A Parametrization of PDFs based on Self-Organizing Maps

Simonetta Liuti
University of Virginia

JFC

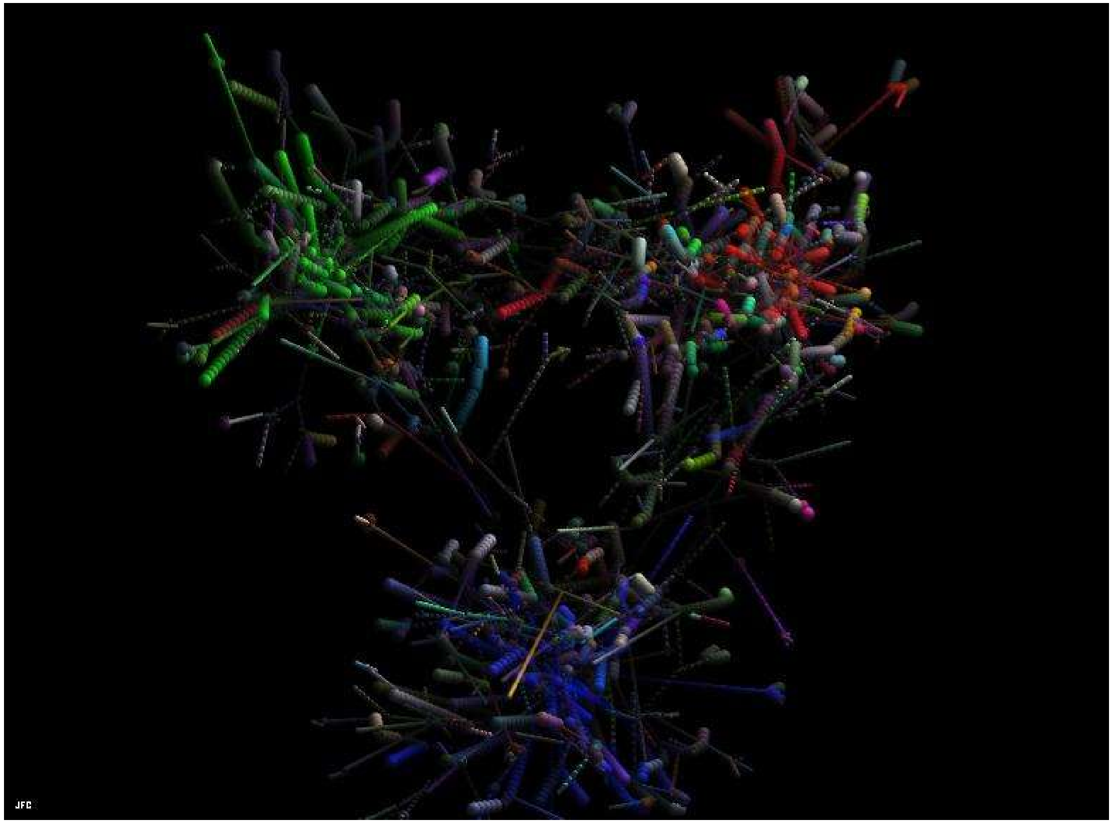
Collaborators

- Saeed Ahmad (graduate student, physics)
- Joe Carnahan (graduate student, CS)
- Heli Honkanen (post-doc, physics)
- Paul Reynolds (Co-PI, CS)
- Swadhin Taneja (graduate student, physics)

Outline

- Introduction to Self-Organizing Maps
- Algorithm
- SOMPDFs: Results
- Comparison with conventional methods and NNPDFs
- Conclusions and Outlook

A new approach: Self-Organizing Maps



A rather large and diverse set of observations is produced that needs to be specifically detected, and compared to patterns predicted theoretically for different momentum, spin, and spatial configurations of the constituents.

Conventional approaches tend to explain such patterns in terms of miscroscopic properties of the theory → forces between two particles

Introduction to Self-Organizing Maps 2

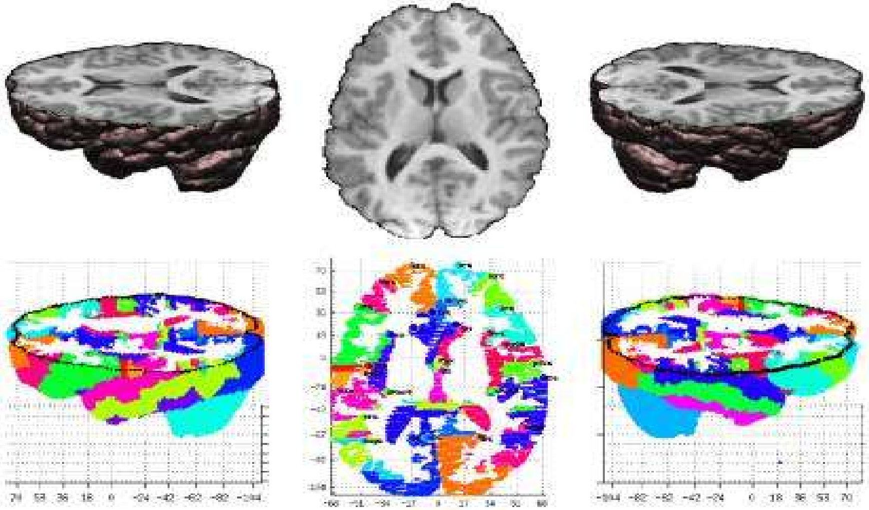
- Idea! Attack the problem from a different perspective

- Study the behavior of multi-particle systems as they evolve from a large and varied number of initial conditions

- Goal can be reached with modern high performance computing

Introduction to Self-Organizing Maps 3

Self-Organizing Maps (SOM) were derived as a mathematical model of these configurations (T. Kohonen, 1981)



Inspired by patterns in cerebral cortex: the detailed topographical order of the neural connections (synapses) form localized maps.

Brain maps are determined both genetically and by experience

"experience" = some projections – growth of axons of neural cells – are developed or stunted with respect to others, different cells are recruited for different tasks

Principles:

1) The neurons behave according to a form of unsupervised self-organization

2) The representation of knowledge assumes the form of a map geometrically organized over the brain so that similar learning functions are associated to adjacent areas

2. Algorithm

Working of SOM

Each cell (neuron) is sensitized to a different domain of vectors:
cell acts as decoder of domain

↓

Initialization → Input vector of dimension **“n”** associated to cell **“i”**:

$$V_i = \left[v_i^{(1)}, \ldots, v_i^{(n)} \right]$$

V_i is given spatial coordinates that define the geometry/topology of a 2D map

Training → Input data:

$$x = \left[\xi^{(1)}, \ldots, \xi^{(n)} \right]$$

x compared to V_i 's with “similarity” metric(L1):

$$|| x - m_i ||$$

(Aggawal et al., 2000)

Location of best match “winner” gives location of response (active cell, all others are passive)

Learning (updating) → cells $\mathbf{V_i}$ that are close up to a certain distance activate each other to “learn” from **x**

...in formulae:

$$V_i(t+1) = V_i(t) + h_{ci}(t)\left[x(t) - V_i(t)\right]$$

t = iteration

c = “winner” cell

i = cell

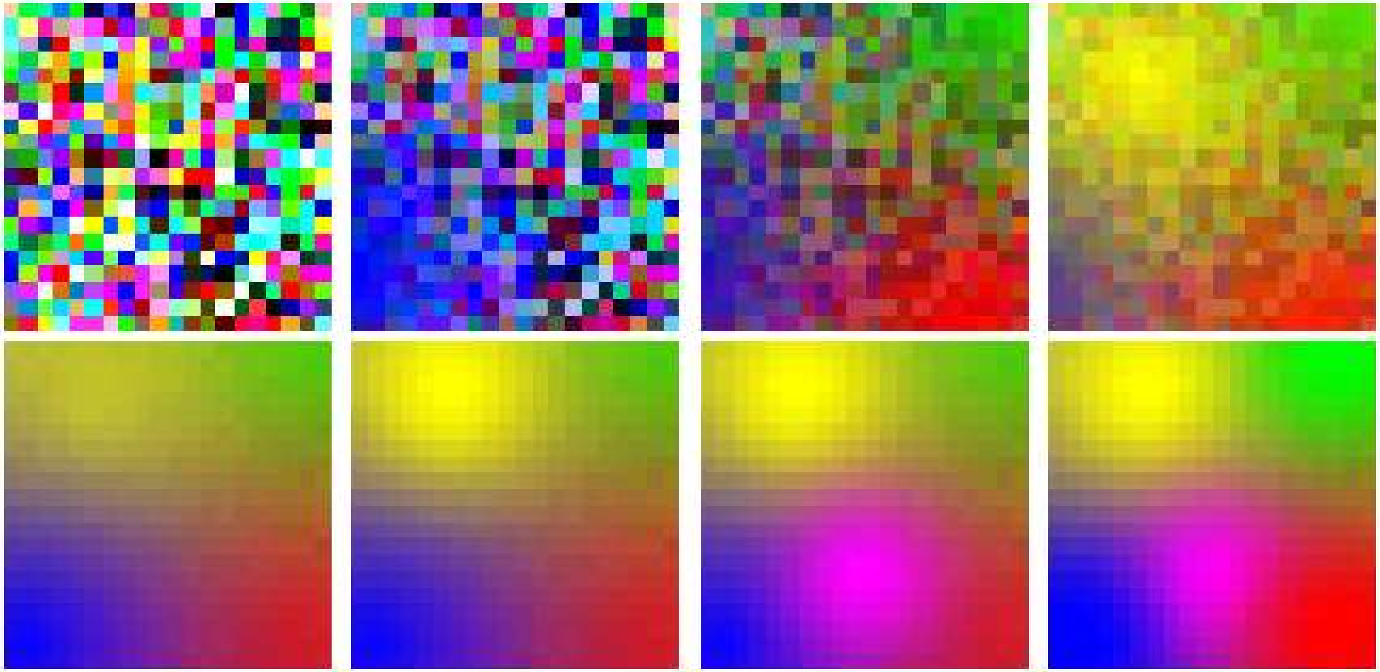
$$h_{ci}(t) \equiv h_{ci}(t, \| r_c - r_i \|)$$

$$h_{ci}(t) = w(t)e^{[-M(r_c,t)^2/r]}$$

$0<w(t)<1$

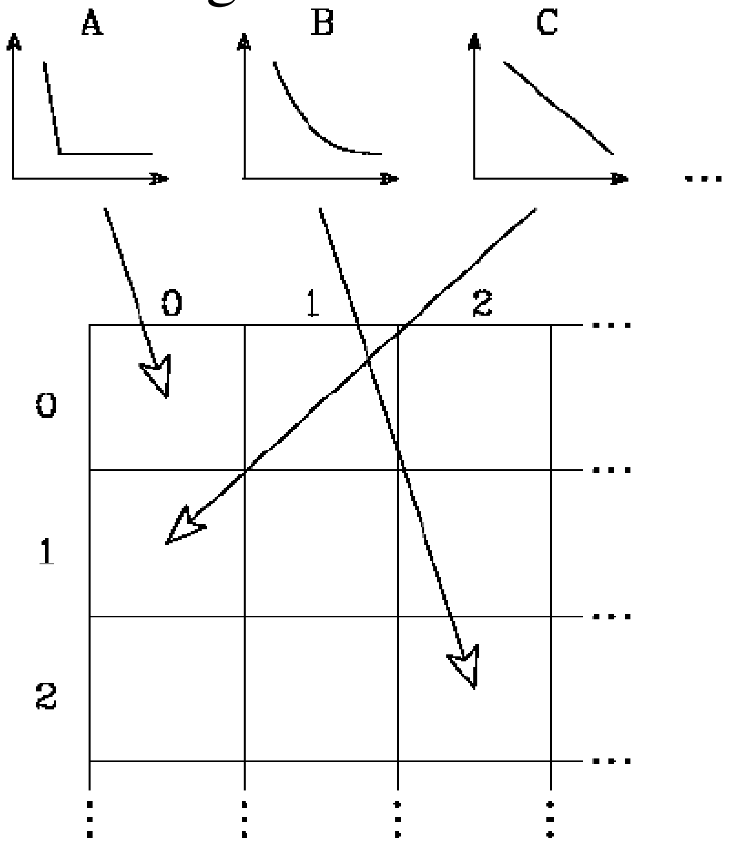
Example 1

“Colors” Example

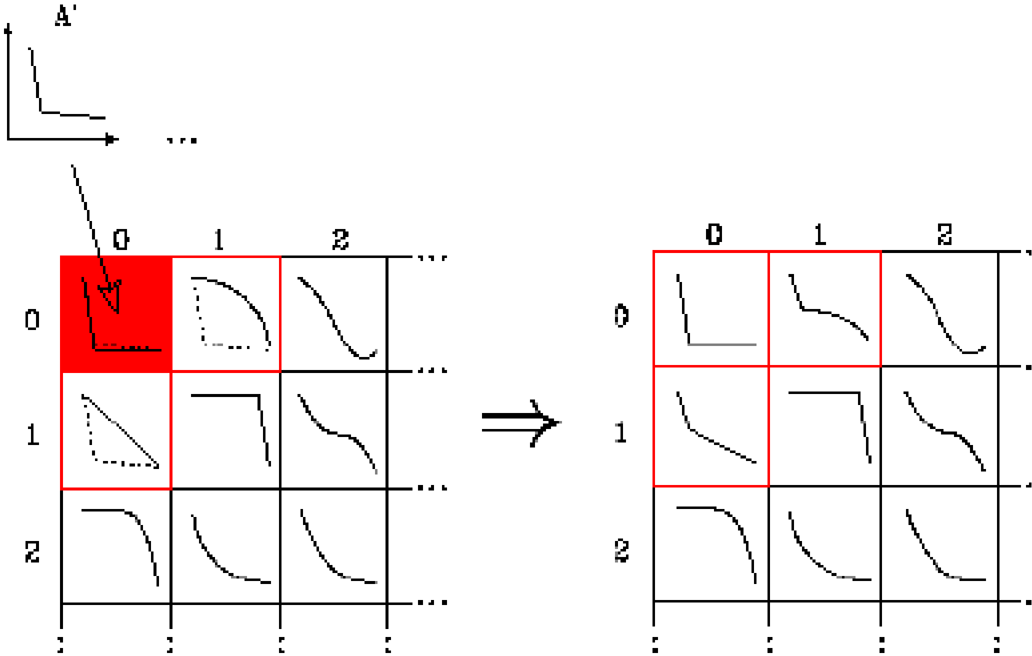


Example 2: Updating

Training

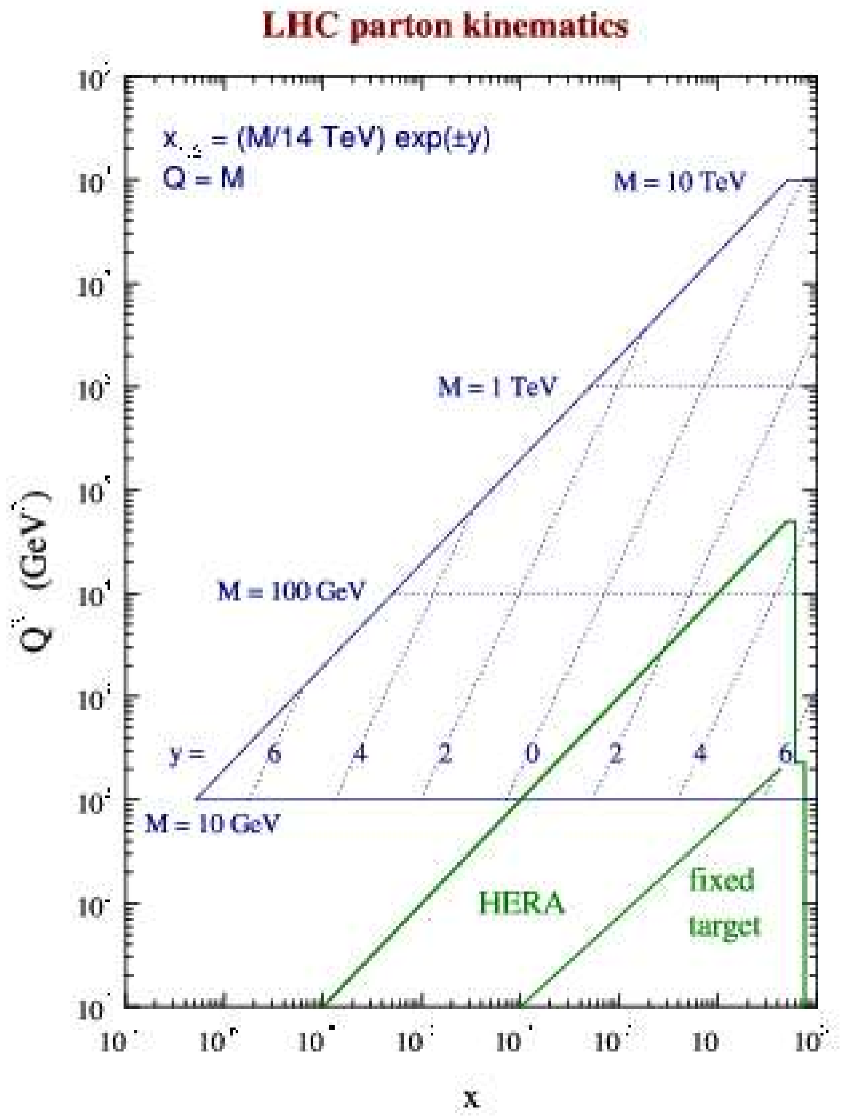


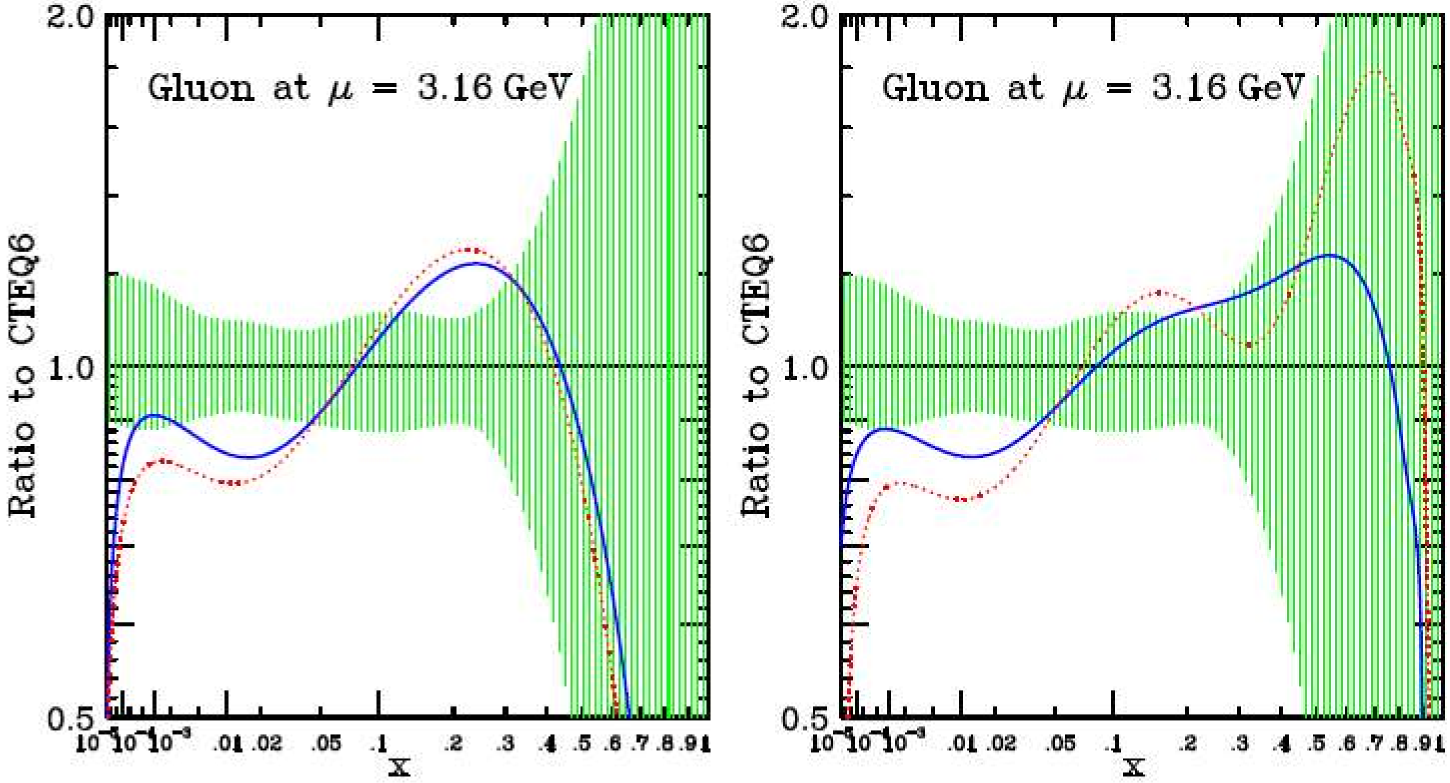
Initialization and Training



Learning

Application to PDF fitting





hep-ph/0201195

In a nutshell

(1) PDFs are clustered in a map according to similarity
(either among PDFs, or observables: structure functions)

(2) PDFs are identified with the code-vectors (decoders)

(3) Map vectors are updated by averaging the data samples
clustered within a neighborhood of the function to be updated

In more detail...

- We cluster stochastically generated PDFs according to the chosen similarity criterion and use the statistical characteristics of the clusters that best match the experimental data, χ^2, to produce a new generation of PDFs and thus guide the fitting process

- We use no functional form for PDFs but use existing distributions to establish an initial range for the GA-type analysis

- Our parameters are the values of PDFs at the initial scale for each flavour at each value of x where the data exist

SOMPDF algorithm

1. Randomly generate some PDFs
2. Smooth and normalize them
3. Cluster them in a SOM
4. Select some of the clusters (e.g based on χ^2) and prepare new random generators
5. Go to 1

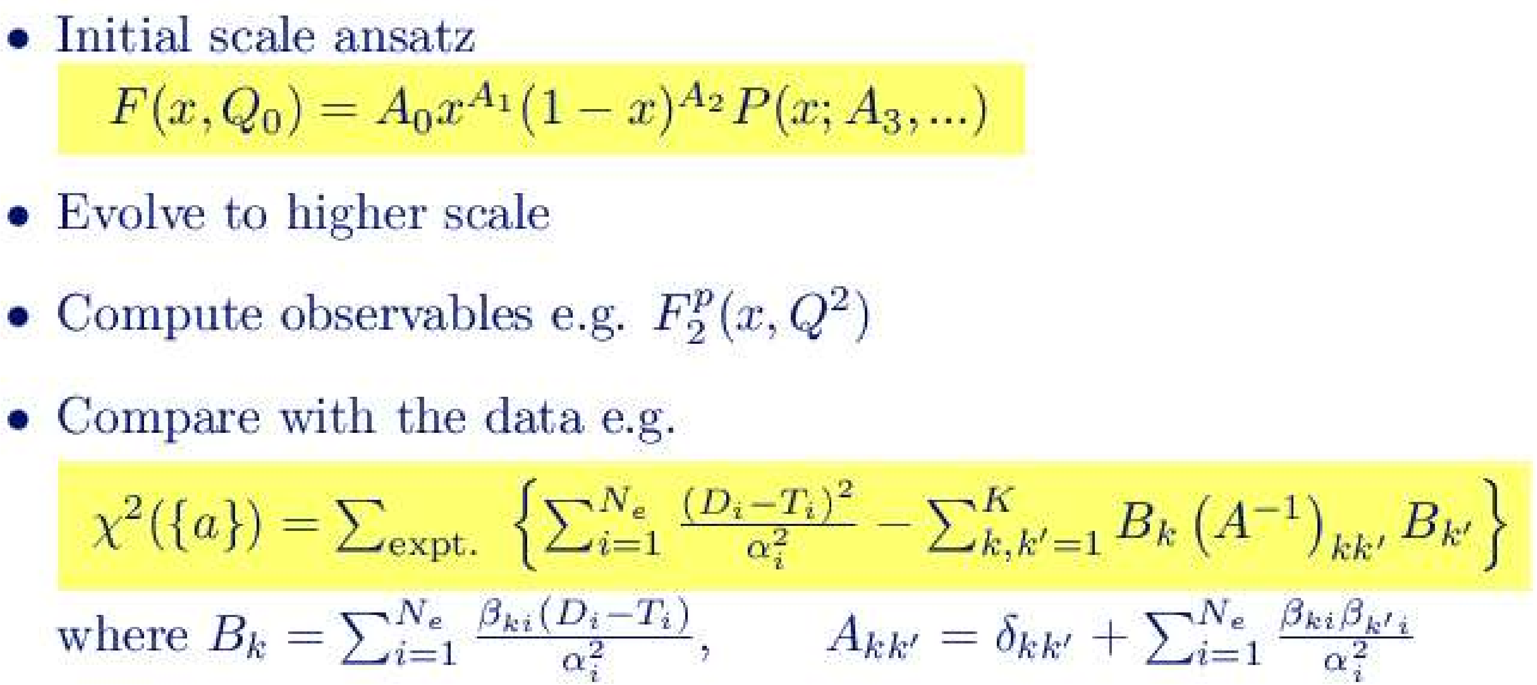
Add more complexity:

3.′ Also produce pseudoPDFs when generating the map

3.″ Insert results from the previous generation into the map if χ^2 is good enough (*elistist selection*)

6. Keep the original generators in the mix

Advantages with respect to “conventional way”:



- Initial scale ansatz

$$F(x, Q_0) = A_0 x^{A_1}(1-x)^{A_2} P(x; A_3, \ldots)$$

- Evolve to higher scale
- Compute observables e.g. $F_2^p(x, Q^2)$
- Compare with the data e.g.

$$\chi^2(\{a\}) = \sum_{\mathrm{expt.}} \left\{ \sum_{i=1}^{N_e} \frac{(D_i - T_i)^2}{\alpha_i^2} - \sum_{k,k'=1}^{K} B_k \left(A^{-1}\right)_{kk'} B_{k'} \right\}$$

where $B_k = \sum_{i=1}^{N_e} \frac{\beta_{ki}(D_i - T_i)}{\alpha_i^2}$, $\qquad A_{kk'} = \delta_{kk'} + \sum_{i=1}^{N_e} \frac{\beta_{ki}\beta_{k'i}}{\alpha_i^2}$

Similarly to NNPDFs we do not depend on a functional form, the “initial bias”, and we can define a faithful estimate of the uncertainty

Advantages over NNPDFs

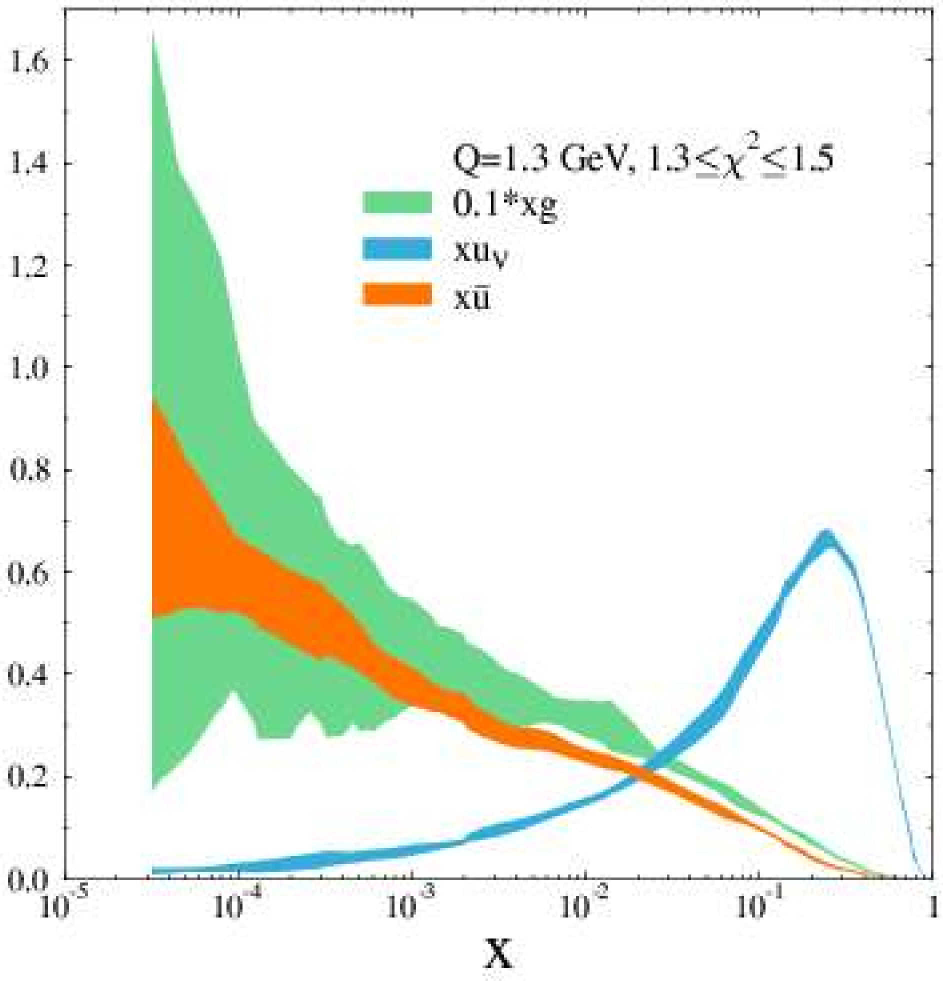
Mechanism responsible for the self-organization of the different representations of information: the response of the network changes in such a way that the location of the cell holding a given response corresponds to a specific input signal.

Geometrical arrangement of information is maintained during the training.

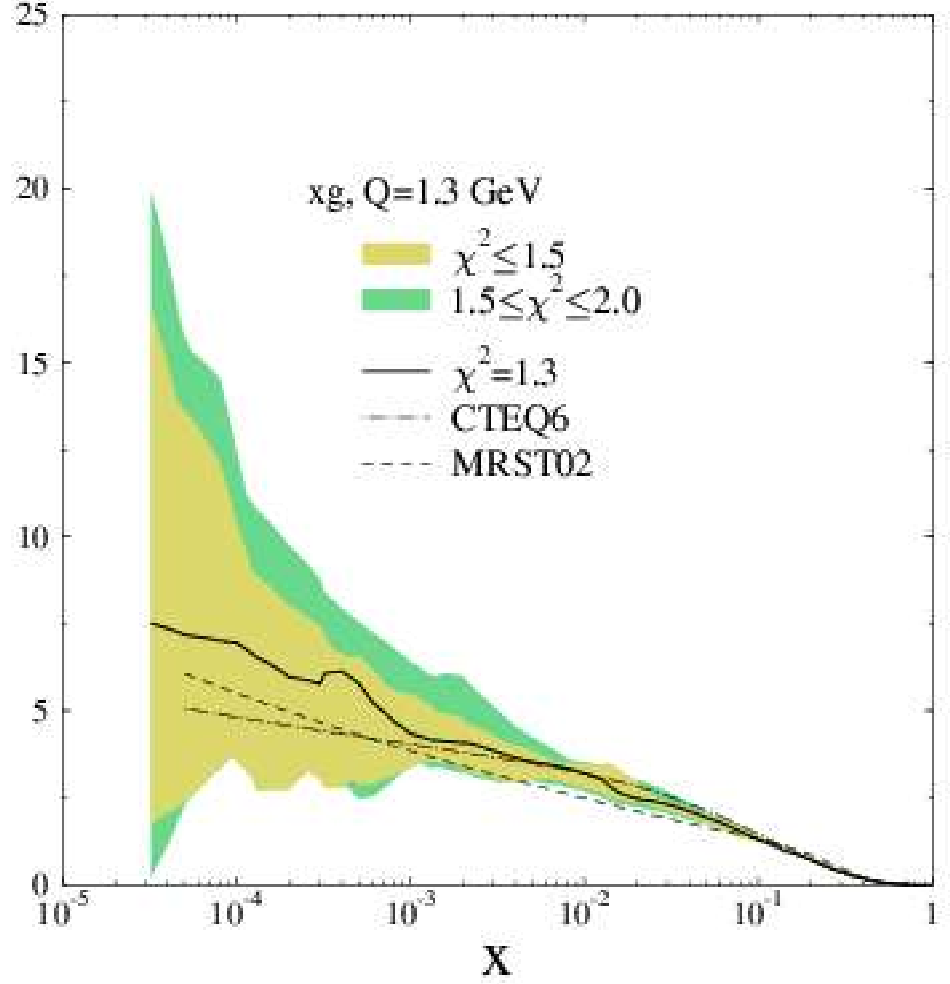
SOM work differently from ANN that do not keep track of the inter-connections among clustering of data at different stages of the network training.

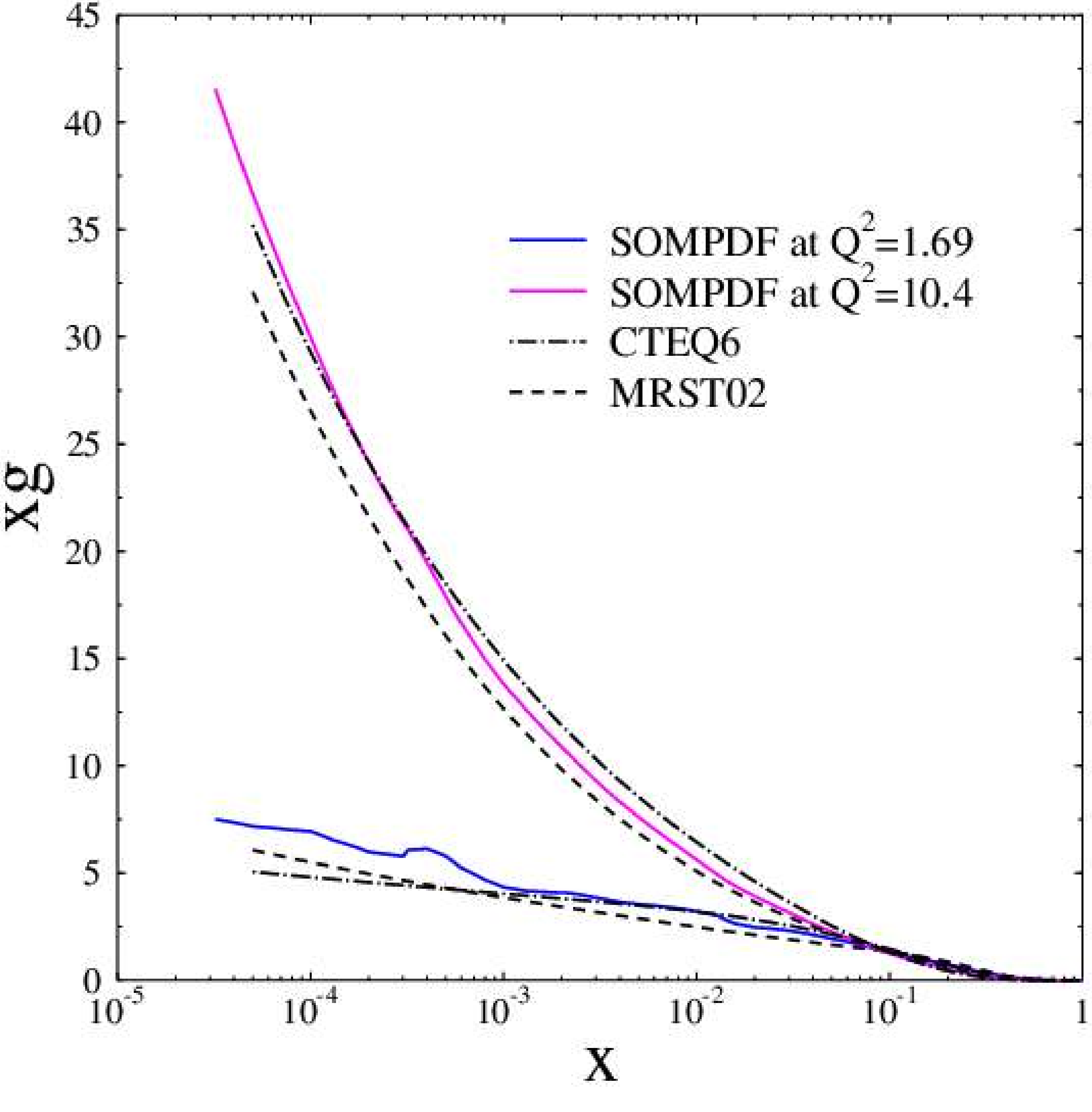
Important because it allows for “user/expert's” intervention: evaluate the impact of possibe theoretical input

Results

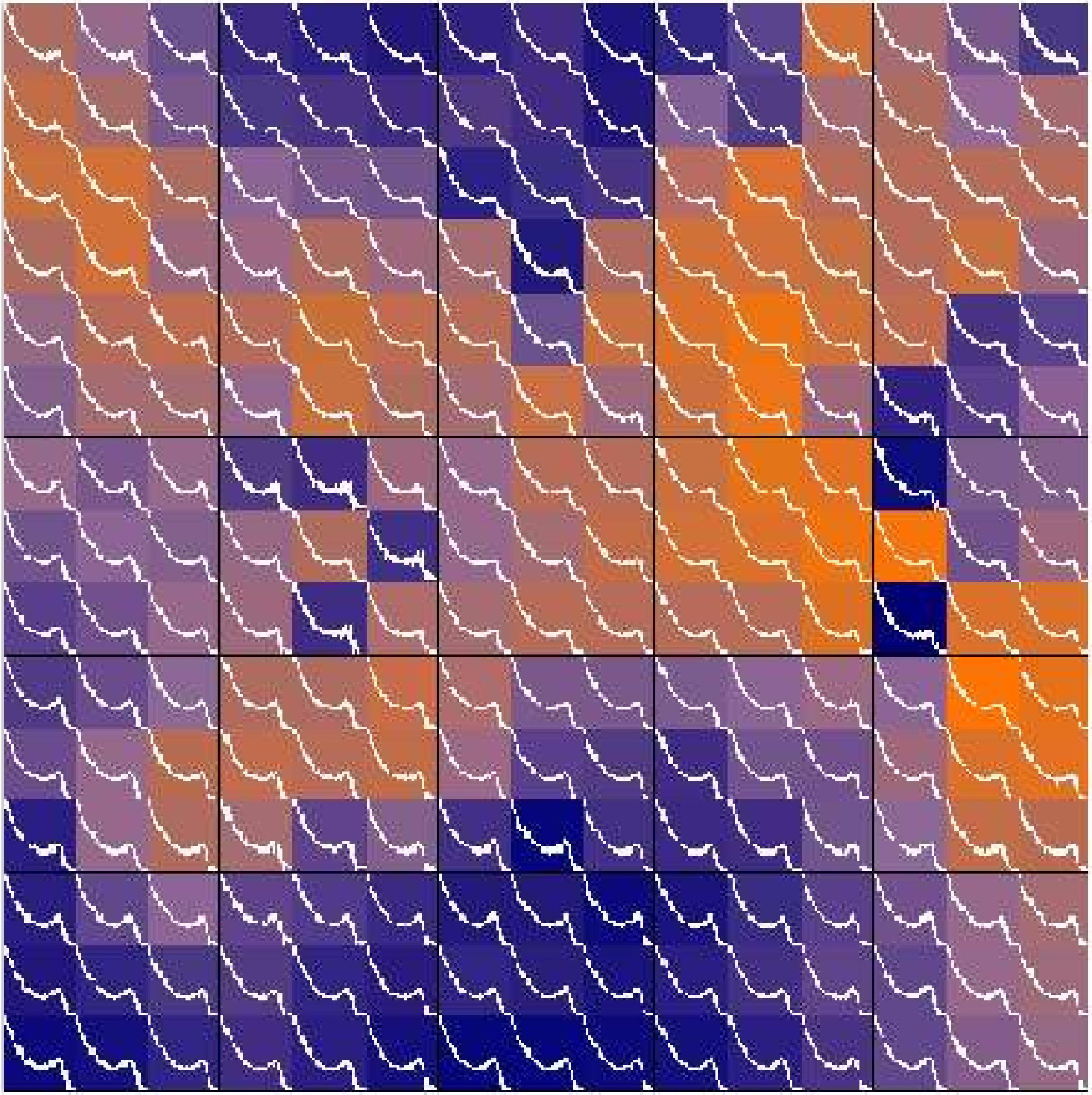


Q=1.3 GeV, 1.3≤χ²≤1.5
0.1*xg
xu_v
xū
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
10⁻⁵
10⁻⁴
10⁻³
10⁻²
10⁻¹
1
x


xg, Q=1.3 GeV
$\chi^2 \leq 1.5$
$1.5 \leq \chi^2 \leq 2.0$
$\chi^2 = 1.3$
CTEQ6
MRST02
25
20
15
10
5
0
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
1
X



45
40
35
30
25
20
15
10
5
0
xg
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
1
x
SOMPDF at Q^2=1.69
SOMPDF at Q^2=10.4
CTEQ6
MRST02



<u>Next step</u>. **Study why the PDFs are arranged in a certain way in the map: introduce “flexbile points” in the analysis**

5. Conclusions and Outlook

- Challenging questions lie ahead for the interpretation of exclusive and semi-inclusive experimental data: quark and gluons momentum, spin, spatial d.o.f, distributions can be accessed in principle but need to be mapped out with new methods
- We presented a new computational method: Self-Organizing Maps (SOM) that works well for proton PDFs
- Future: 1. Apply to nuclear PDFs, semi-inclusive...
- Future: 2. Connection with Complexity Theory?