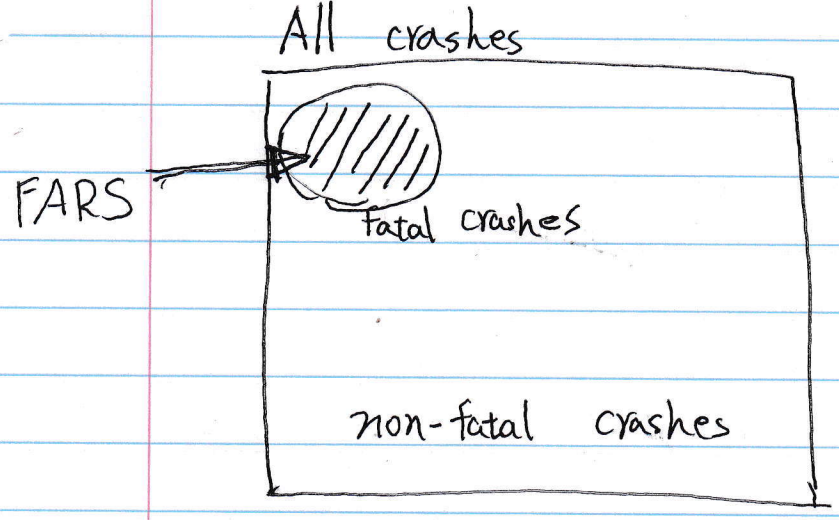
I.



rate of fatality $= \frac{\text{\# of fatalities}}{\text{\# All Crashes}}$, but # of non-fatal crashes is unknown (DOT does not publish non-fatal crash data)

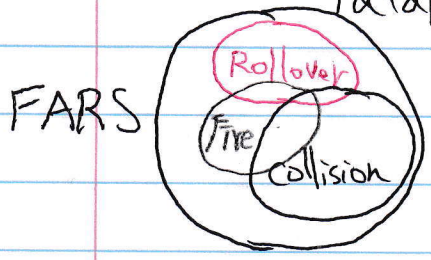
→ The rate of fatality cannot be estimated

Ex:

In a fatal crash, the only occupant is a driver, what is the driver's fatality rate?

→ 100 % (meaningless)

II.



We can select a specific type of fatal crashes and see its proportion in all fatal crashes

① $P(\text{rollover in fatal crashes}) = \frac{\text{\# rollovers}}{\text{\# fatal crashes.}}$

② Given a fatal crash and other information (vehicle body type, model year, driver's gender, road way condition) what is the prob that it's a rollover

III

our data set :

we only consider single-vehicle crashes, and the range of years is from 2014 to 2018 because :

① multi-vehicle crash is hard to analyzed
→ For example : In a two-vehicle fatal crash which driver's age ; which vehicle's model year. should we consider?

② We take 5-year range, since rollover events are rare.

data set : single-vehicle crashes in FARS 2014 – 2018.

IV

① data manipulation : 1. merge person, vehicle, accident files
2. filter data.
3. data cleaning

② check the association between rollover and other variables.

ex:	Rollover	non-rollover
male driver	n_1	n_2
Female driver	n_3	n_4

→ ① We can use plots to present such information
② statistical test :
→ H_0 : driver's gender is associated with rollover
H_a : driver's gender is not associated with rollover
→ use chi-square test.

③ modeling : 1. KNN model
2. logistic regression.
Bring the significant variables in ② into the model

④ model comparison ⑤ conclusion