University of Glasgow

School of Computing Science

INtelligeNt ApplicatiOns oVer Large ScAle DaTa StrEams

INNOVATE

Dr Kostas Kolomvatsos

Research Fellow

Monitoring Meeting MSCA-IF

June 17-18, 2019

Brussels, Belgium

Outline

01 Introduction

02 Objectives

03 Timeline and Initial Plan

04 Research Outcomes

05 Ongoing and Future Work

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

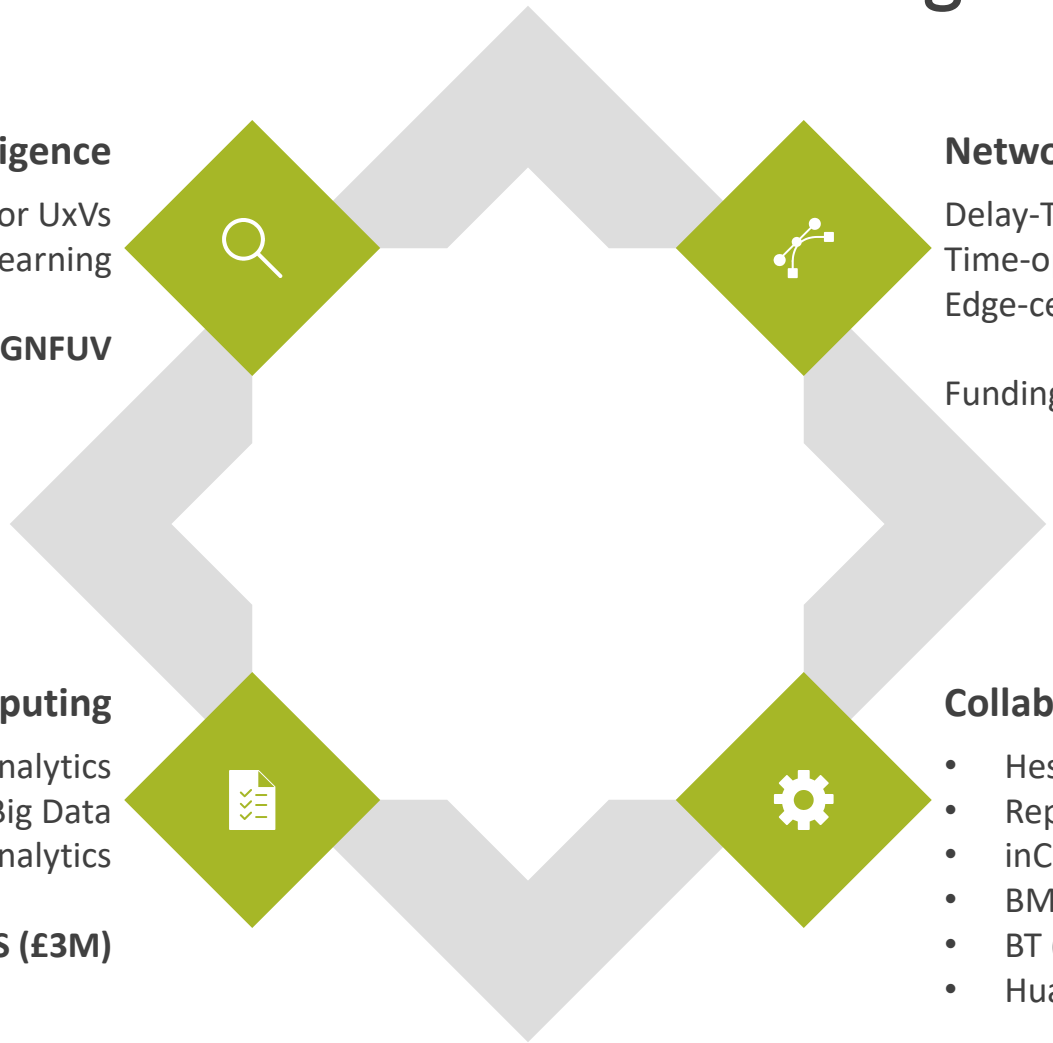
Self-organization Algorithms for UxVs
Edge-centric Statistical Learning

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

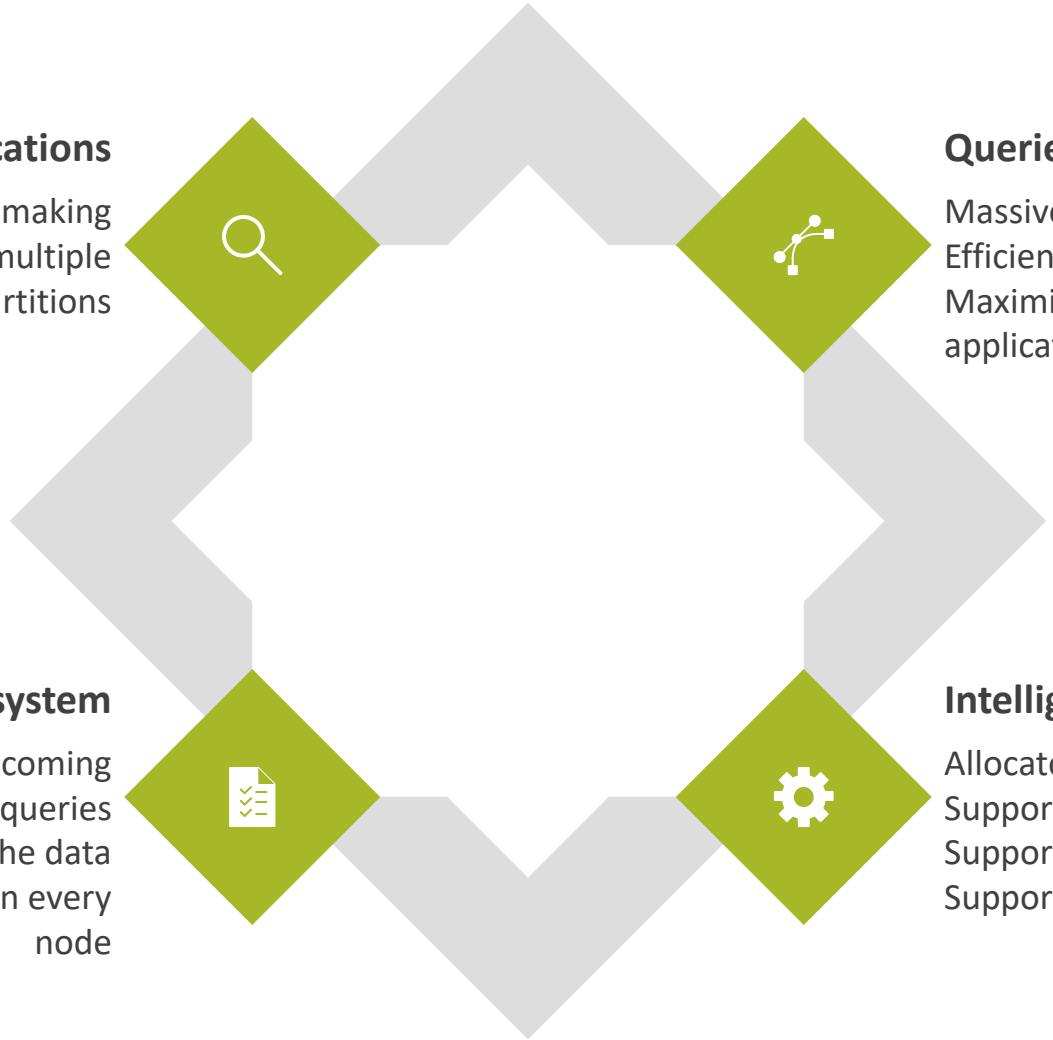
Funding: **UK EPSRC/CLDS (£3M)**

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

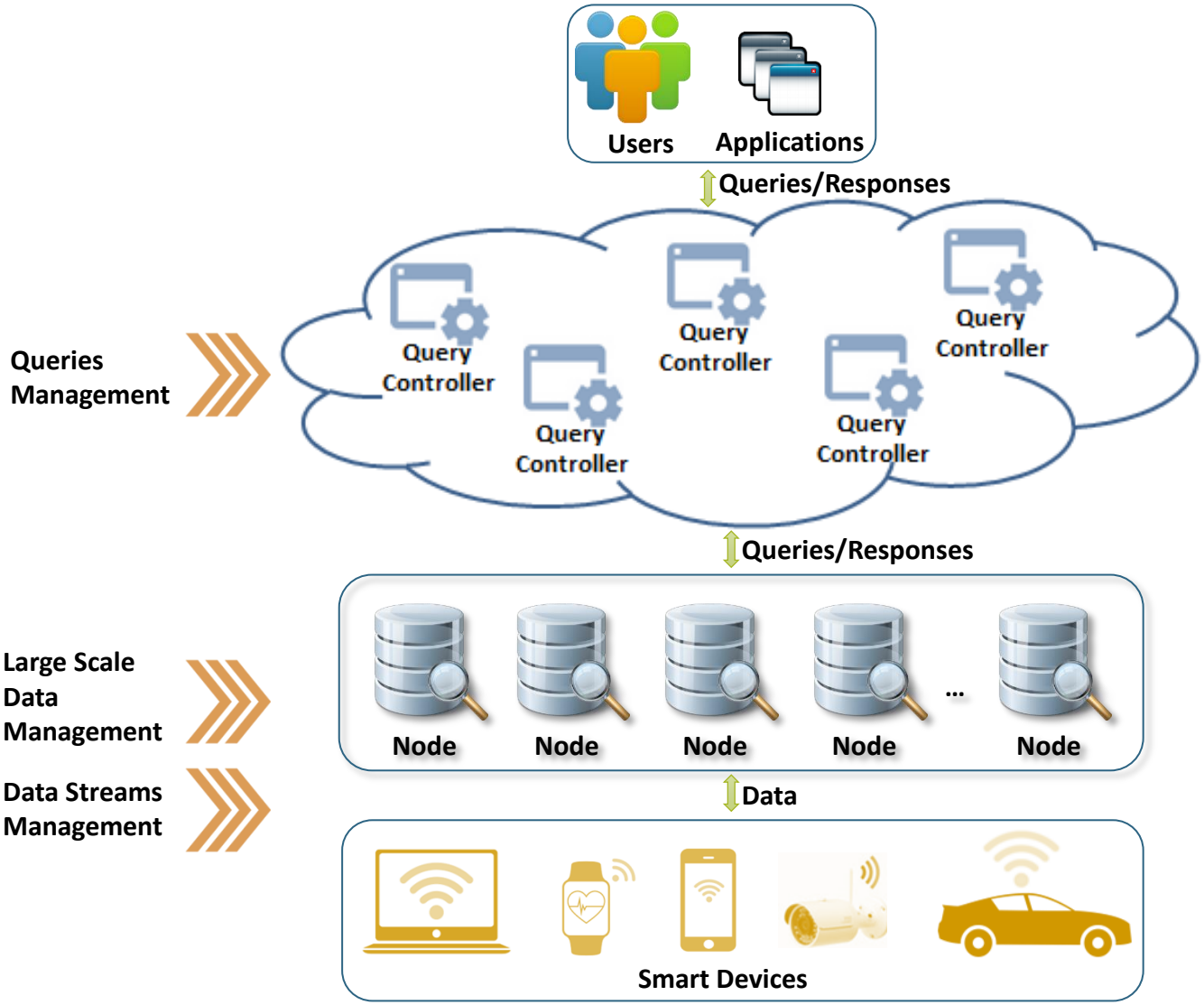
Management of the Ecosystem

Query Controllers (QCs) manage the incoming queries
Distributed nodes host the data
Query Processors (QPs) execute queries in every node

Intelligent Behaviour

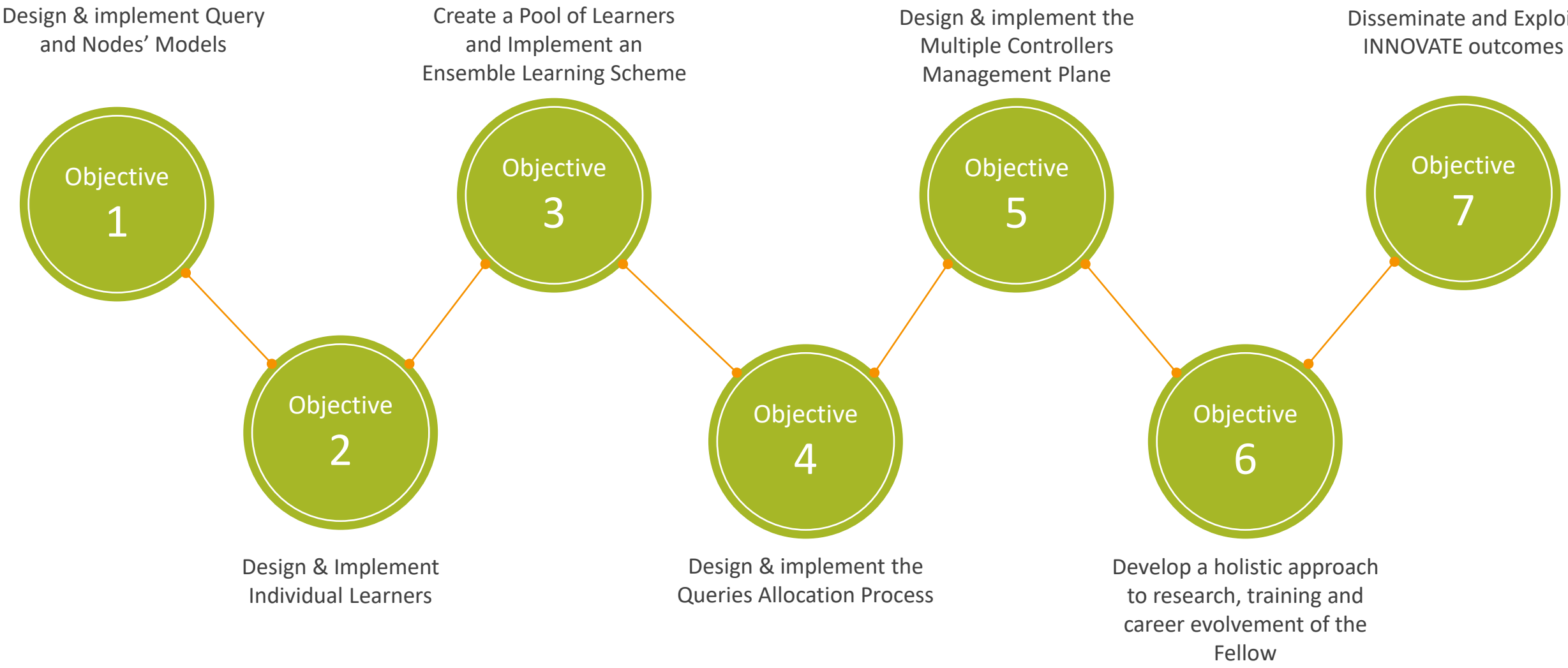
Allocate queries to nodes
Support nodes management
Support data management
Support the behaviour of QCs

INNOVATE Architecture

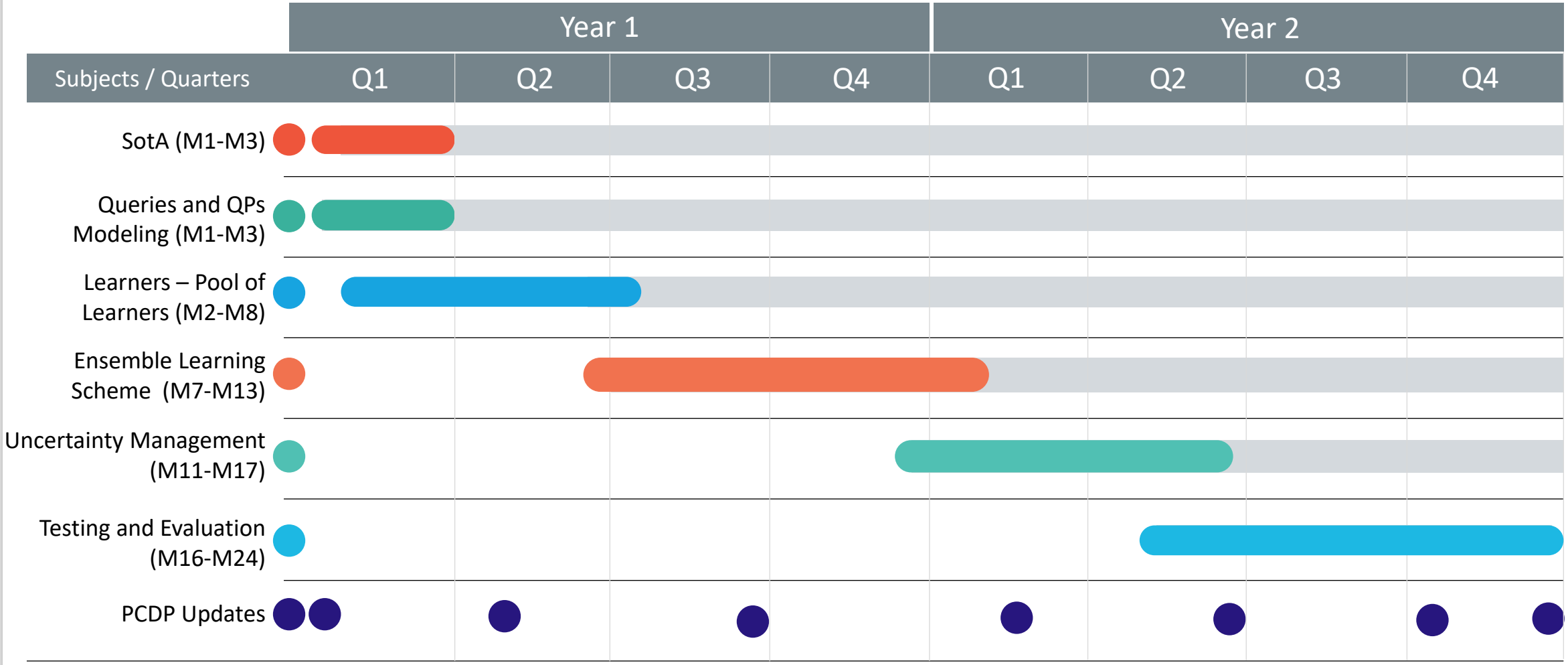


INNOVATE offers intelligent mechanisms for the management of queries, data and distributed nodes

Research Objectives



INNOVATE Timeline



Steps

A HOLISTIC FRAMEWORK

Models, algorithms and methodologies for intelligent query/task management

INTELLIGENCE

Advanced methodologies for the management of the ecosystem

ALLOCATION

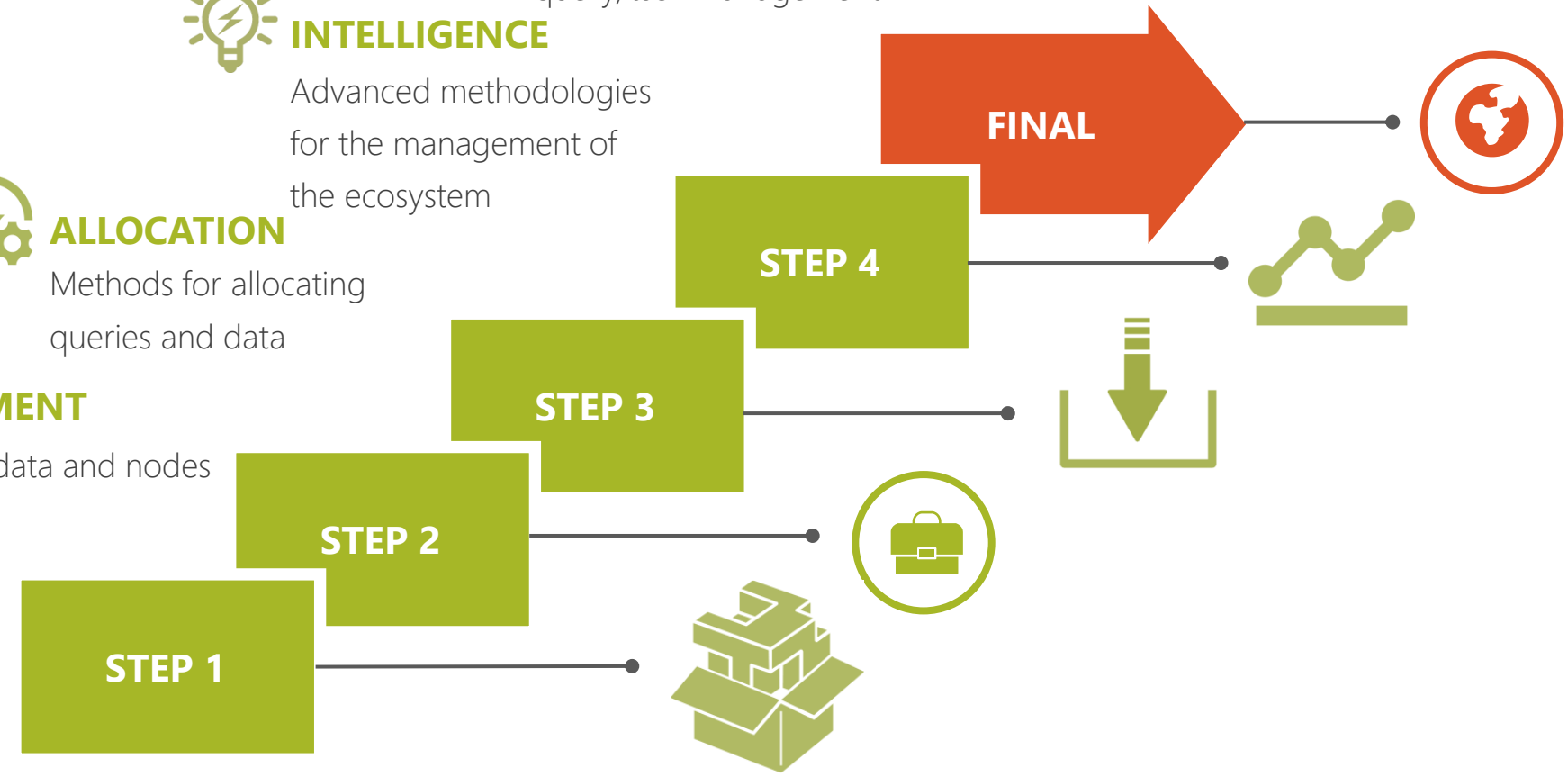
Methods for allocating queries and data

MANAGEMENT

Schemes for data and nodes management

MODELLING

Models for queries and nodes/QPs



Queries and QPs Models

- ✓ We match queries and QPs characteristics
- ✓ Queries
  - ✓ Complexity Class
  - ✓ Deadline
  - ✓ Constraints
- ✓ QPs
  - ✓ Load
  - ✓ Speed of processing
  - ✓ Data present in each node

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)

Queries and QPs Models

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

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

For selecting the appropriate peer, we adopt a ***multi-criteria optimization methodology***
We adopt the VIKOR method

Multi-criteria Query Allocation

- ✓ 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 ***kNN 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*

Data Management

- ✓ 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

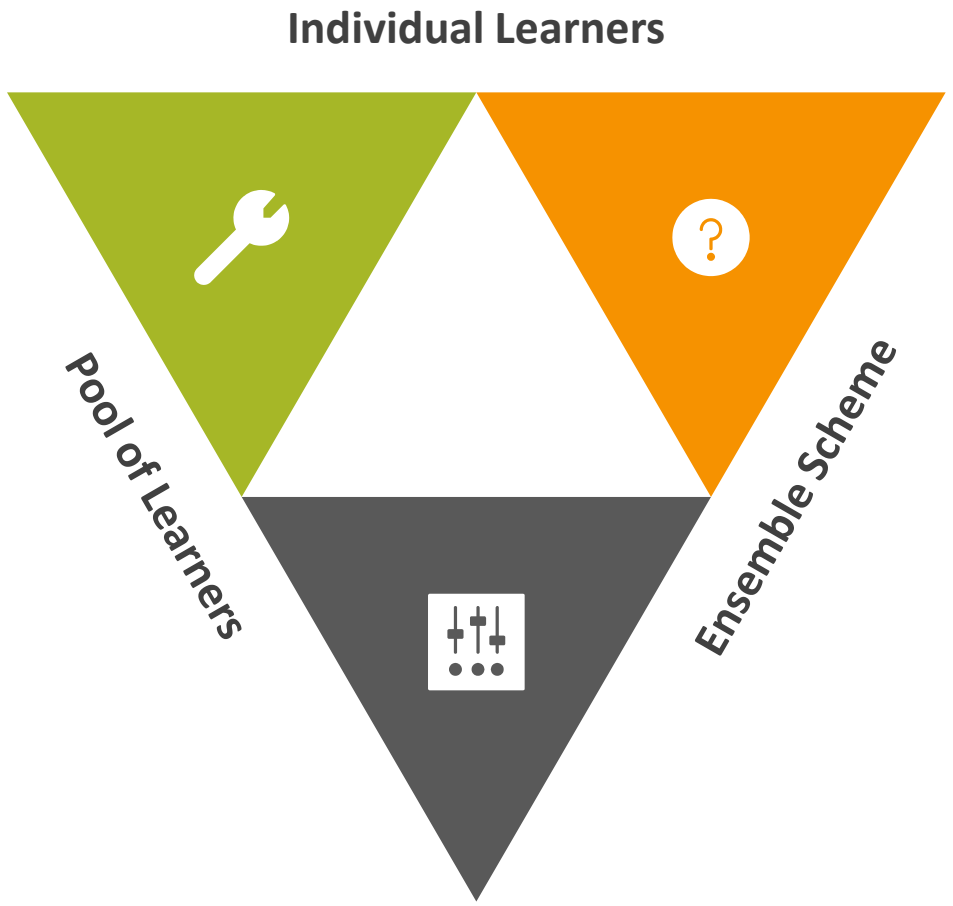
Nodes monitor their internal status (e.g., load, resources)
Nodes monitor the network's performance (e.g.., bandwidth, errors)

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

Ensemble Learning



- ✓ 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***

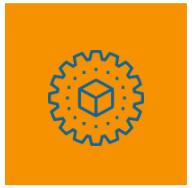
Advanced Models

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

INNOVATE Academic Output

Journal Publications

1. * K. Kolomvatsos, 'A Distributed, Proactive Intelligent Scheme for Securing Quality in Large Scale Data Processing', **Springer Computing**, 2019
2. * K. Kolomvatsos, 'An Efficient Scheme for Applying Updates in Pervasive Computing Applications', **Journal of Parallel and Distributed Computing**, Elsevier, 2019
3. K. Kolomvatsos, C. Anagnostopoulos, 'Multi-criteria Optimal Task Allocation at the Edge', **Elsevier Future Generation Computer Systems**, 2019
4. Kostas Kolomvatsos, Christos Anagnostopoulos, 'An Intelligent Edge-Centric Queries Allocation Scheme based on Ensemble Models', *submitted for review* in **ACM Transactions of Knowledge Discovery from Data**, 2019
5. Kostas Kolomvatsos, Christos Anagnostopoulos, 'A probabilistic Model for Assigning Queries at the Edge', *submitted for review* in **Springer Computing**, 2019
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
7. Kostas Kolomvatsos, Christos Anagnostopoulos, 'Uncertainty-Driven Queries management at the Edge', *in preparation* to be submitted in **Elsevier Fuzzy Sets and Systems**, 2019

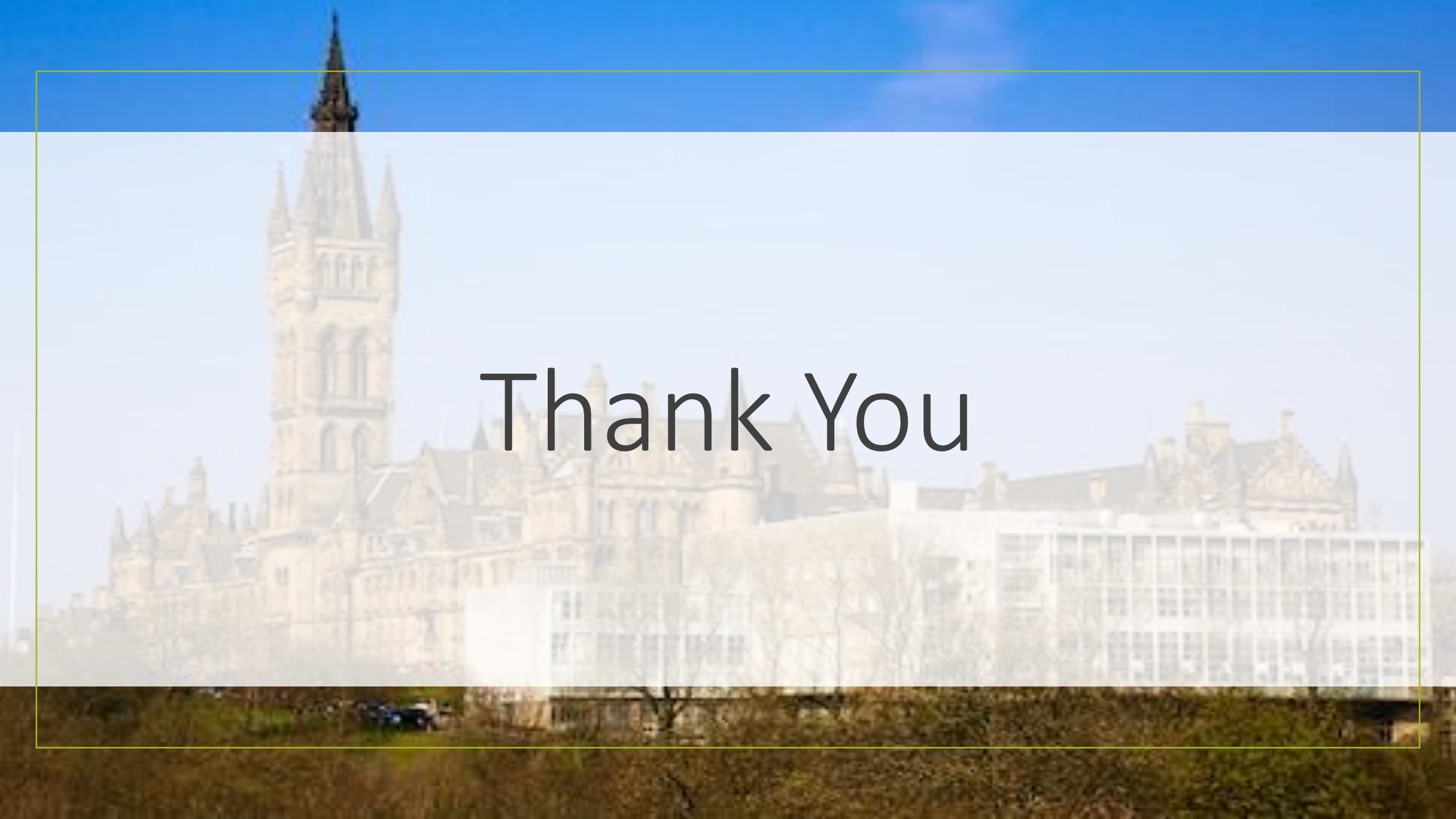
Conferences/Posters/Book Chapters

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
3. E. Aleksandrova, C. Anagnostopoulos, K. Kolomvatsos, 'Machine Learning Model Updates in Edge Computing: An Optimal Stopping Theory Approach', in ***18th IEEE International Symposium on Parallel and Distributed Computing,*** June 5-7, Amsterdam, Netherlands, 2019
4. S. Sagkriotis, K. Kolomvatsos, C. Anagnostopoulos, D. Pezaros, S. Hadjiefthymiades, 'Knowledge-centric Analytics Queries Allocation in Edge Computing Environments', in ***IEEE Symposium on Computers and Communications (ISCC),*** June 29th - July 3rd, Barcelona, Spain, 2019
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
6. Kostas Kolomvatsos, Christos Anagnostopoulos, 'Edge-Centric Queries Stream Management based on an Ensemble Model', *submitted for review* in Springer "Smart Innovation, Systems and Technologies" series volume, 2019

* 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