INtelligeNt ApplicatiOns oVeR Large ScAle DaTa StrEams

INNOVATE

Dr Kostas Kolomvatsos
Research Fellow

Monitoring Meeting MSCA-IF
June 17-18, 2019
Brussels, Belgium
The INNOVATE Team

The Fellow

Dr Kostas Kolomvatsos
Kostas.Kolomvatsos@glasgow.ac.uk
https://www.gla.ac.uk/schools/computing/staff/kostaskolomvatsos/

The Supervisor

Dr Christos Anagnostopoulos
Christos.Anagnostopoulos@glasgow.ac.uk
https://www.gla.ac.uk/schools/computing/staff/christosanagnostopoulos/

The Host Institution

School of Computing Science
University of Glasgow

https://www.gla.ac.uk/schools/computing/

INNOVATE Website: https://sites.google.com/view/mscainnovate/home
ESSENCE

Pervasive & Distributed Intelligence

Distributed Intelligence
Self-organization Algorithms for UxVs
Edge-centric Statistical Learning
Funding: H2020/GNFUV

Network-centric Stream Processing
Delay-Tolerant Data Stream Processing
Time-optimized Task Offloading
Edge-centric Selective Analytics
Funding: H2020/MSCA INNOVATE

Predictive Computing
Query-driven Predictive Analytics
Data Relevance: Relevant Data is Big Data
Dataless Explanation & Exploitation of Analytics
Funding: UK EPSRC/CLDS (£3M)

Collaboration with Industry & Academia
- Hesso Geneve (CH)
- Repado Ltd (CH)
- inCITES Sarl (LU)
- BMW Group Research (DE)
- BT (UK)
- Huawei (CN)

http://www.dcs.gla.ac.uk/essence/
Research Overview

**Query Driven Applications**
Analytics offer the basis for decision making. Analytics should be executed on top of multiple data partitions.

**Queries Management**
Massively allocate queries to distributed datasets. Efficiently aggregate multiple query responses. Maximize the performance and support time critical applications.

**Management of the Ecosystem**
Query Controllers (QCs) manage the incoming queries. Distributed nodes host the data. Query Processors (QPs) execute queries in every node.

**Intelligent Behaviour**
Allocate queries to nodes. Support nodes management. Support data management. Support the behaviour of QCs.
INNOVATE offers intelligent mechanisms for the management of queries, data and distributed nodes.
Research Objectives

1. Design & implement Query and Nodes' Models
2. Design & Implement Individual Learners
3. Create a Pool of Learners and Implement an Ensemble Learning Scheme
4. Design & implement the Queries Allocation Process
5. Design & implement the Multiple Controllers Management Plane
6. Develop a holistic approach to research, training and career evolution of the Fellow
7. Disseminate and Exploit INNOVATE outcomes
## INNOVATE Timeline

<table>
<thead>
<tr>
<th>Subjects / Quarters</th>
<th>Year 1</th>
<th></th>
<th>Year 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q1</td>
<td>Q2</td>
<td>Q3</td>
</tr>
<tr>
<td>SotA (M1-M3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Queries and QPs Modeling (M1-M3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learners – Pool of Learners (M2-M8)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ensemble Learning Scheme (M7-M13)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uncertainty Management (M11-M17)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Testing and Evaluation (M16-M24)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCDP Updates</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Steps

**A HOLISTIC FRAMEWORK**
Models, algorithms and methodologies for intelligent query/task management

**MODELLING**
Models for queries and nodes/QPs

**MANAGEMENT**
Schemes for data and nodes management

**ALLOCATION**
Methods for allocating queries and data

**INTELLIGENCE**
Advanced methodologies for the management of the ecosystem
We propose a model for delivering the complexity class. We propose a *Fuzzy Classification Process* (FCP). The FCP depicts the ‘membership’ of a query in a pre-defined set of classes. We adopt IR techniques. We build on an *ensemble similarity scheme*. We estimate the number of steps required for executing a query. We consider a queue in every node. The size and the rate of the incoming queries/tasks affect the load. Based on the contextual information, we build on the *Probability of Allocation* (PoA). The PoA depicts the ‘ability’ of a QP to execute a query smoothly. The highest PoA(s) win(s).
We also focus on additional contextual information
- Query/task priority
- Available resources
- Status of peer nodes
- Data present locally and in peers
- We propose a local decision making mechanism for allocating queries/tasks

We define the query/task contextual vector
We propose a sequential decision making
Every query/task can be executed locally or at peers

We propose a Bayesian classifier for deciding if a query/task could be executed locally

We define the information vector for peers
We focus on their datasets, the communication cost, the available resources

For selecting the appropriate peer, we adopt a multi-criteria optimization methodology
We adopt the VIKOR method
We extend our findings taking into consideration:

- a more complex decision making scheme
- the 'historical' performance of each node

For deciding a local execution, we adopt a $k$NN classifier.

Peers are selected based on a model retrieved by the utility theory.

We provide formulations for estimating the short term and long term load of each node.

We provide formulations for calculating the probability of a local execution.
We propose a mechanism for data management at every node.

We offer a pre-processing distributed scheme that decides where data should be allocated.

We focus on the accuracy of data.

We want to identify and manage the error between the incoming data and the available datasets.

The proposed scheme proactively ‘prepares’ the data before any query is applied.

We define a model that identifies if the incoming data deviate from the ecosystem.

If not, data are allocated to the appropriate dataset.

If yes, data are rejected.

Our model consists of two parts:

- **The accuracy violation detection scheme (AVDS)**
- **The Partition identification scheme (PIS)**

AVDS calculates the probability of a data vector deviates from the ecosystem.

We provide formulations for delivering the probability based on a *finite mixture of distributions*.

PIS adopts an uncertainty driven decision making.

We propose a *Fuzzy Logic controller* for resulting the appropriate node.
Nodes’ Management

- Nodes convey software and firmware for performing tasks
- We propose a distributed software update scheme
- We avoid the disadvantages of legacy, centralized systems
- Nodes monitor specific KPIs and independently decide when they will initiate the update process

Nodes monitor their internal status (e.g., load, resources)
Nodes monitor the network’s performance (e.g., bandwidth, errors)

We consider proportional and non-proportional metrics
We calculate the reward for each metric realization

We adopt a time-optimized decision making mechanism
We adopt the principles of the Optimal Stopping Theory
We build on the expected reward maximization

Our model exhibits when to stop the monitoring process and initiate the update
We adopt a set of learners
They are trained with real and synthetic data
We propose a **meta-ensemble learning scheme** using the following (ensemble) models:
- AdaBoost
- Stacking
- Bagging
The (sub-)ensemble schemes are combined with the **One-Over-All (OVA) technique**
Advanced Models

**Uncertainty Management**
- We manage the uncertainty about optimal allocations
- We propose the use of Type-2 Fuzzy logic
- We combine Fuzzy Logic with a machine learning model

**Automated Knowledge Extraction**
- We adopt machine learning for generating parts of the Fuzzy Logic model
- We automatically deliver the Type-2 Fuzzy Sets and their membership functions
- We provide mathematical formulations for the new scheme
Ongoing Work

A Probabilistic Model for Allocations
We build on our modeling
We study the expected load of QPs
We propose the concept of the optimal node

Ecosystem Management
We focus on multiple QCs-nodes/QPs
We apply different types of models
We adopt computational intelligence techniques

Data Synopses Management
We propose a scheme for sending data synopses to peers

Extension of the Fuzzy Logic Model
We study the effect of data on the Footprint of Uncertainty (FoU) in Type-2 Systems
We aim to provide a fully automated data driven uncertainty management scheme
Journal Publications

6. Kostas Kolomvatsos, Christos Anagnostopoulos, Maria Koziri, Thanasis Loukopoulos, ‘Proactive & Time-Optimized Data Synopsis management at the Edge’, in preparation to be submitted in IEEE Transactions on Knowledge and Data Engineering, 2019

Conferences/Posters/Book Chapters

1. K. Kolomvatsos, C. Anagnostopoulos, ‘An Edge-Centric Ensemble Scheme for Queries Assignment’, in 8th International Workshop on Combinations of Intelligent Methods and Applications in conjunction with the 30th International Conference on Tools with Artificial Intelligence, Nov. 5-7, Volos, Greece, 2018
2. K. Kolomvatsos, C. Anagnostopoulos, ‘In-Network Edge Intelligence for Optimal Task Allocation’, 30th International Conference on Tools with Artificial Intelligence, Nov. 5-7, Volos, Greece, 2018
5. K. Kolomvatsos, C. Anagnostopoulos, ‘Intelligent Applications over Large-Scale Data Streams’, The Scottish Informatics & Computer Science Alliance (SICSA), DemoFest, Edinburgh, Scotland, Nov. 6th, 2018

* Monograph
INNOVATE in Numbers

- Participation in Supervision Activities
  - Participation in the supervision of 2 MSc and 1 PhD students

- Invited Talks / Guest Lectures
  - Three (3) Invited Talks / Guest Lectures

- Objectives achieved in the 1st Year
  - 70% of the initially planned (total) objectives are fulfilled
Thank You