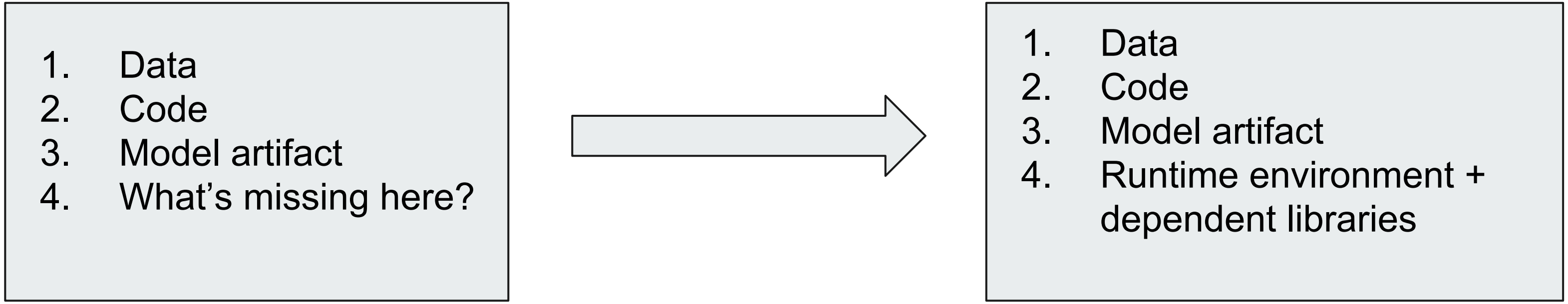
Model deployment

Trung Nguyen
Sep 7th, 2018

Why do we care?

- So far you’ve built models that can make predictions on your local machine.
  - A model is **rarely useful on its own**, on your machine
  -
- To make it more useful, it should:
  - Be reproducible -- anyone should be able to run the model on their own machine
  - Be integrated into a production environment powering a business solution
  -
- E.g. Facial recognition in Facebook photos, video recommendations on Youtube

How to deploy



1. Data
2. Code
3. Model artifact
4. What’s missing here?

Your local machine

1. Data
2. Code
3. Model artifact
4. Runtime environment + dependent libraries

Any other machine (*with proper set up)*

Deployment methods

Today you will learn about:

- Local deployment (for testing purposes only)
  - Your local machine is the web server
- Cloud deployment (Heroku)
  - A remote machine is the web server
- Using AWS Lambda for development and deployment - AWS
  - AWS ECS containers is the web server
- Cognitive services - APIs provided by others

Lecture materials:
https://github.com/trungngv/python-machine-learning-book-2nd-edition/tree/master/code/ch08