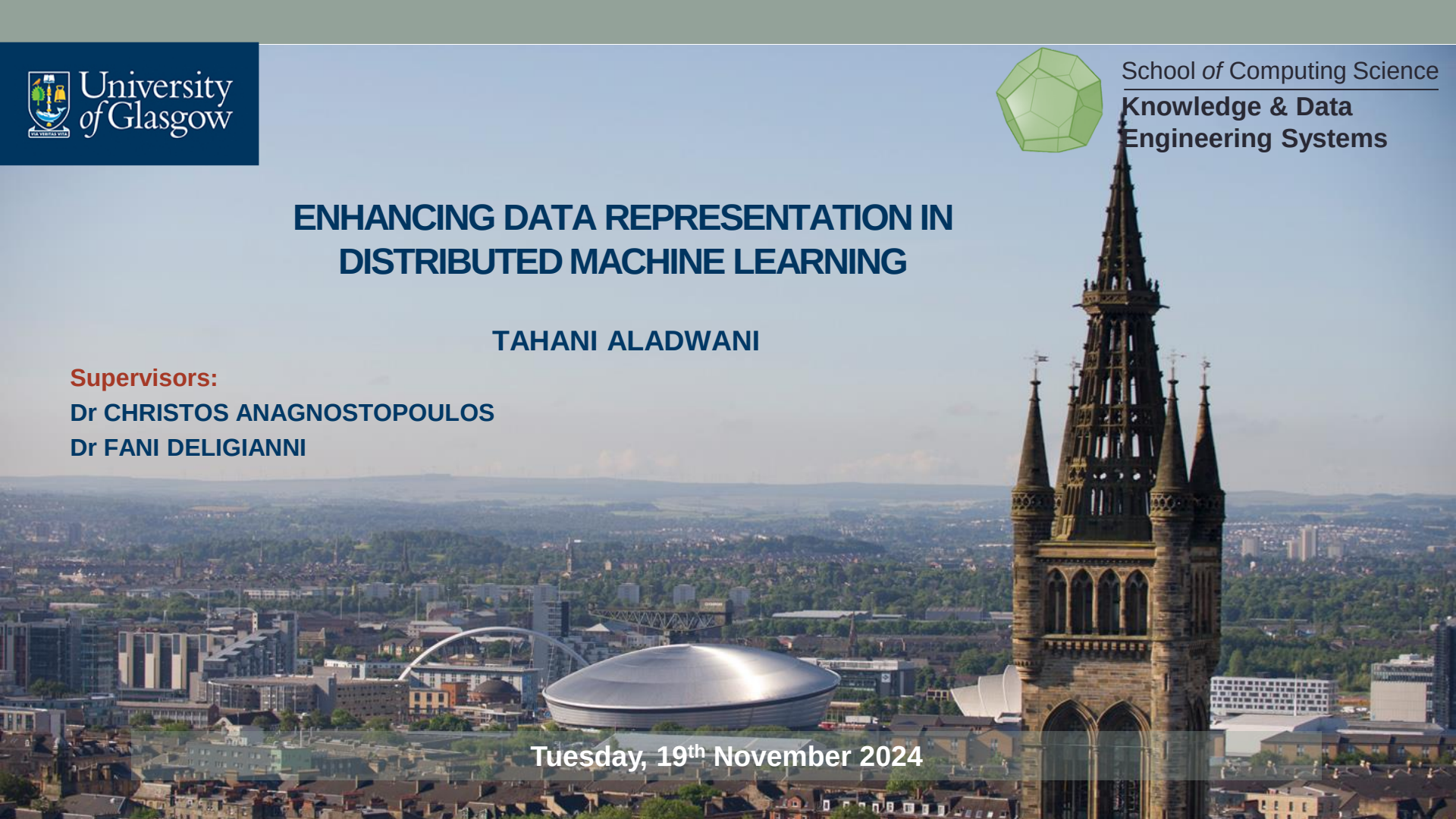


University of Glasgow

School *of* Computing Science

Knowledge & Data Engineering Systems

ENHANCING DATA REPRESENTATION IN DISTRIBUTED MACHINE LEARNING

TAHANI ALADWANI

Supervisors:

Dr CHRISTOS ANAGNOSTOPOULOS

Dr FANI DELIGIANNI

Tuesday, 19th November 2024

University of Glasgow

School of Computing Science
Knowledge & Data Engineering Systems

Distributed Data

Data as valuable source

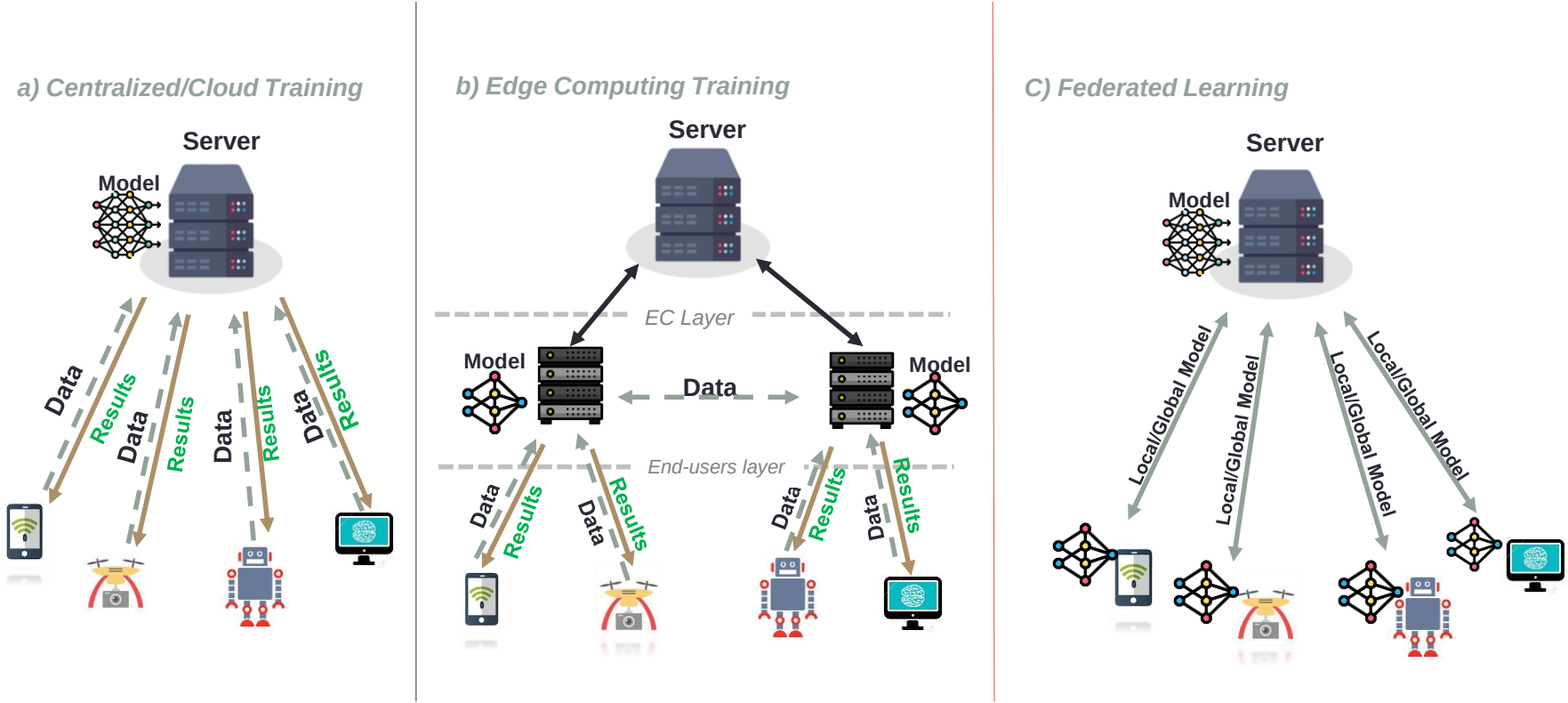
Smart Healthcare Monitoring: Distributed data from wearable devices monitor patient health in real-time and enabling personalized.

Smart Cities: Data from distributed sensors across a city power applications like traffic management, pollution monitoring, and public safety systems, making cities more responsive and efficient.

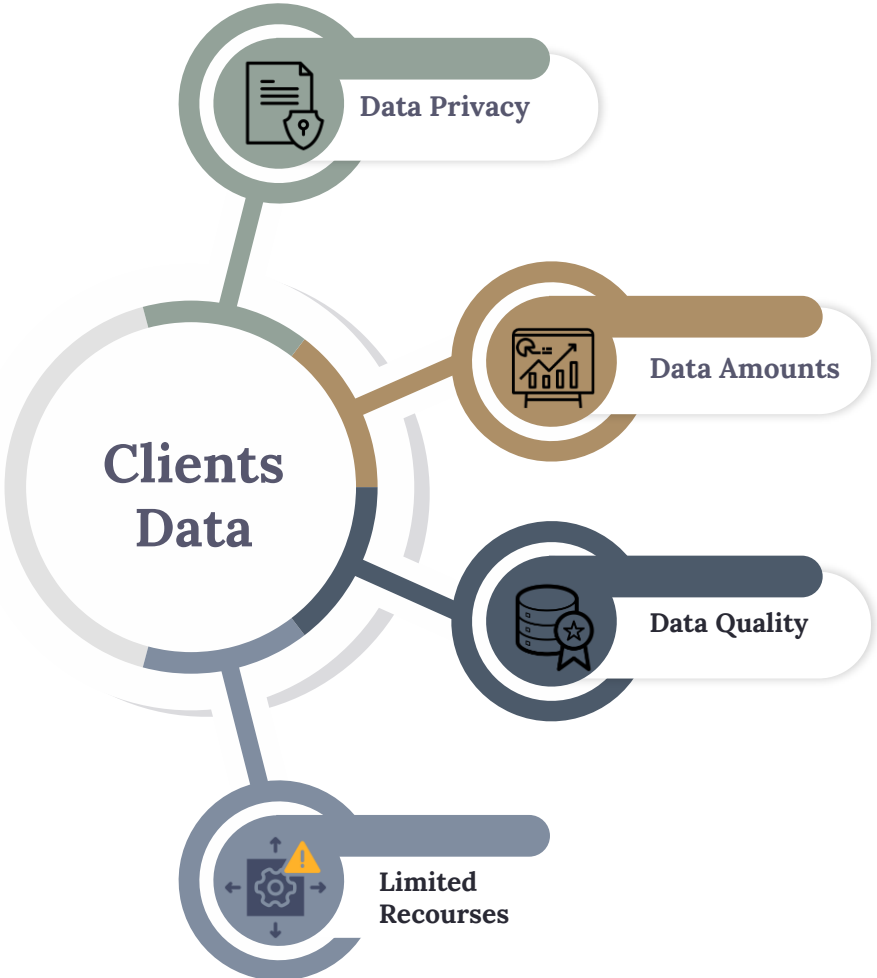
Agricultural Analytics: Data from distributed sensors on farms inform decisions on irrigation, pest control, and crop health.

Public Safety and Surveillance: Data from distributed cameras and sensors can enhance public safety by enabling real-time threat detection and response, useful in areas like surveillance and event security.

The effectiveness of DL models depends on **access** to large, varied datasets

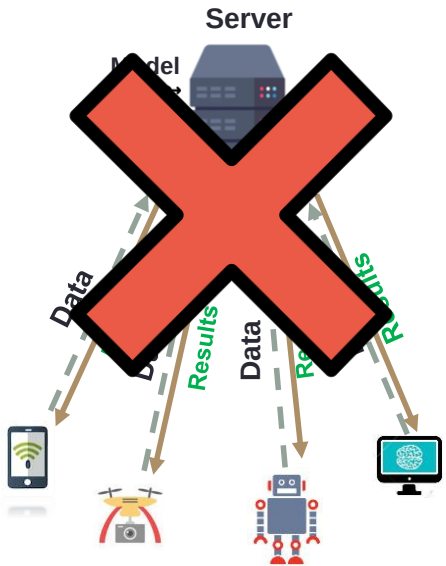


Distributed Data Challenges

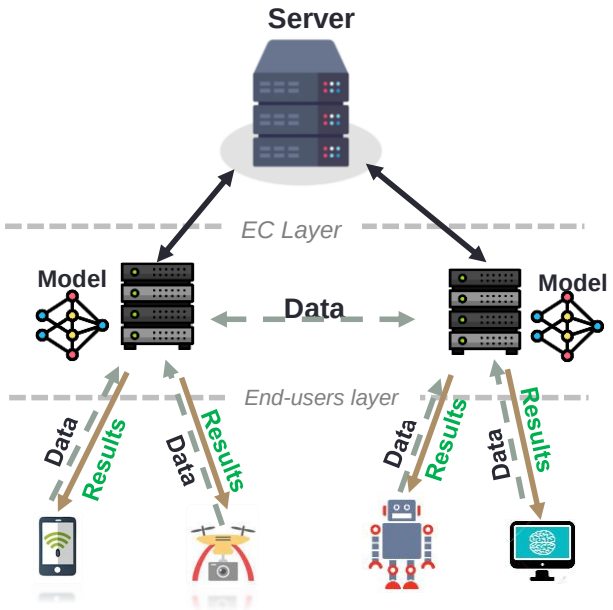


The effectiveness of DL models depends on access to large, varied datasets

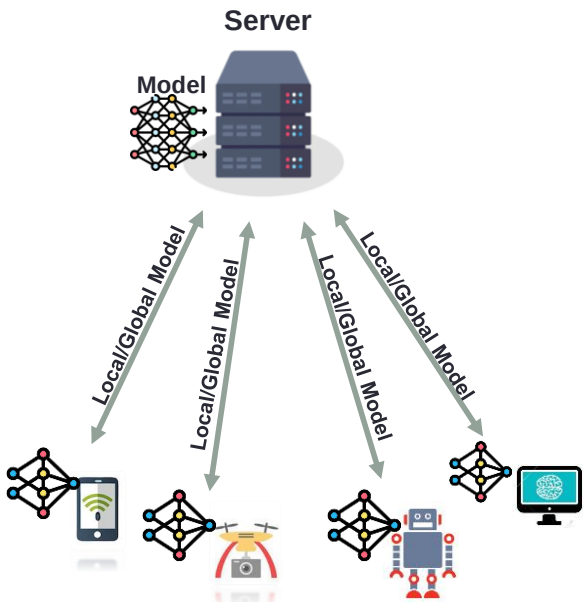
a) Centralized Training



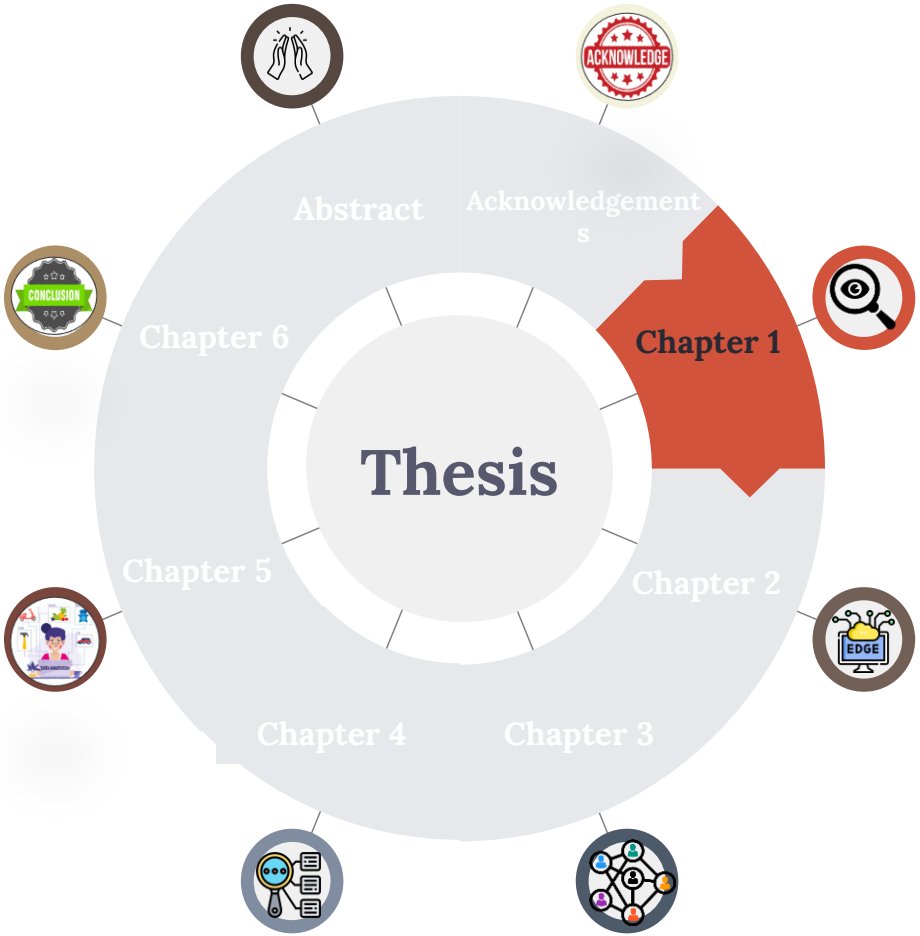
b) Edge computing Training



C) Federated Learning



Thesis Structure



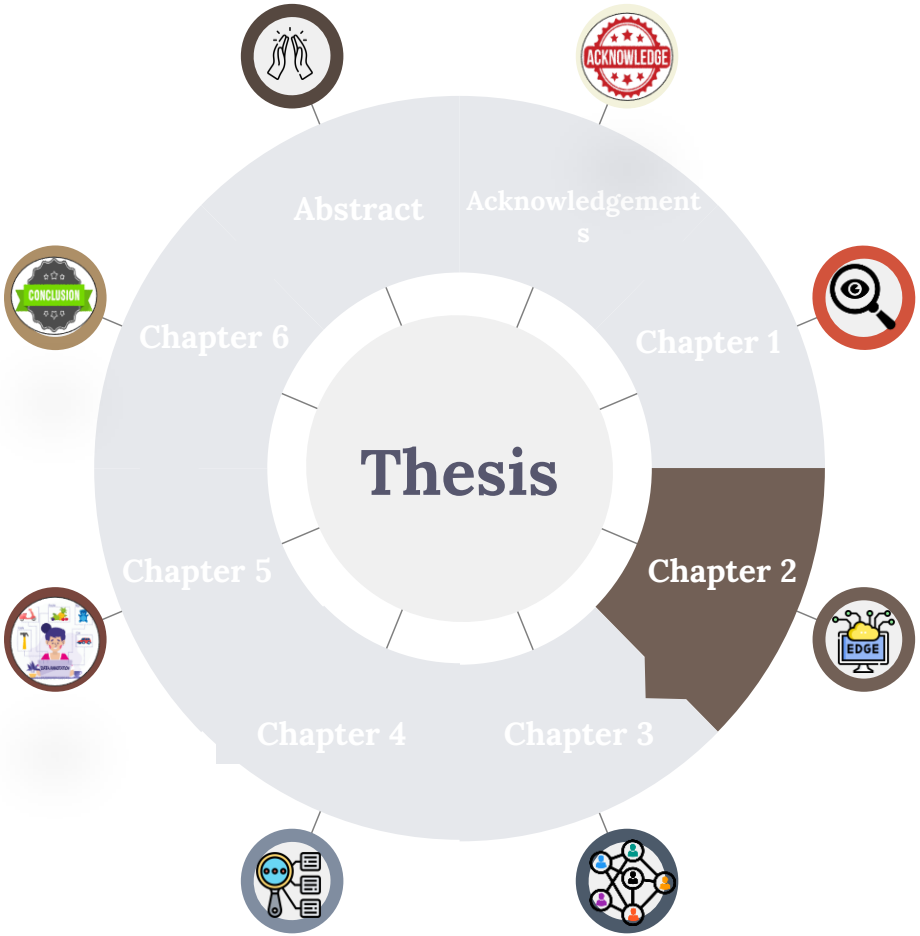
Chapter 1: Overview

Research Goals & Contributions

Research Questions & Solutions

Academic Contribution to the Research Community

Background



Chapter 2: Edge Computing, Data & Tasks Offloading & Caching

Problem

EC servers offer numerous advantages (reduced communication delays, minimized energy consumption, enhanced network stability, improved security, and real-time application support

- **Their resources are still limited**, especially with the rapidly increasing number of requests from end-users.

Challenge: in **selecting efficiently** EC servers for each analytic task.

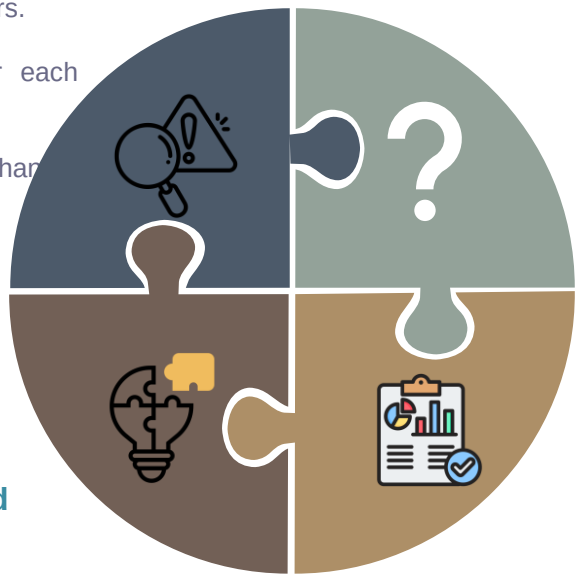
Executing analytic tasks on EC servers requires more than offloading computations;

- involves minimizing repeated tasks and data transfer.
- supporting task offloading & caching popular content for future use is essential.

Solution

A robust approach to optimizing the decision-making process for offloading and caching analytic tasks by utilizing a proposed EC server selection mechanism.

A mechanism based on the Fuzzy Logic Inference (FLI).



Research Question

- **How can we optimize EC server selection to minimize data offloading and reduce redundant task execution?**

Results

- Maximize Offloading Probability to Optimal EC server, considering factors like task popularity and data availability.
- Minimize Data Offloading Rate.
- Increase EC Server Resource Utilization.
- Maximize the chance of selecting the right EC Server for each task: based on the Fuzzy Logic Inference has to enhance decision-making in server selection.

Takeaway…

School *of* Computing Science
Knowledge & Data Engineering Systems

This mechanism significantly reduces data offloading rates, minimizes repeated task offloading, and enhances data privacy.

In many cases, it reduces offloaded data to 5%–15%.

This opens the door to a new research question:

Can we train models in zero-data offloading scenarios where data privacy is strictly regulated (e.g., financial data, or personal information) where sharing data with third parties is prohibited?

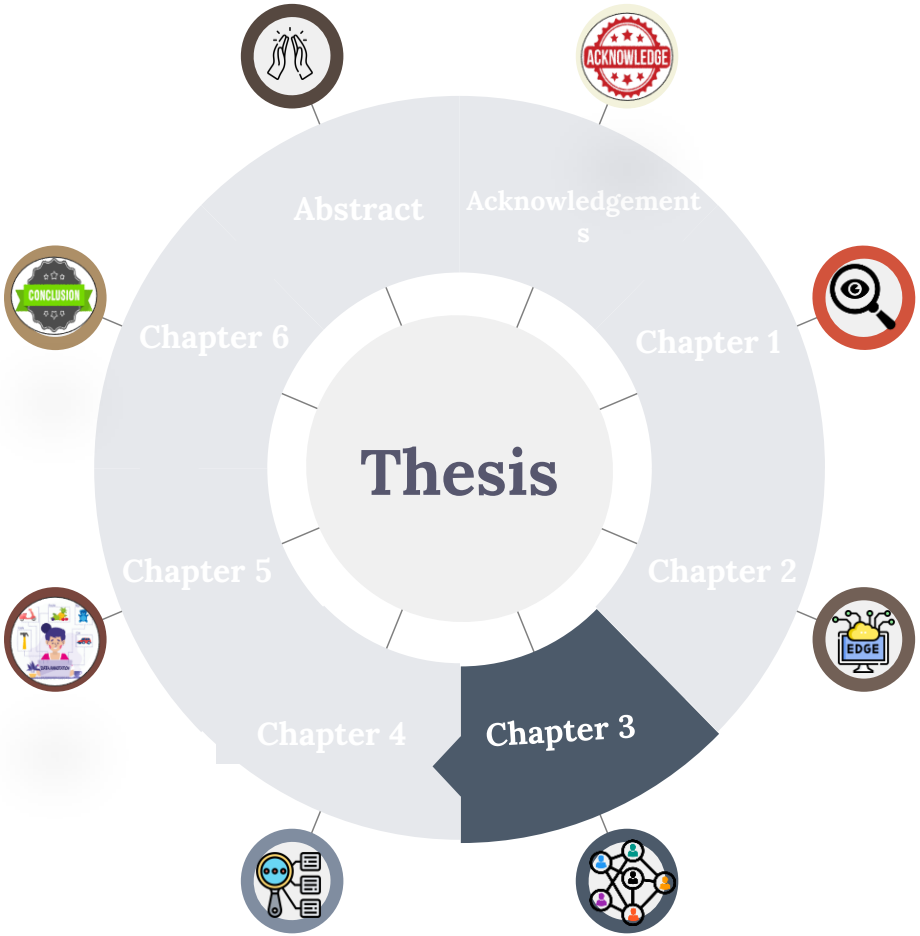
Let's explore the answer to this question in the following chapters.…

Publications for Chapter 2

- ✓ **Tahani Aladwani**, Christos Anagnostopoulos, Konstantinos Kolomvatsos, Ibrahim Alghamdi, "Data-driven analytics task management reasoning mechanism in edge computing." Smart Cities 5.2 (2022): 562-582.

- ✓ **Tahani Aladwani**, Christos Anagnostopoulos, Konstantinos Kolomvatsos, Ibrahim Alghamdi, "Data-Driven Analytics Task Management at the Edge: A Fuzzy Reasoning Approach." 2022 9th International Conference on Future Internet of Things and Cloud (FiCloud). IEEE, 2022.

School *of* Computing Science
Knowledge & Data Engineering Systems



Chapter 3: Node & Relevant Data Selection in Distributed Predictive Analytics: A Query-centric Approach

Problem

Distributed Predictive Analytics (DPA): constructing predictive models based on data distributed across nodes.

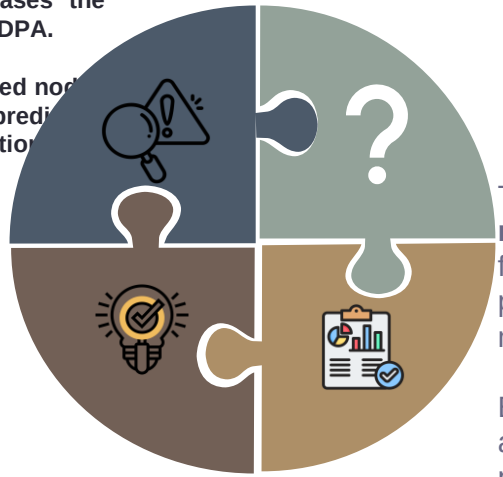
In DPA data collected by nodes are inherently different; each node can have different distributions, volumes, and features space.

This heterogeneity hinders the development of accurate models in a distributed fashion. Many state-of-the-art methods adopt random node selection as a straightforward approach. Such method increases the likelihood of selecting nodes with low-quality or irrelevant data for DPA.

Consequently, it is only after training models over randomly selected nod that the most suitable ones can be identified based on the predi performance. This results in more time and resource consumptio increased network load.

Solution

- ✓ **Query-Centric Node Selection:** Nodes are selected based on the relevance of their data to specific query requirements, rather than random selection. This minimizes irrelevant data use and enhances model performance.
- ✓ **Data Relevance Factors (F1)** Overlapping area between cluster boundaries and query boundaries. **(F2)** Number of samples within each cluster. **(F3)** Number of relevant samples within each cluster.
- ✓ **Ranking Mechanism:** Nodes are ranked based on relevance to ensure that only the most suitable nodes participate in model training. This minimizes resource use and prevents unnecessary data access.



Research Question

- **How can we efficiently select only the relevant data from each chosen node per analytics query when access to data is limited?**

Results

The **query-centric node and data selection mechanism** enhances model accuracy by focusing on data relevant to each query, preventing irrelevant data from degrading model performance.

Effective node selection without full data access, preserving privacy while maintaining prediction accuracy comparable to centralized models.

Communication overhead is reduced by engaging only the most suitable nodes, thus minimizing unnecessary communication rounds.

Takeaway…

This chapter demonstrates that a **query-centric approach to node and data selection in distributed predictive analytics** improves model accuracy, reduces communication overhead, and preserves data privacy. By proactively engaging only relevant nodes and data, the method offers an efficient, scalable solution tailored to the specific requirements of each query.

This mechanism is more effective with low-dimensional data, such as Tabular data.

This opens the door to a new research question:

Can we enhance the data selection mechanism to work with high-dimensional data such as images and videos?

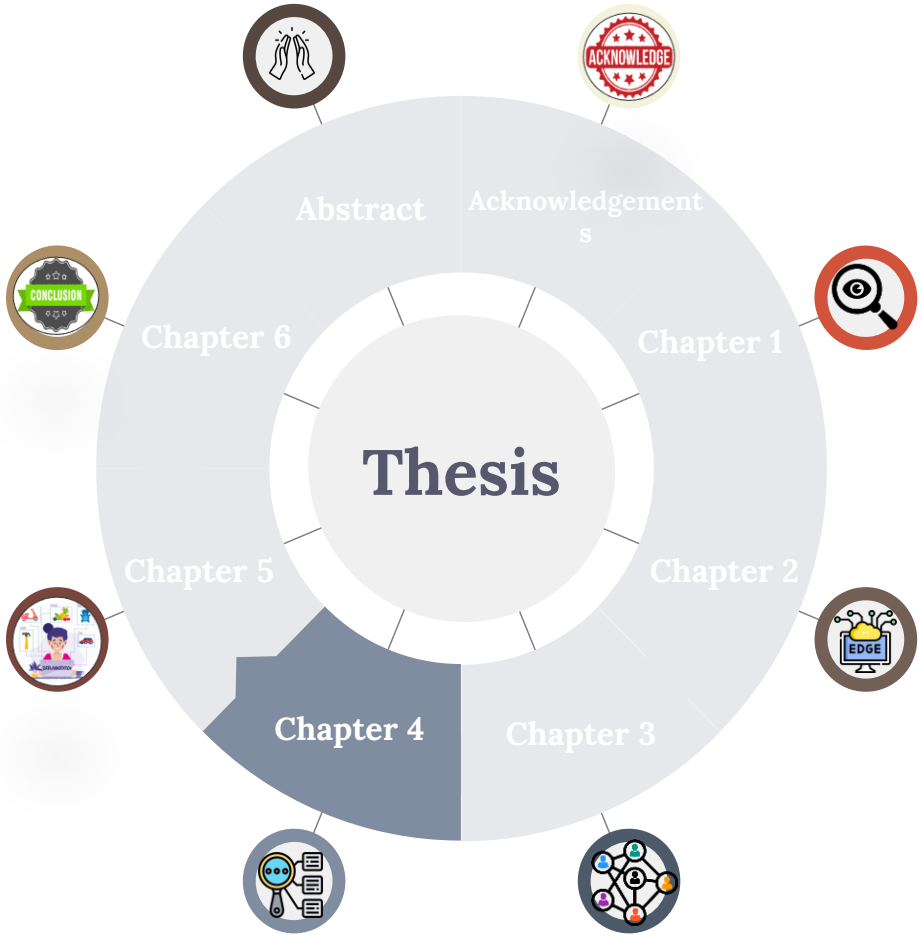
Let's explore the answer to this question in the following chapters….

Publications for Chapter 3

- ✓ **Tahani Aladwani**, Christos Anagnostopoulos, Konstantinos Kolomvatsos, "Node and Relevant Data Selection in Distributed Predictive Analytics: A Query-centric Approach." Journal of Network and Computer Applications, Elsevier, 2024: 104029. (**Q1, (7.7 IF), CORE A**).

- ✓ **Tahani Aladwani**, Christos Anagnostopoulos,Konstantinos Kolomvatsos, Ibrahim Alghamdi, "Query-driven Edge Node Selection in Distributed Learning Environments." IEEE 39th International Conference on Data Engineering Workshops (ICDEW), IEEE, 2023. **(CORE A*).**

School *of* Computing Science
Knowledge & Data Engineering Systems



Chapter 4: Cluster-based & Label-aware Federated Meta-Learning for On-Demand Classification Tasks

Problem

In many applications *where quick decisions are required* the predictions have to be performed in near real-time*.*

Meta-learning *has proved to accelerate model adaptation to arbitrary labels by allowing fine-tuning over small datasets.*

The adoption of meta-learning in FL is not easy due to data and class labels heterogeneity in FL . This issue arises due to various types of distribution shifts among clients including feature, label, and concept distribution shifts.

A single, global FML model proves to be inefficient and impractical accommodate:

(i) any arbitrary classification tasks.

(ii) out-of-distribution labels across clients.

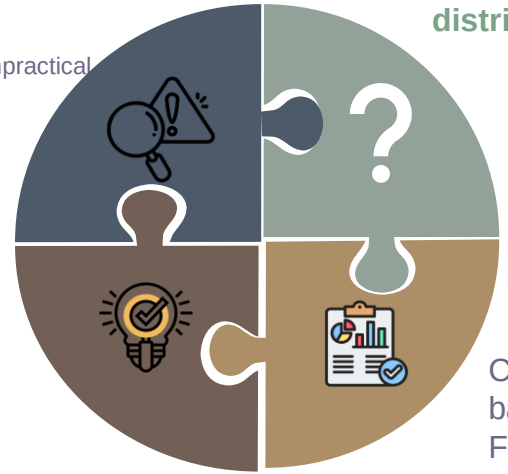
Solution

We introduce **Cluster-based & Label-aware FML** framework (**CL-FML**) that addresses such challenges.

Idea: CL-FML gathers clients together based on label shifting mitigating label imbalance per task.

Main goals:

- ✓ Study the cases of training more than one (reusable) meta-model tailored to available labels $\mathcal{L}_k \subset \mathcal{L}$ of a cluster of clients $\mathcal{C}_k$.
- ✓ Provide compact sized meta-models stored on clients temporarily, to be reused for future tasks.
- ✓ CL-FML copes with **sharing** meta-models among clusters to further fine-tune in case of out of distribution tasks.



Research Question

- **How can we effectively mitigate label shifting and address missing label challenges in FL, while ensuring that the meta-model generalizes to different data distributions?**

Results

Comprehensive experiments against baselines showcase the superiority of CL-FML in terms of final model accuracy, loss, training rounds, and the amount of required data per task.

Takeaway…

The results highlight its effectiveness in achieving high classification accuracy, particularly in settings characterized by label distribution shifts and client data heterogeneity.

CL-FML consistently outperforms traditional federated learning methods, showcasing the importance of its **label-aware clustering and data augmentation strategies** in handling the complexities of distributed machine learning.

This mechanism assumes that all client data is labeled.

This opens the door to a new research question:

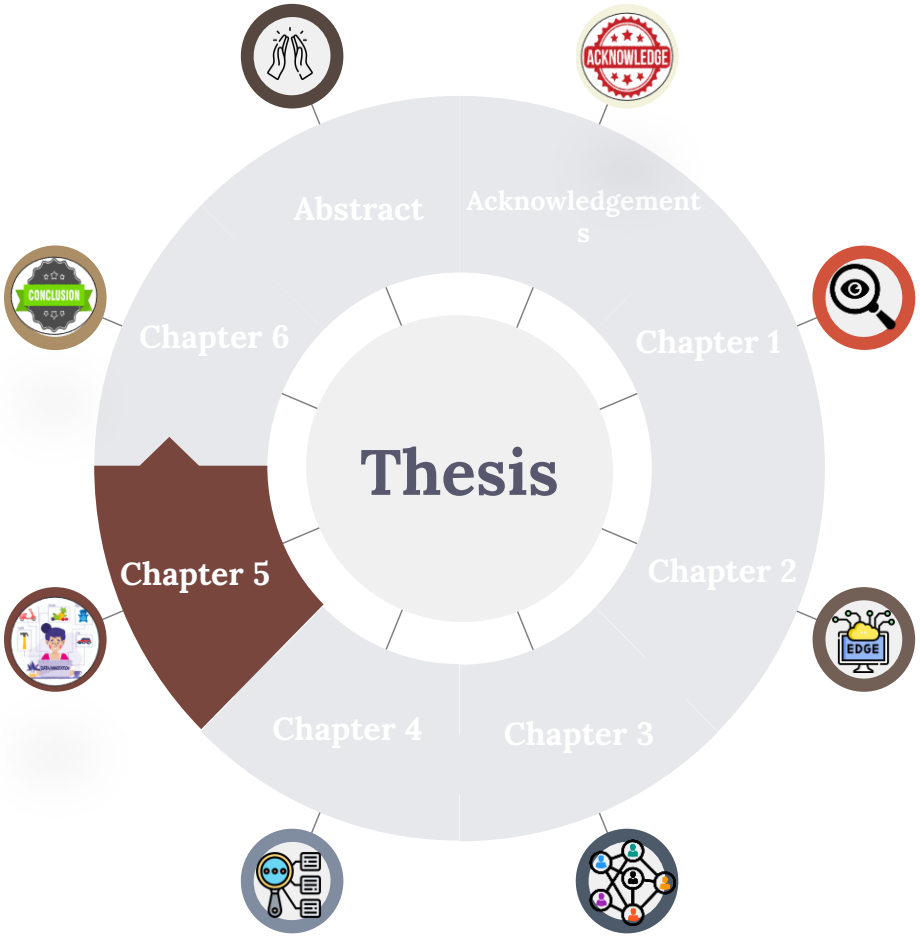
Can we train models with clients who have partially labeled or unlabeled data? For example, clients might lack the motivation to label their data due to costs, time constraints, or limited expertise.

Let’s explore the answer to this question in the following chapters….

Publications for Chapter 4

- ✓ **Tahani Aladwani**, Christos Anagnostopoulos, Shameem P. Parambath, Fani Deligianni, "CL-FML: Cluster-based & Label-aware Federated Meta-Learning for On-Demand Classification Tasks." 11th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2024), IEEE, 2024. **(CORE A).**

School *of* Computing Science
Knowledge & Data Engineering Systems



Chapter 5: The Price of Labelling: A Two-Phase Federated Self-Learning Approach

Problem

Distributed data in real-world scenarios:

- The distributed data in real-world scenarios can be non-IID, leading to common issues such as class imbalance & distribution shift across clients.
- Existence of un-labeled data across clients, due to various factors like limited resources, labeling costs, and human errors.

Challenge: create high-quality pseudo-labels without addressing these issues.

Research Question

- **What is the price of learning a global model using scarce and skewed distributed labelled data, while capitalizing on partially labelled and fully unlabelled data across clients?**

Solution

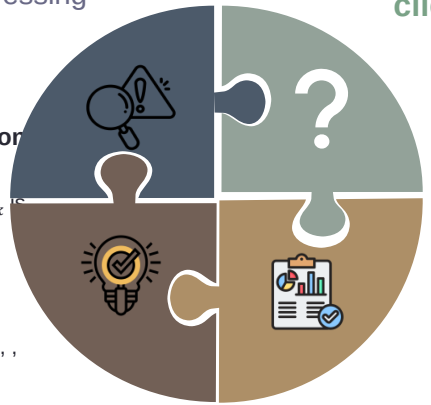
Consider a set of distributed are categorized into three types based on their data:

– Type I clients (**labelled clients**) $n_i \in \mathcal{N}^L \subset \mathcal{N}$, denoted as $\mathcal{D}_i^L = \{(x_k, y_k)\}_{k=1}^{\mathcal{D}_i^L}$, y_k is the label.

– Type II clients (**partially labelled clients**) $n_i \in \mathcal{N}^P \subset \mathcal{N}$ have **labelled and unlabelled** samples, i.e., $\mathcal{D}_i^P = \{(x_k, y_k \vee \perp)\}_{k=1}^{\mathcal{D}_i^P}$, $\perp$.

– Type III clients (**unlabelled clients**) $n_i \in \mathcal{N}^L \subset \mathcal{N}$ have **all samples unlabelled**, , i.e., $\mathcal{D}_i^L = \{(x_k, \perp)\}_{k=1}^{\mathcal{D}_i^U}$.

Focus: labelled samples are much fewer than unlabelled ones, i.e., $|\mathcal{D}^L| \ll |\mathcal{D}^U|$



Results

The 2PFL framework addresses the challenge of training FL models across different types of clients with limited and skewed labeled and unlabelled data.

By leveraging data augmentation, 2PFL leads to improved model performance and accelerates convergence by progressive pseudo-labelling

2PFL exploits labelled, partially labelled and unlabelled data across all types of clients $(\mathcal{N}^L \cup \mathcal{N}^P \cup \mathcal{N}^U)_{n_i \in \mathcal{N}}$ to minimize the loss function $f^L(\theta_G)$, $f^P(\theta_G)$, and $f^U(\theta_G)$ respectively:

$$\min_{\theta_G} f(\theta_G) = \frac{1}{\mathcal{N}^L} \sum_{\ell=1}^{\mathcal{N}^L} \mathcal{L}^L(x_\ell^L, y_\ell^L, \theta_G) + \frac{1}{\mathcal{N}^P} \sum_{\ell=1}^{\mathcal{N}^P} \mathcal{L}^P(x_\ell^P, y_\ell^P, \theta_G) + \frac{1}{\mathcal{N}^U} \sum_{\ell=1}^{\mathcal{N}^U} \mathcal{L}^U(x_\ell^U, y_\ell^U, \theta_G)$$

Takeaway…

Our comprehensive experiments and comparison with state-of-the-art methods highlight that 2PFL consistently outperforms baselines across various performance metrics and datasets. 2PFL exhibits superior convergence speed, accuracy, and data pseudo-labelling rate acquired in each phase. Therefore,

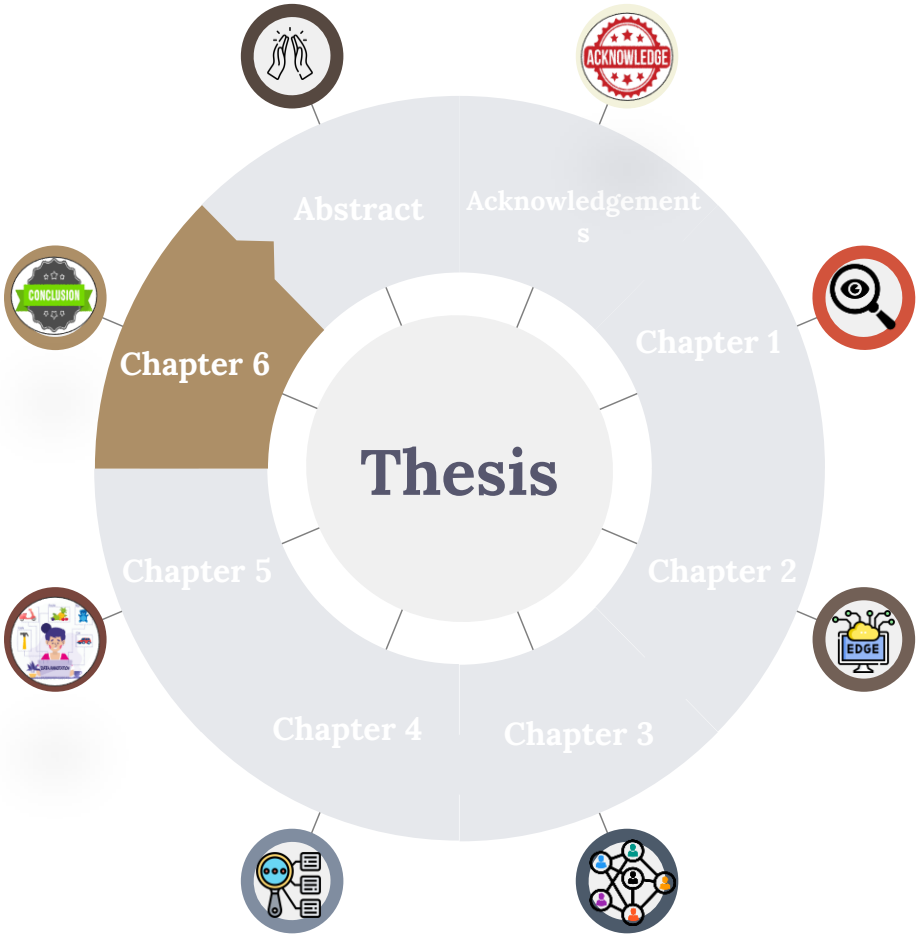
The price for learning a global model with skewed and unlabeled data is minimal with 2PFL'.

Closing the door of my PhD Research.

Opening the door to many questions that I carry with me into my journey as a Research Assistant & Associate

Publications for Chapter 5

- ✓ **Tahani Aladwani**, Christos Anagnostopoulos, Shameem P. Parambath, Fani Deligianni, "The Price of Labelling: A Two-Phase Federated Self-Learning Approach." European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML-PKDD 2024. (**CORE A**).



Chapter 6: Conclusion

Conclusion

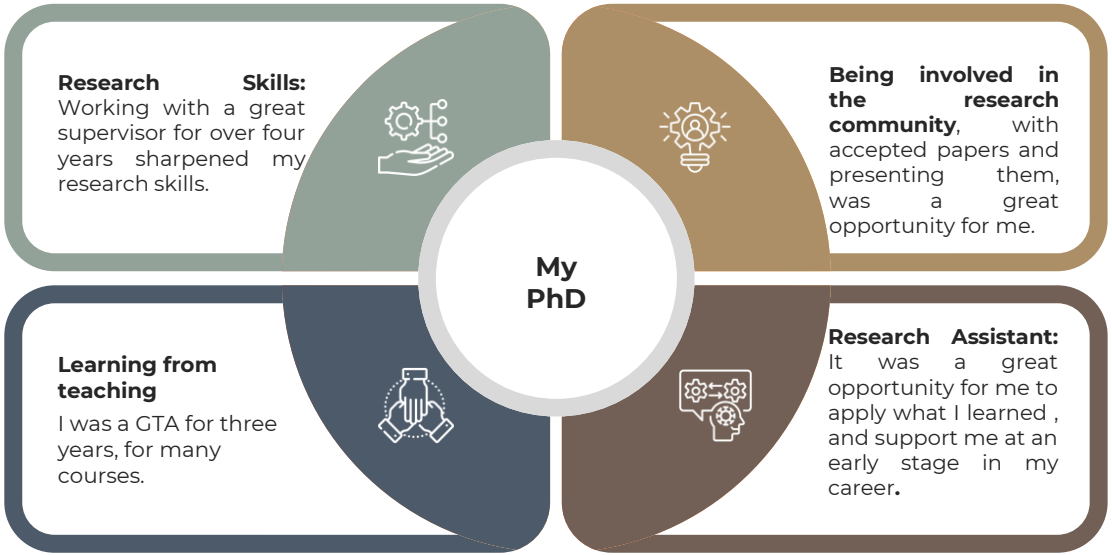
Evolving Data Challenges: The thesis addresses the growing diversity in data, which now varies widely across clients in volume, format, and labeling, making traditional centralized methods less effective.

Decentralized Solutions for Efficiency and Privacy: Adaptable edge computing and federated learning frameworks are proposed to manage data heterogeneity while meeting efficiency, privacy, and accuracy needs.

Enhanced Task Relevance and Reduced Data Transmission: The solutions emphasize reducing data transmission and improving task relevance, which are essential in handling diverse data sources in a distributed setting.

Resilience to Data Scarcity and Skewness: The adaptive methodologies introduced support effective, privacy-conscious model training, especially in scenarios with limited labeled data, aligning with modern data demands.

What I earned during my PhD




University of Glasgow
THE End
School of Computing Science
Knowledge & Data Engineering Systems