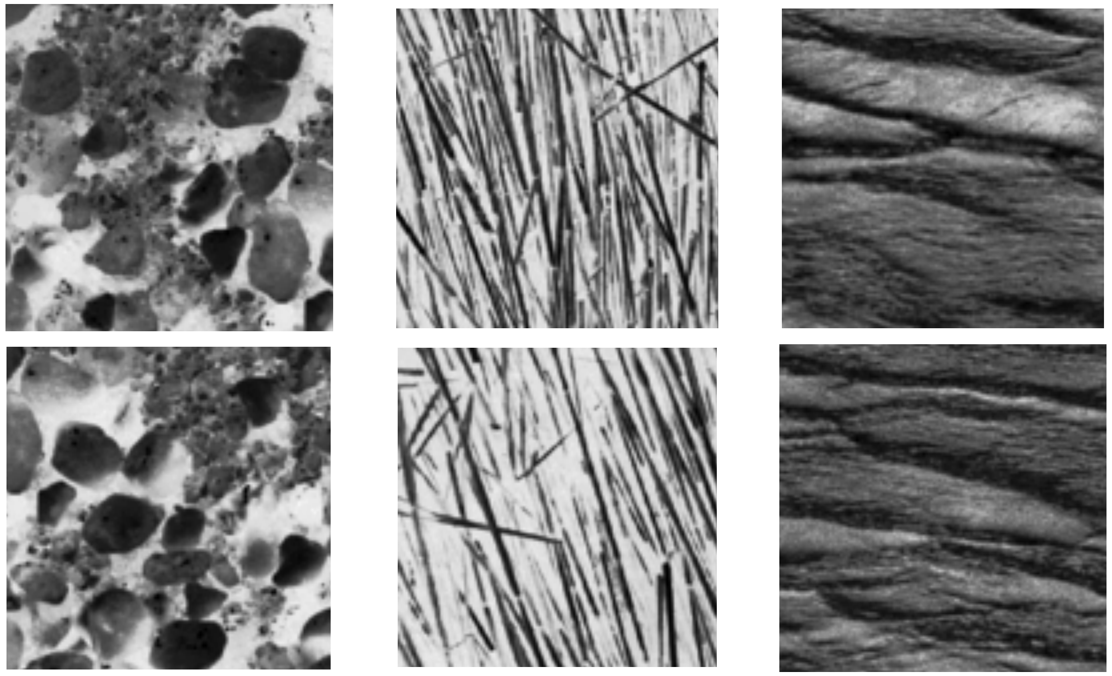
Stat 212b: Topics in Deep Learning
Lecture 12

Joan Bruna
UC Berkeley

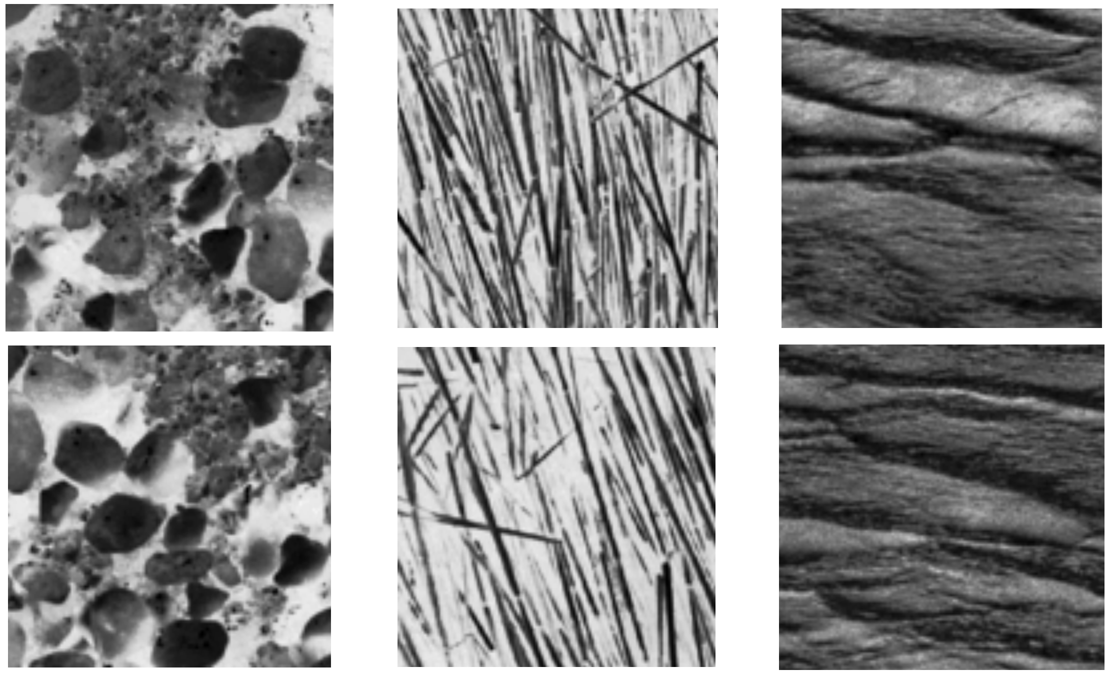
Representation of Stationary Processes

$x(u)$: realizations of a stationary process $X(u)$ (not Gaussian)



Representation of Stationary Processes

$x(u)$: realizations of a stationary process $X(u)$ (not Gaussian)



$$\Phi(X) = \{E(f_i(X))\}_i$$

Estimation from samples $x(n)$: $\widehat{\Phi}(X) = \left\{ \frac{1}{N} \sum_n f_i(x)(n) \right\}_i$

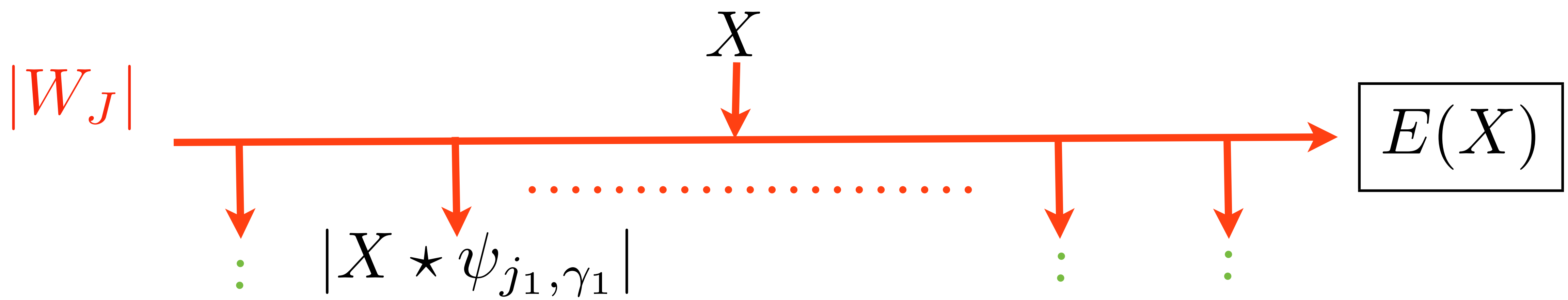
Discriminability: need to capture high-order moments

Stability: $E(\|\widehat{\Phi}(X) - \Phi(X)\|^2)$ small

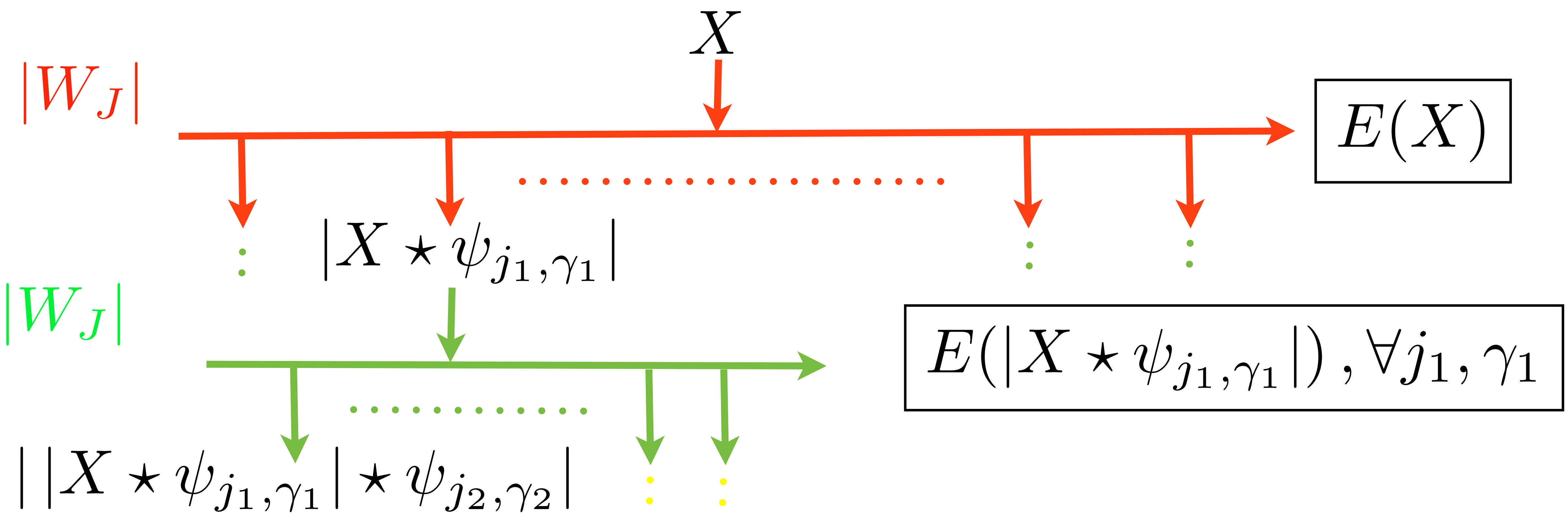
Scattering Moments

$$X$$

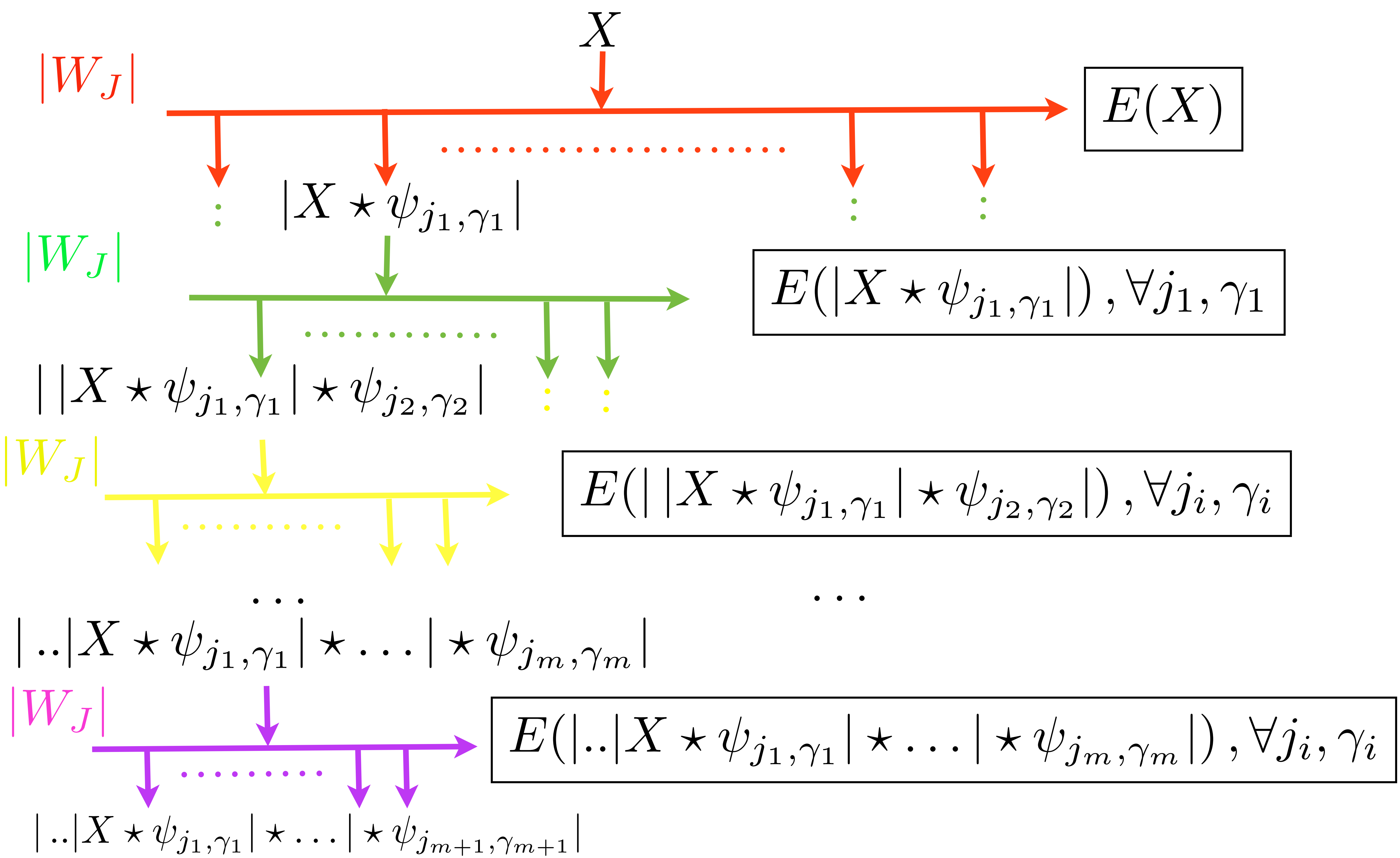
Scattering Moments



Scattering Moments



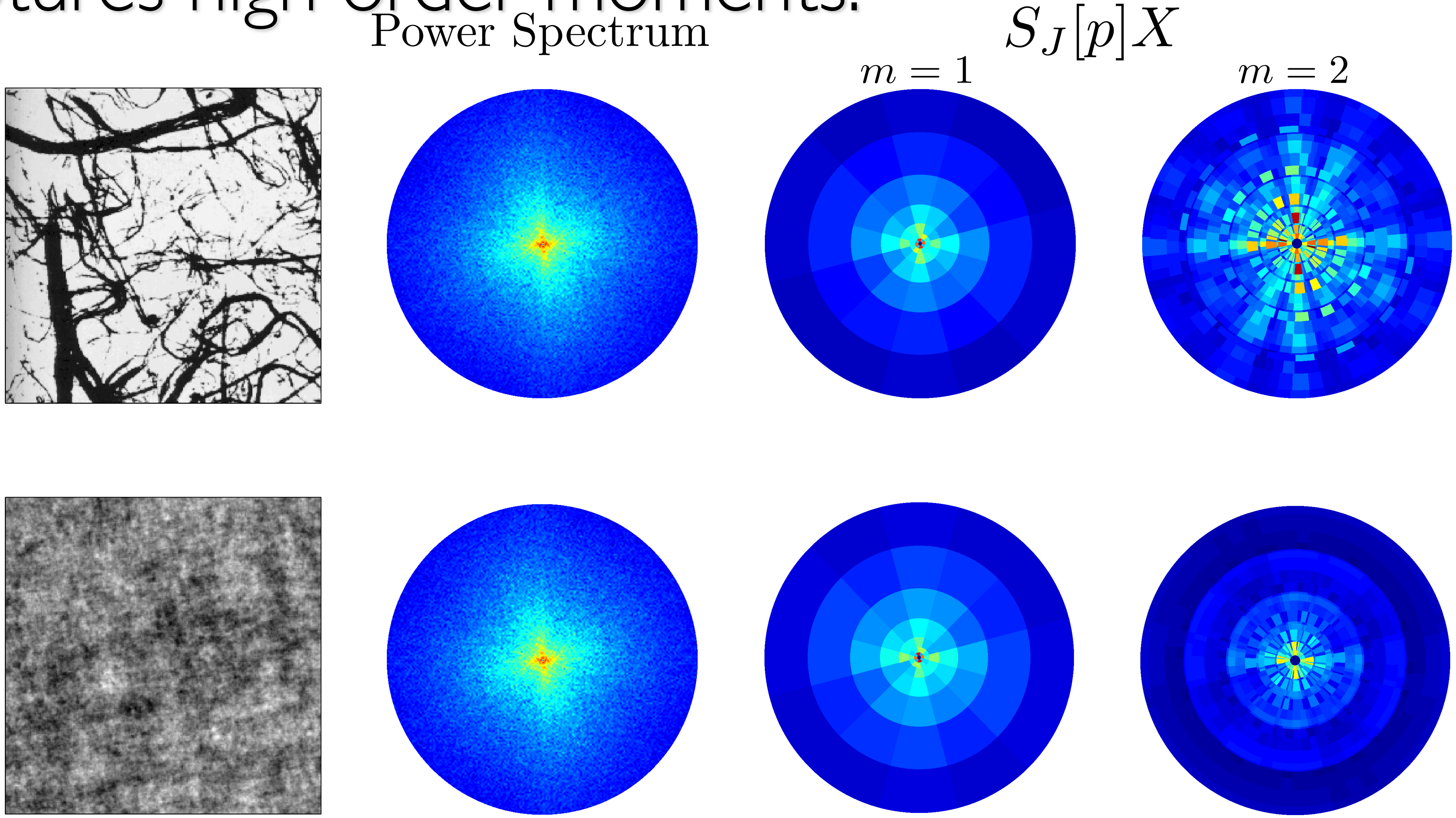
Scattering Moments



Properties of Scattering Moments

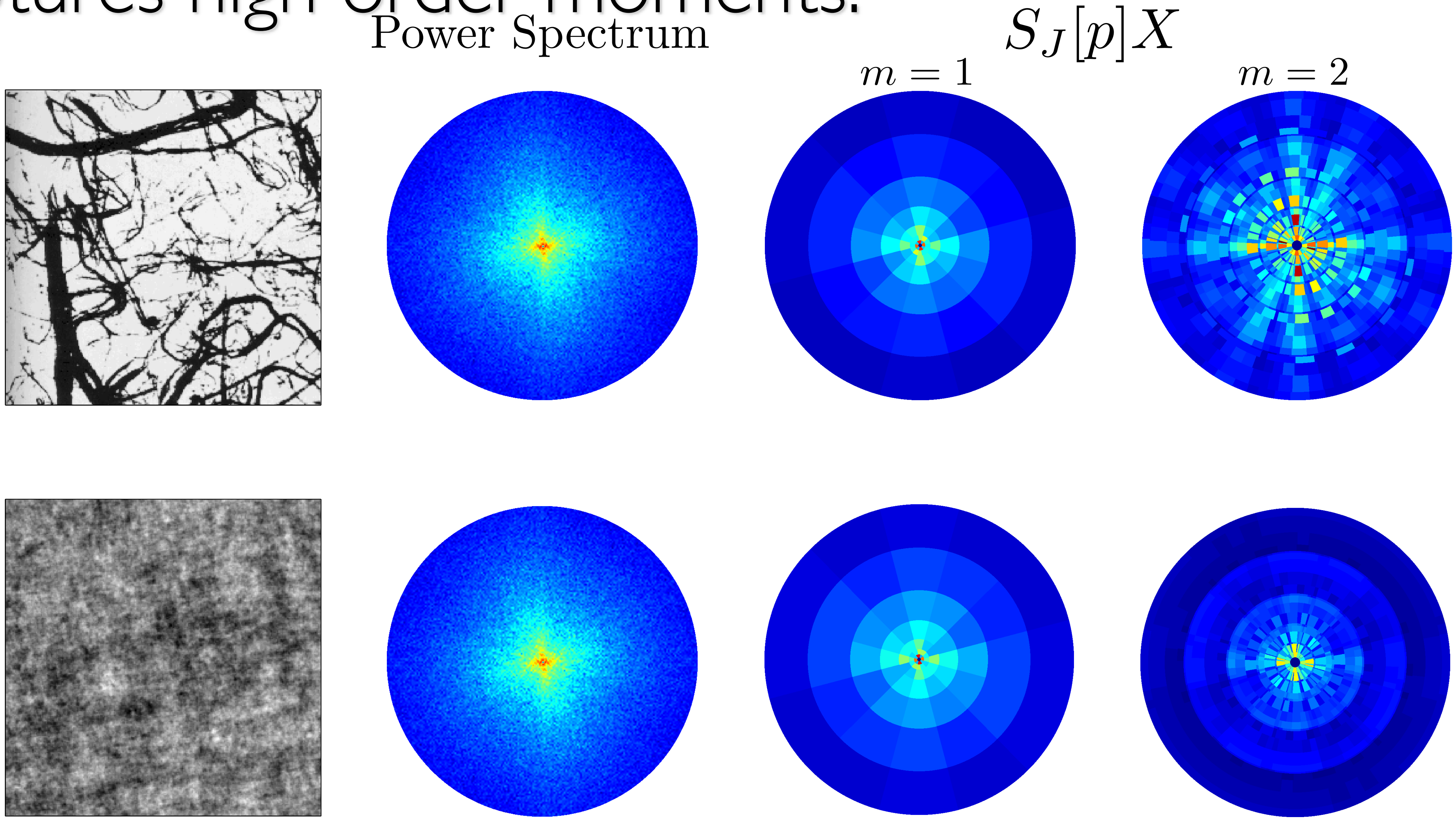
- Captures high order moments:

[Bruna, Mallat, '11,'12]



Properties of Scattering Moments

- Captures high order moments: [Bruna, Mallat, '11,'12]



- Cascading non-linearities is ***necessary*** to reveal higher-order moments.

Consistency of Scattering Moments

Theorem: [B'15] If ψ is a wavelet such that $\|\psi\|_1 \leq 1$, and $X(t)$ is a linear, stationary process with finite energy, then

$$\lim_{N \to \infty} E(\|\hat{S}_N X - SX\|^2) = 0 \ .$$

Consistency of Scattering Moments

Theorem: [B'15] If ψ is a wavelet such that $\|\psi\|_1 \leq 1$, and $X(t)$ is a linear, stationary process with finite energy, then

$$\lim_{N \to \infty} E(\|\hat{S}_N X - SX\|^2) = 0 \ .$$

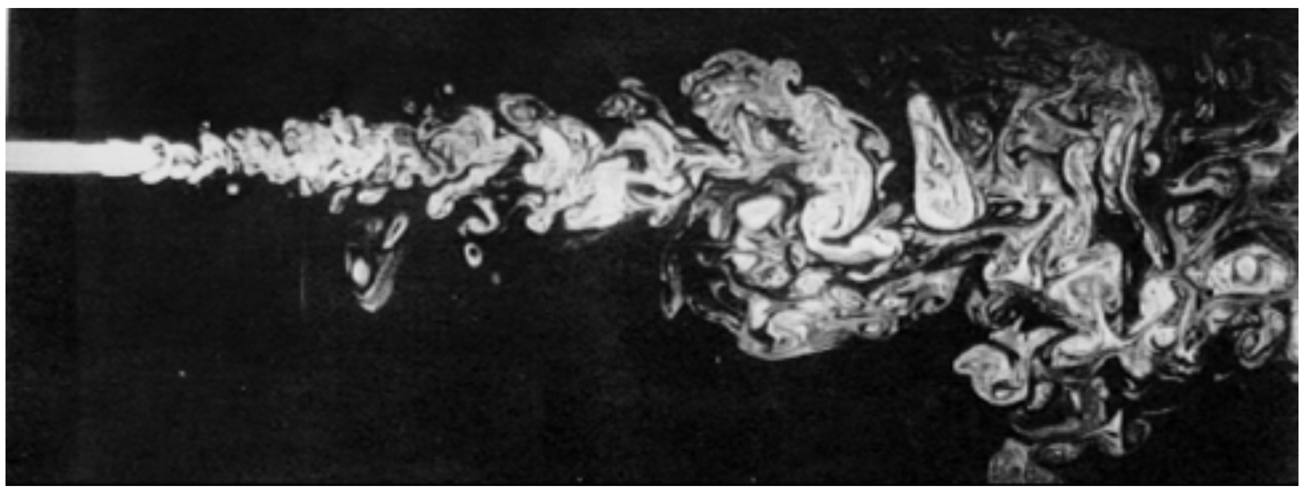
Corollary: If moreover $X(t)$ is bounded, then

$$E(\|\hat{S}_N X - SX\|^2) \leq C \frac{|X|_\infty^2}{\sqrt{N}} \ .$$

- Although we extract a growing number of features, their global variance goes to 0.
- No variance blow-up due to high order moments.
- Adding layers is critical (here depth is log(N)).

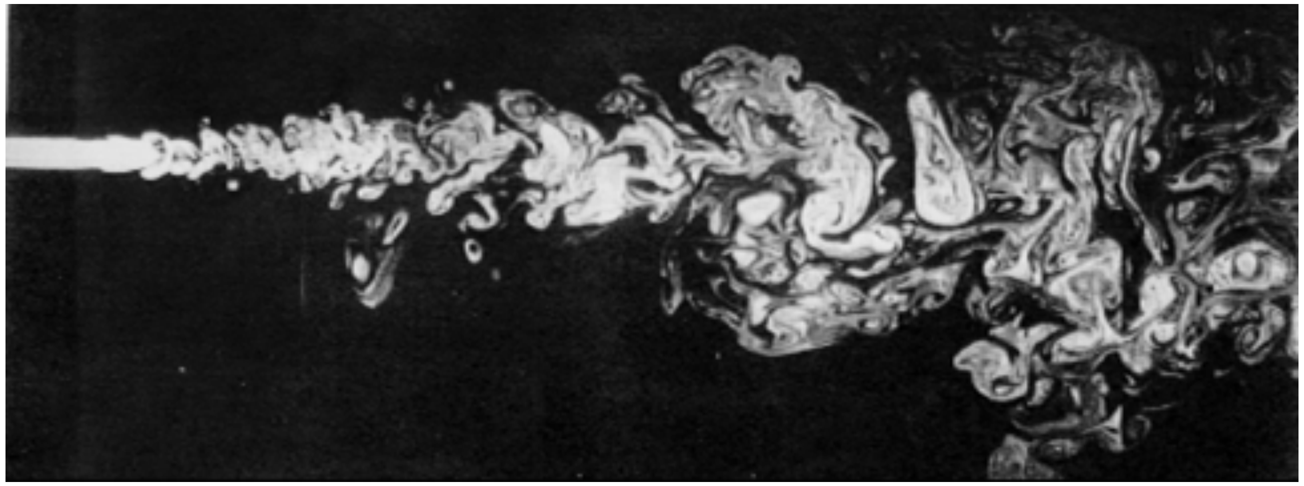
Fractal Processes

- *Motivation*: Find statistical models for chaotic phenomena such as Turbulent flows.



Fractal Processes

- *Motivation*: Find statistical models for chaotic phenomena such as Turbulent flows.



- Kolmogorov "5/3" theory (1941): isotropic energy dissipation induces a power spectrum of the form

$$f_F(\omega) \propto |\omega|^{-5/3} \ .$$

Fractal Processes

- Kolmogorov "5/3" theory (1941): isotropic energy dissipation induces a power spectrum of the form

$$f_F(\omega) \propto |\omega|^{-5/3} \ .$$

Fractal Processes

- Kolmogorov "5/3" theory (1941): isotropic energy dissipation induces a power spectrum of the form

$$f_F(\omega) \propto |\omega|^{-5/3} \ .$$

- This model implies *scale self-similarity:*

$$\{X(st)\} = W_s\{X(t)\}$$

Fractal Processes

- Kolmogorov "5/3" theory (1941): isotropic energy dissipation induces a power spectrum of the form

$$f_F(\omega) \propto |\omega|^{-5/3} \ .$$

- This model implies *scale self-similarity:*

$$\{X(st)\} = W_s\{X(t)\}$$

- Two main families:
  - W_s deterministic: Mono-fractal processes (e.g. Brownian Motion)
  - W_s random: Multifractal processes.
- Multifractality allows the distribution to change with scale: *intermittency.*

Scattering Renormalization

First Order:

$$\tilde{S}X(j_1) = \frac{SX(j_1)}{SX(1)}$$

(Invariance to global amplitude changes)

Second Order:

$$\tilde{S}X(j_1, j_2) = \frac{SX(j_1, j_2)}{SX(j_1)} = \frac{E(||X \star \psi_{j_1}| \star \psi_{j_2}|)}{E(|X \star \psi_{j_1}|)} \ , \ j_1, j_2 \in \mathbb{Z}$$

Renormalisation Properties

- Invariance to Self-similarity:

Proposition: If $\{X(2^j t)\}_t \overset{l}{=} A_j \{X(t)\}_t$, then

$$\forall j_1 \ , \tilde{S}X(j_1, j_2) = \tilde{S}X(j_2 - j_1) \ .$$

Renormalisation Properties

- Invariance to Self-similarity:

Proposition: If $\{X(2^j t)\}_t \stackrel{l}{=} A_j \{X(t)\}_t$, then

$$\forall j_1 \ , \tilde{S}X(j_1, j_2) = \tilde{S}X(j_2 - j_1) \ .$$

- Near Invariance to Fractional Derivatives:

Proposition: If $LX = X \star h$ is such that $\forall j \ \{|X \star L\psi_j|\}_t \stackrel{l}{=} C_j \{|X \star \psi_j|\}_t$, then

$$\tilde{S}X(j_1, j_2) = \tilde{S}(LX)(j_1, j_2) \ .$$

– For wavelets well localized in frequency,

$$D^\alpha \psi_j \approx C_j \psi_j \quad , \text{ hence } \tilde{S}X(j_1, j_2) \approx \tilde{S}D^\alpha X(j_1, j_2) \ .$$

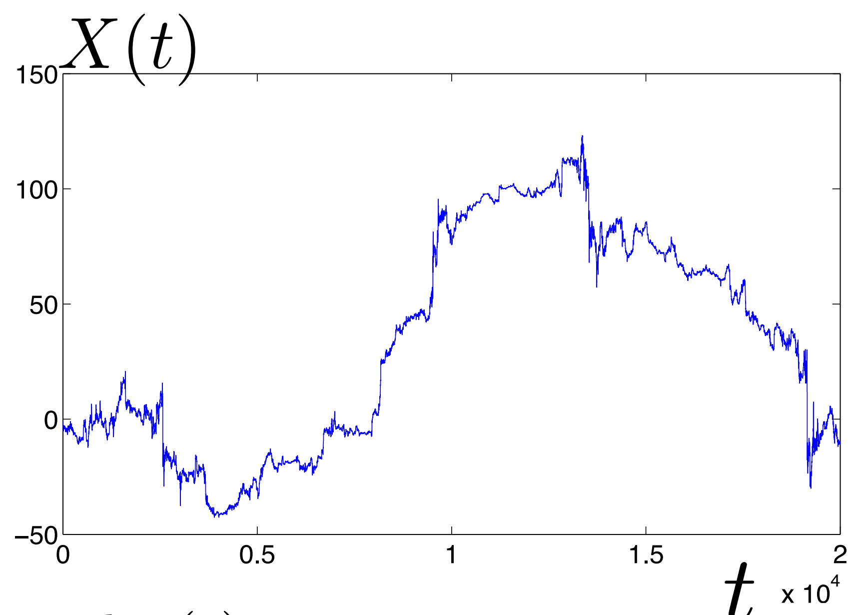
Fractional Derivative Near Invariance

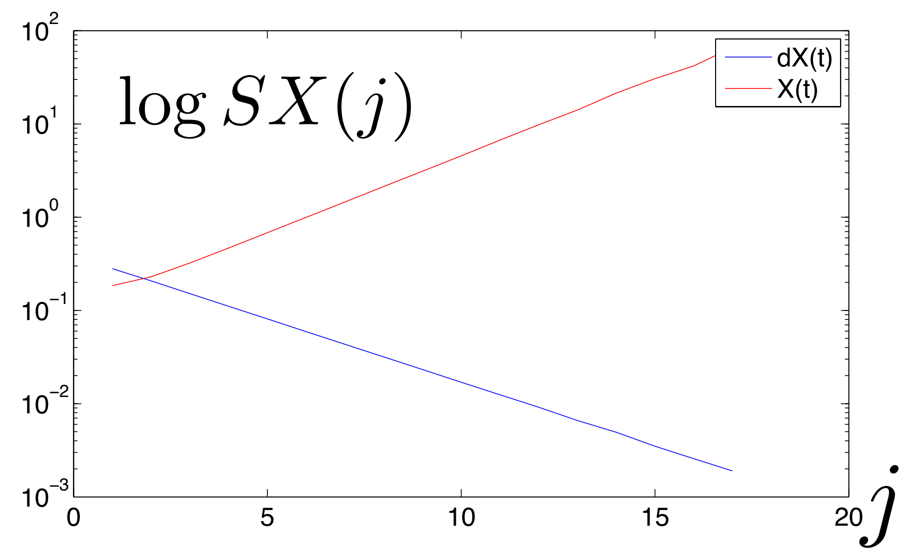
Proposition: If $LX = X \star h$ is such that $\forall j \ \{|X \star L\psi_j|\}_t \stackrel{l}{=} C_j \{|X \star \psi_j|\}_t$, then

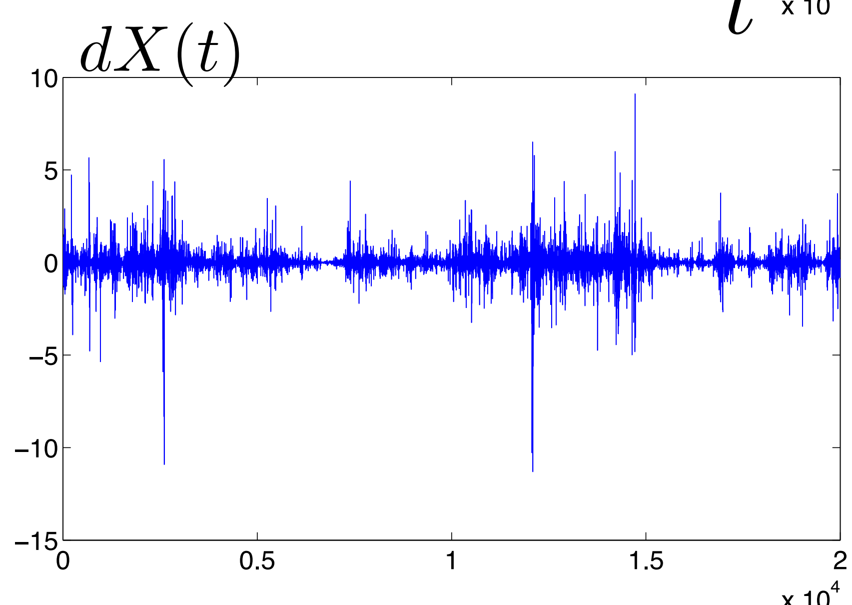
$$\tilde{S}X(j_1, j_2) = \tilde{S}(LX)(j_1, j_2) \ .$$

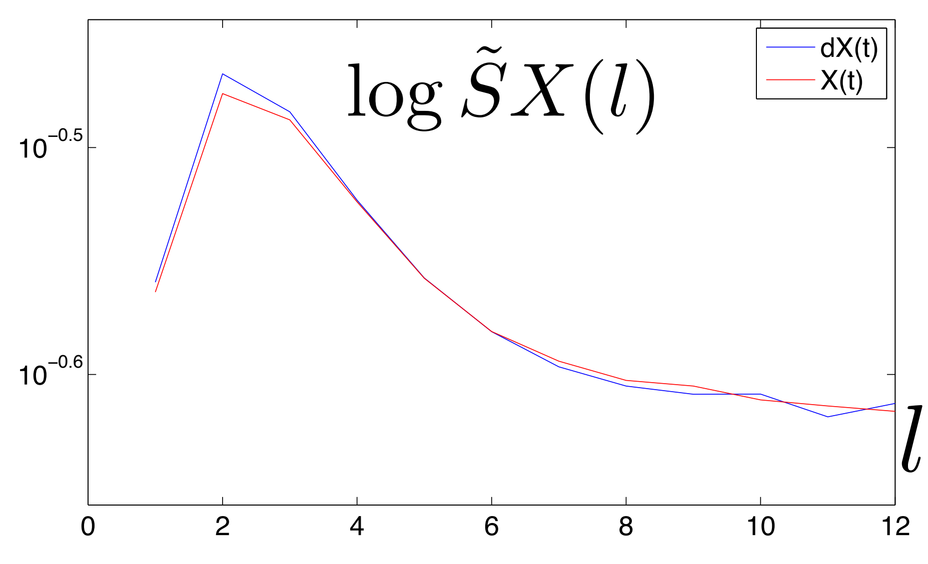
– For wavelets well localized in frequency,

$$D^\alpha \psi_j \approx C_j \psi_j \quad , \text{ hence } \quad \tilde{S}X(j_1, j_2) \approx \tilde{S}D^\alpha X(j_1, j_2) \ .$$





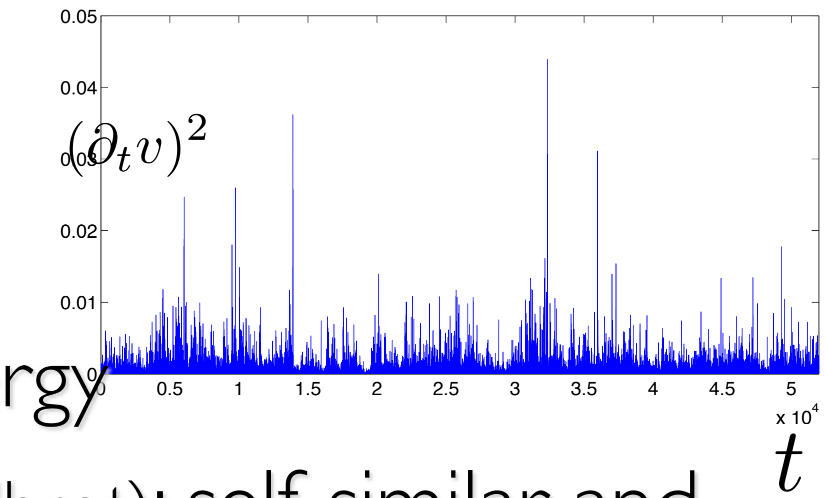




Intermittent Processes

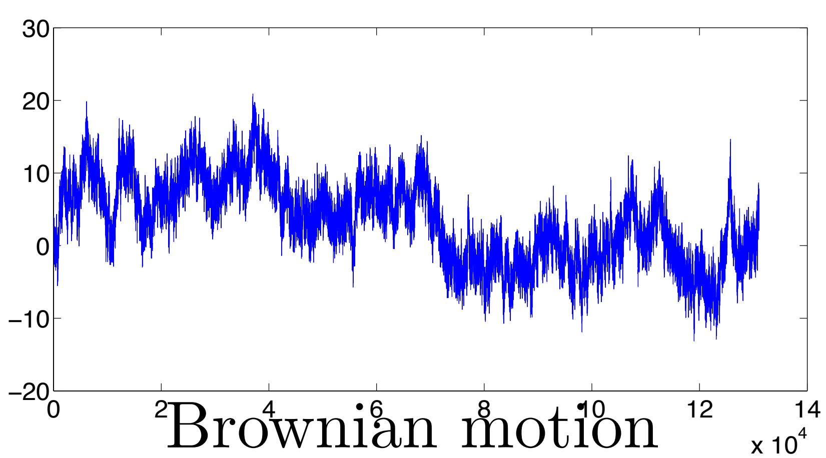
- First Order Decay: Hurst exponent:

$$SX(j) = E(|X \star \psi_j|) \simeq 2^{jH}$$

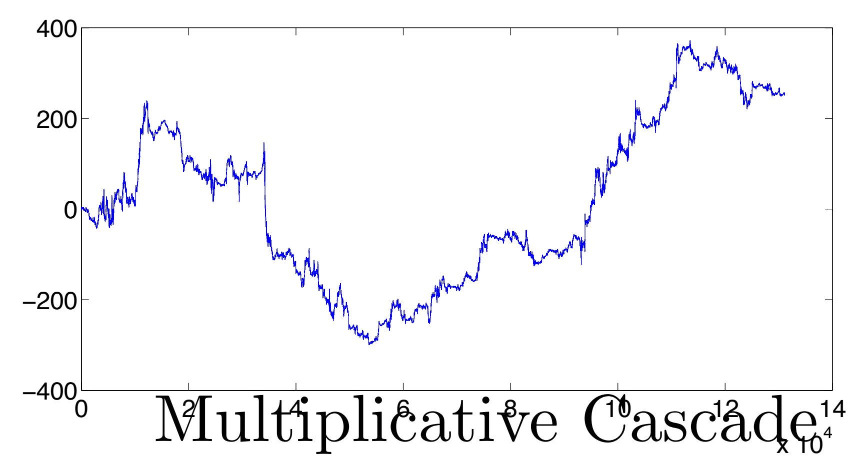


- Intermittency
  - In Turbulence: irregular dissipation of kinetic energy
  - Multiplicative Canonical Cascades (Yaglom, Mandelbrot): self-similar and intermittent (multifractal)
  - Can be defined from q-order wavelet moments:

$$E(|X \star \psi_j|^q) \simeq 2^{j\zeta(q)} \;\; (j \to -\infty) \qquad \text{Intermittency: curvature of } \zeta(q)$$



Brownian motion



Multiplicative Cascade

- How to efficiently measure intermittency?

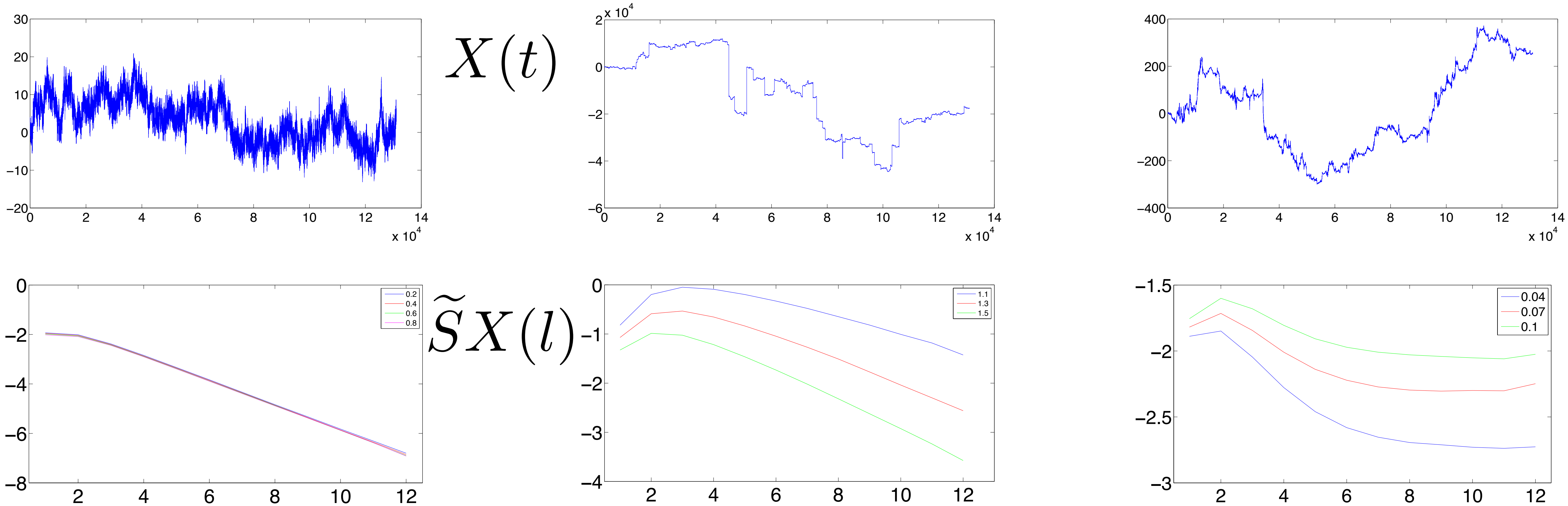
Scattering and Intermittency

Theorem [BBMM'13]:

If $X(t)$ Fractional Brownian Motion, then $\tilde{S}X(l) \simeq 2^{-l/2}$,

If $X(t)$ α-stable Lévy process, then $\tilde{S}X(l) \simeq 2^{l(\alpha^{-1}-1)}$,

If $X(t)$ Multiplicative Random Cascade, then $\tilde{S}X(l) \simeq O(1)$,



$X(t) \sim$ FBM $\qquad X(t) \sim$ Lévy $\qquad X(t) \sim$ MRW

Second Order: Measure of Multiscale Intermittency

Forgery Detection

(from Charlotte dataset)

[with I.Daubechies]

Original?

Forged?

Forgery Detection

First order coefficients: $SX(j, \theta) = E(|X \star \psi_{j,\theta}|)$

Renormalized second order coefficients:

$$\tilde{S}X(j_1, j_2, \theta_1, \theta_2) = \frac{SX(j_1, j_2, \theta_1, \theta_2)}{SX(j_1, \theta_1)}$$

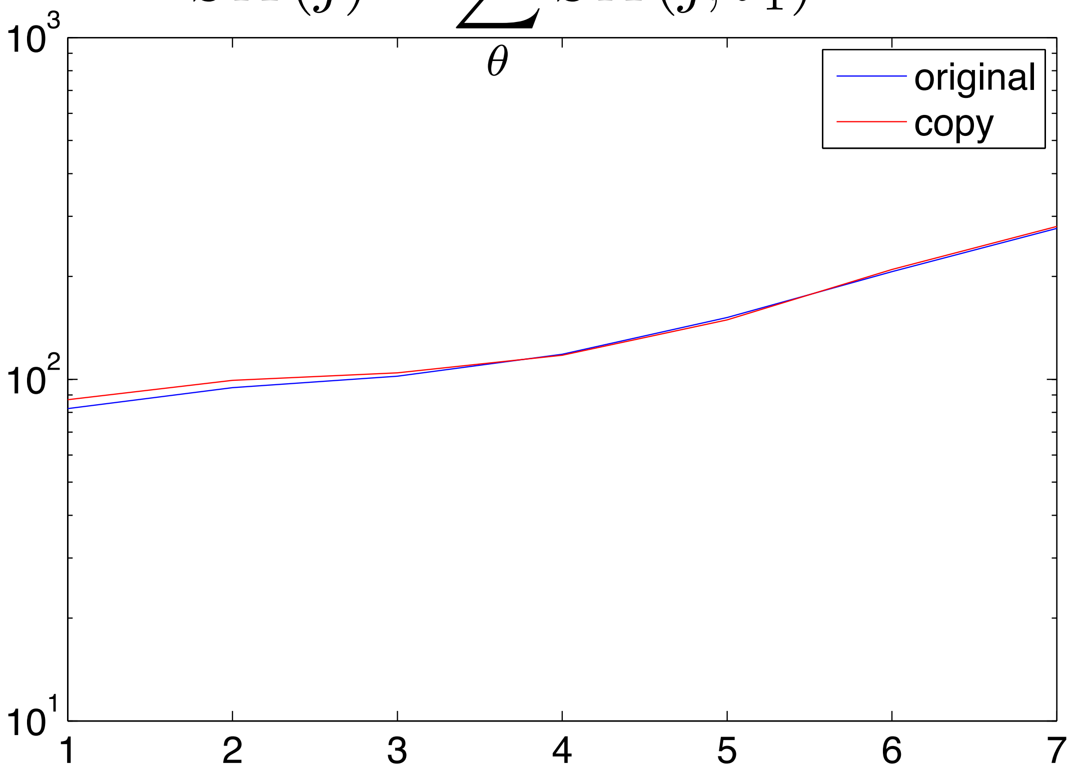
Forgery Detection

First order coefficients: $SX(j,\theta) = E(|X \star \psi_{j,\theta}|)$

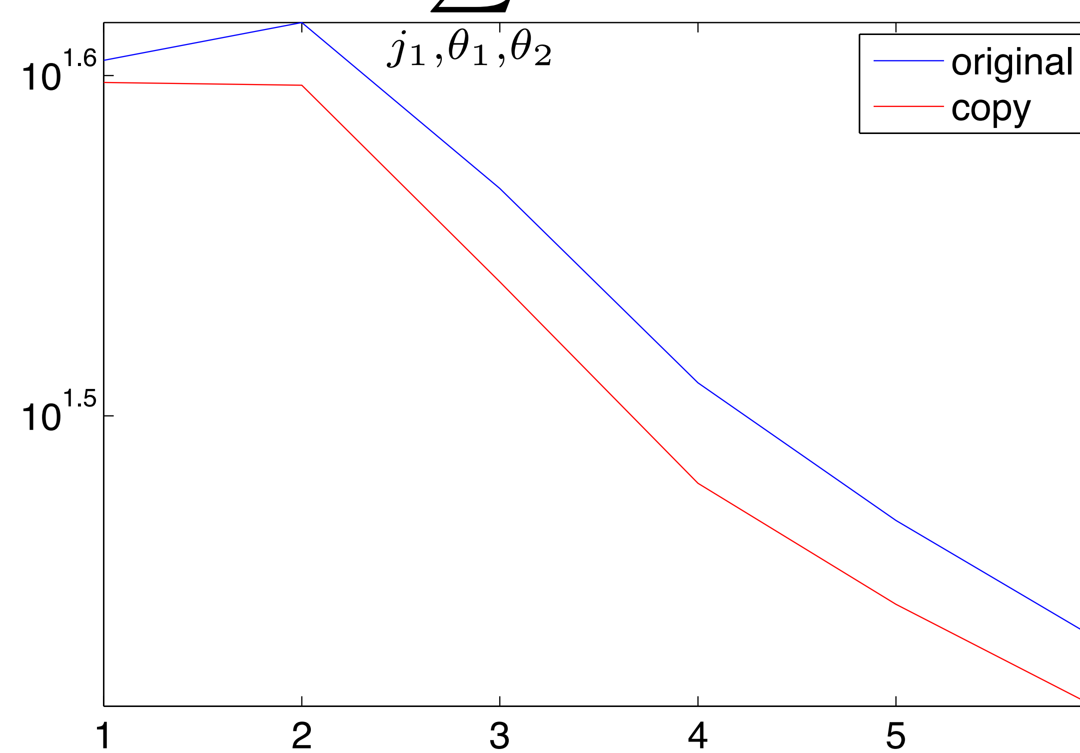
Renormalized second order coefficients:

$$\tilde{S}X(j_1,j_2,\theta_1,\theta_2) = \frac{SX(j_1,j_2,\theta_1,\theta_2)}{SX(j_1,\theta_1)}$$

$$\overline{S}X(j) = \sum_{\theta} SX(j,\theta_1)$$



$$\overline{\tilde{S}}X(l) = \sum_{j_1,\theta_1,\theta_2} \tilde{S}X(j_1,j_1+l,\theta_1,\theta_2)$$



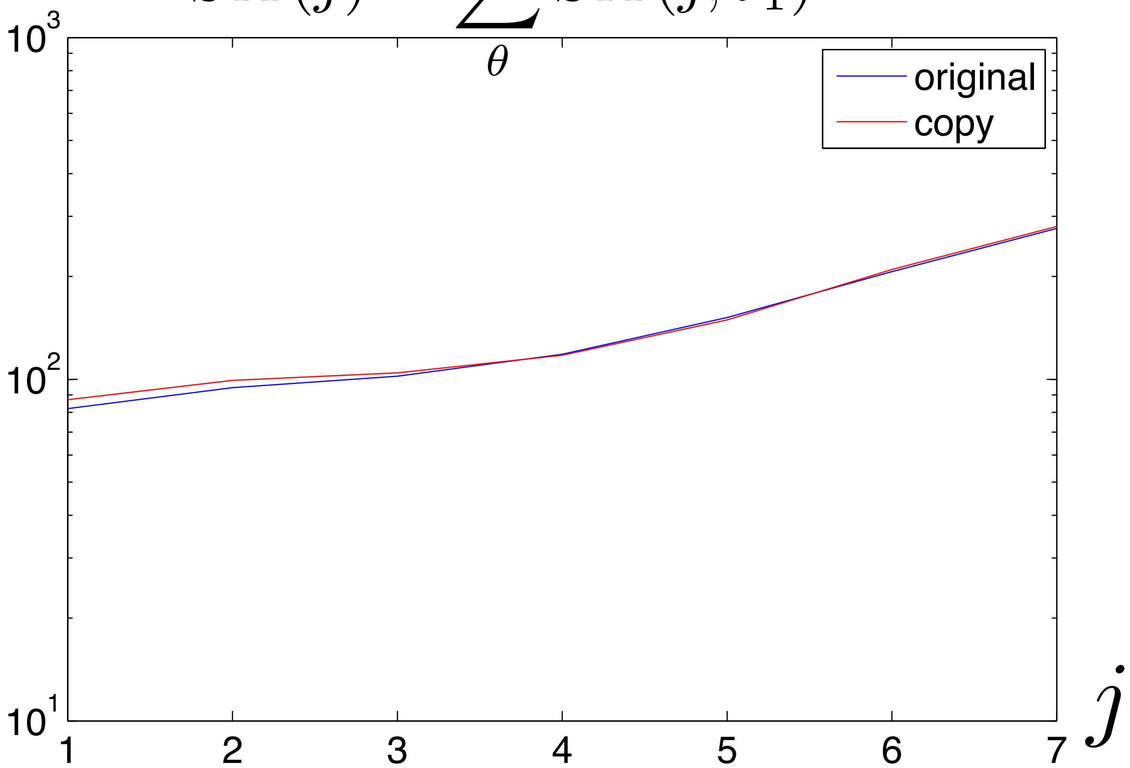
Forgery Detection

First order coefficients: $SX(j,\theta) = E(|X \star \psi_{j,\theta}|)$

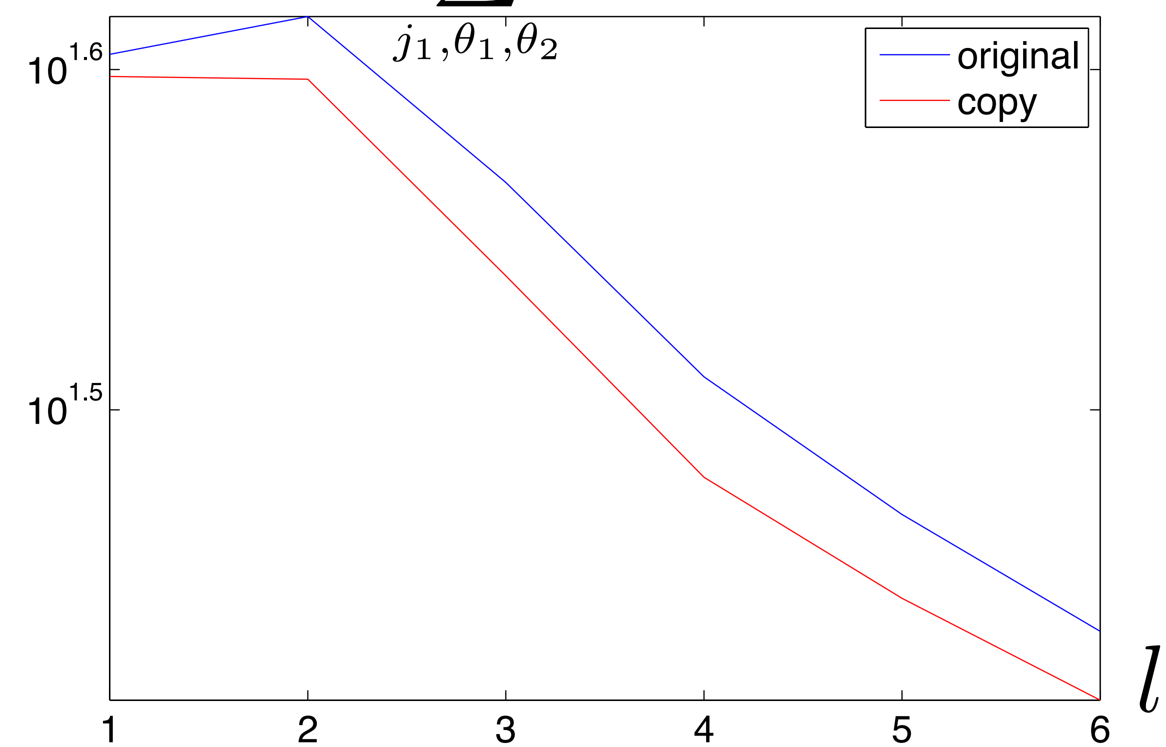
Renormalized second order coefficients:

$$\tilde{S}X(j_1, j_2, \theta_1, \theta_2) = \frac{SX(j_1, j_2, \theta_1, \theta_2)}{SX(j_1, \theta_1)}$$

$$\overline{S}X(j) = \sum_{\theta} SX(j, \theta_1)$$



$$\overline{\tilde{S}}X(l) = \sum_{j_1,\theta_1,\theta_2} \tilde{S}X(j_1, j_1 + l, \theta_1, \theta_2)$$



Wilcoxon RankSum Test
(assuming independent patches)

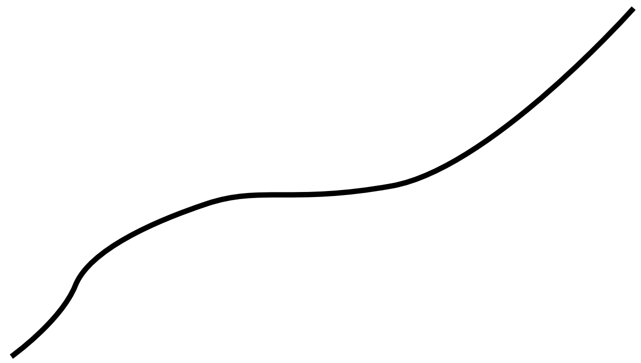
$SX(j,\theta)\ :\ p = 0.54$

$\tilde{S}X(j_1 - j_2, \theta_1, \theta_2)\ :\ p = 0.00025$

"A posteriori" Interpretation

Original

Forged



Geometric regularity: More intermittent

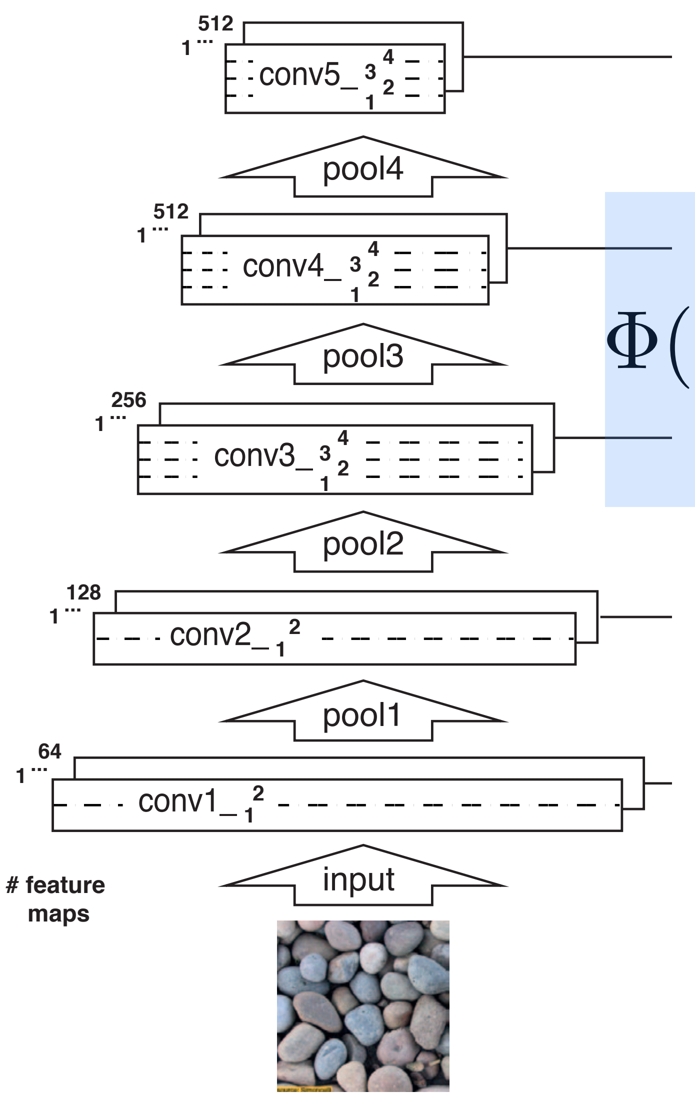
CNNs for Texture Representation

- Q:How to obtain a texture representation from a CNN?

CNNs for Texture Representation

- Q:How to obtain a texture representation from a CNN?
- Simple, yet powerful, idea [Gatys et al.'15]:

Let $(\Phi_1(x)(u_1, \lambda_1), \Phi_2(x)(u_2, \lambda_2), \ldots, \Phi_K(x)(u_K, \lambda_K))$ the outputs of each layer of a pre-trained CNN

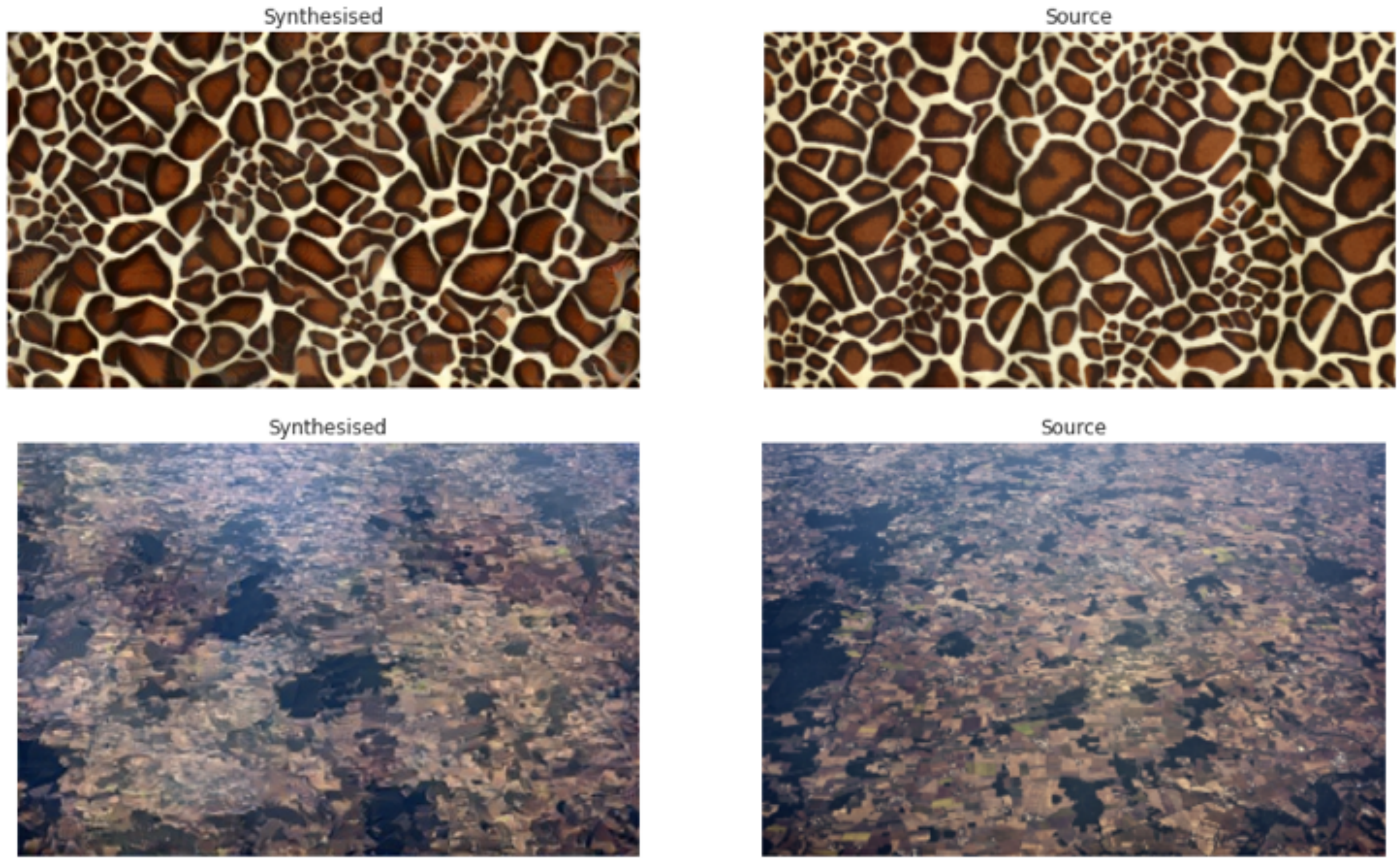


Stationary or "style" representation:

$$\Phi(x) = \left\{ \frac{1}{N_k} \sum_{u_k} \Phi_k(x)(u_k, \cdot)\Phi_k(x)(u_k, \cdot)^T \ , \ k = 1 \leq K \right\}$$

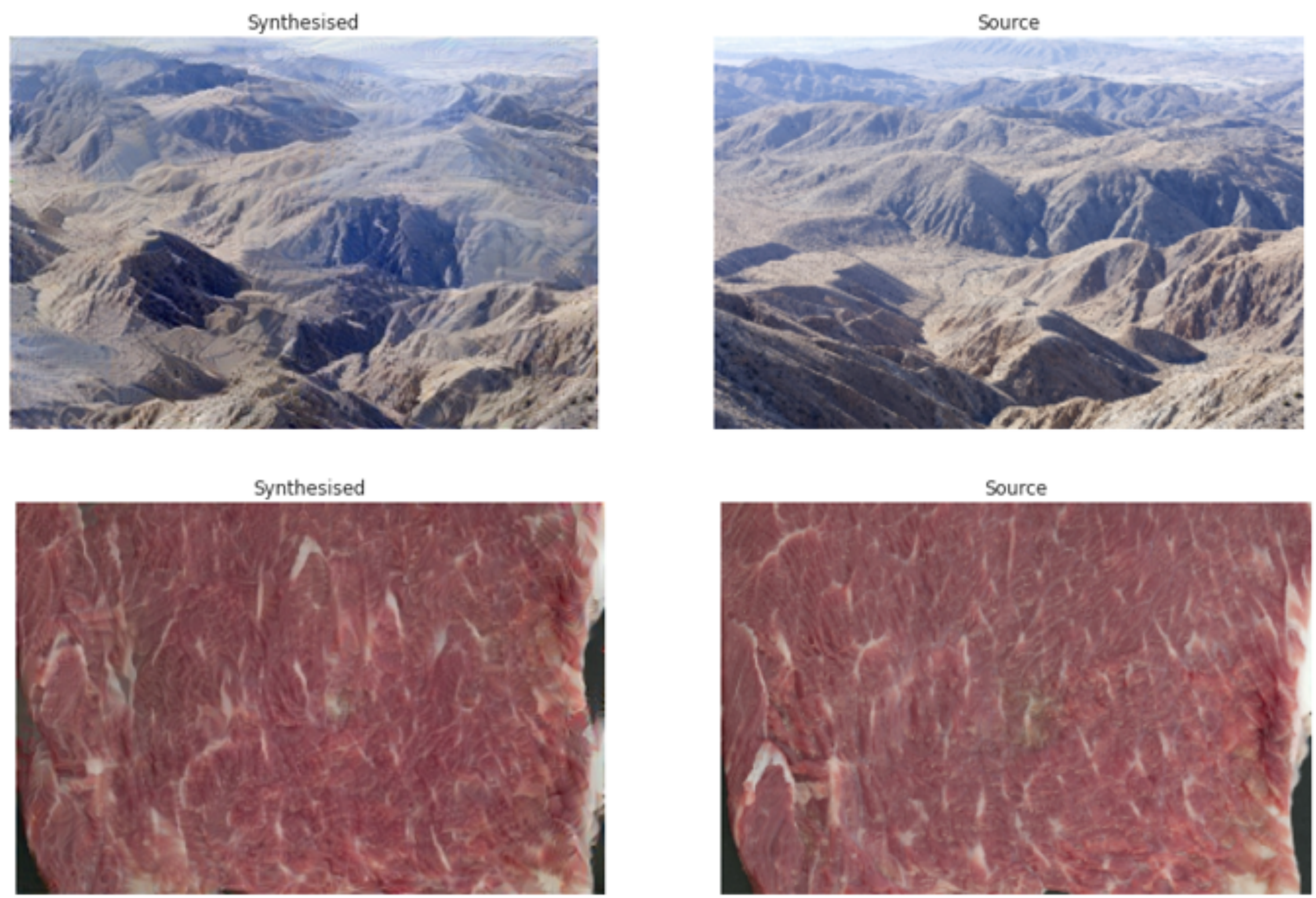
Ergodic Texture Reconstruction

- Scattering Moments of 2nd order capture essential geometric structures with only $O((\log N)^2)$ coefficients.
- However, not all texture geometry is captured.
- Results using a deep VGG network from [Gathys et al, NIPS'15]



Ergodic Texture Reconstruction

- Scattering Moments of 2nd order capture essential geometric structures with only $O((\log N)^2)$ coefficients.
- However, not all texture geometry is captured.
- Results using a deep VGG network from [Gathys et al, NIPS'15]



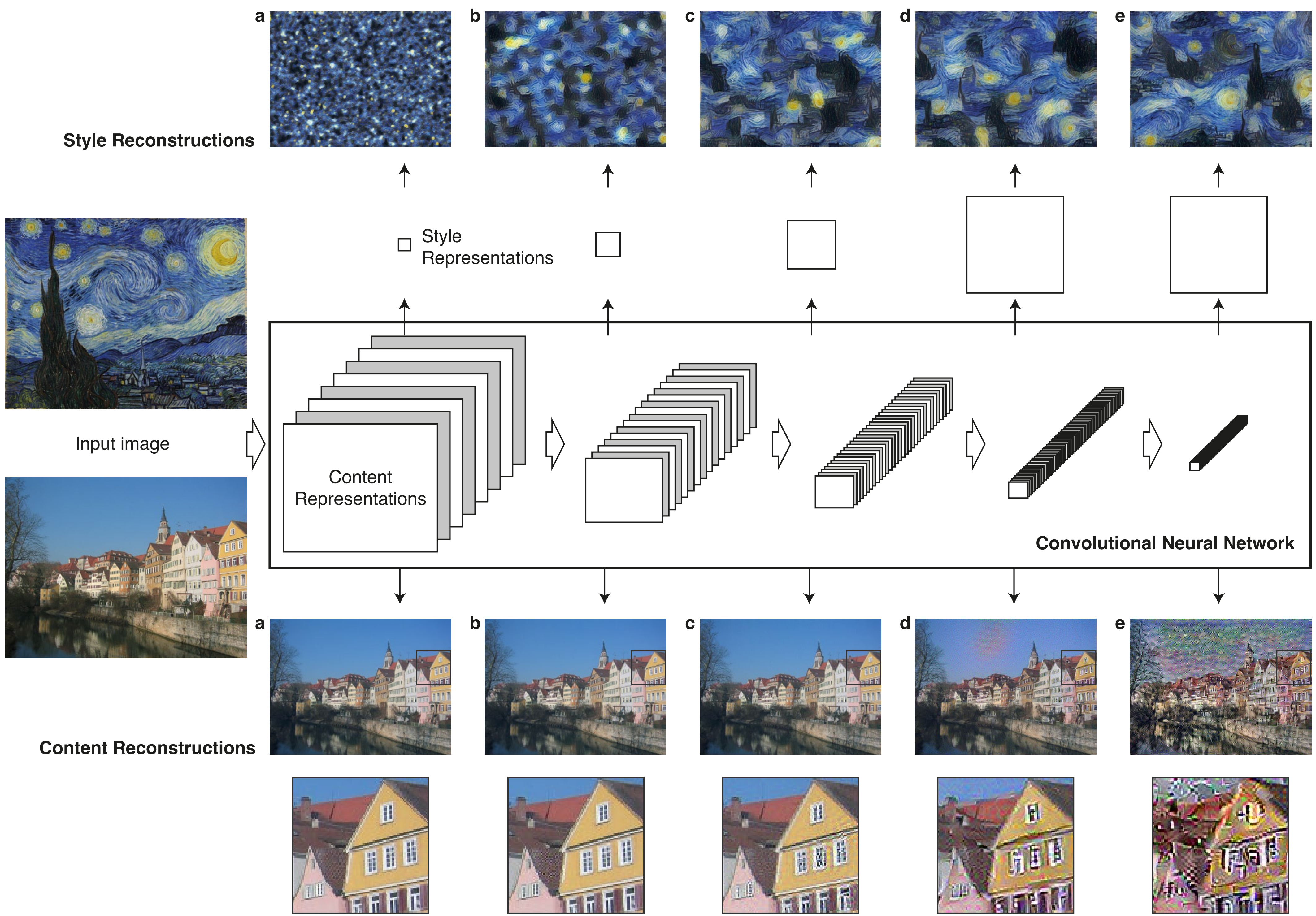
Texture and Geometry

- We have seen that both in the case of scattering and in general CNNs, texture and template/geometry representations use the same nonlinearities
  - We only change the pooling operator to adapt to stationarity.

- Q: Can we disentangle texture and geometry by combining these two representations?

Texture and Geometry

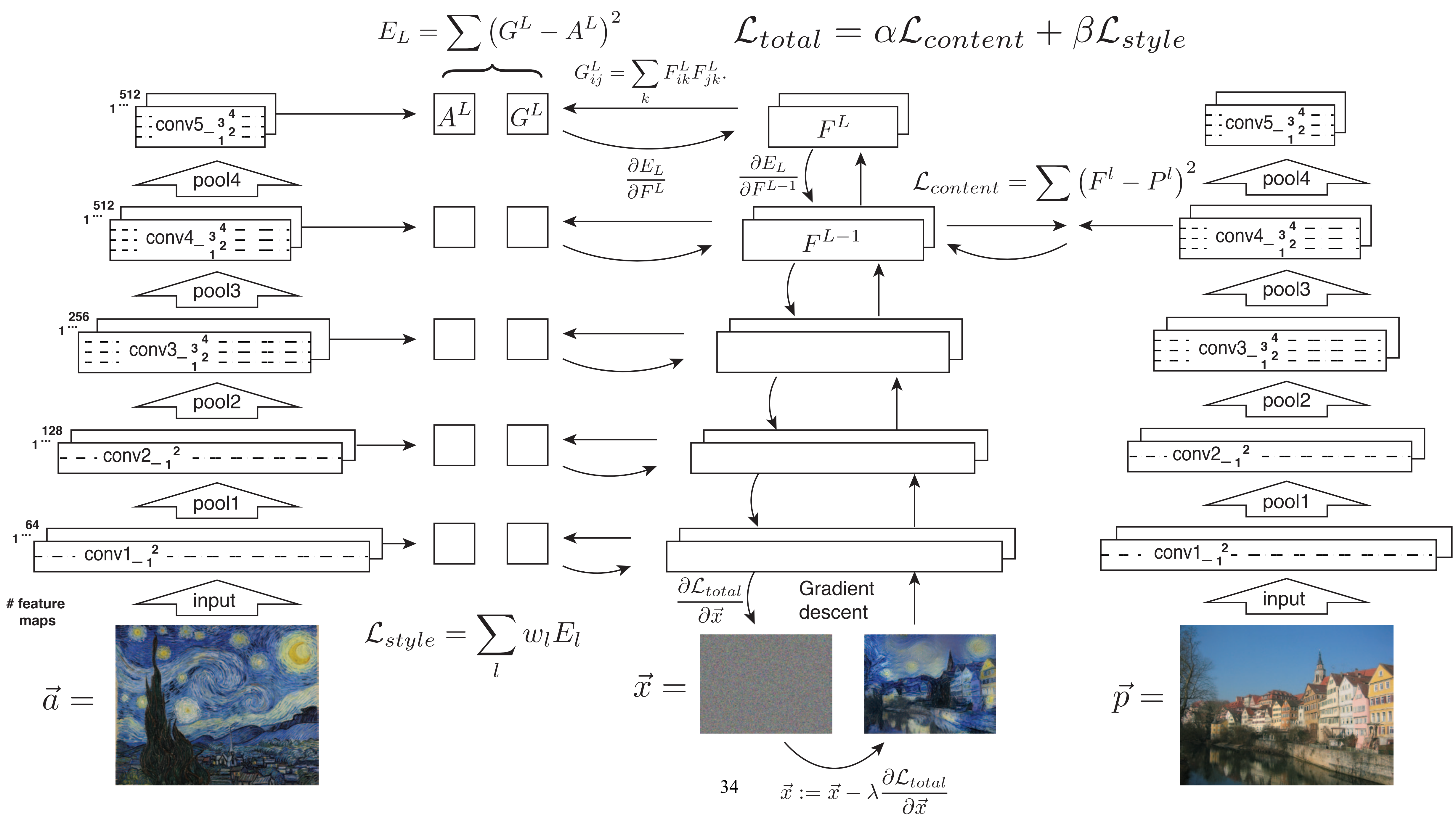
- "StyleNet", Gatys et al,'15.



Texture and Geometry

- "StyleNet", Gatys et al,'15.

Given x_1 and x_2, we look for $\hat{x}$ such that $\Phi_s(x_1) \approx \Phi_s(\hat{x})$ and $\Phi(x_2) \approx \Phi(\hat{x})$.



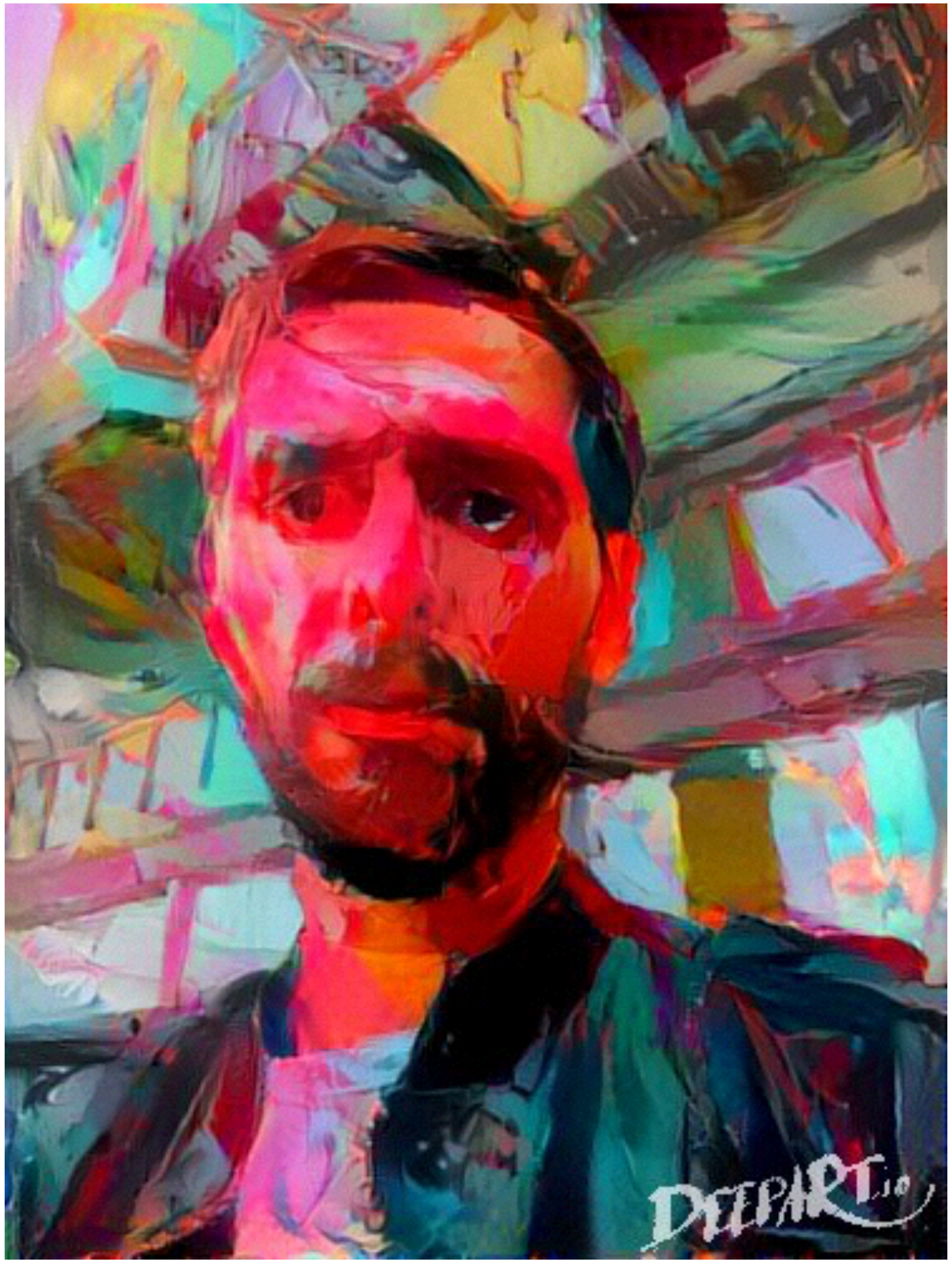
Texture and Geometry

Texture and Geometry

- Advertisement of the MLSS I am co-organizing:

Texture and Geometry

- Check out your own pictures at deepart.io!

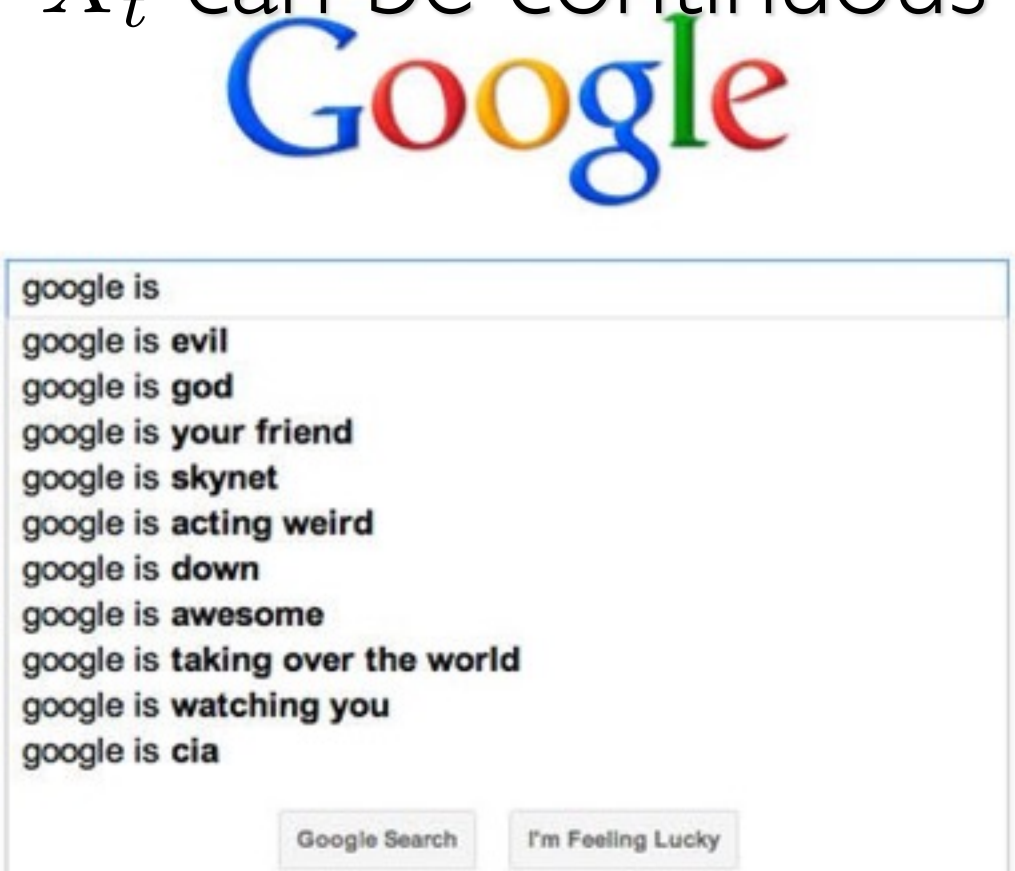


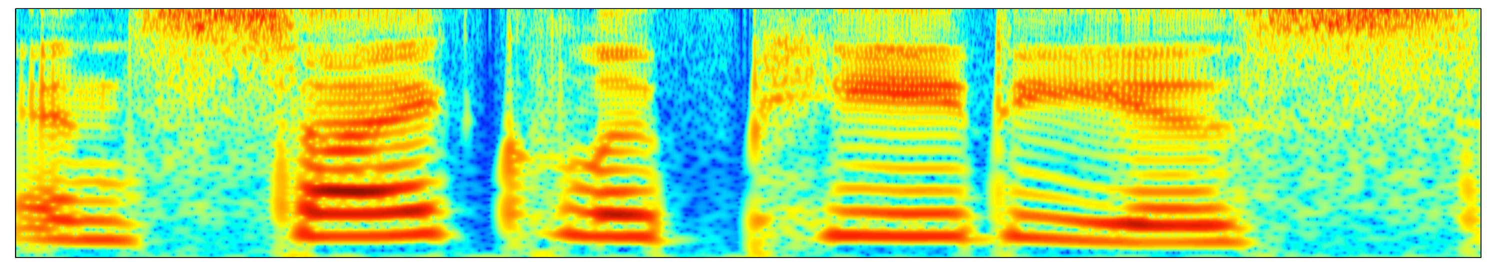
Time Series

- An ordered sequence of (multivariate) random variables:

$$\{X_t\}_{t \in \mathbb{N}}$$

- X_t can be continuous or discrete:





- Important Statistical assumption:

$$p(X_{t+\tau_1}, X_{t+\tau_2}, \ldots, X_{t+\tau_k}) = p(X_{\tau_1}, X_{\tau_2}, \ldots, X_{\tau_k}) \quad , \ \forall \ t, \tau_1, \ldots, \tau_k$$

We say that $\{X_t\}$ is stationary.

Time Series Tasks

- Statistical Modeling:
  - Speech Synthesis, Music generation, etc.
- Forecasting/Prediction:
  - Biostatistics.
  - Financial applications
- Regression/Classification:
  - Sentiment Analysis
  - Action Recognition.
  - Speech Recognition.
  - Machine Translation, Question/Answering.

Curse of Dimensionality

- As t increases, complexity of $P(X_1, \ldots, X_t)$ increases exponentially

- Thus we need to introduce models that have finite amount of capacity.
  - Stationarity implies capacity should be constant in time.

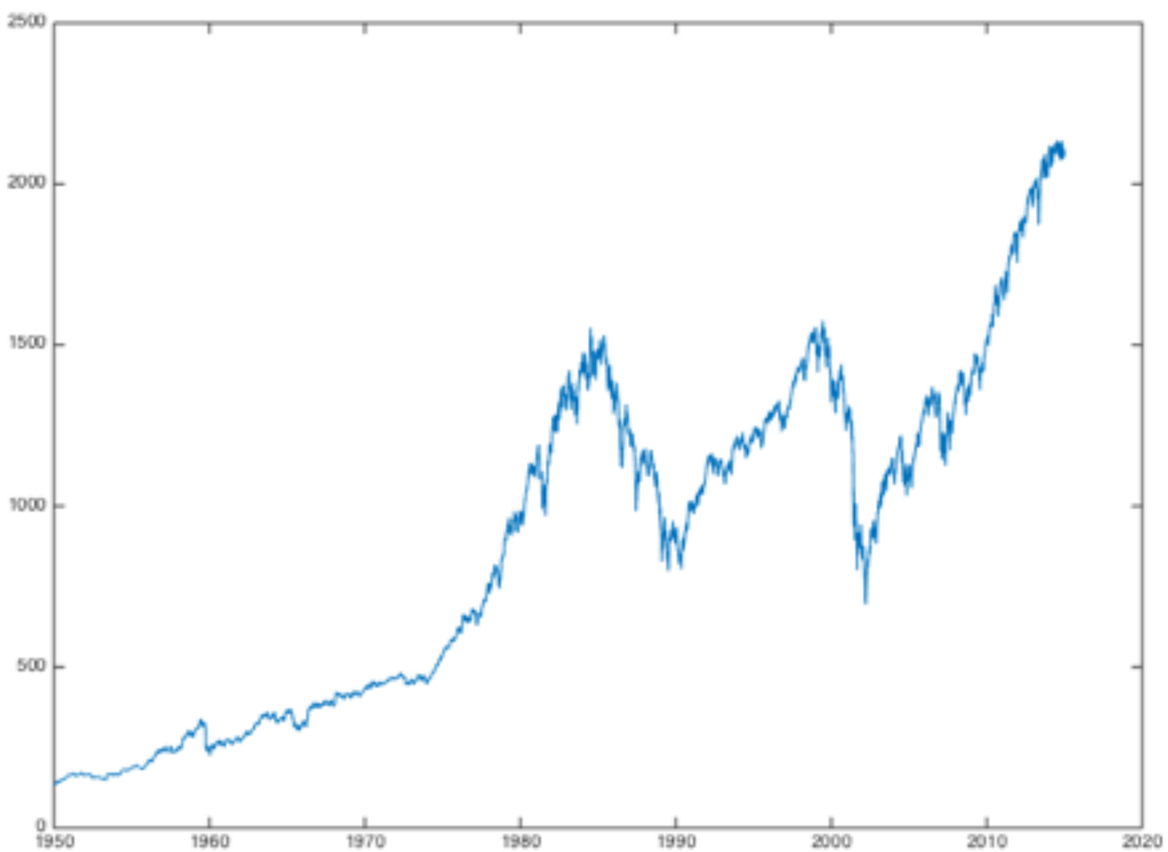
- Q: What does this assumption require/imply?

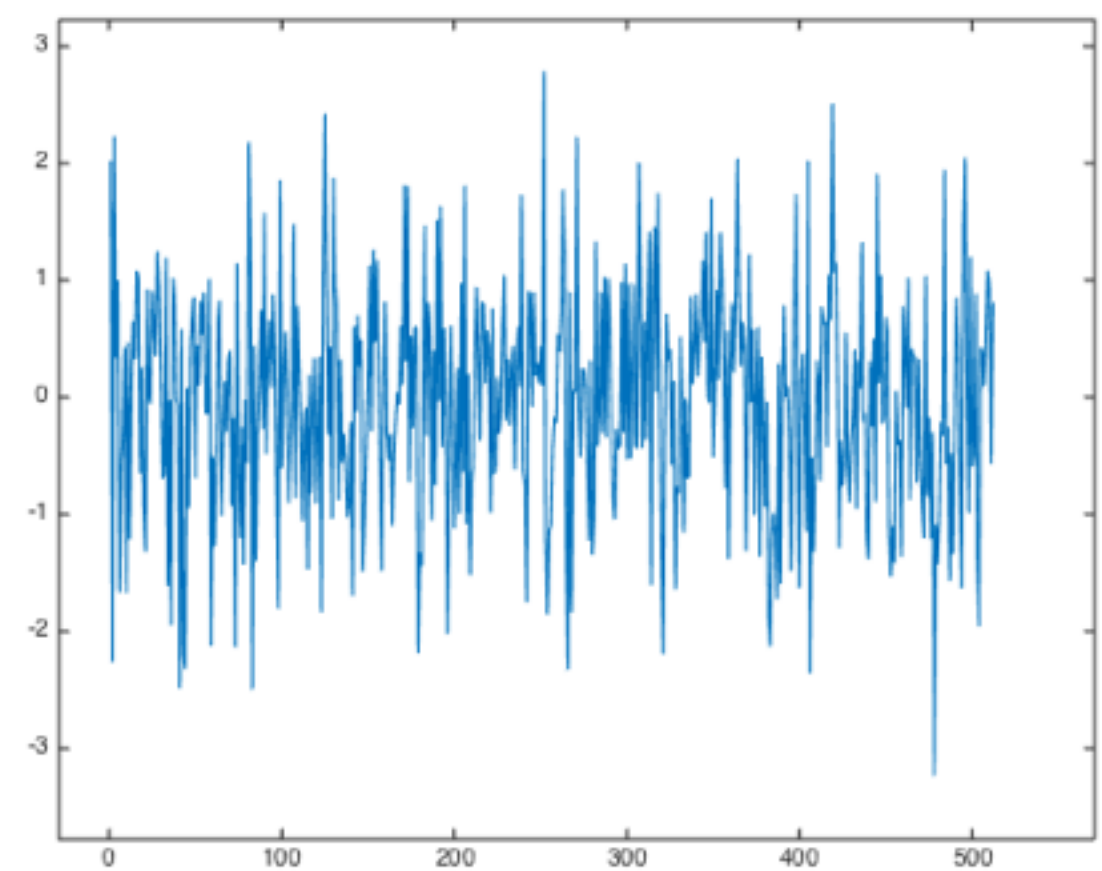
Memory of a Process

- Measure of the statistical dependency between X_t and $X_{t+\tau}$
  - A particularly simple measure is through the second-order moments:

$$\|R_X\|_1 = \sum_k |R_X(k)| \text{ measures decorrelation scale}$$

$$R_X(\tau) \simeq |\tau|^{-\alpha}$$



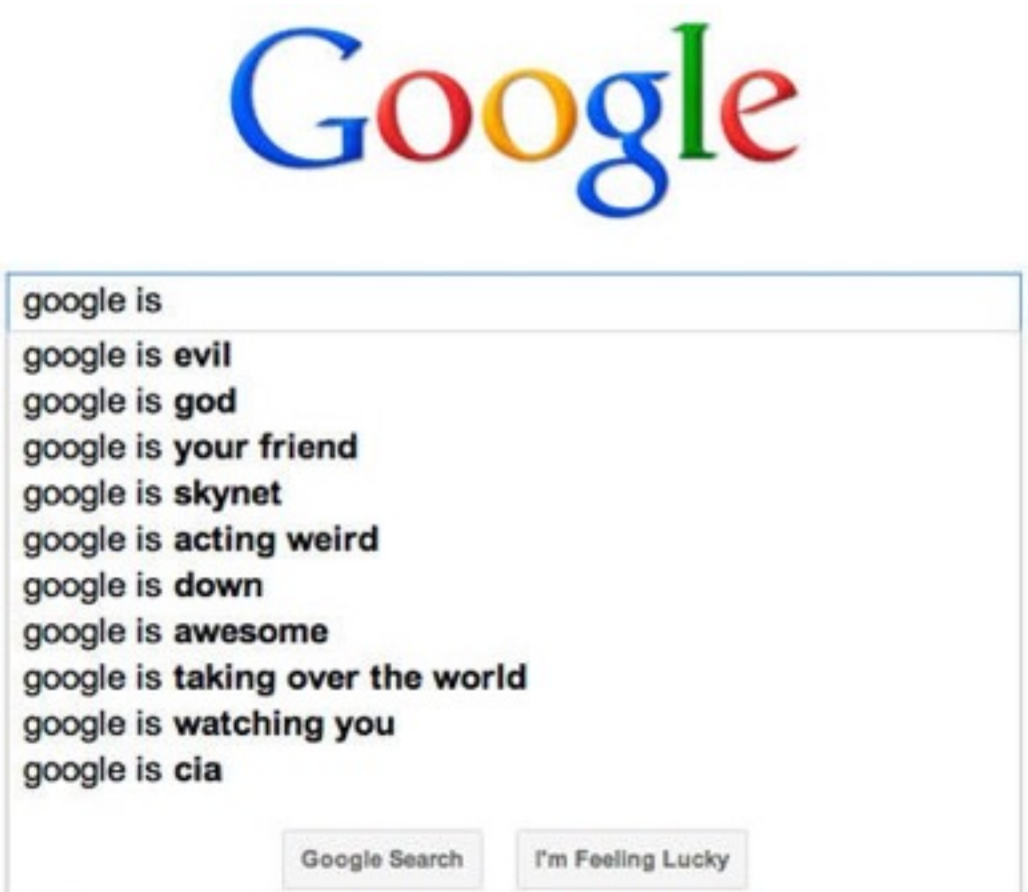


Memory of a Process

- For discrete time series, we can use a divergence between the joint distribution of $(X_t, X_{t+\tau})$ and the product of its marginals:

$$m_X(\tau) = D_{KL}\left(p(X_t, X_{t+\tau}) \mid\mid p(X_t)p(X_{t+\tau})\right)$$

$$m_X(\tau) \simeq |\tau|^{-\alpha}$$



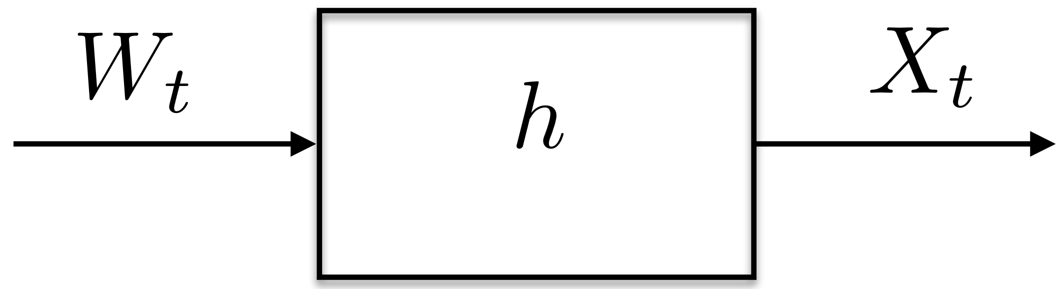
Stationary Time Series Models

- A stationary process with no memory is called a white noise:

$$\{W_t\} \quad iid. \qquad\qquad W_t \sim F_\theta$$

- A general class of stationary processes is obtained by filtering white noise with an integrable kernel:

$$X_t = W_t \star h \ , \ \text{with } \|h\|_1 = \sum_k |h_k| < \infty \ , \ \mathbb{E}W_t = 0 \ .$$



These are called *linear* processes.

Stationary Time Series Models

- Pure Autoregressive Processes (AR(p)):

$$X_t - a_1 X_{t-1} - \dots a_p X_{t-p} = W_t$$

- Moving Average Processes (MA(q)):

$$X_t = W_t + b_1 W_{t-1} + b_q W_{t-q}$$

- ARMA(p,q):

$$X_t - a_1 X_{t-1} - \dots a_p X_{t-p} = W_t + b_1 W_{t-1} + b_q W_{t-q}$$

- Second-order moments are sufficient to fitting parameters (Yule-Walker Equations).

Spectral Theory for Stationary Processes

- Denote by B the *shift* or translation operator: $BX_t = X_{t-1}$

- Then the previous models can be rewritten as

$$X_t - a_1 X_{t-1} - \ldots a_p X_{t-p} = W_t + b_1 W_{t-1} + b_q W_{t-q}$$

$$(1 - a_1 B - \ldots a_p B^p) X_t = (1 + b_1 B + \ldots b_q B^q) W_t$$

$$X_t = \frac{1 + b_1 B + \ldots b_q B^q}{1 - a_1 B - \ldots a_p B^p} W_t$$

Spectral Theory for Stationary Processes

- Denote by B the *shift* or translation operator: $BX_t = X_{t-1}$

- Then the previous models can be rewritten as

$$X_t - a_1 X_{t-1} - \ldots a_p X_{t-p} = W_t + b_1 W_{t-1} + b_q W_{t-q}$$

$$(1 - a_1 B - \ldots a_p B^p) X_t = (1 + b_1 B + \ldots b_q B^q) W_t$$

$$X_t = \frac{1 + b_1 B + \ldots b_q B^q}{1 - a_1 B - \ldots a_p B^p} W_t$$

- This is a convolution:

Suppose h has $q+1$ taps $(h_0, \ldots, h_q)$:

$$X \star h(t) = \sum_{k=0}^{q} h_k X_{t-k} = \sum_{k=0}^{q} h_k B^k X_t = \left(\sum_k h_k B^k \right) X_t$$

Spectral Theory for Stationary Processes

- We cannot easily define a Fourier transform of a stationary process (without random measure theory).

Spectral Theory for Stationary Processes

- We cannot easily define a Fourier transform of a stationary process (without random measure theory)
- But we can easily define the Fourier transform of its autocorrelation:

$$\hat{R}_X(e^{i\omega}) = \sum_k R_X(k)e^{-i\omega k}$$

Spectral Theory for Stationary Processes

- We cannot easily define a Fourier transform of a stationary process (without random measure theory)
- But we can easily define the Fourier transform of its autocorrelation

$$\hat{R}_X(e^{i\omega}) = \sum_k R_X(k) e^{-i\omega k}$$

- In terms of the autocorrelation

$$\hat{R}_X(e^{i\omega}) = \sigma^2 \frac{|1 + b_1 e^{i\omega} + \cdots + b_q e^{iq\omega}|^2}{|1 - a_1 e^{i\omega} - \cdots - a_p e^{ip\omega}|^2}$$

- Zeros and Poles decomposition:

$$\hat{R}_X(e^{i\omega}) = \sigma^2 \frac{\prod_{k \leq q} |e^{i\omega} - z_k|^2}{\prod_{k' \leq p} |e^{i\omega} - p_{k'}|^2}$$

Forecasting

- Q: Given $X_1 = x_1, \ldots, X_t = x_t$, how to estimate X_{t+1}?
- When X_t are continuous random variables, we can consider

$$\mathbb{E}(|\hat{X}_{t+1} - X_{t+1}|^2 \mid X_1, \ldots, X_t)$$

Forecasting

- Q: Given $X_1 = x_1, \ldots, X_t = x_t$, how to estimate X_{t+1}?
- When X_t are continuous random variables, we can consider

$$\mathbb{E}(|\hat{X}_{t+1} - X_{t+1}|^2 \mid X_1, \ldots, X_t)$$

- For general noise models W_t and general nonlinear predictors $\hat{X}_{t+1} = F(X_1, \ldots, X_t)$, no closed form solution.
- Two important exceptions:
  - If W_t is Gaussian then optimal predictor is lineal and explicit.
  - Linear predictors only depend upon correlation measurements: efficient solution (Durbin-Levinson algorithm)

Forecasting

- Q: Given $X_1 = x_1, \ldots, X_t = x_t$, how to estimate X_{t+1}?
- When X_t are continuous random variables, we can consider

$$\mathbb{E}(|\hat{X}_{t+1} - X_{t+1}|^2 \mid X_1, \ldots, X_t)$$

- For general noise models W_t and general nonlinear predictors $\hat{X}_{t+1} = F(X_1, \ldots, X_t)$, no closed form solution.
- Two important exceptions:
  - If W_t is Gaussian then optimal predictor is lineal and explicit.
  - Linear predictors only depend upon correlation measurements: efficient solution (Durbin-Levinson algorithm)
- Limitations
  - Many predictions require a nonlinear component (hysteresis)
  - How to combine information from different sources?

State-space Models

- We can consider a hidden state Y_t with its own internal dynamics:

$$Y_{t+1} = F(Y_t, W_t)$$

W_t: Internal noise modeling uncertainty

- Hidden states influences observations X_t:

$$X_t = G(Y_t, Z_t)$$

Z_t: observational noise

- Q: How to infer the hidden states given observations? i.e $P(Y_t \mid X_1, \ldots, X_t)$

- Only tractable on particular models.

The Kalman Filter

- If we consider Gaussian Noises W_t, Z_t and Linear Dynamics, we have a fully Gaussian model.
- The posterior distribution of hidden states is also Gaussian, and is computed using the *Kalman Filter.*

The Kalman Filter

- If we consider Gaussian Noises W_t, Z_t and Linear Dynamics, we have a fully Gaussian model.
- The posterior distribution of hidden states is also Gaussian, and is computed using the *Kalman Filter.*

- Very useful in Control Theory: it can incorporate control variables.
- Parameter fitting possible with iterative schemes (such as EM algorithm).
- However, this is still a Gaussian model: poor modeling of highly non-linear phenomena.

Hidden Markov Models (HMMs)

- Suppose the hidden state Y_t is now a discrete random variable, taking N possible values.
- We can model $\{Y_t\}_t$ using a *Markov process*:

$$p(Y_1, \ldots, Y_t) = p(Y_1)p(Y_2 \mid Y_1) \ldots p(Y_t \mid Y_1, \ldots, Y_{t-1})$$

$$= p(Y_1) \prod_{i \leq t} p(Y_i \mid Y_{i-1})$$

Hidden Markov Models (HMMs)

- Suppose the hidden state Y_t is now a discrete random variable, taking N possible values.
- We can model $\{Y_t\}_t$ using a *Markov process*:

$$p(Y_1, \ldots, Y_t) = p(Y_1)p(Y_2 \mid Y_1) \ldots p(Y_t \mid Y_1, \ldots, Y_{t-1})$$

$$= p(Y_1) \prod_{i \leq t} p(Y_i \mid Y_{i-1})$$

- The transition probabilities are encoded with the matrix

$$\Pi_{k,l} = P(Y_i = c_k \mid Y_{i-1} = c_l) \ , \ \ k, l = 1, \ldots N$$

- Efficient learning and inference with EM-type algorithms
- Very successful in speech processing among others.

Limitations of HMMs

- The memory of the model is encoded with a state amongst N:
  - This amounts to $\log(N)$ bits.

Limitations of HMMs

- The memory of the model is encoded with a state amongst N:
  - This amounts to $\log(N)$ bits.

- In many high-dimensional systems, the information that the past conveys about the future is considerable
  - Speech Recognition: need to remember utterance, accent, pitch, syntax, etc.
  - Watching movies: remember the characters, the plot.

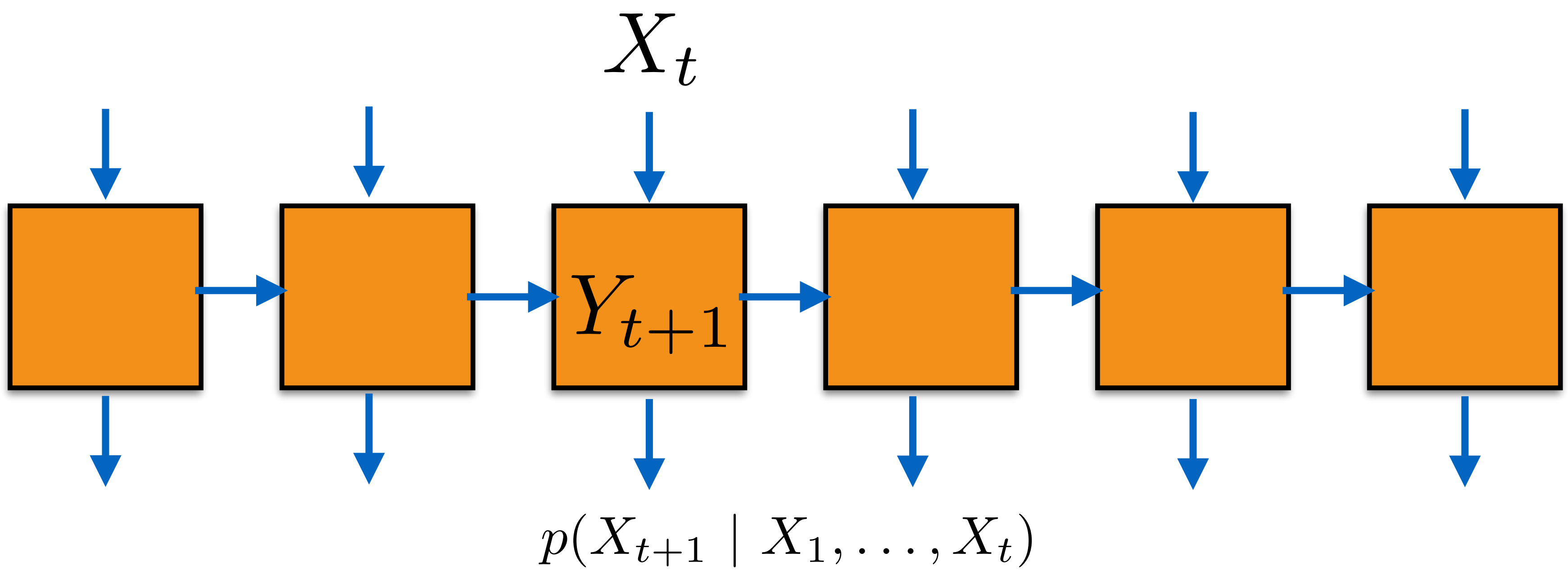
The required number of states grows exponentially with the amount of information.

Recurrent Neural Networks (RNN)

- We can combine the advantages of previous models into a non-linear continuous dynamical system:

$$p(X_1, \ldots, X_t) = \prod_{i \leq t} p(X_i \mid Y_i) \text{ with}$$

$$Y_i = F_\theta(Y_{i-1}, X_{i-1}) \qquad\qquad F_i \in \mathbb{R}^L$$



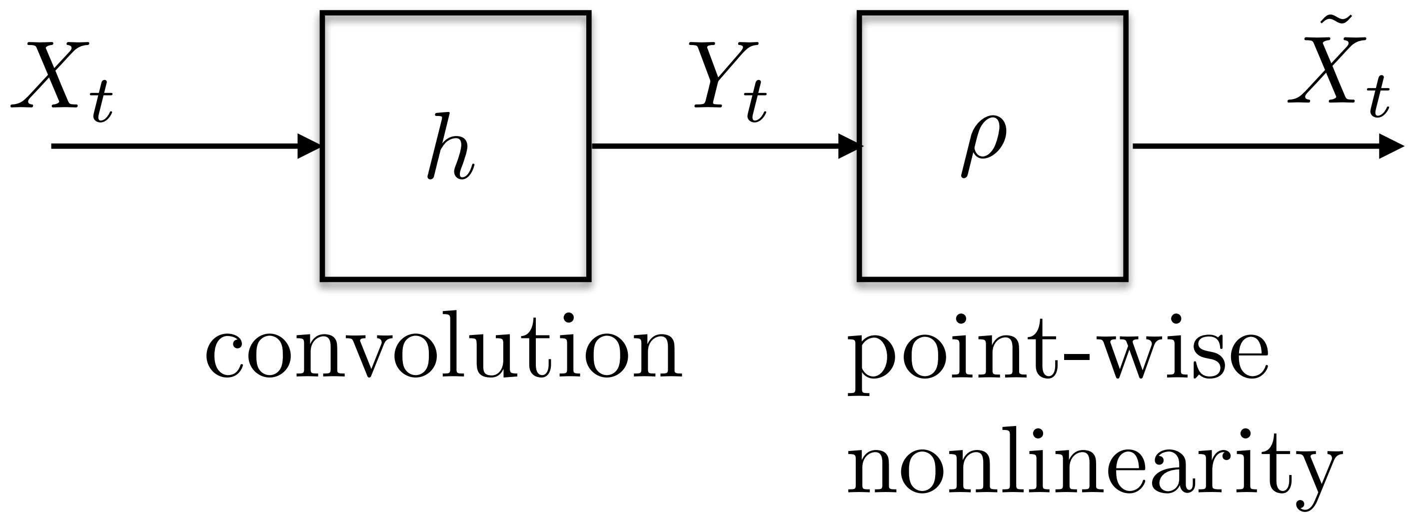
- Typically, we consider $F_\theta(Y_i, X_i) = \rho(A_{Y,Y} Y_{i-1} + A_{Y,X} X_i)$, with ρ a non-expansive point-wise nonlinearity.

RNNs and CNNs

- We can consider a CNN with IIR filters:

$$Y_t - a_1 Y_{t-1} - \ldots a_p Y_{t-p} = X_t \;\leftrightarrow\; Y = X \star h$$

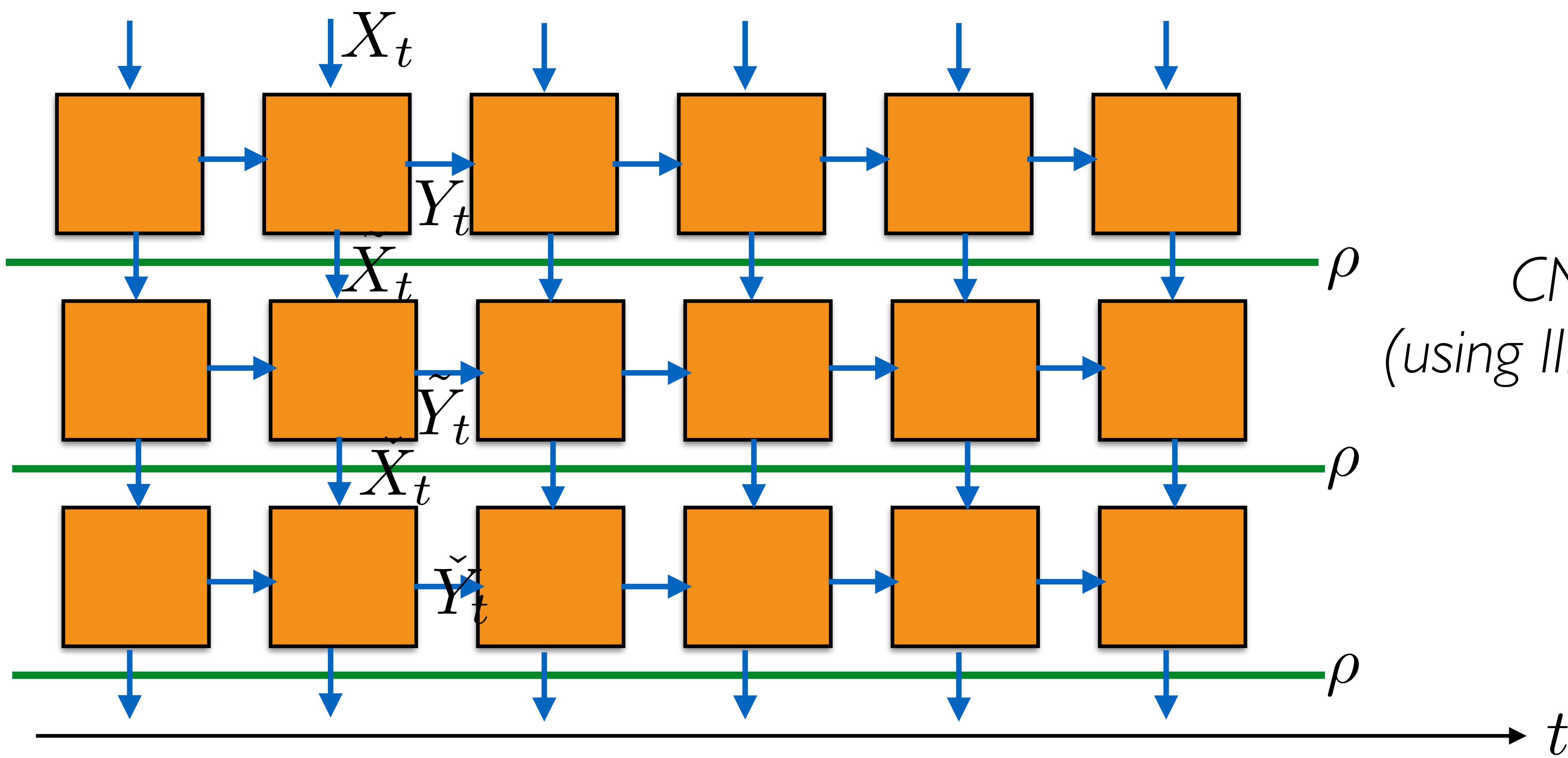
$$\hat{h}(e^{i\omega}) = \frac{1}{\sum_{j \leq p} a_j e^{ij\omega}} = \frac{1}{\bar{a} \prod_{j \leq p} (e^{iw} - z_j)}$$



RNNs and CNNs

- Multivariate IIR filters with multiple layers (with p=1):

$$\begin{cases} Y_t &= A_1 Y_{t-1} + B X_t \\ \tilde{X}_t &= \rho(Y_t) \\ \tilde{Y}_t &= \tilde{A}_1 \tilde{Y}_{t-1} + \tilde{B} \tilde{X}_t \\ \dots \end{cases}$$



CNN
(using IIR filters)

RNNs and CNNs

- RNN: Non-linear recurrence:

$$\begin{cases} Y_t &= \rho(A_1 Y_{t-1} + B X_t) \\ \tilde{X}_t &= C Y_t \\ \tilde{Y}_t &= \rho(\tilde{A}_1 \tilde{Y}_{t-1} + \tilde{B}\tilde{X}_t) \\ \check{X}_t &= \tilde{C}\tilde{Y}_t \\ \cdots \end{cases}$$