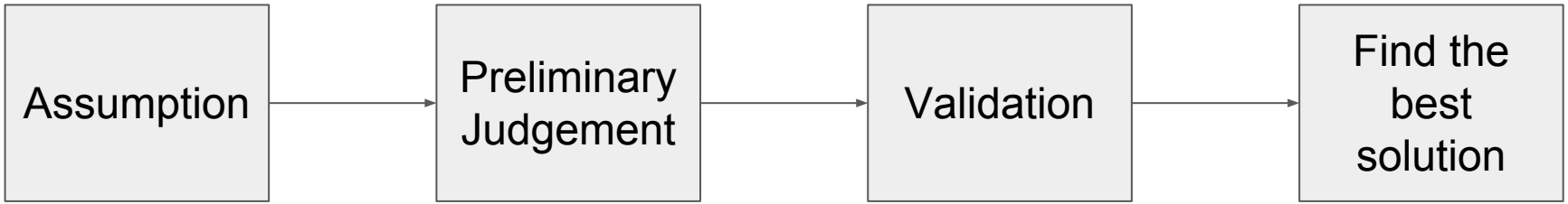
Which model is better?

Prediction procedure



Prediction problem

We want to know the pollen amount in the air but we do not have a pollen counter device.

Problem: find the way to get the pollen amount without counter device. → prediction

1. Assumption

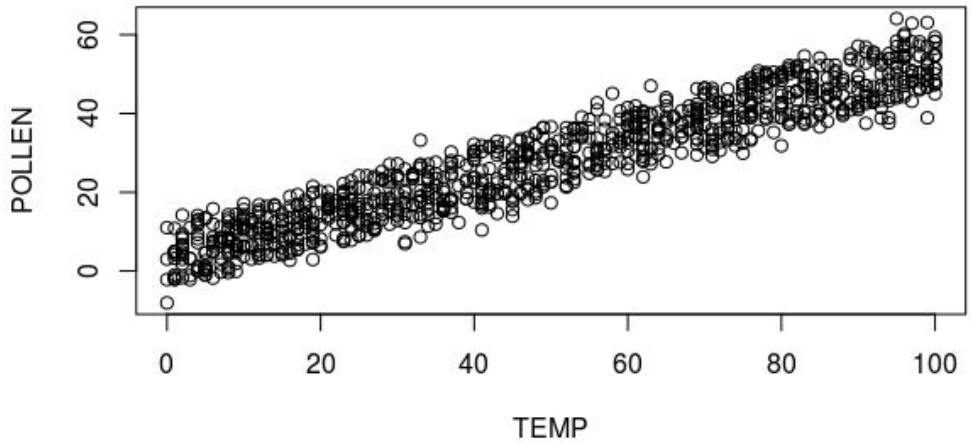
In allergy data set, pollen level amount can be affected by other conditions.

guess: 1. Higher humidity → less pollen in the air
2. Higher temperature → more pollen in the air

There should be relation between humidity and pollen, temperature and pollen.

2. Preliminary judgement -- plot

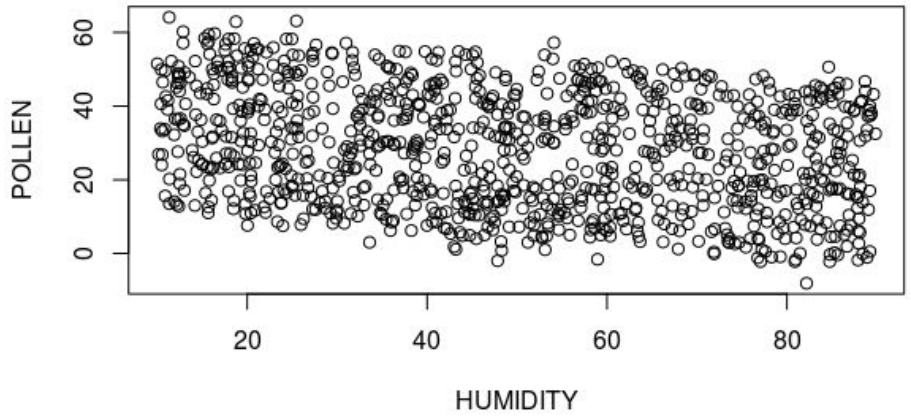
Higher temperature → more pollen?



positive correlation

2. Preliminary judgement -- plot

Higher humidity → less pollen?



negative correlation

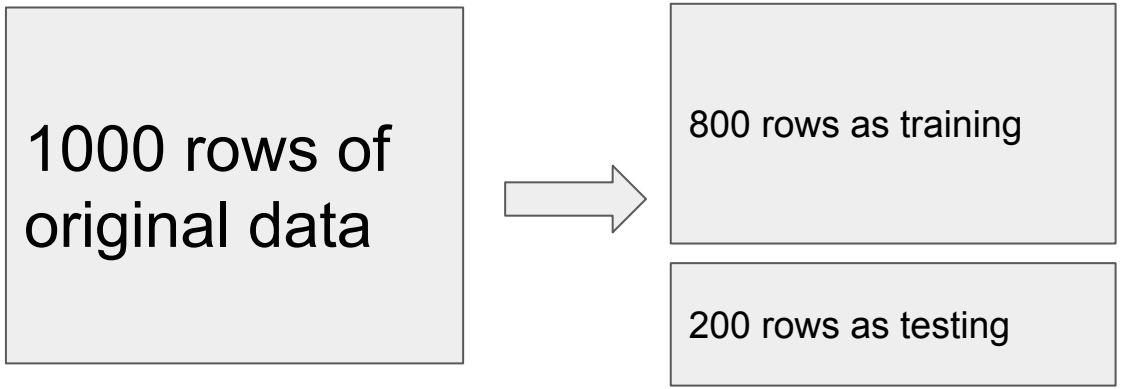
3. Validation

1. Decision tree? Or Linear model? → which one is better?
2. How to find better model? → error rate(or accuracy)
3. How to calculate the prediction error? → compare truthful data and prediction result
4. How to get truthful data? → separate from existing data set

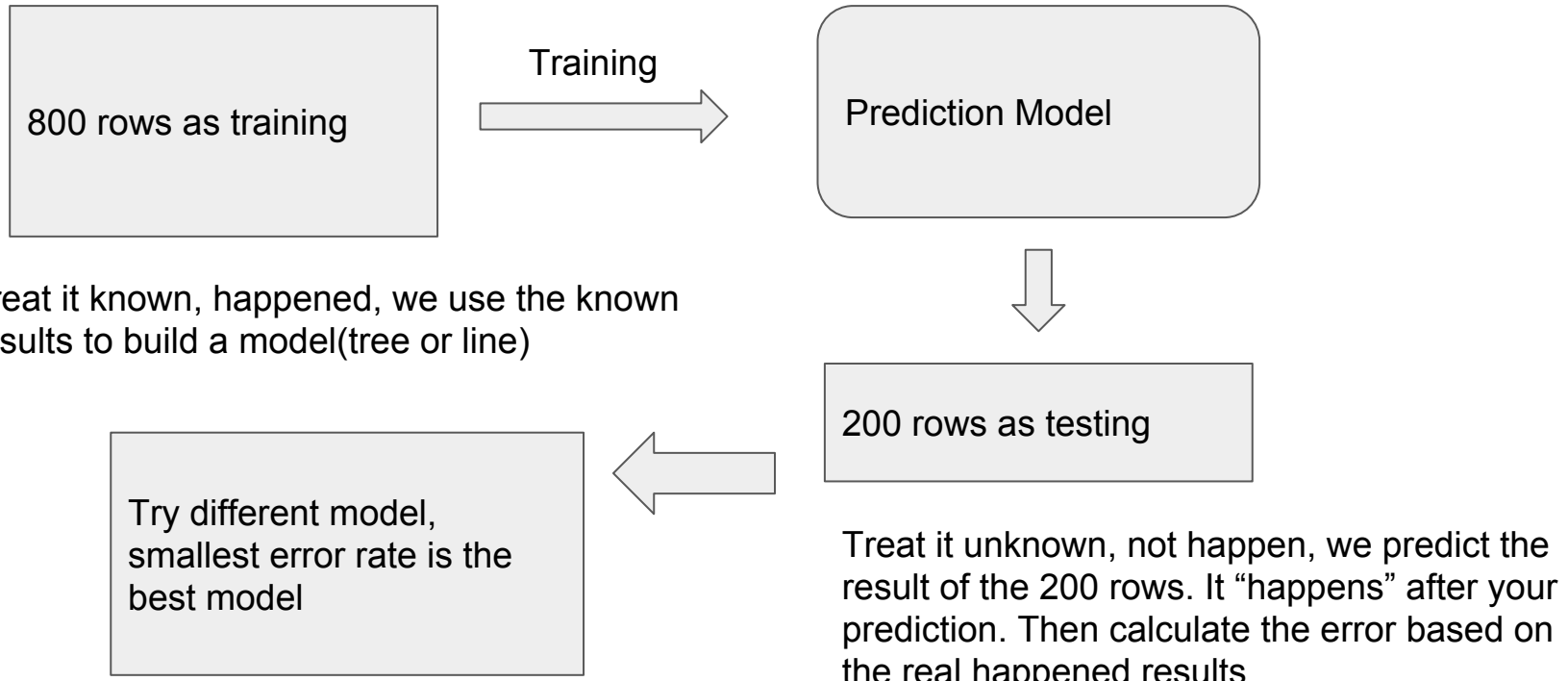
4. Validation

Allergy data set

1000 rows, divide it to two parts. 800 rows and 200 rows.



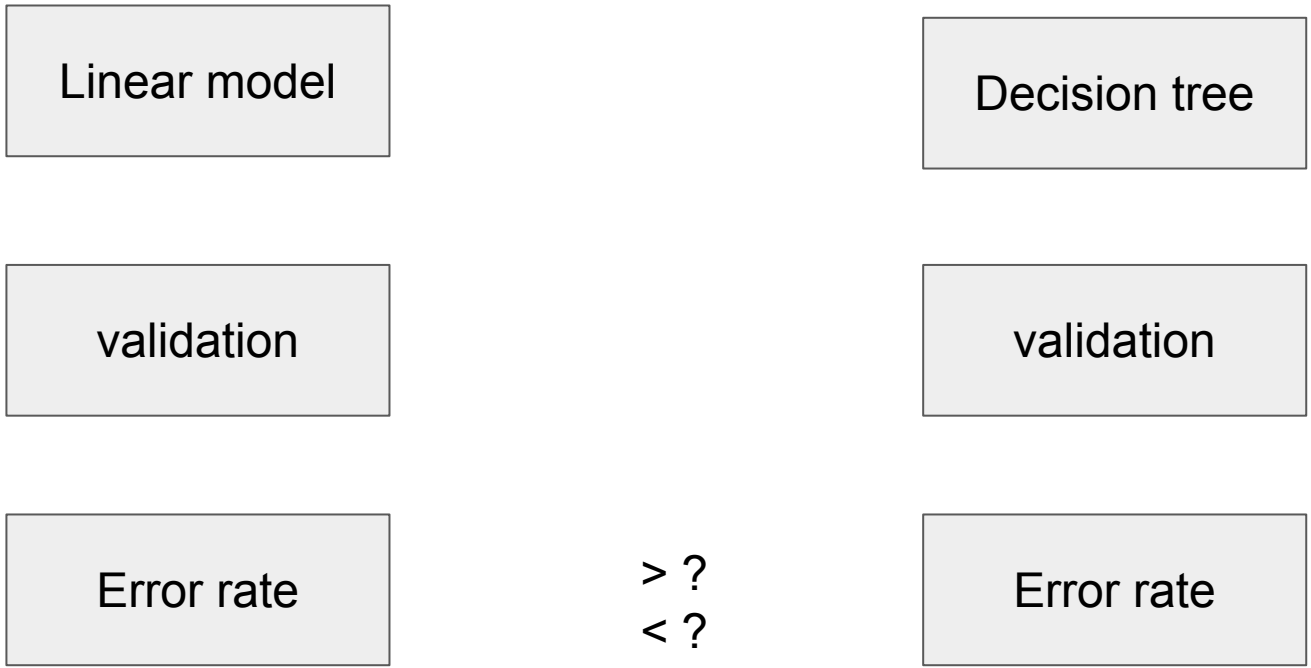
4. Validation



Treat it known, happened, we use the known results to build a model(tree or line)

Treat it unknown, not happen, we predict the result of the 200 rows. It “happens” after your prediction. Then calculate the error based on the real happened results

4. Validation



5. Error for numerical data- MSE

Truth in the separated testing	32 32.4 43.4 23.3
Prediction result	31 32.2 42.1 25

$$error = (32 - 31)^2 + (32.4 - 32.2)^2 + (43.4 - 42.1)^2 + (23.3 - 25)^2$$

6. Cross Validation