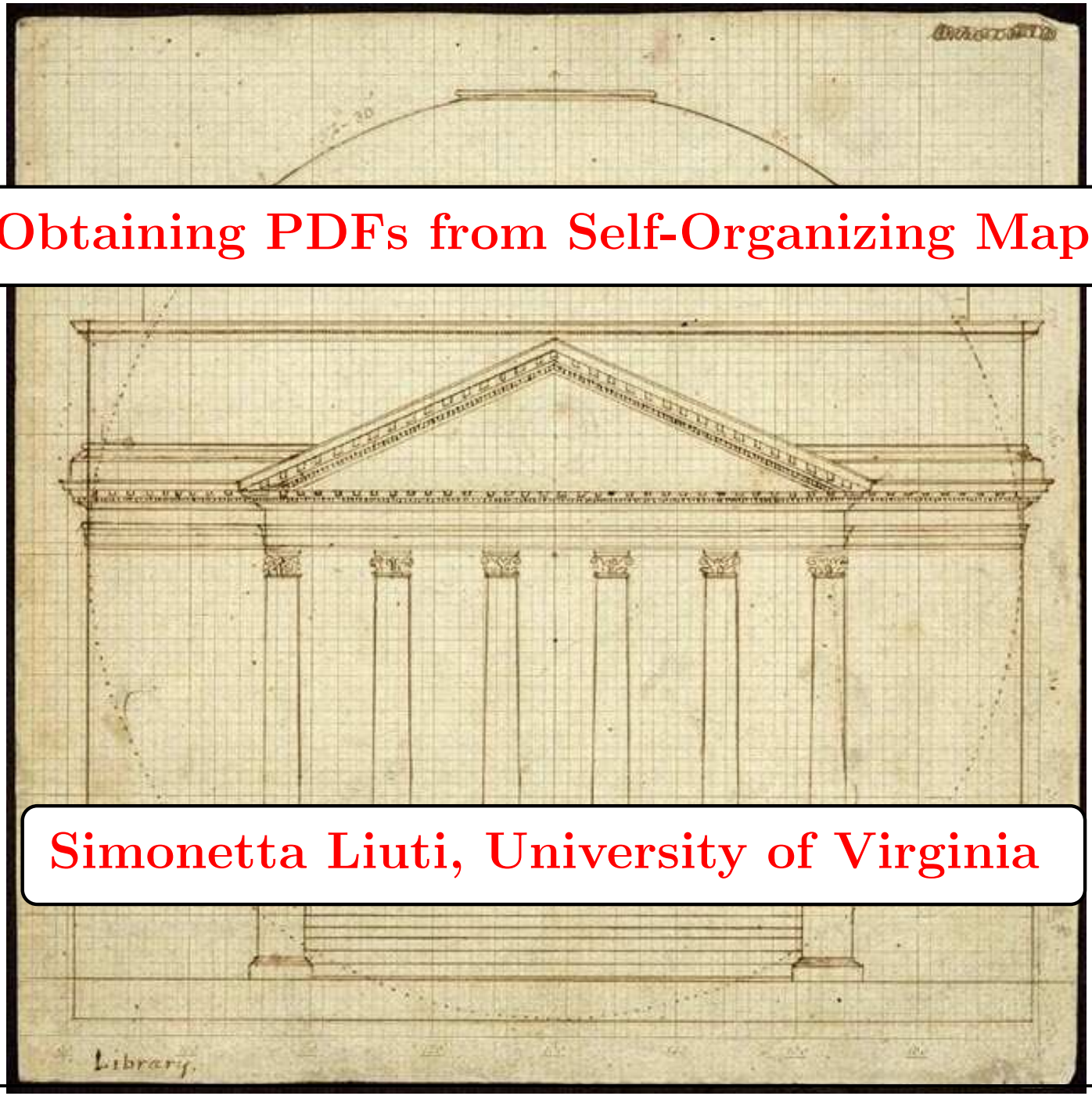


Obtaining PDFs from Self-Organizing Maps

Simonetta Liuti, University of Virginia

Motivation

" a fundamental measurement" and ... "a necessary evil" ...

Jon Pumplin, DIS 2005 on PDFs

"a disease" that would lead scientists into computerized daydreams tangential to the task at hand.

Richard Feynman

An Interdisciplinary Project with CS Colleagues using Dynamic Data Driven Application Systems (DDDAS)

"... a new direction for applications/simulations and measurement methodology"

Yannick Loitière (computer science)

Heli Honkanen (physics)

David Brogan (computer science)

Four Essential Open Questions

1. Parametrization dependence $\Rightarrow$ bias from the functional forms chosen at the initial scale, Q_o^2.
2. Theoretical assumptions $\Rightarrow$ s, $\bar{s}$, c quark content, PQCD evolution (NNLO, large x resummation, non-linearity, ...)
3. Error analysis $\Rightarrow$ ambiguities in the usage of data from different experiments
4. Choice of statistical estimator $\Rightarrow$ global χ^2 is not adequate as shown by inconsistencies from different data sets

Proposed Method

Different research groups have been focusing on each of the four open questions

Instead of addressing points separately

$\Rightarrow$ question and develop alternative to

***"fully automated"* procedure**

1. What does a "bias free" approach buy us?

   - Develop an alternative to standard NN approach relating the continuous input PDFs to finite number of measurements.

2. Theoretical uncertainty:

- Allow for interplay between user and algorithm
- DGLAP is part of training procedure

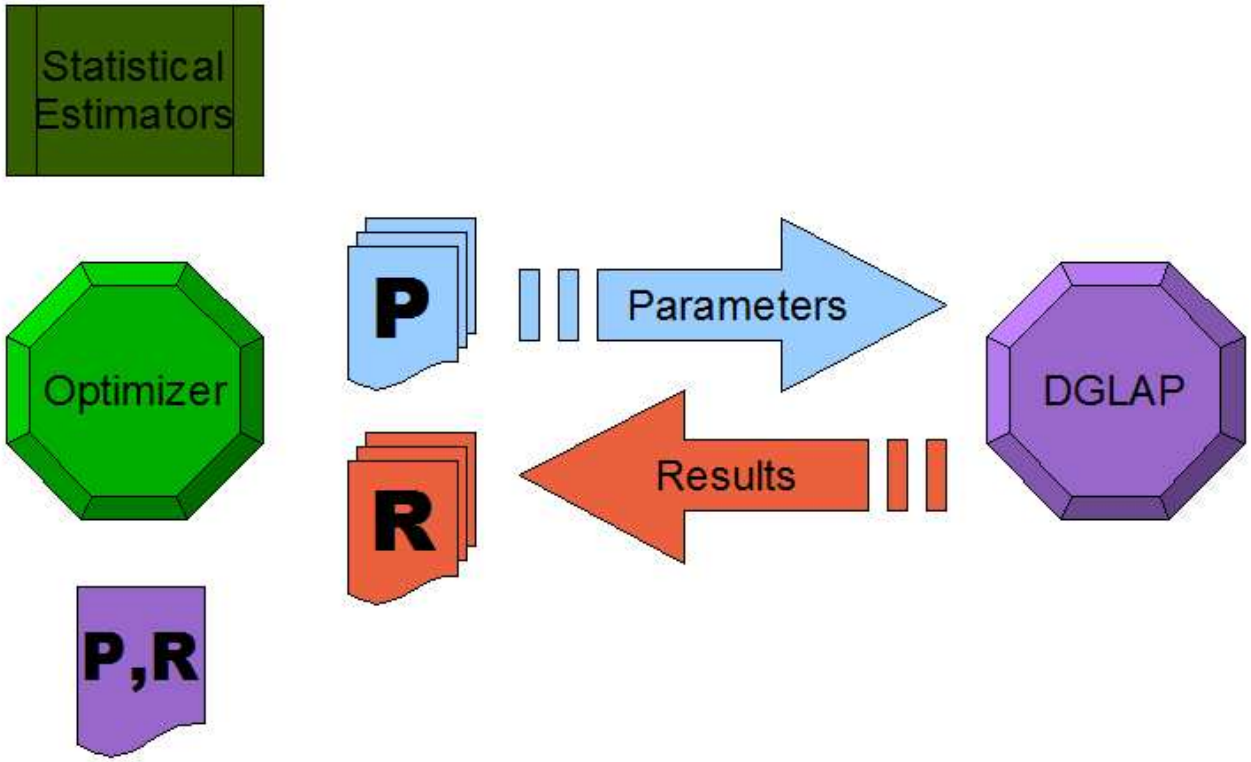
3. Merging different experimental data sets

- Allow for interplay between user and algorithm

4. χ^2 ?

- Require flexibility, *e.g.* allow for introduction of different estimators

Standard Approaches: Automated Procedure



Computational point of view

Problems inherent to fitting procedures (both global fits and NN):

- Sparse Data: the problem is underconstrained from the point of view of an unbiased function fit $\Rightarrow$ as a consequence, the "systematic uncertainty' is not under control
- High Dimensionality
- Non-Linearity $\Rightarrow$ Presence of local minima (?) do not allow to extrapolate from first derivatives.

$\Rightarrow$ ... **Introduce Self-Organizing Maps!**

What are Self-Organizing Maps?

T. Kohonen, Springer-Verlag, 1997

⋆ A SOM is an algorithm that maps in a topologically ordered way the training data onto a neural network.

⋆ The mapping proceeds by selecting the neuron, N_W, that best matches each data sample according to a metric, M_D.

⋆ Each neuron is represented in a two-dimensional grid, with coordinates: $\mathbf{x_i} \equiv (x_1, x_2)$.

⋆ A weighted average of each neuron, N_i in the grid to the data sample is then performed, where the weight, w_i is computed from the distance of N_i to N_W according to a metric, M_G, and a given neighborhood radius. M_G defines the topology of the grid.

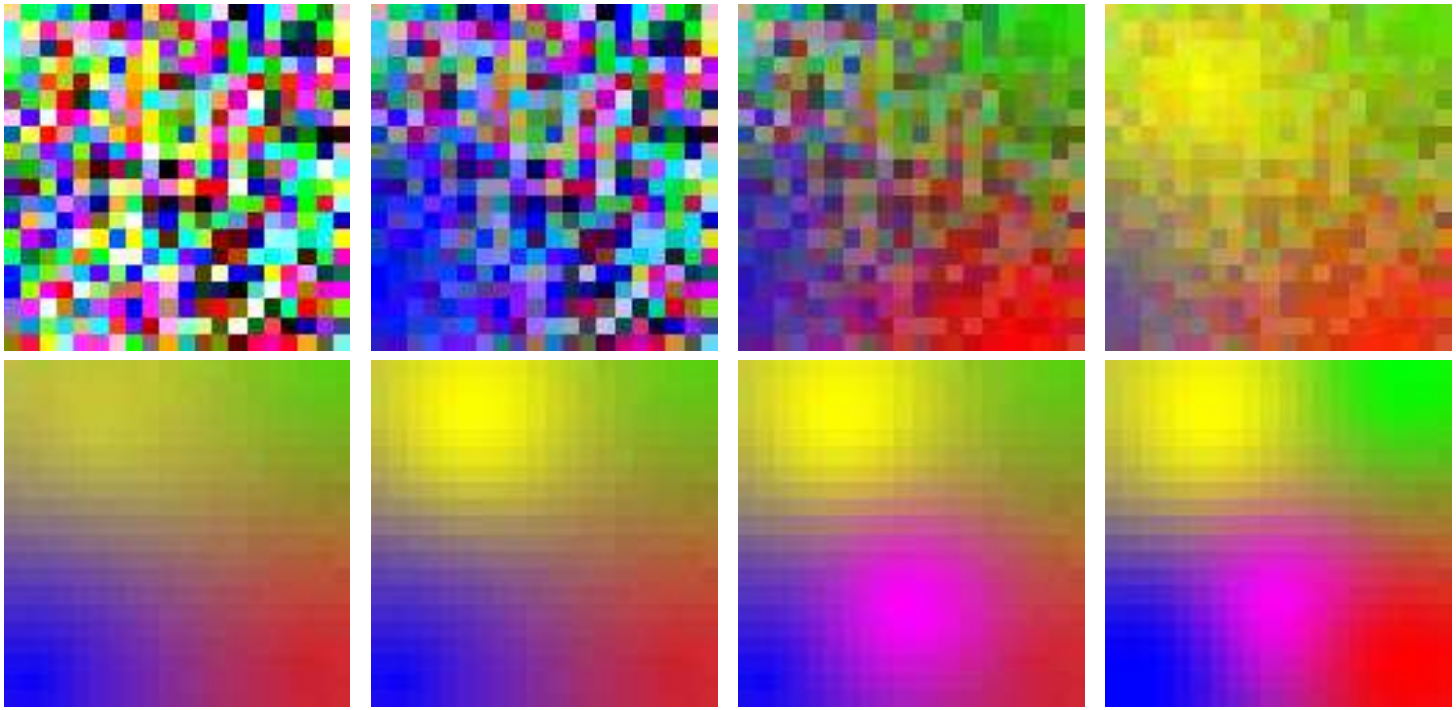
⋆ This procedure is iterated with smaller radii until it saturates.

Application to our case:

- The neurons are the PDFs, and the data are "synthetic data" (randomized samples of the original data).
- The metric M_G that defines the topology of the map is:

$$L_1(\mathbf{x}, \mathbf{y}) = \sum_{j=1,2} \mid x_j - y_j \mid$$

"Colors" Example



How do we apply all this to PDF parametrizations?

Automated version

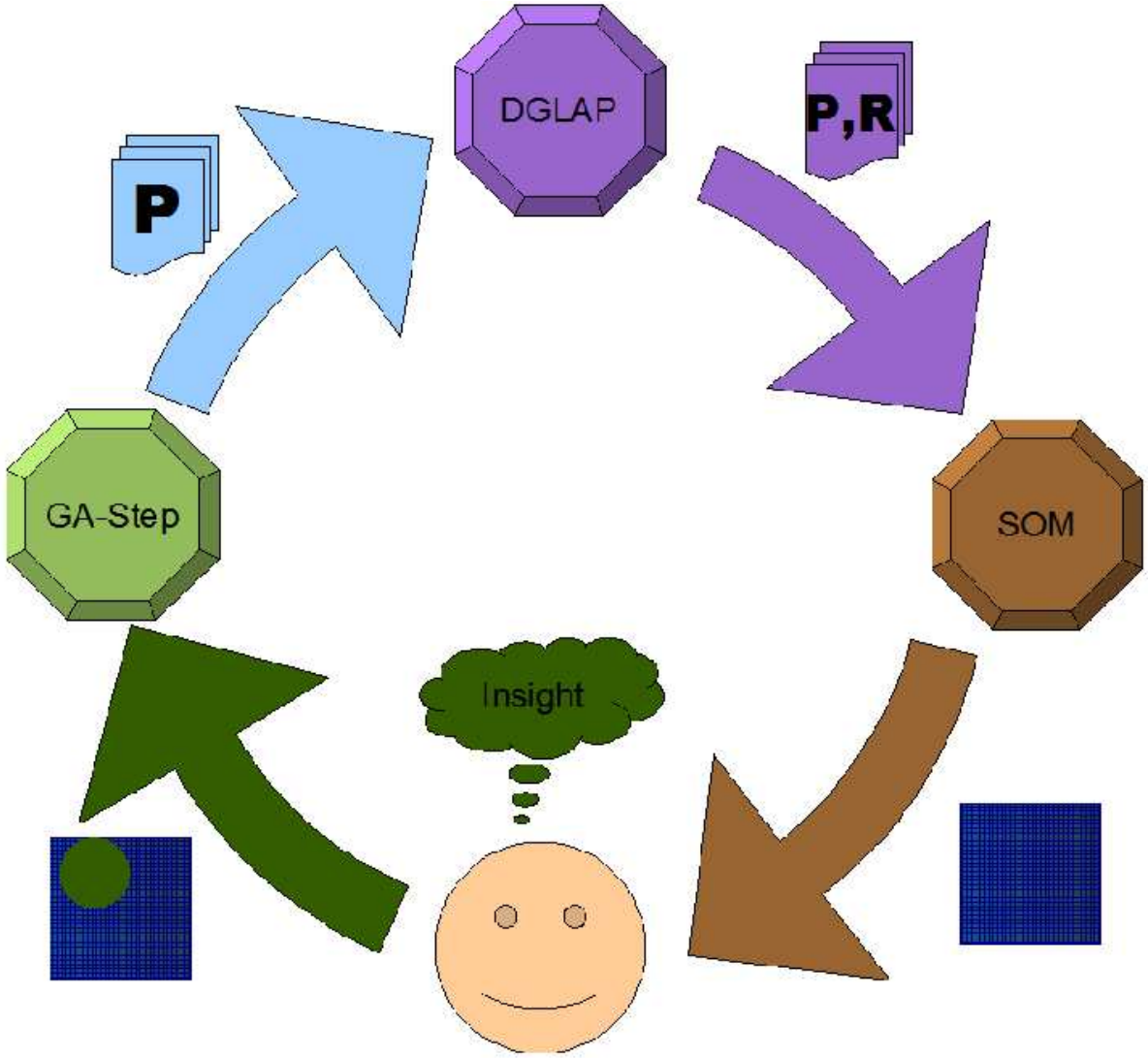
We compare each neuron with the experimental data according to a given criterion, *e.g.* χ^2 and select the best neurons.

Go back to training data and compare using M_D to the fully trained map.

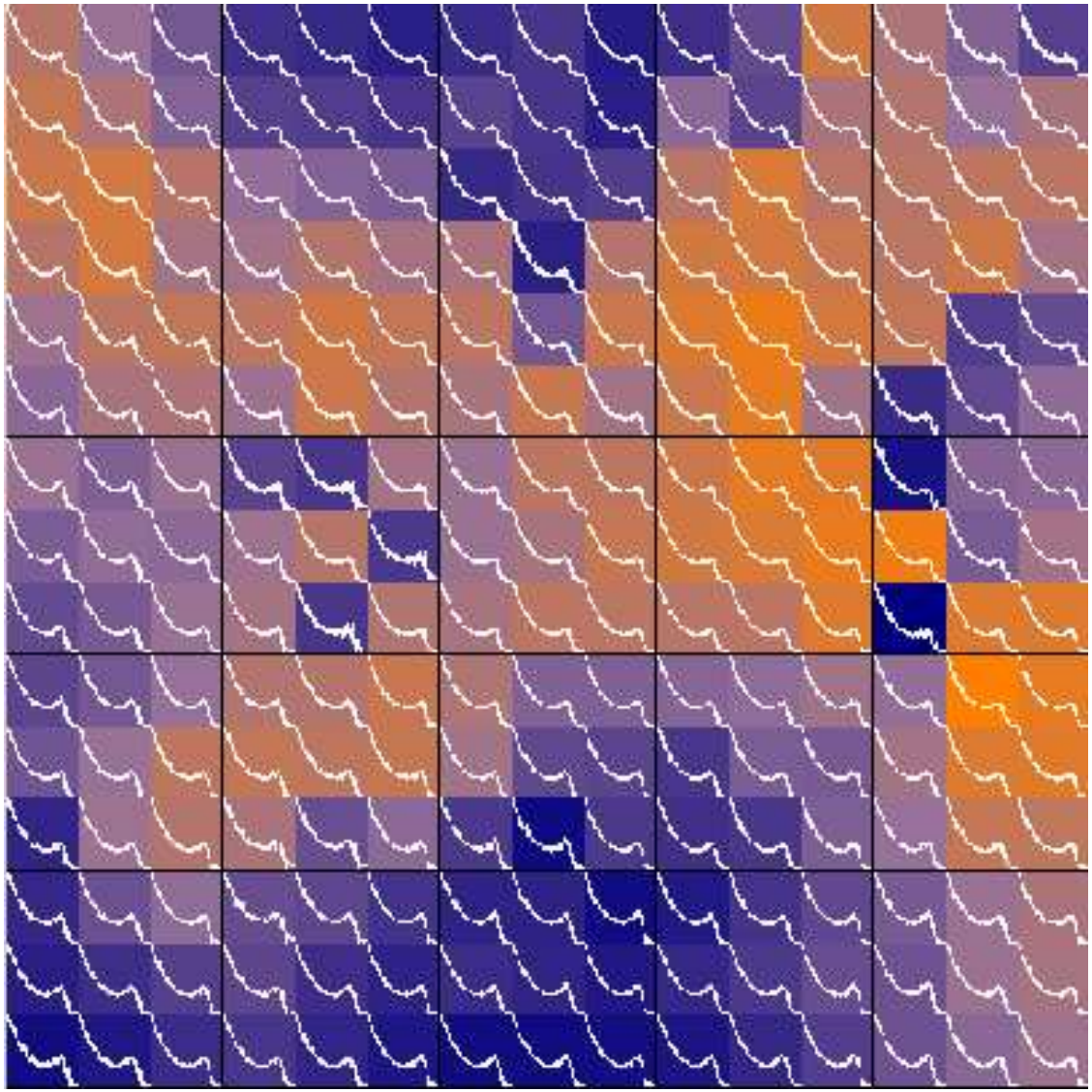
Each neuron is the center of a number of training data samples, center of the cluster. Compute the statistics of the PDFs in each cluster, generate new training data set.

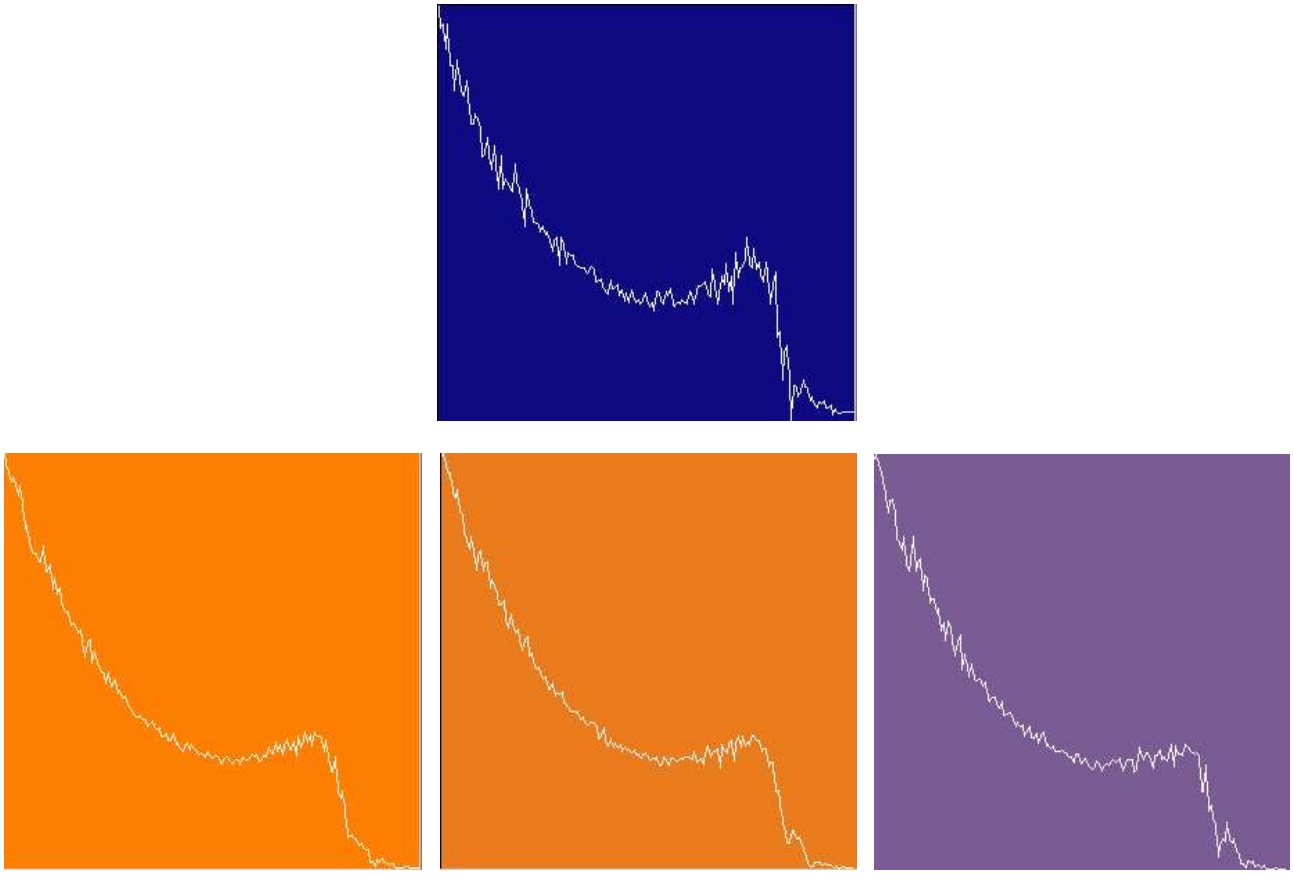
Interactive version (in progress...)

There is also the possibility of using the SOM interactively by selecting other features from the map without focusing on the "global" χ^2.

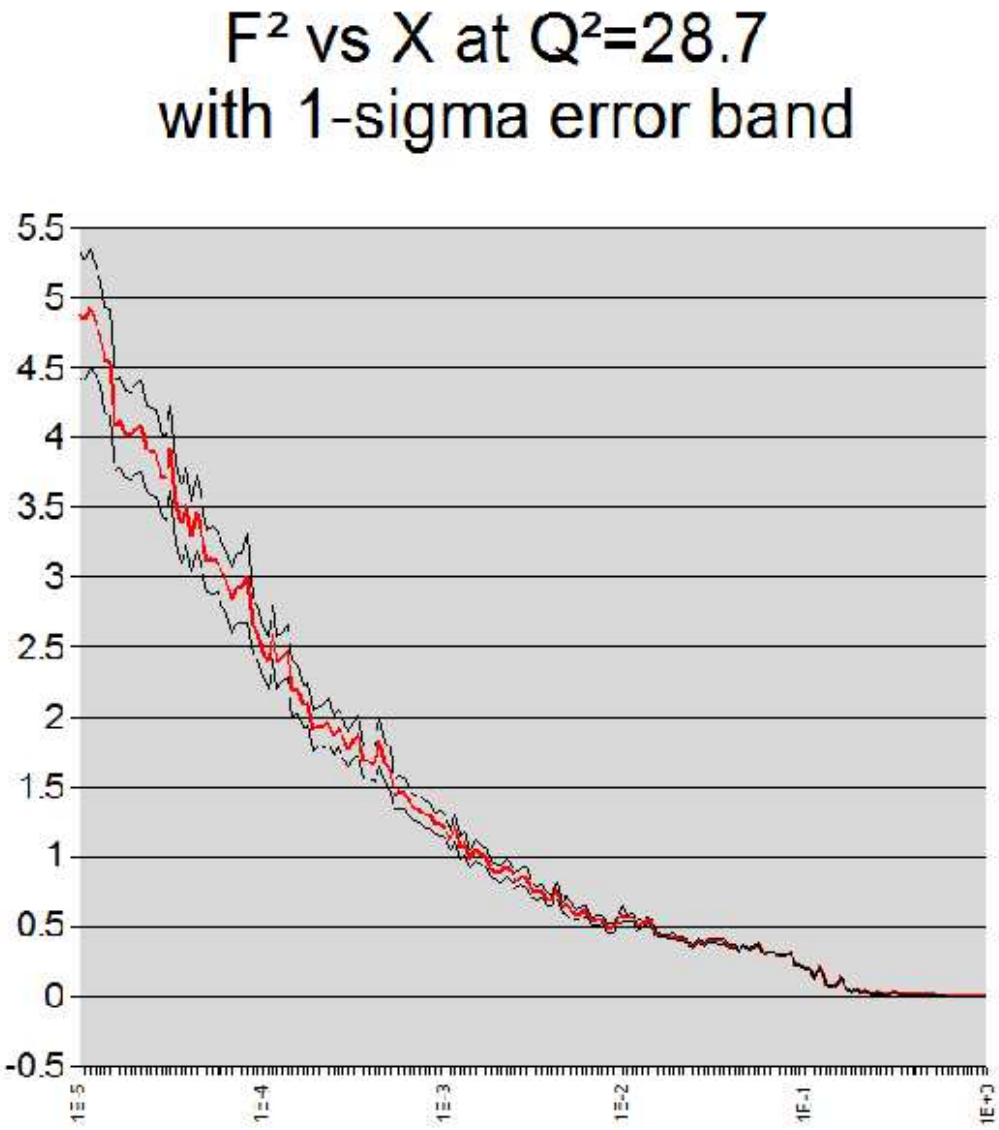

DGLAP
P,R
SOM
Insight
GA-Step
P

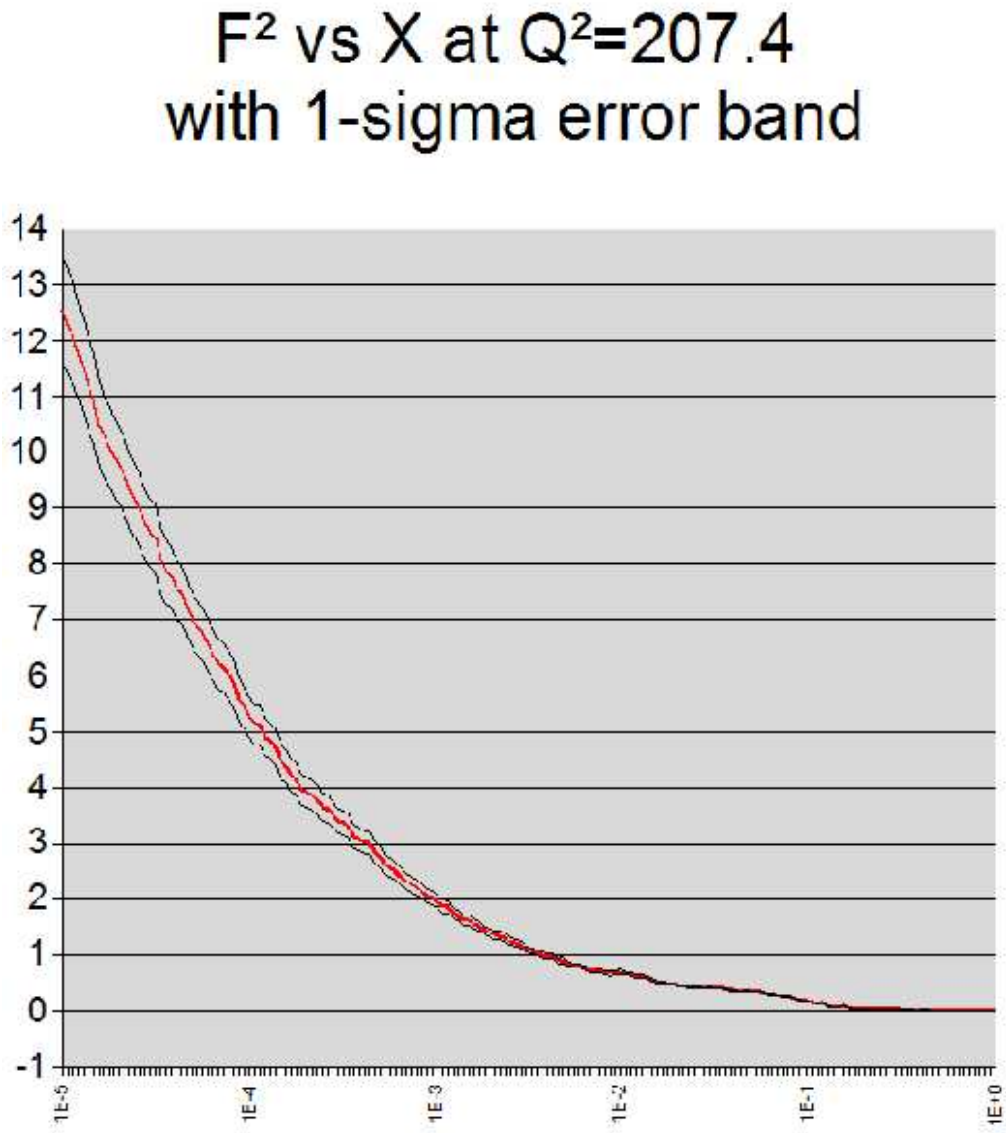
Results

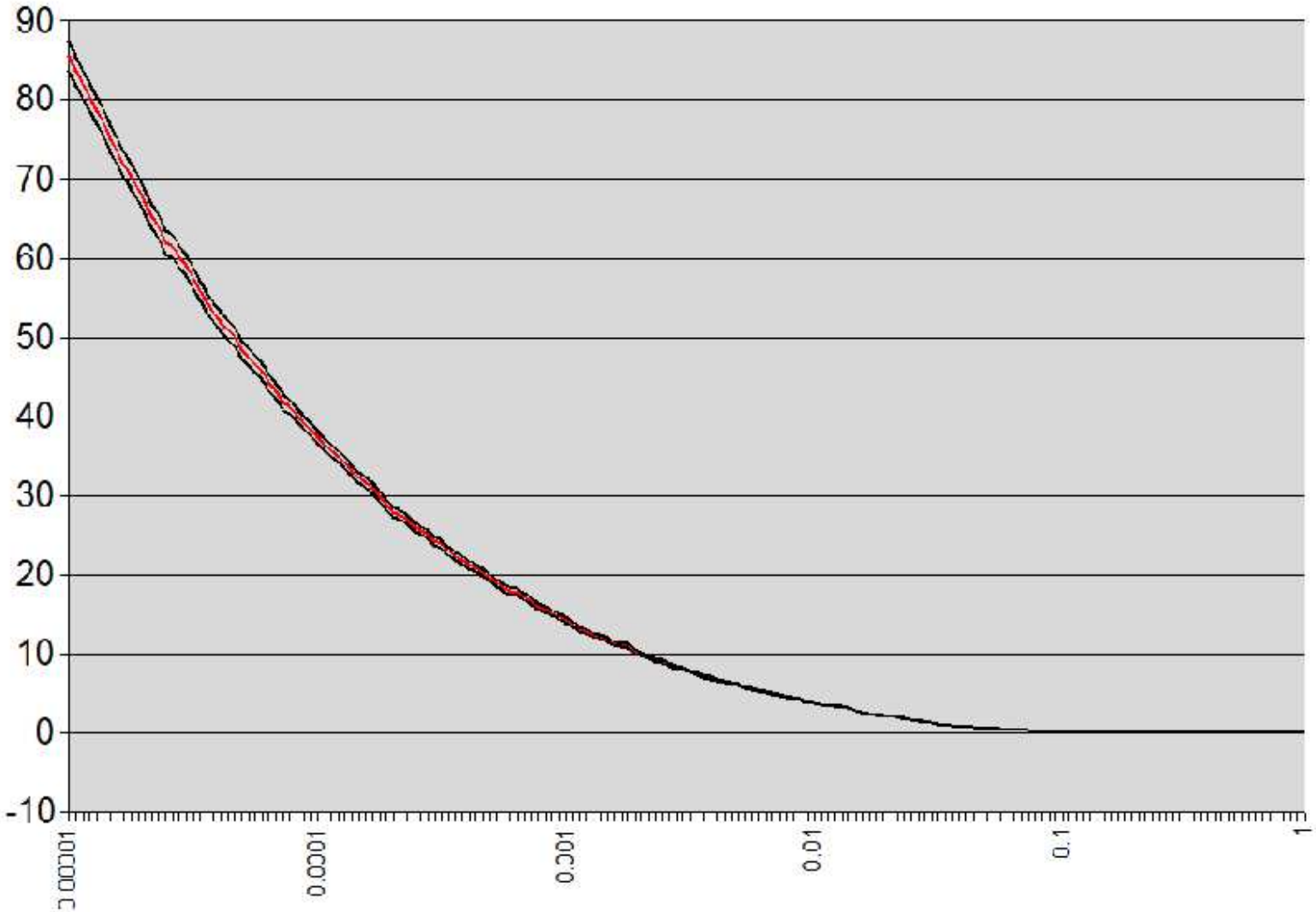




Color is only relative value of χ^2 to the experimental data. We used only DIS data.


F² vs X at Q²=28.7
with 1-sigma error band
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
1E-5
1E-4
1E-3
1E-2
1E-1
1E+0


F² vs X at Q²=207.4
with 1-sigma error band
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
-1
1E-5
1E-4
1E-3
1E-2
1E-1
1E+0


Gluons vs. X at Q²=28.7
90
80
70
60
50
40
30
20
10
0
-10
0.00001
0.0001
0.001
0.01
0.1
1

Conclusions and Outlook

- We proposed SOMPDFs, change of paradigm with respect to NNPDFs aimed at bringing "theory" back in the loop.
- SOMPDFs have the following additional advantages over generic Genetic Algorithms (NNPDFs):
  - Visualization
  - Dimensionality reduction $\Rightarrow$ helps idenifying the role of different parameters
  - Clustering: one can attempt to understand what features of PDFs produce given types of clustering
- We presented first results on $F_2(x, Q^2)$ and $g(x, Q^2)$ using global χ^2 and DIS data.
- Practical Goal: Currently addressing the question of implementing SOMPDFs in actual data analyses (LHC)
- Near Future: Extend to a number of different quantities: Nuclear PDFs, GPDs.