EUROfusion

Machine learning for the Sciences: from Causality Detection to Data Driven Theory

Author: A. Murari and JET contributors

Many Thanks to PMU, JEU, TF leaders, Project leaders, Operator, Secondees and JET contributors, Associations and International Partners

Università di Roma Tor Vergata
Università degli Studi di Roma Tor Vergata
Quantum Electronic and Plasma Physics research group
Department of Industrial Engineering

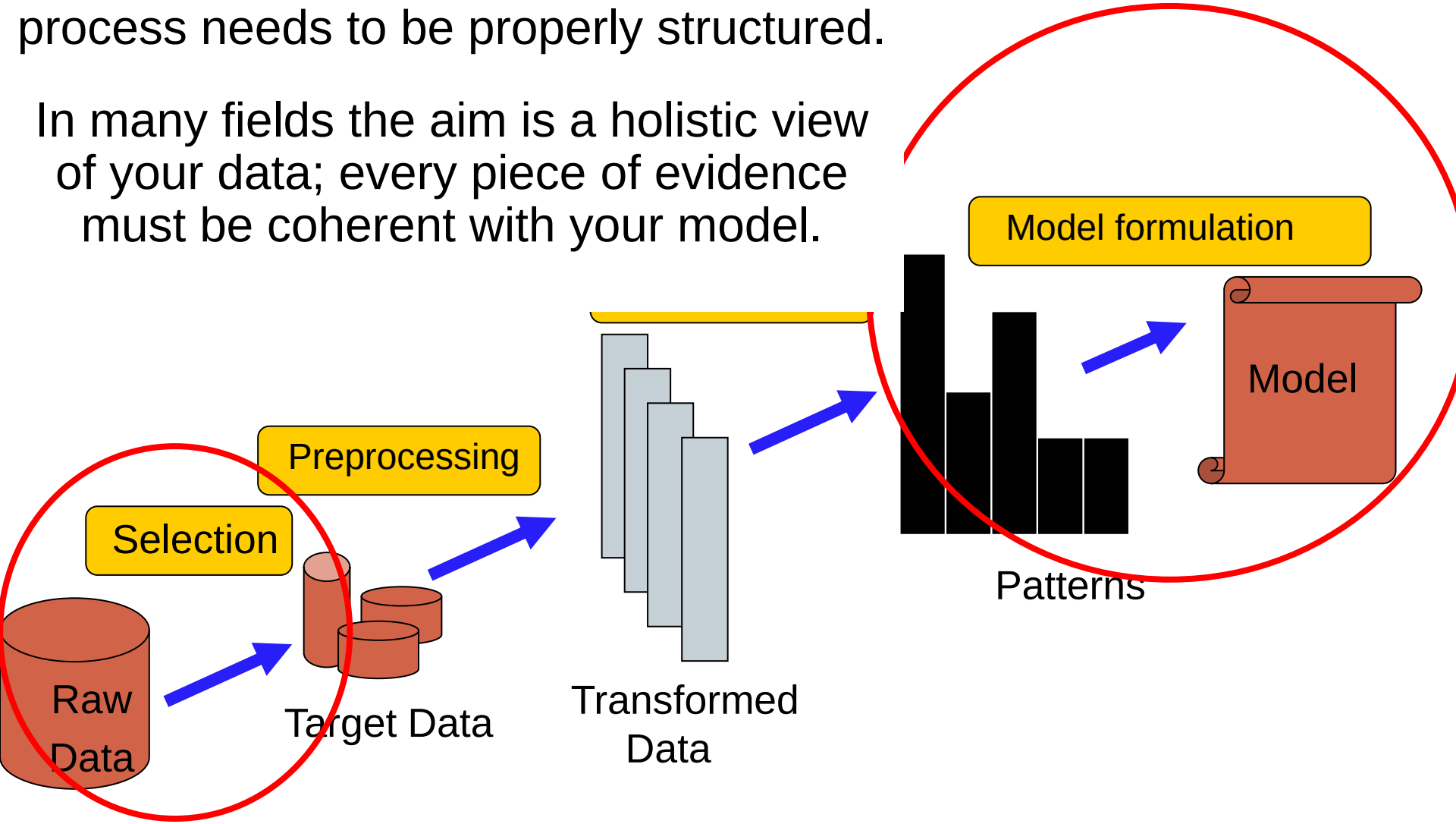
CONSORZIO RFX
Ricerca Formazione Innovazione

This work has been carried out within the framework of the EUROfusion Consortium and has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 633053.
The views and opinions expressed herein do not necessarily reflect those of the European Commission.

Data Analysis: an overview

Given the complexity of the problems and the amount of data, the inference process needs to be properly structured.

In many fields the aim is a holistic view of your data; every piece of evidence must be coherent with your model.

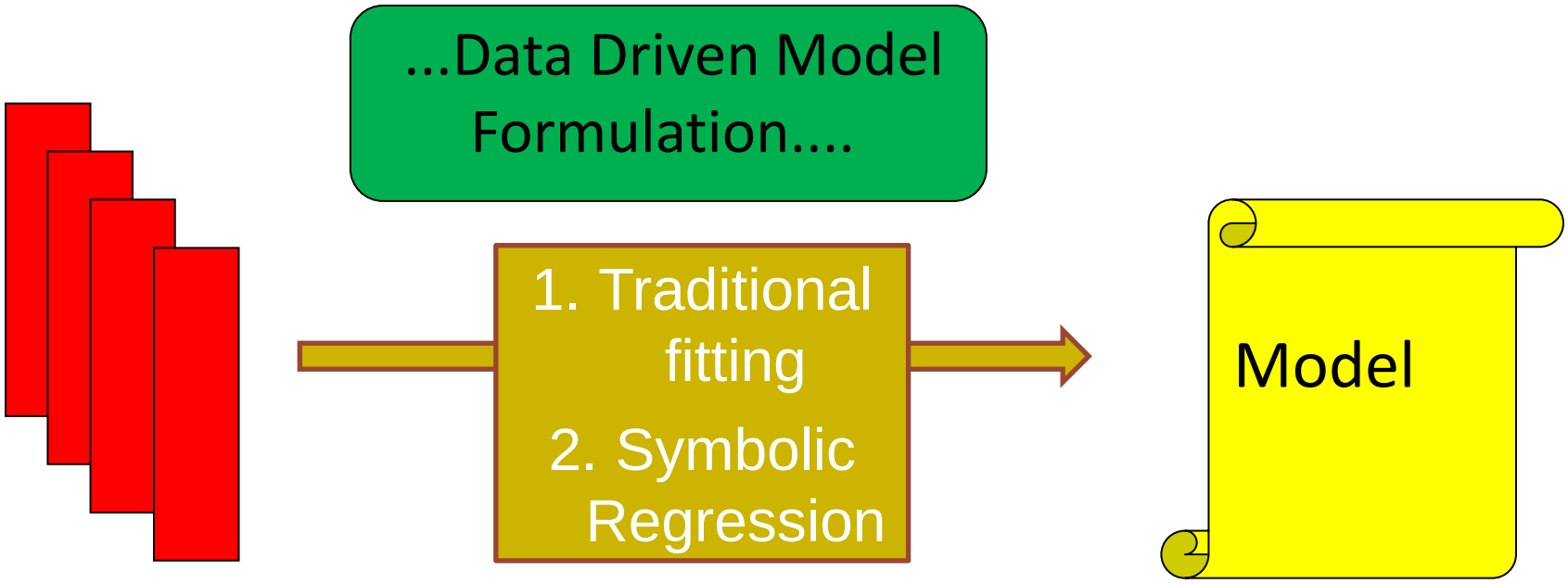


Outline

- Data driven Theory (longitudinal data):
  - Dimensionless variables
  - Scaling laws
- Observational Causality Detection for the analysis of dynamical systems (time series):
  - Causality detection based on manifolds: Cross Convergent Maps
  - Causality detection based on Information Theory: Transfer Entropy
  - Causality detection based on recurrence: Joint Recurrence Plots
- Conclusions

Model Formulation

Logical positioning of the technique



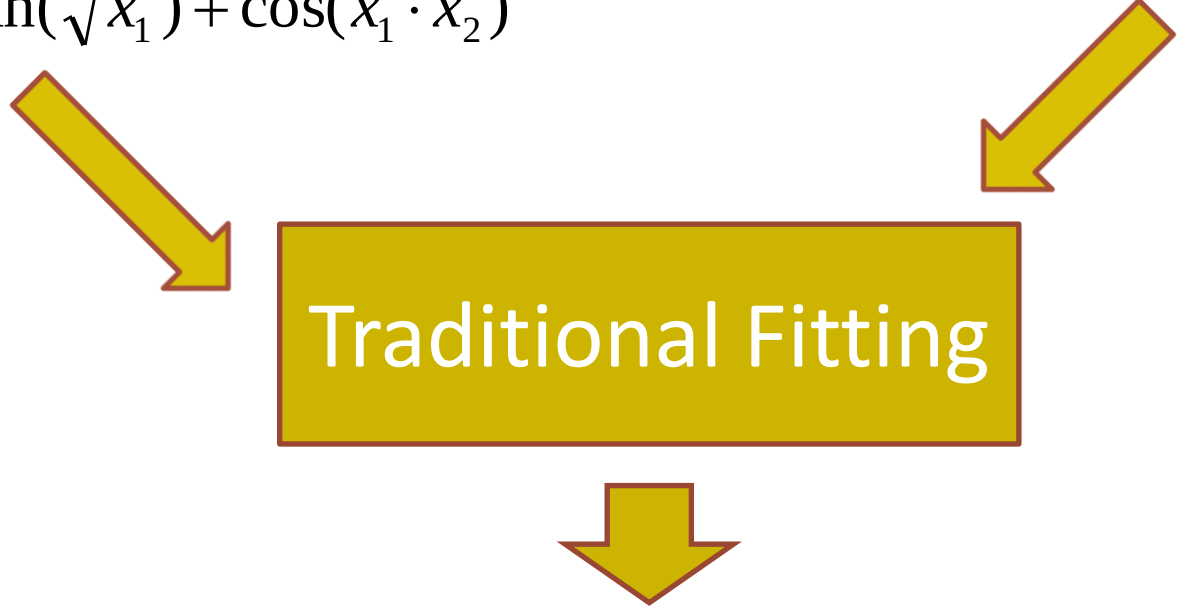
Available Validated Pre-processed Data (tens of Gbytes)

Traditional Fitting

A theoretical model of the independent physical quantity as a function of the regressors must be available.

$$y_{theoretical} = \sin(\sqrt{x_1}) + \cos(x_1 \cdot x_2)$$

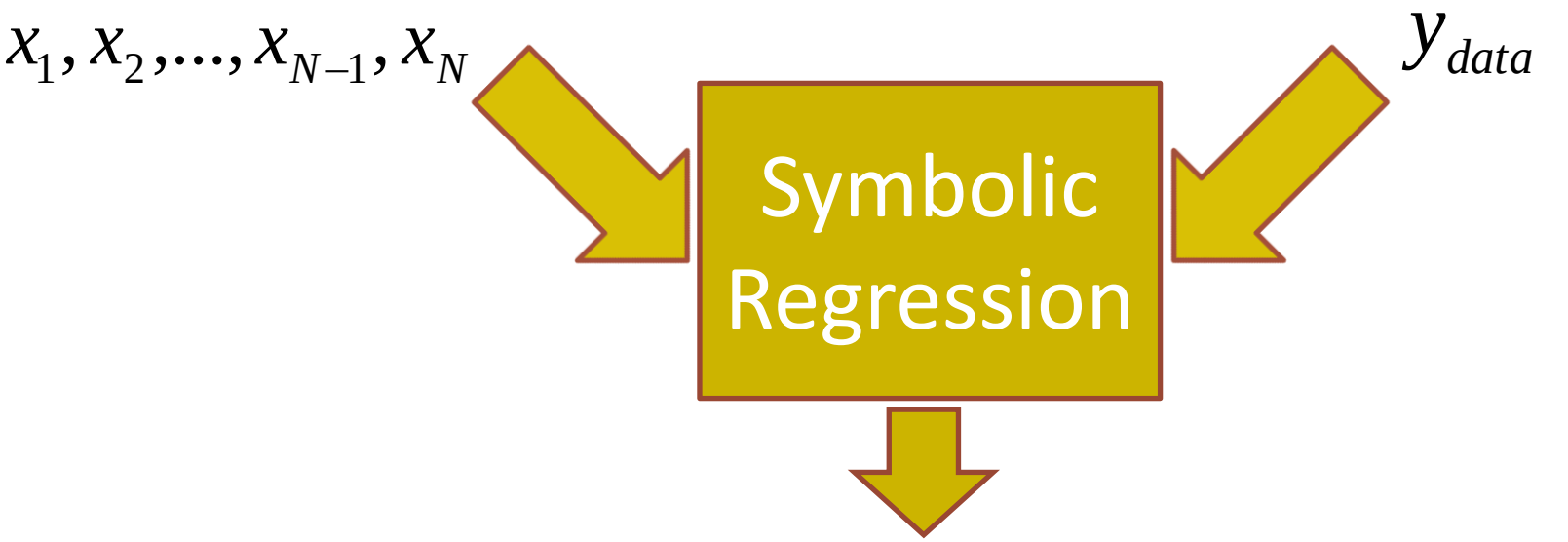
$$x_1, x_2, \ldots, x_{N-1}, x_N$$



$$y_{to\ be\ \ fitted} = \alpha_1 \cdot \sin(\alpha_2 \cdot x_1^{\alpha_3}) + \alpha_4 \cos(\alpha_5 \cdot x_1^{\alpha_6} \cdot x_2^{\alpha_7})$$

Symbolic Regression via Genetic Programming

- On the basis of the data available (selection of the dependent quantity and the regressors) the best mathematical model is provided by SR via GP



Genetic Programming for Symbolic Regression

- Standard procedure of SR via GP:

  1- Generate a random population of individuals (formulas).

  2- Evaluate each individual of the population (formula) with a fitness function (FF).

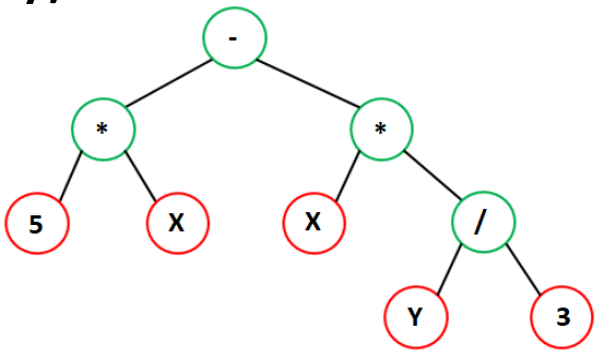
  3- Select the best fitting individuals (parents) to create a new population of individuals (formulas).

  4- Combine the genes (“crossover”) of the chosen parents and implement mutations, obtaining “children”.

  5- Repeat the steps 2 to 4 till an ending condition is fulfilled.

Overview of SR via GP

Formulas are represented as trees: 5x-xy/3.



Fitness Function: AIC & BIC

- Akaike Information Criterion (*AIC*):

$$AIC = 2\log MSE + 2k$$

- Bayesian Information Criterion (*BIC*):

$$BIC = 2\log MSE + k\log n$$

Penalty for models with a higher number of parameters

MSE ≡ Mean Square Error of the residuals,
the differences between
the data and the estimates of the model)

k ≡ number of parameters

n ≡ number of observations

- <u>The preferred model for AIC (BIC) criterion is the one with the minimum value of AIC (BIC)</u>

Identification of dimensionless quantities

Numerical exercises have been performed: synthetic data have been generated using dimensionless equations but only the dimensional quantities have been provided to SR, which has always been able to identify the original dimensionless quantities.

A well-known law connecting dimensionless quantities in fluid dynamics is

$$Pe = Pr \cdot Re$$

The Peclet number Pe is quantifies the ratio between transferred heat by advection and diffusion in a fluid. The Prandtl number Pr is defined as the ration between kinematic and thermal diffusivity; the Reynold number Re takes into account the relative importance of viscosity for internal layers of a fluid.

A noise level up to 30% of the data has been added ot the variables (with Gaussian distribution)

Equation identified by SR via GP: $Pe = 0.99 \cdot PrRe$

Energy Confinement Time

- The energy confinement time τ_E is an important parameter to determine the quality of a thermonuclear plasma

$$\tau_E = W/Pin$$

Where W is the internal energy and Pin the input power

- The transport phenomena leading to τ_E are too complex to model satisfactorily in realistic geometries
- In the last decades databases of empirical values of τ_E have been built collecting data from all the major Tokamaks in the world.
- The ITPA database is an international effort

Power law scalings

- Scaling laws are very important to determine how a system changes with some parameters
- Practical all the scaling laws in fusion are power law monomials:

$$Z = \gamma X^{\alpha} Y^{\beta} \ldots.$$

- Power laws are appropriate in case of self similar processes
- The tool used to extract these scalings from the DBs is log regression, which has been on the market for decades
- The limits of power laws are well known but overlooked: no saturation, interaction only multiplicative, vulnerable to collinearities, valid only in case of Gaussian noise etc

Energy Conf. Time: dimensionless quantities

- ITPA database used to derive the IP98 y2 scaling law
- A similar analysis has been performed for the dimensionless product between the confinement time τ_E and the ion Larmor gyro-frequency to obtain an actual independent scaling (no possible with log regression)

$$\omega \cdot \tau_{AdPL1} = 7.21 \cdot 10^{-8} \frac{M^{0.96} \varepsilon^{0.73} k_a^{3.3}}{\rho^{2.70} \beta^{0.90} \nu^{0.01} q^{3.0}}$$

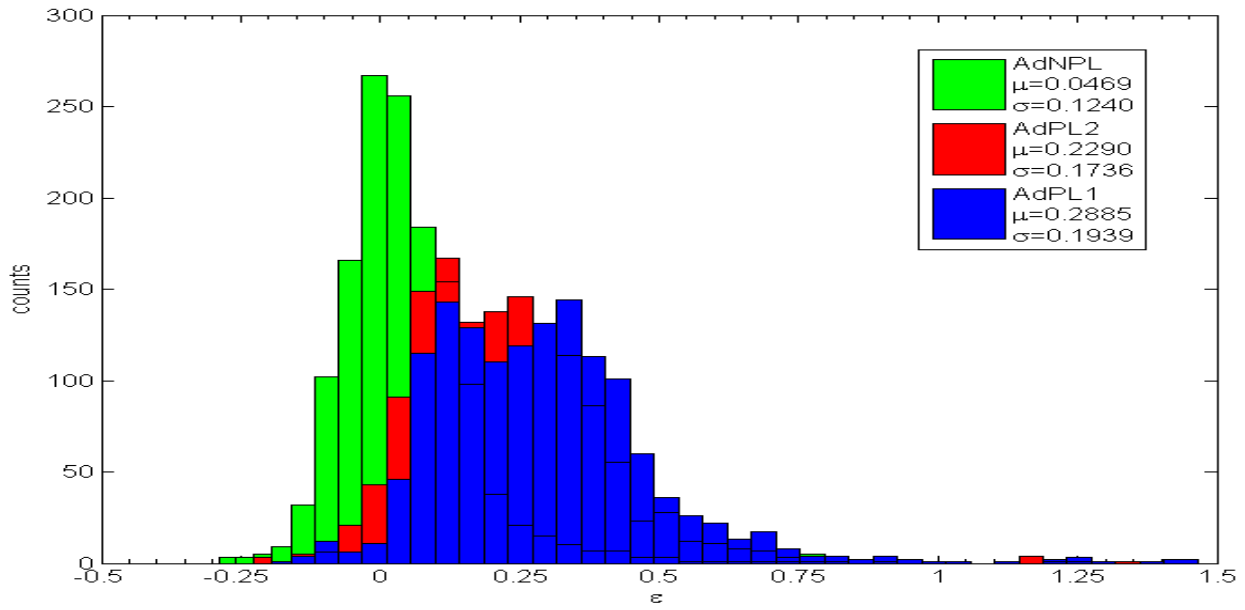
$$\omega \cdot \tau_{AdNPL} = (1.13)^{1.15}_{1.11} \cdot 10^{-6} \frac{k_a^{1.93^{2.12}_{1.70}} \beta^{0.37^{0.41}_{0.35}} M^{0.57^{0.67}_{0.46}}}{\rho^{2.19^{2.22}_{2.16}} \nu^{0.40^{0.42}_{0.39}} q^{0.16^{0.23}_{0.03}}} - (0.072)^{-0.060}_{-0.085} k_a^{1.18^{1.40}_{0.94}} +$$

$$-0.009^{-0.006}_{-0.011} q^{1.08^{1.21}_{0.94}} + 0.15^{-0.17}_{-0.13} M^{0.07^{0.19}_{-0.05}}$$

	AIC	BIC	MSE	KLD
ipb98y2-> AdPL1	-1650.59	-2533.00	0.55	0.33
AdNPL	-13833.00	-13758.91	0.0072	0.056

τ_E: extrapolation to JET

To substantiate the extrapolability of the non power law scalings, the various scalings have been obtained for the small devices and the histograms of the residuals have been calculated for JET

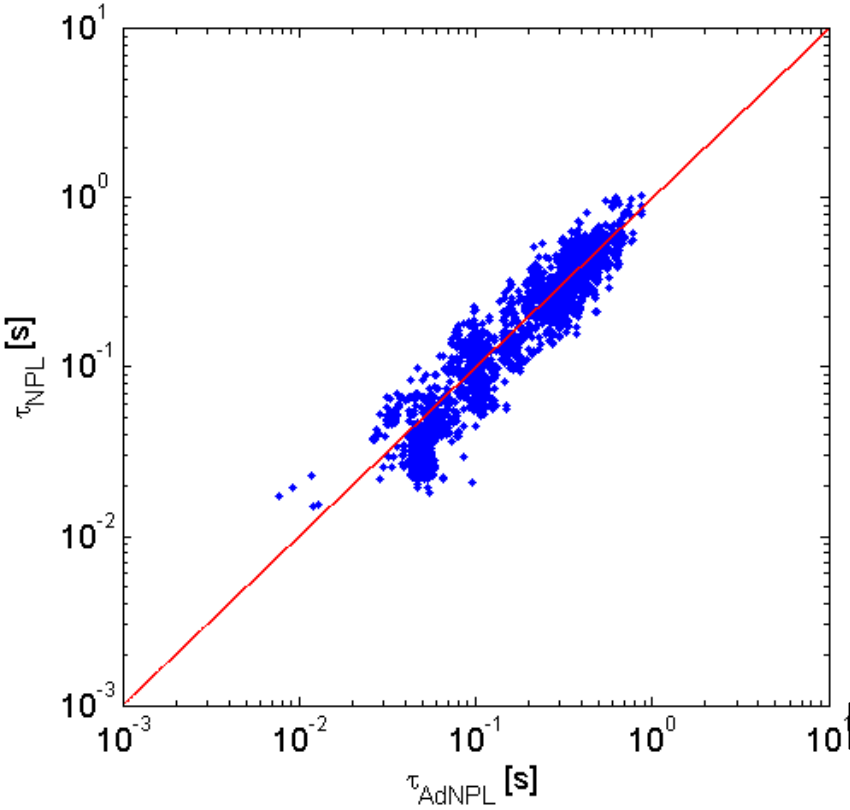


	k	AIC	BIC	MSE	KLD
AdPL1	9	−2930.77	−4505.82.	$12.078 \cdot 10^{-2}$	8.2048
AdPL2	9	−3461.54	−4813.36.	$8.255 \cdot 10^{-2}$	3.8786
AdNPL	14	−5610.85	−5723.52	$1.756 \cdot 10^{-2}$	0.9758

Scalings with dimensional and dimensionless regressors

Indipendent Scalings with dimensional and dimensionless regressors: very good agreement

Excellent independent match and ~20% reduction



Extrapolation to ITER

Equation	τ [s]
AdNPL	$2.97^{3.16}_{2.78}$
NPL	$2.83^{3.31}_{2.42}$
Power laws	3.66

Hundreds of thousands of models have been tested. Agreement between predictions of dimensional and dimensionless scalings

Outline

- Data driven Theory (longitudinal data):
  - Dimensionless variables
  - Scaling laws

- Observational Causality Detection for the analysis of dynamical systems (time series):
  - Causality detection based on manifolds: Cross Convergent Maps
  - Causality detection based on Information Theory: Transfer Entropy
  - Causality detection based on recurrence: Joint Recurrence Plots

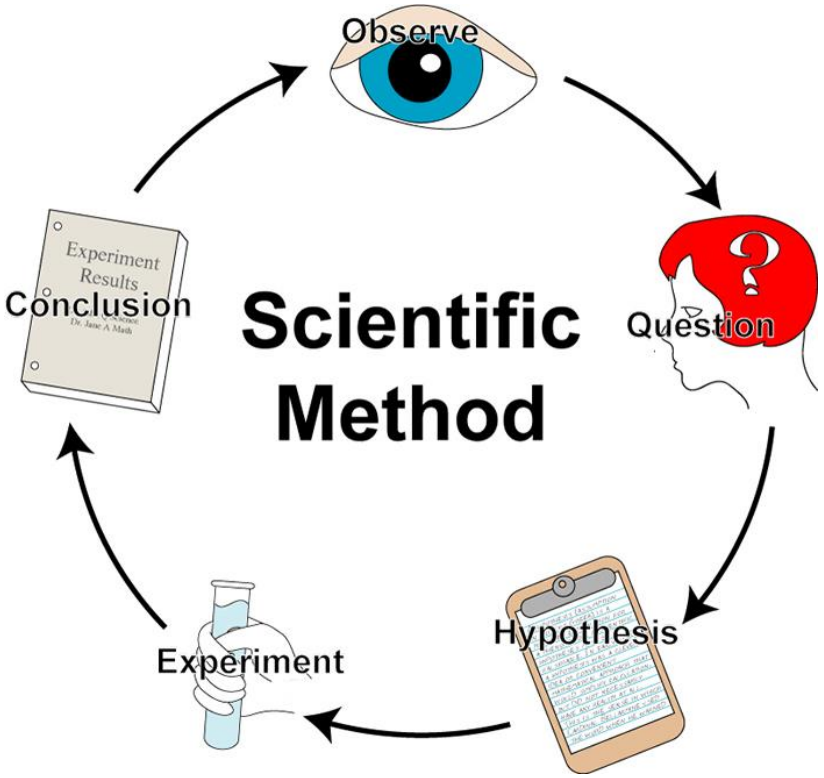
- Conclusions

Observational Causality Detection

- Causality should inform the entire scientific process.

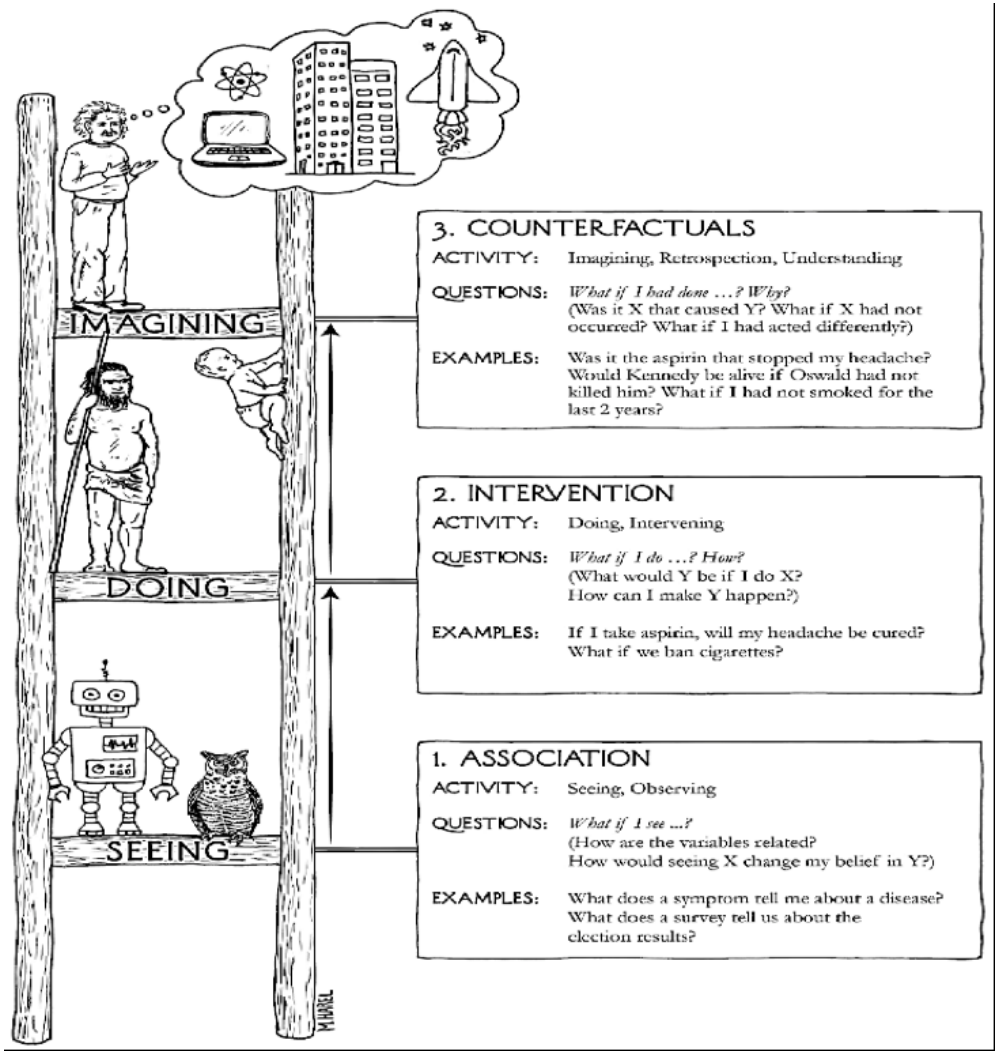
- Observational Causality Detection is meant to derive as much information as possible about the causal relation between quantities from the analysis of data.

- It becomes essential when experiments are impossible but can also contribute to their design



Causality and Data Analysis

The ladder of causality (J.Pearl)



It is not possible to determine causality only from data analysis: interventions (experiments) are indispensable.

On the other hand, good analysis tools are essential.

Association techniques to extract robust indications about causality have been recently developed (C.Granger)

Given the difficulties of the task a <u>multipath approach is proposed: convergence of different numerical methods</u>

Correlation and Causality

- Statistics and machine learning tools have become very powerful in detecting correlations but their paradigms are blind to the distinction between correlation and causality
- This can have catastrophic consequences for control

Prediction of Y given X and Z (assuming values between 0 and 1).

$Y \leftarrow 0.5X + E_y$ where E_y is the additive error affecting Y

$Z \leftarrow Y + E_z$ where E_Z is the additive error affecting Z

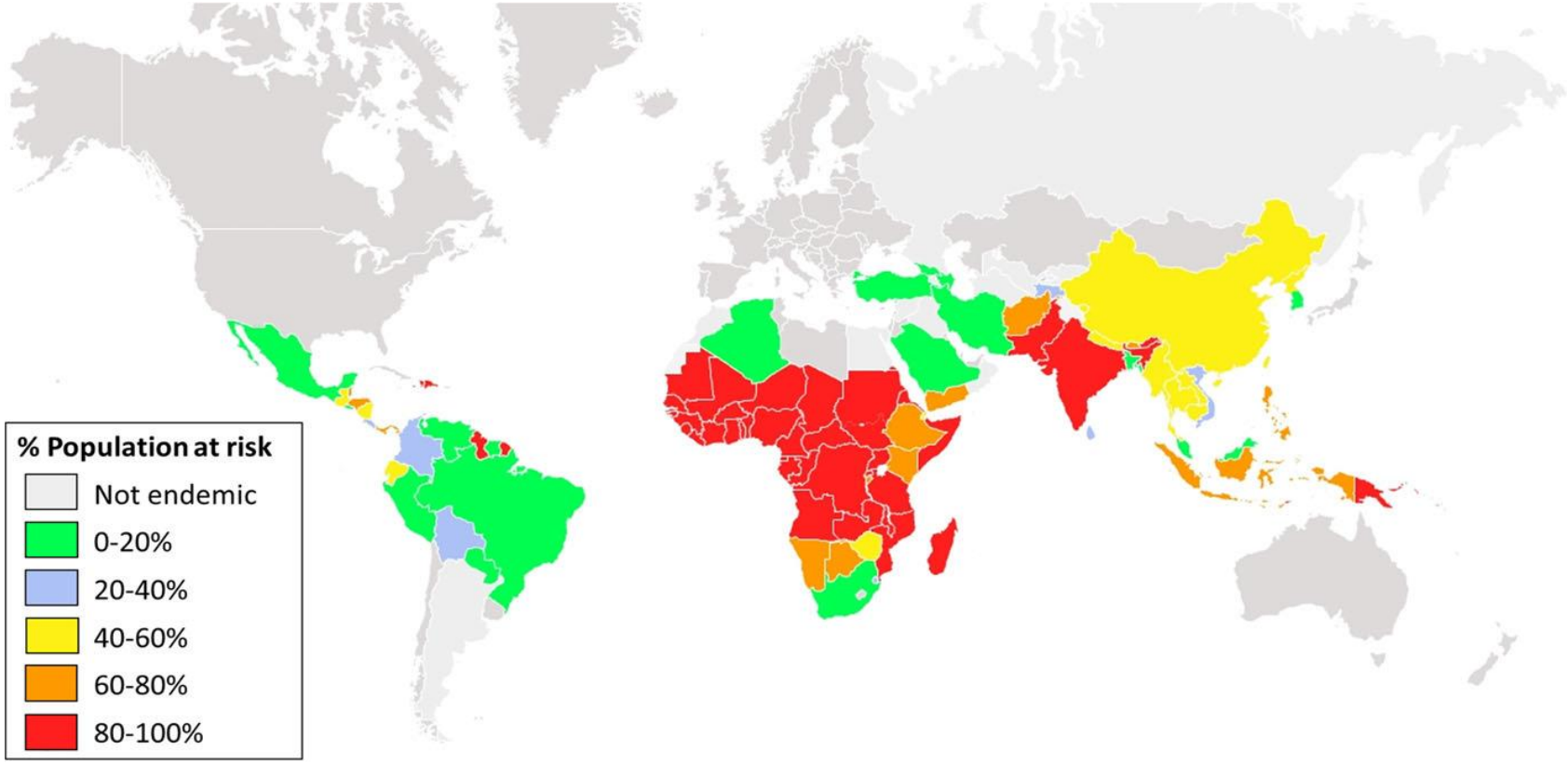
If Z has a better SNR than X, the best regression is:

$$Y = 0.25\ X + 0.5Z$$

Examples of applications: malaria epidemics

Causal influence of climatic factors on malaria epidemics

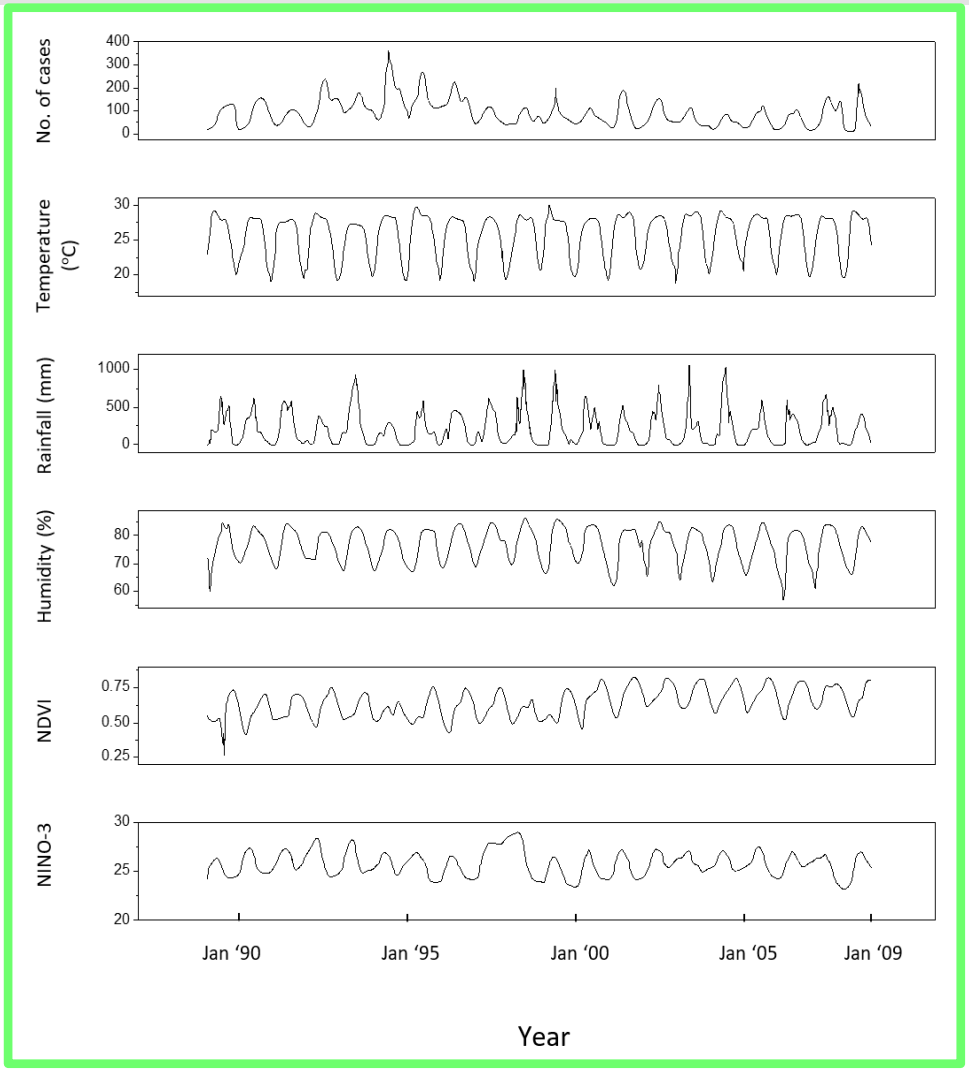
Global populations at risk of malaria, in 2013: ***3.4 billion***



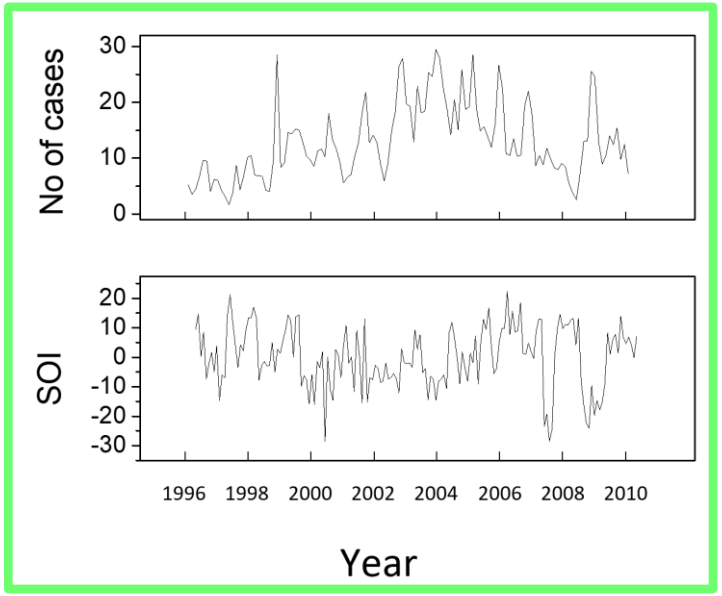
Map generated by the World Health Organization's Malaria Mapper

http://www.worldmalariareport.org/node/68

Climatic factors and malaria epidemics



Time series corresponding to Rangamati (Bangladesh), 1989–2008 [8] - from top bottom: number malaria cases, temperature, rainfall, humidity, NDVI and NINO-3.



Time series corresponding to the malaria cases at the Cayenne General Hospital (French Guyana) and to SOI for the time interval 1996–2009

M. Hanf, et al, Malar. J., 10, 100, 2011.

U. Haque, et al, PLoS ONE, 5(12), e14341, 2010.

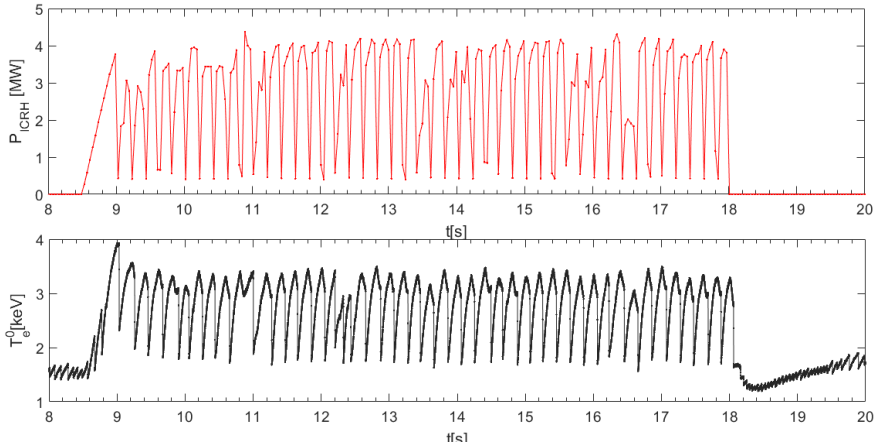
Examples of applications: fusion

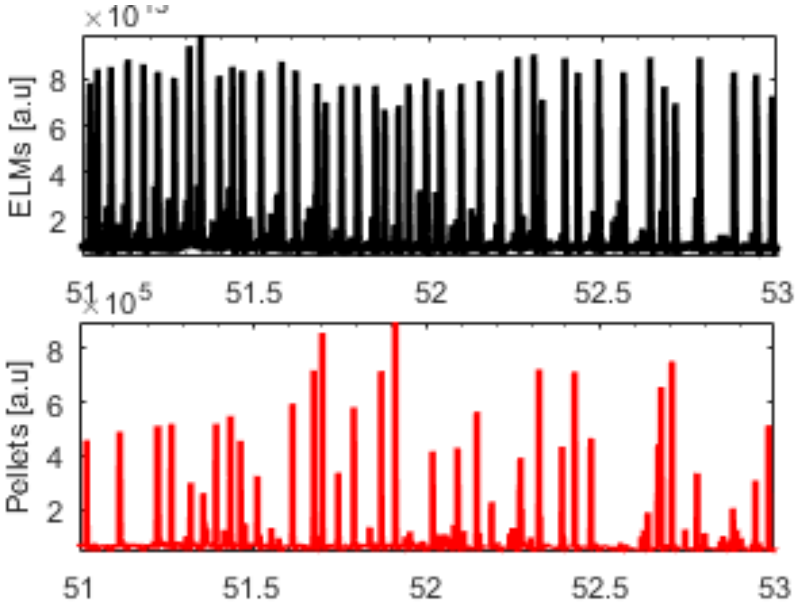
- Let us consider two examples: the problem is to determine which external preturbation (bottom plot) triggers the reaction (top plot)

a) Sawteeth triggering with ICRH (top) *what's the average time interval between the ICRH modulation and the saweteeth crashes?*

b) ELMs pacing (bottom) *how many ELMs have been triggered in the most efficient coupling time?*

- No theoretical model can establish or quantify the synchronization between experimental data.

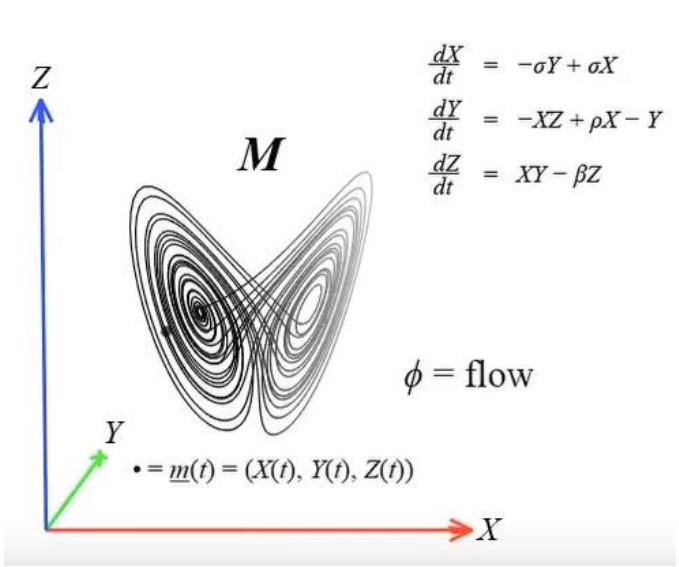




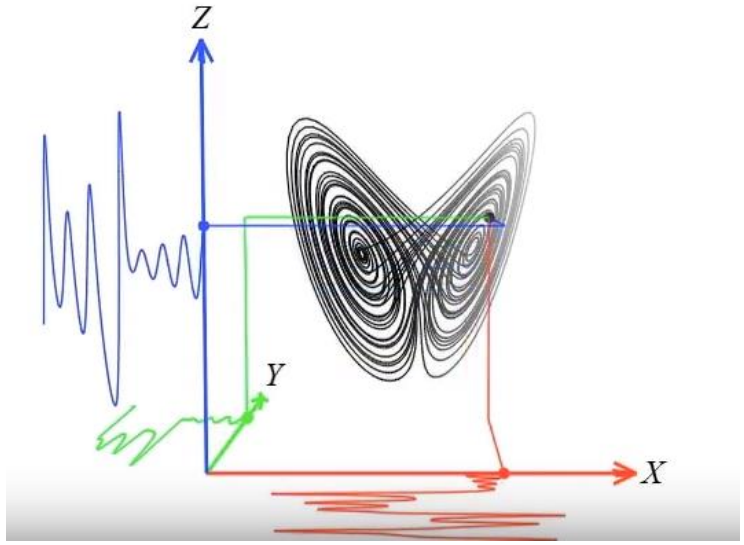
Observational Causality Detection

- Traditionally the techniques for observational causality detection are based on :
    - Conditional independence
    - Modelling with constraints (Occam Razor)
    - Prediction improvement (Granger Causality)
    - Asymmetries between cause and effect
- More recent are based on a state space approach (Taken's theorem)
- Very recent developments exploit a combination of deep learning and complex networks

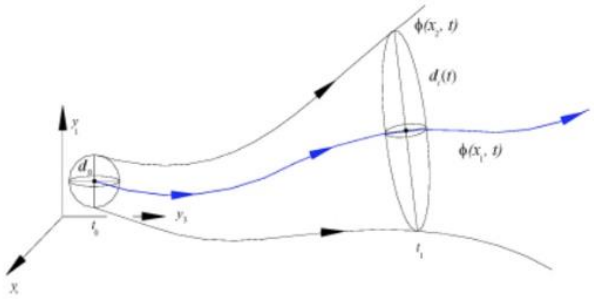
Dynamical Systems and Time Series



Coupled dynamical system (Lorenz)
Manifold M – set of all trajectories

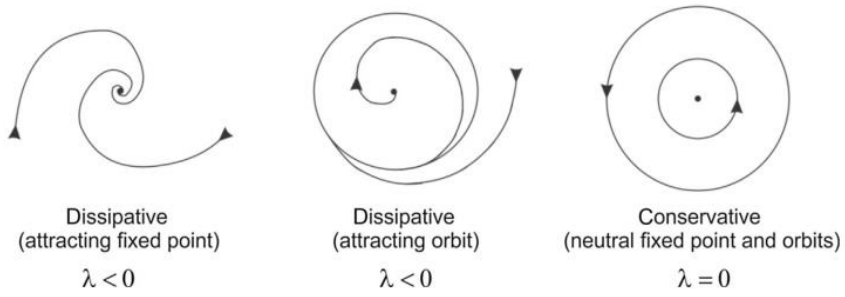


Time series – projection of the manifold on the state bases



- $|\delta \mathbf{Z}(t)| = e^{\lambda t}|\delta \mathbf{Z}(0)|$

Lyapunov exponent



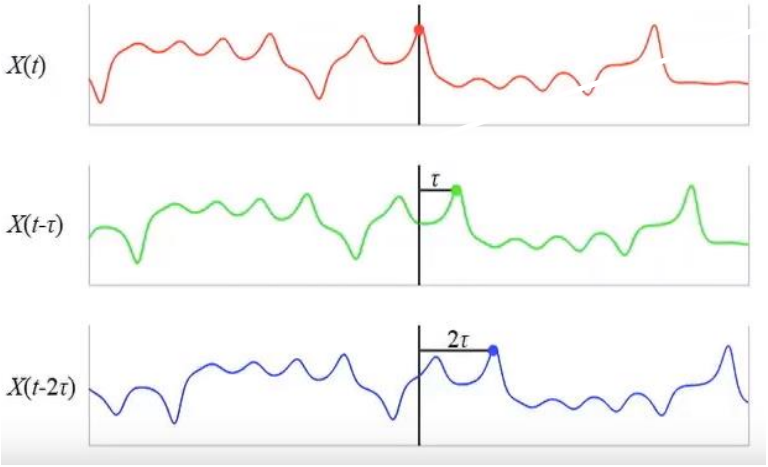
$\Lambda > 0$, the orbit is unstable and chaotic

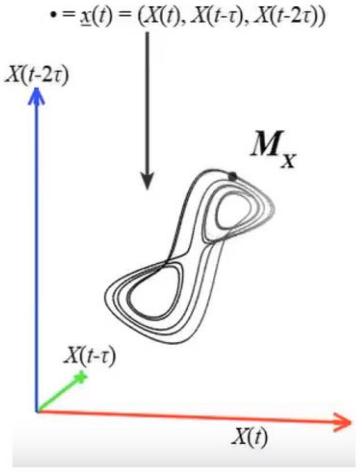
Nearby points, no matter how close, will diverge to any arbitrary distance

Manifold Reconstruction

Takens theorem (1981)

Reconstructing a shadow of the original manifold simply by looking at one of its time series projections using embedding

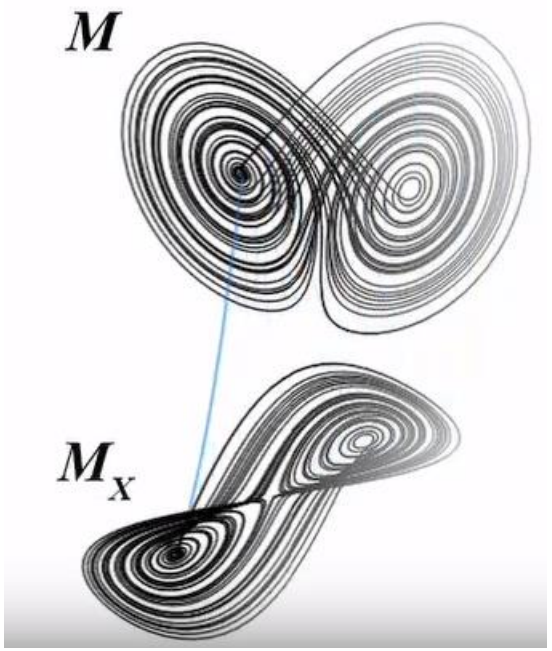




Copies of variable X displaced by τ

The reconstruction preserves the essential mathematical properties of the system:

- Topology of the manifold
- Lyapunov exponents.

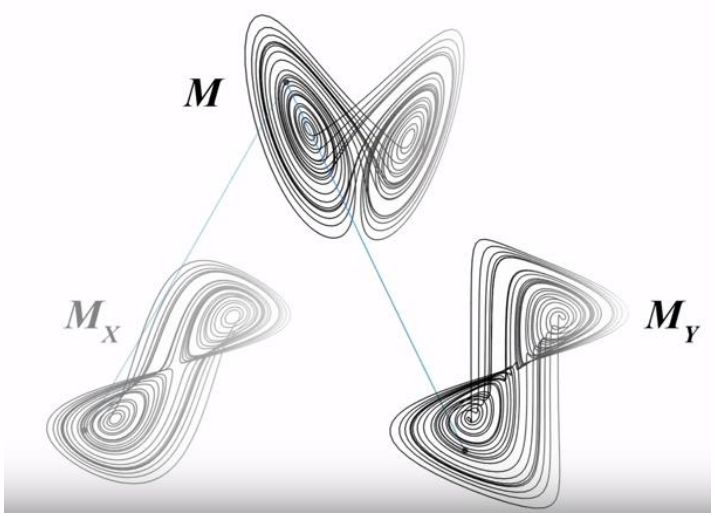


one-to-one correspondence

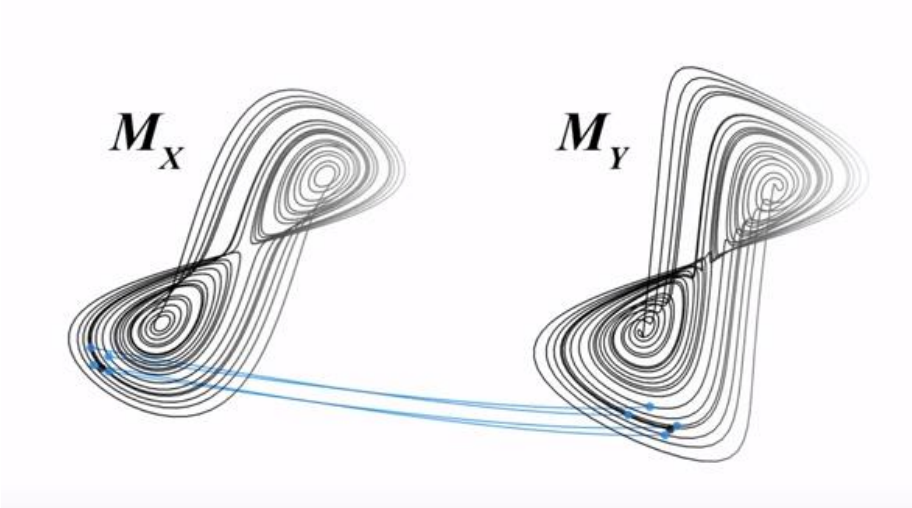
Cross Convergent Maps (1)

Two time series are related by a causal relation if they belong to the same dynamical system

M – original manifold
$\mathbf{M_X}$ – manifold created from time series X
$\mathbf{M_Y}$ – manifold created from time series Y
$\mathbf{M_X}$, $\mathbf{M_Y}$ - diffeomorphic to the original **M**



As $\mathbf{M_X}$ and $\mathbf{M_Y}$ maps one-to-one to the original manifold **M** then they should map one-to-one to each other



Nearest neighbors on $\mathbf{M_X}$ should correspond to nearest neighbors on $\mathbf{M_Y}$

- CCM - determine how good is the correspondence between local neighbourhoods on $\mathbf{M_x}$ and local neighbourhoods on $\mathbf{M_y}$.

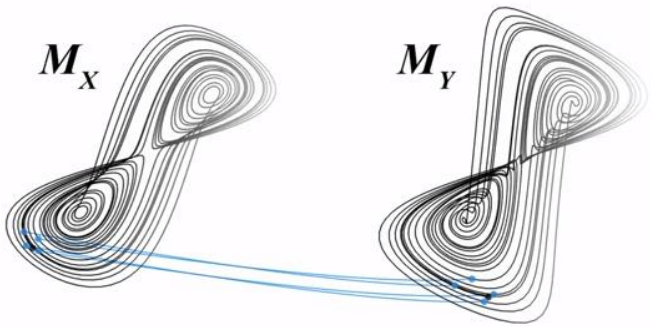
G. Sugihara et al., Science, 2012

Cross Convergent Maps (2)

- determining the embedding dimension **E**
- for each time t, x(t) is the corresponding value in one time series and y(t) in the other
- for each x(t), the **E+1** neighbours are found
- **Nb_1, Nb_2, … , Nb_{E+11}** be the time indices of the nearest neighbors of **x(t)**, ordered from closest to farthest
- Then these time indices are used to construct a putative neighborhood on the manifold of system **y**

The difference between the estimated values $y^{est}(t)$ and the actual values ***y(t)*** is evaluated by the Pearson correlation coefficient:

$$\rho = \frac{COV(Y^{est}, Y)}{\sigma(Y^{est})\sigma(Y)}$$



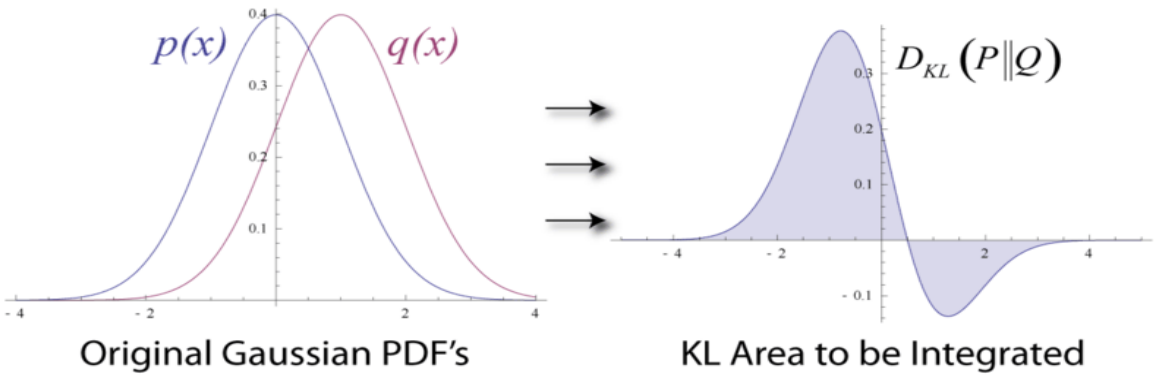
Kullback-Leibler divergence

- For distributions *P* and *Q* of two continuous random variables KL-divergence is defined to be the integral:

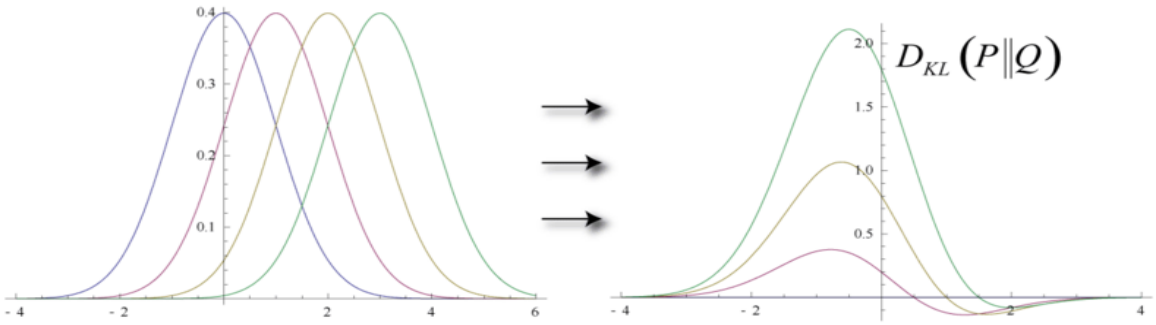
$$KLD(P||Q) = \int p(x) \cdot \ln\left(\frac{p(x)}{q(x)}\right) dx$$

where *p* and *q* denote the densities of *P* and *Q*.

- The Kullback–Leibler divergence is always non-negative and zero if and only if $p = q$.



- The smaller the KLD the closer the densities p and q



Transfer Entropy

- Introducing the formalism of the Markov processes and consequently the transition probabilities, the two observables to be analysed can be written as $i_n^{(k)} = (i_n, \dots, i_{n-k+1})$ and $j_n^{(l)} = (j_n, \dots, j_{n-l+1})$.

- Then the dynamical structure of the relationship between the two can be investigated using the Transfer Entropy, defined as:

$$T_{J\to I} = \sum_n p\left(i_{n+1}, i_n^{(k)}, j_n^{(l)}\right) \log\left(\frac{p\left(i_{n+1}|i_n^{(k)}, j_n^{(l)}\right)}{p\left(i_{n+1}|i_n^{(k)}\right)}\right)$$

The main characteristic of the TE are:

- The more j influences i, the higher the TE, i.e it has to be maximized
- It is time asymmetric, i.e the $TE_{J\to I} \neq TE_{I\to J}$.
- TE takes into account the past history of the signals

Schreiber Physics Review Letters 2010

Recurrence Plots (1)

A recurrence plot (RP) is a plot showing the times at which the phase space trajectory of a dynamical system visits roughly the same area in the phase space

RPs are based on the following matrix representation:

$$R_{ij} = \Theta\left(\varepsilon - \left\|\overrightarrow{x_i} - \overrightarrow{x_j}\right\|\right), \quad \text{i,j=1,...,N}$$

$\overrightarrow{x_i}$, $\overrightarrow{x_j}$ - points in phase space at times ***i*** and ***j***

ε is a predefined threshold

$\Theta(x)$ is the Heaviside function

j

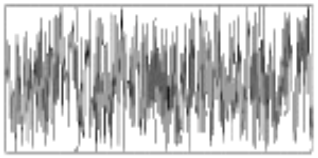
i

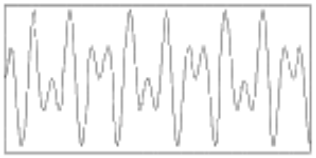
❖ RP depict the collection of pairs of times at which the trajectory returns sufficiently close the same place

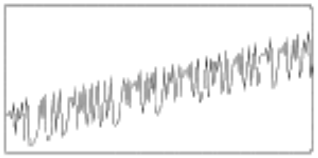
N. Marwan et al., Physics Reports, 438(5-6), 237-329.

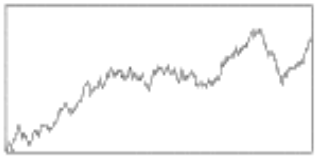
Recurrence Plots (2)

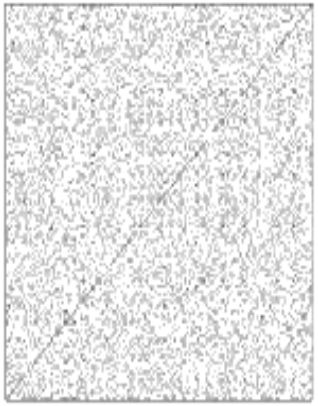
- Measure the recurrences of the trajectories
- Instrument for visualizing the behavior of trajectories in phase space

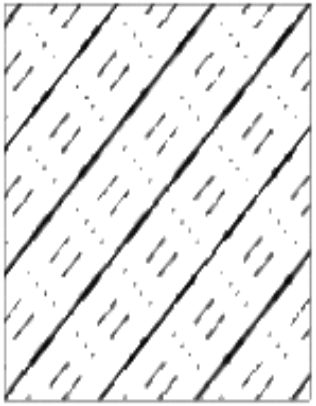


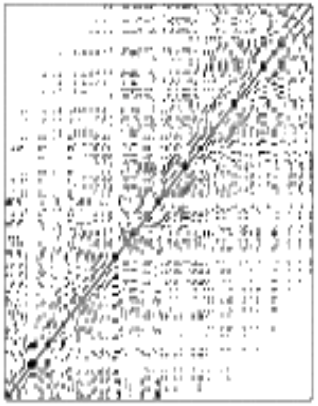












uncorrelated stochastic data (white noise)

harmonic oscillator with two frequencies

chaotic data with linear trend (logistic map)

auto-regressive process

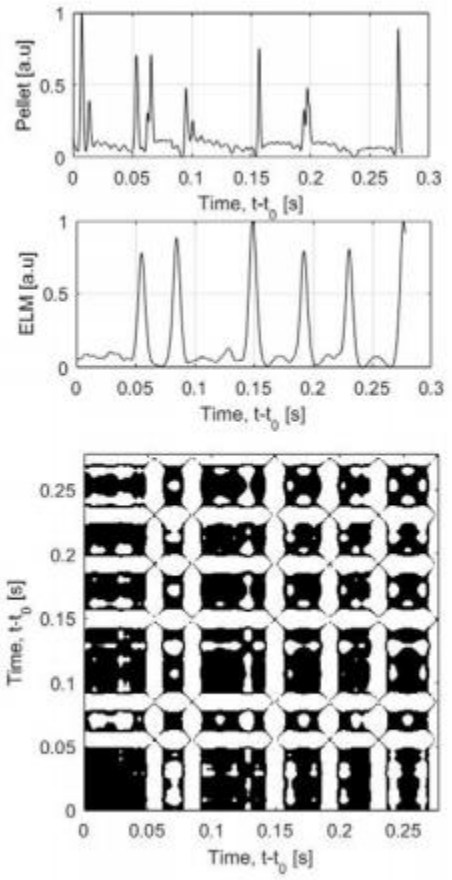
Joint Recurrence Plots

Hadamard product of the recurrence plots of the considered sub-systems

$$JR(i,j) = \Theta\left(\varepsilon_x - \left\|\overrightarrow{x_i} - \overrightarrow{x_j}\right\|\right) \cdot \Theta\left(\varepsilon_y - \left\|\overrightarrow{y_i} - \overrightarrow{y_j}\right\|\right)$$

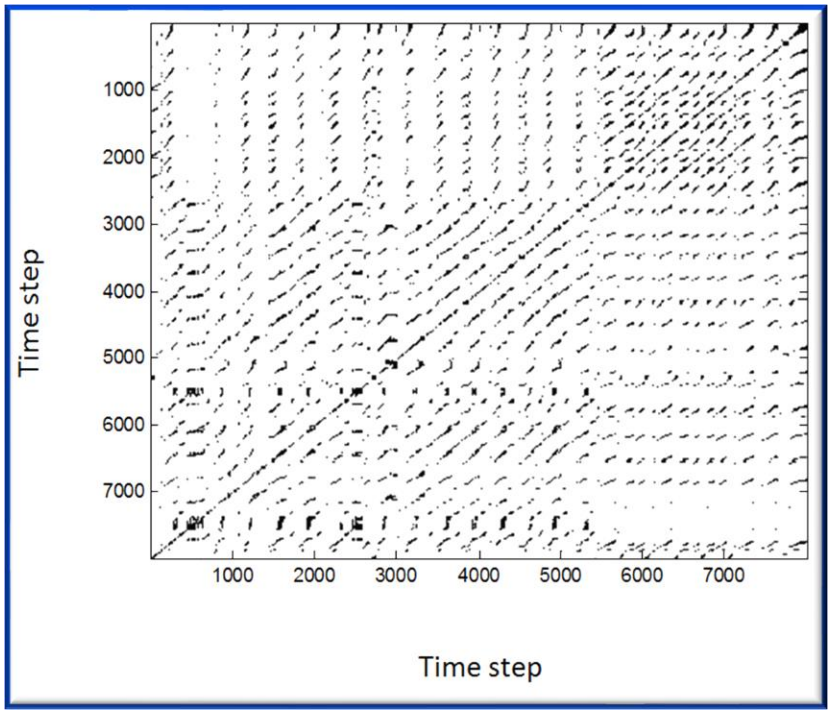
- Compare the simultaneous occurrence of recurrences in two (or more) systems
- Joint recurrence plots can be used to detect phase synchronisation.

RQA measures can be defined also for JRP



JRP between ELMs and pellets

Recurrence Quantification Analysis (RQA)



- Single isolated points correspond to states with a rare occurrence, do not persist of they are characterized by high fluctuation
- Vertical/Horizontal lines correspond to states which do not change significantly during a certain period of time
- Diagonal lines occur when the trajectory visits the same region at different times and a segment of the trajectory runs parallel to another segment.
- Long diagonal structures correspond to similar time evolution of the two processes.

The ***entropy of the diagonal lengths***:

$$ENTR = -\sum_{l=l_{min}}^{N} p(l) \cdot \ln[p(l)]$$

➢ Gives a measure of how much information is needed to recover the system and it reflects the complexity of the RP with respect to the diagonal lines.

J.P. Zbilut et al., Physics Letters A. **171** (3–4): 199–203

Definition of the «Causality Horizon»

The «causality horizon» is the maximum time interval into which two physical quantities are coupled (i.e syncronized) and in which one observable can be thougth as the «drive mechanism» of the second observable.

Pulse number	Regime	CCM Causality horizon [ms]	JRP Causality horizon [ms]	TE Causality horizon [ms]	Triggering [%]	Slowing down time of the ions [ms]
89822	L	51	[53,52]	63±5	80	50 ±10
89826	L	52	54	87±3	71	50 ±10
90005	H	69	72	67±9	42	80 ±20
90006	H	98	[95,93]	75±5; 94±10	72	80 ±20

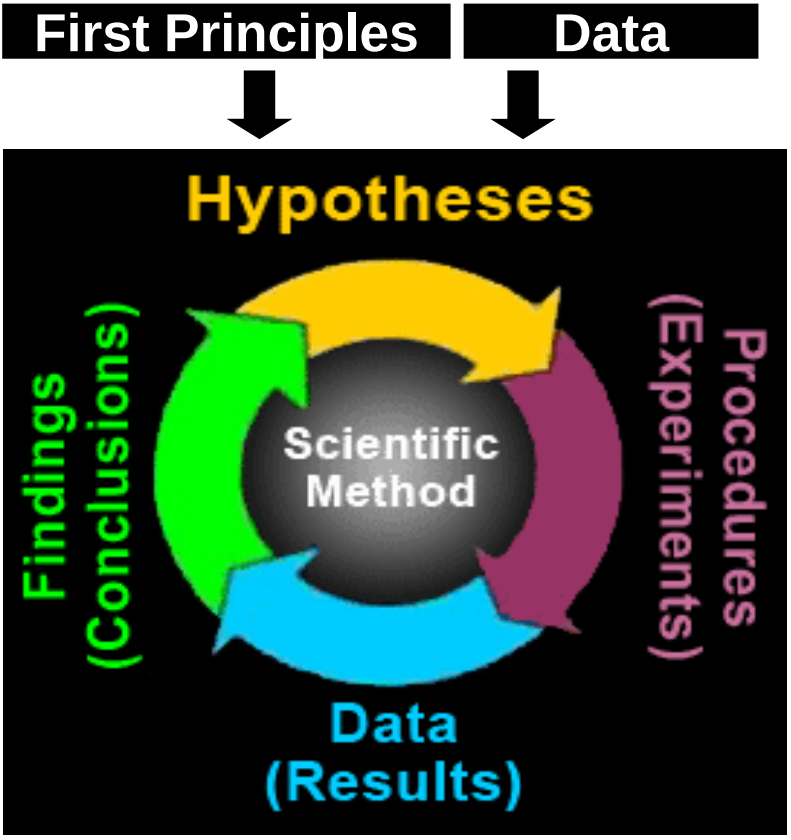
- The three indicators give similar estimates of the **"causality horizon"** and are in excellent agreement with the slowing down time of the ions.

Conclusions and Future Work on Causality

- A set of causality detection techniques for longitudinal data have been adapted to analyse fusion time series.
- Systematic numerical tests have been carried out to prove their performance for realistic time series (clearly outperforming more traditional techniques).
- Additional technique based on complex networks and neural computation are being developed
- A lot of work remains to be done to interpret the various techniques and to assess the relative merits of the various methods and address scalability

Conclusions

The developed tools are meant to complement traditional theory formulation and computer simulations not to replace them.



Data driven methods

1. They try to mathematize also the phase of hypothesis formulation from observations and data (in analogy to hypothesis formulation from first principles)
2. They try to overcome the dichotomy between model testing and theory from first principles (and the division of labour)
3. The observational causality detection tools can help in all phases of the scientific process from experiment design to interpretation

Thank You for Your Attention!

QUESTIONS?