Photo by NASA on Unsplash (retrieved 2023-05-07)

Graph Data Science

The world is full of challenges

Photo by NASA on Unsplash (retrieved 2023-05-07)

The world is full of challenges

Photo by NASA on Un

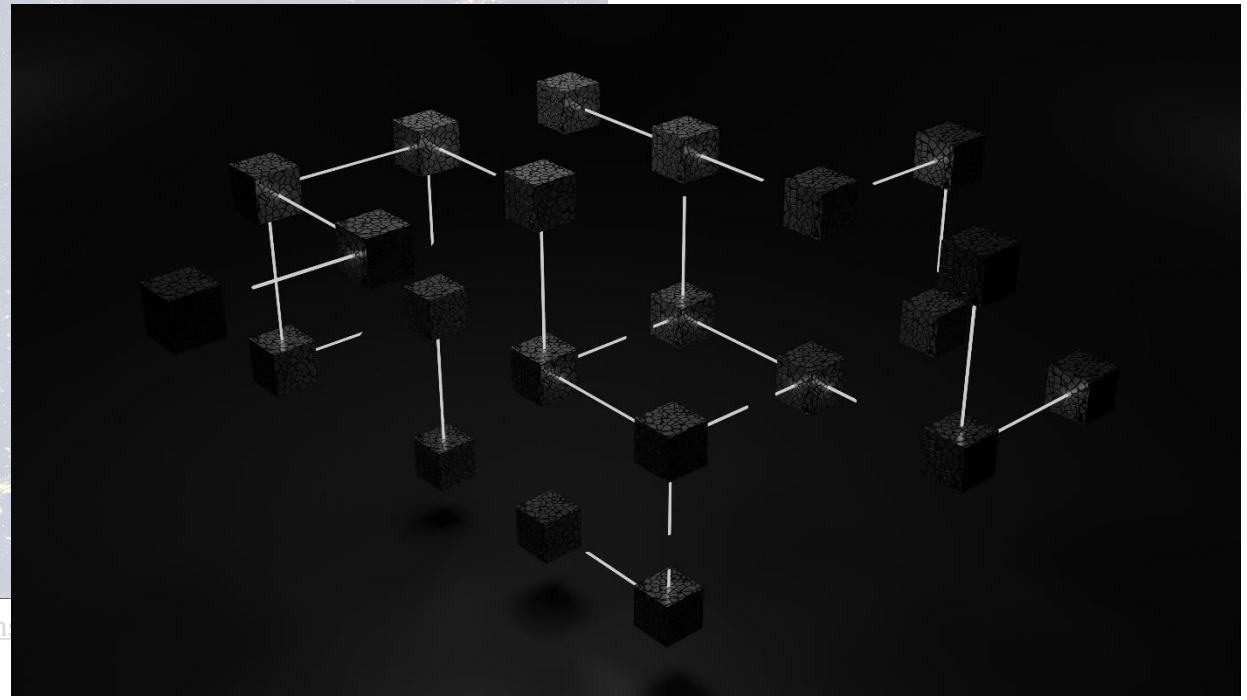


Photo by Shubham Dhage on Unsplash (retrieved 2023-05-07)

The world is full of challenges

Photo by NASA on Un

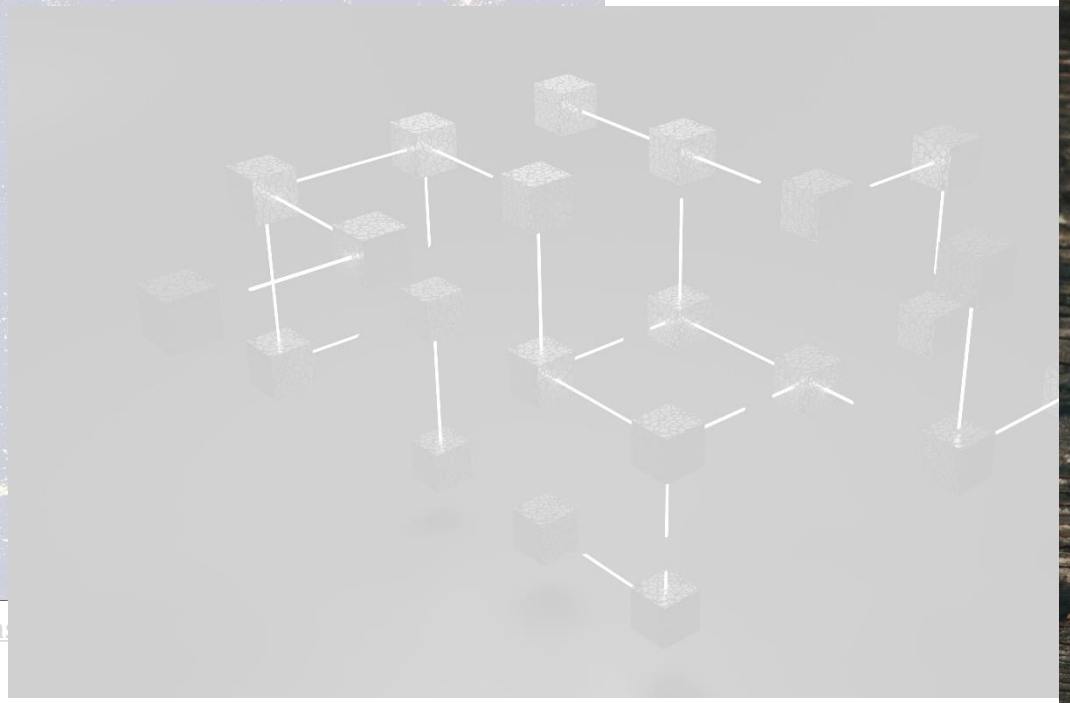


Photo by Shubham Dhage on Unsplash (retrieved 2023-05

Photo by Nathan Dumlao on Unsplash (retrieved 2023-05-07)

The world is full of challenges

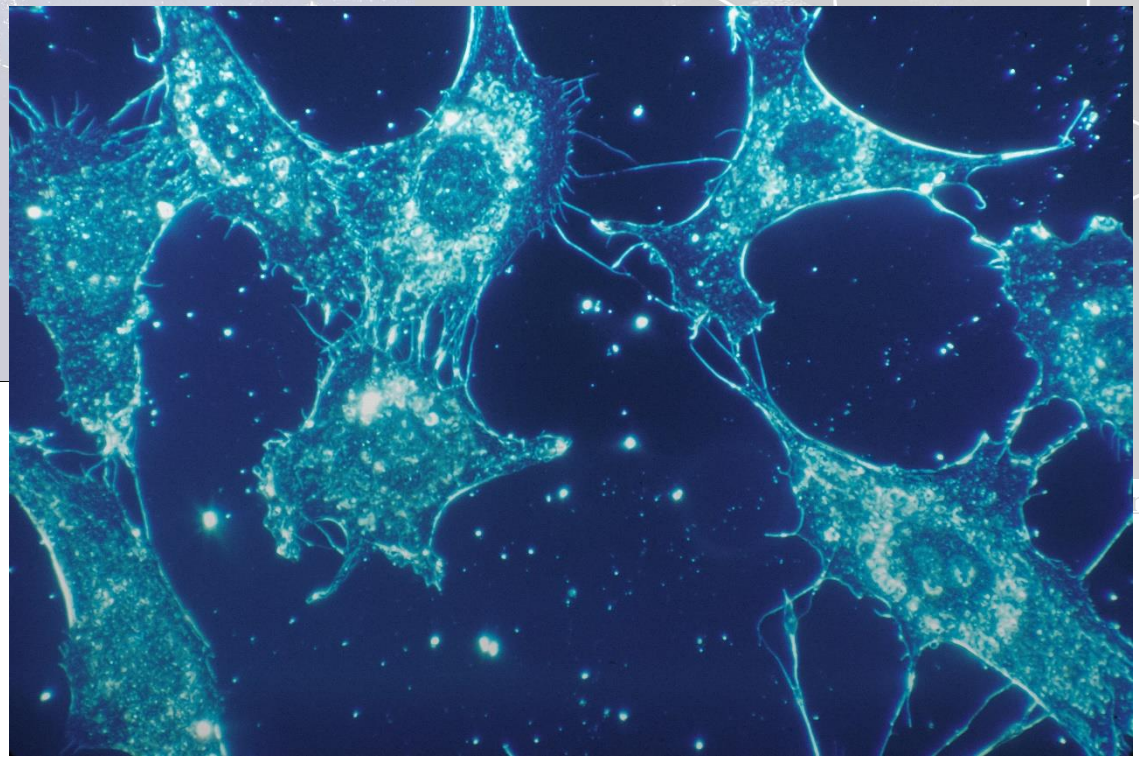


Photo by National Cancer Institute on Unsplash (retrieved 2023-05-07)

Photo by Nathan Dumlao on Unsplash (retrieved 2023-05-07)

The world is full of challenges

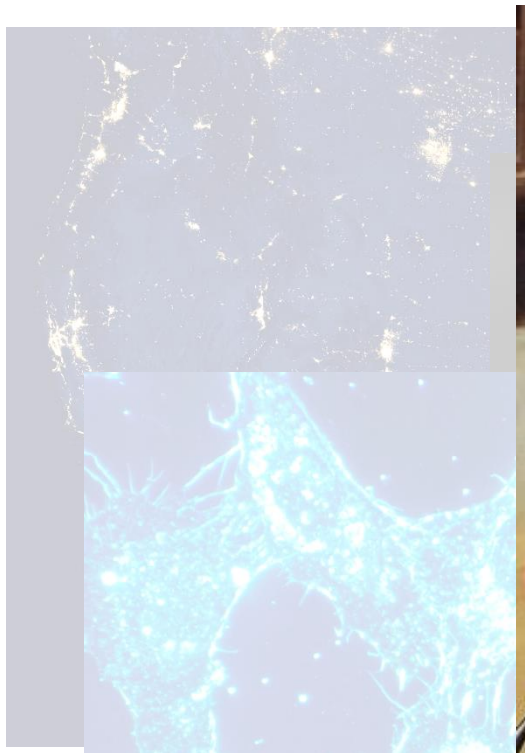


Photo by National Cancer [illegible] Wikimedia (retrieved 2023-05-07)

[illegible] Dumlao on Unsplash (retrieved 2023-05-07)

This course – getting to know each other

About me: Nanotechnology → theoretical physics → complex systems
→ PhD in Computer Science / AI → Sr. Data Scientist at FedEx

My contact: ilja.rausch@pm.me

My LinkedIn: **linkedin.com/in/ilja-rausch/**

Disclaimer: None of the material is linked to FedEx!

This course – getting to know each other

About you

- Your background and interests
  - Engineering (mechanical, industrial, technology and systems)
  - Innovation and design
  - Cybersecurity

This course – getting to know each other

About you

- Your background and interests
  - Engineering (mechanical, industrial, technology and systems)
  - Innovation and design
  - Cybersecurity

- Your knowledge and experience
  - Python
  - AIML
  - Data analysis
  - Business / industry

This course – getting to know each other

About you

- Your background and interests
  - Engineering (mechanical, industrial, technology and systems)
  - Innovation and design
  - Cybersecurity

- Your knowledge and experience
  - Python
  - AIML
  - Data analysis
  - Business / industry

- Motivation to take this course
  - Fun / curiosity
  - Career development
  - FOMO
  - Other

Course overview - expectations

- Attendance: in my view, not 100% mandatory

Course overview - expectations

- Attendance: in my view, not 100% mandatory
- Computer:
  - useful but not a must to follow the lectures
  - strongly recommended for the graded presentation
- Slides: will be shared

Course overview - expectations

- Attendance: in my view, not 100% mandatory

- Computer:
  - useful but not a must to follow the lectures
  - strongly recommended for the graded presentation

- Slides: will be shared

- Grading:
  - one presentation (10 – 15 minutes) per student, 5 minutes Q&A
  - in groups of two; Dates: Jul 3rd - Jul 5th

Course overview - expectations

- Attendance: in my view, not 100% mandatory

- Computer:
  - useful but not a must to follow the lectures
  - strongly recommended for the graded presentation

- Slides: will be shared

- Grading:
  - one presentation (10 – 15 minutes) per student, 5 minutes Q&A
  - in groups of two; Dates: Jul 3rd - Jul 5th
  - all tools are allowed (PowerPoint, LaTeX, Google Slides)
  - strongly invited to do your own literature research
    - Wikipedia is acceptable for images but insufficient for information
    - ChatGPT is allowed but be very very careful

Course overview - expectations

- Attendance: in my view, not 100% mandatory

- Computer:
  - useful but not a must to follow the lectures
  - strongly recommended for the graded presentation

- Slides: will be shared

- Grading:
  - one presentation (10 – 15 minutes) per student, 5 minutes Q&A
  - in groups of two; Dates: Jul 3rd - Jul 5th
  - all tools are allowed (PowerPoint, LaTeX, Google Slides)
  - strongly invited to do your own literature research
    - Wikipedia is acceptable for images but insufficient for information
    - ChatGPT is allowed but be very very careful
  - cover the core notions about the selected topic + limitations
  - in simplified terms (similar to a technical presentation given to the upper management in business)
  - pack the information into a story (make it a fun story if you want); build an imaginary business case

Course overview - topics for presentations

1. Network centrality (PageRank, Katz)
2. Network controllability and control nodes

3. Trees in ML: random forests and gradient boosting
4. GNNs (GraphSAGE, GIN)
5. Reinforcement learning

6. Word/Text embeddings
7. Clustering and vector similarity in ML
8. Dimensionality reduction (PCA and t-SNE)
9. Data sampling techniques (SMOTE)
10. Bayesian hyperparameter tuning

11. Explainable AI (XAI), Feature importance
12. Recommender Systems
13. Multivariate testing and A/B testing
14. Time Series (ARIMA, drift and temporal networks)
15. Meta-heuristics and Genetic Algorithms/ACO/PSO

16. Collective foraging/flocking
17. Site-selection + inhibition in bee colonies
18. Simulating signal propagation on networks

19. Qubits

20. Prompt engineering

Course overview - syllabus

	Session #	Description
WEEK 1	1	Introductory lecture, course overview; some useful tools from statistics
	2	Introduction to Network Science
	3	Introduction to Machine Learning
	4	Neural Networks
	5	Deep Learning architectures: CNNs, Attention, Transformers, BERT
WEEK 2	6	Knowledge Graphs and Graph Neural Networks
	7	Data mining; Quantum computing;
	8	Collective intelligence, simulation and agent-based modeling
	9	Practical tips for software and ML development
	10	Practical tips for working in the private sector (business, industry)
WEEK 3	11	Students' presentations I
	12	Students' presentations II
	13	Students' presentations III
	14	Guided hands-on session for creating a simple PoC with GNNs and Python
	15	Last day of class: grading and general feedback

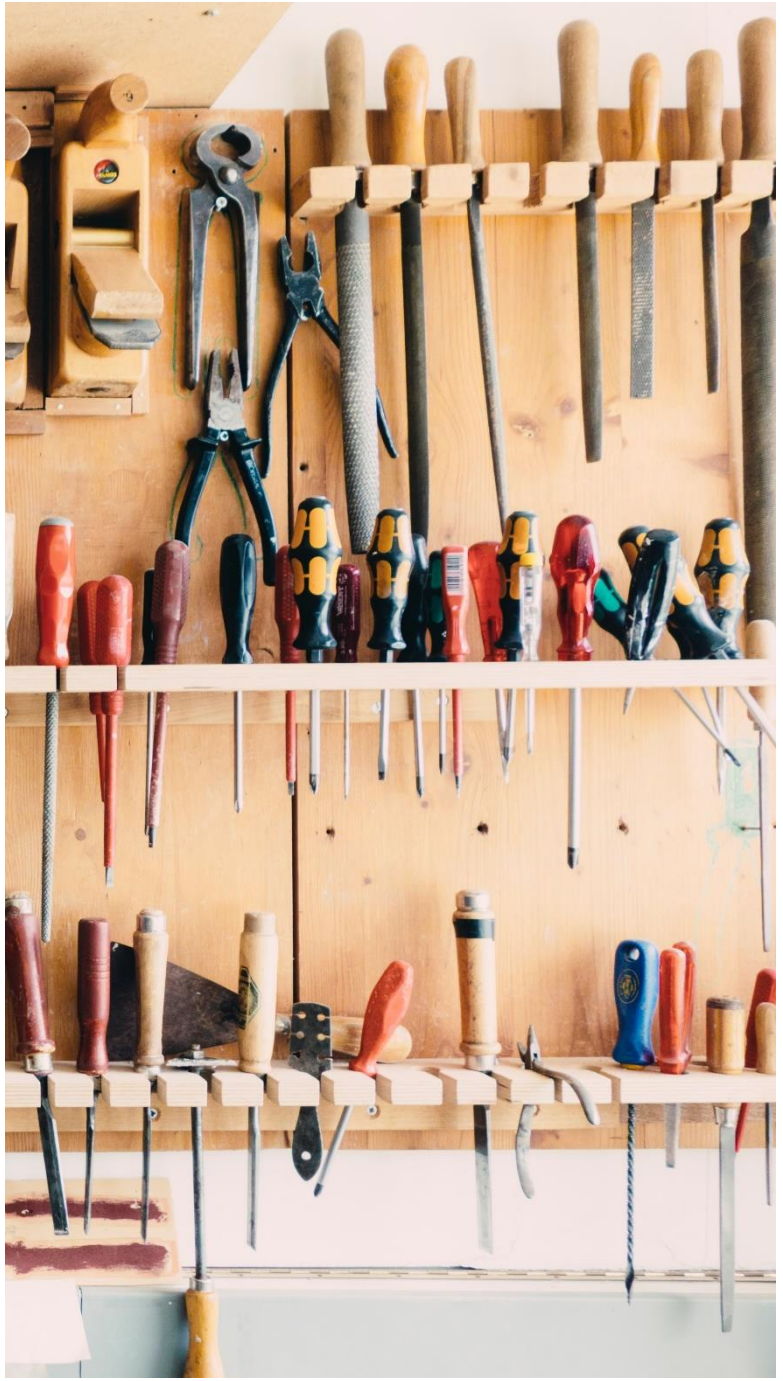


Photo by Barn Images on Unsplash (created 2023-05-13)

Course overview – my promise

- The course will widen your horizon
- Provide a broad (but incomplete) overview of Graph Data Science
- Lower the barrier to entry, enabling you to explore beyond
- Some insights useful for business / industry

Image created with Dall-E 2 (created 2023-05-13)

Useful literature and tutorials

- The Art of Statistics book by D. Spiegelhalter
- Statistics course by J. Blitzstein (Harvard)
- Network science book by A.-L. Barabási

- The Machine and Deep Learning Compendium
- A high-bias, low-variance introduction to Machine Learning for physicists paper by Mehta et al.
- Machine Learning course by Y. Abu-Mostafa

- Neural Networks lecture by F. Marquardt
- Deep Learning course by Andrew Ng
- Attention is all you need publication by Vaswani et al.

- Gentle introduction to GNNs web article by Sanchez-Lengeling et al.
- Machine Learning with Graphs by J. Leskovec (Stanford)

- Collective Motion paper by Viscek et al.
- Corey Shafer YouTube channel
- Scaled Agile Framework material

What is Graph Data Science?

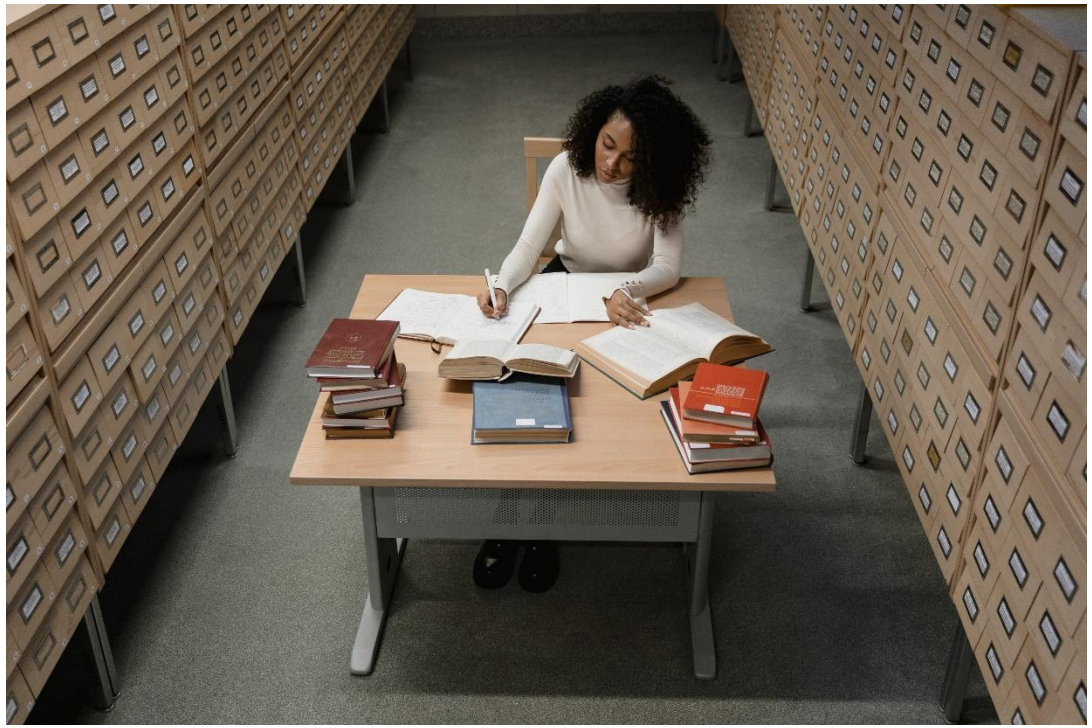


Photo by Tima Miroshnichenko from Pexels (retrieved 2023-05-07)

What is Graph Data Science?

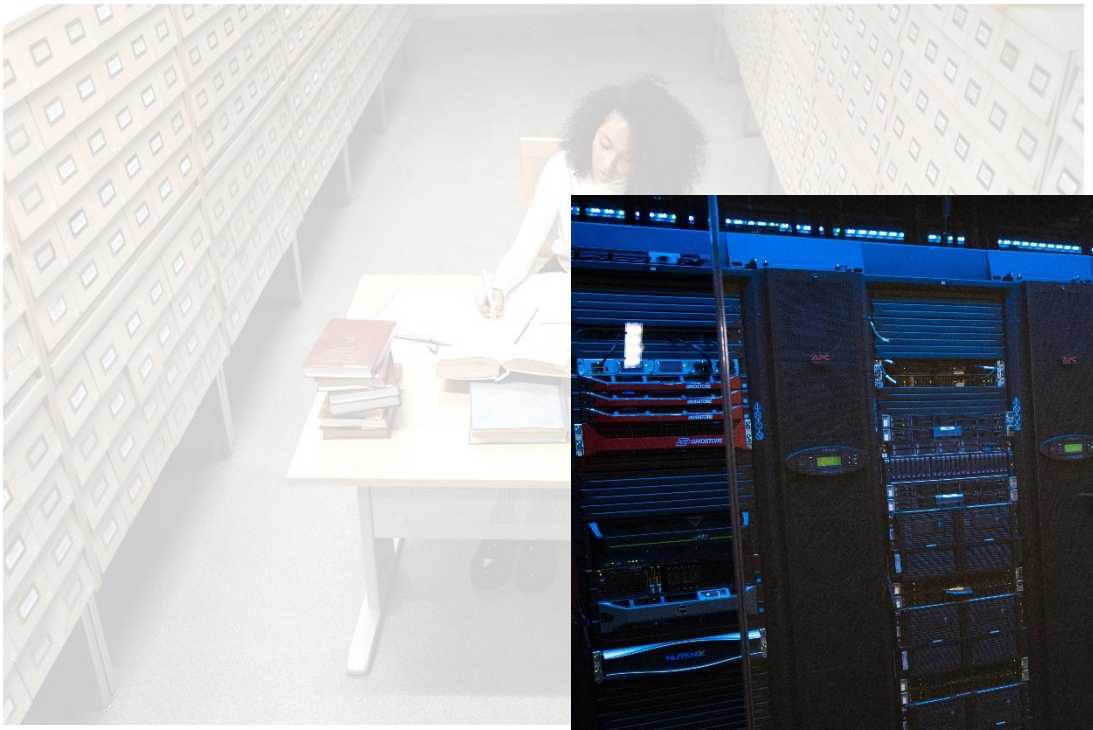


Photo by Tima Miroshnichenko from

Photo by Christina Morillo from Pexels (retrieved 2023-05-07)

What is Graph Data Science?

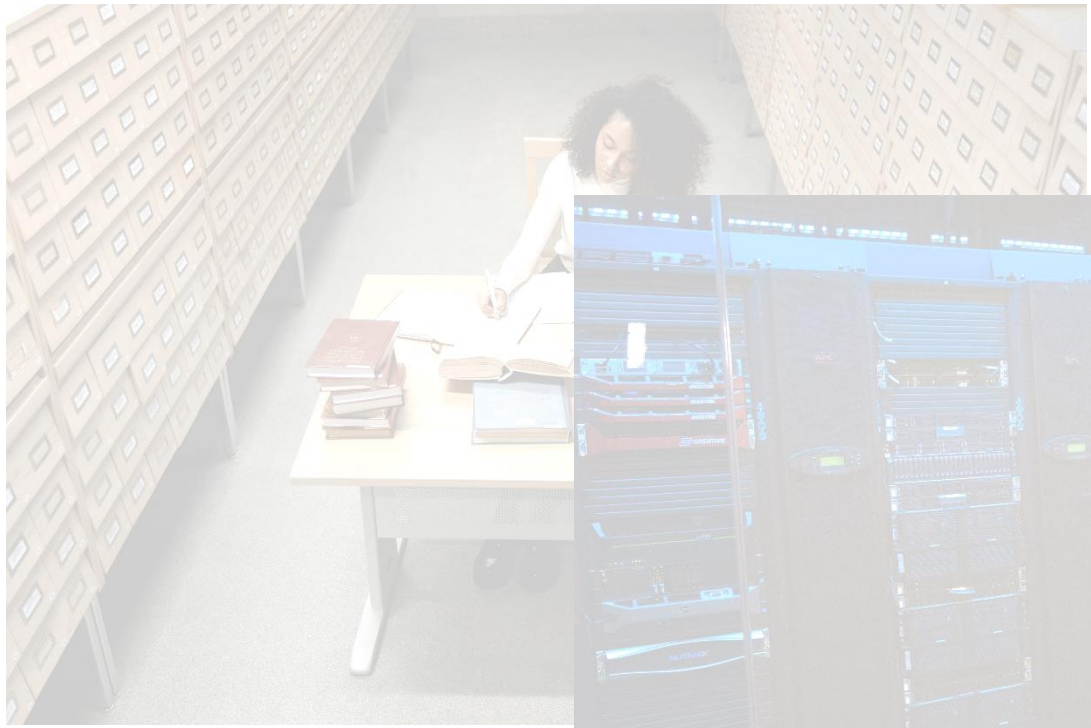


Photo by Tima Miroshnichenko from

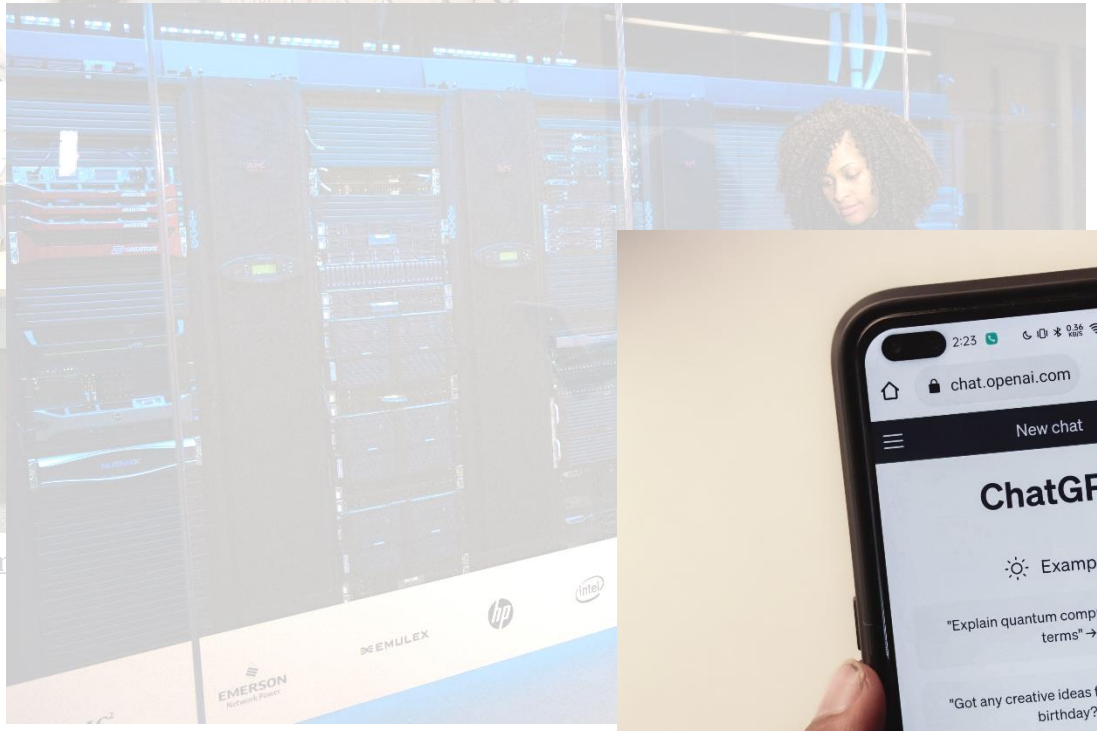


Photo by Christina Morillo from Pexels (

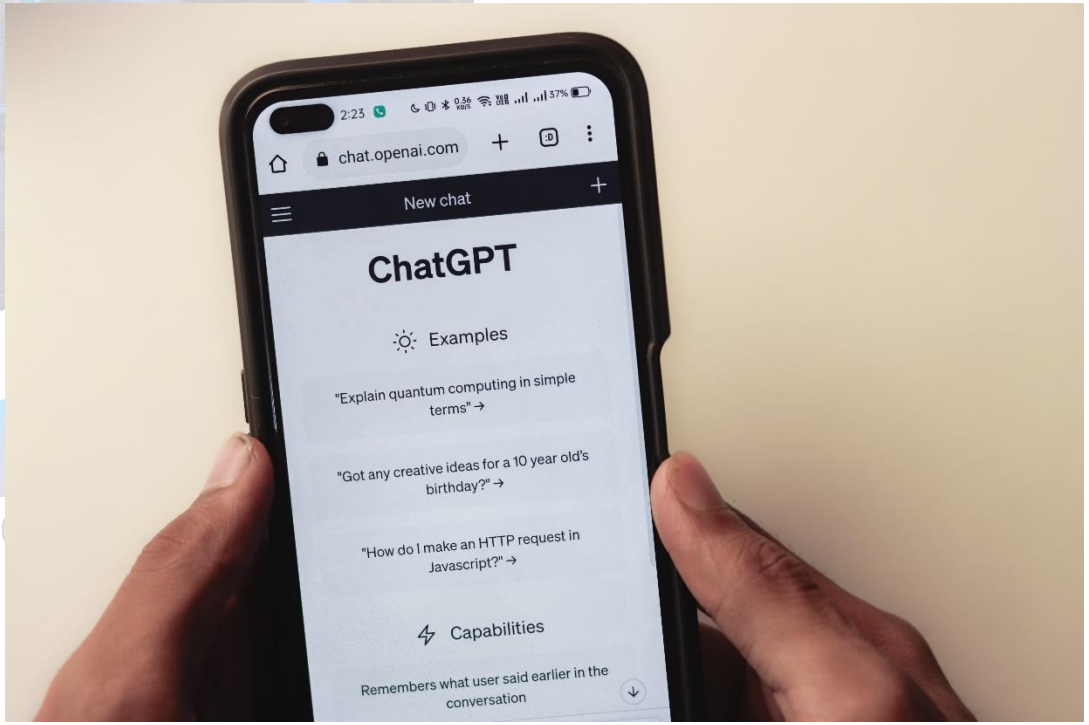


Photo by Sanket Mishra from Pexels (retrieved 2023-05-07)

What is Graph Data Science?

Data analysis

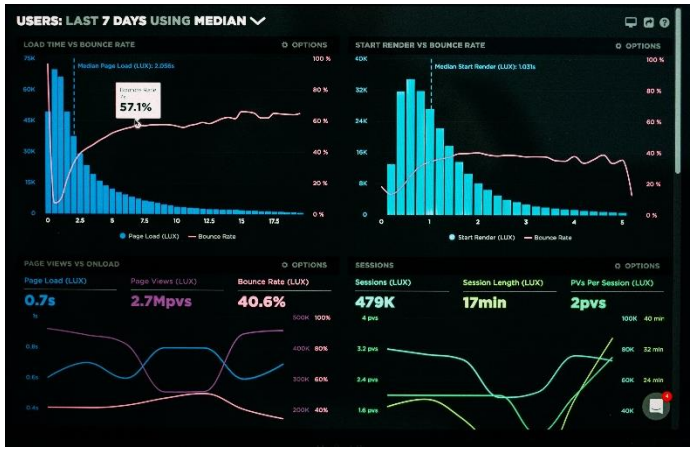


Photo by Luke Chesser on Unsplash (retrieved 2023-05-13)

What is Graph Data Science?

Data analysis

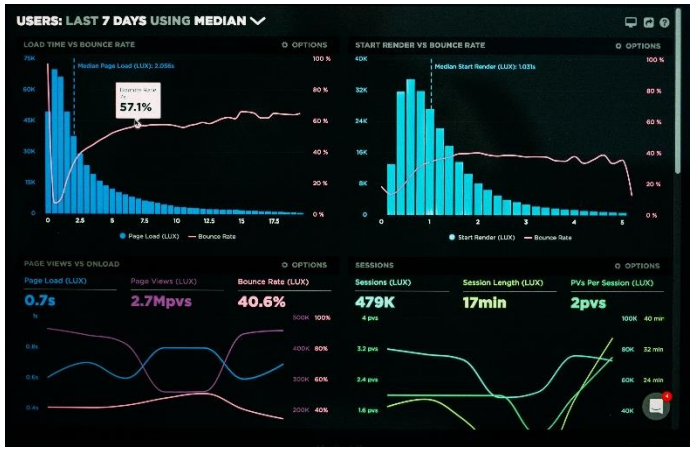


Photo by Luke Chesser on Unsplash (retrieved 2023-05-13)

Data bases & DevOps

Photo by Growtika on Unsplash (retrieved 2023-05-13)

What is Graph Data Science?

Data analysis

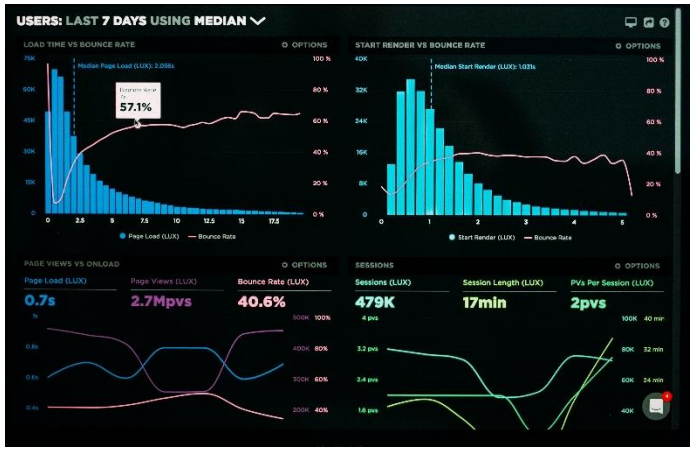


Photo by Luke Chesser on Unsplash (retrieved 2023-05-13)

AI and software dev

Photo by Christopher Gower on Unsplash (retrieved 2023-05-13)

Data bases & DevOps

Photo by Growtika on Unsplash (retrieved 2023-05-13)

What is Graph Data Science?

Data analysis

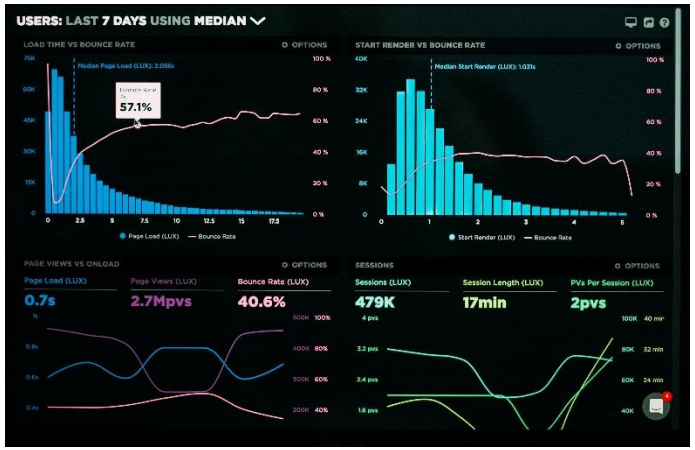


Photo by Luke Chesser on Unsplash (retrieved 2023-05-13)

AI and software dev

Photo by Christopher Gower on Unsplash (retrieved 2023-05-13)

Data bases & DevOps

Photo by Growtika on Unsplash (retrieved 2023-05-13)

Graphs / Networks

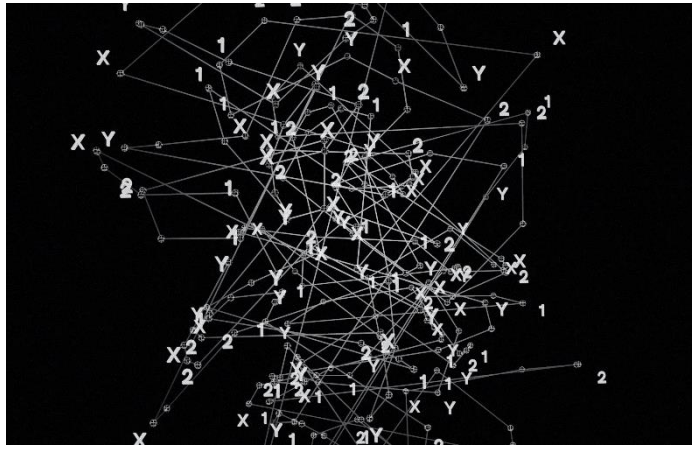
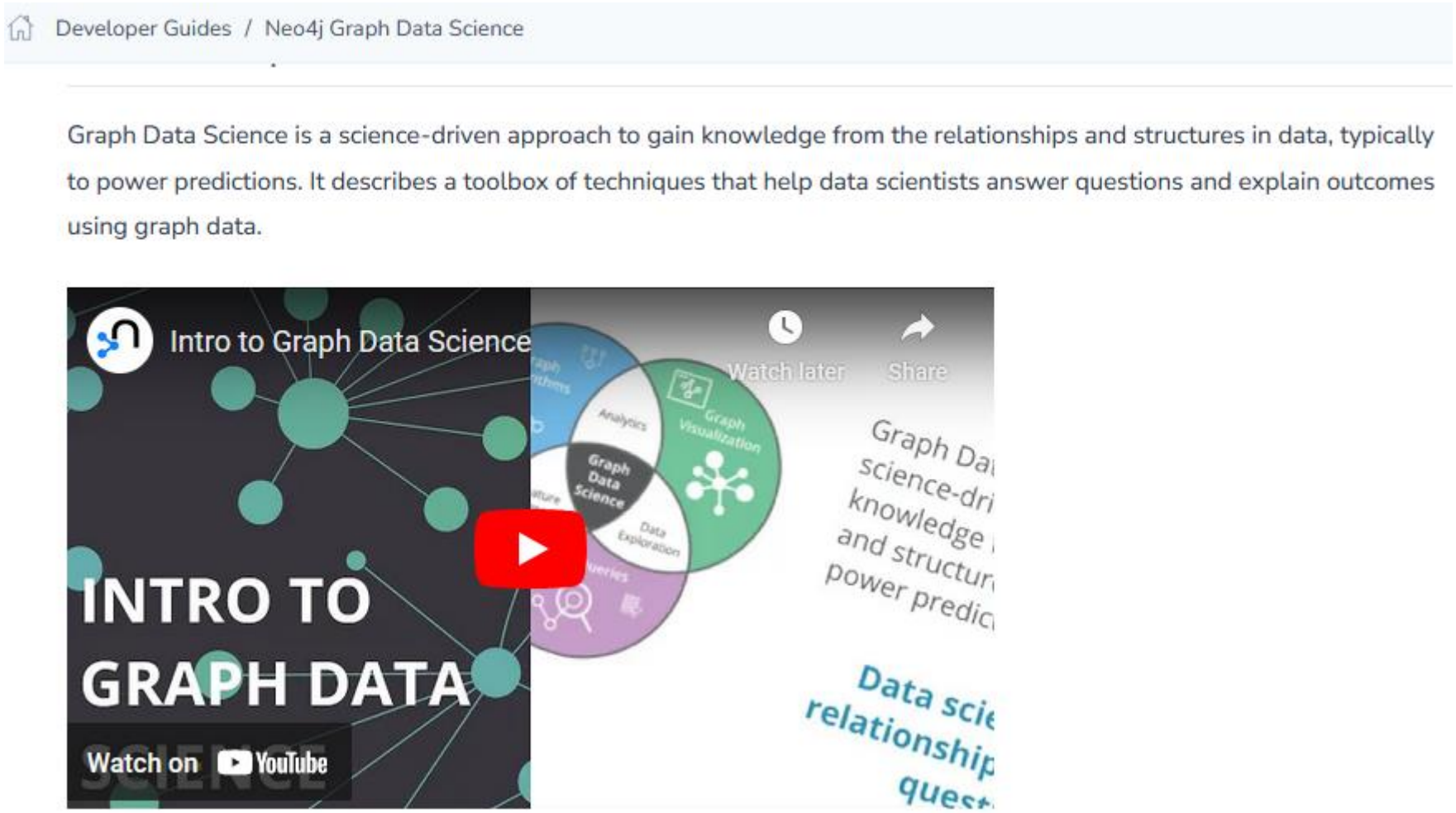


Photo by Resource Database on Unsplash (retrieved 2023-05-13)

What is Graph Data Science?

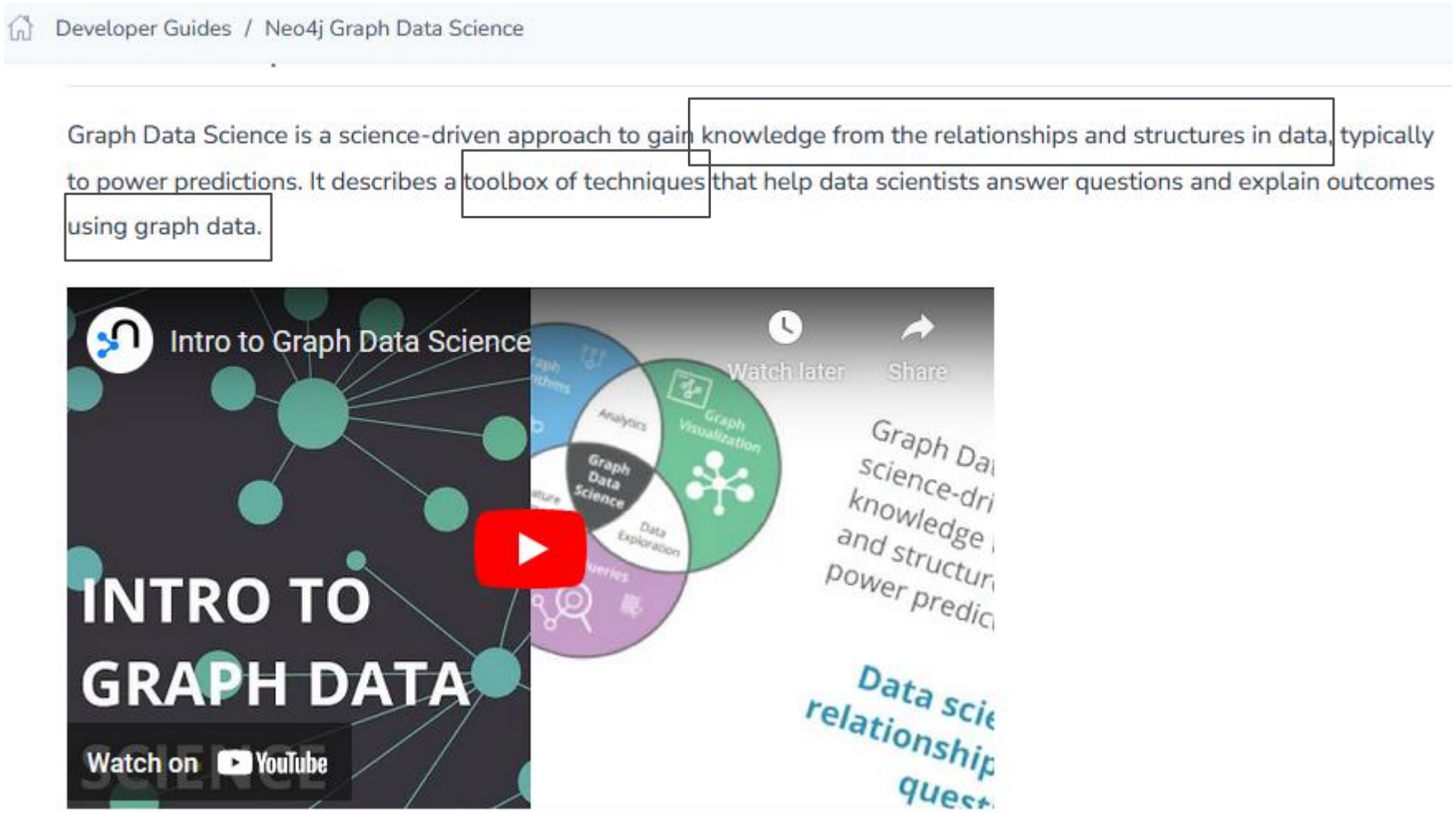
Graph Data Science term coined / promoted by Neo4j



Screenshot from Neo4J.com (retrieved 2023-05-13)

What is Graph Data Science?

Graph Data Science term coined / promoted by Neo4j



Screenshot from Neo4J.com (retrieved 2023-05-13)

What is Graph Data Science?

"Traditional Data Science"

"Graph Data Science"

What is Graph Data Science?

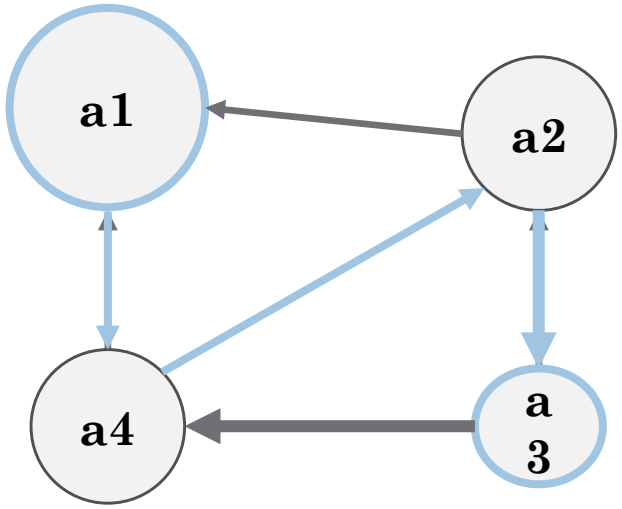
"Traditional Data Science"

structured data

	a1	a2	a3	a4
r1	0	0	0	1
r2	1	0	1	0
r3	0	1	0	1
r4	1	1	0	0

"Graph Data Science"

graph-structured data



What is Graph Data Science?

"Traditional Data Science"

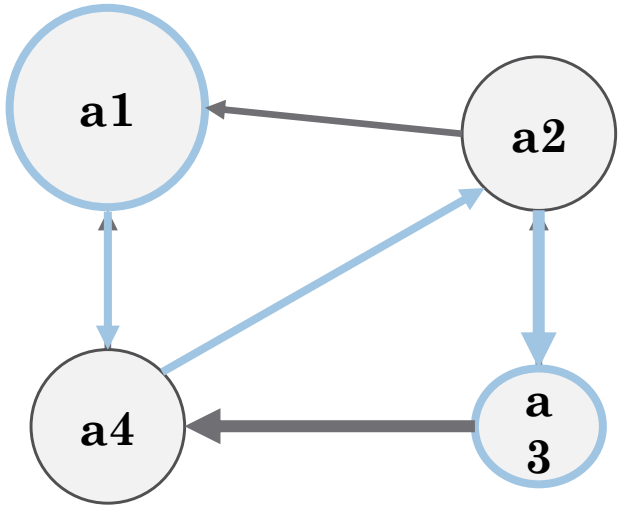
structured data

	a1	a2	a3	a4
r1	0	0	0	1
r2	1	0	1	0
r3	0	1	0	1
r4	1	1	0	0

- Databases
- Neural Networks
- Data mining

"Graph Data Science"

graph-structured data



- Graph databases
- Graph Neural Networks
- Graph mining

What is Graph Data Science?

"Traditional Data Science"

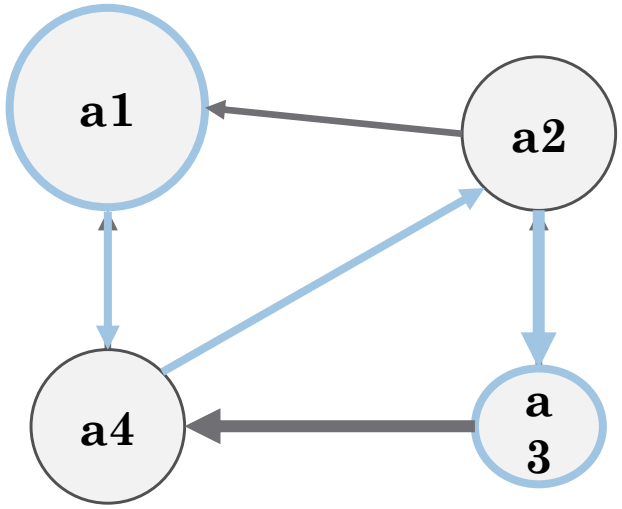
structured data

	a1	a2	a3	a4
r1	0	0	0	1
r2	1	0	1	0
r3	0	1	0	1
r4	1	1	0	0

- Databases
- Neural Networks
- Data mining

"Graph Data Science"

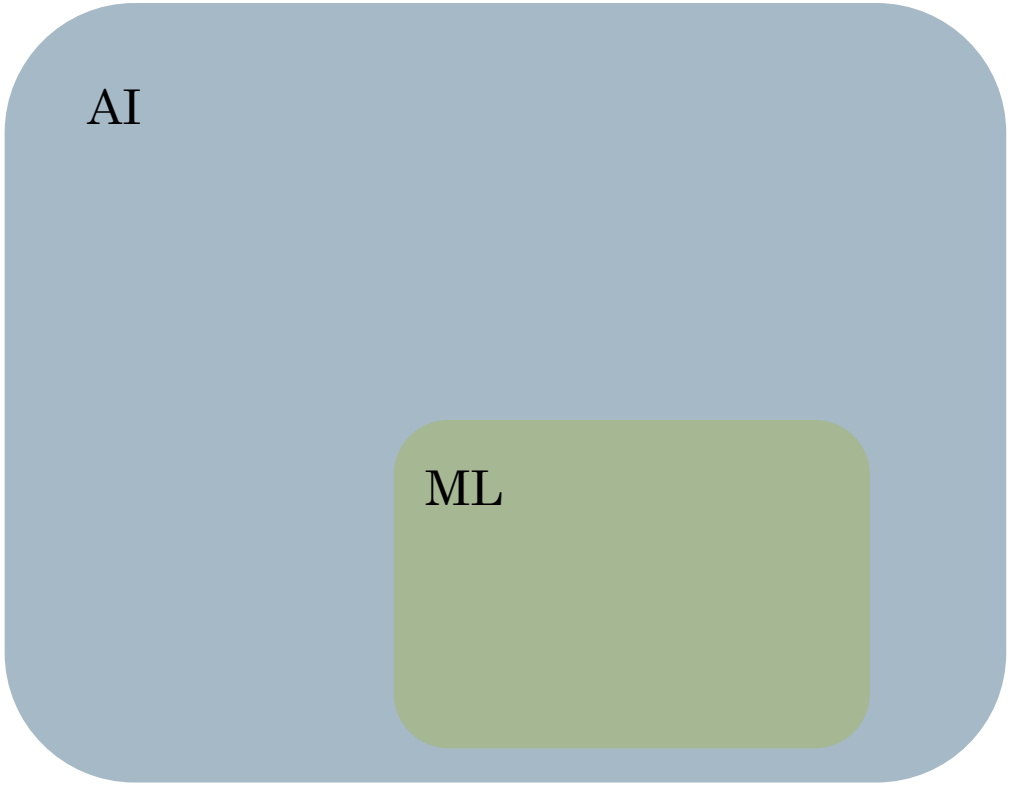
graph-structured data



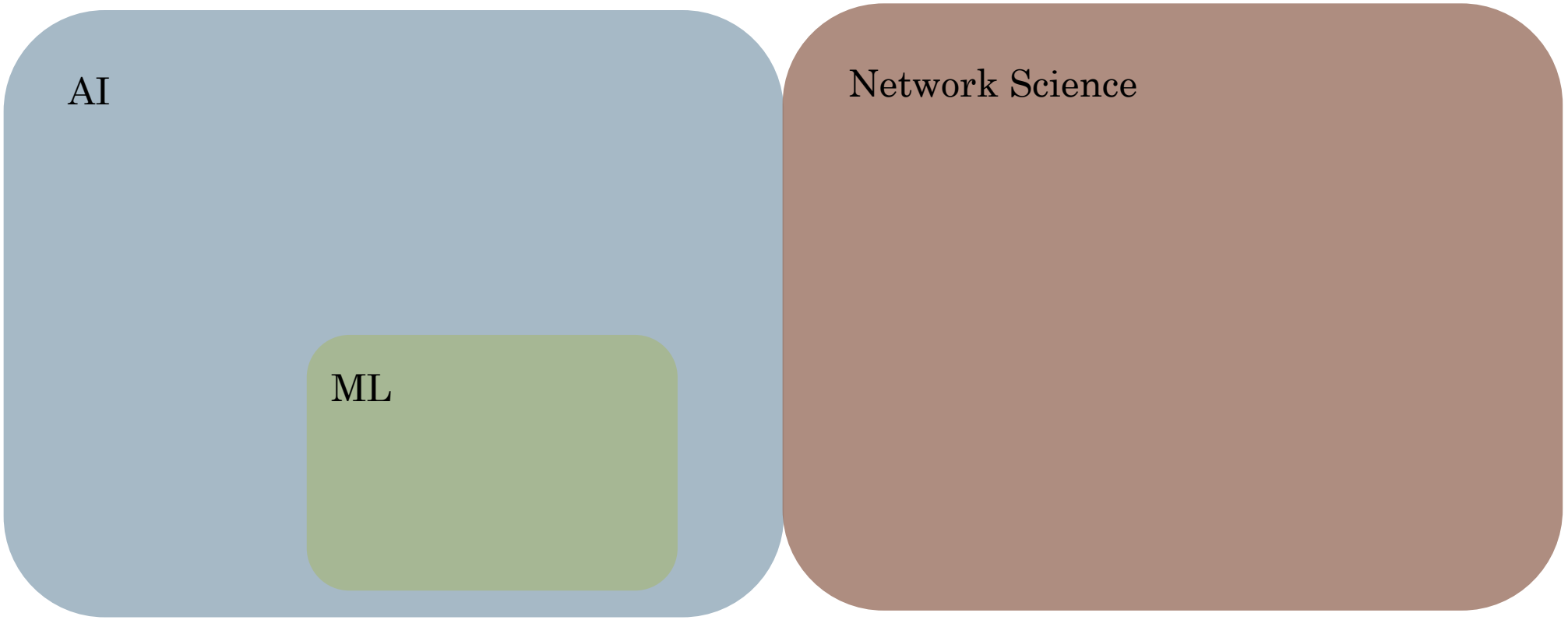
- Graph databases
- Graph Neural Networks
- Graph mining

GDS = Data Science enriched with complex data relations

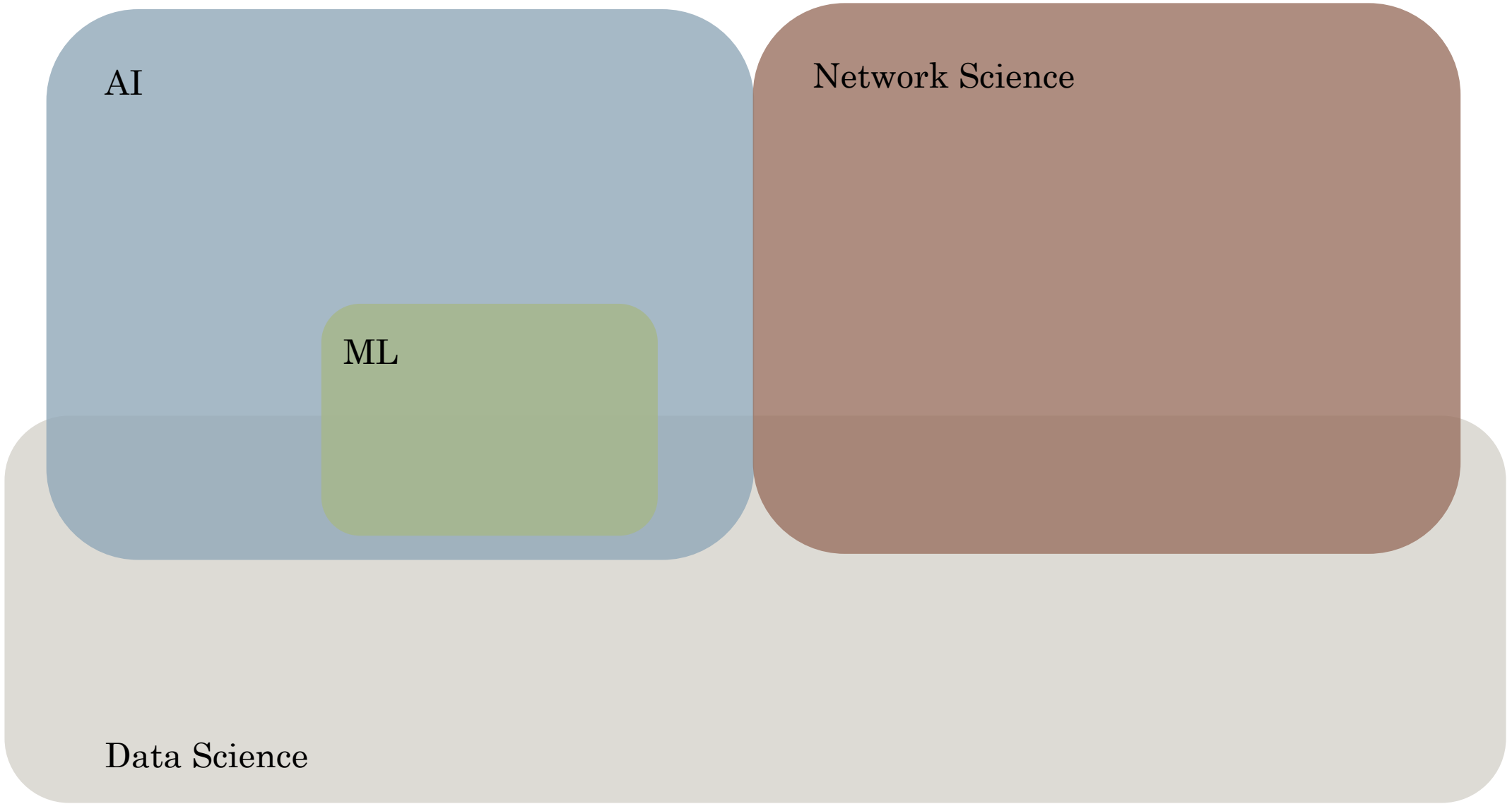
What is Graph Data Science?



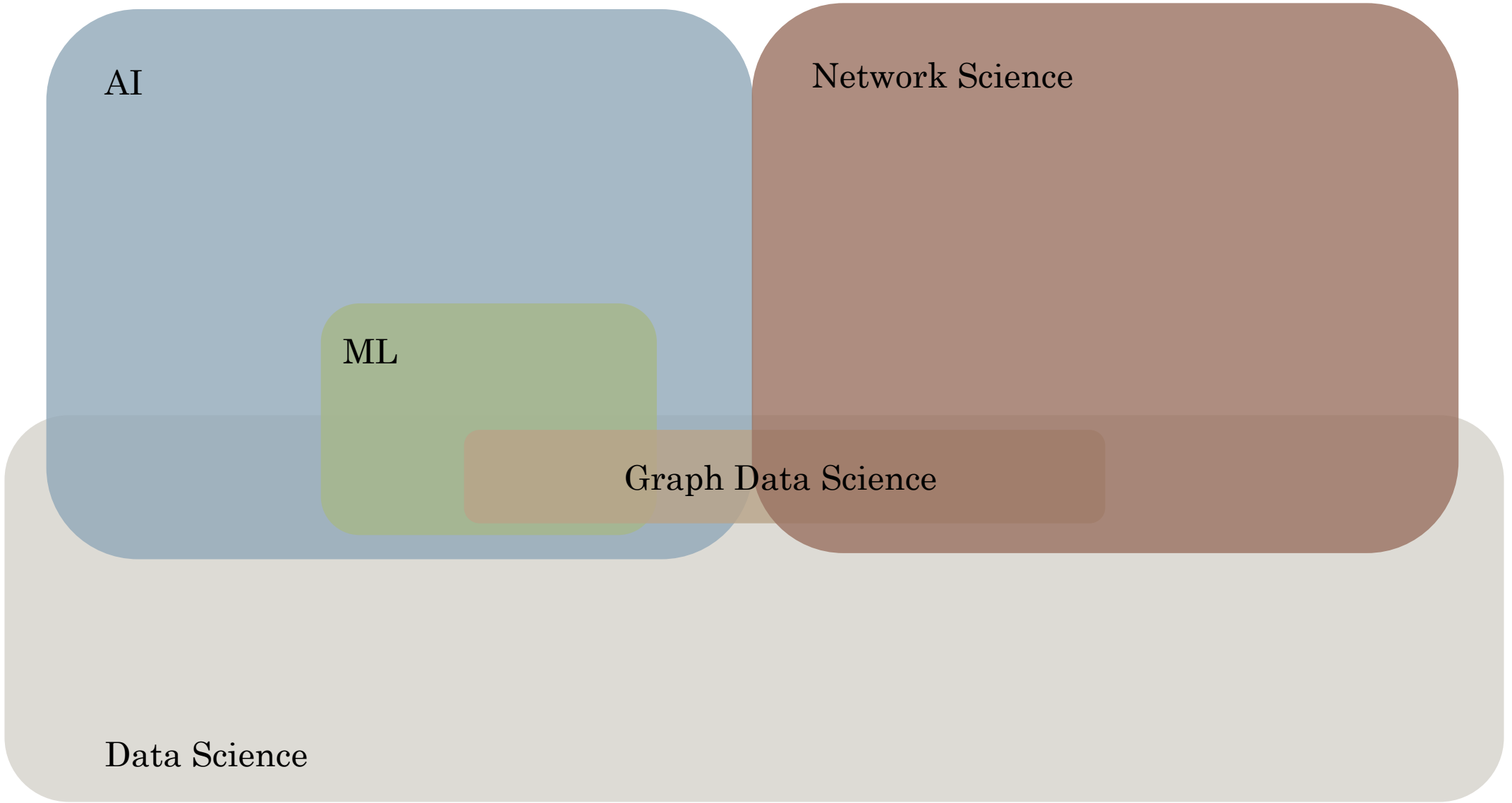
What is Graph Data Science?



What is Graph Data Science?

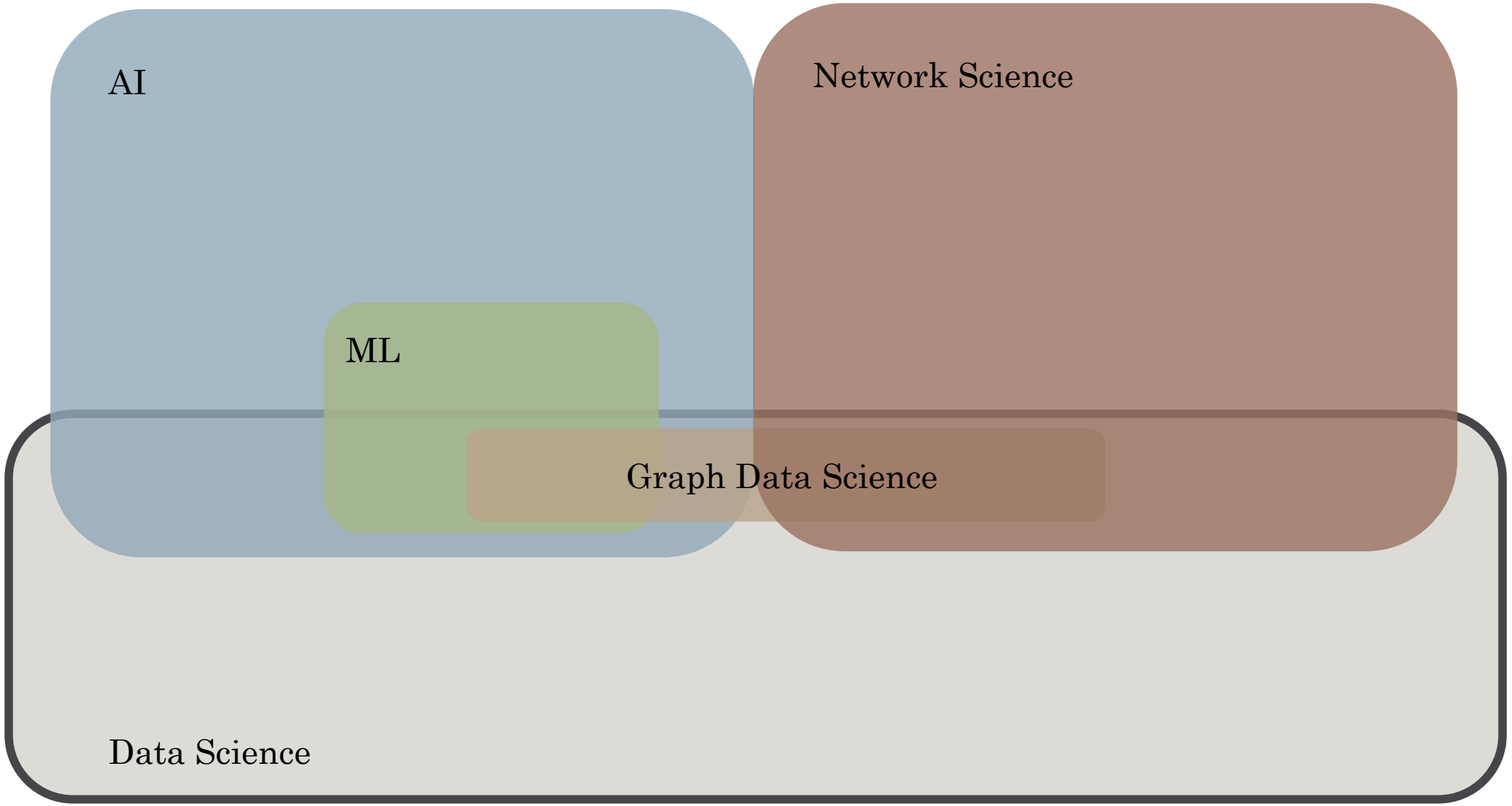


What is Graph Data Science?



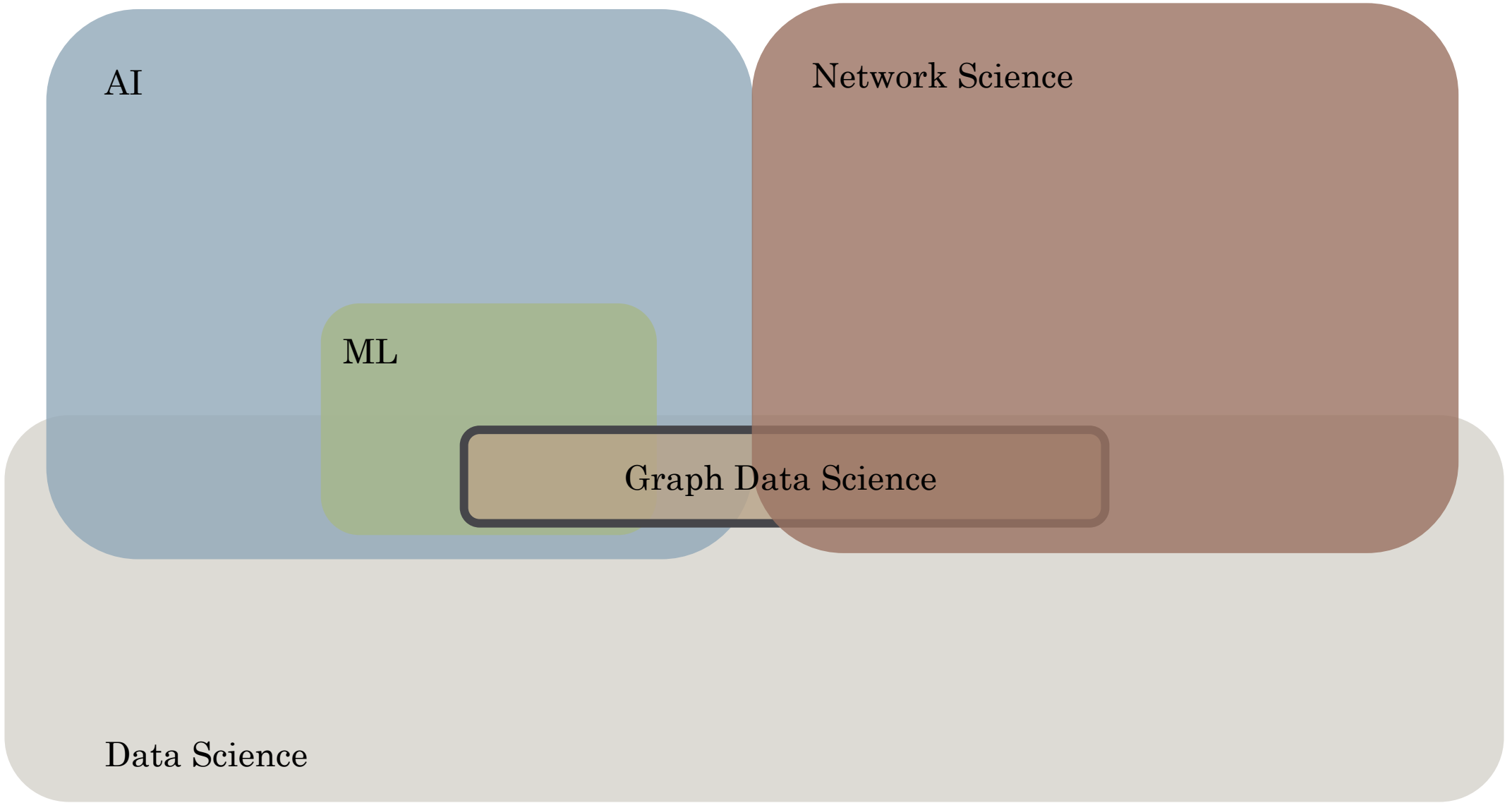
Who is a Graph Data Scientist?

Data Scientist

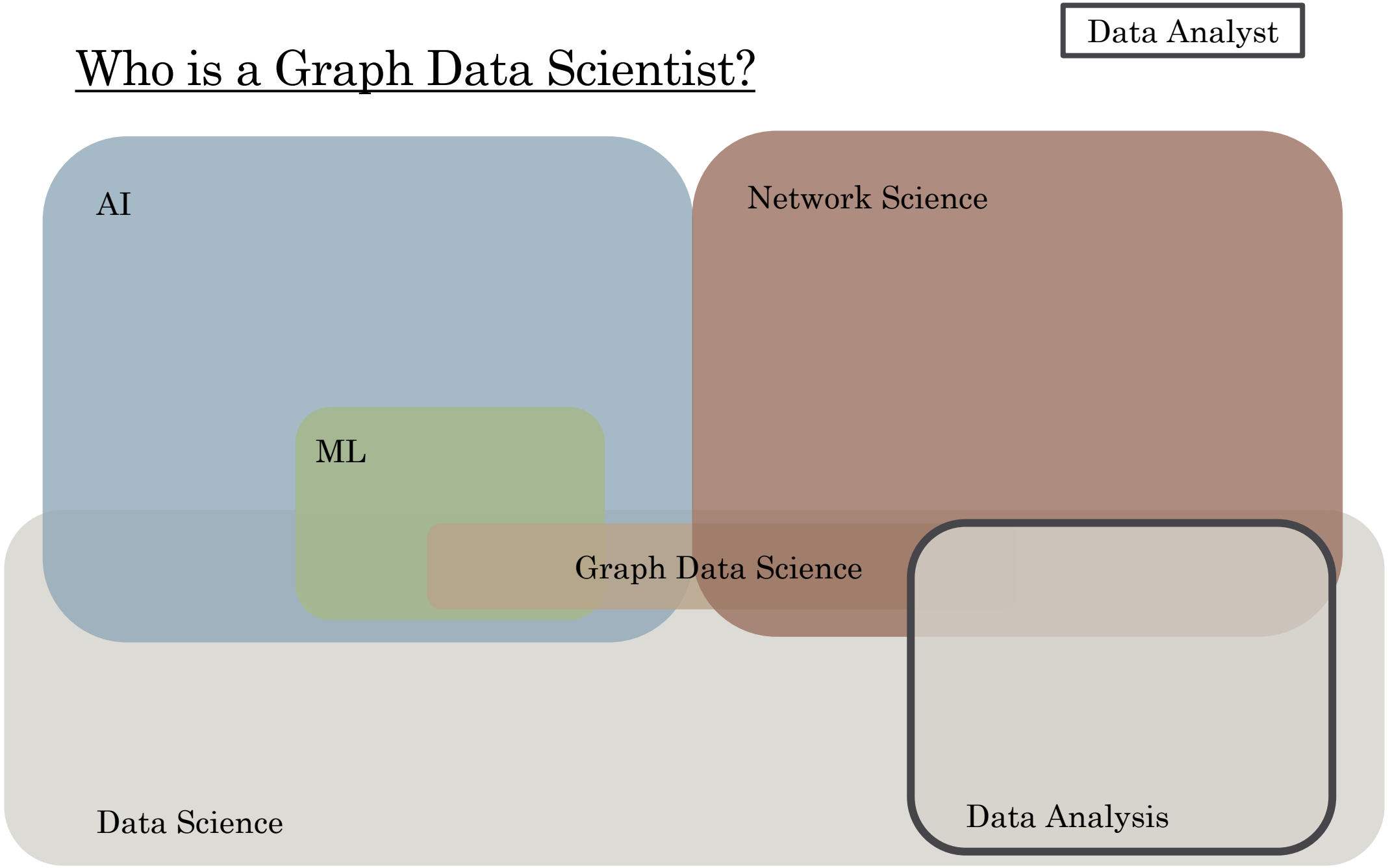


Who is a Graph Data Scientist?

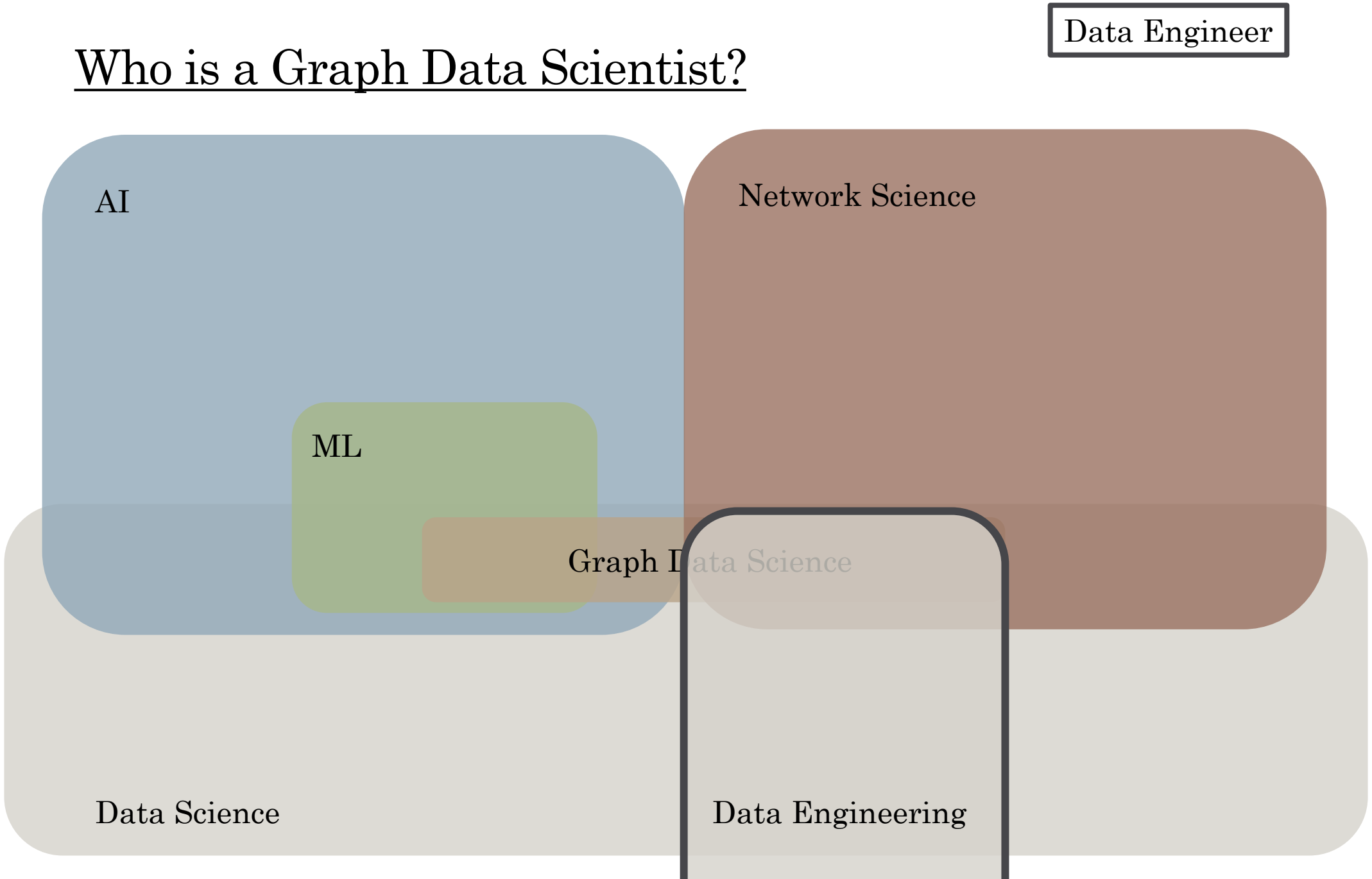
Graph Data Scientist



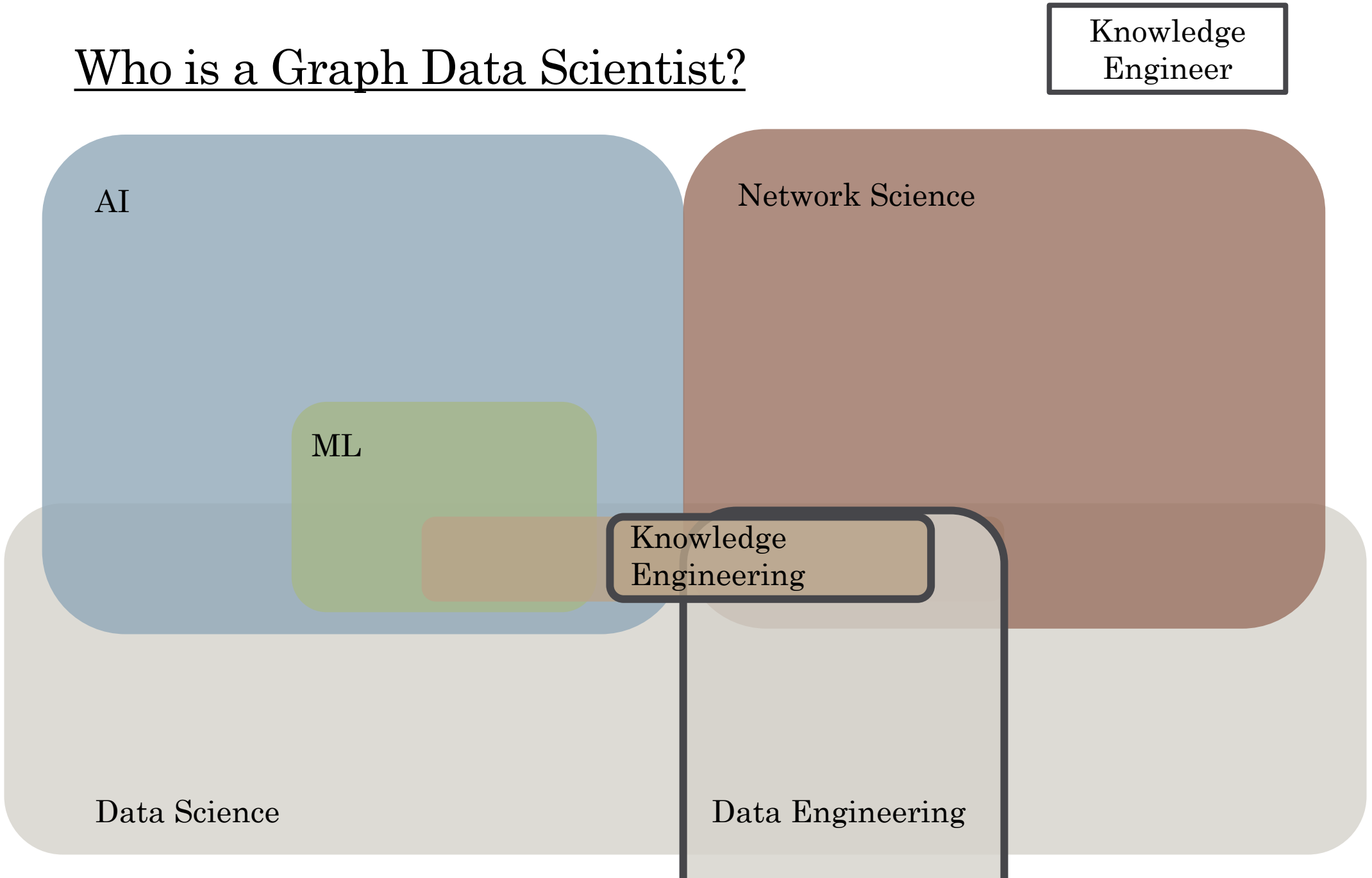
Who is a Graph Data Scientist?



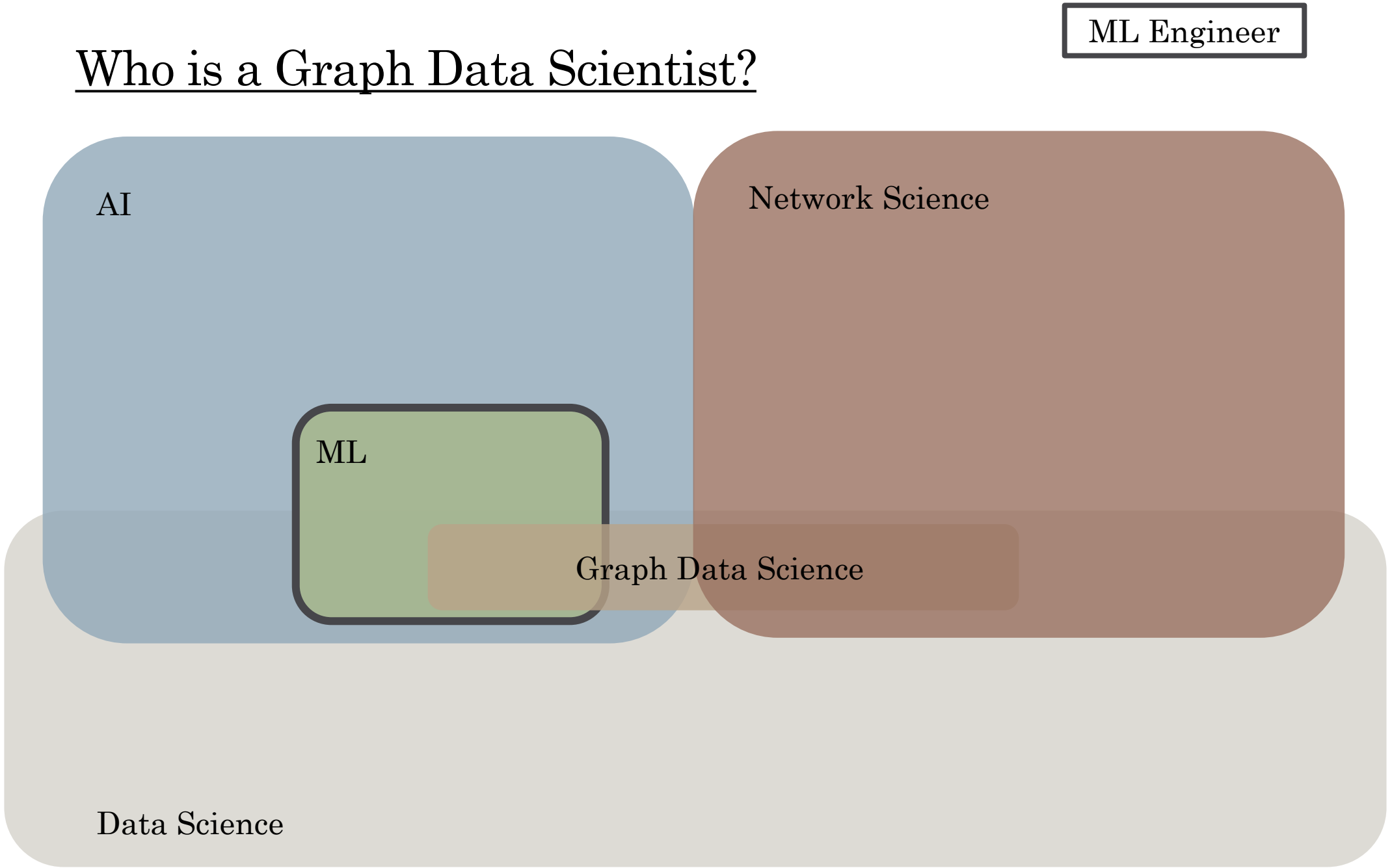
Who is a Graph Data Scientist?



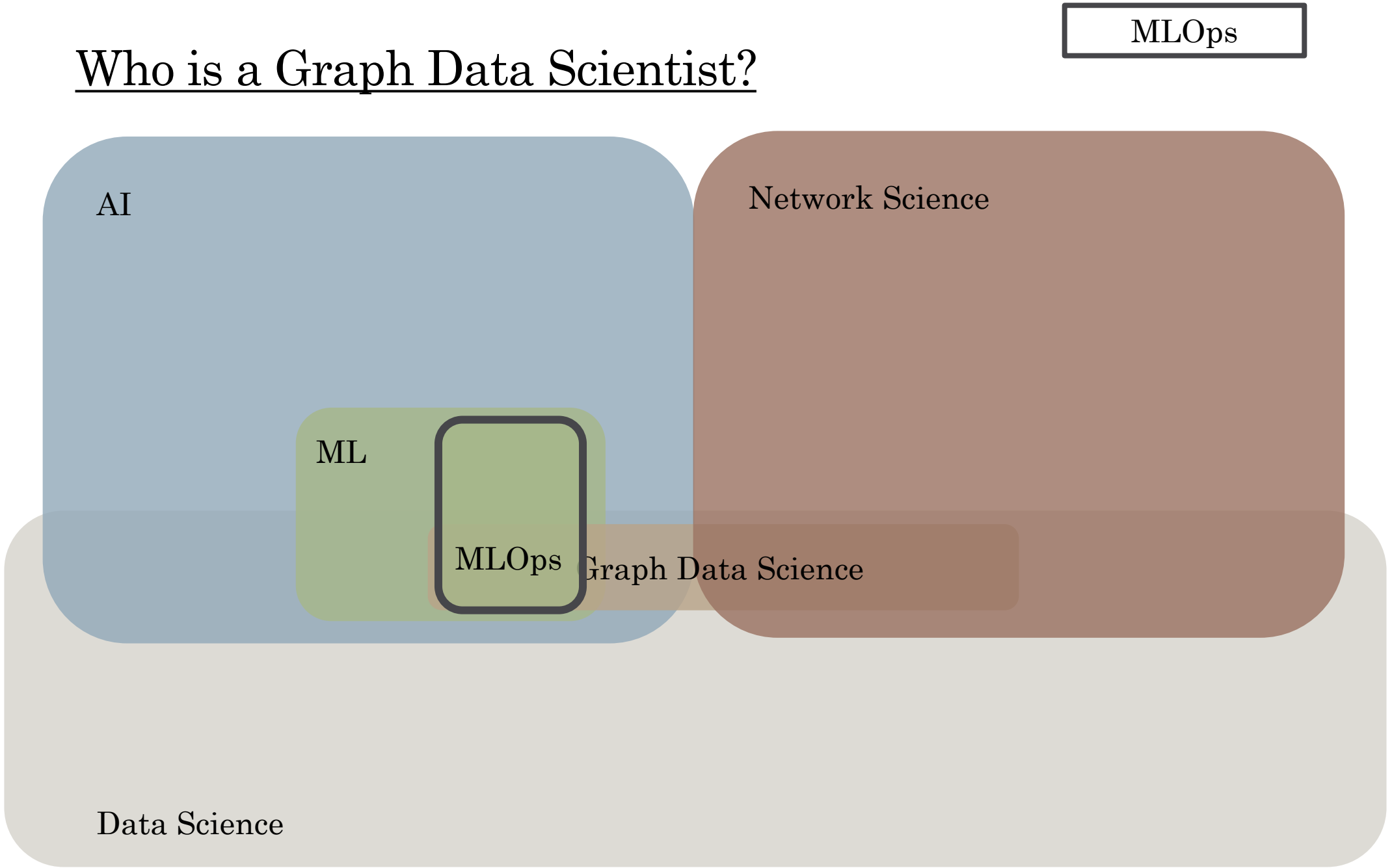
Who is a Graph Data Scientist?



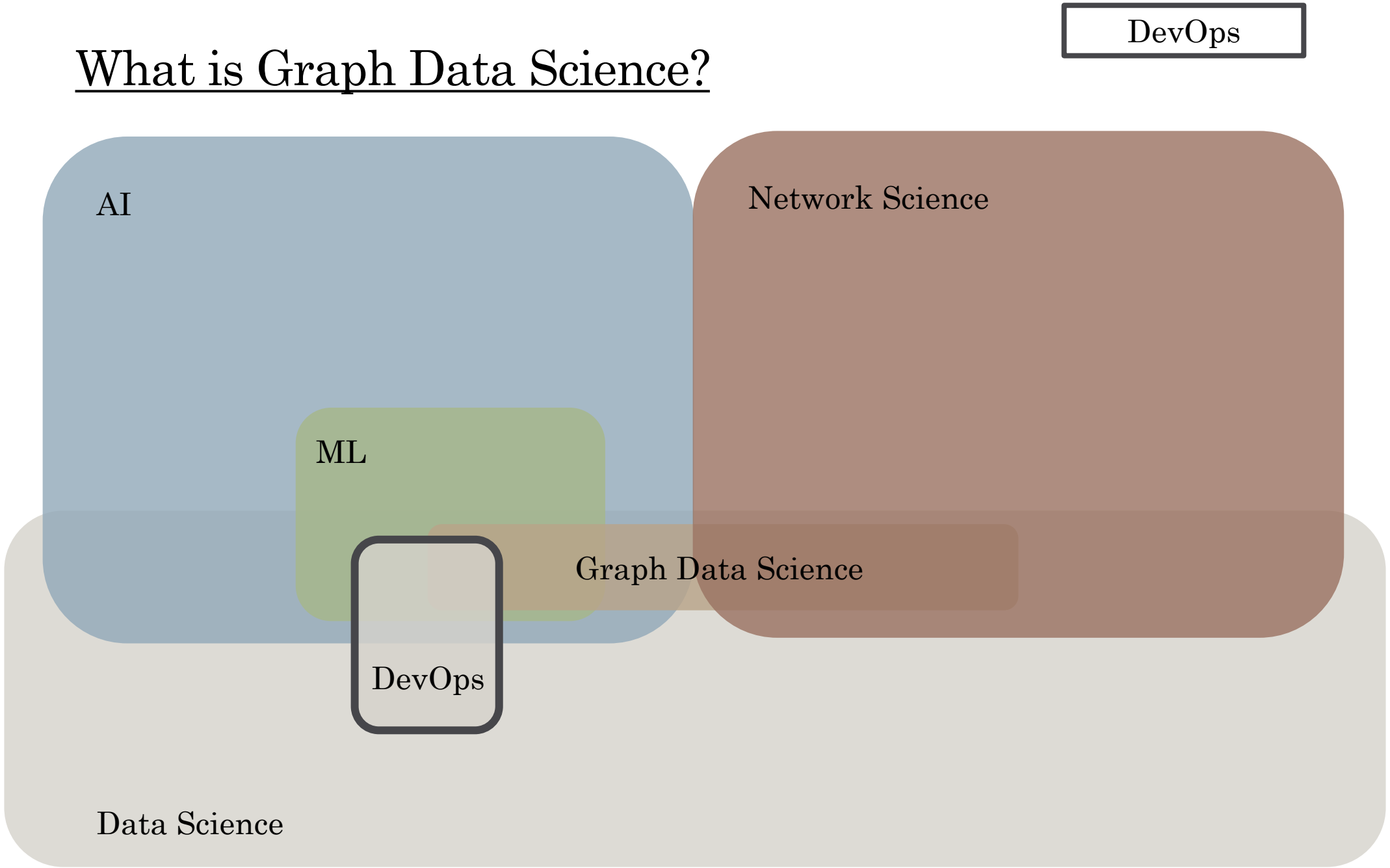
Who is a Graph Data Scientist?



Who is a Graph Data Scientist?



What is Graph Data Science?



A work week in the life of a GDS

Monday	Tuesday	Wednesday	Thursday	Friday
Emails / chat	Emails / chat	Emails / chat	Emails / chat	Emails / chat
• Data analysis • AIML dev • Other software dev • Debugging	Team call Other call/email	Project support activities (e.g. documentation, maintenance, git, venvs, gpu)	• Data analysis • AIML dev • Software dev • Debugging • R&D	Continuous learning Other call/email
Lunch/office chat	Lunch/office chat	Lunch/office chat	Lunch/office chat	Lunch/office chat
Calls / meetings • Manager • Stakeholders • Business • Peers • Knowledge sharing	• Data analysis • AIML dev • Other software dev • Debugging • R&D	Alignment calls Data-related work (engineering, DevOps)	Other call/email Documentation Preparing slides Dashboard dev	• Data analysis • AIML dev • Other software dev • Debugging • R&D

A typical Data Science project workflow

1. Identify the problem / a business need

Icons from Flaticon (retrieved 2023-05-13)

A typical Data Science project workflow

1. Identify the problem / a business need
2. Formulate an idea in analytical ways, think of a solution
3. Implement a small prototype for the solution

Icons from Flaticon (retrieved 2023-05-13)

A typical Data Science project workflow

MONTHS

1. Identify the problem / a business need
2. Formulate an idea in analytical ways, think of a solution
3. Implement a small prototype for the solution
4. Prepare a PoC

Icons from Flaticon (retrieved 2023-05-13)

A typical Data Science project workflow

MONTHS

1. Identify the problem / a business need
2. Formulate an idea in analytical ways, think of a solution
3. Implement a small prototype for the solution

4. Prepare a PoC

5. Promote, validate with business, get feedback, evaluate
6. Improve and test again

Icons from Flaticon (retrieved 2023-05-13)

A typical Data Science project workflow

MONTHS

1. Identify the problem / a business need
2. Formulate an idea in analytical ways, think of a solution
3. Implement a small prototype for the solution

4. Prepare a PoC

A YEAR

5. Promote, validate with business, get feedback, evaluate
6. Improve and test again
7. Create MVP

Icons from Flaticon (retrieved 2023-05-13)

A typical Data Science project workflow

MONTHS

1. Identify the problem / a business need
2. Formulate an idea in analytical ways, think of a solution
3. Implement a small prototype for the solution

4. Prepare a PoC

A YEAR

5. Promote, validate with business, get feedback, evaluate
6. Improve and test again
7. Create MVP

8. Test and improve

Icons from Flaticon (retrieved 2023-05-13)

A typical Data Science project workflow

MONTHS

1. Identify the problem / a business need
2. Formulate an idea in analytical ways, think of a solution
3. Implement a small prototype for the solution

4. Prepare a PoC

A YEAR

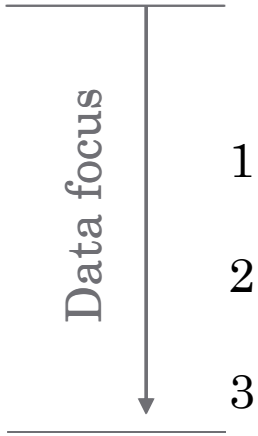
5. Promote, validate with business, get feedback, evaluate
6. Improve and test again
7. Create MVP

YEARS

8. Test and improve
9. Monitor and maintain

Icons from Flaticon (retrieved 2023-05-13)

A typical data modelling pipeline



1. Data collection
2. Data understanding
3. Data cleaning and processing

A typical data modelling pipeline

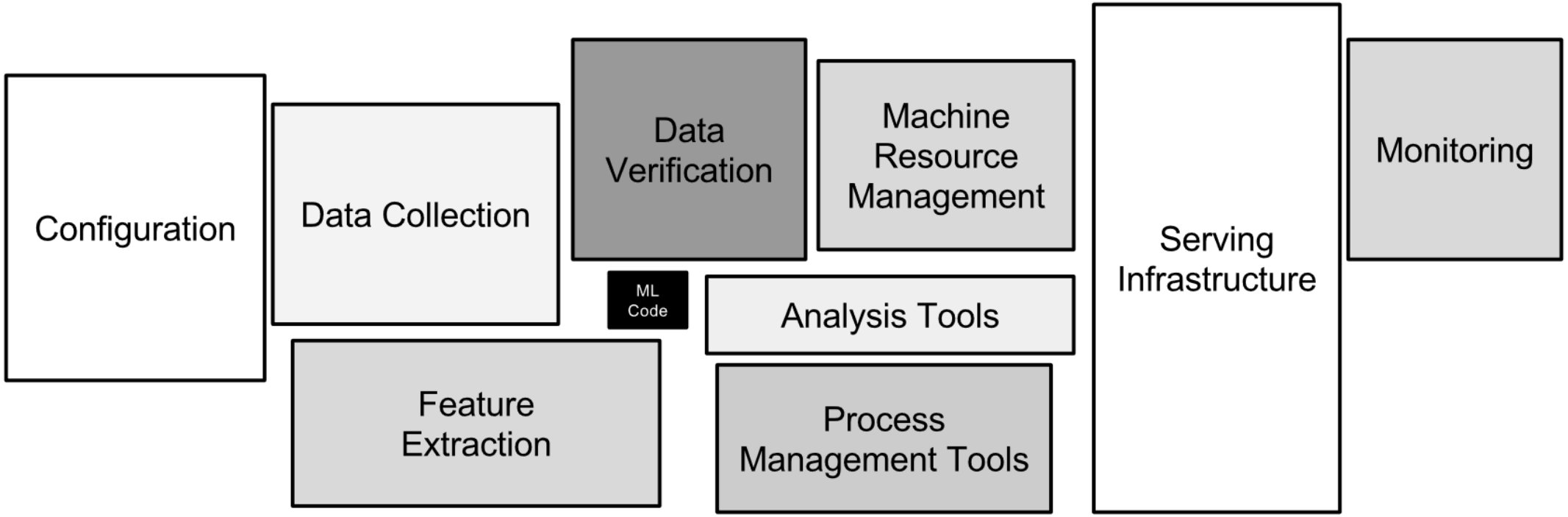
Data focus

1. Data collection
2. Data understanding
3. Data cleaning and processing

Data **and** model focus

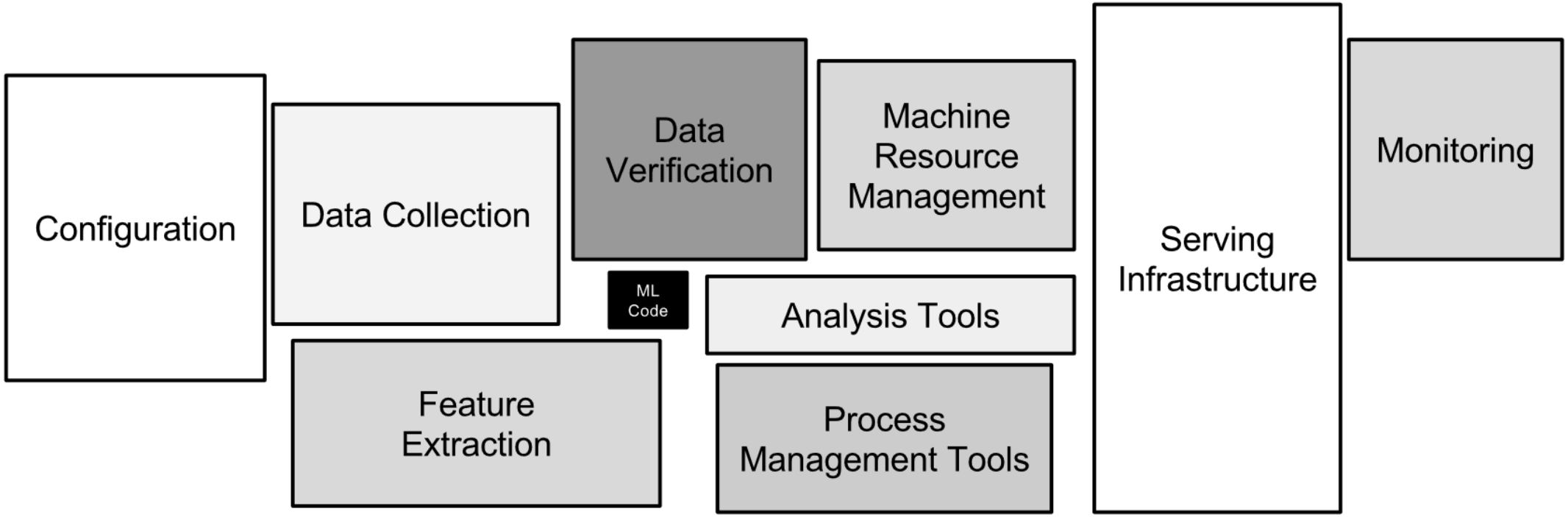
4. Model development and training
5. Model testing and business validation
6. Model deployment (and/or dashboard deployment)
7. Model maintenance (incl. retraining + retesting)

(Hidden) complexity of a data science project



Sculley *et al.*, NeurIPS **28** (2015)

(Hidden) complexity of a data science project

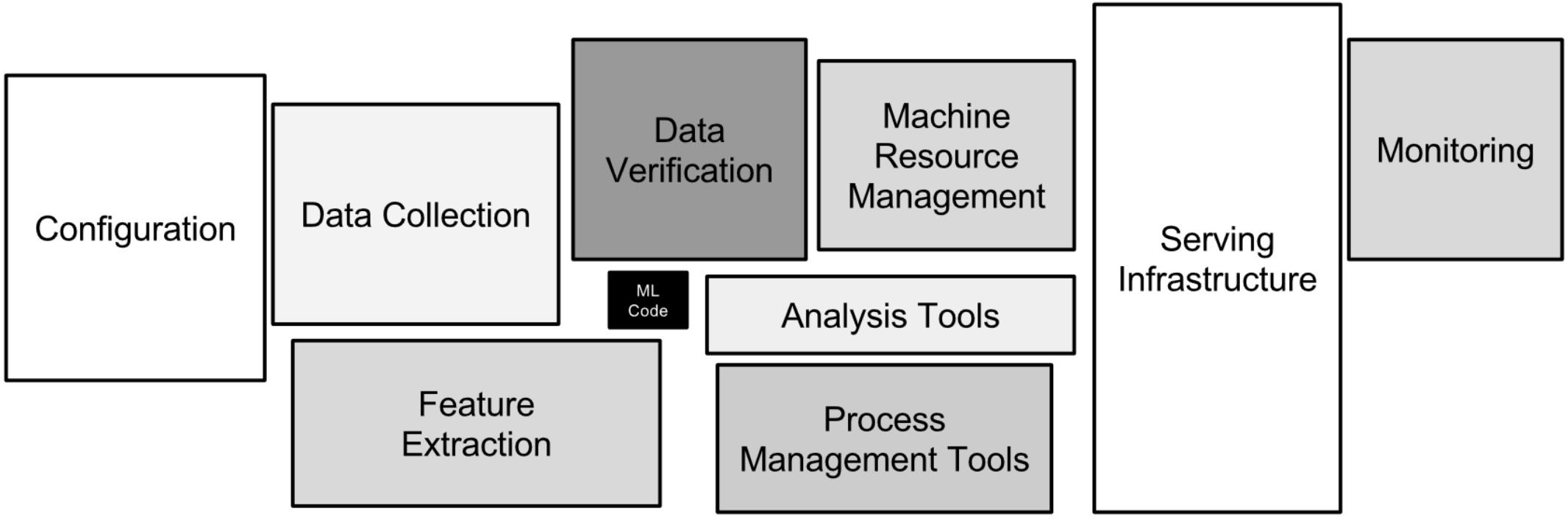


Sculley *et al.*, NeurIPS **28** (2015)

Additional conditions for success:

- IT support (infrastructure, model consumption, platforms)

(Hidden) complexity of a data science project

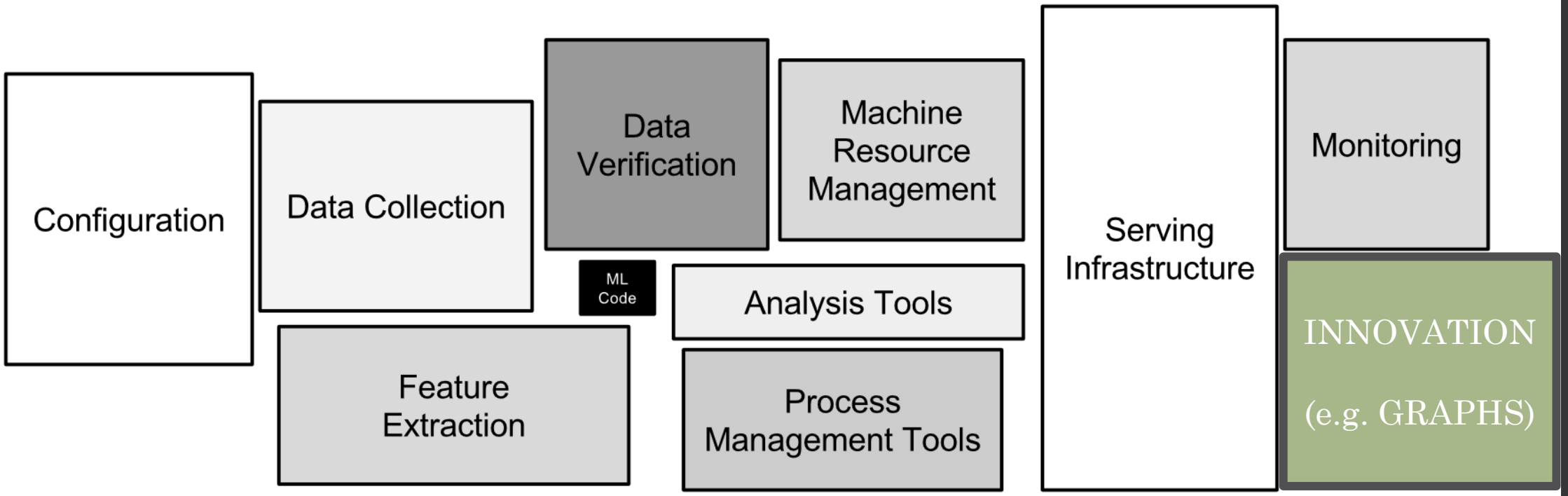


Sculley *et al.*, NeurIPS **28** (2015)

Additional conditions for success:

- IT support (infrastructure, model consumption, platforms)
- Business support (coordination, prioritization, visibility and model promotion, enablement, change management, sponsorship, managing expectations)

(Hidden) complexity of a data science project



Sculley *et al.*, NeurIPS **28** (2015)

Additional conditions for success:

- IT support (infrastructure, model consumption, platforms)
- Business support (coordination, prioritization, visibility and model promotion, enablement, change management, sponsorship, managing expectations)