Applied Data Science and Machine Learning Course

Mohammad Sabik Irbaz
Lead Machine Learning Engineer
Pioneer Alpha Ltd.
sabik@pioneeralpha.com

Amar iSchool

Powered by Pioneer Alpha Ltd.

Lecture – 02

Data Science Pipeline

MSI

TABLE OF CONTENTS

TABLE OF CONTENTS

01 Introduction

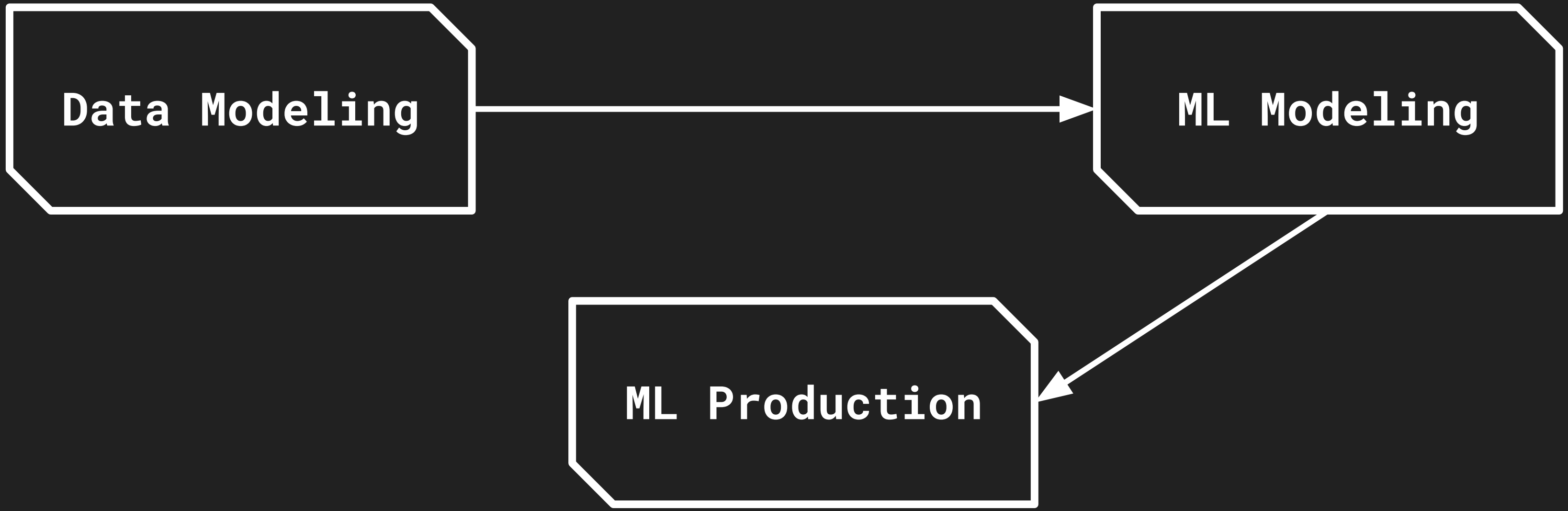
02 Data Modeling

03 ML Modeling

04 ML Production

Introduction

Any data science project can be divided into 3 parts



Data Modeling

ML Modeling

ML Production

MSI

Reality is often disappointing

Introduction

Any data science project can be divided into 3 parts

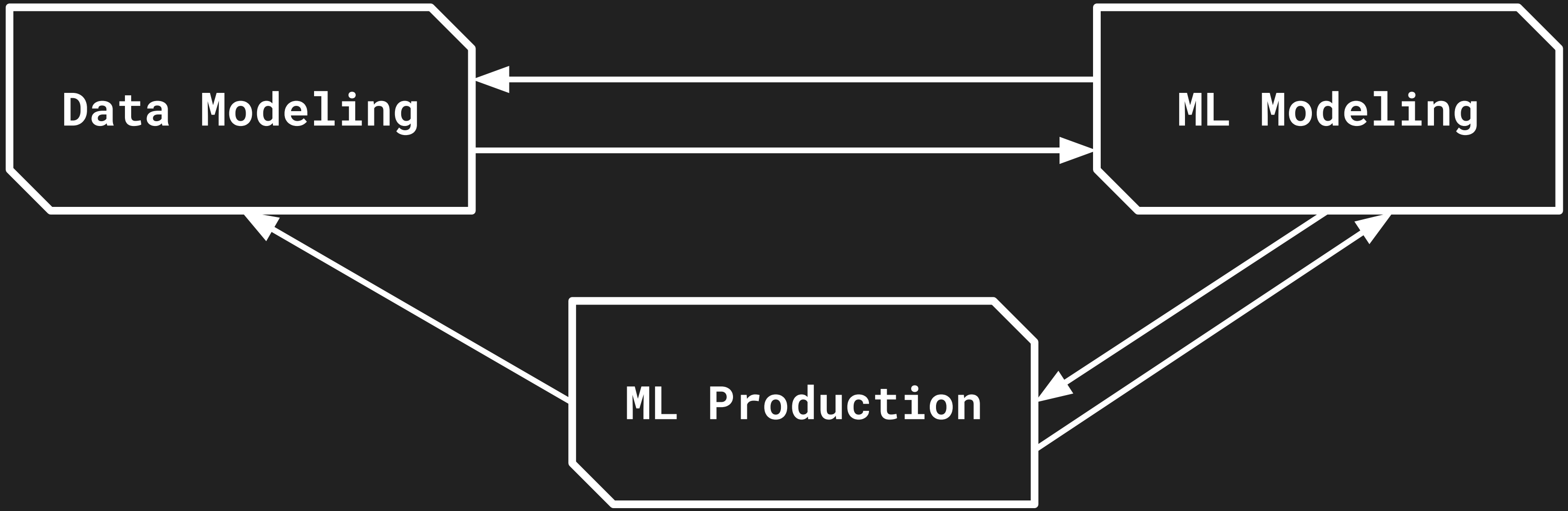


TABLE OF CONTENTS

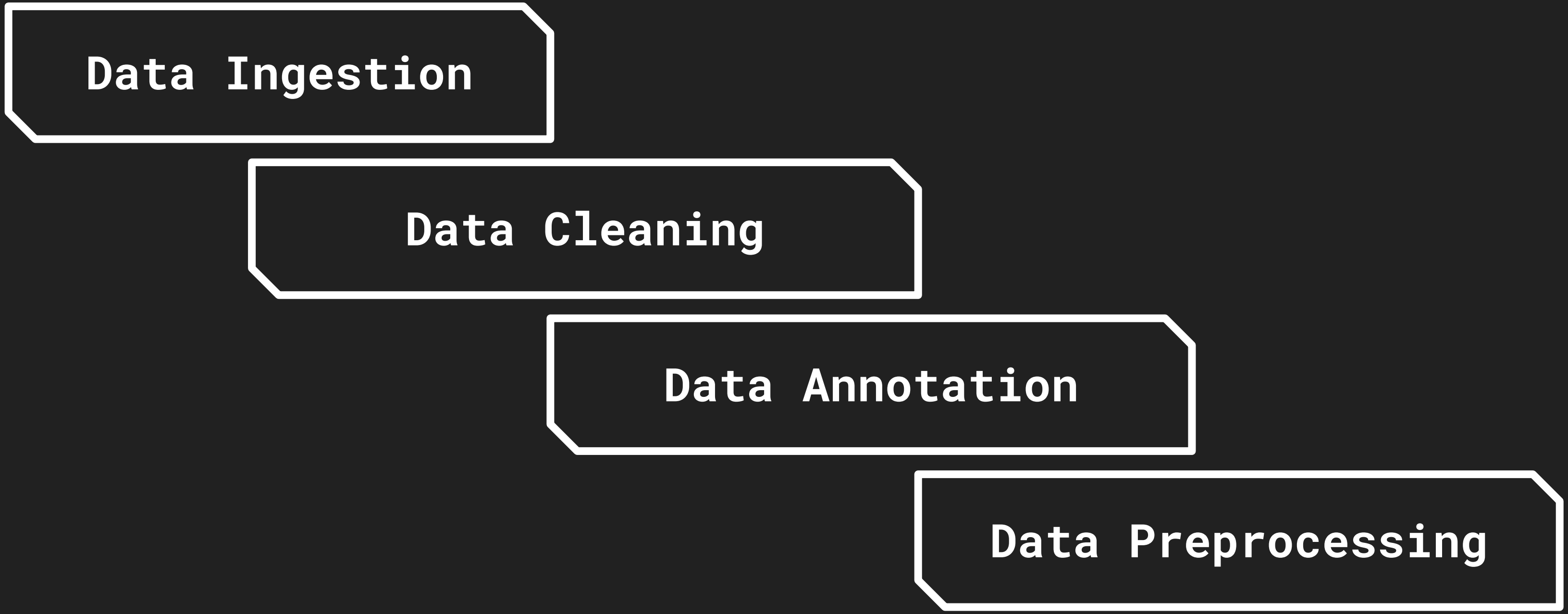
01 Introduction

02 Data Modeling

03 ML Modeling

04 ML Production

Data Modeling



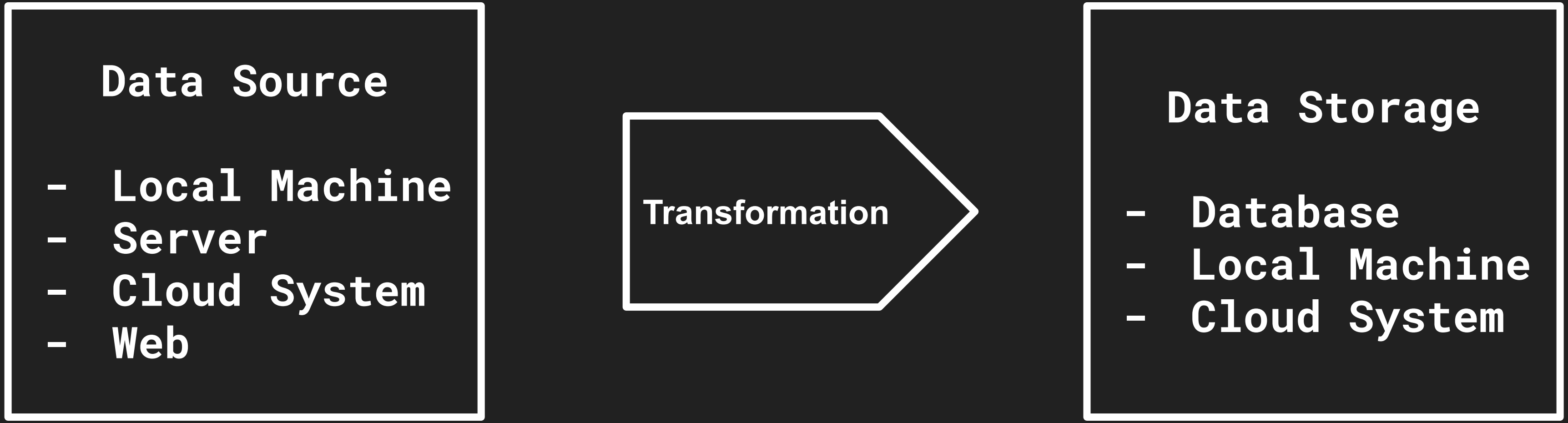
Data Ingestion

Data Cleaning

Data Annotation

Data Preprocessing

Data Ingestion



Data Source

- Local Machine
- Server
- Cloud System
- Web

Transformation

Data Storage

- Database
- Local Machine
- Cloud System

Data Cleaning and Annotation

- Removing Noise
- Removing Data not related to end-goal
- Fix mislabeled data during previous iteration

Annotating Data for Supervised or Semi-supervised Learning

Data Preprocessing

Feature Engineering

Feature Transformation

Augmentation

TABLE OF CONTENTS

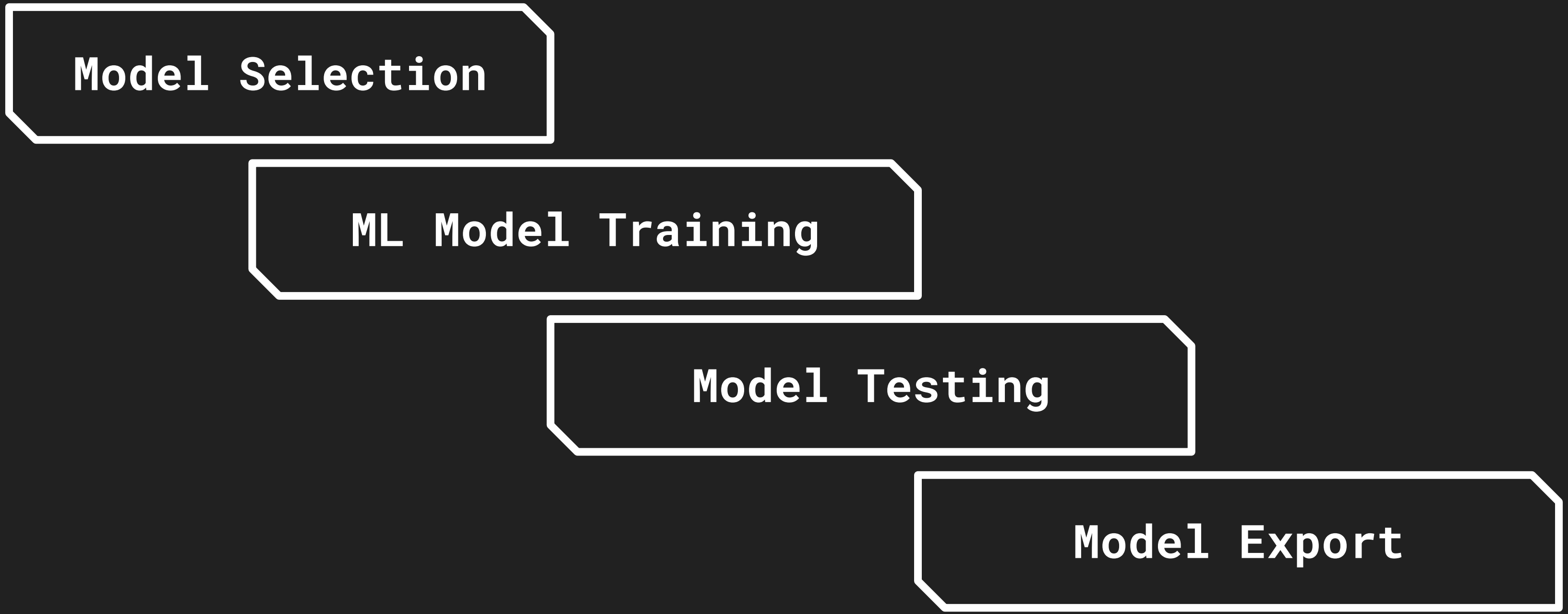
01 Introduction

02 Data Modeling

03 ML Modeling

04 ML Production

ML Modeling



Model Selection

ML Model Training

Model Testing

Model Export

TABLE OF CONTENTS

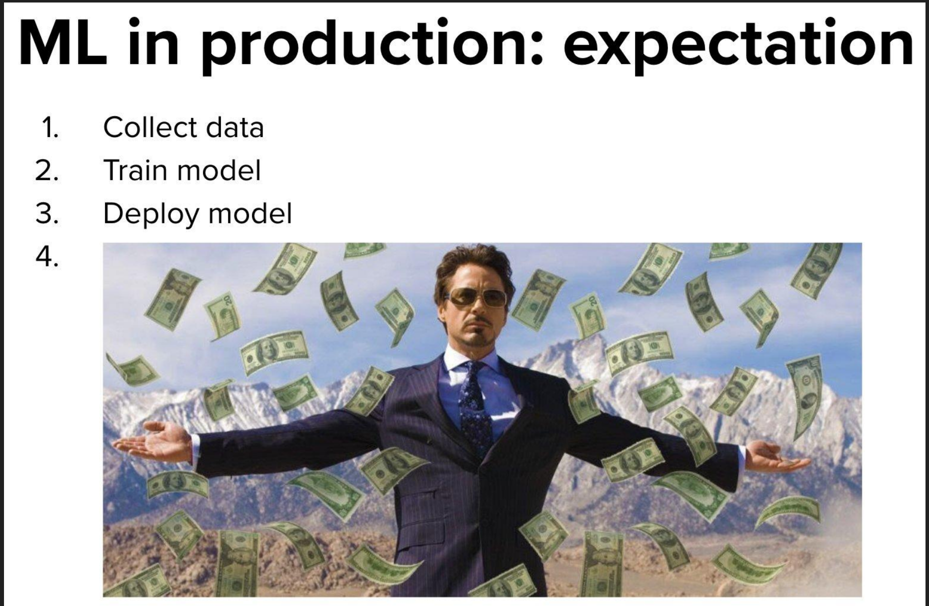
01 Introduction

02 Data Modeling

03 ML Modeling

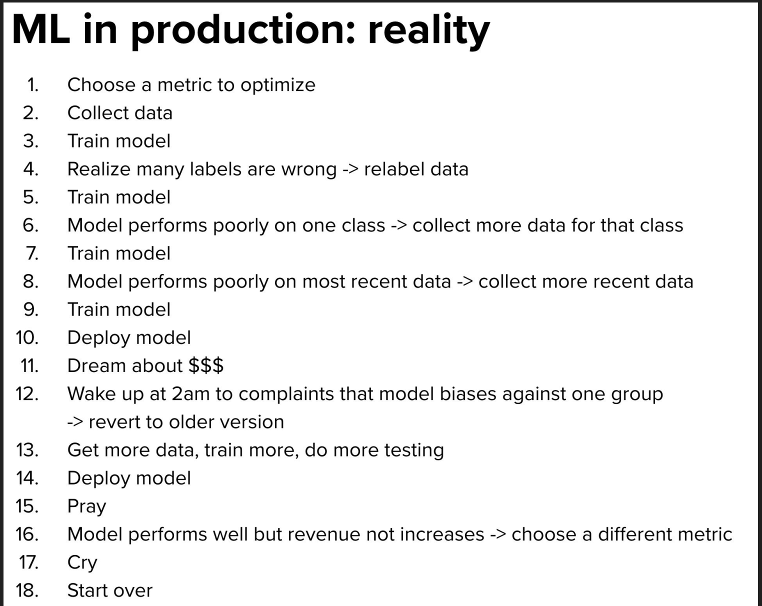
04 ML Production

ML Production



ML in production: expectation

1. Collect data
2. Train model
3. Deploy model
4.



ML in production: reality

1. Choose a metric to optimize
2. Collect data
3. Train model
4. Realize many labels are wrong -> relabel data
5. Train model
6. Model performs poorly on one class -> collect more data for that class
7. Train model
8. Model performs poorly on most recent data -> collect more recent data
9. Train model
10. Deploy model
11. Dream about $$$
12. Wake up at 2am to complaints that model biases against one group -> revert to older version
13. Get more data, train more, do more testing
14. Deploy model
15. Pray
16. Model performs well but revenue not increases -> choose a different metric
17. Cry
18. Start over

Chip Huyen (Teaches MLSys at Stanford University)

Discussion
on
Course Project