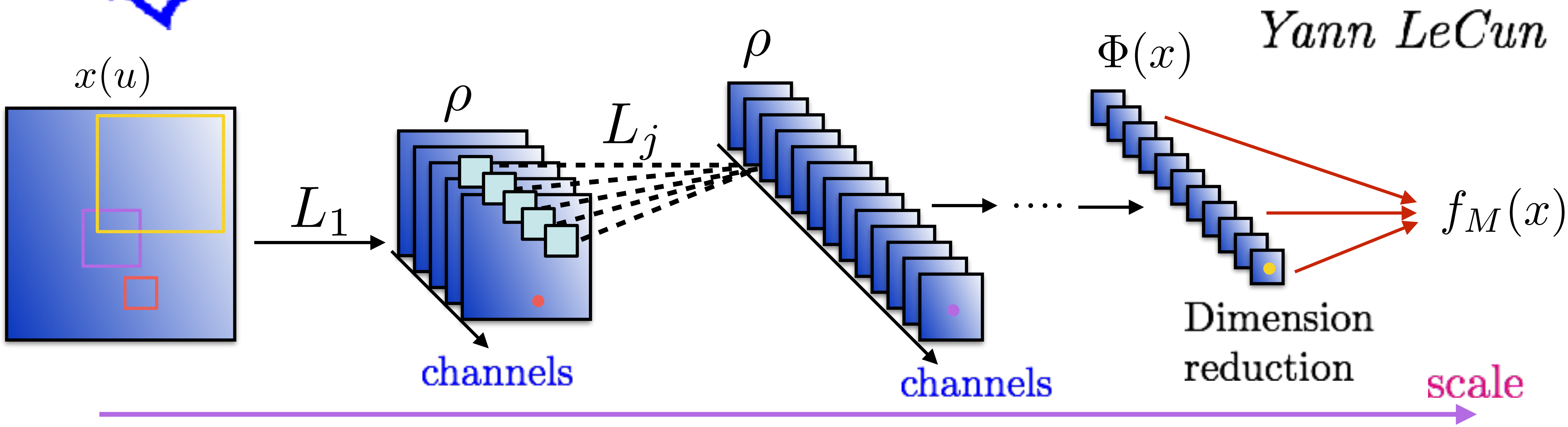
Unsupervised Learning and Inverse Problems with Deep Neural Networks

Joan Bruna, Stéphane Mallat,
Ivan Dokmanic, Martin de Hoop

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

Deep Convolutional Networks



L_j is a sum of spatial convolutions across channels, subsampling

$\rho(u)$ is a scalar non-linearity: $\max(u, 0)$ or $|u|$ or ...

Part I Architecture Simplification: wavelet scattering

Part II Unsupervised learning: generative models

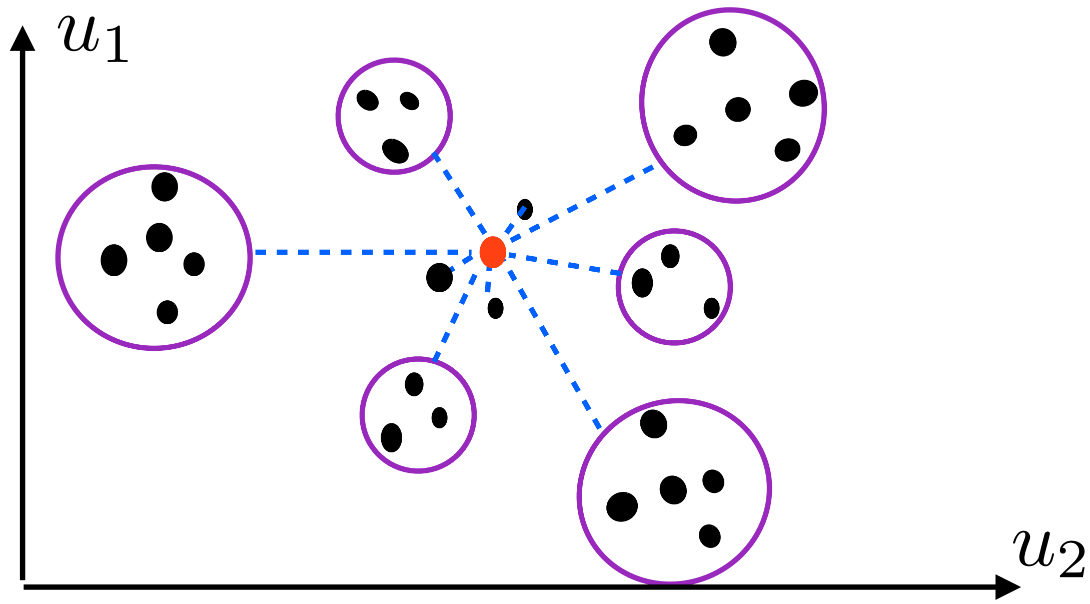
Part III Inverse problems

Dimensionality Reduction Multiscale

- Why can we learn despite the curse of dimensionality ?

  Multiscale structures/interactions

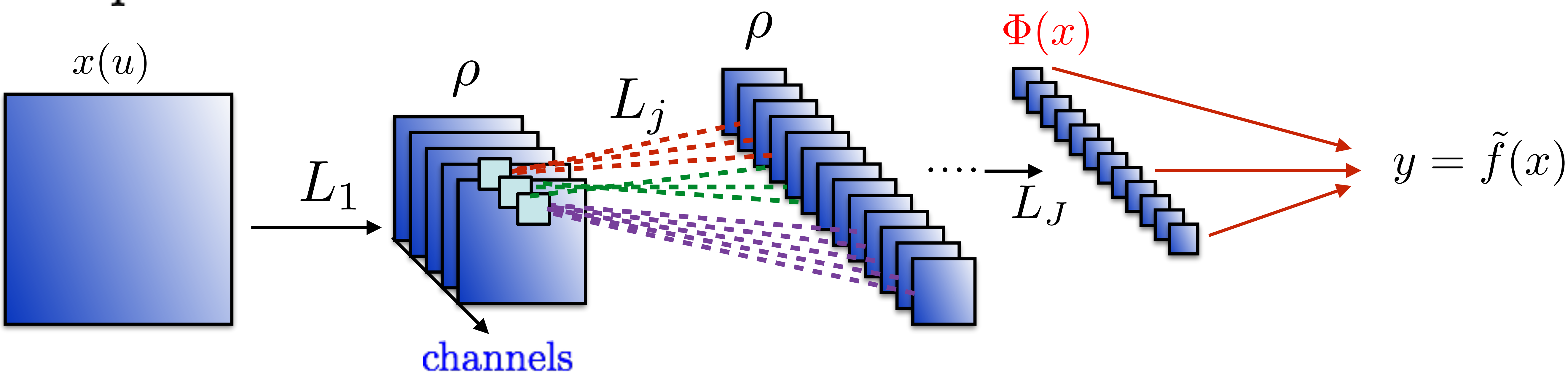
Interactions de d variables $x(u)$: pixels, particules, agents...



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

Deep Convolutional Trees

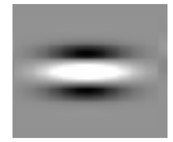
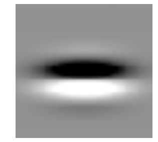
Simplified architecture:



Cascade of convolutions: no channel connections

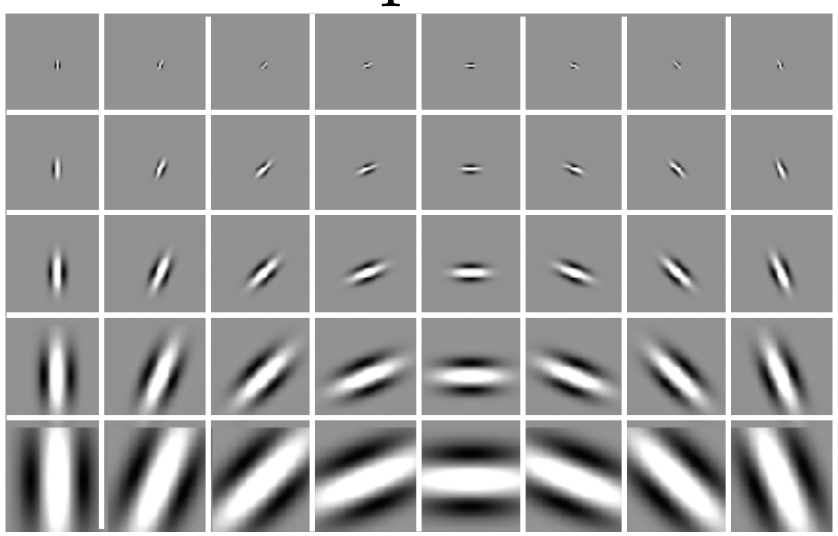
predefined wavelet filters

Scale separation with Wavelets

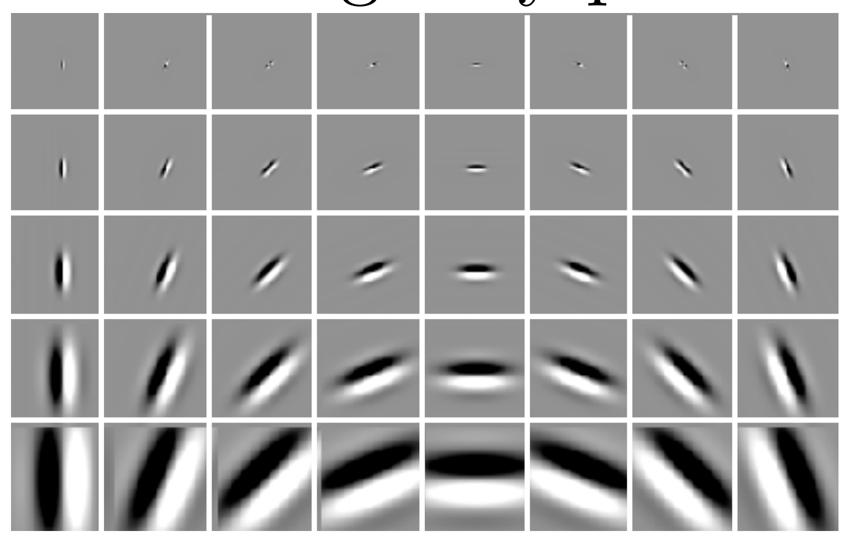
- Wavelet filter $\psi(u)$:  

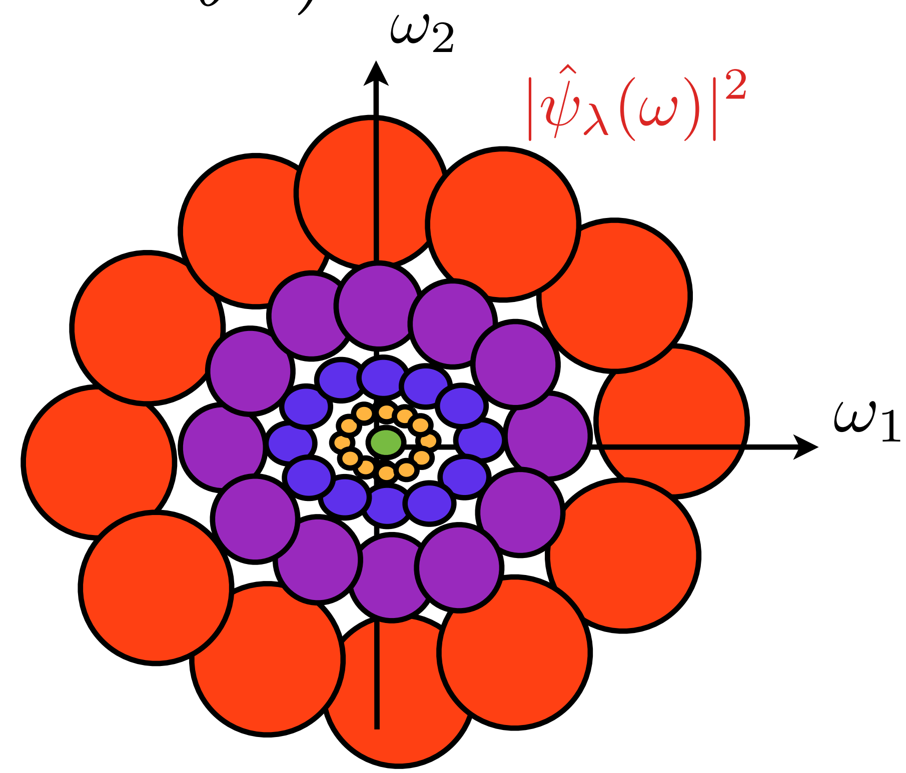
rotated and dilated: $\psi_{2^j,\theta}(u) = 2^{-j}\,\psi(2^{-j} r_\theta u)$

real parts



imaginary parts



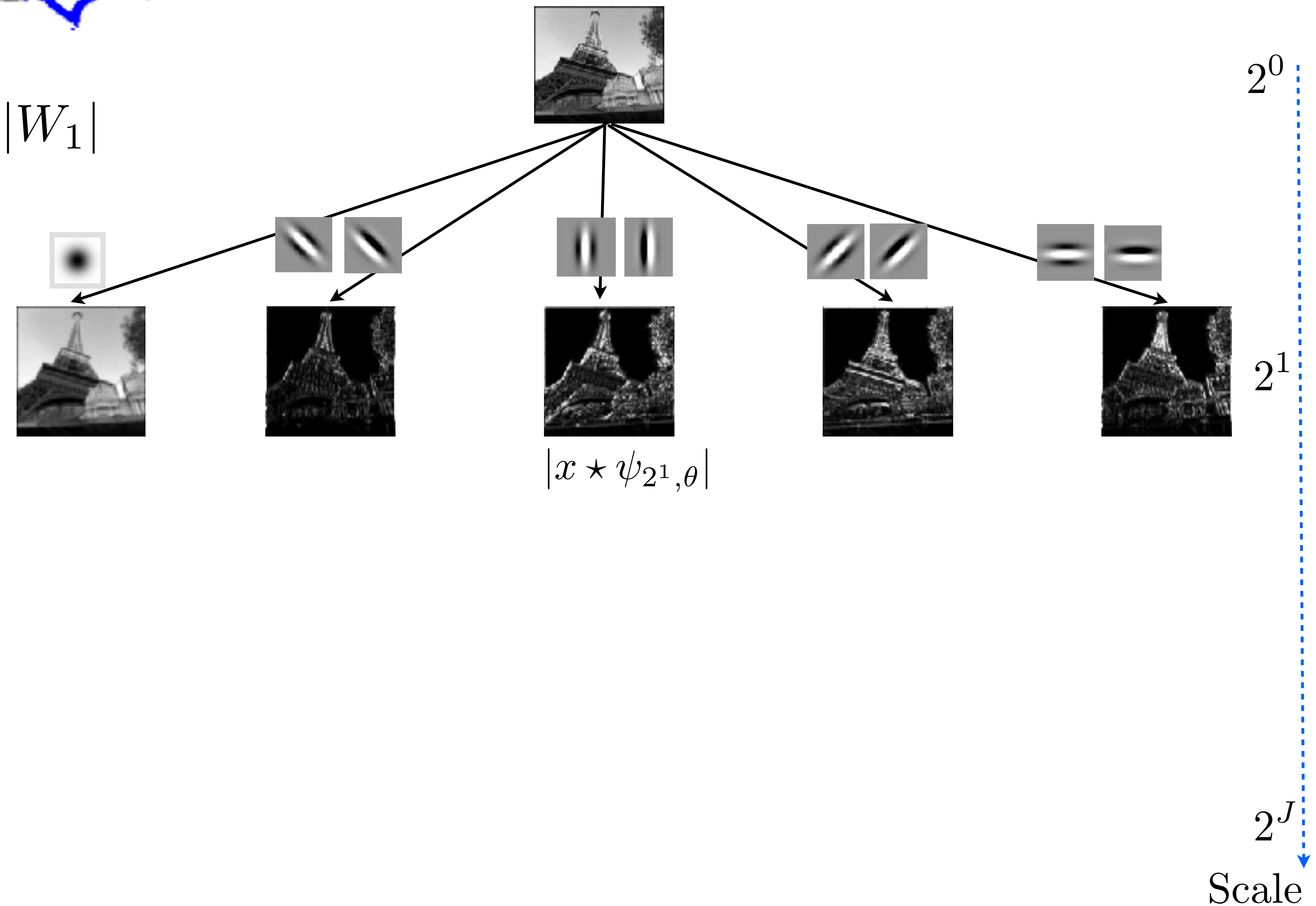


$$x \star \psi_{2^j,\theta}(u) = \int x(v)\,\psi_{2^j,\theta}(u-v)\,dv$$

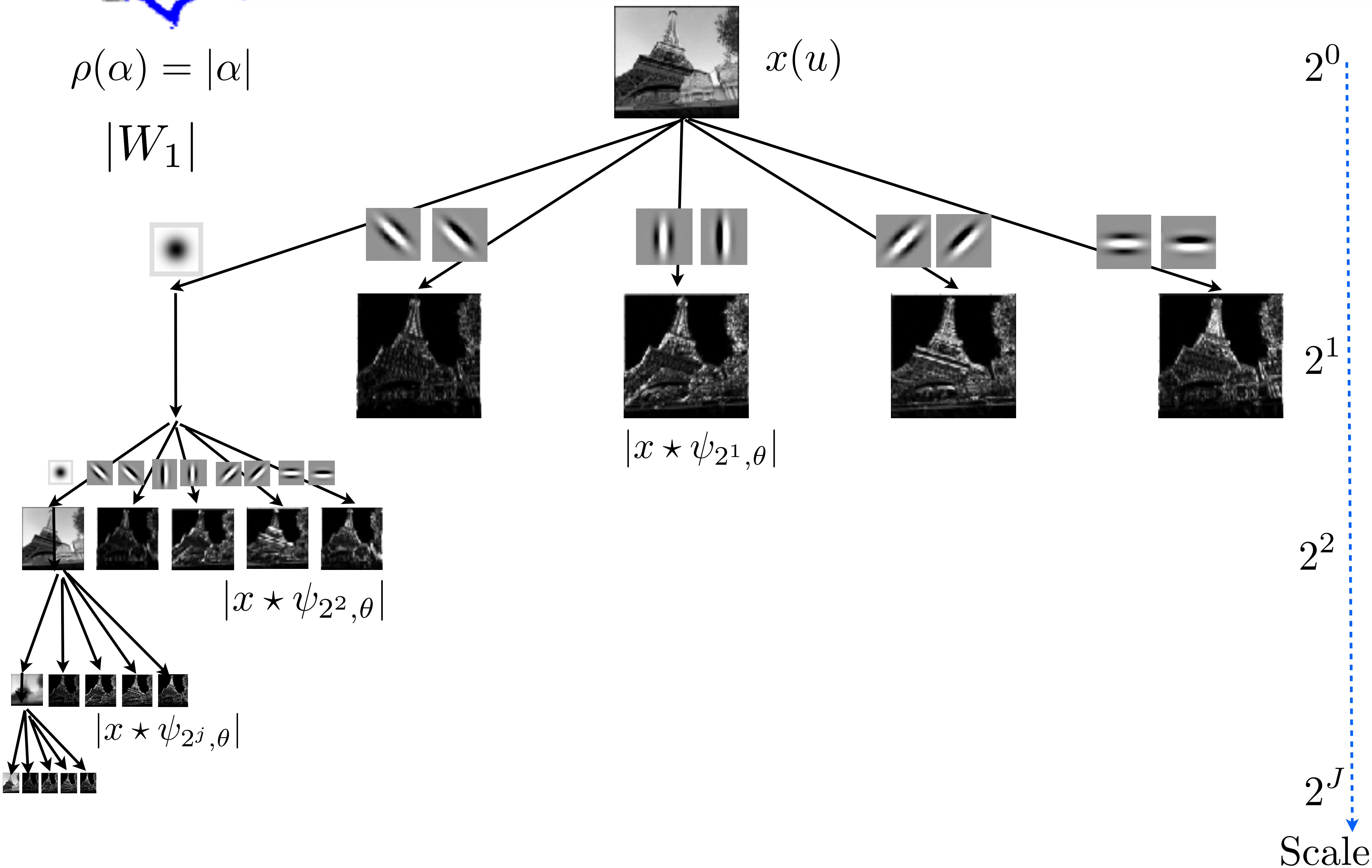
- Wavelet transform: $W x = \begin{pmatrix} x \star \phi_{2^J}(u) \\ x \star \psi_{2^j,\theta}(u) \end{pmatrix}_{j \le J, \theta}$: average; : higher frequencies

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

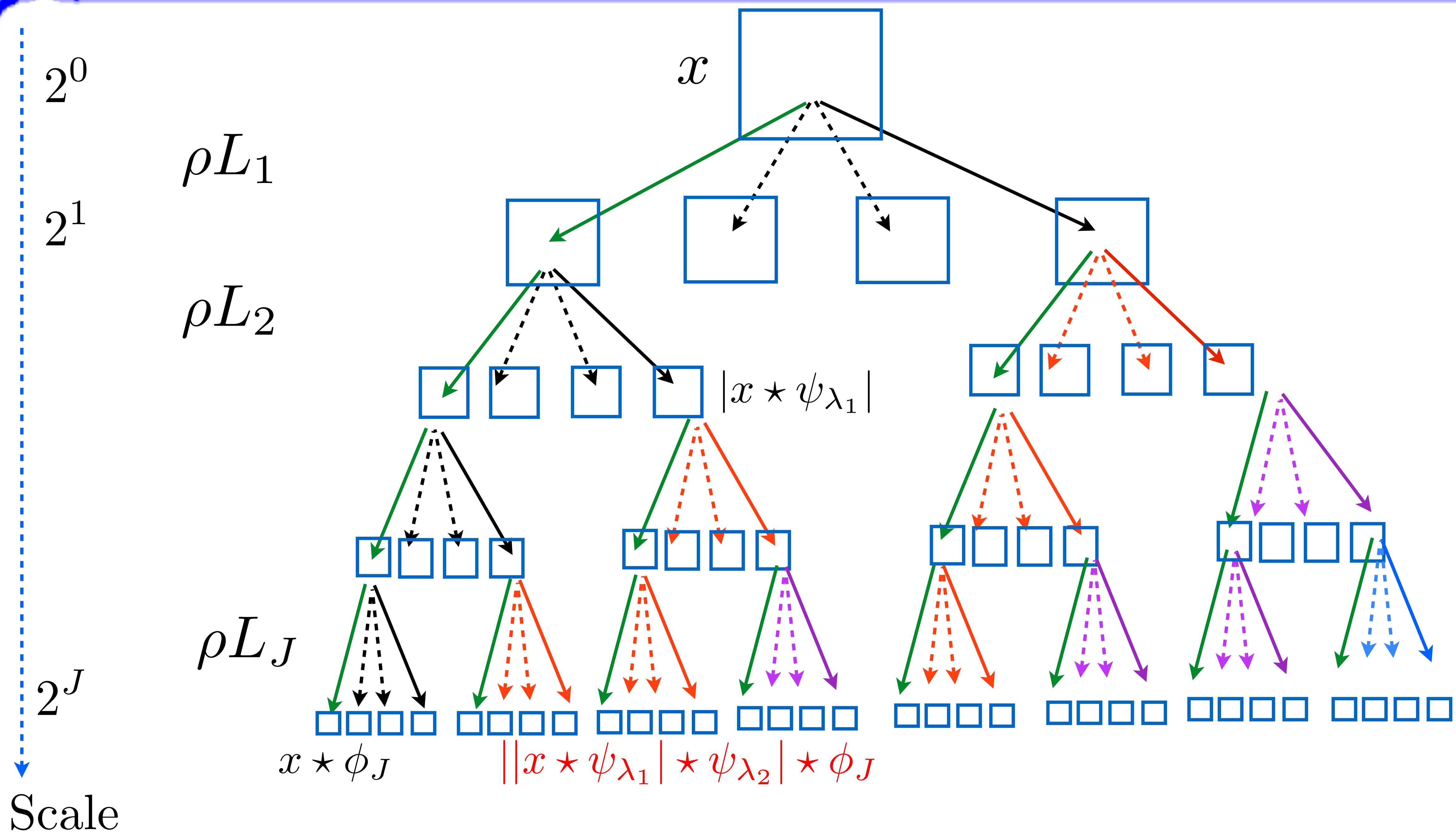
Fast Wavelet Filter Bank



Wavelet Filter Bank



Wavelet Scattering Network



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

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

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

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

Theorem: *For appropriate wavelets, a scattering is*

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

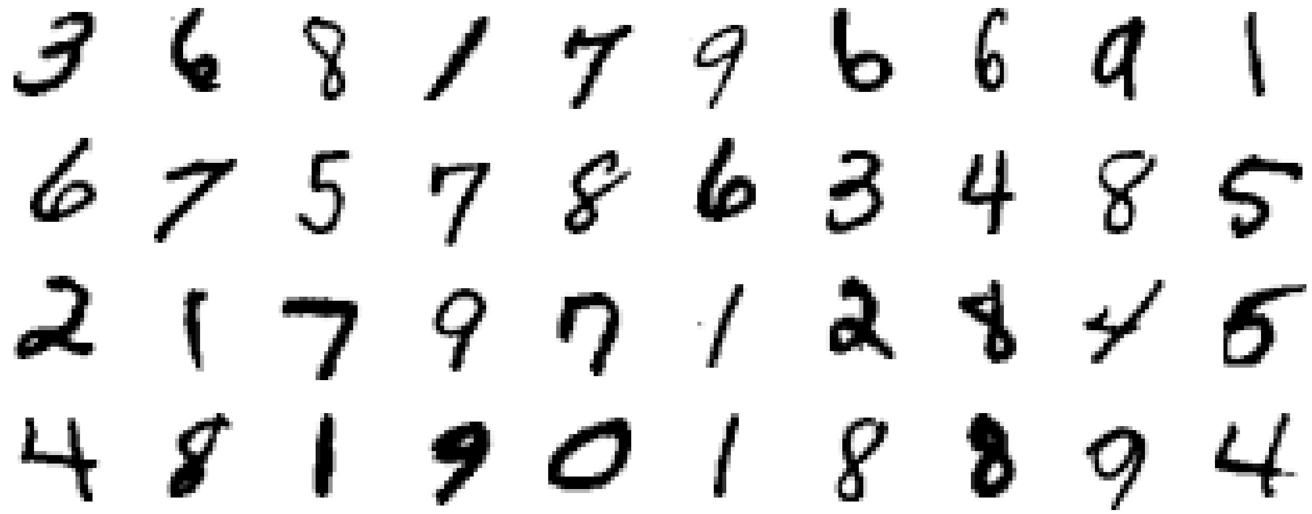
preserves norms $\|S_J x\| = \|x\|$

translations invariance and deformation stability:

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

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

Digit Classification: MNIST



Joan Bruna

x → $S_J x$ → Supervised Linear classifier → $y = f(x)$

Invariants to translations
Linearises small deformations
No learning

Invariants to specific deformations
Separates different patterns

Classification Errors

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

LeCun et. al.

Part II- Unsupervised Learning

joint work with Joan Bruna

Unsupervised learning:

Approximate the probability distribution $p(x)$ of $X \in \mathbb{R}^d$ given P realisations $\{x_i\}_{i \leq P}$ with potentially $P = 1$

Which class of processes can we approximate ?

Stationary Processes

- Ergodic versus non-ergodic (long-range dependance)
- Capture non-Gaussianity: geometry of realisations

Scattering/Deep Net. of a stationary process $X(t)$

$$S_J X = \left(\begin{array}{c} X \star \phi_{2^J}(t) \\ |X \star \psi_{\lambda_1}| \star \phi_{2^J}(t) \\ ||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_{2^J}(t) \\ |||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}| \star \phi_{2^J}(t) \\ ... \end{array} \right)_{\lambda_1,\lambda_2,\lambda_3,...} : \text{ stationary vector}$$

Ergodicity and Moments

Scattering transform of a stationary vector $X \in \mathbb{R}^d$

maximum scale: $2^J = d$

$$S_J X = \begin{pmatrix} d^{-1} \sum_{u=1}^{d} X(u) \\ d^{-1} \| X \star \psi_{\lambda_1} \|_1 \\ d^{-1} \| |X \star \psi_{\lambda_1}| \star \psi_{\lambda_2} \|_1 \\ d^{-1} \| ||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3} \|_1 \\ \ldots \end{pmatrix}_{\lambda_1, \lambda_2, \lambda_3, \ldots}$$

$d \to \infty$

Central limit theorem
with "weak" ergodicity conditions

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

Generation of Random Processes

Scattering transform of a stationary vector $X \in \mathbb{R}^d$

maximum scale: $2^J = d$

$$S_J X = \begin{pmatrix} d^{-1} \sum_{u=1}^{d} X(u) \\ d^{-1} \|X \star \psi_{\lambda_1}\|_1 \\ d^{-1} \||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}\|_1 \\ d^{-1} \|||X \star \psi_{\lambda_2}| \star \psi_{\lambda_2}| \star \psi_{\lambda_3}\|_1 \\ ... \end{pmatrix}_{\lambda_1, \lambda_2, \lambda_3, ...}$$

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

$$S_J \tilde{X} \approx S_J X$$

with random initialisation and gradient descent.

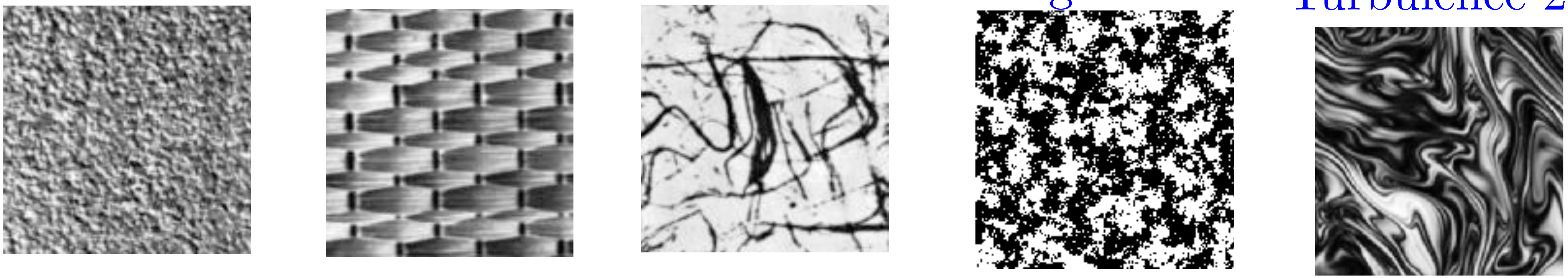
Texture Reconstructions

Joan Bruna

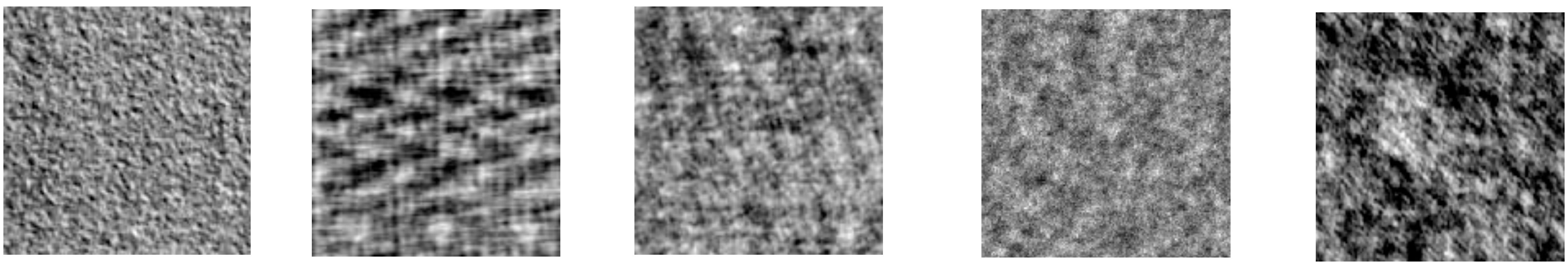
Statistical Physics

Texture of d pixels

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

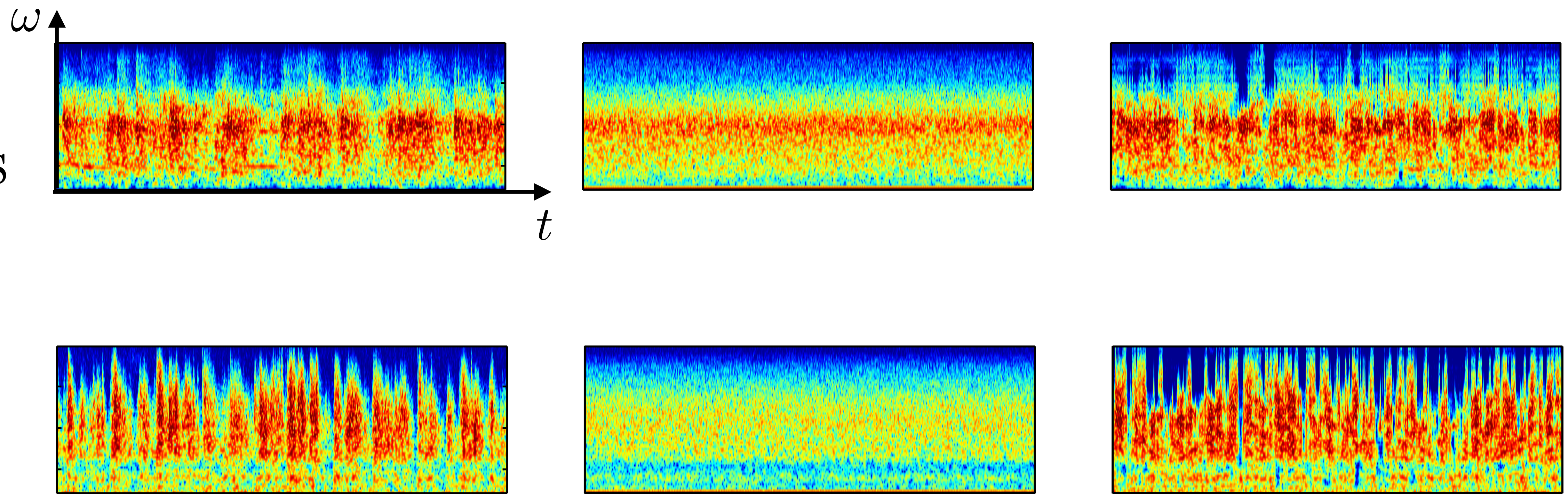
Original

Gaussian in time

Scattering Order 2

Applauds

Paper



Cocktail Party

Max Entropy Canonical Models

- A representation $\Phi(x) = \{\phi_k(x)\}_{k \leq K}$ with $x \in \mathbb{R}^d$

- Canonical distribution $p(x)$ of X satisfies

$$\mu_k = \mathbb{E}(\phi_k X) = \int \phi_k(x)\, p(x)\, dx$$

with maximum entropy: $H(p) = -\int p(x)\, \log p(x)\, dx$

$$\Rightarrow \quad p(x) = Z^{-1} \exp\Big(-\sum_k \theta_k \phi_k(x) \Big)$$

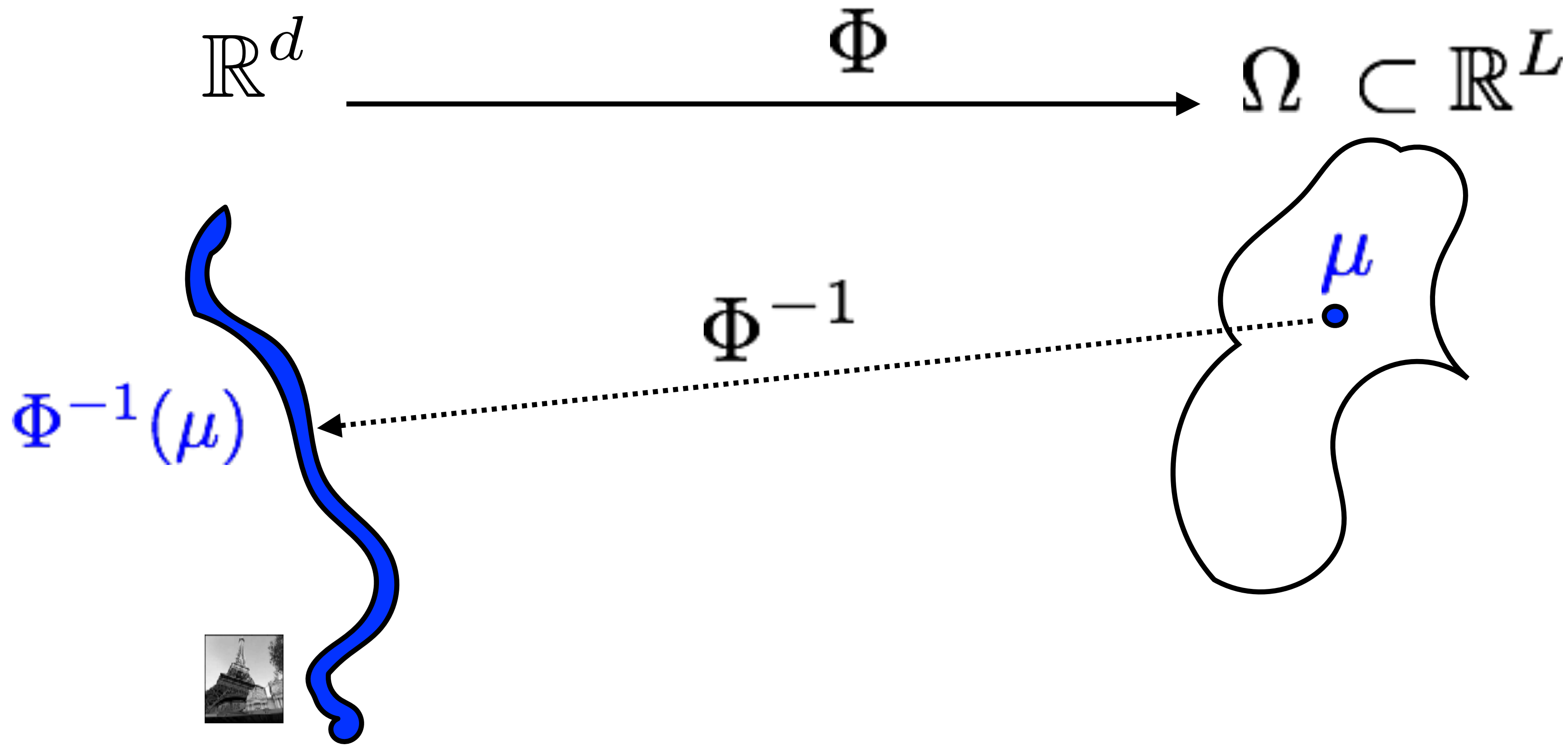
Gaussian, Markov random field models

- **Problem:** in other cases we can't compute the θ_k.

Ergodic Microcanonical Model

- If concentration: $\text{Prob}\Big(|\Phi X - \mu| < \epsilon\Big) \underset{d \to \infty}{\longrightarrow} 1$

$$\text{with } \mu = \mathbb{E}(\Phi X)$$

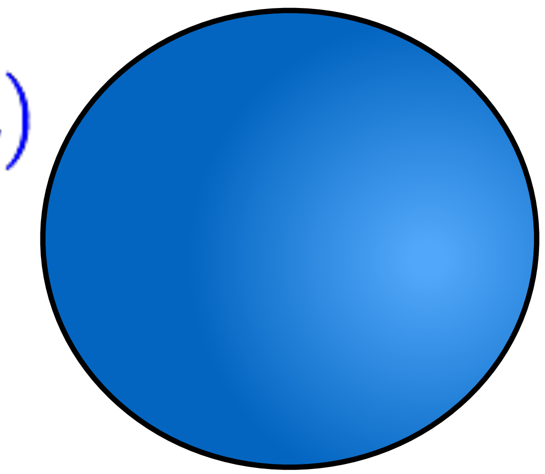


A microcanonical model $\tilde{X}$ has a distribution $\tilde{p}$ of maximum entropy conditioned to $\Phi \tilde{X} = \mu$ which is uniform in $\Phi^{-1}(\mu)$ (if compact)

Uniform Distribution on Balls

- Sphere in $\mathbb{R}^d$ $\quad \Phi x = d^{-1/2}\|x\|_2 = \left(d^{-1}\sum_{k=1}^{d}|x(k)|^2\right)^{1/2} = \mu$

$\Phi^{-1}(\mu)$



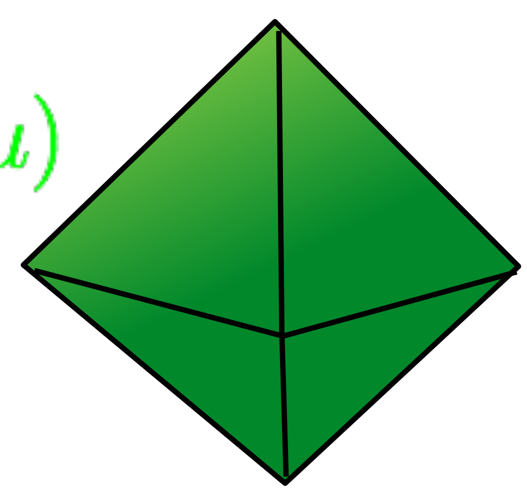
not a low-dimensional manifold !

Borel 1914
Diaconis, Freedman 1987

$$\tilde{X}(1),...,\tilde{X}(K) \underset{d\to\infty}{\longrightarrow} \text{.i.i.d Gaussian} \quad \sim e^{-u^2/2\sigma^2}$$

- Simplex in $\mathbb{R}^d$ $\quad \Phi x = d^{-1}\|x\|_1 = d^{-1}\sum_{k=1}^{d}|x(k)| = \mu$

$\Phi^{-1}(\mu)$



Diaconis, Freedman 1987

$$\tilde{X}(1),...,\tilde{X}(K) \underset{d\to\infty}{\longrightarrow} \text{i.i.d Exponential} \quad \sim e^{-\lambda|u|}$$

Scattering Representation

- Scattering coefficients of order 0, 1 and 2; up to scale 2^J

$$\Phi x = \left\{ d^{-1} \sum_u x(u) \ , \ d^{-1} \|x \star \psi_{\lambda_1}\|_1 \ , \ d^{-1} \| |x \star \psi_{\lambda_1}| \star \psi_{\lambda_2} \|_1 \right\}$$

$\Phi^{-1}(\mu)$ is an intersection of about J^2 polytopes in $\mathbb{R}^d$

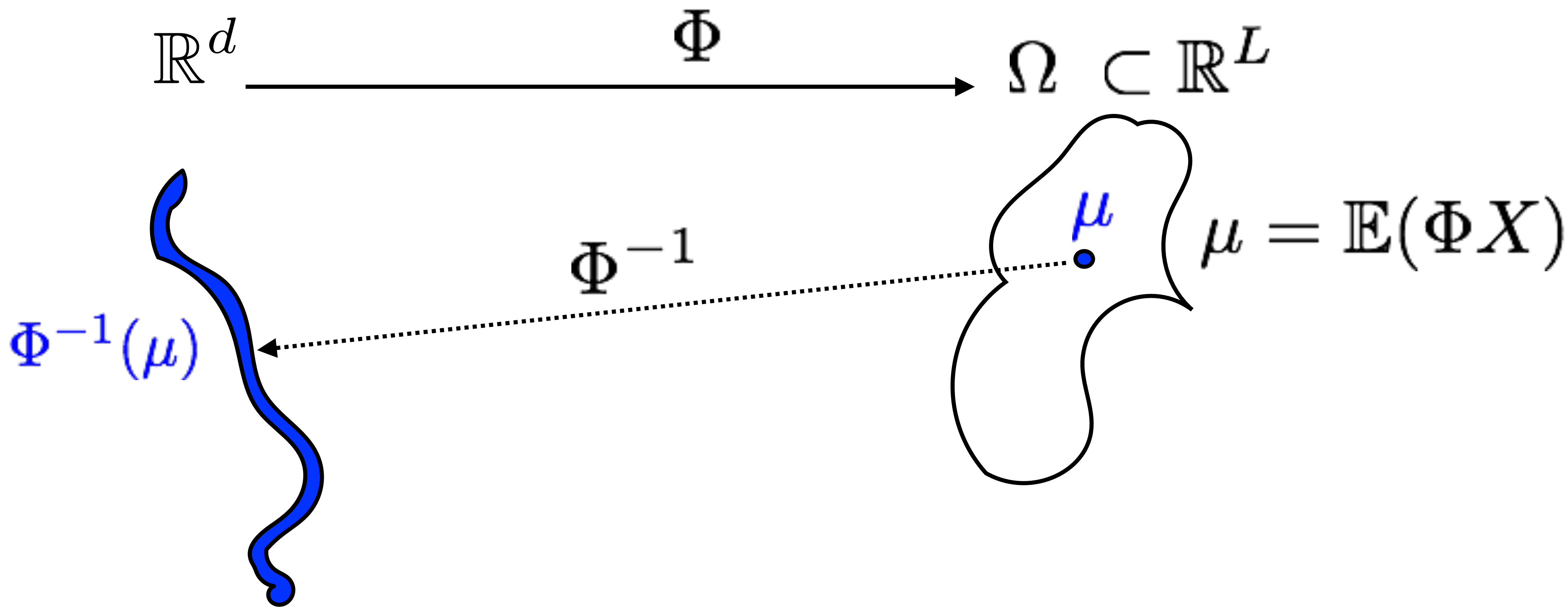
Complex high-dimensional geometry

- Reproduces $\mathbf{l^2}$ norms

$$d^{-1} \|x \star \psi_{\lambda_1}\|_2^2 = d^{-2} \|x \star \psi_{\lambda_1}\|_1^2 + \sum_{\lambda_2} d^{-2} \| |x \star \psi_{\lambda_1}| \star \psi_{\lambda_2} \|_2^2 + \text{higher order}$$

Specify $\{\|x \star \psi_{\lambda_1}\|_2\}_{\lambda_1}$: intersection of $\mathbf{l^2}$ balls

Microcanonical Scattering

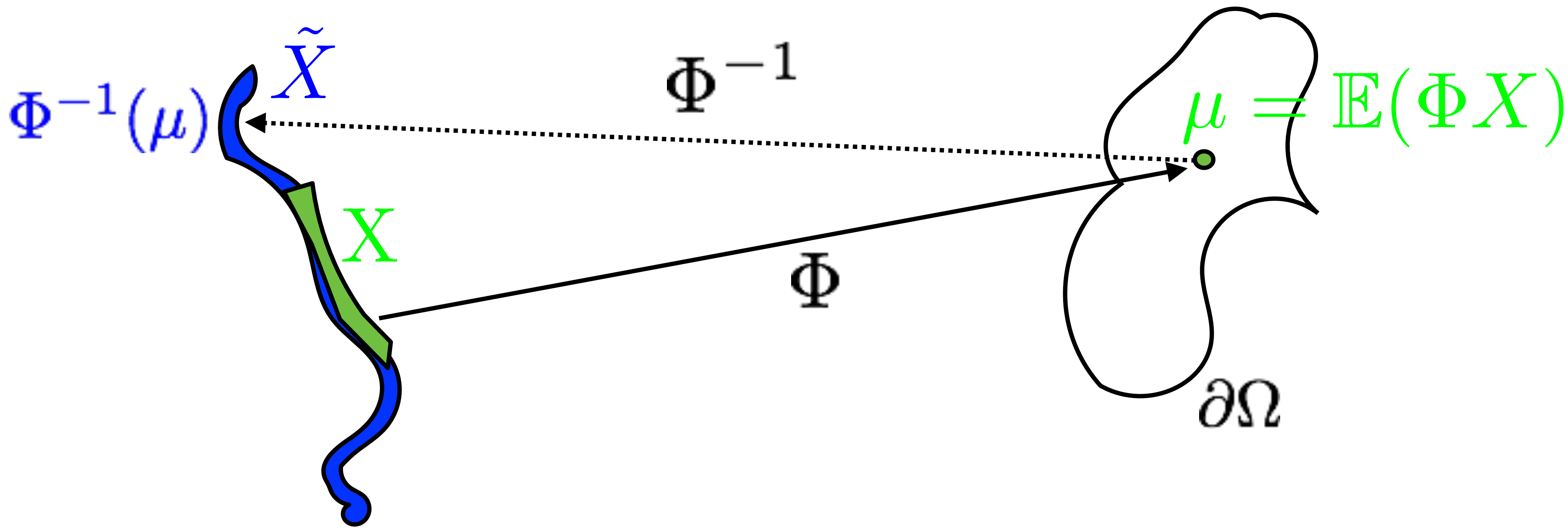


Proposition If $X(u)$ is stationary and

$X(u)$ and $X(v)$ are independent for $|u - v| \geq \Delta$

then $\lim_{d \to \infty} \mathbb{E}(\|\Phi X - \mu\|^2) = 0$

Scattering Approximations



Theorem If $X(u)$ is stationary and

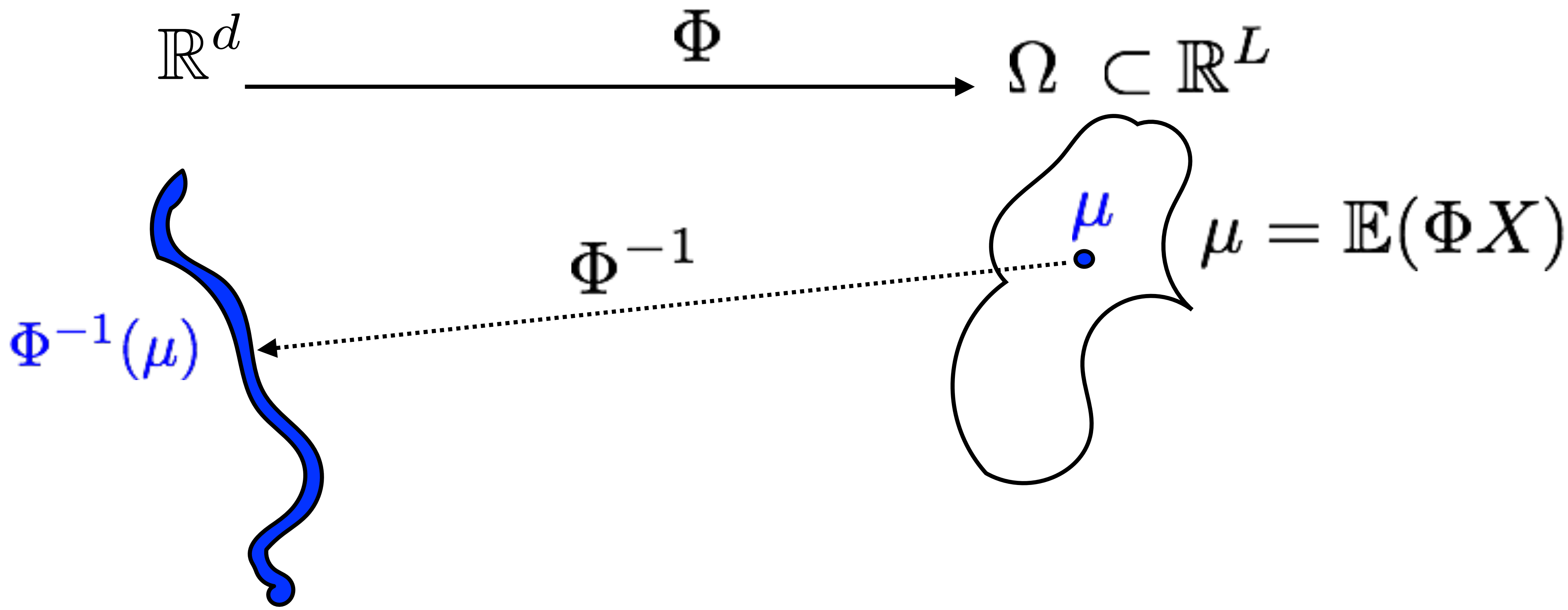
$$X(u) \text{ and } X(v) \text{ are independent for } |u-v| \geq \Delta$$

If Typical of $\tilde{X}$ is typical of X

$$\text{and} \quad \lim_{d \to \infty} \mathbb{E}(|d^{-1} \log p(\tilde{X}) - H(p)|^2) = 0 \quad \text{then}$$

$$\tilde{X}(1), ..., \tilde{X}(K) \text{ converges in probability to } X(1), ..., X(K)$$

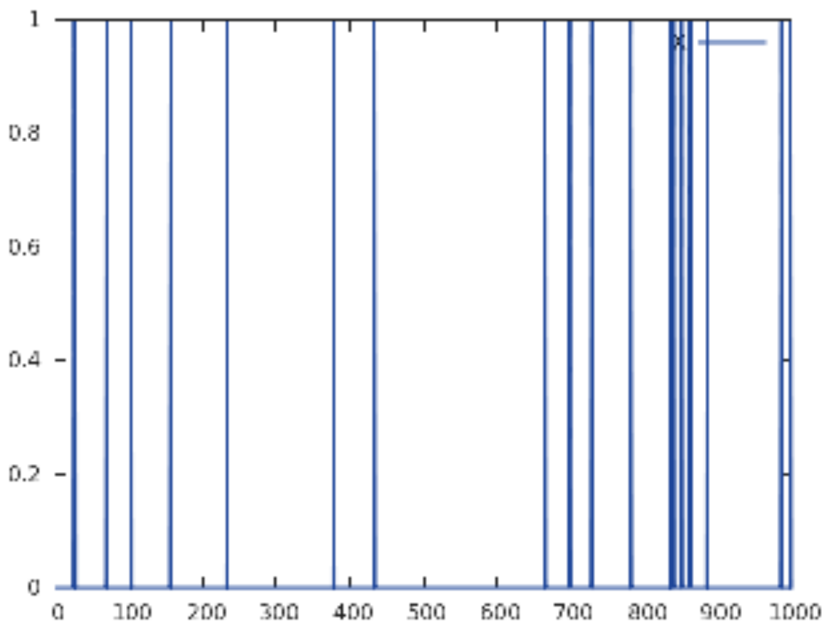
Ergodic Microcanonical Model



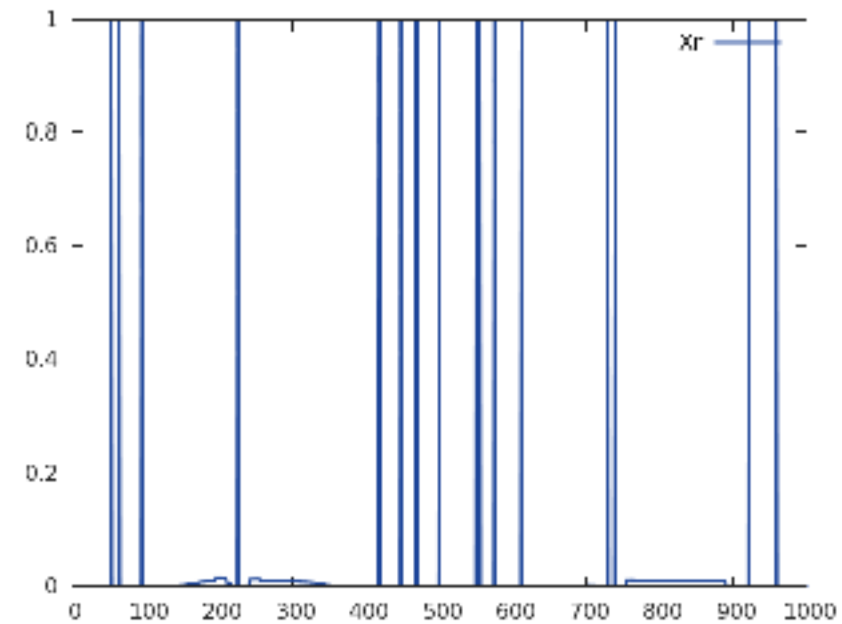
If X is Gaussian stationary
with a bounded and regular spectrum
then for a scattering with appropriate wavelets
$\tilde{X}(1), ..., \tilde{X}(K)$ converges in probability to $X(1), ..., X(K$
up to an arbitrary small error ϵ

Singular Ergodic Processes

Bernoulli X



Scattering Microcanonical $\tilde{X}$



Concentration of ΦX Typical of $\tilde{X}$ is typical of X

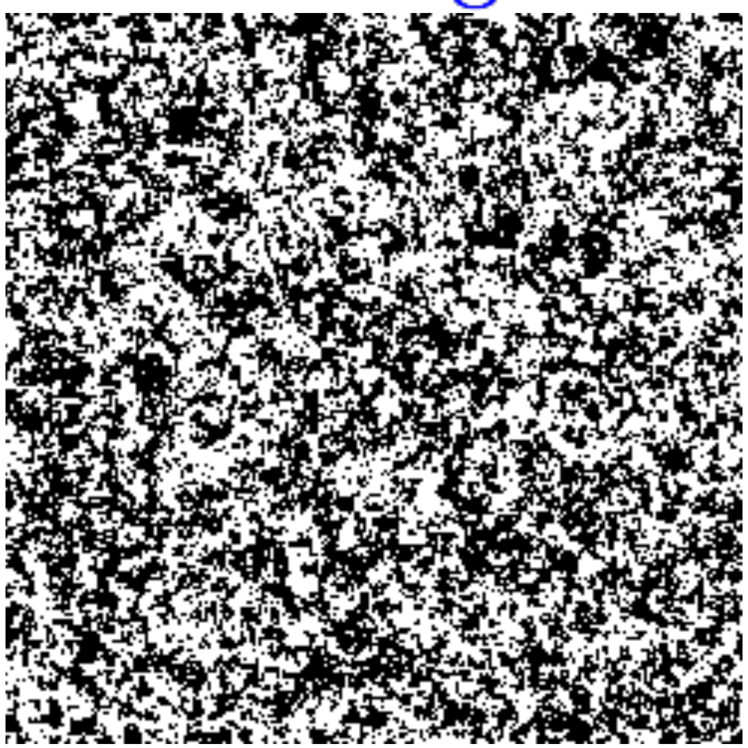
Why ?

Scattering Ising

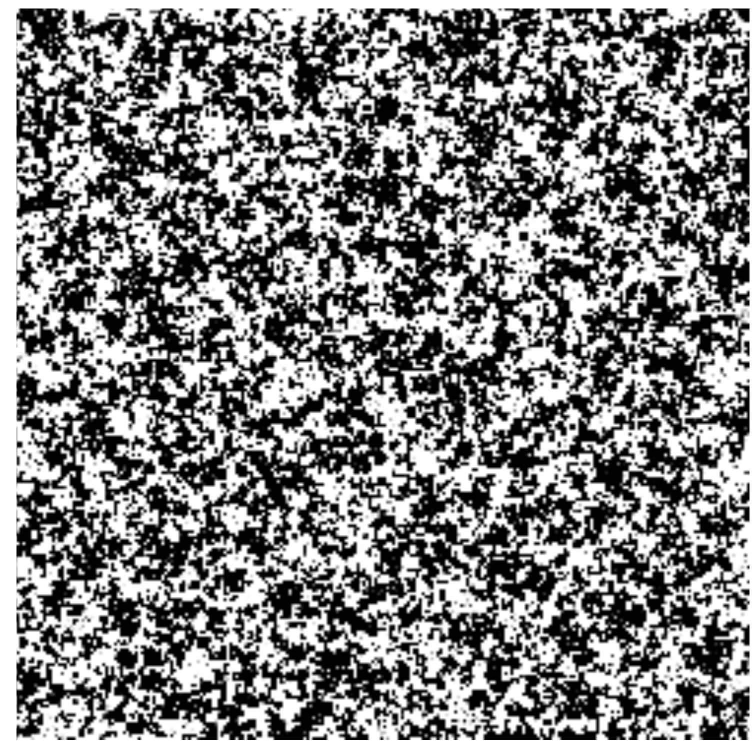
$$x(u) \in \{0,1\} \qquad p(x) = Z^{-1} \exp\Big(\frac{1}{T} \sum_{(u,u') \in C_I} x(u)\, x(u')\Big)$$

Ising X for $T \geq T_{critic}$

Ergodic



Microcanonical Scat $\tilde{X}$



Concentration of ΦX — Typical of $\tilde{X}$ is typical of X

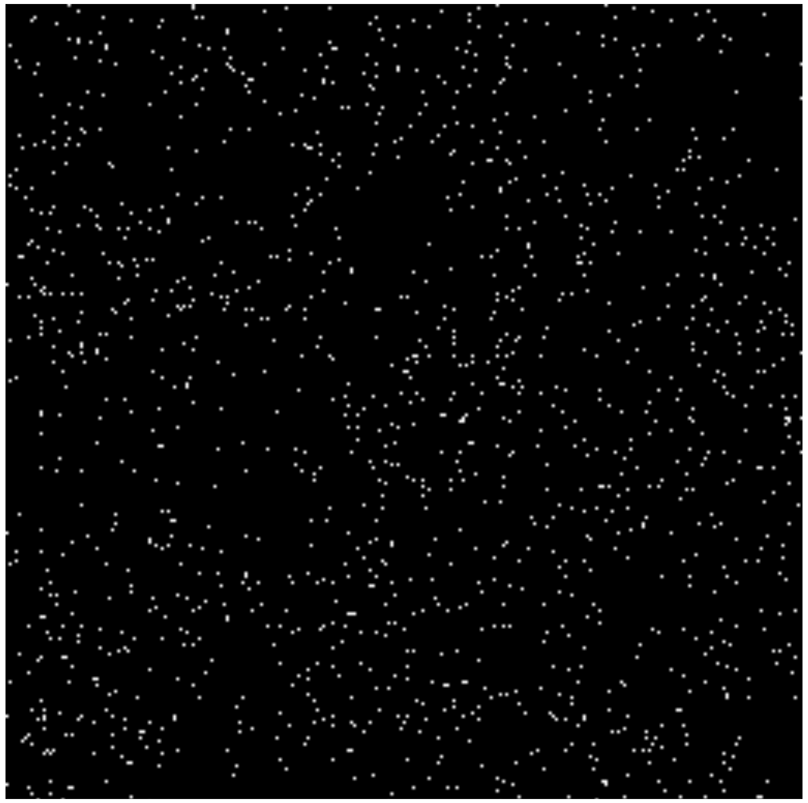
d	$\frac{\mathbb{E}(\|\Phi(X) - \mathbb{E}\Phi(X)\|^2)}{\|\mathbb{E}\Phi(X)\|^2}$	$\frac{\mathbb{E}(\lvert d^{-1} \log p(\tilde{X}) - H(p)\rvert^2)}{H(p)^2}$
2^{12}	$3 \cdot 10^{-4}$	$1 \cdot 10^{-5}$
2^{14}	$1 \cdot 10^{-4}$	$5 \cdot 10^{-6}$

Why ?

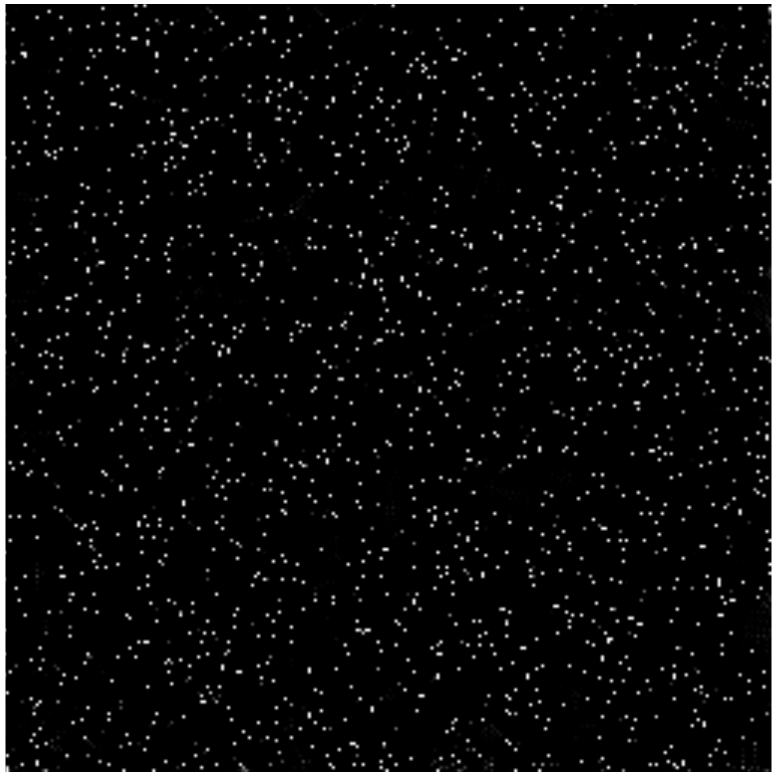
Stochastic Geometry: Cox Process

Bernoulli with random density $\lambda(u)$

Cox X Ergodic



Microcanonical Scat $\tilde{X}$



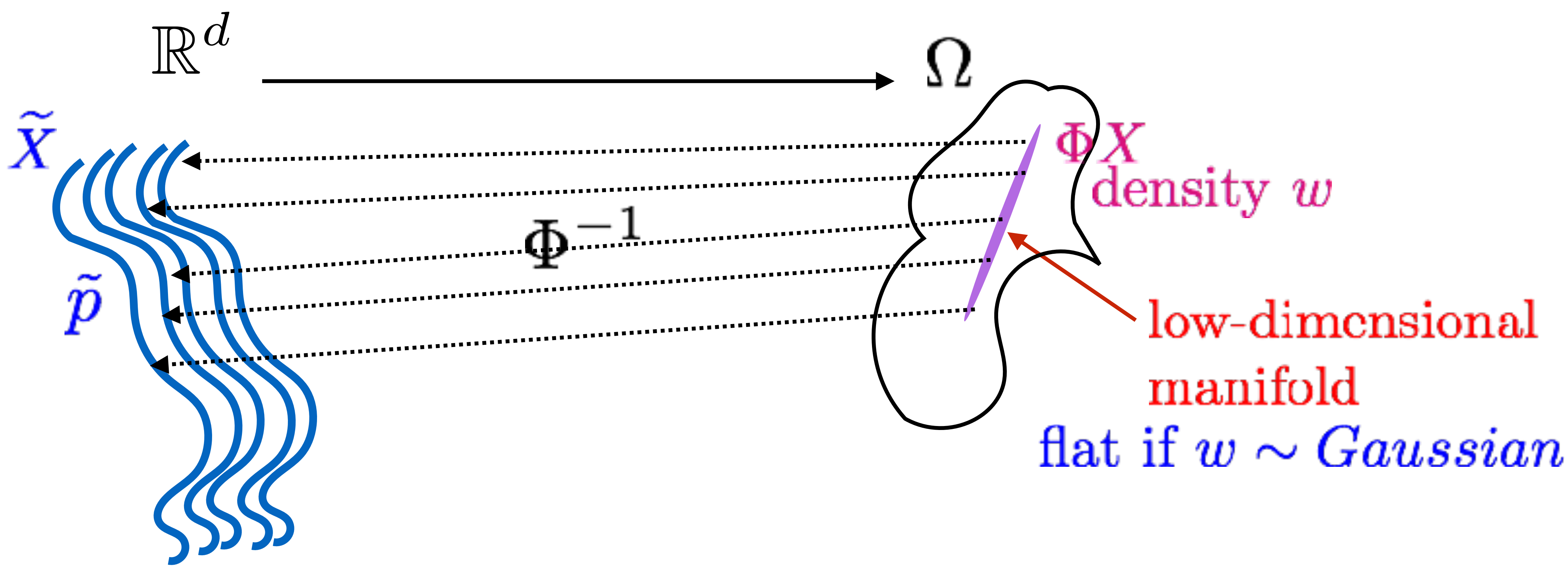
Concentration of ΦX

Typical of $\tilde{X}$ is typical of X

d	$\frac{\mathbb{E}(\|\Phi(X)-\mathbb{E}\Phi(X)\|^2)}{\|\mathbb{E}\Phi(X)\|^2}$	$\frac{\mathbb{E}(\|d^{-1}\log p(\tilde{X})-H(p)\|^2)}{H(p)^2}$
2^{12}	$3 \cdot 10^{-4}$	
2^{14}	$1 \cdot 10^{-4}$	

Non-Ergodic Mixture

- Non-ergodicity: $\Phi(X)$ does not concentrate in all directions

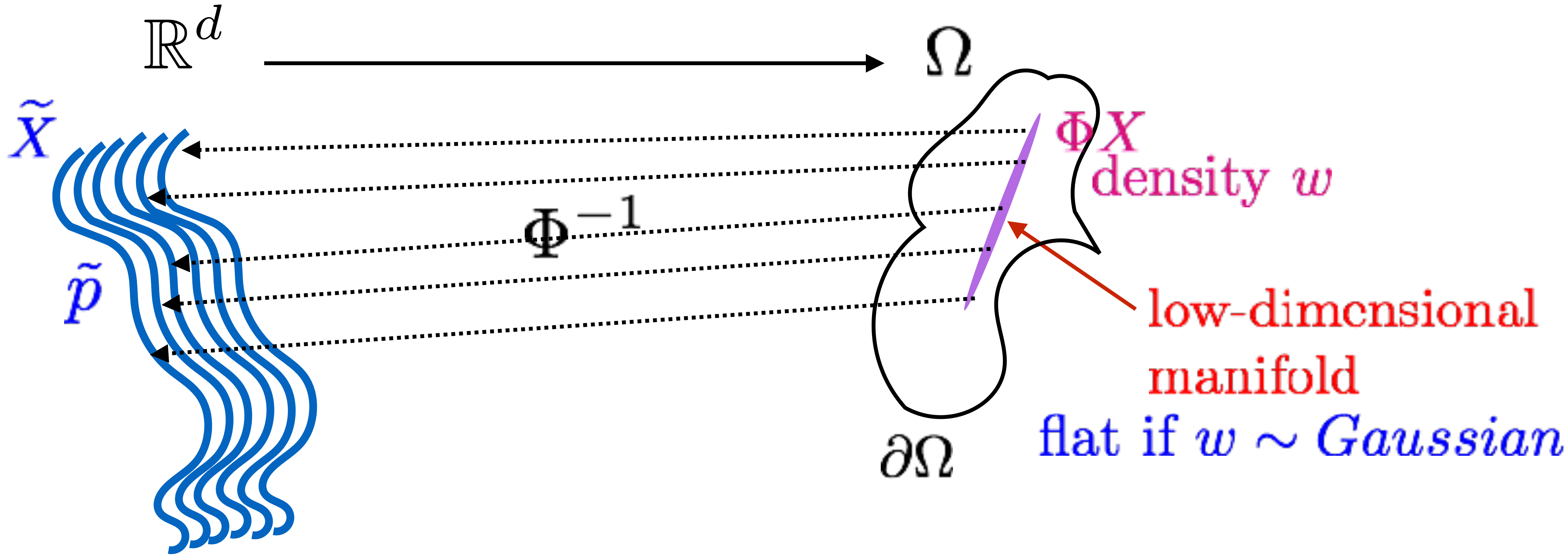


Maximum entropy conditioned to $\Phi\tilde{X}$ having a density w

micro canonical mixture $\tilde{X}$ weighted by the density w of ΦX

Non-Ergodic Microcanonical Mixture

- Non-ergodicity: $\Phi(X)$ does not concentrate in all directions



Theorem A microcanonical mixture has a density $\tilde{p}$ with

$$\tilde{p}(x) = \frac{w(\Phi x)}{h(\Phi x)}$$

with $\quad h(y) = \int_{\Phi^{-1}(y)} |J_L \Phi x|^{-1} \, d\mathcal{H}^{d-L}(x)$

which is singular only if $\Phi x \in \partial\Omega$

Scattering Multifractal Processes

- Multifractal processes with stationary increment have non-ergodic low-frequencies: long-range correlations.

- Wavelet coefficients $X \star \psi_\lambda(u)$ decorrelate at larger scales

- Scattering coefficients of order 0, 1 and 2:

$$\Phi X = \Big\{ d^{-1} \sum_u X(u) \, , \, d^{-1} \| X \star \psi_{\lambda_1} \|_1 \, , \, d^{-1} \| |X \star \psi_{\lambda_1}| \star \psi_{\lambda_2} \|_1 \Big\}$$

non-ergodic ergodic ergodic

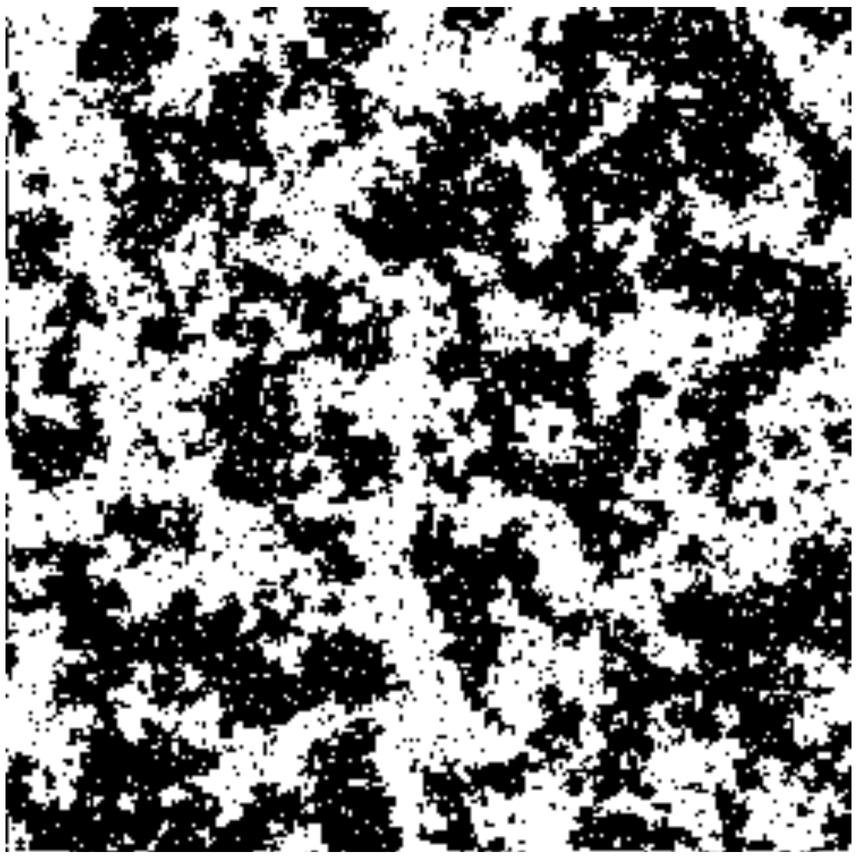
$\Rightarrow$ one-dimensional mixture weight w (non-ergodic part) can be estimated from few examples: manifold.

Scat Ising at Critical Temperature

$$p(x) = Z^{-1} \exp\left(\frac{1}{T} \sum_{(u,u') \in C_I} x(u)\, x(u')\right)$$

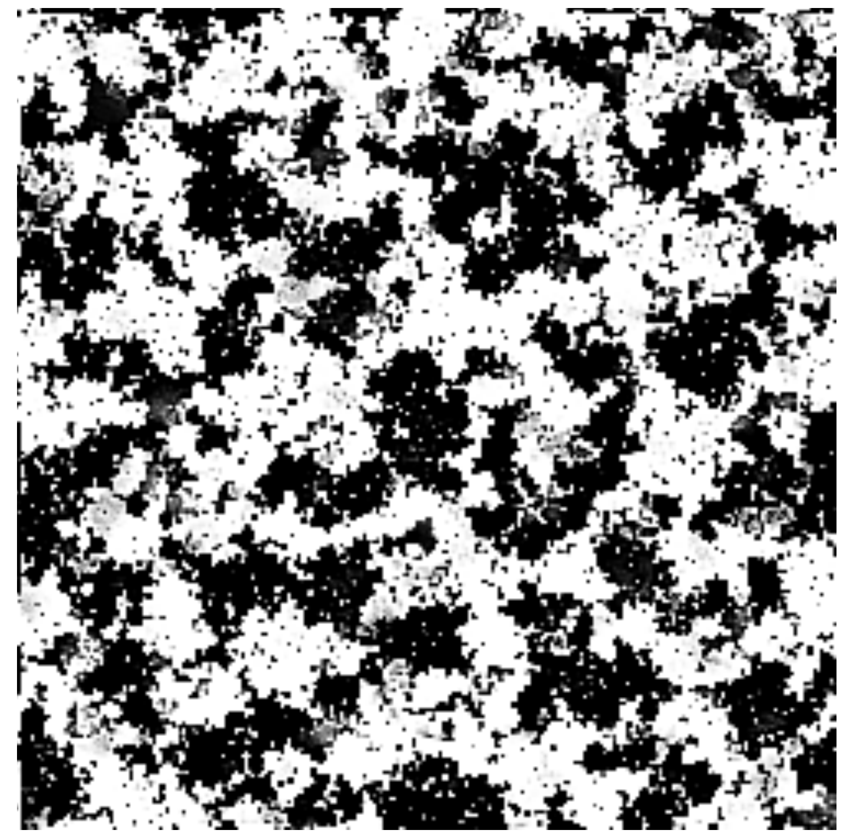
Ising X for $T = T_{critic}$

Non ergodic



Concentration of ΦX without low-freq.

Microcanonical Scat $\tilde{X}$



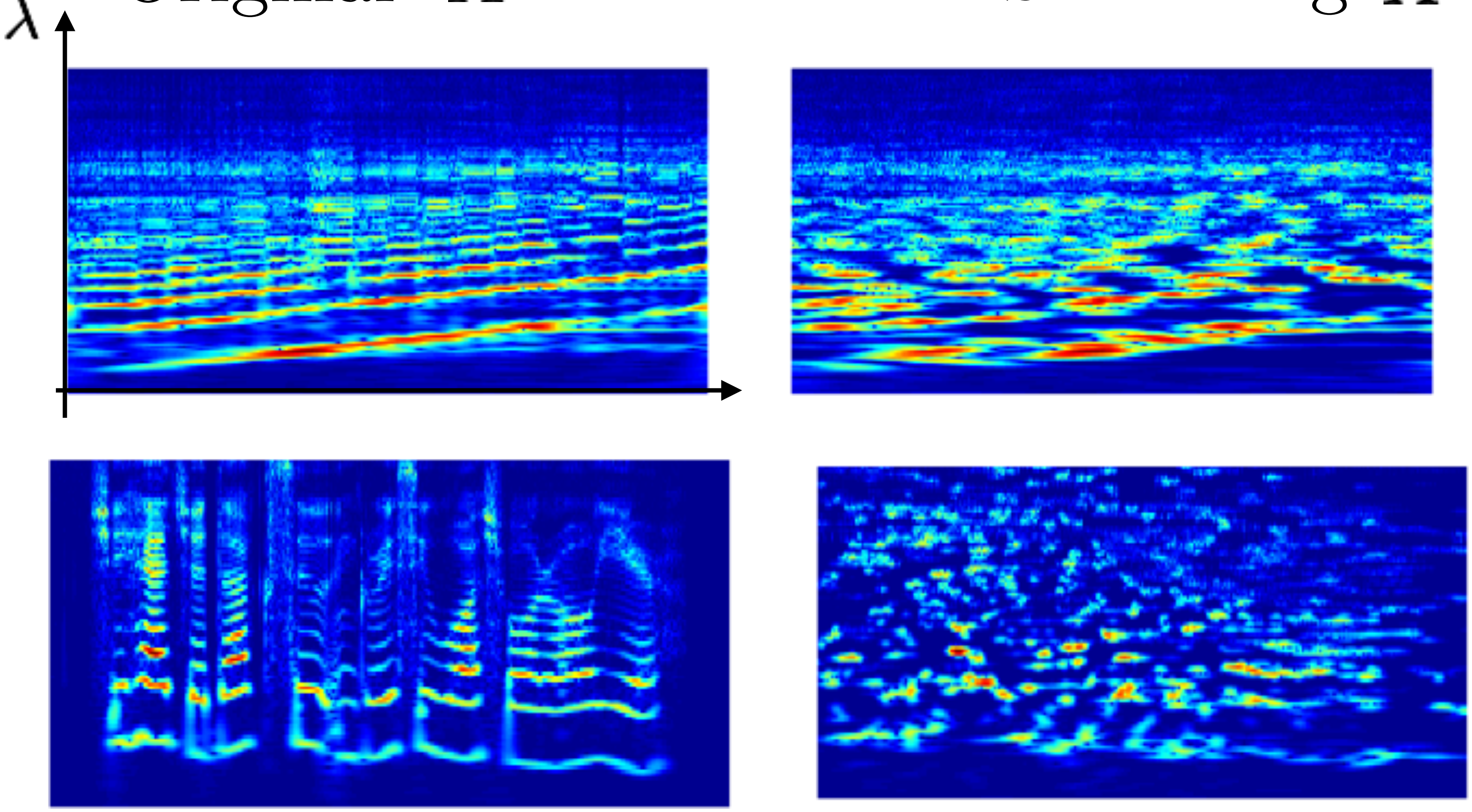
Typical of $\tilde{X}$ is typical of X

d	$\frac{\mathbb{E}(\|\Phi(X) - \mathbb{E}\Phi(X)\|^2)}{\|\mathbb{E}\Phi(X)\|^2}$	$\frac{\mathbb{E}(\|d^{-1}\log p(\tilde{X}) - H(p)\|^2)}{H(p)^2}$
2^{12}	$8 \cdot 10^{-3}$	$2 \cdot 10^{-3}$
2^{14}	$2.5 \cdot 10^{-3}$	$2 \cdot 10^{-4}$

Failures of Audio Synthesis

J. Anden and V. Lostanl

Original X Time Scattering $\tilde{X}$



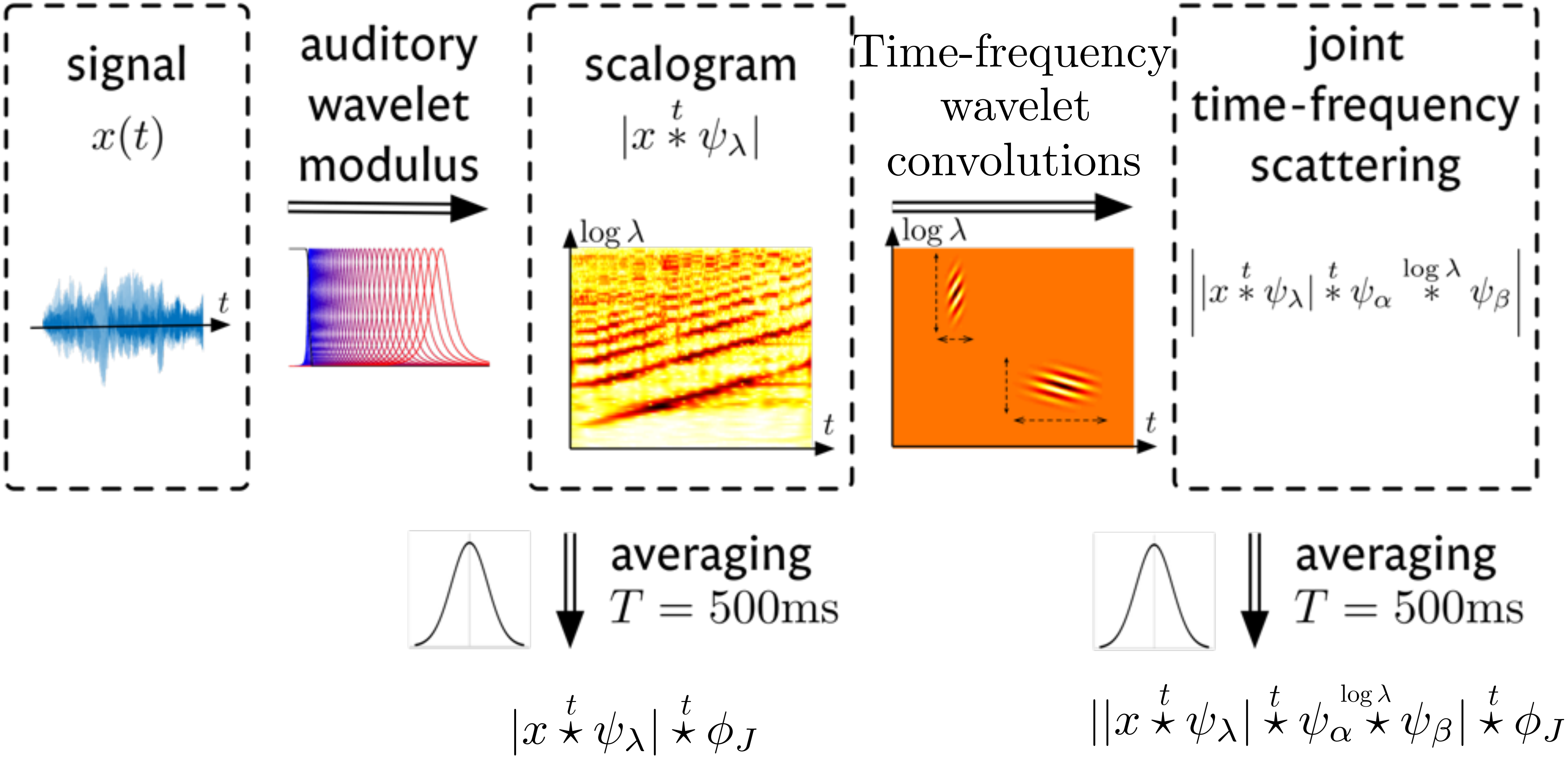
Typical of $\tilde{X}$ is not typical of X

- Missing frequency connections $\Rightarrow$ misalignments

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

Time-Frequency Translation Group

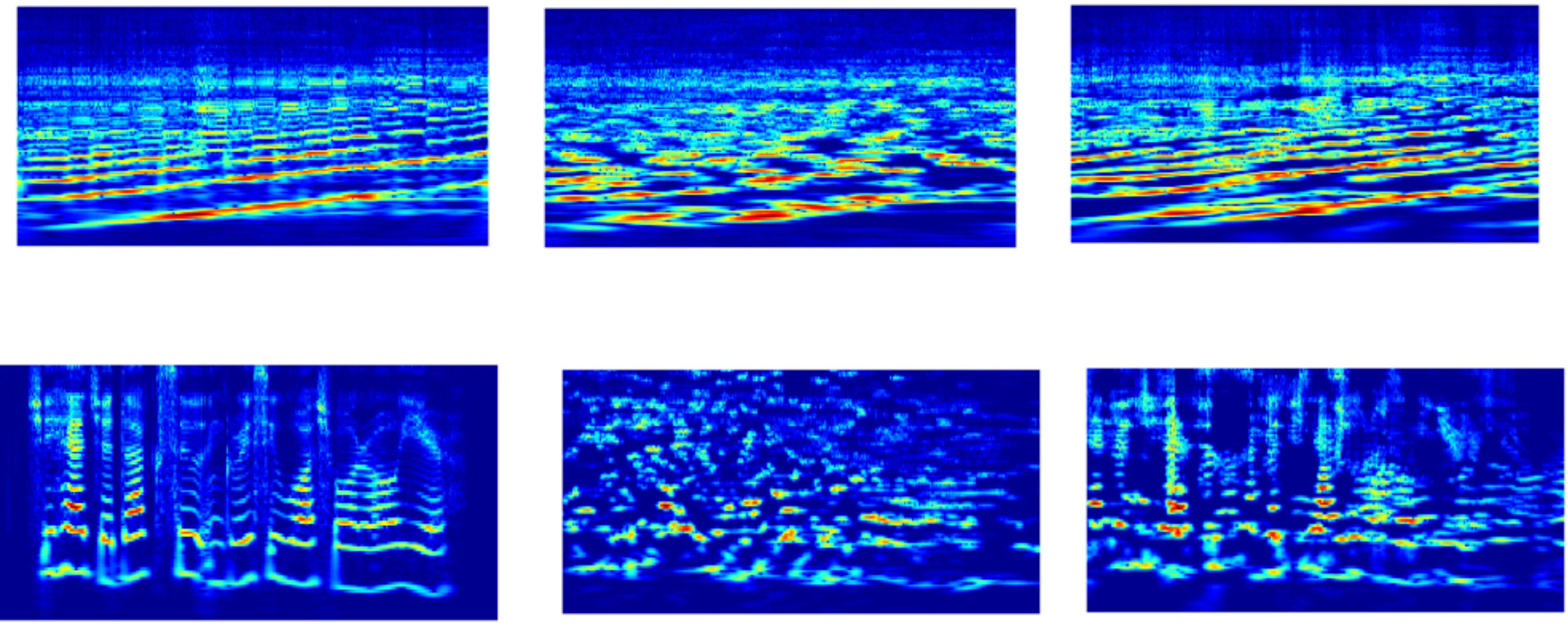
J. Anden and V. Lostanlen



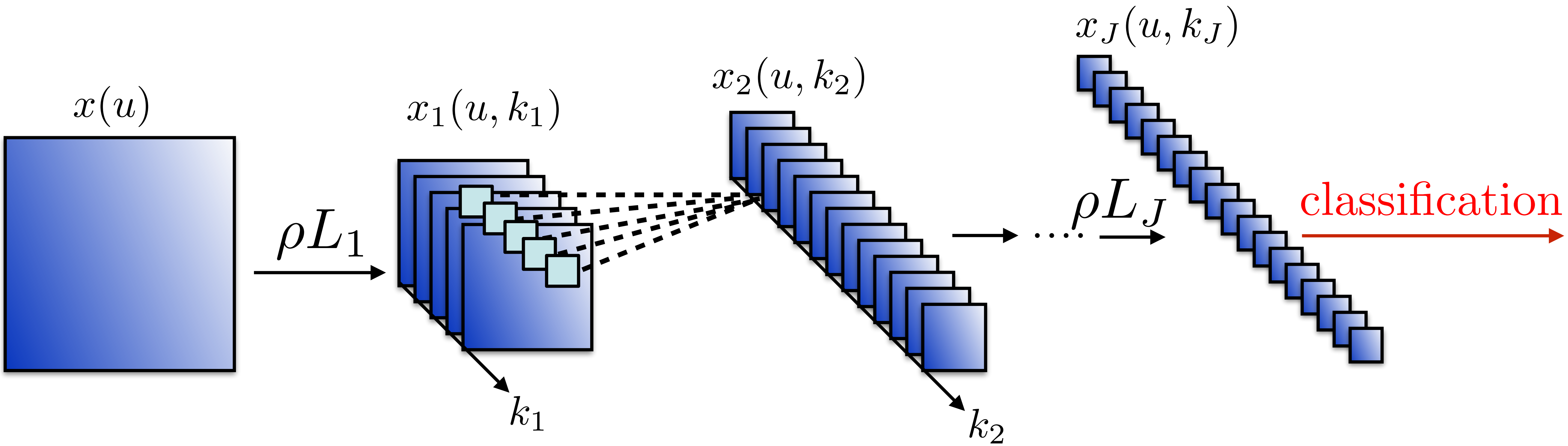
Joint Time-Frequency Scattering

J. Anden and V. Lostanl

Original | Time Scattering | Time/Freq Scattering



Part III- Supervised Learning



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

- L_j is a linear combination of convolutions and subsampling:

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

What is the role of channel connections ?

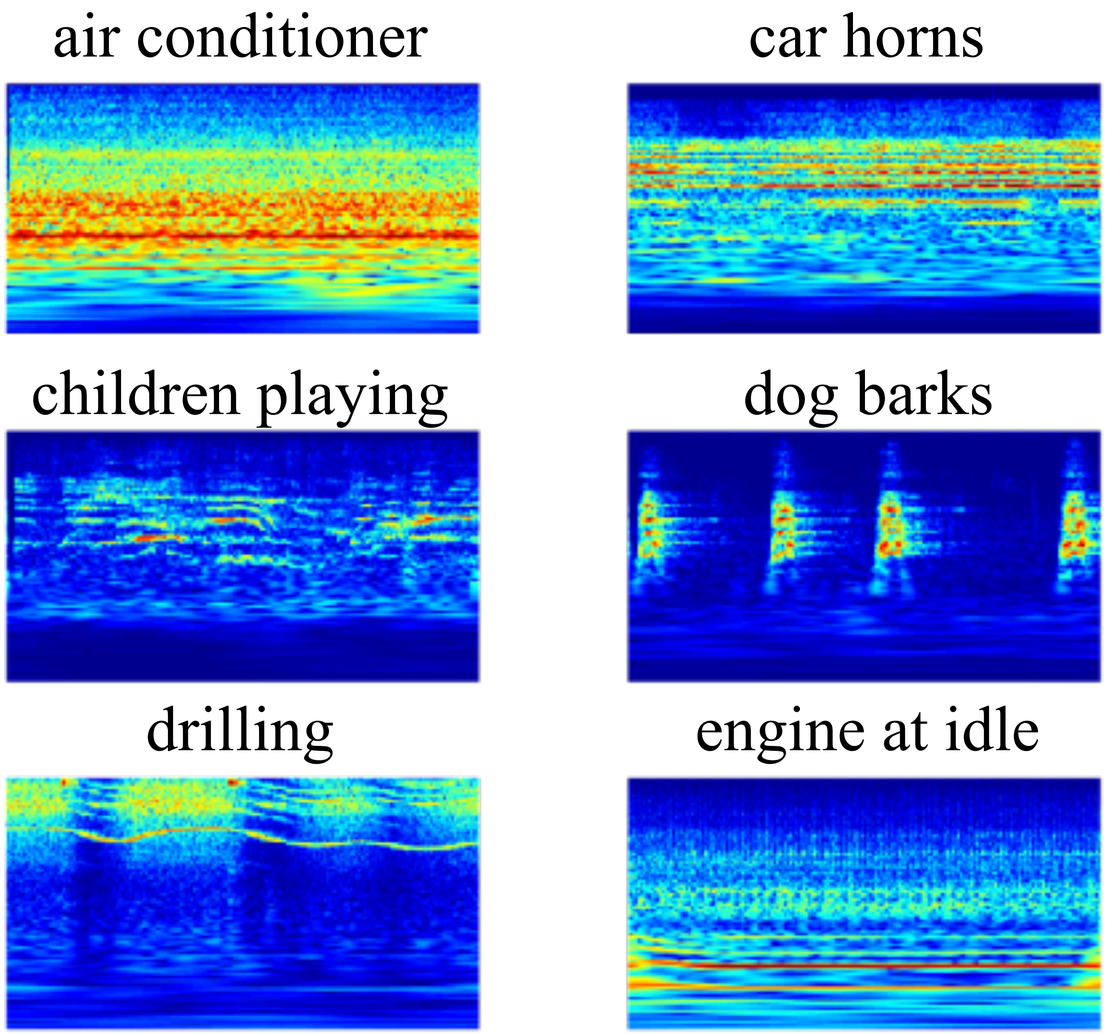
Invariant over groups of operators other than translations

Environmental Sound Classification

J. Anden and V. Lostanlen

$x \longrightarrow$ $S_J x$ $\longrightarrow$ Supervised Linear classifier $\longrightarrow y = f(x)$

No learning



UrbanSound8k: 10 classes
8k training examples

class-wise average error

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

Inverse Scattering Transform

Joan Bruna

- Given $S_J x$ we want to compute $\tilde{x}$ such that:

$$S_J \tilde{x} = \left(\begin{array}{c} \tilde{x} \star \phi_{2^J} \\ |\tilde{x} \star \psi_{\lambda_1}| \star \phi_{2^J} \\ \ldots \\ |||\tilde{x} \star \psi_{\lambda_1}| \star ..| \star \psi_{\lambda_m}| \star \phi_{2^J} \end{array} \right)_{\lambda_1,\ldots,\lambda_m} = \left(\begin{array}{c} x \star \phi_{2^J} \\ |x \star \psi_{\lambda_1}| \star \phi_{2^J} \\ \ldots \\ |||x \star \psi_{\lambda_1}| \star ..| \star \psi_{\lambda_m}| \star \phi_{2^J} \end{array} \right)_{\lambda_1,\ldots,\lambda_m} = S_J x$$

We shall use $m = 2$.

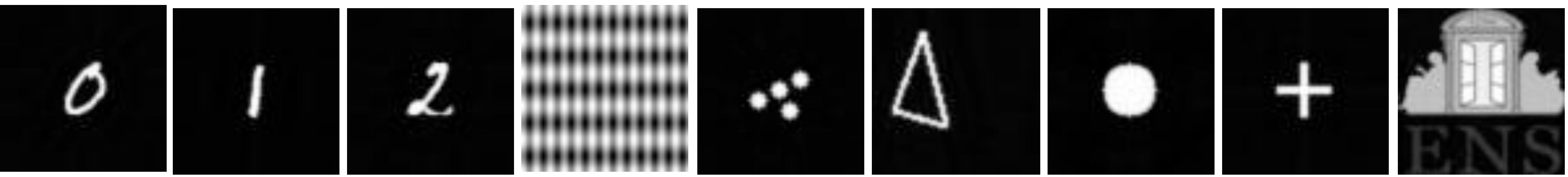
- If $x(u)$ is a Dirac, or a straight edge or a sinusoid then $\tilde{x}$ is equal to x up to a translation.

Sparse Shape Reconstruction

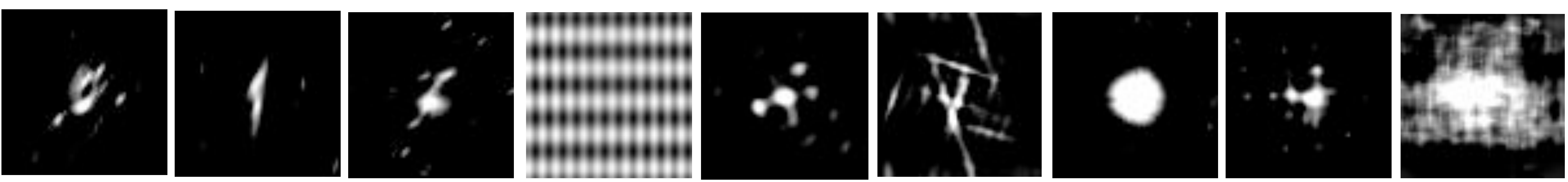
Joan Bruna

With a gradient descent algorithm:

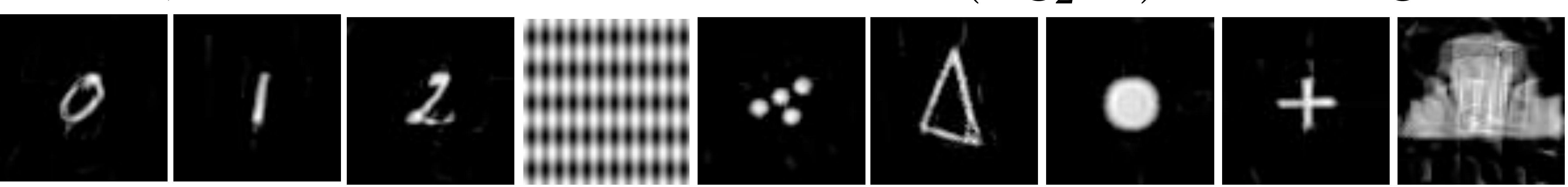
Original images of N^2 pixels:



$m = 1,\ 2^J = N$: reconstruction from $O(\log_2 N)$ scattering coeff.



$m = 2,\ 2^J = N$: reconstruction from $O(\log_2^2 N)$ scattering coeff.



Multiscale Scattering Reconstructions

Original Images
N^2 pixels

Scattering Reconstruction

$2^J = 16$
$1.4\, N^2$ coeff.

$2^J = 32$
$0.5\, N^2$ coeff.

$2^J = 64$

$2^J = 128 = N$

III- Inverse Problems

x

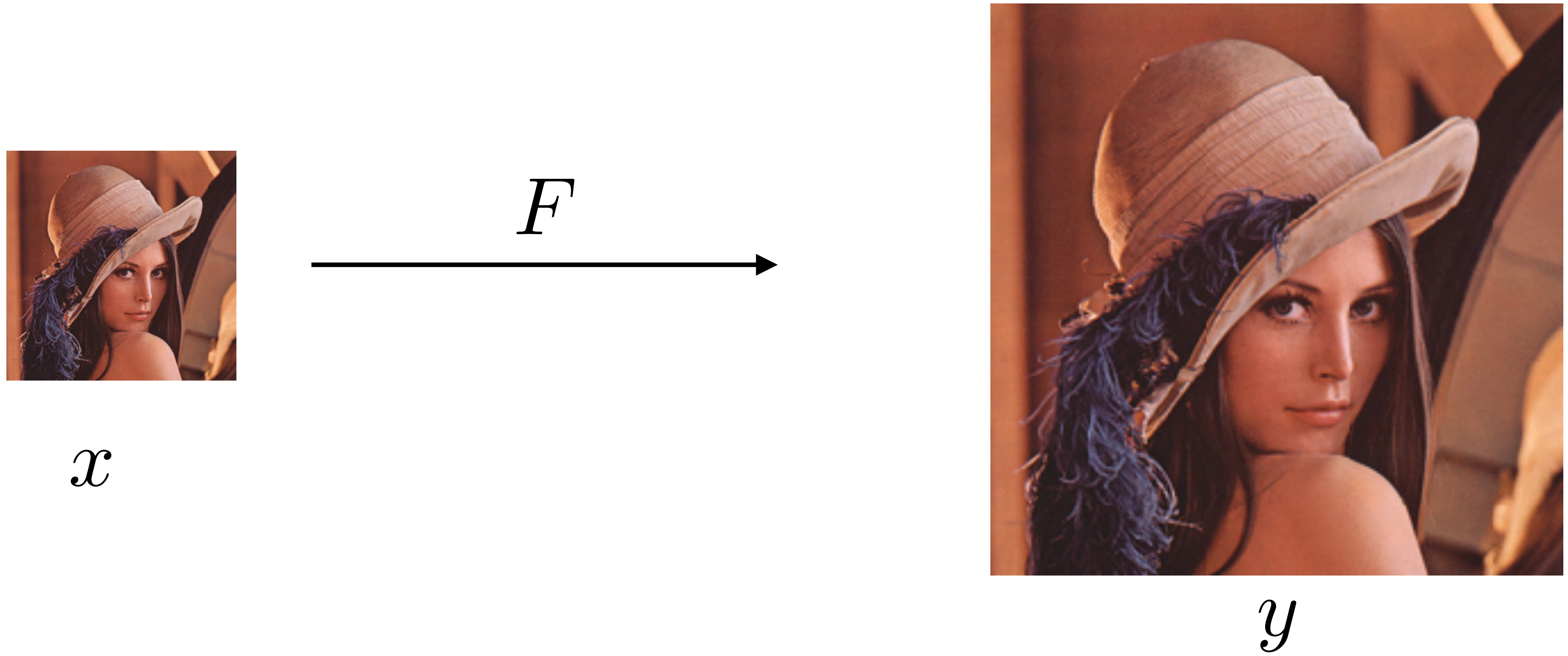
$\xrightarrow{F}$

y

- Best Linear Method: Least Squares estimate (linear interpolation):

$$\hat{y} = (\widehat{\Sigma}_x^{\dagger} \widehat{\Sigma}_{xy}) x$$

Super-Resolution



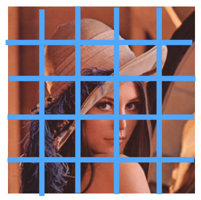
- Best Linear Method: Least Squares estimate (linear interpolation):

$$\hat{y} = (\widehat{\Sigma}_x^{\dagger} \widehat{\Sigma}_{xy})x$$

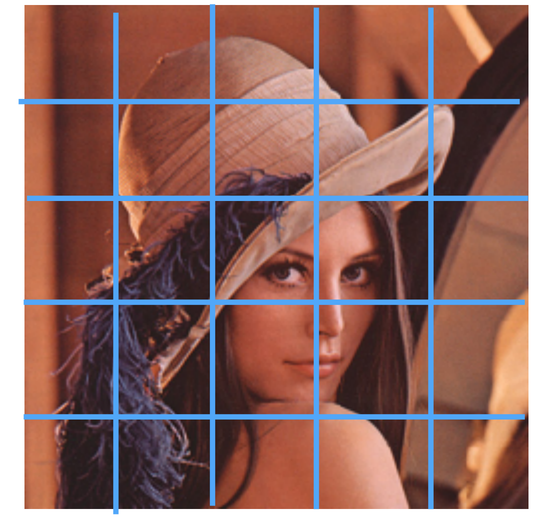
- State-of-the-art Methods:
  - Dictionary-learning Super-Resolution
  - CNN-based: Just train a CNN to regress from low-res to high-res.
  - They optimize cleverly a fundamentally unstable metric criterion:

$$\Theta^* = \arg\min_{\Theta} \sum_i \|F(x_i, \Theta) - y_i\|^2 \quad , \quad \hat{y} = F(x, \Theta^*)$$

Scattering Super-Resolution



x



y

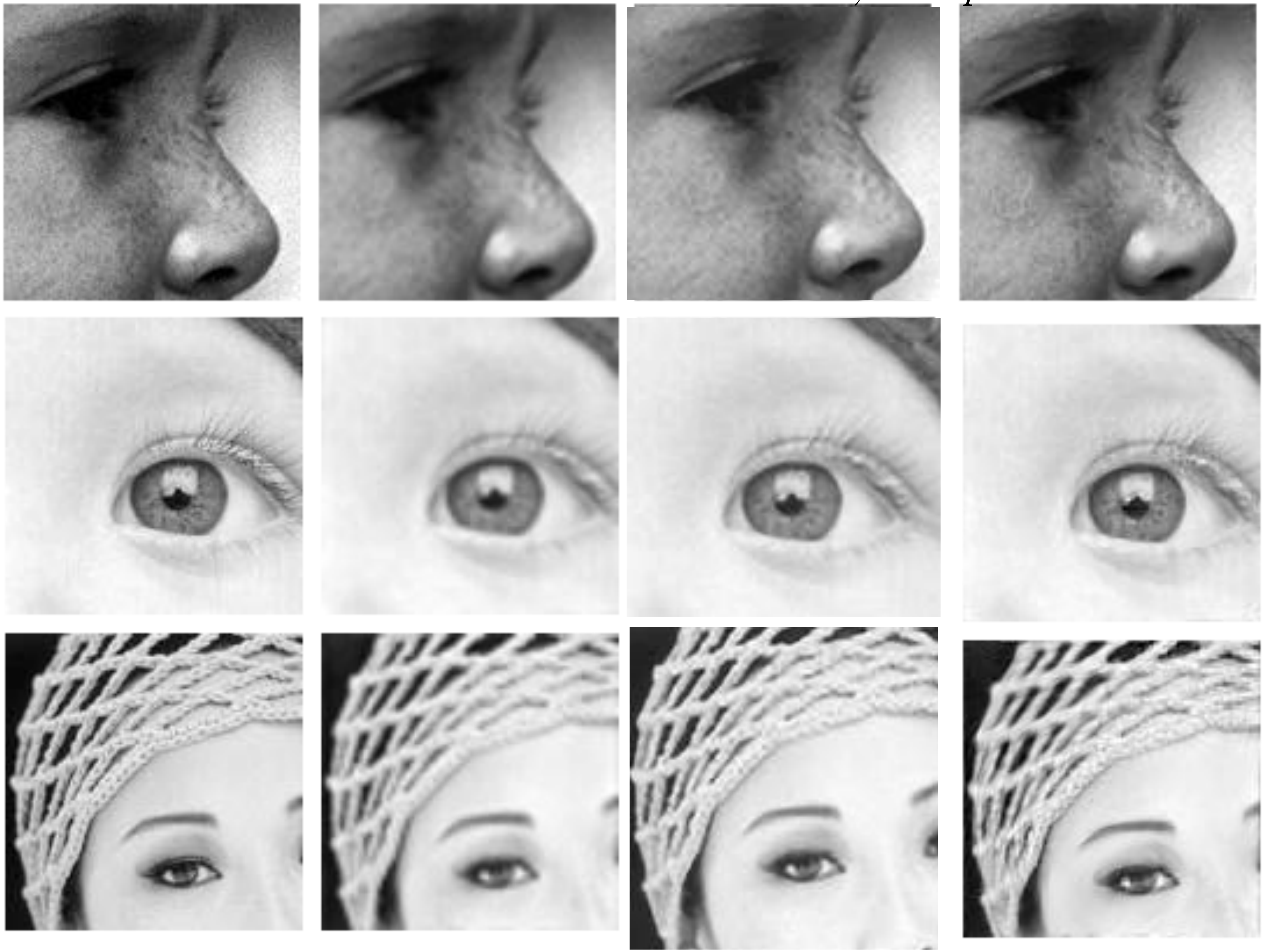
$$S_{L,J}\,x \xrightarrow{\quad F \quad} S_{L-\alpha,J}\,x$$

$$S_{L,J}x = \begin{pmatrix} x \star \phi_{2^J}(u) \\ |x \star \psi_{j_1,k_1}| \star \phi_{2^J}(u) \\ ||x \star \psi_{j_1,k_1}| \star \psi_{j_2,k_2}| \star \phi_{2^J}(u) \end{pmatrix}_{L \leq j_1, j_2 \leq J}$$

- Linear estimation in the scattering domain
- No phase estimation: potentially worst PSNR
- Good image quality because of deformation stability

Super-Resolution Results

J. Bruna, P. Sprechmann



Original | Linear Estimate | state-of-the-art | Scattering

Super-Resolution Results

J. Bruna, P. Sprechmann

Original | Best Linear Estimate | state-of-the-art | Scattering Estimate

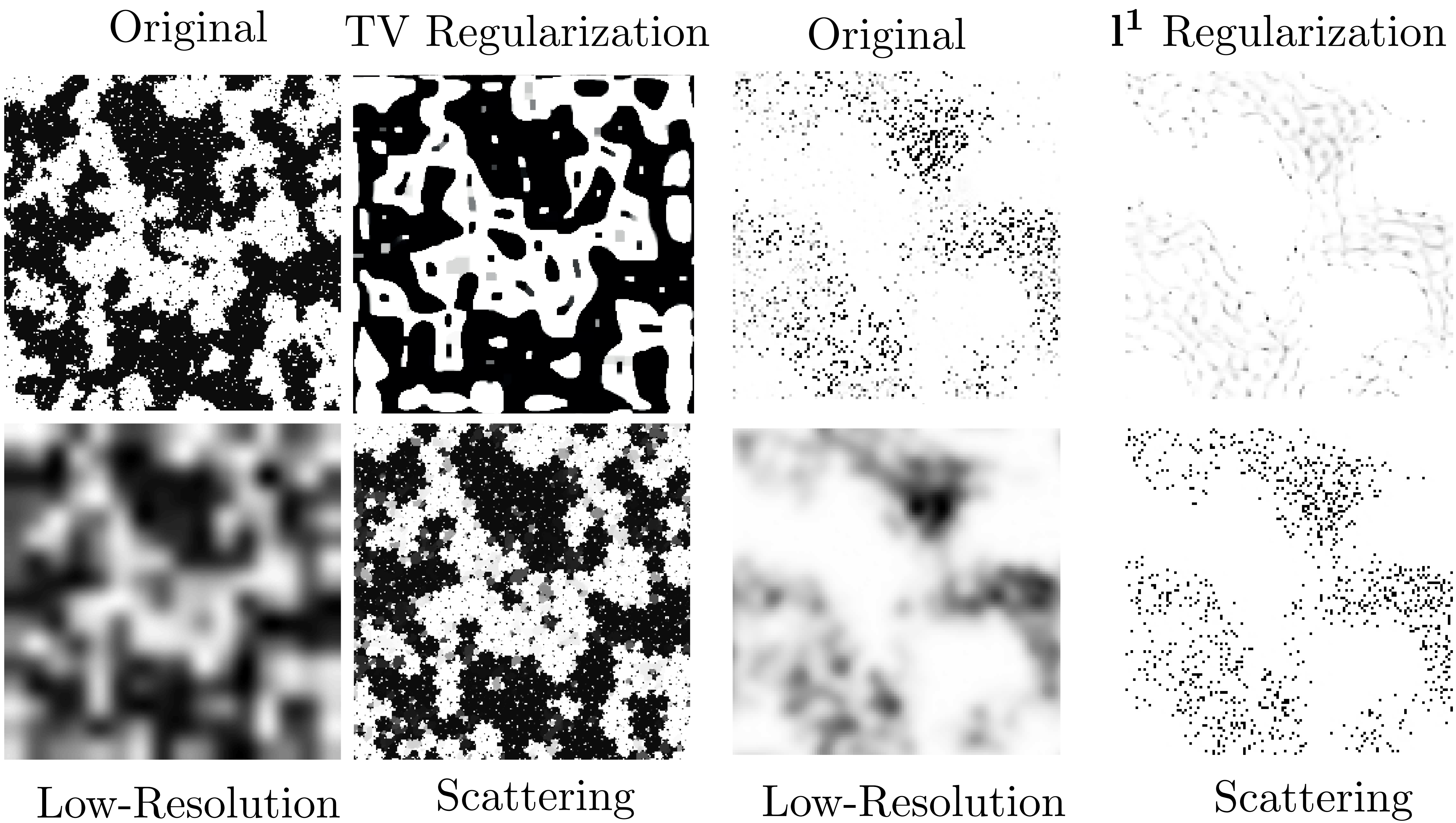
Super-Resolution Results

J. Bruna, P. Sprechmann

Original | Best Linear Estimate | state-of-the-art | Scattering Estimate

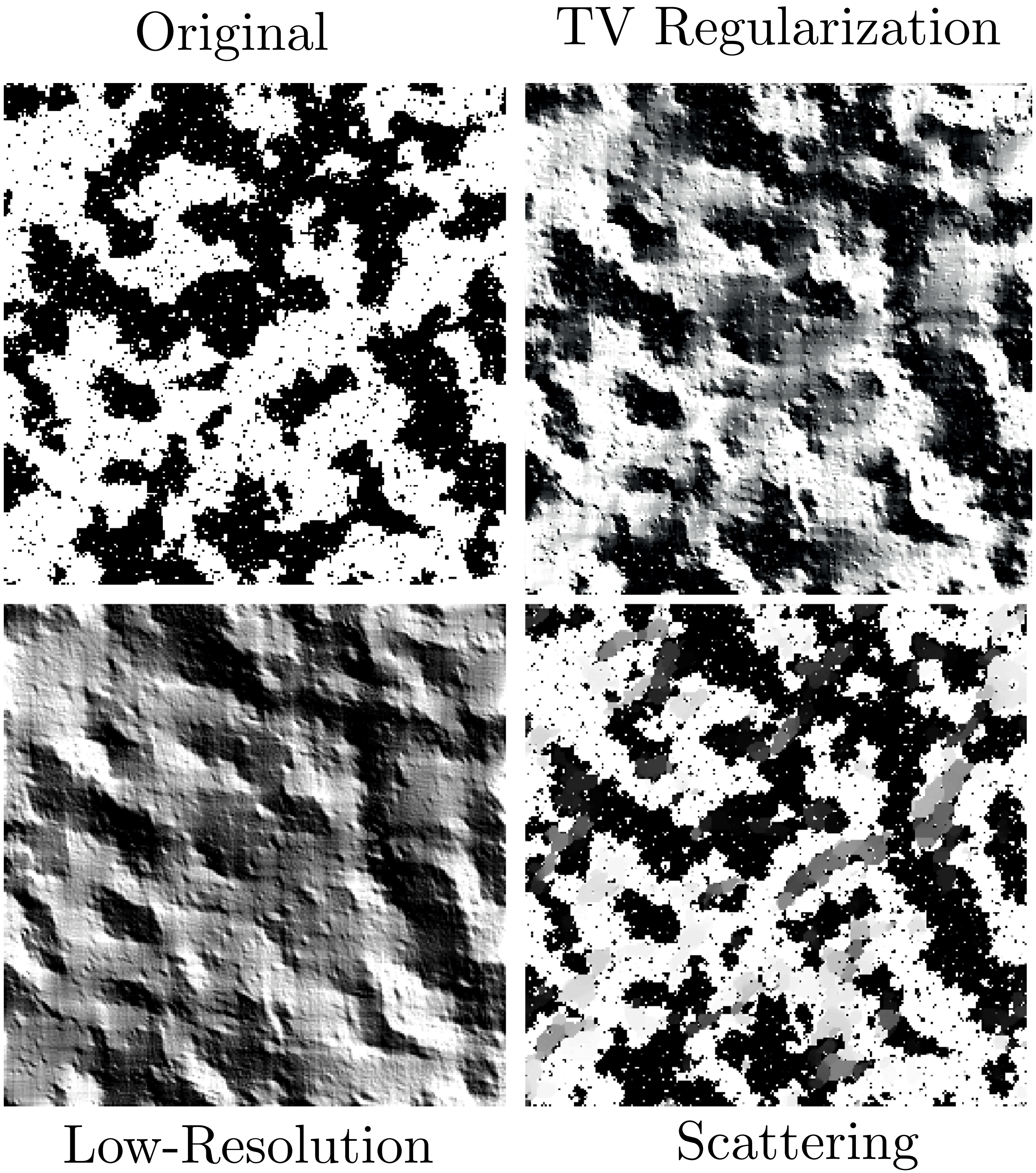
Super-Resolution Results

I. Dokmanic, J. Bruna, M. De Hoop



Tomography Results

I. Dokmanic, J. Bruna, M. De Hoop



Conclusions

- Deep convolutional networks have spectacular high-dimensional and generic approximation capabilities.

- New stochastic models of images for inverse problems.

- Outstanding mathematical problem to understand deep nets:
  - How to learn representations for inverse problems ?

(Not) Understanding Deep Convolutional Networks, arXiv 2016.