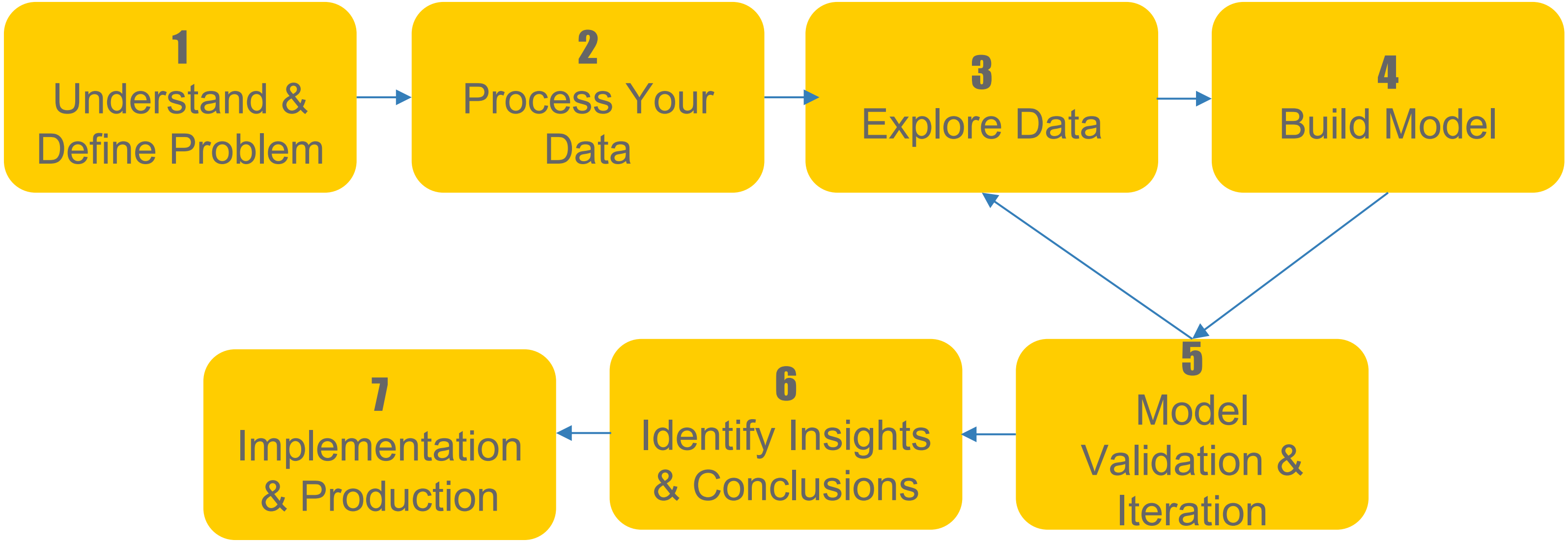
End to End Data Science & Open Ended Questions

Joyce

BITTIGER

来自硅谷的终身学习平台

End to End Data Science



1 Understand & Define Problem

2 Process Your Data

3 Explore Data

4 Build Model

5 Model Validation & Iteration

6 Identify Insights & Conclusions

7 Implementation & Production

1 Understand & Define Problem

Real world data science problems are mostly vaguely defined product & business problems

Ex: 1, How to improve pick-up experience in a Uber ride?

2, Survey showed teenagers are less engaged with Facebook after their parents join FB. What to do?

Data Scientist Need To:

1, **Understand** the problem. Decompose into small problems

2, **Translate** business problem to a quantified data problem

3, **Define** your data problem. Be clear of your **objectives**

2 Process Your Data

Get Data

1, Understand what's the data you need

2, Find data source (if existing, if not existing need to define logging schema and work with engineers to get the data)

Data Preprocessing

1, Validate data (understand definition, quality check, data inconsistency)

2, Clean data (missing data, invalid values, duplicate record, etc)

2, Data Transformation & Aggregation, etc

2.1 Manipulate Missing Values

Check

- How many missing?
- Random or Systematic ?
  - Actions needed if systematically missing (change logging, gather more data)

Treatment Methods

- Drop (not recommended unless very small amount)
- New level
- Mean / Mode /Median Imputation
- Model Imputation

3 Explore Data

Very Important! Spent plenty of time doing exploration before building models

1, Variable identification

- Different data type needs different analysis method
- type of variable: predictors, response
- data type: character, numeric - variable category: continuous, categorical

2, Exploratory visualization (correlation matrix, scatter plot, etc)

- Multi-collinearity (frequently asked)
- Normality (frequently asked)

3, Variable reduction

- Principle Component Analysis (hard to interpret)

4, Variable Creation (feature engineering)

- Good features are usually more important than fancy models
- We need domain knowledge

4 Build Model

Start with Simple Models! Interpretation is often more important than accuracy

1, Validate your assumptions (frequently asked)

2, Split data into Train/Validate/Test (Industry sometimes train/test)

3, Select your model, select your features (understand the pros & cons of each model)

5 Model Validation & Iteration

Evaluation Metrics

1, Define evaluation metrics

- MSE, MAE, Weighted MSE, etc

2, Compare performance of multiple models

3, Tune model for better performance.

- Change model
- Add / delete features, interaction terms
- Change model parameters

6 Identify Insights & Conclusions

Translate model result back to business insights

e.g. which feature is the most important for improving prediction

Sales volume will increase by X% if decrease price by 1%

A subgroup of users are more likely to take more rides if lower average price

Your summary should give **actionable recommendations in business language** (important for take-home)

e.g. The optimized price for product A is X, estimated +Y% revenue lift

Recommend testing a season pass package to user group A

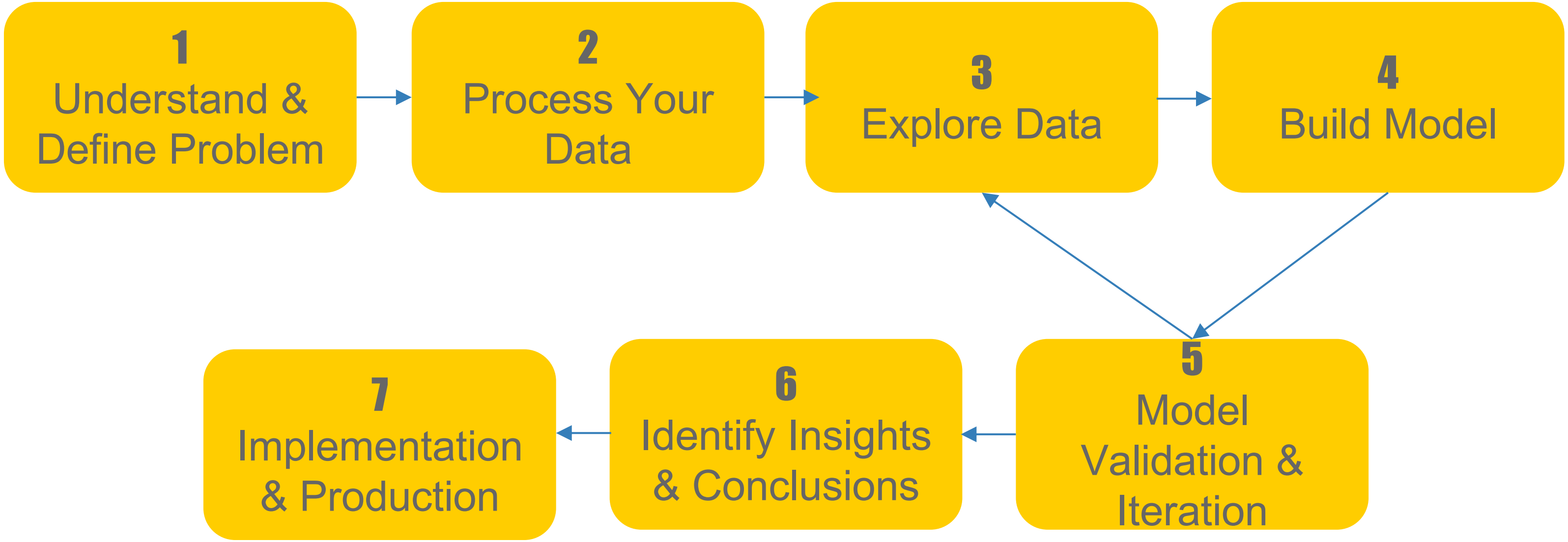
7 Implementation & Production

Data scientists’ work only makes value when it is implemented in practice

Collaborate with cross-functional partners

- Integrate models into the production system (needs more engineering)
  - E.g. Uber real-time driver rider matching algorithms
- Influence business decisions
  - E.g. Leadership decide to develop new product for user group A
- Influence business operations
  - E.g. Marketing team send out coupons to users based on your prediction

Review



1 Understand & Define Problem

2 Process Your Data

3 Explore Data

4 Build Model

5 Model Validation & Iteration

6 Identify Insights & Conclusions

7 Implementation & Production