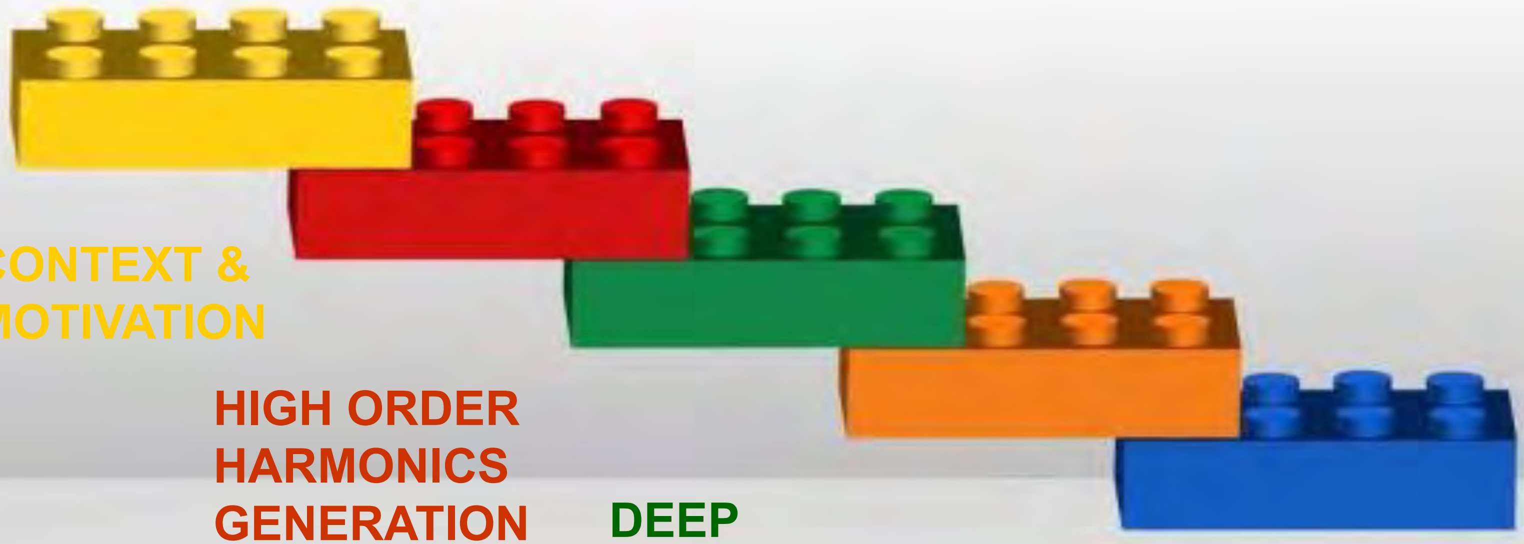
Big Data and Deep Learning Based Predictive Analytics of High Order Harmonics Generation Optimal Scenarios

Andreea MIHAILESCU

GRID CLOUD AND HIGH-PERFORMANCE COMPUTING IN SCIENCE

OUTLINE



CONTEXT & MOTIVATION

HIGH ORDER HARMONICS GENERATION & PARTICLE -IN- CELL

DEEP LEARNING & BIG DATA BASED PREDICTIVE ANALYTICS

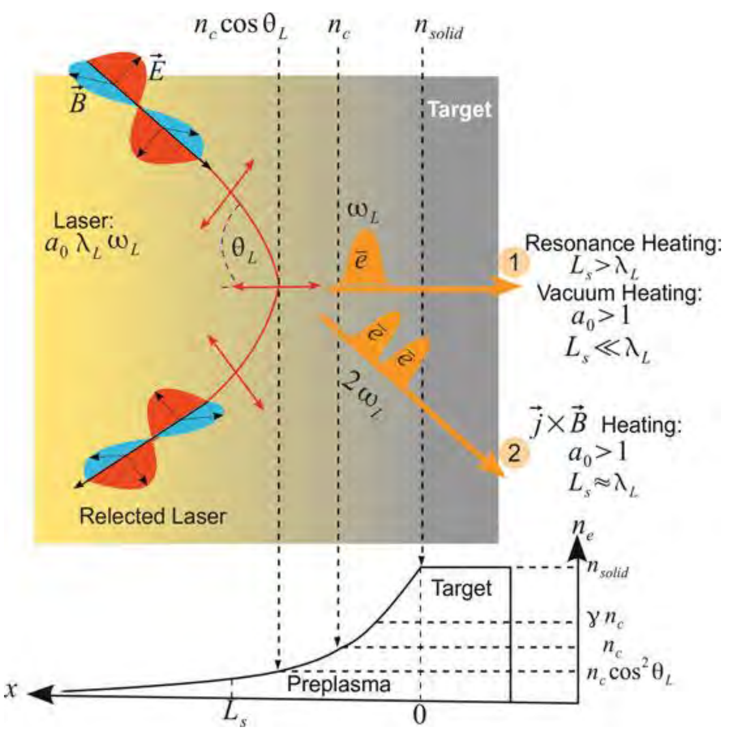
TOWARDS PREDICTING OPTIMAL INTERACTION SCENARIOS WITH DEEP LEARNING

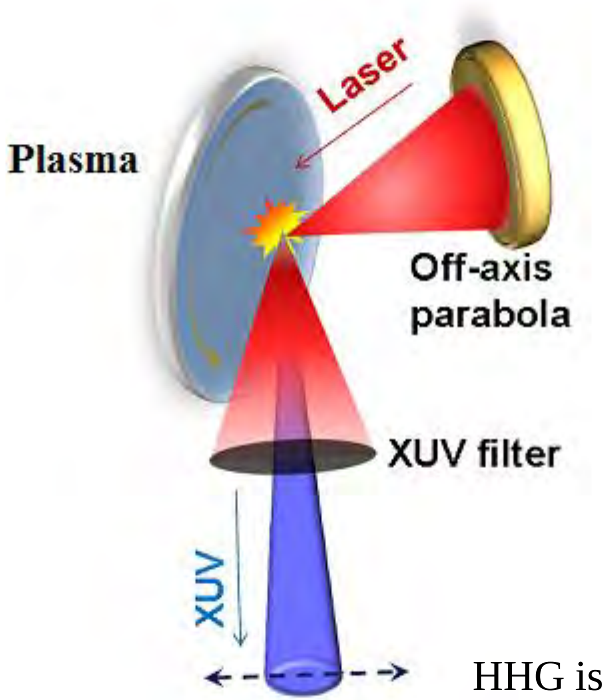
CONCLUSION & PERSPECTIVES

• The increasing availability of ***ultrashort (ps, fs)*** and ***intense*** (*I* > $10^{17} W/cm^2$) laser pulses opened opportunities for challenging investigations of the rapidly heated matter in extreme conditions.

laser energy transfer to matter

• **Traditionally Modelled with:**
Collisionless Vlasov, PIC, Hybrid Codes





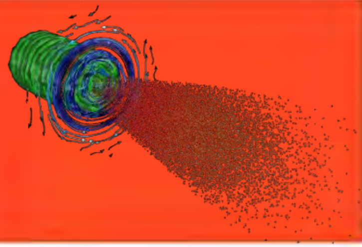
- Higher conversion efficiency than HHG in gases
- Tunable source of XUV/Soft X rays, typically from 100 to 10 nm but even down to 3 nm
- No limitation on the intensity of the incident laser
- Better source from the coherence point of view
- HH exhibit a very tight angular confinement, sometimes with less divergence than that of the fundamental field and near Gaussian beam profiles

HHG is strongly influenced by the laser-plasma interaction itself. The harmonics are a promising tool for obtaining info on plasma parameters such as the local electron density & on the presence of large electric and magnetic fields, plasma waves, and the electron transport inside the target.

- **PIC codes:**

- used in almost all areas of plasma physics,
- suitable for analyzing highly transient processes.

- they are the most intensive computational modelling tool employed in plasma physics



- evolved towards improved performances, incorporating physics packages beyond the traditional method & taking advantage of the advent of HPC systems

- Increased dimensionality towards 3D3V
- Full-parallelization and CUDA versions
- Fully-relativistic
- Incorporating new diagnostic capabilities
- Improved visualization packages
- Object Orientation
- Migration from Fortran to better performing languages

- Statistical noise
- Non-physical instabilities
- Non-conservation
- Numerical heating
- Require considerable computational resources

"In theory there is no difference between theory and practice. In practice there is."

D. Rosenberg & M. Stephens

OBJECTIVE: Move the simulation results as close as possible to the real scenarios.

Since

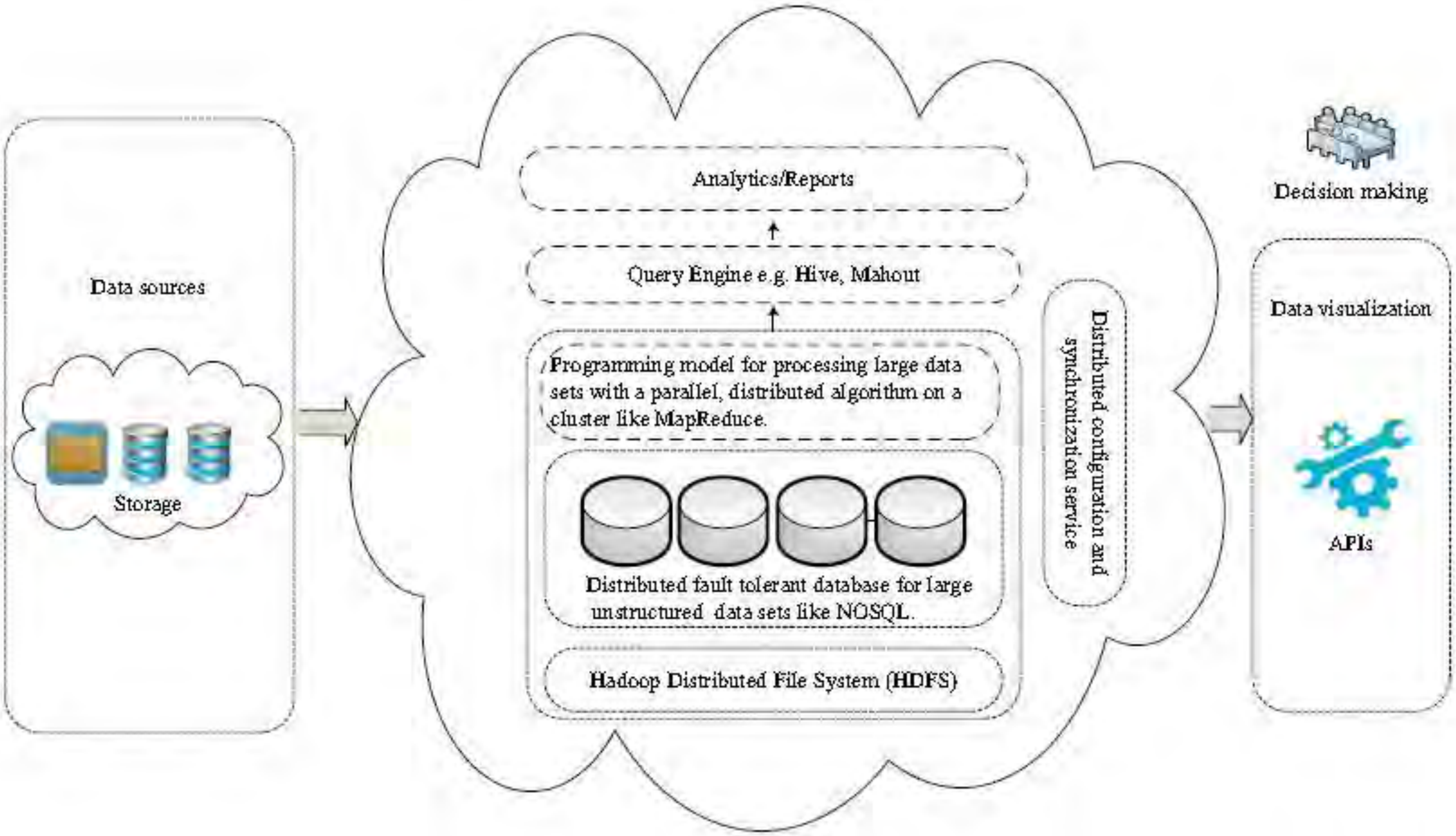
- There is lots of experimental and simulation data already available in the literature

&

- Cutting-edge technologies like the cloud and Big Data have tremendous potential for scientific computing

COMPUTATIONALLY INTENSIVE → DATA INTENSIVE

DEEP LEARNING & BIG DATA BASED PREDICTIVE ANALYTICS



DEEP LEARNING & BIG DATA BASED PREDICTIVE ANALYTICS

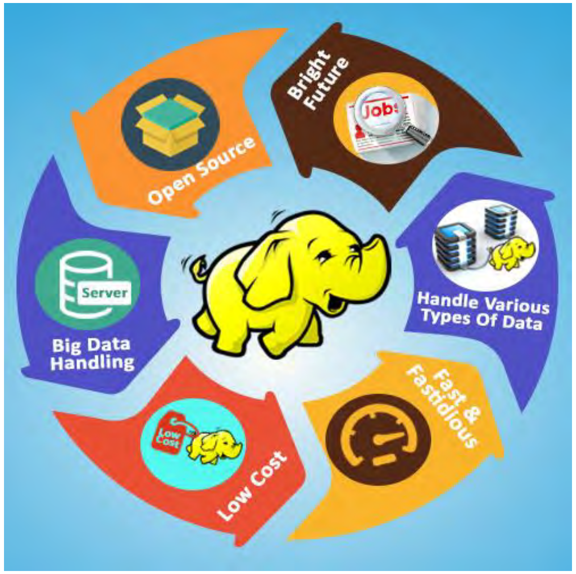
LASER

TARGET

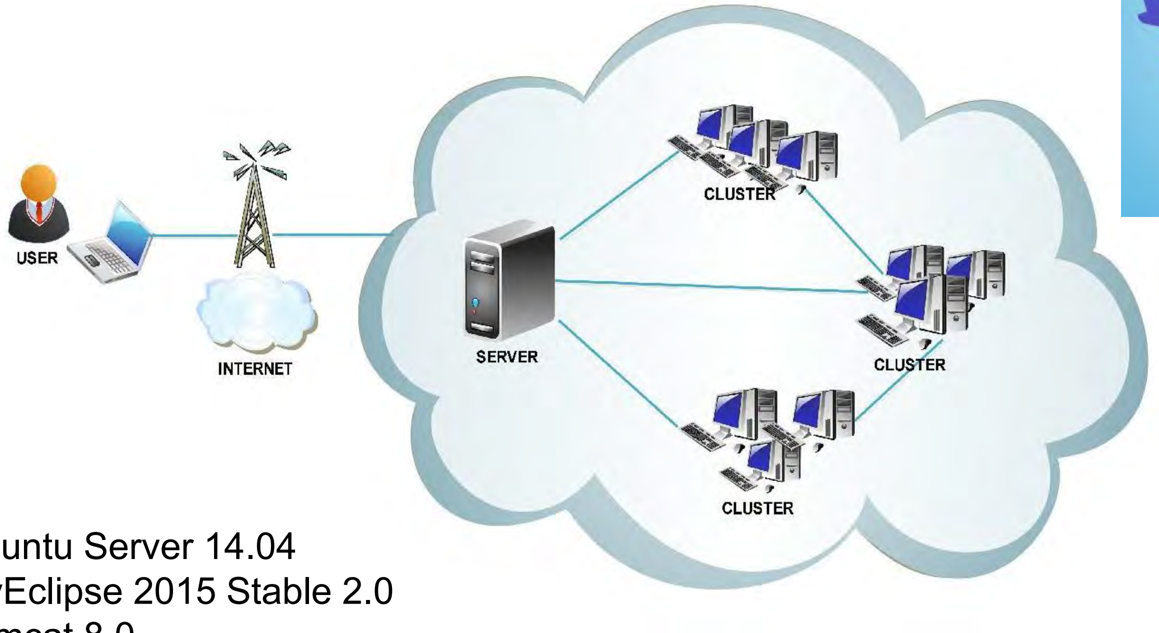
GOAL

PREDICTIVE SYSTEM → OPTIMAL INTERACTION CONDITIONS

DEEP LEARNING & BIG DATA BASED PREDICTIVE ANALYTICS

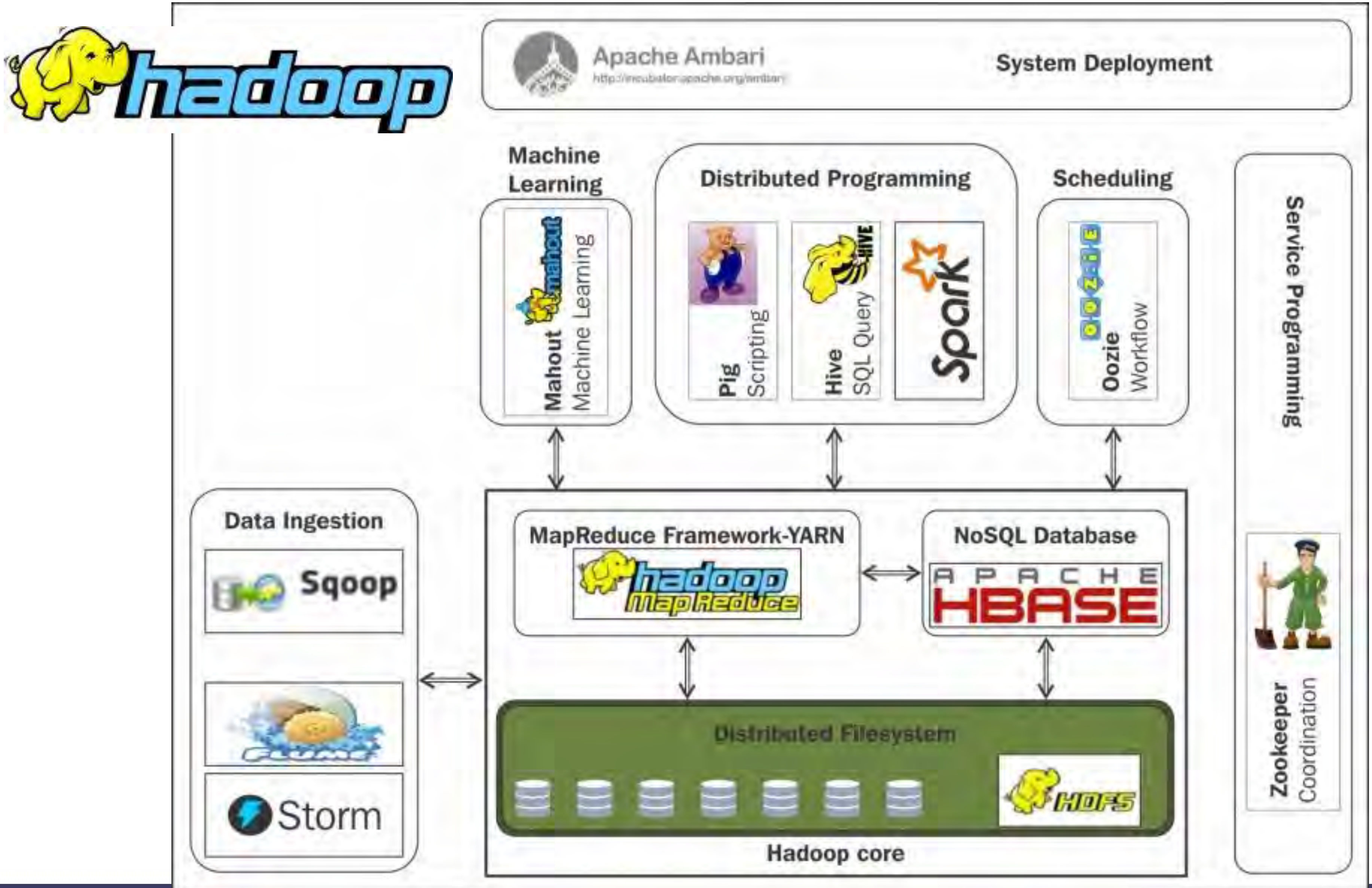


- Communication with users
- User requests
- Manipulation &Treatment of data
- Data Storage

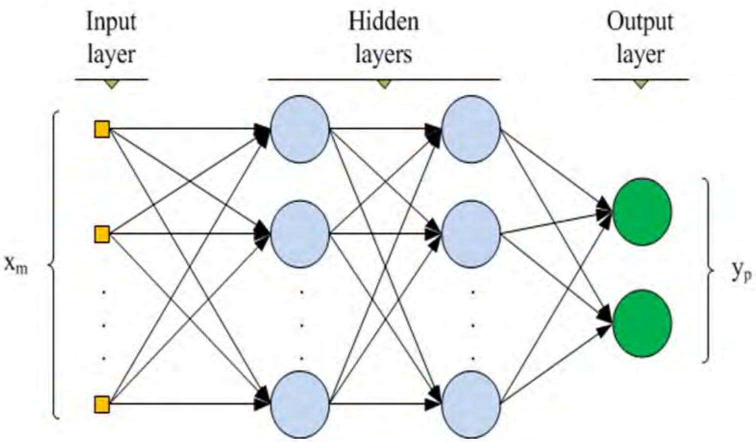


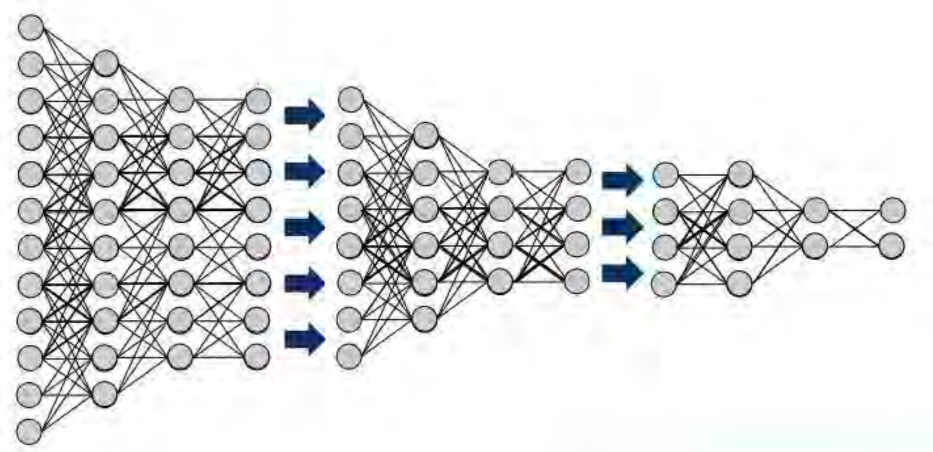
- Ubuntu Server 14.04
- MyEclipse 2015 Stable 2.0
- Tomcat 8.0
- JDK 1.7.0_80
- Hadoop 2.6.0
- Hive 1.2.1

- 1QuadCore CPU
- HDD(500GB,6Gbps, 7200rpm,16MB cache)
- 1000Mbps connectivity card


hadoop
Apache Ambari
System Deployment
Machine Learning
Mahout
Machine Learning
Distributed Programming
Pig
Scripting
Hive
SQL Query
Spark
Scheduling
Oozie
Workflow
Service Programming
Data Ingestion
Sqoop
Flume
Storm
MapReduce Framework-YARN
hadoop Map Reduce
NoSQL Database
APACHE HBASE
Distributed Filesystem
HDFS
Hadoop core
Zookeeper
Coordination

ORDINARY MULTILAYER PERCEPTRON (MLP) → **DEEP LEARNING**





Certain problems are to be expected:

• redundant data can cause overfitting & unrealistic predictions
• adding multiple hidden layers is not feasible
  ordinary CPUs
• layers of sigmoidal units may cause underestimations

- **MLP**
- **Training set**

INPUT VECTOR: laser intensity, laser wavelength, pulse duration, polarization, incidence angle, type of plasma, initial plasma density

DESIRED OUTPUT: maximum observable harmonics order, intensity values for different harmonics, harmonic wavelength, harmonics duration, harmonic conversion efficiency

- **MLP1 configuration:**

INPUT LAYER: 8 Adaline neurons
HIDDEN LAYERS: 2, all sigmoidal neurons, 12 neurons in each layer.
OUTPUT LAYER: 5 sigmoidal neurons

Batch training, Cost Function optimized with Steepest Descent

- **MLP2 configuration:**

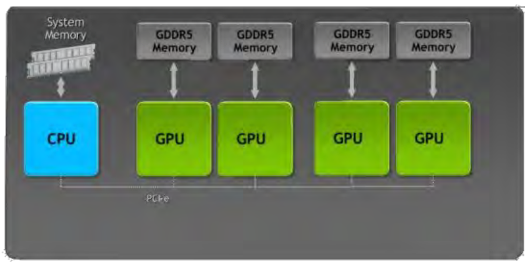
INPUT LAYER: 8 Adaline neurons
HIDDEN LAYERS: 3, all sigmoidal neurons, 10 neurons in each layer.
OUTPUT LAYER: 5 sigmoidal neurons

Batch training, Cost Function optimized with Resilient BKP

DEPLOYING A GRID SEARCH ALGORITHM

INPUT ← NUMBER OF DNNs TO BE TESTED

THEIR CONFIGURATION: LAYERS & UNITS



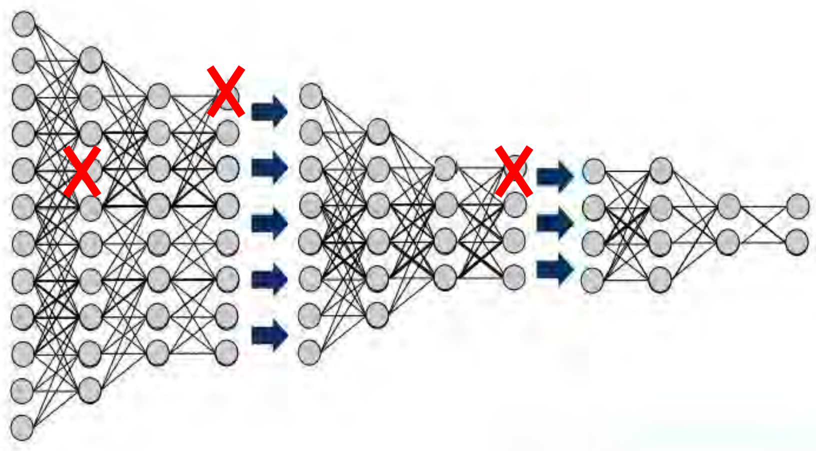
GeForge GTX Titan, 6GB memory

→ OUTPUT

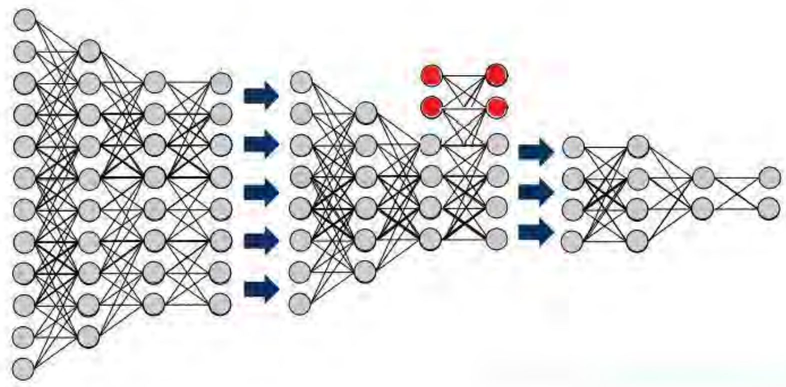
PERFORMANCES PREDICTIONS

DEPLOYING A GRID SEARCH ALGORITHM WITH DROPOUT AND CONSTRUCTIVE LEARNING

DROPOUT



CONSTRUCTIVE LEARNING

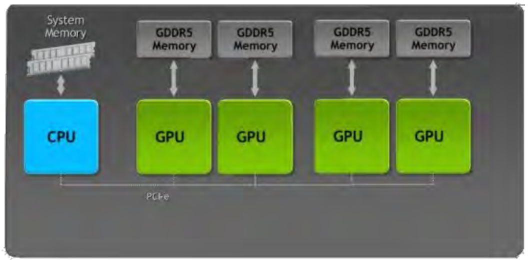


TOWARDS PREDICTING OPTIMAL INTERACTION SCENARIOS WITH DEEP LEARNING

INPUT

NUMBER OF DNNs TO BE TESTED

THEIR CONFIGURATION: LAYERS & UNITS



OUTPUT

PERFORMANCES PREDICTIONS

GTX Titan, 6GB memory, CUDA

MONITOR

Dropout

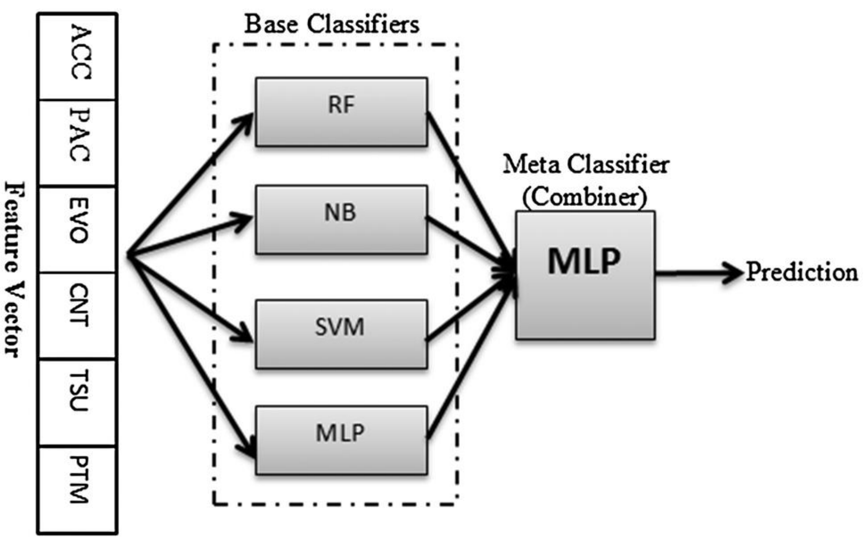
Construct

TOWARDS PREDICTING OPTIMAL INTERACTION SCENARIOS WITH DEEP LEARNING

Ensemble Learning:
The Wisdom of Crowds
(of Machines)

OPTIMIZE WITH “ENSEMBLE LEARNING”

Predicting protein–protein interactions between human and hepatitis C virus *via* ensemble learning
A. Emamjomeh, et al. ***Mol. BioSyst.***, 2014,**10**, 3147-3154



OUTPUT

PERFORMANCES
PREDICTIONS

Choose best 40
& average

Average over all

OUTPUT

TOWARDS PREDICTING OPTIMAL INTERACTION SCENARIOS WITH DEEP LEARNING

HHG INTERACTION SCENARIOS

Laser: intensity $2\cdot10^{18} W/cm^2$

$\lambda = 800nm$

pulse duration $\tau = 150\,fs$

Incidence angle $\alpha = 45$

polarization p

Plasma $n_e = 4n_c = 6.875\cdot10^{21} cm^{-3}$

DNN1: INPUT LAYER: 8 Adaline units
HIDDEN LAYERS: 20, all sigm., 12 neurons in each layer, except for layers 3 (11), 5(15), 6(12), 8(12),11(7).
OUTPUT LAYER: 5 sigmoidals
Batch training+ LM

	Highest Observable Harmonic				
	Max. Ord.	Intensity W/cm^2	Duration fs	Wavelength nm	Conv. Efficiency
Lit. Data [14]	50	$2\cdot10^{11}$	20	16	10^{-7}
PIC Data	58	$2.1\cdot10^{11}$	19	13.8	10^{-7}
MLP1	54	10^{11}	21	14.4	10^{-7}
MLP2	56	10^{11}	20	14.8	10^{-7}

	Highest Observable Harmonic				
	Max Ord.	Intens.	τ	λ	η
DNN1	52	$2.1\cdot10^{11}$	20	16	10^{-7}
DNN2	50	$2.02\cdot10^{11}$	20	15	10^{-7}
EL1	51	$2.18\cdot10^{11}$	19	15.8	10^{-7}
EL2	50.6	$1.98\cdot10^{11}$	21	16.3	10^{-7}

DNN2: INPUT LAYER: 8 Adaline units
HIDDEN LAYERS: 36, all sigm.,
14 neurons in each layer, except
for layers 2 (15), 6(12), 7(13),
9(12),12(16), 16(12), 18(13), 23(15),
24(11), 25(12), 28(12), 30(9), 31(13),
32(12), 35(7).
OUTPUT LAYER: 5 sigmoidals
Batch training+ LM

EL1: values mediated over 40 configurations
EL2: values mediated over all configurations

CONCLUSION & PERSPECTIVES

Flexibility & Modularity.
A versatile predictive system

Big Data = Better Model

Some issues with data redundancy and sparsity

Deep Learning is a performance booster, at the expense of computational complexity &run time

Graph Technology. Migrating from Hadoop to Spark.

Incorporating the GPUs into the cloud

CONCLUSION & PERSPECTIVES