



October 26, 2023

- 3 tracks:
 - Building Applications & APIs
 - ML & AI with Graphs
 - Powerful Visualizations
- 24 hours
- 100 talks

Register for free! - neo4j.com/nodes

neo4j

Building a graph solution

Using a real world digital twin data set

neo4j

Jean-Marc Guerin Jesus Barrasa

Neo4j Field Engineering Team

Agenda

1. Logistics
2. Introduction
3. Use Case Explanation
4. Modeling
5. Building the solution
6. Q & A

Logistics

Material for the workshop:

<https://github.com/guerinjeanmarc/graph-day-bangalore>

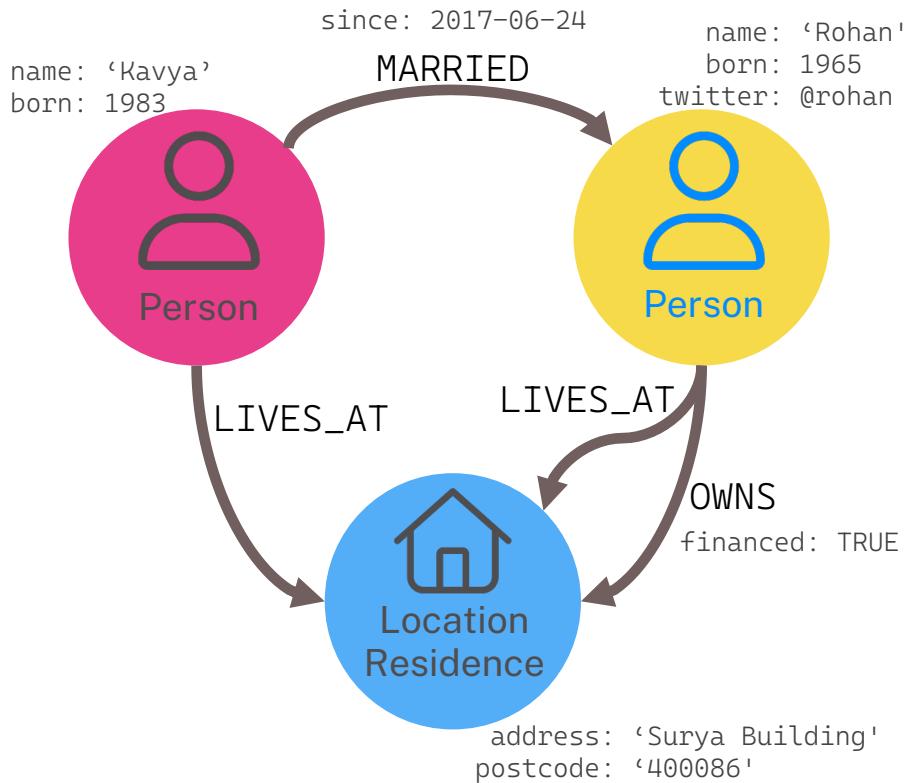


Introduction

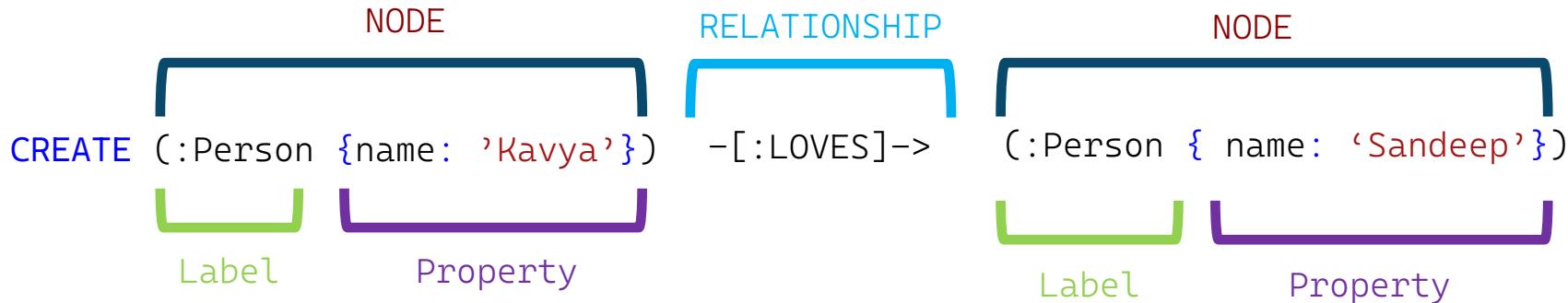
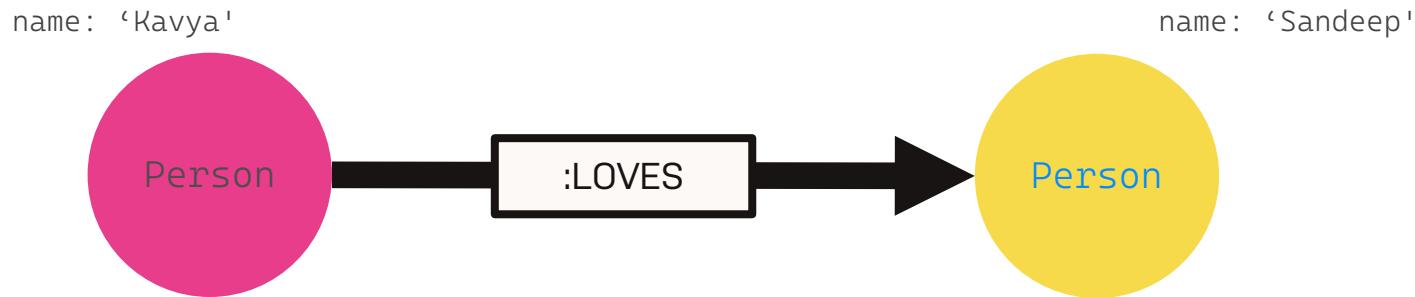
A short overview of the Neo4j Product

Labeled Property Graph Model Components

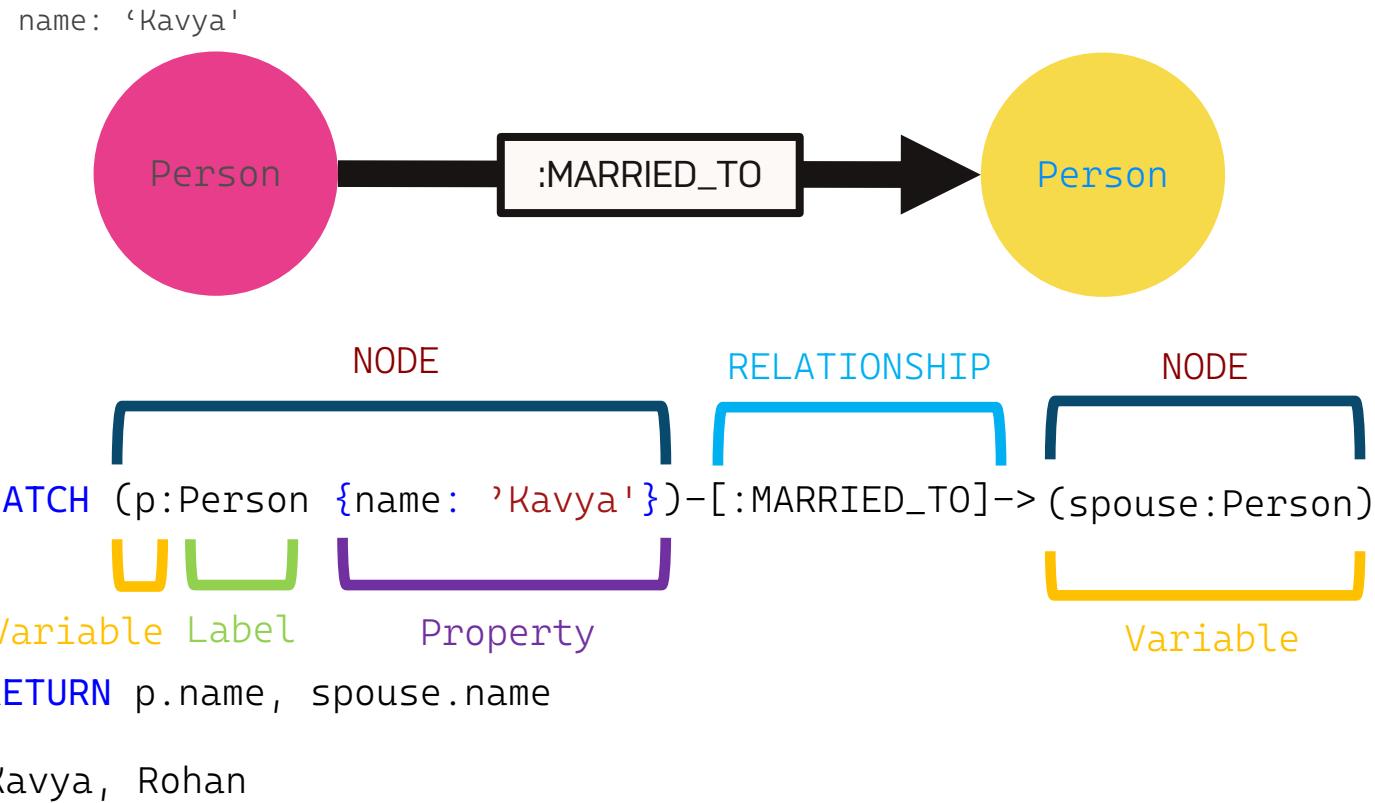
- **Nodes**
 - Represent objects in the graph
- **Labels**
 - Group nodes
 - Shape the domain
- **Relationships**
 - Relate nodes by type and direction
- **Properties**
 - Name-value pairs that can go on nodes and relationships
 - Can have indexes and composite indexes
 - (types: String, Number, Long, Date, Spatial, byte and arrays of those)



Cypher: powerful and expressive query language

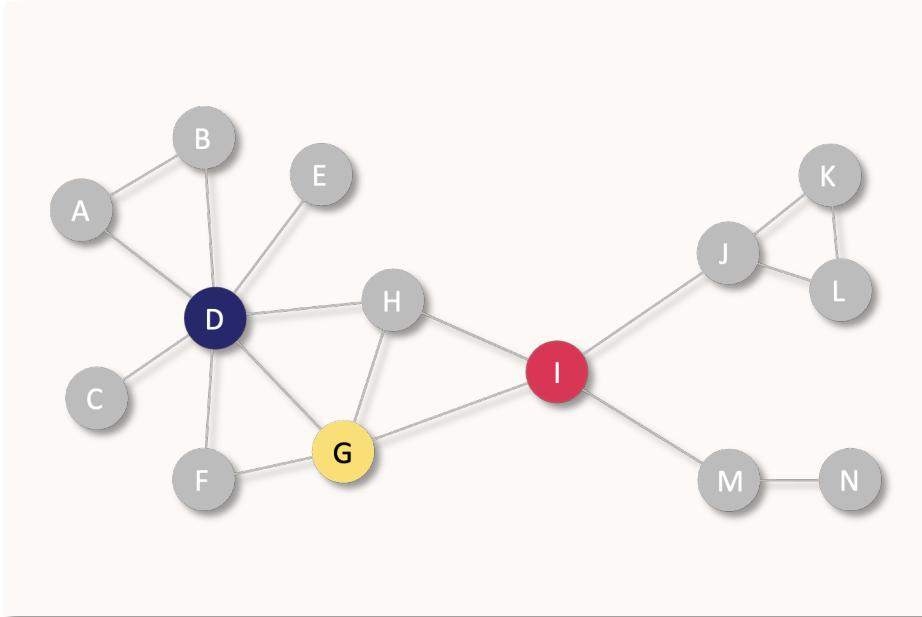


Cypher: Matching



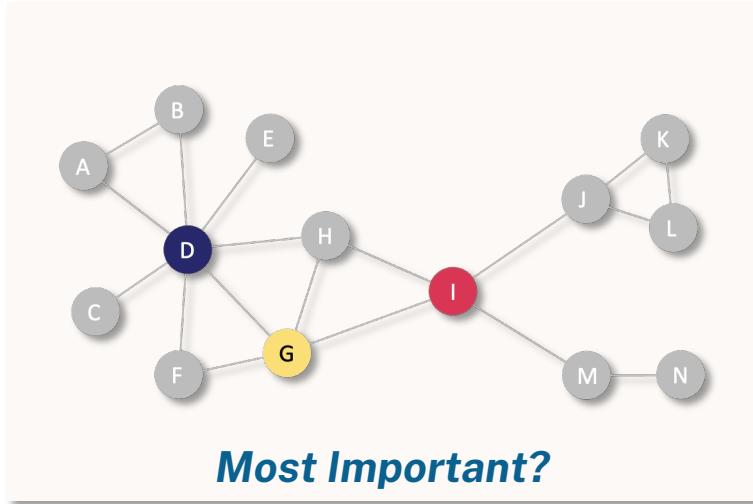
Neo4j Graph Data Science

Pop Quiz



Which of the coloured nodes would be considered the most “important”?

Graphs Contain Implicit Knowledge



D

D has the highest degree centrality (7)

This is the most connected individual in the network. If important is how well you are personally known, you pick D.

G

G has the highest closeness centrality (0.52)

Information will disperse through the network more quickly through this individual. If you need to get a message out rapidly, choose G.

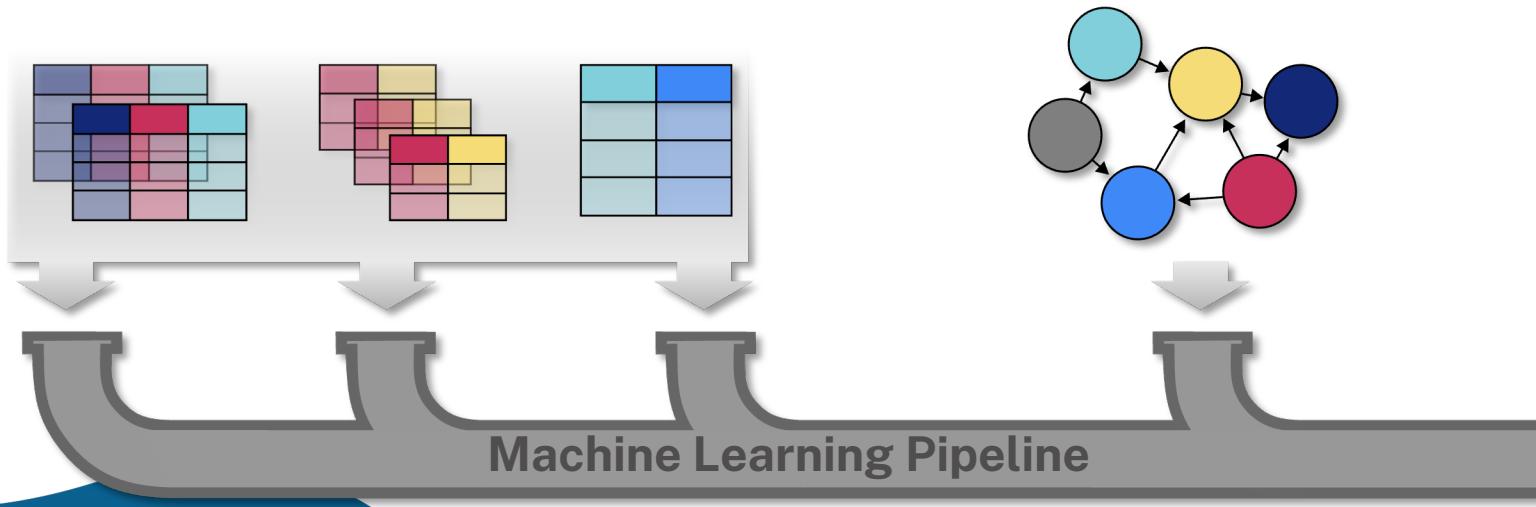
I

I has the highest betweenness centrality (0.59)

This person is an efficient connector of other people. Risk of network disruption is higher if you lose this individual

Better Predictions With Data You Already Have

- Traditional ML ignores network structure because it's difficult to extract
- Graphs use relationships to unlock otherwise unattainable predictions
- Add graph data to existing ML pipelines to increase accuracy



Graph Algorithm Categories



Pathfinding & Search

Finds optimal paths or evaluates route availability and quality



Centrality / Importance

Determines the importance of distinct nodes in the network



Community Detection

Detects group clustering or partition



Link Prediction

Estimates the likelihood of nodes forming a future relationship



Similarity

Evaluates how alike nodes are by neighbours and relationships



Embeddings & ML

Compute low-dimensional vector representations of nodes in a graph, and allow you to train supervised machine learning models

Available Algorithms (as of 2.4)



Pathfinding & Search

- A* Shortest Path
- All Pairs Shortest Path
- Bellman-Ford Single Source
- Breadth & Depth First Search
- Delta-Stepping Single-Source
- Dijkstra Single-Source
- Dijkstra Source-Target
- Minimum Spanning Tree
- K-Spanning Tree
- Random Walk
- Yen's K Shortest Path
- Minimum Directed Steiner Tree



Centrality / Importance

- ArticleRank
- Betweenness Centrality & Approx.
- Closeness Centrality
- Degree Centrality
- Eigenvector Centrality
- Harmonic Centrality
- Hyperlink Induced Topic Search (HITS)
- Influence Maximization (CELF)
- PageRank
- Personalized PageRank



Community Detection

- Conductance Metric
- K-1 Coloring
- K-Core Decomposition
- K-Means Clustering
- Label Propagation
- Leiden Algorithm
- Local Clustering Coefficient
- Louvain Algorithm
- Max K-Cut
- Modularity Optimization
- Speaker Listener Label Propagation
- Strongly Connected Components
- Triangle Count
- Weakly Connected Components



Link Prediction

- Adamic Adar
- Common Neighbors
- Preferential Attachment
- Resource Allocations
- Same Community
- Total Neighbors



Similarity

- K-Nearest Neighbors (KNN)
- Node Similarity
- Filtered KNN & Node Similarity
- Cosine & Pearson Similarity Functions
- Euclidean Distance Similarity Function
- Euclidean Similarity Function
- Jaccard & Overlap Similarity Functions



Embeddings & ML

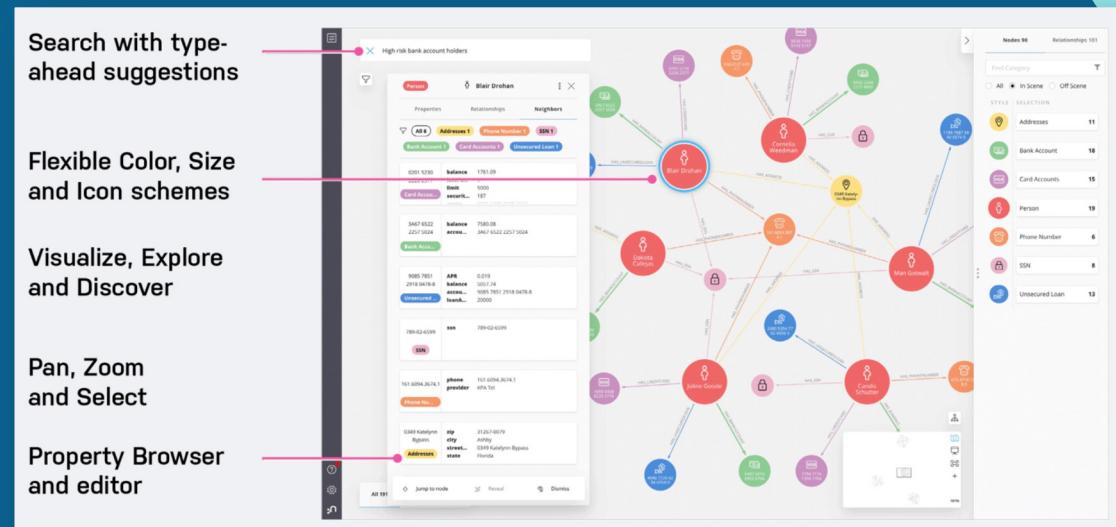
- Fast Random Projection (FastRP)
- FastRP with Property Weights
- GraphSAGE
- Node2Vec
- HashGNN
- Node Classification Pipelines
- Link Prediction Pipelines
- Node Regression Pipelines

Visualisation

Data Visualization with Neo4j Bloom

Neo4j's user-friendly graph database visualization, exploration and collaboration tool.

- Visually explore graphs
- Prototype faster
- Visualize and discover
- Easy for non-technical users



NeoDash - Dashboarding with Graph Data

NeoDash - Neo4j Dashboard Builder

Graph Summit 2023 - Digital Twin Workshop

Graph Summit Digital Twin Overview Shortest Path Calculations Speed vs. Time +

Country Code shown: BE Stations in country BE: 418 Stop Points in country BE: 325 Switches in country BE: 454

Countries

- BE
- CH
- DE
- DK
- ES
- EU
- FR
- IT
- LU
- NL
- SE

Map on Tracks

Report name... +

neo4j

Use Case Explanation

Digital Twin - An Overview

What is a Digital Twin?



A Digital Twin is a digital representation of a ... real-world physical product, system, or process ... that serves as the effectively indistinguishable digital counterpart of it for practical purposes, such as simulation, integration, testing, monitoring and maintenance.

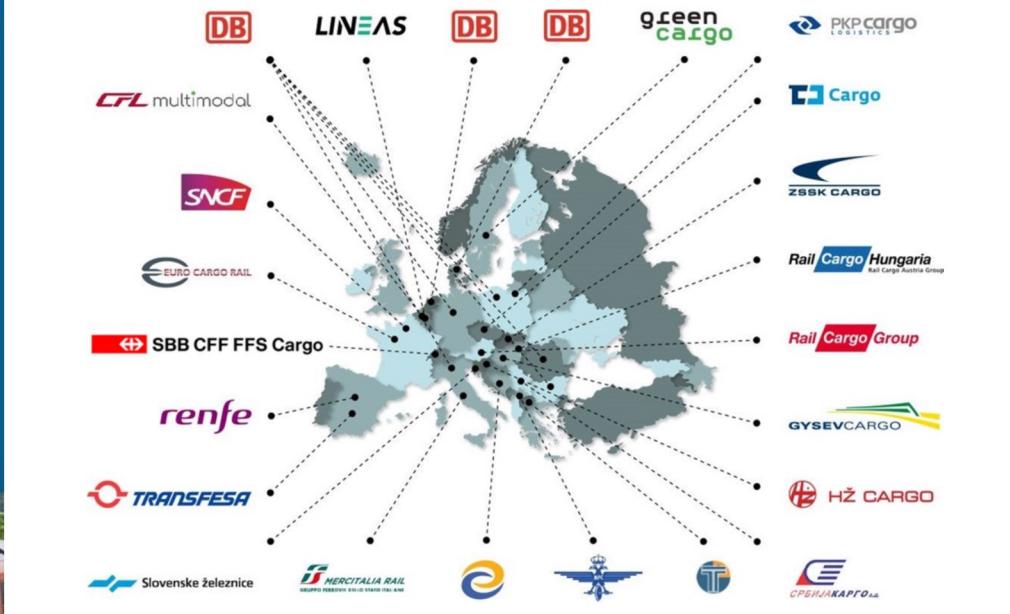


It has been done before

- **Challenge:** Legacy technology could not section and analyse train journeys
- **Solution:** Neo4j Knowledge Graph
- Identify and avoid bottlenecks



EU Rail Network



Why do we need a Digital Twin?



Improved efficiency

Optimize operations and reduce costs by simulating different scenarios and making data-driven decisions.



Enhanced safety

Identify potential hazards and test safety measures to improve safety for passengers and employees.

Predictive maintenance

Monitor asset condition in real-time, predict maintenance needs, and increase asset lifespan.



Improved customer experience

Simulate disruptions and help proactively address issues to enhance the customer experience and increase satisfaction.



Modeling the solution

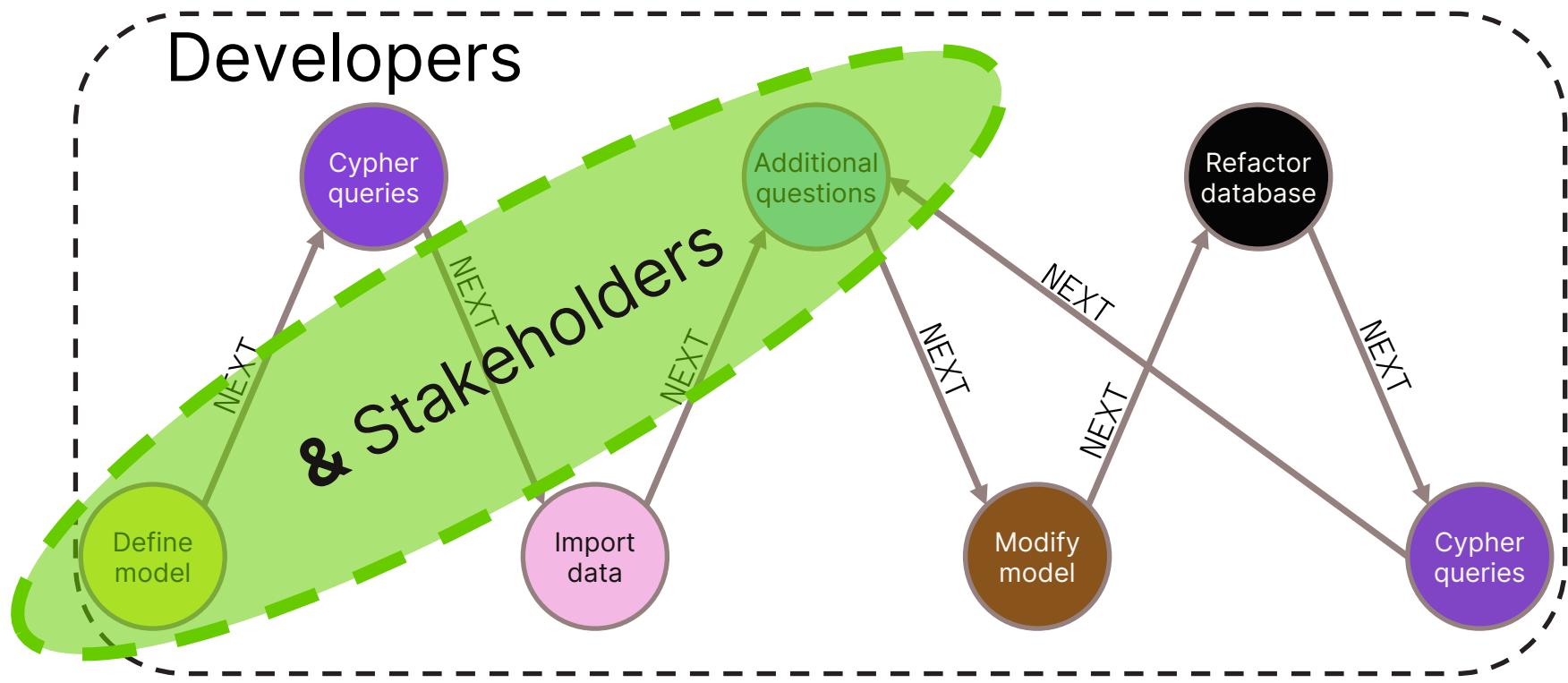
What is graph data modeling?

Collaborative effort where the application domain is analysed by **stakeholders** and **developers** to come up with the optimal model for use with Neo4j.

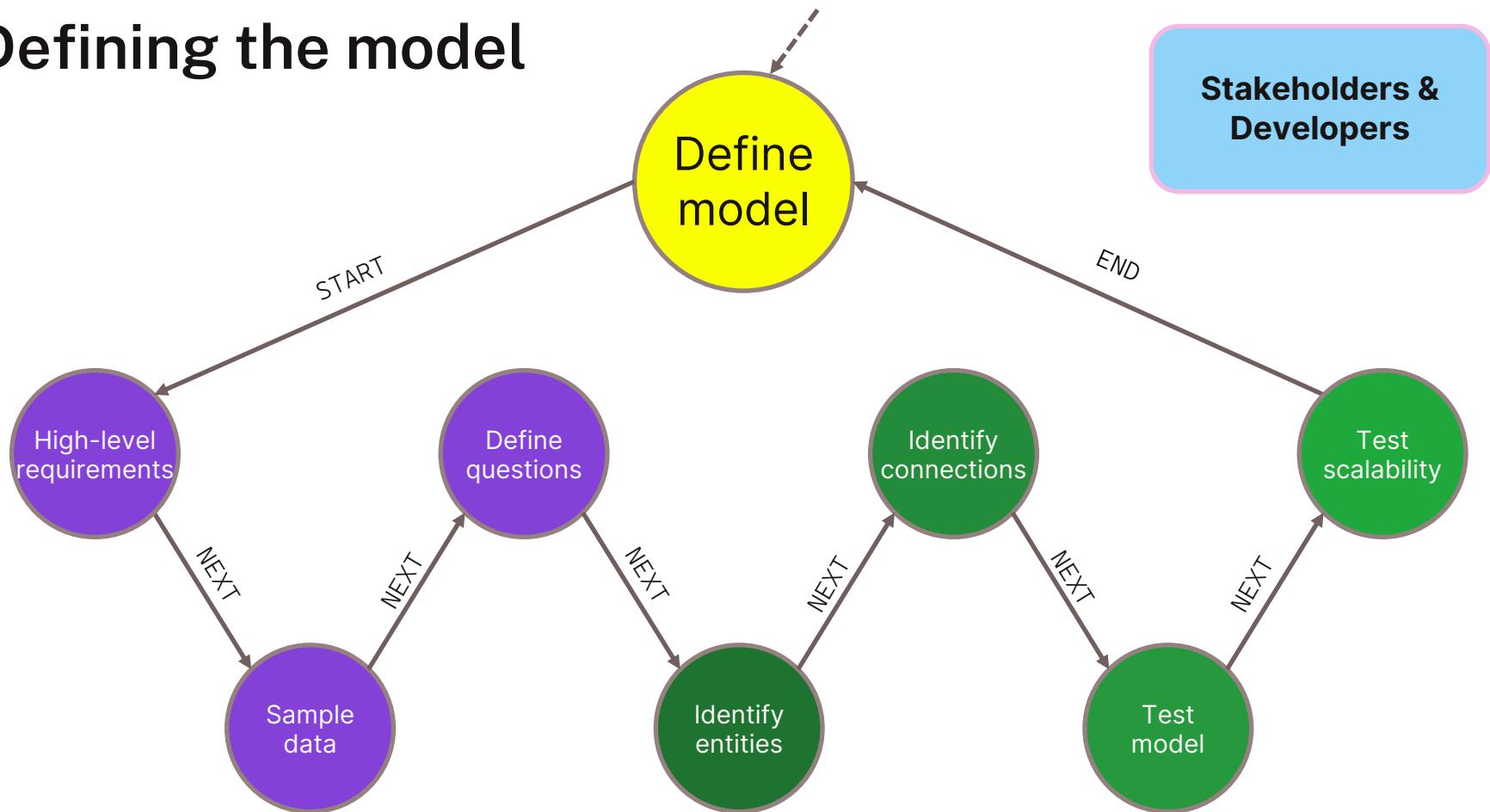
Who are the stakeholders?

- Business analysts
- Architects
- Managers
- Project leaders
- Data Scientists

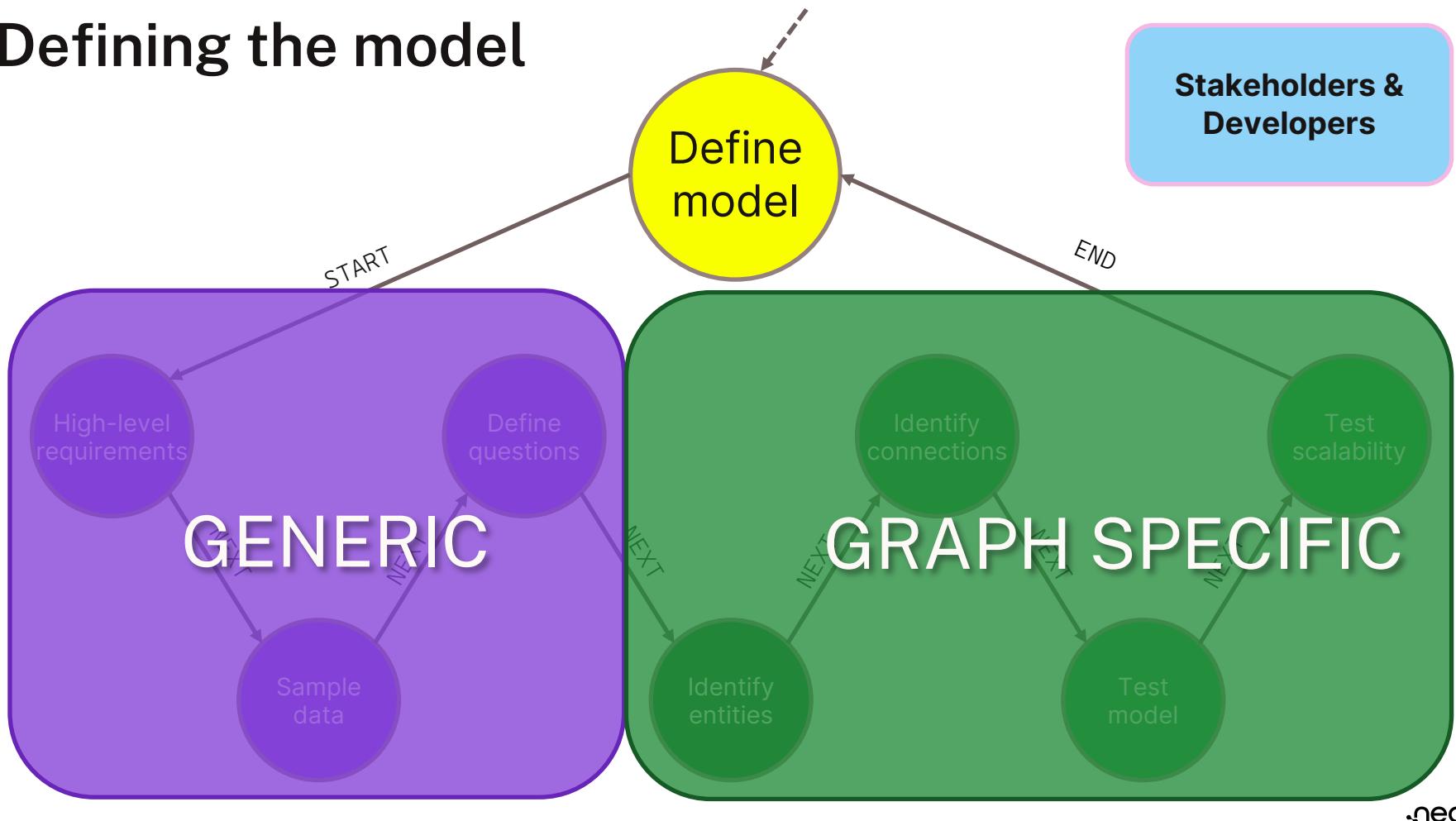
Graph data modeling workflow



Defining the model



Defining the model



Modeling - Step 1

Domain knowledge – High level requirements

We've already got this!

- Normally...
- In this case we've talked about the Domain
- You will have the knowledge of your own Domain

Modeling - Step 2

Sample Data

Get Sample Data



Static Rail Network*

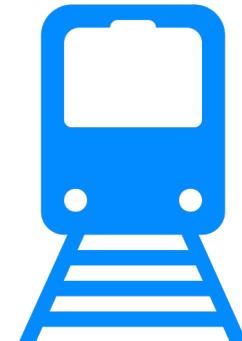
- Sections of lines
 - Length,
 - Speed
- Operational Points (Stations etc)
 - Geolocation information,
- Points of Interest (POI) along lines

*Source: Register of Infrastructure (RINF) - <https://data-interop.era.europa.eu/>

Operational Points (OP) - Data Explanation

CSV Header titles:

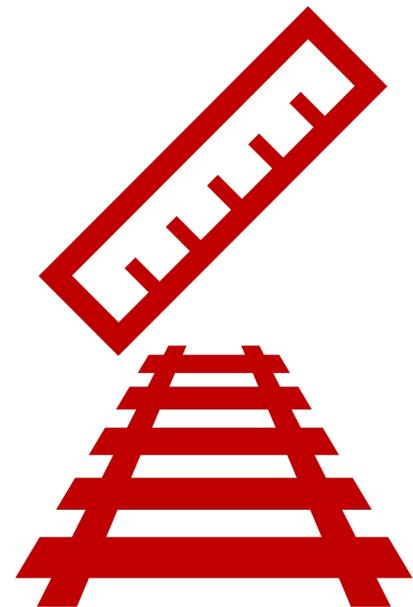
- id (string): The identifier of the OP
- extralabel (string): The type of the OP (Station, Switch, Border Point...)
- name (string): The name of the OP
- latitude (float): The latitude of the OP
- longitude (float): The longitude of the OP



Section Length Data

CSV Header titles:

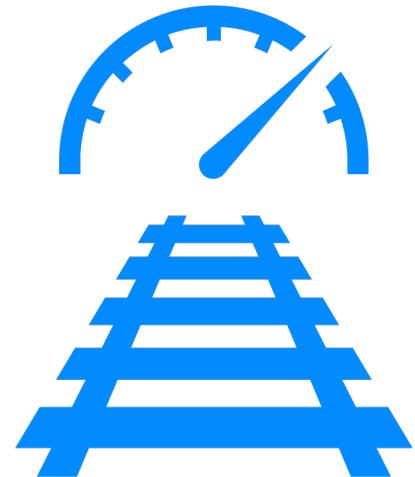
- source (string): The identifier of the *start* OP for the section
- target (string): The identifier of the *end* OP for the section
- sectionlength (float): The distance (in Kilometers) of the section



Section Speed Data

CSV Header titles:

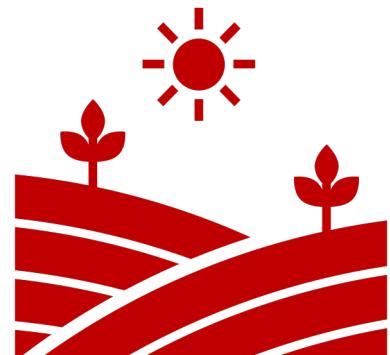
- source (string): The identifier of the *start* OP for the section
- target (string): The identifier of the *end* OP for the section
- trackspeed (float): The maximum speed allowed on that section



Point of Interest (POI) Data

CSV Header titles:

- CITY (string): The name of the City the POI is in, or close to
- POI_DESCRIPTION (string): A short description of of the POI
- LINK_FOTO (string): A link to a picture of the POI
- LINK_WEBSITE (string): A link to a web page about the POI
- LAT (float): The latitude of the POI
- LONG (float): The Longitude of the POI



Modeling - Step 3

Domain Questions

Data Modeling – Example Domain Questions

1. What is the route from Operational Point X to Operational Point Y?
 - What's the quickest way to get a repair crew from Technical Services to a given Switch?
2. What is an alternative route if an Operational Point on a route is closed?
 - A Switch is broken and we need to reroute Trains
3. How many routes are affected if I need to upgrade an Operational Point?
 - A Switch needs to be upgraded to support the network
4. What POIs are along a route?
 - Can we make revenue from referral commissions? Find busier routes during tourism season?

Modeling - Step 4

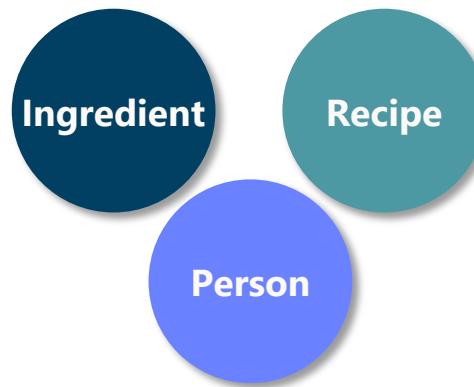
Identifying entities and connections

Identify Entities from Questions

Entities are the **nouns** in the domain questions:

1. What **ingredients** are used in a **recipe**?

2. Who is married to this **person**?

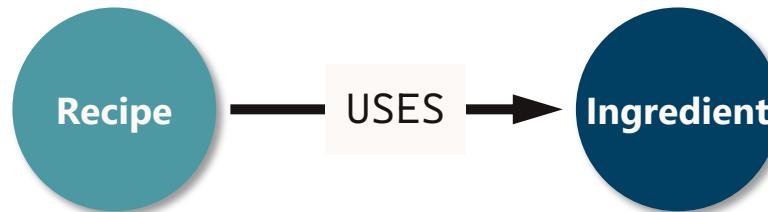


- The generic *nouns* often become labels in the model
- Use *domain knowledge* deciding how to further group or differentiate entities

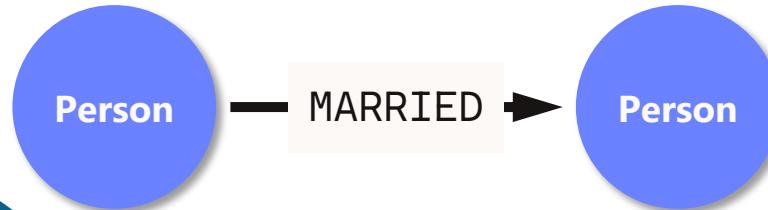
Identify Connections between Entities

Connections are the **verbs** in the domain questions:

- What ingredients are **used** in a recipe?



- Who is **married** to this person?

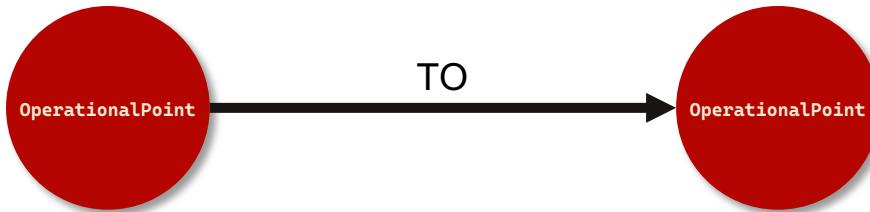


Using our Questions – Question 1

1. What is the route **from Operational Point X to Operational Point Y?**
 - What's the quickest way to get a repair crew from Technical Services to a given Switch?
2. What is an alternative route if an Operational Point on a Section is closed?
 - A Switch is broken and we need to reroute Trains
3. How many routes are affected if I need to upgrade an Operational Point?
 - A Switch needs to be upgraded to support the network
4. What POIs are near Station Operational Points on a Section?
 - Can we make revenue from referral commissions? Find busier routes during tourism season?

Using our Questions – Question 1 – Model

1. What is the route **from Operational Point X to Operational Point Y?**
 - What's the quickest way to get a repair crew from Technical Services to a given Switch?

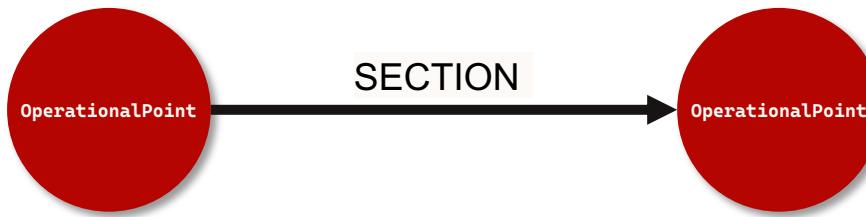


Using our Questions – Question 2

1. What is the route **from Operational Point X to Operational Point Y?**
 - What's the quickest way to get a repair crew from Technical Services to a given Switch?
2. What is an alternative route if an **Operational Point** on a **Section** is closed?
 - A Switch is broken and we need to reroute Trains
3. How many routes are affected if I need to upgrade an Operational Point?
 - A Switch needs to be upgraded to support the network
4. What POIs are near Station Operational Points on a Section?
 - Can we make revenue from referral commissions? Find busier routes during tourism season?

Using our Questions – Question 2 – Model

2. What is an alternative route if an **Operational Point** on a **Section** is closed?
 - A Switch is broken and we need to reroute Trains

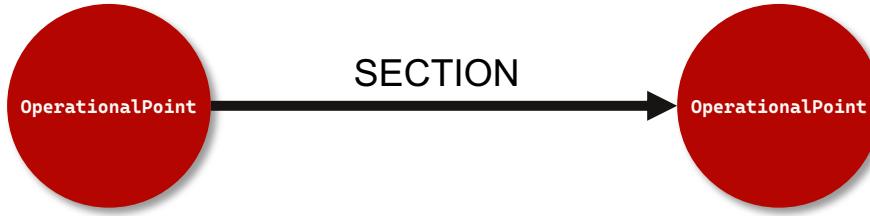


Using our Questions – Question 3

1. What is the route **from Operational Point X to Operational Point Y?**
 - What's the quickest way to get a repair crew from Technical Services to a given Switch?
2. What is an alternative route if an **Operational Point** on a **Section** is closed?
 - A Switch is broken and we need to reroute Trains
3. How many routes are affected if I need to upgrade an **Operational Point**?
 - A Switch needs to be upgraded to support the network
4. What POIs are near Station Operational Points on a Section?
 - Can we make revenue from referral commissions? Find busier routes during tourism season?

Using our Questions – Question 3 – Model

3. How many routes are affected if I need to upgrade an **Operational Point**?
 - A Switch needs to be upgraded to support the network



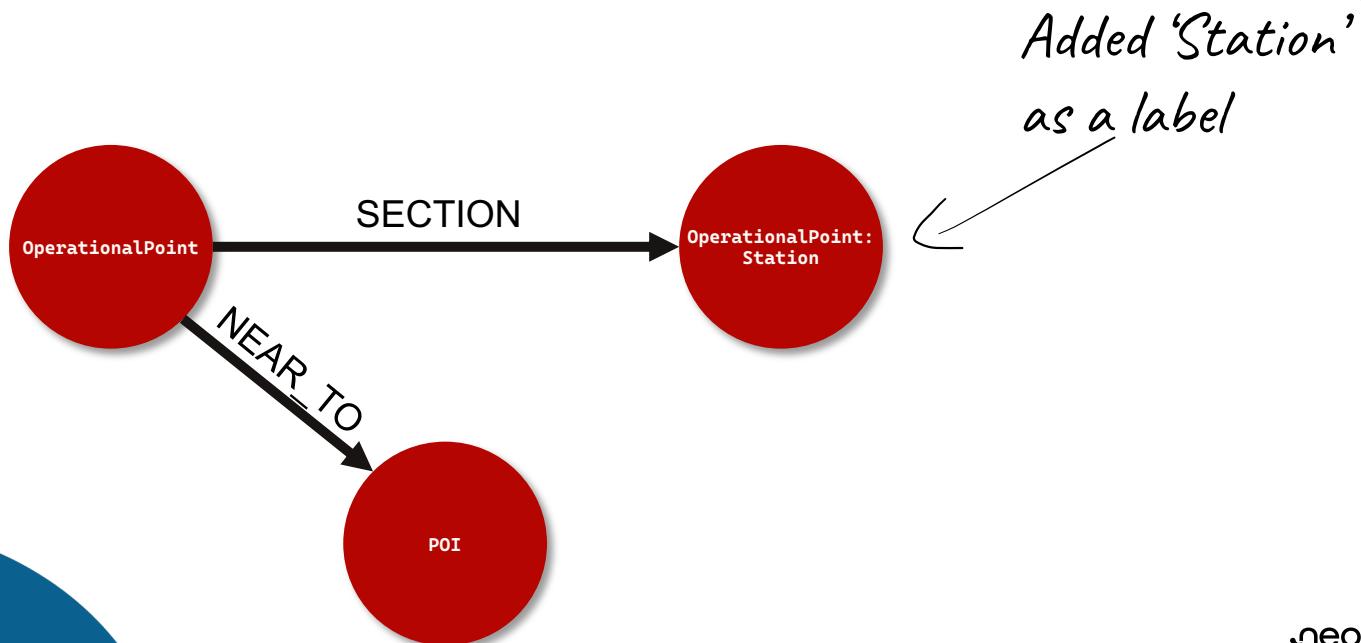
Using our Questions – Question 4

1. What is the route **from Operational Point X to Operational Point Y?**
 - What's the quickest way to get a repair crew from Technical Services to a given Switch?
2. What is an alternative route if an **Operational Point** on a **Section** is closed?
 - A Switch is broken and we need to reroute Trains
3. How many routes are affected if I need to upgrade an **Operational Point**?
 - A Switch needs to be upgraded to support the network
4. What **POIs** are **near Station Operational Points** on a **Section**?
 - Can we make revenue from referral commissions? Find busier routes during tourism season?

Using our Questions – Question 4 – Model

4. What POIs are near Station Operational Points on a Section?

- Can we make revenue from referral commissions? Find busier routes during tourism season?



neo4j

Workshop

Connecting to the demo environment

We created a temporary instance dedicated for the BT Graph Day Bangalore.

Access is through **<https://browser.neo4j.io>**

And use the following connection details:

- Connect url: **neo4j+s://neo4j-bt.graphdatabase.ninja:443**
- Login: **attendeeXX** (where XX corresponds to the number that was assigned to you)
- Password: **attendeeXX** (same as login)

We already preloaded the data for you.

Get the code

Open the Github page:

<https://github.com/guerinjeanmarc/graph-day-bangalore>



Visualisation - NeoDash

NeoDash

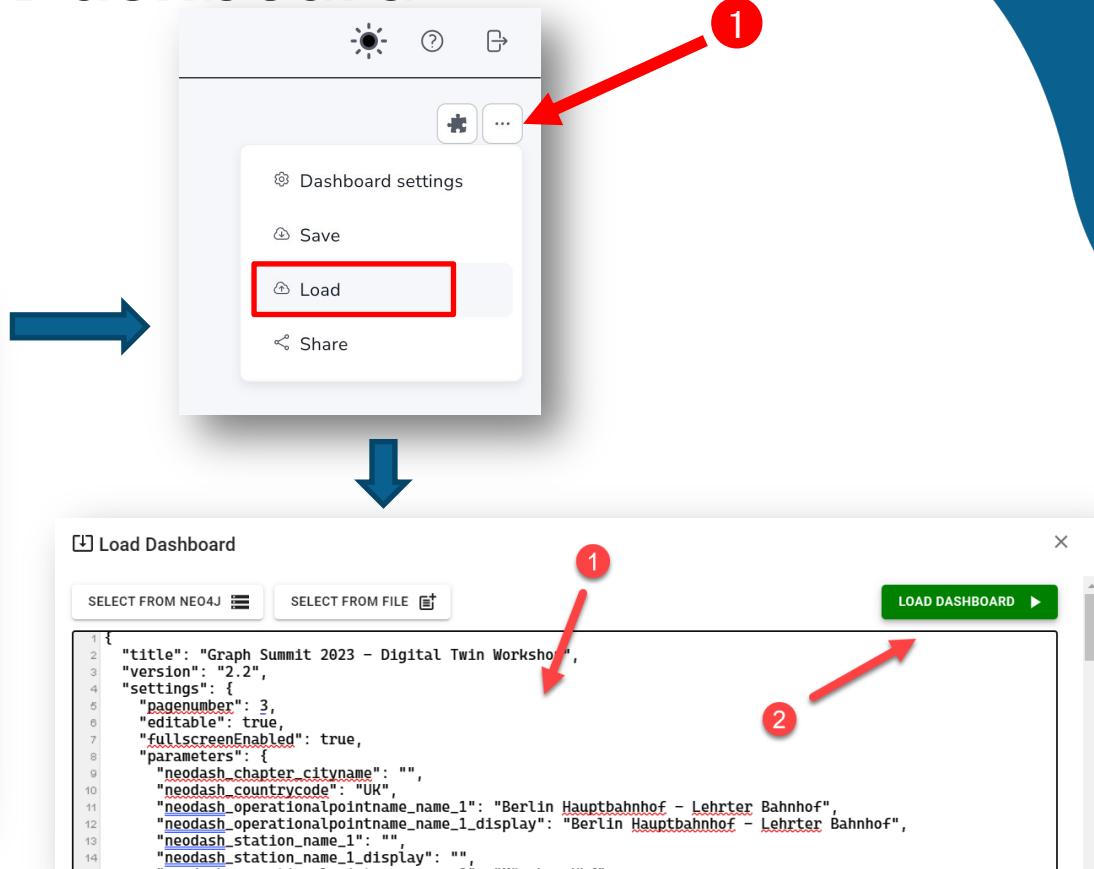
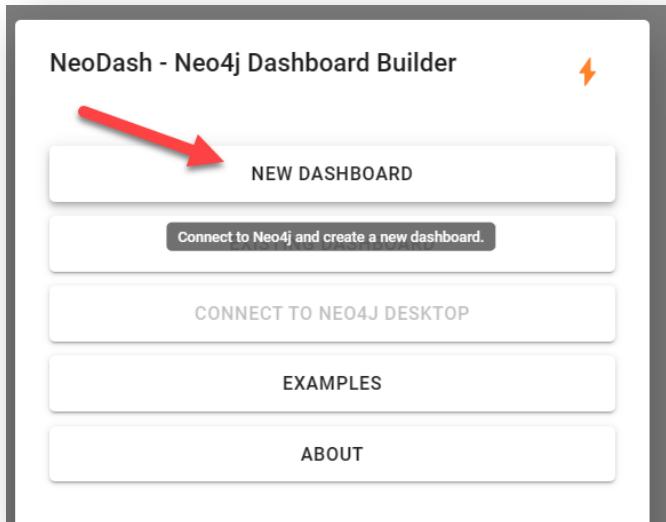
Open NeoDash:

<http://neodash.graphapp.io/>

- Same connection details:
- Connect url: **neo4j+s://neo4j-bt.graphdatabase.ninja:443**
- Login: **attendeeXX**
- Password: **attendeeXX**

NeoDash – Import Dashboard

- Create New Dashboard
- Import from Source files





Q & A

