DATA MINING IN THE IOT ERA: PRACTICAL EXAMPLES AND A PEEK INTO FUTURE DEVELOPMENTS

Michele Rossi

rossi@dei.unipd.it

Dept. of Information Engineering

University of Padova, IT

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

Roadmap

- Environmental Monitoring
- Human Centric Sensing
- Rethinking IoT Protocol Designs

Environmental Monitoring

The Problem at Stake

- Spatio-temporal signal
- Data acquired from field sensors at regular intervals
  - Slotted time
- Study the tradeoff among
  - Signal compression / Transmission
  - Energy Consumption (sensors)
  - Reconstruction accuracy (data collector)

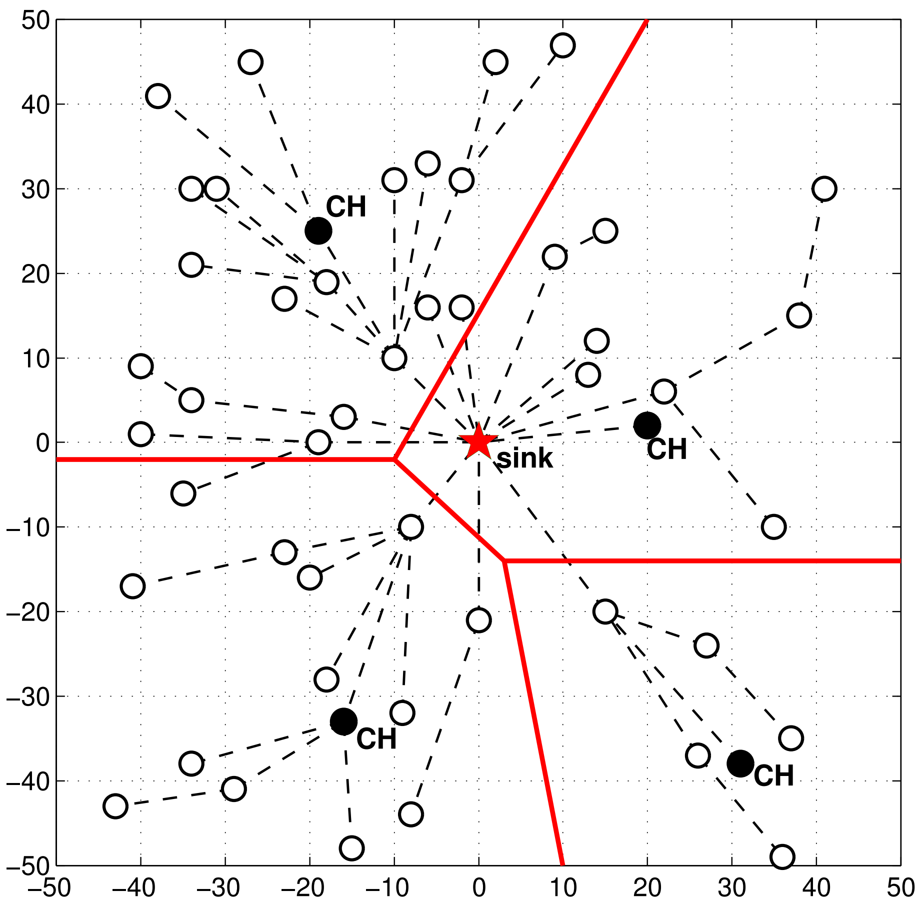
Signals

- Synthetic signals [Z++11]
- Temperature, humidity, solar radiation, wind direction, air temperature, precipitation, etc.
- LUCE and St-Bernard WSN testbeds (EPFL, Lausanne), CitySense WSN testbed (Hardvard Univ. & BBN tech.)
- Meteorological station records from [WML12]
- Our own testbed @DEI (~300 nodes, indoor)

[Z++11] Zordan, G. Quer, M. Zorzi, M. Rossi,"Modeling and Generation of Space-Time Correlated Signals for Sensor Network Fields," *Proc. of IEEE GLOBECOM*, 2011.
[WML12] C. J. Willmott, K. Matsuura, D. R. Legates, "Global Air Temperature and Precipitation: Regridded Monthly and Annual Climatologies," 2012, Center for Climatic Research, Dept. of Geography, Univ. of Delaware, US. [Online]

Network Model

- N nodes uniformly distributed within a square area
- Same TX range for all sensors - shortest path routing



Distributed Source Coding

- Uses clustering
- CH sends data uncompressed
- Other nodes TX compressed data

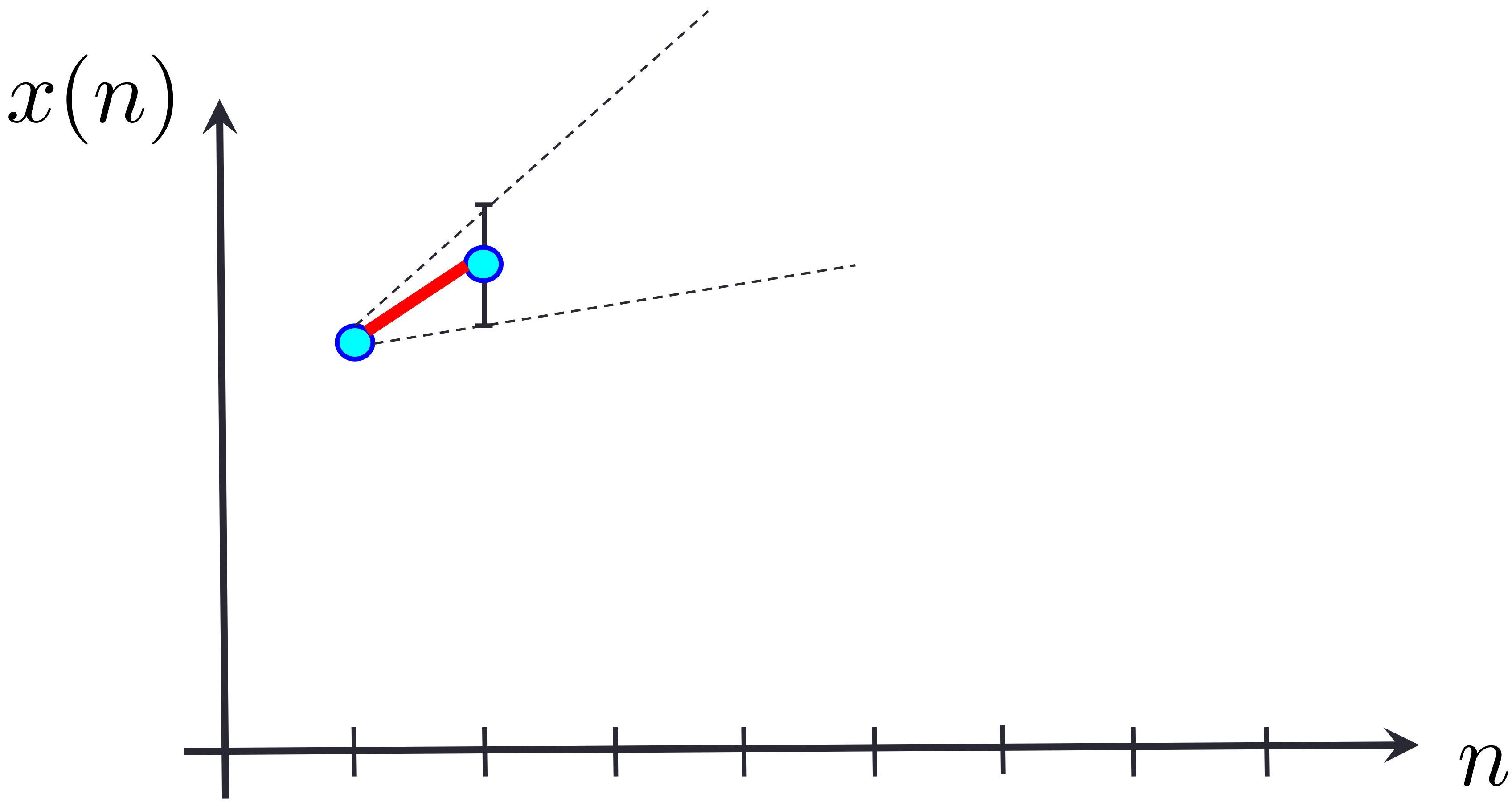
Compressive Sensing (CS)

- Data is sampled from a small number of nodes at each round

DCT & LTC

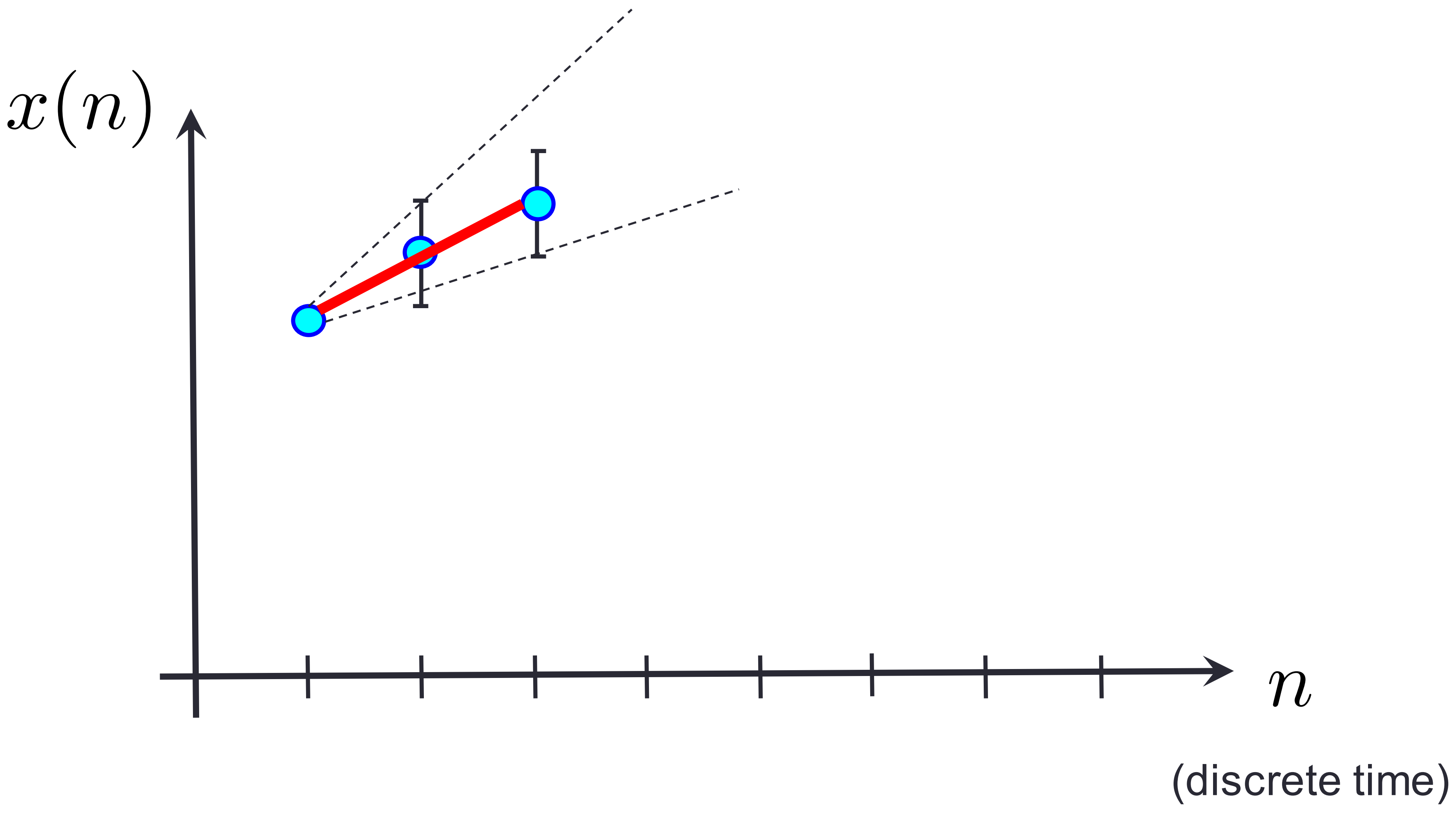
- All nodes independently compress
- They all TX compressed data

LTC (compressing time series)

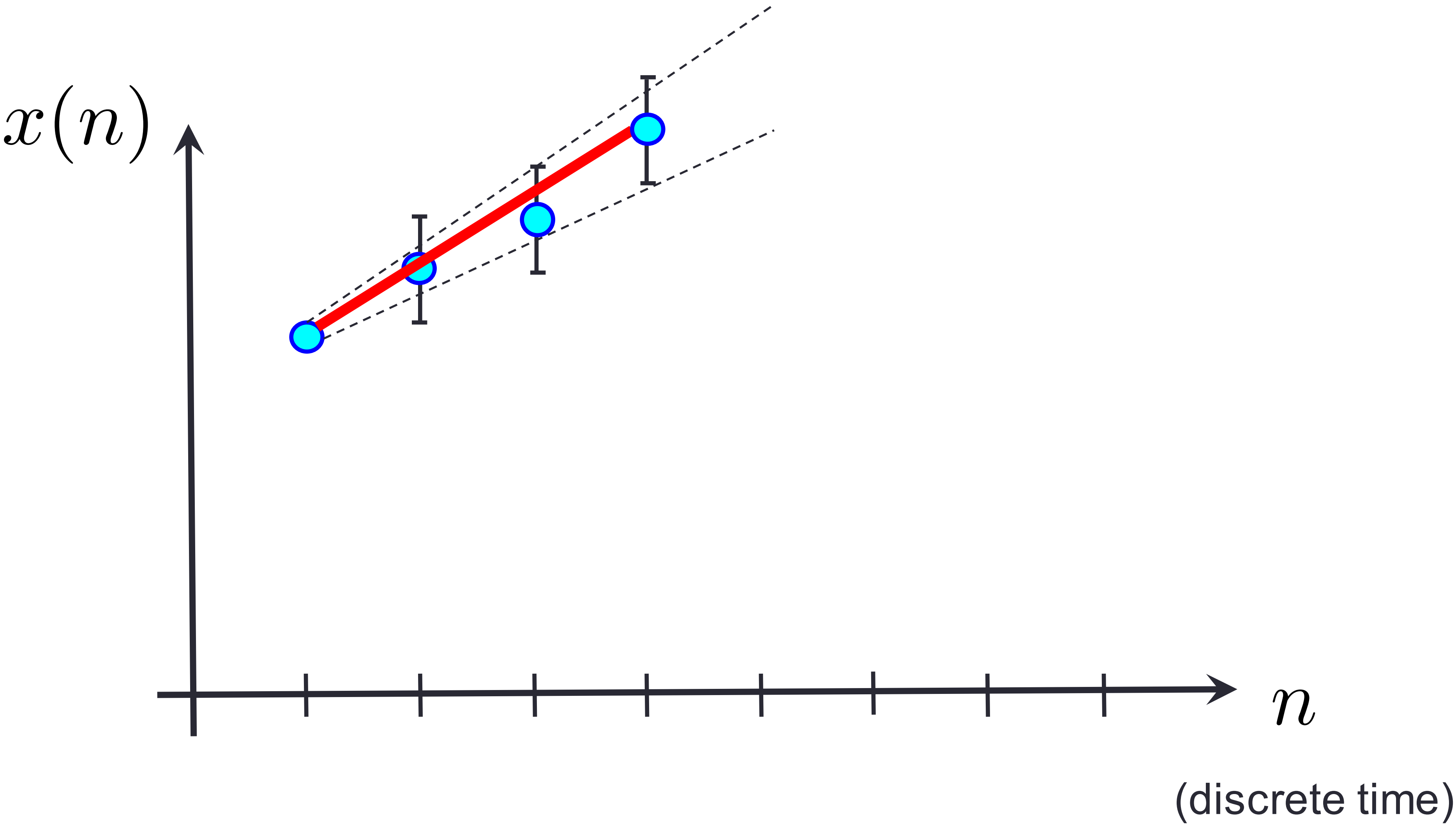


[S++04] T. Schoellhammer, B. Greenstein, M. Wimbrow E. Osterweil, and D. Estrin, “Lightweight temporal compres- sion of microclimate datasets,” in *Proceedings of the IEEE International Conference on Local Computer Net- works (LCN)*, Tampa, FL, US, Nov. 2004.

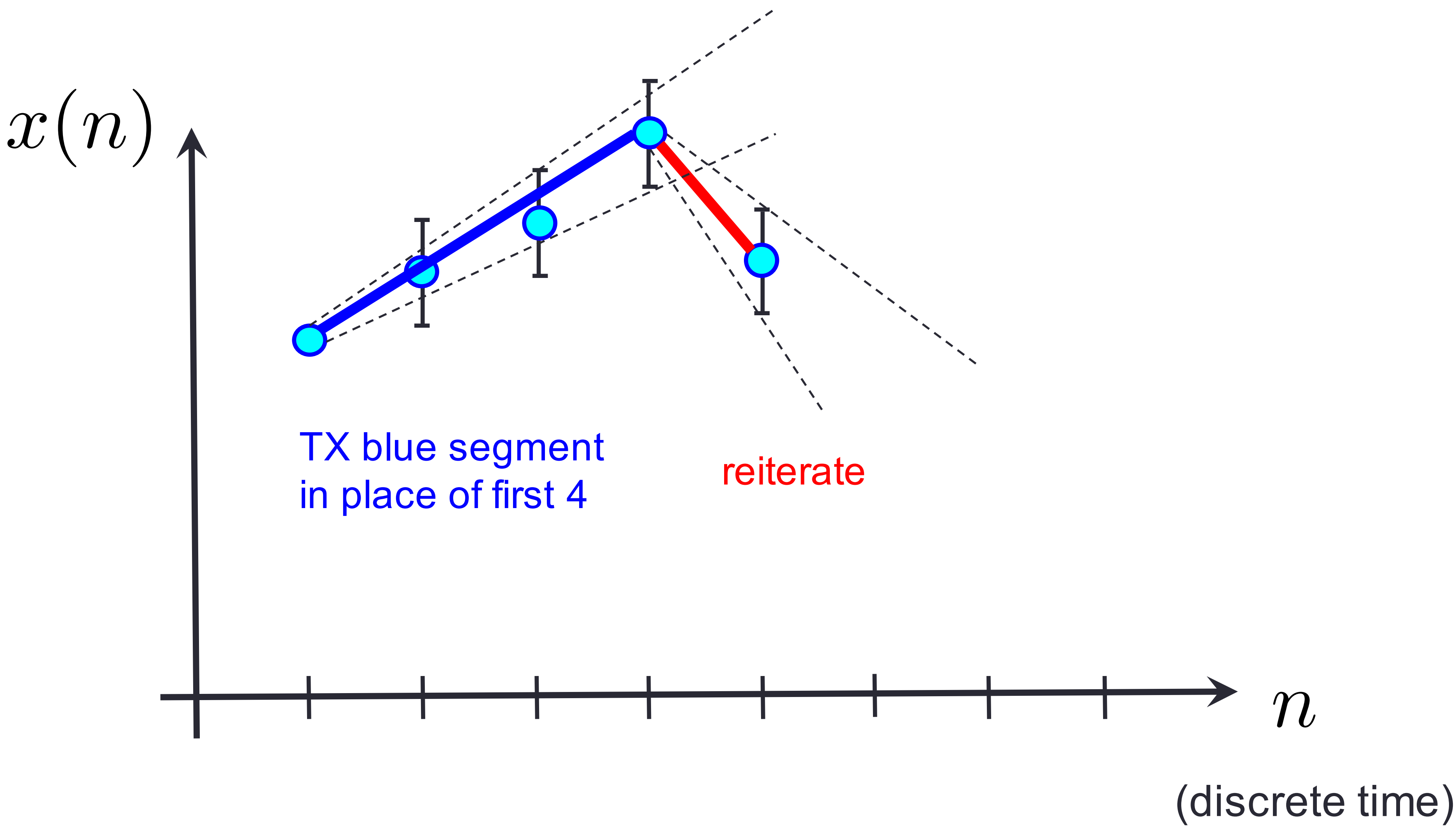
LTC



LTC

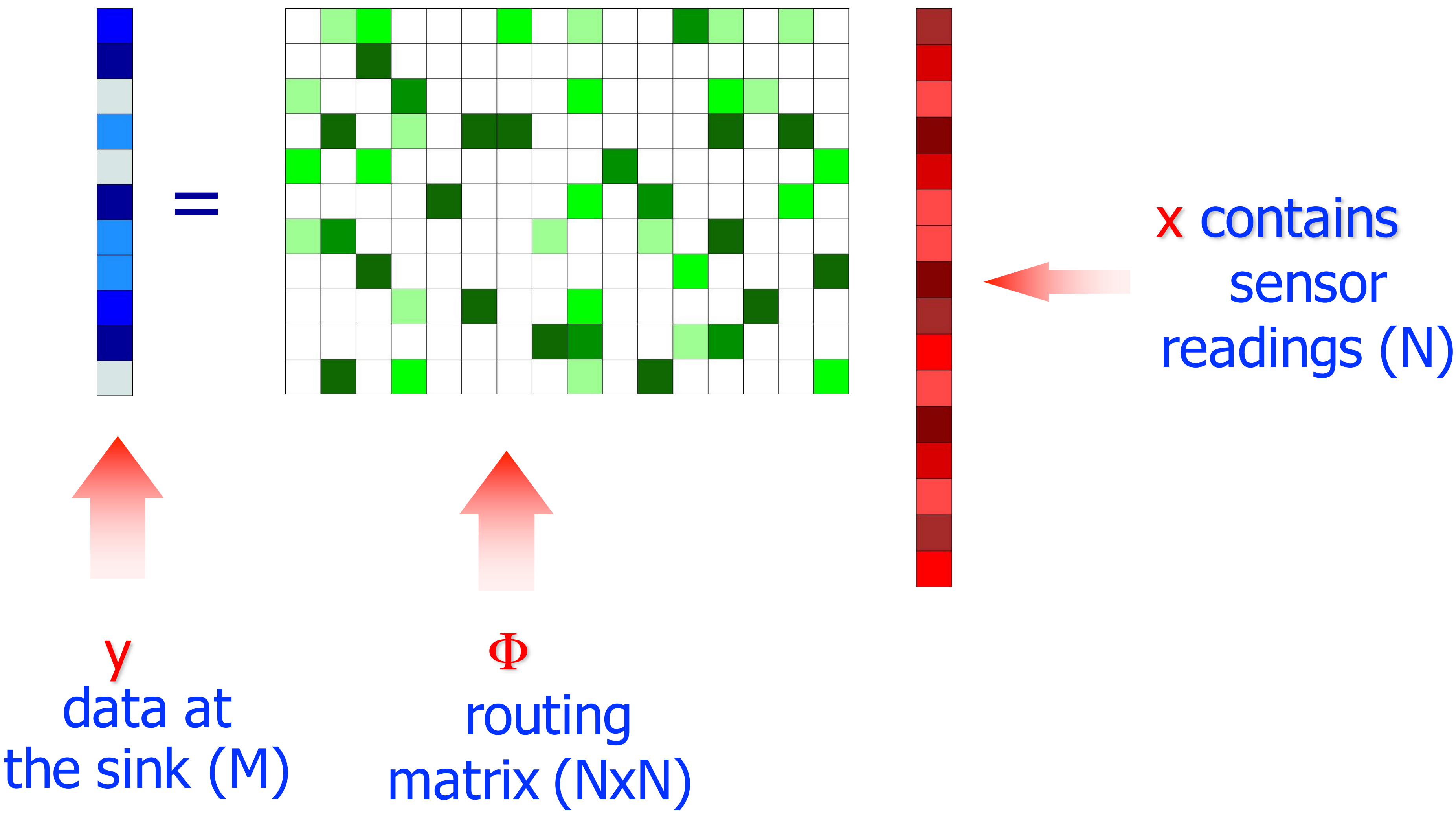


LTC

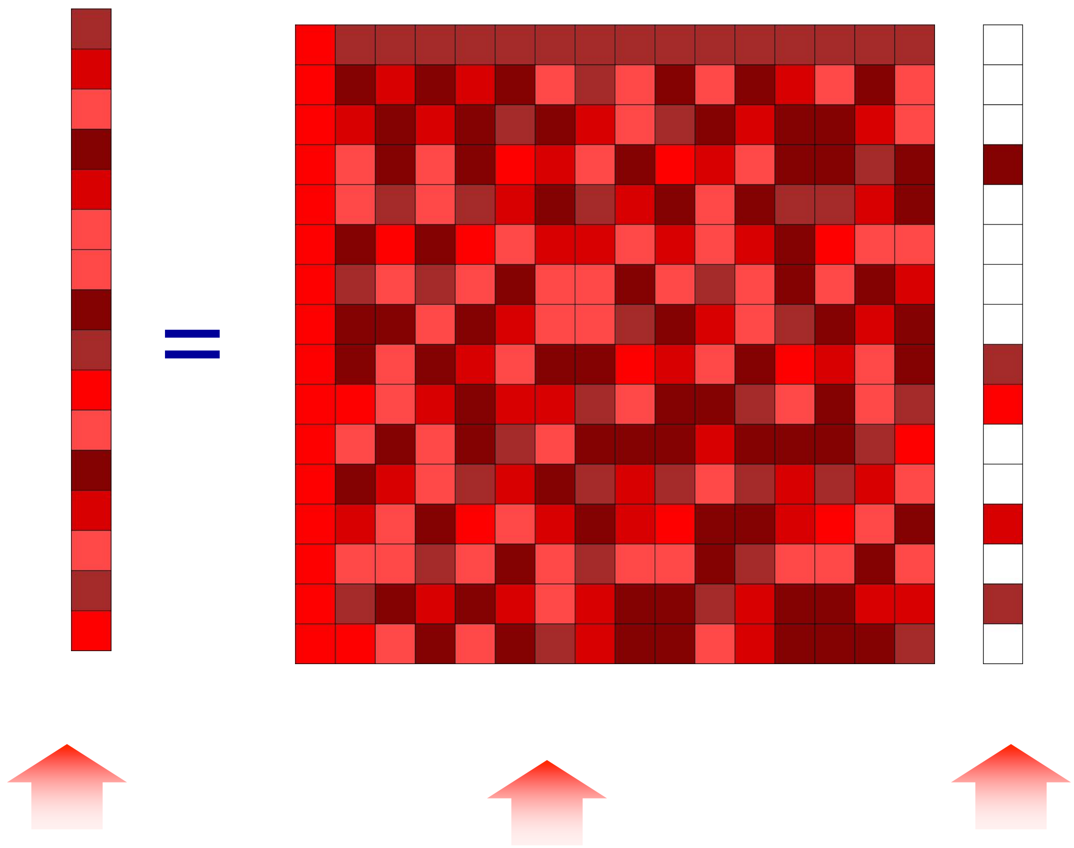


$x(n)$
TX blue segment
in place of first 4
reiterate
n
(discrete time)

Compressive Sensing



Sparse Signal



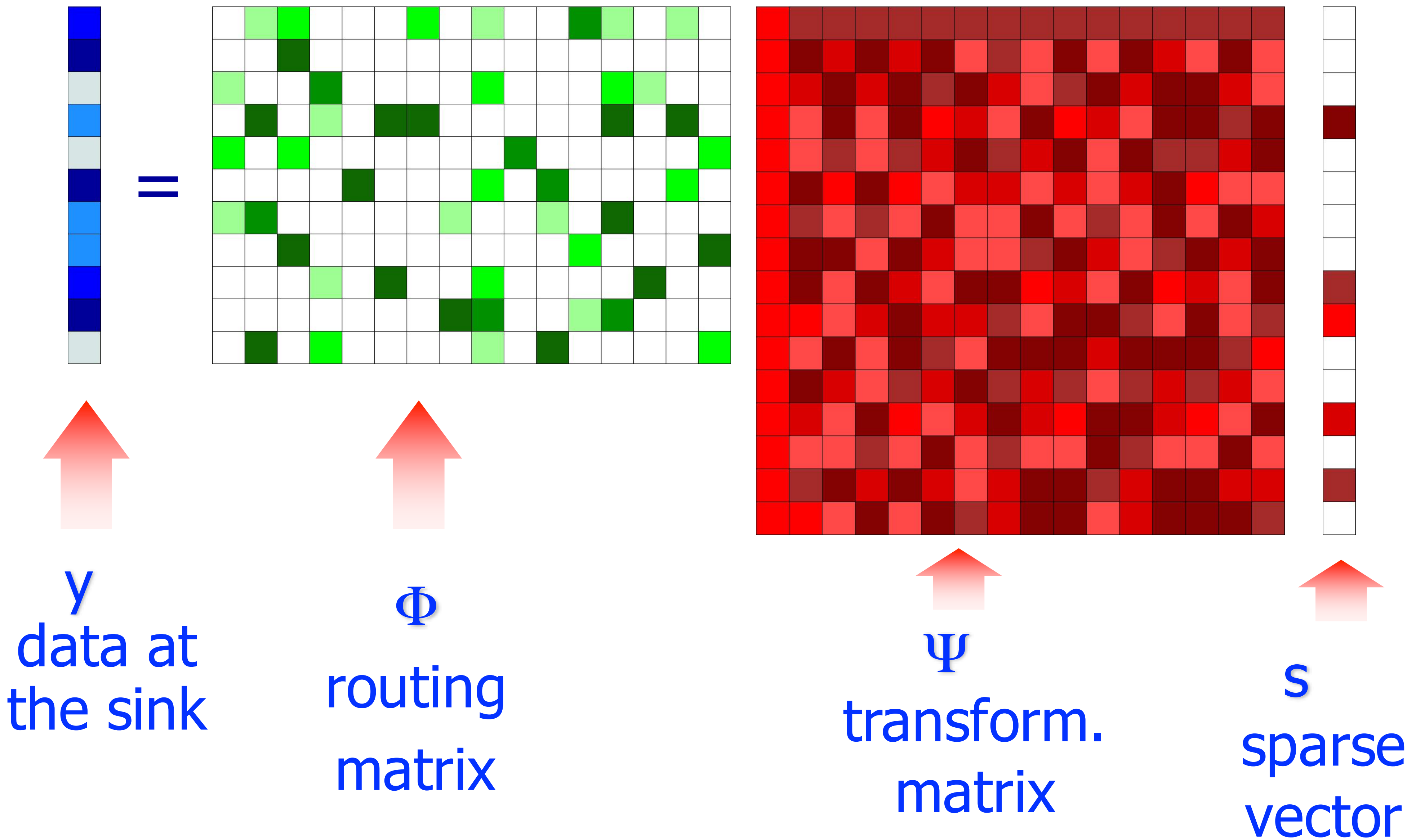
x
sensor readings

Ψ
transform. matrix

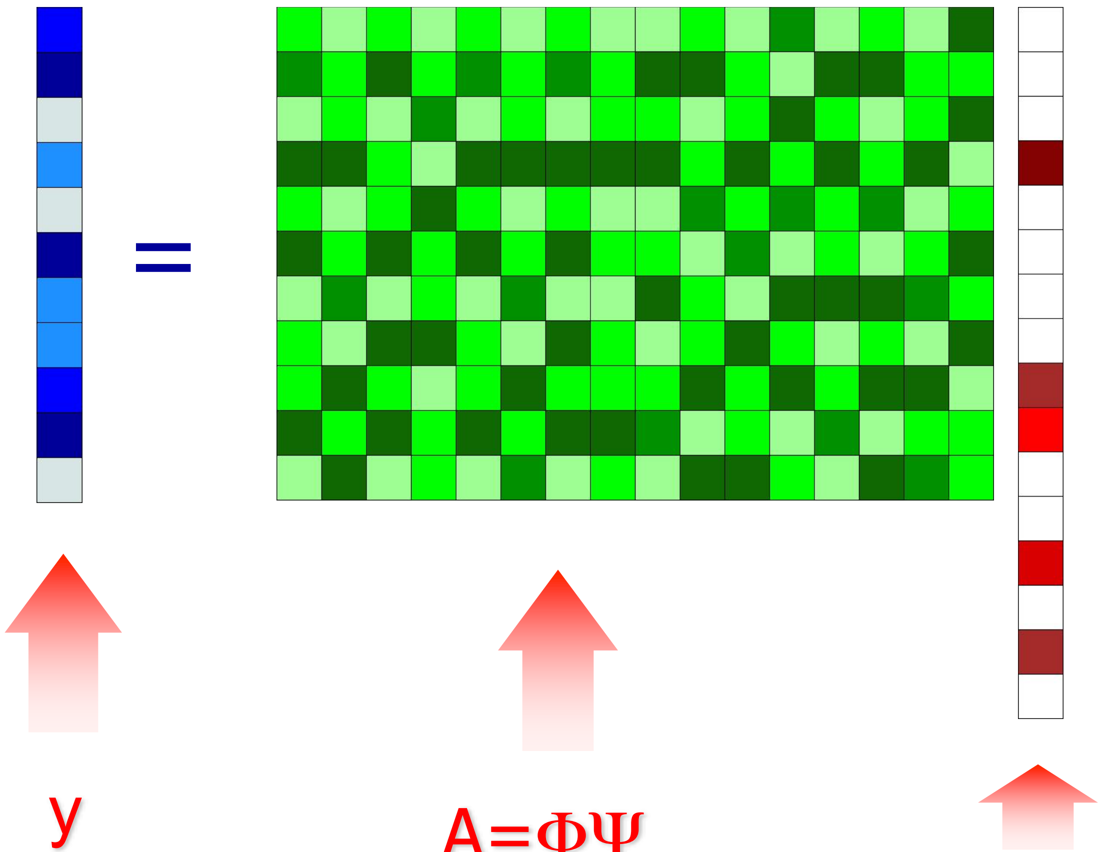
s
sparse vector

- Assumption
  - There exists a basis Ψ that sparsifies the signal
  (the sink must know Ψ)

Replacing “x” with “Ψs”



We finally get



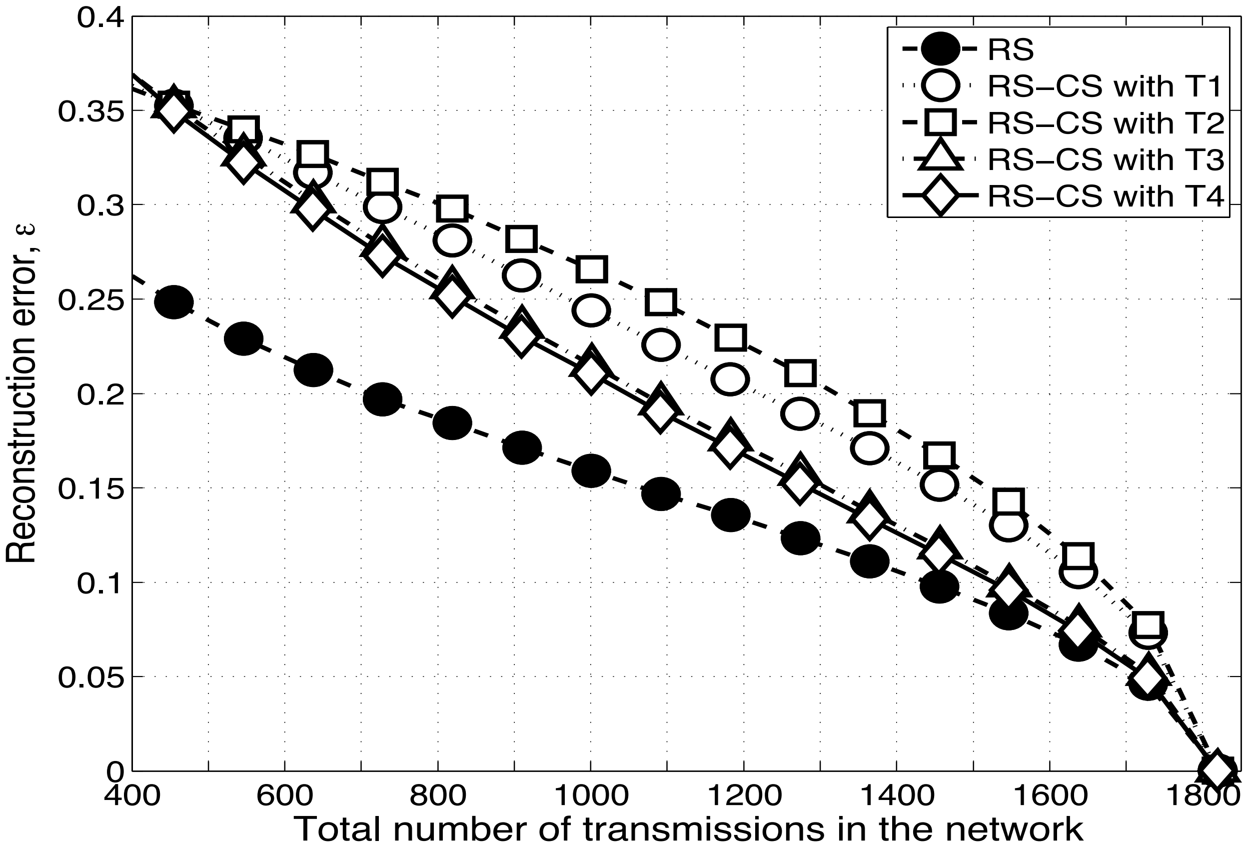
- Linear system,
  M equations,
  N>M unknowns
- N-M deg. of freedom
  (ill-posed problem)
- Norm-1 minimization
- Solvers: NESTA, l1-magic, Subspace Pursuit, …

y
data at the sink
(M values)

A=ΦΨ
MxN matrix
(M<N)

s
sparse vector

Which transform?



Transforms
T1: 2D DCT
T2: Haar Wavelet
T3, T4: our own

[Q++09] G. Quer, R. Masiero, D. Munaretto, M. Rossi, J. Widmer, M. Zorzi, "On the interplay between routing and signal representation for compressive sensing in Wireless Sensor Networks," *Inf. Theory and Applications Workshop (ITA), 2009*.

[S87] D. T. Sandwell, "Biharmonic Spline Interpolation of GEOS-3 and SEASAT Altimeter Data," *Geophysical Research Letters*, vol. 14, no. 2, 1987.

Principal Component Analysis

- Linear transform
  - Change of basis
- Dimensionality reduction
  - Only the M (<N) “strongest” principal components are retained
- Decorrelates the data set
  - Removes second order dependencies
- Columns of transform matrix Ψ are the eigenvectors of C_x
- C_x *covariance matrix* of random process x (N-sized)

Estimate C_x at runtime (sample cov.)
Use PCA as CS transform

PCA – large variance reveals structure

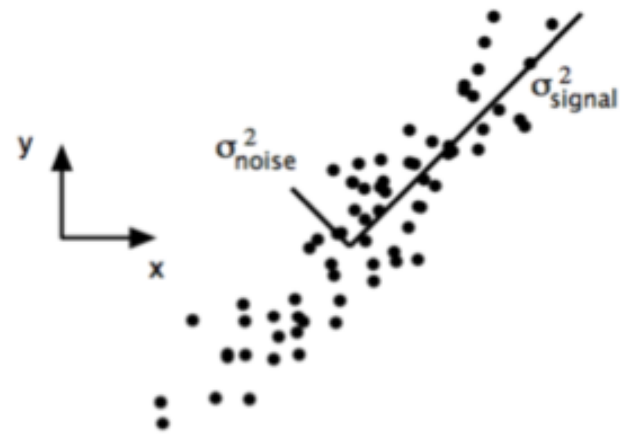


FIG. 2 Simulated data of (x, y) for camera A. The signal and noise variances σ^2_{signal} and σ^2_{noise} are graphically represented by the two lines subtending the cloud of data. Note that the largest direction of variance does not lie along the basis of the recording (x_A, y_A) but rather along the best-fit line.

$$SNR = \frac{\sigma^2_{\text{signal}}}{\sigma^2_{\text{noise}}}$$

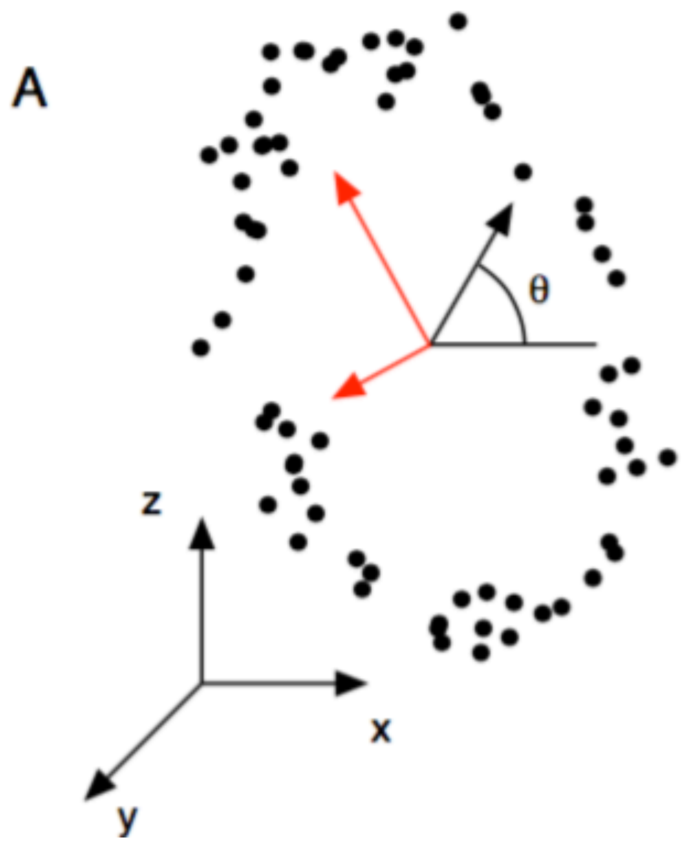
Basic assumption: the direction with highest variance contains the dynamics of interest (this corresponds to having SNR>>1) → direction with highest variance provides the best (linear) fit for the signal

PCA – variance *vs* correlation

FIG. 3 A spectrum of possible redundancies in data from the two separate measurements r_1 and r_2. The two measurements on the left are uncorrelated because one can not predict one from the other. Conversely, the two measurements on the right are highly correlated indicating highly redundant measurements.

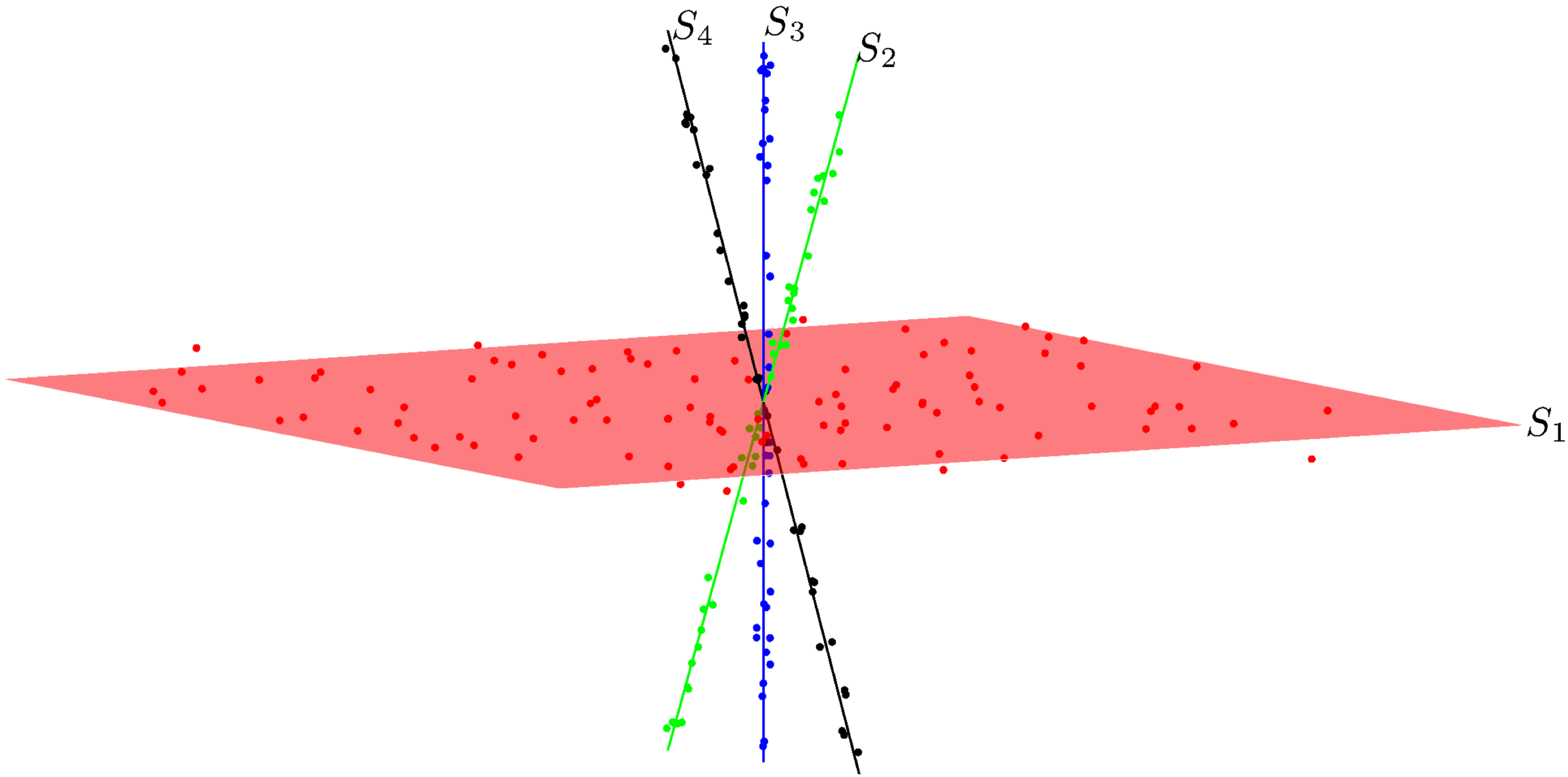
Dimensionality reduction: approximate the signal from 2 to a single variable projecting along the axis with the maximum variance

Where PCA fails



Maximum variance: in these examples, projecting the original data along the directions with maximum variance does not describe a good description of the underlying data structure

In general…



Interesting technique (pointers)

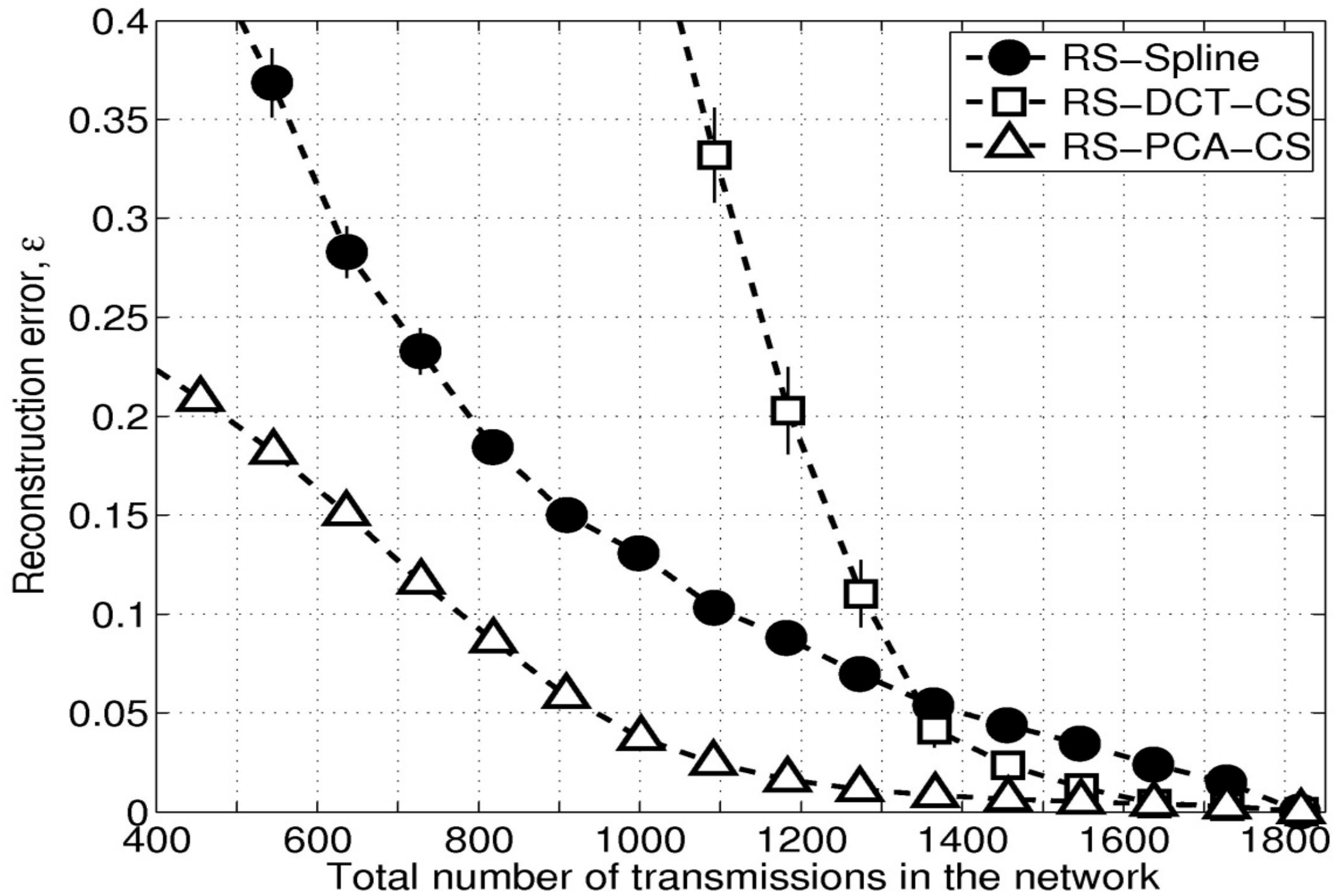
Subspace clustering is the problem of finding a multi-subspace representation that best fits a collection of points taken from a high-dimensional space.

[Chu10] Yi-Hong Chu, Jen-Wei Huang, Kun-Ta Chuang, De-Nian Yang, Ming-Syan Chen, “Density Conscious Subspace Clustering for High-Dimensional Data”, IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 1, 2010.

[Elhamifar13] Eshan Elhamifar, Ren Vidal, “Sparse Subspace Clustering: Algorithm, Theory and Applications”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35, No. 11, 2013.

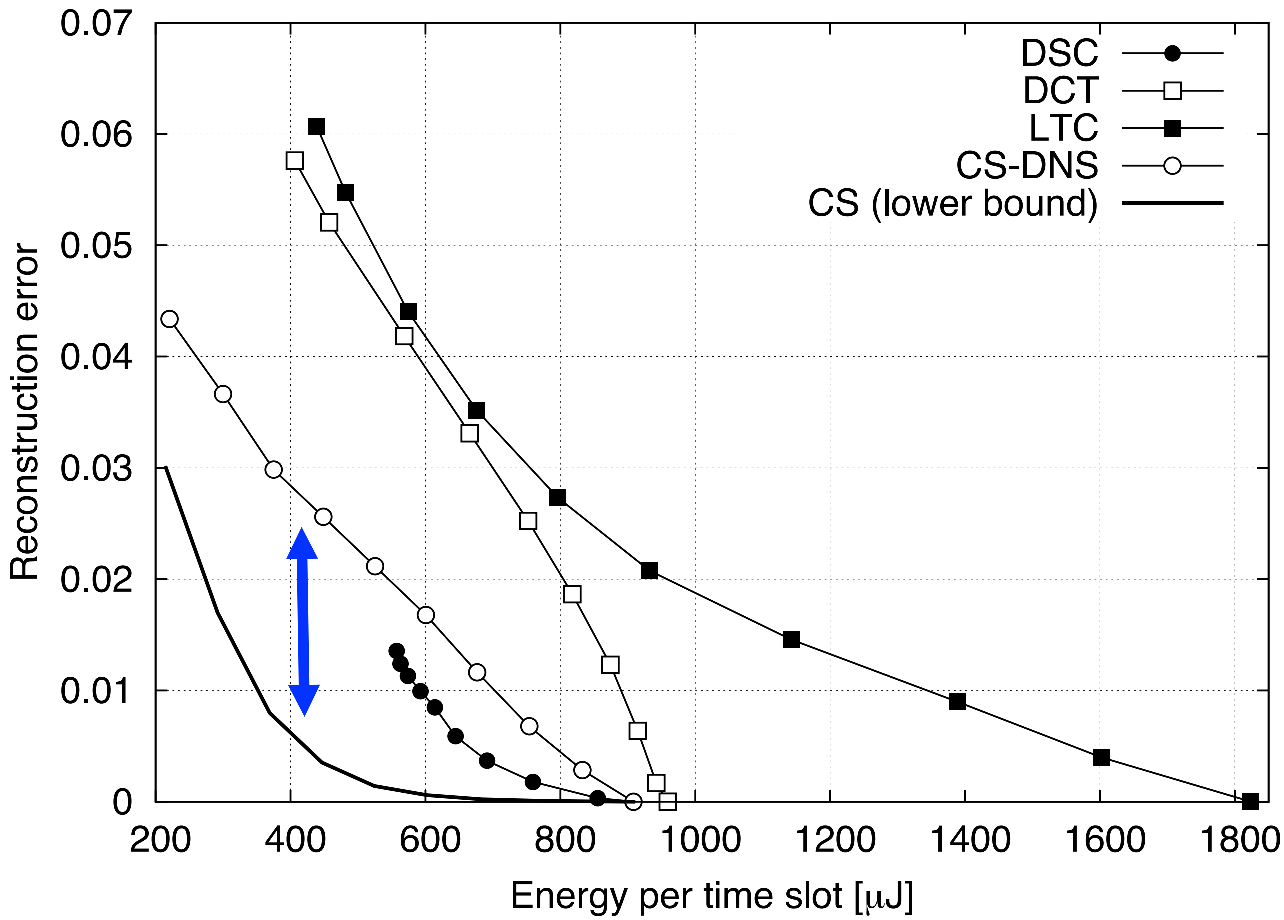
[Hu15] Han Hu, Jianjiang Feng, Jie Zhou, “Exploiting Unsupervised and Supervised Constraints for Subspace Clustering”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 37, No. 10, 2015.

Principal Component Analysis



[Q++12] G. Quer, R. Masiero, G. Pillonetto, M. Rossi, M. Zorzi, "Sensing, Compression and Recovery for WSNs: Sparse Signal Modeling and Monitoring Framework," *IEEE Transactions on Wireless Communications*, Vol. 11, No. 10, 2012.

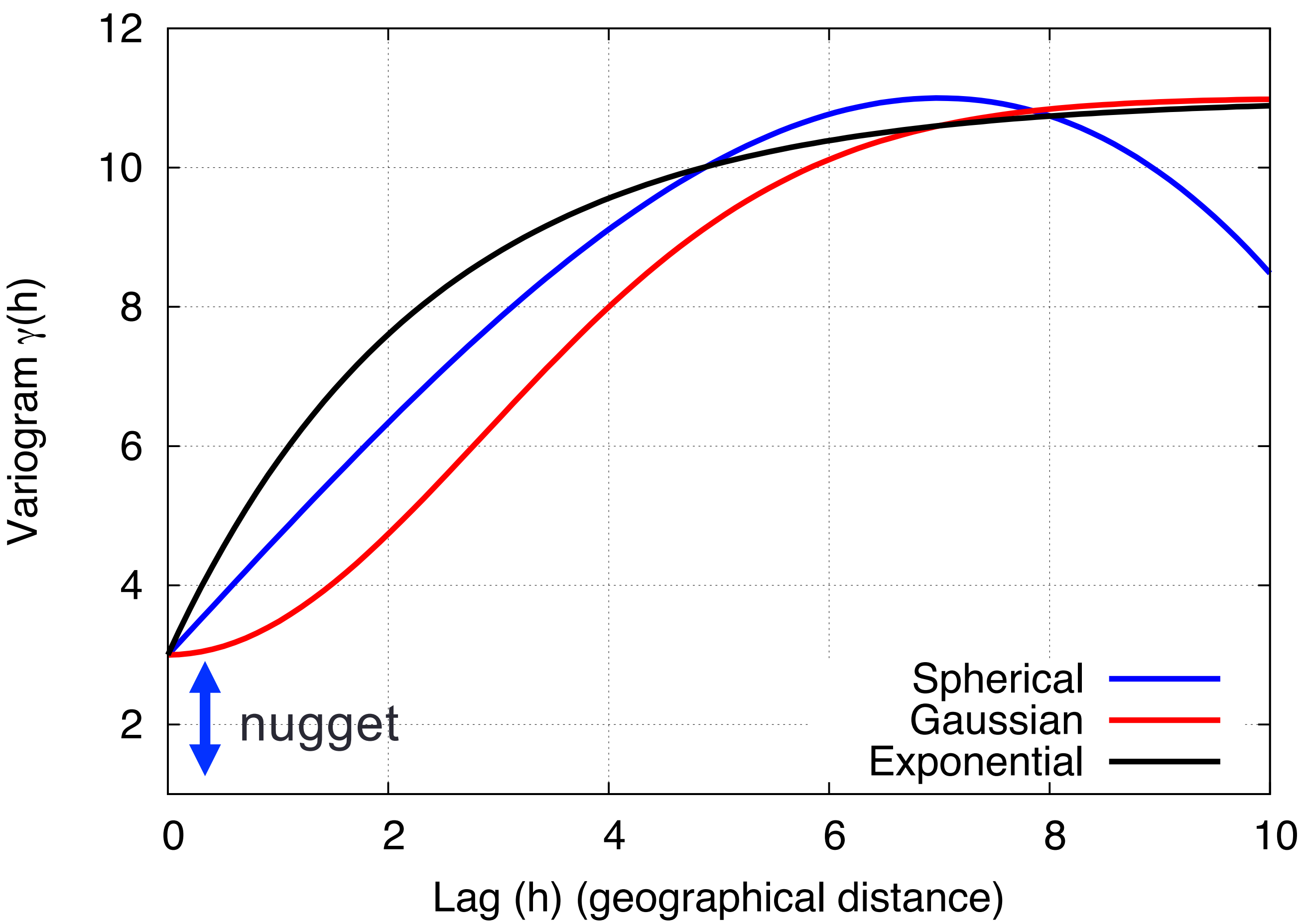
Medium to high spatial correlation



[R++15] M. Rossi, M. Hooshmand, D. Zordan, M. Zorzi, “Evaluating the gap between compressive sensing and distributed source coding in WSN,” *IEEE ICNC*, 2015.

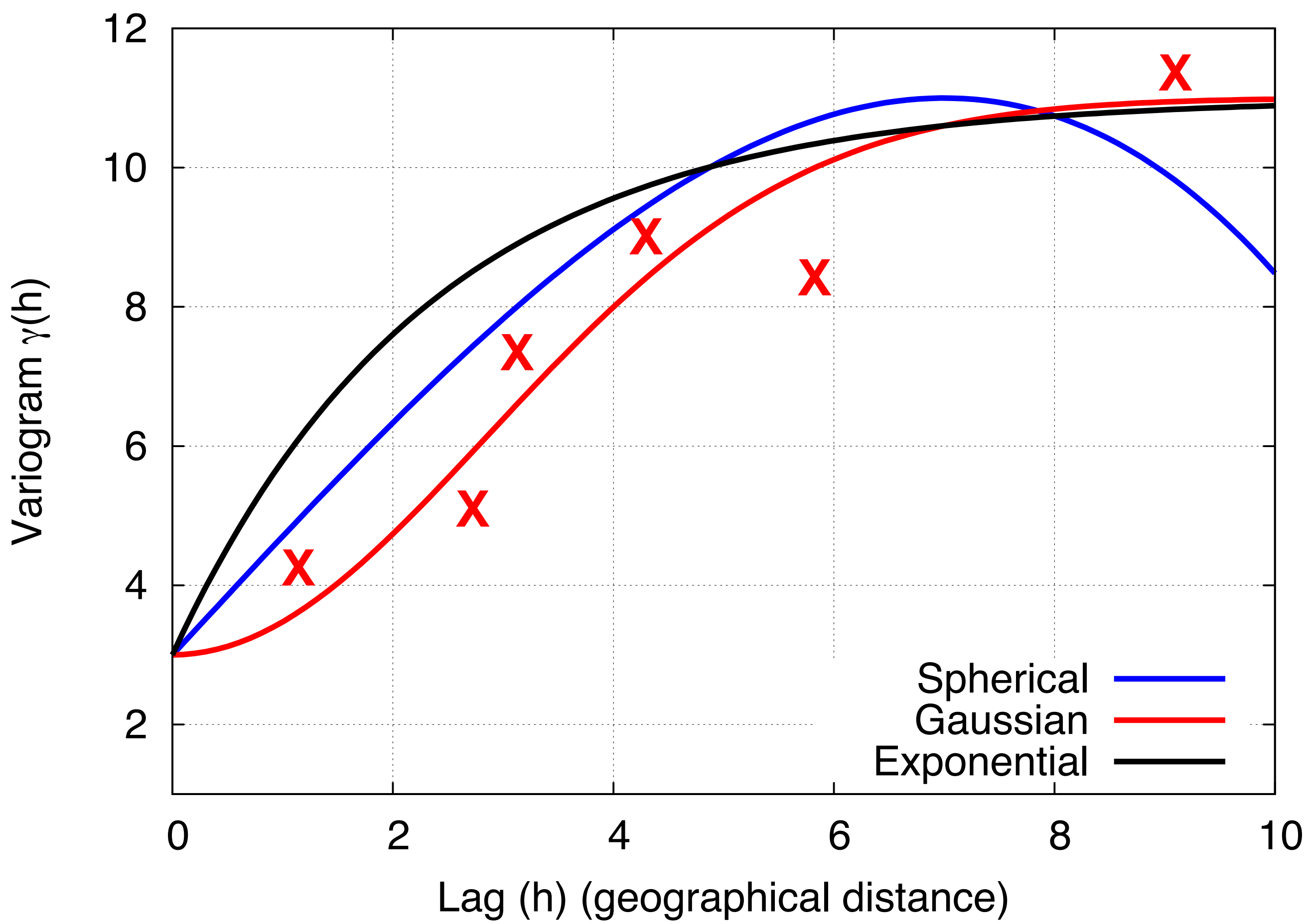
Covariogram theory (geostatistics)

Empirical Variograms (Evs) $\gamma(h) = \frac{1}{2N(h)} \sum_{\|\boldsymbol{p}-\boldsymbol{p}'\|_2=h} (x(\boldsymbol{p}) - x(\boldsymbol{p})')^2$



widely used covariogram functions (three pars)

Covariogram theory (geostatistics)

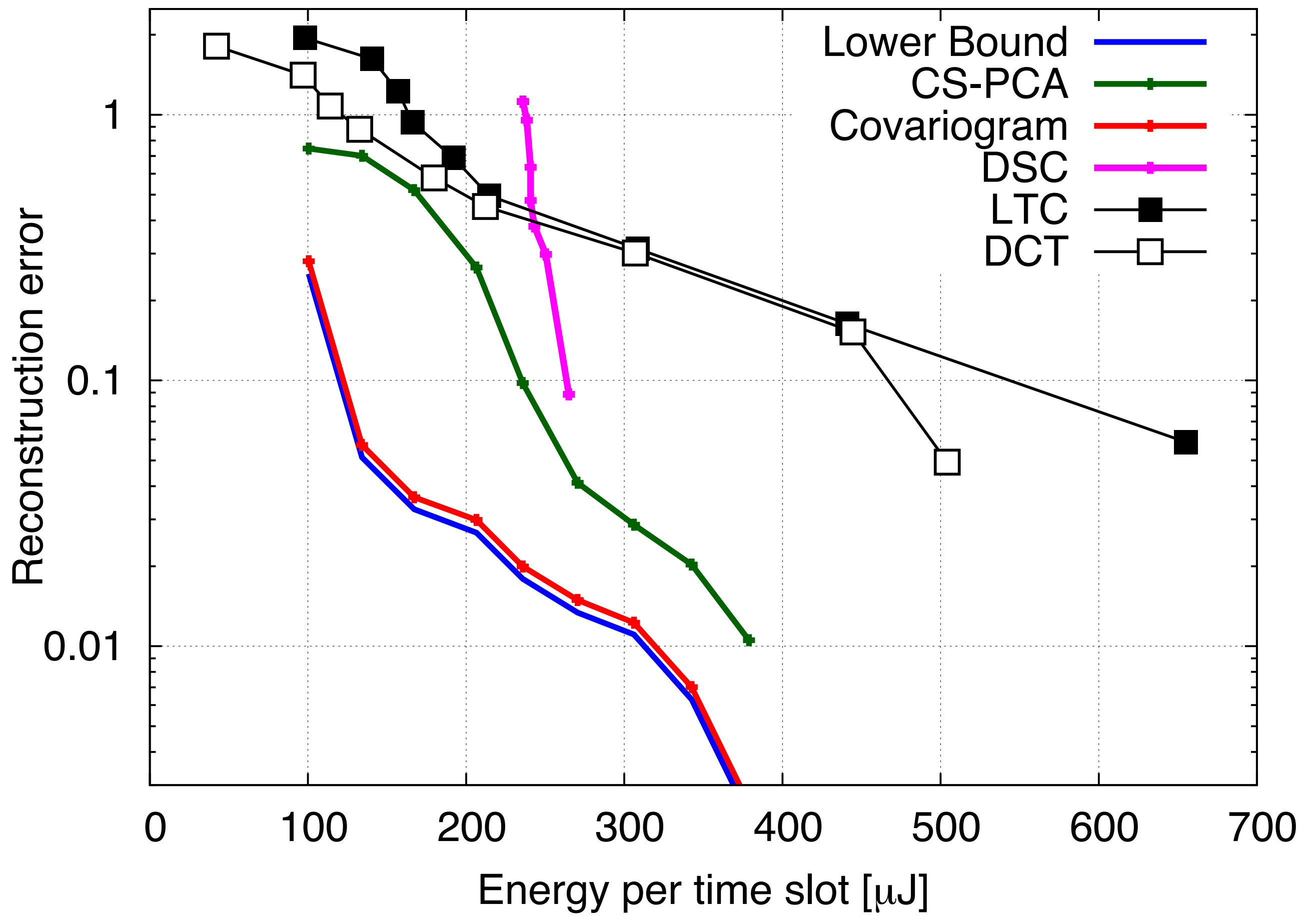


1) Measure EVs between node pairs (same slot TX)
2) h is the geographical distance of nodes pairs
3) Fit EVs with best model
4) Use model to obtain correlation matrix

$$\sigma_{ij} = n + s - \gamma(h_{i,j})$$

5) Use PCA to obtain transform Ψ

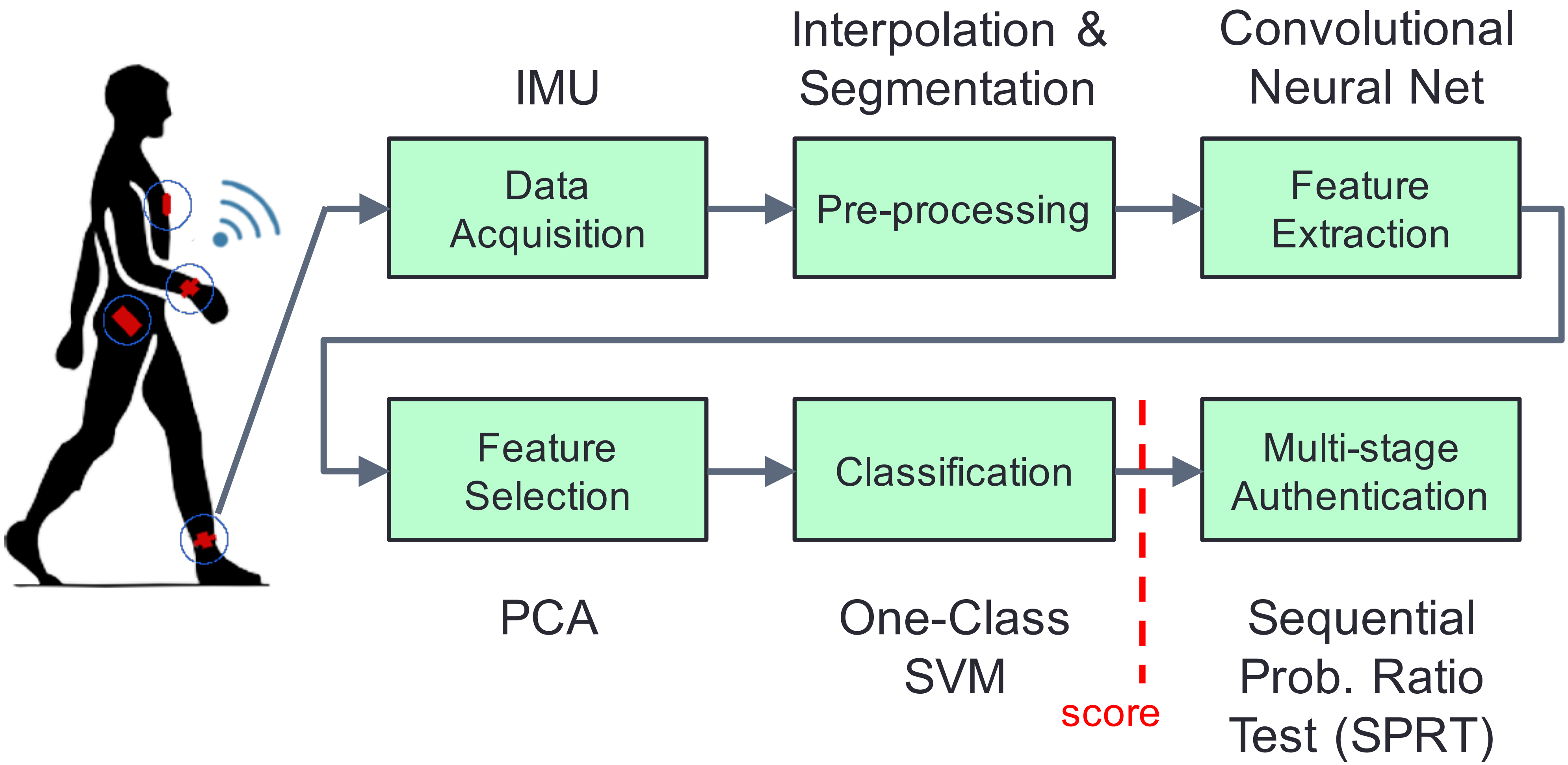
Results: temperature data [WML12]



[H++16] M. Hooshmand, M. Rossi, D. Zordan, M. Zorzi, “Covariogram-based compressive sensing for environmental wireless sensor networks,” *IEEE Sensors Journal*, Vol. 16, No. 6, 2016.

Human Centric Sensing

System Model

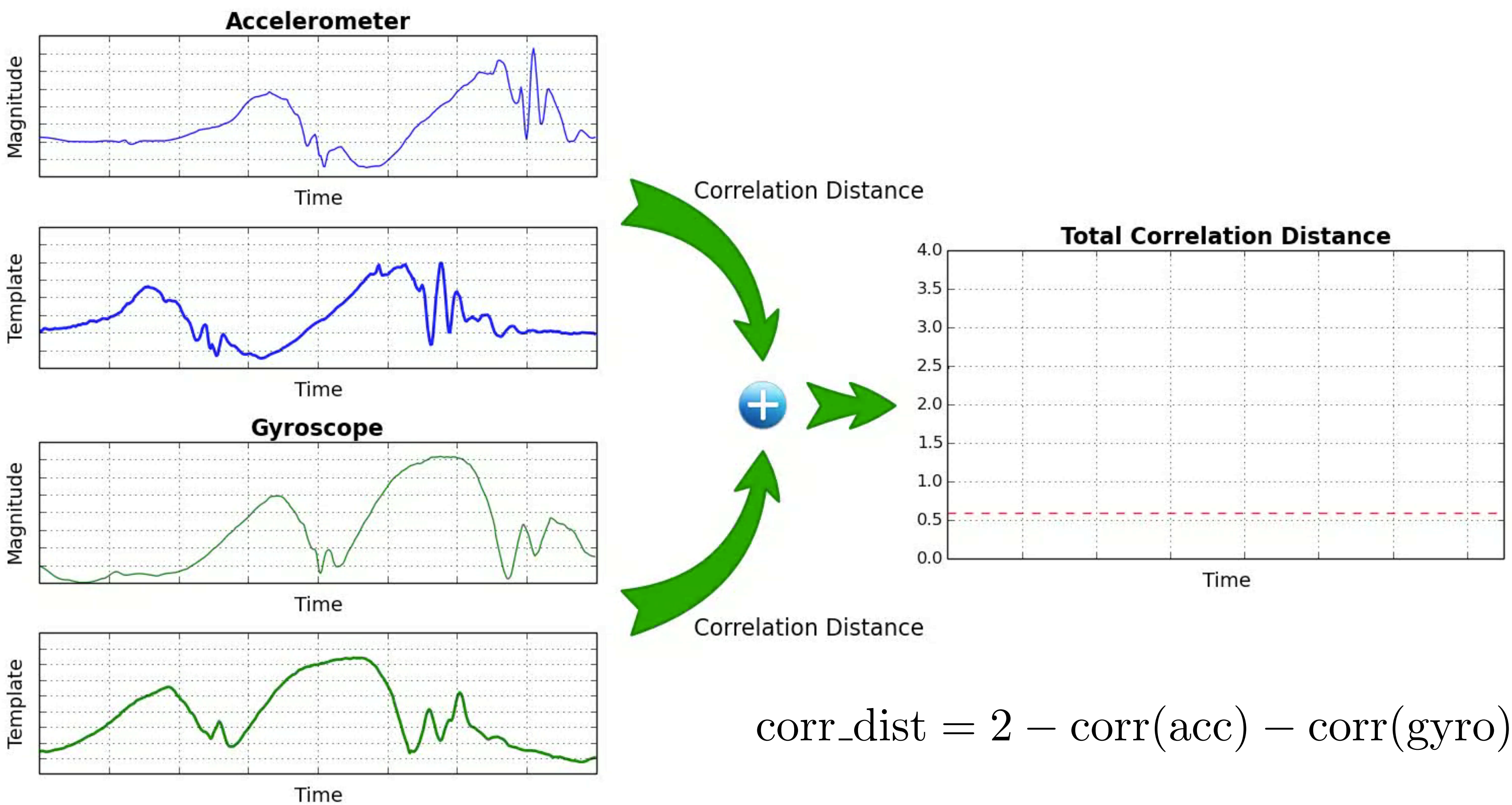


Data Acquisition

Inertial Measurement Unit (IMU)

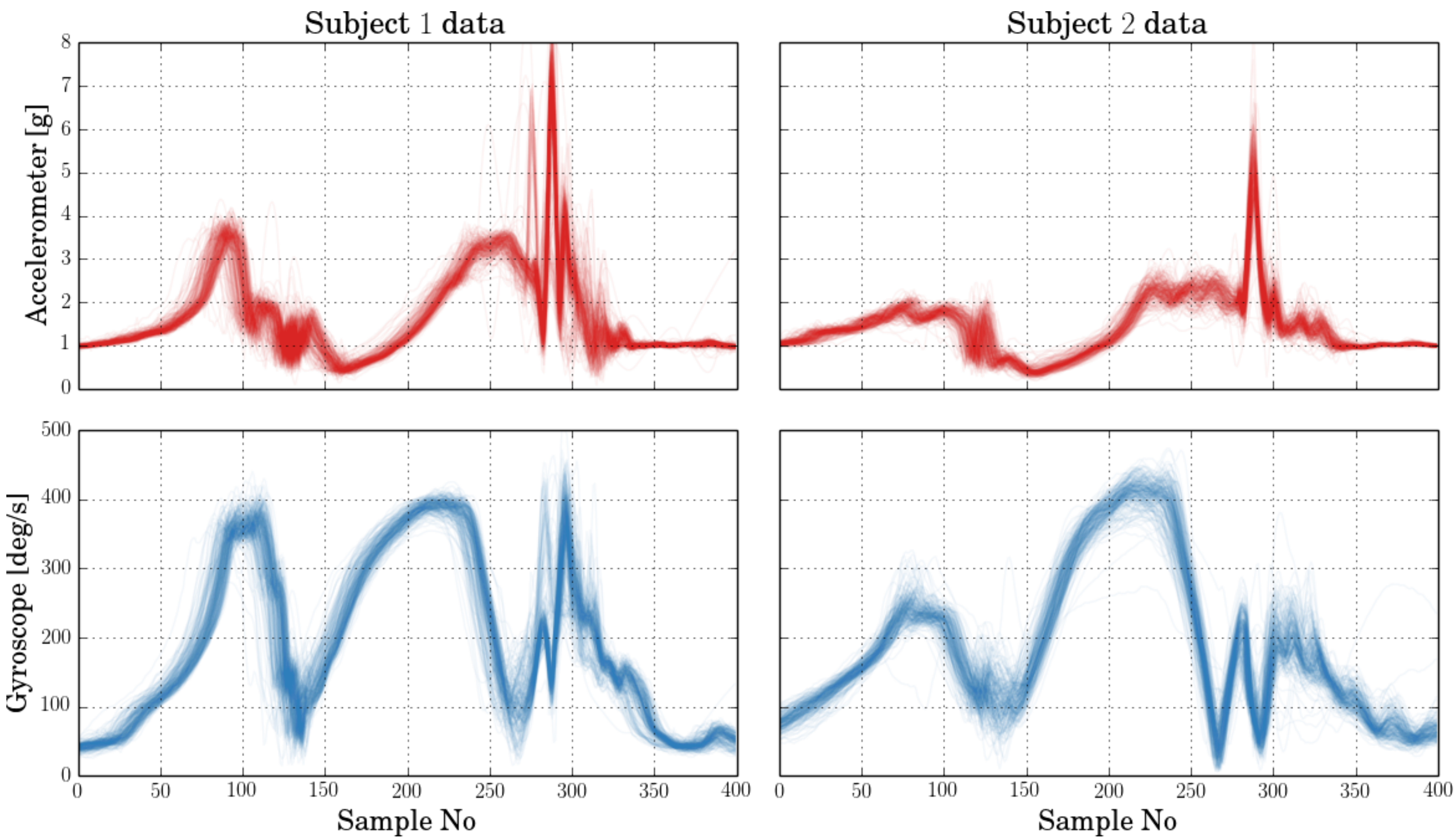
- **Axivity WAX9**
  - **Accelerometer** ±2 / 4 / **8 g** (14 bit resolution)
  - Magnetometer ±1mT (16 bit resolution)
  - **Gyroscope** ±250 / 500 / **2000 dps** (16 bit resolution)
  - Temperature 0 - 65 °C (0.1˚C resolution)
  - Pressure 30-110 kPA (1Pa resolution)
  - Max. sampling frequency: **400 Hz**
  - Bluetooth LE radio

Template-based Segmentation

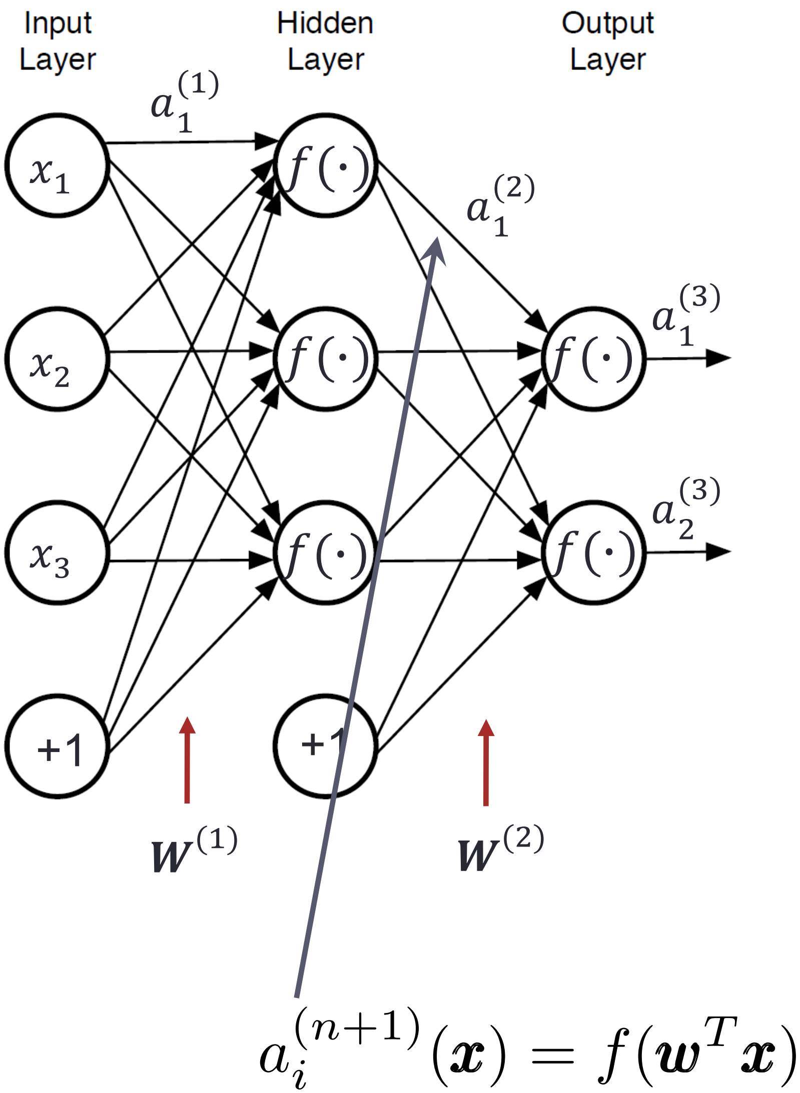


$$\mathrm{corr\_dist} = 2 - \mathrm{corr}(\mathrm{acc}) - \mathrm{corr}(\mathrm{gyro})$$

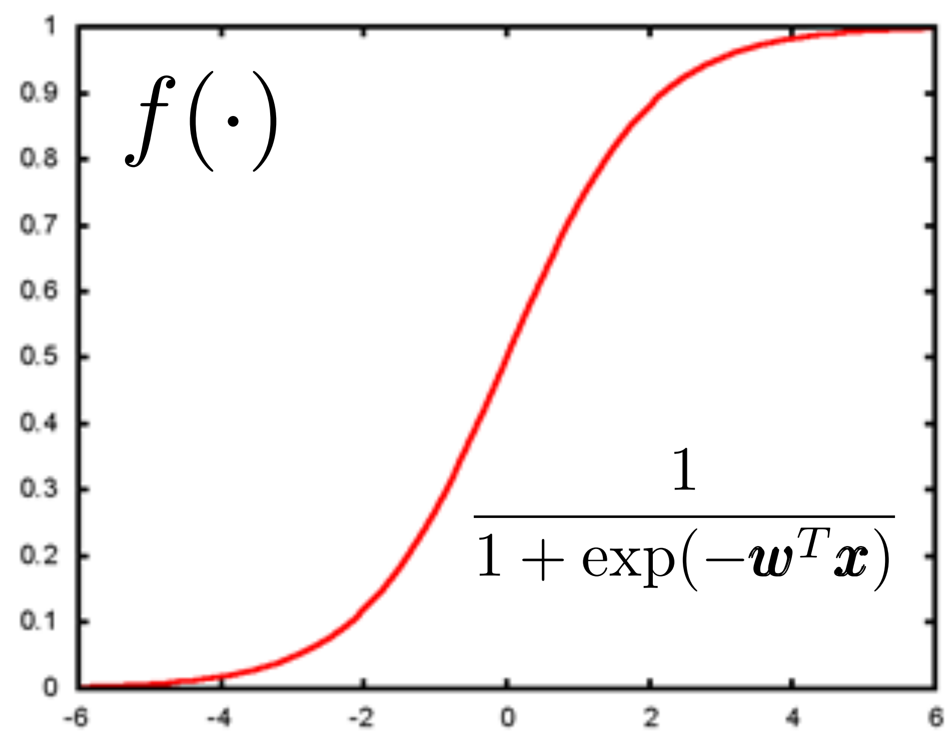
Walking Pattern Examples



Feed-Forward Neural networks (FF-NN)



- Feed Forward NN
- Layers of neurons
- Non-linear activation functions
- Set of weights

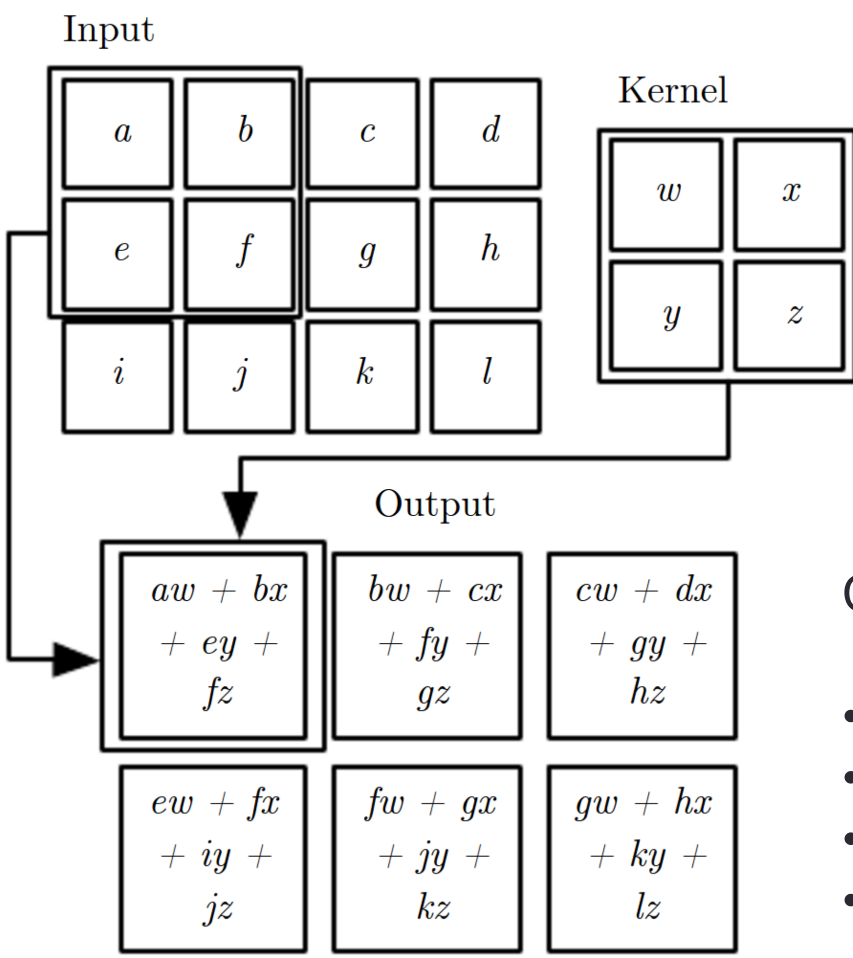


Convolutional Neural Network (CNN)

- Supervised training (15 subjects)
  - 10 minutes of walk, at least 2 sessions
  - 2 minutes for training, 8 to test the CNN
- Implicit extraction of relevant features
  - Automatic feature engineering
- Shared weights
  - Lower number of pars than Feed Forward NN
- Convolution and Max. Pooling
  - Equivariant to translation of features in input space
  - Invariance to other local transforms

Convolutional Neural Network (CNN)

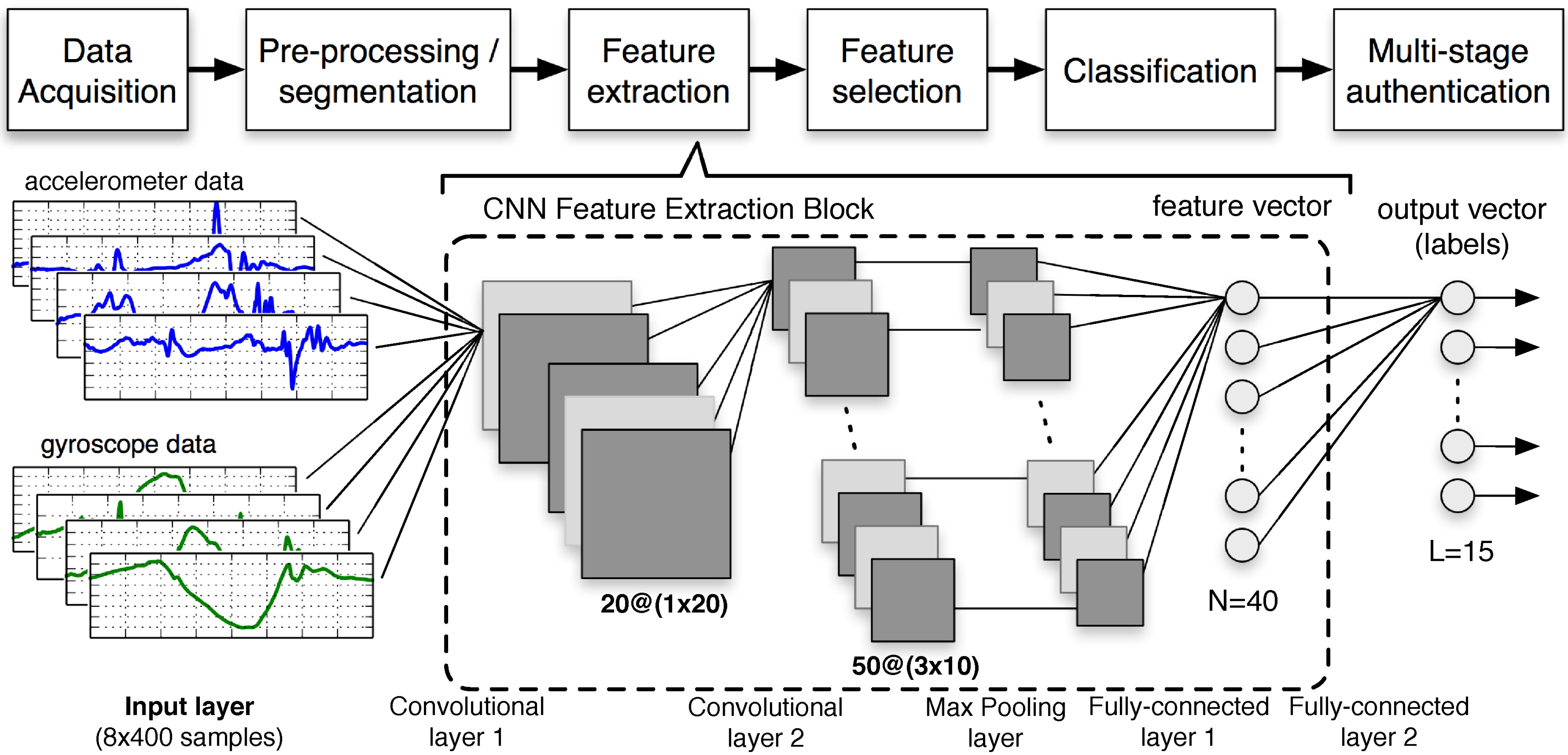
$$\underbrace{S(i,j)}_{\text{Output}} = (I * K)(i,j) = \sum_m \sum_n \underbrace{I(i+m, j+n)}_{\text{Input}} \underbrace{K(m,n)}_{\text{Kernel}}$$



CNN features

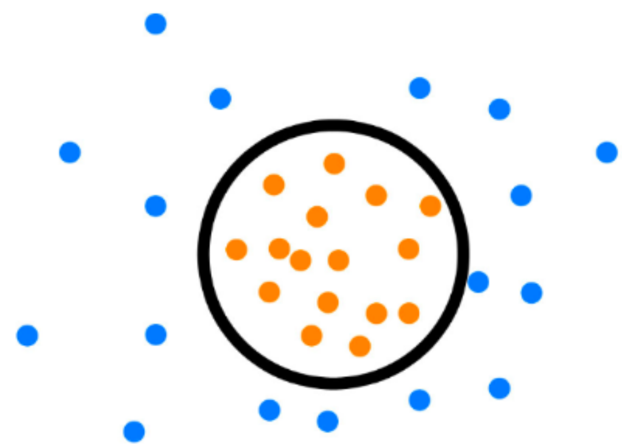
- Kernel reuse
- Memory savings
- Complexity reduction
- Robust to translations

Convolutional Neural Network



One-Class SVM [S+00]

- Only target class (orange) data is available
- Training finds the *boundary* (tick circle)
- Classification output: score
  - Distance from the boundary
- Training
  - We only used data from target user
- Test
  - Data from the negative class were also used
  - Performance assessment



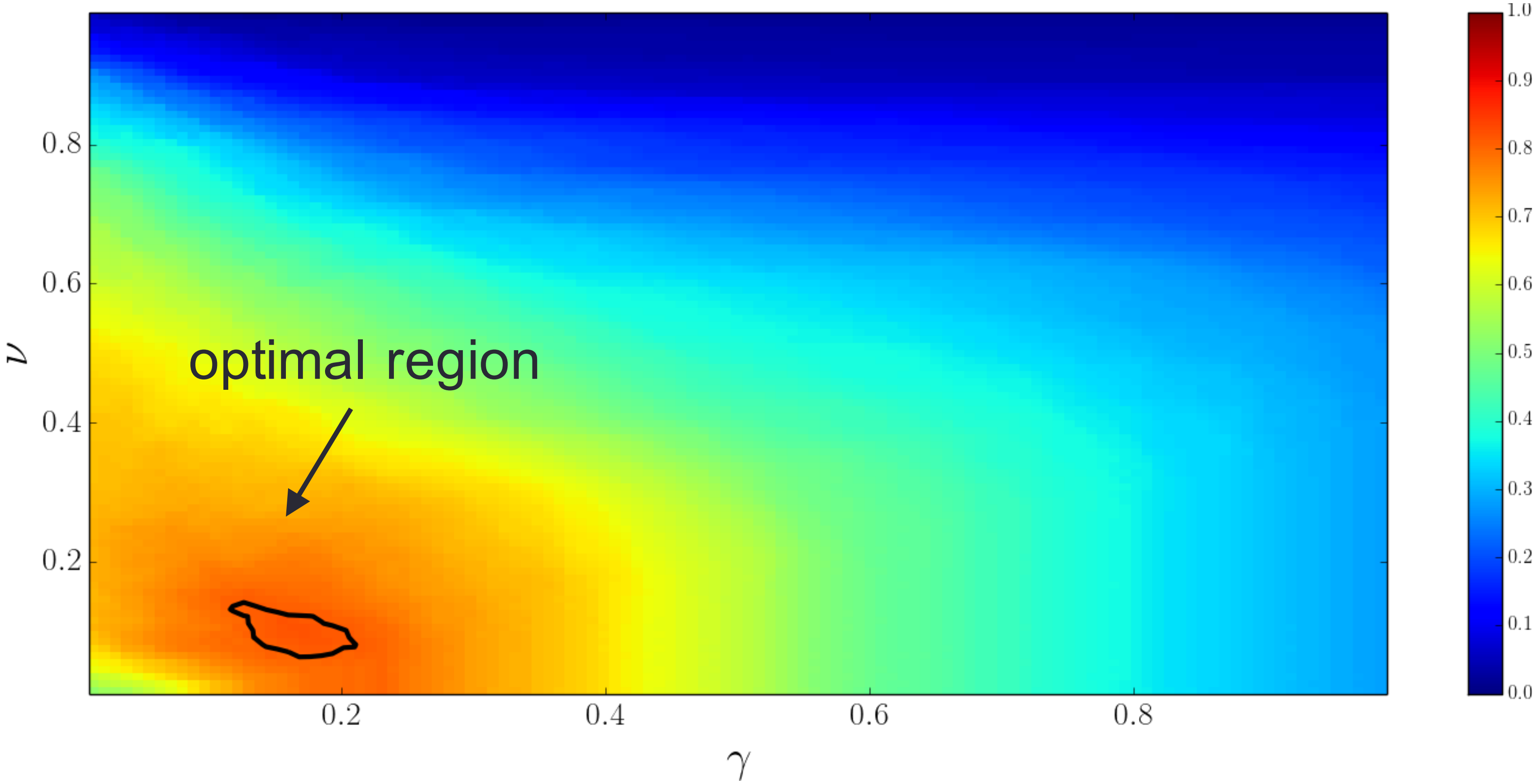
[S++00] B. Schölkopf, J.C. Plattz, J. Shawe-Taylory, A.J. Smolax, Robert C. Williamson, “Estimating the Support of a High-Dimensional Distribution,” Technical Report MSR-TR-99-87, Microsoft Research, Redmond, WA, 2000.

One-Class SVM

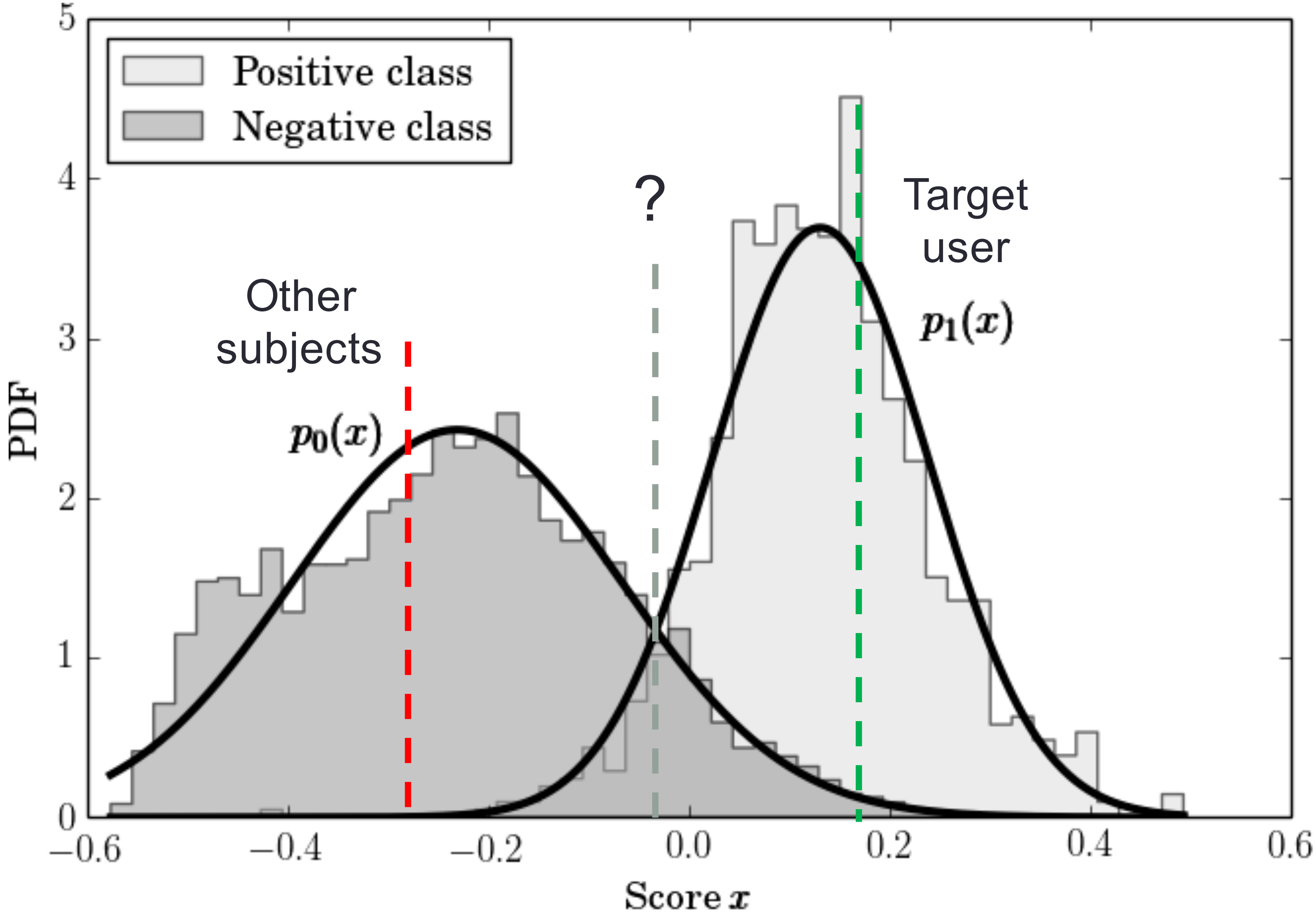
Precision: no. of true classified true / tot. number of positives
Recall: no. of true classified true / tot. number of true

$$F = 2\,\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

F-measure



One-Class SVM (test output)



Sequential Decision Making

Sequential Probability Ratio Test (SPRT) [W47]

$$p_0(x) = p(x|H_0) \qquad H_0 \longrightarrow \text{OTHER USERS' CLASS}$$

$$p_1(x) = p(x|H_1) \qquad H_1 \longrightarrow \text{TARGET CLASS}$$

Log-likelihood Ratio

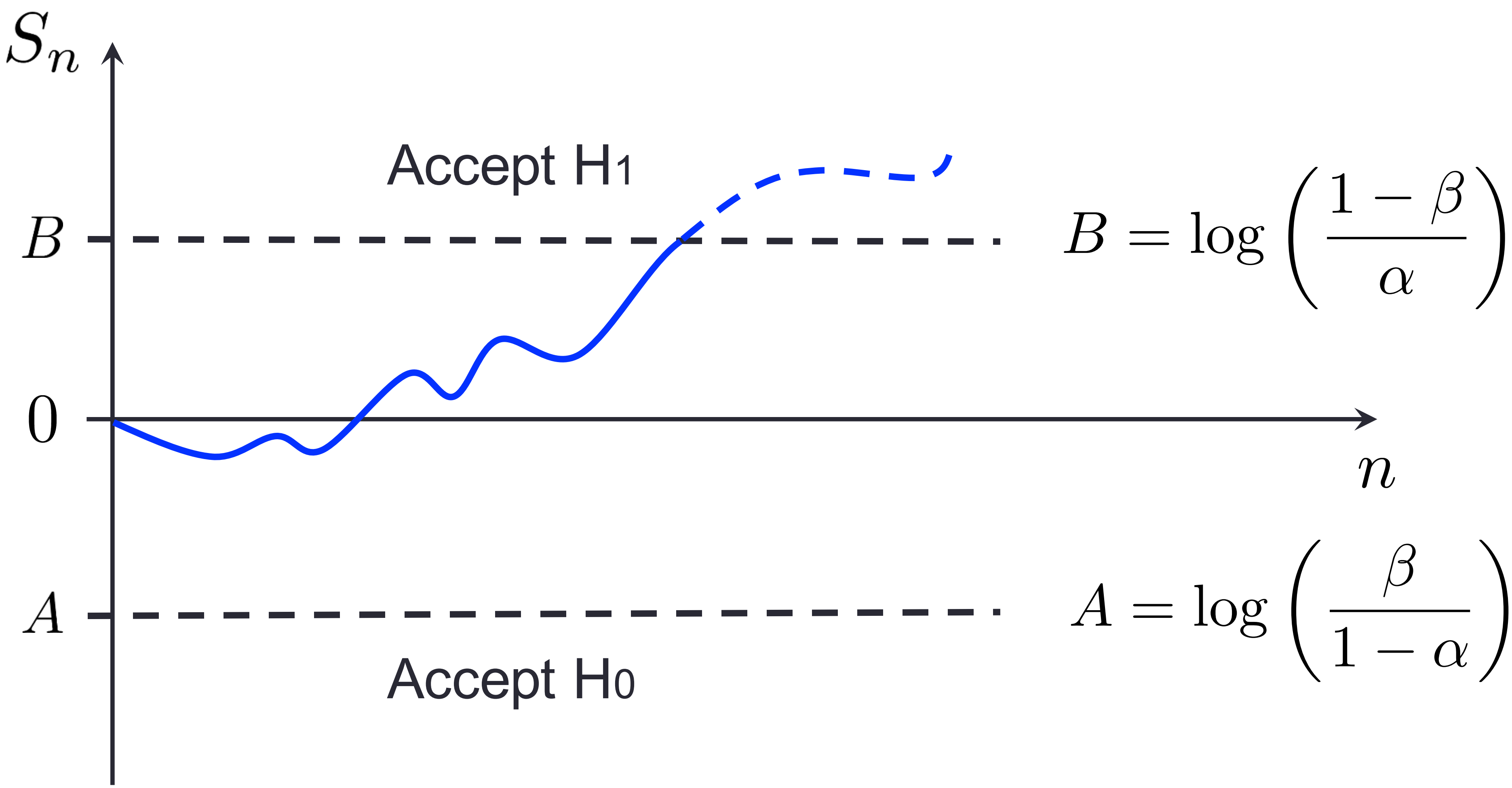
$$S_n = \log\left(\frac{p(x_1,\ldots,x_n|H_1)}{p(x_1,\ldots,x_n|H_0)}\right) = \sum_{i=1}^{n}\log\left(\frac{p_1(x_i)}{p_0(x_i)}\right)$$

$\beta =$ probability of accepting H0 when H1 is true	$\alpha =$ probability of accepting H1 when H0 is true

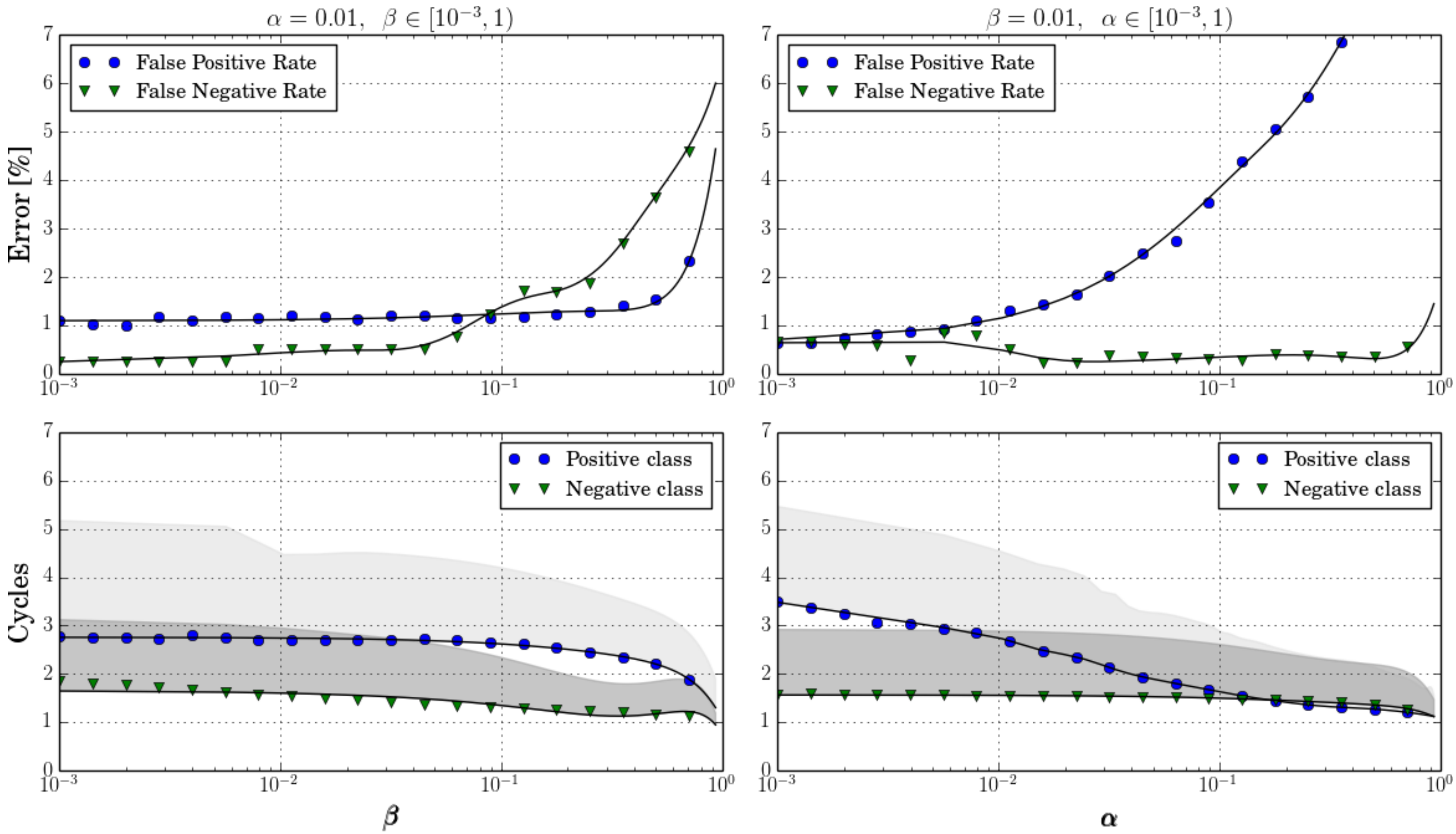
[W47] A. Wald, *Sequential analysis*, Dover, New York, NY, US, 1947.

[TNB15] A. Tartakovsky, I. Nikiforov, M. Basseville, "Sequential Analysis Hypothesis Testing and Changepoint Detection," *CRC Press*, 2015.

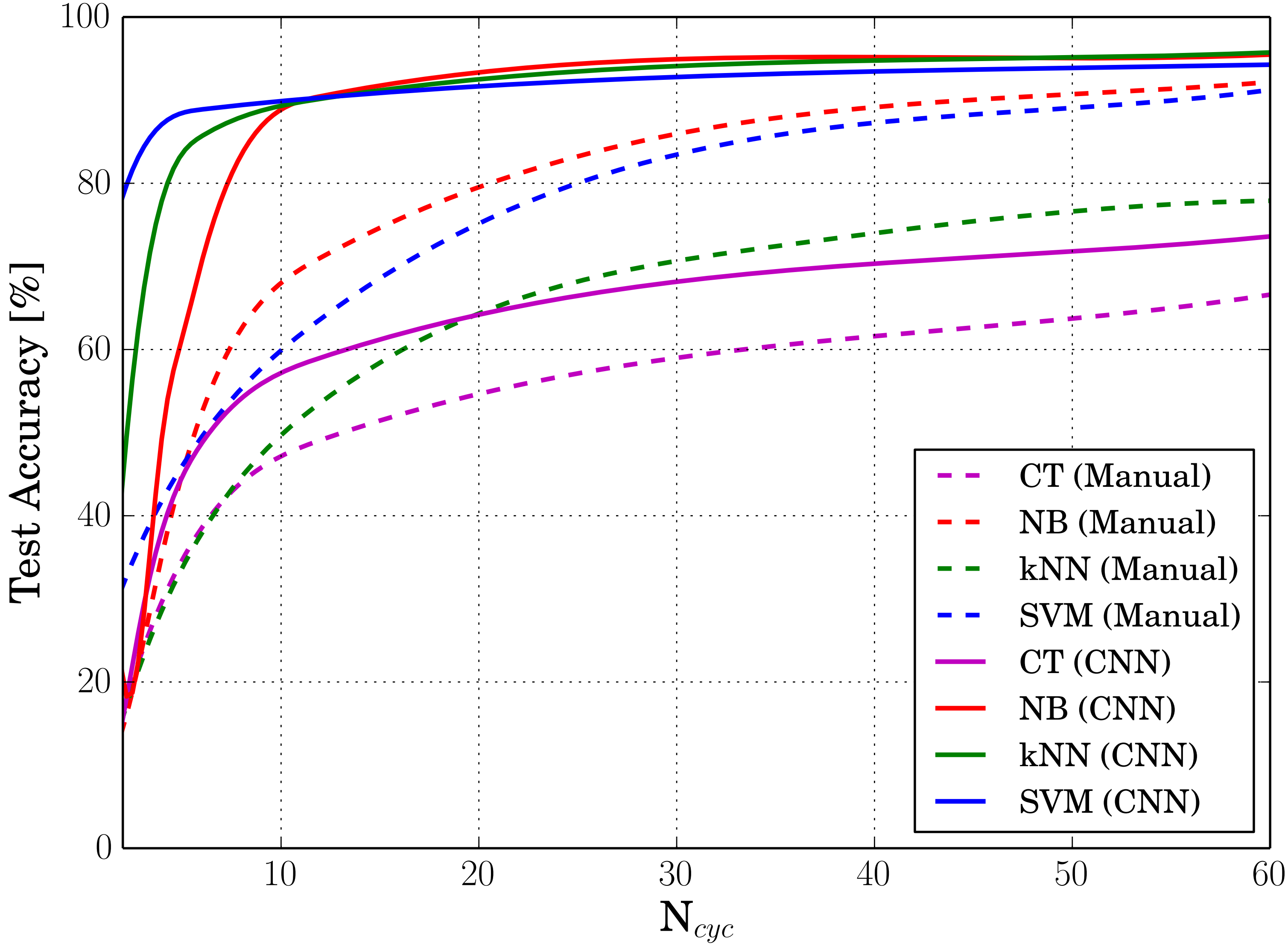
Sequential Analysis



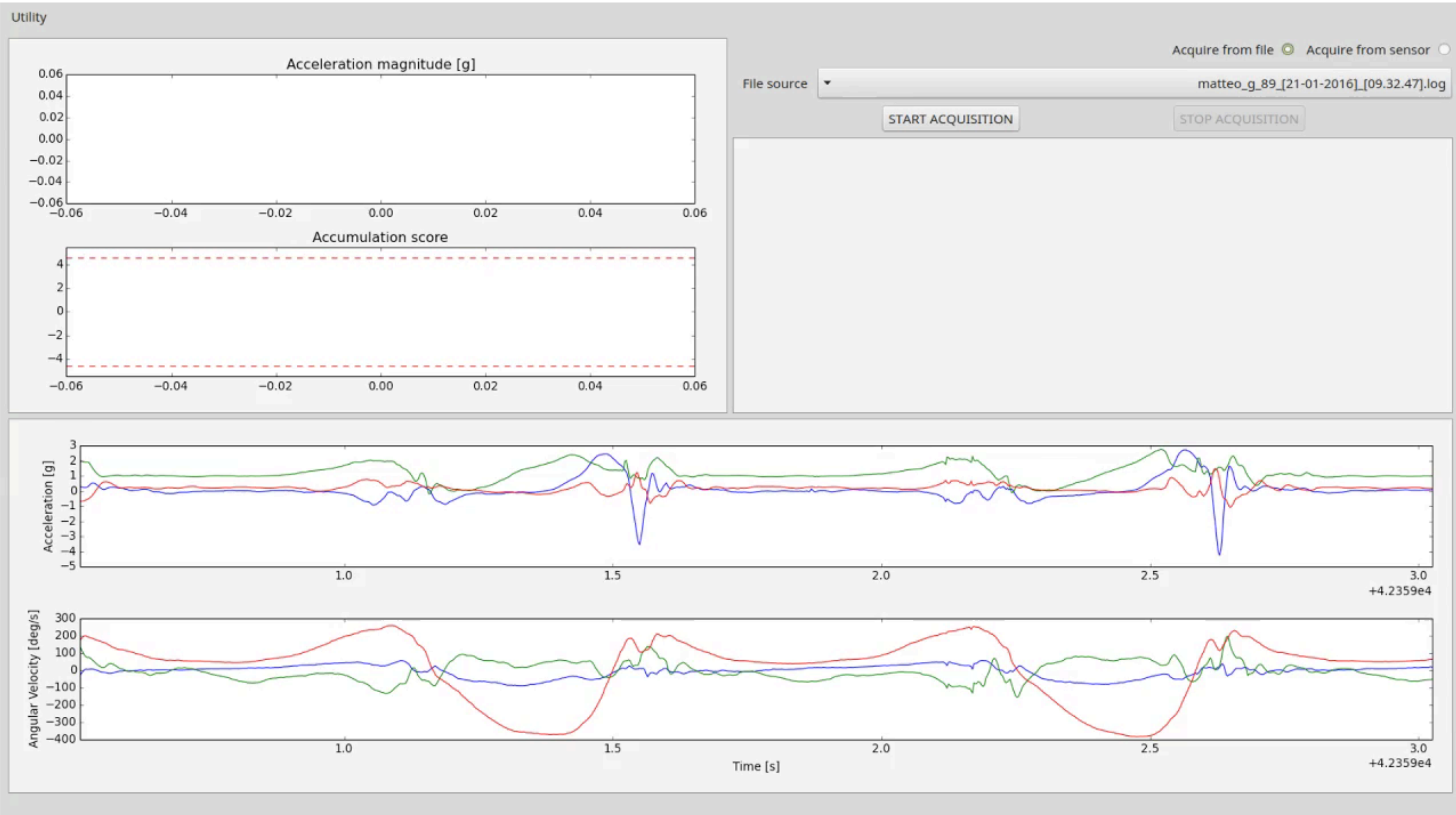
Final Results



Feature extraction CNN *vs* SOTA



Demo


Utility
Acceleration magnitude [g]
Accumulation score
Acquire from file
Acquire from sensor
File source
matteo_g_89_[21-01-2016]_[09.32.47].log
START ACQUISITION
STOP ACQUISITION
Acceleration [g]
Angular Velocity [deg/s]
Time [s]
+4.2359e4

Rethinking IoT Protocol Design

Thus far

- Networking protocol design
  - Routing, channel access
- Common assumptions
  - Backlogged queues
  - Some sort of arrival process
- Often neglected
  - signal-processing (e.g., compression)
  - signal-statistics

Our point

- In Future Networks
  - Functionalities & algorithms will be injected as new software
  - Software Defined Networking
- Signal statistics depend on
  - What is being measured
  - Where
  - When
- Statistics may change as a function of time

⇒ classification, estimation, adaptation

Protocol Reconfiguration

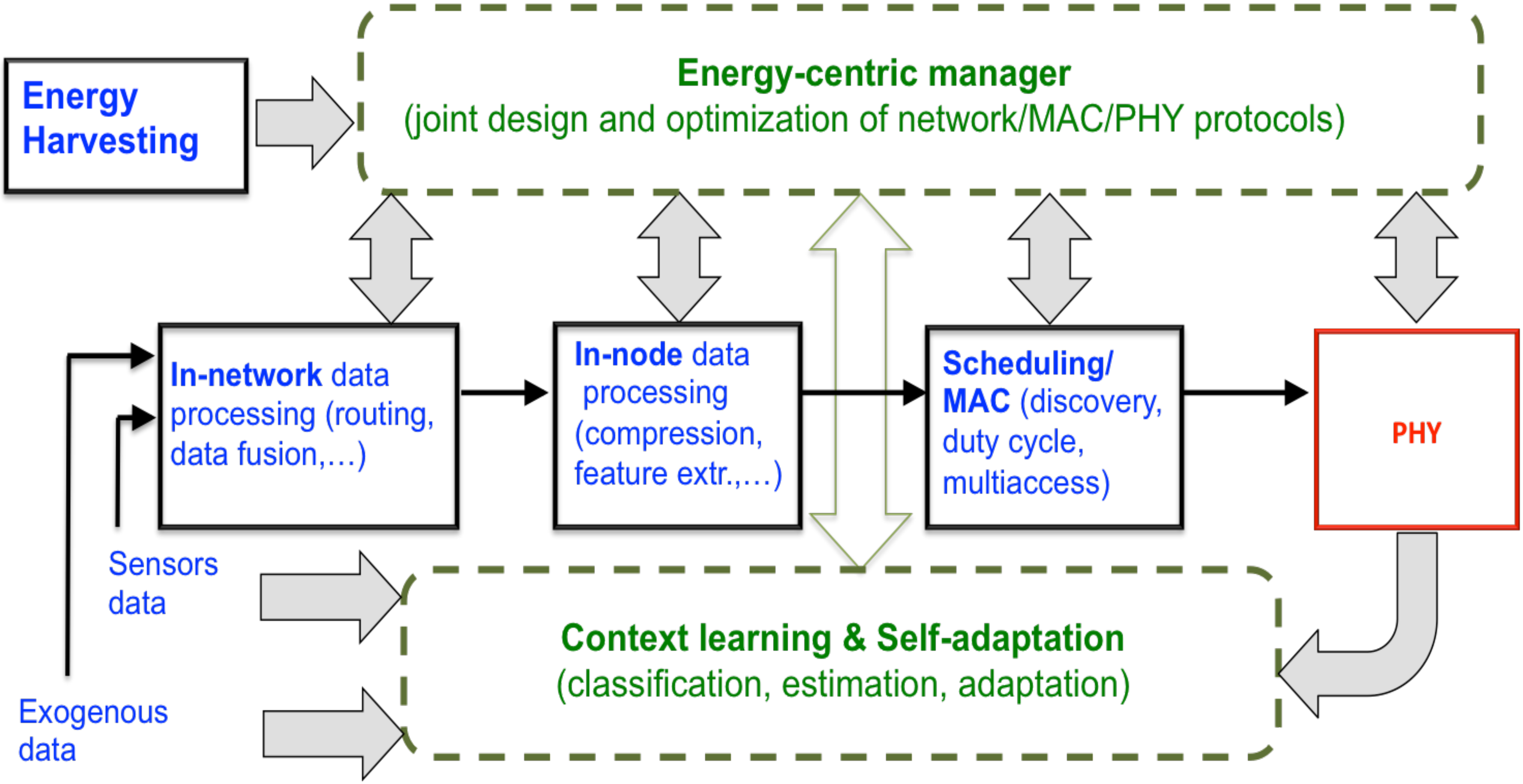
- Channel Access Behavior
  - Periodic *vs* bursty *vs* asynchronous
  - Example: dynamic setting of MAC pars *or*
  - Dynamic switching the MAC protocol in use [AZD16]

- Routing behavior
  - What can be aggregated (data fusion)
  - Make routing decisions based on spatio-temporal statistics

- Routing vs compression at the sources
  - When and how much to compress
  - Depending on: cost of transport, data fidelity

[AZD16] A. Asudeh, G.V. Zàruba, S.K. Das, "A general model for MAC protocol selection in wireless sensor networks," Ad Hoc Networks, Vol. 36, Part. 1, January 2016.

The Big Picture



Data Mining Framework

- Highly Comparative Time Series Analysis (HCTSA)
  - Feature Extraction & Classification framework [FLJ13]
  - Over 9000 features, including:
    - Basic statistics, linear correlation, stationarity, information theoretic and entropy measures, non linear time series analysis, model fits, …
  - Supervised classification using:
    - SVM trained on signal features
    - SVM trained on N principal components of the Feature Matrix
    - Linear classification on single features

[FLJ13] B. D. Fulcher, M. A. Little, N. S. Jones, "Highly comparative time-series analysis: the empirical structure of time series and their methods," *J. Roy. Soc. Interface*, 10, 83, 2013.

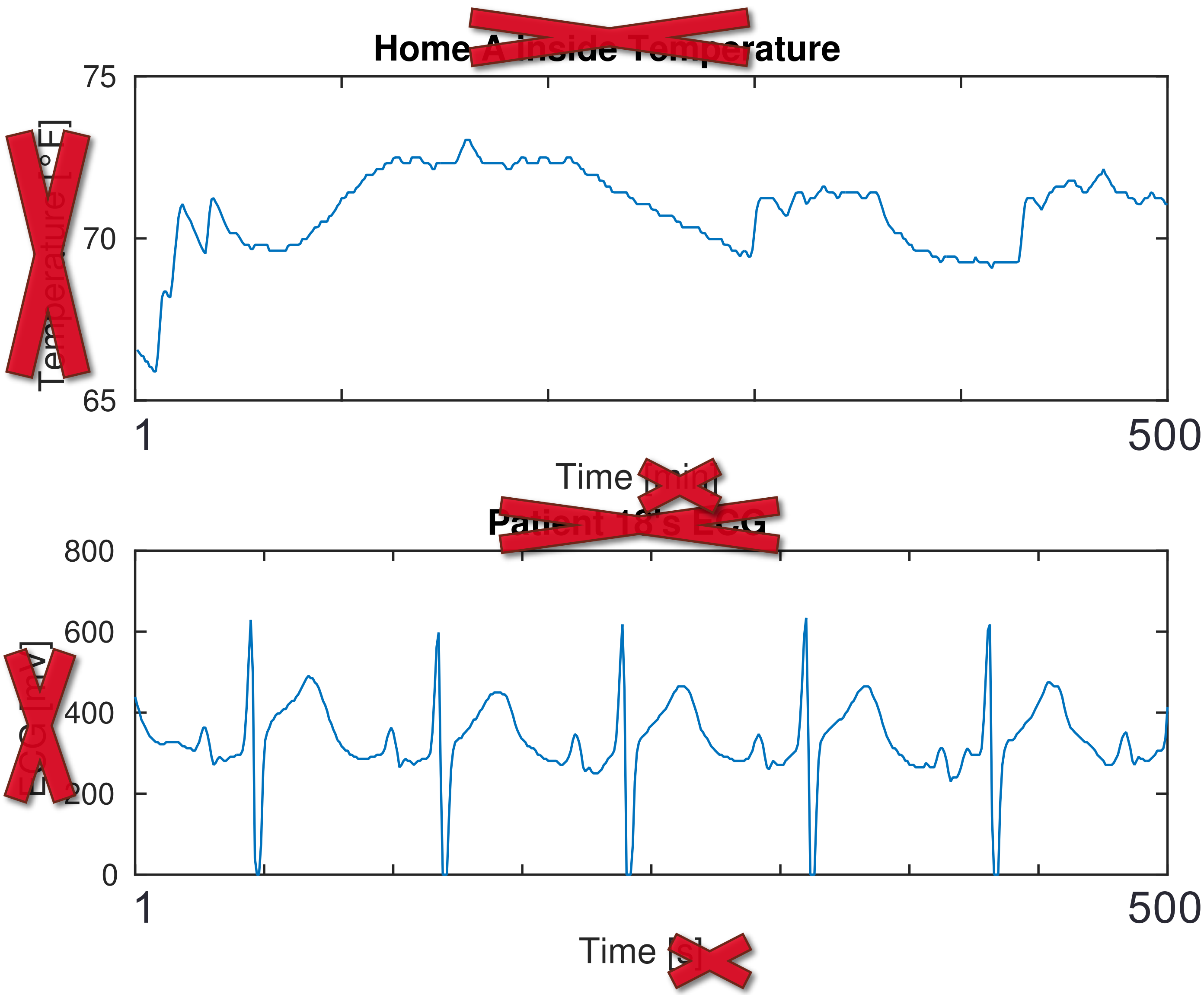
Considered Time Series

- Environmental Data
  - Temperature, Humidity, Wind speed, Wind direction, Rain intensity, Rain rate, Solar irradiation, etc.
- Structural Monitoring
  - Strain gauge sensors
- Biomedical
  - ECG, PPG
- Power Consumption (Households)

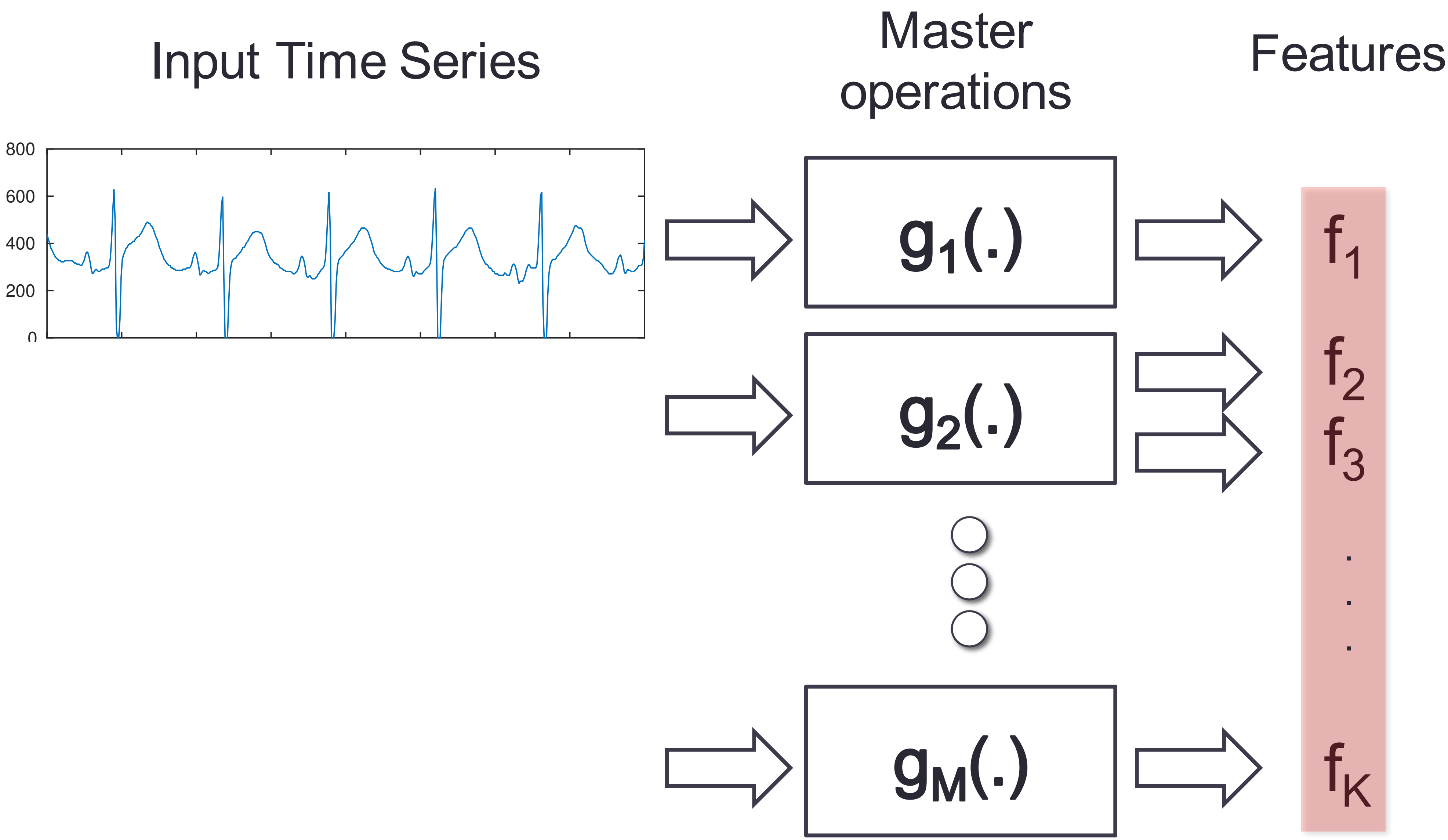
~1400 temporal signals

~7000 windows of 500 samples each

What do we know?

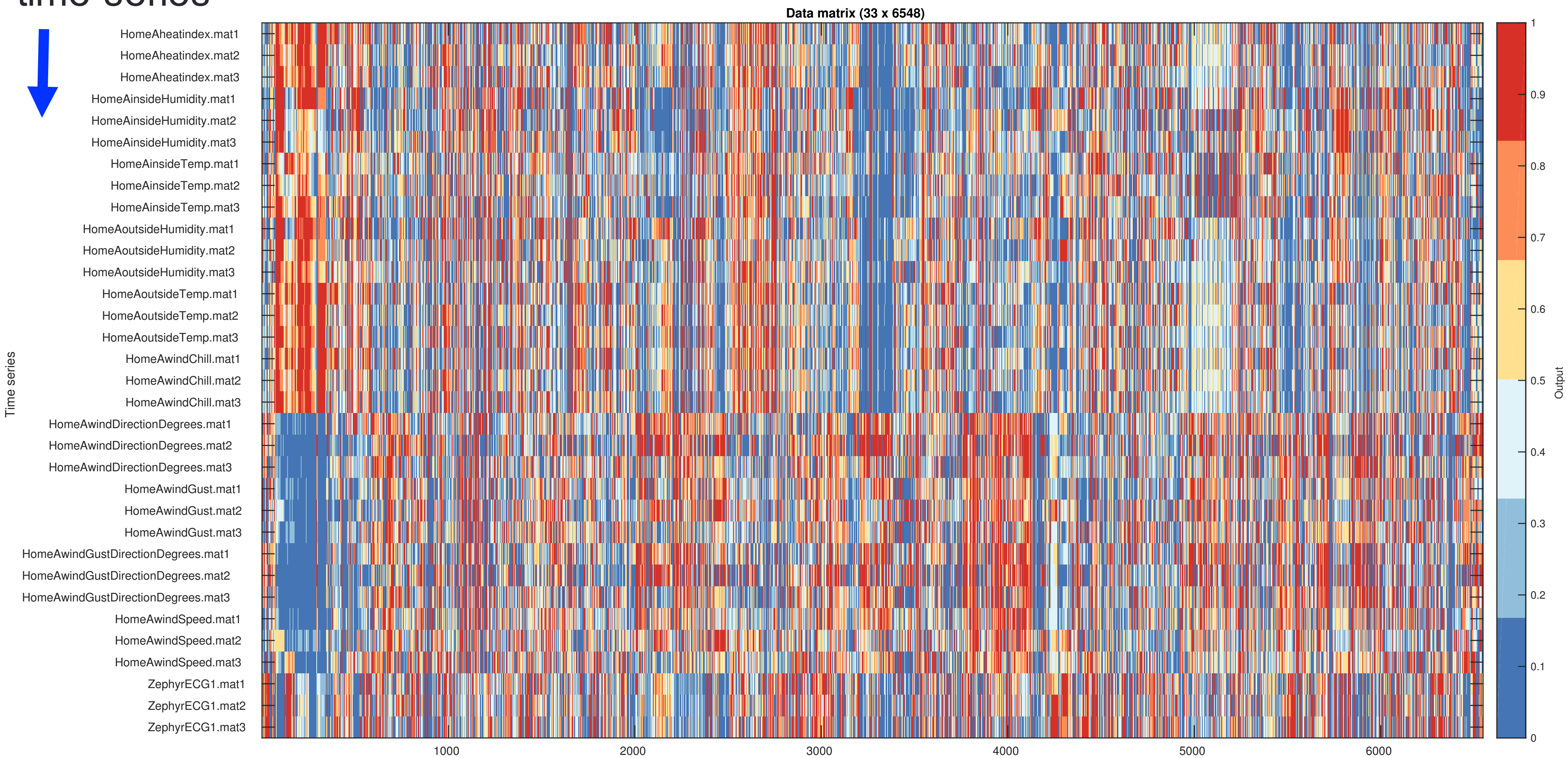


Feature vector



Features (raw)

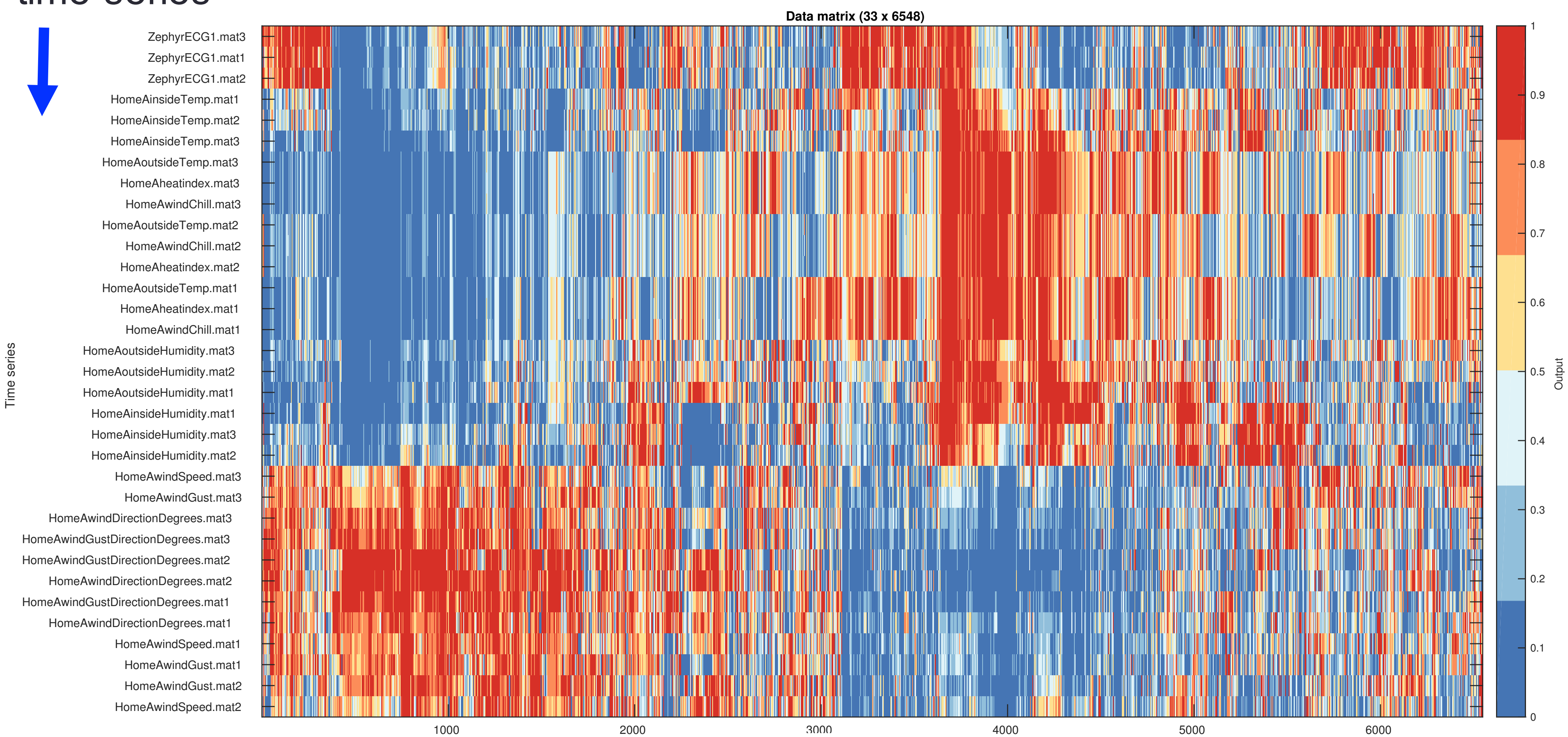
time series



Metrics (6548 features)

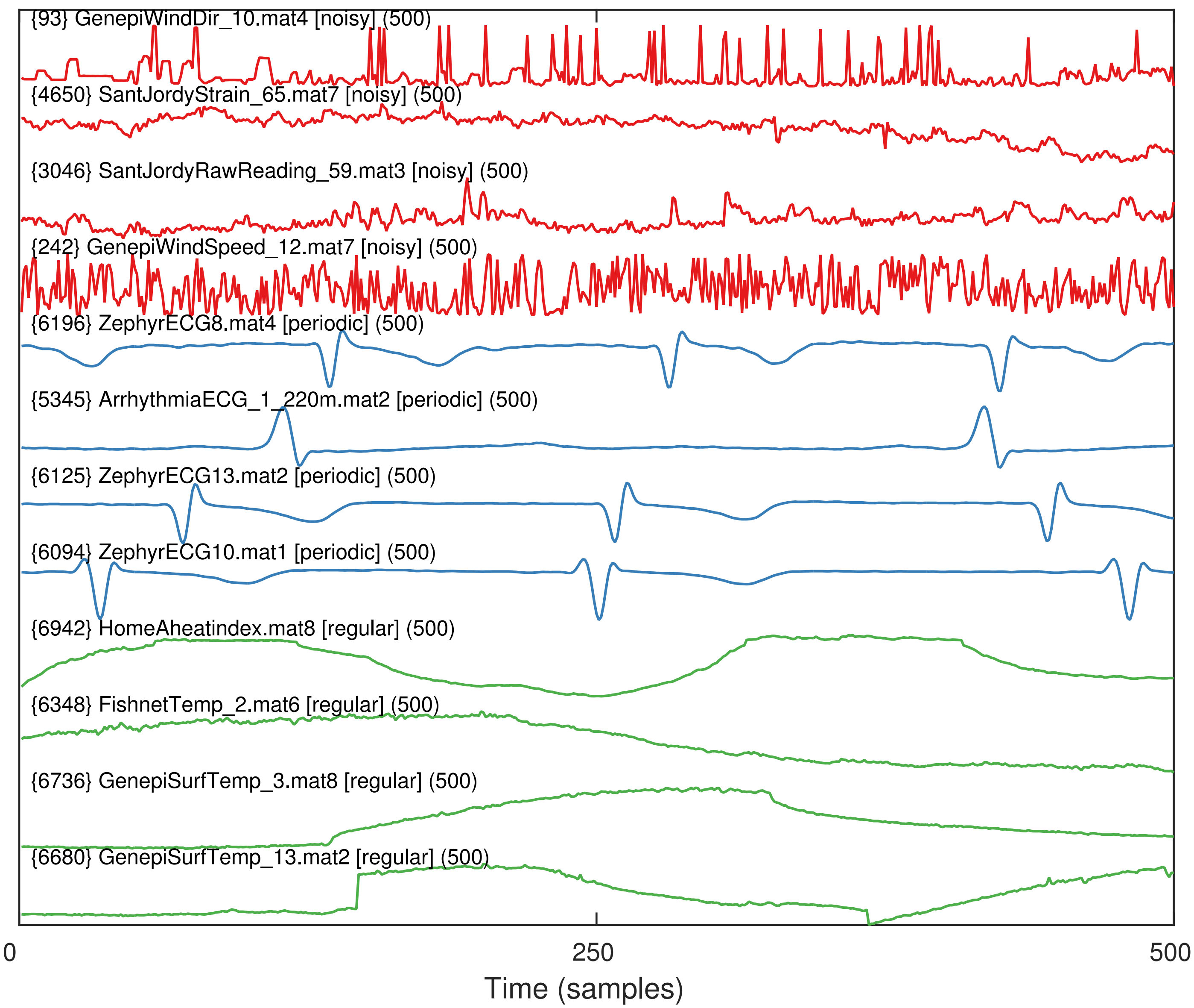
Features (clustered)

time series

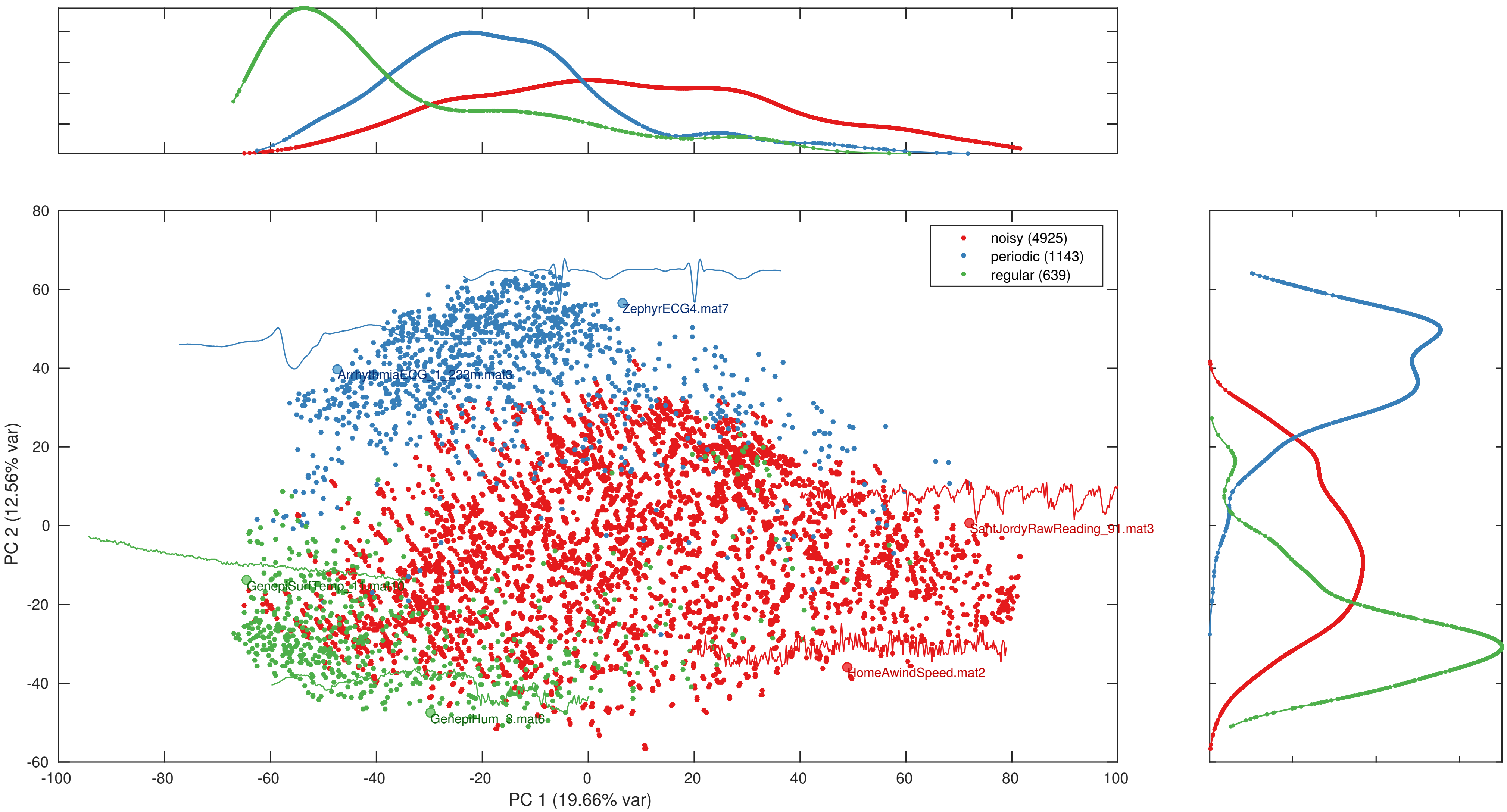


metrics

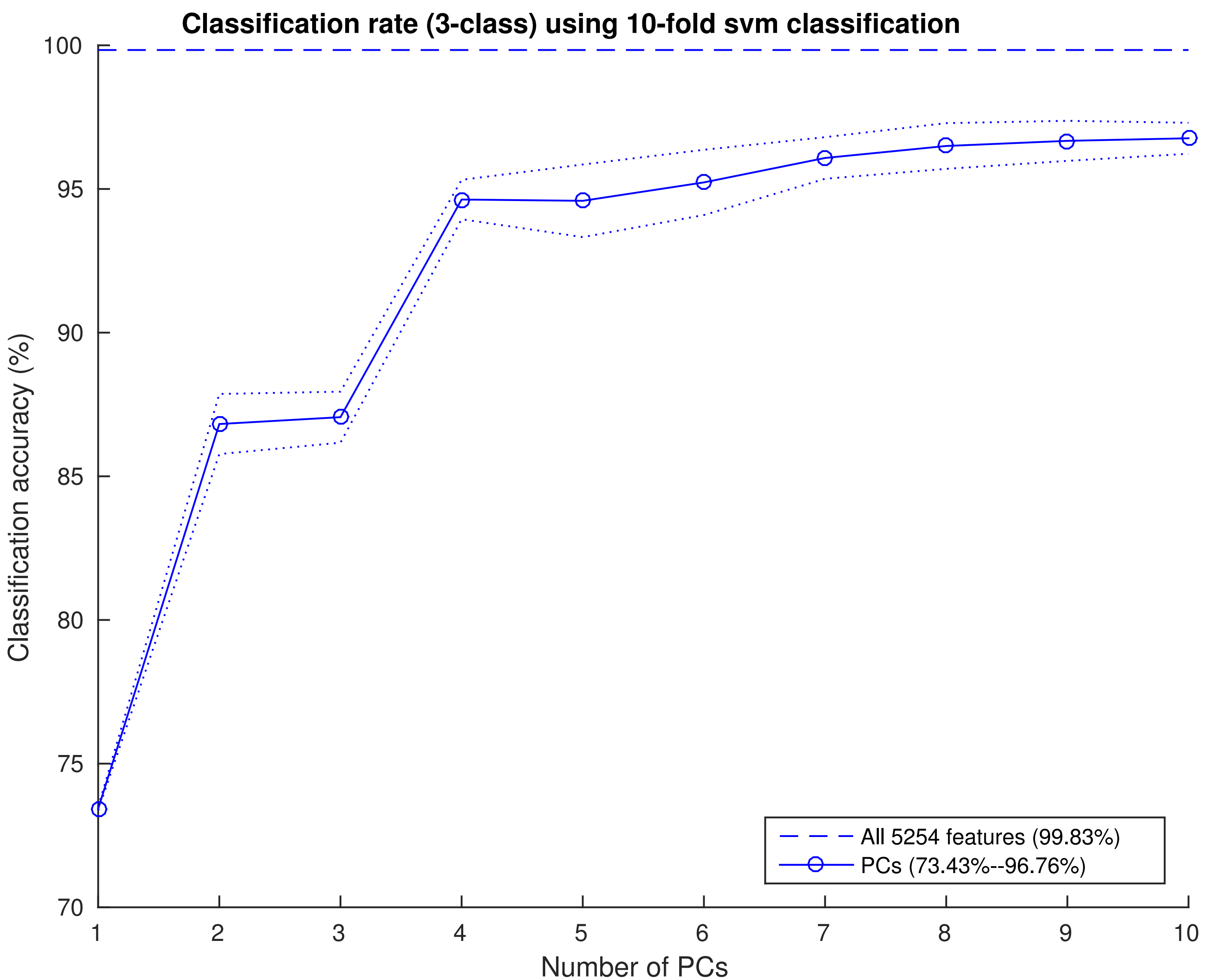
Signal class example



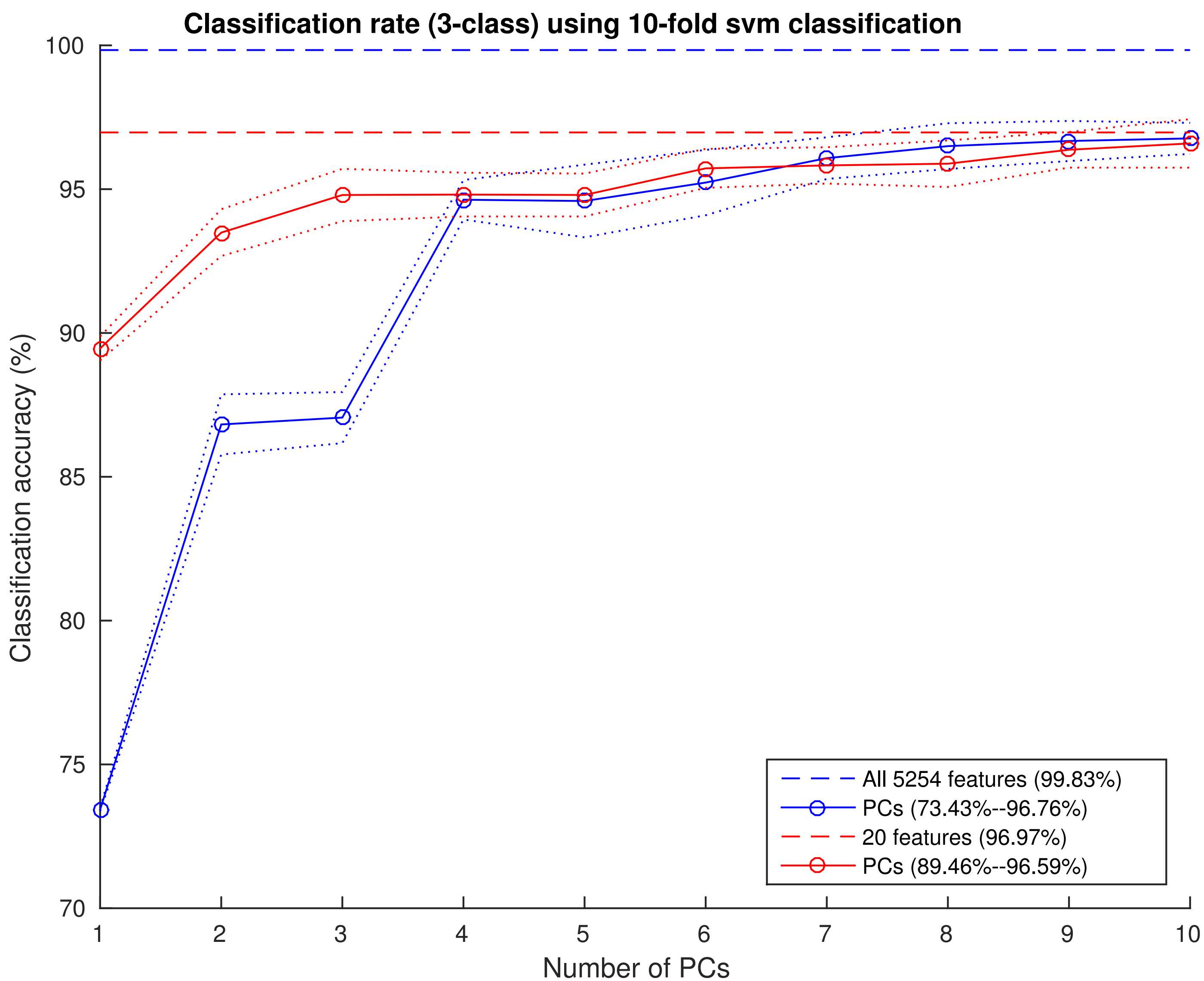
PCA (first two principal components)



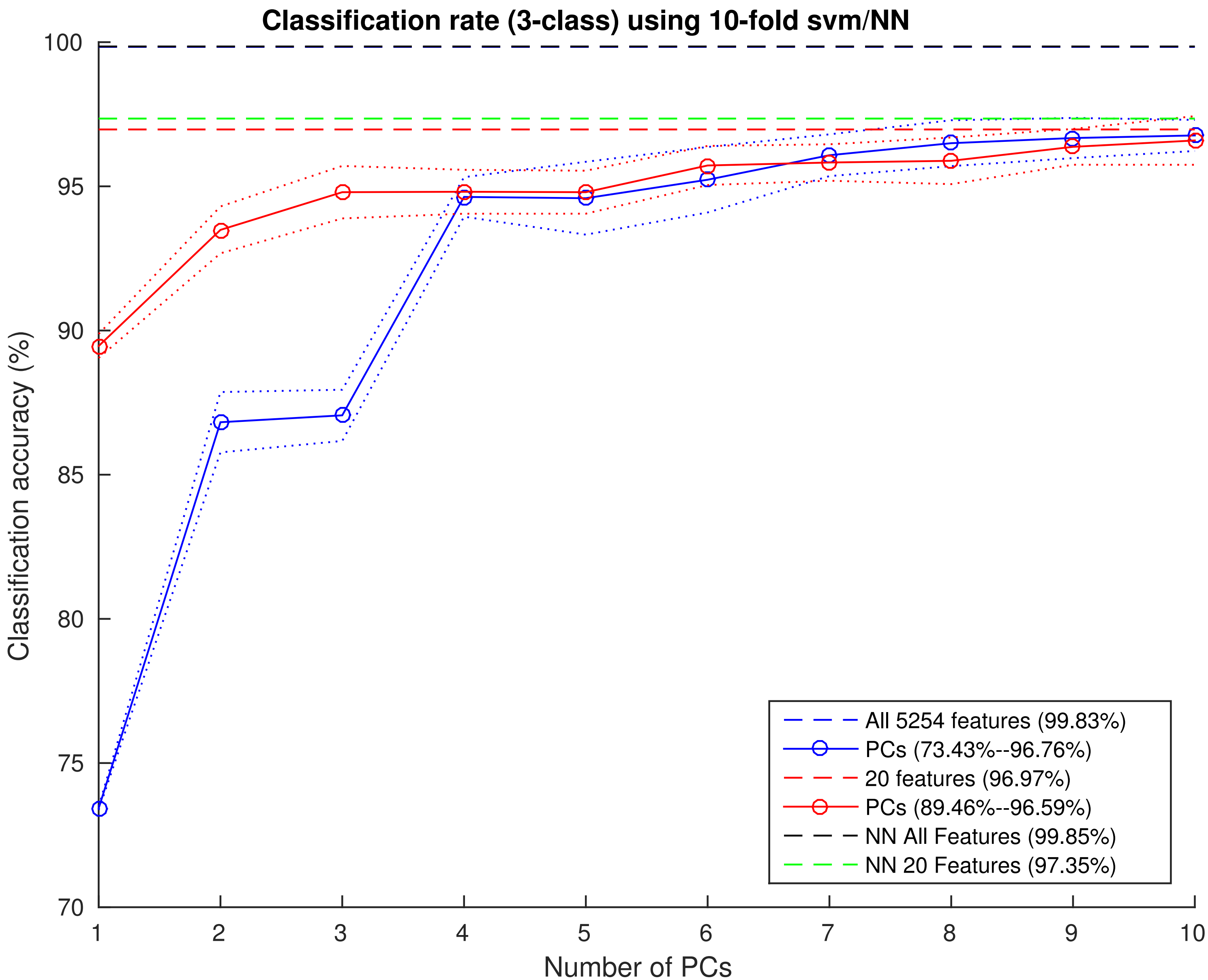
Classification Rate



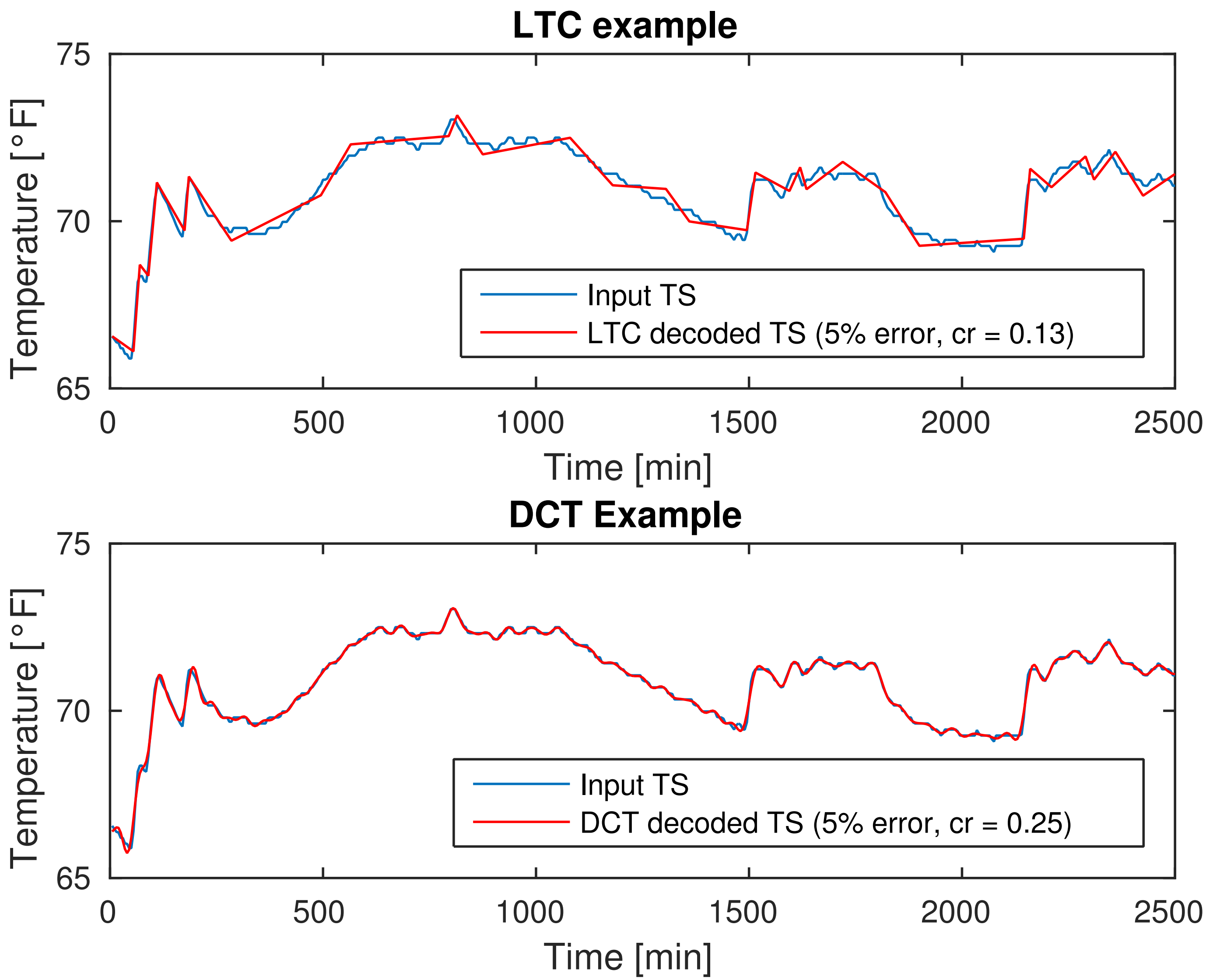
Classification Rate



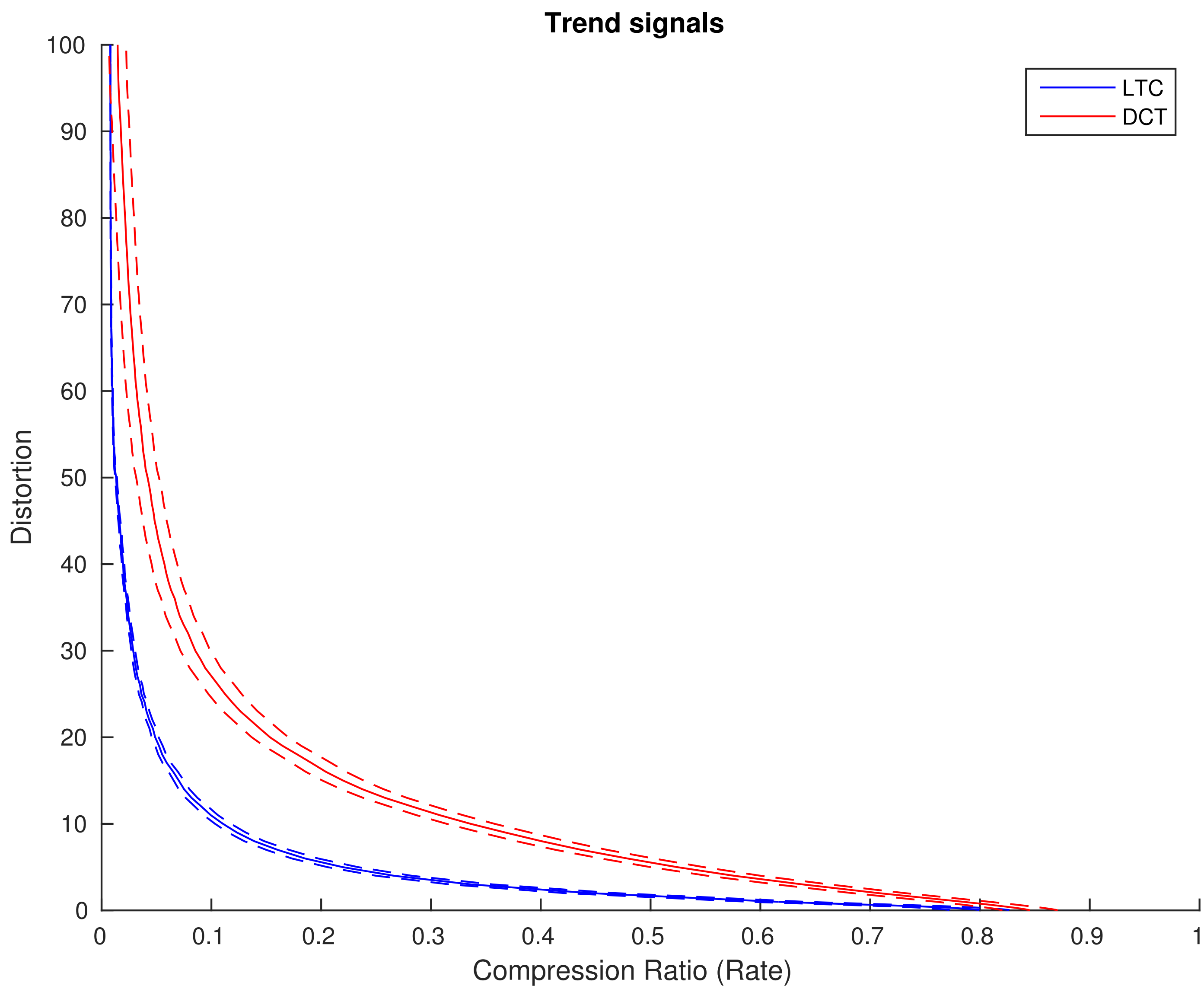
Using Neural Networks (FF-NN)



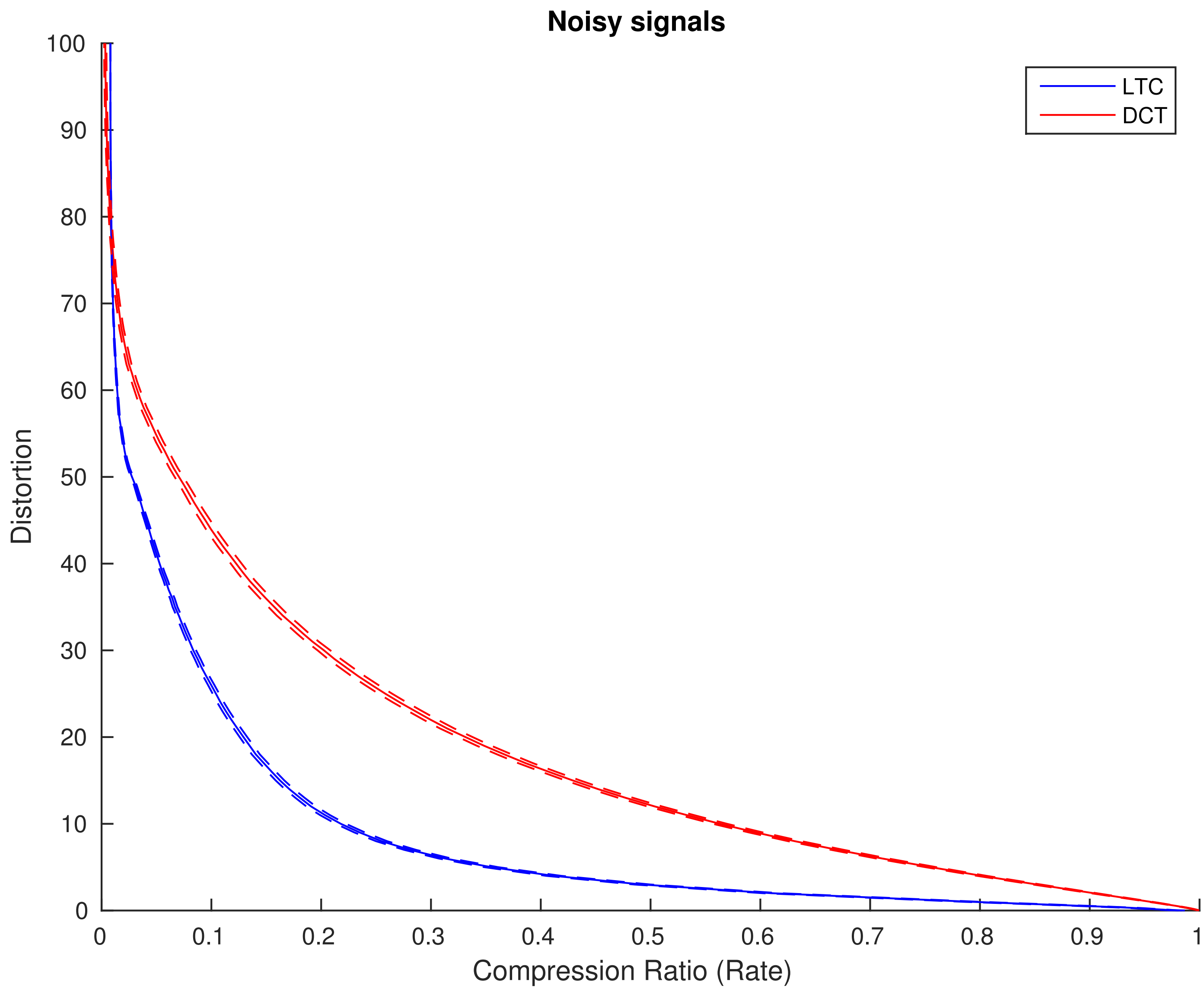
Compression Example (cr=comp/full)



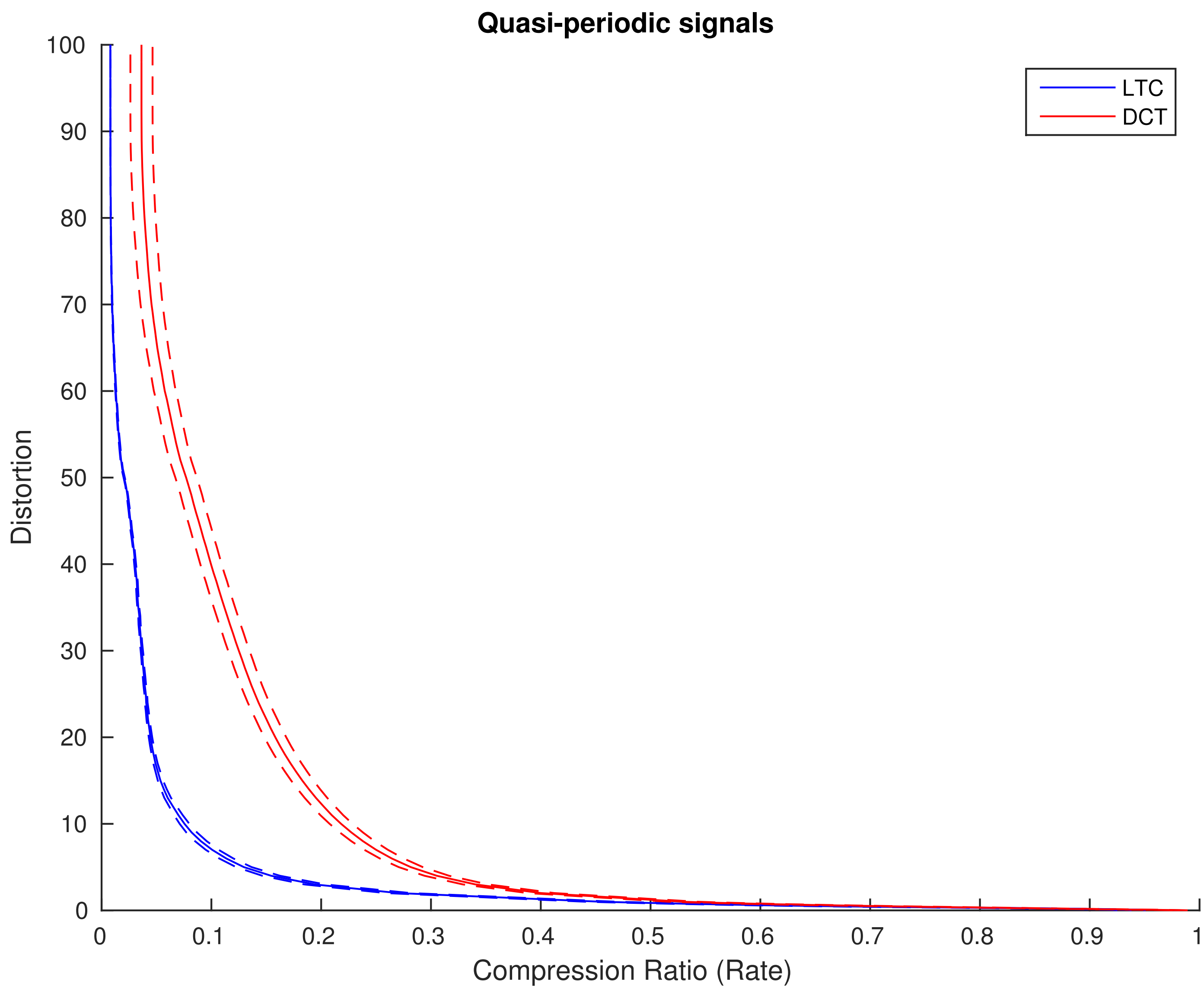
Compression Performance (1/3)



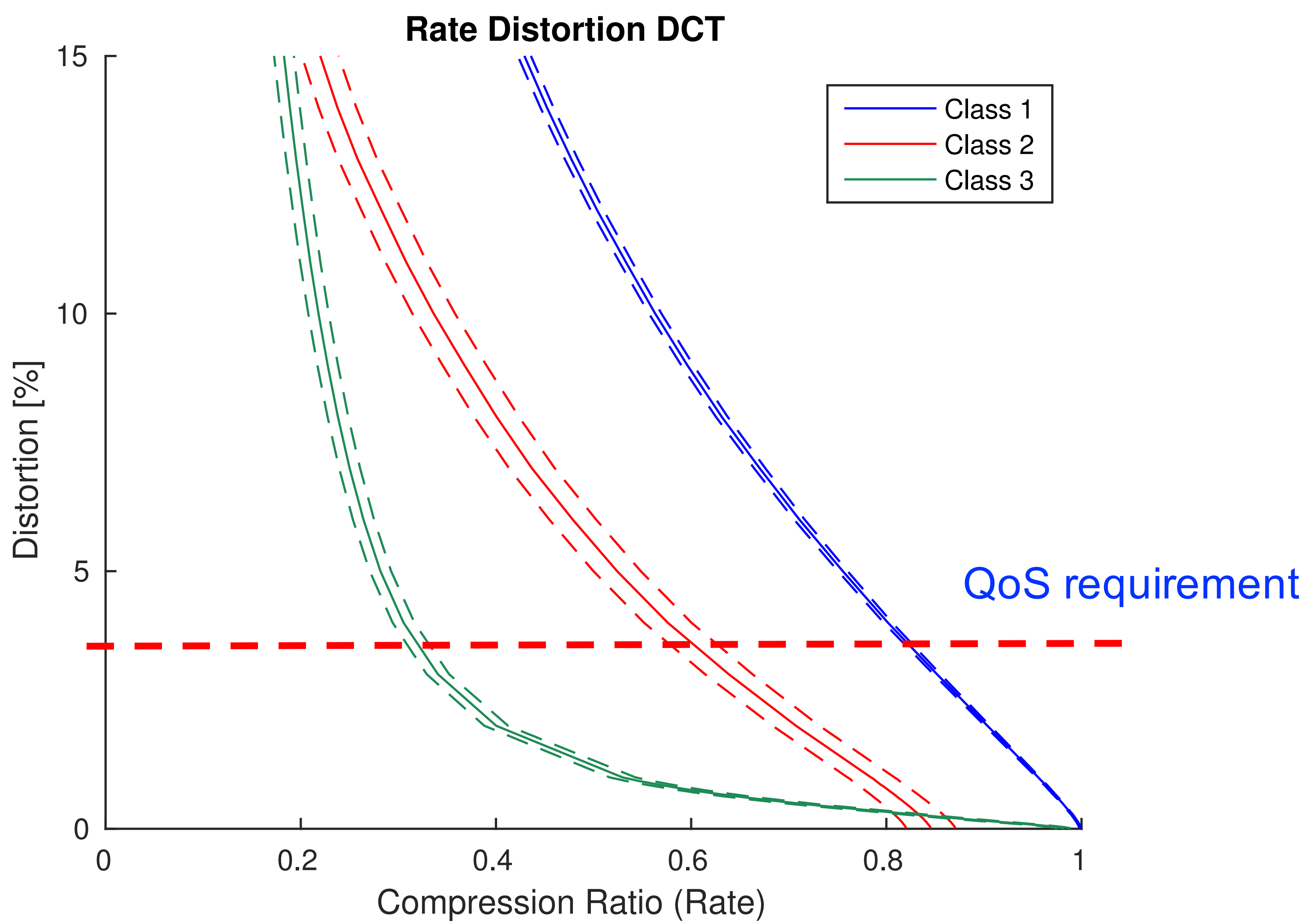
Compression Performance (2/3)



Compression Performance (3/3)



Discrete Cosine Transform Compression Performance



The Work Ahead

- Step1: classify based on compression performance
- Step2: check which features are most meaningful

- Supervised classification
  - SVM
  - Sensor networks

- Unsupervised classification
  - DBSCAN


thank you
danke
рахмат
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
спасибо
mersi
kia ora
barka
welalin
tack
spas
mahalo
tapadh leat
faafetai lava
vinaka
blagodaram
dank je
misaotra
matondo
paldies
grazzi
gracias
спасибі
kiitos
dankie
nanni
nandri
dhanyavad
hvala
maururu
köszönöm
bedankt
bayarlalaa
gràçie
akun
dankon
aĉiū
хвала
asante
manana
obrigada
murakoze
tenki
enkosi
djiere dieuf
tau
mochchakkeram
дякую
mamnun
chokrane
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
go raibh maith agat
sobodi
děkuji
sagolun
sukriya
kop khun krap
arigatô
takk
dakujem
trugarez
obrigado
mèsi
didi madloba
najis tuke
ありがとう
tanemirt
grazie
rahmet
kam sah hamnida
rahmat
terima kasih
diolch
dhanyavadagalu
shukriya
mercê
мерси
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
merci
ευχαριστώ

DATA MINING IN THE IOT ERA: PRACTICAL EXAMPLES AND A PEEK INTO FUTURE DEVELOPMENTS

Michele Rossi

rossi@dei.unipd.it

Dept. of Information Engineering

University of Padova, IT

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE