Machine Learning as a Service

Amit Kapoor

amitkaps.com

Anand Chitpothu

anandology.com

Getting Started

- Download the Repo: https://github.com/amitkaps/full-stack-data-science
- Finish installation
- Run jupyter notebook in the console

Motivation

- Solve a business problem.
- Understand the end-to-end process
- Build a Machine Learning application

"Jack of all trades, master of none, though oft times better than master of one."

Approach

- Simple approach
- Go wide vs. go deep
- Practical and scalable

Schedule

1. **Introduction, Setup** (10 mins)
2. **ML Process, Frame** - *Conceptual* (20 mins)
3. **Acquire, Refine, Explore** - *Coding* (30 mins)
4. **Transform, Model** - *Coding* (40 mins)
   -- Break (15 mins)--
5. **Building an ML Application** - *Conceptual* (10 mins)
6. **Deploy the ML Model as Service** - *Coding* (20 mins)
7. **Wiring the Model** - *Coding* (20 mins)
8. **Wrap-up** (15 mins)

Data-Driven Lens

"Data is a clue to the End Truth"

— Josh Smith

Metaphor

— A start-up providing loans to the consumer

— Running for the last few years

— Now planning to adopt a data-driven lens

What are the **type of questions** you can ask?

Type of Questions

- What is the trend of loan defaults?
- Do older customers have more loan defaults?
- Which customer is likely to have a loan default?
- Why do customers default on their loan?

Type of Questions

- Descriptive
- Inquisitive
- Predictive
- Causal

Data-driven Analytics

- **Descriptive**: Understand Pattern, Trends, Outlier
- **Inquisitive**: Conduct Hypothesis Testing
- **Predictive**: Make a prediction
- **Causal**: Establish a causal link

Prediction Challenge

It's tough to make predictions, especially about the future.

— Yogi Berra

How to make a Prediction?

- **Human Learning**: Make a *Judgement*
- **Machine Programmed**: Create explicit *Rules*
- **Machine Learning**: Learn from *Data*

Machine Learning (ML)

[Machine learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

— *Arthur Samuel*

Machine learning is the study of computer algorithm that improve automatically through experience

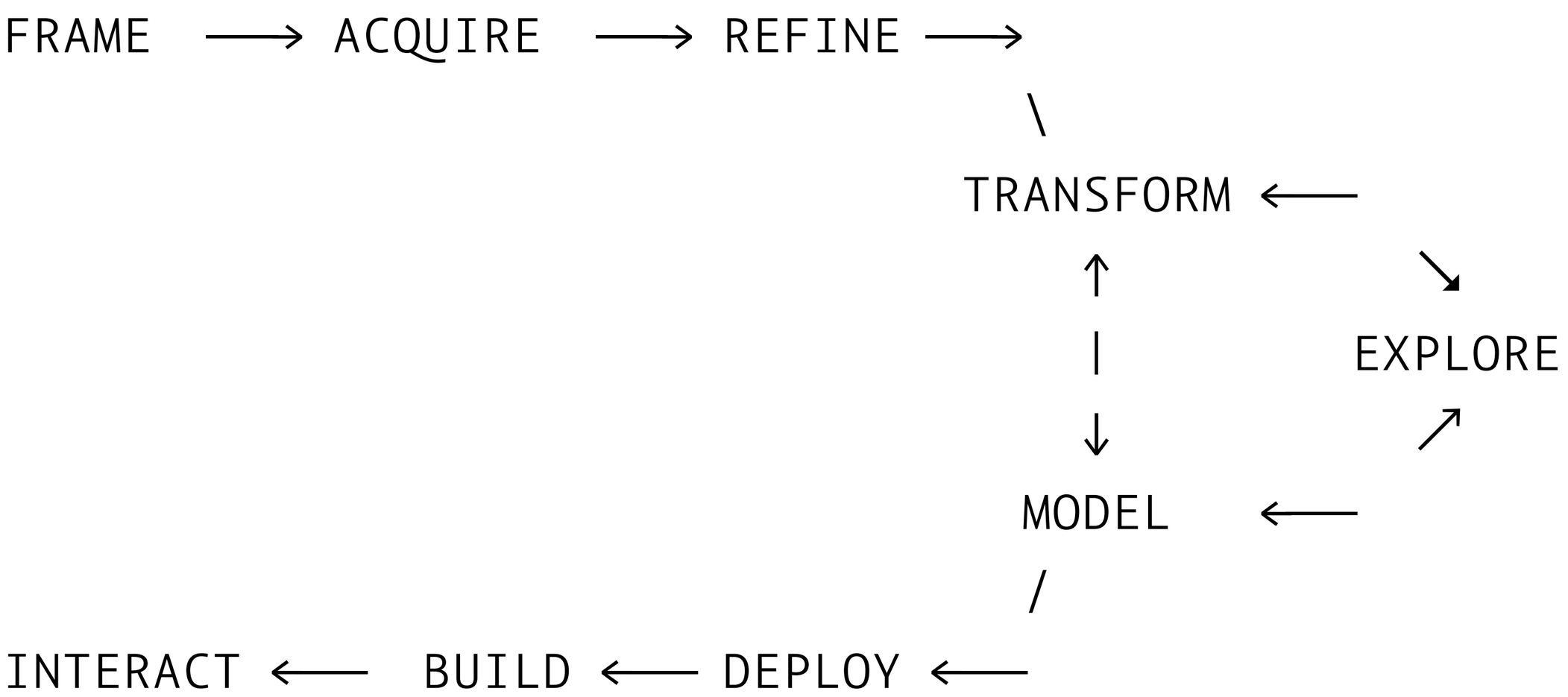
— *Tom Mitchell*

Machine Learning: Essense

- A pattern exists
- It cannot be pinned down mathematically
- Have data on it to learn from

"Use a set of observations (data) to uncover an underlying process"

ML as a Service (MLaaS) Approach



MLaaS Approach

- *Frame*: Problem definition
- *Acquire*: Data ingestion
- *Refine*: Data wrangling
- *Transform*: Feature creation
- *Explore*: Feature selection
- *Model*: Model creation & selection
- *Deploy*: Model deployment
- *Build*: Application building
- *Interact*: User interaction

ML Theory: Data Types

- What are the types of data on which we are learning?
- Can you give example of say measuring temperature?

Data Types e.g. Temperature

- **Categorical**
  - *Nominal*: Burned, Not Burned
  - *Ordinal*: Hot, Warm, Cold
- **Continuous**
  - *Interval*: 30 °C, 40 °C, 80 °C
  - *Ratio*: 30 K, 40 K, 50 K

Data Types - Operations

- **Categorical**
  - *Nominal*: = , !=
  - *Ordinal*: =, !=, >, <
- **Continuous**
  - *Interval*: =, !=, >, <, -, % of diff
  - *Ratio*: =, !=, >, <, -, +, %

Case: Loan Default Prediction

Application Attributes

- **age**: age of the applicant
- **income**: annual income of the applicant
- **year**: no. of years of employment
- **ownership**: type of house owned
- **amount** : amount of loan requested by the applicant

Behavioural Attributes:

- **grade**: credit grade of the applicant

Question - whether the applicant will **default** or not?

Historical Data

default	amount	grade	years	ownership	income	age
0	1,000	B	2.00	RENT	19,200	24
1	6,500	A	2.00	MORTGAGE	66,000	28
0	2,400	A	2.00	RENT	60,000	36
0	10,000	C	3.00	RENT	62,000	24
1	4,000	C	2.00	RENT	20,000	28

Data Types

- **Categorical**
  - *Nominal*: home owner [rent, own, mortgage]
  - *Ordinal*: credit grade [A > B > C > D > E]
- **Continuous**
  - *Interval*: approval date [20/04/16, 19/11/15]
  - *Ratio*: loan amount [3000, 10000]

ML Terminology

Features: $\mathbf{x}$

- age, income, years, ownership, grade, amount

Target: y

- default

Training Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \ldots (\mathbf{x}_n, y_n)$

- historical records

ML Paradigm: Supervised

Given a set of **feature x**, to predict the value of **target** y

Learning Paradigm: **Supervised**

— If y is *continuous* - **Regression**

— If y is *categorical* - **Classification**

Frame

Variables

- age, income, years, ownership, grade, amount, default and interest

— What are the **Features**: **x** ?

— What are the **Target**: y

Frame

Features: $\mathbf{x}$

- age
- income
- years
- ownership
- grade
- amount

Target: y

- default

Acquire

— Simple! Just read the data from csv file

Refine - Missing Value

- **REMOVE** - NAN rows
- **IMPUTATION** - Replace them with something?
  - Mean
  - Median
  - Fixed Number - Domain Relevant
  - High Number (999) - Issue with modelling
- **BINNING** - Categorical variable and "Missing becomes a category*
- **DOMAIN SPECIFIC** - Entry error, pipeline, etc.

Refine - Outlier Treatment

- What is an outlier?
- Descriptive Plots
  - Histogram
  - Box-Plot
- Measuring
  - Z-score
  - Modified Z-score > 3.5

    where modified Z-score = 0.6745 * (x - x_median) / MAD

Explore

- Single Variable Exploration
- Dual Variable Exploration
- Multi Variable Exploration

Transform

Encodings e.g.

- One Hot Encoding
- Label Encoding

Feature Transformation e.g.

- Log Transform
- Sqrt Transform

Model Creation

Types of ML Model

- Linear
- Tree-Based
- Neural Network

Choosing a Model

1. Interpretability
2. Run-time
3. Model complexity
4. Scalability

Tree Based Models

- Easy to interpret
- Little data preparation
- Scales well with data
- White-box model
- Instability – changing variables, altering sequence
- Overfitting

Ensemble Models

Bagging

— Also called bootstrap aggregation, reduces variance

— Uses decision trees and uses a model averaging approach

Random Forest

— Combines bagging idea and random selection of features.

— Similar to decision trees are constructed – but at each split, a random subset of features is used.

Model Selection

How to choose between competing model?

- Error Metric (Business Decision)
- Hyper-Parameter Tuning
- Cross-Validation

If you torture the data enough, it will confess.

— Ronald Case

Challenges

- Data Snooping
- Selection Bias
- Survivor Bias
- Omitted Variable Bias
- Black-box model Vs White-Box model
- Adherence to regulations

Machine Learning as a Service

Amit Kapoor

amitkaps.com

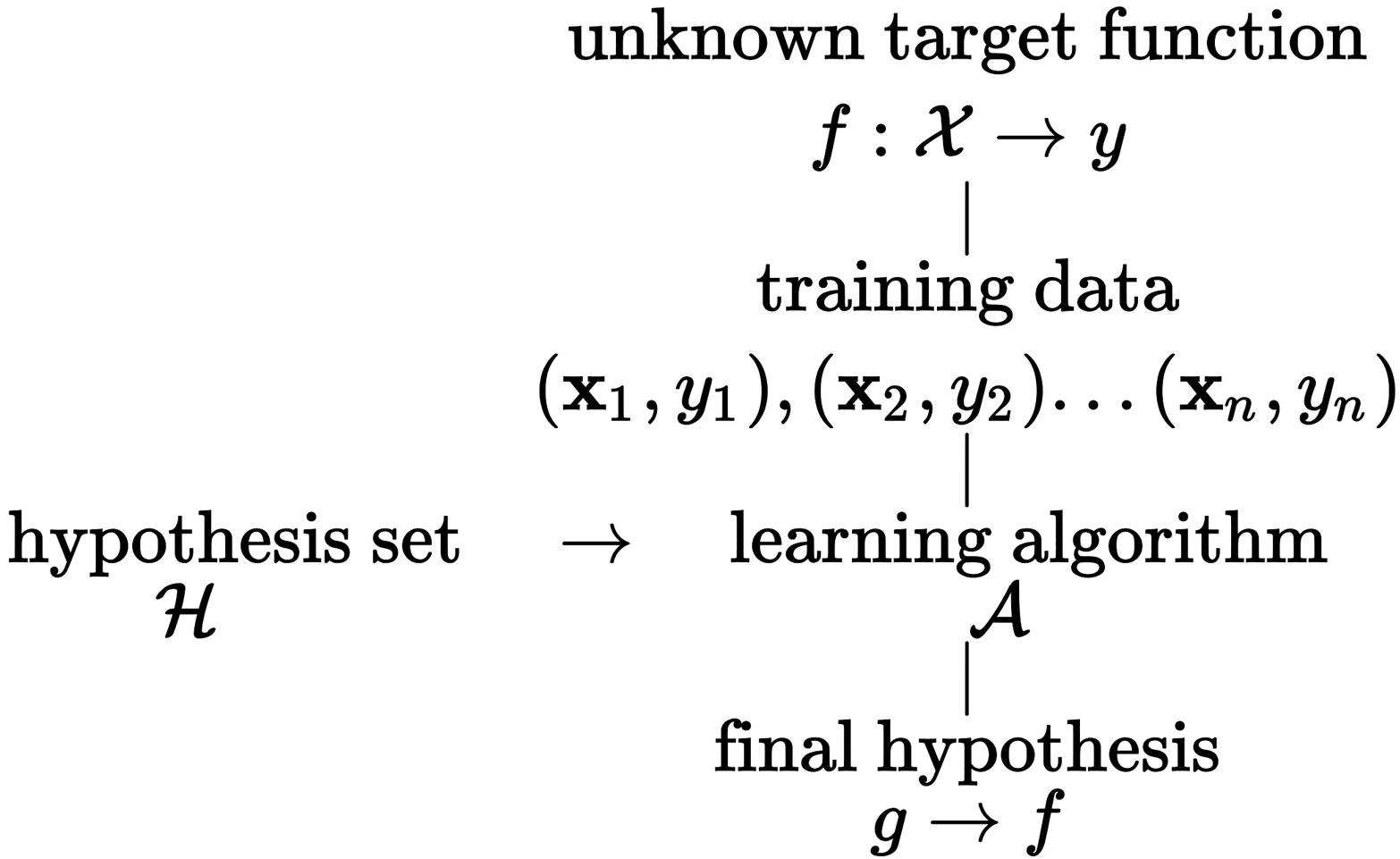
Anand Chitpothu

anandology.com

ML Theory: Formulation

- **Features** $\mathbf{x}$ *(customer application)*
- **Target** y *(loan amount)*
- **Target Function** $f : \mathcal{X} \to y$ (ideal formula)
- **Data** $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \ldots (\mathbf{x}_n, y_n)$ *(historical records)*
- **Final Hypothesis** $g : \mathcal{X} \to y$ (formula to use)
- **Hypothesis Set** $\mathcal{H}$ (all possible formulas)
- **Learning Algorithm** $\mathcal{A}$ (how to learn the formula)

ML Theory: Formulation



ML Theory: Learning Model

The Learning Model is composed of the two elements

- The Hypothesis Set: $\mathcal{H} = \{h\}$ $g \in \mathcal{H}$
- Learning Algorithm: $\mathcal{A}$

ML Theory: Formulation (Simplified)