Università Ca'Foscari Venezia
Dipartimento di Statistica

European Centre for Living Technology

Evolutionary Algorithms for Complex Designs of Experiments and Data Analysis

Irene Poli

Dep. of Statistics, University Ca' Foscari of Venice

European Centre for Living Technology (ECLT)

www.ecltech.org

Research group:

Matteo Borrotti, Davide De March, Davide Ferrari, Michele Forlin, Daniele Orlando, Debora Slanzi, Laura Villanova.

outline

Complex Design of Experiments:

High Dimensionality and High Throughput (HDHT)

Intelligent data:

the evolutionary perspective

Statistical models in the evolution:

the Statistical Evolutionary Experimental Designs (SEEDS) involving small sets and low dimensional data

4 September 2008 | www.nature.com/nature | £10
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
nature
THE BITER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
NATUREJOBS
Minnesota musings
SCIENCE IN THE PETABYTE ERA

Big Data

refers to the immense volume of data that are continuously generated in any area of research, from Biology, to Material Science, Economics, Finance or Environment.

Data are growing in

size, *for the huge number of data provided by the great technological advances (high Throughput);*

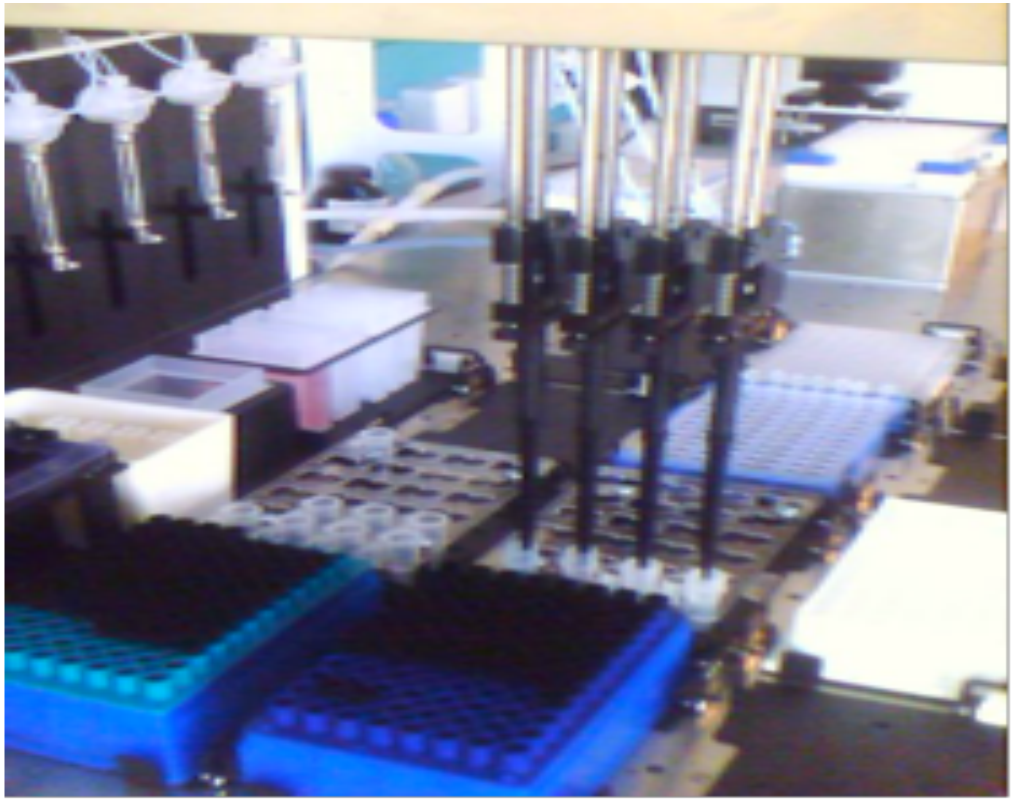
dimensions, *for the very large number of variables that investigators consider in developing research;*

complexity, *for the high level of connectivity that characterizes these data sets.*

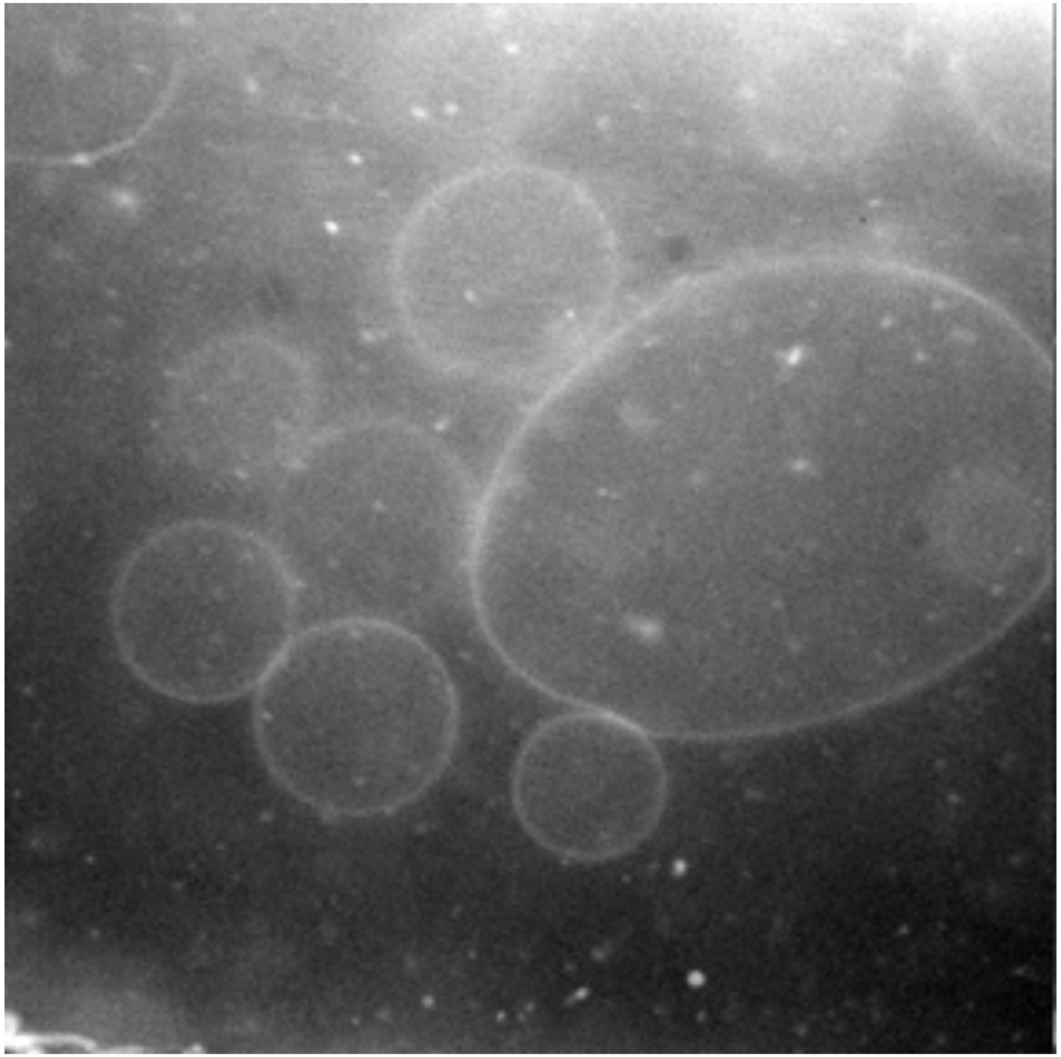
*From such "**Big Data**", how can investigators extract information, how can they find meaning and connections?*

experimentation

High Throughput Robot



The response



Protolife Laboratory, Martin Hanczyc,
EU - PACE project

Q: in HDHT settings how do we *design* the experiments?

how many and which factors should be considered in the investigation;

how many and which levels for each factor,

which interactions among factors; *which network* of interaction

which experimental technology and laboratory protocols to employ.

The Statistical Design of Experiments

and the challenge of *high dimensional data.*

When the number of variables increases the number of experimental points to be explored increases exponentially

Developments in*:*

Feature selection and

Dimensionality reduction: Tibshirani , Donoho,

Johnstone and Titterington; Li, Cook, Fan, Li

Fractional Factorial Design, Response surface, Jones, Myers

Uniform Design: Lin, Sharpe, and Winker

Evolution, as a search engine in HDHT

The idea is to learn from Nature:

how Nature solves complex and complicated problems?

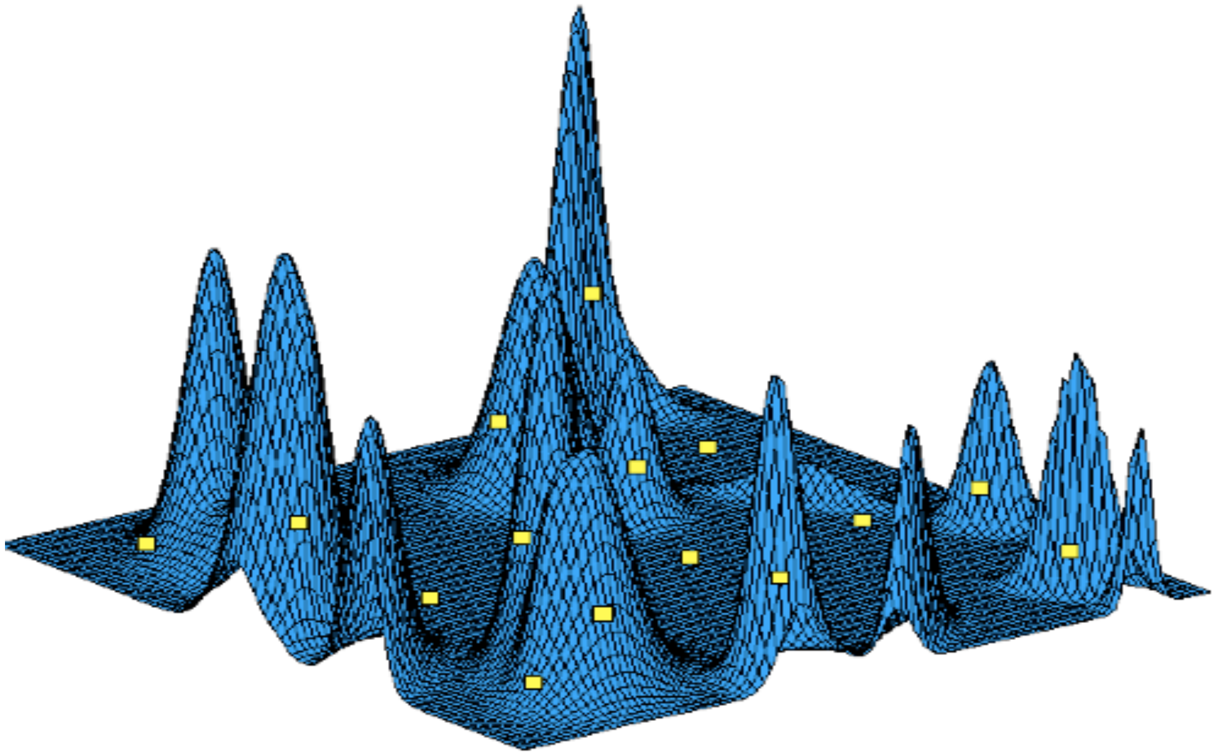
Living systems evolve through generations, learning, adapting, changing in a particular environment and according to a particular target.

The search in huge spaces can then be realized adopting the **Darwinian paradigm of evolution**

The Evolutionary Design

The *design of an experiment*

is a set of experimental points in a multidimensional space *where to …look* for uncovering information on the target of the problem

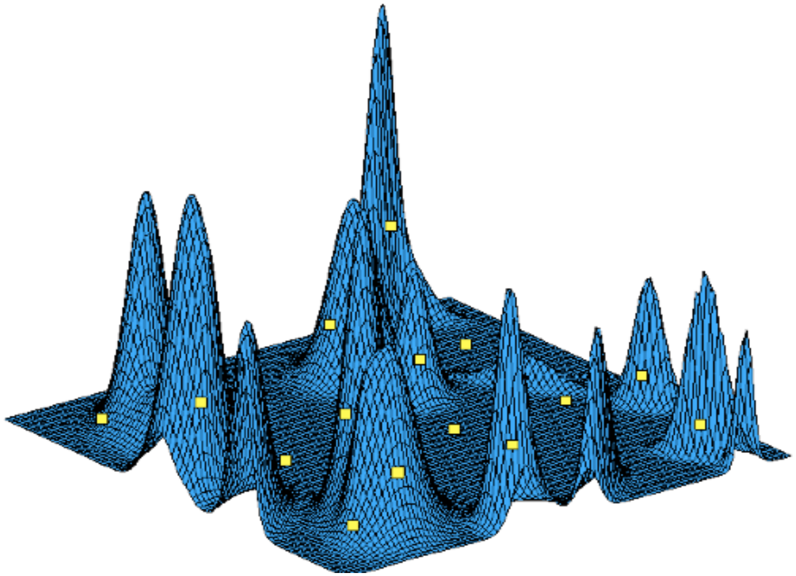


A small, low dimensional, set of sites where to collect information

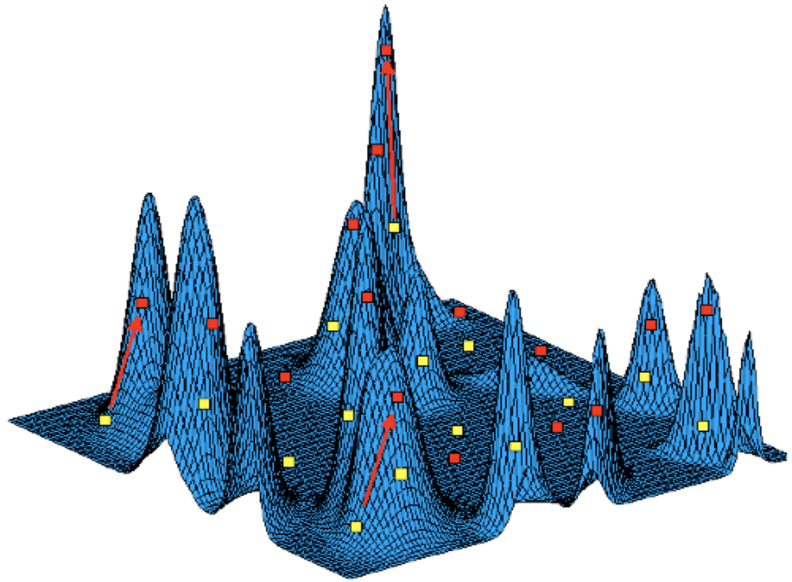
The Evolutionary Design

The design can then be represented as a *population of solutions that can learn, adapt* and then *evolve* through generations.

It is not of an *a priori* choice.



…



How to build the evolutionary design?

The problem:

Let $X = \{x_1, . . . , x_p\}$ be the set of experimental factors, with $x_k \in L_k$, where L_k is the set of the levels for factor k, k = 1, . . . , p.

The experimental space, represented by **Ω**, is the product set $L_1 \times L_2, . . . , \times L_p$.

Each element of **Ω**, namely ω_r , r = 1, . . . , N , *is a candidate solution*, and the experimenter is asked to

find ω_T^* the best combination,

the combination with the maximum (minimum) response value (optimization problem).

Evolution with a Genetic Algorithm, GA

A GA

is an iterative, population-based search procedure.

In designing experiments

the GA evolves a *population of experimental points,*

which are evaluated in their environment and

transformed under *genetic operators,*

to generate a new population experimental points,

... emulating Nature in generating new solutions.

The GA design

An initial *very small set* of experimental points, ***D***^{1} , with different structure composition, is chosen in a random way

Randomness (instead of just prior knowledge) allows the exploration of the space in areas not anticipated by prior knowledge but where interesting new information may reside.

*each element of **D**^{1}* , is a vector of symbols from a given alphabet (binary or decimal or other),

is a candidate solution to be tested.

The GA design

Experimenting D^1, we learn which are the best solutions and their compositions and

with a set of genetic operators (selection, recombination, mutation, ecc..) we can build the successive generations of solutions, i.e. the successive design.

....

D^1 ← Randomly select an initial design from Ω

Conduct the experiment testing each member of D^1 and derive its fitness function value

while termination conditions not met do

D^1_1 ← Select (D^1)

D^1_2 ← Recombine (D^1_1)

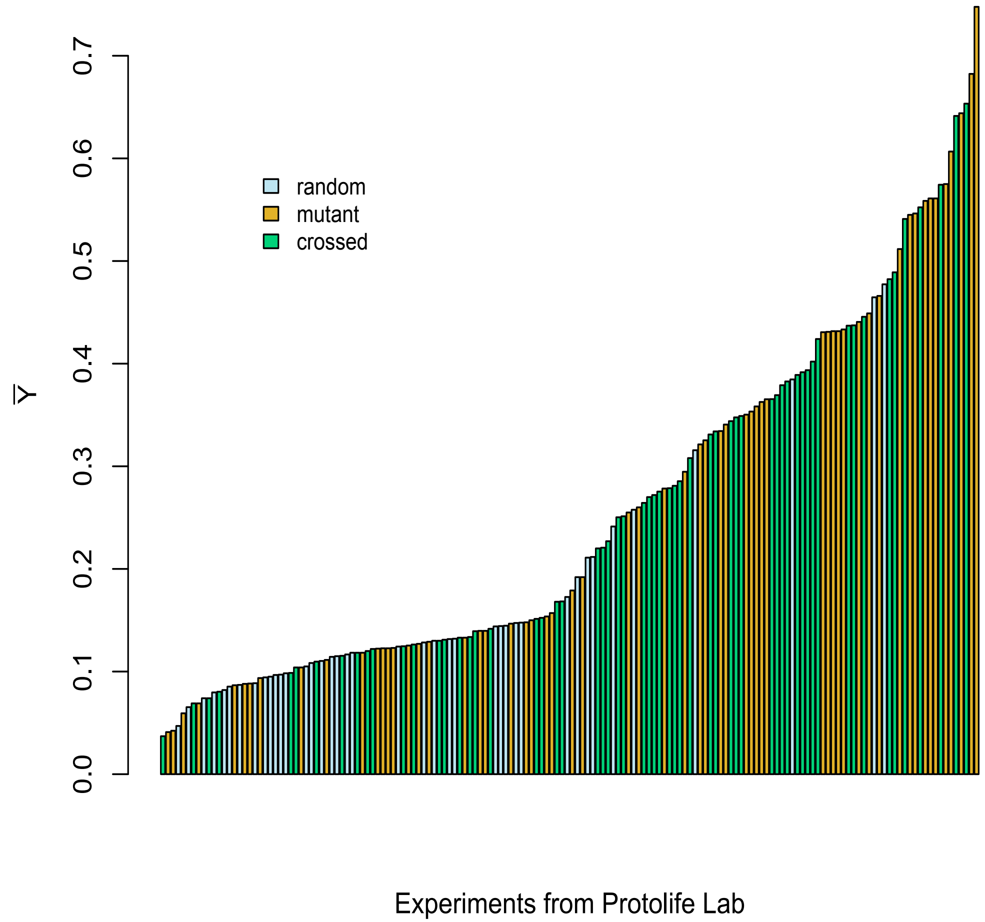
D^1_3 ← Mutate (D^1_2)

D^2 ← ……..

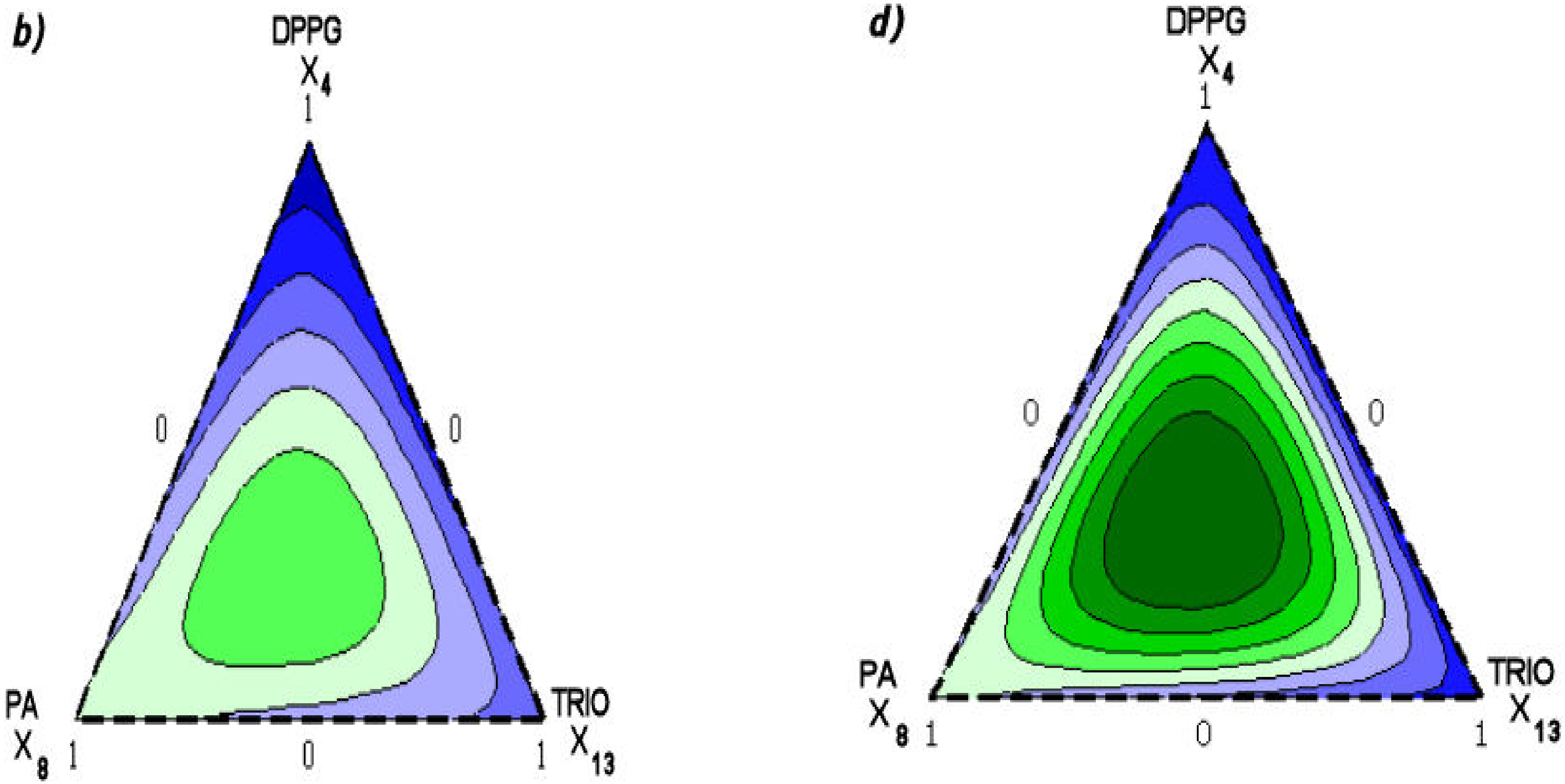
Conduct the experimentation testing each member of D^1_3 endwhile

....

Results from the GA design on **real** experiments

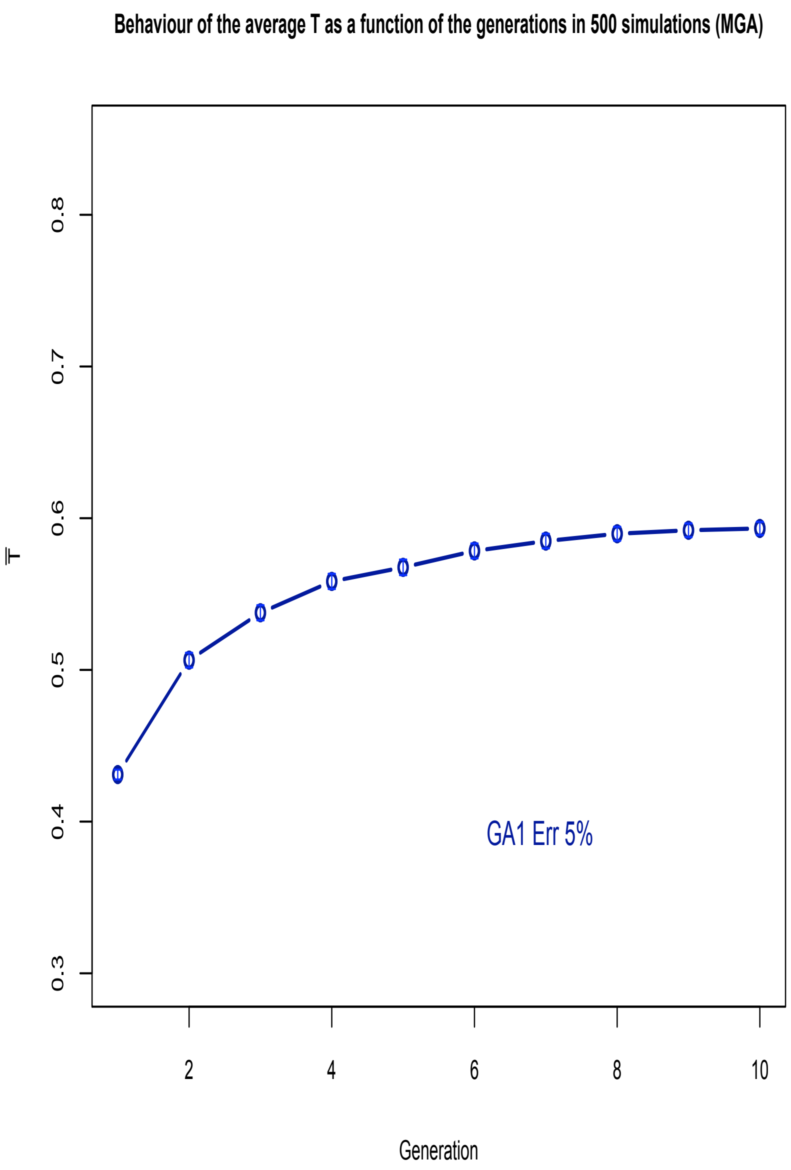


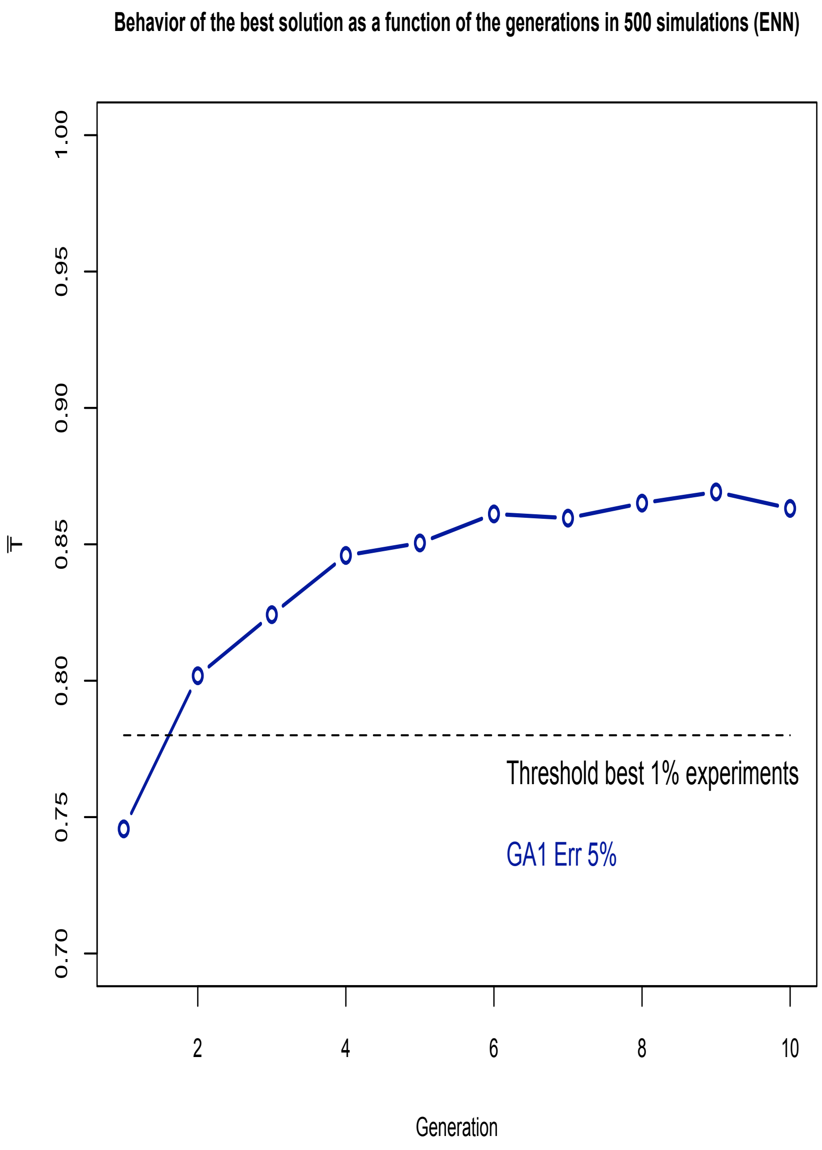
Contour plots



Forlin, Poli, De March, Packard, Serra, 2008, Chemometrics.

Simulated experiments





Statistical models in the evolution?

Can statistical models make a difference in the evolutionary process?

At any generation of experiments, we can build *statistical models* on the dataset and uncover information not considered by the genetic operators.

This *information* can then be embedded in the generating process of the next generation of experiments, providing "more intelligent data"

Finding information and communicating it...

The Statistical Evolutionary Experimental Design

A simulation platform for comparing different evolutionary procedures where ***models*** *lead the evolution of the design.*

The Model Based Genetic Algorithm Design (MGA)

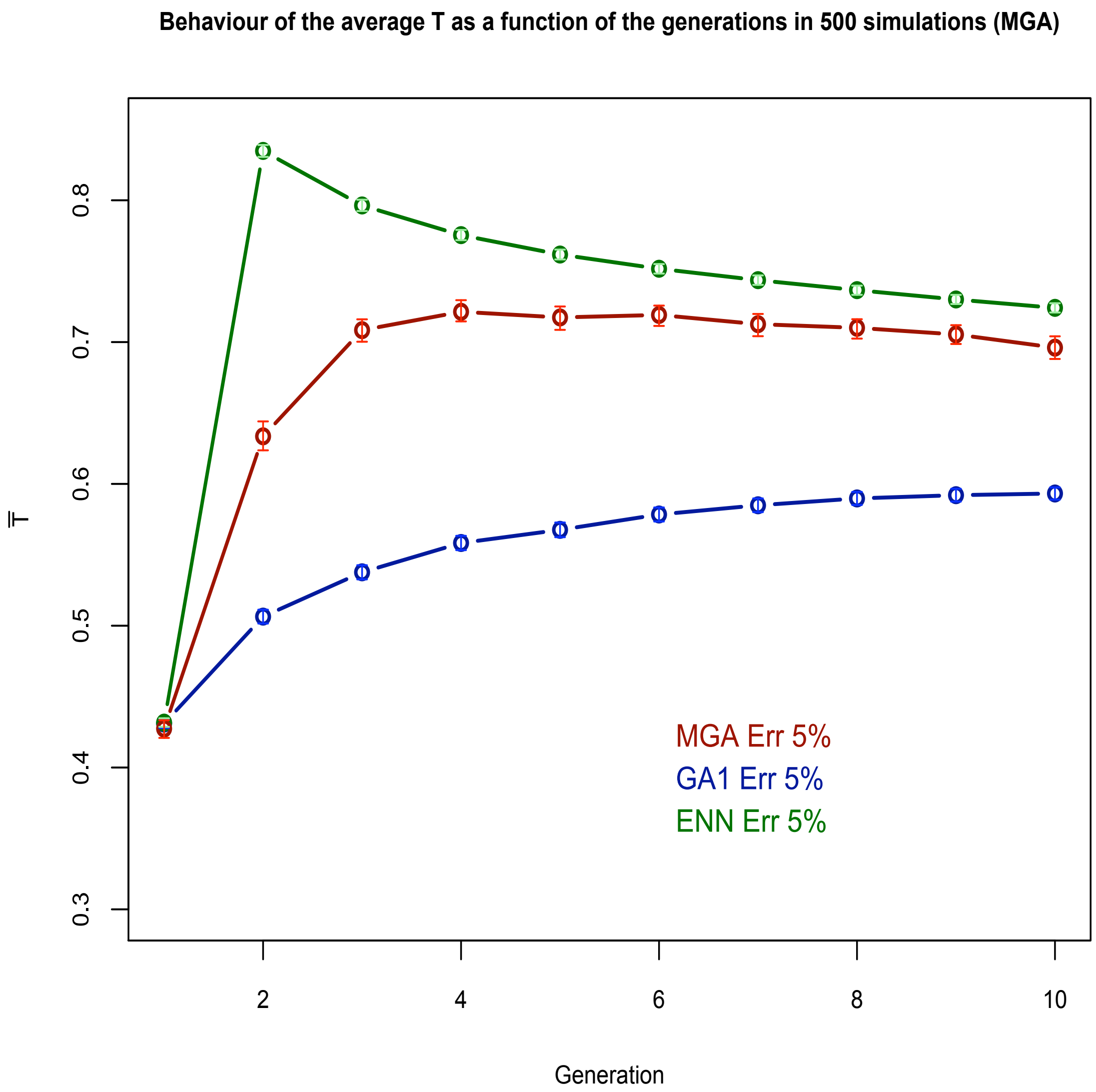
The Evolutionary Neural Networks Design (ENN)

The Evolutionary Bayesian Network Design (EBN)

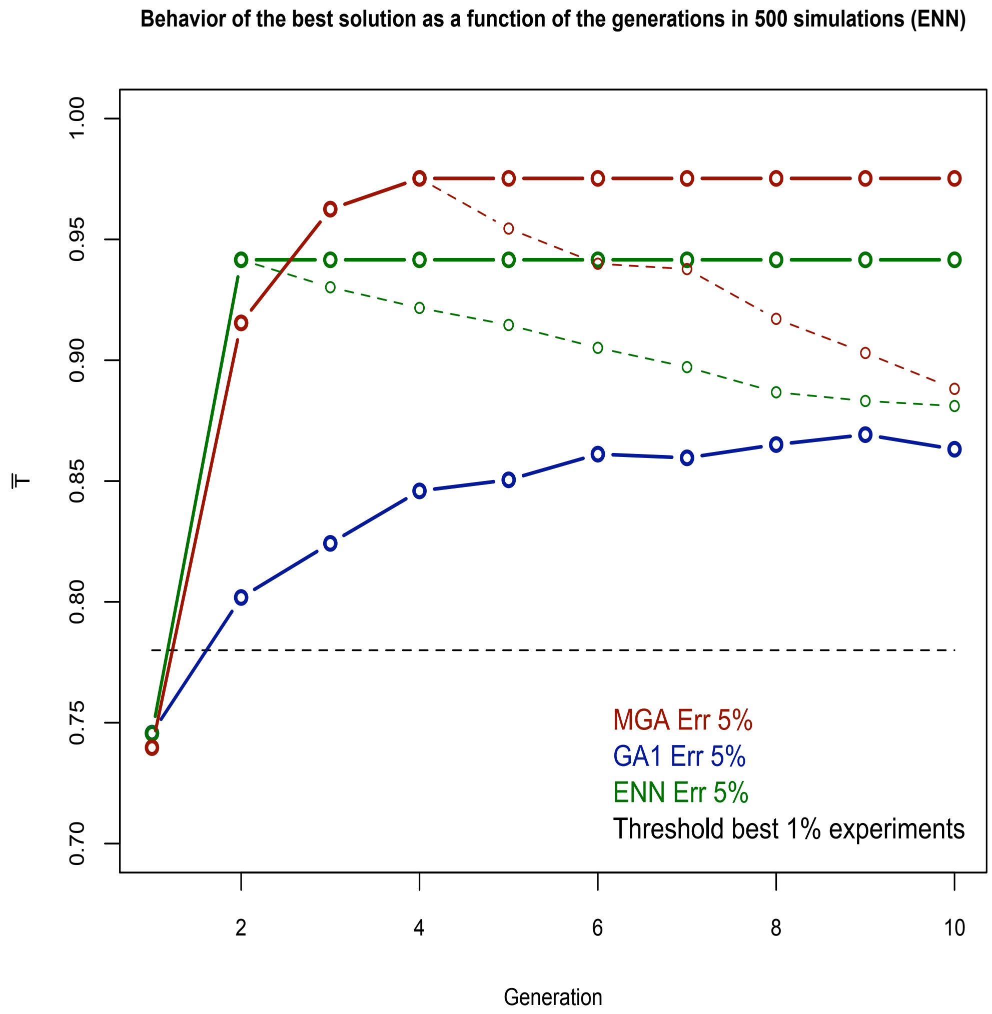
and Ant Colony Design

Particle Swarm Design

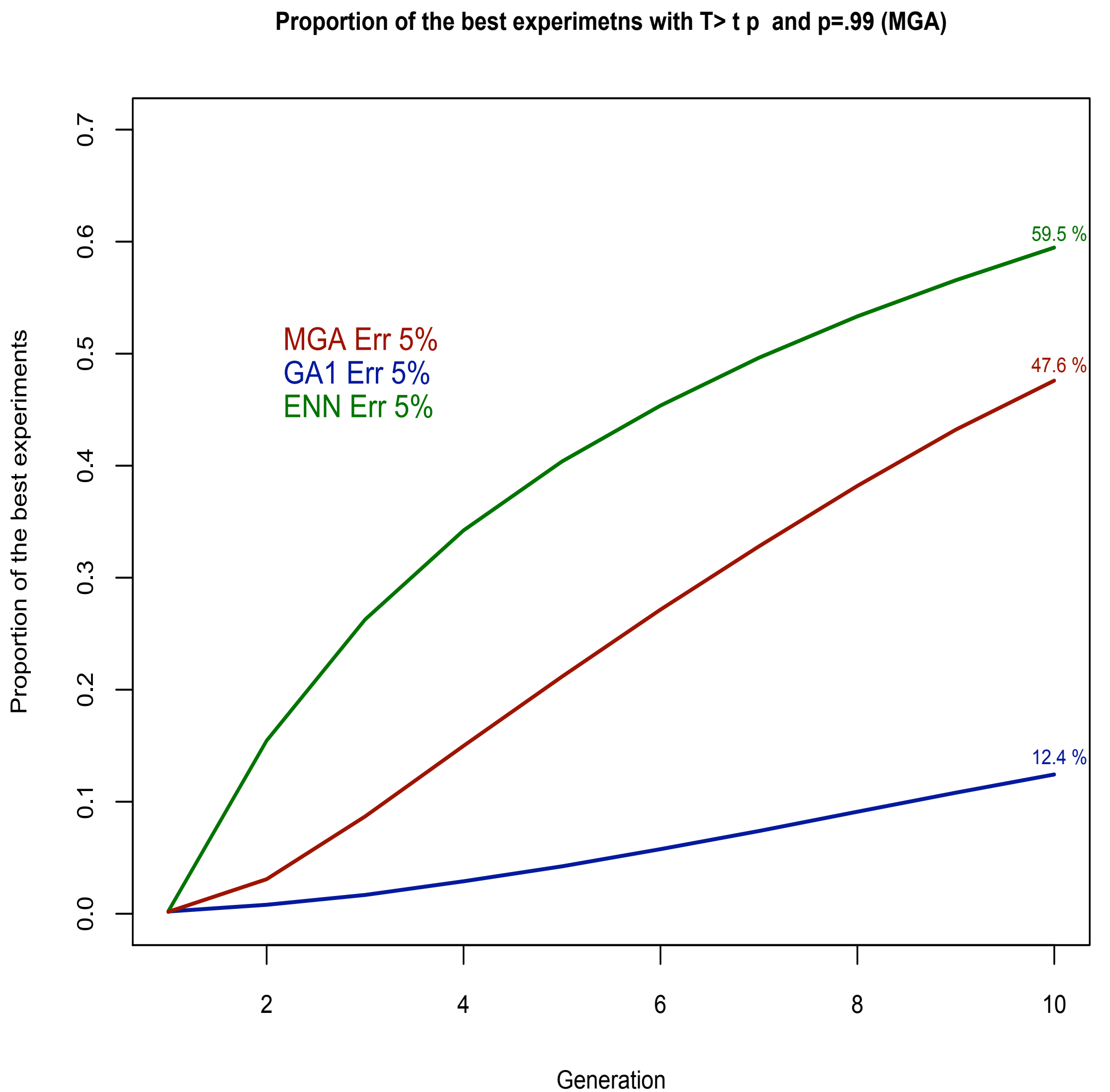
The average experimental response



The best experimental response



Proportion of the best experiments in the class of the 1% best experiments



Conclusions

The evolutionary approach can successfully address the problem of HDHT

The statistical models can lead the evolutionary process generating "more intelligent data"

The Statistical Evolutionary Experimental Designs (SEEDS) can derive designs which are

cheap,

fast

and effective.

D. Slanzi, D. De March, I. Poli*, Probabilistic graphical models in high dimensional systems*, 2009.

D. De March, D. Slanzi, I. Poli, *Evolutionary Algorithms for Complex Experimental Designs,* 2009.

D. De March, M. Forlin, D. Slanzi, I. Poli, *An evolutionary predictive approach to design high dimensional experiments*, 2009.

M. Forlin, *A computational design for high dimensional biochemical experiments*, 2009.

A. Pepelyshev, Poli, I. , Melas, V., *Uniform coverage designs for mixture experiments,* 2009.

D. Slanzi, D. De March, I. Poli, *Evolutionary Probabilistic Graphical Models in High Dimensional Data Analysis,* 2009

I. Poli, Evolutionary Designs of Experiments, 2010.

Thanks

to the research group at ECLT,

to EU for the **PACE** project, and
to Fondazione di Venezia for the **DICE** project.

to the Dept. of Statistics UNIVE,
to Protolife Laboratory,

to you !!!