Highly comparative time-series analysis

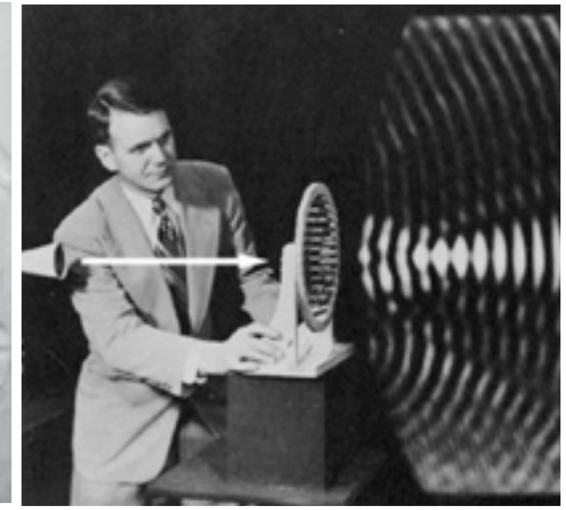
Paris School of Economics, Nov 2015

Max Little,
Aston University

Nick Jones,
Imperial College

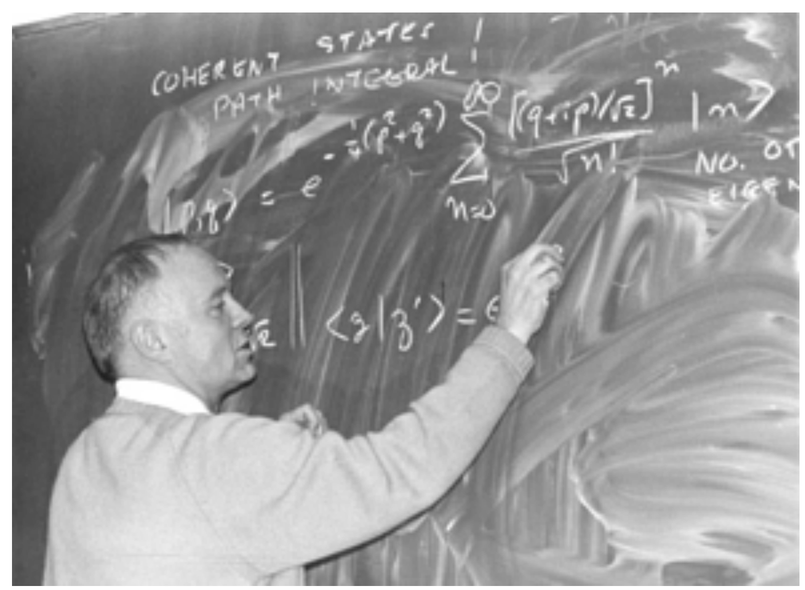
The art of time-series analysis

1. Measure data



---

2. Inspect data thoroughly, talk to domain experts, and manually devise appropriate models and methods based on intuition and experience

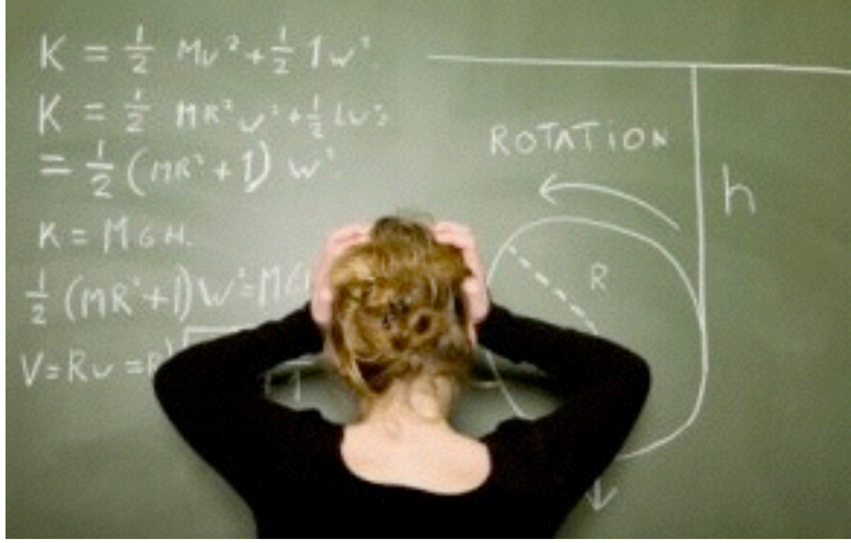


"Do what I did during my PhD"

"Use standard analysis methods from my field"

"Apply a hot new method I read about in Nature"

Worries



- 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)?
- Papers introducing a new method compare their method to an average of 0.91 others, and 1.85 different datasets*.

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

Competing interdisciplinary approaches/opinions

vast and growing volumes of data and methods

“I know someone smart who uses wavelets”

“Everyone knows you can’t apply AR time-series models to nonstationary biomedical data!”

“ARIMA models are a waste of time”

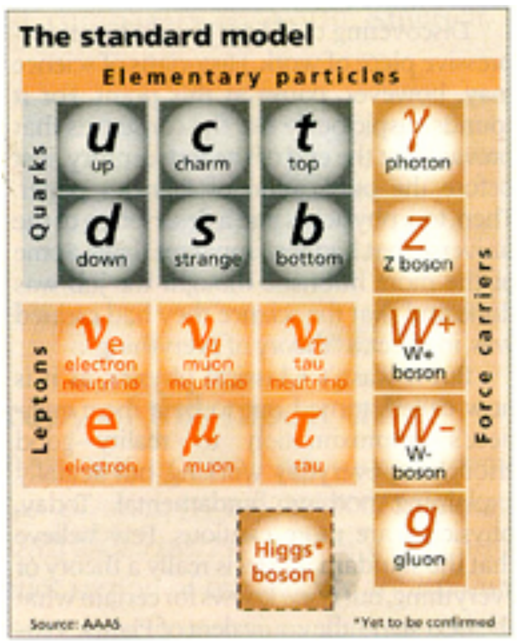
Structuring

Scientific endeavors often focus on structuring libraries of collected information.

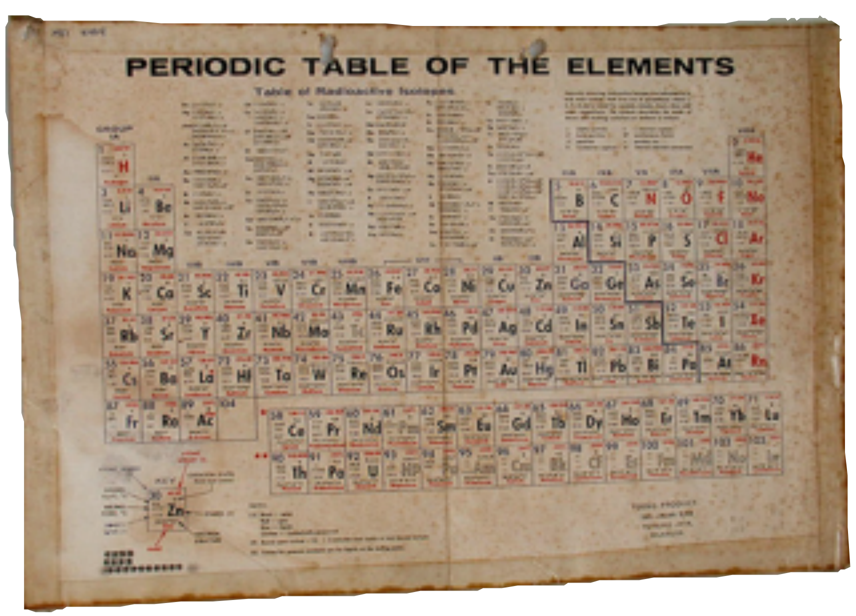
Helps us to understand the complexity in the world

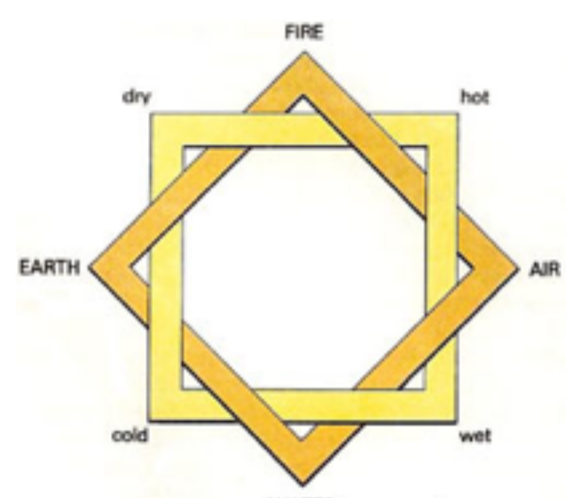
Food

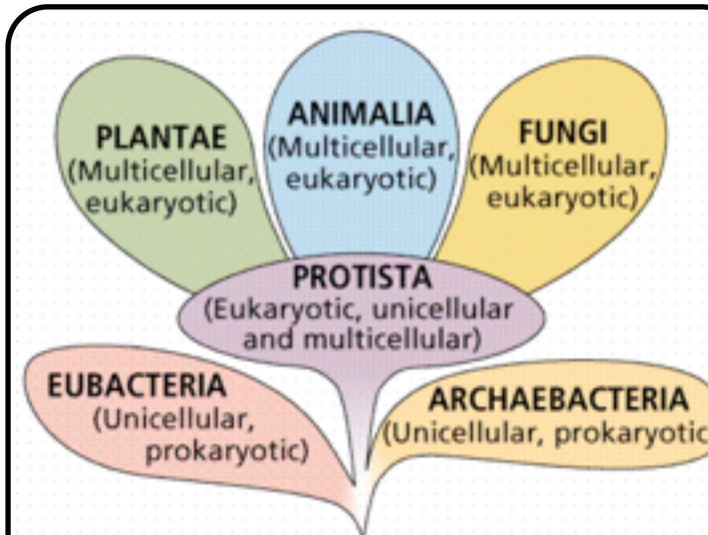
Matter



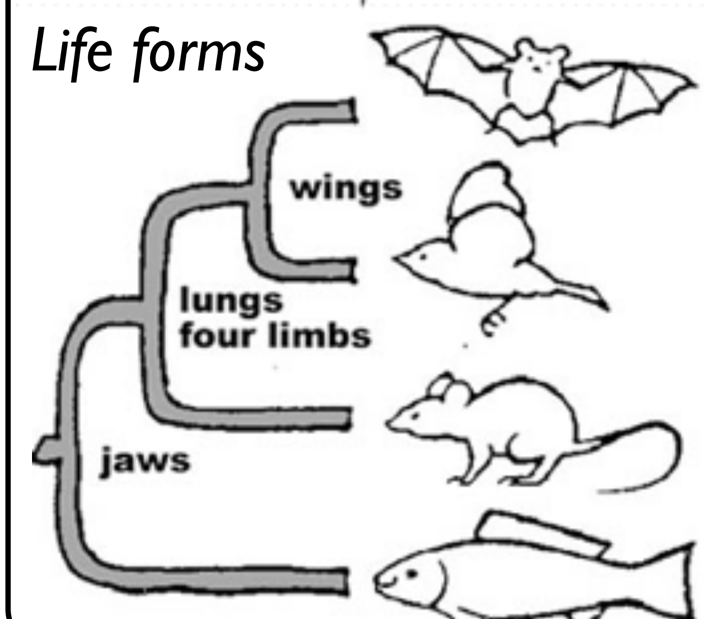
Substances

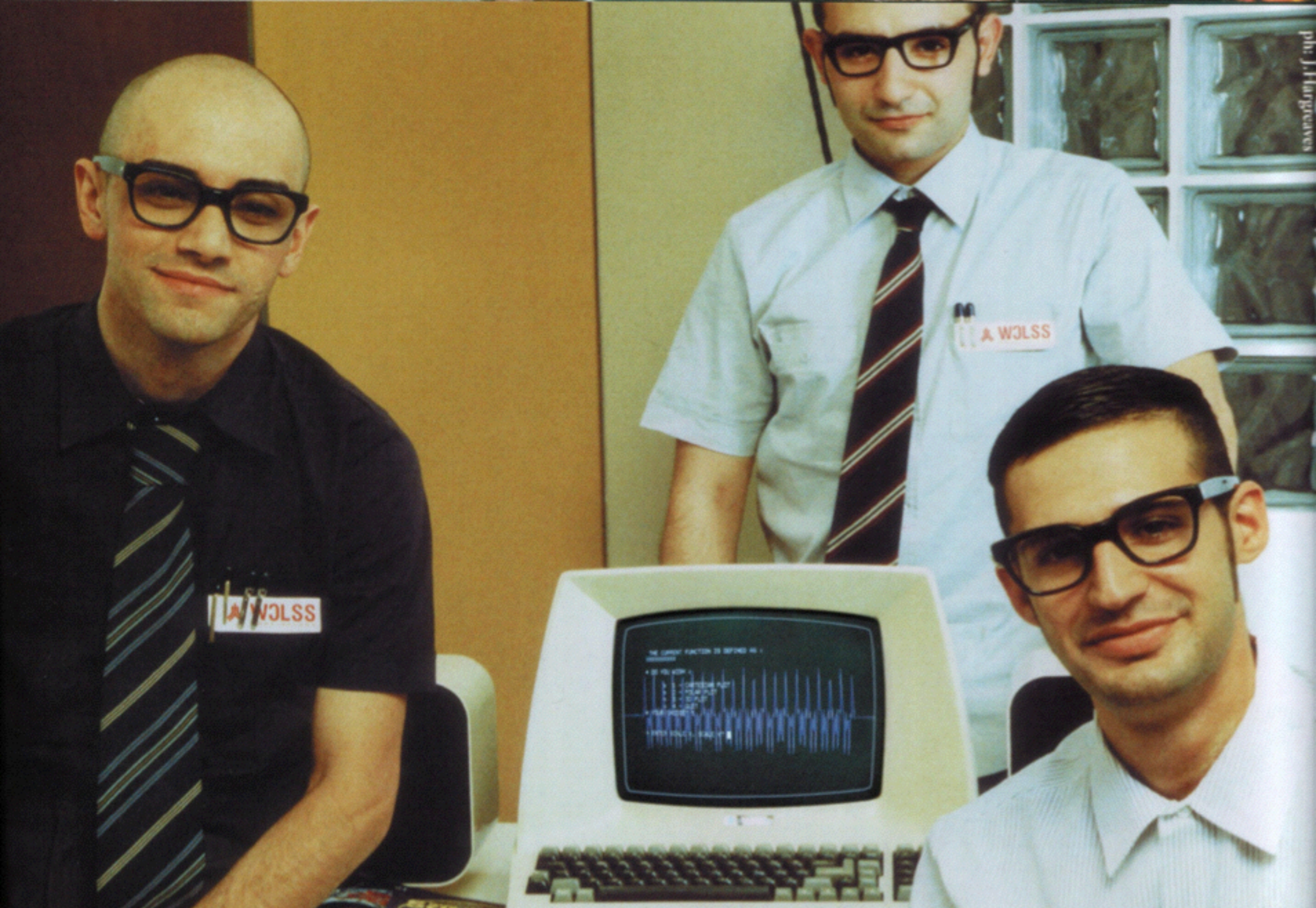






Life forms





ph: J.Hargreaves

LET COMPUTERS DO IT !

First we collect

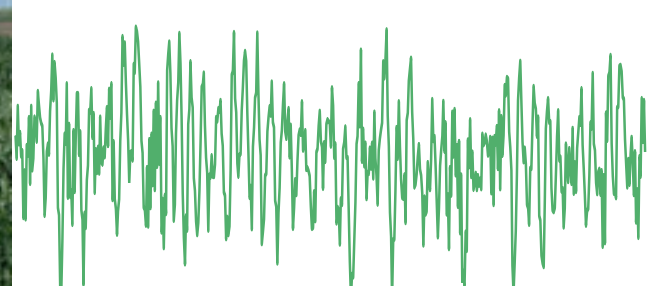
e.g.:

Frolicking in field with net

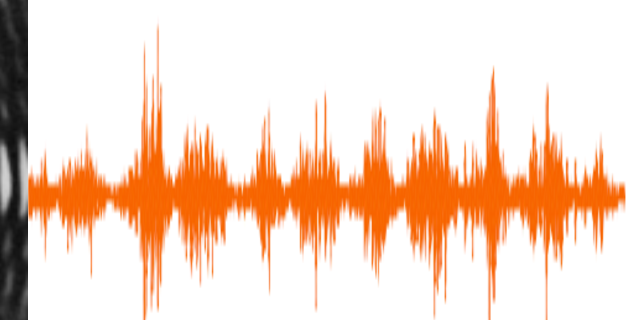
before you know it

Butterfly collection

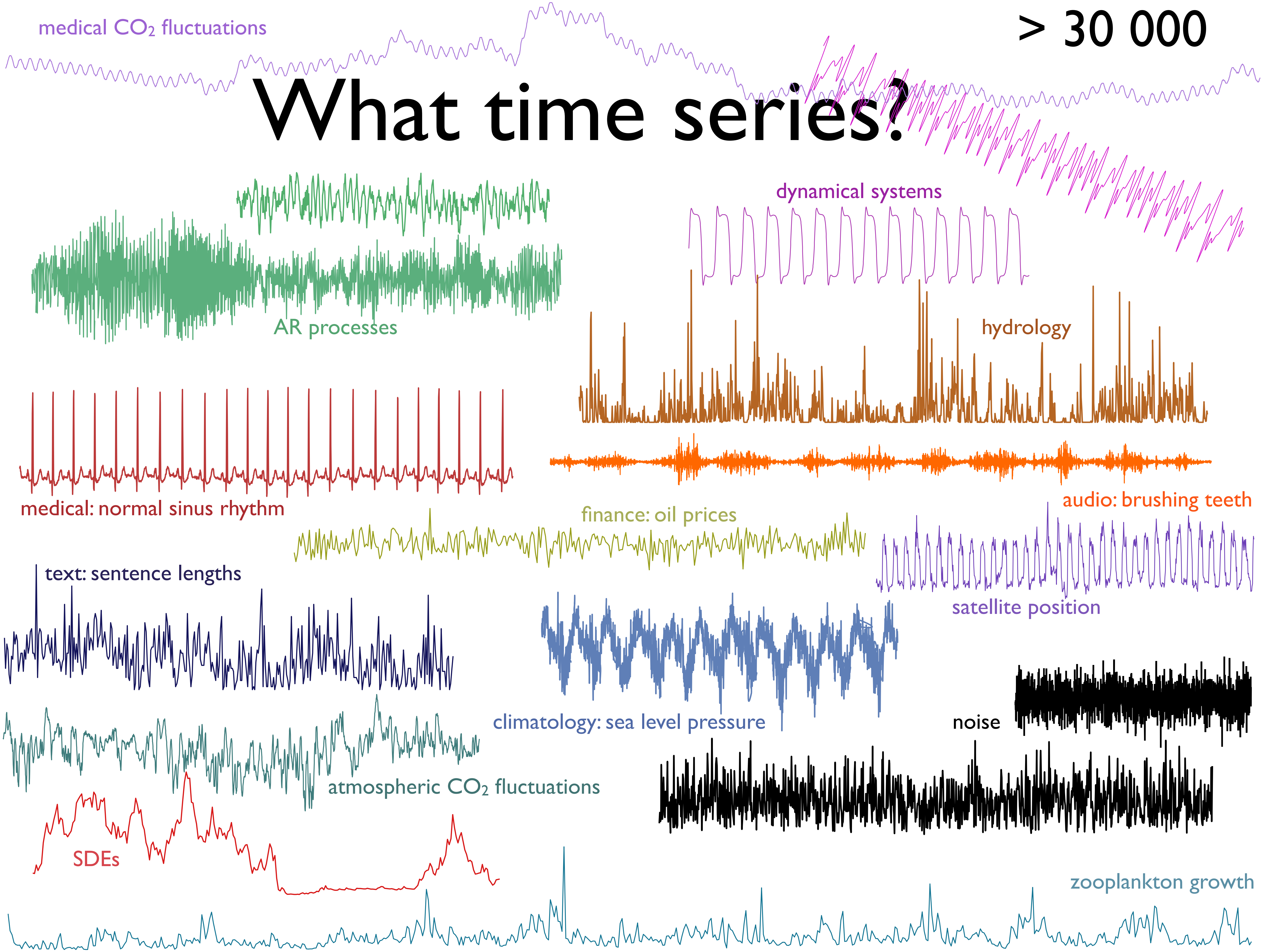
Less fun, but also important: What about our data? What about our methods?



This dude is measuring crops



This dude is measuring sound waves



> 30 000
What time series?
medical CO_2 fluctuations
dynamical systems
AR processes
hydrology
medical: normal sinus rhythm
audio: brushing teeth
finance: oil prices
satellite position
text: sentence lengths
climatology: sea level pressure
noise
atmospheric CO_2 fluctuations
SDEs
zooplankton growth

Analysis methods

Static distribution

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

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

(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

Stationarity

StatAv
Sliding window measures
Bootstraps
Step detection
Distribution comparisons

Model fits

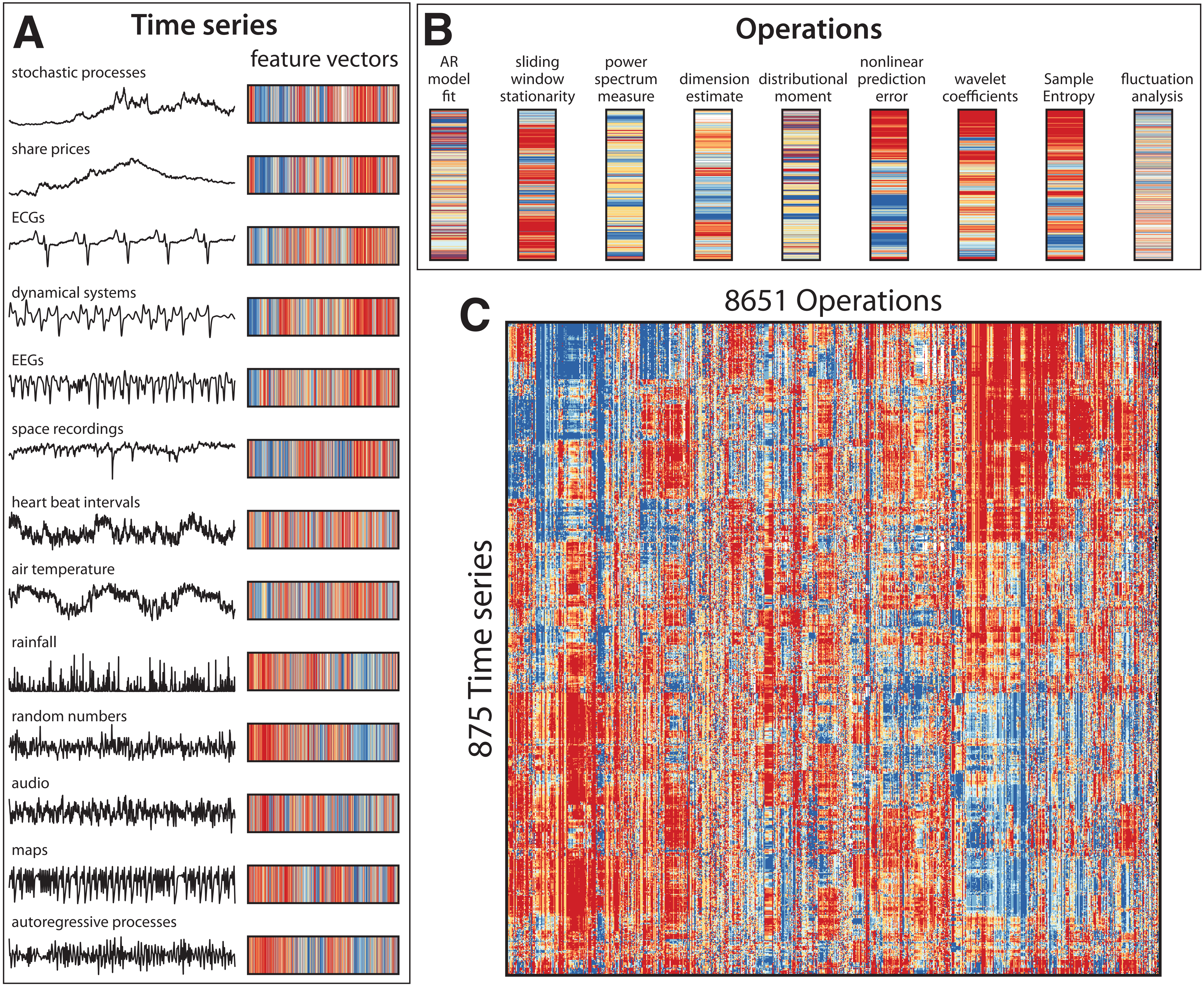
Local prediction
GARCH models
Fourier fits
AR models
Exponential smoothing
State space models
Hidden Markov models
Biased walker simulations
Piecewise splines
ARMA models
Gaussian Processes

Others

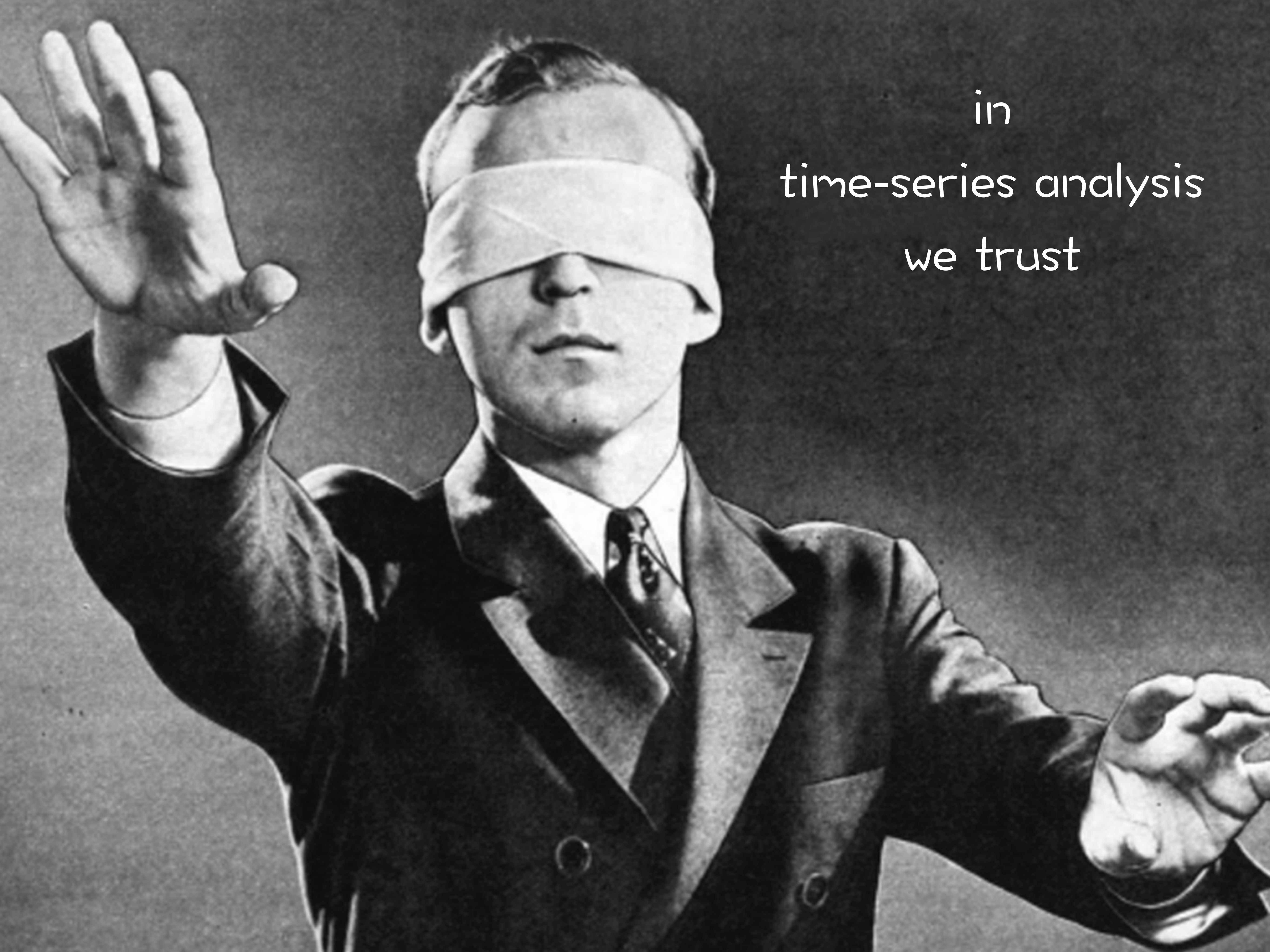
Transition matrices
Local motifs
Dynamical system coupling
Visibility graph
Stick angle distribution
Extreme events
Singular spectrum analysis
Domain-specific techniques

Basis Functions

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

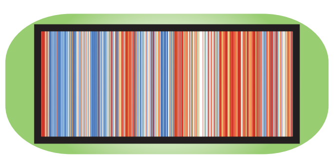

A
Time series
feature vectors
stochastic processes
share prices
ECGs
dynamical systems
EEGs
space recordings
heart beat intervals
air temperature
rainfall
random numbers
audio
maps
autoregressive processes
B
Operations
AR model fit
sliding window stationarity
power spectrum measure
dimension estimate
distributional moment
nonlinear prediction error
wavelet coefficients
Sample Entropy
fluctuation analysis
C
8651 Operations
875 Time series


Time-series analysis 101:
always look at your data


in
time-series analysis
we trust

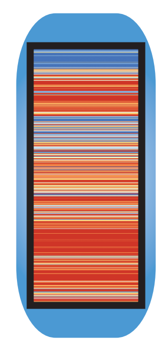
Empirical fingerprints

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



= time series of type 'green'

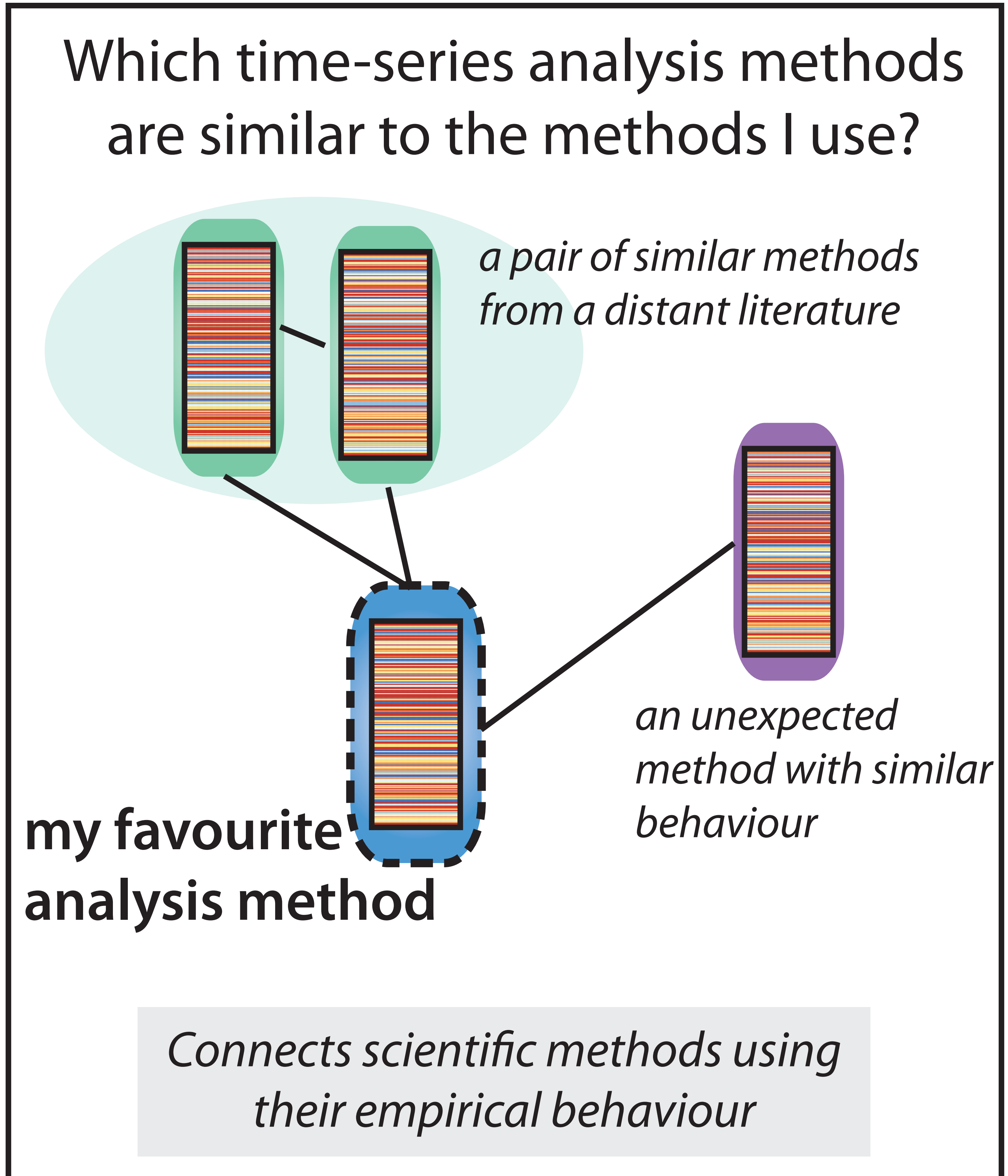
captures properties measured by diverse scientific methods



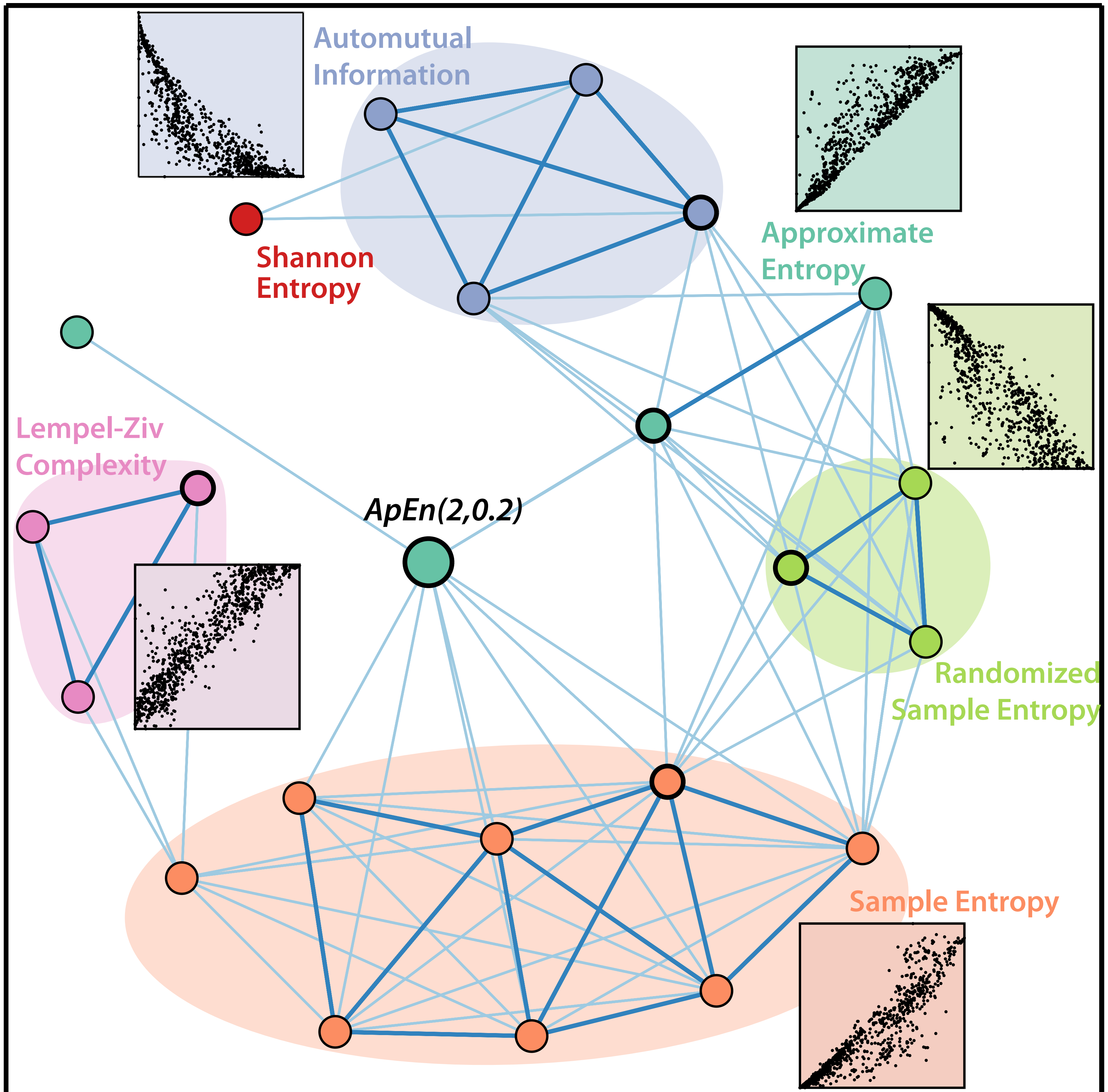
= operation of type 'blue'

captures behaviour across a range of empirical time series

Organizing our methods

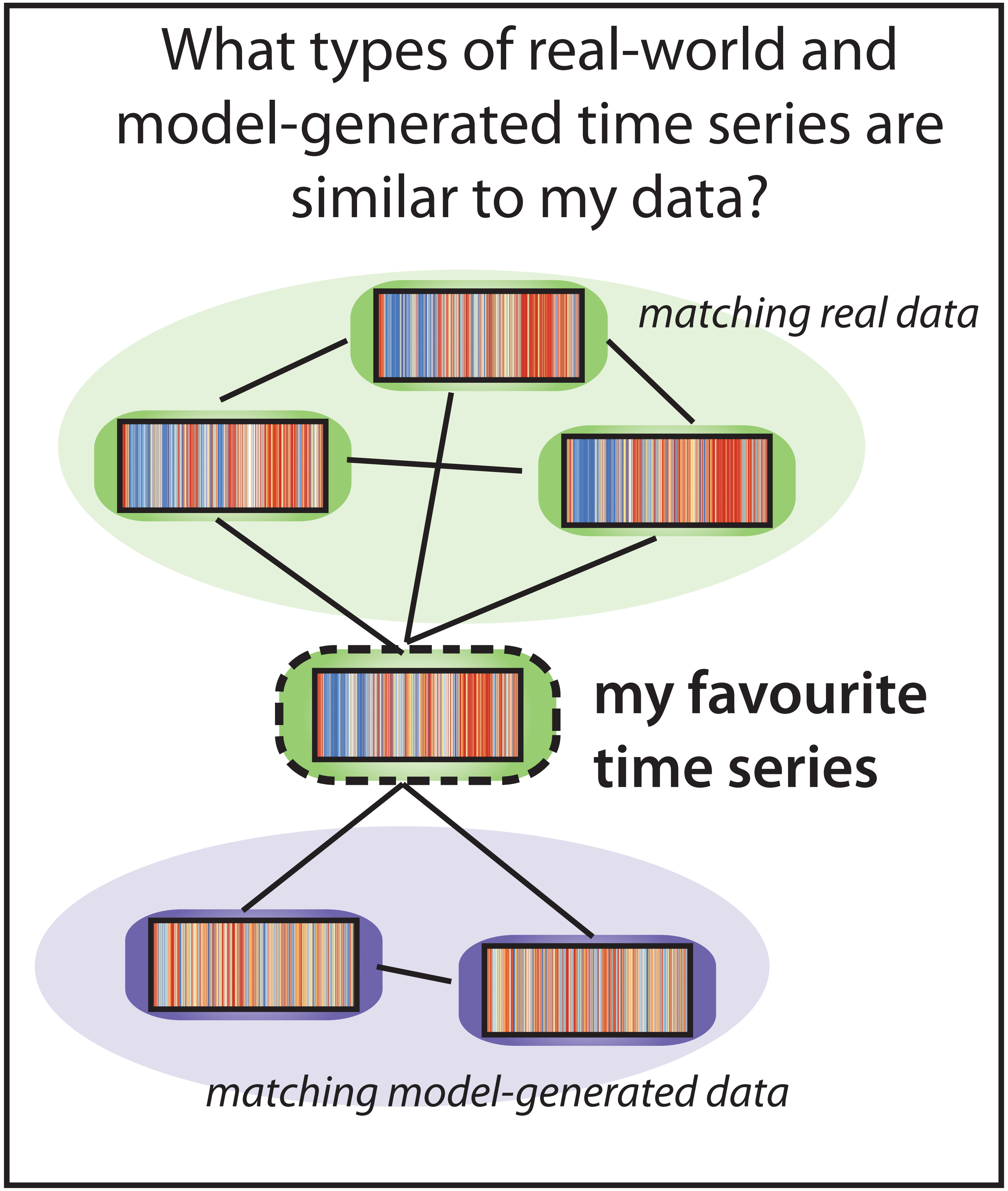


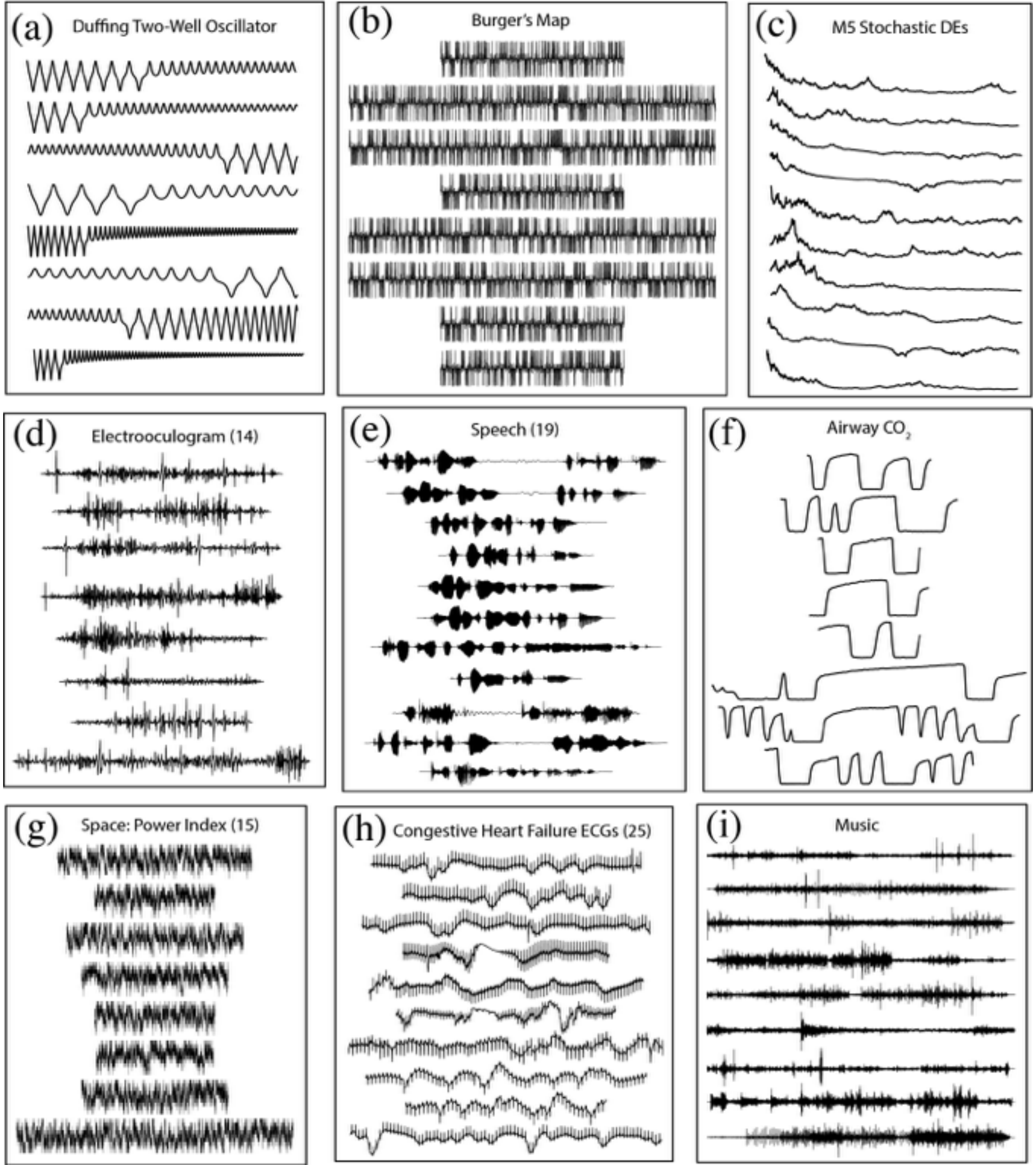
Local neighborhoods



Automatically find interdisciplinary connections between our methods for time-series analysis

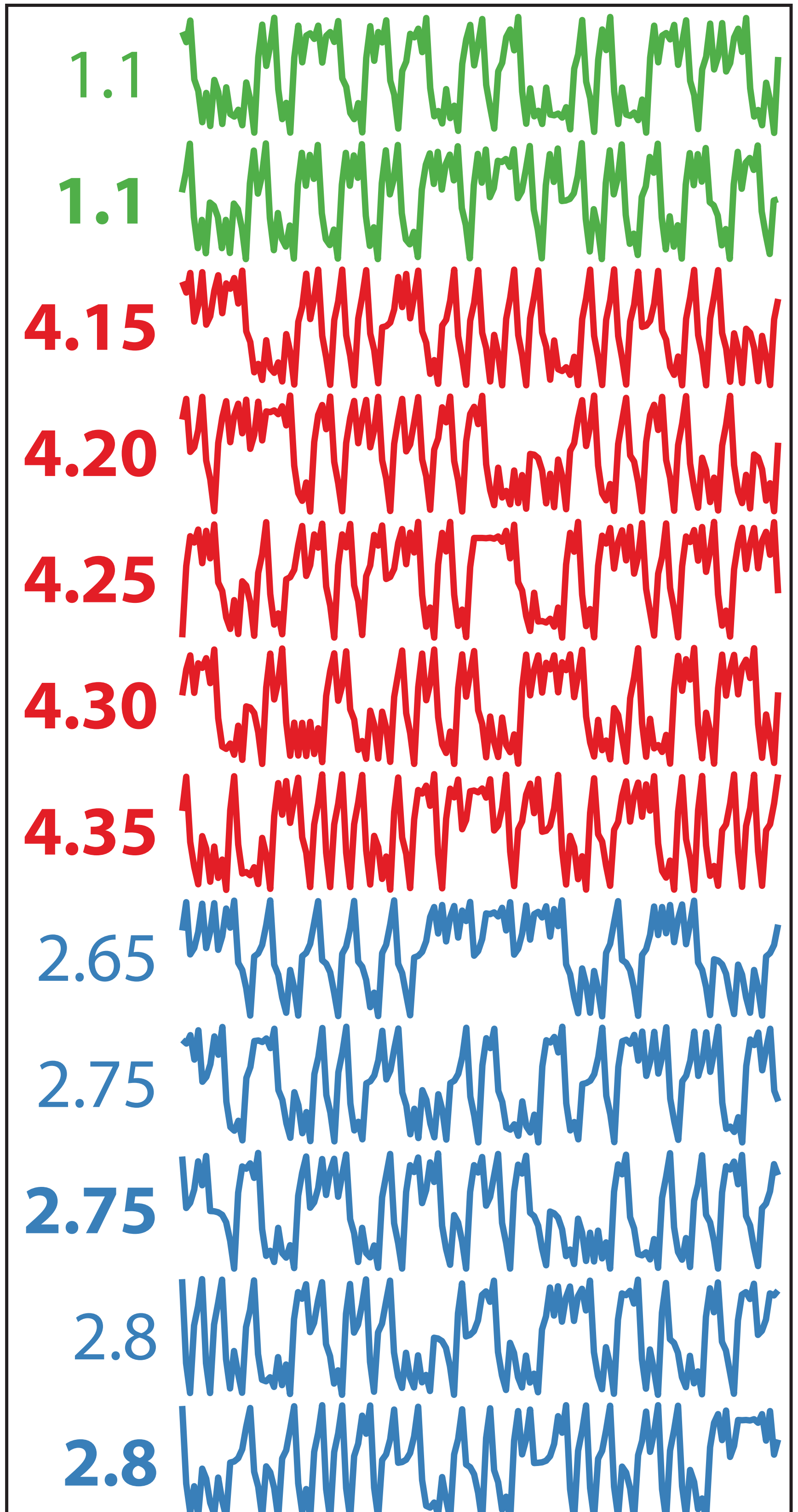
Organizing our data





Clusters of time series group systems with common dynamics

A time-series cluster:



Sine map

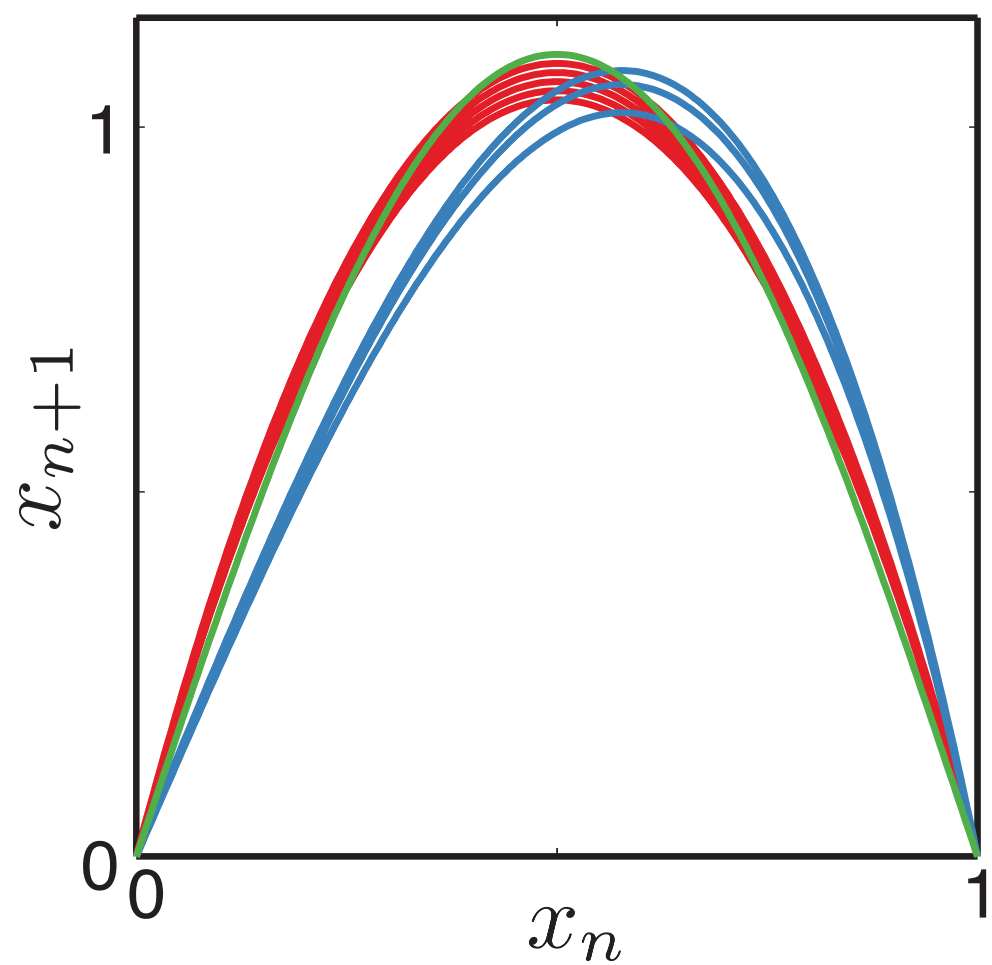
$$x_{n+1} = A\sin(\pi x_n)$$

Asymmetric Logistic map

$$x_{n+1} = Ax_n(1-|x_n|)$$

Cubic map

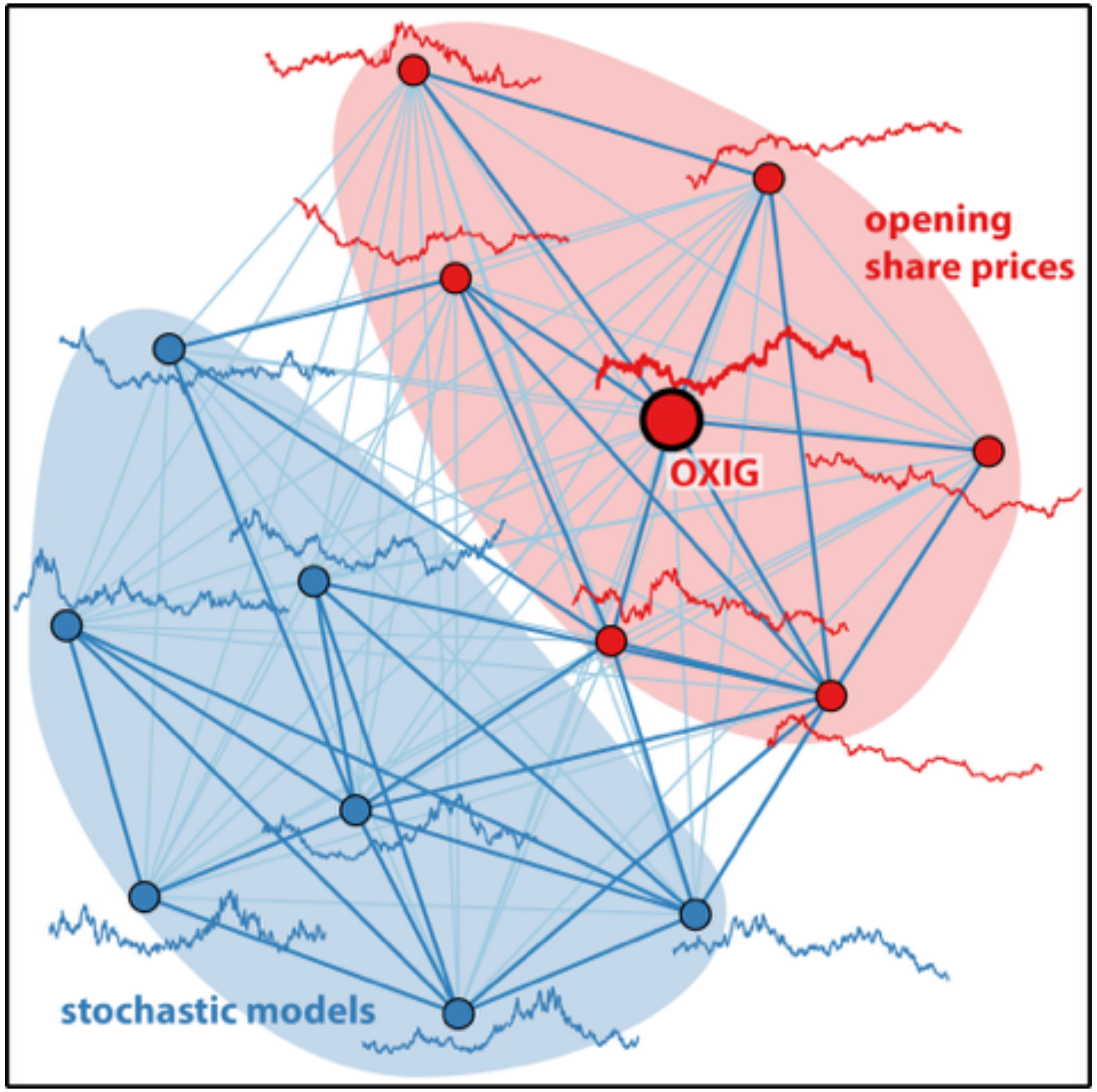
$$x_{n+1} = Ax_n(1-x_n^2)$$

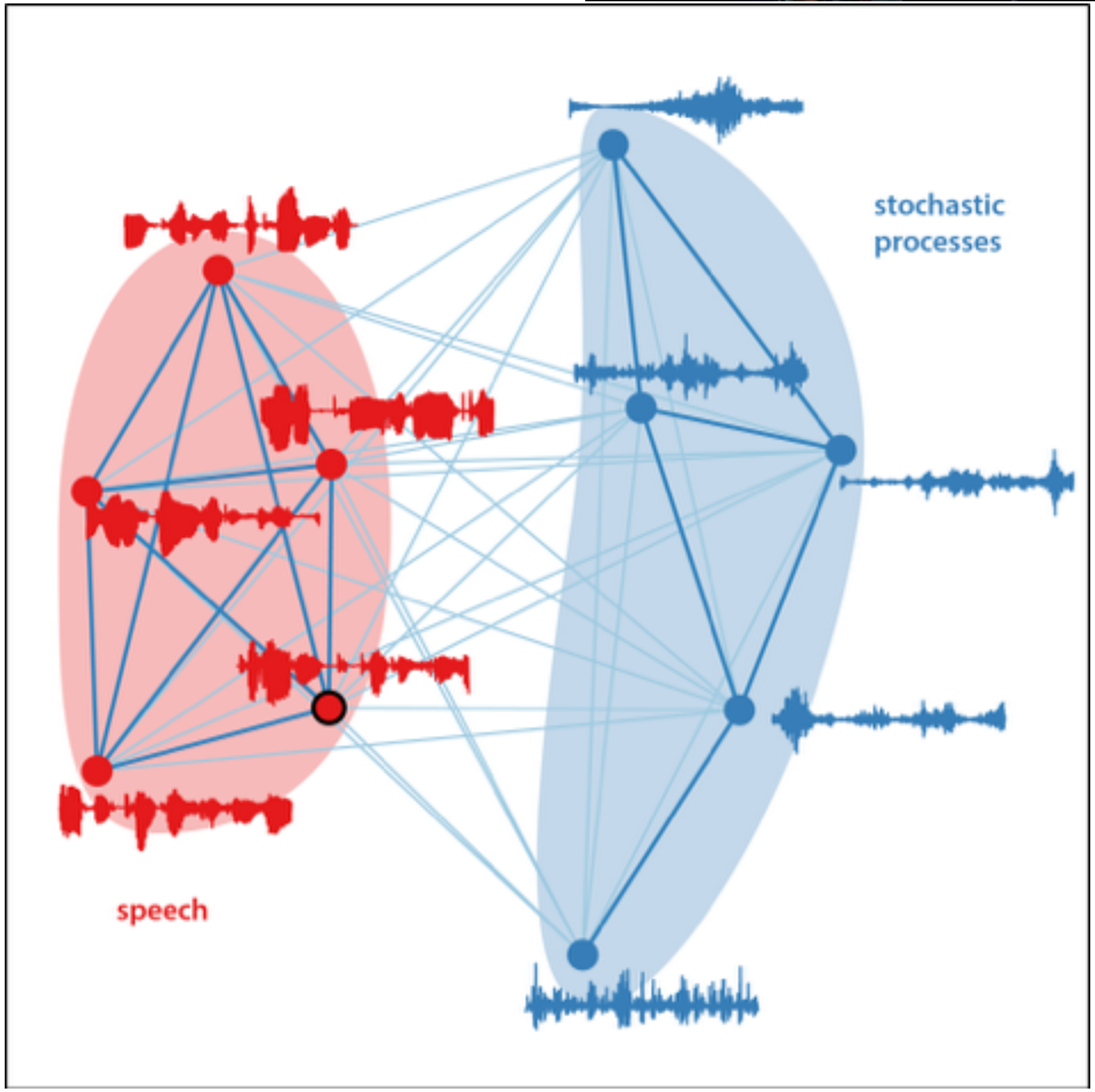


Fishing for data

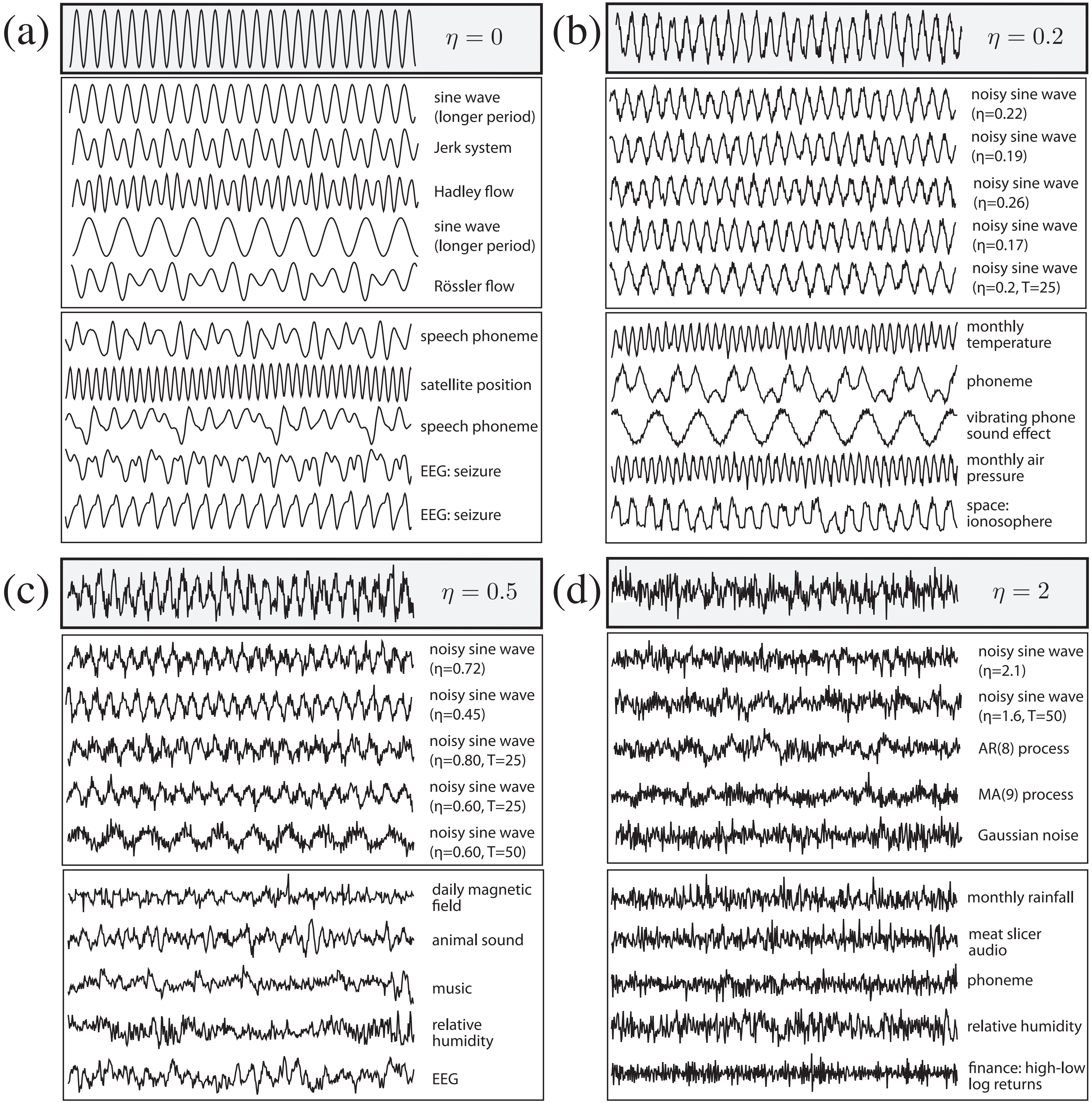
Can gain insights into your data by comparing it to a wealth of data collected in other areas of science

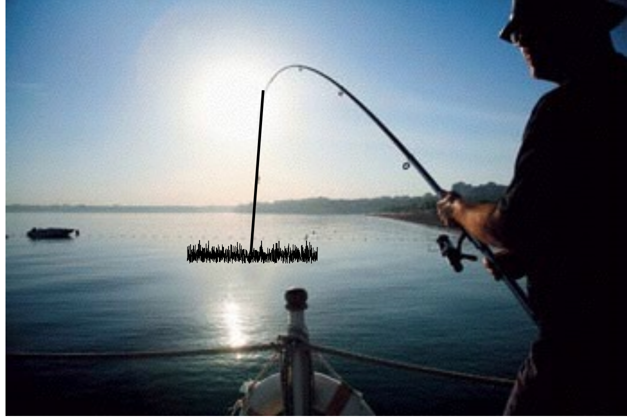
Fishing for data



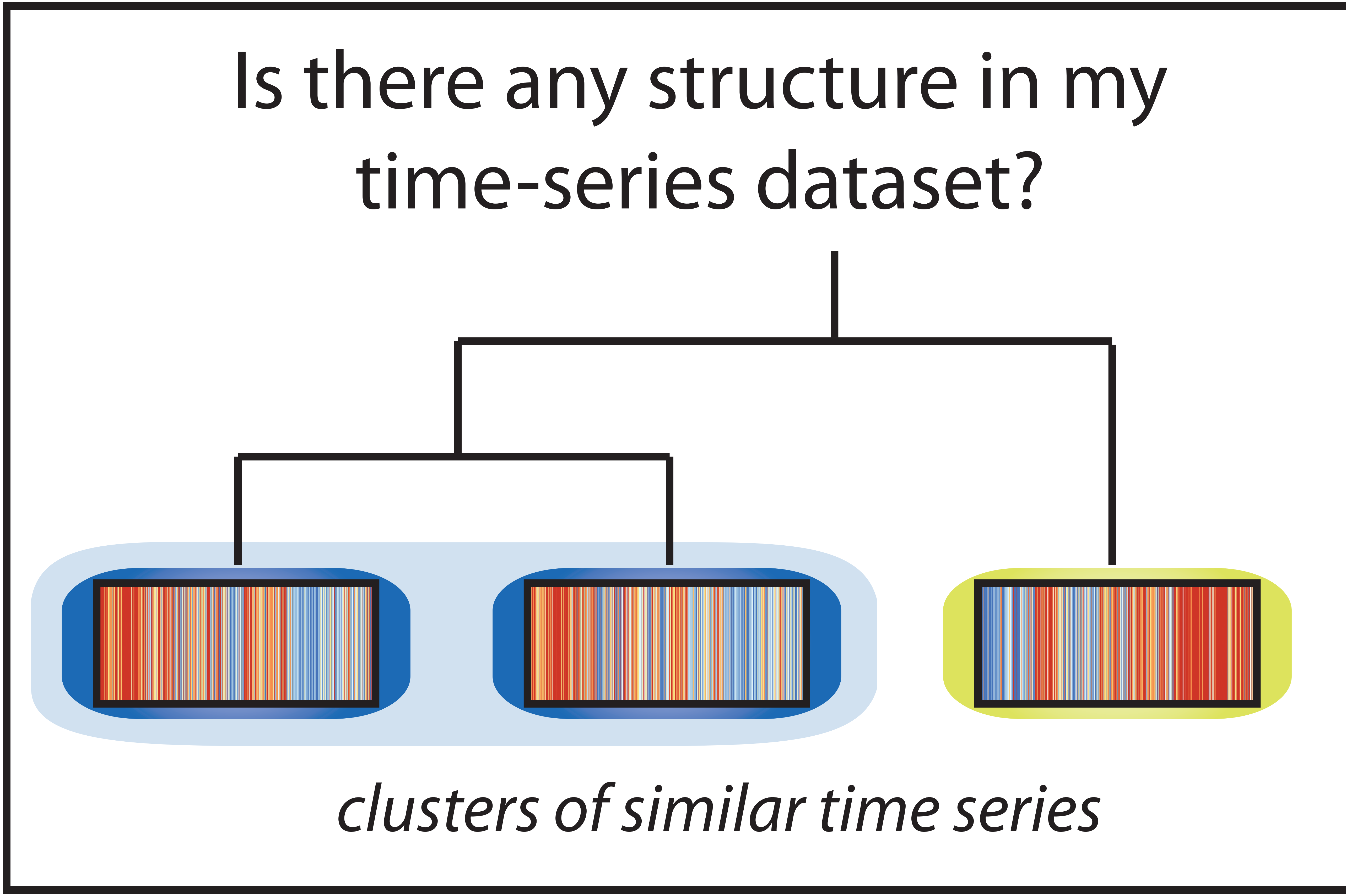


suggest models, or similar real-world processes to our data

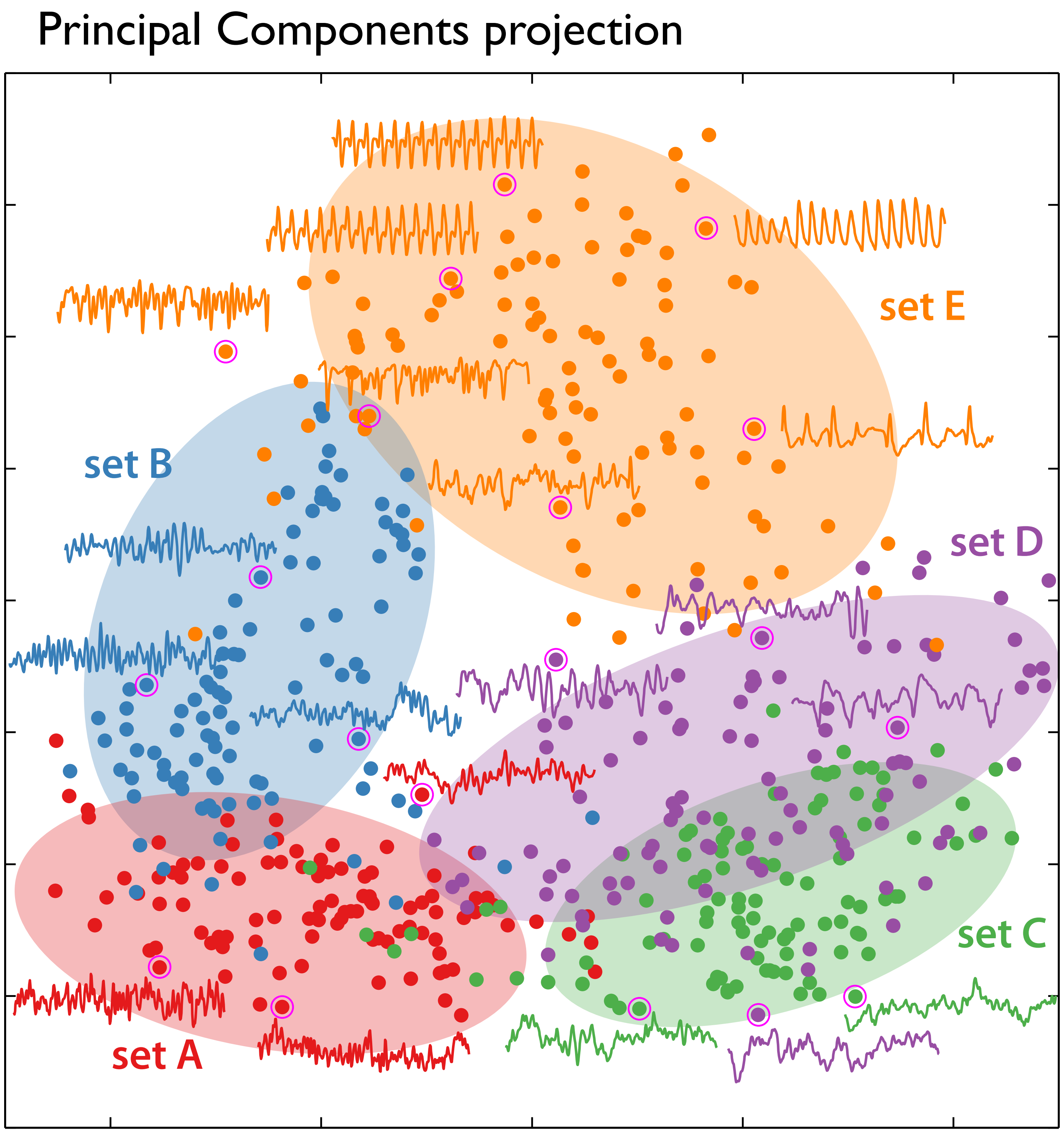


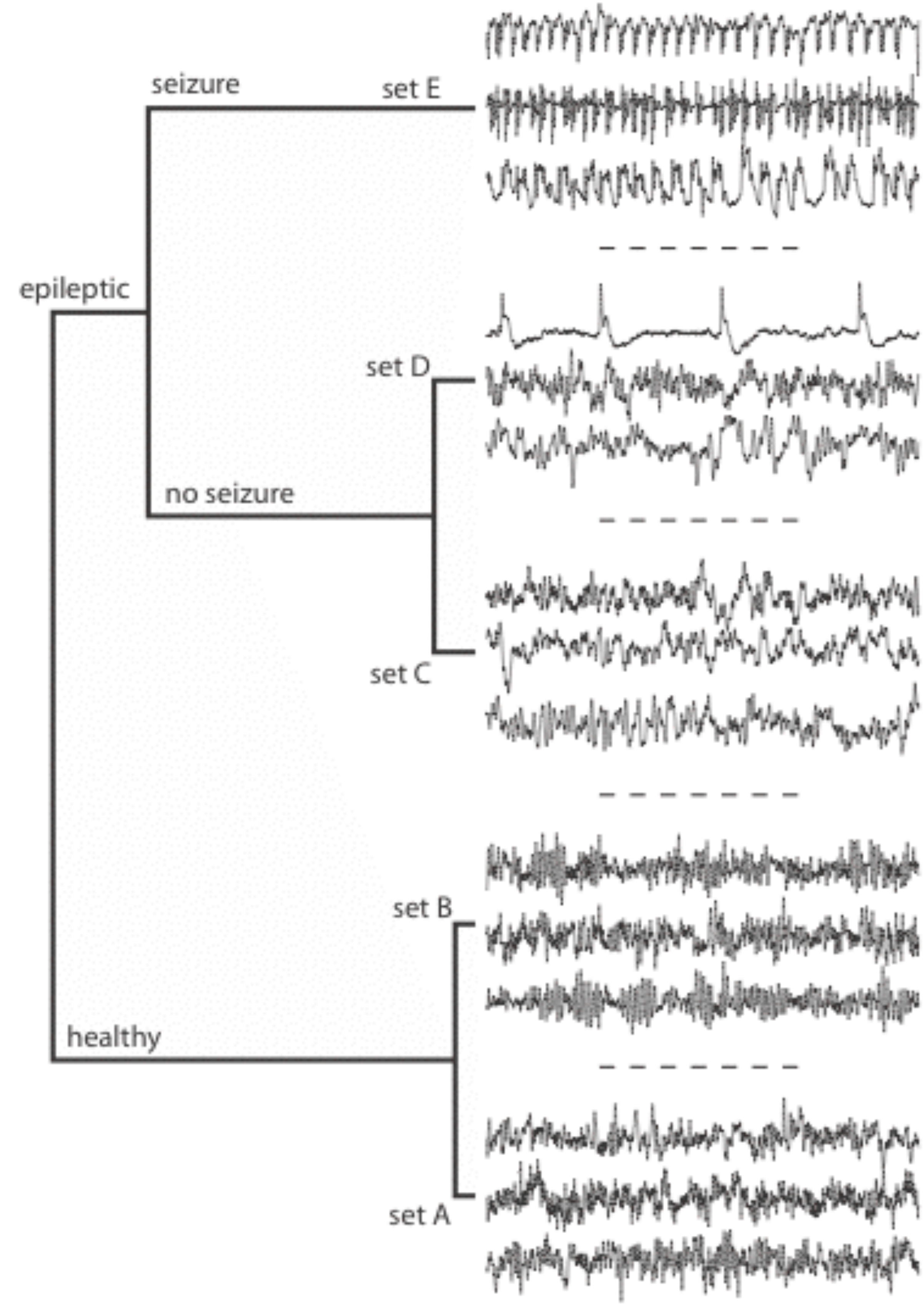


Brings our data and models closer together


Is there any structure in my time-series dataset?
clusters of similar time series

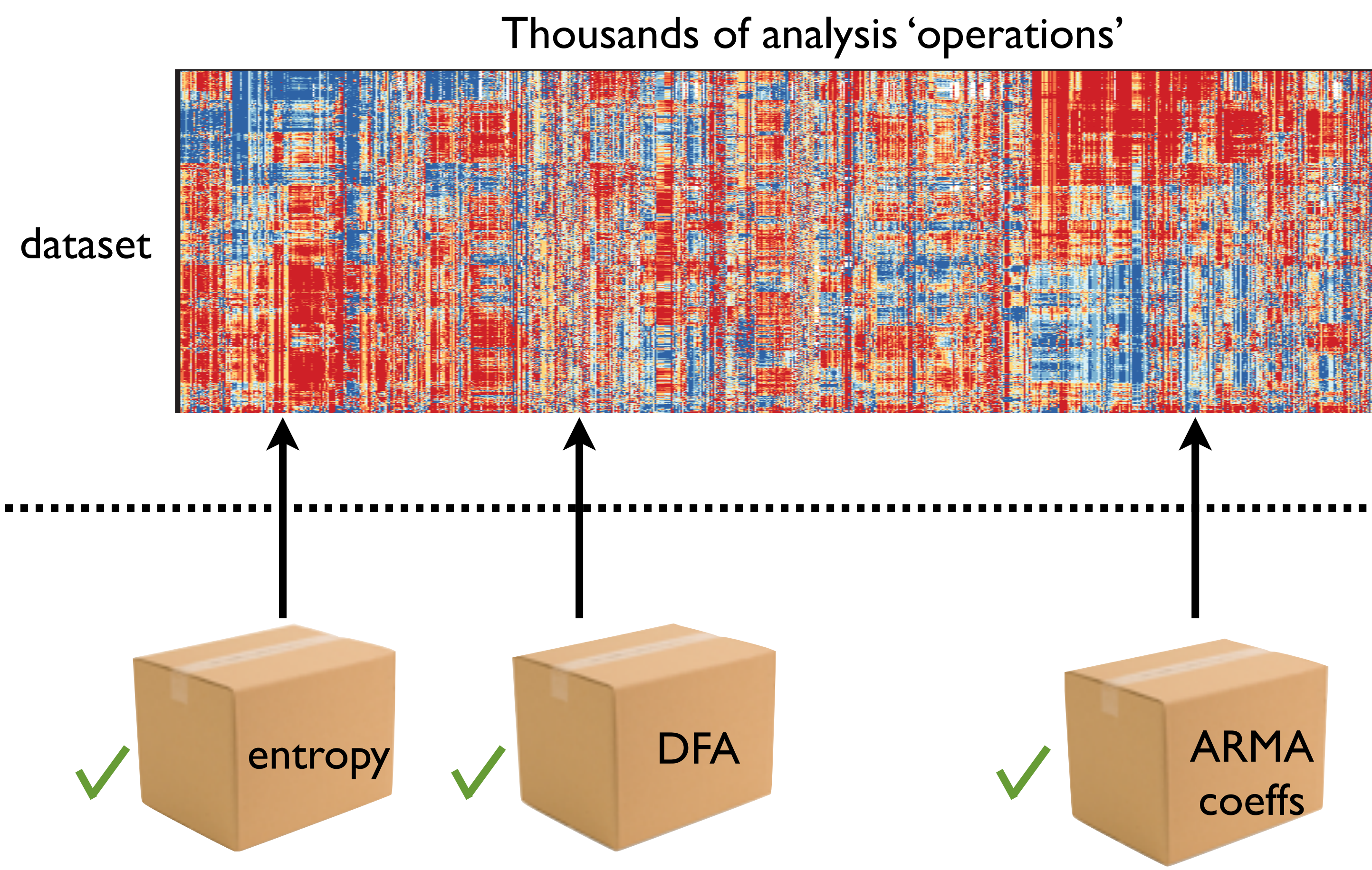
EEGs


Principal Components projection
set E
set B
set D
set C
set A


seizure
set E
epileptic
set D
no seizure
set C
set B
healthy
set A

Highly comparative time-series analysis

1. Compute and compare thousands of analysis methods



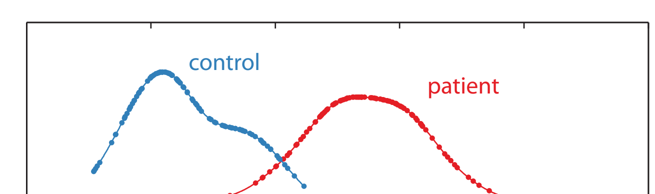
2. Select methods that perform well on your data

3. Interpret new methods to gain insights into your data

“Signals from the patient group are less predictable”

“Single neuron recordings from region X have more outliers and intermittent fluctuations”

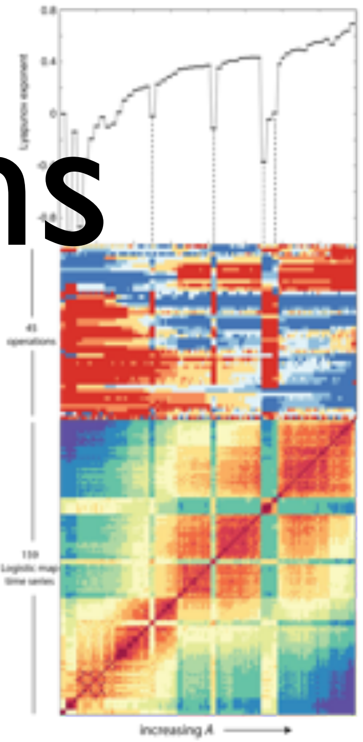
“…”

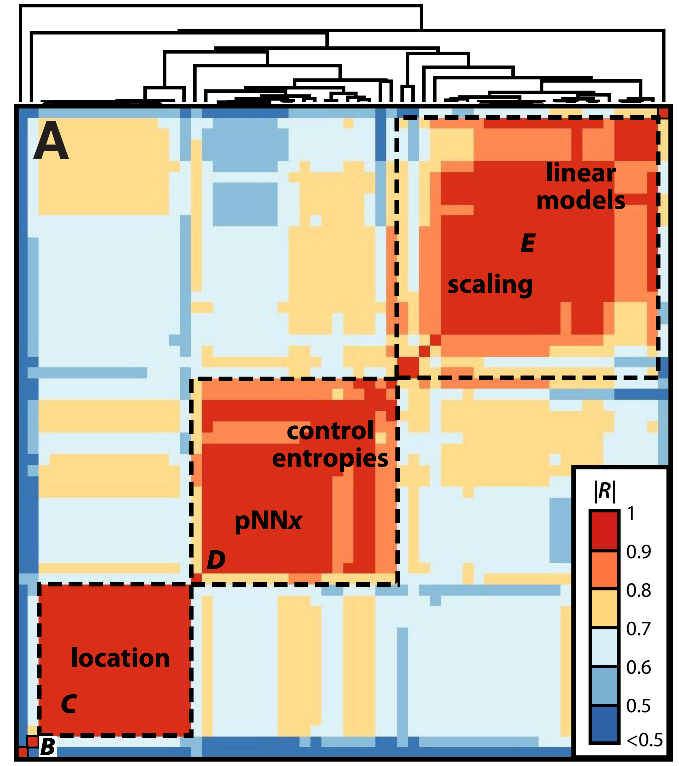


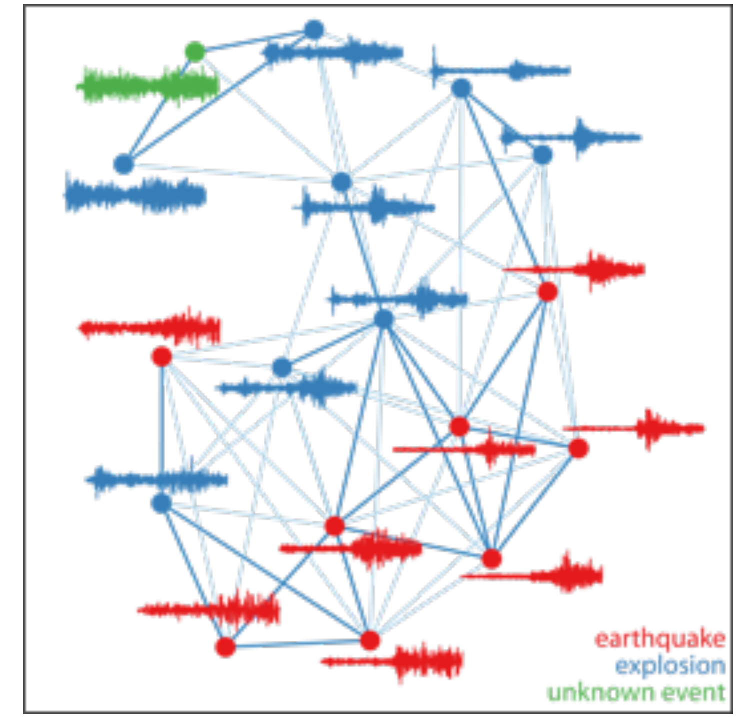
EEG entropy in pre-frontal cortex

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

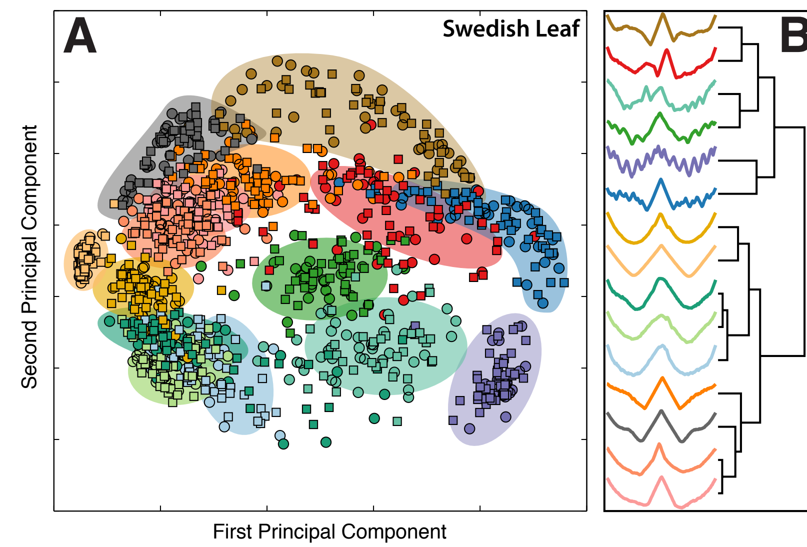
Applications

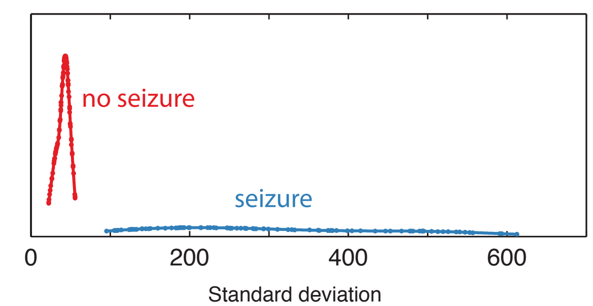


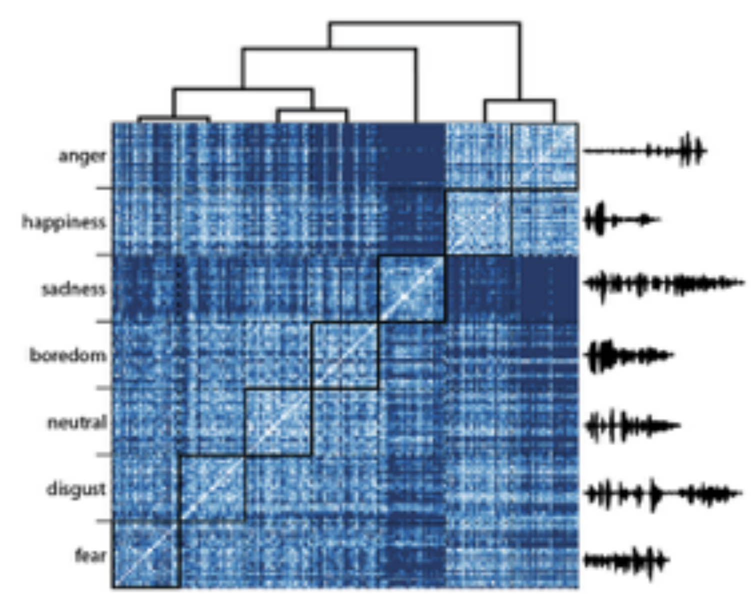


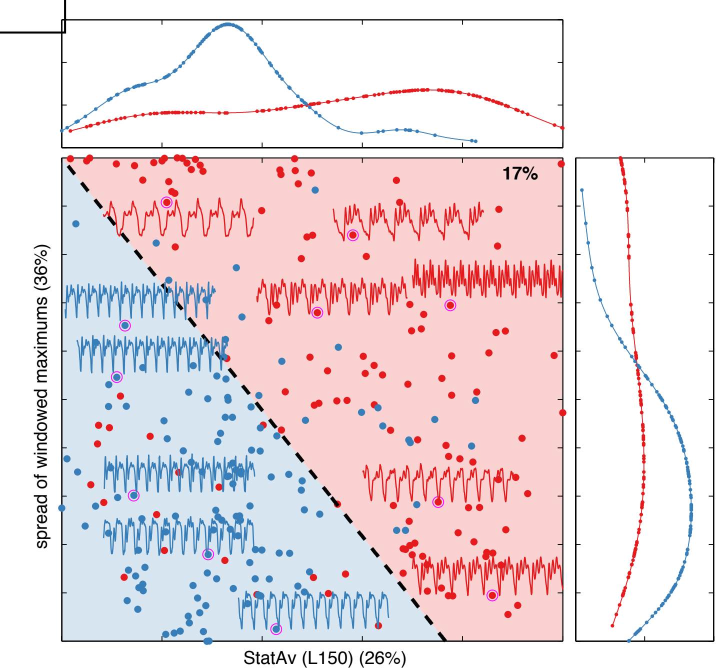


- Seismic data
- Simulated chaos
- Fetal heart rate
- Heart rate intervals
- Parkinsonian speech
- Epileptic EEGs
- Emotional speech







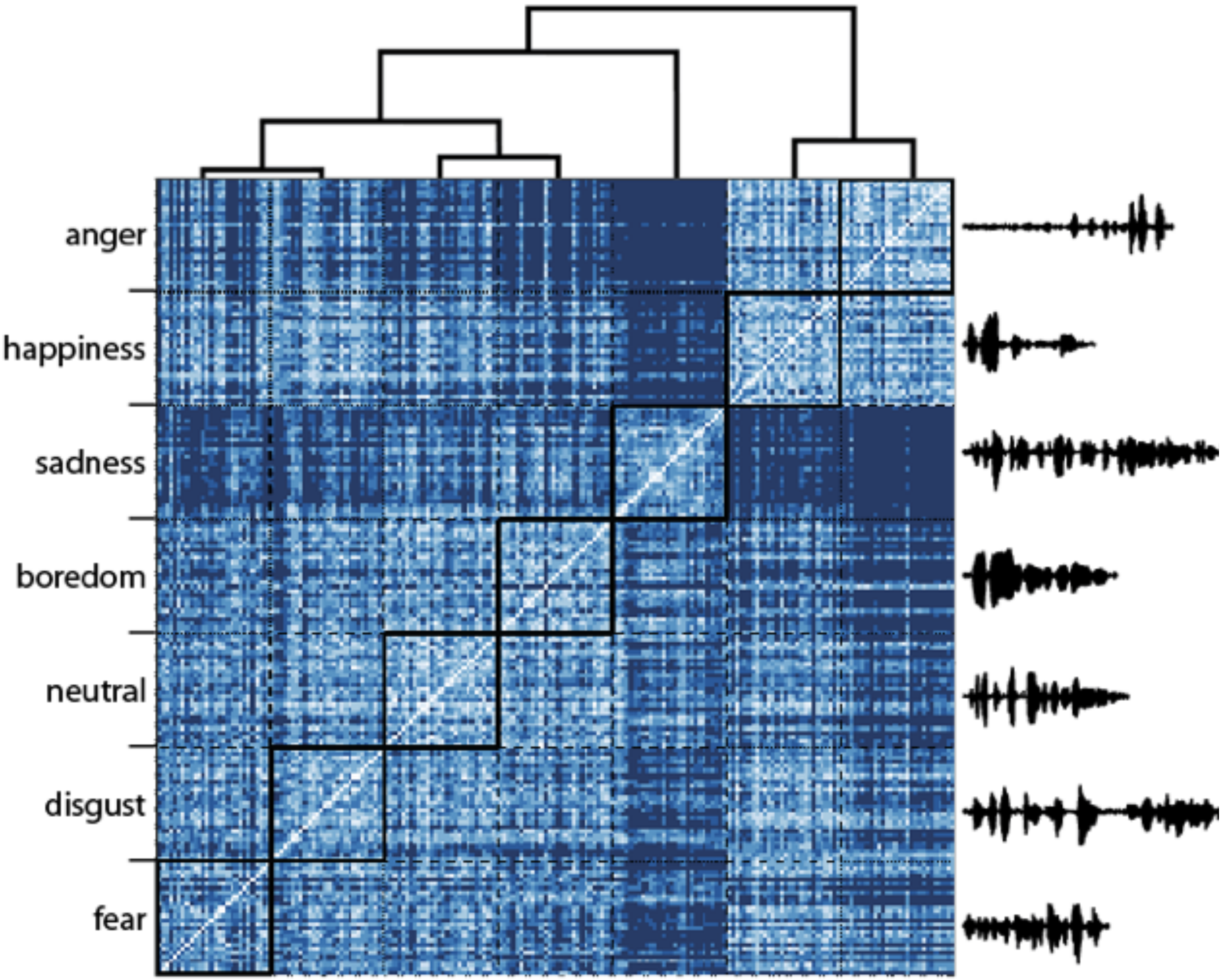


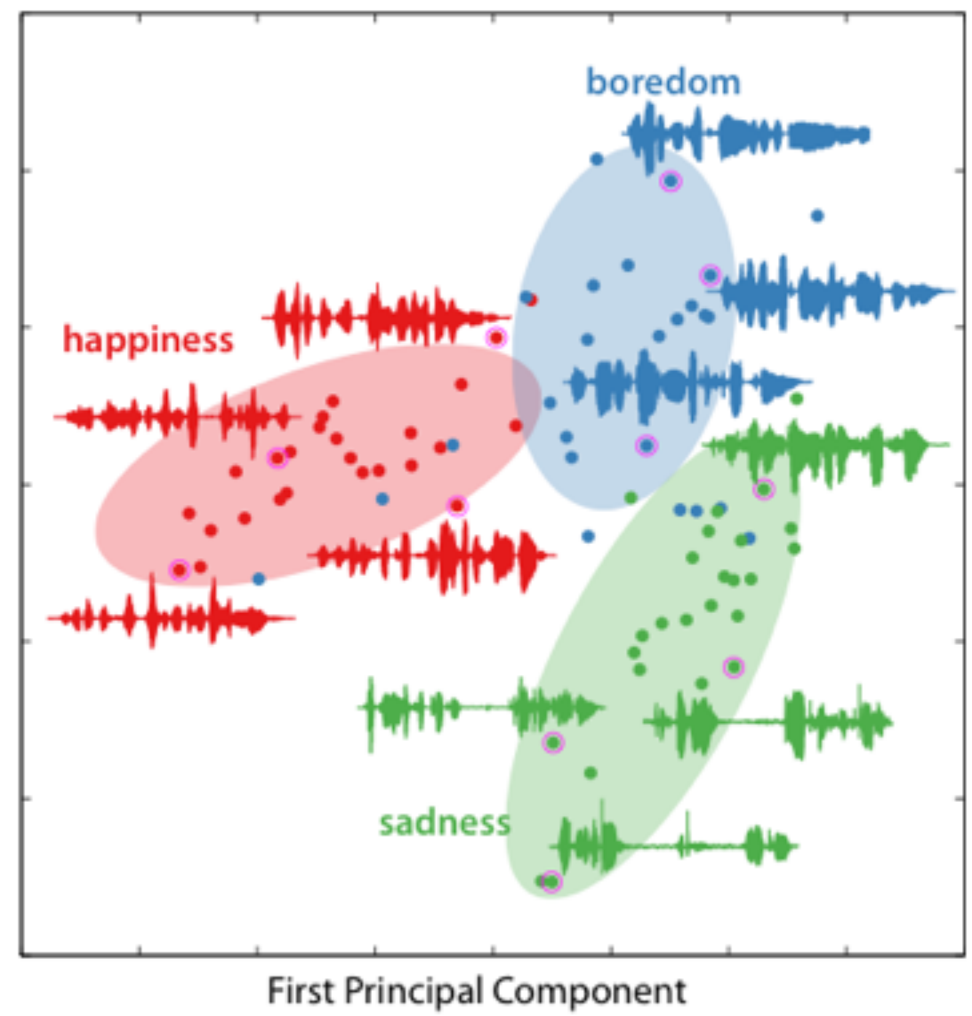
BD Fulcher, NS Jones. *IEEE KDE* (2014), *DOI: 10.1109/TKDE.2014.2316504*

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

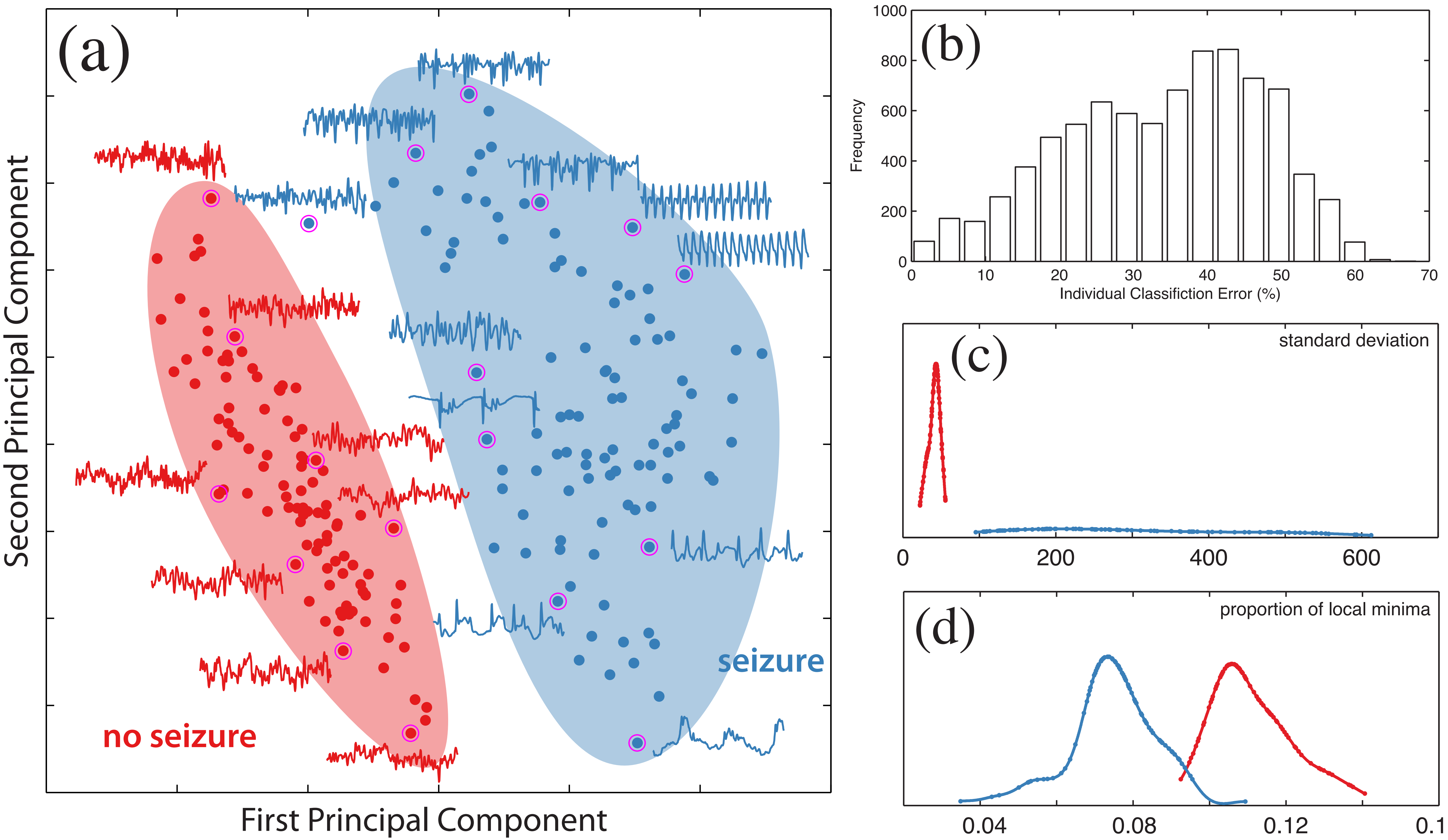
BD Fulcher, AE Georgieva, C Redman, NS Jones, Annual International Conference of the IEEE, EMBC, 3135 (2012), DOI: 10.1109/EMBC.2012.6346629

Identifying emotions in German speech

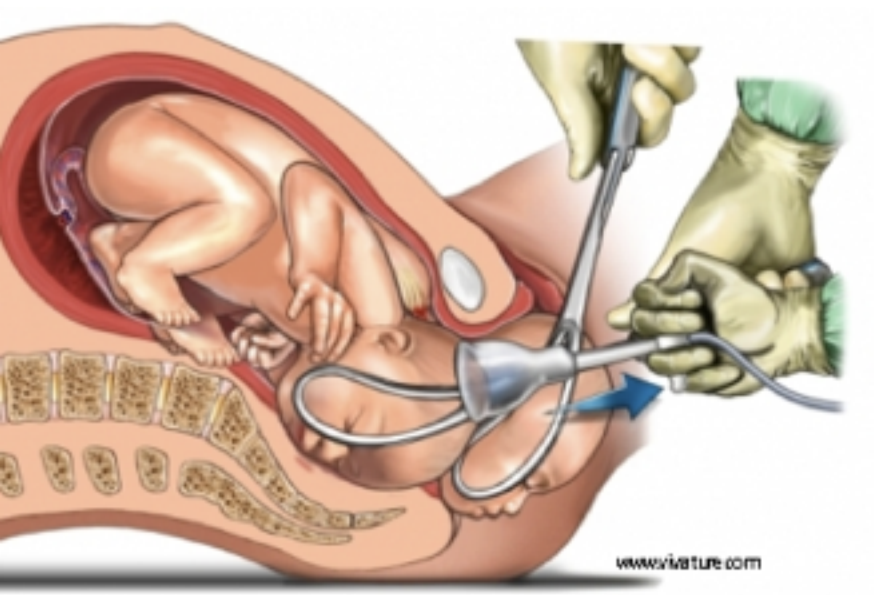




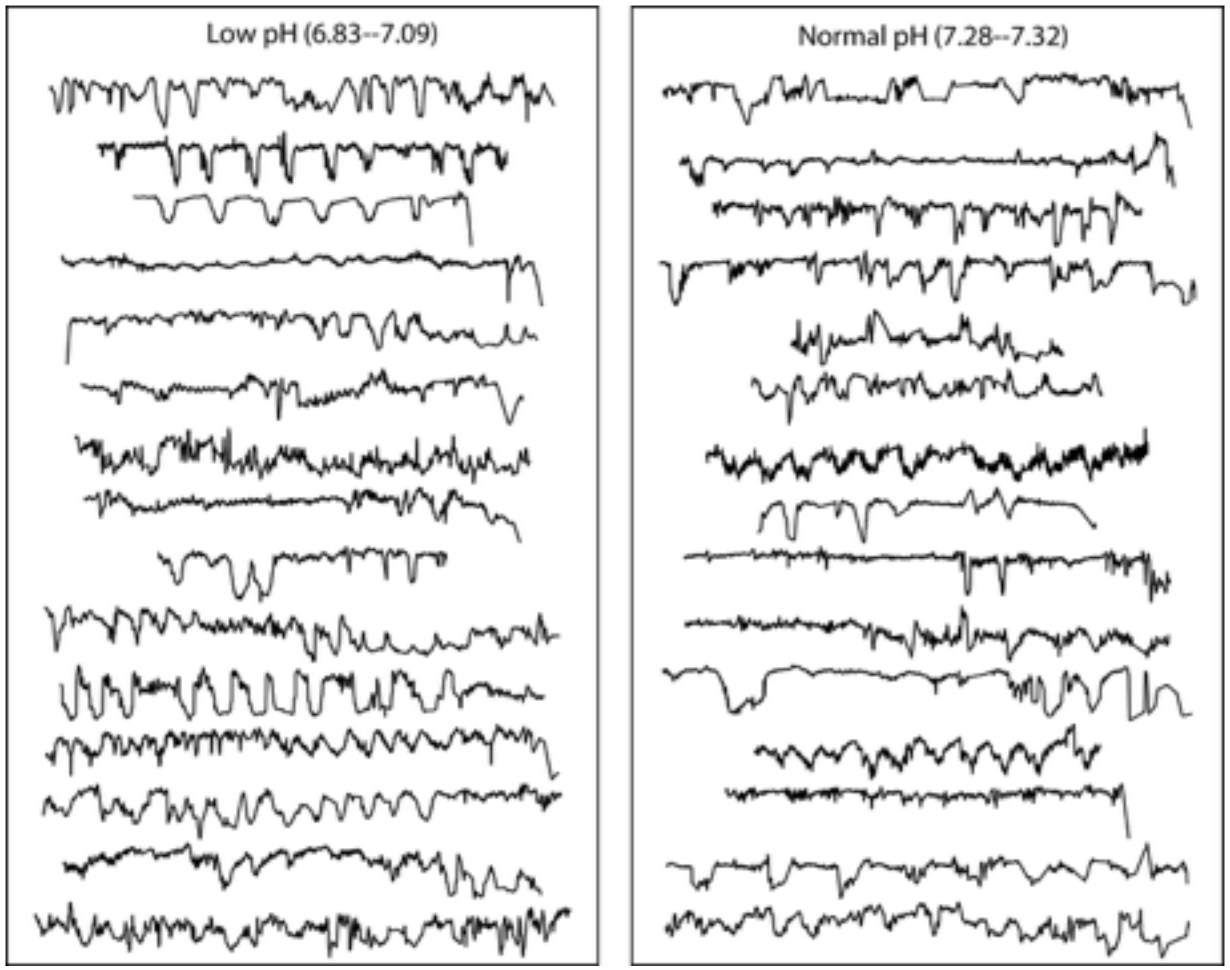
Classifying seizures

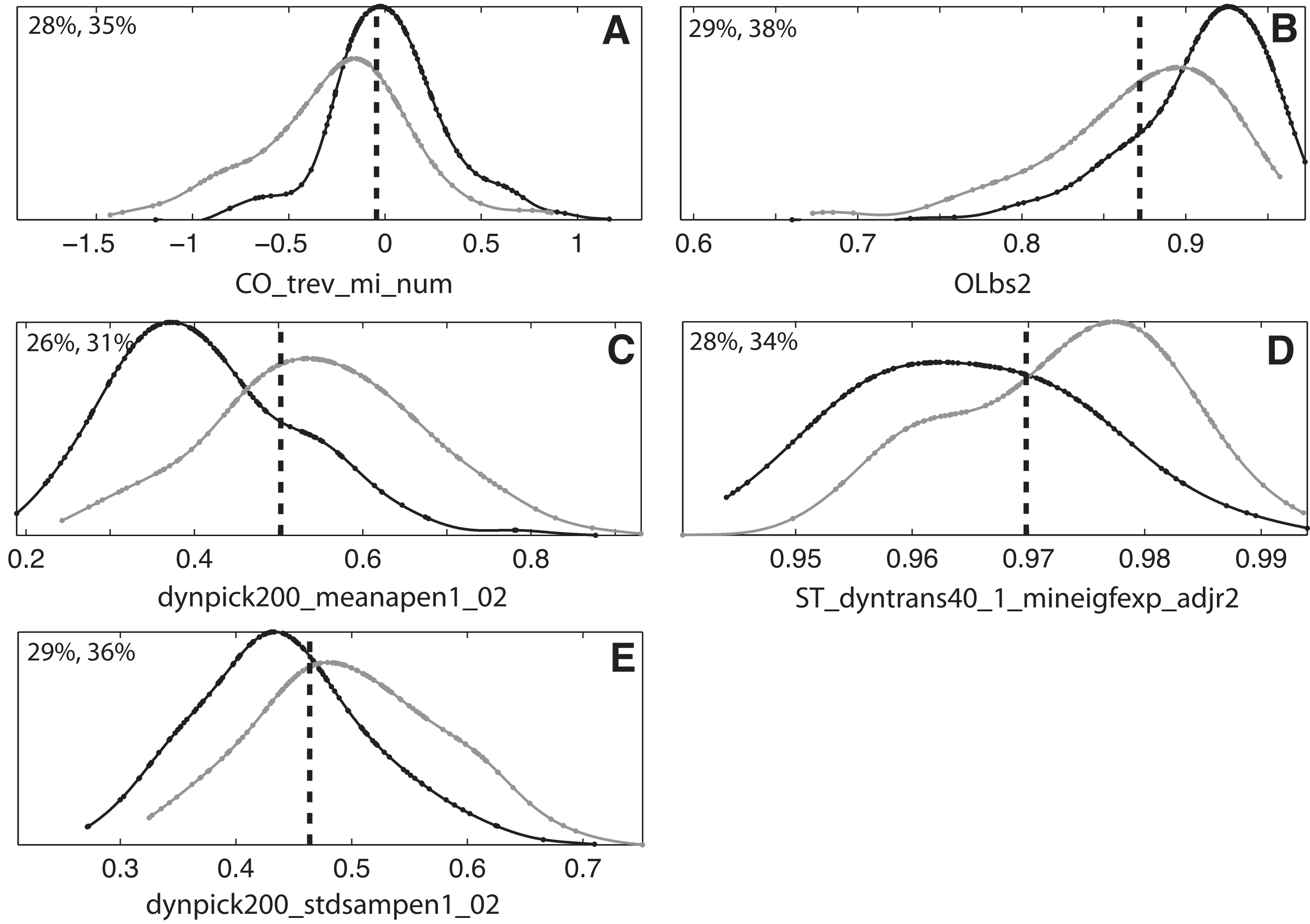


Diagnosis of fetal heart rates



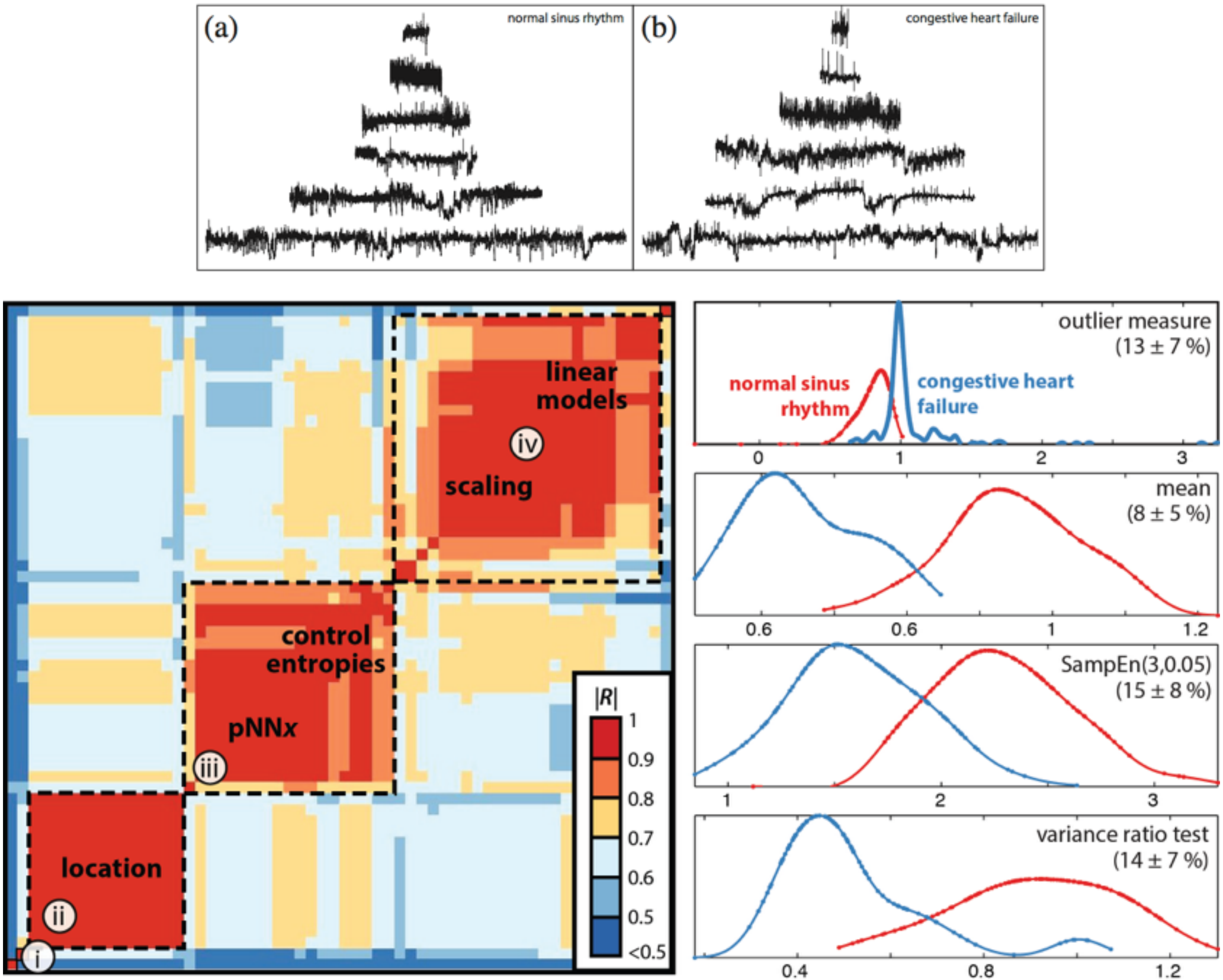
Arterial and venous blood samples



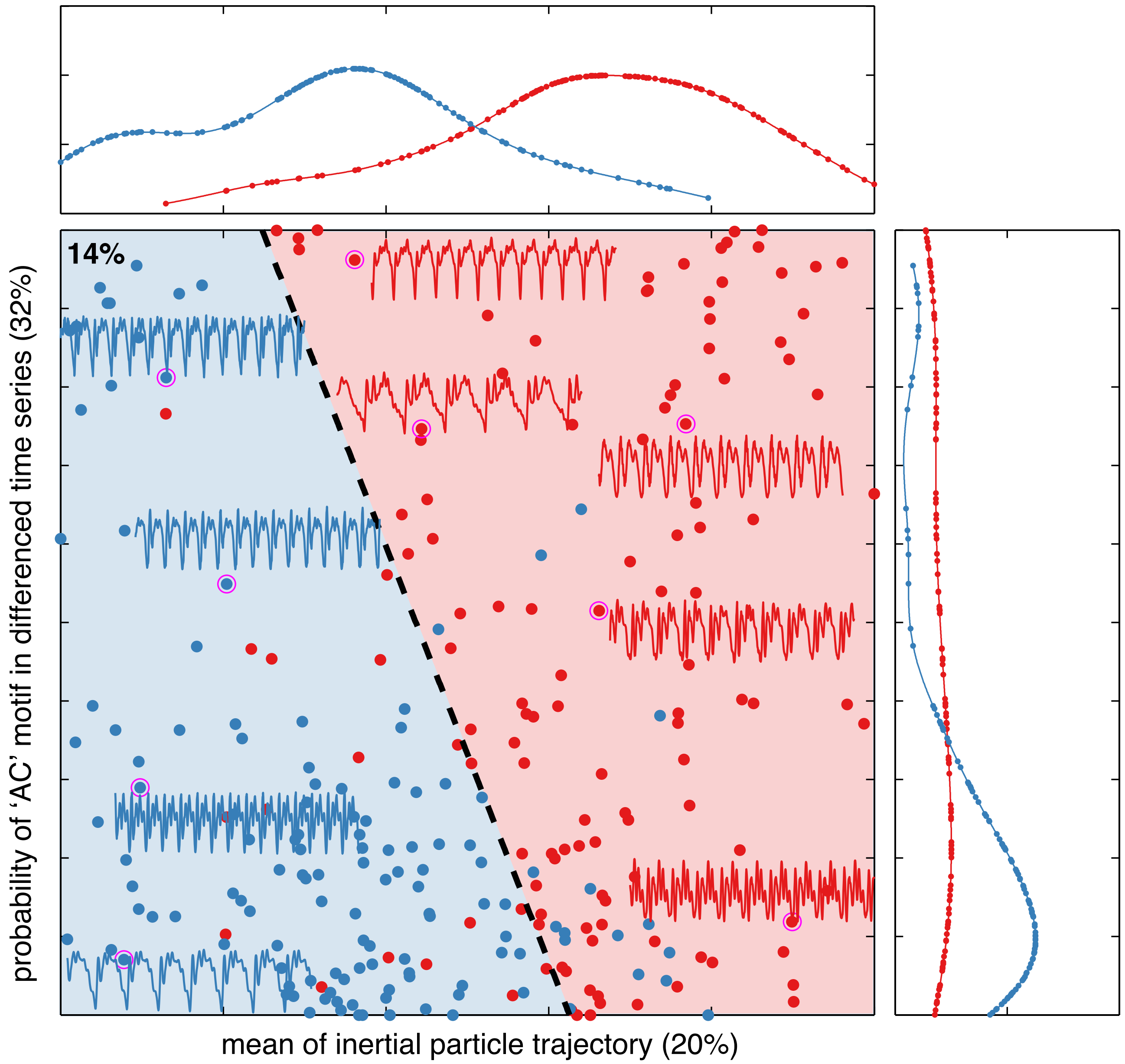


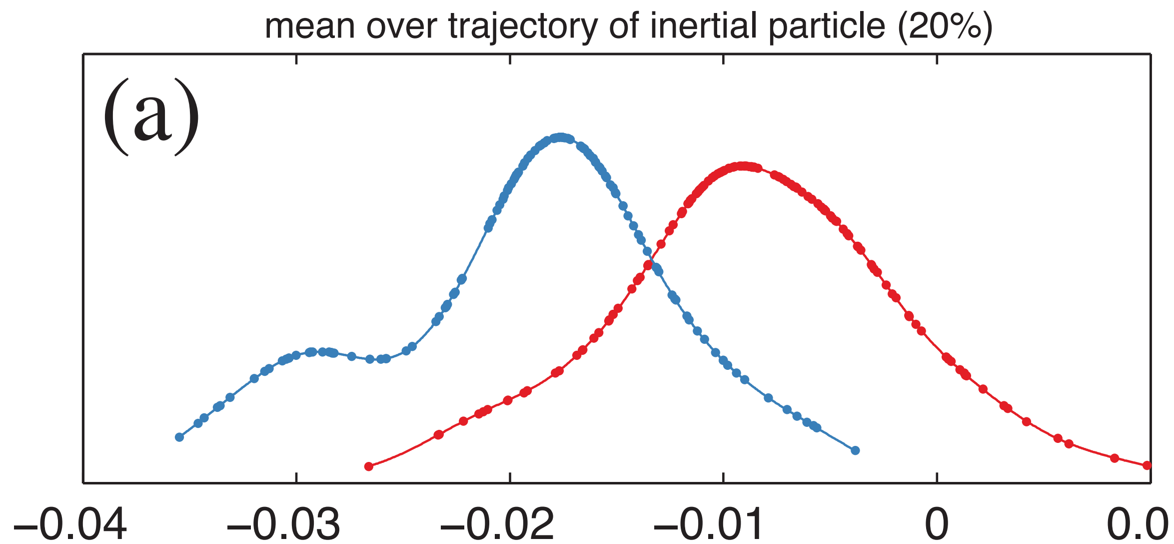
Fulcher, B.D. and Georgieva, A. E. and Redman, C. W G and Jones, N. S., Annual International Conference of the IEEE, EMBC, 3135 (2012), DOI: 10.1109/EMBC.2012.6346629

Heart rate variability



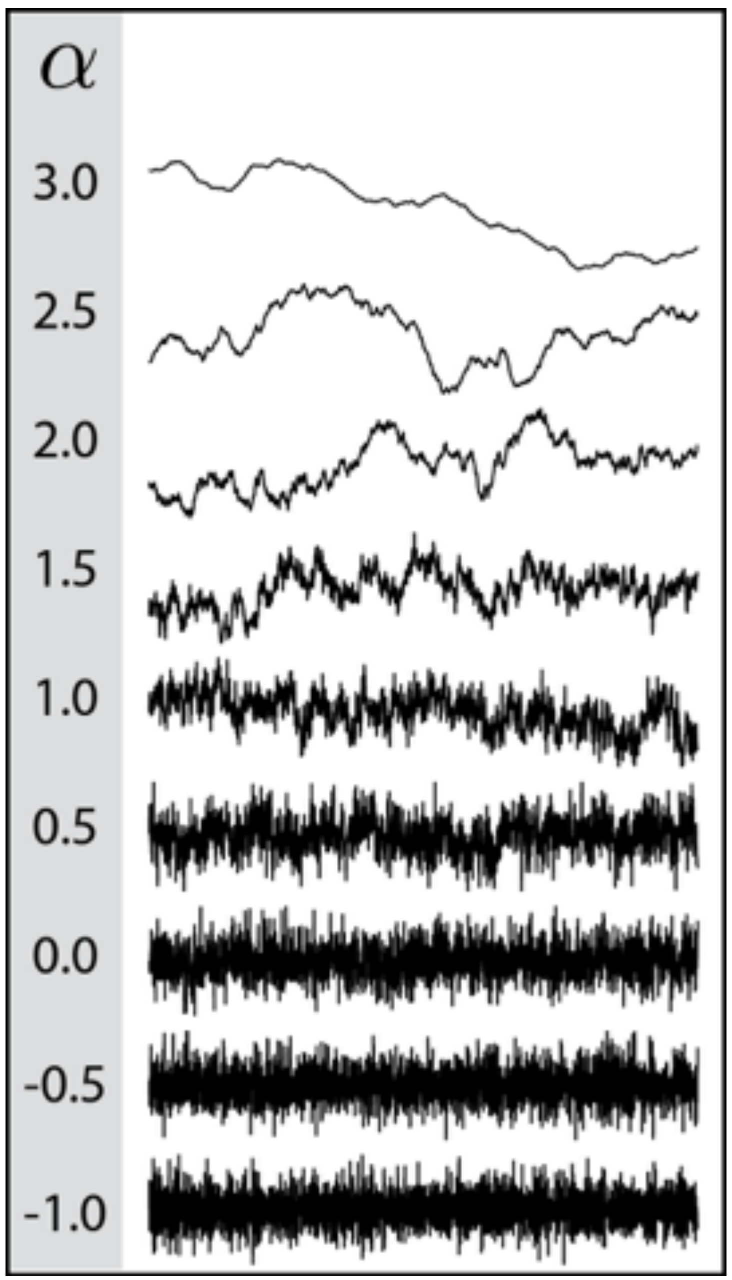
Parkinsonian speech

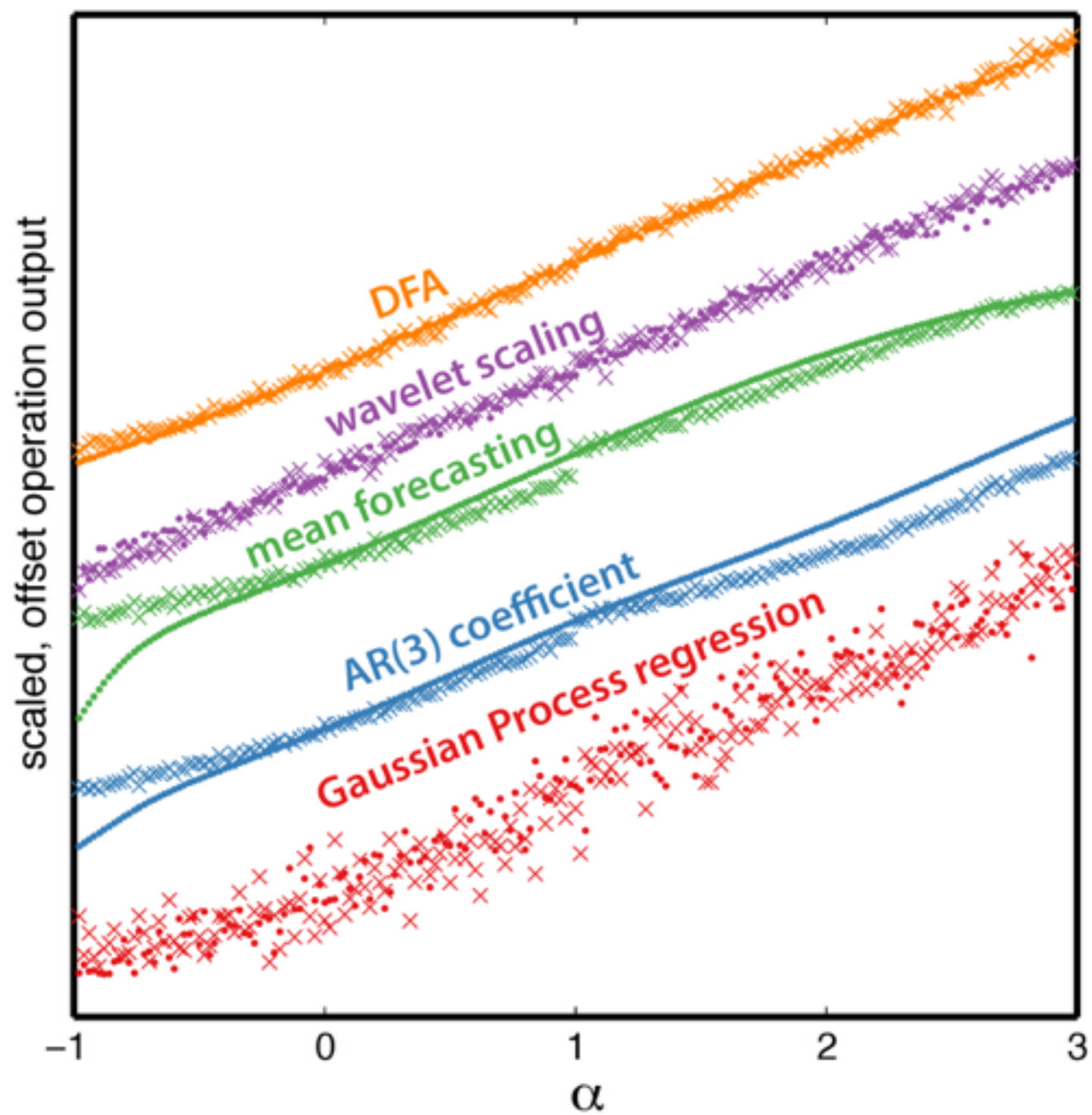




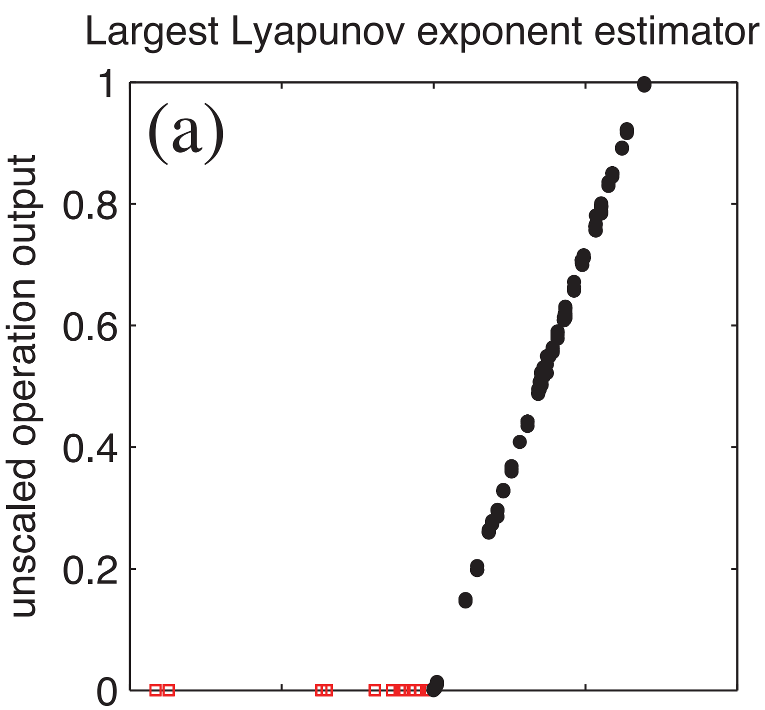
Classifiers combine methods developed in different scientific disciplines

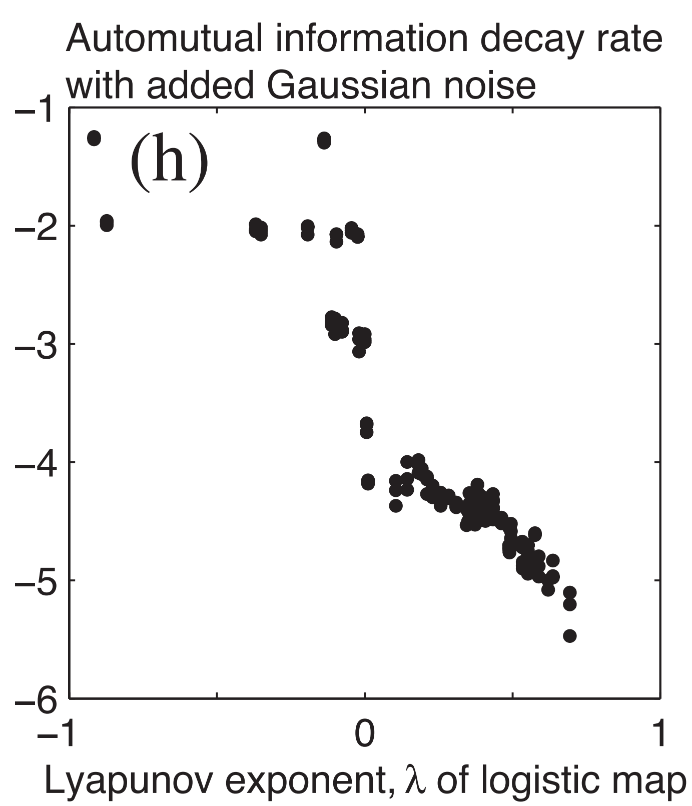
Self-affine time series

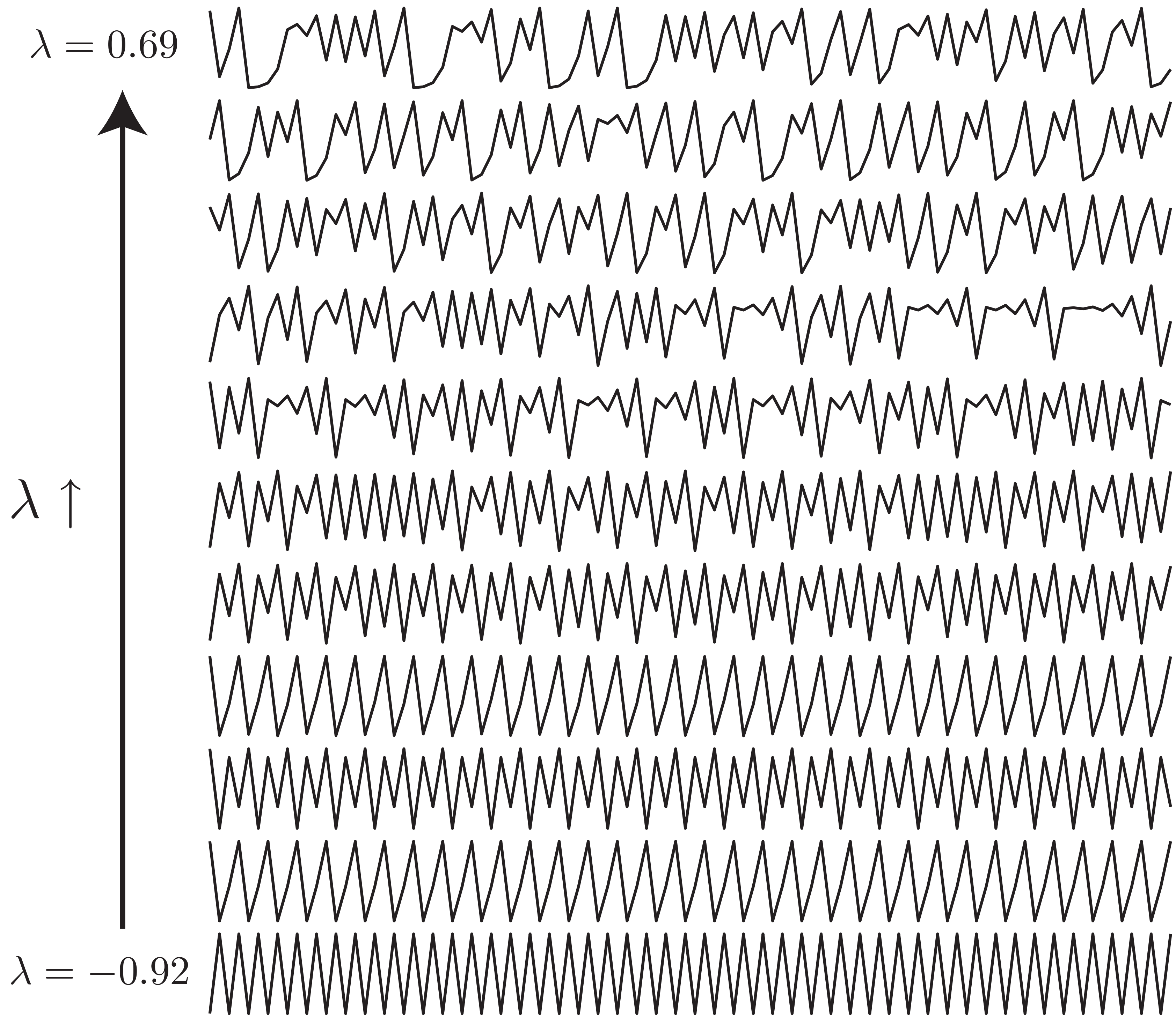




Logistic Map



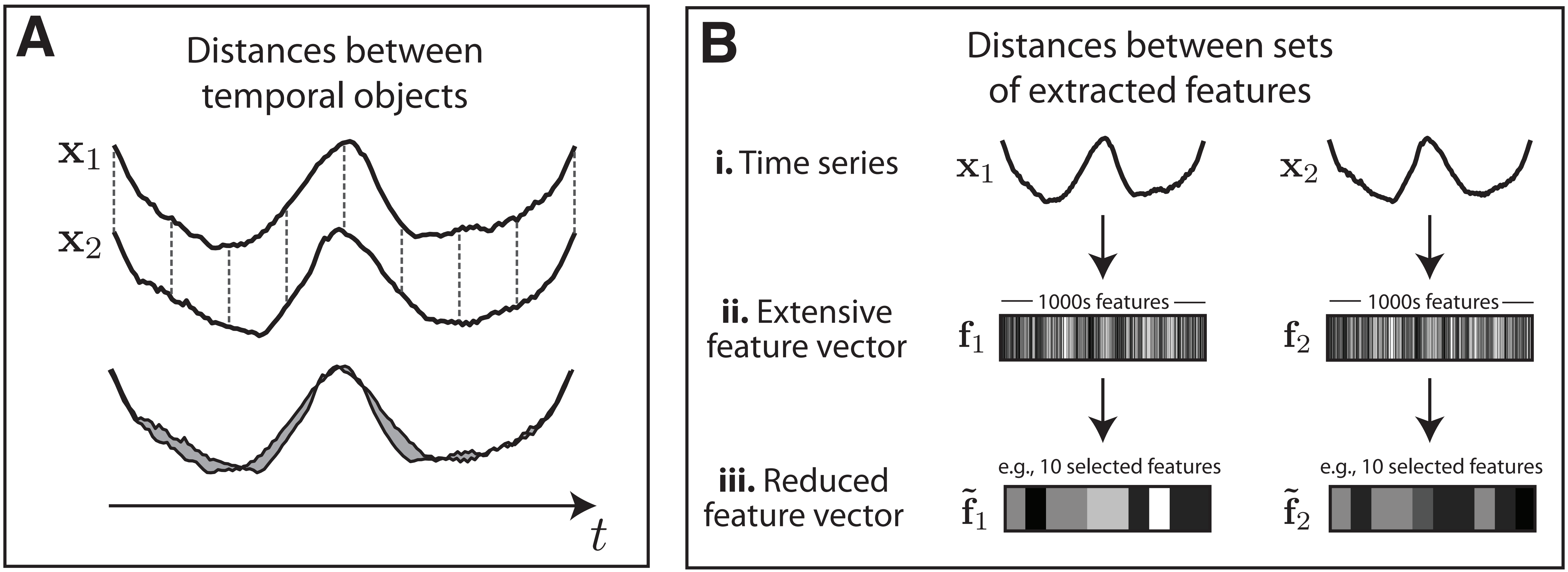




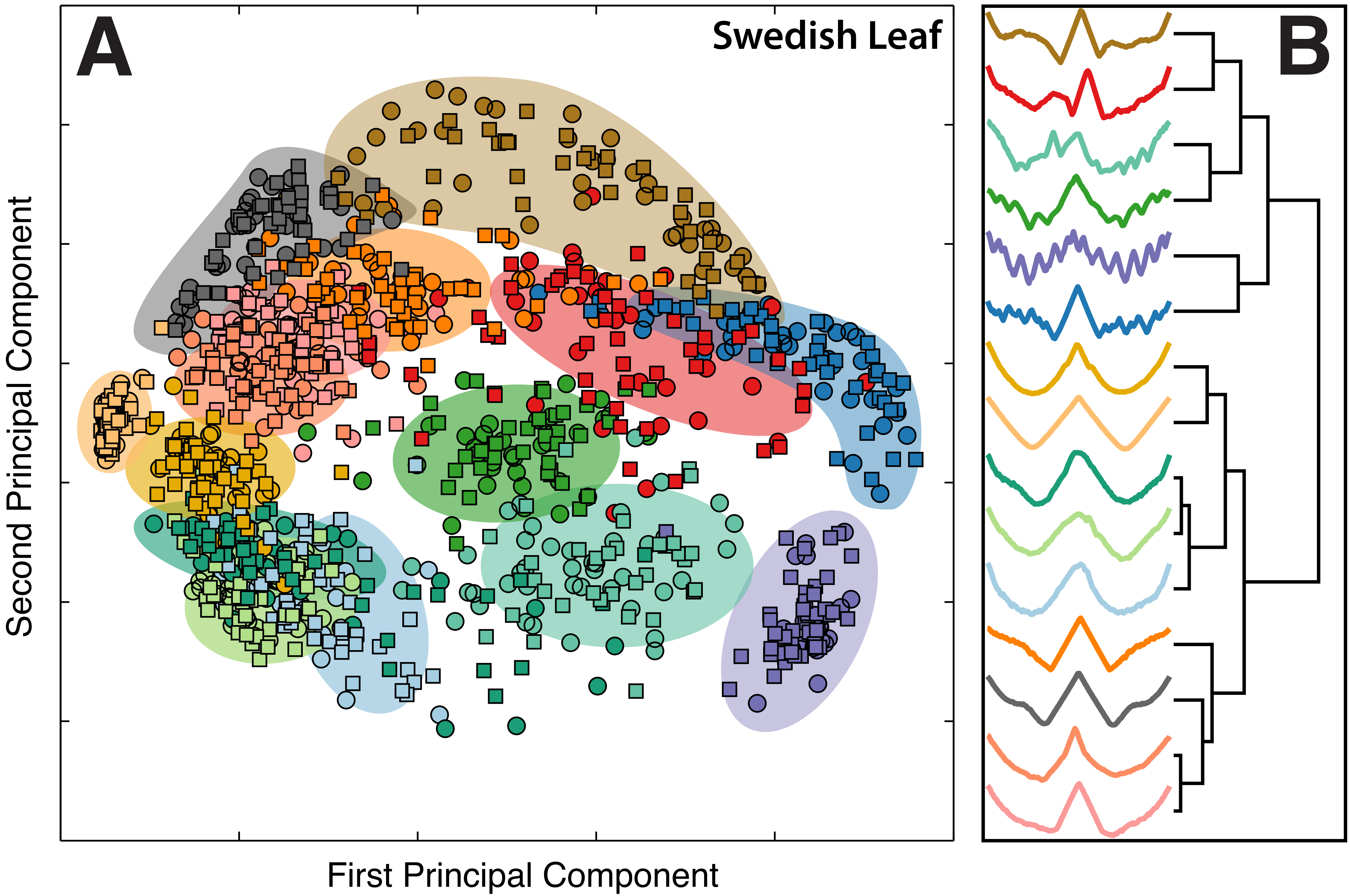
logistic maps: $x_{n+1} = Ax_n(1 - x_n)$

Time-series data mining

Cluster and classify short time-series 'patterns' (functional data)

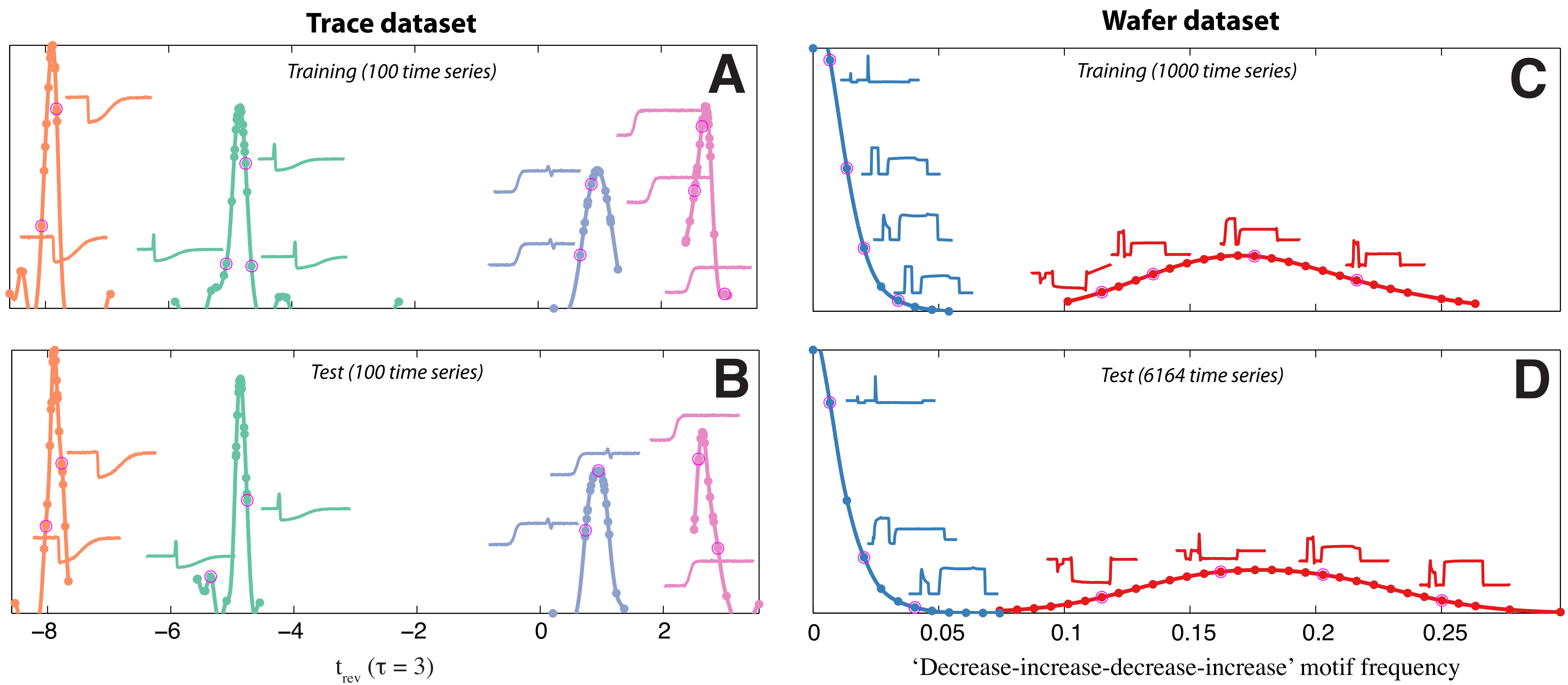


B. D. Fulcher & N. S. Jones, Highly comparative feature-based time-series classification. *IEEE Trans. Knowl. Data Eng.* **26**, 3026–3037 (2014)

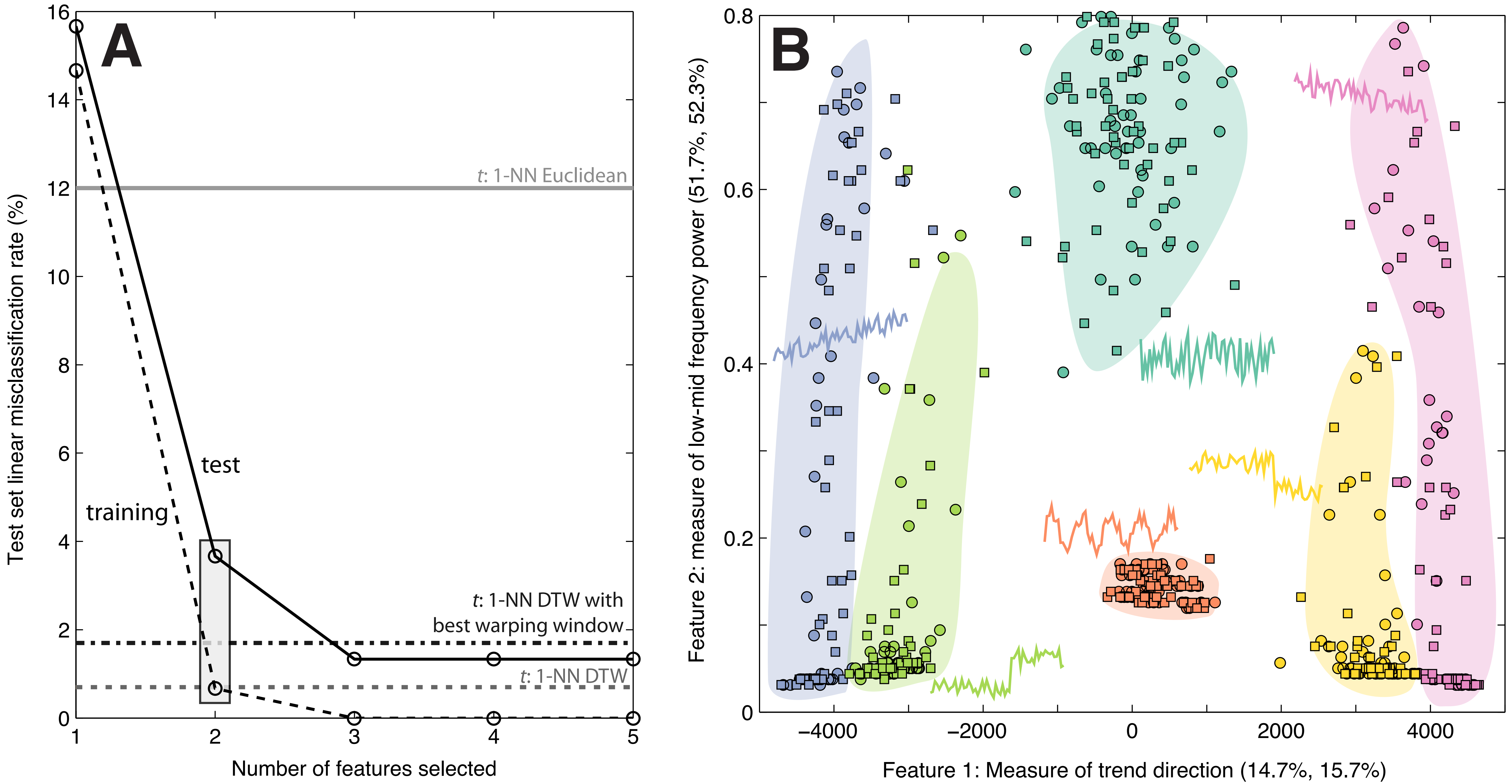

A
Swedish Leaf
B
Second Principal Component
First Principal Component

Single features perform well

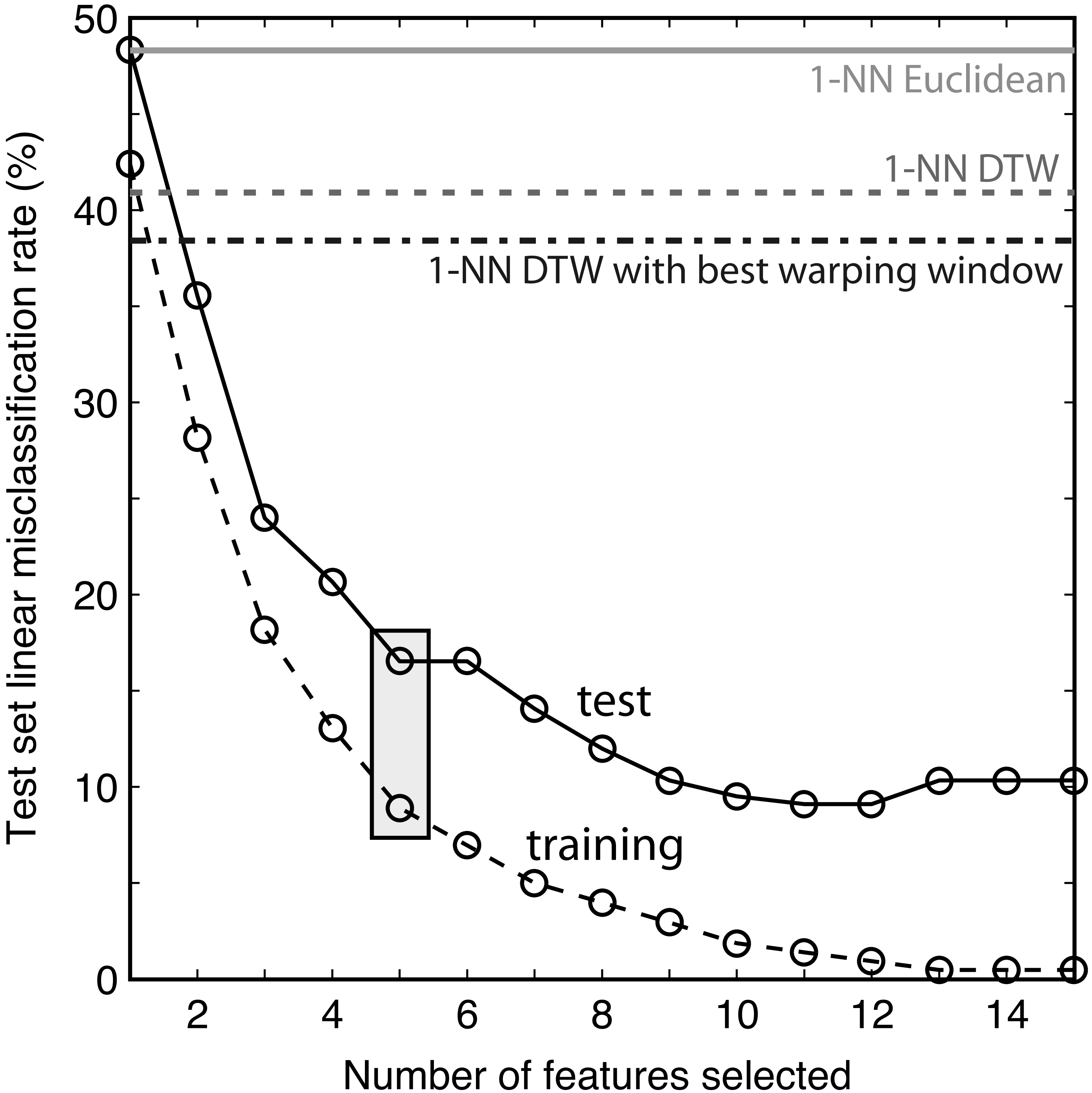
would be difficult to motivate by intuition



Improvements by adding a second feature



Improvements by adding multiple features



automatic

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)

Comp-Engine Time Series

A comparison engine for data and its analysis methods

www.comp-engine.org/timeseries

- Web resource for interdisciplinary scientific collaboration on time-series analysis
- >32,000 views since launching in February 2014
- Explore relationships between ~30,000 time series and ~9,000 analysis operations
- alpha implementation of drag-and-drop

Comp-Engine Time Series

A comparison engine for data and its analysis methods

Time Data Source Archives: Physionet: MGHDB (1089 items)

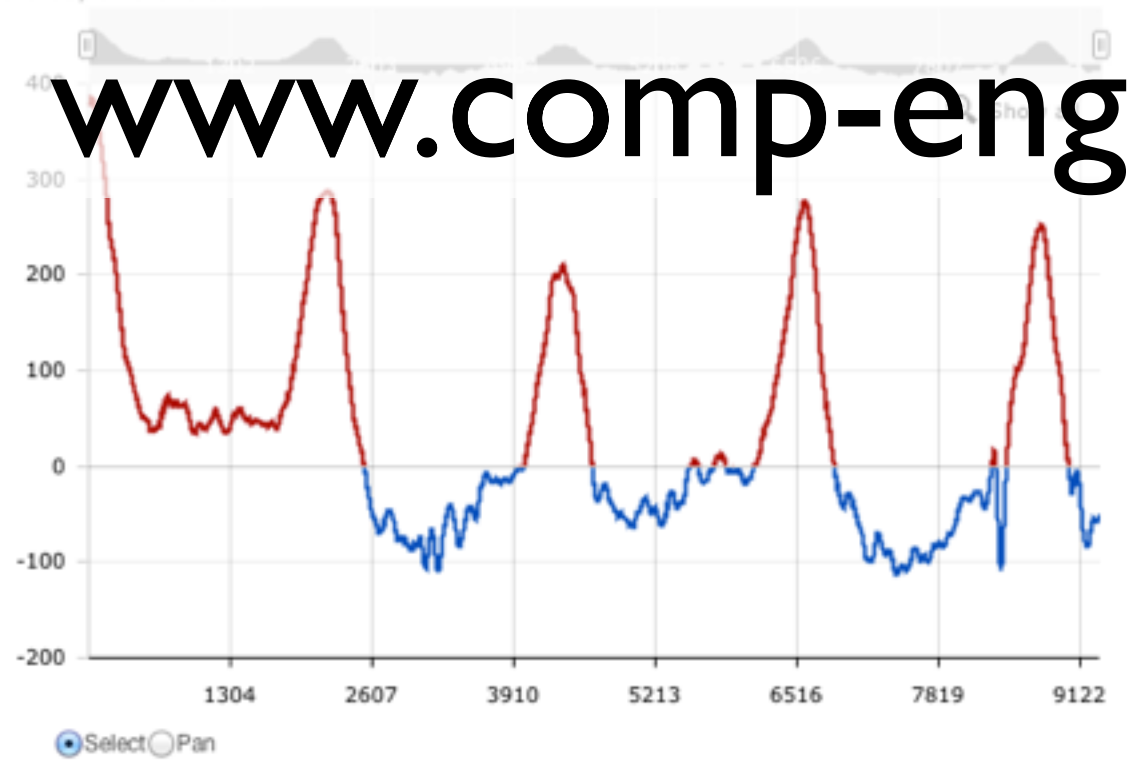
The Massachusetts General Hospital/Marquette Foundation (MGH/MF) Waveform Database is a com
collection of electronic recordings of hemodynamic and electrocardiographic waveforms of patients
units. It is the result of a collaboration between physicians, biomedical engineers and nurses at the M
General Hospital. The database consists of recordings from 250 patients and represents a broad sp
physiologic and pathophysiologic states.
Individual recordings vary in length from 12 to 86 minutes, and in most cases are about an hour lon
The typical recording includes three ECG leads, arterial pressure, pulmonary arterial pressure, centra

www.comp-engine.org/timeseries

MD_mghdb_mgh79_RespImp_SNIP_9145-18444

Share:

Data file: MD_mghdb_mgh79_RespImp_SNIP_9145-18444.dat
Length: 9300
chart by amcharts.com



Tags:

medical, mghdb, physionet, respiratoryimpedance, snip

Categories:

Real-world

Time series measured from real-world systems

Medical

Source:

Data by Category

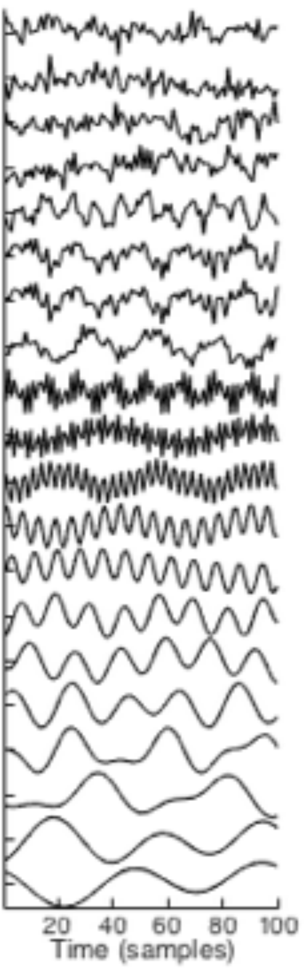
Air pressure Air temperature Animal sounds Astrophysics Audio Autoregressive with noise Beta noise Birdsong Correlated Noise Damped driven pendulum Driven pendulum with dissipation ECG Finance Flow Fietas Stochastic Sine Map Gait High low Like MIX(P) Logistic map Map Medical Meteorology Model M1a Model M5a Model M10a Moving average process Music Nonstationary autoregressive Opening prices Postural sway Powerlaw noise Precipitation rate Real-world Relative humidity Rossler attractor RR SDE models Sound effects Sprott 3D Flows Stochastic processes Synthetic Text Traded volume Uncategorised White noise

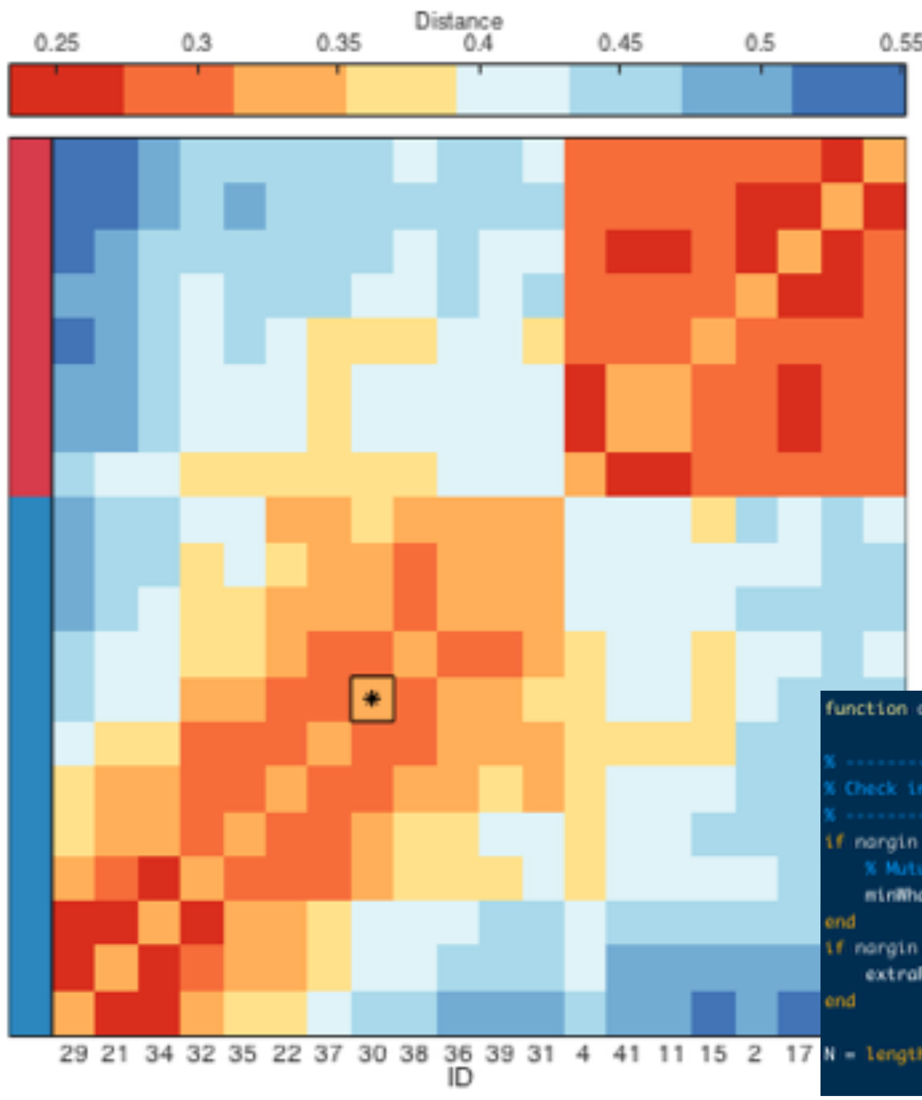
Data by Source

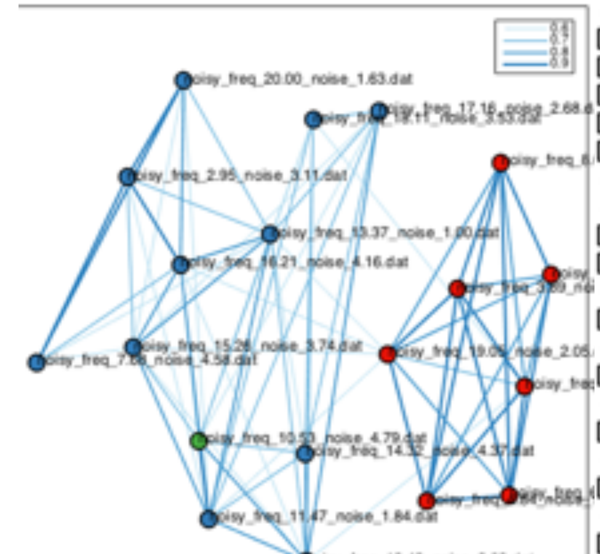
Air Temperature, NCEP/NCAR, CRU Ben Fulcher Simulated Ben generated powernoise Ben iTunes Ben making rmpnoise Ben MA simulations Ben music downsampled Ben Random ... Ben ... Climatic Research Unit, University of East Anglia Driven pendulum Ben Financial log returns Ben Frietas Stochastic Sine Map Ben Google trends Logistic Map A sweep Ben Macaulay Library NCEP/NCAR, CRU Physionet Physionet: CHFDB Physionet: MGHDB Physionet: NESFDB Physionet: NSRDB Physionet RR CHF NSR Precipitation rate, NCEP/NCAR, CRU Project Gutenberg Relative humidity, NCEP/NCAR, CRU SDE Toolbox M1a SDE Toolbox M5a SDE Toolbox M10a SDE Toolbox Simulated Sea level pressure, NCEP/NCAR, CRU Sound Jay SPIDR SPIDR Geomagnetic annual means -- Ionosphere Sprott Conservative Flows Sprott Conservative Maps Ben Sprott Damped driven pendulum Ben Sprott Dissipative Maps Ben Sprott Noninvertible Maps Ben Text processing Ben Time-Series Data Library Timmer nonstationary autoregressive processes Yahoo Finance Yahoo Finance Shares

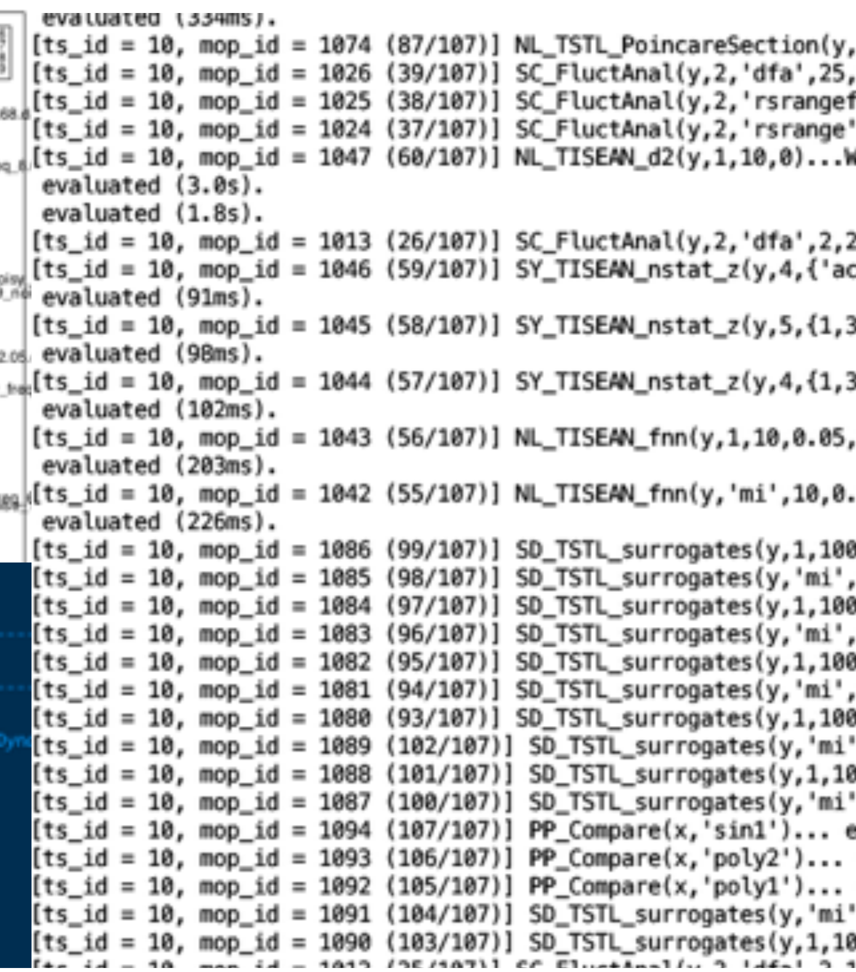
Matlab-based code repository

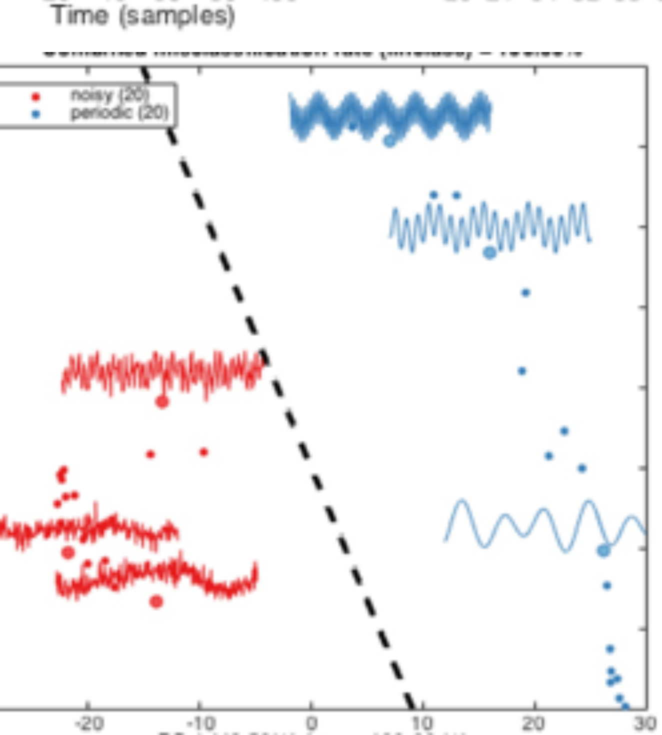
@compTimeSeries

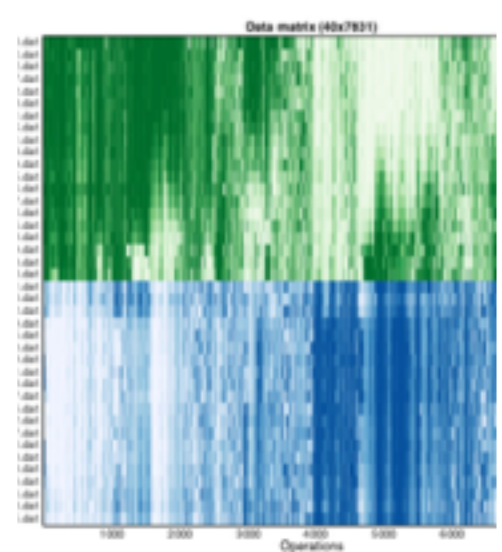


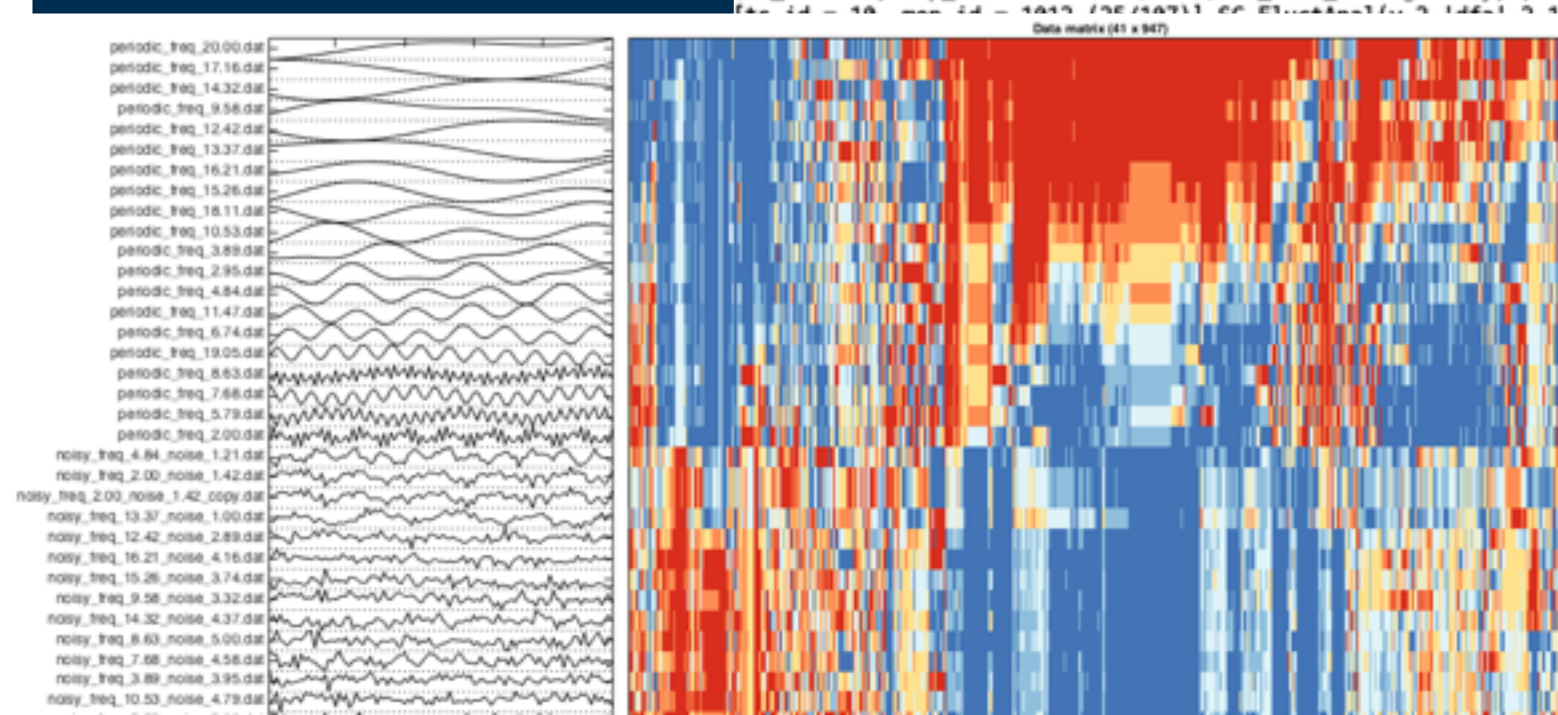


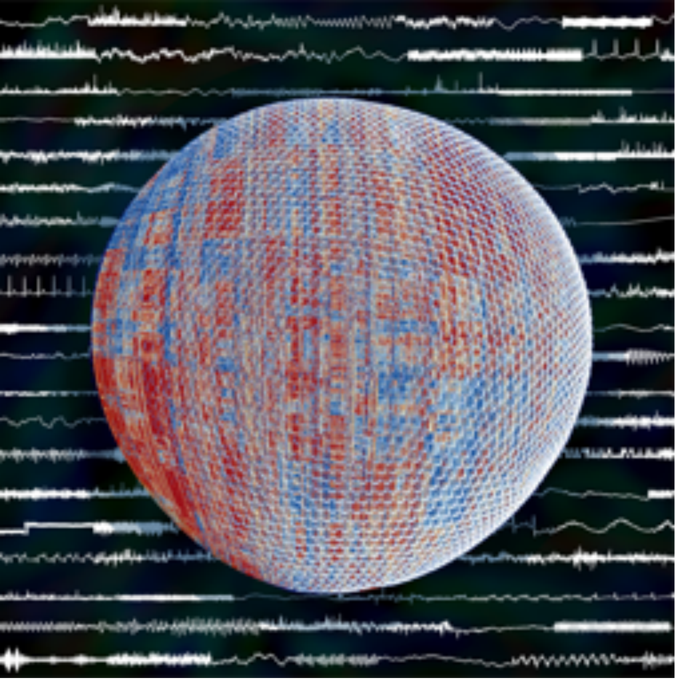




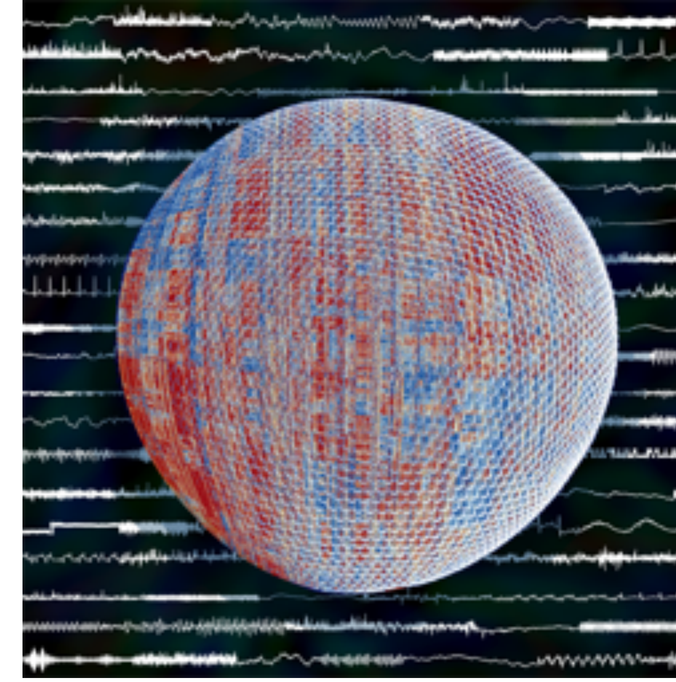




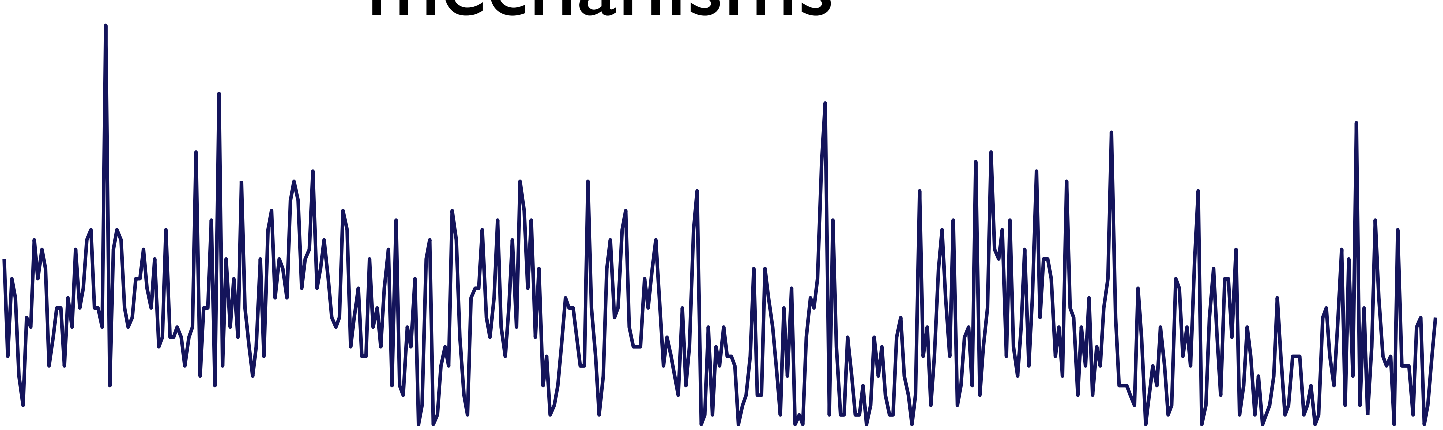


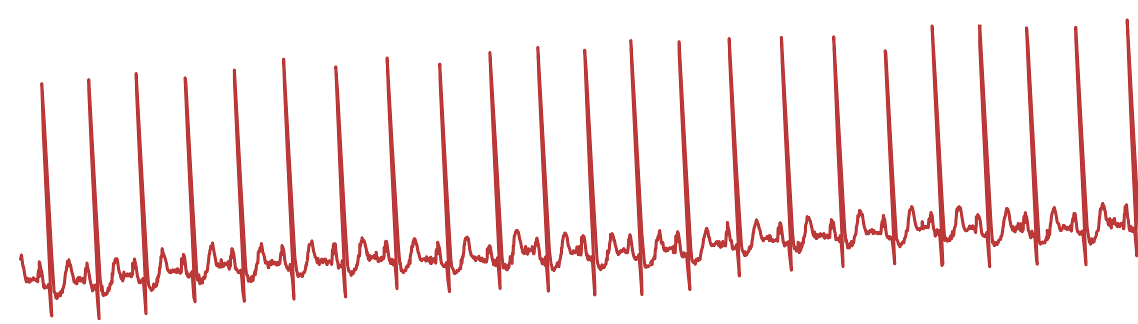


Conclusions



- A semi-automated approach to time-series analysis that compares thousands of interdisciplinary methods
- Can be viewed as a starting point to guide more focused time-series analysis
- Results provide insights into underlying dynamical mechanisms





ben.fulcher@monash.edu

@bendfulcher, @compTimeSeries

www.comp-engine.org/timeseries

B. D. Fulcher, M. A. Little, and N. S. Jones. *J. R. Soc. Interface*, **10**:83 (2013), *DOI: 10.1098/rsif.2013.0048*