« Apprentissage automatique et Big Data: les défis de la mise à l'échelle »

Gianluca Bontempi
ULB Machine Learning Group,
mlg.ulb.ac.be

ULB

MLG
MACHINE LEARNING GROUP

ULB Machine Learning Group

Directors: Pr. Gianluca Bontempi, Pr. Tom Lenaerts,
4 academics, 4 postdocs, 10 PhD students

Research topics

- Big Data Mining
- Scalable machine learning
- Spatio-temporal forecasting
- Bioinformatics and Computational Biology
- Multiagent, game theory

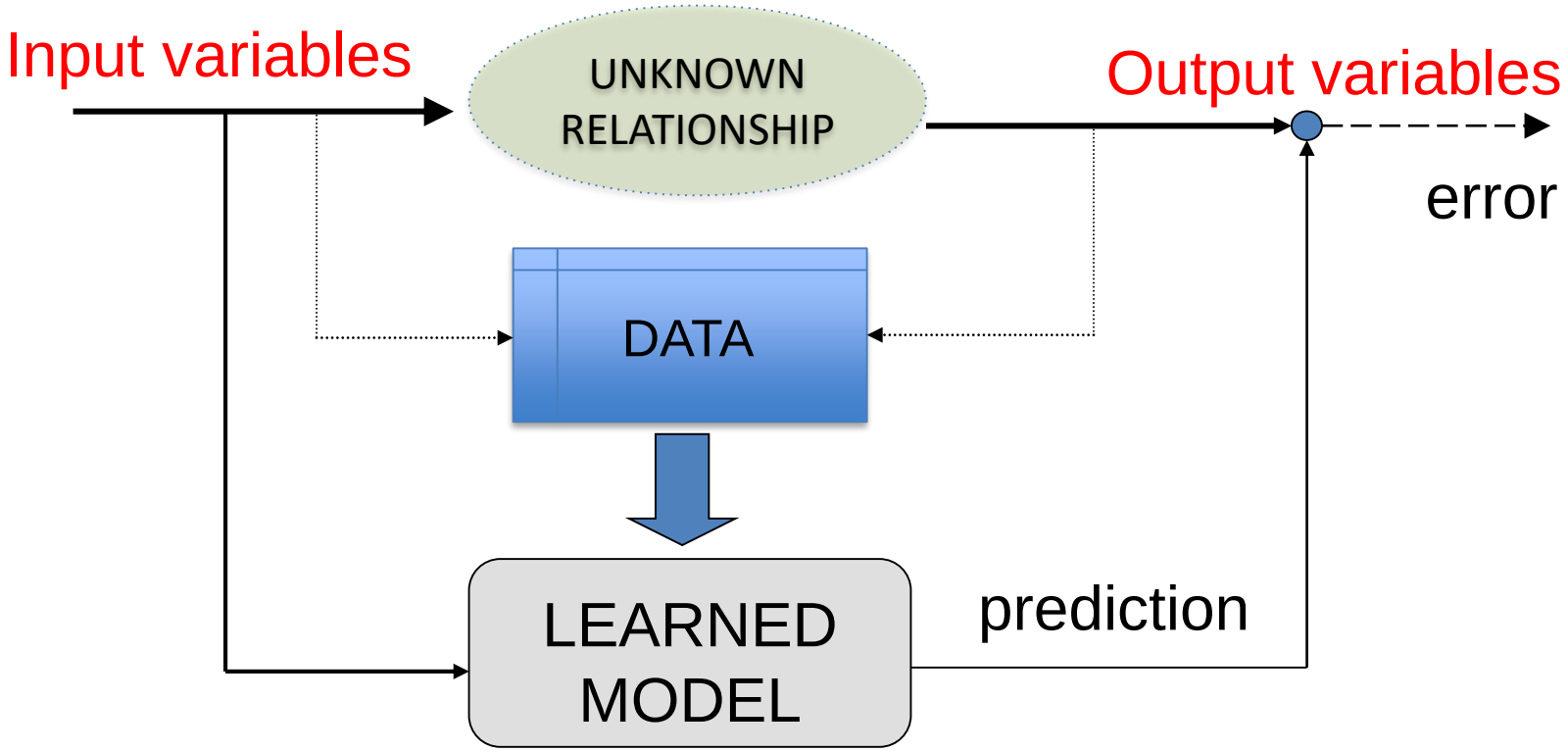
Application domains

- Fraud detection (in collaboration with ATOS Worldline)
- Finance
- Genomics and Biomedical sciences
- Cryptoanalysis, cybersecurity
- Smart cities

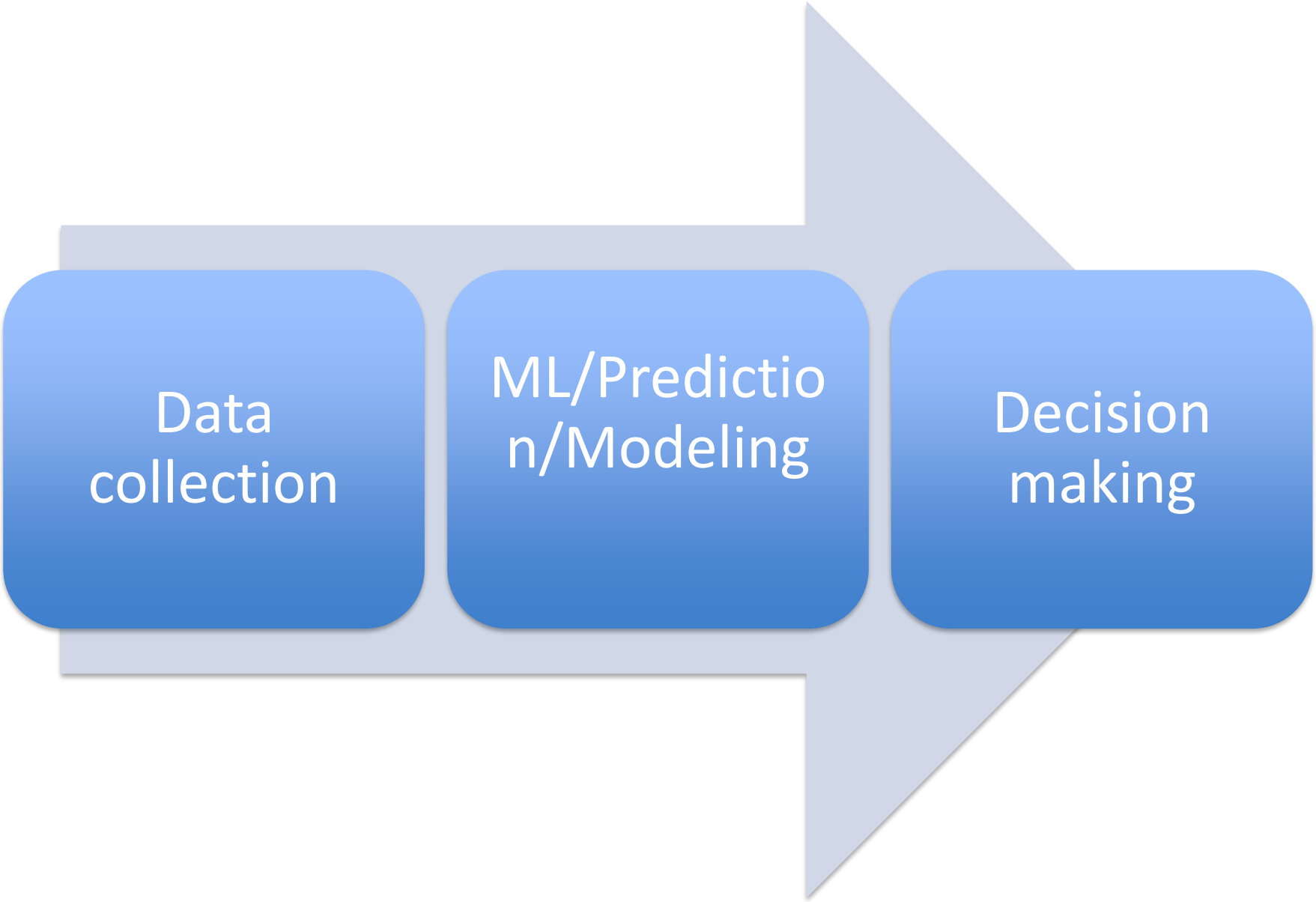
Facilities

- Big data cluster
- Wireless sensors
- Experimental economics lab

ML for the dummies



ML in the decision process



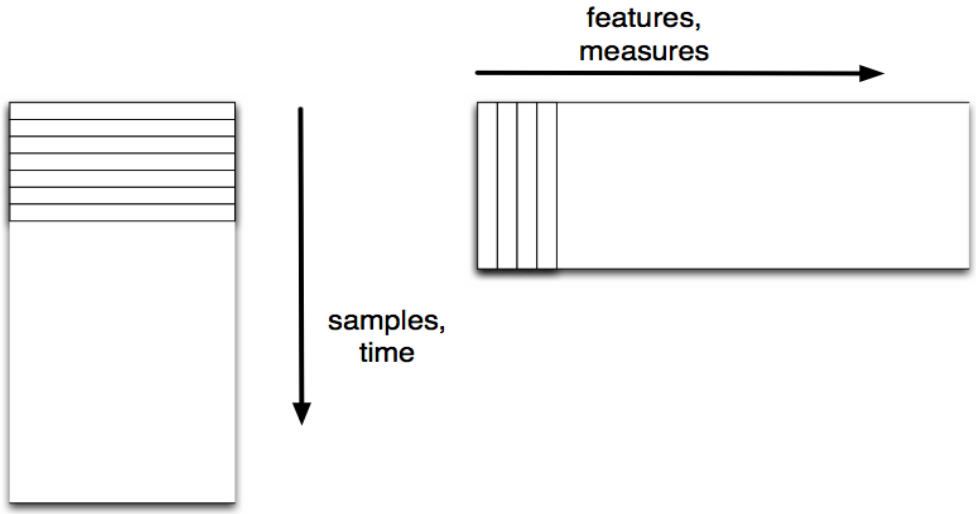
Countless number of applications

- Predict the performance of a aeronautic device on the basis of a set of parameters
- Predict whether you will like a film/movie (collaborative filtering)
- Assign keywords to articles and news in order to better classify them.
- Classifying credit applicants as low, medium, or high risk.
- Determining which home telephone lines are used for Internet access.
- Figuring out which customers are likely to stop being customers (*churn*).
- Estimating the value of a piece of real estate
- Predicting which CARREFOUR clients will be more interested to a discount in Italian products.
- Predict the probability that a company is employing black workers (social anti-fraud detection)
- Classify satellite images in civil and military sites.
- Predict which machine is most likely to be the next to fail.
- Predict the next value of a time series.

Recent MLG projects

- MOBI-AID: Brussels Mobility Advanced Indicator Dashboard
- **BruFence: scalable machine learning for automating defense systems**
- **BRiDGEIris: BRussels big Data platform for sharing and discovery in clinical Genomics**
- Adaptive real-time machine learning for credit card fraud detection.
- ICT4REHAB - Advanced ICT Platform for Rehabilitation
- ARMURS - Automatic Recognition for Map Update by Remote Sensing.
- OASIS - Detection and analysis of social fraud in Social Security Databases.
- Integrating experimental and theoretical approaches to decipher the molecular networks of nitrogen utilisation in yeast.
- TANIA - Système d'aide à la conduite de l'anesthésie.
- PIMAN - Pôle de compétence en Inspection et Maintenance Assistée par langage Naturel.
- Predictive data mining techniques in anaesthesia.
- AIDAR - Adressage et Indexation de Documents Multimédias Assistés par des techniques de Reconnaissance Vocale.
- Time series prediction of Belgian car market.

Challenges of scalable ML



Driving factors for accumulation:

- Vertical: streaming data, sensor measurements, process monitoring, security, financial transactions
- Horizontal: new measurement technologies (sensors technology, sequencing)

Challenges:

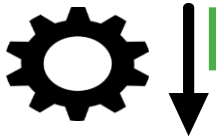
- Vertical: prediction, concept drift detection, model updating
- Horizontal: dimensionality reduction, (causal) feature selection

Horizontal big data: BridgeIRIS

- Genotype (input) phenotype (output) association
- Huge number of variables (10^7), thousands of samples

Input: Variant dataframe

Sample_ID	Chr	Position	Reference	Alternative	Zygosity	Gene Symbol	…
HG03837	12	62114671	T	C	1|0	FAM19A2	…
HG03690	16	19051520	A	G	0|1	TMC7	…
HG02072	22	21829513	TTGTC	T	1|1	TMEM191C	…
HG02052	6	99771540	T	C	0|1	PDCD2	…
…	…	…	…	…	…	…	…

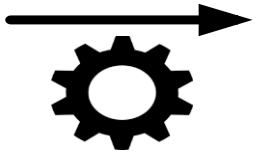


1) Filtering

Genotype matrix

	s_1	s_2	s_3	s_4	s_5	s_6	…
v_1,g_1	1	0	0	1	0	0	…
v_2,g_1	0	0	2	0	0	1	…
v_3,g_1	0	0	1	0	0	0	…
v_4,g_2	0	2	1	0	0	0	…
…	…	…	…	…	…	…	…

2) Scoring



Output: Scoring matrix

Scope	Score	…
g_{732}	14	…
g_{1756}	12	…
g_{91}	12	…
g_{661}	11	…
…	…	…

DiGeST
Distributed Gene & variant Scoring Tool

Define
- Control group
- Case group
- Scoring function

Start ranking

RESTful API

Return
- Variant/Gene ranking
- Additional statistics

Retrieve results

RESTful API

1) Variant filtering

Highlander genomics DB

Samples/Variants genotypes matrix

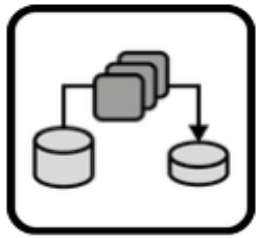
	v_1	v_2	…	v_S
s_1	0	1	…	0
s_2	1	2	…	2
…	…	…	…	…
s_n	1	0	…	2

0: Homozygous REF
1: Heterozygous
2: Homozygous ALT

Currently ~2700 exomes/genomes available.
~11.10^9 variants

2) Apply scoring function

Spark distributed processing



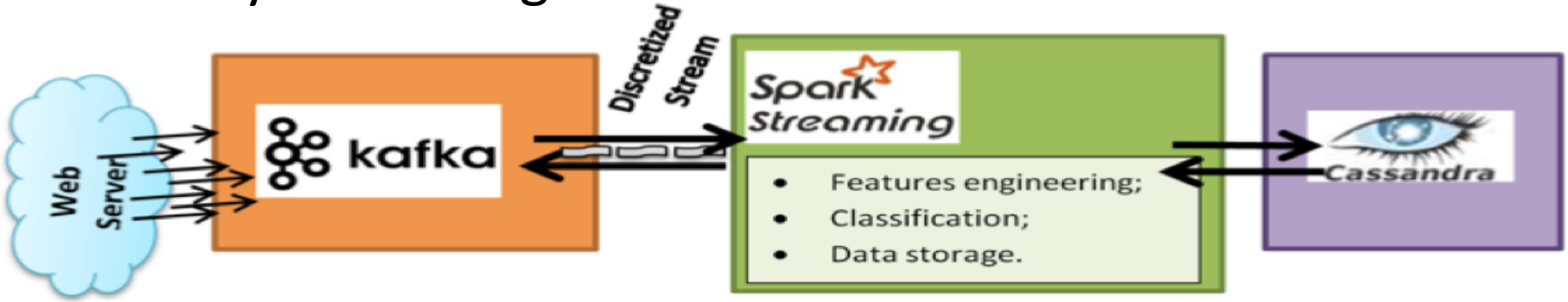
	Rank
v_1	r_1
v_2	r_2
…	…
v_N	r_N

Currently scoring for
- **Variant**
- **Gene**

Monogenic and digenic

Vertical big data: BruFence

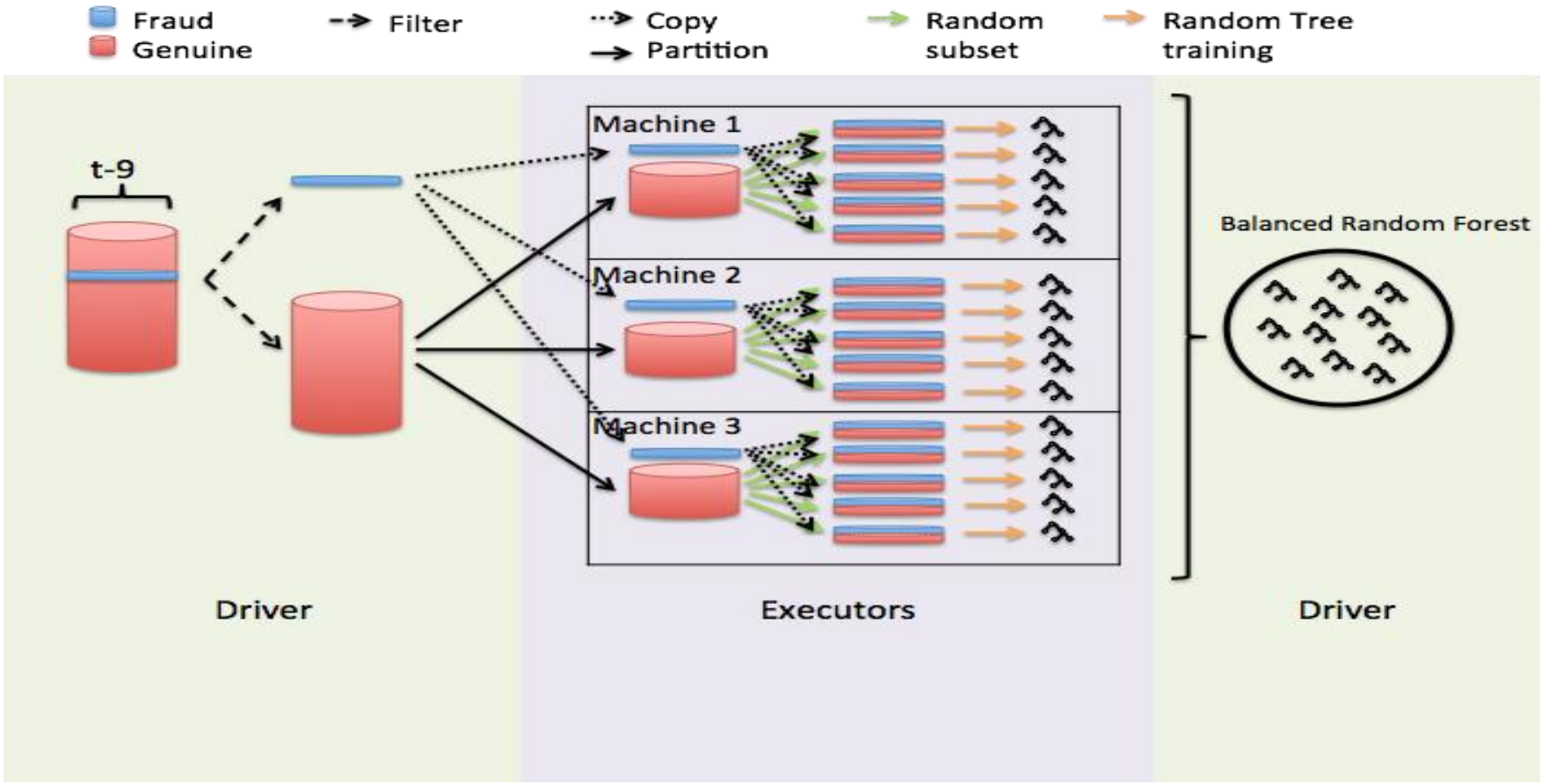
- Credit-card fraud detection in nearly run time
- Massive amounts of streaming data (~200000 tx/day)
- Unbalancedness
- Nonstationarity/ concept drift
- Delayed labeling of transactions



Collaboration with ATOS Worldline

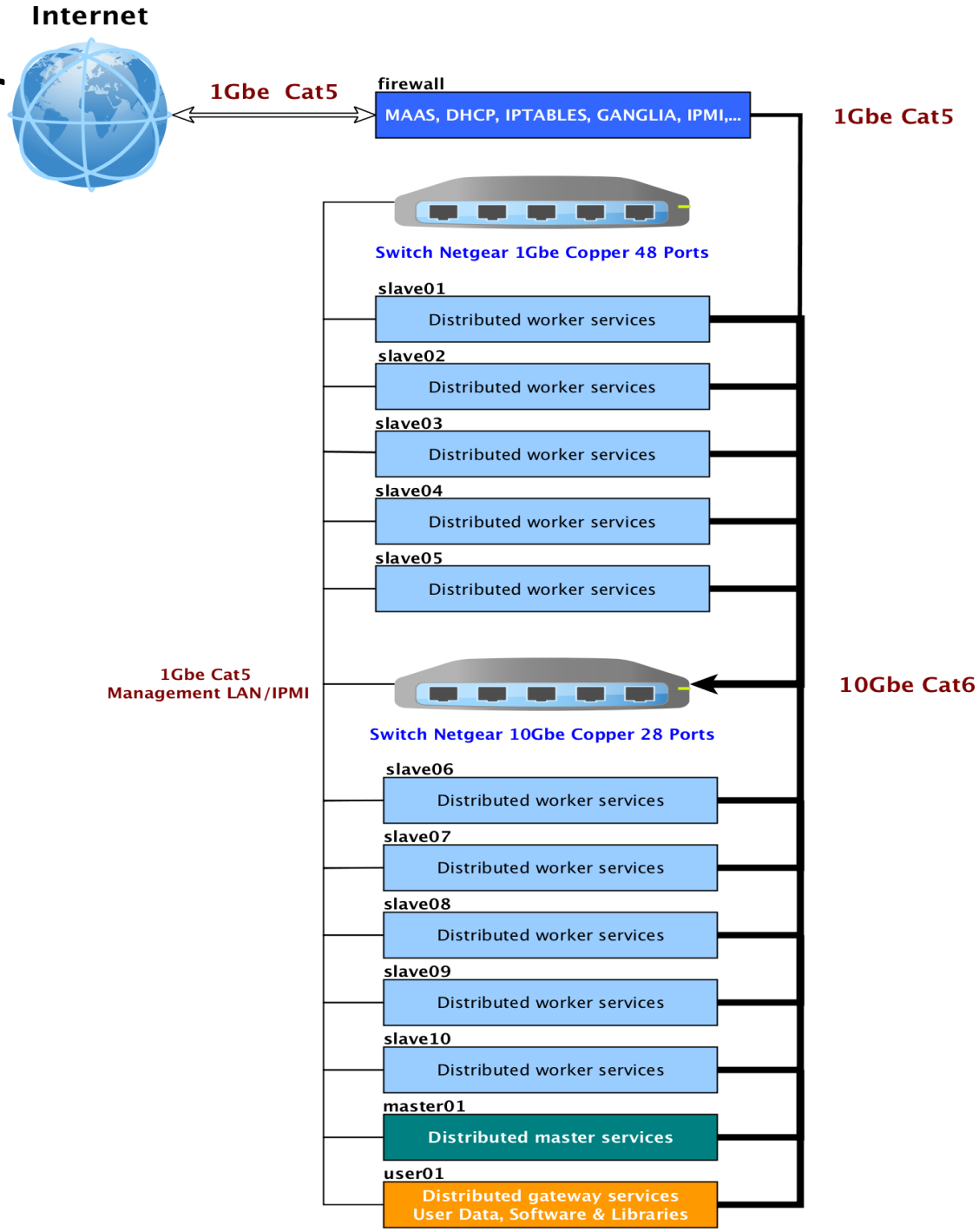
Distributed machine learning

- Map-reduce distribution of state-of-the-art learning algorithms



MLG Big data cluster

- Hardware
  - 240 cores
  - 1.2 TB RAM
  - 260TB Disk
  - 10Gb/s network
- Software
  - Cloudera Hadoop
  - Spark
  - Cassandra
  - H2O



Big data: opportunities and risks

- Opportunities
  - Integration of heterogenous sources of information
  - Continuous learning
  - Better, faster predictive models
  - From analytical to data-driven science
  - Validation based science
- Risks
  - Excessive sense of confidence
  - Spurious causal inference
  - Ethical issues

Opportunities of collaboration

- Internships, Master thesis
- Joint research projects (FIRST Entreprise, Walloon projects)
  - Spatio-temporal forecasting
  - Classification, prediction
  - Big data analysis
  - Dimensionality reduction
  - Analysis of wireless sensor data
- Training
  - Data mining
  - Open source
  - Big data technologies (Spark, Hadoop)

Contact

Pr. Gianluca Bontempi

Pr. Tom Lenaerts

Machine Learning Group, Computer Science Dept. ULB

mlg.ulb.ac.be

www.facebook.com/mlgulb

Interuniversity Institute of Bioinformatics in Brussels

ibsquare.be