MathMOuse: A Mathematical MOdels WarehoUSE to handle both Theoretical and Numerical Data

Context

Time Series in Experimental Sciences
- Numeric technologies evolution → increase in the volume of data to process.
- In automatic control: time series processing → differential equations.
- Increasing volume of data → increase of derived differential equations.

Problems

Database as service for data scientists?
- Time series in the context of databases are stored as raw numerical data
  - → The models resulting from the scientists work are not stored.
  - → Storing and organizing models to ease their management and retrieval became essential.
- A storage of the models would have a better interest
  - → How to exploit the models database to aid the retrieval of models from time series?

Proposal
- MathMOuse: an enriched database storing differential equations
- MathMOuse has a « Query by data » system: time series are used to query the database and retrieve relevant equations
- It contains a GUI to allow user to load models, navigate through them, visualize their data and perform queries, etc

Example of a time series

Timestamp (ms)	Value
0.0000000e+00	0.0000000e+00
1.0000000e+00	9.4105346e-02
2.0000000e+00	1.8452077e-01
3.0000000e+00	2.7139095e-01
4.0000000e+00	3.5485491e-01
5.0000000e+00	4.3504619e-01
6.0000000e+00	5.1209313e-01
7.0000000e+00	5.8611902e-01
8.0000000e+00	6.5724231e-01
9.0000000e+00	7.2557682e-01
...	...

Time Series

Time Series are
- Tables of chronological values
- Usually, measurements (sampling of analog signal)

Times Series as CSV files

Equation example
- Ty'+y=Ku
- T = 25,0
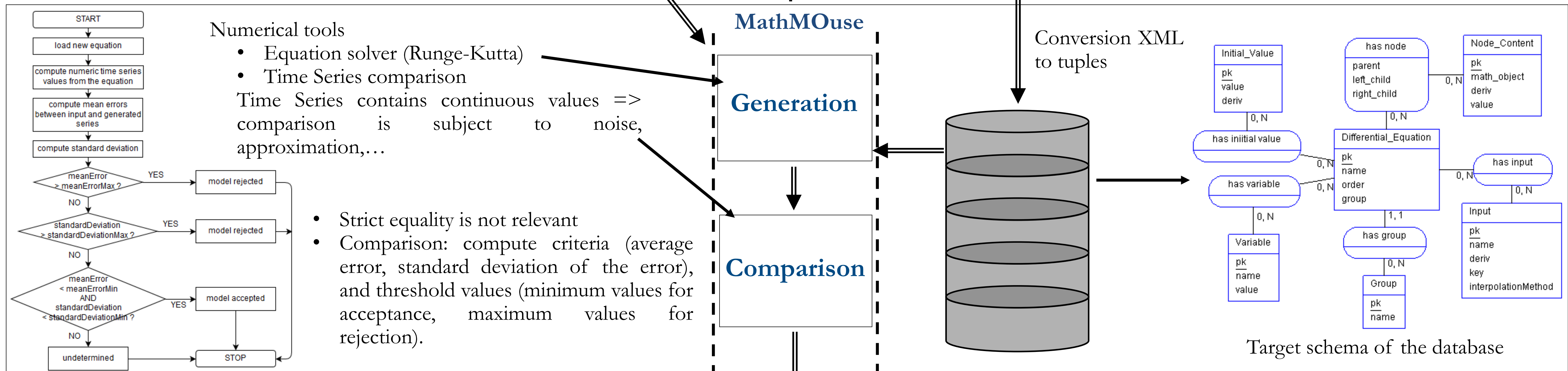- K = 2,4
- y(0) = 0
- u known input function

Models

Models are
- Differential equations
- Ordinary
- Linear (storage structure allow non-linearity, but numerical solving is more complex)

Differential equation model instance always contains
- An explicit formula
- Values for literal variables
- Initial conditions
- Values of the input functions

Equations data as XML files

MathMOuse

Numerical tools
- Equation solver (Runge-Kutta)
- Time Series comparison

Time Series contains continuous values => comparison is subject to noise, approximation,…

Generation

- Strict equality is not relevant
- Comparison: compute criteria (average error, standard deviation of the error), and threshold values (minimum values for acceptance, maximum values for rejection).

Comparison

Conversion XML to tuples

Target schema of the database

Accepted Models

Models with average and standard deviation below minimum acceptance threshold

Example, minimum acceptance 10^{-3}
- Equation: $25,0\, y' + y = 2,4\, u$
- Statistics (Performances):
  - → Average error: $3,9362 \times 10^{-8} < 10^{-3} \Rightarrow$ *accept*
  - → Minimum error: $9,6053 \times 10^{-10}$
  - → Maximum error: $8,2054 \times 10^{-8}$
  - → Standard deviation: $1,9528 \times 10^{-8} < 10^{-3} \Rightarrow$ *accept*

Undetermined Models

Average and/or standard deviation between minimum and maximum thresholds

Example, minimum acceptance 10^{-3}, maximum rejection 1
- Equation: $16,93\, y' + 17,49\, y = 9,52\, u$
- Statistics (Performances):
  - → Average error: $0,8567 (> 10^{-3} \wedge < 1) \Rightarrow$ *unsure*
  - → Minimum error: $0,6740$
  - → Maximum error: $2,0531$
  - → Standard deviation: $0,1042 (> 10^{-3} \wedge < 1) \Rightarrow$ *unsure*

Rejected Models

Average or standard deviation above maximum rejection threshold

Example, maximum rejection 1
- Equation: $10,96\, y^{(5)} + 9,28\, y^{(4)} + 19,95\, y^{(3)} + 11,48\, y^{(2)} + 2,31\, y^{(1)} + 1,82\, y = 0$
- Statistics (Performances):
  - → Average error: $2,0508 \times 10^{15} > 1 \Rightarrow$ *reject*
  - → Minimum error: $6,8534$
  - → Maximum error: $2,0508 \times 10^{15}$
  - → Standard deviation: $2,5146 \times 10^{15} > 1 \Rightarrow$ *reject*

Implementation

Technical Setup
- Docker, Docker-compose
- Postgresql
- Microservices developed in Java
- Broker rabbitmq
- https://forge.lias-lab.fr/projects/mathmouse

Tests

Set of data used
- 100 randomly generated differential equations (order up to 5)
- A set of Time Series, derived from the above random equations, with up to 5000 elements

Parallel computing
- Microservices → comparison and generation can be parallelized
- Prototype tested with different numbers of comparison and generation parallel processes

Execution time, 1 series of 1000 values, 100 models in the database

Participants

Cyrille PONCHATEAU
PhD Student
ponchateau@ensma.fr

Ladjel BELLATRECHE
Prof. at l'ENSMA
bellatreche@ensma.fr

Carlos ORDONEZ
Prof. at University of Houston
ordonez@cs.uh.edu

Mickaël BARON
Research associate
baron@ensma.fr
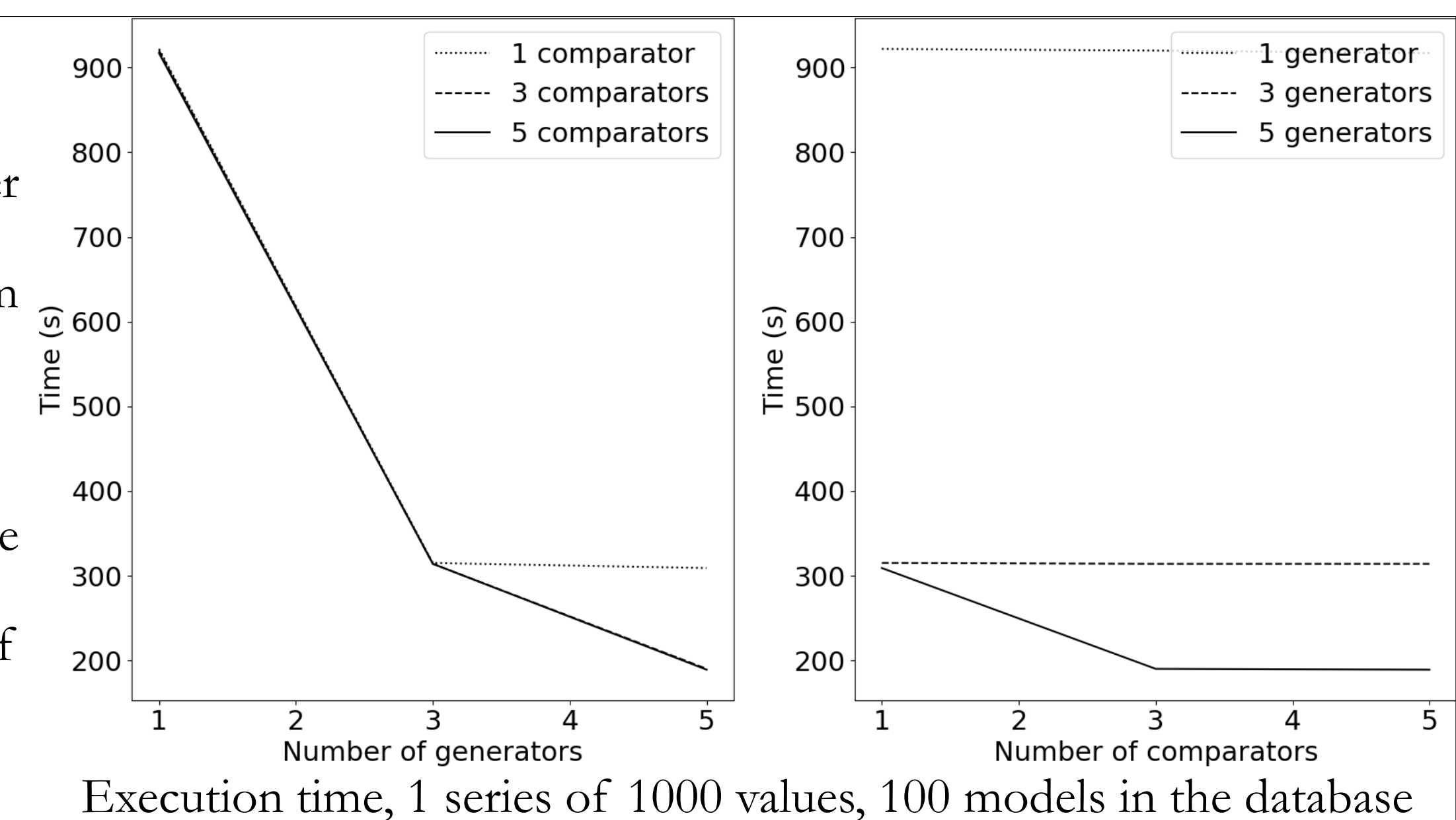
To know more about it…

Related Publications
- Cyrille PONCHATEAU, Ladjel BELLATRECHE, Mickael BARON, Entrepôt de Données dans l'air Data Science : De la Donnée au Modèle, (EDA 2016), June, 2016
- Cyrille PONCHATEAU, Ladjel BELLATRECHE, Carlos ORDONEZ, Mickael BARON, A Database Model for Time Series : From a traditional Data Warehouse to a Mathematical Models Warehouse, 32e journées Bases de Données Avancées (BDA), Poitiers, 2016