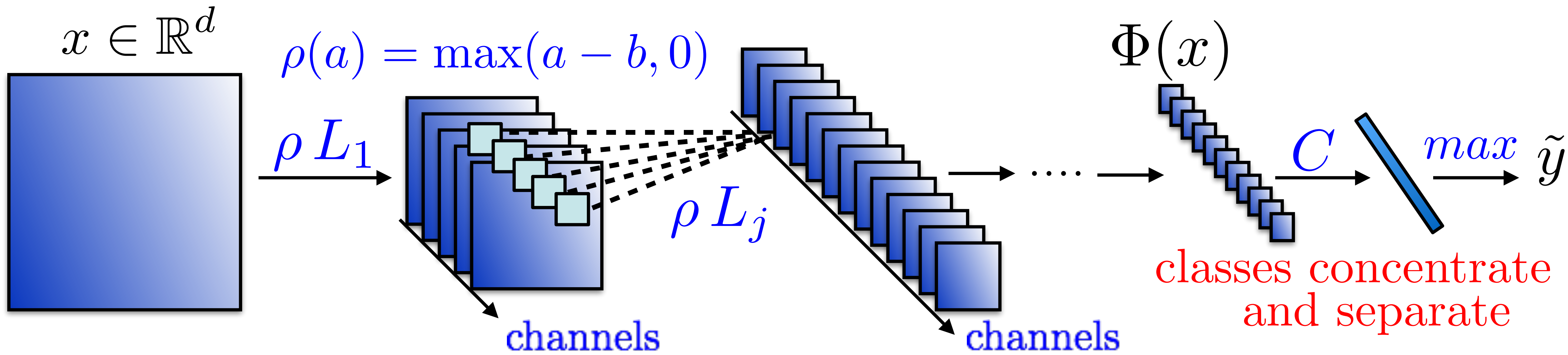
Concentration and Separation in Deep Networks

Stéphane Mallat

Collège de France
École Normale Supérieure, Paris
Flatiron Institute, New York

Deep Convolutional Networks

Classification with deep convolutional networks:



L_j: spatial convolutions and linear combination of channels

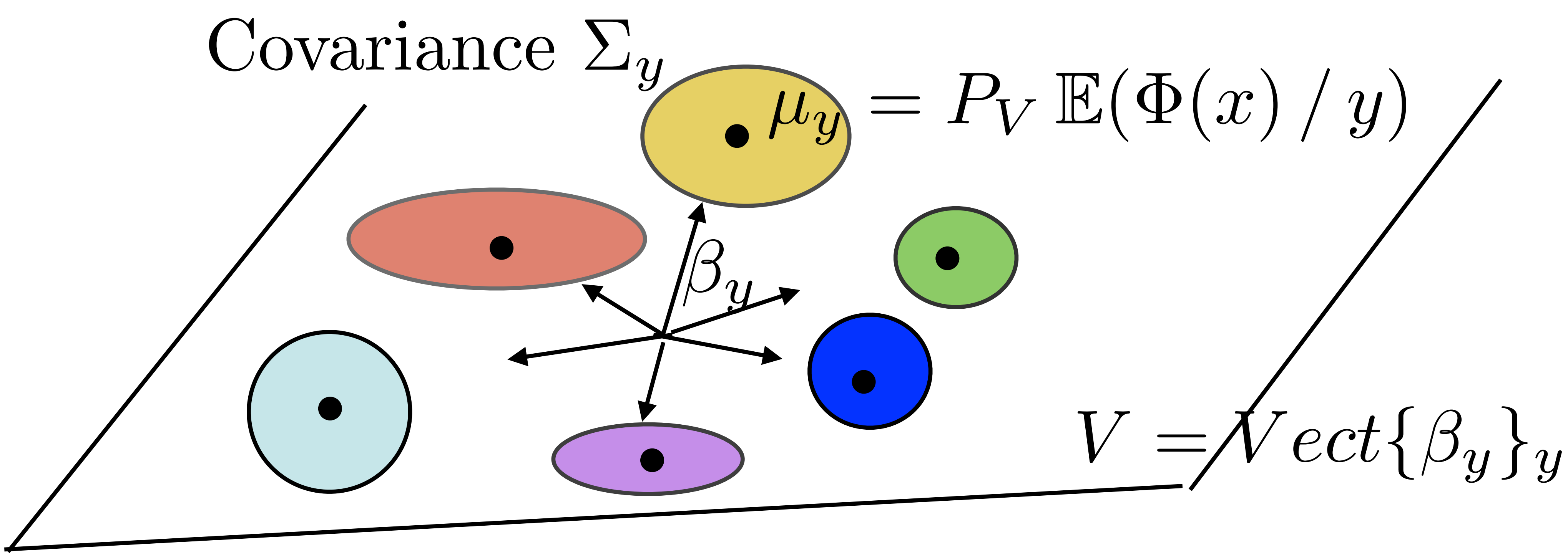
Exceptional results for classification of *images, sounds, language, regressions in physics, signal and image generation...* : not understood

- Issues of robustness and validation in applications: *transport, medecine, sciences...*

- Opportunity for new maths

Linear Classification From $\Phi(x)$

Linear classifier: $\tilde{y} = \arg_y \max \langle \Phi(x), \beta_y \rangle + \alpha_y$



- $\Phi(x)$ must have separated class means μ_y in V

Fisher Ratio: $\mathrm{Trace}(\Sigma_W^{-1} \, \Sigma_B) \xrightarrow[\text{training}]{\text{Neural collapse}} \infty$

V. Papyan
X.Y. Han
D. Donoho

with $\Sigma_B = Ave_y \, (\mu_y - \overline{\mu})(\mu_y - \overline{\mu})^T$ and $\Sigma_W = Ave_y \, \Sigma_y$.

What mechanism leads to this concentration/separation ?

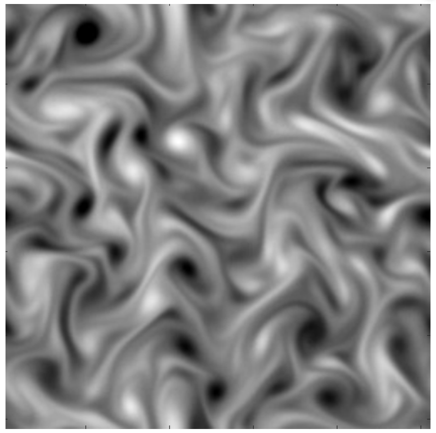
Overview

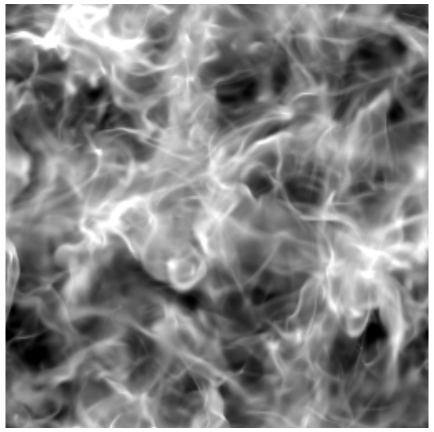
I- Frame separation and contraction in 2-layer nets

II- Concentration and Separation in Statistical Physics:

- Models of non-Gaussian processes

Turbulences:





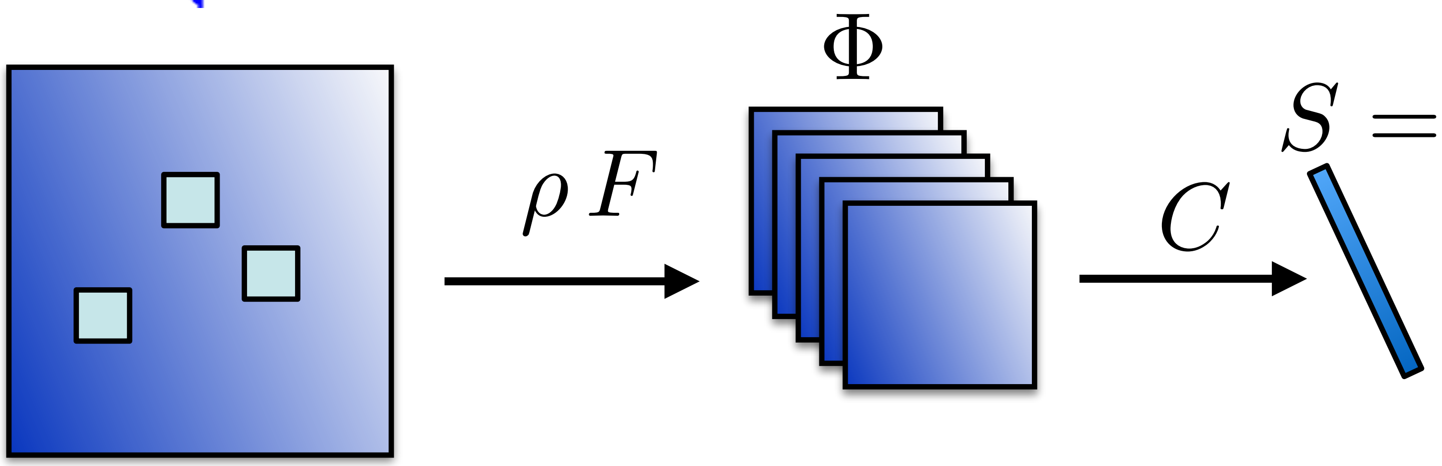
- Wavelet separation and ReLU: scales, orientations and phases

II- Image classification by deep concentration and separation:

- Deep multi scale scattering from priors without learning
- Learning along channels only

Tight Frame Separation & Contraction

John Zarka, Florentin Guth



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

$S = C\,\rho\, F$: 2 layer network with no bias

$$C\Phi(x) = \Big(\langle \Phi(x), \beta_y \rangle\Big)_y$$

$$\Phi(x) = \rho\, F x = \Big(\rho(\langle x, w_n \rangle)\Big)_{n \leq p} \text{ with } p \geq d.$$

Tight frame: $F^T F = Id$ separates along each w_n

contraction: $|\rho(a) - \rho(a')| \leq |a - a'|$

$$\Rightarrow \quad \|\Phi(x) - \Phi(x')\| \leq \|x - x'\| \ : \text{ contraction}$$

Separation and contractions with threshold $t \geq 0$:

Soft-Thresh. $\rho(a) = \mathrm{sign}(a)\, \max(|a| - t, 0)$ *shrinks amplitude* for noise removal

"Stein shrinking estimation"

ReLu $\rho(a) = \max(a - t, 0)$ *shrinks amplitude* *separates sign/phases*

Separation and Contraction

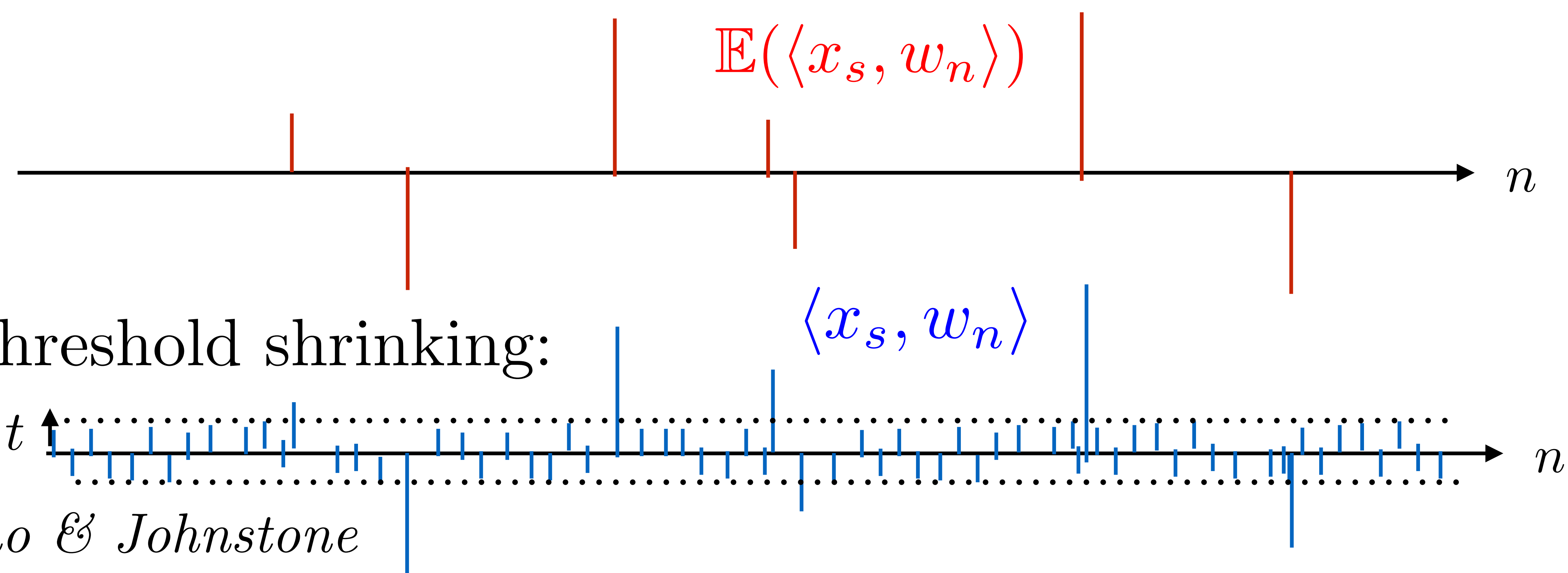
- Let x be a Gaussian mixture of zero mean x_s with covariance Σ_s

  A ReLu $\rho(a) = \max(a, 0)$ can separate the means:

$$(2\pi)^{1/2}\,\mathbb{E}(\rho F x_s) = \mathrm{diag}(F\,\Sigma_s F^T)^{1/2} = \mathbb{E}(|\langle x_s, w_n\rangle|^2)^{1/2}$$

- A soft-thresholding reduces variances and nearly preserves the mean:

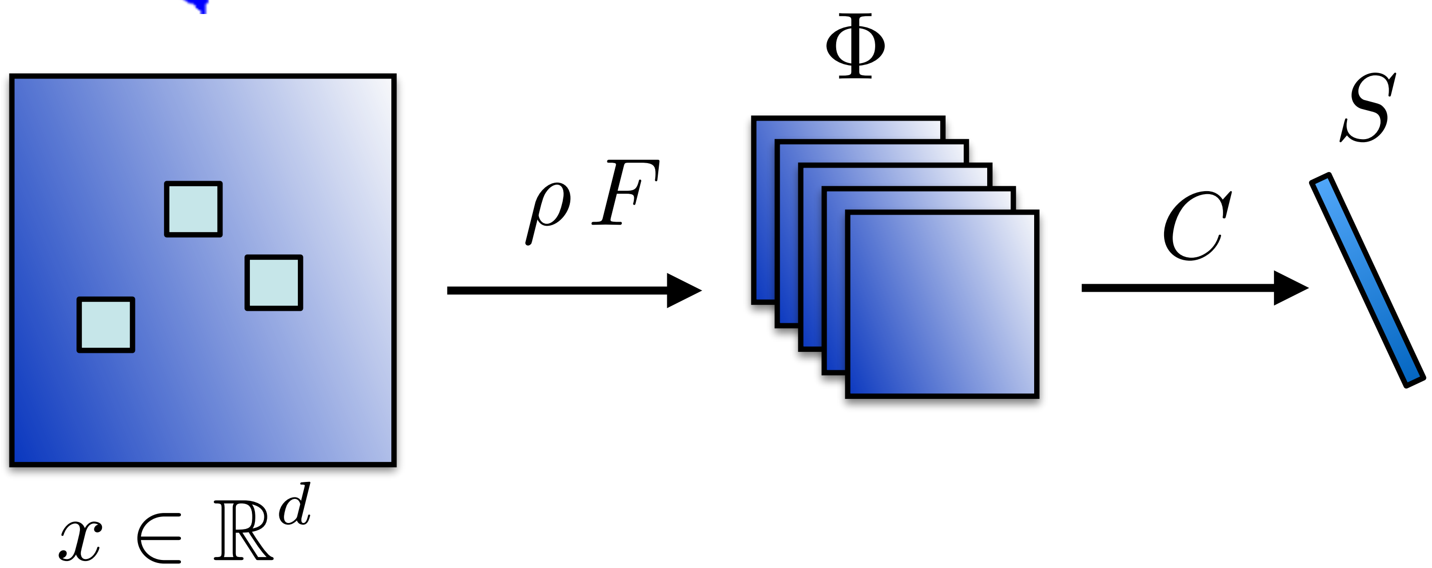
  if $\mu_s = \mathbb{E}(x_s)$ has a sparse representation in F:



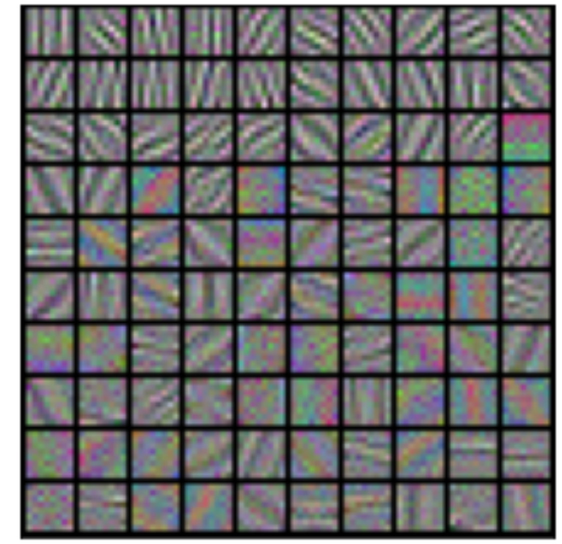
Soft-threshold shrinking:

Donoho & Johnstone

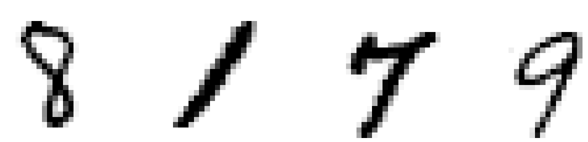
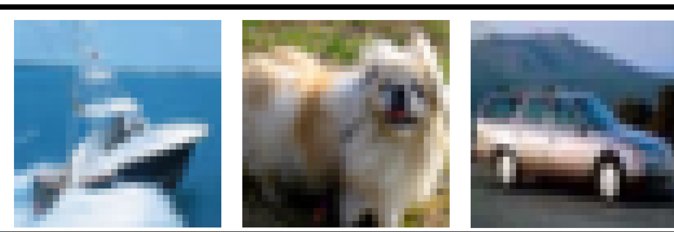
Tight Frame Contraction



Filters of F for CIFAR



- SGD optimisation

		$\Phi(x)$	x	Soft $\rho F x$	ReLu $\rho F x$
MNIST		Error	7.4%	1.4%	1.4%
		Fisher	20	60	60
CIFAR		Error	60%	39%	28%
		Fisher	7	12	15

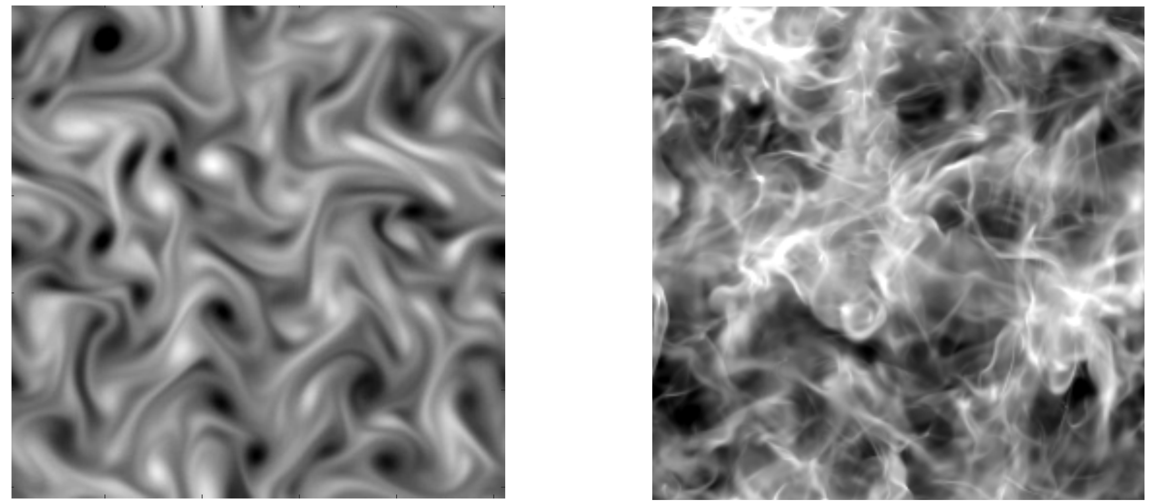
- A soft-thresholding ρ can reduce within class variance and preserve class means μ_y if Fx is sufficiently sparse. *(Donoho Johnstone)* A ReLu ρ also modifies class means.

Do we need to learn the tight frame F ?

II- Statistical Physics

- Characterize a random process and generate samples

Turbulences:



- A *one versus all classification problem*: discriminate typical realisations from all other types of signals.

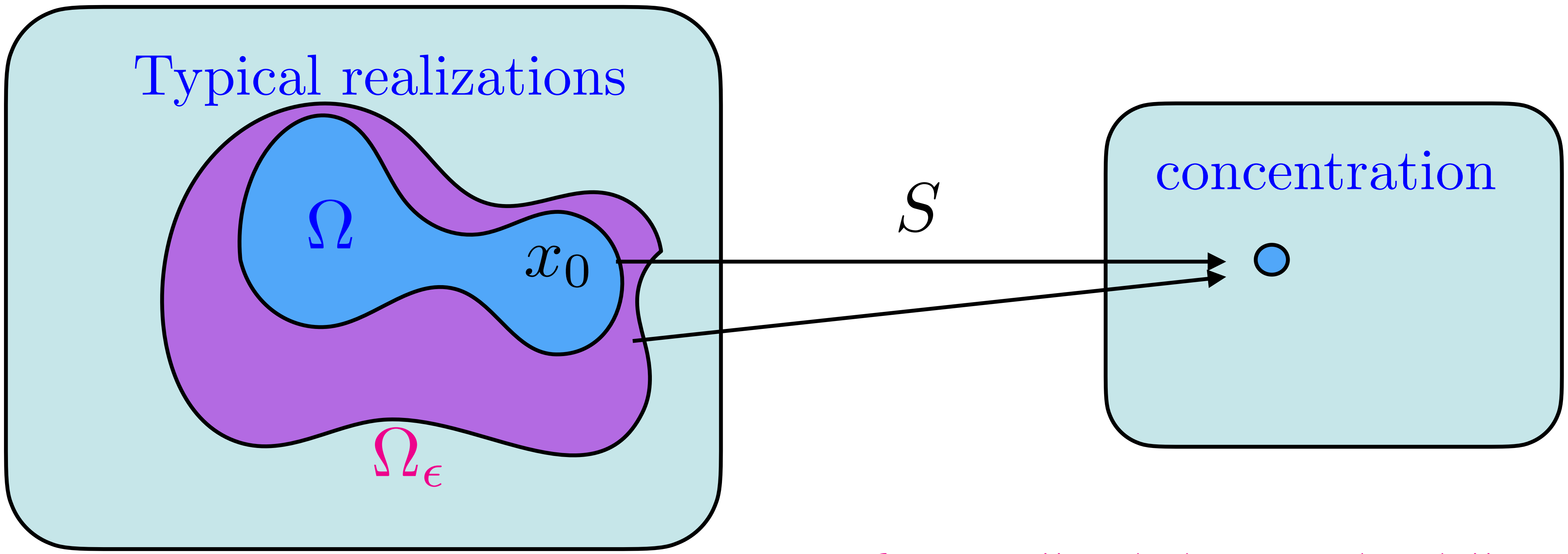
- Characterized through **concentration** of sufficient statistics which **separate** from all others with high probability.

One Versus All: Statistical Physics

Vector of statistics $S(x)$: observable

Concentration: $\mathrm{Prob}_p\Big(\|S(x) - \mathbb{E}_p(S(x))\| > \epsilon\Big) \underset{d\to\infty}{\longrightarrow} 0$

$\Rightarrow$ a realisation x_0 satisfies $S(x_0) \approx \mathbb{E}_p(S(x))$ with high proba.



Microcanonial ensemble: $\Omega_\epsilon = \{x \ : \ \|S(x) - S(x_0)\| \le \epsilon\}$

Maximum entropy model $\tilde{p}$ supported in Ω_ϵ is uniform.

Sufficient model if $\Omega \approx \Omega_\epsilon$: what statistics S ?

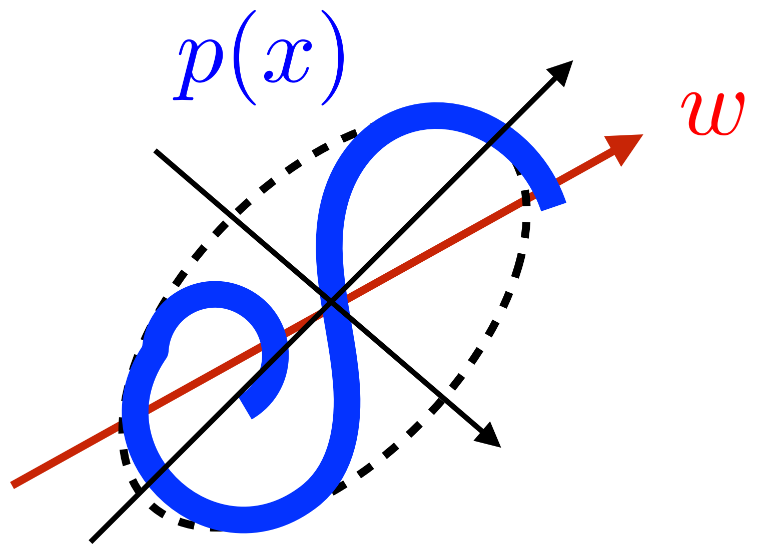
Stat. for Gaussian Stationary Models

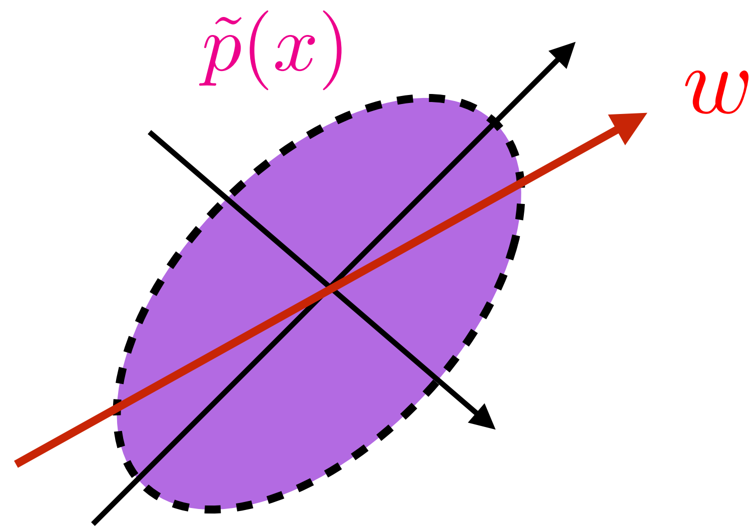
Symmetry prior: $p(x)$ is translation invariant

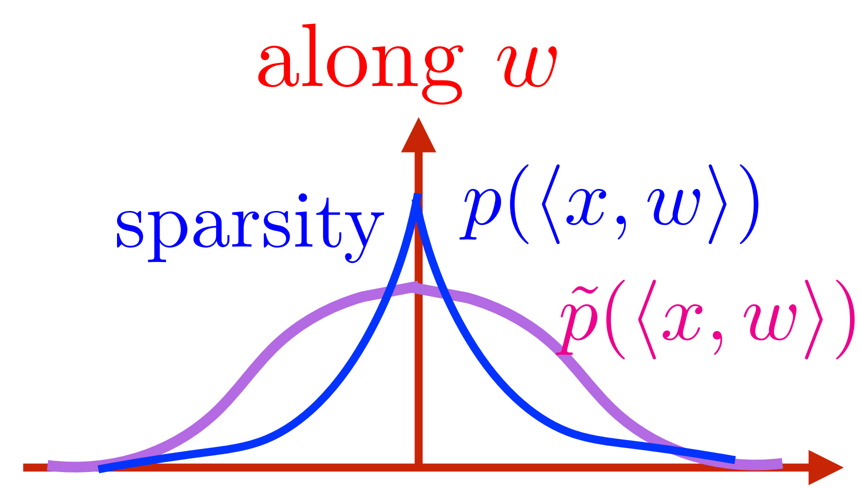
$$S(x) = \Big(d^{-1} \sum_u x(u)\, x(u-\tau) \Big)_\tau$$

empirical covariance
concentrates by spatial averaging

Maximum entropy model $\tilde p$ asymptotically Gaussian: how good ?

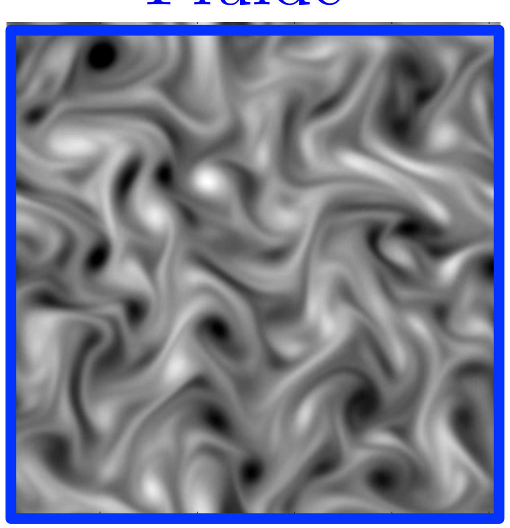




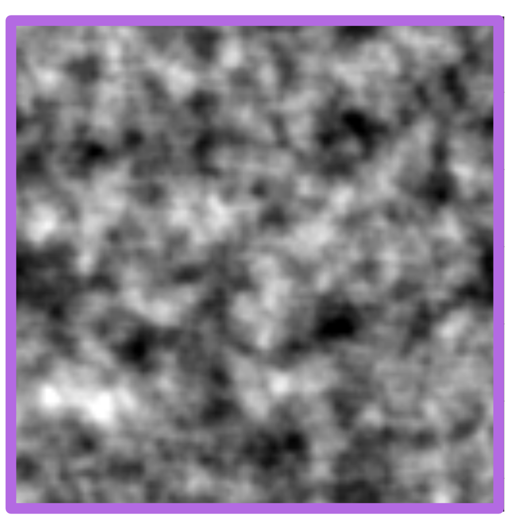


PCA basis: Fourier $\Rightarrow$ Harmonic Analysis

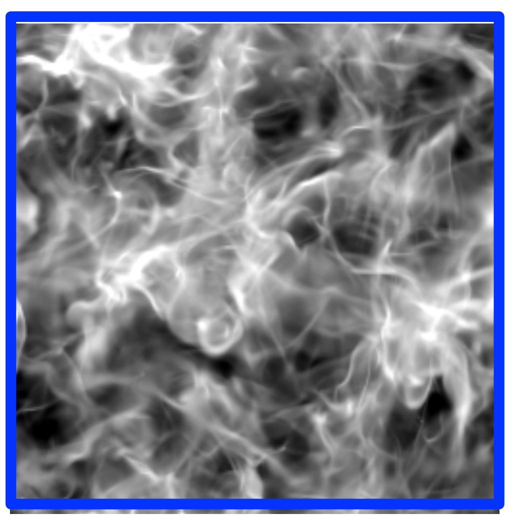
Fluide



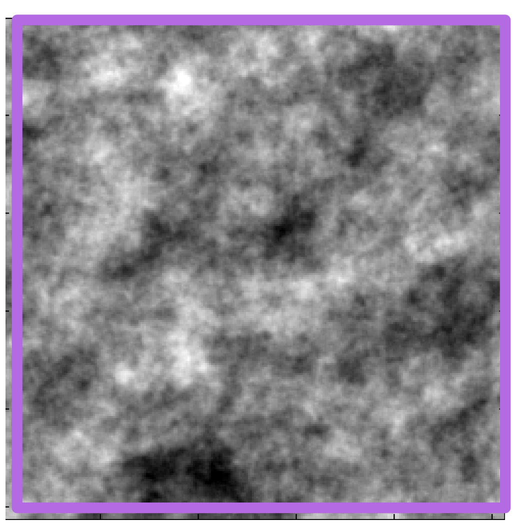
Gaussien



Gaz



Gaussien



Beyond Gaussian Models

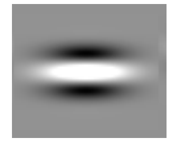
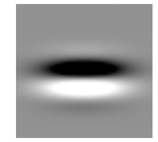
- No reproduction of "coherent structures" because the statistics do not enforce dependancies across frequencies

- Need to capture "patterns" with sparse representations.

- Scale separations with wavelets

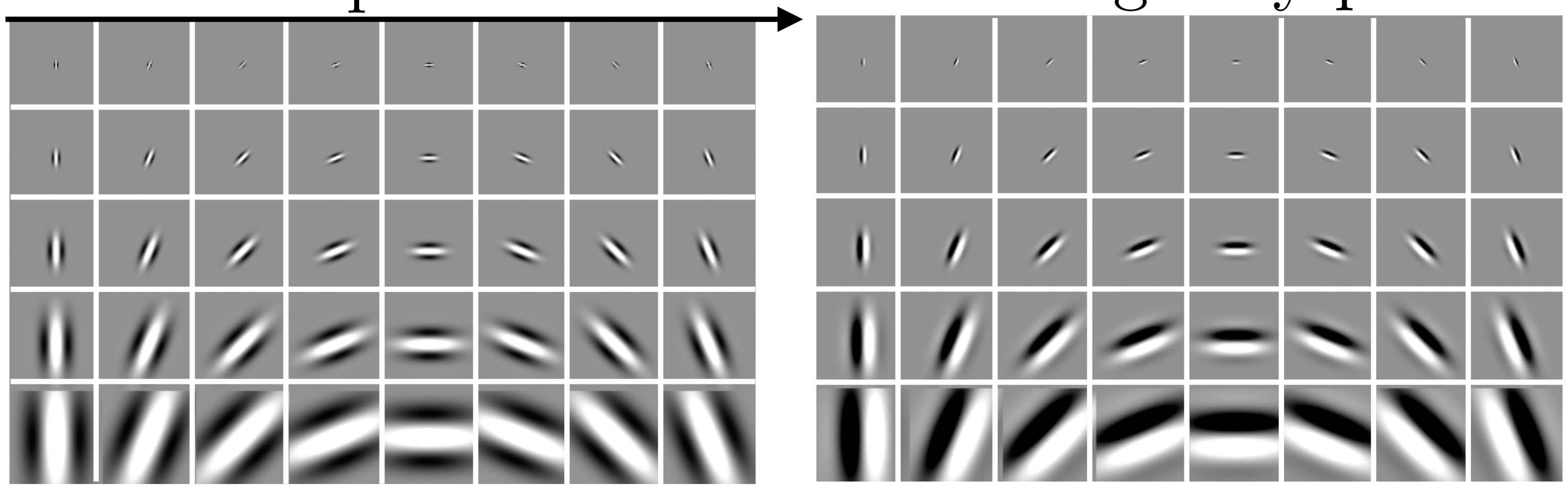
- Role of ReLU to capture scale dependancies

Scale separation with Wavelets

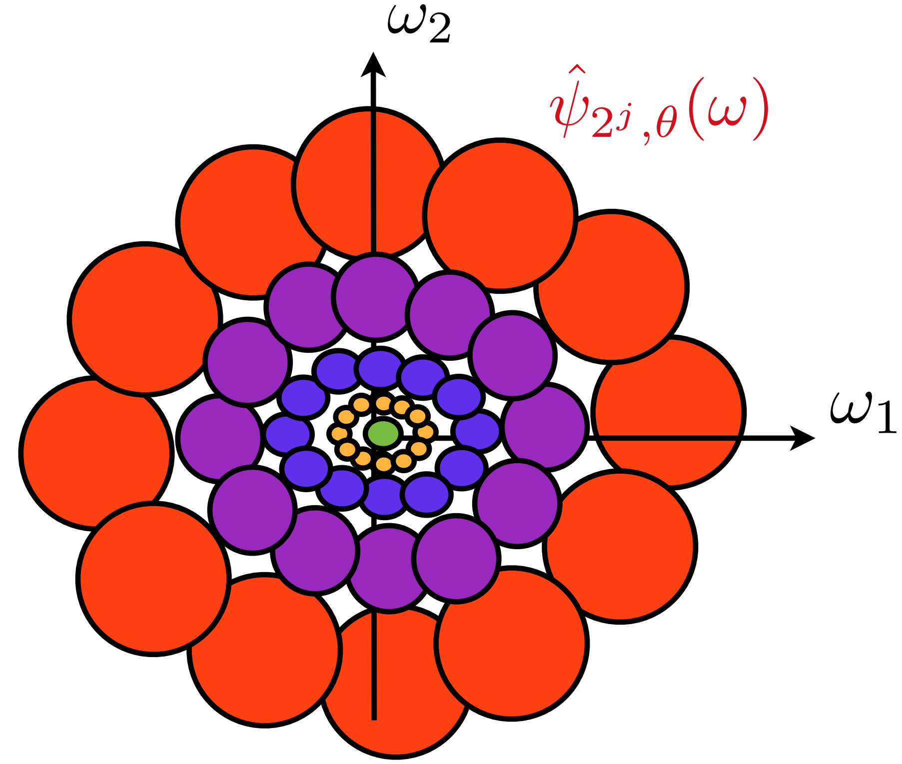
- Wavelet filter $\psi(u)$:  $+\,i$  complex

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

real parts θ imaginary parts



Fourier



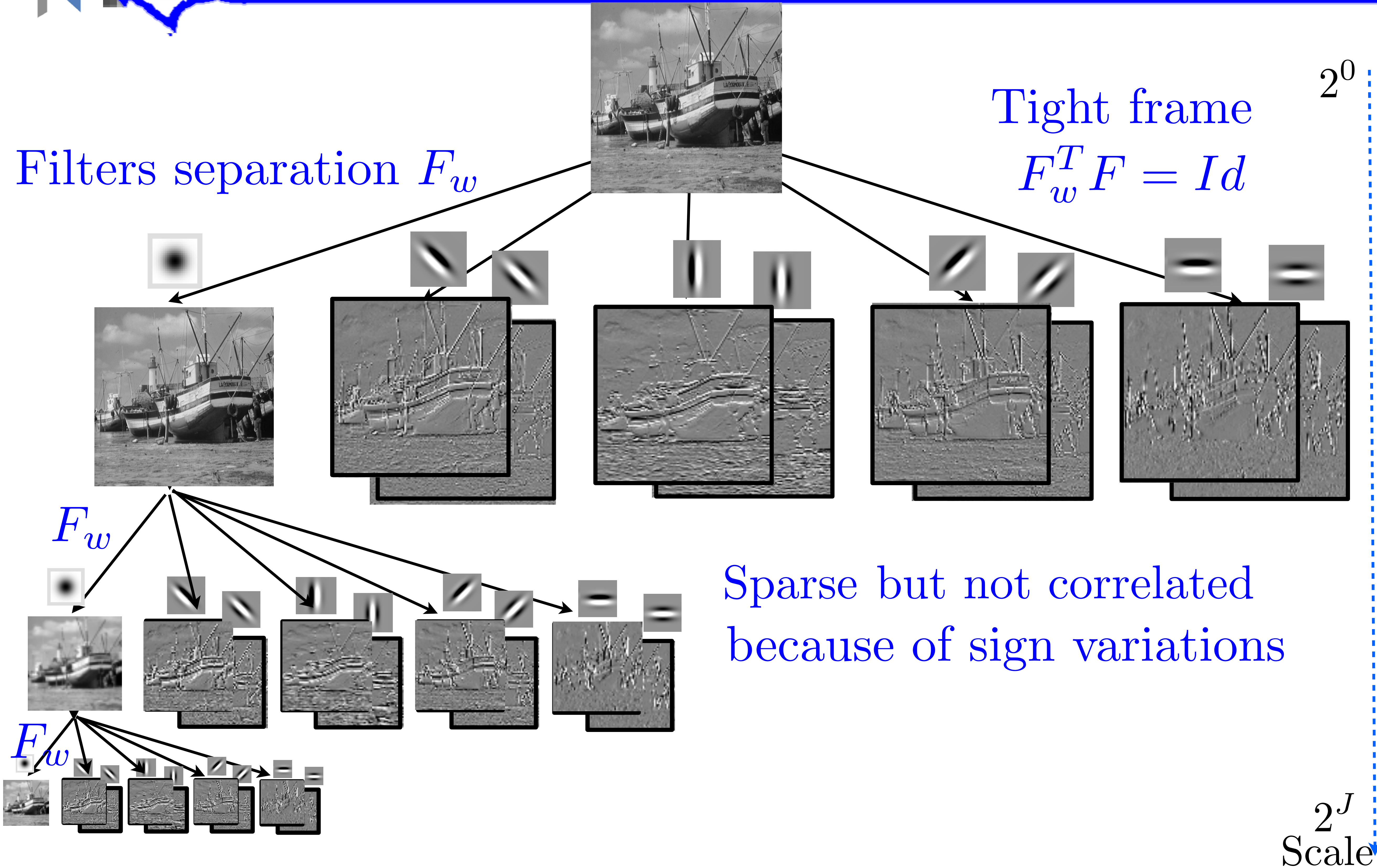
- Wavelet tight frame:

$$Wx(u,\lambda) = x \star \psi_\lambda(u) \xrightarrow{\text{Fourier}} \widehat{x \star \psi_\lambda}(\omega) = \hat{x}(\omega)\,\hat{\psi}_\lambda(\omega)$$

- Not correlated across "channels" if x is stationary:

$$\mathbb{E}\Big(Wx(u,\lambda)\,Wx(u,\lambda')\Big) \approx 0 \ \text{ if } \ \lambda \neq \lambda'$$

Wavelet Filter Bank Algorithm



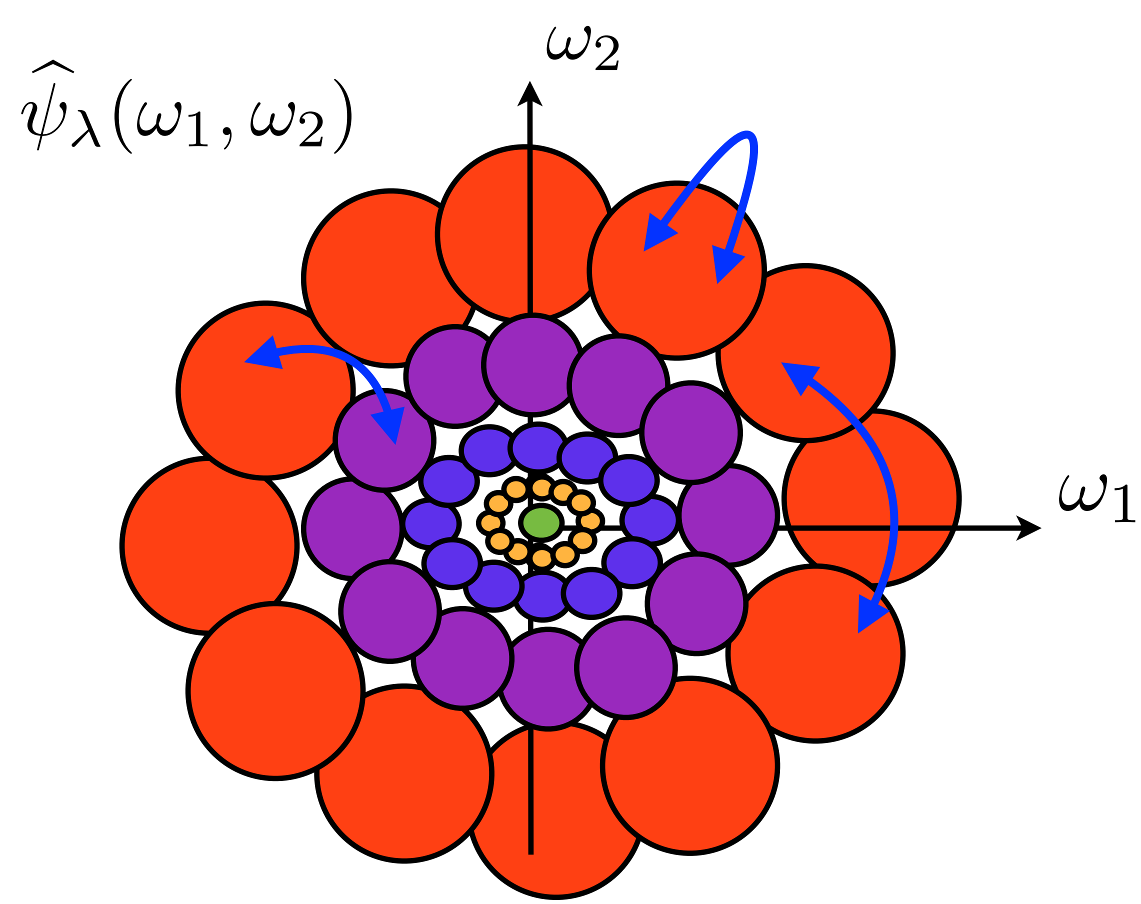
How to capture dependance across scales, angles, phases channels ?

Multiscale Correlation Graph

Correlations across scales/orientations/phases $\lambda = (2^j, \theta), \alpha$ created by a ReLu $\rho(a) = \max(a, 0)$ which separate phases

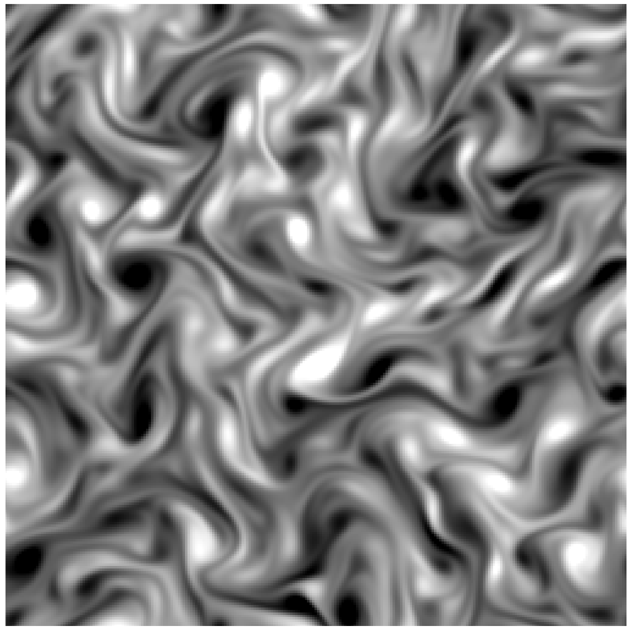
$$S(x) = \Big(\sum_u \rho(x \star \psi_{\alpha,\lambda})\, \rho(x \star \psi_{\alpha',\lambda'}) \Big)_{\substack{\alpha,\lambda \\ \alpha',\lambda'}}$$

Concentration by spatial averaging: dimension $O(\log^2 d)$

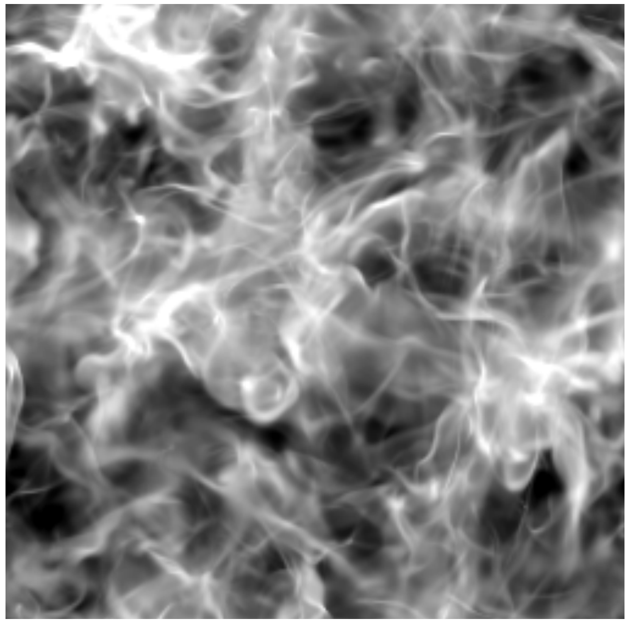


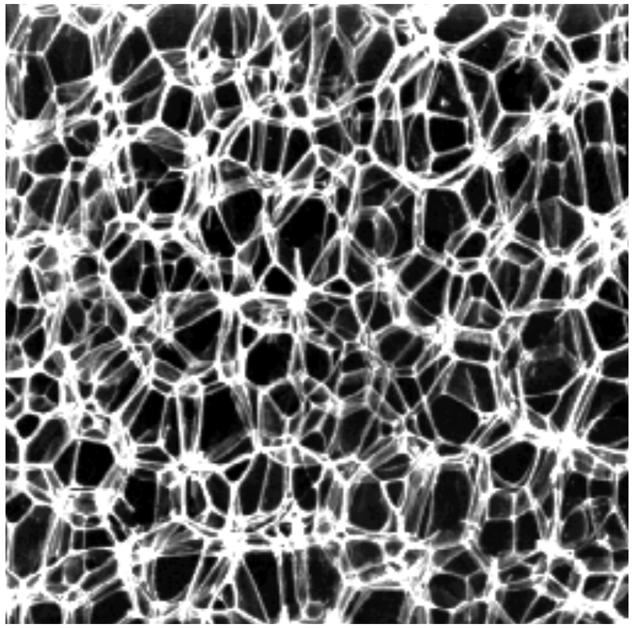
Models of Stationary Processes

x_0

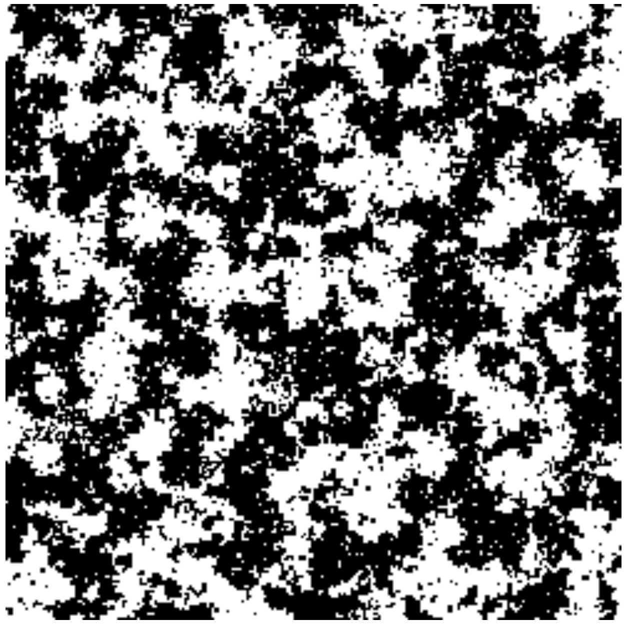


Sixin Zhang





Ising-critical



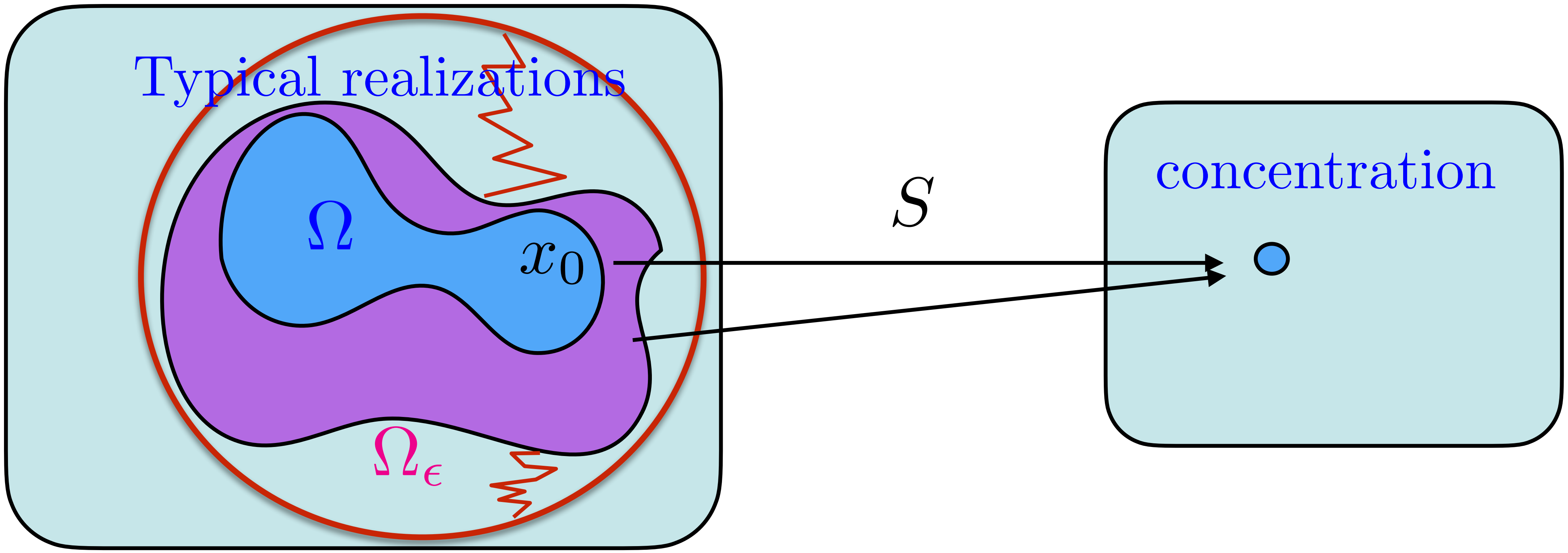
Correlations across scales/orientations/phases $\lambda = (2^j, \theta), \alpha$

$$S(x) = \Big(\sum_u \rho(x \star \psi_{\alpha,\lambda})\, \rho(x \star \psi_{\alpha',\lambda'}) \Big)_{\substack{\alpha,\lambda \\ \alpha',\lambda'}}$$

Maximum entropy models conditioned by $S(x_0)$

Generation by Gradient Descent

Concentration: $\text{Prob}_p\Big(\|S(x) - \mathbb{E}_p(S(x))\| > \epsilon\Big) \underset{d\to\infty}{\longrightarrow} 0$



Microcanonial ensemble: $\Omega_\epsilon = \{x \ : \ \|S(x) - S(x_0)\| \leq \epsilon\}$

Maximum entropy model $\tilde{p}$ supported in Ω_ϵ is uniform.

Generation by sampling $\tilde{p}$: SGD on $\|S(x) - S(x_0)\|$ from white noise

Transport of measure which converges *(J. Bruna)*

Not maximum entropy but same unitary invariants as S

Sampling from Max Entropy Model

S. Zhang, E. Allys, T. Marchand, S. Ho, F. Levrier, F. Boulanger

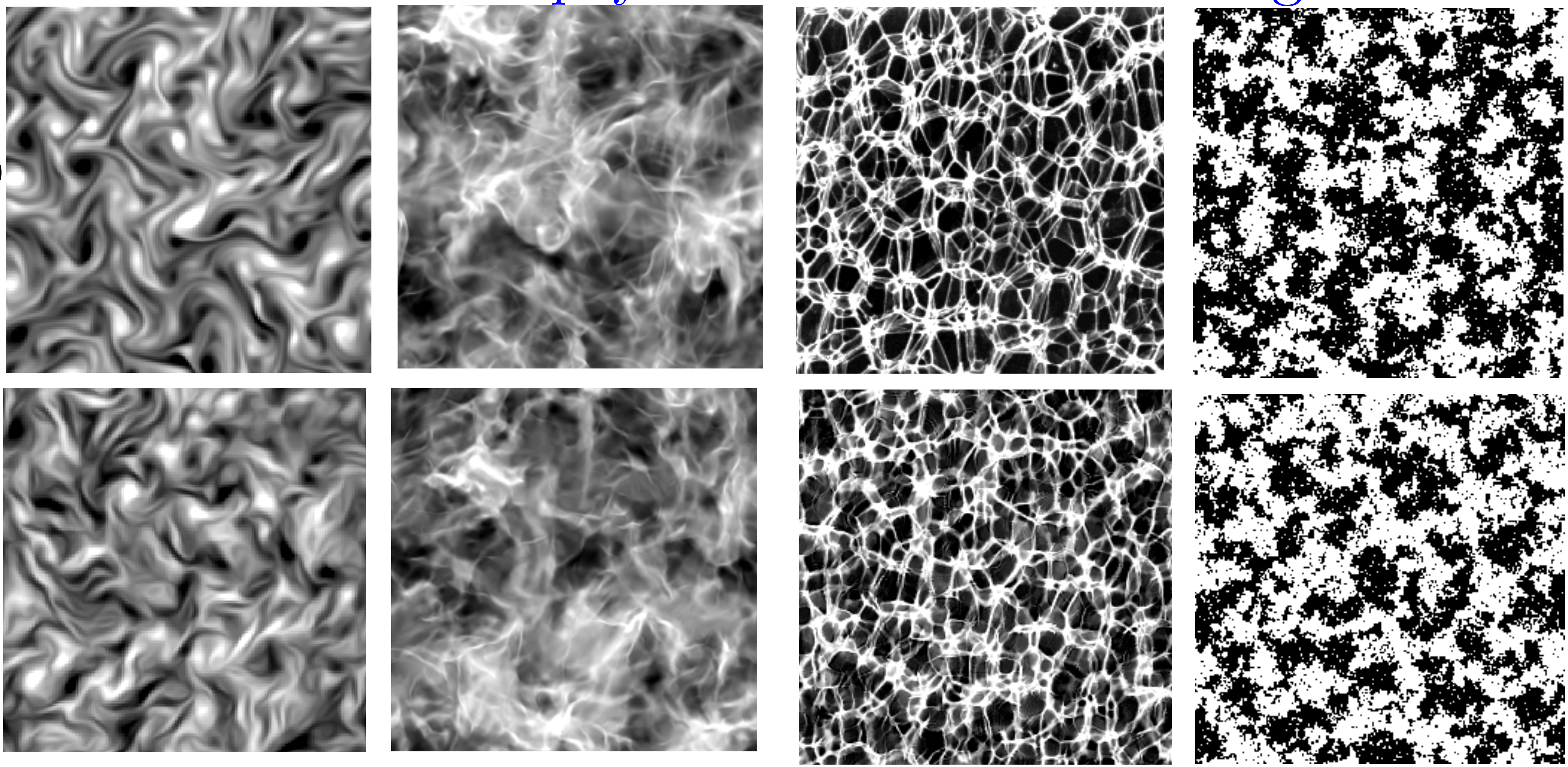
$d = 6\,10^4$

Astrophysics

Ising-critical

x_0

x



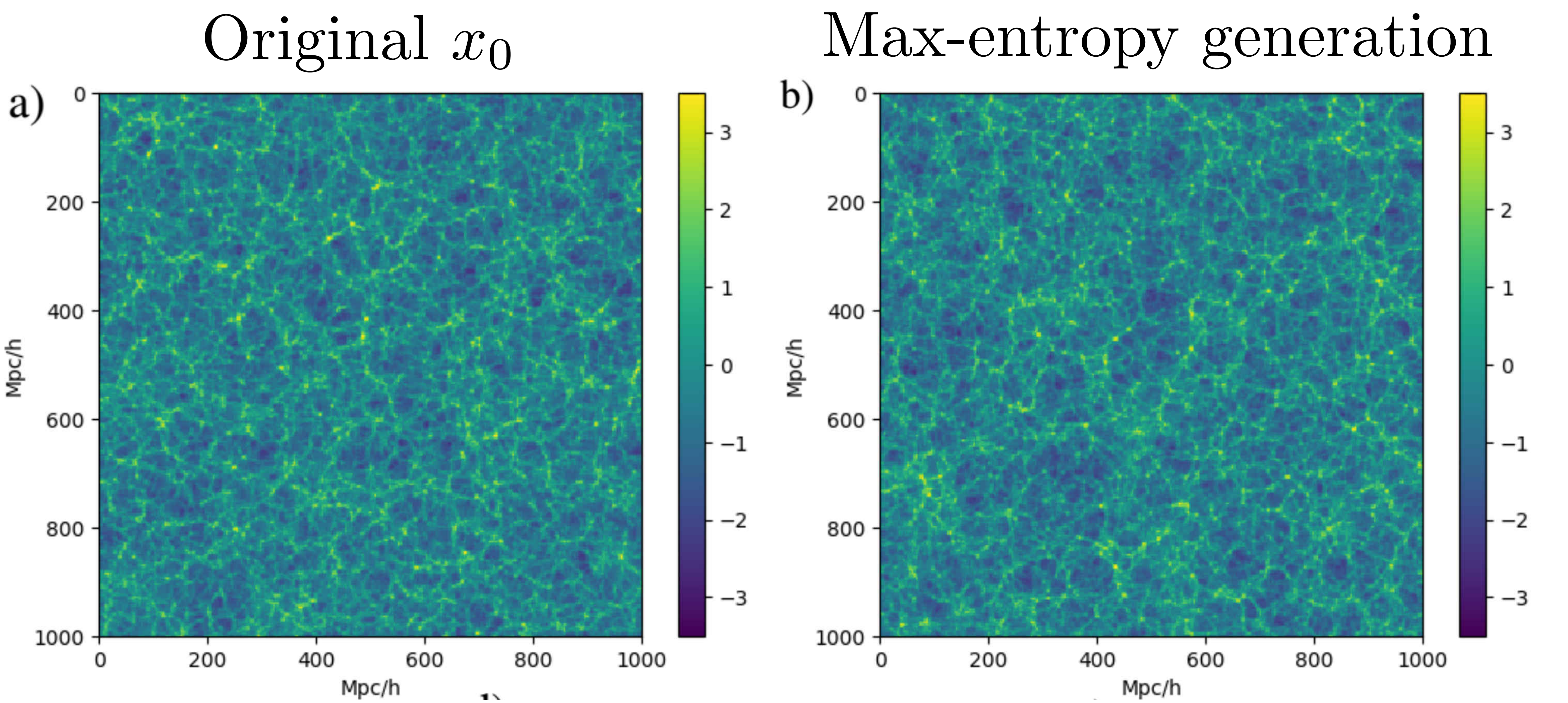
$S(x_0)$ has $2\,10^3$ empirical covariances

Sampled from $S(x_0)$ with SGD algorithm

Generation of Cosmological Models

E. Allys, T. Marchand, J.F. Cardoso, F. Villaescusa, S. Ho, S. Mallat

Generation of matter density fields from rectified wavelet covariances:



- Reproduces high order moments
- Accurate regression of 6 cosmological parameters from $S(x_0)$
- Applications in finance : simulations of markets *R. Morel*

III - Image Classification

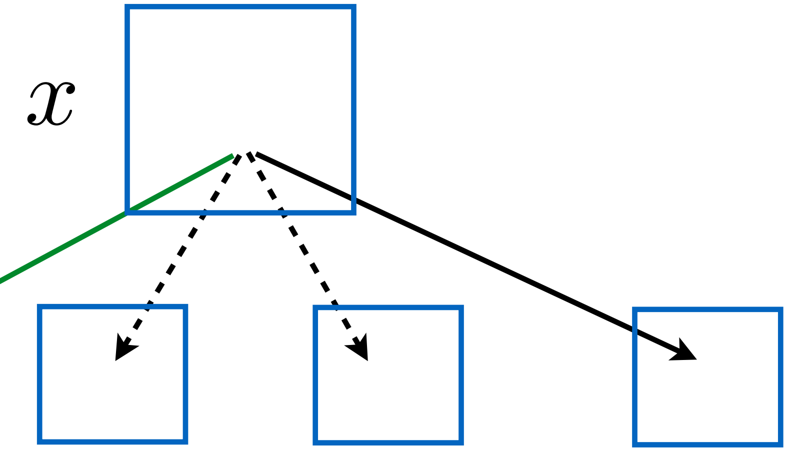
- A deep network progressively separates and concentrates

    - Can we do it from prior without learning ?

    - If not, what needs to be learned ?

Wavelet Scattering Network

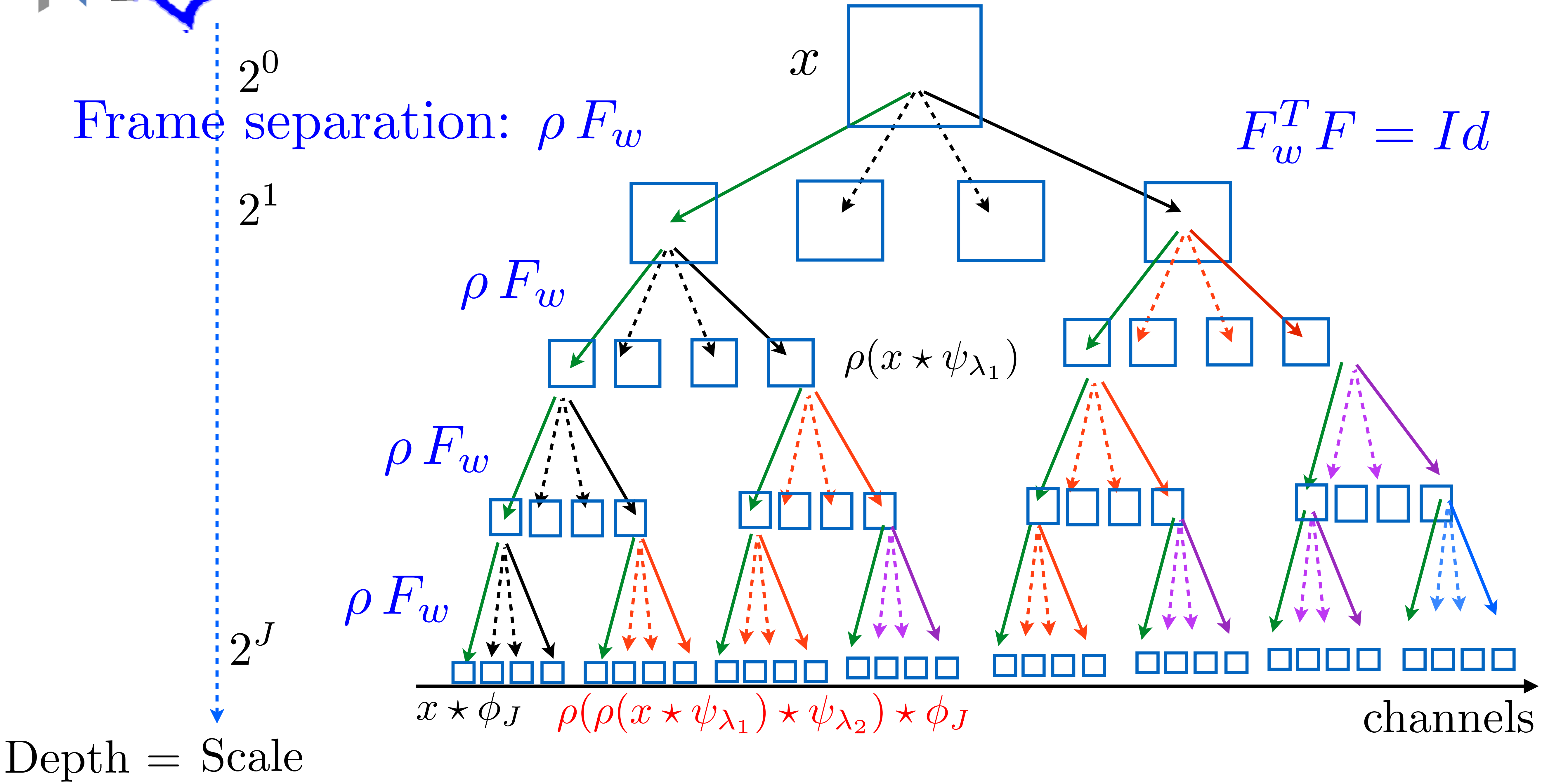
Frame separation: $\rho\, F_w$



$F_w^T F = Id$

ρF_w separates phases and orientations without contraction.

Wavelet Scattering Network



$$\Phi = \rho\, W_J \;\ldots\; \rho\, W_2\, \rho\, W_1 \quad : \text{separation}$$

$$\Phi = (\rho\, F_w)^J \;:\; \text{iterated frame separations}$$

Scatters along progressively more channels

A convolution tree: no channel connections no learning

Scattering Deformation Stability

$$S_J x = \begin{pmatrix} x \star \phi_{2^J} \\ \rho(x \star \psi_{\lambda_1}) \star \phi_{2^J} \\ \rho(\rho(x \star \psi_{\lambda_1}) \star \psi_{\lambda_2}) \star \phi_{2^J} \\ \cdots \end{pmatrix}_{\lambda_1, \lambda_2, \ldots} = \ldots \rho\, W_2\, \rho\, W_1\, x$$

Lipschitz continuity to deformations $D_\tau x(u) = x(u - \tau(u))$

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

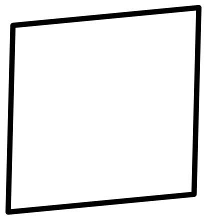
Theorem: there exists $C > 0$ such that

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

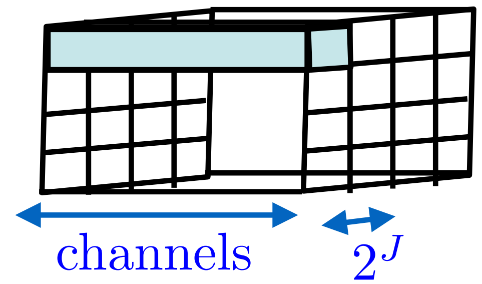
Image Classification

Scat-Net$_J$: J layers

$x \longrightarrow \boxed{\rho(\rho(x \star \psi_\lambda) \star \psi_{\lambda'}) \star \phi_J} \longrightarrow \Phi(x) \longrightarrow \boxed{C} \longrightarrow \tilde{y}$



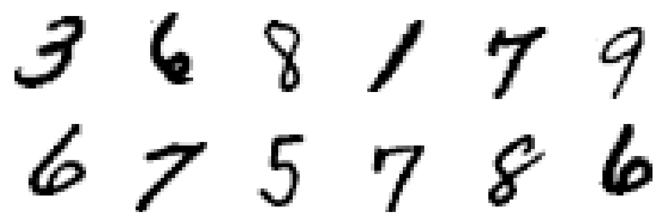
no learning



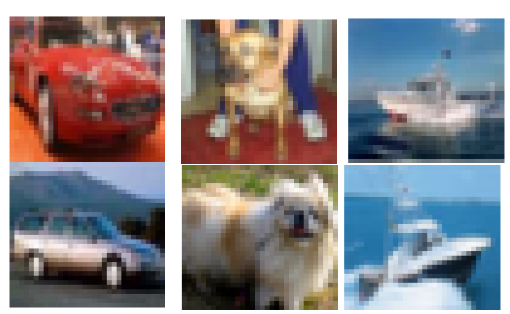
supervised learning

Errors:

MNIST: 28^2
10 classes



CIFAR: 32^2
10 classes



ImageNet: 228^2
10^3 classes
1 million training

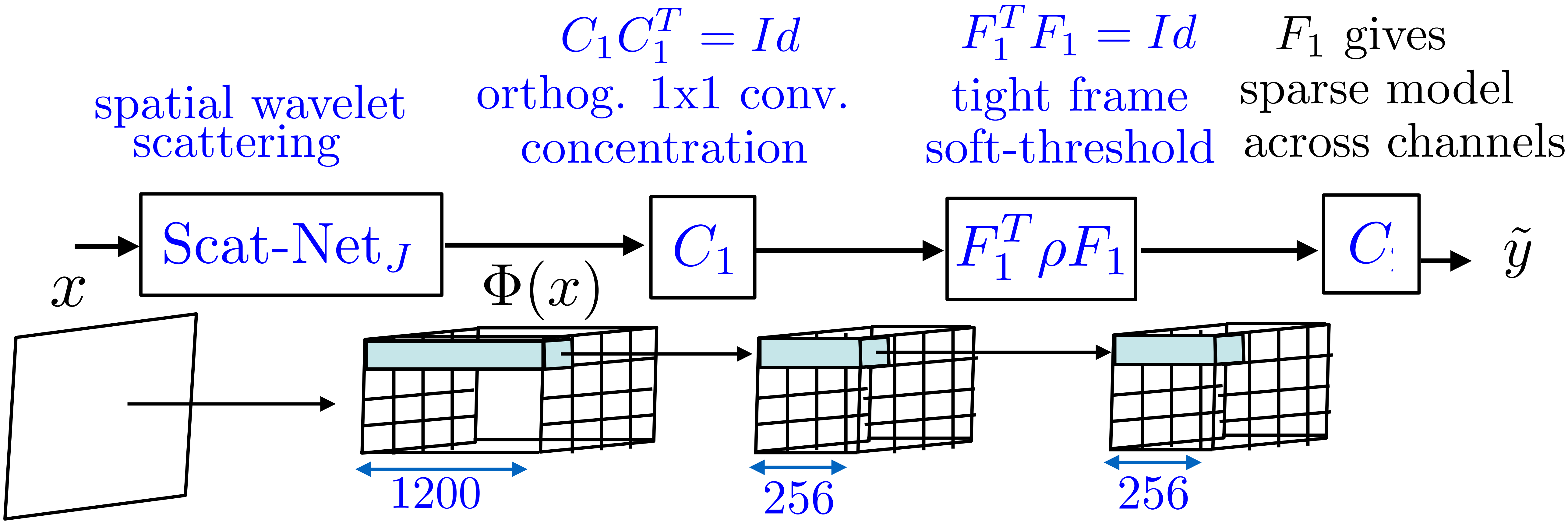
	Scattering	Deep Nets.
$J = 3$	0.5 %	0.5 %
$J = 4$	23%	ResNet-18: 8% ResNet-50: 7.6%
$J = 6$	52 %	AlexNet-7: 20% ResNet-18: 11% Res-Net 50: 7%

What is learned ?

One Concentrated Scattering

John Zarka, Florentin Guth

Frame soft-thresholding along scattering channels:



- SGD optimisation

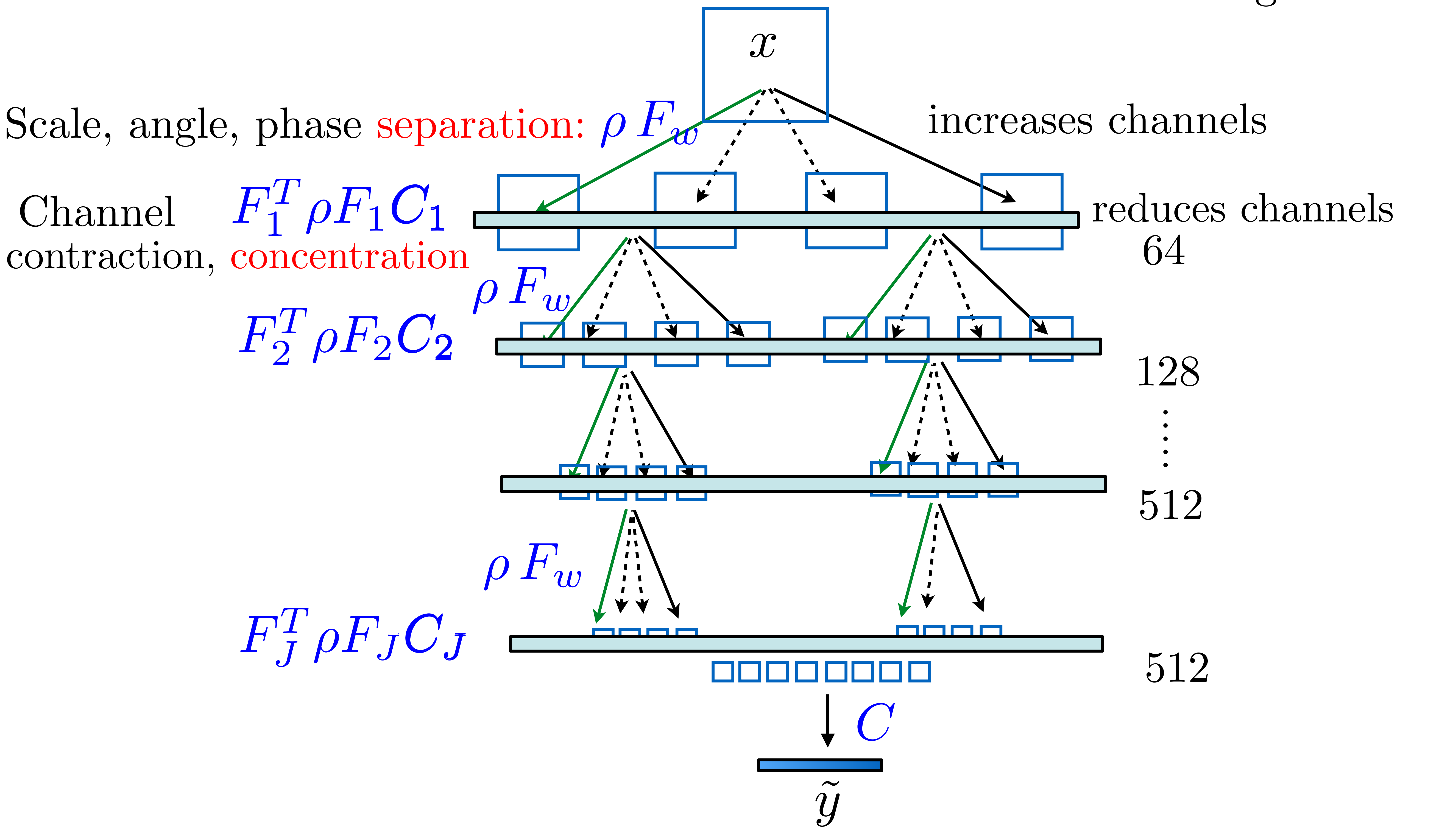
	$\Phi(x)$	Scat.	1CoScat	ResNet-18
CIFAR	Error	27%	18%	8%
	Fisher	22	30	
ImageNet Top 5	Error	60%	30%	11%
	Fisher	2.9	3.4	

Multiscale Concentrated Scattering

Wavelet frame contraction: $\rho \, F_w$ spatial conv, shrinks sign

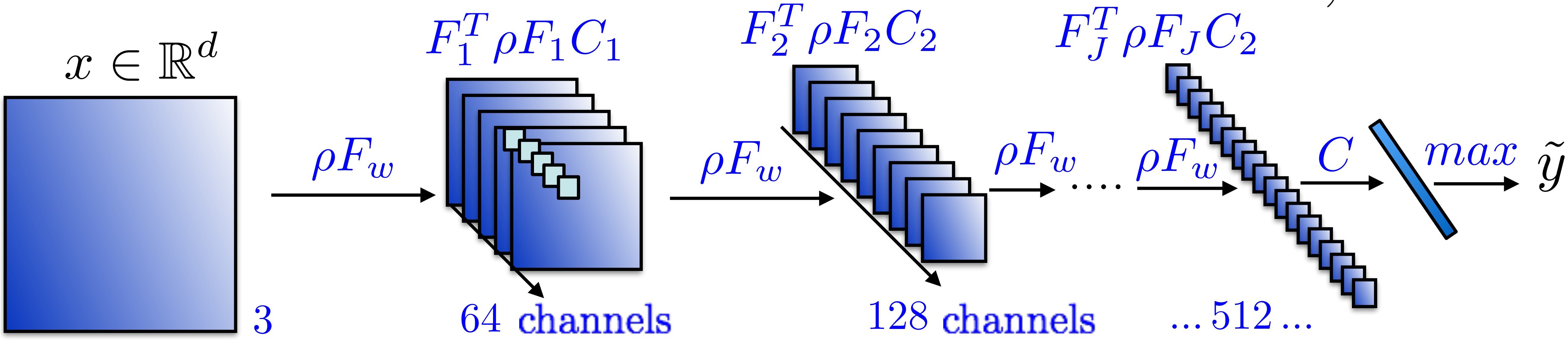
Concentrated frame contraction: $F_j^T \, \rho F_j \, C_j$ shrinks amplitude

1x1 conv. along channels



Concentrated Scattering

J. Zarka, F. Guth



- Network without learning bias
- Learning 1x1 convolutions across scattering channels
- SGD optimisation

	$\Phi(x)$	1CoScat	JCoScat	ResNet-18
CIFAR	Error	18%	7.8%	8%
	Fisher	30	70	
	Depth	5	8	18
ImageNet Top 5	Error	30%	11%	11%
	Fisher	3.4	7.2	
	Depth	7	12	18

What properties of the C_j what geometry ?

Conclusion

- **Deep network separate and concentrate: what mechanism ?**
- Links with statistical physics and large deviations

- Means are separated by separating phases/signs of frame coifs
- Variance can be reduced with tight frame shrinking
- Spatial filtering with wavelet frame is sufficient to separate means across scales, angles and phases.
- State of the art by learning contractions along channels

- What geometry in the scattering domain ?
- Control of ***Fisher concentration ratios*** is an open math. problem.