



UNIVERSITY OF
CAMBRIDGE

Real-world AI-based Systems

Christian Cabrera Jojoa

Research Associate
chc79@cam.ac.uk

Department of Computer Science and Technology

Agenda

PART I: The great AI question.

PART II: Why the question?

PART III: Possible answers...



PART I: The great AI question.



AI in the news...



Can football-playing robots beat the World Cup winners by 2050?

By Benard Delmonaco Jr.
Business reporter

27 September 2021



In RoboCup, the humans leave the pitch before the football game.

For football fans around the globe the pinnacle of the final. But what if that match was ultimately just the prelude between the best humans and the best robots?

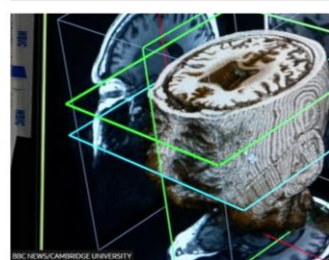
Source: <https://www.bbc.co.uk/news/technology-58158820>



Artificial Intelligence may diagnose dementia in a day

By Paulah Ghosh
Science correspondent

10 August 2021



AI can identify patterns in brain scans even the most expert neurologists cannot see.

Source: <https://www.bbc.co.uk/news/health-57912340>



Would you let a robot lawyer defend you?

By Paulah Ghosh
Business reporter

15 August 2021



AI is increasingly being used by the legal industry.

Could your next lawyer be a robot? It sounds far fetched, but artificial intelligence (AI) software systems - computer programs that can update and "think" by themselves - are increasingly being used by the legal community.

Source: <https://www.bbc.co.uk/news/business-58158820>



UNIVERSITY OF
CAMBRIDGE

AI in the news...



Can football-playing robots beat the World Cup winners by 2050?

By Benard Delmonaco Jr.
Business reporter

27 September 2021



In RoboCup, the humans leave the pitch before the football game.

For football fans around the globe the pinnacle of the final. But what if that match was ultimately just the prelude between the best humans and the best robots?

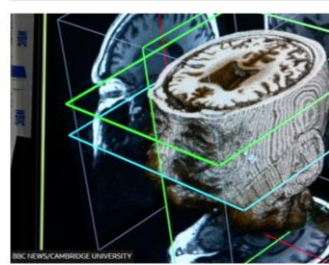
Source: <https://www.bbc.co.uk/news/technology-58158820>



Artificial Intelligence may diagnose dementia in a day

By Paddy Gleeson
Science correspondent

10 August 2021



AI can identify patterns in brain scans even the most expert neurologists cannot see.

Source: <https://www.bbc.co.uk/news/health-57912345>



Would you let you?

By Paddy Gleeson
Business reporter

15 August 2021



AI is increasingly being used by the legal industry.

Could your next lawyer be a robot? It sounds far fetched, but artificial intelligence (AI) software systems - computer programs that can update and "think" by themselves - are increasingly being used by the legal community.

Source: <https://www.bbc.co.uk/news/business-58158820>



Former Go champion beaten by DeepMind retires after declaring AI invincible

'Even if I become the number one, there is no one else who can beat me.'

By James Vincent | Nov 27, 2019, 8:42am EST



Lee Se-dol is seen in 2016 during his matches with the AI program AlphaGo.

Source: <https://www.theverge.com/2019/11/27/google-fires-engineer-who-claims-ai-chatbot-is-sentient>

Google fires software engineer who claims AI chatbot is sentient

Company said Blake Lemoine violated Google policies and that his claims were 'wholly unfounded'



Google say LaMDA is simply a complex algorithm designed to generate convincing human language. Photograph: Andrew Kelly/Reuters

Google has dismissed a senior software engineer who claimed the company's artificial intelligence chatbot LaMDA was a self-aware person.

Google, which placed software engineer **Blake Lemoine** on leave last month, said he had violated company policies and that it found his claims on LaMDA to be 'wholly unfounded'.
Source: <https://www.theguardian.com/technology/2022/jul/23/google-fires-engineer-who-claims-ai-chatbot-is-sentient>



UNIVERSITY OF
CAMBRIDGE

AI in the news...



Can football-playing robots beat the World Cup winners by 2050?

By Benard Delmonaco Jr.
Business reporter
27 September 2021



In RobotCup, the human

For football fans aware of the final. But what if that between the best human

Source: <https://www.bbc.com/news/technology-58158820>



Health

Artificial Intelligence may diagnose dementia in a day

By Pailish Ghosh
Science correspondent
10 August 2021



AI in the news...



Can football-playing robots beat the World Cup winners by 2050?

By Benard Delmonaco Jr.
Business reporter
27 September 2021



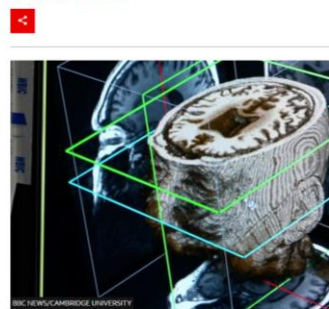
For football fans around the globe the pinnacle of the final. But what if that match was ultimately just the prelude to the best humans and the best robots?

Source: <https://www.bbc.co.uk/news/technology-58158820>



Artificial Intelligence may diagnose dementia in a day

By Paul H. Glendon
Science correspondent
10 August 2021



AI can identify patterns in brain scans even the most expert neurologists cannot see.

Source: <https://www.bbc.co.uk/news/health-57912345>



Would you let you?

By Paul H. Glendon
Business reporter
15 August 2021



AI is increasingly being used by the legal industry.

Could your next lawyer be a robot? It sounds far fetched, but artificial intelligence (AI) software systems - computer programs that can update and "think" by themselves - are increasingly being used by the legal community.

Source: <https://www.bbc.co.uk/news/business-58158820>



Former Go champion beaten by DeepMind retires after declaring AI invincible

'Even if I become the number one, there is no point in playing Go anymore'
By James Vincent
Nov 27, 2019, 8:42am EST



Lee Se-dol is seen in 2016 during his matches with the AI program AlphaGo.

Source: <https://www.theverge.com/2019/11/27/21999999/lee-se-dol-retires>

Google fires software engineer who claims AI chatbot is sentient

Company said Blake Lemoine violated Google policies and that his claims were 'wholly unfounded'



Google say LaMDA is simple language. Photograph: Andre Brouillet

Google has dismissed a software engineer who claimed an artificial intelligence chatbot was sentient.

Google, which placed the chatbot in the hands of a public beta tester, said he had violated company policies.

Source: <https://www.theguardian.com/technology/2021/jan/27/google-fires-engineer-who-claims-ai-chatbot-is-sentient>



La inteligencia artificial lo cambiará todo

Escuche el análisis de Juan Carlos Echeverry sobre este tema que se está tomando el mundo.



La inteligencia artificial lo cambiará todo

Source: https://caracol.com.co/programa/2019/03/1/1/6am_hoy_por_hoy/1552309220_833279.html

Patentan botón para apagar la inteligencia artificial en caso de rebelión

¿Qué sucedería si las computadoras y los robots se rebelaran contra los seres humanos? Google creó una posible solución a las hipótesis de la ciencia ficción.



BBC/GETTY IMAGES

Source: https://caracol.com.co/radio/2016/06/15/tecnologia/1466026084_693727.html

Sounds promising!!!



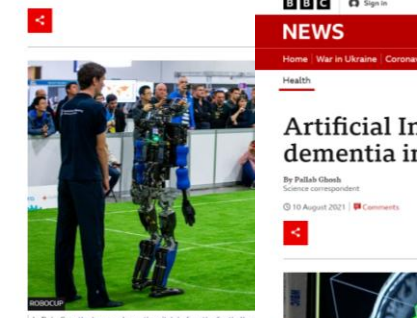
UNIVERSITY OF
CAMBRIDGE

But...



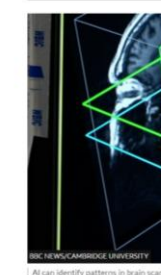
Can football-playing robots beat World Cup winners by 2050?

By Benard Delmonaco Jr.
Business reporter
27 September 2021



In RoboCup, the humans leave the pitch before the football game. For football fans around the globe the pinnacle of the final. But what if that match was ultimately just the prelude between the best humans and the best robots?

Source: <https://www.bbc.co.uk/news/technology-58158820>



Source: <https://www.bbc.co.uk/news/technology-58158820>

Elon Musk just now realizing that self-driving cars are a 'hard problem'

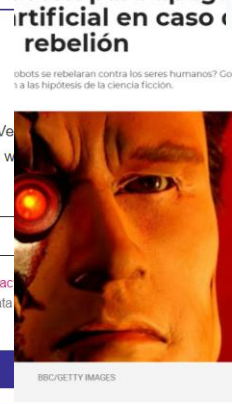
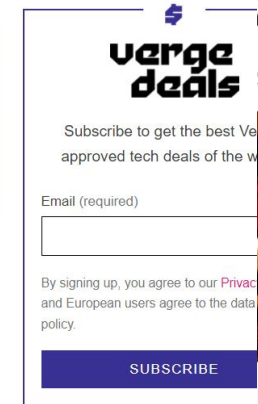
Overpromising and underdelivering
By Andrew J. Hawkins | @andyjayhawk | Jul 5, 2021, 10:53am EDT



Tesla CEO Elon Musk is finally admitting that he underestimated how difficult it is to develop a safe and reliable self-driving car. To which the entire engineering community rose up as one to say, "No duh."

Could your next lawyer be a robot? It sounds far fetched, but artificial intelligence (AI) software systems - computer programs that can update and "think" by themselves - are increasingly being used by the legal community.
Source: <https://www.bbc.co.uk/news/business-58158820>

Source: <https://www.theverge.com/2021/7/5/22563751/tesla-elon-musk-full-self-driving-admission-autopilot-crash>

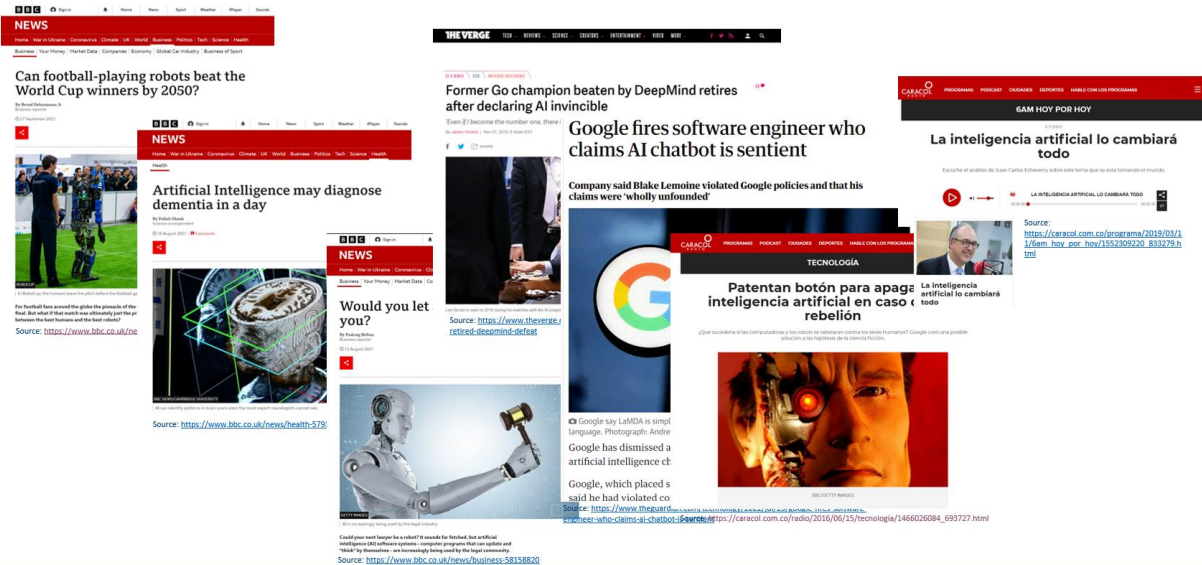


https://www.bbc.com/tecnologia/1466026084_693727.html

It is far from reality...

The Great AI Fallacy

AI in the news...



“The fallacy is associated with an implicit promise that is embedded in many statements about Artificial Intelligence. Artificial Intelligence, as it currently exists, is merely a form of automated decision making. **The implicit promise of Artificial Intelligence** is that it **will be the first wave of automation where the machine adapts to the human**, rather than the human adapting to the machine.”[1]

Neil Lawrence
DeepMind Professor
of Machine Learning
at the University
of Cambridge



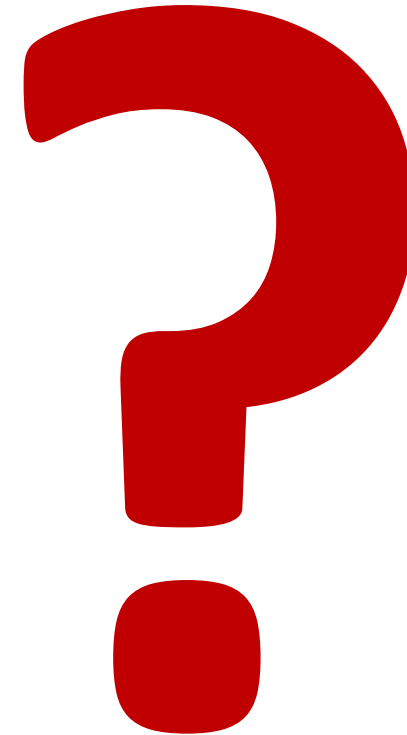
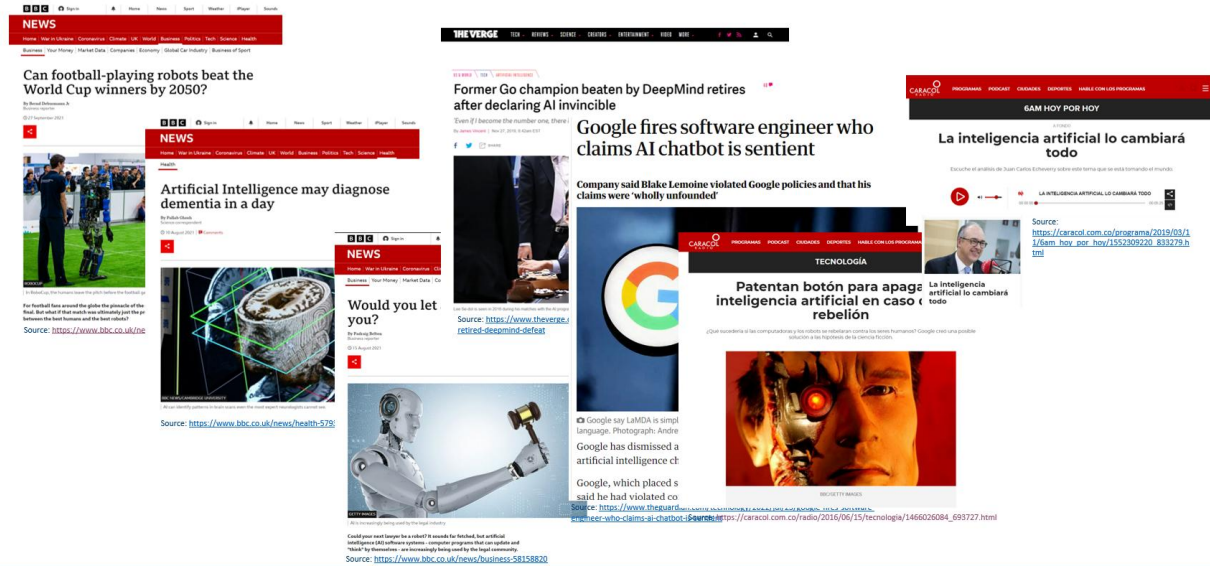
[1] Lawrence N. “The Great AI Fallacy”. Webinar for The Cambridge Network, Apr 2020, Available online: <http://inverseprobability.com/talks/notes/the-great-ai-fallacy.html>



UNIVERSITY OF
CAMBRIDGE

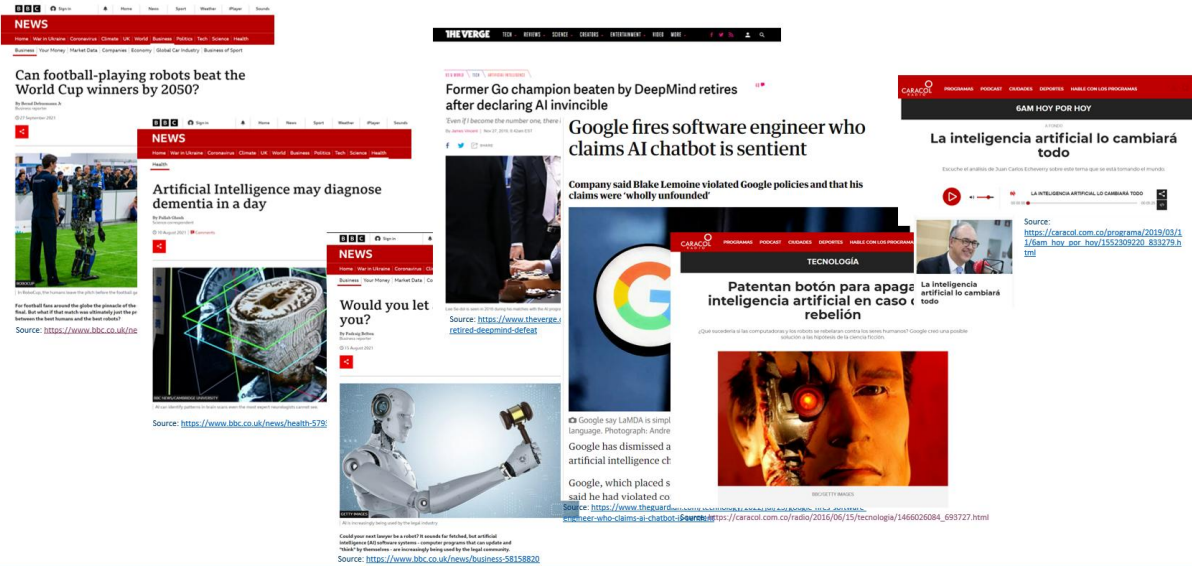
The Great AI Question

AI in the news...



The Great AI Question

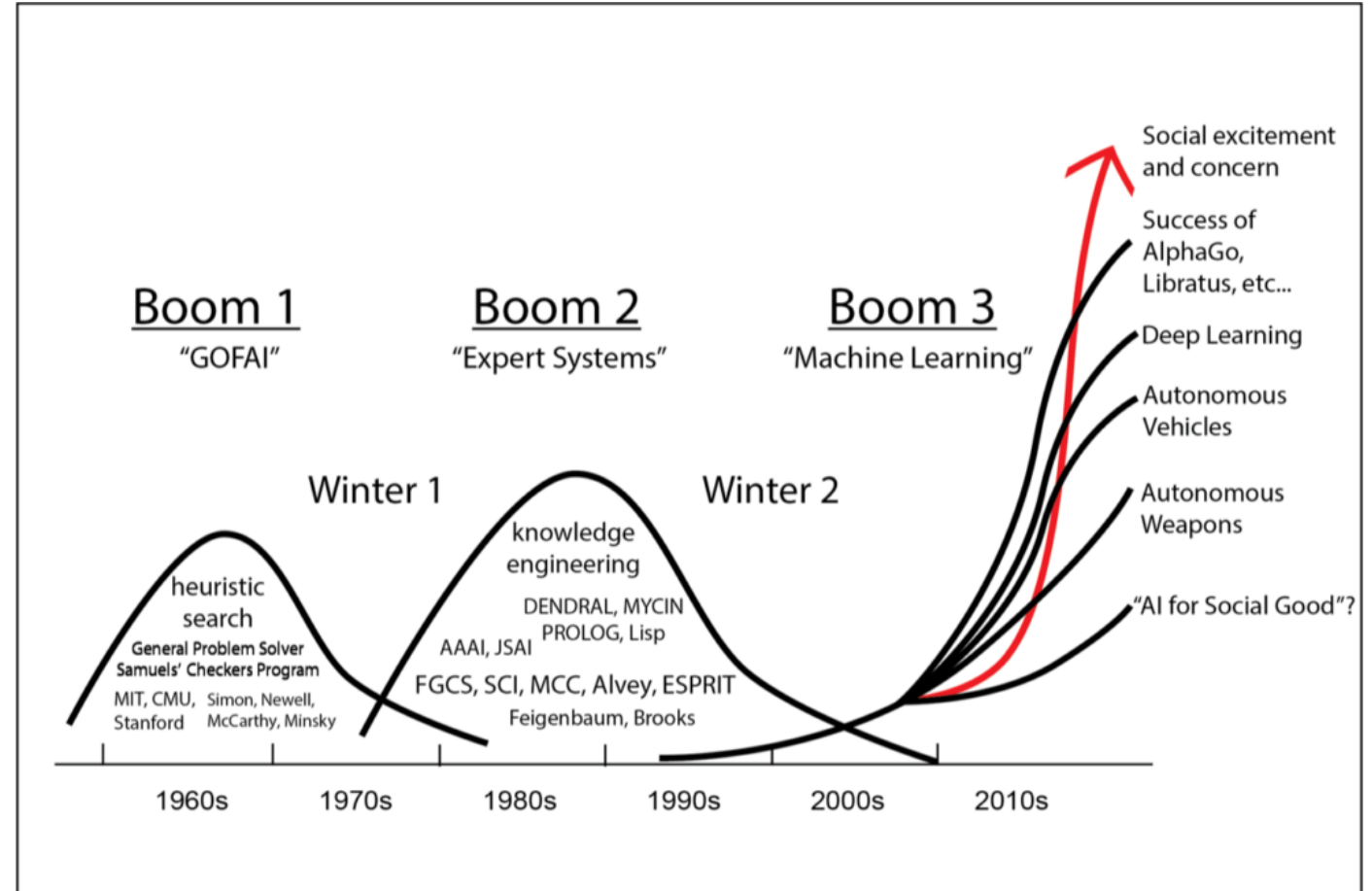
AI in the news...



- The News focus on impressive headlines.
- But they lack scientific rigor:
 - Problems are not well explained.
 - Assumptions are not considered.
 - Experimental setups are ignored.
 - Limitations are not discussed.
- Overwhelmed optimism!

Overwhelmed optimism

- **New technologies usually create this optimism:**
 - There were predictions about Humans conquering other worlds by first decades of 2000s.
- **Then, we have had disappointment periods:**
 - Spatial programs are not as attractive and popular as before.
- **Then, confidence is back again to evaluate actual results:**
 - Global Position System (GPS) is a tool we all use everyday!



Source: <https://hackernoon.com/ai-in-medicine-a-beginners-guide-a3b34b1dd5d7>

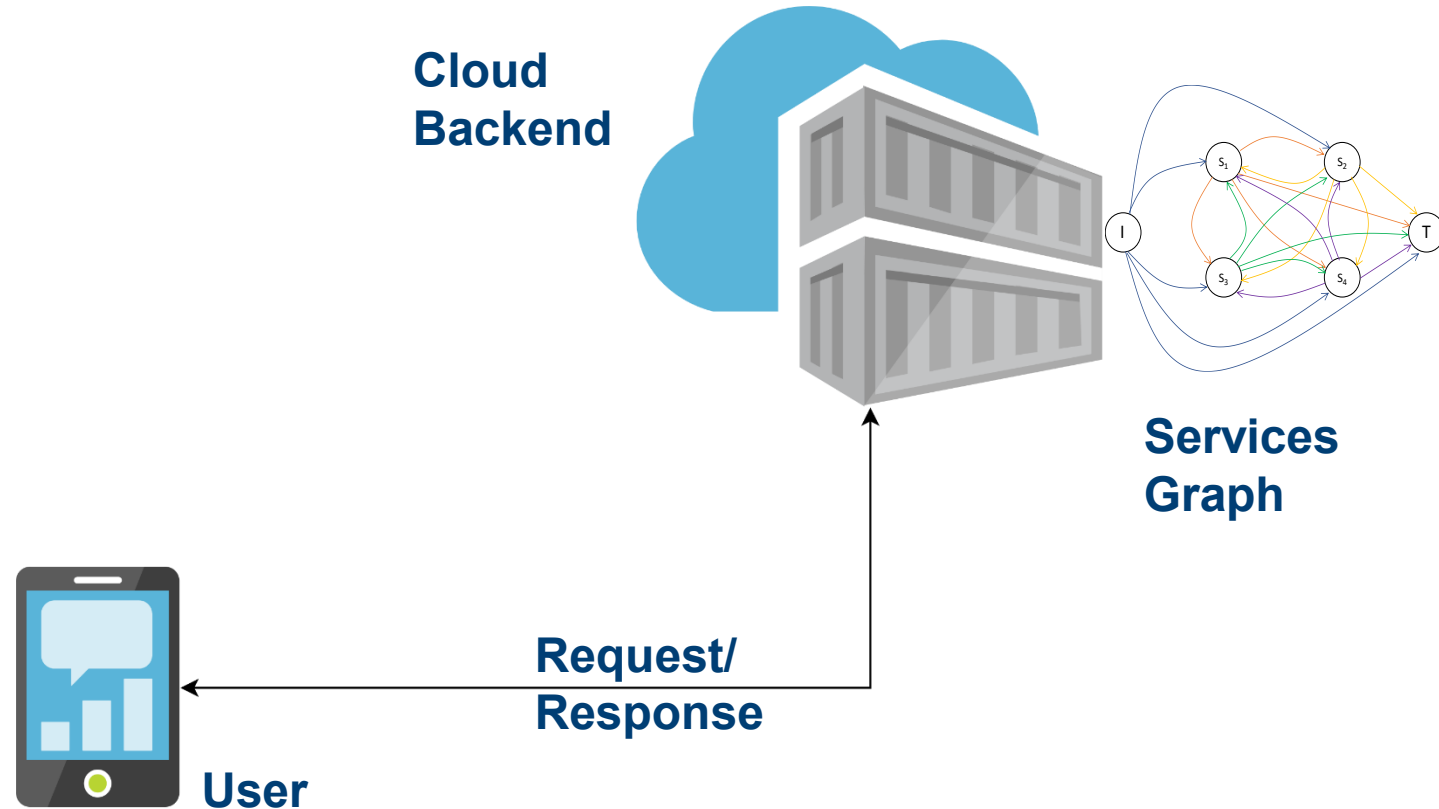


PART II: Why the question?



AI Application - Dynamic Service Placement in Edge Computing

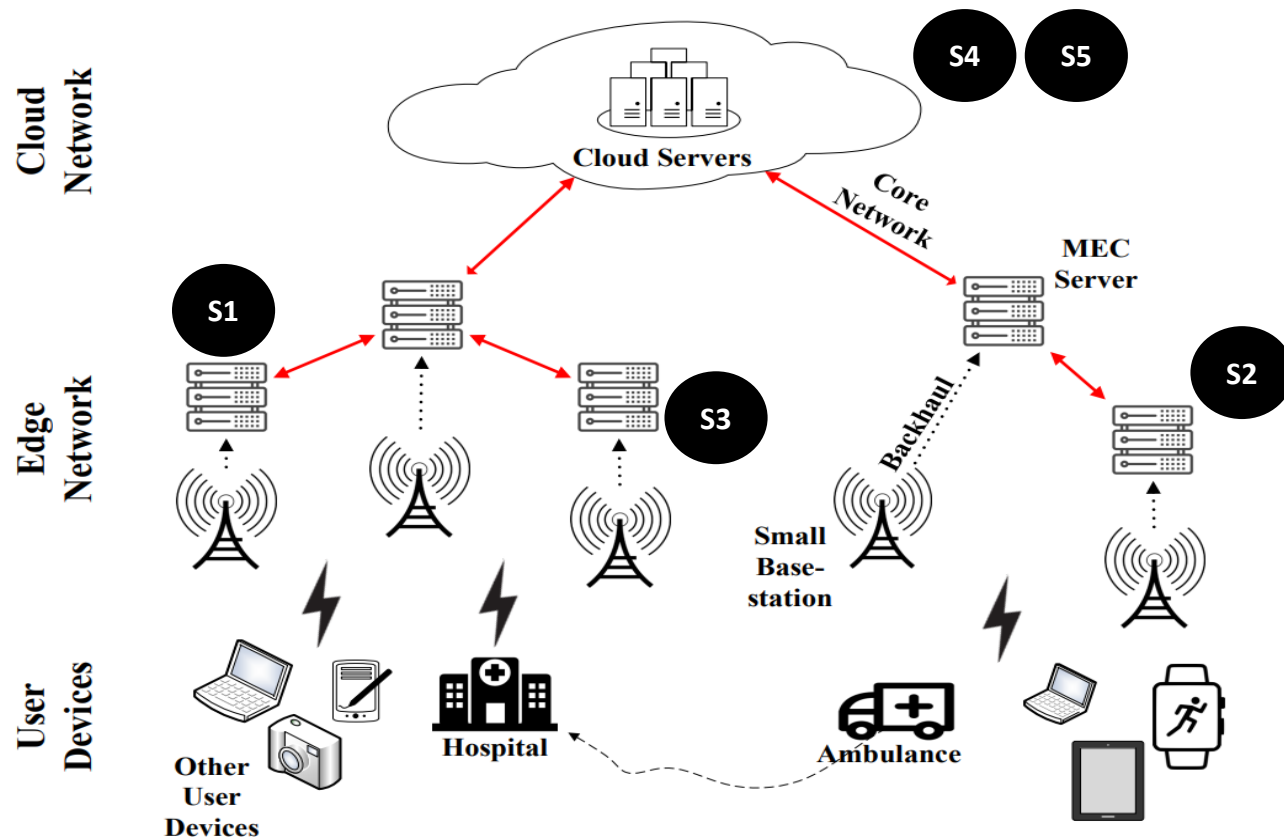
- Current applications are **cloud based** and they follow a **Microservices architecture**.
- There is a **physical distance** between users and the cloud, **which generates additional latency**.
- Some applications do not tolerate such latency.
- **Edge computing** is a new paradigm that enables **low-latency applications**[1].



[1] Malazi, H. T., Chaudhry, S. R., Kazmi, A., Palade, A., Cabrera, C., White, G., & Clarke, S. (2022). Dynamic Service Placement in Multi-access Edge Computing: a Systematic Literature Review. *IEEE Access*.



AI Application - Dynamic Service Placement in Edge Computing



- **Edge servers** are closed to end users, and they can **process data locally**.
- Services are executed on edge servers, but they have **limited resources**.
- The problem is to find the **optimal combination** of services and edge servers that **minimizes latency subject to available resources**.
- It is known as **Service Placement Problem[1]**.

[1] Malazi, H. T., Chaudhry, S. R., Kazmi, A., Palade, A., Cabrera, C., White, G., & Clarke, S. (2022). Dynamic Service Placement in Multi-access Edge Computing: a Systematic Literature Review. *IEEE Access*.



AI Application - Dynamic Service Placement in Edge Computing

- Pareto-optimal[1]

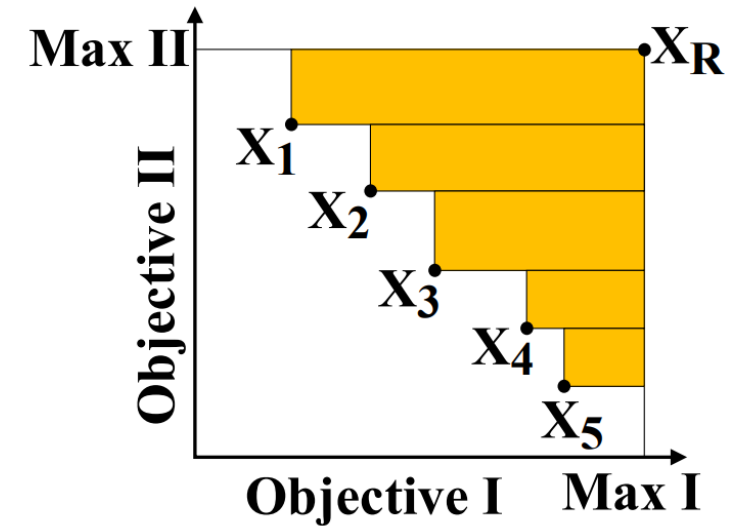
$$\min_j wt_j + lat(D_j) \quad (1)$$

$$\min_i \sqrt{w_p \sigma(RR^{CPU})^2 + w_m \sigma(RR^{RAM})^2} \quad (2)$$

subject to:

$$\sum_{k=1}^n rr_{ik}^{CPU} \leq r_i^{CPU} \quad (3)$$

$$\sum_{k=1}^n rr_{ik}^{RAM} \leq r_i^{RAM} \quad (4)$$

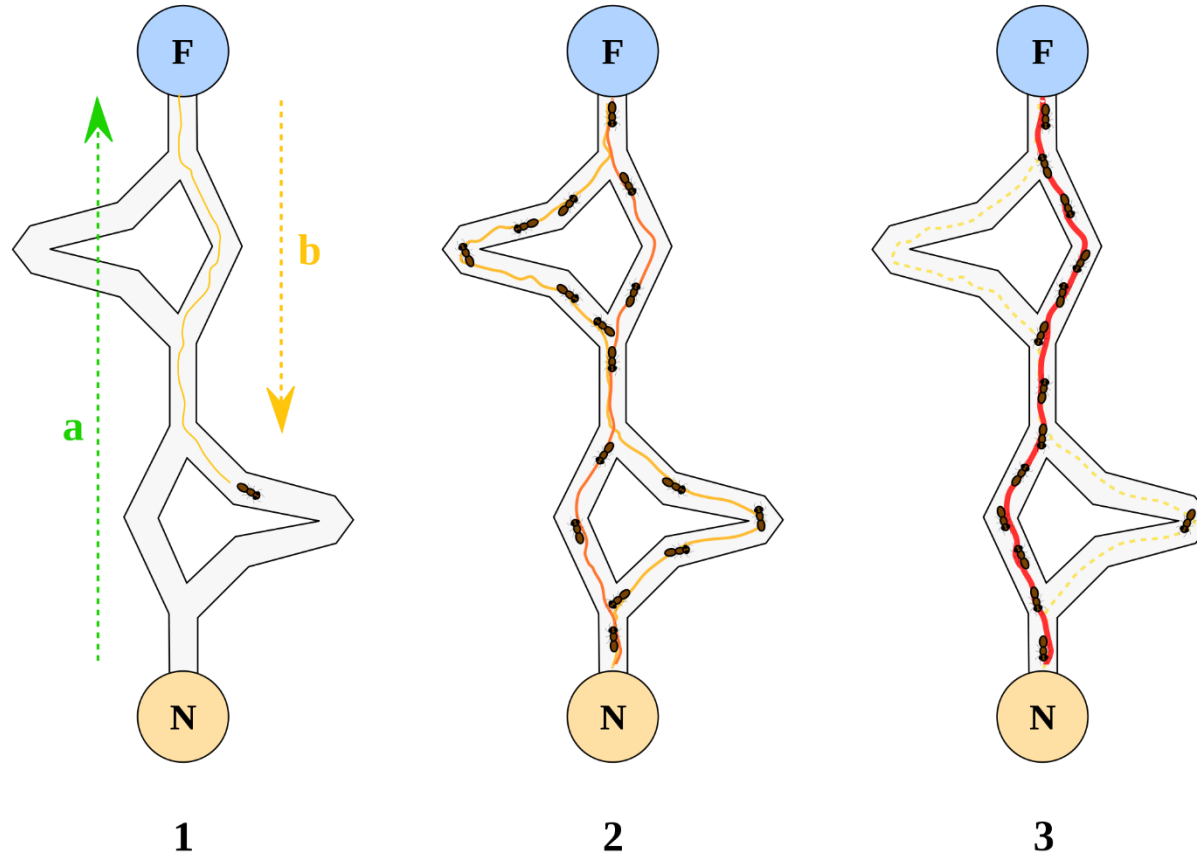


[1] Palade, A., Cabrera, C., White, G., & Clarke, S. (2018, November). Stigmergic service composition and adaptation in mobile environments. In *International Conference on Service-Oriented Computing* (pp. 618-633). Springer, Cham.



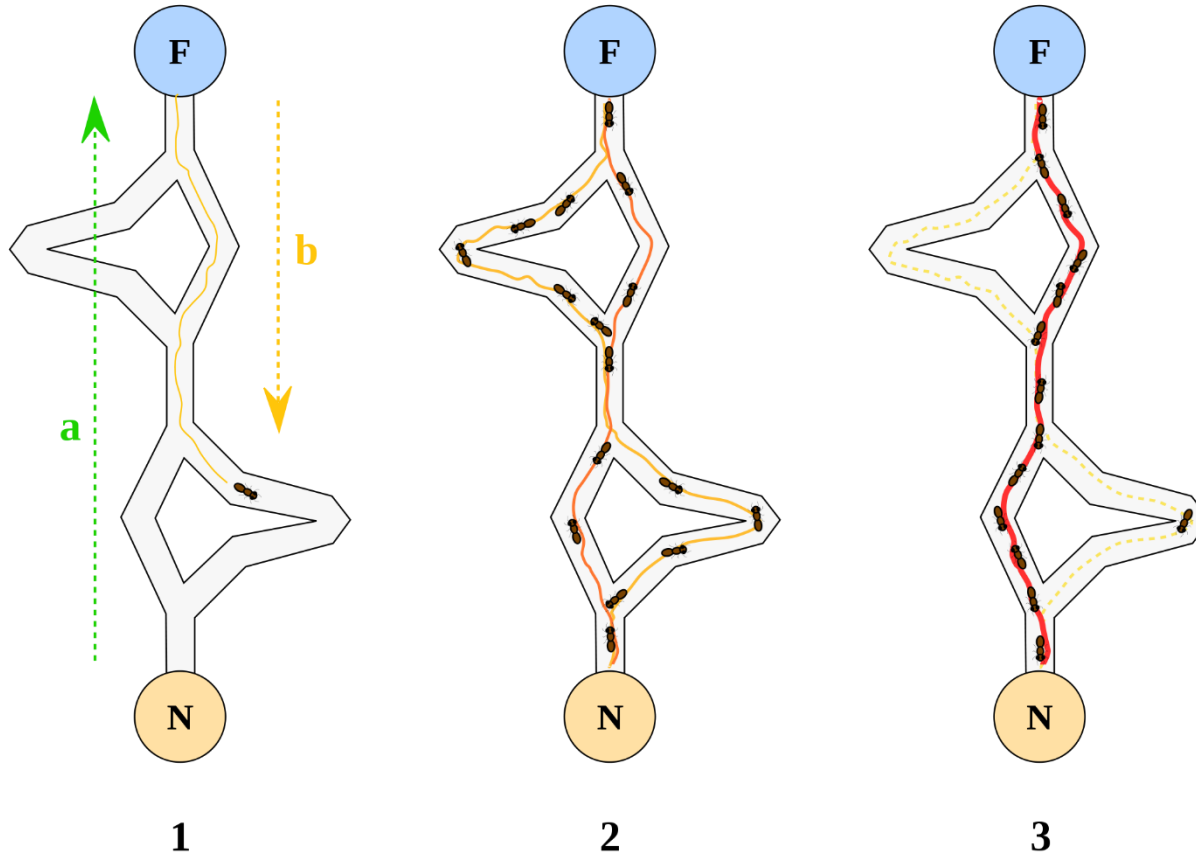
AI Application - Dynamic Service Placement in Edge Computing

- Ant-Colony Optimisation



AI Application - Dynamic Service Placement in Edge Computing

- Ant-Colony Optimisation



Forward movement:

$$p(t, i) = \frac{[\tau_{t,i}]^{\alpha} * [\eta_i]^{\beta}}{\sum_{i=1}^n [\tau_{t,j}]^{\alpha} * [\eta_j]^{\beta}}$$

Backward movement:

$$\tau_{t,i} = \tau_{t,i} + \rho$$



AI Application - Dynamic Service Placement in Edge Computing



AI Application - Dynamic Service Placement in Edge Computing

$App_1 = \langle S, RQ, l, p \rangle$

$App_2 = \langle S, RQ, l, p \rangle$

...

$App_n = \langle S, RQ, l, p \rangle$



$N_1 = \langle R, L, l \rangle$



$N_2 = \langle R, L, l \rangle$



$N_3 = \langle R, L, l \rangle$



$N_4 = \langle R, L, l \rangle$

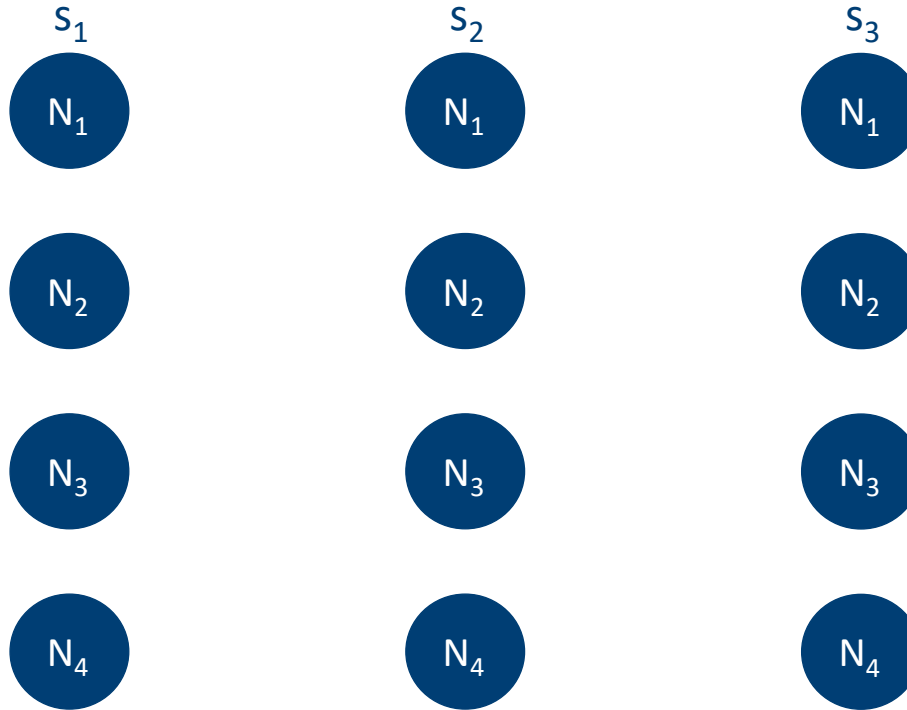


AI Application - Dynamic Service Placement in Edge Computing

$App_1 = \langle S, RQ, l, p \rangle$

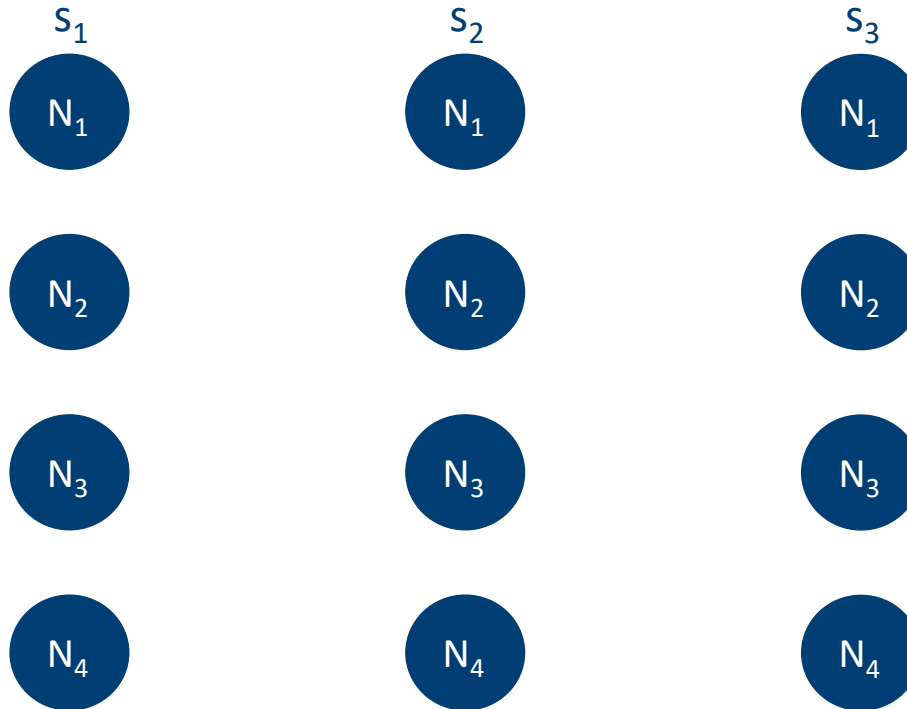
$S = \langle s_1, s_2, s_3 \rangle$

$RQ = \langle R_{f1}, R_{f2}, R_{f3} \rangle$



AI Application - Dynamic Service Placement in Edge Computing

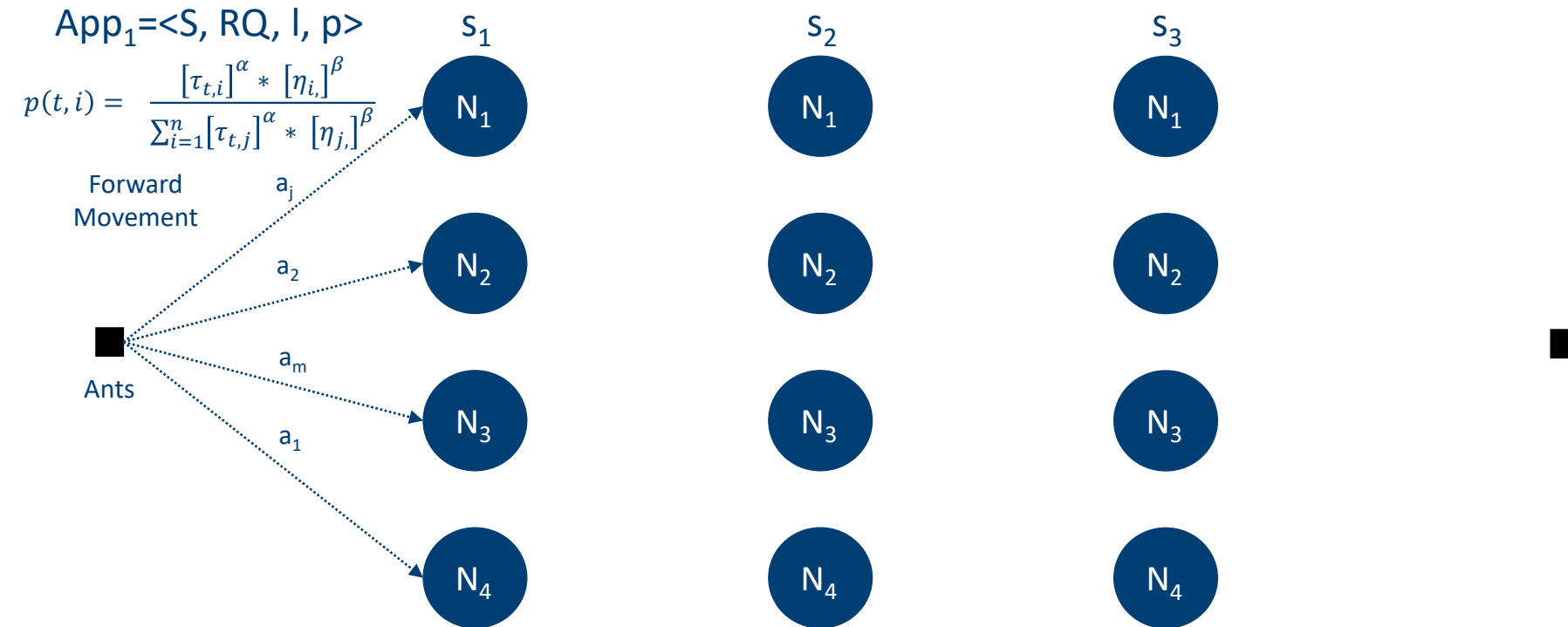
$App_1 = \langle S, RQ, l, p \rangle$



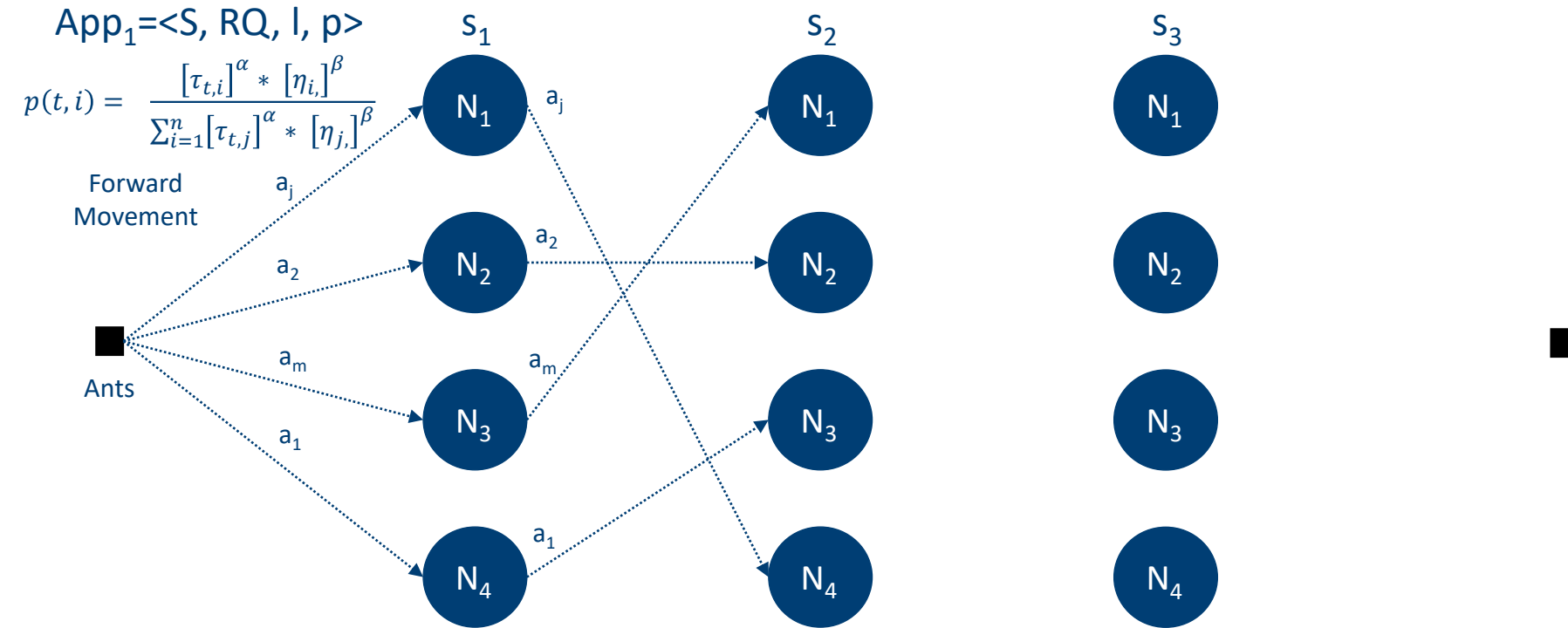
■
 $\langle a_1, a_2, \dots, a_j, \dots, a_m \rangle$
Ants



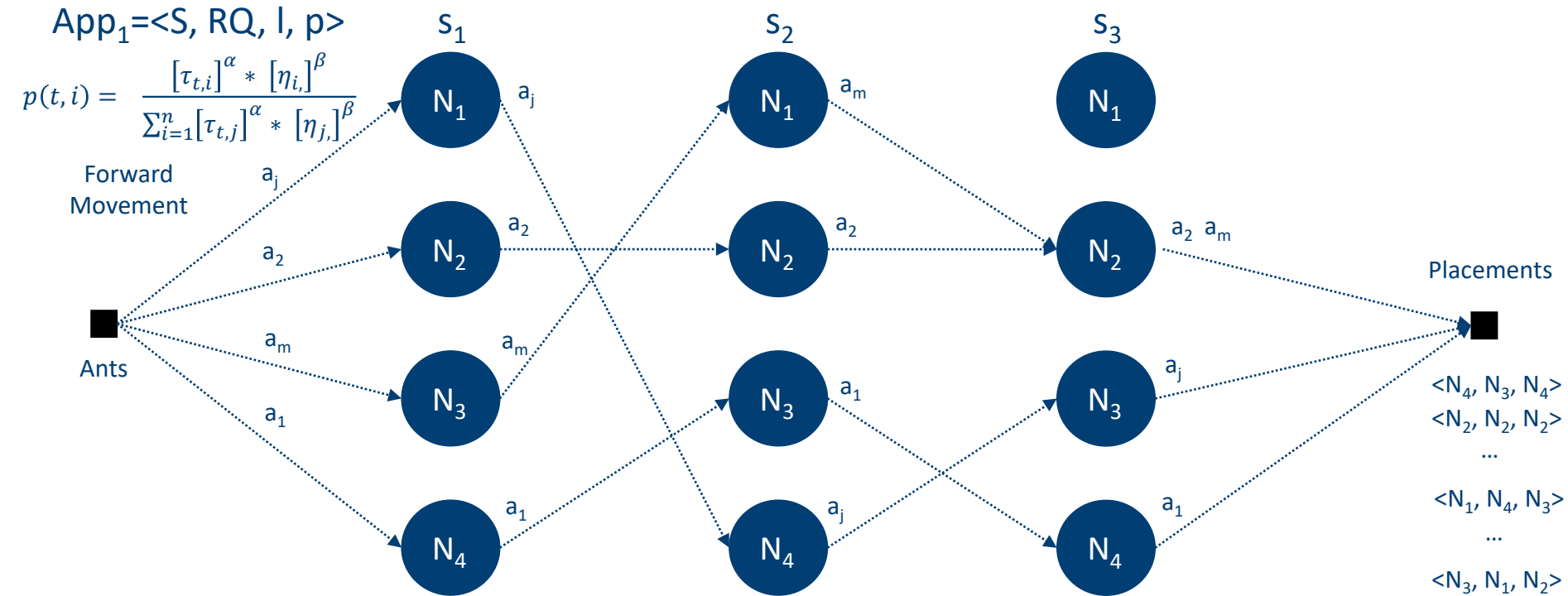
AI Application - Dynamic Service Placement in Edge Computing



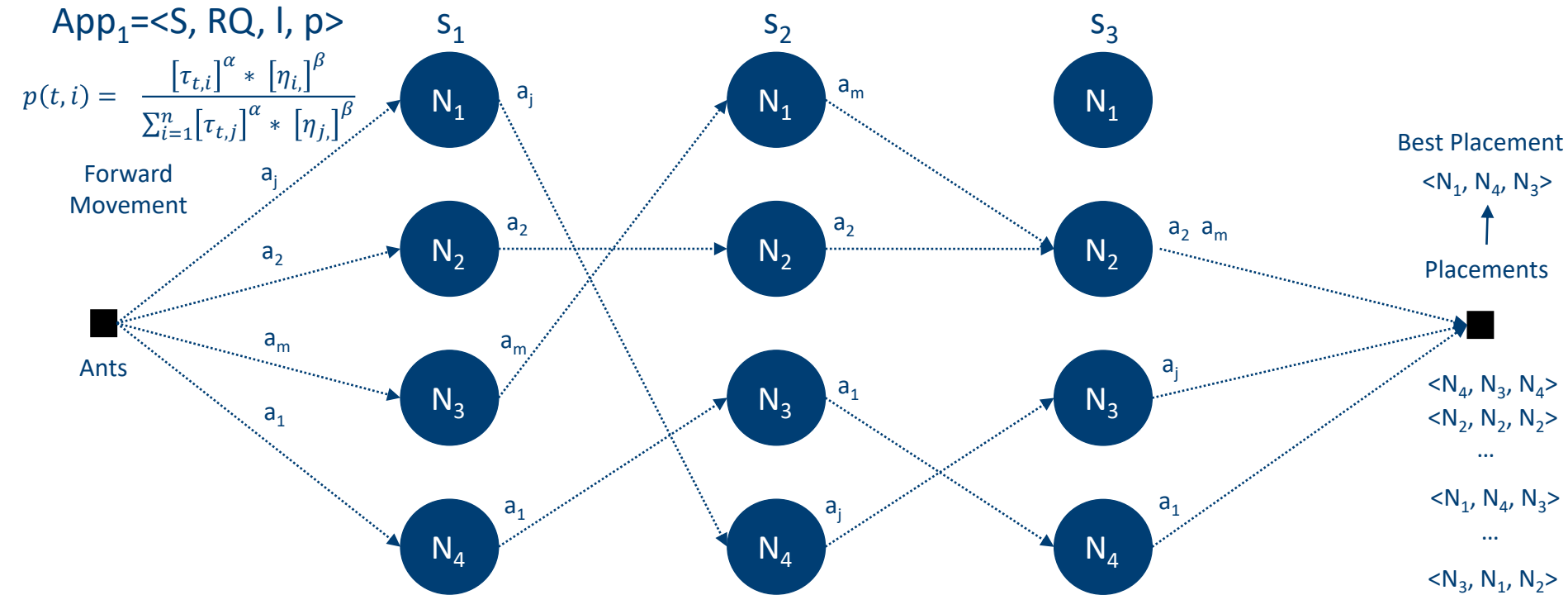
AI Application - Dynamic Service Placement in Edge Computing



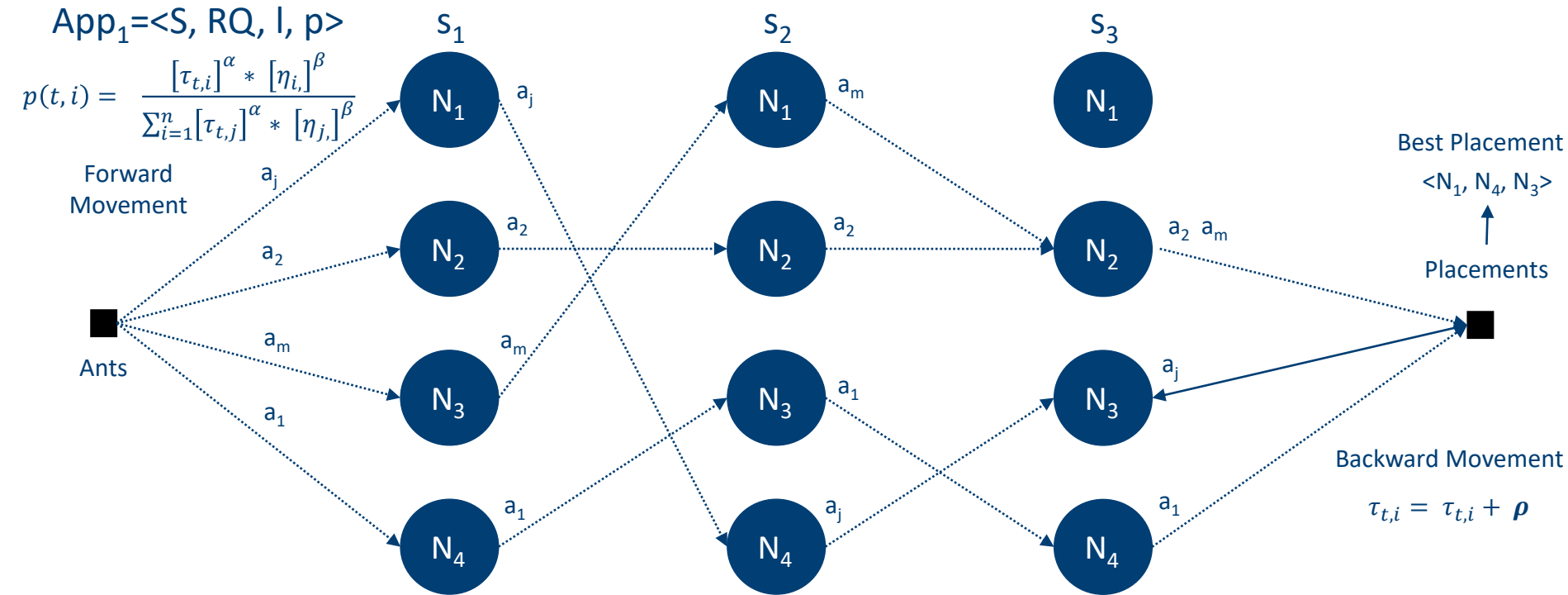
AI Application - Dynamic Service Placement in Edge Computing



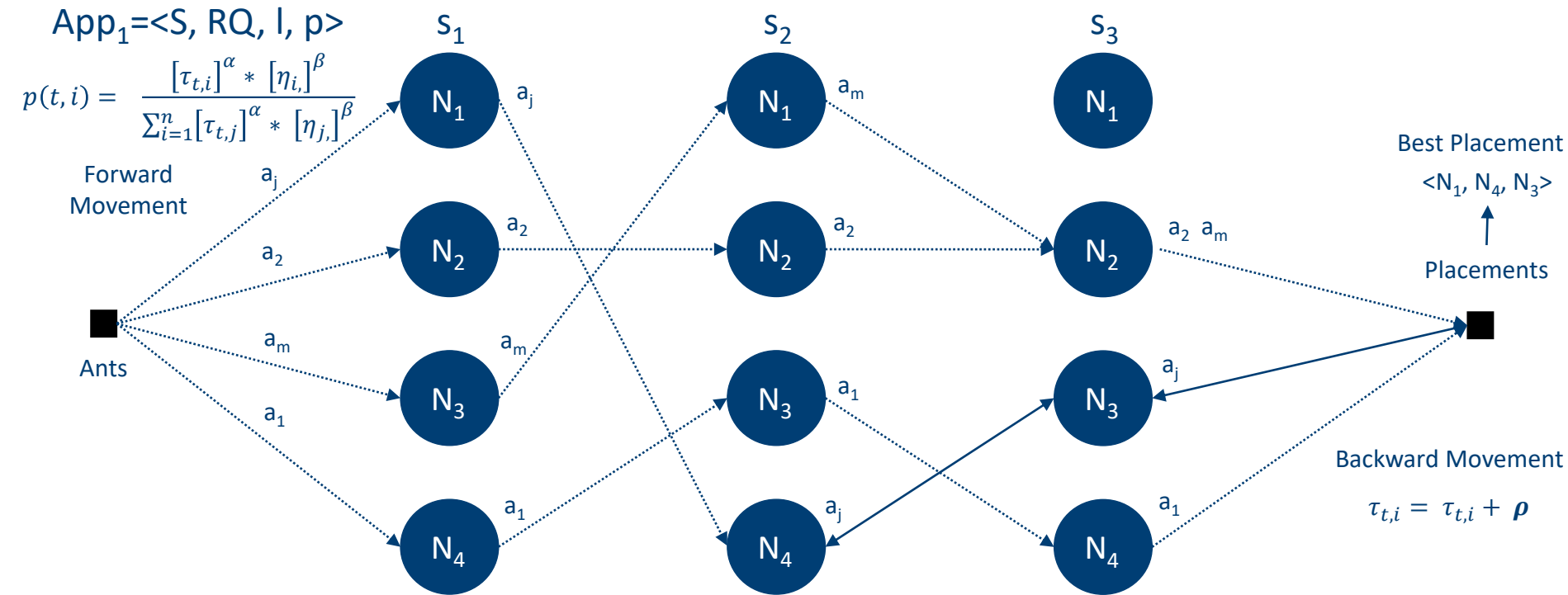
AI Application - Dynamic Service Placement in Edge Computing



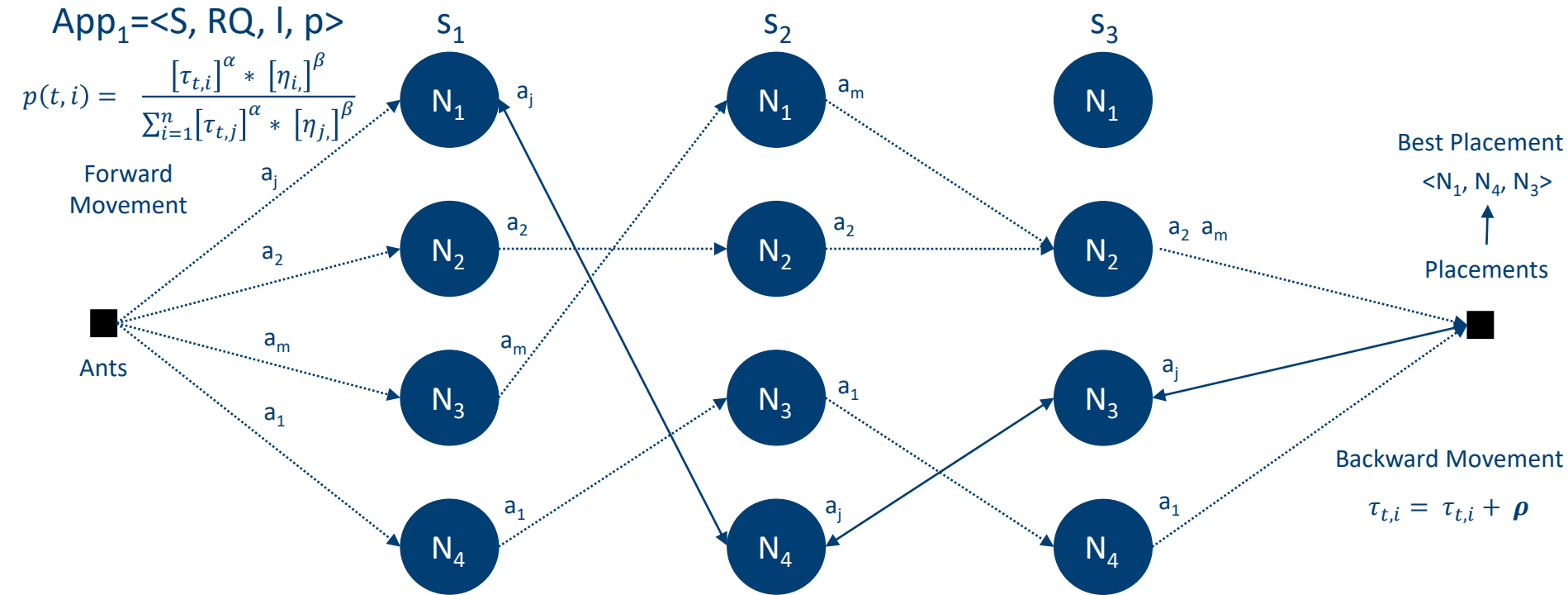
AI Application - Dynamic Service Placement in Edge Computing



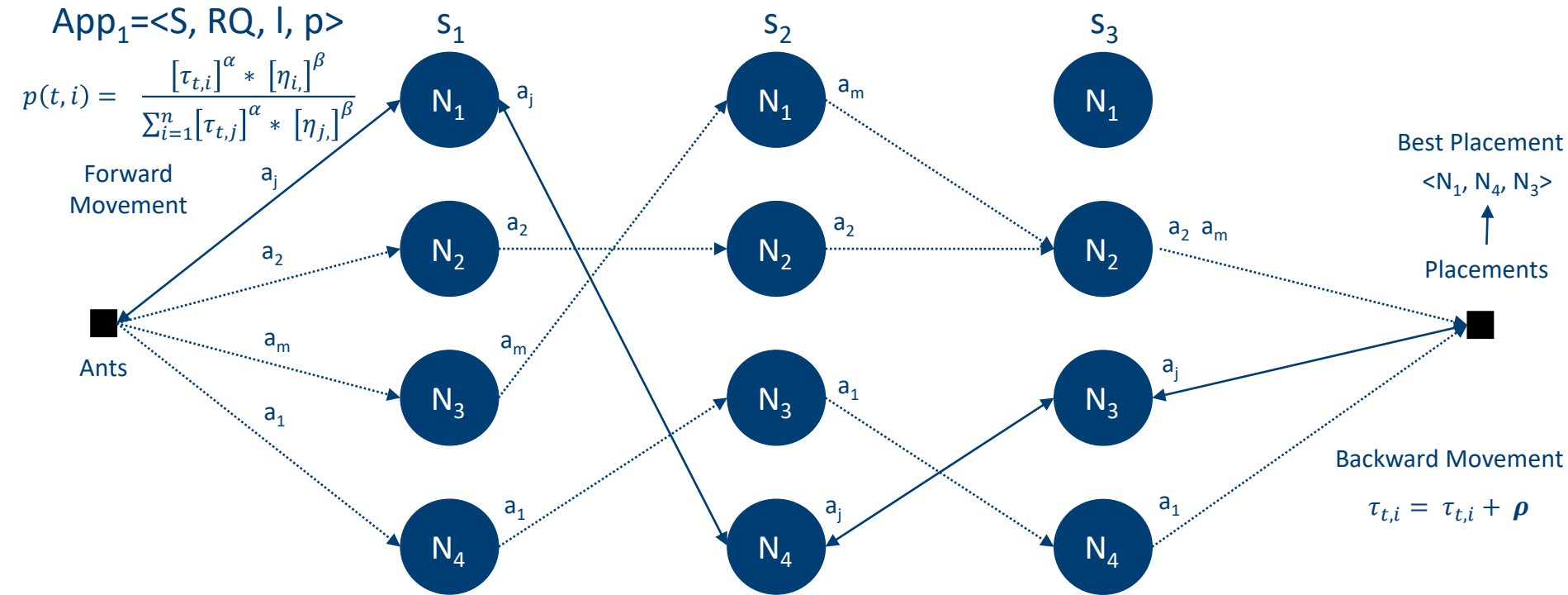
AI Application - Dynamic Service Placement in Edge Computing



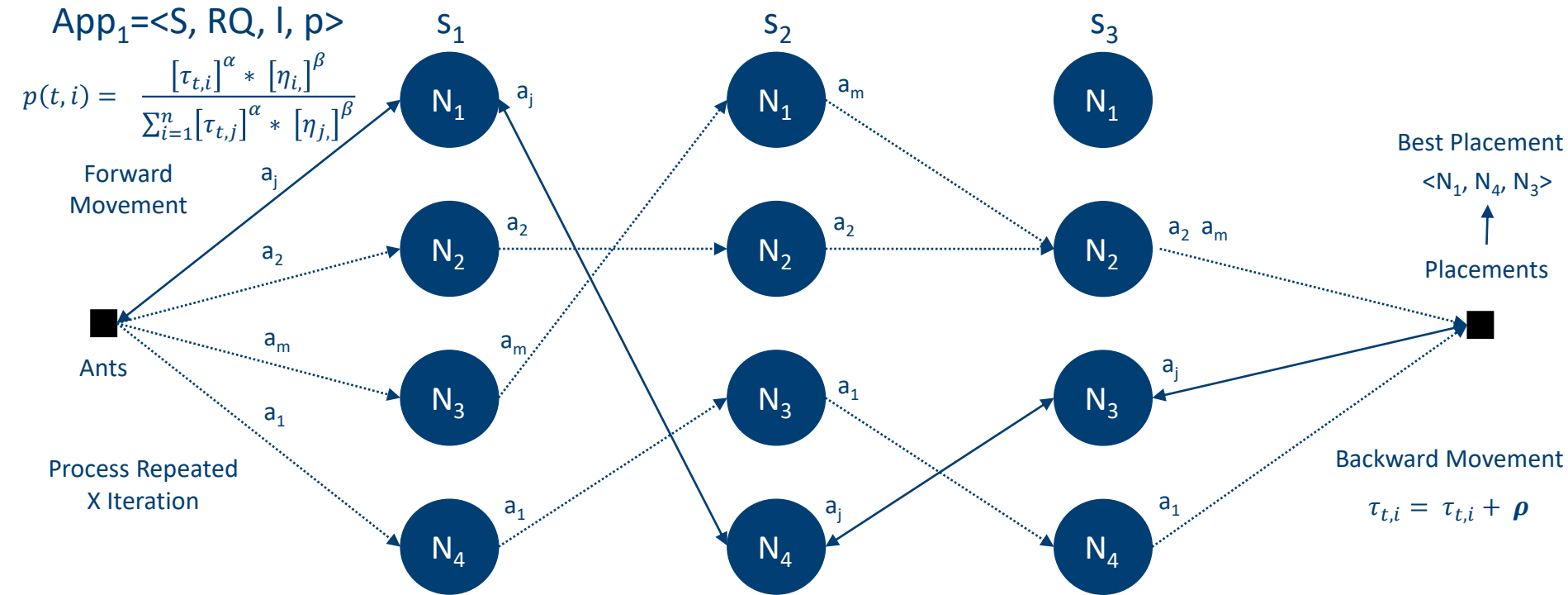
AI Application - Dynamic Service Placement in Edge Computing



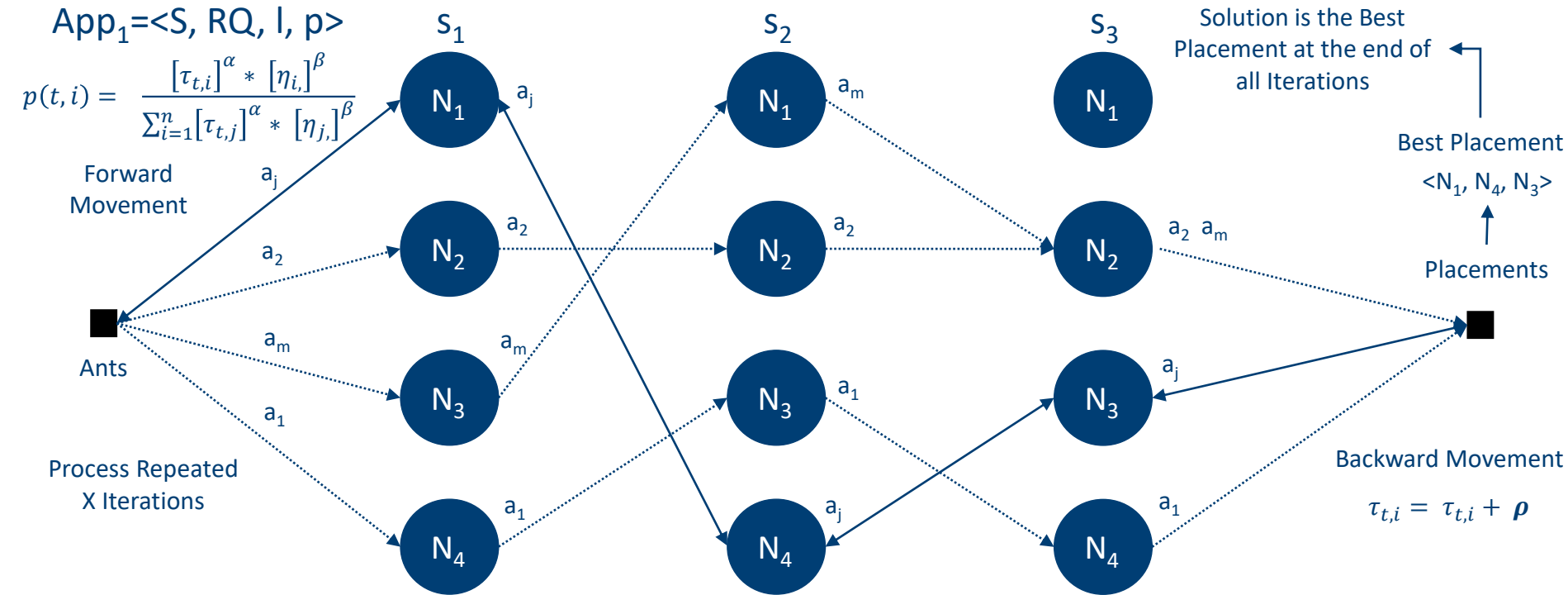
AI Application - Dynamic Service Placement in Edge Computing



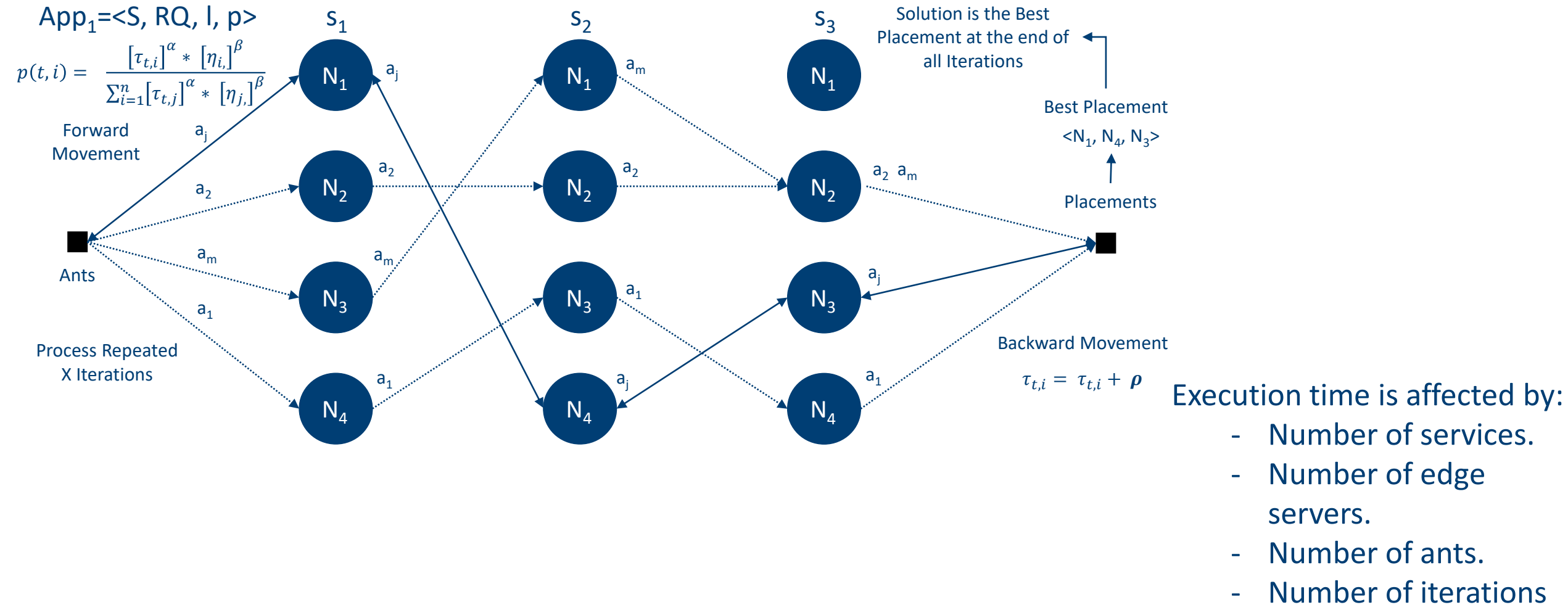
AI Application - Dynamic Service Placement in Edge Computing



AI Application - Dynamic Service Placement in Edge Computing



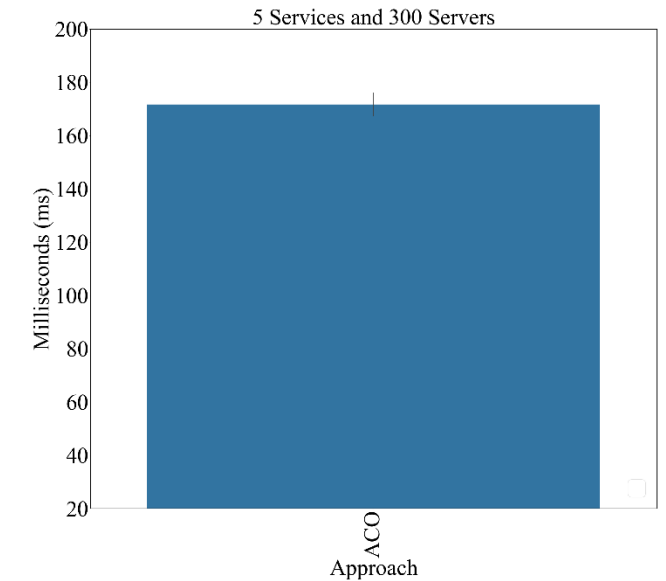
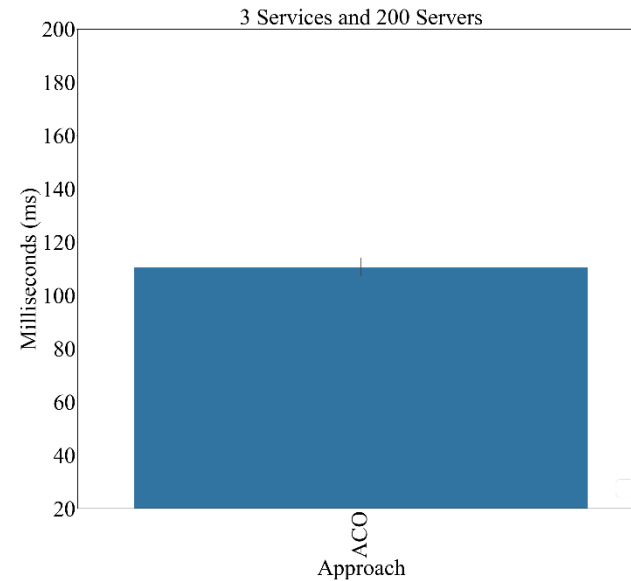
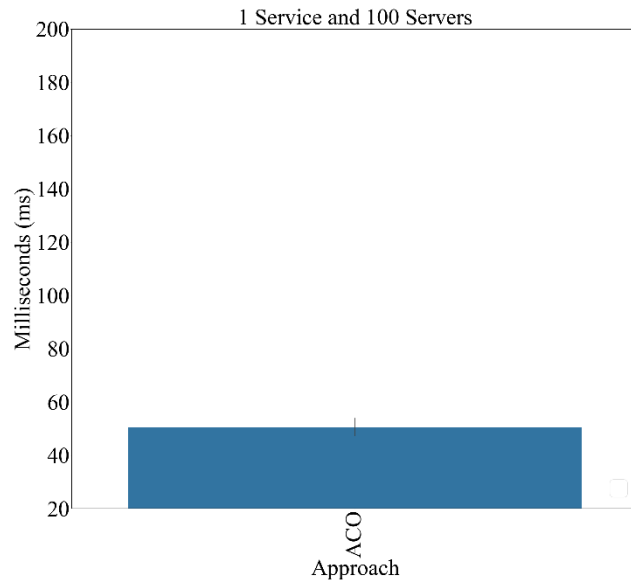
AI Application - Dynamic Service Placement in Edge Computing



AI Application - Dynamic Service Placement in Edge Computing

Smart-city simulation[1]

Number of ants: 10
Number of iterations: 100



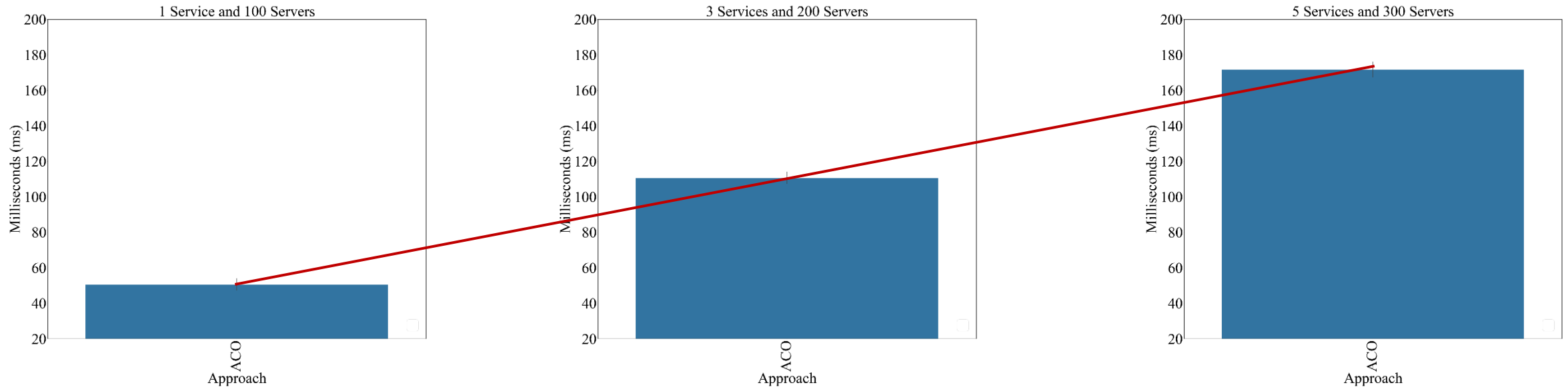
[1] Richerzhagen, B., Stingl, D., Ruckert, J., & Steinmetz, R. (2015, August). Simonstrator: Simulation and prototyping platform for distributed mobile applications. In *The 8th EAI International Conference on Simulation Tools and Techniques (ACM SIMUTOOLS 2015)* (pp. 99-108).



AI Application - Dynamic Service Placement in Edge Computing

Smart-city simulation[1]

Number of ants: 10
Number of iterations: 100

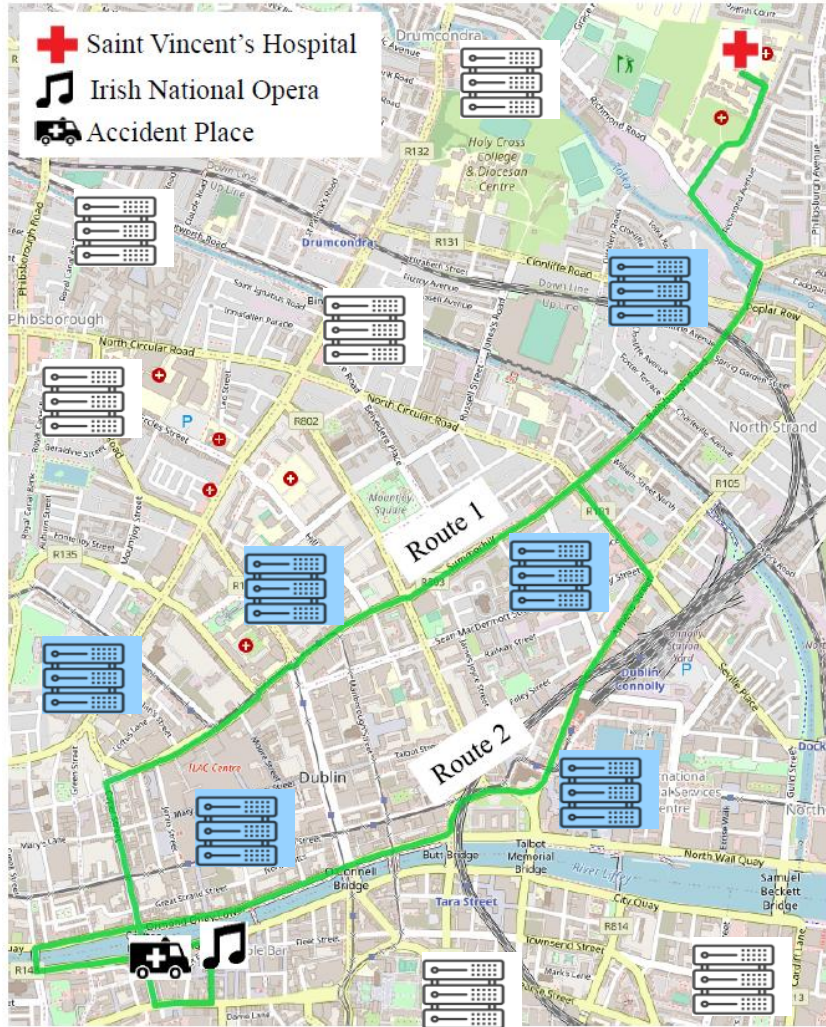


This execution time does not suit low-latency requirements, but that is how ACO is designed!!!

[1] Richerzhagen, B., Stingl, D., Ruckert, J., & Steinmetz, R. (2015, August). Simonstrator: Simulation and prototyping platform for distributed mobile applications. In *The 8th EAI International Conference on Simulation Tools and Techniques (ACM SIMUTOOLS 2015)* (pp. 99-108).



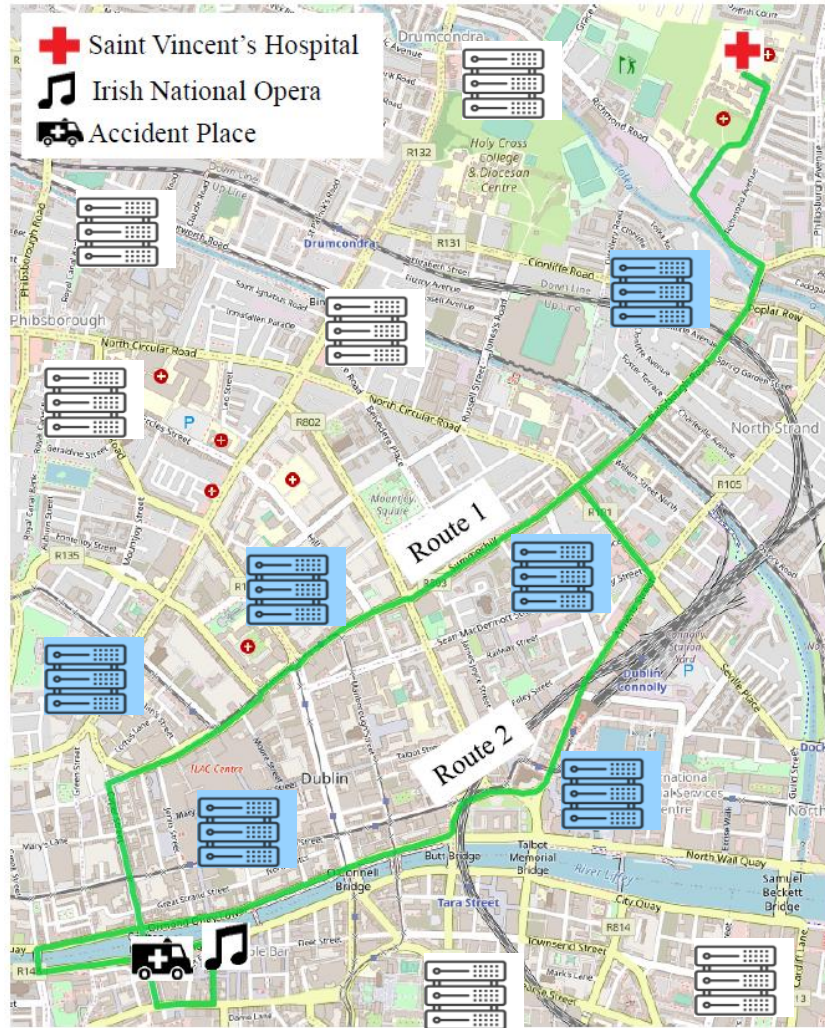
AI Application - Dynamic Service Placement in Edge Computing



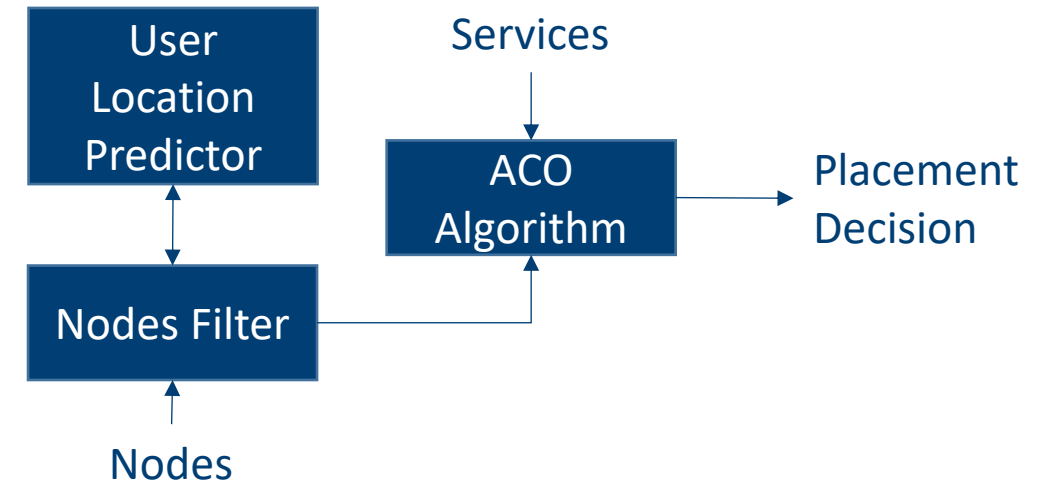
- Variables we cannot reduce:
 - Number of services.
 - Number of iterations.
 - Number of ants.
- **We can reduce the number of servers:**
 - We can pre-select them if we can predict user locations.



AI Application - Dynamic Service Placement in Edge Computing



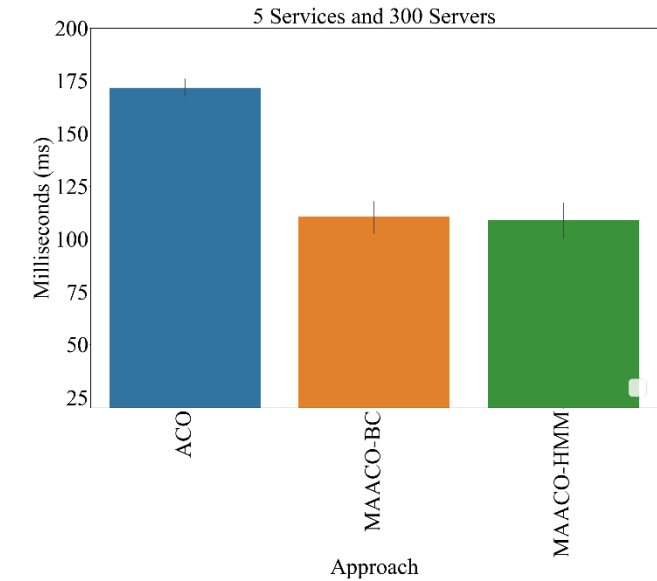
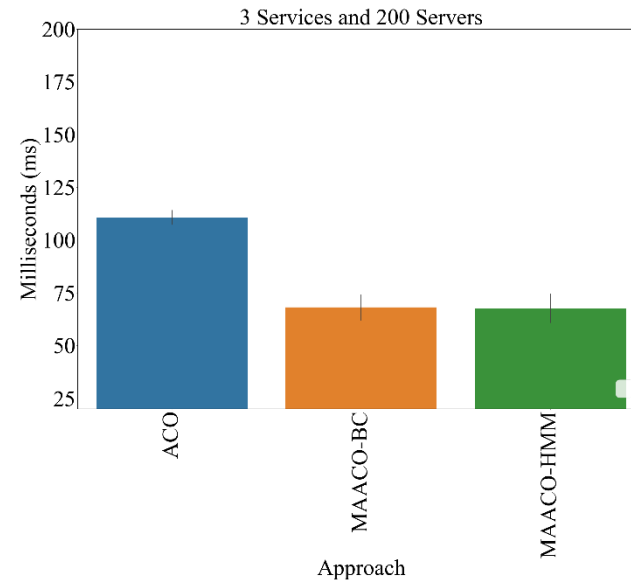
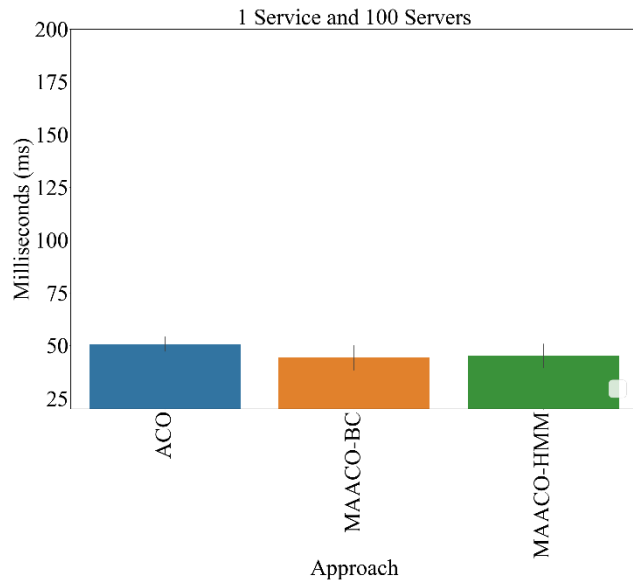
Re-design of the solution:



AI Application - Dynamic Service Placement in Edge Computing

Smart-city simulation[1]

Number of ants: 10
Number of iterations: 100



[1] Richerzhagen, B., Stingl, D., Ruckert, J., & Steinmetz, R. (2015, August). Simonstrator: Simulation and prototyping platform for distributed mobile applications. In *The 8th EAI International Conference on Simulation Tools and Techniques (ACM SIMUTOOLS 2015)* (pp. 99-108).



AI Application - Dynamic Service Placement in Edge Computing

Bayesian Classifier [1]

- Transition matrix depends on the number of links, which are the number of streets in a city.
- A lot of data (i.e., trips) are needed to train the model.
- Training time is now an issue!
- We assumed a limited number of streets in our work.

Hidden Markov Model [2]

- Frequency matrix depends on the number of links, which are the number of streets in a city.
- A lot of data (i.e., trips) are needed to train the model.
- Training time is now an issue!
- We assumed a limited number of streets in our work.

[1] J. P. Epperlein, J. Monteil, M. Liu, Y. Gu, S. Zhuk, and R. Shorten, "Bayesian classifier for route prediction with markov chains," in 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018, pp. 677–682.

[2] Y. Lassoued, J. Monteil, Y. Gu, G. Russo, R. Shorten, and M. Mevissen, "A hidden markov model for route and destination prediction," in 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2017.



AI Application - Dynamic Service Placement in Edge Computing

Bayesian Classifier [1]

- Transition matrix depends on the number of links, which are the number of streets in a city.
- A lot of data (i.e., trips) are needed to train the model.
- Training time is now an issue!
- We **assumed** a limited number of streets in our work.

Hidden Markov Model [2]

- Frequency matrix depends on the number of links, which are the number of streets in a city.
- A lot of data (i.e., trips) are needed to train the model.
- Training time is now an issue!
- We **assumed** a limited number of streets in our work.

**AGAIN, NEW DESIGN DECISIONS ARE NEEDED TO
DEPLOY THESE ML ALGORITHMS IN THE REAL-WORLD!**

[1] J. P. Epperlein, J. Monteil, M. Liu, Y. Gu, S. Zhuk, and R. Shorten, "Bayesian classifier for route prediction with markov chains," in 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018, pp. 677–682.

[2] Y. Lassoued, J. Monteil, Y. Gu, G. Russo, R. Shorten, and M. Mevissen, "A hidden markov model for route and destination prediction," in 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2017.



AI Deployment in the Real-world

- Real-world does not support assumptions.
- Real world environments are usually large, heterogeneous, complex and dynamic.

How to support engineers and developers in the process of deploying AI-based systems in real-world environments?

How to enable the reasoning about AI-based systems in real world environments?

How to automate the repetitive design process?



PART III: Possible answers...



AutoAI Programme

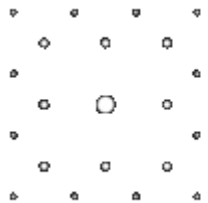
The AutoAI Programme scales our ability to deploy safe and reliable AI solutions, driving innovation in machine learning-enabled techniques for deploying, maintaining, and understanding AI systems. By investigating how to decompose AI systems into their component parts, how to manage data in system development, and how to monitor performance in deployment, **AutoAI will develop a new AI design and engineering paradigm.**



AutoAI Programme



faculty



Ch^{AI}

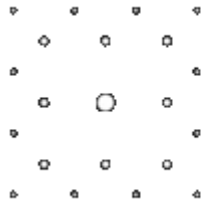


UNIVERSITY OF
CAMBRIDGE

AutoAI Programme



faculty



Ch^{AI}



Data Science Africa - DSA @dsa_org · Jul 19

...



1

4

14



UNIVERSITY OF
CAMBRIDGE

AutoAI Programme



facu

PROBLEM FIRST!!!



Data Science Africa - DSA @dsa_org · Jul 19



1



4



14



UNIVERSITY OF
CAMBRIDGE

AutoAI Programme

- What is the problem?
- Functional vs Non-functional requirements?
- Why do we need AI?
- How would AI impact our requirements?

THE MACHINE LEARNING CANVAS

Designed for:

Designed by:

Date:

Iteration:

PREDICTION TASK  Type of task? Entity on which predictions are made? Possible outcomes? Wait time before observation?	DECISIONS  How are predictions turned into proposed value for the end-user? Mention parameters of the process / application that does that.	VALUE PROPOSITION  Who is the end-user? What are their objectives? How will they benefit from the ML system? Mention workflow/interfaces.	DATA COLLECTION  Strategy for initial train set & continuous update. Mention collection rate, holdout on production entities, cost/constraints to observe outcomes.	DATA SOURCES  Where can we get (raw) information on entities and observed outcomes? Mention database tables, API methods, websites to scrape, etc.
IMPACT SIMULATION  Can models be deployed? Which test data to assess performance? Cost/gain values for (in)correct decisions? <u>Fairness constraint</u> ?	MAKING PREDICTIONS  When do we make real-time / batch pred.? Time available for this + featurization + post-processing? Compute target?		BUILDING MODELS  How many prod models are needed? When would we update? Time available for this (including featurization and analysis)?	FEATURES  Input representations available at prediction time, extracted from raw data sources.
MONITORING  Metrics to quantify value creation and measure the ML system's impact in production (on end-users and business)?				



Version 1.1. Created by Louis Dorard, Ph.D. Licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/). Please keep this mention and the link to ownml.co when sharing.

[OWNML.CO](https://ownml.co)

AutoAI Programme

compute
model + data → prediction



AutoAI Programme

compute
model + data → prediction

- There has been **a lot** of focus on and research **efforts** in learning **models** (i.e., new algorithms) and **compute** (i.e., more powerful machines).
- But, the data has not gotten enough attention!



AutoAI Programme

compute
model + data → prediction

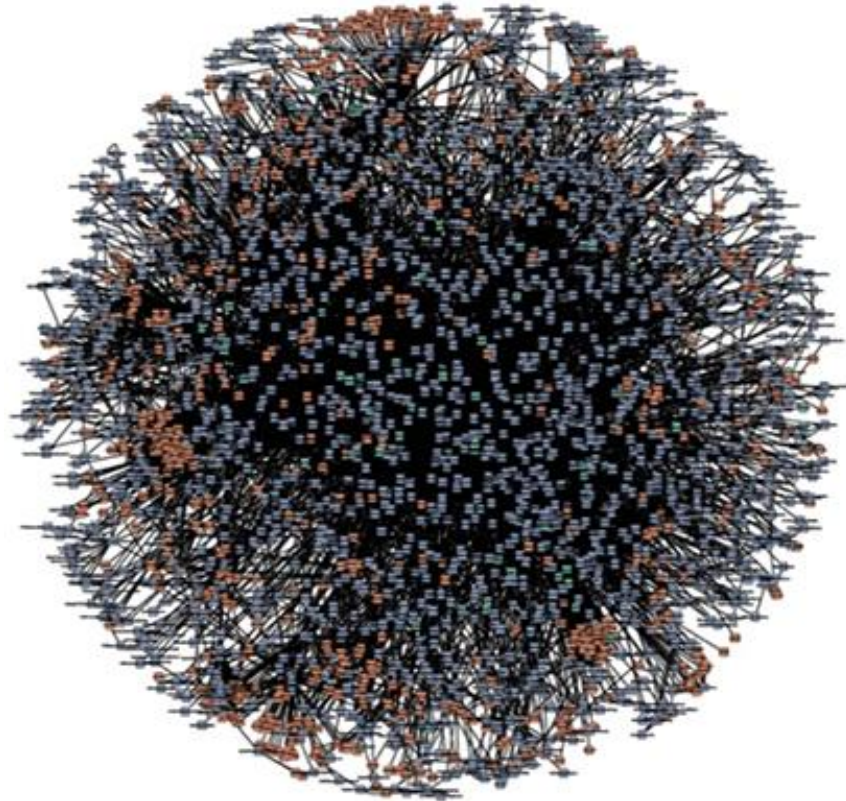
- There has been **a lot** of focus on and research **efforts** in learning **models** (i.e., new algorithms) and **compute** (i.e., more powerful machines).
- But, the data has not gotten enough attention!

The Data Dichotomy:

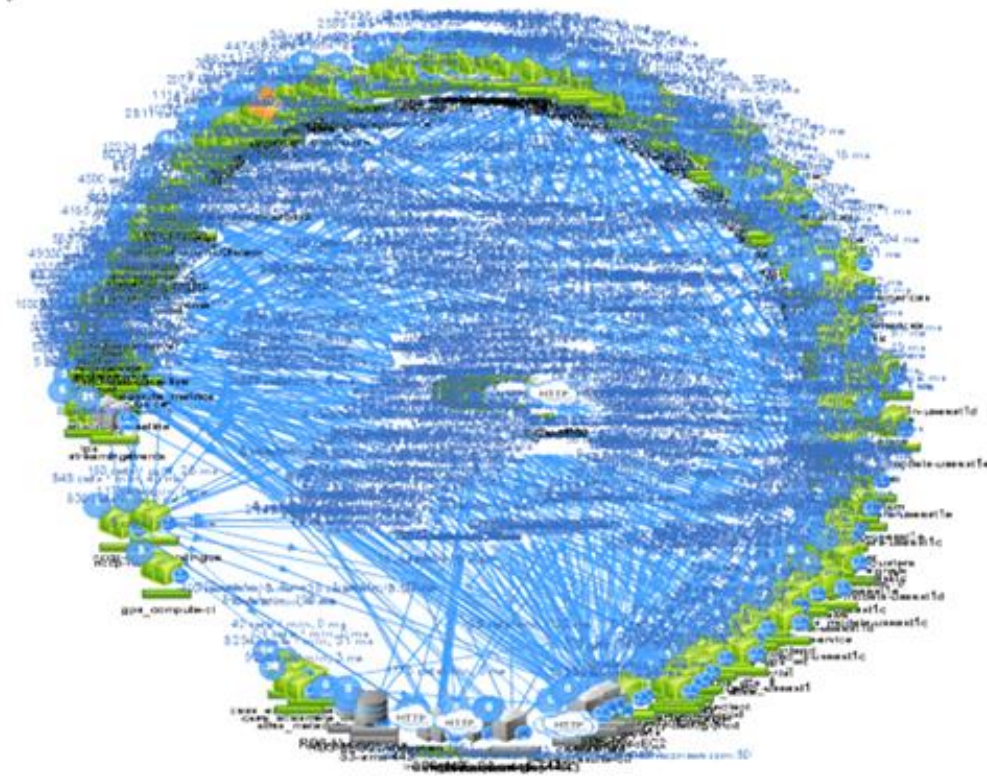
“While data-driven systems are about exposing data, service oriented architectures are about hiding data.”



AutoAI Programme



amazon.com[®]



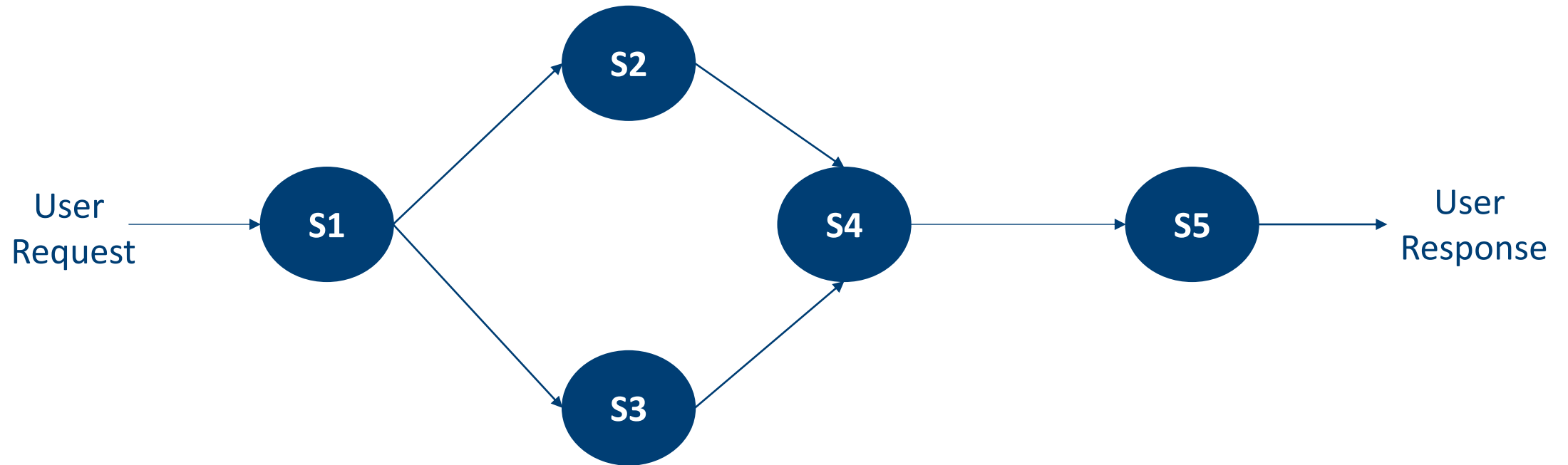
Source: <https://www.divante.com/blog/10-companies-that-implemented-the-microservice-architecture-and-paved-the-way-for-others>



UNIVERSITY OF
CAMBRIDGE

AutoAI Programme

- **Microservices Architecture**

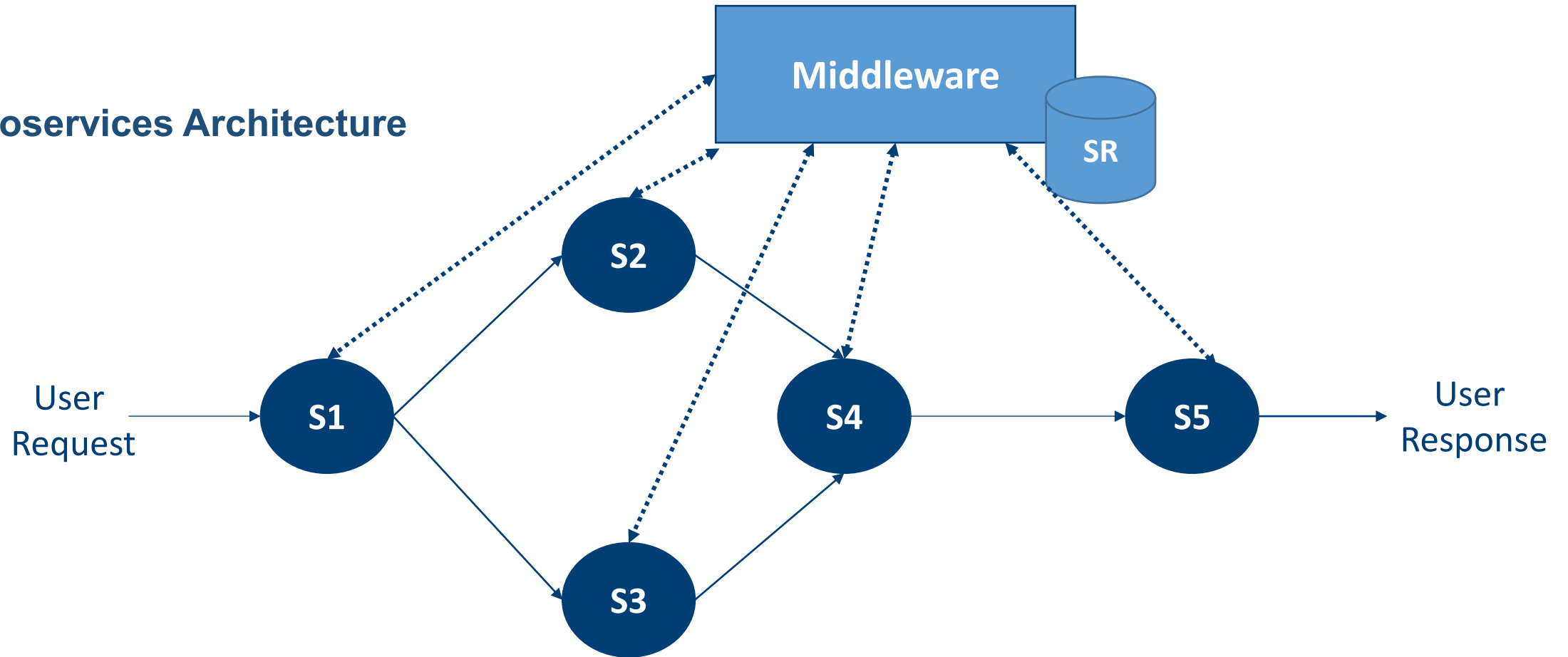


Operations are first!



AutoAI Programme

- Microservices Architecture

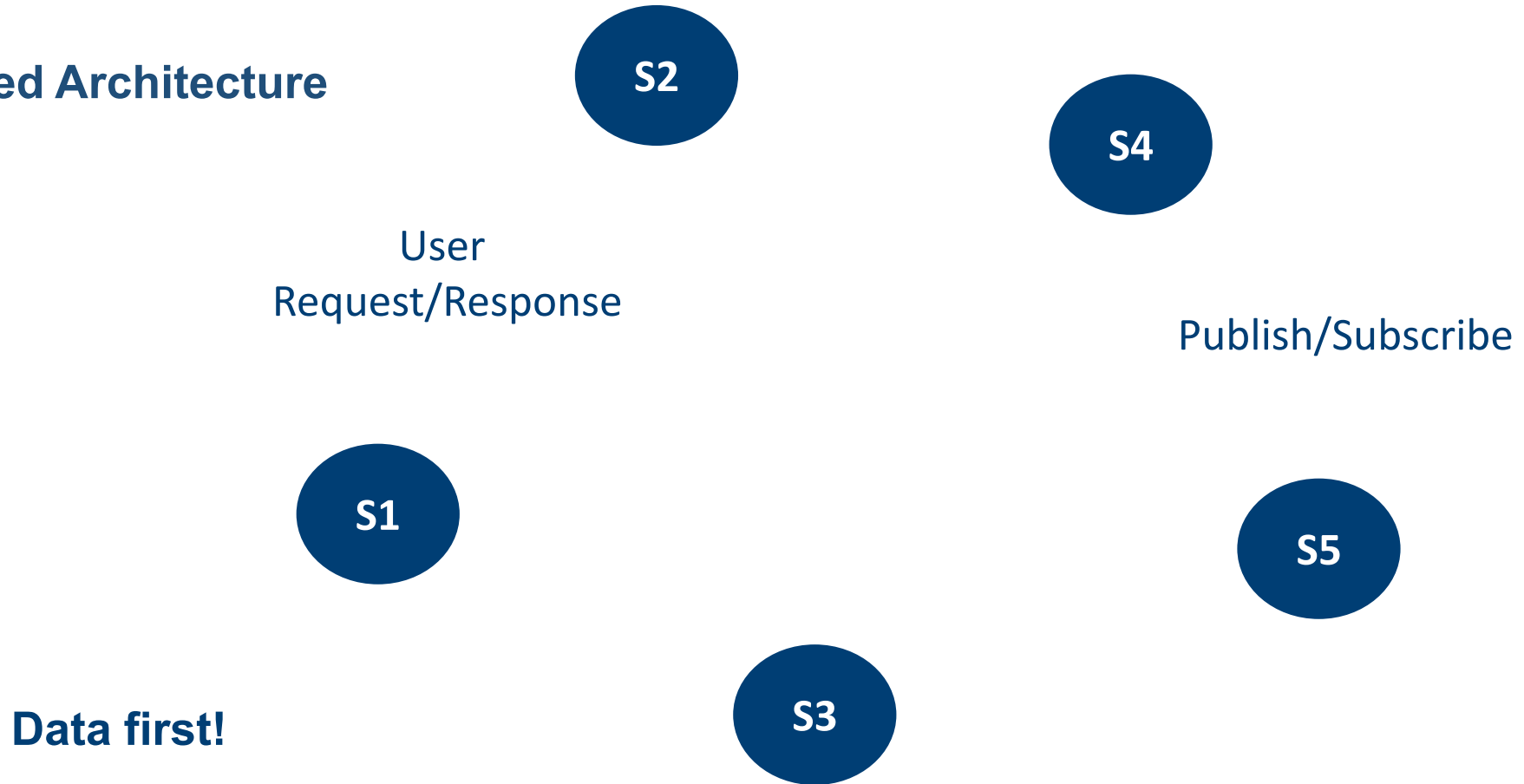


Operations are first!



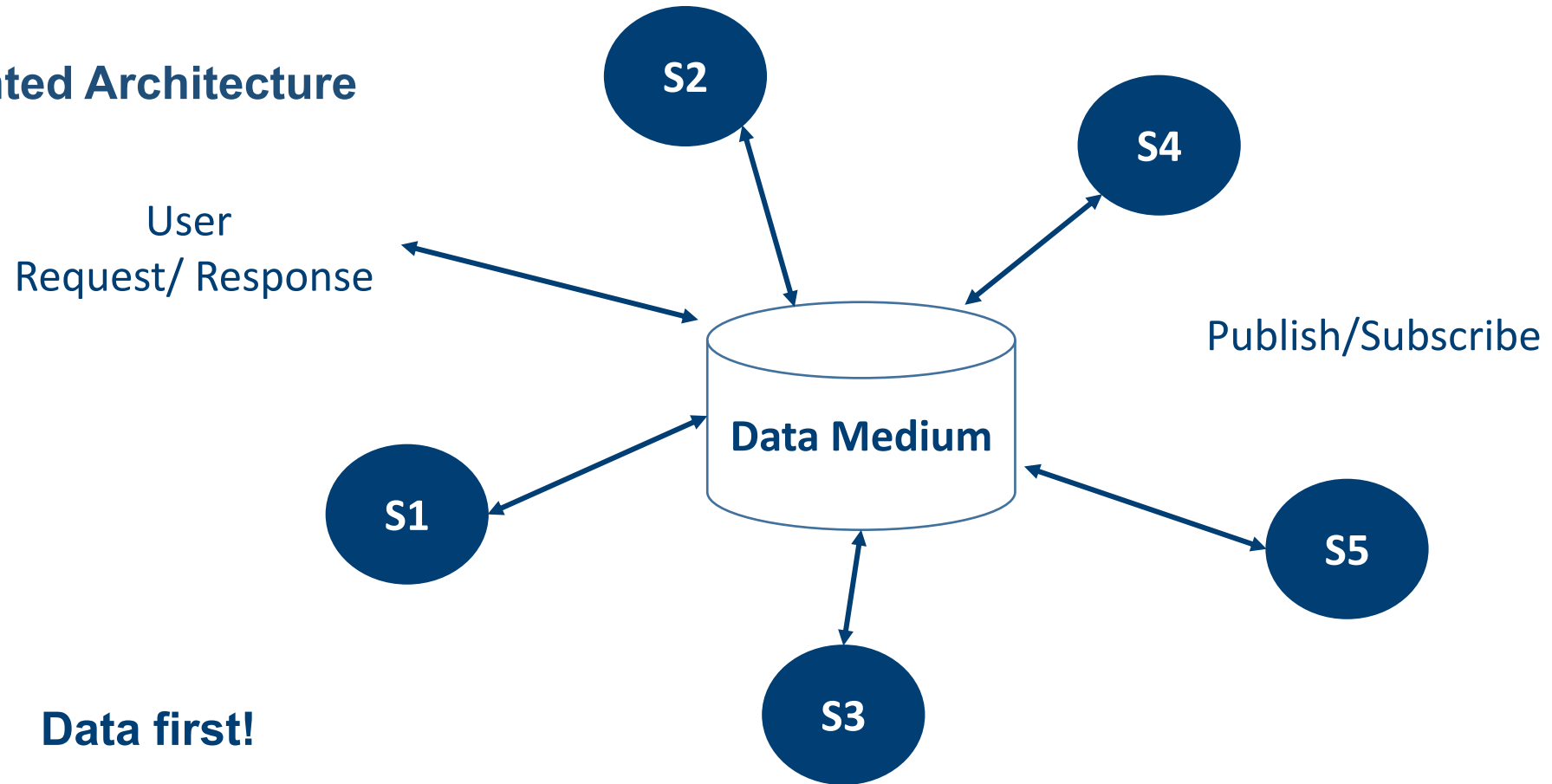
AutoAI Programme

- Data Oriented Architecture



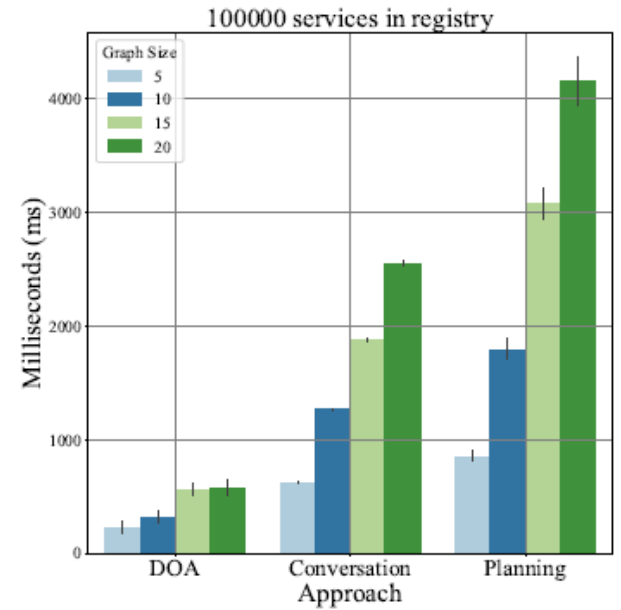
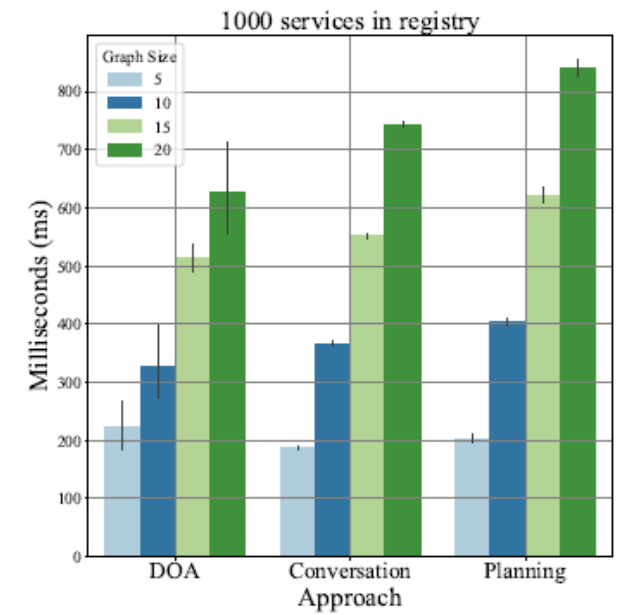
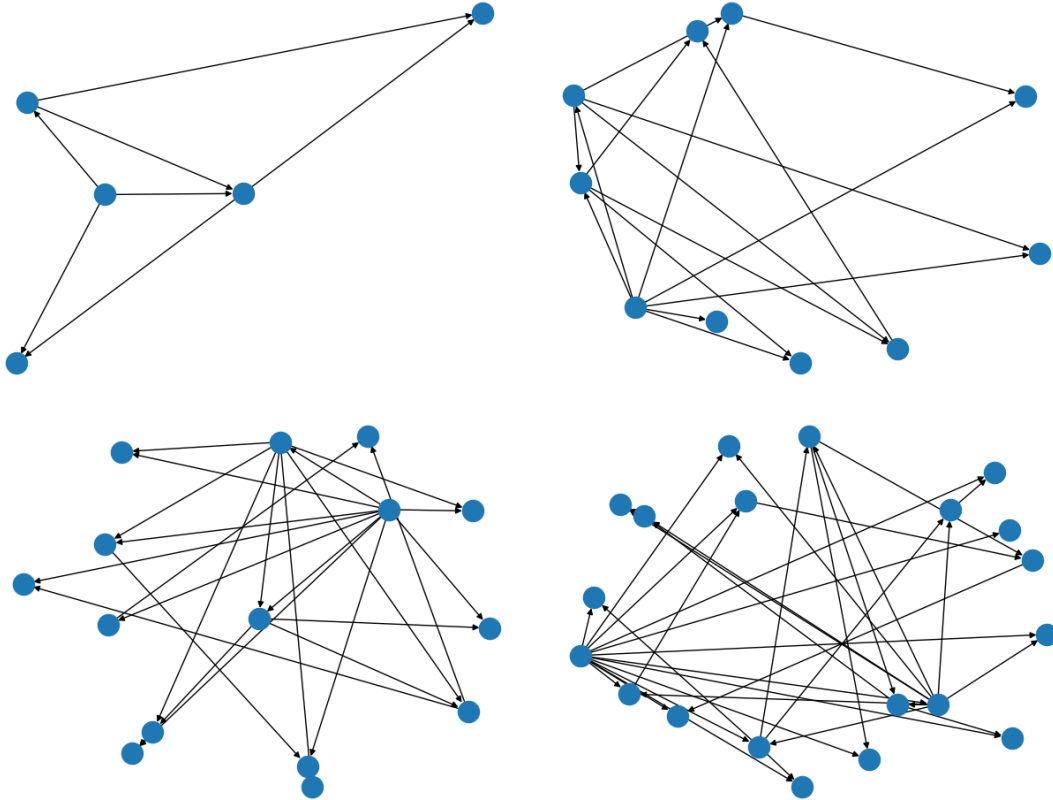
AutoAI Programme

- Data Oriented Architecture



AutoAI Programme

- Data Oriented Architecture



Take aways...

- It is very likely most of the AI popular predictions will not become real.
- AI has potential but we must be careful with the overwhelmed optimism.
- We must approach AI progress in a scientific way.
- **Problems must be always first!**
- The deployment of AI-based systems in real-world environments is hard. (**Design Decisions**)
- New paradigms are needed to realize the potential of AI. (**Data Oriented Architectures**)

Thank you!

chc79@cam.ac.uk

