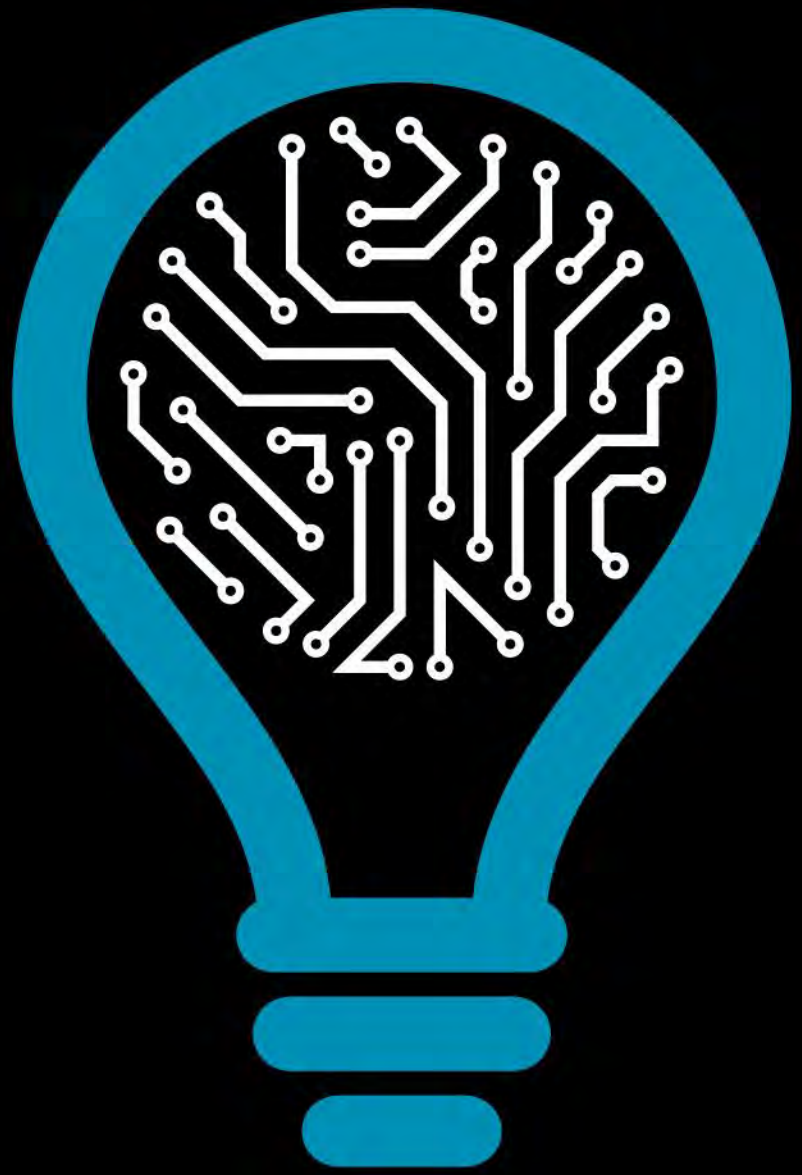
EXCESS OF DATA, LACK OF MODELS

the drive for a cognitive perspective in power systems

Vladimiro Miranda nov 2017



INESCTEC
INSTITUTE FOR SYSTEMS AND COMPUTER ENGINEERING, TECHNOLOGY AND SCIENCE

The present day challenges derive from a flood of data

Past

A growingly complex system with clearly defined borders between layers

A hierarchical paradigm

- Lack of data
- Many models

Models derived from theories, theories derived from knowledge

Present

A growingly complex system with generalized and distributed functions cross-grid

A blended distributed paradigm

- Excess of data
- Few models

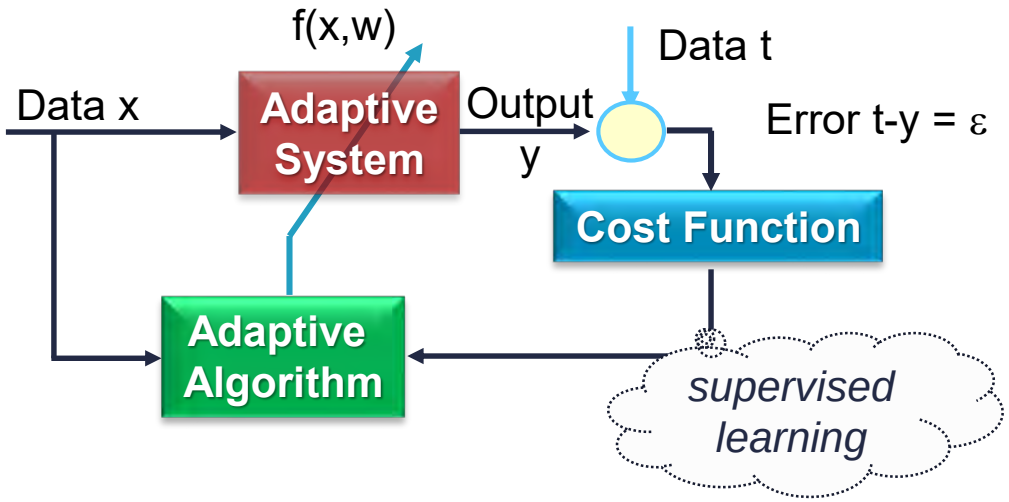
Knowledge derives from models, models derive from data

Learning from real data becomes more important than ever

There are multiple approaches to deal with data mining, but for physical systems the *model learning* approach is still the oldest and one of the most efficient.

Need to decide THREE things:

the mapping function
the cost function, and
the adaptive algorithm.



Because of the complexity, architectures and representation also became important.

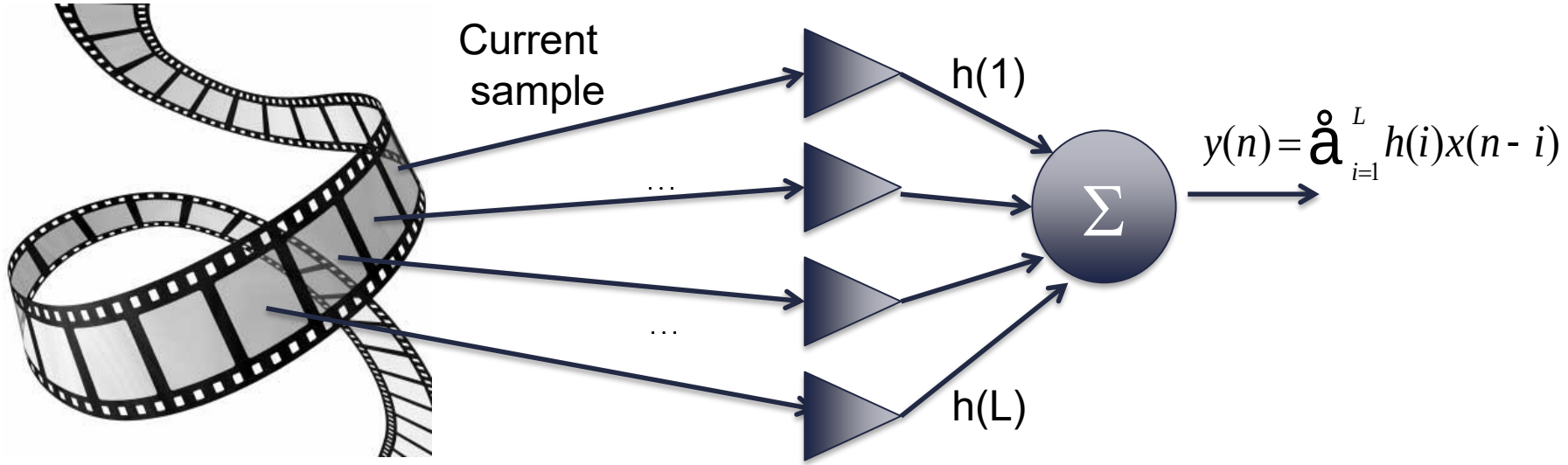
Basically, three processes can be implemented to build a model from data:

- Supervised Learning, Reinforcement Learning and Unsupervised Learning

Courtesy: J.Principe, CNEL

Mapping Function - Feedforward Topology

- We keep using linear feedforward models, the finite impulse response (FIR) filter
- FIR filter, combinatorial model, no context/memory: static mapping.
- But optimization is linear in the weights



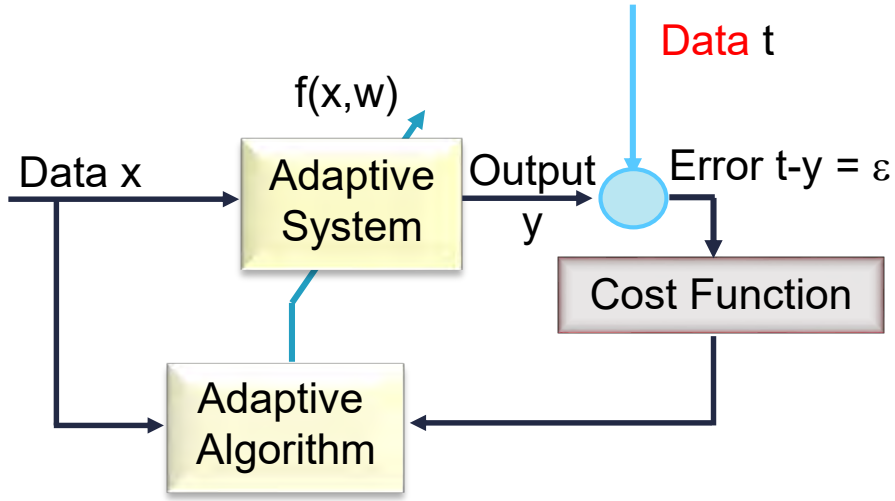
A classical performance criterion is MSE – but is it a good one ????

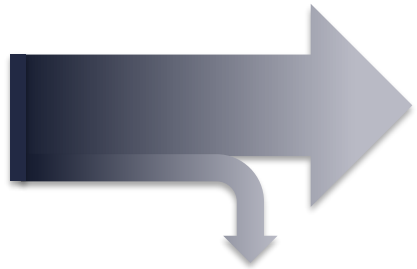
Understanding the cost function under an information theoretic perspective

Information Filters: Given data pairs $\{x_i, t_i\}$

- Optimal Adaptive systems
- Information measures

The information is embedded in the weights of the adaptive system.

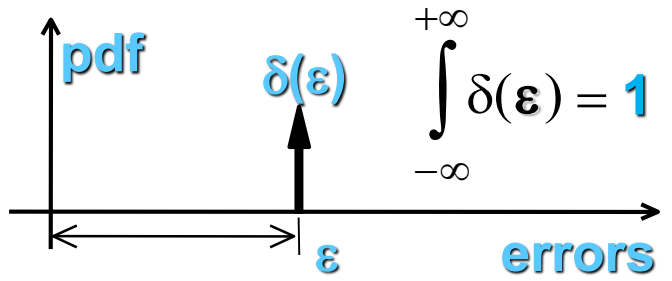




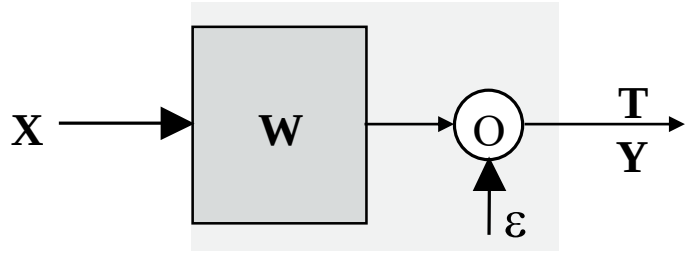
A learning process puts the real world in the adaptive system (as much as possible)

A basic idea: the Dirac function has minimum entropy

The ideal case in model adjustment is when the pdf of errors ε is a Dirac function - all errors equal.



With all errors equal, one could have perfect matching between output Y and target T, by adding a bias to the output neuron.



Minimizing the Entropy of the distribution of errors is a good idea.

Renyi's quadratic entropy is an information measure that leads to convenient models that can be computed.

$$H_{R2} = -\log \sum_{k=1}^{N} p_k^2$$

$$H_{R2} = -\log \int_{-\infty}^{+\infty} f_Y{}^2(z)dz$$

Introducing Correntropy

The Correntropy function may measure how similar (probability) two random variables X and Y are – or how close to zero (0) the difference ε = X-Y is.

The true joint distribution is usually unknown, so an estimate (via Parzen windows) may be obtained from a sample of size N

$$\hat{V}(X,Y) = \frac{1}{N}\sum_{i=1}^{N} G(x_i - y_i, \sigma^2\mathbf{I}) = \frac{1}{N}\sum_{i=1}^{N} G(\varepsilon_i, \sigma^2\mathbf{I}) = \hat{V}(\varepsilon)$$

where $G(.,\sigma^2 I)$ is a Gaussian kernel with variance σ^2.

Maximizing Correntropy tends to the same result as minimizing Entropy while remaining at an average error of zero:

→ Maximum Correntropy Criterion, or MCC

$$max \sum_{i=1}^{m} e^{-\frac{\varepsilon_i^2}{2\sigma_i^2}}$$

Least Squares Regression (MSE – Minimum Square Error) is sensitive to large or gross errors

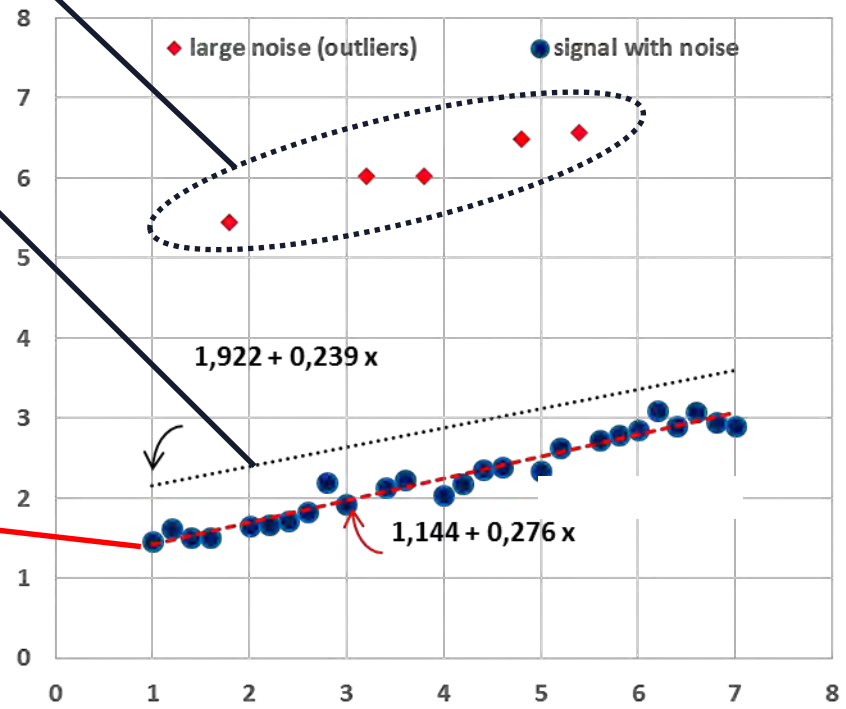
Assume that data is generated by a linear model *f(x)* corrupted by noise z(x)
- a mixture of Gaussians, for example.

$$g(x) = f(x) + z(x) = 1 + 0.3x + z(x)$$

Outliers

This is the regression line with Least Squares – influenced by the presence of outliers

This is the model one would like to estimate - *f(x)*



$$min \sum_{k=1}^{n} \left(g_k - \hat{f}_k(x)\right)^2$$

Regression with outliers: the advantage of Correntropy

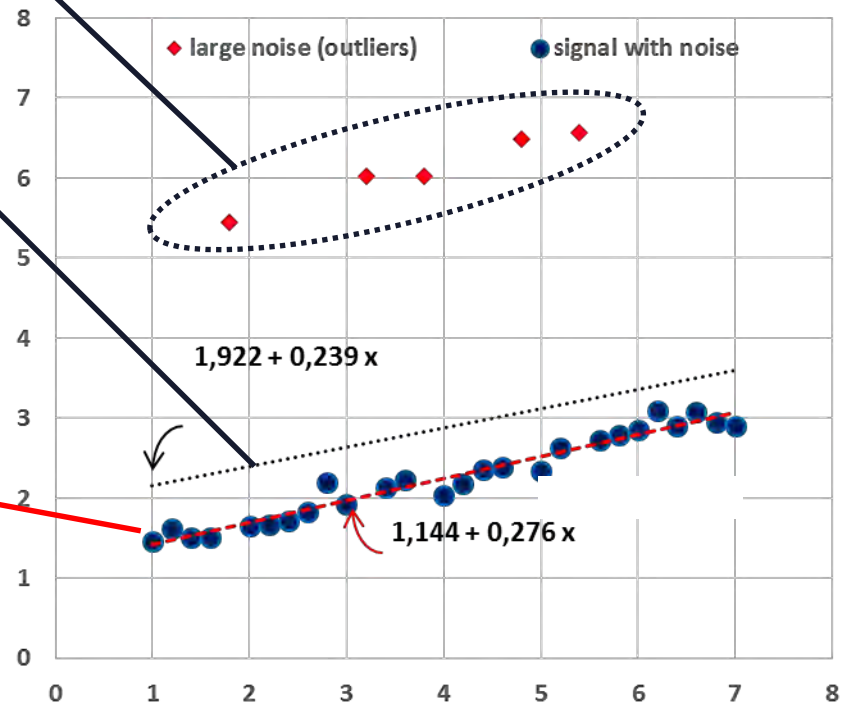
Assume that data is generated by a linear model *f(x)* corrupted by noise z(x)
- a mixture of Gaussians, for example.

$$g(x) = f(x) + z(x) = 1 + 0.3x + z(x)$$

Outliers

This is the regression line with Least Squares – influenced by the presence of outliers

Regression line with MCC and σ = 0.5, ignoring the presence of outliers – very close to the original *f(x)*



$$min \sum_{k=1}^{n} \left(g_k - \hat{f}_k(x)\right)^2$$

$$max \sum_{i=1}^{m} e^{-\frac{\varepsilon_i^2}{2\sigma_i^2}}$$

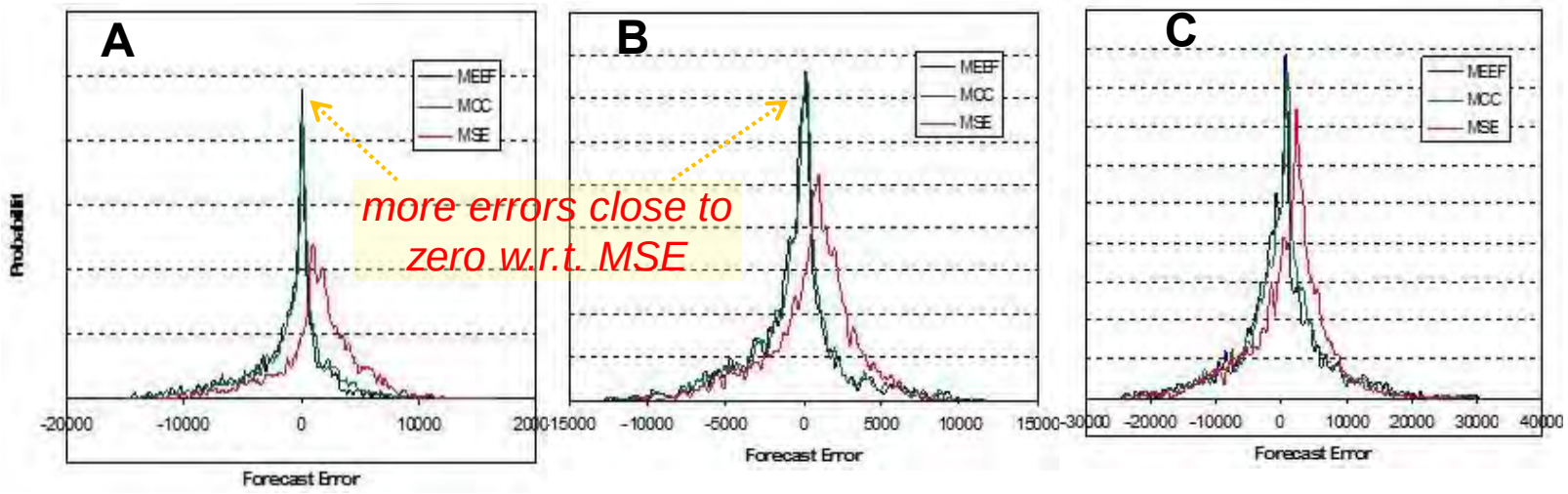
Advances in supervised learning derive from a new interpretation of the cost functions under the light of information theoretic concepts

Abandoning the pervasive *minimum square error (MSE)* criterion, which is an engineering construct, for criteria based on the information content of the pdf of errors. Examples of new cost functions:

MEE: Minimum Error Entropy

MCC: Maximum Correntropy Criterion

In wind power prediction (72 h ahead), predictors trained with INFORMATION measures lead to better results.



Why? Because forecasting the wind is a problem with non-Gaussian errors.

(joint work with J. Principe)

ARGUS PRIMA – Prediction Intelligent Machine *

INESC TEC software package made available:

- PostrgreSQL database
- NN library in C++, implementing several ITL criteria
- Kernel density forecast library in R
- Supporting codes in Python and R

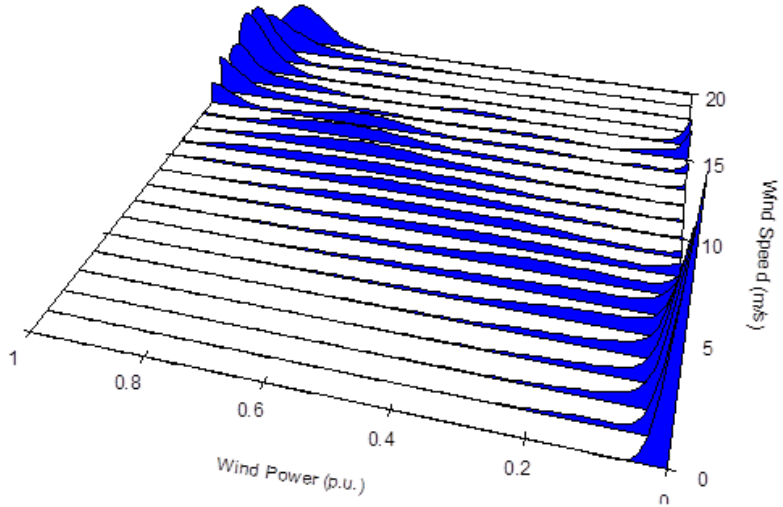
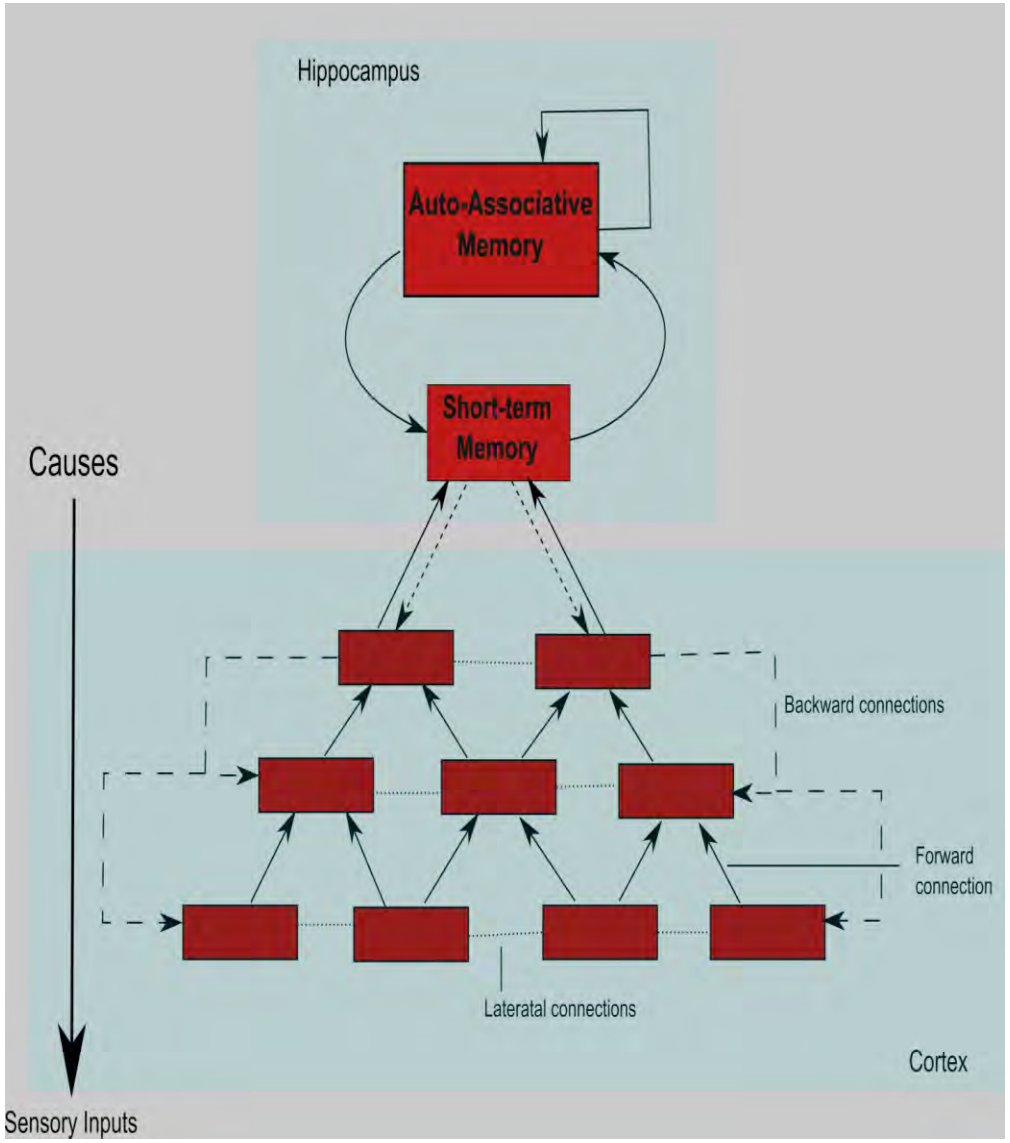


Illustration of uncertainty estimation with KDF (kernel density forecast): stacked conditional probability density function plot for wind power as a function of forecasted wind speed

* made available from ANL (USA) or through INESC TEC (Portugal)

http://www.anl.gov/technology/project/argus-prima-wind-power-prediction

Learning: an analogy to the visual cortex



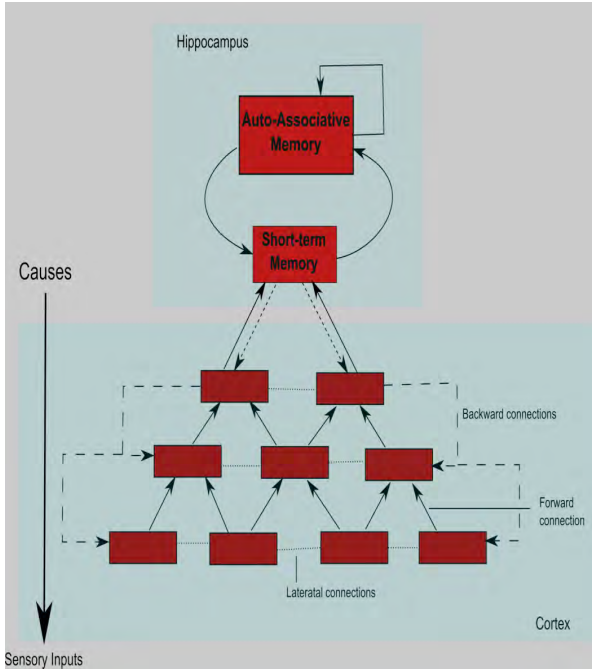
We share Helmholtz' view that cortical function evolved to explain sensory inputs.

As such, we seek to understand the role of processing and stored experience in a machine learning framework for the decoding of sensory input.

(J Principe)

Cognitive architecture for object recognition in video

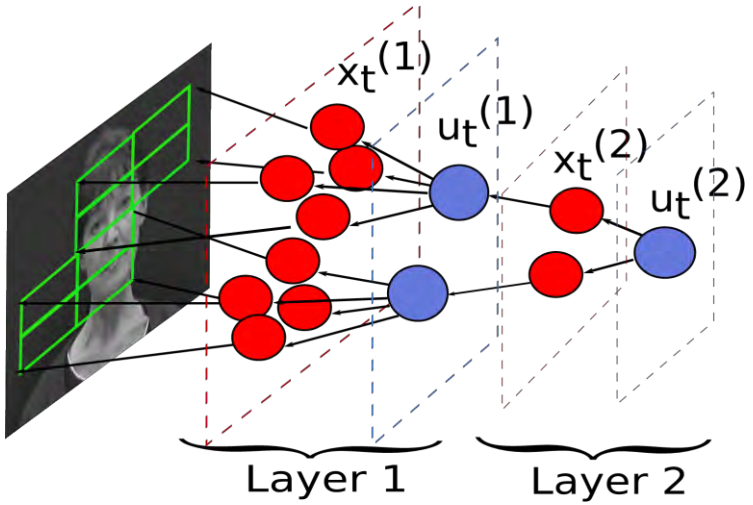
Goal → Develop a bidirectional, dynamical, adaptive, self - organizing, distributed and hierarchical model for sensory cortex processing using approximate Bayesian inference.



J Principe and R Chalasani, “Cognitive Architecture for Sensory Processing”, *Proceedings of the IEEE*, vol 102, #4, 514-525, 2014

A multi-layered architecture is needed for proper learning

- Tree structure with tiling of scene at bottom
- Computational model is uniform within layer and across



- Different spatial scales due to pooling which also slows the time scale in upper layers
- Learning is greedy (one layer at a time)
- This creates a Markov chain across layers

This reminds us of a first half of a deep autoencoder

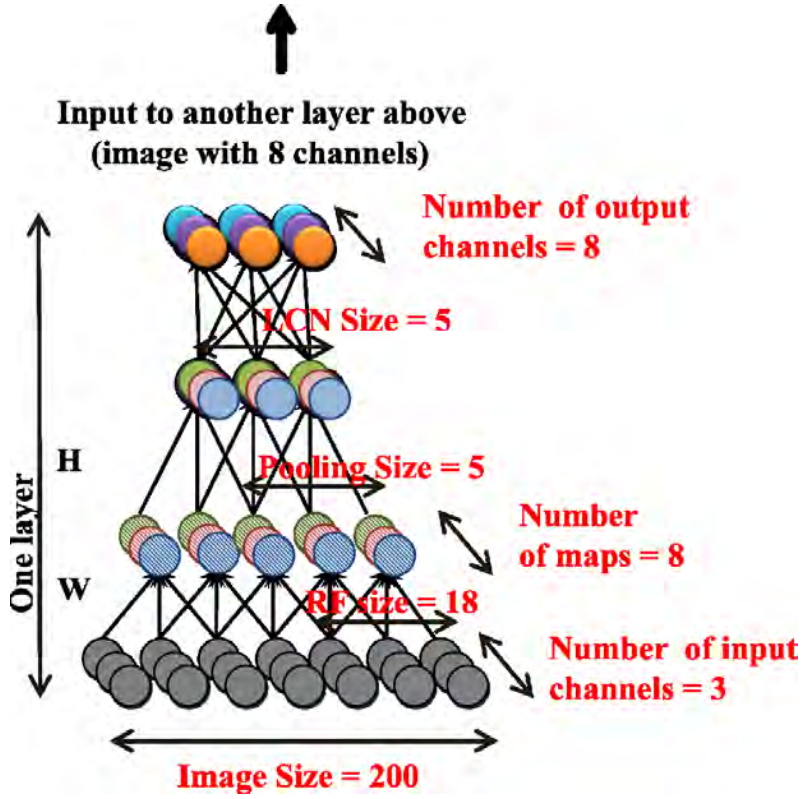
Courtesy: J.Principe, CNEL

Learning macro features with half autoencoders

Experiments with deep sparse networks have shown that macro-features could be learned from data

From 10 million image frames collected randomly from youtube videos, a half-autoencoder trained in unsupervised mode under an ICA criterion learned to identify macro-features such as faces and cats!

[Quoc et al, 2012]



Macro features exist!

Faces… bodies… cats…
Is there a neuron that captures the essential concept?

YES!

cat

body

OLA – Observatory of Latin America

PMU connected in low voltage:

- 22 universities in Brazil
- 3 universities in Chile
- 1 university in Argentina
- Data shipped to the Observatory in Florianópolis

Time-tagged information: like pixels in successive frames in a film

- The “film” captures the dynamics of the power system

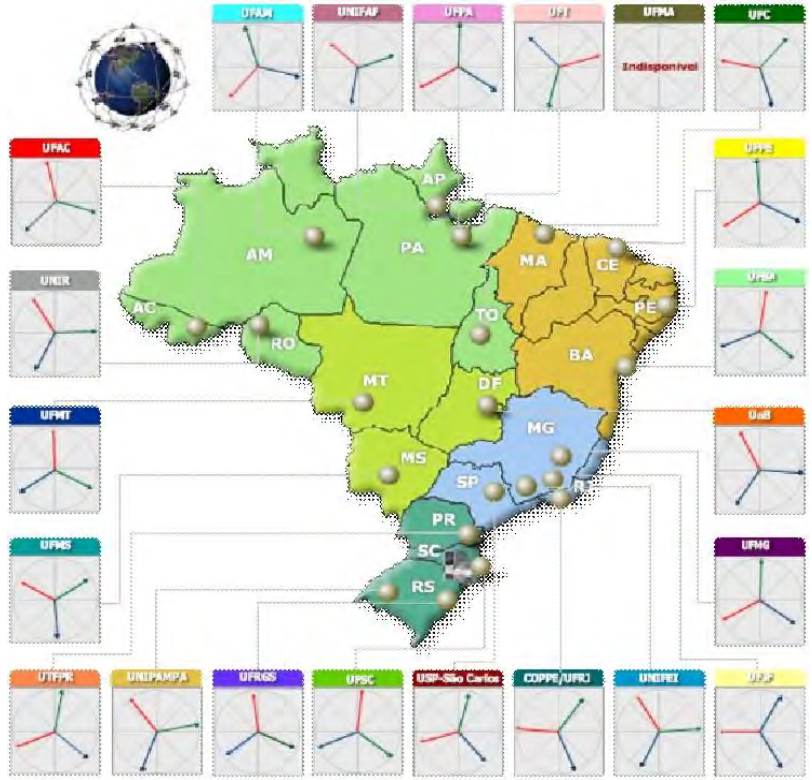
Can we discover invariants?

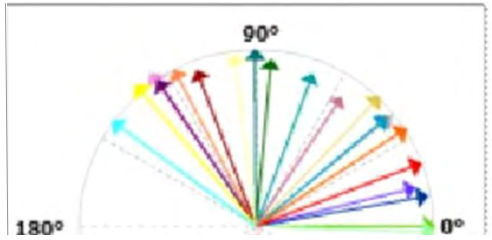
Can we build models from capturing knowledge,

instead of from *a priori* formulations?

This project (MedFasee) is on-going in Brazil.

http://www.medfasee.ufsc.br/temporeal/

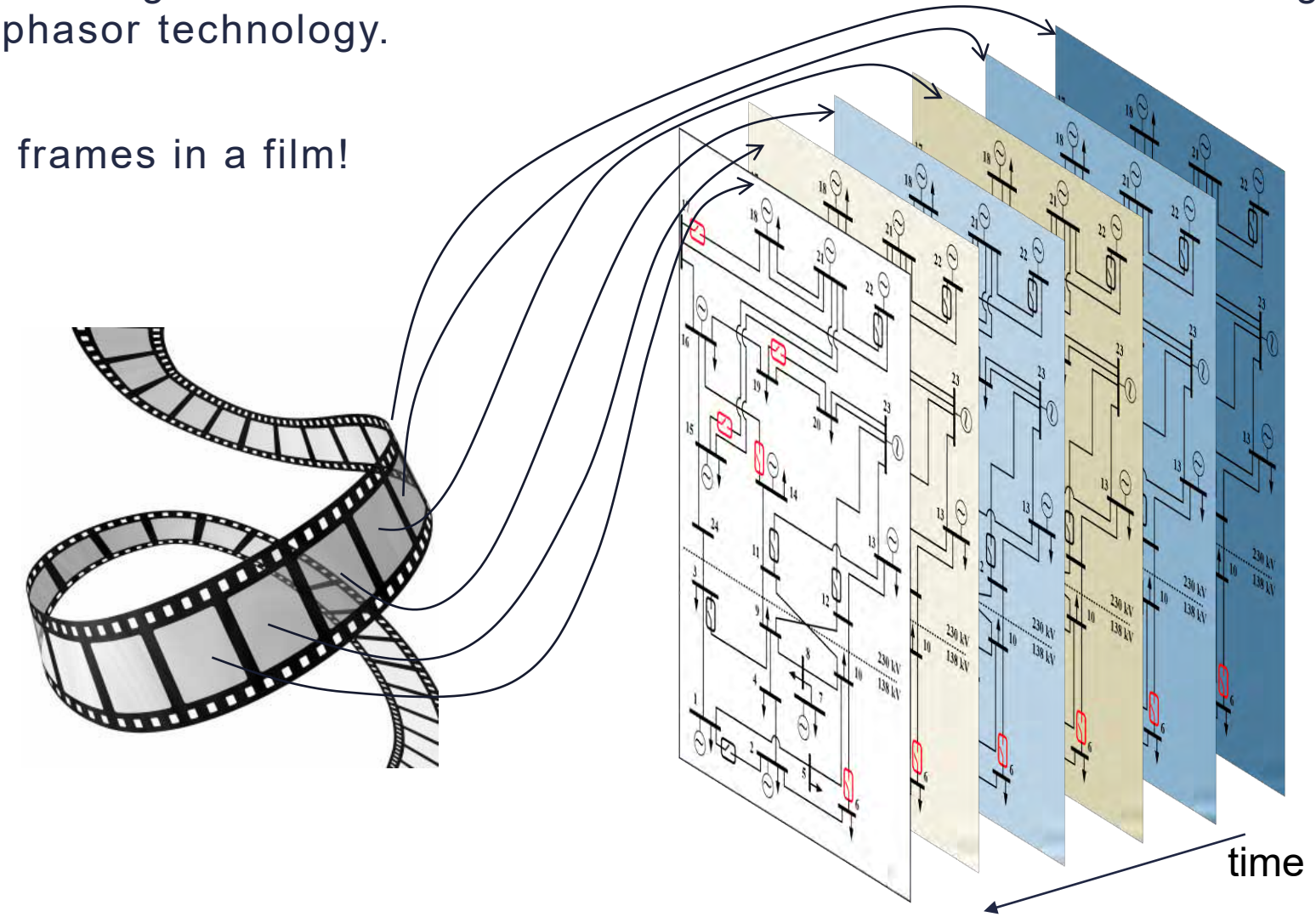




The continuous monitoring of a power system may now benefit from learning instead of relying on static models

All data arriving at the SCADA in a Control Center can become time-tagged with synchrophasor technology.

It is like frames in a film!



Discriminability in Sequence Labeling

Honda/UCSD face data set (20 for training, 39 for testing) using Viola Jones face finding algorithm (on 20x20 patches). Histogram equalization is done. 2 layer model $(16,48)_1$ $(64,100)_2$, 5x5 filters, causes concatenated as features.

Sequence Lengths / Methods	50 frames	100 frames	Full length	Average
MDA [Wang and Chen, 2009]	74.36	94.87	97.44	88.89
AHISD [Cevikalp and Triggs, 2010]	87.18	84.74	89.74	87.18
CHSID [Cevikalp and Triggs, 2010]	82.05	84.62	92.31	86.33
SANP [Hu et al., 2011]	84.62	92.31	**100**	92.31
DFRV [Chen et al., 2012b]	89.74	97.44	97.44	94.87
CDN w/o context	89.74	97.44	97.44	94.87
CDN with context	**92.31**	**100**	**100**	**97.43**

Convolutional Dynamic Models

If this can be done with images, why not with the power system?

Courtesy: J.Principe, CNEL

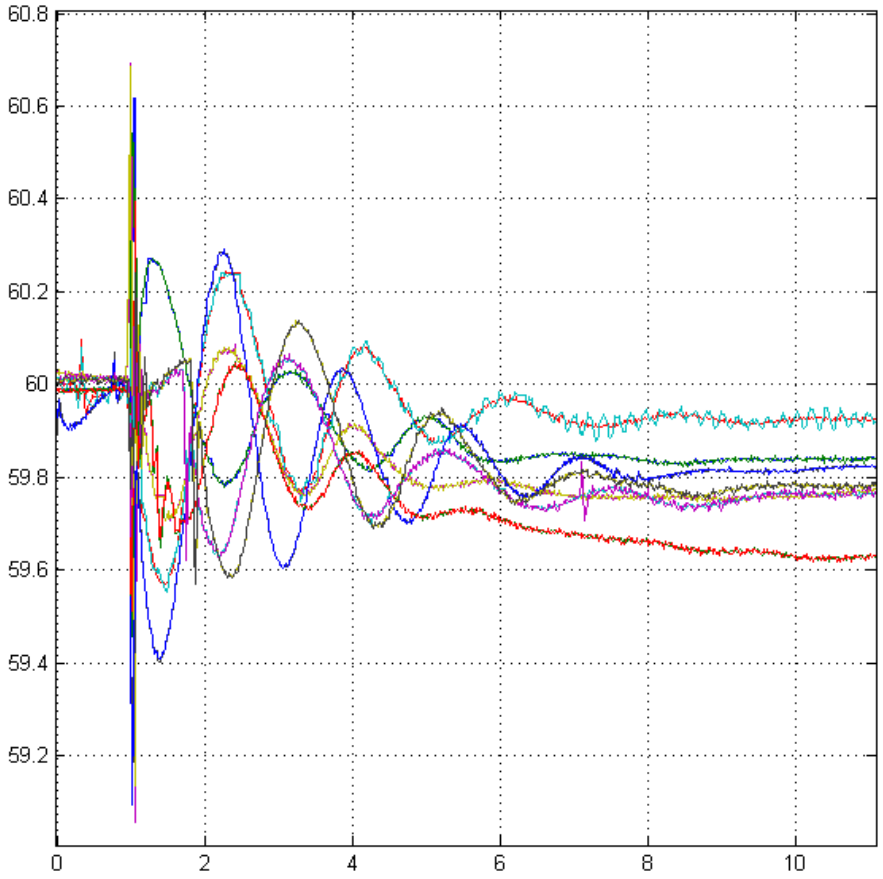
Events in the Brazilian power system

Events collected at several PMUs simultaneously

Frequency data collected at the rate of 1 measurement per 1/60 second

Examples of events stored in the time-tagged database, in 20 seconds long patches (1200 measurements per PMU) – 1 second before and 19 seconds after:

- Generator tripping
- Line tripping
- Load shedding
- Oscillation



Models tested: preliminary work (Decker)

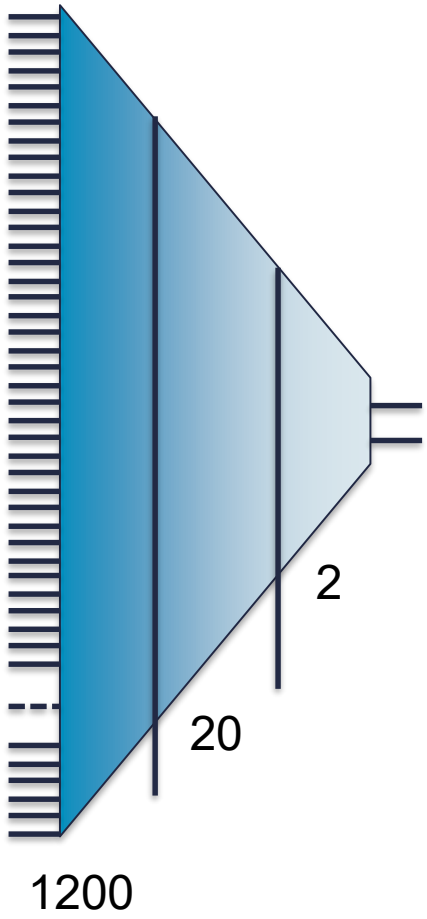
Feedforward ANN, 1 hidden layer

Output layer: 2 neurons

Trained with classical backpropagation

Target output patterns:

Generator loss	1
	0
Load disconnection	0
	1



Only able to 100% reliably distinguish these two events, when fed with data related to multiple events

(configurations with 3, 4 outputs and for other types of events not successful)

New models tested

Deeper feedforward ANN

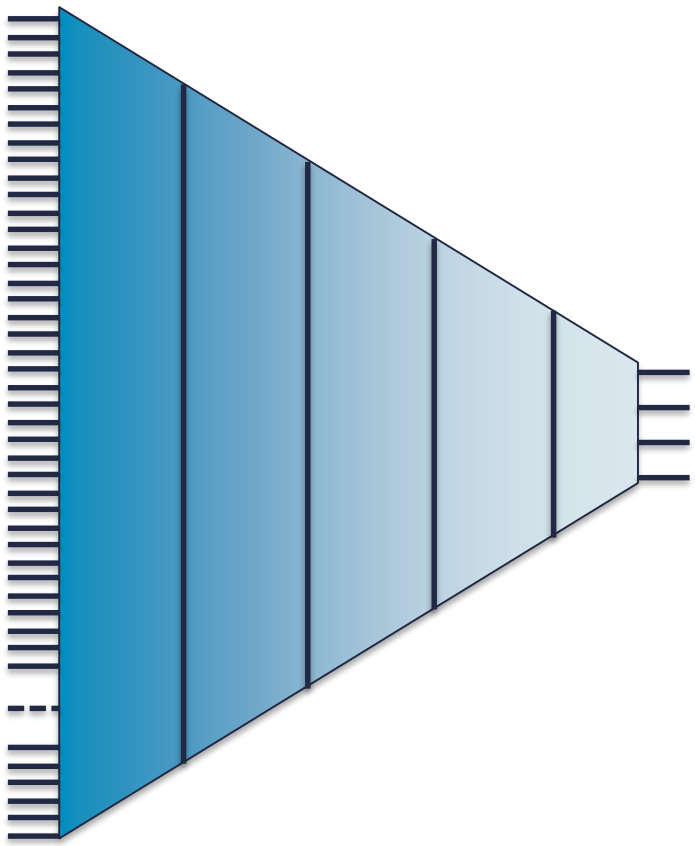
Deeper networks with 4 outputs (1 per event)

0
0
1
0

Two new architectures experimented:

3 hidden layers: 1200 – 500 – 200 – 100 – 4

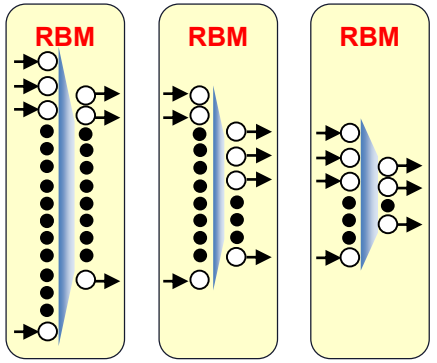
5 hidden layers: 1200 – 500 – 200 – 100 – 50 – 10 – 4

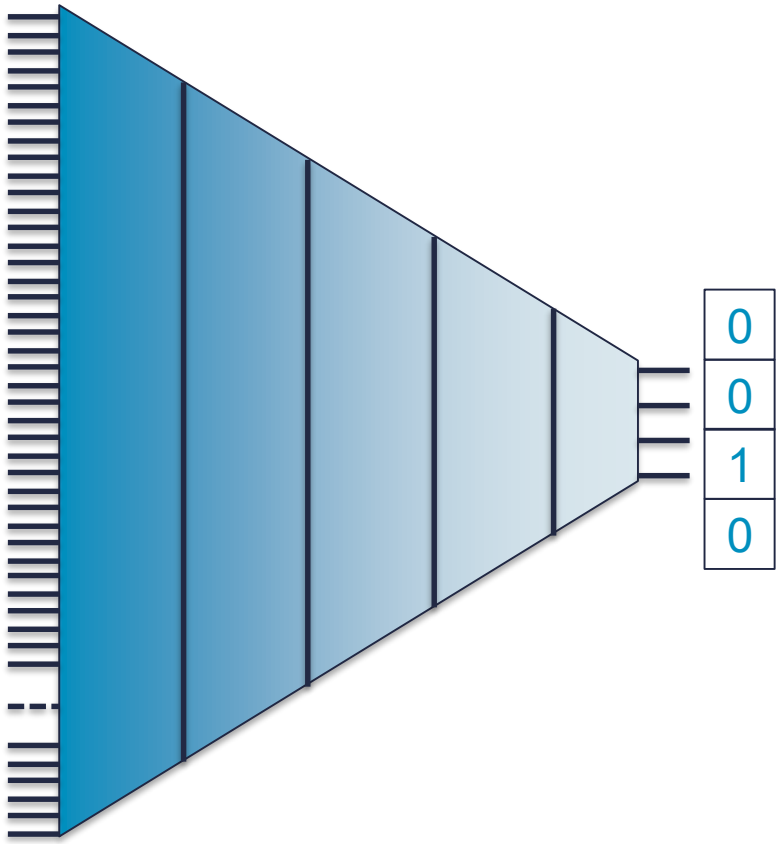


Deeper networks with 4 outputs (1 per event)

A Deep Belief network is trained layer by layer, in unsupervised mode, until the last layer which is trained in a supervised fashion.

The intermediate unsupervised training is supposed to organize and structure information prior to the identification exercise at the outer layer





3 hidden layers: 1200 – 500 – 200 – 100 – 4

New models tested

Convolution Neural Networks

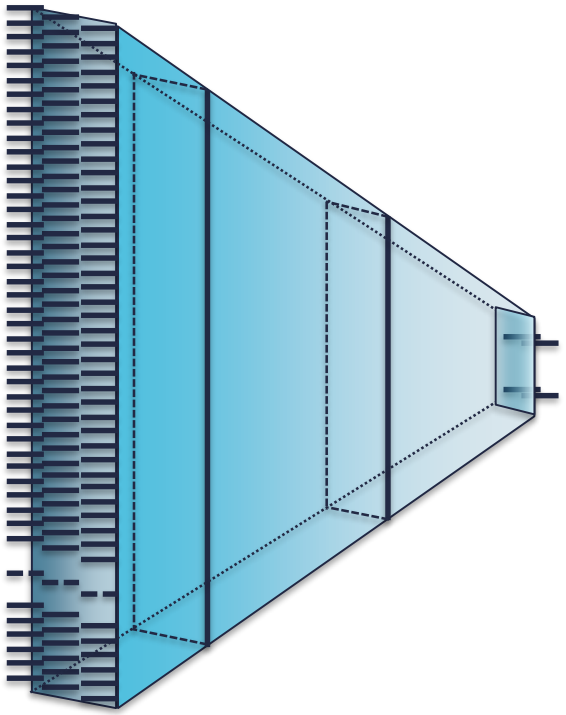
Deeper networks with 4 outputs (1 per event)

0	1
0	0

CNNs mimic the neural structure of the visual cortex

Experiments:

- frames 20 x 60 : 1200 – 400 – 4
- frames 30 x 40 : 1200 – 500 – 4
- inputs are treated as images!

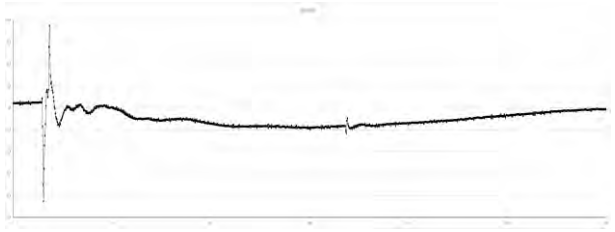


Events in the time domain
perceived as images

Generator tripping

Load shedding

Line tripping



Oscillation

The analogy with the visual cortex is powerful

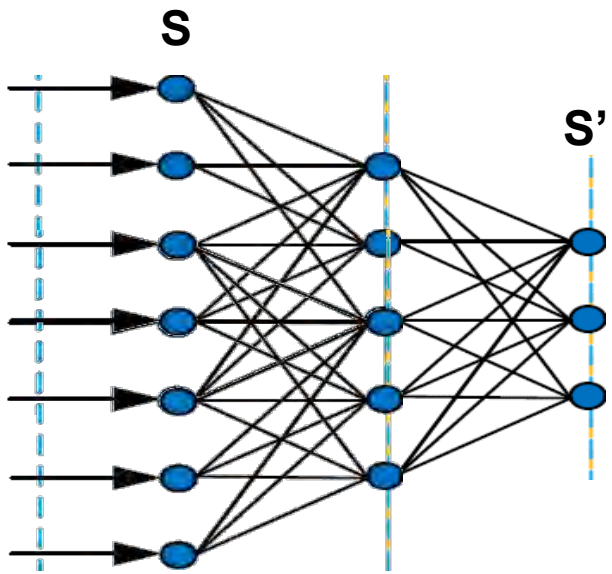
The convolutional neural networks, with an architecture inspired in the visual cortex, and with the input sequences organized as rectangular movie frames, were 100% successful in identifying the 4 events, from a pool of events presented to the network!

model	error in event recognition
Deeper Feedforward ANN	1,5 %
Deep Belief Networks	1,5 %
Convolutional Neural Networks 30x40	**0 %**

These promising results in recognizing events are inspiring!

Deep learning and knowledge discovery

A deep network projects a high dimension space S onto a space of reduced dimension S'.



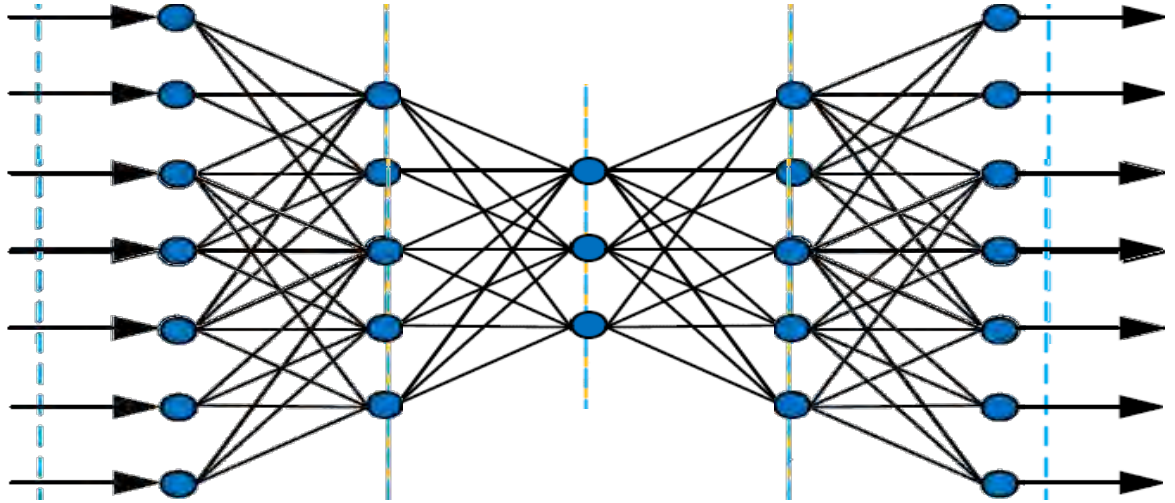
This compression, under a criterion to minimize information loss, keeps what is common among data: the knowledge of macro-features.

Deep learning attempts to uncover knowledge that is implicit yet hidden in data.

The secret is in blurring the details so that only common global features emerge!

From deep networks to autoencoders

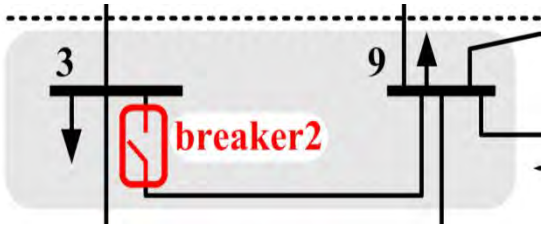
What happens, when one attaches a deep network to an inverse net?



The new network will re-project data onto the original space!

If such re-projection is trained in order to produce an output equal to the input vector (minimizing a function of the error), one gets an auto-associative neural network, or autoencoder.

Identifying breaker status



Is it possible to guess, based on local measurements, the status of a breaker inserted in a network?

CONJECTURE

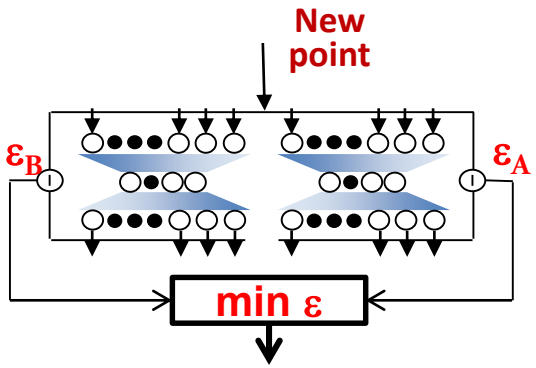
The topology information is hidden (or is diluted) in the values of the electric variables.

So, distinct topology states must become somehow reflected in distinct shapes of the manifold supporting the electric data.

How to unveil such information?

Hint: the breaker status is a macro-feature.

A competitive autoencoder architecture may help discriminating.



Example: topology diagnosis for 1 unknown breaker status

Only power, no voltage information.

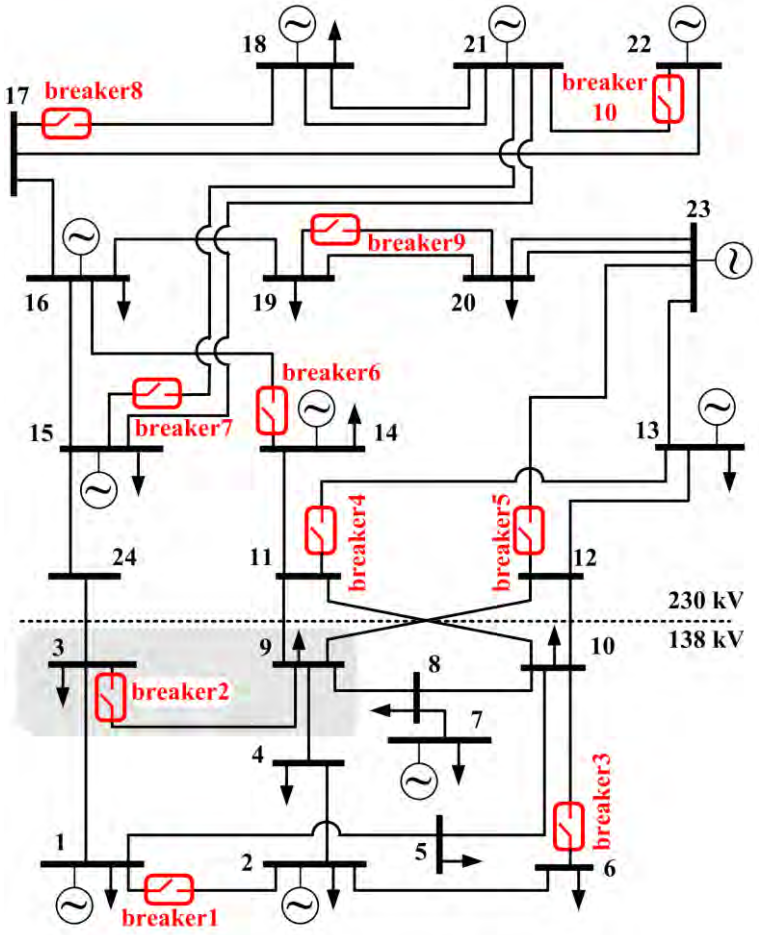
Autoencoders trained in 10.000 cases with random load and generation.

The diagnosis system tested in 10.000 new cases.

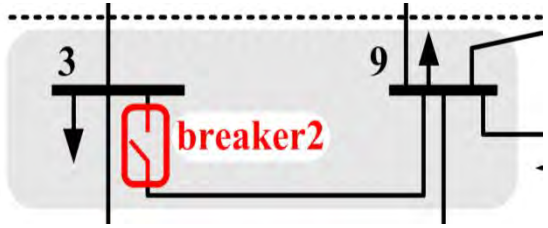
True Positive: br. closed, prediction: closed
False Positive: br. open, prediction: closed
False Negative: br. closed, prediction: open
True Negative: br. open, prediction: open

	TP	FP	FN	TN
output	4919	0	14	5067

Remarkable! Only 14 false negatives in 10,000!
Are they really wrong? Not exactly.



Identifying breaker status

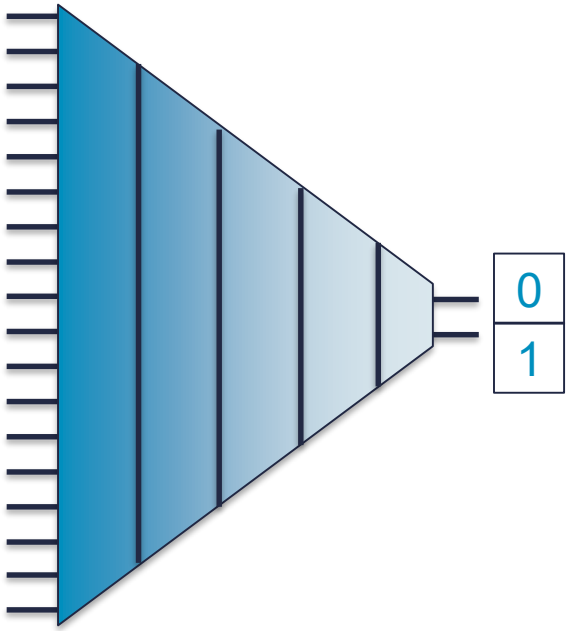


Test in 2346 cases

True Closed States	1230
False Closed States	0
True Open States	1115
False Open States	1
Total Accuracy	99,96 %
Closed State Accuracy	99,92 %
Open State Accuracy	100 %

A deep belief network

18 – 14 – 10 – 6 – 4 – 2



Training implemented in a GPU with clear acceleration

A hyerarchical hybrid model will work fine!

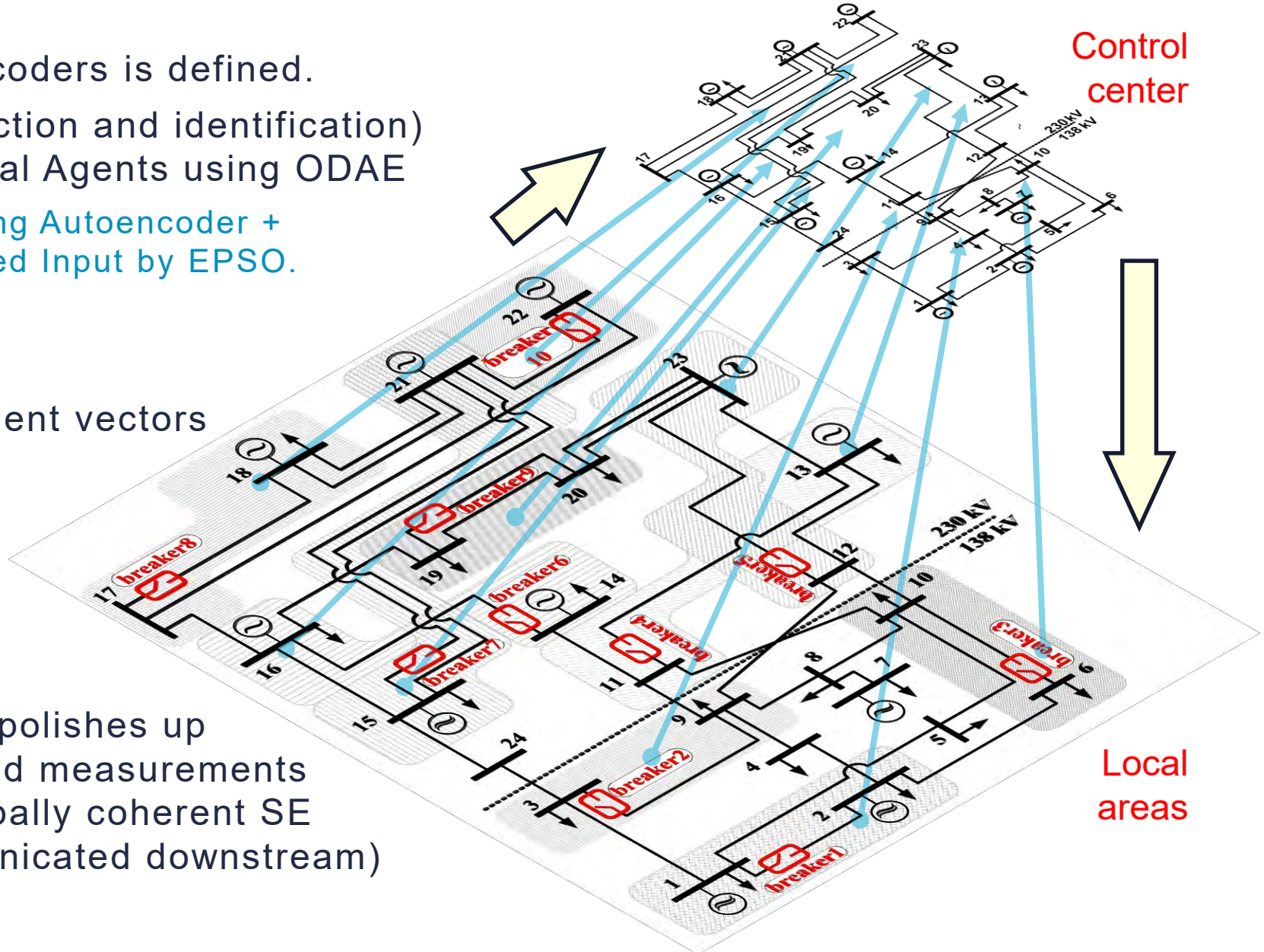
A mosaic of autoencoders is defined.

Local filtering (detection and identification) is performed by Local Agents using ODAE

ODAE – Denoising Autoencoder + Optimized Input by EPSO.

Repaired measurement vectors are communicated upstream.

An ITSE procedure polishes up the received repaired measurements and produces a globally coherent SE (that can be communicated downstream)



Hybrid MCC estimator performance

Tests ***in 10.000 cases*** with larger and larger networks confirm the excellent performance of the new method: the 2-level hybrid ODAE + ITSE architecture.

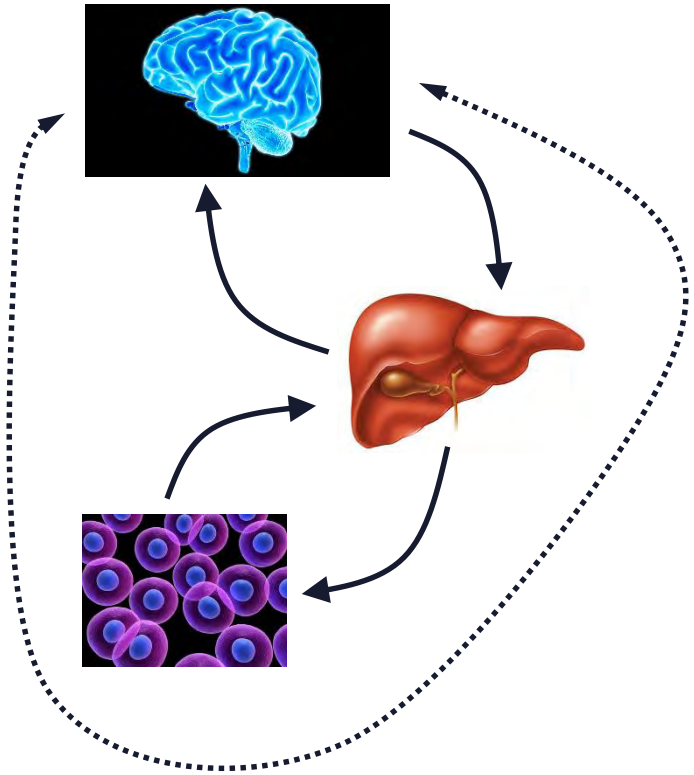
method	Efficiency w/ 2 GE	Mean time $[s]$ w/ 2 GE	Mean time $[s]$ w/o GE
WLS	-	0.1261	0.1239
LNRT	92.17%	0.5088	0.1652
ODAE	**100.00%**	1.3048	0.2071
WLS	-	**0.7219**	
LNRT	91.00%	2.8221	
ODAE	**100.00%**	3.9129	

IEEE 24-bus system

IEEE 118-bus system

Towards OOPS! – the organically organized power system

The biologic metaphor is growingly appropriate:
nested control systems and differenced missions
at distinct hierarchical levels.



Cognitive architectures will be useful for power system monitoring

- A power system is a spatial temporal process with many degrees of freedom.
- We design it piece by piece but have no idea if the current operating point is optimal, how resilient it is, when it is going to fail, etc.
- We simply overdesign it, keep it in steady state and plan the operation off line.
- The complexity of these large man made systems is created by the interactions among the parts – but when we analyze them we use the divide and conquer strategy that exactly throws away the interactions among the parts!
- The current data tsunami allows us to start thinking of building models that learn from data, have nested information exchange and control loops and match classical physical models with knowledge extracted from the real world – in the fashion of an organic system.

INESC TEC
R DR. ROBERTO FRIAS
4200-465 PORTO
PORTUGAL

T +351 222 094 000
F +351 222 094 050
info@inesctec.pt
www.inesctec.pt