HACKWAGON
• ACADEMY •

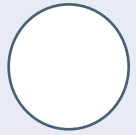
# DATA SCIENCE 101:
# INTRODUCTIONS TO ANALYTICS

# AGENDA

- Data Analytics Vs Data Science
- What brought about the emergence of Data Science?
- Case studies
- Economic benefits of data analytics

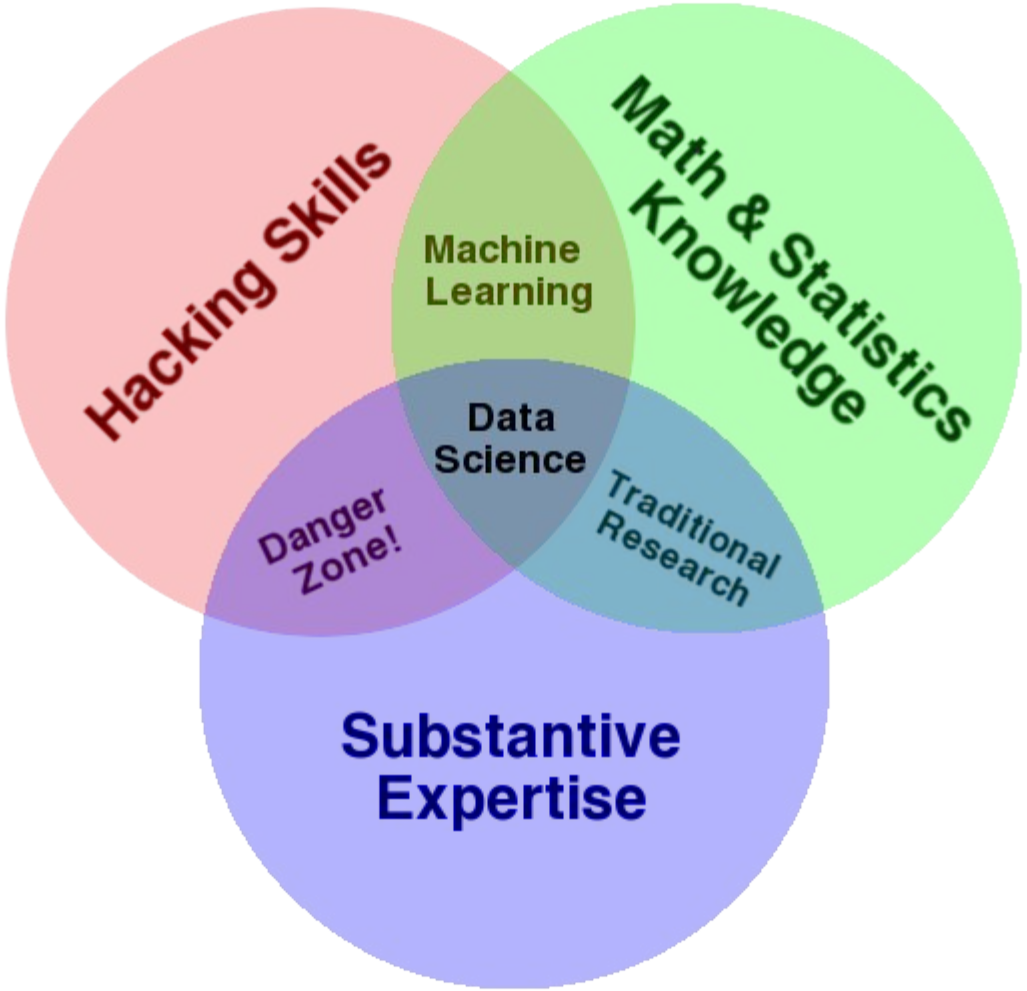
# WHAT IS DATA ANALYTICS & DATA SCIENCE

- **Descriptive Analytics**: summarize raw data and make it something that is interpretable by humans. It describe the past and answer: “What has happened?” (Example: Annual Sales for Shoes in a Adidas)
- **Predictive Analytics**: Using statistical models and forecast techniques to understand the future and answer: “What could happen?” (Example: Predicted sales for shoes in the next quarter)
- **Prescriptive Analytics**:  Using optimization and simulation algorithms to advice on possible outcomes and answer: “What should we do?”
(Example: How to allocate stocks in the right stores to increase sales)
- **Machine Learning**: The science of getting computers to act without being explicitly programmed. Basically Machine Learning is concerned with the methods and models

| | Data Analytics | Data Science |
|---|---|---|
| Descriptive |  |  |
| Predictive |  |  |
| Prescriptive |  |  |
| Machine Learning |  |  |

# THE DATA SCIENCE VENN DIAGRAM

- **Hacking Skills**: Ability to code to develop algorithms to efficiently store, process and visualise data. *(What we teach you in DS101)*
- **Maths & Statistics Knowledge:** Ability to form and apply mathematical models, and summarize these to any given datasets.
- **Substantive Expertise**: Domain expertise to form the right business scenarios and questions



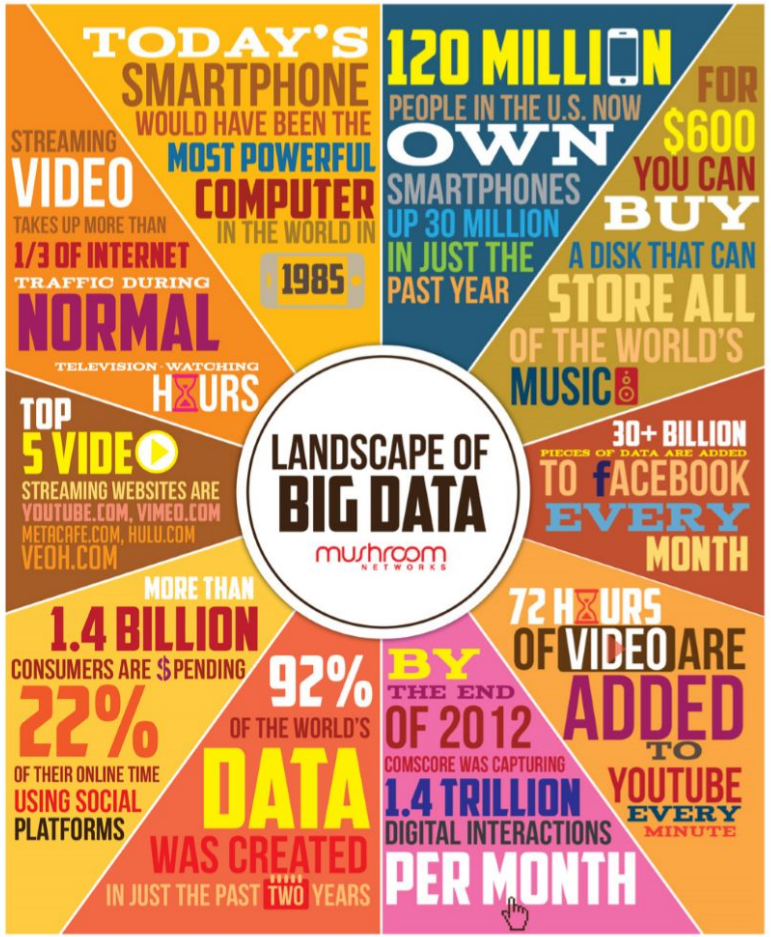

Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

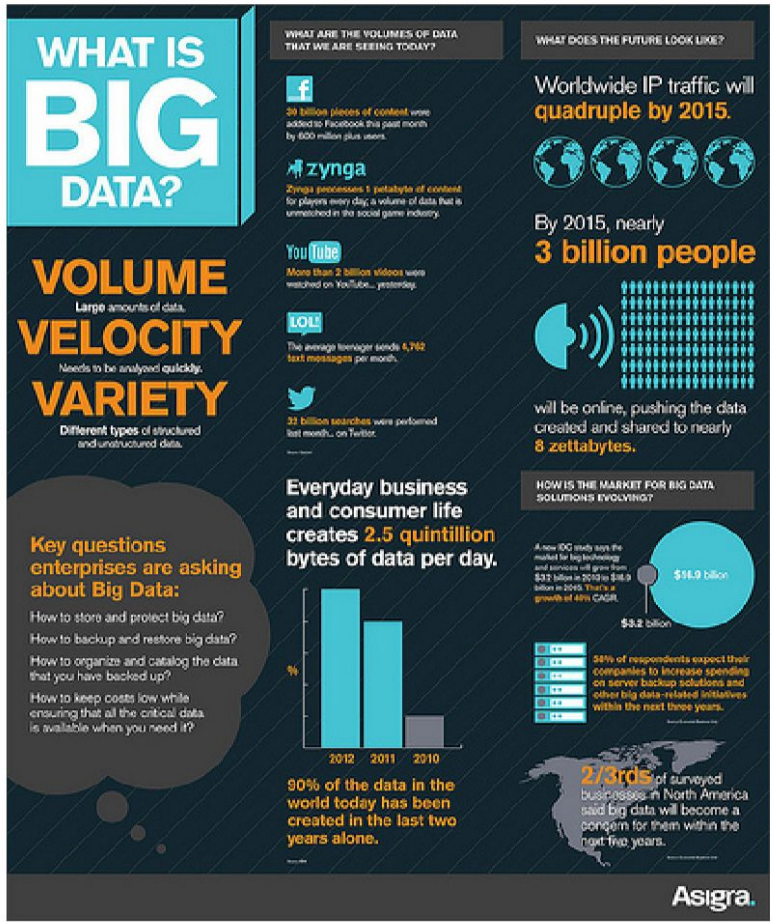
# DATA SCIENCE PROJECT STAGES

| Problem Specification | Data Gathering & Preprocessing | Descriptive Analytics | Machine Learning | Deployment |
|---|---|---|---|---|
| • Understand business scenario<br>• Define the project problem and scope<br>• Define limitations of the project | • Develop a system to gather data<br>• Clean and prepare raw data for processing<br>• Usually the most time consuming stage in a data science project | • Exploratory Data Analysis<br>• Basic understanding of the dataset<br>• Answer the initial assumptions you may have of the data | • Depending on the scope/nature of your project, apply necessary machine learning models to tackle the problem<br>• Train the machine learning model and assess its' performance | • Consult with project stakeholders on suitability of model<br>• Deploy model for live usage |

# EMERGENCE OF DATA ANALYTICS

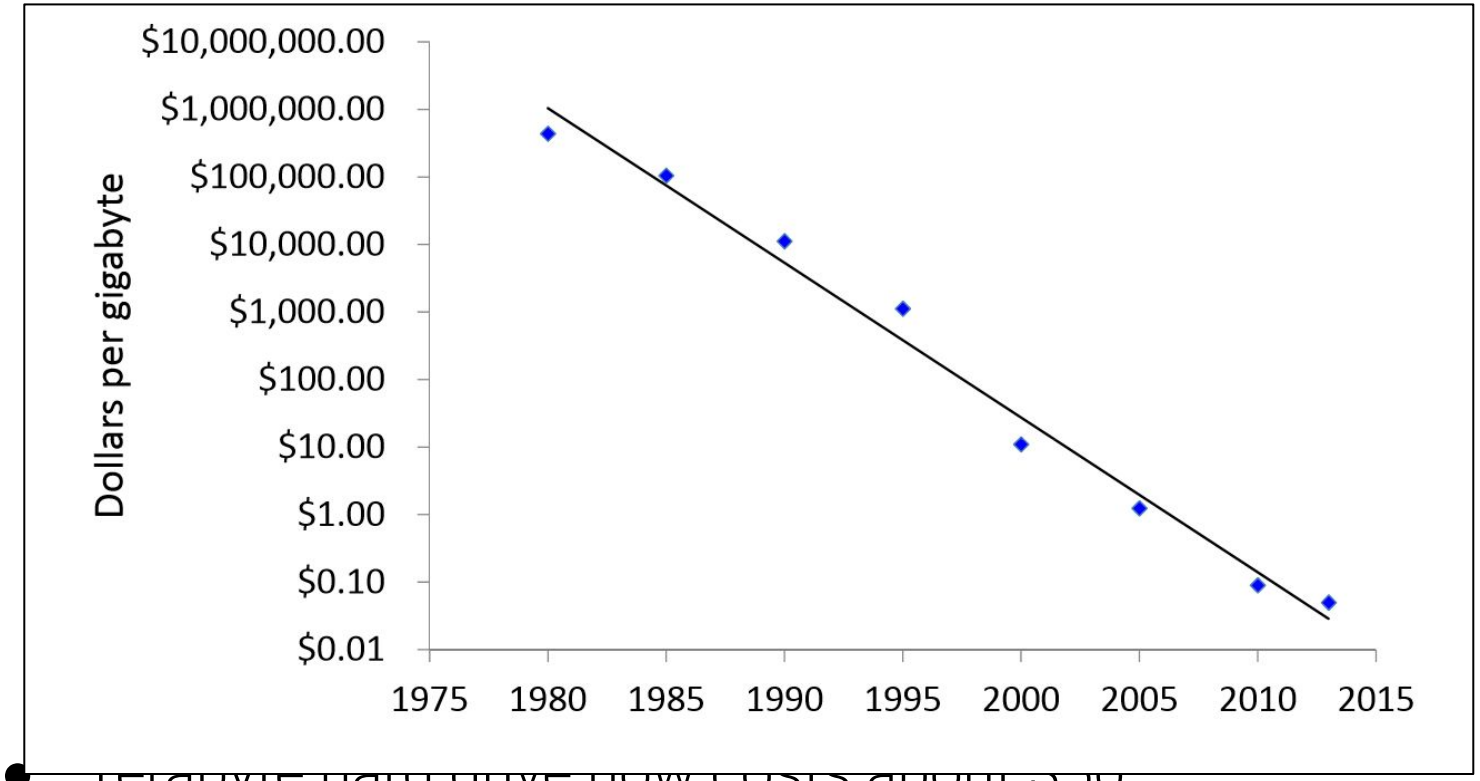
- Surge in available data, which was brought about by cheap cost of computing



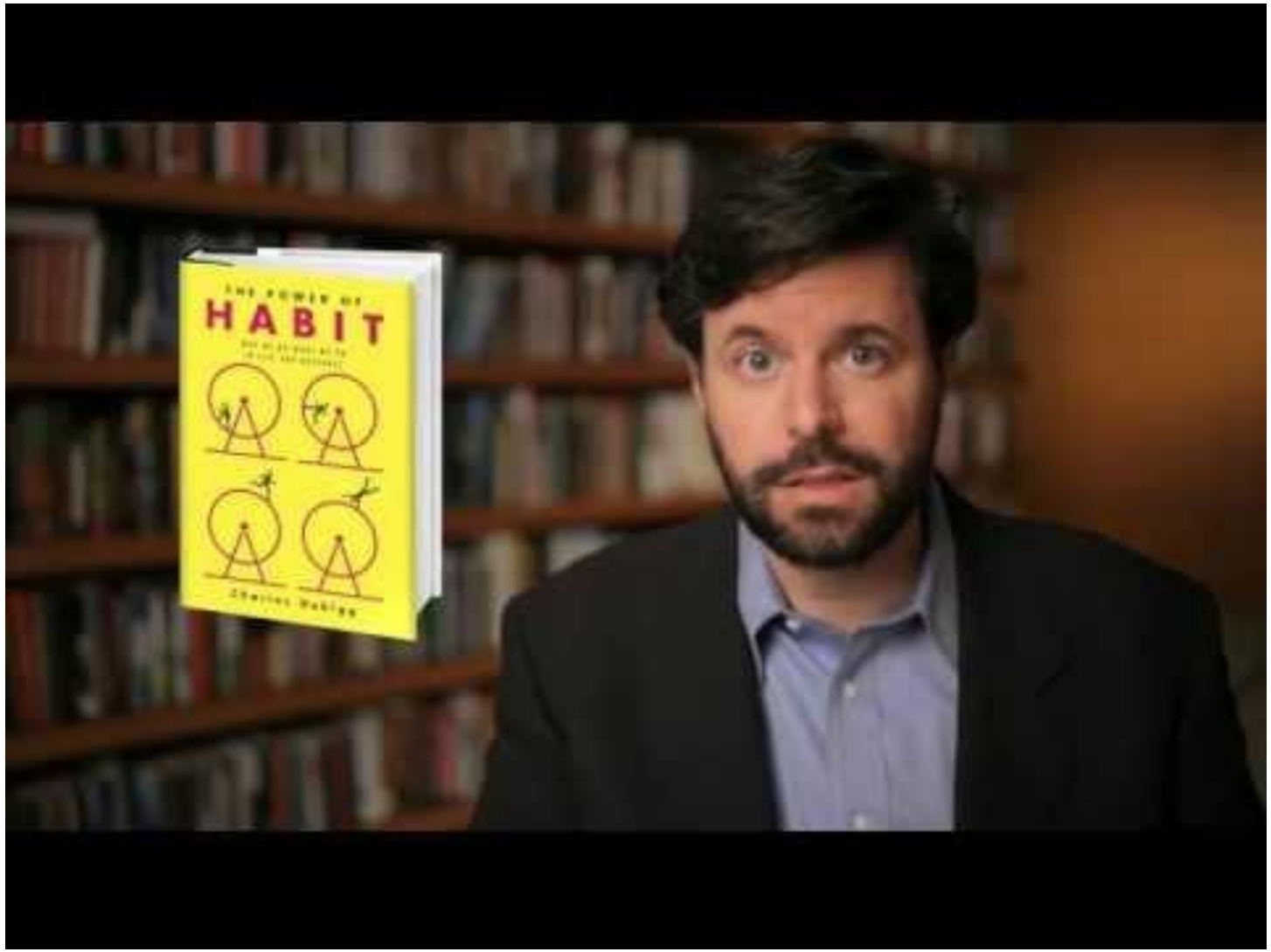
# EMERGENCE OF DATA SCIENCE & ANALYTICS

- Computing cost is plummeting => can throw more computing power to perform intensive predictive analytics




- Terabyte hard drive now costs about $50

# CASE STUDY 1: TARGET TARGETING MOTHERS-TO-BE



# CASE STUDY 1: TARGET TARGETING MOTHERS-TO-BE

## to our shareholders

2012 was an exciting year for Target, as we devoted meaningful resources to driving performance in support of our publicly stated sales and financial goals, while transforming Target to seize the tremendous opportunities we see in the most dynamic and disruptive retail landscape in generations.

Total sales and diluted earnings per share reached new highs of $72.0 billion and $4.52, respectively. We invested $3.3 billion of capital in our U.S. and Canadian businesses, and we returned over $2.7 billion to our shareholders through share repurchase and dividend payments. And our full-year results were right on track with our Long-Range Plan to reach at least $100 billion in sales and $8 in earnings per share in 2017.

In addition to our financial successes, we achieved significant strategic and operational milestones in 2012, including the launch of our first CityTarget stores in Chicago, Seattle, San Francisco and Los Angeles; extending our fresh-food remodel program to another 238 general merchandise stores; achieving record-setting sales penetration through our 5% REDcard Rewards loyalty program; and announcing an agreement to sell our U.S. credit card receivables to TD Bank Group, a strong partner aligned with our goals for portfolio growth and profitability. We also surpassed $4 million per week in charitable giving to support communities we serve.

In 2013, we'll continue to pursue a strategy that is being shaped by our guests' expectations for more shopping flexibility and price transparency, and the rapid pace of change in technology. To ensure that we continue to strengthen our guests' love for our brand and deliver the surprise and delight they have come to expect, we'll leverage our greatest asset, our stores, in combination with increased investment in our digital platforms, to create a seamless, relevant and personalized experience.

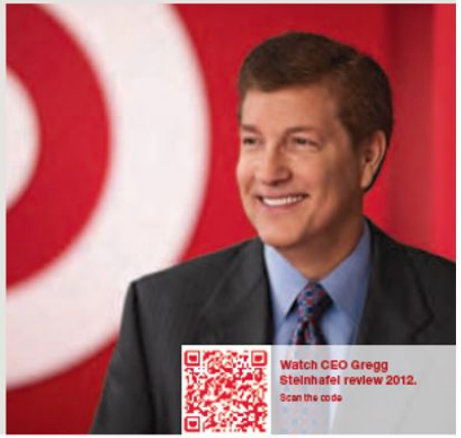
Differentiation with exceptional value, which represents the foundation of our "Expect More. Pay Less." brand promise, will continue to set Target apart in the marketplace. We remain committed to offering a truly unique assortment—through our design partnerships, outstanding portfolio of owned brands and curated selection of signature national brands. And we are equally unwavering in our commitment to provide a compelling value proposition, as we showed by expanding our Price Match Guarantee to include select online competitors. We're also collaborating closely with our vendors on channel-management strategies that sharpen prices and improve selection for our guests.

Meanwhile, in 2013, we will also undertake the largest, single-year store expansion in Target's history. After two years of exceptional dedication and hard work by our team, we've begun opening Target stores in Canada and are on track to open 124 stores across all 10 provinces by year end. In addition, we'll extend our new CityTarget urban format to additional locations in Los Angeles and San Francisco, and, for the first time, to Portland, Oregon.

We are excited by the physical and digital growth ahead of us. By staying focused on creating a superior shopping experience for our guests—whether they are in urban or suburban markets, in the U.S. or Canada, in our stores or digital channels—we believe Target will continue to thrive. And, this strategic clarity, in combination with our powerful brand, gives us confidence in our future: confidence in the values that have guided our company for 50 years, confidence in the talent and passion of our 361,000 team members and confidence in our continued ability to deliver profitable growth for many years to come.

Gregg Steinhafel | Chairman, President and CEO, Target

**Board of Directors Changes:** In March 2013, we welcomed Douglas M. Baker Jr., Chairman and CEO of Ecolab, Inc., and Henrique De Castro, Chief Operating Officer of Yahoo! Inc., to our board of directors. Also in March, Stephen W. Sanger, former Chairman and CEO of General Mills, Inc., retired from our board of directors. We thank Steve for his contributions during his 17 years of service.




- Strong revenue growth from $44B in 2002 to $67B in 2010
- CEO Steinhafel: results due to "**heightened focus on items and categories that appeal to specific guest segments such as mom and baby**"

# CASE STUDY 2: Amy Webb How I Hacked Online Dating

# CASE STUDY 3: CIRCLE LINE ROGUE TRAIN

- Singapore's MRT Circle Line was hit by a spate of mysterious disruptions in 2016, causing much confusion and distress to thousands of commuters.
- From prior investigations by SMRT and LTA: Incidents were caused by some form of signal interference, which led to loss of signals in some trains. The signal loss would trigger the emergency brake safety feature in those trains and cause them to stop randomly along the tracks.
- A team of data analysts and scientists were given a dataset compiled by SMRT:
  - Date and time of each incident
  - Location of incident
  - ID of train involved

Source: https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a

# CASE STUDY 3: CIRCLE LINE ROGUE TRAIN

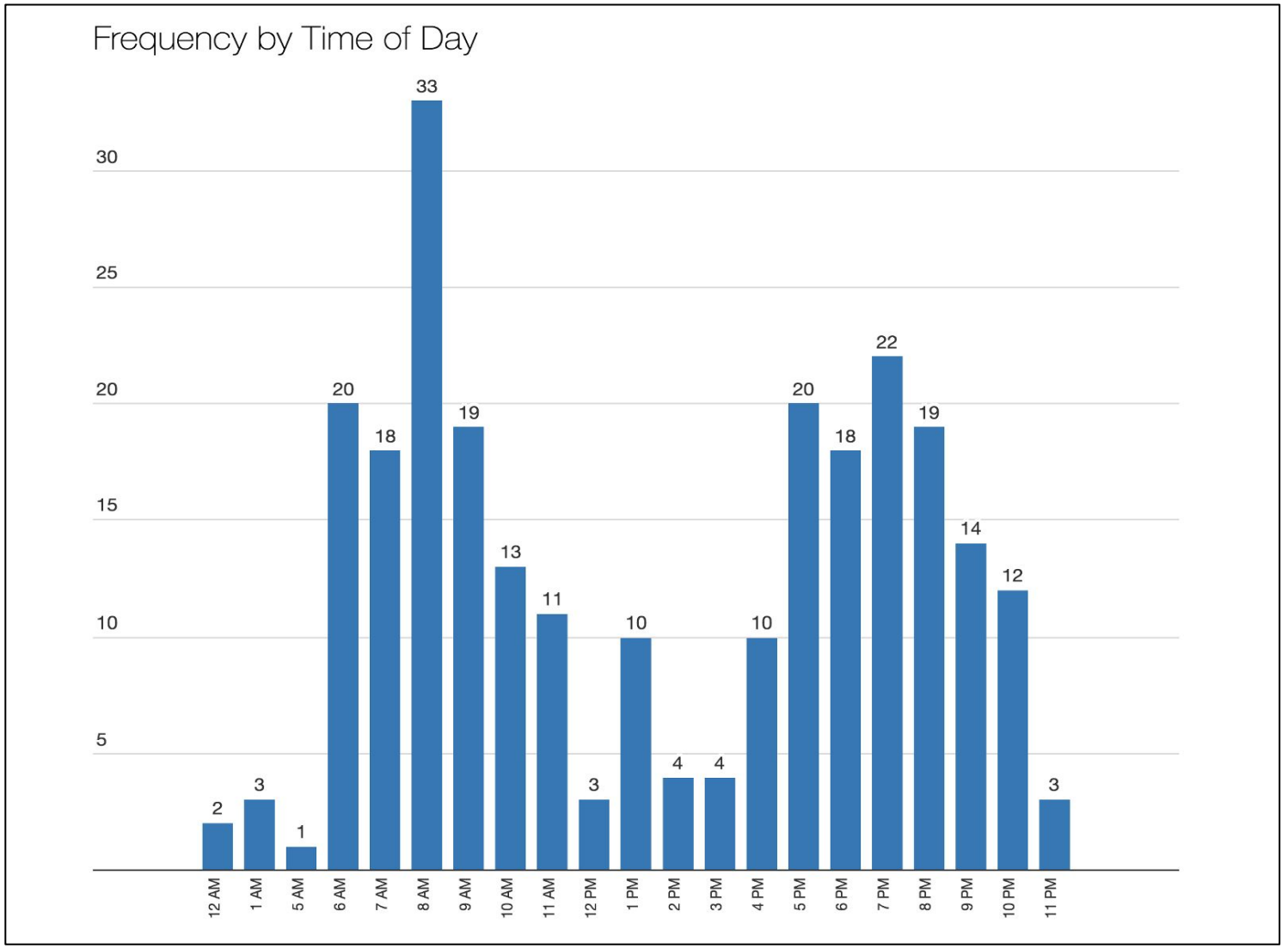
- After performing data cleaning:

| | S/N | S/N.1 | Date | Traffic Date | PV | Time | Bound | Station from | Station to | Event | Remarks | DateTime |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 1 | 2016-08-28 | 2016-08-28 | PV40 | 19:32:00 | OT | KRG | ONH | EB | point track | 2016-08-28 19:32:00 |
| **1** | 2 | 2 | 2016-08-28 | 2016-08-28 | PV53 | 19:39:00 | OT | LBD | PPJ | EB | point track | 2016-08-28 19:39:00 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **257** | 57 | 42 | 2016-11-04 | 2016-11-04 | PV13 | 22:29:00 | IT | SDM | MBT | EB | NaN | 2016-11-04 22:29:00 |
| **258** | 58 | 43 | 2016-11-05 | 2016-11-05 | PV43 | 00:07:00 | IT | TLB | HBF | EVAC | Withdrawal train, no pax on-board | 2016-11-05 00:07:00 |

Source: https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a

# CASE STUDY 3: CIRCLE LINE ROGUE TRAIN

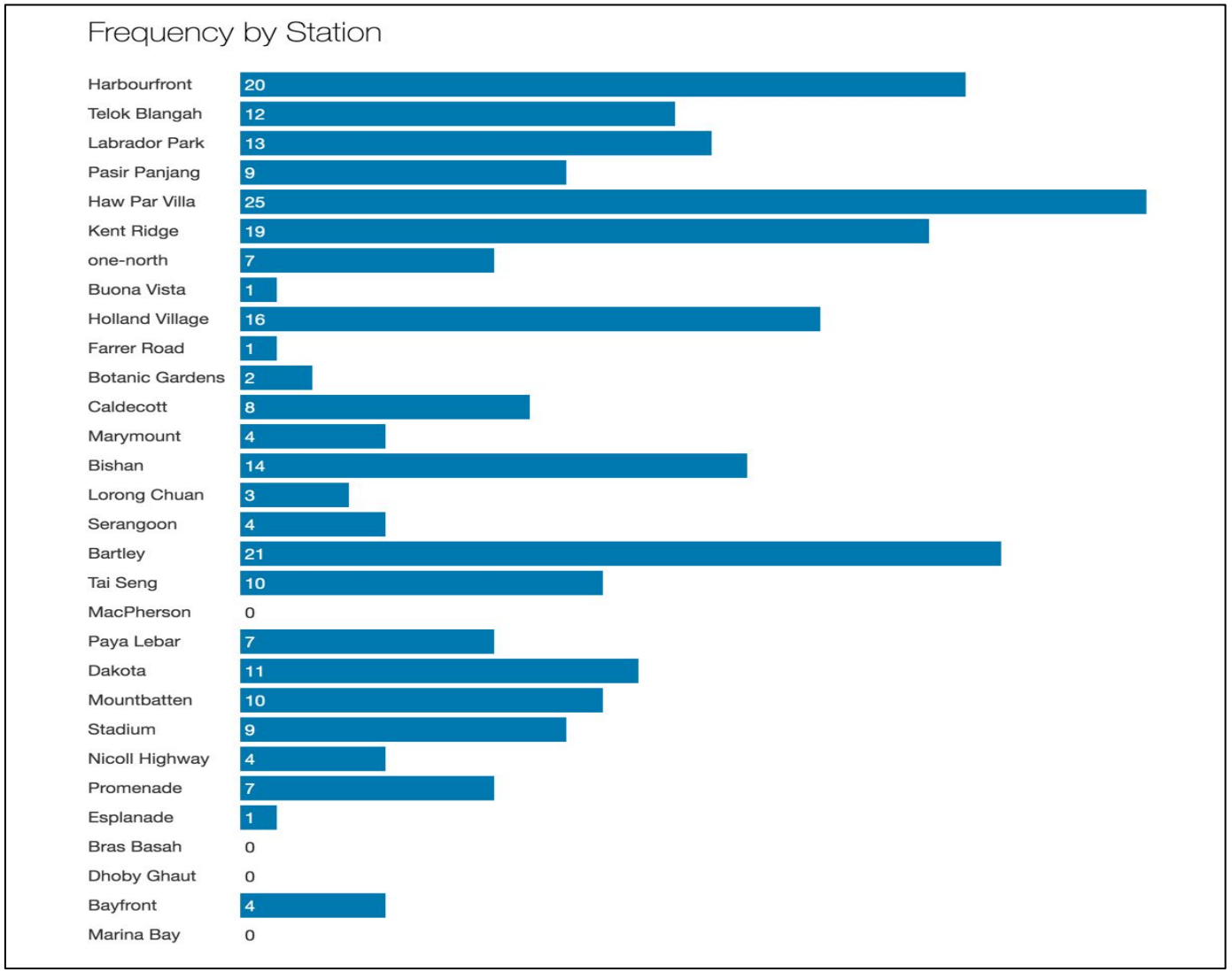
- Initial Visualization: No clear answer
- The incidents were spread throughout a day, and the number of incidents across the day mirrored peak and off-peak travel times




Source: https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a

# CASE STUDY 3: CIRCLE LINE ROGUE TRAIN

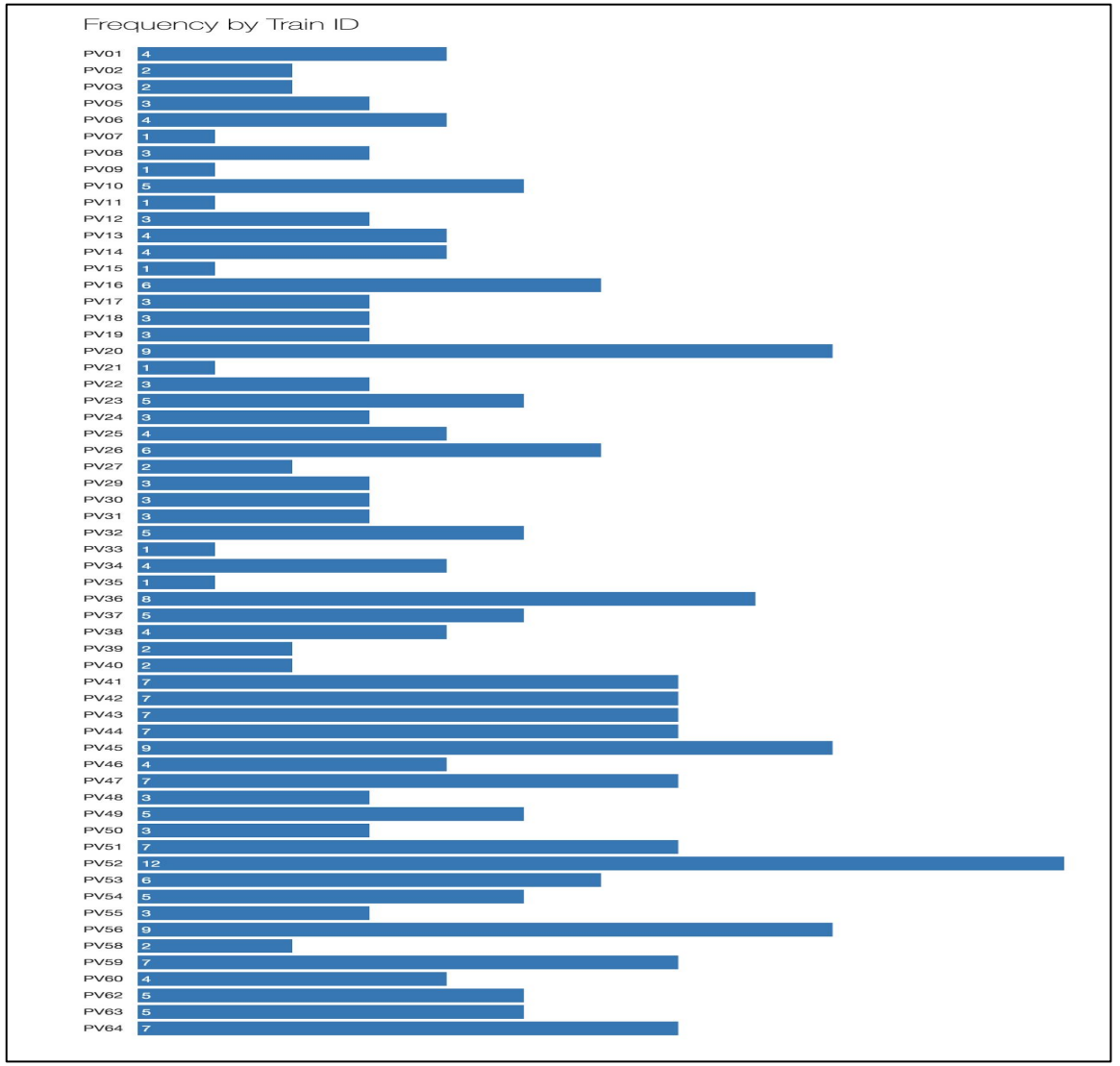
- The incidents happened at various locations on the Circle Line, with slightly more occurrences on the west side




Source: https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a

# CASE STUDY 3: CIRCLE LINE ROGUE TRAIN

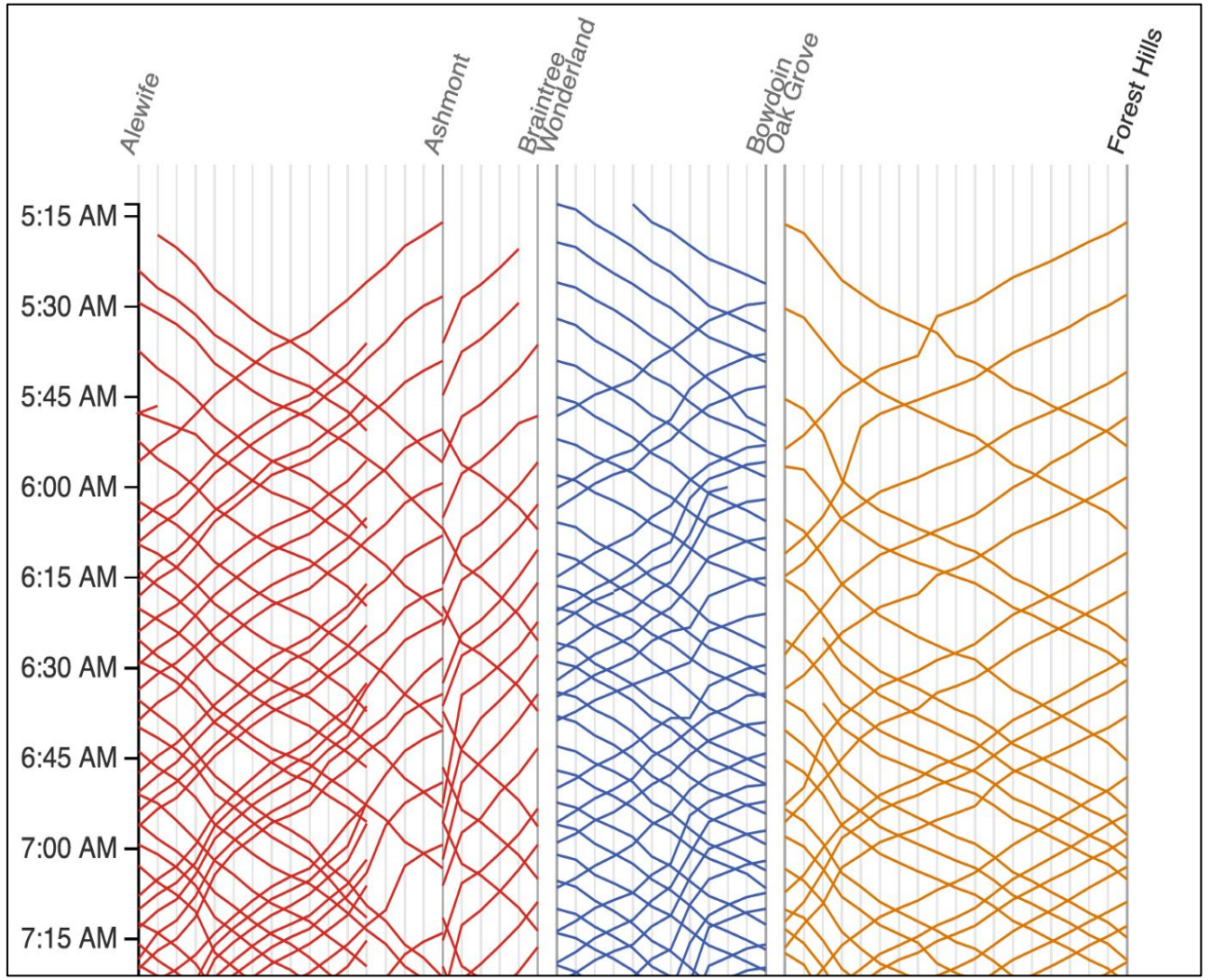
- The signal interferences did not affect just one or two trains, but many of the trains on the Circle Line. “PV” is short for “Passenger Vehicle”




Source: https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a

# CASE STUDY 3: CIRCLE LINE ROGUE TRAIN

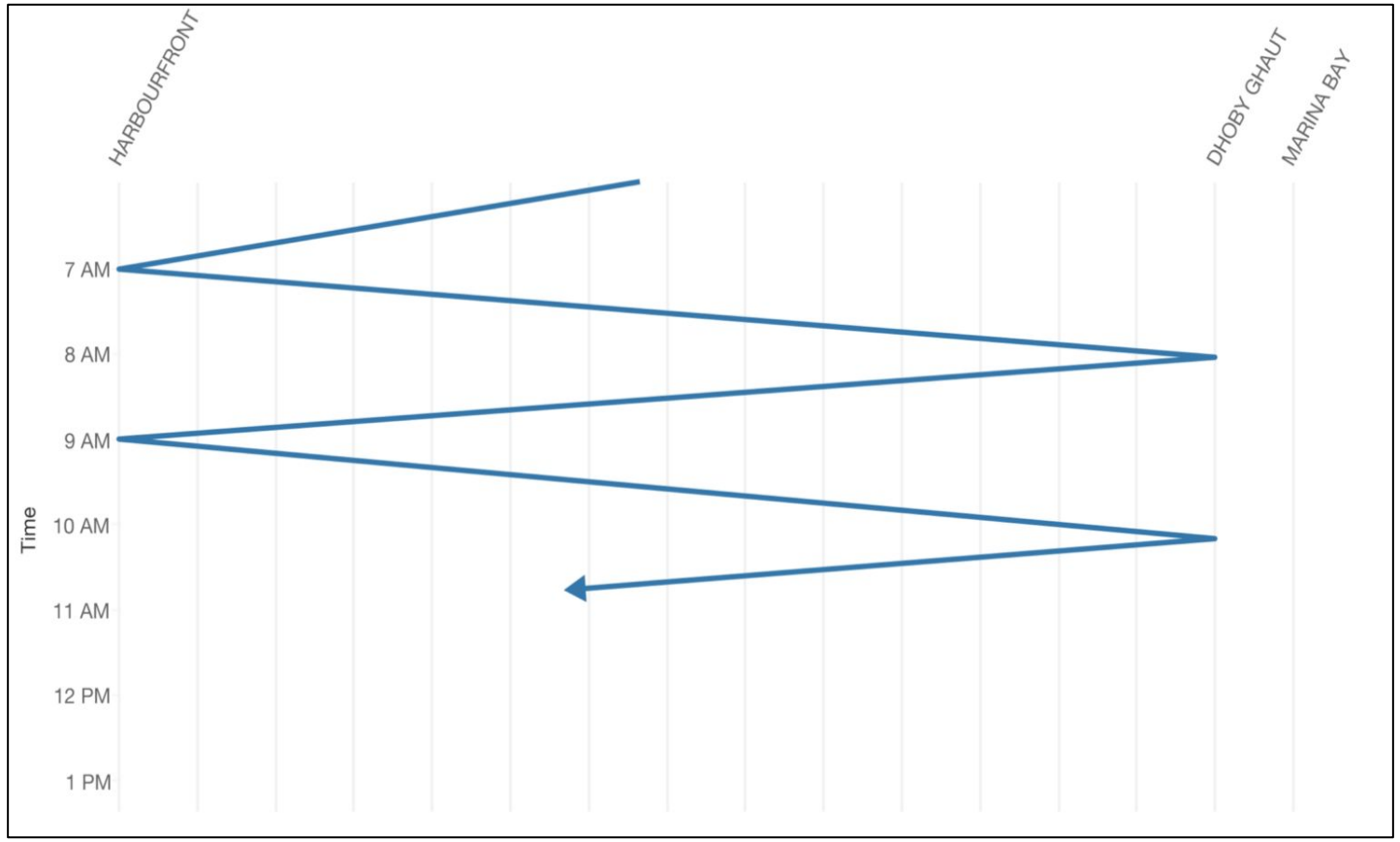
- Next step: Incorporate multiple dimensions into the exploratory analysis
- Inspired by the Marey Chart, which was featured in Edward Tufte's vaunted 1983 classic **The Visual Display of Quantitative Information**. More recently, it was used by Mike Barry and Brian Card for their **extensive visualisation project** on the Boston subway system:




Source: https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a

# CASE STUDY 3: CIRCLE LINE ROGUE TRAIN

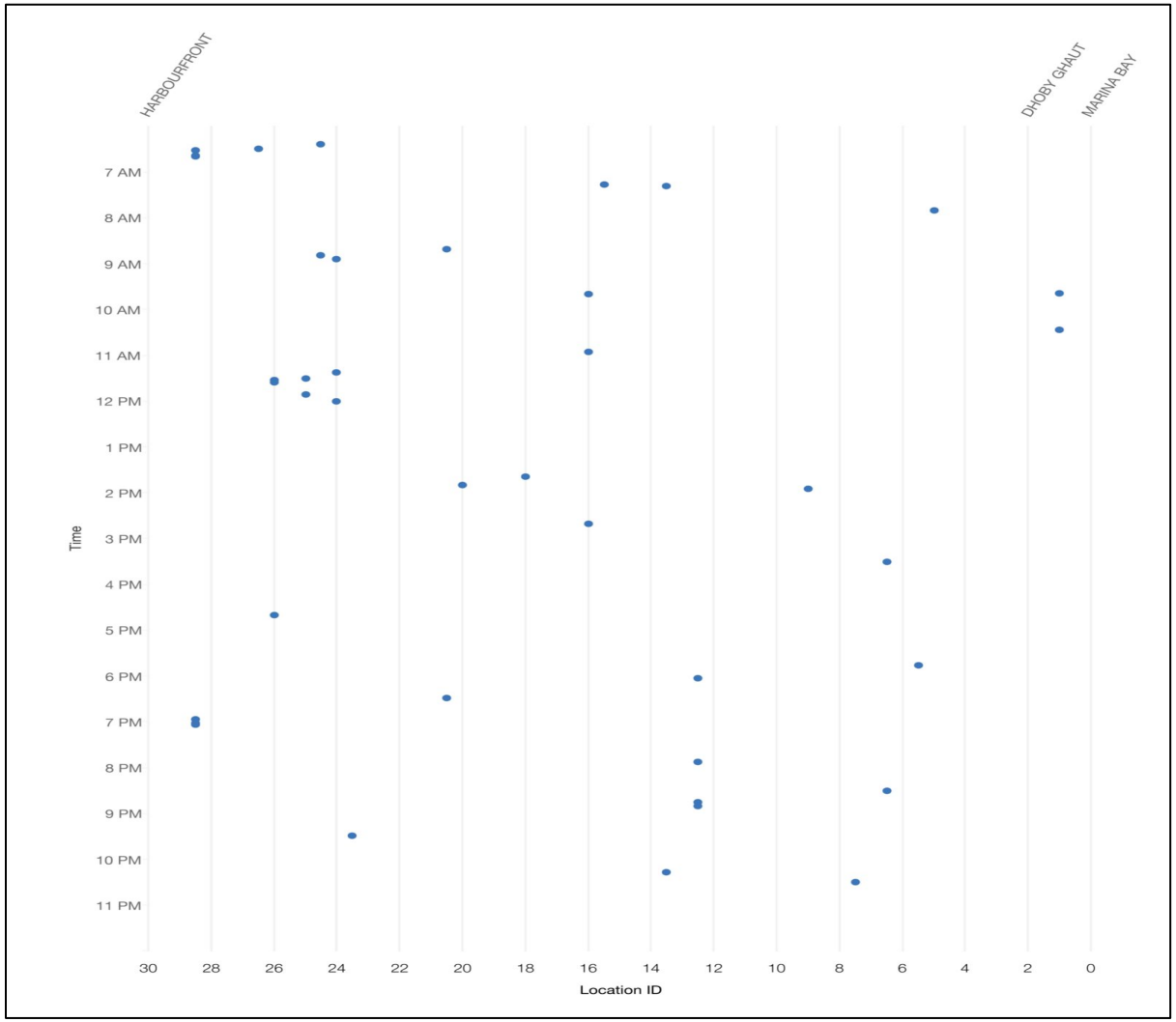
- In this chart, the vertical axis represents time—chronologically from top to bottom—while the horizontal axis represents stations along a train line. The diagonal lines represent train movement
- Under normal circumstances, a train that runs between HarbourFront and Dhoby Ghaut would move in a line similar to this, with each one-way trip taking just over an hour
- **Our intention was to plot the incidents—which are points instead of lines—on this chart**




Source: https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a

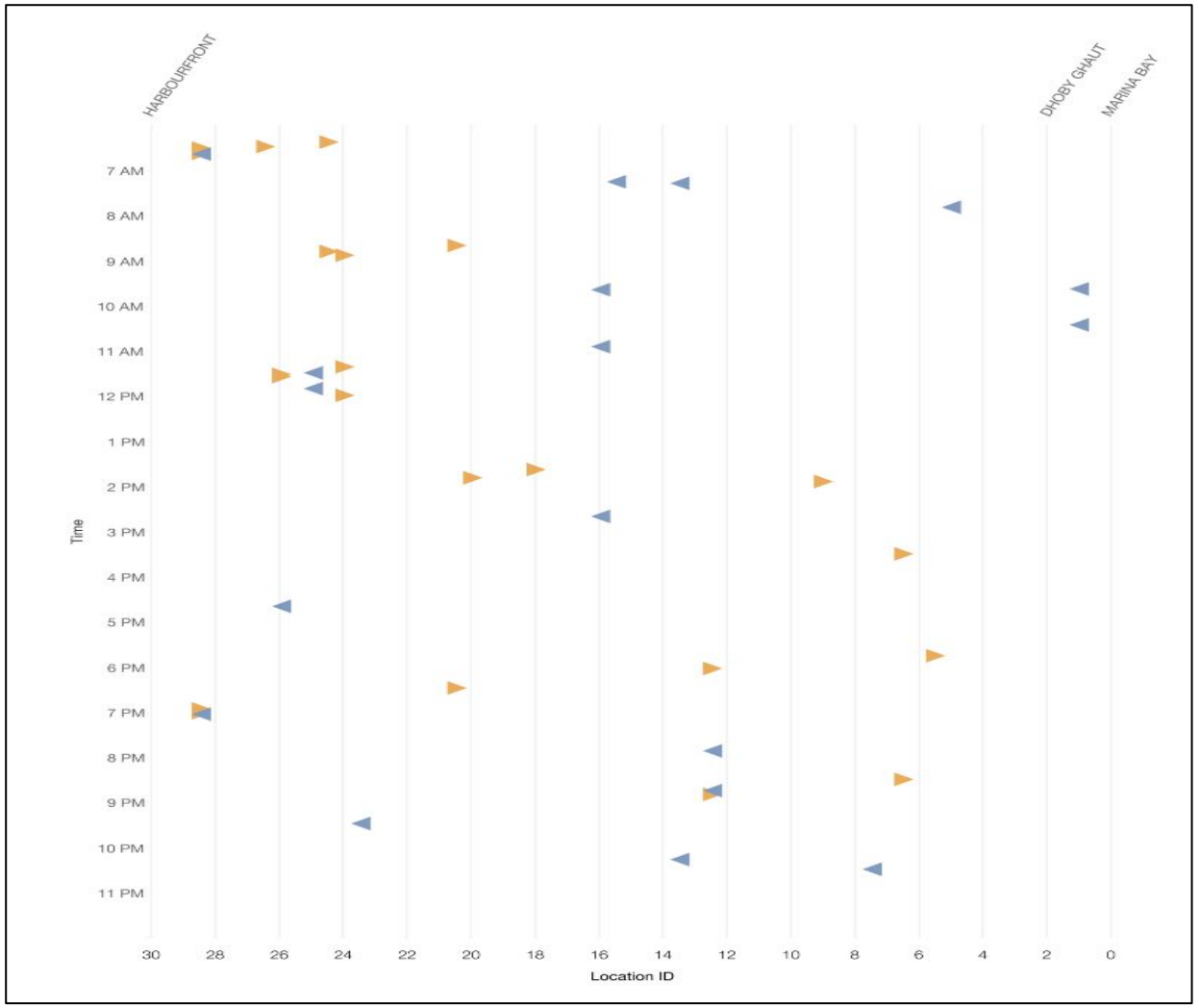
# CASE STUDY 3: CIRCLE LINE ROGUE TRAIN

- Further data processing…
- If the incident occurred between two stations, it would be denoted as 0.5 + the lower of the two station numbers.
- For example, If an incident happened between HarbourFront (number 29) and Telok Blangah (number 28), the location would be “28.5”. This made it easy for us to plot the points along the horizontal axis
- Create a scatterplot of all the emergency braking incidents. Each dot here represents an incident. Once again, we were unable to spot any clear pattern of incidents.



# CASE STUDY 3: CIRCLE LINE ROGUE TRAIN

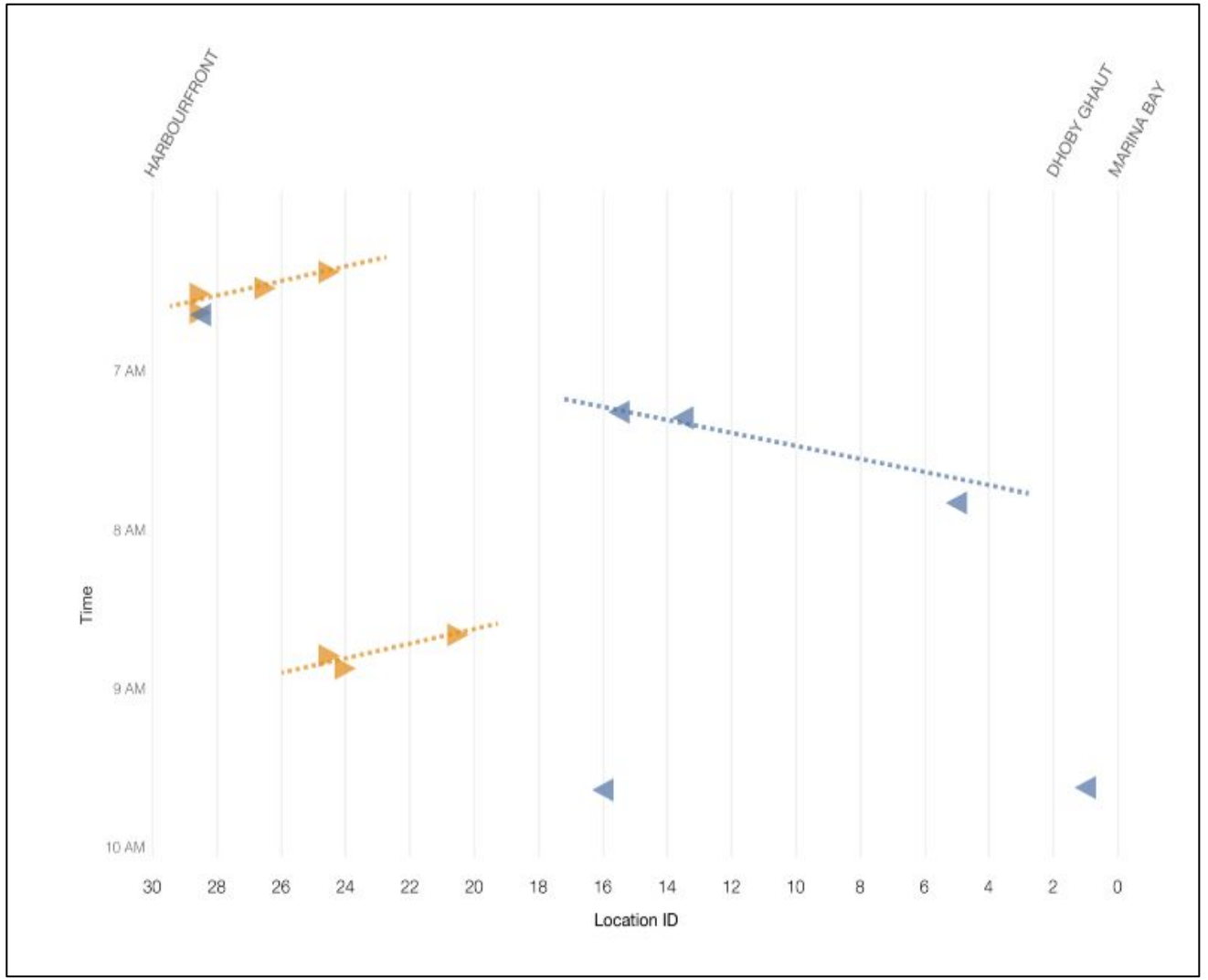
- Add train direction to the chart by representing each incident as a triangle pointing to the left or right, instead of dots
- Still looks random...




Source: https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a

# CASE STUDY 3: CIRCLE LINE ROGUE TRAIN

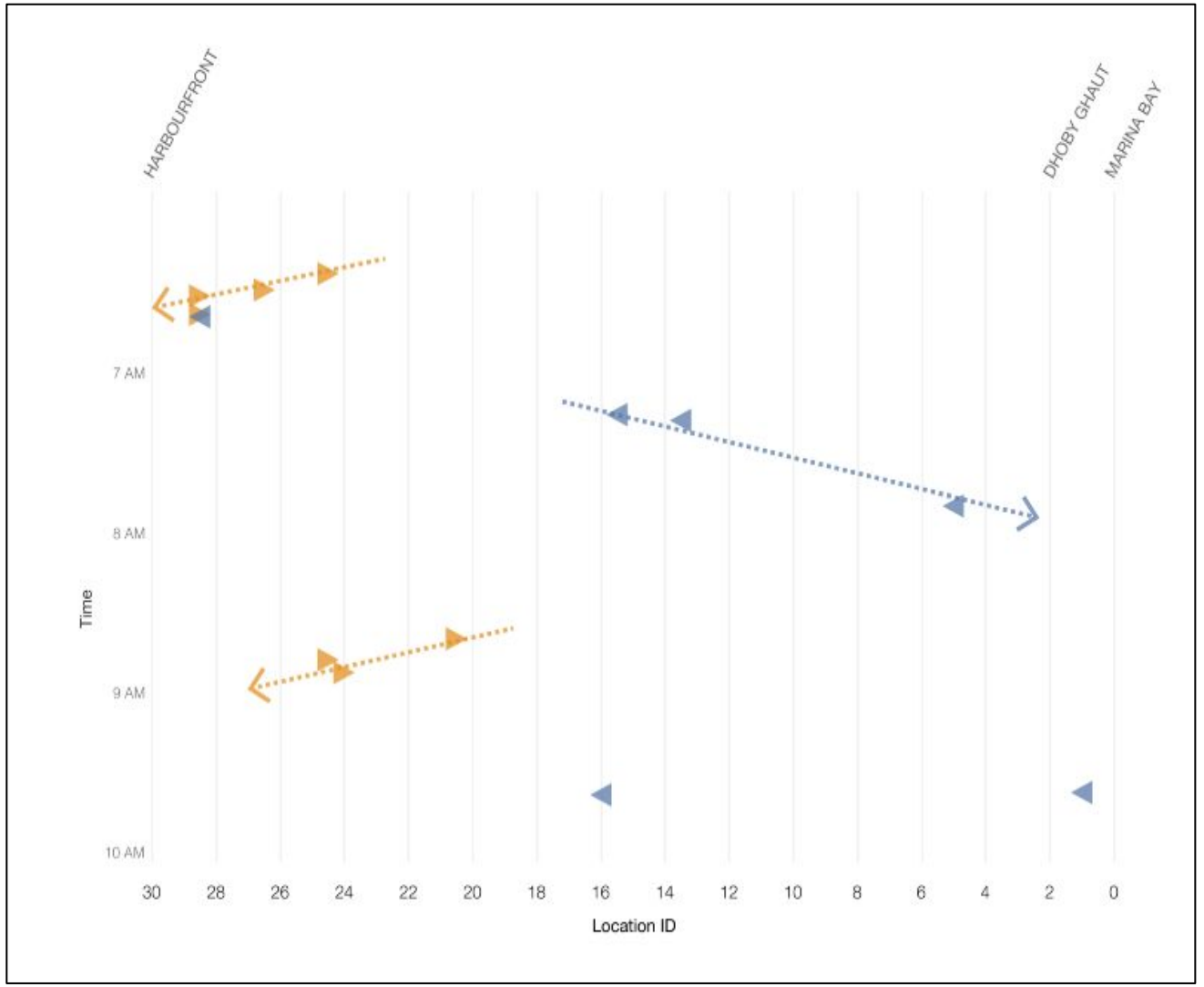
- When zoom in…
- Notice that the breakdowns seem to happen in sequence. When a train got hit by interference, another train behind moving in the same direction got hit soon after...




Source: https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a

# CASE STUDY 3: CIRCLE LINE ROGUE TRAIN

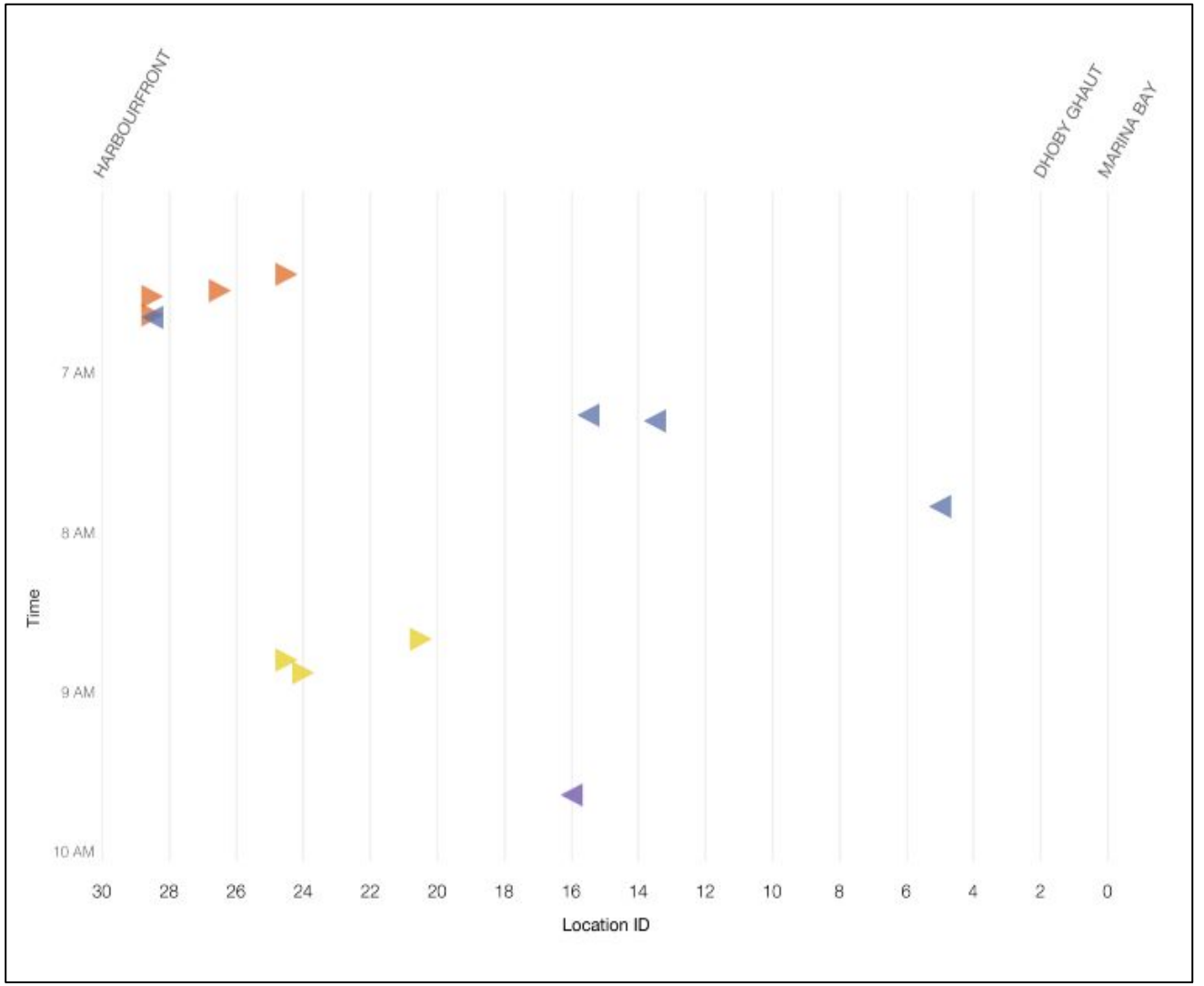
- And the ponder…
- Incidents were happening one after another, in the opposite direction of the previous incident. It seemed almost like there was a “trail of destruction”. **Could it be something that was not in our dataset that caused the incidents?**
- Imaginary lines connecting the incidents looked suspicious. **Could the cause of the interference be a train—in the opposite track?**
- Rouge Train Hypothesis




Source: https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a

# CASE STUDY 3: CIRCLE LINE ROGUE TRAIN

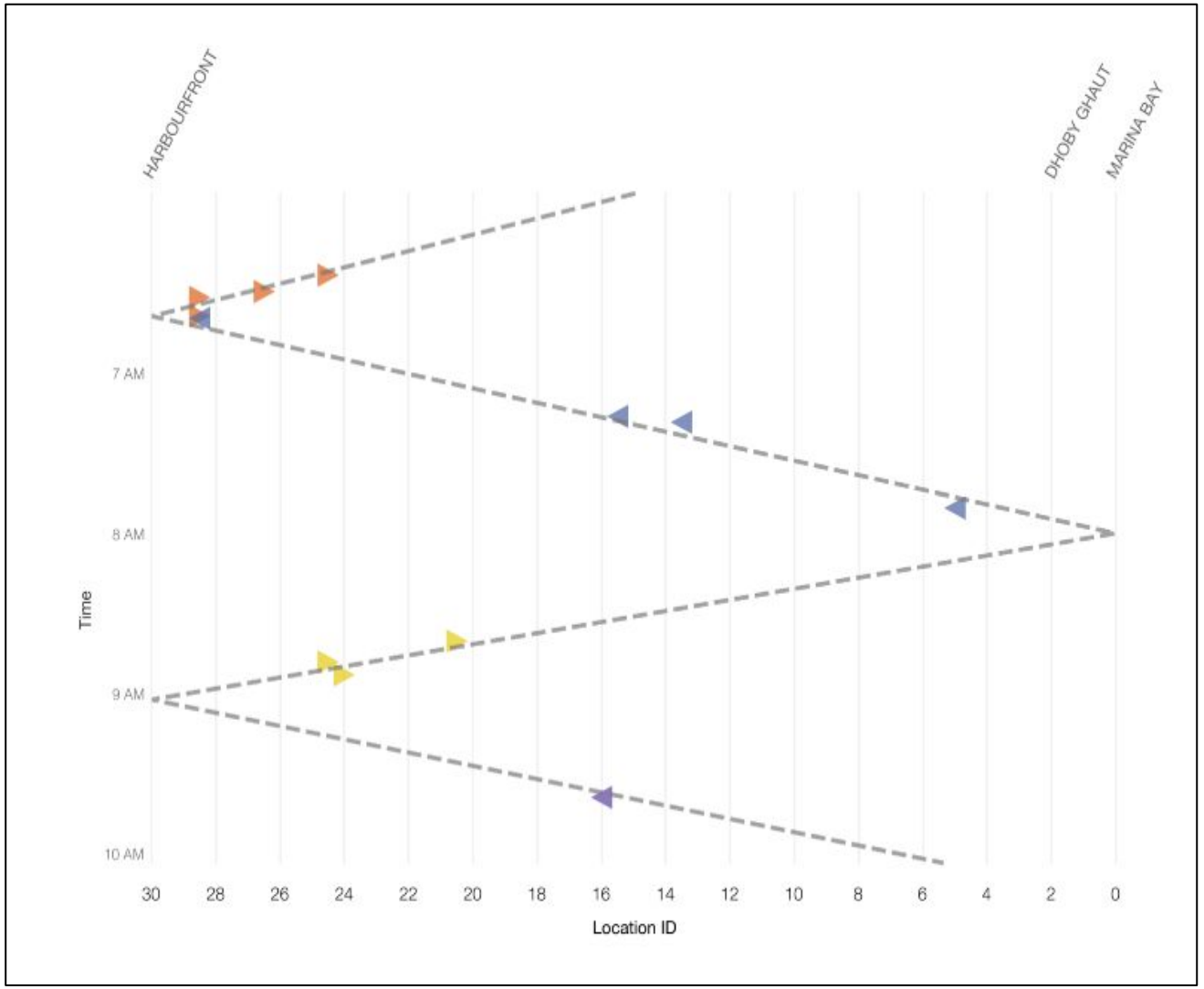
- Test the Rouge Train Hypothesis using clustering algorithm:
  - Group all emergency braking incidents together if they occur up to four minutes apart.
  - Grouped all related pairs of incidents into larger sets. This allowed us to group incidents that could be linked to the same “rogue train”
- **Of the 259 emergency braking incidents in our dataset, 189 cases—or 73% of them—could be explained by the “rogue train” hypothesis**.
- Coloured the incident chart based on the clustering results. Triangles with the same colour are in the same cluster.




Source: https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a

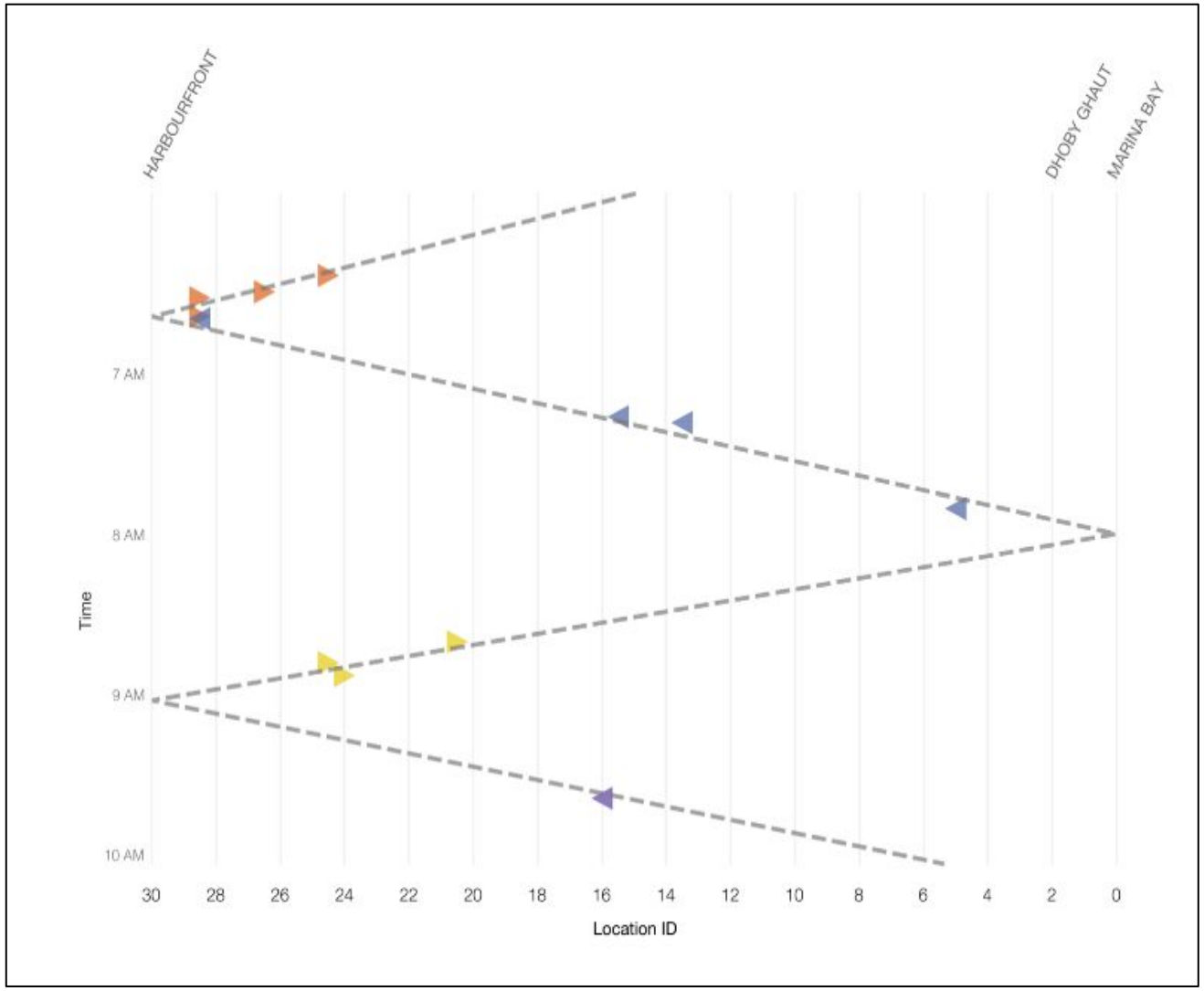
# CASE STUDY 3: CIRCLE LINE ROGUE TRAIN

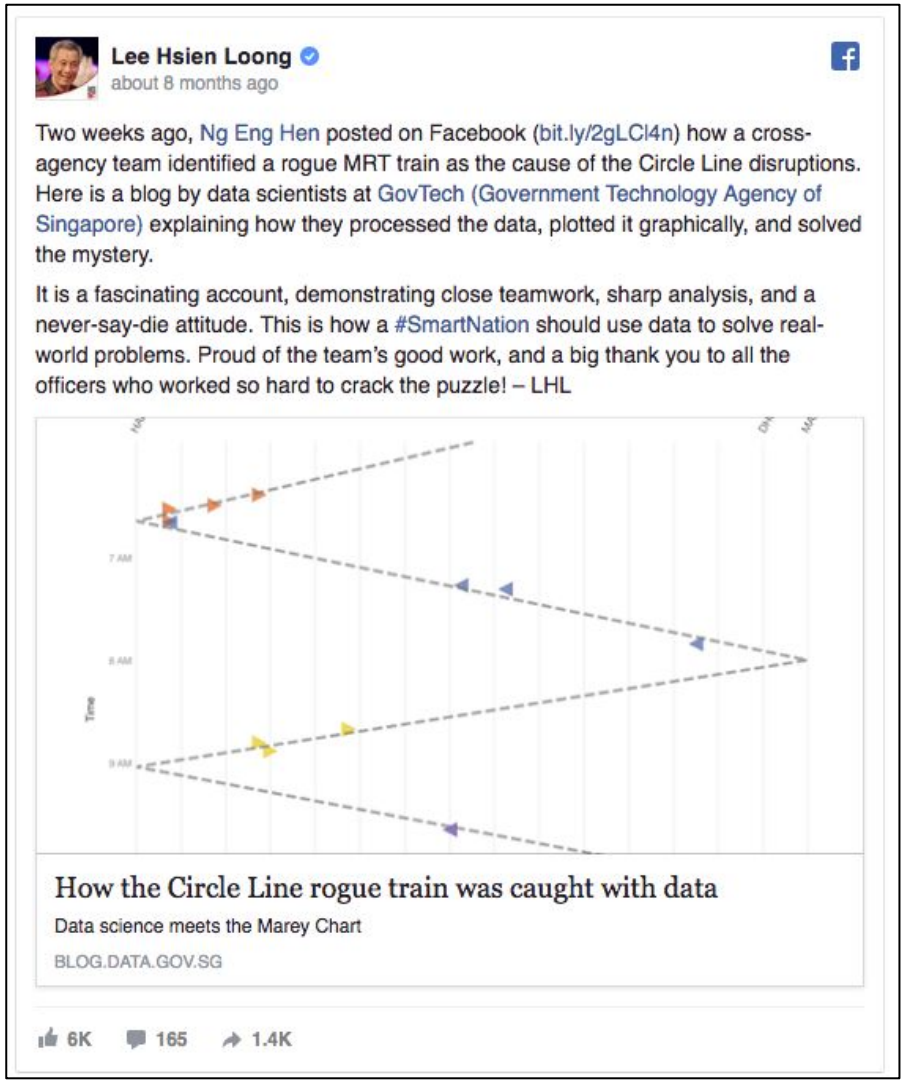
- Eventually they caught the rouge train PV46




Source: https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a

# CASE STUDY 3: CIRCLE LINE ROGUE TRAIN

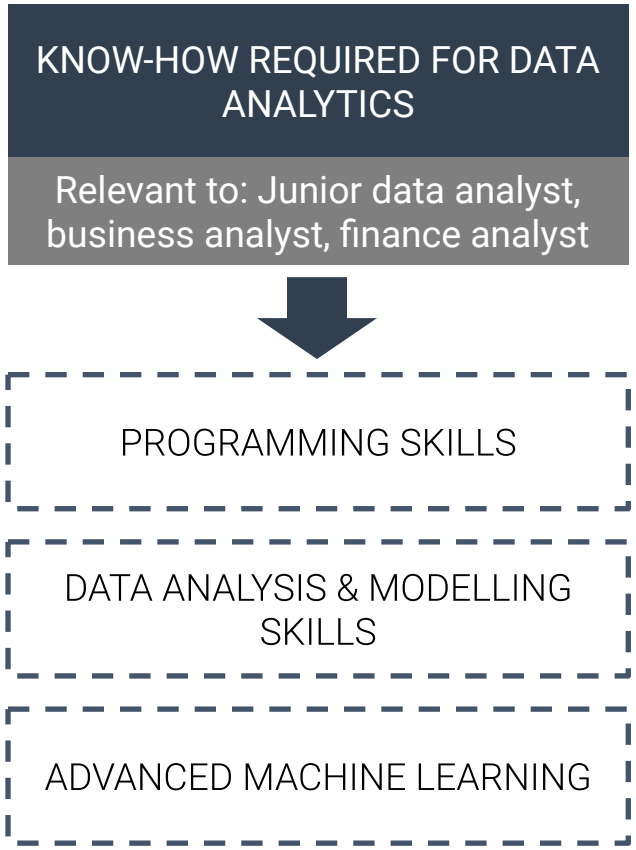



Source: https://blog.data.gov.sg/how-we-caught-the-circle-line-rogue-train-with-data-79405c86ab6a

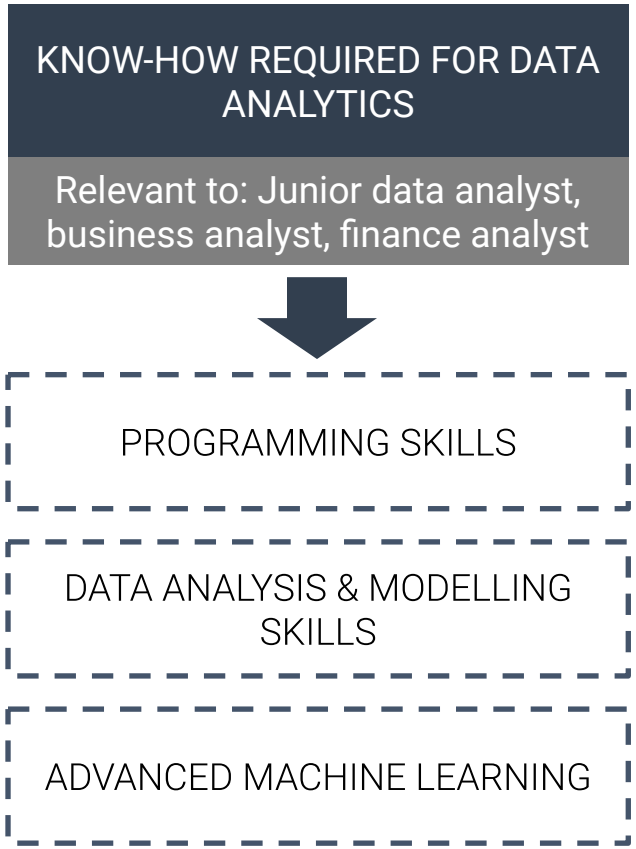
# HOW DOES DS101 TIE IN WITH ALL THIS?

- Today, to be able to properly apply data analytics in any business domain, there are three core skills that are required: programming skills, data analysis & modelling skills, and database management

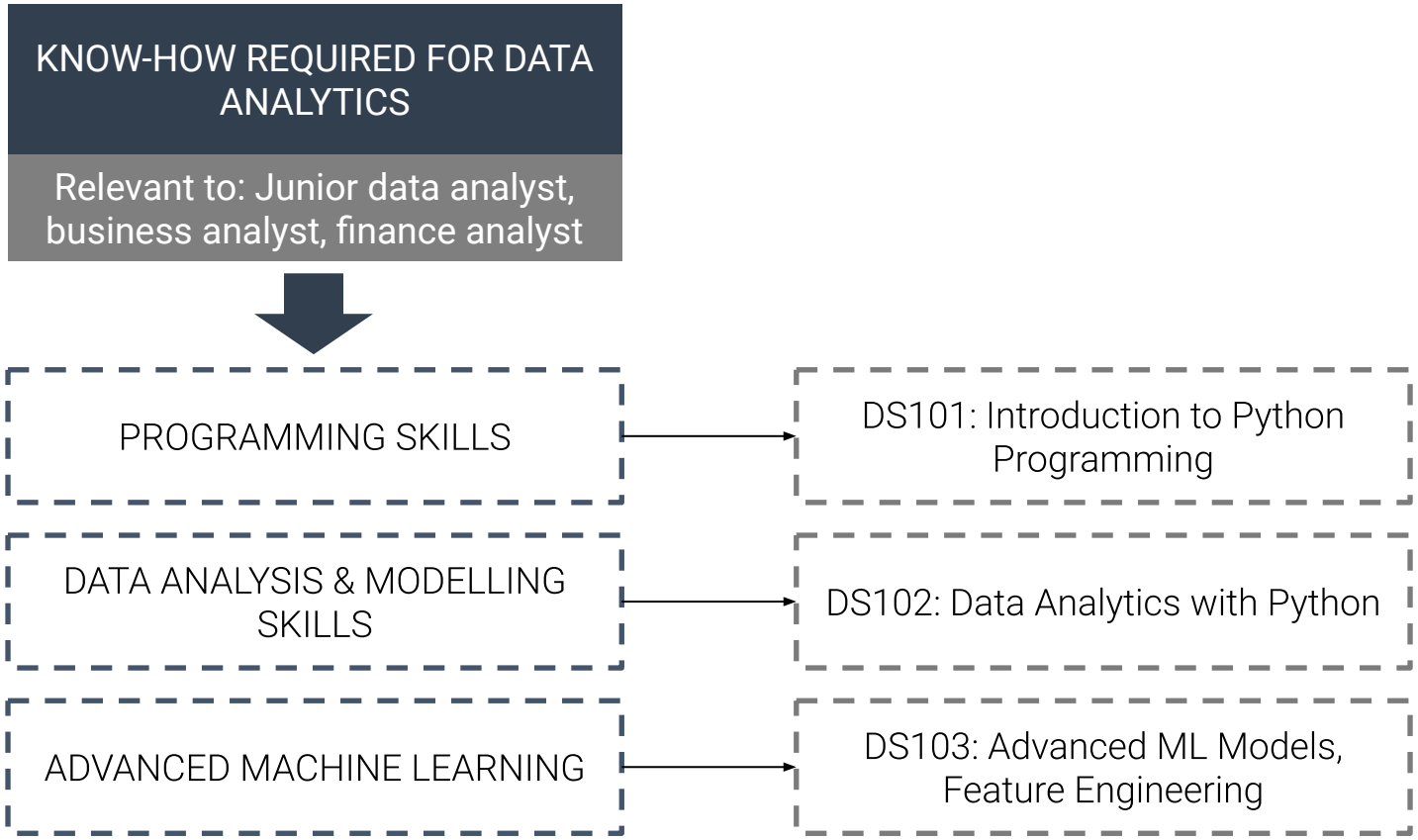


# HOW DOES DS101 TIE IN WITH ALL THIS?

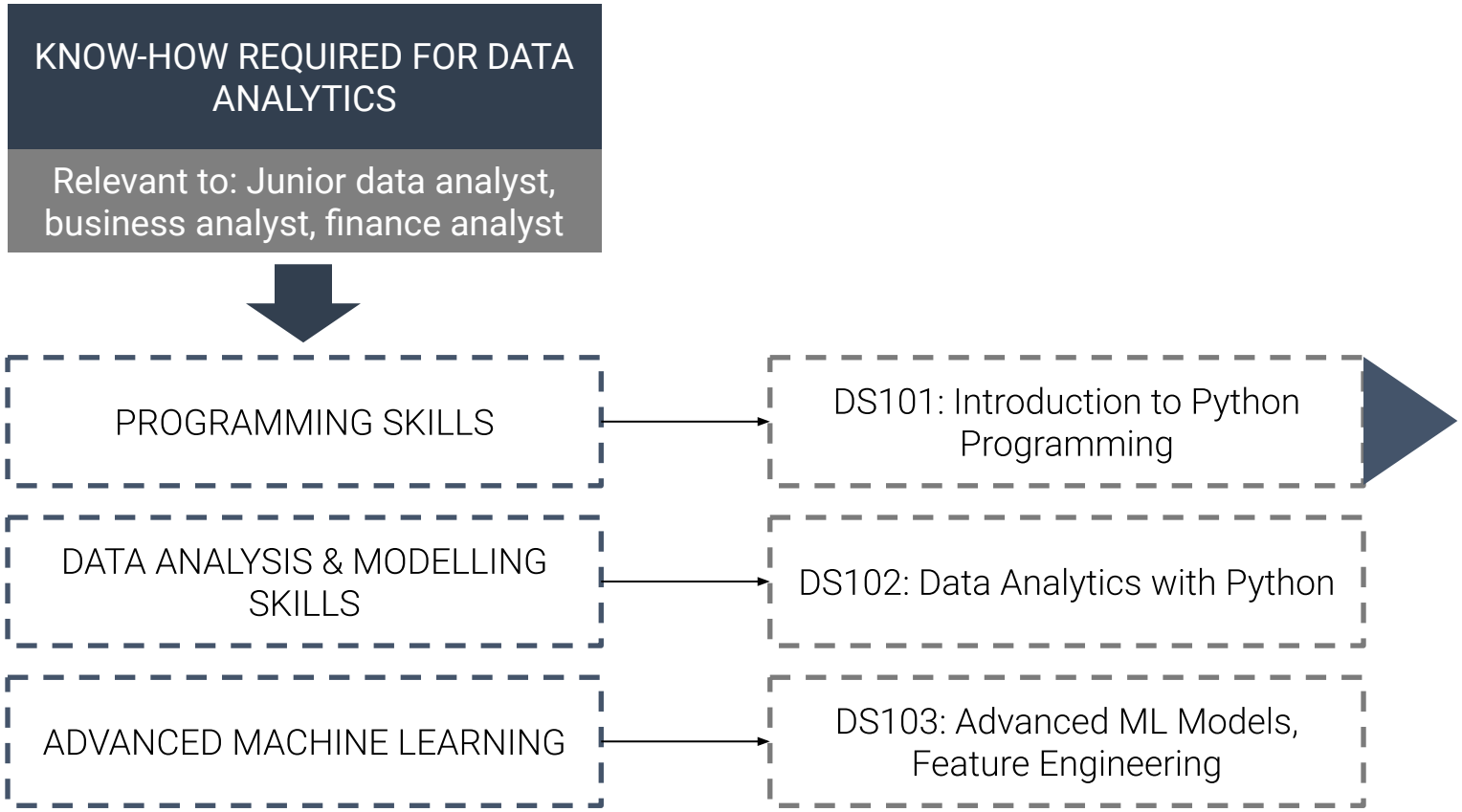
- Obviously, you don't have to acquire all three skills to start having an impact, rather the point is that the sum of these three skills is greater than the sum of its part



# HOW DOES DS101 TIE IN WITH ALL THIS?



# HOW DOES DS101 TIE IN WITH ALL THIS?




- Gain sufficient mastery of python to make use of advanced visualisation and modelling libraries in DS102

- Acquire the ability to read data from CSV into python, and perform:

  - Data cleaning
  - Tally and aggregate useful statistics
  - Visualise data to tell a story

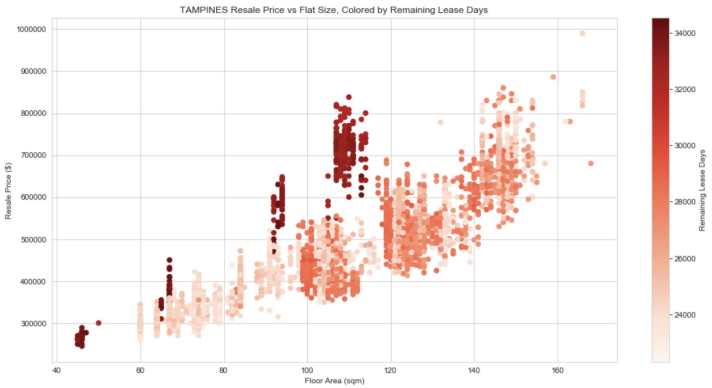
# WHAT WILL THIS COURSE COVER

## WEEK 0 – WEEK 5

- From the level 0 beginner programmer now, you will begin to acquire the syntactical and algorithmic aspect of programming

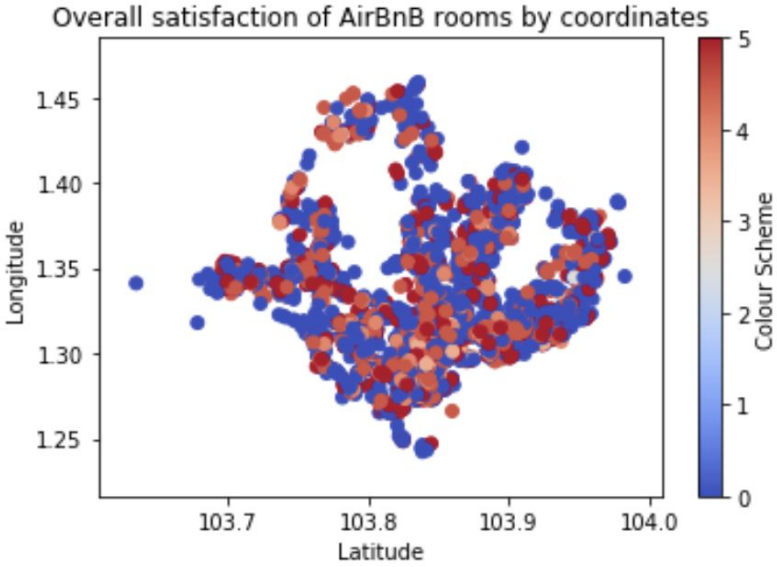
- Throughout this 5 weeks, you will learn how to use python to code various algorithms, starting with very basic algorithms such as how to calculate compound interest, all the way till where you write methods to perform data cleaning & complex aggregation

## WEEK 6: HDB CASE STUDY




You will go through an analytical study of HDB resale flat prices using knowledge learnt in the first 5 weeks - with the aim of reinforcing your understanding of data, data exploration, and descriptive analytics

## WEEK 7: PANDAS




You should reach to level where you can use Python libraries such as Pandas to understand how companies (like Airbnb and Lazada) make use of methods such as association-based recommendation to increase the probability of cross-selling to consumers!