Analyzing time series using feature spaces

Ben Fulcher

NII Shonan Meeting

February 2018

‘Feature-Based Time-Series Analysis’, ***arXiv*** 1709.08055 (2017)

What is a time series?

Repeated measurements taken over time

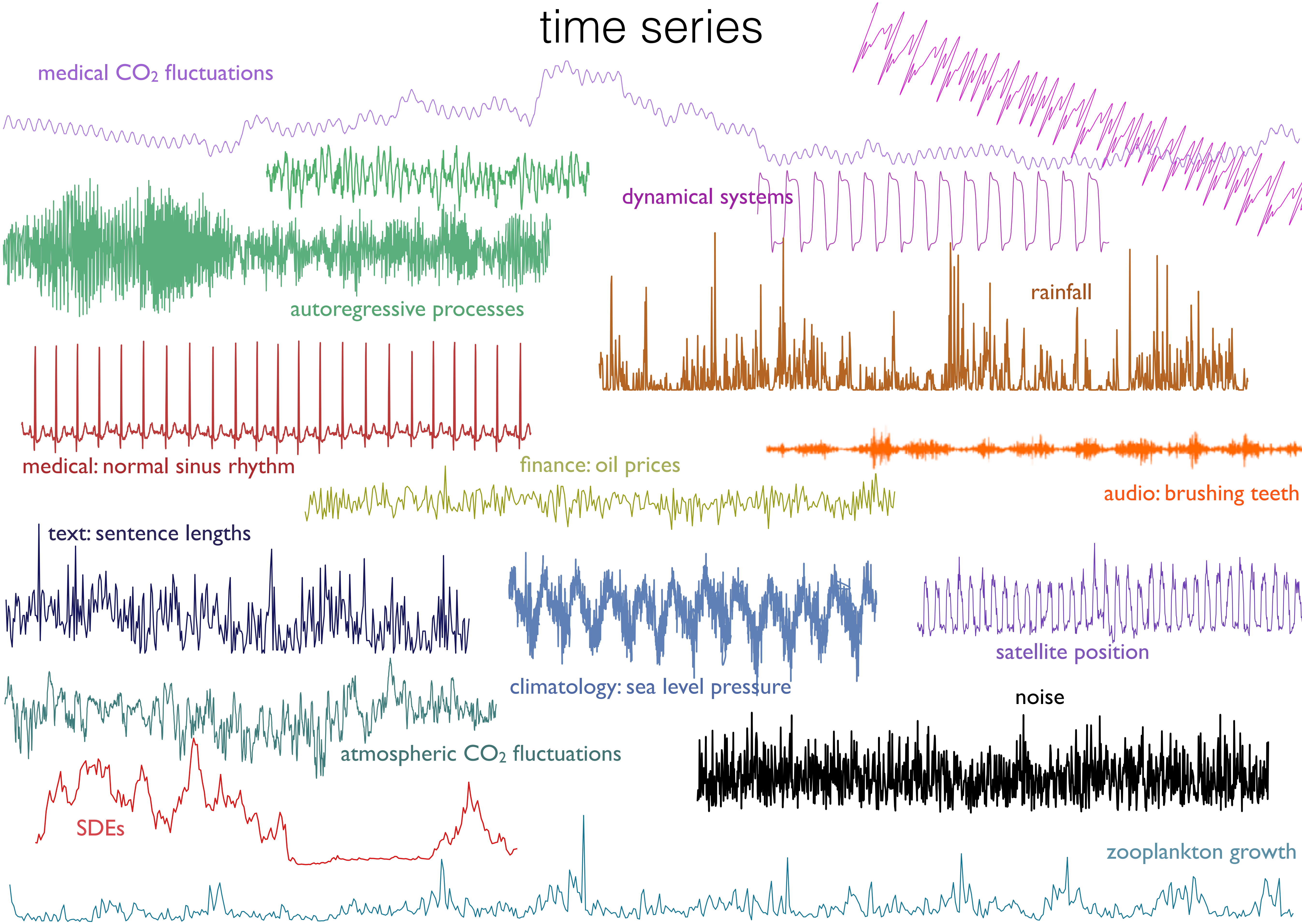
Often the sampling rate is fixed:

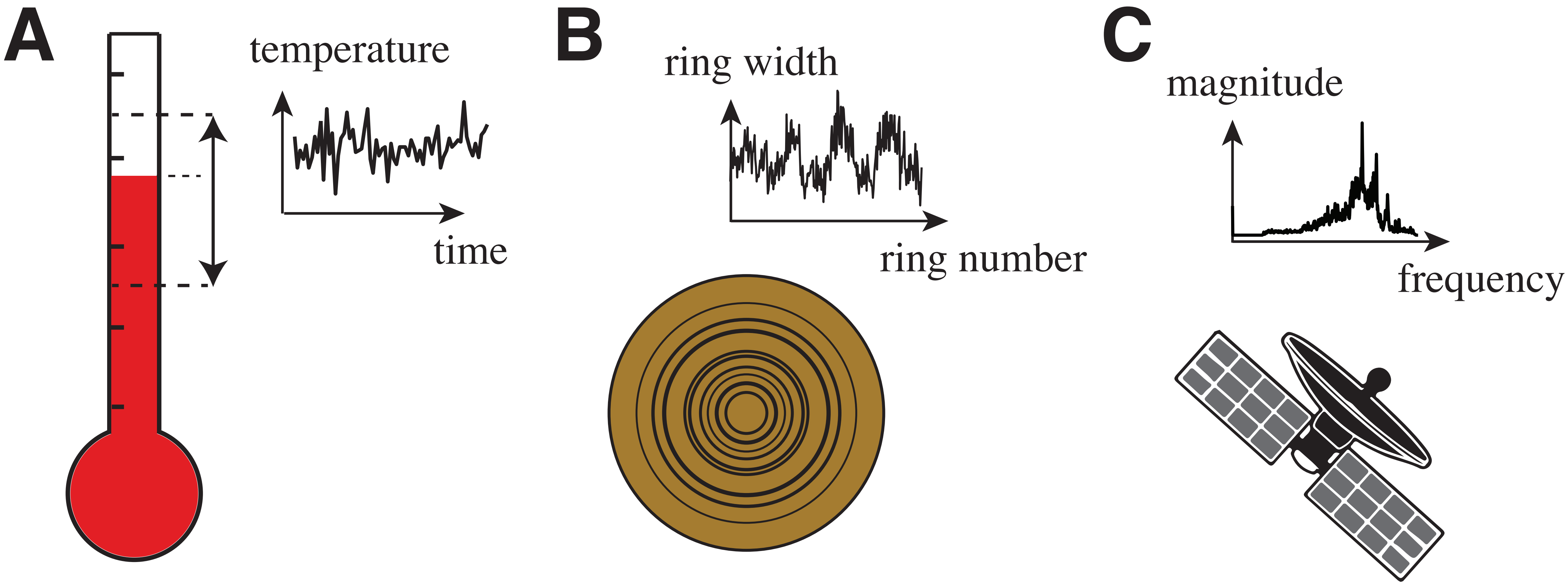
$$(x_1, x_2, ..., x_N)$$

$$(0, \Delta t, 2\Delta t, ..., (N-1)\Delta t)$$

- *seismology* (e.g., recordings of earthquake tremors)
- *biochemistry* (e.g., cell potential fluctuations)
- *biomedicine* (e.g., recordings of heart rate dynamics)
- *ecology* (e.g., animal population levels over time)
- *astrophysics* (e.g., radiation dynamics)
- *meteorology* (e.g., air pressure recordings)
- *economics* (e.g., inflation rates variations)
- *human-machine interfaces* (e.g., gesture recognition from accelerometer data)
- *industry* (e.g., quality control sensor measurements on a production line)

Many of our measurements of the world are in the form of time series

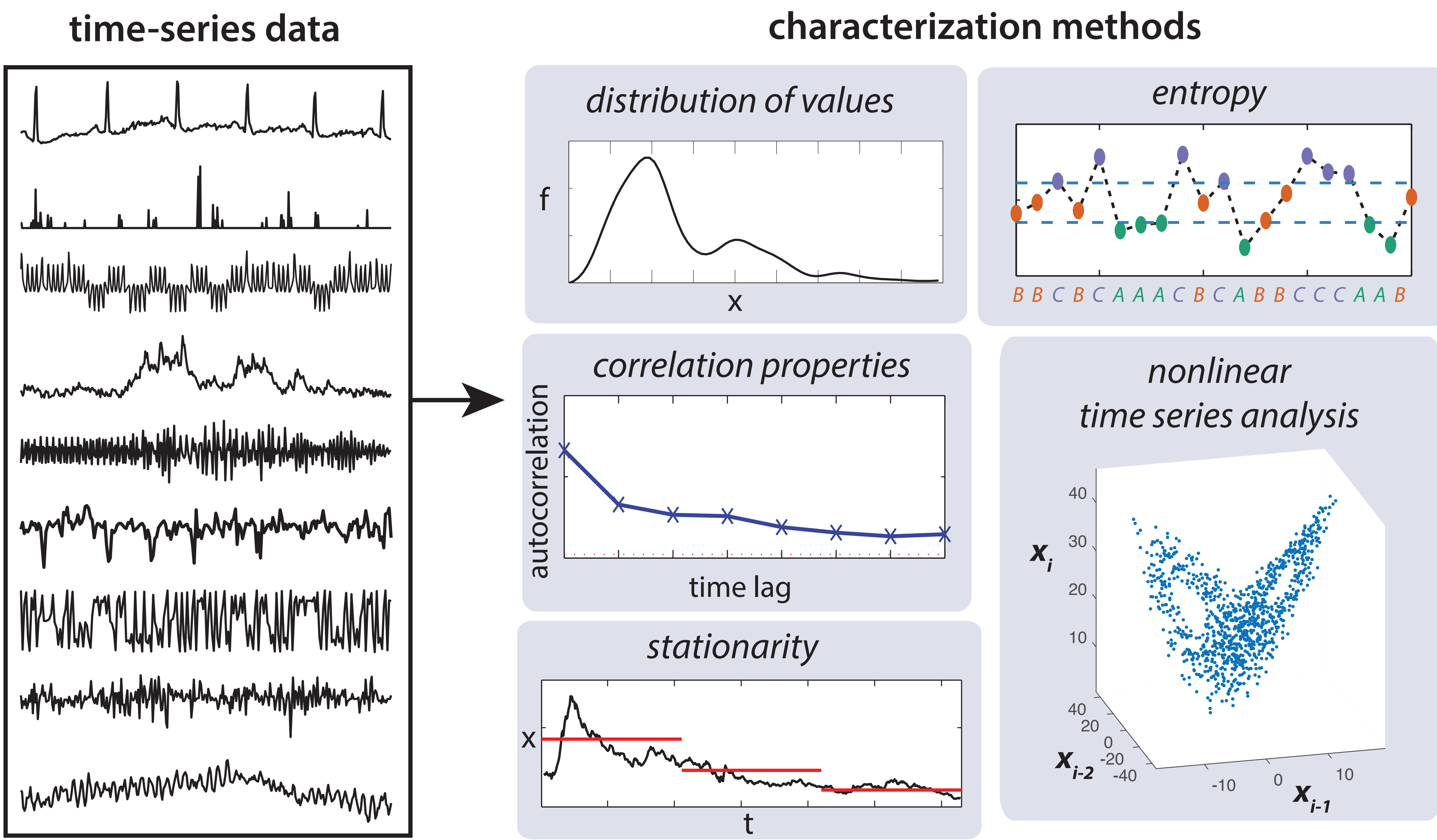




Focus on *univariate time series* where a simple quantity is measured over time

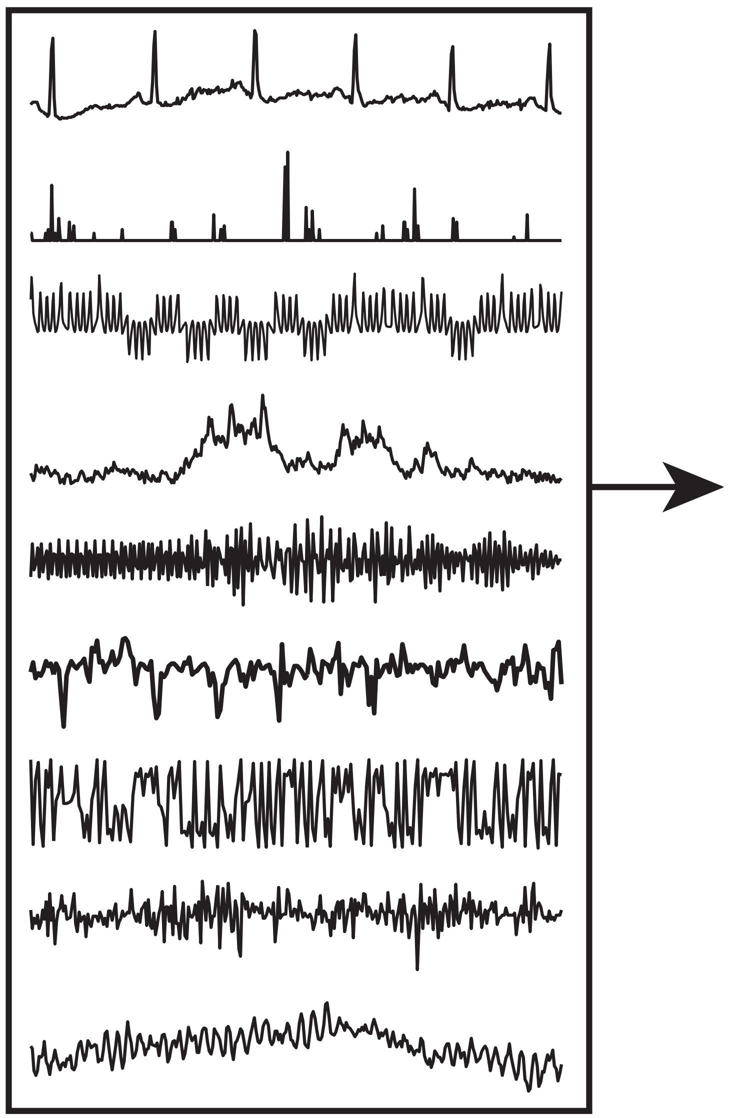
Characterizing time series

How can I reduce complex time-varying patterns to informative summary statistics?



Global features $f : \mathbb{R}^N \to \mathbb{R}$

Methods for characterizing time series can be represented as algorithms that capture time-series properties as real numbers



How fast is it varying? $\tilde{x}_k = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} x_n e^{2\pi i k n/N}$

How variable? $s_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$

How autocorrelated through time?

$$C(\tau) = \langle x_t x_{t+\tau} \rangle = \frac{1}{s_x^2 (N-\tau)} \sum_{t=1}^{N-\tau} (x_t - \bar{x})(x_{t+\tau} - \bar{x})$$

How predictable? $\Phi^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \ln C_r^m(i)$

How stationary? $\frac{\mathrm{std}(\{\overline{x_{1:w}}, \overline{x_{w+1:2w}}, \ldots, \overline{x_{(m-1)w+1:mw}}\})}{\mathrm{std}(x)}$

Connects empirical dynamics to theoretical ideas

Large and diverse methodological literature

the *hctsa* library

Static distribution

Quantiles
Trimmed means
Fits to standard distributions
Outliers
Moments
Entropy
Rank-orderings
Standard deviation

Stationarity

StatAv
Sliding window measures
Bootstraps
Step detection
Distribution comparisons

Basis Functions

Wavelet transform
Peaks of power spectrum
Spectral measures
Power in frequency bands

Correlation

Linear autocorrelation
Decay properties
Additive noise titration
Nonlinear autocorrelations
Time reversal asymmetry
Generalized self-correlation
Recurrence structure
Autocorrelation robustness
Scaling and fluctuation analysis
Permutation robustness
Local extrema
Seasonality tests
Zero crossing rates

Model Fitting

Local prediction
GARCH models
Fourier fits
AR models
Exponential smoothing
State space models
Hidden Markov models
Biased walker simulations
Piecewise splines
ARMA models
Gaussian Processes

(Phys) Nonlinear

2D embedding structure
TSTOOL
TISEAN
Fractal dimension
Correlation dimension
Taken's estimator
Poincaré sections
Surrogate data
Nonlinear prediction error
Lyapunov exponent estimate
False nearest neighbors

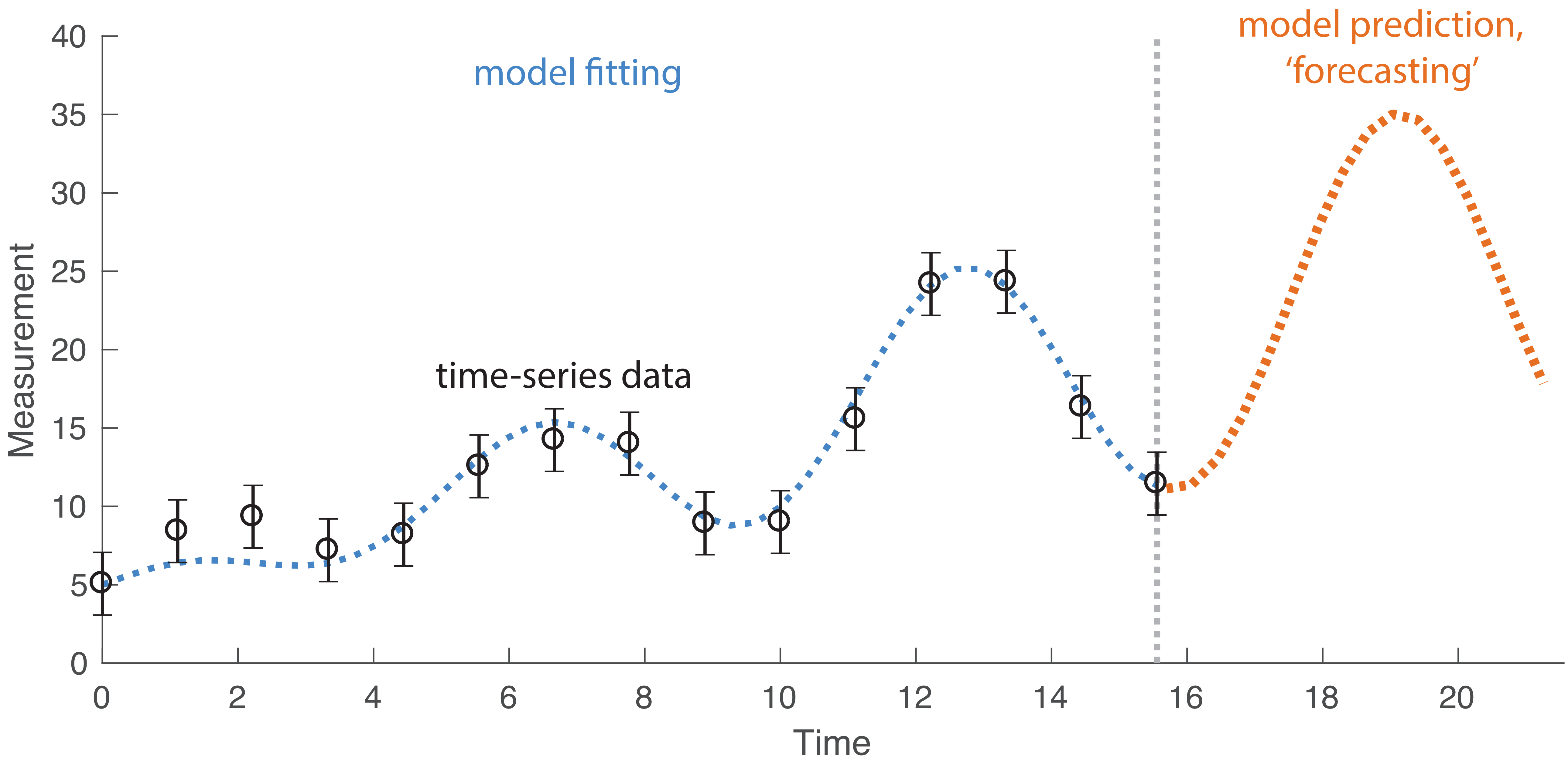
Information Theory

Sample Entropy
Automutual information
Entropy rate
Approximate Entropy
Tsallis entropies

Others

Transition matrices
Local motifs
Dynamical system coupling
Visibility graph
Stick angle distribution
Extreme events
Singular spectrum analysis
Domain-specific techniques

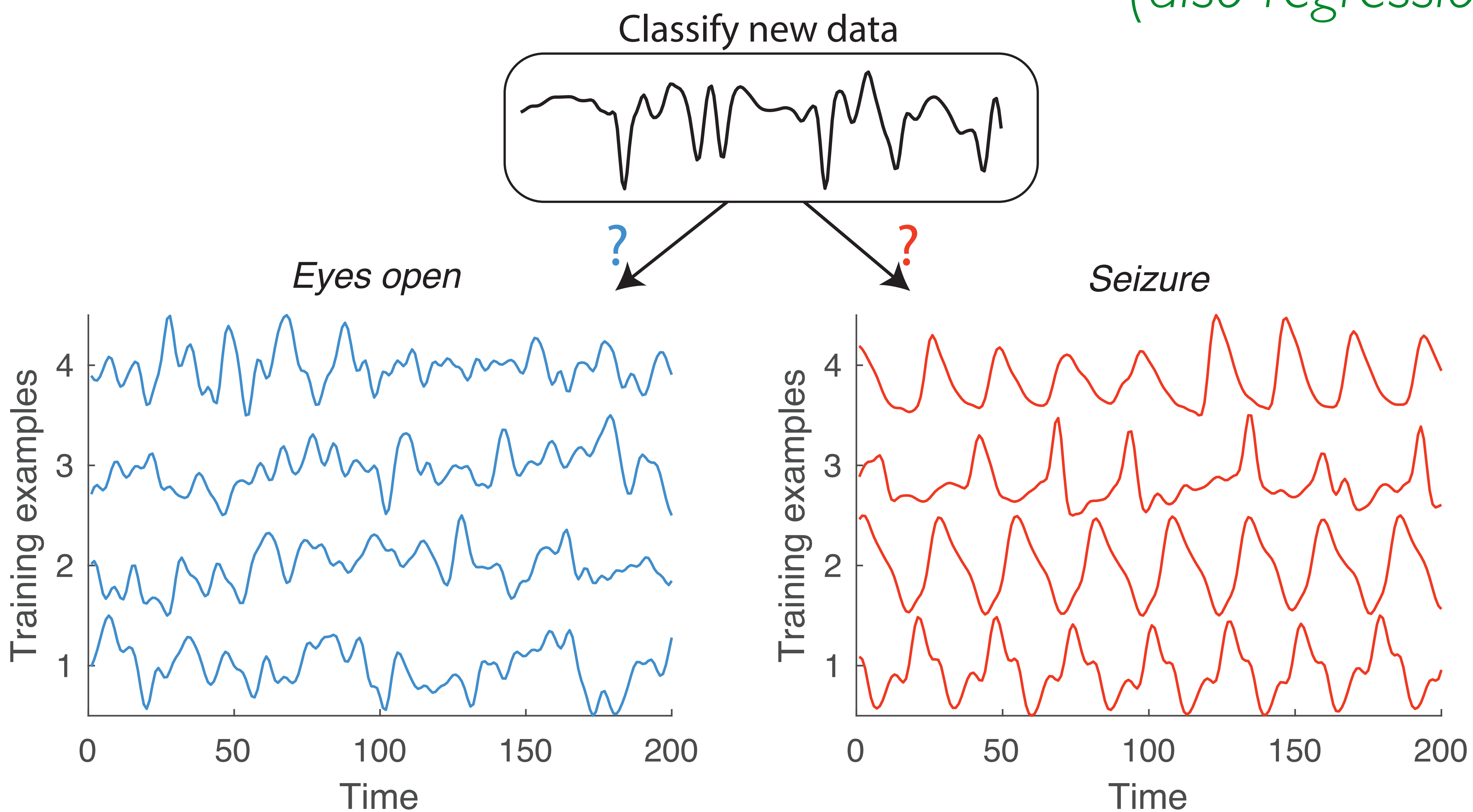
Time-series forecasting



Well-suited to a feature-based representation?

Time-series classification

(also regression)

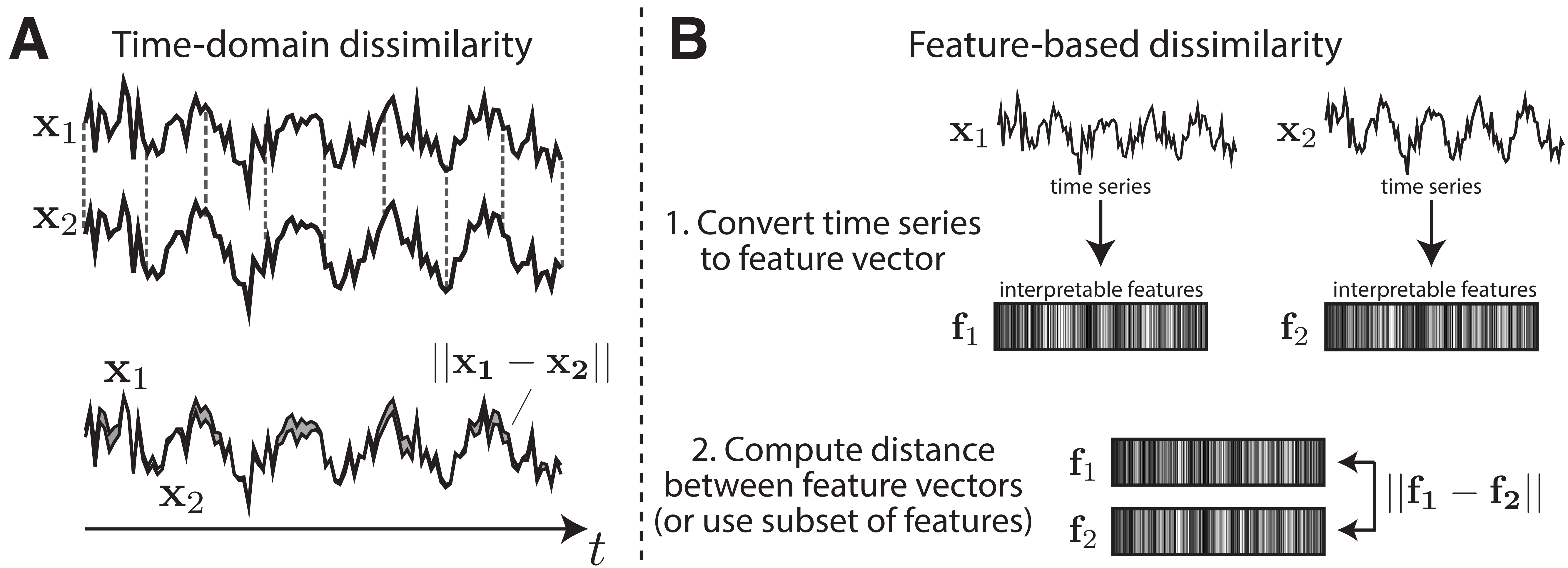


Discriminative features tell you what to measure, and allow you to interpret why.

Time-series classification

Many problems rely on a definition of similarity/distance between pairs of time series

query by content: locate known patterns of interest in a time-series database;
anomaly detection: detect unusual patterns in a time-series database, such as unusual (possibly fraudulent) patterns of credit card transactions
motif discovery: identify commonly recurring subsequences in a time series
clustering: time series are organized into groups;
classification: distinguish different labeled classes of time series from each other



"The crucial problem is not the classificator function (linear or nonlinear), but the selection of well-discriminating features. In addition, the features should contribute to an understanding […]"
—Timmer et al., 1997.

To improve performance:

(i) transform time series into useful (e.g., feature-based) spaces and use simple classifiers

(ii) develop and apply new and complex classifiers in spaces that may not best represent the data for the application

Selecting global features

Global features can be a useful way to represent time series, but there are a huge variety of features…

So how do I pick methods for my data? ***it's an art***

"Do what I did during my PhD"

"Use standard analysis methods from my field"

"Apply a hot new method introduced in PNAS last week"

- Is your proposed method best, or can another (perhaps simpler) method outperform it?
- Are 'new' methods really new, or do they reproduce the performance of existing methods (e.g., from another field, or developed in the past)? *Is any progress being made?*
- Comparison required, but not done in practice (an average of 0.91 other methods, and 1.85 different datasets*).

*Keogh, E. and Kasetty, S., Data Min. Knowl. Disc. **7**, 349 (2003)

Competing interdisciplinary approaches

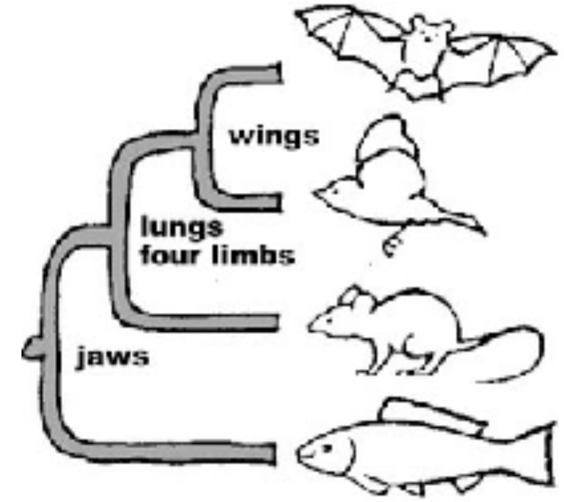
vast libraries of methods lead to diverse opinions

"I know someone smart who uses wavelets"

"Everyone knows you can't apply AR models to nonstationary biomedical data!"

"ARIMA models are a waste of time"

Solution?



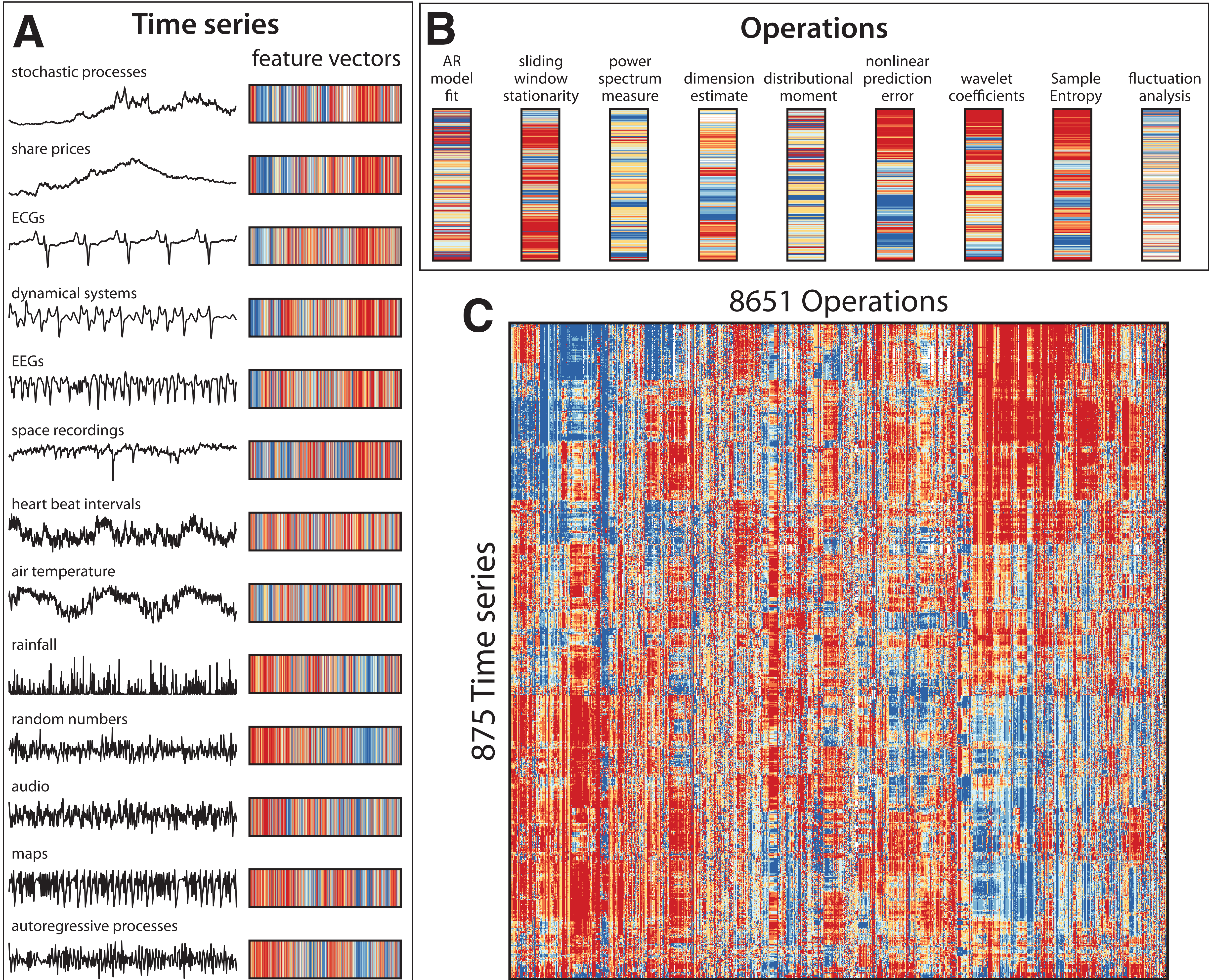
Collect many scientific time series

Collect many scientific time-series analysis methods

Use performance of methods on data to
organize our methods

Use properties of data as measure by the methods to
organize our data

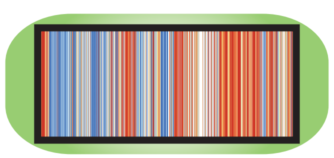
Learn the empirical structure of time series and their methods



Fulcher, B. D., Little, M. A. & Jones, N. S. (2013). *J. Roy. Soc. Interface* **10,** 20130048.

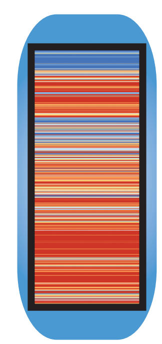
Empirical fingerprints

A flexible, powerful, and data-driven means of comparing time series, and analysis methods.



= time series of type 'green'

captures properties measured by diverse scientific methods

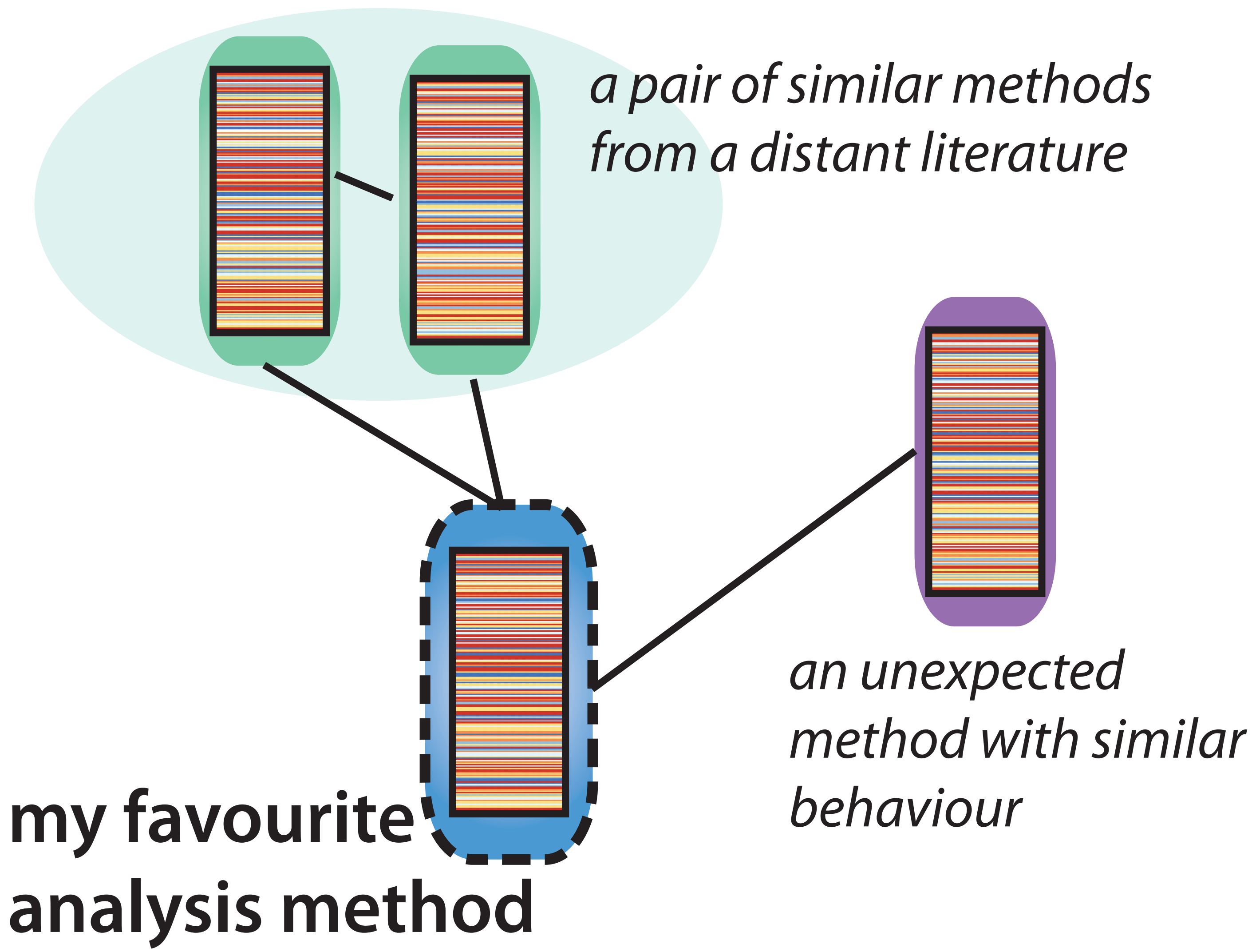


= operation of type 'blue'

captures behaviour across a range of empirical time series

Organizing our methods

Which time-series analysis methods are similar to the methods I use?



Connects scientific methods using their empirical behaviour *long-range scaling*

power spectral density

linear time-series models

stationarity

distribution

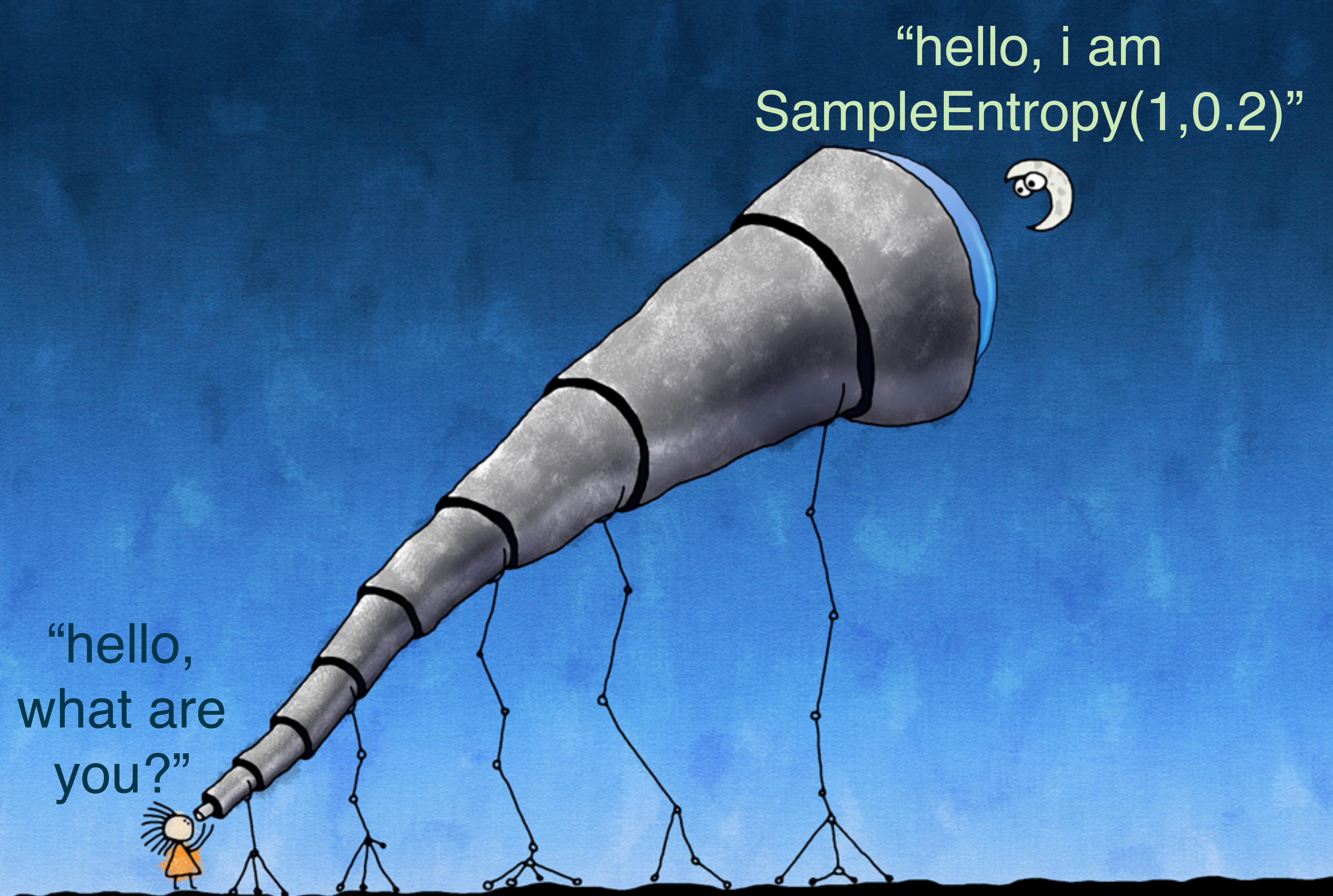
entropy

correlation dimension

complexity

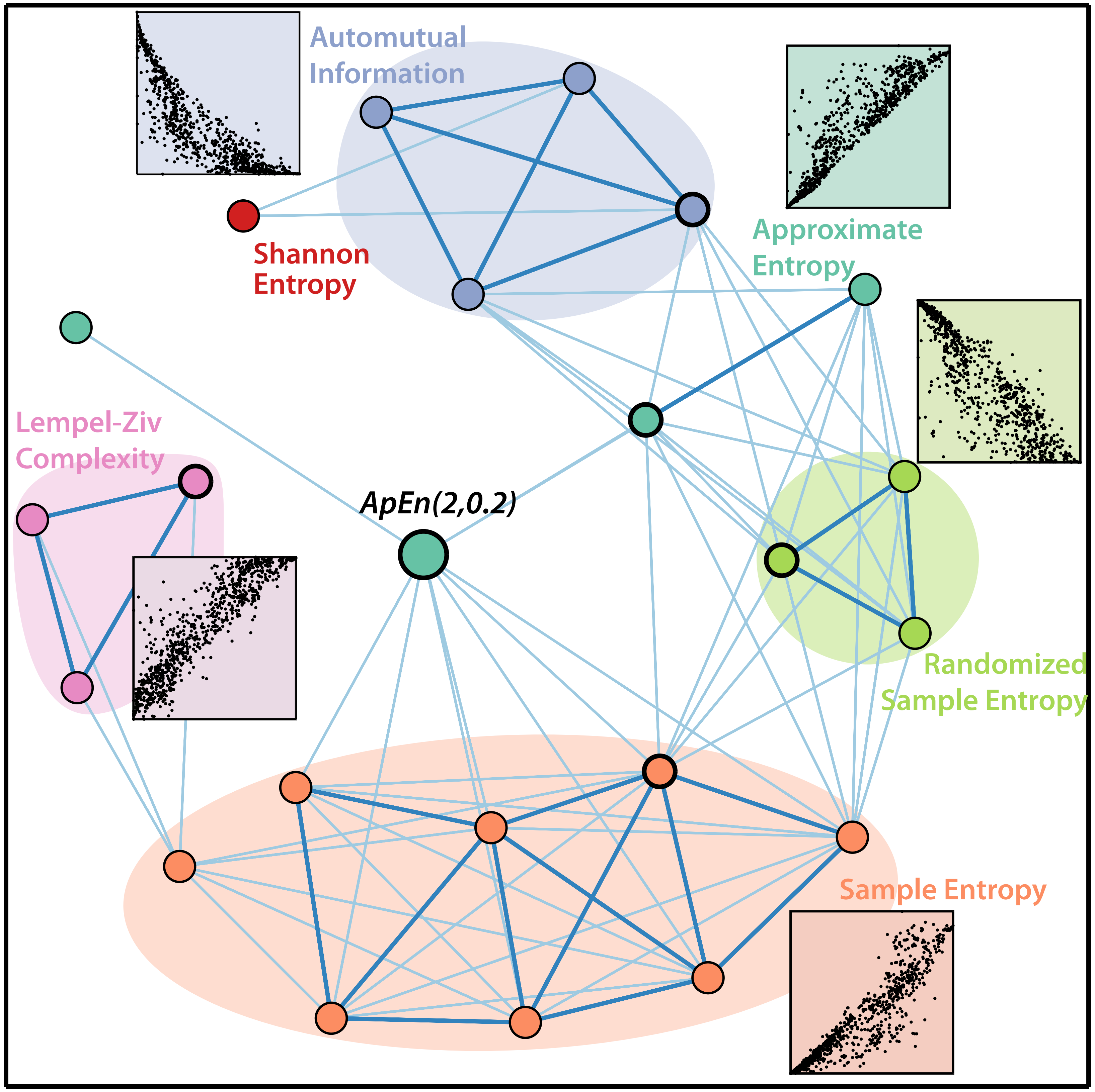
information theory

BIG PICTURE

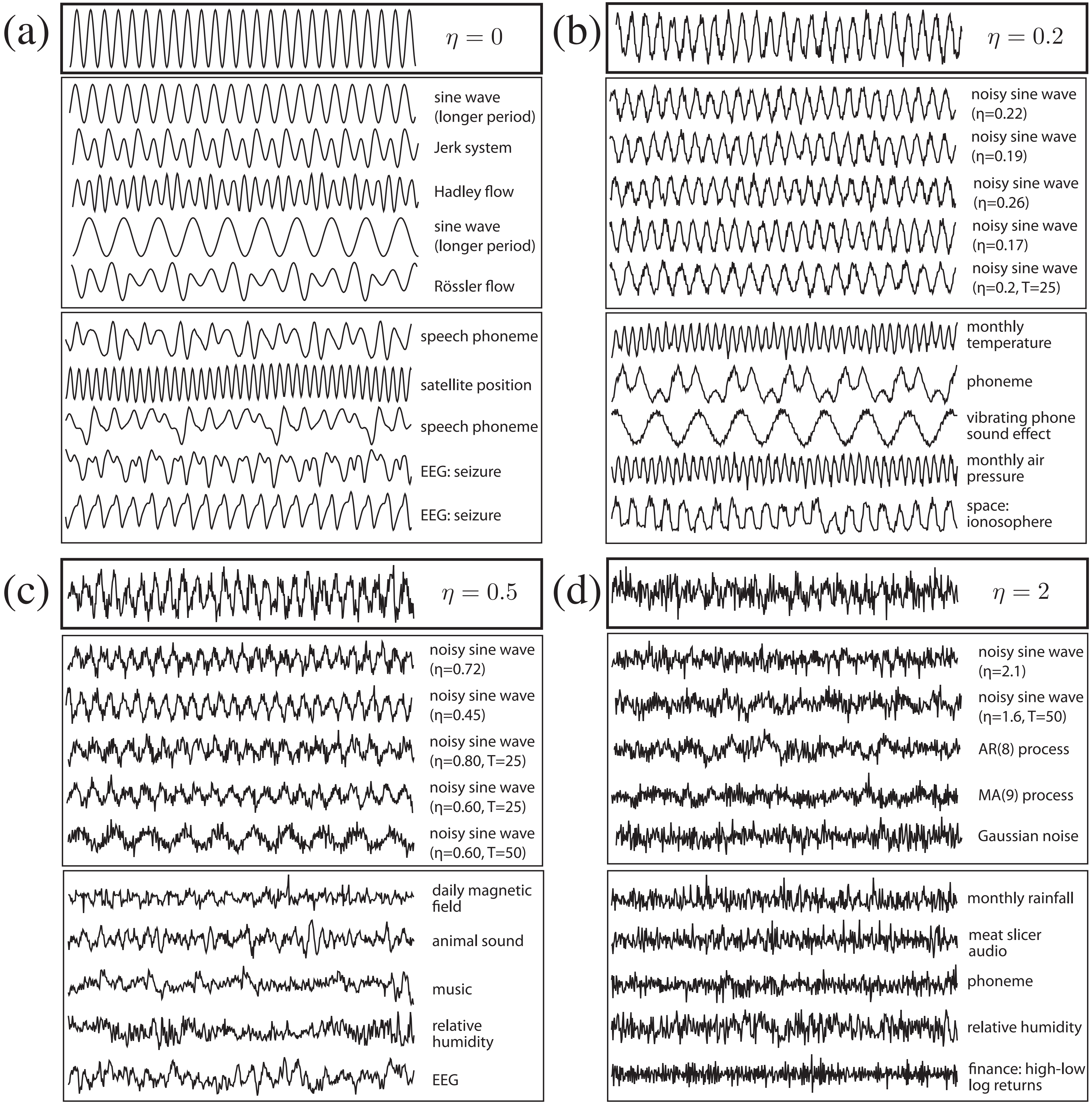


ZOOMING IN

Local Neighborhoods

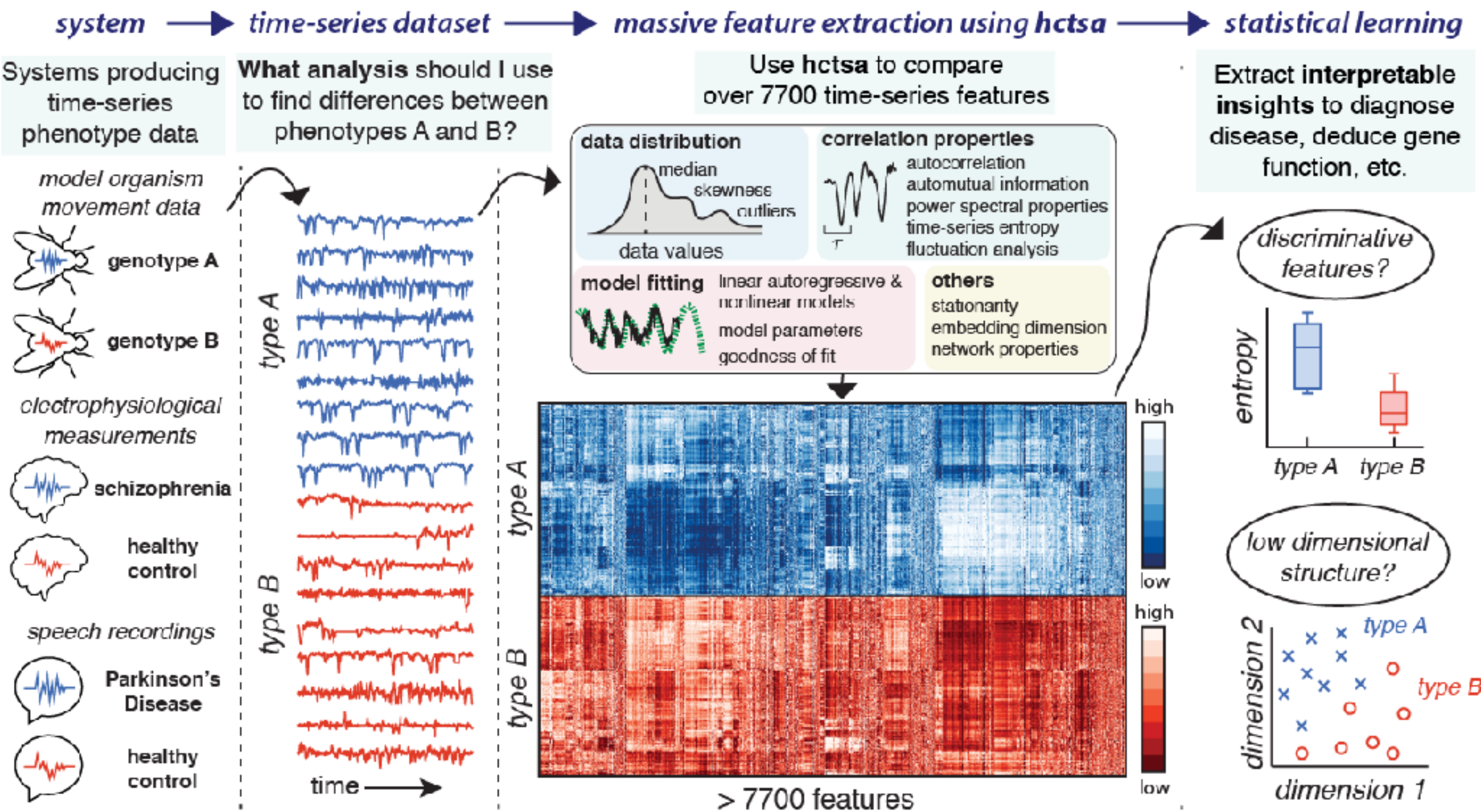


Discover interdisciplinary connections between our methods for time-series analysis



Brings our data and models closer together

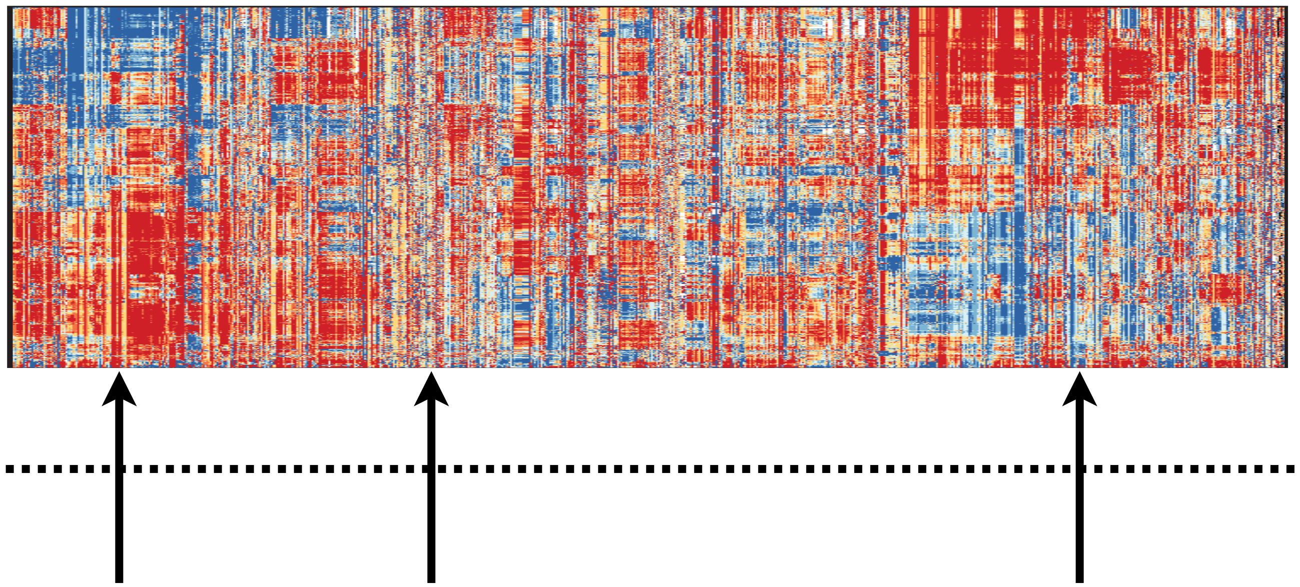
Highly comparative time-series analysis for classification



A very general problem: what method should I use?

Highly comparative time-series analysis

1. Compute and compare thousands of analysis methods

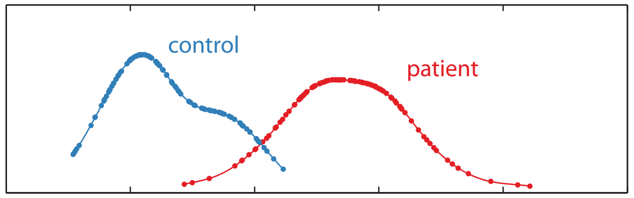


2. Select methods that perform well on your data

3. Interpret new methods to gain insights into your data

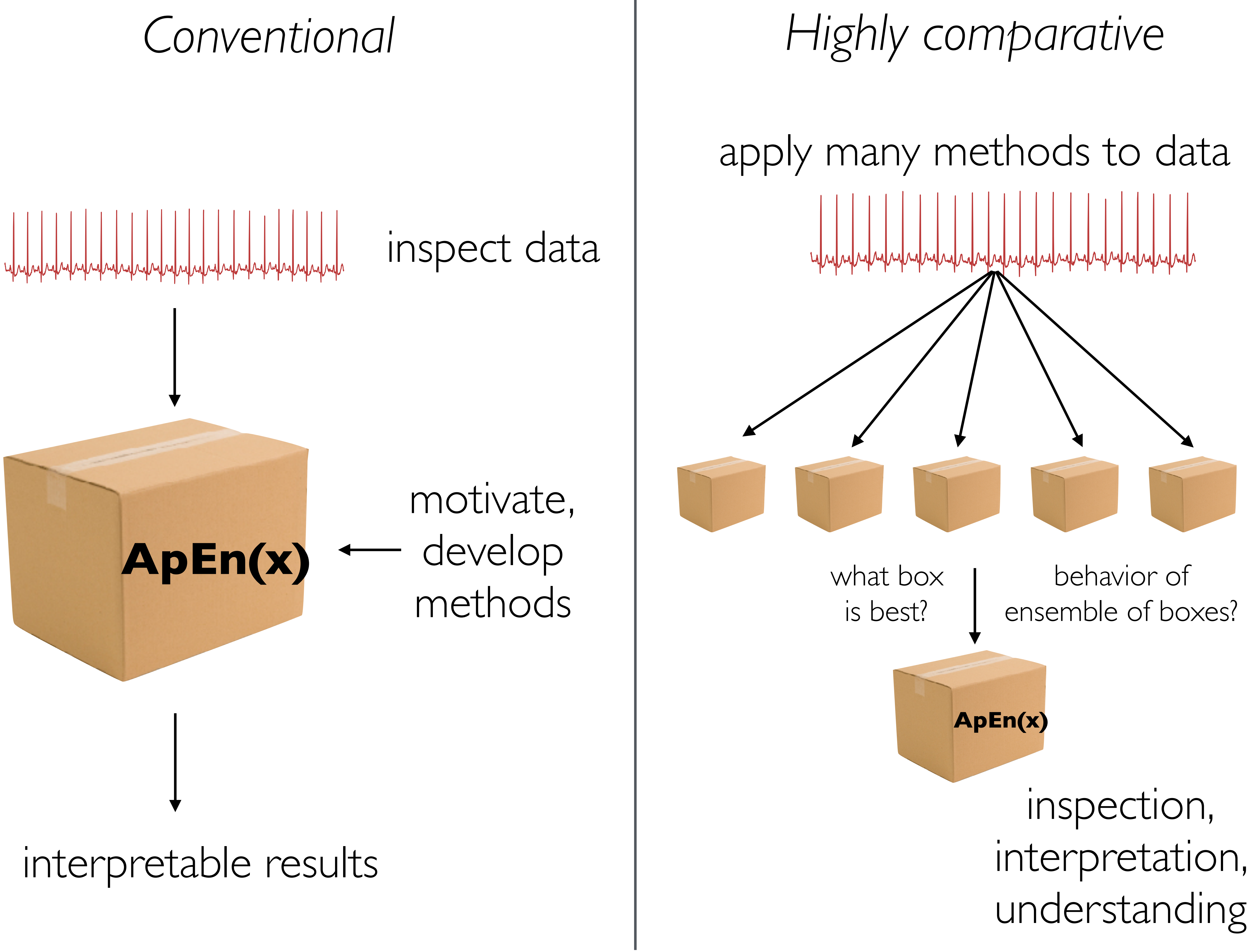
"Signals from the patient group are less predictable"

"Single neuron recordings from the frontal lobe have more outliers and intermittent fluctuations"

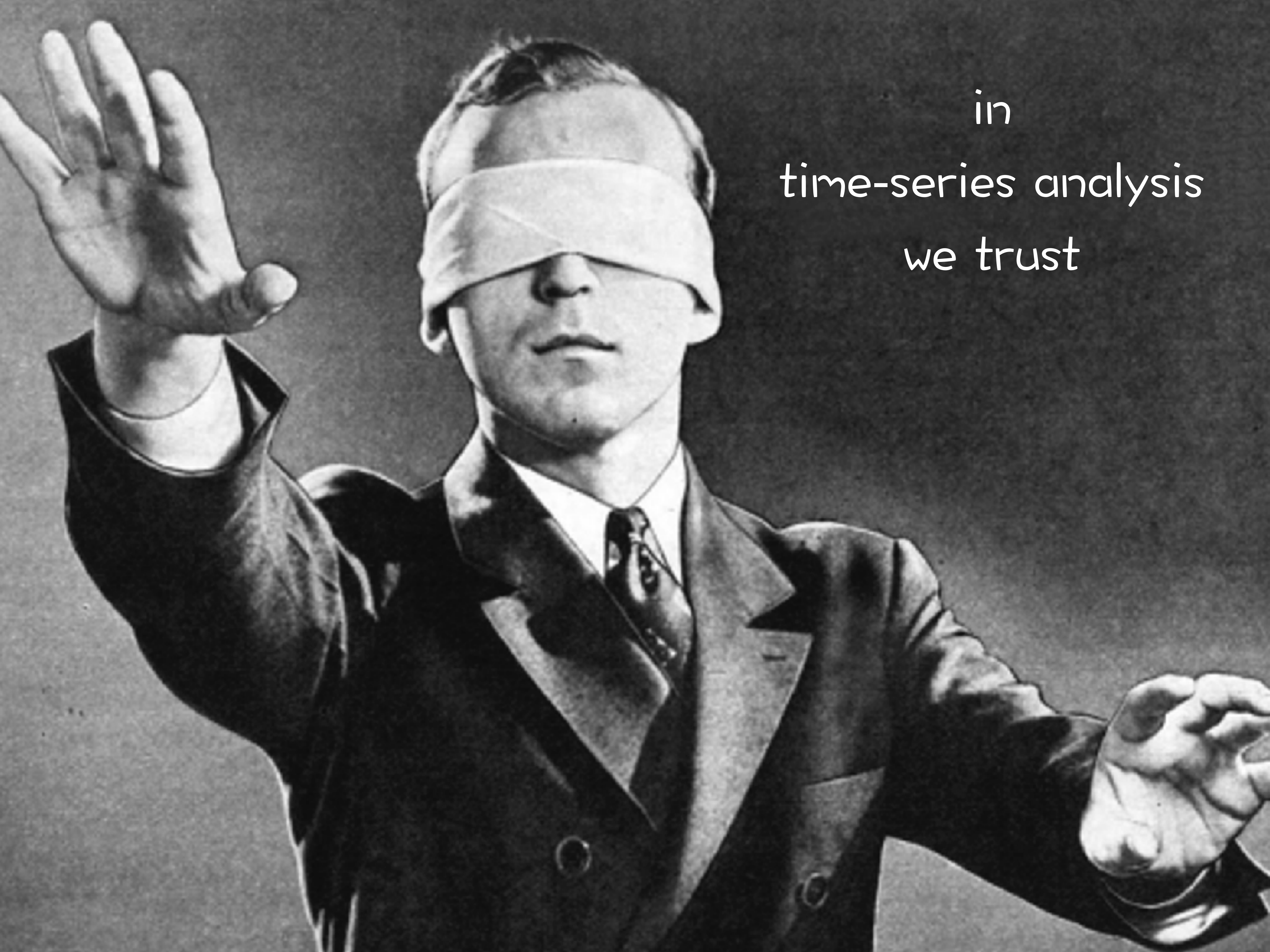


EEG entropy in pre-frontal cortex

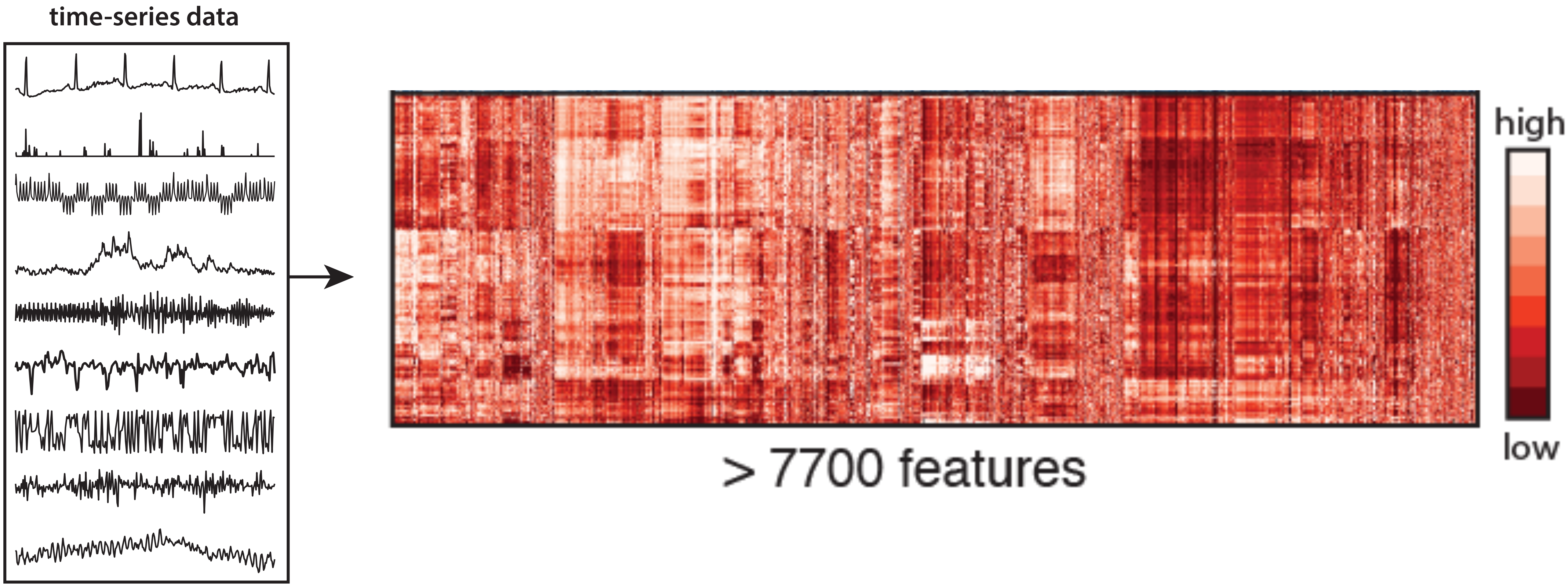
BD Fulcher, MA Little, and NS Jones. *J. R. Soc. Interface*, **10**:83 (2013), *DOI: 10.1098/rsif.2013.0048*


Conventional
inspect data
ApEn(x)
motivate,
develop
methods
interpretable results
Highly comparative
apply many methods to data
what box
is best?
behavior of
ensemble of boxes?
ApEn(x)
inspection,
interpretation,
understanding

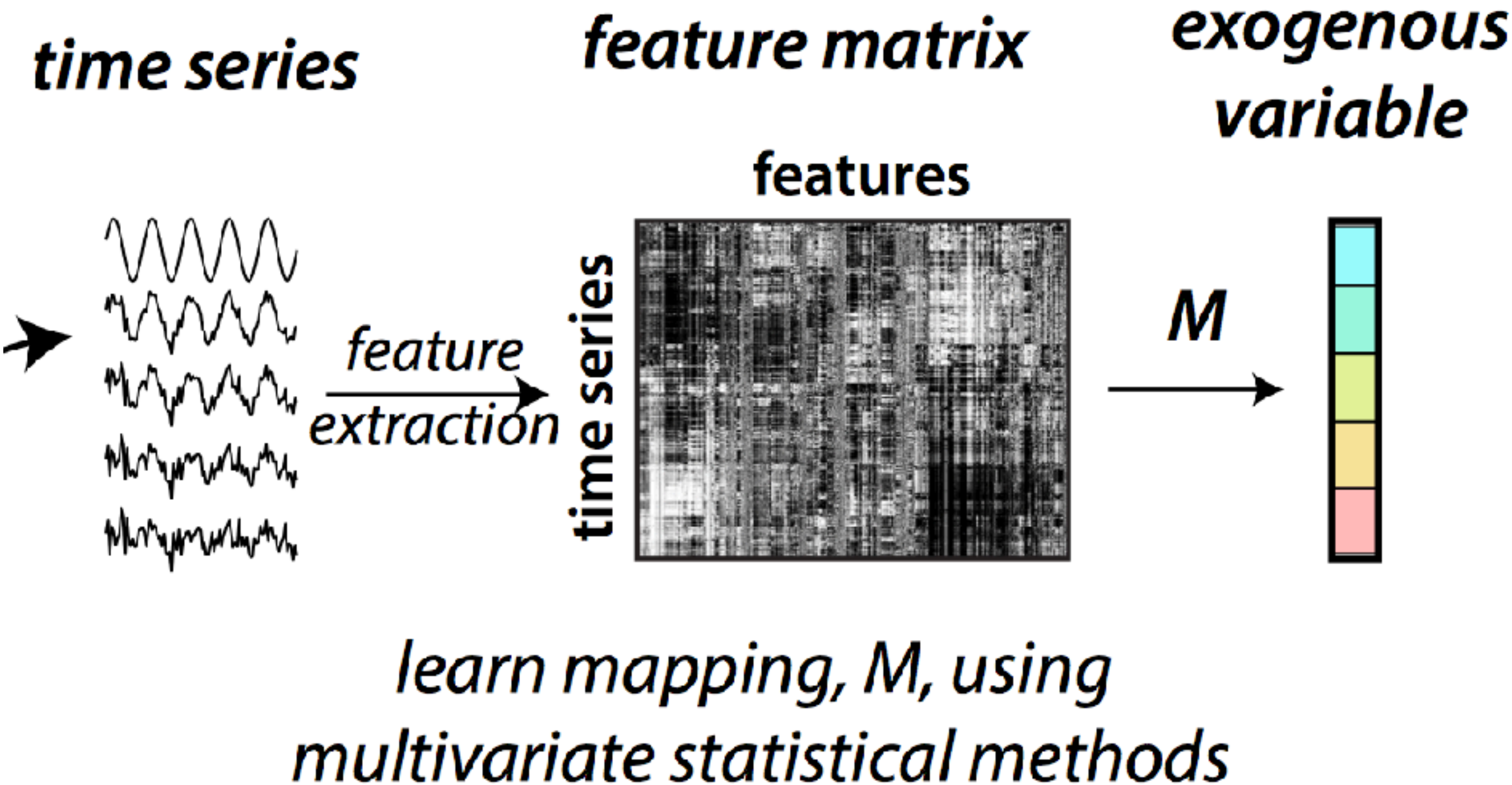

Time-series analysis 101:
always look at your data


in
time-series analysis
we trust

Mapping to features opens algorithmic possibilities



Converts to a static problem; access to traditional algorithms for statistical learning: regression, classification, …

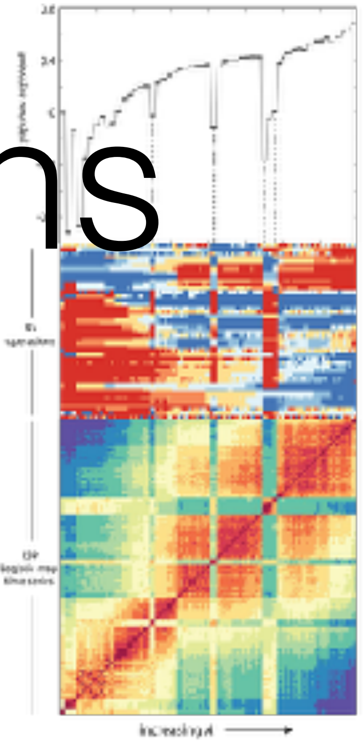


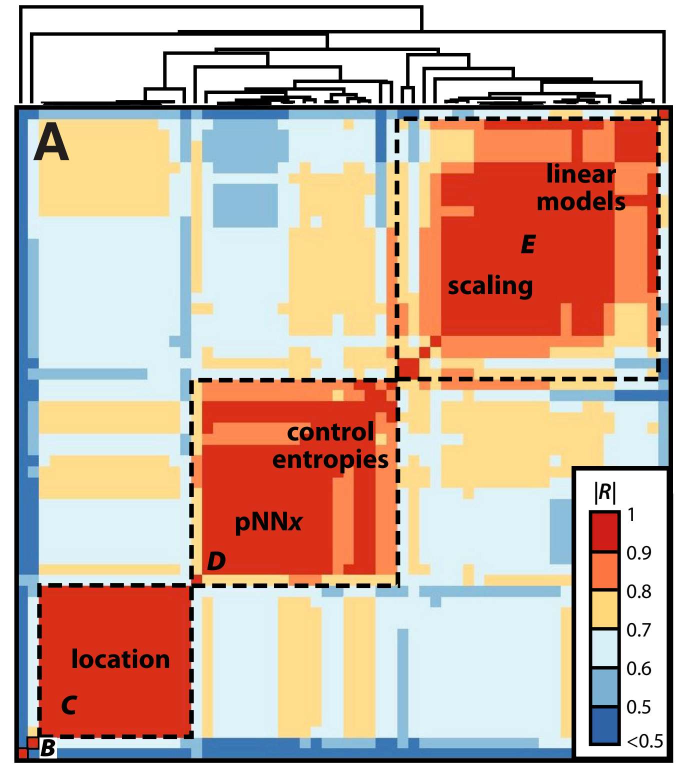
On arXiv next week:

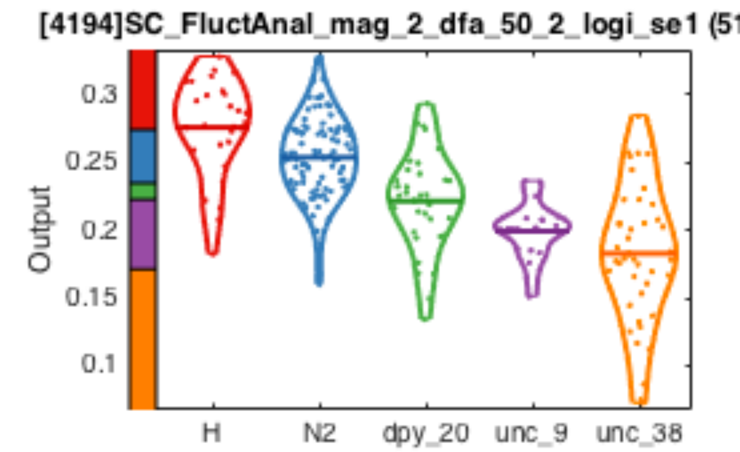
Regression of Exogenous Variables from Time Series.

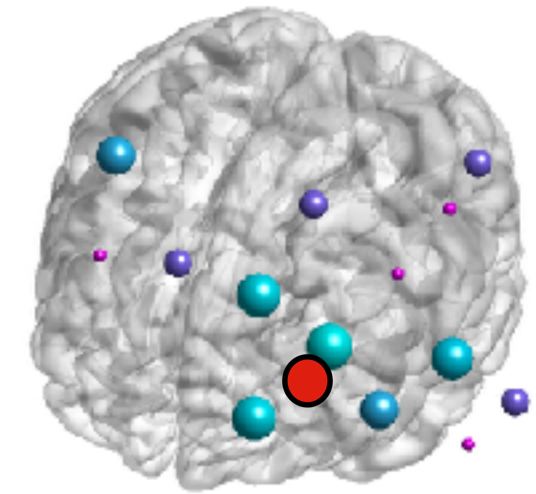
Applications

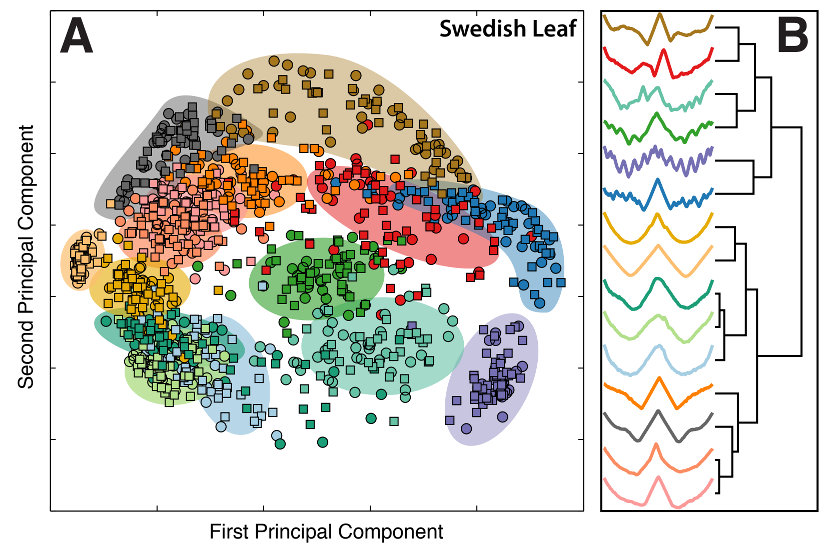
- Seismology
- Heart rate intervals
- Fetal heart rates
- Emotional speech
- Parkinsonian speech
- Epileptic EEG
- Mouse fMRI
- Learning genotypes from worm/fly movement data
- Astrophysical light curve data
- Simulated: criticality, chaos

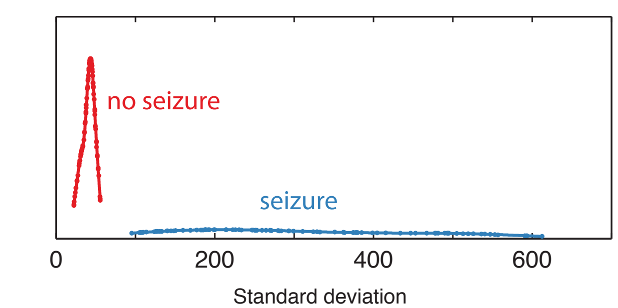


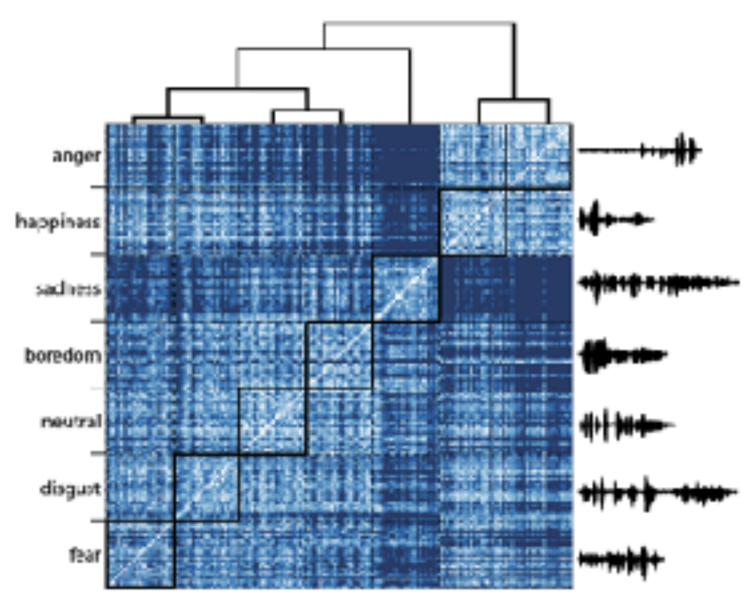


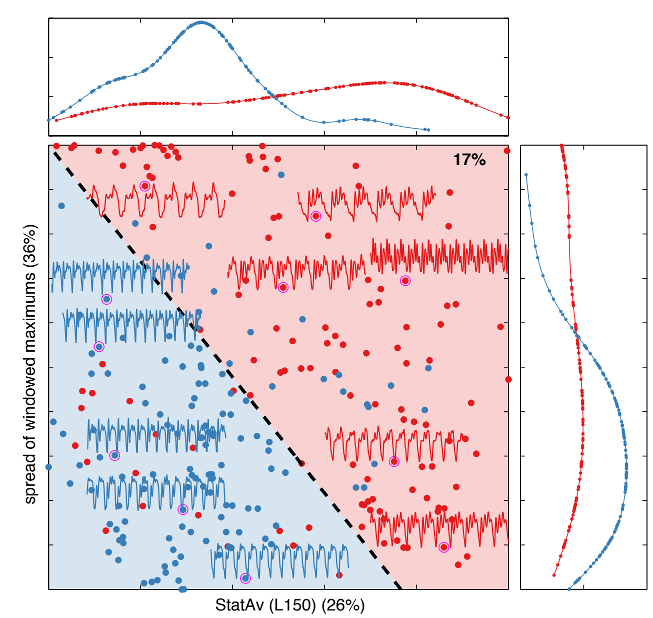




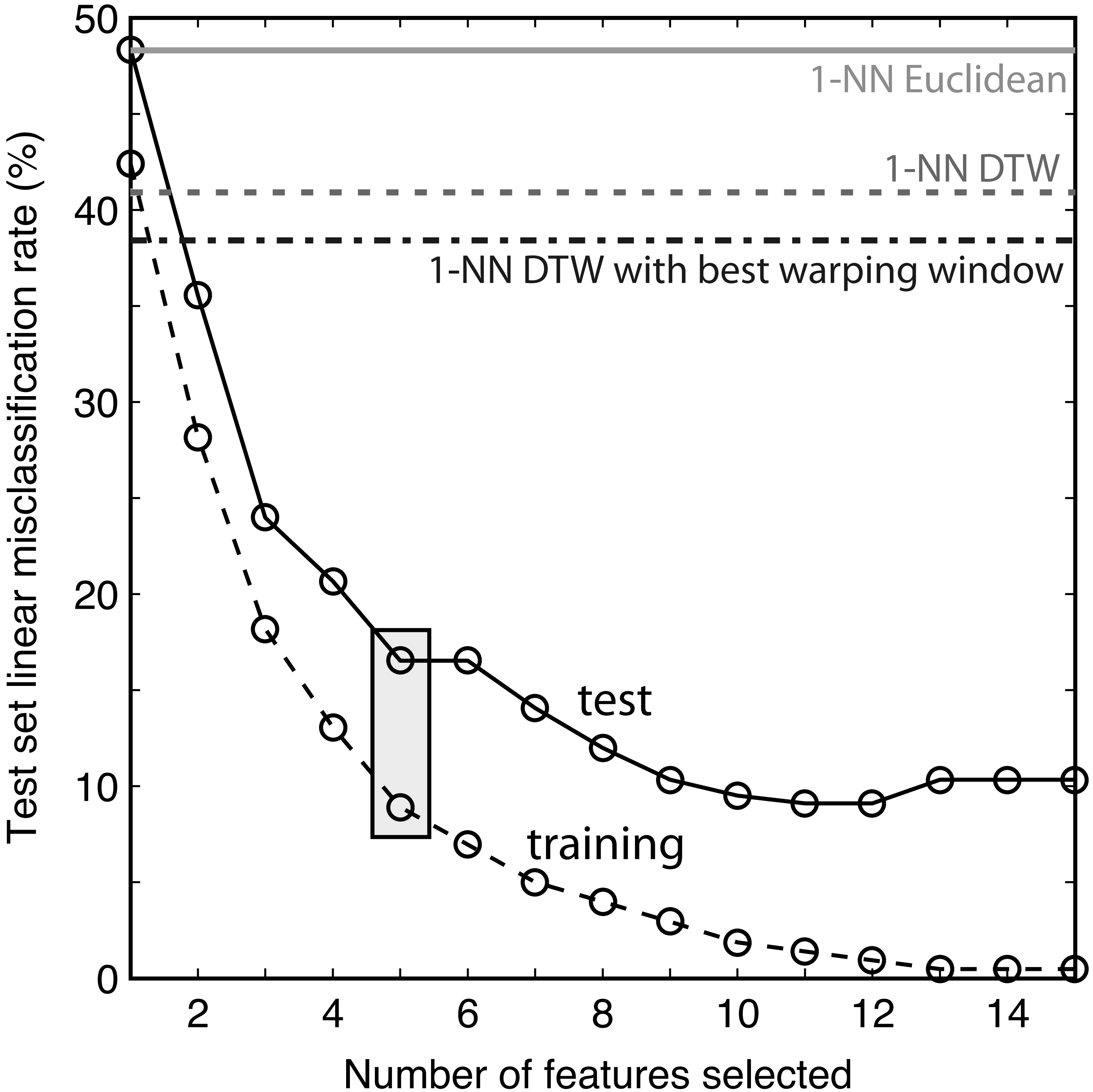








Learning feature spaces for time-series classification



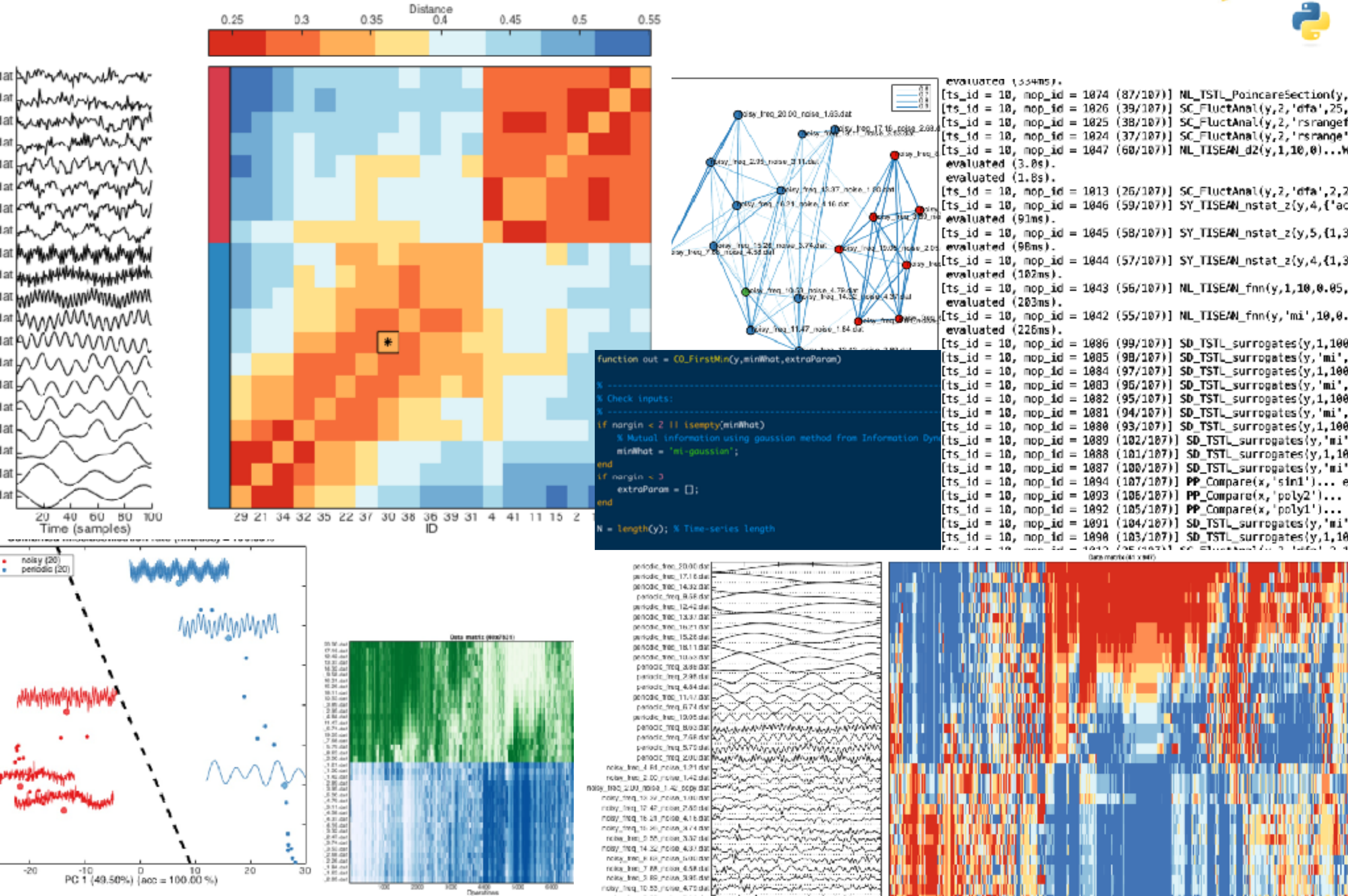
systematic, automatic, data-driven

massive dimensionality reduction

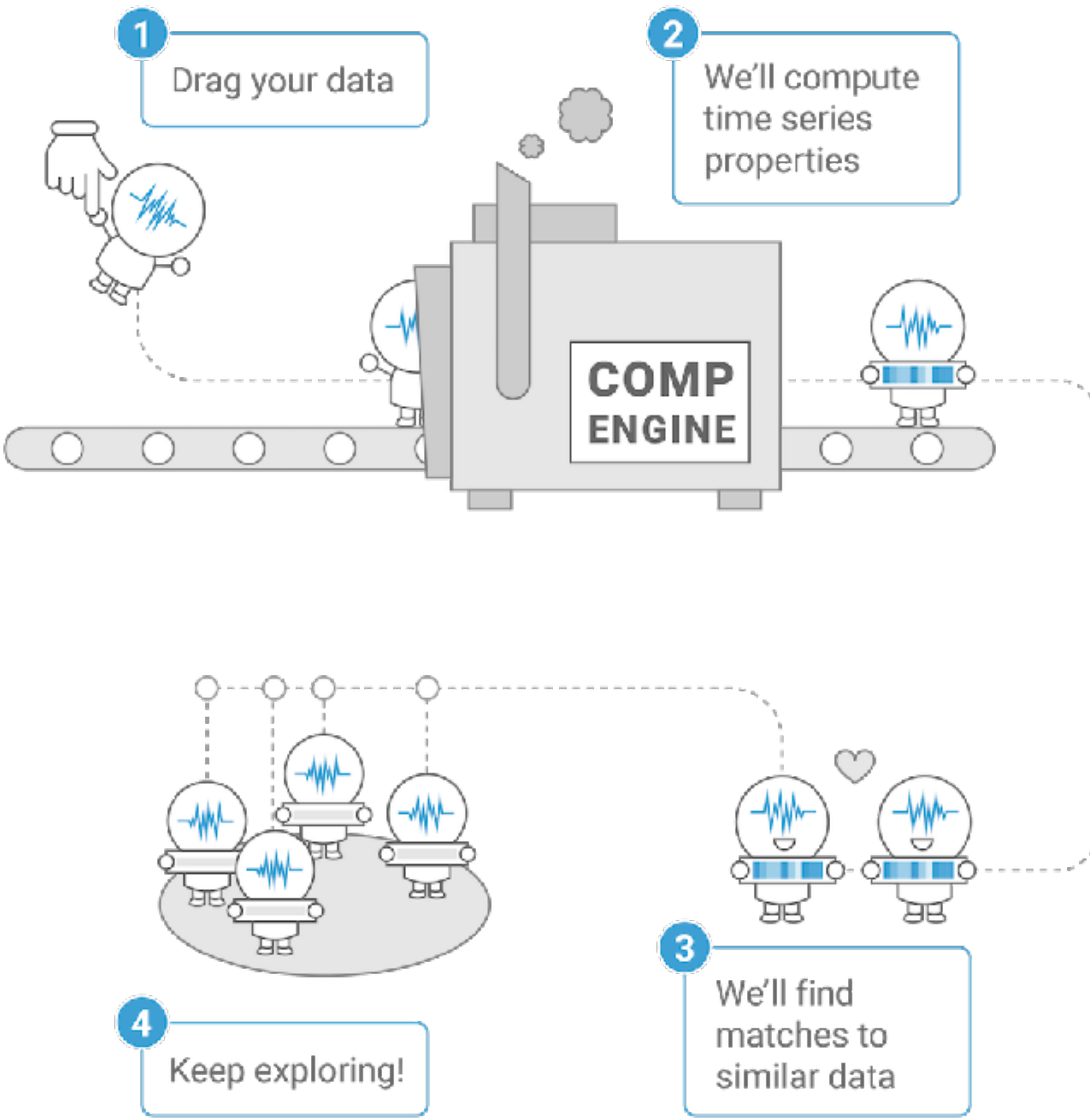
fast classification of new examples

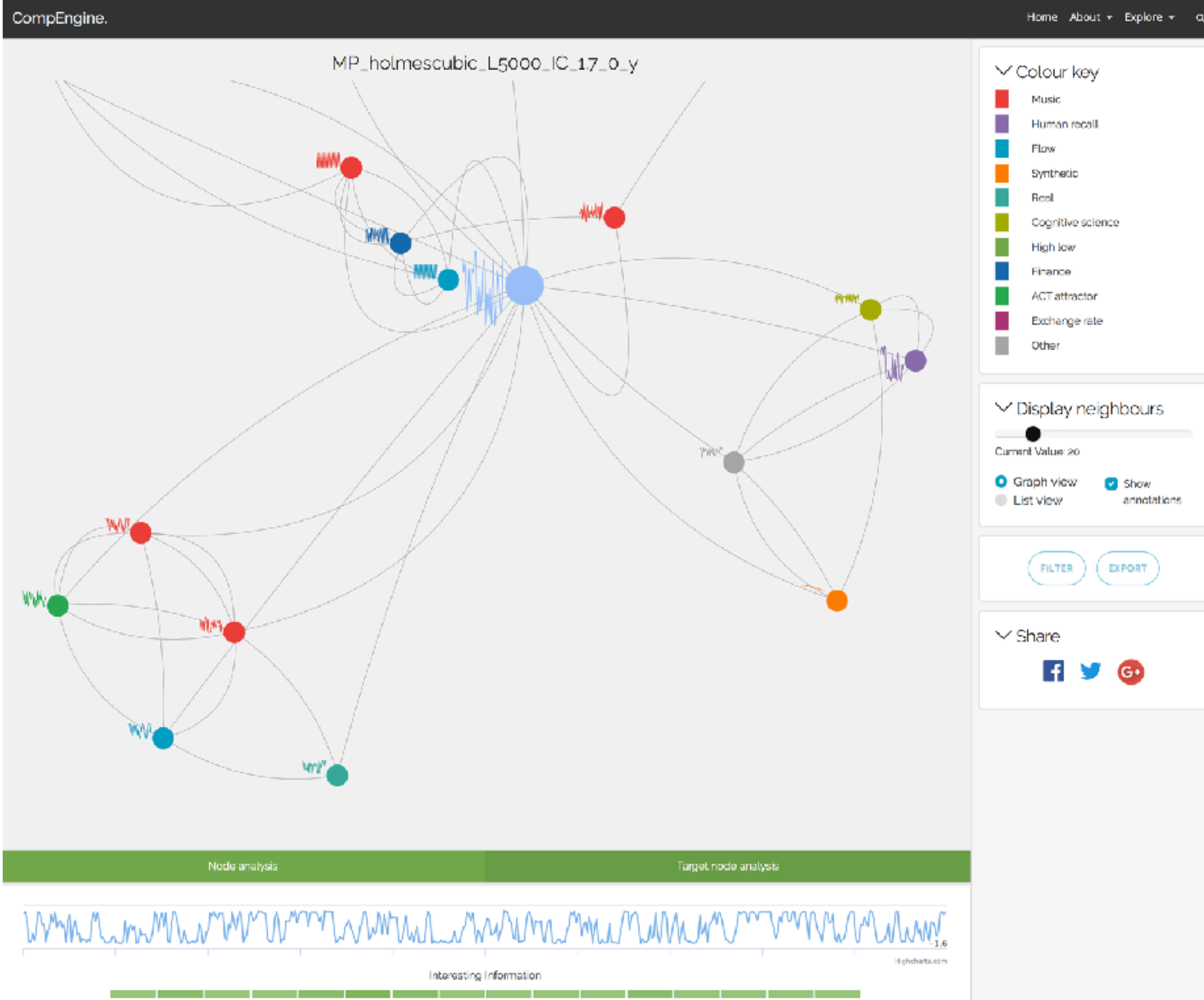
diverse, interpretable features

B. D. Fulcher & N. S. Jones, Highly comparative feature-based time-series classification. *IEEE Trans. Knowl. Data Eng.* **26**, 3026–3037 (2014) https://github.com/benfulcher/hctsa



New, interactive *compEngine* website out this month!

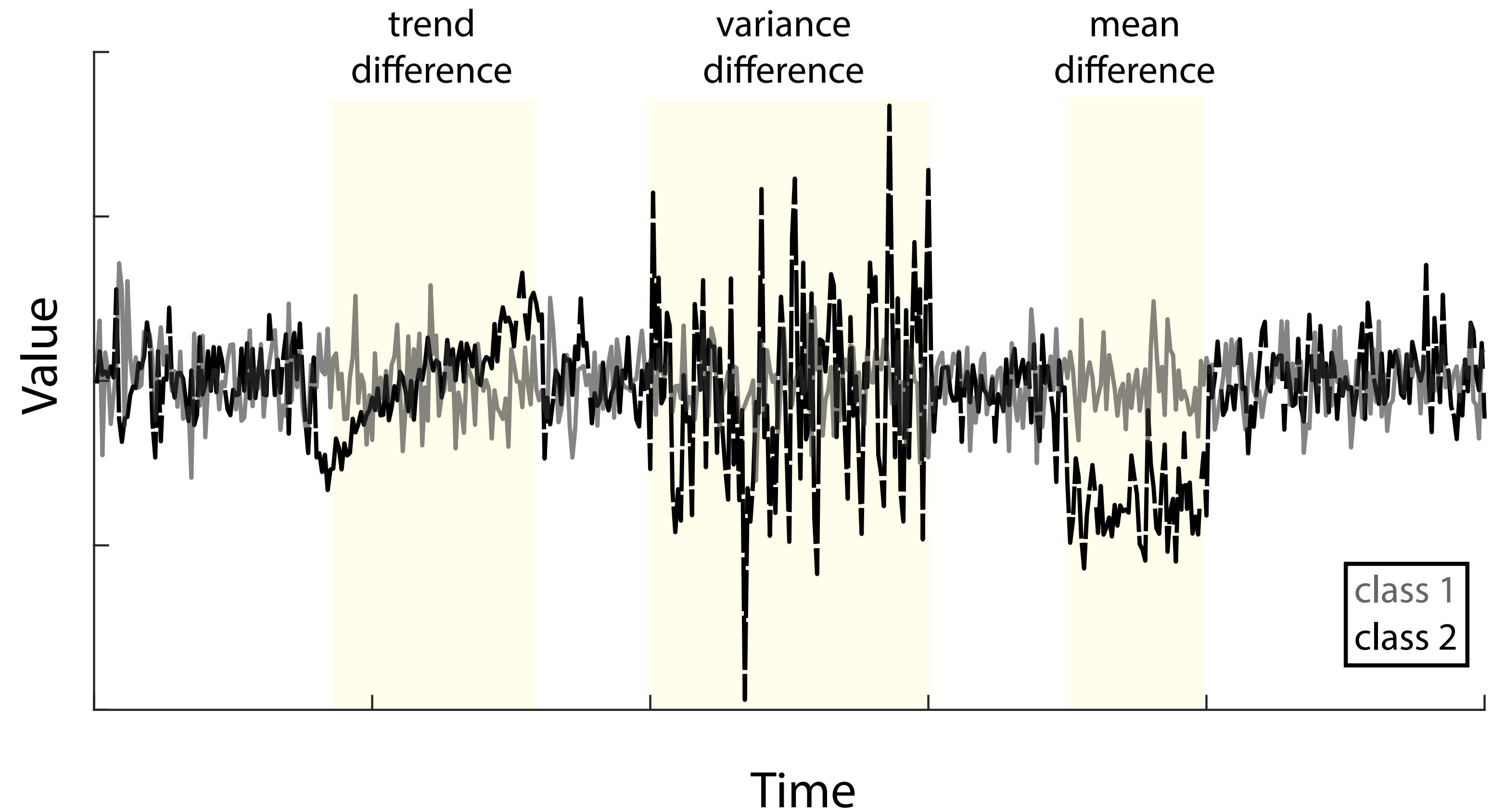



CompEngine.
Home
About
Explore
MP_holmescubic_L5000_IC_17_0_y
Colour key
Music
Human recall
Flow
Synthetic
Real
Cognitive science
High low
Finance
ACT attractor
Exchange rate
Other
Display neighbours
Current Value: 20
Graph view
List view
Show annotations
FILTER
EXPORT
Share
Node analysis
Target node analysis
Interesting Information

Beyond global features…

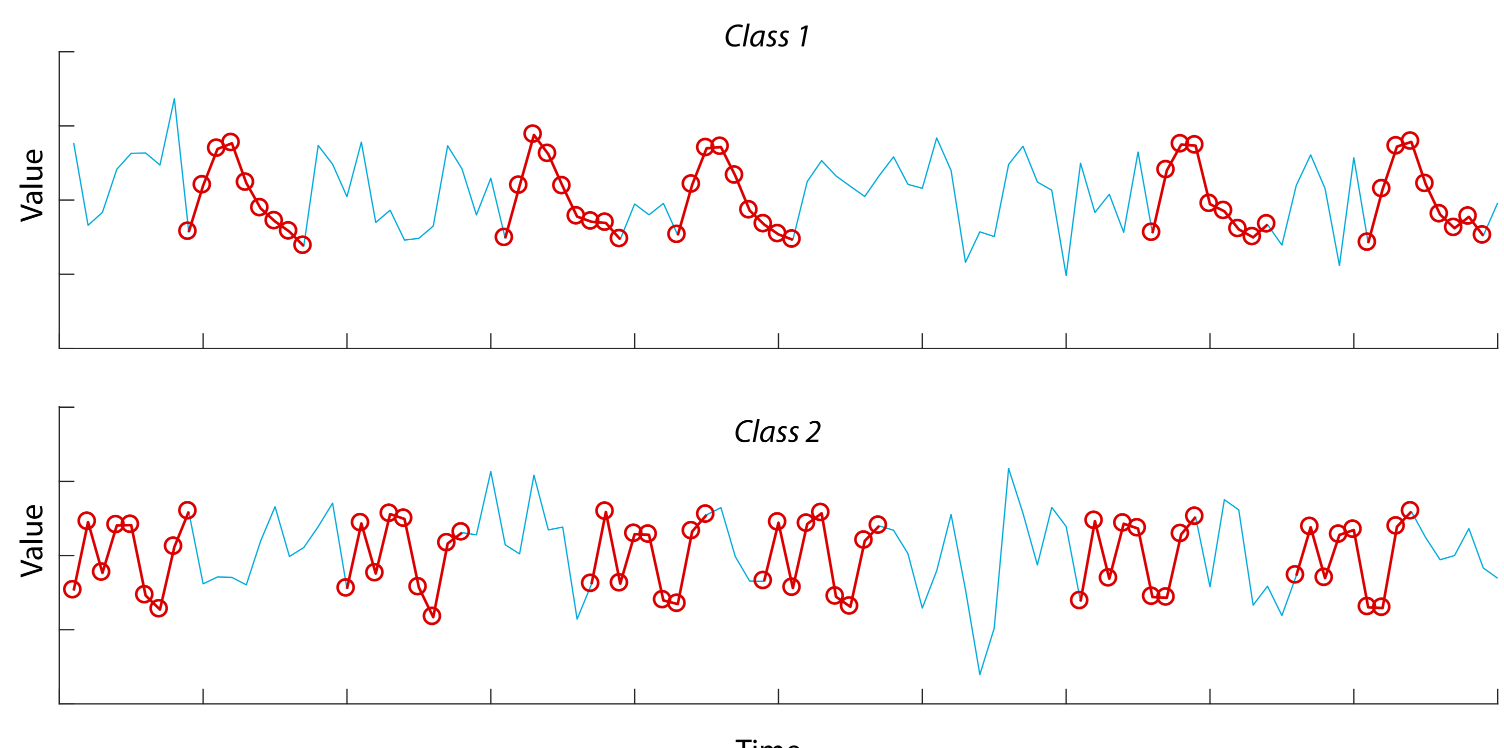
Interval features

simple features measured in sub-interval



Pattern dictionaries

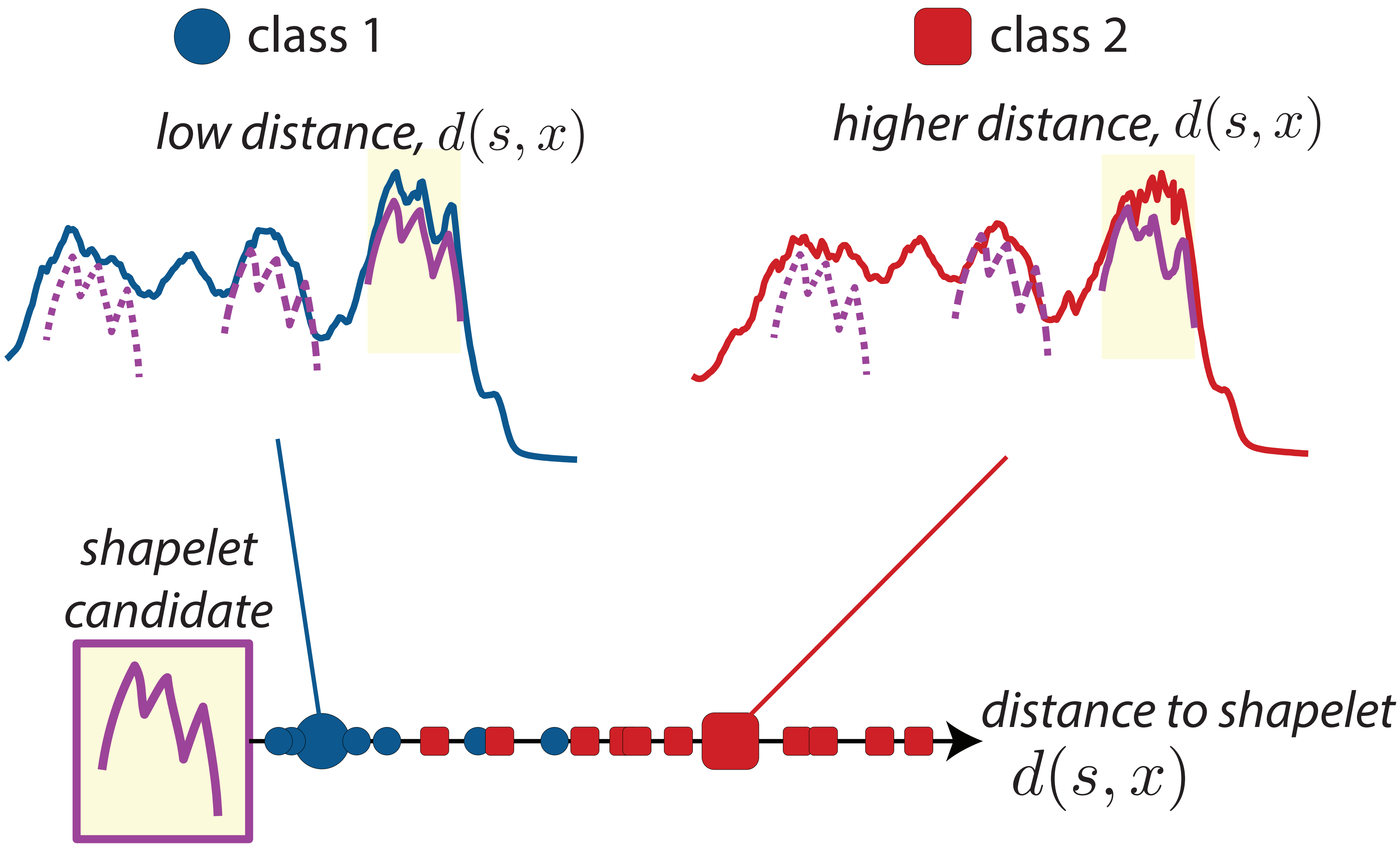
characteristic repeating motifs (e.g., organism movement)



Beyond global features…

Shapelets

matching predictive subsequences

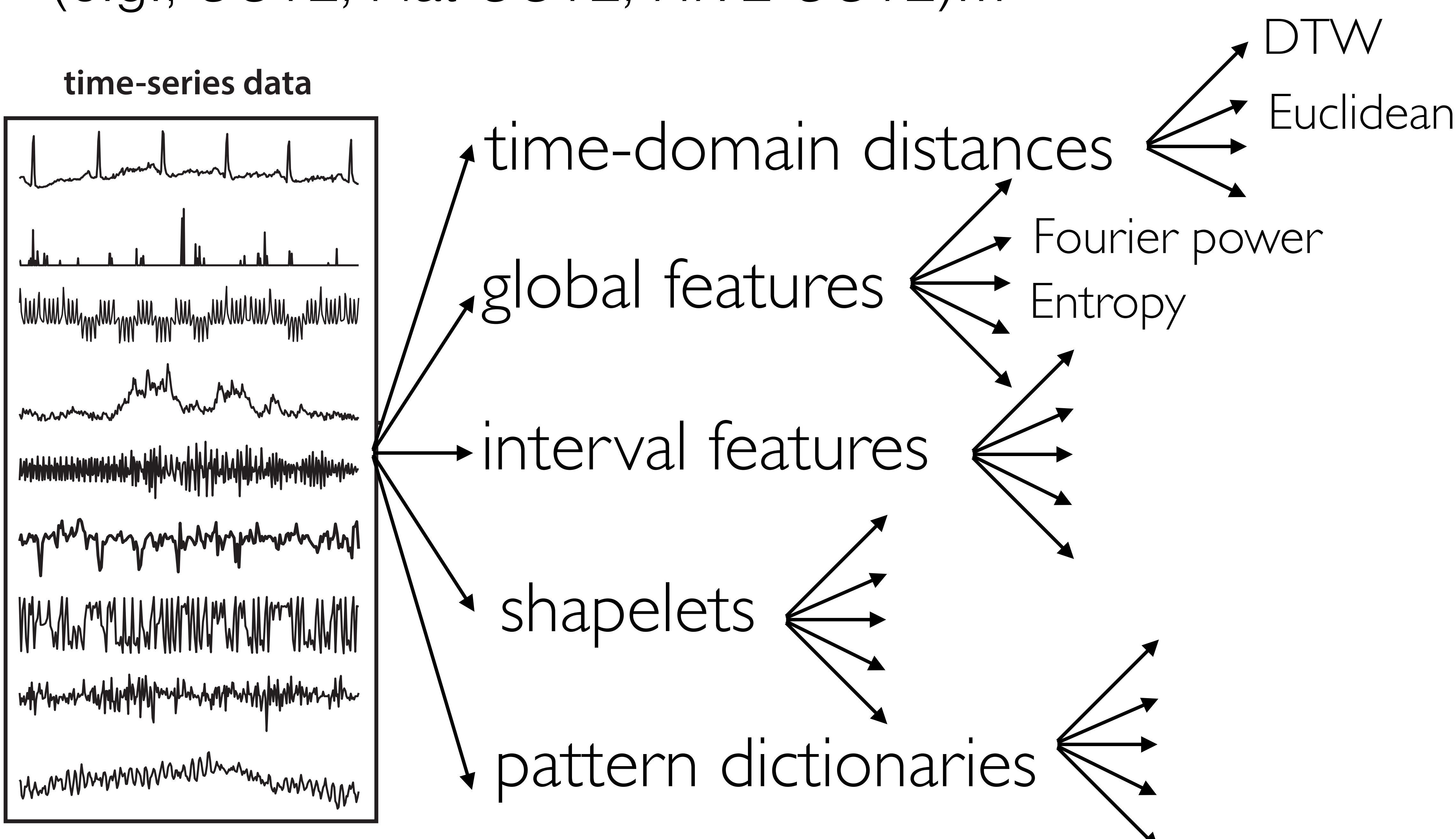


Hybrid approaches

- Define distance metrics containing multiple distance terms
  - E.g., Reweight time-series distances by global feature values
- Define features as the set of time-domain distances to a set of training time series.

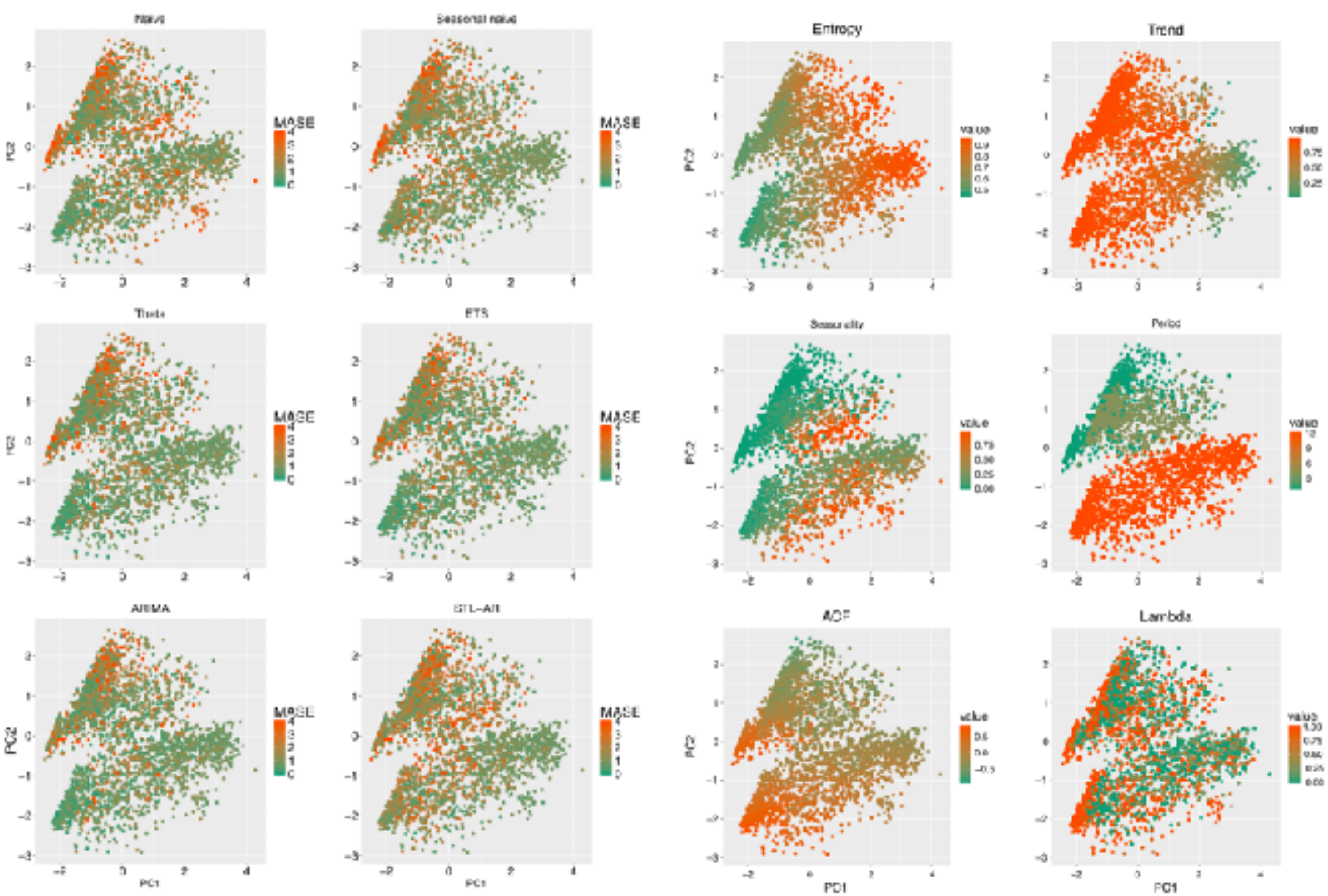
So what representation to use?

- So many to choose from! Ideally methods are motivated by a specific question asked of data (*quantifiable outcome metric*).
- Ensemble methods developing to combine myriad representations (e.g., *COTE, Flat-COTE, HIVE-COTE*)…



Feature-based forecasting

- Do time-series display characteristic patterns from the past that help us to predict its values in the future?
- Are some types of time series suited to different types of forecasting algorithms?




International Journal of Forecasting 33 (2017) 345–358

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

ELSEVIER

Visualising forecasting algorithm performance using time series instance spaces

Yanfei Kang [a,*], Rob J. Hyndman [b], Kate Smith-Miles [c]

[a] School of Economics and Management, Beihang University, Beijing, 100191, China
[b] Department of Econometrics and Business Statistics, Monash University, Clayton VIC 3800, Australia
[c] School of Mathematical Sciences, Monash University, Clayton VIC 3800, Australia

Two Major Take Homes

The most useful methods for a given task are determined by the **structure of the data** (e.g., whether time series are of the same length, are phase-aligned, or whether class differences are global or restricted to specific intervals), and the **context of the problem** (e.g., whether accuracy or interpretability is more important, and what type of understanding would best address the domain question of interest).

Modern statistical methods can help us to leverage decades of interdisciplinary research on algorithm development to ***tailor our methods to our data***. Feature-based learning allows analysts to leverage the power and sophistication of diverse interdisciplinary methods to glean ***interpretable understanding*** from their data.

What Next?

Algorithm development *requires comparison* to demonstrate progress. We need to *properly structure* combined research efforts.

How to construct optimal *(reduced)* feature spaces? [e.g., to classify data, to understand which algorithms suit which types of data]

How best to compare multiple data representations? [COTE]

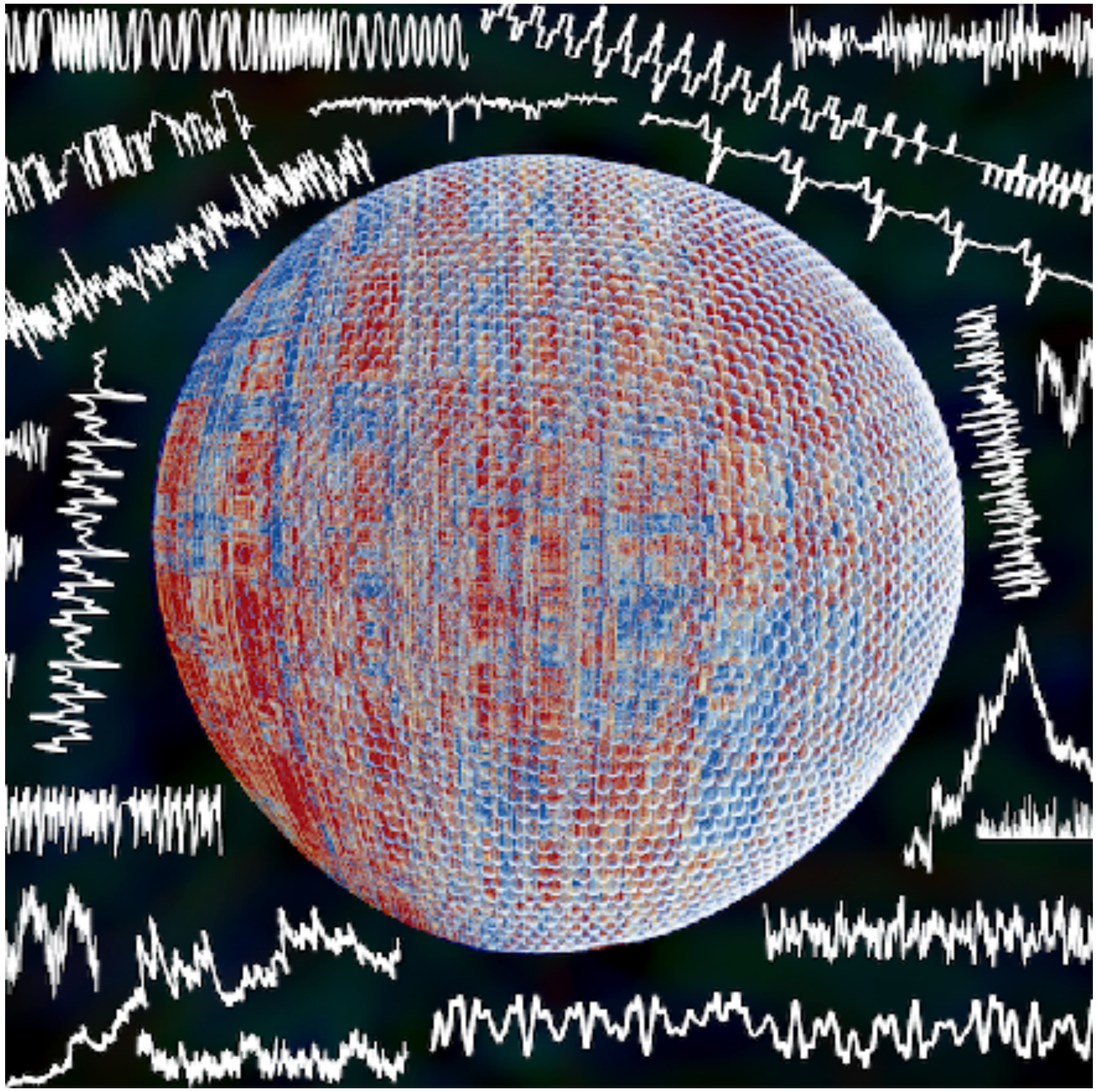
We should be focusing our efforts on developing methods for the types of data we see in the world — are we applying the right types of methods to our data? Are we studying sufficiently diverse data? Can we generate new types of time series?

Can we leverage large scientific feature libraries to learn more about the *processes* underlying a time-series dataset?

Can we apply similar methods developed (lessons learned) for time series/sequential to other data objects? — multivariate time series, unevenly-sampled time series, complex networks, point clouds, images, etc.

ACKNOWLEDGEMENTS

ben.fulcher@sydney.edu.au
@bendfulcher
@compTimeSeries
benfulcher
www.benfulcher.com
www.comp-engine.org/timeseries



THE UNIVERSITY OF SYDNEY

Nick Jones
Imperial College London

Key references:

- **Fulcher, B. D. (2017) Feature-based time-series analysis, *arXiv* 1709.08055 [+ *Feature Engineering,* CRC Press (2018)]**
- Fulcher, B. D., & Jones, N. S. (2017). *hctsa*: A Computational Framework for Automated Time-Series Phenotyping Using Massive Feature Extraction. *Cell Systems*, **5**, 527–531.
- Fulcher, B. D. & Jones, N. S. (2014) Highly comparative feature-based time-series classification. *IEEE Trans. Knowl. Data Eng.* **26,** 3026.
- Fulcher, B. D., Little, M. A. & Jones, N. S. (2013) Highly comparative time-series analysis: the empirical structure of time series and their methods. *J. Roy. Soc. Interface* **10,** 20130048.
- Christ, M., Fulcher, B. D., Kempa-Liehr, A. W. Regression of Exogenous Variables from Time Series. *arXiv* (~next week).
- Sethi, S. S., Zerbi, V., Wenderoth, N., Fornito, A., & Fulcher, B. D. (2017). Structural connectome topology relates to regional BOLD signal dynamics in the mouse brain. *Chaos* **27**, 047405.
- Fulcher, B. D., Georgieva, A. E., Redman, C. W. G. & Jones, N. S. (2012) Highly comparative fetal heart rate analysis. *2012 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 3135.