La Malédiction Mathématiques des Données en Grande Dimension

Stéphane Mallat

École Normale Supérieure

Un Changement de Paradigme

- Pratique classique de la science et de la connaissance:

Observations

↓

Intelligence d'experts

↓

Modélisation

max 10 variables explicatives

↓

optimisation des variables
sur quelques données

- Lent, couteux, et difficile pour des problèmes complexes.

Un Changement de Paradigme

Masses de données

↓

Apprentissage automatique

↓

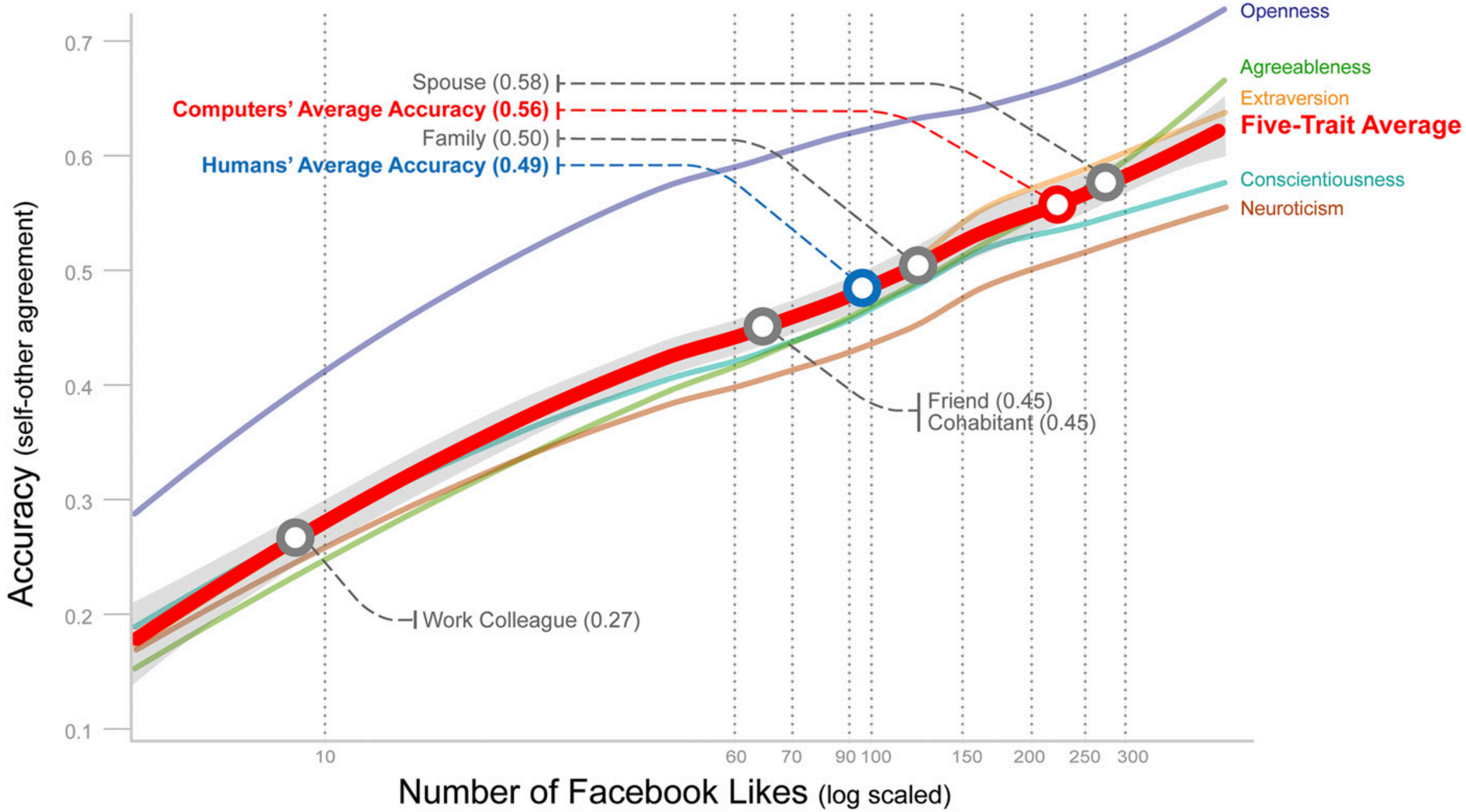
Solution: vote de beaucoup d'indicateurs faibles

- Intelligence: nécessite beaucoup de données et de mémoire.

- Applications:
  - Diagnostiques (médical, industriel)
  - Reconnaissance (images, sons, langage naturel...)
  - Prédictions (sciences sociales, marketing, finance, sciences...)

Un Changement de Paradigme

Personality Evaluation (Proc. Nat. Acad. Sciences 2014)



Le Menu

- Malédiction de la dimensionalité
- Enjeux mathématiques de l'apprentissage
- Paysage scientifique et industriel

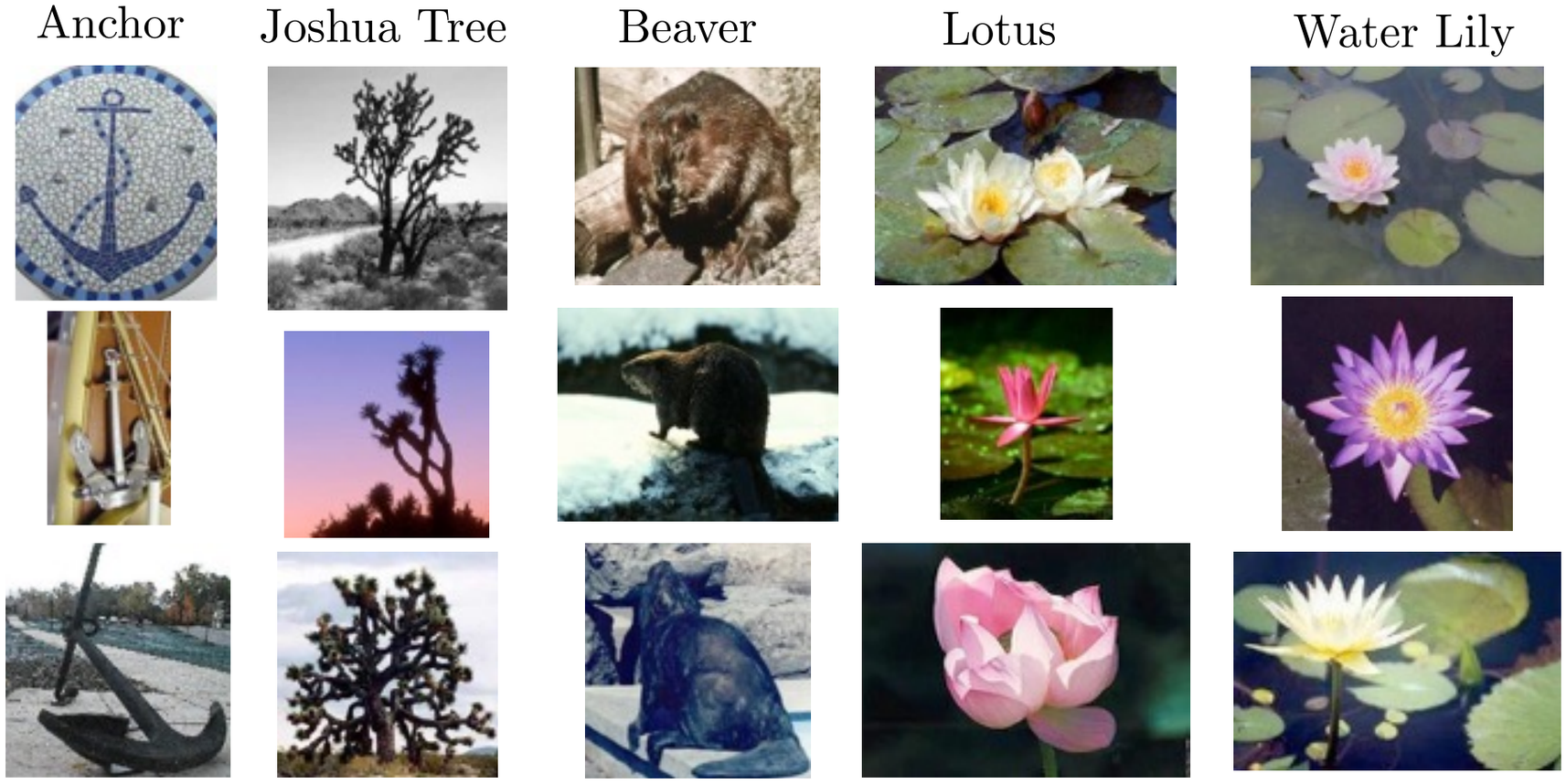
High Dimensional Learning

- 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

Need to find informative invariants

High Dimensional Learning

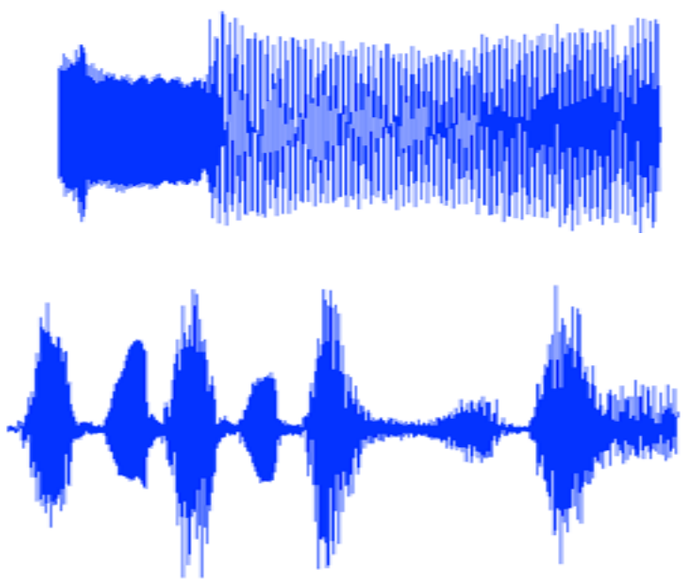
- High-dimensional $x = (x(1), ..., x(d)) \in \mathbb{R}^d$:

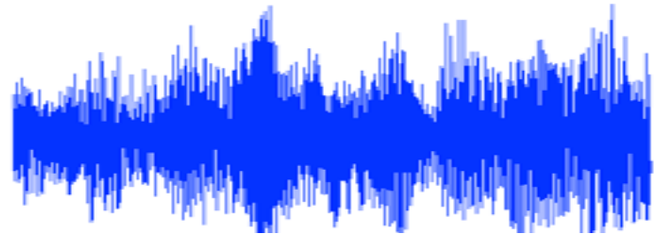
- **Classification:** estimate a class label $f(x)$

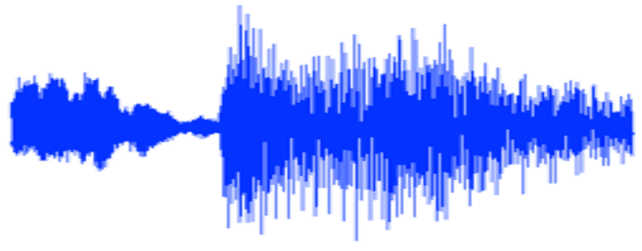
  given n sample values $\{x_i\,,\, y_i = f(x_i)\}_{i \leq n}$

Audio: instrument recognition

Huge variability inside classes







High Dimensional Learning

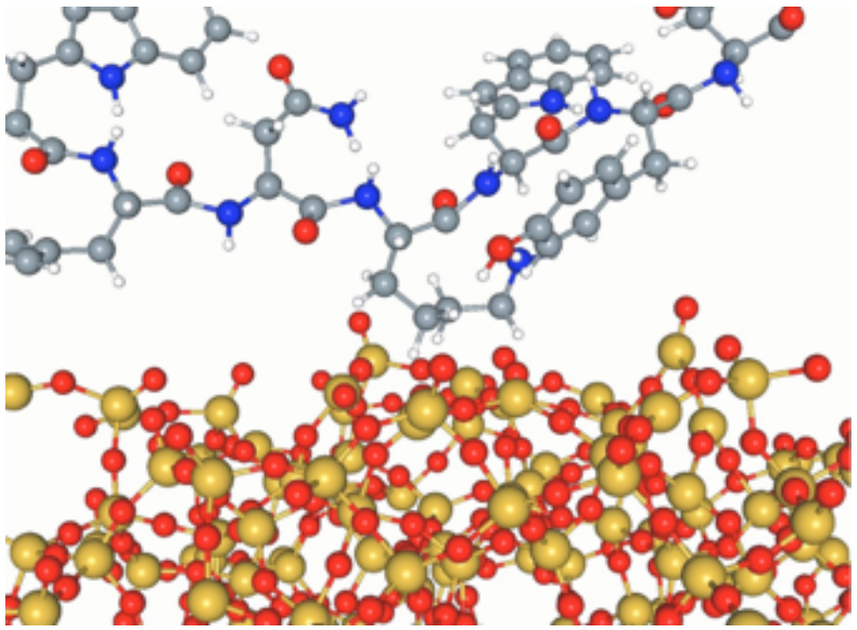
- High-dimensional $x = (x(1), ..., x(d)) \in \mathbb{R}^d$:

- **Regression:** approximate a *functional* $f(x)$
  given n sample values $\{x_i\,,\, y_i = f(x_i) \in \mathbb{R}\}_{i \leq n}$

  Physics: energy $f(x)$ of a state vector x

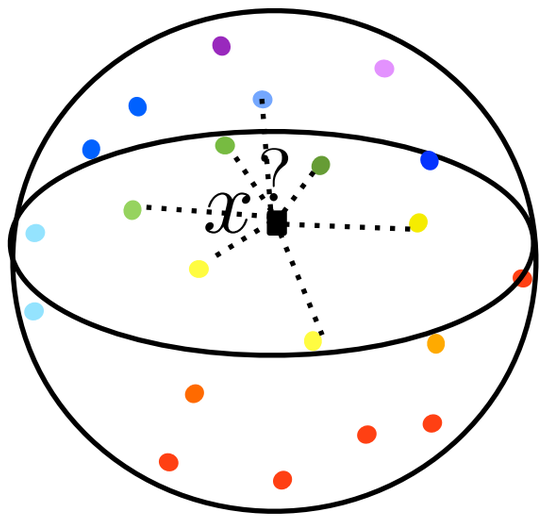
Astronomy

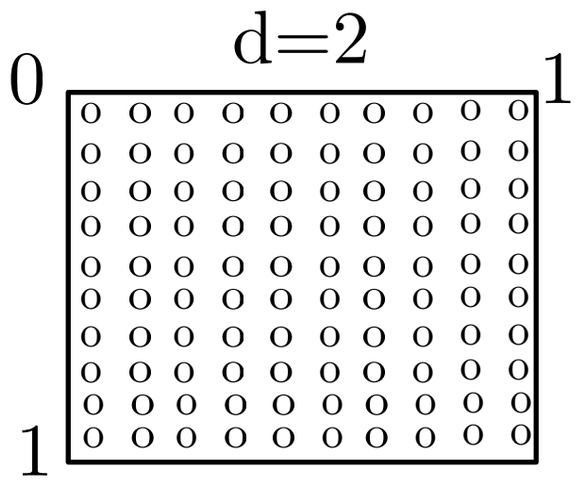
Quantum Chemistry



Curse of Dimensionality

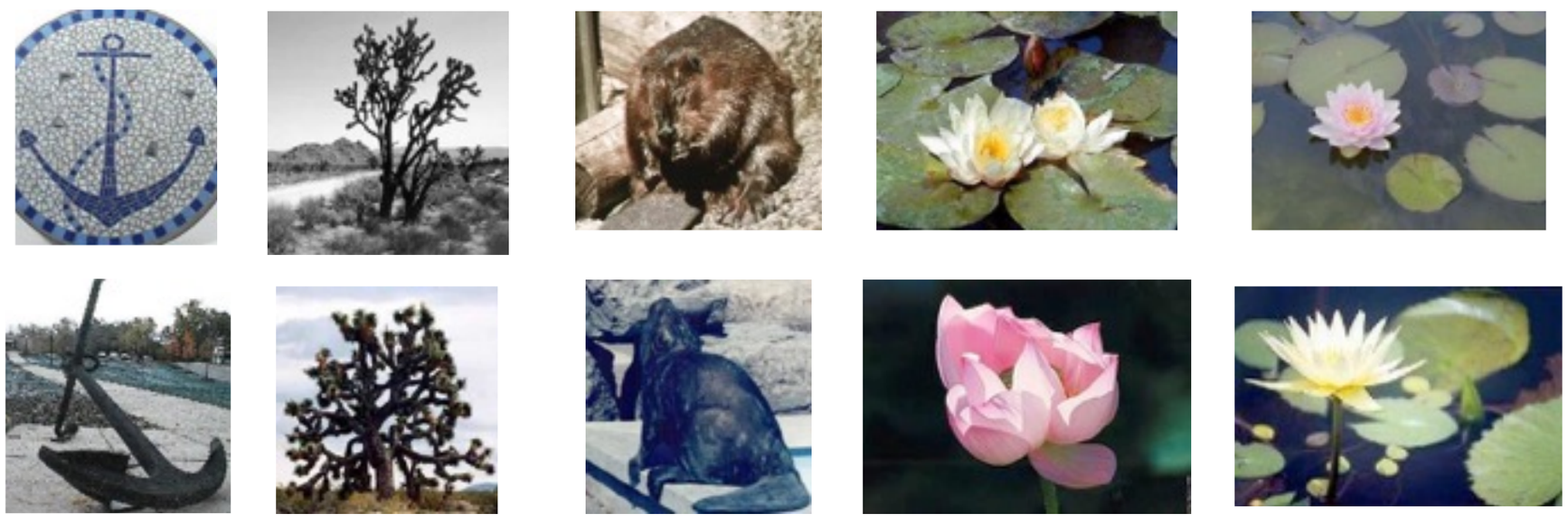
- $f(x)$ can be approximated from examples $\{x_i\,,\, f(x_i)\}_i$ by local interpolation if f is regular and there are close examples:





- To cover $[0,1]^d$ at a distance 10^{-1} we need 10^d points

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



Learning by Euclidean Embedding

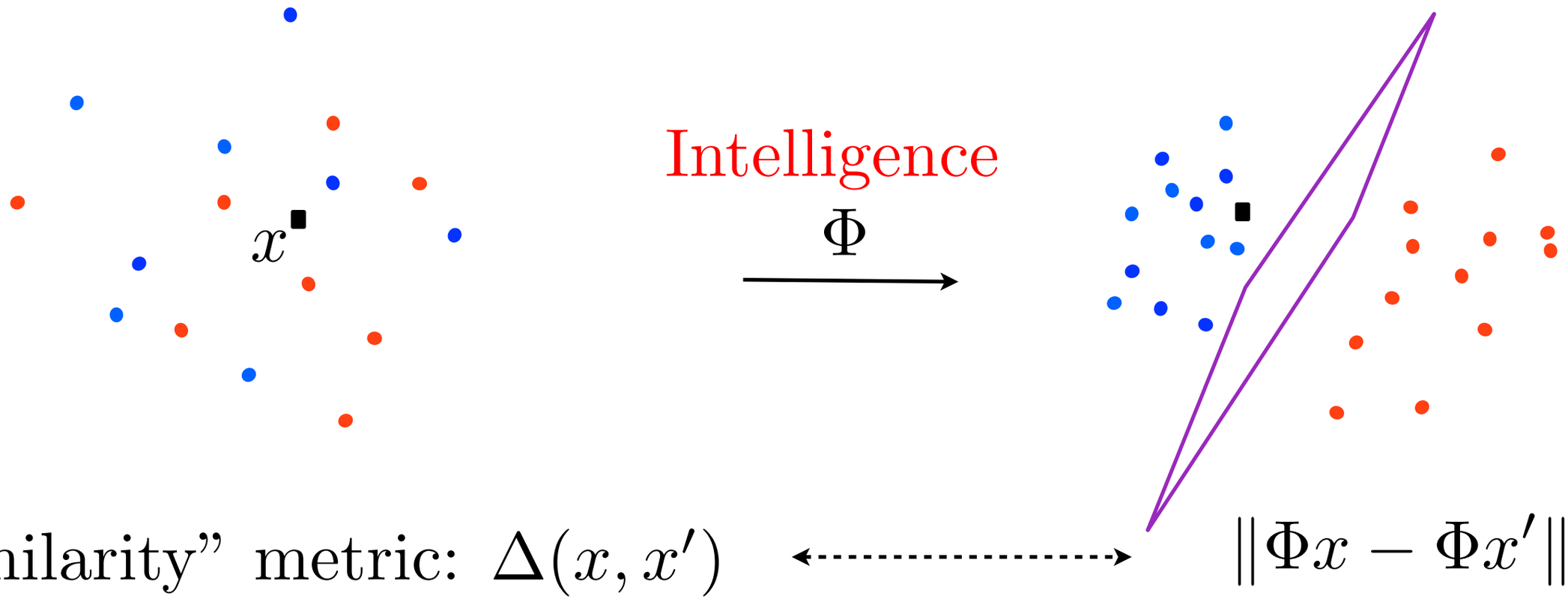
Data: $x \in \mathbb{R}^d$

$\|x - x'\|$: non-informative

Representation

$\Phi x \in \mathcal{H}$

Linear Classifier



Intelligence

Φ

"Similarity" metric: $\Delta(x, x')$ $\longleftrightarrow$ $\|\Phi x - \Phi x'\|$

Equivalent Euclidean metric:

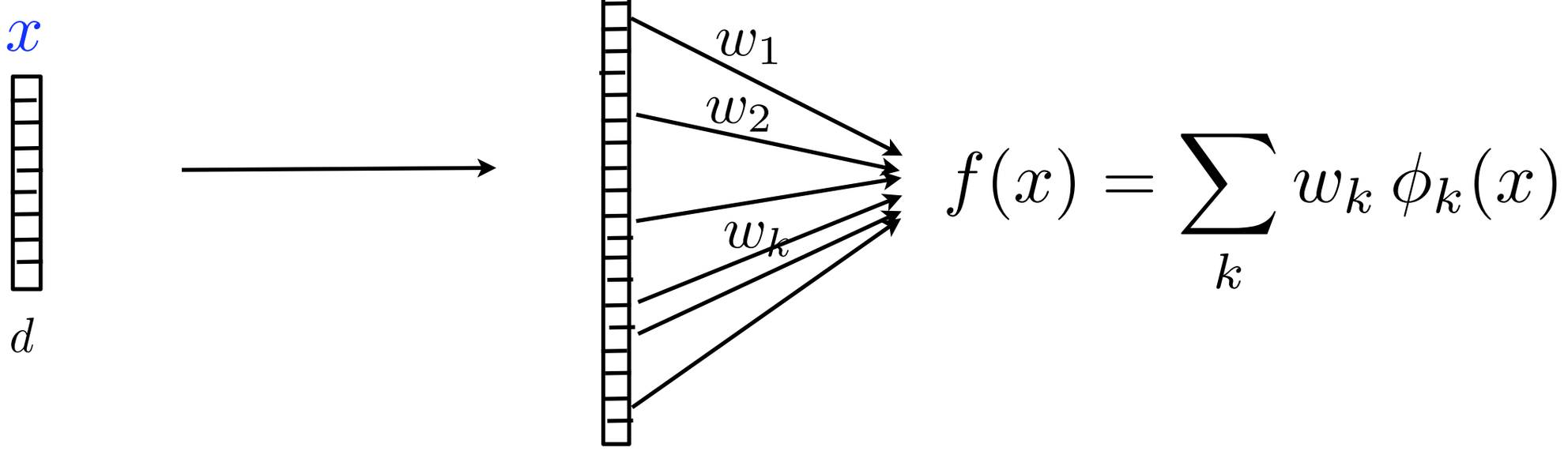
$$\Delta(x, x') \approx \|\Phi x - \Phi x'\|$$

Aggregation of Weak Features

Feature vector

$$\Phi(x) = \{\phi_n(x)\}_n$$

Input Data



$$f(x) = \sum_k w_k \, \phi_k(x)$$

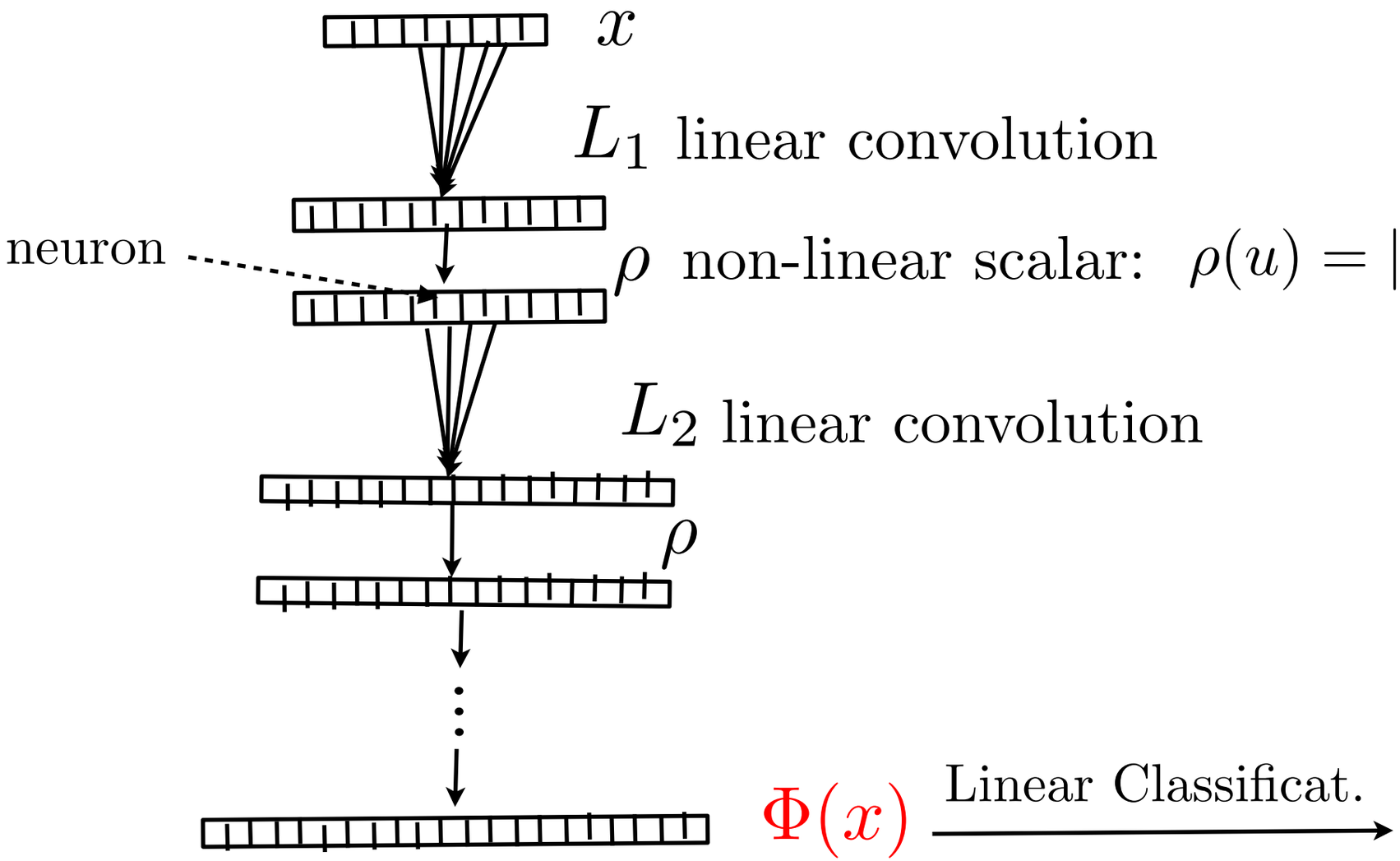
- Selection and vote of many weak features.

How to define Φ ?

Deep Convolution Neworks

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

Hierachical Invariants



Optimize the linear operators L_k: over 10^9 parameters

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

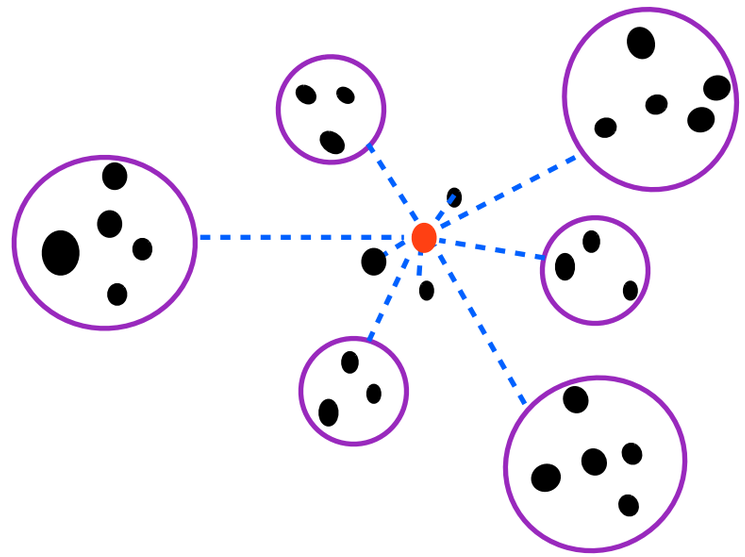
Products by FaceBook, IBM, Google, Microsoft, Yahoo...

Why does it work so well ?

Multiscale Interactions

- A system of d particles involves d^2 interactions

- Multiscale separation into $O(\log^2 d)$ interactions



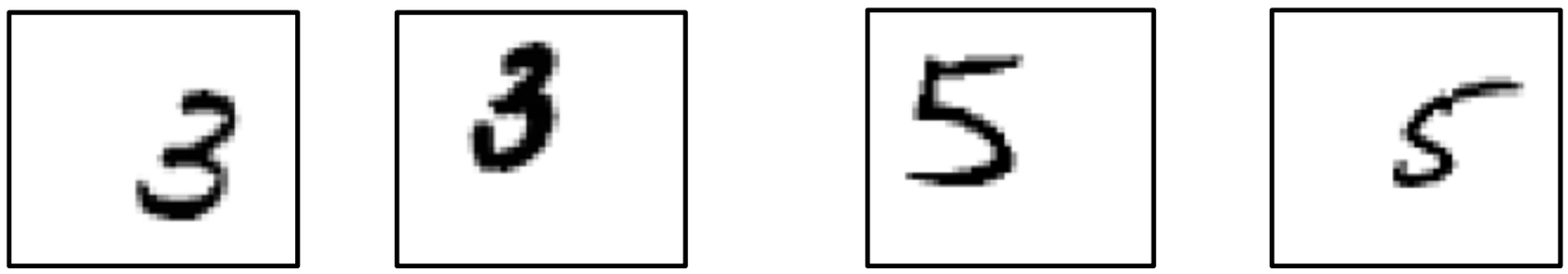
- Multiscale analysis: in mathematics and physics.

Learning Metric Transformations

- Patterns are too diverse to memorize them all
- We learn interactions and transformations (forces in physics)

Geometric shapes: deformation metric

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

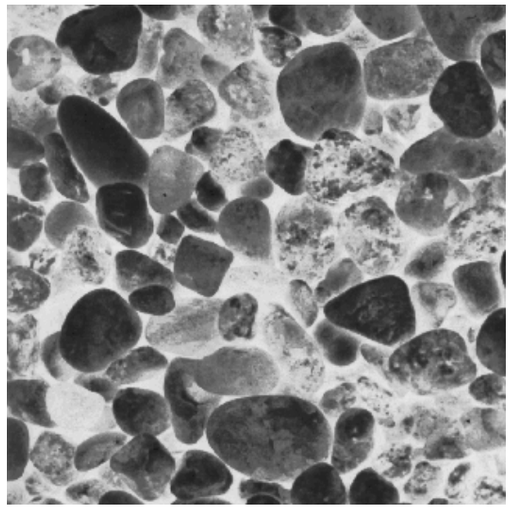


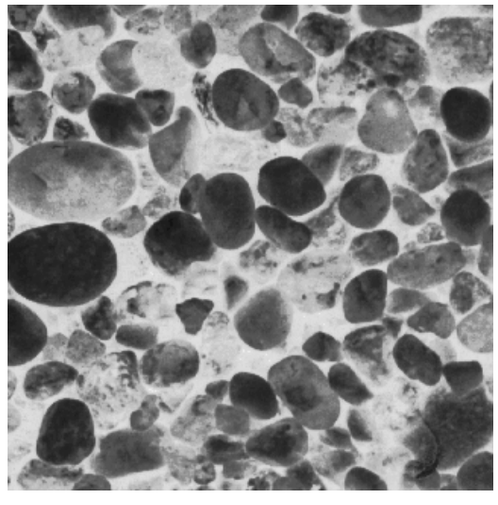
Learn groups of operators

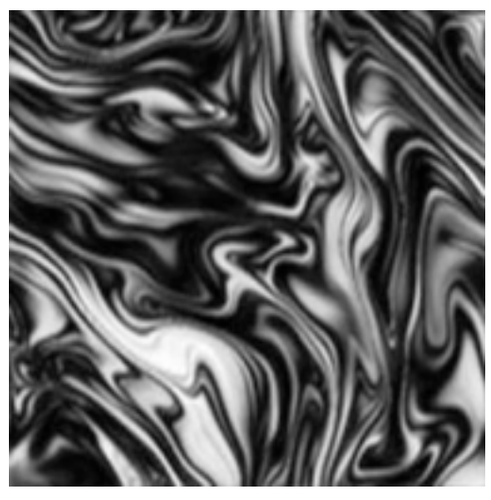
Learning Metric Transformations

- Patterns are too diverse to memorize them all
- We learn interactions and transformations (forces in physics)

Stationary Textures







2D Turbulence

Learn probabilistic metrics

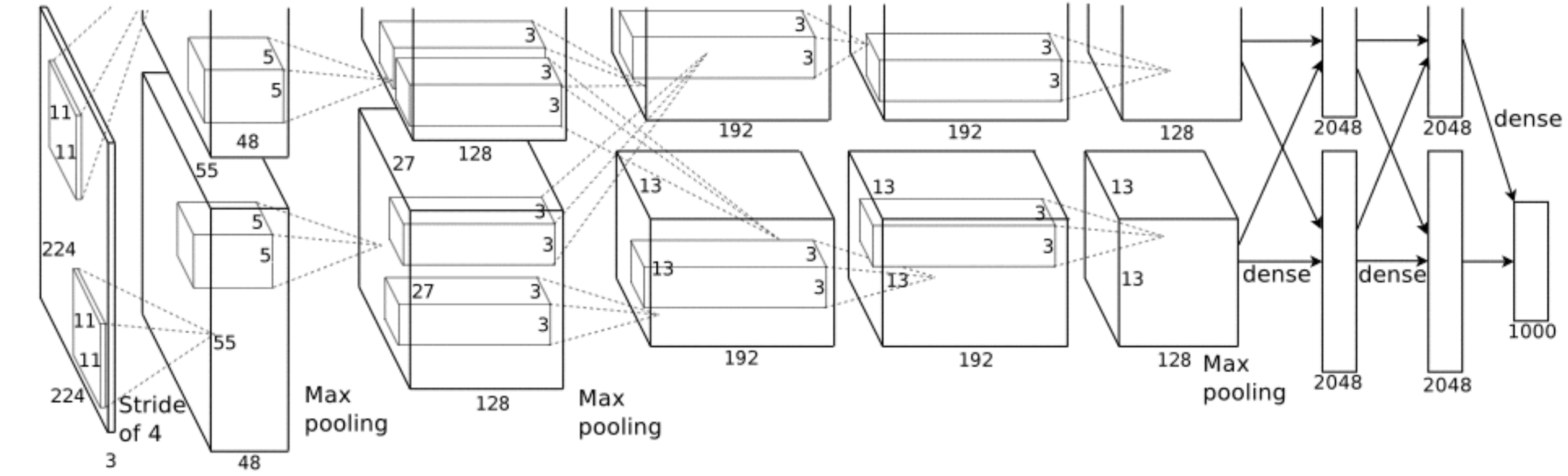
ImageNet Data Basis

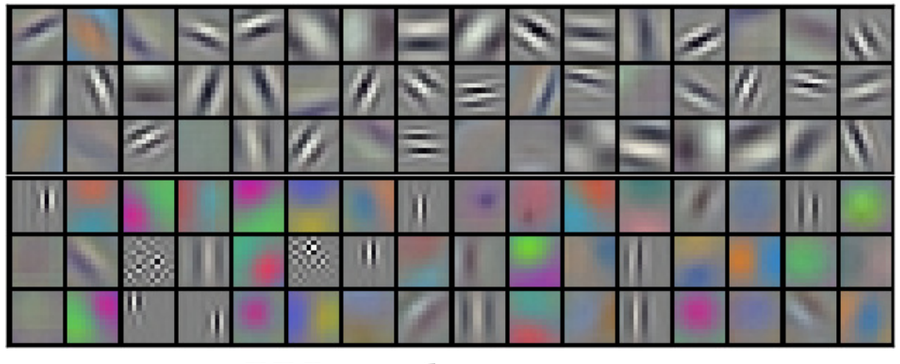
- Data basis with 1 million images and 2000 classes

Alex Deep Convolution Network

A. Krizhevsky, Sutsever, Hinton

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



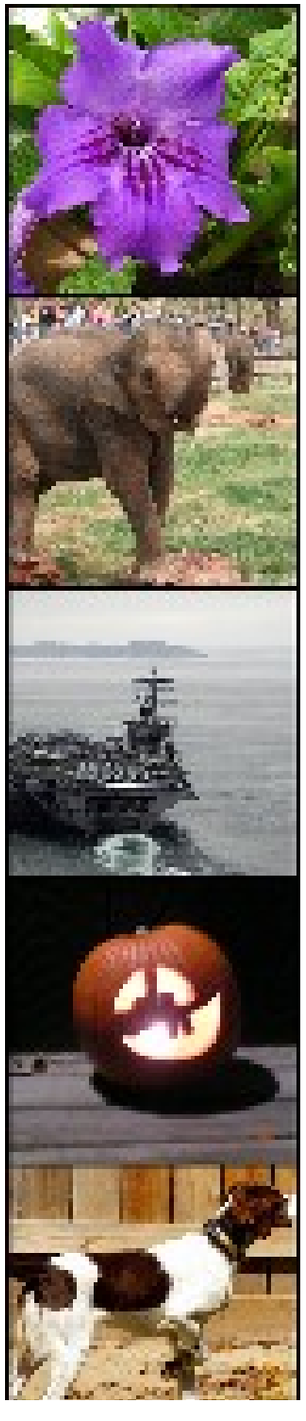


Wavelets

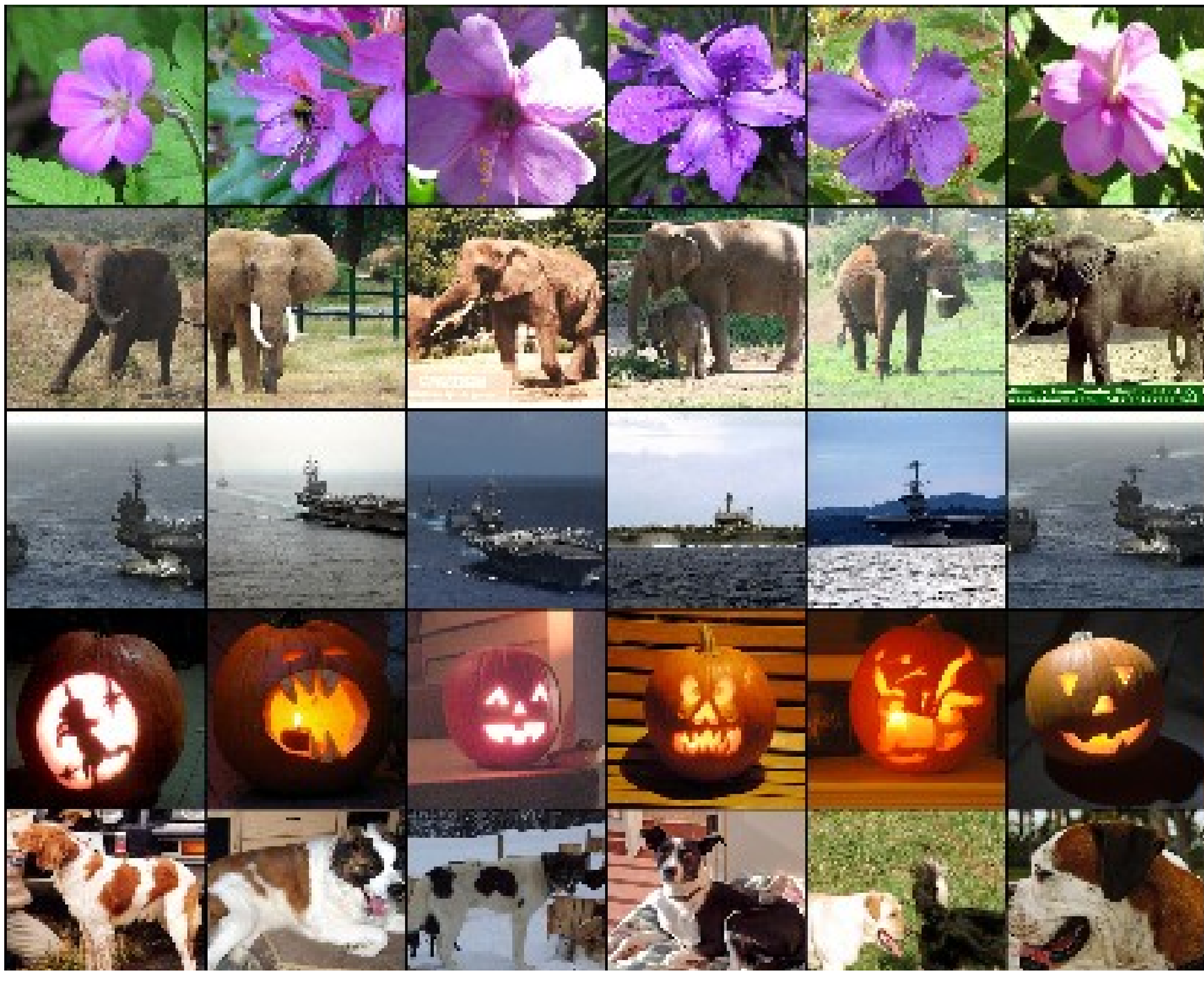
Image Classification

Image Retrieval

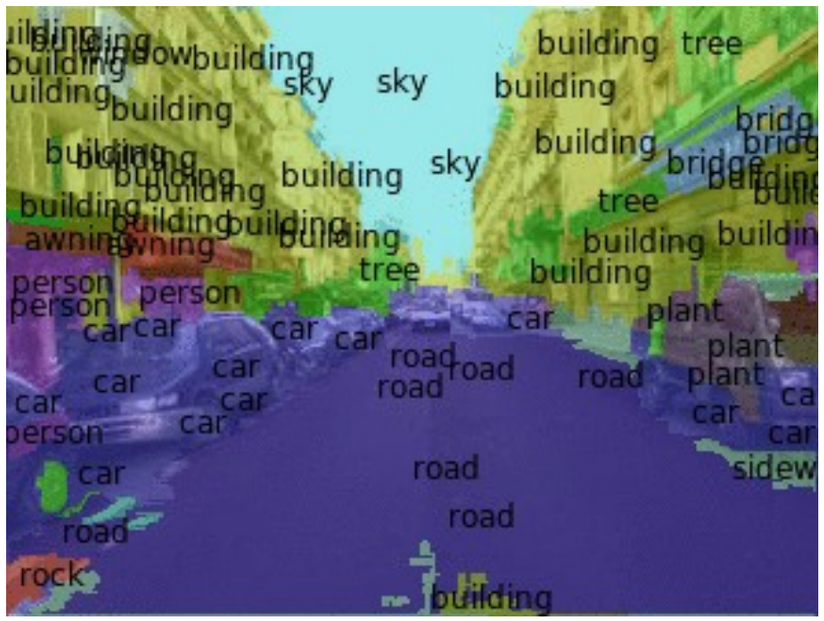
TEST IMAGE



RETRIEVED IMAGES



Scene Labeling

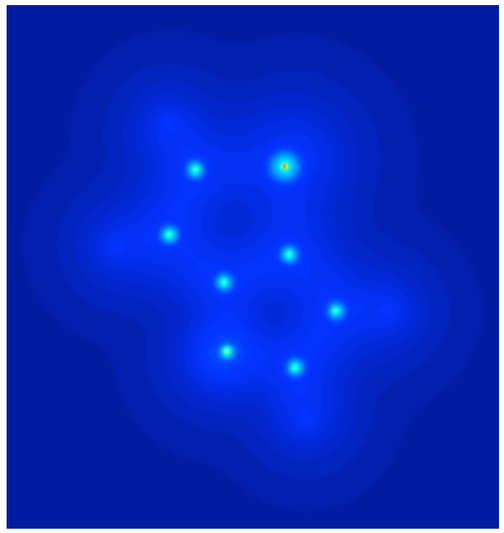
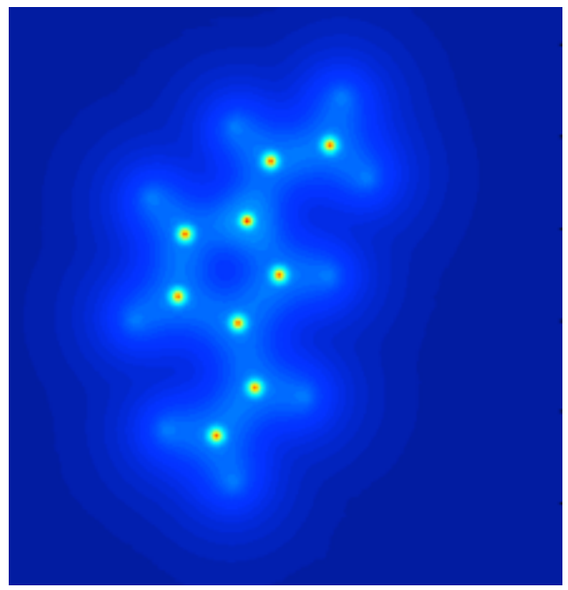
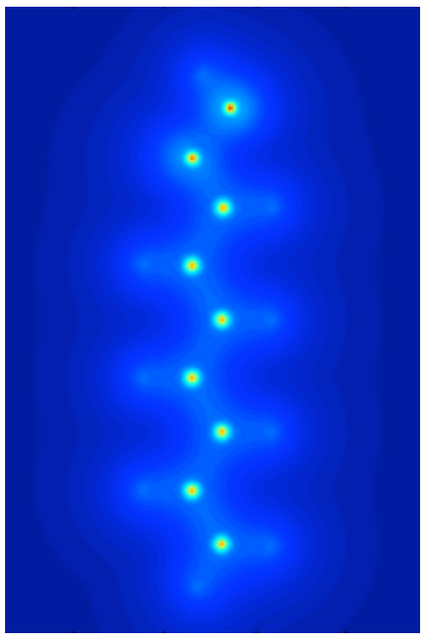
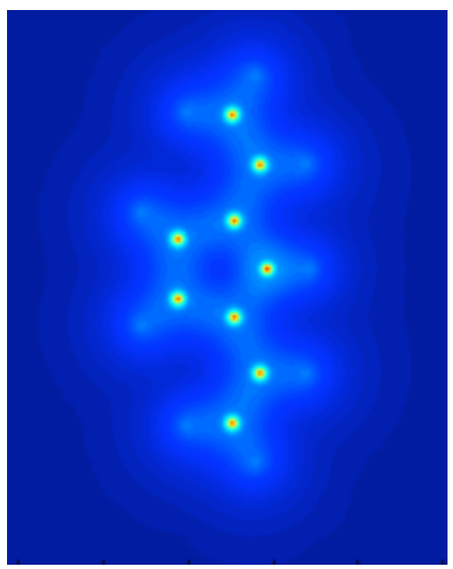
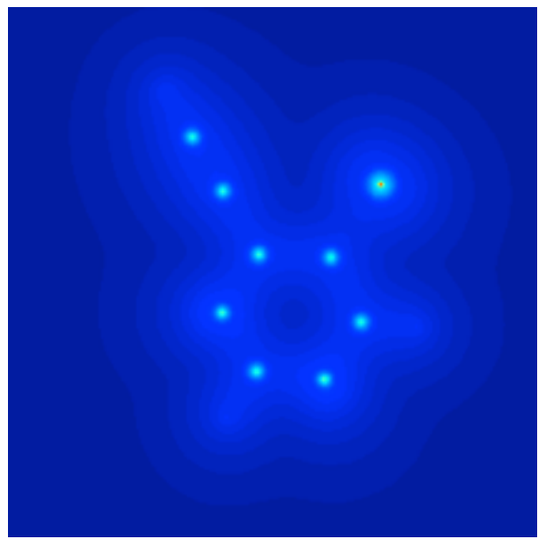
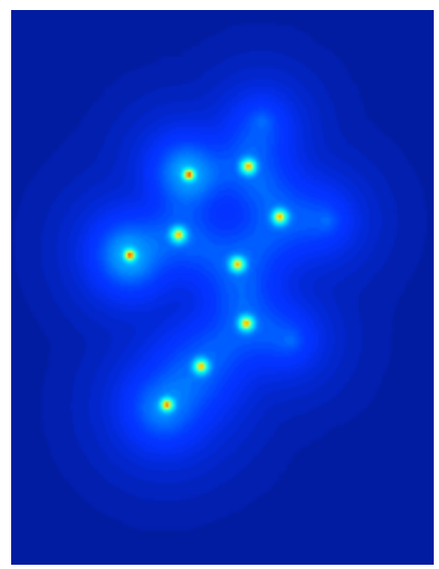


Quantum Chemistry

Regression of molecule energy from atomic positions without solving Schroedinger equation

Organic molecules with

Hydrogne, Carbon
Nitrogen, Oxygen
Sulfur, Chlorine

Paysage Scientifique

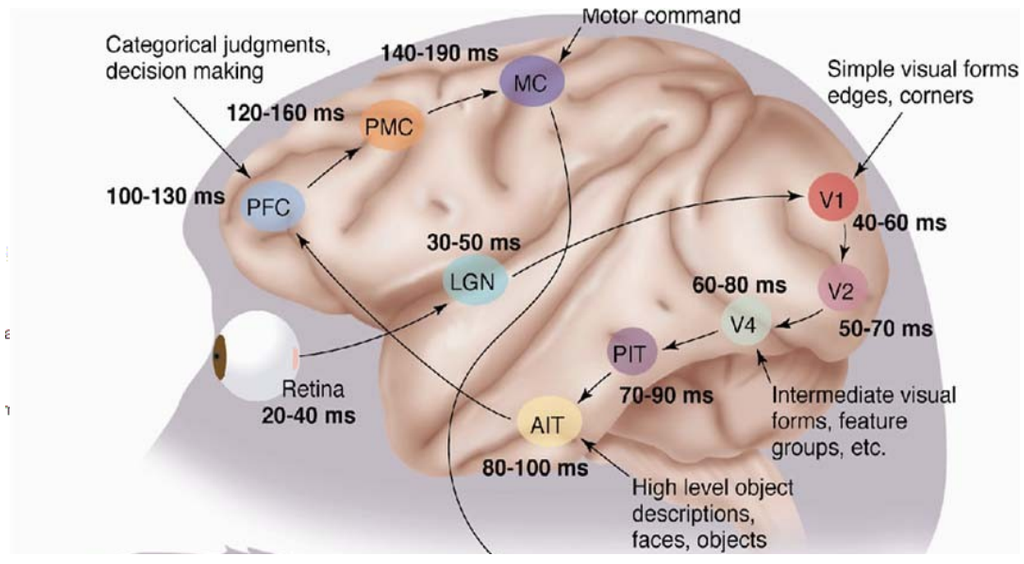
Informatique
Algorithmiciens

Physique
Chimie

Mathématiques

Applications
Médicales
Industrielles
Services...

Neurosciences



Sciences Sociales

Linguistique

Philosophie

- L'ensemble des sciences vont être profondemment impactés

Paysage Industriel

Utilisateurs:
Sociétés
Individus

Réseaux

Terminaux:
Ordinateurs
Téléphones
Objets connectés

Apple, Samsung
Huawei, Google

Internet
Google, FaceBook
Amazon, Microsoft

Applications/Services

Start-up, IBM,...

- A terme, l'intelligence ira dans les terminaux
- Pas d'acteur Français important: prise de conscience lente