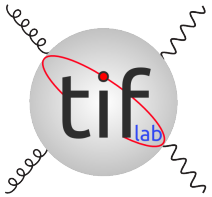


BIG DATA

STEFANO FORTE
UNIVERSITÀ DI MILANO & INFN

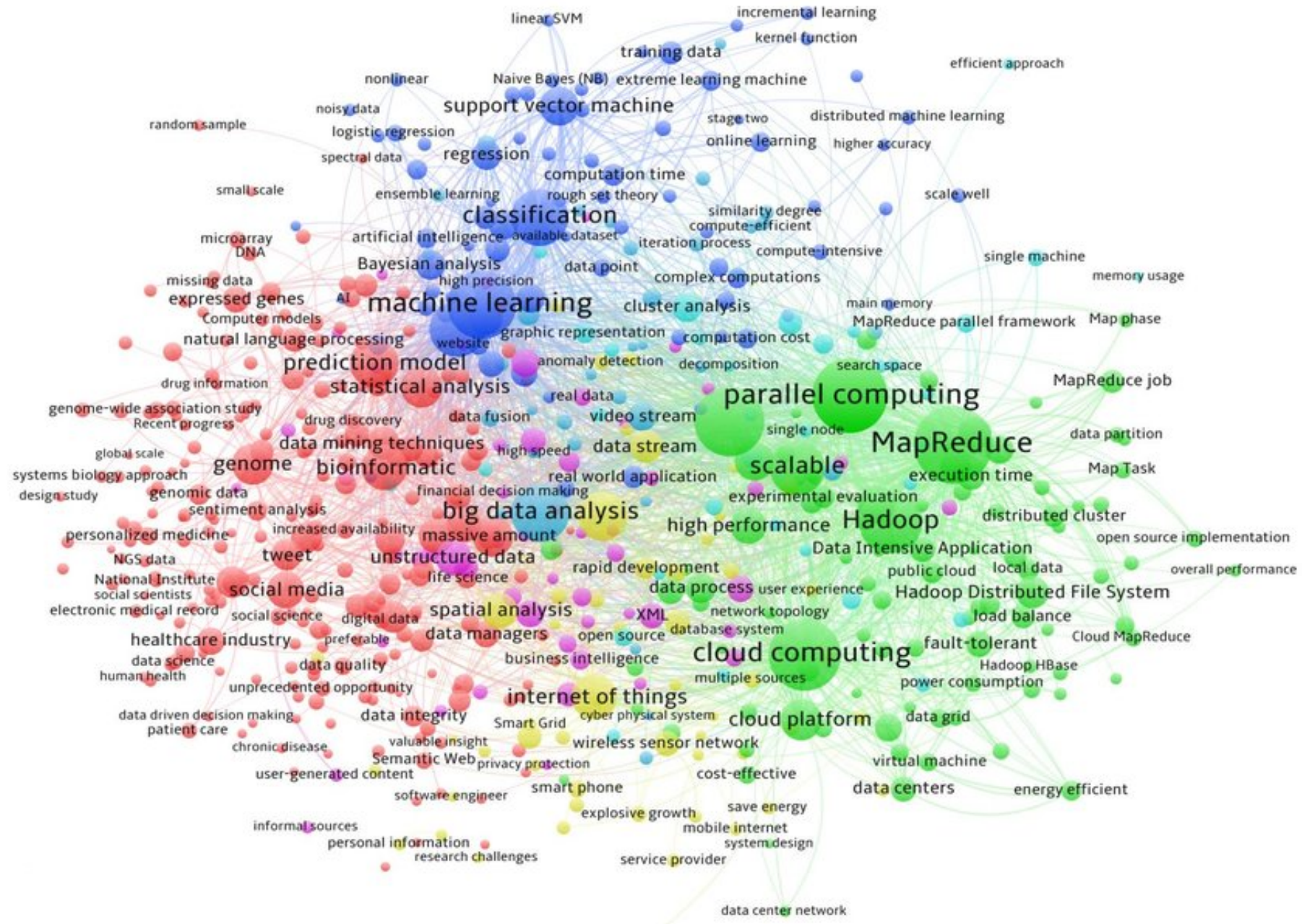
UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI FISICA

SULLE SPALLE DEI GIGANTI

BOLOGNA, 20 MAGGIO 2021

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 740006

BIG DATA?

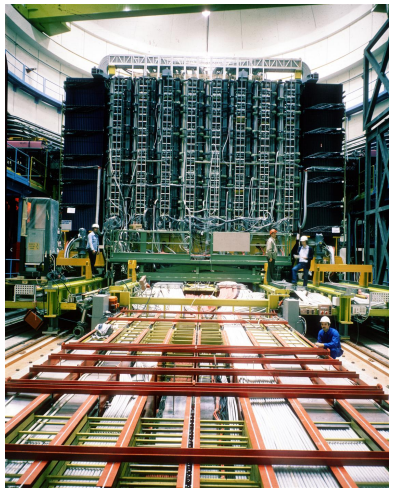

linear SVM
incremental learning
kernel function
training data
efficient approach
nonlinear
Naive Bayes (NB)
extreme learning machine
noisy data
support vector machine
stage two
distributed machine learning
random sample
logistic regression
online learning
higher accuracy
spectral data
regression
computation time
small scale
ensemble learning
rough set theory
scale well
classification
similarity degree
compute-efficient
microarray
DNA
artificial intelligence
available dataset
iteration process
compute-intensive
single machine
Bayesian analysis
data point
missing data
high precision
complex computations
memory usage
expressed genes
AI
machine learning
cluster analysis
main memory
Computer models
MapReduce parallel framework
Map phase
natural language processing
website
graphic representation
computation cost
prediction model
anomaly detection
decomposition
search space
drug information
statistical analysis
real data
MapReduce job
parallel computing
genome-wide association study
Recent progress
drug discovery
data fusion
video stream
single node
data mining techniques
high speed
data stream
MapReduce
data partition
global scale
genome
bioinformatic
scalable
systems biology approach
real world application
execution time
Map Task
design study
genomic data
financial decision making
experimental evaluation
sentiment analysis
big data analysis
high performance
Hadoop
distributed cluster
personalized medicine
increased availability
massive amount
open source implementation
tweet
unstructured data
Data Intensive Application
NGS data
rapid development
public cloud
local data
overall performance
National Institute
life science
social media
data process
user experience
Hadoop Distributed File System
social scientists
electronic medical record
social science
digital data
spatial analysis
XML
network topology
load balance
healthcare industry
preferable
data managers
open source
database system
Cloud MapReduce
data science
business intelligence
cloud computing
fault-tolerant
human health
data quality
multiple sources
Hadoop HBase
power consumption
unprecedented opportunity
internet of things
data driven decision making
data integrity
cyber physical system
cloud platform
data grid
patient care
Smart Grid
valuable insight
wireless sensor network
chronic disease
Semantic Web
privacy protection
virtual machine
user-generated content
cost-effective
smart phone
data centers
energy efficient
software engineer
explosive growth
save energy
informal sources
mobile internet
personal information
system design
research challenges
service provider
data center network

LA SCOPERTA DEI BOSONI W E Z (1984)

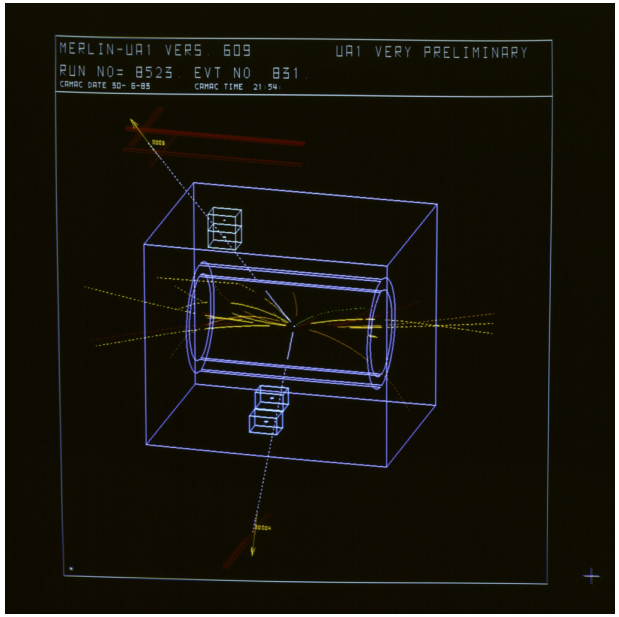
"BIG SCIENCE" MA NON ANCORA BIG DATA

UA1: $\sim$ 100 MEMBRI

UA1: lungh. 6m, diam. 2.3m



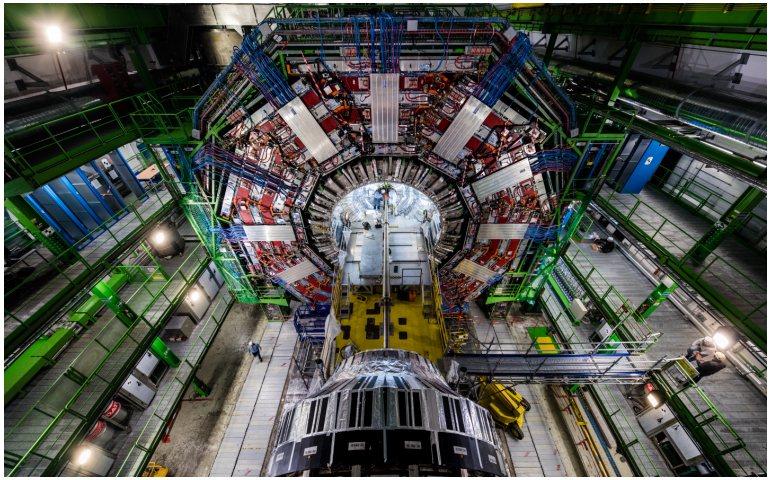
EVENTO: DECINE DI TRACCE



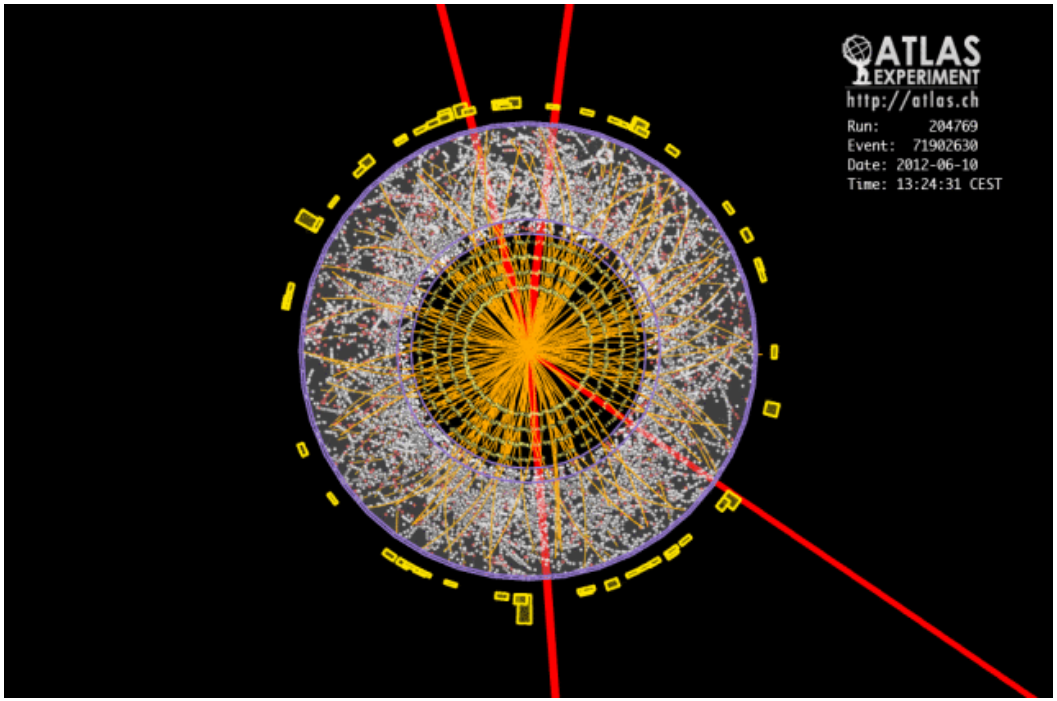
LA SCOPERTA DEL BOSONE DI HIGGS (2012)

CMS: $\sim$ 2500 MEMBRI

CMS: lungh. 21m, diam. 15m

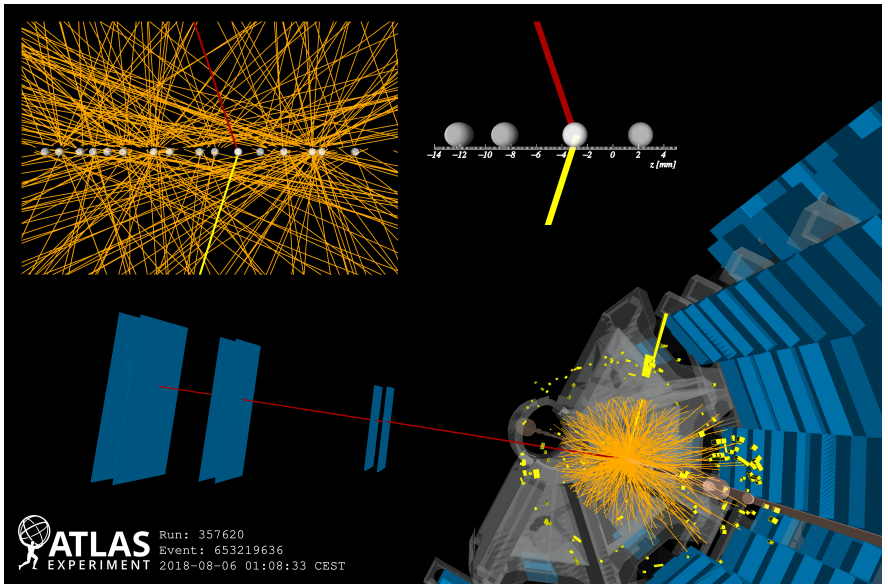


EVENTO: MIGLIAIA DI TRACCE

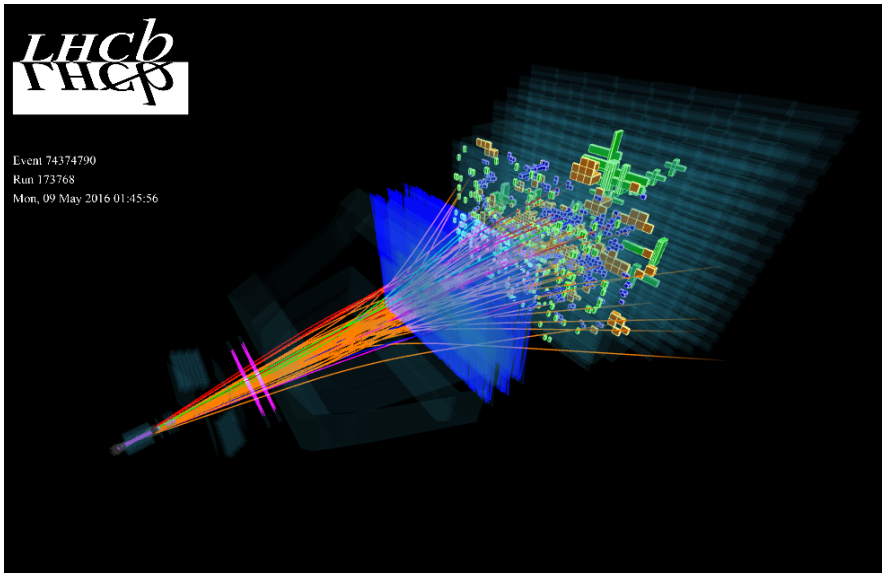


BIG DATA!

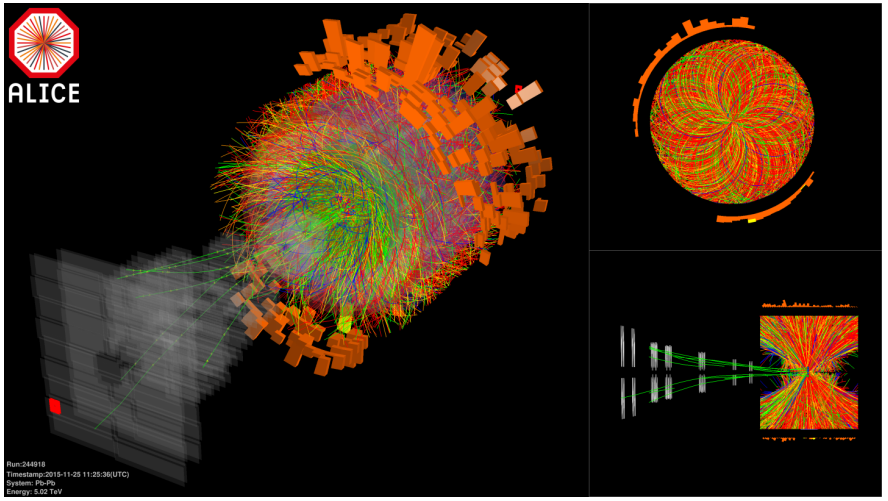
ATLAS



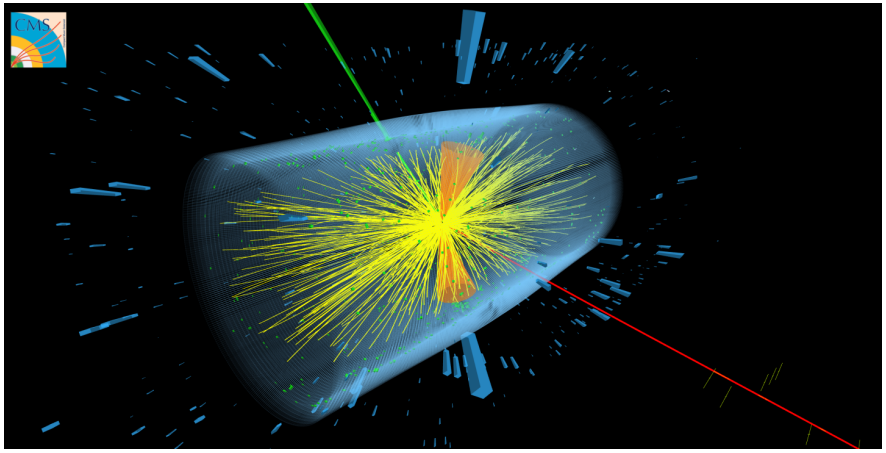
LHCb



ALICE

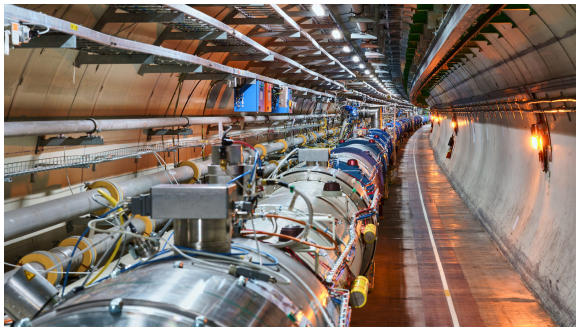


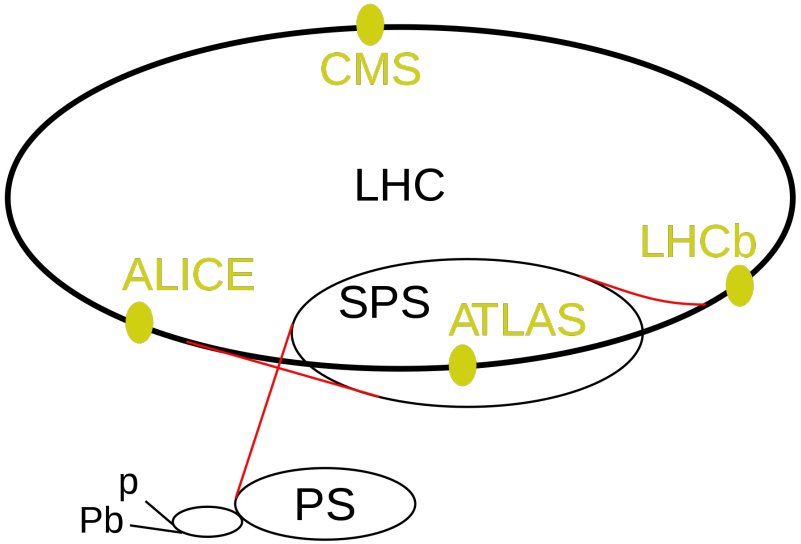
CMS

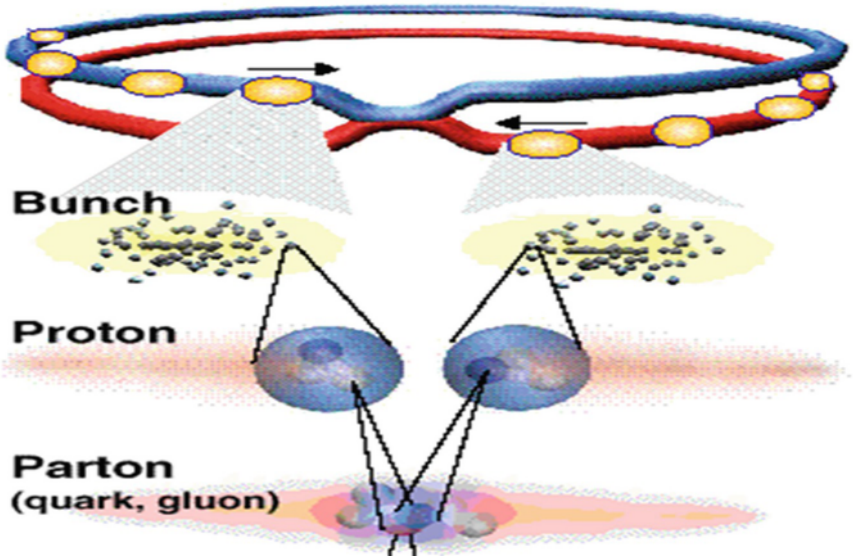


CIRCA 1 MILIARDO COLLISIONI/SEC; CIRCA 100 PETABYTE/ANNO

LHC: URTI FRA PROTONI

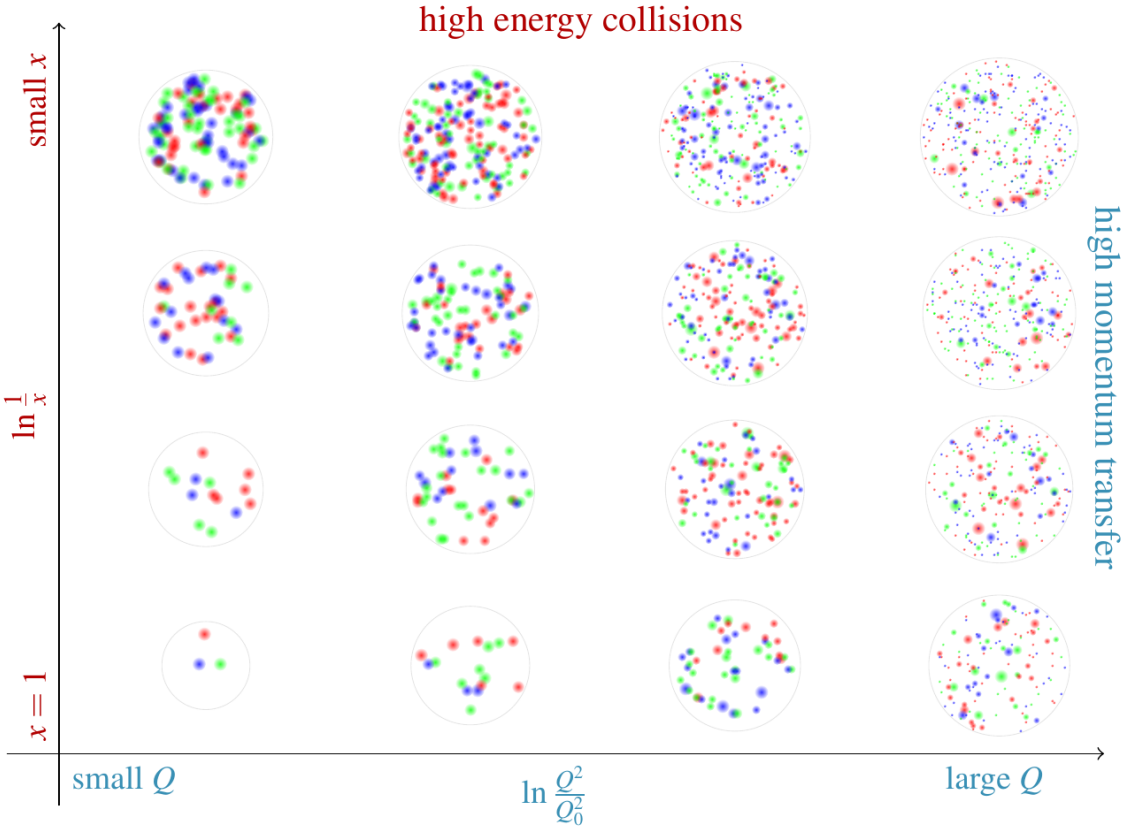


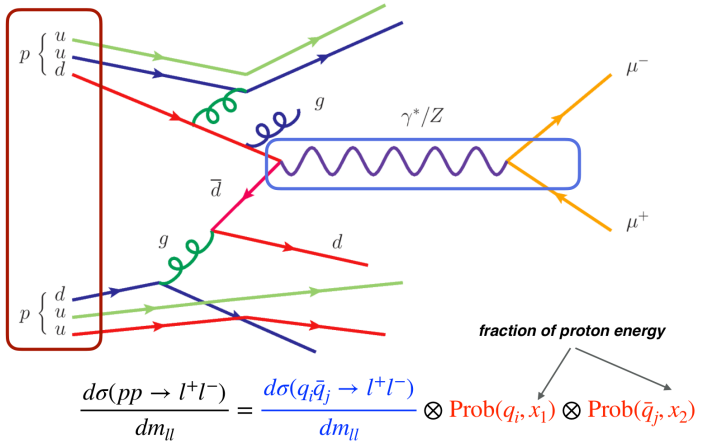




IL PROTONE NON È UNA PARTICELLA ELEMENTARE!

DENTRO IL PROTONE





PROBABILITÀ DI PROBABILITÀ:

INFINITI^2 PARAMETRI

DA ESTRARRE DAI DATI

$\Leftrightarrow$

QCD: UN SOLO

PARAMETRO LIBERO Λ

SMALL DATA

IL PROTONE NEL 1984...

EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CERN-EP/85-108
11 July 1985

W PRODUCTION PROPERTIES AT THE CERN SPS COLLIDER

UA1 Collaboration, CERN, Geneva, Switzerland

Aachen[1]-Amsterdam (NIKHEF)[2]-Annecy (LAPP)[3]-Birmingham[4]-CERN[5]-
Harvard[6]-Helsinki[7]-Kiel[8]-London (Imperial College[9] and Queen Mary College[10])-Padua[11]-
Paris (Coll. de France)[12]-Riverside[13]-Rome[14]-Rutherford Appleton Lab.[15]-
Saclay (CEN)[16]-Victoria[17]-Vienna[18]-Wisconsin[19] Collaboration

The corresponding experimental result for the 1984 data at $\sqrt{s} = 630$ GeV is

$$(\sigma \cdot B)_W = 0.63 \pm 0.05\ (\pm 0.09)\ \text{nb}.$$

This is in agreement with the theoretical expectation [14] of $0.47^{+0.14}_{-0.08}$ nb. We note that the 15%

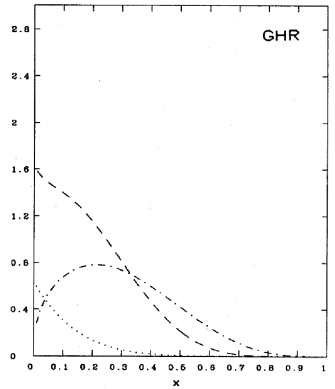


FIG. 25. Parton distributions of Glück, Hoffmann, and Reya (1982), at $Q^2=5$ GeV^2: valence quark distribution $x[u_v(x)+d_v(x)]$ (dotted-dashed line), $xG(x)$ (dashed line), and q_v (dotted line).

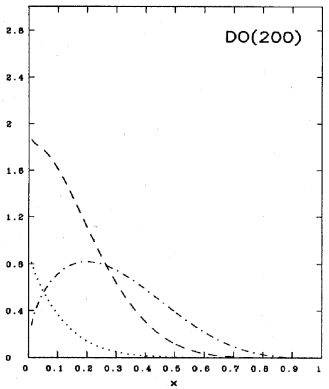


FIG. 27. "Soft-gluon" ($\Lambda=200$ MeV) parton distributions of Duke and Owens (1984) at $Q^2=5$ GeV^2: valence quark distribution $x[u_v(x)+d_v(x)]$ (dotted-dashed line), $xG(x)$ (dashed line), and $q_v(x)$ (dotted line).

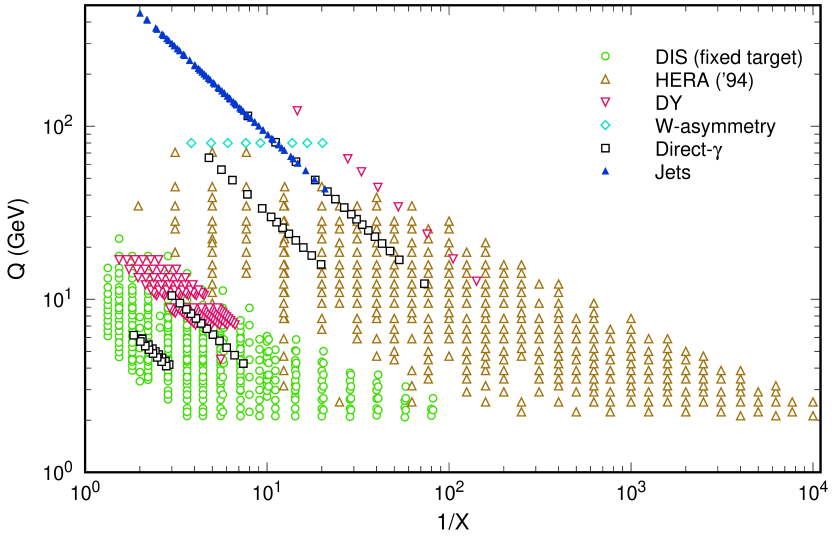
TABLE 2
Values (in nb) of the total cross sections for $W^{\pm}$ and Z^0 production

$\sqrt{s}$ (GeV)	$W^+ + W^-$ GHR	$W^+ + W^-$ DO1	$W^+ + W^-$ DO2	Z^0 GHR	Z^0 DO1	Z^0 DO2	$\frac{\sigma(W^+ + W^-)}{\sigma(Z^0)}$ GHR	$\frac{\sigma(W^+ + W^-)}{\sigma(Z^0)}$ DO1	$\frac{\sigma(W^+ + W^-)}{\sigma(Z^0)}$ DO2
540	4.2	4.3	4.1	1.3	1.3	1.2	3.1	3.4	3.5
700	6.2	6.3	6.1	2.0	1.9	1.8	3.1	3.3	3.4
1000	9.5	9.5	9.6	3.1	3.0	2.9	3.1	3.2	3.3
1300	12.5	12.5	12.9	4.0	3.9	3.9	3.1	3.2	3.3
1600	15.5	15.6	16.5	5.0	4.8	5.0	3.1	3.2	3.3

- SEMPLICE MODELLO CON DUE-TRE PARAMETRI
- NESSUNA STIMA DELL'ERRORE
- ACCURATEZZA QUALITATIVA

...E NEL 2000

DATASET CTEQ5 (1999)



- QUALCHE CENTINAIA DI DATI
- PARAMETRIZZAZIONE AD-HOC
- FIT MULTIPARAMETRICO CON INCERTEZZA
- ACCURATEZZA SEMI-QUANTITATIVA

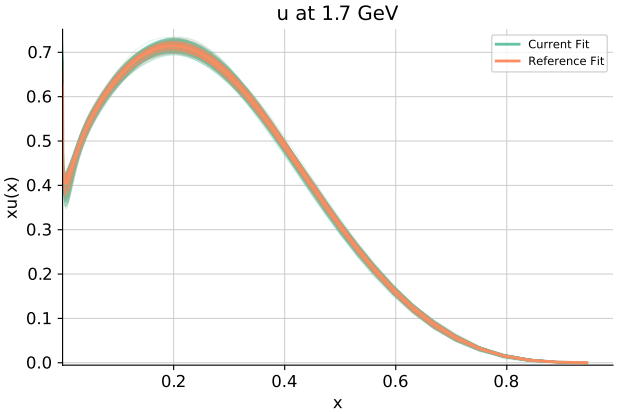
PARAMETRIZZAZIONE CTEQ5 (2006)

$$x\,f(x,Q_0) = A_0\,x^{A_1}\,(1-x)^{A_2}\,e^{A_3 x}\,(1+e^{A_4}\,x)^{A_5}, \quad \text{(6 funzioni , 22 parametri)}$$

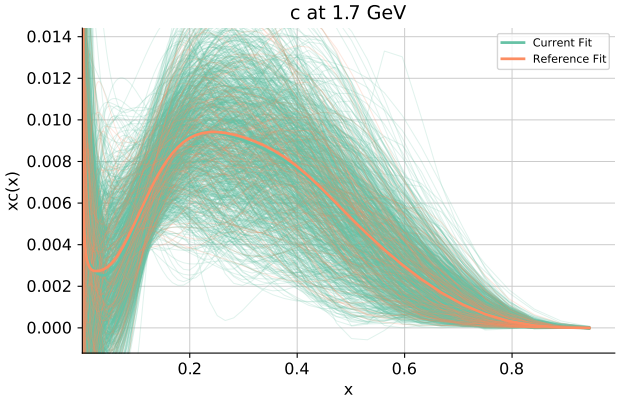
TOWARDS BIG DATA

NNPDF (2002-2017)

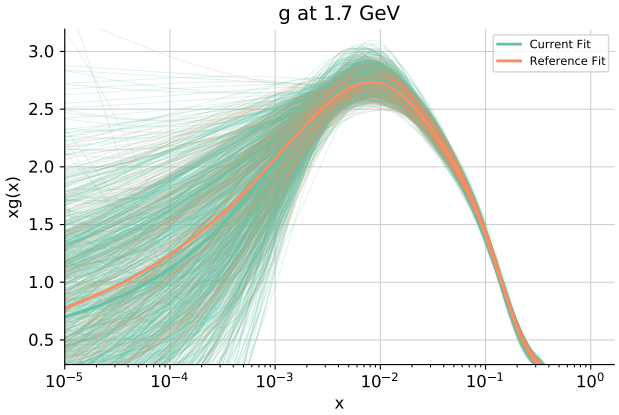
QUARK UP



QUARK CHARM



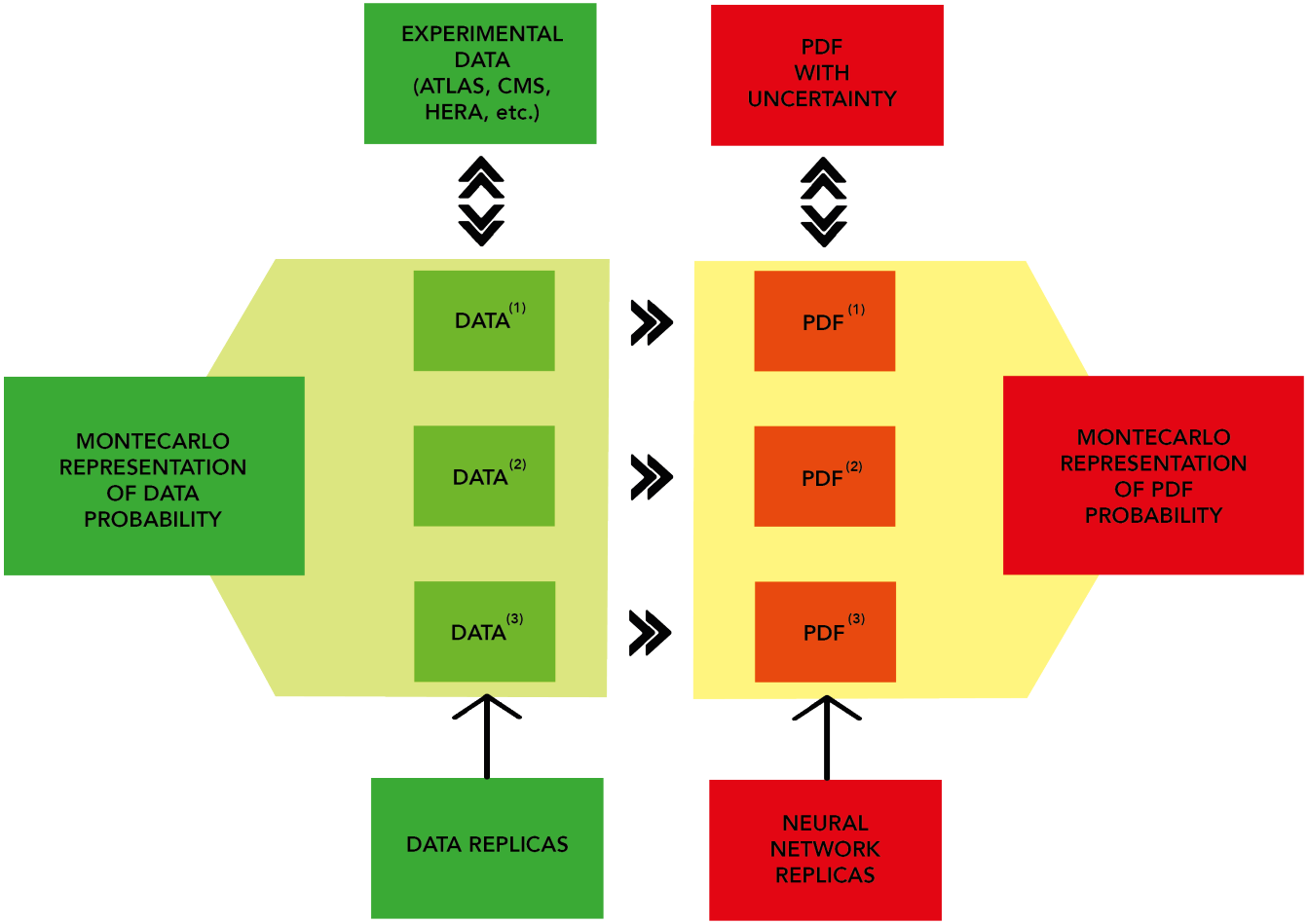
GLUONE



RAPPRESENTAZIONE MONTECARLO $\Leftrightarrow$ DISTRIBUZIONE DI PROBABILITÀ

IL MONTECARLO FUNZIONALE

REPLICHE DI FUNZIONI $\Leftrightarrow$ PROBABILITÀ IN UNO SPAZIO DI FUNZIONI

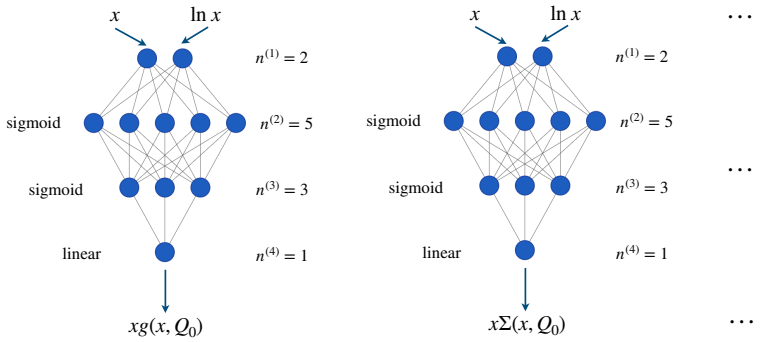


PDF OUTPUT: $f_i^{(a)}(x,\mu)$; i =up, antiup, down, antidown, strange, antistrange, charm, gluone;

$j = 1, 2, \ldots N_{\rm rep}$

RETI NEURALI

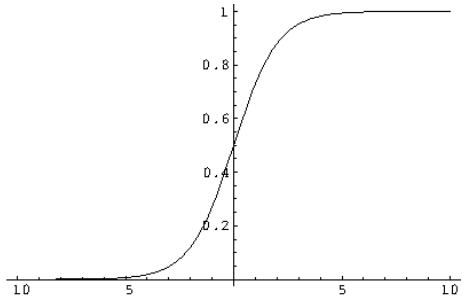
ARCHITETTURA



$$F^{(i)}_{\rm out}(\vec{x}_{\rm in}) = F\left(\sum_j \omega_{ij} x^j_{\rm in} - \theta_i\right)$$

PARAMETERI

FUNZIONE DI ATTIVAZIONE

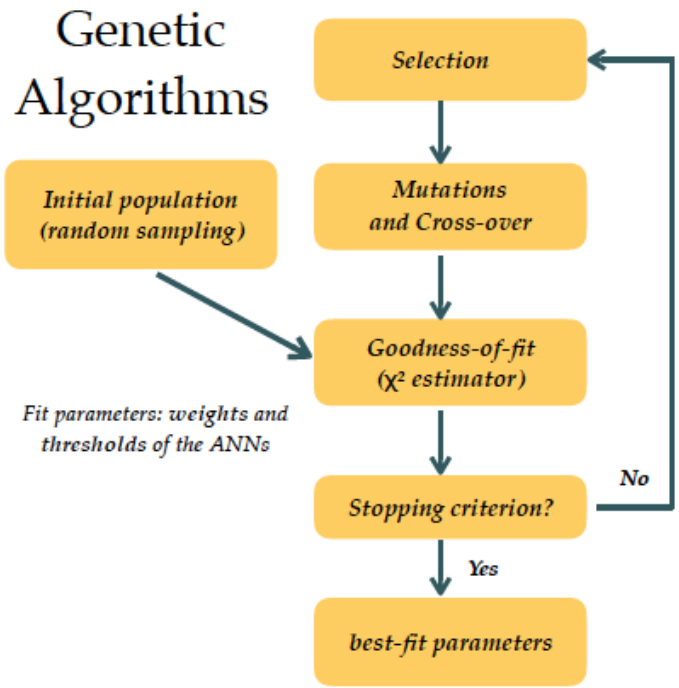


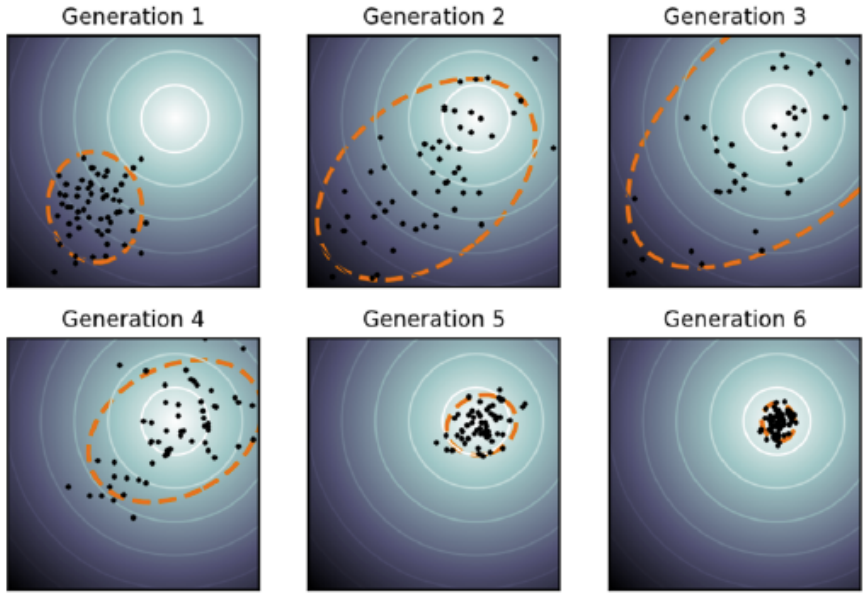
- PESI ω_{ij}
- SOGLIE θ_i
- INTERPOLANTE UNIVERSALE
- PUÒ RIPRODURRE QUALUNQUE FORMA FUNZIONALE
- COMPLESSITÀ CRESCE DURANTE L'ADDESTRAMENTO

NNPDF1.0-3.1: $2 - 5 - 3 - 1$ NN PER OGNI PDF: $37 \times 8 = 296$ PARAMETETRI

ALGORITMI GENETICI

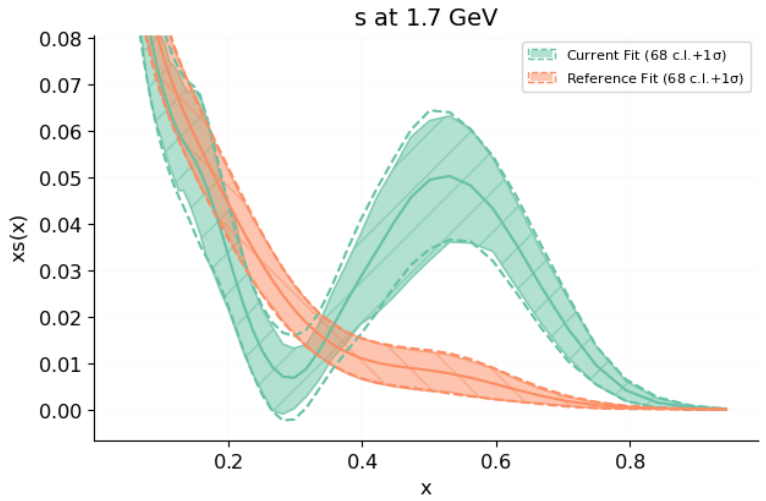
- MUTAZIONE CASUALE DEI PARAMETRI DELLA RETE
- SELEZIONE DEL PIÙ ADATTO



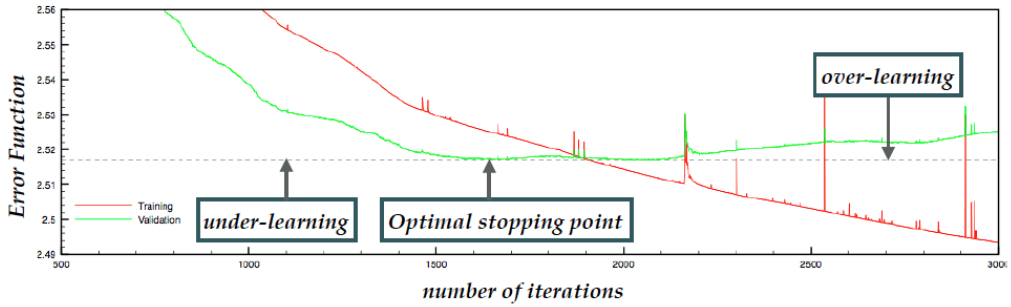


CONVALIDA INCROCIATA

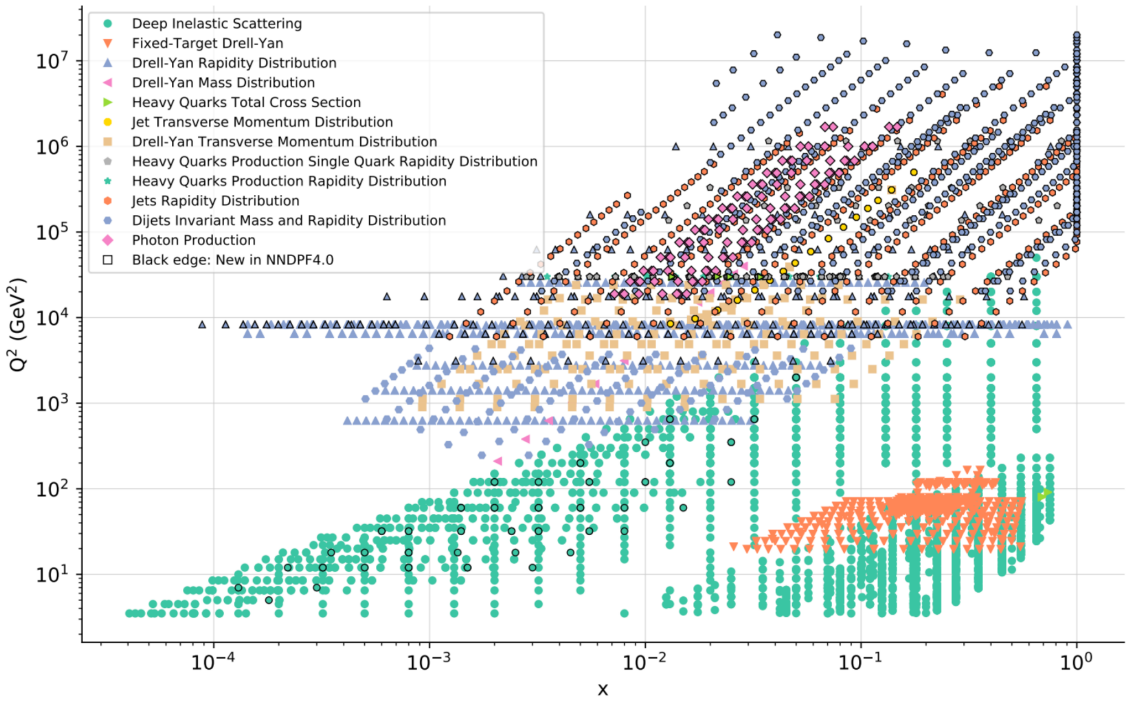
- LA RETE NEURALE
  PUÒ RIPRODURRE TUTTO
- ANCHE IL RUMORE!



- DATI DIVISI CASUALMENTE
  DUE GRUPPI: ADDESTRAMENTO
  E CONVALIDA
- RETI NEURALI OTTIMIZZATE SUI
  DATI DI ADDESTRAMENTO
- QUALITÀ DETERMINATA DAL
  GRUPPO DI CONVALIDA



DATI



- $\sim$ 50 ESPERIMENTI: $\sim$ 5000 DATI SPERIMENTALI
- 1000 REPLICHE, 8 RETI NEURALI
- $\sim$ 20000 GENERAZIONI ALGORTIMO GENETICO
- 80 MUTANTI PER GENERAZIONE

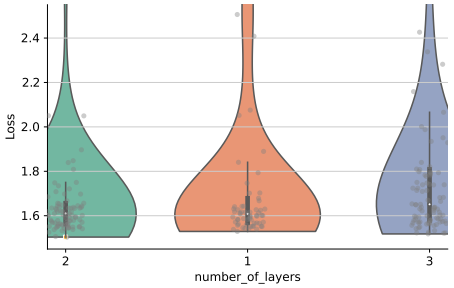
$\sim 10^{13}$ PREDIZIONI TEORICHE PER UNA DETERMINAZIONE TIPICA

BIG DATA!

IPEROTTIMIZZAZIONE

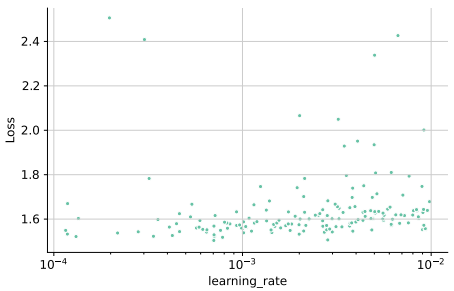
CHI SCEGLIE LA METODOLOGIA?

numero di strati

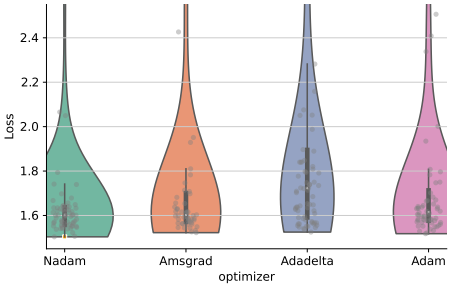


inizializzazione

apprendimento



ottimizzatore

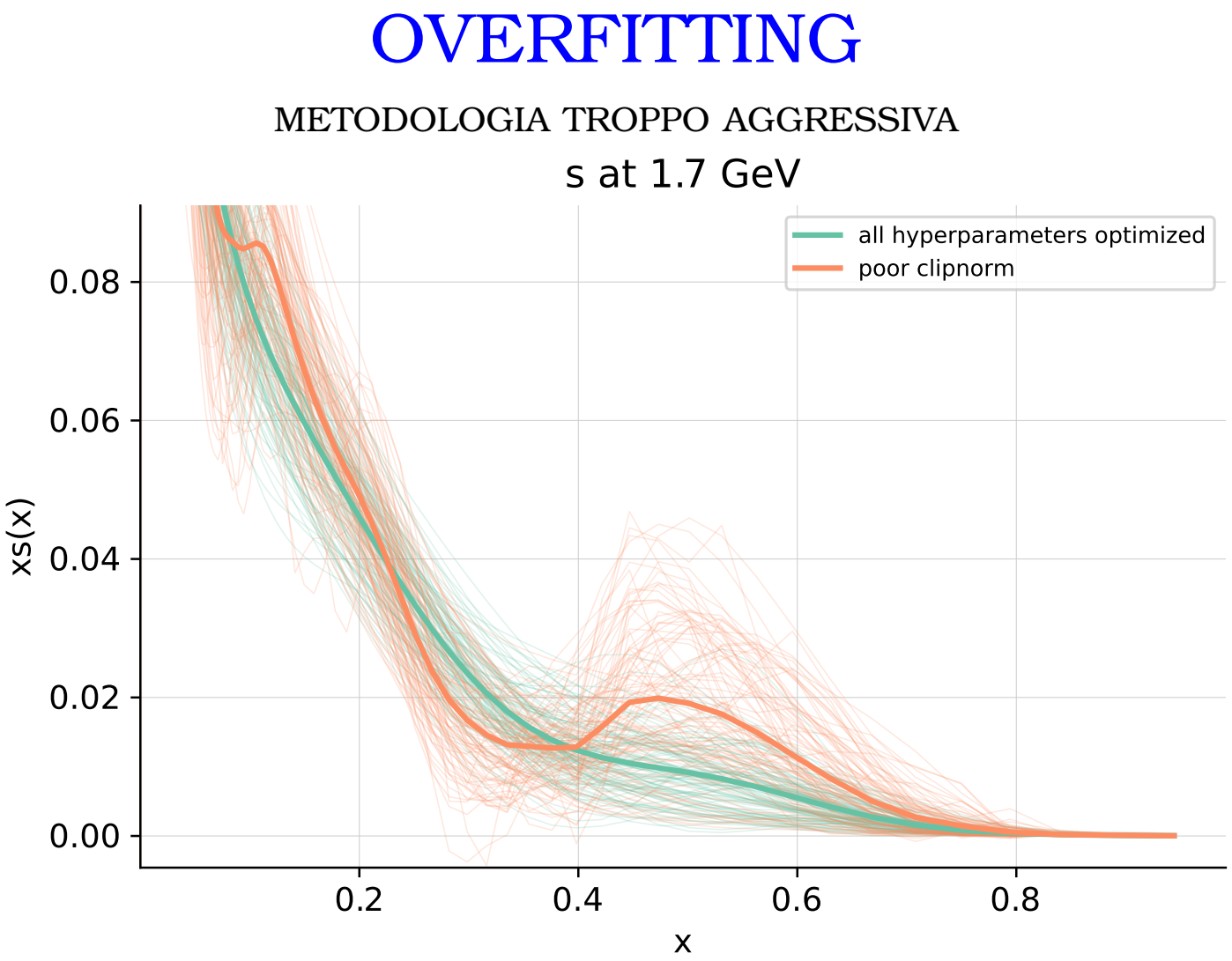


HYPEROPT PARAMETERS

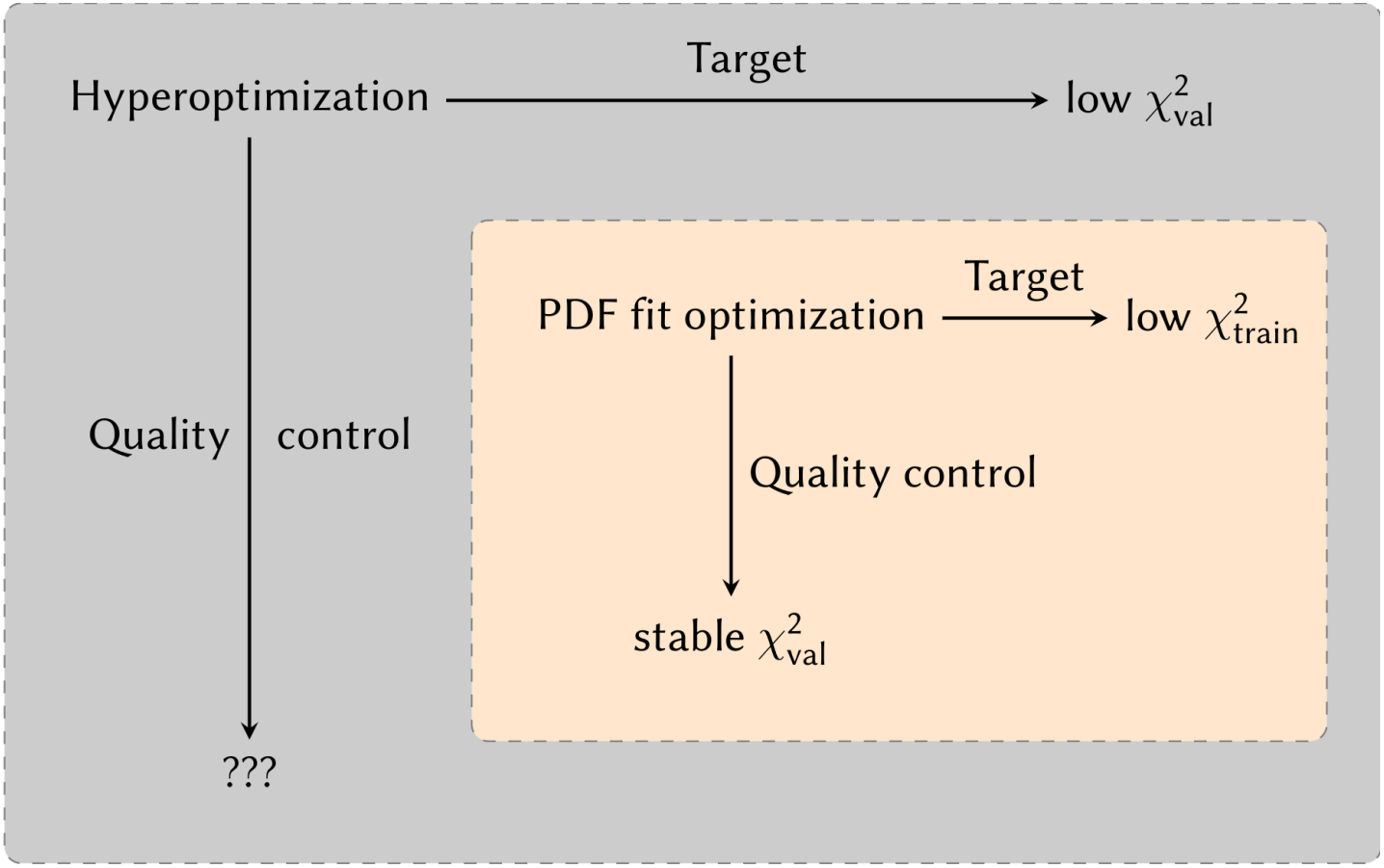
Neural Network	Fit options
Number of layers	Optimizer
Size of each layer	Initial learning rate
Dropout	Maximum number of epochs
Activation functions	Stopping Patience
Initialization functions	Positivity& integrability multiplier

OVERFITTING

METODOLOGIA TROPPO AGGRESSIVA

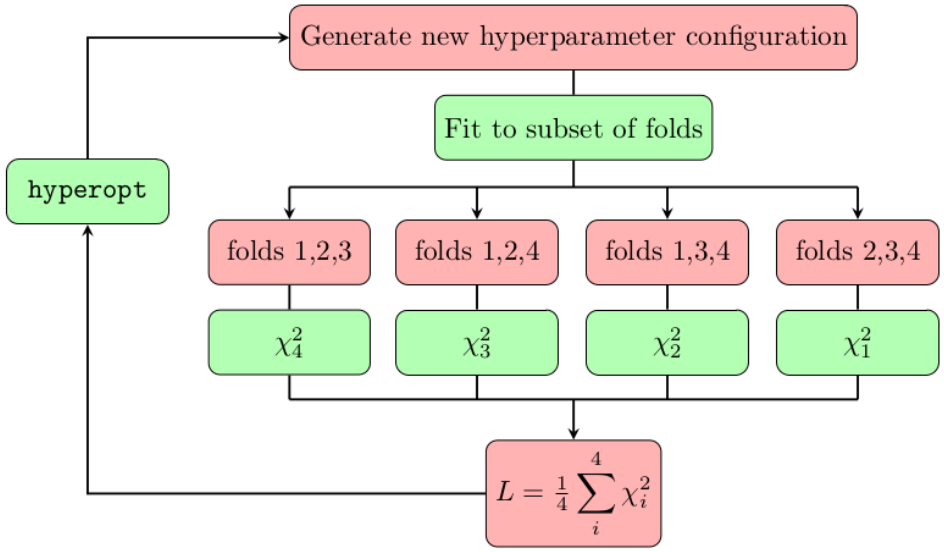


QUAL È LA METODOLOGIA OTTIMALE?



MANCA IL CRITERIO DI QUALITÀ

K-FOLDING

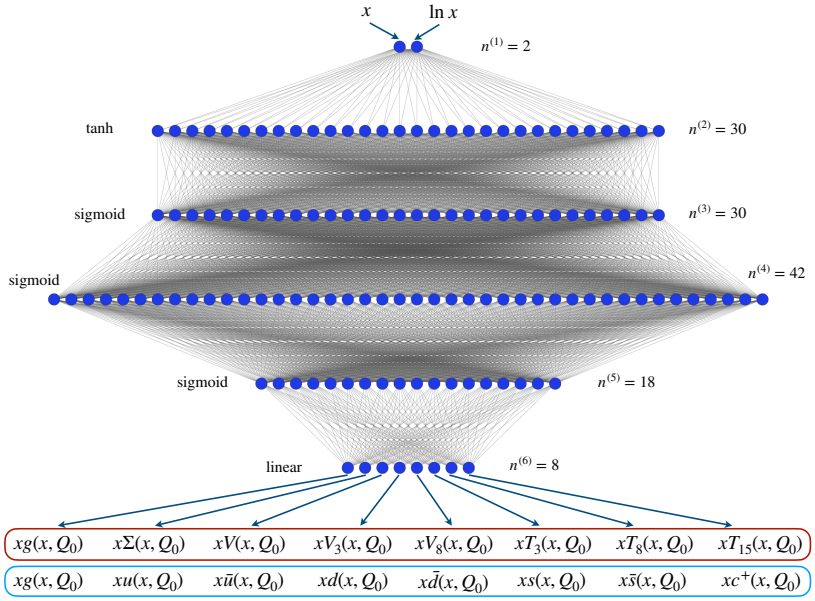


- DATI SUDDIVISI IN n GRUPPI (FOLDS)
- UN GRUPPO ESCLUSO A TURNO
- FIT MIGLIORE $\Rightarrow$ GENERALIZZA CORRETTAMENTE AI DATI ESCLUSI

RISULTATI

15000 CICLI DI IPEROTTIMIZZAZIONE; 4 FOLDS

UNA ARCHITETTURA TESTATA



FINAL NNPDF4.0 SET-UP

AFTER HYPEROPTIMIZATION

PARAMETER	VALUE
ARCHITECTURE	25-20-8
ACTIVATION	HYPERBOLIC TANGENT
INITIALIZER	`glorot_normal`
OPTIMIZER	`Nadam`
CLIPNORM	$6{\cdot}10^{-6}$
LEARNING RATE	$2.6{\cdot}10^{-3}$
MAX EPOCHS	$17{\cdot}10^{3}$
STOPPING PATIENCE	10% OF MAX EPOCHS
INITIAL POSITIVITY λ	185
INITIAL INTEGRABILITY λ	10

IL PROTONE 2021

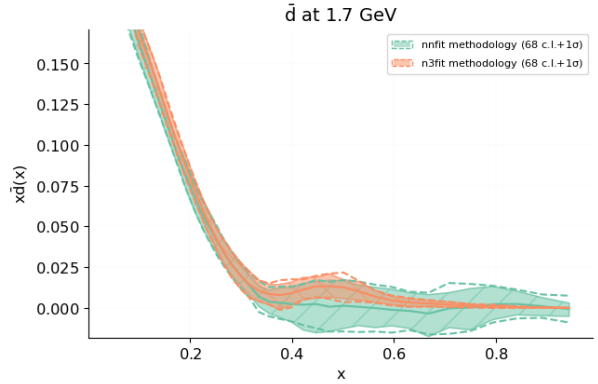
IL CODICE

AVERAGE FITTING TIME PER REPLICA AND USE OF RESOURCES
SAME DATASET FOR OLD AND NEW METHODOLOGIES IN CPU AND GPU
CPU: INTEL(R) CORE(TM) I7-4770 AT 3.40GHZ; GPU: NVIDIA TITAN V

	NNPDF31 CODEBASE	NNPDF40 CODEBASE IN CPU	NNPDF40 CODEBASE IN GPU
TIME	15.2 H.	38 $\pm$ 5 MIN.	6.6 MIN.
RAM USE	1.5 GB	6.1 GB	NA

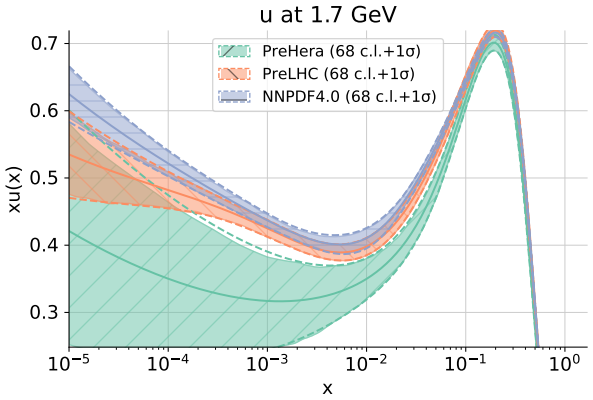
LE PDF

ANTIDOWN: NNPDF3.1 VS NNPDF4.0



IL "FUTURE TEST"

QUARK UP: DATI 1995, 2005, 2021



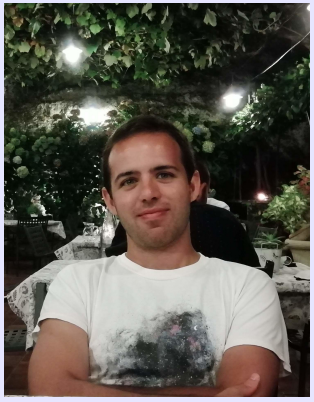
BIGGER DATA?

WHAT NEXT?

- REINFORCEMENT LEARNING
- IMPARARE LA DISTRIBUZIONE DI PROBABILITÀ
- ELEVATA PARALLELIZZAZIONE

N^{3}PDF

A. Candido



S. Carrazza

J. Cruz Martinez

F. Hekhorn

K. Kudashkin

T. Rabemananjara

C. Schwan

R. Stegeman