Feature Engineering for Graph-based Analysis of Recurrent Behavior in Biosignal Ensembles

Dr. George Tzagkarakis

Foundation for Research and Technology-Hellas
Institute of Computer Science

XAI-TS Workshop 2023
Turin, 18 September 2023

"More data beats clever algorithms,
*but **better data** beats more data"*

Peter Norvig
Director of Research @ Google

"Coming up with features is difficult, time-consuming
*requires **expert knowledge**.*
"Applied ML" is basically feature engineering"

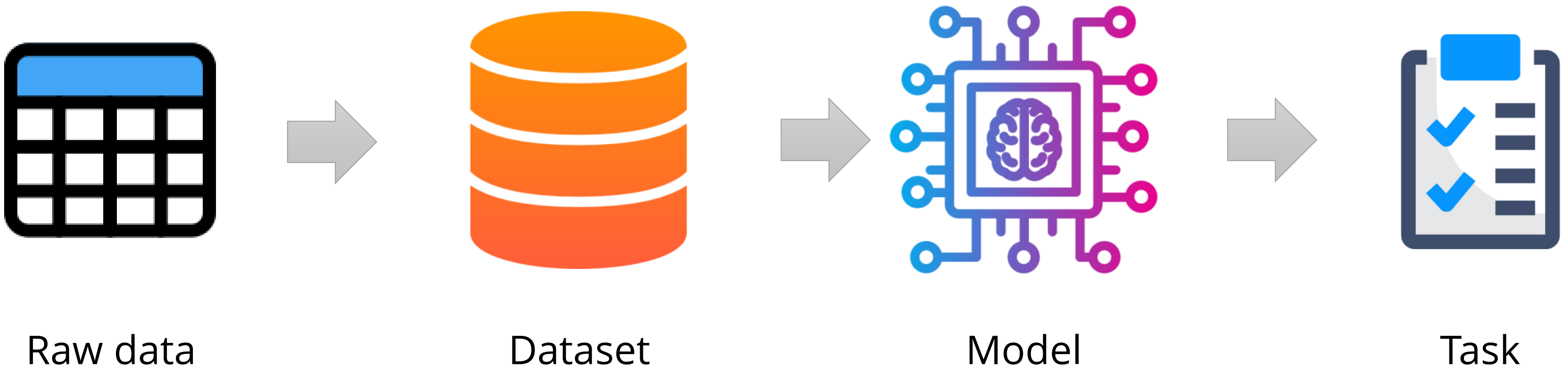
Andrew Ng
Founder of DeepLearning.AI

Outline

- Machine Learning workflow
- Feature Engineering
- Biosignal Analysis: Domain knowledge
  - Graph representations
- Target Application #1: Detection of epileptic seizures (:: exploit statistics)
  - Alpha-stable models
  - Graph filtering
- Target Application #2: Classification of NPSLE patients (:: exploit recurrence)
  - Recurrence quantification analysis
- Conclusions

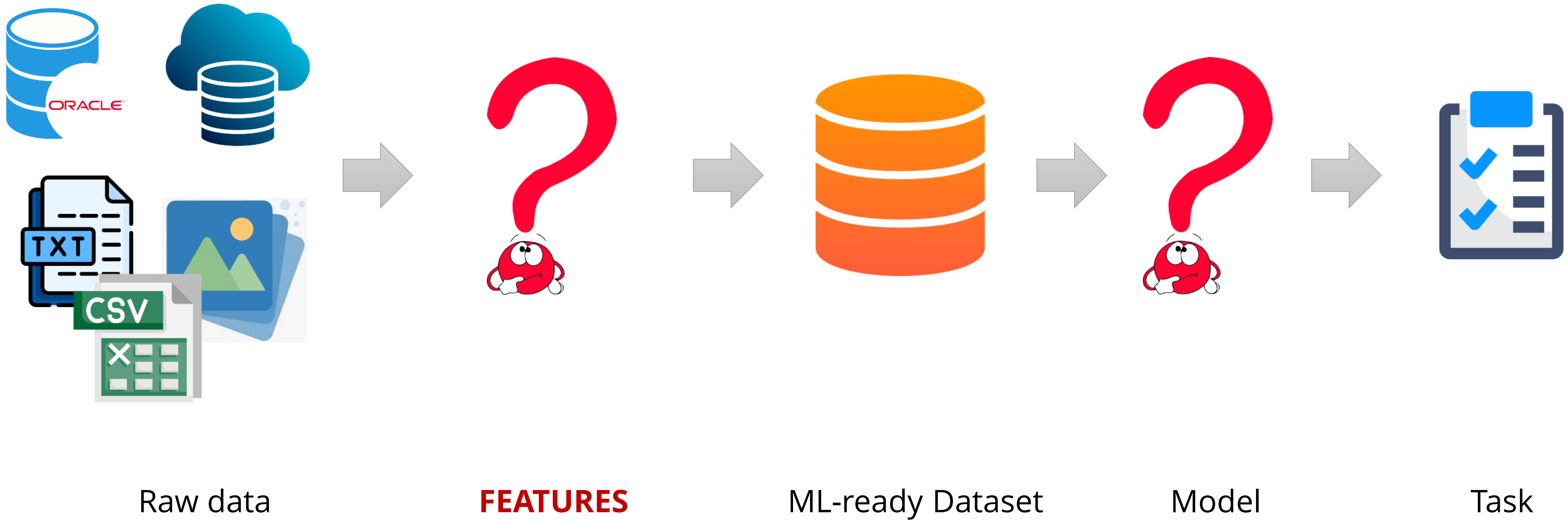
ML workflow

The dream…



ML workflow

The reality…



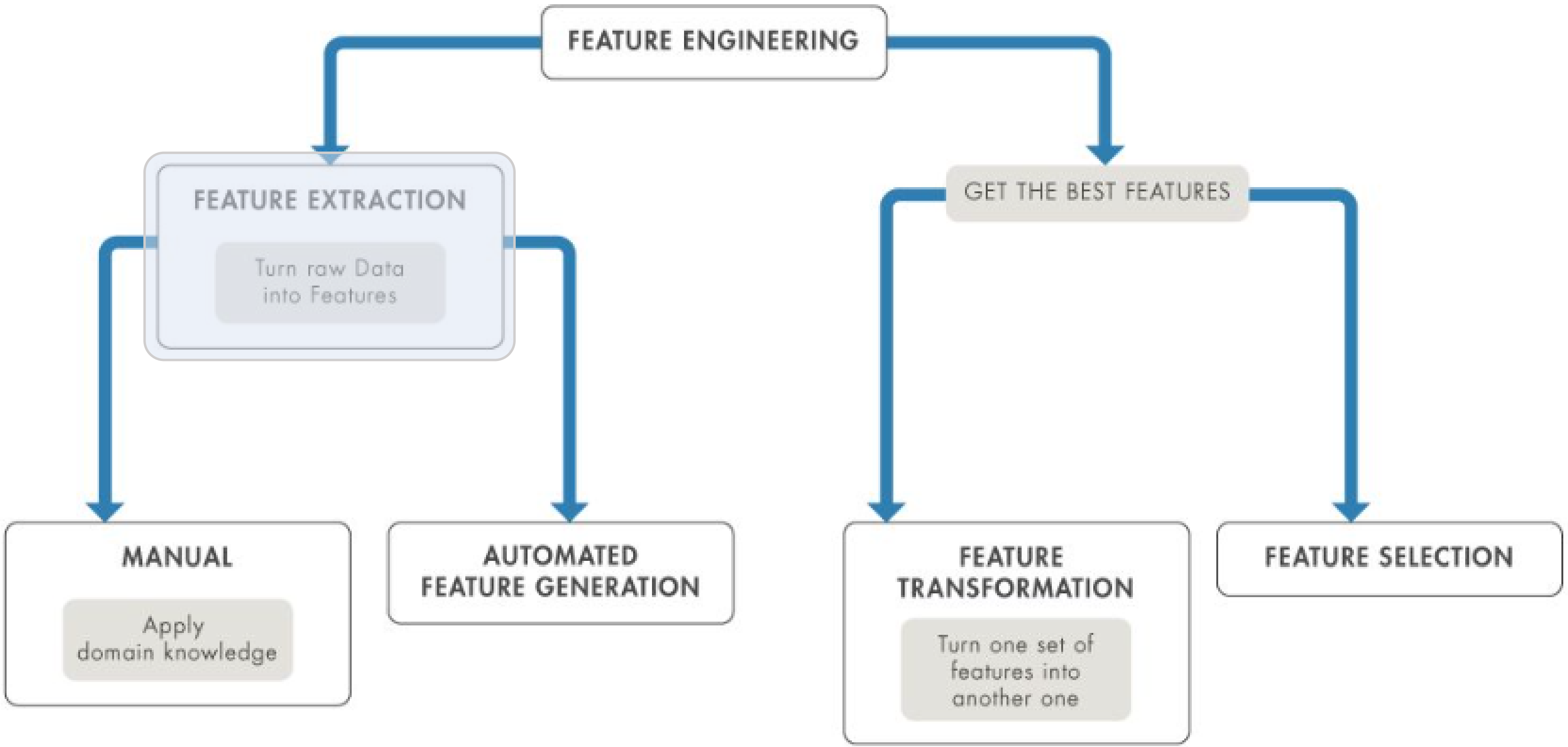
Feature engineering (FE)

"FE is the process of turning raw data into features that ***better represent*** *the underlying problem to the models, resulting in improved model accuracy on unseen data"*

- FE is difficult since extracting features from signals requires **deep domain knowledge**
- Finding the best features fundamentally remains an **iterative process**, even if we apply automated methods

Feature engineering (FE)

FE encompasses one or more of the following steps:



Feature engineering (FE)

Manual vs Automated Feature Extraction

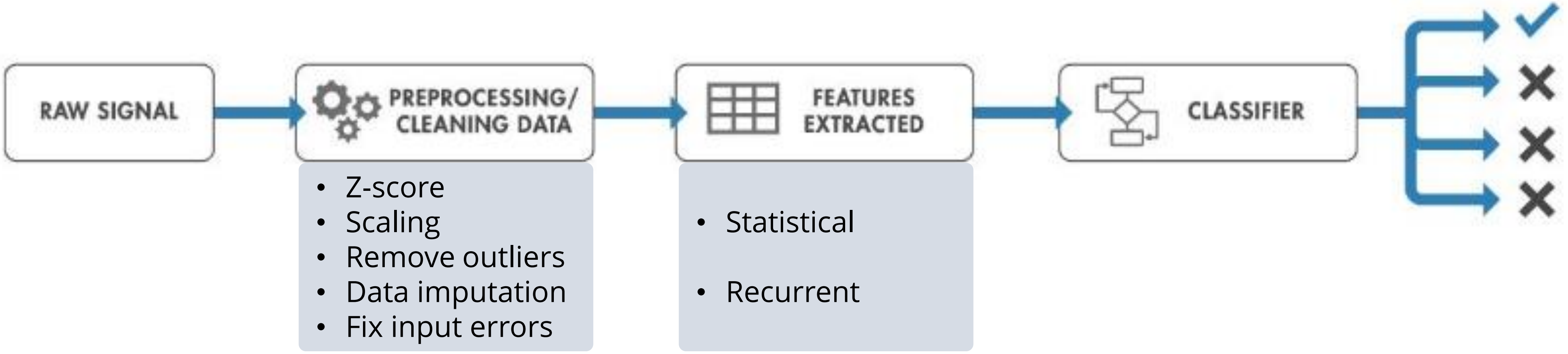
- **Manual:** generate features that are relevant for a given problem (e.g. mean of a signal window); Good understanding of the background or **domain** is a big plus
- **Automated:** use specialized algorithms or deep networks to extract features automatically from signals; Useful to move quickly from raw data to developing ML algorithms

Images vs Time Series Feature Extraction

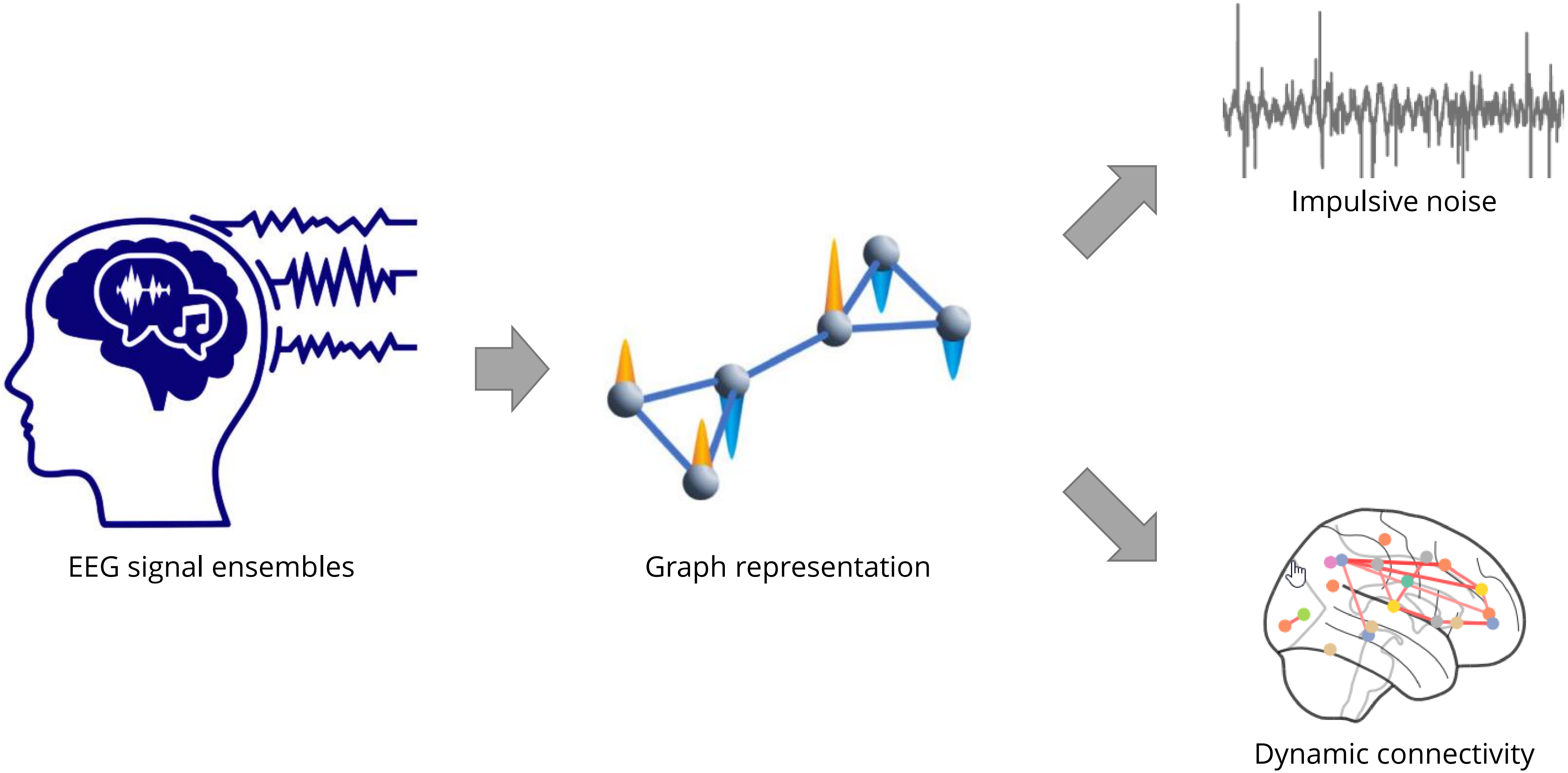
- **Images:** it has been largely replaced by the first layers of deep networks
- **Time series:** it remains the first **challenge** that requires significant expertise

Feature extraction for time series

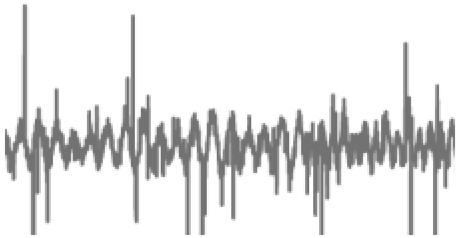
- Feature extraction identifies the most **discriminating characteristics** in signals, which a ML/DL algorithm can more easily consume
- ML/DL training directly with raw signals often yields poor results because of the **high data rate** and **information redundancy**



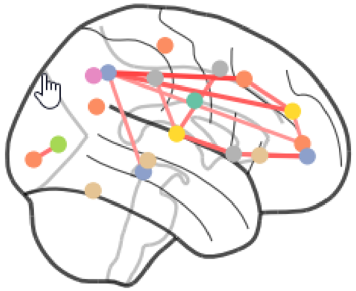
Biosignal Analysis: Domain knowledge



Biosignal Analysis: Domain knowledge



Impulsive noise



Dynamic connectivity

Alpha-Stable Graph Filtering

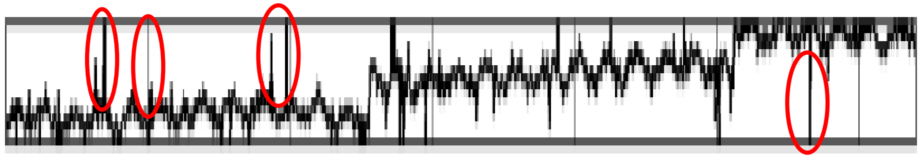
Recurrence Quantification Analysis

Target App #1
Detection of epileptic seizures

[Joint work w/ Dr. Anastasia Pentari]

Impulsive noise

- EEG signals are often corrupted by **impulsive** noise (e.g. electronic equipment, subject motion, etc.) (ref. [1])



- Denoising is a critical issue; significant "peaky" patterns must be preserved
- Noise part is often characterized by non-Gaussian (**heavy-tailed**) statistics
- Existing methods:
  - Per-signal filtering (Wavelet-based, ICA-based, etc.)
  - Better adapt to Gaussian noise statistics
- Exploit graph structure of EEG signal to account for intra-/inter-channel dependencies ⇨ **Graph filters**
  - L2-based formulation (2^{nd} order moments)

Graph representation of EEG signals

- Signal model w/ additive observation noise

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{N \times K}$$

Data matrix: $\mathbf{X}$

i-th electrode's signal: $\mathbf{x}_i$

electrodes: N

samples/electrode: K

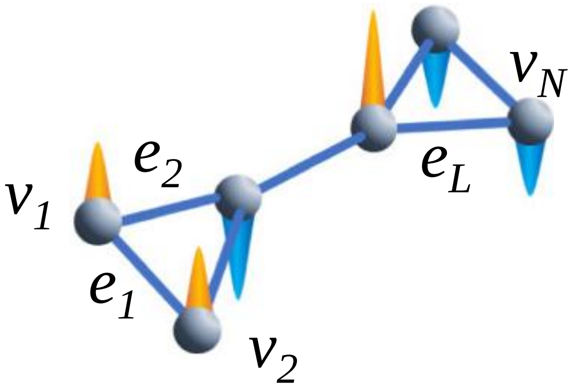
$$\mathbf{X} = \mathbf{S} + \mathbf{W}$$

Noiseless data matrix: $\mathbf{S}$

Noise term: $\mathbf{W}$

Graph representation of EEG signals

- Graph representation



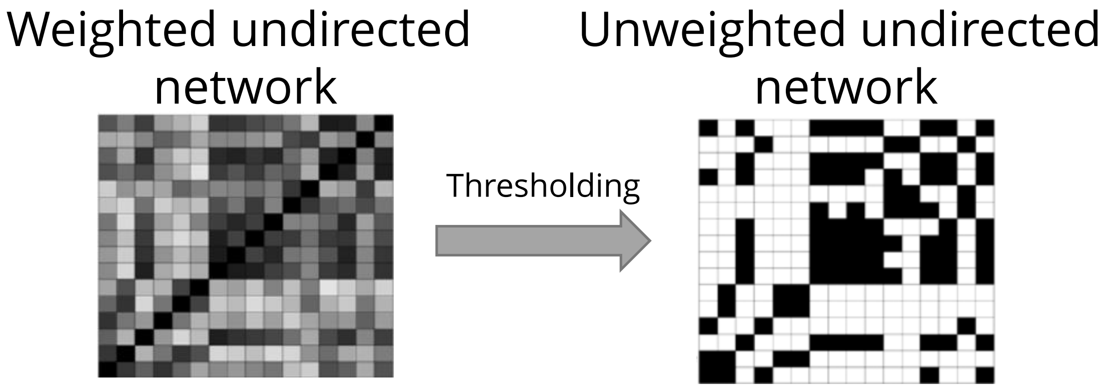
$$\mathbf{X} = \mathbf{S} + \mathbf{W}$$

$$G = (V, E)$$

Adjacency matrix
(capture interrelations among the nodes)

$$\mathbf{A} \in \mathbb{R}^{N \times N}$$

- Correlation matrix $\mathbf{A}$ (descriptor of brain's functional connectivity)



Alpha-stable models

- Powerful tool in accurately modelling impulsive phenomena (e.g. medical imaging, communications, finance, etc.)
- Lack of closed-form expressions for the pdf (except for Gaussian, Cauchy and Lévy)
- Modelling signal statistics via **symmetric alpha-stable** (SaS) distributions

$$\mathbf{X} = \mathbf{S} + \mathbf{W}$$

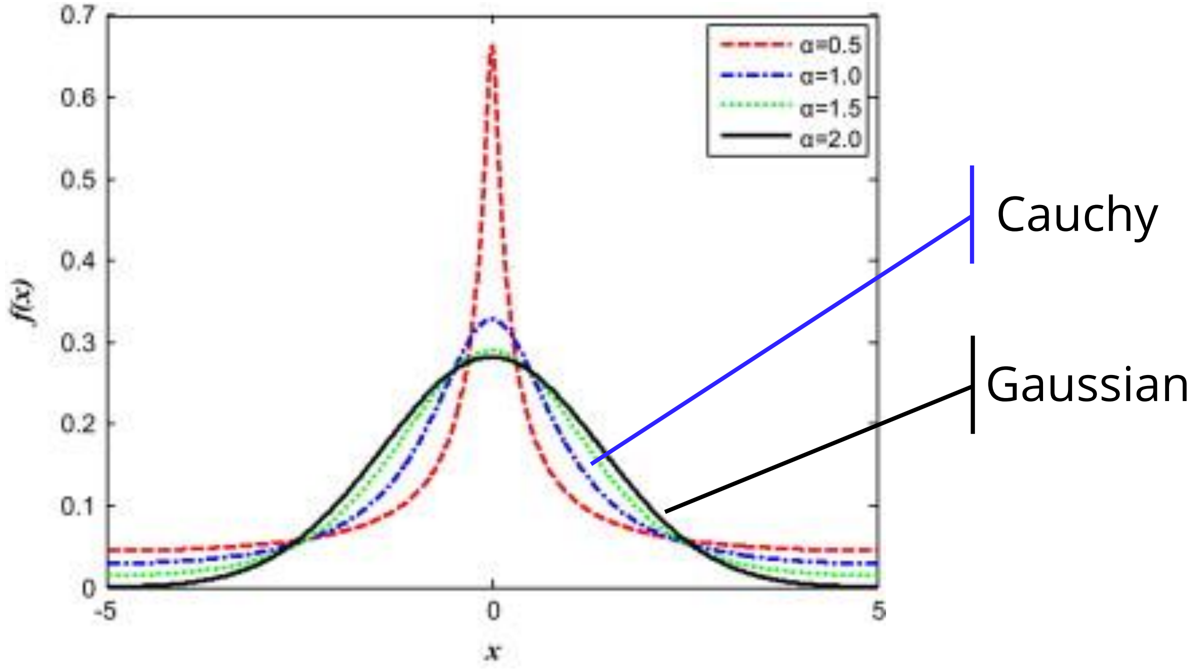
$$f_\alpha(x;\gamma,\delta) = \frac{1}{\gamma} h\left(\frac{x-\delta}{\gamma};\alpha\right)$$

$$h(x;\alpha) = \frac{1}{\pi}\int_0^\infty \cos(xt)e^{-t^\alpha}\,\mathrm{d}t$$

Model parameters	
$\alpha \in (0,2]$	characteristic exponent
$\gamma > 0$	dispersion
$\delta \in \mathbb{R}$	location

Alpha-stable models

- Examples of SaS distributions



- Max Likelihood estimation of model parameters; reliable, tightest possible confidence intervals (ref. [2])

Alpha-stable models

- All moments of order $p < \alpha$ exist; **Fractional lower order moments** (FLOMs)

$$X \sim f_\alpha(\gamma, \delta = 0) \quad \Rightarrow \quad \mathbb{E}\{|X|^p\} = \left(C(p,\alpha)\cdot\gamma\right)^p, \quad 0 < p < \alpha$$

$$\left(C(p,\alpha)\right)^p = \frac{\Gamma\left(1-\frac{p}{\alpha}\right)}{\cos\left(\frac{\pi}{2}p\right)\Gamma(1-p)}$$

- Quantify degree of dependence between two SaS variables X, Y, via the **covariation** (analogue of covariance)

$$[X,Y]_\alpha = \frac{\mathbb{E}\{XY^{<p-1>}\}}{\mathbb{E}\{|Y|^p\}}\gamma_Y^\alpha$$

$$z^{<a>} = |z|^a \mathrm{sign}(z)$$

Discrete case: FLOM-based covariation estimator

$$c_{ij}^{\mathrm{FLOM}} = \frac{\sum_{k=1}^K x_{i,k}|x_{j,k}|^{p-1}\mathrm{sign}(x_{j,k})}{\sum_{k=1}^K |x_{j,k}|^p}\gamma_{\mathbf{x}_j}^\alpha, \quad \mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^K$$

Alpha-stable models

- FLOM-based adjacency matrix (non-negative, symmetric)

$$\mathbf{A} = \frac{|\mathbf{C}_{\text{FLOM}}| + |\mathbf{C}_{\text{FLOM}}^T|}{2}$$

- **Note**: selection of an appropriate FLOM order, p, is critical towards better adapting to the underlying degree of impulsiveness
- Calculate p as a function of α, by minimizing the std of the FLOM-based covariation estimator (ref. [3])
  - **Convention**: the optimal p value is the mean over all channels

α	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
p_{opt}	0.52	0.56	0.58	0.61	0.64	0.69	0.72	0.76	0.81	0.88	0.98

$p \lesssim \alpha/2$

Graph filtering as Lp-regularized optimization

- Existing formulation: L2-optimization regularized by graph total variation (ref. [4])

$$\hat{\mathbf{S}} = \underset{\mathbf{S}\in\mathbb{R}^{N\times K}}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\mathbf{S}-\mathbf{X}\|_2^2 + \frac{1}{2}b\|\mathbf{S}-\mathbf{AS}\|_2^2 \right\}$$

Data fidelity term — Smoothness term

$$\hat{\mathbf{S}} = (\mathbf{I} + b(\mathbf{I}-\mathbf{A})^H(\mathbf{I}-\mathbf{A}))^{-1}\mathbf{X}$$

- Limitation: 2nd order moments inappropriate for SaS models

Graph filtering as Lp-regularized optimization

- **Our scope**: Suppress the effects of heavy-tailed impulsive noise ⇨ Employ Lp (quasi)norms ($p < 2$); direct relation with FLOMs

$$\mathbb{E}\left\{|X|^p\right\} \simeq \lim_{N\to\infty} \frac{1}{N}\sum_{t=1}^{N} |x_t|^p = \lim_{N\to\infty} \frac{1}{N}\|\mathbf{x}\|_p^p$$

$$\hat{\mathbf{S}} = \underset{\mathbf{S}\in\mathbb{R}^{N\times K}}{\operatorname{argmin}} \left\{ \frac{1}{2}\|\mathbf{S}-\mathbf{X}\|_p^p + \frac{1}{2}b\|\mathbf{S}-\mathbf{A}\mathbf{S}\|_p^p \right\}$$

Data fidelity term

Smoothness term

- Challenge: Highly non-convex problem; singularity of gradient

Graph filtering as Lp-regularized optimization

- Use Lp,ϵ approximation of Lp (quasi)norm $||\mathbf{x}||_{p,\epsilon}^{p} = \sum_{j=1}^{N} \left(|x_j|^2 + \epsilon\right)^{p/2}$

$$\hat{\mathbf{S}} = \underset{\mathbf{S}\in\mathbb{R}^{N\times K}}{\operatorname{argmin}} \underbrace{\left\{\frac{1}{2}||\mathbf{S}-\mathbf{X}||_{p,\epsilon}^{p} + \frac{1}{2}b||\mathbf{S}-\mathbf{AS}||_{p,\epsilon}^{p}\right\}}_{Q_3}$$

- **Final implementation**: joint iterative reweighted least squares (IRLS) & Lp,ϵ
  - Better preserves both low- and high-amplitude EEG samples

Graph filtering as Lp-regularized optimization

Inputs: $\mathbf{X}$, $\mathbf{A}$, I_{max}, b, ε

Outputs: $\hat{\mathbf{S}}$

Initialization: $\mathbf{S}^{(0)} = \mathbf{X} + \varepsilon$

for t = 1:I_{max} **do**

  for k = 1:K **do**

    $\mathbf{r} = \mathbf{S}(:,k)^{(t)} - \mathbf{X}(:,k)$

    $\mathbf{w} = |\mathbf{r} + \varepsilon|^{(p-2)/2}$

    $\mathbf{D} = \mathrm{diag}(\mathbf{w})$

IRLS $\mathbf{S}(:,k)^{(t)} = \left((\mathbf{D}^2 + b(\mathbf{I} - \mathbf{A})^H(\mathbf{I} - \mathbf{A}))^{-1}\mathbf{D}^2\right)\mathbf{X}(:,k)$

  end for

Lp,ϵ $\mathbf{S}^{(t+1)} = \mathbf{S}^{(t)} + \left(\nabla^2 Q_3(\mathbf{S}^{(t)})\right)^{-1} \nabla Q_3(\mathbf{S}^{(t)})$

end for

$\hat{\mathbf{S}} = \mathbf{S}^{(Imax)}$

Experimental evaluation

- **Test case 1**: Synthetic noise of varying impulsiveness added to "noise-free" EEG ensembles
  - 32-channel EEGs
  - $\alpha \in \{1.1 : 0.3 : 2\}$, $\gamma = 1$
  - 100 Monte Carlo runs; results averaged over all channels and MC runs
  - Performance metrics

$$\mathrm{SER}(\mathbf{s}, \hat{\mathbf{s}}) = 10 \log_{10}\left(\frac{\|\mathbf{s}\|_2^2}{\|\mathbf{s}-\hat{\mathbf{s}}\|_2^2}\right)$$

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

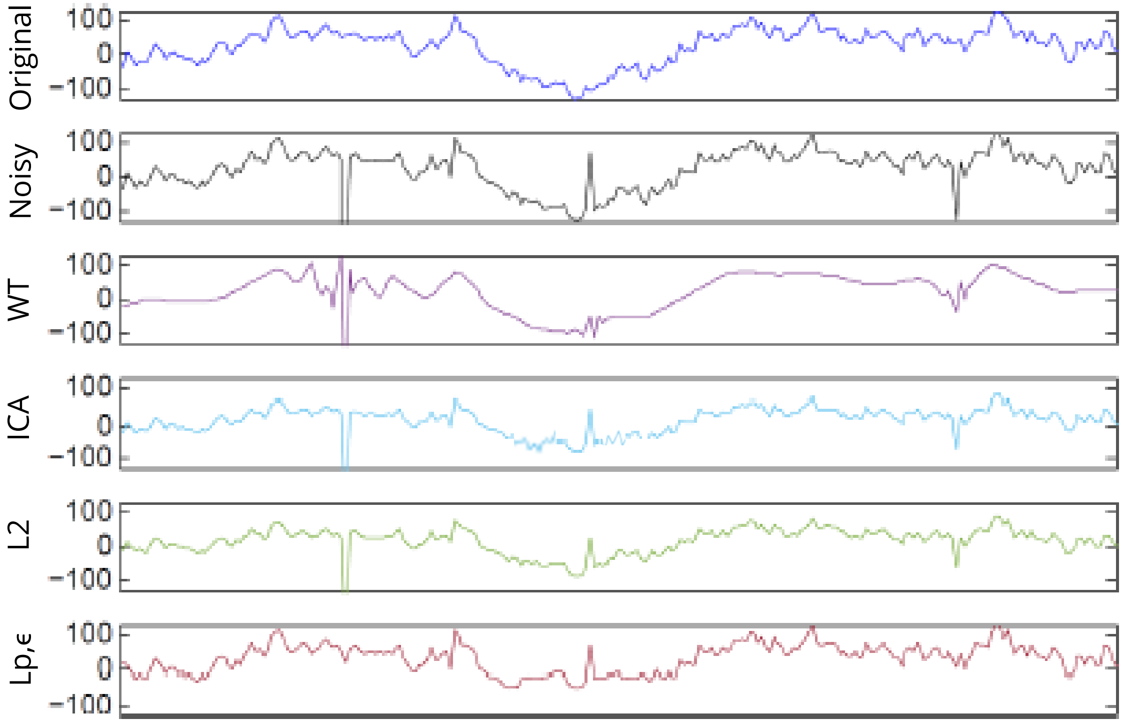
$$c_1 = (k_1 L)^2$$

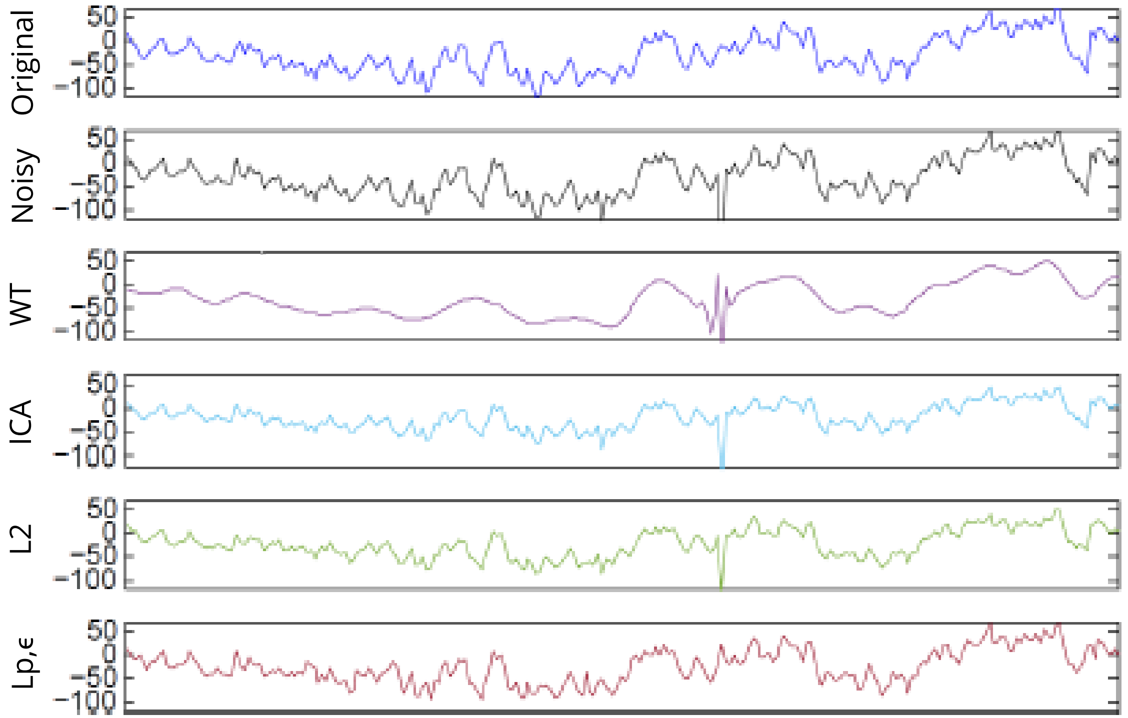
$$c_2 = (k_2 L)^2$$

$$L = 100 \quad , \quad k_1 = k_2 = 0.05$$

Experimental evaluation

- **Test case 1**: Synthetic noise of varying impulsiveness added to "noise-free" EEG ensembles





Experimental evaluation

- **Test case 2**: Filtering of raw data followed by a **classification** task
  - 32-channel EEGs; ground truth (noise-free) signals are unknown
  - 15 “normal” subjects (non-epileptic), 15 “abnormal” subjects (epileptic)
  - Non-overlapping windows of length 256 are selected for each subject
  - 3 filtering methods: “WT”, “L2”, “Lp,ϵ”
  - 5 features: {mean, std, mean($\pm 10\% \cdot$ max_ampl), min_ampl, max_ampl}
  - kNN classifier (2/3 training, 1/3 testing subjects); majority voting to identify the dominant class of each testing subject’s windows

Model parameters

WT	L2	Lp,ϵ
db8	b = 0.01	b = 0.1
2-level		ε = 0.01
		I_{max} = 10

Experimental evaluation

- **Test case 2**: Filtering of raw data followed by a **classification** task

Classification accuracy

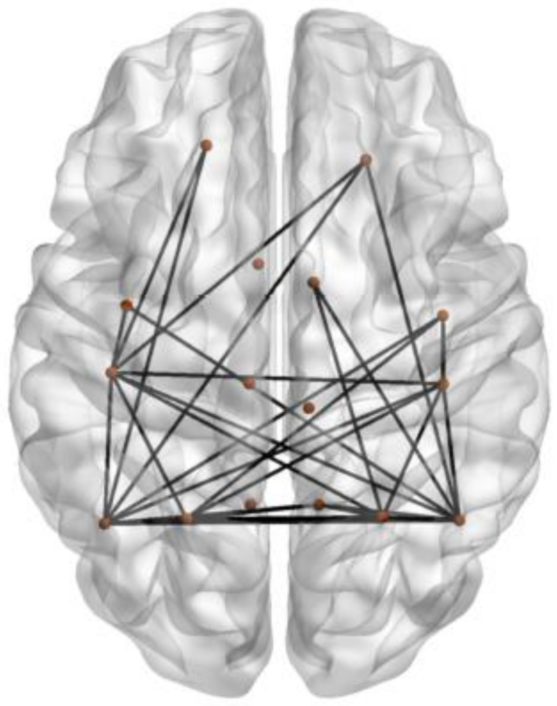
A	Original	WT	L2	Lp,ϵ
-	50%	50%		
Correlation			60%	
Covariation				**90%**

Target App #2
Classification of NPSLE patients

[Joint work w/ Prof. Akis Simos (Evolutionary Neuropsychology) and Dr. Anastasia Pentari]

Brain functional connectivity

- Brain function is highly dynamic ⇨ Assessing functional associations in neurophysiological activity between two or more brain regions is challenging
- Associations as evidence of **functional connectivity** (FC) between regions
- fMRI studies: measure temporal variations in brain activity
- **Limitations**: non-stationary behavior of brain signals is not accounted for; predictability of the signal recorded in one brain region is not captured from the signal recorded in other regions
- **Hypothesis**: brain signals are characterized by the phenomenon of **recurrence**, i.e., similar situations of a dynamic system should evolve in a similar manner



Clinical test case

- Neuropsychiatric systemic lupus erythematosus (NPSLE): disorder that is characterized by a variety of neuropsychiatric symptoms in the absence of remarkable brain injuries
- **Idea**: capture **recurrent** dynamic characteristics of rs-fMRI time series, which could serve as complementary indices of functional connectivity
- We adopt a Region-of-Interest (ROI) approach focusing on abnormal dynamic connectivity between 16 frontoparietal regions (eight in each hemisphere)
- Participants: 45 patients diagnosed with NPSLE (Rheumatology outpatient clinic, University Hospital of Heraklion), 35 age-matched and gender-matched healthy volunteers

Recurrence quantification analysis

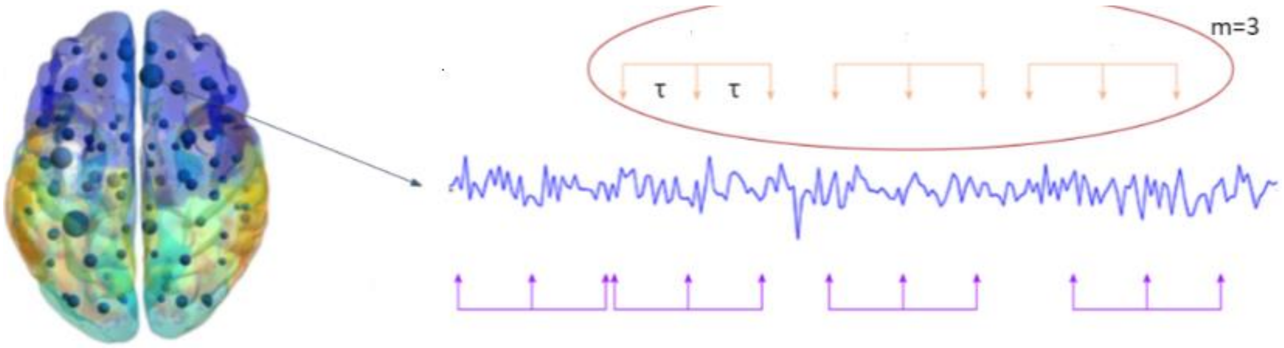
Develop advanced techniques for extracting and characterizing the inherent complex dynamic structures apparent in scientific data

Recurrence is a fundamental feature of nonlinear dynamical systems
It is a time the trajectory returns to a location it has visited before

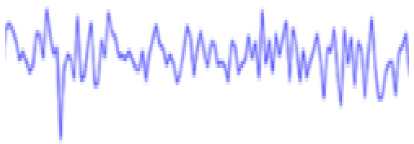
Visualize and analyze recurrences to understand and characterize the dynamics of complex nonlinear systems

Recurrence quantification analysis

- **Recurrence Plot (RP)**
  - Depicts the (local) neighborhood structure
  - Captures time indices at which phase space trajectories return to a neighborhood
  - Visualize recurrences based on a binary recurrence matrix



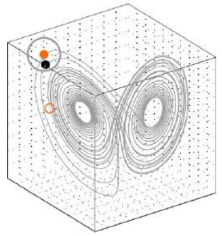
Original time series



$\mathbf{r} = \{r_i\}_{i=1}^{n}$

→ Time-Delay Embedding →

States (Phase Space)

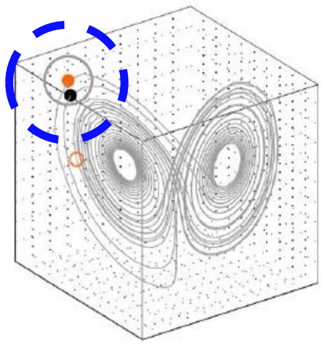


$\mathbf{x}_i = [r_i, r_{i+\tau}, \ldots, r_{i+(m-1)\tau}] \quad i = 1, \ldots, N$

Recurrence quantification analysis

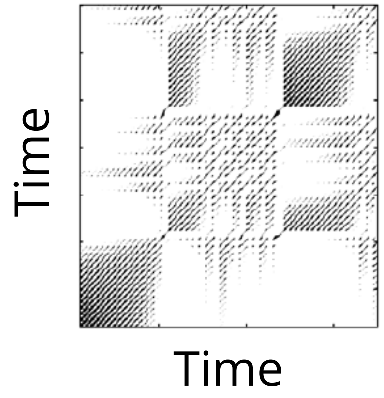
- **Recurrence Plot (RP)**

States (Phase Space)



$$\mathbf{x}_i = [r_i, r_{i+\tau}, \ldots, r_{i+(m-1)\tau}]$$

Binary Recurrence Matrix



$$\mathbf{R}_{i,j} = \Theta\left(\varepsilon - d(\mathbf{x}_i, \mathbf{x}_j)\right) \quad i, j = 1, \ldots, N$$

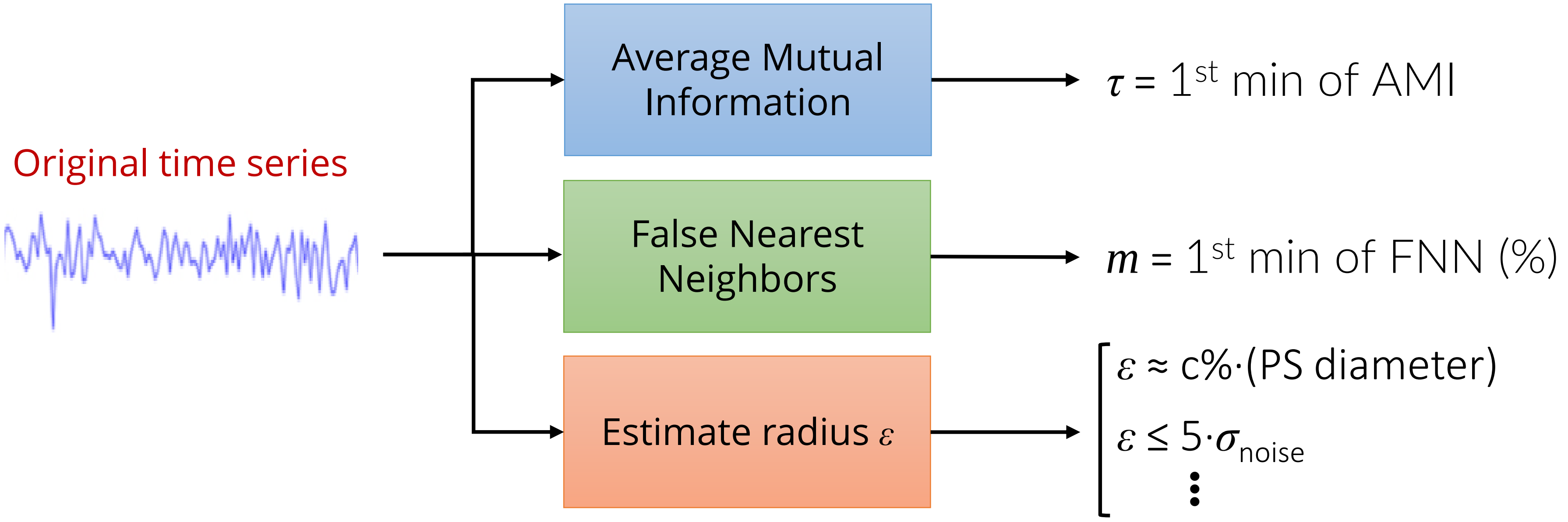
Critical Parameters	
m	Embedding dimension
τ	Delay
N (= n-(m-1)τ)	Number of states

$$\Theta(n) = \begin{cases} 1, & \text{if } n \geq 0 \\ 0, & \text{if } n < 0 \end{cases} \longrightarrow \mathbf{R}_{i,j}(\varepsilon) = \begin{cases} 1, & d(\mathbf{x}_i, \mathbf{x}_j) \leq \varepsilon \quad ● \\ 0, & \text{otherwise} \quad ○ \end{cases}$$

Recurrence quantification analysis

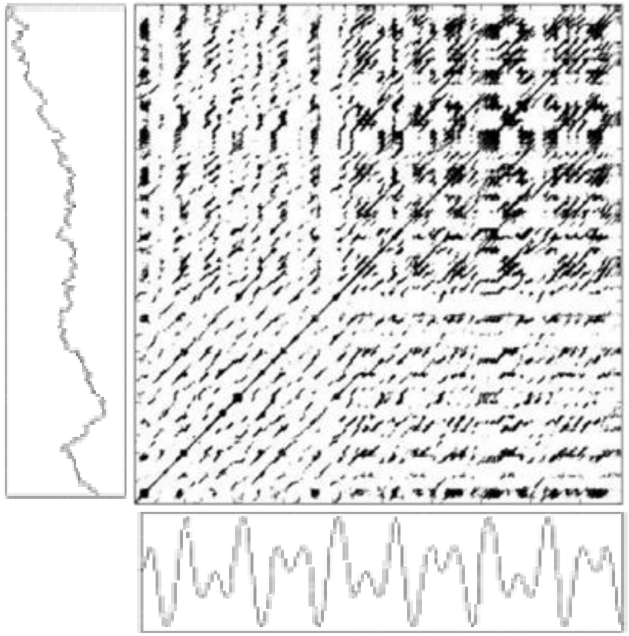
- Selection of embedding parameters

Inaccurate choice of m, τ, ε can hinder the discovery of low-dimensional dynamics or produce false positive indication of chaotic structure



Cross recurrence plot (CRP)

- Study and quantify the interaction of two distinct systems due to coupling
- Estimate time-synchronization profiles and detect co-movements
- Analyze dependencies between two distinct systems using CRP; Visualize times at which a state in one system occurs simultaneously in the second



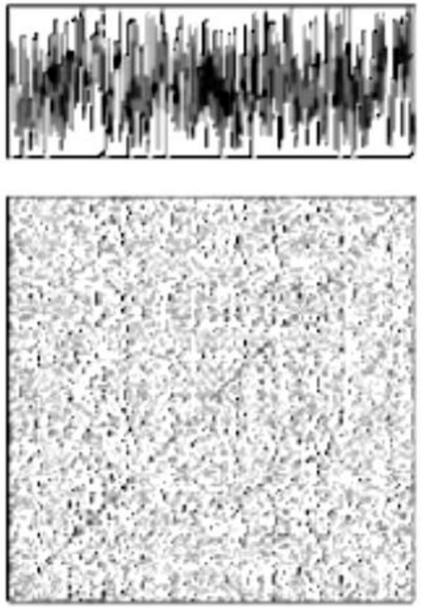
$$\mathbf{CR}_{i,j}(\varepsilon) = \Theta\left(\varepsilon - \|\mathbf{x}_i - \mathbf{y}_j\|_p\right) \qquad i = 1, \ldots, N, \quad j = 1, \ldots, M$$

CR is not necessarily square (in general $N \neq M$)

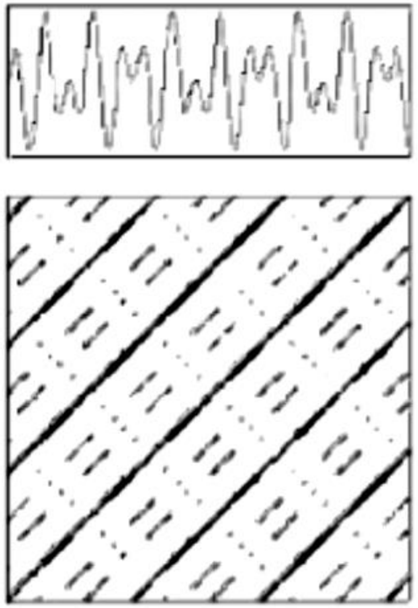
If embedding parameters differ, use the ***higher embedding***

Properties of (C)RP

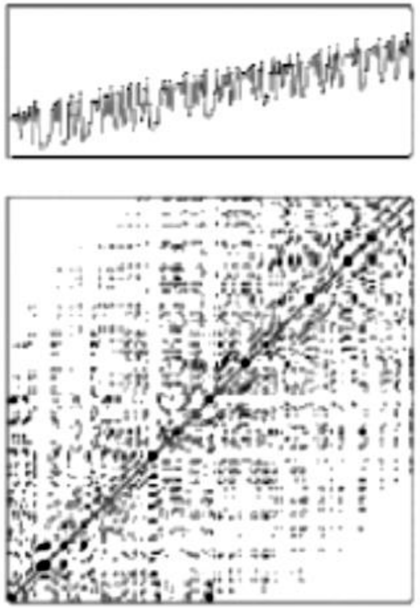
- Several linear and curvilinear structures appear in (C)RPs; Give hints about the time evolution of the high-dimensional phase space trajectories
- (C)RPs applied on short and non-stationary data
- Large-scale and small-scale structures



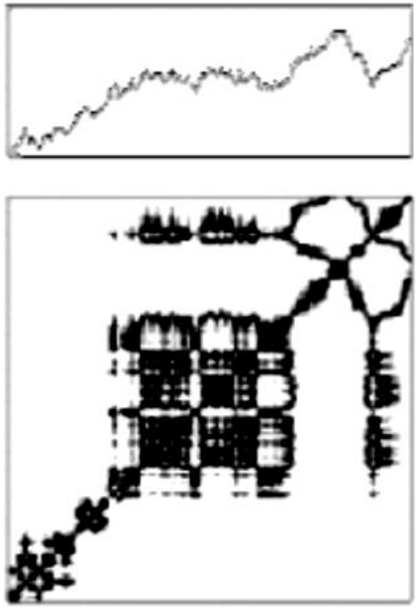
White noise
(uncorrelated data)



Harmonic oscillation
(2 frequencies)



Logistic map
(chaotic data + linear trend)



AR process

Recurrence quantification analysis (RQA)

- Enhance the visual interpretation of (C)RPs by quantifying small-scale patterns with appropriate measures of complexity
- (C)RQA: nonlinear data analysis method, which quantifies the number and duration of recurrences of a (pair of) dynamical system(s) occurring in its (their) phase space trajectory
- (C)RQA measures: categorized according to the structures they are based on (e.g. diagonal, vertical lines); depend on threshold ε

RQA measures

Based on recurrence density

Recurrence Rate (**RR**)	RR for CRP (**CC**)
$$RR(\varepsilon) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{R}_{i,j}(\varepsilon)$$	$$CC(\varepsilon) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{CR}_{i,j}(\varepsilon)$$

RQA measures

Based on diagonal lines

Histogram of diagonal lines of length l : $P(\varepsilon, l) = \sum_{i=1}^{N} \sum_{j=1}^{N} (1 - \mathbf{R}_{i-1,j-1}(\varepsilon)) (1 - \mathbf{R}_{i+l,j+l}(\varepsilon)) \prod_{k=0}^{l-1} \mathbf{R}_{i+k,j+k}(\varepsilon)$

Determinism (**DET**)

$$DET = \frac{\sum_{l=l_{min}}^{N} l\,P(l)}{\sum_{l=1}^{N} l\,P(l)}$$

Measure of determinism (or predictability) of a system

Avg diagonal length (**L**)

$$L = \frac{\sum_{l=l_{min}}^{N} l\,P(l)}{\sum_{l=l_{min}}^{N} P(l)}$$

Measures average time that two trajectory segments are close (mean prediction time)

Longest diagonal (**L_{max}**)

$$L_{max} = \max_{i=1,\dots,N_l} \{l_i\}$$

The faster the trajectory segments diverge, the shorter are the diagonal lines, thus L_{max}

Divergence (**DIV**)

$$DIV = \frac{1}{L_{max}}$$

The faster the trajectory segments diverge, the higher is the value of DIV

RQA measures

Based on diagonal lines

Histogram of diagonal lines of length l : $P(\varepsilon, l) = \sum_{i=1}^{N} \sum_{j=1}^{N} (1 - \mathbf{R}_{i-1,j-1}(\varepsilon)) (1 - \mathbf{R}_{i+l,j+l}(\varepsilon)) \prod_{k=0}^{l-1} \mathbf{R}_{i+k,j+k}(\varepsilon)$

Entropy (**ENTR**)

$$ENTR = -\sum_{l=l_{min}}^{N} p(l) \ln(p(l))$$

$= \frac{P(l)}{N_l}$

Measure of RP complexity w.r.t. diagonal lines (e.g. ENTR is low for uncorrelated processes)

Trend (**TREND**)

$$TREND = \frac{\sum_{t=1}^{\tilde{N}} \left(t - \frac{\tilde{N}}{2}\right) \left(RR_t - \overline{RR_t}\right)}{\sum_{t=1}^{\tilde{N}} \left(t - \frac{\tilde{N}}{2}\right)^2}$$

$$RR_t = \frac{1}{N-t} \sum_{i=1}^{N-t} \mathbf{R}_{i,i+t}$$

Measures non-stationarity in terms of recurrence point density of the diagonals parallel to LOI, as a function of time distance t between these diagonals and the LOI

CRQA-based feature extraction

- 6 CRQA measures (features) are selected: RR, DET, L, L_{max}, ENTR, V_{max}
- Relative sensitivity of CRQA features in differentiating NPSLE patients vs healthy volunteers is compared against:
  - Nodal RQA measures (per ROI)
  - Conventional static FC metric (zero-order Pearson correlation between two ROIs (total of 120 unique connections for 16 ROIs)

Classification performance

- SVM classifier

FE method	Precision	Recall	F1 score
CRQA-based FC	**0.98**	**0.94**	**0.96**
RQA-based FC	0.91	0.90	0.90
Static FC	0.74	0.78	0.76

Conclusions & Key remarks

- Feature Engineering unlocks **hidden insights**: most improvement will probably come from thinking carefully what we put into our models
- This can be (semi-)automated but still one of the true arts in Data Science
- No amount of complex modeling can compensate for poor-quality data; Feature engineering is the first line of defense to enhance **data quality**
- **Domain knowledge** is a powerful ally in practice
- Dimensionality reduction is vital; Feature engineering not only adds valuable attributes but also helps **reduce dimensionality** (it isn't just about computational efficiency; it's about reducing noise, overfitting, and making models more interpretable)

THANKS FOR WATCHING!

Get in Touch

gtzag@ics.forth.gr

N. Plastira 100, GR70013 Heraklion, Crete

+30 2810 391753

https://bit.ly/3VkUWYx

References

[1] S. Supriya *et al.*, “Analyzing EEG signal data for detection of epileptic seizure: Introducing weight on visibility graph with complex network feature,” *in Proc. 27th Australasian Database Conf.*, 2016.

[2] J. Nolan, “Numerical calculation of stable densities and distribution functions,” *Commun. Statist.-Stoch. Models*, vol. 13, pp. 759–774, 1997.

[3] G. Tzagkarakis *et al.*, “Compressive sensing using symmetric alpha-stable distributions for robust sparse signal reconstruction,” *IEEE Trans. Signal Process.*, vol. 67, no. 3, pp. 808–820, 2019.

[4] S. Chen *et al.*, “Signal denoising on graphs via graph filtering,” *IEEE Global Conf. Signal & Inf. Process.*, Atlanta, GA, pp. 872–876, 2014.

[5] E. A. Allen *et al.*, “Tracking whole-brain connectivity dynamics in the resting state,” *Cerebral Cortex*, vol. 24, pp. 663–676, 2014.

[6] S. S. Menon and K. A. Krishnamurthy, “Comparison of static and dynamic functional connectivities for identifying subjects and biological sex using intrinsic individual brain connectivity,” *Sci. Rep. 9*, 5729, 2019.

[7] M. Bianciardi *et al.*, “Model-free analysis of brain fMRI data by recurrence quantification,” *Neuroimage*, vol. 37, no. 2, pp. 489–503, 2007.

[8] N. Marwan and J. Kurths, “Cross recurrence plots and their applications,” *Math. Physics Research at the Cutting Edge*, pp. 101–139, 2004.