Deep Neural Network

Mathematical Mysteries

for High Dimensional Learning

Stéphane Mallat

École Normale Supérieure
www.di.ens.fr/data

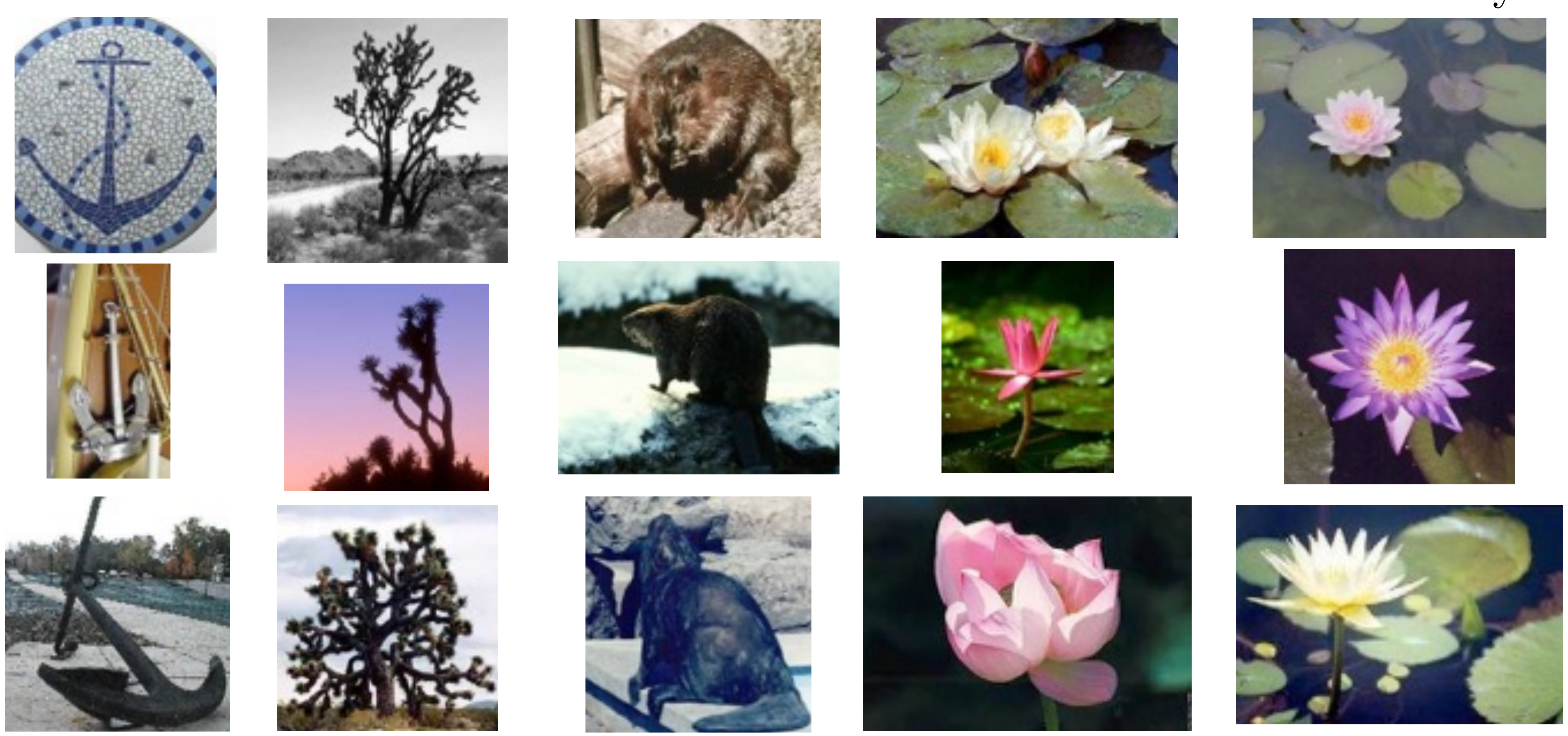
High Dimensional Learning

- High-dimensional $x = (x(1), ..., x(d)) \in \mathbb{R}^d$:

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

Image Classification $\quad d = 10^6$

Anchor | Joshua Tree | Beaver | Lotus | Water Lily



Huge variability inside classes

Find invariants

High Dimensional Learning

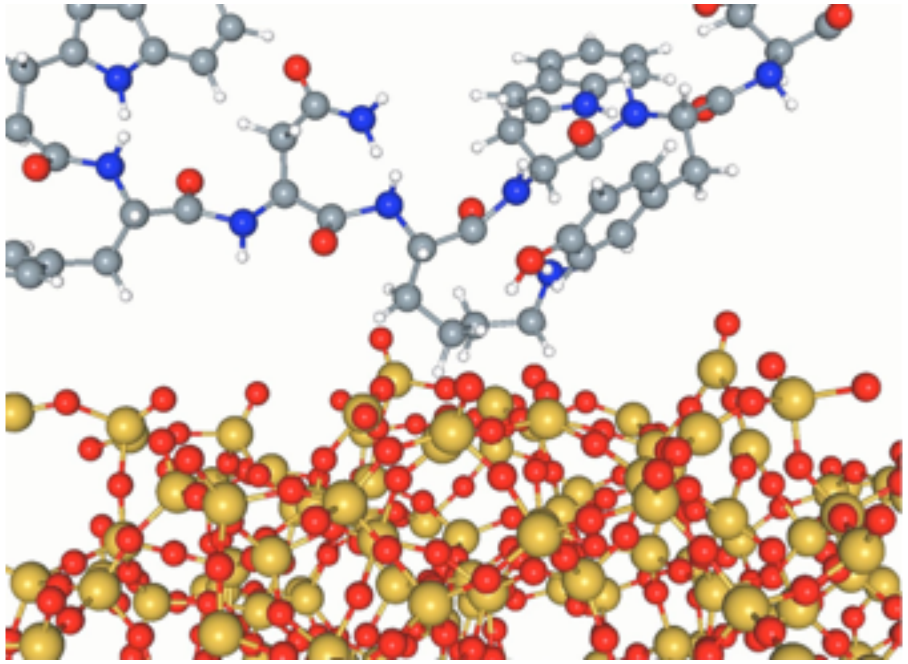
- High-dimensional $x = (x(1), ..., x(d)) \in \mathbb{R}^d$:

- **Regression:** approximate a *functional* $f(x)$
  given n sample values $\{x_i\ ,\ y_i = f(x_i) \in \mathbb{R}\}_{i \leq n}$

  Physics: energy $f(x)$ of a state vector x

Astronomy

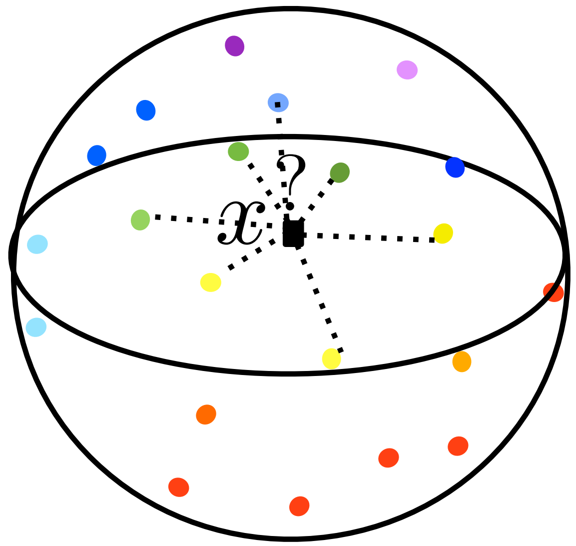
Quantum Chemistry

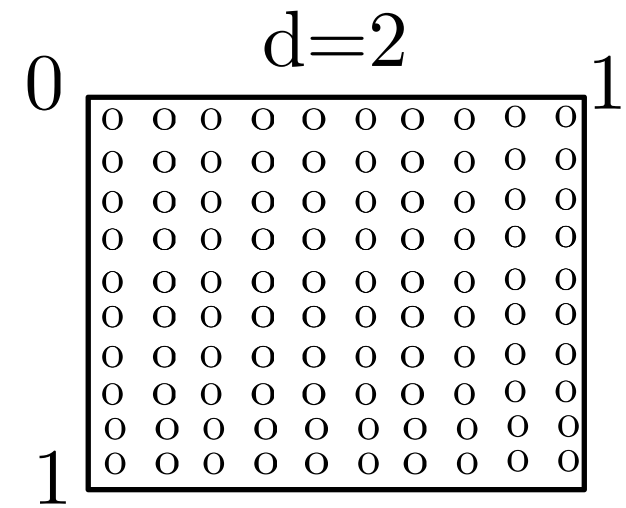


Importance of symmetries.

Curse of Dimensionality

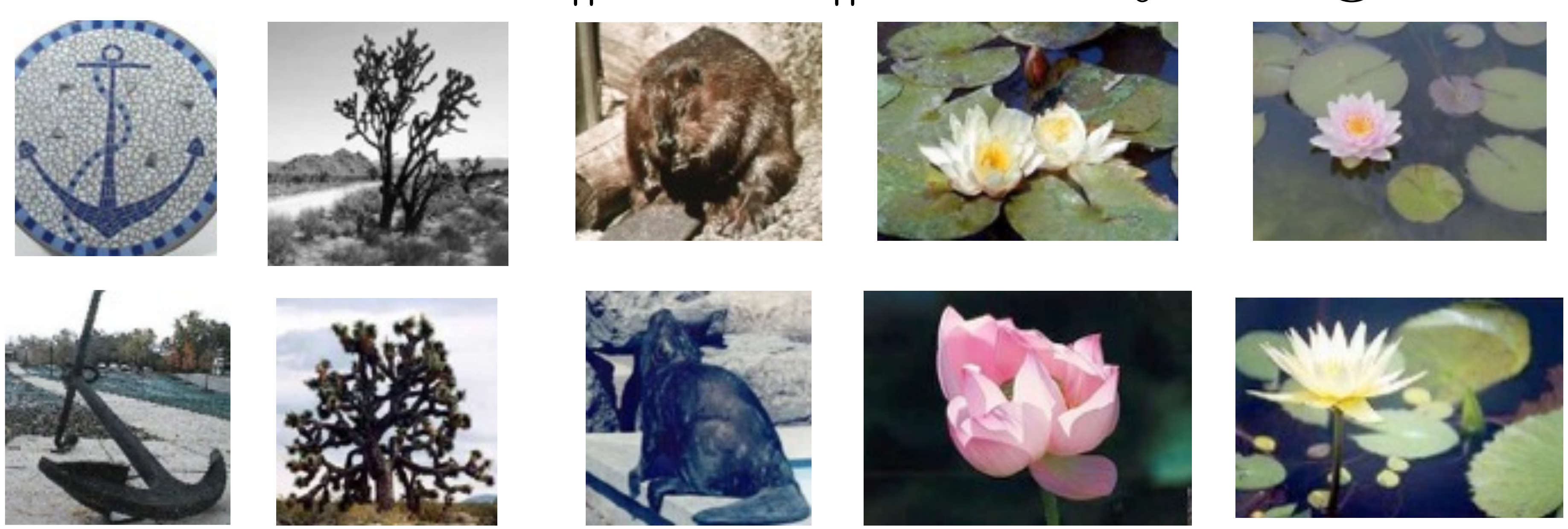
- $f(x)$ can be approximated from examples $\{x_i\,,\,f(x_i)\}_i$ by local interpolation if f is regular and there are close examples:





- Need ϵ^{-d} points to cover $[0,1]^d$ at a Euclidean distance ϵ

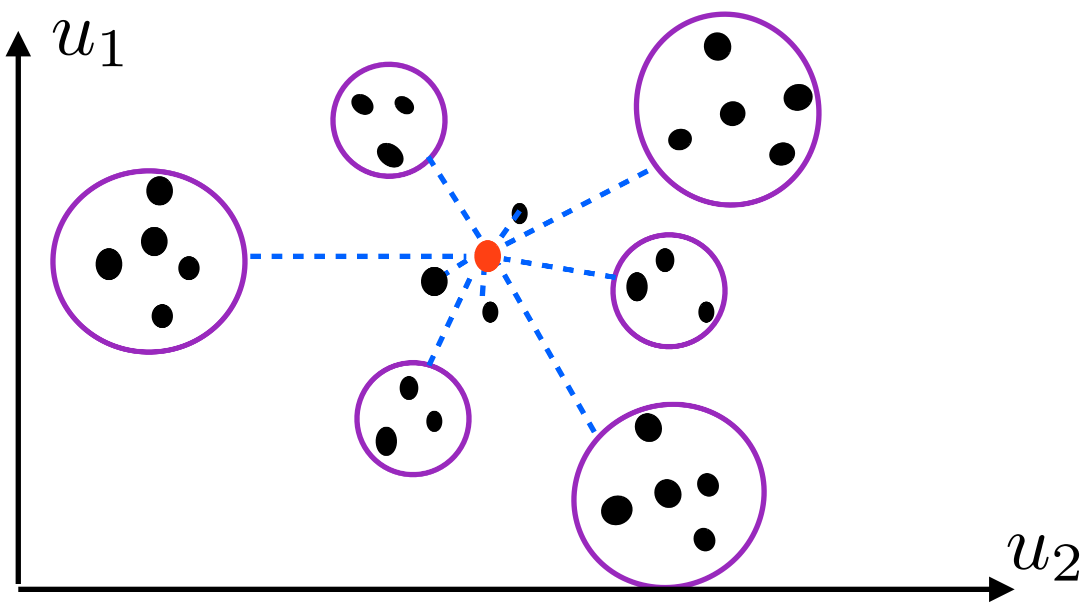
Problem: $\|x - x_i\|$ is always large



Multiscale Separation

- Variables $x(u)$ indexed by a low-dimensional u: time/space... pixels in images, particles in physics, words in text...

- Mutliscale interactions of d variables:



From d^2 interactions to $O(\log^2 d)$ multiscale interactions.

- Multiscale analysis: wavelets on groups of symmetries.
  hierarchical architecture.

Overview

- 1 Hidden Layer Network, Approximation theory and Curse
- Kernel learning
- Dimension reduction with change of variables
- Deep Neural networks and symmetry groups
- Wavelet Scattering transforms
- Applications and many open questions

Understanding Deep Convolutional Networks, arXiv 2016.

Learning as an Approximation

- To estimate $f(x)$ from a sampling $\{x_i\ ,\ y_i = f(x_i)\}_{i \leq M}$

  we must build an M-parameter approximation f_M of f.

- Precise sparse approximation requires some "regularity".

- For binary classification $f(x) = \left\{ \begin{array}{cl} 1 & \text{if } x \in \Omega \\ -1 & \text{if } x \notin \Omega \end{array} \right.$

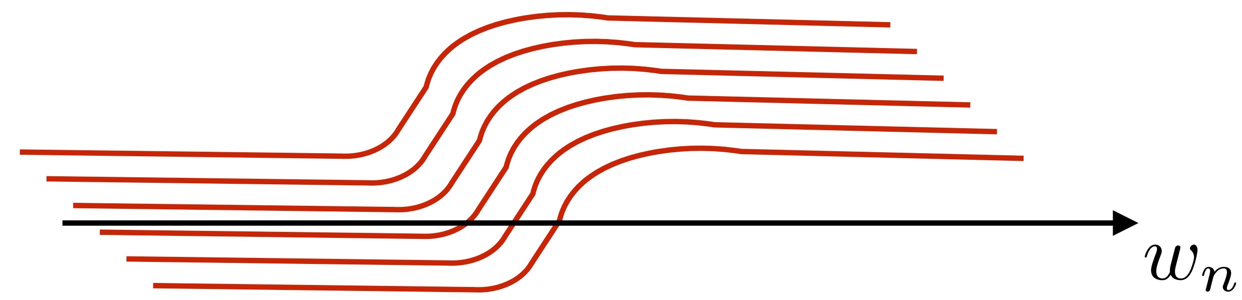
  $$f(x) = \mathrm{sign}(\tilde{f}(x))$$

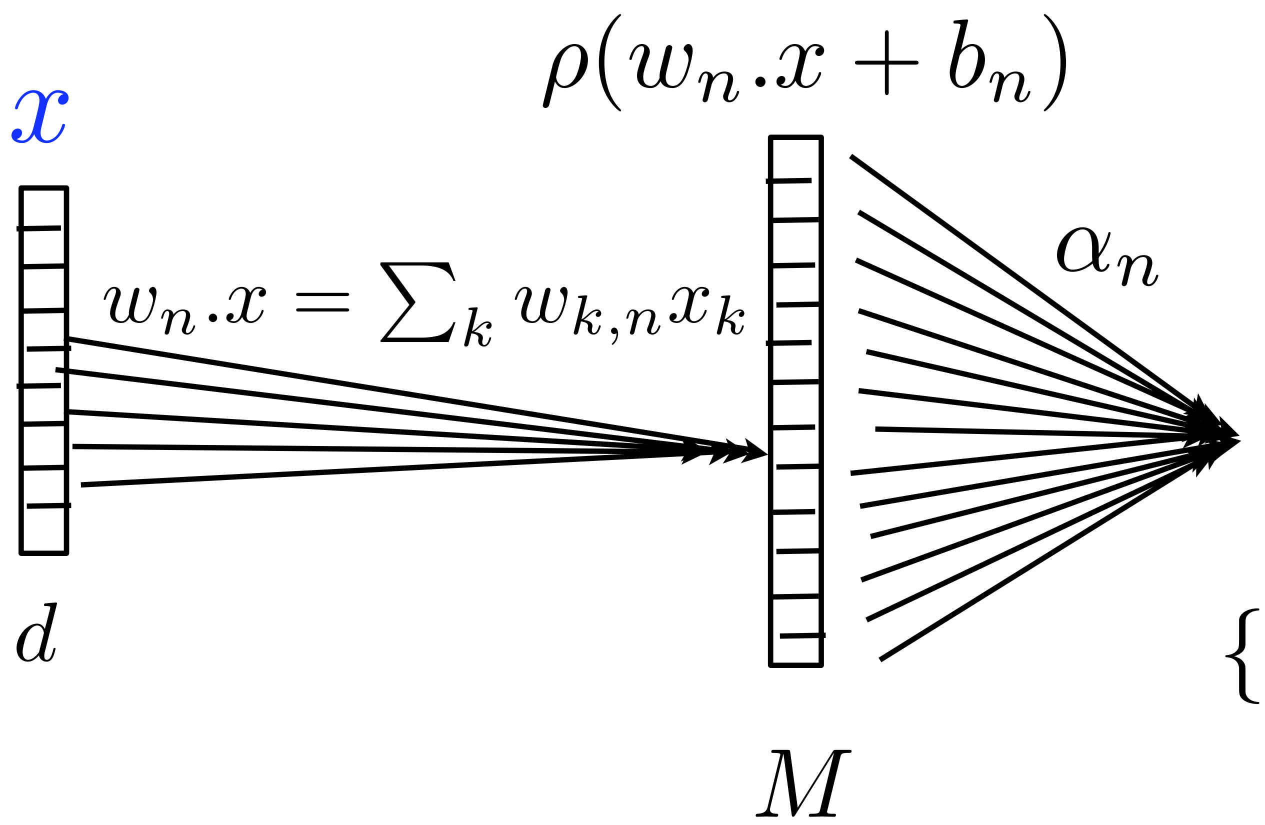
  where $\tilde{f}$ is potentially regular.

- What type of regularity ? How to compute f_M ?

1 Hidden Layer Neural Networks

One-hidden layer neural network: ridge functions $\rho(x.w_n + b_n)$





$$f_M(x) = \sum_{n=1}^{M} \alpha_n \, \rho(w_n.x + b_n)$$

$\{w_{k,k}\}_{k,n}$ and $\{\alpha_n\}_n$ are learned

non-linear approximation.

Cybenko, Hornik, Stinchcombe, White

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

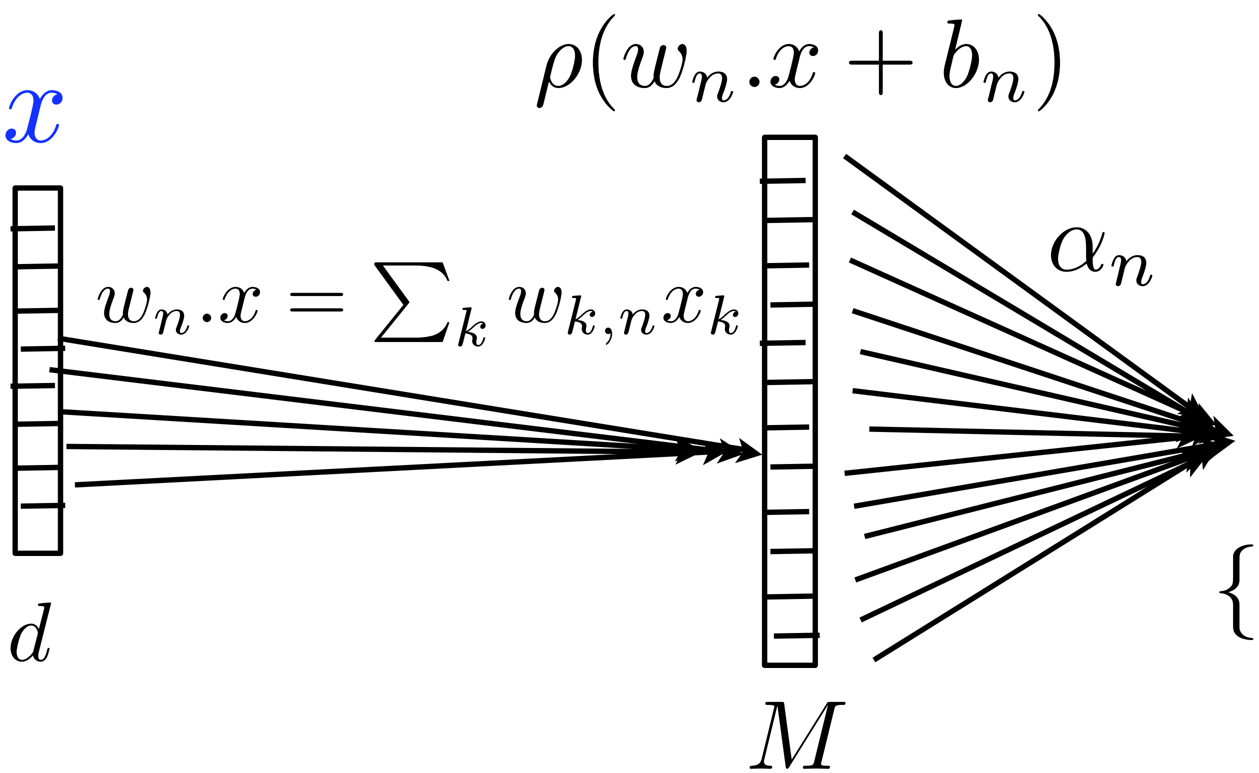
and appropriate choices of $w_{n,k}$ and α_n:

$$\forall f \in \mathbb{L}^2[0,1]^d \qquad \lim_{M \to \infty} \|f - f_M\| = 0 \; .$$

No big deal: curse of dimensionality still there.

1 Hidden Layer Neural Networks

One-hidden layer neural network:



$$f_M(x) = \sum_{n=1}^{M} \alpha_n \, \rho(w_n.x + b_n)$$

$\{w_{k,k}\}_{k,n}$ and $\{\alpha_n\}_n$ are learned non-linear approximation.

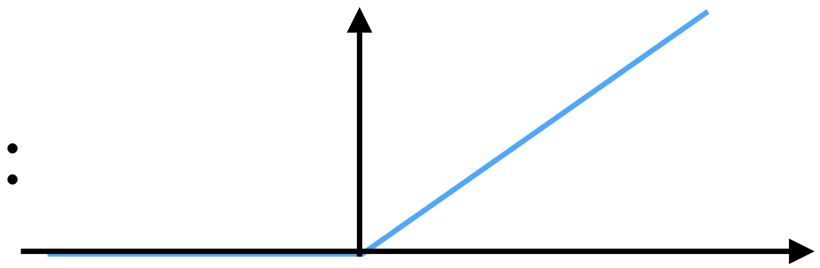
Fourier series: $\rho(u) = e^{iu}$

$$f_M(x) = \sum_{n=1}^{M} \alpha_n \, e^{i w_n . x}$$

For nearly all ρ: essentially same approximation results.

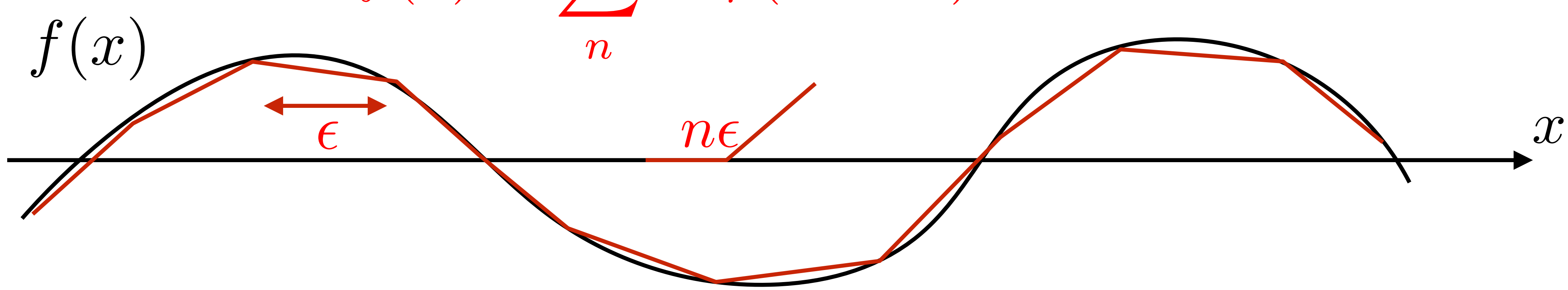
Piecewise Linear Approximation

- Piecewise linear approximation:



$\rho(u) = \max(u, 0)$

$$\tilde{f}(x) = \sum_n a_n \, \rho(x - n\epsilon)$$



If f is Lipschitz: $|f(x) - f(x')| \leq C\,|x - x'|$

$$\Rightarrow \quad |f(x) - \tilde{f}(x)| \leq C\,\epsilon.$$

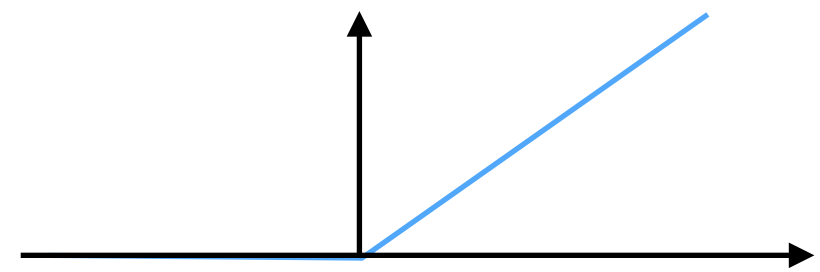
Need $M = \epsilon^{-1}$ points to cover $[0, 1]$ at a distance ϵ

$$\Rightarrow \|f - f_M\| \leq C\,M^{-1}$$

Linear Ridge Approximation

- Piecewise linear ridge approximation: $x \in [0,1]^d$

$$\tilde{f}(x) = \sum_n a_n \, \rho(w_n . x - n\epsilon)$$



$\rho(u) = \max(u, 0)$

If f is Lipschitz: $|f(x) - f(x')| \leq C \, \|x - x'\|$

Sampling at a distance ϵ:

$$\Rightarrow \quad |f(x) - \tilde{f}(x)| \leq C \, \epsilon .$$

need $M = \epsilon^{-d}$ points to cover $[0,1]^d$ at a distance ϵ

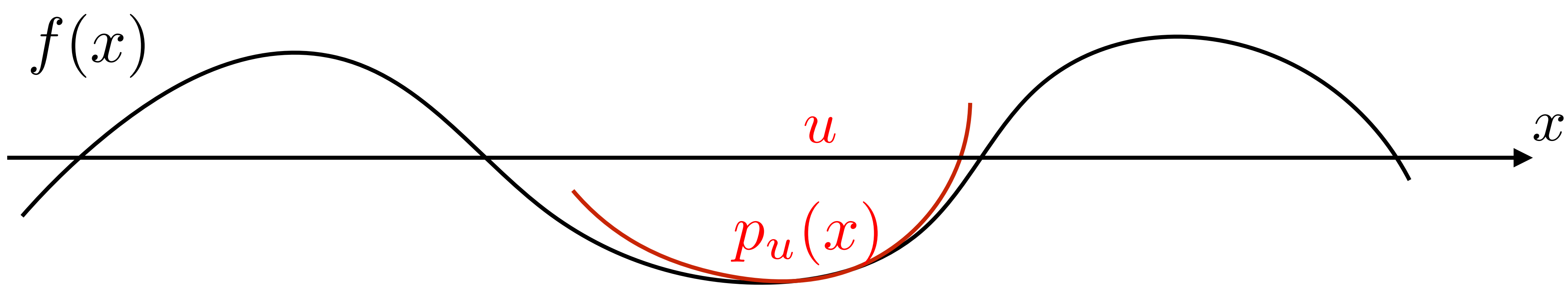
$$\Rightarrow \|f - f_M\| \leq C \, M^{-1/d}$$

Curse of dimensionality!

Approximation with Regularity

- What prior condition makes learning possible ?
- Approximation of regular functions in $\mathbf{C}^s[0,1]^d$:

$$\forall x, u \quad |f(x) - p_u(x)| \leq C\,|x-u|^s \quad \text{with} \quad p_u(x) \quad \text{polynomial}$$



$$|x-u| \leq \epsilon^{1/s} \quad \Rightarrow \quad |f(x) - p_u(x)| \leq C\,\epsilon$$

Need $M^{-d/s}$ point to cover $[0,1]^d$ at a distance $\epsilon^{1/s}$

$$\Rightarrow \|f - f_M\| \leq C\,M^{-s/d}$$

- Can not do better in $\mathbf{C^s}[0,1]^d$, not good because $s \ll d$.

Failure of classical approximation theory.

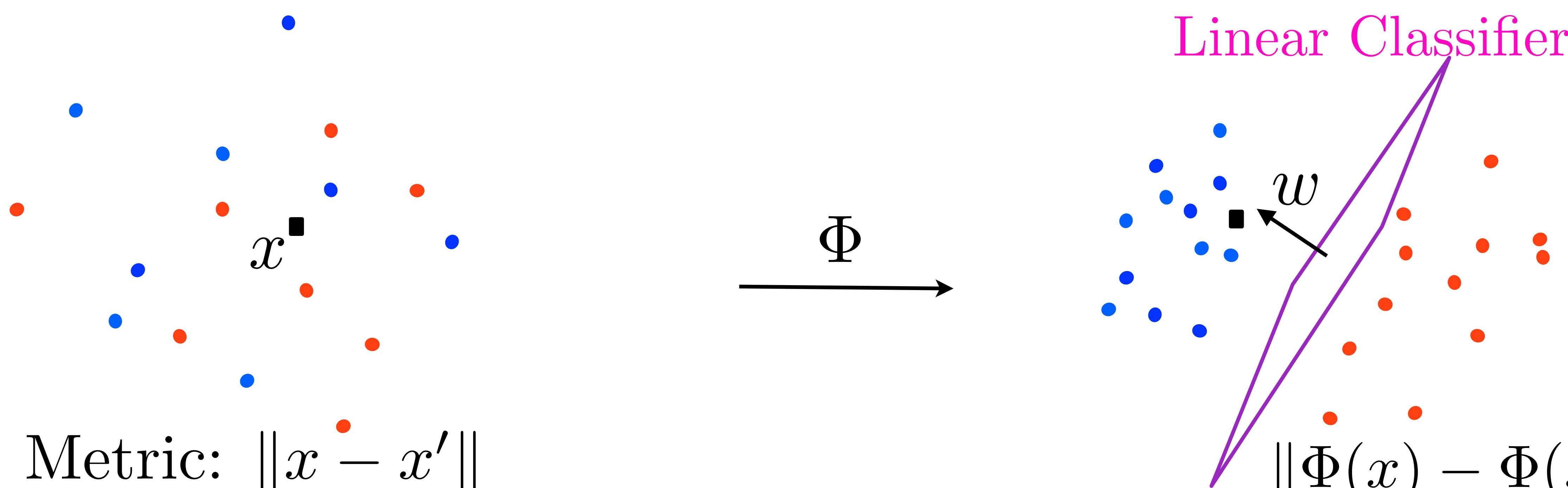
Kernel Learning

Change of variable $\Phi(x) = \{\phi_k(x)\}_{k \leq d'}$

to nearly linearize $f(x)$, which is approximated by:

$$\tilde{f}(x) = \underbrace{\langle \Phi(x)\,,\, w \rangle}_{\textbf{1D projection}} = \sum_k w_k\, \phi_k(x)\ .$$

Data: $x \in \mathbb{R}^d$ $\qquad\qquad$ $\Phi(x) \in \mathbb{R}^{d'}$



Metric: $\|x - x'\|$ $\qquad\qquad$ $\|\Phi(x) - \Phi(x')\|$

- How and when is possible to find such a Φ ?
- What "regularity" of f is needed ?

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 $< N$.

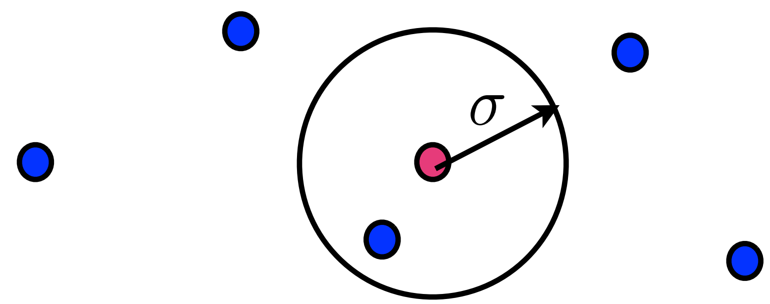
$\Rightarrow$ Choose Φ increasing dimensionality !

Problem: generalisation, overfitting.

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

$\Phi(x)$ is of dimension $d' = \infty$

If σ is small, nearest neighbor classifier type:



Reduction of Dimensionality

- Discriminative change of variable $\Phi(x)$:

$$\Phi(x) \neq \Phi(x') \ \text{ if } \ f(x) \neq f(x')$$

$$\Rightarrow \quad \exists \tilde f \ \text{ with } \ f(x) = \tilde f(\Phi(x))$$

- If $\tilde f$ is Lipschitz: $|\tilde f(z) - \tilde f(z')| \leq C\,\|z - z'\|$

$$z = \Phi(x) \quad \Leftrightarrow \quad |f(x) - f(x')| \leq C\,\|\Phi(x) - \Phi(x')\|$$

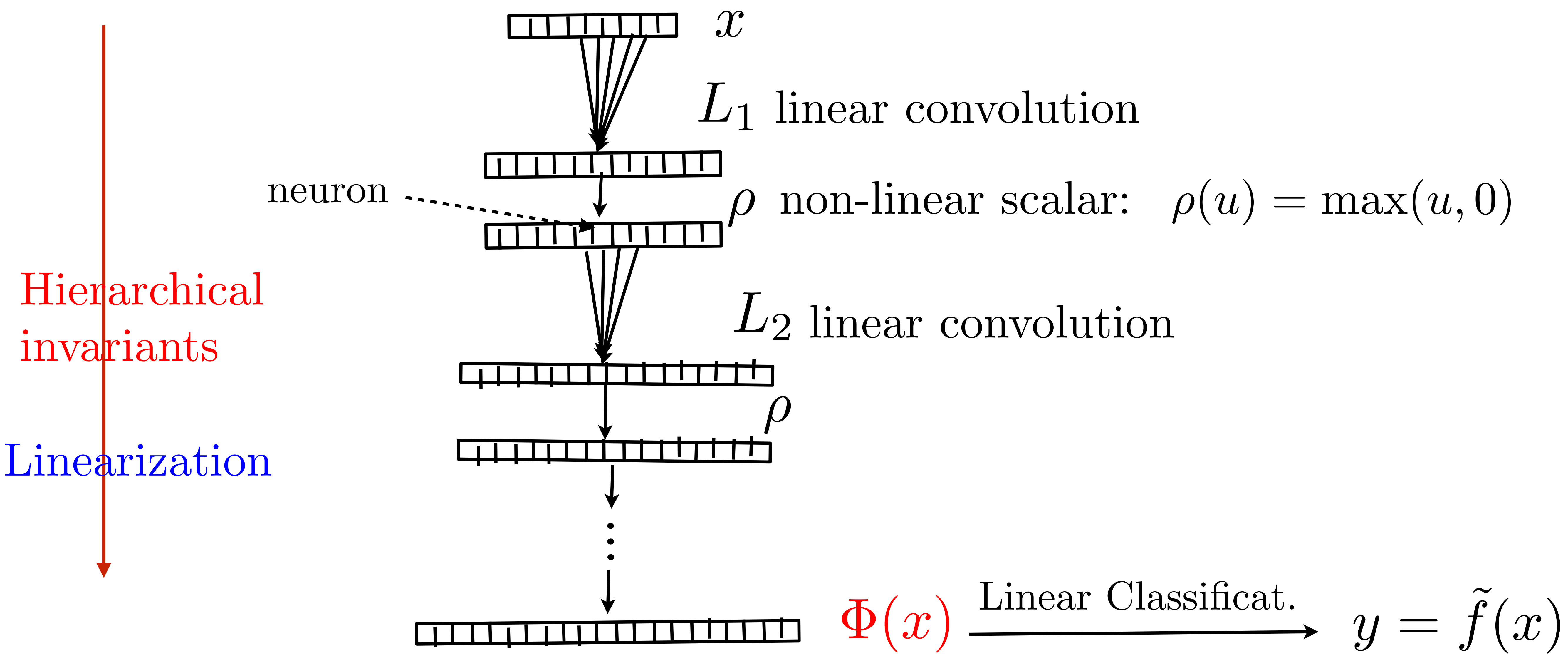
Discriminative: $\|\Phi(x) - \Phi(x')\| \geq C^{-1}\,|f(x) - f(x')|$

- For $x \in \Omega$, if $\Phi(\Omega)$ is bounded and a low dimension d'

$$\Rightarrow \|f - f_M\| \leq C\,M^{-1/d'}$$

Deep Convolution Neworks

- The revival of neural networks: *Y. LeCun*



Optimize L_j with architecture constraints: over 10^9 parameters

Exceptional results for *images, speech, language, bio-data...*

Why does it work so well ? **A difficult problem**

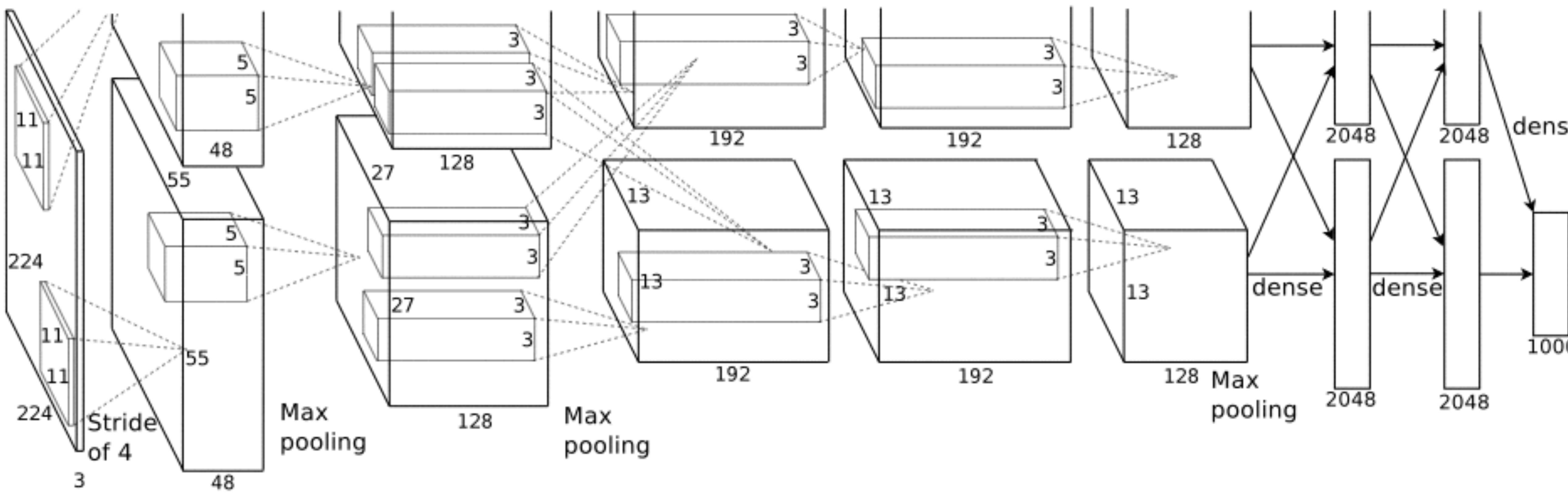
ImageNet Data Basis

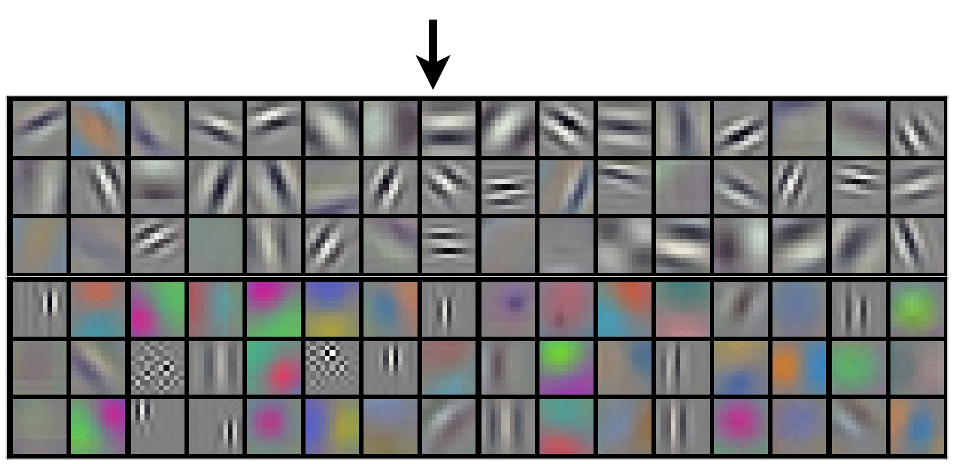
- Data basis with 1 million images and 2000 classes

Alex Deep Convolution Network

A. Krizhevsky, Sutsever, Hinton

- Imagenet supervised training: 1.2 10^6 examples, 10^3 classes
  15.3% testing error in 2012

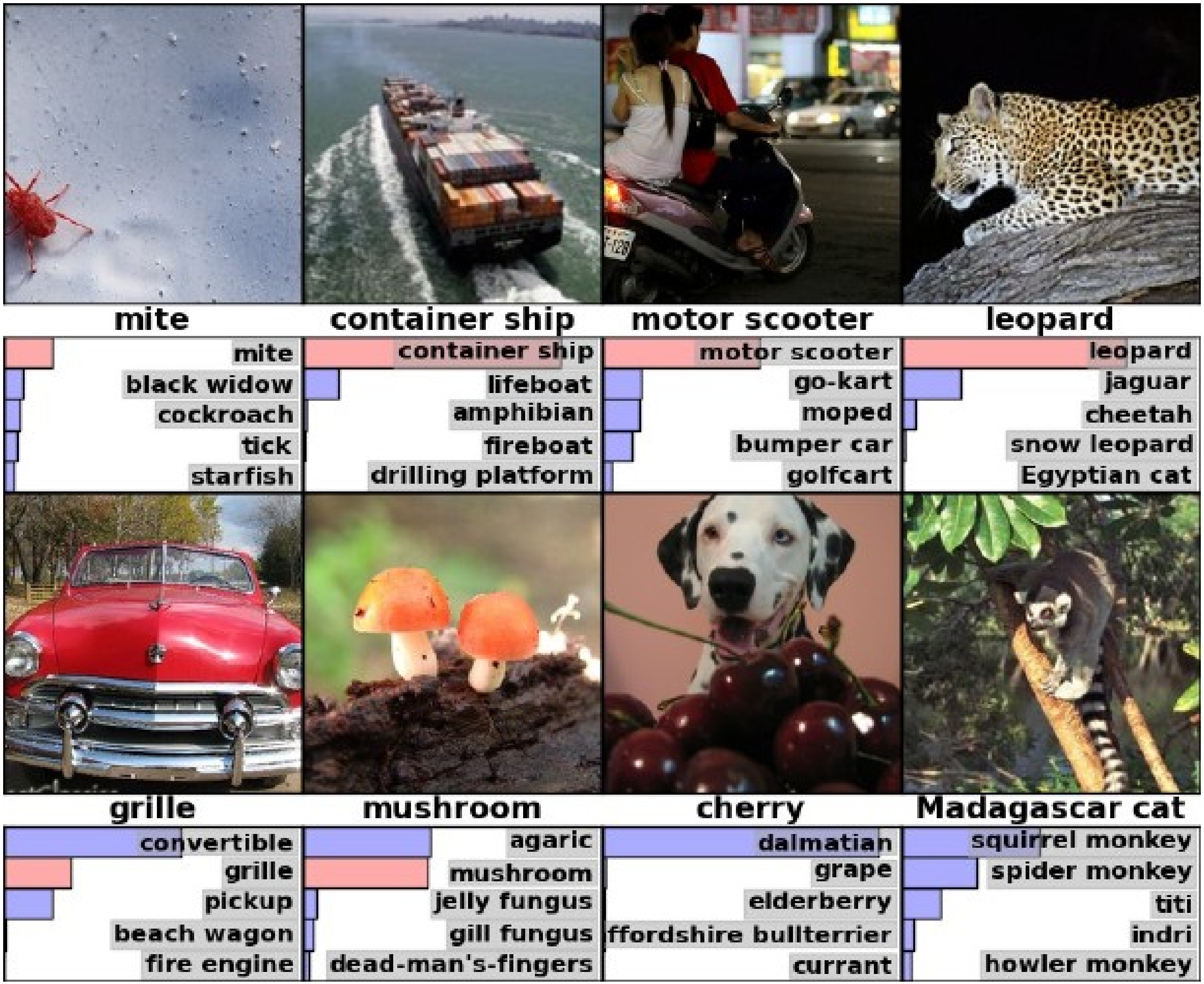




Wavelets

New networks with 5% errors.
Up to 150 layers!

Image Classification

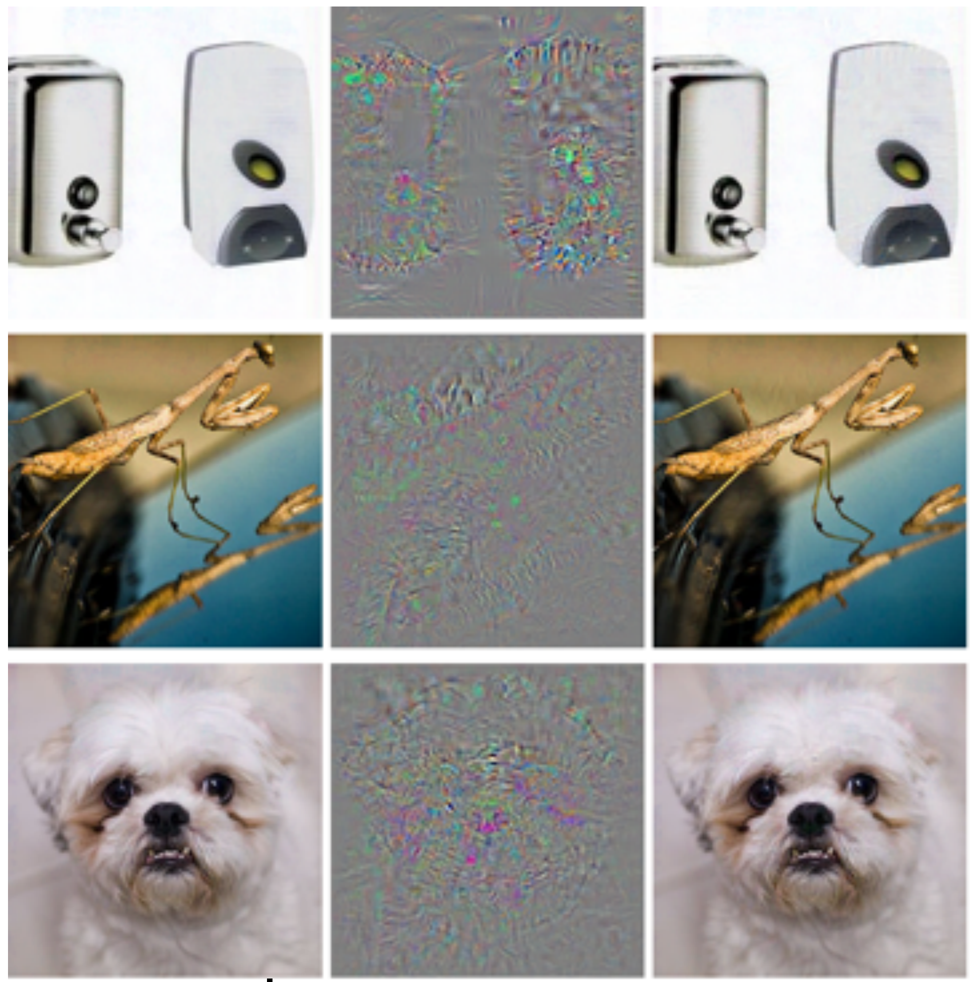


Scene Labeling / Car Driving

Why Understading ?

Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, Fergus

$$x \quad + \quad \epsilon \quad = \quad \tilde{x} \qquad \text{with} \quad \|\epsilon\| < 10^{-2}\|x\|$$

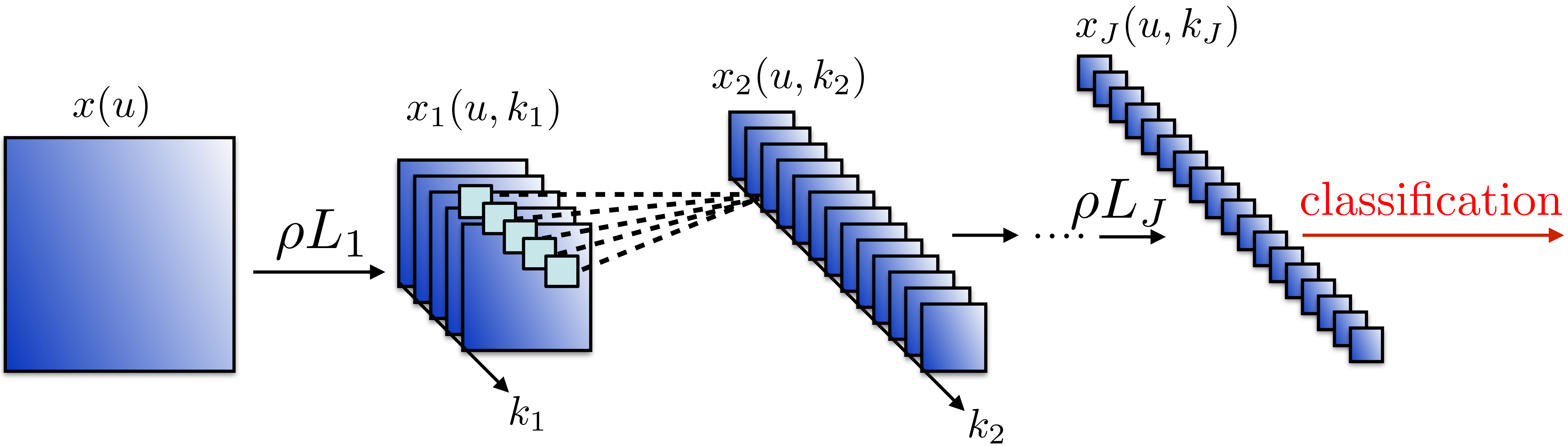


correctly classified

classified as ostrich

- Trial and error testing can not guarantee reliability.

Deep Convolutional Networks



$$x_j = \rho\, L_j\, x_{j-1}$$

- L_j is a linear combination of convolutions and subsampling:

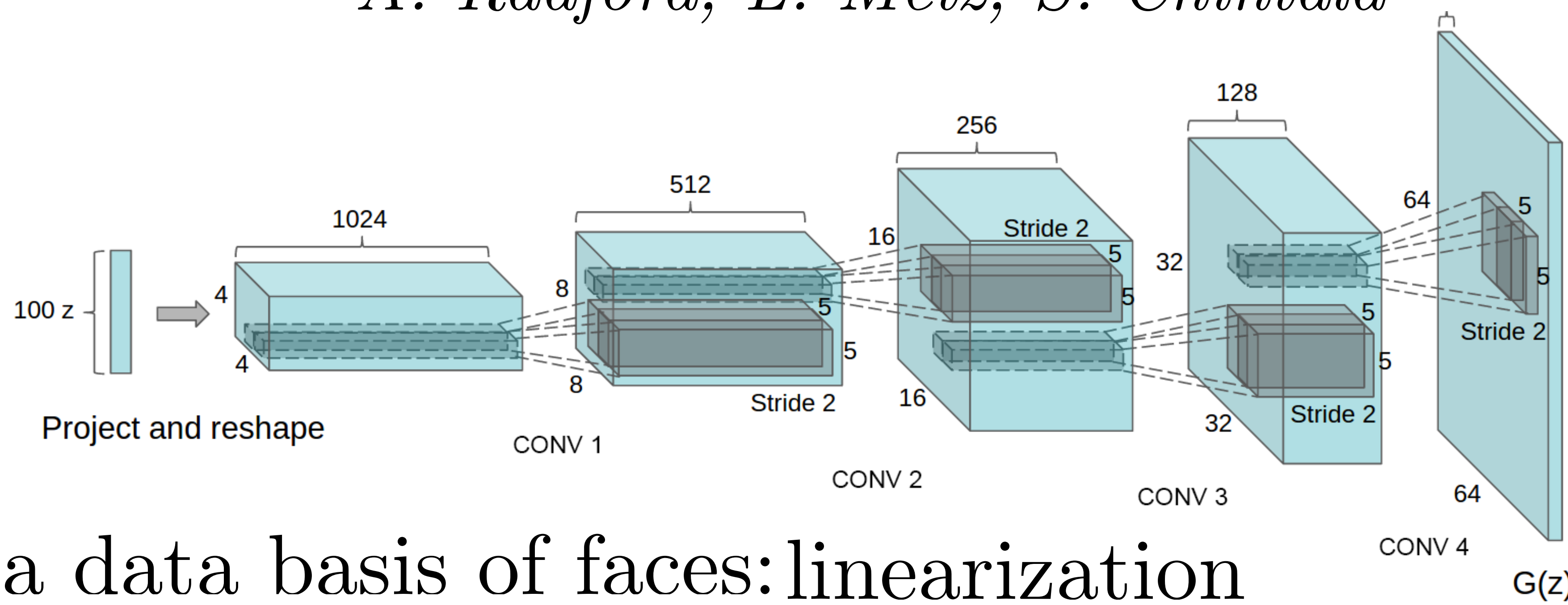
$$x_j(u, k_j) = \rho\Big(\underbrace{\sum_k x_{j-1}(\cdot, k) \star h_{k_j,k}(u)}_{\text{sum across channels}}\Big)$$

- ρ is contractive: $|\rho(u) - \rho(u')| \leq |u - u'|$

$$\rho(u) = \max(u, 0) \text{ or } \rho(u) = |u|$$

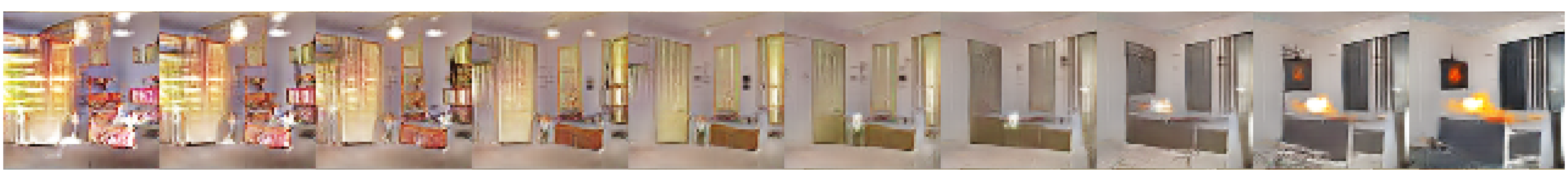
Linearisation in Deep Networks

A. Radford, L. Metz, S. Chintala

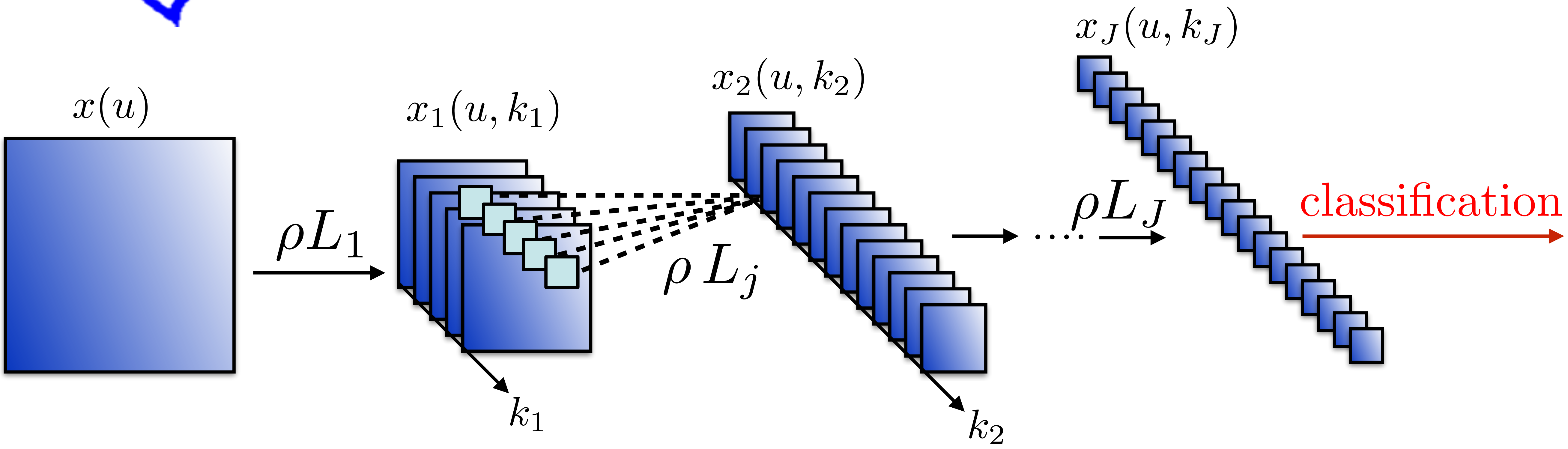


- Trained on a data basis of faces: linearization

- On a data basis including bedrooms: interpolaitons



Many Questions



- Why convolutions ? Translation covariance.
- Why no overfitting ? Contractions, dimension reduction

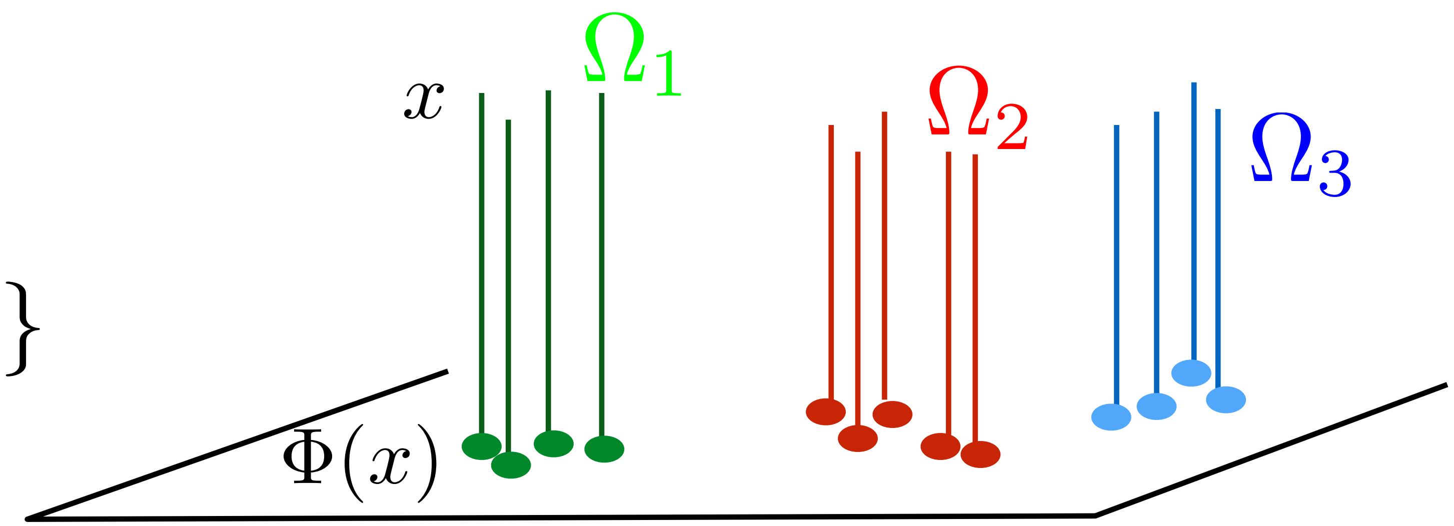
- Why hierarchical cascade ?
- Why introducing non-linearities ?
- How and what to linearise ?
- What are the roles of the multiple channels in each layer ?

Linear Dimension Reduction

Classes

Level sets of $f(x)$

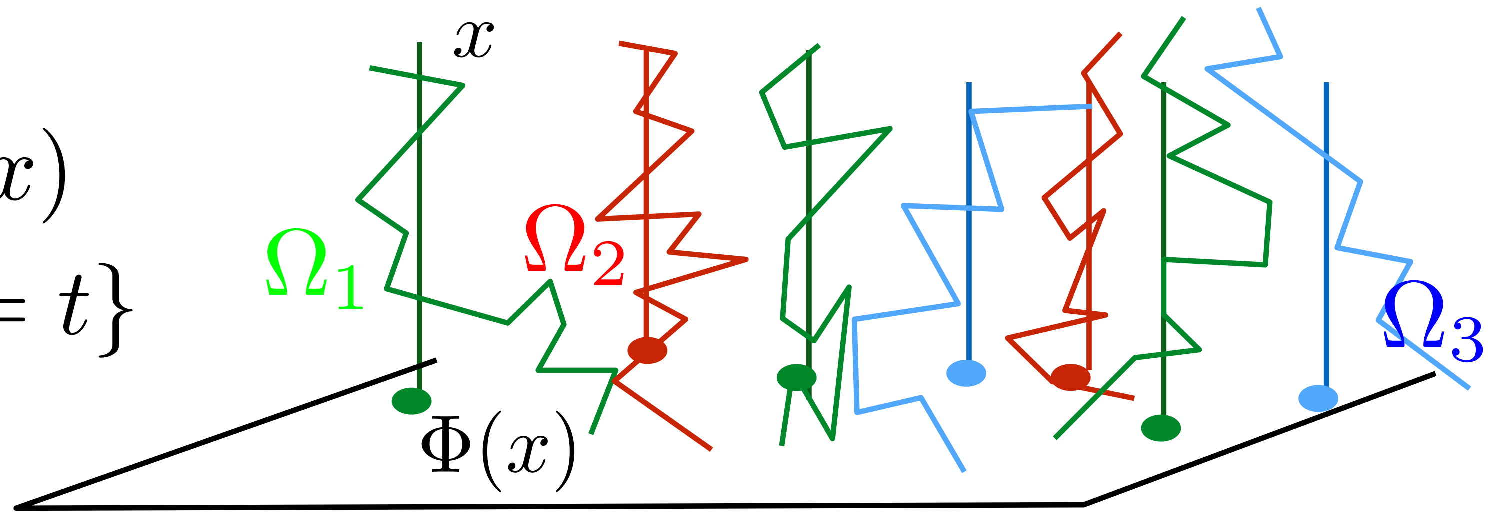
$$\Omega_t = \{x \ : \ f(x) = t\}$$



If level sets (classes) are parallel to a linear space then variables are eliminated by linear projections: *invariants*.

Linearise for Dimensionality Reduction

$$\textit{Classes}$$
$$\textit{Level sets of } f(x)$$
$$\Omega_t = \{x \ : \ f(x) = t\}$$

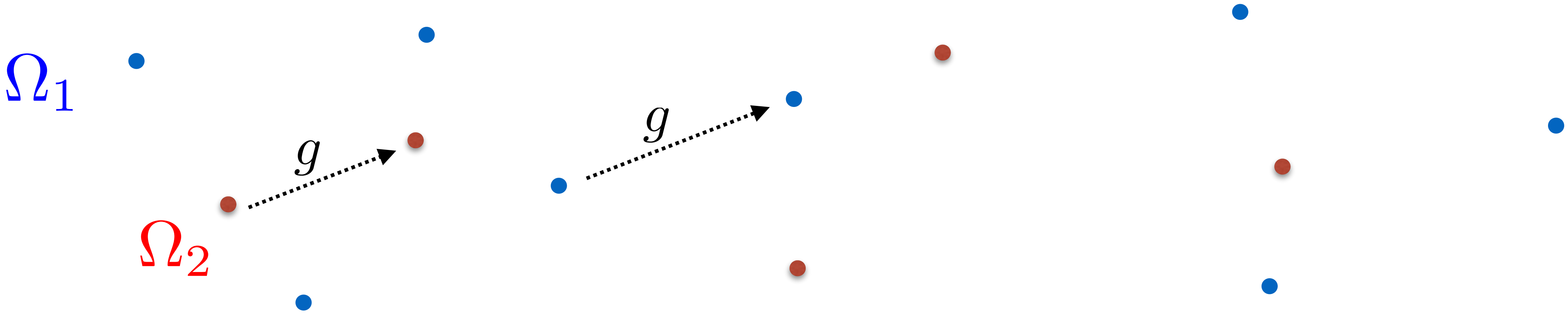


- If level sets Ω_t are not parallel to a linear space
  - Linearise them with a change of variable $\Phi(x)$
  - Then reduce dimension with linear projections
- Difficult because Ω_t are high-dimensional, irregular, known on few samples.

Level Set Geometry: Symmetries

- Curse of dimensionality $\Rightarrow$ not local but global geometry

  Level sets: classes, characterised by their global symmetries.



- A symmetry is an operator g which preserves level sets:

$$\forall x \quad , \quad f(g.x) = f(x) \; : \; \text{global}$$

If g_1 and g_2 are symmetries then $g_1.g_2$ is also a symmetry

$$f(g_1.g_2.x) = f(g_2.x) = f(x)$$

Groups of symmetries

- $G = \{$ all symmetries $\}$ is a group: unknown

$$\forall (g, g') \in G^2 \quad \Rightarrow g.g' \in G$$

Inverse: $\quad \forall g \in G \;\; , \quad g^{-1} \in G$

Associative: $\quad (g.g').g'' = g.(g'.g'')$

If commutative $\; g.g' = g'.g \;$: Abelian group.

- Group of dimension n if it has n generators:

$$g = g_1^{p_1} \, g_2^{p_2} \, ... \, g_n^{p_n}$$

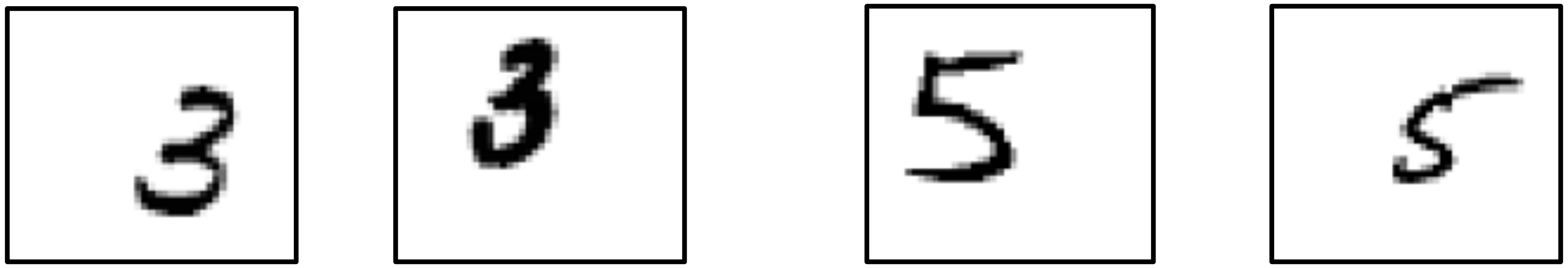
- Lie group: infinitely small generators (Lie Algebra)

Translation and Deformations

- Digit classification:

$x(u)$ $\qquad$ $x'(u) = x(u - \tau(u))$

Ω_3

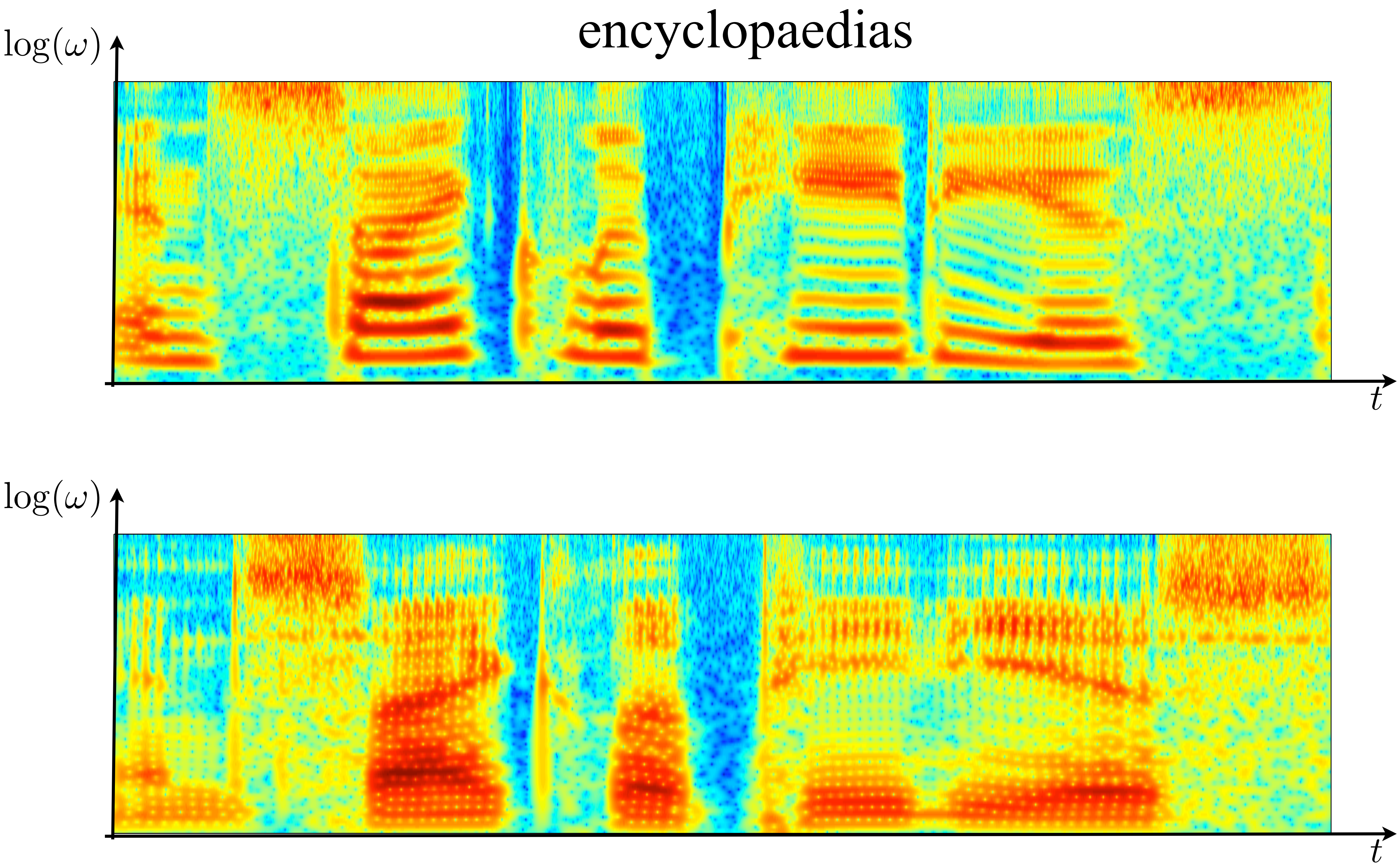


Ω_5

- Globally invariant to the translation group: small
- Locally invariant to small diffeomorphisms: huge group

Video of Philipp Scott Johnson

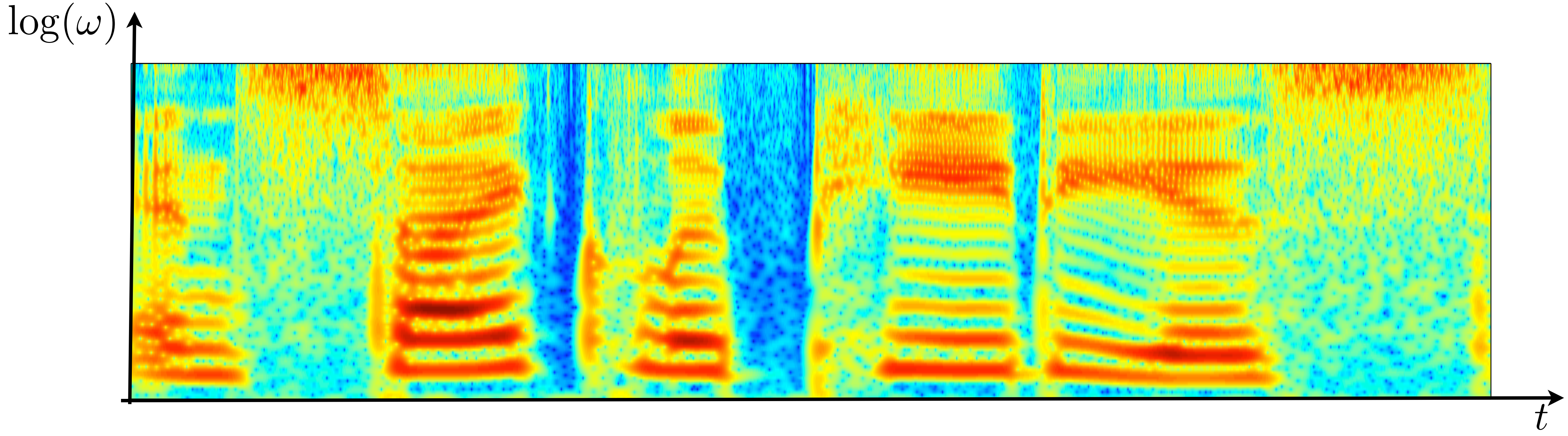
Frequency Transpositions



H : Heisenberg group of "time-frequency" translations

Frequency Transpositions

Time and frequency translations and deformations:

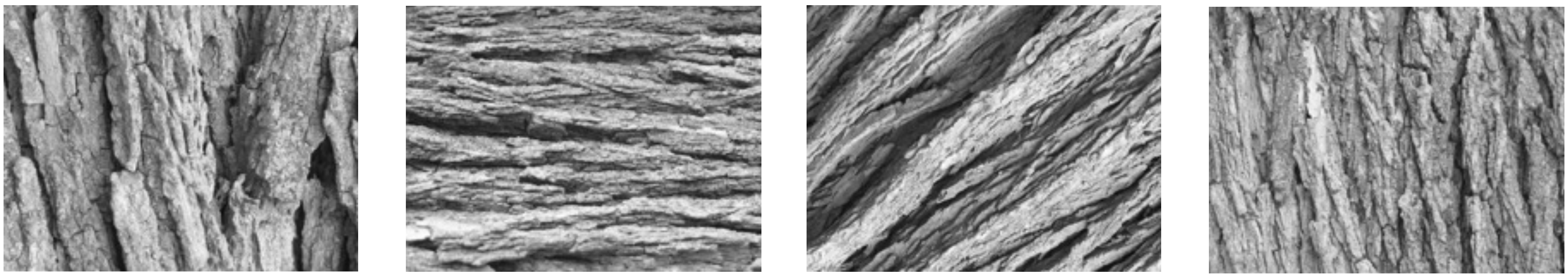


- Frequency transposition invariance is needed

  for speech recognition not for locutor recognition.

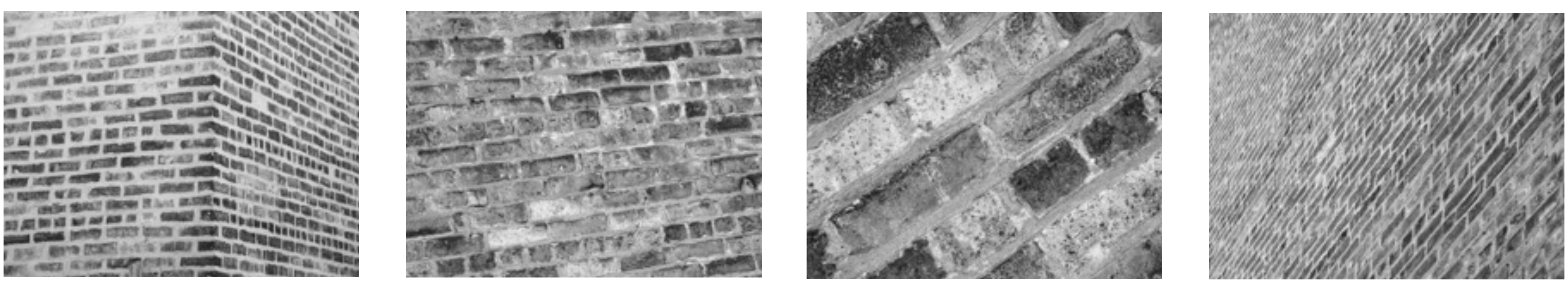
Rotation and Scaling Variability

- Rotation and deformations



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

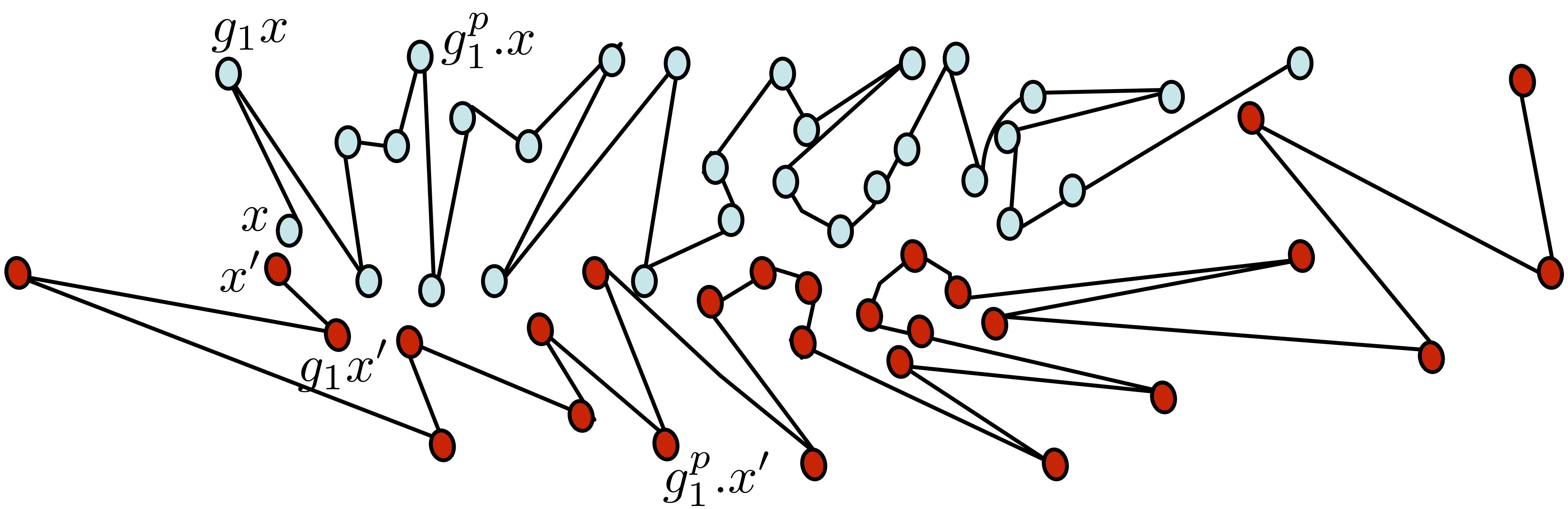
- Scaling and deformations



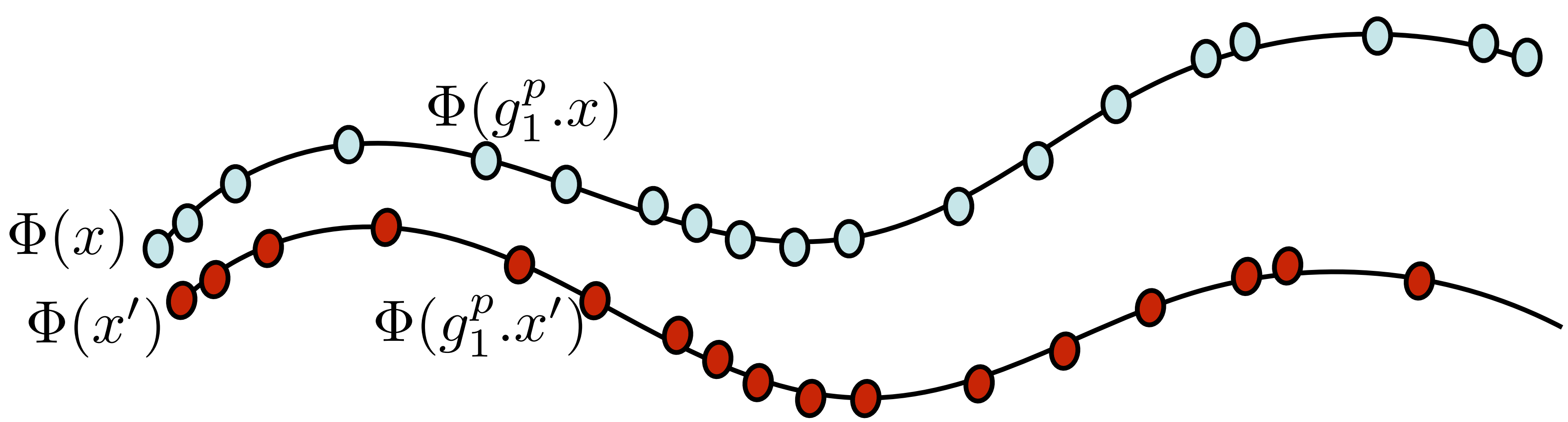
Group: $\mathbb{R} \times \mathrm{Diff}(\mathbb{R})$

Linearize Symmetries

- A change of variable $\Phi(x)$ must linearize the orbits $\{g.x\}_{g \in G}$



- Linearise symmetries with a change of variable $\Phi(x)$

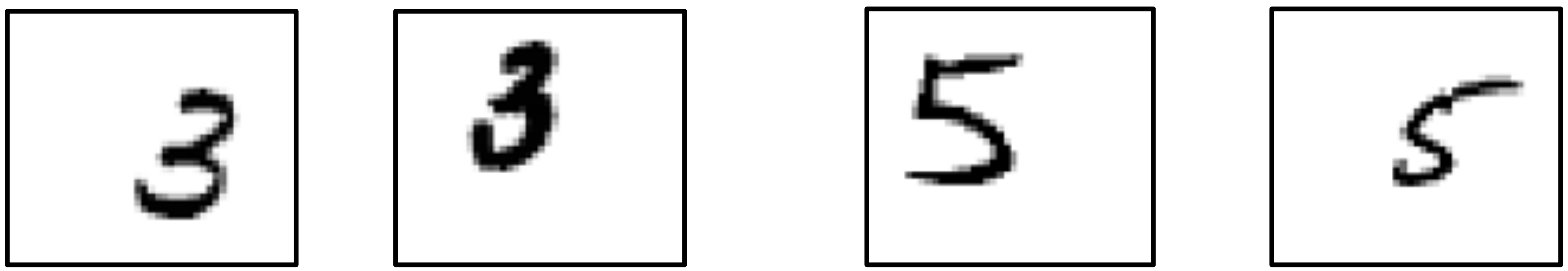


- Lipschitz: $\forall x, g \ : \ \|\Phi(x) - \Phi(g.x)\| \leq C \, \|g\|$

Translation and Deformations

- Digit classification:

$x(u)$ $x'(u)$



- Globally invariant to the translation group

- Locally invariant to small diffeomorphisms

Linearize small diffeomorphisms:
$\Rightarrow$ Lipschitz regular

Video of Philipp Scott Johnson

Translations and Deformations

- Invariance to translations:

$$g.x(u) = x(u-c) \quad \Rightarrow \quad \Phi(g.x) = \Phi(x)\ .$$

- Small diffeomorphisms: $g.x(u) = x(u - \tau(u))$

  Metric: $\|g\| = \|\nabla\tau\|_\infty$ maximum scaling

  Linearisation by Lipschitz continuity

$$\|\Phi(x) - \Phi(g.x)\| \leq C\,\|\nabla\tau\|_\infty\ .$$

- Discriminative change of variable:

$$\|\Phi(x) - \Phi(x')\| \geq C^{-1}\,|f(x) - f(x')|$$

Fourier Deformation Instability

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

$$x_c(t) = x(t-c) \quad \Rightarrow \quad \hat{x}_c(\omega) = e^{-ic\omega}\, \hat{x}(\omega)$$

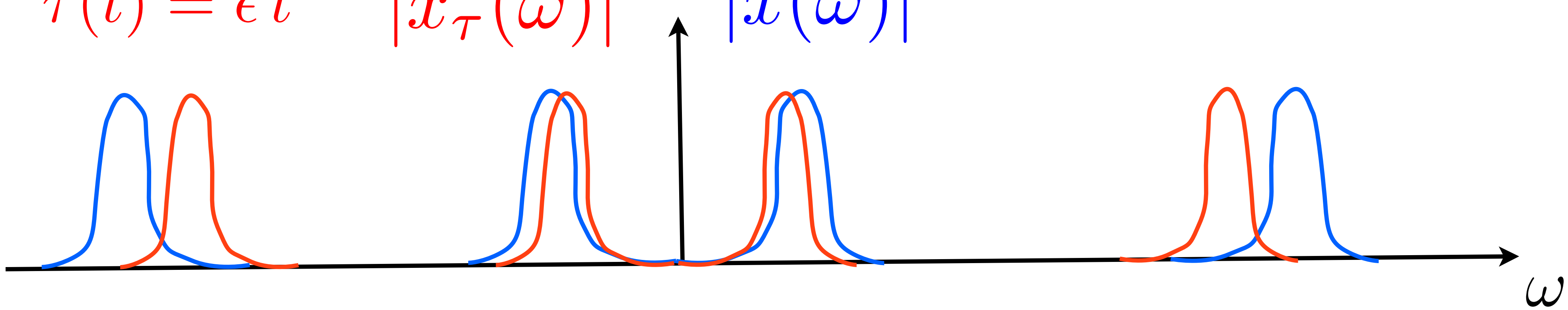
The modulus is invariant to translations:

$$\Phi(x) = |\hat{x}| = |\hat{x}_c|$$

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

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

$\tau(t) = \epsilon\, t$ $\quad |\widehat{x}_\tau(\omega)|$ $\quad |\widehat{x}(\omega)|$



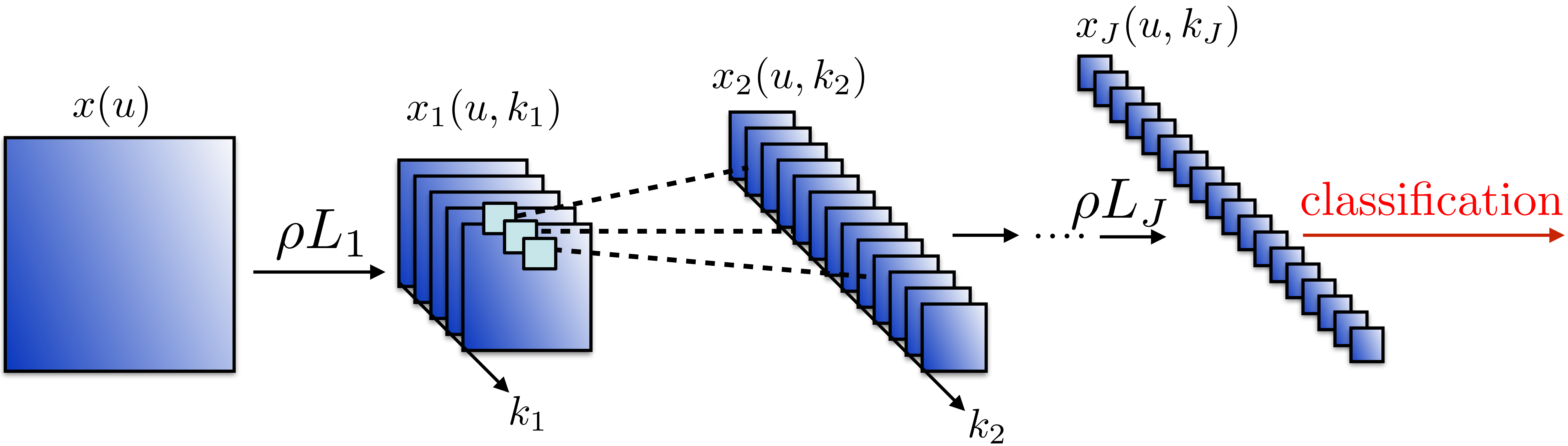
$$\Rightarrow \quad \big\| |\hat{x}| - |\hat{x}_\tau| \big\| \gg \|\nabla \tau\|_\infty\, \|x\|$$

Deep Neural Network Mathematical Mysteries for High Dimensional Learning

Stéphane Mallat

École Normale Supérieure
www.di.ens.fr/data

Deep Convolutional Trees



$$x_j = \rho\, L_j\, x_{j-1}$$

L_j is composed of convolutions and subs samplings:

$$x_j(u, k_j) = \rho\Big(x_{j-1}(\cdot, k) \star h_{k_j, k}(u)\Big)$$

No channel communication: how far can we go ?

Why hierachical cascade ?

Translations and Deformations

- Invariance to translations:

$$g.x(u) = x(u-c) \quad \Rightarrow \quad \Phi(g.x) = \Phi(x)\ .$$

- Small diffeomorphisms: $g.x(u) = x(u - \tau(u))$

Metric: $\|g\| = \|\nabla \tau\|_\infty$ maximum scaling

Linearisation by Lipschitz continuity

$$\|\Phi(x) - \Phi(g.x)\| \leq C\, \|\nabla \tau\|_\infty\ .$$

- Discriminative change of variable:

$$\|\Phi(x) - \Phi(x')\| \geq C^{-1}\, |f(x) - f(x')|$$

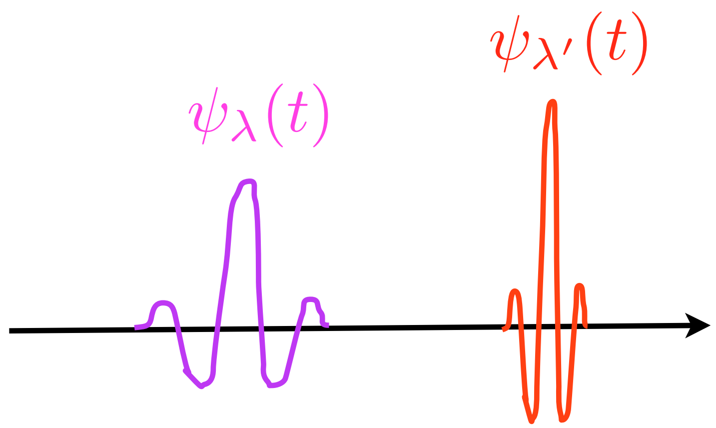
Overview Part II

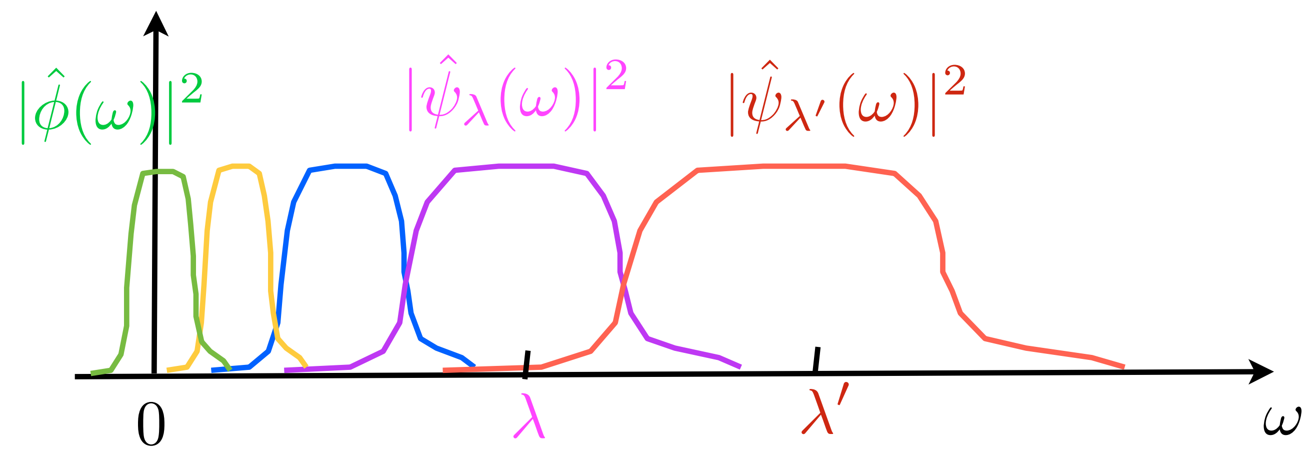
- Wavelet Scattering transform along translations
- Generation of textures and random processes
- Channel connections for more general groups
- Image and audio classification with small training sets
- Quantum chemistry
- Open problems

Understanding Deep Convolutional Networks, arXiv 2016.

Multiscale Wavelet Transform

- Dilated wavelets: $\psi_\lambda(t) = 2^{-j/Q}\,\psi(2^{-j/Q}t)$ with $\lambda = 2^{-j/Q}$





Q-constant band-pass filters $\hat\psi_\lambda$

$$x \star \psi_\lambda(t) = \int x(u)\,\psi_\lambda(t-u)\,du \;\Rightarrow\; \widehat{x \star \psi_\lambda}(\omega) = \widehat{x}(\omega)\,\widehat{\psi}_\lambda(\omega)$$

- Wavelet transform: $Wx = \begin{pmatrix} x \star \phi_{2^J}(t) \\ x \star \psi_\lambda(t) \end{pmatrix}_{\lambda \le 2^J}$: average ; : higher frequencies

Preserves norm: $\|Wx\|^2 = \|x\|^2$.

Why Wavelets ?

- Wavelets are uniformly stable to deformations:

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

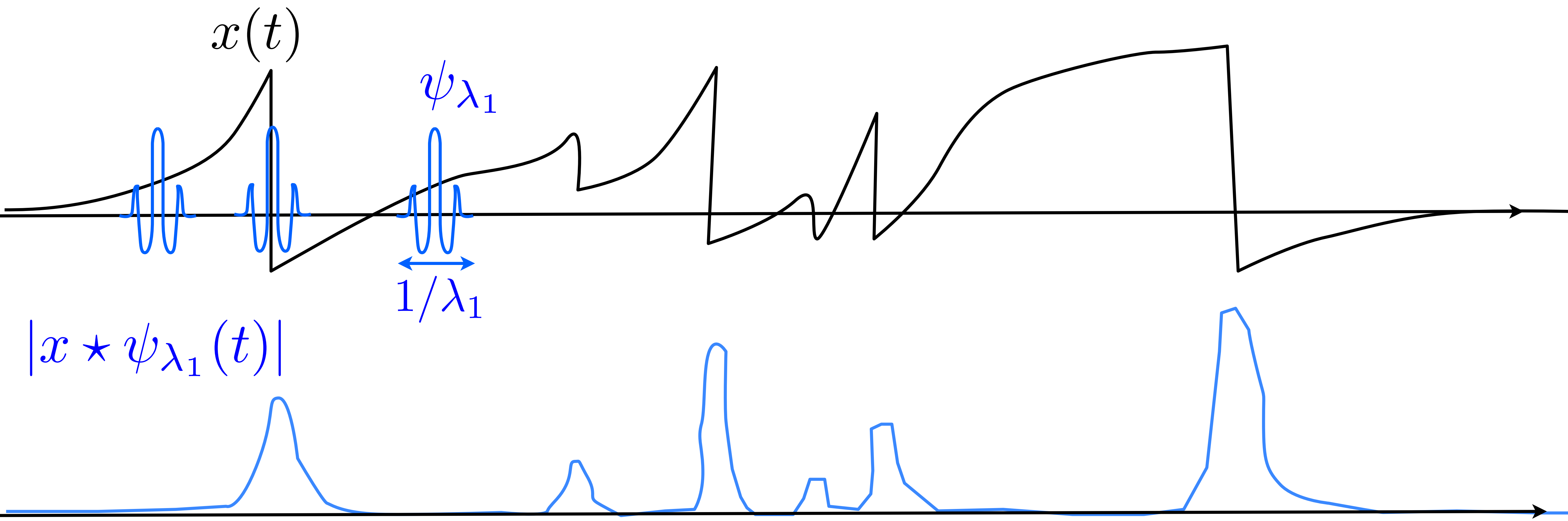
$$\|\psi_\lambda - \psi_{\lambda,\tau}\| \leq C \sup_t |\nabla \tau(t)| \ .$$

- Wavelets separate multiscale information.

- Wavelets provide sparse representations.

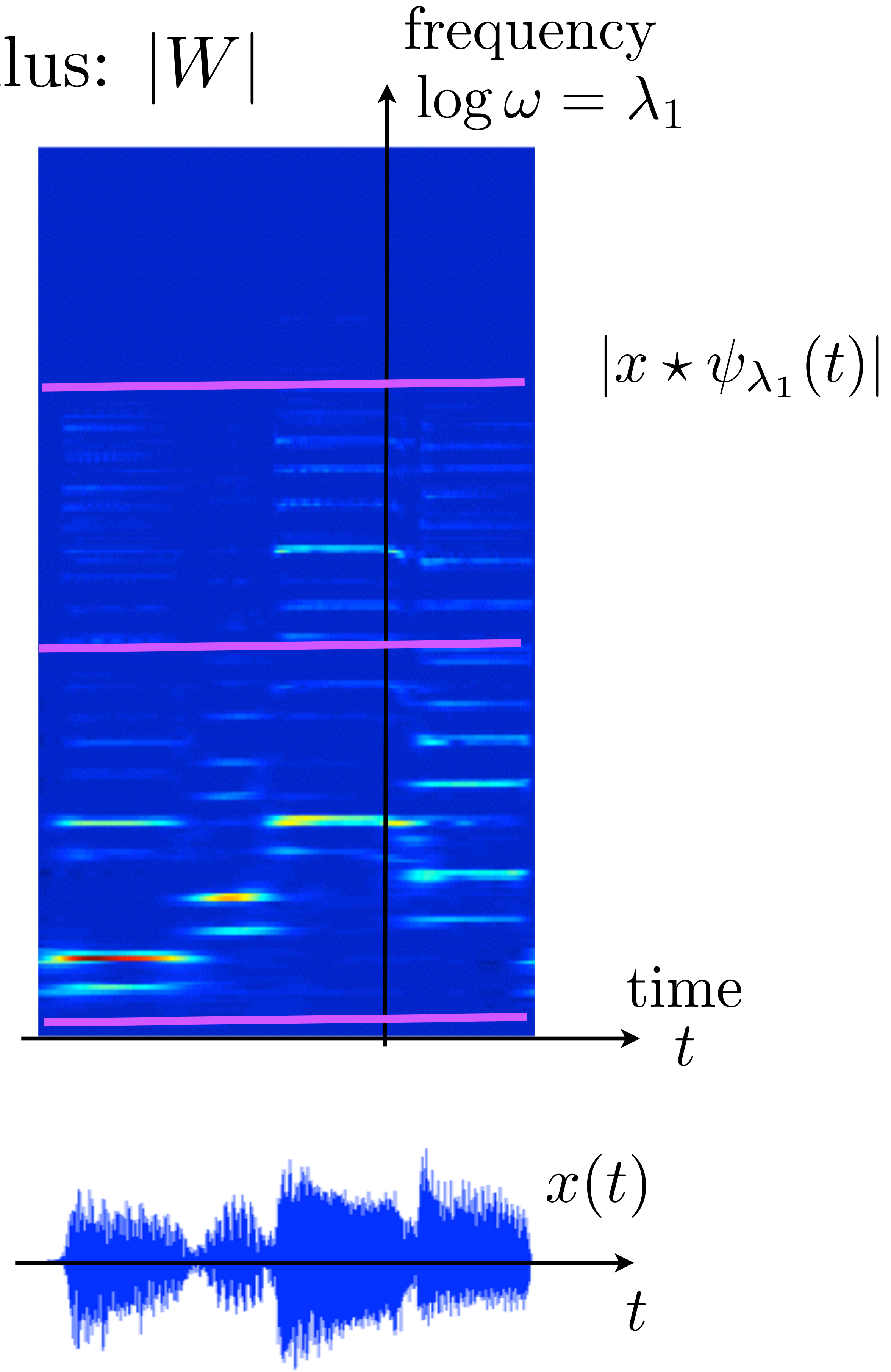
Singular Functions

$$|x \star \psi_{\lambda_1}(t)| = \left| \int x(u) \psi_{\lambda_1}(t-u) \, du \right|$$



Time-Frequency Fibers

Wavelet transform modulus: $|W|$

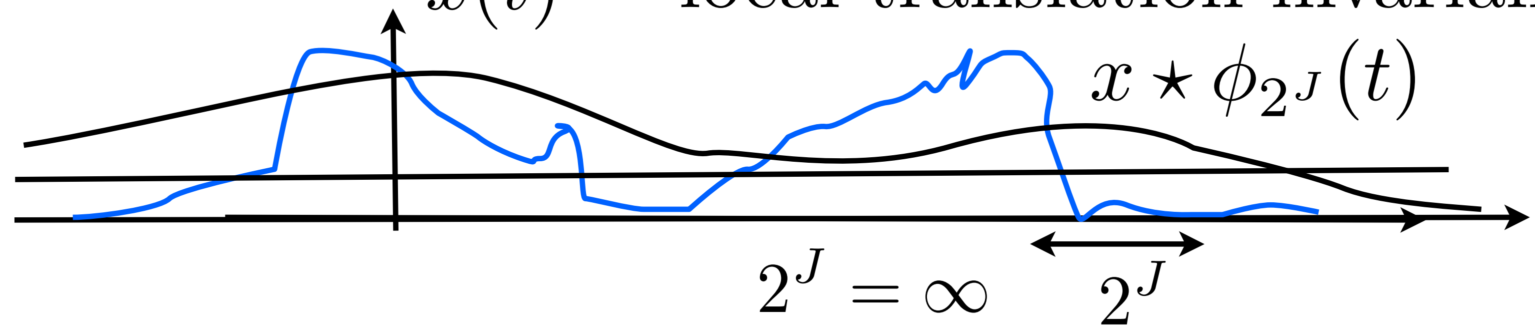


Wavelet Translation Invariance

First wavelet transform

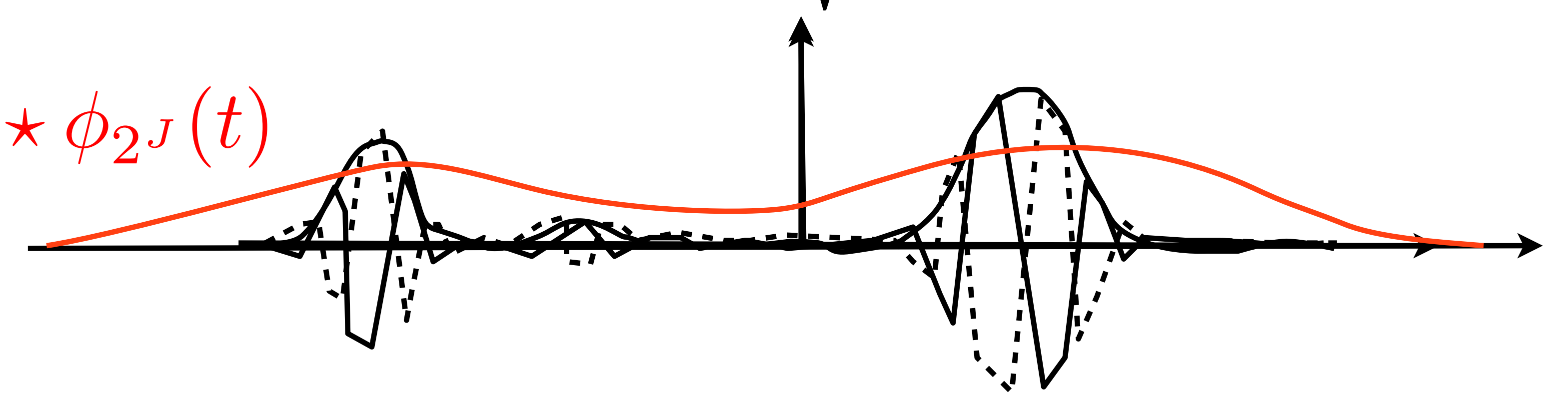
$$W_1 x = \begin{pmatrix} x \star \phi_{2^J} \\ x \star \psi_{\lambda_1} \end{pmatrix}_{\lambda_1}$$

$$|W_1| x = \begin{pmatrix} x \star \phi_{2^J} \\ |x \star \psi_{\lambda_1}| \end{pmatrix}_{\lambda_1}$$



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

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

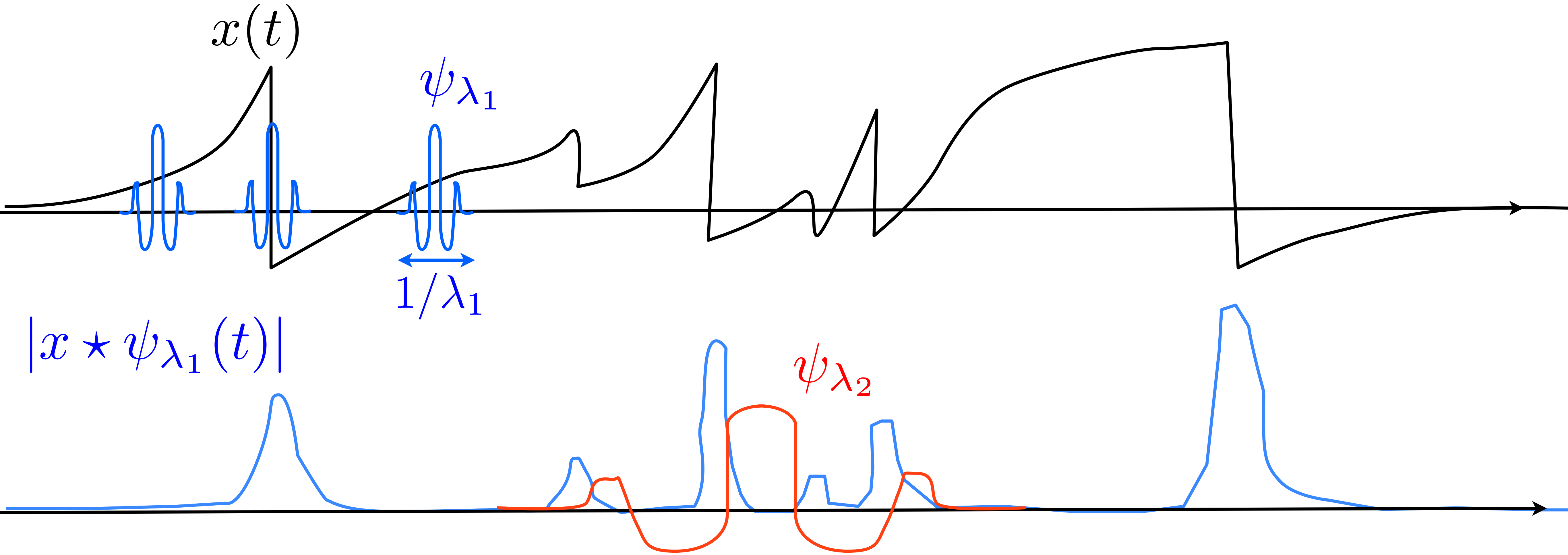


Second wavelet transform modulus

$$|W_2|\ |x \star \psi_{\lambda_1}| = \begin{pmatrix} |x \star \psi_{\lambda_1}| \star \phi_{2^J}(t) \\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}(t)| \end{pmatrix}_{\lambda_2}$$

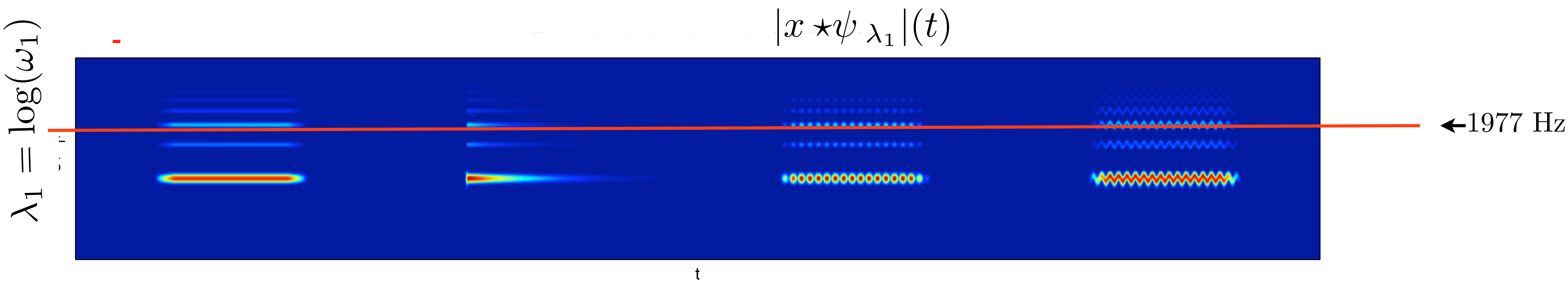
Singular Functions

$$|x \star \psi_{\lambda_1}(t)| = \left| \int x(u)\psi_{\lambda_1}(t-u)\,du \right|$$

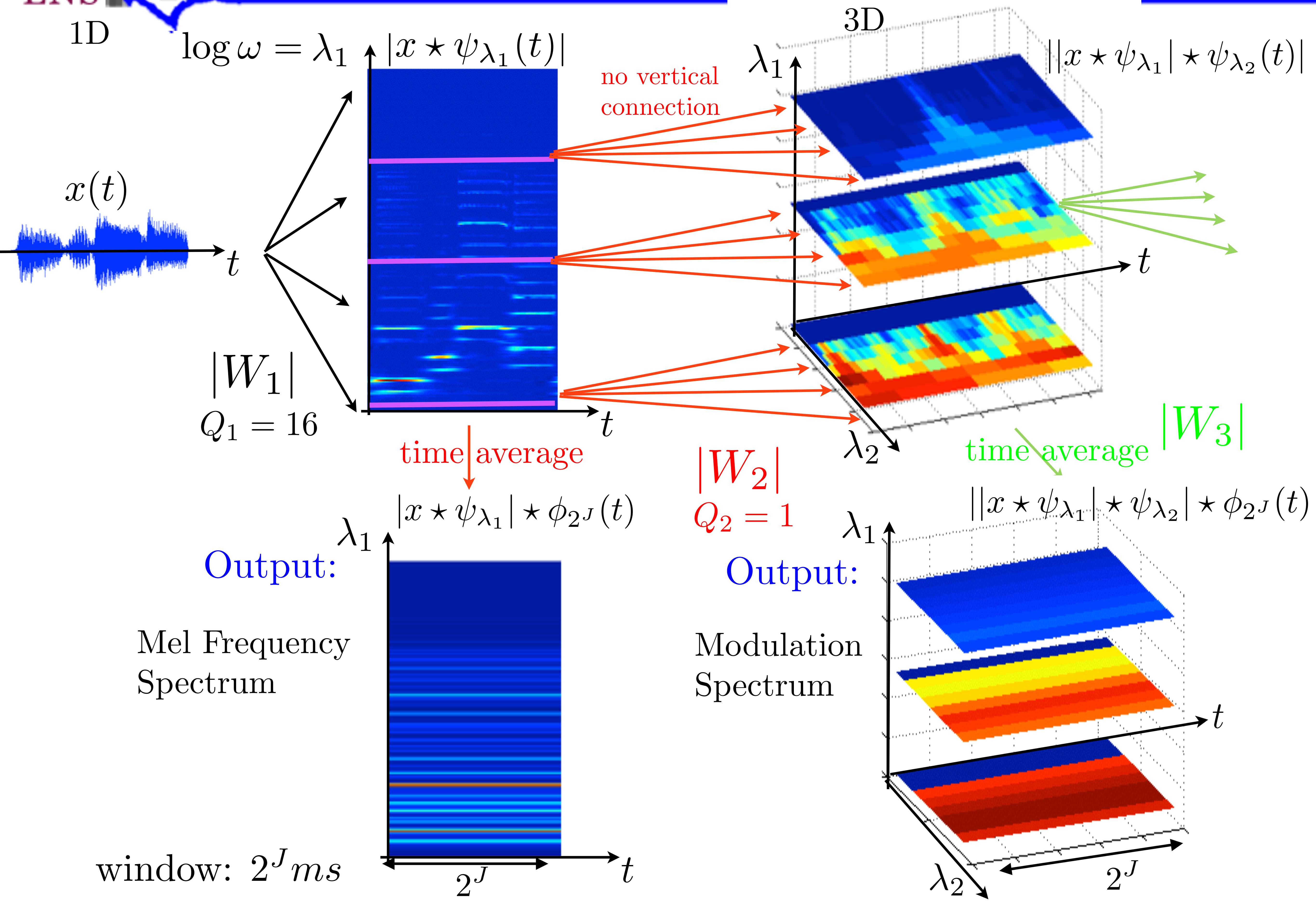


Amplitude Modulation

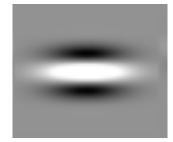
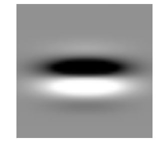
Harmonic sound: $x(t) = a(t)\, e \star h(t)$ with varying $a(t)$



ScatteringConvolution Network

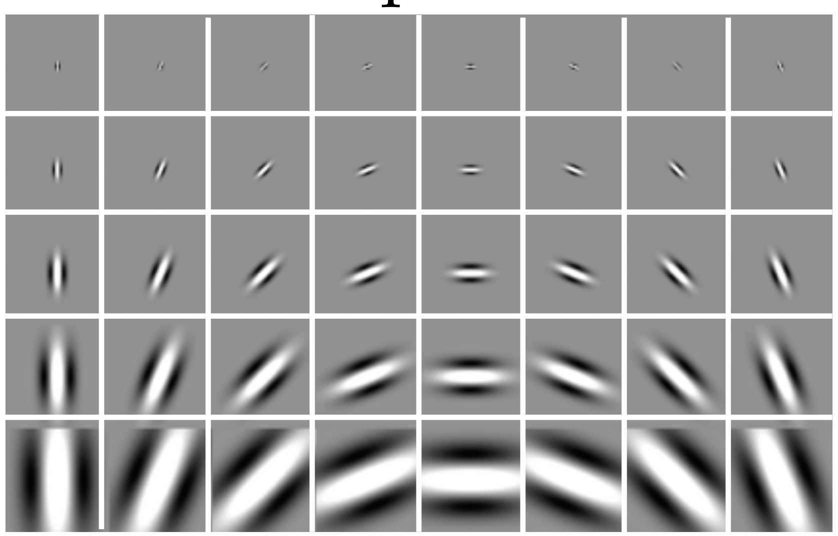


Scale separation with Wavelets

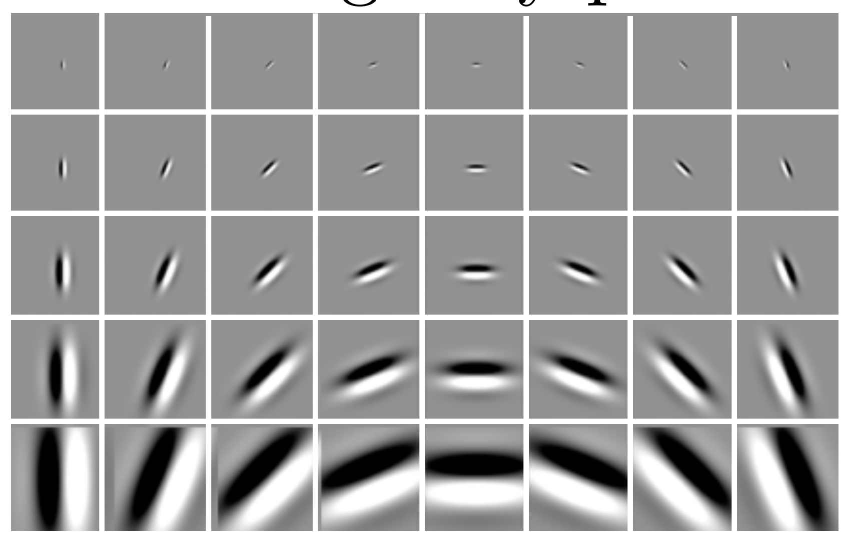
- Wavelet filter $\psi(u)$:  $+\,i$ 

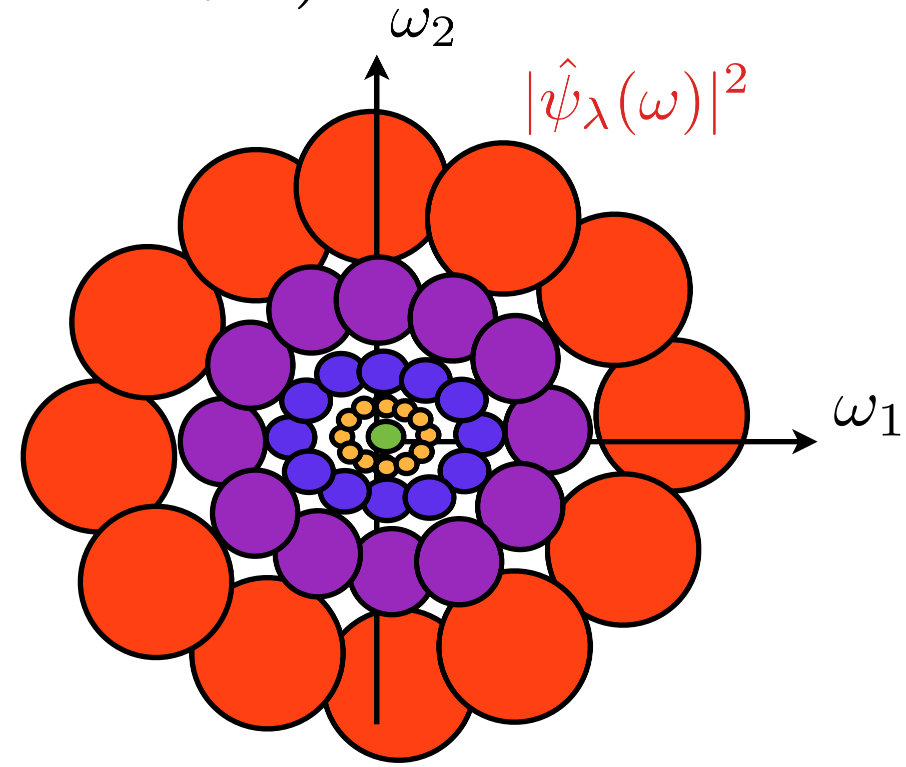
rotated and dilated: $\psi_{2^j,\theta}(u) = 2^{-j}\,\psi(2^{-j} r_\theta u)$

real parts



imaginary parts





$$x \star \psi_{2^j,\theta}(u) = \int x(v)\,\psi_{2^j,\theta}(u-v)\,dv$$

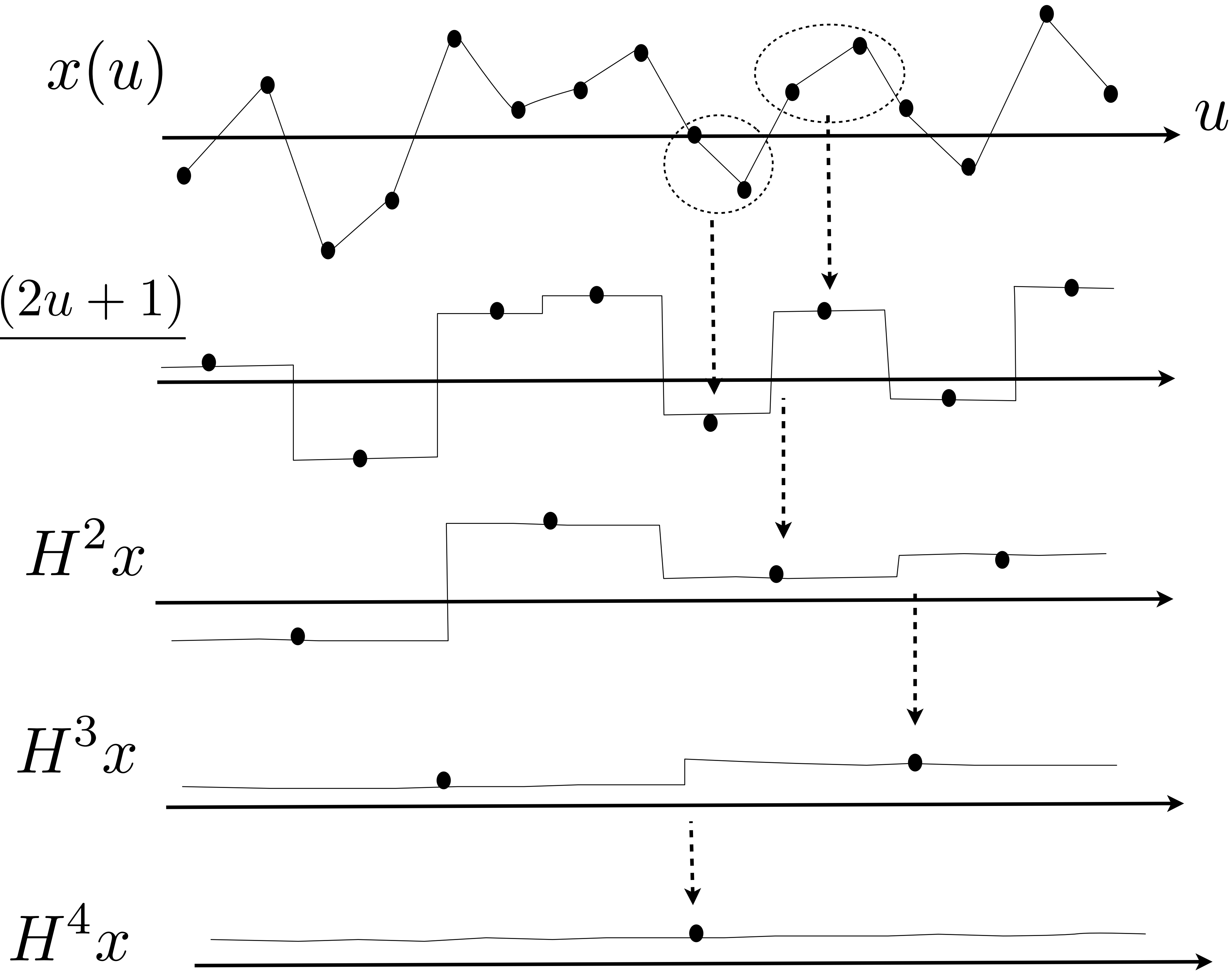
- Wavelet transform: $Wx = \left(\begin{array}{c} x \star \phi_{2^J}(u) \\ x \star \psi_{2^j,\theta}(u) \end{array} \right)_{j \le J, \theta}$: average / : higher frequencies

Preserves norm: $\|Wx\|^2 = \|x\|^2$.

Averaging Pyramid

- Multiscale averaging by cascade of pair averaging:

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



Haar Filtering

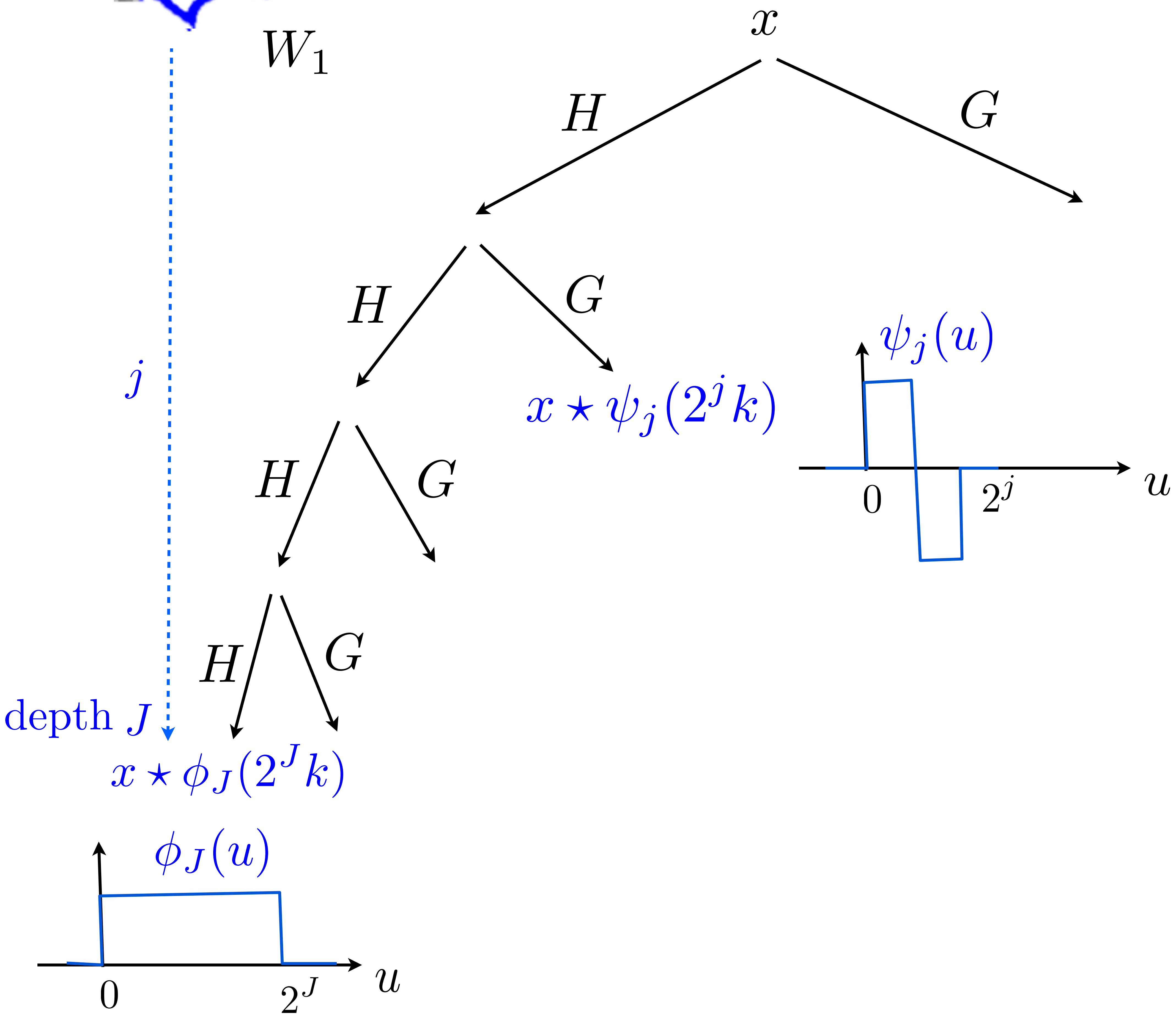
$$\{x(u)\}_{u \leq d} \xrightarrow{H} \left\{\frac{x(2u)+x(2u+1)}{\sqrt{2}}\right\}_{u \leq d/2}$$

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

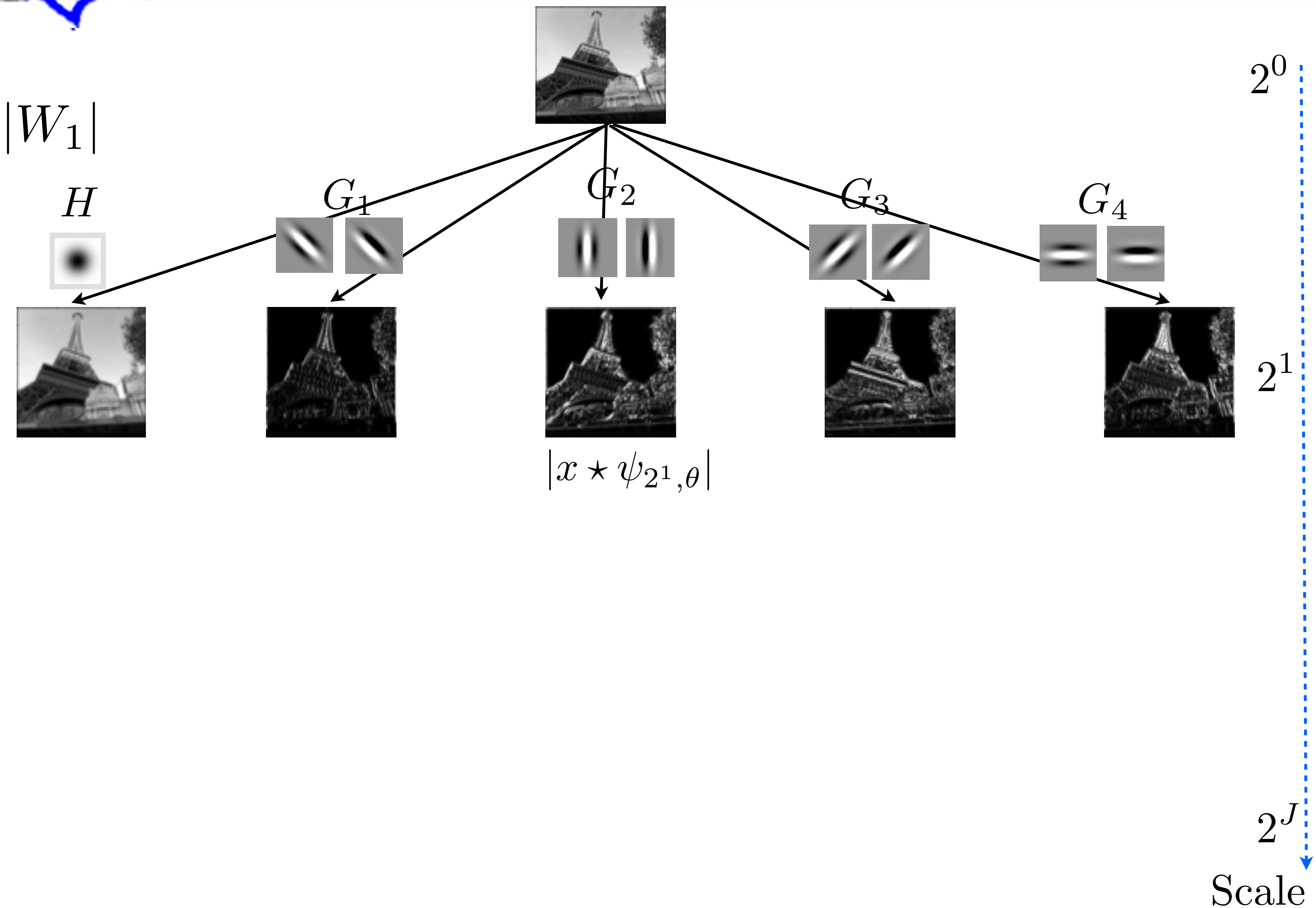
$$Hx(u) = x \star h(2u) \quad \text{and} \quad Gx(u) = x \star g(2u)$$

where h is a low frequency and g is a high frequency filter.

Haar Wavelet Transform



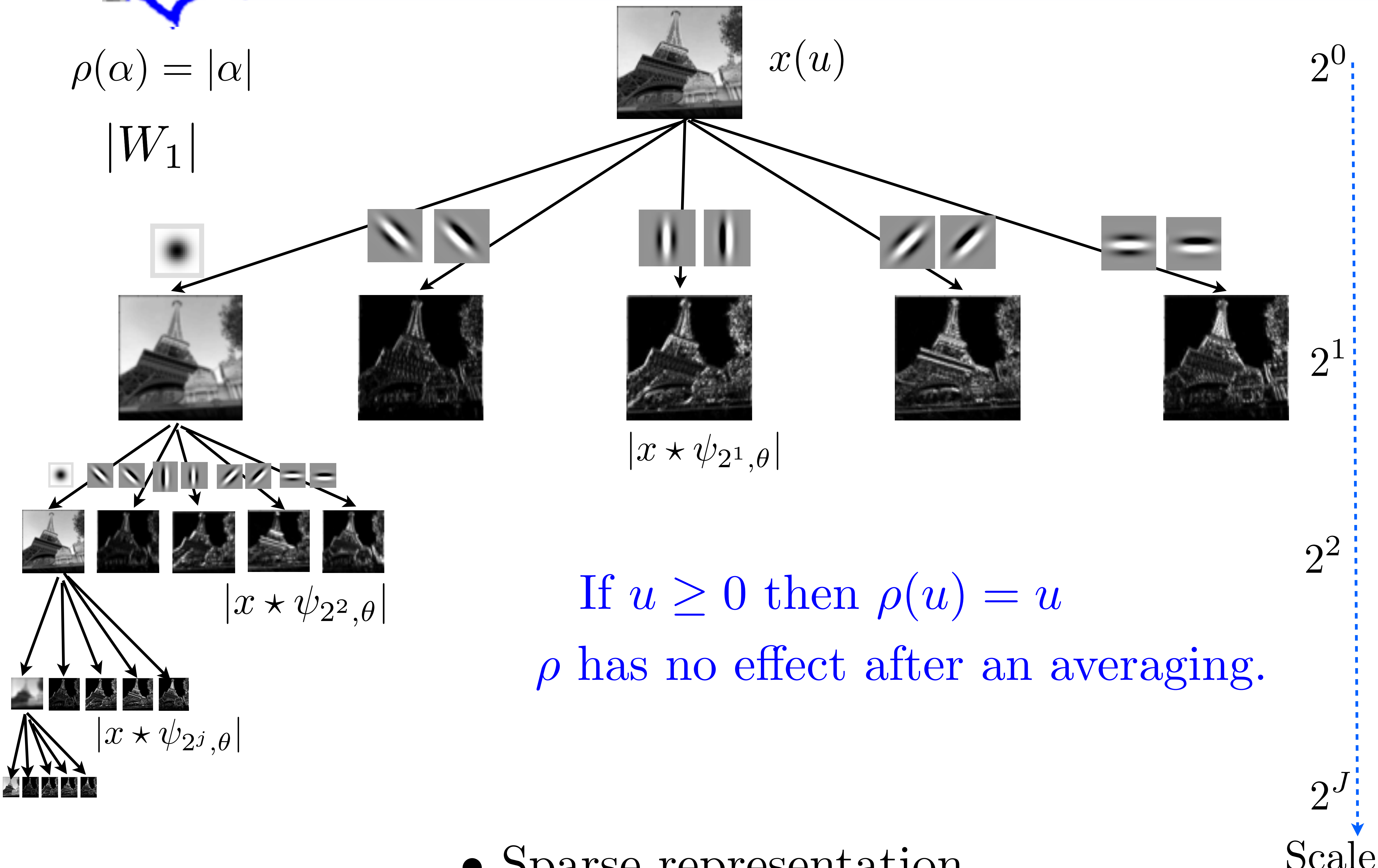
Fast Wavelet Filter Bank



Wavelet Filter Bank

$\rho(\alpha) = |\alpha|$

$|W_1|$

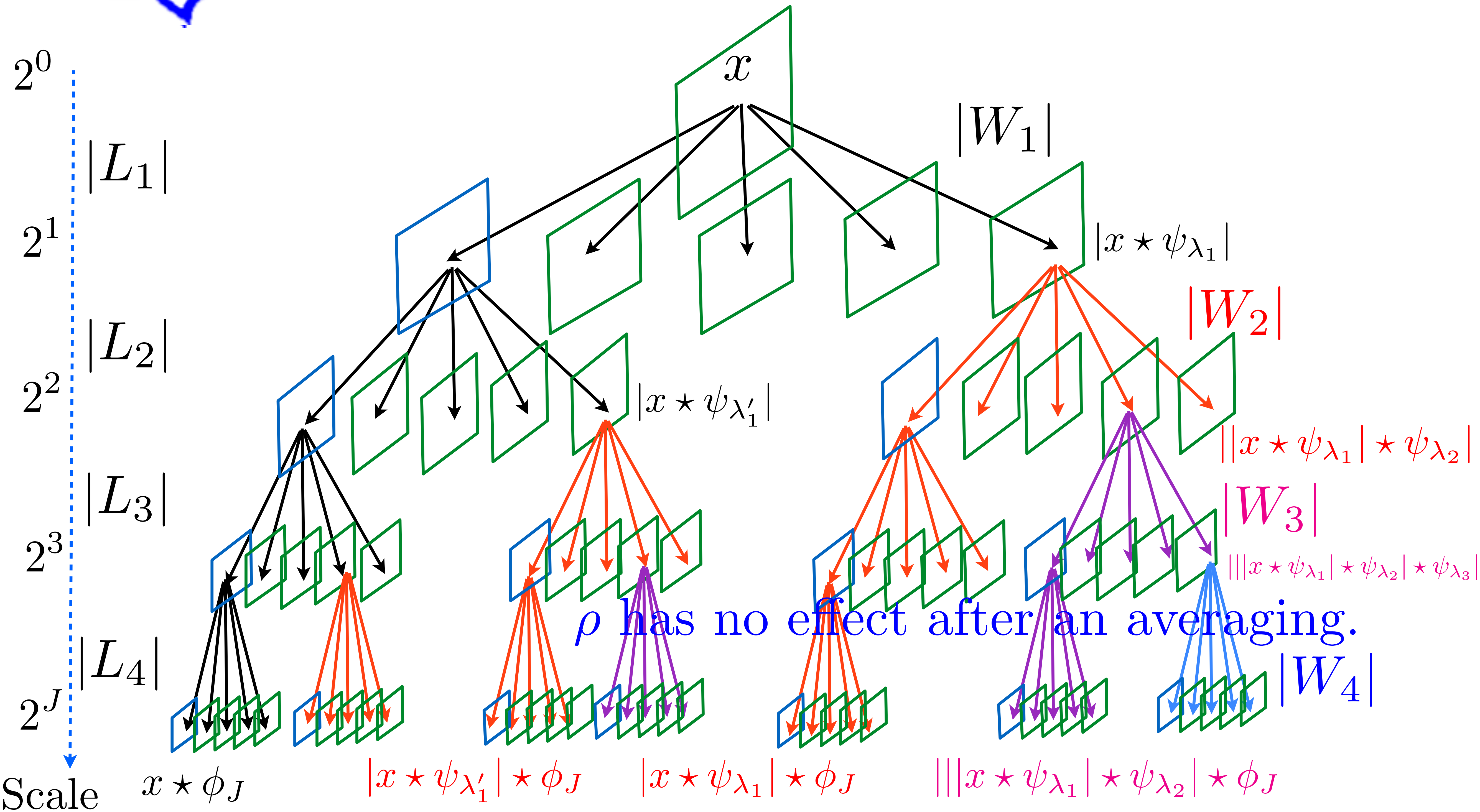


If $u \geq 0$ then $\rho(u) = u$

ρ has no effect after an averaging.

- Sparse representation

Wavelet Convolution Network Tree



$$S_4 x = |L_4|\,|L_3|\,|L_2|\,|L_1| x = |W_4|\,|W_3|\,|W_2|\,|W_1|\,x$$

Contraction

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

$$\rho(u) = |u|$$

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

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

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

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

Scattering Properties

$$S_J x = \begin{pmatrix} x \star \phi_{2^J} \\ |x \star \psi_{\lambda_1}| \star \phi_{2^J} \\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_{2^J} \\ |||x \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi_{2^J} \\ \dots \end{pmatrix}_{\lambda_1, \lambda_2, \lambda_3, \dots} = \dots |W_3| \, |W_2| \, |W_1| \, x$$

$$\|W_k x\| = \|x\| \Rightarrow \| |W_k| x - |W_k| x' \| \leq \|x - x'\|$$

Lemma: $\|[W_k, D_\tau]\| = \|W_k D_\tau - D_\tau W_k\| \leq C \, \|\nabla \tau\|_\infty$

Theorem: *For appropriate wavelets, a scattering is*

contractive $\|S_J x - S_J y\| \leq \|x - y\|$ ($\mathbf{L^2}$ *stability*)

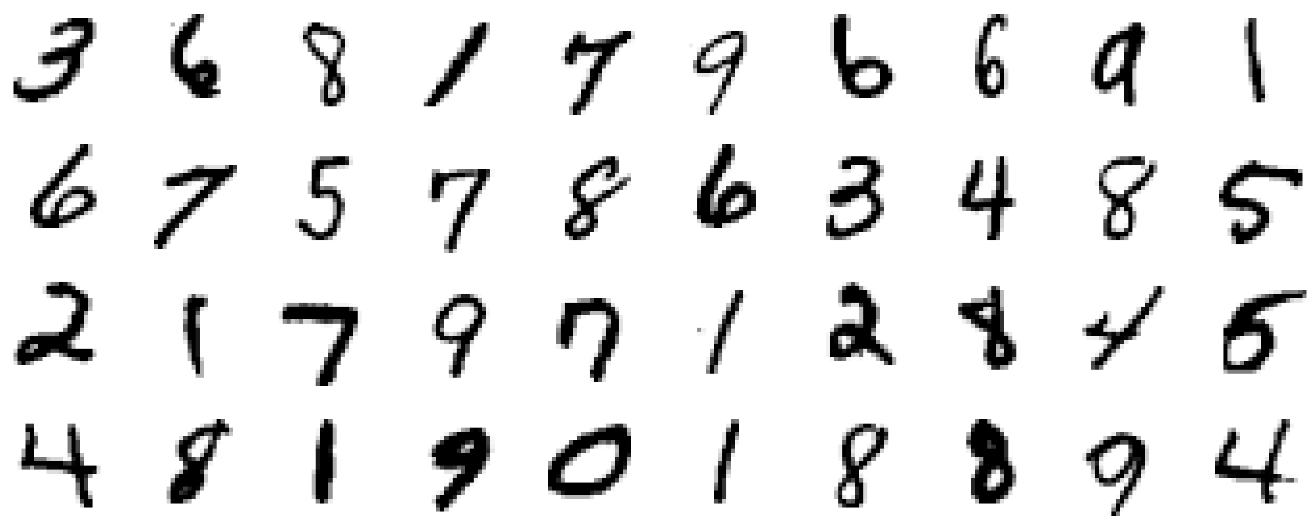
preserves norms $\|S_J x\| = \|x\|$

translations invariance and deformation stability:

if $D_\tau x(u) = x(u - \tau(u))$ *then*

$$\lim_{J \to \infty} \|S_J D_\tau x - S_J x\| \leq C \, \|\nabla \tau\|_\infty \, \|x\|$$

Digit Classification: MNIST



Joan Bruna

$$x \longrightarrow \boxed{S_J x} \longrightarrow \boxed{\begin{array}{c}\text{Supervised}\\ \text{Linear classifier}\end{array}} \longrightarrow y = f(x)$$

Invariants to translations
Linearises small deformations
No learning

Invariants to specific deformations
Separates different patterns

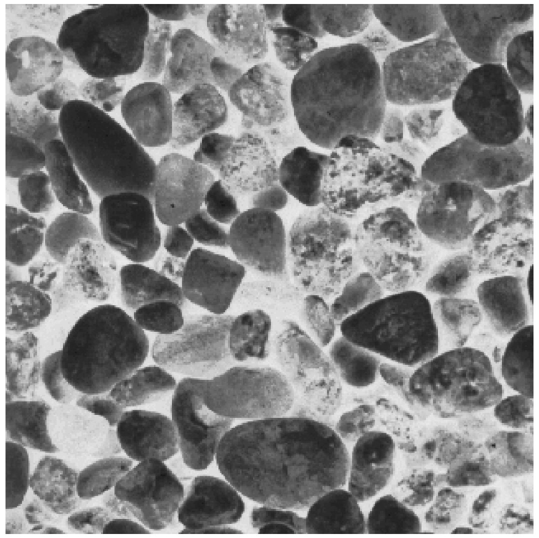
Classification Errors

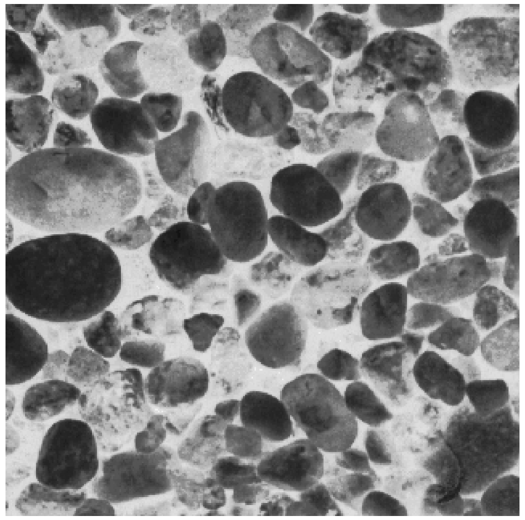
Training size	Conv. Net.	Scattering
50000	0.4%	0.4%

LeCun et. al.

Classification of Stationary Textures

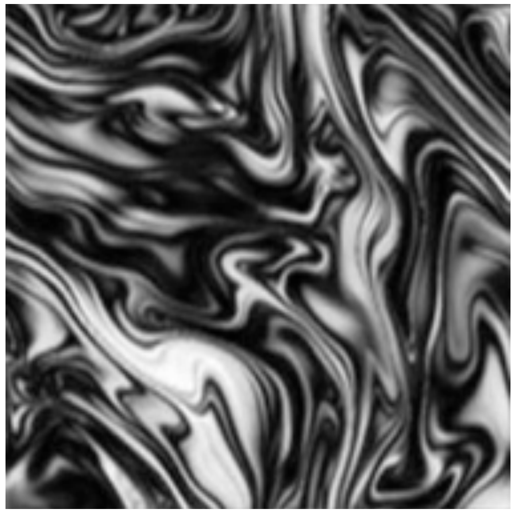
Ω_1





2D Turbulence

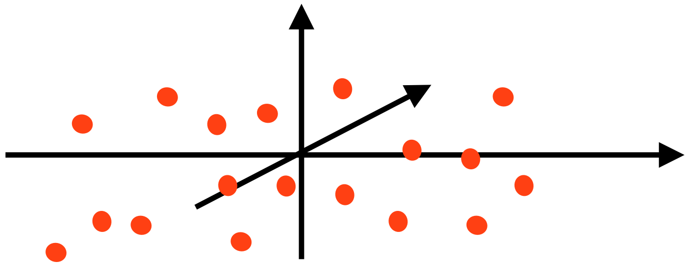
Ω_2



- What stochastic models ?

  Non Gaussian with long-range dependance.

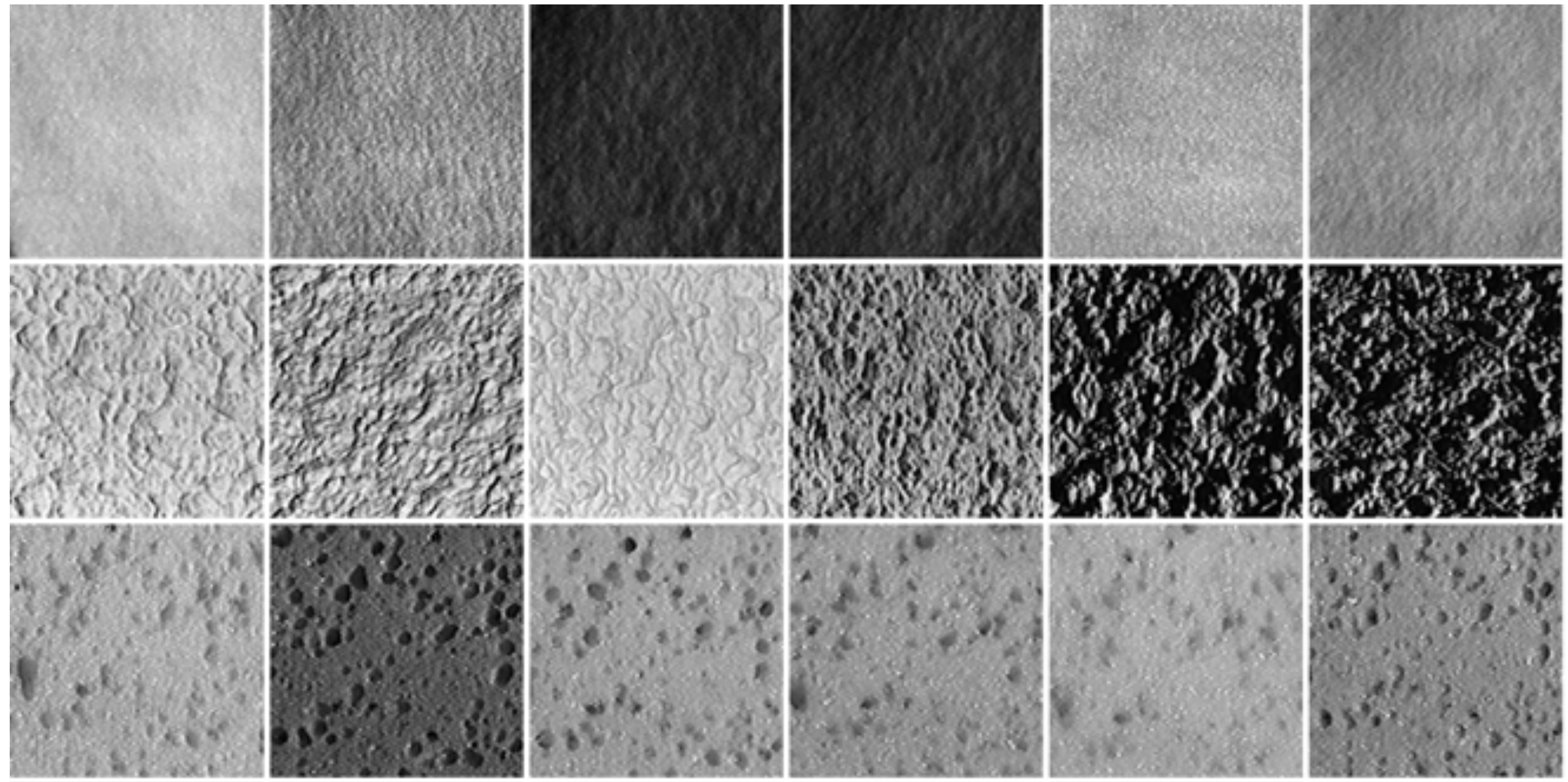
- Can we "Gaussianize" (linearize) such distributions in a reduced dimensional space ?



Classification of Textures

J. Bruna

CUREt database



$$x \longrightarrow \boxed{S_J x} \longrightarrow \boxed{\text{Supervised Linear classifier}} \longrightarrow y = f(x)$$

Classification Errors

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

Scattering Moments of Processes

The scattering transform of a stationary process $X(t)$

$$S_J X = \begin{pmatrix} X \star \phi_{2^J}(t) \\ |X \star \psi_{\lambda_1}| \star \phi_{2^J}(t) \\ ||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_{2^J}(t) \\ |||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi_{2^J}(t) \\ \cdots \end{pmatrix}_{\lambda_1,\lambda_2,\lambda_3,\ldots} \quad : \text{ stationary vector}$$

$J \to \infty$ $\downarrow$

Central limit theorem
with "weak" ergodicity conditions

J. Bruna

Gaussian distribution: $\mathcal{N}\Big(\mathbb{E}(SX), \Sigma_J \to 0\Big)$

$$\mathbb{E}(SX) = \begin{pmatrix} \mathbb{E}(X) \\ \mathbb{E}(|X \star \psi_{\lambda_1}|) \\ \mathbb{E}(||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|) \\ \mathbb{E}(|||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}|) \\ \cdots \end{pmatrix}_{\lambda_1,\lambda_2,\lambda_3,\ldots} \quad : \text{ scattering moments}$$

Scattering Moments of Processes

The scattering transform of a stationary process $X(t)$

$$S_J X = \begin{pmatrix} X \star \phi_{2^J}(t) \\ |X \star \psi_{\lambda_1}| \star \phi_{2^J}(t) \\ ||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_{2^J}(t) \\ |||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi_{2^J}(t) \\ \dots \end{pmatrix}_{\lambda_1,\lambda_2,\lambda_3,\dots} \quad : \text{stationary vector}$$

$J \to \infty$

Central limit theorem
with "weak" ergodicity conditions

Gaussian distribution: $\mathcal{N}\Big(\mathbb{E}(SX), \Sigma_J \to 0\Big)$

- Reconstruction: compute $\tilde{X}$ which minimises

$$\|S_J \tilde{X} - S_J X\|^2$$

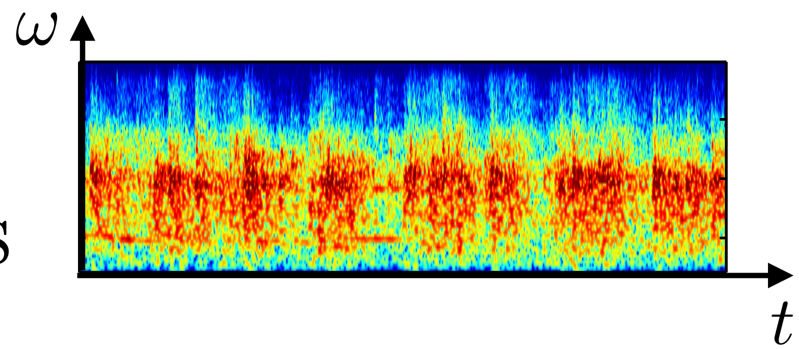
- Gradient descent

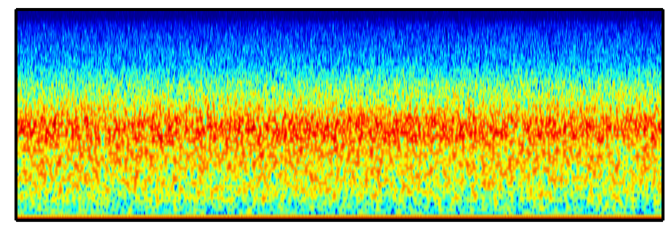
Representation of Audio Textures

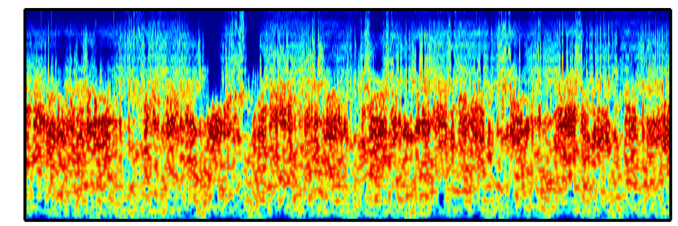
Joan Bruna

Original | Gaussian in time | Gaussian in scattering

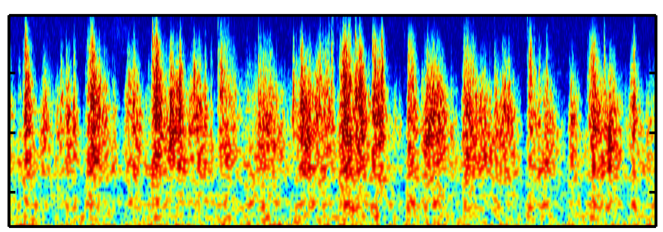
Applauds

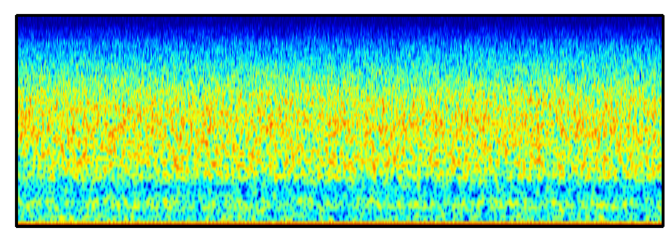


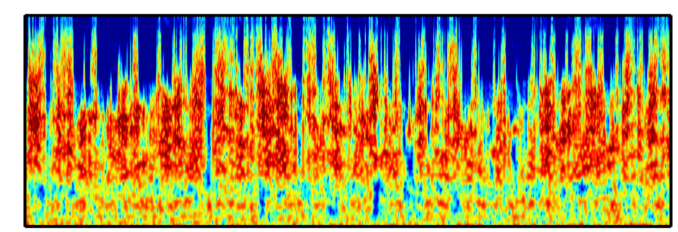




Paper







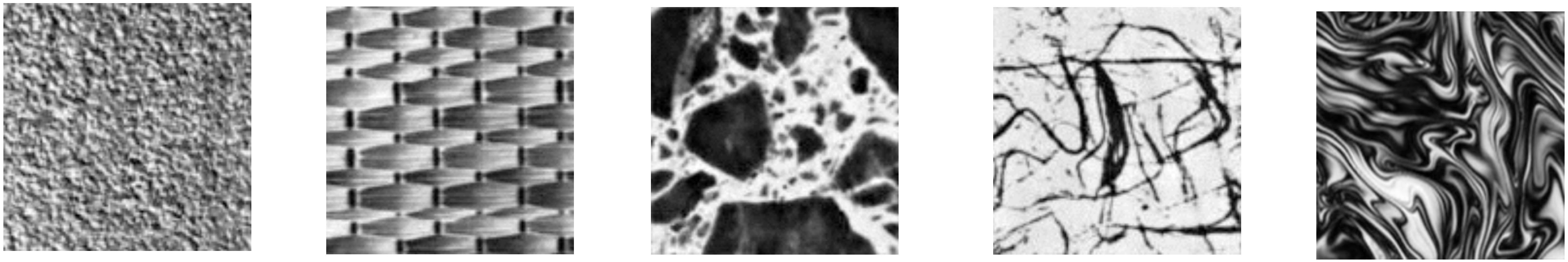
Cocktail Party

Ergodic Texture Reconstructions

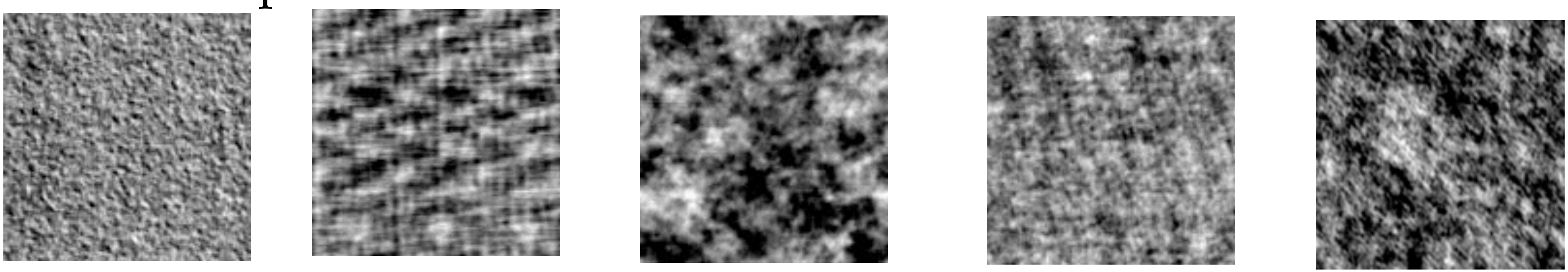
Joan Bruna

Textures of N pixels

2D Turbulence

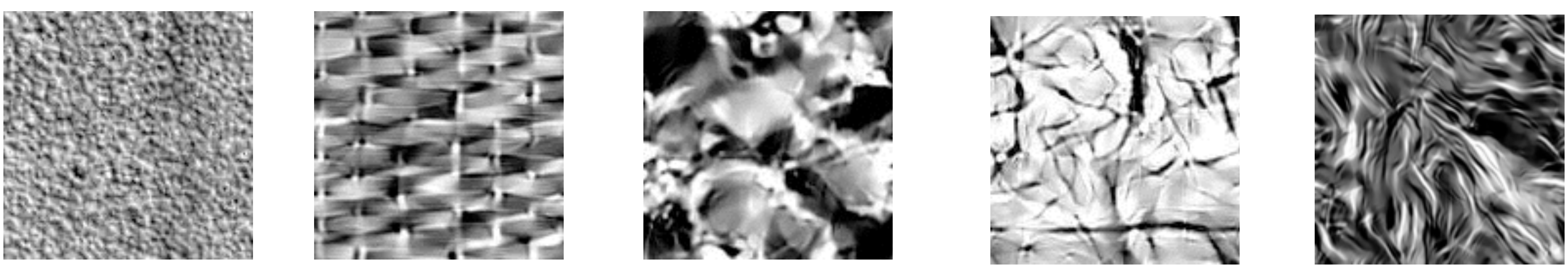


Gaussian process model with N second order moments



Second order Gaussian Scattering: $O(\log N^2)$ moments

$\mathbb{E}(|x \star \psi_{\lambda_1}|) \quad , \quad \mathbb{E}(||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|)$



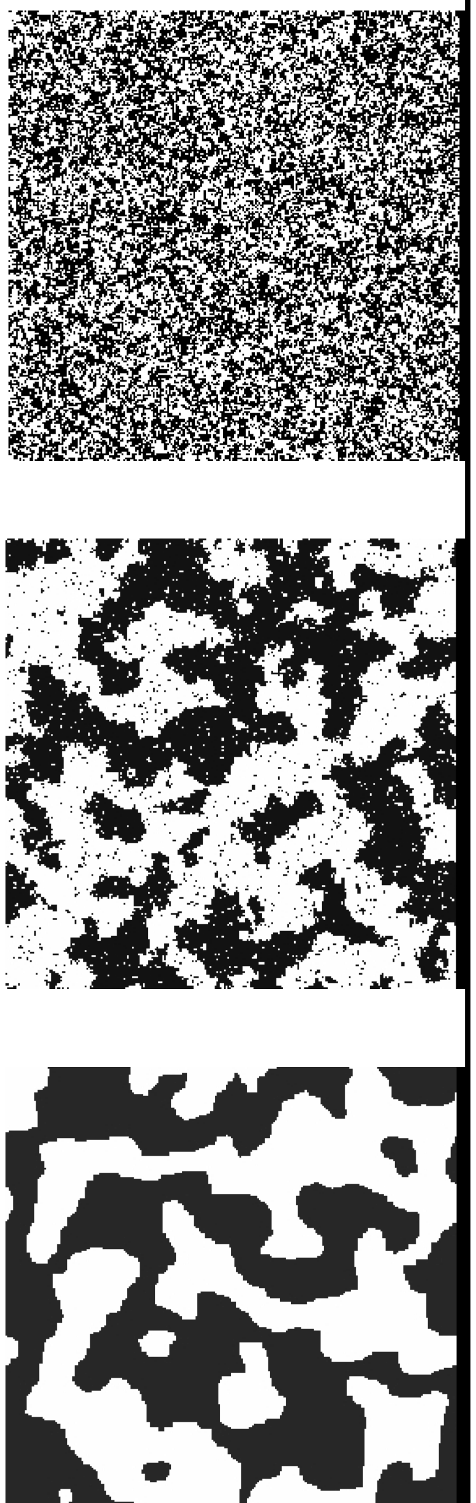
Ising Model and Inverse Problem

Bruna, Dokmanic, Maarten de Hoop

$$p(x) = Z_\beta^{-1} \exp\Big(-\beta \sum_{i,j} J_{i,j}\, x(i)\, x(j) \Big) \quad \text{with} \quad x(i) = \pm 1$$

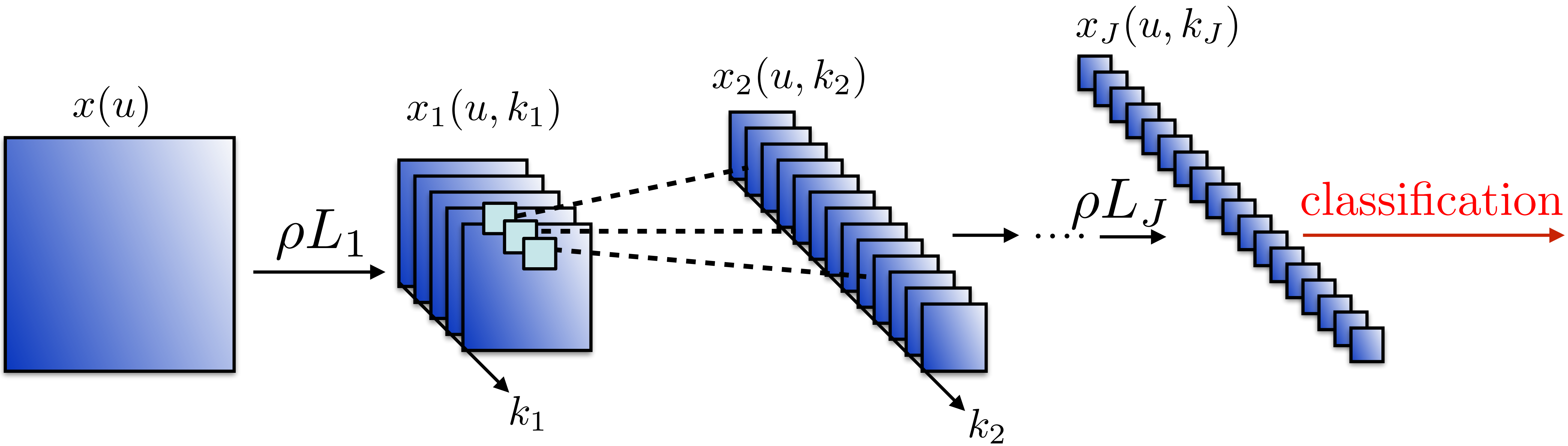
Ising	Gaussian scattering	low-resolution	TV optim.	Scat pred.

β_c



β

Deep Convolutional Trees



$$x_j = \rho \, L_j \, x_{j-1}$$

L_j is composed of convolutions and subs samplings:

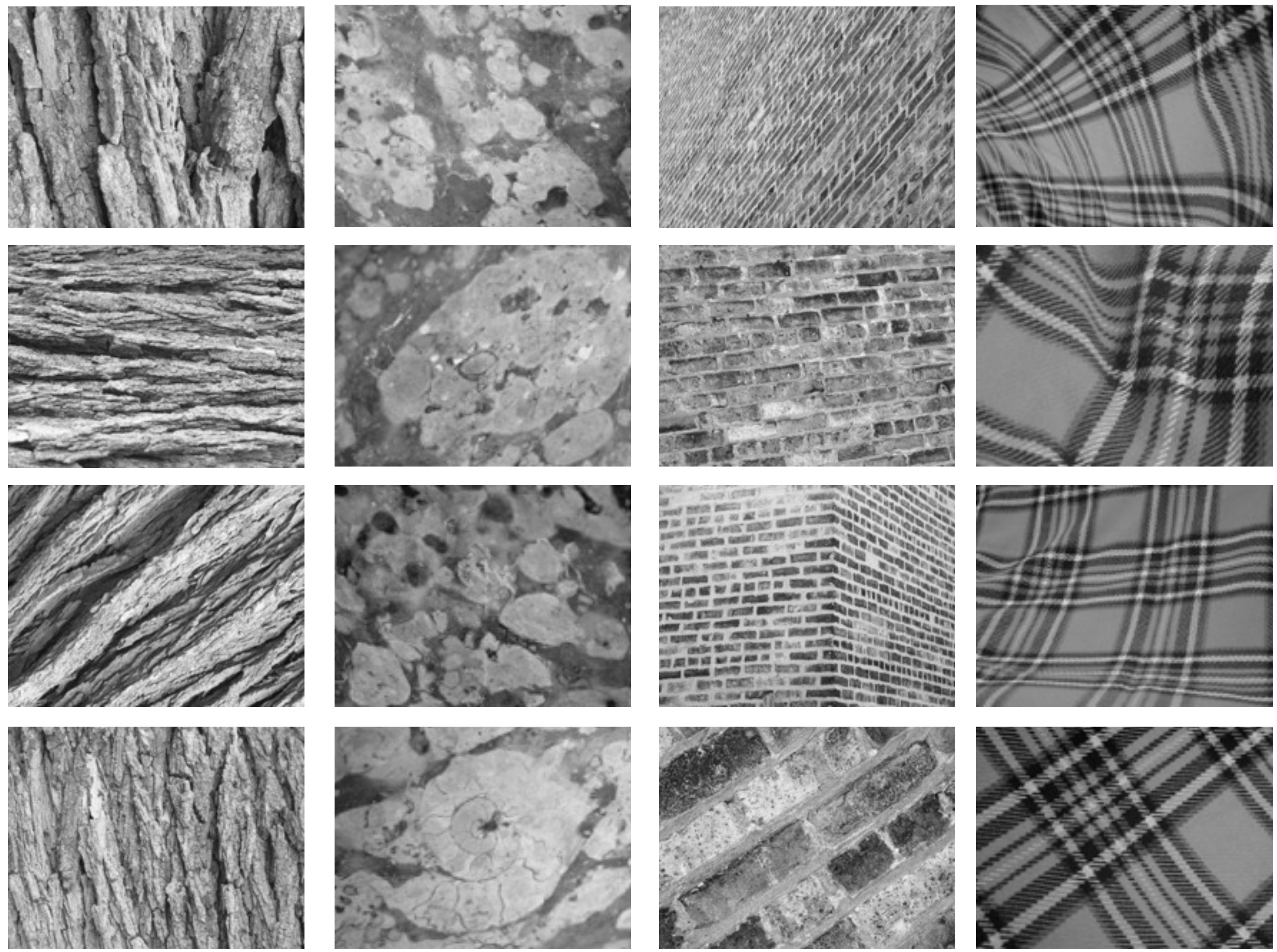
$$x_j(u, k_j) = \rho\Big(x_{j-1}(\cdot, k) \star h_{k_j, k}(u) \Big)$$

No channel communication: what limitations ?

Rotation and Scaling Invariance

Laurent Sifre

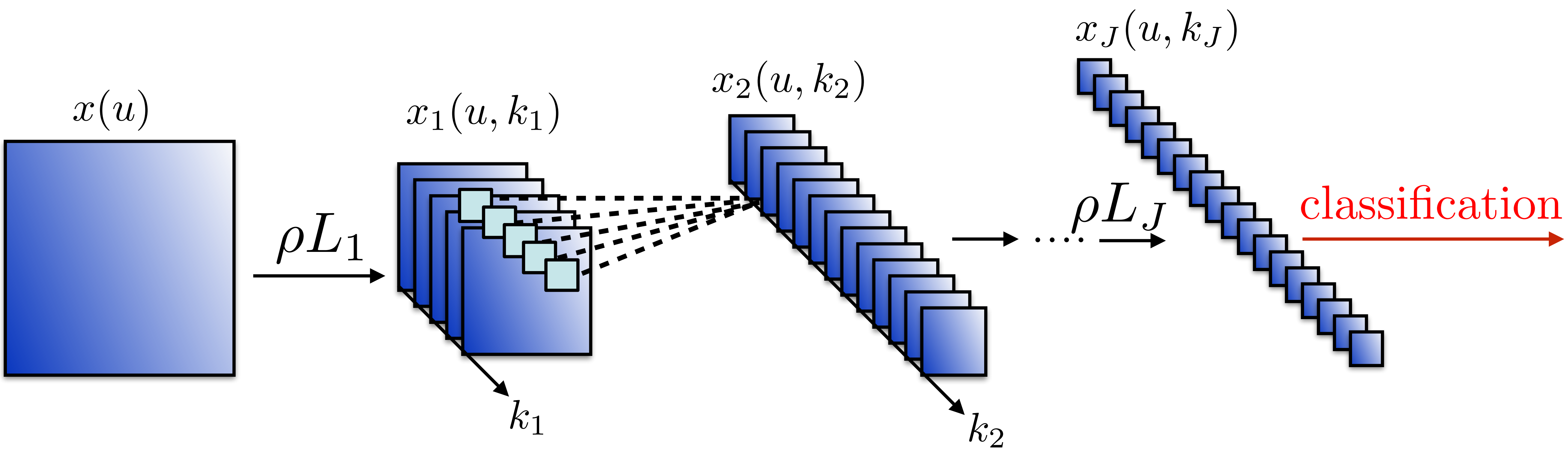
UIUC database:
25 classes



Scattering classification errors

Training	Scat. Translation
20	20 %

Deep Convolutional Networks



$$x_j = \rho \, L_j \, x_{j-1}$$

- L_j is a linear combination of convolutions and subsampling:

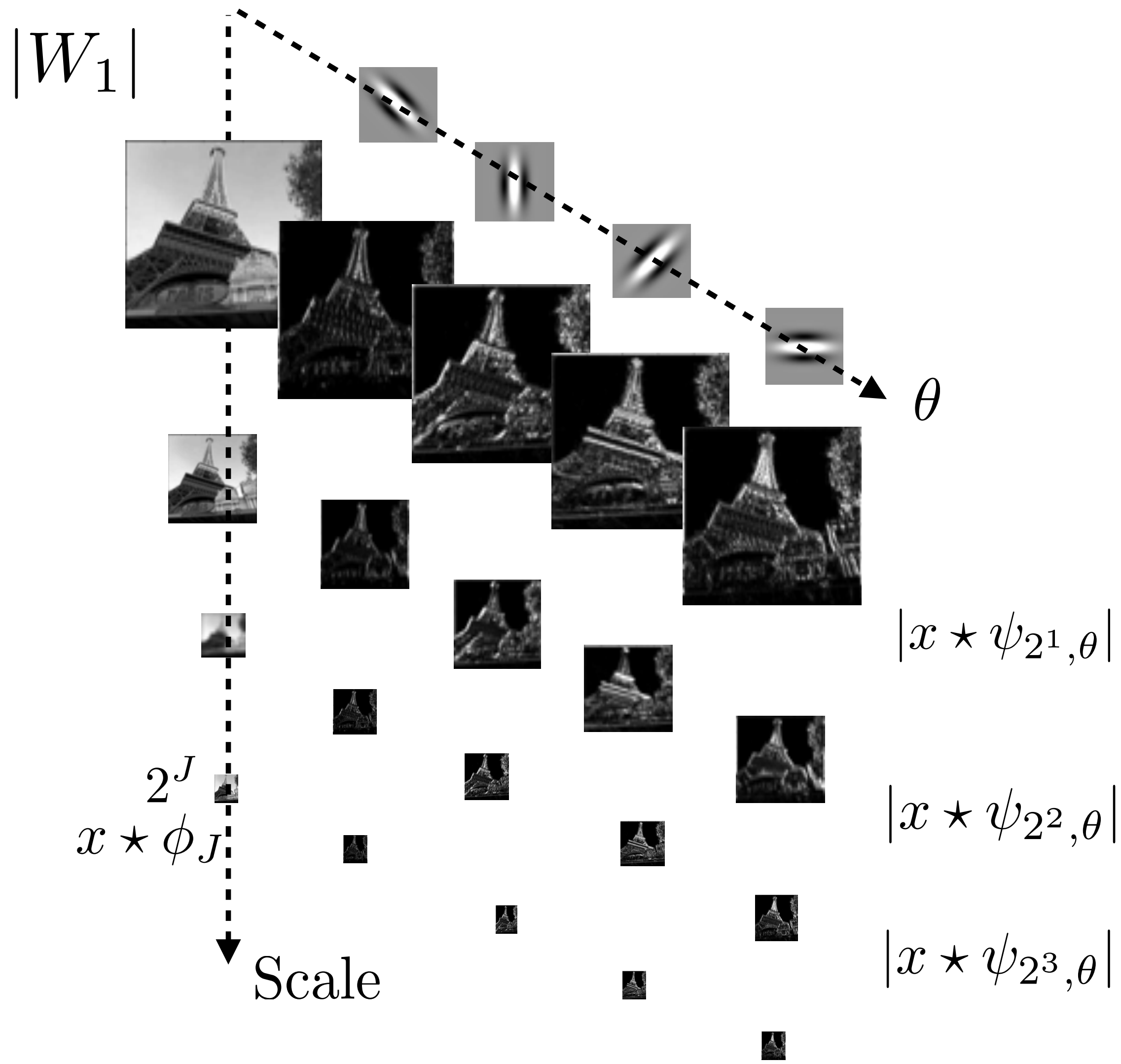
$$x_j(u, k_j) = \rho\Big(\underbrace{\sum_k}_{\text{sum across channels}} x_{j-1}(\cdot, k) \star h_{k_j, k}(u) \Big)$$

What is the role of channel connections ?

Linearize other symmetries beyond translations.

Rotation Invariance

- Channel connections linearize other symmetries.



- Invariance to rotations are computed by convolutions along the rotation variable θ with wavelet filters.

  $\Rightarrow$ invariance to rigid mouvements.

Extension to Rigid Mouvements

Laurent Sifre

Need to capture the variability of spatial directions.

- Group of rigid displacements: translations and rotations

- Action on wavelet coefficients:

rotation & translation

rotation & translation , angle translation

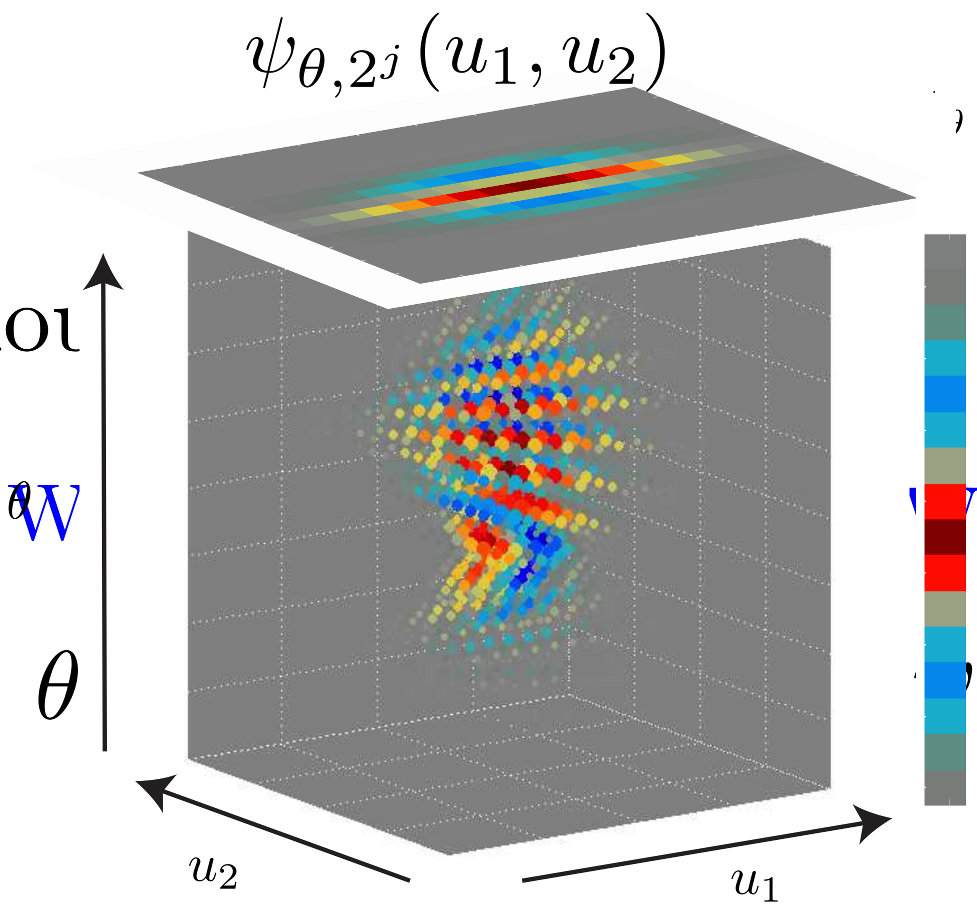
$$
\begin{tikzcd}
x(r_\alpha(u)) \arrow[r] & \boxed{|W_1|} \arrow[r] \arrow[d] & x_j(r_\alpha u, \theta) = |x \star \psi_{2^j,\theta}(u)| \\
& \int x(u)du &
\end{tikzcd}
$$

Extension to Rigid Mouvements

Laurent Sifre

- To build invariants: second wavelet transform on $\mathbf{L^2}(G)$: convolutions of $x_j(u,\theta)$ with wavelets $\psi_{\lambda_2}(u,\theta)$

$$x \circledast \psi_\lambda(u,\theta) = \int_0^{2\pi} \left(\int_{\mathbb{R}^2} x(u',\theta')\, \psi_{\theta,2^j}(r_{-\theta'}(u-u')) \right) \psi_{2^k}(\theta-\theta') d\theta' dt'$$



- Scattering on rigid mou

Wavelets on Translations W Wavelets on Rigid Mvt.

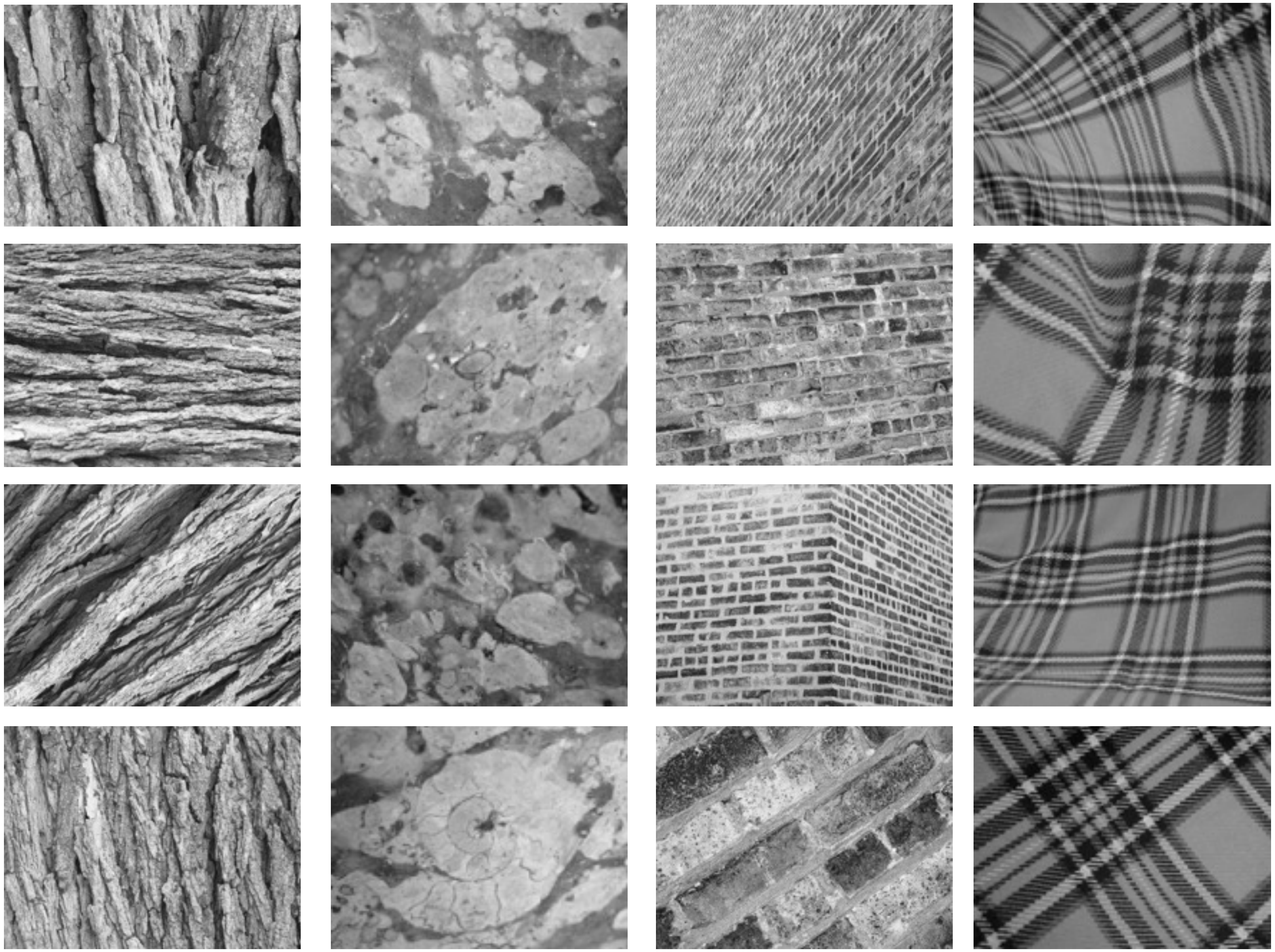
$x(u) \longrightarrow \boxed{|W_1|} \longrightarrow x_j(u, \theta$ $\quad |\psi_{\lambda_2}(v,\theta)| \longrightarrow \boxed{|W_3|}$

$\int x(u)du \quad \int x_j(u,\theta)\, du d\theta \quad \int |x_j \circledast \psi_{\lambda_2}(v,\theta)| du d\theta$

Rotation and Scaling Invariance

Laurent Sifre

UIUC database:
25 classes



Scattering classification errors

Training	Scat. Translation	Scat. Rigid Mouvt.
20	20 %	**0.6**%

Learning Physics: N-Body Problem

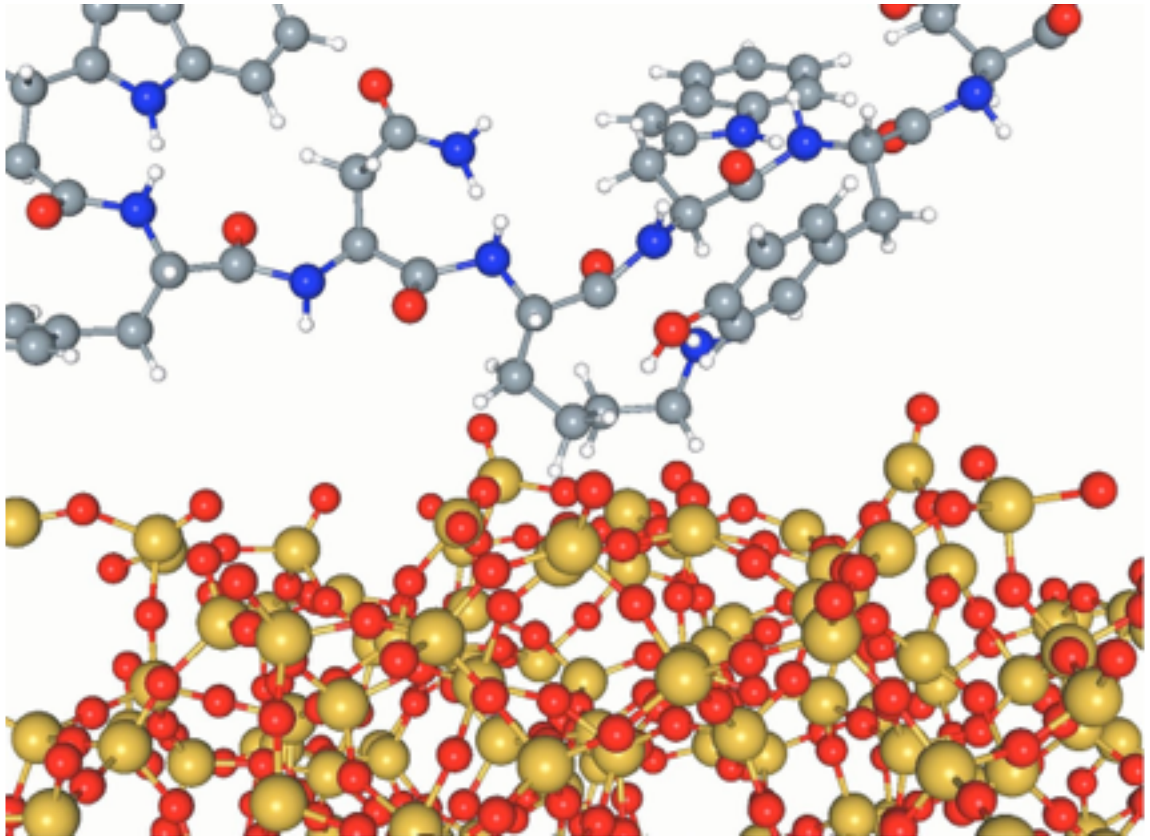
N. Poilvert
Matthew Hirn

- Energy of d interacting bodies:

  Can we learn the interaction energy $f(x)$ of a system with $x = \Big\{ \text{positions}, \text{values} \Big\}$?

Astronomy

Quantum Chemistry



Density Functional Theory

Kohn-Sham model:

$$E(\rho) = T(\rho) + \int \rho(u)\, V(u) + \frac{1}{2}\int \frac{\rho(u)\rho(v)}{|u-v|} du dv + E_{xc}(\rho)$$

Molecular energy	Kinetic energy	electron-nuclei attraction	electron-electron Coulomb repulsion	Exchange correlat. energy

At equilibrium:

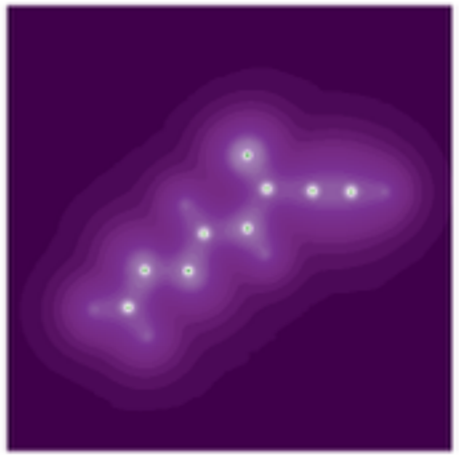
$$f(x) = E(\rho_x) = \min_{\rho} E(\rho)$$

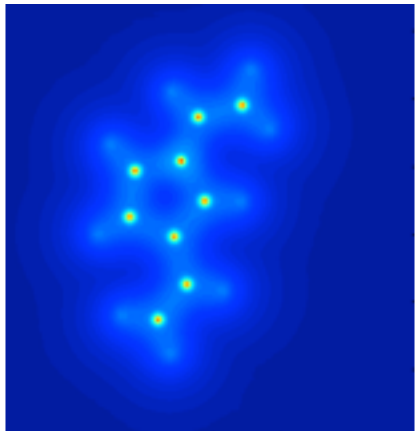
Quantum Chemistry Invariants

Quantum chemistry: $f(x)$ is invariant to rigid mouvements, stable to deformations.

Depends on the true electronic density (Kohn-Sham)

Ground state electronic density computed with Schroedinger





- Can we estimate $f(x)$ from a naive electronic density ?

Density $\tilde{\rho}_x$ computed as a sum of blobs

Quantum Regression

N. Poilvert
Matthew Hirn

- Linear regressions computed with invariant change of variables:

$$\Phi x = \{\phi_n(\tilde{\rho}_x)\}_n \; : \left| \begin{array}{c} \text{Fourier modulus coefficients and squared} \\ \text{or} \\ \text{scattering coefficients and squared} \end{array} \right.$$

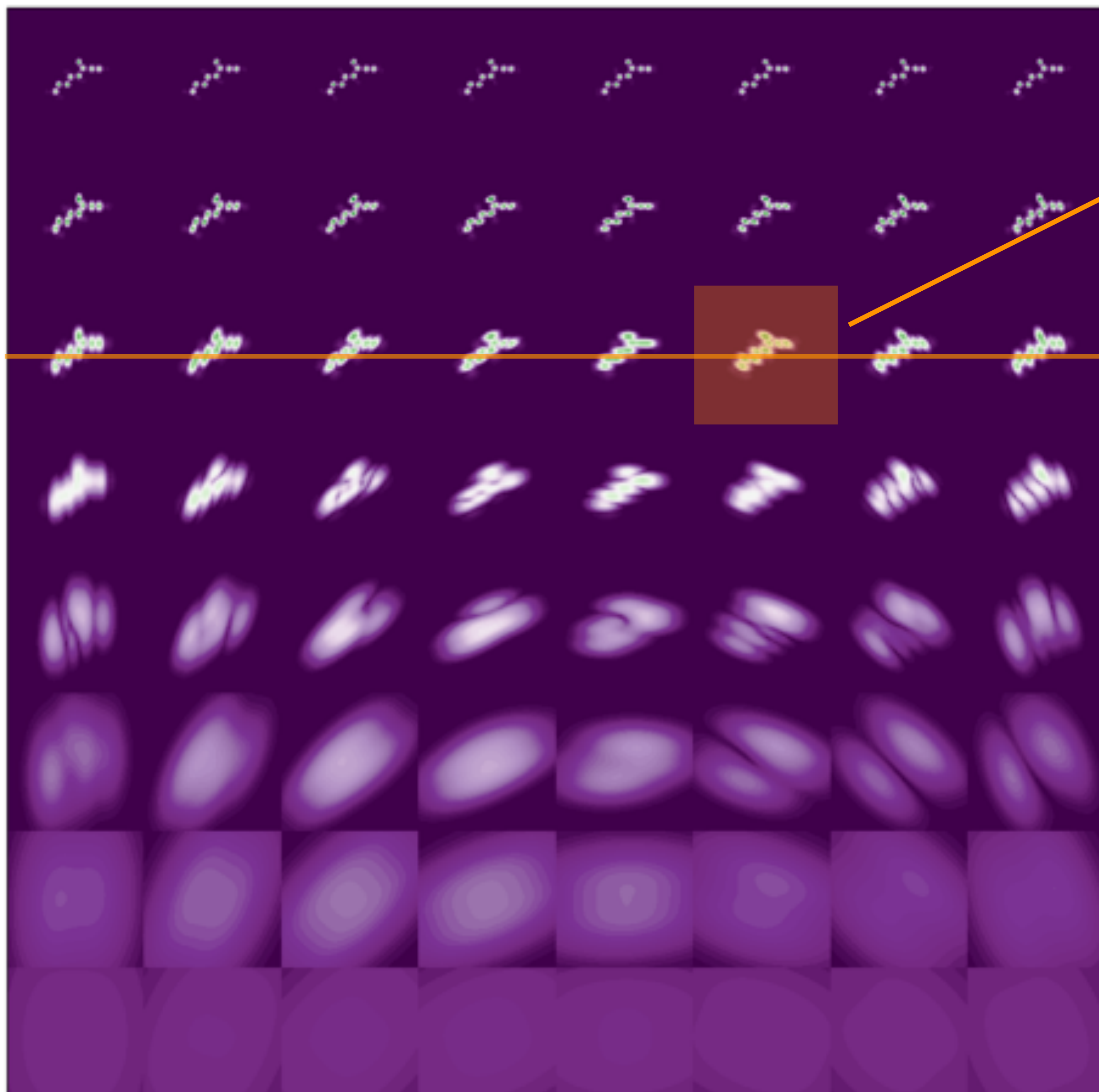
$$f_M(x) = \sum_{k=1}^{M} w_k \, \phi_{n_k}(\tilde{\rho}_x)$$

Regression coefficients w_k: equivalent potential.

Scattering Dictionary

$\rho(u)$

$|\rho * \psi_{j_1,\theta_1}(u)|$



Scales j_1

Rotations θ_1

2nd Order Interferences

Recover translation variability:

$$|\rho * \psi_{j_1,\theta_1}| * \psi_{j_2,\theta_2}(u)$$

Recover rotation variability:

$$|\rho * \psi_{j_1,\cdot}(u)| \circledast \overline{\psi}_{l_2}(\theta_1)$$

Combine to recover
roto-translation variabiltiy:

$$||\rho * \psi_{j_1,\cdot}| * \psi_{j_2,\theta_2}(u) \circledast \overline{\psi}_{l_2}(\theta_1)|$$

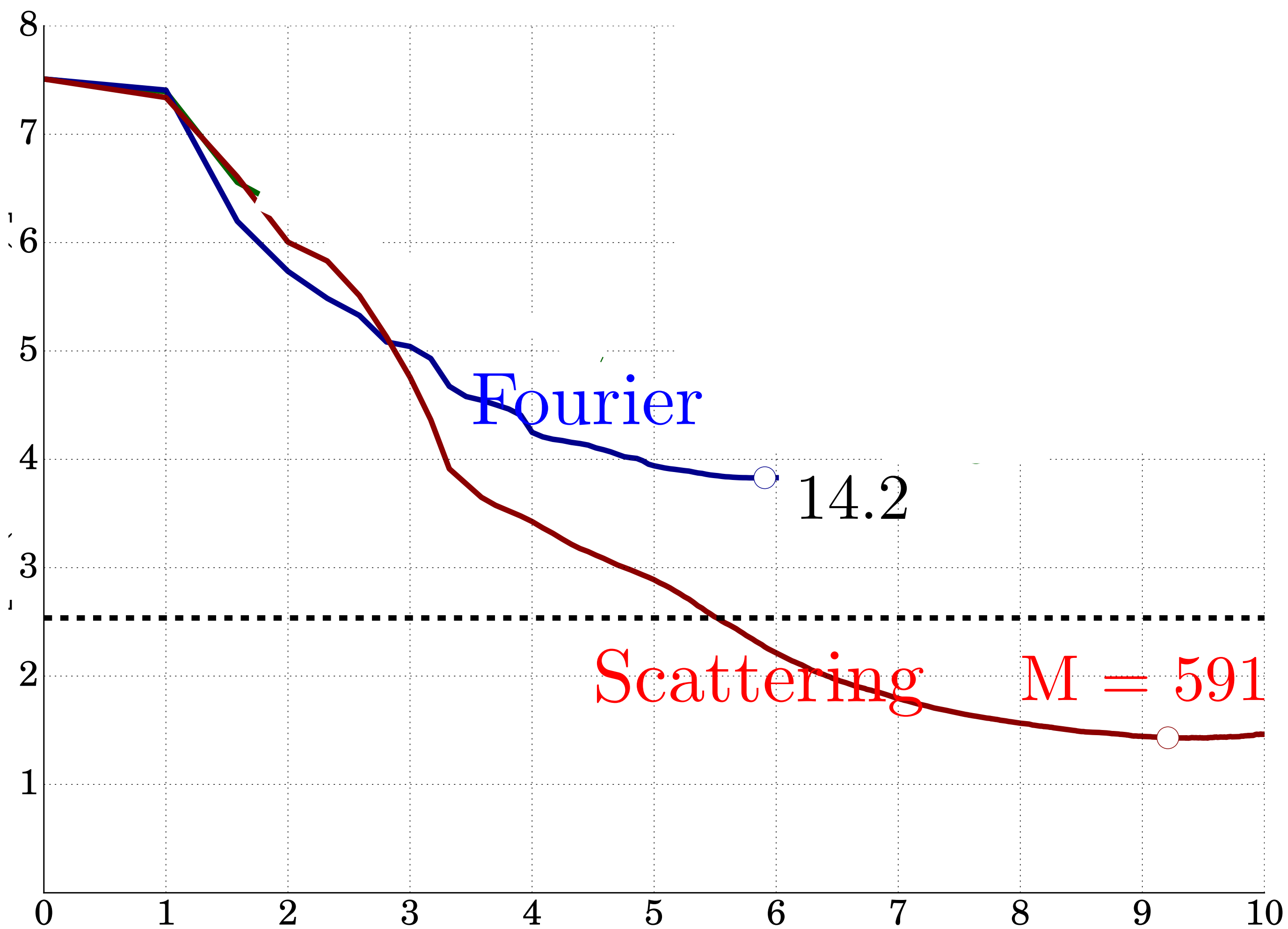
Scattering Regression

Data basis $\{x_i\,,\,f(x_i)\}_{i \leq N}$ of 4357 planar molecules

$$\text{Regression:} \quad f_M(x) = \sum_{m=1}^{M} w_m\, \phi_{k_m}(\tilde\rho_x)$$

Interaction terms
across scales

Testing error
$2^{-1} \log_2 \mathbb{E}|f_M(x) - y(x)|^2$



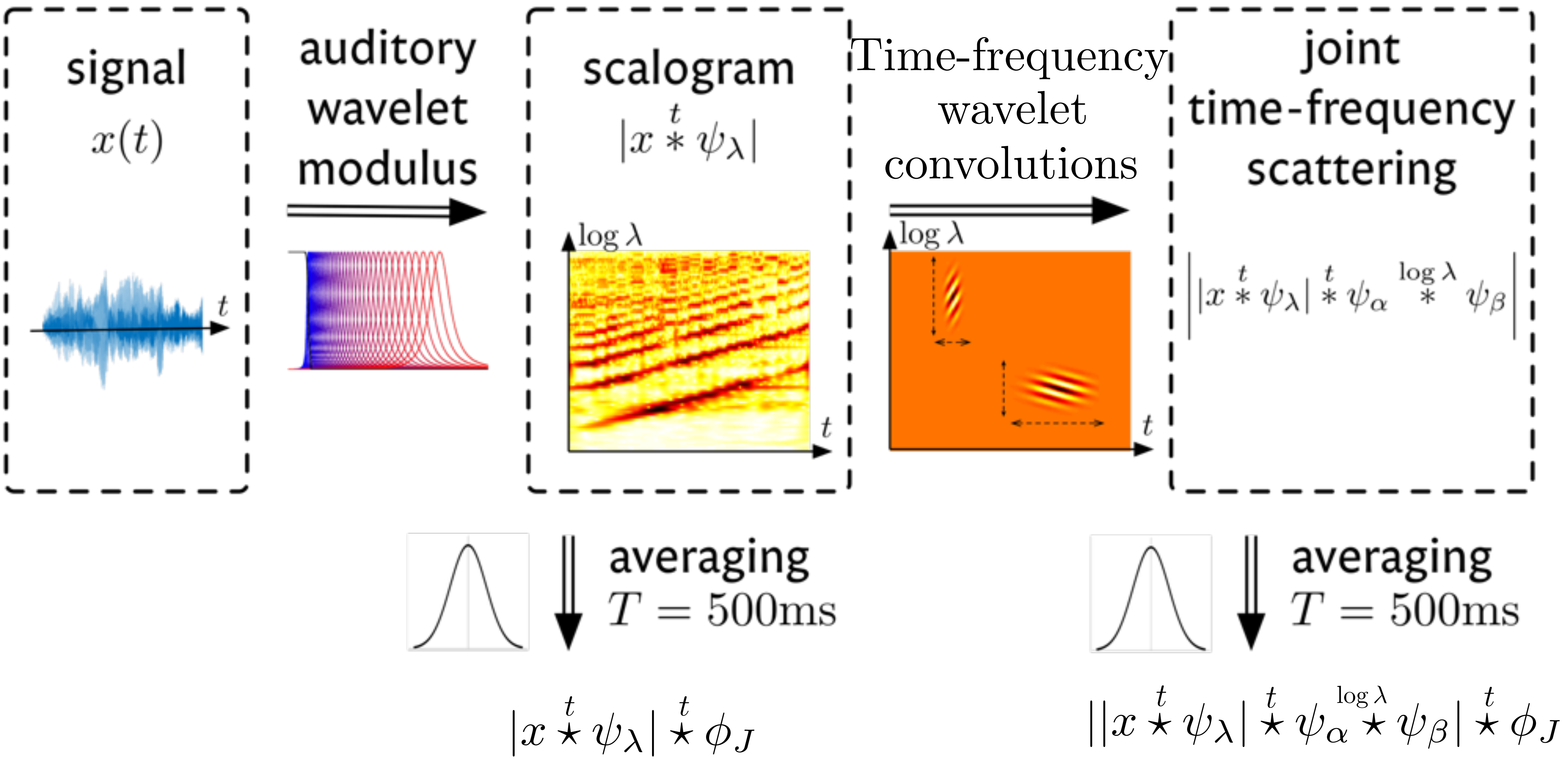
5.8 : State of the art

1.8 kcal/mol

$\log_2 M$

Time-Frequency Translation Group

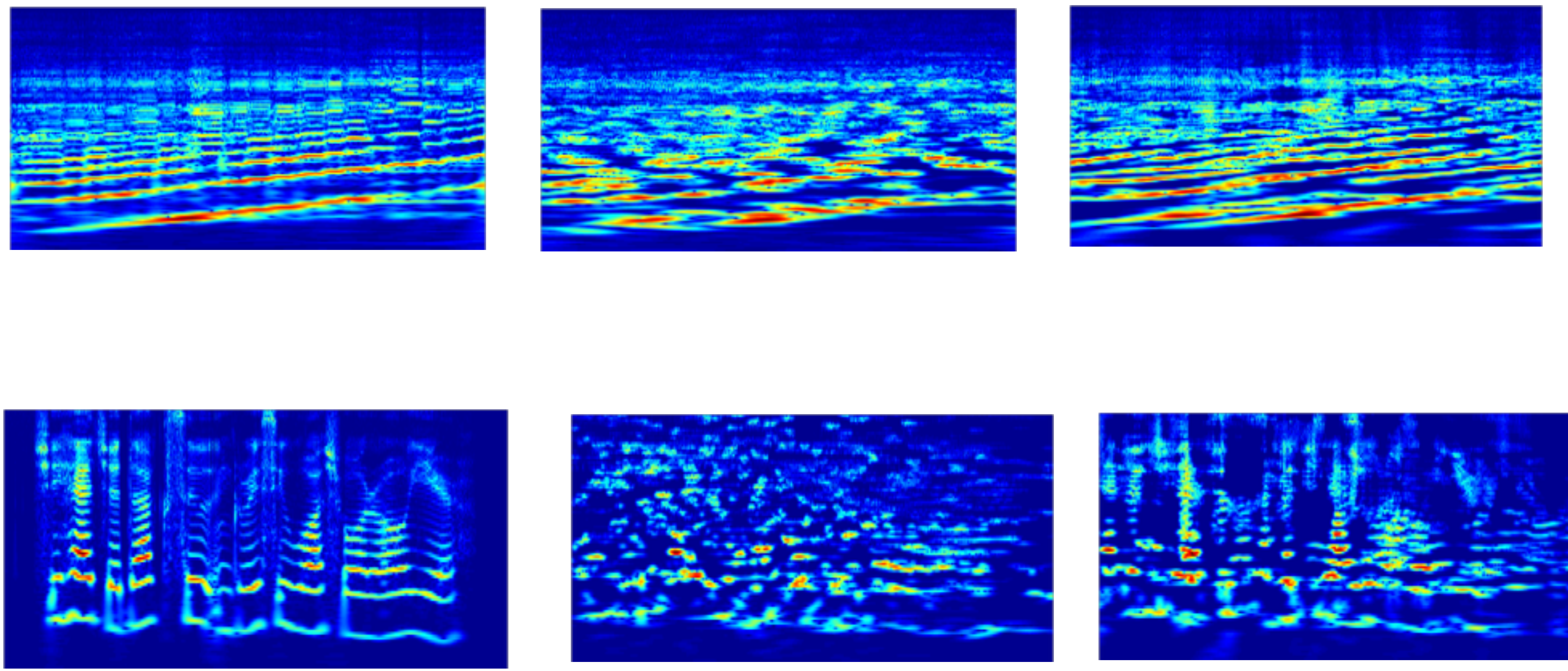
J. Anden and V. Lostanlen



Joint Time-Frequency Scattering

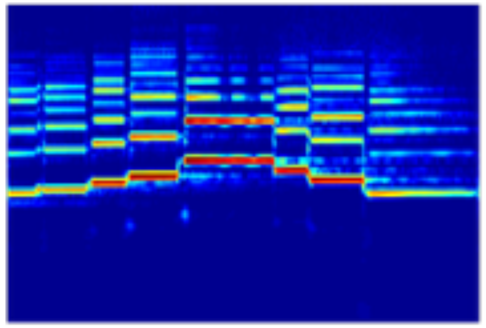
J. Anden and V. Lostanl

Original | Time Scattering | Time/Freq Scattering

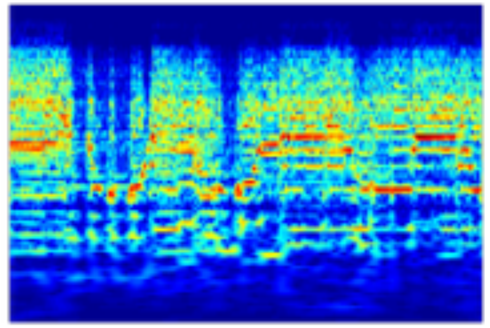


Musical Instrument Classificaiton

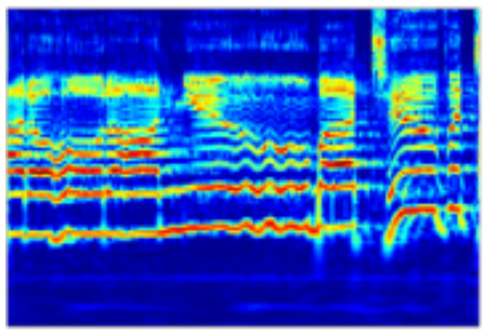
clarinet



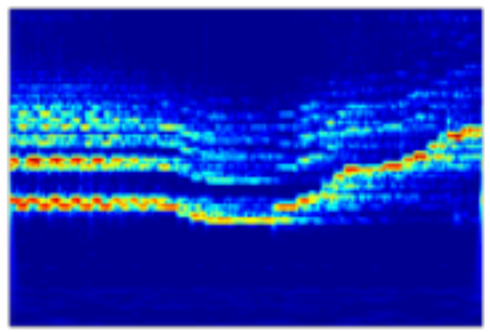
electric guitar



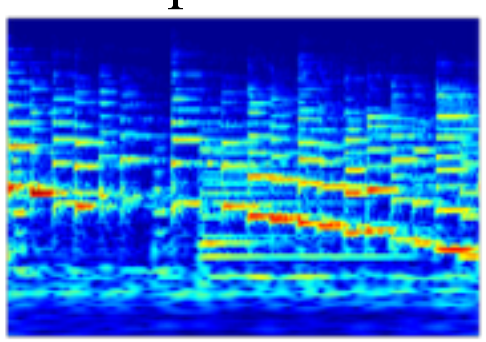
female singer



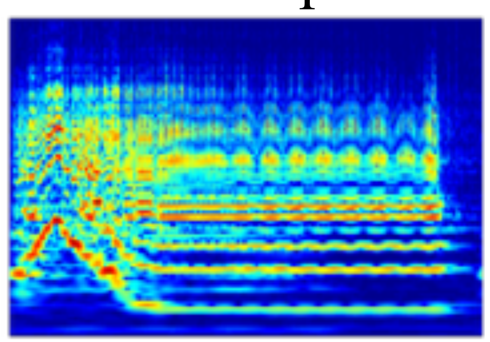
flute



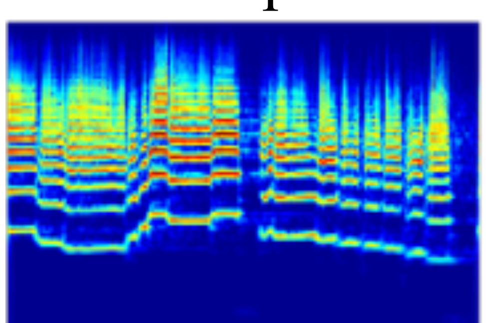
piano



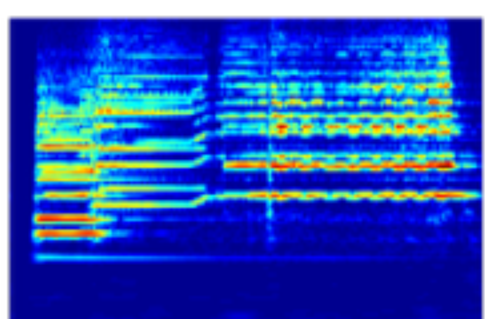
tenor saxophone



trumpet



violin



J. Anden and V. Lostanlen

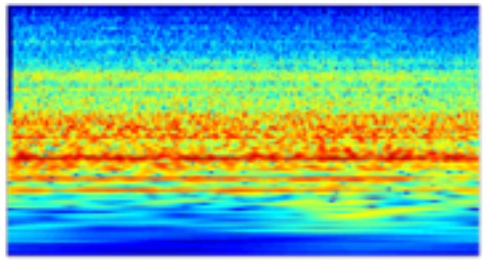
MedleyDB: 8 classes
10k training examples

class-wise average error

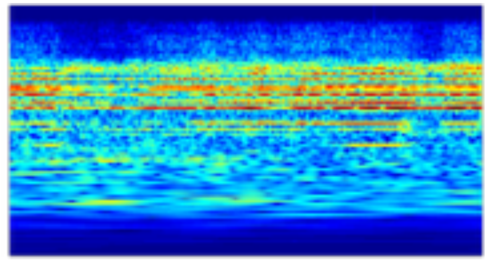
MFCC audio descriptors	0,39
time scattering	0,31
ConvNet	0,31
time-frequency scattering	0,18

Environmental Sound Classification

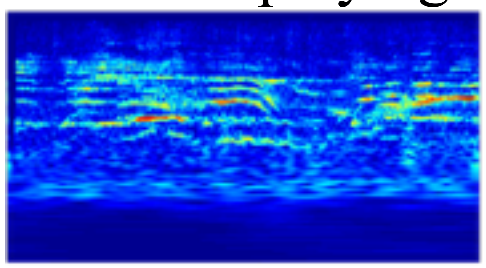
air conditioner



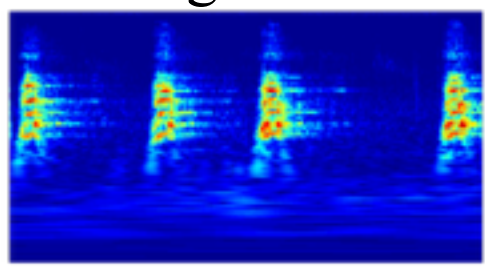
car horns



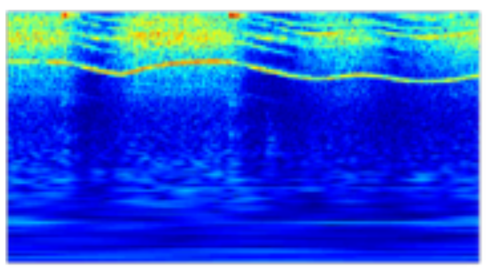
children playing



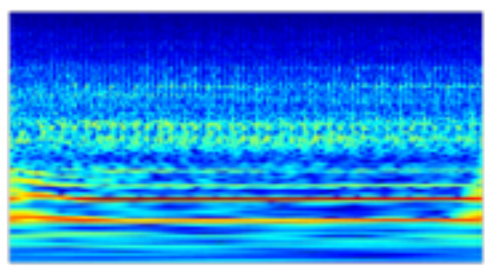
dog barks



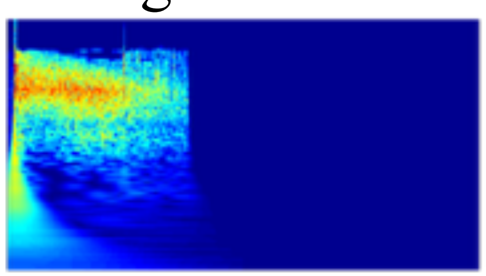
drilling



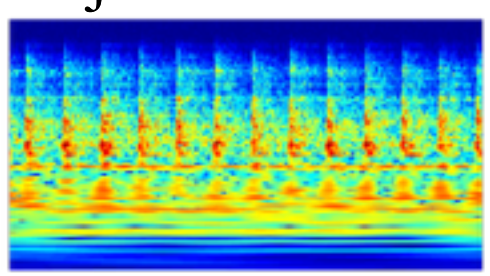
engine at idle



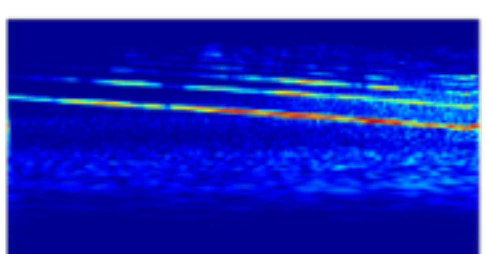
gunshot



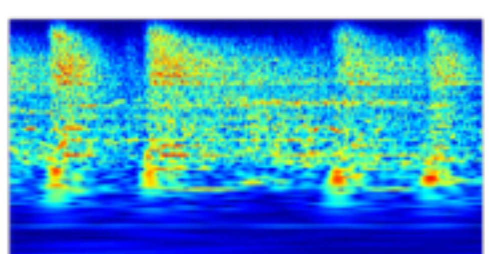
jackhammer



siren



street music



J. Anden and V. Lostanlen

UrbanSound8k: 10 classes
8k training examples

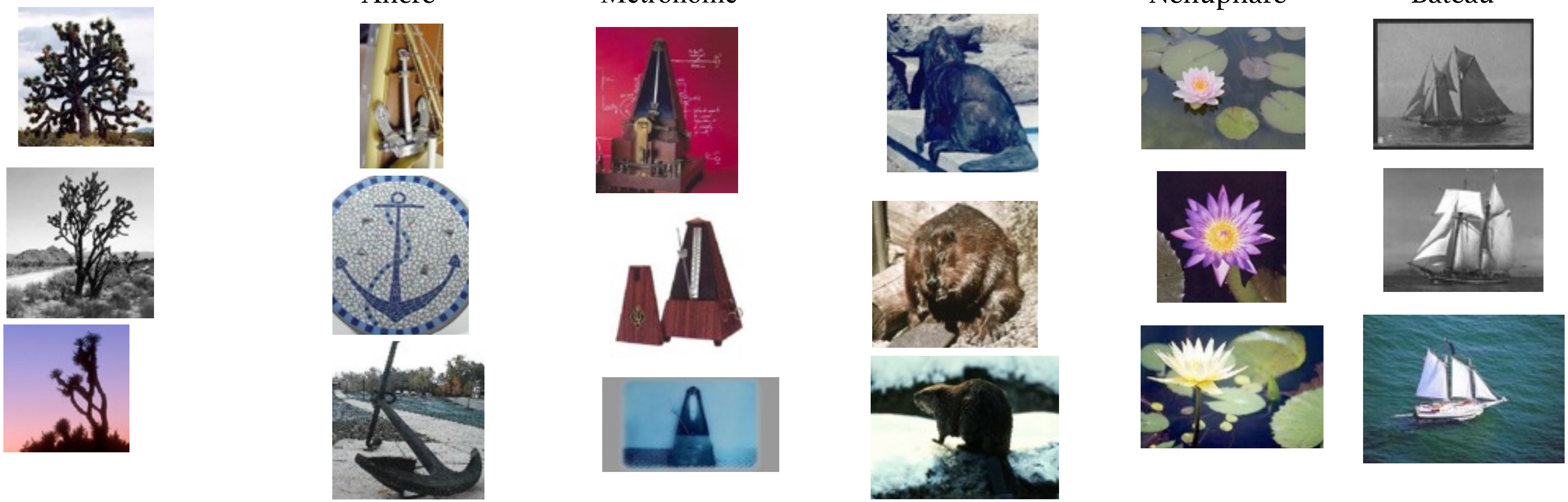
class-wise average error

MFCC audio descriptors	0,39
time scattering	0,27
ConvNet (Piczak, MLSP 2015)	0,26
time-frequency scattering	0,2

Complex Image Classification

Edouard Oyallon

Arbre de Joshua | Ancre | Metronome | Castore | Nénuphare | Bateau



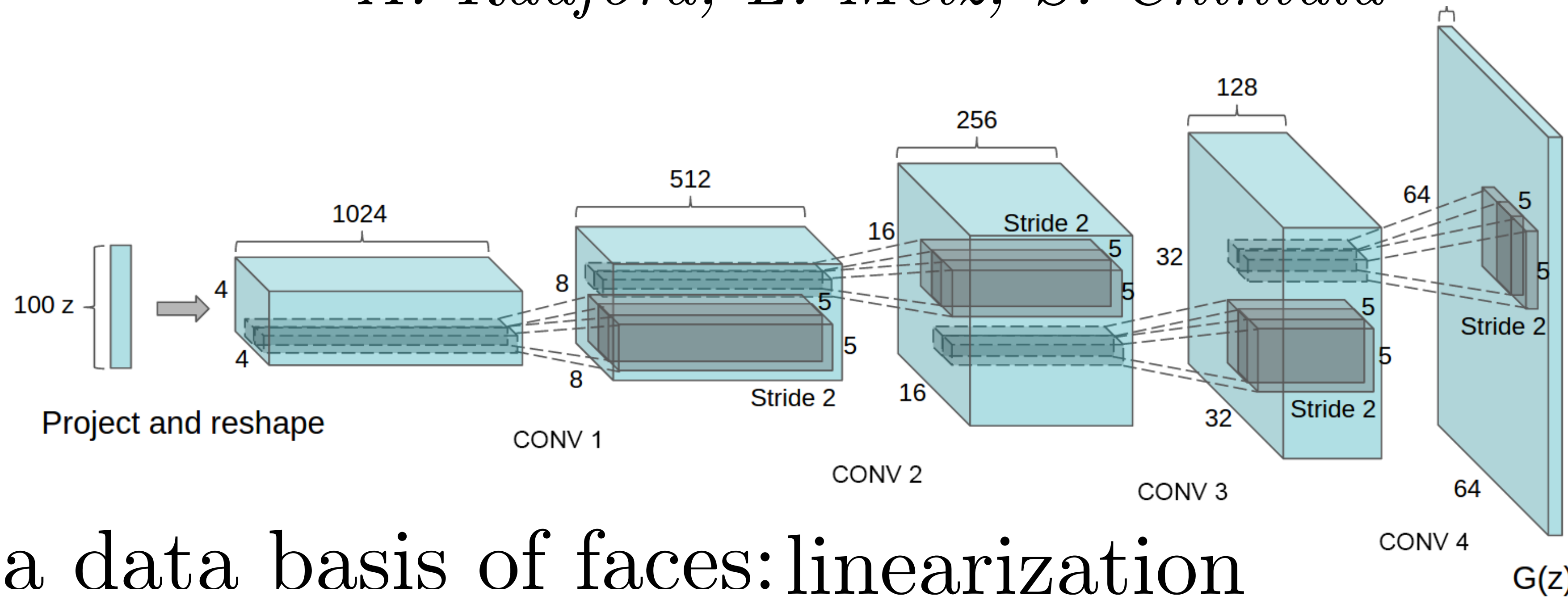
$$x \longrightarrow \boxed{S_J x} \longrightarrow \boxed{\text{Supervised Linear classifier}} \longrightarrow y = f(x)$$

No learning

Data Basis	Deep-Net	Scat/Unsupervised
CIFAR-10	**7%**	20%

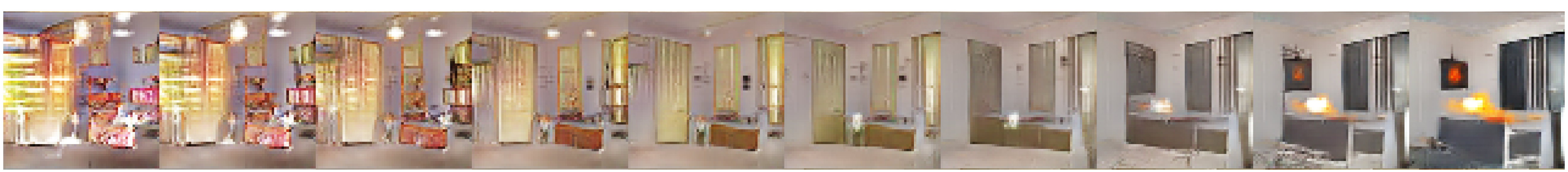
Linearisation in Deep Networks

A. Radford, L. Metz, S. Chintala

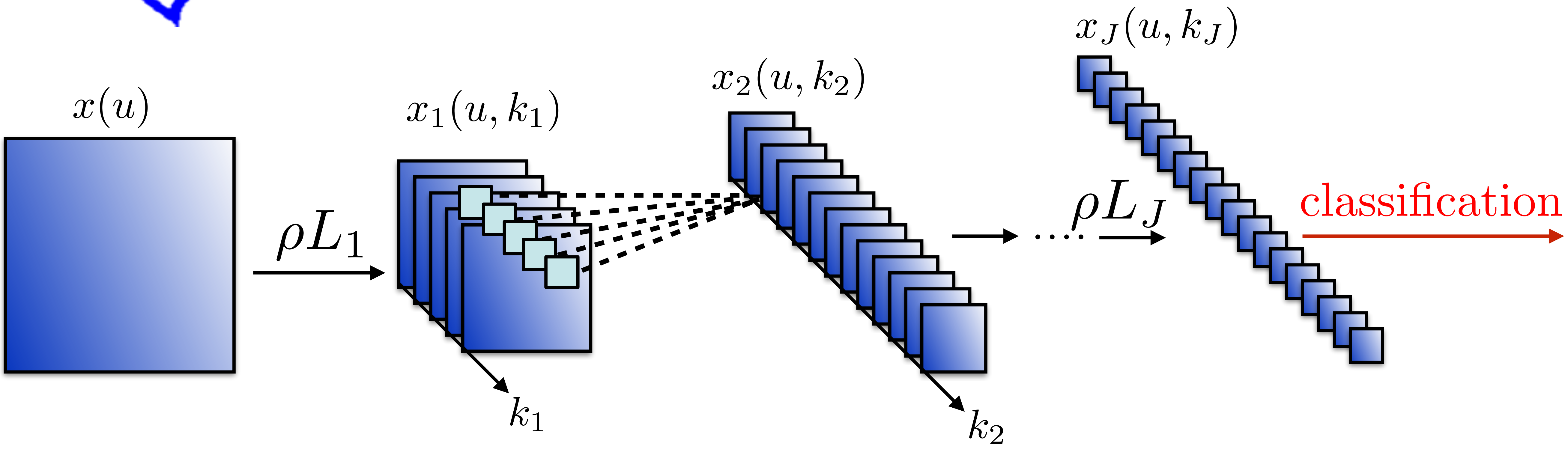


- Trained on a data basis of faces: linearization

- On a data basis including bedrooms: interpolaitons



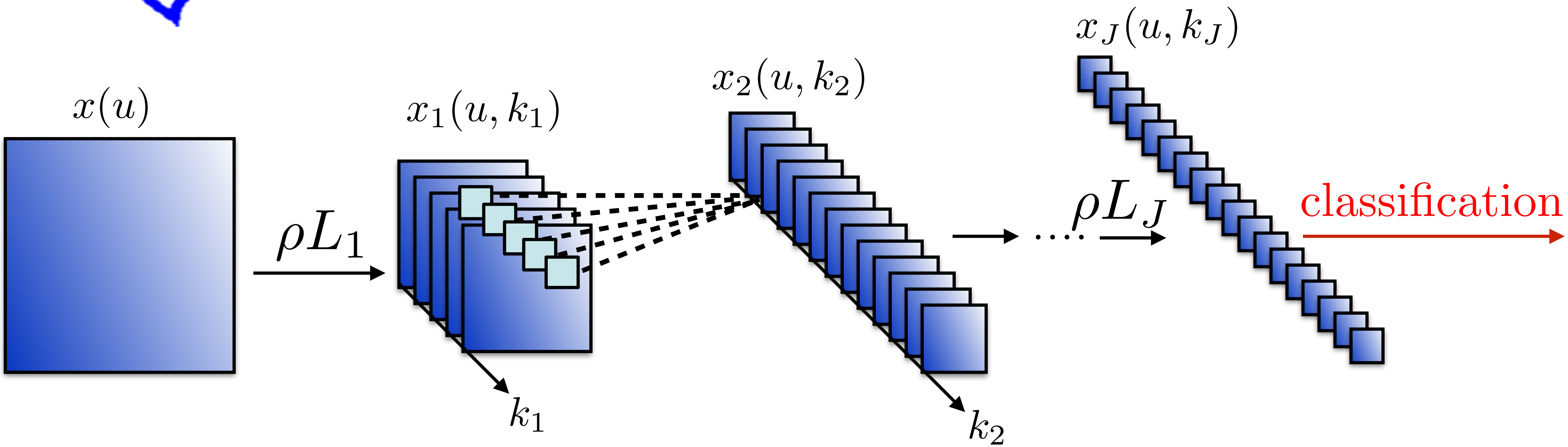
Deep Convolutional Networks



- The convolution network operators L_j have many roles:
  - Linearize non-linear transformations (symmetries)
  - Reduce dimension with projections
  - Memory storage of « characteristic » structures

- Difficult to separate these roles when analyzing learned networks

Open Problems



- Can we recover symmetry groups from the matrices *Lj* ?
- What kind of groups ?
- Can we characterise the regularity of *f(x)* from these groups ?
- Can we define classes of high-dimensional « regular » functions that are well approximated by deep neural networks ?
- Can we get approximation theorems giving errors depending on number of training exemples, with a fast decay ?

Conclusions

- Deep convolutional networks have spectacular high-dimensional approximation capabilities.
- Seem to compute hierarchical invariants of complex symmetries
- Used as models in physiological vision and audition
- Close link with particle and statistical physics
- Outstanding mathematical problem to understand them:

  notions of complexity, regularity, approximation theorems…

Understanding Deep Convolutional Networks, arXiv 2016.