Machine Learning
Diagnostic Workflow

BMI701 Introduction of Biomedical Informatics
Lab Session 9

Wei-Hung Weng

November 30, 2016

HMS DBMI — MGH LCS

Machine Learning

- Optimize a performance criterion using example data or past experience
- Mathematically speaking: given data X, we want to learn a function mapping $f(X)$ for certain purpose
- $f(x) =$ a label $y \rightarrow$ classification
- $f(x) =$ a set Y in $X \rightarrow$ clustering
- $f(x) = p(x) \rightarrow$ distribution estimation
- ML techniques tell us how to produce high quality f(x), given certain objective and evaluation metrics

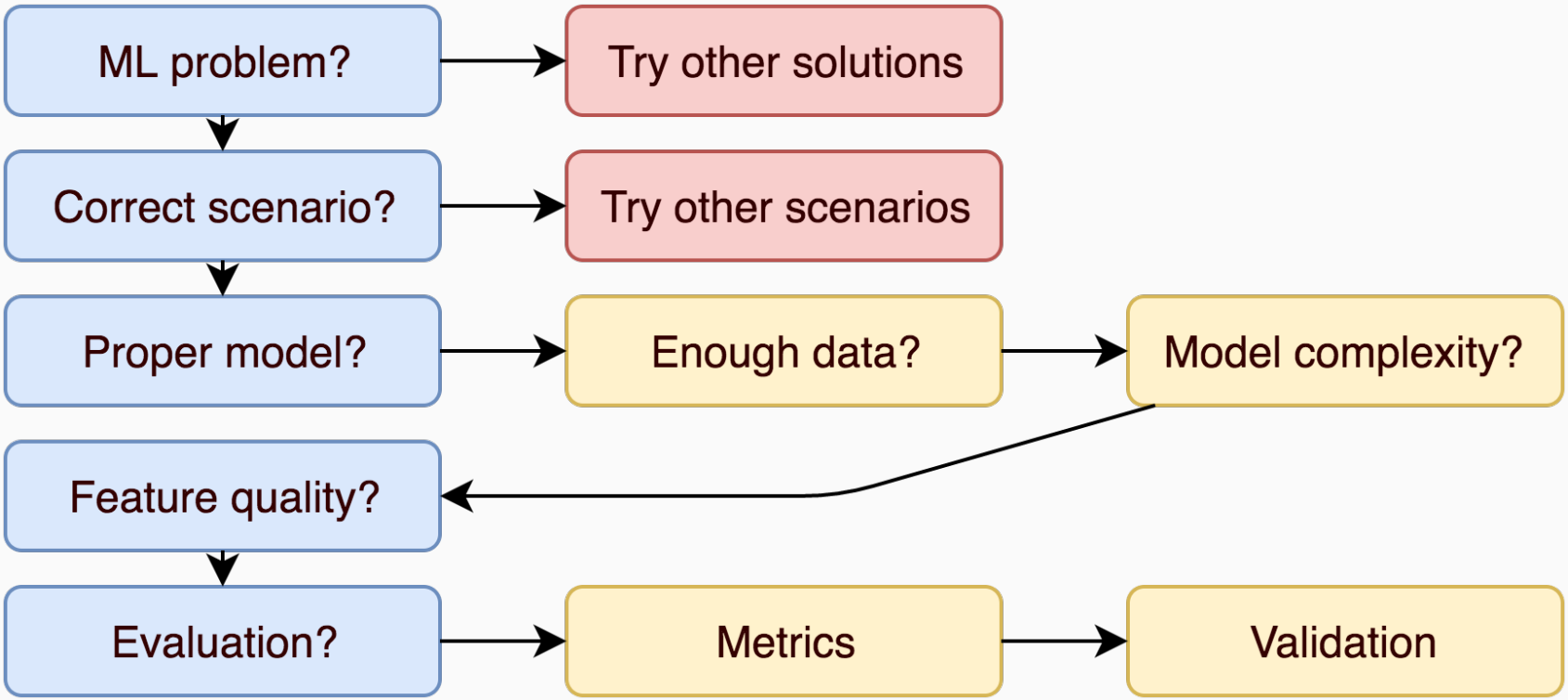
Courtesy by Prof SD Lin (NTU)

Machine Learning Three Steps

1. Modeling
    - function, features, parameters
2. Goodness of Function
    - Loss function
    - Gradient descent, learning rate, SGD
    - Local/global optima
3. Pick the 'Best' Function
    - Bias/variance trade-off
    - Overfitting (complex model)
    - Training/testing error
    - Regularization

- Why My Machine Learning Models Fail (meaning prediction accuracy is low)?

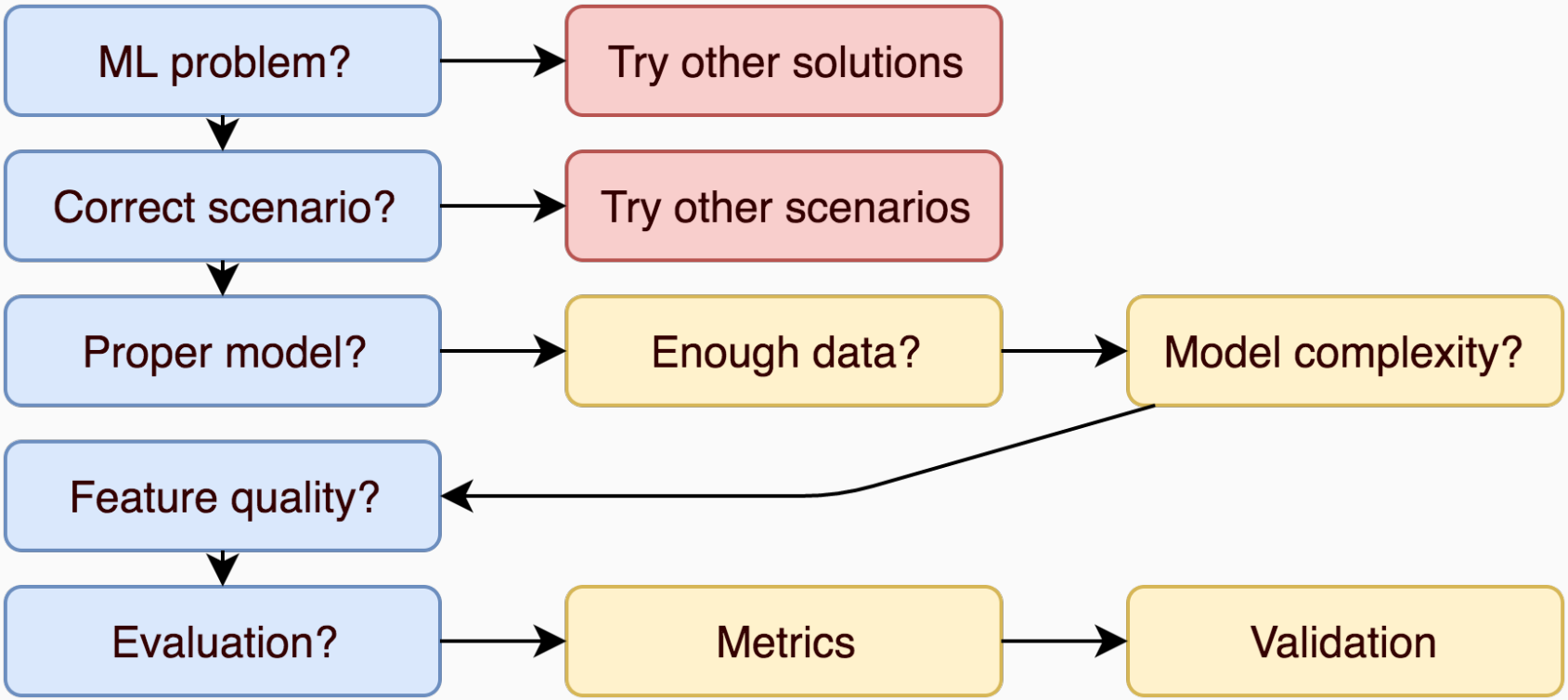
How to Create an Effective ML Model?



Checkpoint 1: ML Task?

- Are you sure Machine Learning is the best solution for your task?
- Too simple
  - Simple mapping: single word translation
  - Writing rules
- Too hard
  - X and y are independent

How to Create an Effective ML Model?



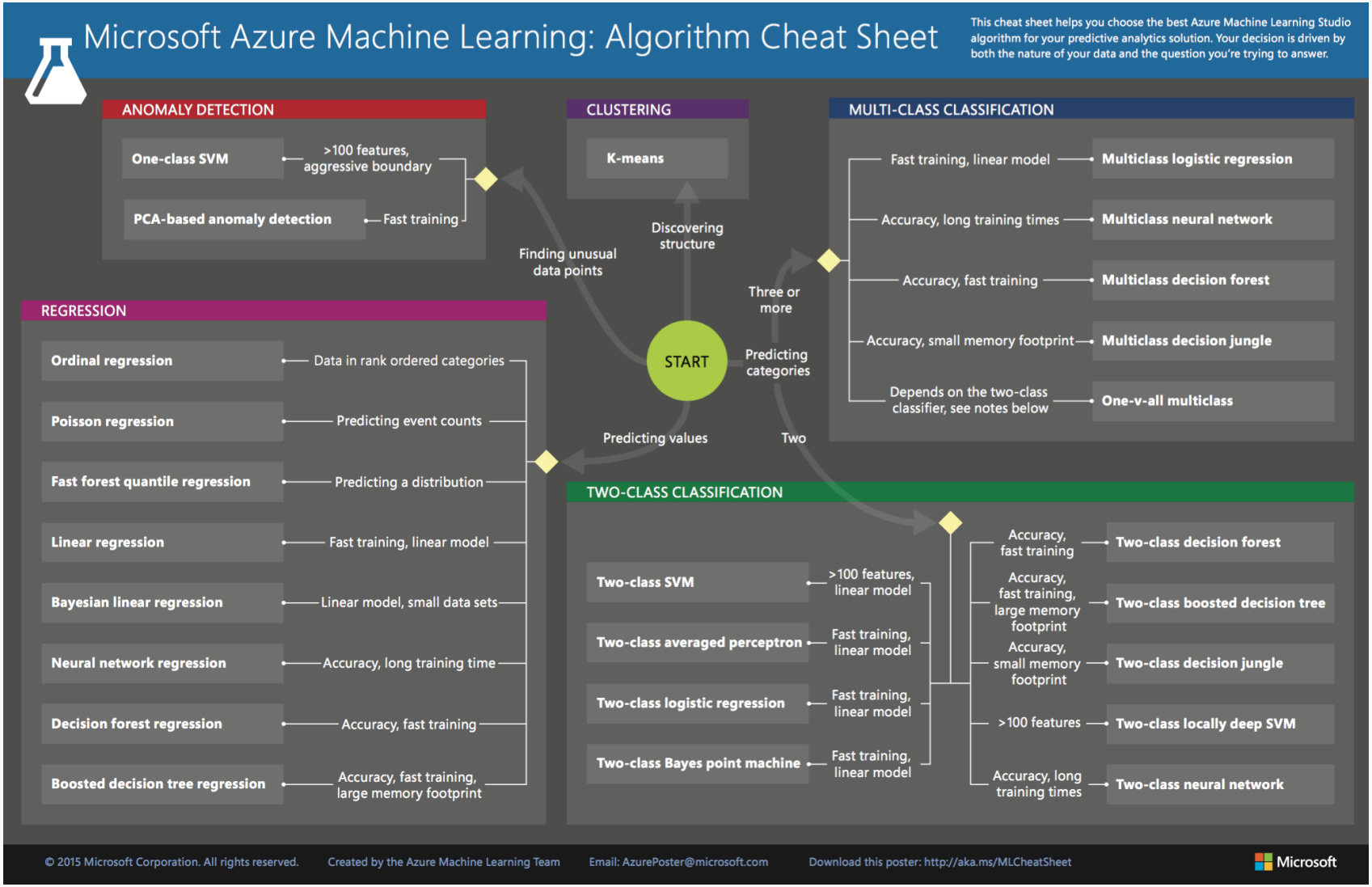
Checkpoint 2: What Kind of ML Scenario?

- Supervised: regression (real value), classification (categorical), structured learning
  - Classification: linear, non-linear (deep learning, SVM, decision tree, kNN)
    - Binary: spam
    - Multiclass: document classification
  - Multilabel learning
  - Structured learning: voice recognition, face recognition
- Semi-supervised
  - Active learning

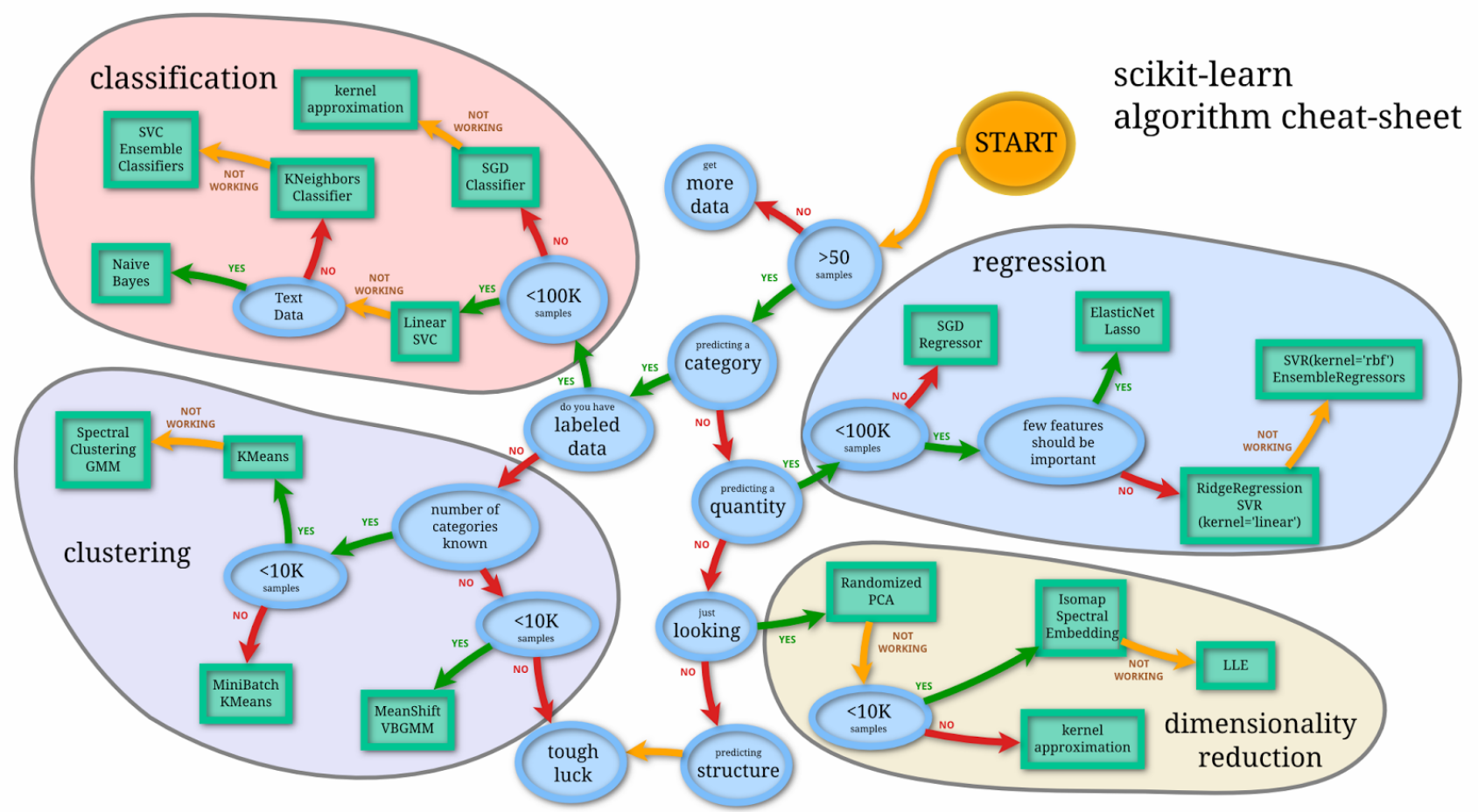
Checkpoint 2: What Kind of ML Scenario?

- Unsupervised: learning without teacher
  - Clustering, distribution learning, pattern learning, Bayesian, association rule, probabilistic graphical model
  - Machine reading, drawing
- Reinforcement: learning from critics
  - RL is a decision making process with states/actions/rewards
  - Goal is to find an optimal policy to guide the decision
  - AlphaGo = supervised + reinforcement
- Transfer learning, online learning, ...

Machine Learning Cheat Sheet

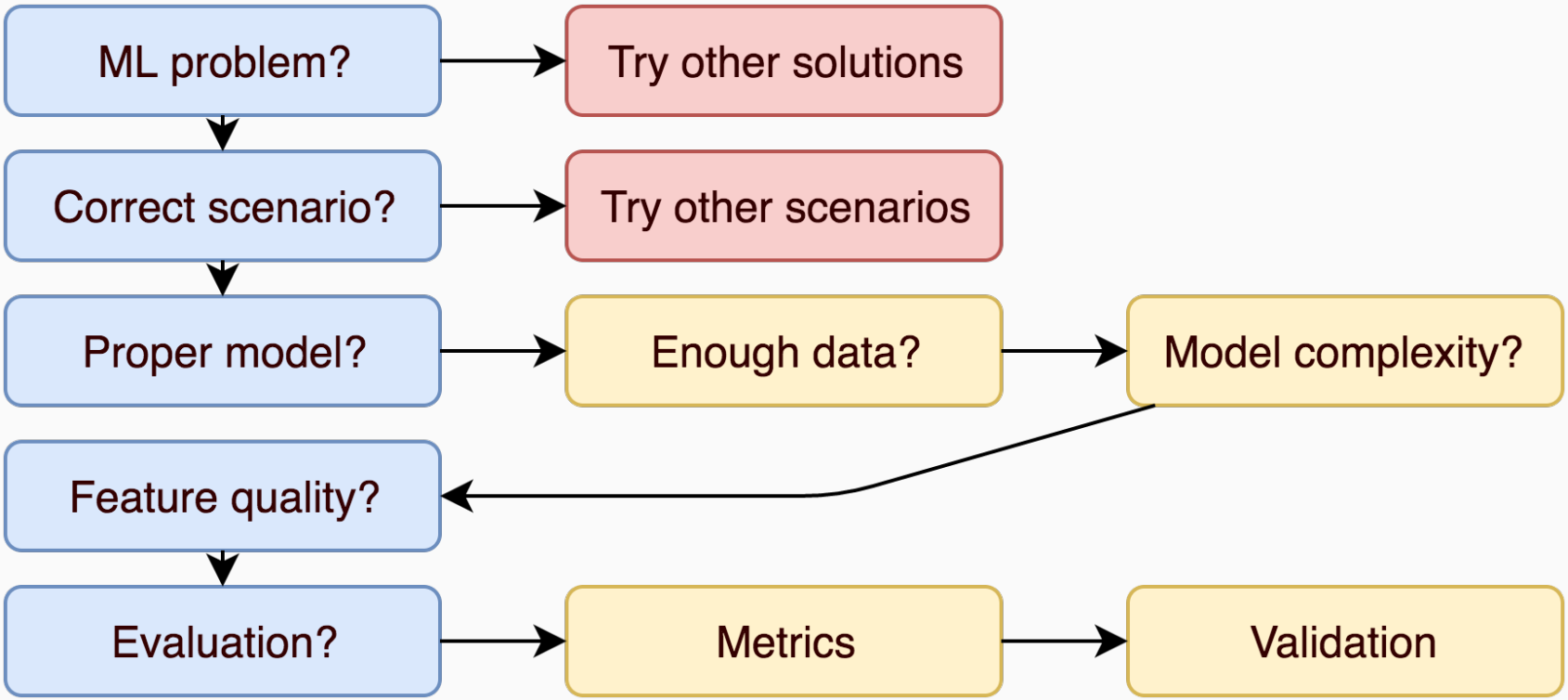


Machine Learning Cheat Sheet



Peekaboo

How to Create an Effective ML Model?



Checkpoint 3: Proper Model?

- Size of data
  - Small $\rightarrow$ linear model
  - Large $\rightarrow$ linear or non-linear
- Sparsity of data
  - Sparse $\rightarrow$ more tricks
  - Dense $\rightarrow$ light algorithm to save memory and computing power
- Balance of data
  - Imbalanced $\rightarrow$ weighting for minority class
- Quality of data (noise, missing values, ...)
  - Use different loss function. e.g. 0/1 loss or L2 are more robust to noise than hinge loss or exponential loss

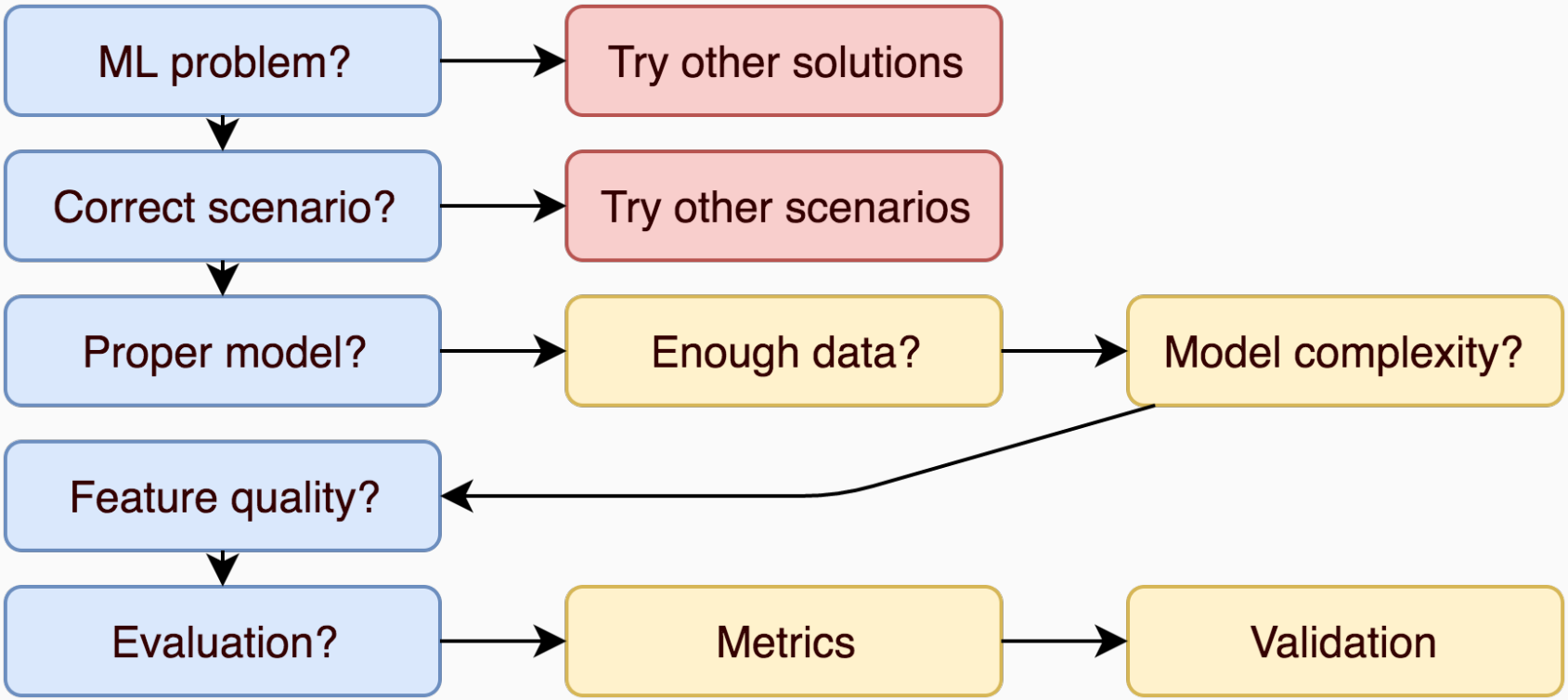
Checkpoint 4: Enough Data?

- Draw a learning curve
- The performance metric should converge if the data is sufficient

Checkpoint 5: Model too Complicated?

- Draw model complexity / loss function curve
- Training and validation (testing) error

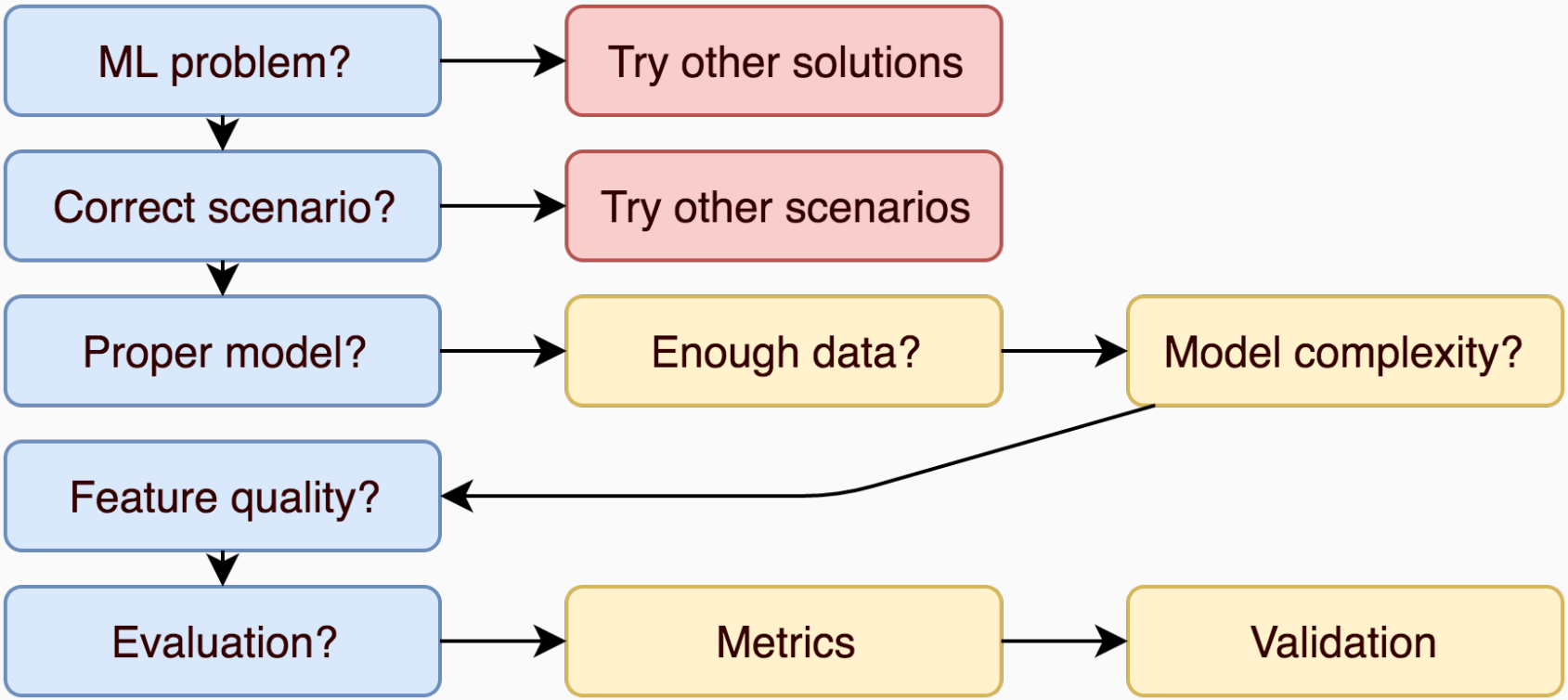
How to Create an Effective ML Model?



Checkpoint 6: Quality of Features?

- Feature engineering is the best strategy to improve performance
- Domain knowledge, human judgment
- If you don't know, then add it
- Algorithms can help you see whether it is useful or not
- χ^2 filtering, LASSO, elastic net, ...
- Different encoding, combined features, ...

How to Create an Effective ML Model?



Checkpoint 7: Correct Evaluation?

- Metrics
  - Accuracy is not always the best way
  - AUC
  - Precision, recall, F1 score
- Training, validation, testing
  - Cross-validation
  - Validation set for learning hyperparameter of your model!

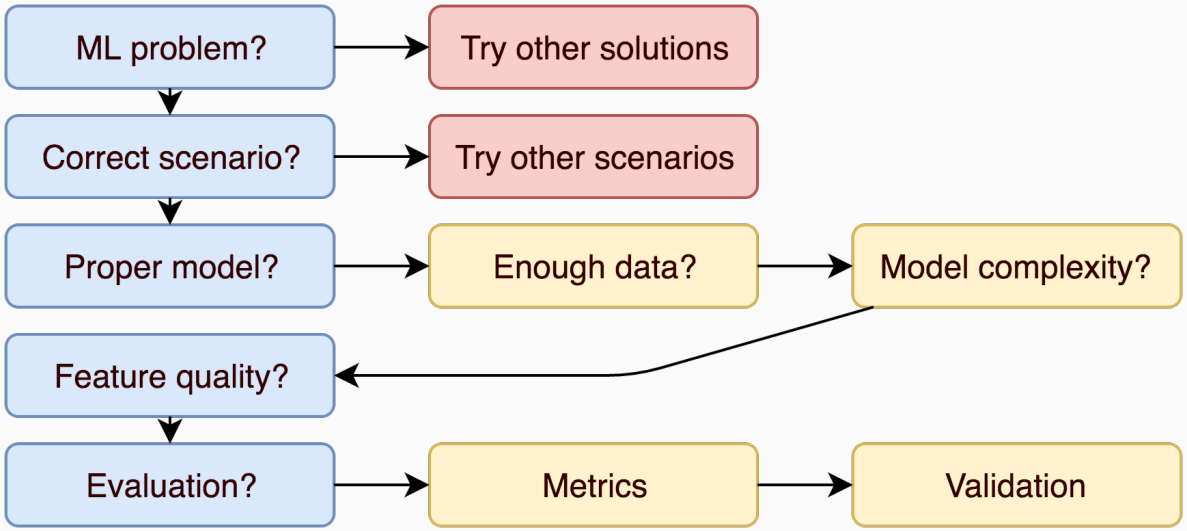
FAQ

- I have a lot of unlabeled data
  - Find expert to label them
  - Semi-supervised learning
  - Transfer learning: from other domains
  - Active learning: actively select a subset unlabeled data for expert (oracle) to annotate
- How to avoid overfitting?
  - Occam's Razor
  - Regularization: reduce the complexity
- How to boost the performance?
  - Feature engineering
  - Model combination (if they are different)

Practical Machine Learning

- R package
  - `caret, e1071, randomforest, rpart, ...`
  - Demo
- Python
  - `scikit-learn, gensim, theano, tensorflow`

Summary



- Contact
  - Github repository
  - ckbjimmy@gmail.com
  - Linkedin