





Towards a Tube Twin

🕒 Last Updated Time	@August 15, 2022 4:00 PM
👤 Last Updated By	 Paul dos Santos
📌 Status	Prioritized
👥 Proposers	 Tariq Desai
👥 Consulted	
👥 Informed	
☰ Objectives	Development of competencies and networks
☑ Ready for review	☑
📌 Estimation - Effort	4
📌 Estimation - Cost	1 📉
📌 Estimation - Skills	4
📌 Estimation - Impact	3
Σ Category	Big Project
Σ Effort Category	High
Σ Impact Category	High
Σ Priority Score	5
☑ Review finalised	☑
Σ Review Status	Review Complete



Please refer to the further detail provided on the [✓ Project Idea Checklist](#) page if anything is unclear.

The purpose of completing this checklist is to validate that the idea makes sense to put R&D effort into. The idea can be as broad or specific as makes sense - the actual projects derived from this will be added to the R&D backlog for prioritisation.

Some of the questions may be difficult to answer on your own with the information that is available to you right now; this is fine! Please involve others to get some additional opinions and advice to be sure you've covered as many aspects of what the proposal entails as possible.

If you feel a question isn't relevant for your proposal, try reframe the question. If it still doesn't work, simply indicate it's "not applicable".

Project Summary

Provide an elevator pitch summary of the project here.

- The London Underground (aka "The Tube") is a network of train stations which connects the city.
- Via an open data API, Transport for London (TfL) publishes data showing in fifteen minute increments the number of people entering and exiting every station.
- Understanding the behaviour of this transport network presents many similar challenges to the ones we face in understanding water and waste distribution networks.
- A project could answer several questions, some more speculative and harder than others:
 - How well can use the available data to forecast passenger counts on train lines and stations throughout the year?

- What knock-on effects do special events (eg. football games, holidays, concerts) have on the network?
- What is the relationship like between passenger counts at different stations? Can we do short-term forecasting of load at one station given observations of loads elsewhere in the network?
- Has there been a shift in the way people have used the Tube since before the pandemic? (Not just in aggregate, but also in the popularity of certain lines, or relationship between counts at different stations.)
- The Elizabeth line has been introduced recently. Can we tell if there's been an effect on the passenger counts at other lines as a result of the development?
- Do we notice any long-term trends in passenger loads across the different lines and stations, perhaps related to developments in the “catchment” areas of different stations
 - Highly speculative: Given what we know about where people live and commute in the city, if we had to add another line or station to the network (such as, in the simplest case, a line connecting two stations which already exist) where would the optimal placing be?
- Can we simulate the network and uncover plausible risks to the network (“what if by some bad luck we had an unusually busy, but not too far-fetched, load of incoming passengers at stations X and Y - would that cause station Z to be totally overloaded?”)

Project Checklist

1. Who are the proposers?

1.1 Correct delegation

List who the project proposer(s) are. Provide some detail about their function/role which highlights that they have the suitable context. Even if the proposer don't ultimately end up working on the

project directly, they should be comfortable supporting the project team as required.

- @Tariq Desai Happy to support on network modelling and forecasting, having tackled some similar problems in Utilities
- @Jan Linde is our programme lead in digital twins and will be interested in understanding the outcomes of this project

2. The use case

2.1 Output-focused ideation

Explain the outputs of the project. Provide as much detail as possible and explain why this output is valuable. Ask yourself, “Is this the end or the means?” If it’s the means, don’t discuss it yet.

- A graph representation of the network (either via a simple Python library like `networkx` or if there is more substantial data engineering expertise in the team, and if it makes technical sense, via a graph database)
- A passenger count *forecast* at each station (entry and exit), in 15 minutes increments, based on time of day, and counts at other stations
 - A study of the relationships between counts at different stations (can we put a distribution on load at one station, given load elsewhere?)
- A study of the effect of significant events (holidays, football games, etc) on traffic
- Further steps towards other harder questions set out in the problem summary

2.2 Source of inspiration

Explain what inspired this proposal. Provide any relevant details that highlight a particular challenge or opportunity this addresses.

- We tackle very similar questions in Utilities, and will be tackling similar ones as we build more advanced “Digital Twin” capabilities. It would help to have a different test

system in which to improve our intuition for the behaviour of physical distribution networks

2.3 Appropriate task for ML/AI

Explain why there is an ML/AI component needed. Provide any information that is available to substantiate the appropriateness. Include examples where relevant.

- Demand forecasting is a classical ML/AI problem
- Doing such an exercise across a distribution network would be more challenging

2.4 User Experience

Explain who your users are and how they will use the solution/outputs. Detail anything special we should know about the intended users that can aid in the project design and will ensure smooth adoption.

- Users are railway technicians trying to understand the behaviour of their system
- Commuters could also find the forecasts valuable
- Final conclusions should be presented in some form (perhaps via dashboard) which allows key insights of forecasts to be understood by users

2.5 Ethical considerations

Explain how the project impacts all people it could affect. Who might be harmed by the outputs of this project? Who stands to benefit?

- Railway planners might benefit, as well as commuters, who might be able to plan their journeys to avoid periods of high congestion
- Can't imagine any harm will be incurred as this forecast improves

3. Reality check

3.1 Reasonable expectations

Explain the expectations of the output and considerations around the impact of mistakes.

- Mistaken forecasts could lead to unnecessary mitigation from users. But it shouldn't be a problem at this stage.

3.2 Possible in production

Explain what kinds of technology resources are needed to get this project into production. Is it part of the output to be production ready or will this be fulfilled by some other team after the concept has been proven? Do you expect the resource requirements to grow over time? What happens if there are more users a year from now?

- The computational demands, even of a relative small network such as the London underground, might lead to issues. Potentially simplifications of the network could be studied, such a aggregating sites on the same lines outside central London

3.3 Data availability

Explain what data is available and how it will be accessed/created.

- Data is available here from TfL:

Our open data

This feed provides access to realtime Tube data, including: A summary train prediction service A detailed train prediction service Station status Line status Powered by Windows Azure How often we publish a fresh copy of the feed 30 secs Maximum time allowed between capturing and displaying the feed 30 secs

<https://tfl.gov.uk/info-for/open-data-users/our-open-data?intcmp=3671#on-this-page-2>

3.4 Data volume

Explain the volume of data, how quickly it gets out of date and if there's a time component. Include any detail others have provided to ensure you can confidently say you can get enough data for what you're trying to do.


- Data volume should be sufficient (need to double check the time over which historical data is available)

3.5 Required technology & compute

Explain what processing power is needed for both R&D and production. Indicate how this needs to scale with more users.

-

3.6 Skills needed

List the skills you think will be needed to make the project a success. See  Project Skills for reference.

- Data engineering (graph databases) - potentially
- Some experience or interest in learning more about graph and network algorithms
- Multivariate time series forecasting

3.7 Ground truth

Explain what the ground truth outputs in this project are. Does it exist or still need to be created?

- Validation data can be set aside in the data provided by the API

3.8 Logging sanity

Explain how confident you are that we can map input to output correctly.

- Highly

3.9 Logging quality

Explain any data quality issues you foresee. Detail how confident you are in the data and add information that may be useful in understanding/confirming the data quality.

- Missing days
- Distributional issues resulting from the effects of the pandemic on commuter patterns

3.10 Indifference curves

List mistakes that could be in the eventual outputs of the project and provide an indication of how 'bad' the errors are relative to one another from a business perspective. Not all mistakes carry the same impact.

- N/A

4. Performance metrics

4.1 Simulation

Explain if the project would benefit from any form of simulation, e.g. UI mocks and synthetic data. Detail why this is a value-add upfront if required.

- Simulation will likely be a core feature of this project. It will allow us to try out plausible scenarios and test the resilience of the system, as well as

4.2 Metric creation

Explain what your performance metric is. Detail how we can measure the project works as intended. Explain why this metric is robust.

- standard forecast metrics (such as MSE or MASE or probabilistic equivalents)

4.3 Metric review

Explain how the performance metric has been reviewed to ensure it can't be misinterpreted/abused/gamed in some way.

- N/A

4.4 Metric-loss comparison

If this project has a AI/ML component, explain how the performance metric correlates with a standard loss function. If it doesn't, explain how you anticipate we can craft a custom metric that behaves as expected.

- We are, in the first instance, looking at standard metrics here

4.5 Population

Explain the statistical population of interest and detail any specific exclusions.

- Might have to exclude pandemic months

5. Testing criteria to overcome human biases

5.1 Minimum performance

Explain what the minimum acceptance criteria are for this project. If we can't meet this criteria are you comfortable terminating the project?

- Graph representation of the network with univariate forecasts for each station
- Happy to terminate if this is impossible

6. Value creation

6.1 Value drivers

What benefit is gained by investing effort into this idea? What are the drivers of value creation that this enables? Use this information to aid in the impact estimation.


- Adding to test cases for distribution network analysis within the EXPLORE team
- TfL is building a digital twin of aspects of the network - perhaps we could pitch the work to them if it's very good




Checklist Review












To be completed by at least 2 reviewers that did not complete the checklist above.

Here you are indicating that you're comfortable there is enough information available in this document for a given section for the project to be added to the candidate list.

Make comments directly on the sections above and update the page properties as necessary.

- Reviewer 1:  Paul dos Santos
- Reviewer 2: *Add name here*
- Reviewer 3: *Add name here*

-  Sufficient information provided
-  Insufficient information provided
-  Not Applicable

		Reviewer 1	Reviewer 2	Reviewer 3
1.	Who are the proposers?			
1.1	Correct delegation			
2.	The use case			
2.1	Output-focused ideation			
2.2	Source of inspiration			
2.3	Appropriate task for ML/AI			
2.4	User experience			
2.5	Ethical considerations			
3.	Reality check			
3.1	Reasonable expectations			
3.2	Possible in production			
3.3	Data availability			
3.4	Data volume			
3.5	Required technology & compute			
3.6	Skills needed			
3.7	Ground Truth			

		Reviewer 1	Reviewer 2	Reviewer 3
3.8	Logging sanity	✓		
3.9	Logging quality	✓		
3.10	Indifference curves	✓		
4.	Performance metric			
4.1	Simulation	✗		
4.2	Metric creation	✓		
4.3	Metric review	✗		
4.4	Metric-loss comparison	✓		
4.5	Population	✓		
5.	Testing criteria to overcome human biases			
5.1	Minimum performance	✓		
6.	Value creation			
6.1	Value drivers	✓		

Additional notes:

- @Paul dos Santos → @Tariq Desai I really like this idea. This is a crazy rich API!

Network assets

- Bus stops and routes
- Coach parking sites/locations
- Pier locations
- Road network
- Bridges, tunnels, road barriers - height restrictions
- Cycling routes
- Oyster ticket stop locations (for topping up cards)
- Station topology - including information on lifts, toilets, etc

Real-time assets

- Journey planning (multi-modal routing)
- Air quality/atmospheric emissions
- Tube departure boards, line status and station status
- Bus & river arrivals
- Road-side message signs

Aggregated historic assets

- Cycling hire trip data
- Walking times between adjacent stations
- Walking times for journeys that could be quicker to walk
- Network stats
 - Busiest times on trains and in stations
 - Passenger counts
 - Origin-destination surveys

What if we tried to twin the full London transport network as a longer term ambition?

Using OSM ([where this Transport for London data has been integrated into](#)) and GTFS data ([from Transitland](#)) the solution can be easily scaled out to other metropolitan areas without a specific reliance on the Transport for London API. If you then combine population information and calibrate the network flows with relevant origin-destination data you can build a very detailed twin using something like **Simulation of Urban MObility** (or something similar we develop in-house). To start, we could consider simply create a Tube simulation for a given line, extend out to the full Tube network and then add other transport modes iteratively.

- Minimal bus stop example: <https://sumo.dlr.de/docs/Tutorials/PublicTransport.html>
- GTFS Tutorial: <https://sumo.dlr.de/docs/Tutorials/GTFS.html>
- Railways: <https://sumo.dlr.de/docs/Simulation/Railways.html>

The main reason I'm proposing we think about the broader transportation network is it helps contextualise the Tube and can aid in scenario-based modelling.