Mathematical Mysteries of Deep Neural Networks

Joan Bruna, Stéphane Mallat,
Edouard Oyallon, Vincent Lostanlen

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

High Dimensional Learning

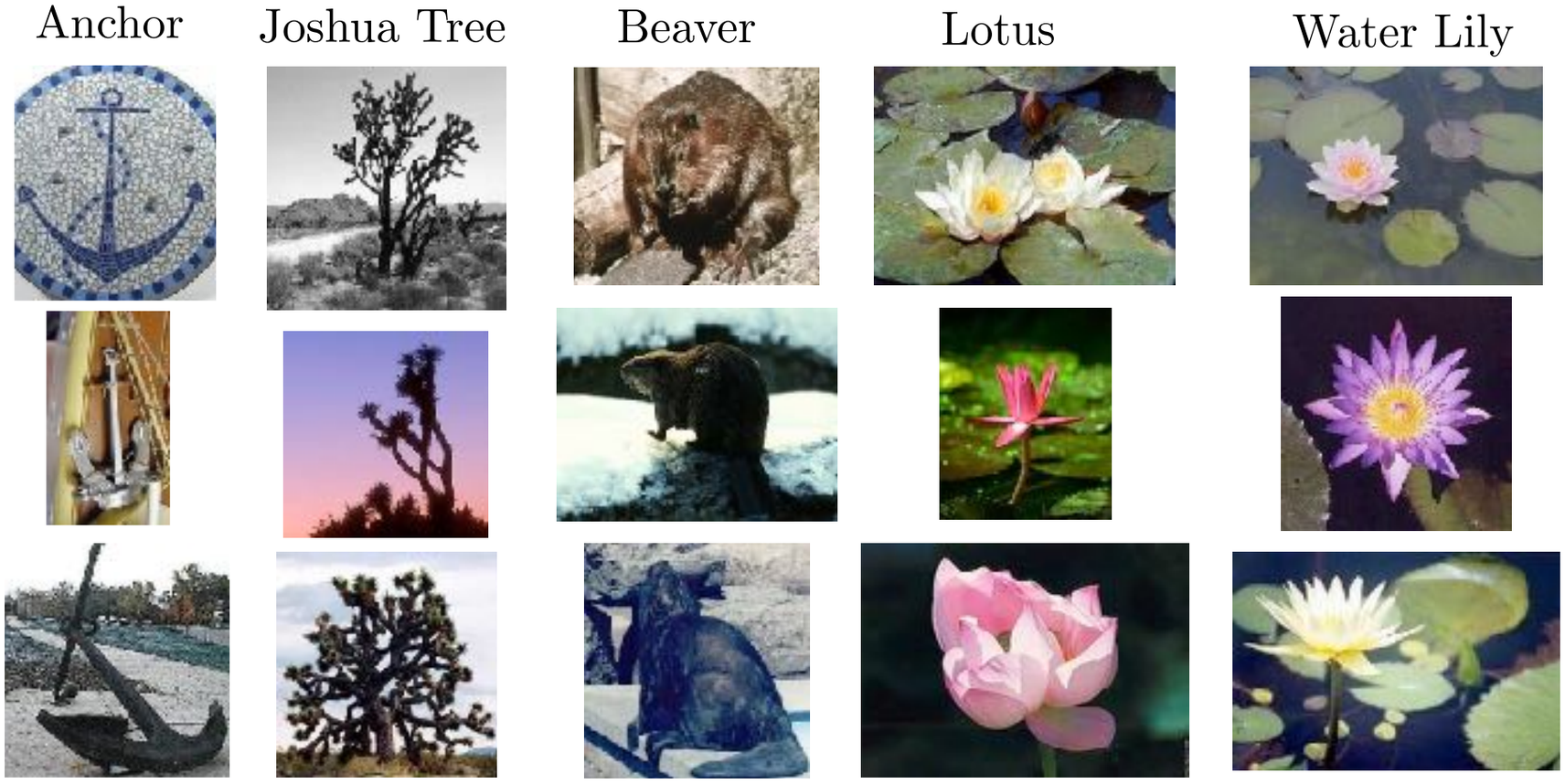
- High-dimensional $x = (x(1), ..., x(d)) \in \mathbb{R}^d$:

- **Classification:** estimate a class label $f(x)$

  given n sample values $\{x_i\,,\, y_i = f(x_i)\}_{i \leq n}$

Image Classification $d = 10^6$

Huge variability inside classes



Find invariants

High Dimensional Learning

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

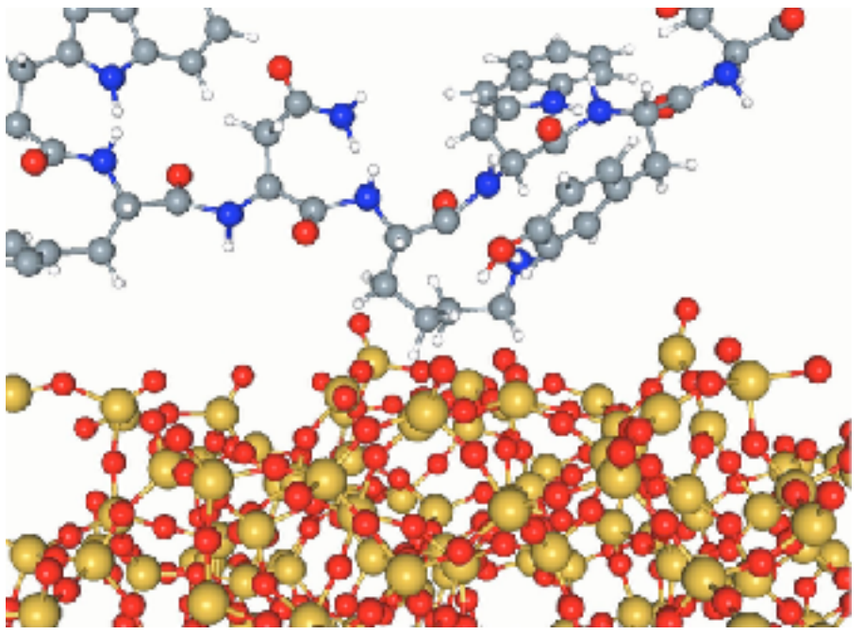
- **Regression:** approximate a *functional* $f(x)$

  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

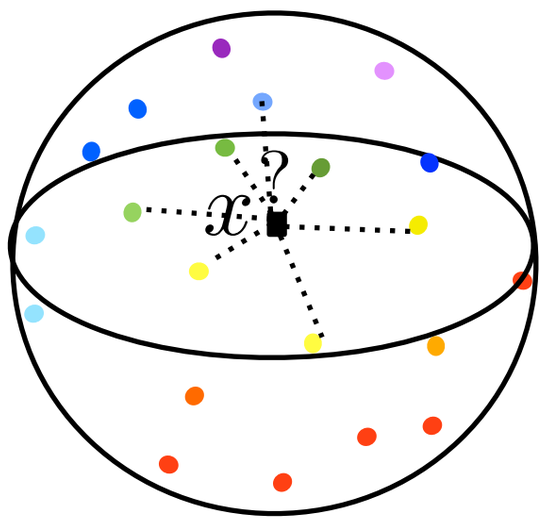
Quantum Chemistry



Importance of symmetries.

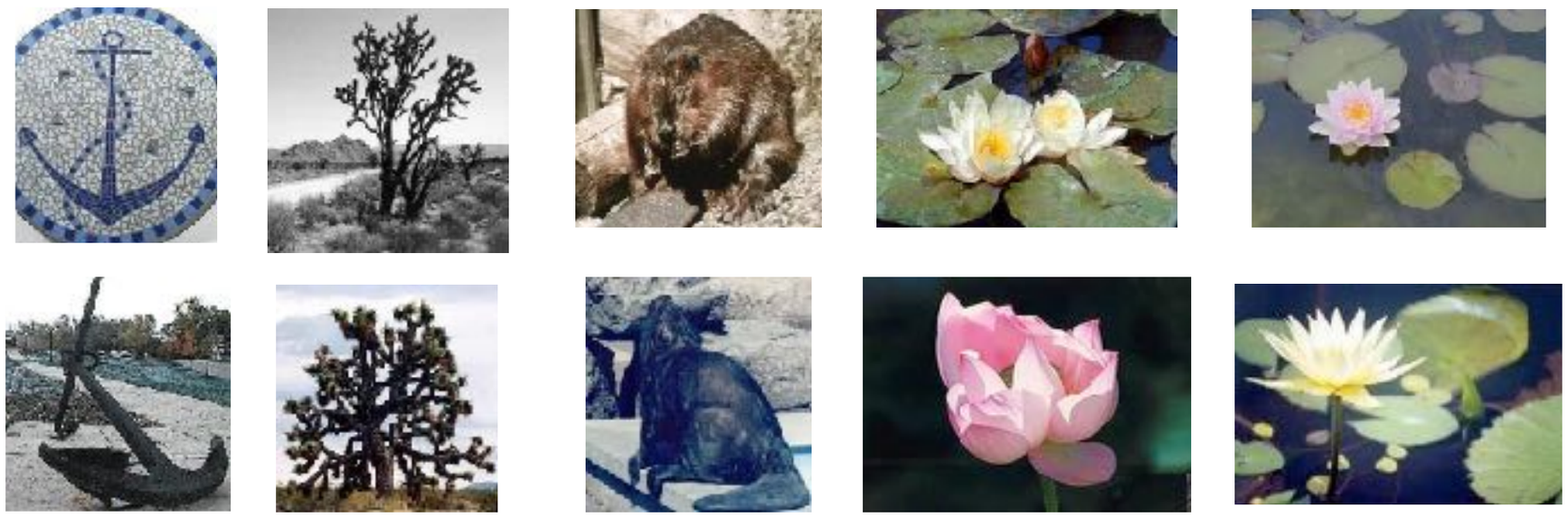
Curse of Dimensionality

- $f(x)$ 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



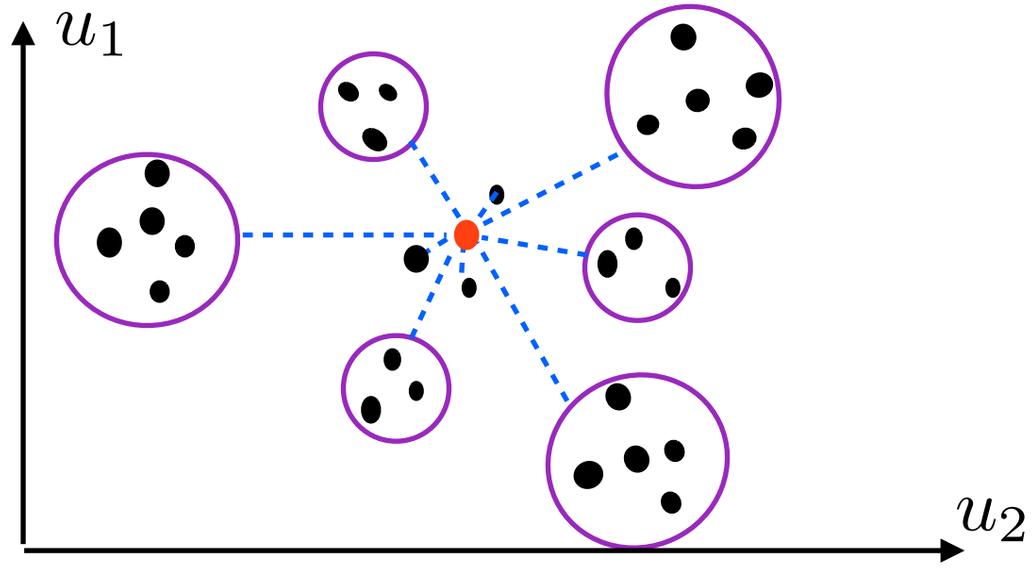
- Why can we approximate f ?

Dimensionality Reduction
Multiscale

- Why can we learn despite the curse of dimensionality ?

Multiscale structures/interactions

Interactions de d variables $x(u)$: pixels, particules, agents...



Regroupement of d interactions in $O(\log d)$

Kernel Classifiers

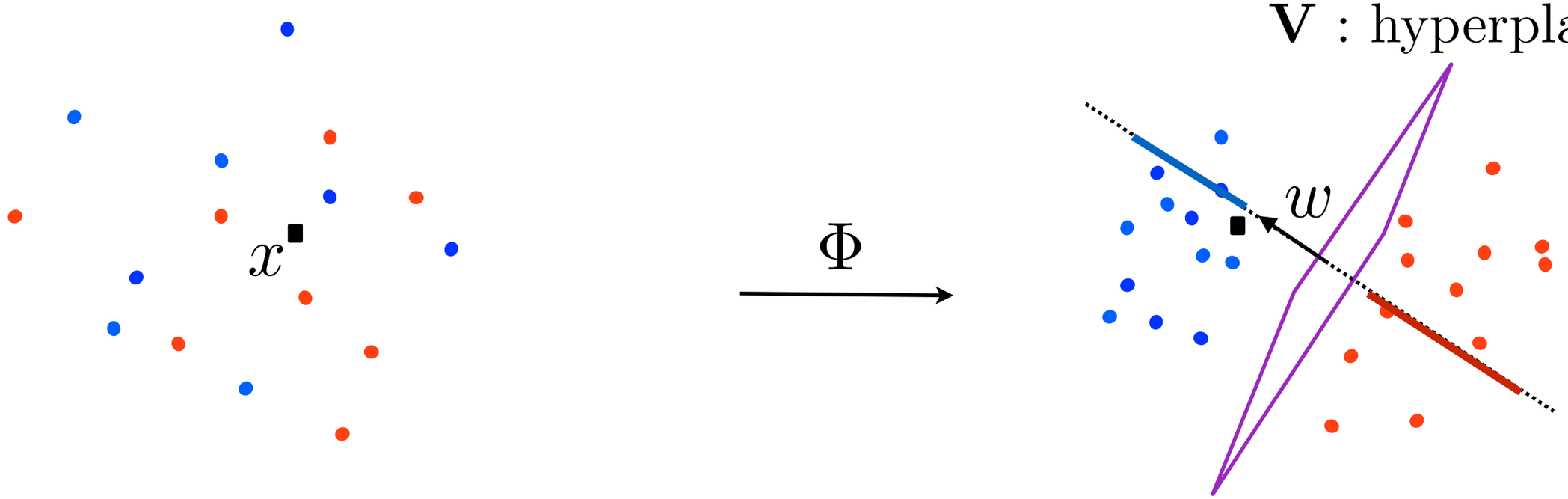
1. Find a change of variable $\Phi(x) = \{\phi_k(x)\}_{k \leq d'}$: linearization separation

2. Linear projection $\langle \Phi(x), w \rangle = \sum_k w_k \, \phi_k(x)$: invariant.

Data: $x \in \mathbb{R}^d$

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

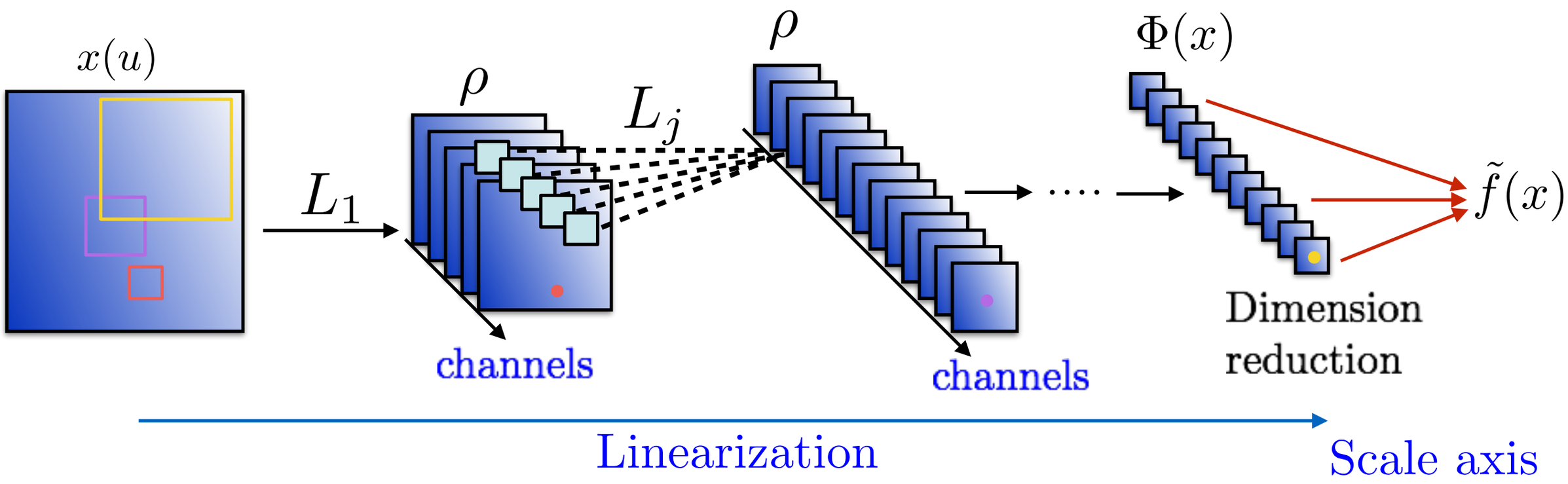
$\mathbf{V}$: hyperplane



- How and when is possible to find such a Φ ?

Deep Convolution Neworks

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



L_j: sums of linear convolutions $\quad \rho(\alpha) = \max(\alpha, 0)\ ,\ |\alpha|\ ,\ \arctan(\alpha)$

Optimize L_j by propagation of errors on training exemples

$$\text{Training error} = \sum_i |\tilde{f}(x_i) - f(x_i)|^2$$

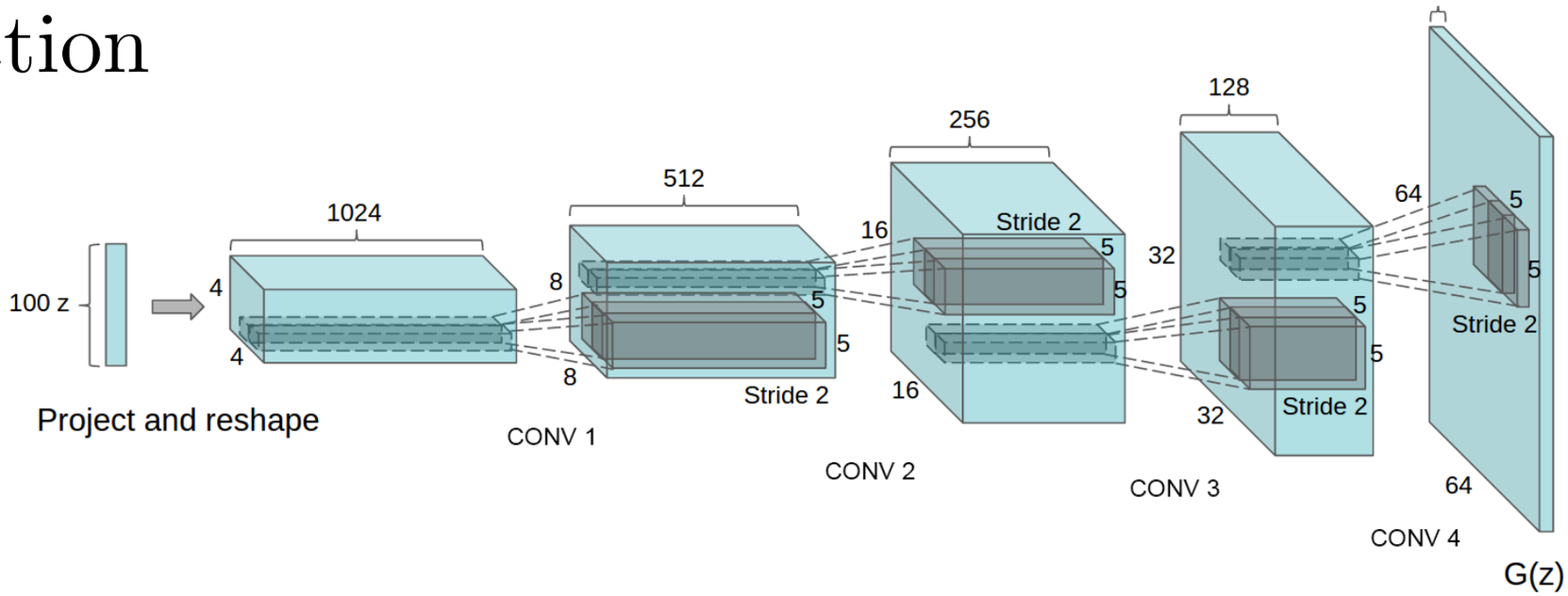
Exceptional results for *images, speech, language, bio-data...*

Why does it work so well ?

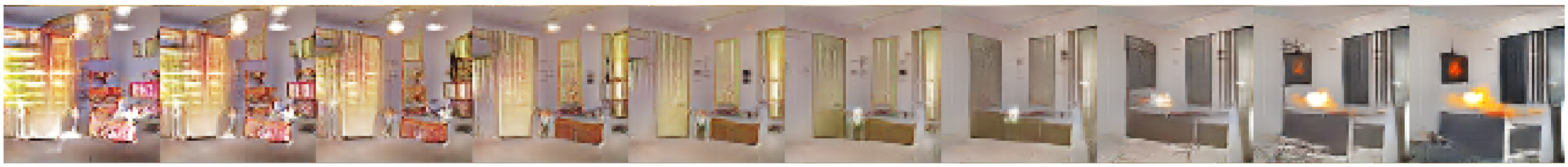
Linearisation in Deep Networks

A. Radford, L. Metz, S. Chintala

- Reconstruction



- On a data basis including bedrooms: interpolaitons



Overview

- Simplified architecture: multiscale wavelet scattering

- Unsupervised learning: statistical physics

- Supervised learning from images to quantum chemistry

- Structuring Deep Networks

Linearise for Dimensionality Reduction

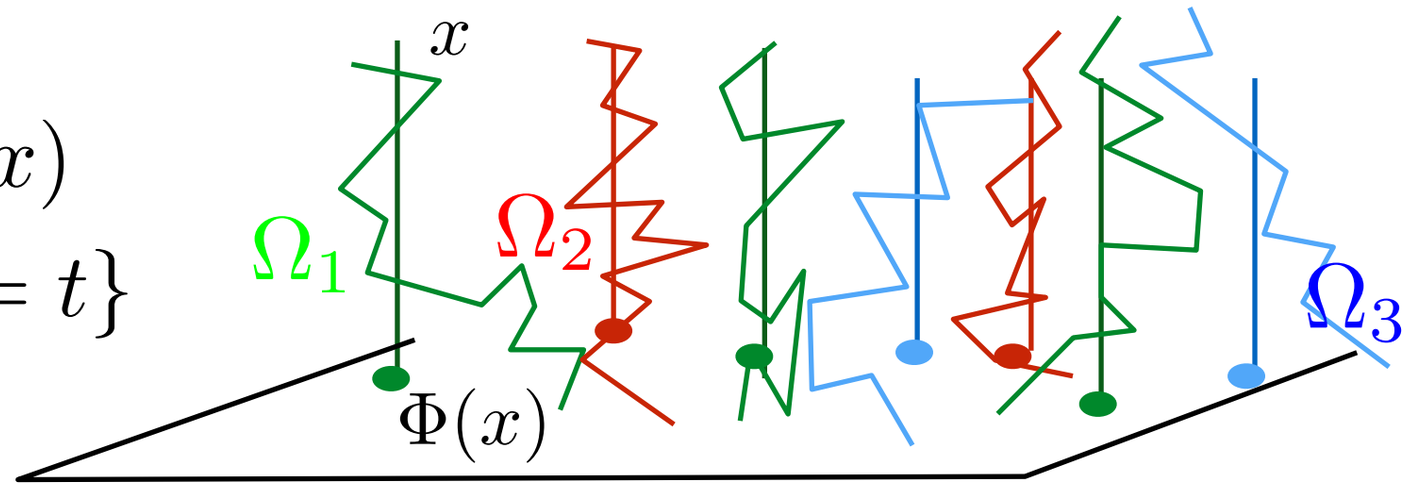
• We want to reduce the dimension of x
with a discriminative lower dimensional representation $\Phi(x)$:

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

Classes

Level sets of $f(x)$

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



• Dimension reduction in two steps:

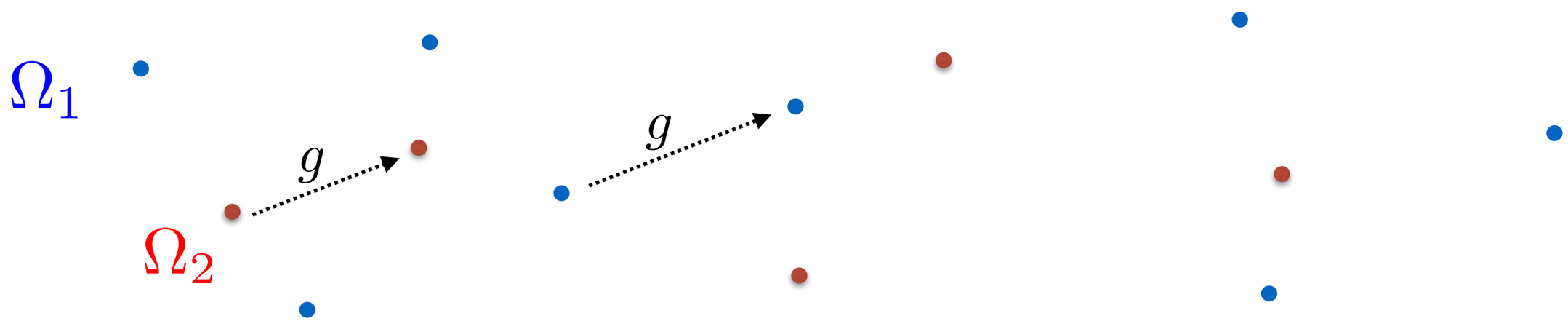
- Linearize level sets Ω_t How ?

- Reduce dimension with linear projections

Symmetries

- Dimensionality curse: geometry of few far away points

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



- Symmetry group of f preserve $\Omega_t = \{x \ : \ f(x) = t\}$

$$G = \{g \ : \ f(g.x) = f(x)\}$$

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

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

- Φ discriminative means that:

- What are the symmetry groups of f ?

$$\Phi(g.x) = \Phi(x) \quad \Rightarrow \quad f(g.x) = f(x)$$

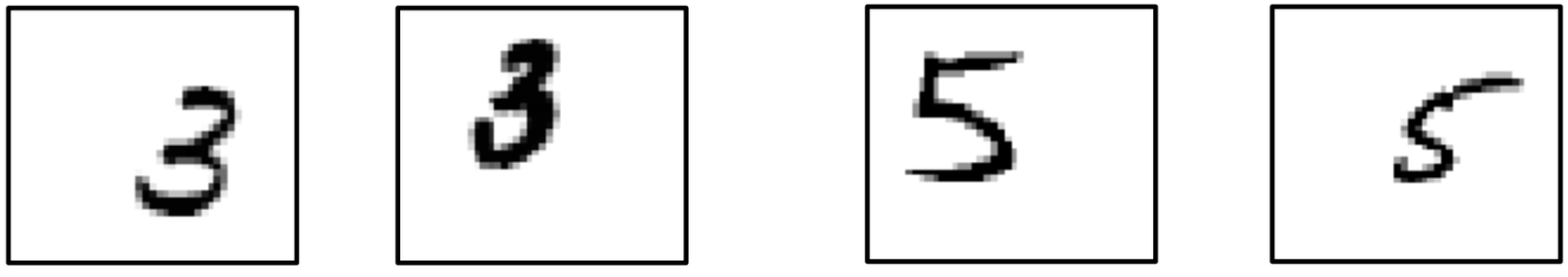
- How to adapt Φ ?

symmetry group of Φ included in symmetry group of f

Translation and Deformations

- Digit classification:

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

Ω_3  Ω_5

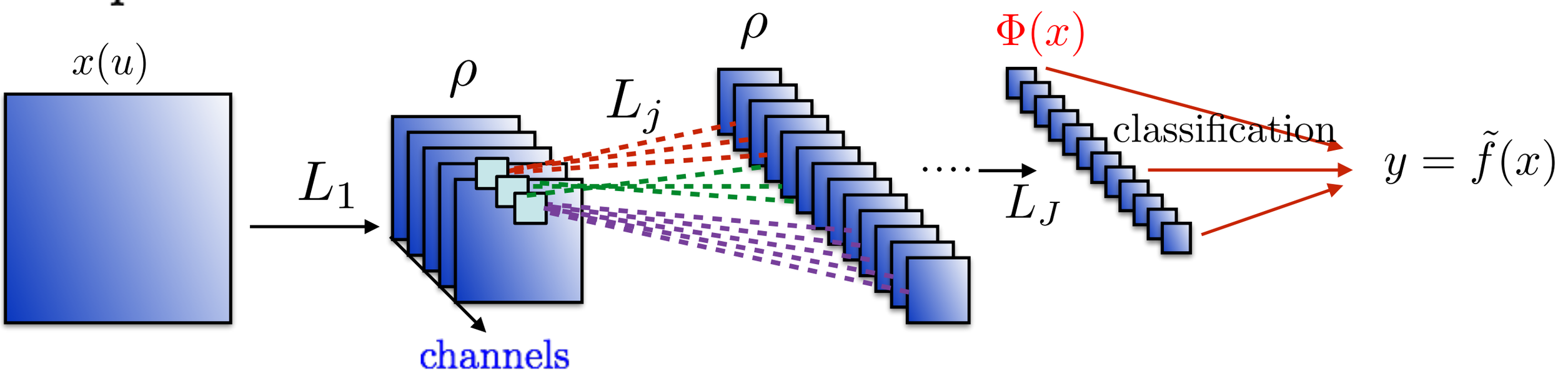
- Globally invariant to the translation group: small
- Locally invariant to small diffeomorphisms: huge group

Linearize small diffeomorphisms:

Video of Philipp Scott Johnson

Deep Convolutional Trees

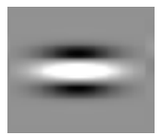
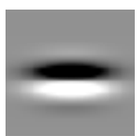
Simplified architecture:



Cascade of convolutions: no channel connections

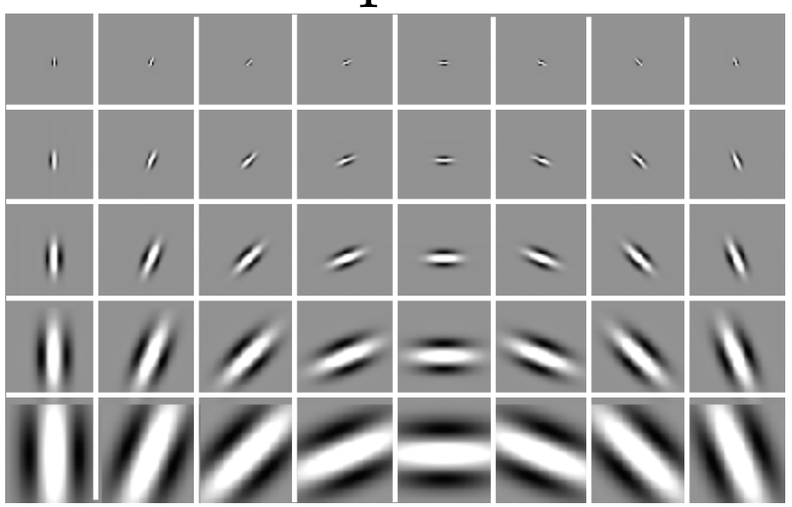
predefined wavelet filters

Scale separation with Wavelets

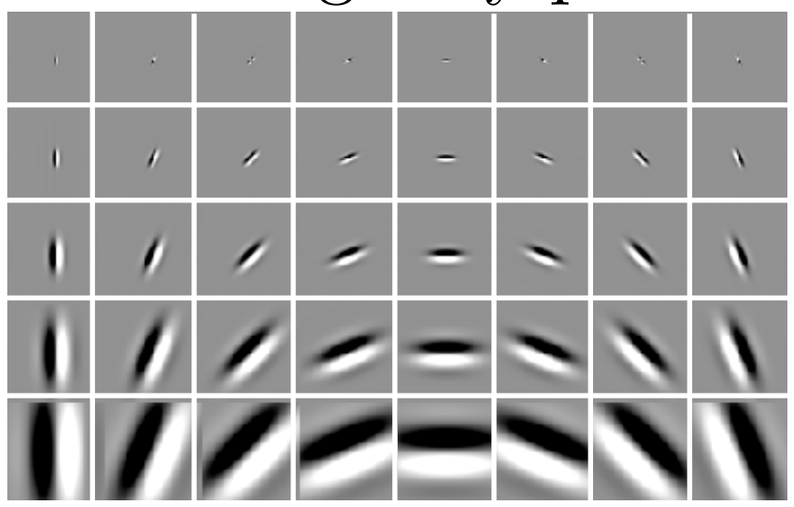
- Wavelet filter $\psi(u) =$  $+ i$ 

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

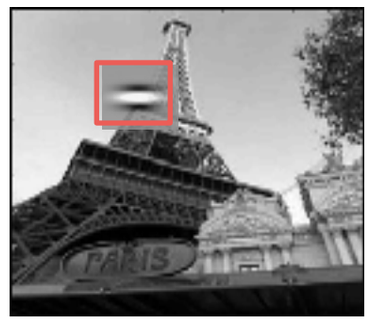
real parts

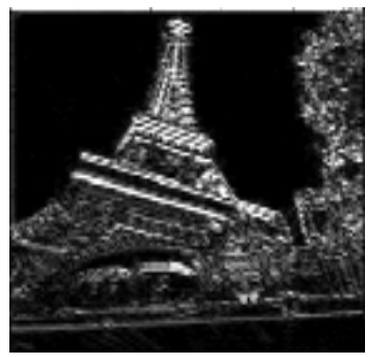


imaginary parts



$x(u)$





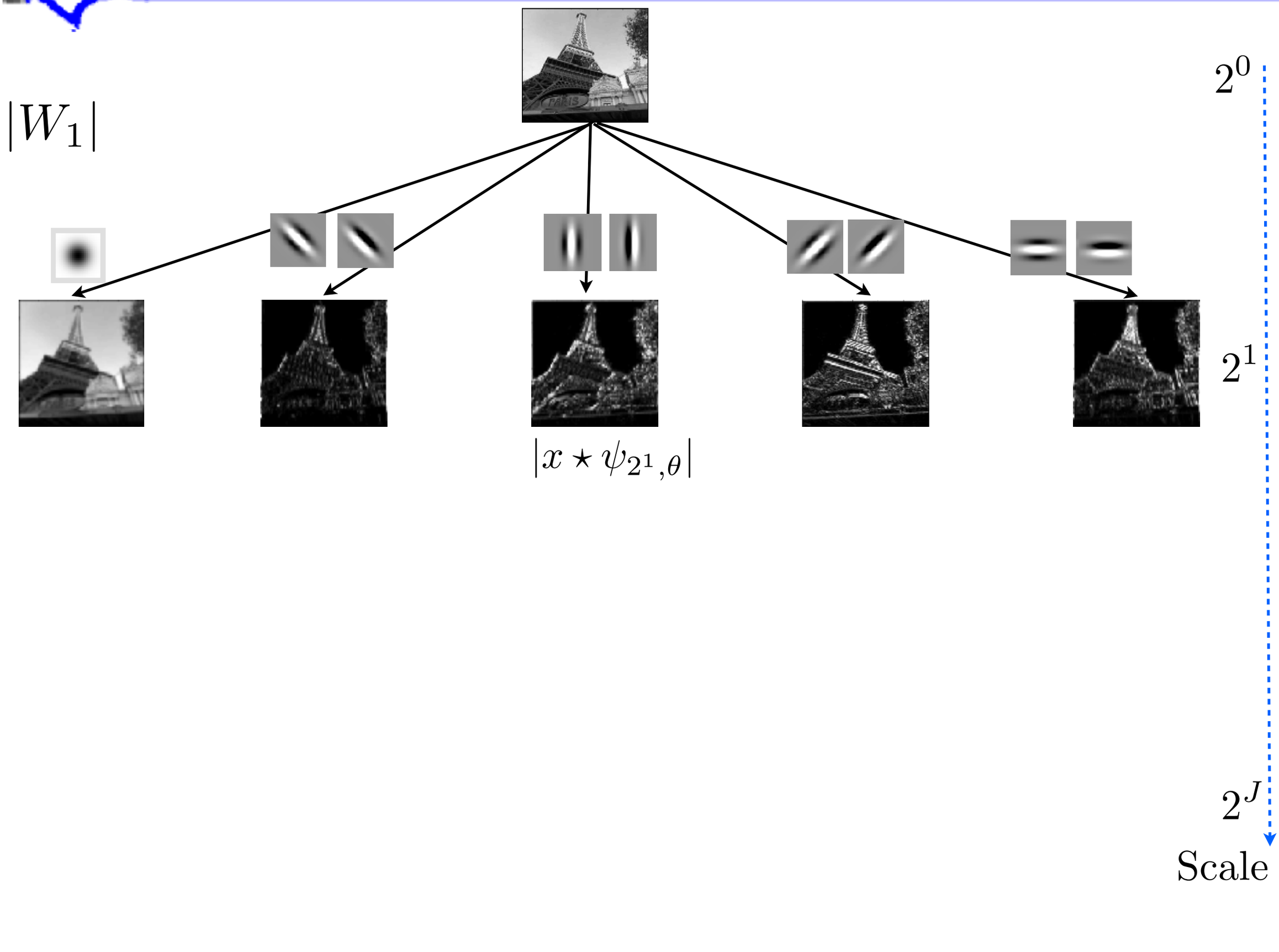
$$x \star \psi_{2^j,\theta}(u) = \int x(v)\,\psi_{2^j,\theta}(u-v)\,dv$$

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

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

Stable to deformations

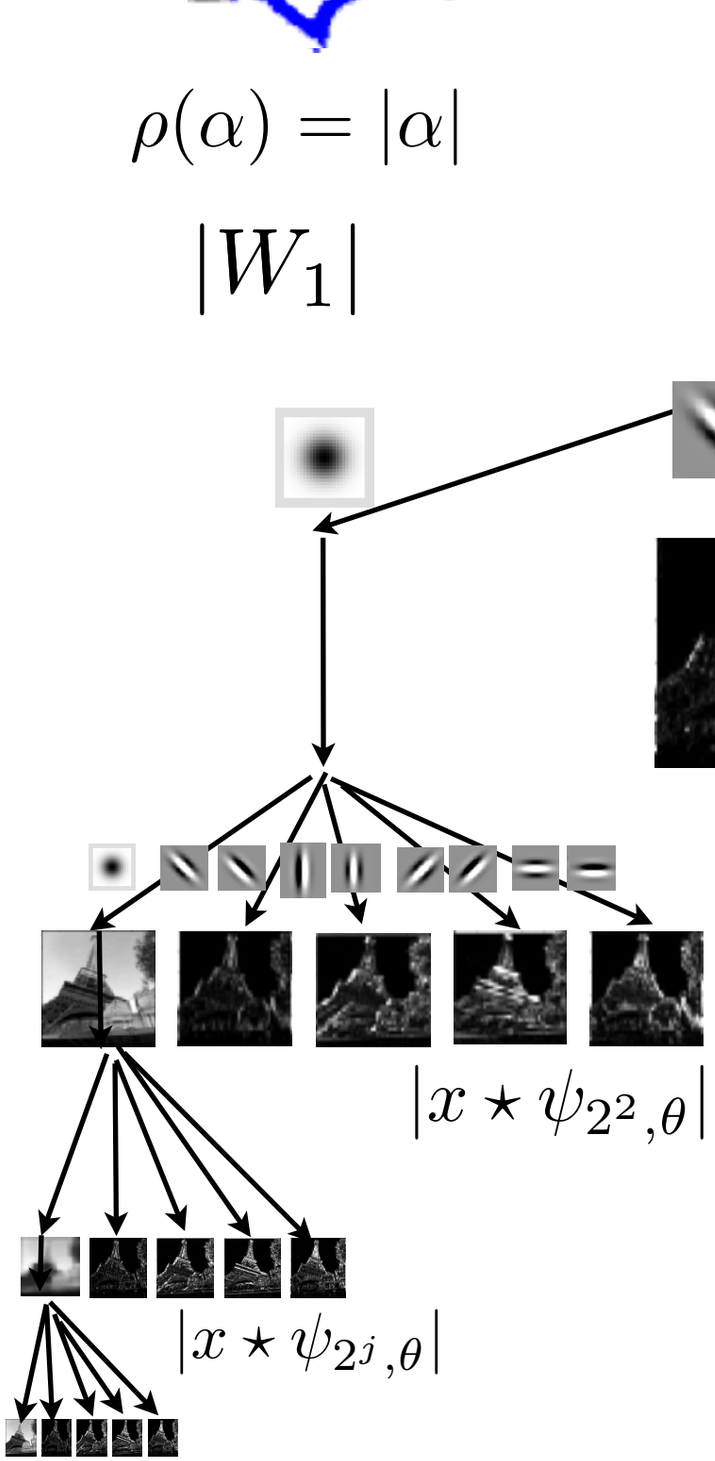
Fast Wavelet Filter Bank



Wavelet Filter Bank

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

$|W_1|$



$x(u)$

$|x \star \psi_{2^1,\theta}|$

$|x \star \psi_{2^2,\theta}|$

$|x \star \psi_{2^j,\theta}|$

2^0

2^1

2^2

2^J

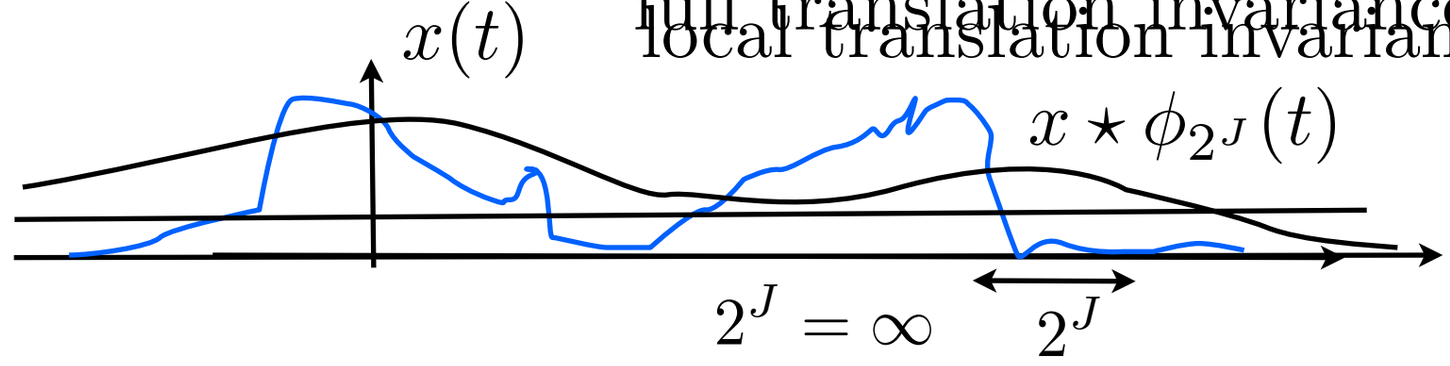
Scale

Wavelet Translation Invariance

First wavelet transform

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

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



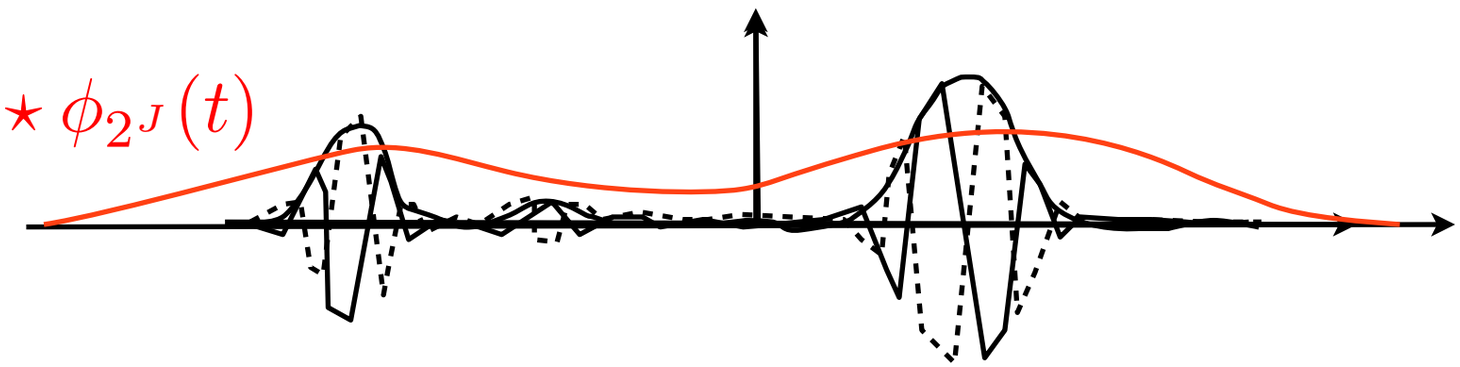
full translation invariance

local translation invariance

Lost high frequencies: $x \star \psi_{\lambda_1}(t)$

Eliminate the phase: $|x \star \psi_{\lambda_1}(t)|$

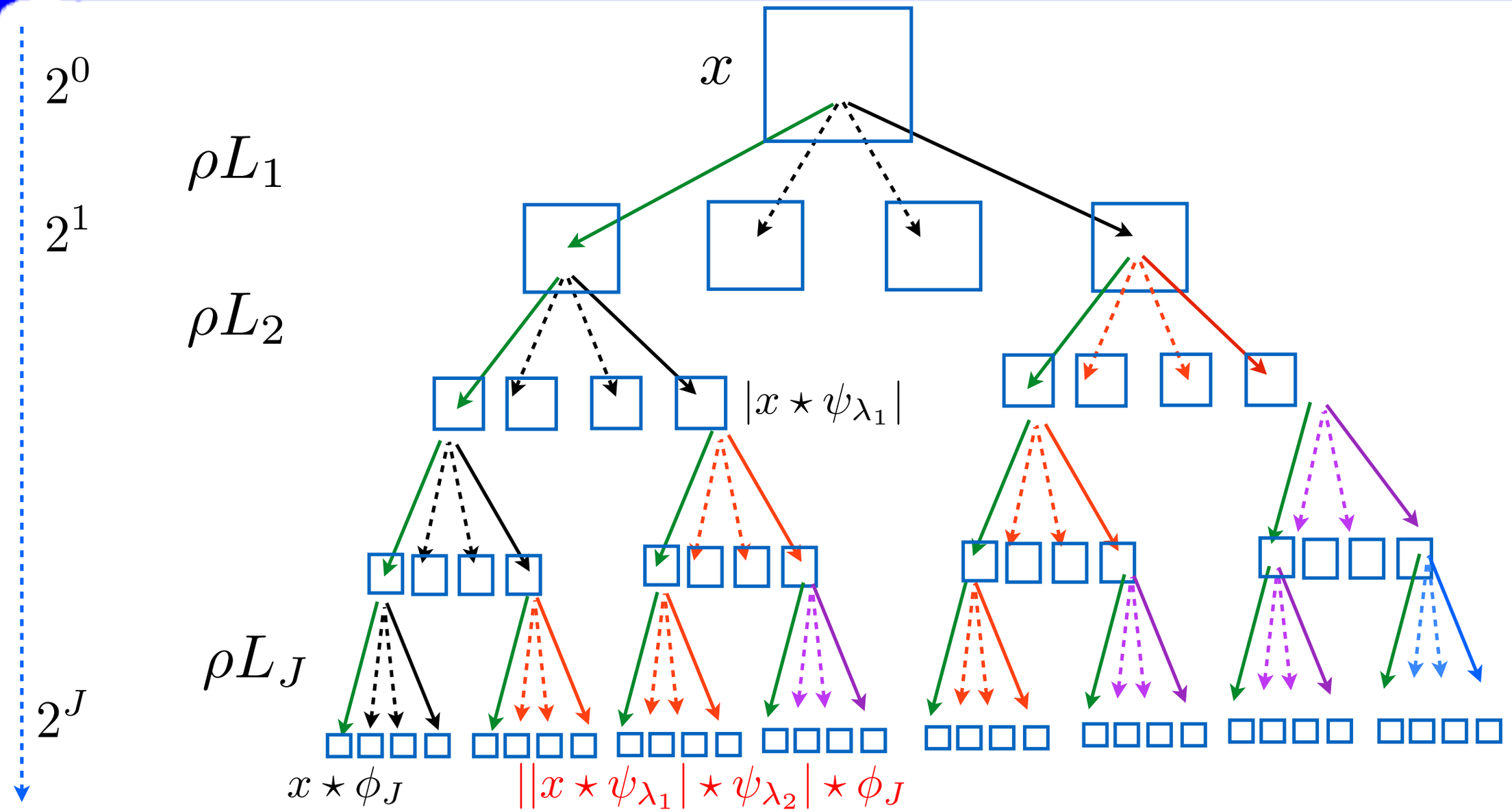
Invariant: $|x \star \psi_{\lambda_1}| \star \phi_{2^J}(t)$



Need to recover lost high frequencies: $|x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}(t)$

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

Wavelet Scattering Network



$$S_J = \rho W_1 \quad \rho W_2 \quad \cdots \quad \rho W_J$$

$$\rho(\alpha) = |\alpha| \qquad S_J x = \left\{ |||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2} \star \ldots| \star \psi_{\lambda_m}| \star \phi_J \right\}_{\lambda_k}$$

Interactions across scales

Scattering Properties

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

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

Theorem: *For appropriate wavelets, a scattering is*

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

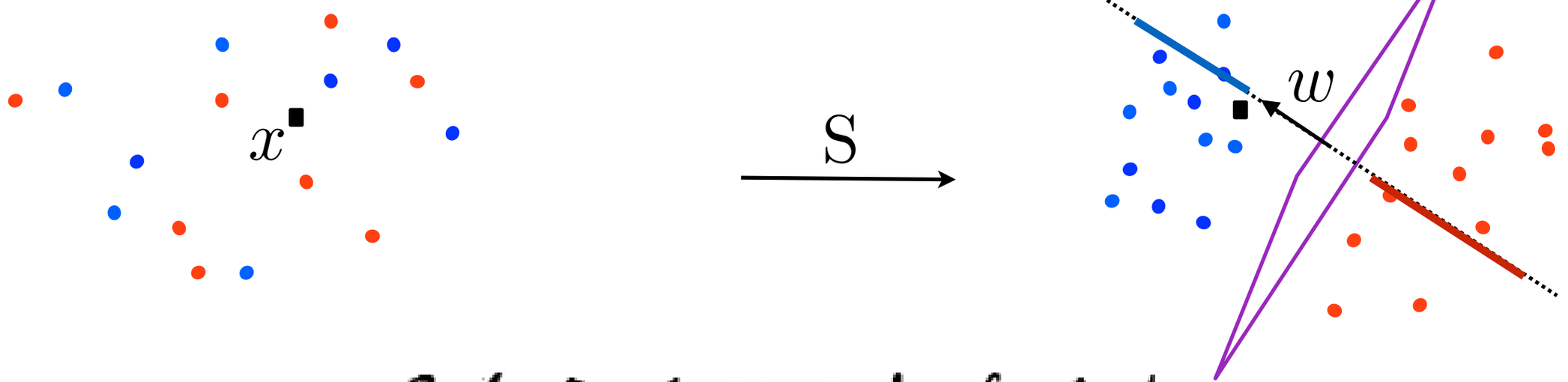
preserves norms $\|S_J x\| = \|x\|$

translations invariance and deformation stability:

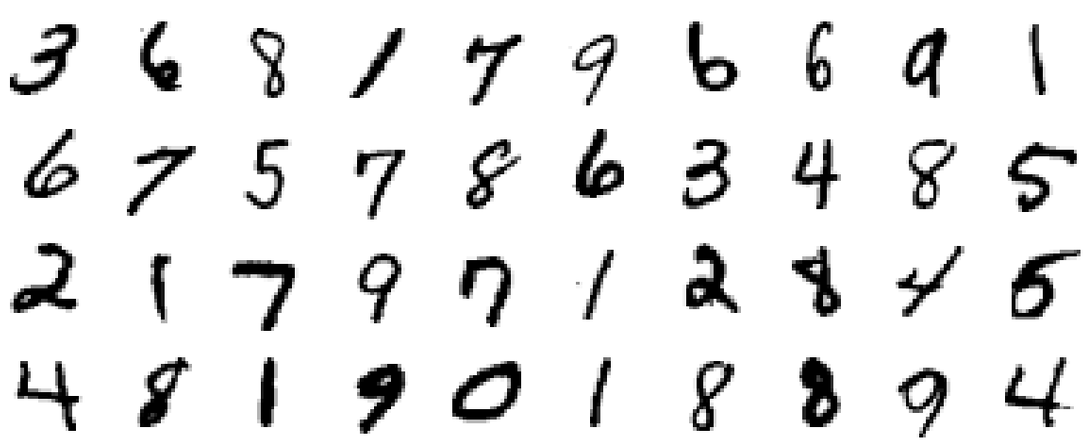
if $D_\tau x(u) = x(u - \tau(u))$ *then*

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

Image Classification

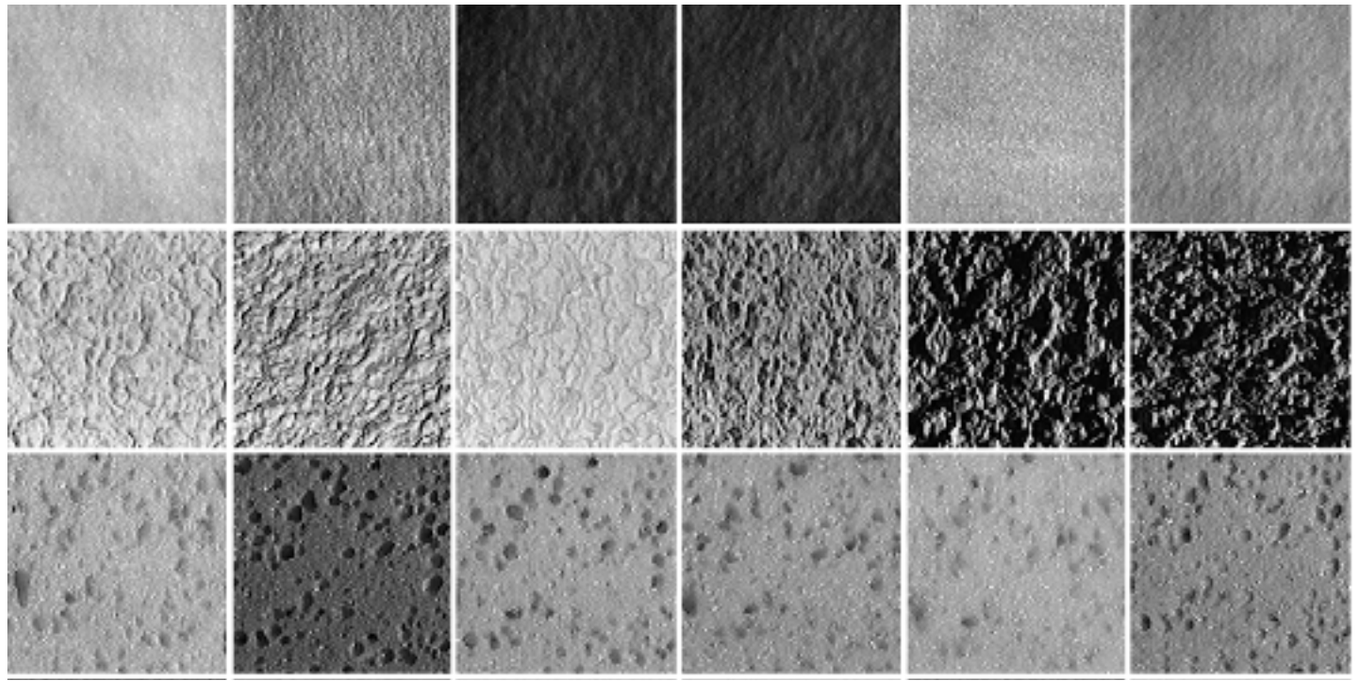


Digit Recognition



0.4% *errors*

CUREt
61 classes



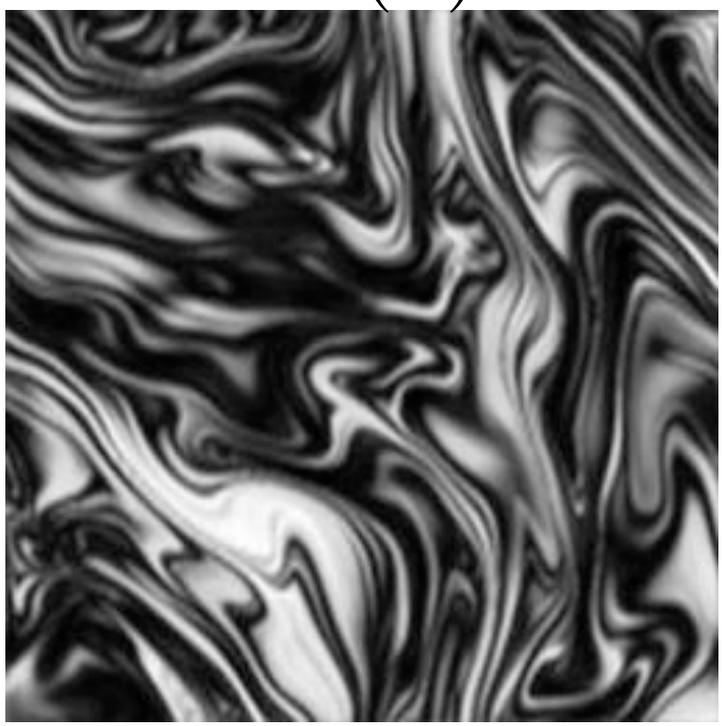
0.2% *errors*

Unsupervised Learning

Joan Bruna

- Estimate the probability density $p(x)$ of $X(u)$ from few realisations $\{x_i(u)\}_i$

$X(u)$



d

Scattering of a stationary process $X(u)$

$$S_J X = \begin{pmatrix} X \star \phi_{2^J}(u) \\ |X \star \psi_{\lambda_1}| \star \phi_{2^J}(u) \\ ||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_{2^J}(u) \\ \dots \end{pmatrix}$$

if $2^J = d$

Scattering moments

$$= \begin{pmatrix} d^{-1} \sum_{u=1}^{d} X(u) \\ d^{-1} \|X \star \psi_{\lambda_1}\|_1 \\ d^{-1} \||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}\|_1 \\ \dots \end{pmatrix} \xrightarrow[d \to \infty]{\text{if ergodicity}} \bar{\mu} = \begin{pmatrix} \mathbb{E}(X) \\ \mathbb{E}(|X \star \psi_{\lambda_1}|) \\ \mathbb{E}(||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|) \\ \dots \end{pmatrix}$$

- How to estimate the probability density $p(x)$ of X ?

Canonical Maximum Entropy

Given a vector of scattering moments:

$$\mathbb{E}(SX) = \begin{pmatrix} \mathbb{E}(X) \\ \mathbb{E}(|X \star \psi_{\lambda_1}|) \\ \mathbb{E}(||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|) \\ \dots \end{pmatrix}_{\lambda_1, \lambda_2, \dots} = \Big(\mathbb{E}(\phi_m(X)\Big)_m$$

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

$$\mathbb{E}(\phi_m(X)) = \int_{\mathbb{R}^N} \phi_m(x)\ p(x)\, dx = \bar{\mu}_m$$

with a maximum entropy $H_{\max} = -\int p(x)\ \log p(x)\, dx$ is

$$p(x) = \frac{1}{Z}\ \exp\Big(\sum_m \beta_m\, \phi_m(x)\Big)$$

Multiscale Hamiltonian with scale interactions

Canonical Maximum Entropy

Given a vector of scattering moments:

$$\mathbb{E}(SX) = \begin{pmatrix} \mathbb{E}(X) \\ \mathbb{E}(|X \star \psi_{\lambda_1}|) \\ \mathbb{E}(||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|) \\ \dots \end{pmatrix}_{\lambda_1, \lambda_2, \dots} = \Big(\mathbb{E}(\phi_m(X)\Big)_m$$

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

$$\mathbb{E}(\phi_m(X)) = \int_{\mathbb{R}^N} \phi_m(x)\ p(x)\, dx = \bar{\mu}_m$$

with a maximum entropy $H_{\max} = -\int p(x)\ \log p(x)\, dx$ is

$$p(x) = \frac{1}{Z}\ \exp\Big(\sum_m \beta_m\, \phi_m(x)\Big)$$

Multiscale Hamiltonian with scale interactions

Numerically too expansive to compute Lagrange multipliers β_m

Microcanonical Sampling

Joan Bruna

- Given a single realisation of X :

$$SX = \left\{ d^{-1} \sum_u X(u), d^{-1} \|X \star \psi_{\lambda_1}\|_1, d^{-1} \| |X \star \psi_{\lambda_1}| \star \psi_{\lambda_2} \|_1 \right\} \approx \mathbb{E}(SX).$$

- A microcanonical max entropy process $\tilde{X}$ satisfies

$$\|S\tilde{X} - SX\| \leq \epsilon$$

Theorem (*H. Georgii*)

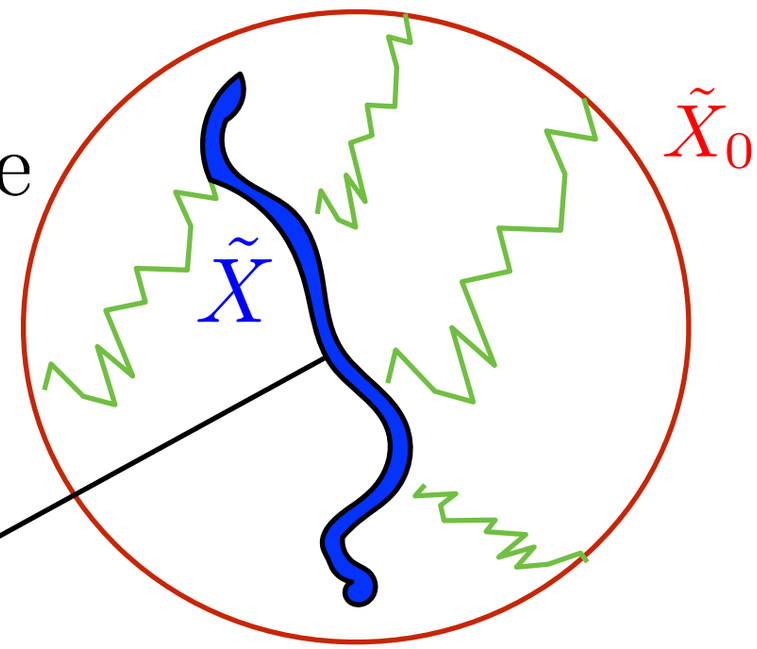
For scattering, the micro and macrocanonical processess converge to the same Gibbs measure when d goes to ∞

Algorithm:

Initialized with $\tilde{X}_0$ Gaussian white noise

Iteratively reduce $\|S\tilde{X}_n - SX\|^2$ with gradient descent

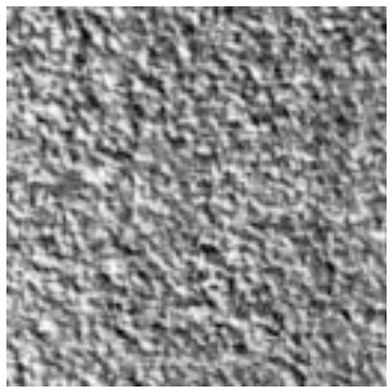
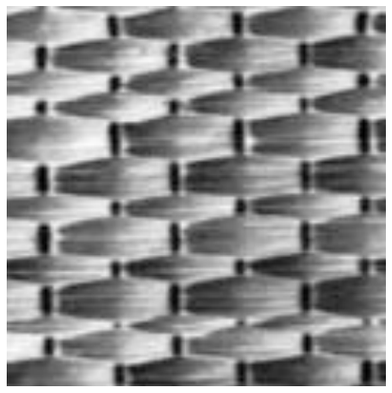
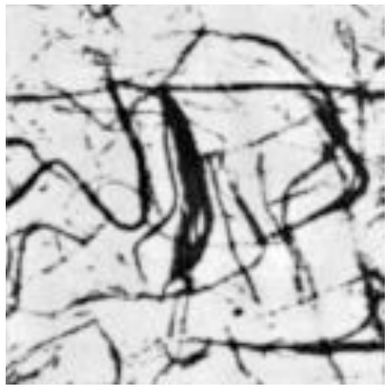
$$\{x \ : \ \|Sx - SX\| \leq \epsilon\}$$



Texture Reconstructions

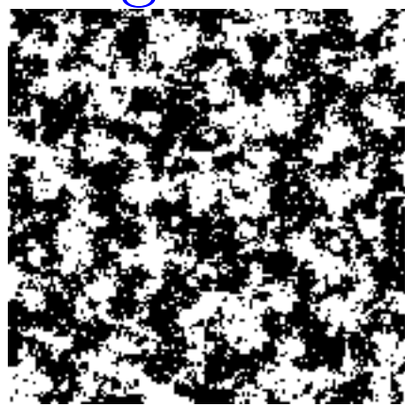
Joan Bruna

Texture of d pixels

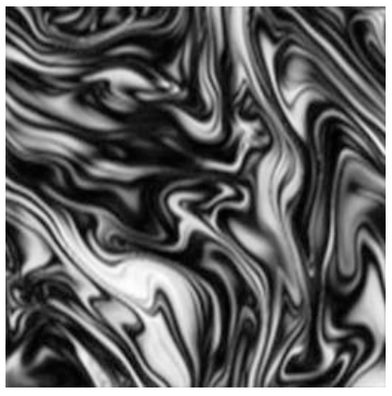




Statistical Physics

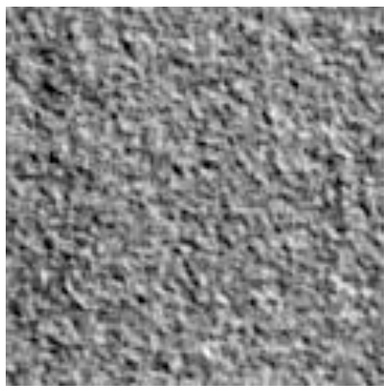
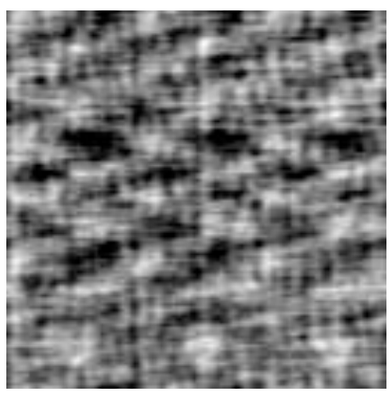
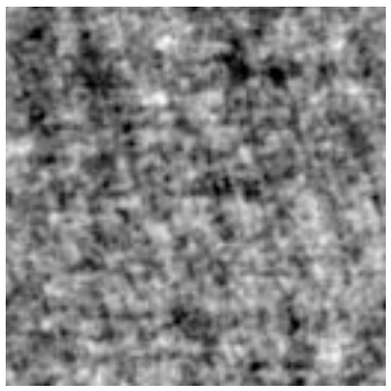
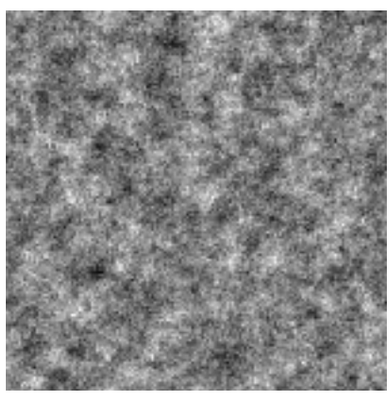
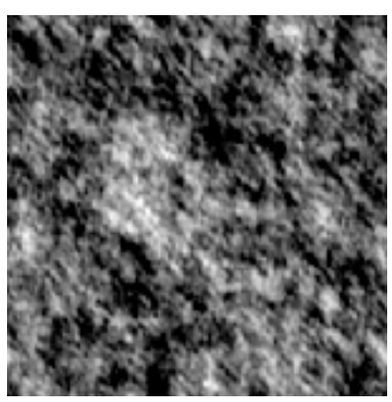
Ising-critical



Turbulence 2D

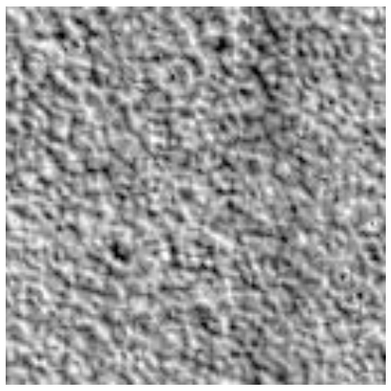
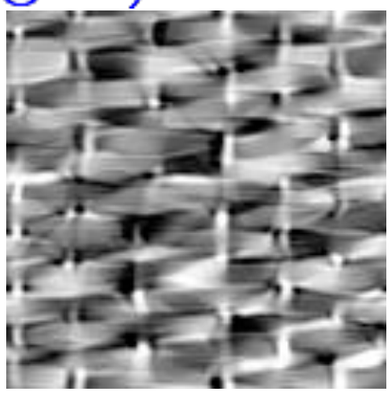
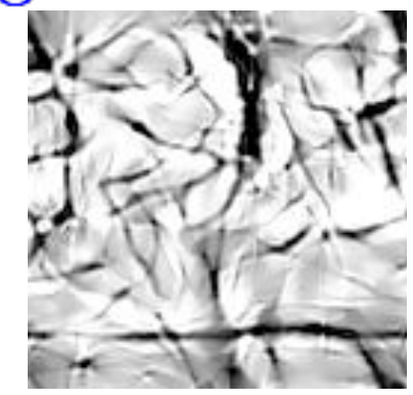
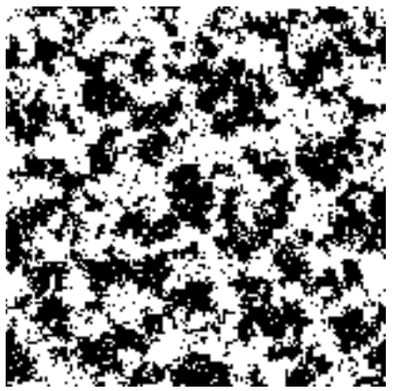


Gaussian process model with d second order moments

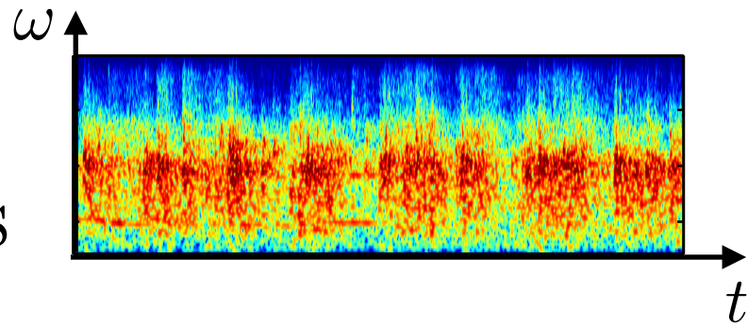
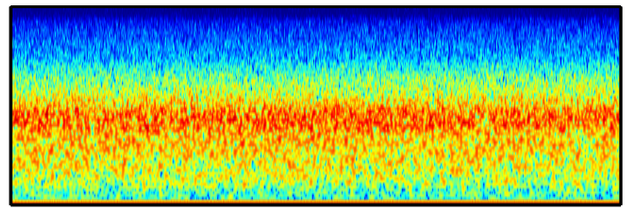
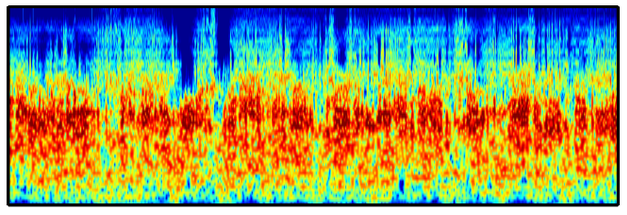
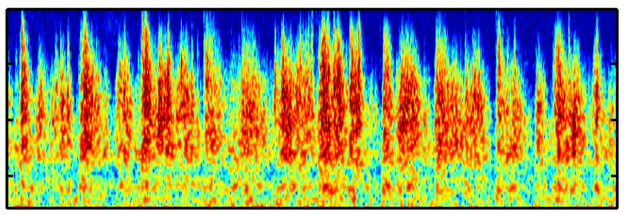
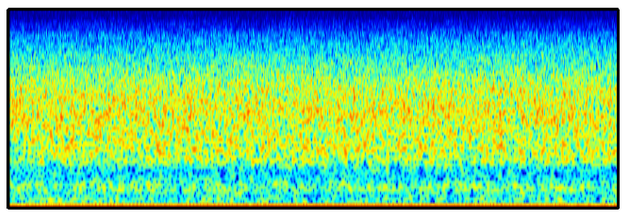
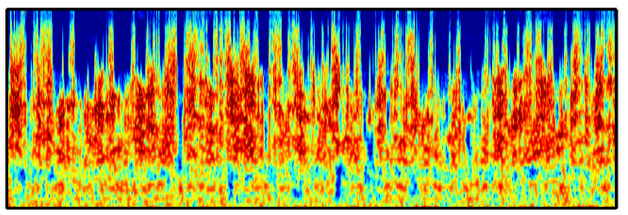
Reconstructions from $\|X \star \psi_{\lambda_1}\|_1$ and $\||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}\|_1$

$O(\log^2 d)$ scattering coefficients


Representation of Audio Textures

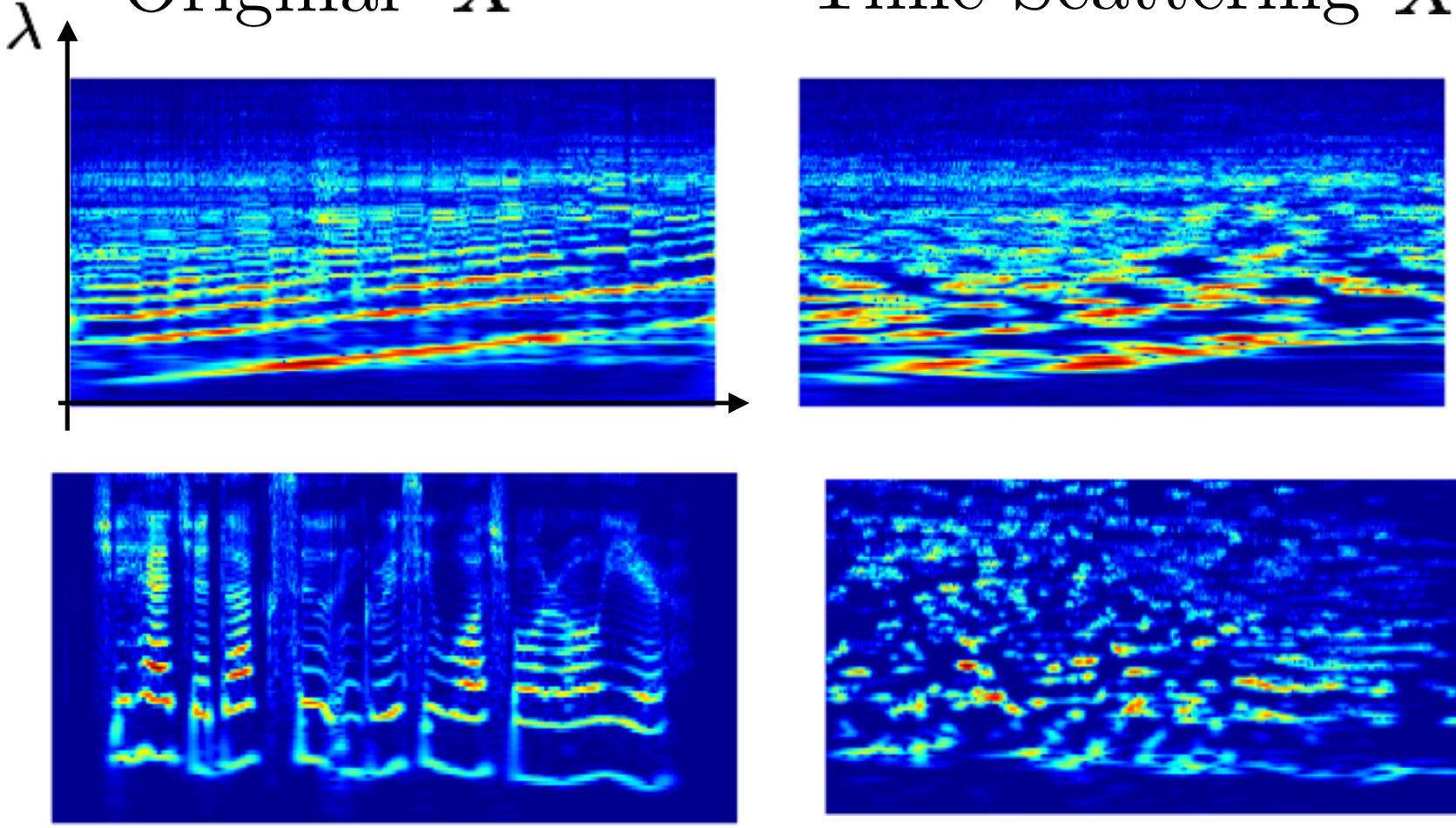
Joan Bruna

	Original	Gaussian in time	Scattering Order 2
Applauds			
Paper			
Cocktail Party			

Failures of Audio Synthesis

J. Anden and V. Lostanlen

Original X Time Scattering $\tilde{X}$

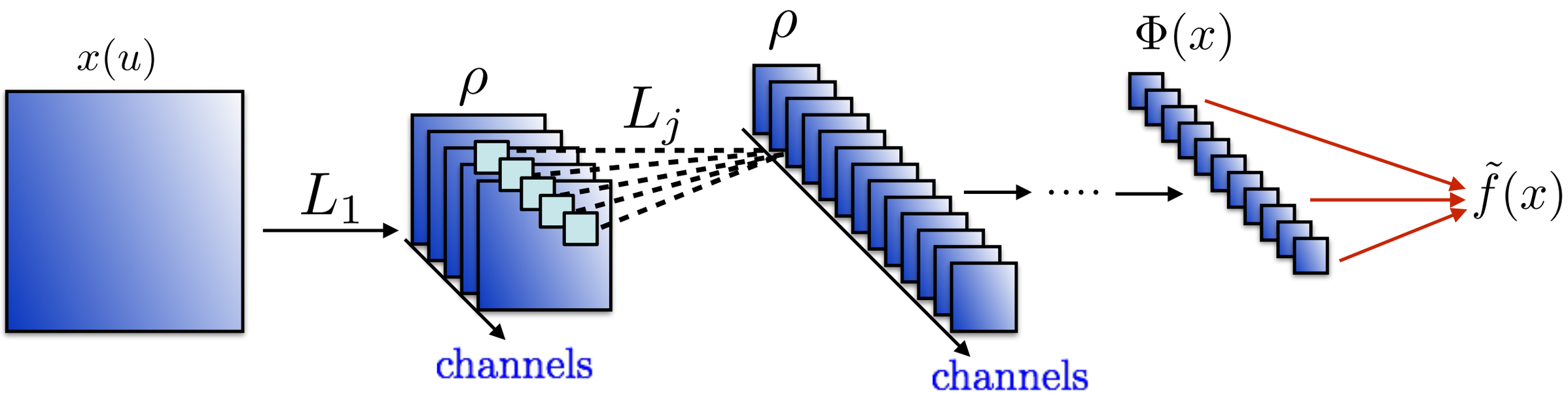


Typical of $\tilde{X}$ is not typical of X

- Missing frequency connections $\Rightarrow$ misalignments

$\Rightarrow$ incorporate two-dimensional translations in time-frequency

Channel Connections



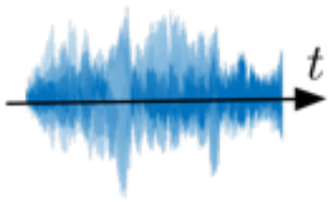
What is the role of channel connections ?

Time-Frequency Translation Group

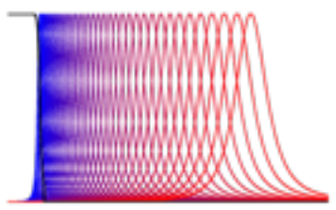
J. Anden and V. Lostanlen

signal

$x(t)$

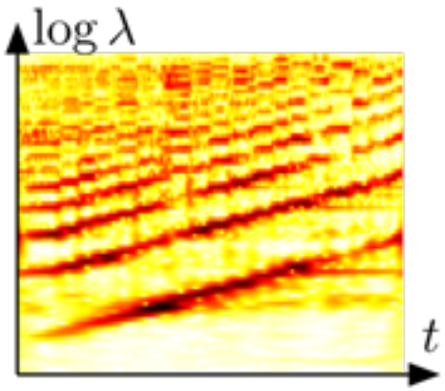


auditory wavelet modulus

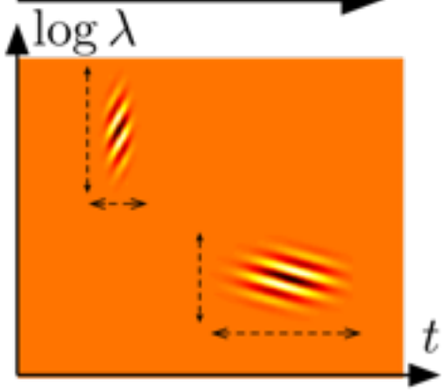


scalogram

$|x \overset{t}{*} \psi_\lambda|$

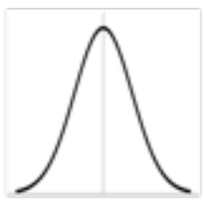


Time-frequency wavelet convolutions



joint time-frequency scattering

$\left| |x \overset{t}{*} \psi_\lambda| \overset{t}{*} \psi_\alpha \overset{\log \lambda}{*} \psi_\beta \right|$



averaging

$T = 500\text{ms}$

$|x \overset{t}{\star} \psi_\lambda| \overset{t}{\star} \phi_J$

averaging

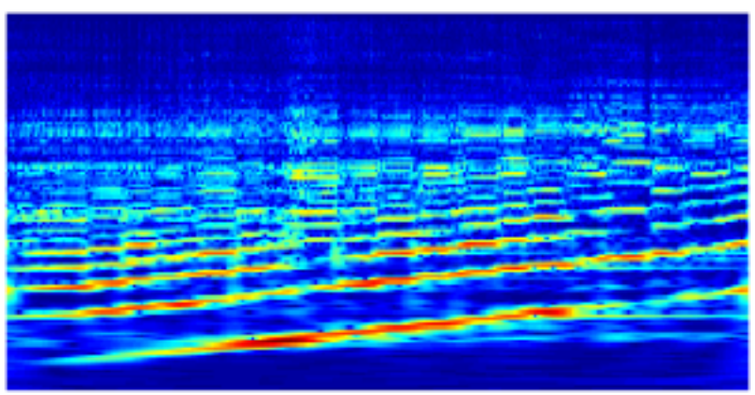
$T = 500\text{ms}$

$||x \overset{t}{\star} \psi_\lambda| \overset{t}{\star} \psi_\alpha \overset{\log \lambda}{\star} \psi_\beta| \overset{t}{\star} \phi_J$

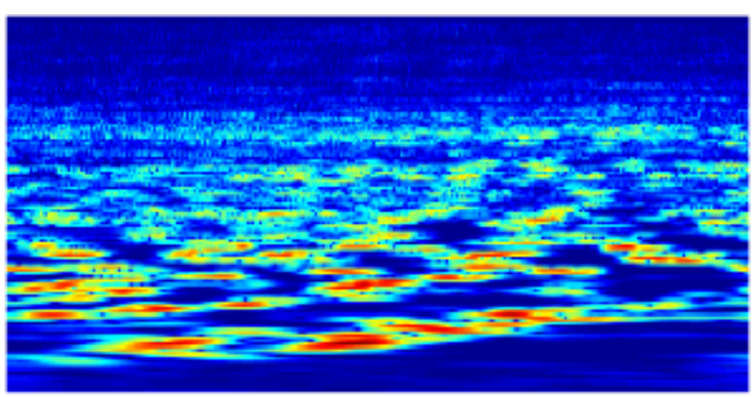
Joint Time-Frequency Scattering

J. Anden and V. Lostanl

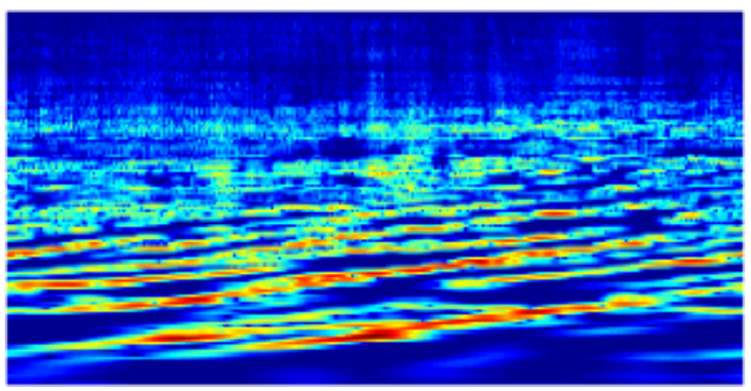
Original

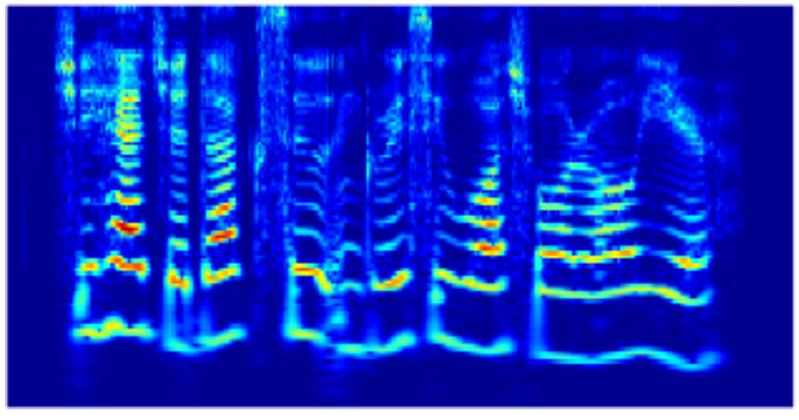


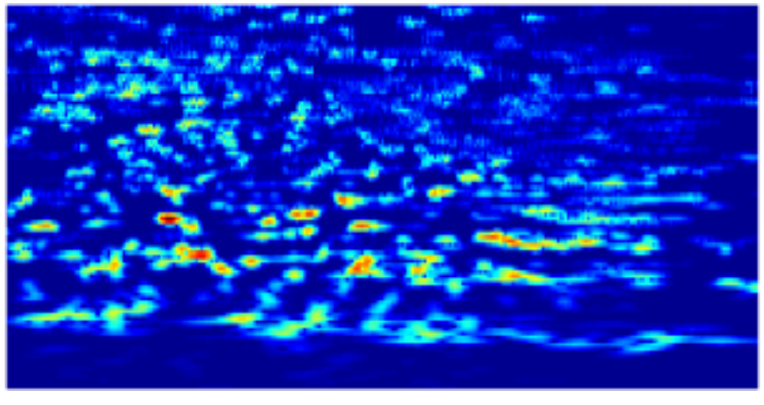
Time Scattering

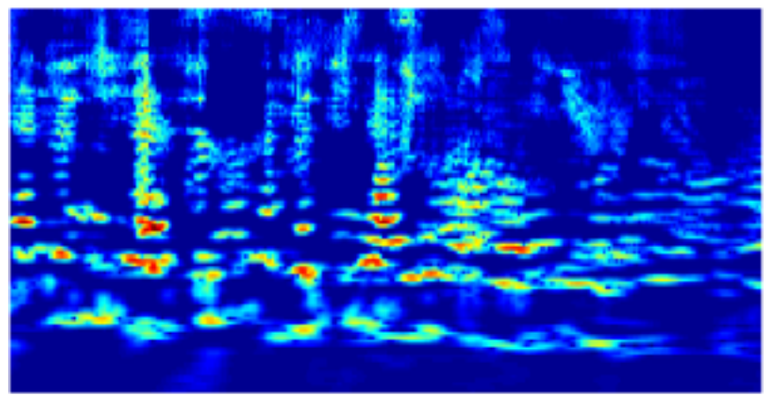


Time/Freq Scattering



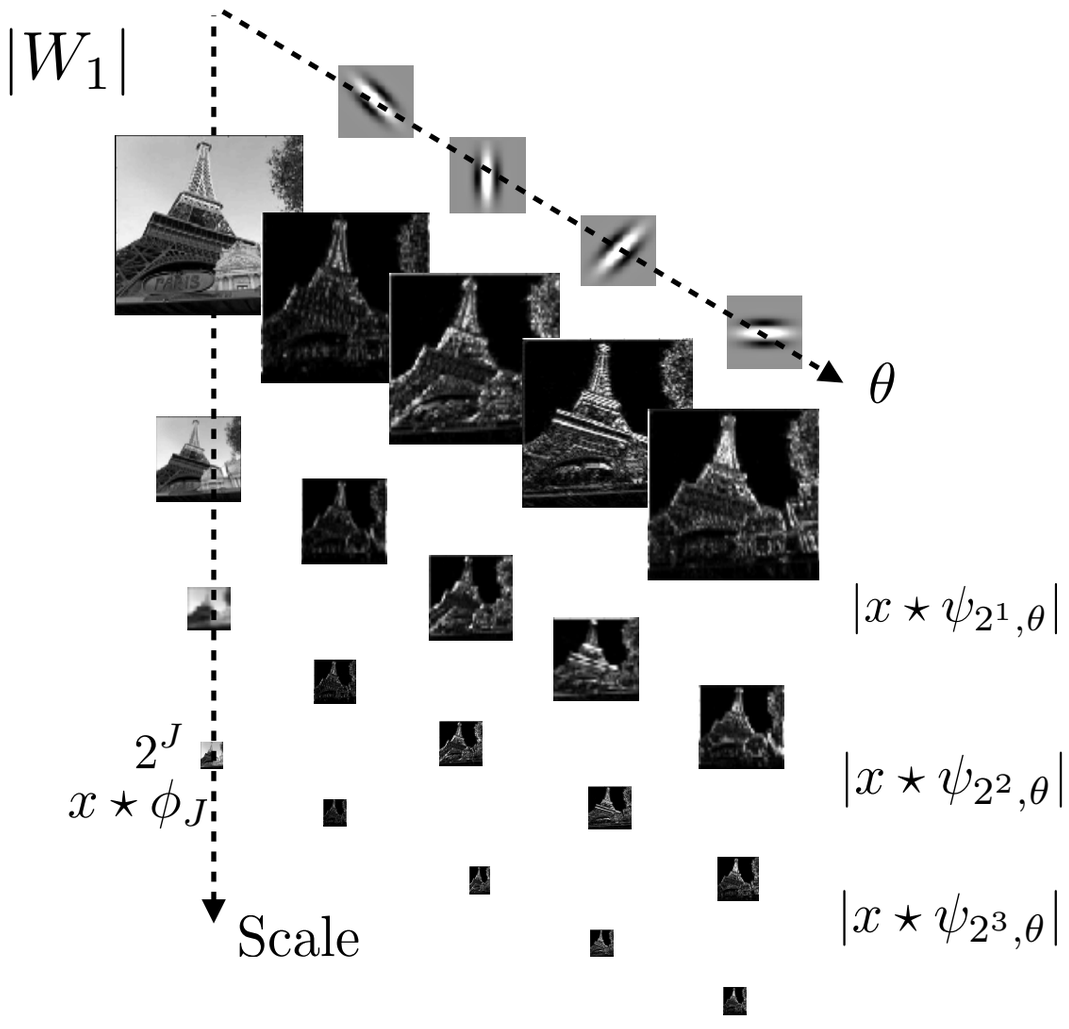






Symmetries: Rotation Invariance

- 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.

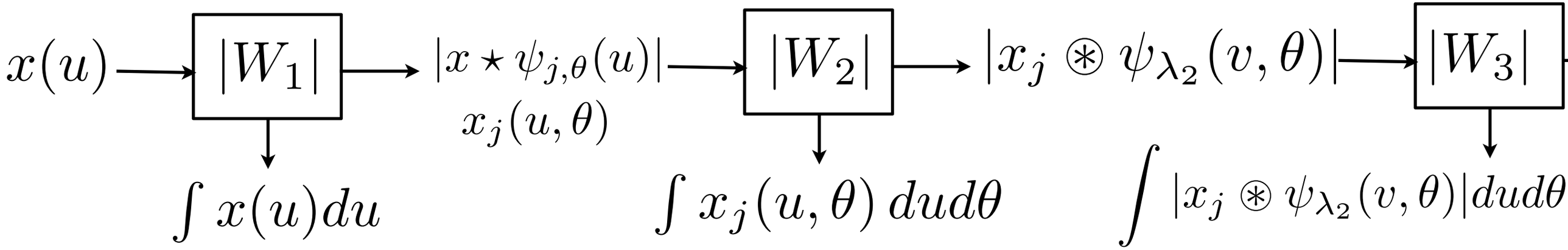
Extension to Rigid Mouvements

Laurent Sifre

- Group of rigid displacements: translations and rotations

- Scattering on rigid mouvements:

Wavelets on Translations Wavelets on Rigid Mvt. Wavelets on Rigid Mvt

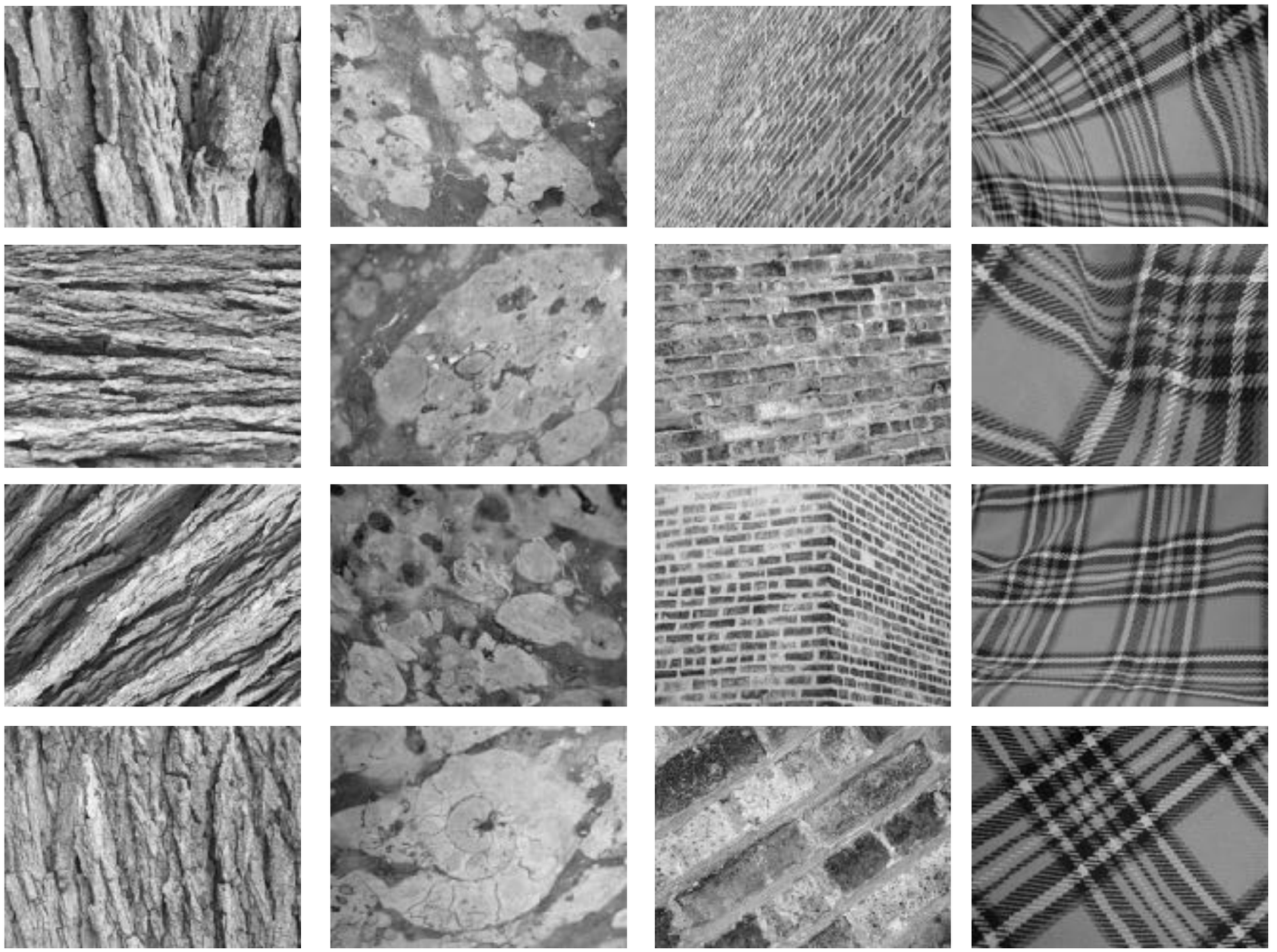


$$x \circledast \psi_\lambda(u,\theta) = \int_0^{2\pi} \int_{\mathbb{R}^2} x(u',\theta')\,\psi_{\theta,2^j}(u-u')\psi_{2^k}(\theta-\theta')\,d\theta' du'$$

Rotation and Scaling Invariance

Laurent Sifre

UIUC database:
25 classes



Scattering classification errors

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

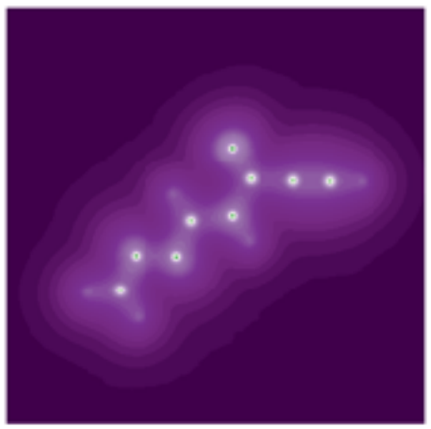
Learning Physics: N-Body Problem

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

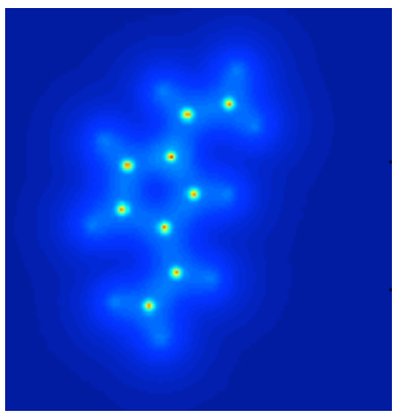
Quantum chemistry: $f(x)$ is invariant to rigid mouvements, stable to deformations.

The energy depends upon the electronic density (Kohn-Sham)

Ground state
electronic density
computed with Schroedinger



$\rho_x(u)$



$\rho_x(u)$

Quantum Regression

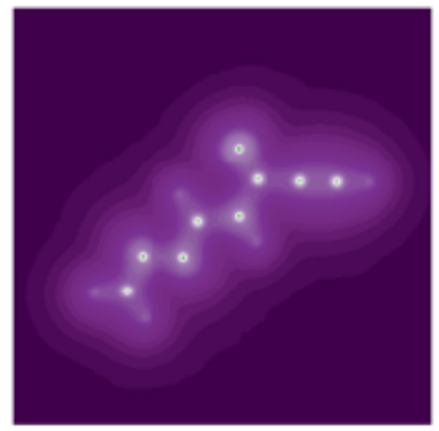
N. Poilvert

Matthew Hirn

- Compute $f(x)$ from isolated atomic densities without interactions:

$\tilde{\rho}_x$: sum of individual densities

ρ_x: ground state elecronic density



- 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.
carrying chemical properties

Scattering Regression

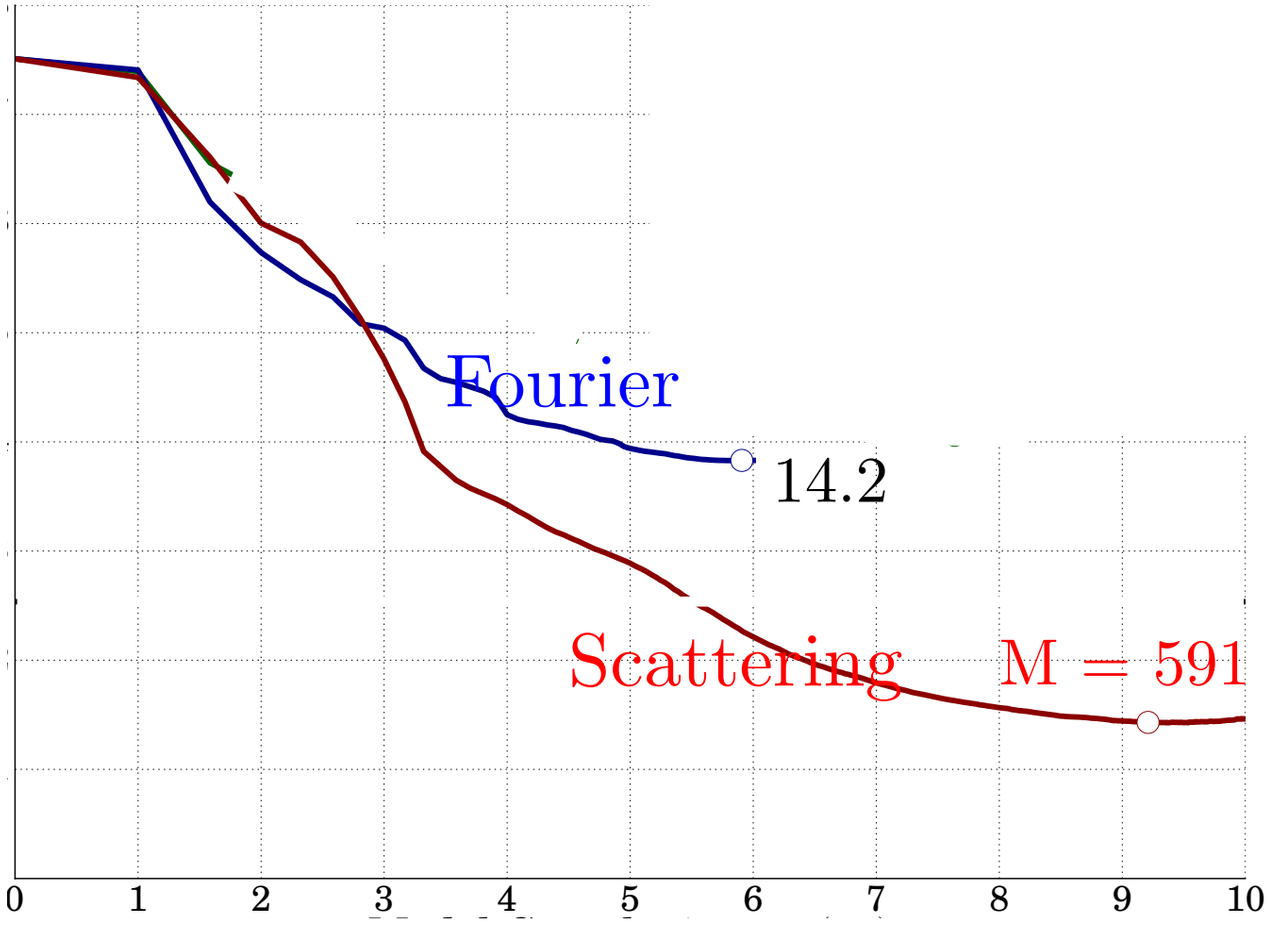
Eickenberg, Exarchakis, Hirn

Data basis $\{x_i\,,\ f(x_i)\}_{i \leq N}$ of 7000 3D molecules

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

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

Interaction terms
across scales



1.2 kcal/mol
1.0: Deep Tensor Networks
(*Schutt et. al.*)

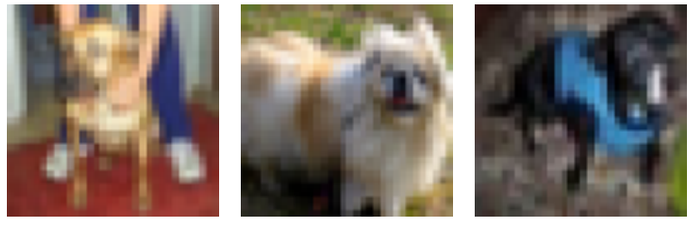
Image Classification: CIFAR-10

Edouard Oyallon

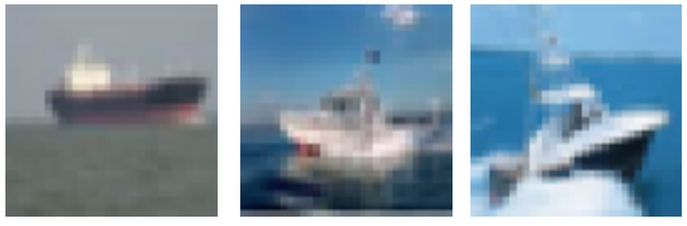
10 classes, $50\,10^3$ labeled training images, of 32×32 pixels

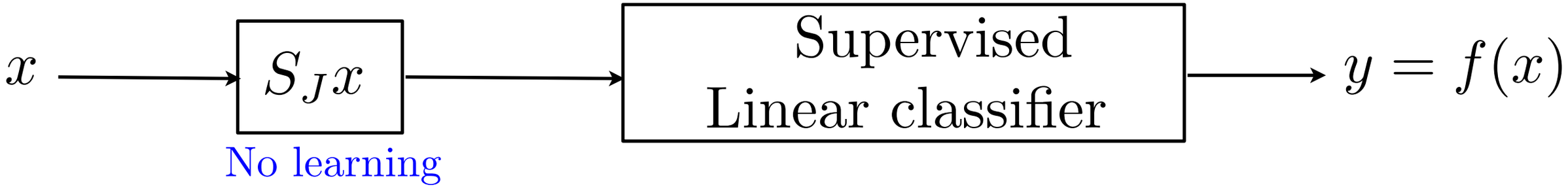
Cars

Dogs



Ships





Data Basis	Deep-Net	Scattering
CIFAR-10	**7%**	20%

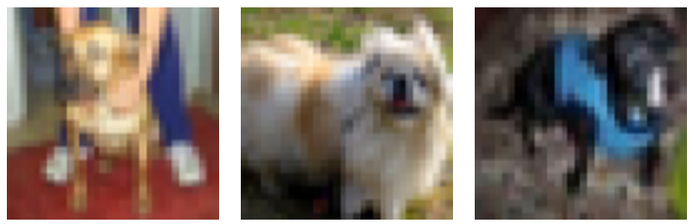
Image Classification: CIFAR-10

Oyallon, Belivovsky, Zagoruyko

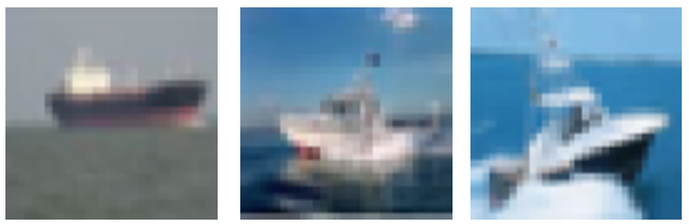
10 classes, $50\,10^3$ labeled training images, of 32×32 pixels

Cars

Dogs



Ships

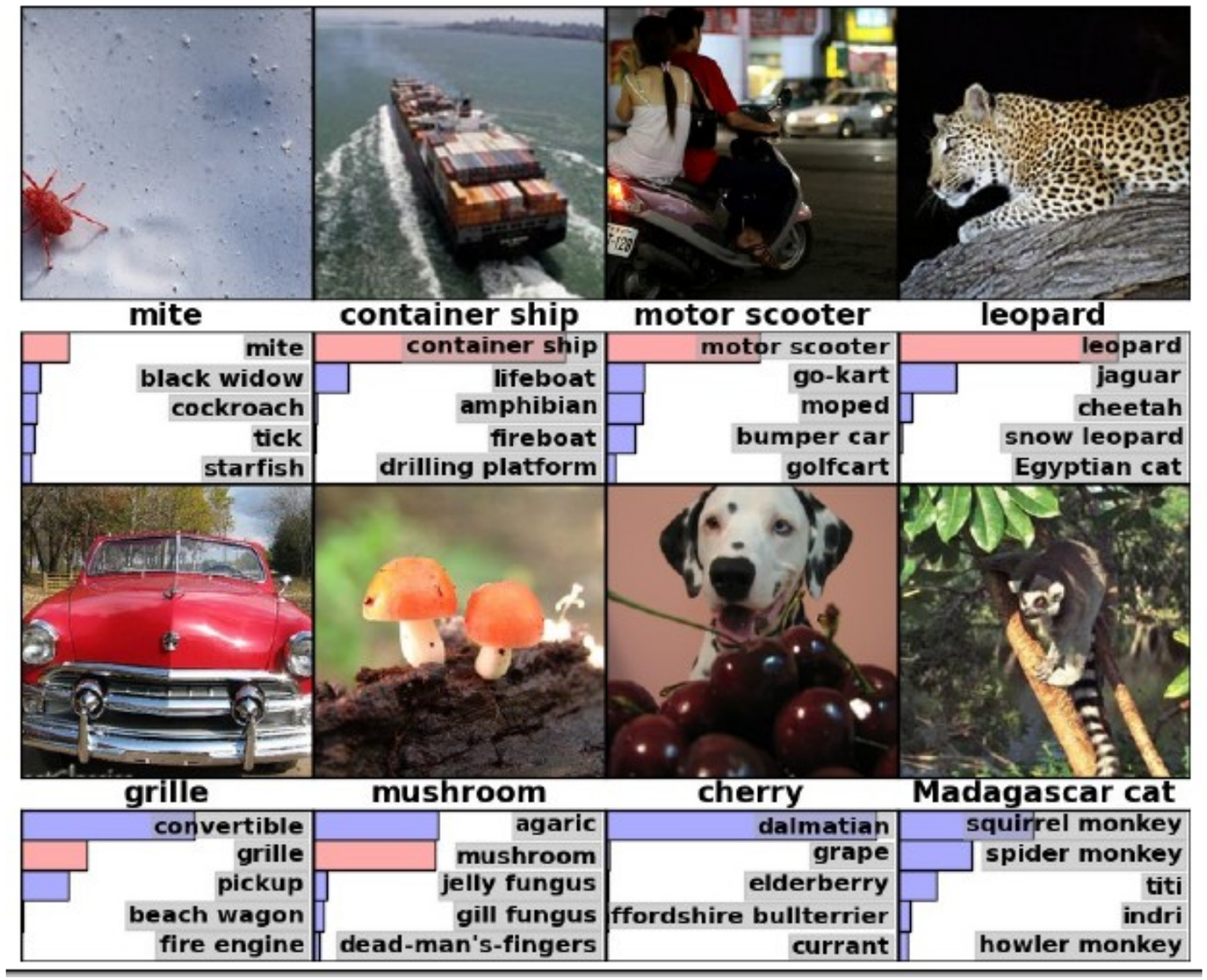


Data Basis	Deep-Net	Scat. + CNN
CIFAR-10	7%	7%

Image Classification: ImageNet 2012

Oyallon, Belivovsky, Zagoruyko

1000 classes, 1.2 million labeled training images, of 224×224 pixels



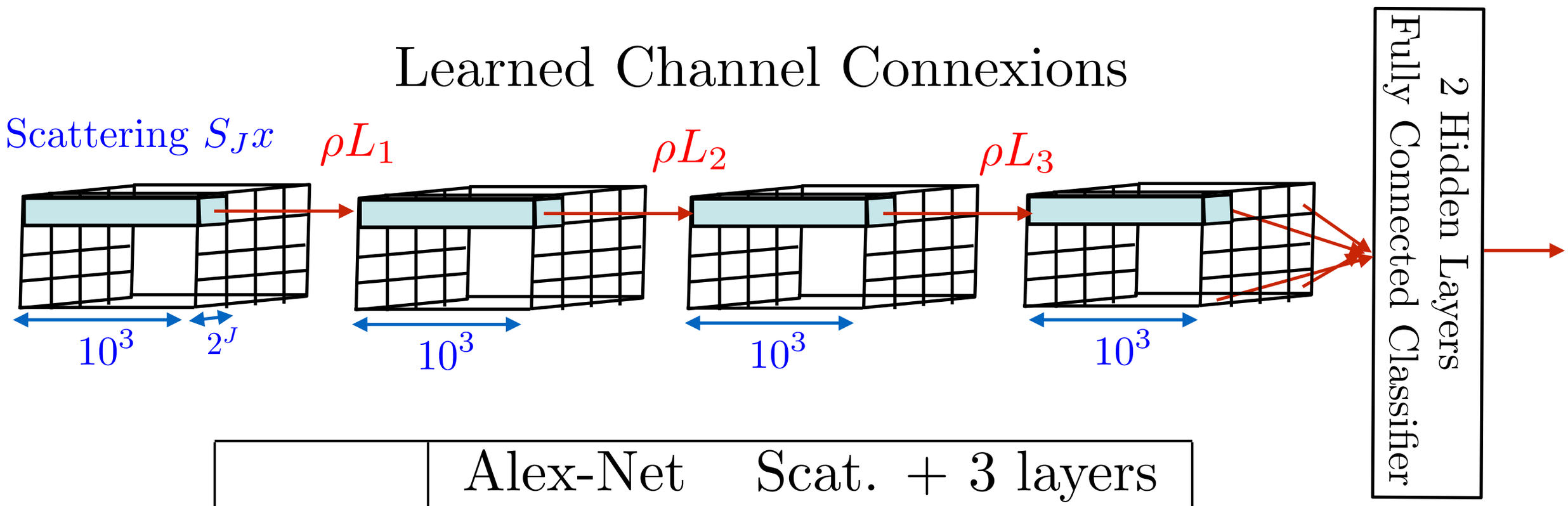
	Res-Net	Scat. + Res-Net
Top 1	30%	30%
Top 5	11%	11%

Structured Network: ImageNet 2012

Oyallon, Belivovsky, Zagoruyko

1000 classes, 1.2 million labeled training images, of 224×224 pixels

Learned Channel Connexions



	Alex-Net	Scat. + 3 layers
Top 1	43%	43%
Top 5	20%	20%

20 times faster

- Which invariants are learned and computed with the L_j ?
- Are the L_j storing some form of memory ?

Conclusions

- Deep convolutional networks have spectacular high-dimensional approximation capabilities. Seem to learn complex symmetries

- Can be further structured to use prior information.

- Close link with particle and statistical physics

- Outstanding mathematical problems to understand them:
  what are the classes of « learnable » functions and processes ?
  notion of complexity, approximation theorems…

Understanding Deep Convolutional Networks, arXiv 2016.