An Engine for Comparative Time-Series Analysis

"the taming of the zoo"

Ben Fulcher,

Max Little, Nick Jones

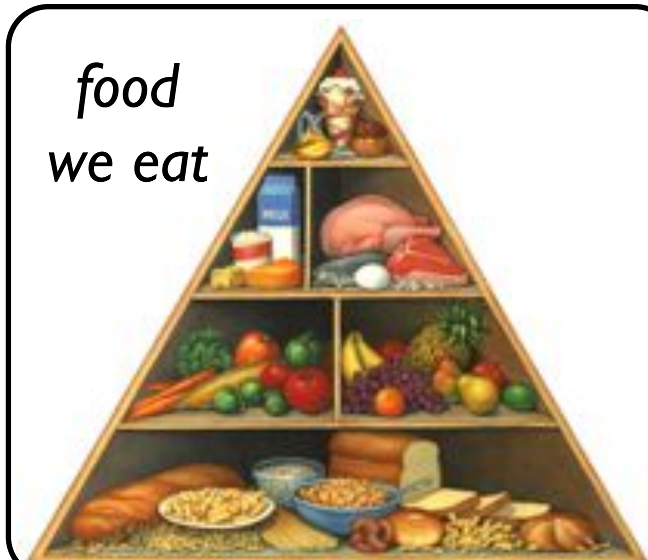
May 2011

Organizing

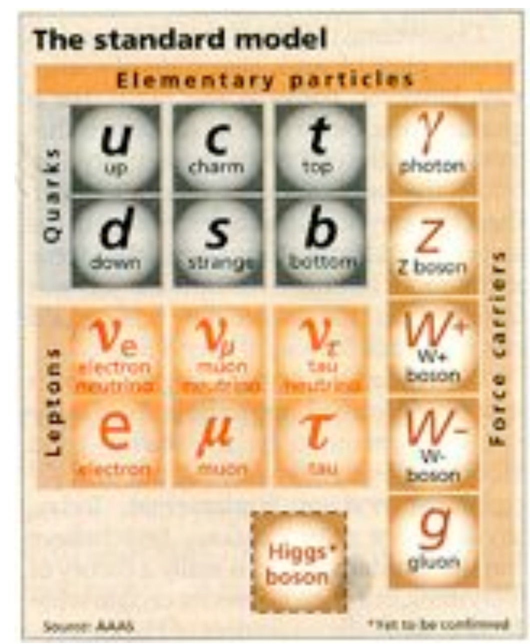
Scientific endeavors often focus on structuring libraries of collected information.

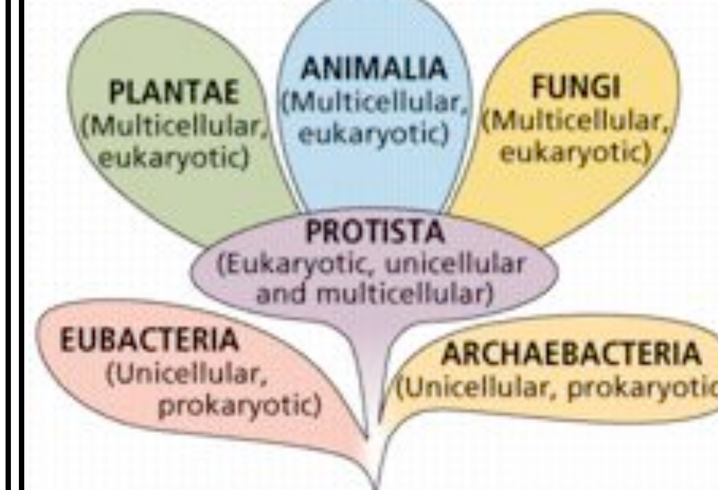
personalities

food we eat

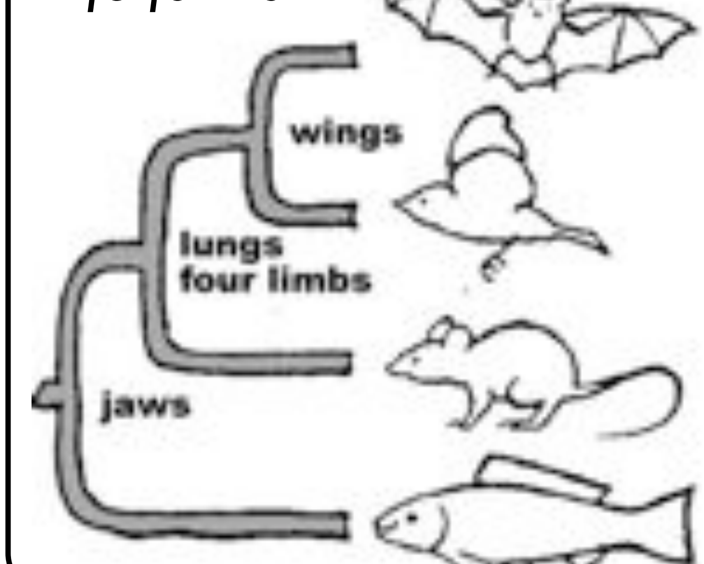


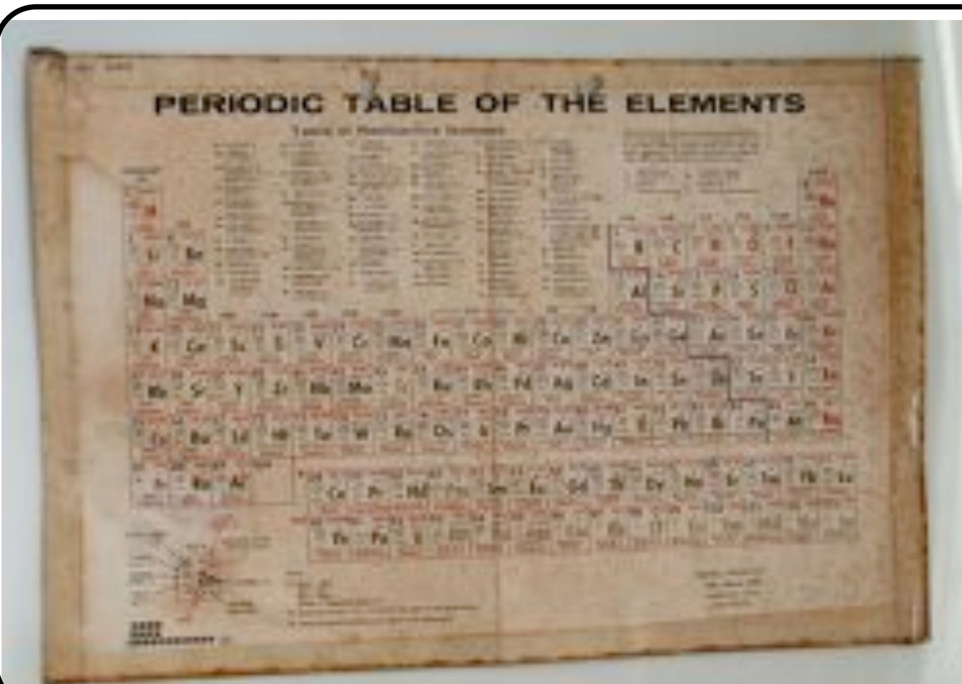
matter



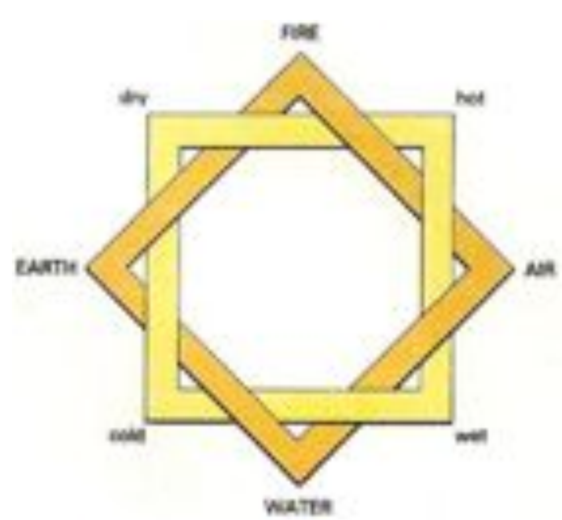


life forms

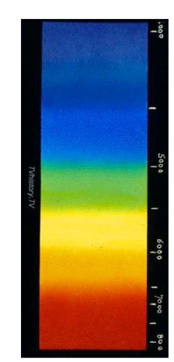




macroscopic substances we encounter

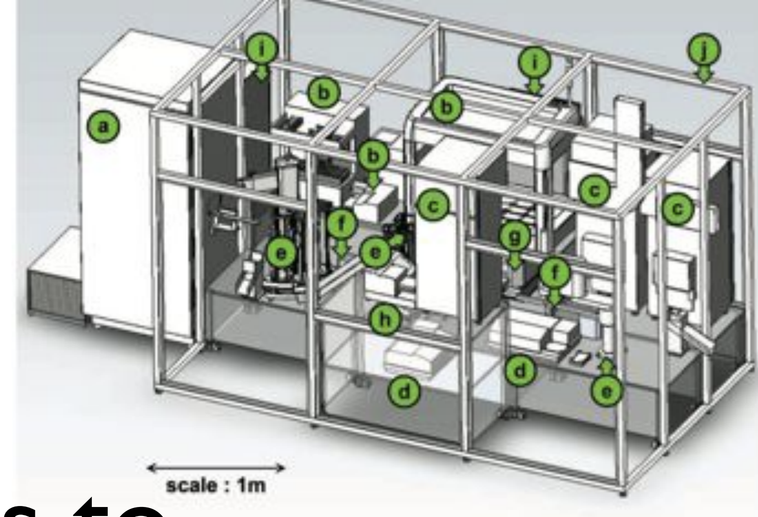


color



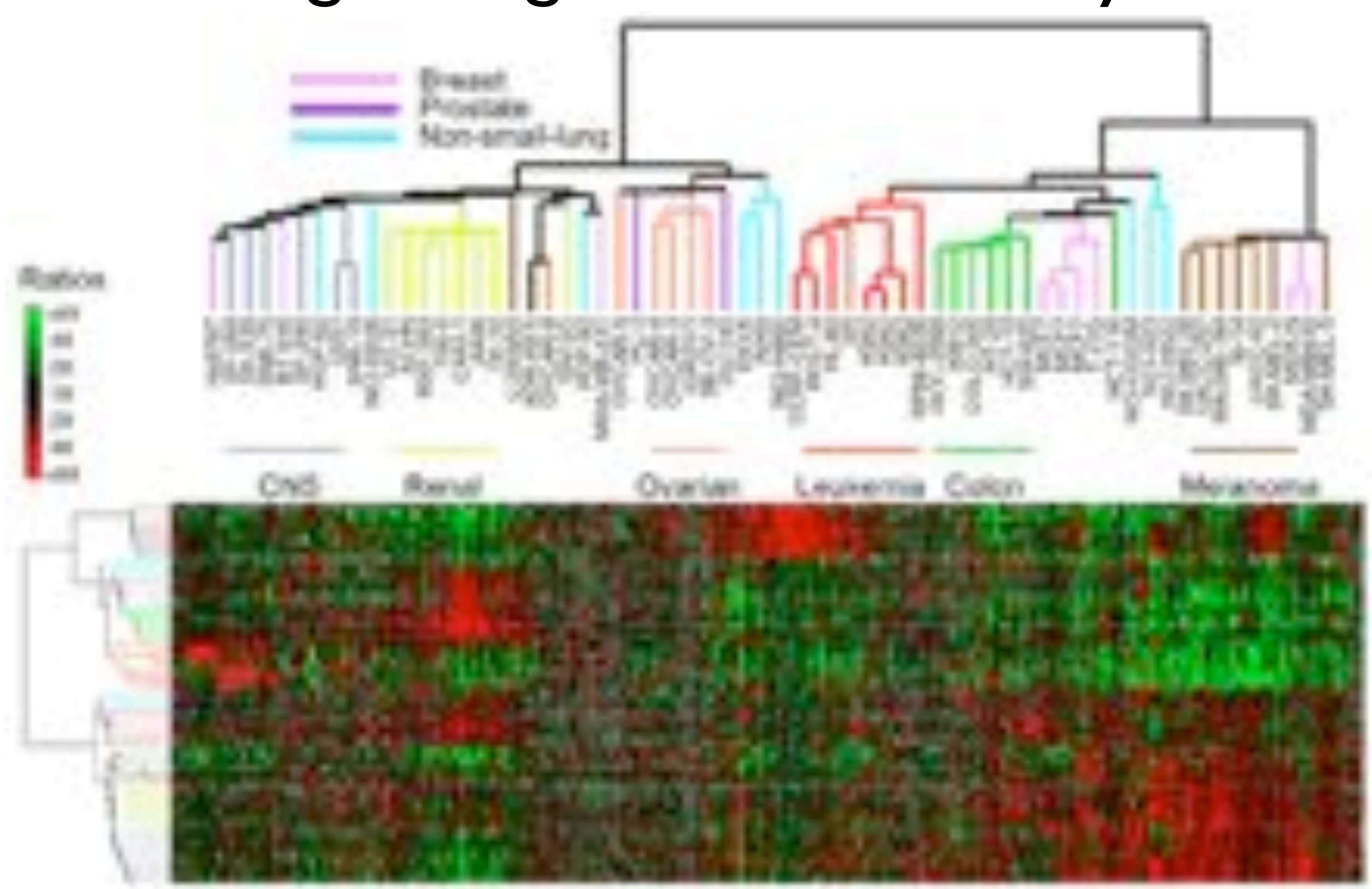
this helps us understand the complexity in the world

Organizing



Statistical learning techniques allow us to organize and understand things on a greater scale than ever before.

e.g., the genetic microarray

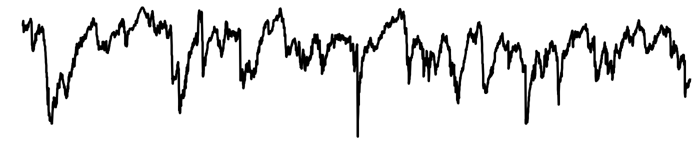


Organizing Science

- Our data and our methods are also objects that require organization.
- How do we make sense of the time-series data that we observe in the world?

We construct a comparative framework for time-series analysis

Challenges



Time Series Analysis:

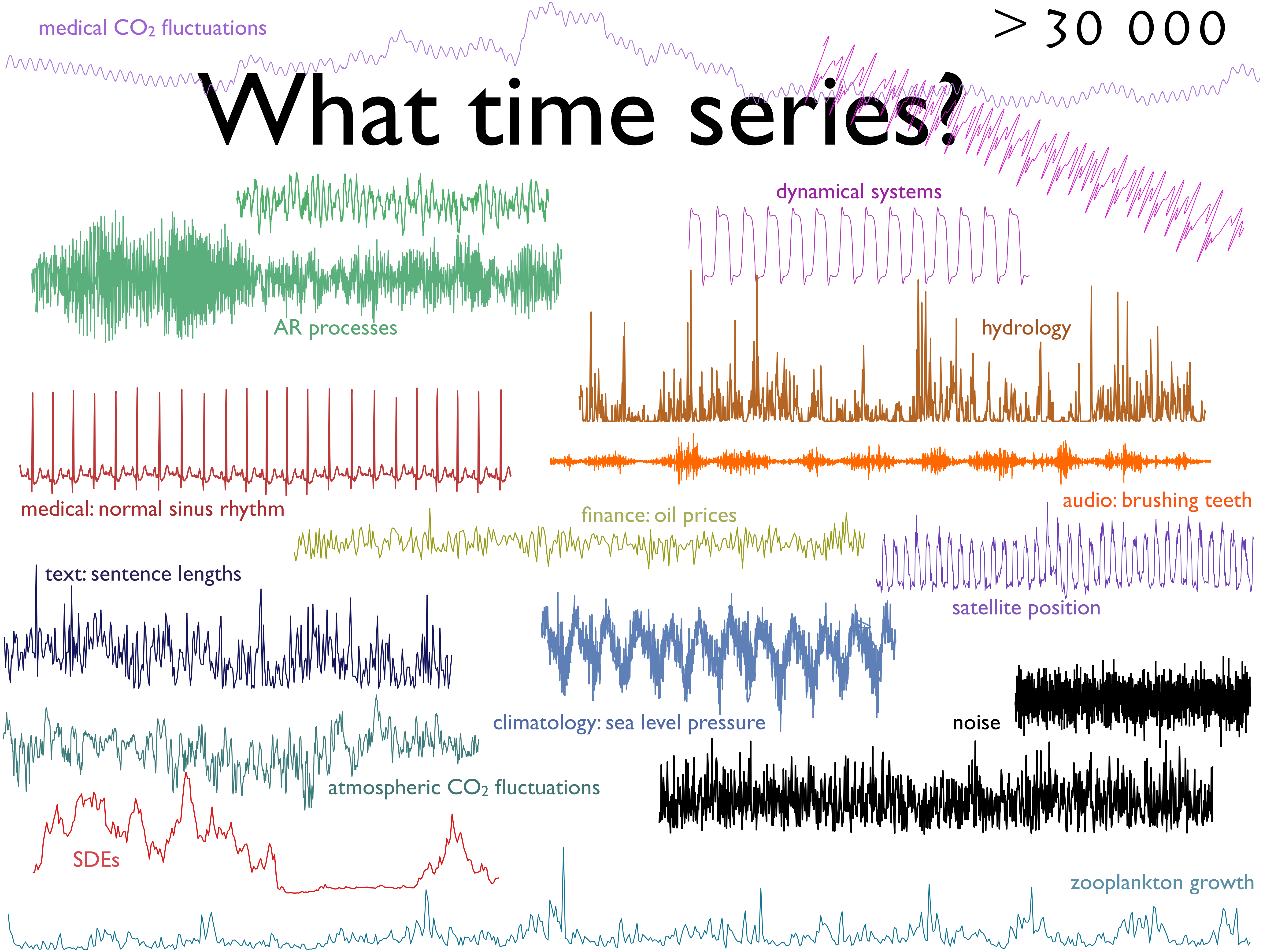
- Pervasive importance in science
- Huge quantities of data
- Vast and growing quantity of methods
- Interdisciplinary boundaries

Structure

- Framework
- Structure of methods for time-series analysis
- Structure of empirical time series
- Utility for specific applications
- Constrained time-series datasets

What time series?

$> 30\ 000$



$> 9\ 000$

What operations?

Basic statistics

trimmed means
zero crossings
standard deviation
outliers
local extrema

Stationarity

StatAv
sliding windows
bootstraps
distribution comparisons

Static distribution

quantiles
moments
fits to standard distributions
hypothesis tests

Basis Functions

wavelet transform
power spectrum peaks
spectral measures
low frequency power

Correlation

linear autocorrelations
decay properties
automutual information
dependence on additive noise
nonlinear autocorrelations
time reversal asymmetry
generalized self-correlation function
recurrence structure
autocorrelation robustness
fluctuation analysis: scaling
randomization robustness
recurrence plots
seasonality testing

Model fits

primitive forecasting
GARCH modeling
Fourier fits
step-ahead dependence
exponential smoothing
AR models
state space models
hidden Markov models
'walker' statistics
piecewise splines
ARMA modeling
Gaussian Processes

Nonlinear

2D embedding structure
TSTOOL
TISEAN
fractal dimension
correlation dimension
Taken's estimator
Poincaré sections
surrogate data
nonlinear prediction error
Lyapunov exponent
false nearest neighbours

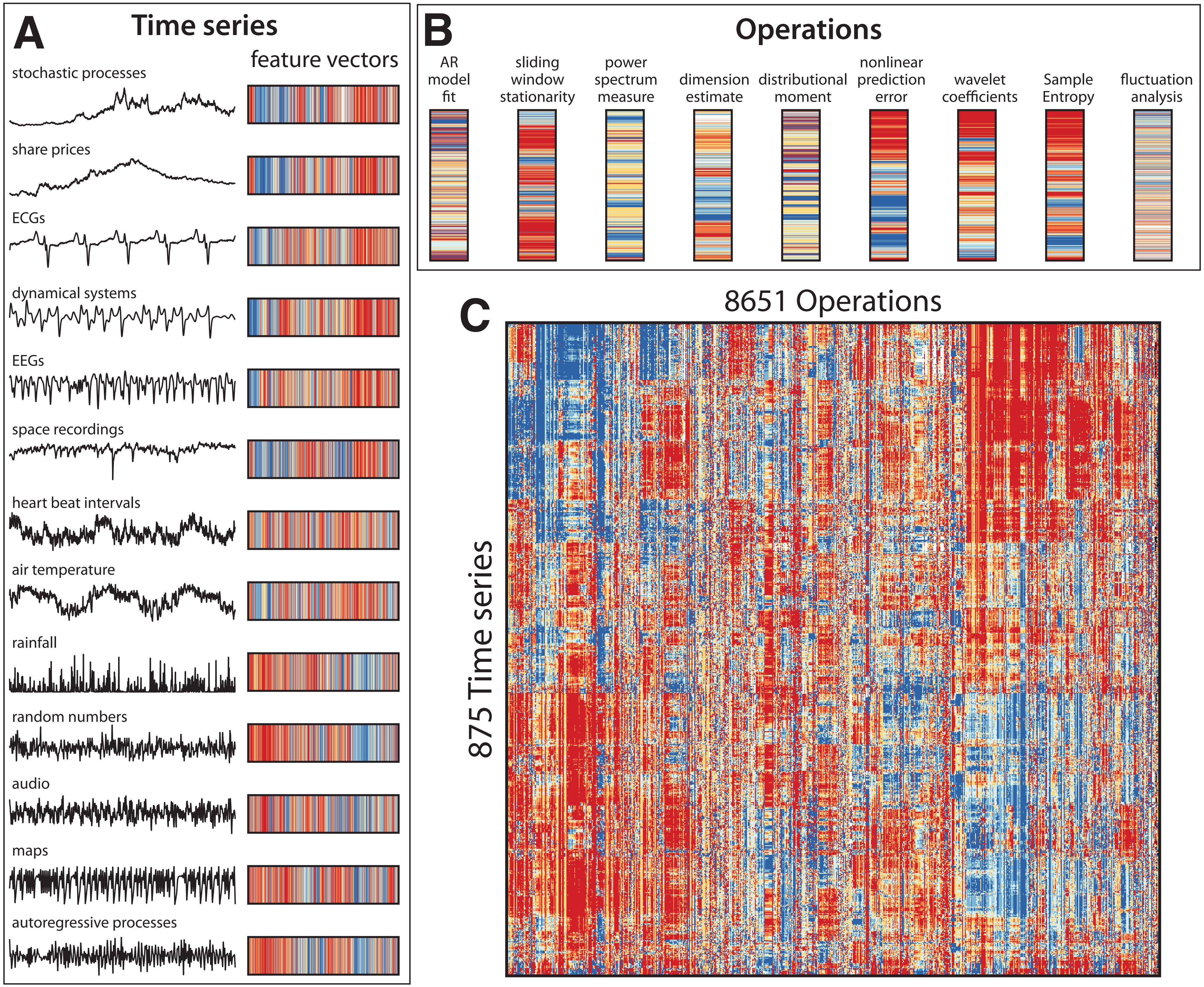
Others

course-grained transition matrices
motif distributions
couple to dynamical systems
visibility graph
stick angle distribution
step detection algorithms
extreme events
drifting mean tests
PCA of embedded signal
domain-specific standard metrics

Information Theory

SampEn
distributional entropies
conditional entropies
binned entropies
kernel smoothed entropies
ApEn
Tsallis entropies

Design Matrix



Organizing Our Methods

- We organize operations using their outputs on a diverse range of empirical time series.
- Clustering allows us to form reduced sets of operations that capture the dominant types of behavior in our database.
- Gives structure to an interdisciplinary field.

long-range scaling

power spectral density

linear models

stationarity

variance

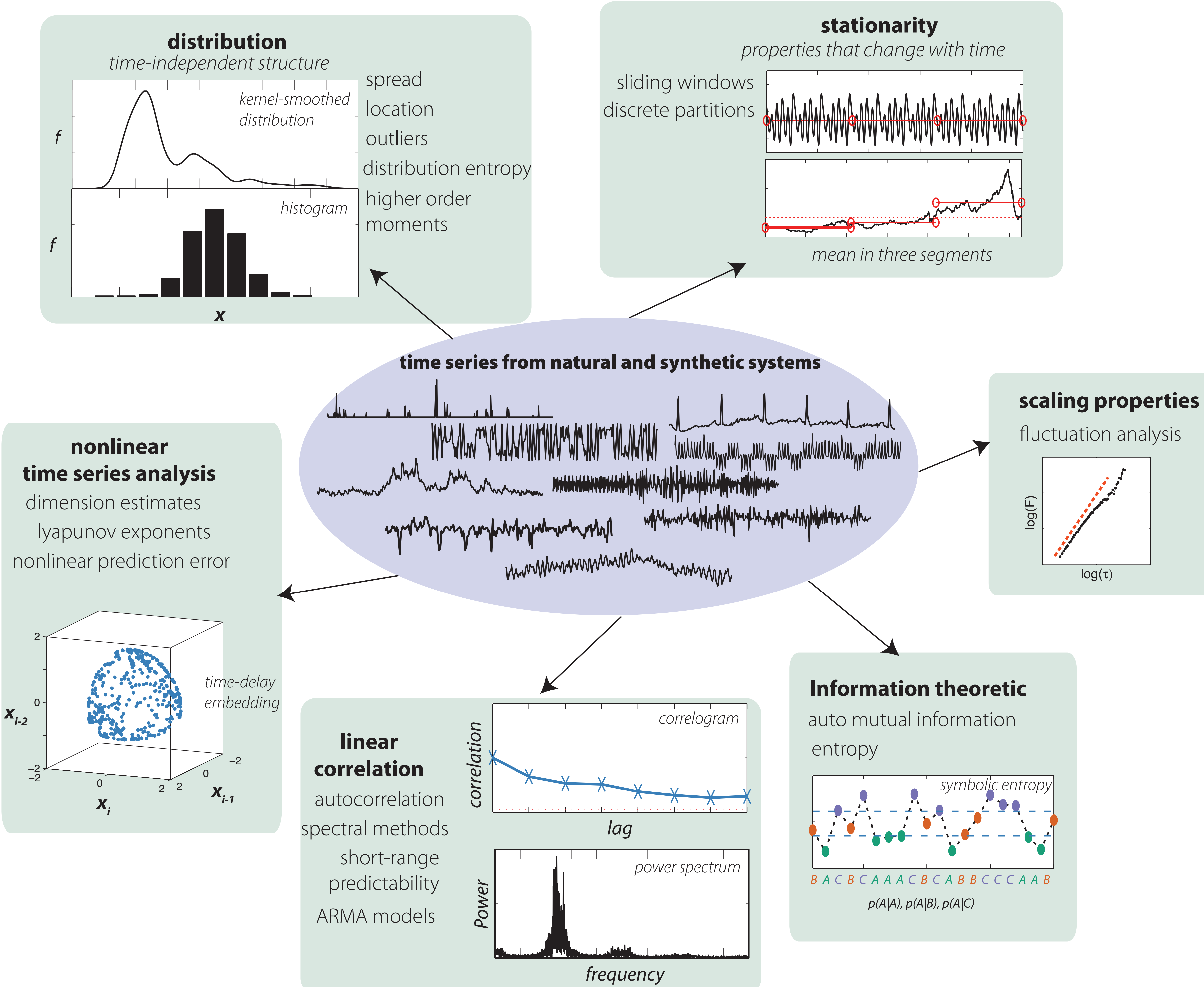
entropy

correlation dimension

complexity

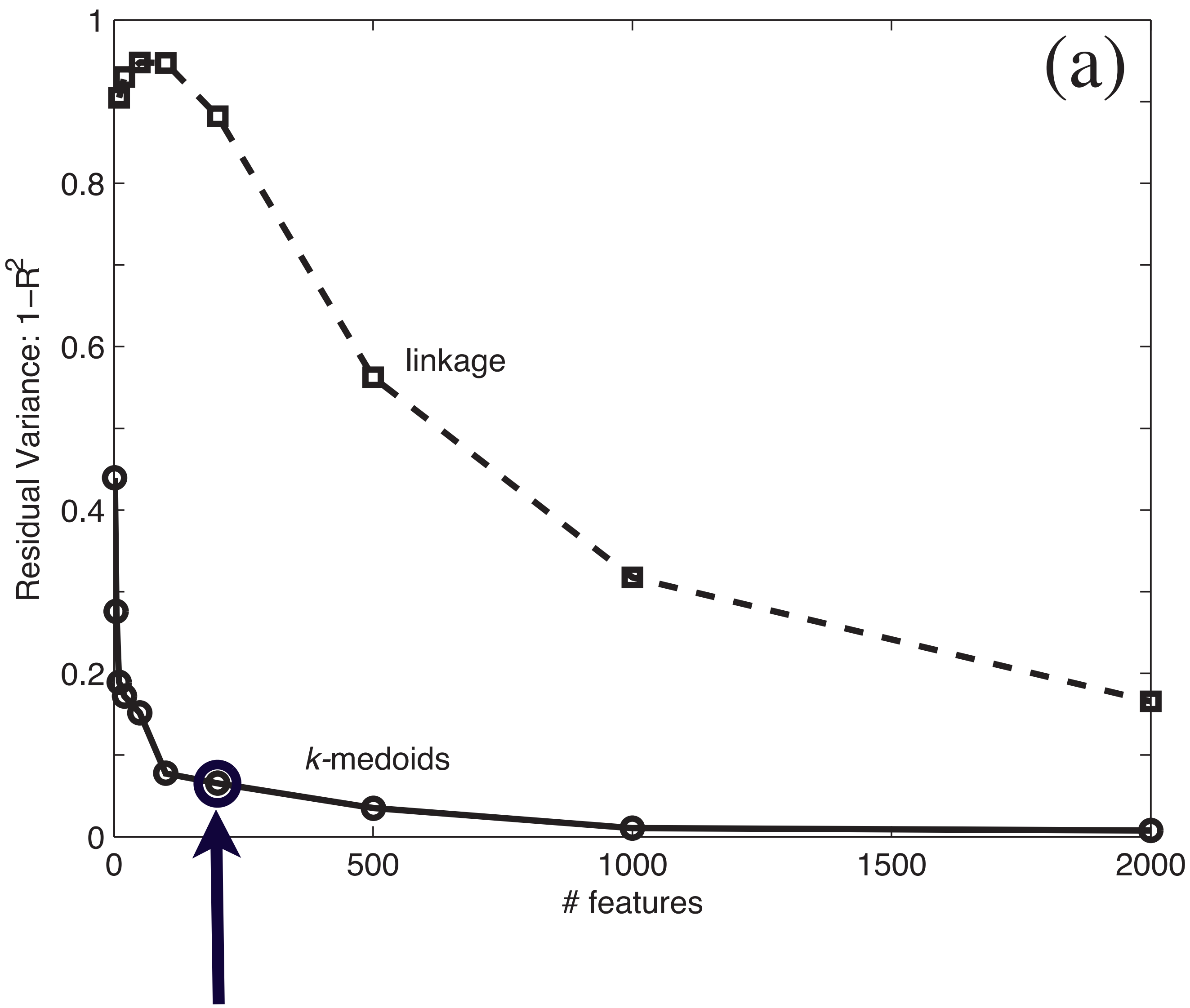
information theory

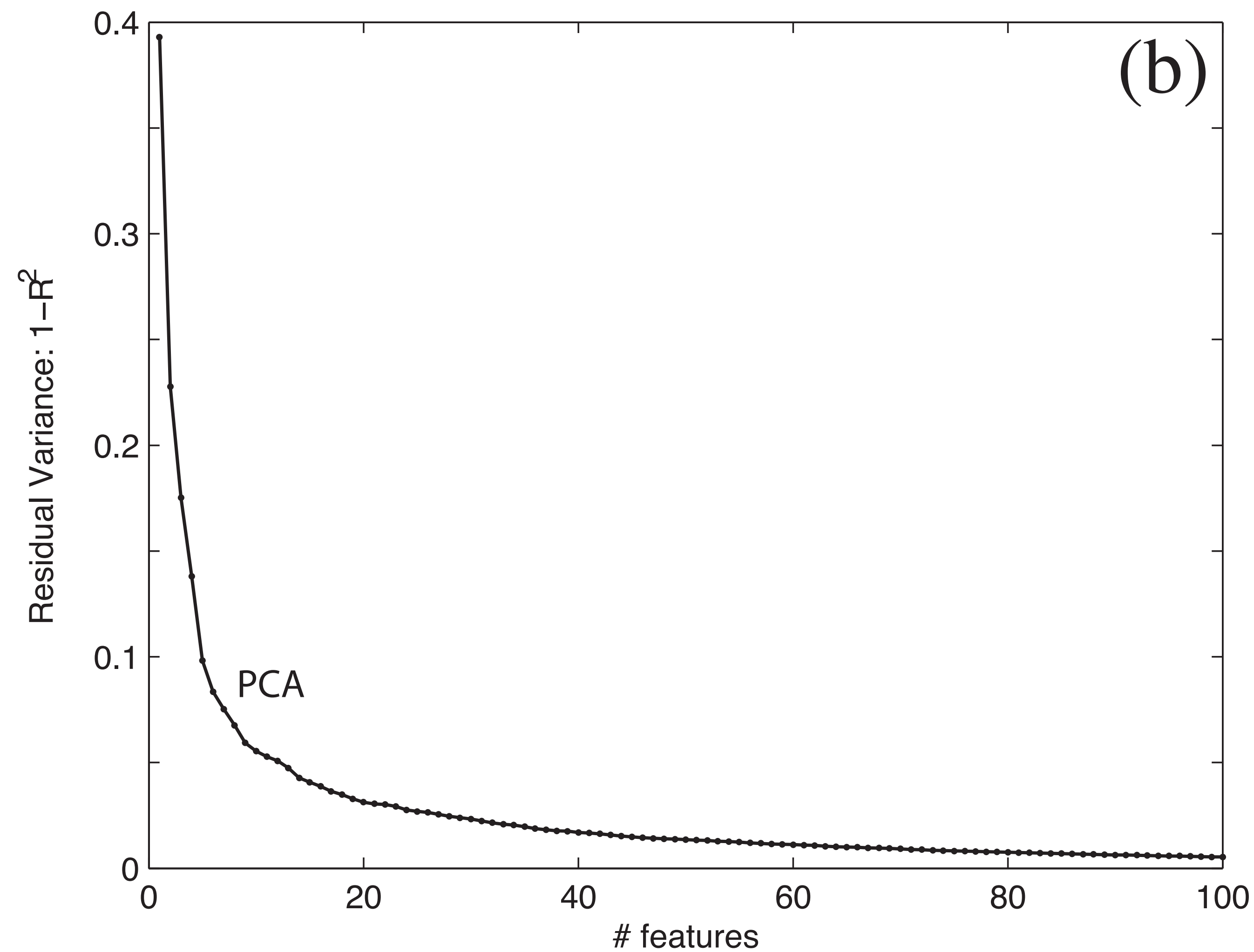
BIG PICTURE



distribution
time-independent structure
kernel-smoothed distribution
histogram
f
x
spread
location
outliers
distribution entropy
higher order moments
stationarity
properties that change with time
sliding windows
discrete partitions
mean in three segments
time series from natural and synthetic systems
scaling properties
fluctuation analysis
log(F)
log(τ)
nonlinear time series analysis
dimension estimates
lyapunov exponents
nonlinear prediction error
time-delay embedding
x_{i-2}
x_{i-1}
x_i
linear correlation
autocorrelation
spectral methods
short-range predictability
ARMA models
correlogram
correlation
lag
power spectrum
Power
frequency
Information theoretic
auto mutual information
entropy
symbolic entropy
B A C B C A A A C B C A B B C C C A A B
p(A|A), p(A|B), p(A|C)

How many operations are needed to efficiently summarize the structure we observe in empirical signals?

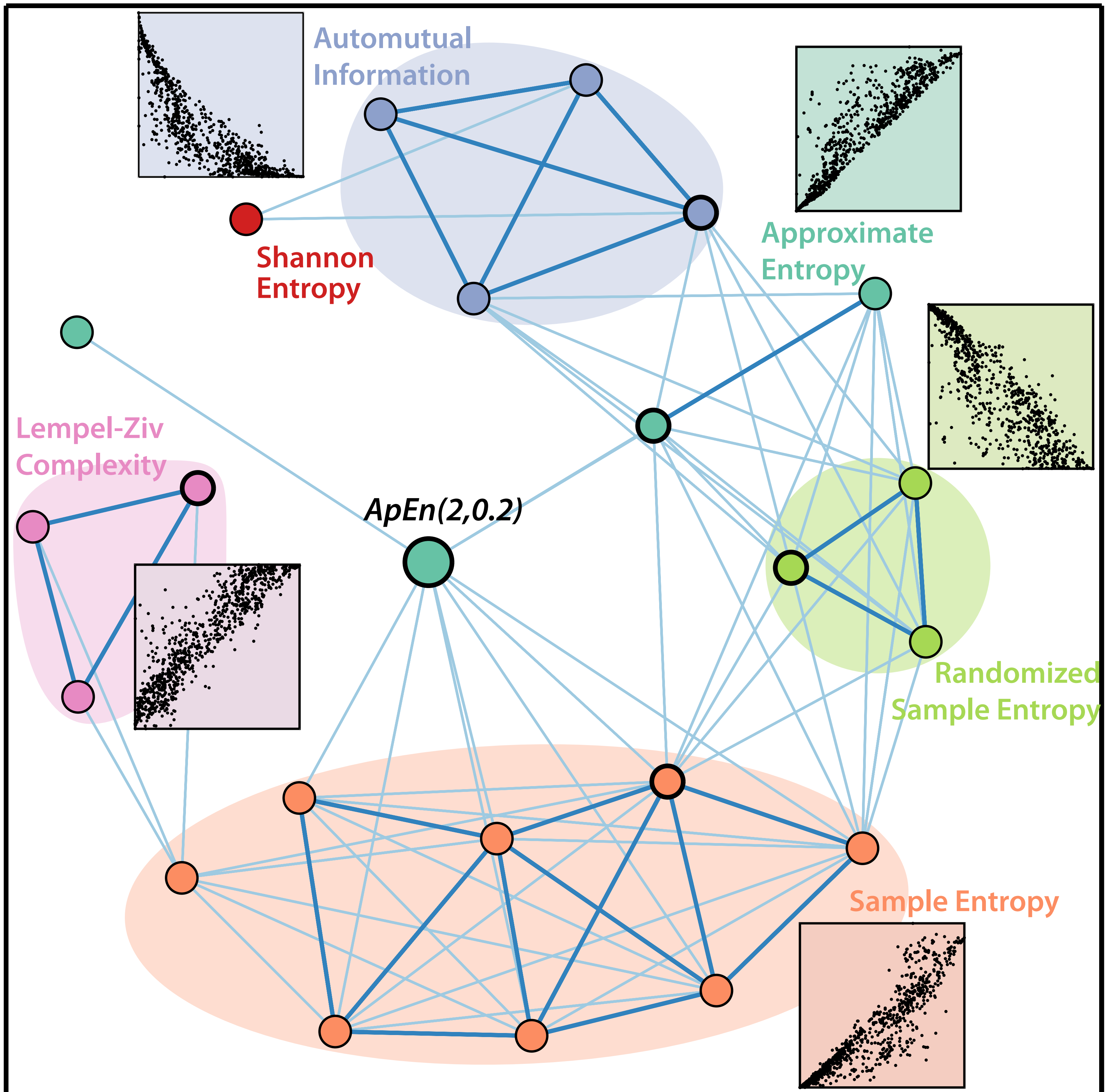




200 operations provide an efficient and interpretable summary

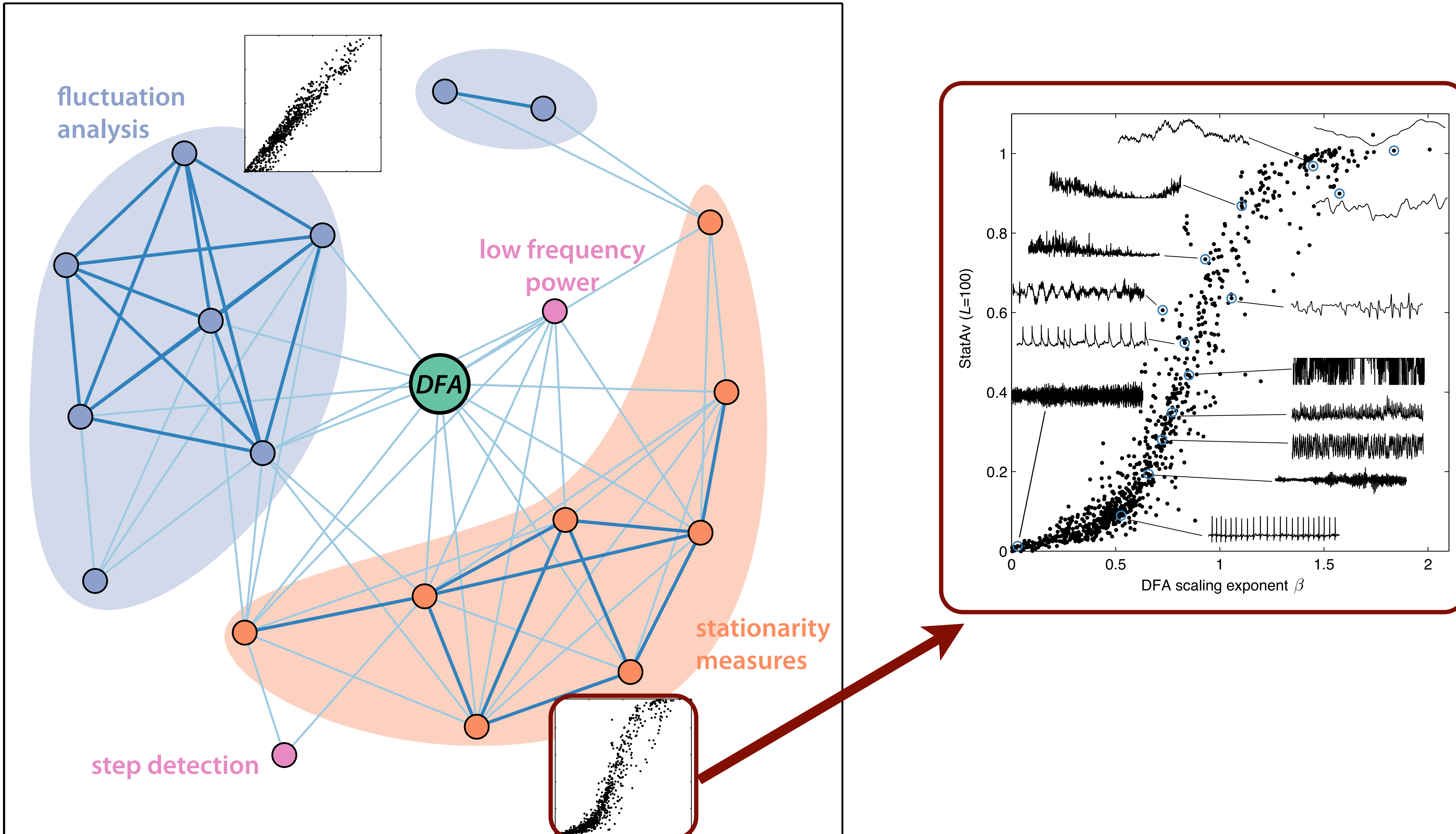
ZOOMING IN

Local Neighborhoods

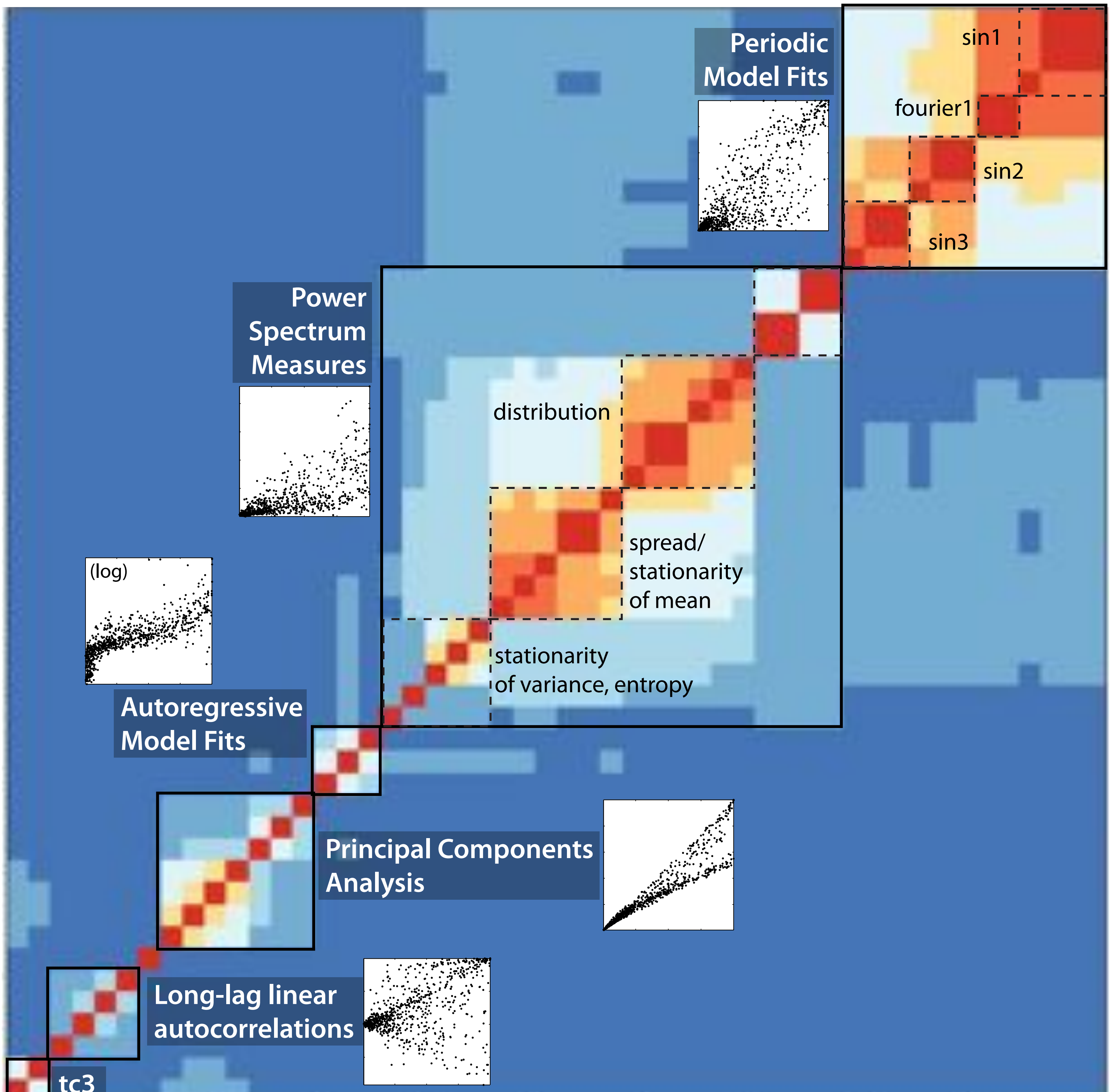


Organize our methods for time-series analysis

Local Neighborhoods

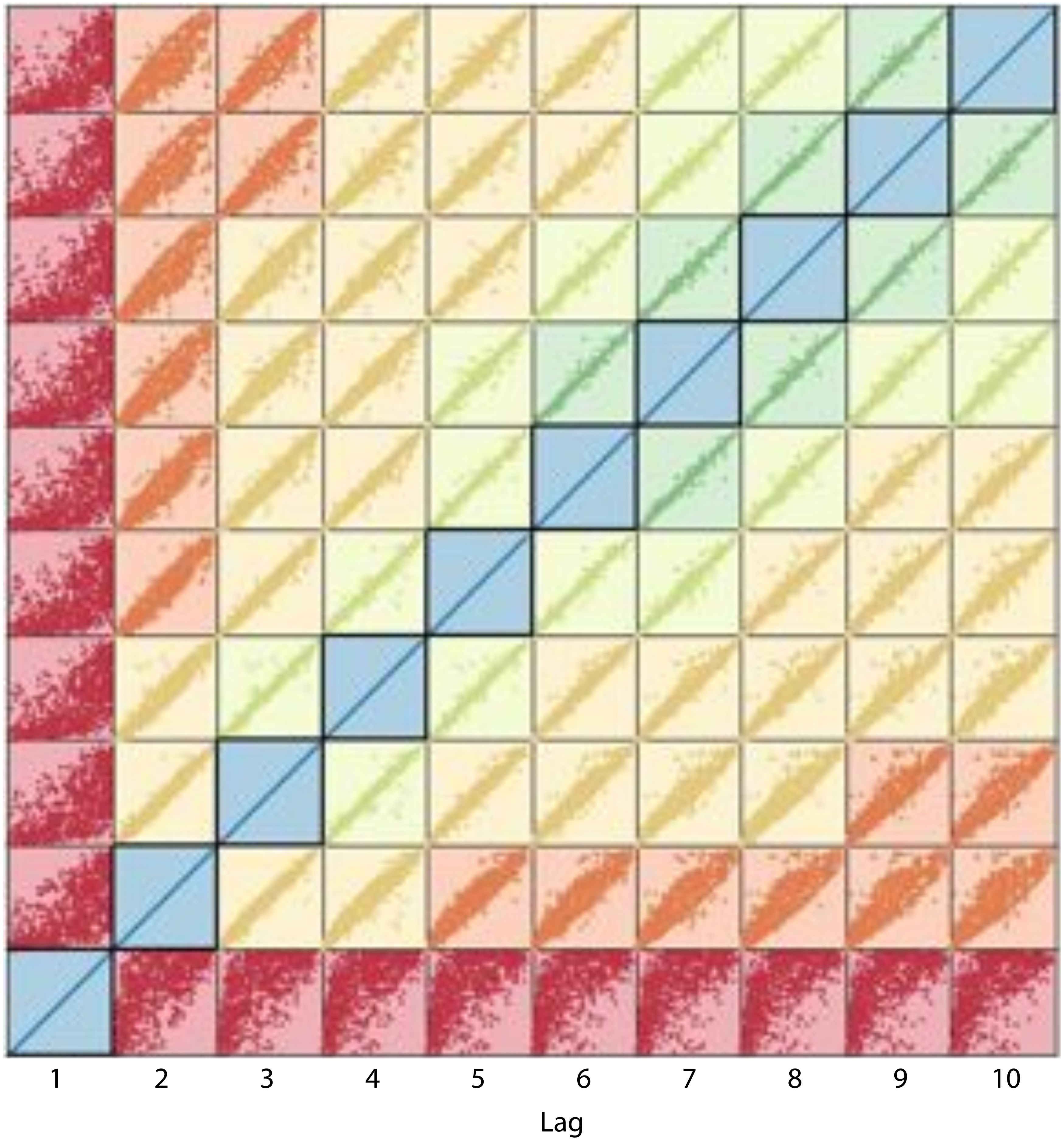


Local Neighborhoods

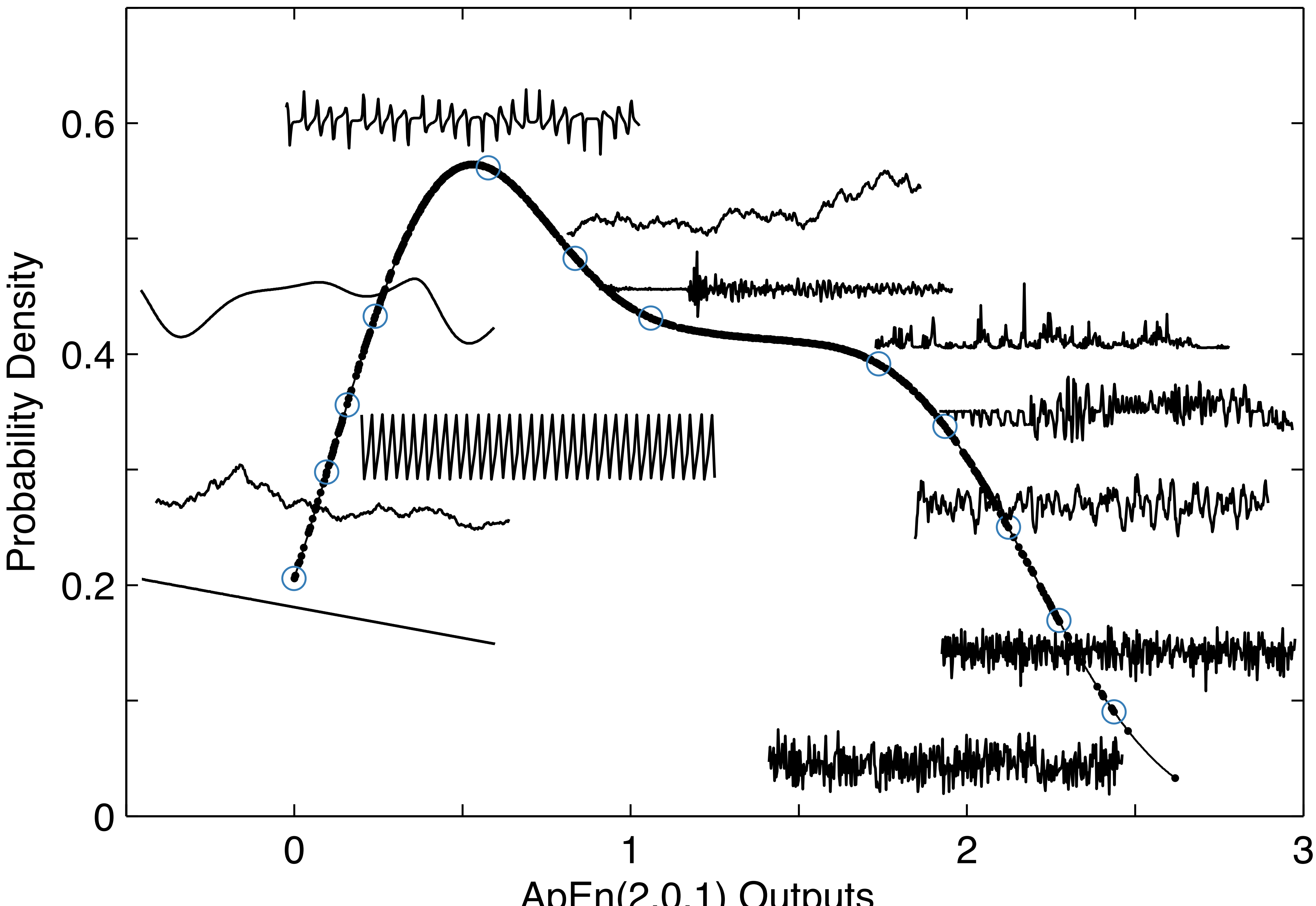


Organize our methods for time-series analysis

Automutual Information Measures

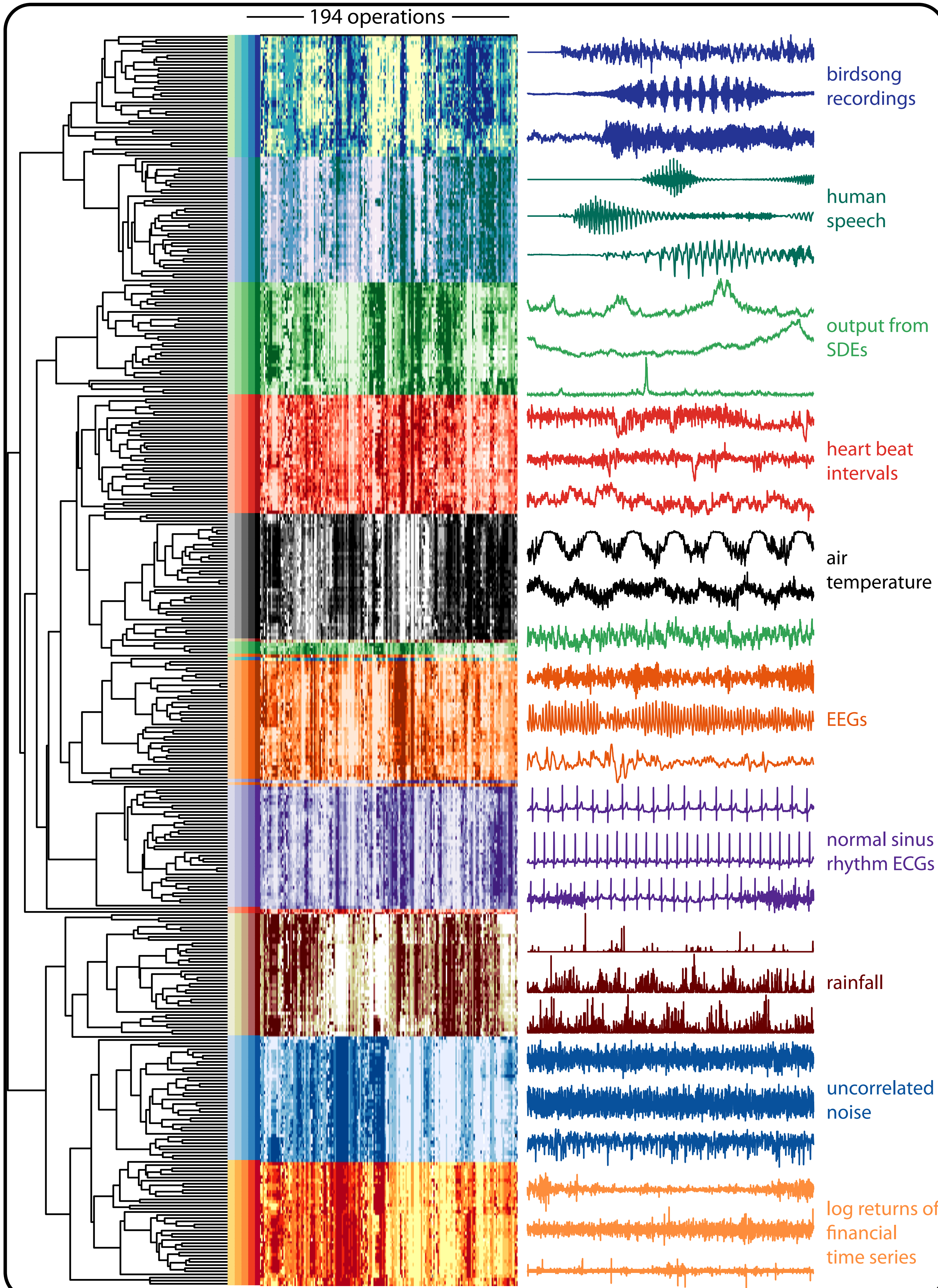


Visualize behavior

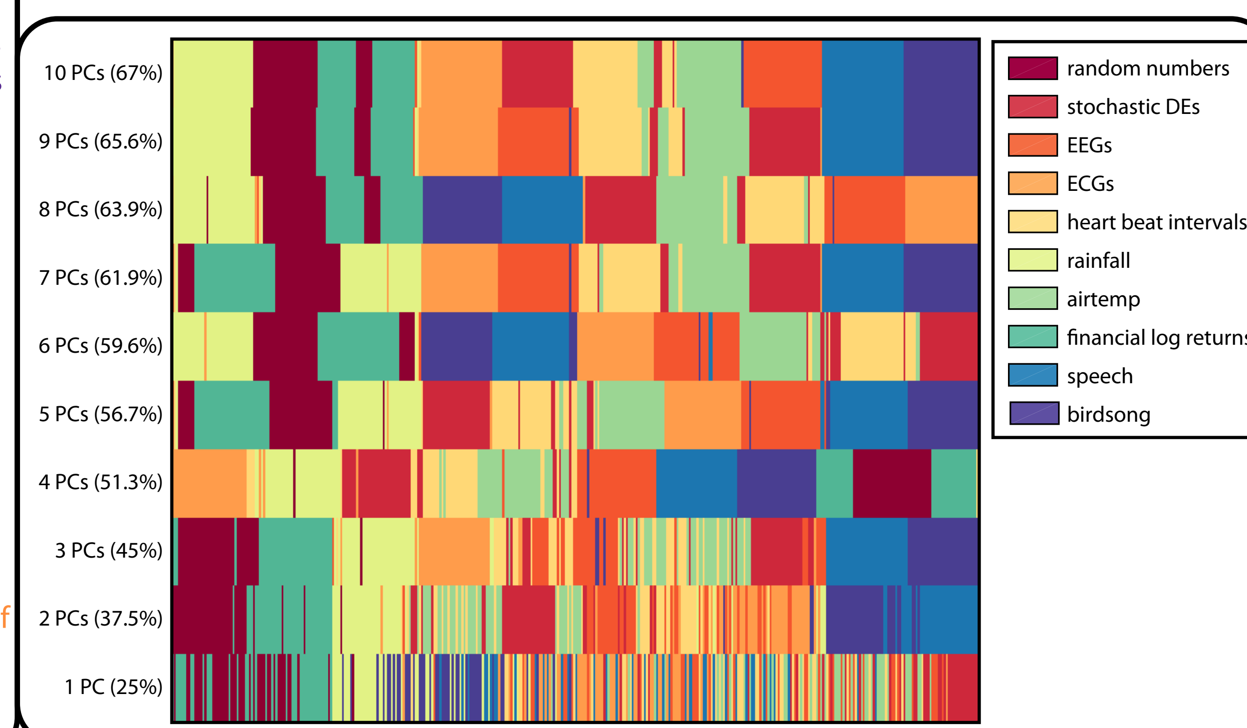


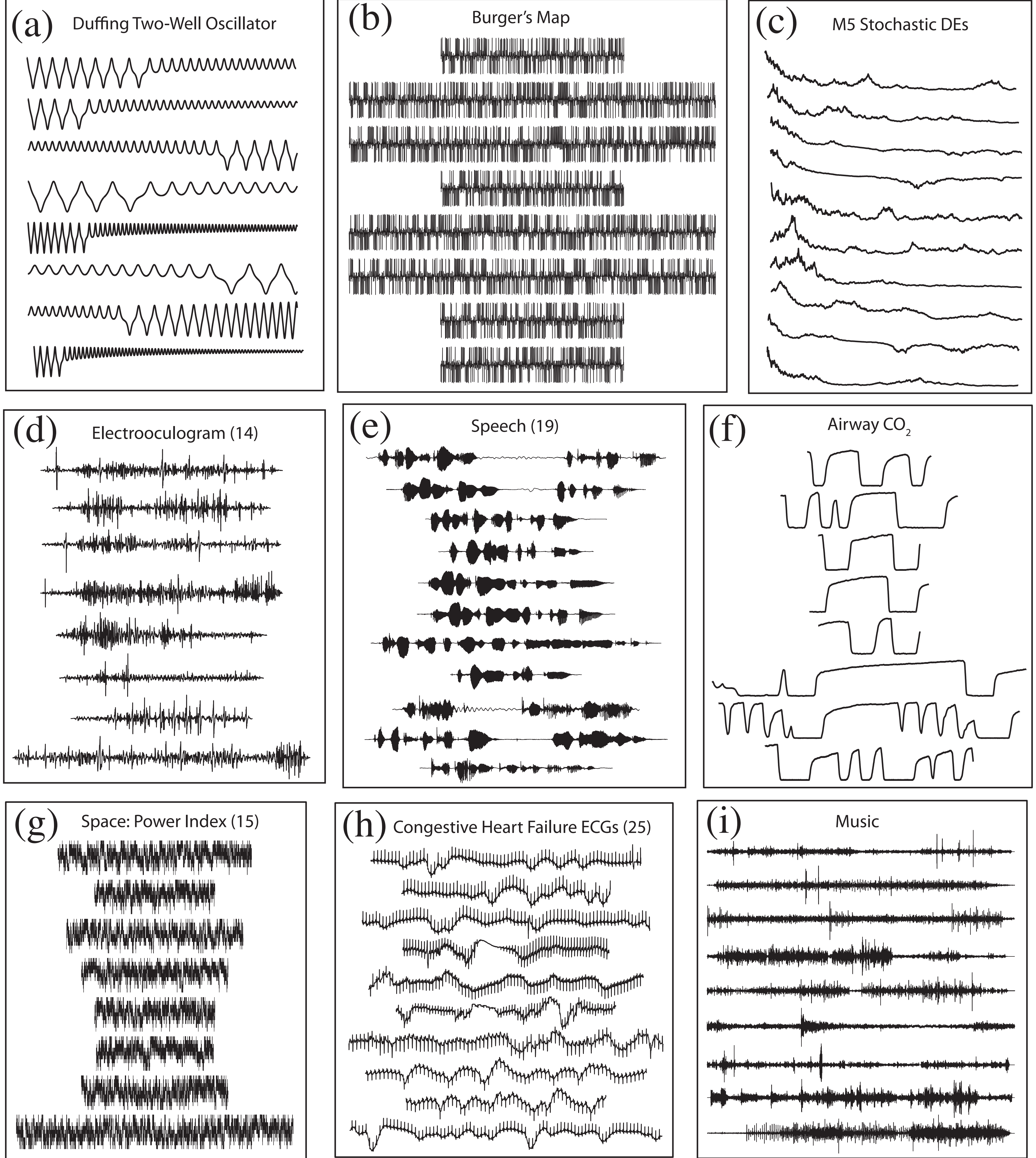
Organizing Our Data

- Our reduced interdisciplinary set of operations is a powerful summary of the structure in empirical time series
- Links between real-world and synthetically-generated time series encourage a unified, collaborative framework for understanding the dynamics in time series

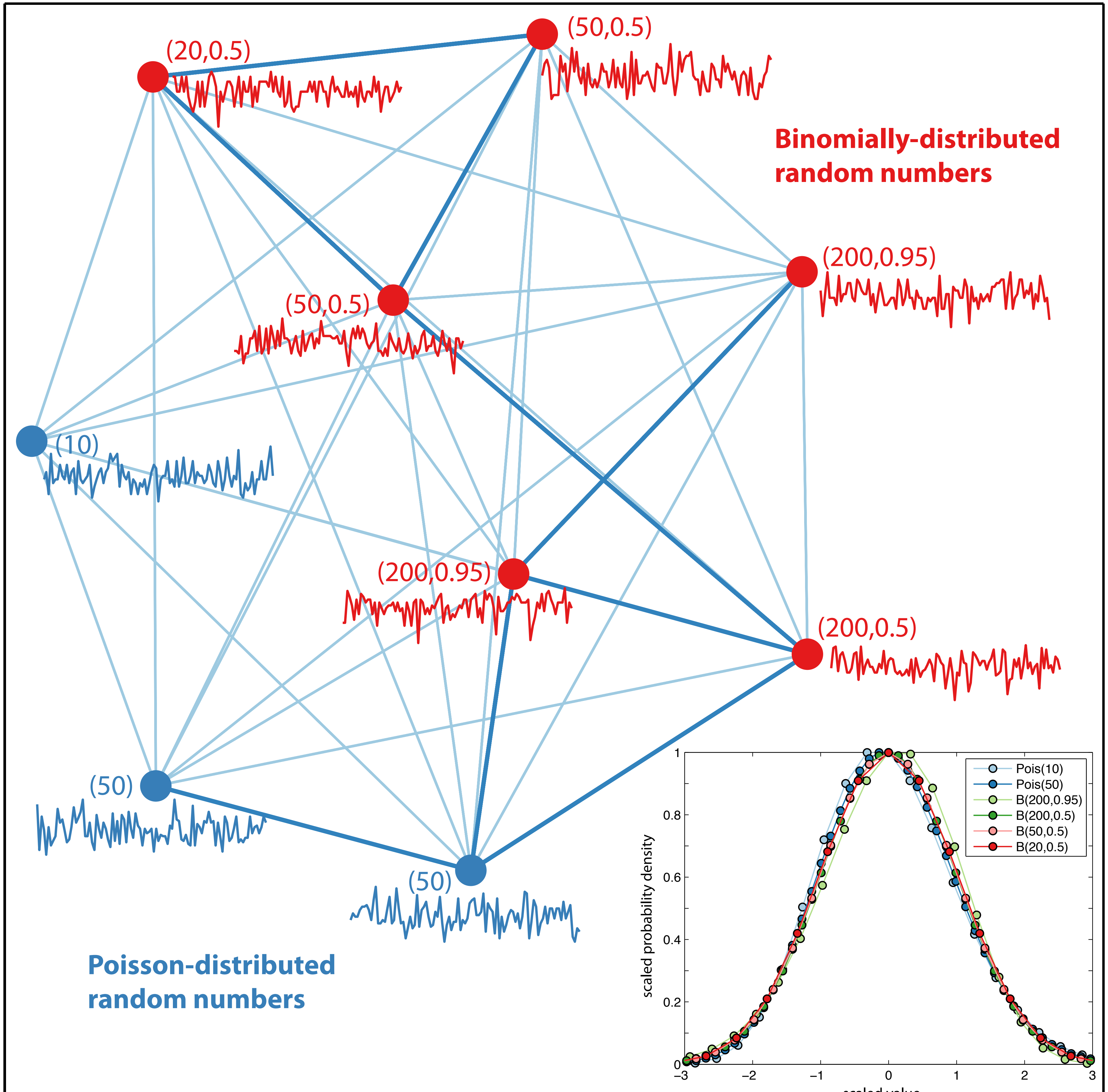


broad classes of dynamics are distinguished

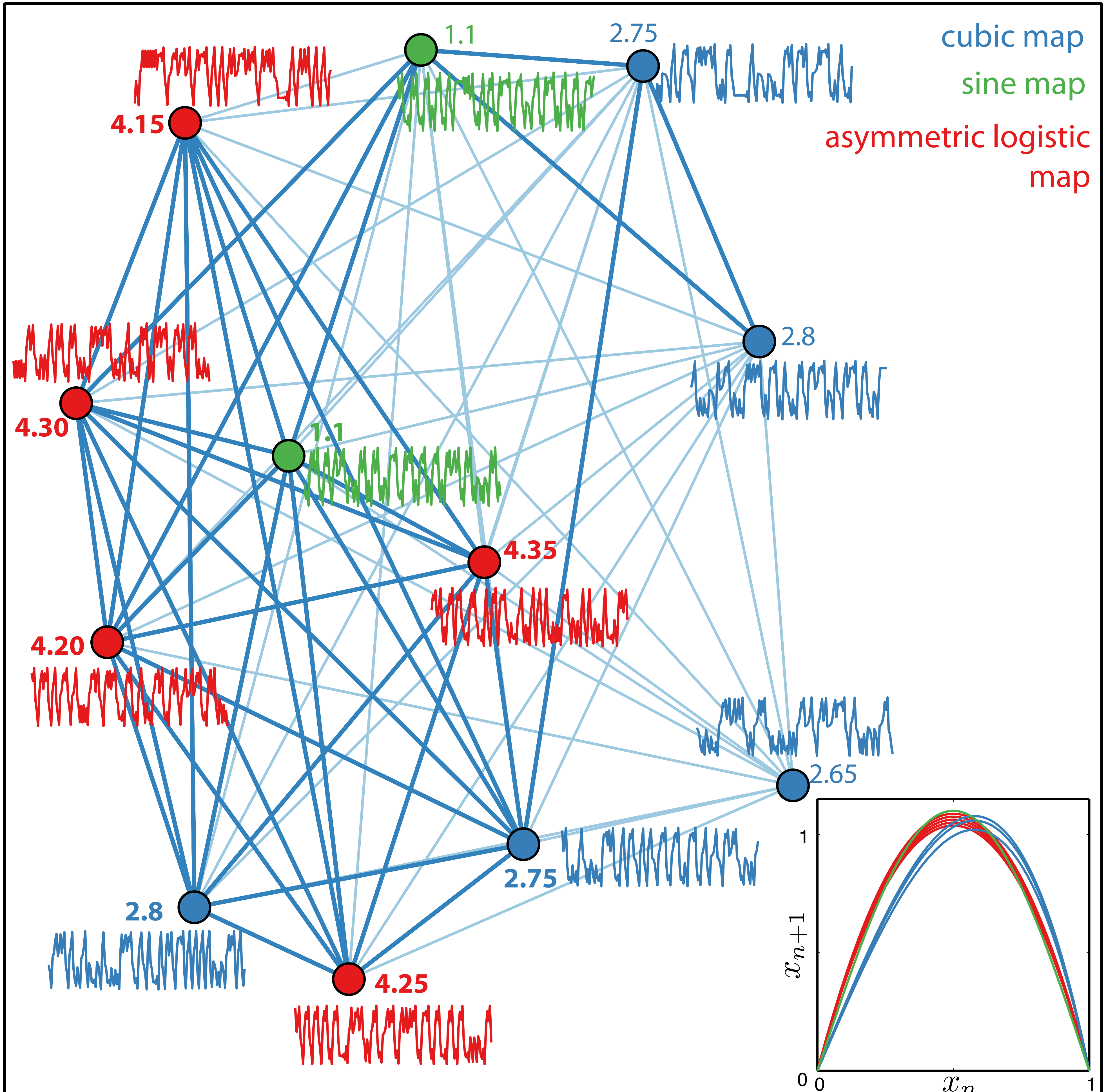




(a) Duffing Two-Well Oscillator
(b) Burger's Map
(c) M5 Stochastic DEs
(d) Electrooculogram (14)
(e) Speech (19)
(f) Airway CO2
(g) Space: Power Index (15)
(h) Congestive Heart Failure ECGs (25)
(i) Music



Binomially-distributed random numbers
Poisson-distributed random numbers
(20,0.5)
(50,0.5)
(200,0.95)
(50,0.5)
(10)
(200,0.95)
(200,0.5)
(50)
(50)
Pois(10)
Pois(50)
B(200,0.95)
B(200,0.5)
B(50,0.5)
B(20,0.5)
scaled probability density
scaled value



cubic map
sine map
asymmetric logistic map
1.1
2.75
4.15
2.8
4.30
1.1
4.35
4.20
2.65
2.75
2.8
4.25
x_{n+1}
x_n
0
1
0
1

Time-series models?

- Real data are recordings of dynamics
- Time-series models generate dynamics with a known mechanism

$$dx/dt = y$$
$$dy/dt = z$$
$$dz/dt = -z - (T - R + Rx^2)y - Tx$$

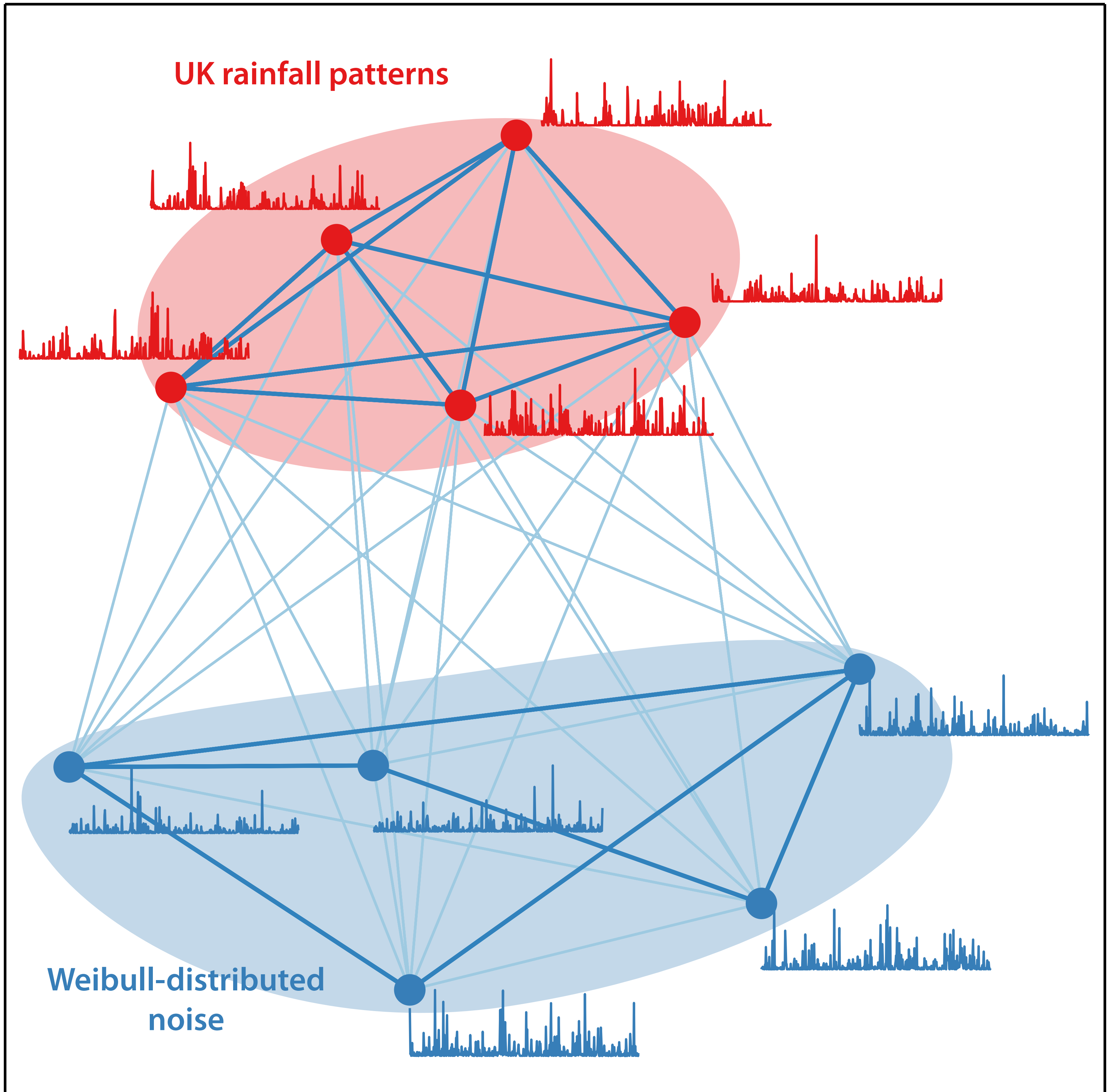
speech recording

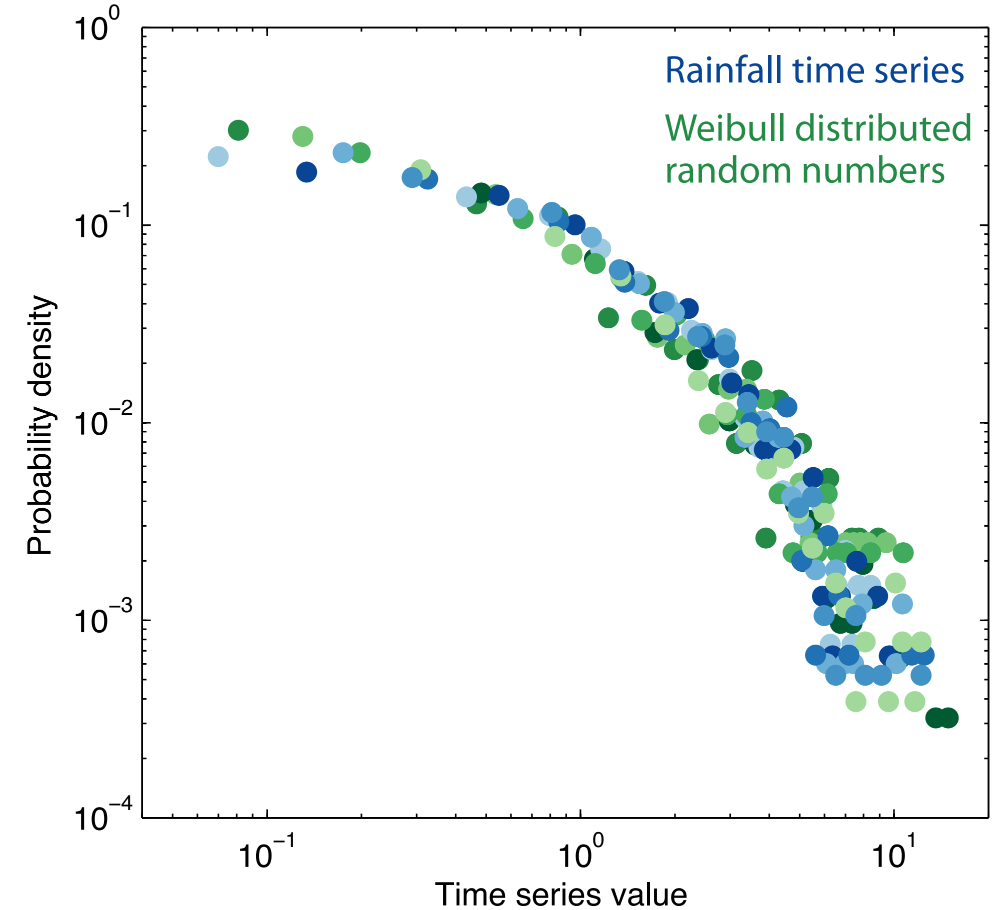
"he did it"

"he did it"

Pointing the finger

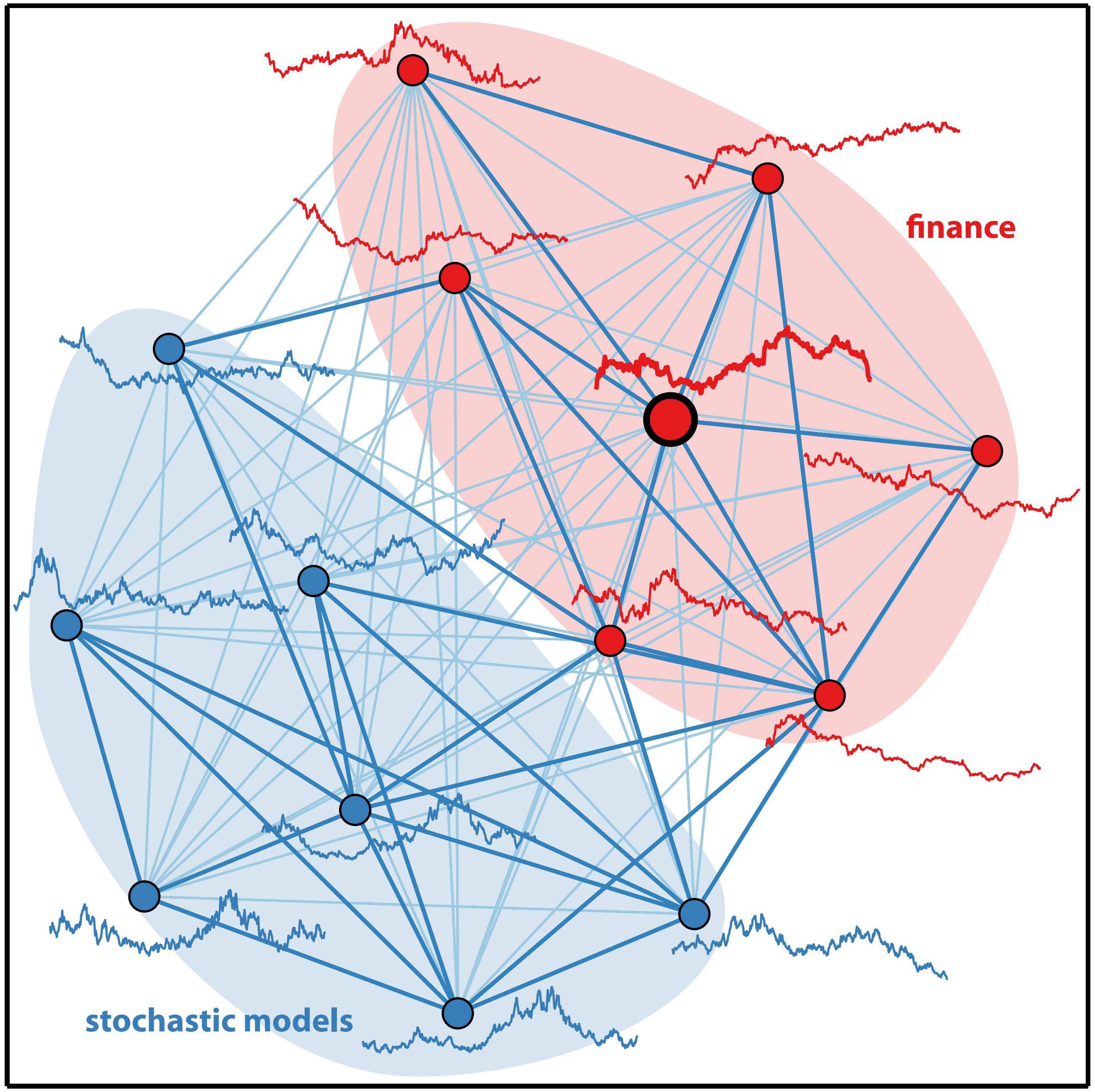
Similarity Search

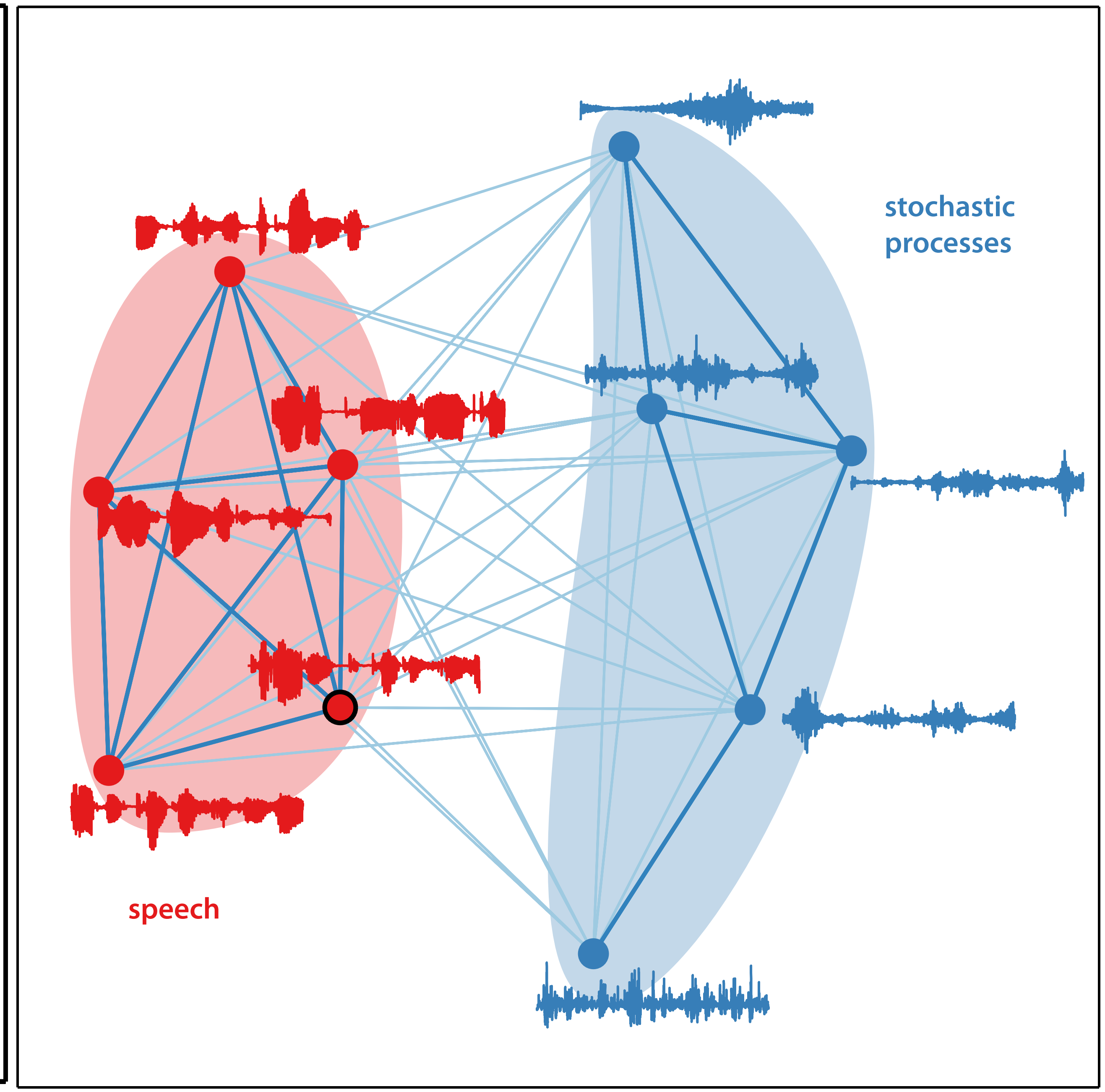




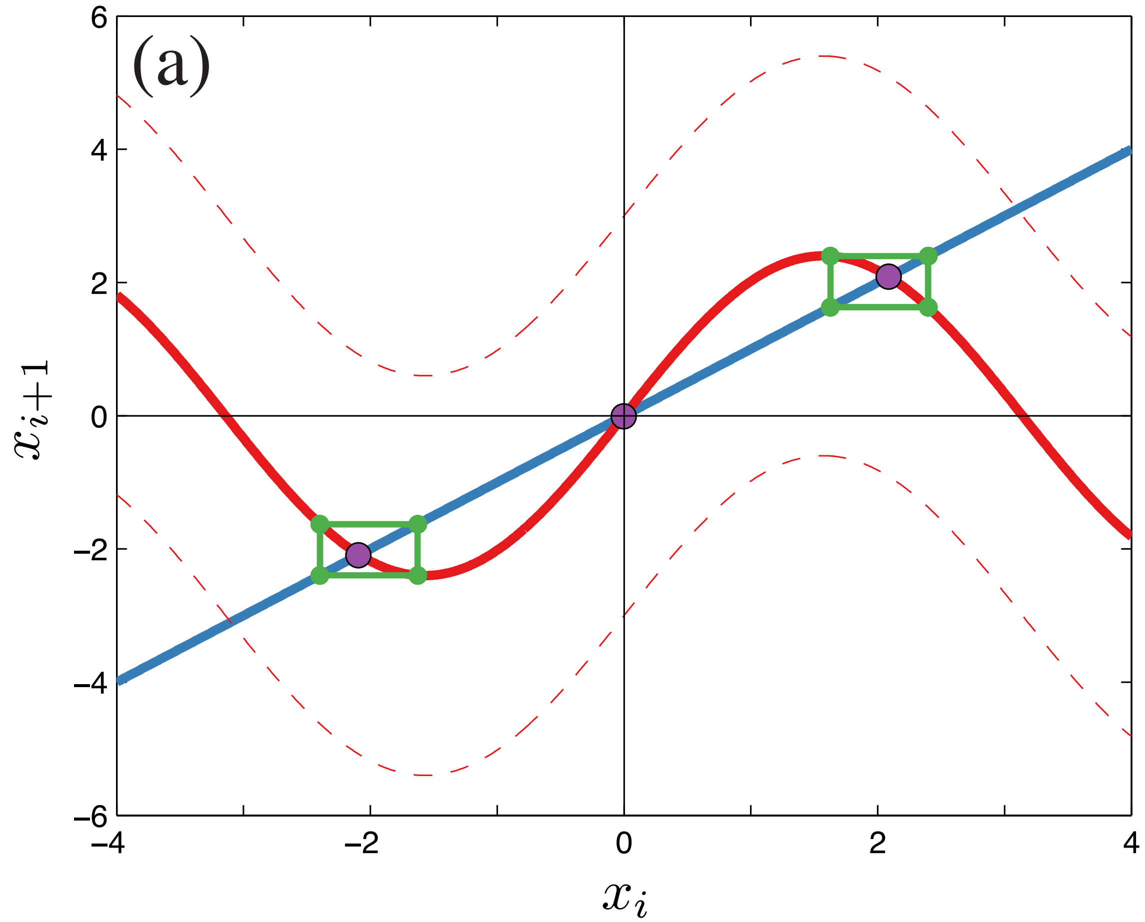
suggest models for our data

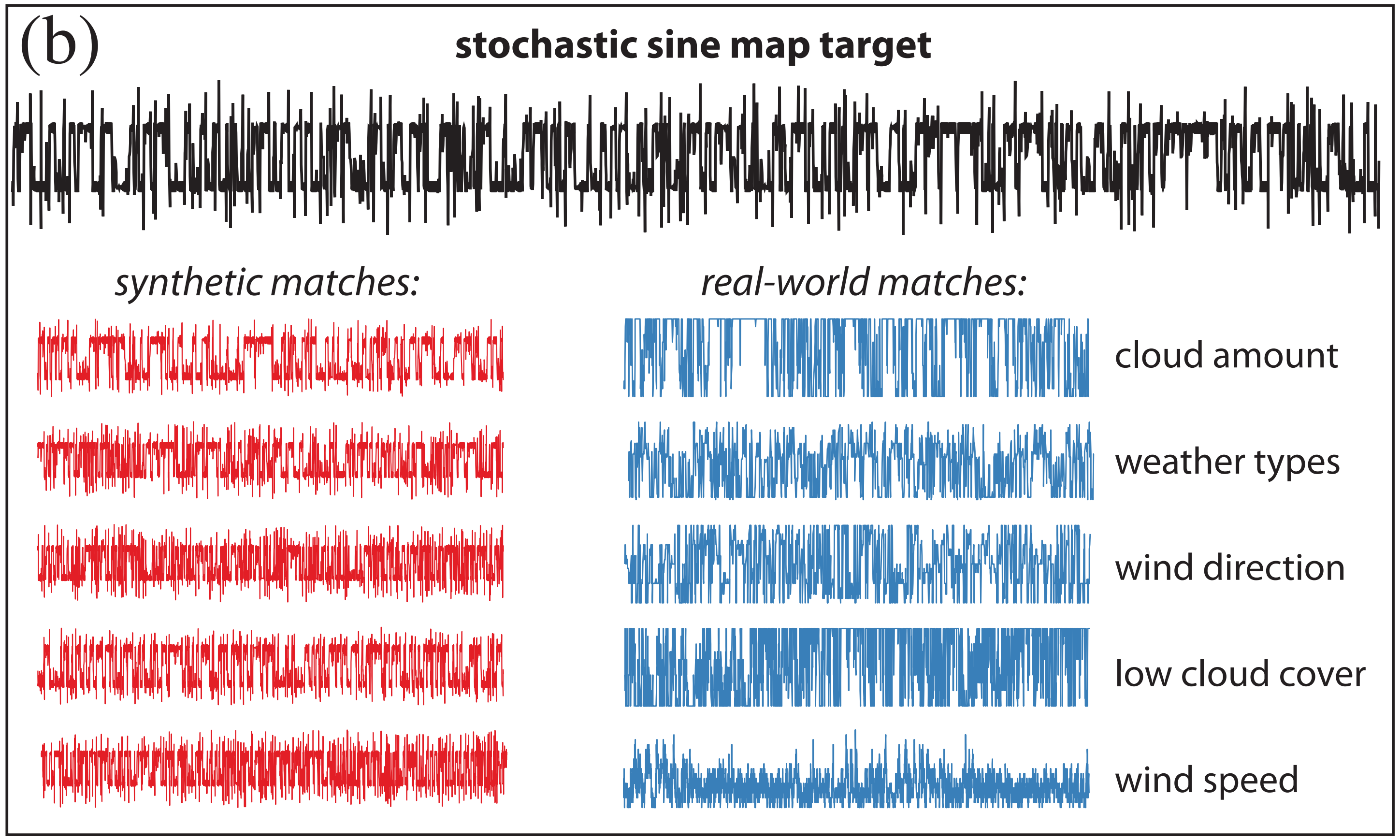
Similarity Search



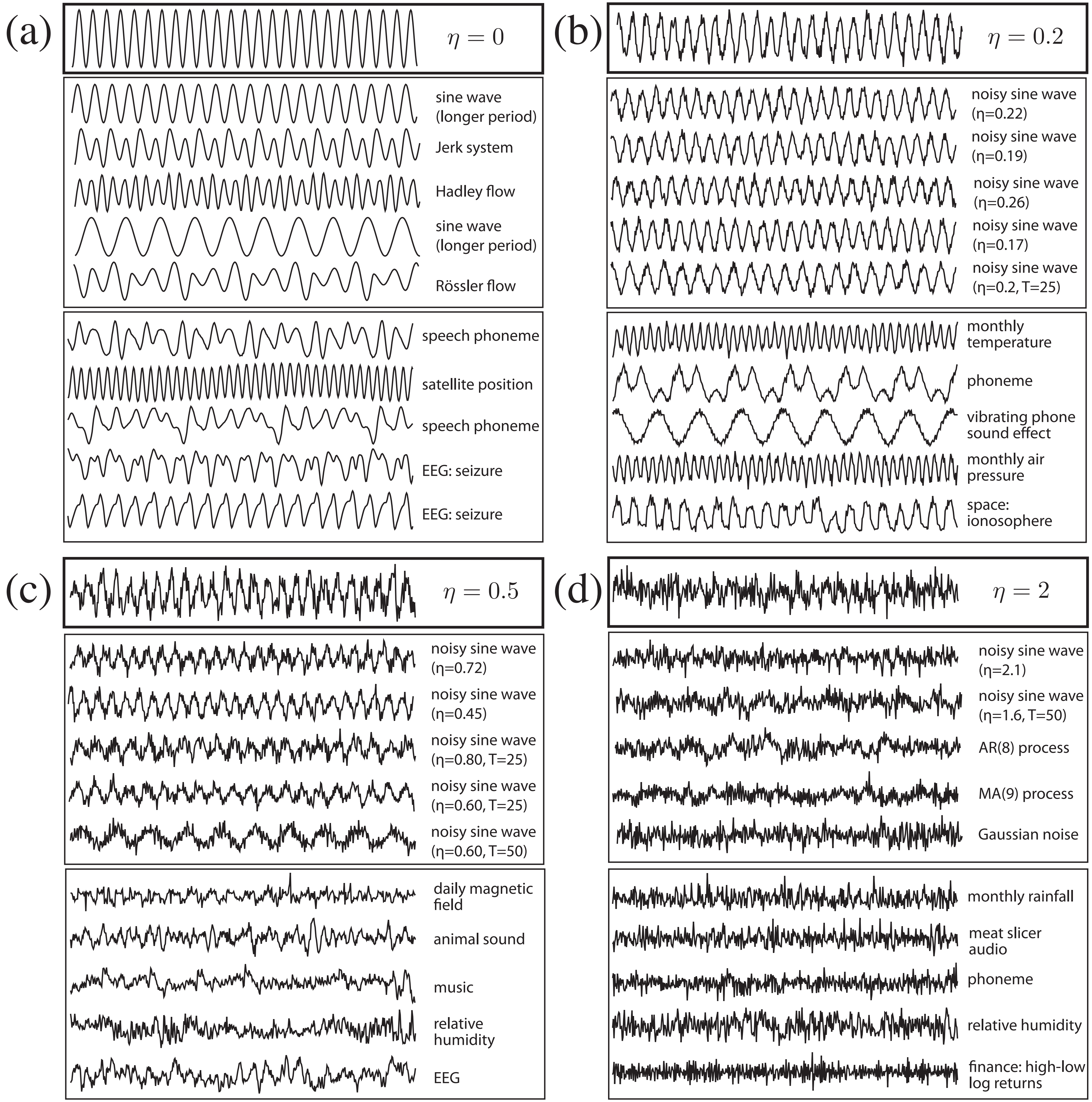


suggest models
for our data





suggest data for
our models

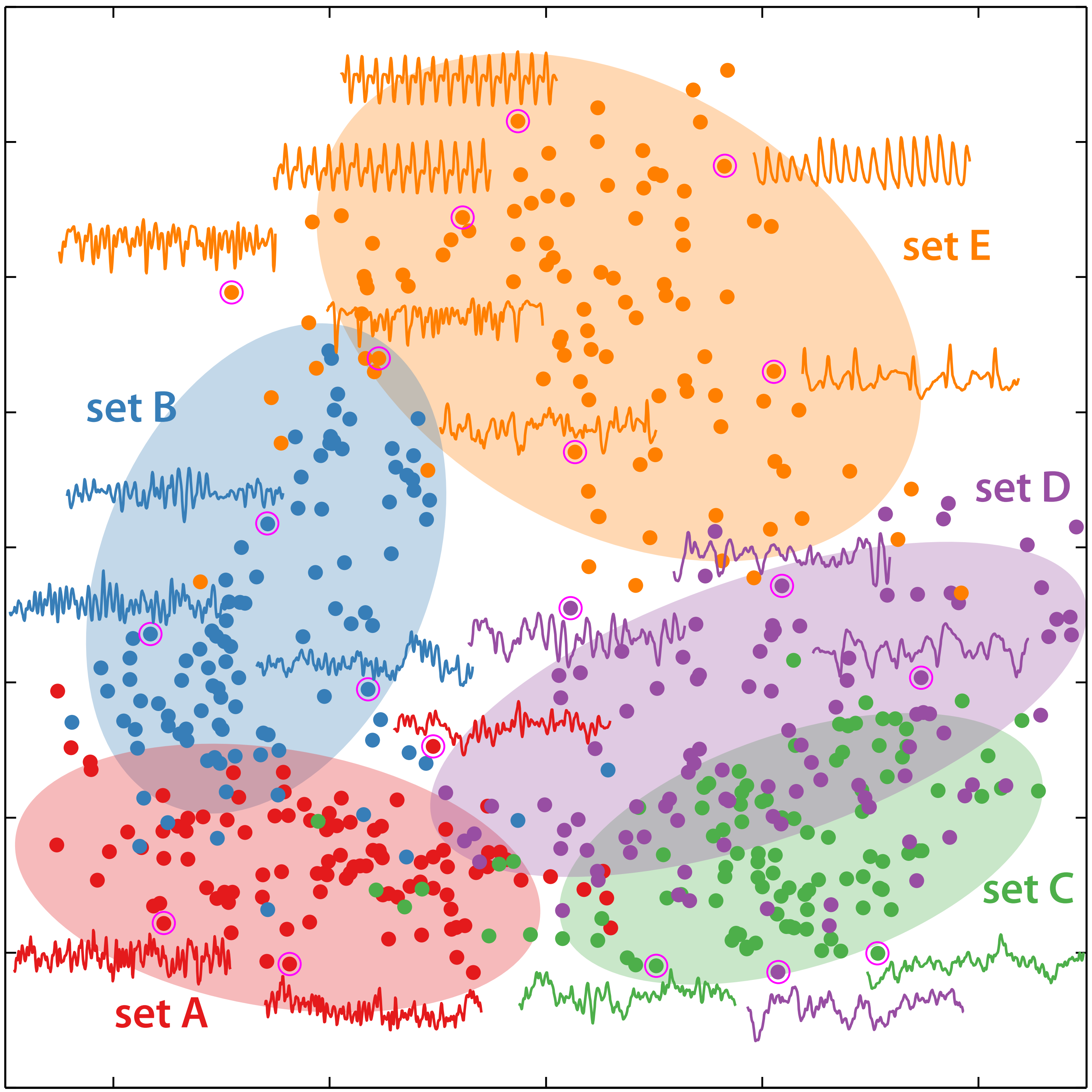


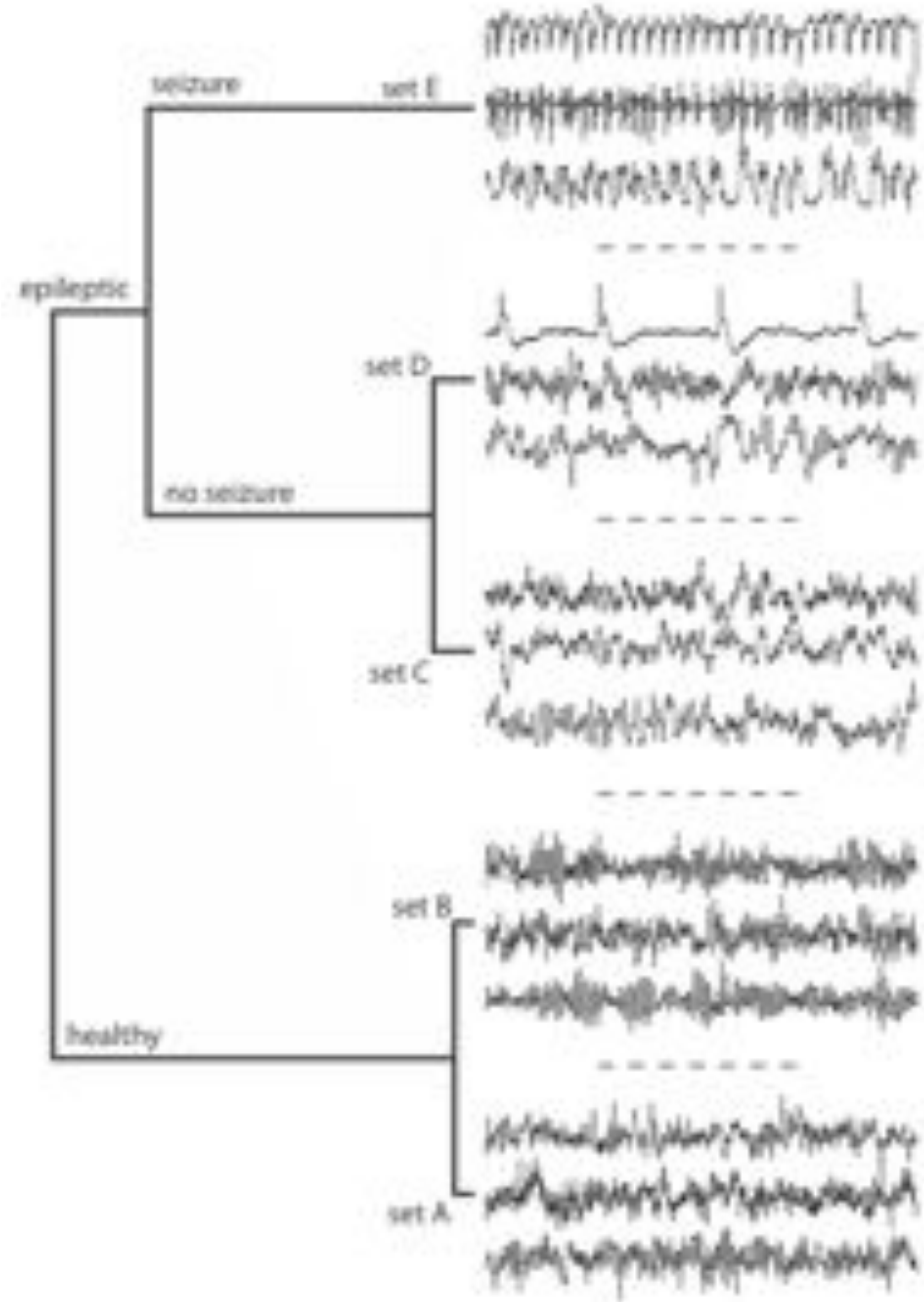
(a)
$\eta = 0$
sine wave (longer period)
Jerk system
Hadley flow
sine wave (longer period)
Rössler flow
speech phoneme
satellite position
speech phoneme
EEG: seizure
EEG: seizure
(b)
$\eta = 0.2$
noisy sine wave (η=0.22)
noisy sine wave (η=0.19)
noisy sine wave (η=0.26)
noisy sine wave (η=0.17)
noisy sine wave (η=0.2, T=25)
monthly temperature
phoneme
vibrating phone sound effect
monthly air pressure
space: ionosophere
(c)
$\eta = 0.5$
noisy sine wave (η=0.72)
noisy sine wave (η=0.45)
noisy sine wave (η=0.80, T=25)
noisy sine wave (η=0.60, T=25)
noisy sine wave (η=0.60, T=50)
daily magnetic field
animal sound
music
relative humidity
EEG
(d)
$\eta = 2$
noisy sine wave (η=2.1)
noisy sine wave (η=1.6, T=50)
AR(8) process
MA(9) process
Gaussian noise
monthly rainfall
meat slicer audio
phoneme
relative humidity
finance: high-low log returns

Applications

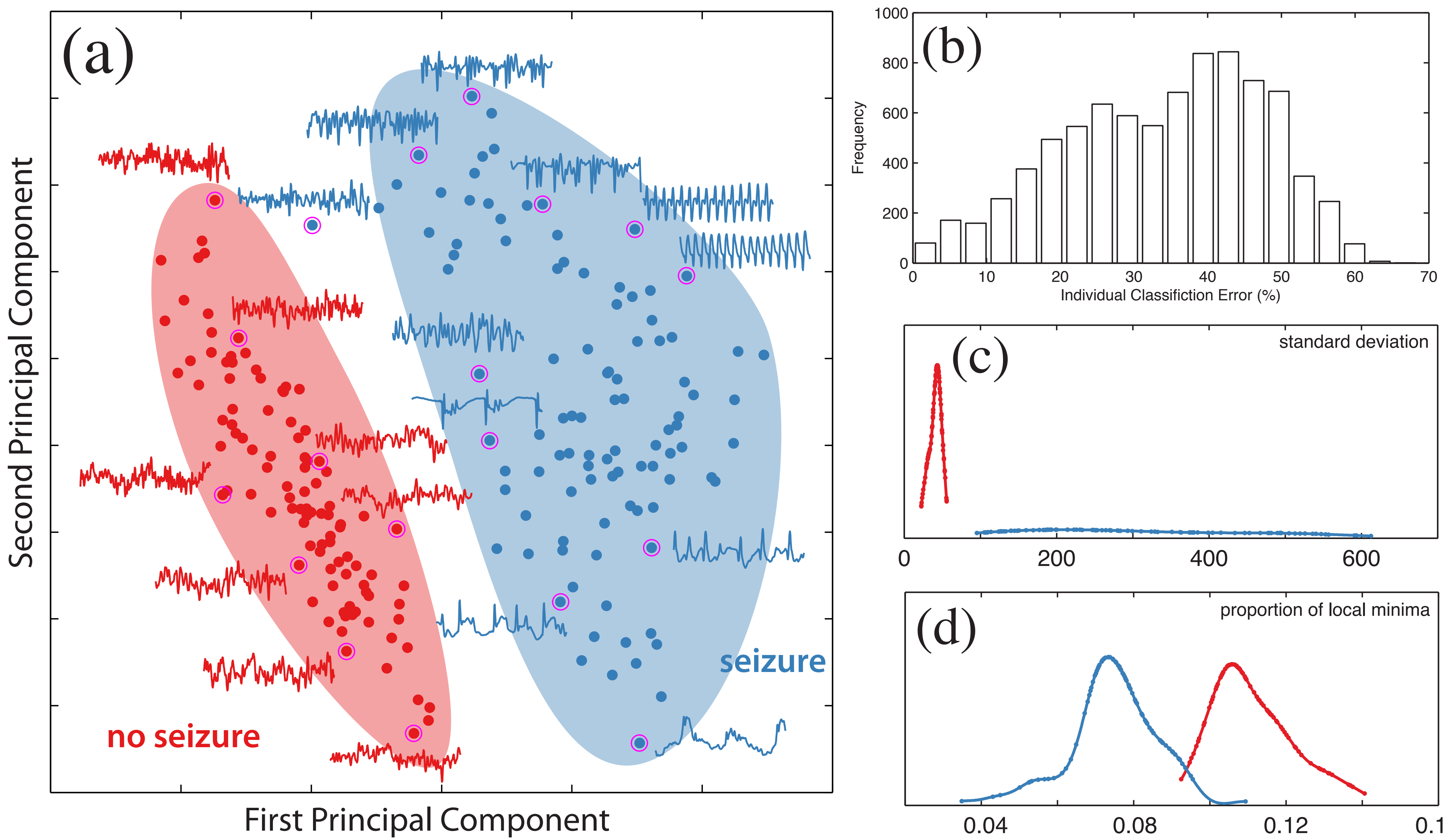
- Drawing on a rich, interdisciplinary database of methods for time-series analysis allows datasets to be analyzed in new ways
- Reveal structure using PCA
- Select interpretable families of useful methods for a given classification/ regression task

EEGs

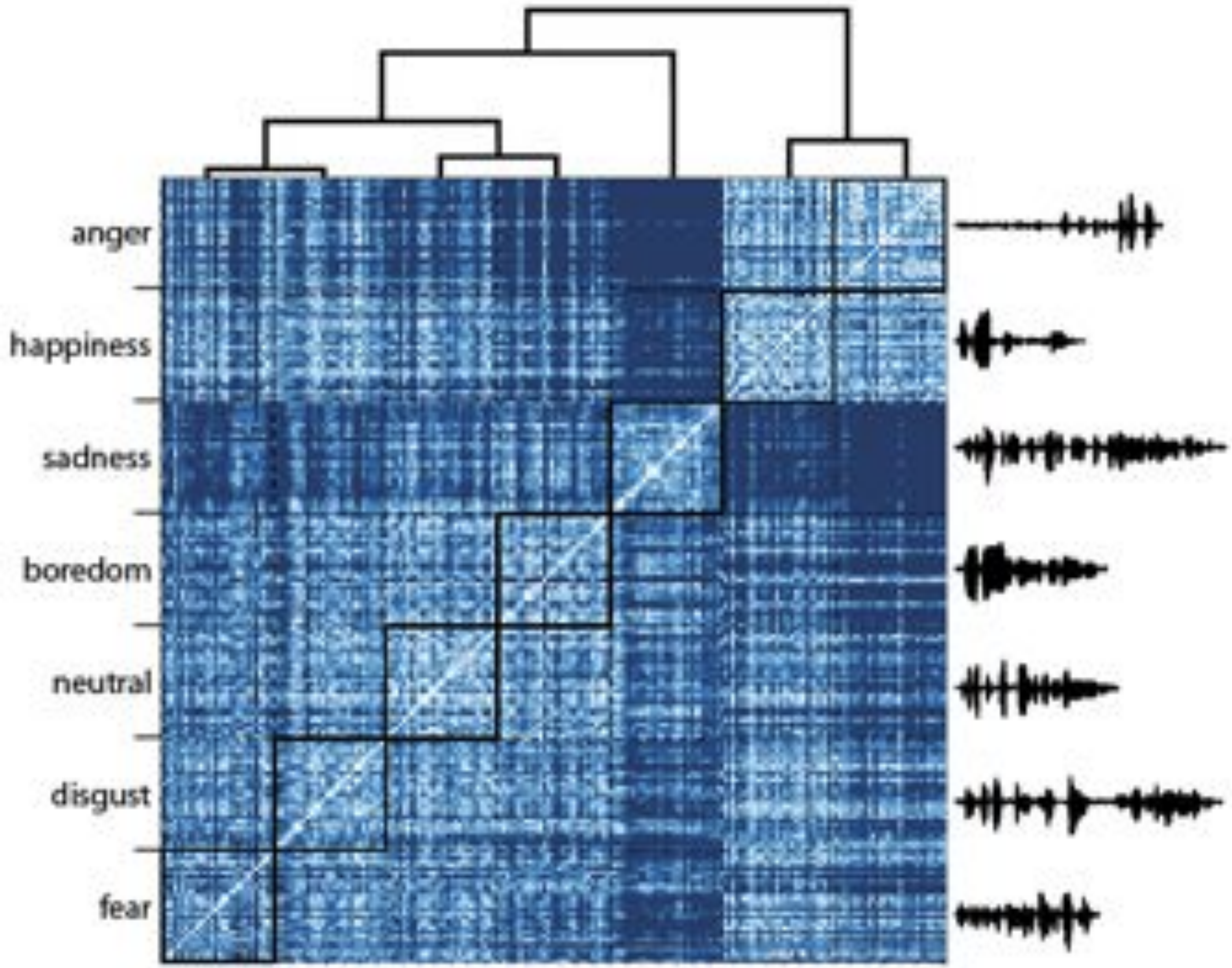

set E
set B
set D
set C
set A


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

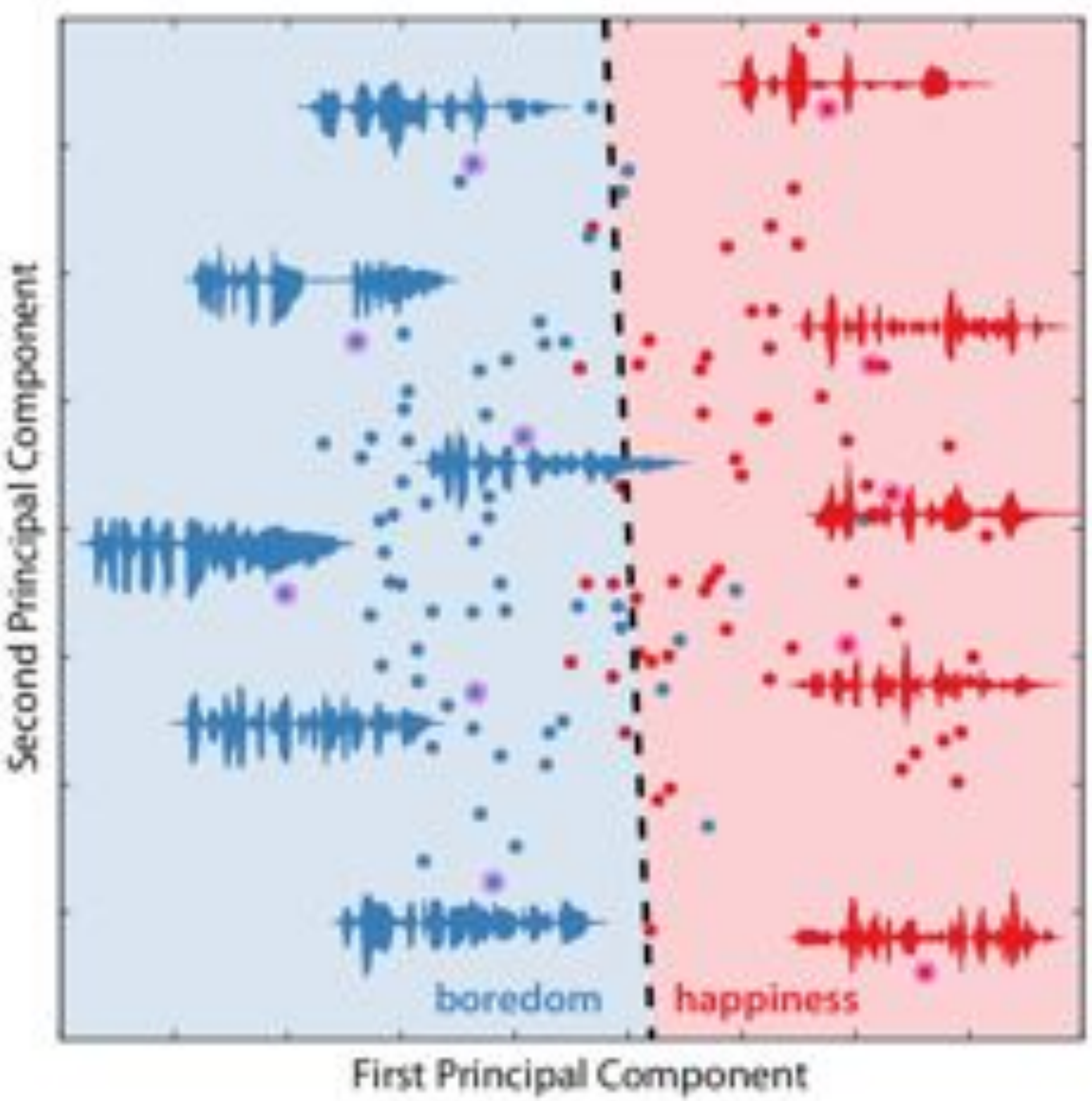
Distinguishing seizures

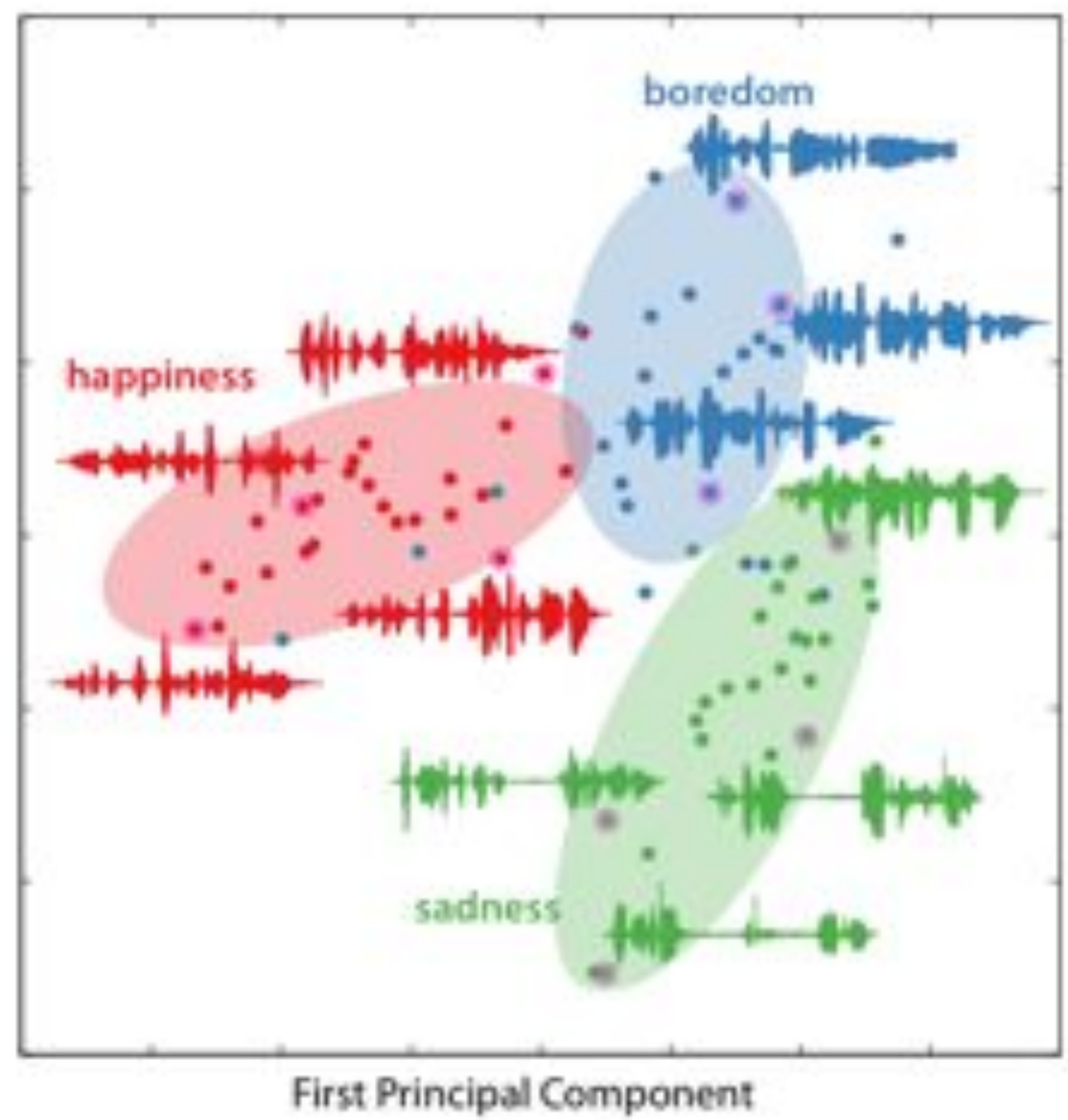


Emotional Speech

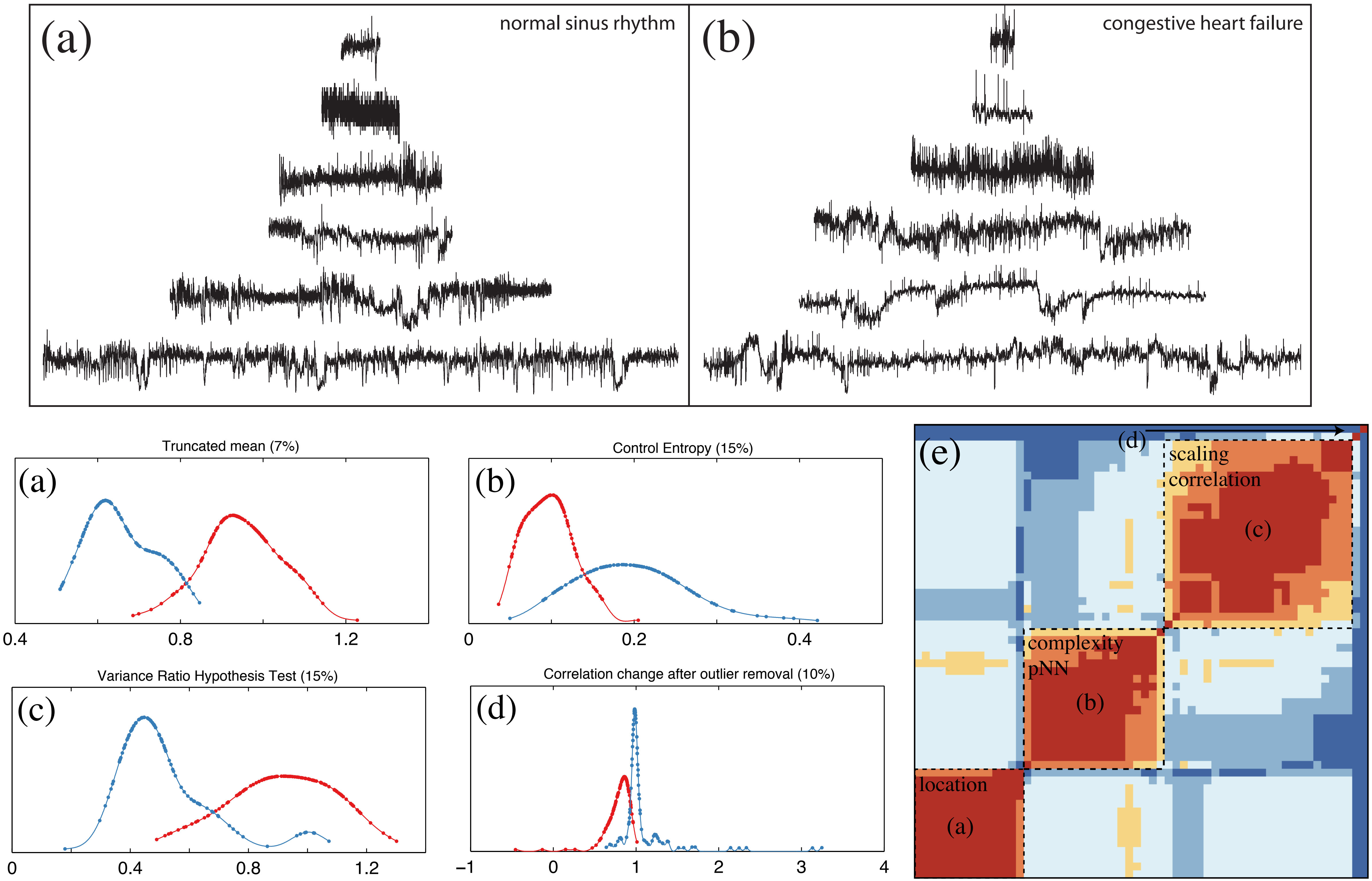


Emotional Speech

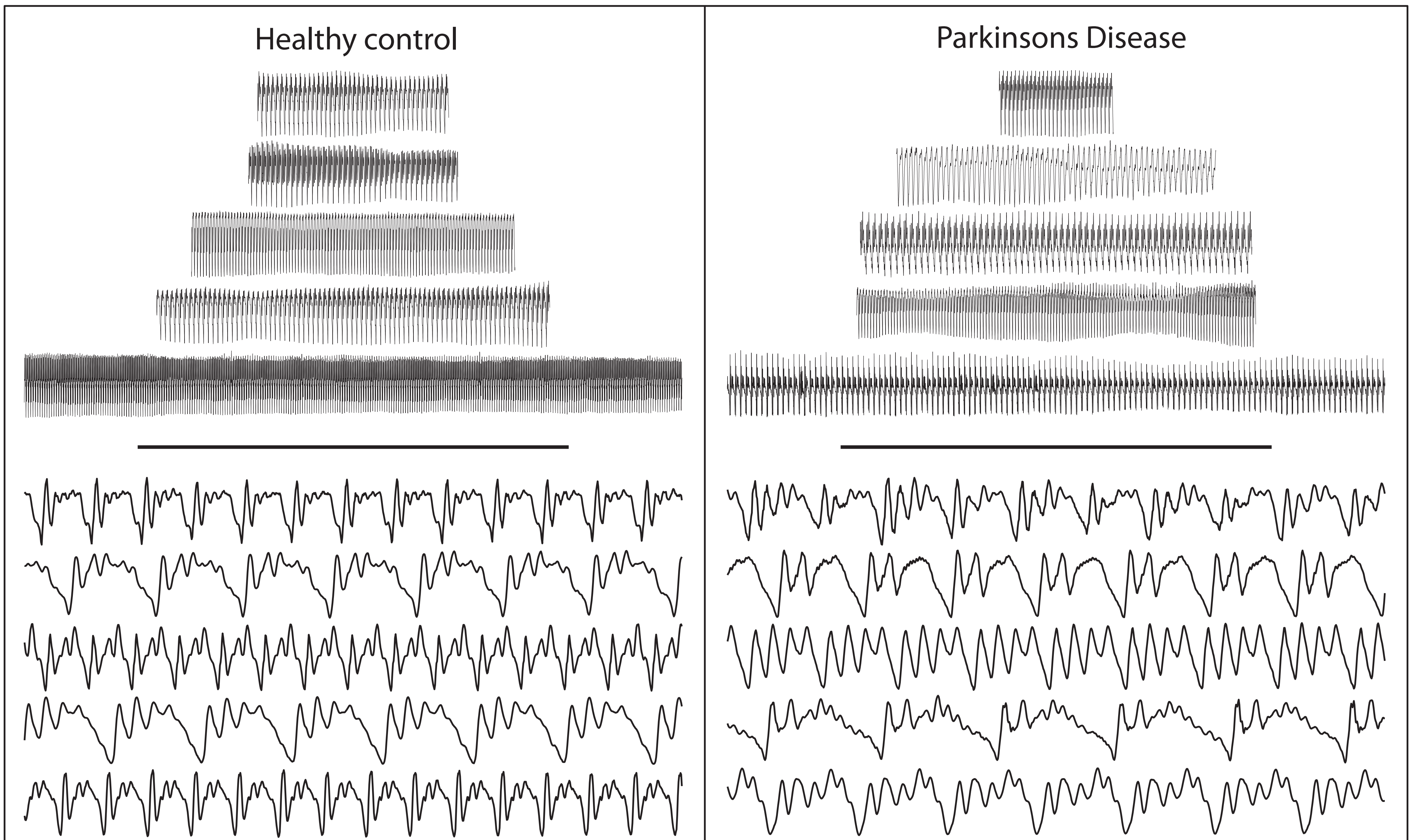




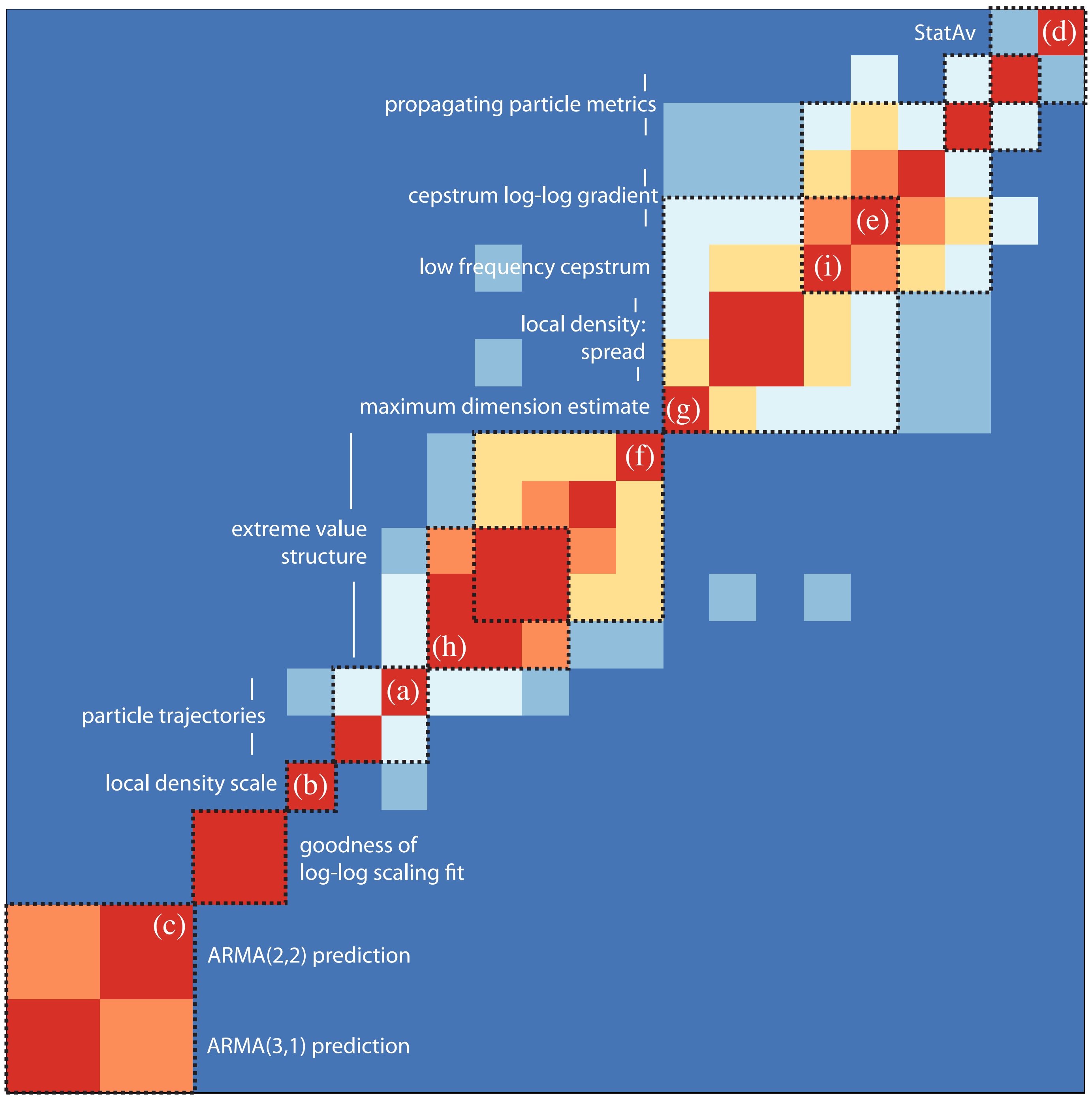
Heart Rate Variability

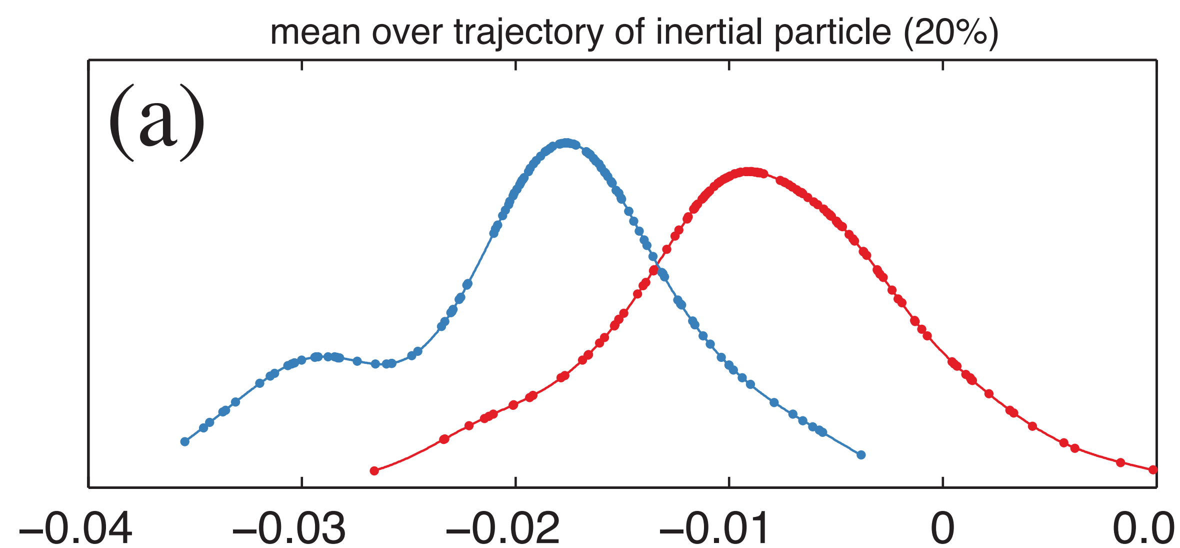


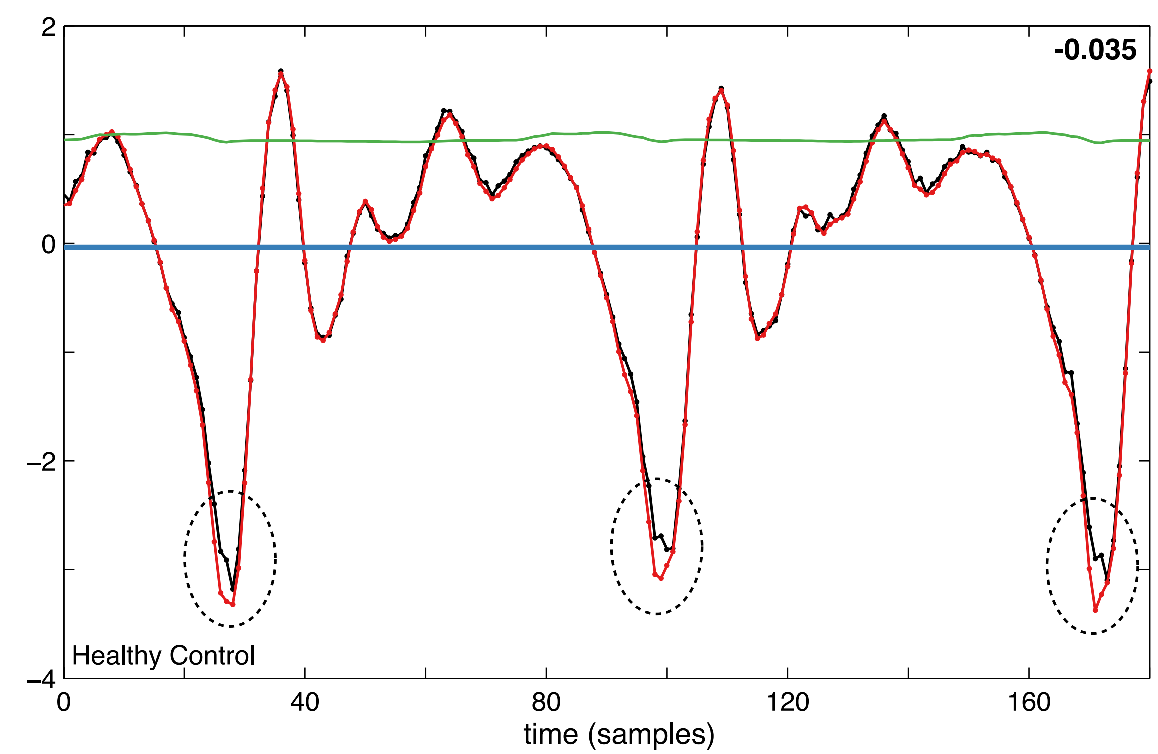
Parkinson's Disease Speech



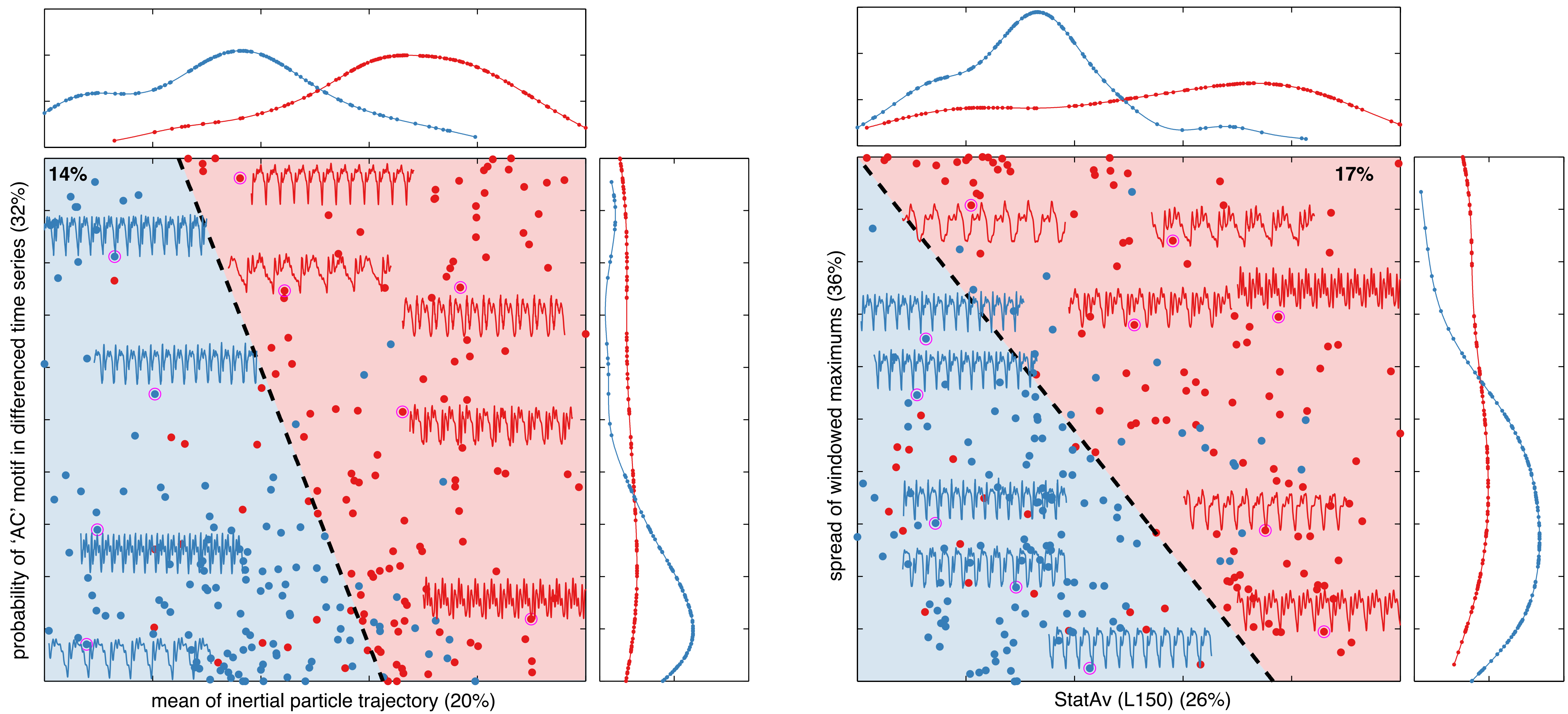
Parkinsonian Speech





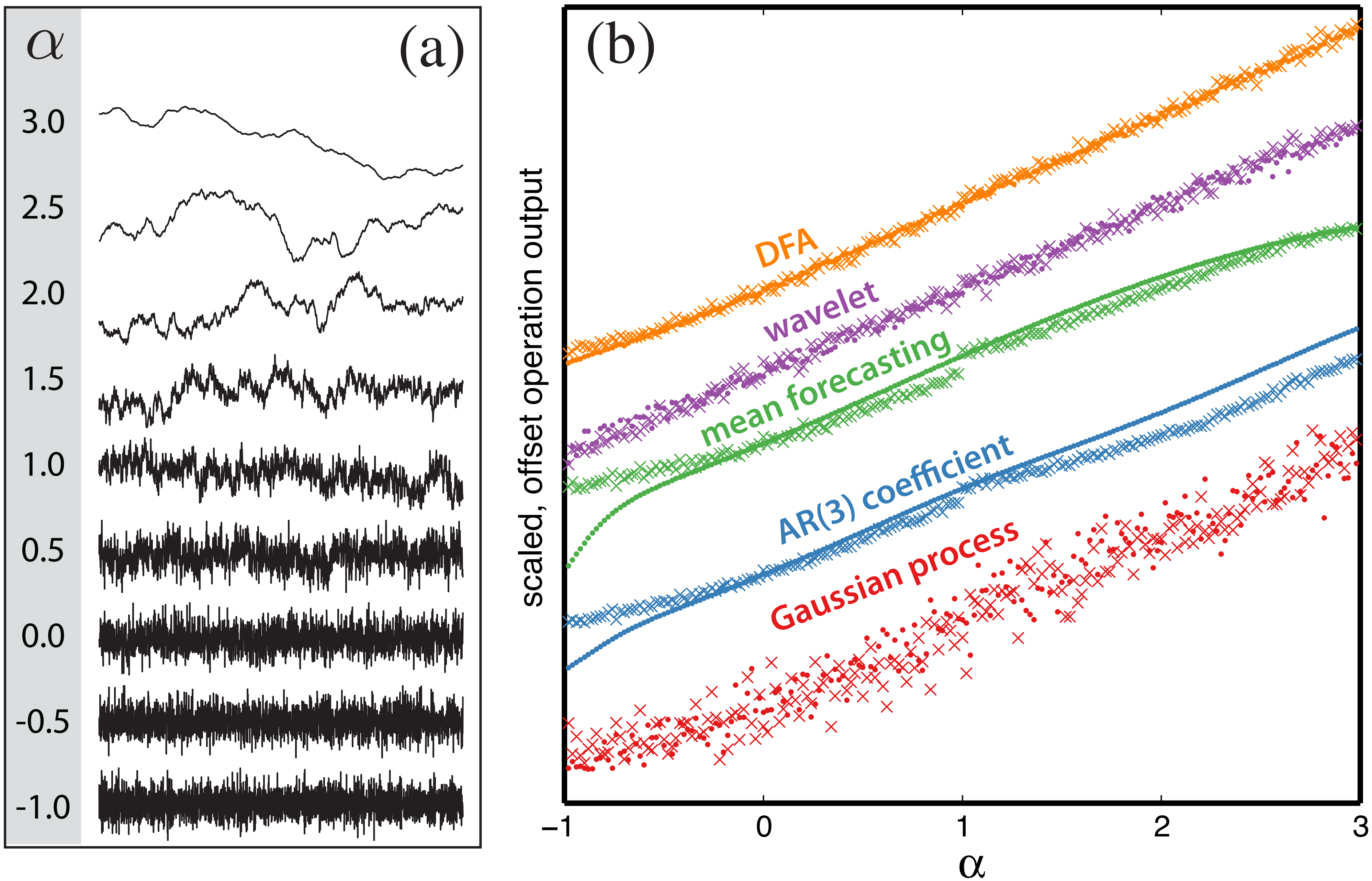


Parkinsonian Speech

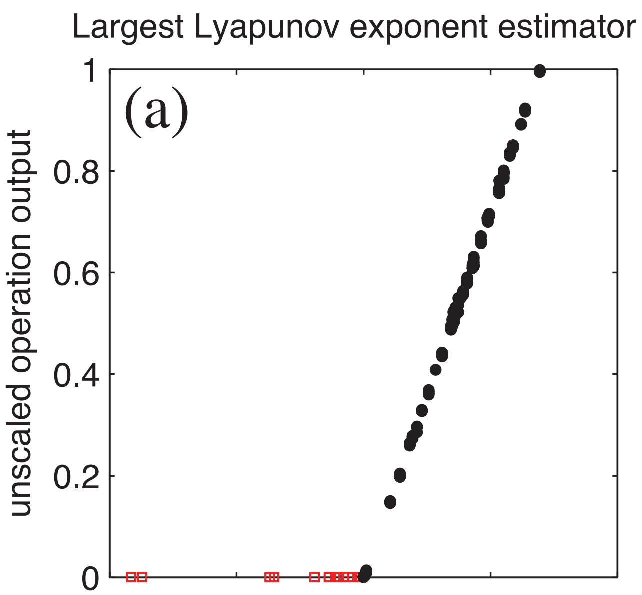


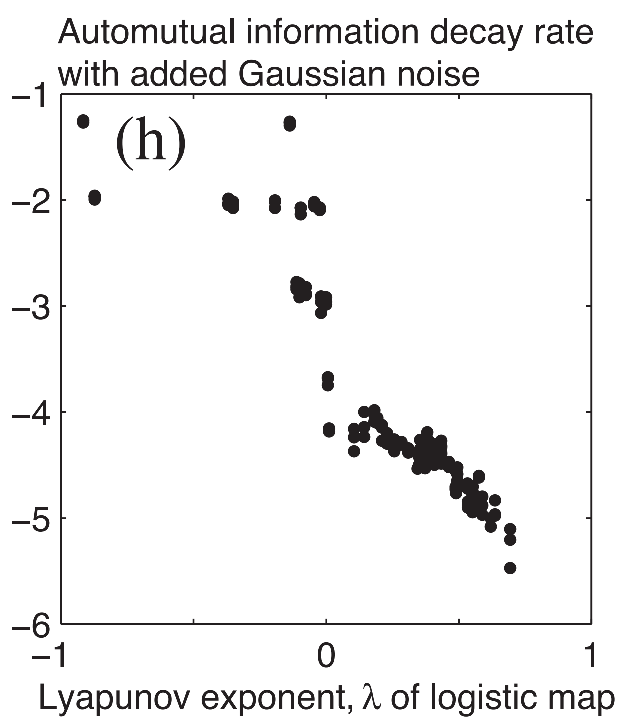
classifiers mix methods developed in different disciplines

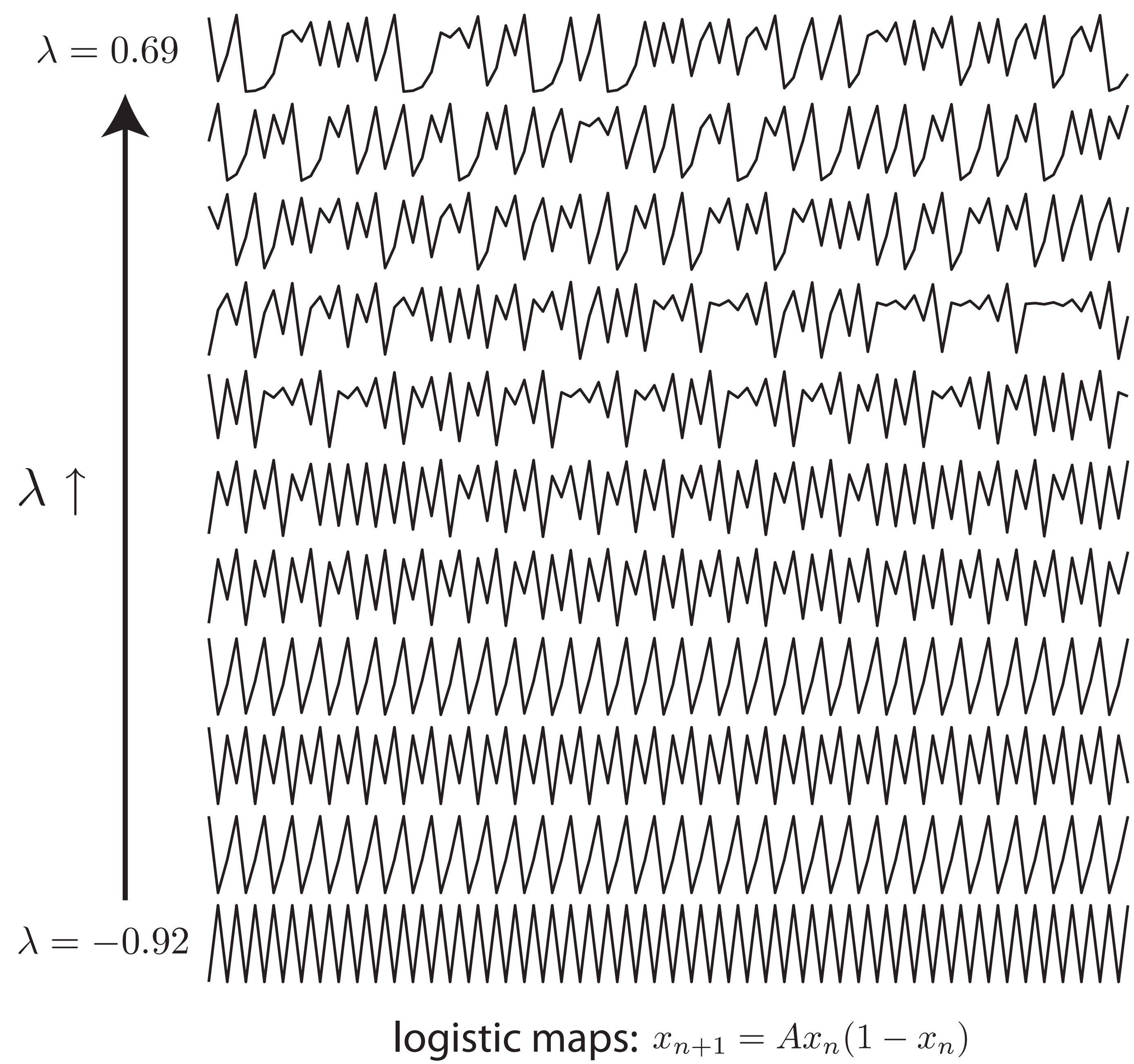
Self-Affine Time Series

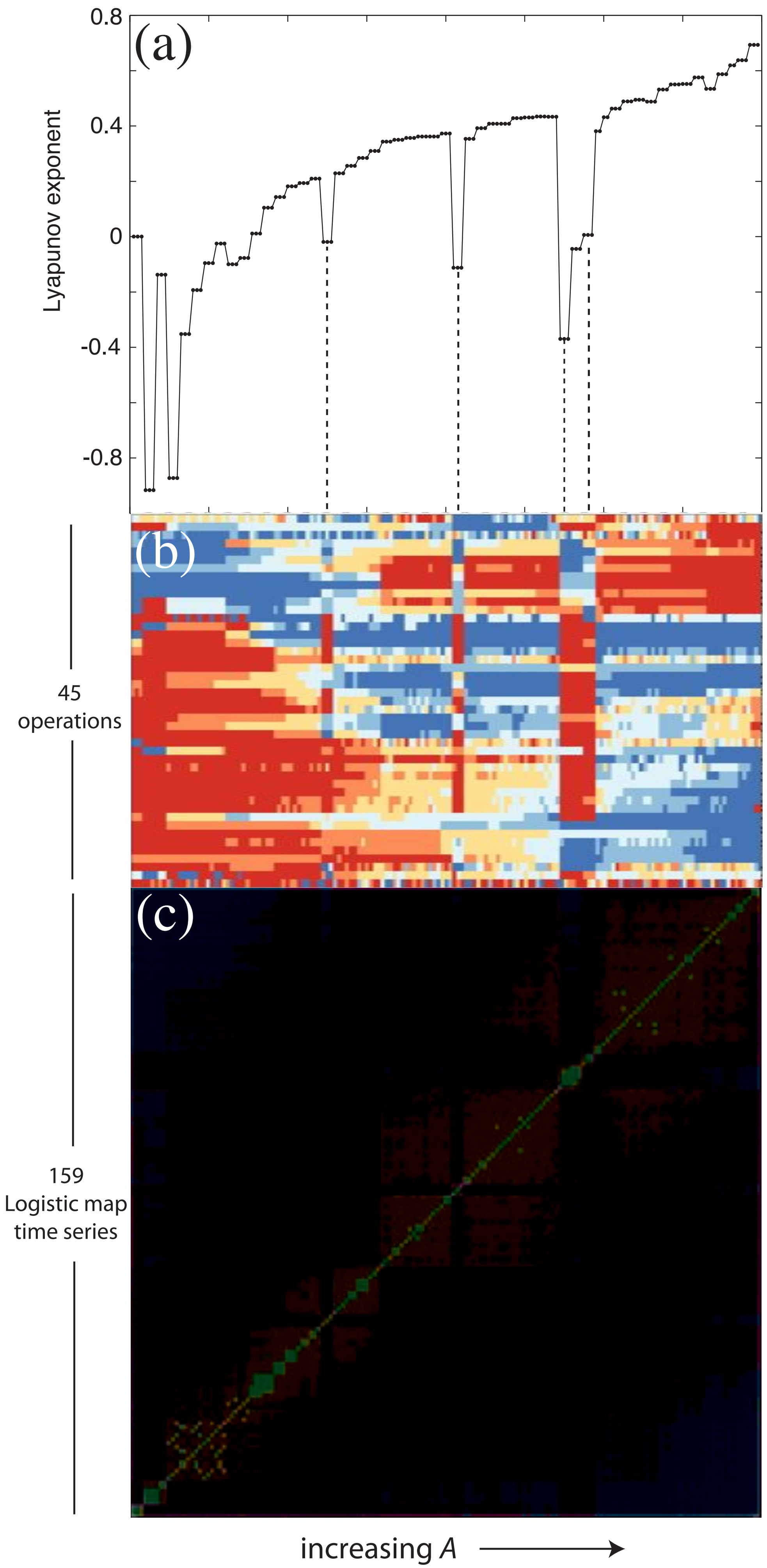


Logistic Map Regression







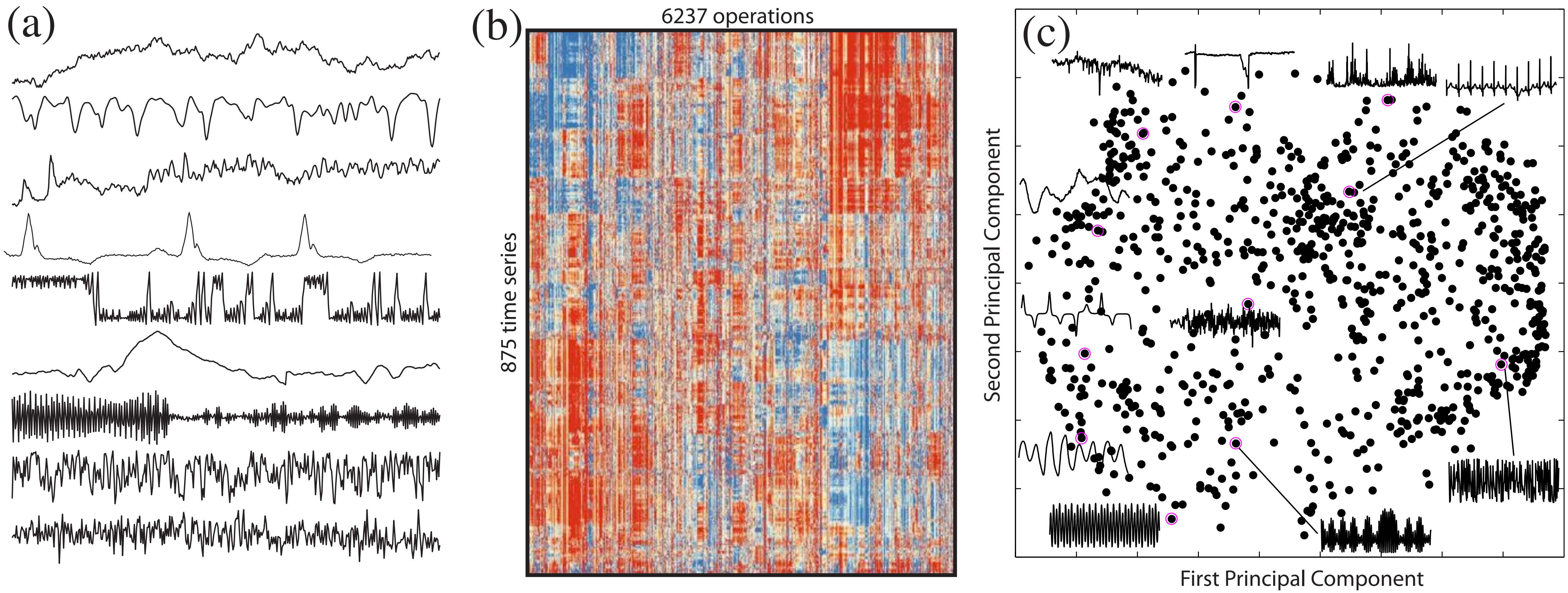


Logistic Map

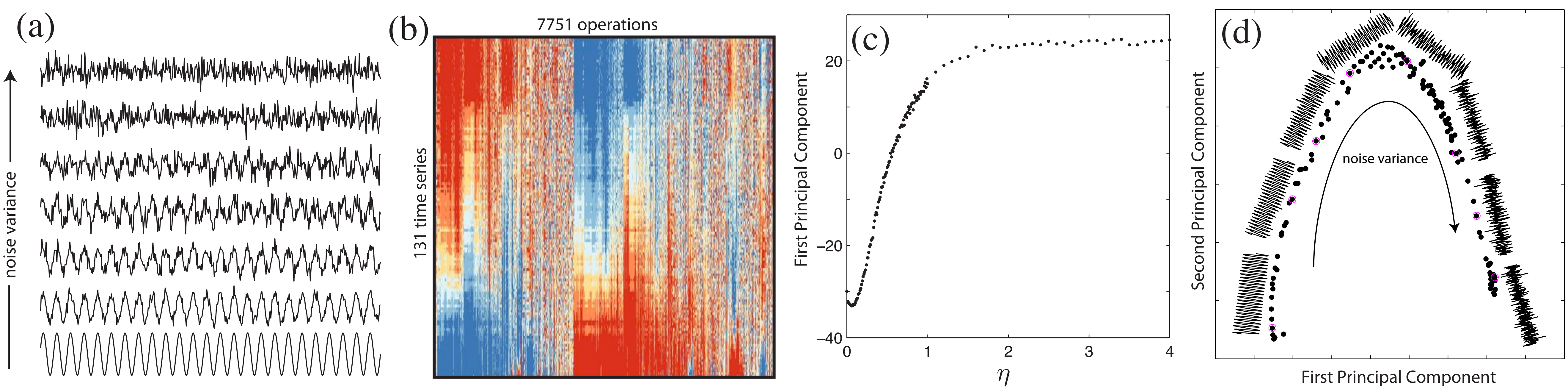
Constrained systems

- We’ve seen redundancy in set of methods for natural signals.
- What about systems that can be fully described by a small number of parameters?
- The structure of our database can hint at this.

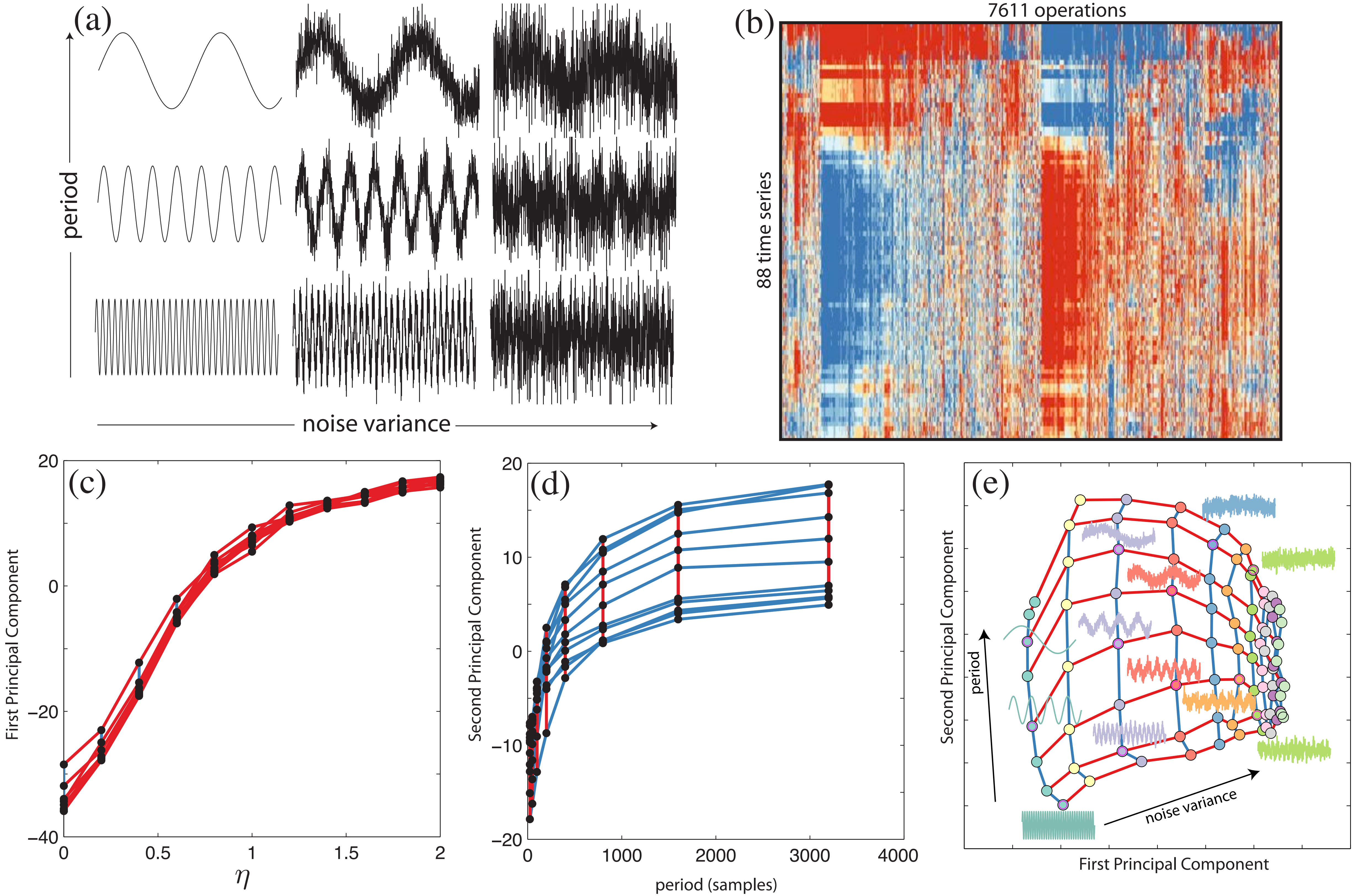
Many parameters



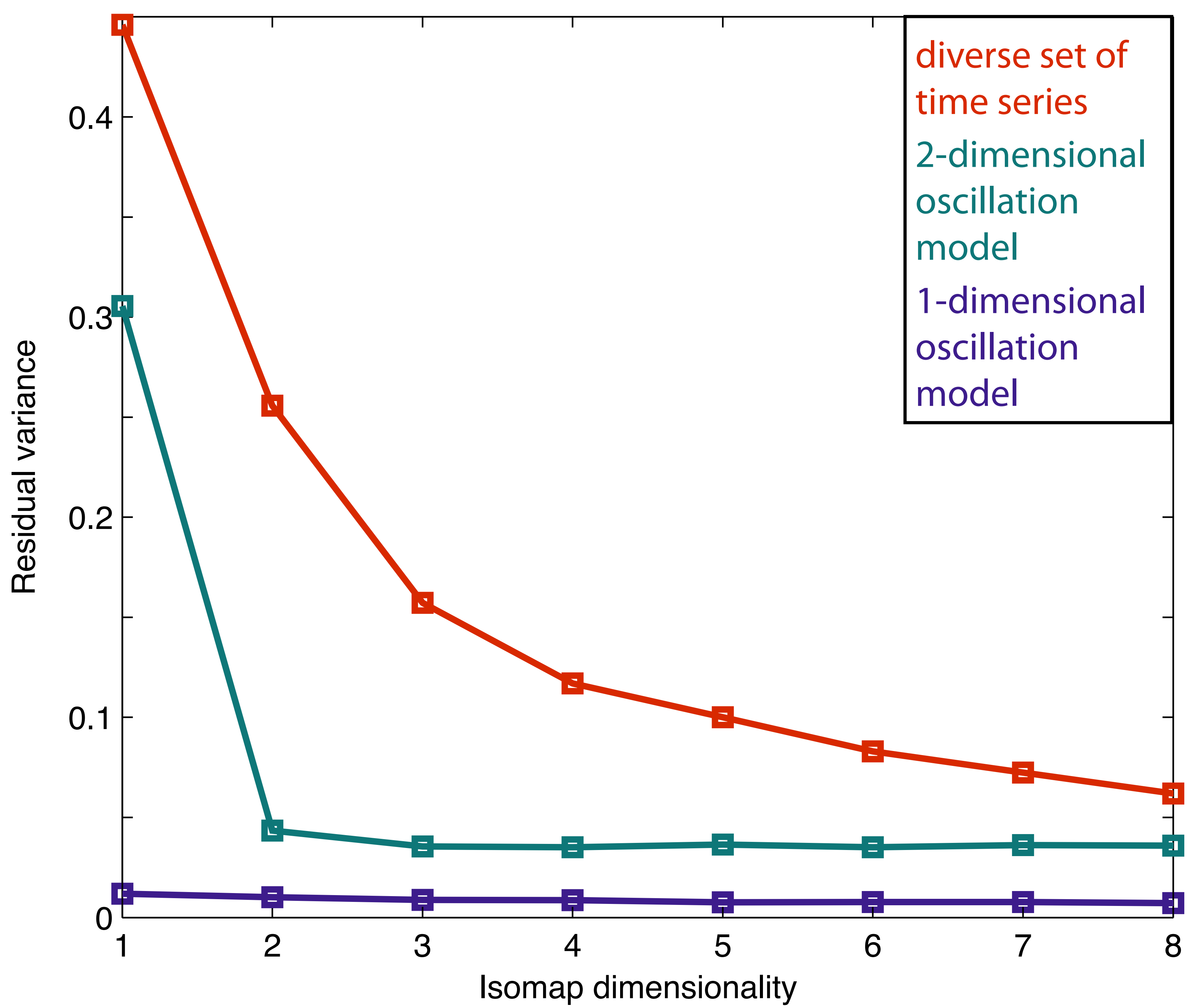
One parameter



Two parameters



Isomap can quantify this



Conclusions

- Empirical organization of the methods we use in science
- Empirical organization of the time series and models we study in science
- Automatic classification and regression with the ability to give insights into underlying dynamics