## Using data to create livable cities

1:56  
Alexey  
This week, we'll talk about using data to create livable cities. We have a special guest today, Rachel Lim. Rachel is an urban data scientist dedicated to creating livable cities through the innovative use of data. Welcome, Rachel!

2:31  
Rachel  
Thank you! I'm happy to be here today. I've benefited greatly from the DataTalks.Club courses, and I'm excited to share my experiences.

2:41  
Alexey  
We're happy to have you. Before diving into our main topic, could you tell us about your career journey so far?

## Rachel's career journey: from geography to urban data science

2:52  
Rachel  
Yes, I'm currently working as a data engineer in Singapore, focusing on creating livable cities using data. My background is in geography — I have a bachelor's degree in geography and a master's in urban data science. I blend qualitative and quantitative analysis to tackle urban challenges.

I began my career in data science, applying analytics and machine learning to various transportation projects, such as bike-sharing analytics to address indiscriminate parking and road defect detection using computer vision. These projects allowed me to make a tangible impact in cities. Seeing my work lead to real-world solutions motivated me to become a transport scientist. I focus on analyzing travel patterns to support long-term planning in Singapore. Recently, I transitioned to data engineering after completing the DataTalks.Club data engineering course. By diving into data foundations and building data platforms, I aim to optimize AI applications for creating livable cities.

4:02  
Alexey  
That sounds amazing! How was the course?

4:05  
Rachel  
It was really good. It covered a lot about building data pipelines and using tools like Apache Kafka, which was quite an eye-opener. The course was very relevant and helpful in my transition to my current role.

## What does a transport scientist do?

4:20  
Alexey  
You mentioned you were a transport scientist. That’s an interesting title. What exactly does a transport scientist do, and what types of organizations need this role?

4:47  
Rachel  
Transport scientists are usually needed in the public sector, especially in government agencies involved in transportation planning. The role involves applying data science in a practical way to public transport and transportation planning. Another sector where transport scientists are valuable is transport consultancy, such as firms like Sam Schwartz. It's essentially about applying data science within an urban context to improve transportation systems.

5:23  
Alexey  
So, is it about planning where to put bus stops, how often buses should run, and similar things?

## Short-term and long-term transportation planning

5:34  
Rachel  
Yes, that’s part of it. We separate our work into short-term and long-term planning. In the short term, we look at bus routes, service frequencies, travel patterns, and how well services are meeting users' needs. In the long term, we use travel pattern data to make projections and plan for future infrastructure needs, such as additional roads or rail lines and their alignment.

## Data sources for transportation planning in Singapore

6:14  
Alexey  
I guess each bus in Singapore has sensors to track its location and passenger load, right? This data helps you see if certain routes are overcrowded and need more frequent service?

6:47  
Rachel  
Exactly. We use a combination of data sources. Buses are equipped with GPS transponders, allowing us to track their locations and times at each bus stop. This helps us identify issues like bus bunching, where multiple buses arrive at the same stop simultaneously. Ideally, buses should be spaced out to optimize service.

On the demand side, we look at fare card data to understand where passengers are tapping in and out, giving us a clearer picture of travel demand.

7:36  
Alexey  
What do you mean by "tapping in and out"?

7:40  
Rachel  
In Singapore, we use a fare card system similar to London's Oyster card or New York's MetroCard. Passengers tap their card when they enter and exit public transportation, like trains and buses.

7:59  
Alexey  
So there's a card and a reader? That makes sense. In Berlin, people can just hop on a bus without any interaction. Sometimes there's a fare check, but it's not consistent.

8:20  
Rachel  
Yes, that approach makes it more challenging to collect travel data. You’d need to rely on video surveillance and computer vision to analyze passenger flow, which is more complex than just processing fare card events.

## Rachel's motivation for combining geography and data science

8:38  
Alexey  
Definitely. So, what motivated you to work at the intersection of geography and data science?

8:55  
Rachel  
Growing up in Singapore, I was fascinated by the systems shaping our cities. I witnessed firsthand how the city rapidly expanded its MRT (Mass Rapid Transit) network and developed new housing estates. This sparked my interest in urban planning and geography.

I did an internship at the Center for Livable Cities, which deepened my understanding of sustainable urban environments. During this time, I attended the World City Summit, where city leaders from around the world shared ideas on improving urban spaces.

This inspired me to study geography, focusing on urban design, geocomputation, and geospatial analytics. Eventually, this led me to pursue a master's in urban informatics at New York University, specializing in applying data science in urban contexts. This combination of experiences has shaped my current role.

## Urban design and its connection to geography

10:19  
Alexey  
That's interesting. My knowledge of geography is mostly from school, where we learned things like capital cities and natural features. I didn't realize geography could involve urban design. Is urban design about planning new districts, including schools and parks?

11:26  
Rachel  
Yes, but it goes beyond just planning where things go. It’s about designing environments that are livable. This involves making streets walkable, deciding on the width of streets, placing sidewalks, and using elements like planter boxes to separate pedestrians from traffic. It's also about creating safe, welcoming spaces where people want to linger and interact, fostering a sense of community.

12:06  
Alexey  
I see. Why is this field still called geography when it covers urban design and other aspects?

12:20  
Rachel  
Geography isn't just about physical features like mountains or rivers. It also includes human geography, which focuses on people, migration, and population changes. It’s about how these factors interact with physical spaces. Geography is fascinating because it’s connected to the real world — what we study directly influences how we live and interact with our environments.

## Defining a livable city

13:12  
Alexey  
So far, we’ve talked about what makes a city livable — parks, pedestrian zones, traffic management, and fostering community. How would you define a livable city?

13:49  
Rachel  
A livable city is one where people feel connected to their community and have opportunities to thrive. In terms of the built environment, this can include efficient public transport with a well-connected network of buses, trains, and bike-sharing programs. It also means safe, pedestrian-friendly streets with dedicated bike lanes and walkways.

Other aspects are affordable housing, proximity to essential services, and green spaces. Beyond physical infrastructure, digital infrastructure plays a role too. This includes online access to government services, digital safety, and platforms that connect residents, recognizing that people spend a lot of time in digital spaces now.

## Livability of Singapore and urban planning

15:30  
Alexey  
How livable is Singapore, in your opinion? I've never been there, but it's on my list.

15:48  
Rachel  
Singapore has made significant progress. Initially, we focused on developing housing estates, but now there's a greater emphasis on placemaking — creating spaces where people can gather and enjoy. We're also converting certain streets into car-free zones, improving walkability and cycling infrastructure. It's a journey; some parts of the city are more livable than others, but we're working to expand these spaces.

16:35  
Alexey  
Singapore is geographically small and densely populated, so I imagine land use has to be very efficient.

16:55  
Rachel  
Absolutely. Singapore practices a process called Master Planning, which involves planning 15 years ahead. This ensures that amenities and infrastructure are effectively integrated over time.

17:16  
Alexey  
Interesting. I live in Berlin, and I think the city is fairly livable. It has good public transport, though cycling infrastructure could be improved. In Moscow, it's harder for people with disabilities to get around, whereas Berlin is more accommodating. It’s fascinating to see the practical application of geography and data science. Could you share more about what you do as a transport scientist and now as a data engineer?

## Role of data science in urban and transportation planning

18:24  
Rachel  
Planning a city requires deliberate effort, and data science plays a critical role in improving livability by offering insights and supporting data-driven decision-making. By analyzing data collected throughout the city, we can optimize services, enhance public safety, and promote sustainability.

In Singapore, we collect a lot of public data, which we share on open data platforms. This enables collaboration with citizen developers, students, and research institutions. Our data sources include transportation data like fare card usage, as well as census and survey data. This helps us understand travel patterns and conduct transport modeling to plan future infrastructure, such as rail lines.

We're also increasingly using data from the private sector, like mobility data from ride-sharing apps. This gives us additional insights into how people move around the city, beyond just public transportation.

## Predicting travel patterns for future transportation needs

20:31  
Alexey  
That's interesting because you mentioned predicting where people will move. If I understood correctly, for instance, if more people start moving from one part of Singapore to another, you can anticipate this trend. Maybe a district is becoming more popular, so it's gradually getting more populated. You want to predict these patterns to plan accordingly, right? Like adding a new bus line or increasing the frequency of existing buses?

21:09  
Rachel  
Yes, that's right. Singapore is quite small, so we actively plan how housing estates will develop. This could involve building new housing estates or renewing and rejuvenating existing areas. By doing so, we can estimate how many people will live in a particular district. From there, we use past travel patterns and data on new modes of transport to predict future movements. This helps us plan ahead for necessary transportation services, such as adding new bus lines or stops to ensure these areas are well-connected to the rest of Singapore.

## Data collection and processing in transportation systems

22:02  
Alexey  
I see. As a data scientist, you can't do much without data. You mentioned various data types, like sensor data, movement patterns, and ride-hailing information. All of this needs to be collected, processed, and analyzed. This must involve having sensors on buses and other physical means of data collection, right? Then, this data needs to be sent to a platform, aggregated, processed, and perhaps stored in a data warehouse for data scientists to use. As a data engineer, you're probably involved in these steps. Can you tell us more about what happens behind the scenes?

23:01  
Rachel  
Yes, we work with a combination of data sources. We gather GPS data from ride-hailing companies and public transport, along with fare card information about when and where people are tapping in and out. In our data pipelines, we have an end-to-end system that aggregates this information, stores it in a data warehouse, and processes it so that it's suitable for downstream data analysis. We don't just need real-time data; we also require historical data to do projections. Long-term data allows us to track patterns over time, which is crucial for providing stable and reliable insights.

## Use of real-time data for traffic management

24:02  
Alexey  
Are there situations where you actually need real-time data as well?

24:09  
Rachel  
Yes, real-time data is often needed for managing operations. For example, to monitor the reliability of services, real-time data is essential. Another use case is tracking traffic flow during specific events. In Singapore, we host the F1 night race, which takes place on a street circuit, so parts of the roads are closed. We use taxi data as a proxy to understand how traffic is flowing around the closed areas. By monitoring how quickly the GPS coordinates from taxis are moving, we can detect congestion and adjust traffic management strategies accordingly.

25:07  
Alexey  
What actions are taken if there's a traffic jam?

25:10  
Rachel  
It depends on the location. If a traffic jam occurs on an expressway, we have cameras that monitor these areas, pinpointing where the issue is. Recovery services, like tow trucks or other assistance, are dispatched to manage the situation and clear the blockage.

25:29  
Alexey  
So it could involve police officers managing traffic or using other traffic marshals to control the situation?

25:35  
Rachel  
Yes, a combination of traffic marshals and recovery services are deployed to clear any blockages and ensure that traffic can flow smoothly again.

25:45  
Alexey  
In Berlin, we often have events like half marathons or marathons that require large streets to be closed for half a day. This can be quite inconvenient for drivers who need to find alternative routes. These events cover large distances, so multiple roads must be blocked. Do you have similar events in Singapore, and how do you handle them?

26:30  
Rachel  
Yes, we host marathons and similar events in Singapore. These events often use a mix of roads and our network of park connectors, which are pathways that connect different parks and are free of vehicles. This allows most of the marathon routes to take place without significantly disrupting traffic on main roads. The impact on traffic is minimized because only a few roads need to be closed, thanks to these park connectors.

## Incorporating generative AI into data engineering

27:06  
Alexey  
That makes sense. Having connected parks where people can run without needing to wait for traffic lights is an excellent way to make a city more livable. I'm into running myself, and finding a route without having to wait for traffic lights is always a challenge. It's great that Singapore has planned these aspects so well. Do you participate in organizing these events, or are you more focused on transportation in your role?

27:59  
Rachel  
My role is more focused on transportation. While I don't directly manage these events, I work on data preparation and building data pipelines, ensuring that transportation systems run smoothly. Increasingly, I'm also looking at ways to incorporate generative AI into data engineering, such as enabling users to query databases without needing SQL knowledge.

28:43  
Alexey  
How often do people need to make these types of queries?

28:48  
Rachel  
Quite often. As data engineers or scientists, we frequently receive requests to extract data or perform specific analyses. Building tools that allow users to access insights on their own would free up our time for more innovative work. I believe that subject matter experts should drive data science and analysis because they best understand their needs and the problems they want to solve. When problems and requests are passed through multiple people, the original intent can sometimes get diluted. Clear, well-defined problem statements are essential for effective analysis.

29:53  
Alexey  
Who are the people making these requests, usually? I assume they are subject matter experts who aren't necessarily technical, right?

## Data analysis for transportation policies

30:09  
Rachel  
Yes, they could be subject matter experts, like those working in policy-making. They need data to develop data-driven policies.

30:21  
Alexey  
Can you give an example of a policy that might need data analysis?

30:27  
Rachel  
Sure, one example is analyzing fare card data to determine appropriate transportation pricing. We might want to know how many people use concession cards, like those for senior citizens or students, to evaluate the effectiveness of a new policy or fare adjustment.

31:11  
Alexey  
What is a concession card?

31:00  
Rachel  
A concession card provides discounted fares. In Singapore, for instance, senior citizens and students can use these cards to pay reduced rates on public transportation.

31:11  
Alexey  
So, policy questions might include things like what happens if we increase fares or introduce a new ticket option? For example, in Berlin, there are discounted bundles of tickets. You might need data to predict the impact of similar changes.

31:37  
Rachel  
Exactly. In Singapore, we have monthly passes offering unlimited travel for a fixed price, with specific options for students. We might analyze how many students use these passes and how effective they are in encouraging public transport use.

32:05  
Alexey  
And since people need to tap their cards to use transportation, you have all this data. But policy specialists might not know how to query it. Are you working on making this easier for them, perhaps through a chat interface that translates their questions into SQL queries?

32:35  
Rachel  
Yes, that's something we're looking to build. It's still in development, but we're working on tools that can help extract information and perform analysis more intuitively, potentially using plain language queries.

32:53  
Alexey  
Is this a project you're actively working on now?

32:59  
Rachel  
Yes, it's one of the projects we're currently focusing on.

## Technologies used in text-to-SQL projects

33:19  
Alexey  
What technologies or approaches are you using for this project? It sounds like a fascinating application of large language models (LLMs), and it might interest our LLM Zoomcamp students to learn about its practical uses.

33:48  
Rachel  
In the text-to-SQL space, we're looking at using metadata from data warehouses and data catalogs. We chunk this information and create a vector database, then apply a large language model (LLM) on top of it. The LLM takes user queries in plain English, translates them into SQL statements, and returns outputs to the users.

34:00  
Alexey  
So, it's like a retrieval-augmented generation (RAG) setup. A user provides a plain English query, and you use context from metadata to generate the SQL query. The LLM then executes the query and returns the results. Is that how it works?

34:43  
Rachel  
Yes, exactly. We use RAG techniques. The metadata helps the LLM understand which tables to refer to for the correct information extraction.

35:09  
Alexey  
How often do the queries generated by this process fail or need correction?

35:18  
Rachel  
I don't have exact numbers, but it does happen. Success largely depends on effective prompt engineering. Providing sufficient examples of text-to-SQL conversions helps guide the LLM. It’s also important to restrict certain types of SQL commands, like insert, update, or delete statements, to prevent unintended database modifications.

## Handling large datasets and transportation data in Singapore

36:12  
Alexey  
That's a fascinating project. I'm always keen to learn about new LLM use cases because there are so many possibilities. It's great that you're exploring this space too. This project is in an early development phase, right?

36:30  
Rachel  
Yes, it’s still in early development.

36:32  
Alexey  
How large are the datasets you're working with for these projects? Besides the text-to-SQL project, are there other more established projects you're involved in? You mentioned transportation data — like fare card data. How large are these datasets?

37:11  
Rachel  
We collect a significant amount of data. For fare card data alone, we gather millions of passenger flow records daily in Singapore. This provides numerous data points for analyzing passenger movements, identifying peak travel times, and assessing route popularity. This data is crucial for optimizing routes and fare structures.

For example, in Singapore, the Public Transport Council implemented a "morning pre-peak fare" policy, offering savings for commuters who travel before 7:45 AM. This aims to shift demand away from peak hours, balancing train load and encouraging earlier travel. By analyzing fare card data, we can evaluate the effectiveness of such policies.

38:34  
Alexey  
Regarding these millions of data points you collect, how are they processed? Do you use standard tools like Kafka, data lakes, and data warehouses? Also, how does data from the buses get collected? Are there transmitters on the buses that send data to Kafka or another system?

39:04  
Rachel  
Yes, we have sensors on the buses that collect location information. We also gather data from fare card systems at the entry and exit points.

39:14  
Alexey  
Is this data collected in real-time? For example, the moment I tap my card, does an event get recorded immediately, or is the data only gathered at the end of the day?

39:27  
Rachel  
The data is collected in real-time, but aggregation happens afterward. In Singapore, we use a system that defines a ride and a journey. Our fare structure allows commuters to make multiple transfers within a 45-minute period, which is considered one journey. We calculate fares based on the total distance traveled. So, while data is collected in real-time, processing and aggregation occur after a time lag. This allows us to combine all this information before storing it in a data warehouse. We use tools like Kafka and Apache Spark for processing.

40:17  
Alexey  
So, the events from the bus are sent immediately to your system. You don't need to wait until the bus completes its shift to connect and download the data, right?

40:36  
Rachel  
Yes, that's correct.

40:40  
Alexey  
That's great. You use Kafka, Apache Spark, and other common tools in data engineering, right?

40:48  
Rachel  
Yes, that's right.

41:00  
Alexey  
Have you ever had to go on a bus to fix a sensor?

41:04  
Rachel  
No, I haven't needed to do that.

41:08  
Alexey  
So, the sensors are quite reliable, right?

41:08  
Rachel  
Yes, generally they are reliable. However, part of data engineering involves detecting data quality issues or anomalies. For instance, if a transponder on a bus isn’t sending data correctly, we need to identify that issue. While I haven't personally gone to a bus to fix a sensor, maintaining data quality is a critical aspect of data engineering.

41:46  
Alexey  
I'm considering whether to dive deeper into traditional data engineering topics or explore generative AI applications. What would you like to focus on next?

42:11  
Rachel  
We could go more into how AI is being used.

## Generative AI applications beyond text-to-SQL

42:17  
Alexey  
Do you have other AI applications besides the text-to-SQL tool?

42:28  
Rachel  
The text-to-SQL tool is my main focus right now. However, generative AI has other potential applications, such as creating synthetic data. In projects where we don't have a full data set, generative AI could help generate synthetic data. Additionally, I believe generative AI could redefine user interfaces, making information retrieval more conversational and intuitive, moving away from traditional keyword searches to semantic searches. This approach could change how people search for things, not just in e-commerce but in various domains, like planning what to buy for a gift.

43:55  
Alexey  
I'm becoming more accustomed to using tools like ChatGPT with voice recognition. It's very convenient. Regarding synthetic data, how effective is generative AI at creating data, especially numerical data or time series data?

44:42  
Rachel  
I was speaking generally about the potential of generative AI for creating synthetic data, especially when dealing with complex or sensitive data sets. Generative AI could help create synthetic versions that mask confidential information while retaining essential characteristics.

## Publishing public data and maintaining privacy

45:26  
Alexey  
Since Singapore releases a lot of public data on open platforms, sometimes you might need to edit this data before publishing it, right?

45:40  
Rachel  
The current publicly shared data is collected from various systems, and we mask sensitive information, such as fare card numbers, before publishing.

## Recommended datasets and projects for data engineering beginners

45:52  
Alexey  
Where can people find this public data?

46:00  
Rachel  
Two main platforms provide public data in Singapore: data.gov.sg and DataMall. Data.gov.sg aggregates data from various government bodies, covering areas like rainfall, air pollution, transportation, and more.

46:20  
Alexey  
There are categories like arts, education, economy, environment, housing, health, social, transport, and real-time APIs. I assume I should look under the transport category to find relevant data, right?

46:41  
Rachel  
Yes, you’ll find transportation data, such as air travel and geospatial information, under the transport category.

46:50  
Alexey  
If someone is starting the next cohort of a data engineering course in January, what kinds of projects would you recommend they try using these data sets?

47:05  
Rachel  
One useful data set could be car parking data, as we collect real-time parking transaction data. It's a large and dynamic data set, ideal for real-time data ingestion, storage in a data warehouse or data lake, and subsequent analysis.

47:36  
Alexey  
How do I find the car parking data?

47:47  
Rachel  
Car parking data is typically part of our dynamic data sets.

47:51  
Alexey  
There are real-time APIs. Maybe that’s what I need to look for. Could you send me the link later?

48:08  
Rachel  
Sure, I can send it to you. Another valuable data set is real-time taxi data, which can be useful for examining data engineering processes. In previous data engineering courses, we used the New York taxi data set, which is aggregated. Real-time data offers another layer of complexity and learning opportunities.

48:35  
Alexey  
I see public transport capacity, taxi population data, and more, some of which are updated regularly. You've sent me another link, right?

49:03  
Rachel  
Yes, I've sent another link, specifically for transportation data under the dynamic data sets, including real-time taxi availability and parking information.

## Recommended resources for learning urban data science

49:16  
Alexey  
Great, I'll add this link to our description. Thanks. I don't see any questions from the audience at the moment. I'm wondering if someone wants to study what you do — urban data science and transport planning — what resources, books, or courses would you recommend?

50:22  
Rachel  
The courses by DataTalks.Club are a good primer. For a deeper understanding of urban data science, the book "The Death and Life of Great American Cities" by Jane Jacobs is a classic. It critiques traditional urban planning and advocates for vibrant, livable cities through community-based approaches and human-scale design. Another book is "Happy City: Transforming Our Lives Through Urban Design" by Charles Montgomery. This book blends urban planning with psychology and sociology, exploring how thoughtful urban design can improve happiness and well-being. Both books offer insights into why we study urban data and plan cities thoughtfully.

51:42  
Alexey  
Those sound interesting. The first one is "The Death and Life of Great American Cities," and the second is "Happy City," right?

52:00  
Rachel  
Yes, that's correct.

52:05  
Alexey  
I remember visiting the United States, where getting around without a car is challenging, except in cities like New York or Boston. It would be great if some of the practices from Europe or Singapore were implemented there. I recall once being picked up by the police while walking down a road without sidewalks because I wasn't used to the idea that you need a car to move around.

52:40  
Rachel  
That's quite amusing!

52:42  
Alexey  
Anyway, Rachel, thank you so much for joining us today and sharing your experiences. I’ve learned a lot. These topics were new to me, and it was enlightening to hear about your work. Thanks for being here, and thanks to everyone else who joined us today. This was our first episode after a break, and it was a great start.

53:09  
Rachel  
Thank you for having me.

53:11  
Alexey  
I'm happy to hear about your transition to data engineering and that our course helped. It means a lot to see success stories like yours. Thanks again!

53:26  
Rachel  
Thank you.

53:28  
Alexey  
Have a wonderful rest of your day, Rachel, and to everyone else — see you around!