A Mathematical Quest for Information Processing

Stéphane Mallat

Collège de France

COLLÈGE DE FRANCE 1530

A Century of Information

What is information in mathematics ?

1921
Fisher Information
Math. Statistics

1948
Shannon Information Theory
Proba. Concentration

How to represent data and analyse information ?

1940's
Wiener Gaussian Models
Fourier Analysis

1990's
Wavelet Sparse Models
Functional Analysis

2021
Neural networks
High-dim. Geometry
Maths not understood

Math Foundations of Statistics

1921 report of Ronald Fisher to the Royal Academy

"The object of statistical methods is the reduction of data"

"Represent the whole of the relevant data information"

- Model data $\{x_t\}_{t \leq n}$ as independant samples of a distribution $p_\theta(x_t)$ parameterised by θ.

  Gaussian example: $p_\theta(x) \sim e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ with $\theta = (\mu, \sigma)$.

- Consistent estimator $\hat{\theta} \to \theta$ as n tends to ∞

- Maximum likelihood estimator $\hat{\theta}$ of θ given $\{x_t\}_{t \leq n}$

$$\hat{\theta} \text{ maximises } p_\theta(x_1, ..., x_n) = \prod_t p_\theta(x_t)$$

Fisher Information

- Amount of information carried by data with probability p_θ on the unknown parameter θ: curvature of $\log p_\theta$

$$I(\theta) = \mathbb{E}\Big[\Big(\frac{\partial \log p_\theta(x)}{\partial \theta}\Big)^2\Big]$$

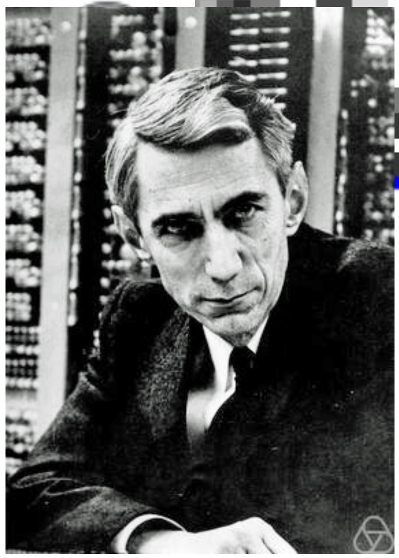
- **Cramer-Rao Bound** on parameter estimation (1940's):

Theorem If $\mathbb{E}(\hat\theta) = \theta$ then

$$\mathbb{E}(\hat\theta - \theta)^2 \geq \frac{1}{I(\theta)}$$

The Fisher information controls the uncertainty to estimate θ

What family of parametrised probabilities $\{p_\theta\}_\theta$?



Shannon Information Theory

Concentration in high dimension

n independent random variables $X = (X_1, ..., X_n)$

with same probability distribution $p(X) = \prod_t p(X_t)$

Law of large numbers: *Entropy*

$$-\frac{1}{n}\log p(X) = -\frac{1}{n}\sum_{t=1}^{n}\log p(X_t) \xrightarrow[n \to \infty]{} H = \mathbb{E}\Big(-\log p(X_t)\Big)$$

Asymptotic Equipartion Theorem

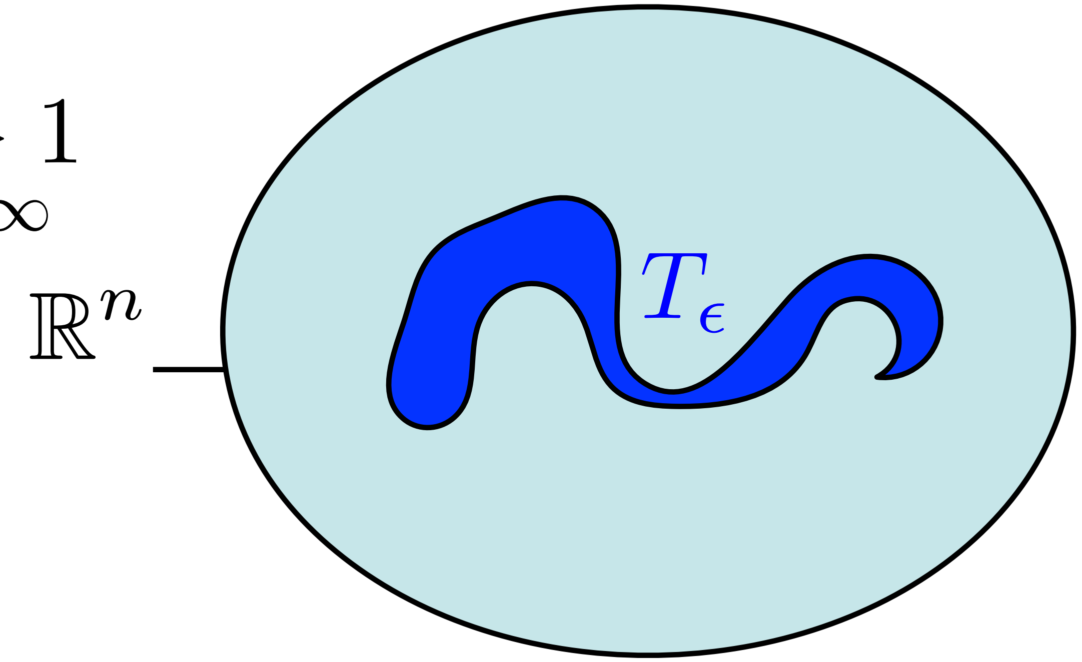
For an ergodic stationary process $\{X_t\}_t$

$$-\frac{1}{n}\log p(X_1, ..., X_n) \xrightarrow[n \to \infty]{} H \quad \text{with probability 1}$$

Typical Sets

Typical set: $T_\epsilon = \left\{ x \in \mathbb{R}^n \ : \ |n^{-1} \log p(x) - H| \le \epsilon \right\}$

Concentration: $\text{Prob}(X \in T_\epsilon) \underset{n \to \infty}{\longrightarrow} 1$



If $x \in T_\epsilon$ then $p(x) \sim 2^{-nH}$

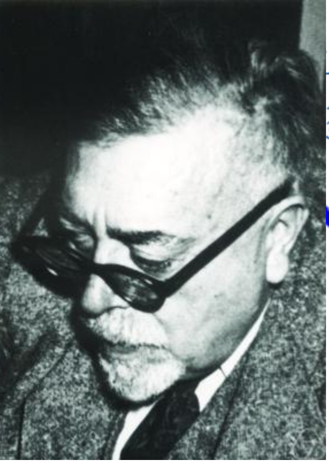
If X is quantised then $|T_\epsilon| \sim 2^{nH}$

nH is the minimum average number of bits to code X

Considerable impact:

- Coding: telecommunication and data storage
- Statistical physics (thermodynamic entropy)
- Large Deviation Theory (*Donsker-Varadhan 1960's*)

How to specify the geometry of Typical sets ?



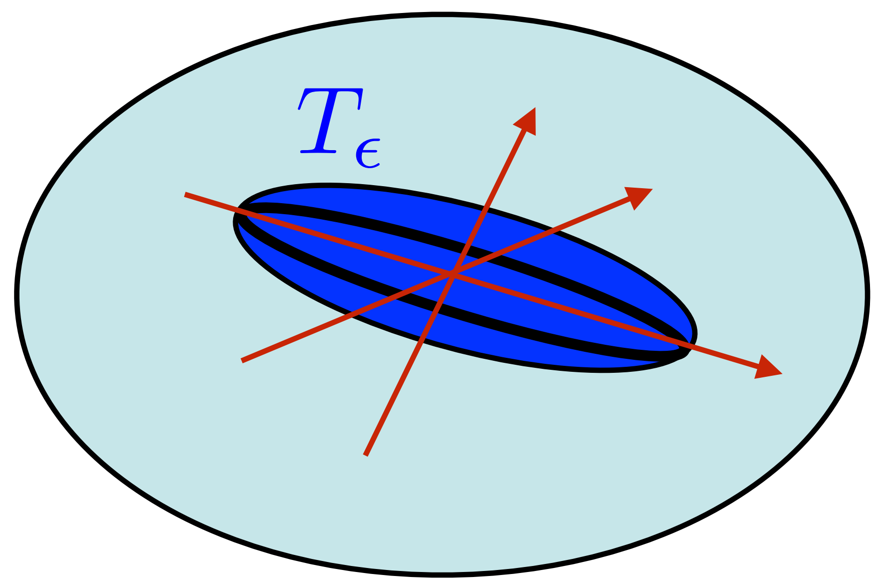
Wiener

Gaussian Stationary Processes

$$p_\Theta(x) = Z^{-1} \exp\Big(-\tfrac{1}{2}\langle x, \Theta x\rangle\Big)$$

where Θ is a positive matrix of parameters

Typical sets T_ϵ are ellipsoids whose principal axes are vectors vectors of an orthonormal basis $\mathcal{B}$ which diagonalises Θ.



- If $X_1 ..., X_t, ...$ is stationary, i.e. $p(x)$ is invariant to time-shift then $\mathcal{B}$ is a Fourier basis: $X_t = \sum_\omega \tilde{X}_\omega \, e^{it\omega}$

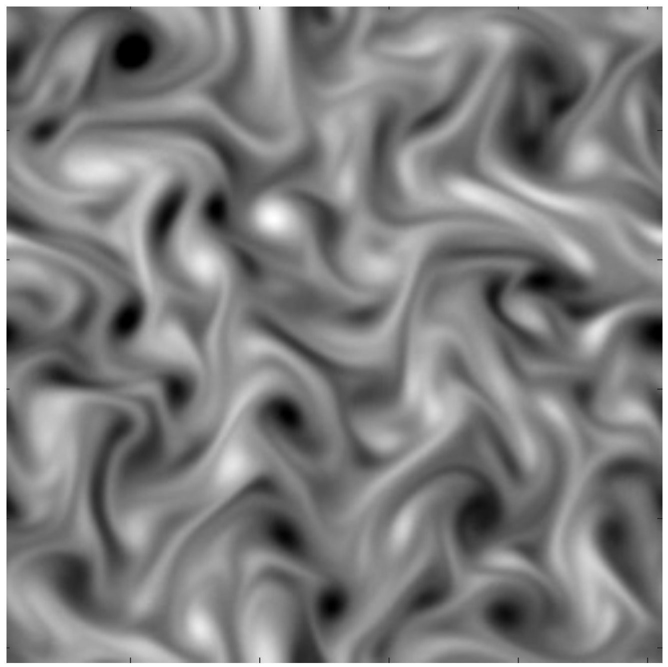
  Limit of continuous time: spectral representation
  Typical sets: balls of weighted Fourier spaces (Sobolev).

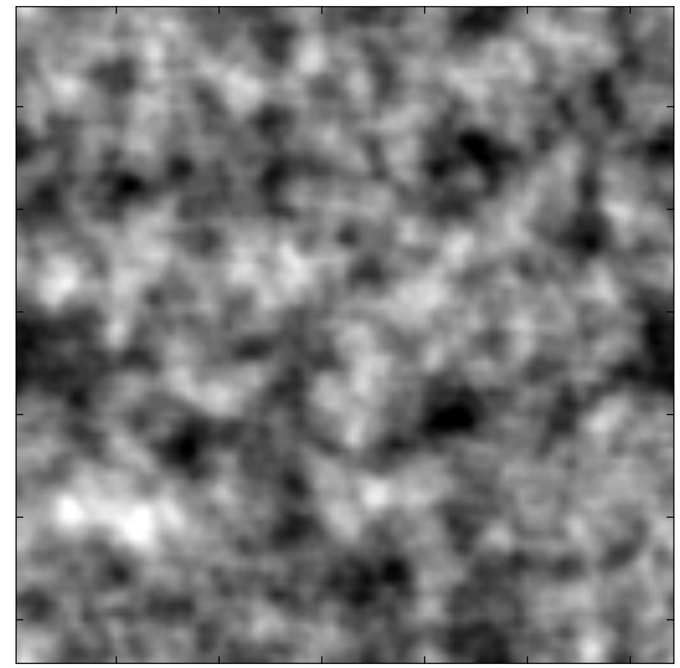
Non Gaussian Ergodic Processes

- Non Gaussianity: transients, intermittency, crises, edges...

Original | Gaussian model (*Kolmogorov*) (1941)

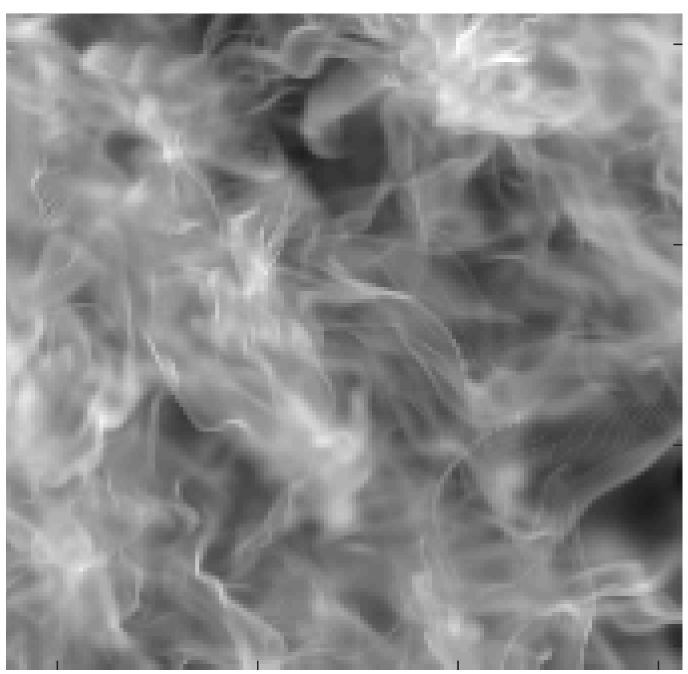
Fluid Turbulence

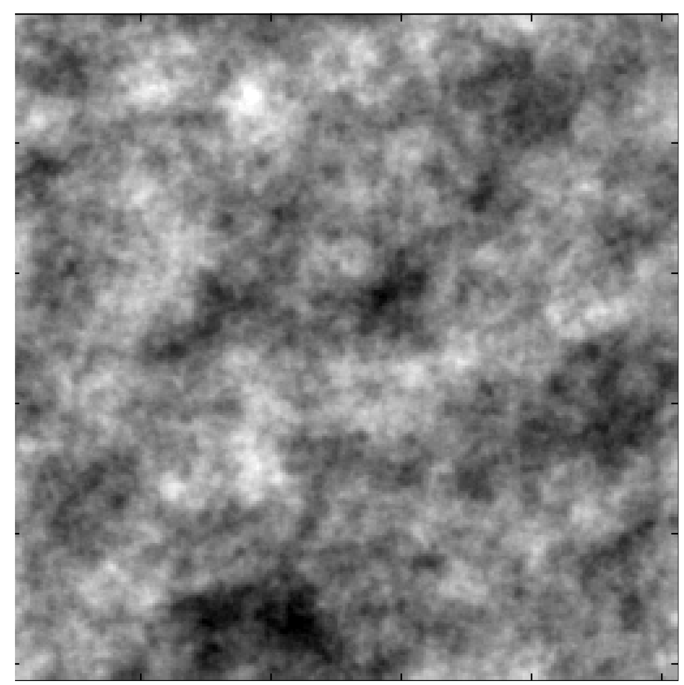




Same power spectrum

Cosmologic Turbulence



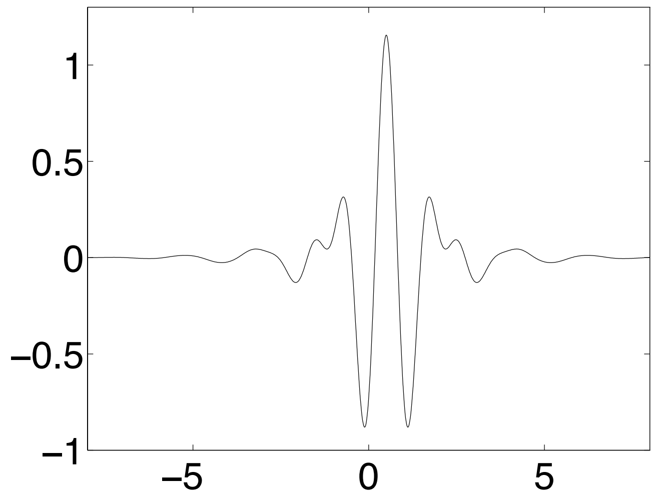


How to represent X and identify typical sets ?

Sparse Wavelet Representations

- Represent transient phenomena with localised wavelets.
- Sparse representations in wavelet bases (1980-90's):

Meyer wavelet $\psi(t)$



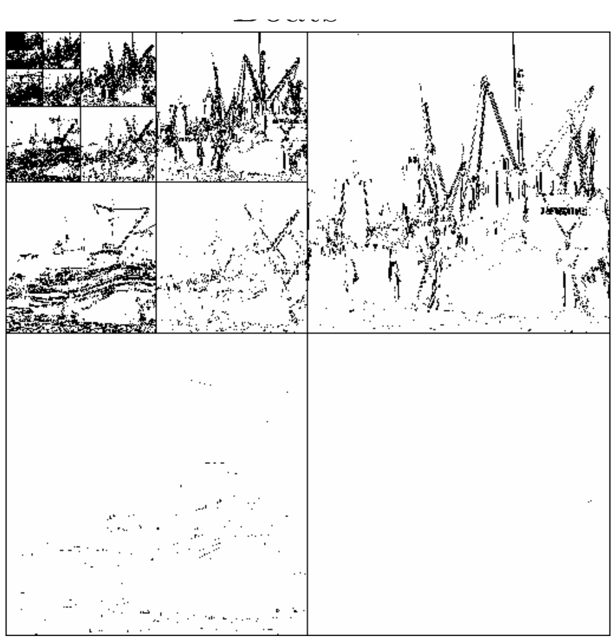
Orthonormal basis of $\mathbf{L^2}(\mathbb{R})$

$$\left\{\psi_{j,n}(t) = 2^{-j/2}\psi(2^j t - n)\right\}_{j,n}$$

$$X = \sum_{j,n} \langle X, \psi_{j,n}\rangle\, \psi_{j,n}$$

X

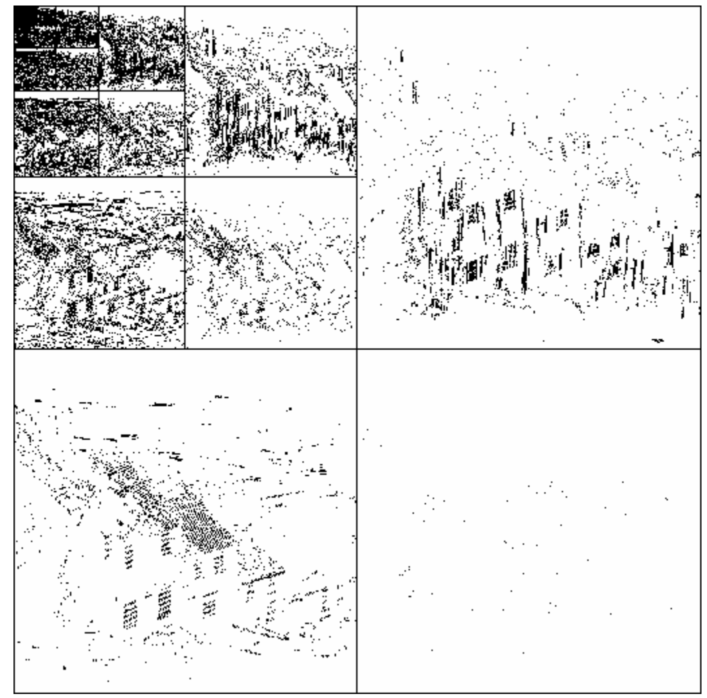
Sparse wavelet coefficients



Compression and Typical Sets

Original

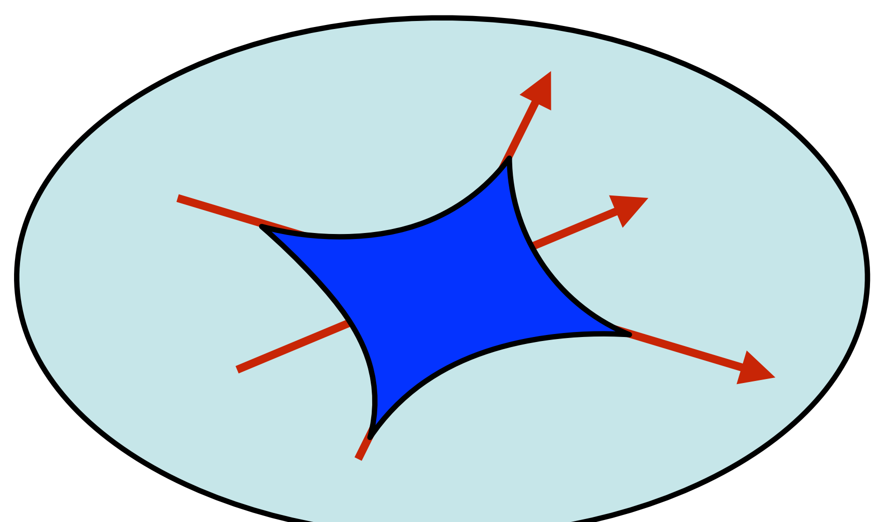
Sparse
Wavelet coefficients



JPEG-2000
Compressed by 40

Typical sets T_ϵ : bounded weighted ℓ^p norms of wavelet coef.

balls of Besov spaces



Still too crude to model geometric image structures: what else ?

New Frontier: Neural Networks

McCulloch and Pitts (1943)

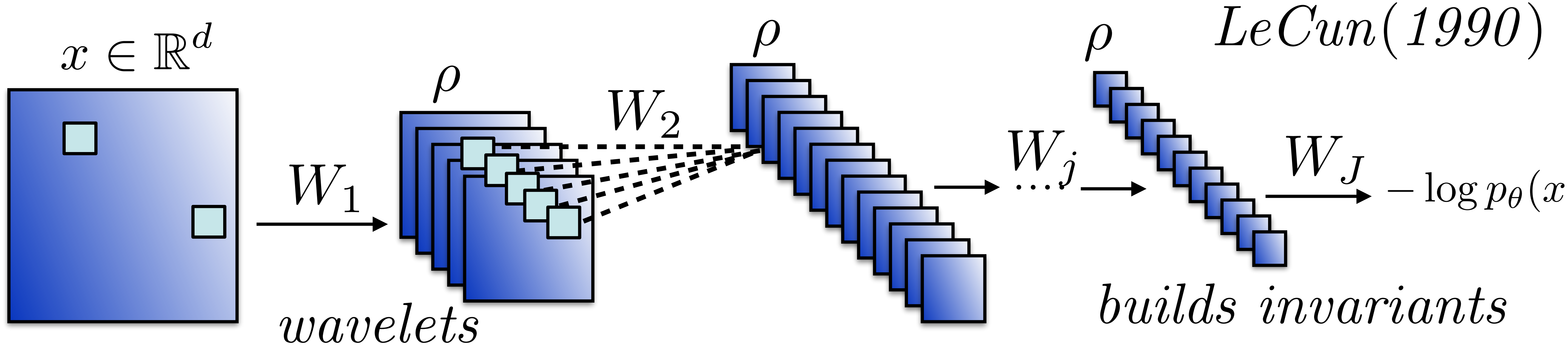
- Alternate linear operators and a pointwise non-linearity:

$$-\log p_\theta(x) = W_J \, \rho \, W_{J-1} \ldots \rho W_2 \, \rho \, W_1 x$$

with a rectifier $\rho(\alpha) = \max(\alpha, 0)$
and $\theta = (W_j)_{1 \leq j \leq J}$ are matrices optimised by
maximising the data *likelihood* with a gradient descent.

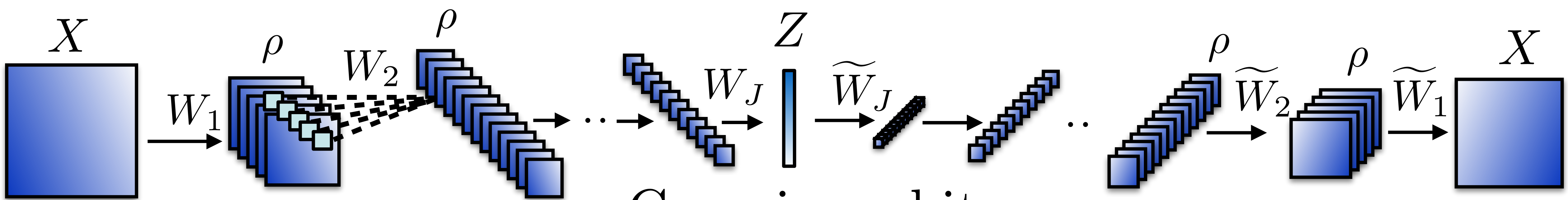
- Convolutional architectures: shift-invariant operators W_j

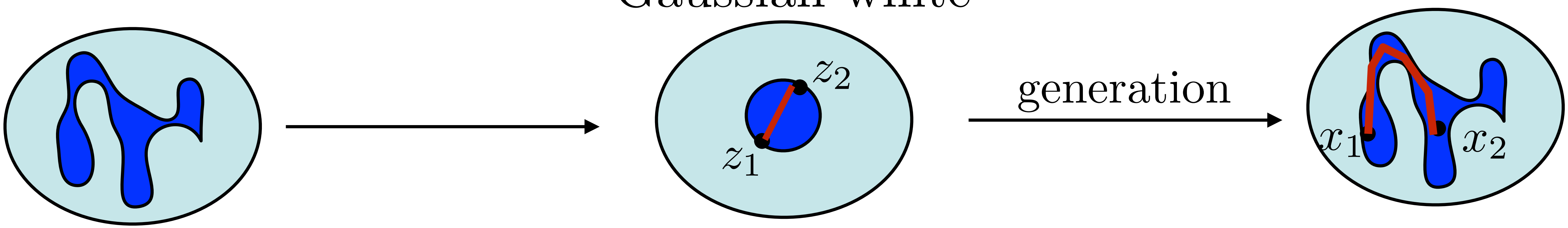
LeCun(1990)



Models and Generation of Processes

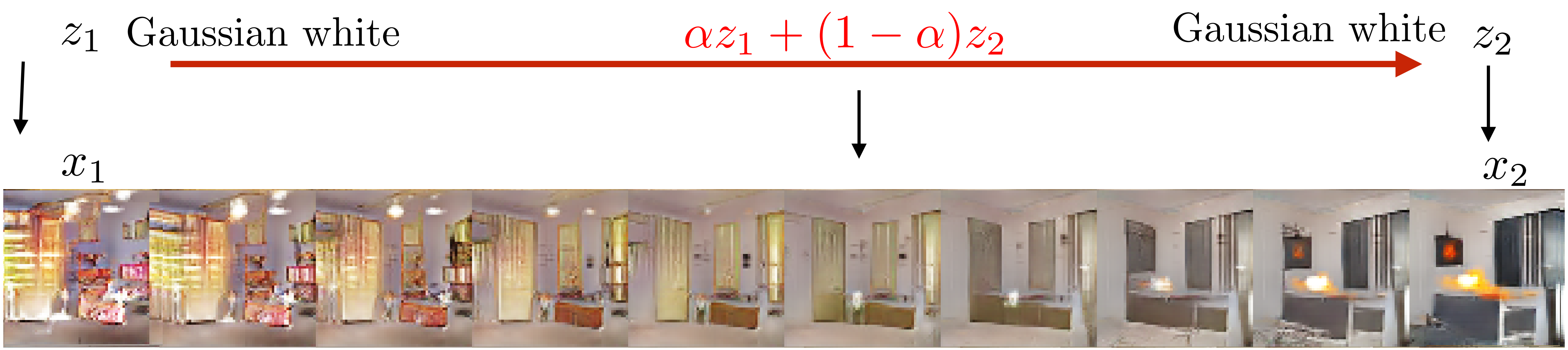
mapping into a Gaussian and inversion





Can generate complex ergodic processes including turbulences

Beyond ergodicity: generation from images of bedrooms



What class of random processes / transport / function spaces ?

High Resolution Generation

T. Karras, T. Aila, S. Laine, J. Lehtinen

Generated from Hollywood celebrities data basis

Outstanding Mathematical Questions

- Information processing is about high-dimensional geometry.

- Neural networks have spectacular ability to process information, but mathematically not understood.

- Major societal issue because of critical AI applications: *medical, transport, decision making...*

- Outstanding questions, from *statistics* to:
  - Probability and concentration
  - Functional and harmonic analysis
  - Geometry and group theory
  - Optimisation and high-dimensional transport