Aim

1. Data split into train and test

In [16]:

```
import os
from datetime import datetime

import numpy as np
import pandas as pd
```

In [2]:

```
DATA_ROOT = f"../data"
```

In [3]:

```
df = pd.read_pickle(f"{DATA_ROOT}/fselect/accidents_raw.pkl")
```

In [7]:

```
df.head(3)
```

Out[7]:

	ID	Source	TMC	Start_Time	Distance(mi)	Description	Side	City	Cou
0	A-1	MapQuest	201.0	2016-02-08 05:46:00	0.01	Right lane blocked due to accident on I-70 Eas...	R	Dayton	Montgom
1	A-2	MapQuest	201.0	2016-02-08 06:07:59	0.01	Accident on Brice Rd at Tussing Rd. Expect del...	L	Reynoldsburg	Fran
2	A-3	MapQuest	201.0	2016-02-08 06:49:27	0.01	Accident on OH-32 State Route 32 Westbound at ...	R	Williamsburg	Clerm

3 rows × 38 columns

In [13]:

```
# convert time to datetime
df["Start_Time"] = pd.to_datetime(df["Start_Time"])

# The task is to predict the impact of accident on traffic from January 202
df = df.sort_values(by=["Start_Time"], ascending=True, ignore_index=True)
```

In [27]:

```
test_data_start_date = pd.to_datetime("2020-01-01")
test_data_end_date = pd.to_datetime("2020-07-01")
print(test_data_end_date)
```

```
2020-07-01 00:00:00
```

In [26]:

```
df
```

Out[26]:

	ID	Source	TMC	Start_Time	Distance(mi)	Description	Side	
0	A-2478859	Bing	0.0	2016-02-08 00:37:08	3.230	Between Sawmill Rd/Exit 20 and OH-315/Olentang...	R	
1	A-1	MapQuest	201.0	2016-02-08 05:46:00	0.010	Right lane blocked due to accident on I-70 Eas...	R	D
2	A-2478860	Bing	0.0	2016-02-08 05:56:20	0.747	At OH-4/OH-235/Exit 41 - Accident.	R	D
3	A-2	MapQuest	201.0	2016-02-08 06:07:59	0.010	Accident on Brice Rd at Tussing Rd. Expect del...	L	Reynolc
4	A-2478861	Bing	0.0	2016-02-08 06:15:39	0.055	At I-71/US-50/Exit 1 - Accident.	R	Cin
...	...	...	...	...	...	...	...	
3513612	A-560284	MapQuest	201.0	2020-06-30 22:51:25	0.000	Right lane blocked due to accident on I-195 We...	R	Provi
3513613	A-561183	MapQuest	241.0	2020-06-30 22:52:02	0.000	Lane blocked due to accident on Dallas North T...	R	
3513614	A-561279	MapQuest	201.0	2020-06-30 22:52:37	0.000	Accident on US-31A Nolensville Pike at Allied Dr.	R	Na
3513615	A-561184	MapQuest	241.0	2020-06-30 22:56:52	0.000	Lane blocked due to accident on I-30 Westbound...	R	
3513616	A-560480	MapQuest	201.0	2020-06-30 23:18:09	0.000	Accident on I-264 Watterson Expy Westbound at ...	R	Lou

3513617 rows × 38 columns

In [38]:

```
df_test = df[
    (df["Start_Time"] >= test_data_start_date) & (df["Start_Time"] < test_da
]

df_test.shape
```

Out[38]:

```
(539187, 38)
```

In [39]:

```
df_train = df.iloc[~df.index.isin(df_test.index)]

df_train.shape
```

Out[39]:

```
(2974430, 38)
```

In [40]:

```
# check if train test split has any intersections

set(df_train["ID"]).intersection(set(df_test["ID"]))
```

Out[40]:

```
set()
```

In [43]:

```
os.makedirs(f"{DATA_ROOT}/train/raw/", exist_ok=True)
os.makedirs(f"{DATA_ROOT}/test/raw/", exist_ok=True)
```

In [45]:

```
df_train.reset_index(drop=True).to_pickle(f"{DATA_ROOT}/train/raw/data.pkl")
df_test.reset_index(drop=True).to_pickle(f"{DATA_ROOT}/test/raw/data.pkl")
```