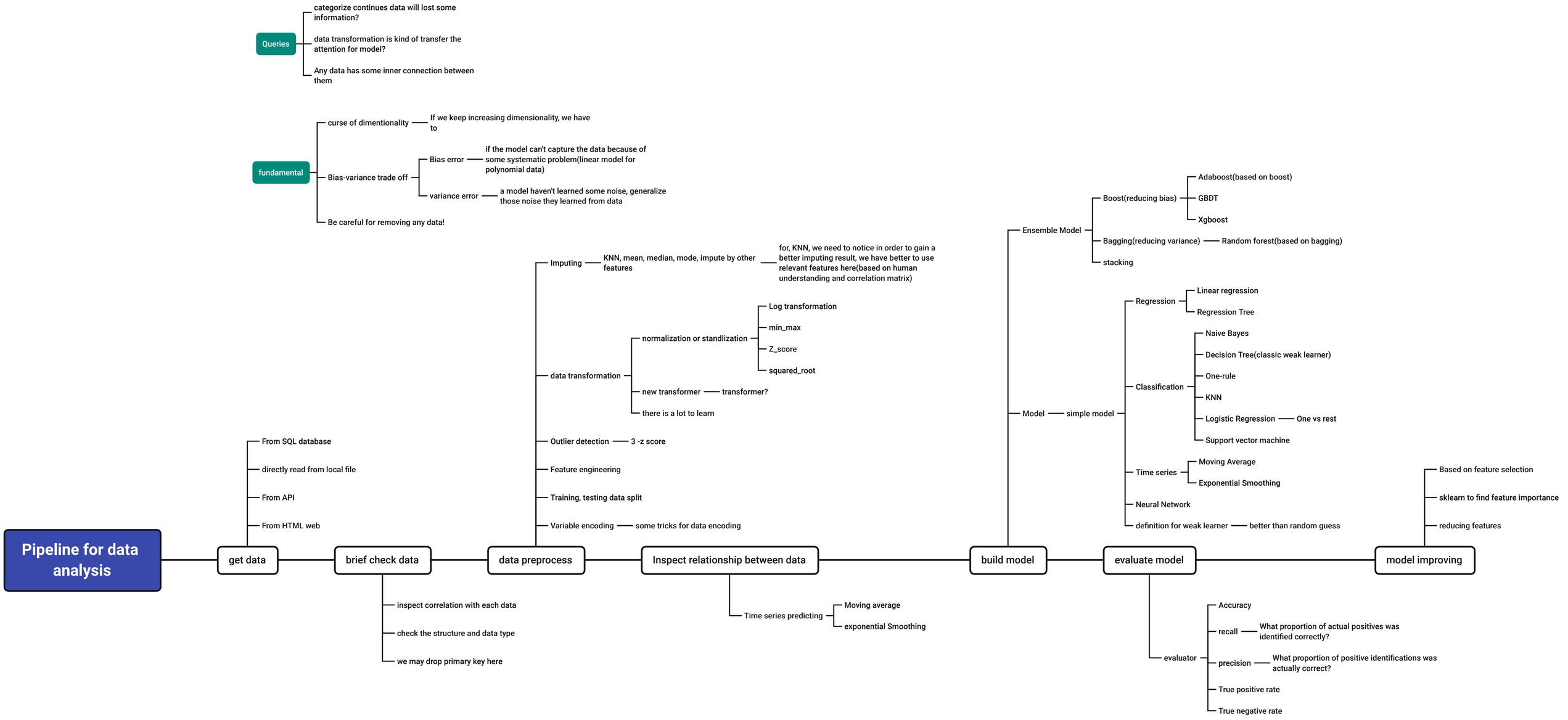
Queries
categorize continues data will lost some information?
data transformation is kind of transfer the attention for model?
Any data has some inner connection between them
fundamental
curse of dimentionality
If we keep increasing dimensionality, we have to
Bias-variance trade off
Bias error
if the model can't capture the data because of some systematic problem(linear model for polynomial data)
variance error
a model haven't learned some noise, generalize those noise they learned from data
Be careful for removing any data!
Pipeline for data analysis
get data
From SQL database
directly read from local file
From API
From HTML web
brief check data
inspect correlation with each data
check the structure and data type
we may drop primary key here
data preprocess
Imputing
KNN, mean, median, mode, impute by other features
for, KNN, we need to notice in order to gain a better imputing result, we have better to use relevant features here(based on human understanding and correlation matrix)
data transformation
normalization or standlization
Log transformation
min_max
Z_score
squared_root
new transformer
transformer?
there is a lot to learn
Outlier detection
3 -z score
Feature engineering
Training, testing data split
Variable encoding
some tricks for data encoding
Inspect relationship between data
Time series predicting
Moving average
exponential Smoothing
build model
Ensemble Model
Boost(reducing bias)
Adaboost(based on boost)
GBDT
Xgboost
Bagging(reducing variance)
Random forest(based on bagging)
stacking
Model
simple model
Regression
Linear regression
Regression Tree
Classification
Naive Bayes
Decision Tree(classic weak learner)
One-rule
KNN
Logistic Regression
One vs rest
Support vector machine
Time series
Moving Average
Exponential Smoothing
Neural Network
definition for weak learner
better than random guess
evaluate model
evaluator
Accuracy
recall
What proportion of actual positives was identified correctly?
precision
What proportion of positive identifications was actually correct?
True positive rate
True negative rate
model improving
Based on feature selection
sklearn to find feature importance
reducing features