### What is the use case?

1. **For NJ transit:** to predict the delay time and arrange the trains accordingly
2. **For city planners:** provide transportation needs and infrastructure information to better design the station and transportation routes
3. **For customers:** arrange schedules before 2 hours

### How could data make a difference in answering this question?

1. Data can detect the time pattern and spatial pattern and the line pattern about delay time, **for NJ transits**, they can use the pattern insights to arrange the trains in advance for more than a few days.
2. Data can know the attribution for each independent variable concerning the delay time, **for NJ transits**, they can use the attribution to accordingly improve the user experience.
3. Data can tell which station has periodic delay pattern, **for city planners**, they can use this insights to improve the station design.
4. **For the customers**, data can make a precise prediction for the delay before 2-3 hours of the depart.

### Do you have a sense for the business-as-usual decision making?

**The business-as-usual model is taking MTA (MTA New York City Transit) as an example.** [**LINK!**](https://ibo.nyc.ny.us/iboreports/we-are-being-held-momentarily-how-much-time-and-money-are-new-york-city-subway-riders-losing-to-delays-october-2017.html)

a. The MTA uses several indicators to measure the on-time performance of the subway system. Among these are the **subway wait assessment**, the **terminal on-time performance**, and the **number and cause of terminal delays**.

b. Weekday Wait Assessment. The MTA carries out its subway wait assessment by calculating the observed length of time between trains and comparing it with the scheduled interval. Service **meets the standard** if the actual headway is no greater than 25 percent over the scheduled headway. Service gaps are **classified as minor** (between 25 and 50 percent over the scheduled headway), **medium** (between 50 and 100 percent), and **major** (more than 100 percent over the scheduled headway).

c. The maximum headway on individual subway lines during the morning and evening rush is 10 minutes, in which case the gap is considered **minor** if the interval between trains is between 12.5 minutes and 15 minutes; **medium** if the gap is between 15 minutes and 20 minutes; and **major** if the gap exceeds 20 minutes.

d. Each train that arrives more than five minutes late to its terminal is assigned a reason for the delay. While individual trains may be delayed for multiple reasons, the MTA assigns just **one cause for each delay**. The category “overcrowding” is typically the greatest source of delays.

### What datasets have you identified to help you answer this question?

1. NJ Transit train-record dataset (From Kaggle)
2. NJ Station (From NJ opendata platform to get the geometry distribution)
3. OSM Landmark Distance AND census tract data (to get the variables about "overcrowding mentioned above")

### What kind of model would you build and what is the dependent variable?

1. Linear regression model with temporal, spatial independent variables
2. Passion regression model to compare with linear regression model
3. Dependent variable: Delay minutes for every station in each train ID

### How will you validate this model ?

1. **Cross-validation** will be used to measure the goodness of fit.
2. **Business Performance**

Delay Repay is a national scheme that train companies use to compensate you for unexpected delays and cancellations. If you arrive 15 minutes or more late at your destination because of a delay or cancellation to a Southern service, you can claim Delay Repay compensation. Delay Repay is a national scheme train companies use to compensate passengers for delays or cancellations.

### How do you think that stakeholders would want to consume this data?

1. **For NJ transit:** to predict the delay time and arrange the trains accordingly, decrease the business loss. (Increase the satisfices and decrease the refund rate)
2. **For city planners:** provide transportation needs and infrastructure information to better design the station and transportation routes
3. **For customers:** arrange schedules before 2 hours

### What are the use cases for your app and what should the app do?

1. **App for NJ Transits**

Quantitively measure the delay for different lines and situations

Quantitively measure the effect of the rearrangement

Arrange the schedule more accurately (like increase or decrease the operation time during certain interval)

1. **App for consumers**

Modify the schedule in 2 hours before

Optimize the transportation plan (like take which train at when to where)