MAP670R

Advanced Topics in Deep Learning

Edouard Oyallon

edouard.oyallon@cnrs.fr

CNRS, Sorbonne University

Overview of the lectures

- 24/02/22 - Lecture 1: Symmetry, Invariance & Groups (3h30)
- 03/03/22 - Lecture 2: Scattering Transform & Non-Euclidean data. (2h30) + 1h lab
- 10/03/22 - Lecture 3: Approximation of Shallow NNs and Lazy training (2h30) + 1h lab
- 17/03/22 - Lecture 4: Generalisation properties of DNNs.
- 24/03/22 - Lecture 5: TBD
- 31/03/22 - Poster presentation of the Projects

Grade= 50%(1 homework+1 lab) + 50% project

Groups of 2: homework and projects have to done by groups of 2!
Projects: Pick a research article from a list or an academic paper of your choice (please validate it with me)
Project grading procedure: via a poster (as in academic conferences), 5-10 min of presentations + 5 min of questions. The quality of the poster will be graded.

A poster is about A1 format (and can simply be a collection of 6-8 A4 pages)

Homework is out and due in 2 weeks (March 10th), as well as project choices.

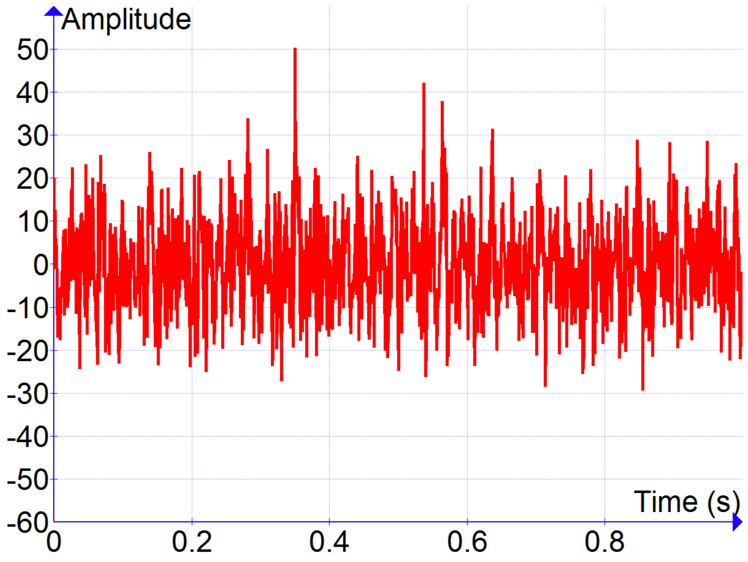
Generic statements

- Announcements will be held on the website, and sometimes by email.
- For each lecture, you'll find some reference papers, lecture notes, slides.
- A google spreadsheet will be dedicated to the projects.
- No correction for the lab will be sent.

Signal Processing meets Deep Learning

- **Signal processing goal**: analysing, generating or altering the digitalisation of observations obtained from a sensor.

Relies a lot on Fourier Analysis!



- **Deep Learning goal**: solving signal processing tasks with neural networks.

Traditionally understood through the lens of Machine Learning

Lecture 1: Symmetry, Invariance & Groups

Edouard Oyallon
edouard.oyallon@cnrs.fr
CNRS, Sorbonne University

Machine Learning & Deep Learning for Information Access

Objective of the current lecture

- Understanding the challenges in high-dimensional classification

- Understanding the concepts of covariance, invariance and linearisation

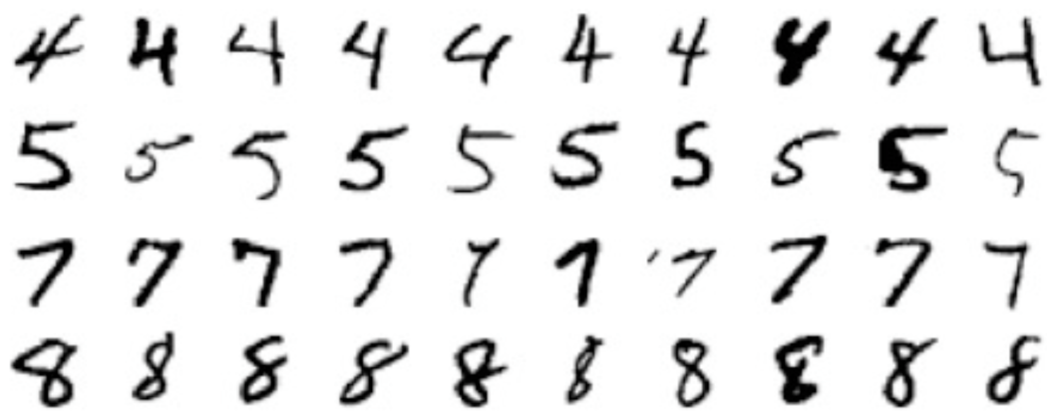
- Linking Machine Learning and Signal Processing

- Introducing the Scattering Transform

We will discuss widely the Scattering Transform.

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

- Successfully used in several applications:
  - Digits



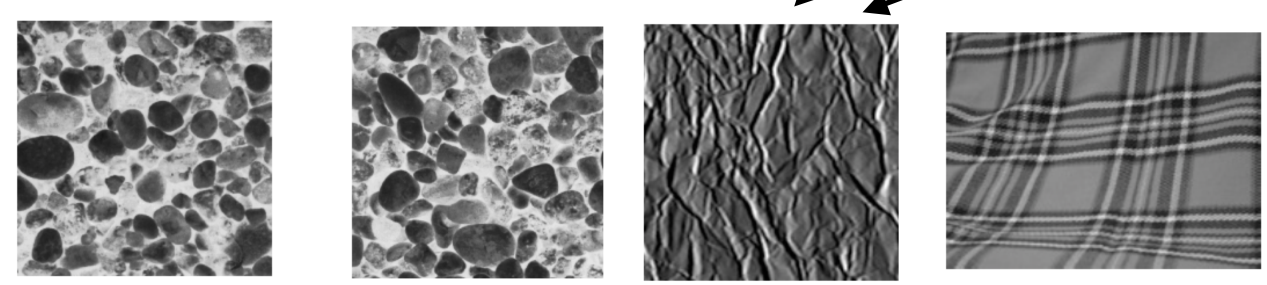
All variabilities are known

Small deformations +Translation

  - Textures

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

Rotation+Scale

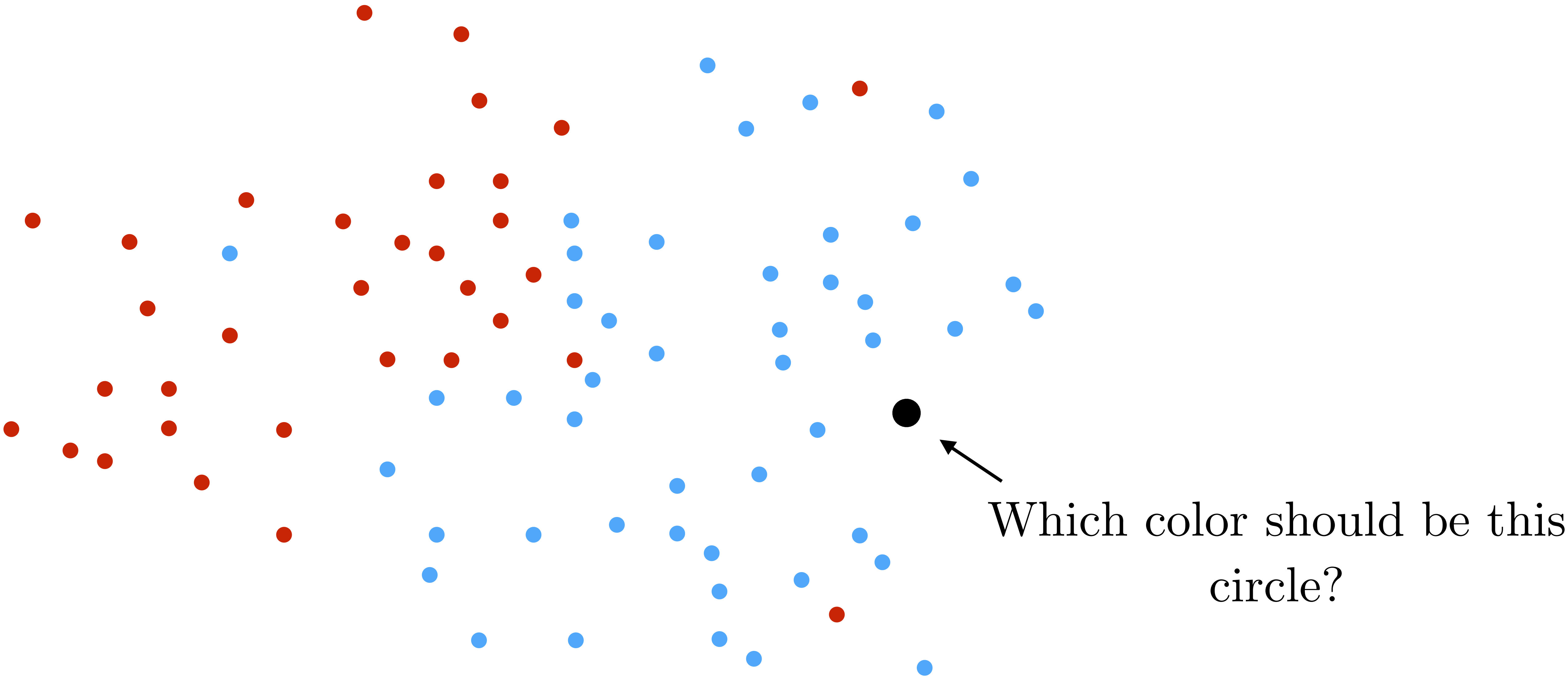


- We will see that the design of the scattering transform is guided by the euclidean group.

- **Goal of a Scattering Transform**: removing undesirable (group) variabilities

General comments about Deep Learning

High-dimensional classification



An example of supervised task: classification

<u>Why is this picture bad?</u>

Supervised task

$\mathcal{X} = \mathbb{R}^2$ Samples space

$\mathcal{Y} = \{\bullet, \bullet\}$ Labels

Input data $x \in \mathcal{X}$ $\xrightarrow{\Phi?}$ Output data $\Phi(x) \approx y \in \mathcal{Y}$

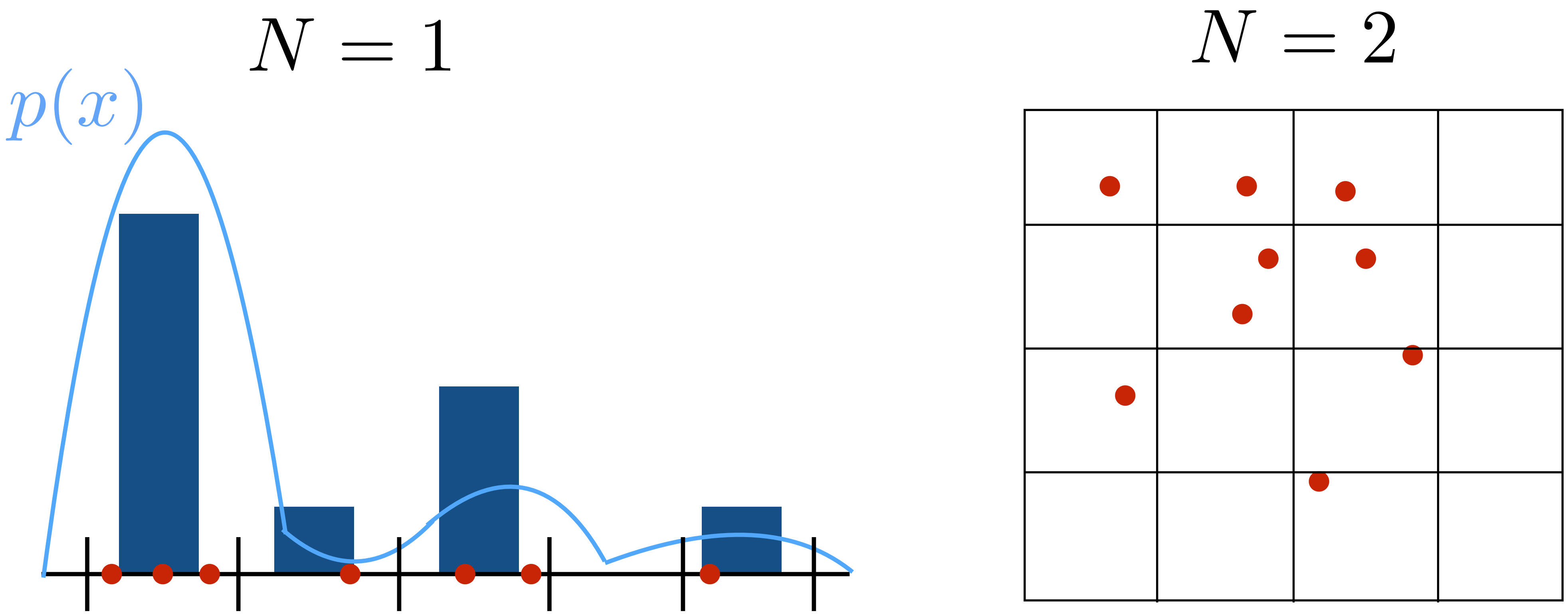
- Estimating a label $\boldsymbol{y}$ from a sample x, by training a model Φ on a training set. Validation of the model is done on a different test set.

- Examples: prediction, regression, classification,...

- Best setting: dimensions of x and y is small, $\mathcal{X}$ large

High dimensional images

- PdFs are difficult to estimate in high dimension.



- For a fixed number of points and bin size, as N increases, the bins are likely to be empty.

Curse of dimensionality:
occurs in many machine learning problems

Very high-dimensional images

- Curse of dimensionality!

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

Estimation problem

Training set to predict labels

"Rhino"

Large datasets. . .

- ImageNet 2012: (350GB)
  1 million training images, 1 000 classes
  400 000 test images
  Large coloured images of various sizes

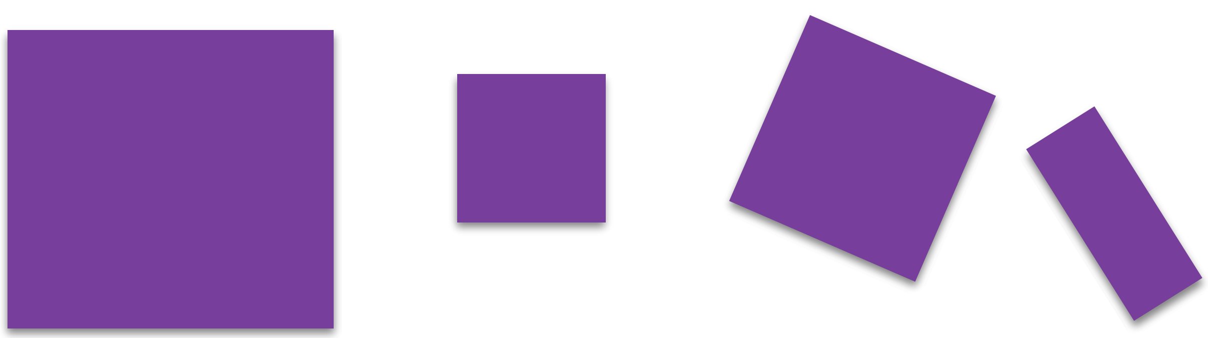
Ref.: image-net.org

- Labels obtained via Amazon Turk (complex process that requires human labelling)

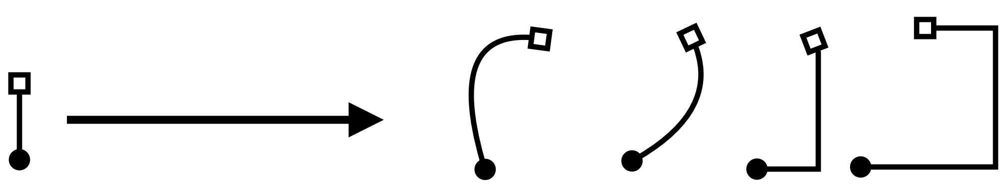
Difficult problems due to Image variabilities

Geometric variability

Groups acting on images:
translation, rotation, scaling



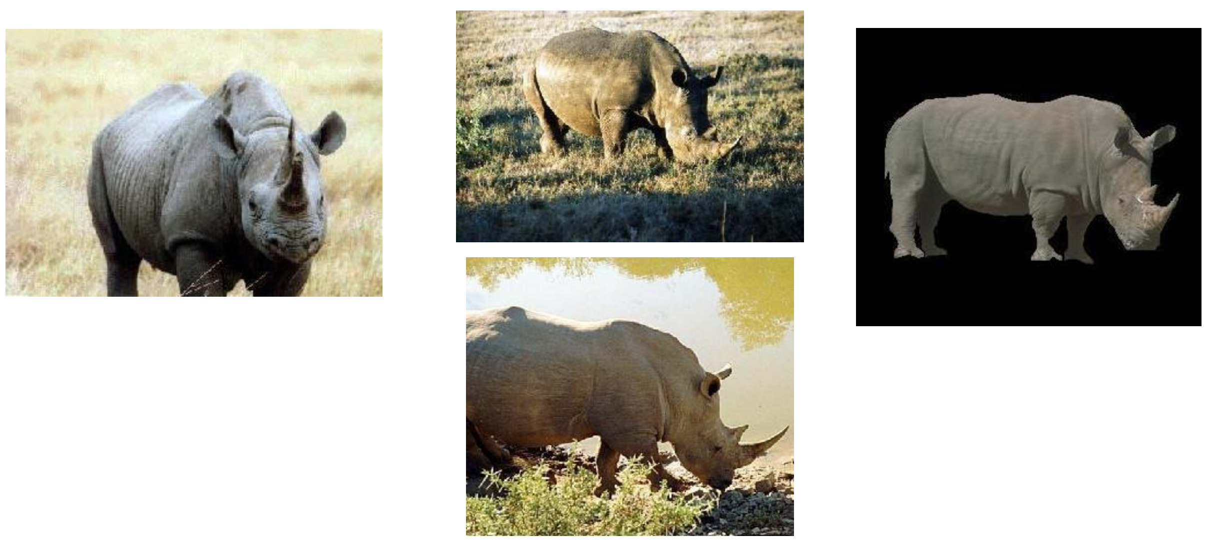
Other sources : luminosity, occlusion,
small deformations



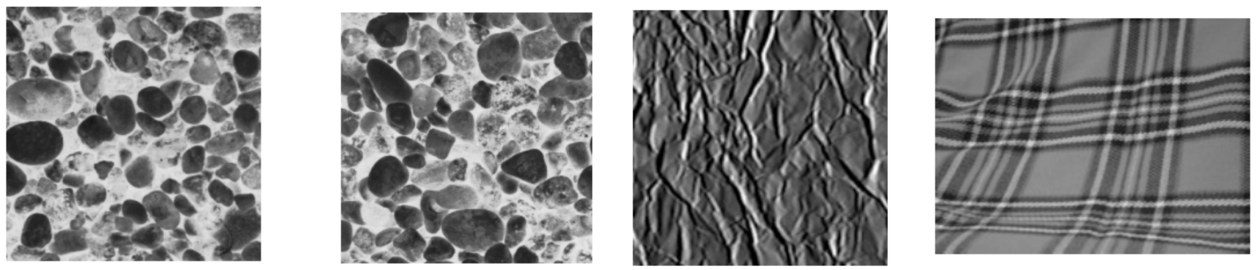
Class variability

Intraclass variability

<u>Not informative</u>



Extraclass variability

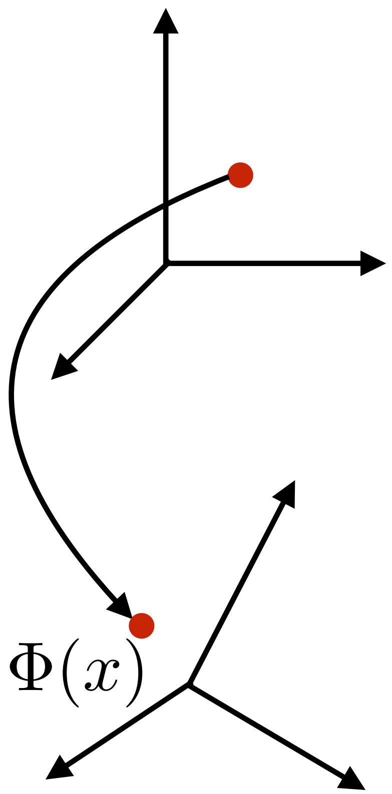


High variance: hard to reduce!

Desirable properties of a representation

- **Invariance** to group G of transformation (e.g. roto-translation):

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



- **Stability** to noise

$$\forall x, y, \|\Phi(x) - \Phi(y)\|_2 \leq \|x - y\|_2$$

- **Reconstruction** properties

$$y = \Phi(x) \iff x = \Phi^{-1}(y)$$

- **Linear separation** of the different classes

$$\forall i \neq j, \|E(\Phi(X_i)) - E(\Phi(X_j))\|_2 \gg 1$$

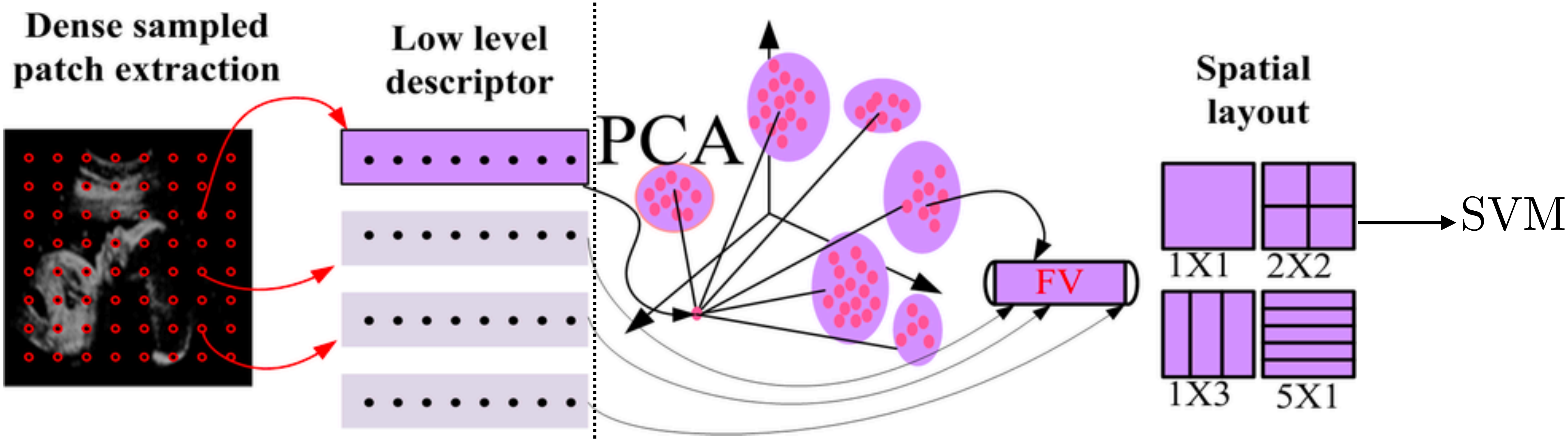
$$\forall i, \sigma(\Phi(X_i)) \ll 1$$

Can be difficult to handcraft..

Typical Vision pipelines (prior 2010)

Not Learned | Mostly Learned

Image Extractor → Image Descriptor → Descriptor Encoder → Classifier

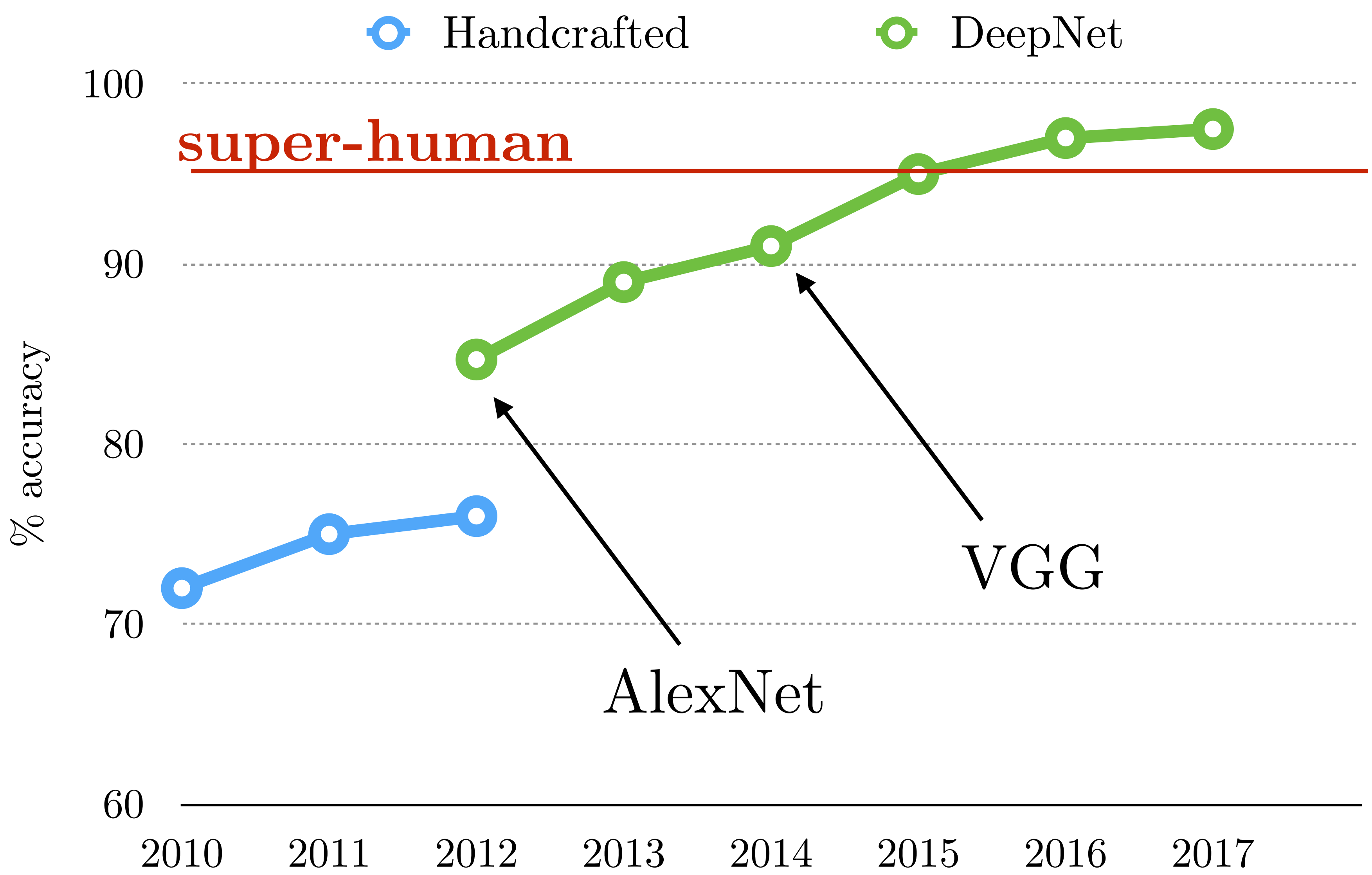


Is ImageNet solvable?

Years of research. . .

Of course?

- Huge gap thanks to deep neural networks.

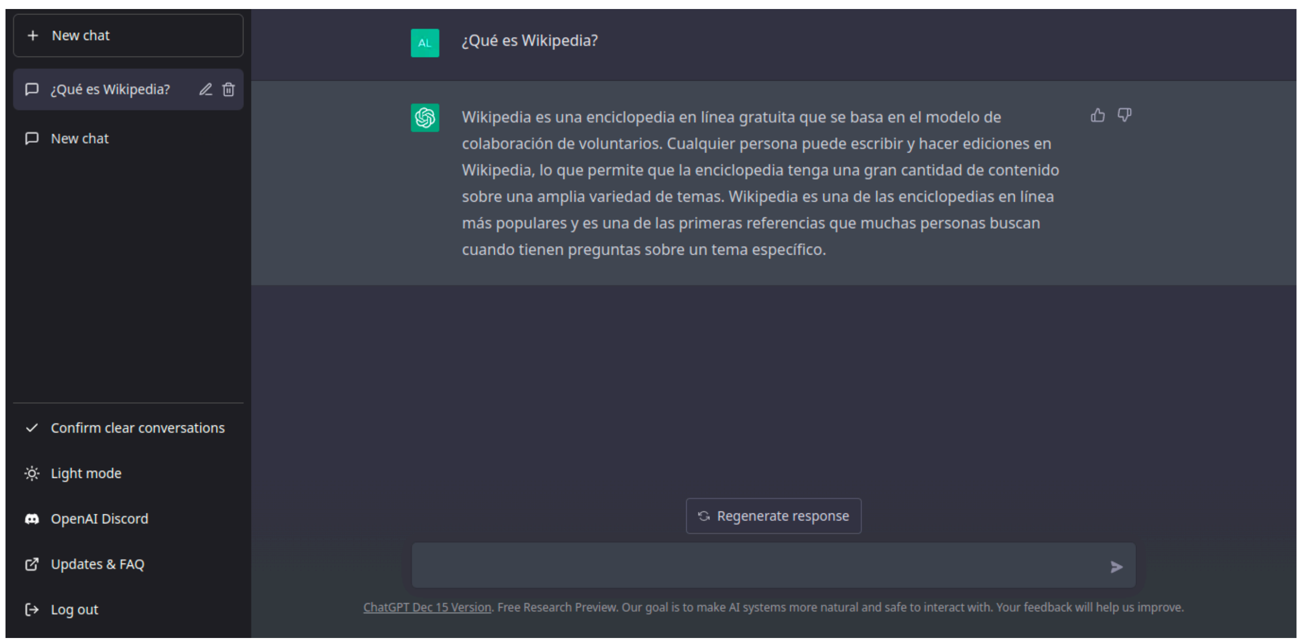


top5 - ImageNet

ImageNet:

1 million training images, 1 000 classes
400 000 test images
Large coloured images of various sizes

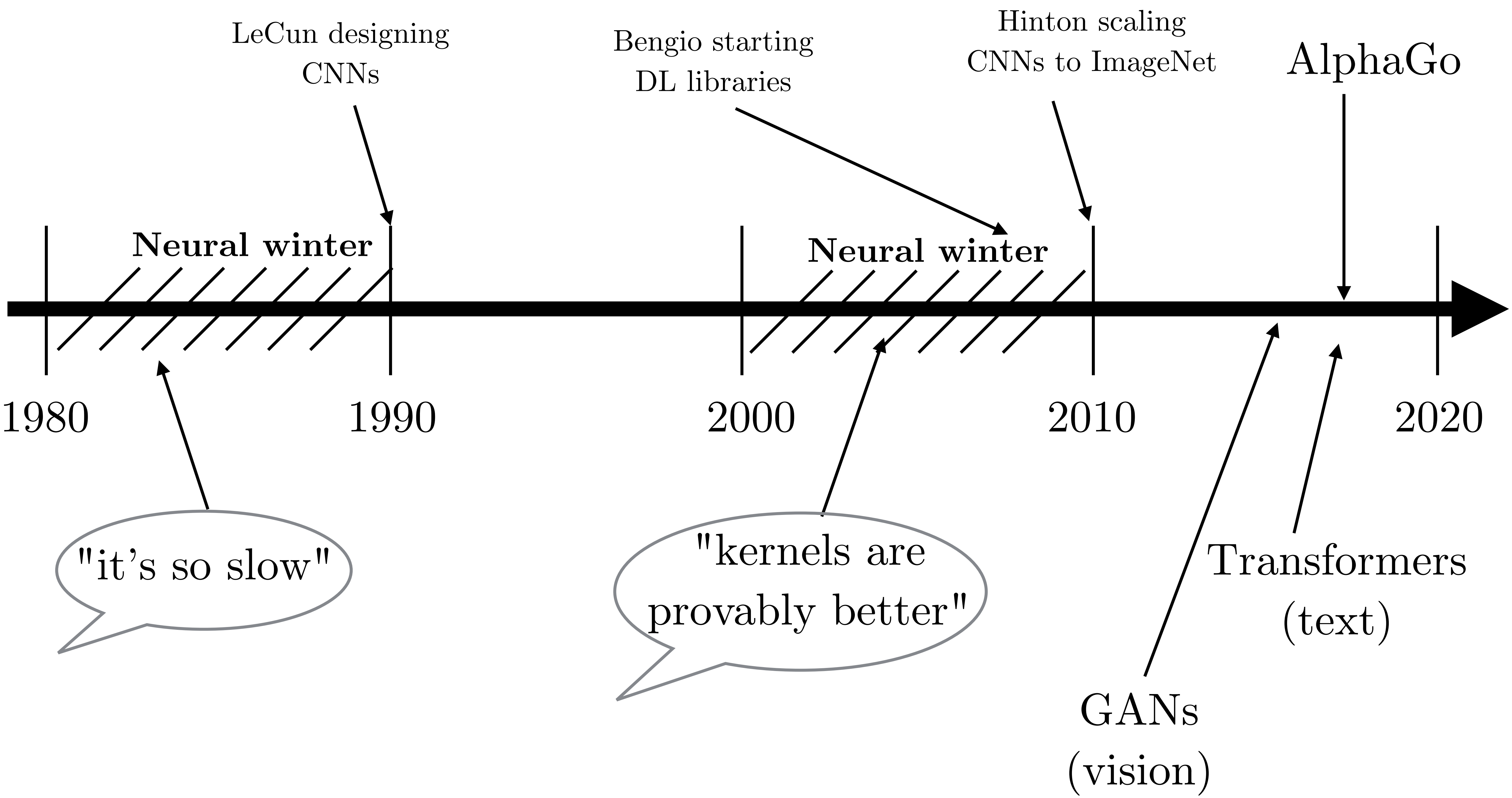
Theory for good performances?



Spectacular results

(let's not spend too much time here, everybody is convinced by their supremacy.)

A biased history of Deep Learning



Multi Layer Perceptron

- We'll write a J-1-hidden layer neural network of depth J, with affine operators $W_1, ..., W_J$:

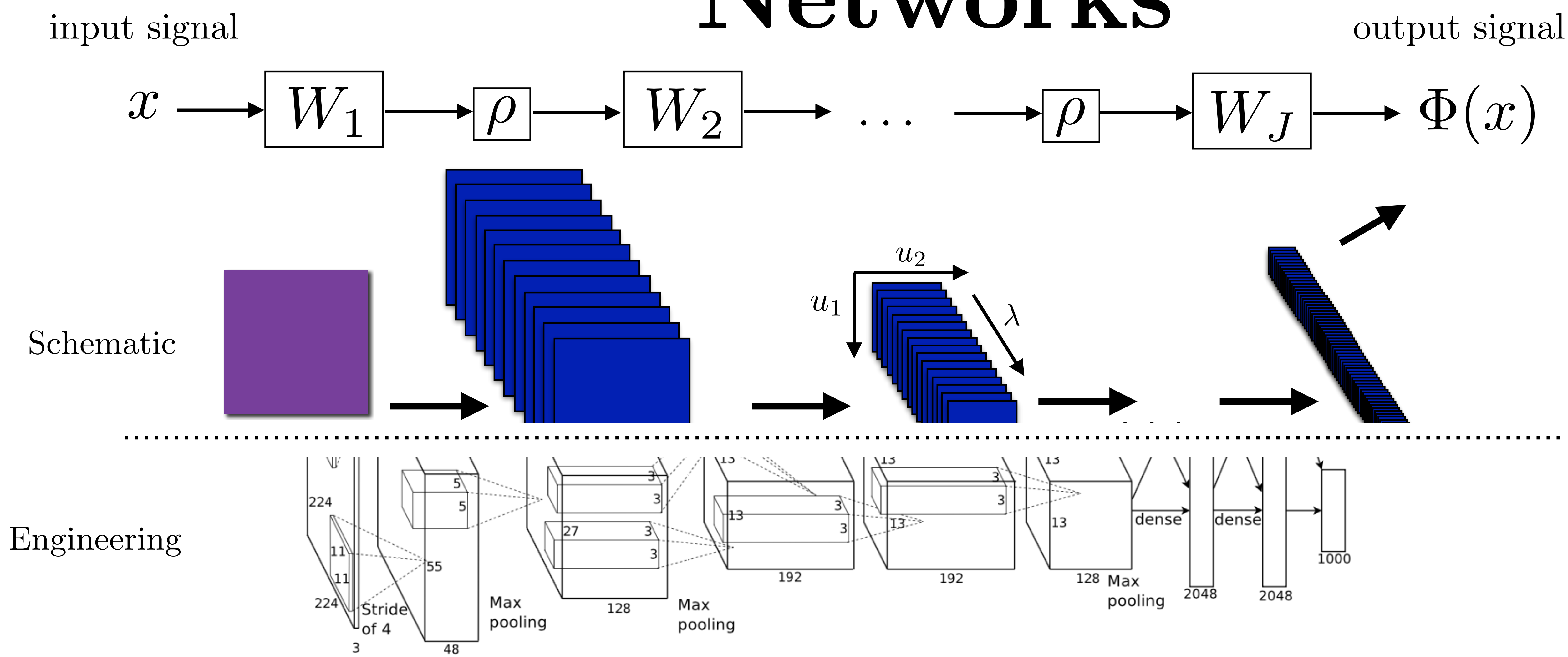
$$\Phi x = W_J \rho W_{J-1} \rho ... W_1 x$$

- Where, $\rho : \mathbb{R} \to \mathbb{R}$ is a non-linear function that we extend to a point-wise non-linear operator via:

$$[\rho(x)]_i = \rho(x_i)$$

- An additional parameter is the maximal width "K" of each layer.

Convolutional Neural Networks



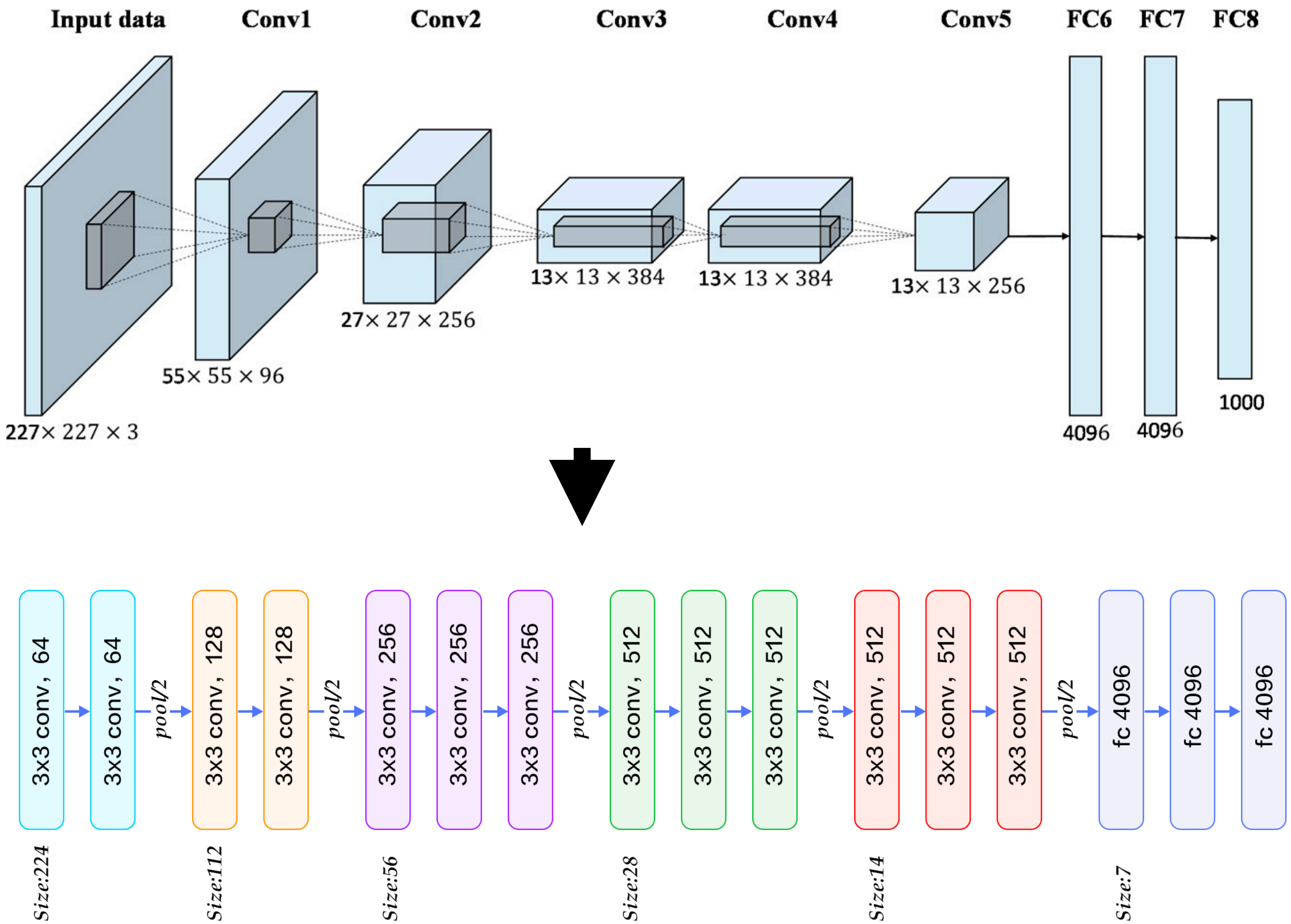
Each layer: $x_{j+1} = \rho W_j x_j$

learned kernel

that leads to: $x_{j+1}(u, \lambda_{j+1}) = \rho\left(\sum_{\lambda_j} \left(x_j(., \lambda_j) \star w_{\lambda_j, \lambda_{j+1}}\right)(u)\right)$

where: $\rho(x) = \max(0, x)$ s.t. $|\rho(x) - \rho(y)| \leq |x - y|$

From AlexNet to VGG to ResNet



From 7x7 convolutions to 3x3 convolutions.

\+ Less down-sampling

For an image of size N

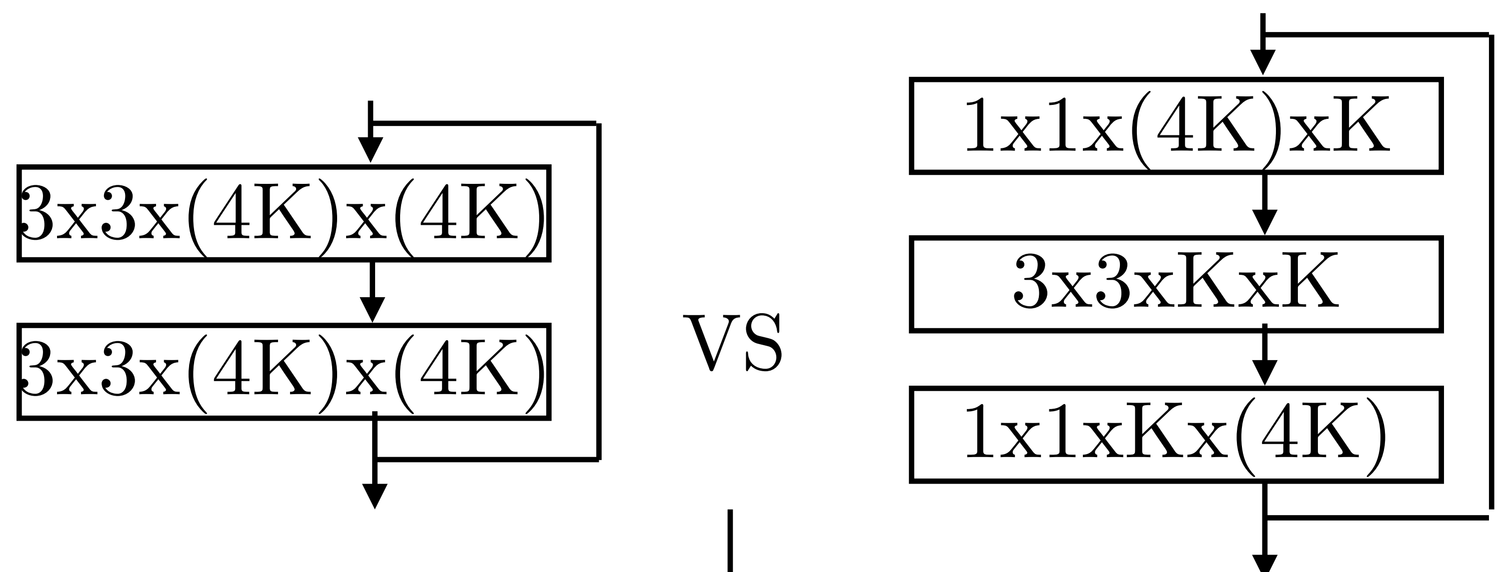
Kernel size	3x3	7x7	3x3>3x3>3x3
Receptive field	3	7	7
# params	9	49	27
Complexity	9N	49N	27N

From VGG to ResNet

Bottlenecks as a cheap way to increase dimension >> only helpful for Deeper CNNs

34-layer	50-layer
7×7, 64, stride 2	
3×3 max pool, strid	
[3×3, 64; 3×3, 64] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3
[3×3, 128; 3×3, 128] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4
[3×3, 256; 3×3, 256] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×6
[3×3, 512; 3×3, 512] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3
average pool, 1000-d fc,	

	plain	ResNet
18 layers	27.94	27.88
34 layers	28.54	**25.03**



Complexity	290K^2	17K^2
#params	290K^2	17K^2

Take home message: tricks to maximise the utility of a GPU to train bigger CNNs.

Today study:

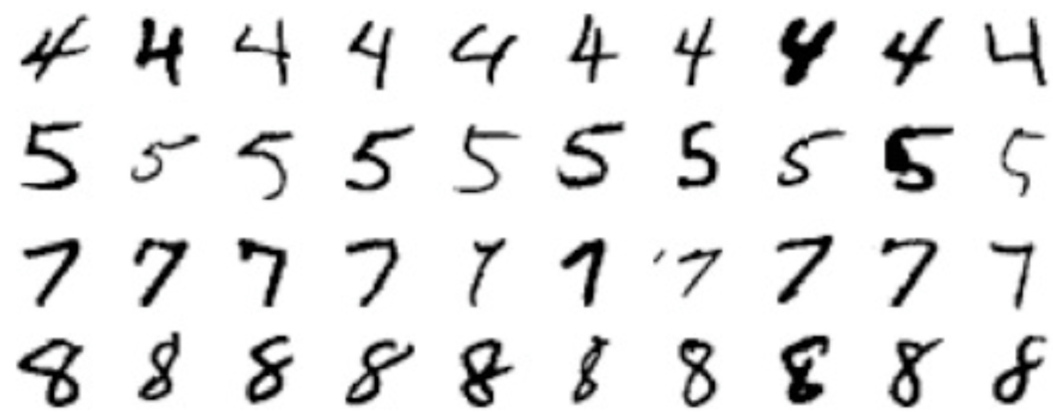
We will discuss widely the Scattering Transform (2).

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

- Successfully used in several applications:

All variabilities are known

  - Digits

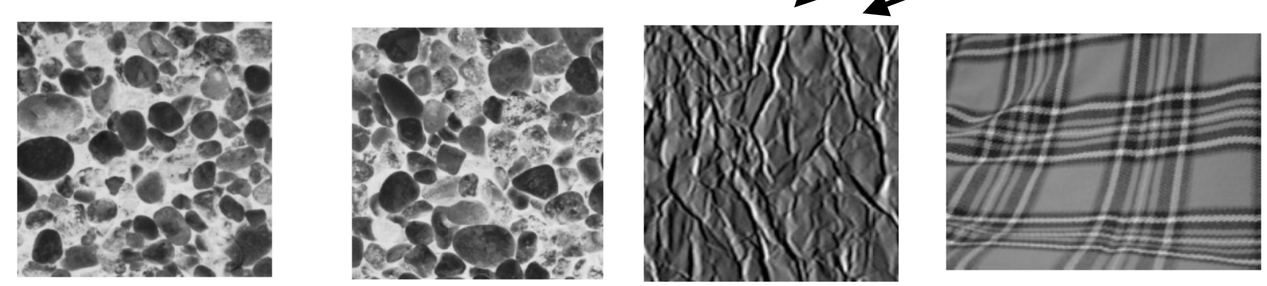


Small deformations +Translation

Rotation+Scale

  - Textures

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



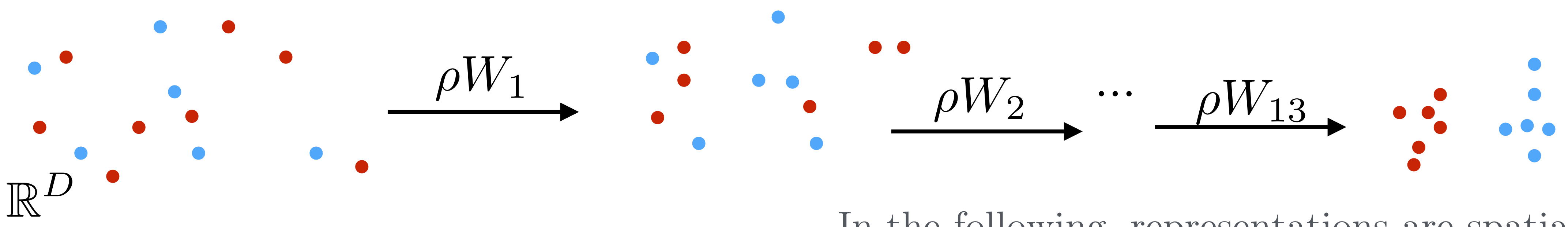
- The design of the scattering transform is guided by the euclidean group

- To which extent can we compete with other architectures on more complex problems (e.g. variabilities are more complex)?

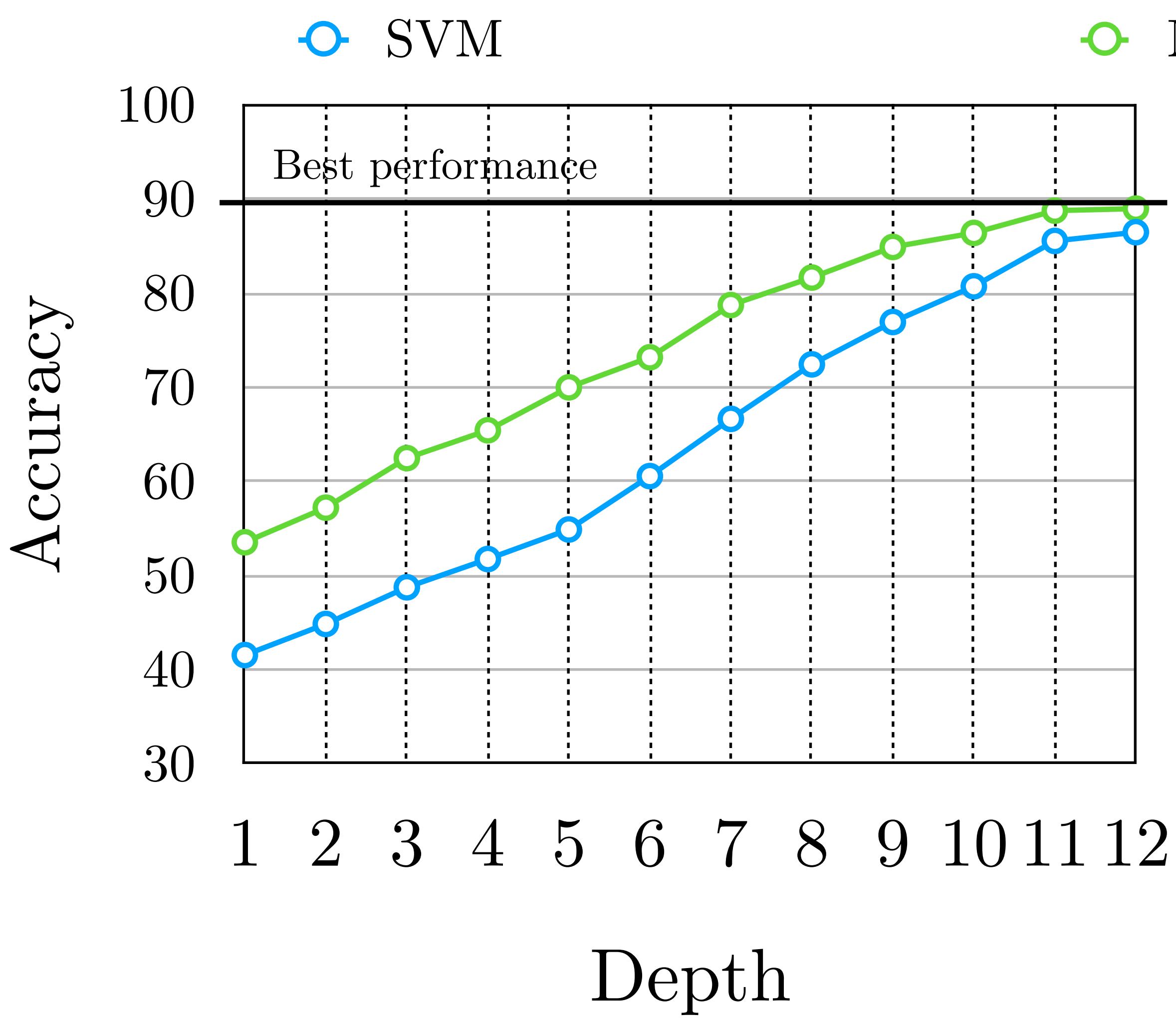
Symmetries, linearisation

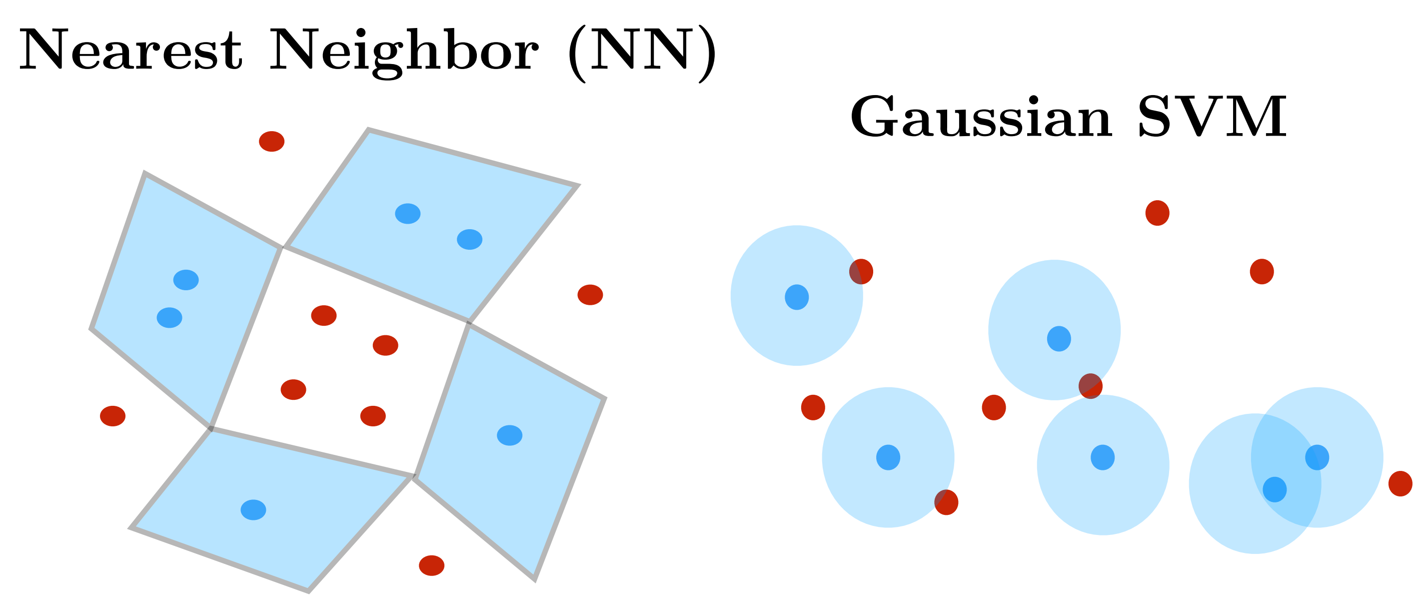
Empirical observation: Progressive separability

- **Typical CNN exhibits a progressive contraction & separation, w.r.t. the depth:**



In the following, representations are spatially averaged.





Localised classifiers

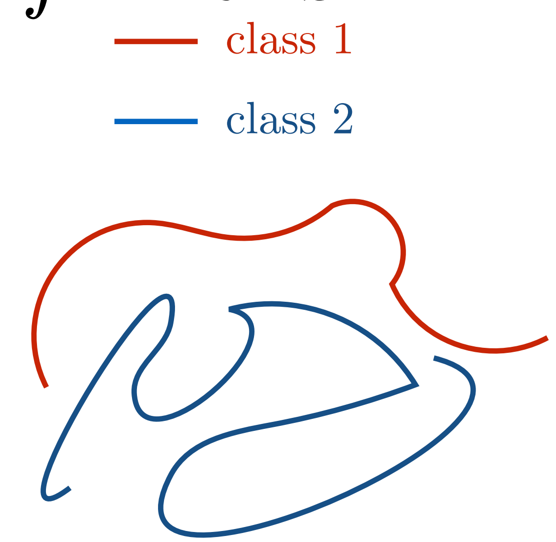
Ref.: Building a Regular Decision Boundary with Deep Networks, EO

- **How can we explain it?**

Symmetries

- Consider $f : \mathcal{X} \to \mathbb{R}$.
  We say that $\mathcal{L} : \mathcal{X} \to \mathcal{X}$ is a symmetry of f if it is invertible and:

$$f(\mathcal{L}x) = f(x)$$



- We can consider the group of symmetries:

$$G = \{\mathcal{L} \text{ invertible}, f(\mathcal{L}x) = f(x)\}$$

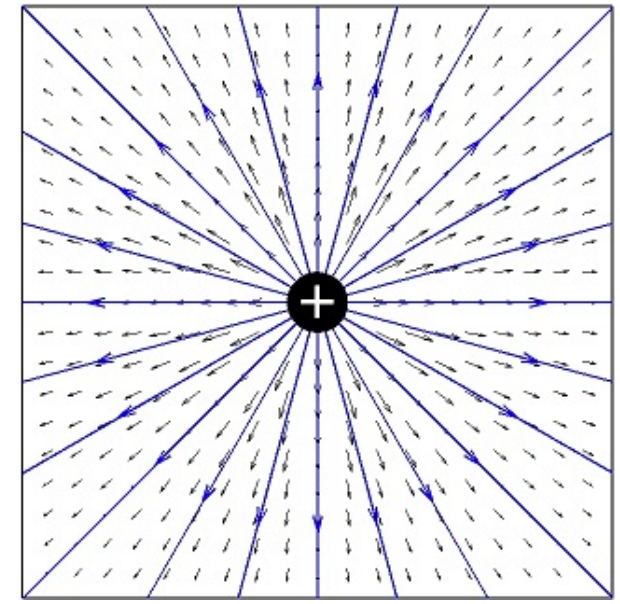
- Without any constraints on G, the action is transitive and thus f is completely characterised by G, as:

$$f^{-1}(f(x)) = \{\mathcal{L}x, \mathcal{L} \in G\}$$

As for any $\{u, v\}$, one can get: $\mathcal{L}x = \begin{cases} u, & \text{if } x = v \\ v, & \text{if } x = u \\ x, & \text{otherwise} \end{cases}$

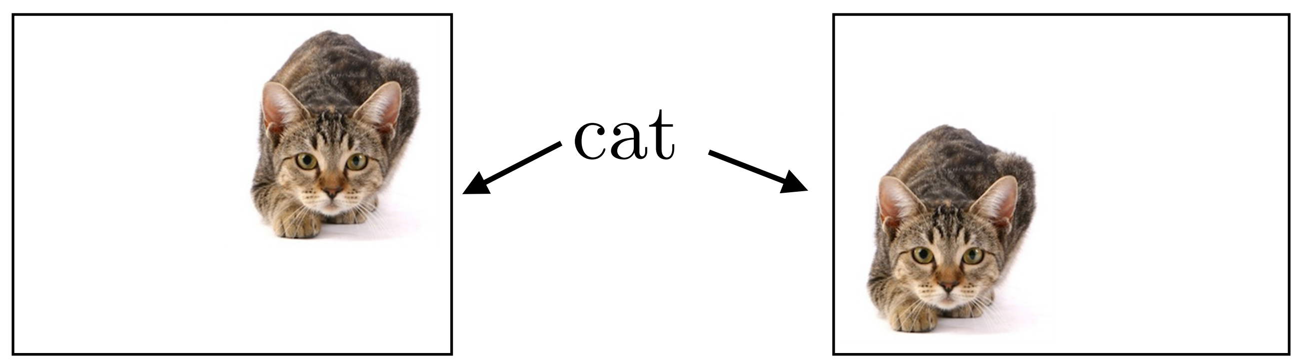
Examples of symmetries

- In physics: $r_\theta E(u) \triangleq E(r_{-\theta} u)$



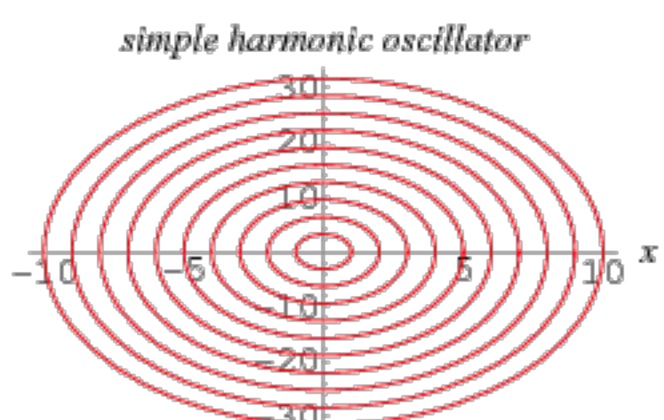
via $$E(u) = \frac{q}{4\pi\epsilon_0 \|u - u_0\|}$$

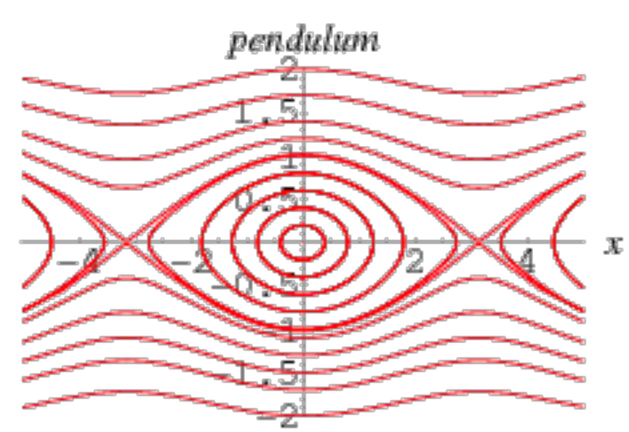
- In machine learning: $\mathcal{L}_a x(u) \triangleq x(u - a)$



- With ODEs:

$$\mathcal{L}_t y(u) \triangleq y(u - t)$$
$$y' = F(y)$$

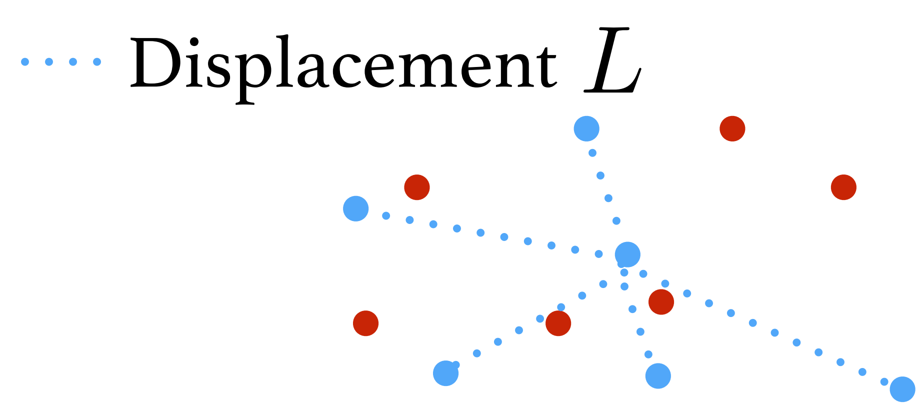




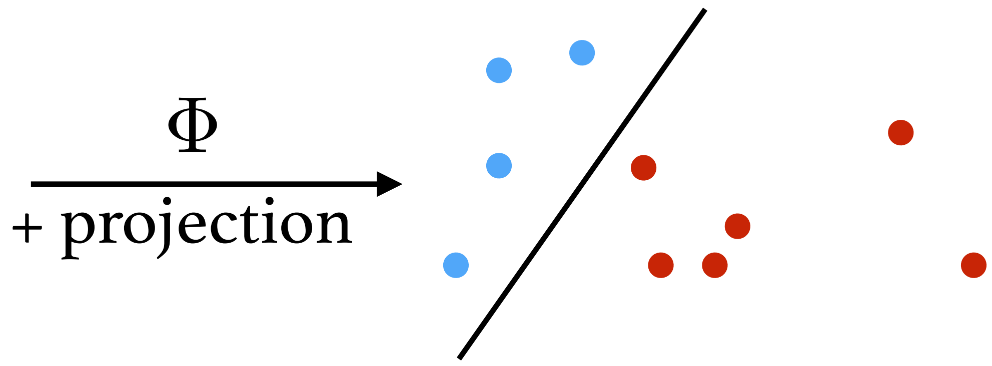
Linearization: Lipschitz gives differentiability

- Weak differentiability property, via Rademacher theorem:

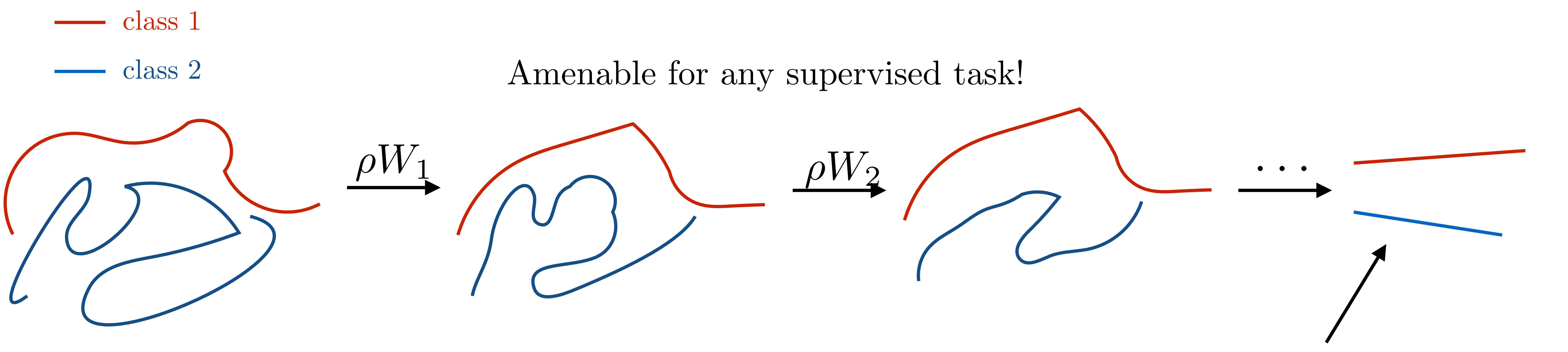
$$\sup_L \frac{\|\Phi L x - \Phi x\|}{\|Lx - x\|} < \infty \Rightarrow \exists \text{ ''weak'' } \partial_x \Phi$$
$$\Rightarrow \Phi L x \approx \Phi x + \underbrace{\partial_x \Phi L}_{\text{A linear operator}} + o(\|L\|)$$



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



Flattening the level sets (= classification symmetries)



Ref.: Understanding Deep Convolutional Networks, Mallat, 2016

- How to linearize? Ex.: Gâteaux differentiability

$$\exists C_x,\ \sup_{\mathcal{T}} \frac{\|\Phi x - \Phi \mathcal{T} x\|}{\|\mathcal{T}\|} < C_x \quad \Rightarrow \exists \partial \Phi_x :\ \Phi \mathcal{T} x \approx \Phi x + \partial \Phi_x . \mathcal{T}$$

- However, exhibiting $\mathcal{T}$ can be difficult. (*curse of dimensionality*)

Ex.: linear translations $\mathcal{T}_a(x)(u) \triangleq x(u+a)$, yet non linear case?

Flattening the space: progressive manifold?

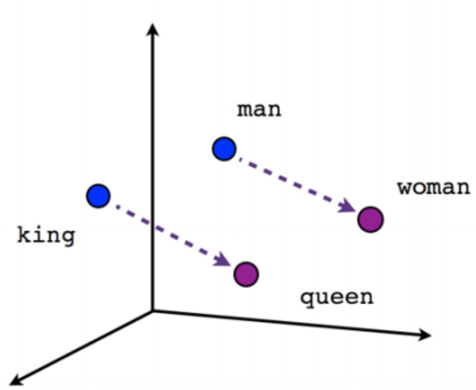
- Parametrize variability on synthetic data: $L_\theta, \theta \in \mathbb{R}^d$ and observe it after PCA

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

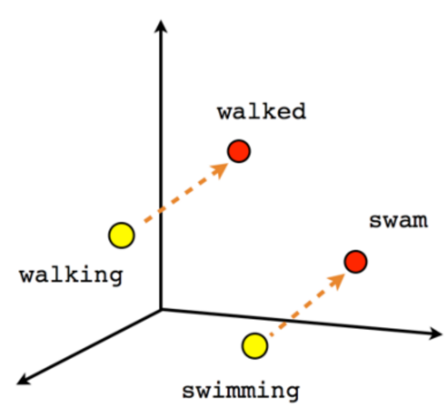
(c) Object color (d) Background color

(a) Lighting (b) Scale

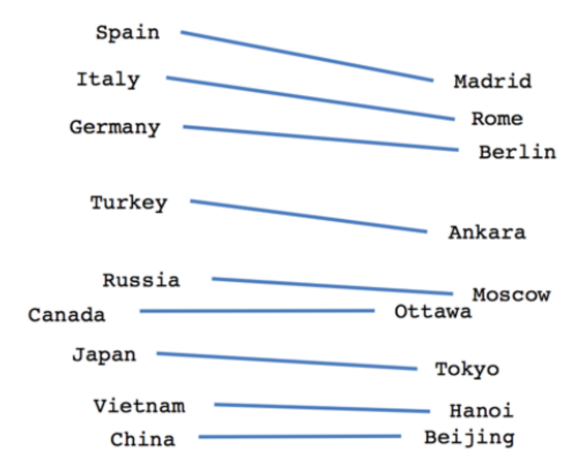
- Data tends to live on flattened space. Tangent space?



Male-Female



Verb tense



Country-Capital

Difficult to find evidences of such phenomenon more formally

Mathematical Toolbox

Reminders about Hilbert Space

We will always work in a Hilbert Space. . .

Hilbert space

- $(\mathcal{H}, \langle ., . \rangle)$ is a (real or complex) Hilbert space, if it is complete, for the norm:

$$\|x\| = \sqrt{\langle x, x \rangle}$$

- A linear operator T is bounded, if:

$$\|Tx\| \leq \|T\|\|x\|$$

Its adjoint is defined via: $\forall x, y \in \mathcal{H}, \langle Tx, y \rangle = \langle x, T^* y \rangle$

- If $TT' = T'T = \mathbf{I}$ and T is bounded, then T' is bounded. We write: $\mathcal{U}(\mathcal{H}) = \{T, TT^* = T^*T = \mathbf{I}\}$

- The spectrum of T is defined as:

$$\mathrm{Sp}(T) = \{\lambda, T - \lambda\mathbf{I} \text{ has no inverse.}\}$$

Compact operator: spectral theorem

- T is compact if $\overline{T\mathcal{B}(0,1)}$ is compact (note that it is automatically bounded). In this case, its spectrum is countable and:

(i) $\mathcal{H} = \overset{\perp}{\bigoplus_{n\in\mathbb{N}}} \ker(T - \lambda_n \mathbf{I})$

(ii) $\forall \lambda \neq 0, \dim\ker(T - \lambda\mathbf{I}) < \infty$

(iii) $\overline{\{\lambda_n\}} \subset \{\lambda_n\} \cup \{0\}$

- *A simple characterisation*: T is compact if and only if it is the limit of compact operators. In particular, if $\dim(T\mathcal{H}) < \infty$, then T is compact.

Reminders about integration

Fourier Tools super useful to this class (sometimes tricky)
and the notion of Integral Operators.

Integral Operator

- An example of operator is given on $L^2(\mathcal{X})$, with Integral Operators:

$$Kf(u) = \int_t k(u,t)f(t)\,dt$$

- This is indeed an operator of $L^2(\mathcal{X})$ if for example:

$$\exists C > 0, \int_{t,u} |k(u,t)|^2\,dt \leq C$$

- Here, the adjoint is given by: $K^*f(t) = \int_u f(u)\overline{k(u,t)}\,du$
- The kernel of K^*K is given by $w(u,t) = \int_z \bar{k}(z,u)k(z,t)\,dz$

and: $\|Kf\|^2 = \int_u |Kf(u)|^2\,du = \int_u \overline{f(u)}(K^*Kf)(u)\,du$

Schur Test

- Estimating the norm of a kernel will be crucial in the following...
- We have the **Schur test**:

Let: $$Kf(u) = \int_v k(u,v) f(v)\, dv$$

If $$\int_u |k(u,v)|\, du \leq C_1 \quad \text{and} \quad \int_v |k(u,v)|\, dv \leq C_2$$

then: $$\|K\| \leq \sqrt{C_1 C_2}$$

Convolutions in $\mathbb{R}^d$

- Here, $\mathcal{X} = \mathbb{R}^d$ and remind that:

$$f \star g(x) \triangleq \int_{\mathbb{R}^d} f(x-y)g(y)\, d\mu(y)$$

- (Young's inequality) If: $\dfrac{1}{r} + 1 = \dfrac{1}{p} + \dfrac{1}{q}$, $f \in L^p(\mathbb{R}^d), g \in L^q(\mathbb{R}^d)$

  then:

$$\|f \star g\|_r \leq \|f\|_p \|g\|_q$$

- Setting of interest in this class: $f \in L^2$, g fast decay, then

If $\mathcal{L}_a x \triangleq x(u-a)$ and $Wx = x \star \psi$ then $\mathcal{L}_a W = W \mathcal{L}_a$

Reminder about Fourier

$$\mathcal{F} : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$$

$$\mathcal{F}x(\omega) \triangleq \hat{x}(\omega) \triangleq \int_{\mathbb{R}^d} e^{-i\omega^T u} x(u)\, du$$

Isometry: $\|\mathcal{F}x\|_2 = \|x\|_2$

Hermitian symmetry: f real implies that $\hat{f}^*(x) = \hat{f}(-x)$

$$x \star y(u) \triangleq \int_{\mathbb{R}^d} x(u-t)y(t)\, dt$$

$$x \star y(u) \xrightarrow{\mathcal{F}} \hat{x}(\omega)\hat{y}(\omega)$$

$$\frac{d}{du}x(u) \xrightarrow{\mathcal{F}} i\omega\hat{x}(\omega)$$

$$x_a(u) \triangleq x(u-a) \xrightarrow{\mathcal{F}} e^{-i\omega^T a}\hat{x}(\omega)$$

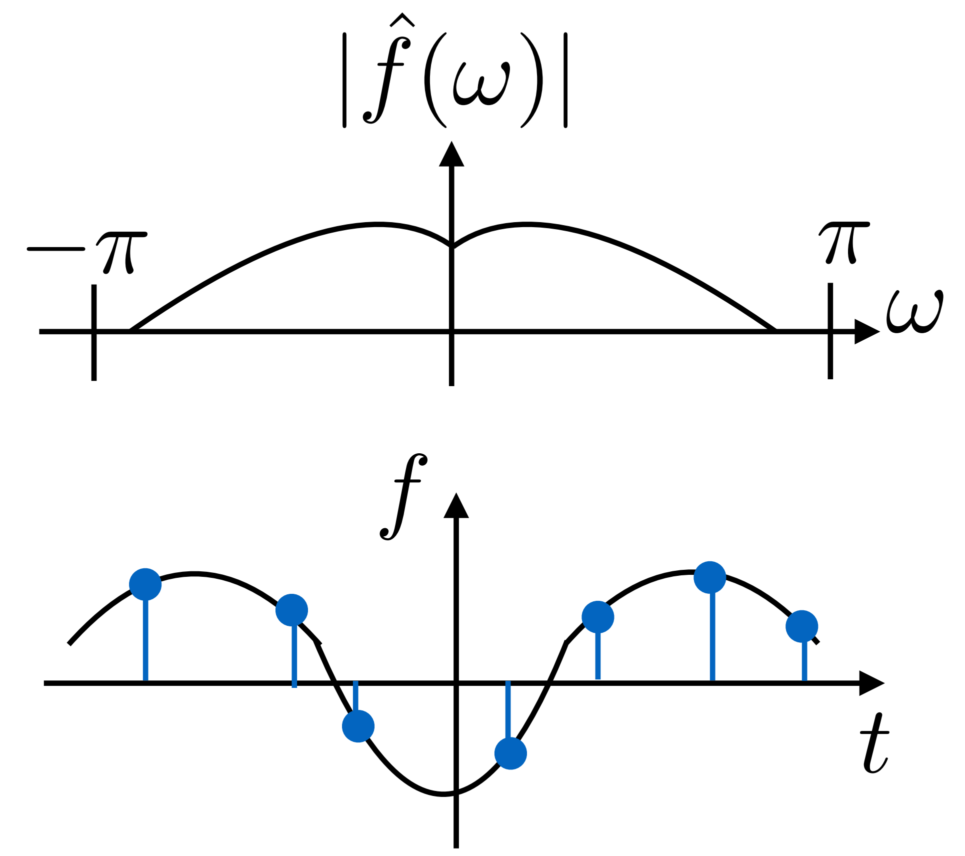
Discrete image to continuous image

- An image x corresponds to the discretisation of a **physical** anagogic signal (light!) and is thus continuous by nature.

Say we want to estimate f with:

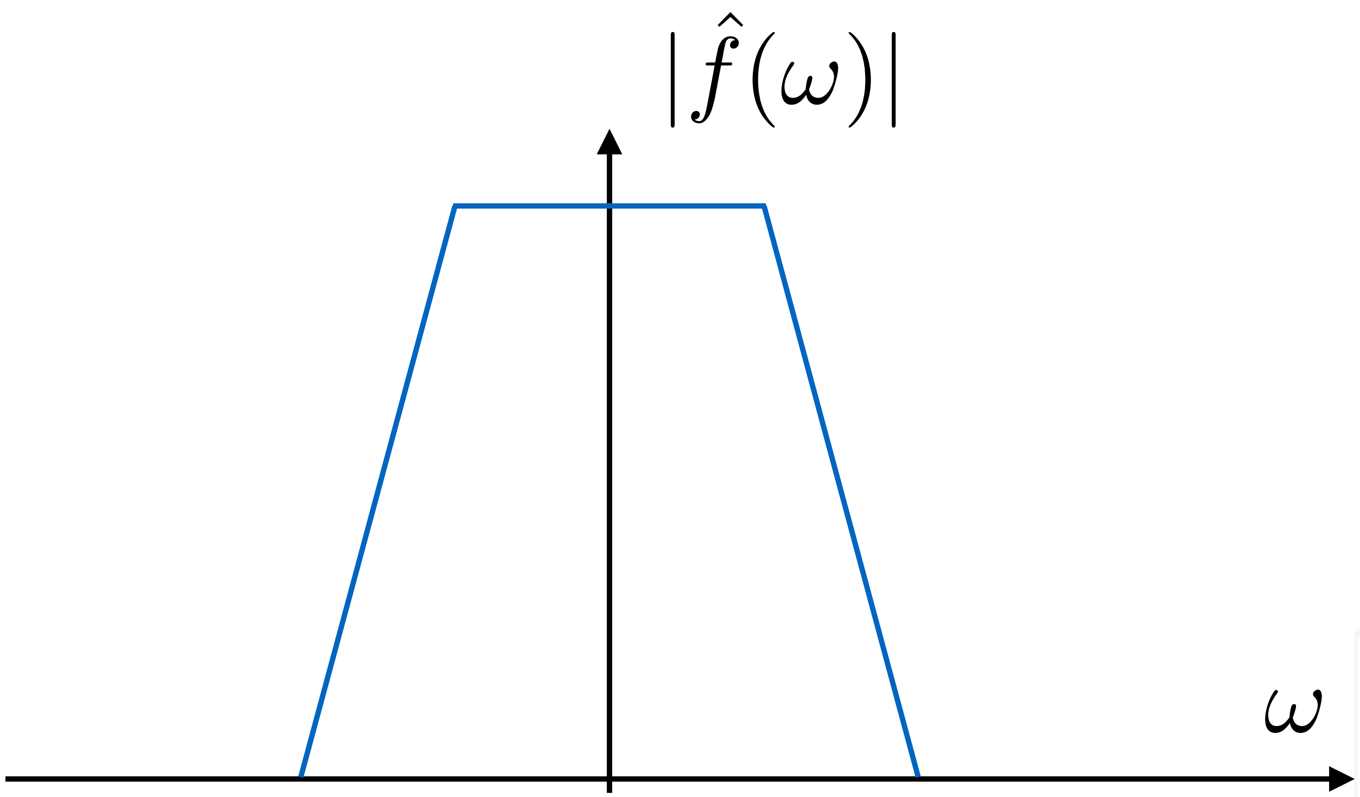
$$\tilde{f}(t) = \sum_{n=-\infty}^{\infty} f(n)\delta_{t-n}$$

Only valid if $\text{support}(\hat{f}) \subset [-\pi, \pi]$

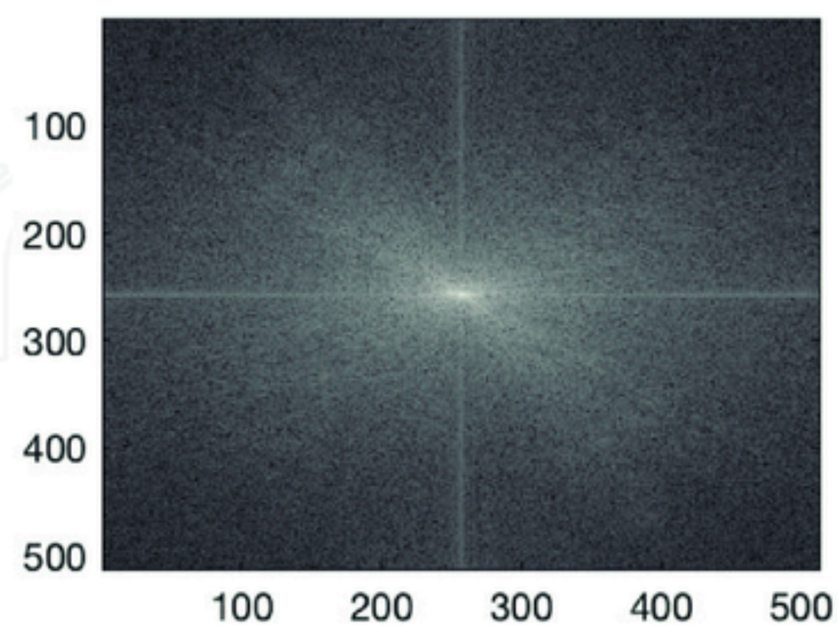


Why is Fourier analysis useful?

$$|\widehat{\mathcal{L}_a f - f}(\omega)| = |\hat{f}(\omega)(e^{i\omega^T a} - 1)| \leq |\hat{f}(\omega) \sin \omega^T a| \leq |\hat{f}(\omega) w^T a|$$



(a)



(b)

Convolutions!

Convolutional Kernel

- For illustration purpose, consider

$$Kf(u) = \int_{\mathbb{R}^d} f(v)\psi(u-v)\,dv = (f \star \psi)(u)$$

- Then,

$$K^*f(v) = \int_{\mathbb{R}^d} f(u)\bar{\psi}(u-v)\,du = (f \star \breve{\psi})(v)$$

where: $\breve{\psi}(u) = \bar{\psi}(-u)$

$$K^*Kf = \breve{\psi} \star \psi \star f \quad \text{and} \quad \widehat{\breve{\psi} \star \psi}(\omega) = |\hat{\psi}(\omega)|^2$$

leads to:

$$\|Kf\|^2 = \int_{\mathbb{R}^d} |\hat{f}(\omega)|^2 |\hat{\psi}(\omega)|^2\,d\omega = \langle f, K^*Kf \rangle$$

Convolutional Frame

- Consider $Wx = \{x \star \psi_n\}_n$ with norm $\|Wx\|^2 = \sum_n \|x \star \psi_n\|^2$
- We say that W is a convolutional frame if:

$$A\|x\|^2 \leq \sum_n \|x \star \psi_n\|^2 \leq B\|x\|^2$$

or

$$A \leq \sum_n |\hat{\psi}(\omega)|^2 \leq B$$

- Furthermore, the frame is tight if $A = B$.

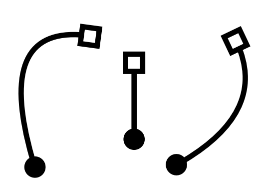
Covariance via convolution

- We say that L is covariant with W if $WL = LW$

- We say that A is invariant to L if $AL = A$

- If W (e.g., convolution), ρ (e.g., point-wise non-linearity) are covariant and if A is invariant to L then
$$\Phi x = AW_J \rho W_{J-1} \rho W_{J-2} ... W_1 x$$
is invariant. Indeed:
$$\Phi L x = ALW_J \rho ... W_1 x = \Phi x$$

- It is also possible to have only an approximate covariance and one measure it via the norm of:
$$[W, L] = WL - LW$$

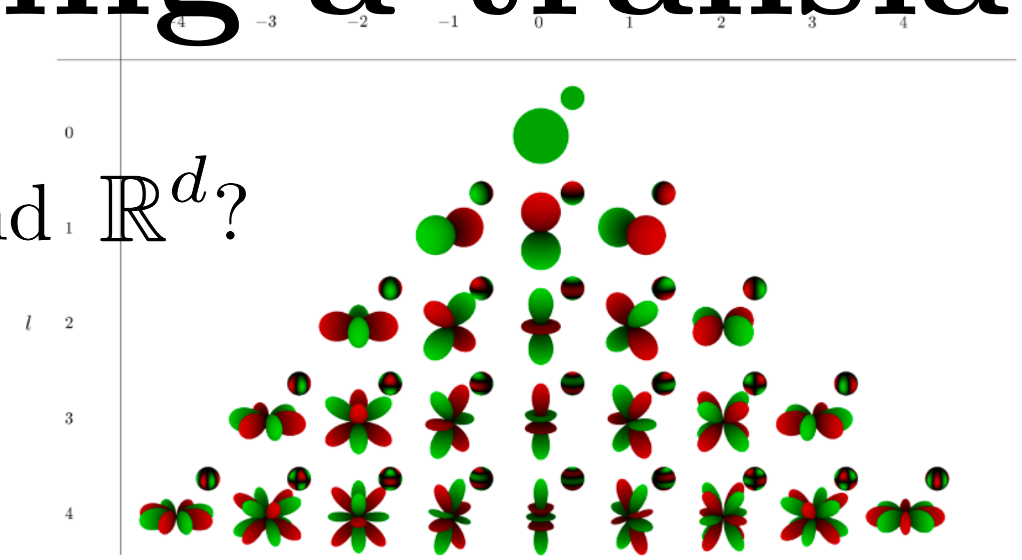
example: deformation 

Group theory for analysing convolutions

How can we design and characterise convolutions along a group?

Fourier on a circle, decomposing a translation

How can extend Fourier beyond $\mathbb{R}^d$?



- Derivation is the infinitesimal generator of translation...

$$L_a(e^{i.t})(\omega) = e^{i(\omega+a)t} = e^{i\omega t}e^{iat}$$

$\mathrm{span}(e^{i.t})$ is stable by translation...

$$\widehat{L_{-a}x}(\omega) = \hat{x}(\omega)e^{i\omega a} = \sum_n \frac{a^n}{n!}(i\omega)^n \hat{x}(\omega)$$

$$x(u+a) = \sum_n \frac{a^n}{n!} x^{(n)}(u)$$

Groups

- We remind that a group is a set G equipped with . and a neutral element e $s.t.$ $\forall x, \exists x^{-1} : x.x^{-1} = x^{-1}.x = e$
- Examples are given by: $\mathbb{R}^d, \mathbb{F}_p, SO_d(\mathbb{R}), SU_d(\mathbb{C}), ...$
- We'll assume all our groups are equipped with an invariant distance (not restrictive for compact groups) $\forall g, d(g.g', g.g'') = d(g', g'')$
- In practice, we'll discuss only: $\mathbb{R}^d, [0, 2\pi]^k$ product/semi product of those

Haar measure on a group

- If G is locally compact, there exists a non-0 measure unique (up to multiplication) measure

$$\forall g \in G, \quad \mu(A) = \mu(g.A)$$

where $a.x(g) \triangleq L_a x(g) \triangleq x(a^{-1}.g)$

- For compact/abelian groups, the measure is unimodular:

$$\forall g \in G, \; \mu(g.A) = \mu(A.g)$$

- We write:

$$L^2(G) = \{f \text{ measurable}, \int_G |f(g)|^2 \, d\mu\}$$

Convolution along a group

- Again, introduce:

$$(a \star b)(g) \triangleq \int_G a(\tilde{g})b(\tilde{g}^{-1}g)$$

- (Young's inequality)
  For $a \in L^p(G), b \in L^q(G), \frac{1}{r} + 1 = \frac{1}{p} + \frac{1}{q}$, we get:

$$\|a \star b\|_r \leq \|a\|_p \|b\|_q$$

- $a \star b = b \star a$ if and only if G is commutative.

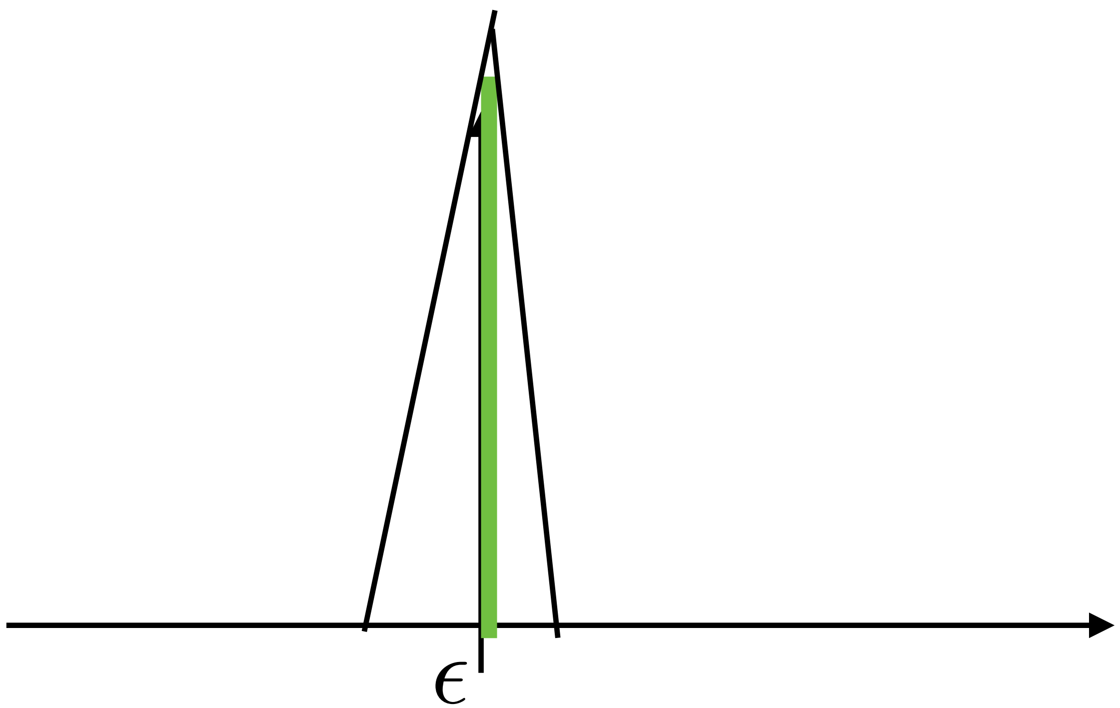
Can we recover a notion of Fourier? Invariance?

Unit approximation

- Convolution in $\mathbb{R}^d$ has no neutral elements.

- Yet, there exists a sequence $(\delta_n)_n, \delta_n \geq 0, \mathrm{supp}(\delta_n) \to 0$

  $\|\delta_n \star f - f\|_1 \to 0$ and $\|\delta_n \star f - f\| \to 0$



- In $\mathbb{R}^d$, if $\hat{\delta}_n(x) = e^{-\frac{\|x\|^2}{2n^2}}$ then $\|\delta_n\|_1 = 1$, $\delta_n \in L^2(\mathbb{R}^d)$ and $\delta_n \geq 0$

Covariant operators

- Let $W : L^1(\mathbb{R}^d) \to L^1(\mathbb{R}^d)$ be a bounded operator, s.t.:

(i) $W\mathcal{L}_a = \mathcal{L}_a W, \forall a$

(ii) $\exists f \in L^1(G), W\delta_n \to f$

$$\iff \quad Wx = x \star f$$

with $f \in L^1(\mathbb{R}^d)$

Invariant operators

- Let $A : L^1(G) \to \mathbb{R}$ be a bounded operator, then:

$$A\mathcal{L}_a = A, \forall a \qquad \Longleftrightarrow \qquad \exists \lambda, Ax = \lambda \int_G x(g)\, d\mu(g)$$

Unitary representations of groups

- We say that $\rho : G \to \mathcal{U}(\mathcal{H})$ is a representation if it is a continuous morphism. Note that potentially, here: $\dim \mathcal{H} = \infty$

- This will be our main tool to analyse convolutions, via:
$$\begin{aligned} \rho : G &\to \mathcal{U}(L^2(G)) \\ g &\to (f \to L_g f) \end{aligned}$$

- And thus, if W is covariant with translations $WL_g = L_g W$ then the characteristic subspace are stabilised.

- What can we say about those invariant subspaces? Favorable case: matrix are diagonal.

Invariant and irreducible subspaces

- Def.: $F \subset \mathcal{H}$ is an invariant subspace of a representation ρ if it is closed and:
$$\forall g, \rho(g)F \subset F$$
- F is invariant if and only if $F^{\perp}$ is invariant.
- Def.: ρ is irreducible on $\mathcal{H}$ if its only invariant subspaces are $\mathcal{H}$ and $\{0\}$. We also say that $\mathcal{H}$ is irreducible.
- Ideally, we would like to write $\mathcal{H}$ as $\bigoplus_{n \in \mathbb{N}} \mathcal{H}_n$ s.t. $\rho(g)_{|\mathcal{H}_n}$ is irreducible.

Compact abelian group

- Let's give a couple of examples in the compact abelian case.

- Example 1: $\mathbb{R}^N$, with $\mathcal{F}_N : \mathbb{R}^N \to \mathbb{R}^N$ and $\mathcal{L}x[n] \triangleq x[n+1]$

  and $\mathcal{F}_N x[k] = \sum_{n=0}^{N} x[n] e^{-2i\pi k \frac{n}{N}}$

  $\mathcal{H}_n = \text{span}\{k \to e^{2i\pi k \frac{n}{N}}\}$

- Example 2: $L^2([0,1])$, with $\mathcal{F} : L^2([0,1]) \to \ell^2(\mathbb{Z})$ and $\mathcal{L}_a x(u) \triangleq x(u-a)$

  and $\mathcal{F}x[n] = \frac{1}{2\pi}\int_0^{2\pi} x(u) e^{-2in\pi u}\, du$

  $\mathcal{H}_n = \text{span}\{u \to e^{2i\pi n u}\}$

Commutative groups, compact, finite dimension

- Let $\rho : G \to \mathcal{U}(\mathcal{H})$ be a group action.

- Theorem (Peter-Weyl): Assume G is compact. Then,

$$\mathcal{H} = \bigoplus_{n \in \mathbb{N}} \mathcal{H}_n \quad \text{with} \quad \dim \mathcal{H}_n < \infty$$

  where each subspace $\mathcal{H}_n$ is an invariant subspace of ρ, ie:

$$\forall g, \rho(g)\mathcal{H}_n \subset \mathcal{H}_n$$

- Theorem: If G is also abelian, then $\dim \mathcal{H}_n = 1$

TLDR: Compact abelian groups behave like $[0, 2\pi]^d$

Invariant Representations with the Scattering Transform

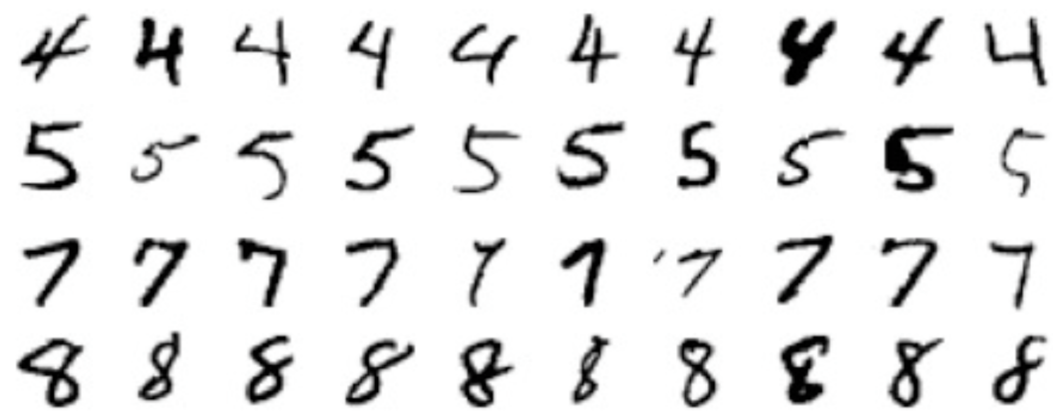
Models for natural signals

We will discuss widely the Scattering Transform.

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

- Successfully used in several applications:
  - Digits
  - Textures

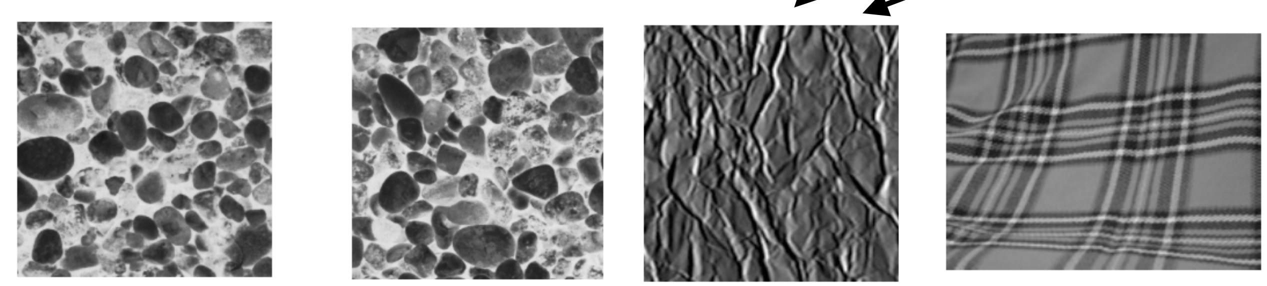
All variabilities are known



Small deformations +Translation

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

Rotation+Scale



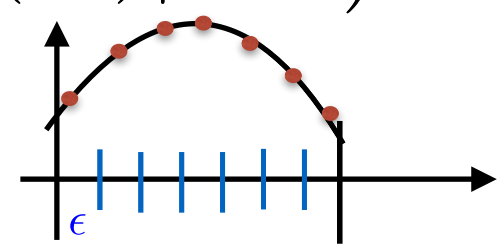
- The design of the scattering transform is guided by the euclidean group.

- A scattering transform is a combination of complex-valued wavelets and modulus non-linearity.

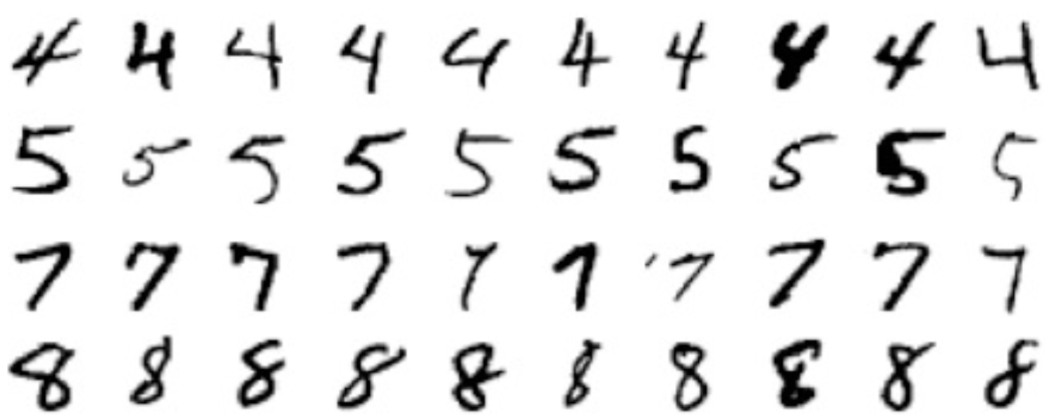
Model on the data: low dimensional manifold hypothesis?

- Low dimensional manifold: dimension up to 6. Not higher:

Property: if $f : \mathbb{R}^D \to [0,1]$ is 1-Lipschitz, then let $N_\epsilon = \arg\inf_N \sup_{i \leq N} (|f(x) - f(x_i)| < \epsilon)$.
Then $N_\epsilon = \mathcal{O}(\epsilon^{-D})$



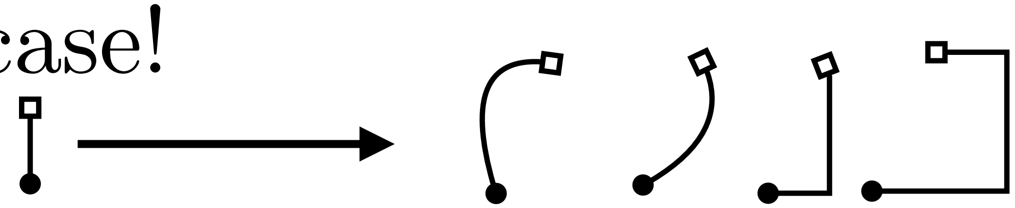
- Can be true for MNIST…



All variabilities are known

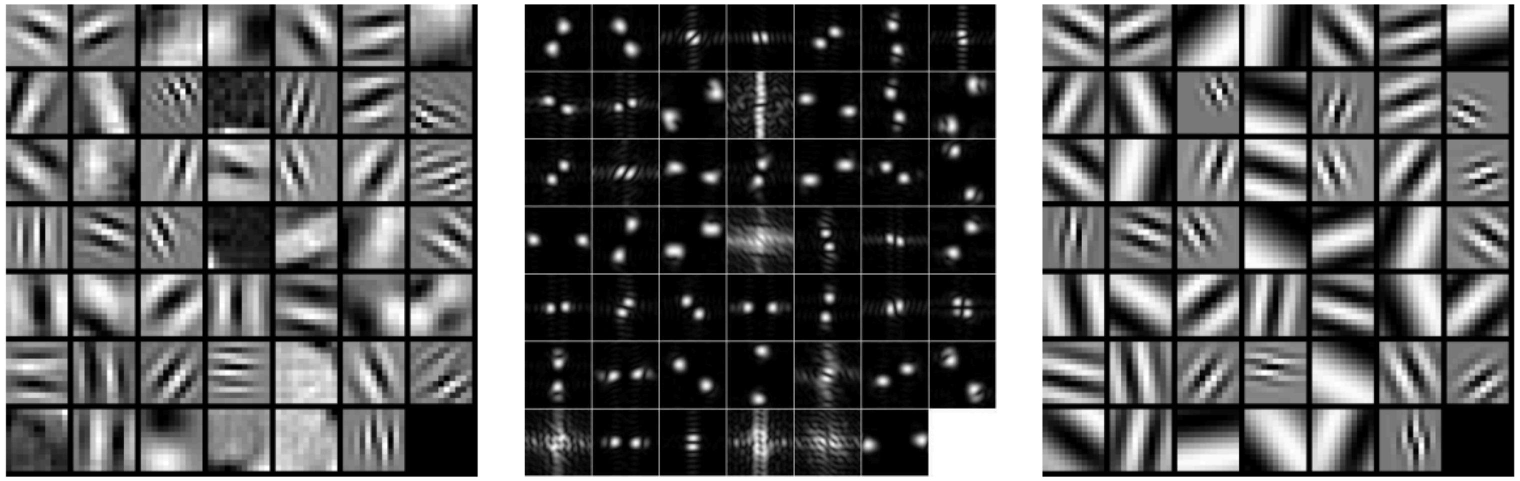
Small "limited" deformations +Translation

- Yet high dimensional deformations are an issue in the general case!



Model for the first layer

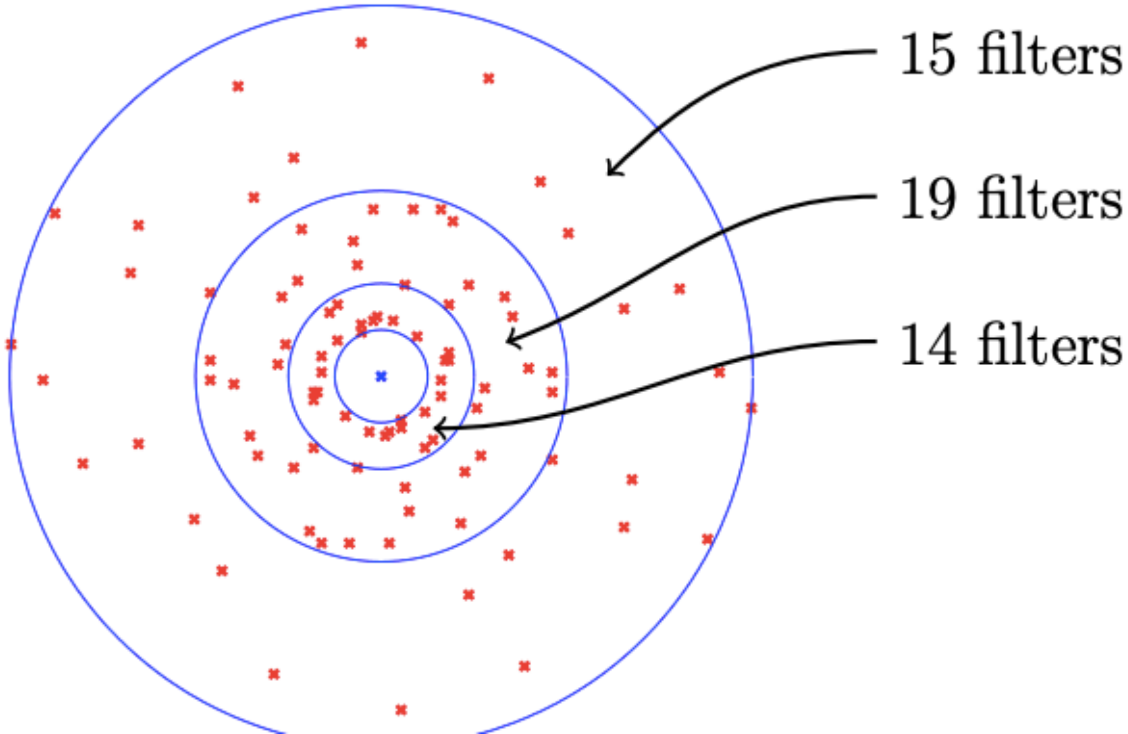
$$\psi_{C,D,\xi}(u) = Ce^{-u^T Du} e^{iu^T \xi}$$



Ref.: I Waldspurger's phd

- Consider Gabor filters and fit the model.

This principle is core in many models (V1, Scattering,...)



Model for the second layer

Ref.: I Waldspurger's phd

First layer:

$$\psi_\lambda(u)$$

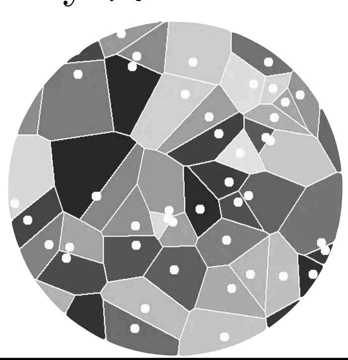
Second layer:

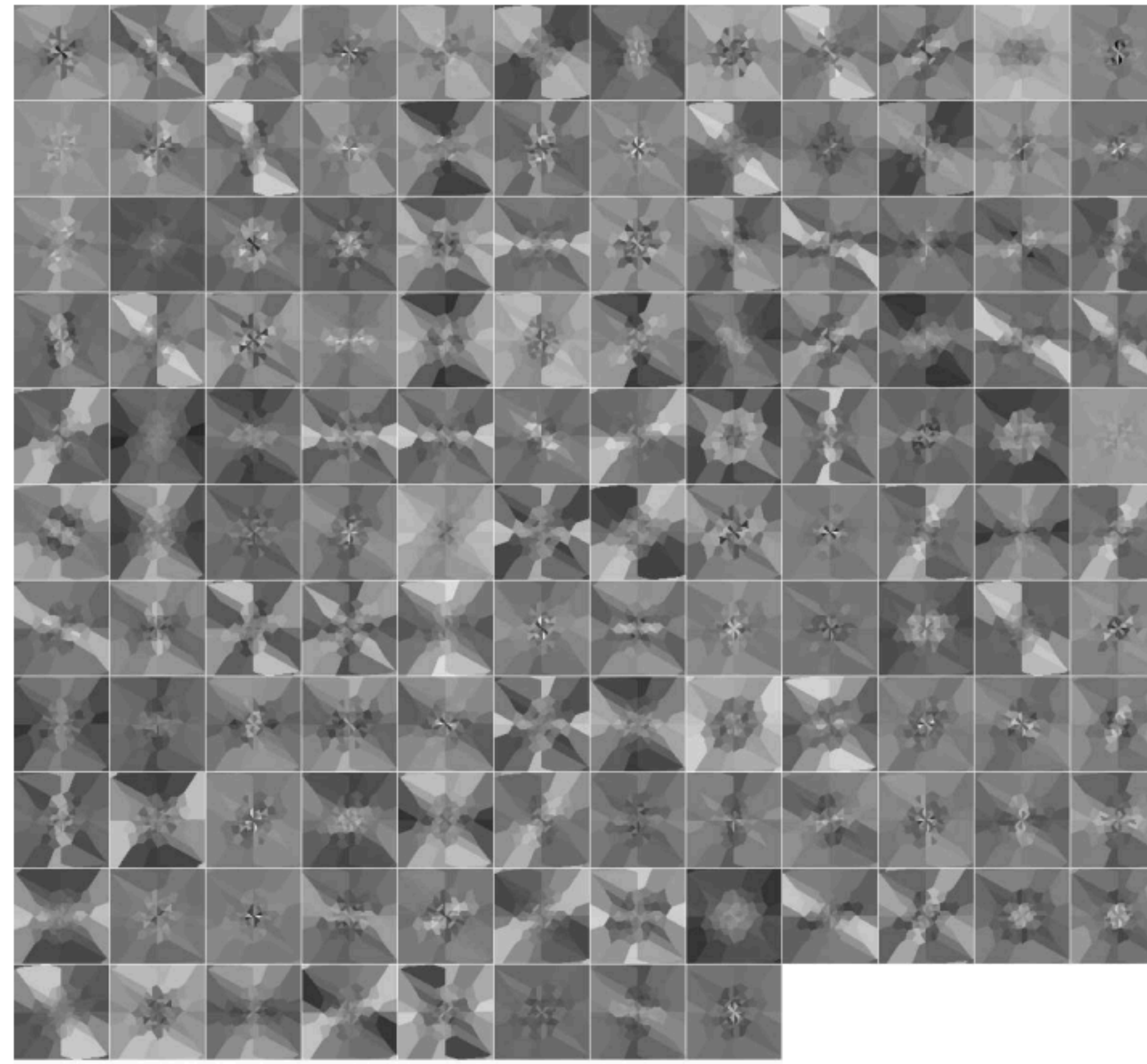
$$\psi(u,\lambda) \approx \phi^1(u) \times \phi^2(\lambda)$$

Recombines along λ

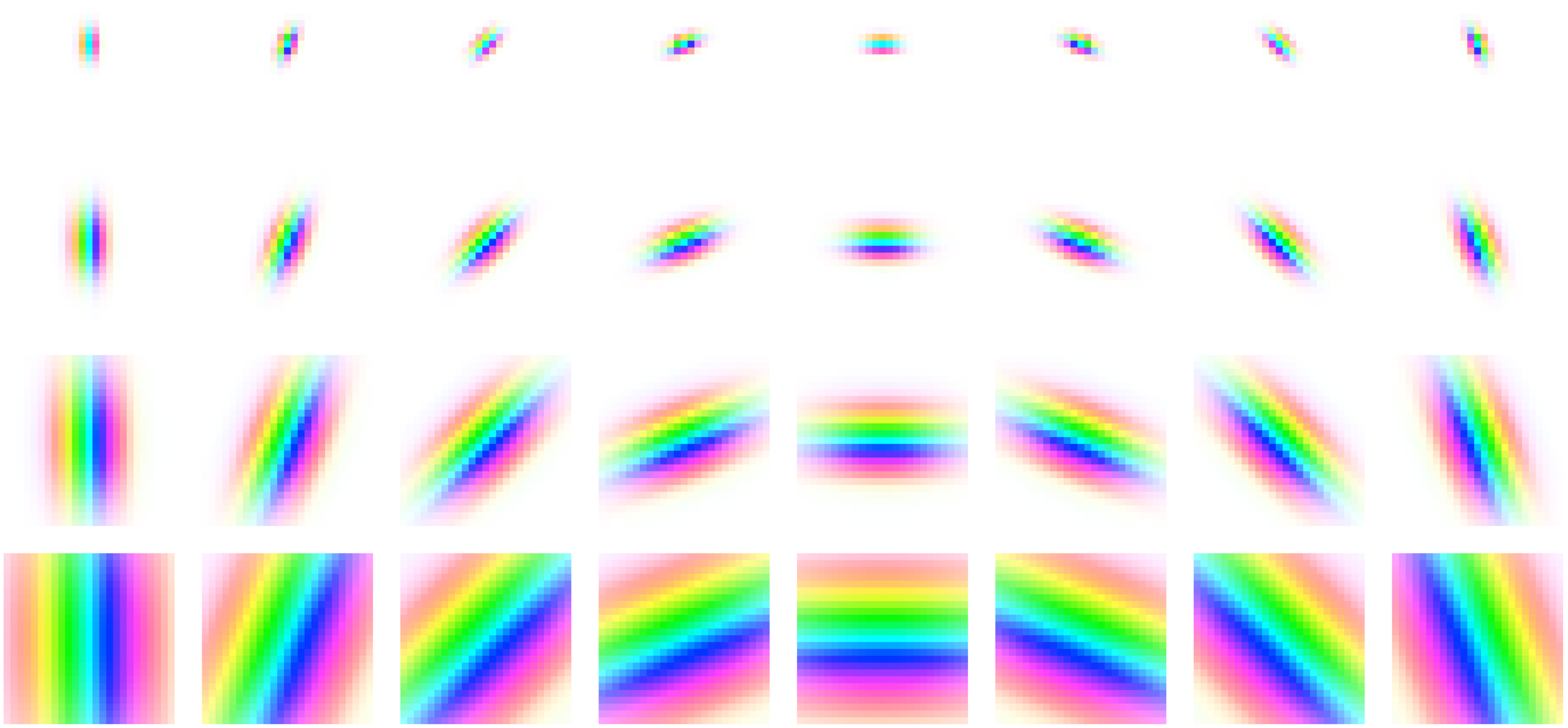
Why was this possible?
We were aware of the topology
of the previous layer!

Take the Voronoi diagram associated to central frequency λ and color according to $\phi^2(\lambda)$





Visualisation of ϕ^2
in the frequency plane
(by reindexing along frequency topology)



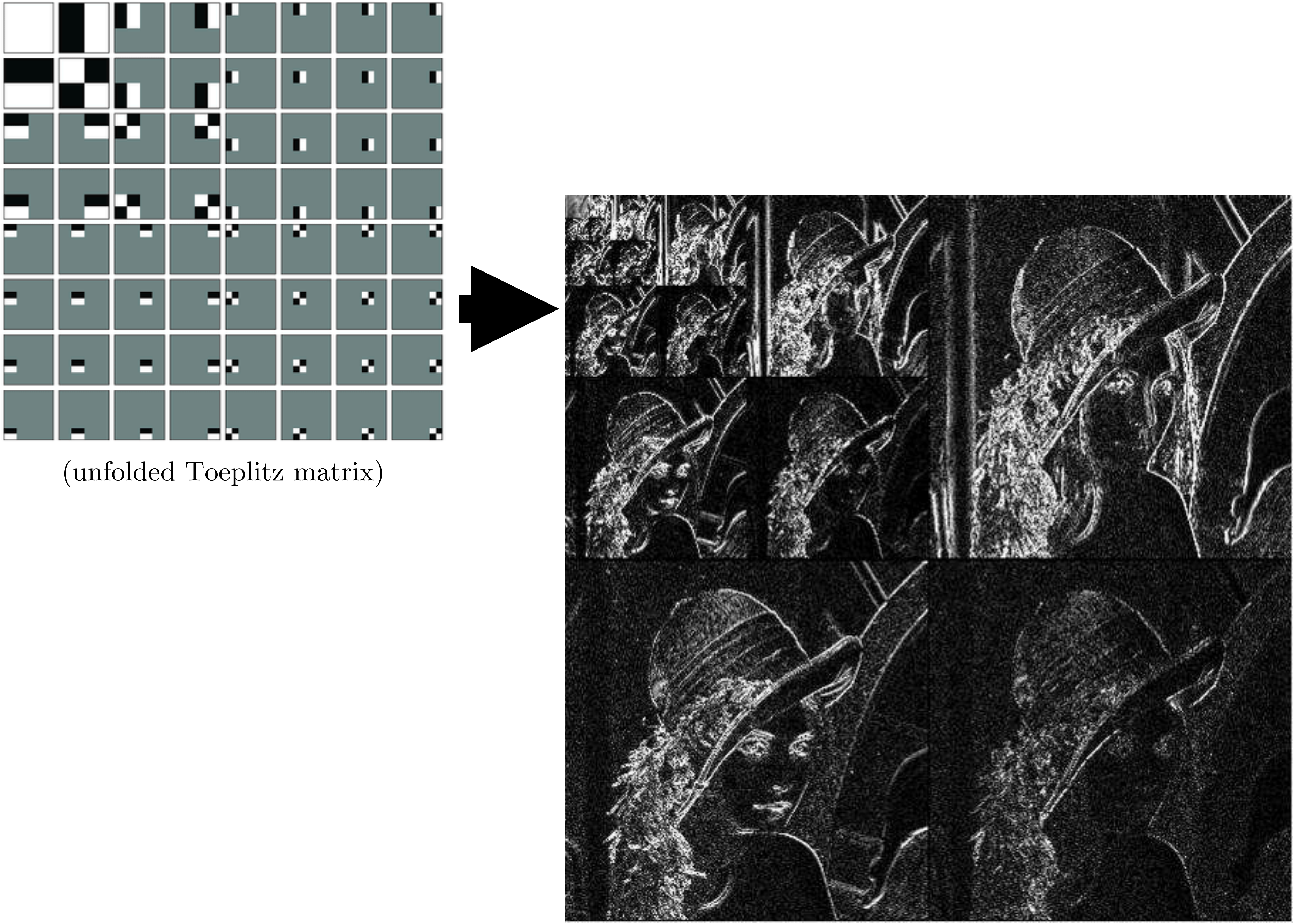
$$\psi(u) = \frac{1}{2\pi\sigma} e^{-\frac{\|u\|^2}{2\sigma}} (e^{i\xi . u} - \kappa)$$

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

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

The Gabor wavelet

Another example: Haar Wavelets

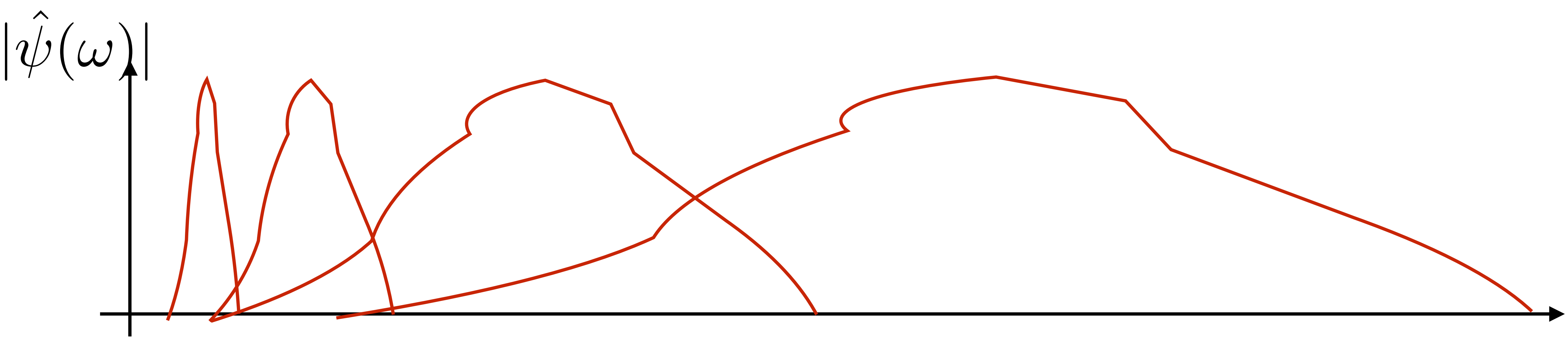


(unfolded Toeplitz matrix)

Wavelets on the real line

- $\psi \in L^1(\mathbb{R})$ is a wavelet iff $\int \psi(u)du = 0$ and $\int |\psi|^2(u)du < \infty$
- Typically localised in time and frequency, via Heisenberg principle

$$\psi_j(u) = \frac{1}{2^j}\psi(\frac{u}{2^j}) \xrightarrow{\mathcal{F}} \hat{\psi}_j(\omega) = \hat{\psi}(2^j\omega)$$



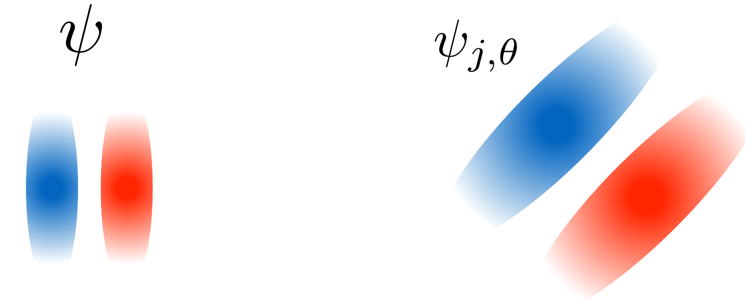
Wavelet transform

2D-Wavelets

- ψ is a wavelet iff $\int \psi(u)du = 0$ and $\int |\psi|^2(u)du < \infty$
- Typically localised in space and frequency.
- Rotation, dilation of a wavelets:

Group action!

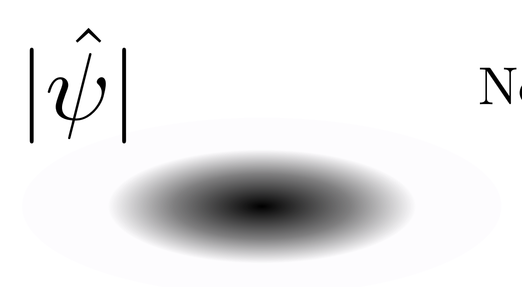
$$\psi_{j,\theta} = \frac{1}{2^{2j}}\psi(\frac{r_\theta(u)}{2^j})$$



- Design wavelets selective to **rotation** variabilities.

Isotropic

VS



Non-Isotropic

2D-Wavelet Transform

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

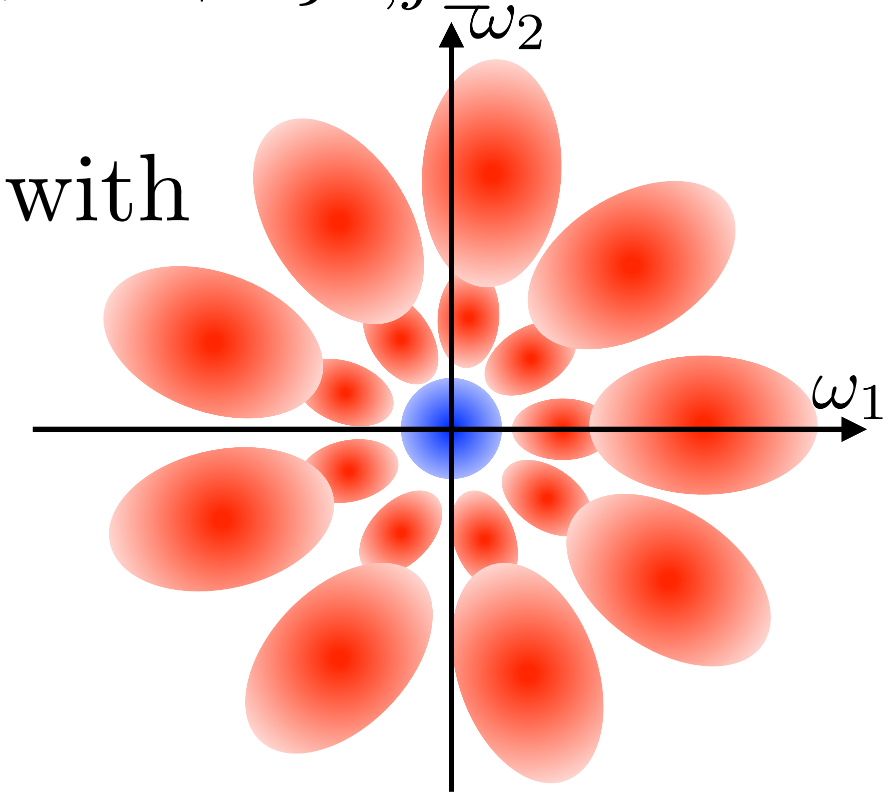
- Isometric and linear operator of L^2, with

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

- Covariant with translation L_a:

$$WL_a = L_a W$$

- $\|x \star \psi\|_1$ is small. (*sparsity*)



Ref.: Group Invariant Scattering, Mallat S

Admissible wavelets

- A family of wavelets $\{\psi_\lambda\}_{\lambda \in \Lambda_J}$ and low-pass filter ϕ_J is $\epsilon-$admissible if:

$$(1-\epsilon)\|x\|^2 \leq \sum_{\lambda \in \Lambda} \|x \star \psi_\lambda\|^2 + \|x \star \phi_J\|^2 \leq \|x\|^2$$

or

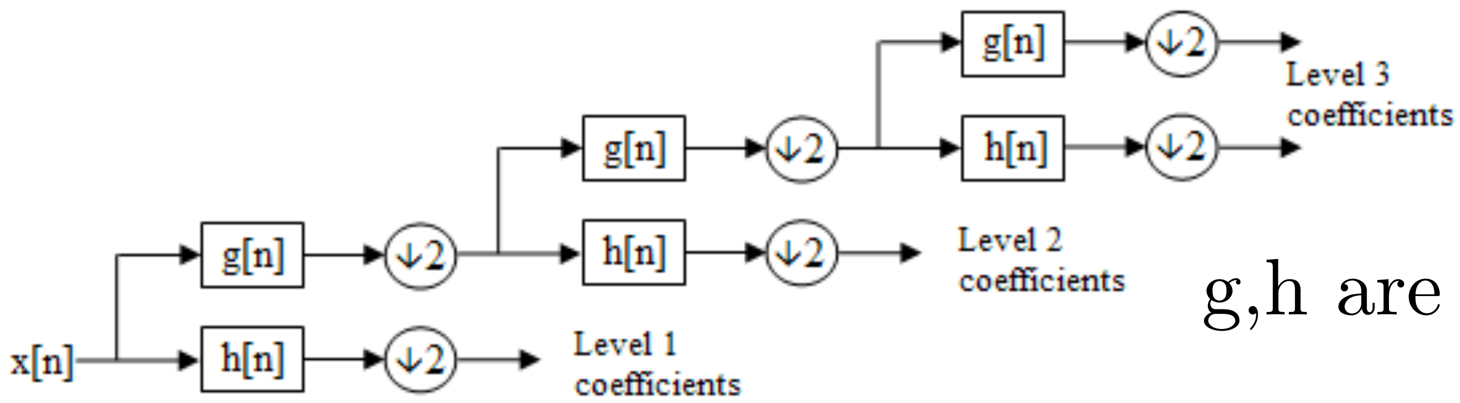
$$(1-\epsilon) \leq \sum_{\lambda \in \Lambda} |\hat{\psi}_\lambda|^2(\omega) + |\hat{\phi}_J|^2(\omega) \leq 1$$

- In practice, one adapts ϕ_J and we use:

$$\Lambda = \{(j,\theta) \in \mathbb{Z} \times SO_d(\mathbb{R}), j \leq J\}$$

Wavelet Transform implementation as a CNN

Implementation of a Fast Wavelet Transform algorithm



g,h are 3x3 filters

VGG implementation: