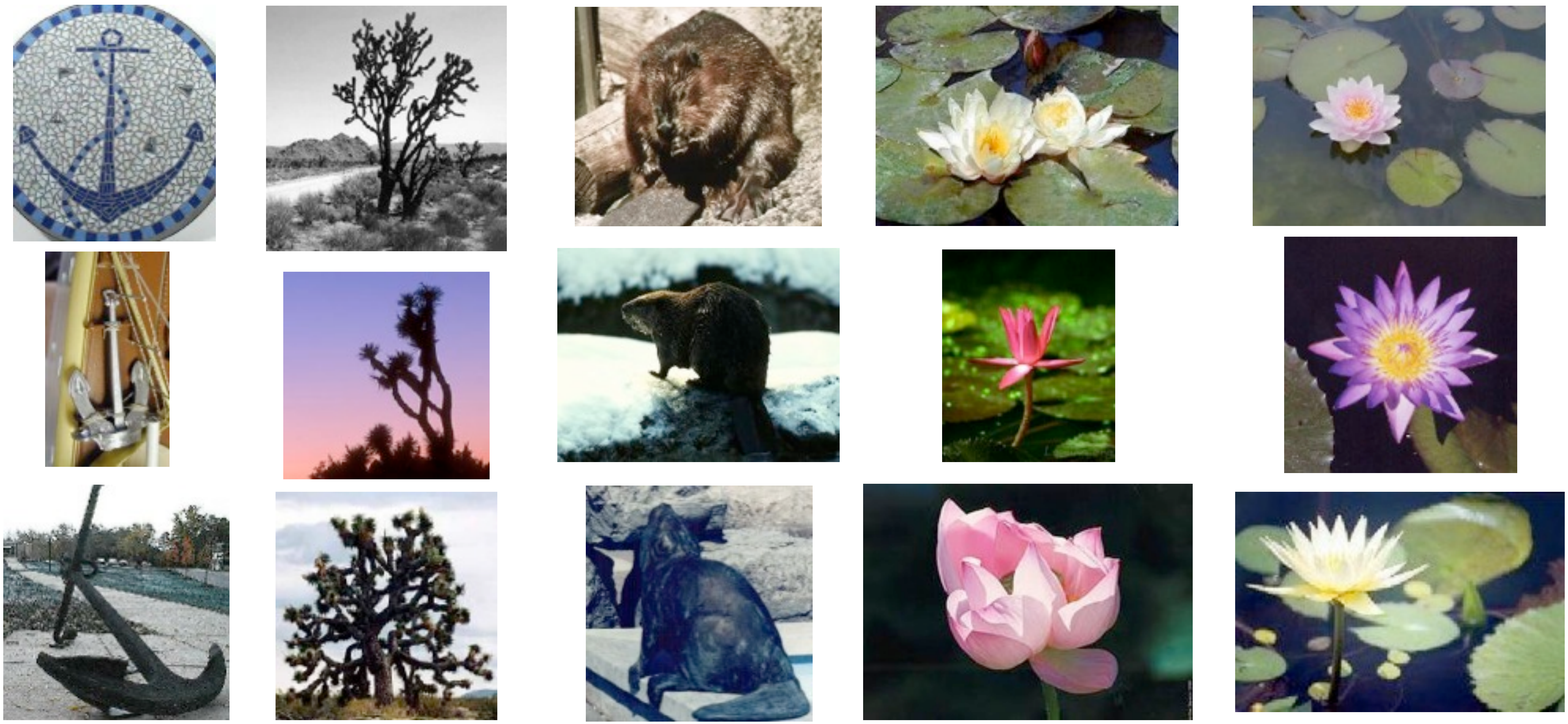
  10^{100}: plus que le nombre d'atomes dans l'univers.

$\Rightarrow$ on ne peut apprendre que des choses assez simples

High Dimensional Classification

CalTech 101

Anchor | Joshua Tree | Beaver | Lotus | Water Lily

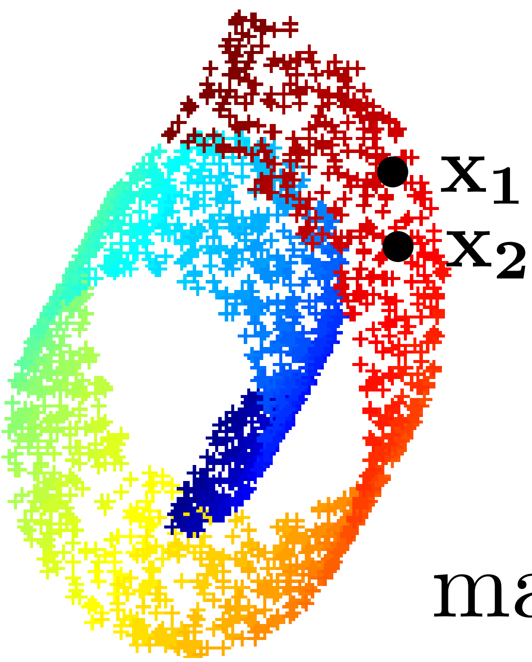


- Considerable variability in each class.
- Euclidean distances are meaningless
- Need to find **discriminative invariants**.

Signal Representation

- Signals x belong to subsets Ω of $\mathbb{R}^d$

- No dimensionality curse when Ω is low-dimensional

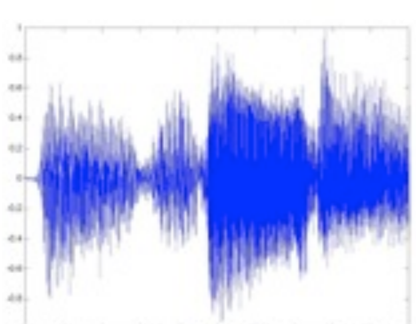


$\|x_1 - x_2\|$ is a good local measure of similarity

Finding Ω is a signal representation issue:

manifold learning or sparse dictionary representation

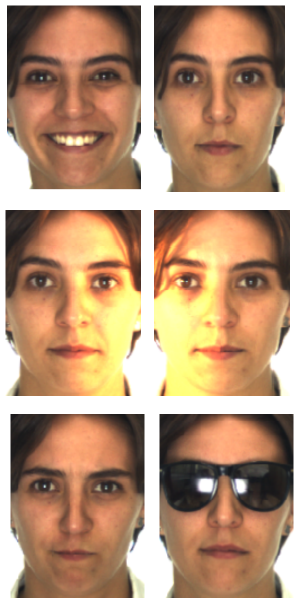
- For complex signals, Ω is most often high-dimensional:



$\Rightarrow$ Ω must be reduced depending on f.

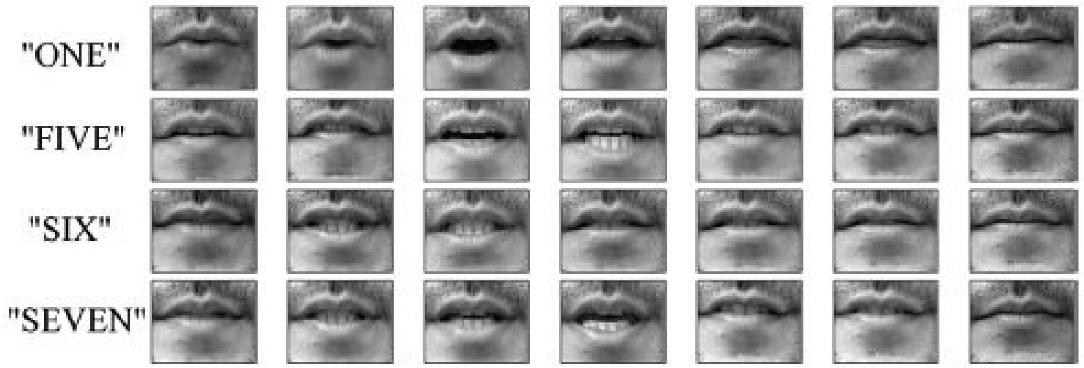
Low-Dimensional Data

- Face variations

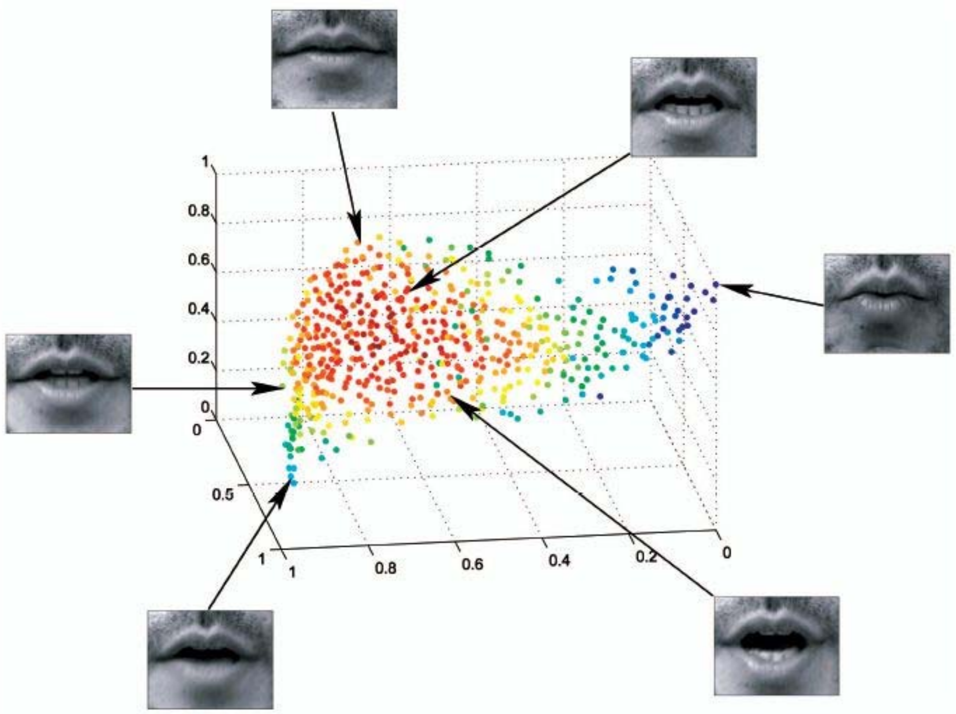


- Rigid motions

- Lips motion

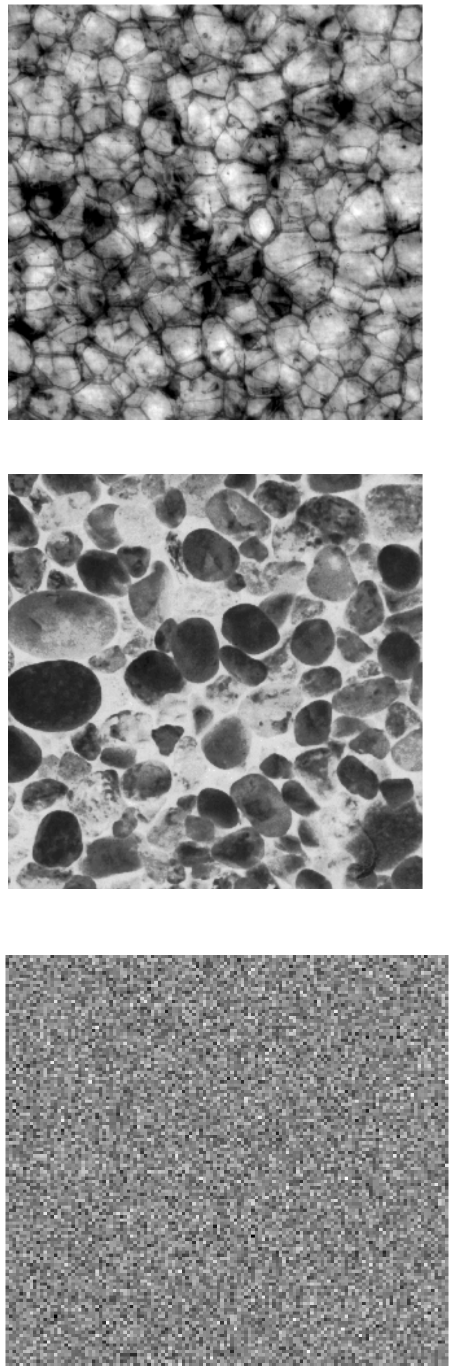


- Identify the manifold where the data lies.



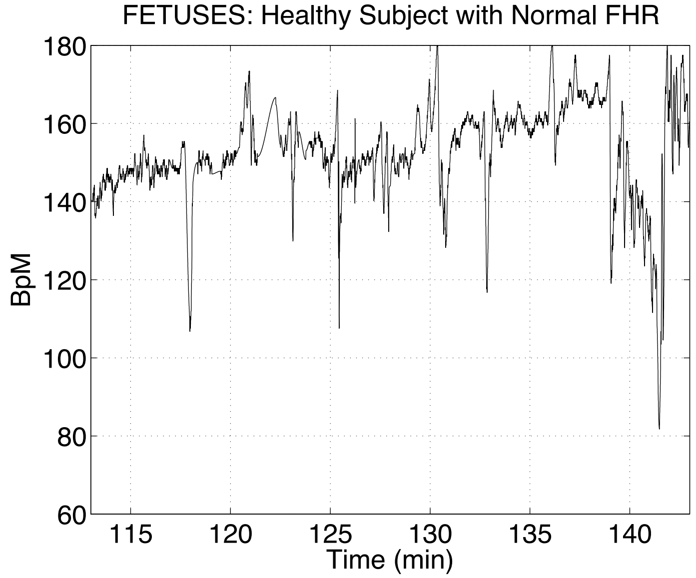
High-Dimensional Data

Textures

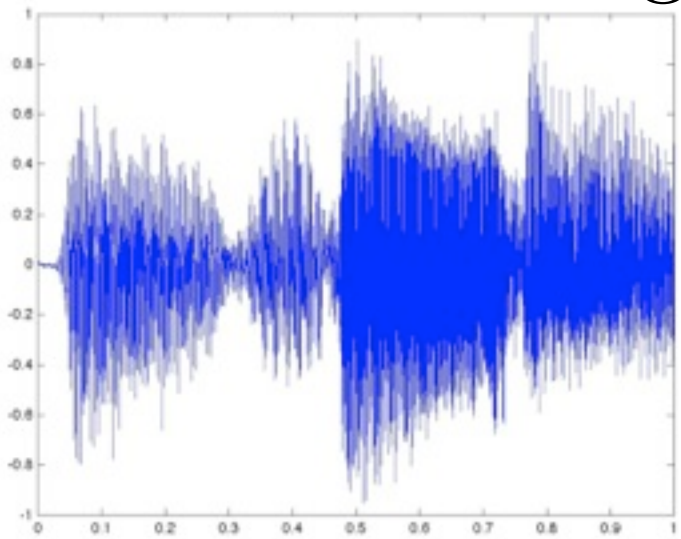


Beaver

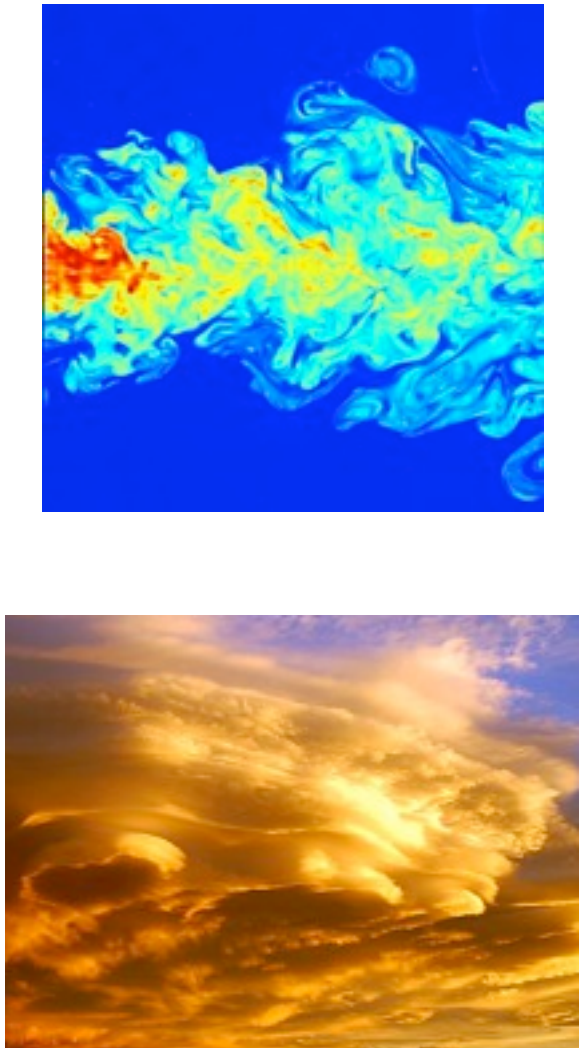
Electro-Cardiograms



Audio Recordings



Turbulences



- Need to eliminate irrelevant variability: compute invariants.

Linear and Kernel Classifiers

- Classifications can be reduced to multiple binary classifications $\text{sign}(f(x)) = \pm 1$

Training samples: $\{(x_i, y_i)\}_i$

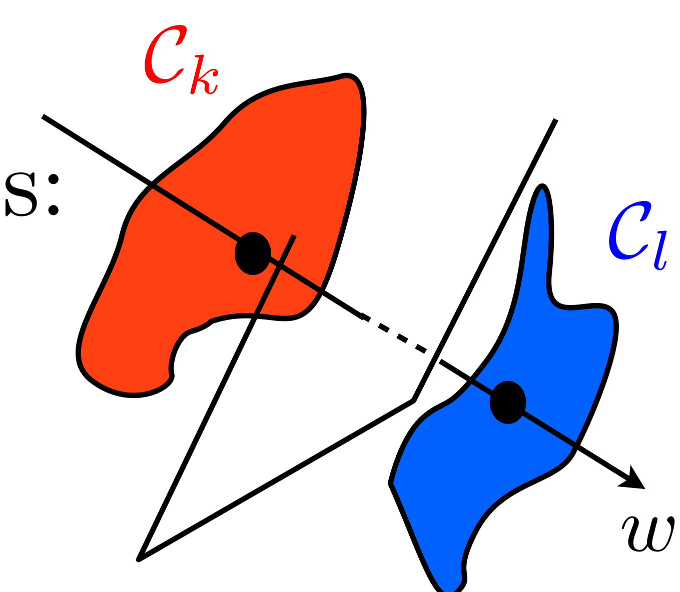
Representation

Supervised linear classification

$$\underset{\mathbb{R}^d}{x} \longrightarrow \boxed{\Phi} \longrightarrow \overset{\mathbb{R}^{d'}}{\Phi x} = \{\phi_n(x)\}_{n \leq d'} \longrightarrow \boxed{\langle \Phi x, w \rangle + b \overset{?}{\geq} 0} \longrightarrow \text{class}$$

$d' \gg d$

Hyperplance separation between pairs of classes:



$$f(x) = \langle \Phi x, w \rangle + b = \sum_n w_n \, \phi_n(x) + b$$

- (1) How to optimize (w, b) to minimize "errors" ?

SVM: $f(x)$ depends on kernel values $K(x, x_i) = \langle \Phi(x), \Phi(x_i) \rangle$.

- (2) How to define Φ to get linear discriminative invariants ?

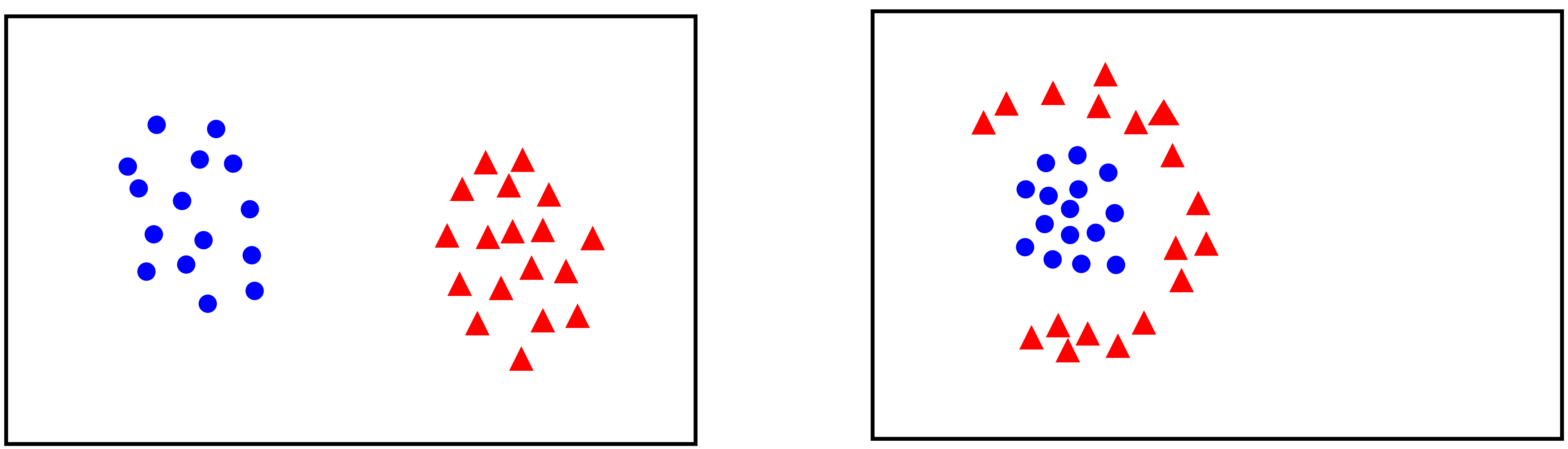
How to choose the Hyperplane ?

Given training data $(\mathbf{x}_i, y_i)$ for $i = 1 \dots N$, with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$, learn a classifier $f(\mathbf{x})$ such that

$$f(\mathbf{x}_i) \begin{cases} \geq 0 & y_i = +1 \\ < 0 & y_i = -1 \end{cases}$$

i.e. $y_i f(\mathbf{x}_i) > 0$ for a correct classification.

Linear decision boundary. Non-linear decision boundary.

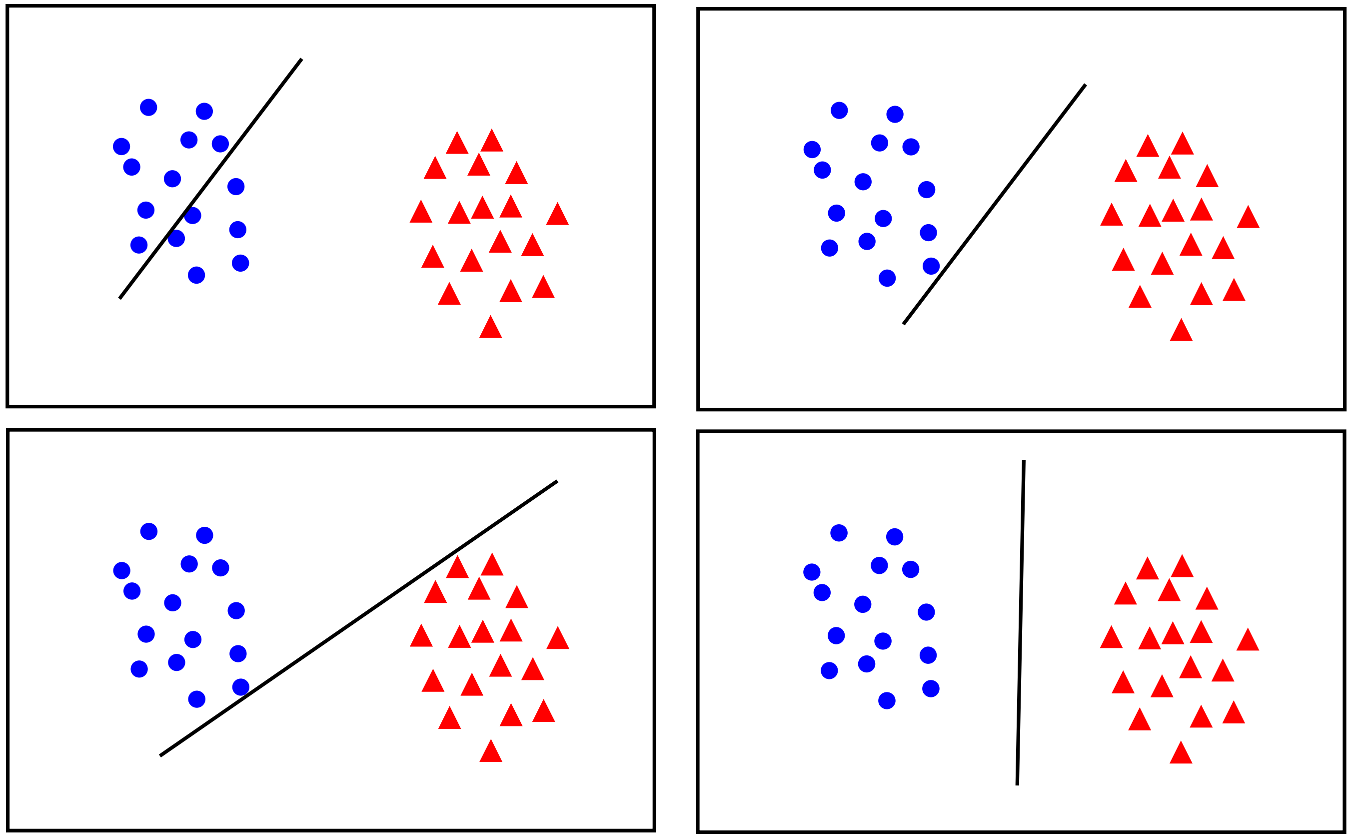


How to choose the Hyperplane ?

Given training data $(\mathbf{x}_i, y_i)$ for $i = 1 \ldots N$, with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$, learn a classifier $f(\mathbf{x})$ such that

$$f(\mathbf{x}_i) \begin{cases} \geq 0 & y_i = +1 \\ < 0 & y_i = -1 \end{cases}$$

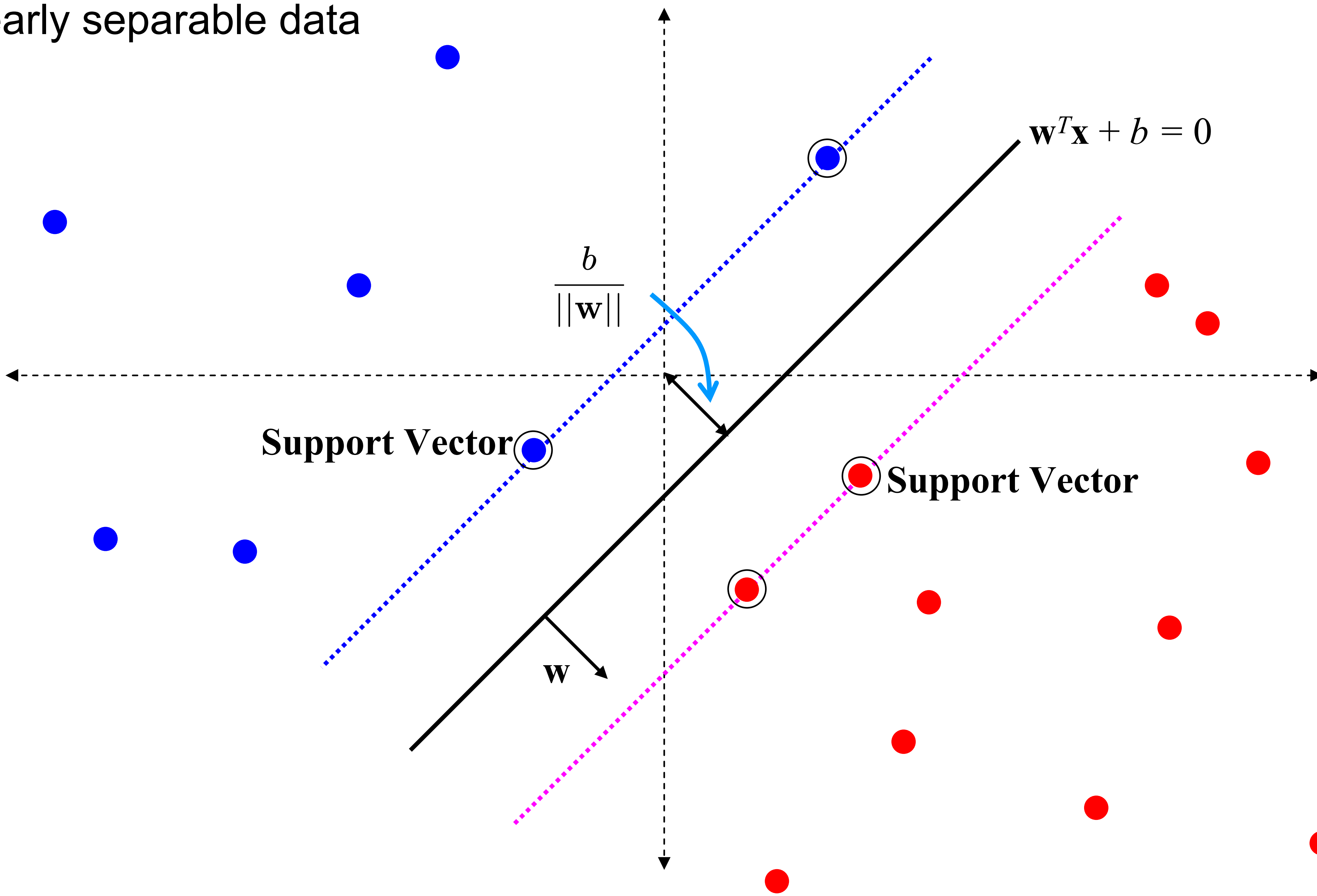
i.e. $y_i f(\mathbf{x}_i) > 0$ for a correct classification.



• maximum margin solution: most stable under perturbations of the inputs

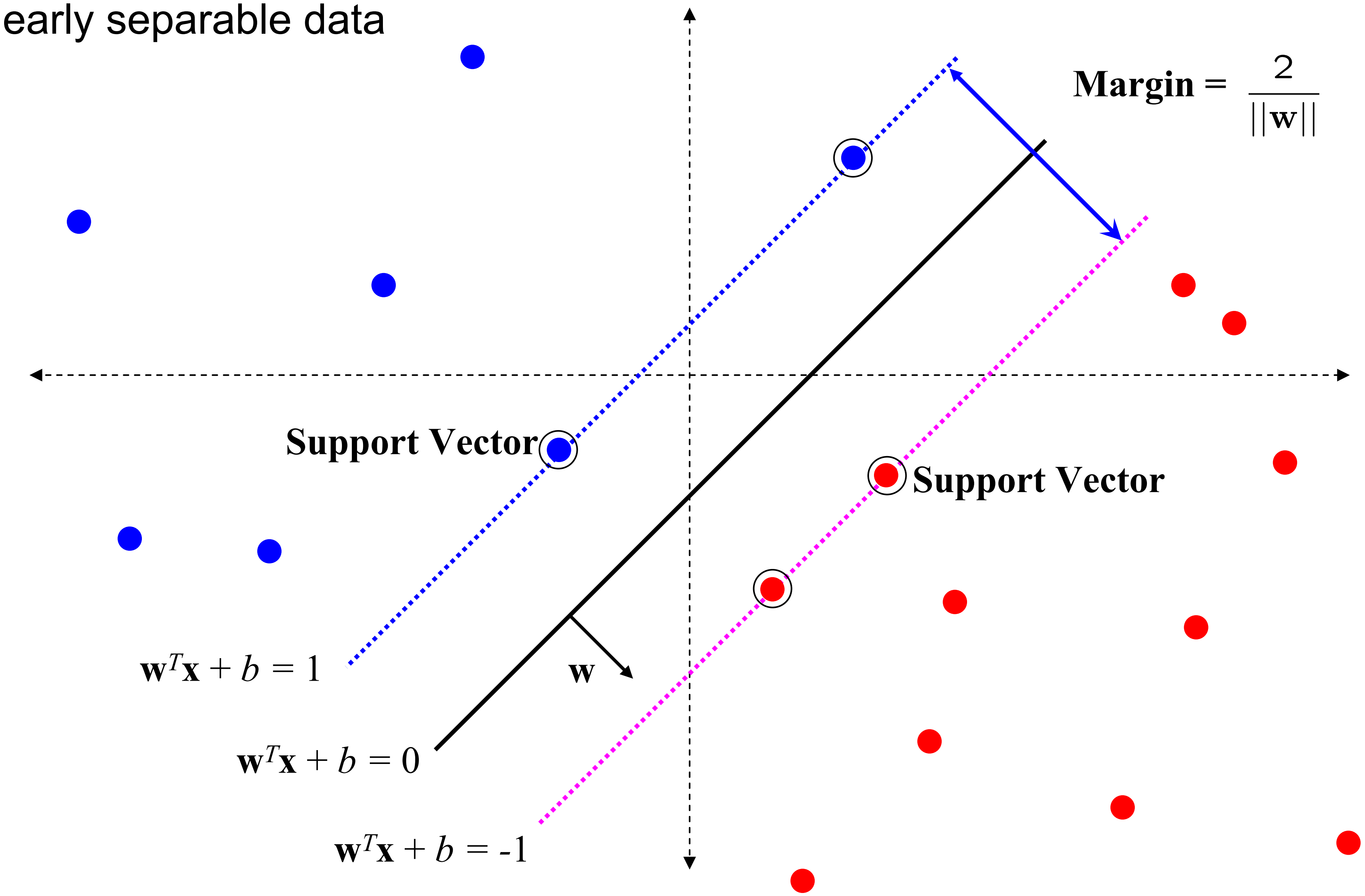
Support Vector Machine

linearly separable data



SVM Convex Optimization

linearly separable data



- Find w with a convex quadratic optimization:

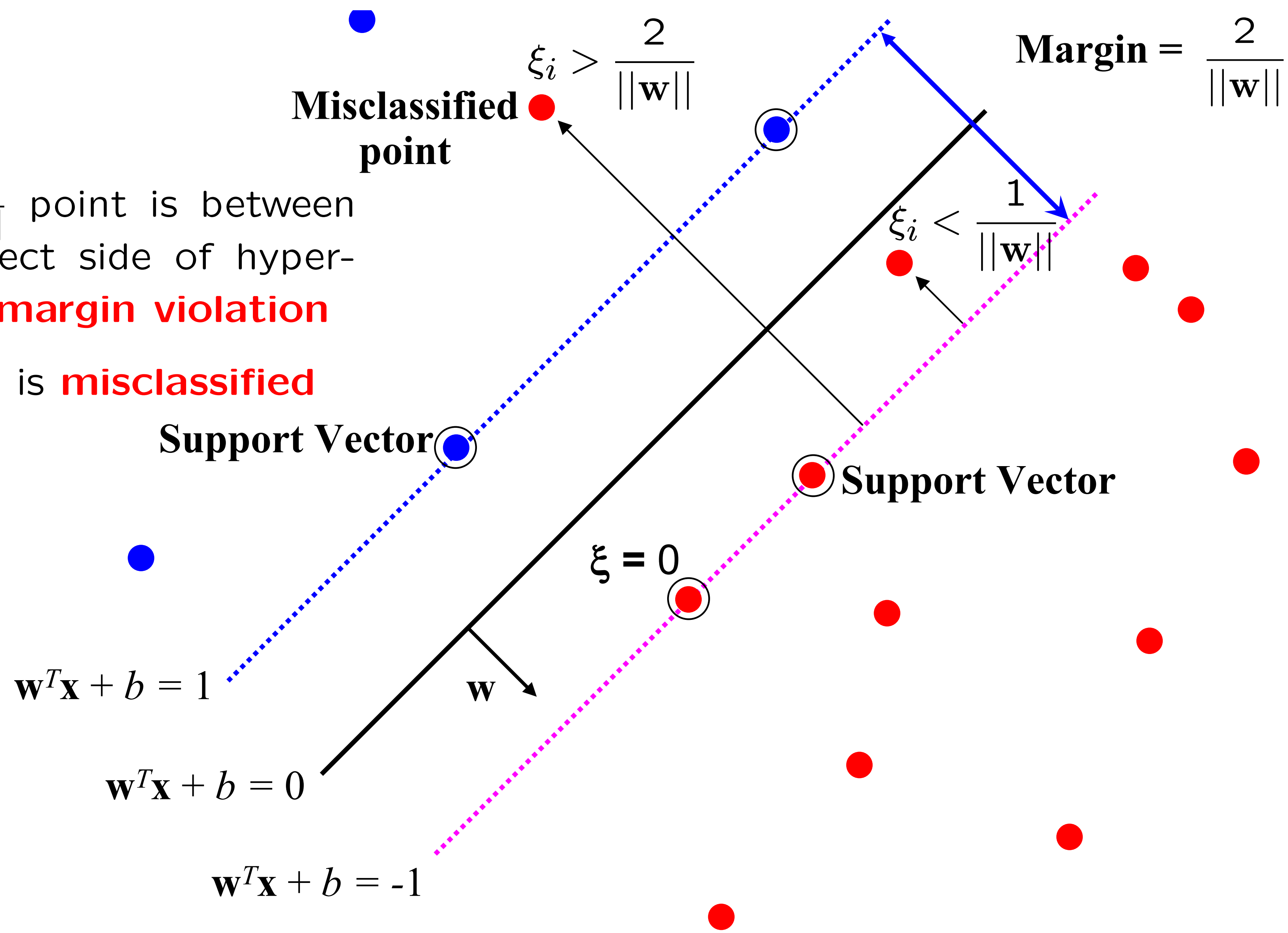
$$\min_w \|w\|^2 \text{ subject to } \forall i \; y_i(w^t x_i + b) \geq 1$$

$$\text{Solution} \; : \; f(x) = \sum_i \alpha_i \, y_i \, (x_i^t x) + b$$

Soft Margin Minimization

$\xi_i \geq 0$

- for $0 < \xi \leq \frac{1}{||\mathbf{w}||}$ point is between margin and correct side of hyper-plane. This is a **margin violation**
- for $\xi > \frac{1}{||\mathbf{w}||}$ point is **misclassified**



$$\min_{w,\xi_i} ||w||^2 + C \sum_i \xi_i$$

$$\text{subject to} \quad \forall i \; y_i(w^t x_i + b) \geq 1 - \xi_i$$

$$\text{Solution} \;\; : \;\; f(x) = \sum_i \alpha_i \, y_i \, (x_i^t x) + b$$

Kernel Support Vector Machine

$$f(x) = w^t x + b = \sum_i \alpha_i \, y_i \, (x_i^t x) + b.$$

- Replacing x by its representation $\Phi(x)$:

$$f(x) = w^t \Phi(x) + b = \sum_i \alpha_i \, y_i \Big(\Phi(x_i)^t \Phi(x) \Big) + b.$$

- Kernel trick: $K(x_i, x) = \Phi(x_i)^t \Phi(x)$ similarity measure

$$f(x) = \sum_i \alpha_i \, y_i \, K(x_i, x) + b.$$

Non-linear decision boundary.

- How to choose $\Phi(x)$ or equivalently $K(x, x')$ (Mercer thm.)

Mercer Theorem

A kernel is symmetric if $K(x,x') = K(x',x)$ and positive if

$$\forall c_i \in \mathbb{R} \;\; \forall x_i \in \mathbb{R}^d \quad , \quad \sum_i \sum_j K(x_i, x_j)\, c_i\, c_j \geq 0$$

Theorem If $K(x,x')$ is continuous, symmetric, positive then

there exists Φ from $\mathbb{R}^d$ to a Hilbert space $\mathcal{H}$ such that

$$K(x,x') = \Phi(x)^t \Phi(x') = \langle \Phi(x), \Phi(x') \rangle$$

Example: Gaussian kernel $K(x,'x) = \exp\left(\frac{-\|x-x'\|^2}{2\sigma^2}\right)$

Increase Dimensionality

Proposition: There exists a hyperplane separating any two subsets of N points $\{\Phi x_i\}_i$ in dimension $d' > N+1$ if $\{\Phi x_i\}_i$ are not in an affine subspace of dimension $< N$.

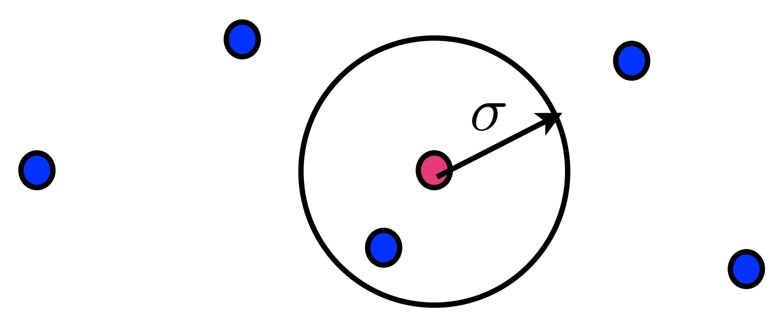
$\Rightarrow$ Choose Φ increasing dimensionality !

Problem: generalisation.

Example: Gaussian kernel $K(x', x) = \exp\left(\frac{-\|x-x'\|^2}{2\sigma^2}\right)$

$K(x', x) = \langle \Phi(x'), \Phi(x) \rangle$ where $\Phi x \in \mathcal{H}$ inifinite dimensional.

If σ is small, nearest neighbor classifier type:



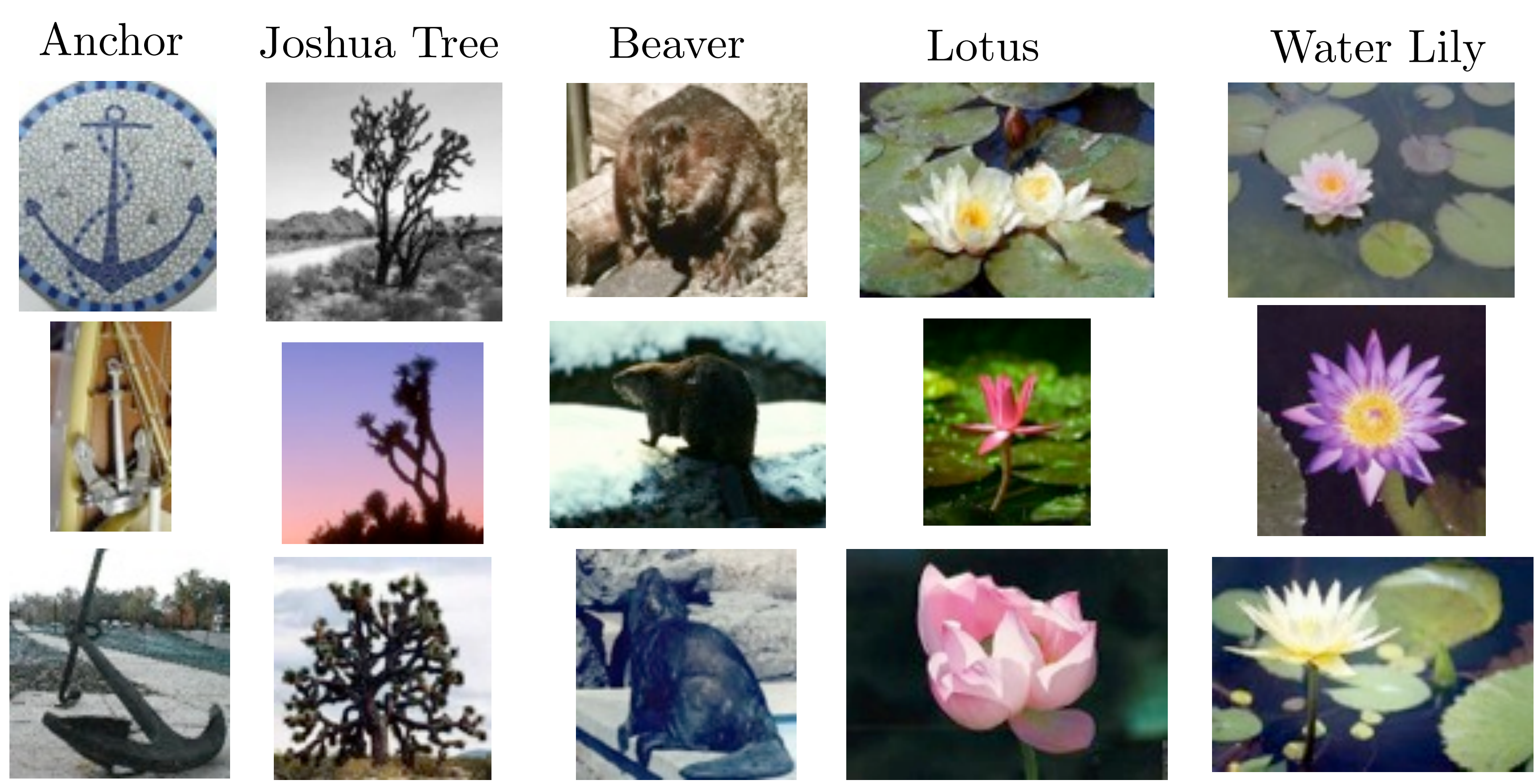
High-Dimensional Curse

- We want to learn $f(x)$ with $x \in \Omega$ and $\dim(\Omega) = d$ very big from n examples $\{x_i\, ,\, f(x_i)\}_{i \leq n}$

- Curse of dimensionality: if $n \ll 2^d$ then for "most" x:

$$\min_i \|x - x_i\| \text{ is large}$$

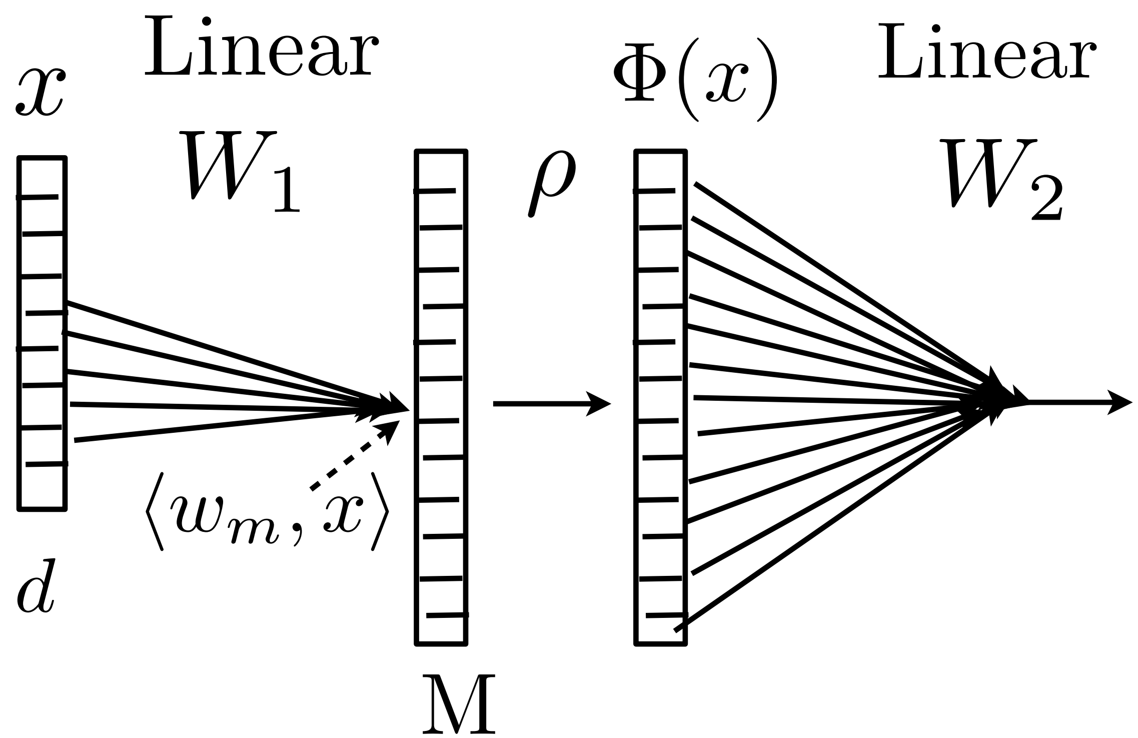
$\Rightarrow$ $f(x)$ can not be computed with a local interpolation



1 Layer Neural Network

- Ridge function approximations: $\Phi(x) = \Big\{\rho(\langle w, x\rangle\Big\}_w$

  $\rho(t) = e^{it}$, $\max(t, 0)$, $\mathrm{argtan}(t)$, $\psi(t)$, ...



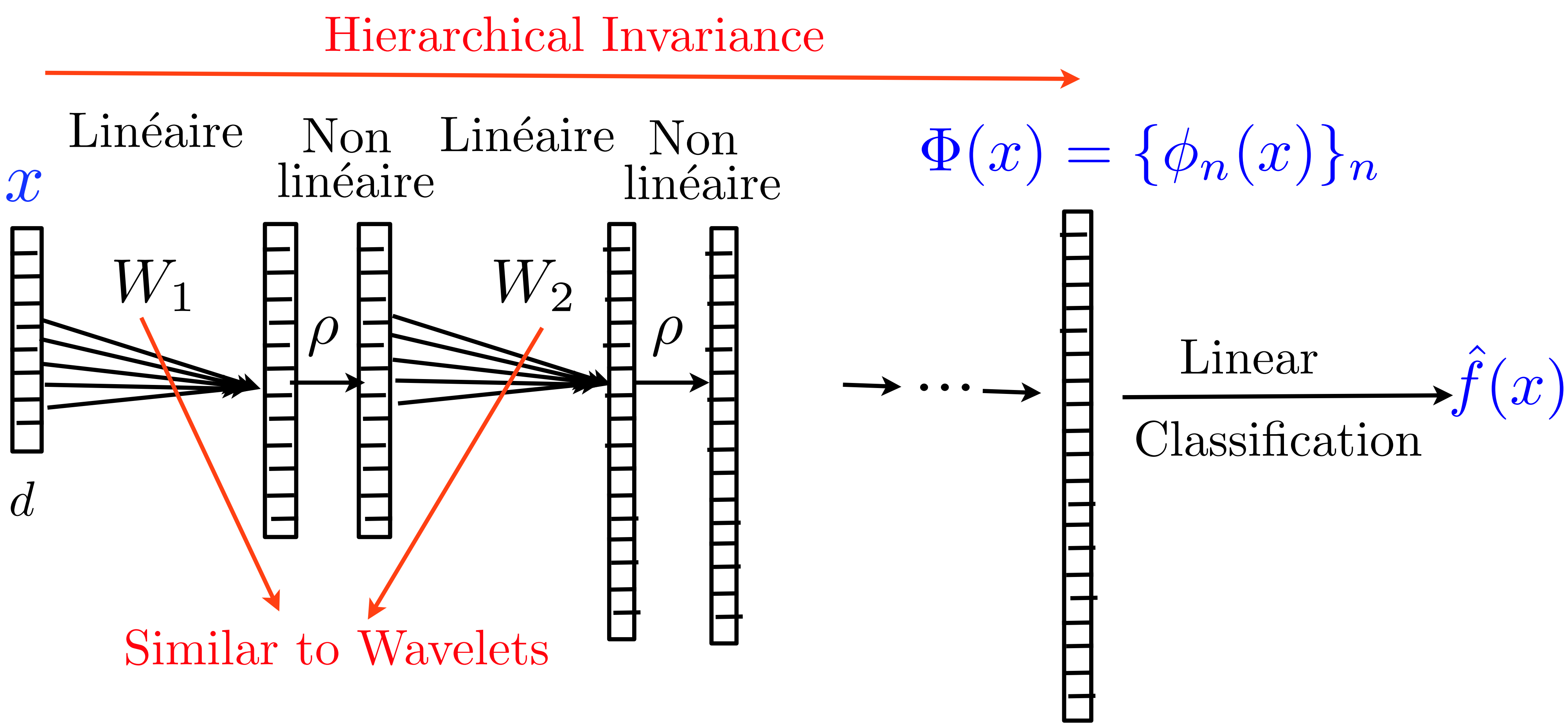
$$f_M(x) = \sum_{m=1}^{M} a_m\, \rho(\langle w_m, x\rangle)$$

Theorem: For "resonnable" bounded $\rho(u)$

If $f \in \mathbf{C}^\alpha[0,1]^d$ then $\|f - f_M\| \leq C\, M^{-\alpha/d}$

Deep Neural Neworks

- The revival of an old idea *(G. Hinton, Y. LeCun)*



Gradient descent learning of the W_k: more than 10^9 parameters

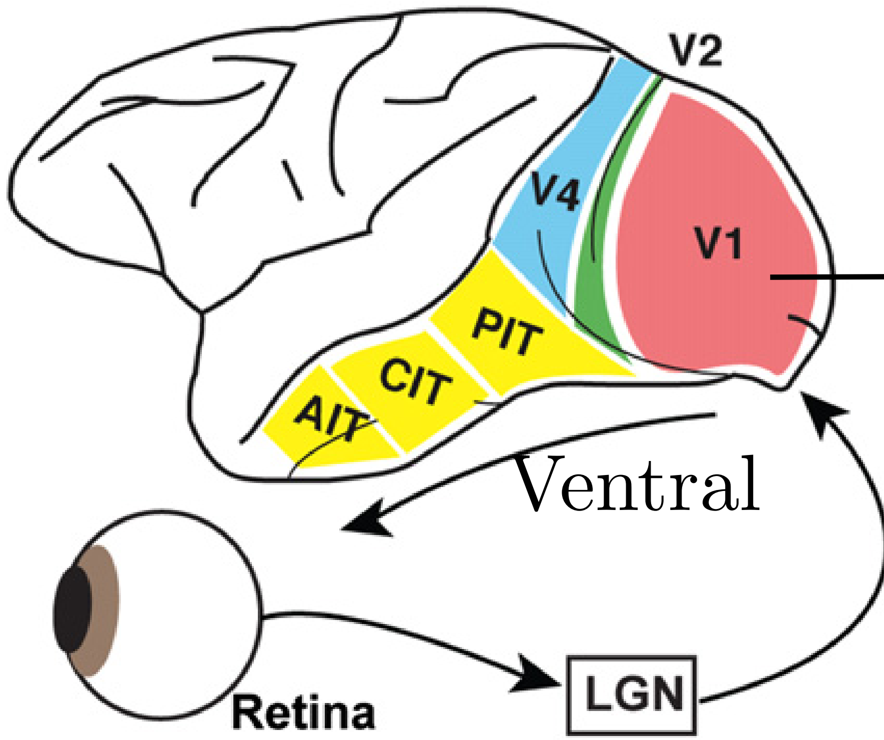
ImageNet (10^6 images and 10^3 classes): 17% error

Images, speech, bio-data: FaceBook, IBM, Google, Microsoft, Yahoo...

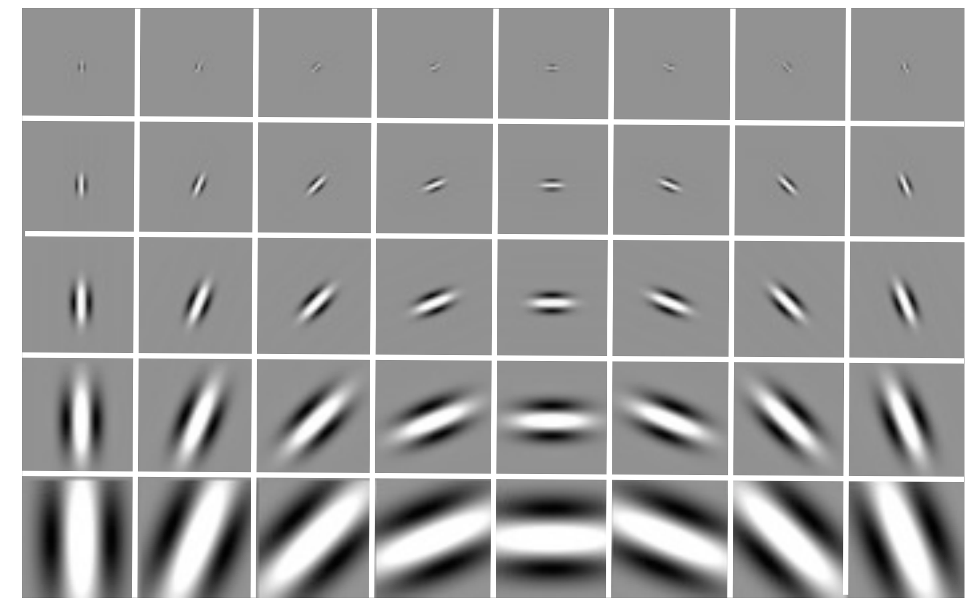
Why does it work ?

Neurophysiologie de la Perception

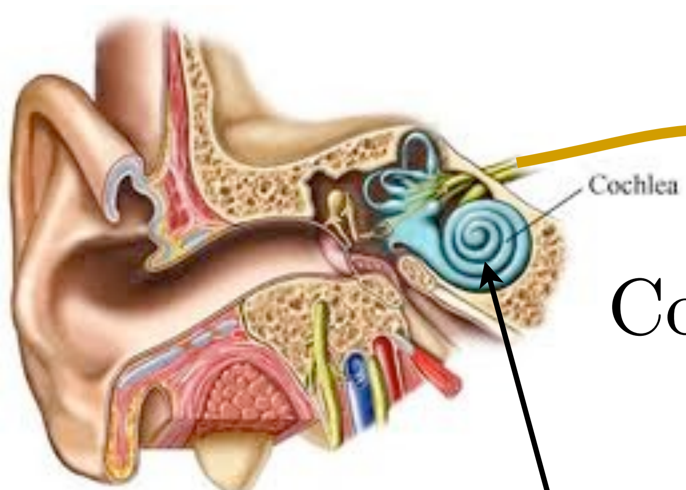
Vision



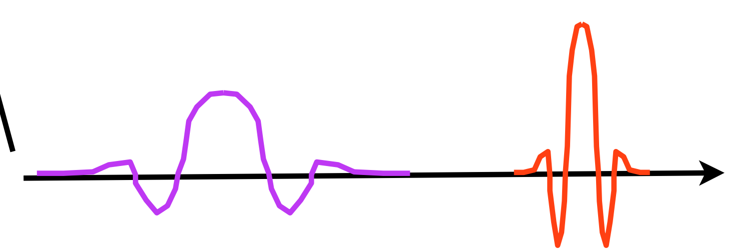
Hubel, Wiesel
Cellules simples modélisées par des ondelettes

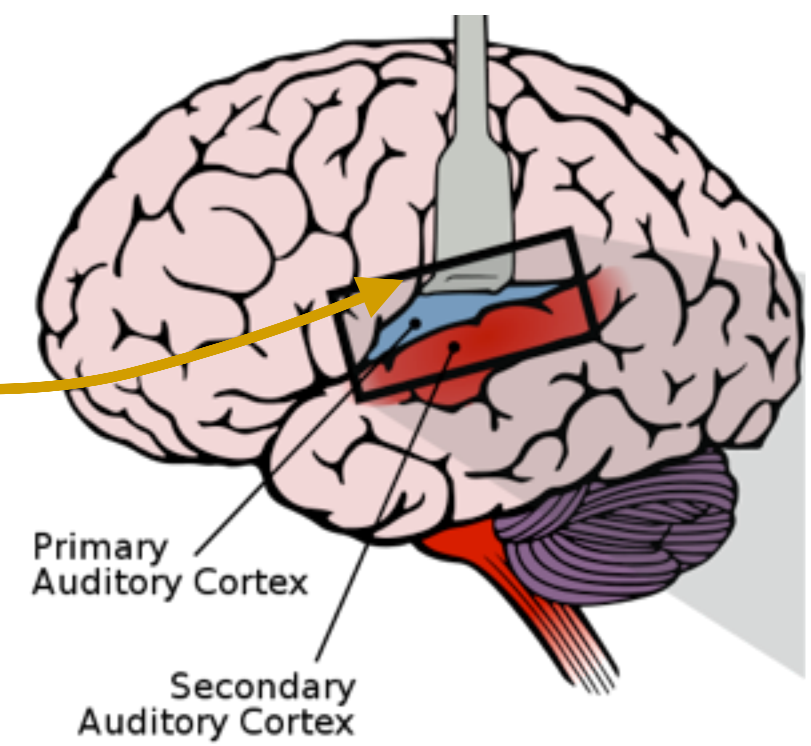


Audition



Cochlée: **ondelettes**





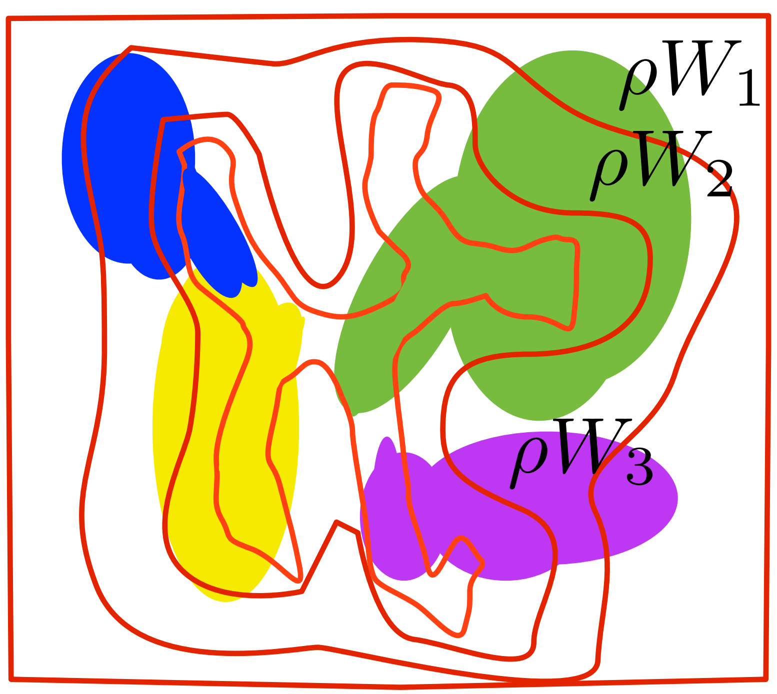
Iterated Contractions

- Reduce the space volume with iterated contractions

$$\Phi x = \rho\, W_m \ldots \rho\, W_2\, \rho\, W_1 x$$

  - W_k preserve distances: $\|W_k x - W_k x'\| = \|x - x'\|$
  - ρ is a contraction

- Iterative space contraction: reduce intra-class variability but avoid reducing class distances: margin condition.



- How to choose the W_k ?

Volume Reduction

- Contract Ω with an operator Φ such that:

  - $\forall x \in \Omega \quad , \quad \min_i \|\Phi(x) - \Phi(x_i)\|$ is small

  - $f(x)$ is regular relatively to $d(x, x') = \|\Phi(x) - \Phi(x')\|$

$$\forall x, x' \quad |f(x) - f(x')| \leq C \, \|\Phi(x) - \Phi(x')\|$$

Regression
margin condition: $\|\Phi(x) - \Phi(x')\| \geq C^{-1} |f(x) - f(x')|$

then $f(x)$ can be locally interpolated:

$$f(x) \approx \sum_{i=1}^{n} \alpha_i \, e^{\frac{\|\Phi(x) - \Phi(x_i)\|^2}{2\sigma^2}}$$

Volume Reduction

- Contract Ω with an operator Φ such that:

  - $\forall x \in \Omega \quad , \quad \min_i \|\Phi(x) - \Phi(x_i)\|$ is small

  - $f(x)$ is regular relatively to $d(x, x') = \|\Phi(x) - \Phi(x')\|$

$$\forall x, x' \quad |f(x) - f(x')| \leq C \, \|\Phi(x) - \Phi(x')\|$$

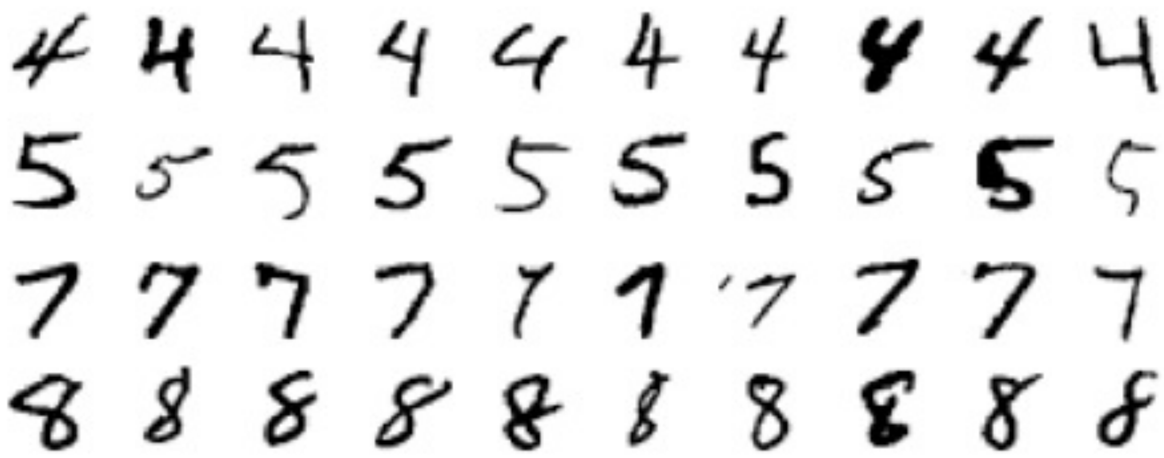
Classification margin condition: $\|\Phi(x) - \Phi(x')\| \geq C^{-1}$ if $f(x) \neq f(x')$

then $f(x)$ can be locally estimated:

$$f(x) \approx \mathrm{sign}\Big(\sum_{i=1}^{n} \alpha_i \, e^{\frac{\|\Phi(x) - \Phi(x_i)\|^2}{2\sigma^2}} \Big)$$

II- Translations and Deformations

- Patterns are translated and deformed



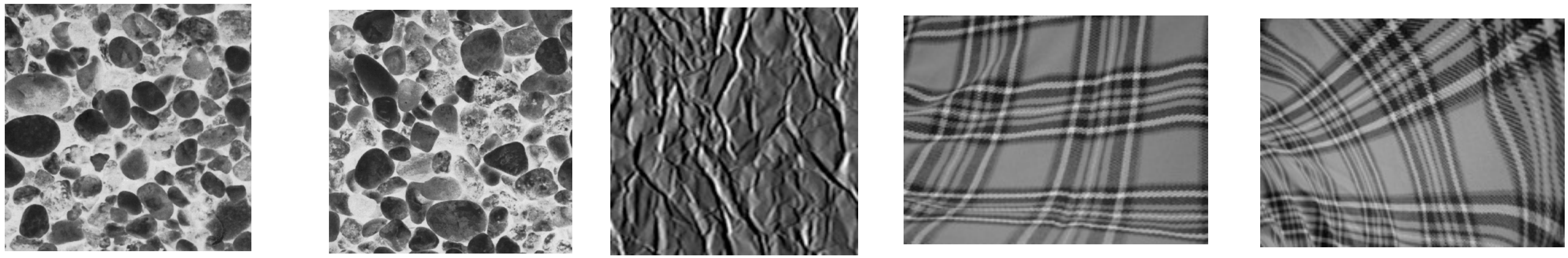
Invariance to Translations
Two dimensional group: $\mathbb{R}^2$

Deformations are actions of diffeomorphisms: infinite group.

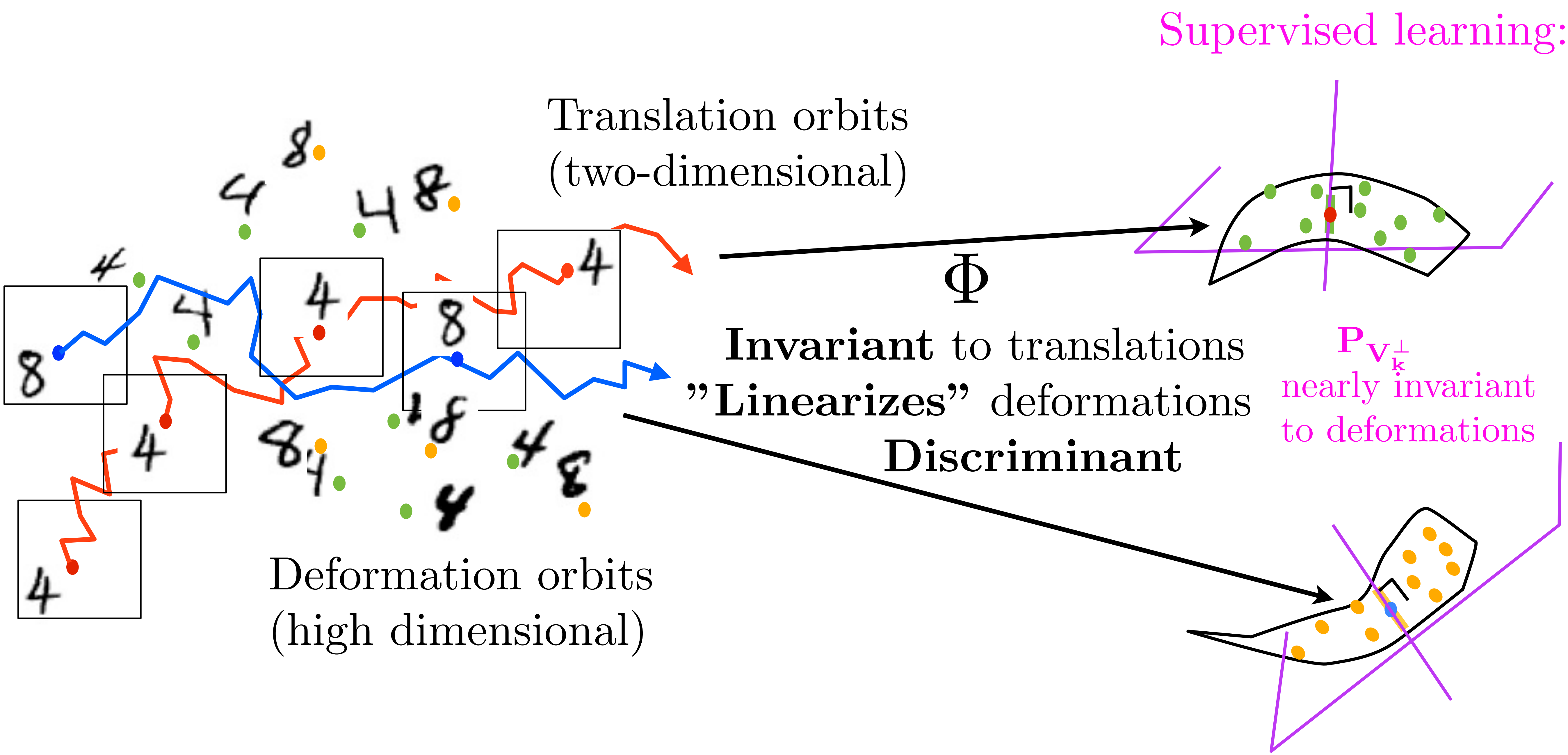
Each digit is invariant to a specific set of small deformations

- Textures are stationary (translation invariant) processes

with deformations



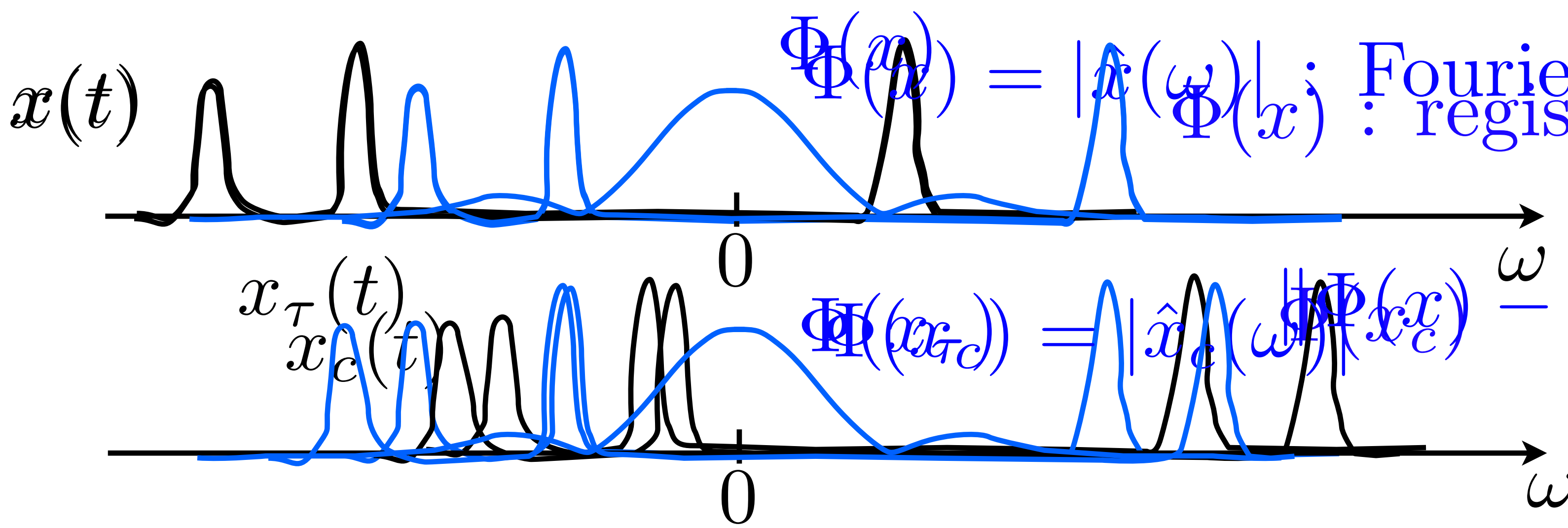
Translation and Deformations



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

$\|\Phi(x) - \Phi(x_\tau)\| \gg \sup_t |\tau'(t)| \, \|x\|$

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:

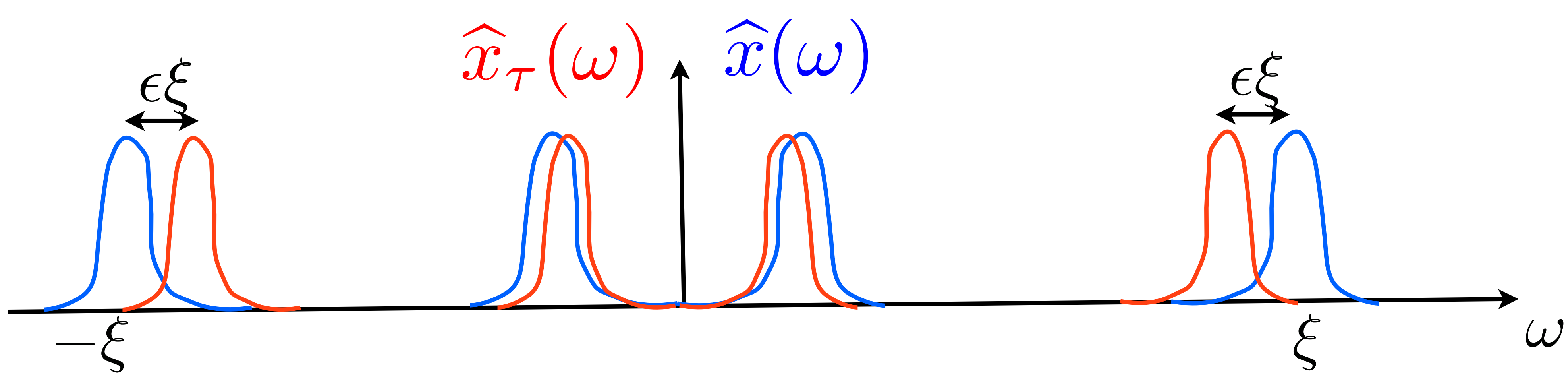
$$\text{if } x_c(t) = x(t-c) \text{ then } |\widehat{x_c}(\omega)| = |\hat{x}(\omega)|$$

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

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

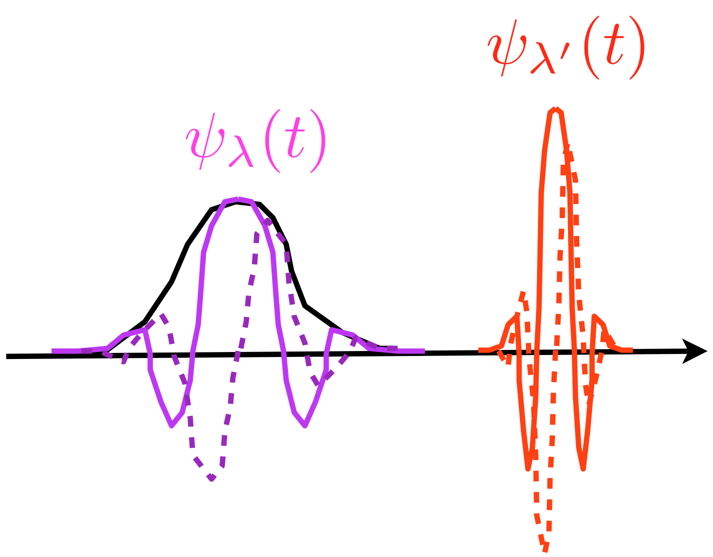
Example: If $\tau(t) = \epsilon t$ then $x_\tau(t) = x((1-\epsilon)t)$

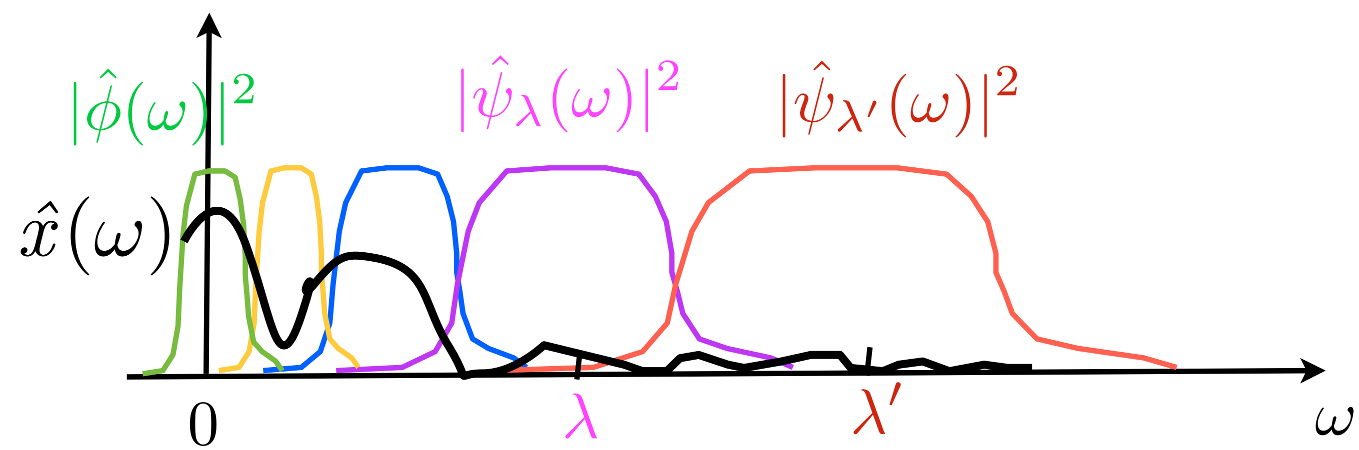
$$\Rightarrow \quad \widehat{x}_\tau(\omega) = (1-\epsilon)^{-1}\, \widehat{x}((1-\epsilon)^{-1}\omega)$$



Scale Separation with Wavelets

- Complex wavelet: $\psi(t) = \psi^a(t) + i\,\psi^b(t)$
- Dilated: $\psi_\lambda(t) = 2^{-j}\,\psi(2^{-j}t)$ with $\lambda = 2^{-j}$.





- Wavelet transform:

$$Wx = \begin{pmatrix} x \star \phi(t) \\ x \star \psi_\lambda(t) \end{pmatrix}_{t,\lambda} \begin{matrix} \longrightarrow \text{ averaging} \\ \longrightarrow \text{ high frequencies} \end{matrix}$$

Unitary: $\|Wx\|^2 = \|x\|^2$.

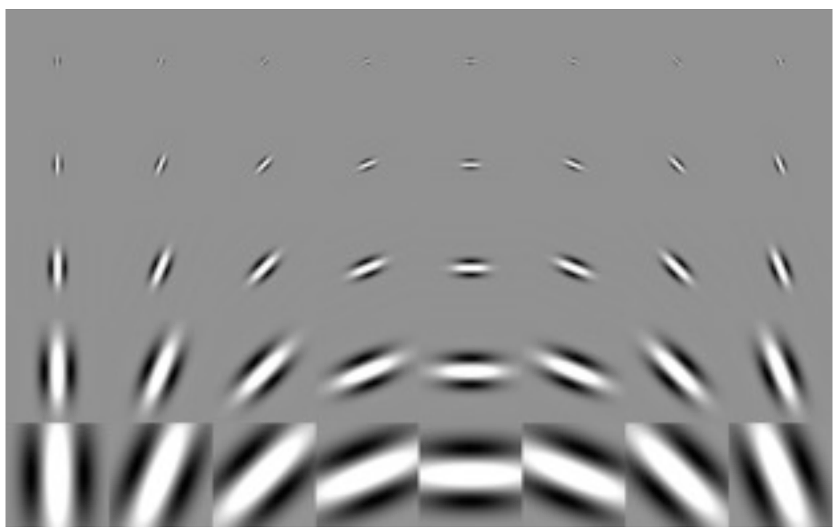
Monday, June 23, 14

Scale and Direction Separation in 2D

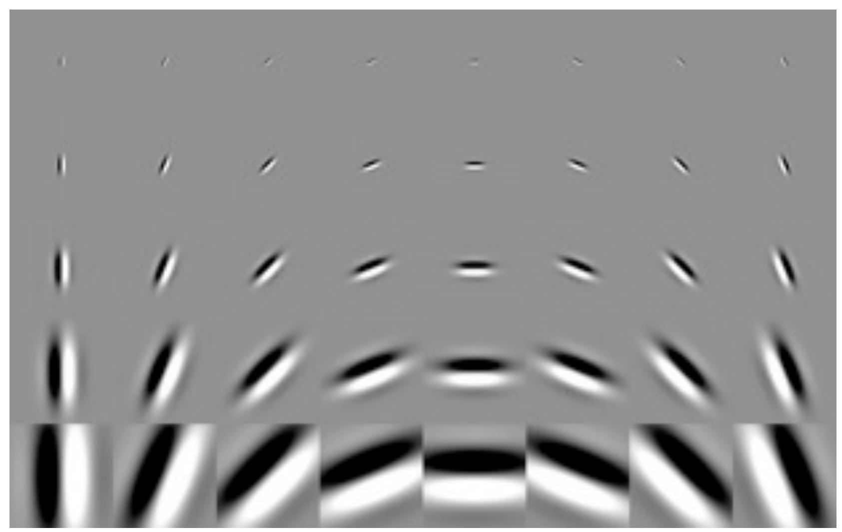
- Complex wavelet: $\psi(t) = \psi^a(t) + i\,\psi^b(t) \quad , \quad t = (t_1, t_2)$

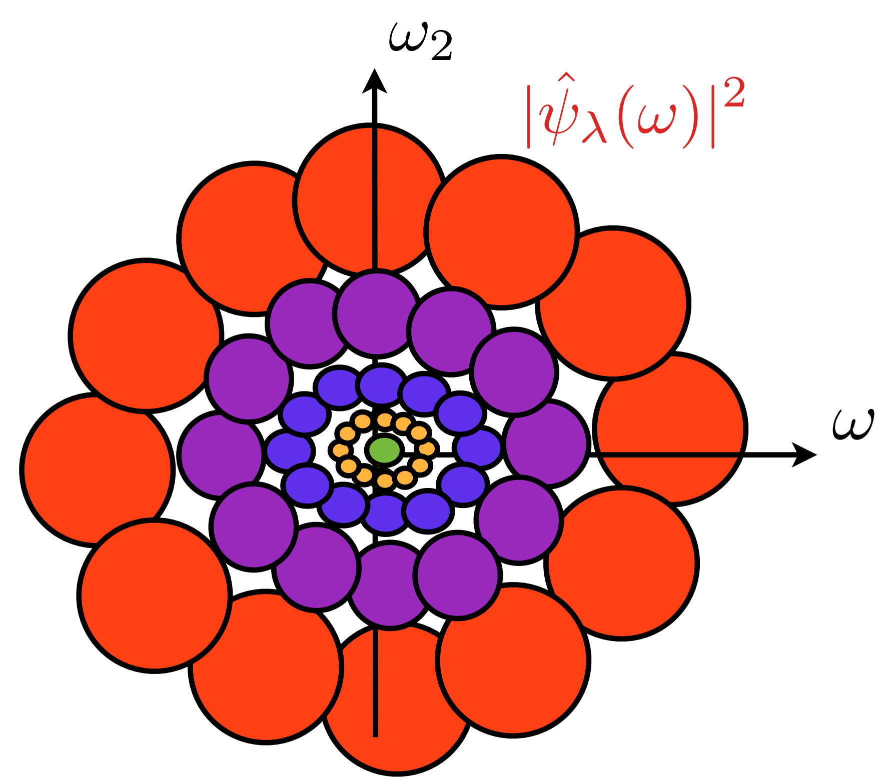
  rotated and dilated: $\psi_\lambda(t) = 2^{-j}\,\psi(2^{-j} r_\theta t)$ with $\lambda = (2^j, \theta)$

real parts



imaginary parts





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

  Unitary: $\|Wx\|^2 = \|x\|^2$.

Wavelet Tight Frames in L2

Functions in $\mathbf{L^2}(\mathbb{R}^d)$: $\|x\|^2 = \int |x(t)|^2\, dt < \infty$

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

Proposition: *(Littlewood-Paley)*

The wavelet transform is a tight frame for $x \in \mathbf{L^2}(\mathbb{R}^d)$

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

if and only if for almost all ω.

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

Why Wavelets ?

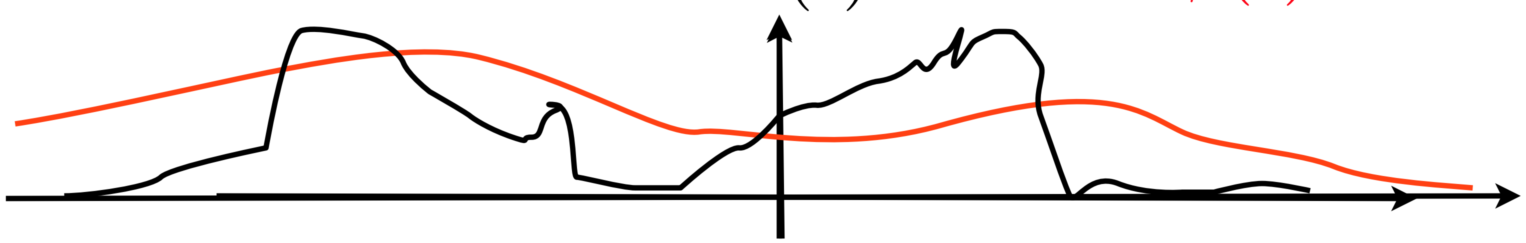
- The wavelet dictionary $\{\psi_\lambda(t-u)\}_{t,\lambda}$ is translation invariant.

- 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



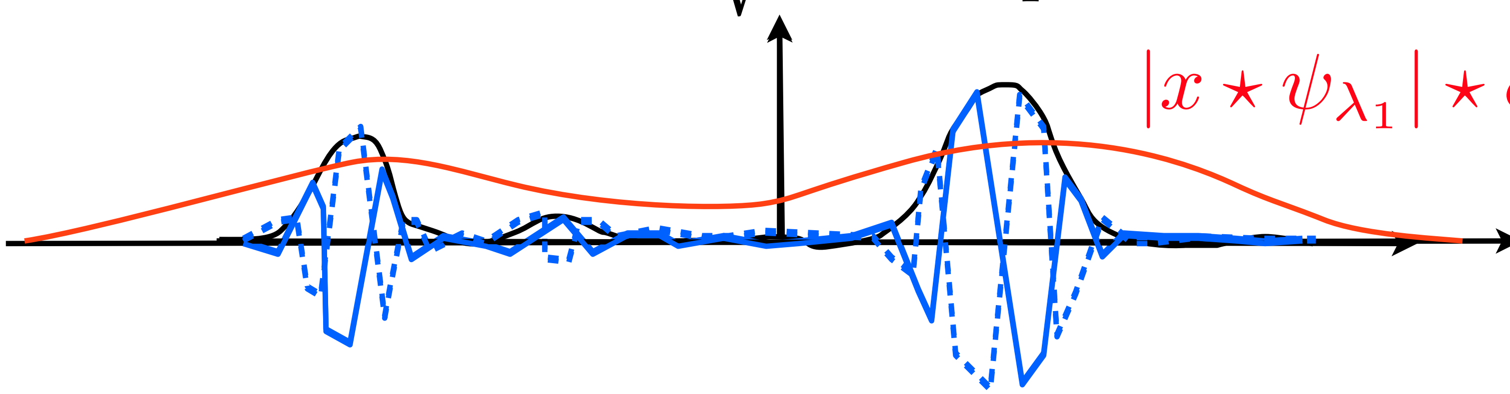
Wavelet transform modulus

W

$|W|$

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

$$|x \star \psi_{\lambda_1}(t)| = \sqrt{|x \star \psi^a_{\lambda_1}(t)|^2 + |x \star \psi^b_{\lambda_1}(t)|^2}$$



Unitary: $\|Wx - Wx'\| = \|x - x'\|$

Contraction: $||a| - |b|| \leq |a - b|$

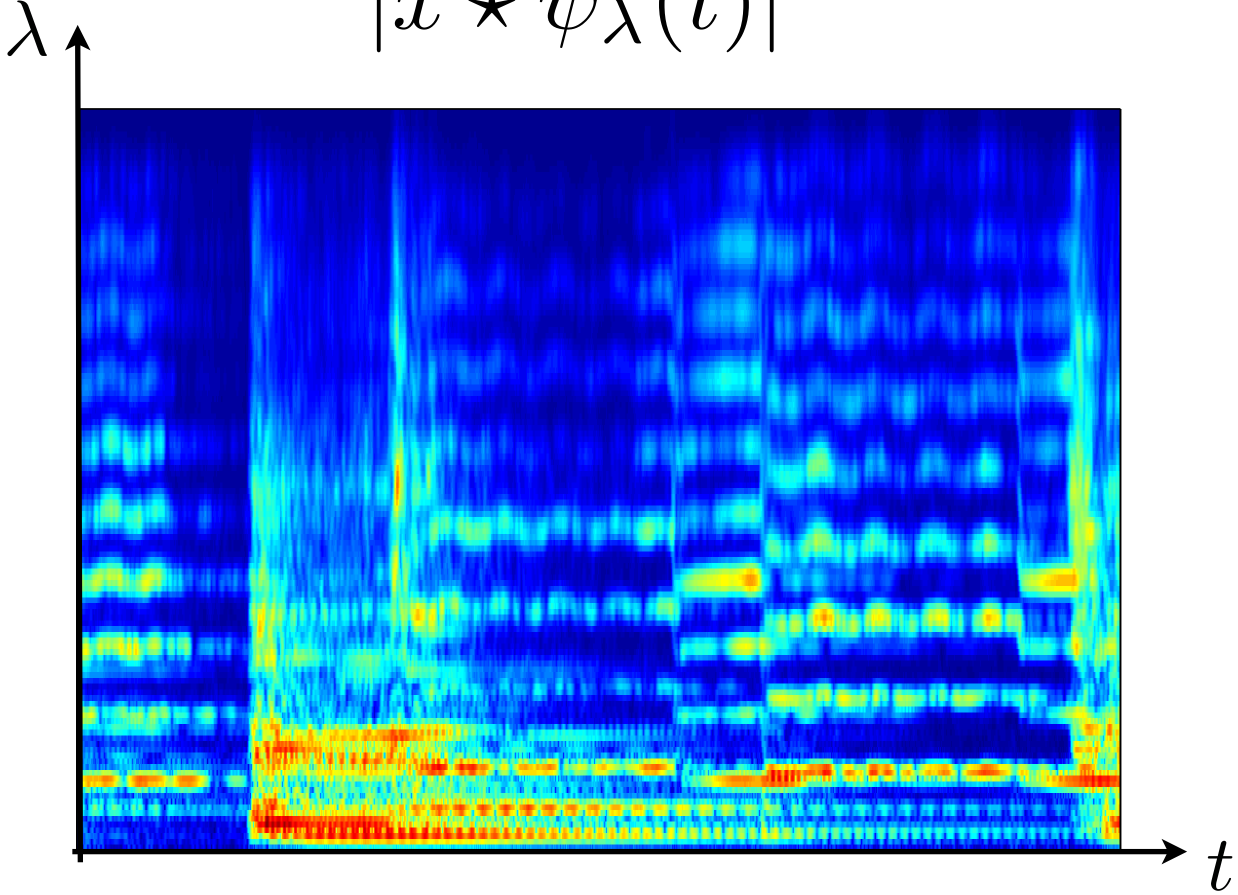
$$\Rightarrow \quad \||W|x - |Wx'|\| \leq \|x - x'\|$$

Preserves the norm: $\||W|x\| = \|x\|$

Wavelet Stabilization

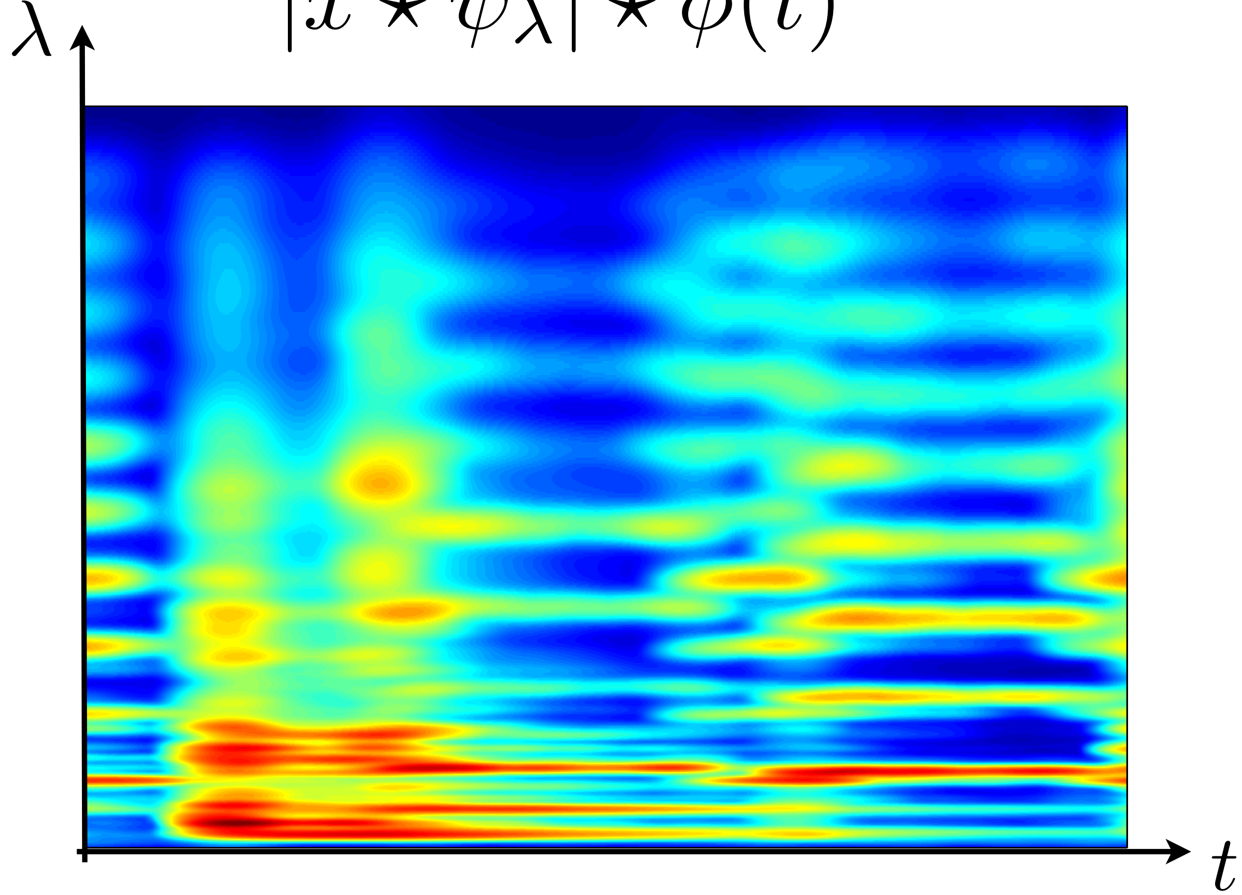
Wavelet time-frequency

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



Time averaging on **370ms**

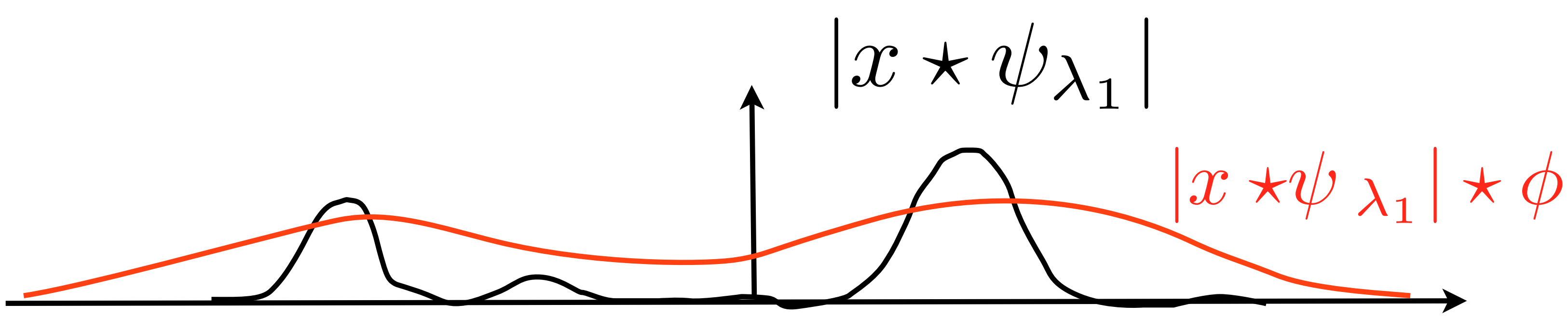
$|x \star \psi_\lambda| \star \phi(t)$



Locally invariant to translations and stable to deformations

but loss of information.

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:

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

Translation Invariance

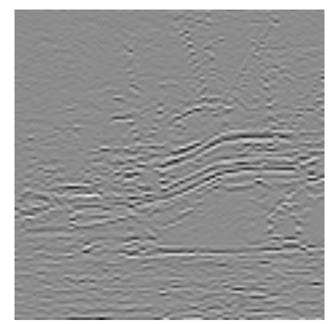
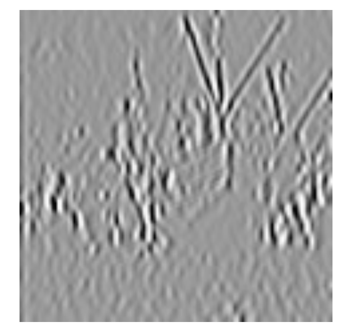
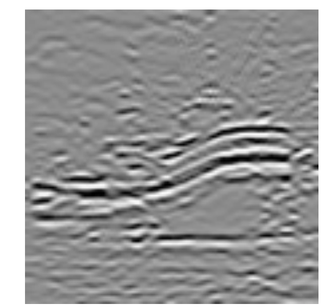
$x(t)$

Local invariant
by translation

$x \star \phi$

Wavelet transform
W_1

$x \star \psi_{\lambda_1}(t)$ $\quad$ $x \star \psi_{\lambda'_1}(t)$ $\quad$ $x \star \psi_{\lambda''_1}(t)$ $\quad$ $x \star \psi_{\lambda'''_1}(t)$

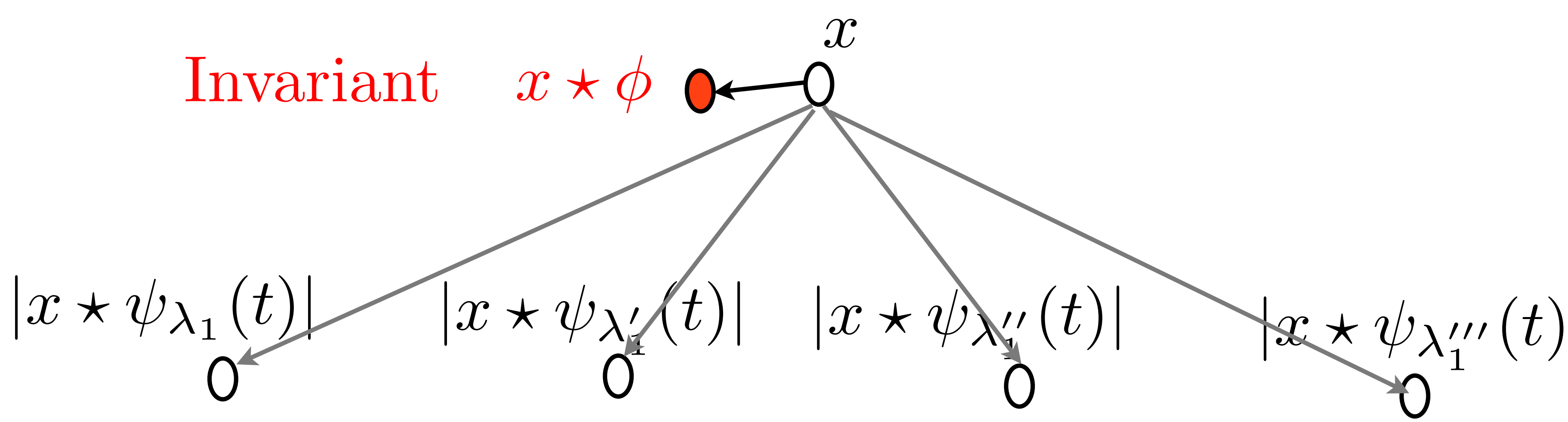





scale and orientation
separation

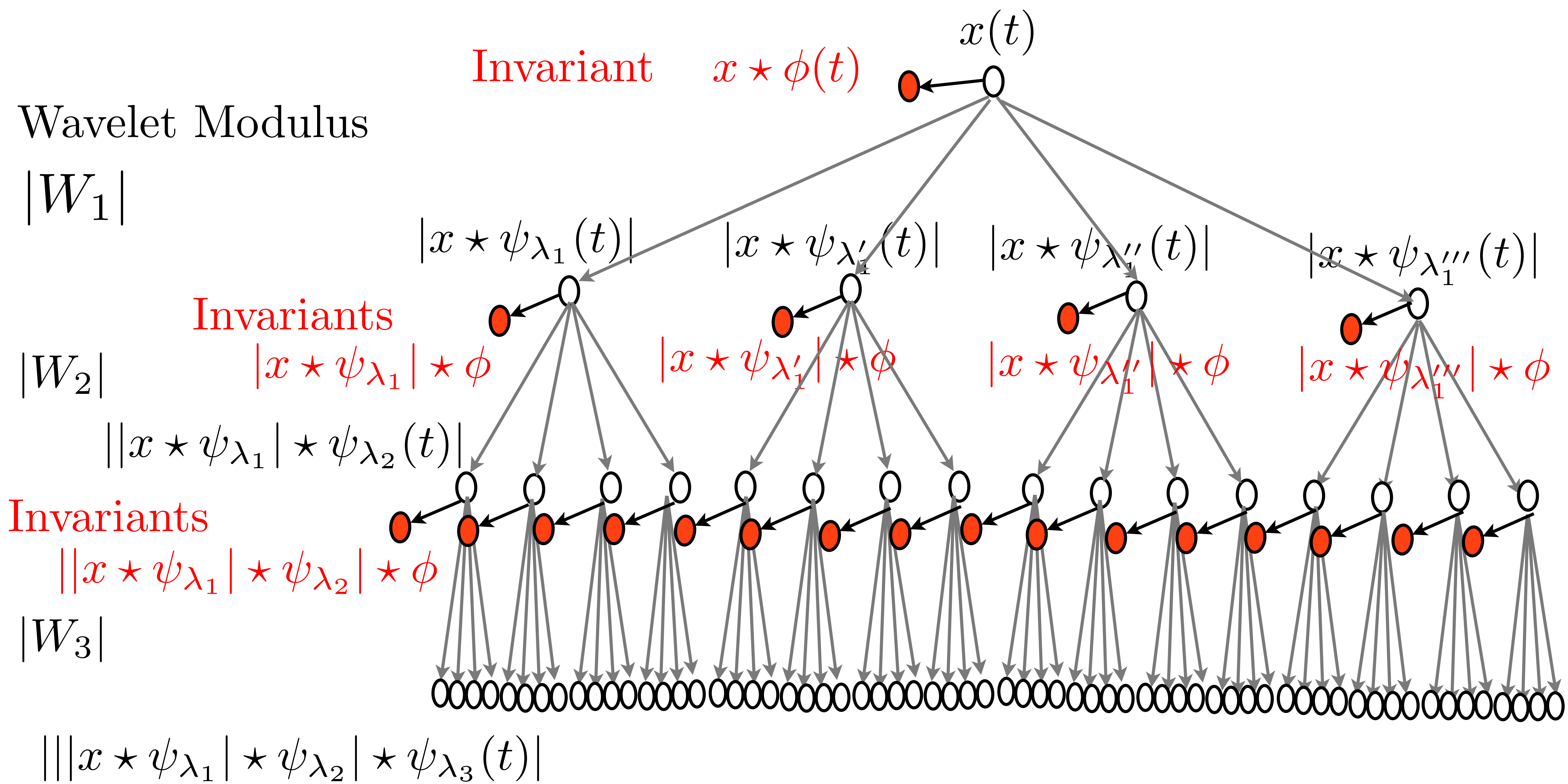
Scattering Neuronal Network

Wavelet Modulus

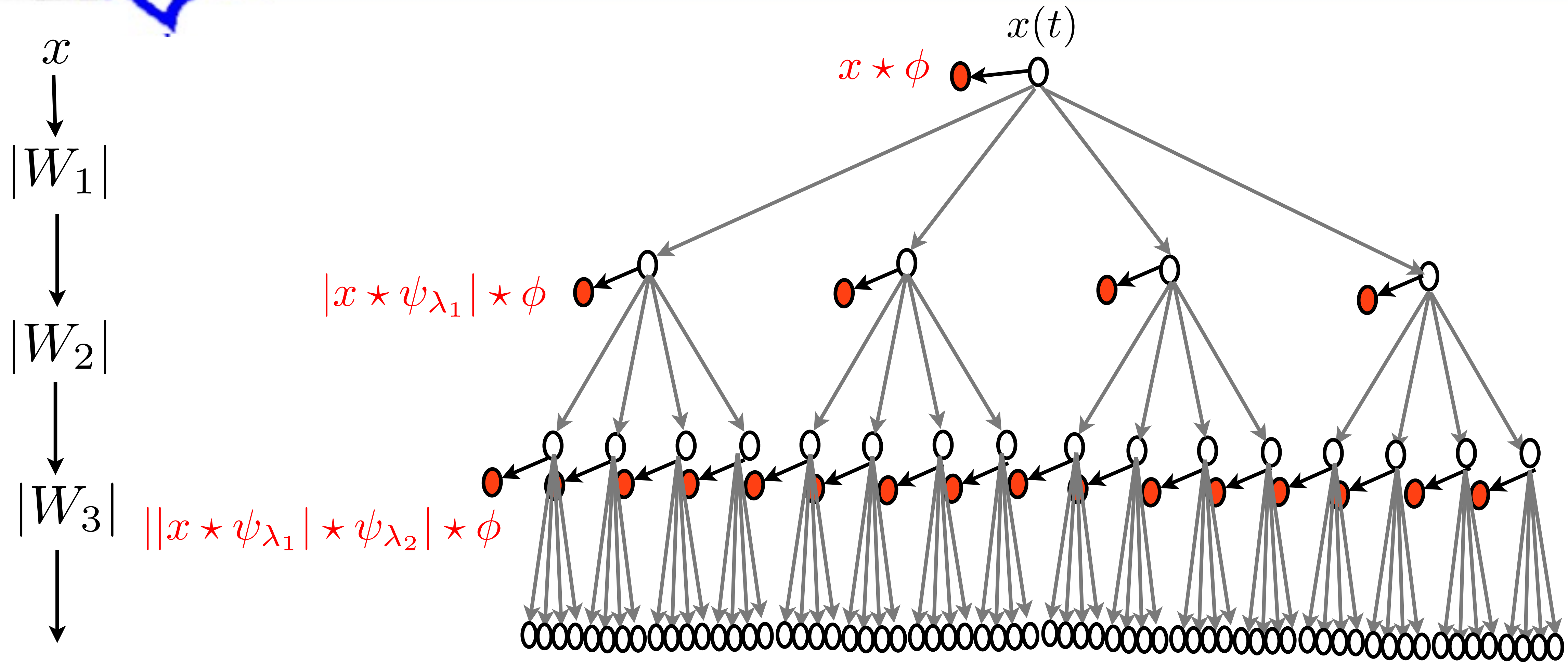
$|W_1|$



Scattering Neuronal Network



Wavelet Scattering



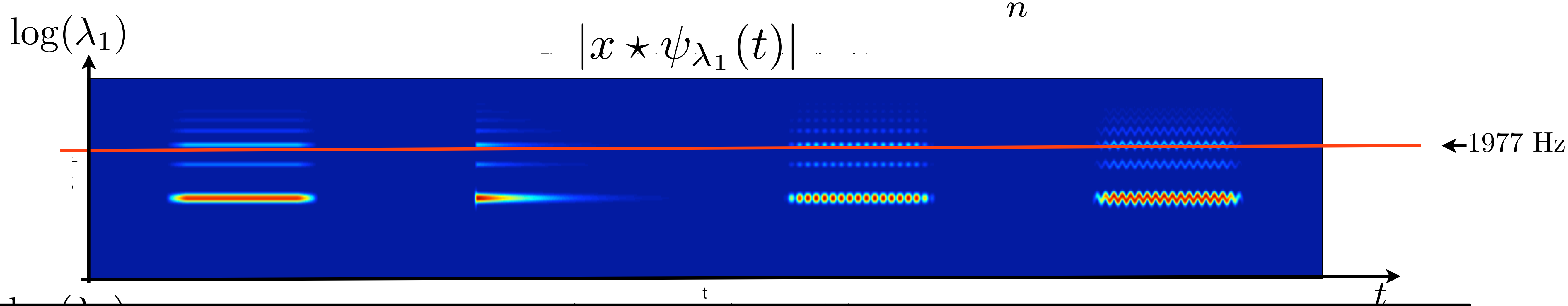
$$\text{Scattering:} \quad Sx = \left(\begin{array}{ccc} S_0(x) & = & x \star \phi \\ S_{\lambda_1}(x) & = & |x \star \psi_{\lambda_1}| \star \phi \\ S_{\lambda_1,\lambda_2}(x) & = & ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi \\ S_{\lambda_1,\lambda_2,\lambda_3}(x) & = & |||x \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi \\ \ldots & & \end{array} \right)_{\lambda_1,\lambda_2,\lambda_3}$$

Theorem:

$$\|Sx - Sx'\| \leq \|x - x'\| \quad \text{et} \quad \|Sx\| = \|x\|$$

Amplitude Modulation

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



Scattering Properties

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

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

Theorem: *For appropriate wavelets, a scattering is*

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

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

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

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

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.

Image Scattering Transforms

Image

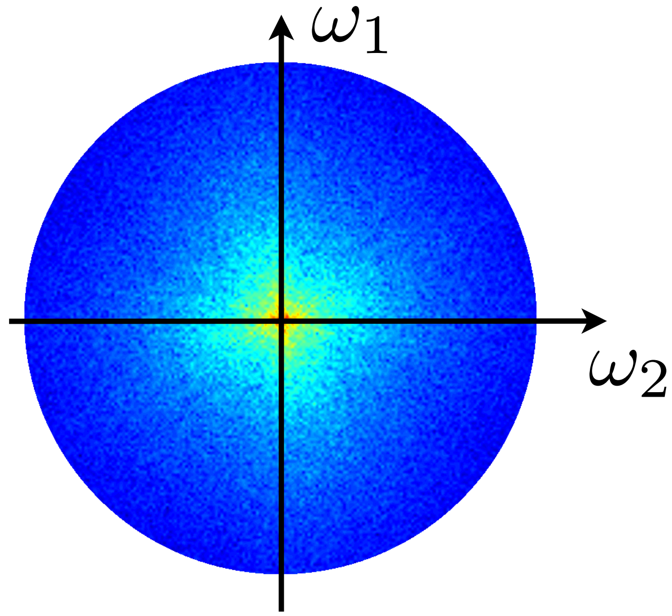
$x(t)$

$t = (t_1, t_2)$

Fourier Modulus

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

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



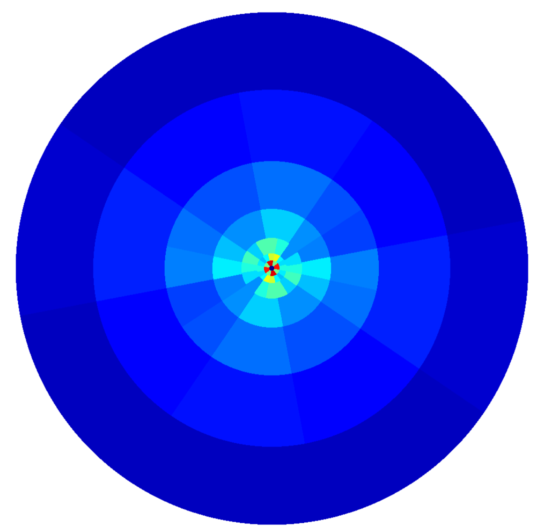
Scattering $\phi(t) = 1$

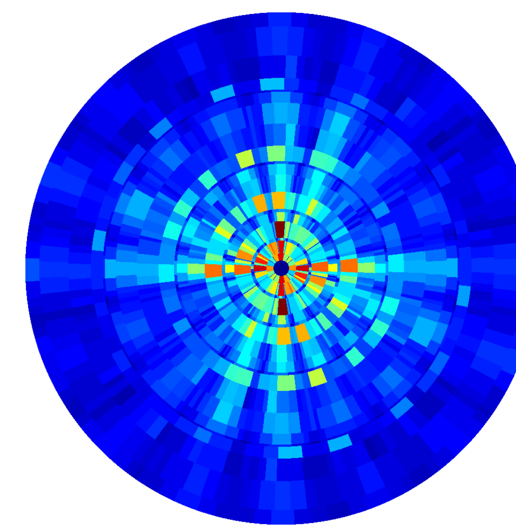
$|x \star \psi_{\lambda_1}| \star \phi$ $\qquad$ $||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi$

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

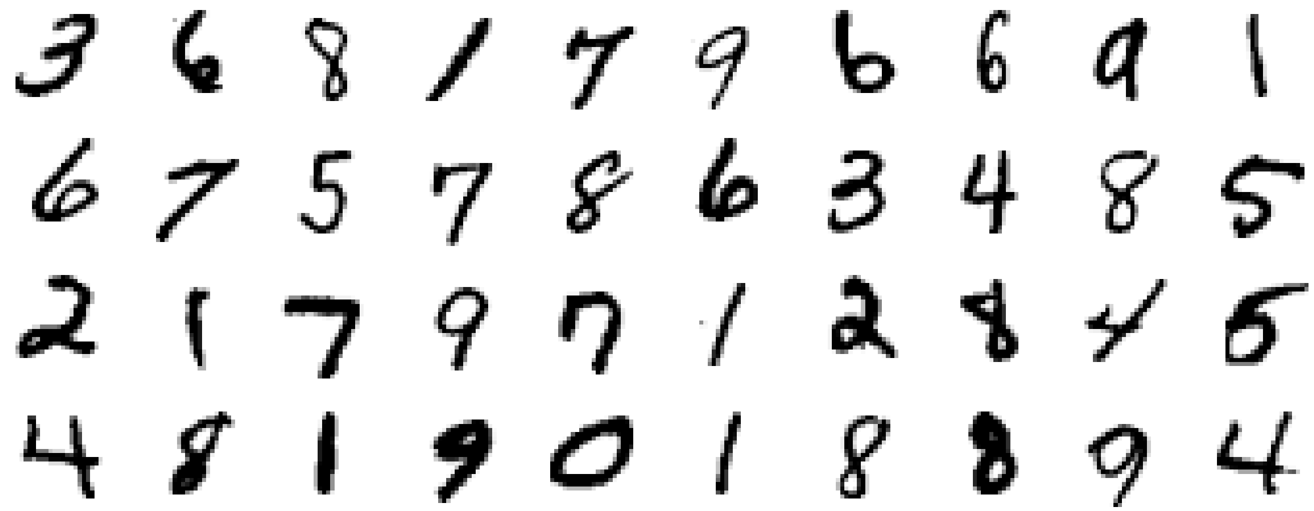
$\lambda_1 = 2^{j_1}\, r_{\theta_1}$ $\qquad$ $\lambda_1 = 2^{j_1}\, r_{\theta_1}$

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





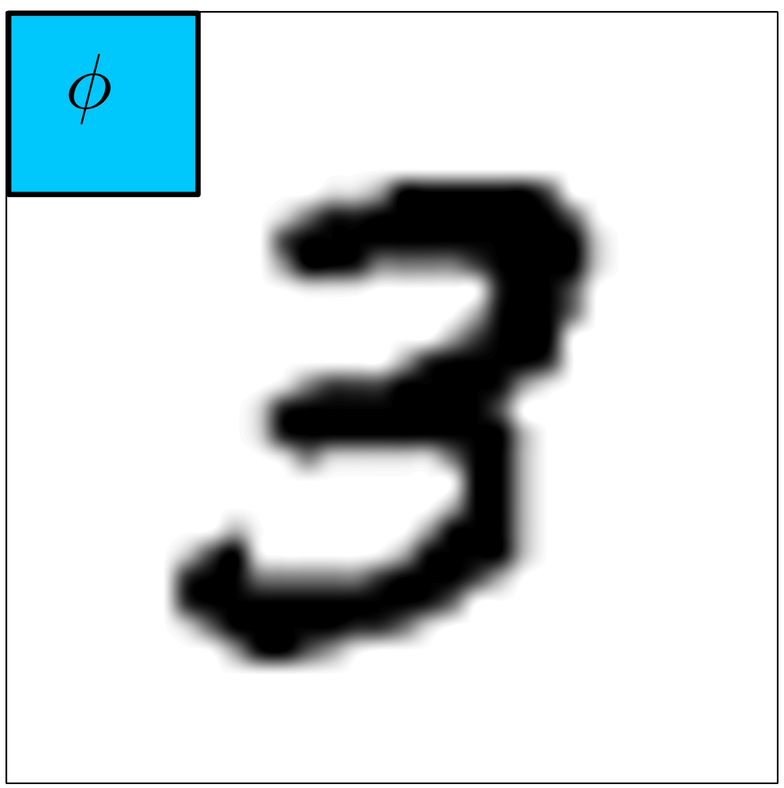
Digit Classification: MNIST



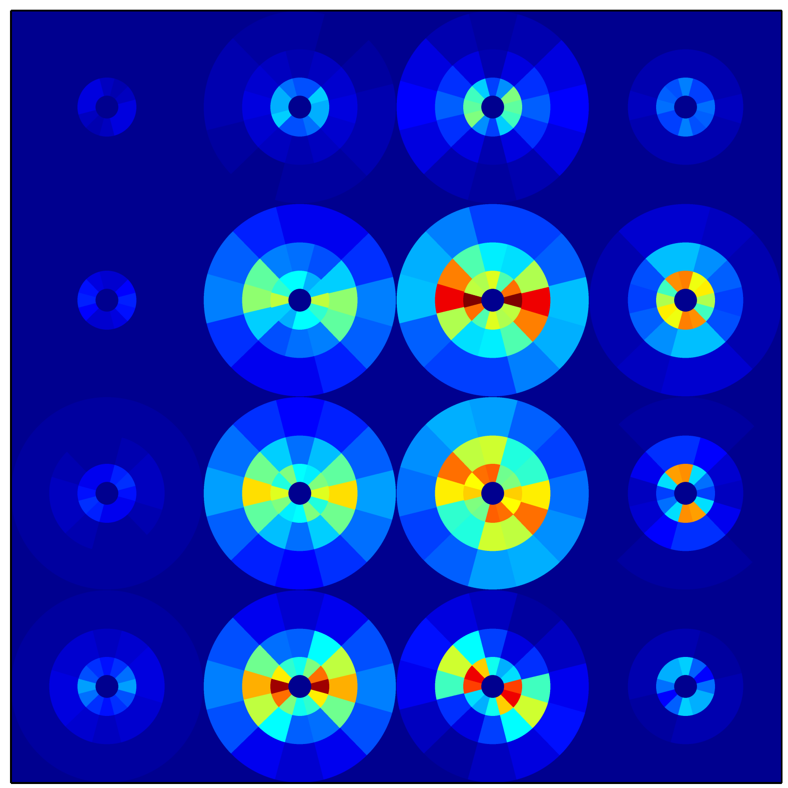
Digit Classification: MNIST

Second order Scattering Sx:

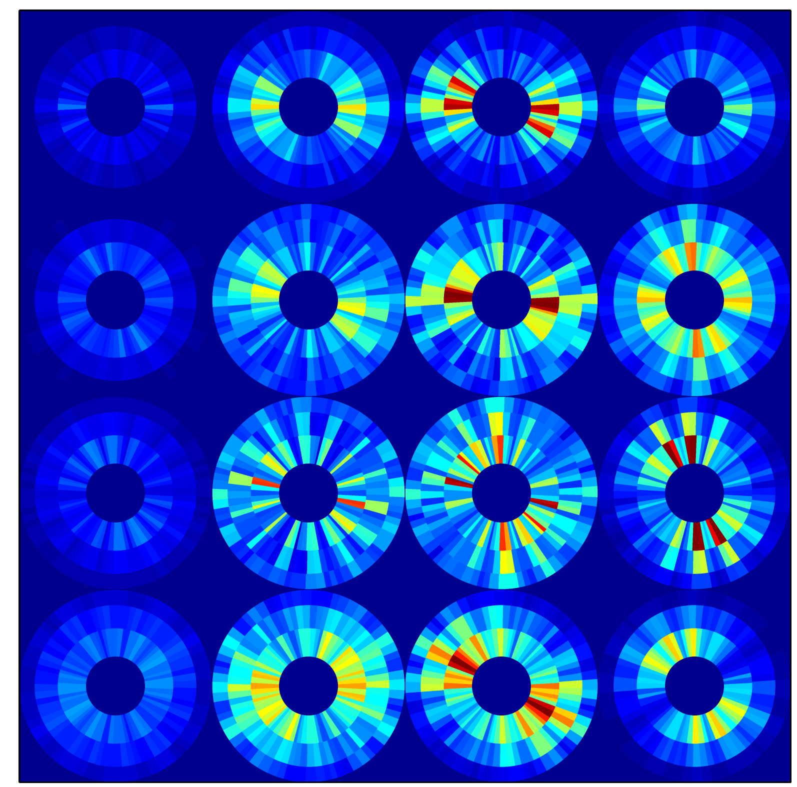
2^J $x(t)$



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



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

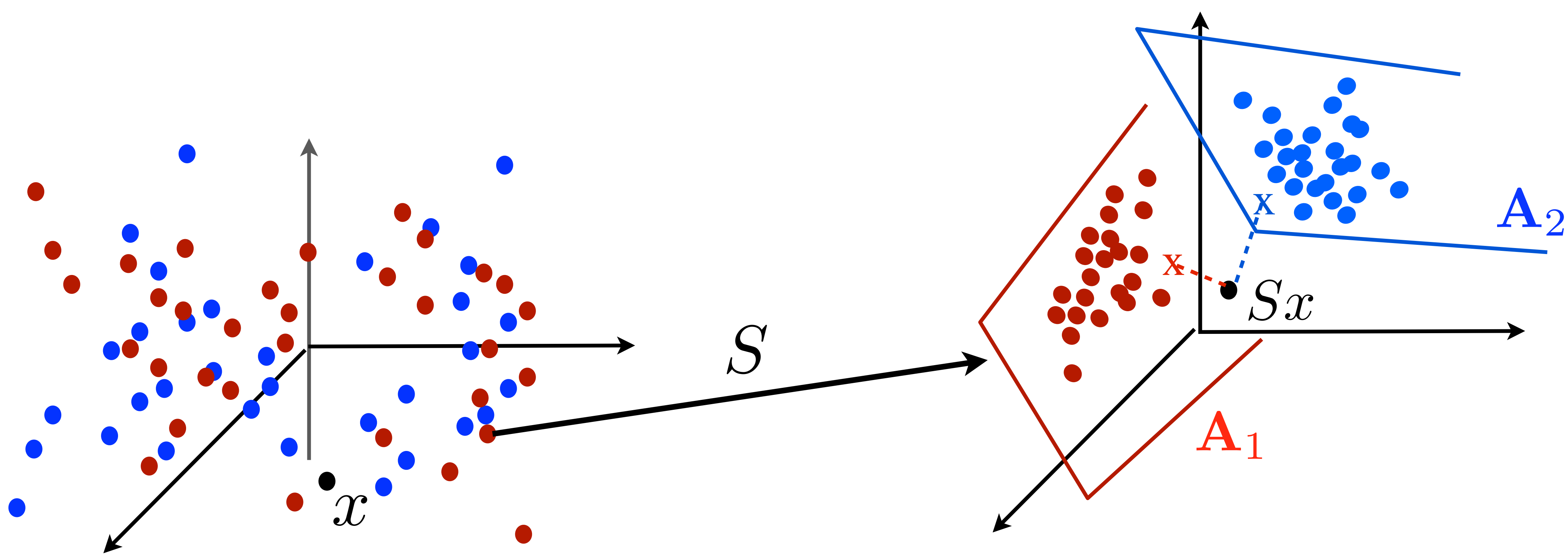


Affine Space Classification

Joan Bruna

- Each class is represented by a random process X_k

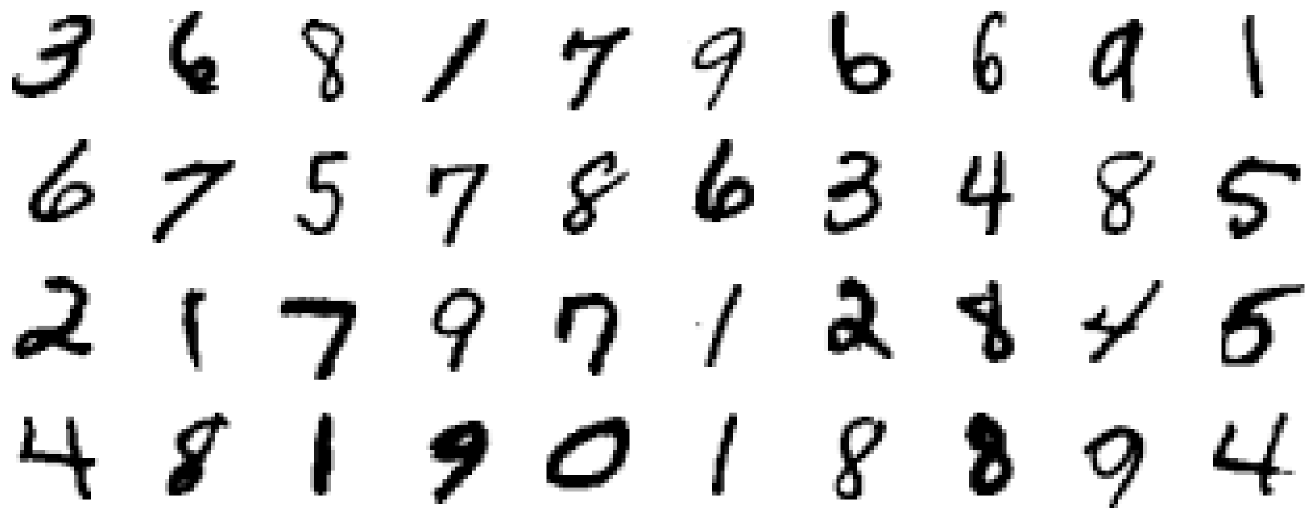
  The support of SX_k is approximated by a low-dimensional affine space $\mathbf{A}_k$ computed with a PCA.



$$\hat{k}(x) = \arg\max_{k} \|Sx - P_{\mathbf{A_k}} Sx\| \ .$$

Digit Classification: MNIST

Joan Bruna



$x \longrightarrow \boxed{Sx} \longrightarrow$ Supervised Linear Classifier: PCA/SVM $\longrightarrow y$

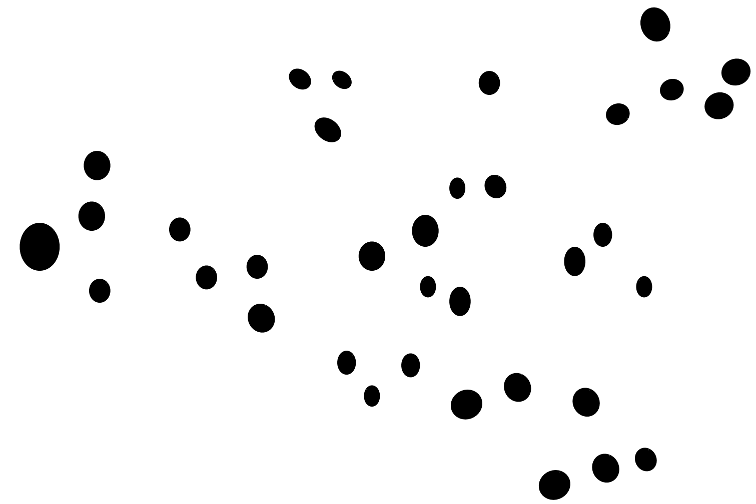
Classification Errors

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.

Many Body Interactions

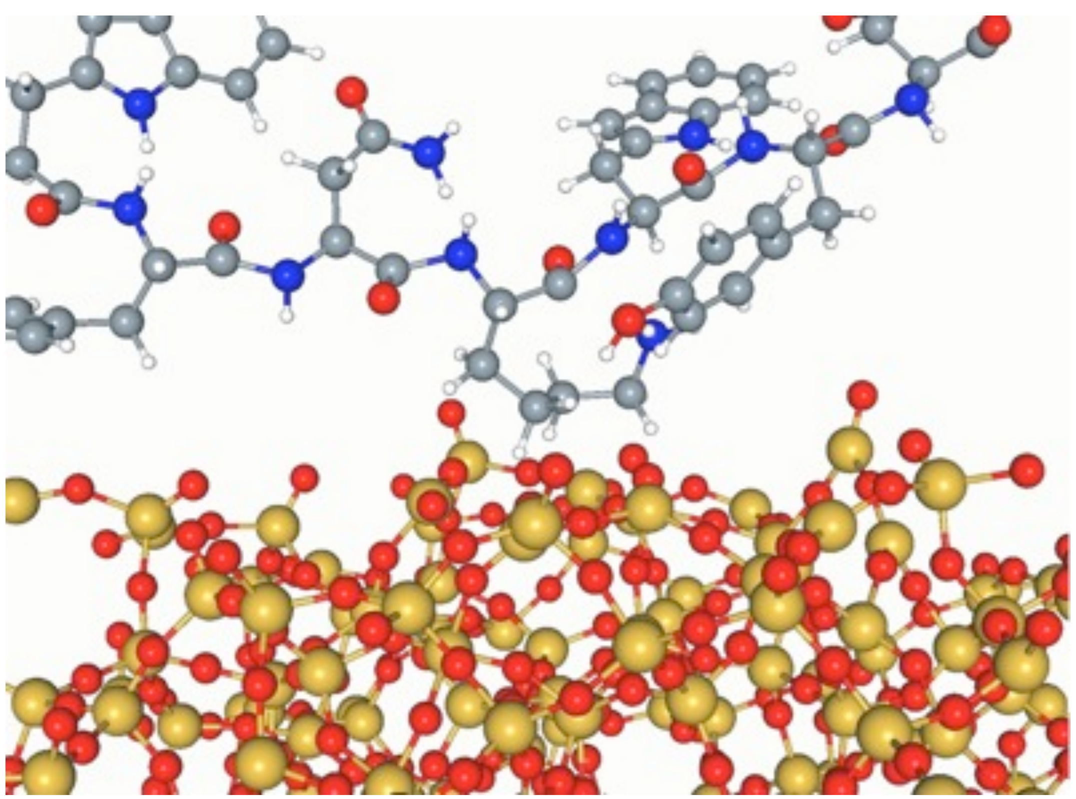
Long range interactions:
each body interacts
with the d others



Interaction energy $f(x)$ of a system $x = \left\{ \text{positions, values} \right\}$

Astronomy Masses

Quantum Chemistry Charges



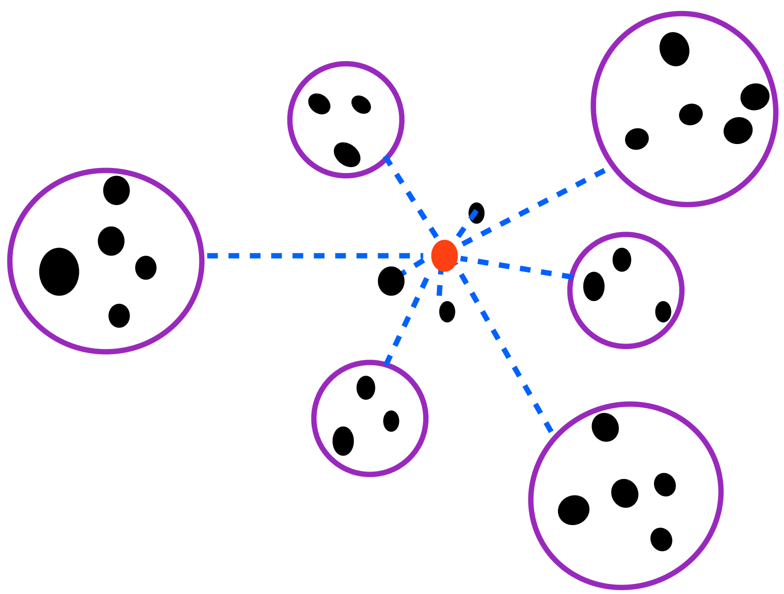
Many Body Interactions

N. Poilvert
Matthew Hirn

- Energy of d interacting bodies:

$$f(x) = \sum_{k=1}^{d} \sum_{k'=1}^{d} \frac{q_k\, q_{k'}}{|p_k - p_{k'}|^{\beta}} \quad \text{with} \quad x(u) = \sum_{k=1}^{d} q_k\, \delta(u - p_k)$$

Fast multipoles: each particle interacts with $O(\log d)$ groups *(Rocklin, Greengard)*

Potential $|r|^{-\beta}$ $\Rightarrow$ 

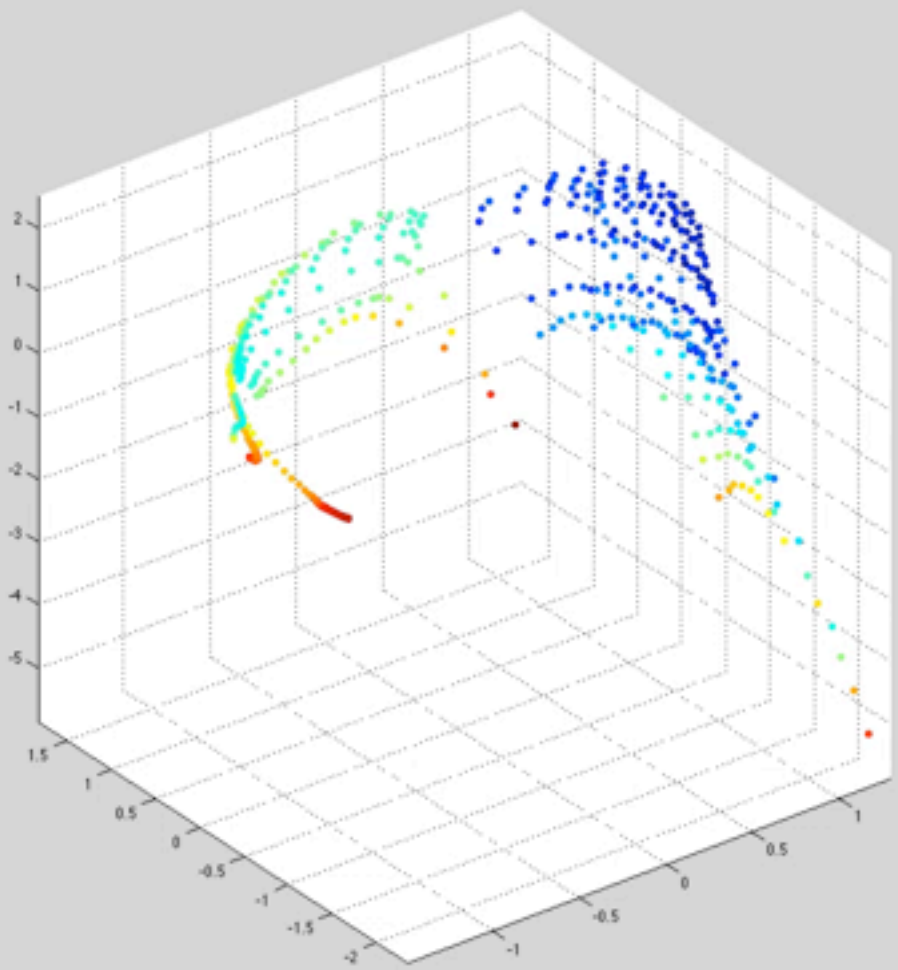
Theorem: For any $\epsilon > 0$ there exists wavelets with

$$f(x) = \sum_{m=0}^{M} \sum_{\lambda_1, \lambda_m} \alpha(\lambda_1, .., \lambda_m)\, S^2 x(\lambda_1, ..., \lambda_m)(1 + \epsilon)$$

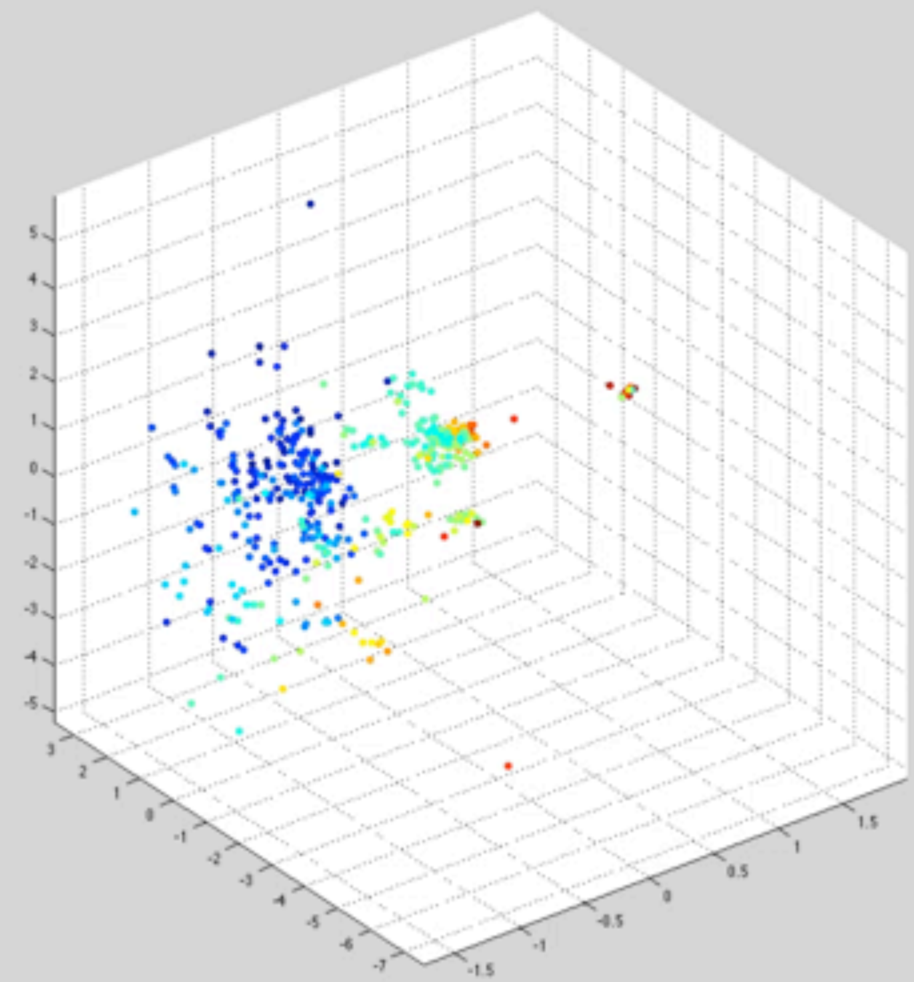
Quantum Chemistry

- Energies $f(x)$ de différentes configurations x de H_2, H_3 et H_4 plongées sur une variété tri-dimensionelle:

Scattering Representation



Fourier Representation

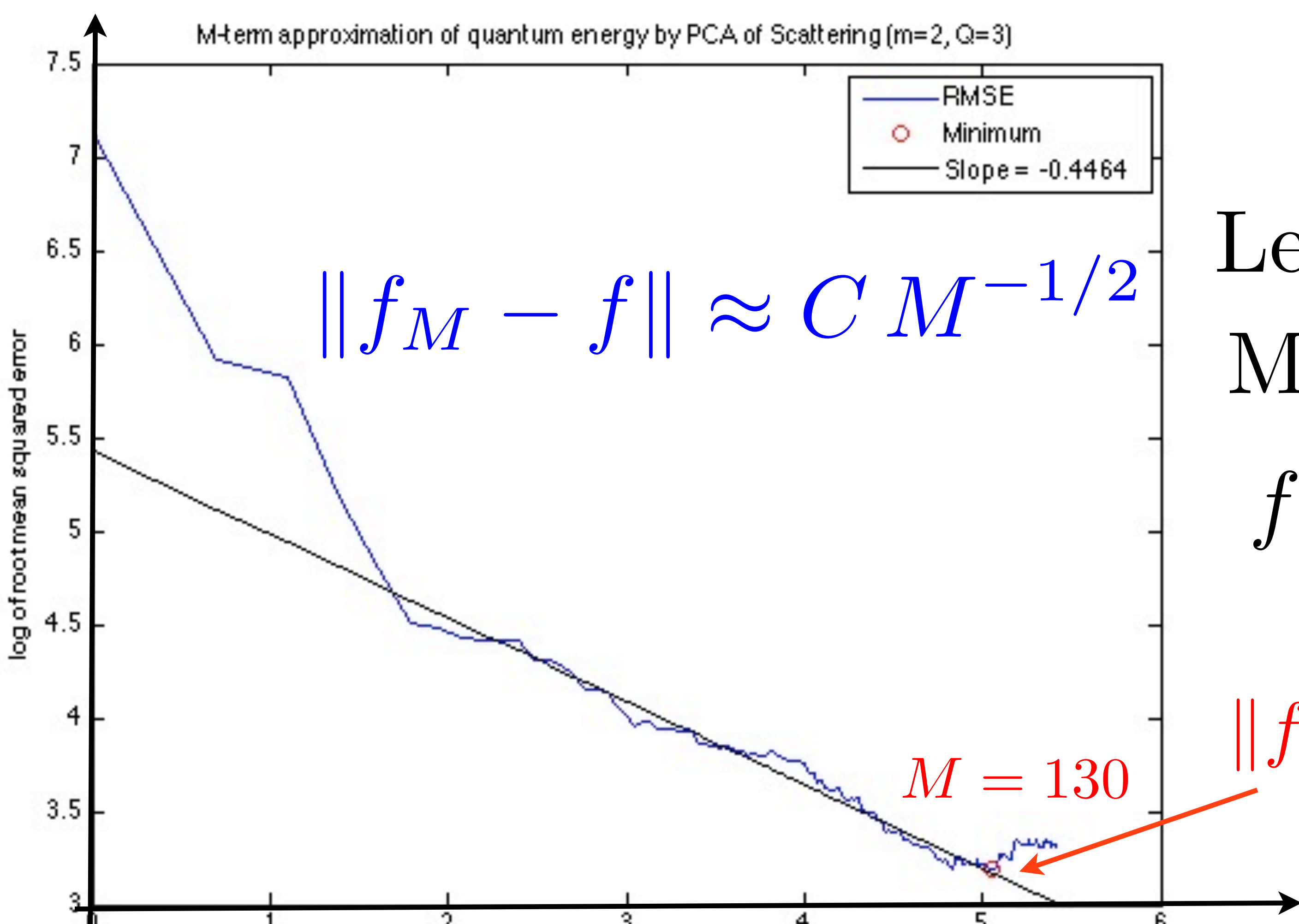


Quantum Chemistry

Matthew Hirn

- Complex orbital interactions: no analytical energy $f(x)$.
- Estimation from $n = 700$ nearly $2D$ molecules: $\{x_i \,,\, f(x_i)\}_{i \leq n}$
- Best M dimensional approximation f_M of f calculated from scattering vectors $\Big\{ Sx(p) \,,\, S^2x_(p) \Big\}_p$

$\log \|f - f_M\|$: M-dimensional scattering error



Learn physics from examples

Multiscale approximation:

$$f(x) \approx \sum_p \alpha_p\, Sx(p) + \beta_p S^2 x(p)$$

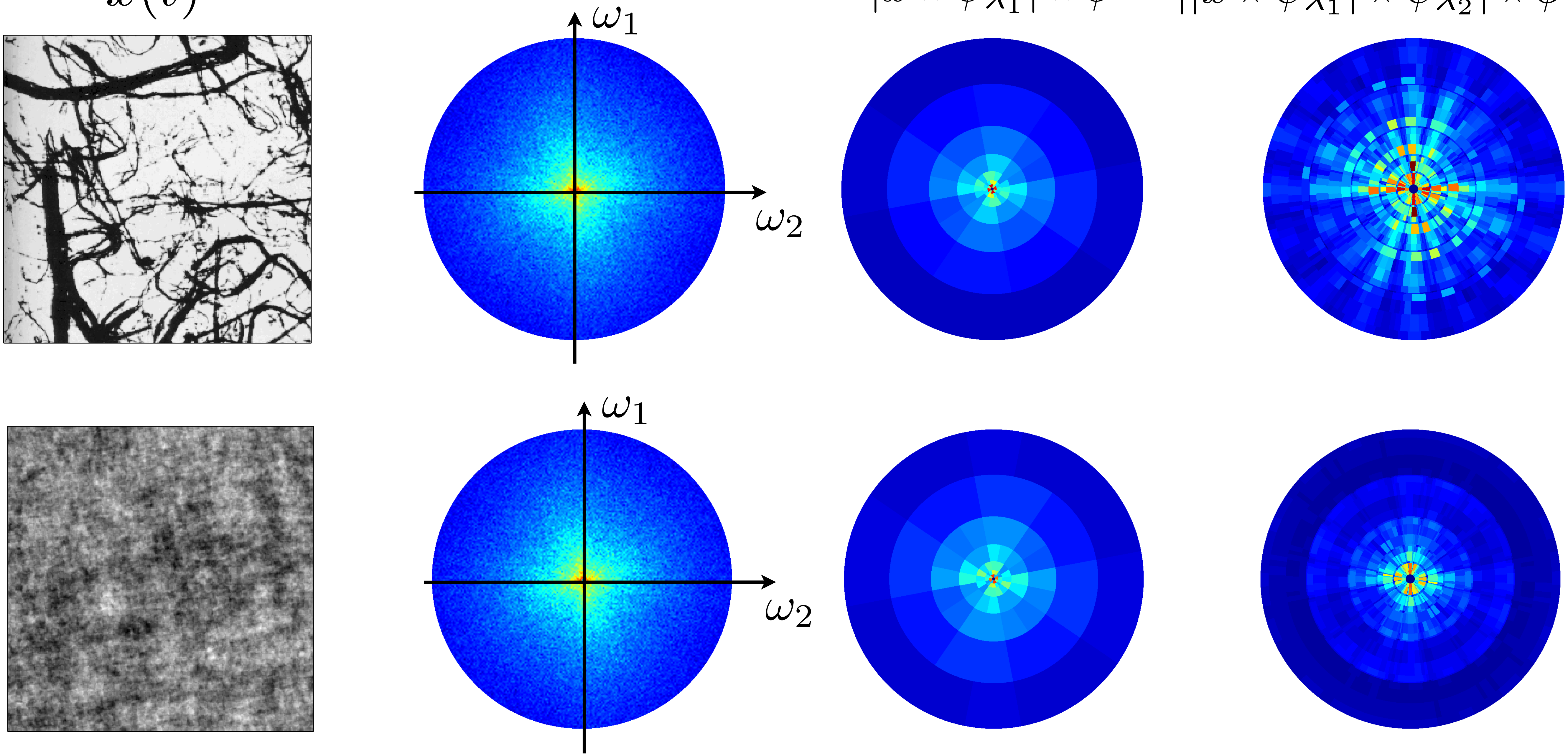
$\|f - f_M\| = 22\,\mathrm{kcal/mole}$

$\log M$

Textures with Same Spectrum

$x(t)$: stationary process

Textures $x(t)$ | Fourier Power Spectrum | Wavelet Scattering $|x \star \psi_{\lambda_1}| \star \phi$ | $||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi$



window size = image size

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 :

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

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

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

Wavelet Tight Frames in L2

Functions in $\mathbf{L^2}(\mathbb{R}^d)$: $\|x\|^2 = \int |x(t)|^2 \, dt < \infty$

Wavelet transform: $Wx = \left(\begin{array}{c} x \star \phi(t) \\ x \star \psi_\lambda(t) \end{array} \right)_{t,\lambda}$

Proposition: *(Littlewood-Paley)*

The wavelet transform is a tight frame for $x \in \mathbf{L^2}(\mathbb{R}^d)$

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

if and only if for almost all ω.

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

Wavelet Frames of Processes

Stationary processes $X(t)$ with $\mathbb{E}(|X(t)|^2) < \infty$.

Wavelet transform: $WX = \begin{pmatrix} \mathbb{E}(X) \\ X \star \psi_\lambda(t) \end{pmatrix}_{t,\lambda}$

Proposition: *(Littlewood-Paley)*

The wavelet transform preserves the variance of stationary X

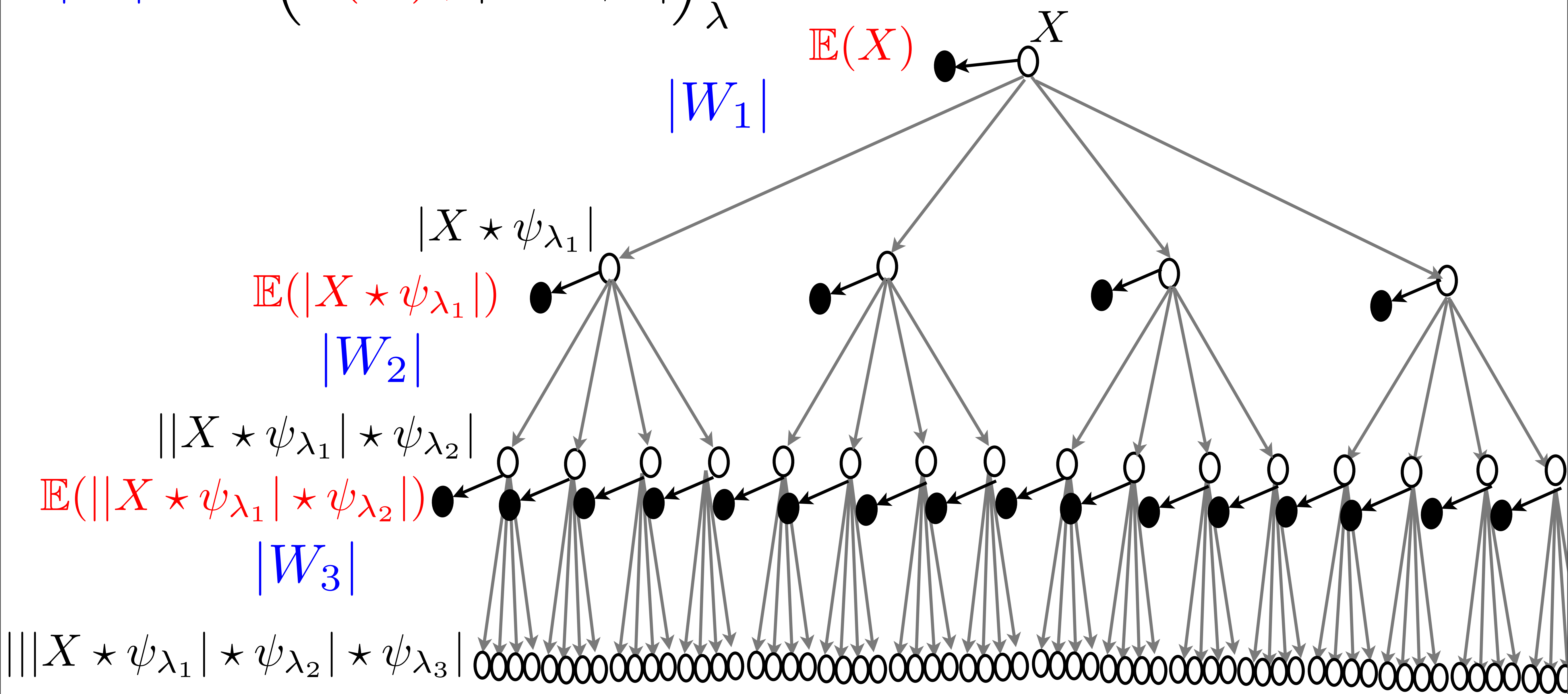
$$\mathbb{E}(X)^2 + \sum_\lambda \mathbb{E}(|X \star \psi_\lambda|^2) = \mathbb{E}(|X|^2)$$

if and only if for almost all ω.

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

Expected Scattering Transform

$$|W|X = \Big(\mathbb{E}(X)\,,\ |X \star \psi_\lambda|\Big)_\lambda$$



- S preserves is contractive because each $|W_k|$ are contractive

Expected Scattering Transform

$X(t)$ stationary process:

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

$$\|\overline{S}X\|^2 = \mathbb{E}(X)^2 + \sum_{m=1}^{\infty} \sum_{\lambda_1,...,\lambda_m} \mathbb{E}\Big(|||X \star \psi_{\lambda_1}| \star ...| \star \psi_{\lambda_m}|\Big)^2$$

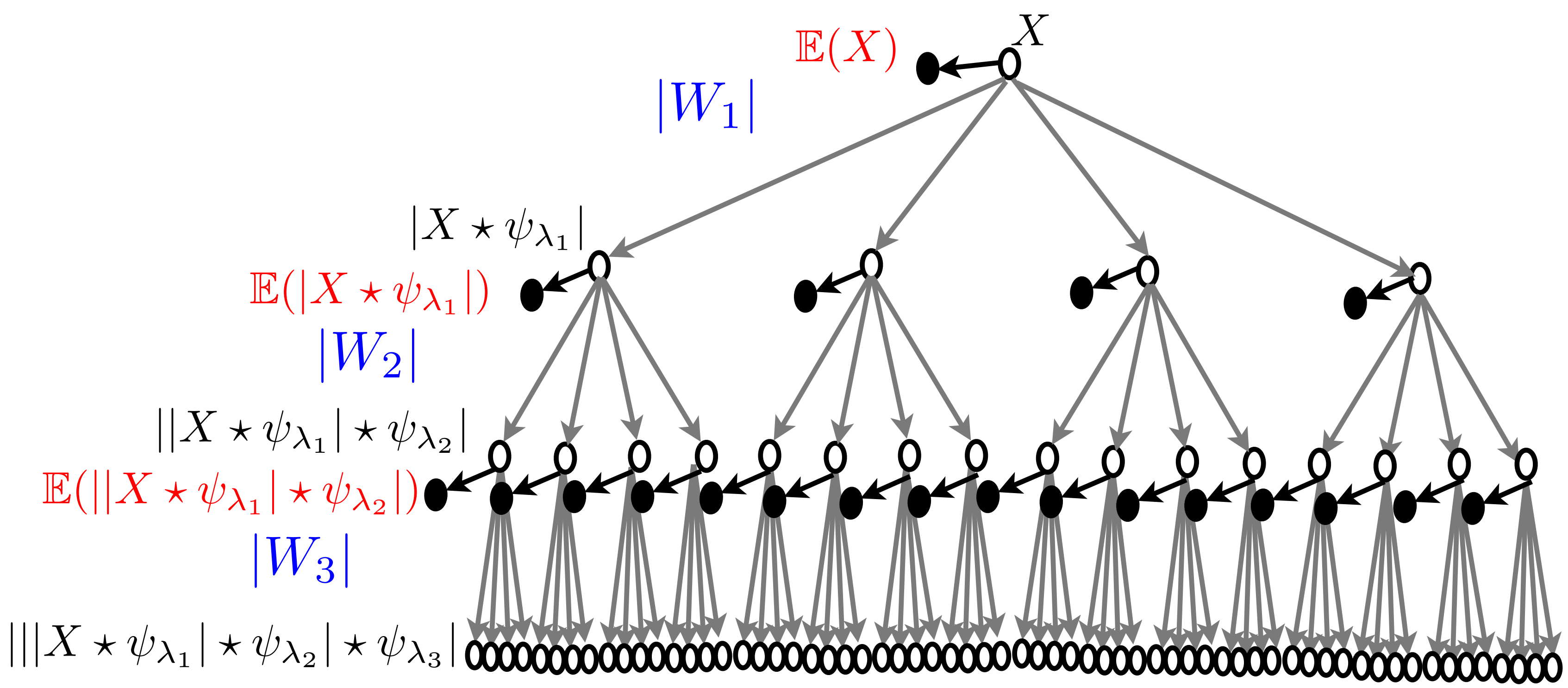
Theorem: *A scattering is*

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

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

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

Expected Scattering Transform



Theorem For any stationary X, equivalent propositions:

(i) The scattering transform is mean-square consistent.

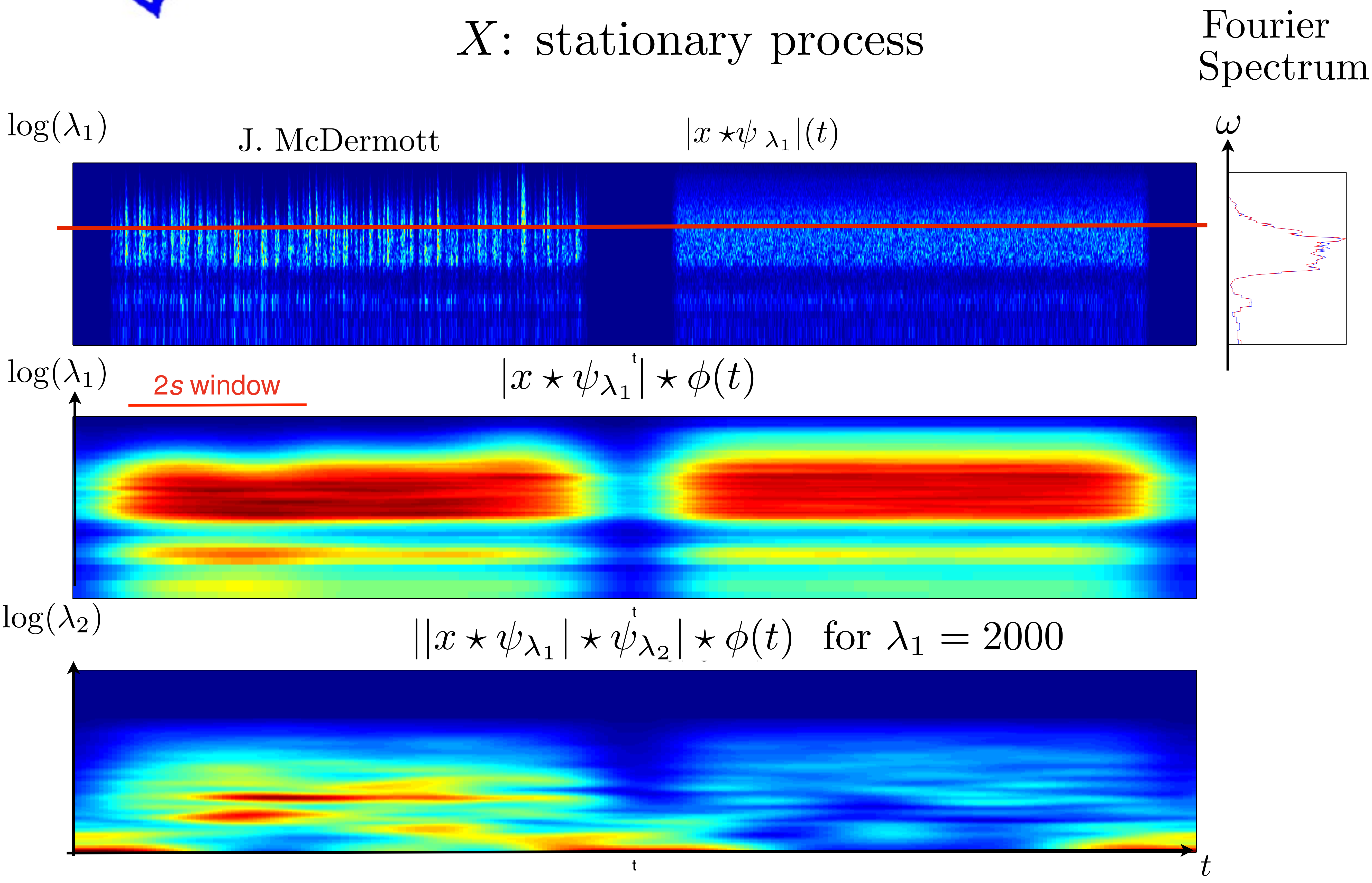
(ii) $\|\overline{S}X\|^2 = E(|X|^2)$

(iii) $$\lim_{m\to\infty} \sum_{\lambda_1,\ldots,\lambda_m} \mathbb{E}\Big(|||X \star \psi_{\lambda_1}|\ldots \star \psi_{\lambda_m}|\Big)^2 = 0$$

- Numerically always verified but not proved.

Sounds with Same Spectrum

X: stationary process



Representation of Random Processes

- An expected scattering is a non-complete representation

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

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: $\quad p(x) = \frac{1}{Z} \exp\Big(\sum_{m=1}^{\infty} \lambda_m \, . \, U_m x \Big)$

Representation of Audio Textures

Joakim Anden Joan Bruna

- $x \in \mathbb{R}^d$ realization of a stationary process

- Gaussien model: covariance $\Rightarrow$ d Fourier spectrum coefficients

- Scattering model Sx of second order: $\log^2 d$ coefficients

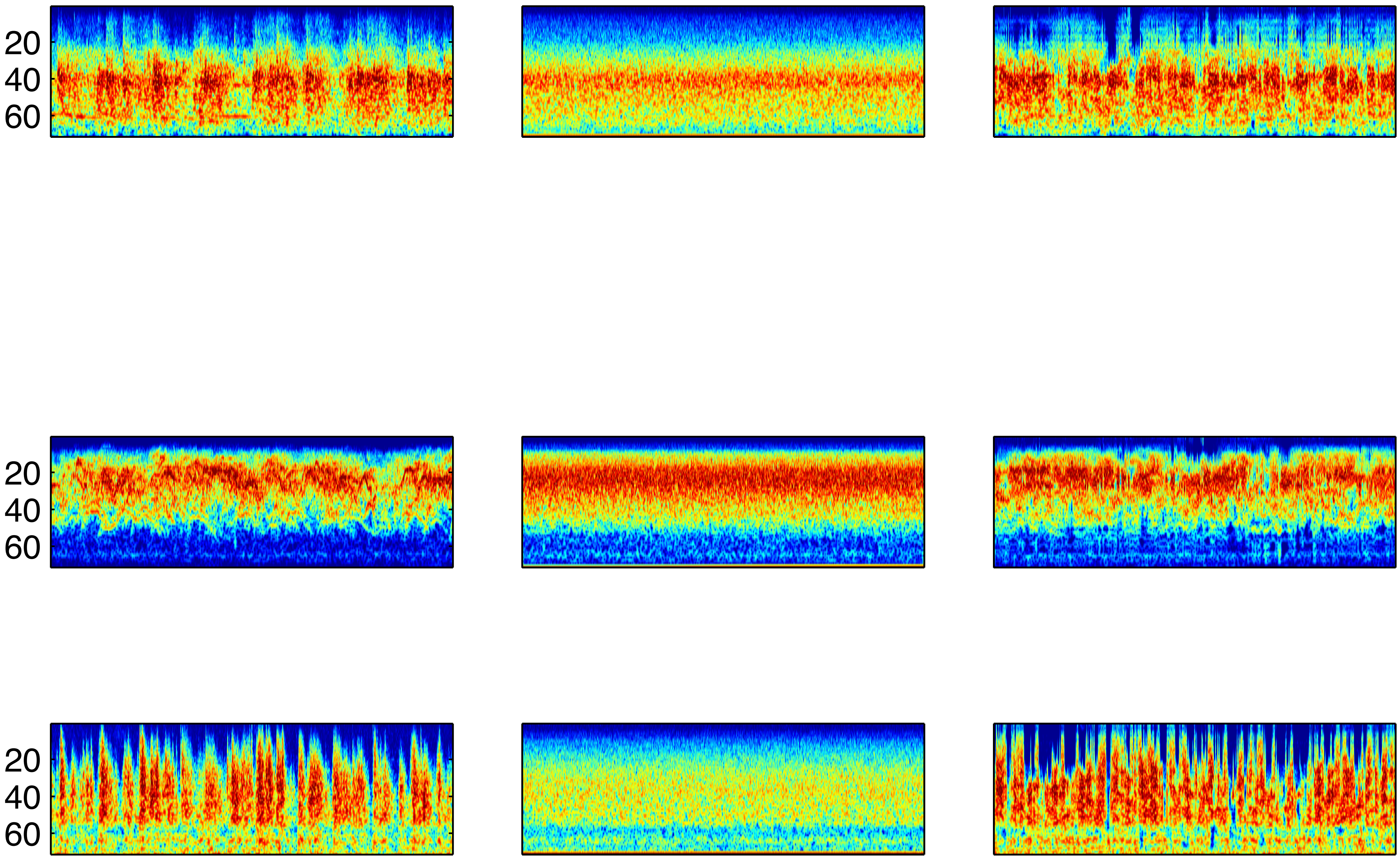
Sample $X(t)$ so that $\|SX - Sx\|$ is small

	Original	Gaussien/Fourier	Scattering
Water			
Paper			
Cocktail Party			

Synthesis Examples

J. Bruna

original [McDermott & Simoncelli'11]	Gaussian	1st+2nd order scattering $K = 500$

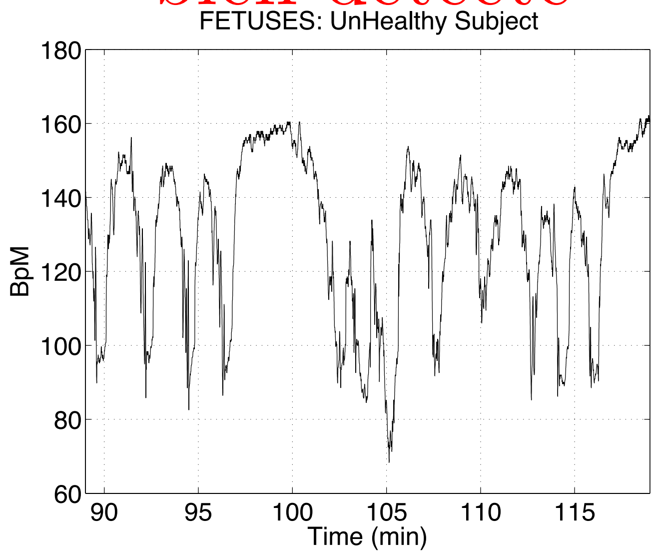


Classification d'ECG

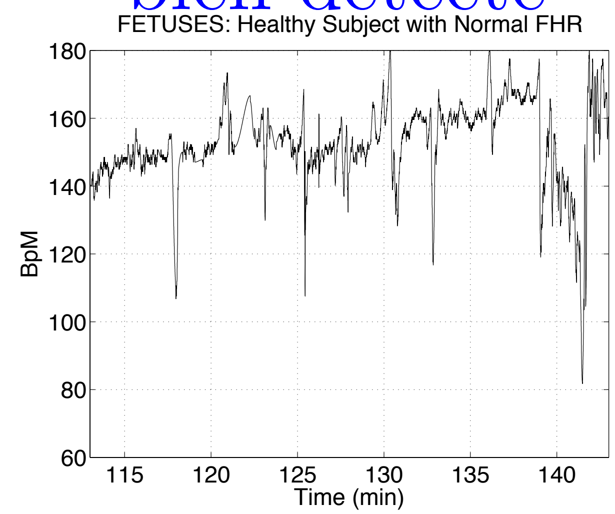
P. Abry, J. Anden, V. Chudacek, M. Doret, R. Talmon

- Mesure du niveau de stress d'un Fetus avant accouchement

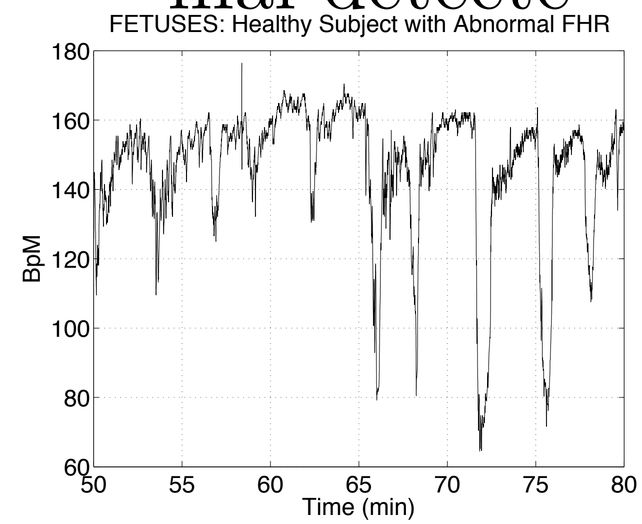
Stress
bien détecté



Sain
bien détecté

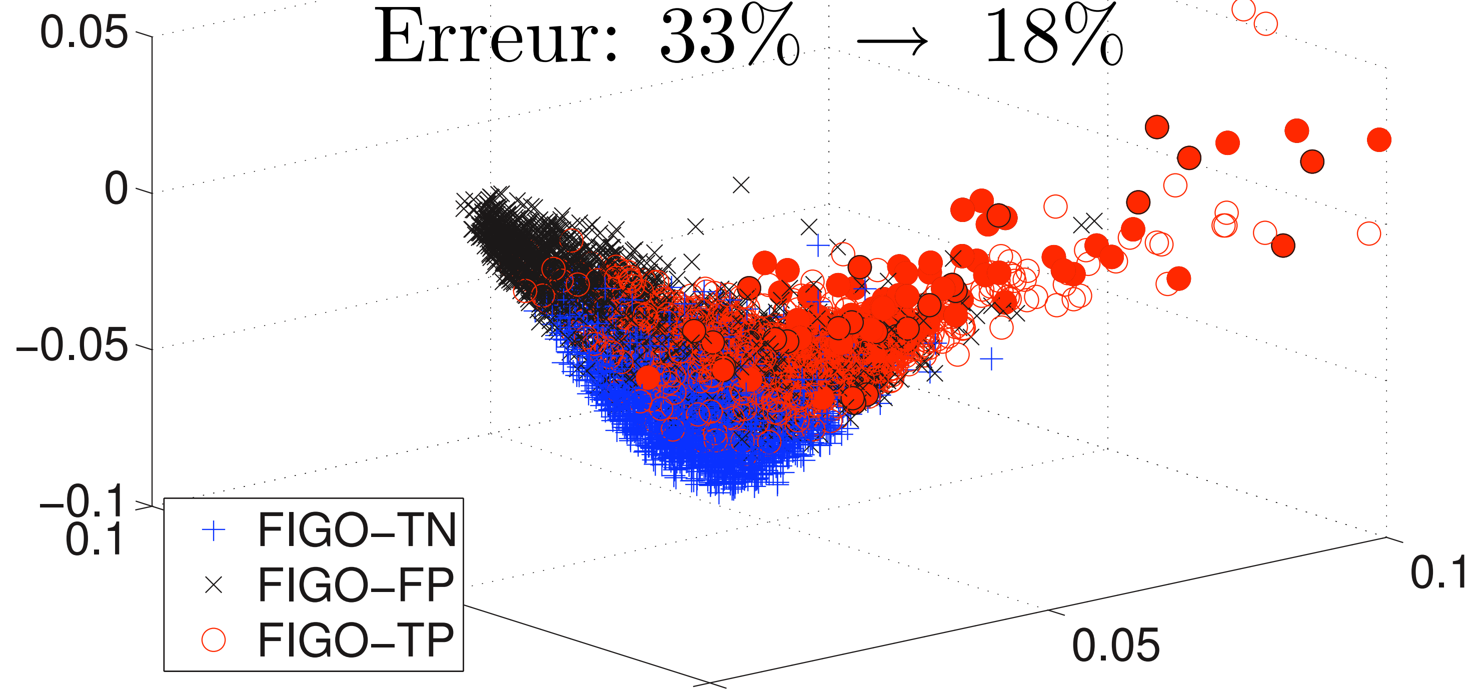


Sain
mal detecté



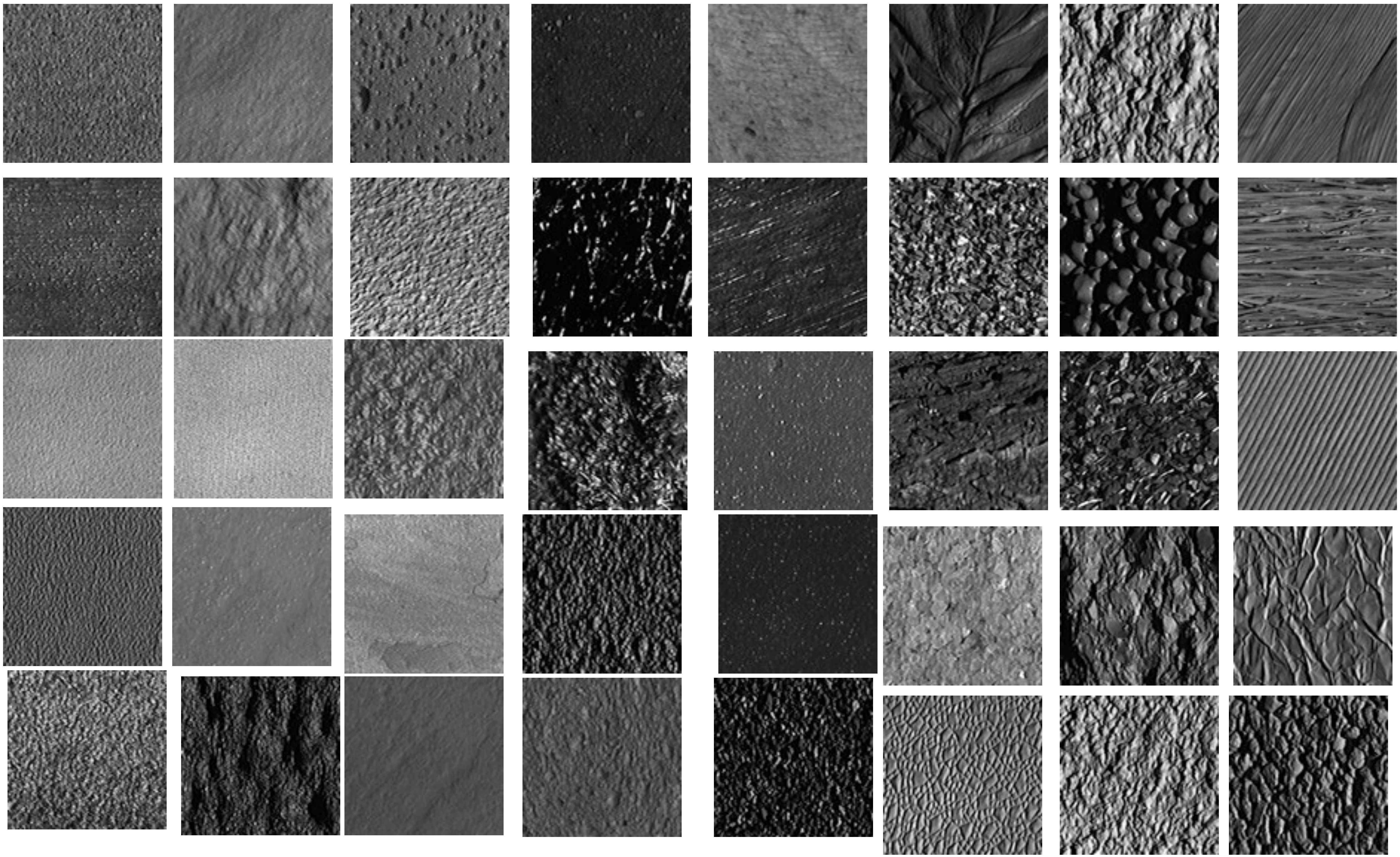
Variété dans l'espace des coefficients Sx

Erreur: $33\% \rightarrow 18\%$



Classification of Textures

40 classes of CureT



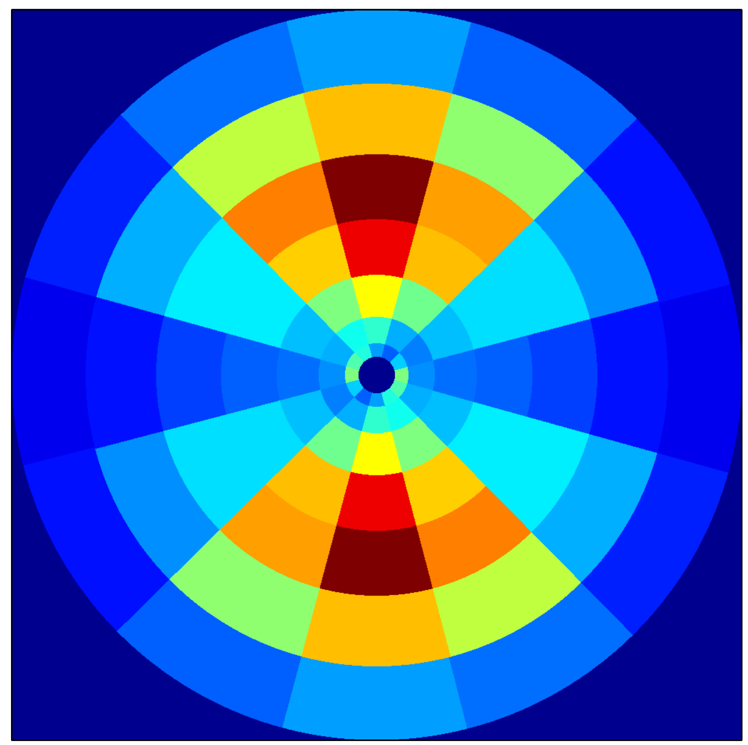
Classification of Textures

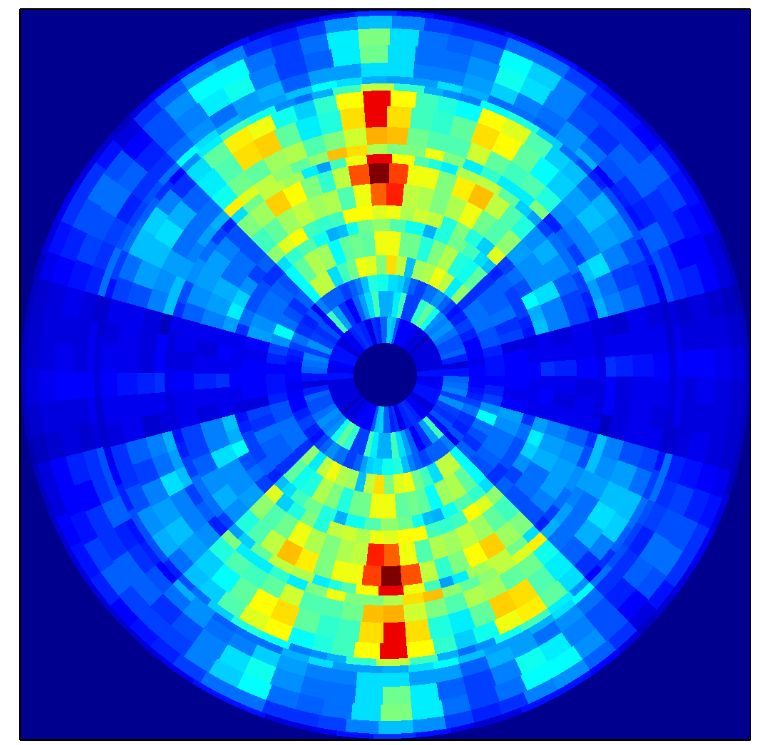
Expected Scattering
estimated with $\phi = 1$

X

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

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



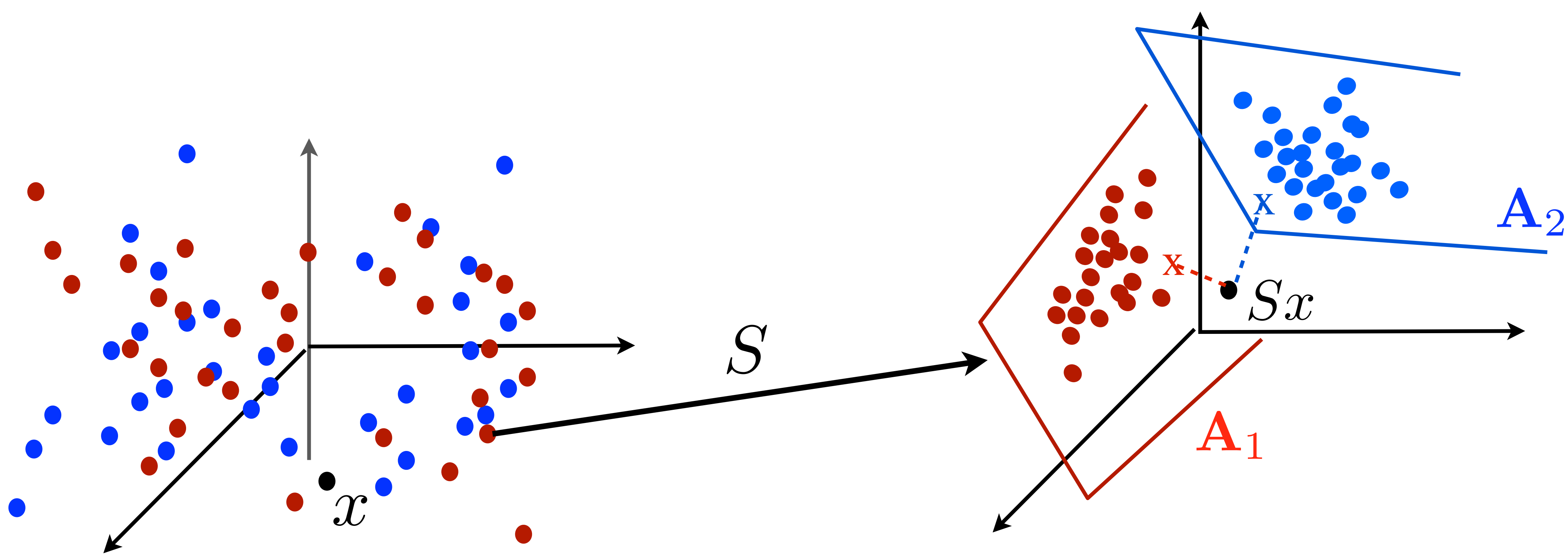


Affine Space Classification

Joan Bruna

- Each class is represented by a random process X_k

  The support of SX_k is approximated by a low-dimensional affine space $\mathbf{A}_k$ computed with a PCA.

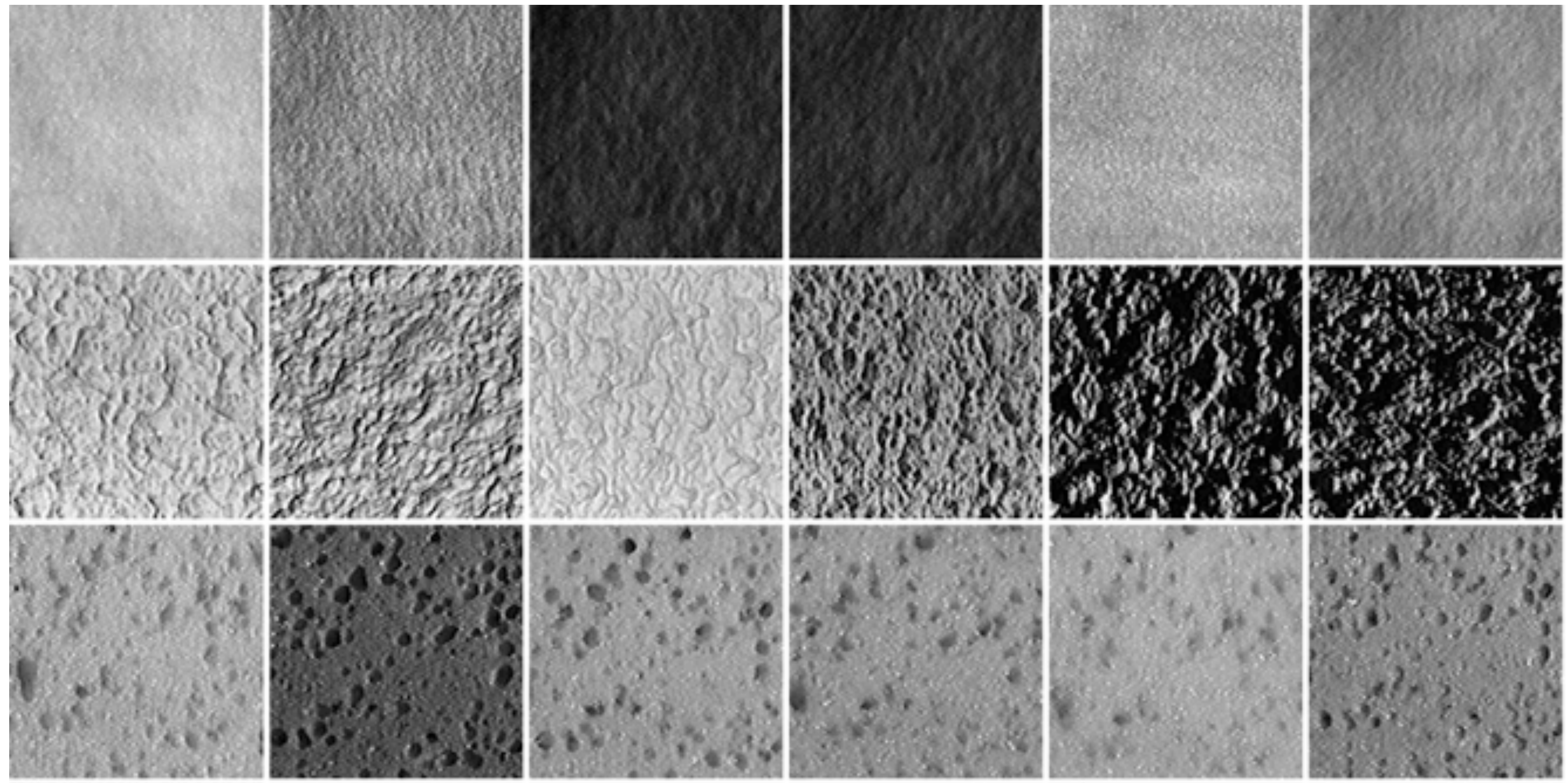


$$\hat{k}(x) = \arg\max_k \|Sx - P_{\mathbf{A_k}} Sx\| \ .$$

Classification of Textures

J. Bruna

CUREt database
61 classes

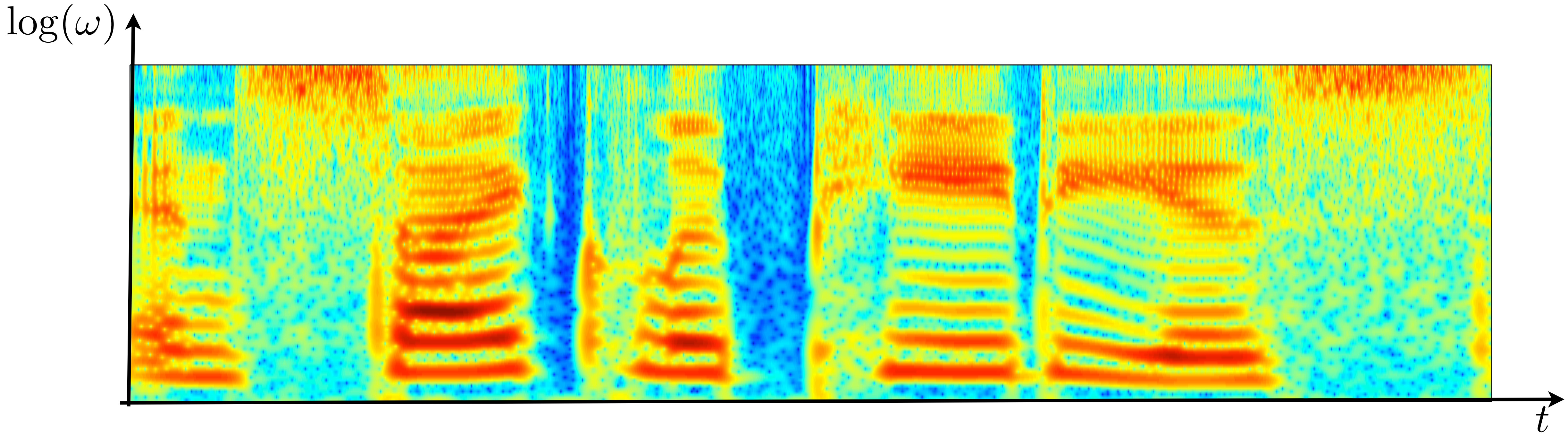


$x \rightarrow$ $\boxed{Sx}$ $\rightarrow$ $\boxed{\text{Supervised Linear Classifier: PCA/SVM}}$ $\rightarrow y$

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

Frequency Transpositions

Time and frequency translations and deformations:



- Frequency transposition invariance is needed

  for speech recognition not for locutor recognition.

Transposition Invariance

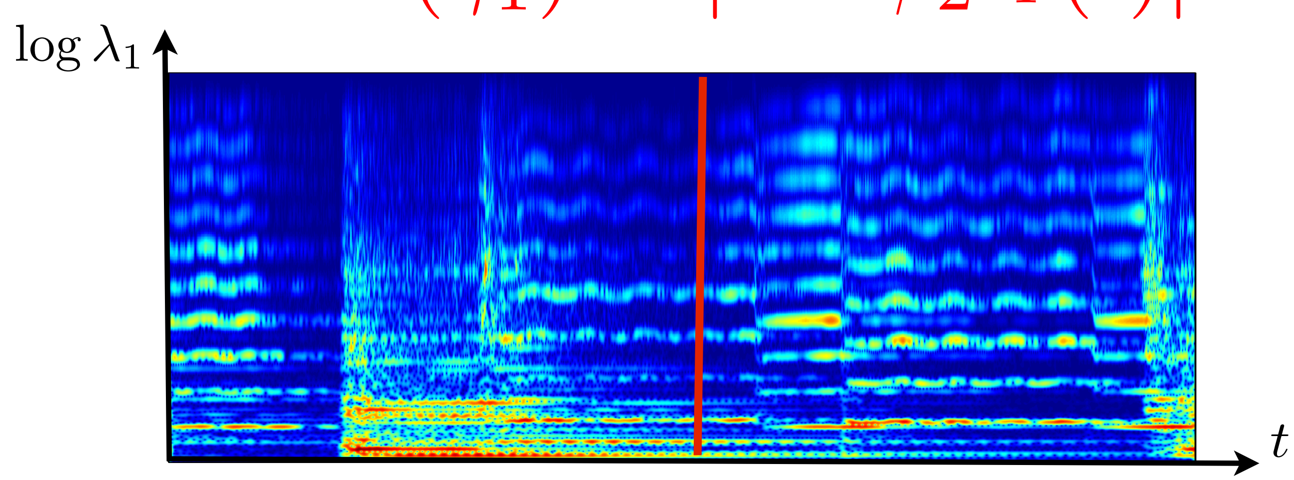
J. Anden

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

$x(t) \longrightarrow$ Scattering along t $\longrightarrow$ 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

Phone Classification (TIMIT)

J.Anden

- Training on 3696 phrases (139868 phones) and and testing on 192 phrases (7201 phones)
- Each phone is classified using a Gaussian kernel SVM.

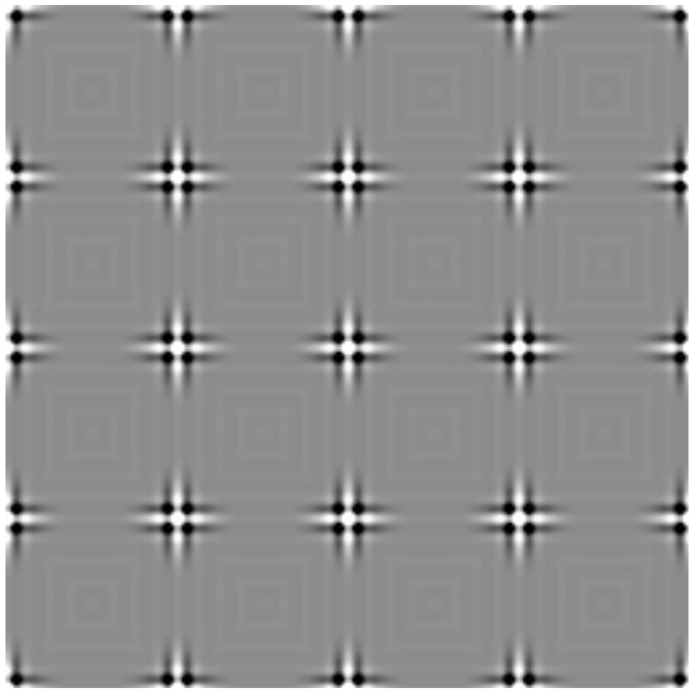
$T = 32$ ms

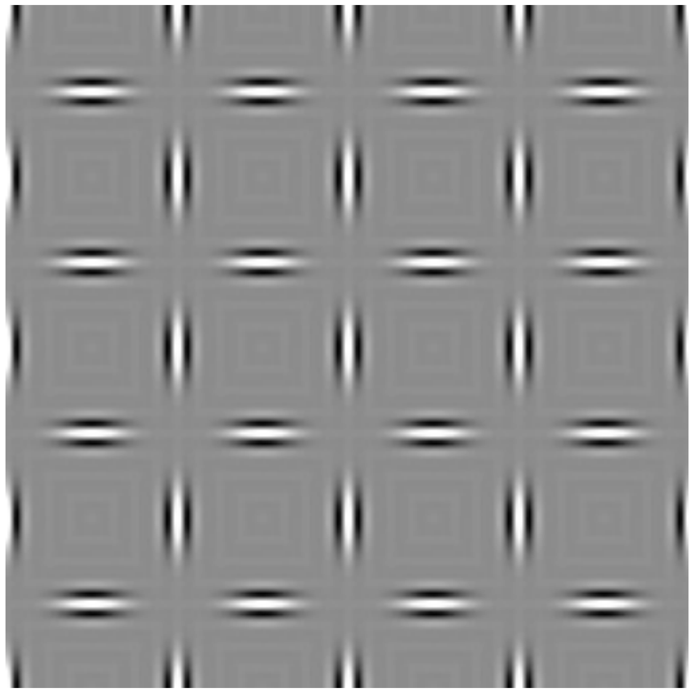
Feature Set	Error (%)
Δ-MFCC (32 ms)	19.3
State of the art (excl. scattering)	16.7
Time Scat., m = 1	18.5
Time Scat., m = 2	17.7
Time & Freq. Scat., m = 2	16.5

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:

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

- Inverse: $(r,t)^{-1} = (r^{-1}, -r^{-1}t)$.

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

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

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

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

Scattering on a Lie Group

L. Sifre

- How to define a wavelet transform of $x(r,t) \in \mathbf{L^2}(G)$?
- One can define separable complex wavelets $\overline{\psi}_{\lambda_2}(r,t) \in \mathbf{L^2}(G)$

$$W_2 x = \left(\begin{array}{c} x \circledast \overline{\phi}(r,t) \\ x \circledast \overline{\psi}_{\lambda_2}(r,t) \end{array} \right)_{\lambda_2, r, t} \quad \text{is tight frame of } \mathbf{L^2}(G).$$

$$x \circledast \overline{\psi}_\lambda(g) = \int_G x(g')\, \overline{\psi}_\lambda(g'^{-1} g)\, dg'$$

$$\|x\|^2 = \int_G |x(g)|^2 dg = \|x \circledast \phi\|^2 + \sum_{\lambda_2} \|x \circledast \psi_{\lambda_2}\|^2$$

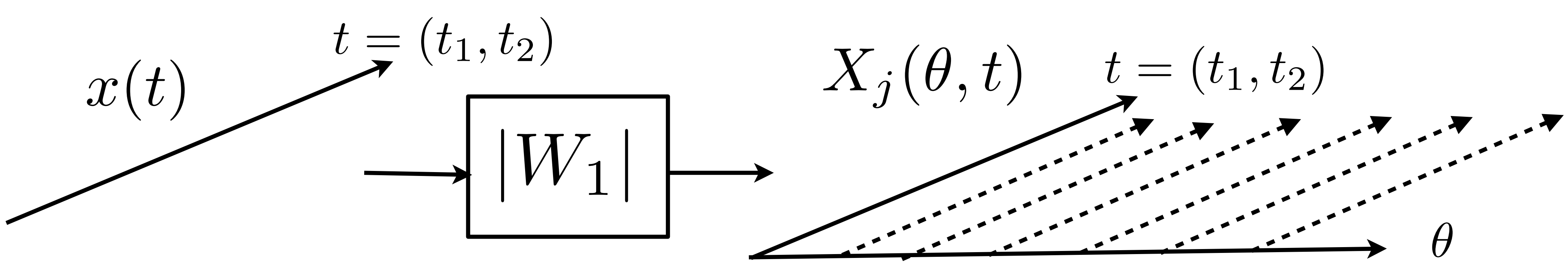
Translation Invariance

Laurent Sifre

translation $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ translation

$$x \rightarrow \boxed{|W_1|} \rightarrow |x \star \psi_{2^j r_\theta}(t)| = X_j(\underbrace{\theta, t}_{\text{3D variable}}) \rightarrow \boxed{|W_2|} \rightarrow |X_j(\theta, .) \star \psi_{\lambda_2}(t)|$$

$$\boxed{|W_1|} \downarrow x \star \phi(t) \qquad\qquad \boxed{|W_2|} \downarrow X_j(\theta, .) \star \phi(t)$$

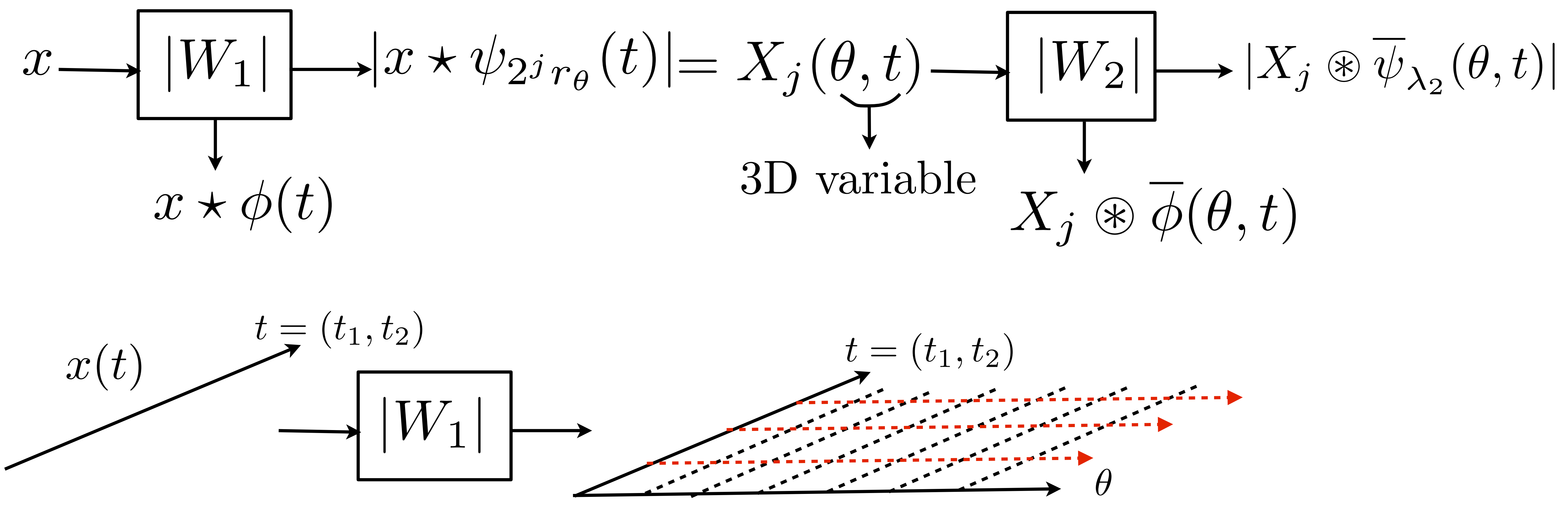


- Convolutions along translation parameter: t

Rotation-Translation Invariance

Laurent Sifre

translation roto-translation



- Convolutions along translation parameter: t

  Convolutions along rotation parameter: θ

Rotation-Scale Invariance

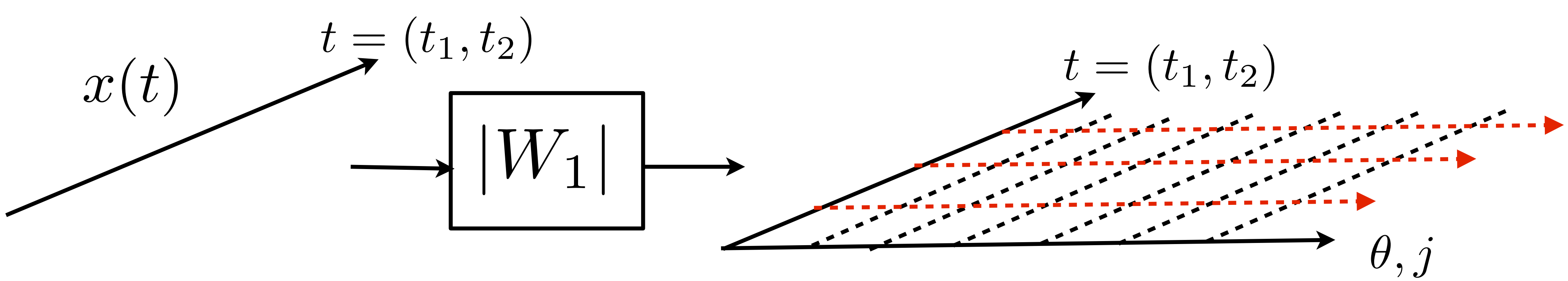
Laurent Sifre

translation

scalo-roto-translation

$$x \longrightarrow \boxed{|W_1|} \longrightarrow |x \star \psi_{2^j r_\theta}(t)| = X(\underbrace{j, \theta, t}_{\text{4D variable}}) \longrightarrow \boxed{|W_2|} \longrightarrow |X \circledast \overline{\psi}_{\lambda_2}(j, \theta, t)|$$

$$x \star \phi(t) \qquad\qquad X \circledast \overline{\phi}(j, \theta, t)$$



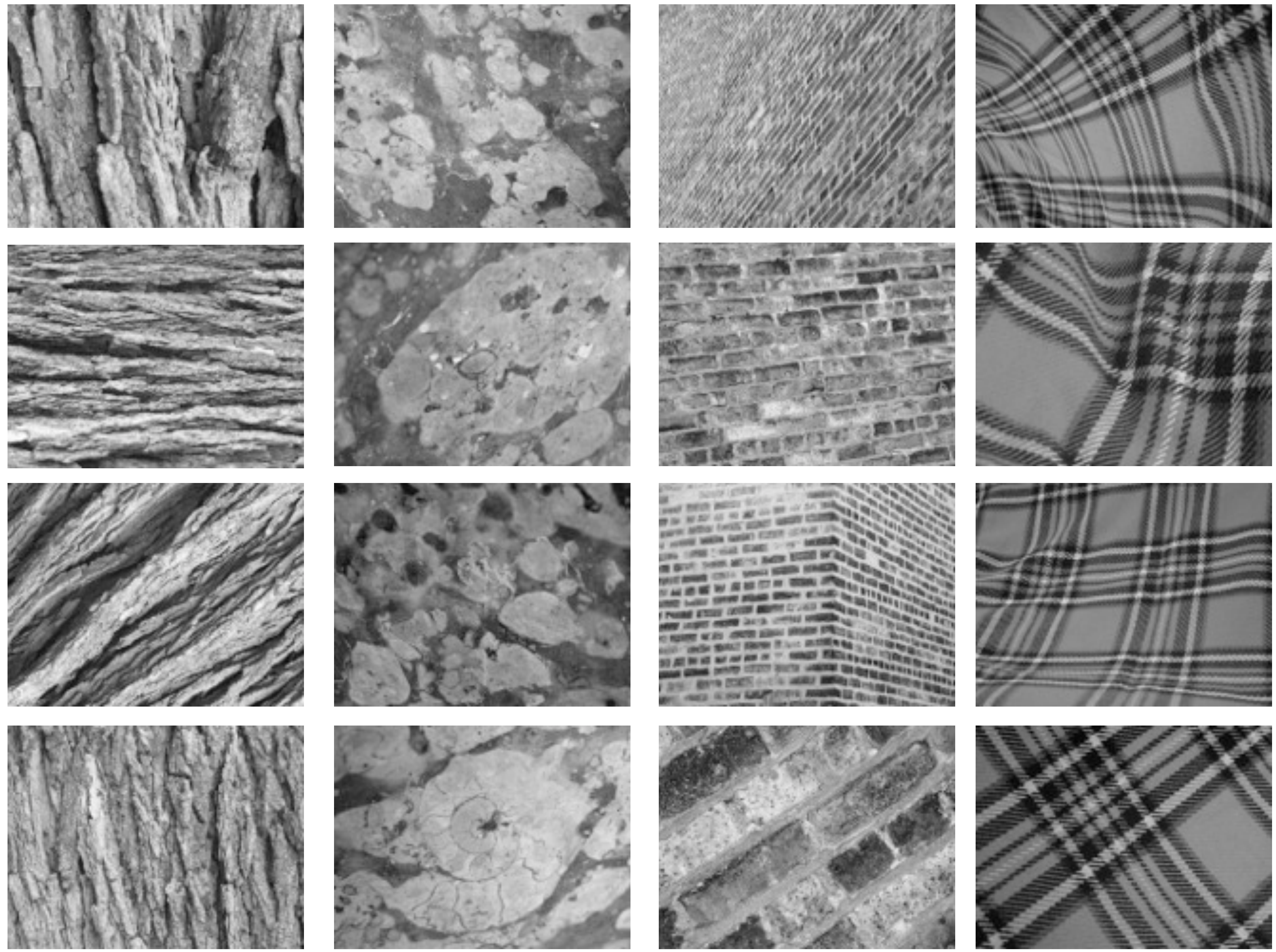
- Convolutions along translation parameter: t

  Convolutions along rotation and scale parameters: θ , j

Rotation and Scaling Invariance

Laurent Sifre

UIUC database:
25 classes



Scattering classification errors

Training	Translation	Transl + Rotation	+ Scaling
20	20 %	2%	**0.6**%

$x \rightarrow \boxed{Sx} \rightarrow$ Supervised Linear Classifier: PCA/SVM $\rightarrow y$

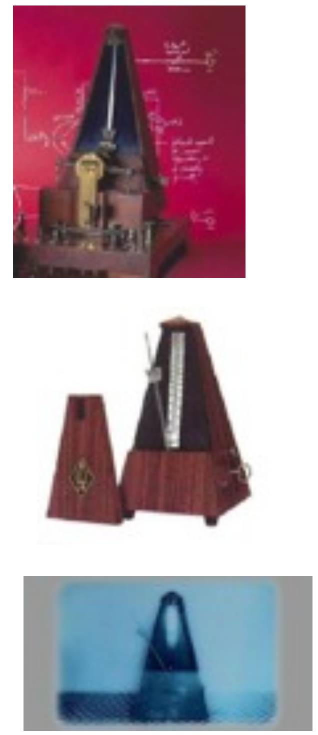
Complex Source of Variability

CalTech 101/256 data-basis:

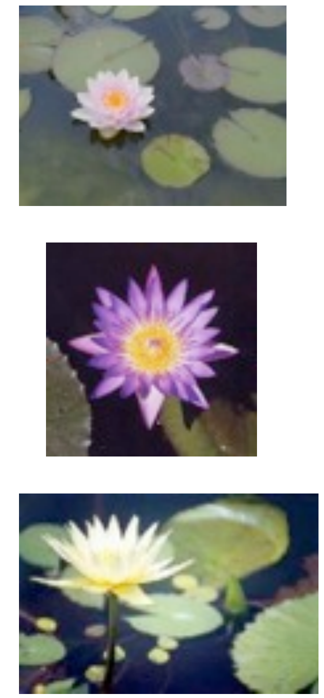
Arbre de Joshua

Ancre

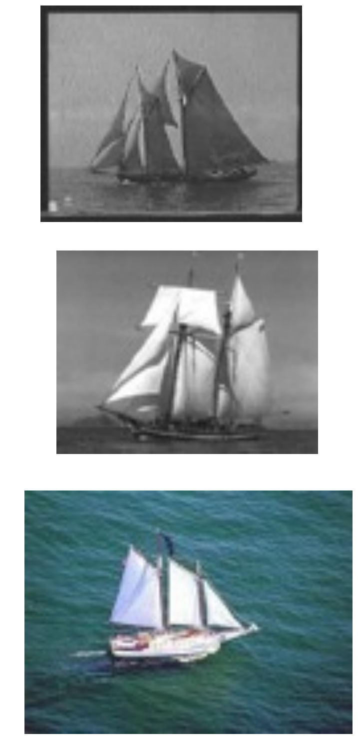
Metronome



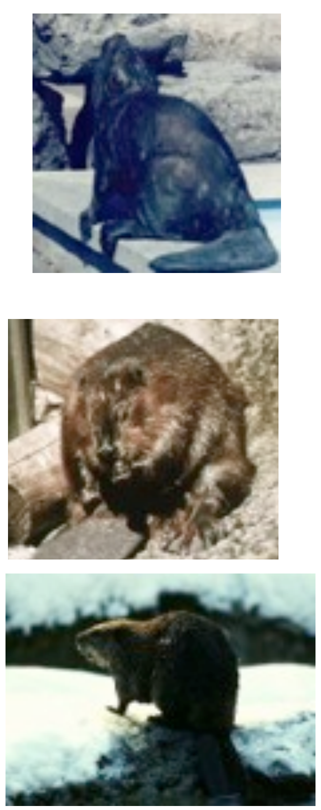
Nénuphare



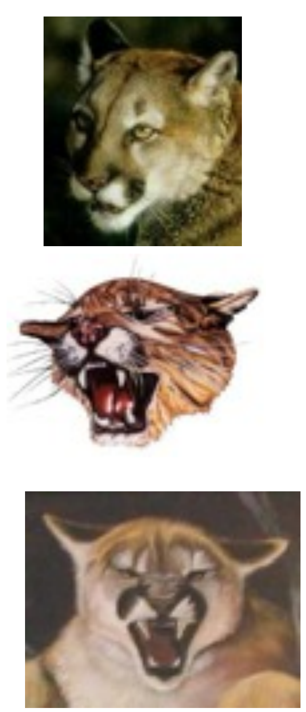
Bateau



Castore

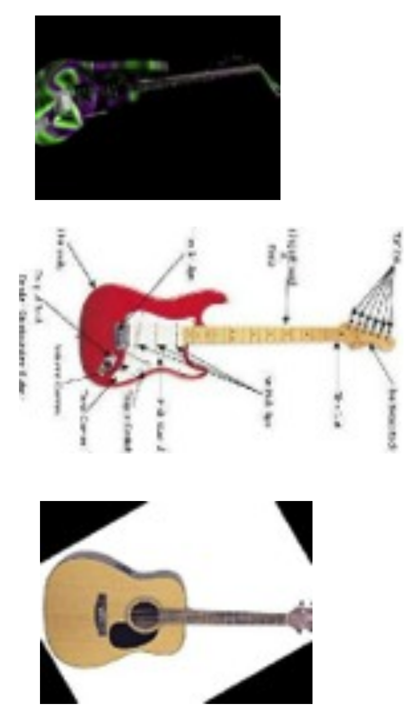


Puma

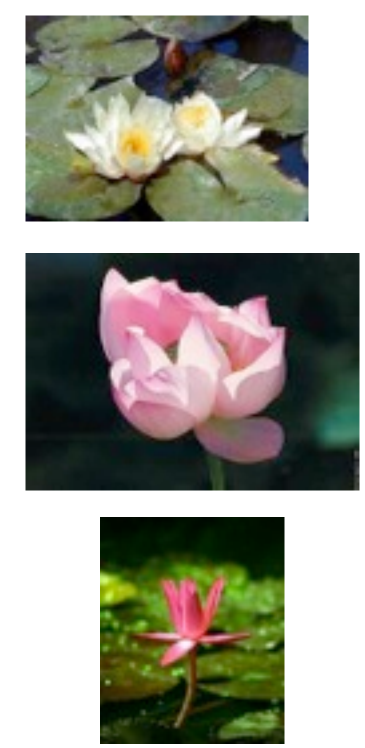


Buddha

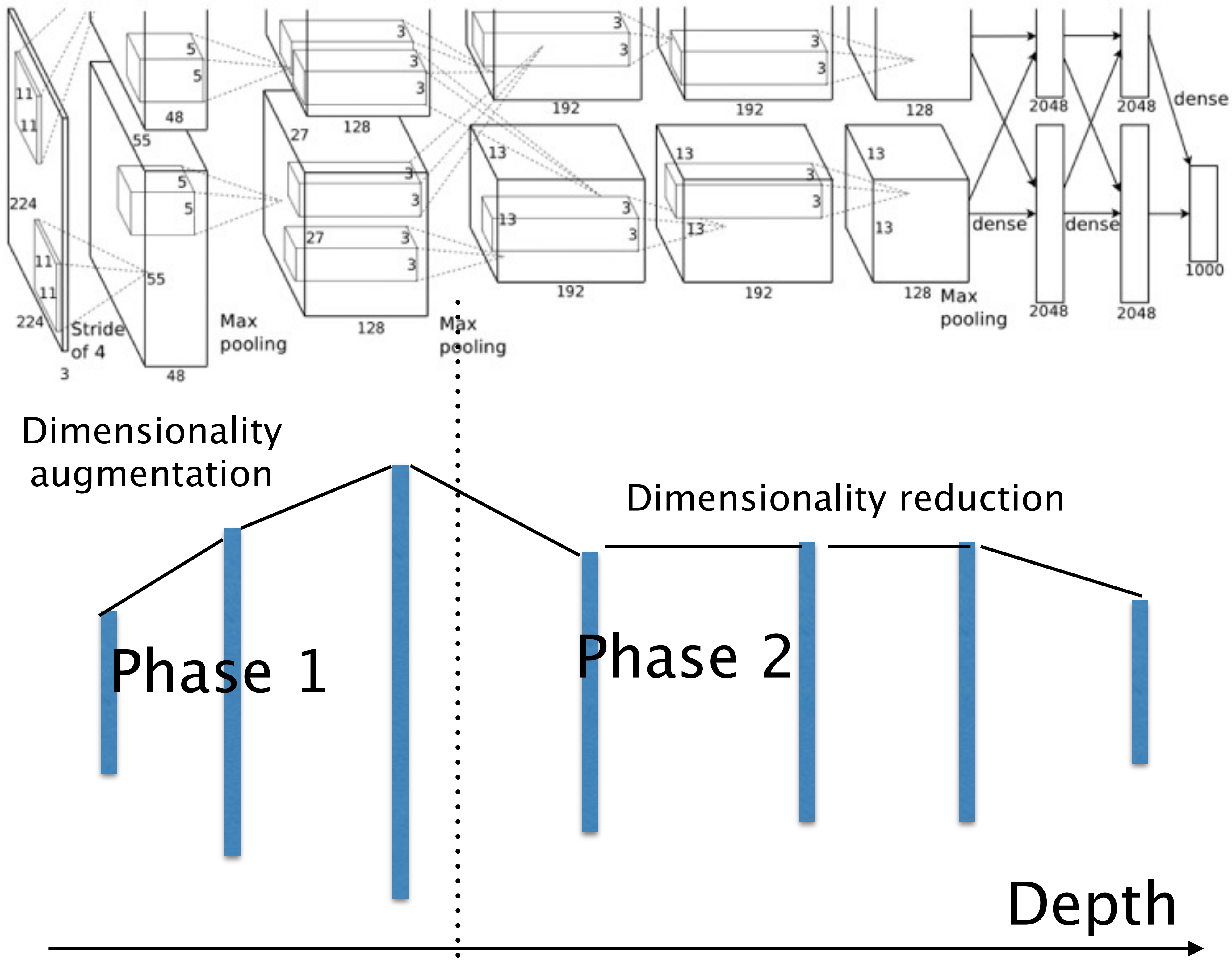
Guitare



Lotus



Alex Deep Neural Network

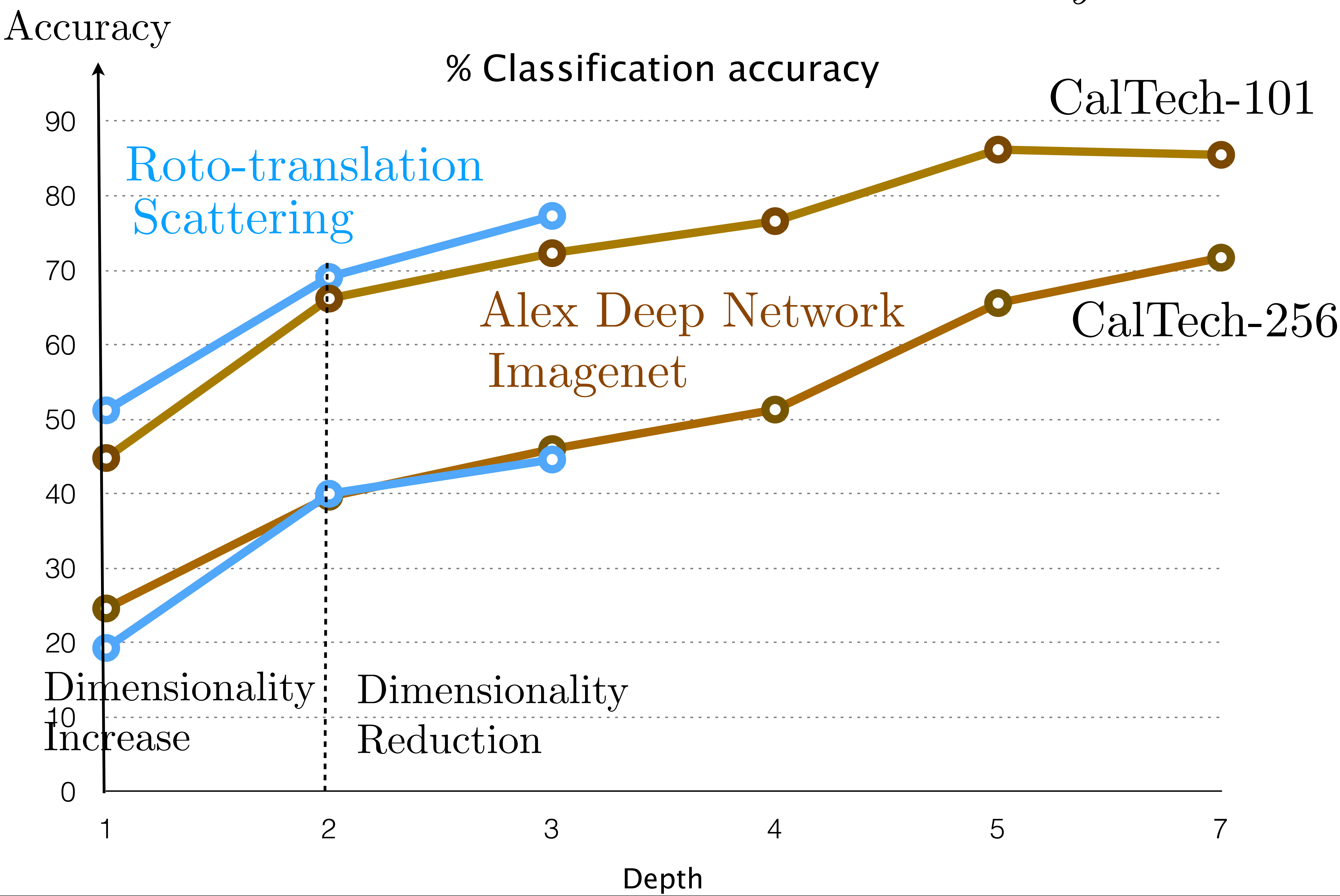


Problems/Datasets in Computer Vision

- Imagenet
  - 14,000,000 images (1,000,000 with bounding box annotations)
  - 20000 categories

Cal-Tech Classification

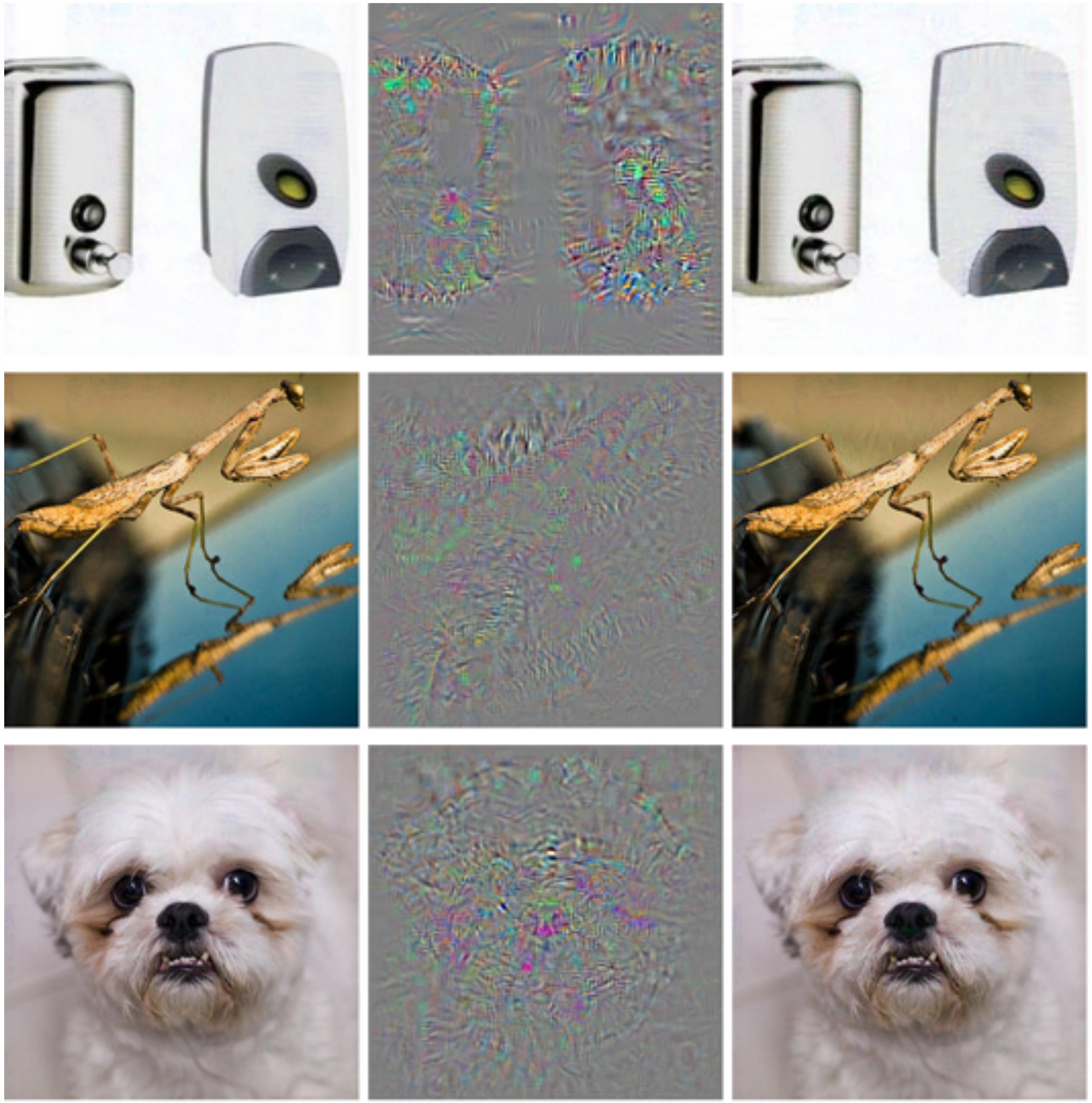
E. Oyallon



Instabilities of Deep Networks

[J. Bruna Szegedy *et al, ICLR'14*]

x $\tilde{x}$



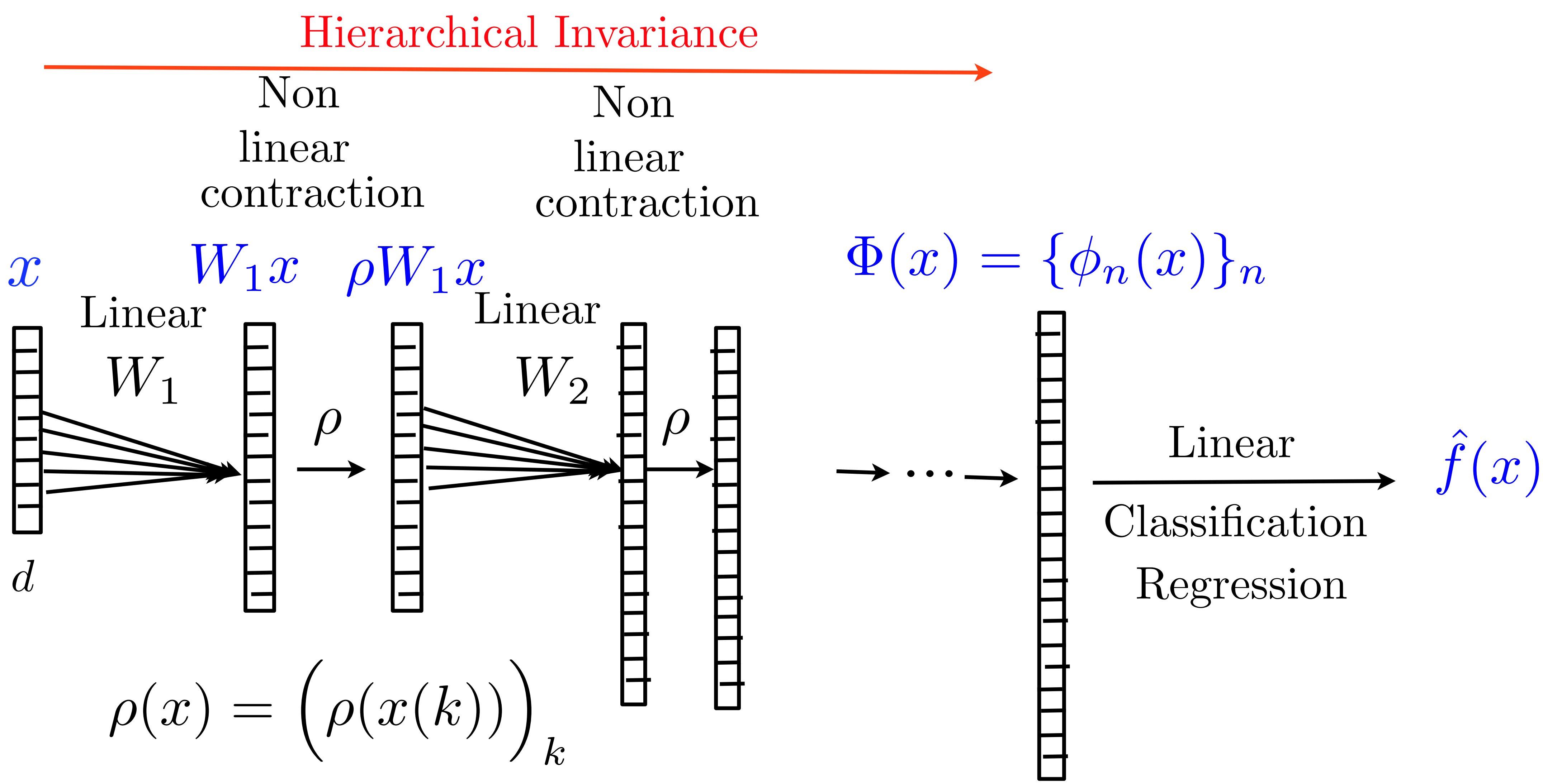
correctly classified

classified as ostrich

Alex Krizhevsky's Imagenet
8 layer Deep ConvNet

$$\|x - \tilde{x}\| < 0.01\|x\|$$

Deep Neural Neworks



$$\rho(x) = \Big(\rho(x(k))\Big)_k$$

For example: $\rho(u) = \max(0, u)$ or $\rho(u) = |u|$

Convolution networks: $W_m x(k) = \{x \star g_l(2k)\}_{l \leq K}$

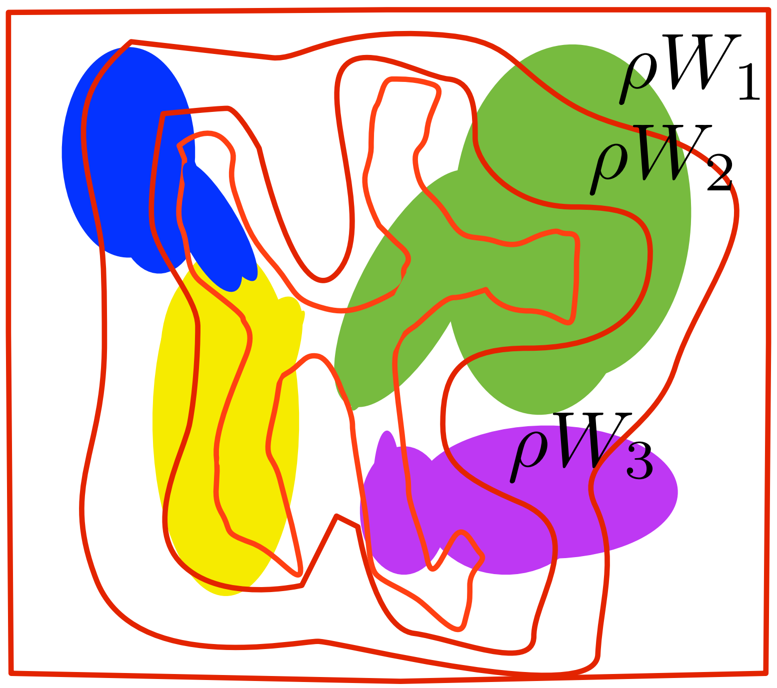
Iterated Contractions

- Reduce the space volume with iterated contractions

$$\Phi x = \rho\, W_m \ldots \rho\, W_2\, \rho\, W_1 x$$

  - W_k preserve distances: $\|W_k x - W_k x'\| = \|x - x'\|$
  - ρ is a contraction

- Iterative space contraction: reduce intra-class variability but avoid reducing class distances: margin condition.

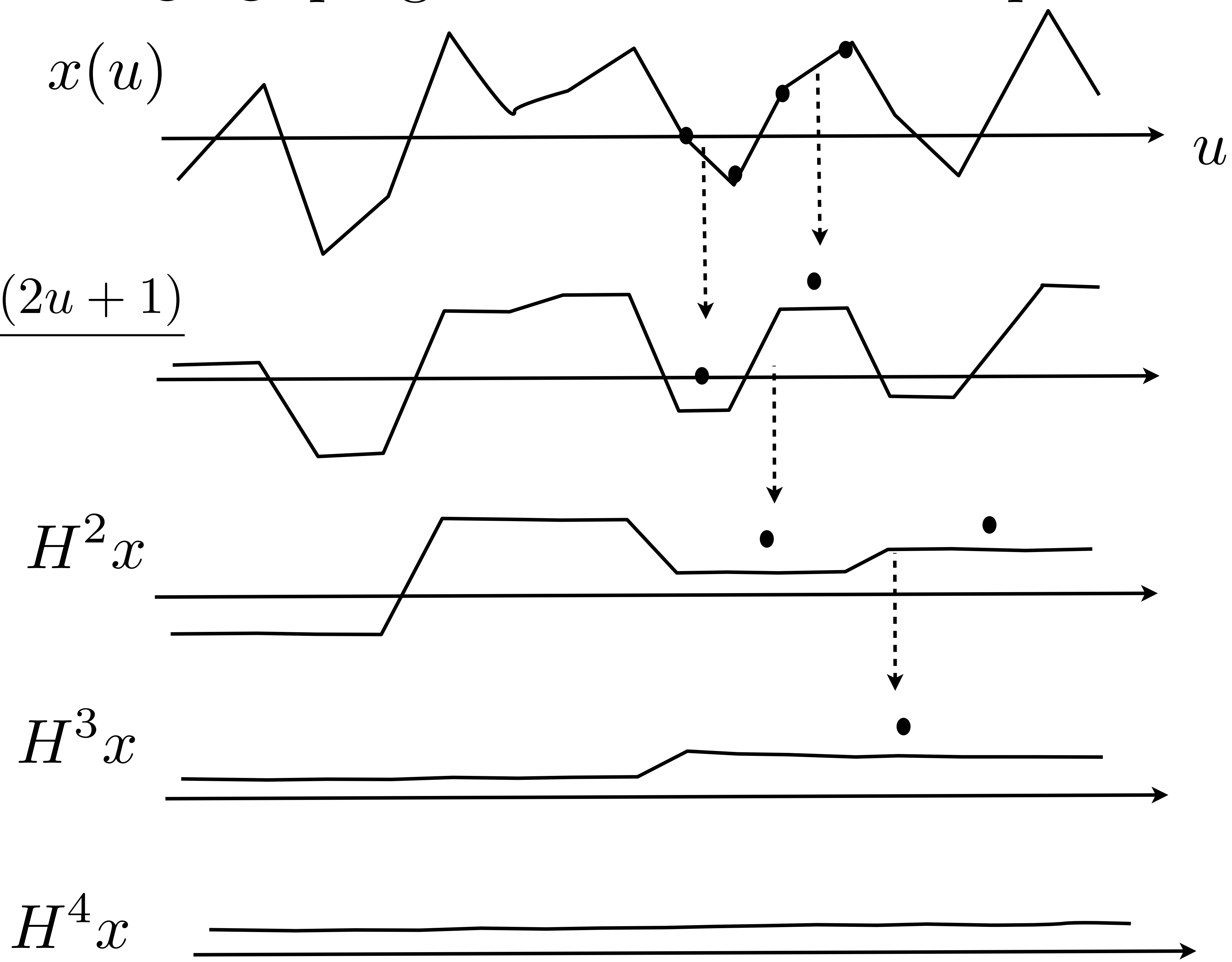


- How to choose the W_k ?

Hierarchical Averaging

- Linear translation invariance by averaging.
- Hierarchical averaging: progressive invariant computation

$$Hx(u) = \frac{x(2u) + x(2u+1)}{2}$$



Haar Filtering

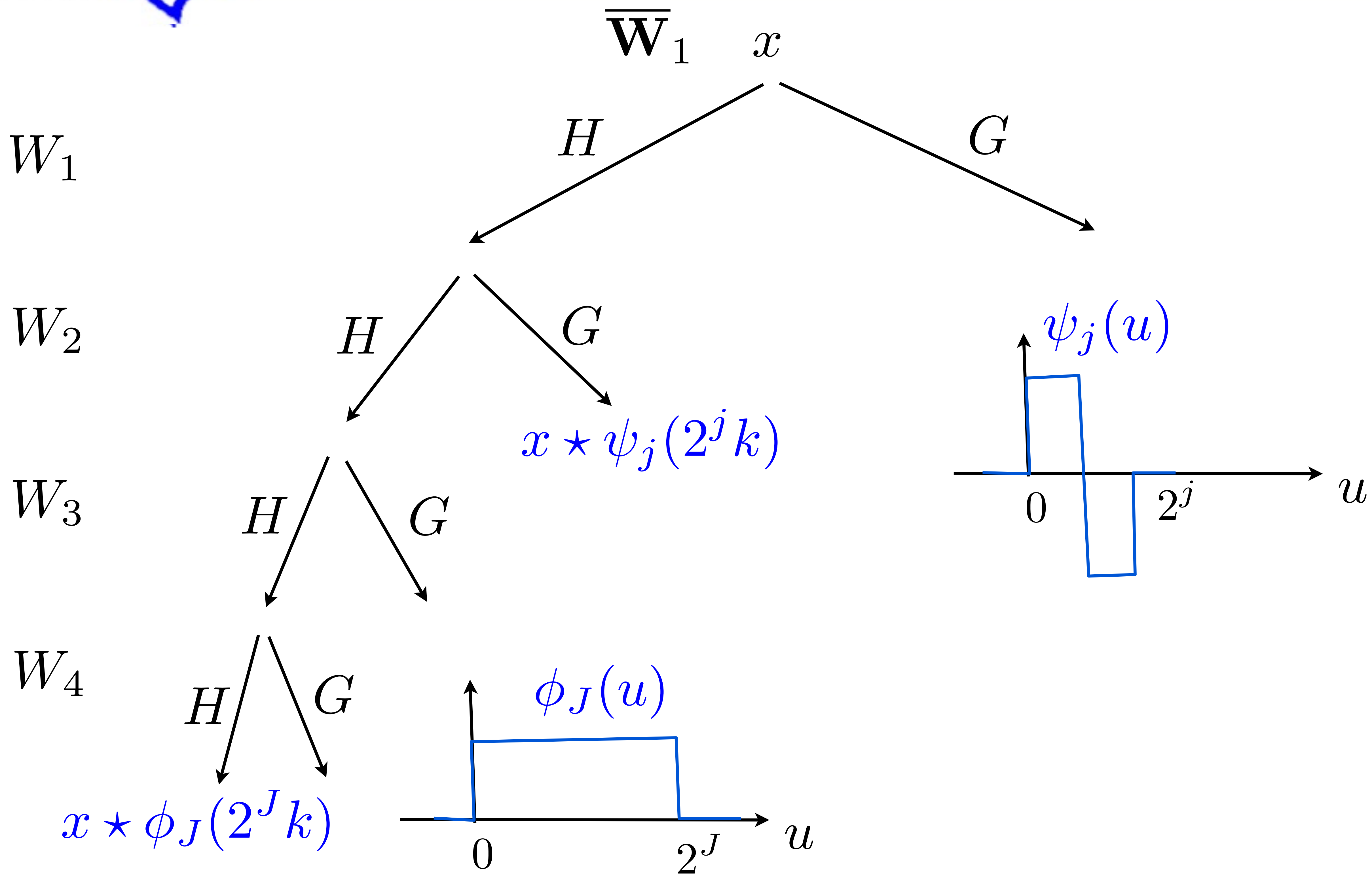
$$\{x(u)\}_{u \leq d}$$

$$H \swarrow \qquad \searrow G$$

$$\left\{ \frac{x(2u) + x(2u+1)}{\sqrt{2}} \right\}_{u \leq d/2} \qquad \left\{ \frac{x(2u) - x(2u+1)}{\sqrt{2}} \right\}_{u \leq d/2}$$

$W = (H, G)$ is an orthogonal operator

Haar Wavelet Basis



$$\left\{ \phi_J(u - 2^J k) \;,\quad \psi_j(u - 2^j k) \right\}_{j > J, k} \quad \text{orthonormal basis of } \mathbf{L^2}[0,1]$$

Haar Filtering

$$
\begin{tikzcd}
 & \{x(u)\}_{u\leq d} \arrow[dl, "H"'] \arrow[dr, "G"] & \\
\left\{\frac{x(2u)+x(2u+1)}{\sqrt{2}}\right\}_{u\leq d/2} & & \left\{\frac{x(2u)-x(2u+1)}{\sqrt{2}}\right\}_{u\leq d/2}
\end{tikzcd}
$$

$W=(H,G)$ is an orthogonal operator

Haar Modulus

$$\{x(u)\}_{u \le d}$$

$$\left\{ \frac{x(2u) + x(2u+1)}{\sqrt{2}} \right\}_{u \le d/2} \xleftarrow{\;H\;} \{x(u)\}_{u \le d} \xrightarrow{\;\rho G = |G|\;} \left\{ \frac{|x(2u) - x(2u+1)|}{\sqrt{2}} \right\}_{u \le d/2}$$

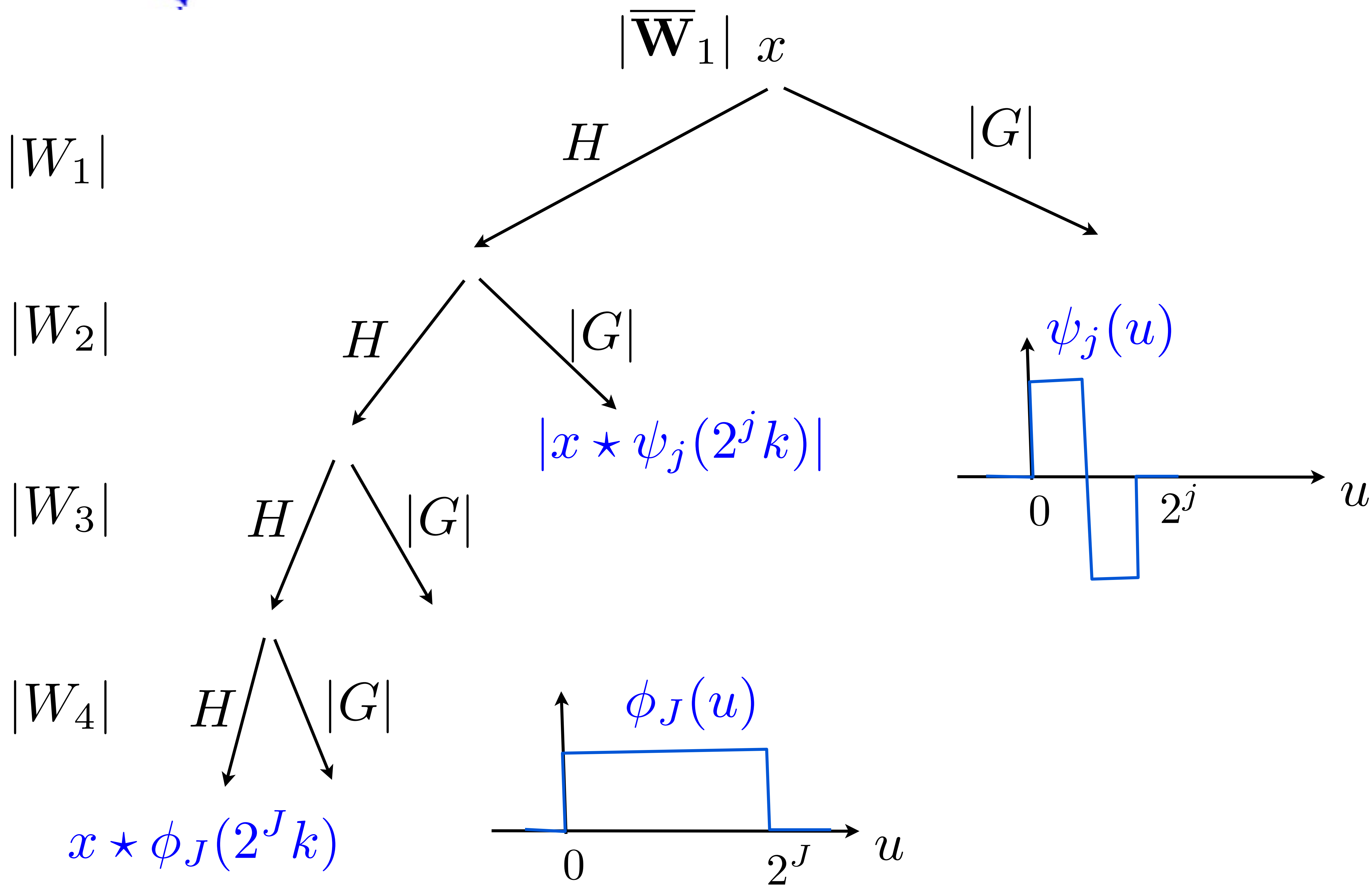
$$|W| = (H, |G|) \text{ is contracting}$$

$$\Big(\frac{a+b}{\sqrt{2}}, \frac{|a-b|}{\sqrt{2}} \Big) : \text{ permutation invariant of } (a,b)$$

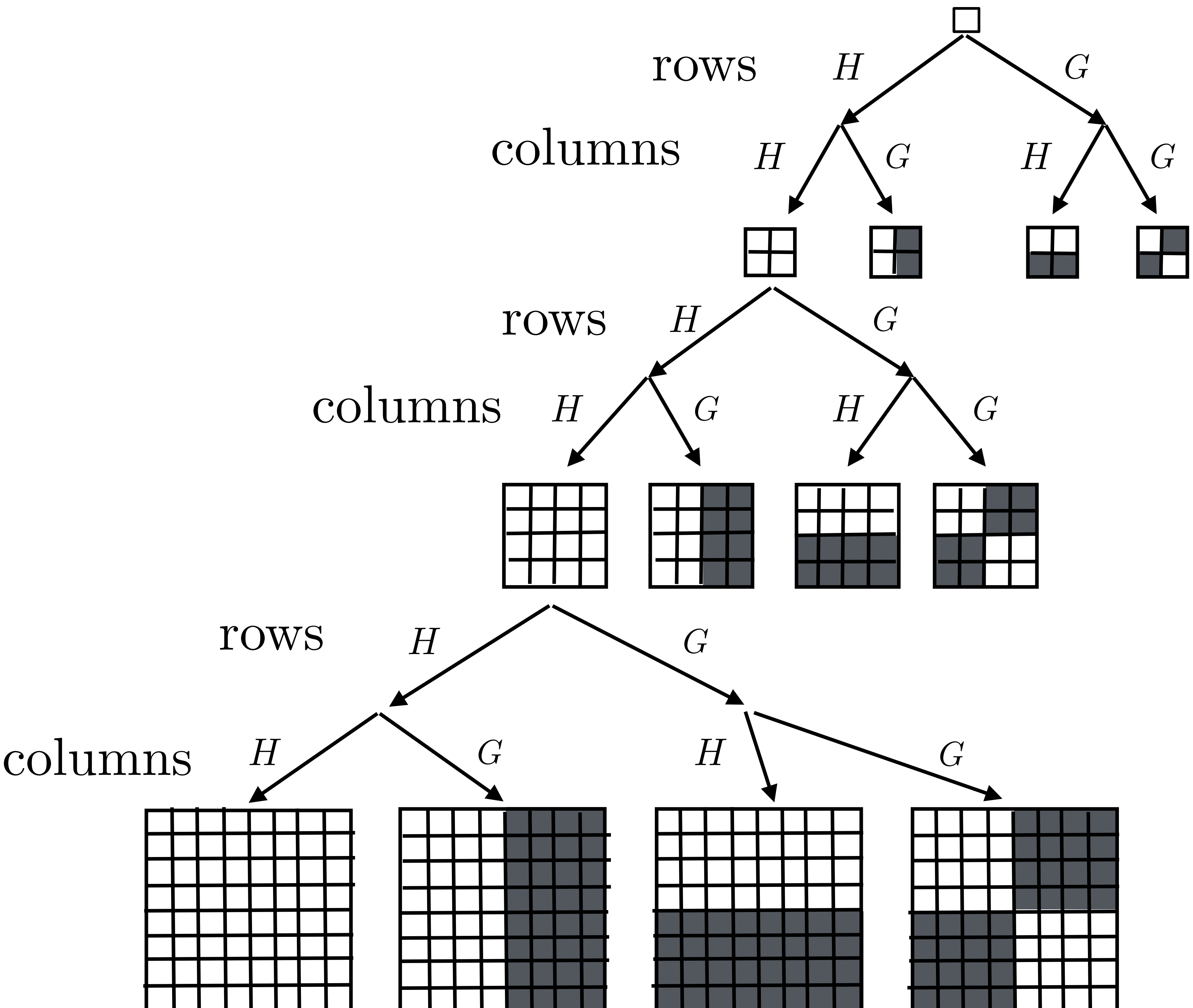
$$\max(a,b) = \frac{a+b}{2} + \frac{|a-b|}{2}$$

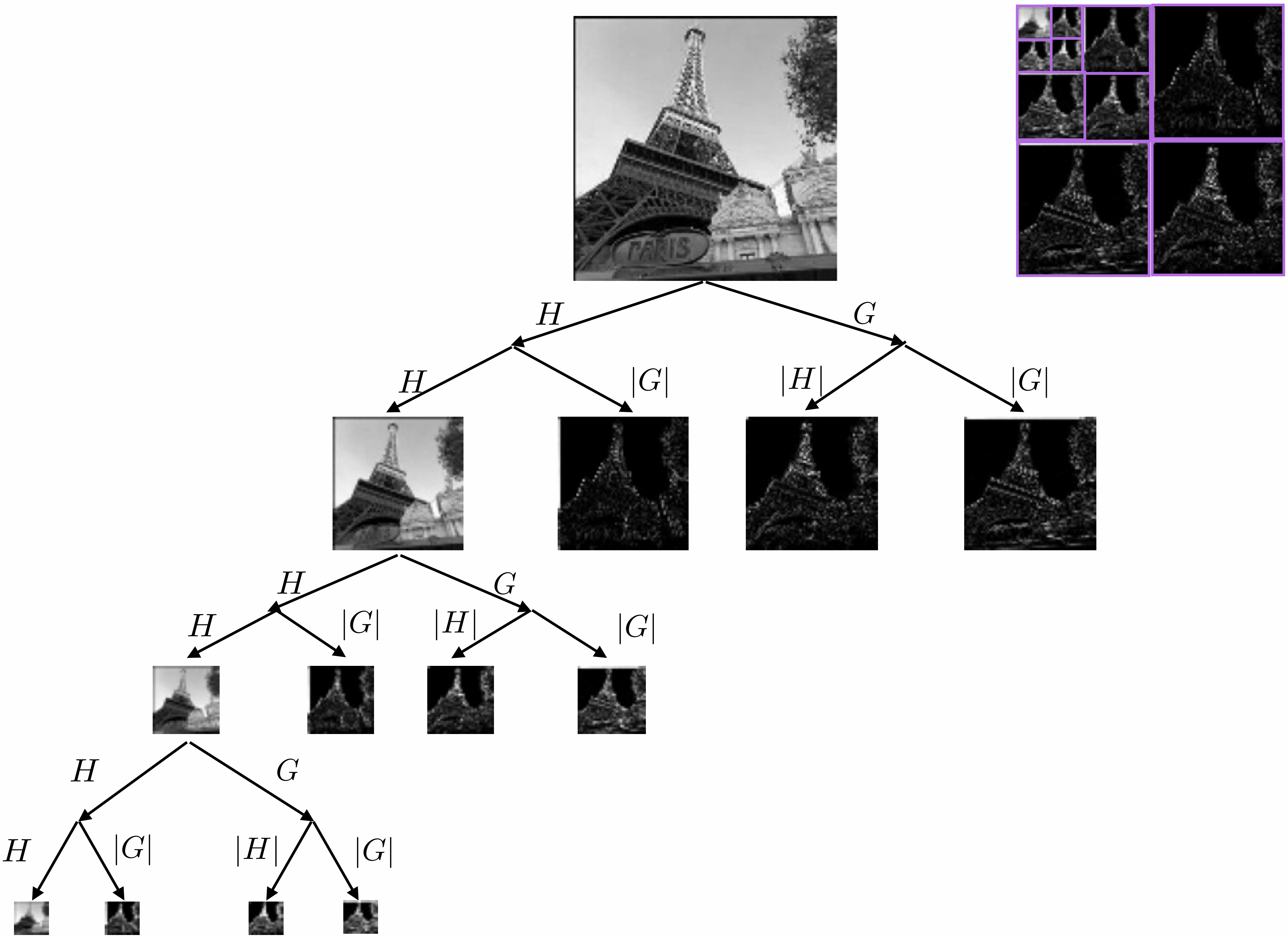
$$\min(a,b) = \frac{a+b}{2} - \frac{|a-b|}{2}$$

Haar Wavelet Modulus



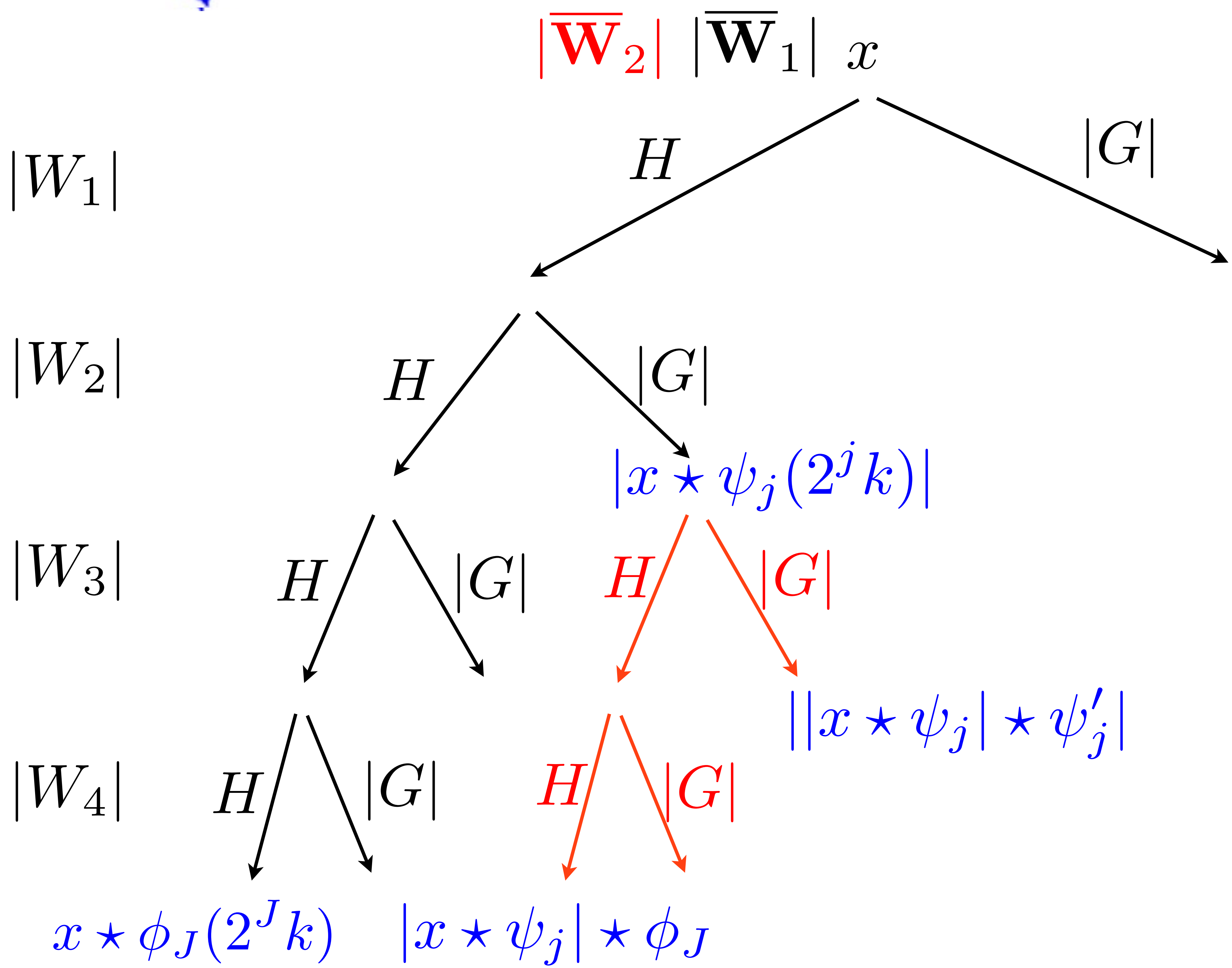
Haar Basis of Images



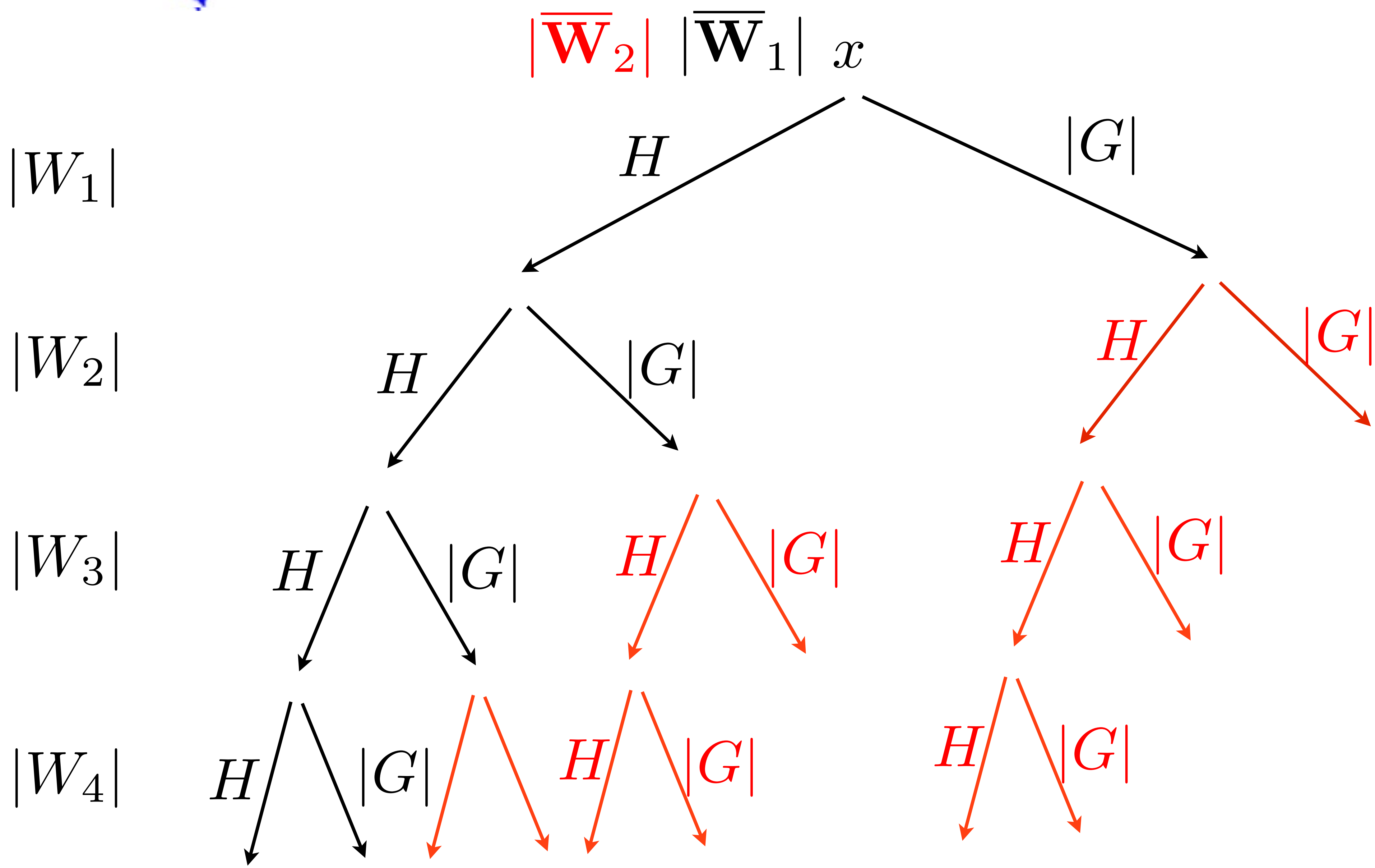


H
G
H
|G|
|H|
|G|
H
G
H
|G|
|H|
|G|
H
G
H
|G|
|H|
|G|

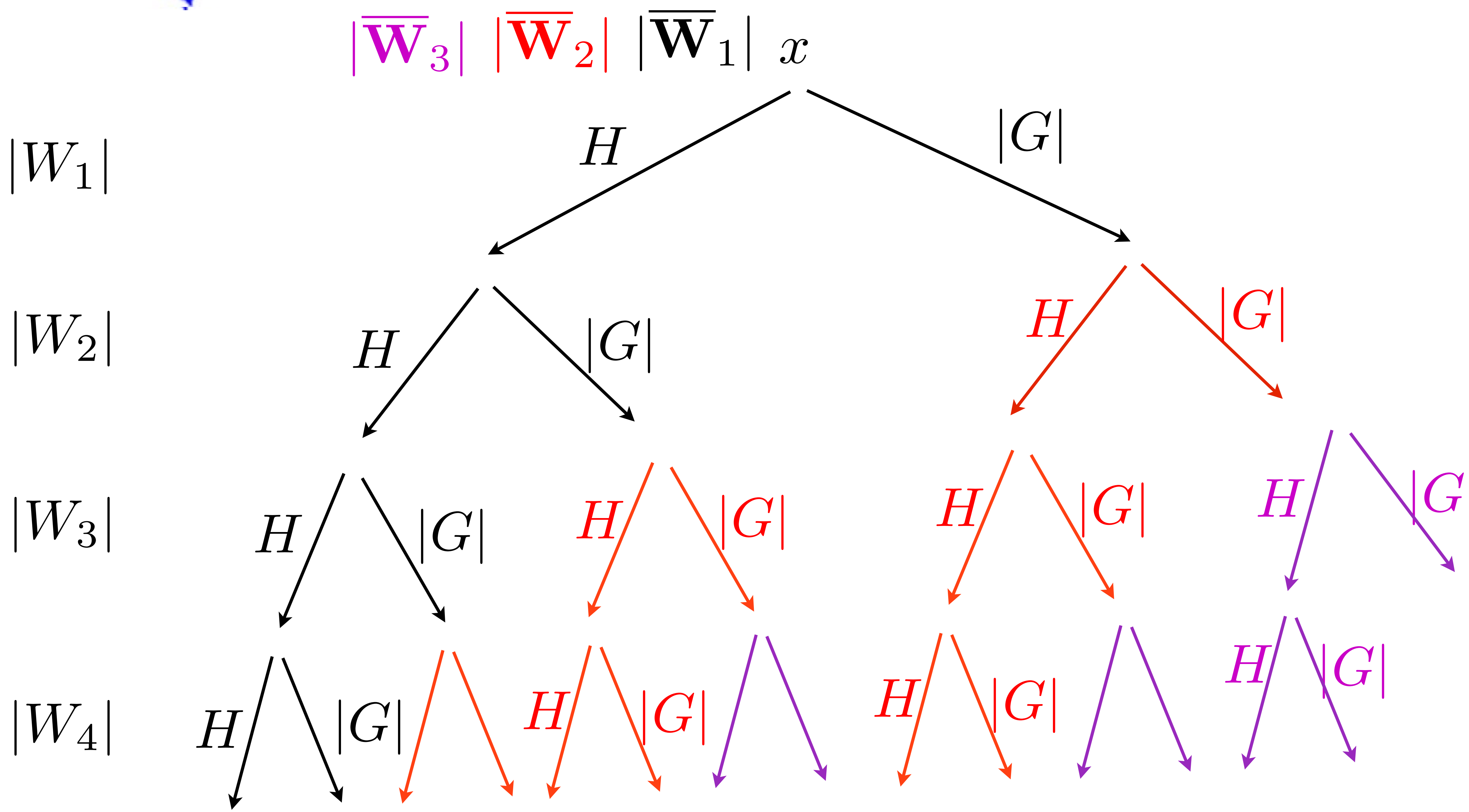
Haar Wavelet Modulus



Haar Wavelet Modulus

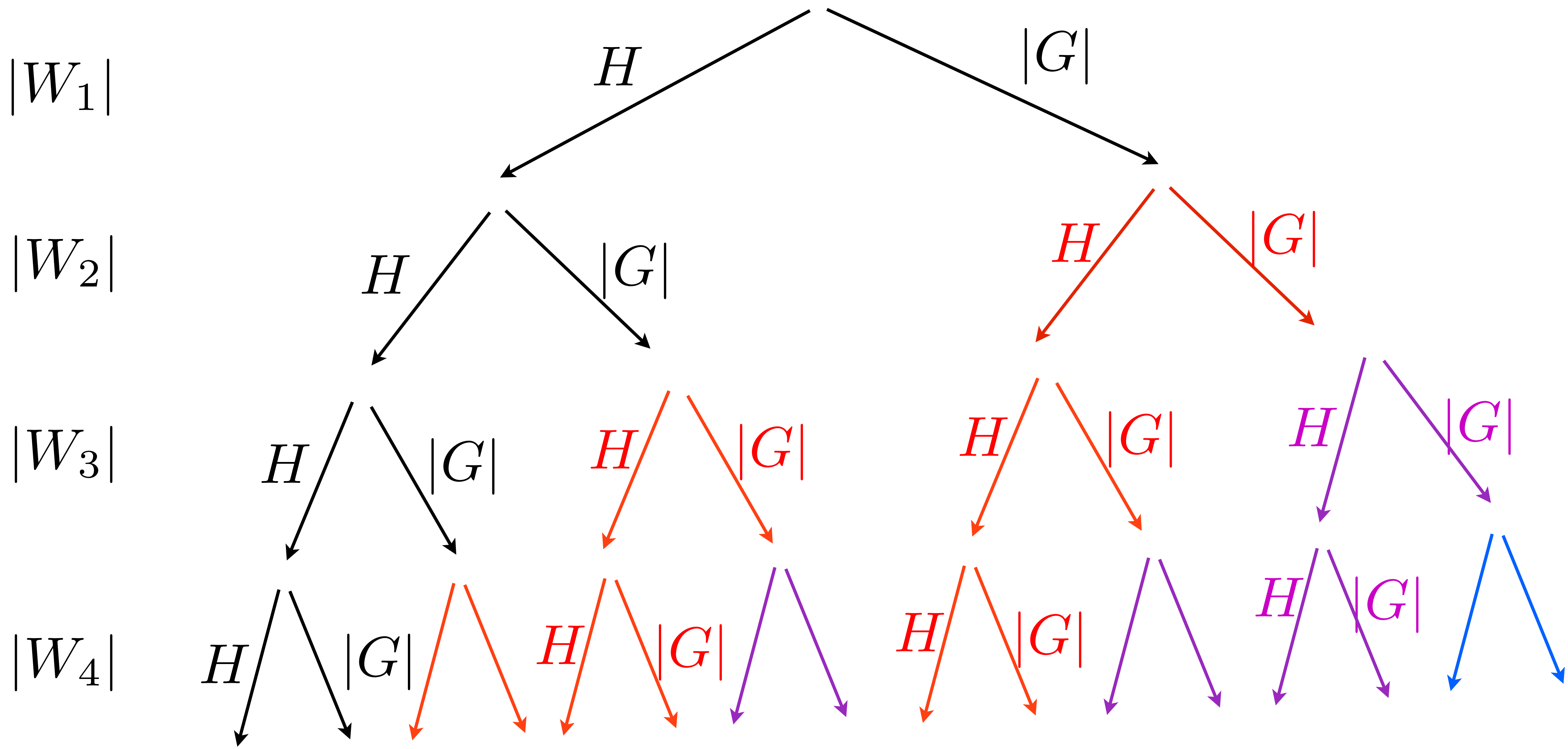


Haar Wavelet Modulus



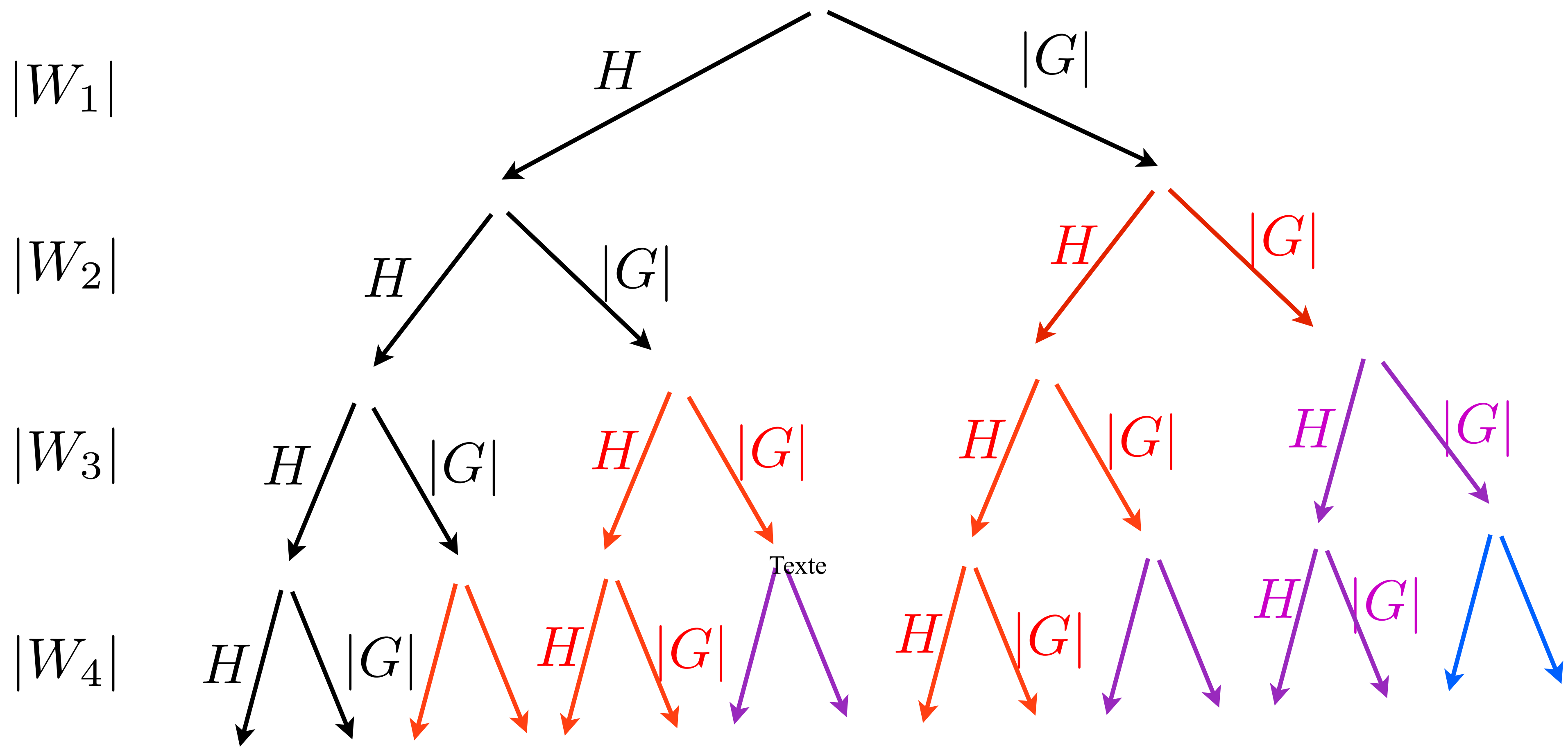
Haar Wavelet Modulus

$$Sx = |\overline{\mathbf{W}}_4|\ |\overline{\mathbf{W}}_3|\ |\overline{\mathbf{W}}_2|\ |\overline{\mathbf{W}}_1|\ x$$



Haar Wavelet Scattering

$$Sx = |\overline{\mathbf{W}}_4|\,|\overline{\mathbf{W}}_3|\,|\overline{\mathbf{W}}_2|\,|\overline{\mathbf{W}}_1|\;x$$



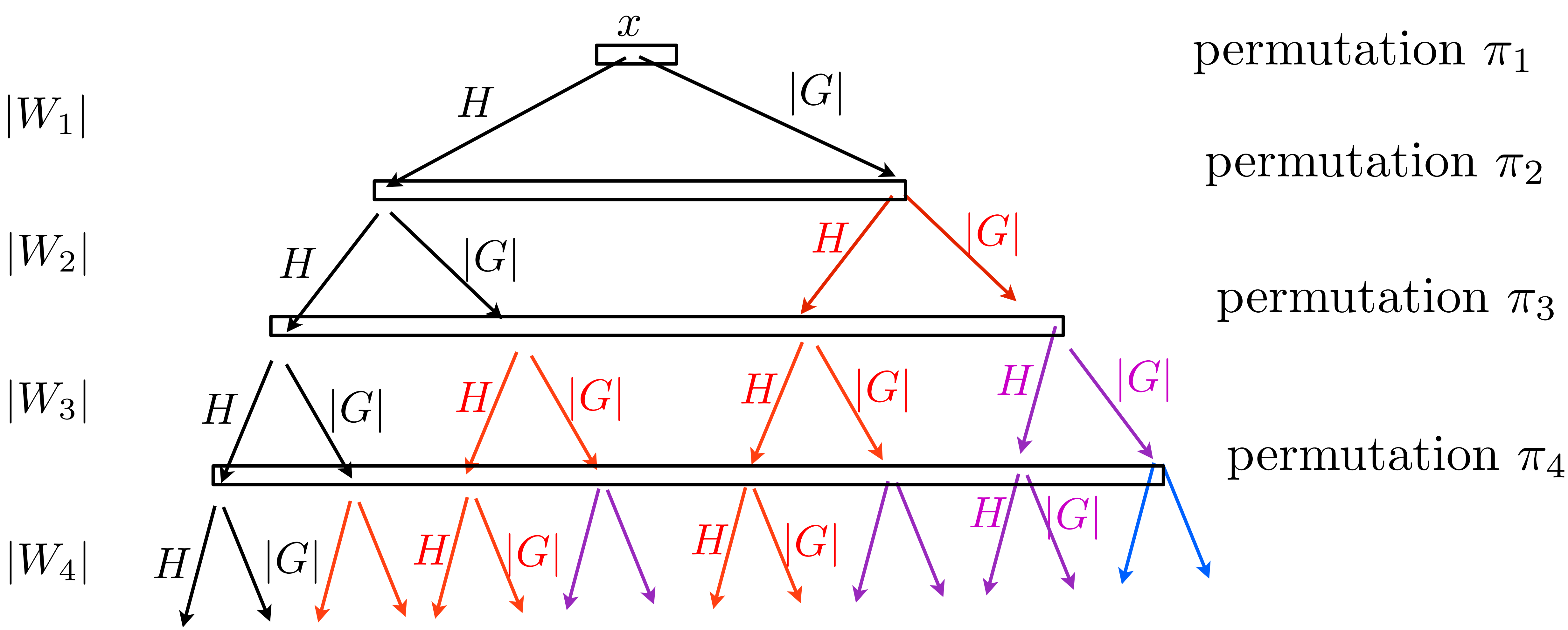
$$S = |W_4|\,|W_3|\,|W_2|\,|W_1| = |\overline{\mathbf{W}}_4|\,|\overline{\mathbf{W}}_3|\,|\overline{\mathbf{W}}_2|\,|\overline{\mathbf{W}}_1|$$

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

Theorem: $\|Sx - Sx'\| \leq \|x - x'\|$ et $\|Sx\| = \|x\|$

$$\|Sx_\tau - Sx\| \leq C\,\|\nabla\tau\|_\infty\,\|x\| \;\;?$$

Wavelet Scattering



$$S = |W_4|\,|W_3|\,|W_2|\,|W_1|$$

- What is happening if (H, G) are changed ? different wavelets.
- How to change the invariant ? change convolutions

with permutations $\quad S = \pi_4\,|W_4|\,\pi_3\,|W_3|\,\pi_2\,|W_2|\,\pi_1\,|W_1|$

Learning with Haar

Xu Chen, Xiu Cheng

- Learning convolutions/permutations reduces to pairing.
- Haar filtering of coefficient pairs:

$$\Big\{x(k)\Big\}_{k\le d} \overset{\text{pairing}}{\longrightarrow} \Big\{x(k), x(k')\Big\}_{(k,k')} \overset{W}{\longrightarrow} \Big\{\frac{x(k)+x(k')}{\sqrt{2}} \; , \; \frac{x(k)-x(k')}{\sqrt{2}}\Big\}_{(k,k')}$$

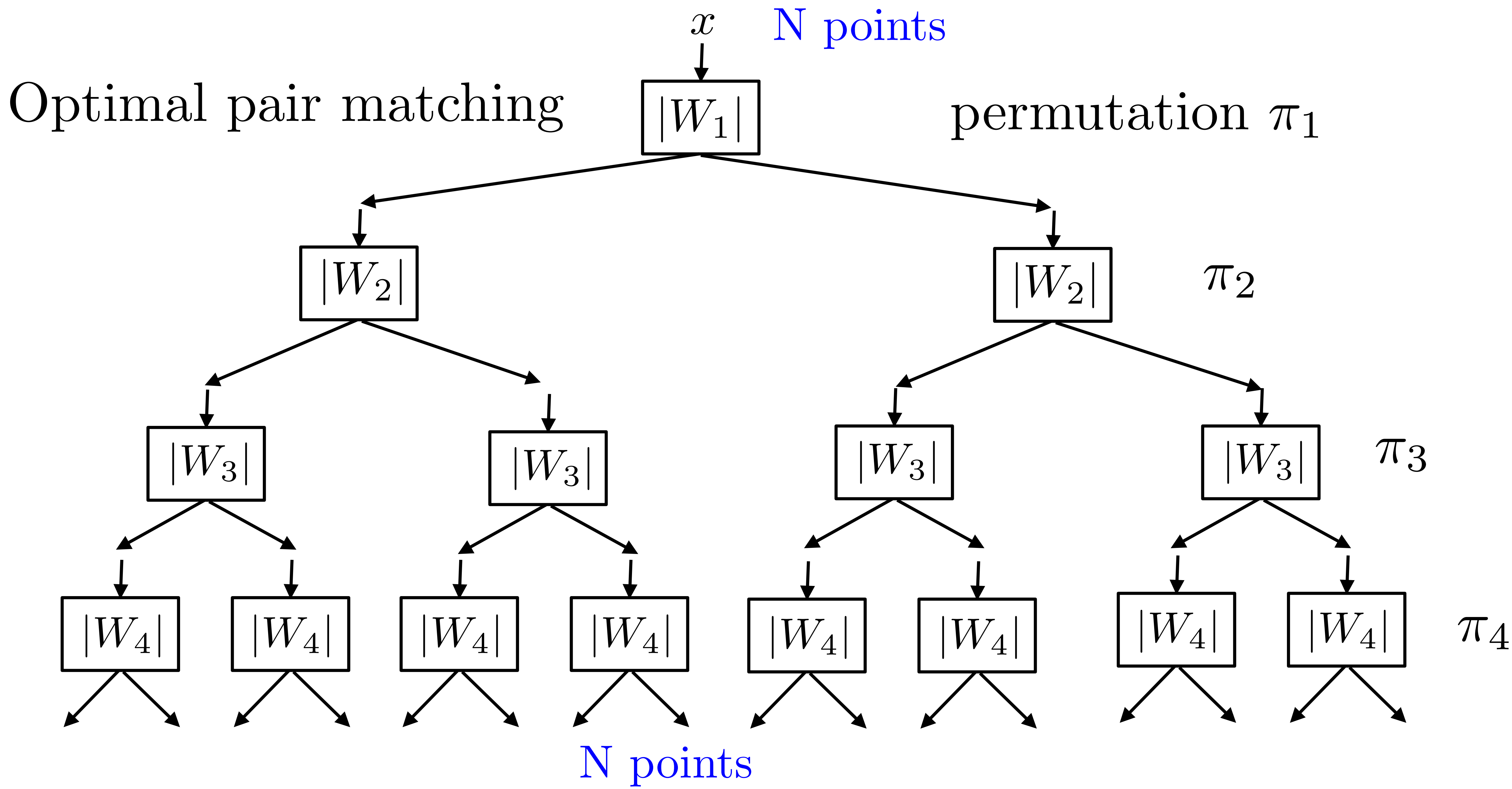
- Permutation invariant contraction:

$$\Big\{x(k)\Big\}_{k\le d} \overset{\text{pairing}}{\longrightarrow} \Big\{x(k), x(k')\Big\}_{(k,k')} \overset{|W|}{\longrightarrow} \Big\{\frac{x(k)+x(k')}{\sqrt{2}} \; , \; \frac{|x(k)-x(k')|}{\sqrt{2}}\Big\}_{(k,k')}$$

- Learn pairing $\{(k,k')\}$: low-dimensional problem (no curse) with optimal matching algorithms.

Learned Haar Scattering

The pairing defining each W_m is learned from data.

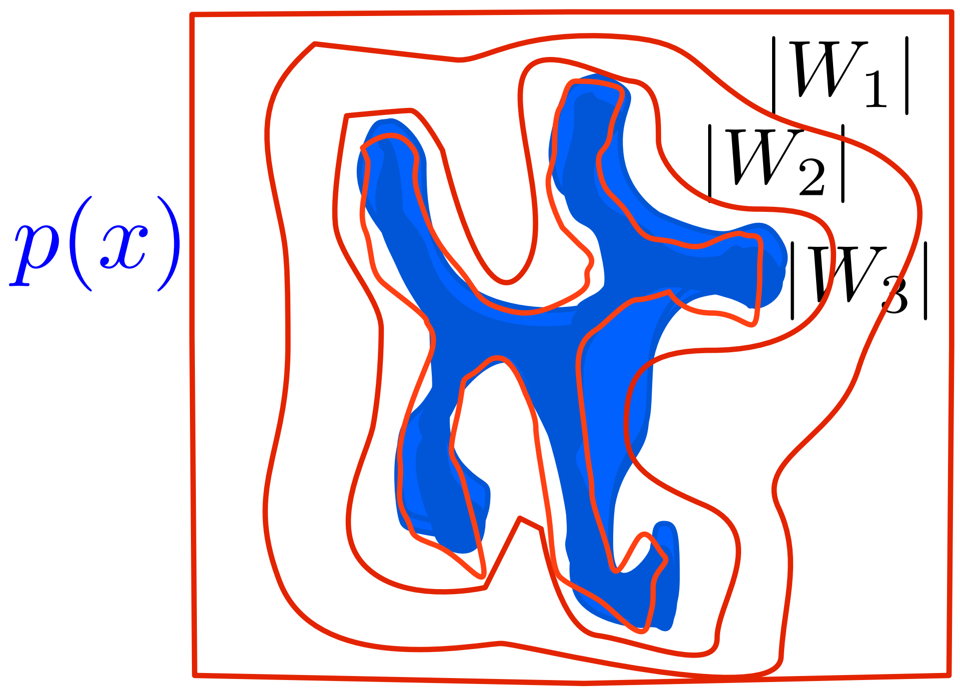


Learned Haar Scattering: $Sx = \prod_{k=1}^{4} |W_k| \pi_k x$

Multiple trees: increases dimension

Unsupervised Space Contraction

- Learn $S = \prod_m |W_m|$
- **Unsupervised**: minimise the data volume reduction



- Pair matching algorithm finds the pairing which maximizes the data volume: minimises a mixed $\mathbf{l^2/l^1}$ sparsity norm.
- Sparsity minimises contraction:

$$||a| - |b|| = |a - b| \text{ if } a = 0 \text{ or } b = 0.$$

Learning with Optimal Contraction

- We want to maximize

$$\sigma^2(SX) = \sigma^2(\prod_{m=1}^{J} |W_m| X).$$

- Greedy: for increasing m finds W_m which maximizes

$$\sigma^2(|W_m| X_{m-1}) \quad \text{with} \quad X_{m-1} = \prod_{k=1}^{m-1} |W_k| X$$

$$\sigma^2(|W_m| X_{m-1}) = \mathbb{E}(\| |W_m| X_{m-1} \|^2) - \| \mathbb{E}(|W_m| X_{m-1}) \|^2$$

$\Rightarrow$ find a grouping which minimizes $\| \mathbb{E}(|W_m| X_{m-1}) \|^2$

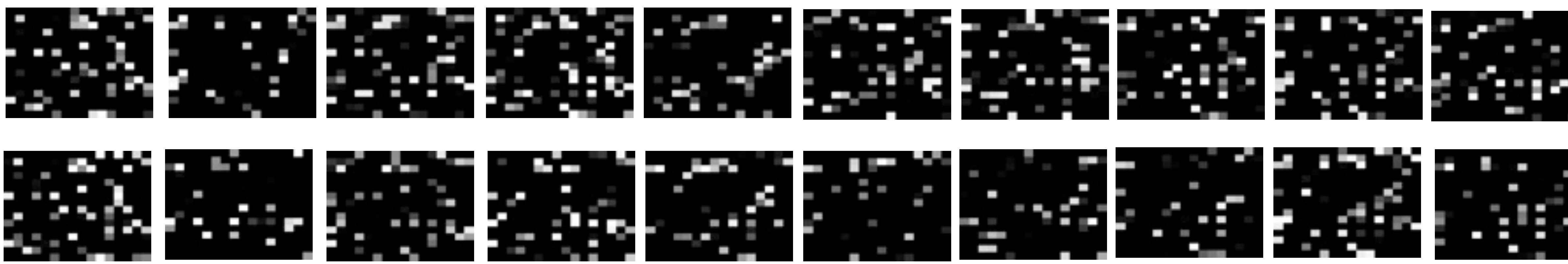
sparsity $\mathbf{l^2/l^1}$ norm: build discriminative features which are sparsely activated across realizations.

Digit Image Classification

Xu Chen, Xiu Cheng

Examples of MNIST written digits

Permutation of digit image pixels:



Unsupervised learning W_m for $1 \leq m \leq 4$ yields Haar wavelets of size 2^4

Reordered Haar pairing: 100% connected for first 3 levels $m = 1, 2, 3$

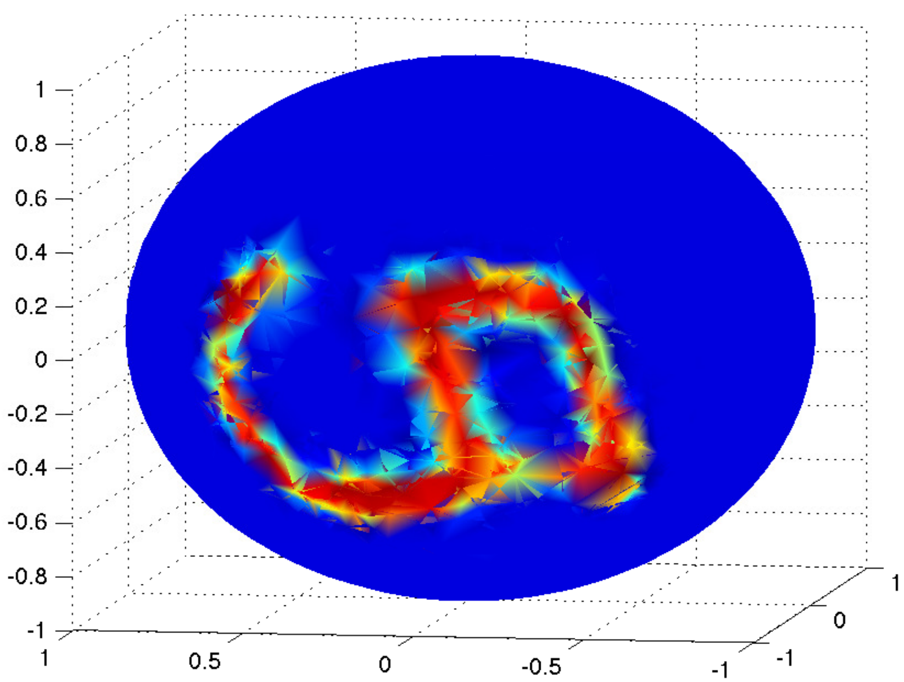
85% for $m = 4$ and 65% for $m = 5$

Learned Haar Scattering : $\mathbf{0.9}\%$ errors

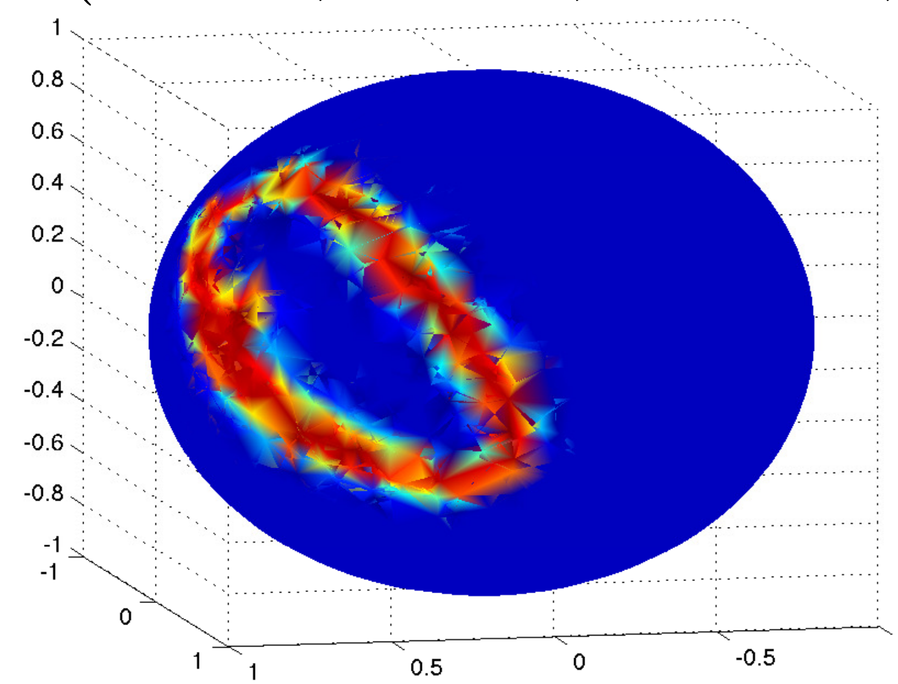
Haar with Rotations

- Rotate MNIST digits on a $3D$ sphere

(Bruna, Szlam, Zaremb, LeCun)



(a)



(b)

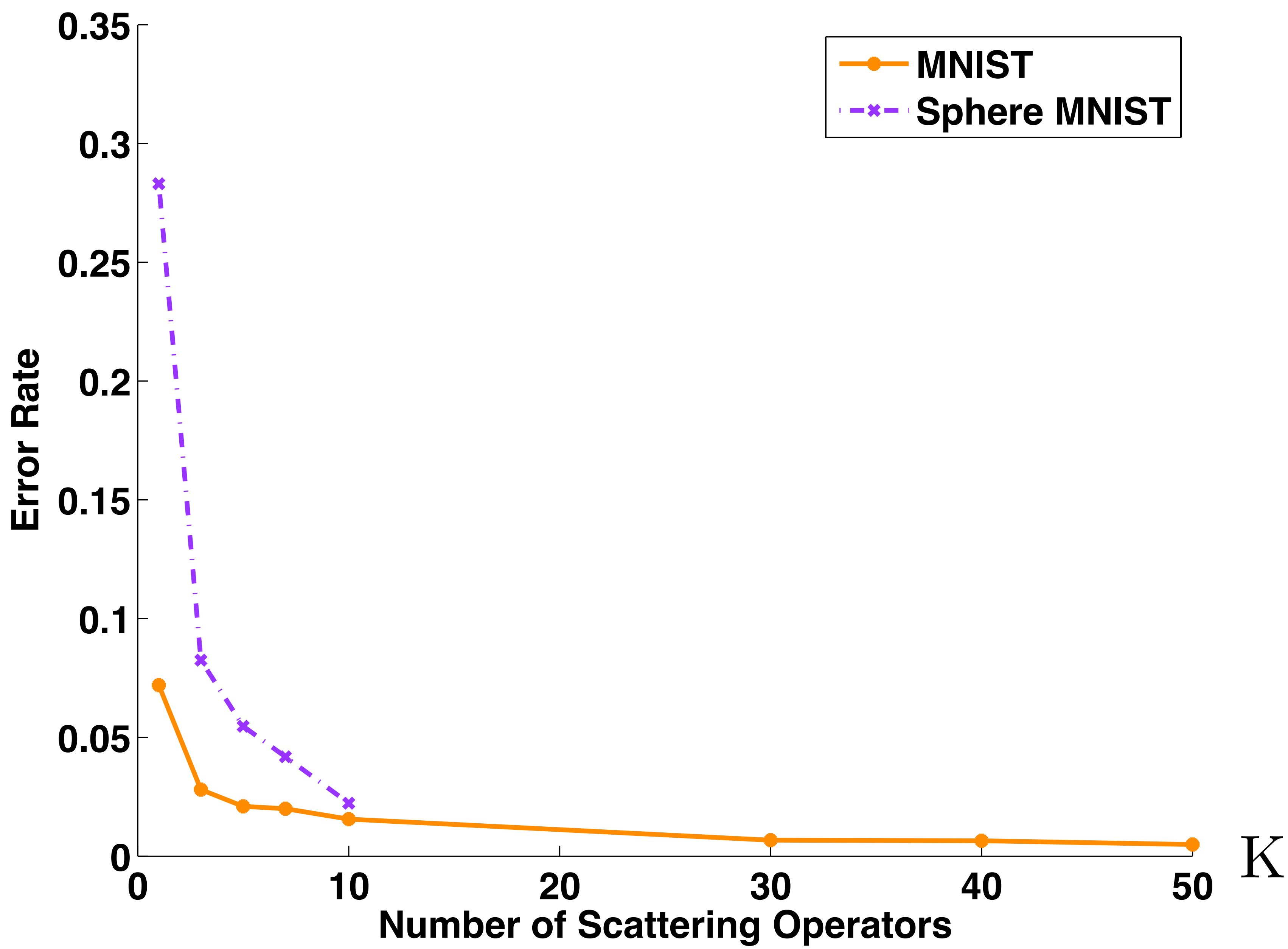
- Haar scattering: does not know translations and rotations

Nearest neighbor	Fully connect. 2 layers	Local connect. 2 layers	Learned Haar Scattering
19%	5.6%	6%	**2.2** %

Bagging Scattering Vectors

- The training set $\{x_i, f(x_i)\}_i$ is divided in K groups example
  A Haar scattering representation $S_k x$ is learned from each g

  The aggregation $Sx = \{S_k x\}_{k \leq K}$ is a vector size KN.



Conclusion

- Éfficacité remarquable des réseaux de neurones profonds:
  - Séparation d'échelles par ondelettes: scattering
  - Métriques invariantes et stables par difféomorphismes
  - Modèles de processus stationaires intermittents

- Apprentissage non-supervisé par contractions itérées.

- Grand potentiel à l'interface traitement du signal/apprentissage.

Papiers et Softwares Matlab:

www.di.ens.fr/data/scattering