Colloquium du CERMICS

Mathematical Mysteries of Deep Neural Networks

Stéphane Mallat (Collège de France et École Normale Supérieure)

20 septembre 2019

Mathematical Mysteries of Deep Neural Networks

Tomas Anglès, Roberto Leonarduzi,
Stéphane Mallat, Louis Thiry,
John Zarka, Sixin Zhang

Collège de France
École Normale Supérieure

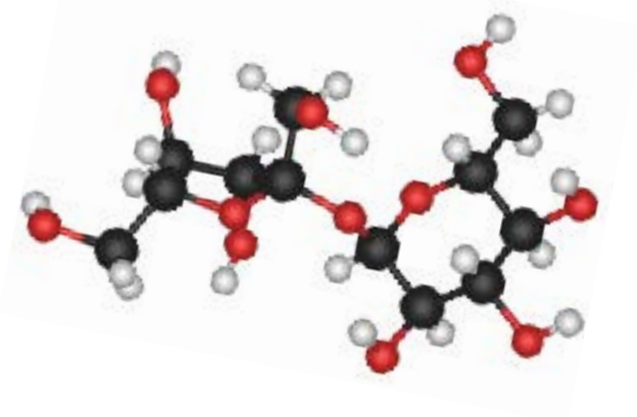
High-Dimensional Approximations

What regularity properties lead to low-dimensional approximations of $f(x)$ for a high-dimensional $x \in \mathbb{R}^d$ in physics and machine learning ?

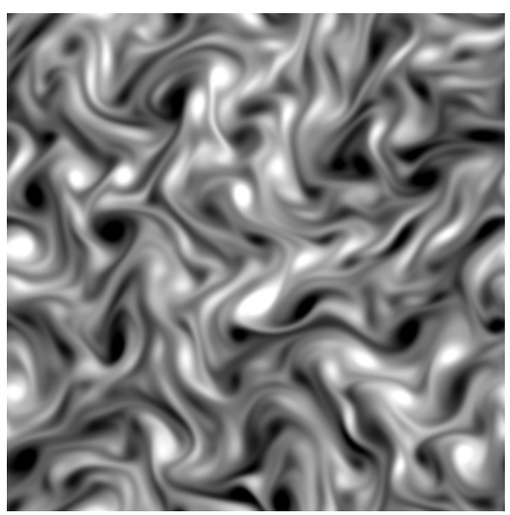
- $f(x)$: class of an image x having $d = 10^6$ pixels

or

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

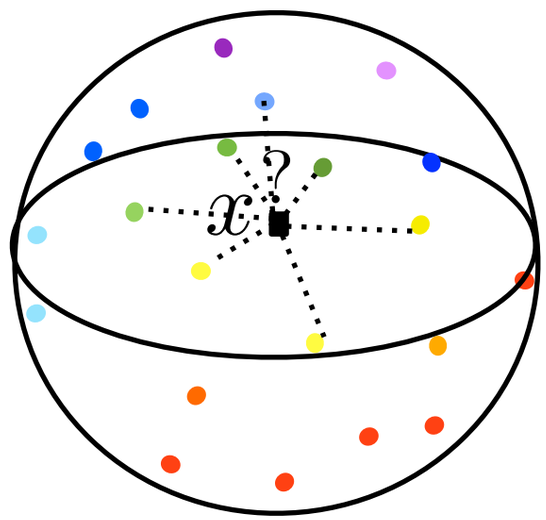


- $f(x) = p(x)$ a probability density.



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:



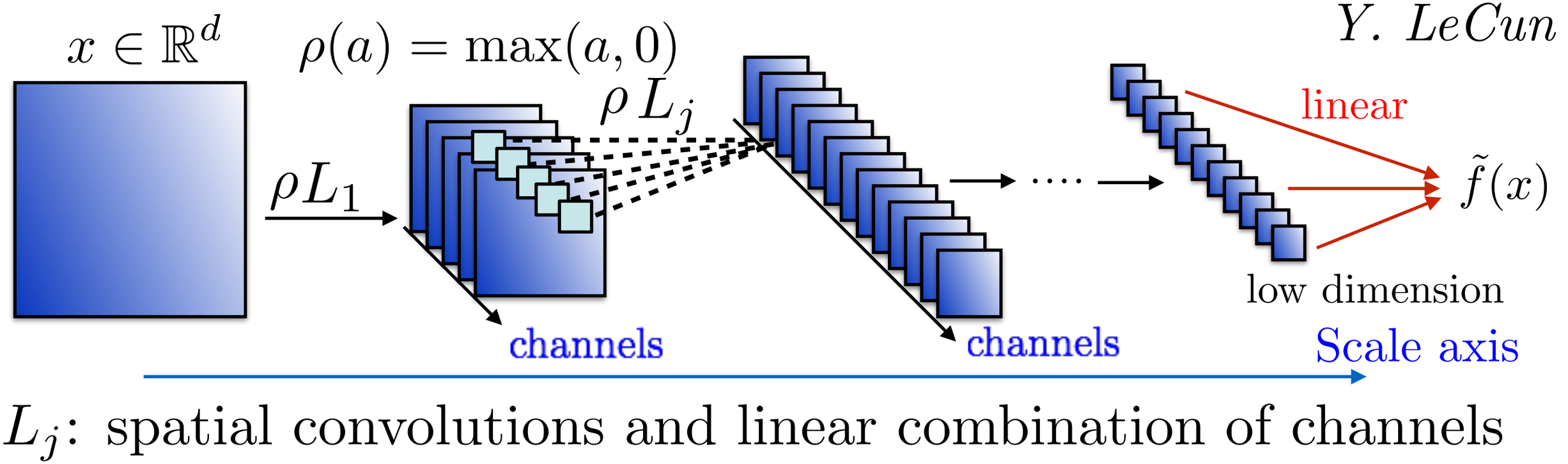
- Need $n \geq \epsilon^{-d}$ points to cover $[0,1]^d$ at a Euclidean distance ϵ

  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 ?

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 interpretable.

To create simpler interpretable networks:

What underlying regularity is captured and how ?

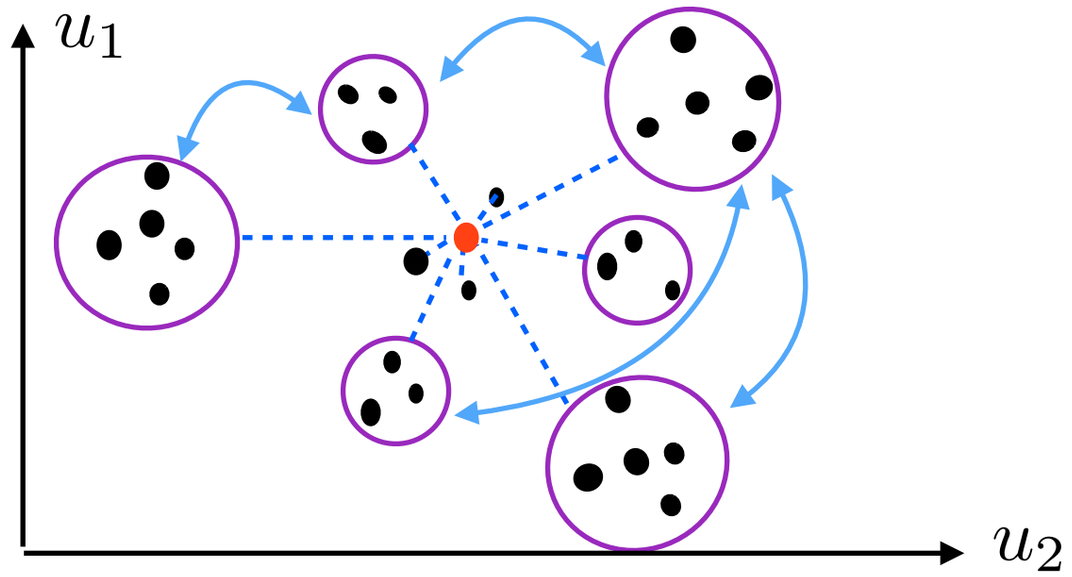
3 ingredients: Multiscale, Linearize group actions, Sparse

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

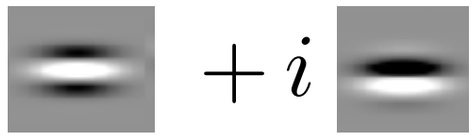
Overview: Simpler Networks

- Scale separation with wavelets and interactions through phase

- Linear scale interaction models and invariants in:
  - Statistical physics for turbulence
  - Quantum chemistry and image classification

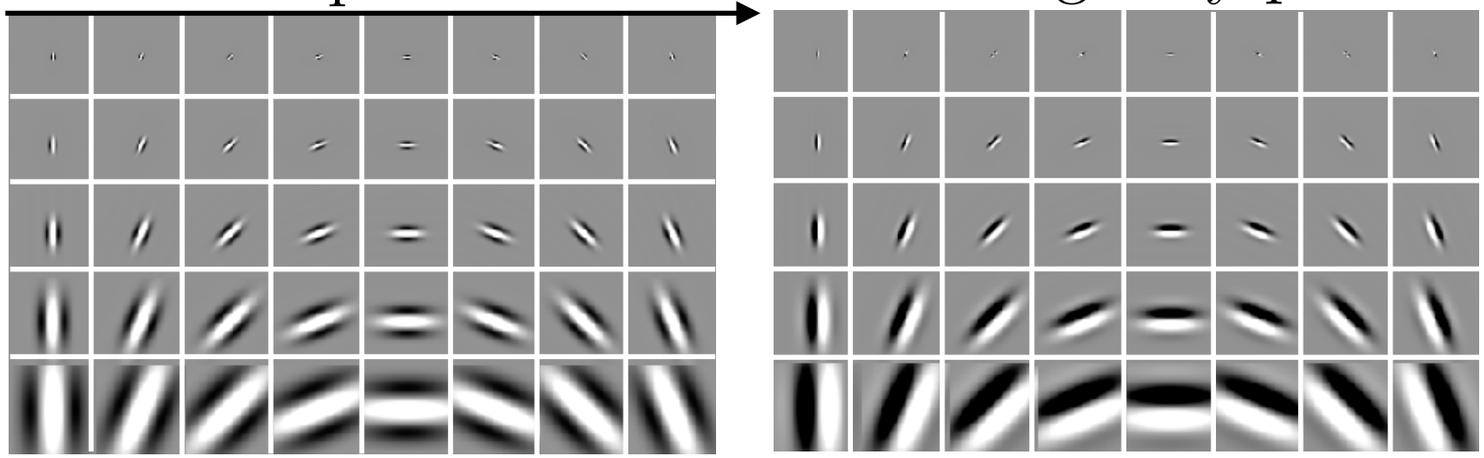
- Non-linear scale interactions models with sparse dictionaries in:
  - Classification of complex structures as in ImageNet
  - Generation of non-ergodic processes

Scale separation with Wavelets

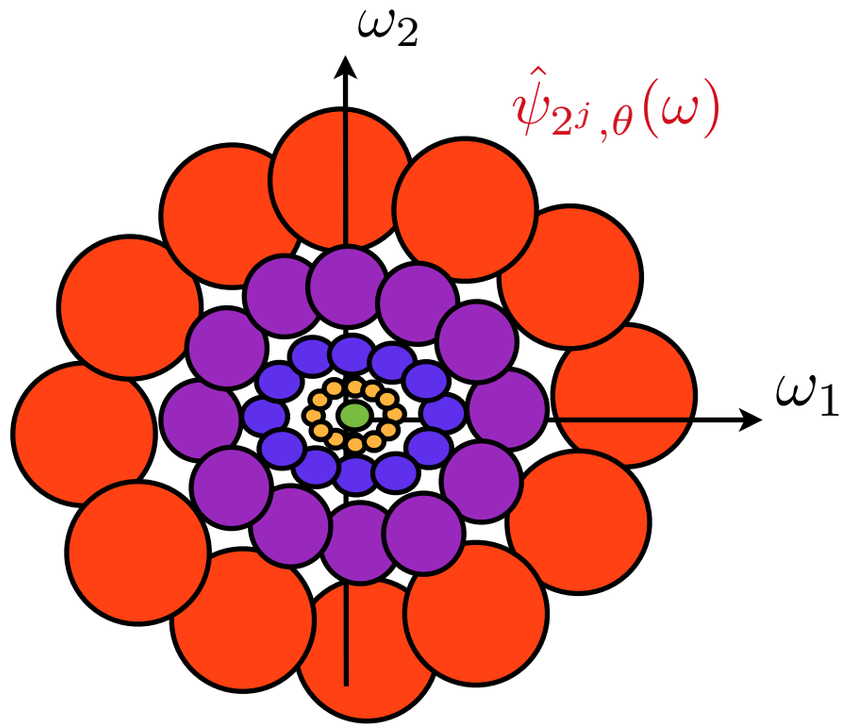
- Wavelet filter $\psi(u)$:  $+ i$

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

real parts θ imaginary parts



Fourier



- Wavelet transform: invertible

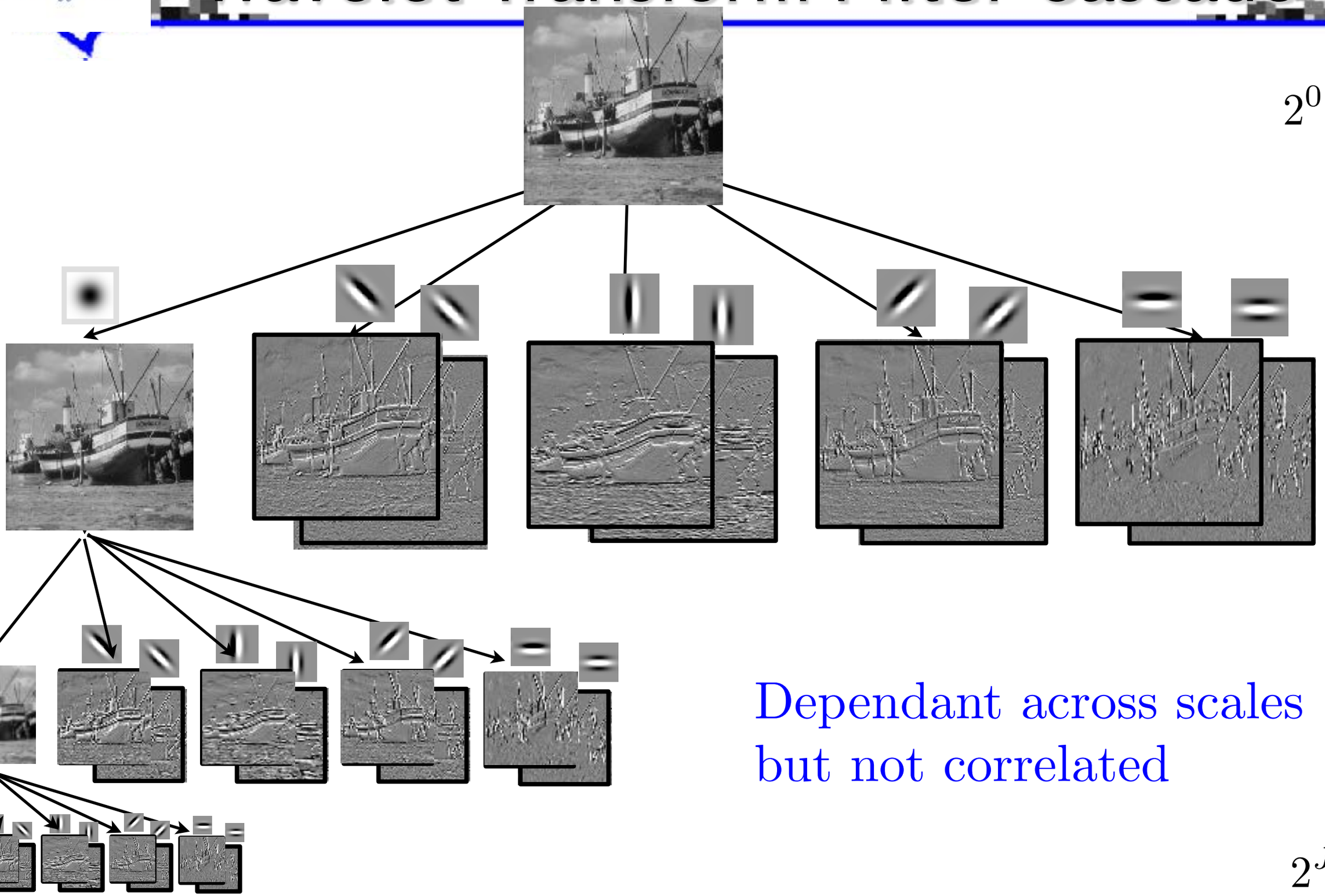
$$Wx = \Big(x \star \psi_\lambda\Big)_\lambda$$

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

- 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

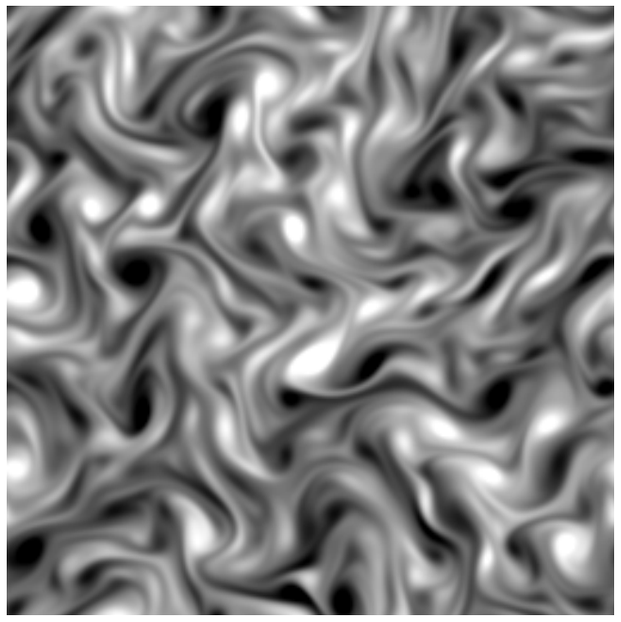


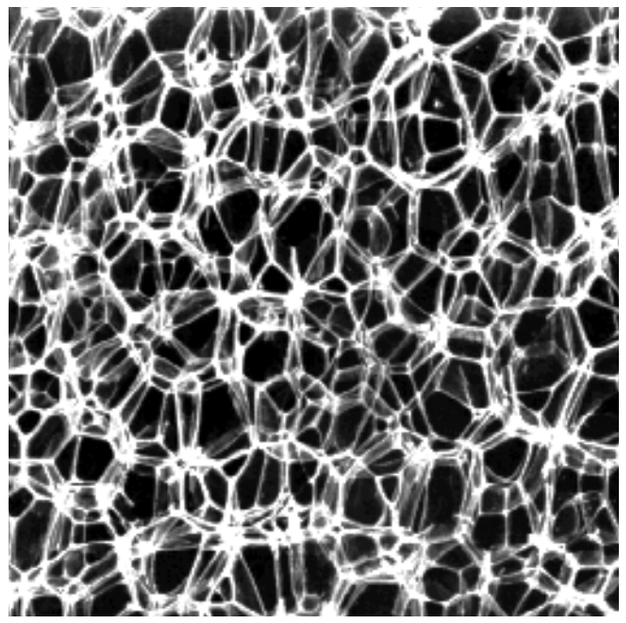
Dependant across scales
but not correlated

Stat. Physics of Stationary Proc.

What stochastic models for turbulence ?

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





Prior: stationary $\Leftrightarrow$ $p(x)$ is invariant to translations.

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

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

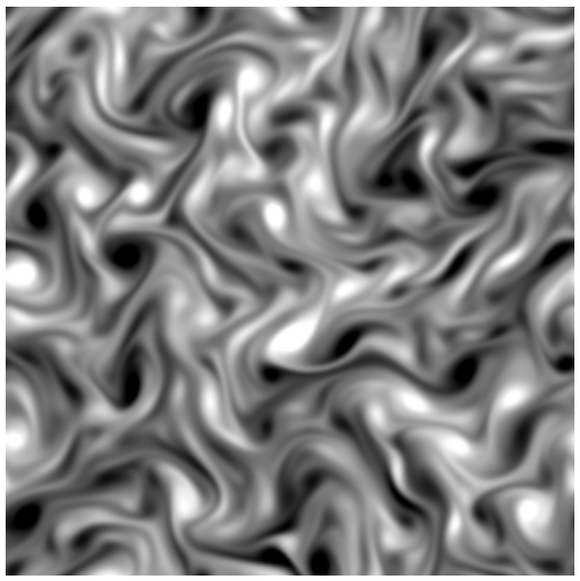
With $M = d$ second order moments:

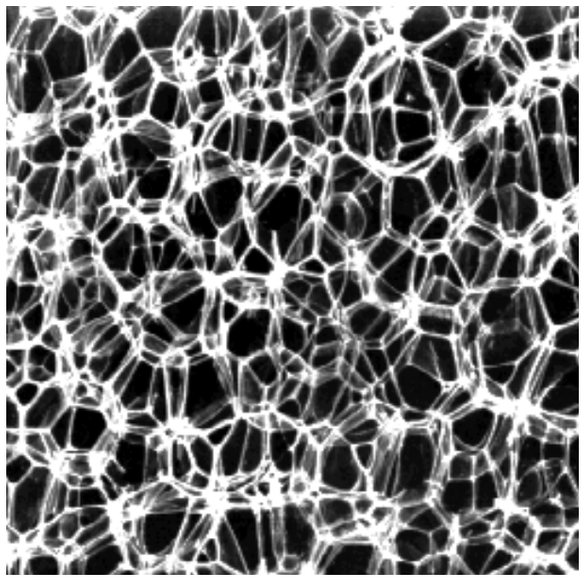
$\phi_m(x) = \sum_u x(u)x(u-m) \quad \Rightarrow$ $\tilde{p}(x)$ is a Gaussian distribution

Gaussian Models of Stationary Proc.

What stochastic models for turbulence ?

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

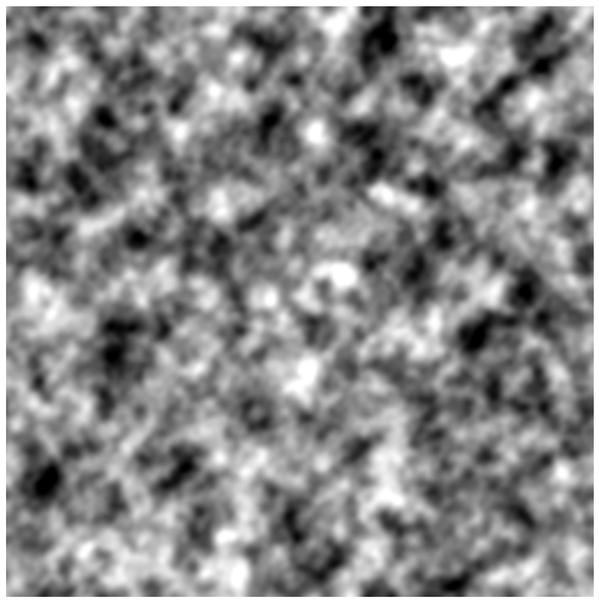


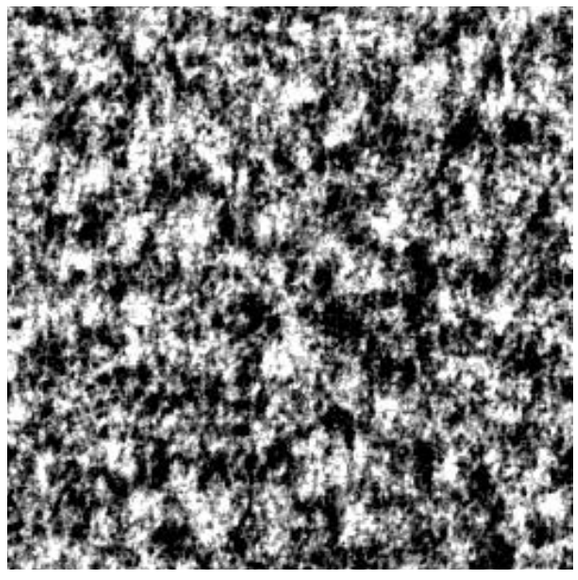


$\tilde{p}(x)$ is a Gaussian distribution

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

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

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

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

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

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:

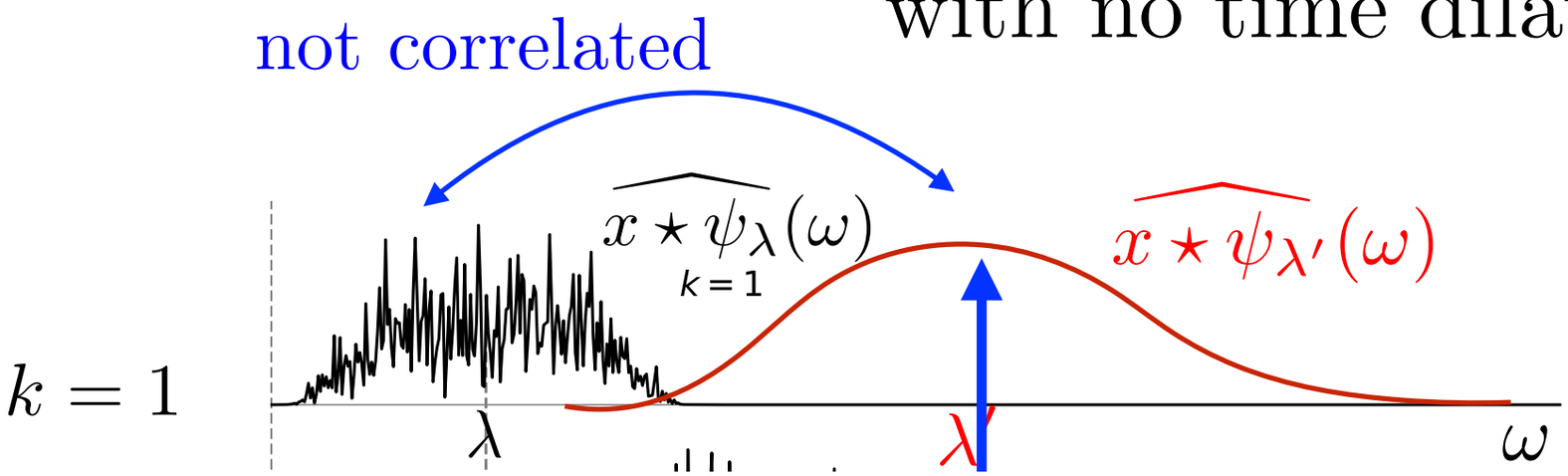
$$\widehat{U}x(u,k,\lambda) = \hat{\gamma}(k)\,|x \star \psi_\lambda(u)|\, e^{ik\,\varphi(x\star\psi_\lambda(u))}$$

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

Frequency Transpositions

Phase harmonics: $|x \star \psi_\lambda(u)|\, e^{i\,k\,\varphi(x \star \psi_\lambda(u))}$

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



Phase Harmonics

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

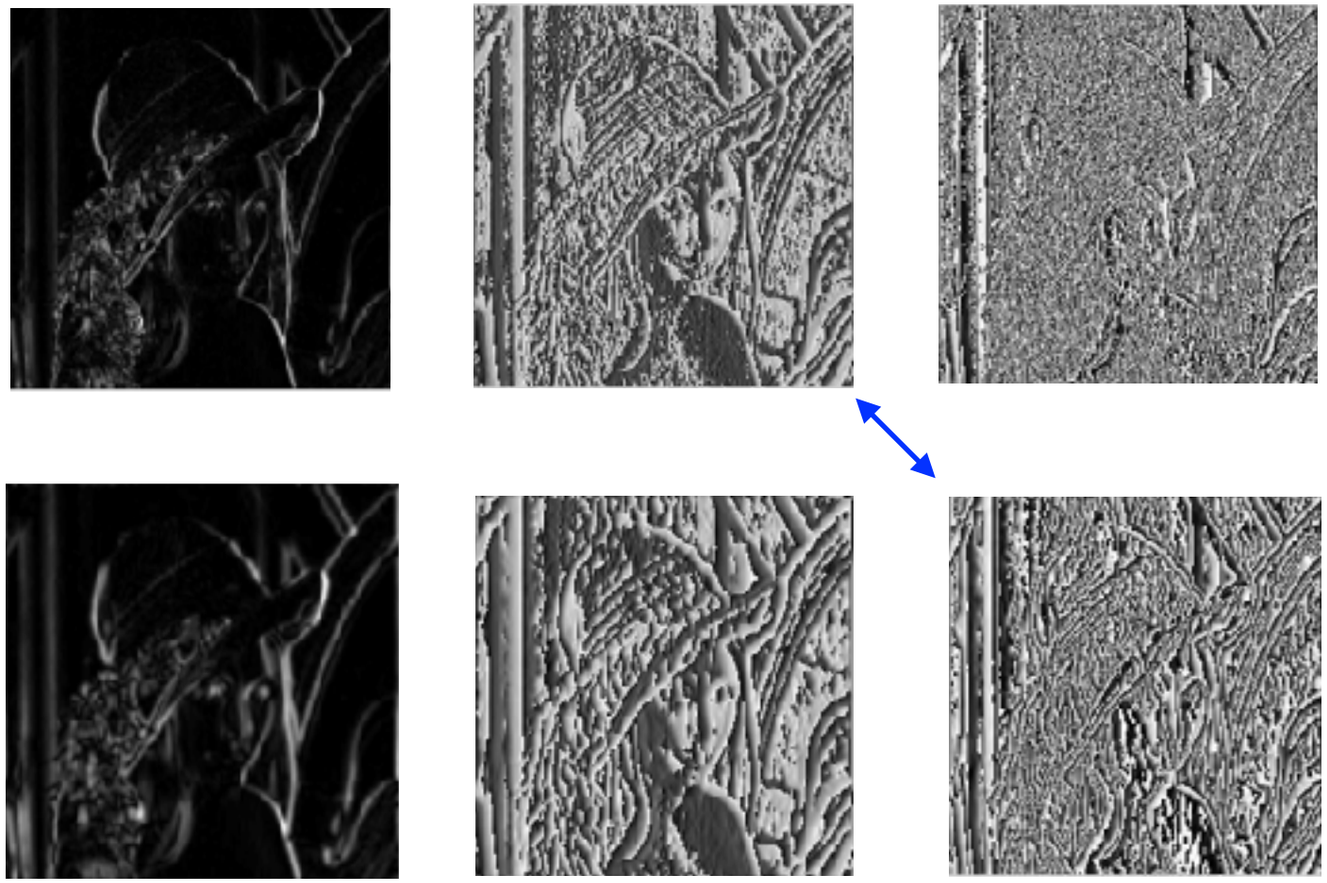
Rectified Wavelet Coefficients

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

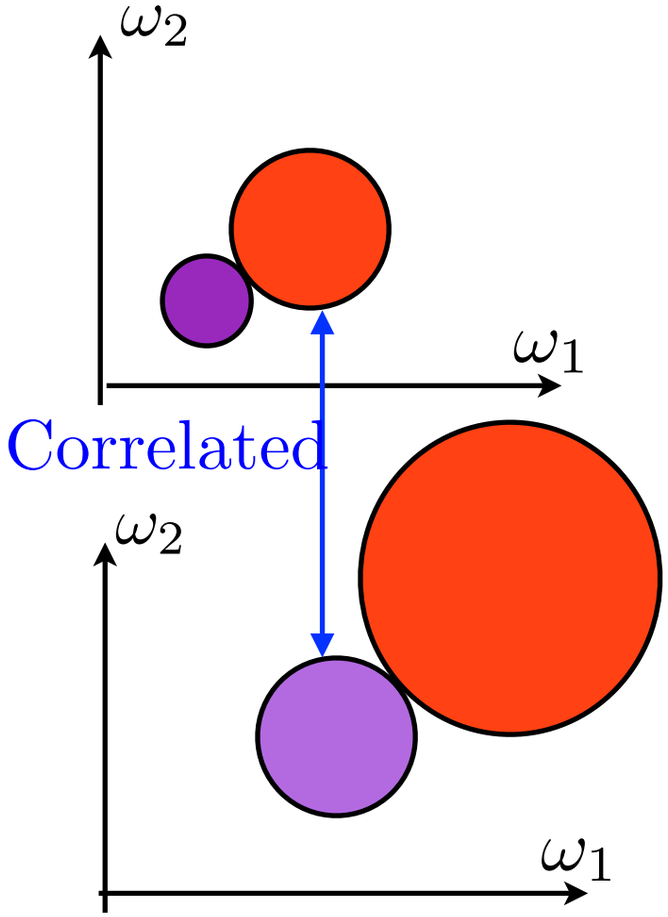
$$\rho(x \star \psi_{\alpha,\lambda}) \xrightarrow[\text{along } \alpha]{\text{Fourier transform}} c_k\,|x \star \psi_\lambda|\,e^{ik\,\varphi(x\star\psi_\lambda)}$$

Harmonics

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



$k = 2$

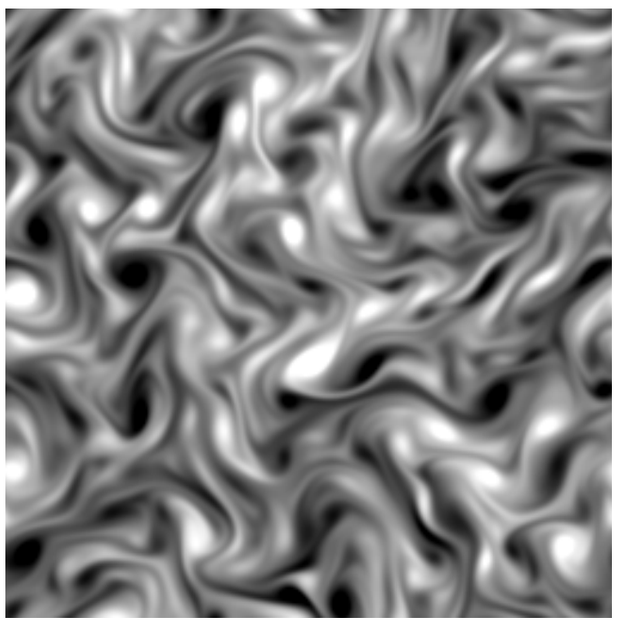


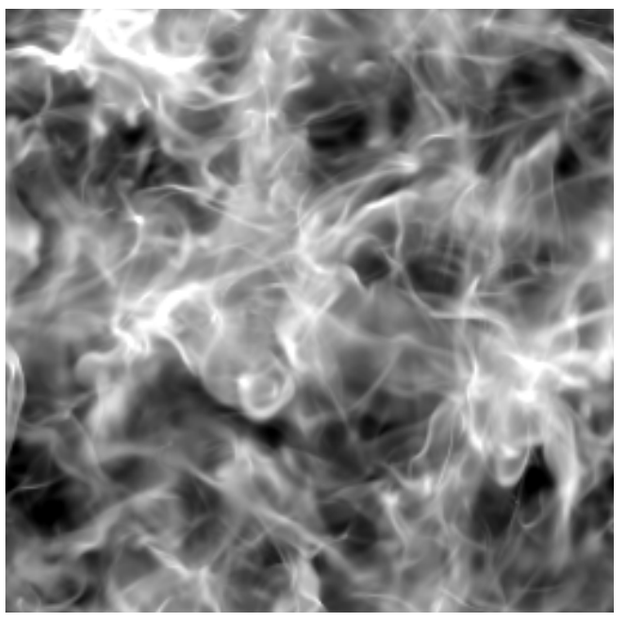
Phase harmonics:
Frequency transpositions

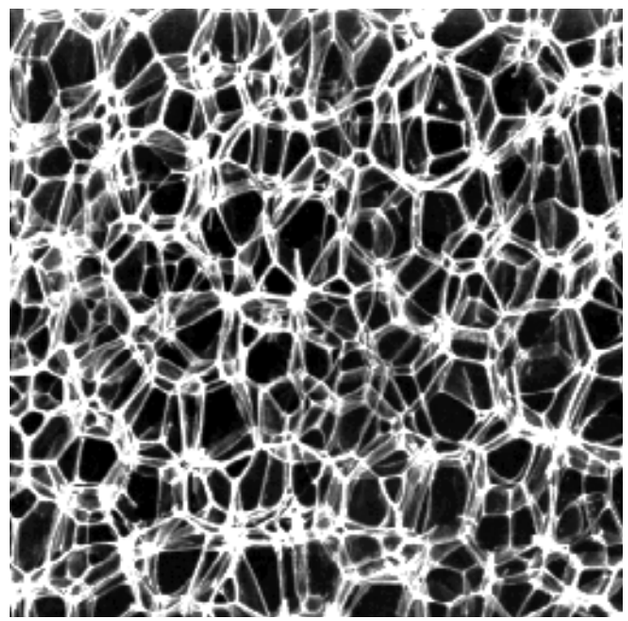
Models of Stationary Processes

Sixin Zhang

x







Maximum entropy distribution conditioned by

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

$$\phi_m(x) = \sum_u |x \star \psi_\lambda(u)|\, e^{ik\varphi(x\star\psi_\lambda(u))}\, |x \star \psi_{\lambda'}(u)|\, e^{-ik'\varphi(x\star\psi_{\lambda'}(u))}$$

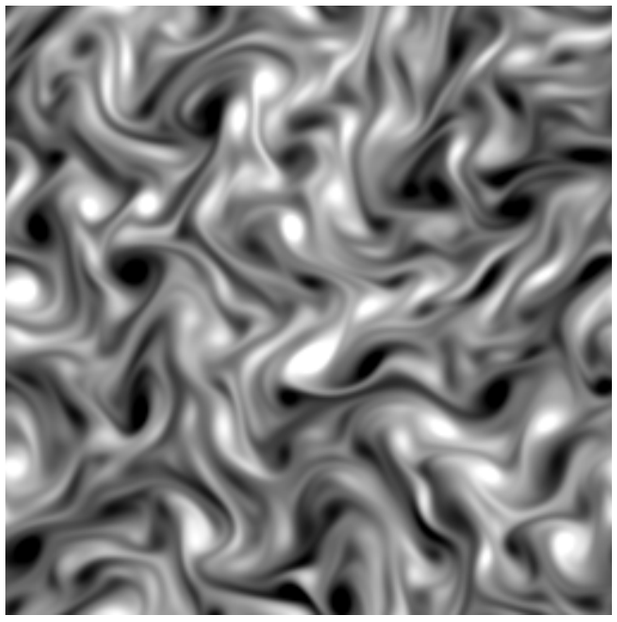
$$\tilde{p}(x) = \mathcal{Z}^{-1} \exp\Big(- \sum_{m=1}^{M} \beta_m\, \phi_m(x) \Big)$$

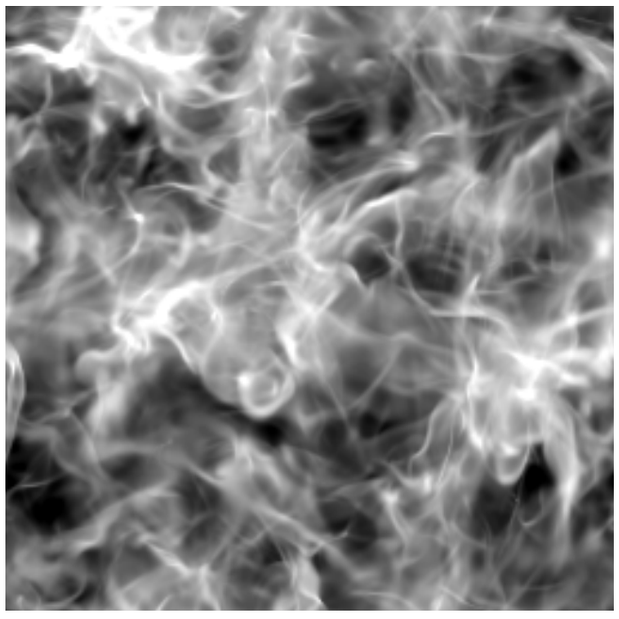
Ergodic Stationary Processes

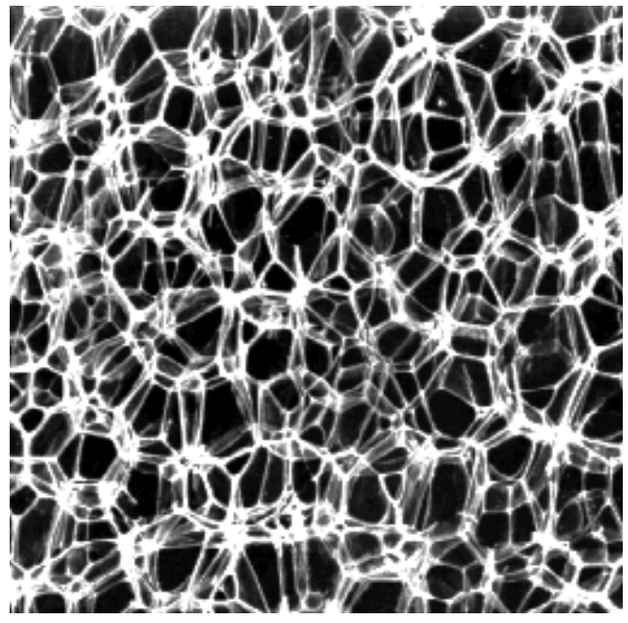
S. Zhang, J. Bruna, E. Allys, F. Levrier, F. Boulanger

$d = 6\,10^4$

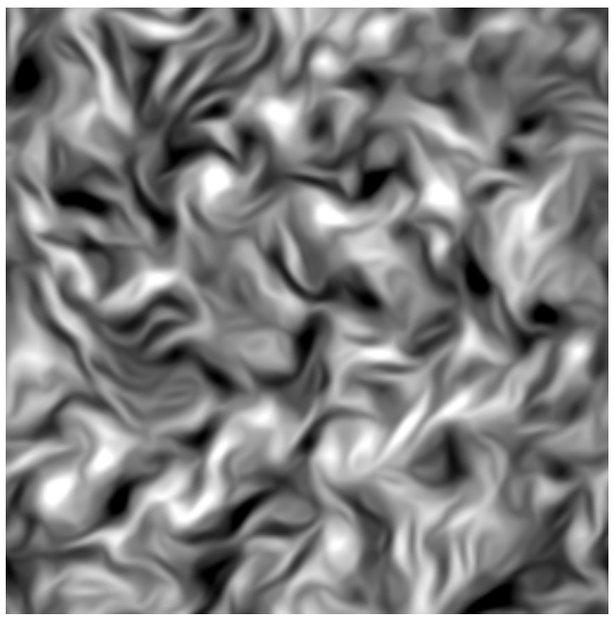
x

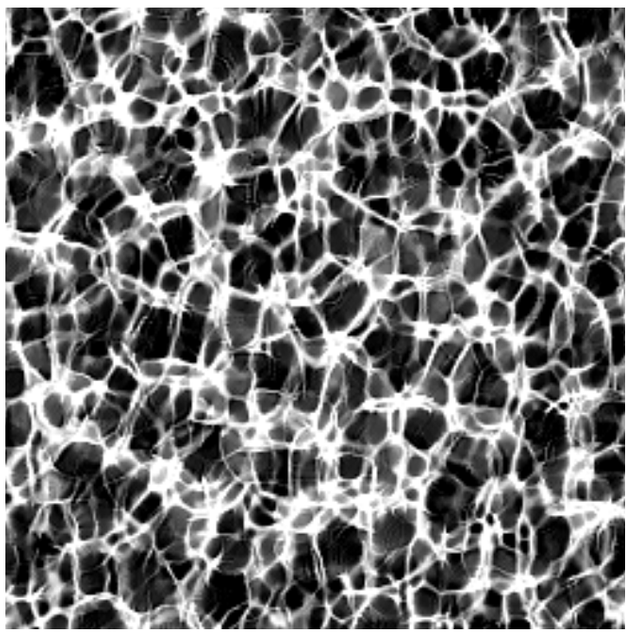






$\tilde{x}$





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

Phase coherence is restored

How much physics are these models capturing ?

What about non-ergodic processes ?

Scattering Wavelet Coefficients

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:

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

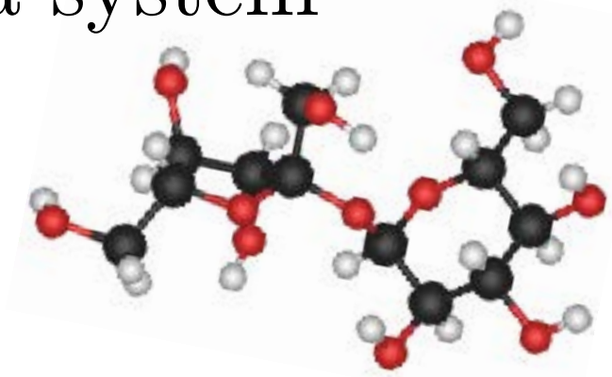
- Linearize small deformations

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 $x = \Big\{\text{positions}, \text{charges}\Big\}$?



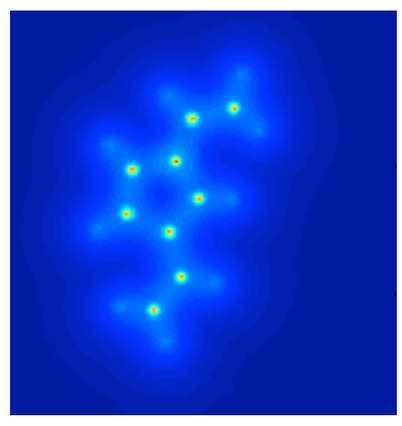
Symmetries:

$f(x)$ 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



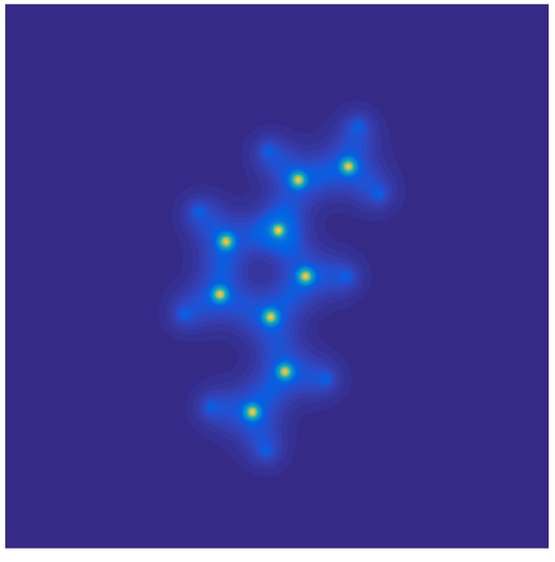
Dirac Electronic Density

- 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:

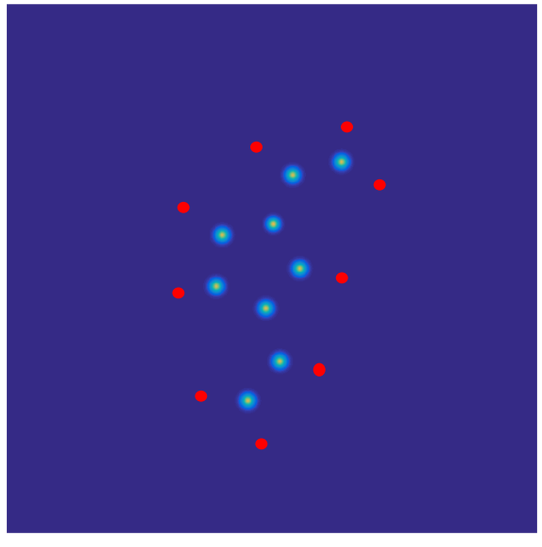
$$x(u) = \sum_{k=1}^{d} z_k\,\delta(u - r_k)$$

Electronic density



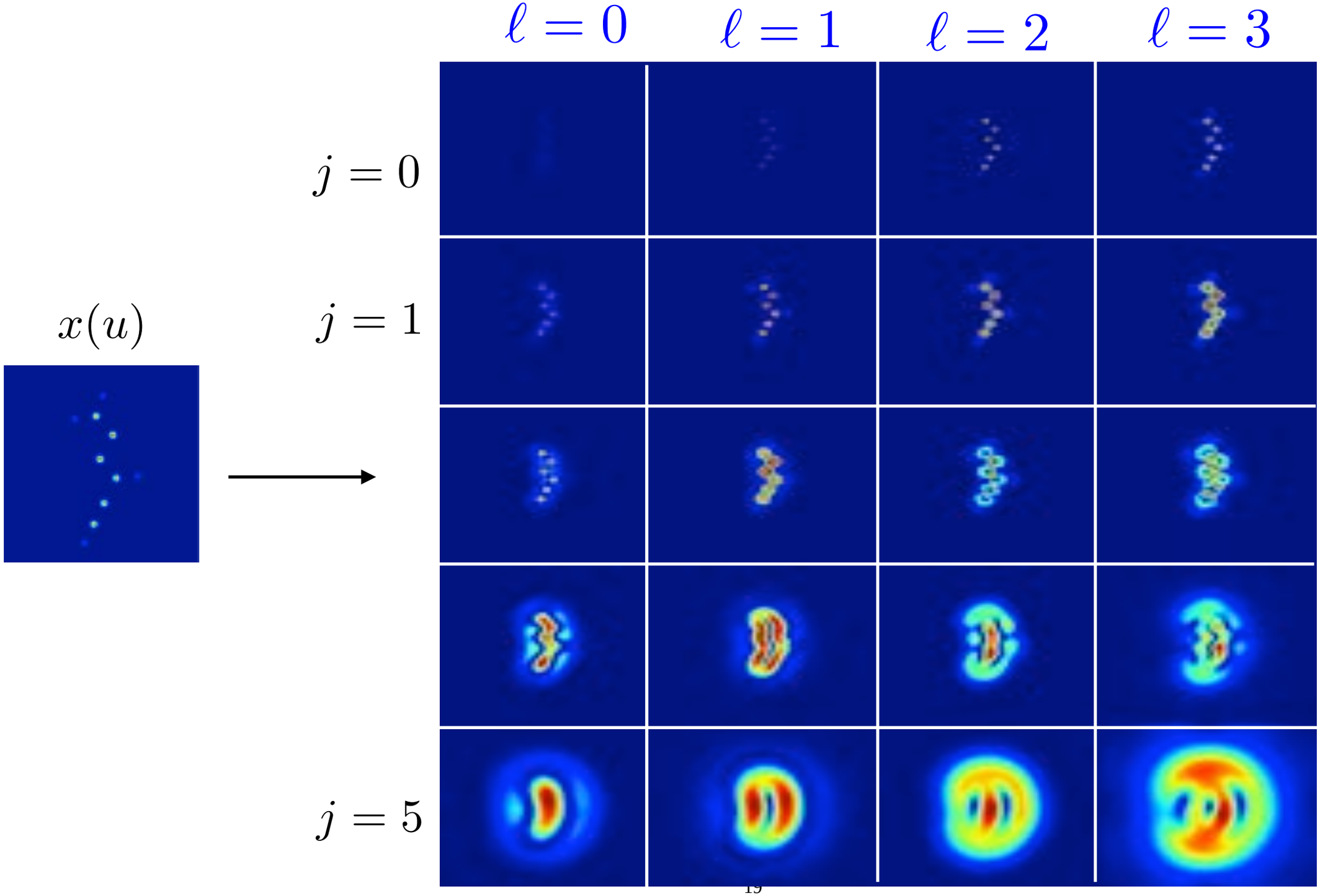
Dirac density

$x(u)$



Harmonic Wavelet Interferences

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



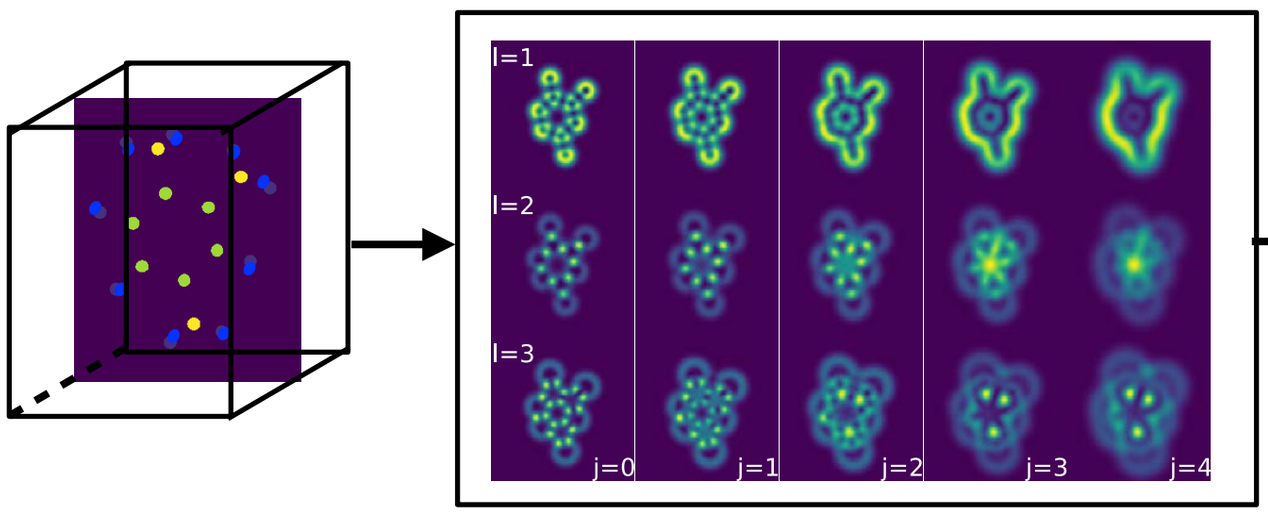
Scattering Energy Regression

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

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

Invariants to
Translations
Rotations

x



convolutional neural network without learning

Linear
Averaging

Sx

$O(\log^2 d)$

Linear
Regression

$\tilde{f}(x)$

supervised
learning

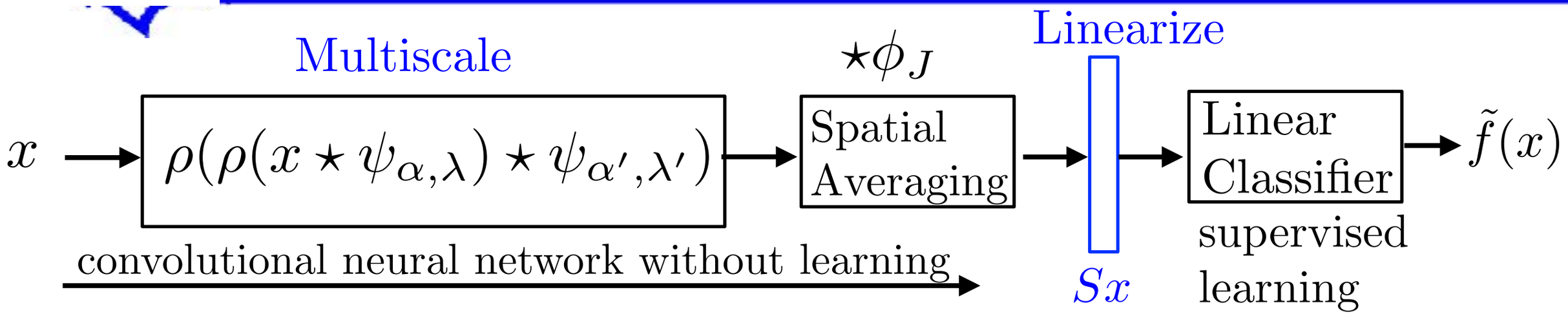
QM9: Data basis of 130.000 organic molecules with C, H, O, N, F with DFT atomisation energies

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

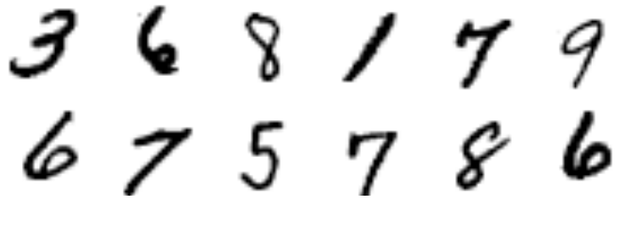
But small molecules with at most 29 atoms and 9 heavy ones

Image Classification

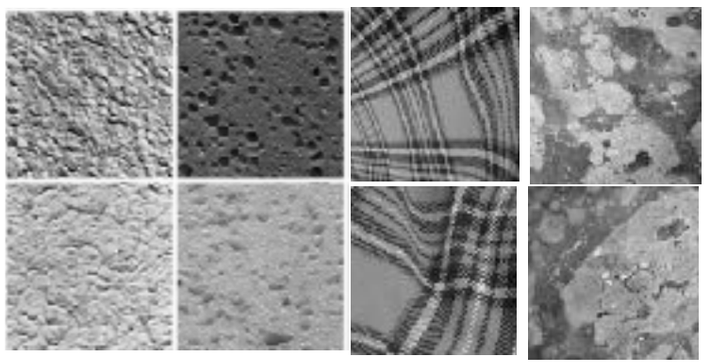
J. Bruna



Digits
10 classes



Textures
60 classes



ImageNet
10^3 classes

Errors: Scattering | Deep Nets.

0.5 % | 0.5 %

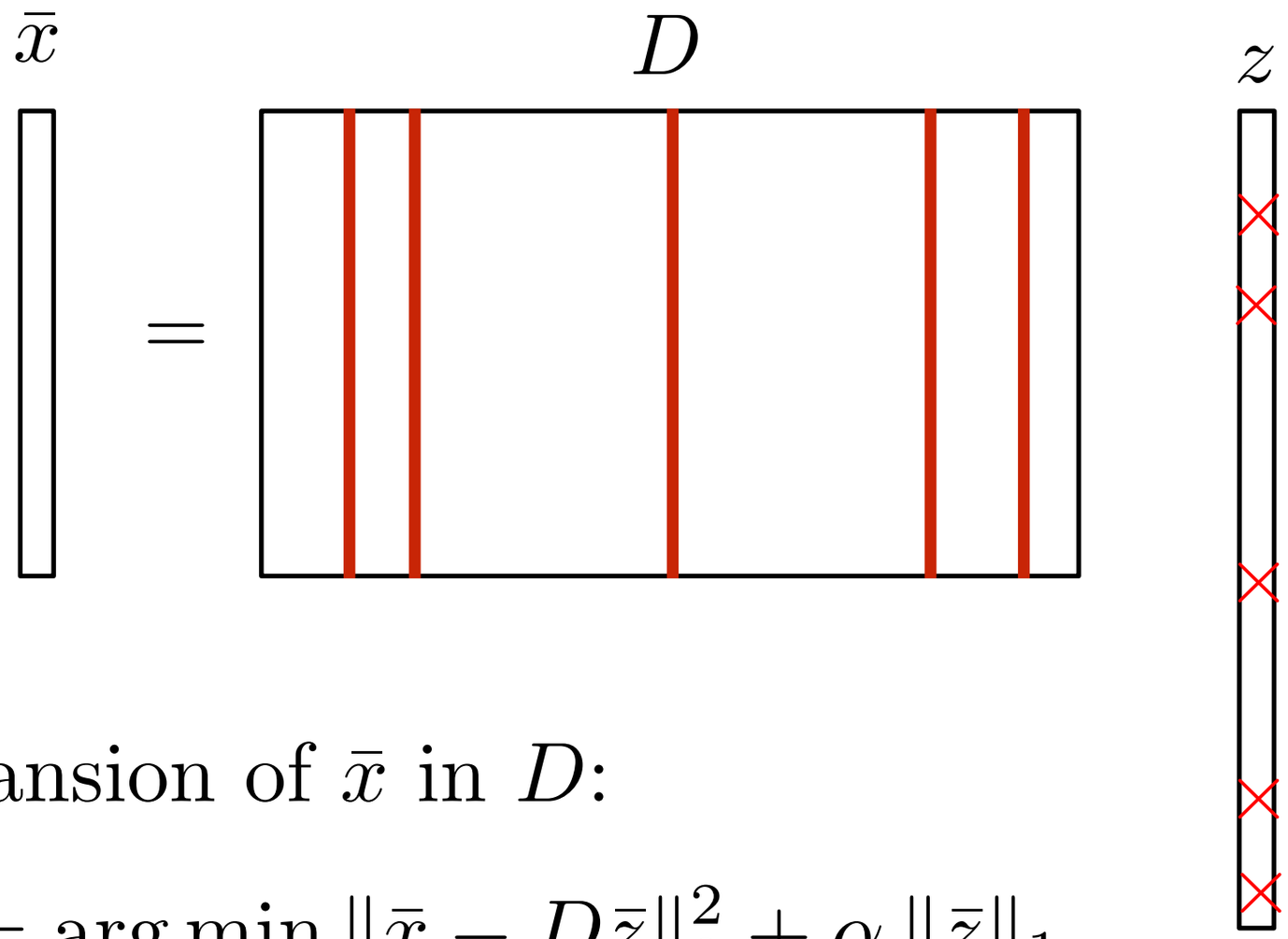
60% | AlexNet 20% : 2012

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:

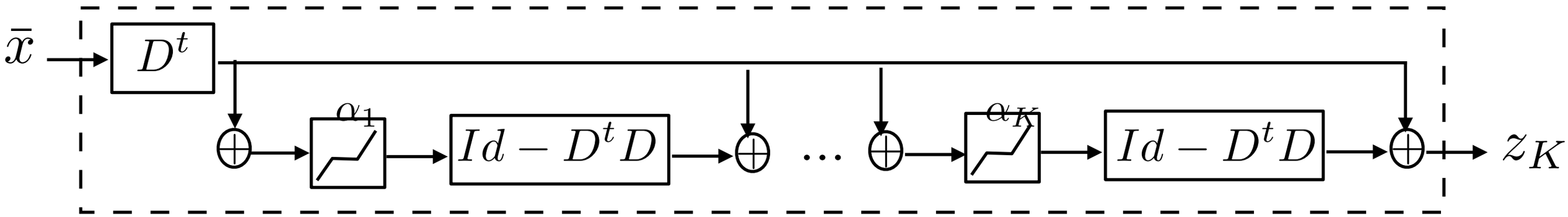
$$z = \arg\min_{\bar{z}} \|\bar{x} - D\bar{z}\|_2^2 + \alpha_k \|z\|_1$$

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:

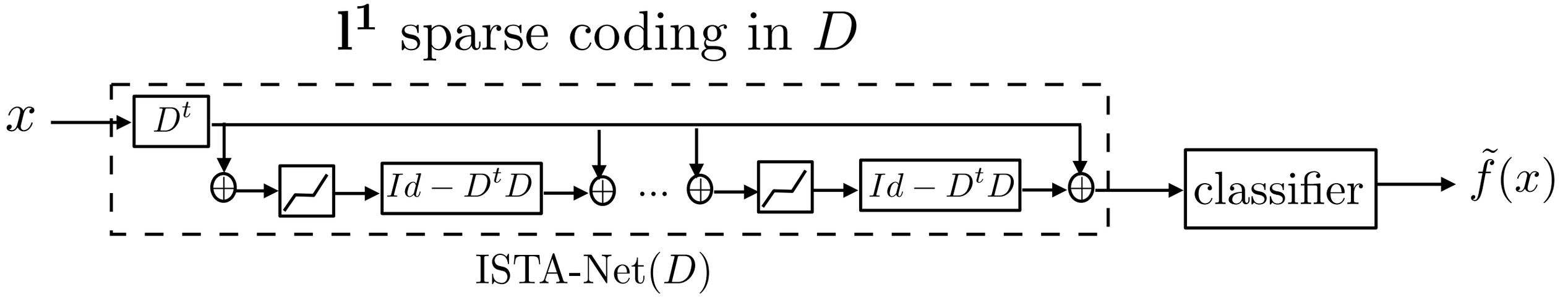


ISTA-Net(D)

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

$x \rightarrow \boxed{\rho(\rho(x \star \psi_{\alpha,\lambda}) \star \psi_{\alpha',\lambda'})} \rightarrow \boxed{\text{Spatial Averaging}} \xrightarrow{Sx} \boxed{\text{Classifier}} \rightarrow \tilde{f}(x)$ 60% error

Learned sparse coding

$x \rightarrow \boxed{\text{ISTA-Net } D} \xrightarrow{z} \boxed{\text{Classifier}} \rightarrow \tilde{f}(x)$ 50% error

18% error

Multiscale Scattering Linearize Learned sparse coding

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

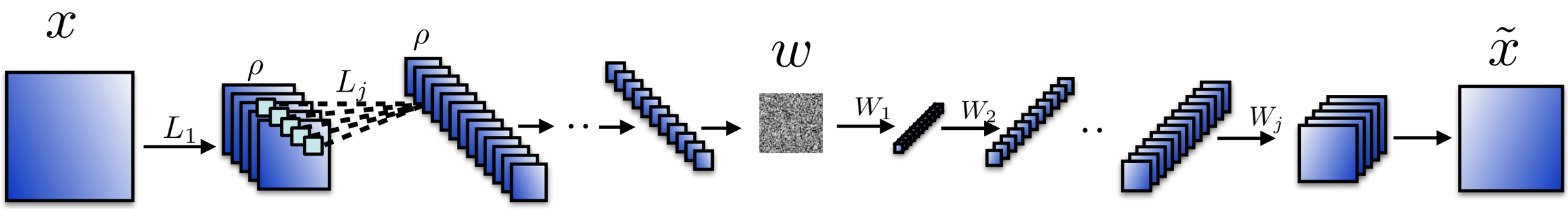
Interpretable network : patterns stored in D

AlexNet: 20% error

Why such an error reduction ? Linearize classes in separate spaces

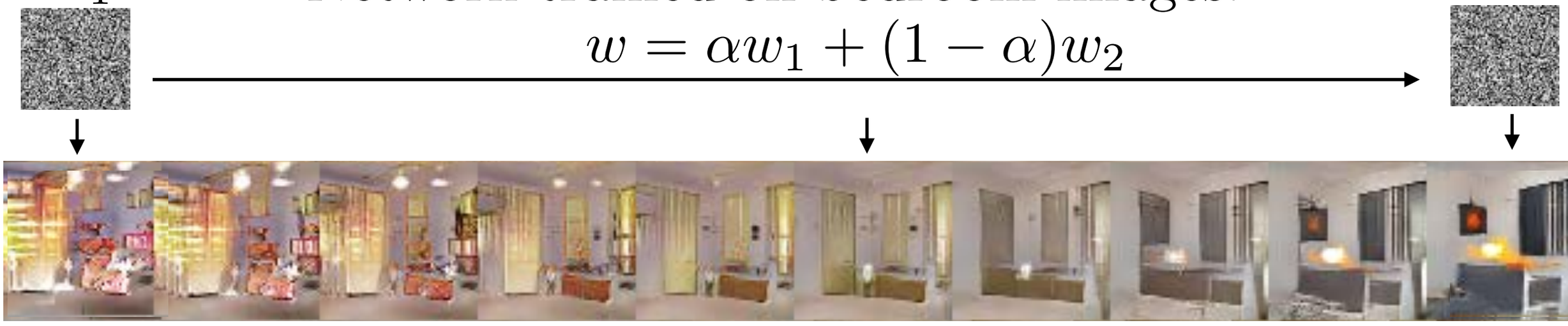
Non Ergodic Processes

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

Encoder

Multiscale

Nearly Gaussian

Low-dimensional Gauss. white noise

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

L : linear mixing central limit theorem

Inversion:

$w = LUx$  $\rightarrow$ **Generator** $\rightarrow$ 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

Inversion:

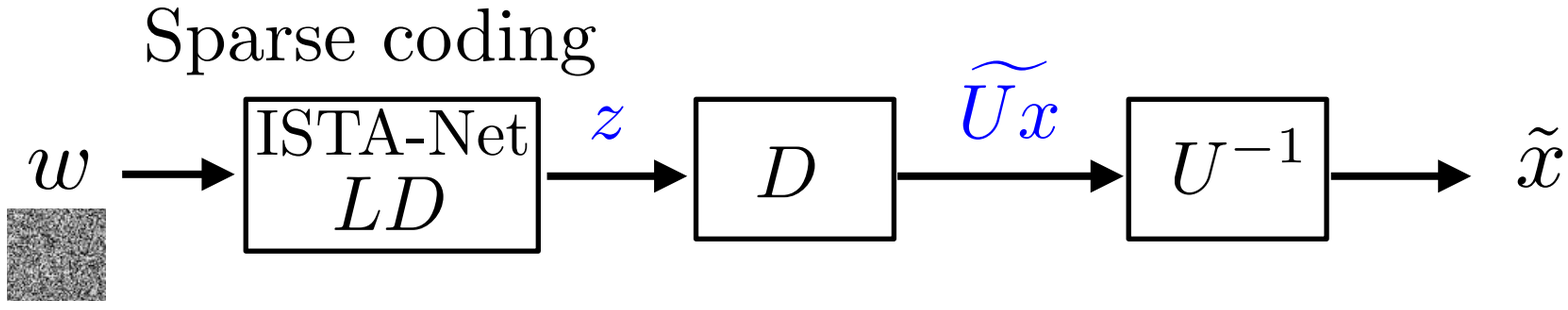
$$w = L\,U x$$

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 $\{x_i\}_i$

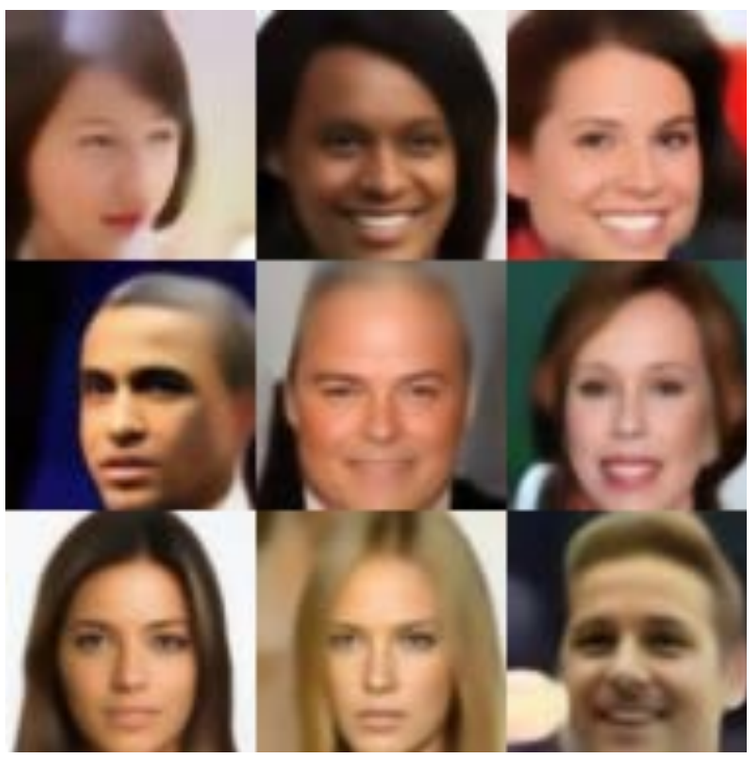
Training Reconstruction

Celebrities Data Basis

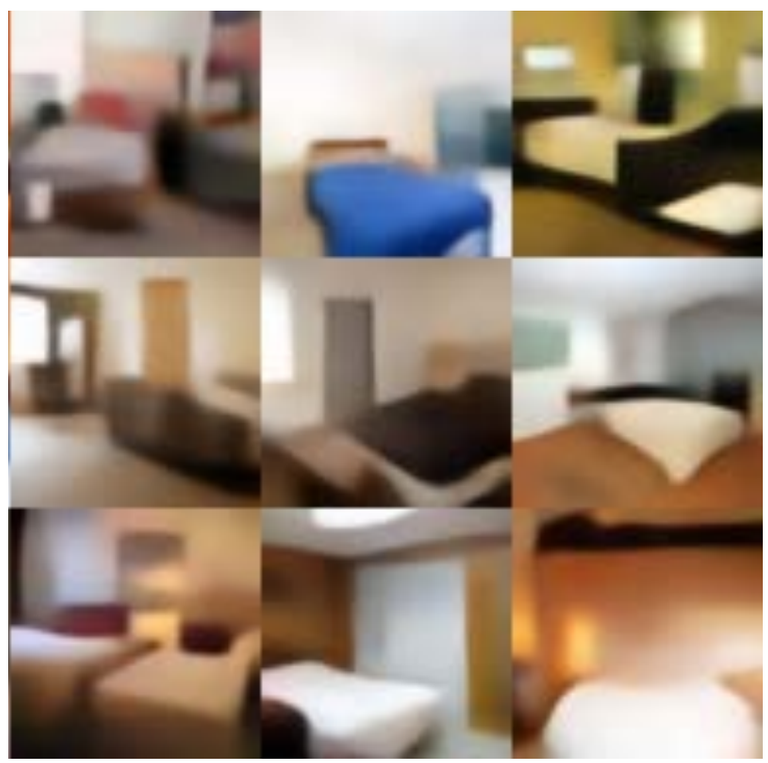
Tomas Angles

Bedrooms

x_i



$\tilde{x}_i$



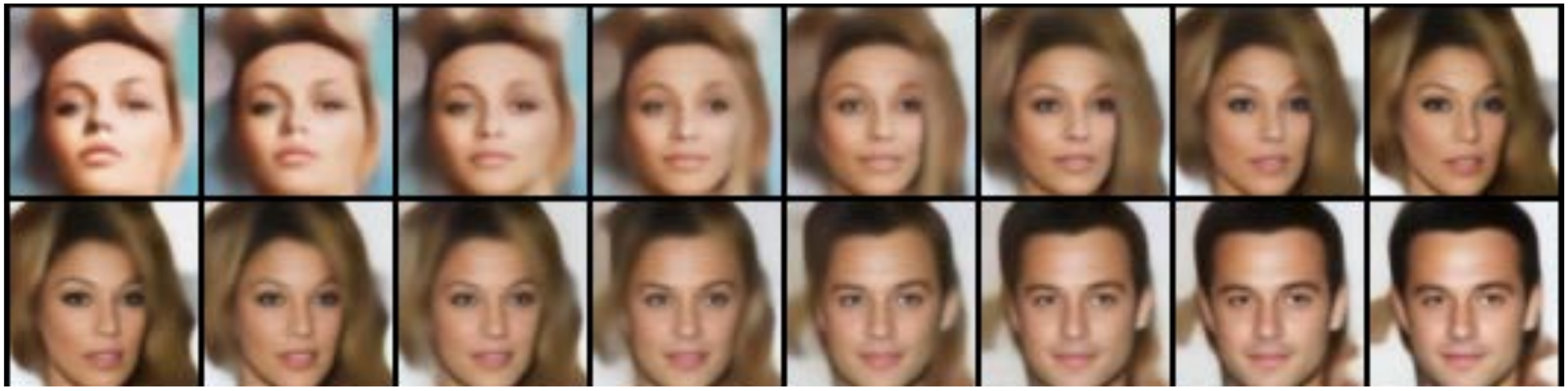
Generative Interpolations

Tomás Angles

Celebrities

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

w_1 → w_2

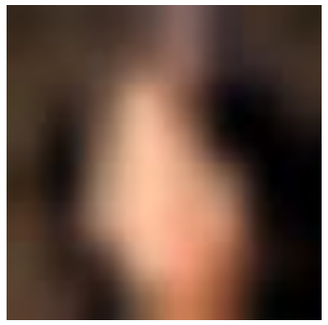


Coarse Grain Model

Tomas Angles

original

averaged



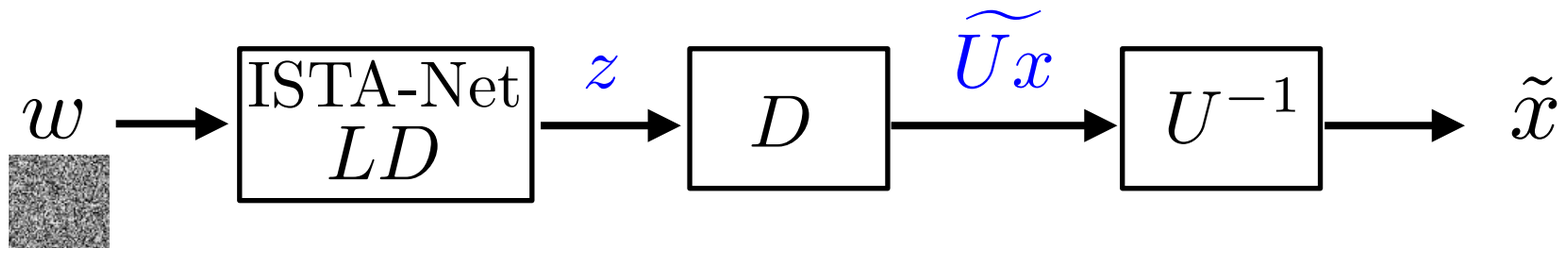
Noise w

Generator

Syntheses with different input noises

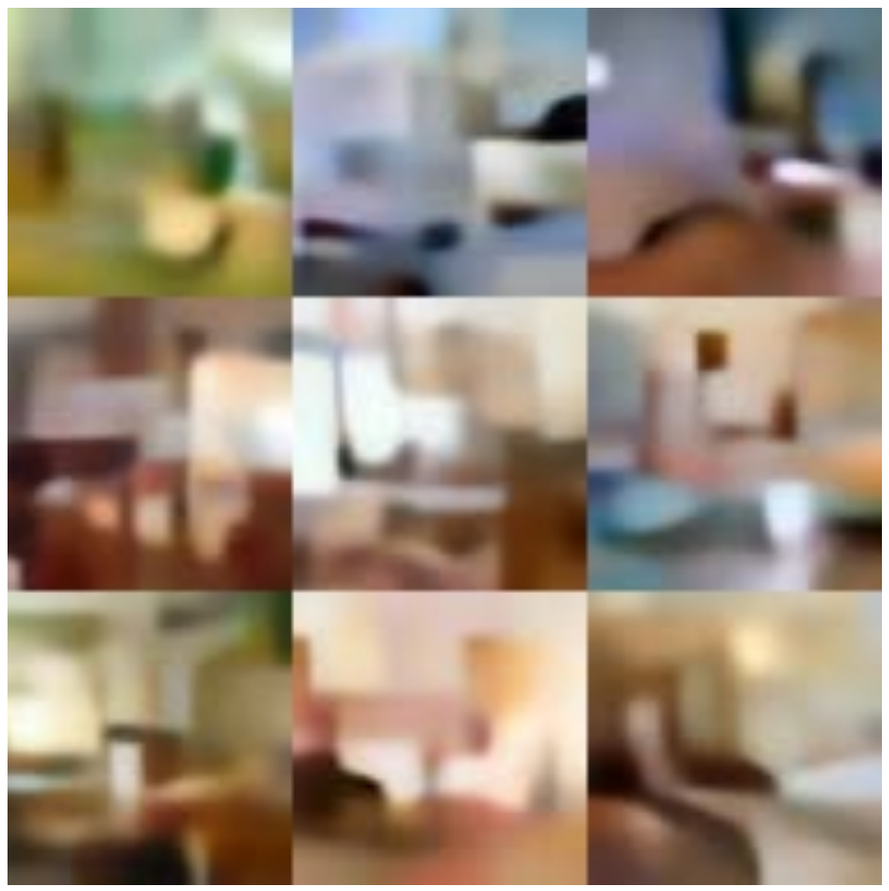
Random Generations from Noise

Tomás Angles



Celebrities

Bedrooms



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.