BABEŞ-BOLYAI UNIVERSITY
Faculty of Mathematics and Computer Science

Master specialization

High Performance Computing and Big Data Analytics

Virginia Niculescu – June 2020

HPC & BDA

WHY

HIGH PERFORMANCE COMPUTING

TOGETHER WITH

BIG DATA ANALYTICS

?

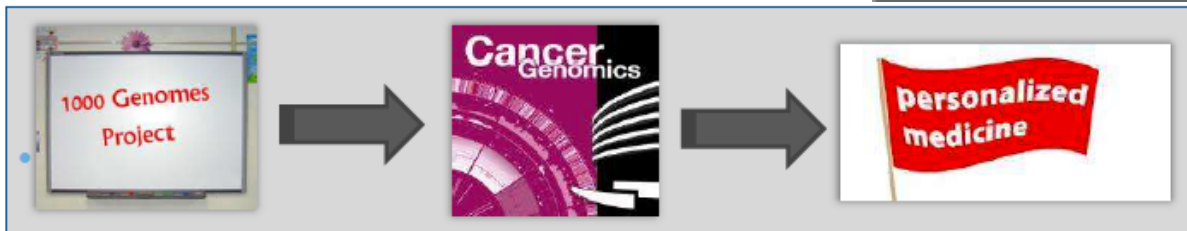# HPC (always) uses *big data*

- For a long period *big data* has been synonymous with

    High Performance Computing (HPC)

- HPC -> intensive computations with a big volume of data

- HPC leaded to the progress and evolution in High Performance Hardware

- *Simulation*: An Original (*big data*) HPC domain

    - Weather forecasting/climate modeling
    - Physics - molecular dynamics
    - Engineering & design
    - Finance
    - Life sciences: genomics, drug discovery…
    - Security
    - Automotive
    - …
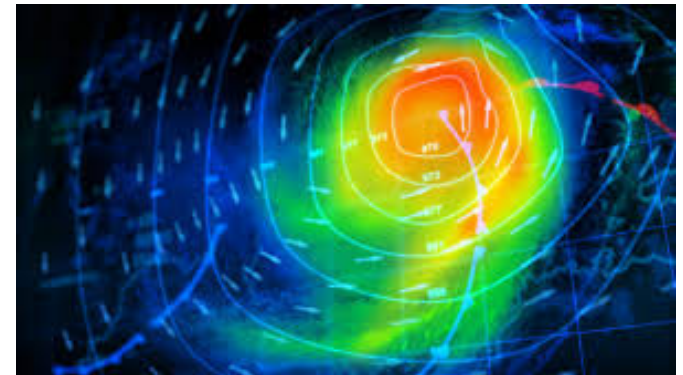
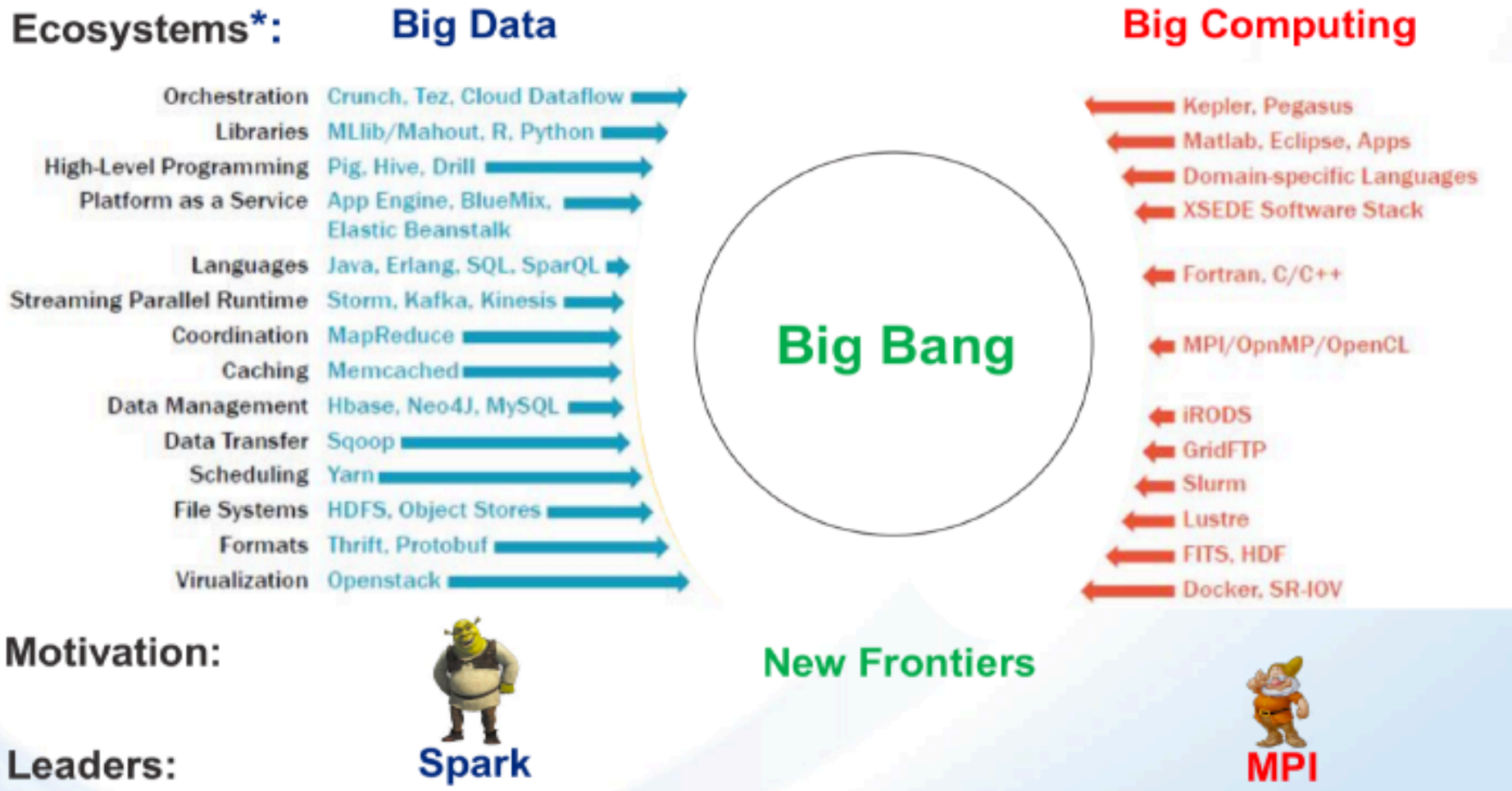# Analytics: a newer, complementary *big data* domain - BDA

- Analytics methods applied to established HPC domains in industry, government, academia ->

- High-end commercial analytics pushing up into HPC

- The path from science to industry and commerce can be relatively short:

Closing a Gap between Big Data and Big Computing

Ecosystems*: **Big Data**   **Big Computing**

	Big Data		Big Computing
Orchestration	Crunch, Tez, Cloud Dataflow →	←	Kepler, Pegasus
Libraries	MLlib/Mahout, R, Python →	←	Matlab, Eclipse, Apps
High-Level Programming	Pig, Hive, Drill →	←	Domain-specific Languages
Platform as a Service	App Engine, BlueMix, Elastic Beanstalk →	←	XSEDE Software Stack
Languages	Java, Erlang, SQL, SparQL →	←	Fortran, C/C++
Streaming Parallel Runtime	Storm, Kafka, Kinesis →		
Coordination	MapReduce →	←	MPI/OpnMP/OpenCL
Caching	Memcached →		
Data Management	Hbase, Neo4J, MySQL →	←	iRODS
Data Transfer	Sqoop →	←	GridFTP
Scheduling	Yarn →	←	Slurm
File Systems	HDFS, Object Stores →	←	Lustre
Formats	Thrift, Protobuf →	←	FITS, HDF
Virualization	Openstack →	←	Docker, SR-IOV

Big Bang

Motivation: **New Frontiers**

Leaders: **Spark**   **MPI**

*G. Fox at al. HPC-ABDC High Performance Computing Enhanced Apache Big Data Stack, CCGrid, 2015

Hardware - flow of interest

Data-Intensive HPC

ℵ *HPC vendors* are targeting commercial markets, driven by opportunity

Boundaries tend to dissolve

ℵ *Commercial vendors* are moving up to HPC, driven by customers

High-End Commercial Analytics

AI needs HPC and HPC needs AI

- **Scaling brings AI progress!**
  - Best AI algorithms scale with systems!
  - Make models bigger
  - Use more data
  - Reduce research cycle time

- **Machine learning for *scientific applications* is a growing area!**
  - *In the high energy and astrophysics communities, simulation could mean examining images using <u>deep learning</u>*
  - Machine Learning in HPC Environments 2019 - the 5th edition
    - held in conjunction with SuperComputing SC'19

HPC <-> BDA

- **HPC** world
    - realizes that there are more things in data storage than just files
- **BDA** world
    - realizes that HPC can really speedup analytics
        =>
        - All major public Cloud services now have an HPC offering.
        - Many academic HPC centers offer Cloud infrastructures and BDA tools.

- **HPDA — High Performance Data Analytics**
    - using HPC for data-intensive challenges
    - simulation + newer high-performance analytics

HPC & BDA — master specialization

Topics covered

Semester	High Performance Computing	Data Science
I	Distributed Operating Systems Concurrent Processes Modelling	Machine Learning Advanced Data Analysis
II	Parallel Programming Grid, Cluster and Cloud Computing	Data Visualization Formal Concept Analysis Data Streaming
III	Distributed & GPU Programming Optimisation models	Big Data Processing Data Mining Statistics
IV	Research Project Research Internship Dissertation Thesis	

৯ Computing infrastructure

Babeş-Bolyai University's HPC hybrid infrastructure (http://hpc.cs.ubbcluj.ro)
• HPC Cluster: IBM NextScale
• Private Cloud: IBM Flex System

Useful information

- Master presentation:
  - http://www.cs.ubbcluj.ro/invatamant/programe-academice/masterat/calcul-de-inalta-performanta-si-analiza-volumelor-mari-de-date/

- Syllabus:
  - https://www.cs.ubbcluj.ro/apps/fise/viewSyllabi.php?an=2020&lang=ro&specializare=HPC

- Admission procedure:
  - http://www.cs.ubbcluj.ro/admitere/nivel-masterat/admiterea-la-facultatea-de-matematica-si-informatica-nivel-master/

Sustained by Bosch starting from 2020 !!!

High Performance Computing and Big Data Analytics

Master Program in:

• Data Science and Big Data

• Parallel and Distributed Programming

Master Program organized by Babeș-Bolyai University, Faculty of Mathematics and Computer Science in partnership with Bosch Engineering Center Cluj