Dữ Liệu Lớn

Advices for applying Machine Learning

Thân Quang Khoát

khoattq@soict.hust.edu.vn

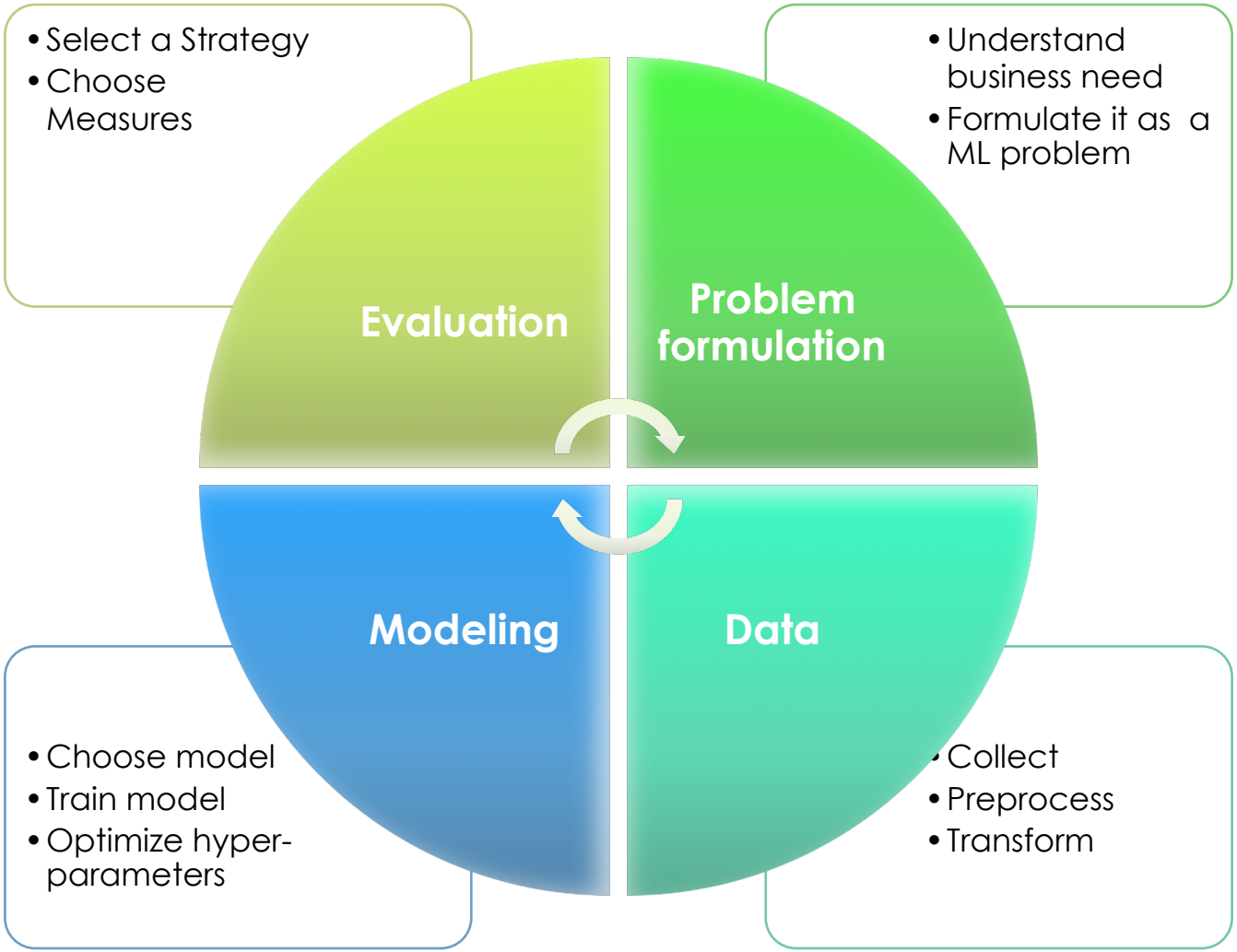
Viện Công nghệ thông tin và Truyền thông
Trường Đại Học Bách Khoa Hà Nội

Năm 2019

Nội dung khoá học

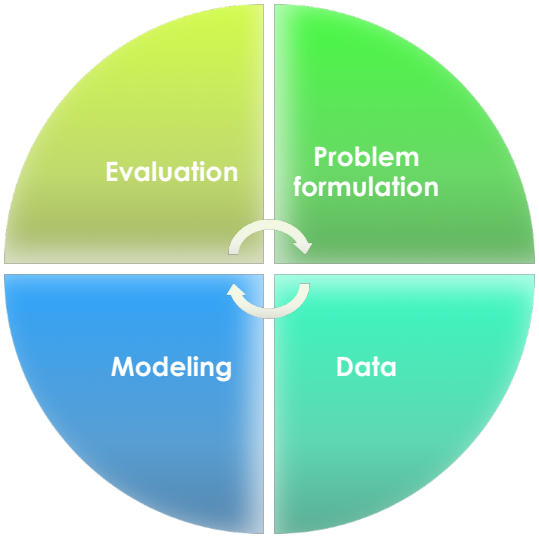
- Overview of data analytics/science
- Basic statistics
- Python and programming tools
- Exploratory data analysis
- Data integration and preprocessing
- Prediction with machine learning
- Data visualization
- **Evaluation of analysis results**
- Basics of natural language processing
- Anomaly detection
- Big data analysis
- Capstone project

Key steps in buiding a ML system



Problem formulation

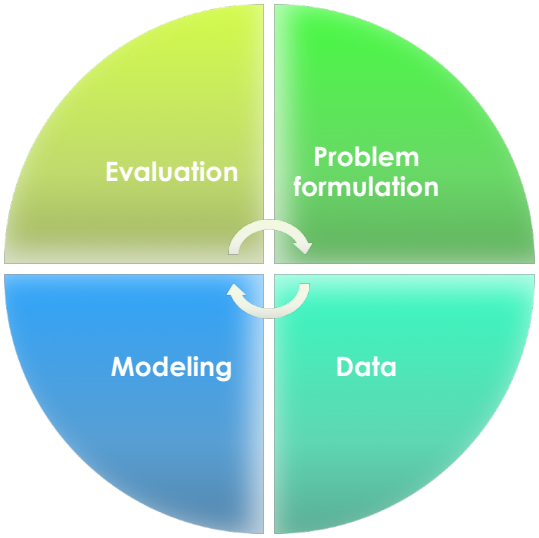
- Understand the business need clearly
- Formulate it as a machine learning problem
  - Supervised >< Unsupervised? How many classes?
  - Classification >< Regression?
  - …



- Exercise:
  - Anomaly detection: *detect which internet packages are attacks?*
  - Work in groups, formulate, identify pros/cons

Evaluation

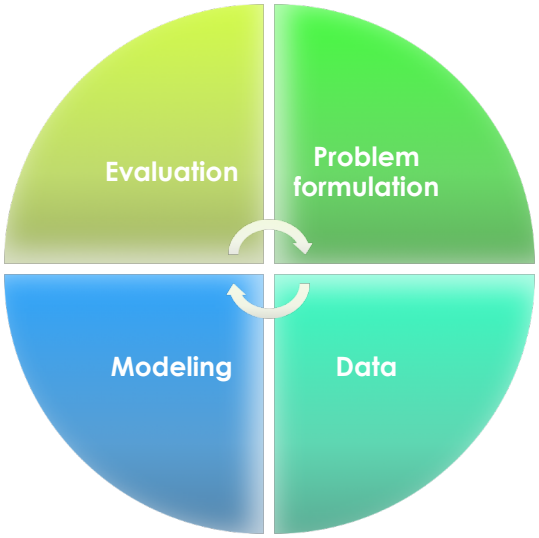
- How to make a reliable evaluation on the performance?
  - A good strategy: training, testing, and model selection
  - A correct measure of performance
  - Evaluation details



- Exercise:
  - Sentiment analysis: *detect what sentiment in a sentence is there?*
  - What measure for evaluation should be used?

Modeling

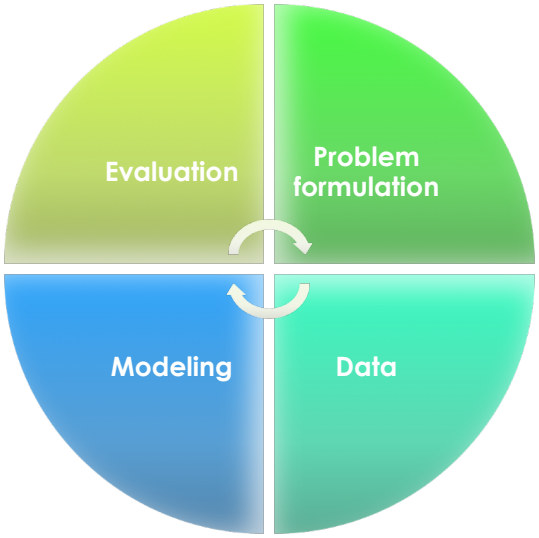
- Remember the "No-free-lunch Theorem".
- Try a simple model first
- Complex models are not always needed



- Ensemble of methods often performs as well as the best

Data

- Make sure you teach your model what you want it to learn
  - Select training data carefully
- Big data is not always needed
  - Training data should characterize the key properties of the whole space
- Unsupervised data might help much
  - Pre-training from unsupervised data



References

- Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009.
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.
- https://chatbotnewsdaily.com/10-more-lessons-learned-from-building-real-life-ml-systems-part-i-b309cafc7b5e