Interpretable Deep Neural Networks

Stéphane Mallat,

Collège de France
École Normale Supérieure

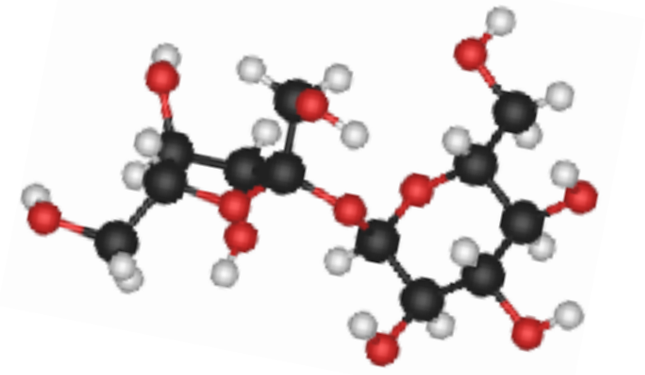
High-Dimensional Learning

- Supervised training: estimate $f(x)$ from $\{x_i, y_i = f(x_i)\}_{i \leq n}$.

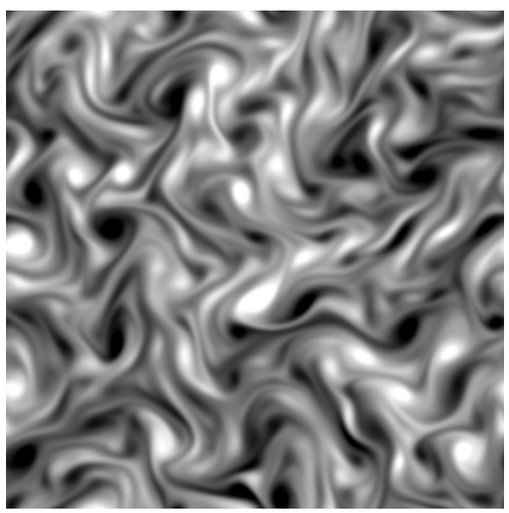
$f(x)$: class of an image $x \in \mathbb{R}^d$ having $d = 10^6$ pixels

or

energy of a physical system in a state $x \in \mathbb{R}^d$



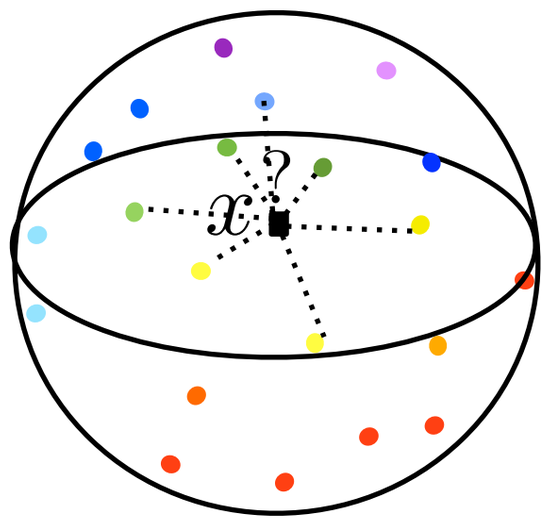
- Unsupervised training: estimate the probability $p(x)$ from $\{x_i\}_{i \leq n}$



What regularity of $f(x)$ or $p(x)$ lead to accurate estimations ?

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 $n \geq \epsilon^{-d}$ points to cover $[0,1]^d$ at a Euclidean distance ϵ

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

- To estimate $f(x)$ when x is in a high-dimensional Ω requires *strong regularity* of f in Ω : what regularity ?

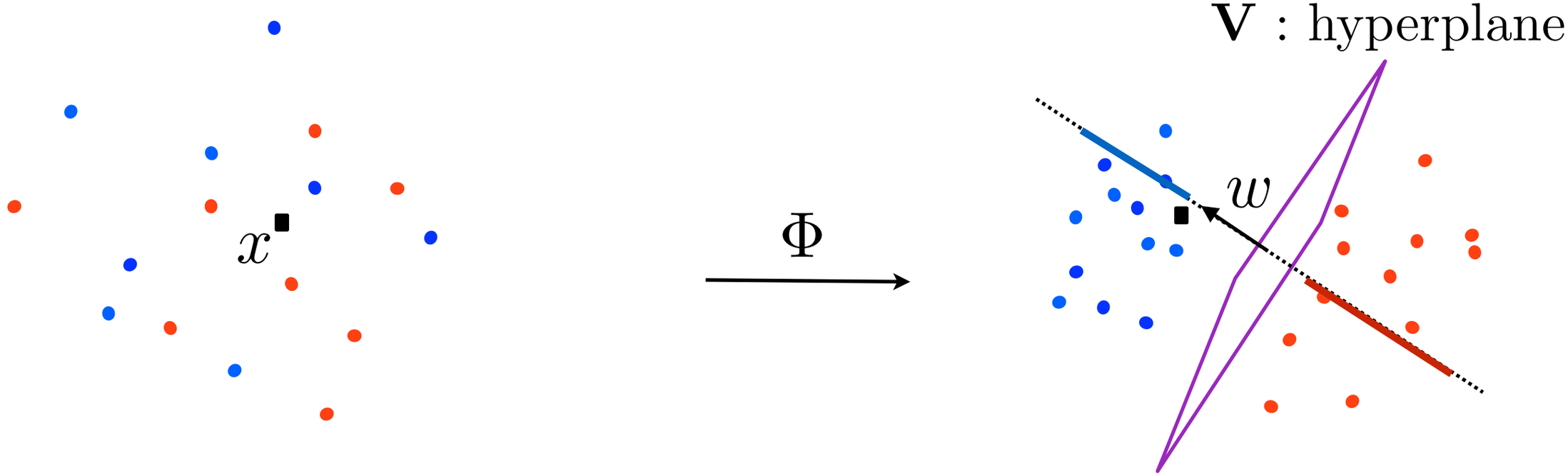
Kernel Classifiers

Change of variable $\Phi(x) = \{\phi_k(x)\}_{k \leq d'}$ to nearly linearize class boundaries, and approximate $f(x)$ by:

$$\tilde{f}(x) = \text{sign}(\langle w, \Phi(x) \rangle + b) = \text{sign}\Big(\sum_k w_k \, \phi_k(x) + b\Big)$$

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

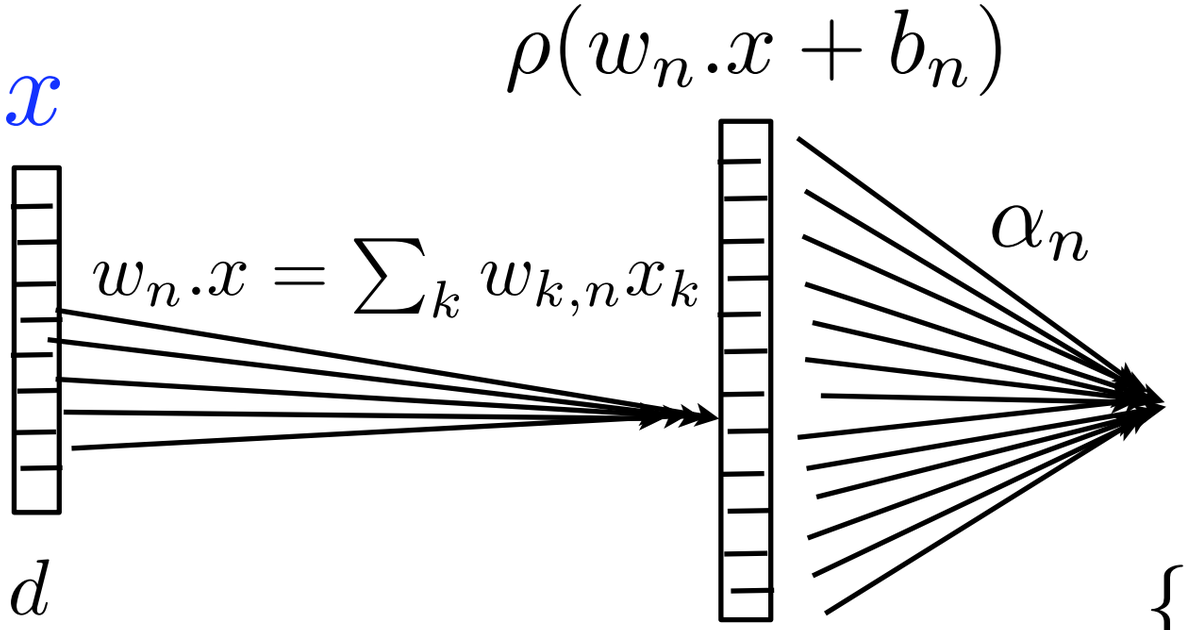
$\Phi(x) \in \mathbb{R}^{d'}$

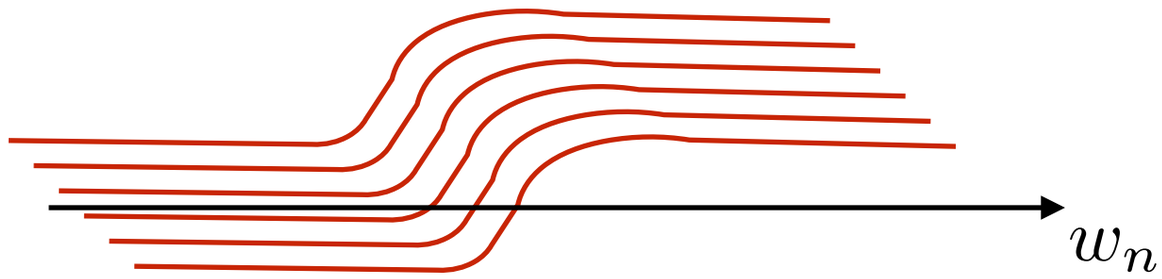


- How and when is possible to find such a Φ ?

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, Pinkus, Schocken...

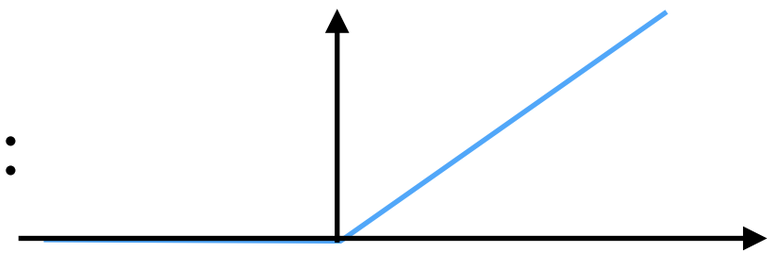
Theorem: For any continuous, non polynomial $\rho(u)$ and appropriate choices of $w_{n,k}$ and α_n:

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

No big deal: curse of dimensionality still there.

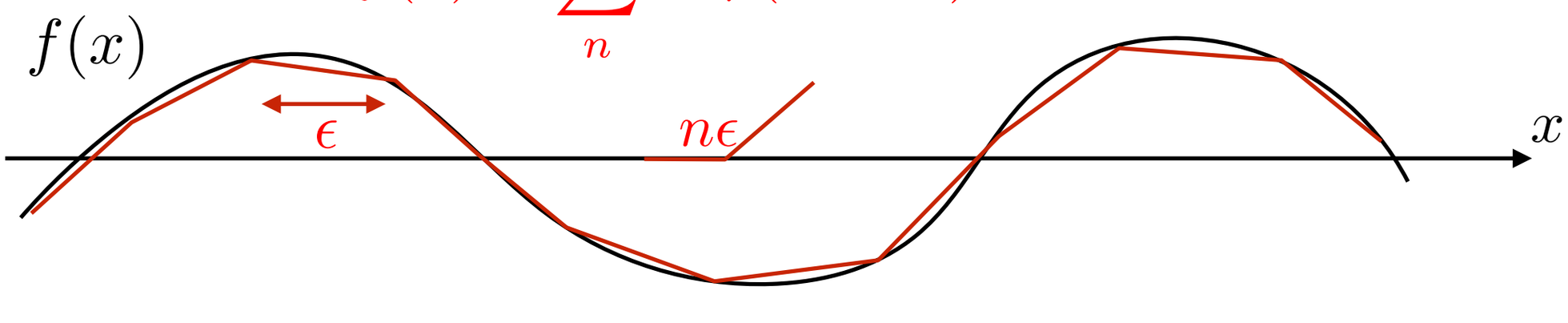
Piecewise Linear Approximation

- Piecewise linear approximation:



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

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



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

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

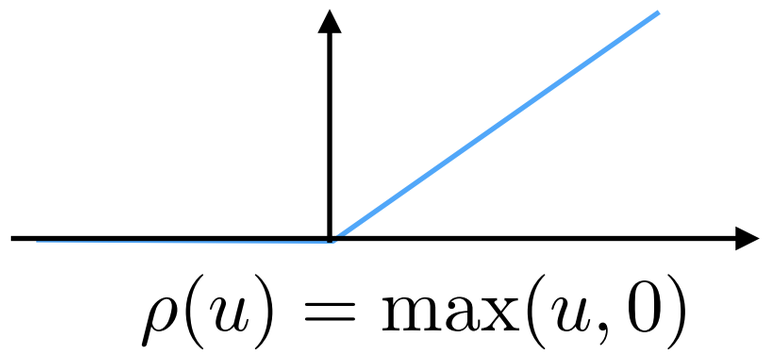
Need $M = \epsilon^{-1}$ points to cover $[0, 1]$ at a distance ϵ

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

Linear Ridge Approximation

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

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



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

Sampling at a distance ϵ:

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

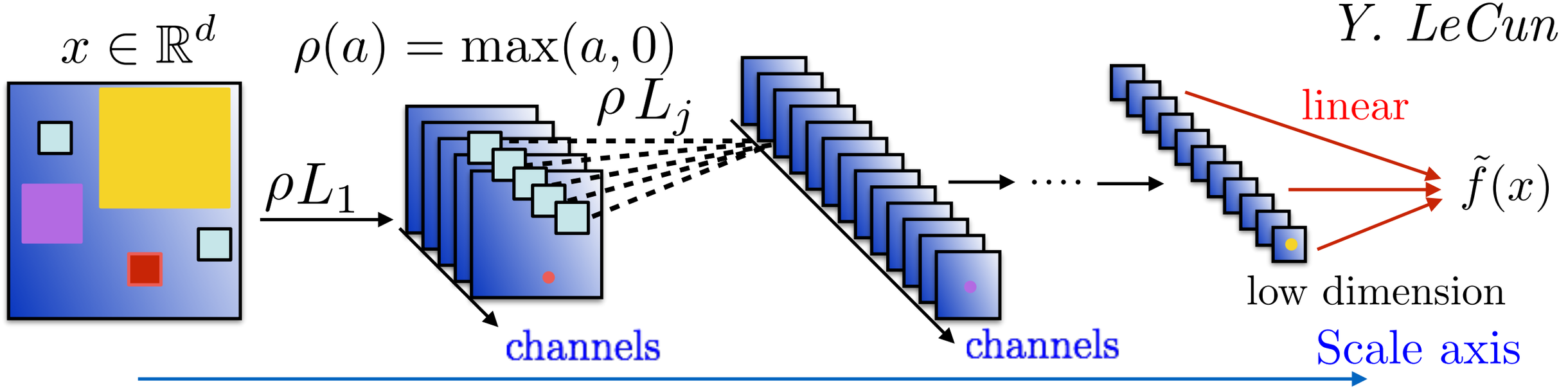
need $M = \epsilon^{-d}$ points to cover $[0,1]^d$ at a distance ϵ

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

Curse of dimensionality!

Deep Convolutional Network

- Deep convolutional neural network to predict $y = f(x)$:



L_j: spatial convolutions and linear combination of channels

Exceptional results for classification of *images, sounds, language, regressions in physics, signal and image generation...*

but not understood: optimisation and approximation problems.

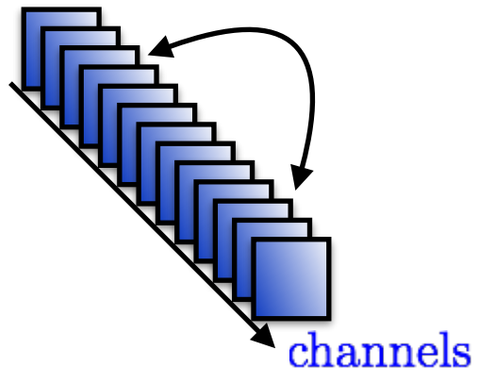
To create simpler interpretable networks:

What underlying regularity is captured and how ?

3 ingredients: Multiscale, Linearize group actions, Sparse

Statistical Models from 1 Example

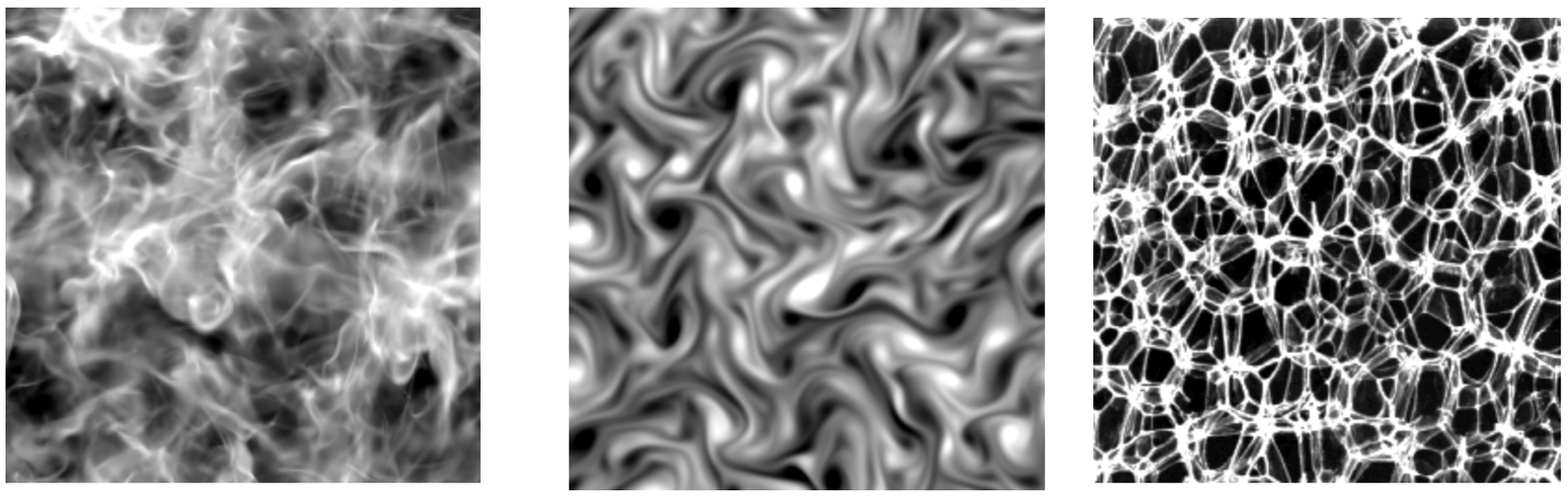
M. Bethdge et. al.



- Supervised network training (ex: on ImageNet)
- For 1 realisation x of X, compute each layer
- Compute autocorrelation across channels
- Synthesize $\tilde{x}$ having similar statistics

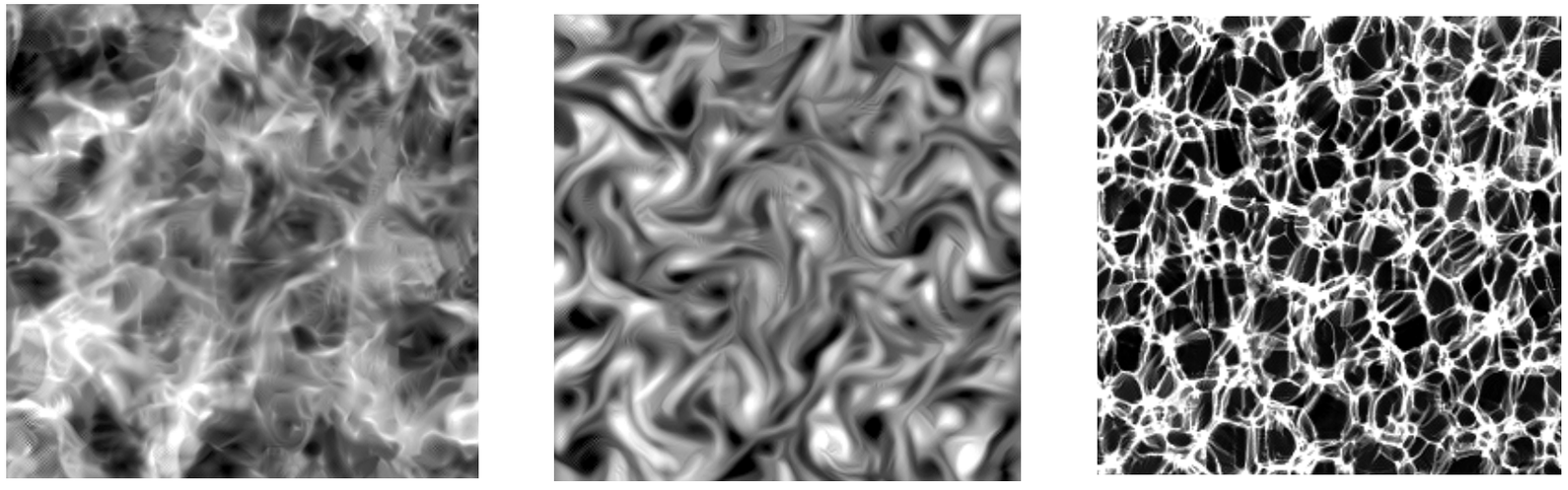
x

$6\,10^4$ pixels



$\tilde{x}$

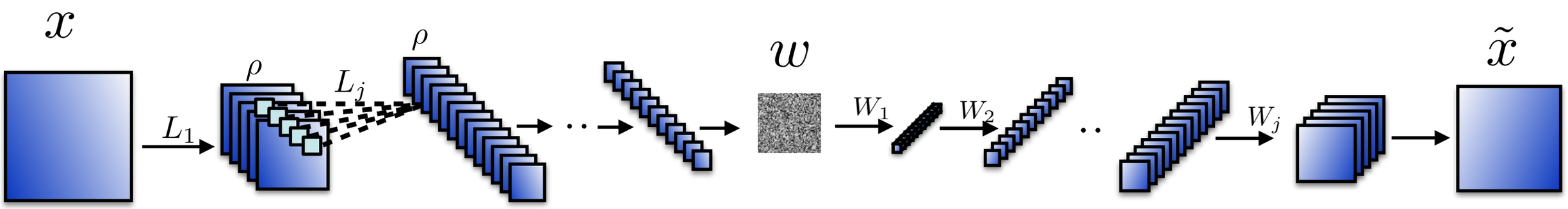
$2\,10^5$ correlations



What mathematical interpretation ?

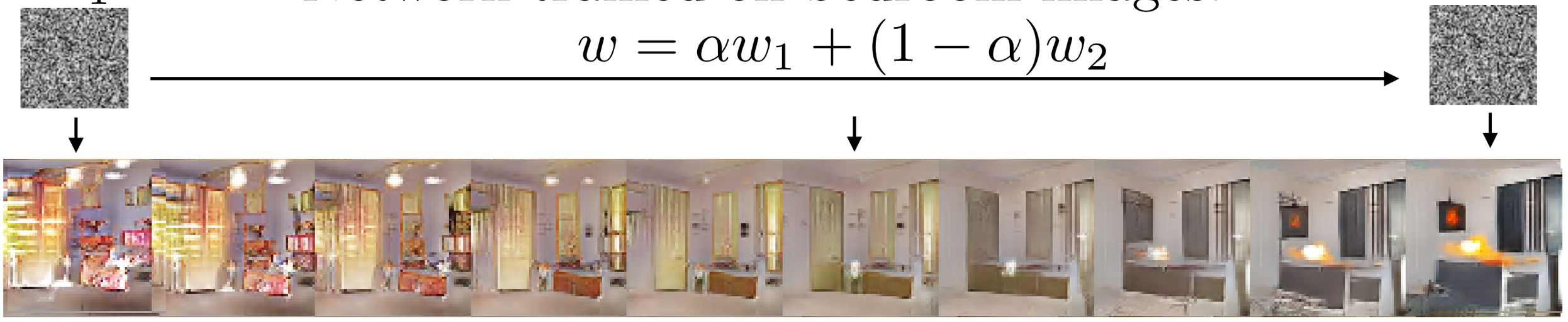
Non Ergodic Processes

autoencoder: trained on n examples $\{x_i\}_{i \leq n}$

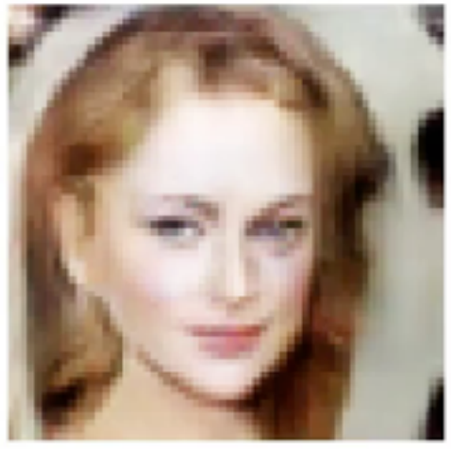


Network trained on bedroom images:

$$w = \alpha w_1 + (1 - \alpha) w_2$$



Network trained on faces of celebrities:



What mathematical interpretation ?

Overview: Simpler Networks

- What is the use of the different network « components » ?

- Scale separation with wavelets and interactions through phase

- Relu: without learning
  - Statistical physics for turbulence
  - Quantum chemistry and simple image classification

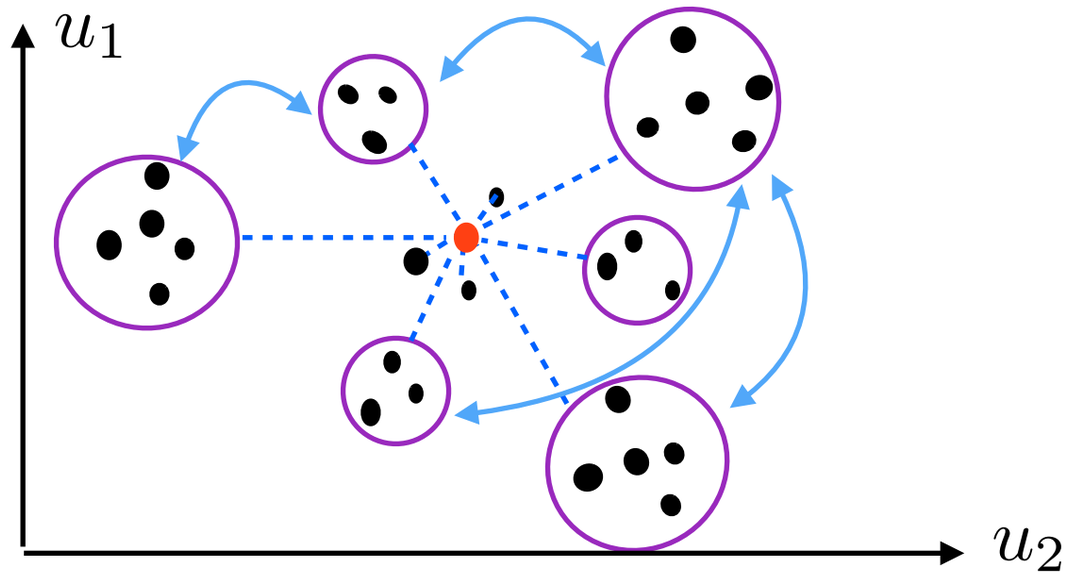
- Relu for learning: with sparse dictionaries
  - Classification of complex structures as in ImageNet
  - Generation with autoencoders

Scale Separation and Interactions

- **Dimension reduction:**

Interactions de d bodies represented by $x(u)$: particles, pixels...

Interactions across scales



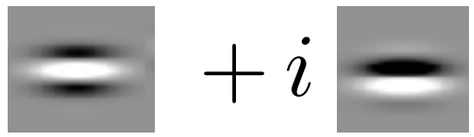
Multiscale regroupement of interactions of d bodies into interactions of $O(\log d)$ groups.

Scale separation $\Rightarrow$ wavelet transforms.

How to capture scale interactions ?

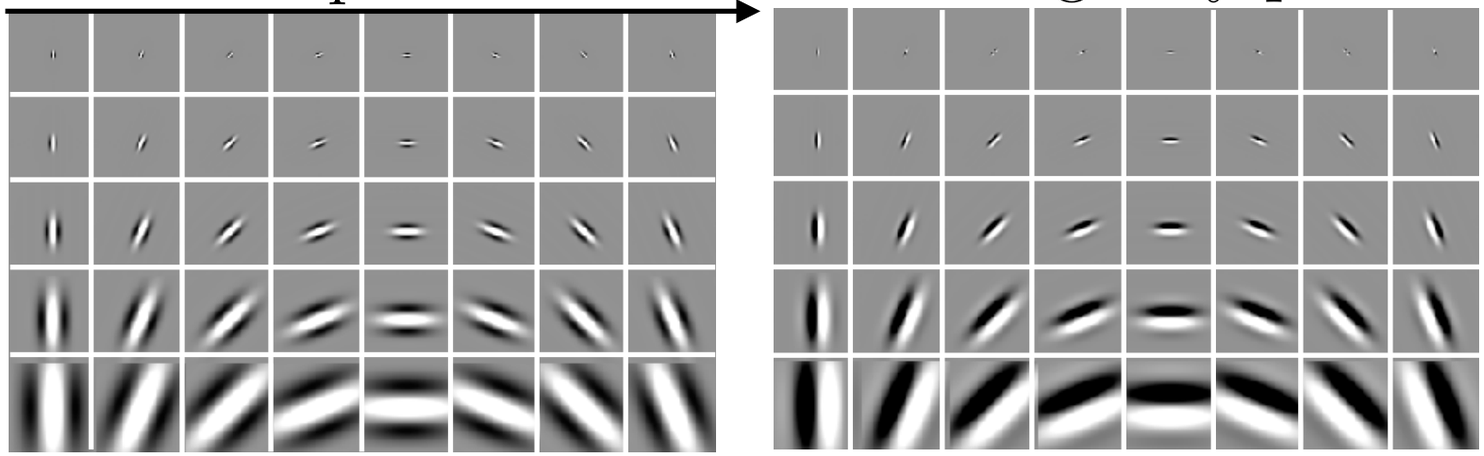
Critical harmonic analysis problems since 1970's

Scale separation with Wavelets

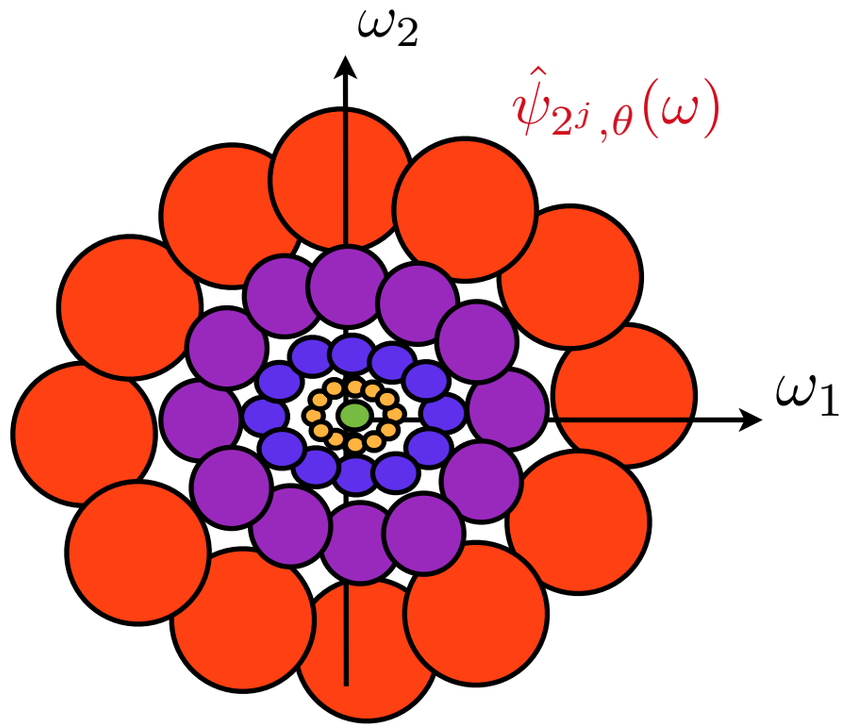
- Wavelet filter $\psi(u)$:  $+ i$

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

real parts θ imaginary parts



Fourier



- Wavelet transform: invertible

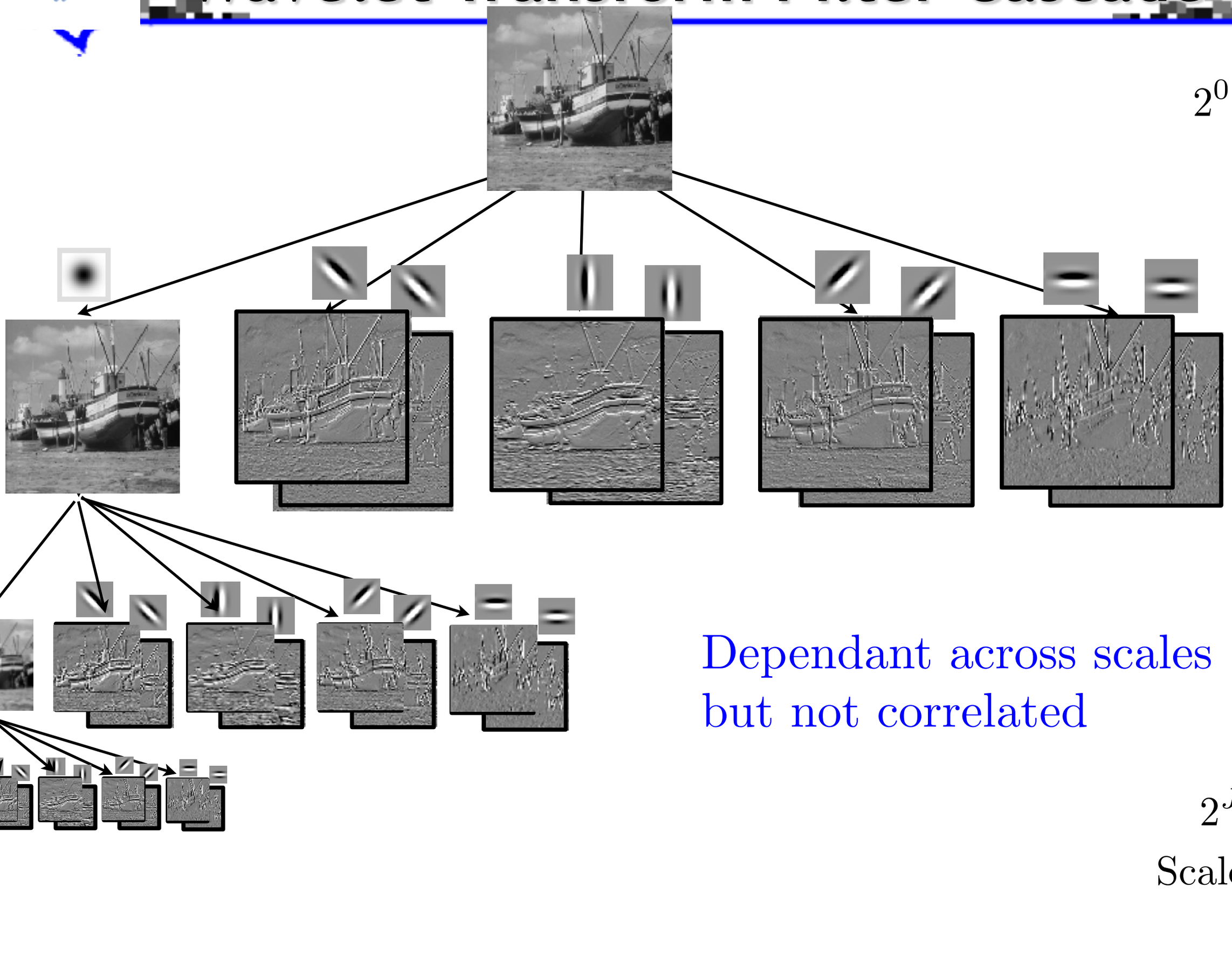
$$Wx = \Big(x \star \psi_\lambda\Big)_\lambda$$

$$\widehat{x \star \psi_\lambda}(\omega) = \hat{x}(\omega)\,\hat{\psi}_\lambda(\omega)$$

- Zero-mean and no correlations across scales: problem!

$$\sum_u x \star \psi_\lambda(u)\, x \star \psi^*_{\lambda'}(u) = \sum_\omega |\widehat{x}(\omega)|^2\; \psi_\lambda(\omega)\, \psi_\lambda(\omega)^* \approx 0 \quad \text{if} \quad \lambda \neq \lambda'$$

Wavelet Transform Filter Cascade

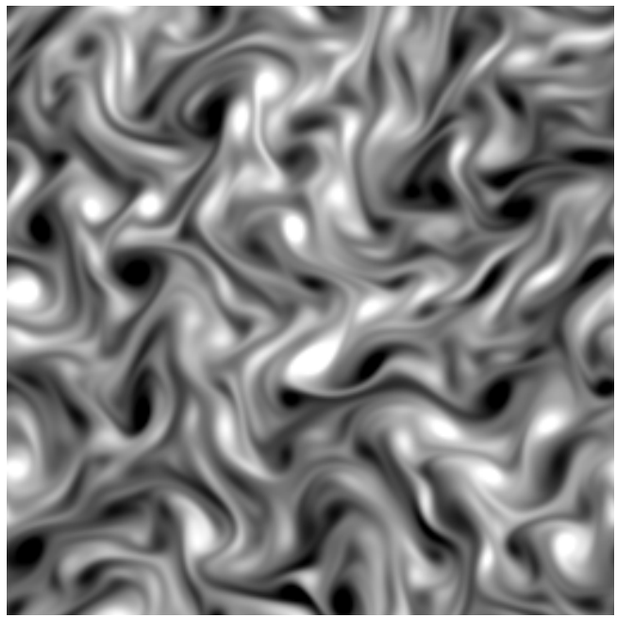


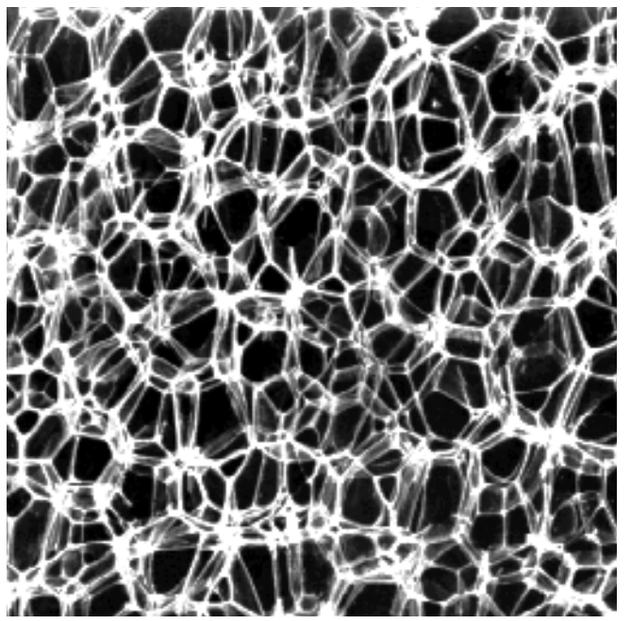
Dependant across scales
but not correlated

Stat. Physics of Stationary Proc.

What stochastic models for turbulence ?

$$d = 6\,10^4 \qquad x$$





Maximum entropy distribution $\tilde{p}$ conditioned by M moments

$$\max -\int \log \tilde{p}(x)\,\tilde{p}(x)\,dx \quad \text{with} \quad \mathbb{E}(\phi_m(x)) = \mu_m \quad 1 \le m \le M$$

$$\Rightarrow \qquad \tilde{p}(x) = \mathcal{Z}^{-1} \exp\Big(-\sum_{m=1}^{M} \beta_m\,\phi_m(x) \Big)$$

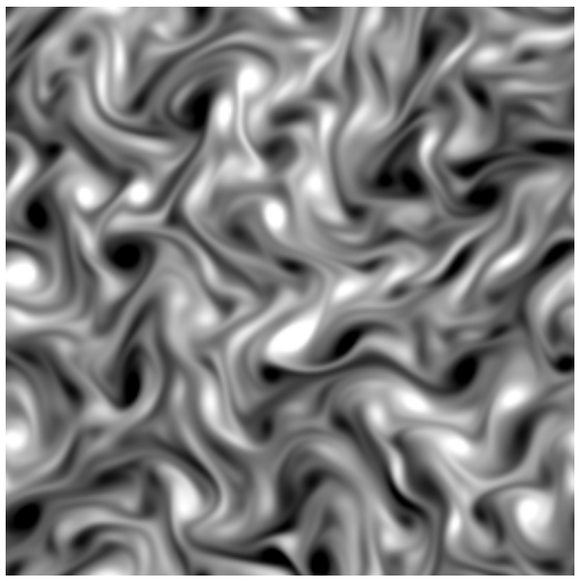
With $M = d$ second order moments:

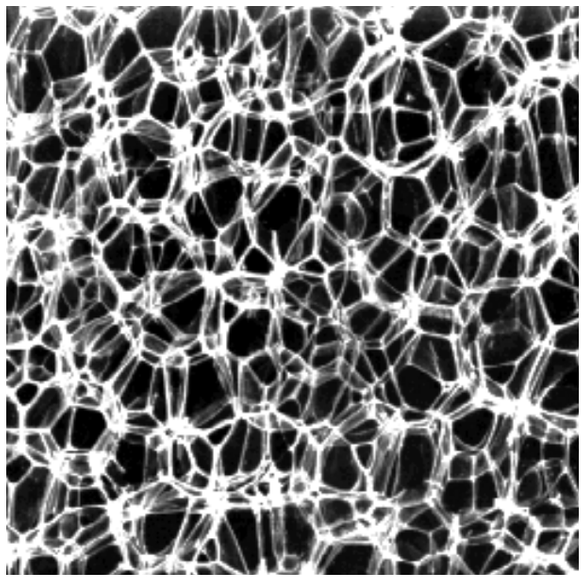
$\phi_m(x) = \sum_u x(u)x(u-m) \quad \Rightarrow \tilde{p}(x)$ is a Gaussian distribution

Gaussian Models of Stationary Proc.

What stochastic models for turbulence ?

$$d = 6\,10^4$$

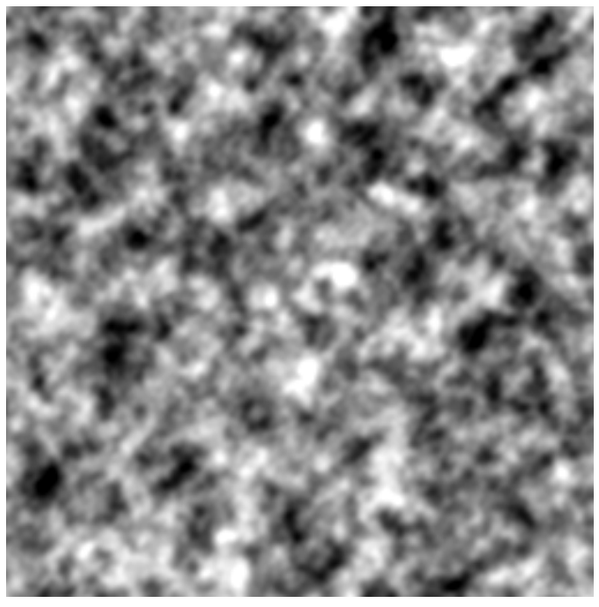


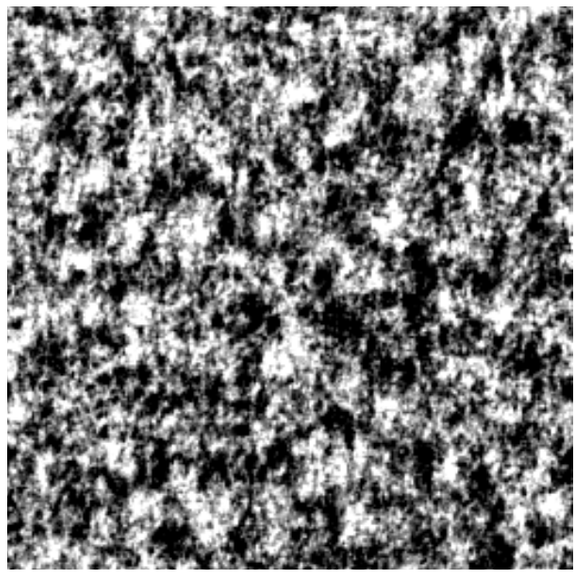


$\tilde{p}(x)$ is a Gaussian distribution

x

$\tilde{x}$





No correlation is captured across scales and frequencies.
Random phases.

How to capture non-Gaussianity and long range interactions ?

Failure of high order moments. Deep net generations look better.

Rectifiers act on Phase

Gaspar Rochette, Sixin Zhang

- Real wavelets of phase α: $\psi_{\alpha,\lambda} = \text{Real}(e^{-i\alpha}\,\psi_\lambda)$

  Rectifier: $\rho(a) = \max(a, 0)$

  $$Ux(u,\alpha,\lambda) = \rho(x \star \text{Real}(e^{i\alpha}\,\psi_\lambda)) = \rho(\text{Real}(e^{i\alpha}\,x \star \psi_\lambda))$$

  $$x \star \psi_\lambda = |x \star \psi_\lambda|\, e^{i\varphi(x\star\psi_\lambda)}$$

  Homogeneous: $\rho(\alpha a) = \alpha\,\rho(a)$ if $\alpha > 0$

  $$Ux(u,\alpha,\lambda) = |x \star \psi_\lambda|\,\rho(\cos(\alpha + \varphi(x \star \psi_\lambda))$$

A Relu computes phase harmonics:

Theorem : Fourier transform along the phase α:

$$\widehat{U}x(u,k,\lambda) = \hat{\gamma}(k)\,|x \star \psi_\lambda(u)|\,e^{ik\,\varphi(x\star\psi_\lambda(u))}$$

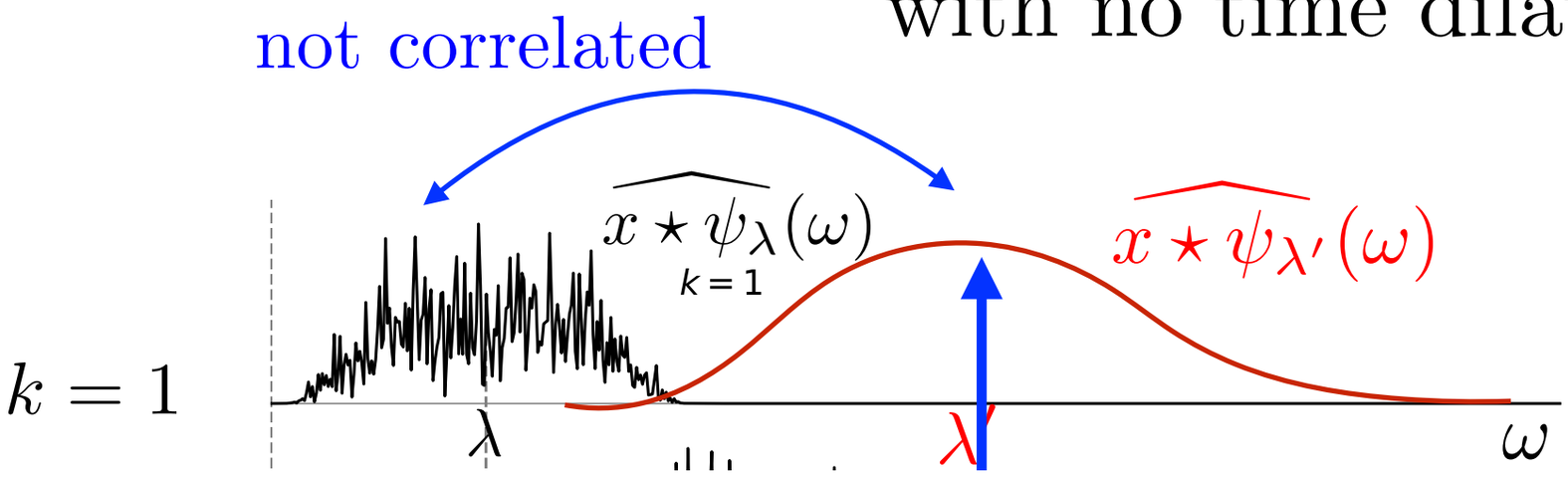
with $\gamma(\alpha) = \rho(\cos\alpha)$ for any homogeneous non-linearity ρ.

Frequency Transpositions

Real wavelets: $\psi_{\alpha,\lambda} = \mathrm{Real}(e^{-i\alpha}\,\psi_\lambda)$ and $\rho(a) = \max(a, 0)$

$$\rho(x \star \psi_{\alpha,\lambda}) \xrightarrow[\text{along } \alpha]{\text{Fourier transform}} c_k\,|x \star \psi_\lambda|\,e^{ik\,\varphi(x\star\psi_\lambda)}$$

Performs a non-linear frequency dilation / transposition with no time dilation



Phase Harmonics

Correlated if $k\lambda \approx \lambda'$

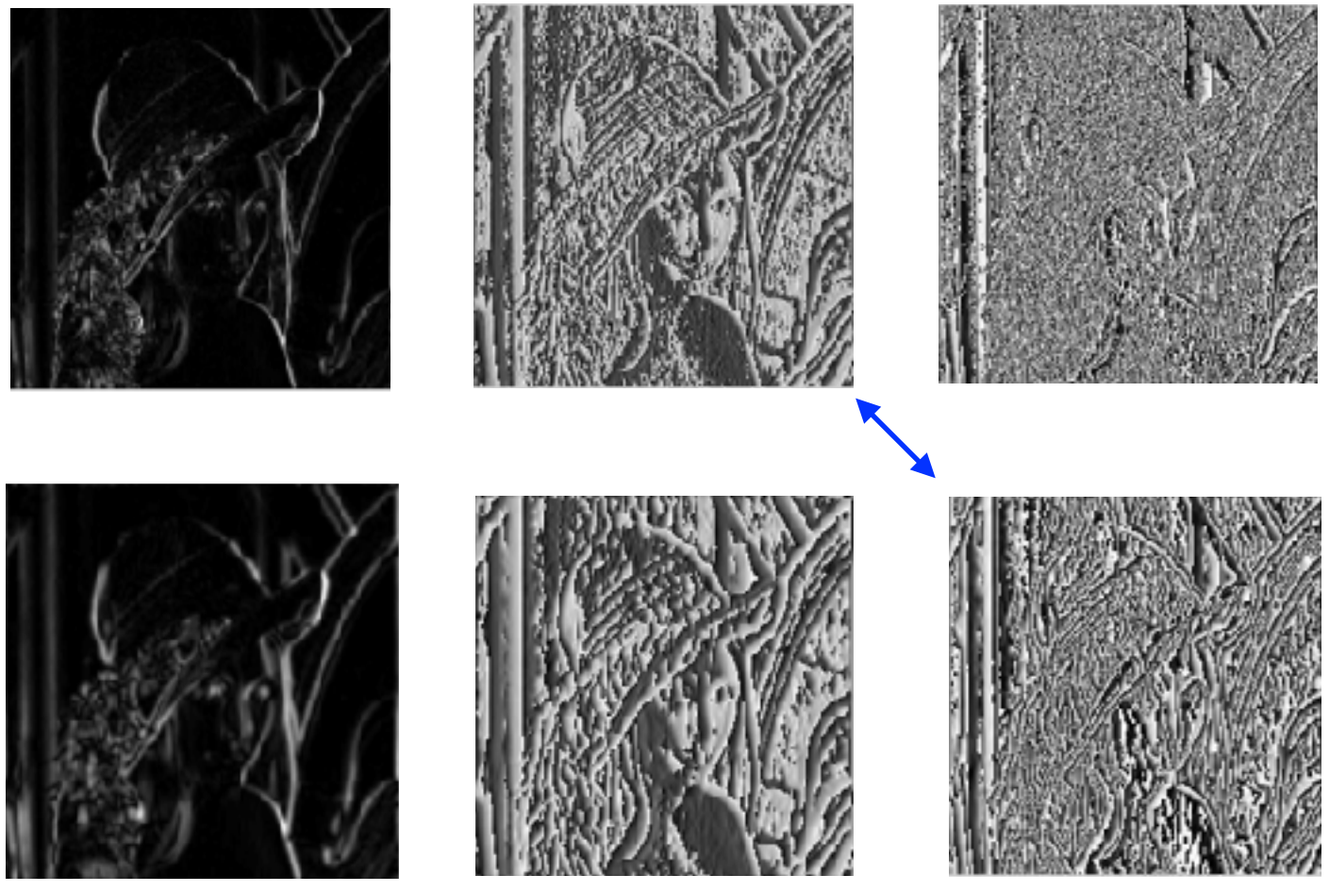
Rectified Wavelet Coefficients

Real wavelets: $\psi_{\alpha,\lambda} = \mathrm{Real}(e^{-i\alpha}\,\psi_\lambda)$ and $\rho(a) = \max(a, 0)$

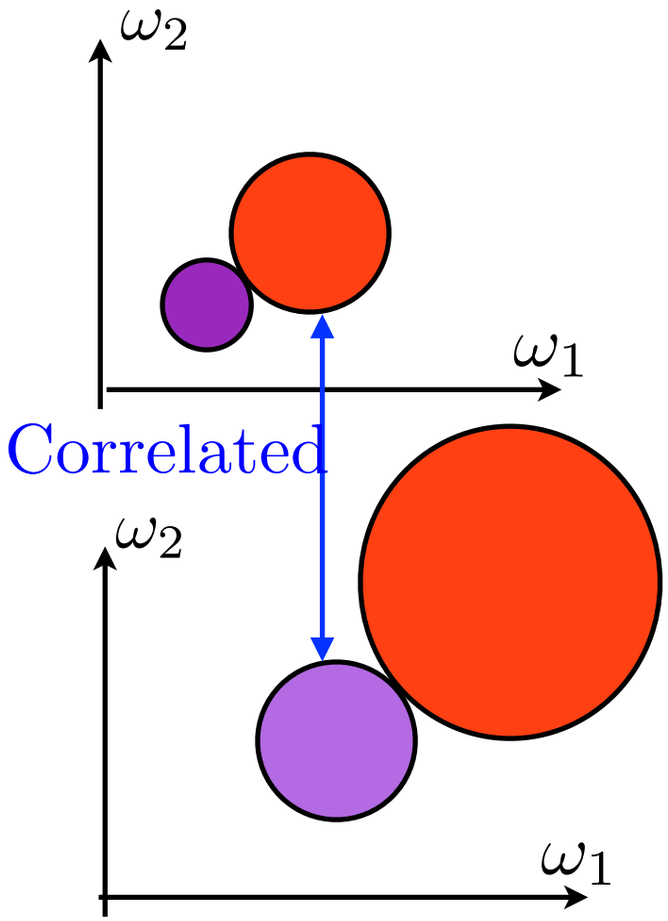
$$\rho(x \star \psi_{\alpha,\lambda}) \xrightarrow[\text{along } \alpha]{\text{Fourier transform}} c_k\,|x \star \psi_\lambda|\,e^{ik\,\varphi(x\star\psi_\lambda)}$$

Harmonics

$|x \star \psi_\lambda|$ $\quad$ $\varphi(x \star \psi_\lambda)$ $\quad$ $k\,\varphi(x \star \psi_\lambda)$



$k = 2$

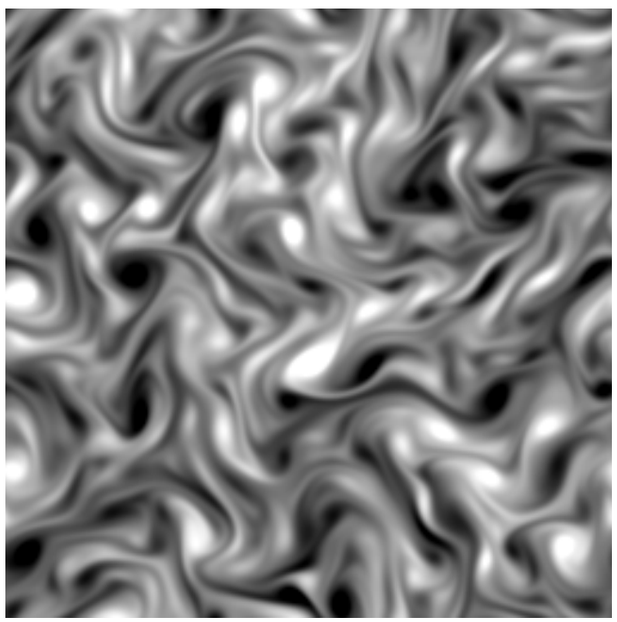


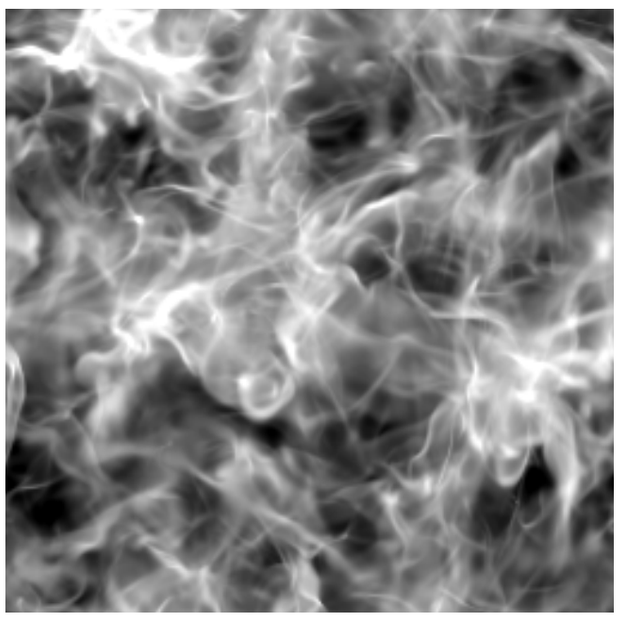
Phase harmonics:
Frequency transpositions

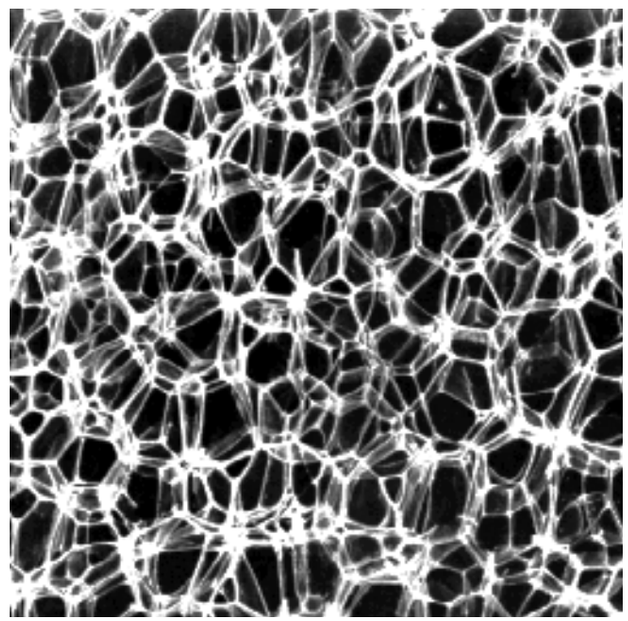
Models of Stationary Processes

Sixin Zhang

x







Maximum entropy distribution conditioned by

$M = O(\log^2 d)$ wavelet harmonic correlations $\mathbb{E}(\phi_m(x))$

$$\phi_m(x) = \sum_u \rho(x \star \psi_{\alpha,\lambda}(u))\, \rho(x \star \psi_{\lambda',\alpha'}(u))$$

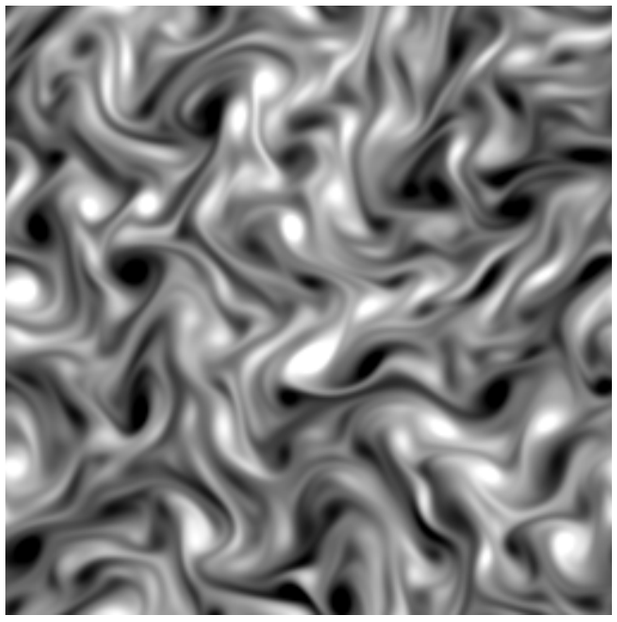
$$\tilde{p}(x) = \mathcal{Z}^{-1} \exp\Big(- \sum_{m=1}^{M} \beta_m\, \phi_m(x) \Big)$$

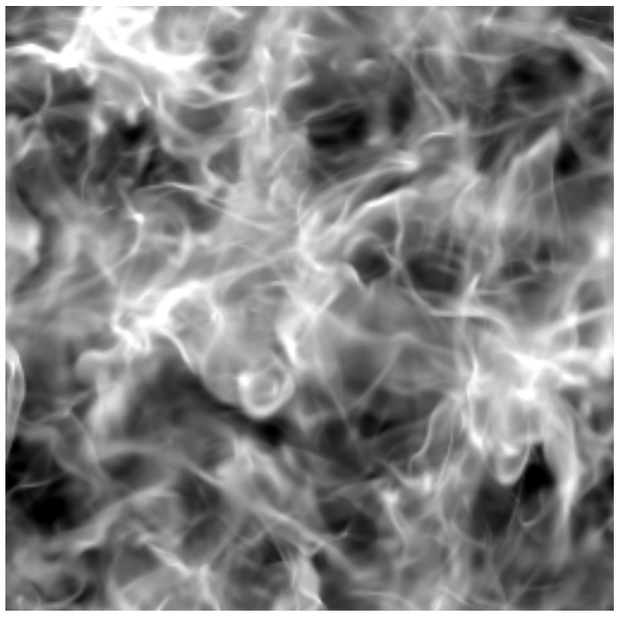
Ergodic Stationary Processes

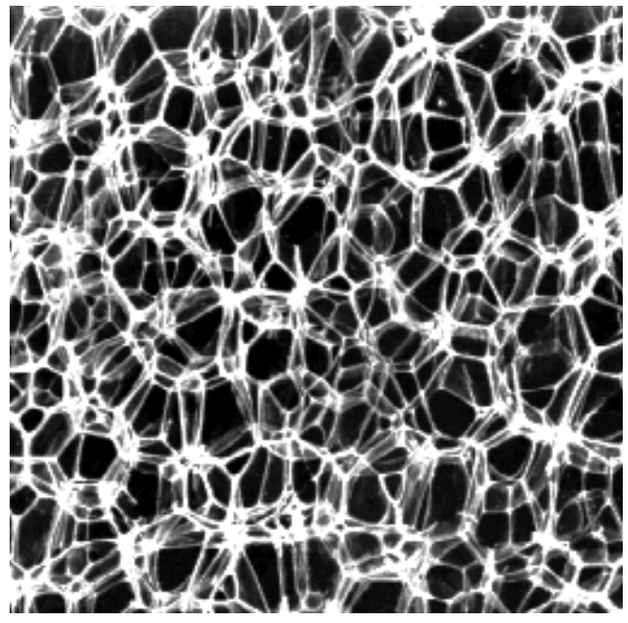
S. Zhang, J. Bruna, E. Allys, F. Levrier, F. Boulanger

$d = 6\,10^4$

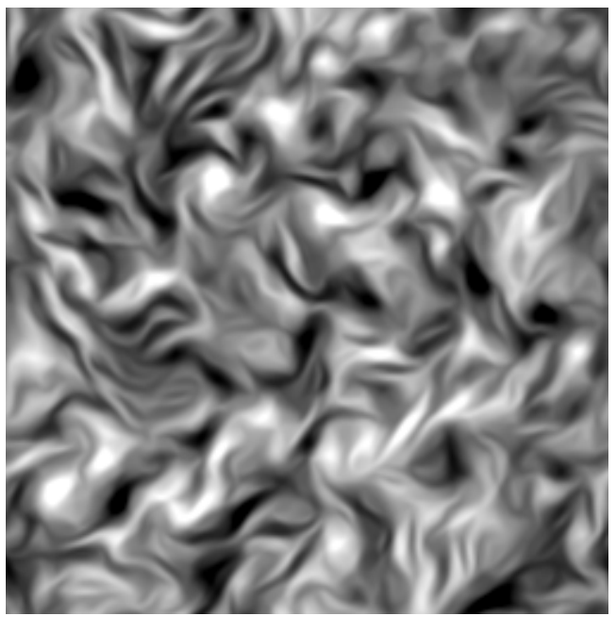
x

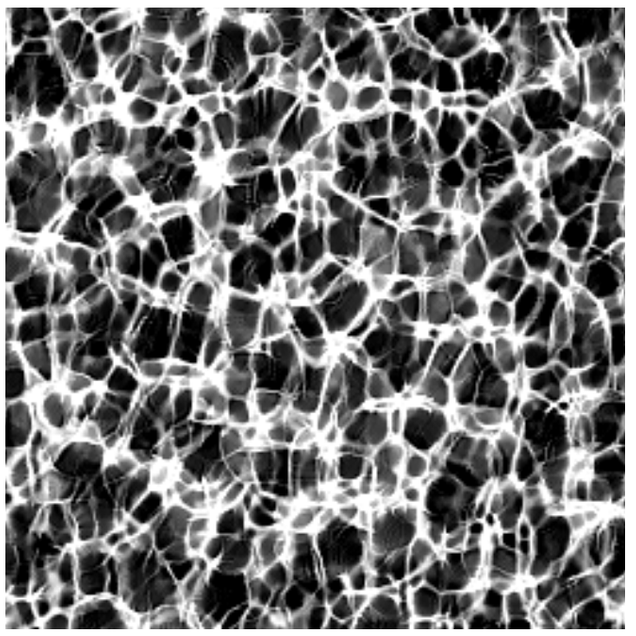






$\tilde{x}$





$M = 3\,10^3$
number
of moments

Phase coherence is restored

How much physics are these models capturing ?

What about non-ergodic processes ?

Classification: Scattering Wavelets

Classification: invariance by translation by spatial averaging

$$\begin{pmatrix} x \star \phi_{2^J}(2^J n) \\ \rho(x \star \psi_{\alpha,\lambda}) \star \phi_J(2^J n) \end{pmatrix}_{\alpha,\lambda}$$

Recover the information loss with a second layer:

$$S_J x = Ux \star \phi_J = \begin{pmatrix} x \star \phi_{2^J}(2^J n) \\ \rho(x \star \psi_{\alpha,\lambda}) \star \phi_J(2^J n) \\ \rho(\rho(x \star \psi_{\alpha,\lambda}) \star \psi_{j',\alpha'}) \star \phi_J(2^J n) \end{pmatrix}_{\alpha,\lambda,\alpha',\lambda'}$$

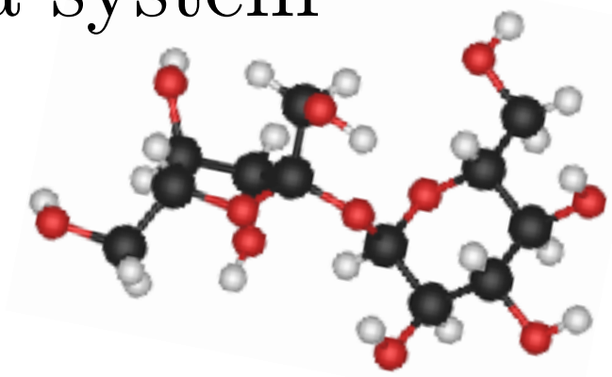
- Linearize small deformations

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

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

Quantum Chemistry: N-Body Problem

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



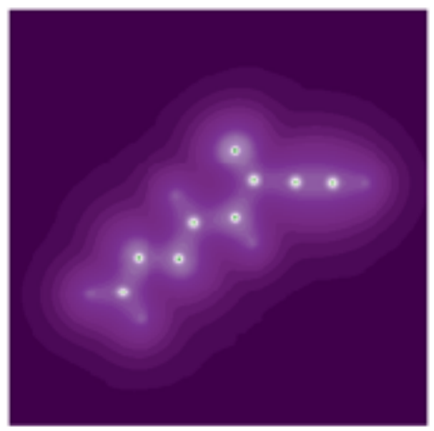
Symmetries:

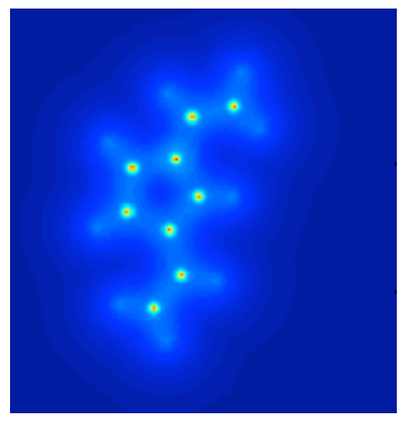
$f(x)$ is invariant to translations and rotations,

multiscale interactions: chemical bounds, Van der Waal forces...

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

Ground state
electronic density
computed with Schroedinger





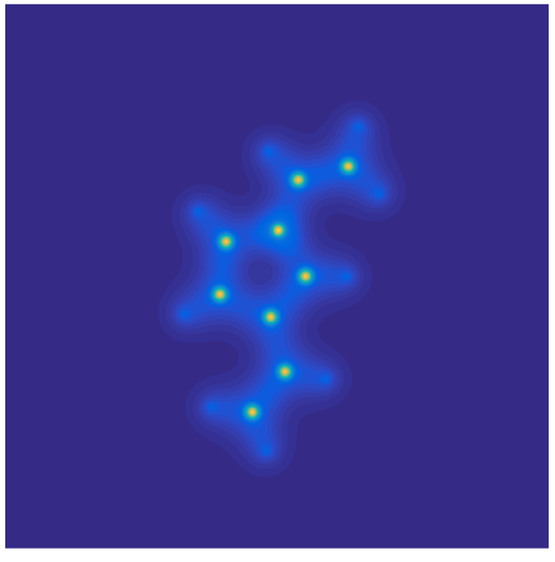
Dirac Electronic Density

- We do not know the electronic density at equilibrium.

- The molecular state $\{z_k\,,\,r_k\}_{k\leq d}$ is represented by Diracs located at r_k weighted by charges z_k:

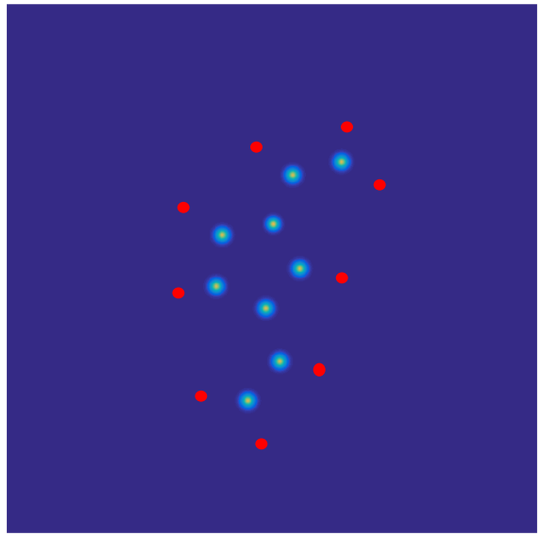
$$x(u) = \sum_{k=1}^{d} z_k\,\delta(u - r_k)$$

Electronic density



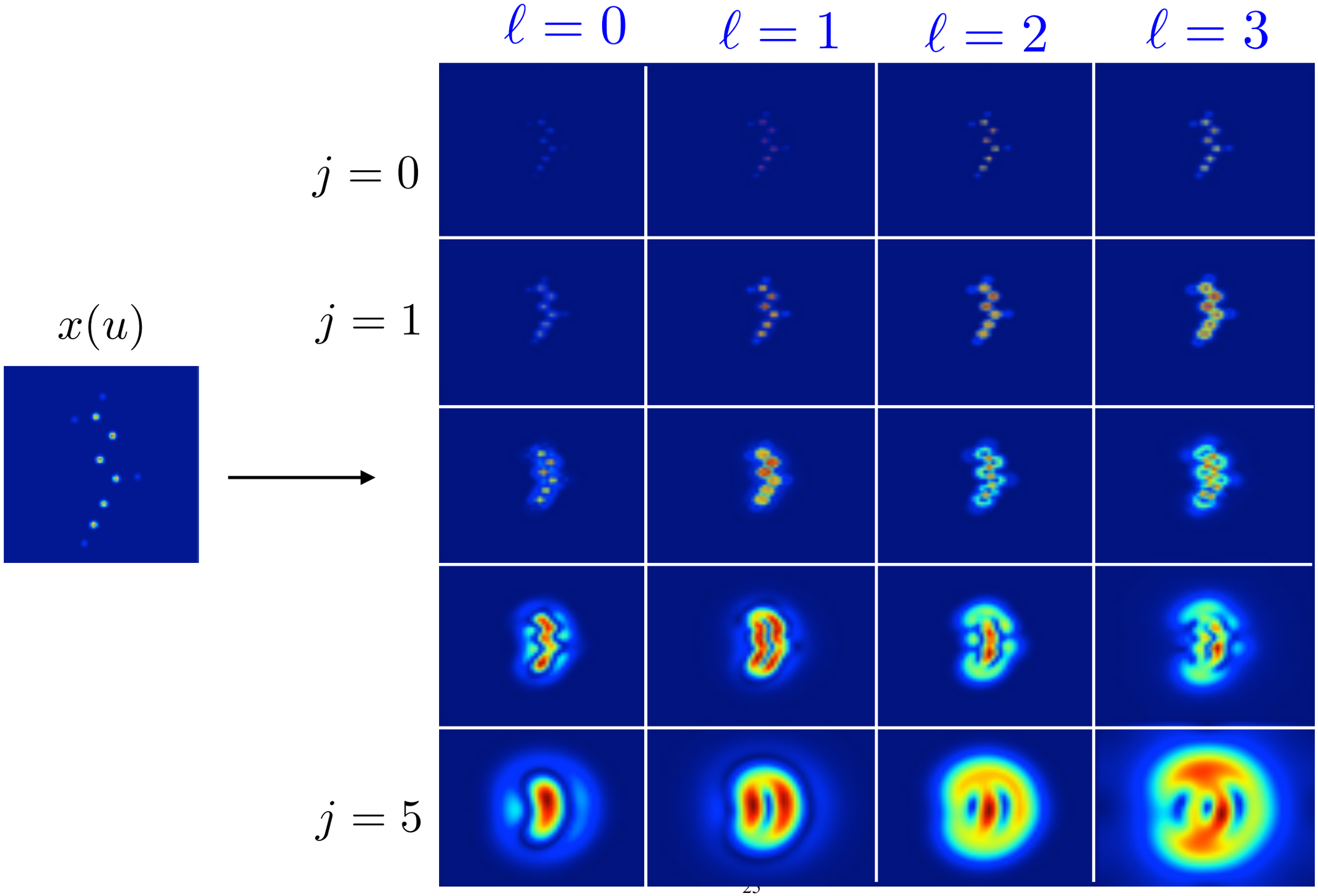
Dirac density

$x(u)$



Harmonic Wavelet Interferences

$$x = \sum_k z_k \delta(u - r_k) \Rightarrow \rho\Big(x \star \psi_{2^j,\ell}(u)\Big) = \rho\Big(\sum_k z_k \, \psi_{2^j,\ell}(u - r_k)\Big)$$



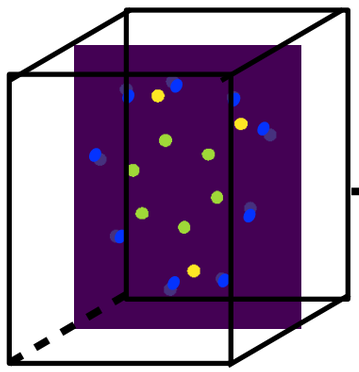
Scattering Energy Regression

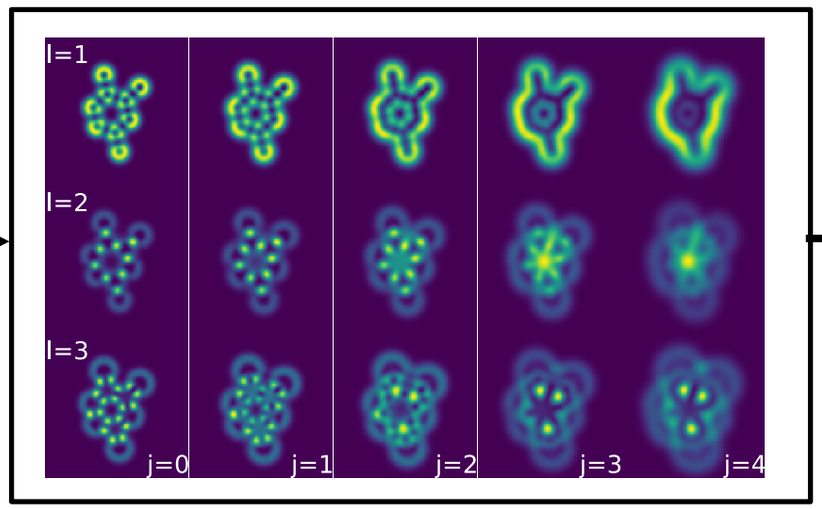
M. Eickenberg, G. Excarchakis M. Hirn, N. Poilvert, L. Thiry

Invariants to
Translations
Rotations

$$\rho(\rho(x \star \psi_{\alpha,\lambda}) \star \psi_{\alpha',\lambda'})$$

x





convolutional neural network without learning

Linear Averaging

Sx

$O(\log^2 d)$

Linear Regression

$\tilde{f}(x)$

supervised learning

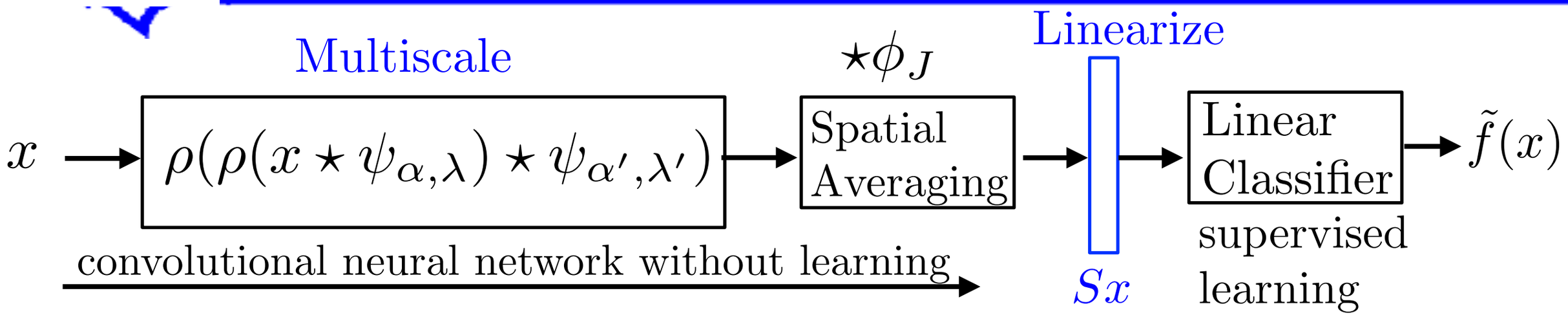
QM9: Data basis of 130.000 organic molecules with C, H, O, N, F with DFT atomisation energies

Regression error $\sim$ 0.5 kcal/mol $\sim$ Deep Nets.

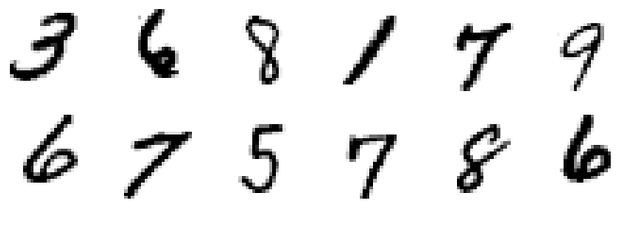
But small molecules with at most 29 atoms and 9 heavy ones

Image Classification

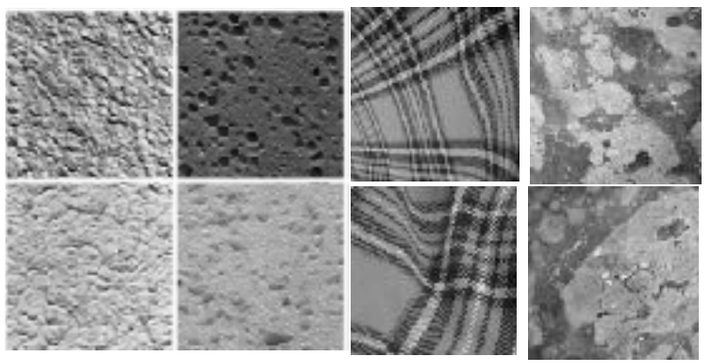
J. Bruna



Digits
10 classes



Textures
60 classes



ImageNet

10^3 classes

Errors: Scattering | Deep Nets.

0.5 | %

0.5 | %

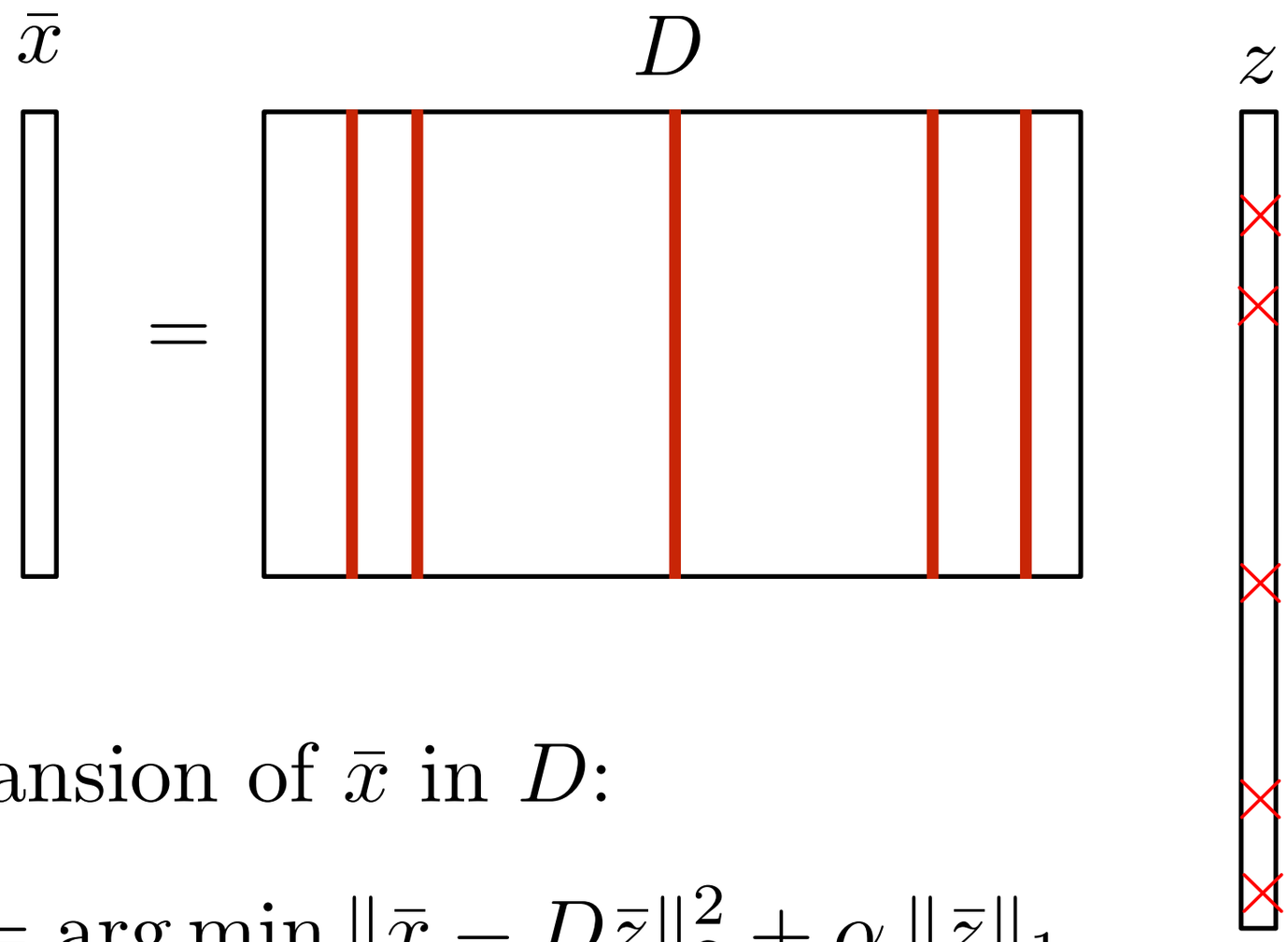
60% | AlexNet 20% : 2012

What is learned ?

Sparse Dictionary Representation

- Need to learn "sparse informative patterns"

  Pattern representations with sparse dictionary expansions:



Sparse $\mathbf{l^1}$ expansion of $\bar{x}$ in D:

$$z = \arg\min_{\bar{z}} \|\bar{x} - D\bar{z}\|_2^2 + \alpha \|\bar{z}\|_1$$

- How to minimise this convex cost ?

Homotopy ISTA Network

- Homotopy algorithms decrease the multiplier α_k:

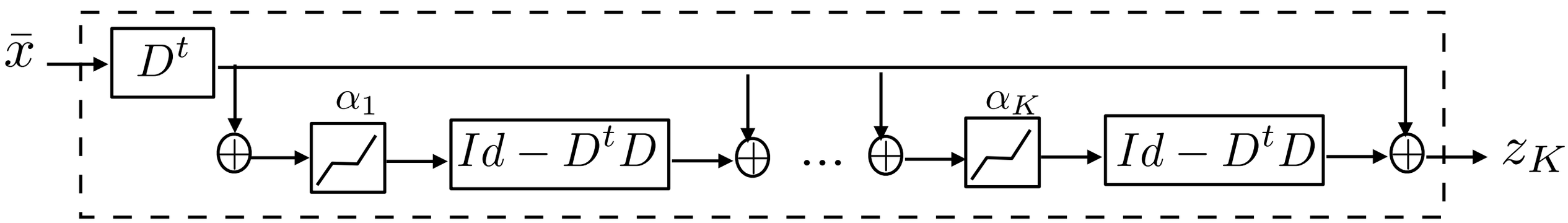
$$z = \arg\min_{\bar{z}} \|\bar{x} - D\bar{z}\|_2^2 + \alpha_k \|z\|_1$$

with an iterated soft-threshold decreasing thresholds:

$$z_{k+1} = T_{\alpha_k}(D^t\bar{x} + (I - D^tD)z_k) \xrightarrow[k \to \infty]{\alpha_k \sim \gamma^{-k}} z$$

where $T_\alpha(a) = \text{sign}(a)\ \max(|a| - \alpha, 0)$ is a soft-thresholding.

- Implemented with a convolutional network of depth K:

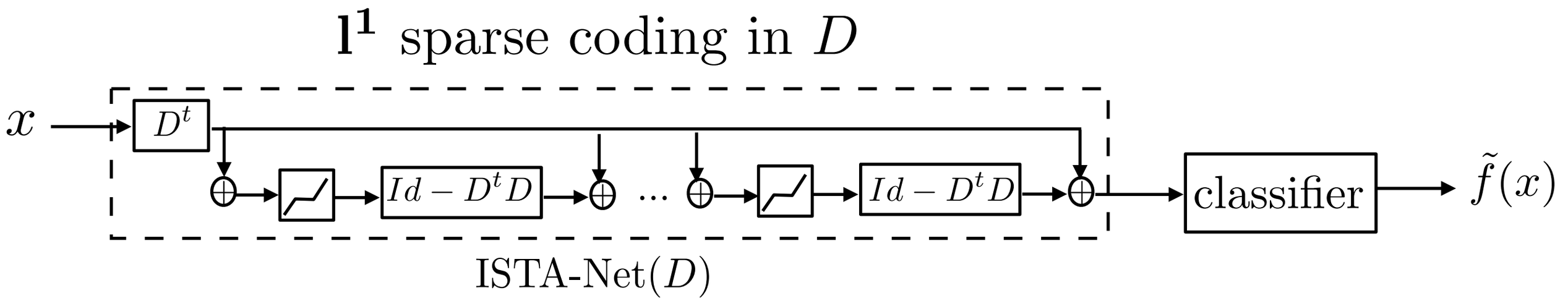


ISTA-Net(D)

with a convolutional dictionary D

Dictionary Learning for Classification

- Deep network with sparse coding and classification:



- Optimize the dictionary D and the classifier to minimize the classification loss over a supervised data basis $\{x_i, y_i\}_i$:

$$\text{Loss}(D) = \sum_i \text{loss}(y_i, \tilde{f}(x_i))$$

- Stochastic gradient descent

ImageNet Classification

J. Zarka, L. Thiry, T. Angles

ImageNet
10^3 classes

$$x \to \boxed{\rho(\rho(x \star \psi_{\alpha,\lambda}) \star \psi_{\alpha',\lambda'})} \to \boxed{\text{Spatial Averaging}} \xrightarrow{Sx} \boxed{\text{Classifier}} \to \tilde{f}(x)$$

60% error

Learned sparse coding

$$x \to \boxed{\text{ISTA-Net } D} \xrightarrow{z} \boxed{\text{Classifier}} \to \tilde{f}(x)$$

50% error

18% error

Multiscale Scattering — Linearize — Learned sparse coding

$$x \to \boxed{\rho(\rho(x \star \psi_{\alpha,\lambda}) \star \psi_{\alpha',\lambda'})} \to \boxed{\text{Spatial Averaging}} \xrightarrow{Sx} \boxed{\text{ISTA-Net } D} \xrightarrow{z} \boxed{\text{Classifier}} \to \tilde{f}(x)$$

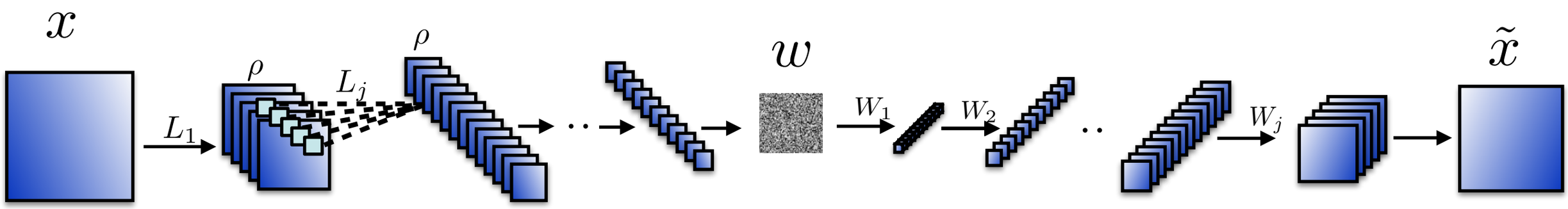
Interpretable network : patterns stored in D

AlexNet: 20% error

Why such an error reduction ? Linearize classes in separate spaces

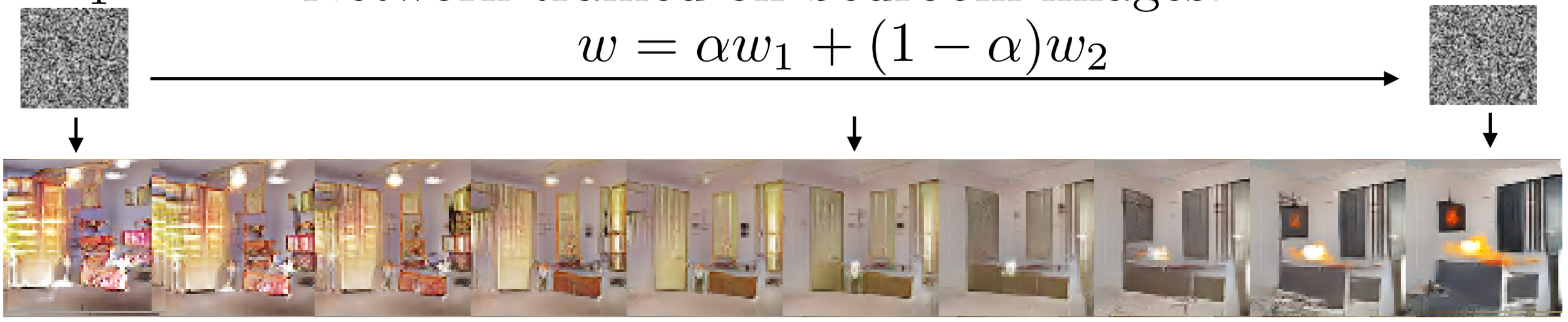
Non Ergodic Processes

autoencoder: trained on n examples $\{x_i\}_{i \leq n}$

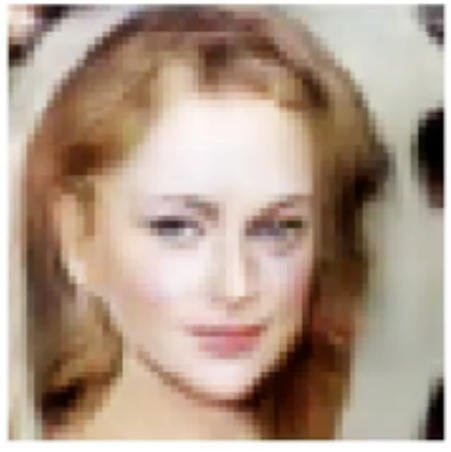


Network trained on bedroom images:

$$w = \alpha w_1 + (1 - \alpha) w_2$$



Network trained on faces of celebrities:



Generation as an Inverse Problem

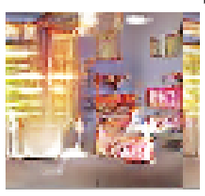
Tomas Angles *Florentin Guth*

Encoder

Multiscale

Nearly Gaussian

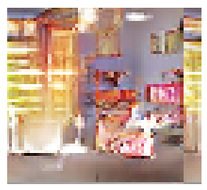
Low-dimensional Gauss. white noise



x → $\rho(\rho(x \star \psi_{\alpha,\lambda}) \star \psi_{\alpha',\lambda'})$ → Ux → Spatial Averaging → Linear Whitening →  $w = LUx$

L : linear mixing central limit theorem

Inversion:

$w = LUx$  → **Generator** → x 

U has a linear inverse U^{-1} : $\rho(a) + \rho(-a) = a$

L is non-invertible linear projector

Regularization: inversion in a dictionary D where Ux is sparse

Compute z such that $Ux = Dz$ where z is sparse

Non-linear multiscale model

Generation as an Inverse Problem

Tomas Angles Florentin Guth

Inversion:

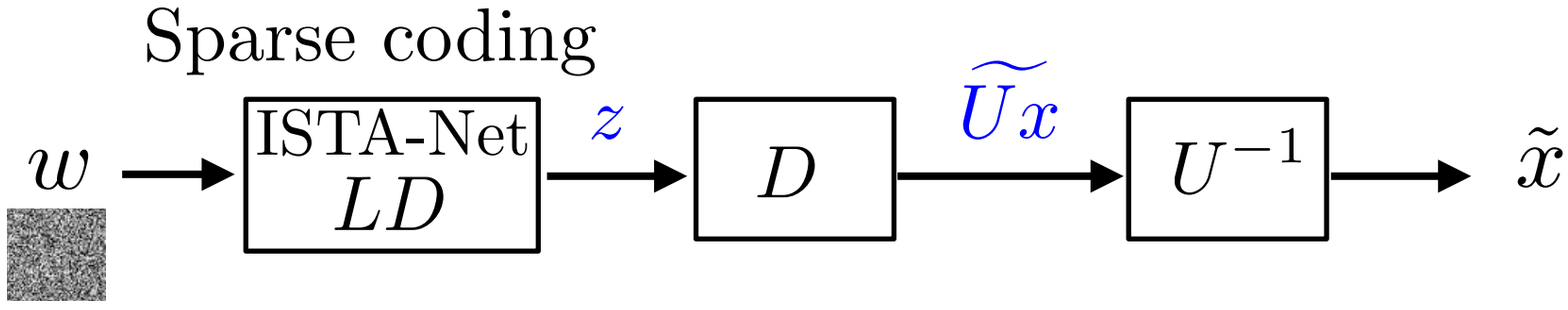
$$w = LUx$$

U has a linear inverse U^{-1}, L is non-invertible linear projector

Inversion in a dictionary D where Ux is sparse:

$$Ux = Dz \quad \Rightarrow \quad w = LUx = LDz$$

compute sparse z from w in LD



- How to optimise the dictionary D ?

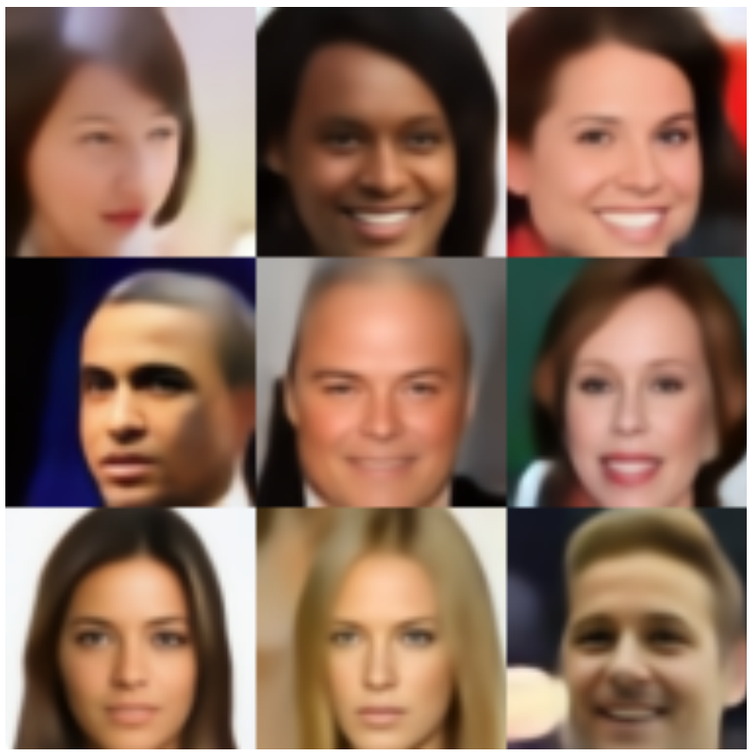
  Learn the dictionary D by minimizing $\sum_i \|x_i - \tilde{x}_i\|^2$

  with a stochastic gradient descent on a training set $\{x_i\}_i$

Training Reconstruction

Celebrities Data Basis

x_i



$\tilde{x}_i$

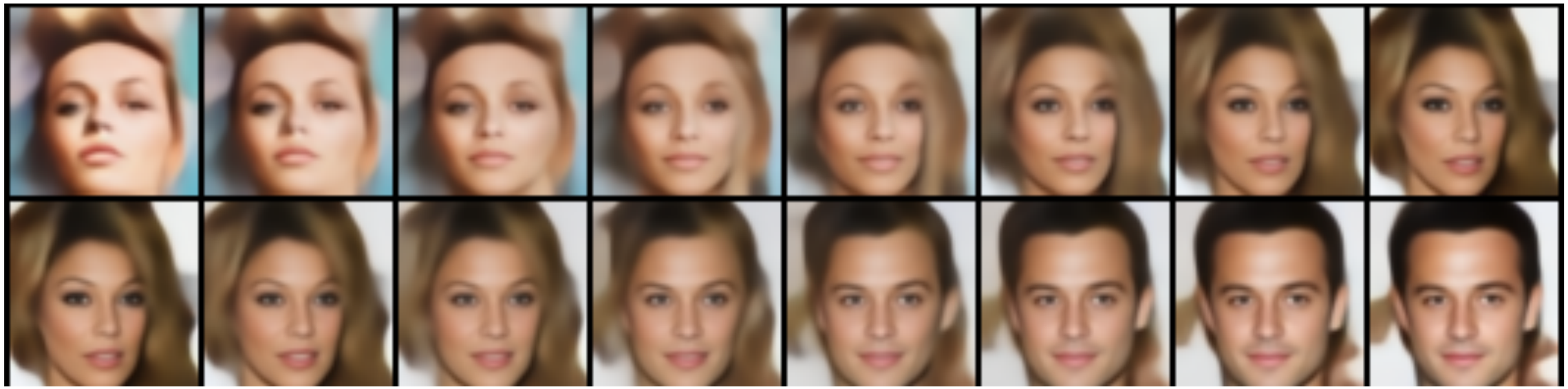
Bedrooms

Generative Interpolations

Tomas Angles Florentin Guth

Celebrities

$$w_1 \longrightarrow w = \alpha w_1 + (1-\alpha) w_2 \longrightarrow w_2$$

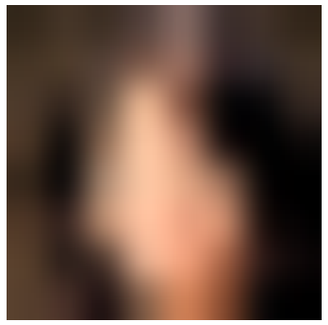


Coarse Grain Model

Tomas Angles

original

averaged



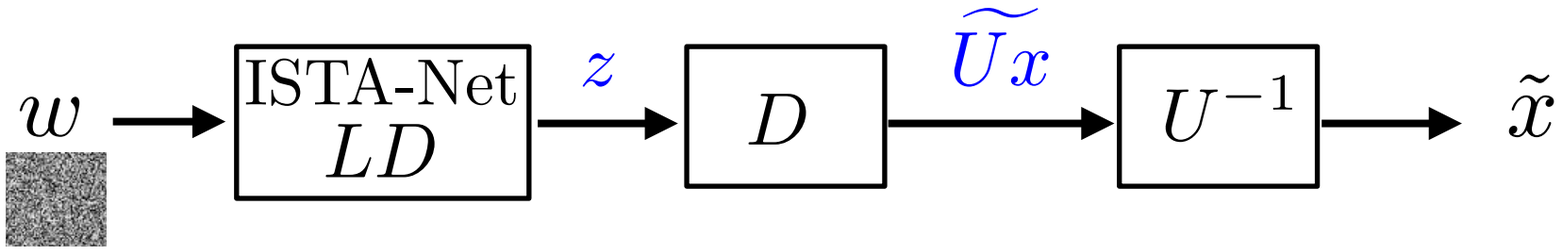
Noise w

Generator

Syntheses with different input noises

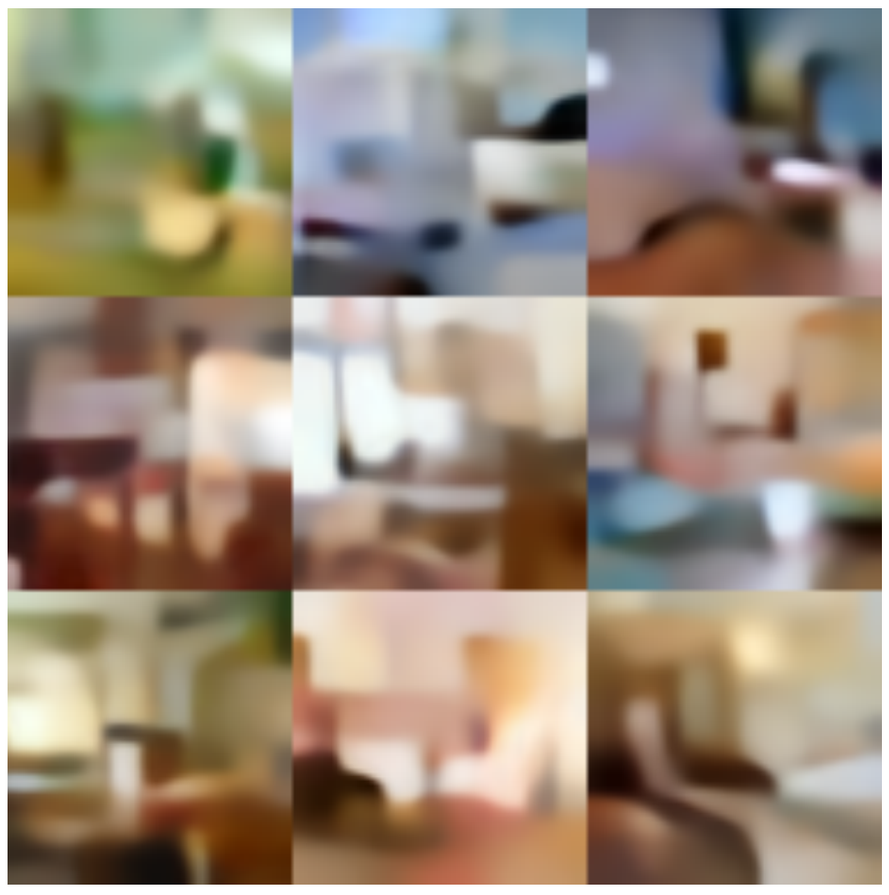
Random Generations from Noise

Tomas Angles Florentin Guth



Celebrities

Bedrooms



Resultats of GAN's

Generated from Hollywood celebrities data basis

Conclusion

- Deep neural network are complex computational machines whose flexibility can be compared with Turing machines.

- A Relu on multiscale filters can produce scale interactions: creates phase harmonics, it may also be used to compute sparse representations, or piecewlinear approximations.

- One can define structured networks which are interpretable: similar to a structured program, with state of the art results.

- Still need functional analysis models and approximation theorems with decay rates.

References

- Understanding Deep Neural Networks: arXiv 1601.04920

- Multiscale Sparse Microcanonical Networks: arXiv 1801.02013

- Generating Networks as Inverse Problems: arXiv 1805.06621

- Deep Network Classification by Scattering and Dictionary Learning: arXiv

- Wavelet Phase Harmonic Covariance Models of Stationary Processes: arXiv