UNIVERSITY OF CAMBRIDGE

Real-world AI-based Systems

Christian Cabrera Jojoa
Research Associate
chc.79@cam.ac.uk

Department of Computer Science and Technology

Agenda

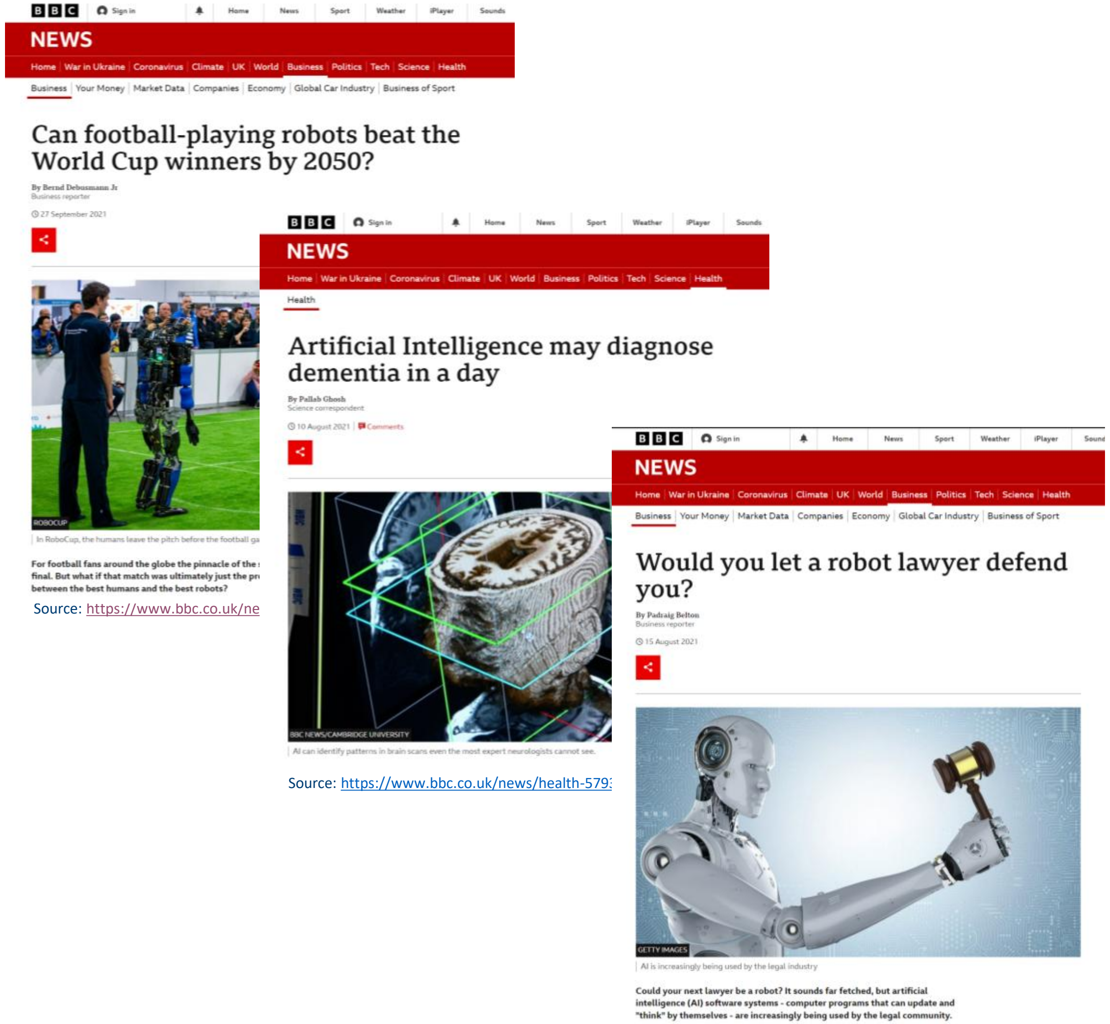
PART I: The great AI question.

PART II: Why the question?

PART III: Possible answers…

PART I: The great AI question.

AI in the news…


BBC
Sign in
Home
News
Sport
Weather
iPlayer
Sounds
NEWS
Home | War in Ukraine | Coronavirus | Climate | UK | World | Business | Politics | Tech | Science | Health
Business | Your Money | Market Data | Companies | Economy | Global Car Industry | Business of Sport
ROBOCUP

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

By Bernd Debusmann Jr
Business reporter

27 September 2021

In RoboCup, the humans leave the pitch before the football ga

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

Source: https://www.bbc.co.uk/ne

BBC
Sign in
Home
News
Sport
Weather
iPlayer
Sounds
NEWS
Home | War in Ukraine | Coronavirus | Climate | UK | World | Business | Politics | Tech | Science | Health
Health
Comments
BBC NEWS/CAMBRIDGE UNIVERSITY

Artificial Intelligence may diagnose dementia in a day

By Pallab 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-5793

BBC
Sign in
Home
News
Sport
Weather
iPlayer
Sound
NEWS
Home | War in Ukraine | Coronavirus | Climate | UK | World | Business | Politics | Tech | Science | Health
Business | Your Money | Market Data | Companies | Economy | Global Car Industry | Business of Sport
GETTY IMAGES

Would you let a robot lawyer defend you?

By Padraig Belton
Business reporter

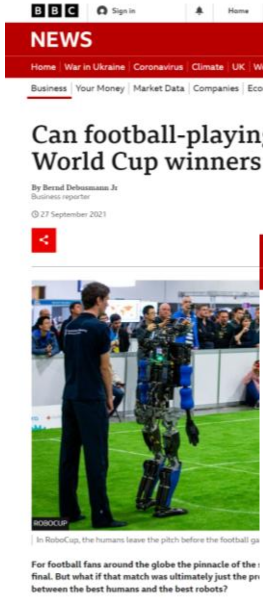
15 August 2021

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

AI in the news…


BBC
Sign in
Home
News
Sport
Weather
iPlayer
Sounds
NEWS
Home | War in Ukraine | Coronavirus | Climate | UK | World | Business | Politics | Tech | Science | Health
Business | Your Money | Market Data | Companies | Economy | Global Car Industry | Business of Sport
ROBOCUP

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

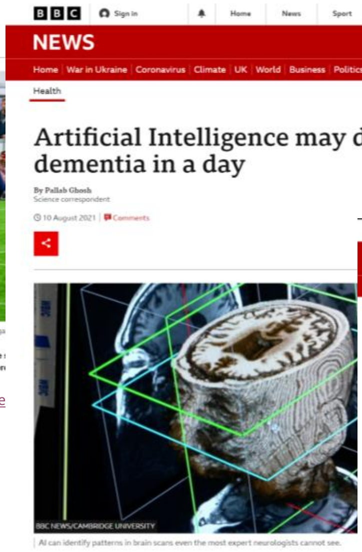
By Bernd Debusmann Jr
Business reporter

27 September 2021

In RoboCup, the humans leave the pitch before the football ga

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

Source: https://www.bbc.co.uk/ne


BBC
Sign in
Home
News
Sport
Weather
iPlayer
Sounds
NEWS
Home | War in Ukraine | Coronavirus | Climate | UK | World | Business | Politics | Tech | Science | Health
Health
BBC NEWS/CAMBRIDGE UNIVERSITY

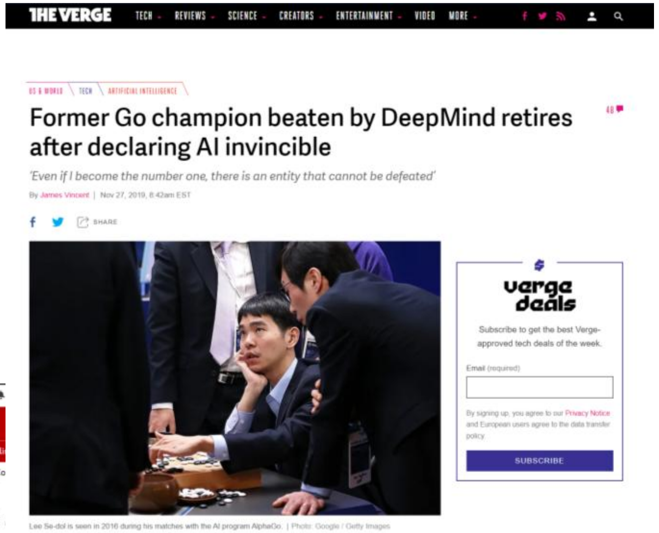
Artificial Intelligence may diagnose dementia in a day

By Pallab Ghosh
Science correspondent

10 August 2021 | Comments

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

Source: https://www.bbc.co.uk/news/health-579


THE VERGE
TECH
REVIEWS
SCIENCE
CREATORS
ENTERTAINMENT
VIDEO
MORE
US & WORLD \ TECH \ ARTIFICIAL INTELLIGENCE
48
SHARE
verge deals
Subscribe to get the best Verge-approved tech deals of the week.
Email (required)
By signing up, you agree to our Privacy Notice and European users agree to the data transfer policy.
SUBSCRIBE

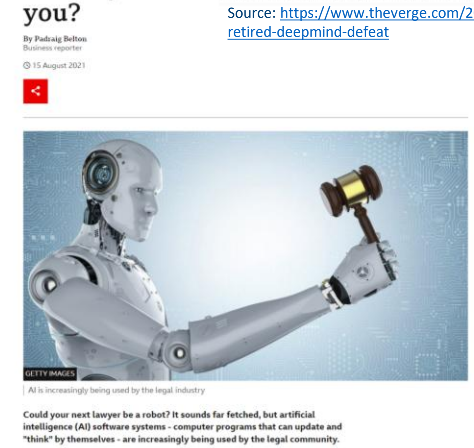
Former Go champion beaten by DeepMind retires after declaring AI invincible

'Even if I become the number one, there is an entity that cannot be defeated'

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

Lee Se-dol is seen in 2016 during his matches with the AI program AlphaGo. | Photo: Google / Getty Images

Source: https://www.theverge.com/2019/11/27/20985260/ai-go-alphago-lee-se-dol-retired-deepmind-defeat


BBC
Sign in
NEWS
Home | War in Ukraine | Coronavirus | Cli
Business | Your Money | Market Data | Co
GETTY IMAGES

Would you let you?

By Padraig Belton
Business reporter

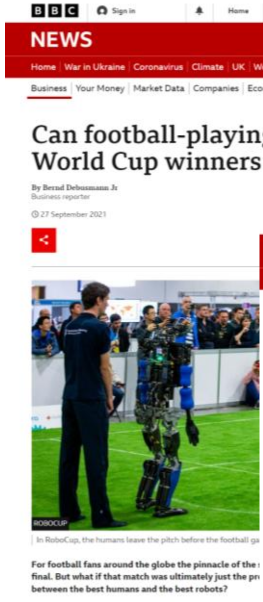
15 August 2021

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

AI in the news…



BBC NEWS

Home | War in Ukraine | Coronavirus | Climate | UK | World | Business | Politics | Tech | Science | Health

Business | Your Money | Market Data | Companies | Economy | Global Car Industry | Business of Sport

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

By Bernd Debusmann Jr
Business reporter

27 September 2021

ROBOCUP

In RoboCup, the humans leave the pitch before the football ga

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

Source: https://www.bbc.co.uk/ne

BBC NEWS

Home | War in Ukraine | Coronavirus | Climate | UK | World | Business | Politics | Tech | Science | Health

Health

Artificial Intelligence may diagnose dementia in a day

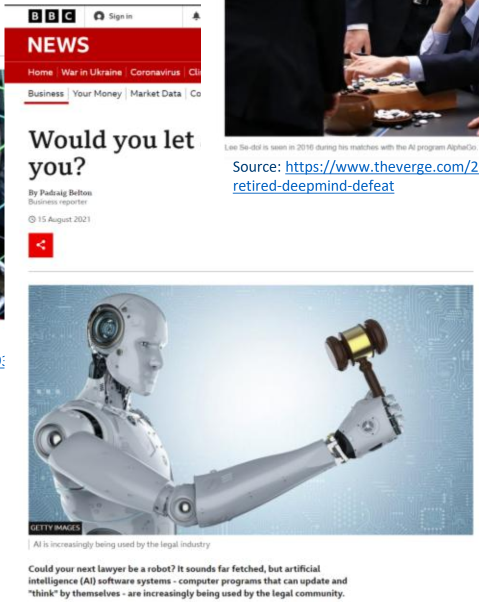
By Pallab Ghosh
Science correspondent

10 August 2021 | Comments

BBC NEWS/CAMBRIDGE UNIVERSITY

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

Source: https://www.bbc.co.uk/news/health-579



BBC NEWS

Home | War in Ukraine | Coronavirus | Cl

Business | Your Money | Market Data | Co

Would you let you?

By Padraig Belton
Business reporter

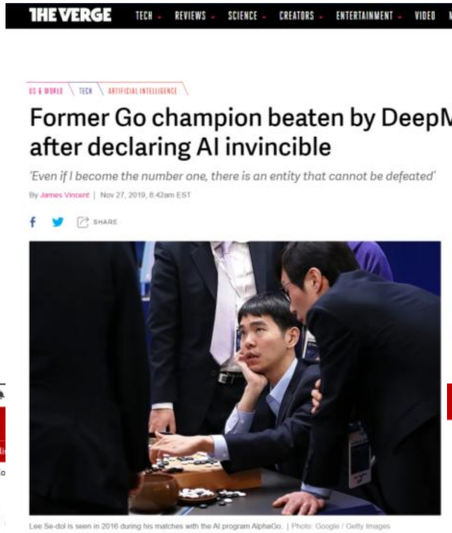
15 August 2021

GETTY IMAGES

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



THE VERGE — TECH · REVIEWS · SCIENCE · CREATORS · ENTERTAINMENT · VIDEO · MORE

US & WORLD \ TECH \ ARTIFICIAL INTELLIGENCE

Former Go champion beaten by DeepMind retires after declaring AI invincible

'Even if I become the number one, there is an entity that cannot be defeated'

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

SHARE

verge deals

Subscribe to get the best Verge-approved tech deals of the week.

Email (required)

Lee Se-dol is seen in 2016 during his matches with the AI program AlphaGo. | Photo: Google / Getty Images

Source: https://www.theverge.com/2019/11/27/20985260/ai-g retired-deepmind-defeat

CARACOL RADIO — PROGRAMAS · PODCAST · CIUDADES · DEPORTES · HABLE CON LOS PROGRAMAS

6AM HOY POR HOY

A FONDO

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

La inteligencia artificial lo cambiará todo

Source: https://caracol.com.co/programa/2019/03/11/6am_hoy_por_hoy/1552309220_833279.html

CARACOL RADIO — PROGRAMAS · PODCAST · CIUDADES · DEPORTES · HABLE CON LOS PROGRAMA

TECNOLOGÍA

Patentan botón para apaga inteligencia artificial en caso rebelión

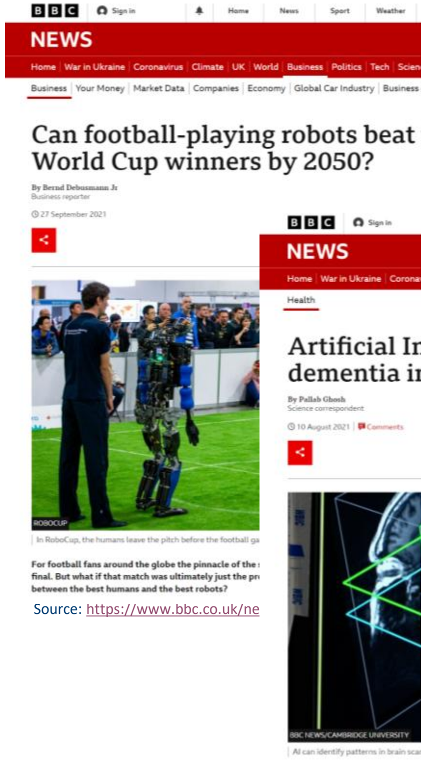
¿Qué sucedería si las computadoras y los robots se rebelaran contra los seres humanos? Google creó una posible solución a las hipótesis de la ciencia ficción.

BBC/GETTY IMAGES

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

Sounds promising!!!

But…


BBC
Sign in
Home
News
Sport
Weather
NEWS
Home | War in Ukraine | Coronavirus | Climate | UK | World | Business | Politics | Tech | Scien
Business | Your Money | Market Data | Companies | Economy | Global Car Industry | Business

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

By Bernd Debusmann Jr
Business reporter

27 September 2021

In RoboCup, the humans leave the pitch before the football ga

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

Source: https://www.bbc.co.uk/ne

BBC
Sign in
NEWS
Home | War in Ukraine | Corona
Health

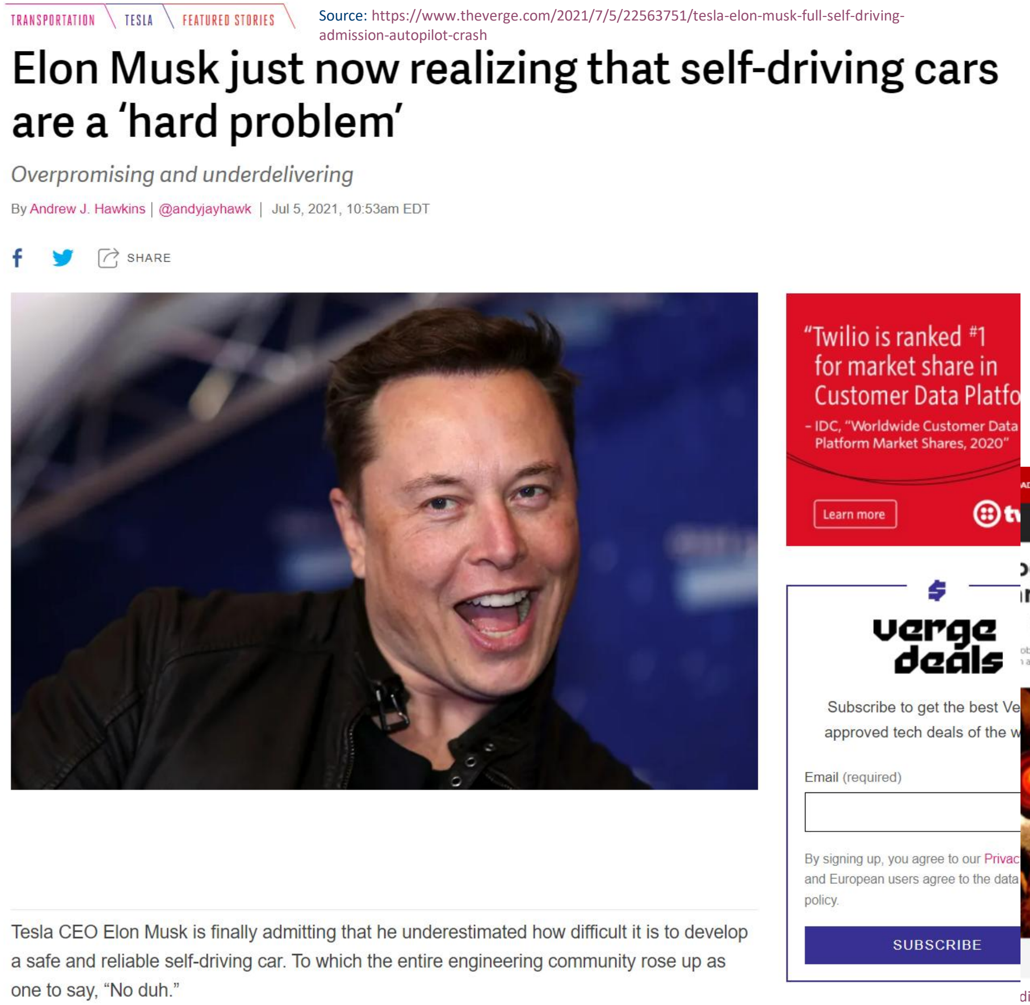
Artificial Ir
dementia ir

By Pallab Ghosh
Science correspondent

10 August 2021 | Comments

AI can identify patterns in brain scar

Source: https://www



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

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."

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/2019/03/11/6am_hoy_por_hoy/1552309220_833279.html

botón para apaga
rtificial en caso
rebelión

obots se rebelaran contra los seres humanos? Google creó una posible
a las hipótesis de la ciencia ficción.

dio/2016/06/15/tecnologia/1466026084_693727.html

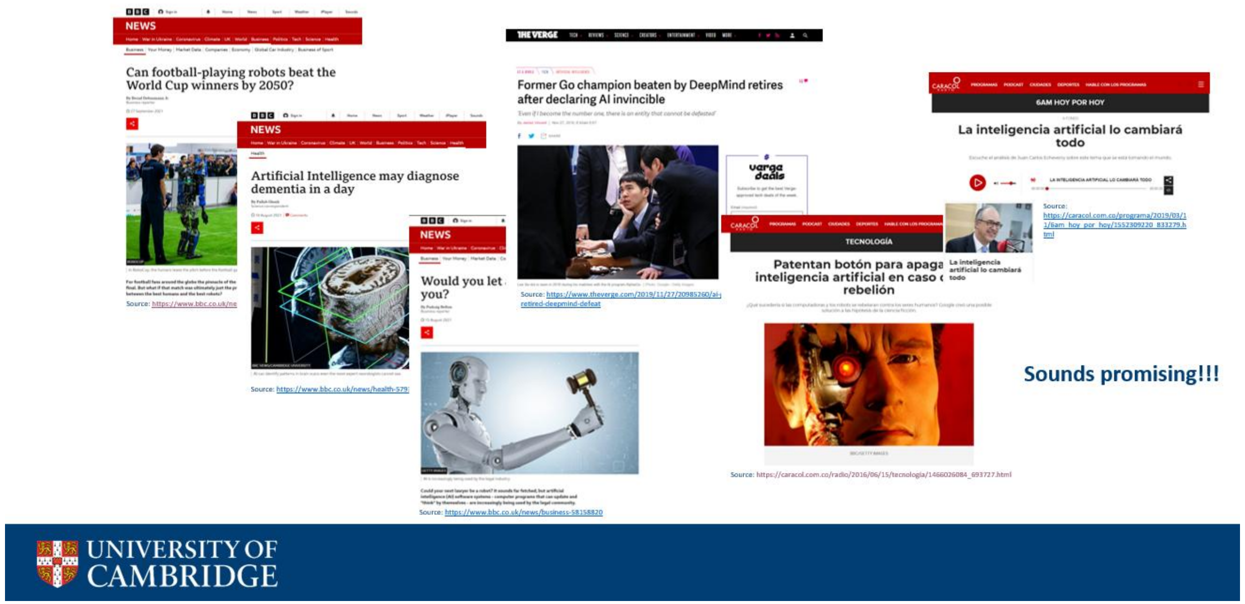
Could your next lawyer be a robot? It sounds far fetched, but artificial intelligence (AI) software systems - computer programs that can update and "think" by themselves - are increasingly being used by the legal community.

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

It is far from reality…

The Great AI Fallacy

AI in the news…



BBC
NEWS
Home | War in Ukraine | Coronavirus | Climate | UK | World | Business | Politics | Tech | Science | Health
THE VERGE
verge deals
CARACOL
PROGRAMAS PODCAST CIUDADES DEPORTES HABLE CON LOS PROGRAMAS
6AM HOY POR HOY
TECNOLOGÍA
LA INTELIGENCIA ARTIFICIAL LO CAMBIARÁ TODO
UNIVERSITY OF CAMBRIDGE

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

Source: https://www.bbc.co.uk/ne

Artificial Intelligence may diagnose dementia in a day

Source: https://www.bbc.co.uk/news/health-579

Would you let … you?

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 an entity that cannot be defeated'

Source: https://www.theverge.com/2019/11/27/20985260/ai-retired-deepmind-defeat

La inteligencia artificial lo cambiará todo

Source: https://caracol.com.co/programa/2019/03/11/6am_hoy_por_hoy/1552309220_833279.html

Patentan botón para apaga… inteligencia artificial en caso … rebelión

La inteligencia artificial lo cambiará todo

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

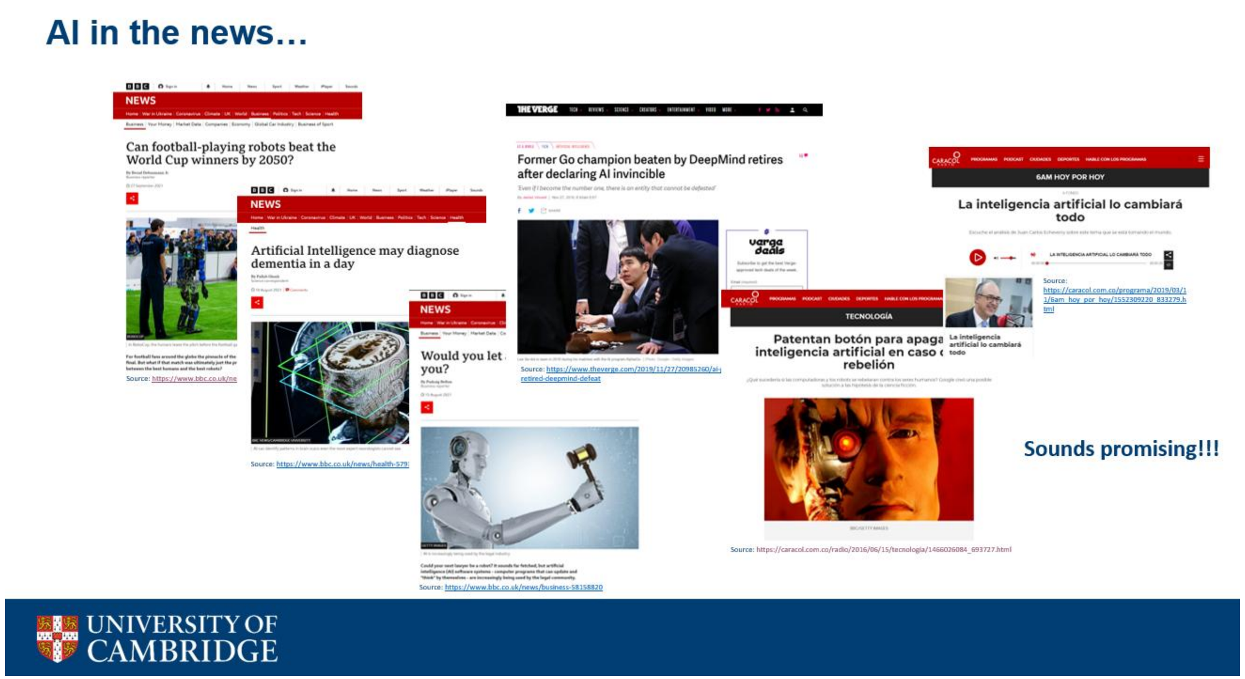
Sounds promising!!!

"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

The Great AI Question



AI in the news…

BBC NEWS

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

Source: https://www.bbc.co.uk/ne

BBC NEWS

Artificial Intelligence may diagnose dementia in a day

Source: https://www.bbc.co.uk/news/health-579

BBC NEWS

Would you let you?

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

THE VERGE

Former Go champion beaten by DeepMind retires after declaring AI invincible

'Even if I become the number one, there is an entity that cannot be defeated'

Source: https://www.theverge.com/2019/11/27/20985260/ai-retired-deepmind-defeat

verge deals

CARACOL RADIO

TECNOLOGÍA

Patentan botón para apaga inteligencia artificial en caso rebelión

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

CARACOL RADIO

6AM HOY POR HOY

La inteligencia artificial lo cambiará todo

LA INTELIGENCIA ARTIFICIAL LO CAMBIARÁ TODO

La inteligencia artificial lo cambiará todo

Source: https://caracol.com.co/programa/2019/03/11/6am_hoy_por_hoy/1552309220_833279.html

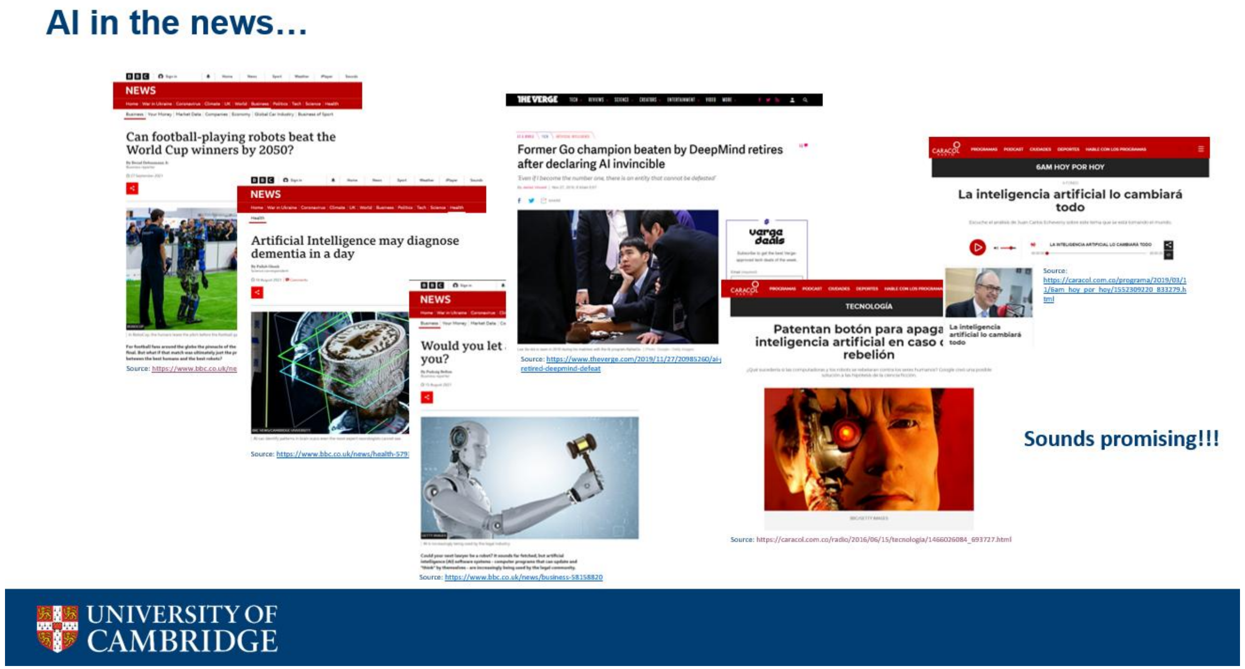
Sounds promising!!!

UNIVERSITY OF CAMBRIDGE

?

UNIVERSITY OF CAMBRIDGE

The Great AI Question



AI in the news…

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

Source: https://www.bbc.co.uk/ne

Artificial Intelligence may diagnose dementia in a day

Source: https://www.bbc.co.uk/news/health-579

Would you let you?

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 an entity that cannot be defeated'

Source: https://www.theverge.com/2019/11/27/20985260/ai-retired-deepmind-defeat

6AM HOY POR HOY

La inteligencia artificial lo cambiará todo

Source: https://caracol.com.co/programa/2019/03/11/6am_hoy_por_hoy/1552309220_833279.html

TECNOLOGÍA

Patentan botón para apaga inteligencia artificial en caso rebelión

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

Sounds promising!!!

UNIVERSITY OF CAMBRIDGE

- **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.

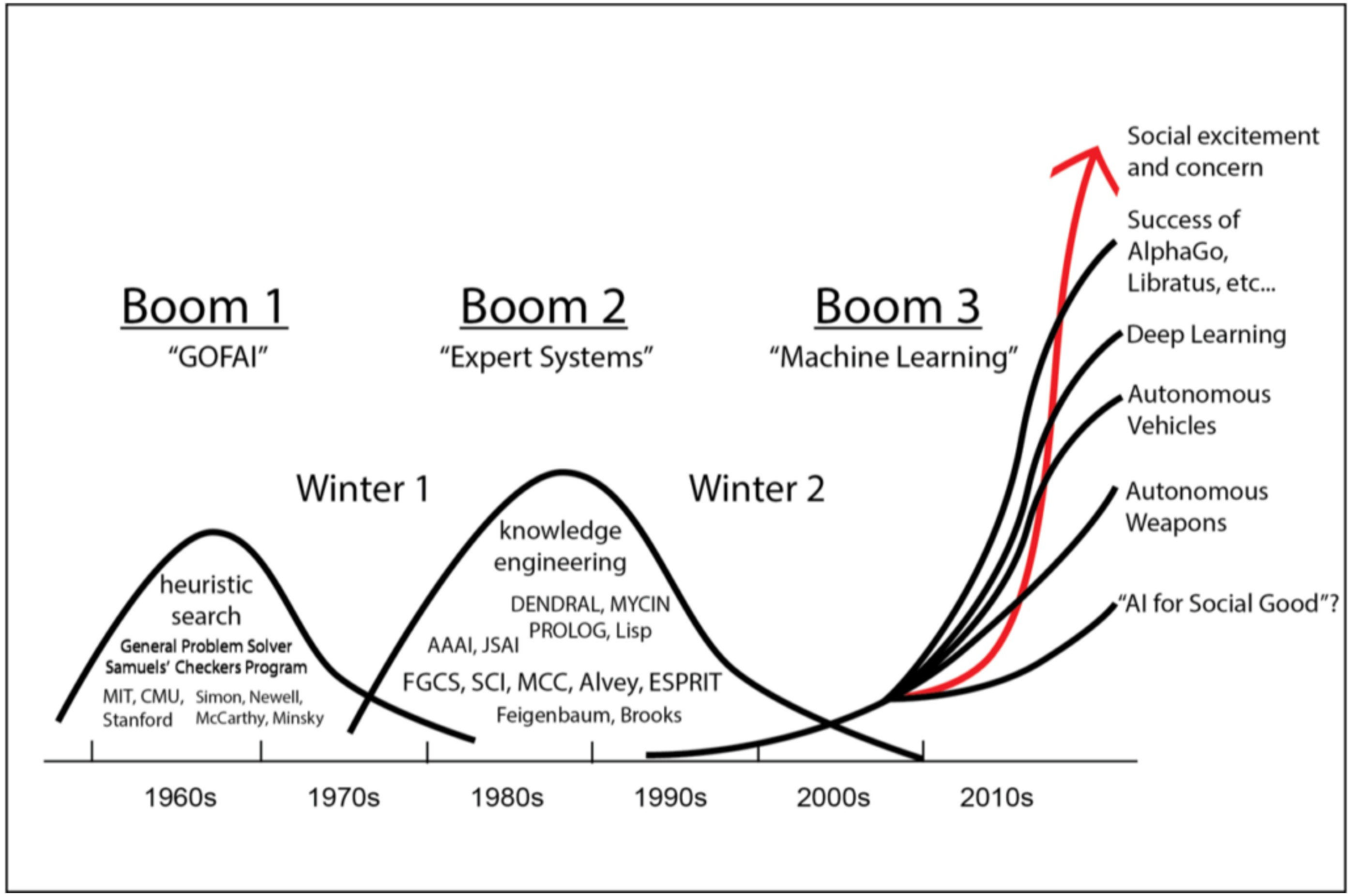
- **Overwhelmed optimism!**

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?

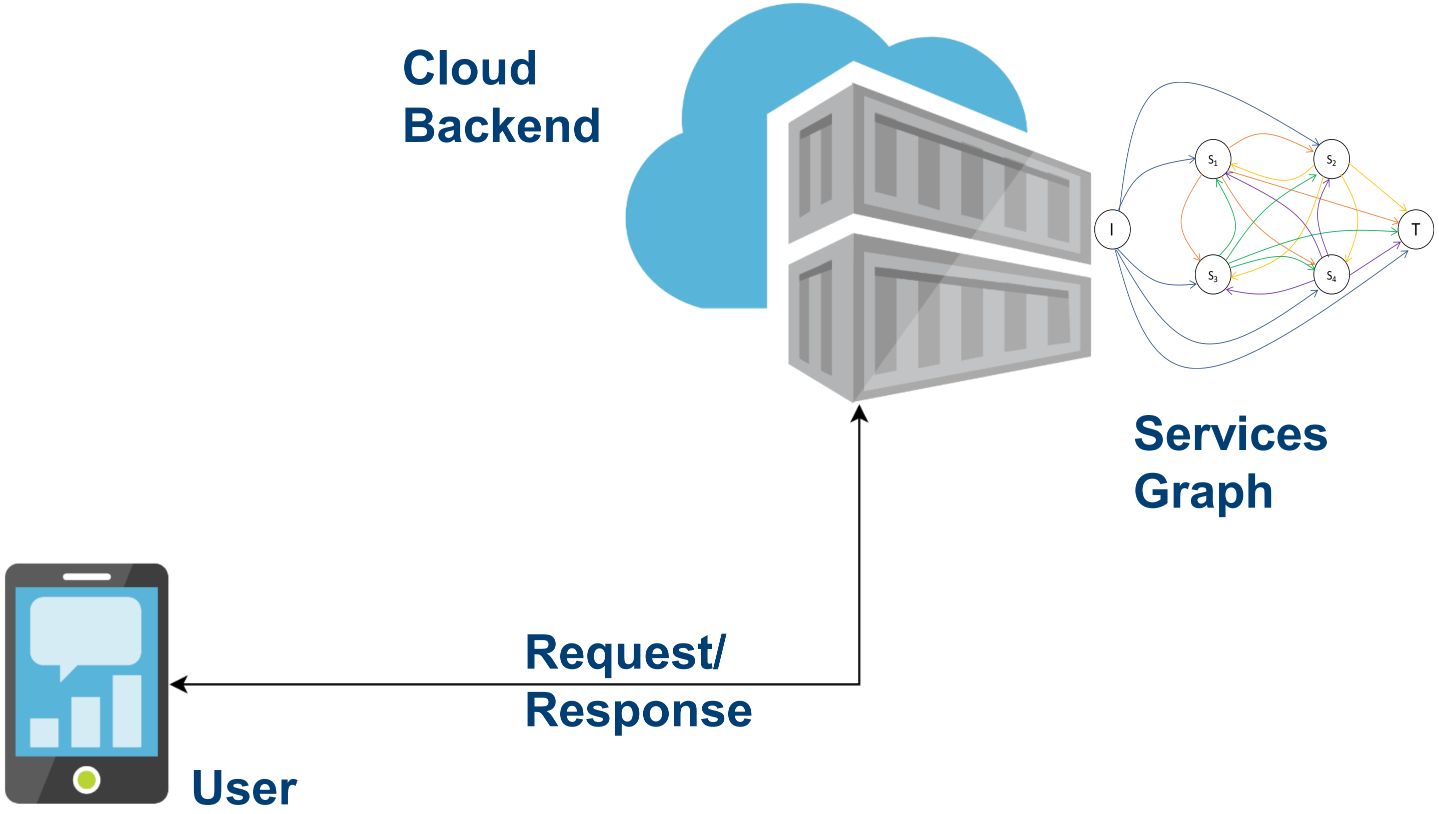
AI Application - Dynamic Service Placement in Edge Computing

- 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].**



Cloud Backend

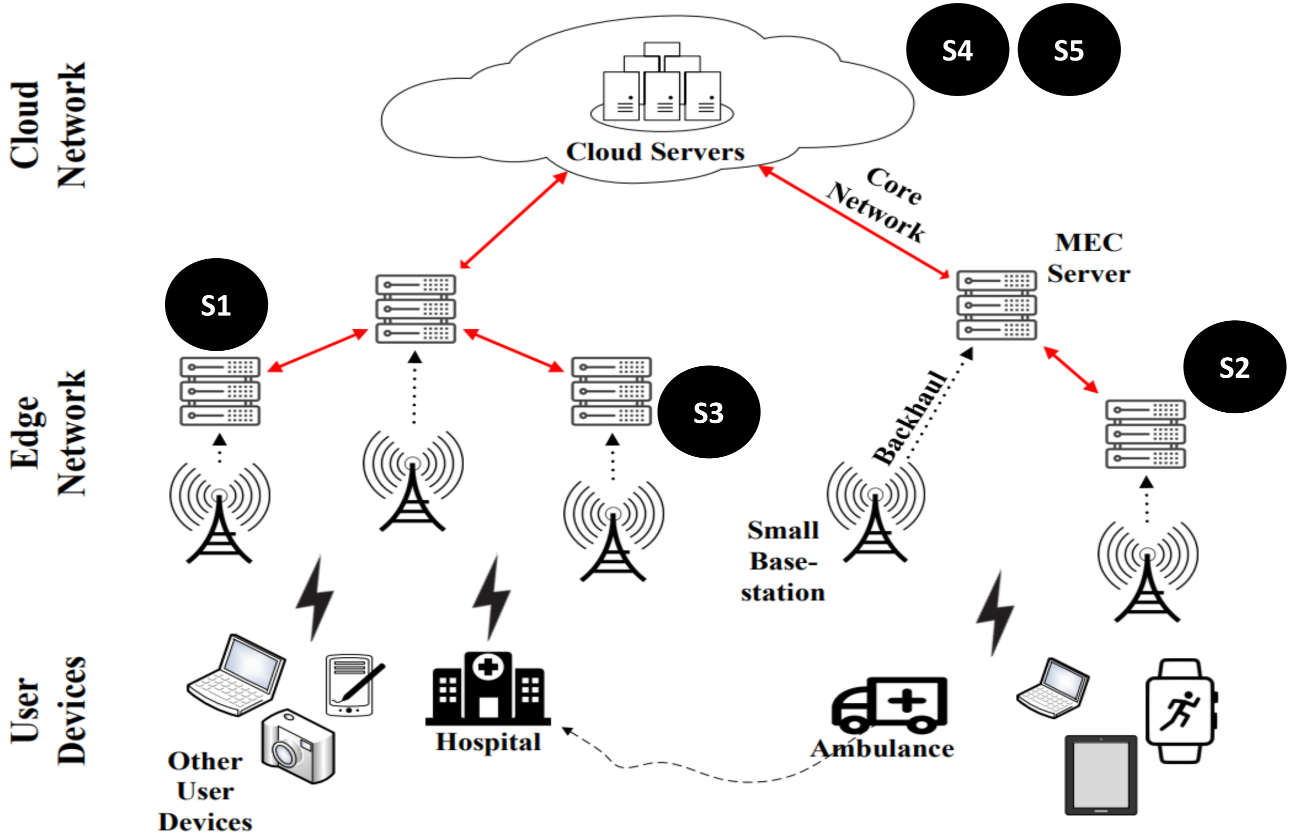
Services Graph

Request/ Response

User

[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

$$\min_{j} \quad wt_j + lat(D_j) \tag{1}$$

$$\min_{i} \sqrt{w_p \sigma(RR^{CPU})^2 + w_m \sigma(RR^{RAM})^2} \tag{2}$$

$$\text{subject to:} \quad \sum_{k=1}^{n} rr_{ik}^{CPU} <= r_i^{CPU} \tag{3}$$

$$\sum_{k=1}^{n} rr_{ik}^{RAM} <= r_i^{RAM} \tag{4}$$

[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

$$\min_{j} \quad wt_j + lat(D_j) \qquad (1)$$

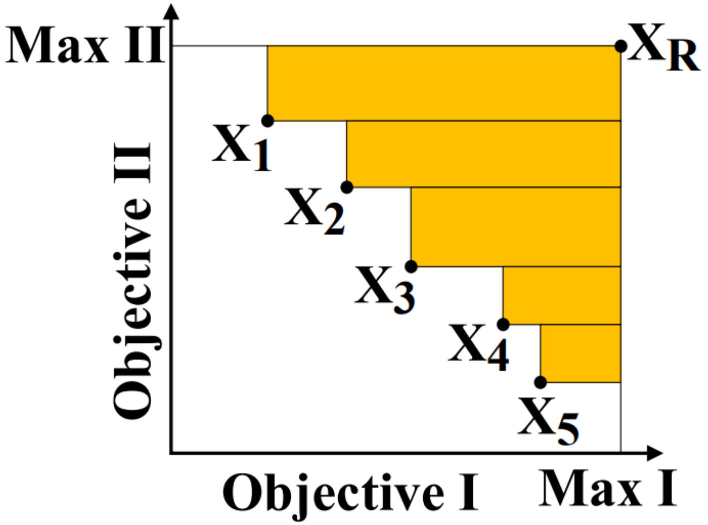
$$\min_{i} \sqrt{w_p \sigma(RR^{CPU})^2 + w_m \sigma(RR^{RAM})^2} \qquad (2)$$

subject to:

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

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

- Pareto-optimal[1]



[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

$$\min_{j} \quad wt_j + lat(D_j) \qquad (1)$$

$$\min_{i} \sqrt{w_p \sigma(RR^{CPU})^2 + w_m \sigma(RR^{RAM})^2} \qquad (2)$$

$$\text{subject to:} \quad \sum_{k=1}^{n} rr_{ik}^{CPU} <= r_i^{CPU} \qquad (3)$$

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

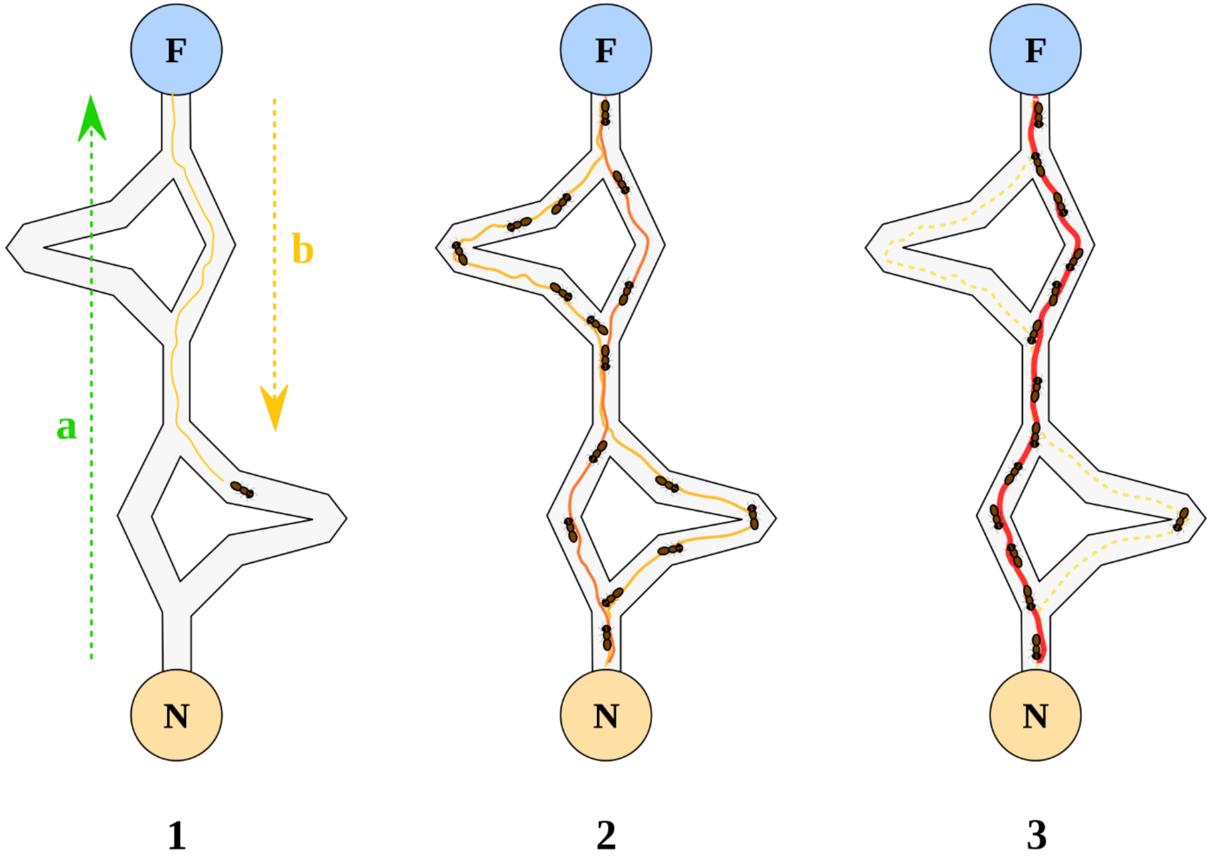
- Related work.
  - **Exact algorithms** take too long.
  - **Approximation** and **heuristic** algorithms have problems finding the optimal solution.
  - **Meta-heuristics** are promising solutions for this optimisation problem.
- We explored a **bio-inspired** meta-heuristic [2].
  - Ant colony optimization algorithm.
  - Inspired in how ants find the shortest path to food.
  - **Reinforcement learning.**

[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*.

[2] Cabrera C., Svorobej, S., Palade, A., Kazmi, A., & Clarke, S. (2022). MAACO: A Dynamic Service Placement Model for Smart Cities. *IEEE Transactions on Services Computing*.

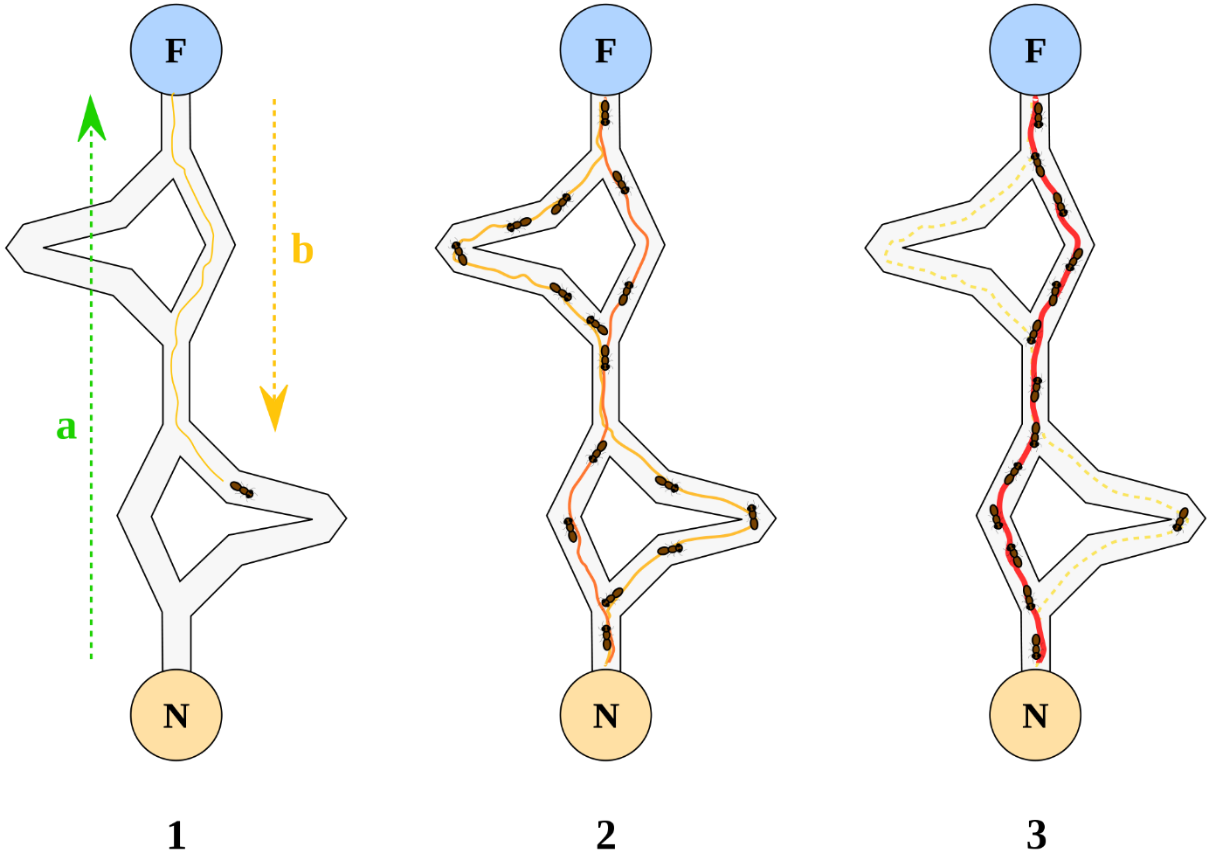
AI Application - Dynamic Service Placement in Edge Computing

- Ant-Colony Optimisation



AI Application - Dynamic Service Placement in Edge Computing

- Ant-Colony Optimisation



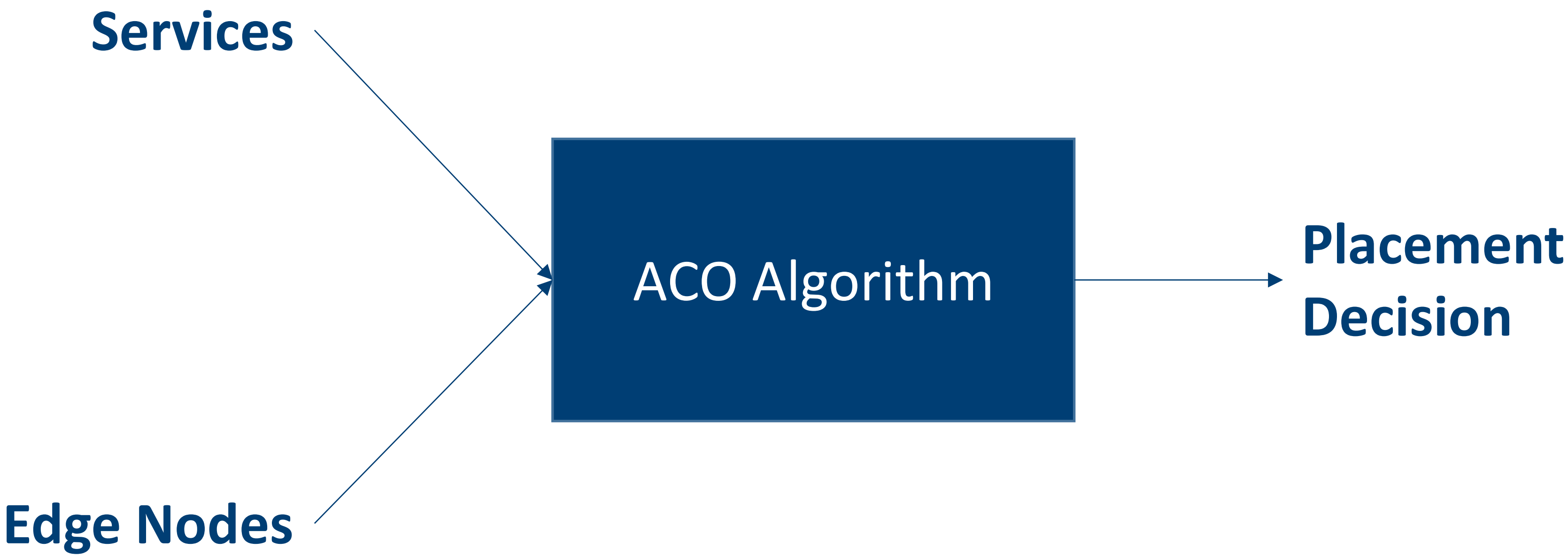
Forward movement:

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

Backward movement:

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

AI Application - Dynamic Service Placement in Edge Computing



Services

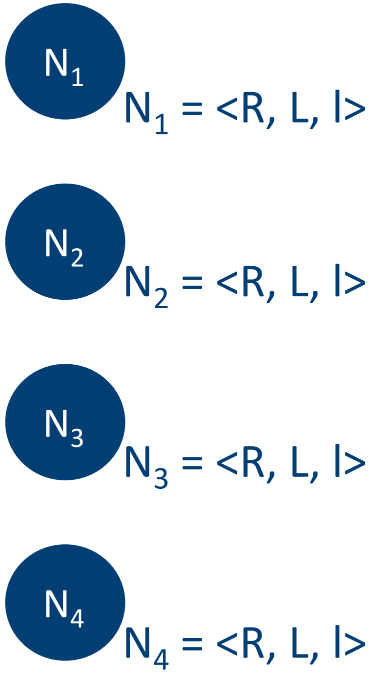
Edge Nodes

ACO Algorithm

Placement Decision

AI Application - Dynamic Service Placement in Edge Computing

App_1=<S, RQ, l, p>
App_2=<S, RQ, l, p>
…
App_n=<S, RQ, l, p>



N_1 = <R, L, l>

N_2 = <R, L, l>

N_3 = <R, L, l>

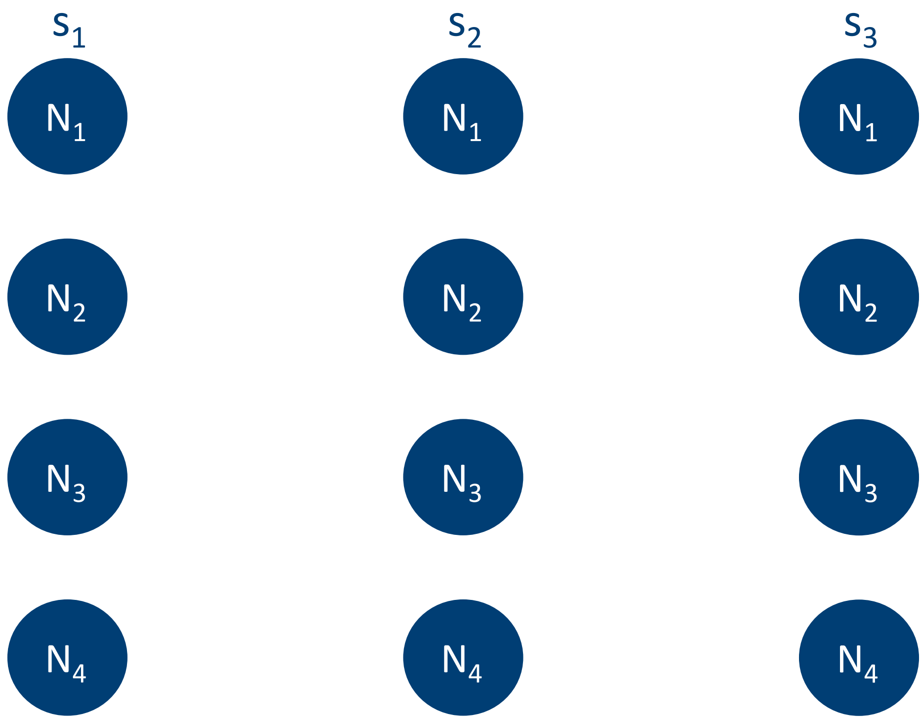
N_4 = <R, L, l>

AI Application - Dynamic Service Placement in Edge Computing

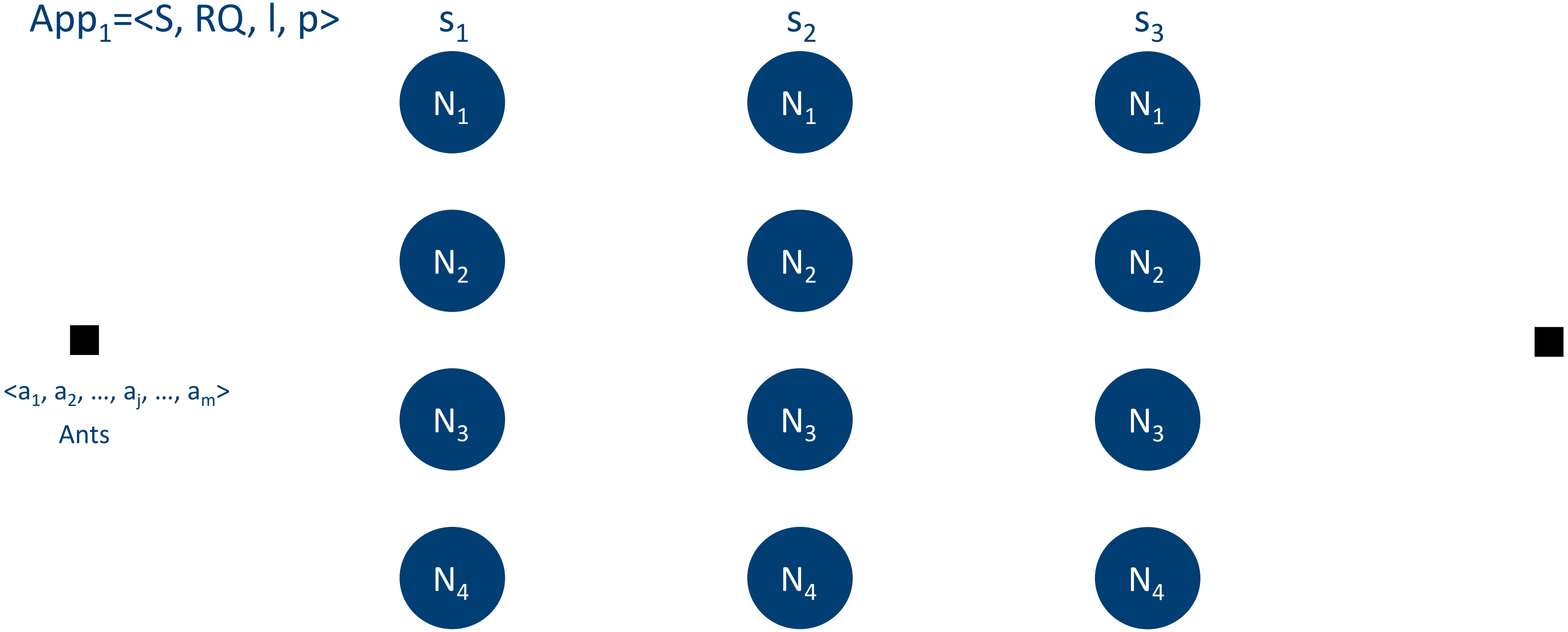
App$_1$=<S, RQ, l, p>

S=<s_1, s_2, s_3>

RQ=<R_{f1}, R_{f2}, Rf_{f3}>



AI Application - Dynamic Service Placement in Edge Computing



App_1=<S, RQ, l, p>

s_1 s_2 s_3

N_1 N_1 N_1

N_2 N_2 N_2

N_3 N_3 N_3

N_4 N_4 N_4

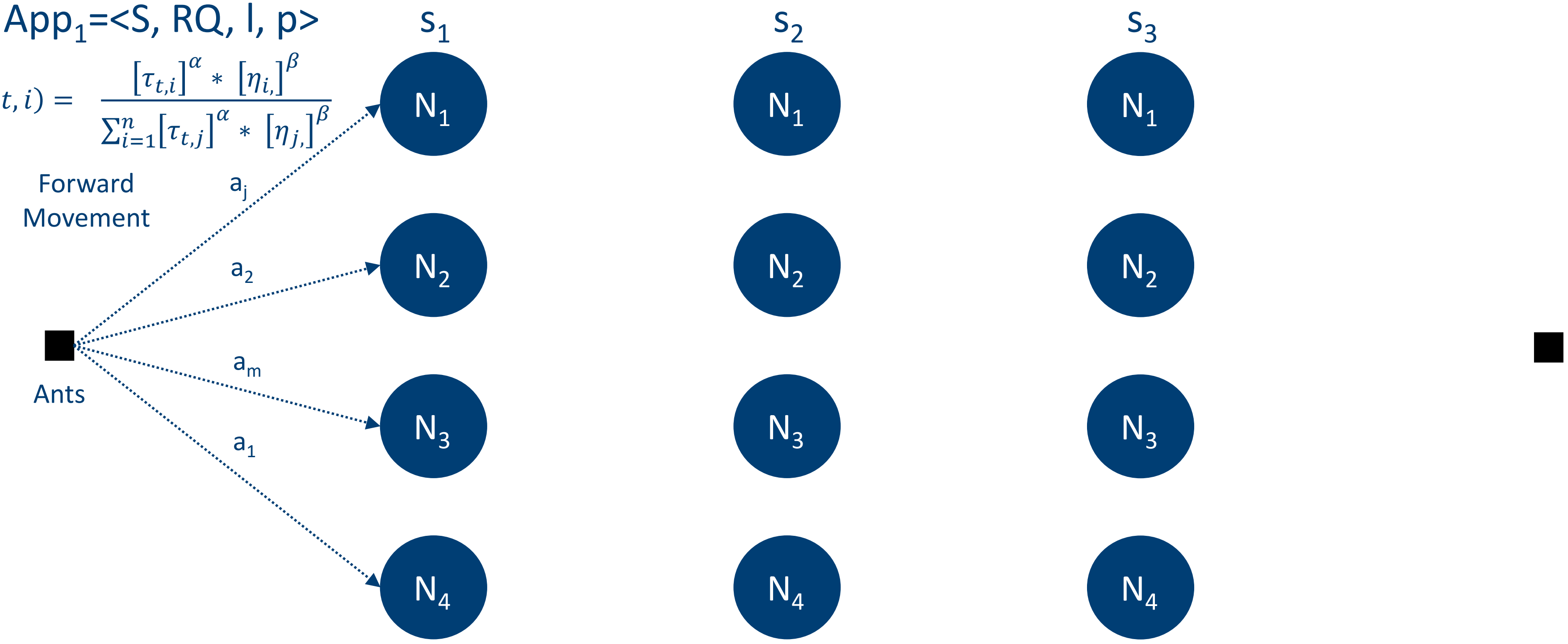
<a_1, a_2, …, a_j, …, a_m>

Ants

AI Application - Dynamic Service Placement in Edge Computing

App$_1$=<S, RQ, l, p>

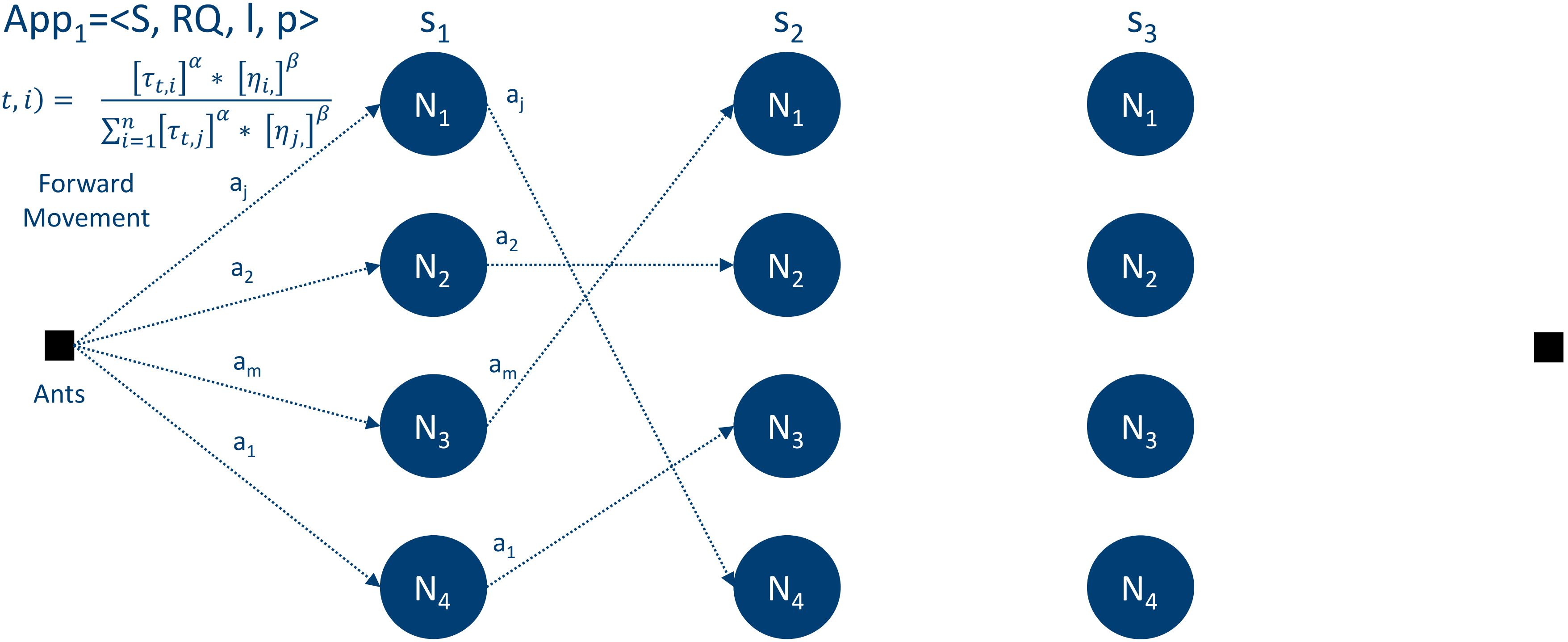
$$p(t,i) = \frac{[\tau_{t,i}]^{\alpha} * [\eta_{i,}]^{\beta}}{\sum_{i=1}^{n}[\tau_{t,j}]^{\alpha} * [\eta_{j,}]^{\beta}}$$



AI Application - Dynamic Service Placement in Edge Computing

App_1=<S, RQ, l, p>

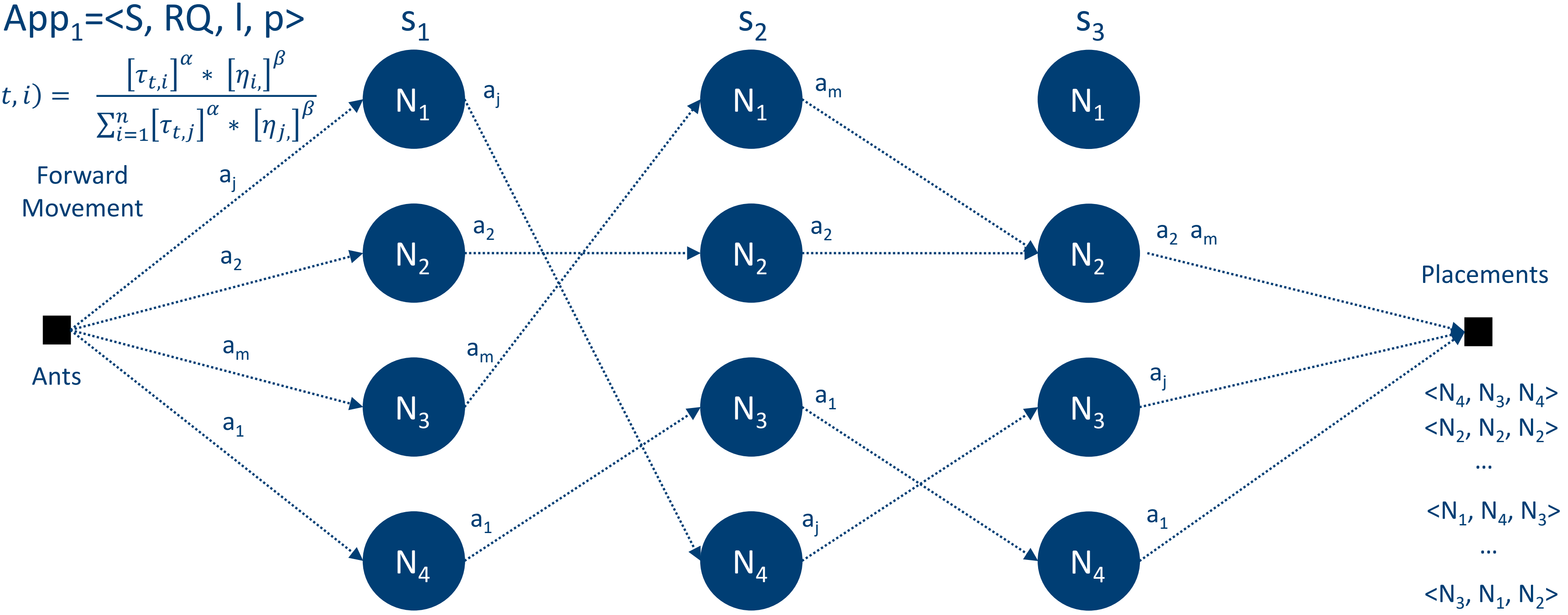
$$p(t,i) = \frac{[\tau_{t,i}]^{\alpha} * [\eta_{i,}]^{\beta}}{\sum_{i=1}^{n} [\tau_{t,j}]^{\alpha} * [\eta_{j,}]^{\beta}}$$



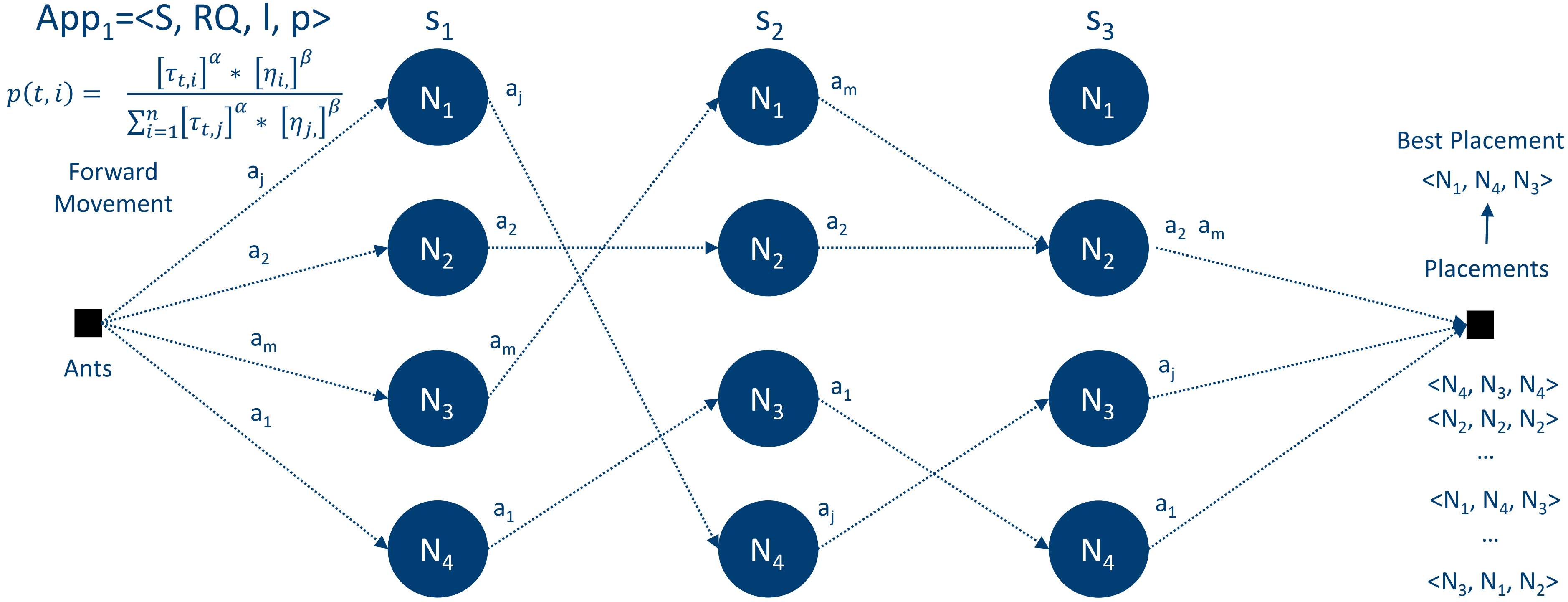
AI Application - Dynamic Service Placement in Edge Computing

App_1=<S, RQ, l, p>

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



AI Application - Dynamic Service Placement in Edge Computing



App$_1$=<S, RQ, l, p>

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

Best Placement

<N$_1$, N$_4$, N$_3$>

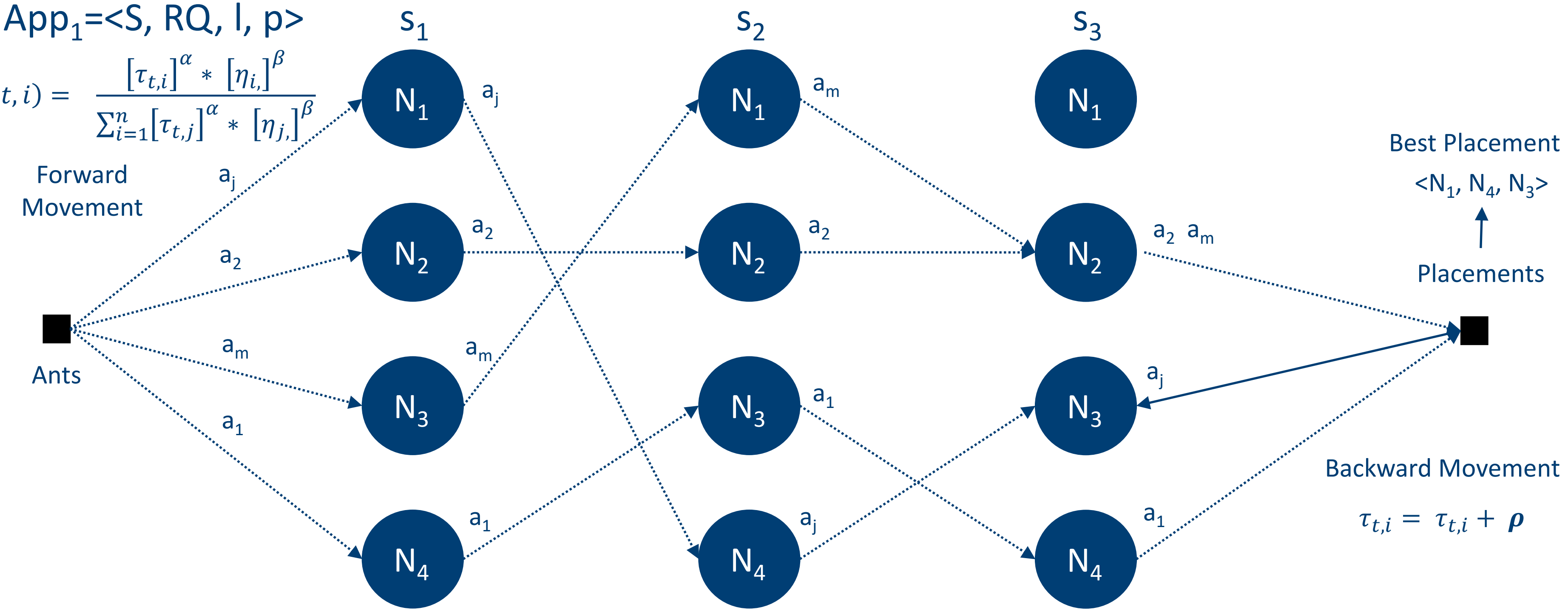
Placements

<N$_4$, N$_3$, N$_4$>
<N$_2$, N$_2$, N$_2$>
…
<N$_1$, N$_4$, N$_3$>
…
<N$_3$, N$_1$, N$_2$>

AI Application - Dynamic Service Placement in Edge Computing

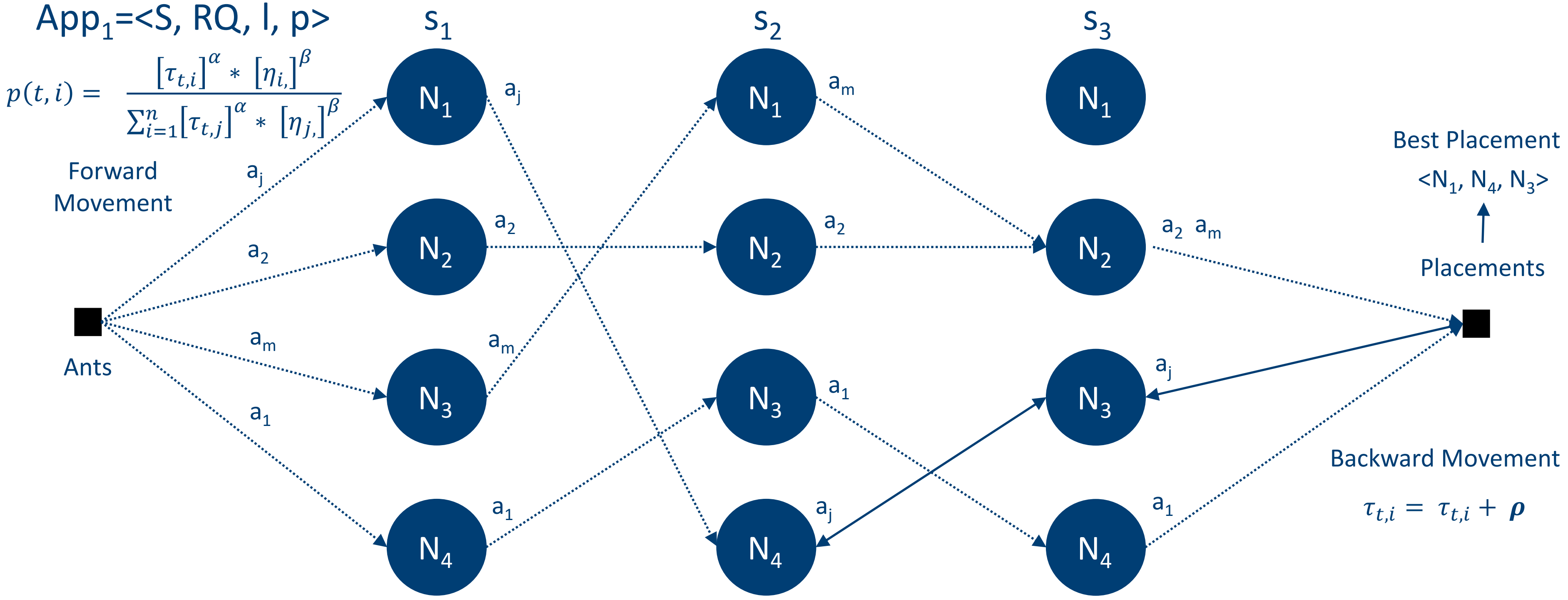
App$_1$=<S, RQ, l, p>

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



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

AI Application - Dynamic Service Placement in Edge Computing



App$_1$=<S, RQ, l, p>

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

Forward Movement

Ants

Best Placement

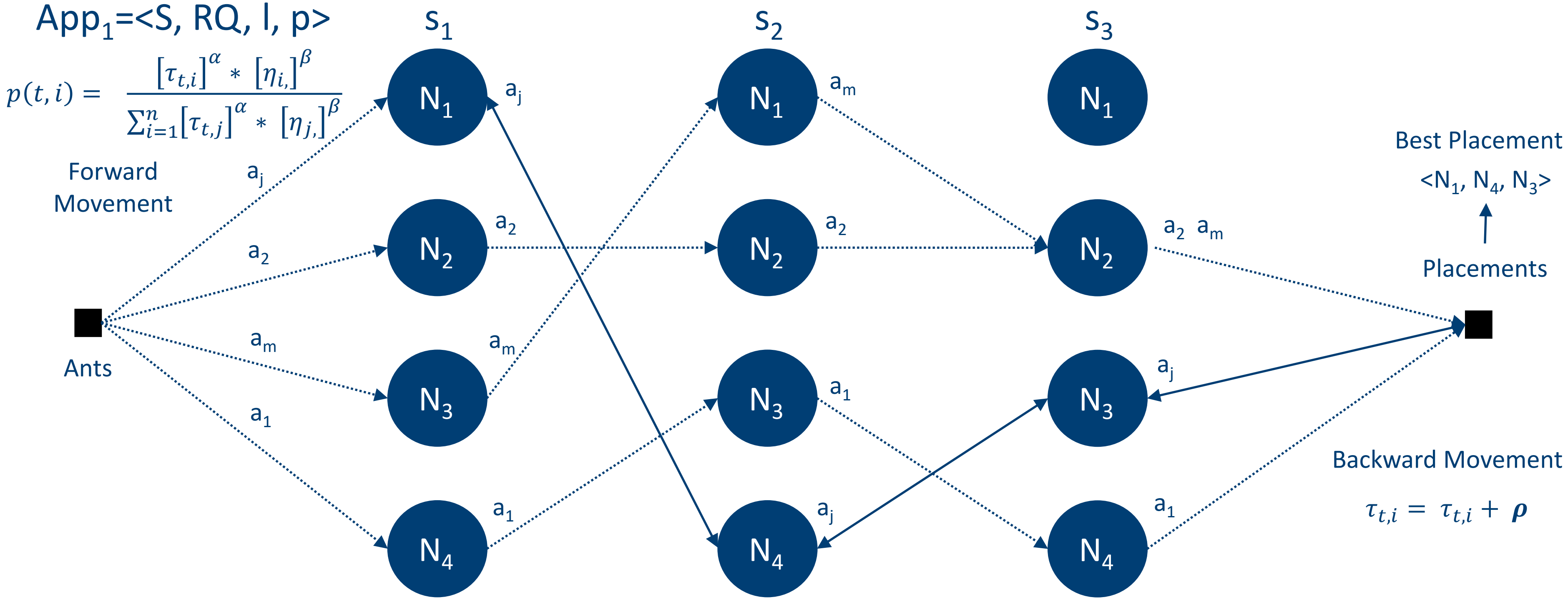
<N$_1$, N$_4$, N$_3$>

Placements

Backward Movement

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

AI Application - Dynamic Service Placement in Edge Computing



App$_1$=<S, RQ, l, p>

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

Forward Movement

Best Placement

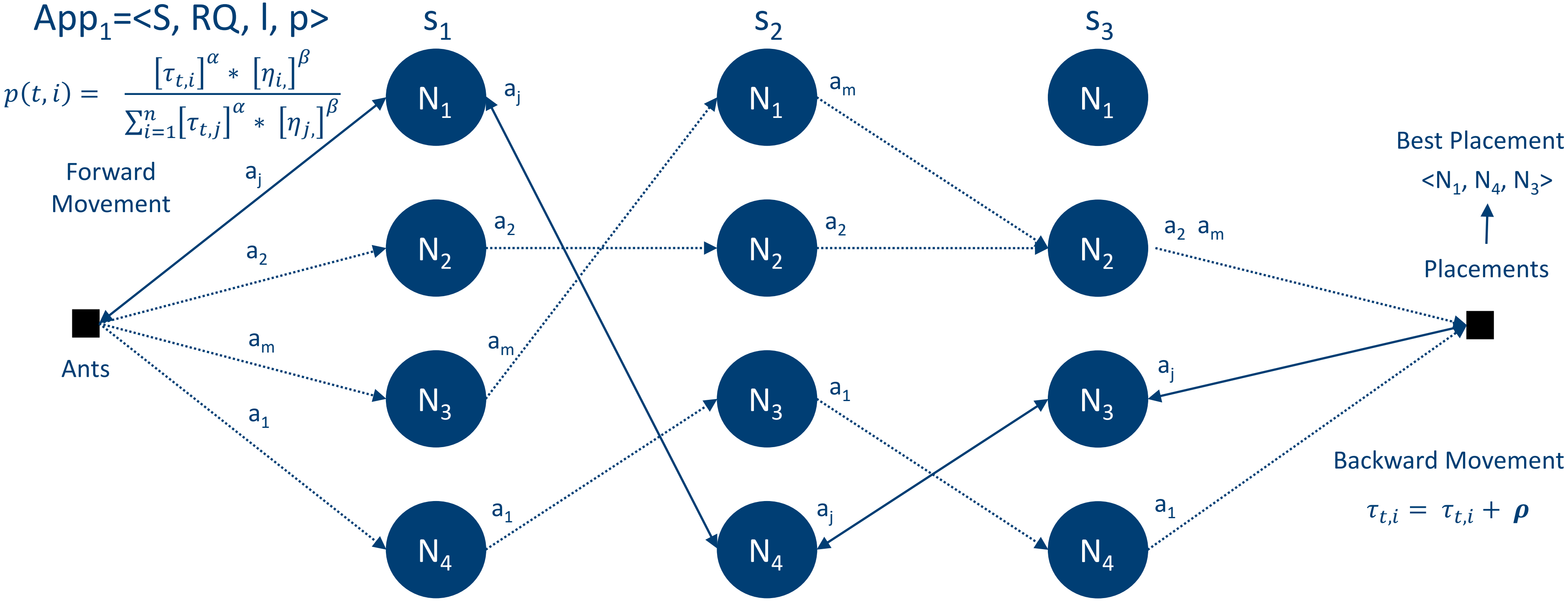
<N$_1$, N$_4$, N$_3$>

Placements

Backward Movement

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

AI Application - Dynamic Service Placement in Edge Computing



App$_1$=<S, RQ, l, p>

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

Forward Movement

Ants

Best Placement

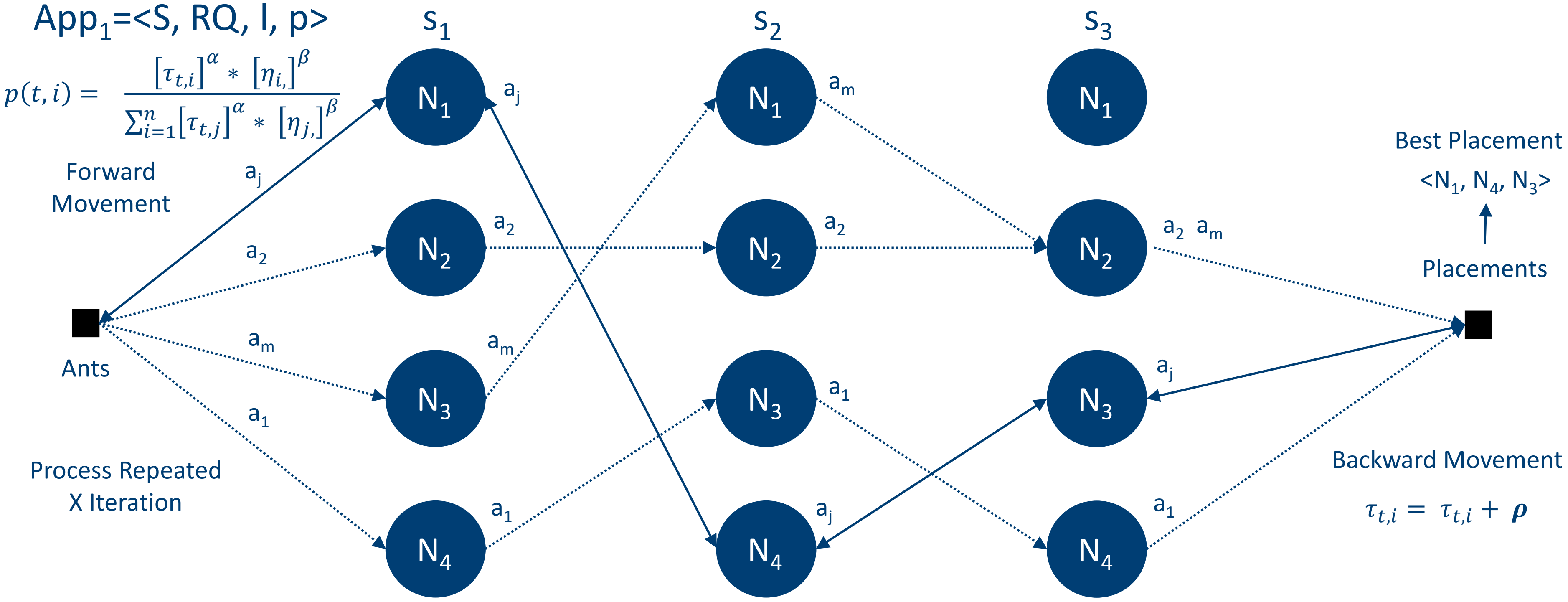
<N$_1$, N$_4$, N$_3$>

Placements

Backward Movement

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

AI Application - Dynamic Service Placement in Edge Computing



App_1=<S, RQ, l, p>

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

Forward Movement

Ants

Process Repeated X Iteration

Best Placement

<N_1, N_4, N_3>

Placements

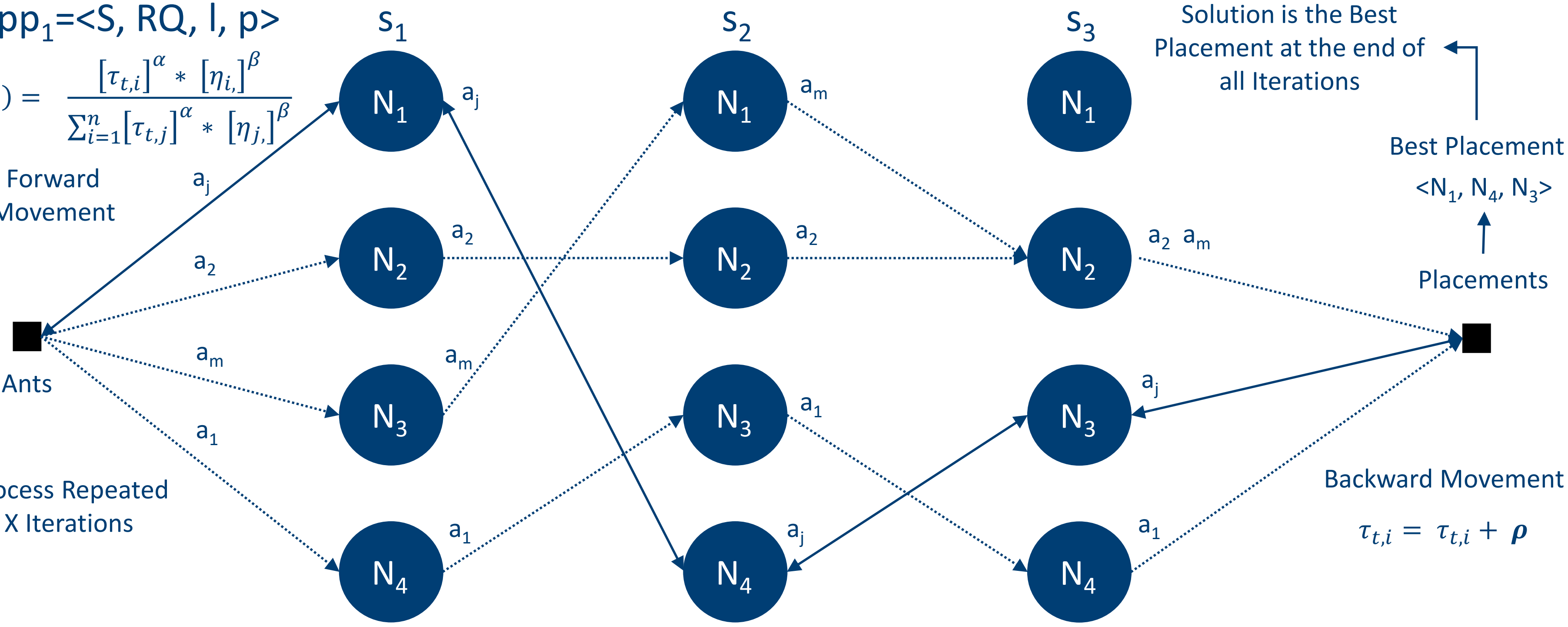
Backward Movement

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

AI Application - Dynamic Service Placement in Edge Computing

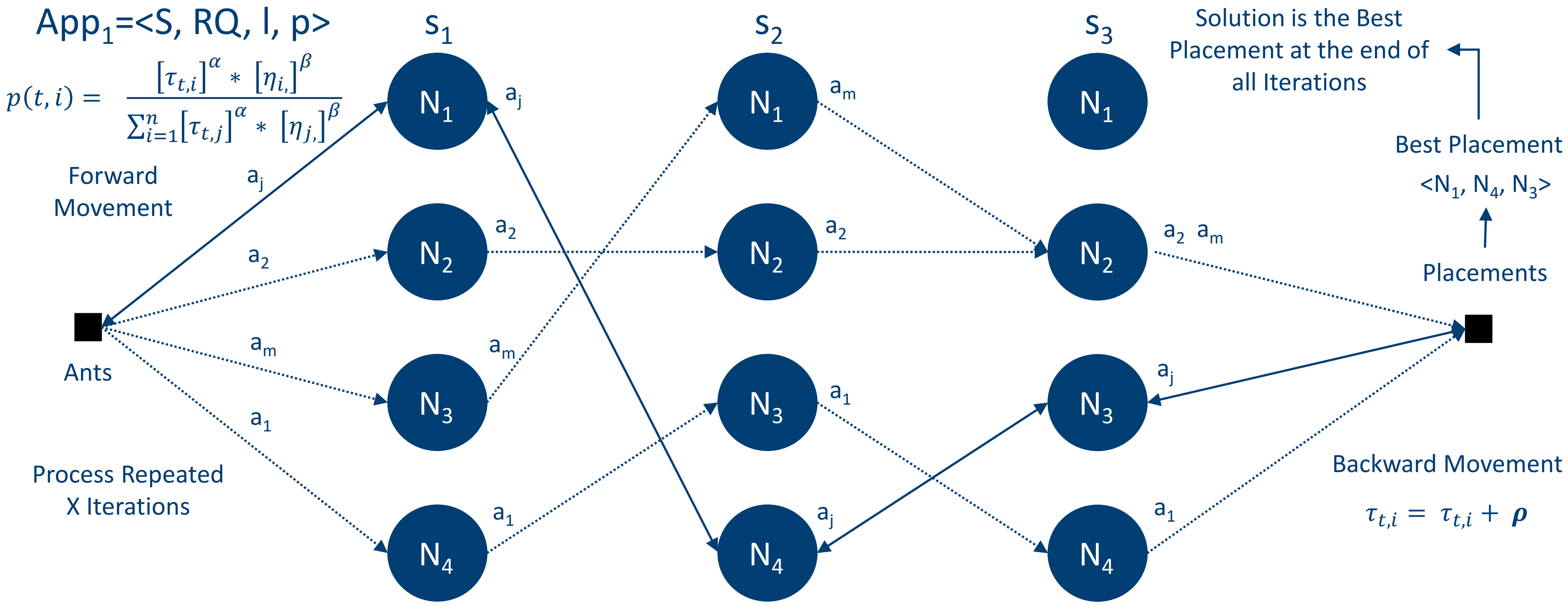
App_1=<S, RQ, l, p>

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



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

AI Application - Dynamic Service Placement in Edge Computing



App_1=<S, RQ, l, p>

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

Forward Movement

Ants

Process Repeated X Iterations

Solution is the Best Placement at the end of all Iterations

Best Placement

<N_1, N_4, N_3>

Placements

Backward Movement

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

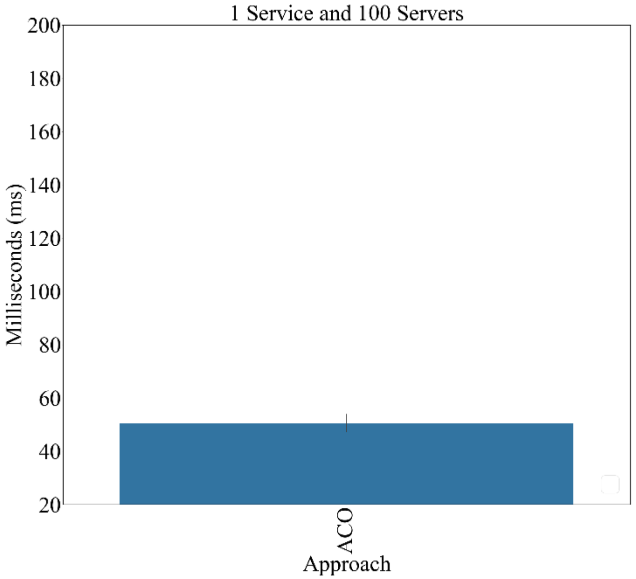
Execution time is affected by:

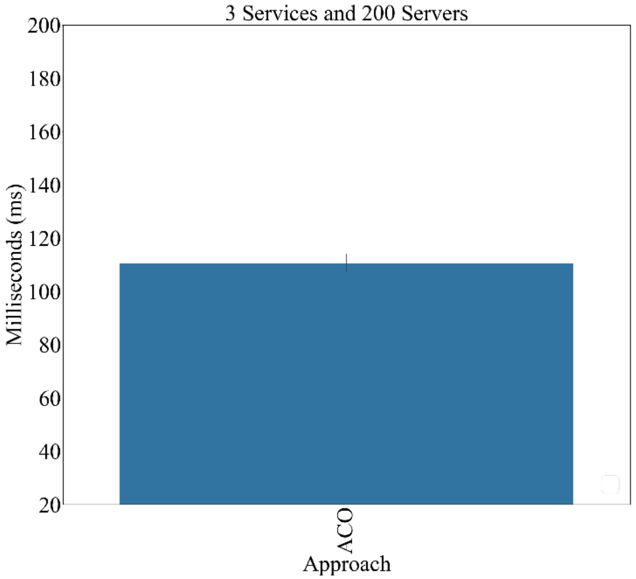
- Number of services.
- Number of edge servers.
- Number of ants.
- Number of iterations

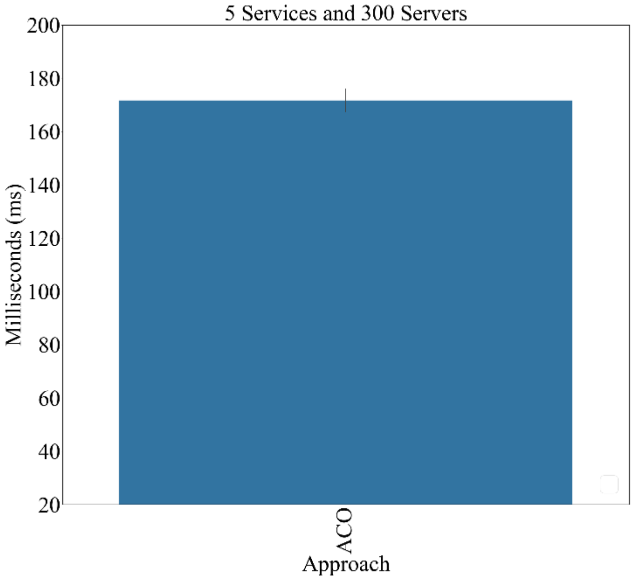
AI Application - Dynamic Service Placement in Edge Computing

Smart-city simulation[1]

Number of ants: 10
Number of iterations: 100


1 Service and 100 Servers


3 Services and 200 Servers

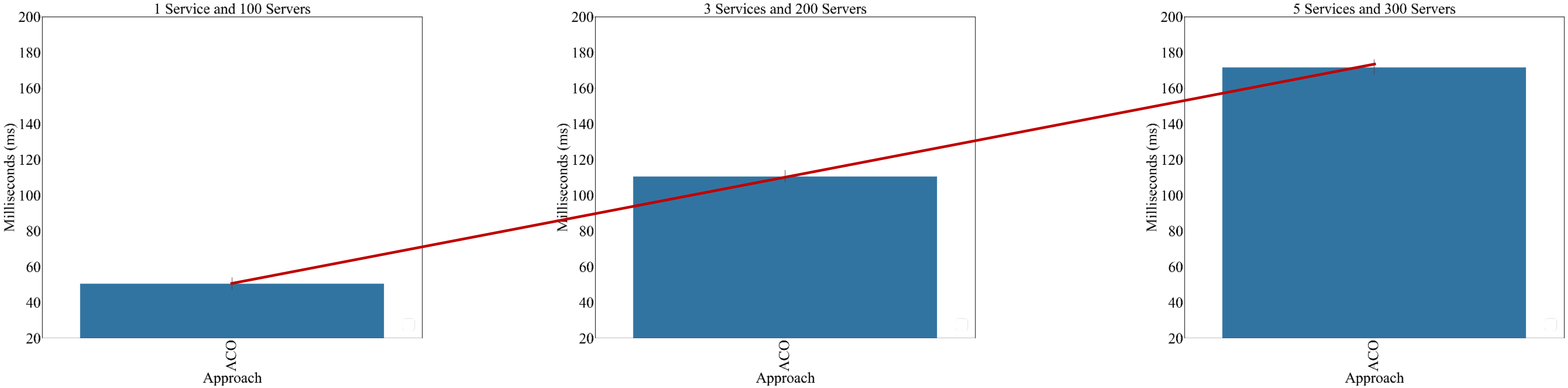

5 Services and 300 Servers

[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]

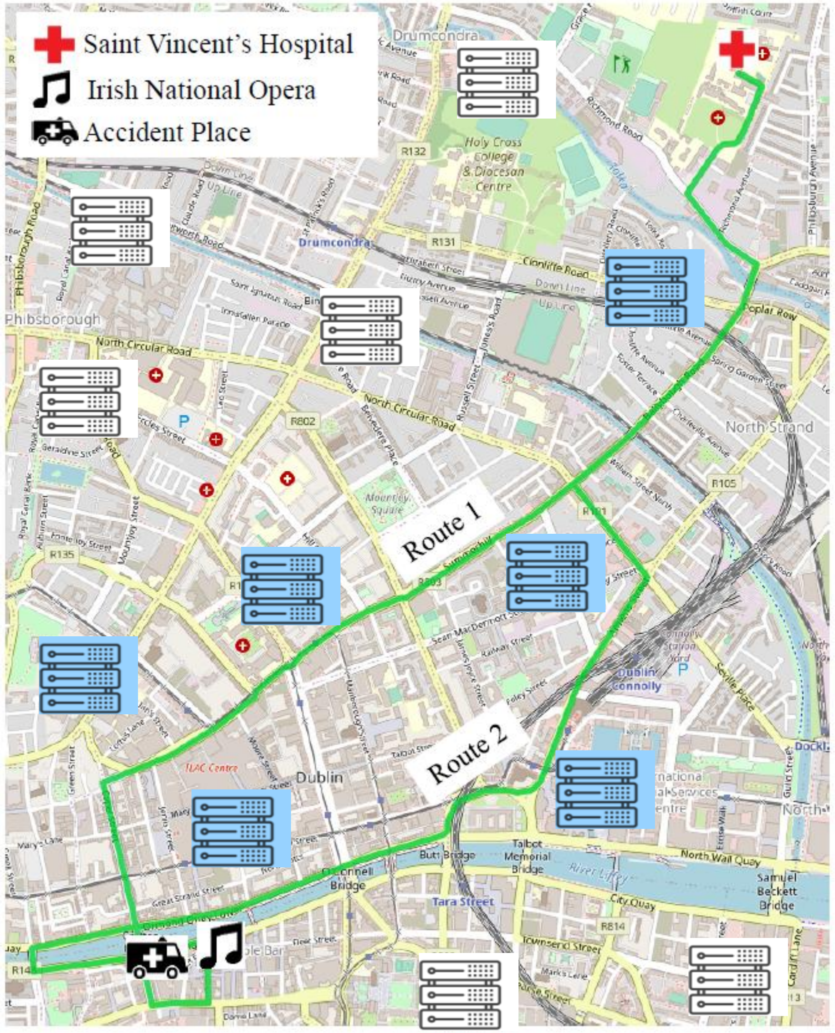
Number of ants: 10
Number of iterations: 100



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

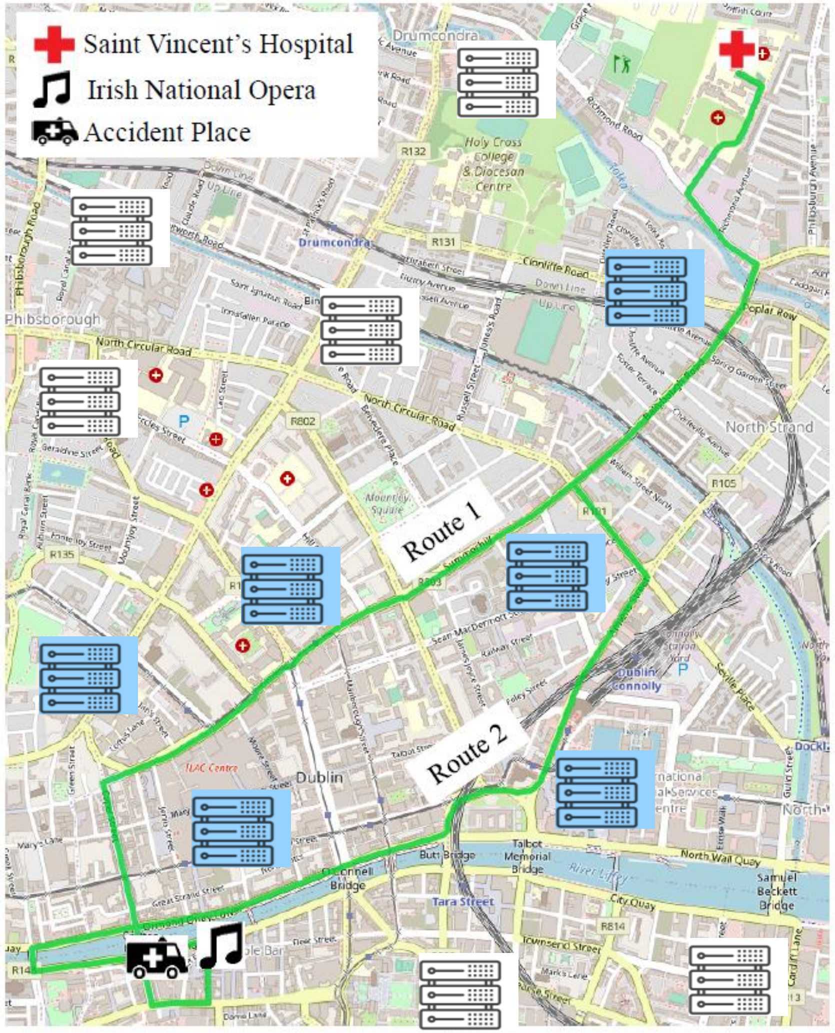


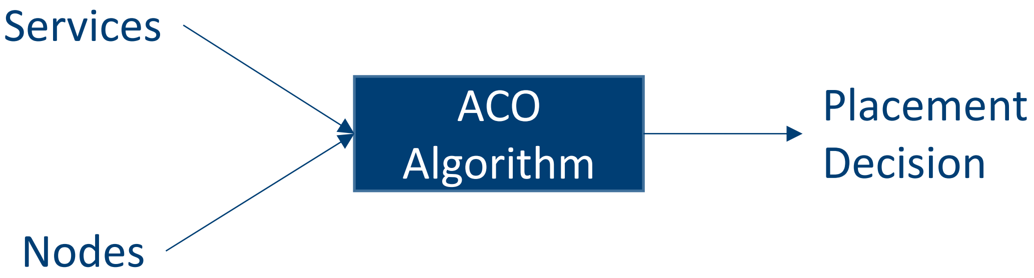
- Variables we cannot reduce:
  - Number of services.
  - Number of iterations.
  - Number of ants.

- **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





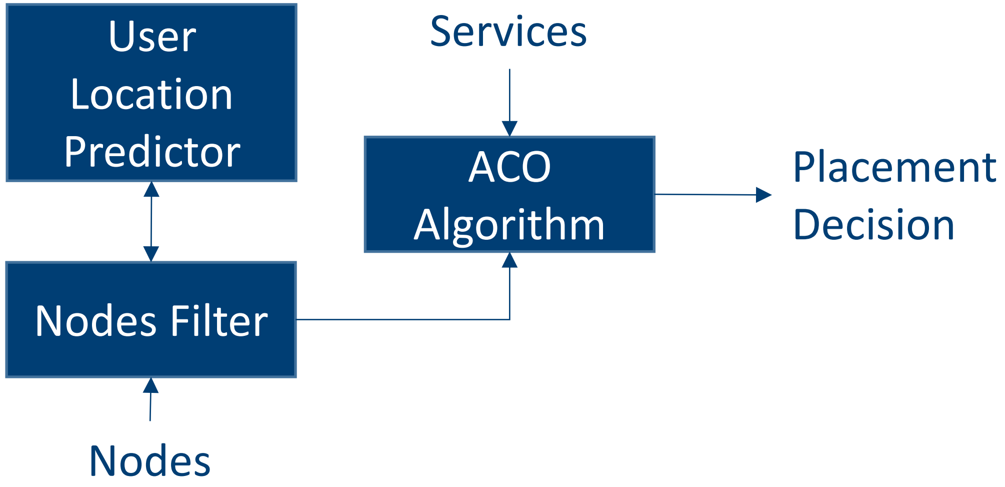
Services

Nodes

ACO Algorithm

Placement Decision

Re-design of the solution:



User Location Predictor

Nodes Filter

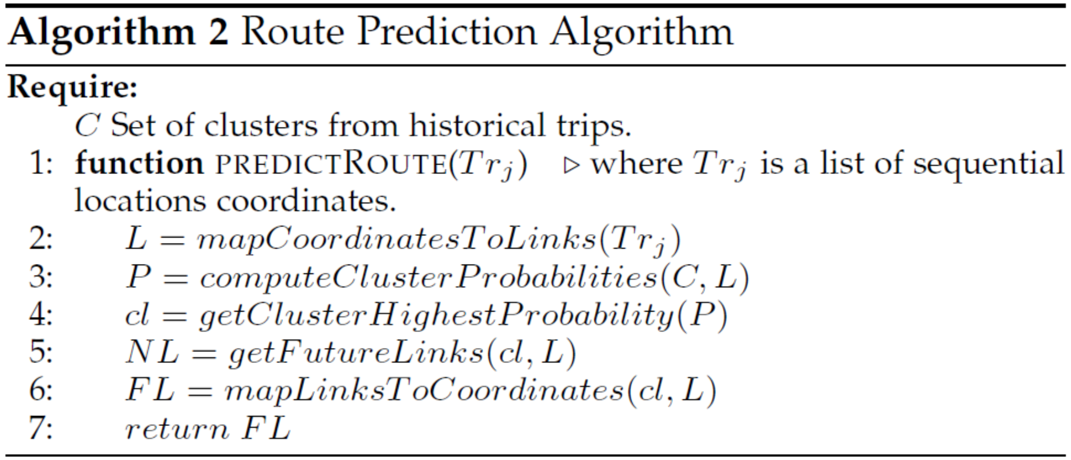
Nodes

Services

ACO Algorithm

Placement Decision

AI Application - Dynamic Service Placement in Edge Computing



Algorithm 2 Route Prediction Algorithm

```
Require:
    $C$ Set of clusters from historical trips.
function PREDICTROUTE($Tr_j$)    ▷ where $Tr_j$ is a list of sequential locations coordinates.
    $L = mapCoordinatesToLinks(Tr_j)$
    $P = computeClusterProbabilities(C, L)$
    $cl = getClusterHighestProbability(P)$
    $NL = getFutureLinks(cl, L)$
    $FL = mapLinksToCoordinates(cl, L)$
    $return\ FL$
```

- The goal is to select the **edge servers** that are **close to** the current and future **user's location.**

- Two approaches were explored: **Bayesian Classifier[1]** and **Hidden Markov Model[2]**.

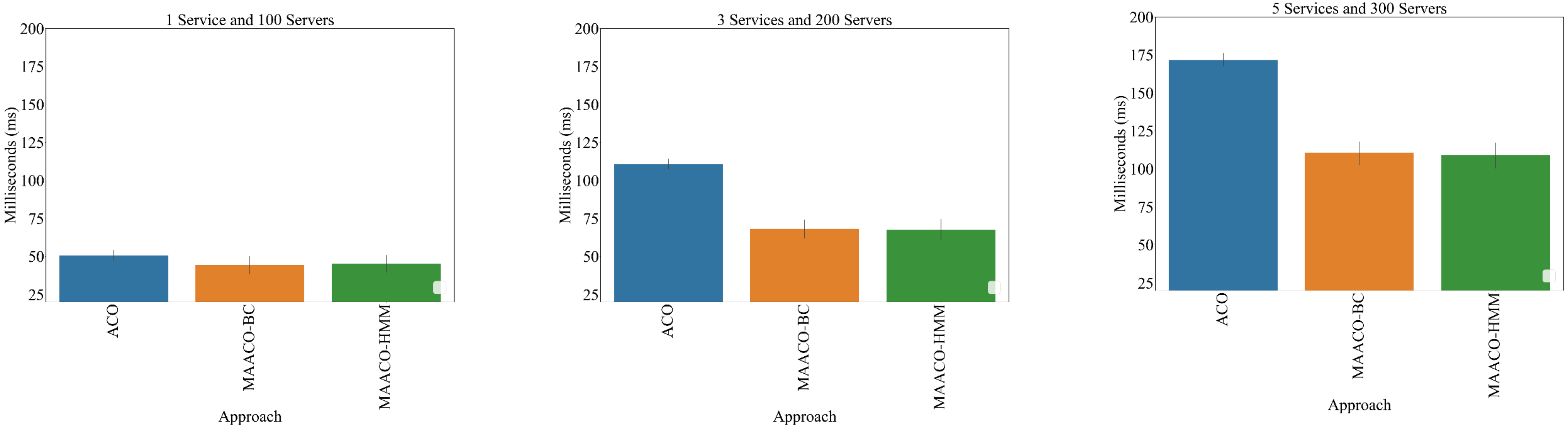
[1] J. P. Epperlein, J. Monteil, M. Liu, Y. Gu, S. Zhuk, and R. Shorten, "Bayesian classifier for route prediction with markov chains," in 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018, pp. 677–682.

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

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]

Set of clusters of historical trips C.

Transition matrix for each cluster cl_k. $\mathrm{m} \times m$ dimensions where m is the total number of unique links in historical trips.

The output of the training maps:

$$(L, cl_k) \rightarrow P(L, cl_k)$$

Prediction of a trip belonging to cl_k:

$$p(l_0, \dots, l_t | cl_k) = p(l_0, \dots, l_{t-1} | cl_k)$$

Initial probability:

$$p(l_0 = i | cl_k) = \frac{|L_m | l_0 = i \wedge L_m \in cl_k|}{|L_m | L_m \in cl_k|}$$

Hidden Markov Model [2]

Set of clusters of historical trips C.

Frequency matrix $F_{m \times n}$, where m is the total number of unique links in historical trips and n is the number of clusters.

Prob. of being on link l, if current trip belongs to cl:

$$p(l | C = cl) = \frac{F_{l,cl}}{\sum_{i=1}^{m} F_{i,cl}}$$

Prob. of being in cl knowing current link l :

$$p(C = cl | l) = \frac{F_{l,cl}}{\sum_{j=1}^{n} F_{l,j}}$$

Initial probability:

$$p(cl_k | l_0, \dots, l_t) = p(l_t | cl_k) p(cl_k | l_0, \dots, l_{t-1})$$

[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,” n 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.

[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,” n 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,” n 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

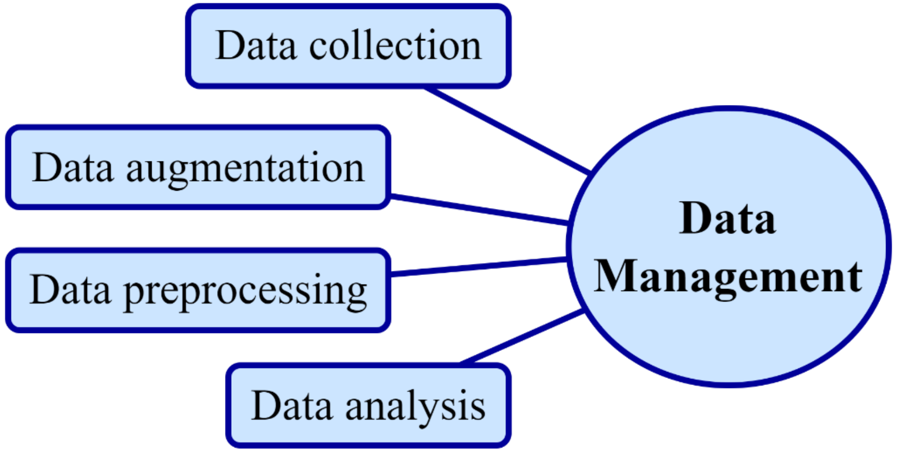
The AutoAI Programme scales our ability to deploy safe and reliable AI solutions, driving innovation in machine learning-enabled techniques for deploying, maintaining, and understanding AI systems. By investigating how to decompose AI systems into their component parts, how to manage data in system development,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

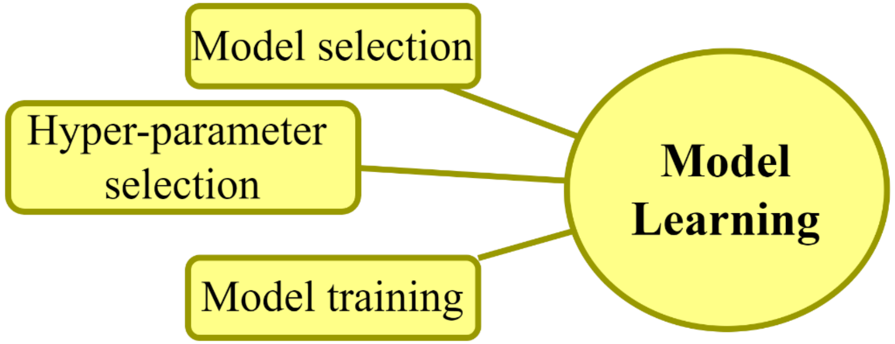
AutoAI Programme

$$model + data \xrightarrow{compute} prediction$$

- 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 lack of focus on data causes several challenges when deploying ML…**

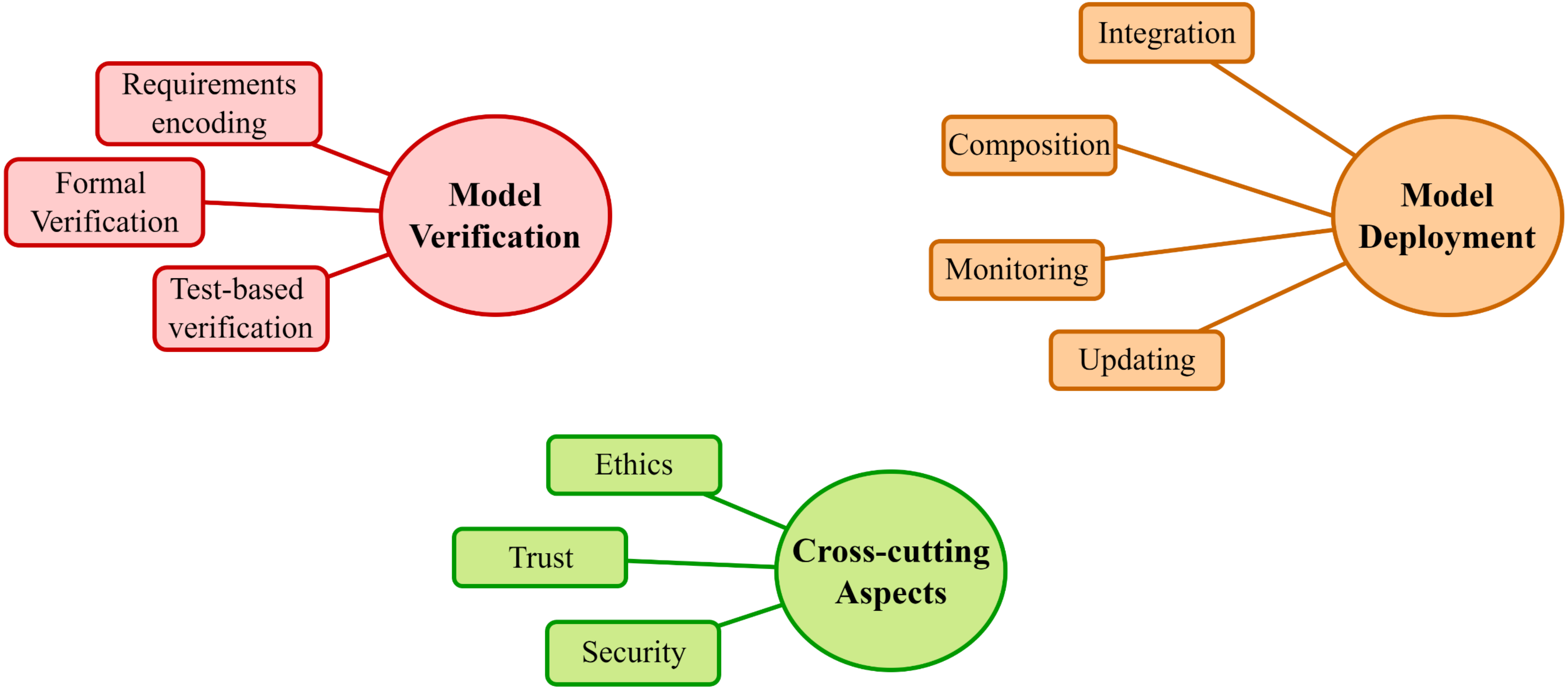
ML Challenges in Real World[1]





[1] Paleyes, A., Urma, R. G., & Lawrence, N. D. (2020). Challenges in deploying machine learning: a survey of case studies. *arXiv preprint arXiv:2011.09926*.

ML Challenges in Real World[1]



Model Verification

- Requirements encoding
- Formal Verification
- Test-based verification

Model Deployment

- Integration
- Composition
- Monitoring
- Updating

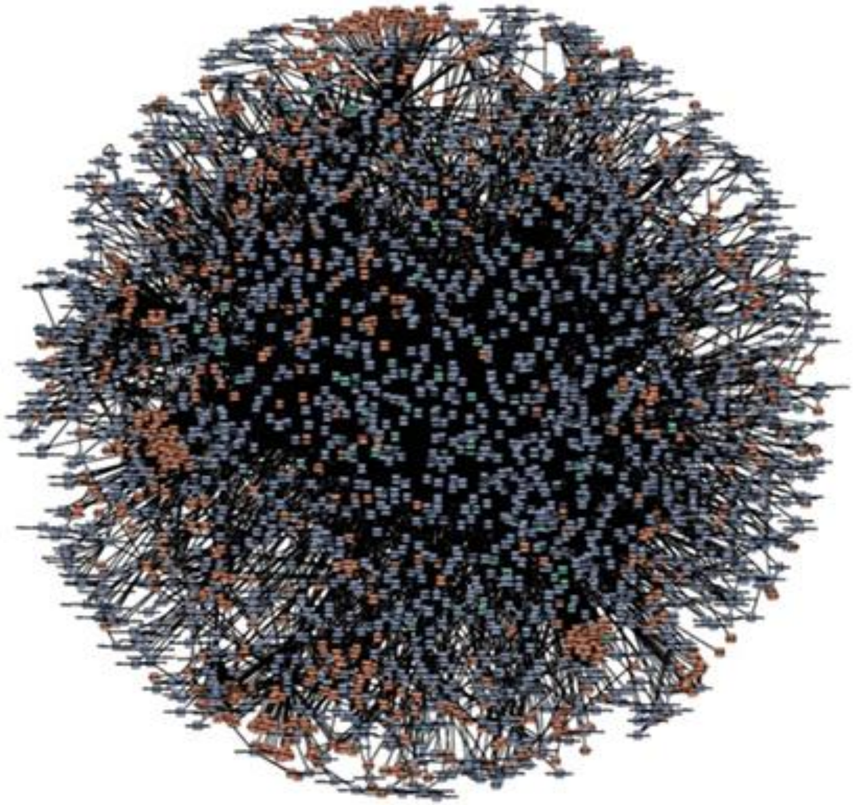
Cross-cutting Aspects

- Ethics
- Trust
- Security

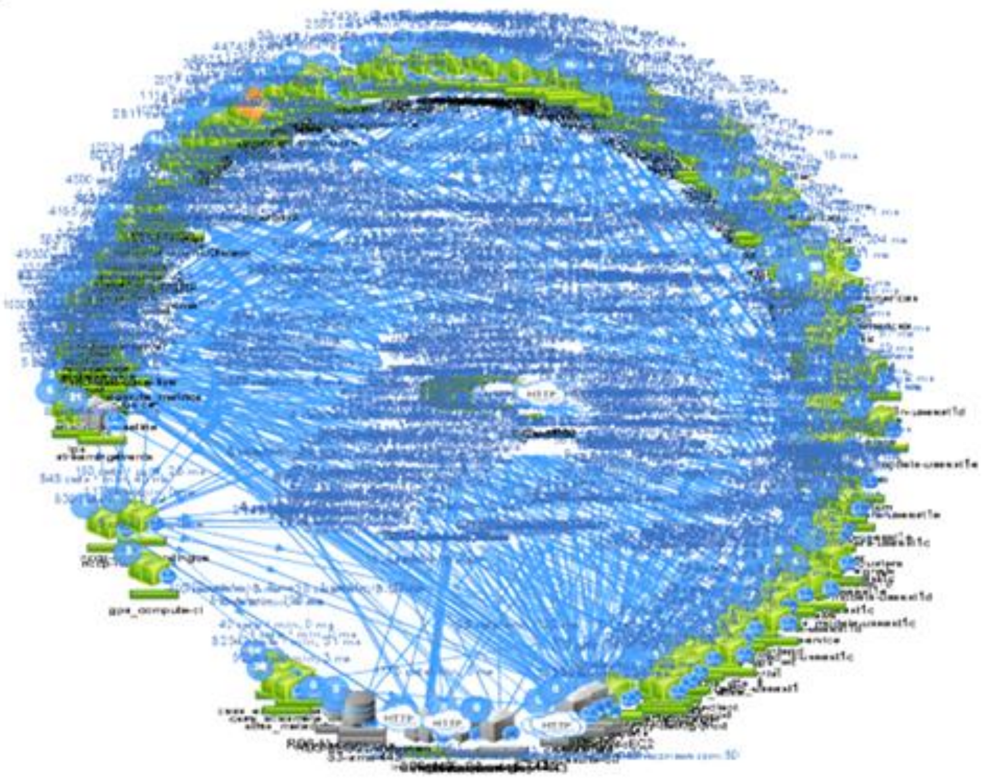
[1] Paleyes, A., Urma, R. G., & Lawrence, N. D. (2020). Challenges in deploying machine learning: a survey of case studies. *arXiv preprint arXiv:2011.09926*.

Systems Design

Microservices Architecture



amazon.com

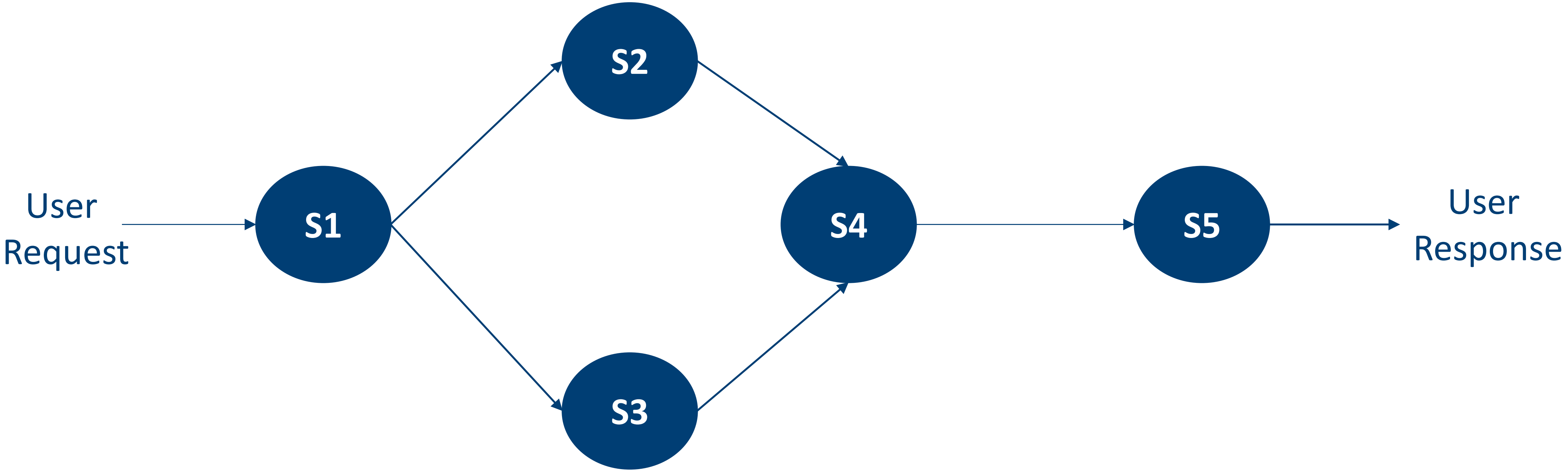


NETFLIX

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

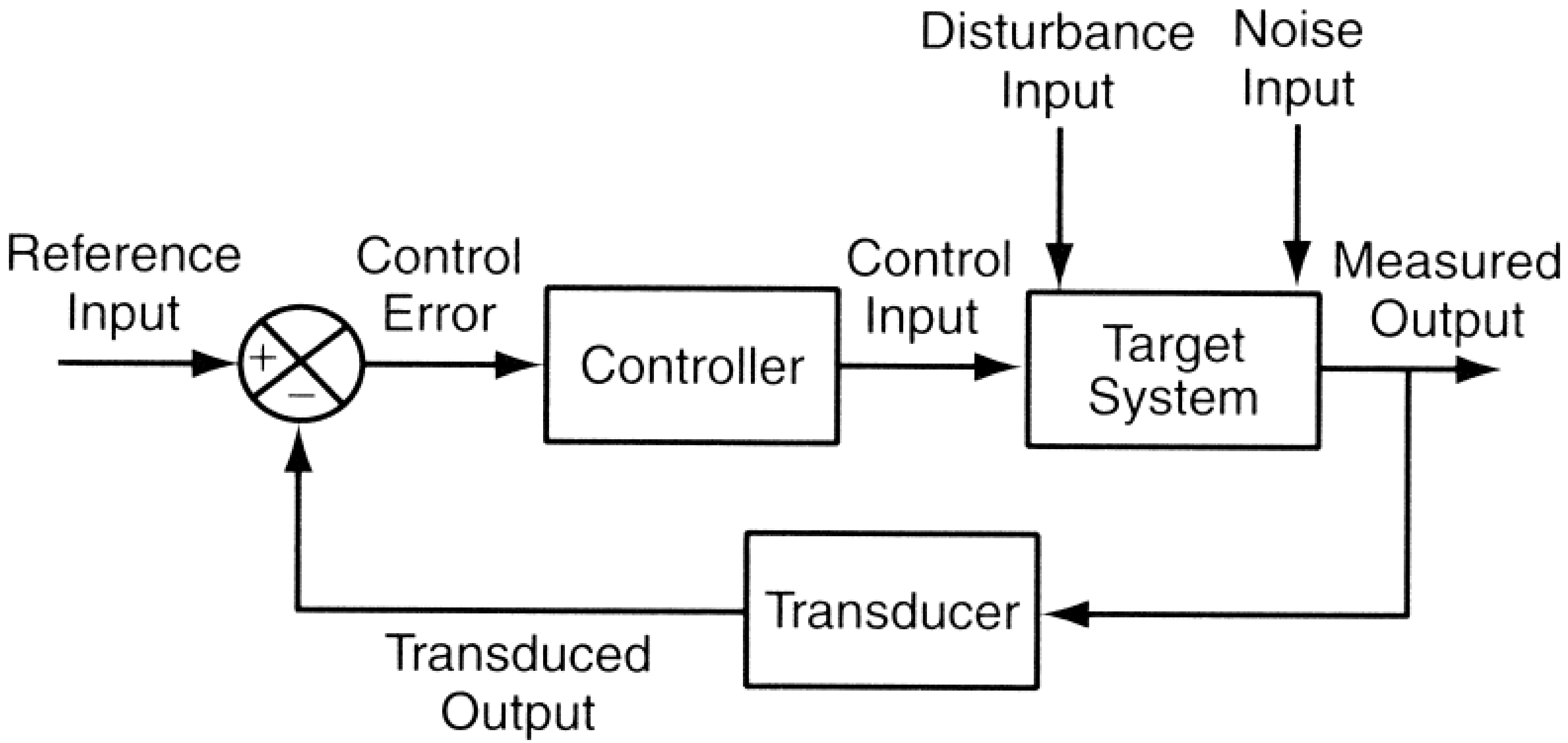
Systems Design

Microservices Architecture



Systems Design

Control System



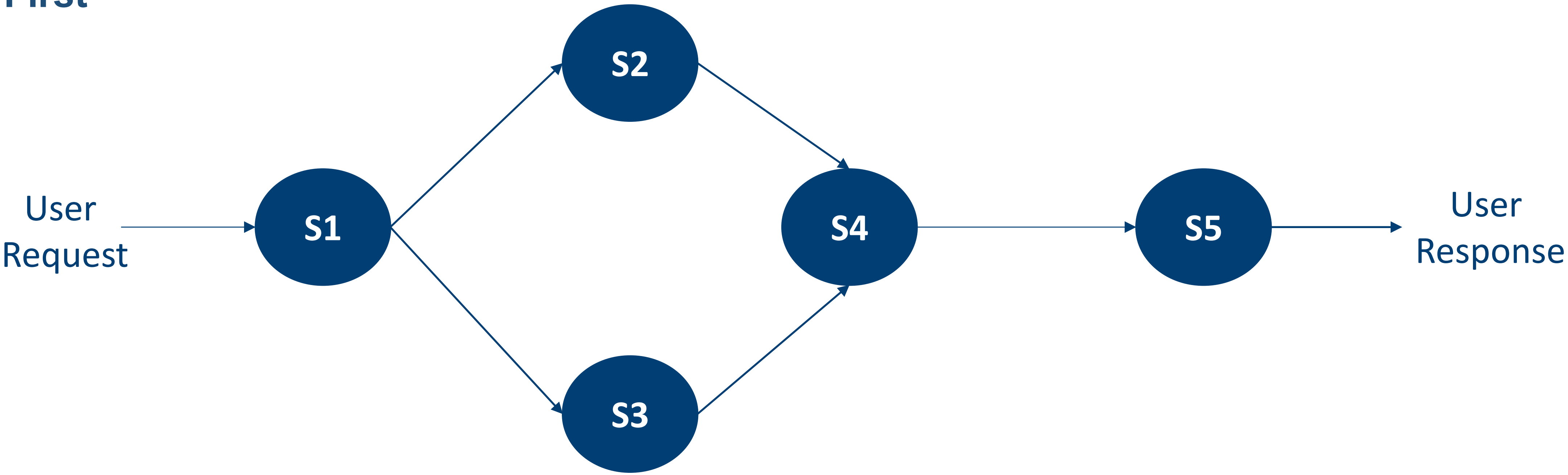
Source: https://www.eecs.umich.edu/courses/eecs373.w05/lecture/control.html

Ongoing Work – Data-Oriented Architectures

- Architecture principles:
  - **Data First.**
  - **Decentralised Architecture.**
  - **Open Architecture.**

Ongoing Work – Data-Oriented Architectures

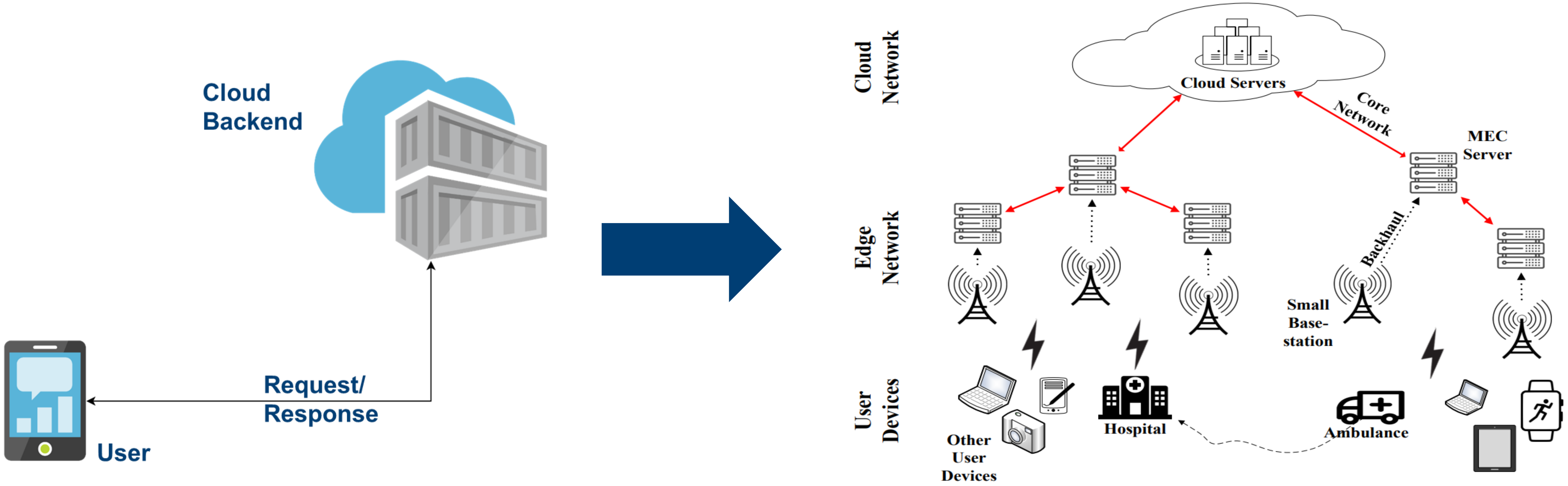
- Data First



Data is primary and the operations on data are secondary.

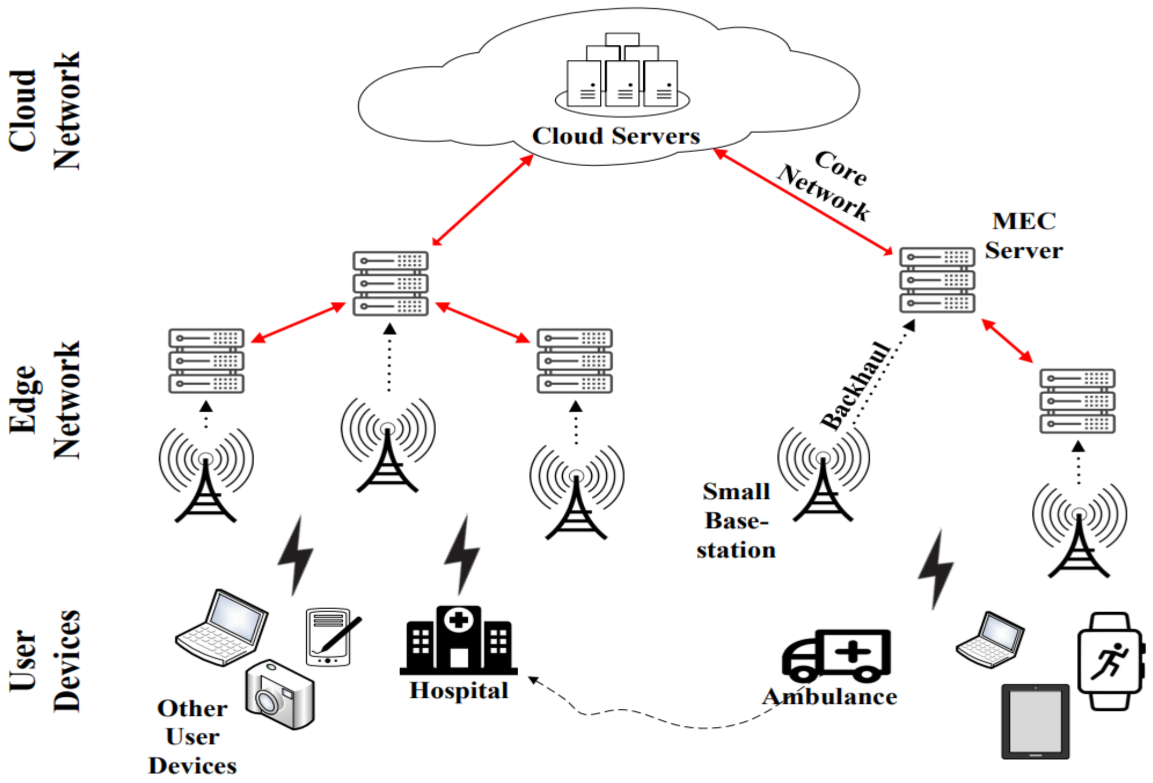
Ongoing Work – Data-Oriented Architectures

- Decentralised architecture.



Ongoing Work – Data-Oriented Architectures

- Open architecture.



Ongoing Work – Data-Oriented Architectures

- **Initial Exploration – SOA vs FBP[1,2]**

We explored programming paradigms that can enable DOAs and compare them against classic SOA.

- **Flow Based Programming:**

Dataflow programming paradigm that defines software applications as a set of processes which exchange data via connections that are external to those processes.

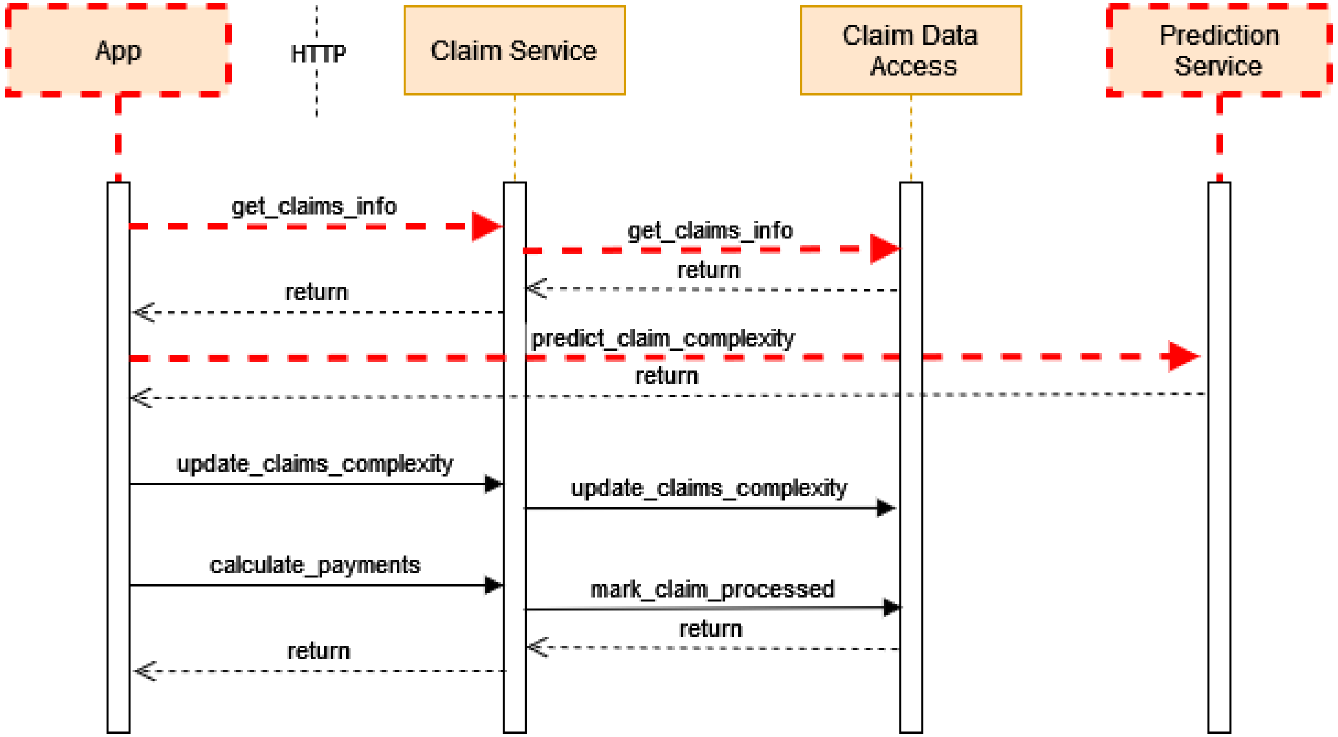
- **We implemented different applications with SOA and FBP and then measured code quality.**

[1] Paleyes, A., Cabrera, C., & Lawrence, N. Towards better data discovery and collection with flow-based programming. In *2021 Neurips Data-Centric AI Workshop (DCAI)*

[2] Paleyes, A., Cabrera, C., & Lawrence, N. An Empirical Evaluation of Flow Based Programming in the Machine Learning Deployment Context. 1st International Conference on AI Engineering – Software Engineering for AI. To Appear, May 2022.

Ongoing Work – Data-Oriented Architectures

- Initial Exploration – SOA vs FBP[1,2]

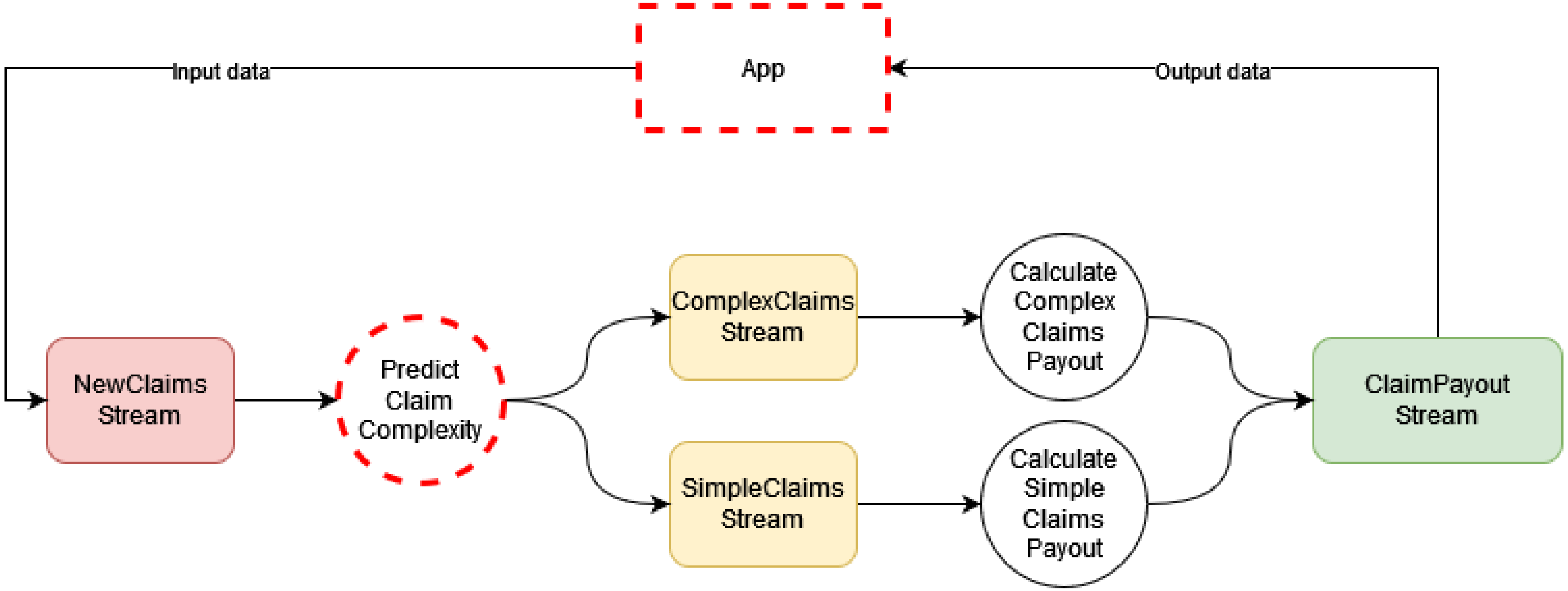


[1] Paleyes, A., Cabrera, C., & Lawrence, N. Towards better data discovery and collection with flow-based programming. In *2021 Neurips Data-Centric AI Workshop (DCAI)*

[2] Paleyes, A., Cabrera, C., & Lawrence, N. An Empirical Evaluation of Flow Based Programming in the Machine Learning Deployment Context. 1st International Conference on AI Engineering – Software Engineering for AI. To Appear, May 2022.

Ongoing Work – Data-Oriented Architectures

- Initial Exploration – SOA vs FBP[1,2]

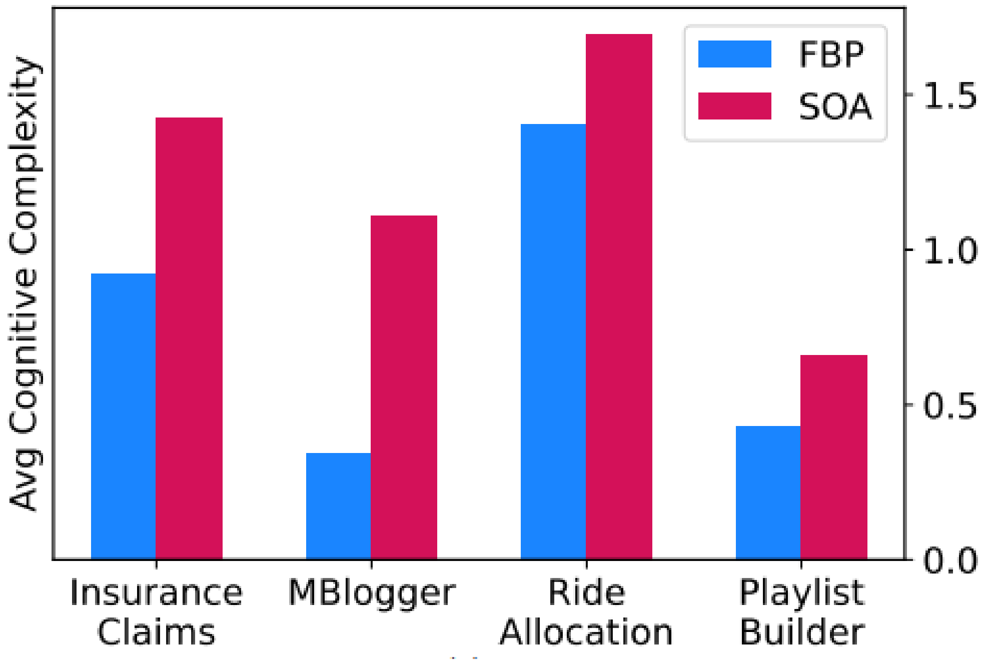


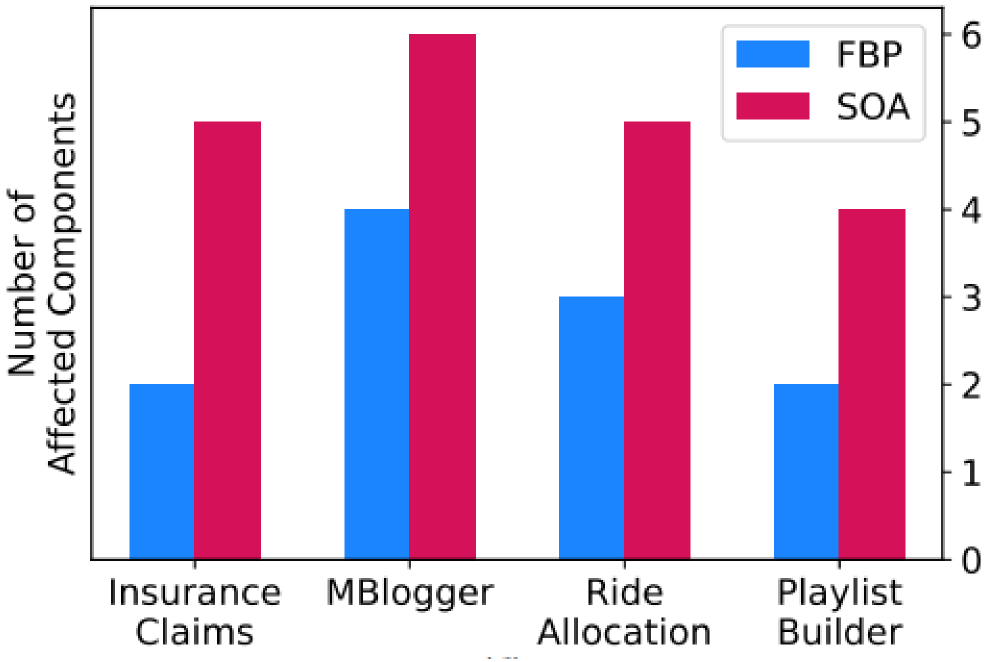
[1] Paleyes, A., Cabrera, C., & Lawrence, N. Towards better data discovery and collection with flow-based programming. In *2021 Neurips Data-Centric AI Workshop (DCAI)*

[2] Paleyes, A., Cabrera, C., & Lawrence, N. An Empirical Evaluation of Flow Based Programming in the Machine Learning Deployment Context. 1st International Conference on AI Engineering – Software Engineering for AI. To Appear, May 2022.

Ongoing Work – Data-Oriented Architectures

- Initial Exploration – SOA vs FBP[1,2]





[1] Paleyes, A., Cabrera, C., & Lawrence, N. Towards better data discovery and collection with flow-based programming. In *2021 Neurips Data-Centric AI Workshop (DCAI)*

[2] Paleyes, A., Cabrera, C., & Lawrence, N. An Empirical Evaluation of Flow Based Programming in the Machine Learning Deployment Context. 1st International Conference on AI Engineering – Software Engineering for AI. To Appear, May 2022.

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.

- The deployment of AI-based systems in real-world environments is hard.

- New paradigmns are needed to realize the potential of AI.

Thank you!

chc79@cam.ac.uk