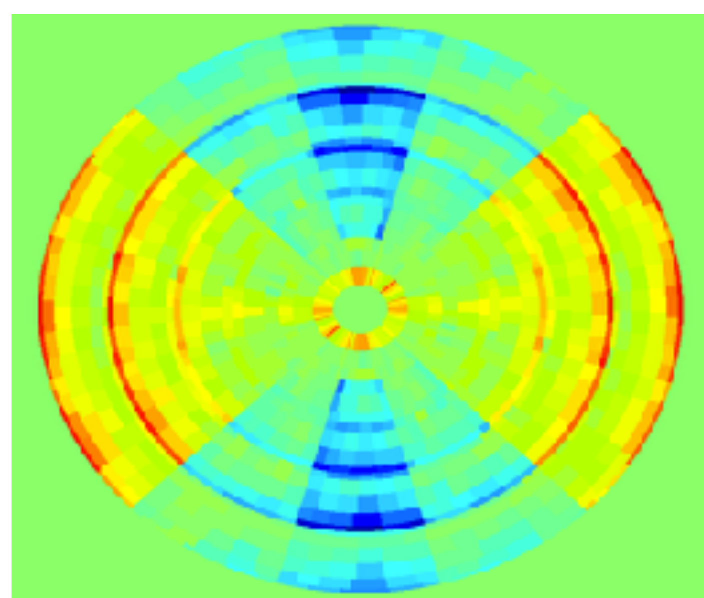
Organizing Deep Networks

Edouard Oyallon



advisor: Stéphane Mallat

following the works of Laurent Sifre, Joan Bruna, …

collaborators: Eugene Belilovsky, Sergey Zagoruyko, Bogdan Cirstea, Jörn Jacobsen, …

Classification of signals

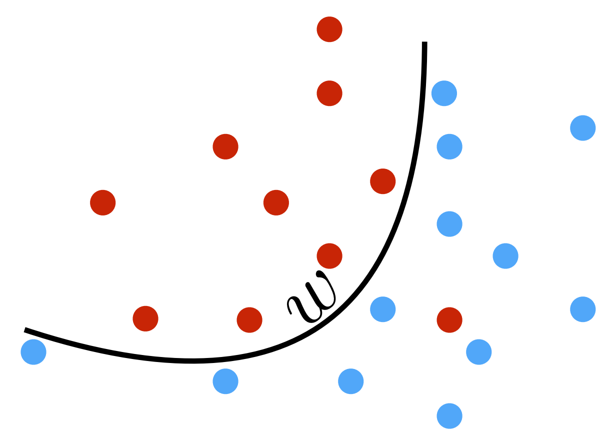
- Let $n > 0$, $(X, Y) \in \mathbb{R}^n \times \mathcal{Y}$ random variables
- **<u>Problem</u>**: Estimate $\hat{y}$ such that $\hat{y} = \arg\inf_{\tilde{y}} \mathbb{E}(|\tilde{y}(X) - Y|)$
- We are given a training set $(x_i, y_i) \in \mathbb{R}^n \times \mathcal{Y}$ to build $\hat{y}$
- Say one can write $\hat{y} = \text{Classifier}(\Phi x)$, Classifier being built with $(\Phi x_i, y_i)$
- 3 ways to build Φ:

Supervised	Unsupervised	Predefined
$(x_i, y_i)_i$	$(x_i)_i$	Geometric priors

$\mathcal{Y} = \{\cdot, \cdot\}$

$n = 2$

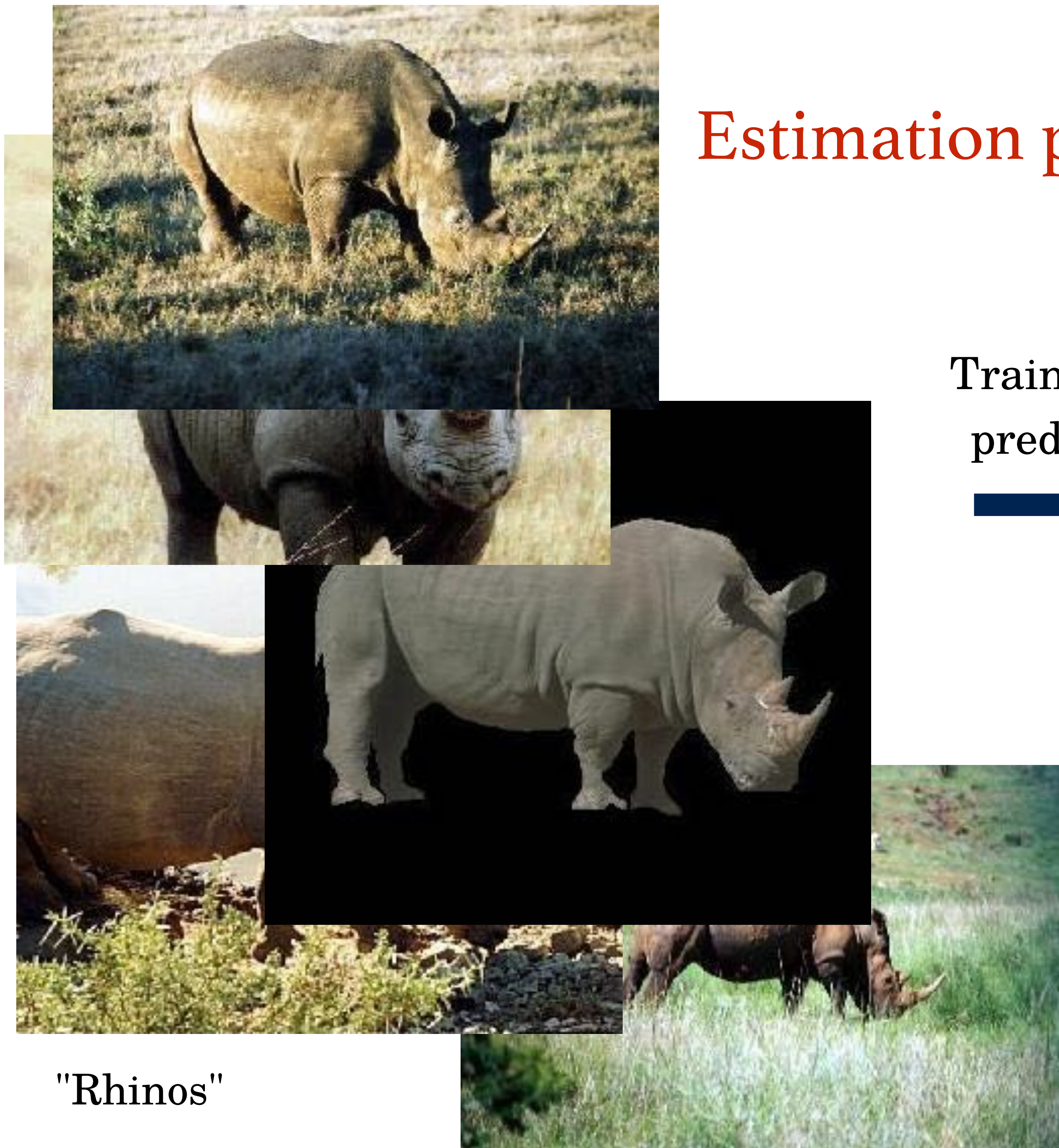
Classifier w



High Dimensional classification

$$(x_i, y_i) \in \mathbb{R}^{224^2} \times \{1, ..., 1000\}, i < 10^6 \longrightarrow \hat{y}(x)?$$

Estimation problem



"Rhinos"

Training set to predict labels

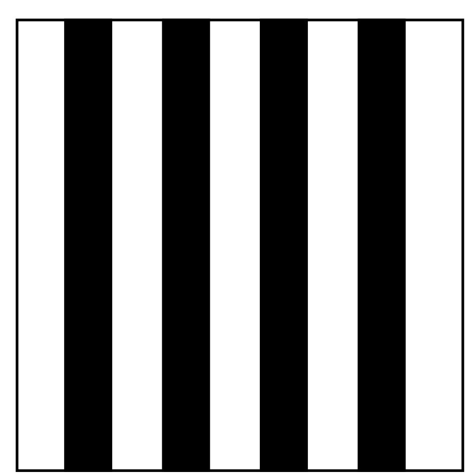
"Rhino"

Not a "rhino"

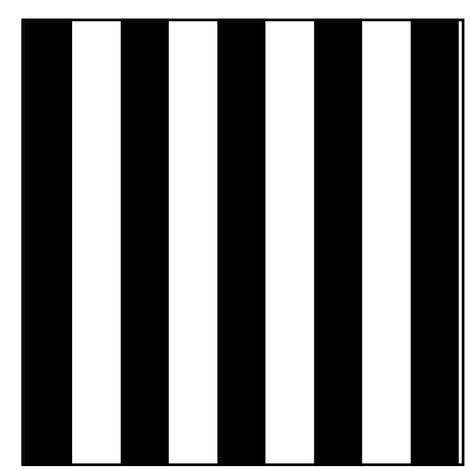
High-dimensional variabilities

- **Claim**: In $\mathbb{R}^n, n \gg 1$, the variance is huge.
  Ex.:
  $$X \sim \mathcal{N}(0, I_n) \text{ then } \exists C > 0, \forall n, \mathbb{P}(\|X\| \geq t)) \leq 2e^{-\frac{t^2}{Cn}}$$
  $$\mathbb{E}(X) = 0$$
- **Claim**: Small deformations (not parametric) can have huge effects:
  Ex.: $x \in L^2(\mathbb{R}^n)$,$\tau \in \mathcal{C}^\infty$ define $L_\tau x(u) = x(u - \tau(u))$
  $$\tau(u) = \epsilon, \mathcal{C} \subset \mathbb{R}^2, \|1_\mathcal{C} - L_\tau 1_\mathcal{C}\| = 2$$
- The variance is **high**, and the bias is **difficult to estimate**. There are also **few available samples**… How to handle that?

$$\|x - y\|_2 = 2 \longleftarrow$$



x

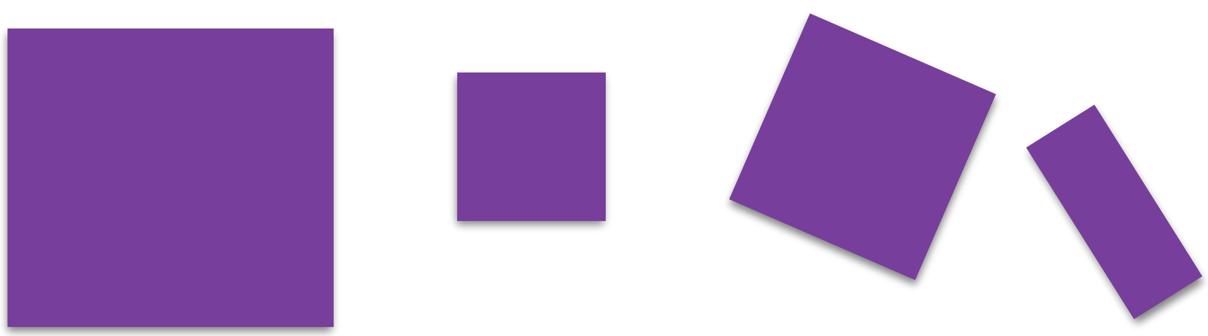


y

Image variabilities

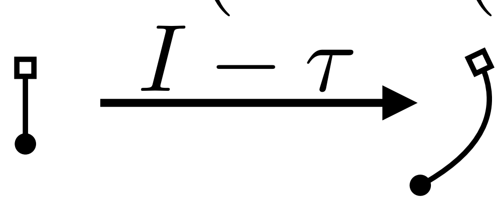
Geometric variability

Groups acting on images:
translation, rotation, scaling



Other sources : luminosity, occlusion,
small deformations

$$L_\tau x(u) = x(u - \tau(u)), \tau \in \mathcal{C}^\infty$$



Class variability

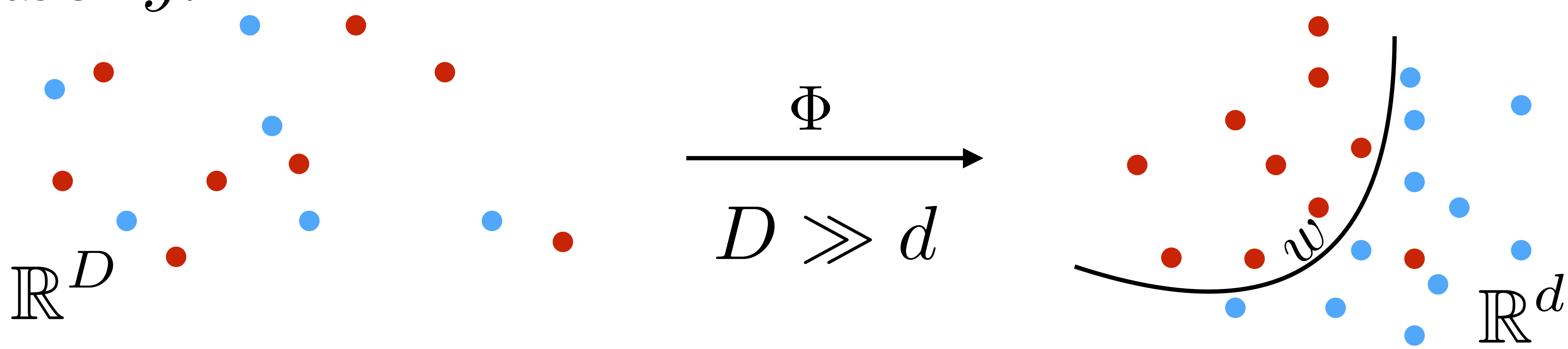
Intraclass variability
Not informative

Extraclass variability

High variance: how to reduce it?

Fighting the curse of dimensionality

- **Objective:** building a representation Φx of x such that a simple (say euclidean) classifier $\hat{y}$ can estimate the label y:

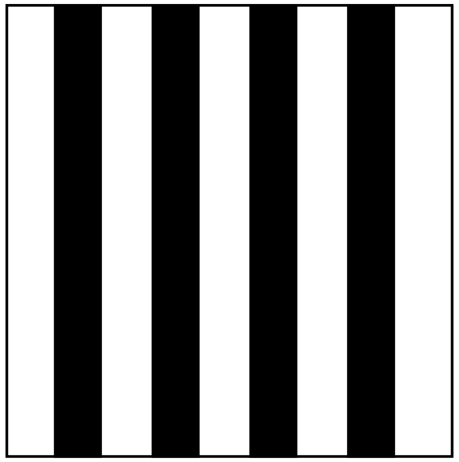
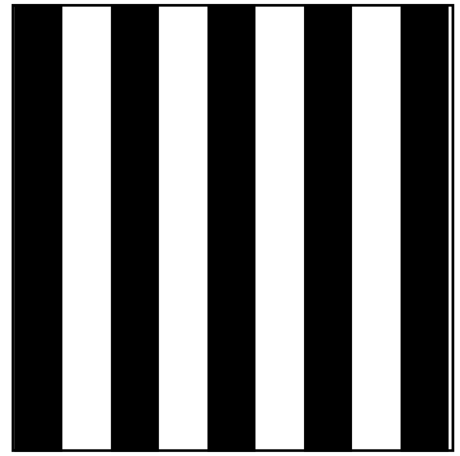


- Designing Φ consist of building an approximation of a low dimensional space which is regular with respect to the class:

$$\|\Phi x - \Phi x'\| \lll 1 \Rightarrow \hat{y}(x) = \hat{y}(x')$$

- **Necessary** dimensionality reduction

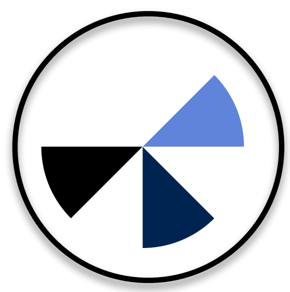
Translation

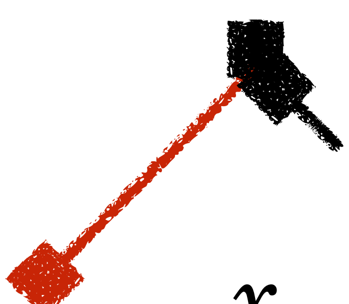
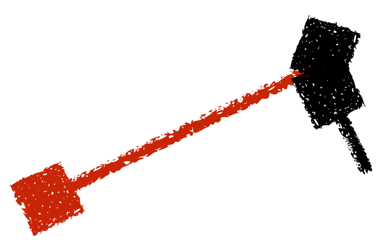
x y

$$\|x - y\|_2 = 2$$

Rotation

Averaging is the key to get invariants

x y

Averaging makes euclidean distance meaningful in high dimension

An example: Invariance to translation

$$\boxed{\text{Translation operator} \quad L_a x(u) = x(u-a)}$$

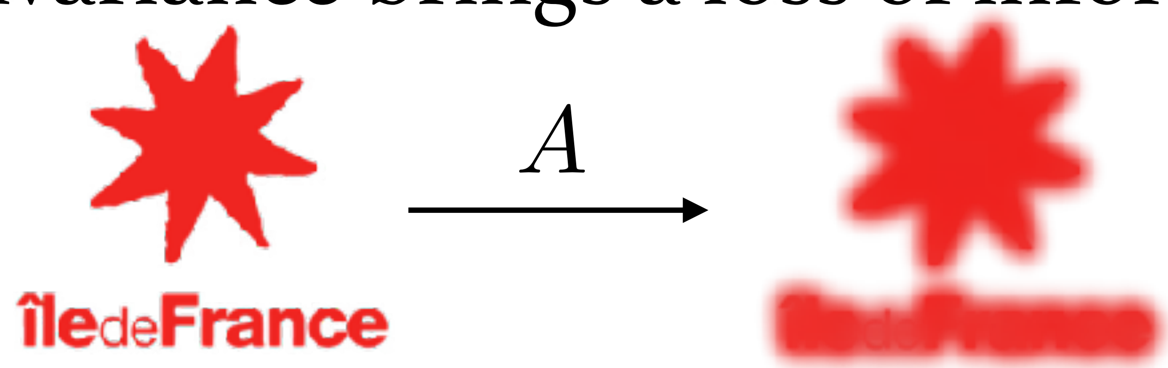
- In many cases, one wish to be invariant globally to translation, a simple way is to perform an averaging:

$$Ax = \int L_a x da = \int x(u) du$$

It's the 0 frequency!

$$AL_a = A$$

- Even if it can be localized, the averaging keeps the low frequency structures: the invariance brings a loss of information!

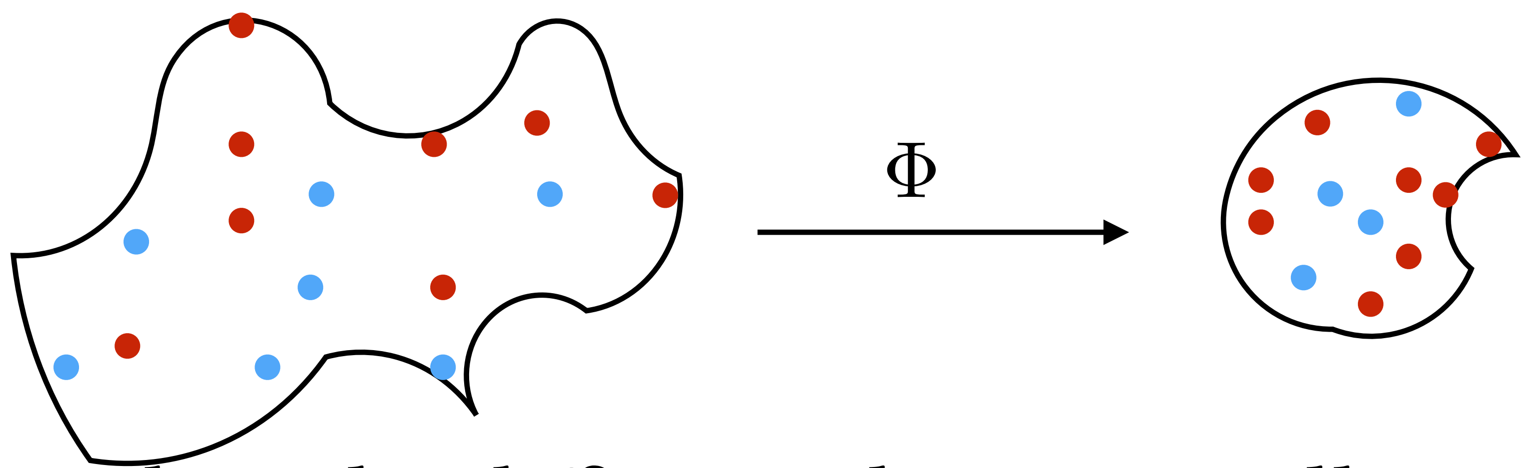


- Bias issue! How do we recover the missing information?

Necessary mechanism: Separation - Contraction

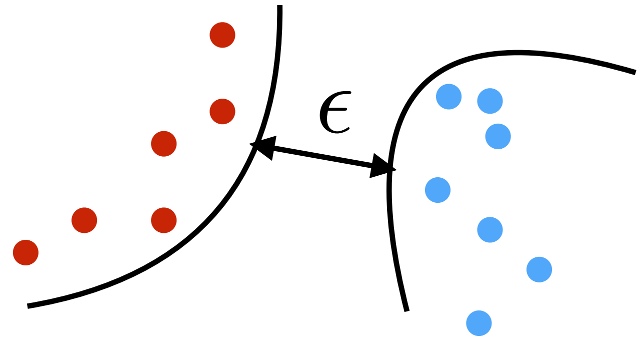
- In high dimension, typical distances are huge, thus an appropriate representation must contract the space:

$$\|\Phi x - \Phi x'\| \leq \|x - x'\|$$



- While avoiding the different classes to collapse:

$$\exists \epsilon > 0, y(x) \neq y(x') \Rightarrow \|\Phi x - \Phi x'\| \geq \epsilon$$



Deep learning: Technical breakthrough

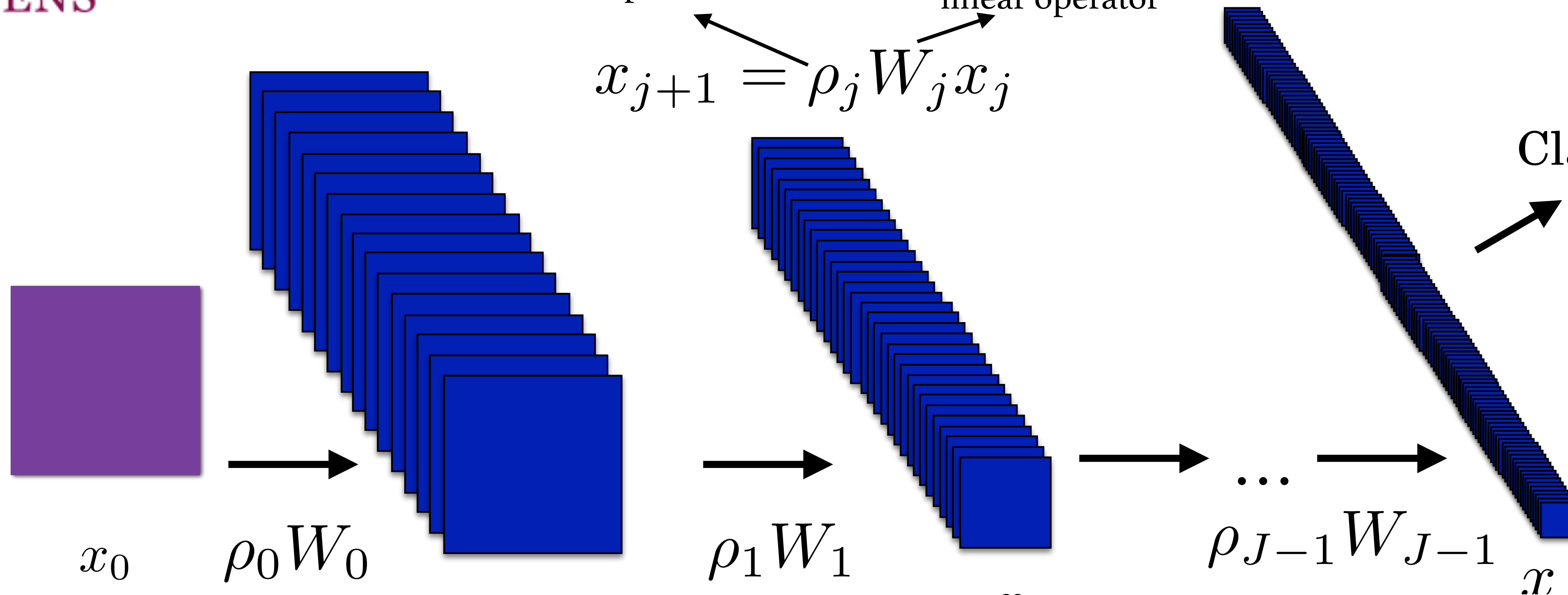
- Deep learning has permitted to <u>**solve**</u> a large number of task that were considered as extremely challenging for a computer.
- The technique that is used is **generic** and its success implies that it reduces those sources of variability.
- Previous properties hold for deep learning.
- **How, why?**

DATA

non-linear operator

linear operator

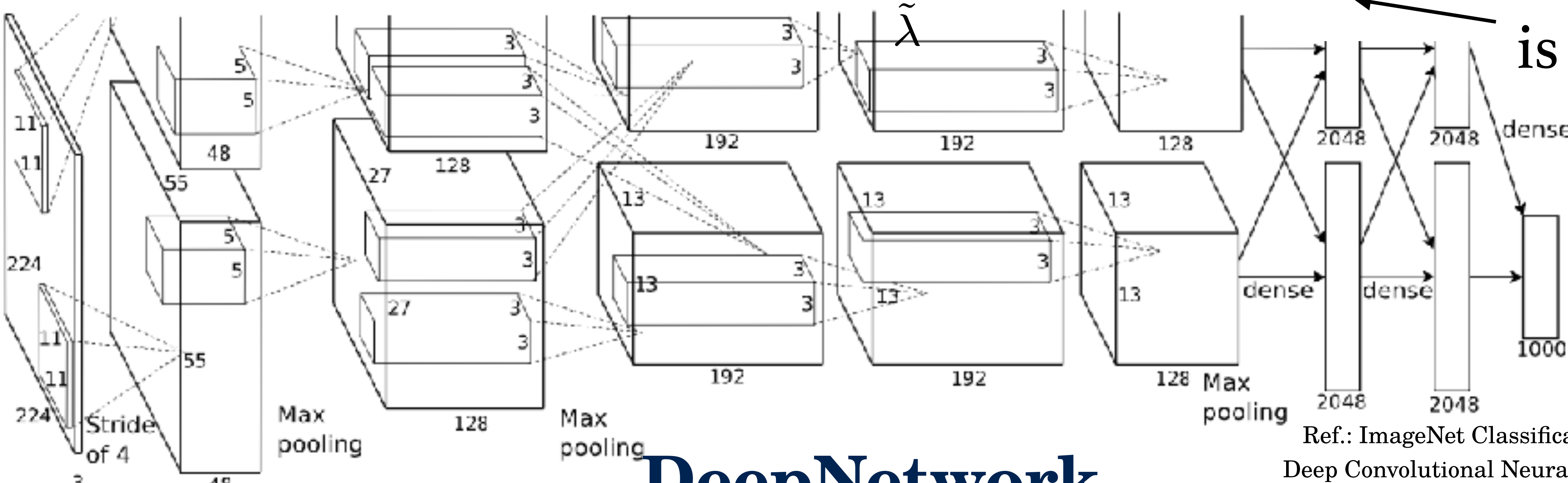
$$x_{j+1} = \rho_j W_j x_j$$



x_1 x_2

$$x_{j+1}(u,\lambda) = \rho\Big(\sum_{\tilde\lambda} x_j(.,\tilde\lambda) \star w_{j,\lambda,\tilde\lambda}(u)\Big)$$

The kernel is learned

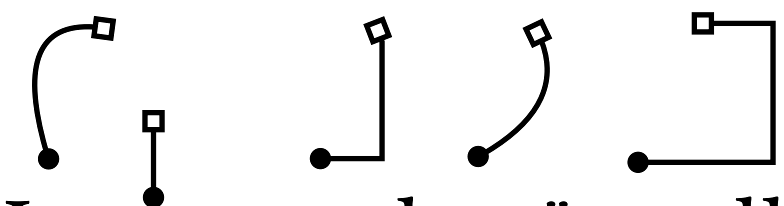


DeepNetwork

Ref.: ImageNet Classification with Deep Convolutional Neural Networks. A Krizhevsky et al.

Why mathematics about deep learning are important

- **Pure black box.** Few mathematical results are available. Many rely on a "manifold hypothesis". Clearly wrong:
  Ex: stability to diffeomorphisms



- No **stability results.** It means that "small" variations of the inputs might have a large impact on the system. And this happens.
  Ref.: Intriguing properties of neural networks. C. Szegedy et al.

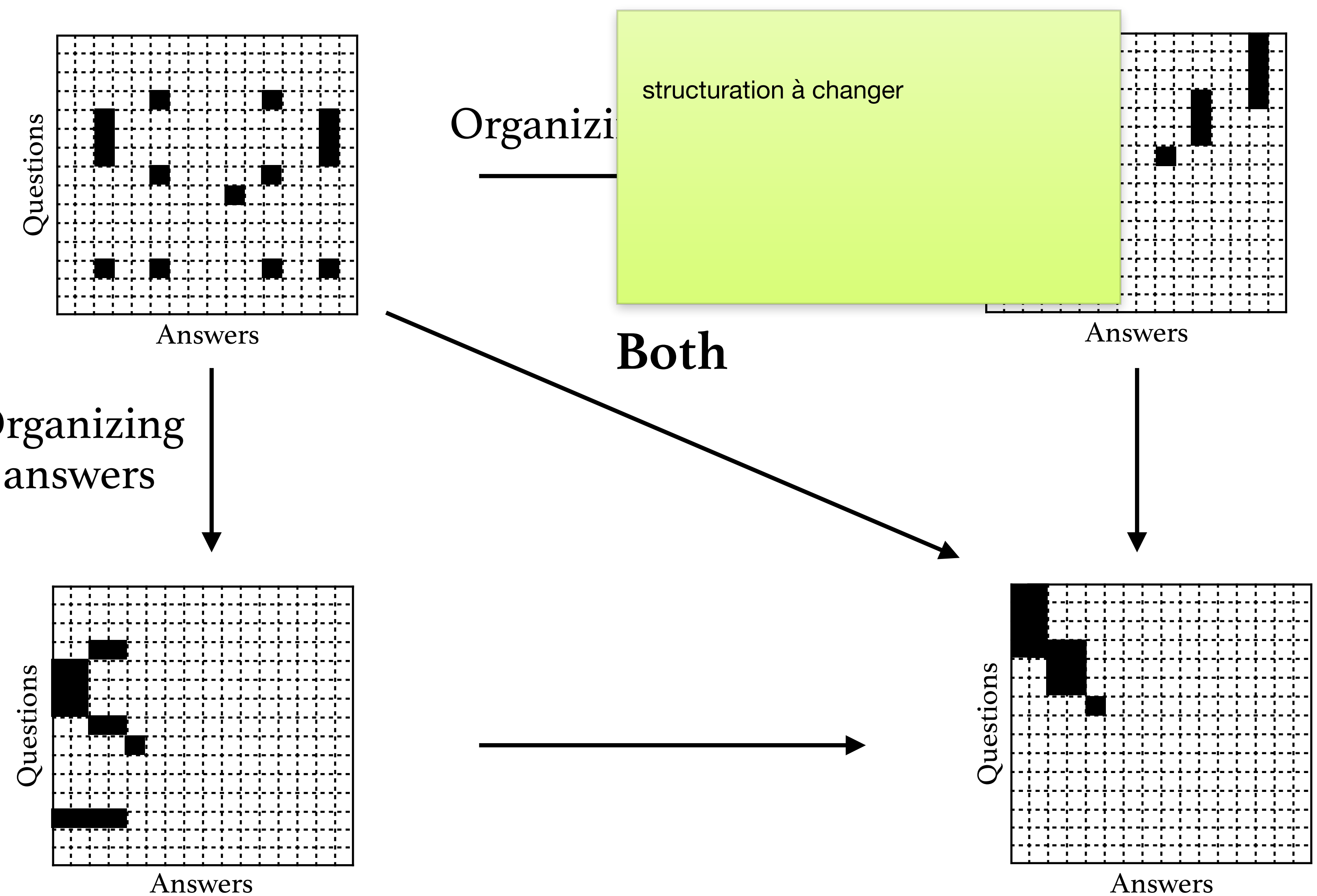
- **No generalisation result.** Rademacher complexity can not explain the generalization properties.
  Ref.: Understanding deep learning requires rethinking generalization C. Zhang et al.

- Shall we learn **each layer** from scratch? (geometric priors?) The deep cascade makes features are hard to **interpret**
  Ref.: Deep Roto-Translation Scattering for Object Classification. EO and S Mallat

Organization is a key

- Consider a problem of questionnaires: people answer to 0 or 1 to some question. What does structuration means?

Ref.: Harmonic Analysis of Digital Data Bases
Coifman R. et al.

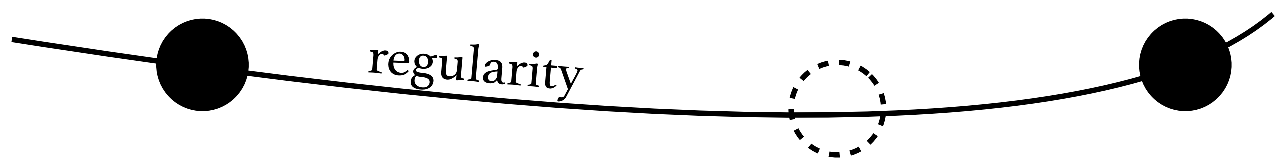


In general, works tackle only one of the aspect

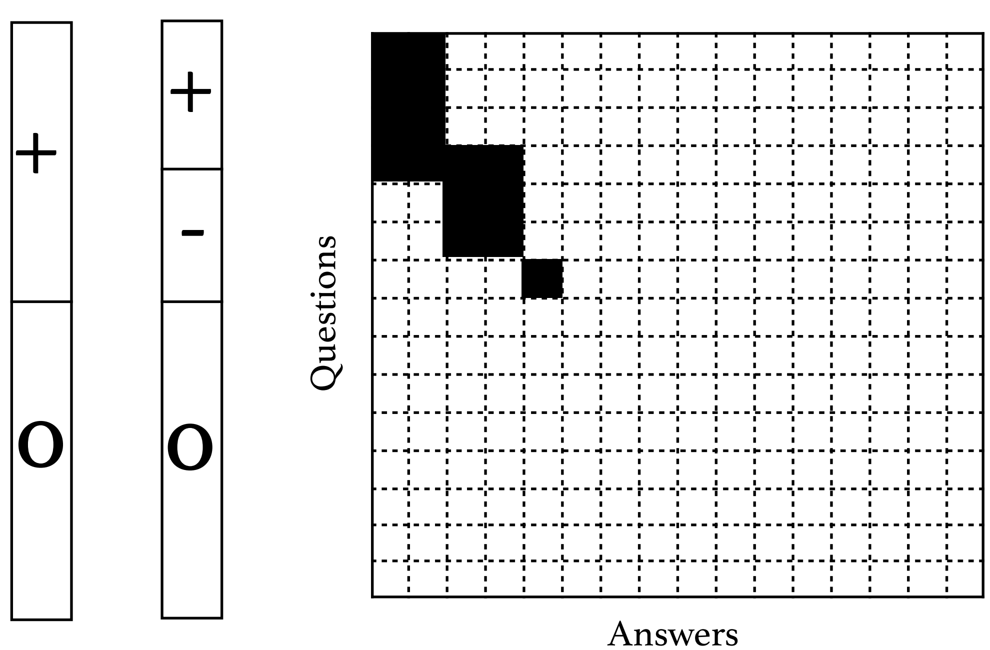
neighbours become meaningful: **local metrics**

Organization permits creation of invariance

- As (all) the sources of regularities are obtained, interpolating new points is possible (**in statistical terms: generalisation property!**)



- In the previous case, one can build a discriminative and invariant representation: Haar wavelets on graphs for example.



Ref.: Harmonic Analysis of Digital Data Bases
Coifman R. et al.

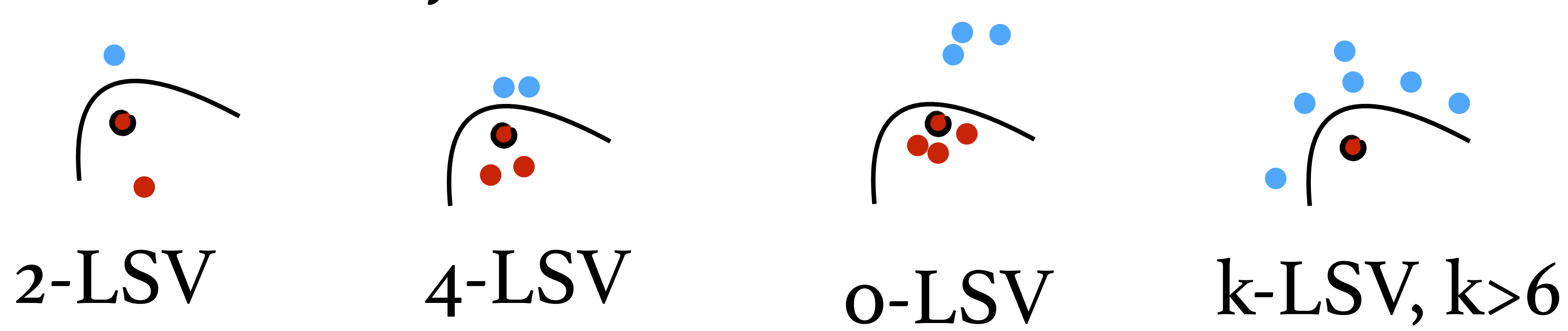
Organising the CNN representation: Local Support Vectors

Ref.: Building a Regular Decision Boundary with Deep Networks
EO

- Let's consider a CNN of depth J.

Local dimension is intractable!

- Local Support Vectors of order k at depth j: representations at depth j that are well classified by a k-NN but not by a l-NN for l<k



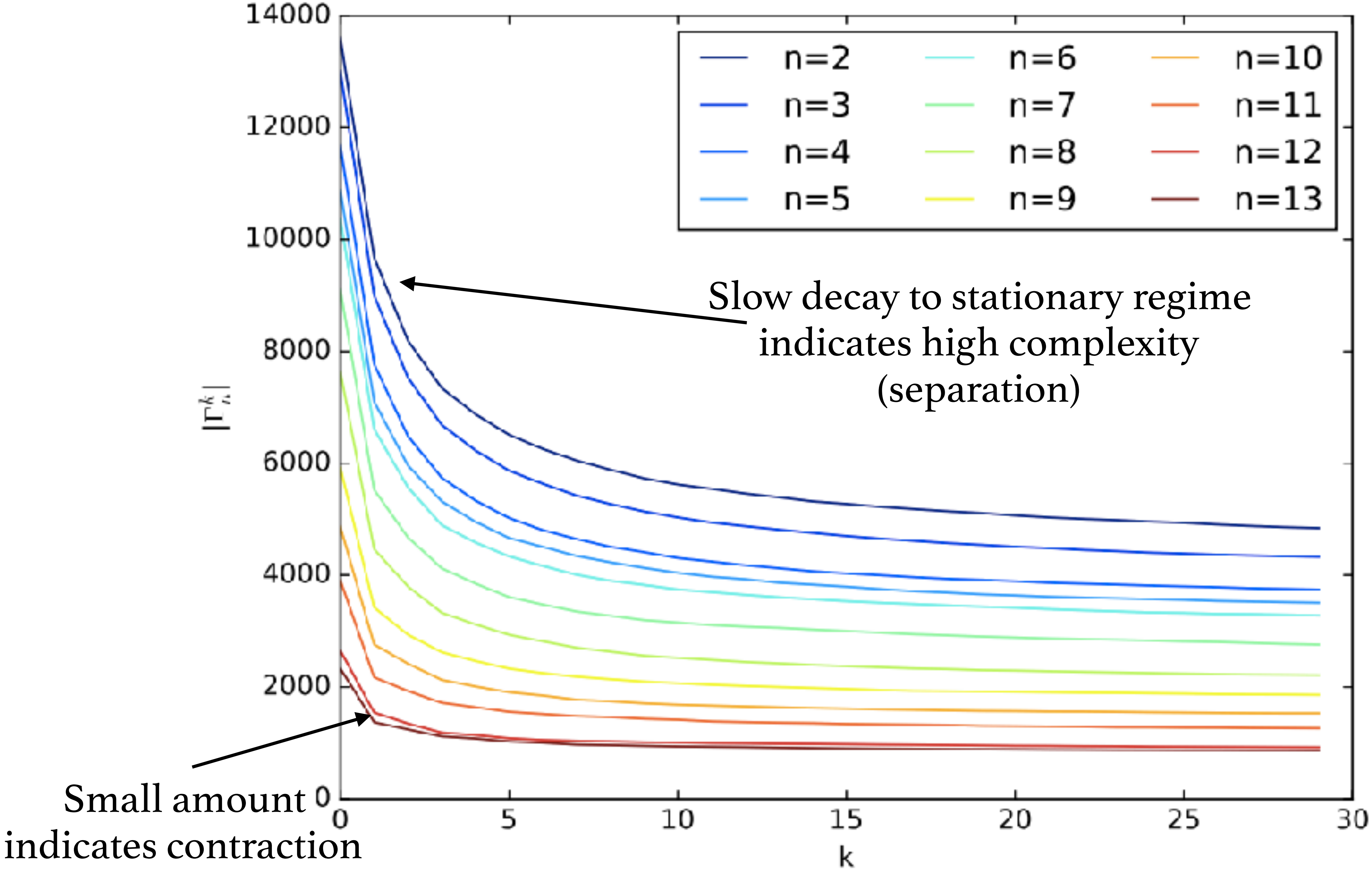
- They give a measure of the separation-contraction via:

$$\Gamma_j^{k+1} = \{x_j \in \Gamma_j^k | \text{card}\{y(x_j^{(l)}) \neq y(x_j^{(l)}), l \leq k+1\} > \frac{k}{2}\}$$

$x_j^{(l)}$: l-NN at depth j

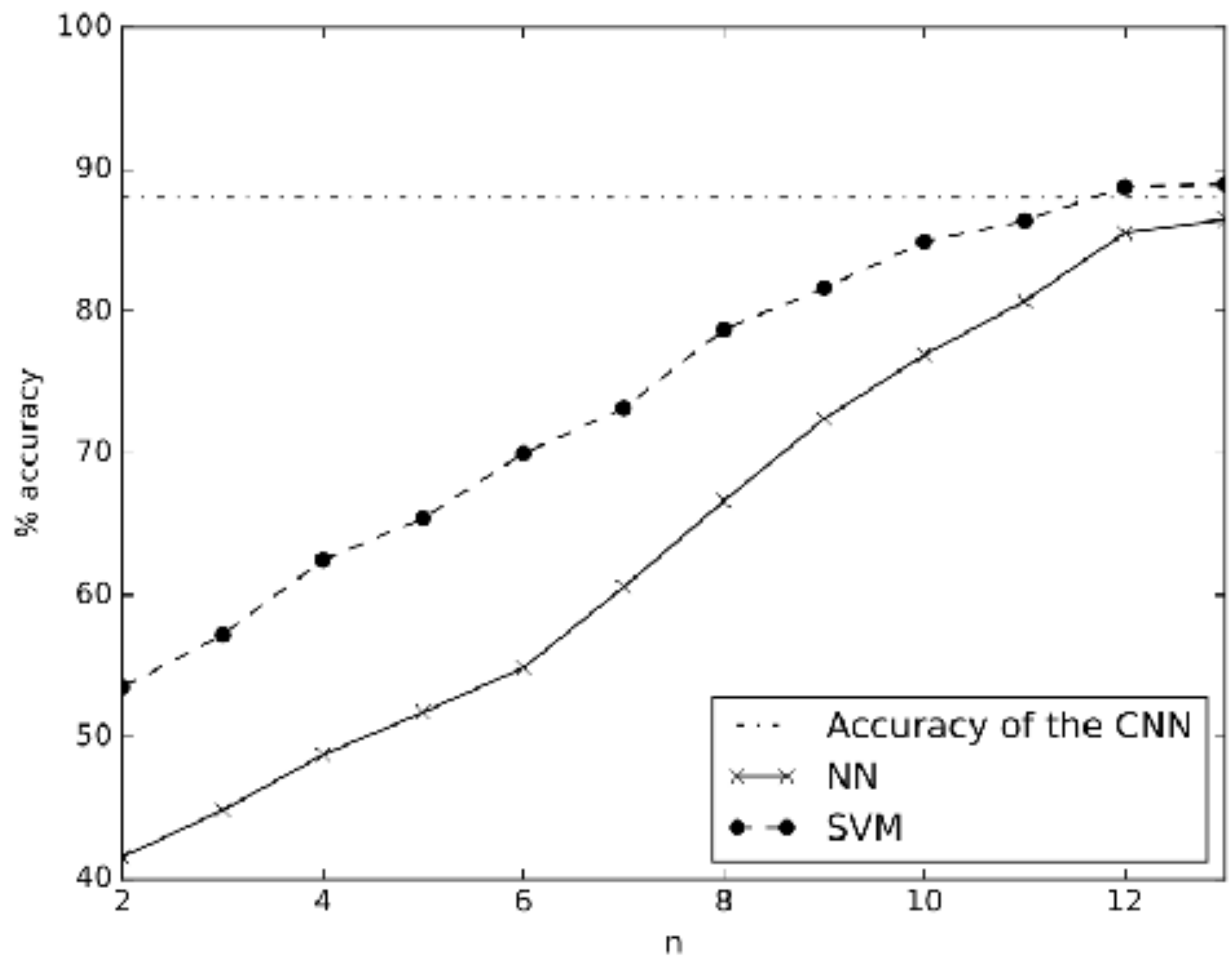
Complexity measure

of k-local support vectors at different depth n



An organisation of the representation

- There is a progressive localisation which explains why a 1-NN (or a Gaussian SVM) works better with depth:



linear metrics are more meaningful in low dimension

- How do the representation got localized? **Necessary** variability reduction

Identifying the variabilities?

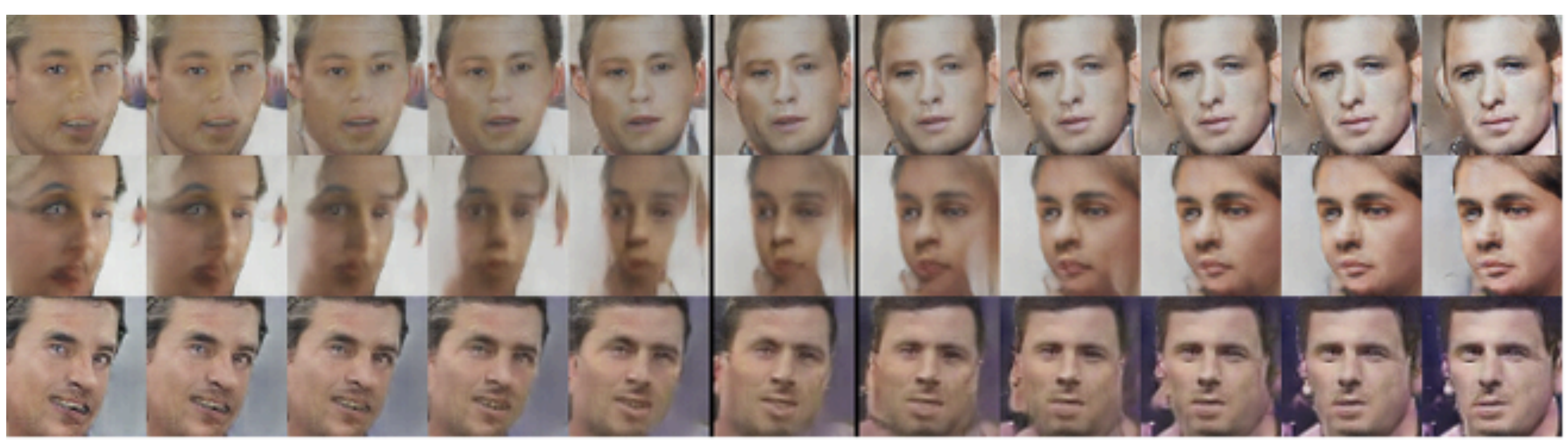
- Several works showed a Deepnet exhibits some covariance:

(c) Object color (d) Background color

(a) Lighting (b) Scale

Ref.: Understanding deep features with computer-generated imagery, M Aubry, B Russel

- Manifold of faces at a certain depth:



Ref.: Unsupervised Representation Learning with Deep Convolutional GAN, Radford, Metz & Chintalah

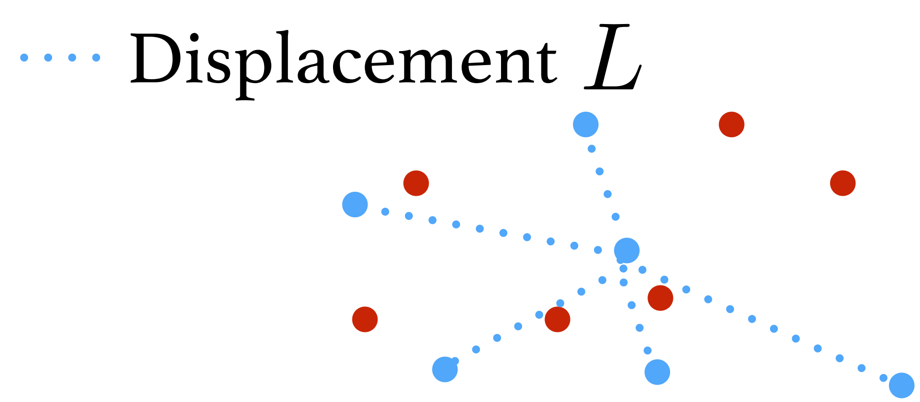
- Can we use these?

Linearizing variabilities

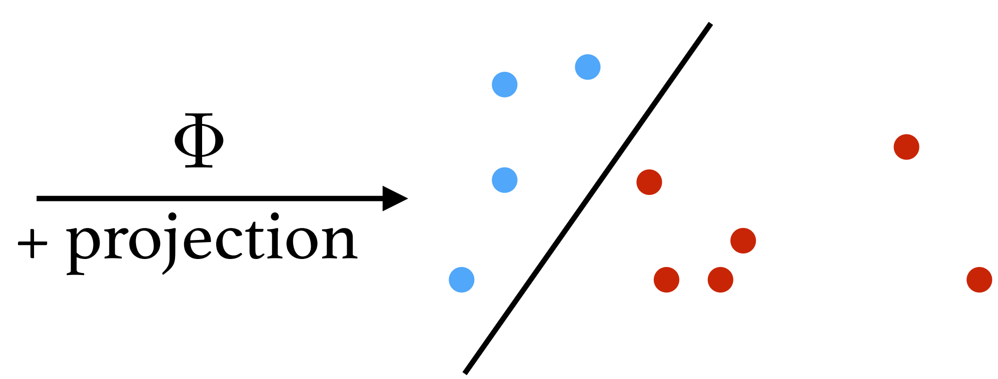
- Weak differentiability property:

$$\sup_L \frac{\|\Phi L x - \Phi x\|}{\|L x - x\|} < \infty \Rightarrow \exists \text{ "weak" } \partial_x \Phi$$

$$\Rightarrow \Phi L x \approx \Phi x + \underbrace{\partial_x \Phi L}_{\text{A linear operator}} + o(\|L\|)$$



- A linear projection (to kill L) build an invariant



example:
Scattering Transform

Symmetry group hypothesis

Ref.: Understanding deep convolutional networks
S Mallat

- To each classification problem corresponds a canonic and unique symmetry group G: ⟵ High dimensional

$$\forall x, \forall g \in G, \Phi x = \Phi g.x$$

- We hypothesise there exists **Lie** groups and CNNs such that:

$$G_0 \subset G_1 \subset ... \subset G_J \subset G$$

$$\forall g_j \in G_j, \phi_j(g_j.x) = \phi_j(x) \text{ where } x_j = \phi_j(x)$$

- Examples are given by the euclidean group:

$$G_0 = \mathbb{R}^2, G_1 = G_0 \ltimes SL_2(\mathbb{R})$$

Structuring the input with the Scattering Transform

- Scattering Transform S_J is a local descriptor of neighbourhood of amplitude 2^J.

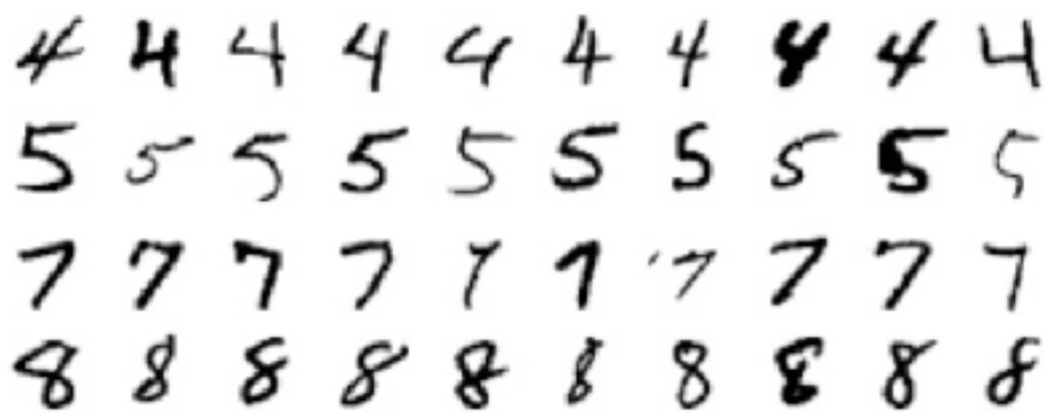
- It is a representation built via geometry with limited learning. (~SIFT)

Ref.: Invariant Convolutional Scattering Network, J. Bruna and S Mallat

- Successfully used in several applications:

All variabilities are known

  - Digits



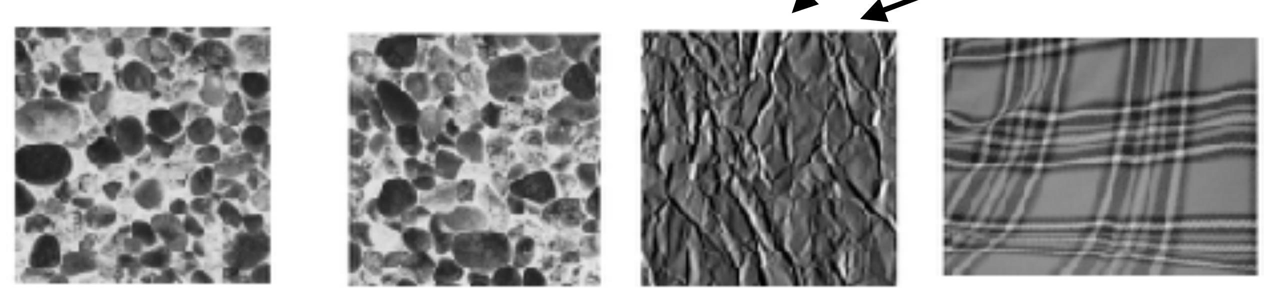
Small deformations

+Translation

Rotation+Scale

Ref.: Rotation, Scaling and Deformation Invariant Scattering for texture discrimination, Sifre L and Mallat S.

  - Textures



Wavelets

- Wavelets help to describe signal structures. ψ is a wavelet iff

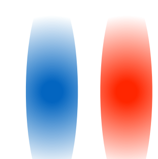
$$\psi \in \mathcal{L}^2(\mathbb{R}^2, \mathbb{C}) \text{ and } \int_{\mathbb{R}^2} \psi(u)du = 0$$

- They are chosen localised in space and frequency.

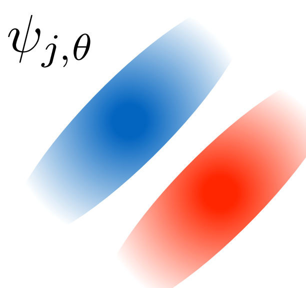
- Wavelets can be dilated in order to be a **multi-scale** representation of signals, **rotated** to describe rotations.

$$\psi_{j,\theta} = \frac{1}{2^{2j}} \psi\left(\frac{r_\theta(u)}{2^j}\right)$$

ψ



$\psi_{j,\theta}$



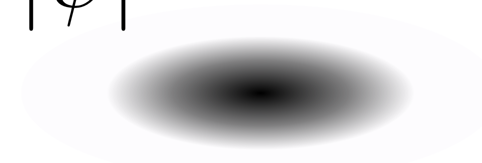
- Design wavelets selective to an **informative** variability.

Isotropic

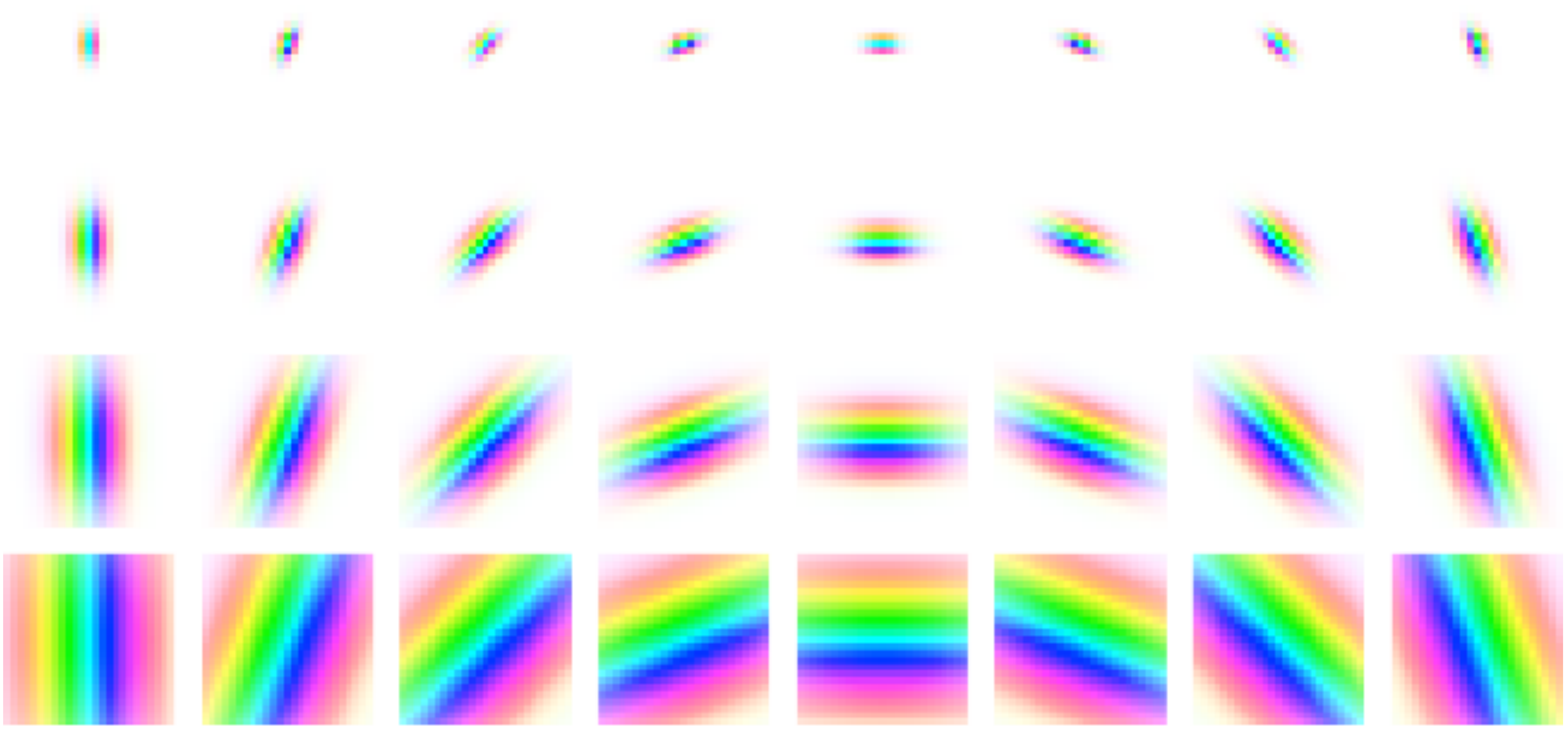
$|\hat{\psi}|$

VS

$|\hat{\psi}|$



Non-Isotropic



$$\psi(u) = \frac{1}{2\pi\sigma} e^{-\frac{\|u\|^2}{2\sigma}} (e^{i\xi . u} - \kappa)$$

$$\phi(u) = \frac{1}{2\pi\sigma} e^{-\frac{\|u\|^2}{2\sigma}}$$

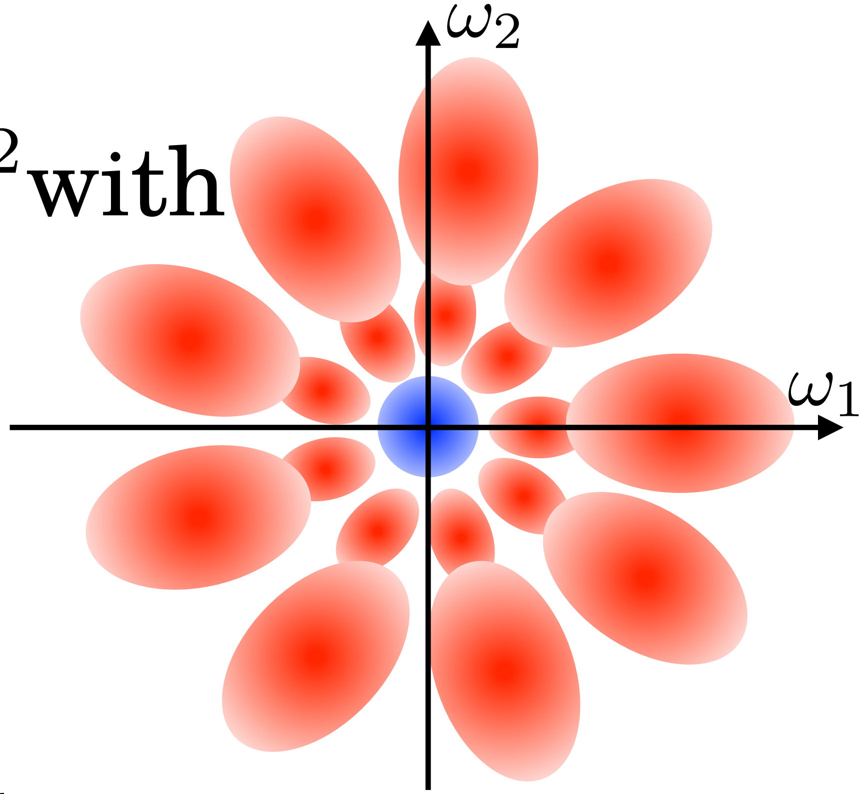
Heisenberg principle!
Good localisation in space and Fourier

(for sake of simplicity, formula are given in the isotropic case)

The Gabor wavelet

Wavelet Transform

- Wavelet transform : $Wx = \{x \star \psi_{j,\theta}, x \star \phi_J\}_{\theta, j \leq J}$



- Isometric and linear operator of L^2, with

$$\|Wx\|^2 = \sum_{\theta, j \leq J} \int |x \star \psi_{j,\theta}|^2 + \int x \star \phi_J^2$$

- Covariant with translation L_a:

$$WL_a = L_a W$$

- Nearly commutes with diffeomorphisms

$$\|[W, L_\tau]\| \leq C \|\nabla \tau\|$$

Ref.: Group Invariant Scattering, Mallat S

- A good baseline to describe an image!

Filter bank implementation of a Fast WT

Ref.: Fast WT, Mallat S, 89

- Assume it is possible to find h and g such that

$$\hat{\psi}_\theta(\omega) = \frac{1}{\sqrt{2}}\hat{g}_\theta(\frac{\omega}{2})\hat{\phi}(\frac{\omega}{2}) \quad \text{and} \quad \hat{\phi}(\omega) = \frac{1}{\sqrt{2}}\hat{h}(\frac{\omega}{2})\hat{\phi}(\frac{\omega}{2})$$

- Set:

$$x_j(u,0) = x \star \phi_j(u) = h \star (x \star \phi_{j-1})(2u) \quad \text{and}$$

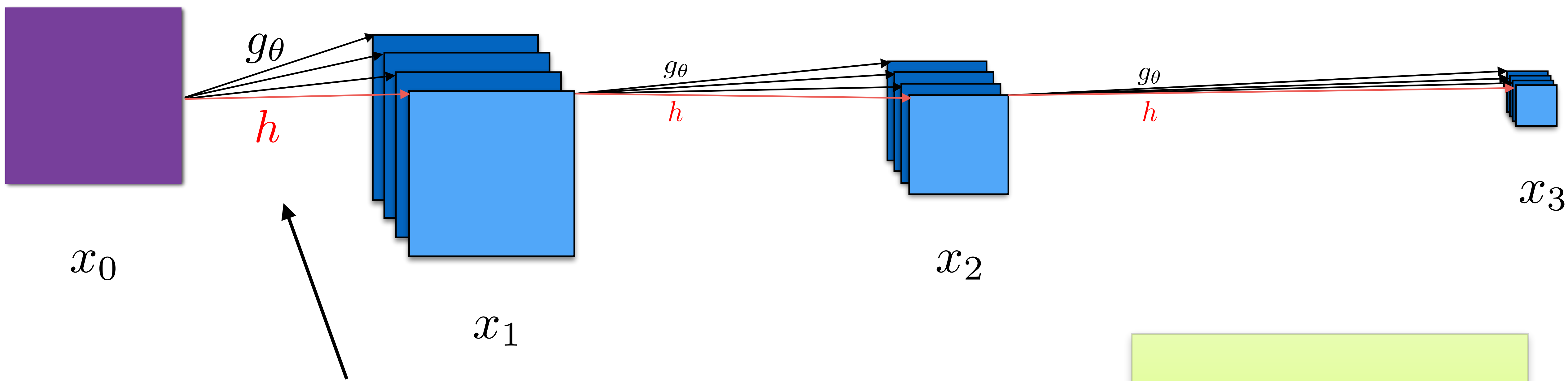
$$x_j(u,\theta) = x \star \psi_{j,\theta}(u) = g_\theta \star (x \star \phi_{j-1})(2u)$$

- The WT is then given by $Wx = \{x_j(.,\theta), x_J(.,0)\}_{j\leq J,\theta}$

- A WT can be interpreted as a **deep cascade** of linear operator, which is approximatively verified for the Gabor Wavelets.

$$\hat\phi_j = \frac{1}{\sqrt{2}}\hat h(\frac{\cdot}{2})\hat\phi_{j-1}$$

$$\hat\psi_{j,\theta} = \frac{1}{\sqrt{2}}\hat g_\theta(\frac{\cdot}{2})\hat\phi_{j-1}$$



There is an oversampling

$h \geq 0$

step by step of the construction(and add modulus also)

Implementation of a WT

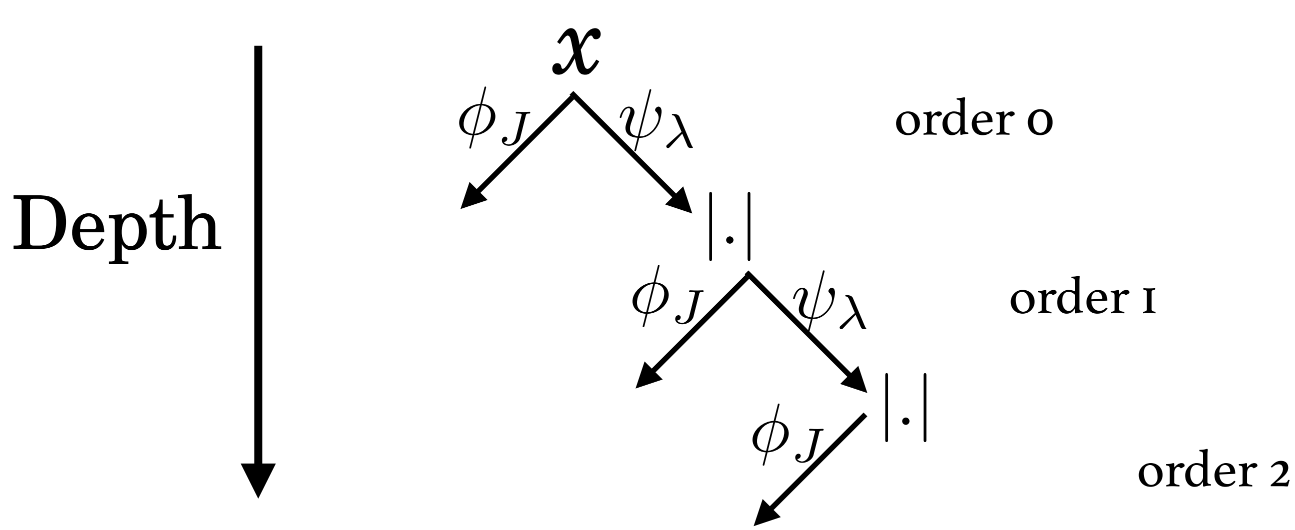
Scattering Transform

- Scattering transform at scale J is the cascading of complex WT with modulus non-linearity, followed by a low pass-filtering:

Ref.: Group Invariant Scattering, Mallat S

$$S_J x = \{x \star \phi_J,\ |x \star \psi_{\lambda_1}| \star \phi_J,\ ||x \star \psi_{\lambda_1}| \star \psi_{\lambda_2}| \star \phi_J\}$$

with $\lambda_i = \{j_i, \theta_i\}, j_i \leq J$

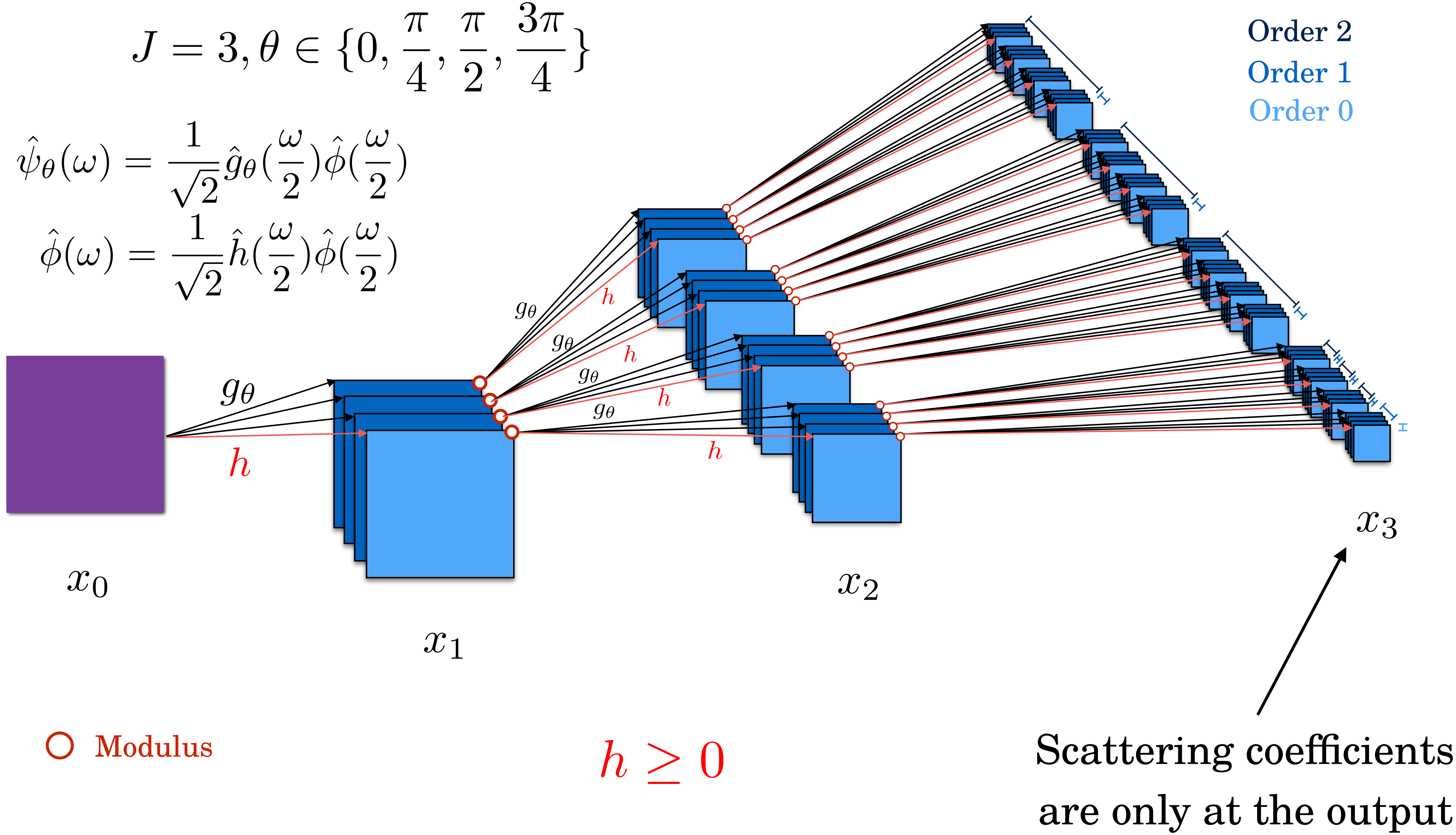


- **Mathematically** well defined for a large class of wavelets.

$$J=3, \theta \in \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$$

$$\hat{\psi}_\theta(\omega)=\frac{1}{\sqrt{2}}\hat{g}_\theta(\frac{\omega}{2})\hat{\phi}(\frac{\omega}{2})$$

$$\hat{\phi}(\omega)=\frac{1}{\sqrt{2}}\hat{h}(\frac{\omega}{2})\hat{\phi}(\frac{\omega}{2})$$



$$h \geq 0$$

Scattering as a CNN

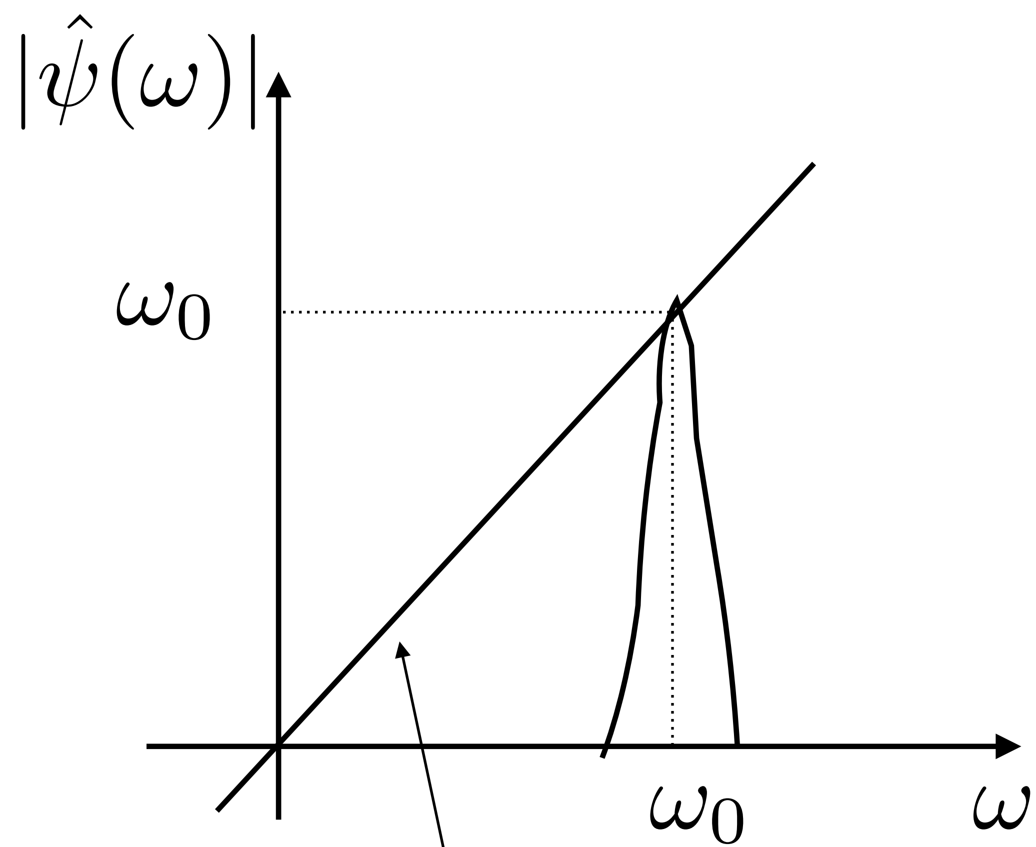
Ref.: Deep Roto-Translation Scattering for Object Classification. EO and S Mallat

Analytic wavelets and modulus?

- For any translations :

Ref.: Group Invariant Scattering, Mallat S

$$
\begin{aligned}
\widehat{L_a x \star \psi}(\omega) &= e^{i\omega^T a}\hat{x}(\omega)\hat{\psi}(\omega)\\
&= \sum_n \frac{(i\omega^T a)^n}{n!}\hat{x}(\omega)\hat{\psi}(\omega)\\
&\approx \sum_n \frac{(i\omega_0^T a)^n}{n!}\hat{x}(\omega)\hat{\psi}(\omega)\\
&= e^{i\omega_0^T a}\widehat{x \star \psi}(\omega)
\end{aligned}
$$

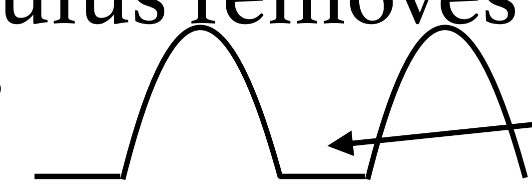


$$\omega^T a\hat{\psi}(\omega) \approx \omega_0^T a\hat{\psi}(\omega)$$

the infinitesimal generator
of translations is the derivative...

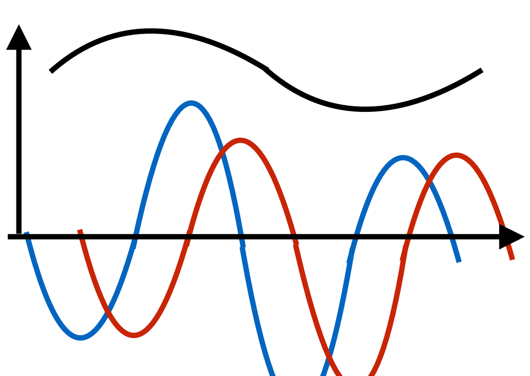
- A modulus removes the phase!

Rectificatio



Phase artifact

Modulu



Real and
imaginary

Non-linear projection

Information loss
Reconstruction

Ref.: Mallat S, Bruna J

x

$$\arg\inf_{y} \|S_3 x - S_3 y\| \longrightarrow$$

invariance up to
2^3 pixels

$\tilde{y}$

Wavelets on Lie group

- **Discovering more complex groups** is necessary to build more complex invariants:

Ref.: Deep Roto-Translation Scattering for Object Classification. EO and S Mallat

Translation Scattering does not see the difference

$$\mathbb{R}^2 \hookrightarrow SO_2(\mathbb{R}) \rtimes \mathbb{R}^2 \hookrightarrow \dots$$

- A wavelet is defined by $\psi \in L^2(G), \hat{\psi}(e) = 0$ and can be dilated via $\psi_\lambda = L_\lambda \psi$

- **<u>Theorem</u>**: Let G be a compact Lie group, for appropriate mother wavelet ψ and Λ then

$$Wx = \{\int_G x, x \star^G \psi_\lambda\}_{\lambda \in \Lambda}$$

is an isometry and covariant with the action of G

Ref.: Stein, E. M. Topics in harmonic analysis related to the Littlewood-Paley theory.

- **<u>Proposition</u>**: W almost commutes with deformations but is not invariant to translation...

$$\|[W, L_\tau]\| \leq C\|\tau\|$$

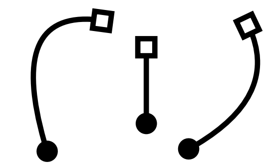
Ref.: Group Invariant Scattering, Mallat S

An ideal input for a modern CNN

$$x \longrightarrow \boxed{S_J} \longrightarrow \boxed{CNN} \longrightarrow$$

Deformations

$$L_\tau x(u) = x(u - \tau(u))$$



- Scattering is stable:
$$\|S_J x - S_J y\| \leq \|x - y\|$$
- Linearize small deformations:
$$\|S_J L_\tau x - S_J x\| \leq C \|\nabla \tau\| \|x\|$$
- Invariant by local translation:
$$|a| \ll 2^J \Rightarrow S_J L_a x \approx S_J$$
- For $\lambda, u,\ S_J x(u, \lambda)$ has a topology that is **structured** by $SO_2(\mathbb{R})$, and this structures the first layer also:

Ref.: Scaling the Scattering Transform: Deep Hybrid Networks EO, E Belilovsky, S Zagoruyko

$$\text{if } \forall u \forall g \in SO_2(\mathbb{R}), g.x(u) \triangleq x(g^{-1}u) \text{ then,}$$

$$S_J(g.x)(u, \lambda) = S_J x(g^{-1}u, g^{-1}\lambda) \triangleq g.S_J x(u, \lambda)$$

How much learning is really required?

Ref.: Deep Roto-Translation Scattering for Object Classification. EO and S Mallat

Dataset	Type	Paper	Accuracy
Caltech101	Scattering		79.9
	Unsupervised	Ask the locals	77.3
	Supervised	DeepNet	***91.4***
CIFAR100	Scattering		56.8
	Unsupervised	RFL	54.2
	Supervised	DeepNet	***65.4***

Identical Representation

10^4 images
101 classes
256×256 color images

CALTECH

CIFAR

5.10^4 images
100 classes
32×32 color images

Group representations are competitive with representations learned from data without labels

Benchmarking ImageNet

Ref.: Scaling the Scattering Transform: Deep Hybrid Networks
EO, E Belilovsky, S Zagoruyko

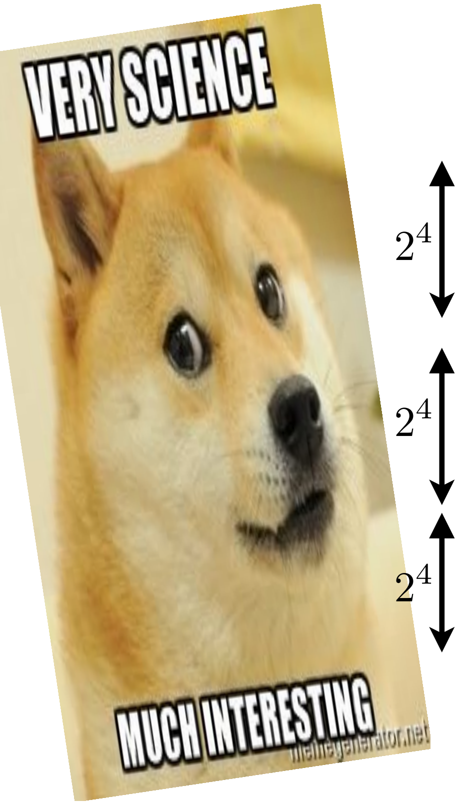
$$x \longrightarrow \boxed{S_J} \longrightarrow \boxed{ResNet} \longrightarrow$$

- Cascading a modern CNN leads to almost state-of-the-art result on Imagenet2012:

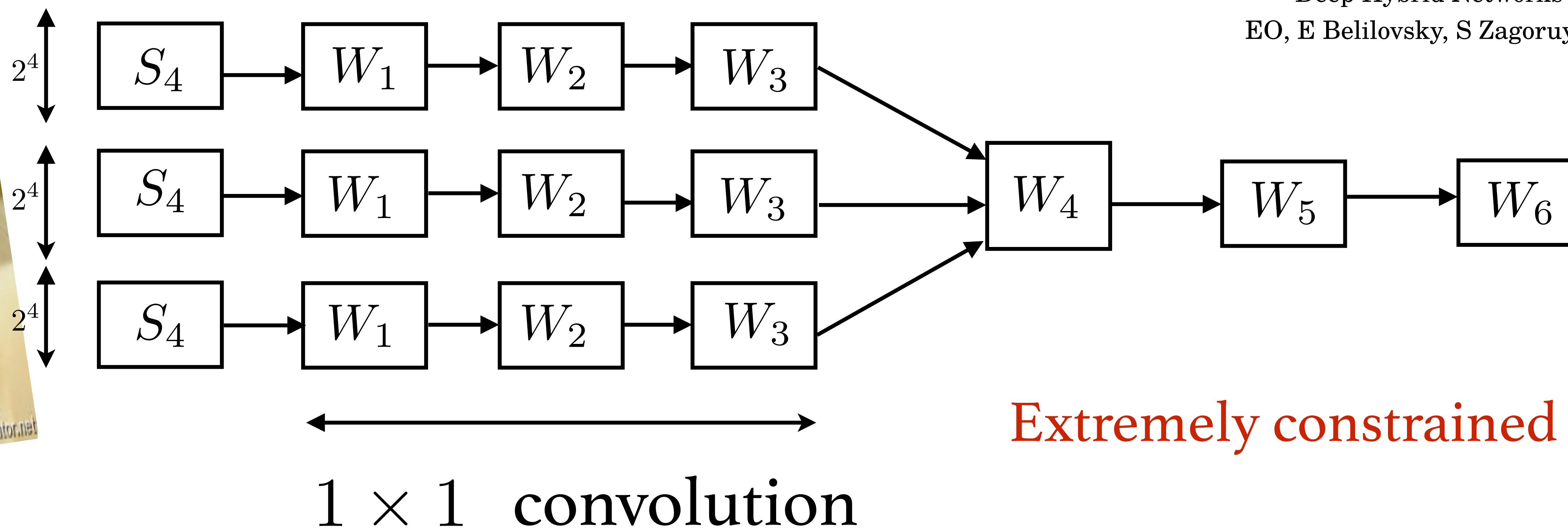
Method	Top 1	Top 5	Params
AlexNet	56.9	80.1	61M
VGG-16	68.5	88.7	138M
Scat + Resnet-10 (ours)	68.7	88.6	12.8M
Resnet-18 (ours)	68.9	88.8	11.7M
Resnet-200	**78.3**	**94.2**	64.7M

- Demonstrates no loss of information + Less layers

Shared Local Encoder



Ref.: Scaling the Scattering Transform: Deep Hybrid Networks
EO, E Belilovsky, S Zagoruyko



Extremely constrained

- It is equivalent to encode the non-overlapping scattering patches: the output of the 1x1 is a **local descriptor** of an image that leads to AlexNet performances.

Method	Top 1	Top 5
FV + FC	55.6	78.4
FV + SVM	54.3	74.3
AlexNet	56.9	**80.1**
Scat + SLE	**57.0**	79.6

Good generalization on Caltech101

Benchmarking Small data

Ref.: Scaling the Scattering Transform:
Deep Hybrid Networks
EO, E Belilovsky, S Zagoruyko

- Adding geometric prior regularises the CNN input, in the particular case of limited samples situations, **without reducing** the number of parameters.

- State-of-the-art results on STL10 and CIFAR10:

STL10: 5k training, 8k testing, 10 classes
+100k unlabeled(not used!!)

Method	Accuracy
Supervised methods	
Scat + WRN 19-8	$\mathbf{76.0 \pm 0.6}$
CNN	70.1 ± 0.6
Unsupervised methods	
Exemplar CNN	75.4 ± 0.3
Stacked what-where AE	74.33
Hierarchical Matching Pursuit (HMP)	64.5 ± 1
Convolutional K-means Network	60.1 ± 1

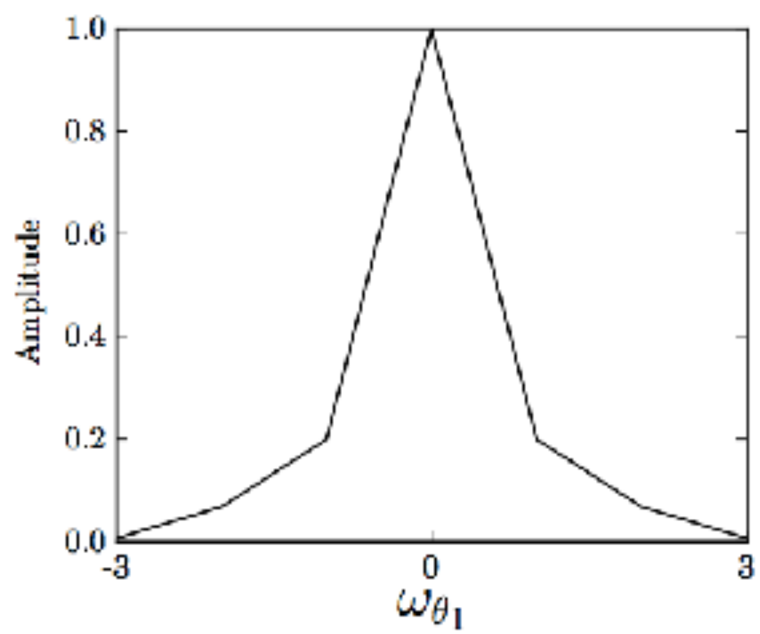
Cifar10, 10 classes
keeping 100, 500 and 1000 samples
and testing on 10k

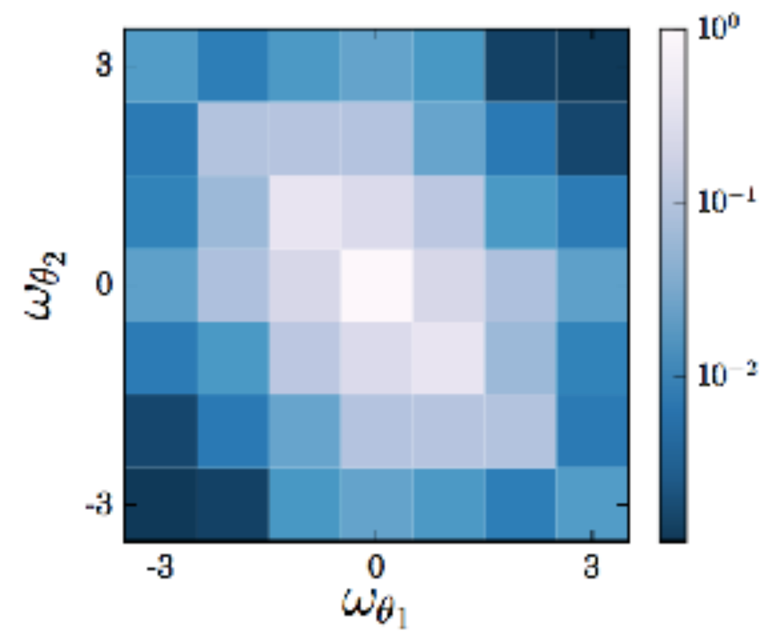
Method	100	500	1000
WRN 16-8	34.7 ± 0.8	46.5 ± 1.4	60.0 ± 1.8
Scat + WRN 12-8	$\mathbf{38.9 \pm 1.2}$	$\mathbf{54.7 \pm 0.6}$	$\mathbf{62.0 \pm 1.1}$

Invariance to rotation

Ref.: Scaling the Scattering Transform: Deep Hybrid Networks
EO, E Belilovsky, S Zagoruyko

- We evaluate the angular energy propagated for given frequencies: $\Omega(\omega_{\theta_1}, \omega_{\theta_2}) = \sum |W_1(., \omega_{\theta_1}, \omega_{\theta_2})|^2$





- They are all localised in the low-frequency domain: **invariance to rotation** is learned. (supports symmetry group hypothesis)

Multiscale Hiearchical CNN

- Can we structure the next layers?

Ref.: Multiscale Hierarchical Convolutional Networks
J Jacobsen, EO, S Mallat, Smeulders AWM

- Introduce a CNN that is convolutional along each direction, finally averaged:

$$x_{j+1} = \rho_j W_j x_j$$

$$x_{j+1}(v_1, ..., v_j, v_{j+1}) = \rho_j(x_j \star^{v_1,...,v_j} \psi_{v_{j+1}})(v_1, ..., v_j)$$

$$x_J = \sum_{v_j, j \leq J-2} x_{J-1}(v_1, ..., v_{J-1})$$

- For x_j, we refer to the variable v_j as an attribute that discriminates previously obtained tensor.

- W_j performs an averaging along v_{j-2}.

Flattening the variability

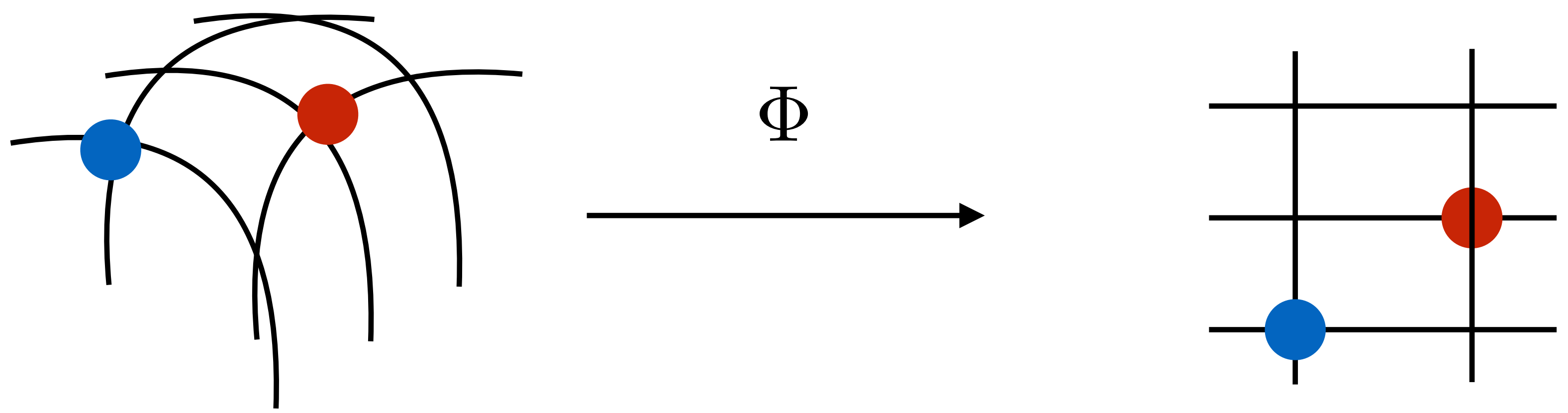
- An explicit invariant of any translations along $(v_1, ..., v_j)$ is built.

Ref.: Multiscale Hierarchical Convolutional Networks
J Jacobsen, EO, S Mallat, Smeulders AWM

- Completely structures the axis of the "channels" via convolutions.

- It aims at mapping the symmetries of $\Phi x = x_J$ into the translations along $G_j = \mathbb{R}^j, j \leq J$.

Organizing the channels indexes



Reducing the number of parameters

Ref.: Multiscale Hierarchical Convolutional Networks
J Jacobsen, EO, S Mallat, Smeulders AWM

CIFAR10

Model	# Parameters	% Accuracy
Hiearchical CNN	0.098M	91.43
Hiearchical CNN (+)	0.34M	92.50
All-CNN	1.3M	92.75
ResNet	0.27M	91.25
Network in Network	0.98M	91.20
WRN-student	0.17M	91.23
FitNet	2.5M	91.61

This implies an effective structuration

CIFAR100

Model	# Parameters	% Accuracy
Hiearchical CNN	0.25M	62.01
Hiearchical CNN (+)	0.89M	63.19
All-CNN	1.3M	66.29
Network in Network	0.98M	64.32
FitNet	2.5M	64.96

Organization of the representation?

Ref.: Multiscale Hierarchical Convolutional Networks
J Jacobsen, EO, S Mallat, Smeulders AWM

- We observe that representations at several layers are translated:

Conclusion

- Structuration should be the topic of future research to improve Deep neural networks
- Check my webpage for softwares and papers: http://www.di.ens.fr/~oyallon/

Thank you!