Understanding (or not)

Deep Convolutional Networks

Stéphane Mallat

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

Deep Neural Networks

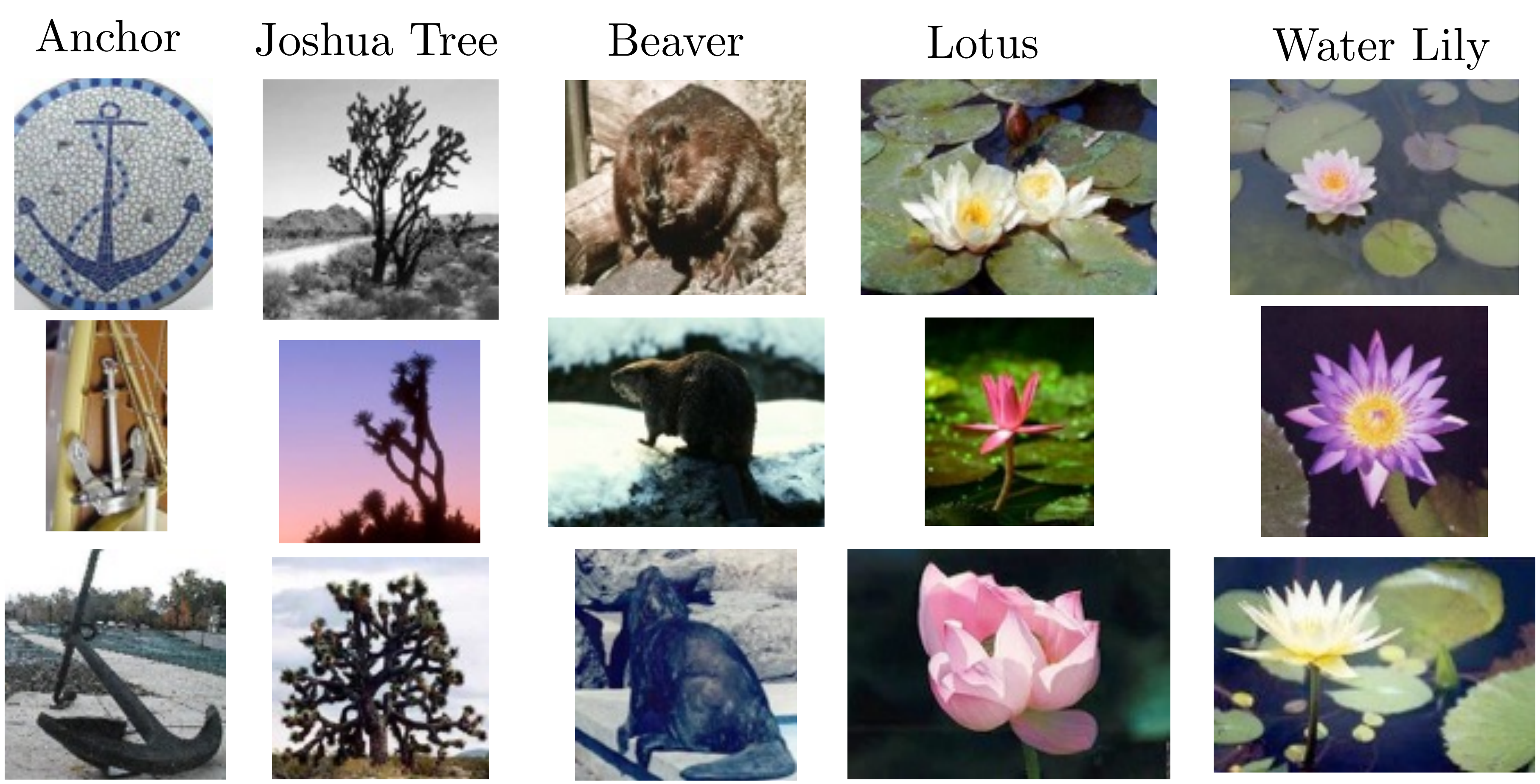
- Approximations of high-dimensional functions from examples, for classification and regression.
- **Applications:** computer vision, audio and music classification, natural language analysis, bio-medical data, unstructured data…
- **Related to:** neurophysiology of vision and audition, quantum and statistical physics, linguistics, …
- **Mathematics:** statistics, probability, harmonic analysis, geometry, optimization. *Little is understood*.

High Dimensional Learning

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

- **Classification:** estimate a class label $f(x)$ given n sample values $\{x_i \, , \, y_i = f(x_i)\}_{i \leq n}$

Image Classification $d = 10^6$

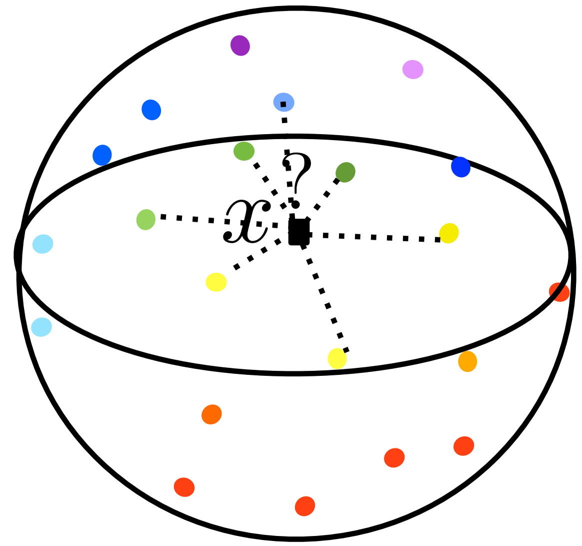


Huge variability inside classes

Find invariants

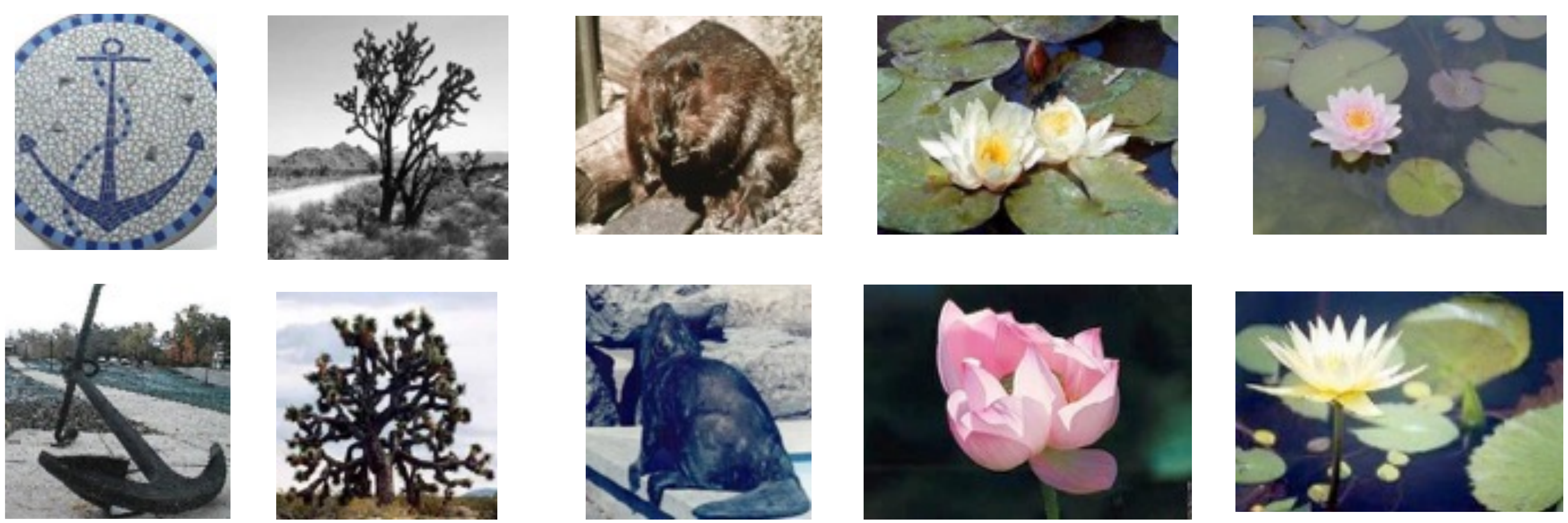
Curse of Dimensionality

- $f(x)$ can be approximated from examples $\{x_i\,,\, f(x_i)\}_i$ by local interpolation if f is regular and there are close examples:



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

$$\Rightarrow \|x - x_i\| \text{ is always large}$$



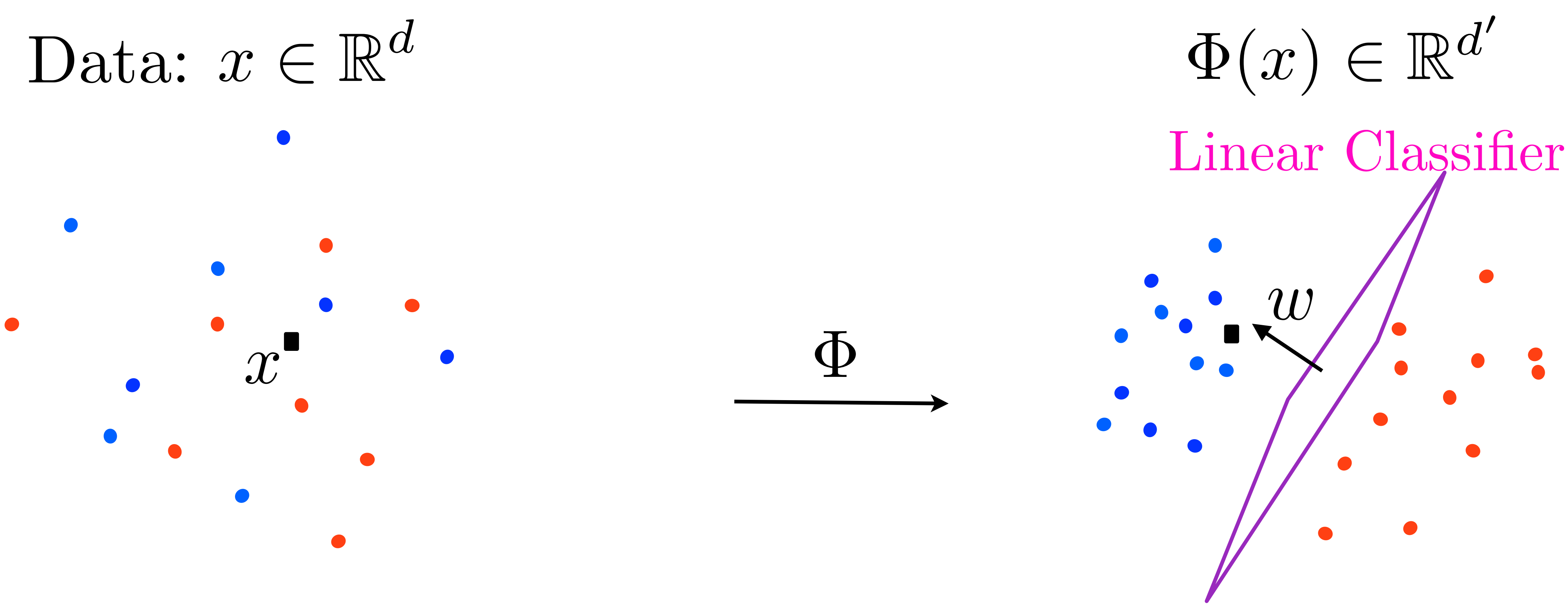
Huge variability inside classes

Linearisation by Change of Variable

Change of variable $\Phi(x) = \{\phi_k(x)\}_{k \leq d'}$

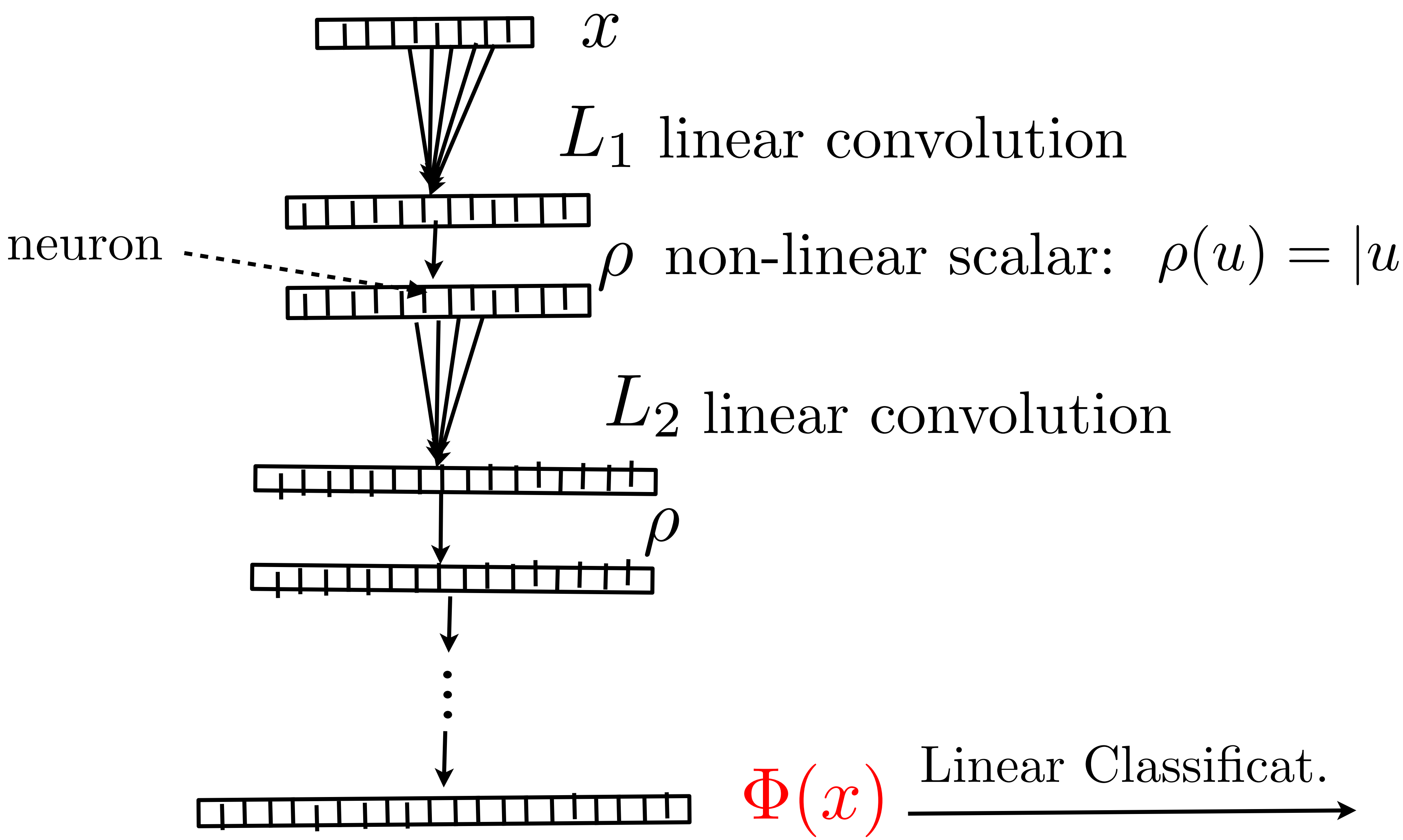
to nearly linearize $f(x)$, which is approximated by:

$$\tilde{f}(x) = \underbrace{\langle \Phi(x)\,,\, w \rangle}_{\textbf{1D projection}} = \sum_k w_k\, \phi_k(x)\ .$$



Deep Convolution Neworks

- The revival of an old (1950) idea: *Y. LeCun, G. Hinton*



Optimize L_j with architecture constraints: over 10^9 parameters

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

Products by FaceBook, IBM, Google, Microsoft, Yahoo...

Why does it work so well ?

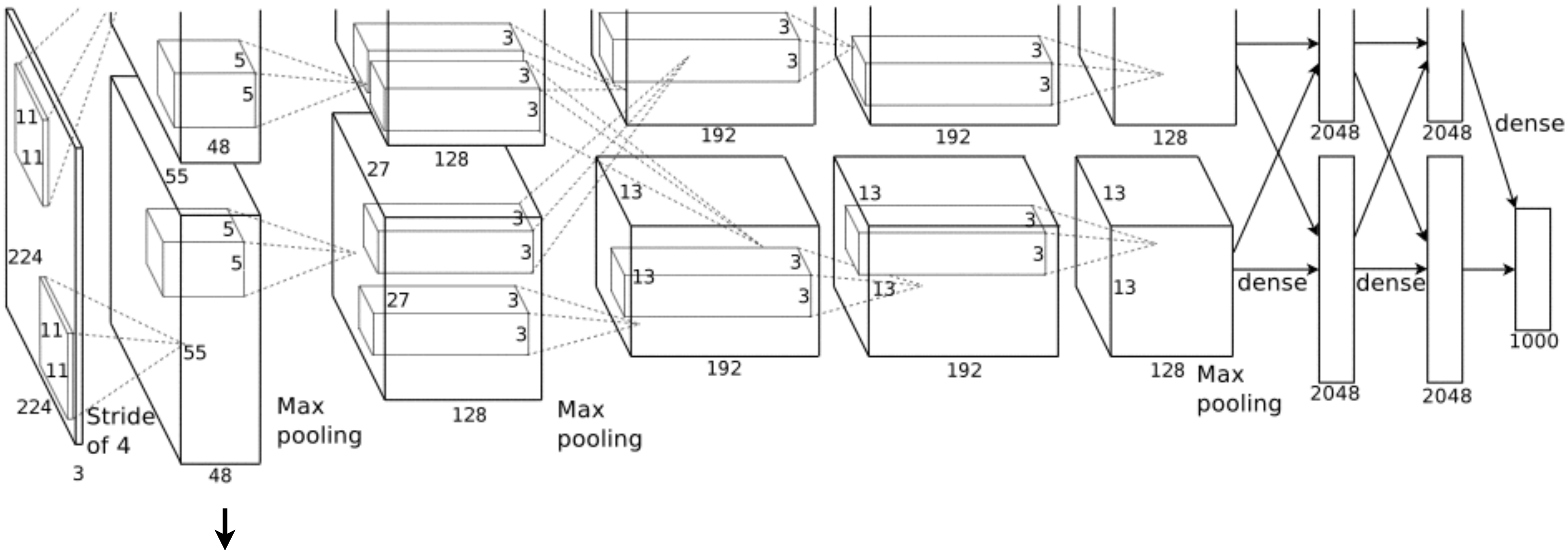
ImageNet Data Basis

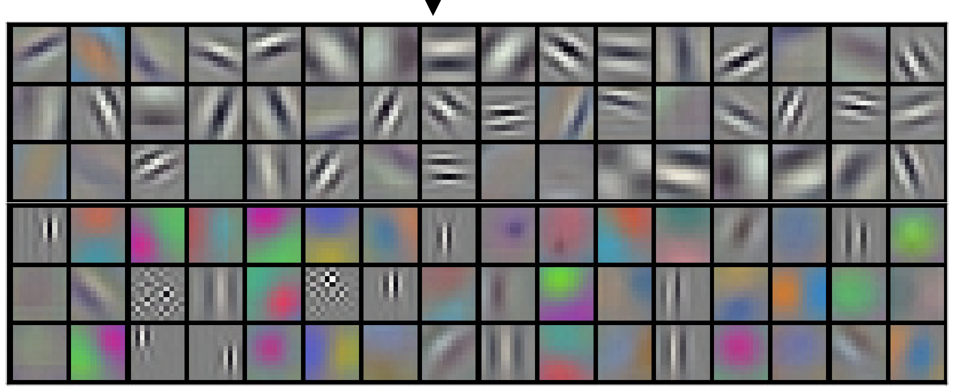
- Data basis with 1 million images and 2000 classes

Alex Deep Convolution Network

A. Krizhevsky, Sutsever, Hinton

- Imagenet supervised training: 1.2 10^6 examples, 10^3 classes
  15.3% testing error in 2012

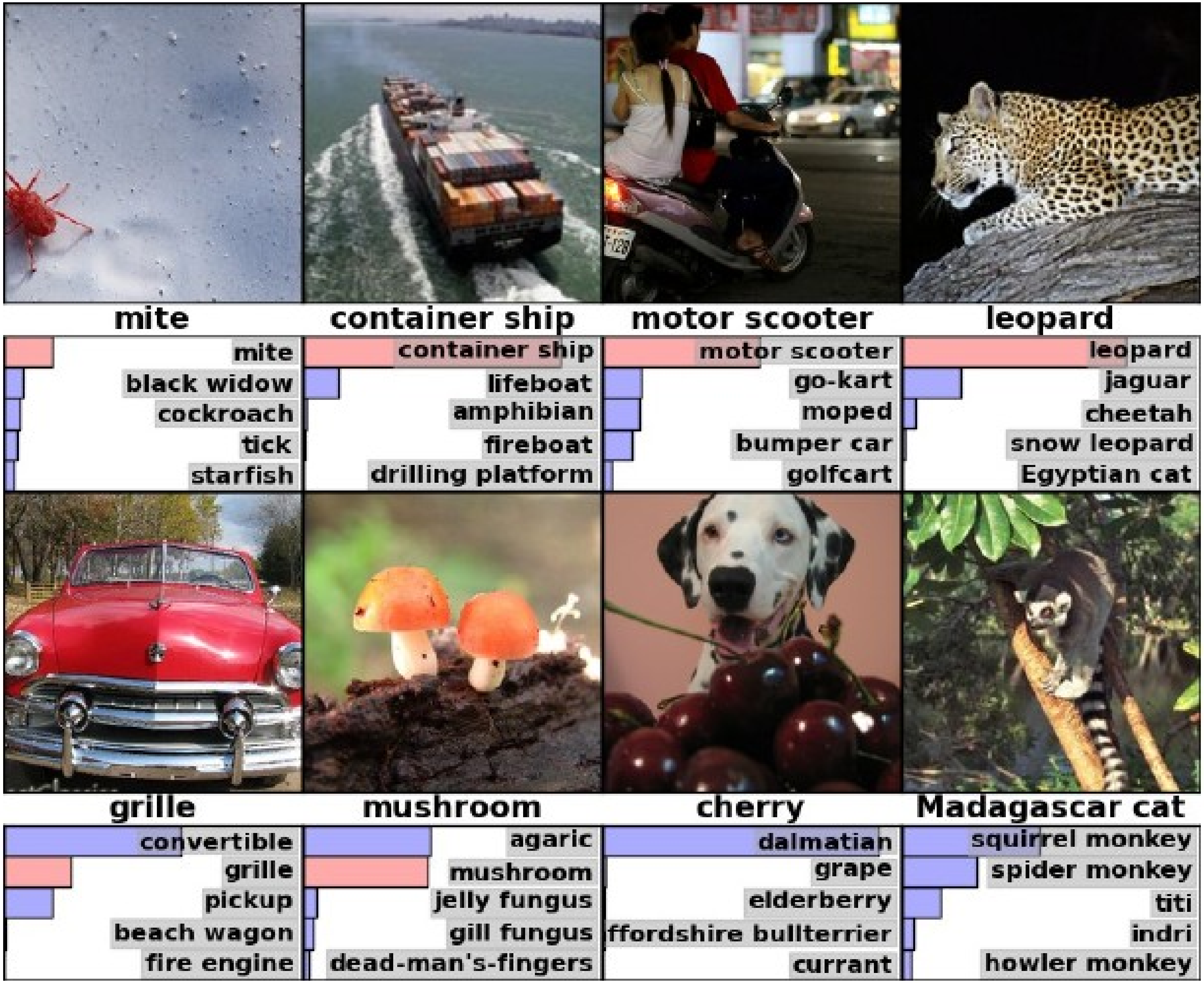




Wavelets

New networks with 5% errors.
with 150 layers!

Image Classification



Scene Labeling / Car Driving

Overview

- Linearisation of symmetries
- Deep convolutional networks architectures
- Simplified convolutional trees: wavelet scattering
- Deep networks: contractions, linearization and separations

Separation and Linearization with Φ

- Separation: change of variable $f(x) = \overline{f}(\Phi(x))$

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

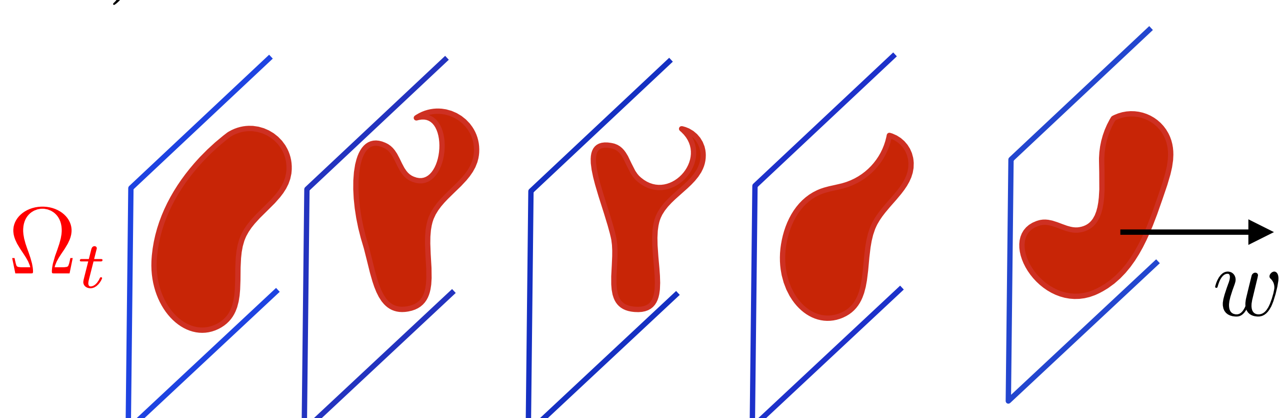
$$\overline{f}(z) \text{ is Lipschitz} \quad \Leftrightarrow \quad \|\Phi(x) - \Phi(x')\| \geq \epsilon \, |f(x) - f(x')|$$

- Linearization: $\overline{f}(z) = \langle w, z \rangle$

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

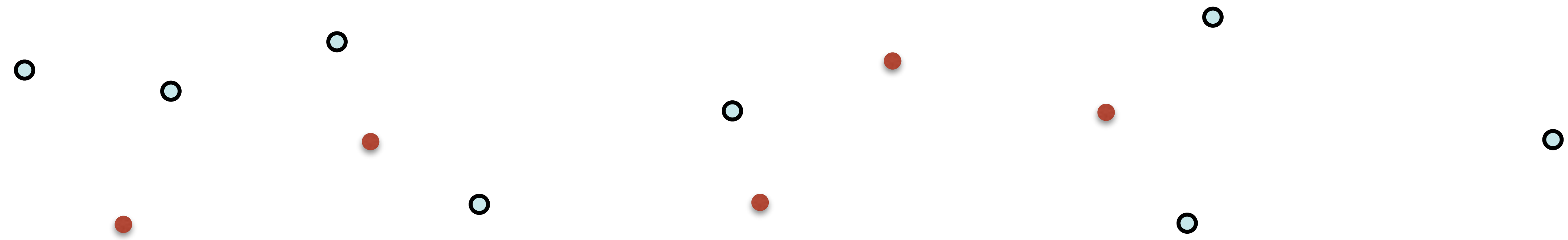
$$\forall x \in \Omega_t \quad , \quad f(x) = \langle \Phi(x), w \rangle = t$$

$\Phi(\Omega_t)$ for all t are in parallel linear spaces



Linearization of Symmetries

- No local estimations because of dimensionality curse



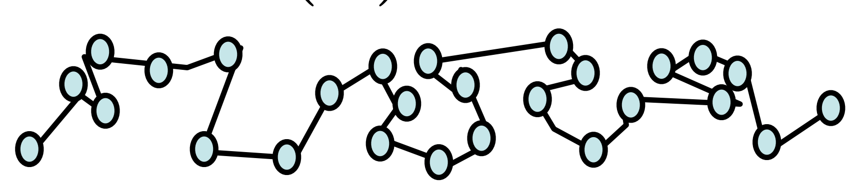
- A symmetry is an operator g which preserves level sets:

$$\forall x \quad , \quad f(g.x) = f(x) \; : \; \text{global}$$

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

$\Rightarrow$ groups G of symmetries: high dimensional

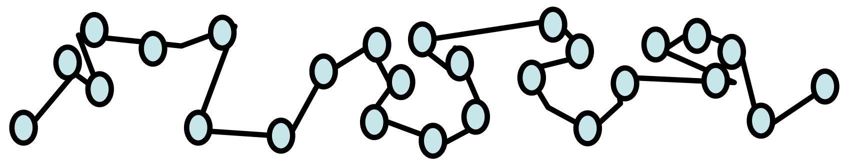
- A change of variable $\Phi(x)$ must linearize the orbits $\{g.x\}_{g\in G}$



Problem: find the symmetries and linearise them.

Contract Linearize Symmetries

- A change of variable $\Phi(x)$ must linearize the orbits $\{g.x\}_{g \in G}$



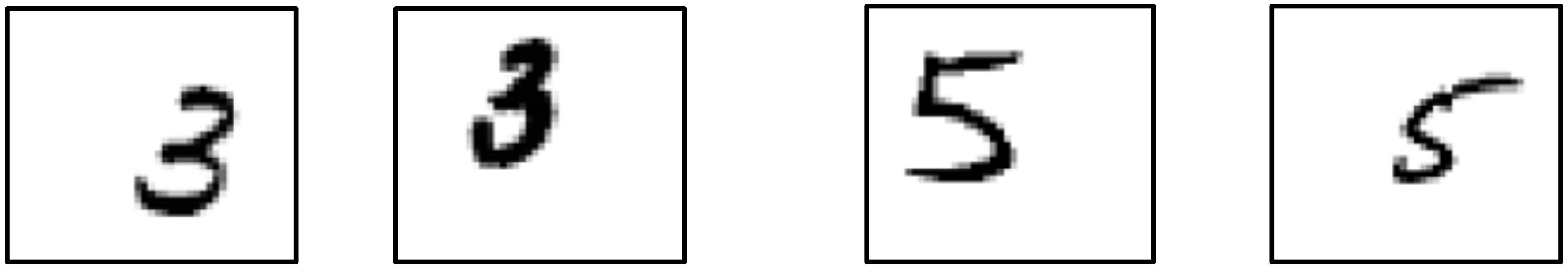
Problem: find the symmetries and linearise them.

- Regularize the orbit, remove high curvature: linearisation

Translation and Deformations

- Digit classification:

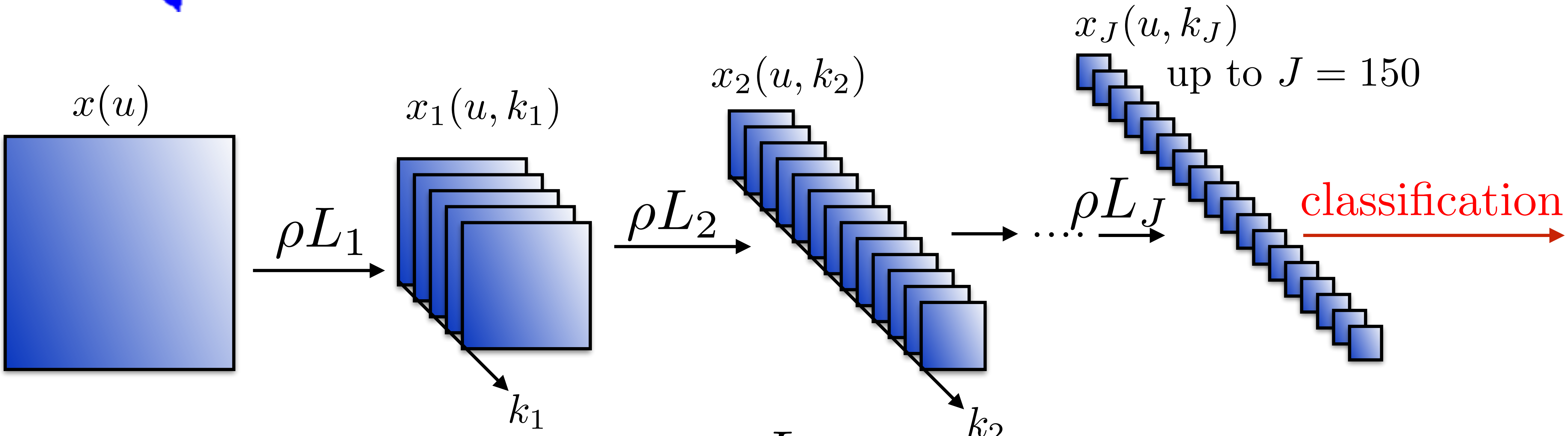
$x(u)$ $x'(u)$



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

Video of Philipp Scott Johnson

Deep Convolutional Networks



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

- ρ is a pointwise contractive non-linearity:

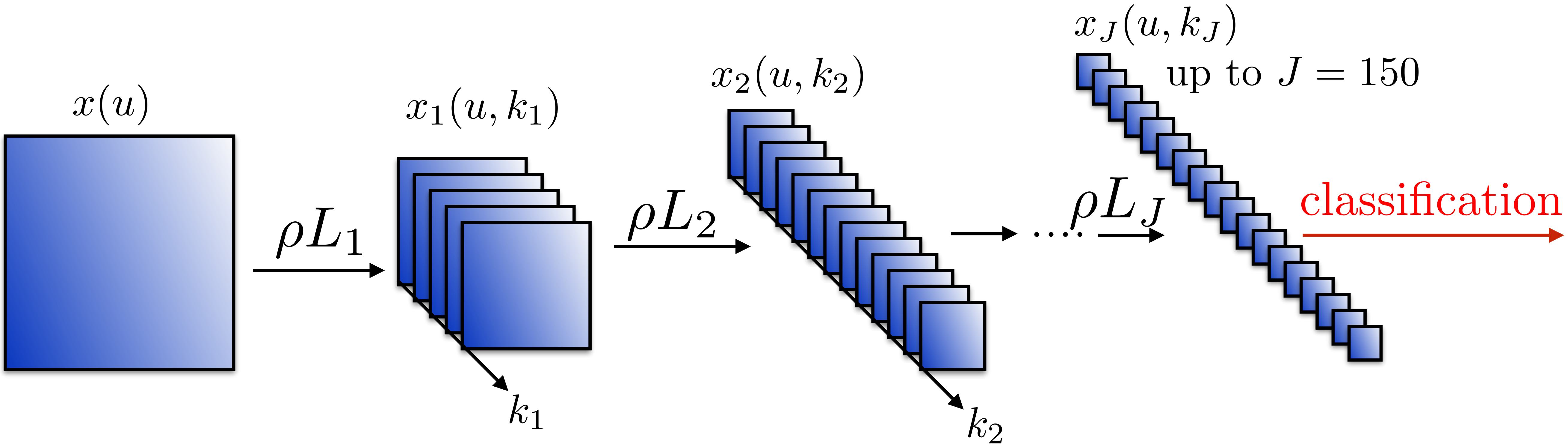
$$\forall (\alpha, \alpha') \in \mathbb{R}^2 \quad , \quad |\rho(\alpha) - \rho(\alpha')| \leq |\alpha - \alpha'|$$

  Examples: $\rho(u) = \max(u, 0)$ or $\rho(u) = |u|$.

- Optimisation of the L_j to minimise the training error with stochastic gradient descent and back-propagation.

- What is the role of the linear operators L_j and of ρ ?

Deep Convolutional Networks



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

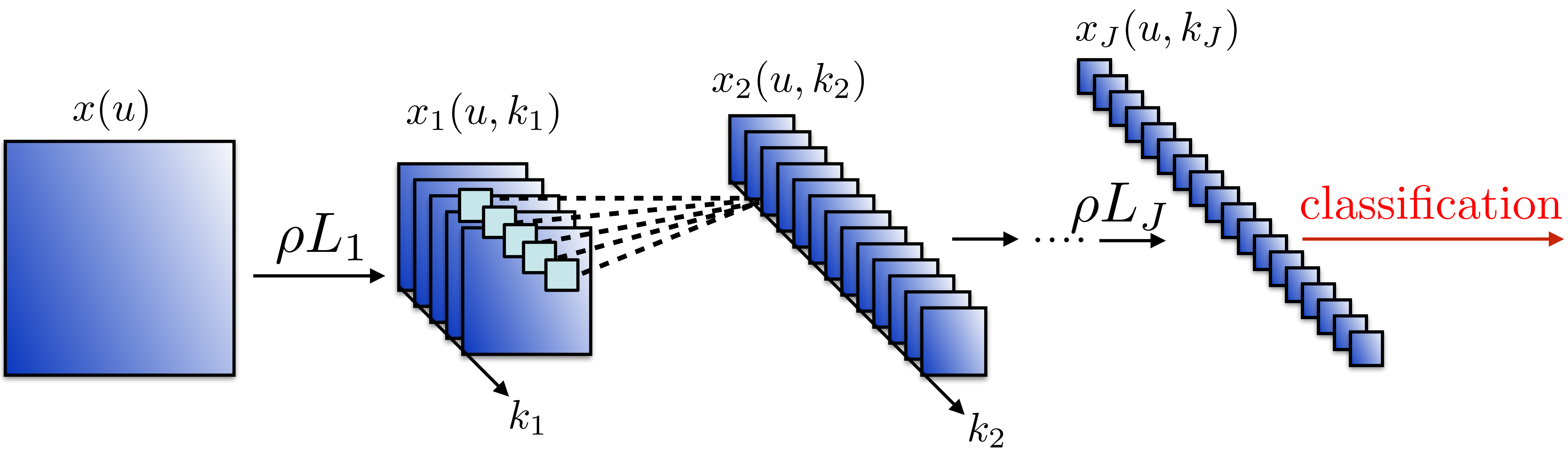
L_j has several roles:

- L_j eliminates useless linear variable: dimension reduction
- L_j computes appropriate variables contracted by ρ

Linearizes and computes invariants to groups of symmetries

- L_j is a linear preprocessing for the next layers

Deep Convolutional Networks



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

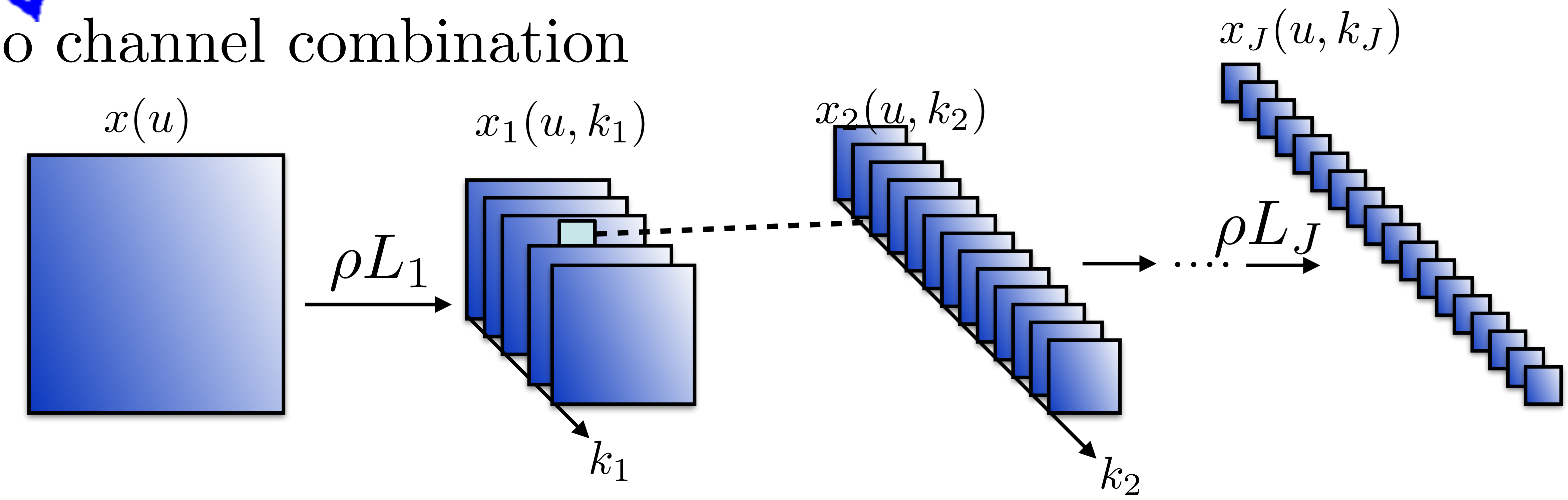
- L_j is a linear combination of convolutions and subsampling:

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

- Optimization of $h_{k_j,k}(u)$ to minimise the training error

Simplified Convolutional Networks

- No channel combination



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

- L_j is a linear combination of convolutions and subsampling:

$$x_j(u, k_j) = \rho\Big(x_{j-1}(\cdot, k) \star h_{k_j, k_{j-1}}(u) \Big)$$

no channel interaction

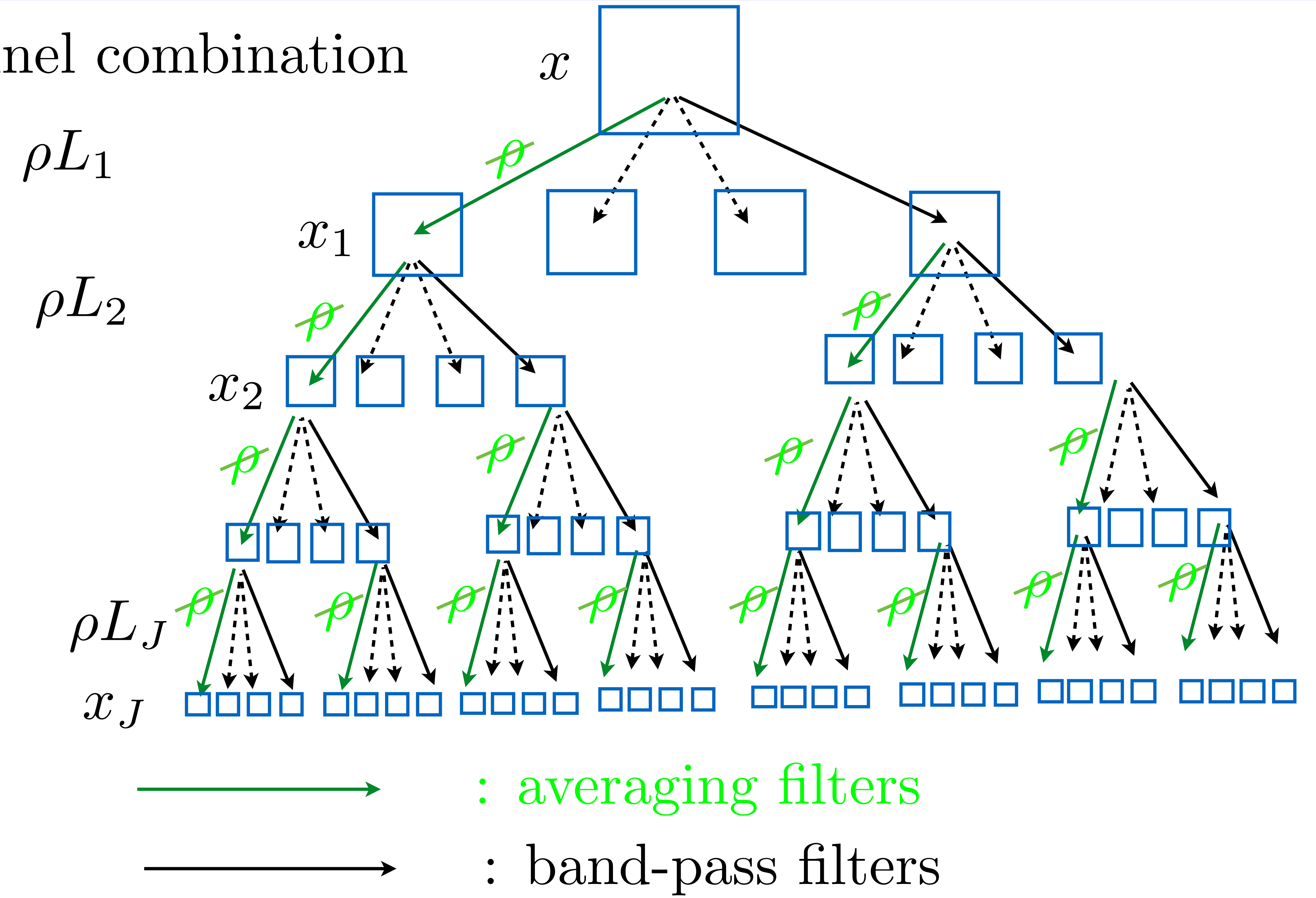
- If $\alpha \geq 0$ then $\rho(\alpha) = \alpha$

  $\Rightarrow$ if $h_{k_j, k_{j-1}}$ is an averaging filter then

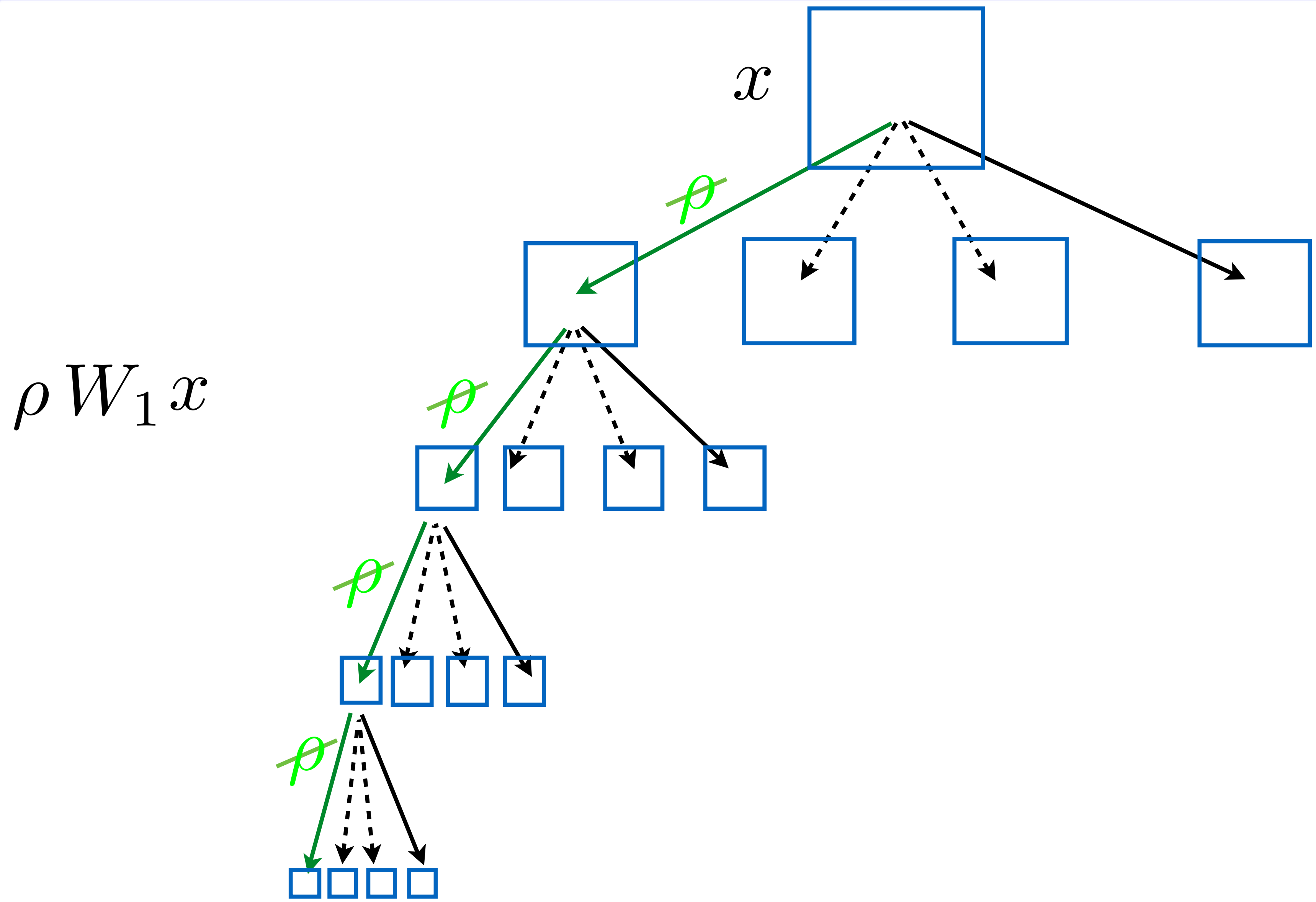
$$x_j(u, k_j) = x_{j-1}(\cdot, k) \star h_{k_j, k_{j-1}}(u)$$

Convolution Tree Network

- No channel combination



Wavelet Transform

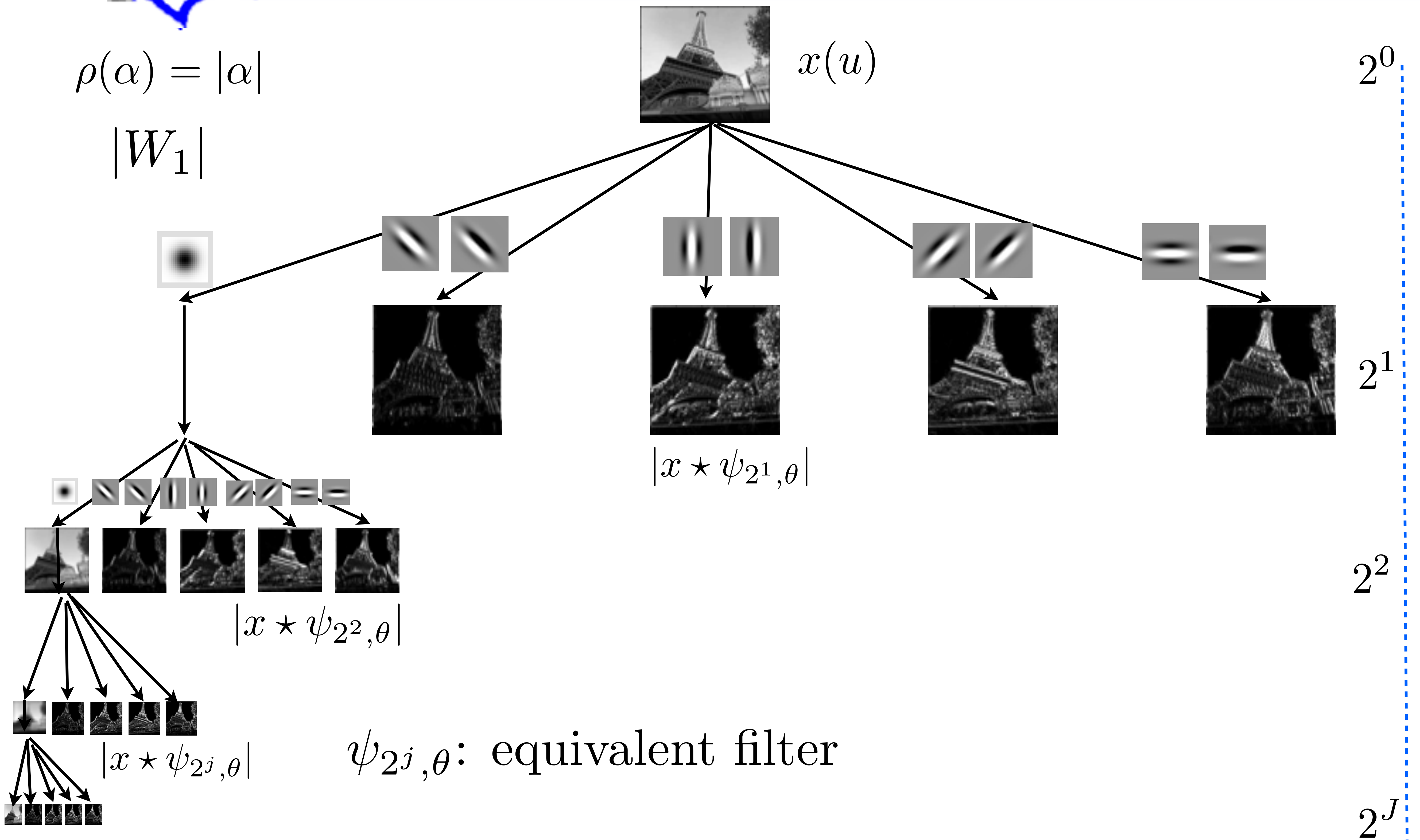


: averaging filters

: band-pass filters

W_1 : cascade of low-pass filters and a band-pass filter

Wavelet Filter Bank



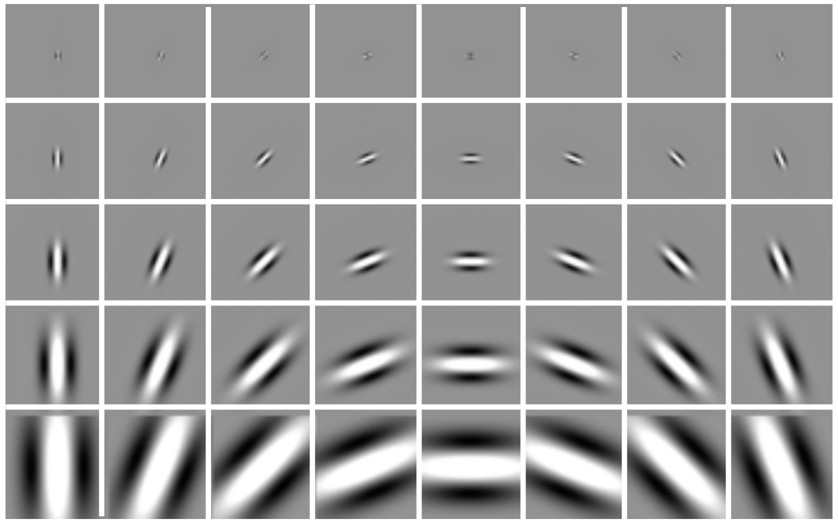
- Sparse representation

Scale separation with Wavelets

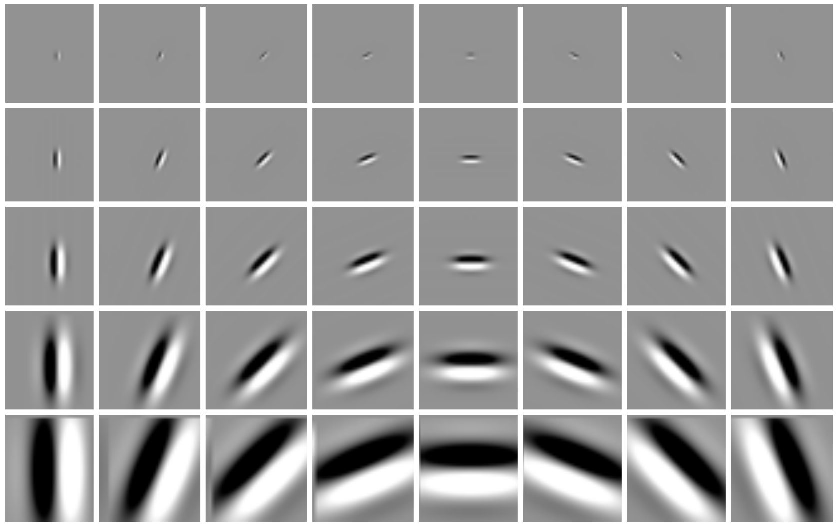
- Complex wavelet: $\psi(u) = g(u) \exp i\xi u \quad , \quad u \in \mathbb{R}^2$

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

real parts



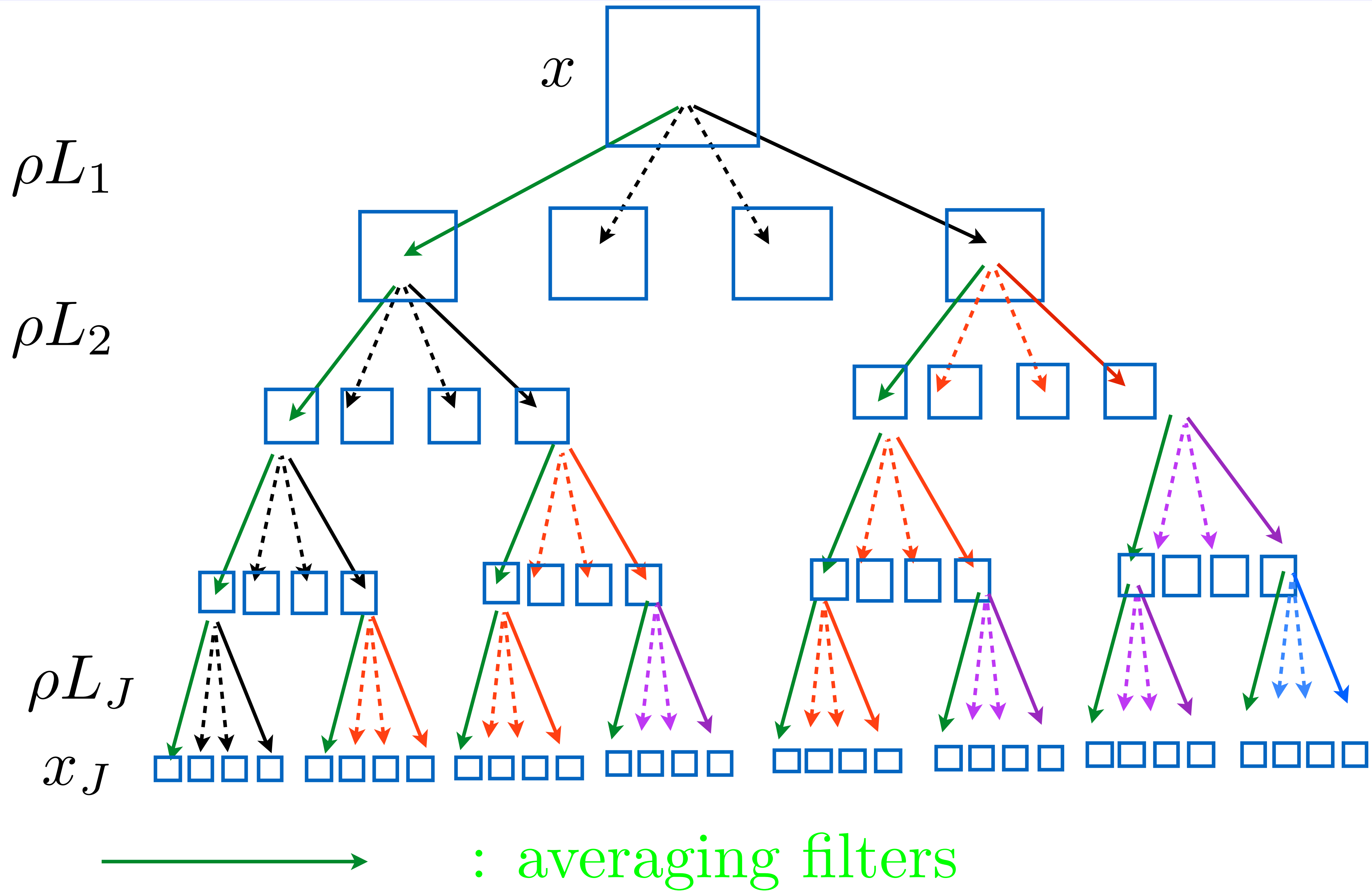
imaginary parts



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

$|x \star \psi_{2^j,\theta}(u)|$: eliminates phase which encodes local translation

Wavelet Scattering Network



$$x_J = \rho W_1 \ \rho W_2 \ \cdots \ \rho W_J \ x$$

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

$$Sx = \Big\{ |||x \star \psi_{2^{j_1},\theta_1}| \star \psi_{2^{j_2},\theta_2}| \star \ldots | \star \psi_{2^{j_m},\theta_m}| \star \phi_J \Big\}_{j_k,\theta_k}$$

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

Theorem: *For appropriate wavelets, a scattering is*

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

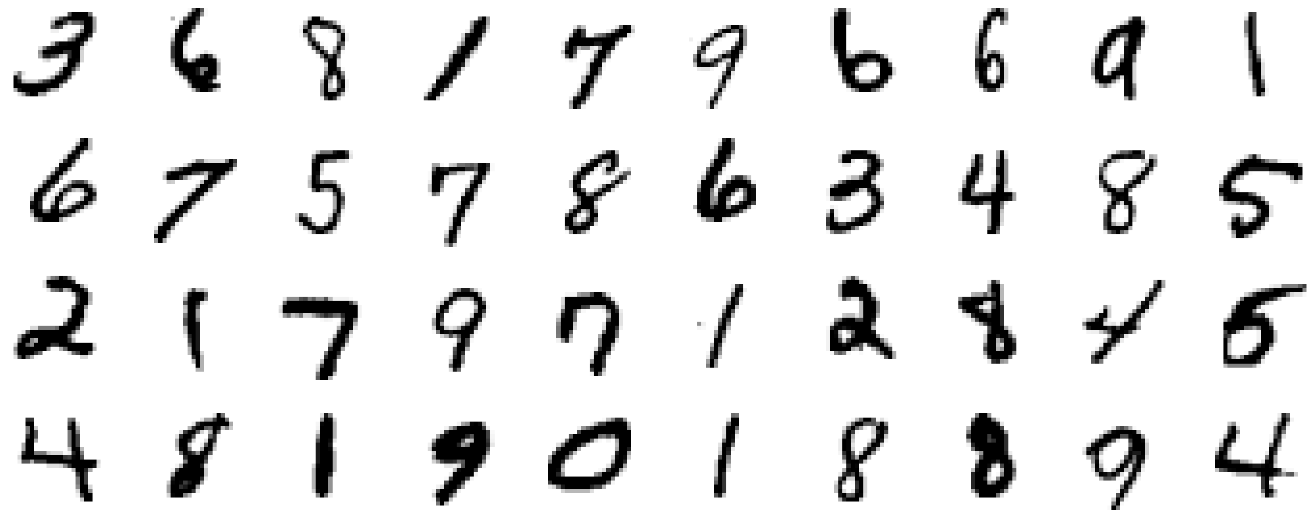
translations invariance and linearizes small deformations:

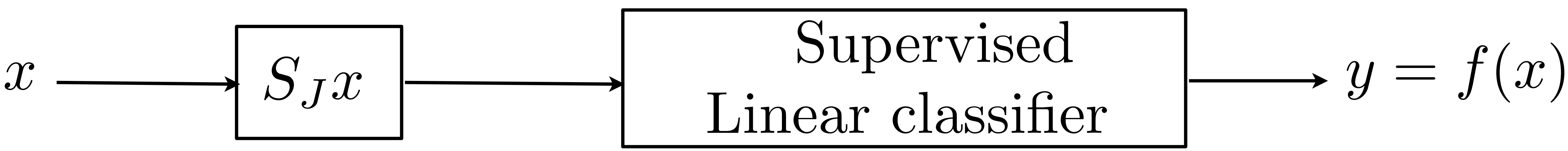
if $D_\tau x(u) = x(u - \tau(u))$ *then*

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

Digit Classification: MNIST

Joan Bruna





Invariants to translations
Linearises small deformations
No learning

Invariants to specific deformations
Separates different patterns

Classification Errors

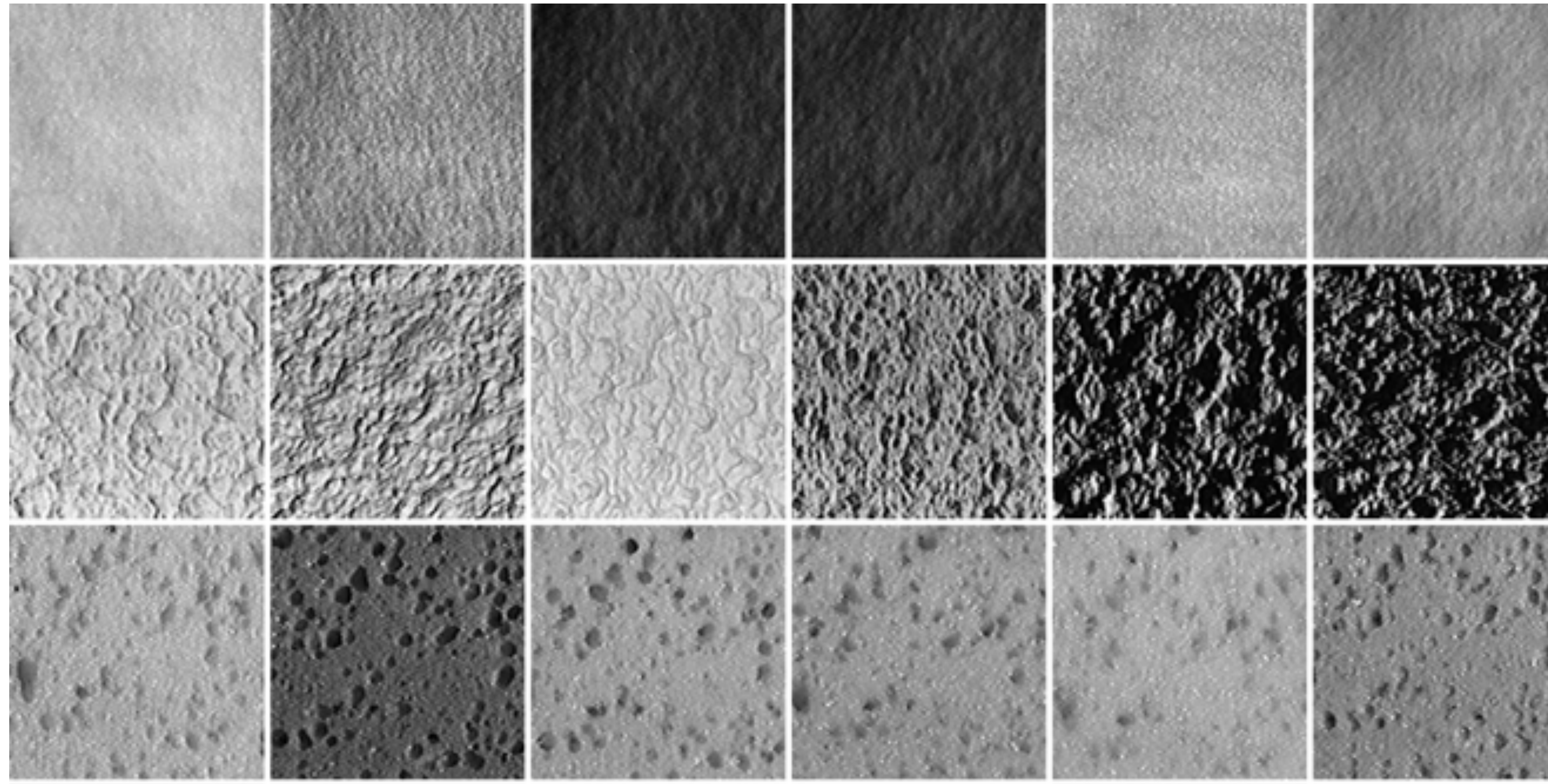
Training size	Conv. Net.	Scattering
50000	0.5%	**0.4%**

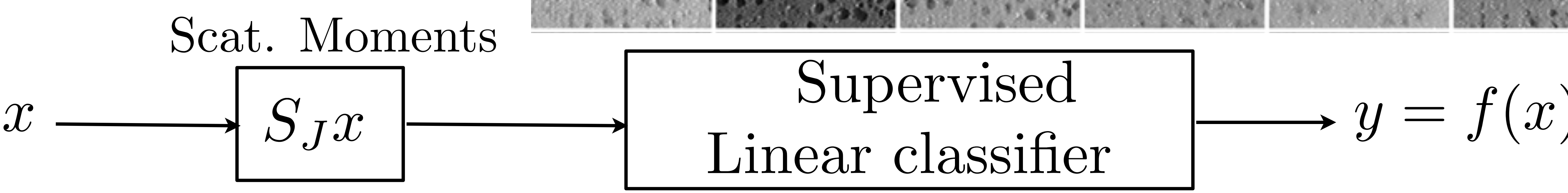
LeCun et. al.

Classification of Textures

J. Bruna

CUREt database
61 classes





Classification Errors $\quad 2^J$ = image size

Training per class	Fourier Spectr.	Histogr. Features	Scattering
46	1%	1%	**0.2** %

Reconstruction from Scattering

- Second order scattering:

$$S_J x = \Big\{ x \star \phi_J \,,\, |x \star \psi_{2^{j_1},\theta_1}| \star \phi_J \,,\, ||x \star \psi_{2^{j_1},\theta_1}| \star \psi_{2^{j_2},\theta_2}| \star \phi_J \Big\}$$

If x has N^2 pixels and $J = \log_2 N$: translation invariant

then $S_J x$ has $O([\log_2 N]^2)$ coefficients.

- If $x(u)$ is a stationary process

$$S_J x \approx \Big\{ \mathbb{E}(x) \,,\, \mathbb{E}(|x \star \psi_{2^{j_1},\theta_1}|) \,,\, \mathbb{E}(||x \star \psi_{2^{j_1},\theta_1}| \star \psi_{2^{j_2},\theta_2}|) \Big\}$$

- Gradient descent reconstruction:

given a random initialisation x_0 iteratively update x_n

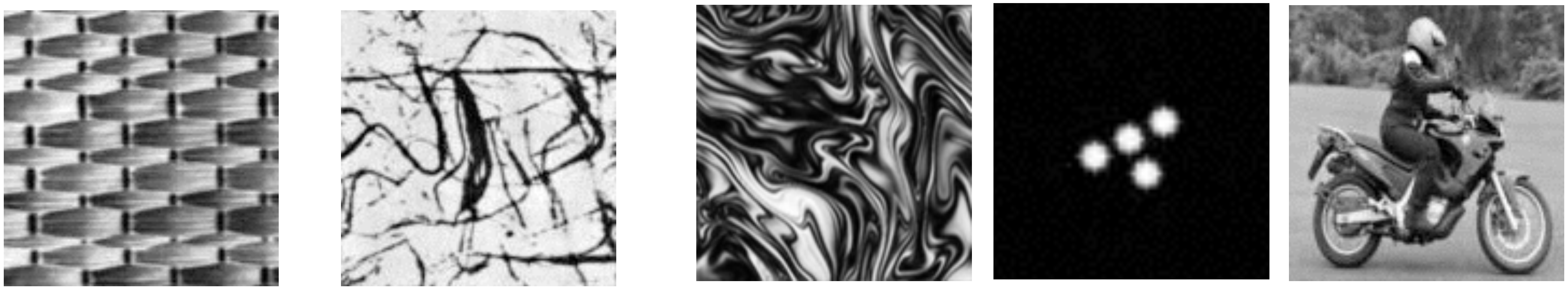
to minimise $\|S_J x - S_J x_n\|$

Translation Invariant Models

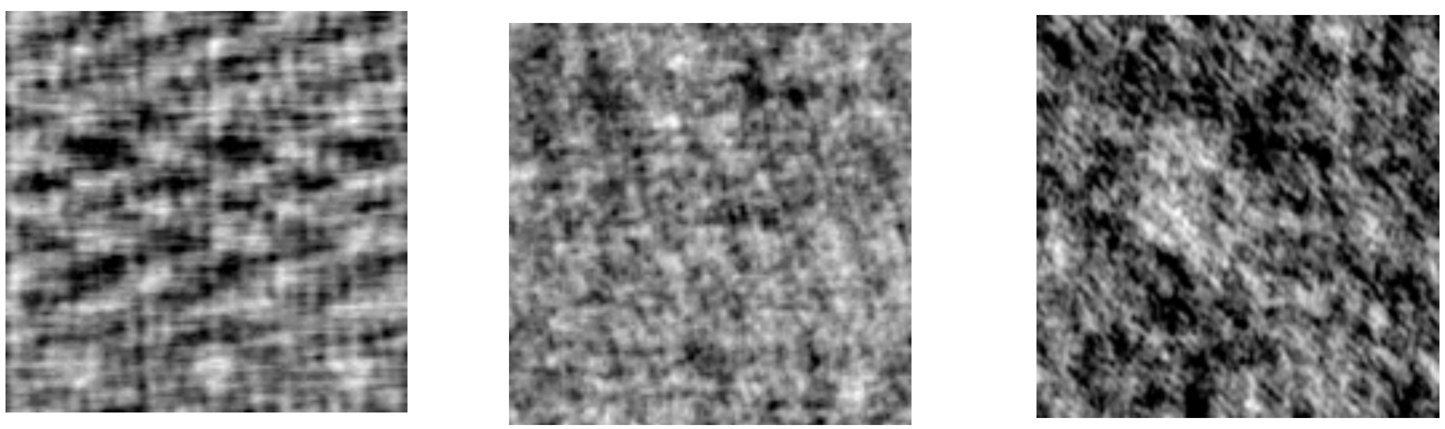
Joan Bruna

Original Textures

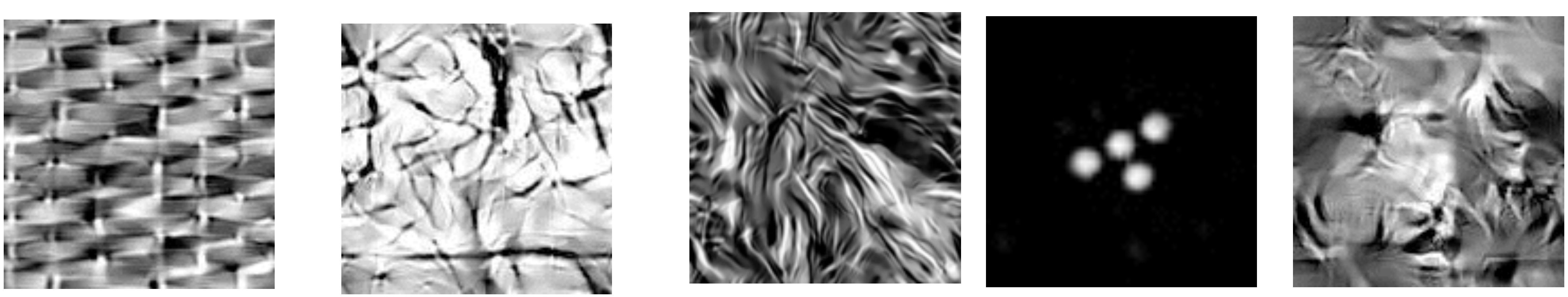
2D Turbulence   Sparse



Gaussian process model with same second order moments



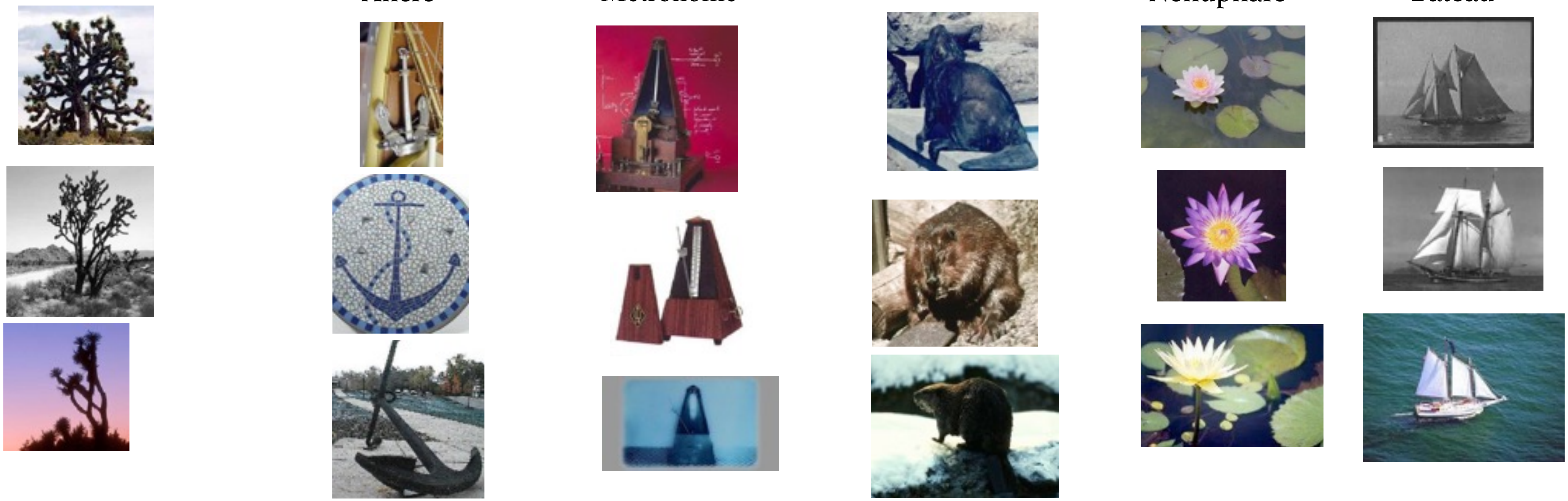
From $O((\log_2 N)^2)$ scattering coefficients of order 2



Complex Image Classification

Edouard Oyallon

Arbre de Joshua | Ancre | Metronome | Castore | Nénuphare | Bateau



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

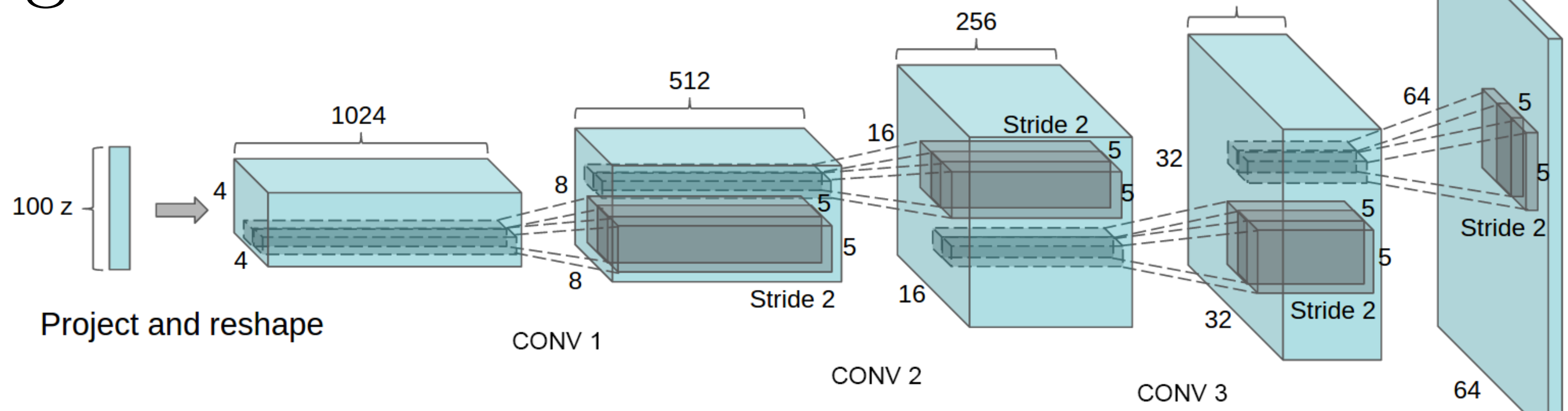
No learning

Data Basis	Deep-Net	Scat/Unsupervised
CIFAR-10	**7%**	20%

Generation with Deep Networks

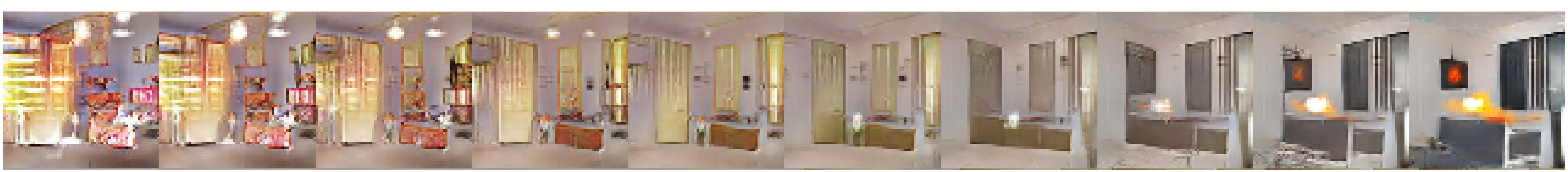
A. Radford, L. Metz, S. Chintala

- Unsupervised generative models with convolutional networks



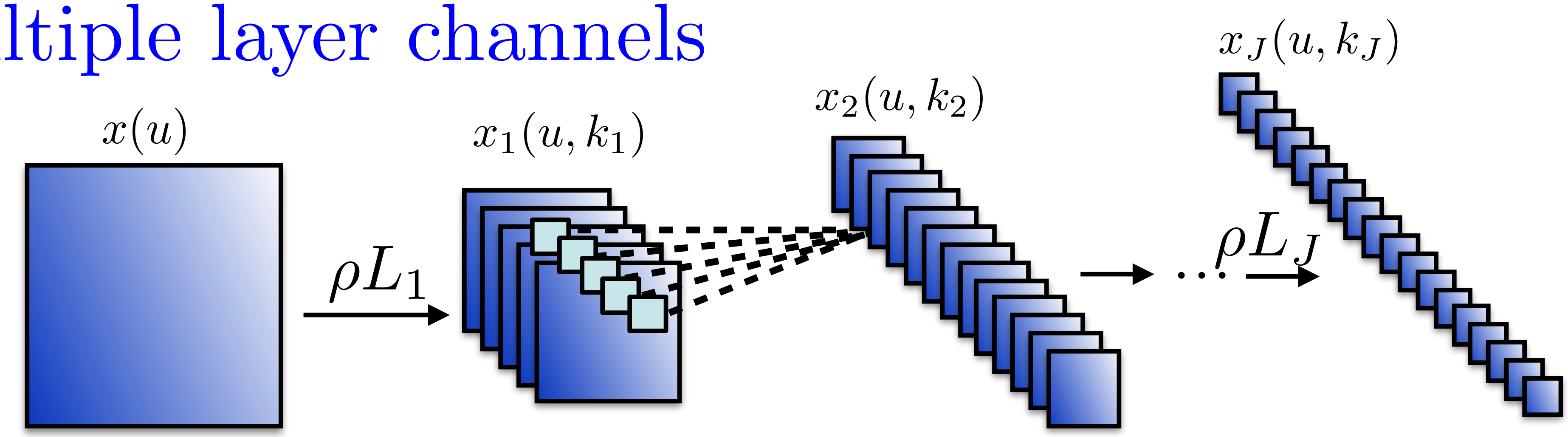
- Trained on a data basis of faces: linearization

- On a data basis including bedrooms: interpolaitons



Contractions and Separations

- Combining multiple layer channels



- A deep network progressively contracts the space while preserving margins across classes:

$$\|x_{j-1} - x'_{j-1}\| \geq \epsilon \ \text{ if } \ f(x) \neq f(x') \ .$$

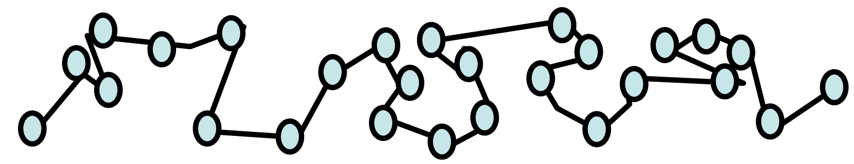
$$x_j = \rho \, L_j \, x_{j-1}$$

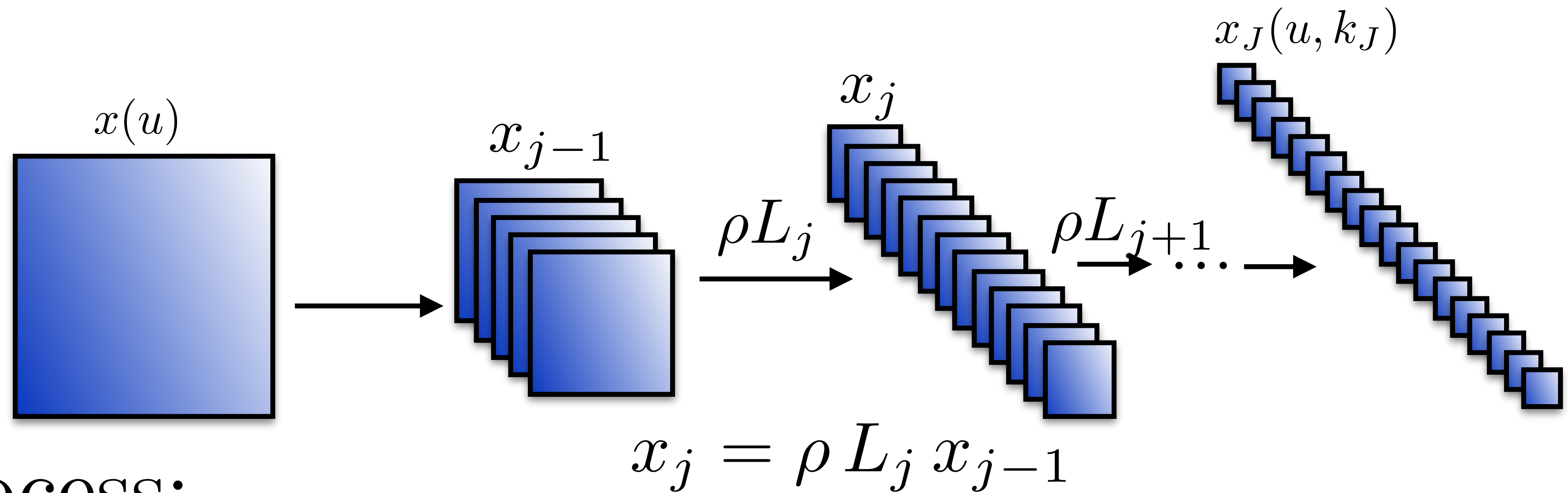
$$\Rightarrow \quad \|\rho L_j x_{j-1} - \rho L_j x'_{j-1}\| \geq \epsilon \ \text{ if } \ f(x) \neq f(x') \ .$$

$\Rightarrow$ contract in directions along which f remains constant.

From Translations to Symmetries

- The value of f remains constant along an orbit $\{g.x_{j-1}\}_{g \in G}$ of a group G of symmetries.



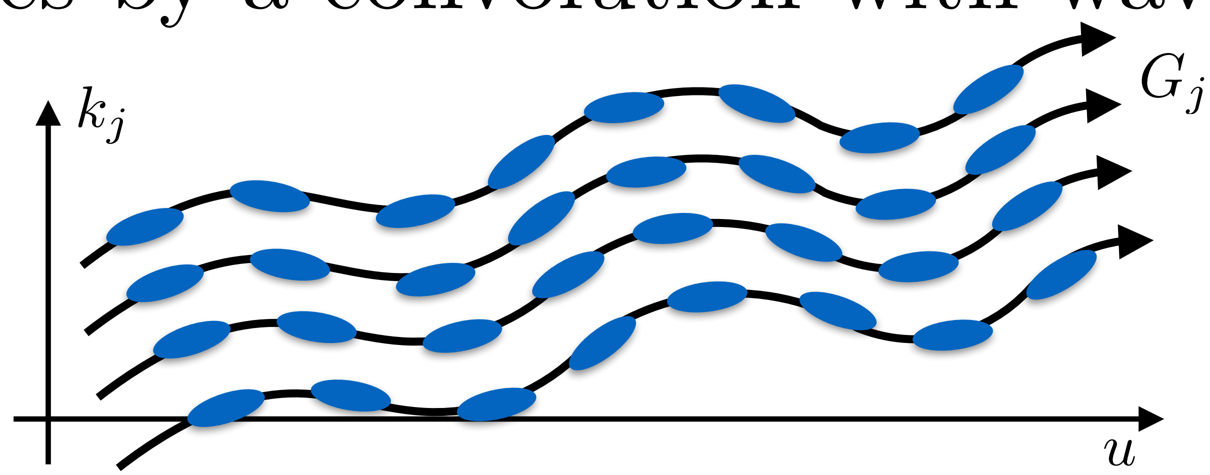


- A two step process:

ρL_j transforms the orbit of x_{j-1} in a parallel transport in x_j:

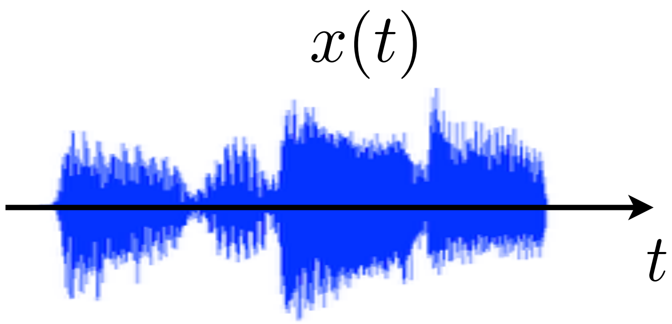
$$g.x_j(v) = x_j(g.v) \; .$$

ρL_{j+1} linearizes by a convolution with wavelets along fibers

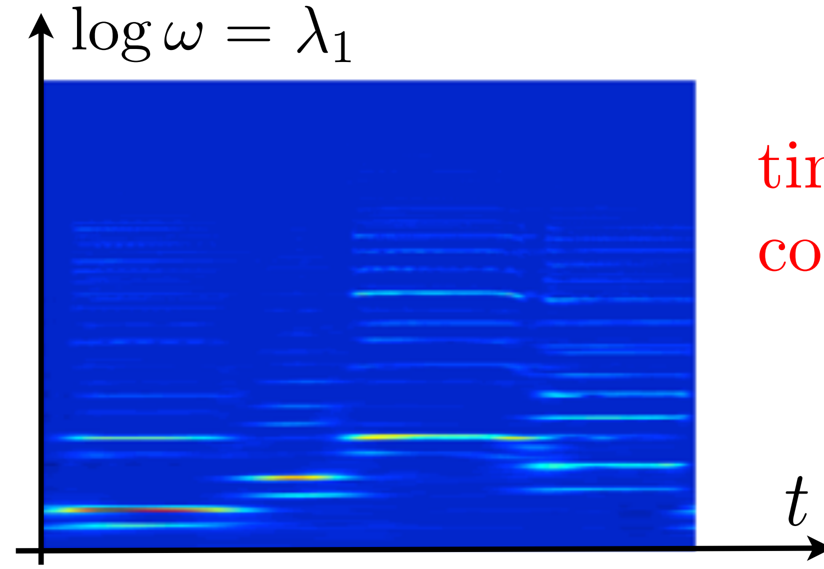


Time-Frequency Fibers

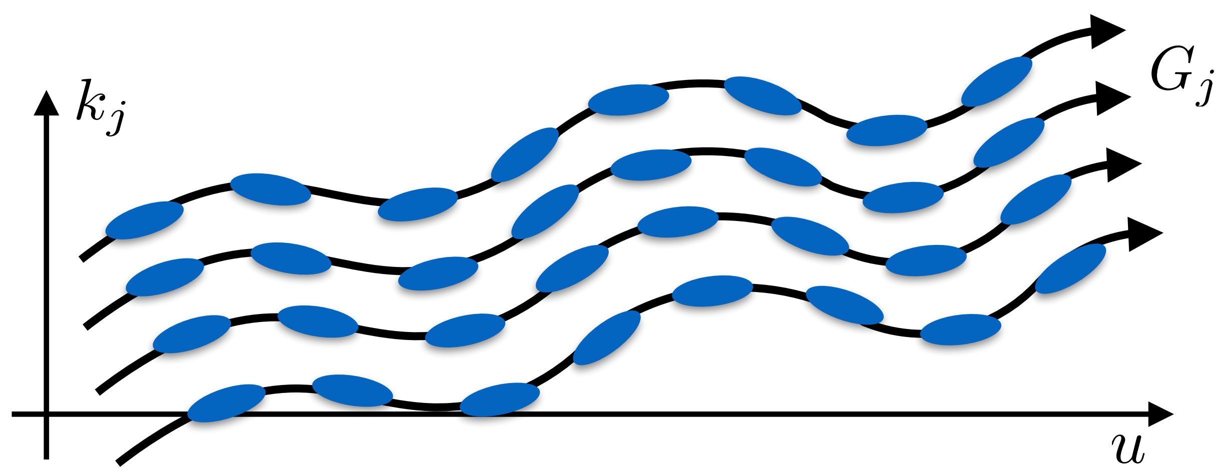
$$x_1(t, \lambda_1) = |x \star \psi_{\lambda_1}(t)|$$



time convolutions

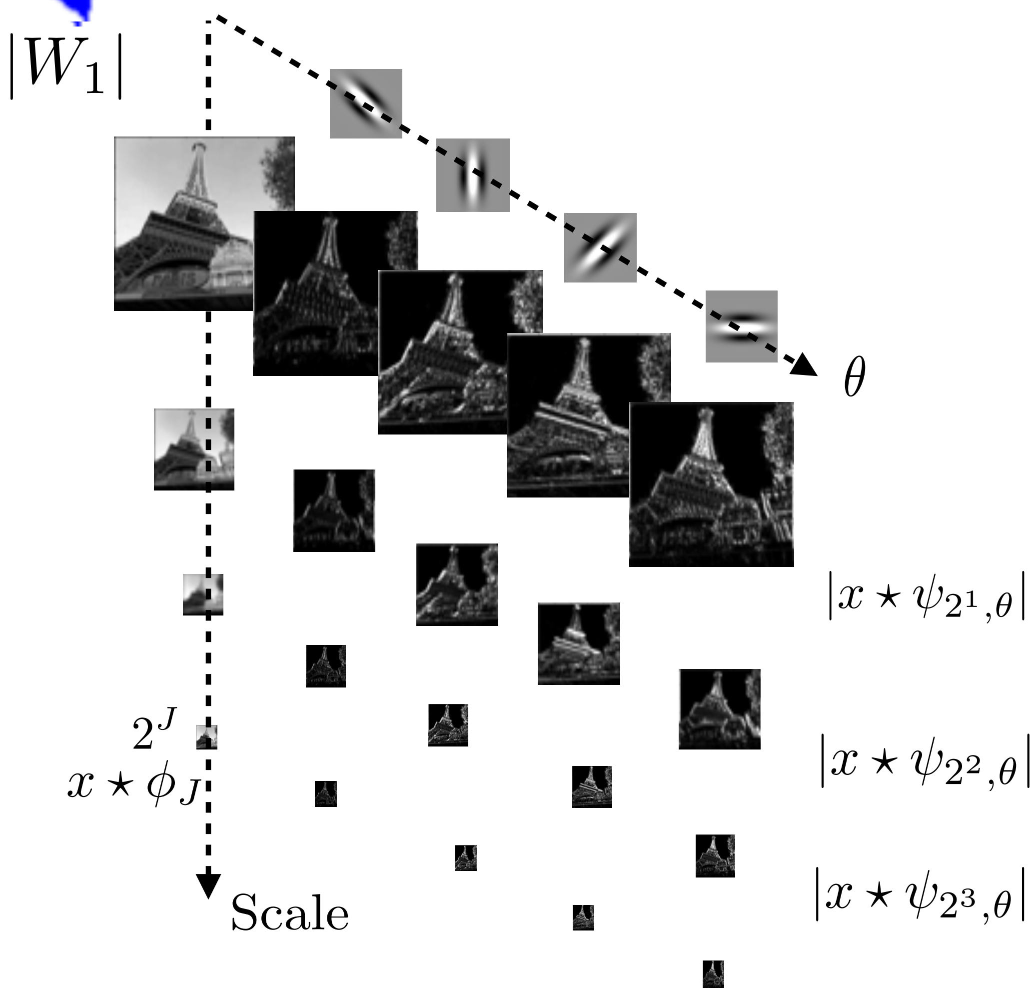


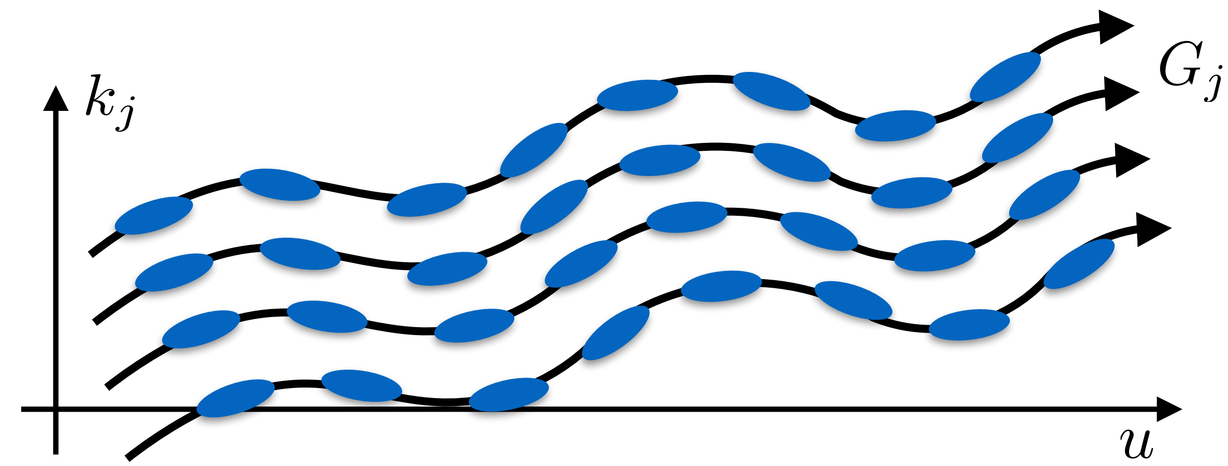
time-frequency convolutions



- Applied to audio classification

Scale-Rotation-Translation Fibers





Scaling and rotations defines a parallel transport in $(u, \theta, 2^j)$

Linear covariant operators: convolutions on the group

- Applied to object recognition

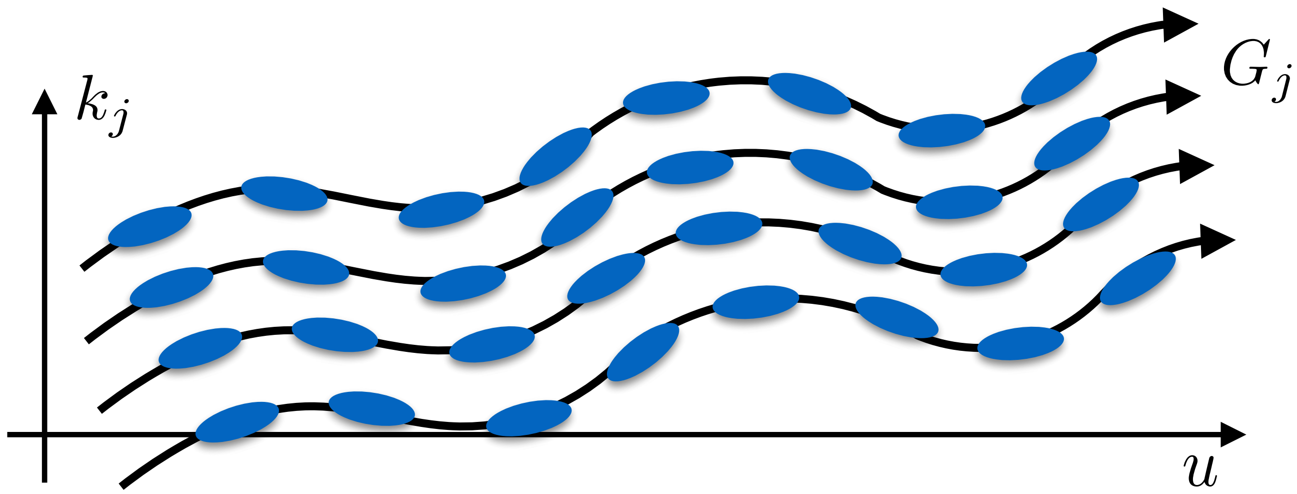
Separate Support Vectors

- Support vectors are pairs x_{j-1}, x'_{j-1} with

$$\|x_{j-1} - x'_{j-1}\| \approx \epsilon \;\text{ and }\; f(x) \neq f(x') \;.$$

Their distance must not be reduced.

- The operator ρL_j must separate them in different fibers:



$\Rightarrow$ sparse representations along fibers

$\Rightarrow$ the rows of L_j encodes the support vectors

Memory of discriminative patterns

Complex optimization

- The operators L_j have many roles:
  - Transform symmetries into transport within network layers
  - Convolutions along fibers to linearize symmetries and reduce dimensions
  - Separate support vectors along different fibers: sparsity

- Difficult to separate these roles when analyzing learned networks

Conclusions

- Deep neural networks have spectacular high-dimensional approximation capabilities.
- They seem to compute hierarchical invariants of complex symmetries
- They store memory
- Neurophysiological models of audition and vision
- Outstanding mathematical problem to understand them:
  notions of complexity, regularity, approximation theorems…