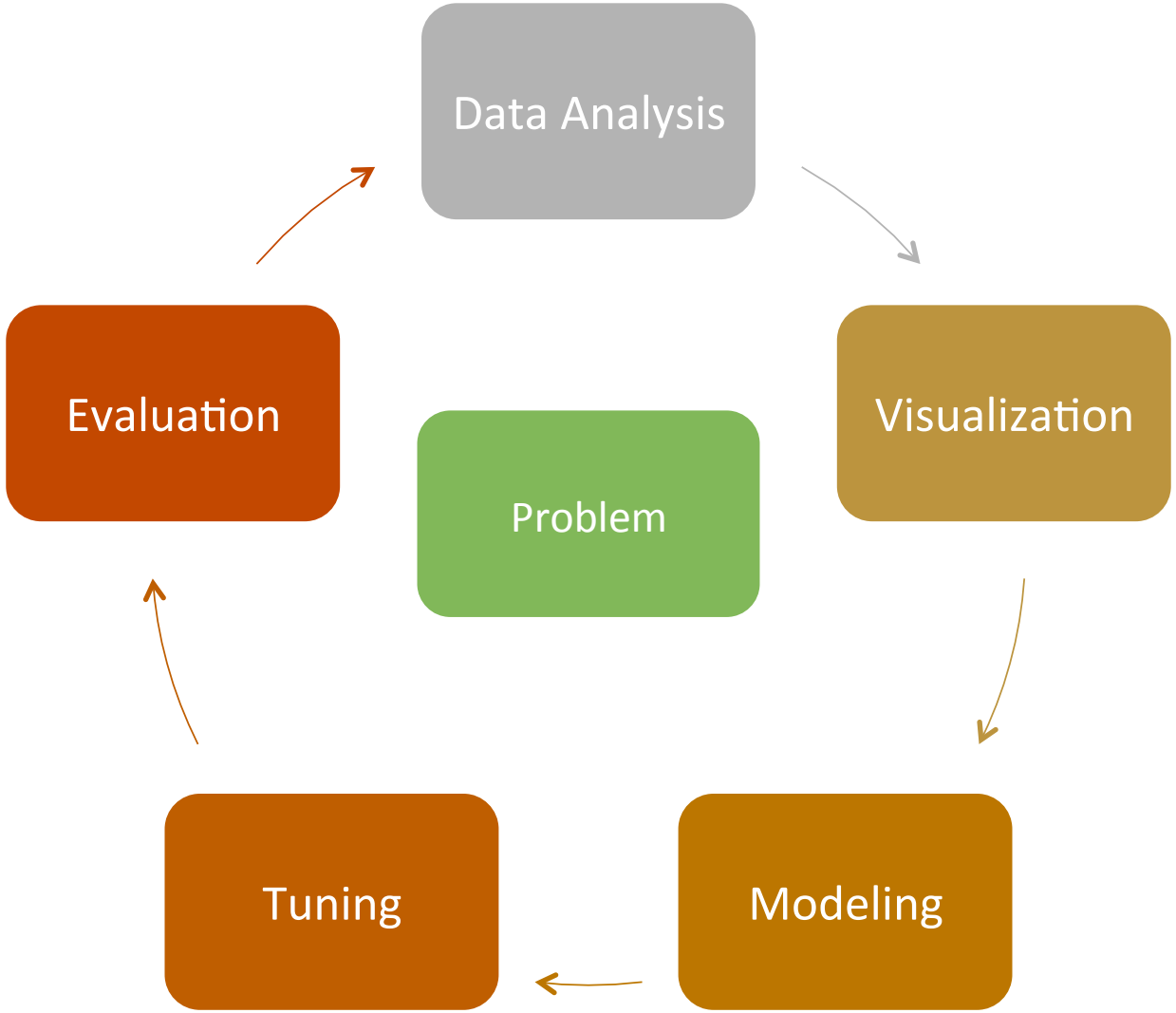
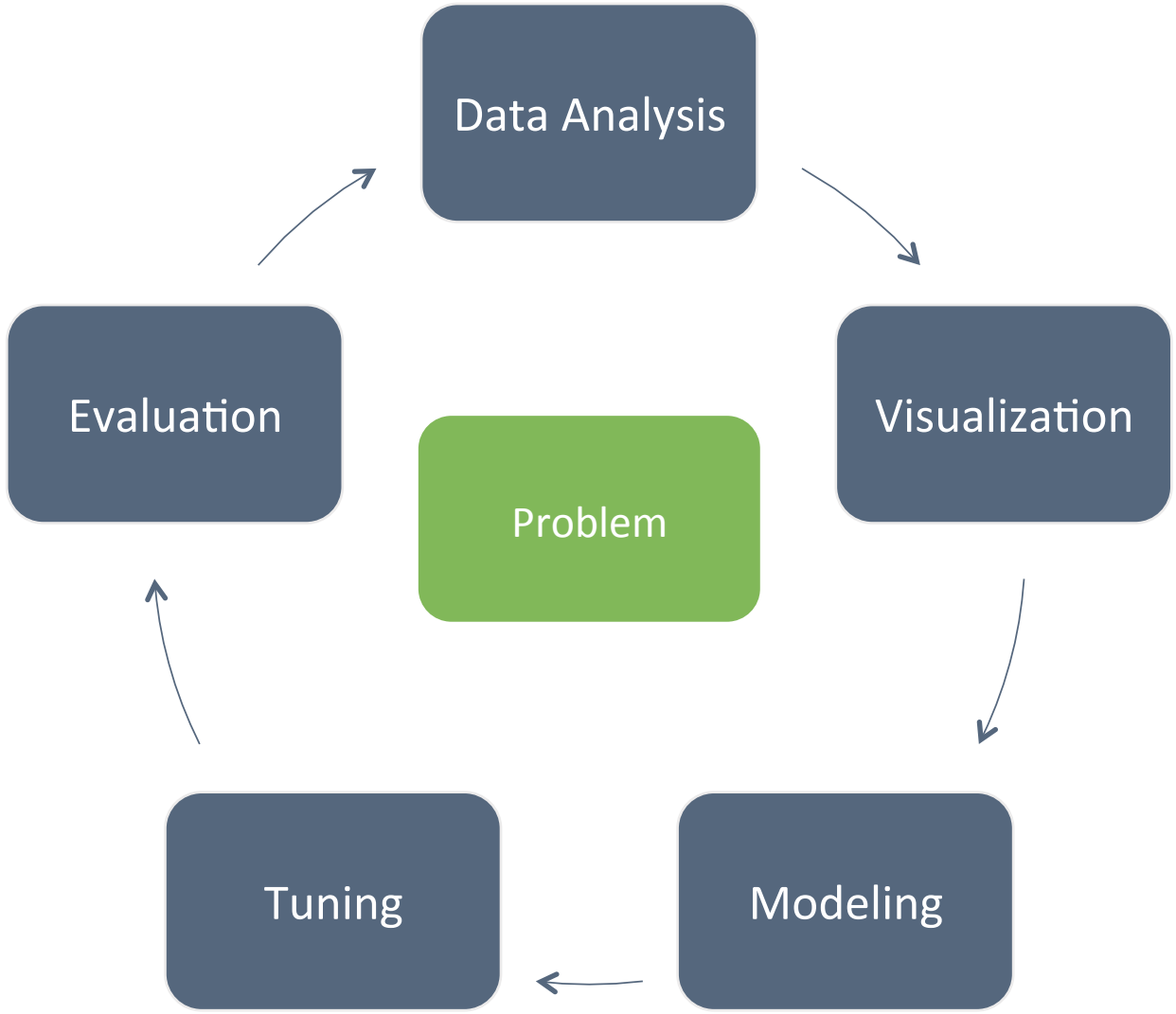
Data analysis lesson/experience



Data analysis lesson/experience

Problem

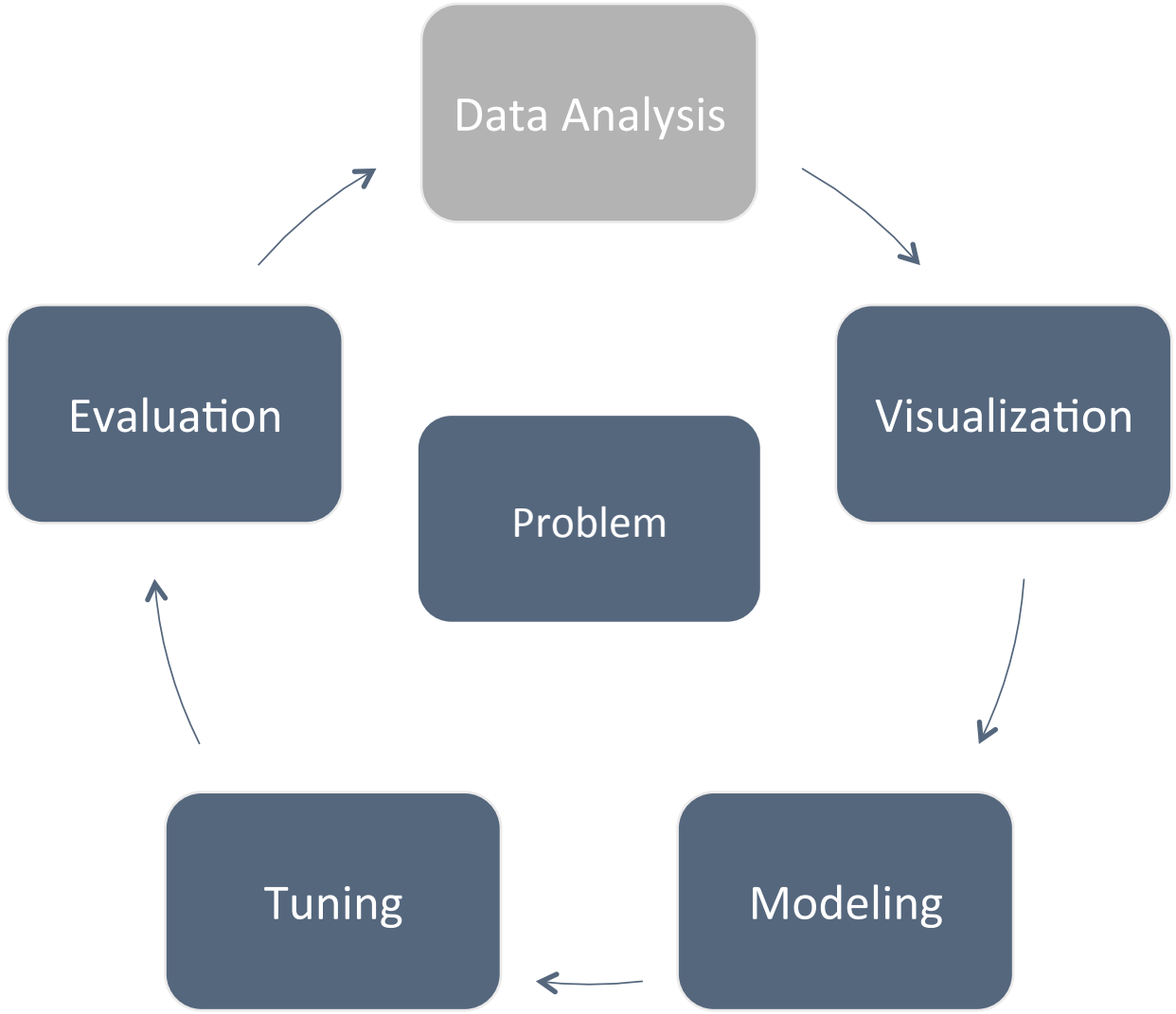
- Study the problem carefully.
- Construct the scenery.
- Define the goal.
- Figure out the challenge and contribution.
- Consider the data can be utilized to solve the problem or not.



Data analysis lesson/experience

Data Analysis

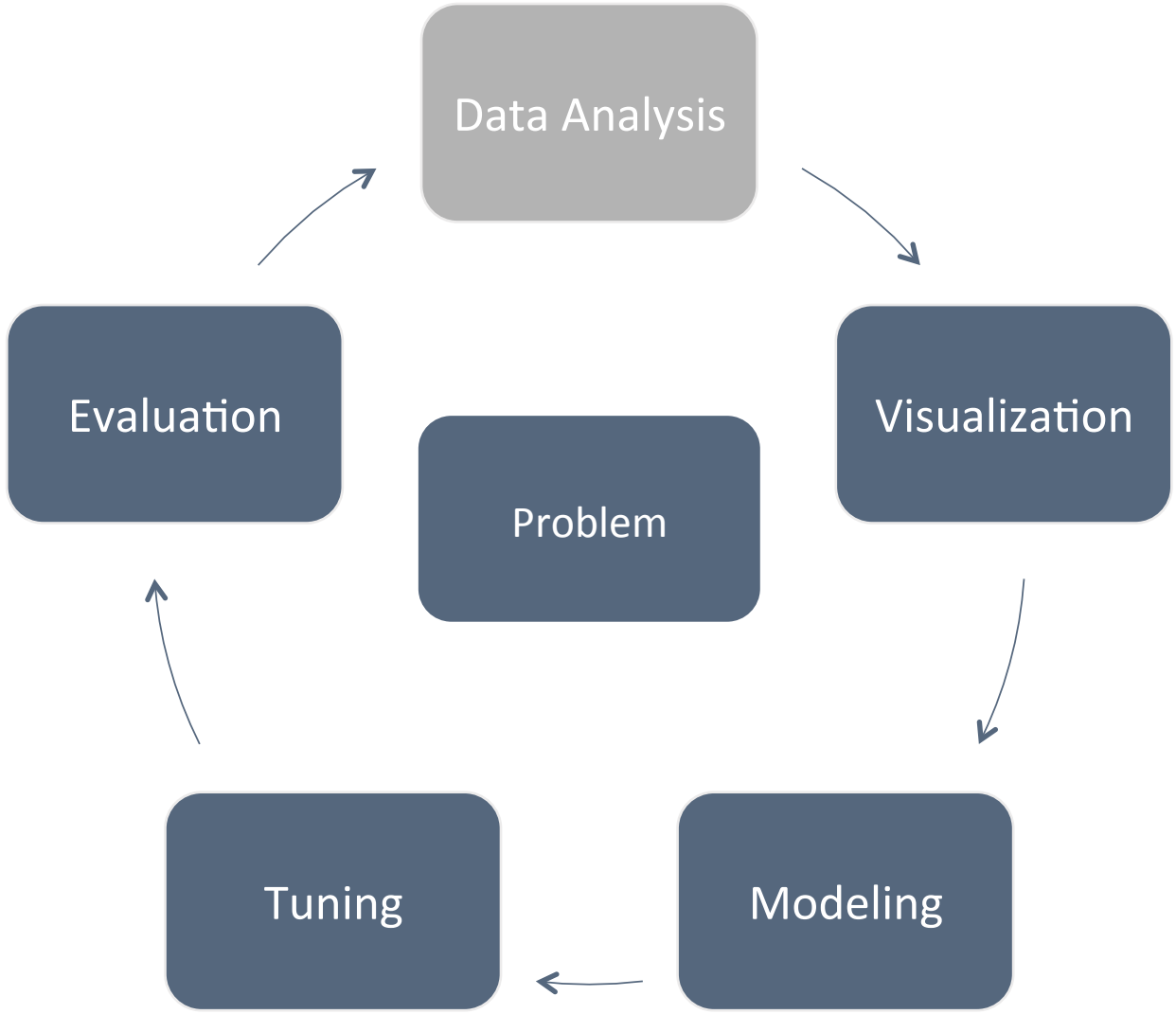
- Knowing the data, involving the range, type, std., mean, maximum, minimum mutual information and so on…
- Scale to [0 1] but Information loss problem!
- What is the data about?
- How was the data collected?
- Who is the data collector?
- What is the way to complain about the data?
- Who should be responsible for the data?



Data analysis lesson/experience

Data Analysis

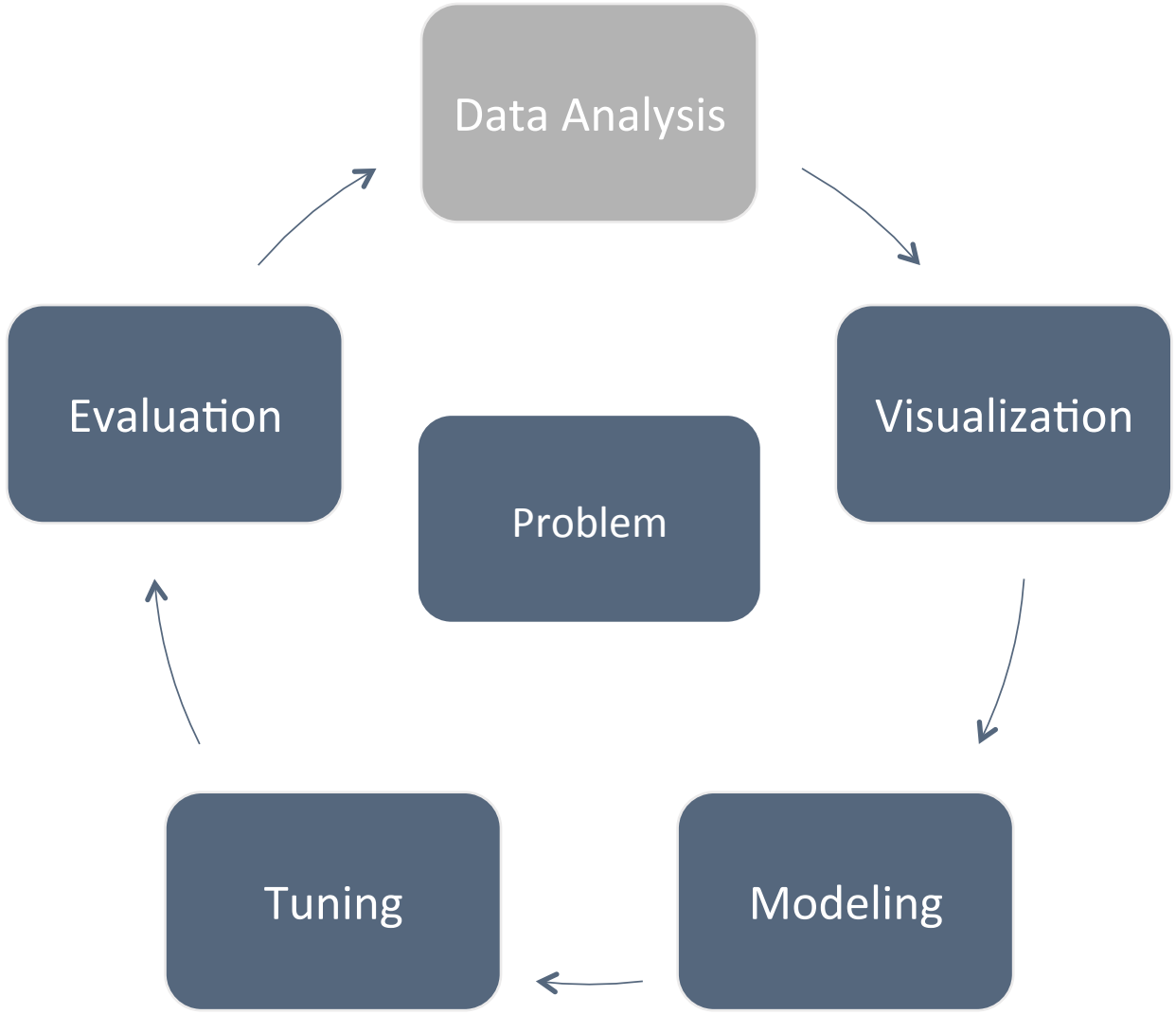
- Understanding the data.
  - Type of each feature.
  - Numerical or categorical?
  - If your method is numerical data only, use m numbers to represent an m-category attribute while preprocessing. Domain of each feature.
  - Distribution of each feature.



Data analysis lesson/experience

Data Analysis

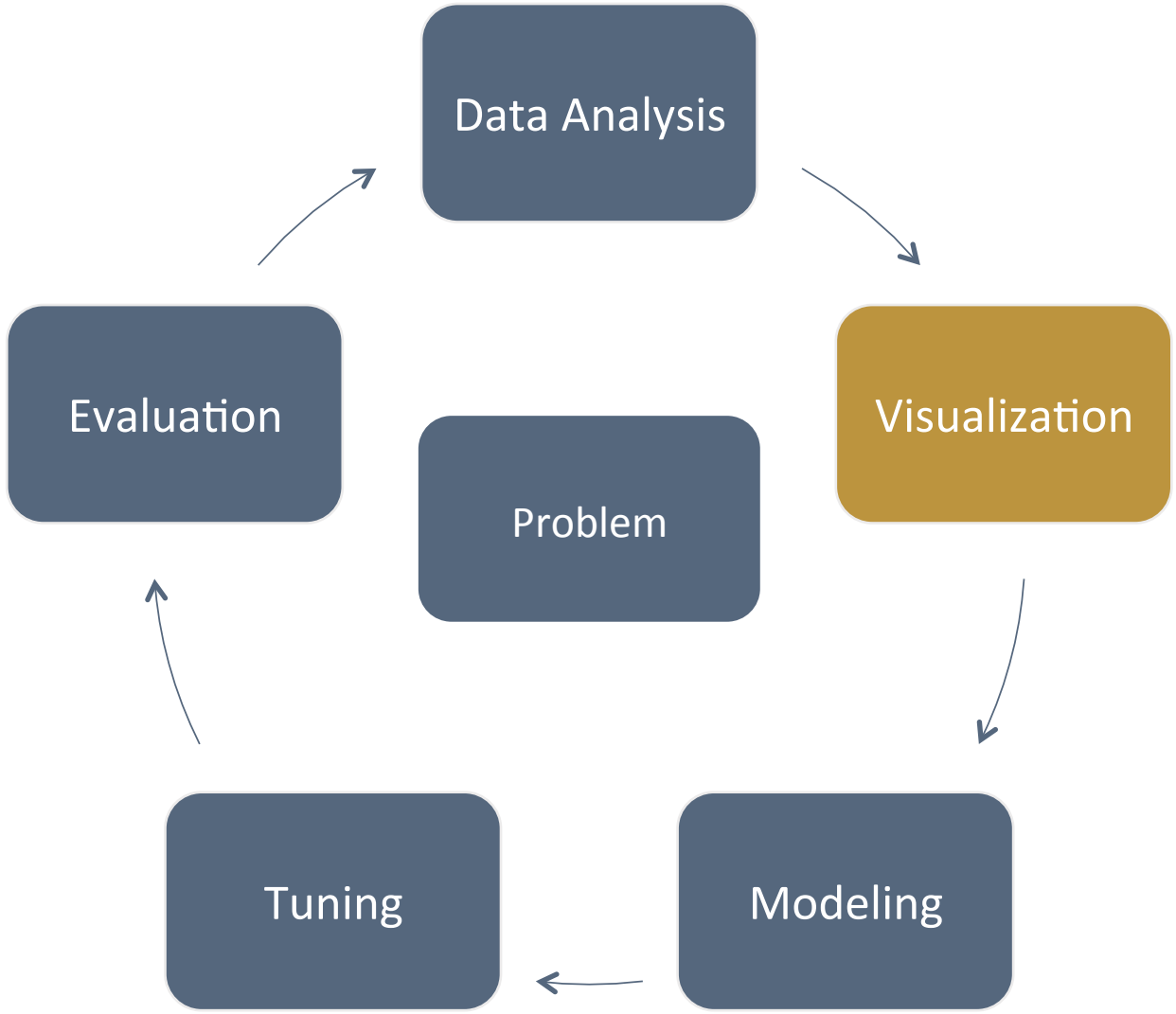
- Understanding the data.
  - Format of the data.
  - Encoding of the data.
  - Newline of the data.
  - Missing value of the data.
  - Type of the data.
  - Anything wired of the data.



Data analysis lesson/experience

Visualization

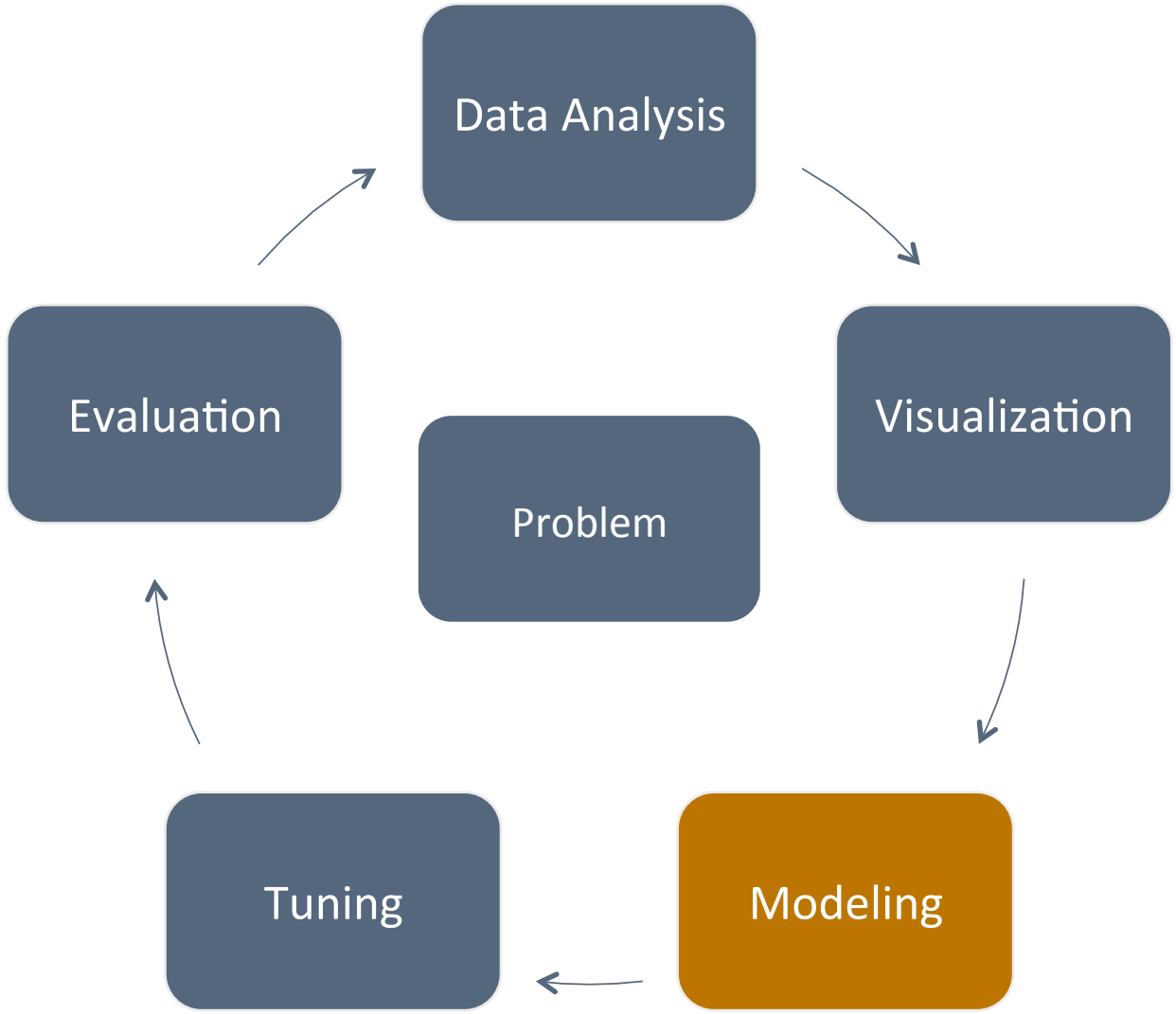
- Try to visualize the data in three dimensional space
- Try linear dimension reduction first.
- Try to visualize the data in three dimensional space
  - Use image to make more sense about the data.
  - Prepare beautiful image for report.



Data analysis lesson/experience

Modeling

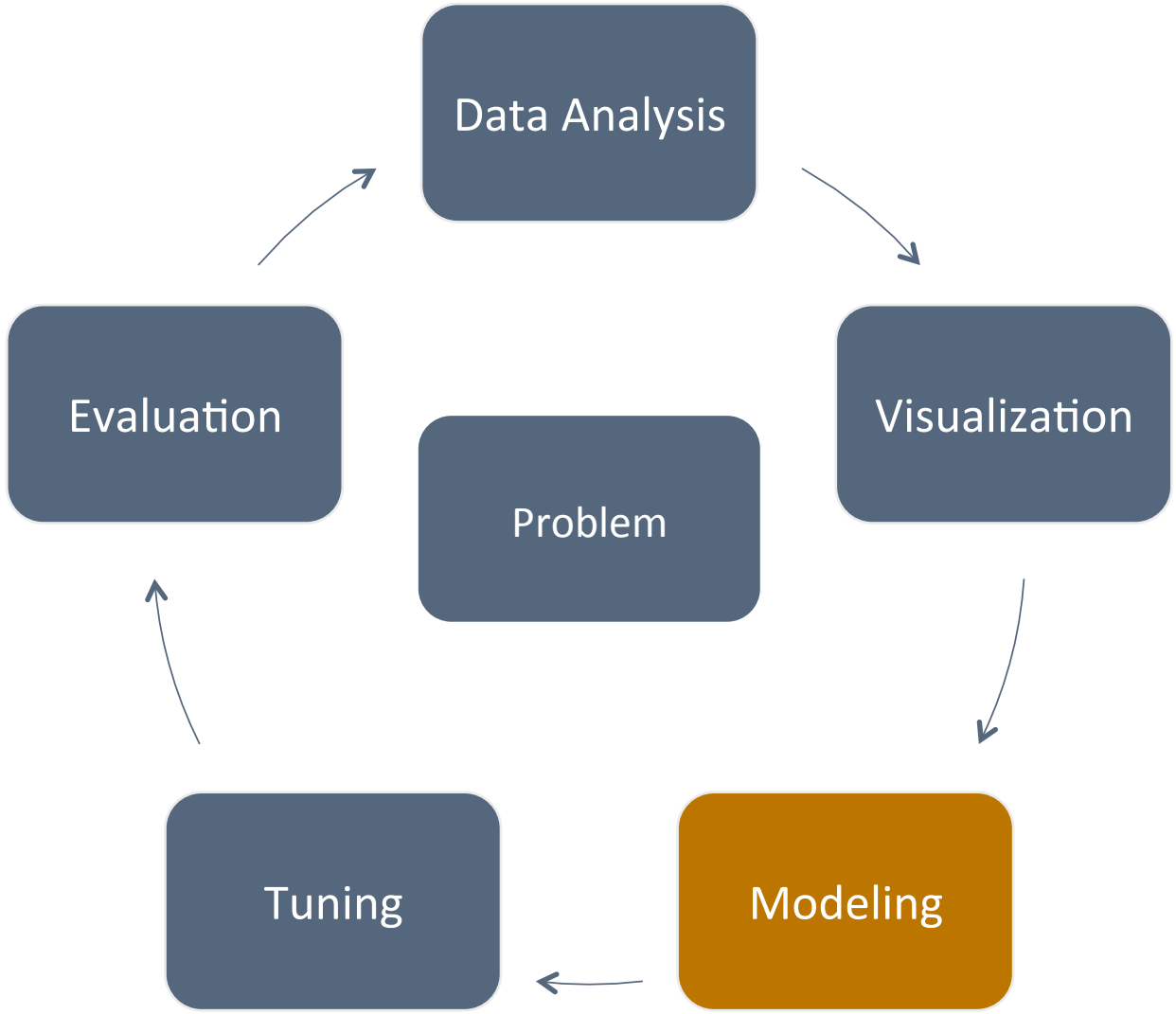
- Treat it as a binary classification problem first. Make sure the member of team, the setting of the positive and negative instance are the same.
- You can combine different classifiers. That is "ensemble" the classifiers.
- If you have some preliminary result look at the misclassified points carefully.



Data analysis lesson/experience

Modeling

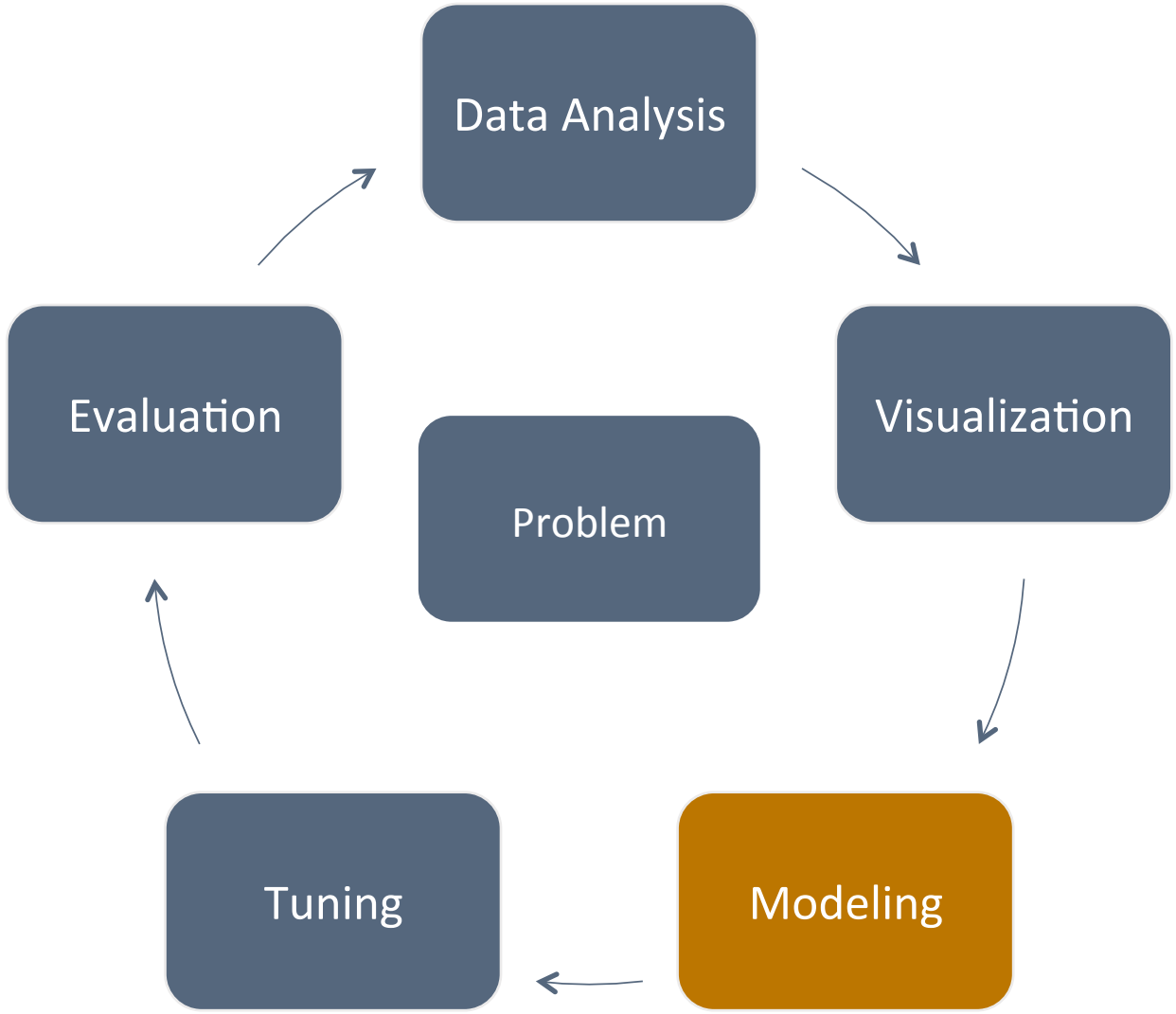
- Treat the problem as a binary classification problem first.
  - Try linear SVM first.
  - How much nonlinear SVMs can improve?
  - If you have some preliminary result look at the misclassified points carefully.
  - Try dimension reduction. I think kernel SIR is a very useful tool.
  - Check result can achieve the goal or not.



Data analysis lesson/experience

Modeling

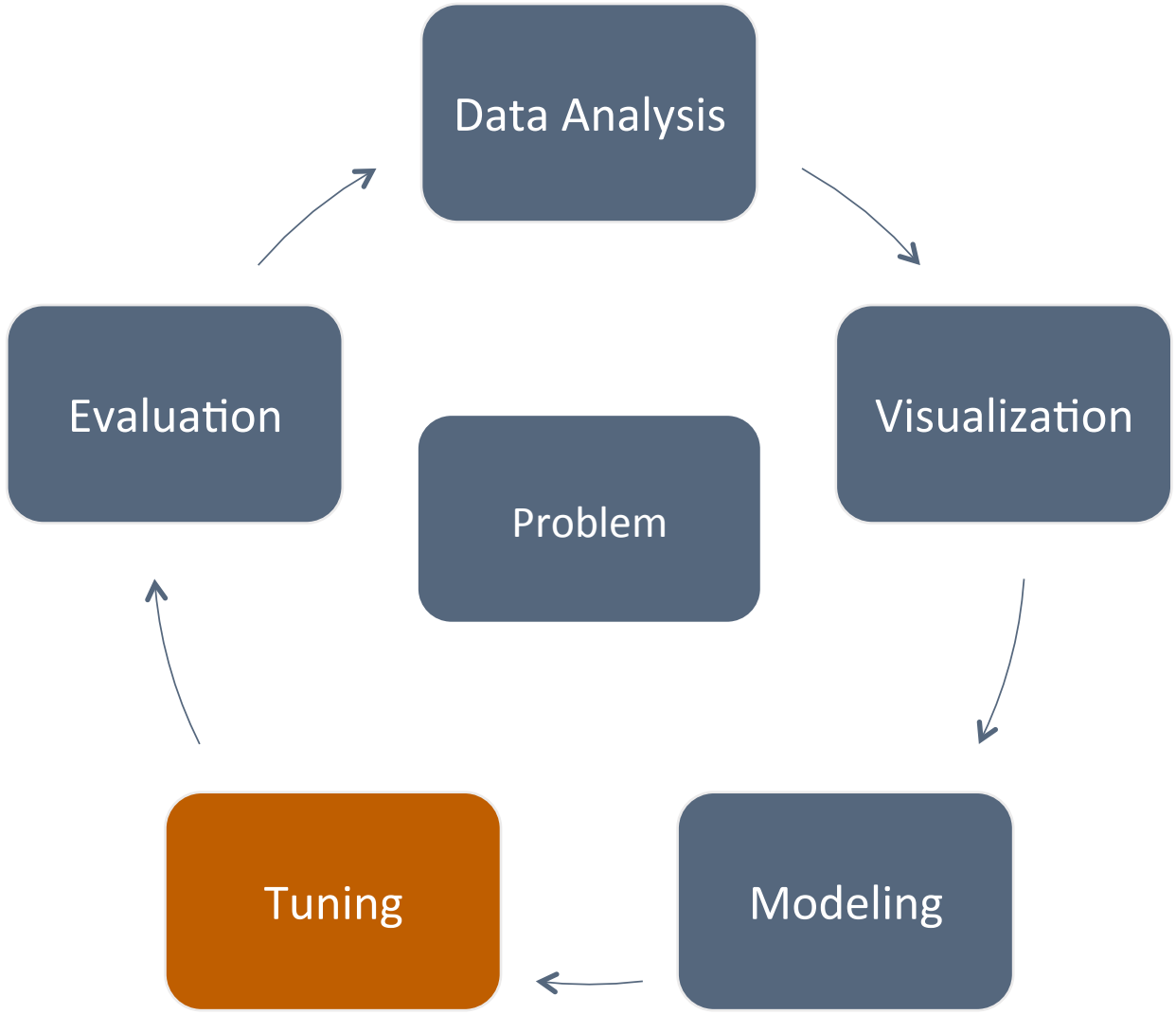
- Treat the problem as a binary classification problem first.
  - Find anything wired of the result.
  - Complain to data collector.
- Use simple algorithm to make base line.
  - Try decision tree, KNN, linear regression first.
  - Maybe even simple algorithm can help for selecting feature.



Data analysis lesson/experience

Tuning

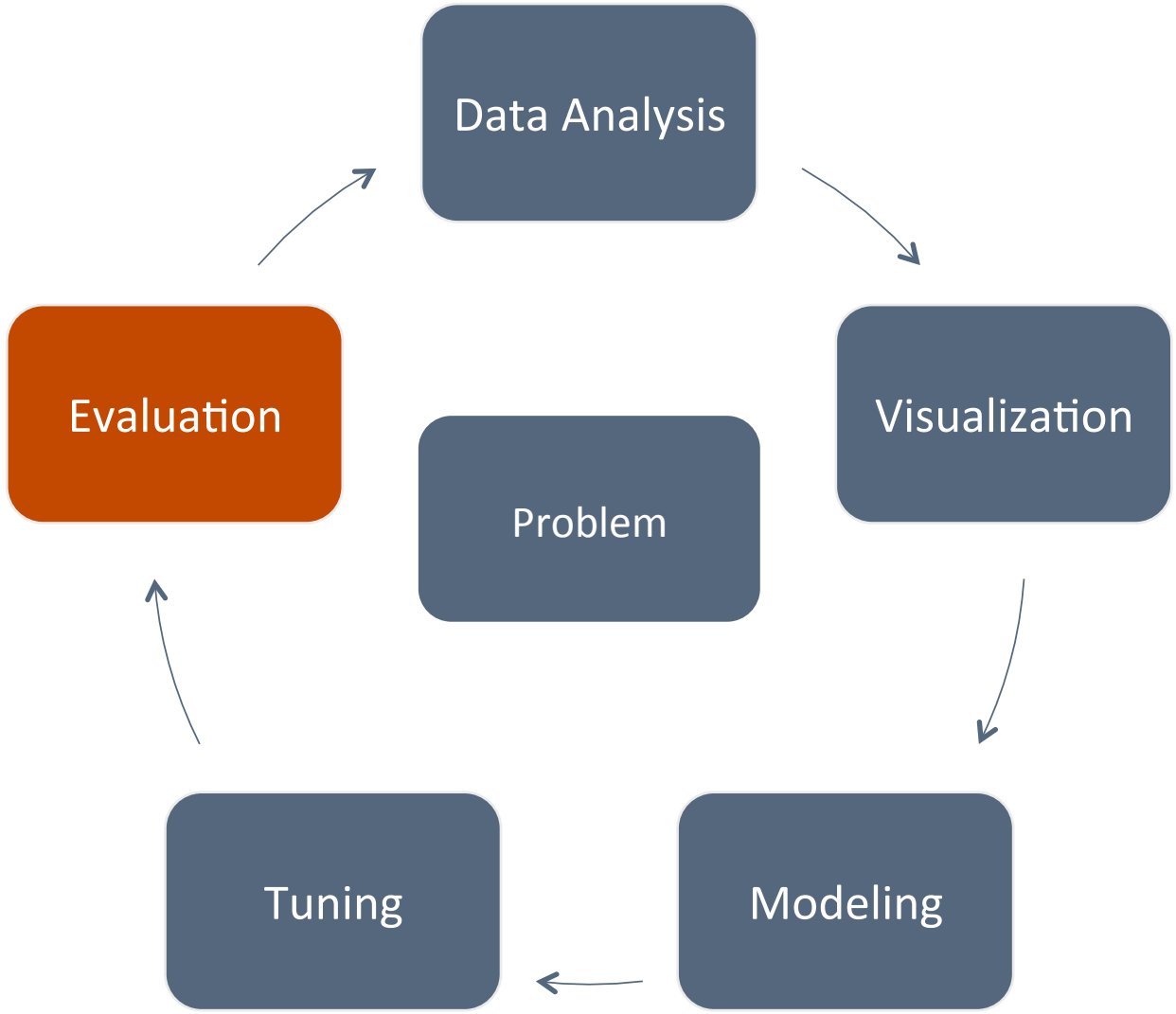
- Programing on auto-tuning or semi-auto- tuning.
- Write down the tuning parameter by case.



Data analysis lesson/experience

Evaluation

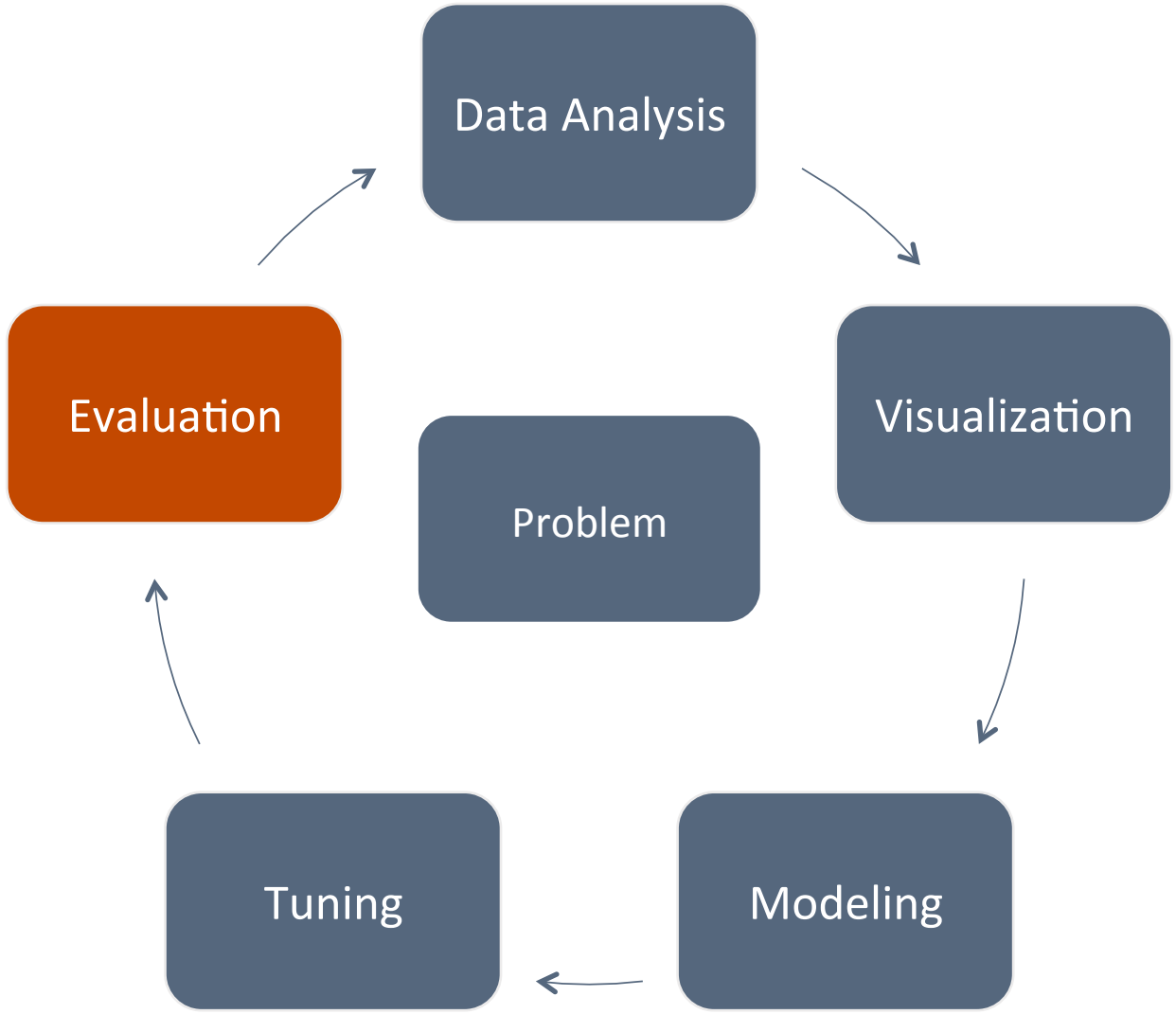
- Evaluating Models
- Hold-Out Testing Sets
- K-fold Cross validation
- Leave-One-Out Cross Validation
- Classifier Accuracy
- Confusion Matrix: Sensitivity, Specificity, Precision, NPV, FPR, FDR, FDR, Accuracy, F1 score, MCC.



Data analysis lesson/experience

Evaluation

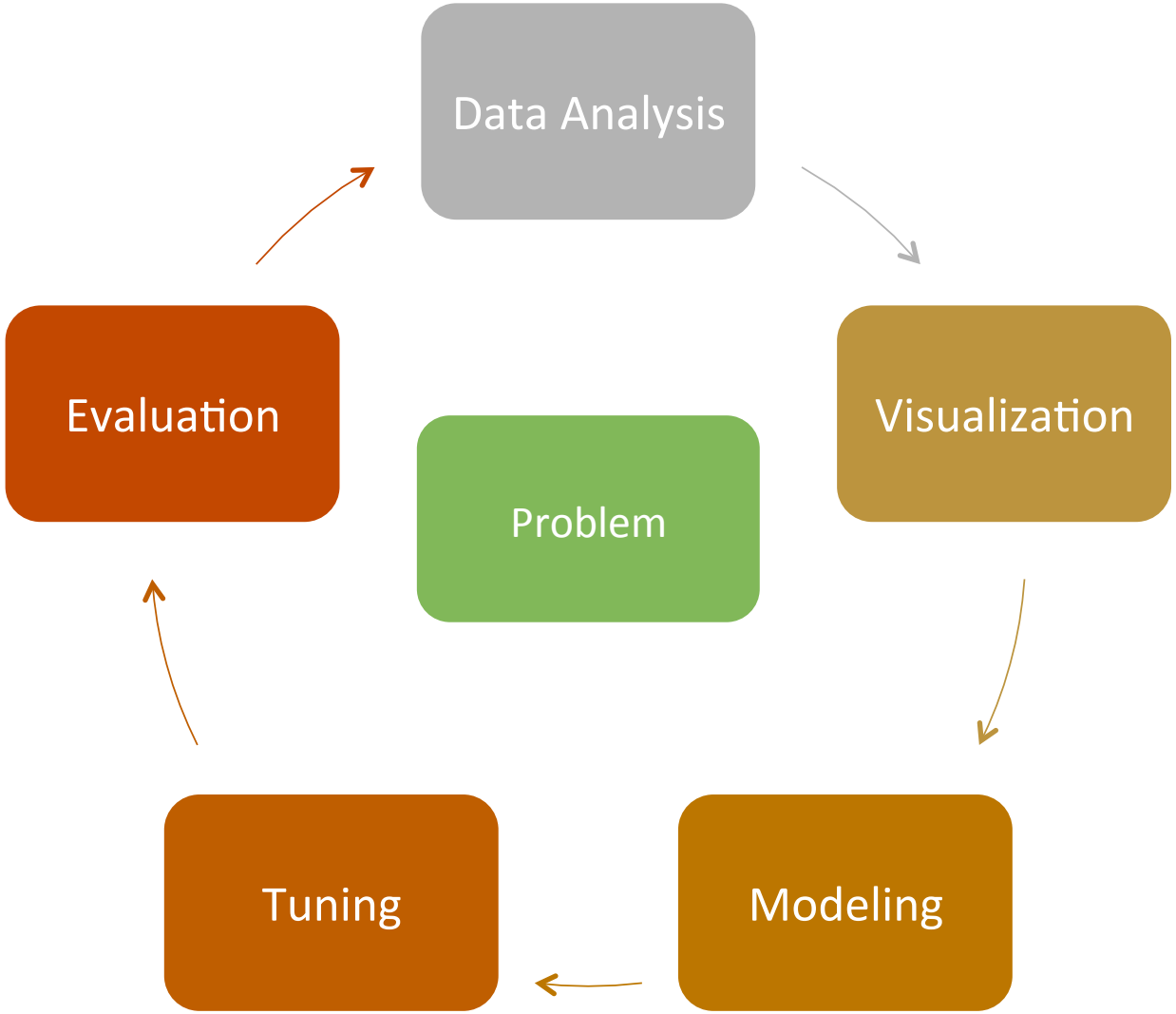
- Measure beyond simple accuracy.
- The Mahalanobis Distance.
- The Kolmogorov-Smirnov Distance.
- ROC Curves & Area under the Curves.



Data analysis lesson/experience

Note

- Try WEKA with different algorithms.
- Make sure you did NOT make any mistake in EVERY step.
- Compare each other.
- You have to WRITE down EVERY step and record the parameters.
- Make sure you are able to REPEAT your experiments.
- Backup your code and make sure the version.



Data analysis lesson/experience

Note

- Use API instead of software to improve controllability.
- You have to use scriptable tool so that repeat your experiment easily and treat script as step of experiment simply.
- Understanding the algorithms, whose basis of cognitive.
- Probably don't worry about the CPU time first.
- Make sure the member of team whose opinions are the same.