EUROfusion

Data Driven Theory to Support Model Formulation and the Design of New Experiments

Authors: A. Murari, E.Peluso, M.Lungaroni, P.Gaudio, T.Craciunescu, J.Vega, and M. Gelfusa

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

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

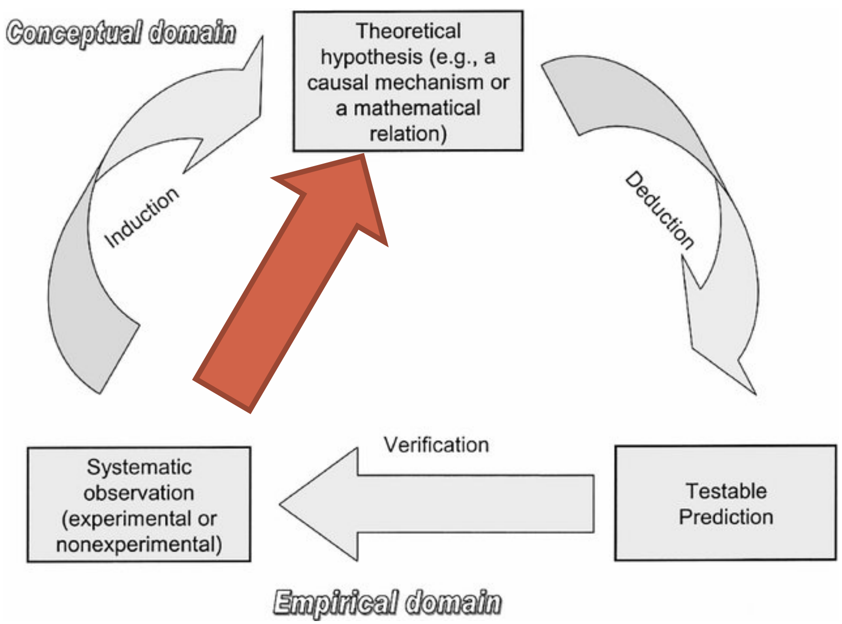
Ciemat

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.

Scientific cycle

- The “scientific process” relies on the formulation of testable predictions, which implies a dialectic relation between two domains: conceptual and empirical.



- The deduction step is very well formalised
- The induction step is more an art than a science and would benefit from: 1) more flexible tools for knowledge discovery 2) a more solid mathematization of the procedures. Data Driven Theory

1. A. Murari et al, Entropy 2017, 19, 569; doi:10.3390/e19100569
2. A. Murari et al Nuclear Fusion, Volume 57, Number 1 November 2016
3. A. Murari et al 2016 Nucl. Fusion 56 076008
4. A. Murari et al 2017 Nucl. Fusion 57 126057
5. A.Murari et at Nuclear Fusion, Volume 56, Number 2 (2015)
6. A. Murari et al Plasma Physics and Controlled Fusion (2015),57 (1),

1. A. Murari et al 2016, Nuclear Fusion 56
2. A. Murari et al. Nuclear. Fusion **57** (2017) 016024 2017,
3. A. Murari et al. Nuclear Fusion , **57,** Number 12, September 2017
4. A. Murari et al Nuclear Fusion, Volume 58, Number 5, March 2018

Data Deluge and Measurements

•The amount of data produced by modern societies is enormous

•JET can produce more than 55 Gbytes of data per shot (potentially about 1 Terabyte per day). Total Warehouse: almost 0.5 Petabytes

•ATLAS can produce up to about 10 Petabytes of data per year

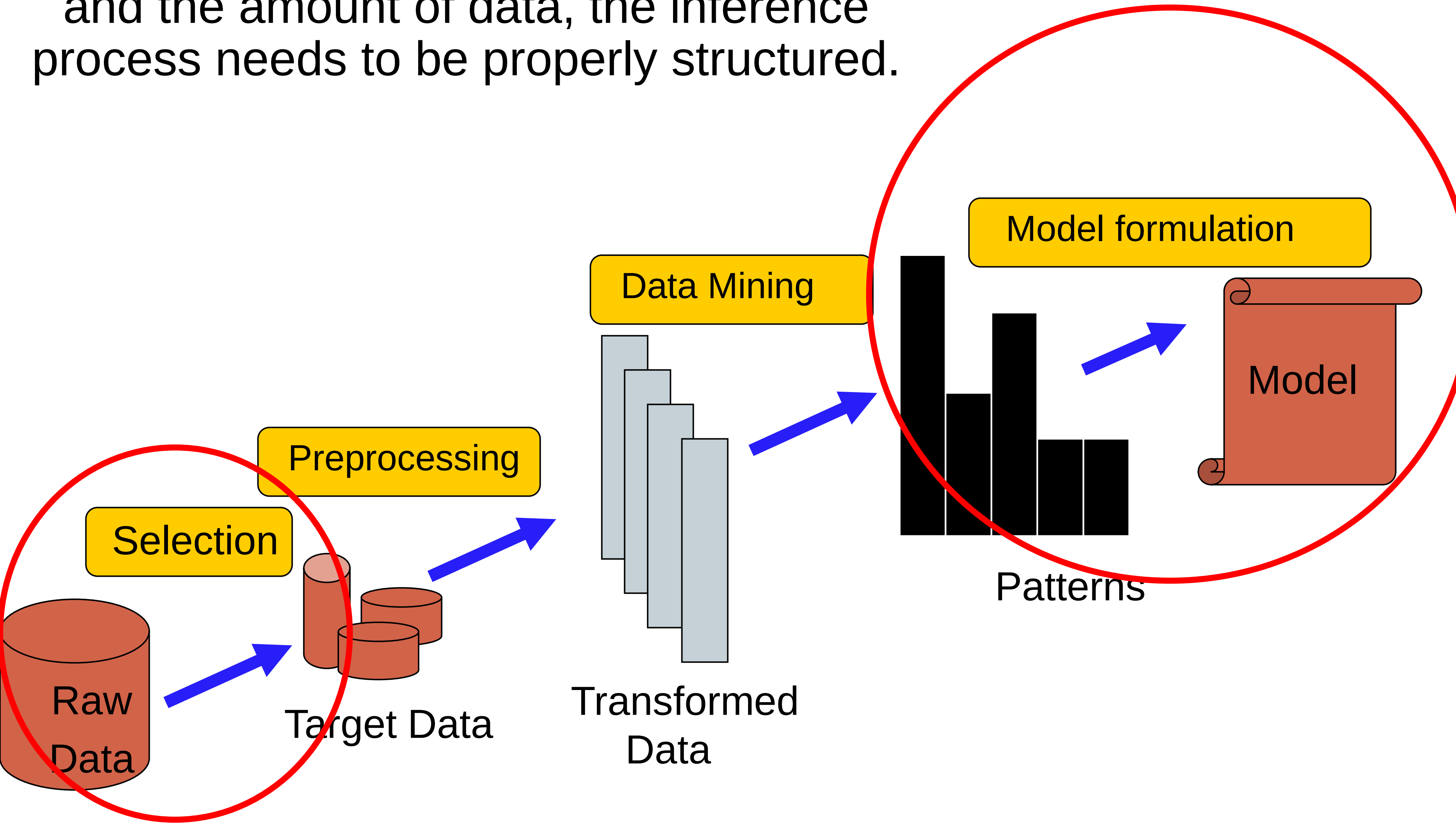
•Hubble Space Telescope in its prime sent to earth up to 5 Gbytes of data per day

•Commercial DVD 4.7 Gbytes (Blue Ray 50 Gbytes).

These amounts of data cannot be analysed manually in a reliable way. Given the complexity of the phenomena to be studied, there is scope for the development of new data analysis tools particularly in support to theory formulation!!

Data Analysis: an overview

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



Outline

I. Symbolic Regression/Genetic programming to extract models directly from the data for better "physics fidelity" and interpretability

II. Numerical tests: identif. dimensionless quantities

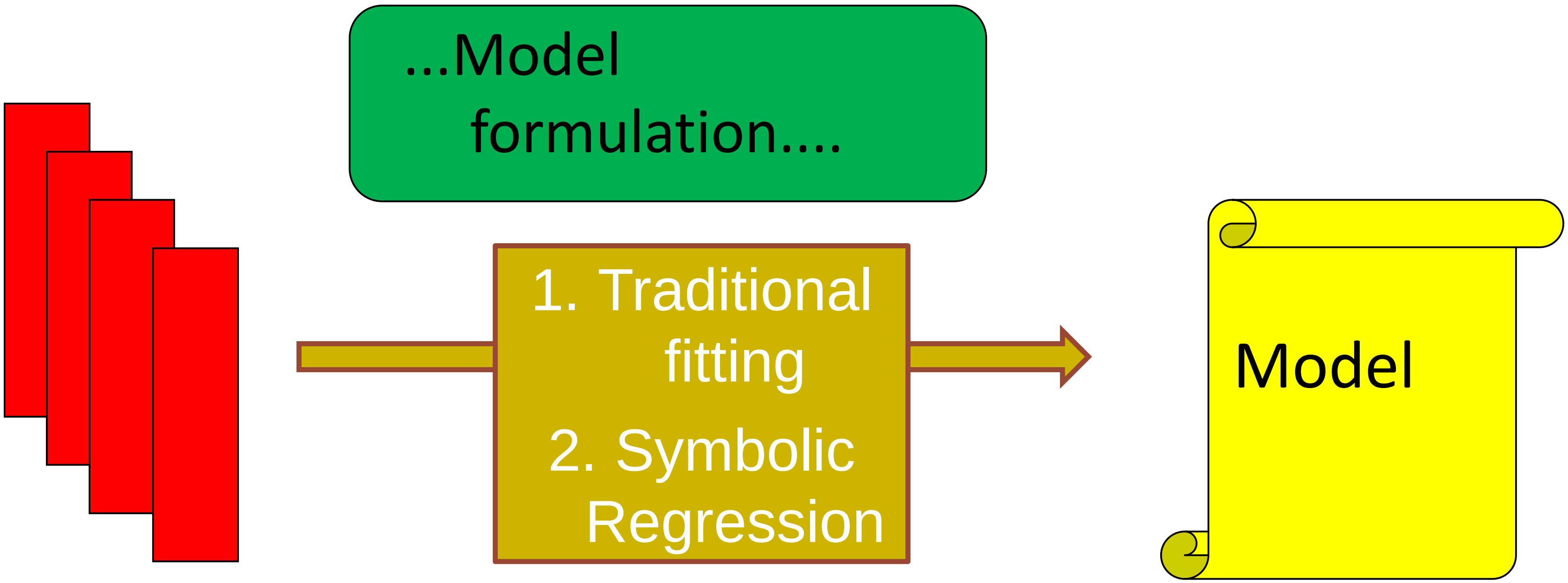
III. Scaling laws (energy confinement time τ_E): exploratory application

IV. Identification boundary between safe and disruptive regions of the operational space: interpretative appl.

V. 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.

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

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

Traditional Fitting

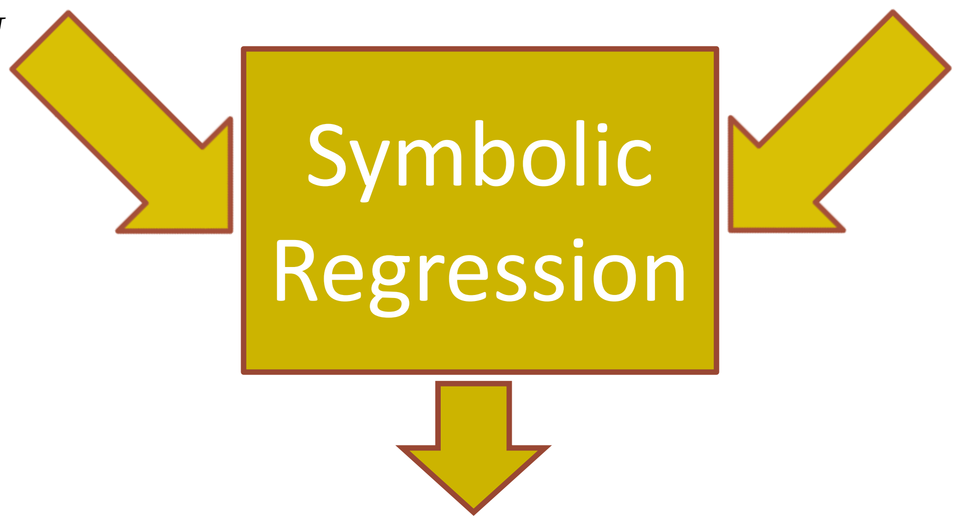
$$y_{to\ be\ fitted} = \alpha_1 \cdot \sin(\alpha_2 \cdot x_1^{\alpha_3}) + \alpha_4 \cos(\alpha_5 \cdot x_1^{\alpha_6} \cdot x_2^{\alpha_7})$$

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

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

$$y_{data}$$



$$y_{\mathrm{mod}\ by\ SR} = \alpha_1 \cdot \sin(\alpha_2 \cdot x_1^{\alpha_3}) + \alpha_4 \cos(\alpha_5 \cdot x_1^{\alpha_6} \cdot x_2^{\alpha_7})$$

Genetic Algorithms 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).

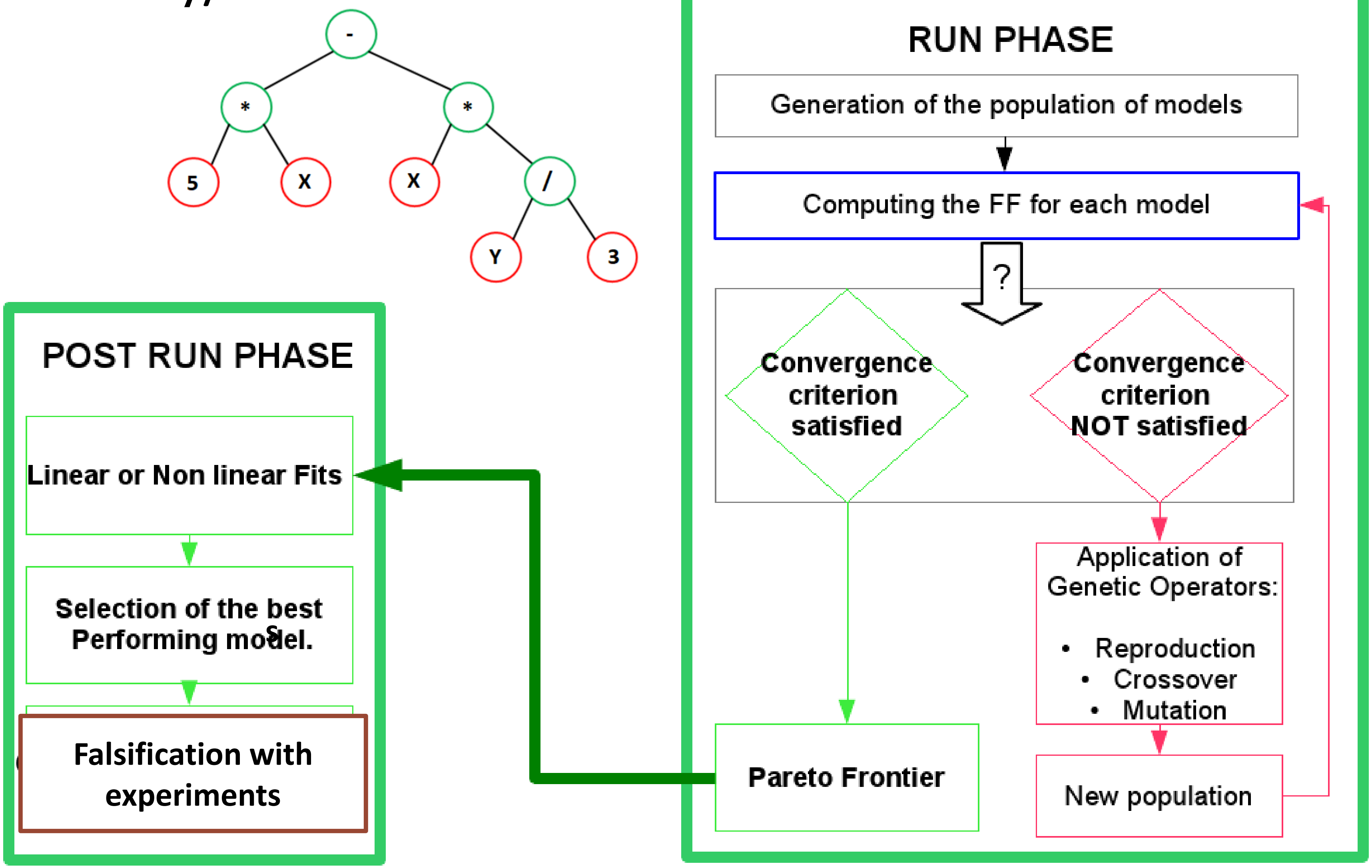
  3- Select the best fitting individuals (parents) to create a new population of trees (formulas).

  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>

Outline

I. Symbolic regression/Genetic programming to extract models directly from the data for better "physics fidelity" and interpretability

II. Numerical tests: identif. dimensionless quantities

III. Scaling laws (energy confinement time τ_E): exploratory

IV. Identification boundary between safe and disruptive regions of the operational space

V. Conclusions

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: $Pe = 0.99 \cdot PrRe$

Outline

I. Symbolic regression/Genetic programming to extract models directly from the data for better “physics fidelity” and interpretability

II. Numerical tests: extraction dimensionless quantities

III. Scaling laws (energy confinement time τ_E): exploratory

IV. Identification boundary between safe and disruptive regions of the operational space

V. Conclusions

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)

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

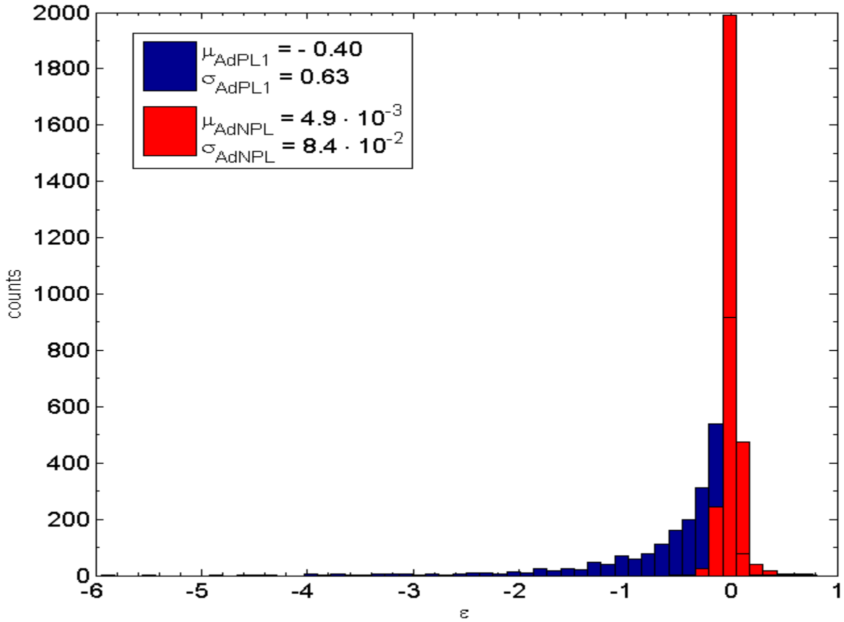
$$\omega \cdot \tau_{AdNPL} = (1.13)_{1.11}^{1.15} \cdot 10^{-6} \frac{k_a^{1.93_{1.70}^{2.12}} \beta^{0.37_{0.35}^{0.41}} M^{0.57_{0.46}^{0.67}}}{\rho^{2.19_{2.16}^{2.22}} \nu^{0.40_{0.39}^{0.42}} q^{0.16_{0.03}^{0.23}}} - (0.072)_{-0.085}^{-0.060} k_a^{1.18_{0.94}^{1.40}} +$$

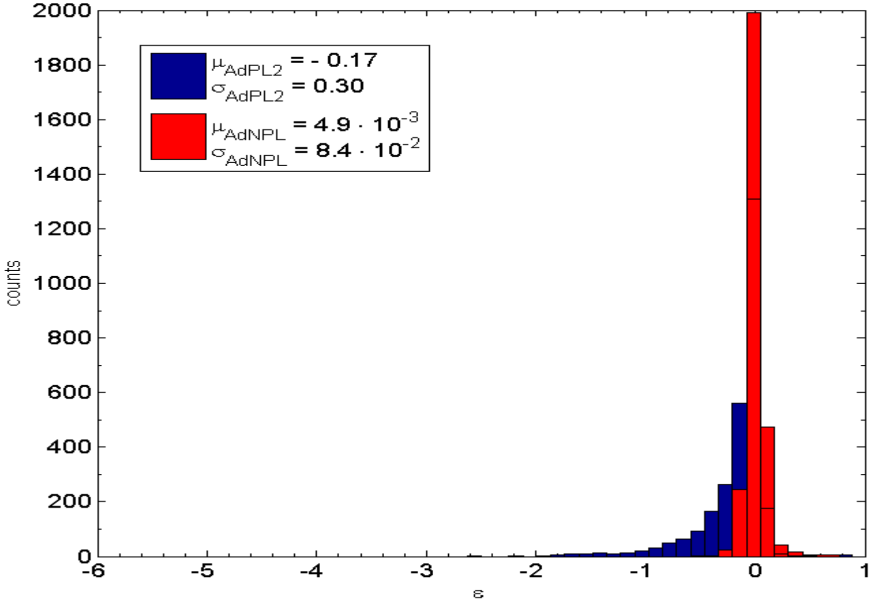
$$-0.009_{-0.011}^{-0.006} q^{1.08_{0.94}^{1.21}} + 0.15_{-0.13}^{-0.17} M^{0.07_{-0.05}^{0.19}}$$

	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

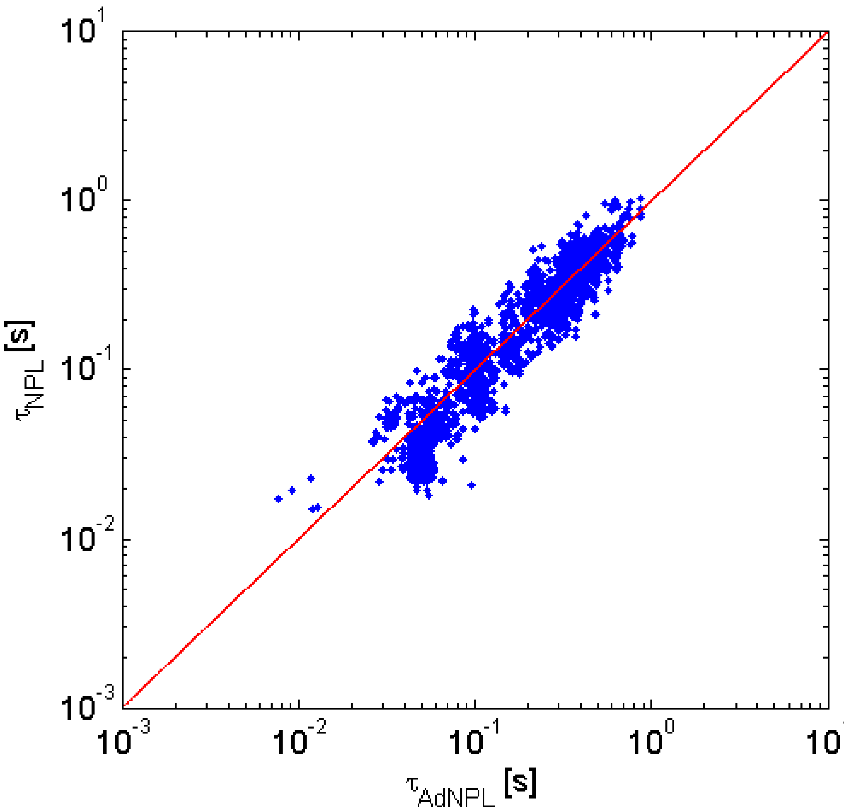




	k	AIC	BIC	MSE	KLD
AdPL1	9	−2930.77	−4505.82	$12.078 \cdot 10^{-2}$	8.2048
AdPL2	9	−3461.54	−4813.36	$8.255 \cdot 10^{-2}$	3.8786
AdNPL	14	−5610.85	−5723.52	$1.756 \cdot 10^{-2}$	0.9758

Indipendent Scalings with dimensional and dimensionless regressors: very good agreement

Excellent independent match and ~20% reduction



Extrapolation to ITER

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

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

Outline

I. Symbolic regression/Genetic programming to extract models directly from the data for better “physics fidelity” and interpretability

II. Numerical tests: extraction dimensionless quantities

III. Scaling laws (energy confinement time τ_E): exploratory

IV. Identification boundary between safe and disruptive regions of the operational space

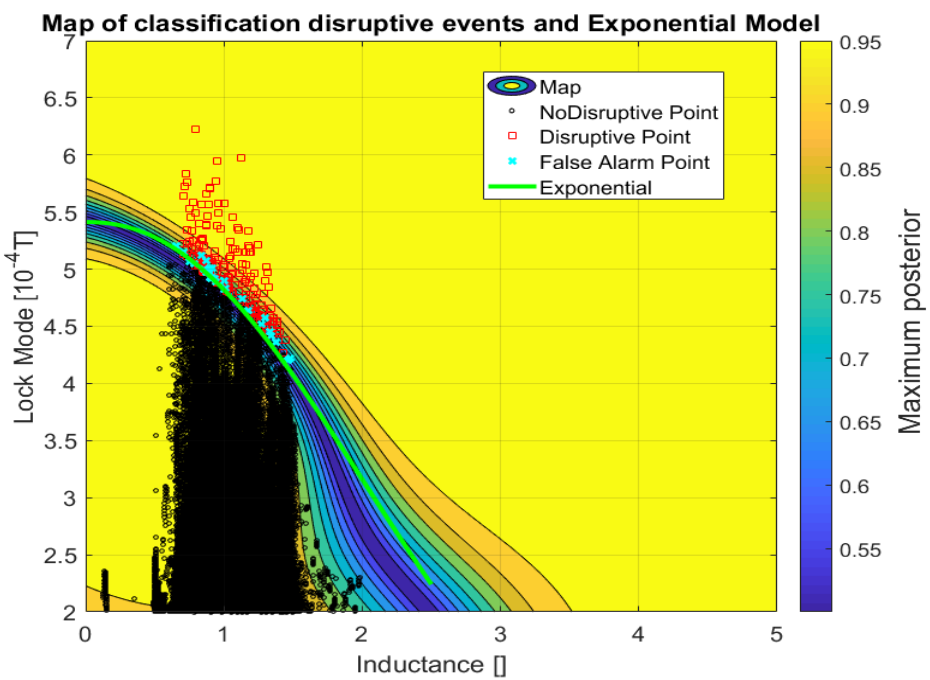
V. Conclusions

Boundary between disruptive and safe regions

Given the difficulties to develop theoretical models from first principles for disruption predictions, machine learning tools have been deployed since quite some time.

Typical criticisms: they are difficult to interpret and have poor “physics fidelity”

Probabilistic SVM with RBF kernel



Plot of the disruptive probability in the plane locked mode internal inductance The safe region is the one within the closed curve (black points safe, red disruptions, light blue false alarms).

Light blue curve: 60% probability threshold.

Threshold 60 %: 98% Success Rate of and 2.8% of False Alarms

Equation obtained with Symbolic Regression

Symbolic Regression has been deployed to regress the points on the frontier. The best equation found is not a power law: LM is the locked mode amplitude and l_i the internal inductance

$$LM(l_i) = a_0 \exp(a_1 l_i^{a_2})$$

$a_0 = 5.4128 \pm 0.0031;$

$a_1 = -0.11614 \pm 0.00085;$

$a_2 = 2.21 \pm 0.011;$

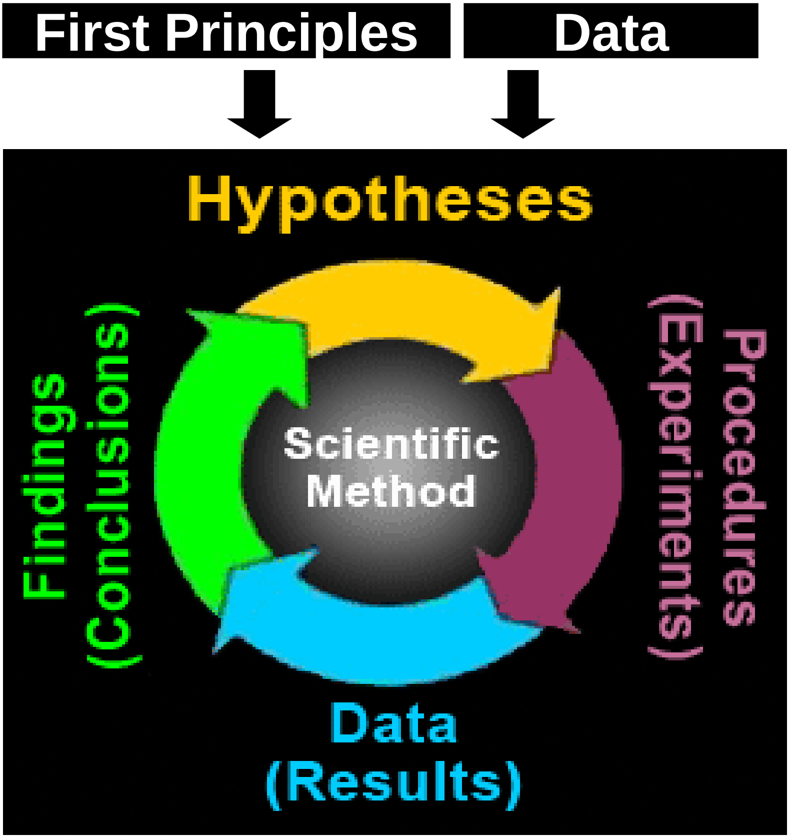
Probability Thershold	Success rate	Tardy	Early	Missed	False
60	97.9 % (183/187)	2.1 % (4/187)	0 % (0/187)	0 % (0/187)	2.8 % (29/1020)

The found formula reproduces exactly the performance of the original probabilistic SVM with 60% threshold. SVM model consists f tens of Gaussians centered on the support vectors.

SR via GP model better for both interpretability and physics fidelity

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. They are more powerful than traditional tools and can be used both in exploratory and interpretative ways (and design of new experiments)

Thank You for Your Attention!

QUESTIONS?

Essential Aspects of Genetic Programming

Three aspects are fundamental in Genetic Programming:

- Genes or Knowledge representation (how to represent formulas)
- Fitness Function (FF)
- Criteria to validate the results

A. Murari et al Nuclear Fusion, Volume 52, Number 6 May 2012 doi.org/10.1088/0029-5515/52/6/063016

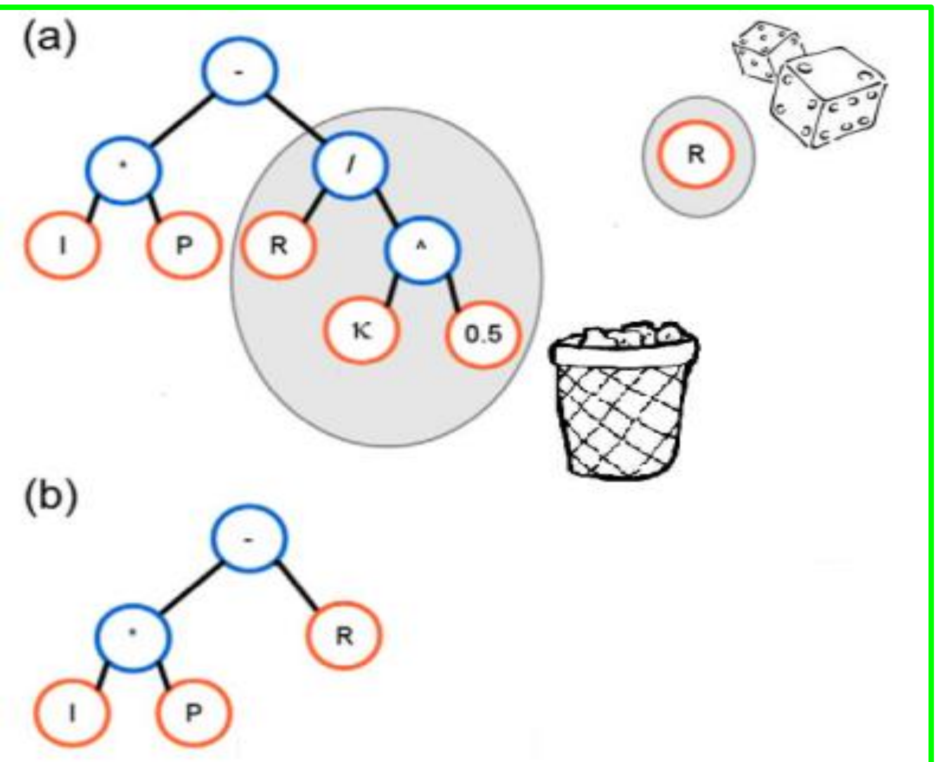
A. Murari *et al* 2013 *Nucl. Fusion* **53** 043001 doi:10.1088/0029-5515/53/4/043001

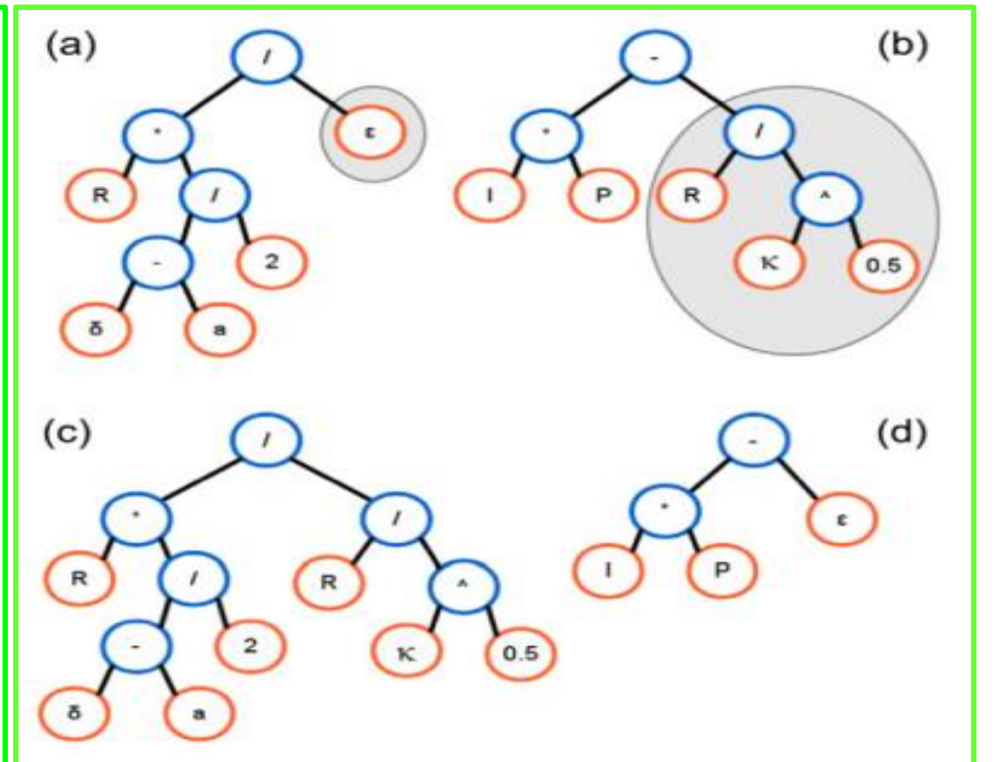
A.Murari et at Nuclear Fusion, Volume 56, Number 2 (2015) doi:10.1088/0029-5515/56/2/026005

A. Murari et al Plasma Physics and Controlled Fusion (2015),**57** (1), doi: 10.1088/0741-3335/57/1/014008

Formulas

- Formulas are represented as trees
- Trees have a very high representational capability
- This representation allows an easy implementation of genetic steps: copy, mutation, cross over





Basis functions

Function class	List
Arithmetic	constants,+,-,*,/
Trigonometric	sin(x_i), cos(x_i),tan(x_i),asin(x_i),atan(x_i)
Exponential	exp(x_i),log(x_i),power(x_i, x_j), power(x_i,c)
Squashing	logistic(x_i),step(x_i),sign(x_i),gauss(x_i),tanh(x_i), erf(x_i),erfc(x_i)
Boolean	equal(x_i, x_j),less(x_i, x_j), less_or_equal(x_i, x_j), greater(x_i, x_j), greater_or_equal(x_i, x_j), if(x_i, x_j, x_k), and(x_i, x_j),or(x_i, x_j),xor(x_i, x_j), not(x_i)
Other	min(x_i, x_j),max(x_i, x_j),mod(x_i, x_j),floor(x_i),ceil(x_i), round(x_i),abs(x_i)

Identification of dimensionless quantities

The above dimensionless quantities can be written as:

$$Re = \frac{\rho \cdot u \cdot d}{\mu} \qquad\qquad Pr = \frac{c_p \mu}{k}$$

Where:

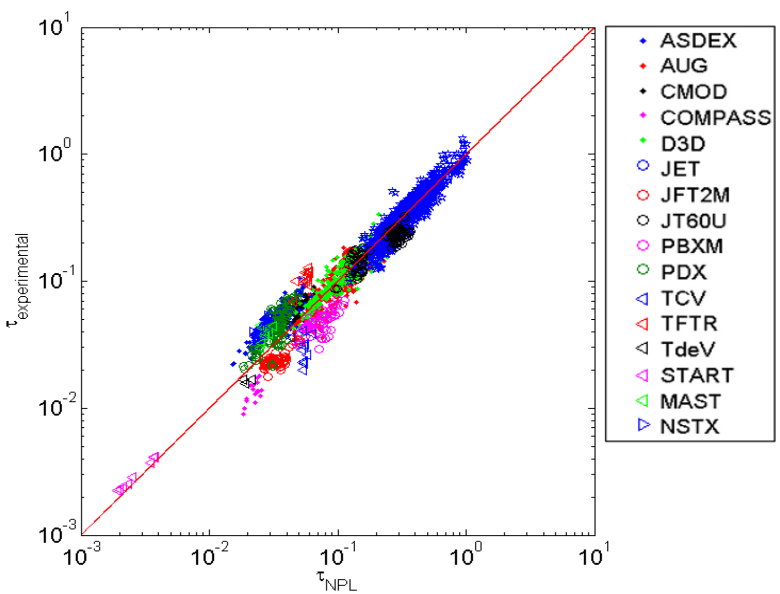
1. μ is the dynamic viscosity
2. k is the thermal conductivity
3. c_p is the specific heat
4. ρ is the density
5. u is the velocity of the fluid
6. d is a characteristic linear dimension of the object moving in the fluid

Energy Confinement Time

Dimensional scaling law of confinement time (characteristic time measuring the rate at which the plasma loses energy)

ITPA database in Carbon. Comparison with traditional IPB98y2

The scaling law obtained with SR via GP is not a power law and has better statistical indicators



ipb98y2 [6]	$\tau_E = 5.62 \cdot 10^{-2} I^{0.93} B^{0.15} n^{0.41} M^{0.19} R^{1.97} \varepsilon^{0.58} k_a^{0.78} P^{-0.69}$
NPL	$\tau_E = 3.67_{3.66}^{3.69} \cdot 10^{-2} I^{1.01_{0.99}^{1.02}} R^{1.73_{1.71}^{1.75}} k_a^{1.45_{1.41}^{1.49}} P^{-0.74_{-0.75}^{-0.72}} h(n,B)$ $h(n,B) = n^{0.45_{0.44}^{0.46}} \cdot \left(1+e^{-9.40_{-9.69}^{-9.11} \cdot (n/B)^{-1.37_{-1.41}^{-1.32}}}\right)^{-1}$

	AIC	BIC	MSE [10^{-3} s^2]	KLD
ipb98y2	-19416.86	-19362.86	1.866	0.0337
NPL	-19660.03	-19599.04	1.724	0.0254

[4]A. Murari, E. Peluso et al, Plasma Phys. Control. Fusion, 57(1), 2015, doi::/10.1088/0741-3335/57/1/014008
[5] E. Peluso, A. Murari,,et al, 41[st] EPS Conference on Plasma Physics, 2014, P 2.029
[6] McDonald D.C et al , *Nucl. Fusion* (2007), 47:147–174

Additional topics and future developments

- SR via GP presents various advantages compared to log regression: no constrained to produce power laws, no assumptions about the noise distribution, less vulnerable to collinearity etc.
- A priori information can be integrated at various levels: election of the basis functions, tree structure, correlation between branches etc.
- More advanced versions of SR via GP are available (Pareto Frontier, better treatment of the errors with Geodesic Distance on Gaussian Manifolds etc.)
- The same techniques can be applied to the results of complex simulations performed using supercomputers.
- Another interesting application is the support to experimental design