Introduction to AI, ML and Deep Learning

PG Level AP in AI&MLOps Cohort 2

Deepak Subramani

Assistant Professor

Dept. of Computational and Data Science

Indian Institute of Science Bengaluru

The Learning Process

1
- AI/ML Fundamentals
- Ops Fundamentals
- Programming

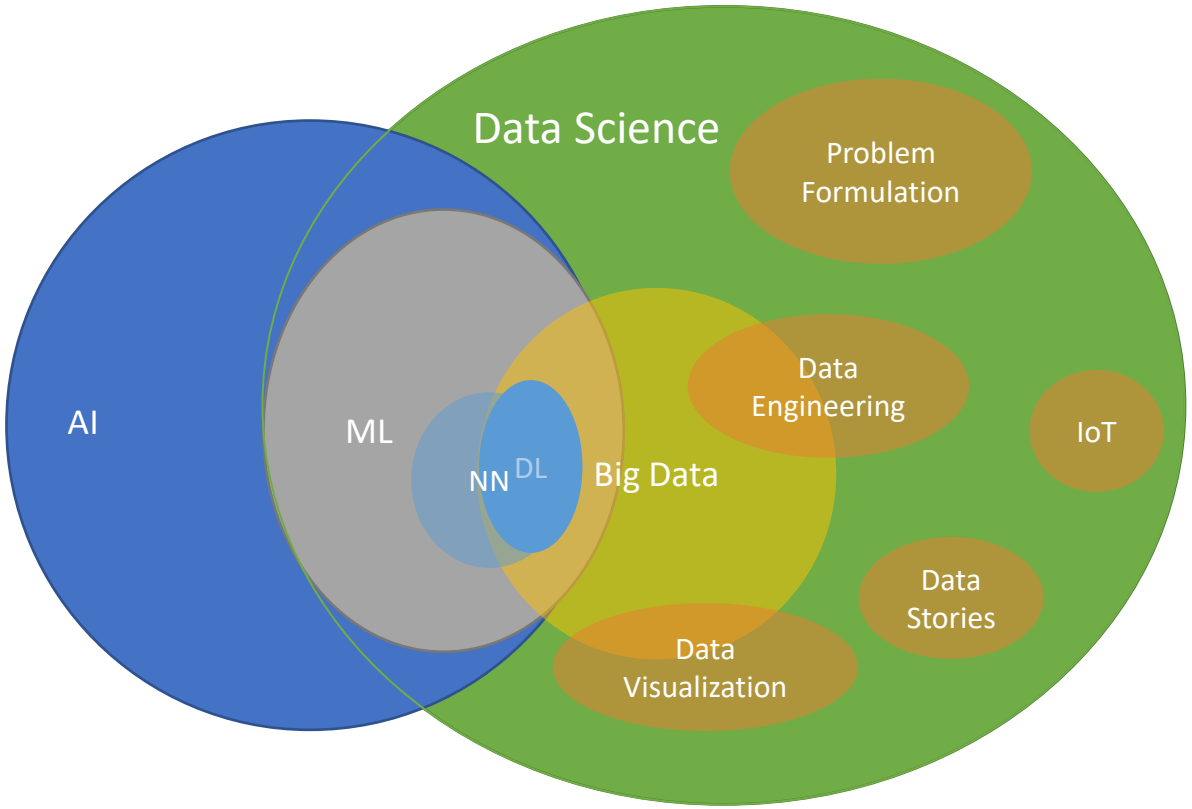
3
- Learning Tools
- Using Tools

Image Courtesy: 123rf

2
- ML Algorithms
- DL Algorithms
- Cloud MLOps
- MLOps at Scale

4
- Projects
- Portfolio

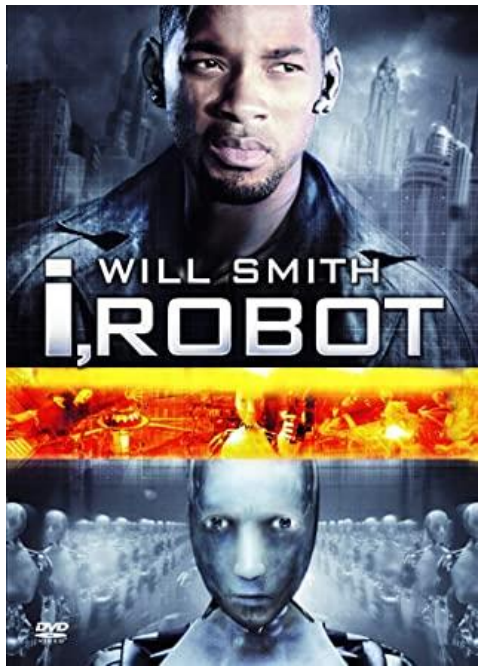
Data Science: ML/AI/DL – What is it?

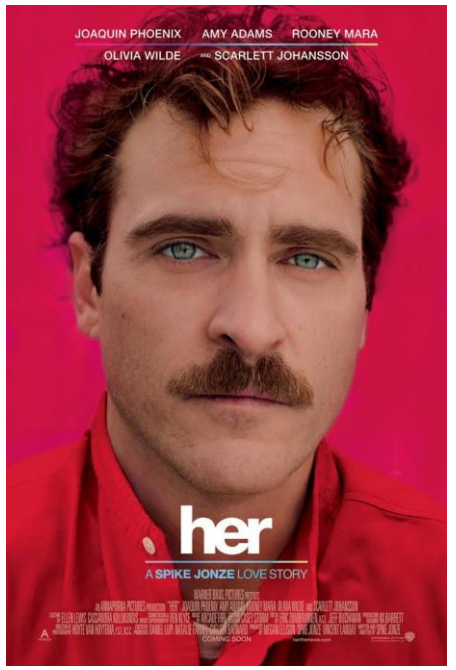


- Data Science is an umbrella term
- It is the full building that we showed
- It has foundation, pillars, floors, walls, interiors, maintenance
- One can focus on a part of the building and develop deep expertise
- But should know the breadth as well

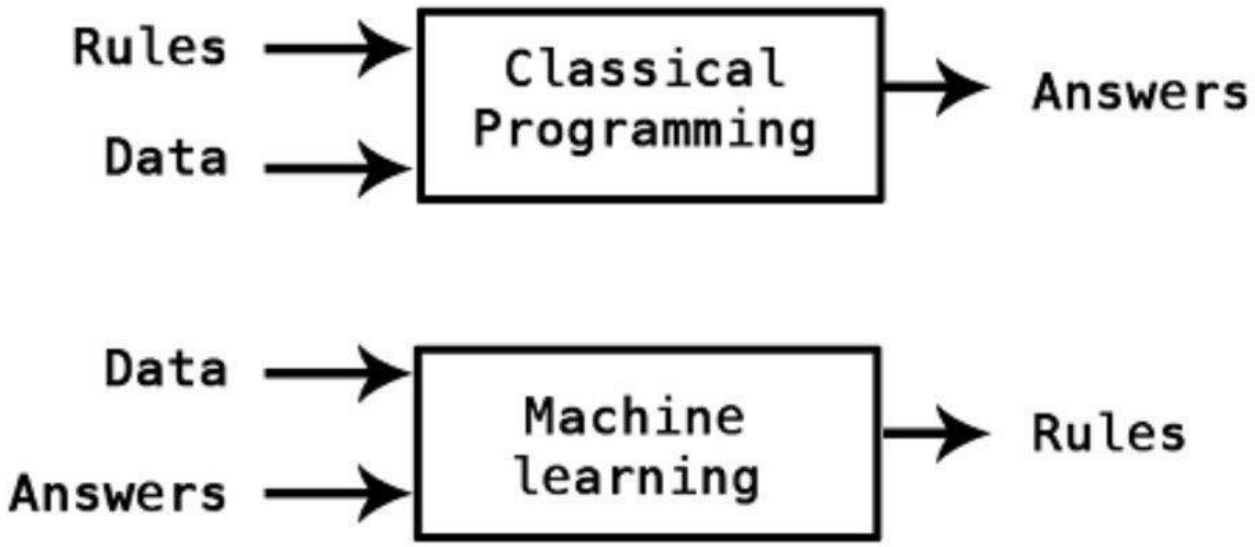
Artificial Intelligence

- AI: The ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings.
- Can be data-driven or model-driven (rule-based)
- Artificial General Intelligence is the ultimate goal in AI research





Classical Programming vs ML

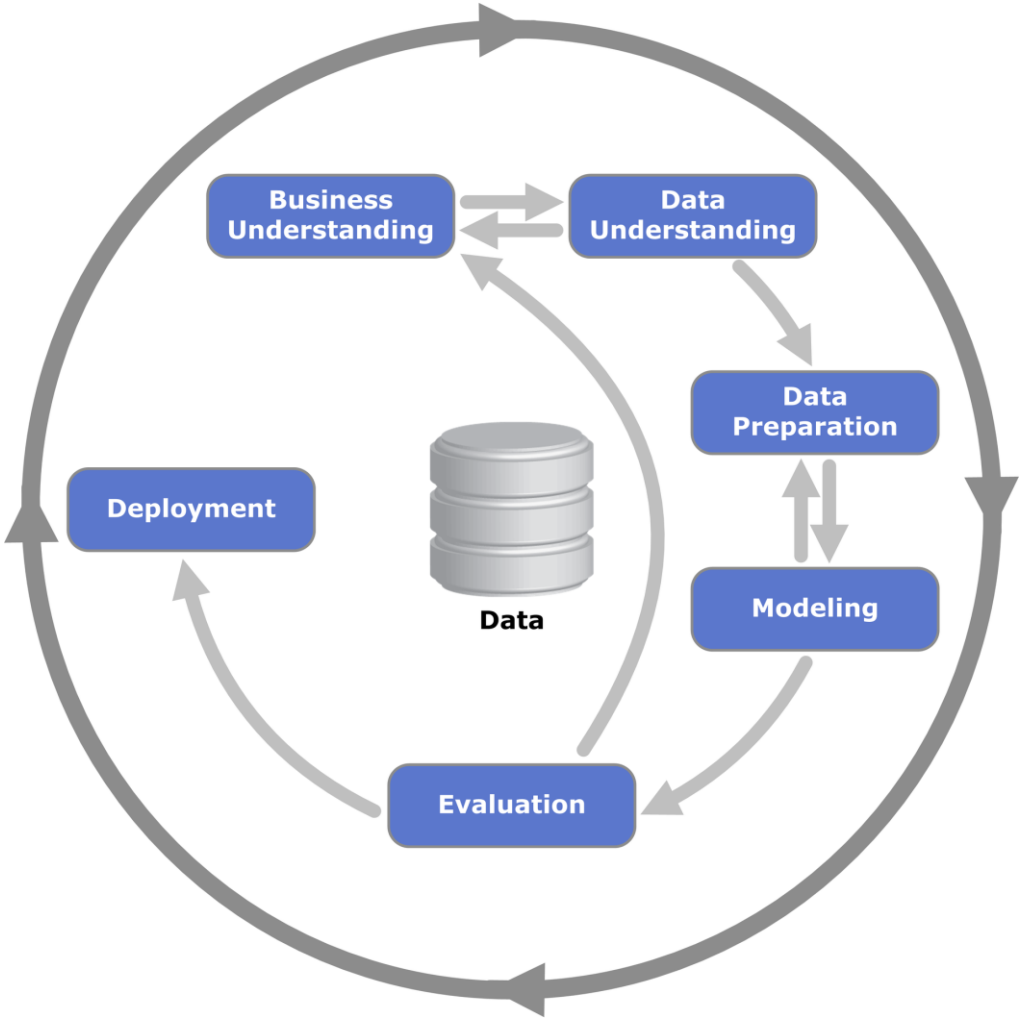


The AI/ML Workflow

1. Frame the AI problem by looking at the business need
   a. Identify subproblems (One/more of the 5 tasks a computer can do)
   b. Establish a current baseline (What is currently done?)
   c. Define success
2. Gather the data and do Data Munging/Wrangling + Baselines
   a. Explore the data
   b. Clean data and prepare for the downstream ML models
   c. Establish a data, domain and SoTA baseline
3. Explore different models, improve them through Cross Validation and perhaps new model design
4. Form an ensemble of multiple models and solutions
5. Present your solution
   a. Say a story with the data
6. Deploy

CRISP-DM

- Cross Industry Standard Process for Data Mining
- Initiative in the mid 90s by European Strategic Programme on Research in Information Technology (ESPRIT)
- The key ideas are in our 6-step process as well



Types of Data

- Tabular Data
  - Most common form
  - Arises in almost all business use cases
  - Usually number of data points x features
- Timeseries Data
  - Tabular but at different times (a logical ordering in time)
- Image Data
  - Increasing in recent years
  - Usually number of data points x height x width x sensor channels
  - Time series of image data is video data
  - Vision Tasks
- Text Data
  - Language tasks
  - Usually text corpus – Needs to be converted to number – How?
- Speech Data
  - Language tasks
  - Usually recording corpus – Signal Processing
- Knowledge Point? – Scalars, Vectors, Matrix, Tensors

Continuous vs Categorical Data

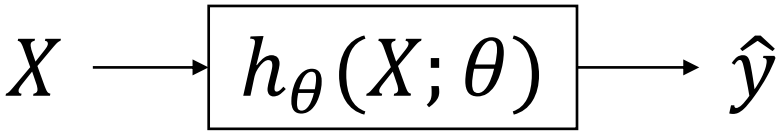
- Continuous Data – mm of rainfall tomorrow
- Categorical data – Will it rain or no?
- How to reason about categories?
- We will use the language of probability and statistics to answer these questions

ML: Mental Model

Data that can be collected

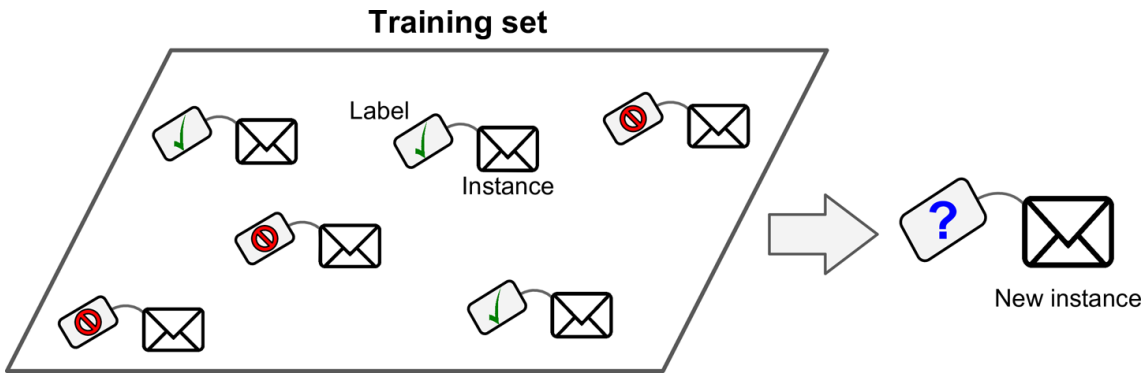
Quantity that must be predicted to make money

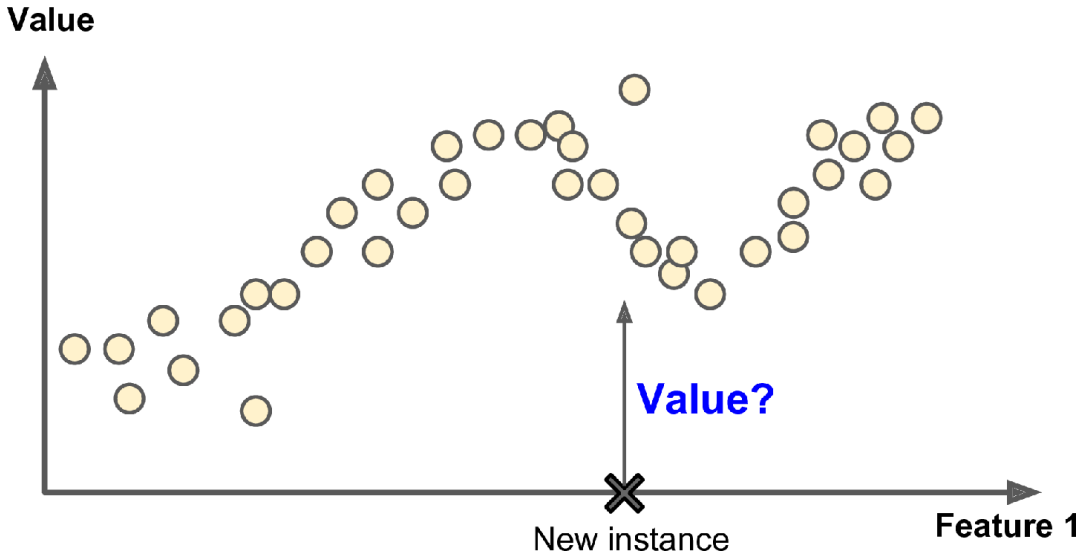
Data that can be collected

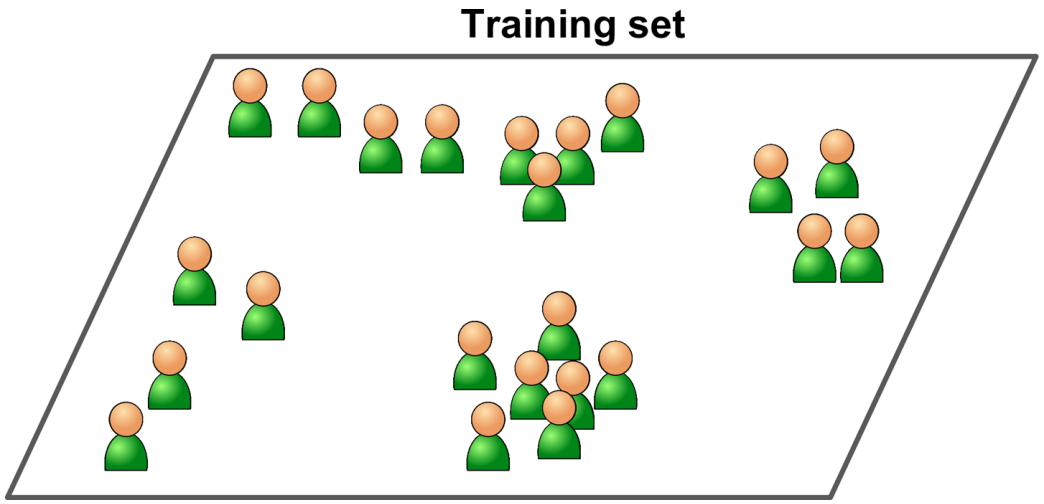


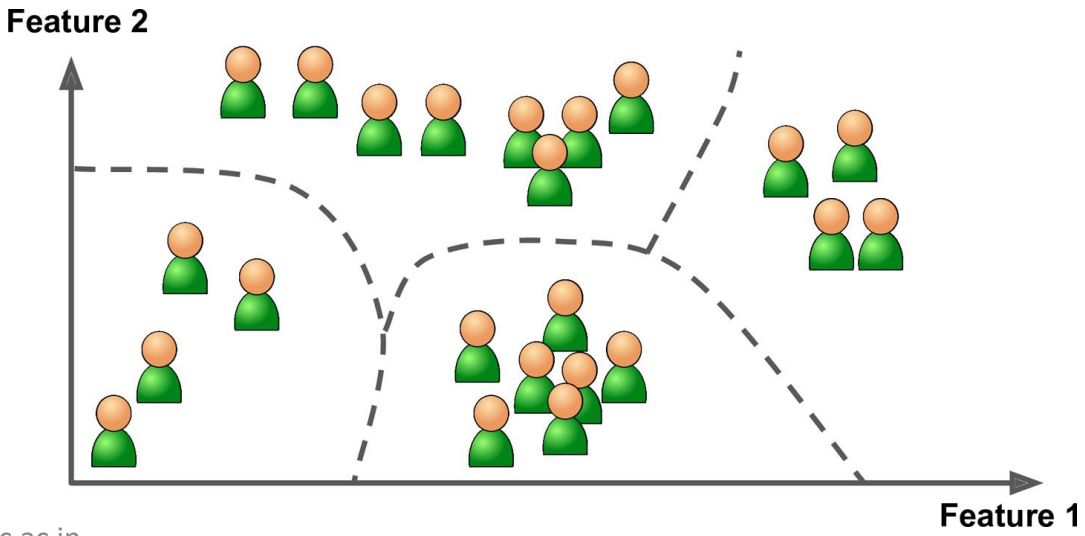
Machine's Prediction

Tasks in ML/DS/AI: Visual Introduction



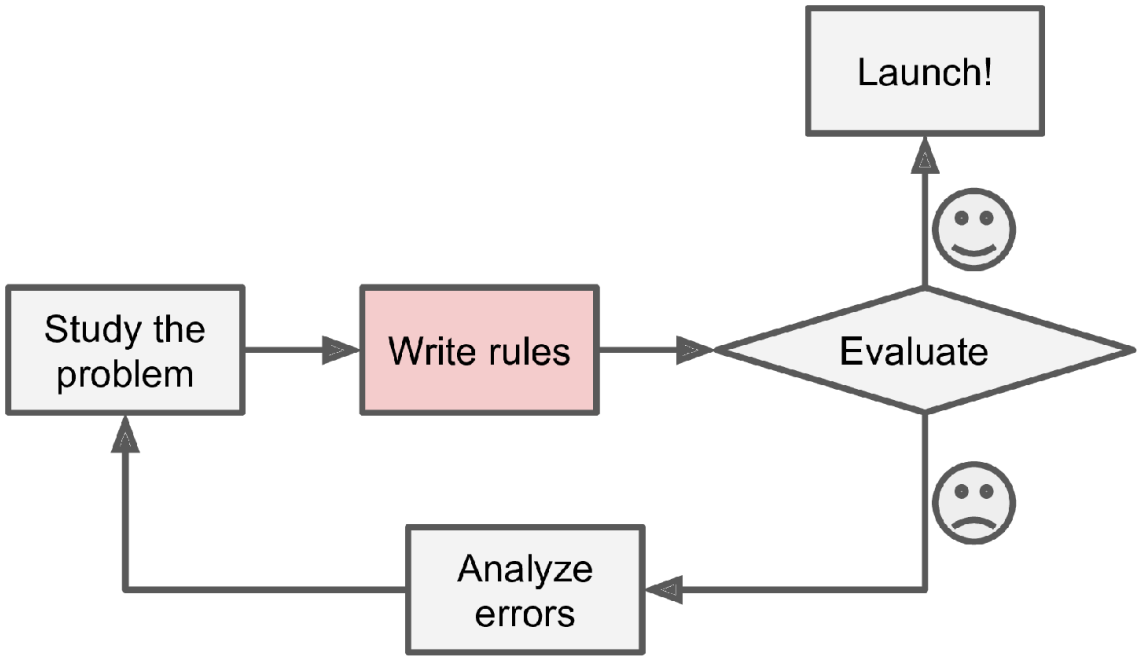


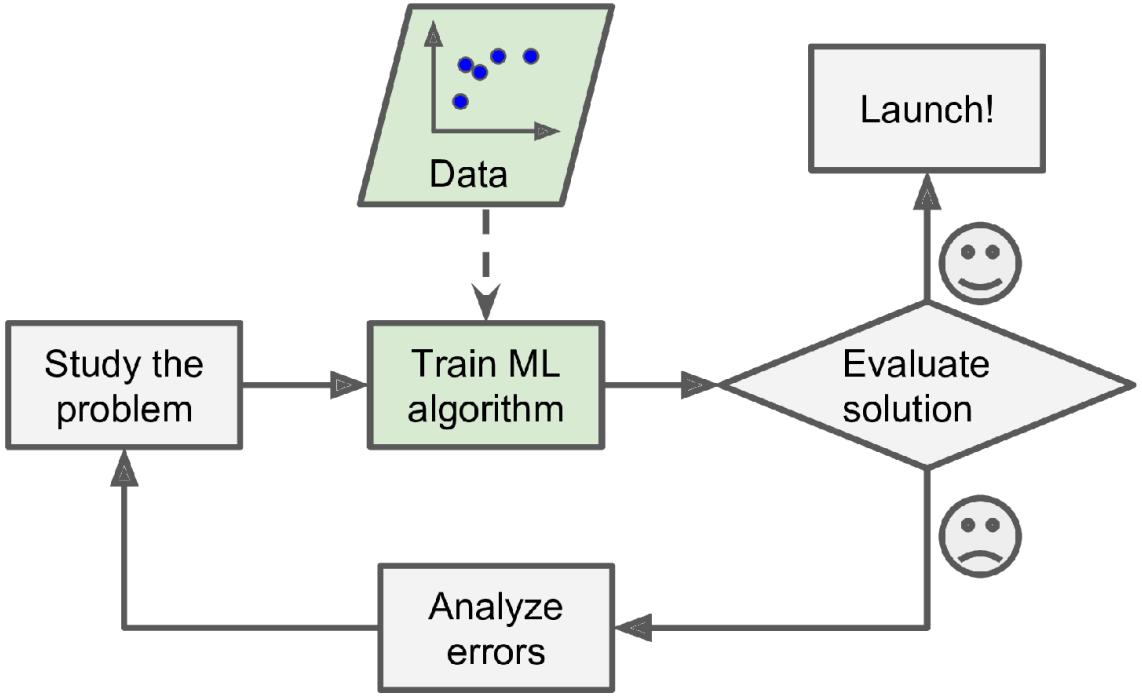




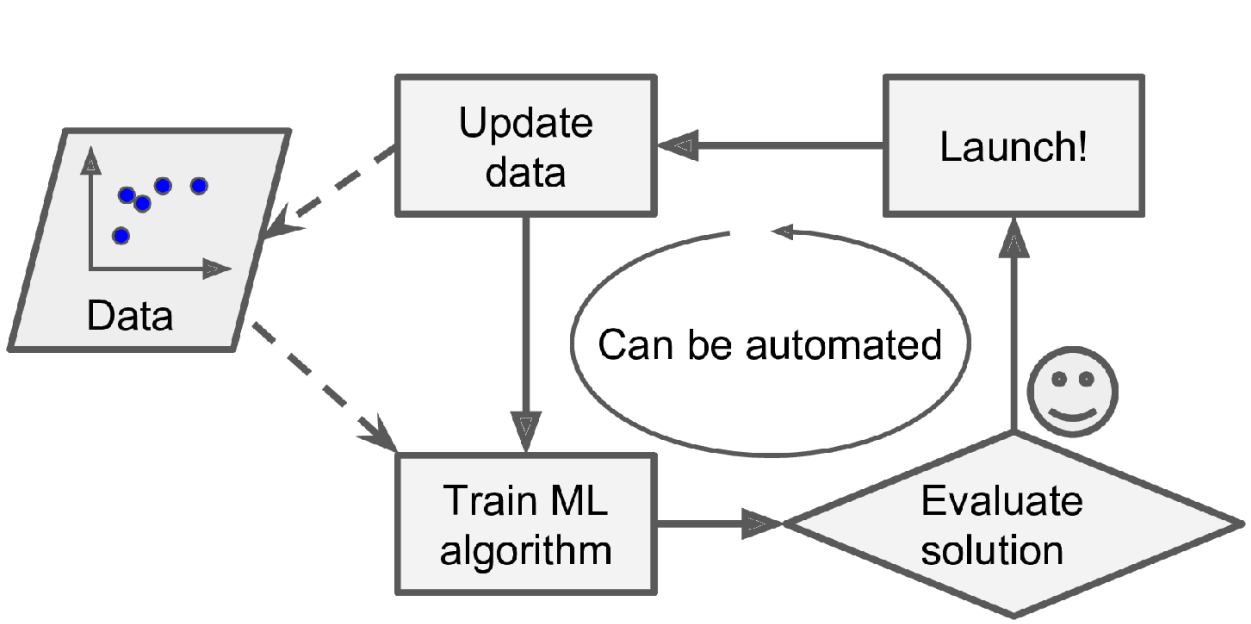
Deepak Subramani, deepakns@iisc.ac.in

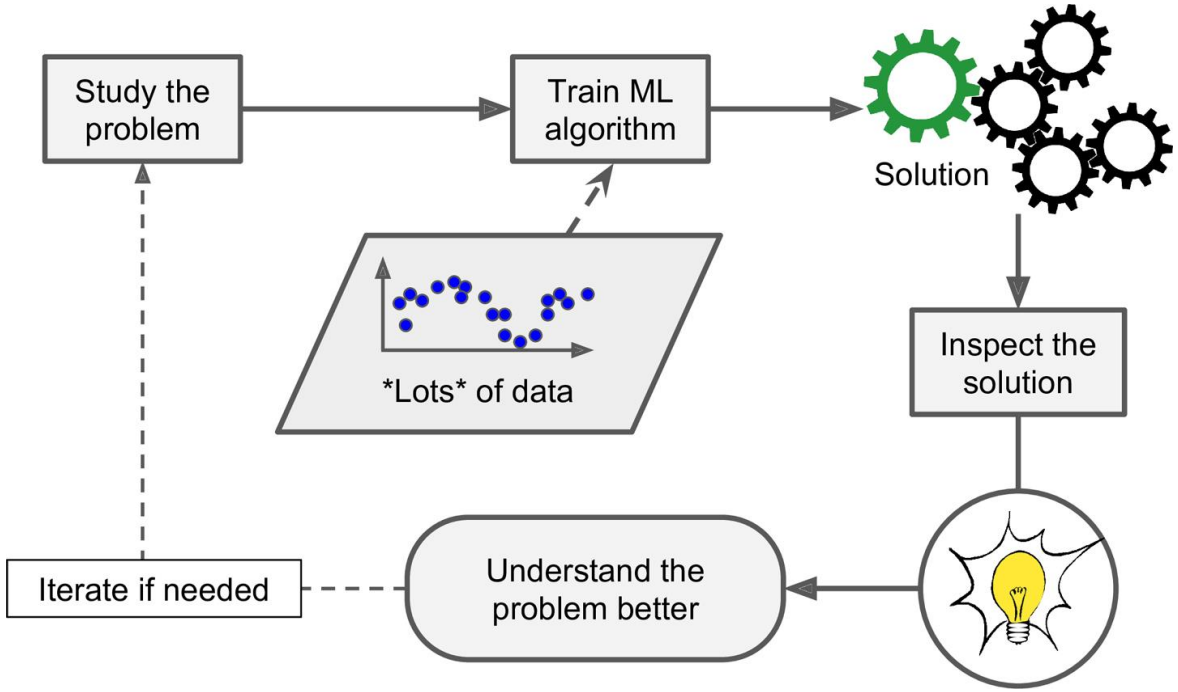
Traditional Approach vs ML





Uses of ML


Update data
Launch!
Data
Can be automated
Train ML algorithm
Evaluate solution

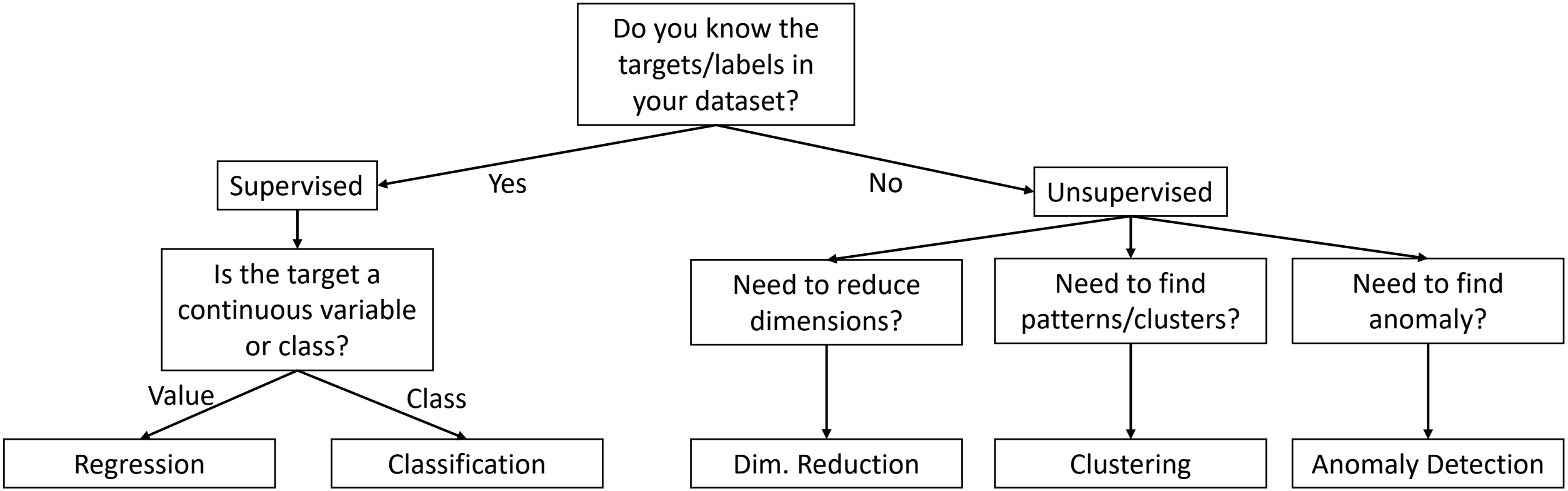

Study the problem
Train ML algorithm
Solution
Lots of data
Inspect the solution
Iterate if needed
Understand the problem better

How to train?

- We need the help of numerical optimization
- What to optimize?
  - Vector Calculus
- Stochastic Gradient Descent Algorithm
  - Stochastic – Need probability
  - Gradient – Need Calculus
  - Descent – Need Algebra/Optimization
- Tensor Calculus for Deep Neural Networks

Deepak Subramani, deepakns@iisc.ac.in

Summary of ML



Audience Poll

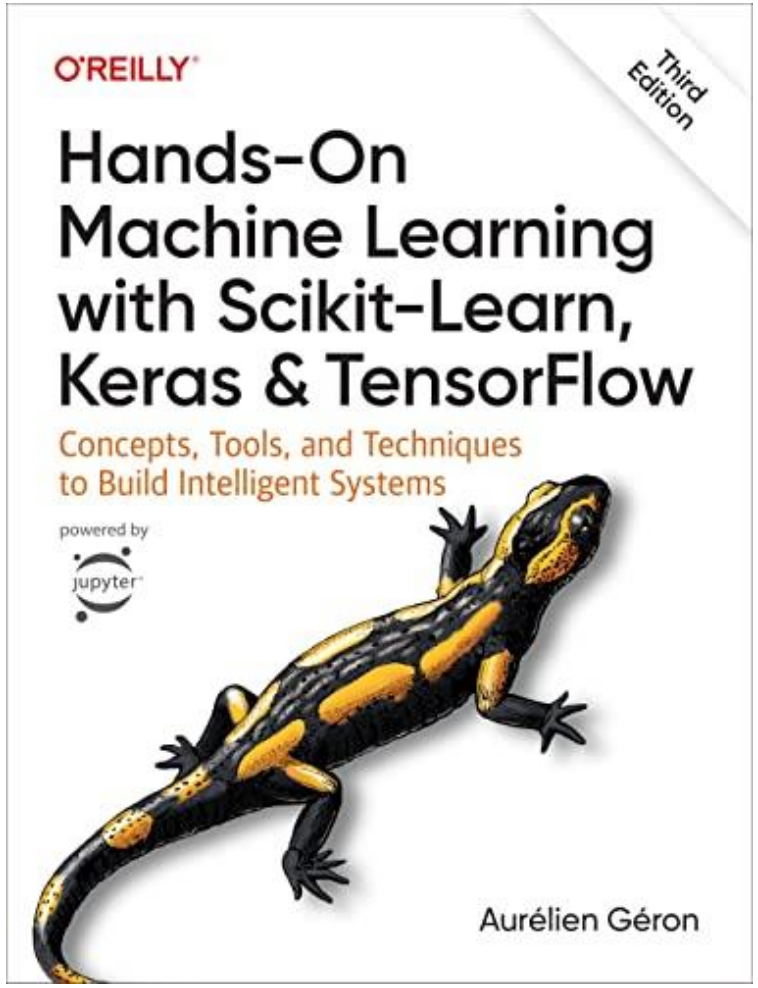
1. What are the supervised learning tasks
   - Clustering,
   - Regression,
   - Anomaly Detection,
   - Density Estimation
2. What are the unsupervised learning tasks
   - Classification,
   - Regression,
   - SVM,
   - Clustering
3. How many overall steps did we discuss was in the AI Workflow?
   - 1
   - 3
   - 6
   - 7

Sources of ML Data

- There are several open data repositories for learning ML/DL
- UCI Repository - https://archive.ics.uci.edu/ml/index.php
- Google Dataset Search - https://datasetsearch.research.google.com/

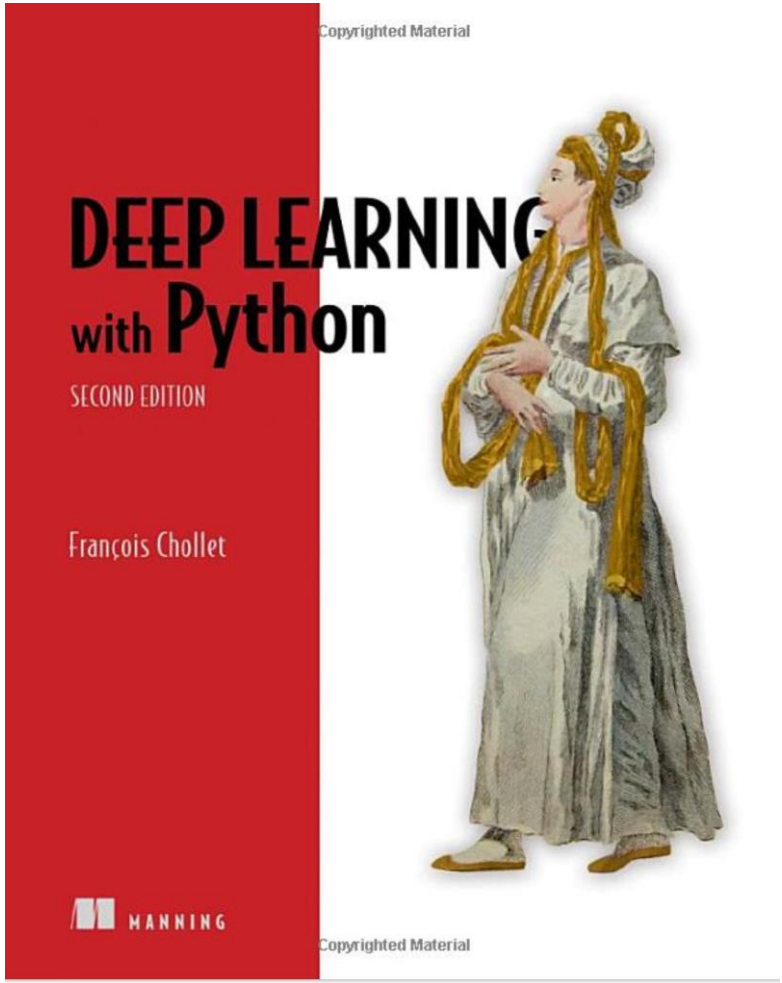
- For our illustration and as a first case study, we will use the California Housing Prices dataset from Geron Textbook Chapter 2

Text Book 1



https://www.amazon.in/Hands-Machine-Learning-Scikit-Learn-TensorFlow-ebook/dp/B0BHCFNY9Q/

Text Book 2

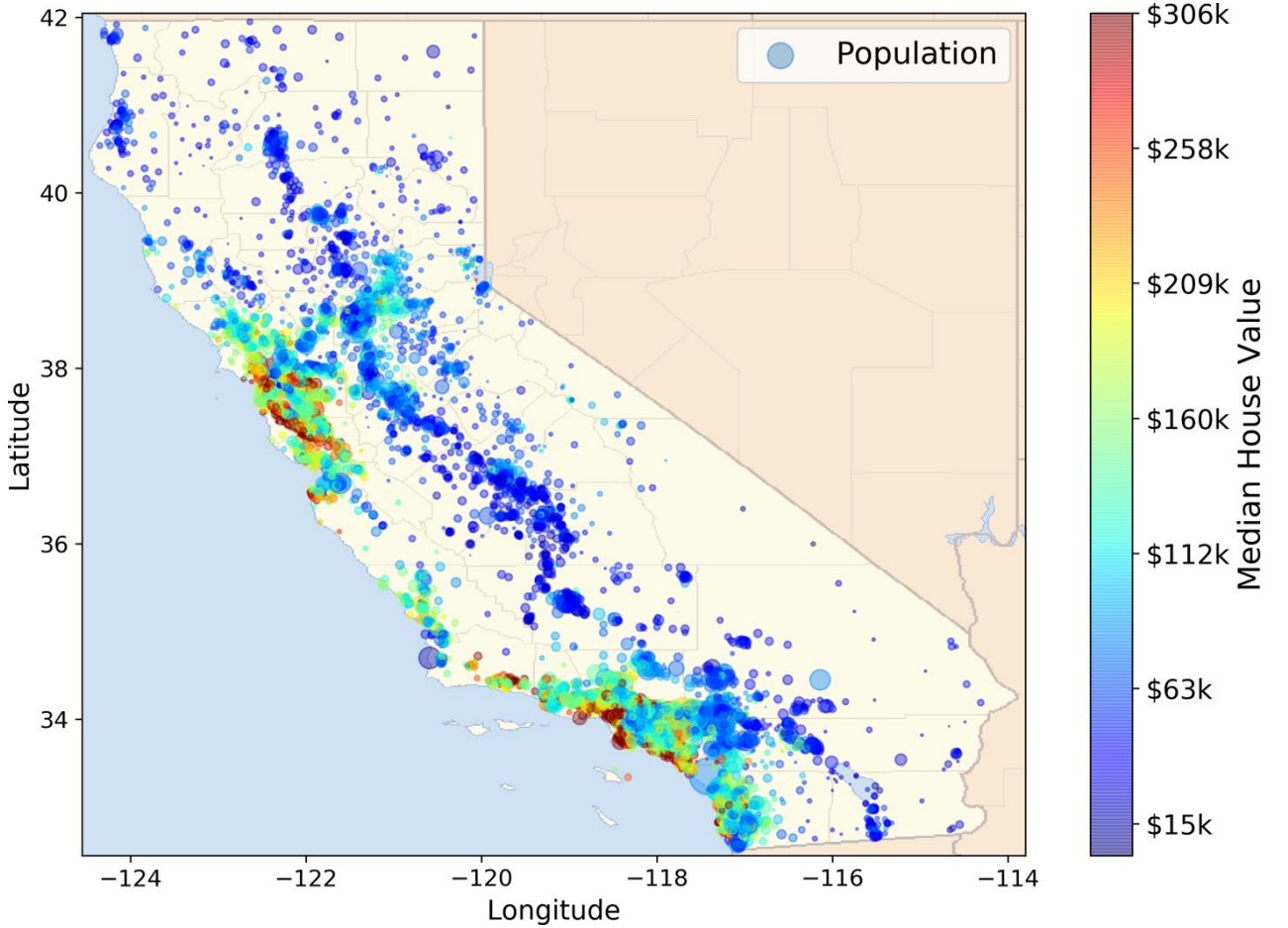


https://www.amazon.in/Learning-Python-Second-Fran%C3%A7ois-Chollet-ebook/dp/B09K81XLN1/

Housing Corporation: Case Study

See the accompanying Colab Notebook

THE HOUSING CORPORATION

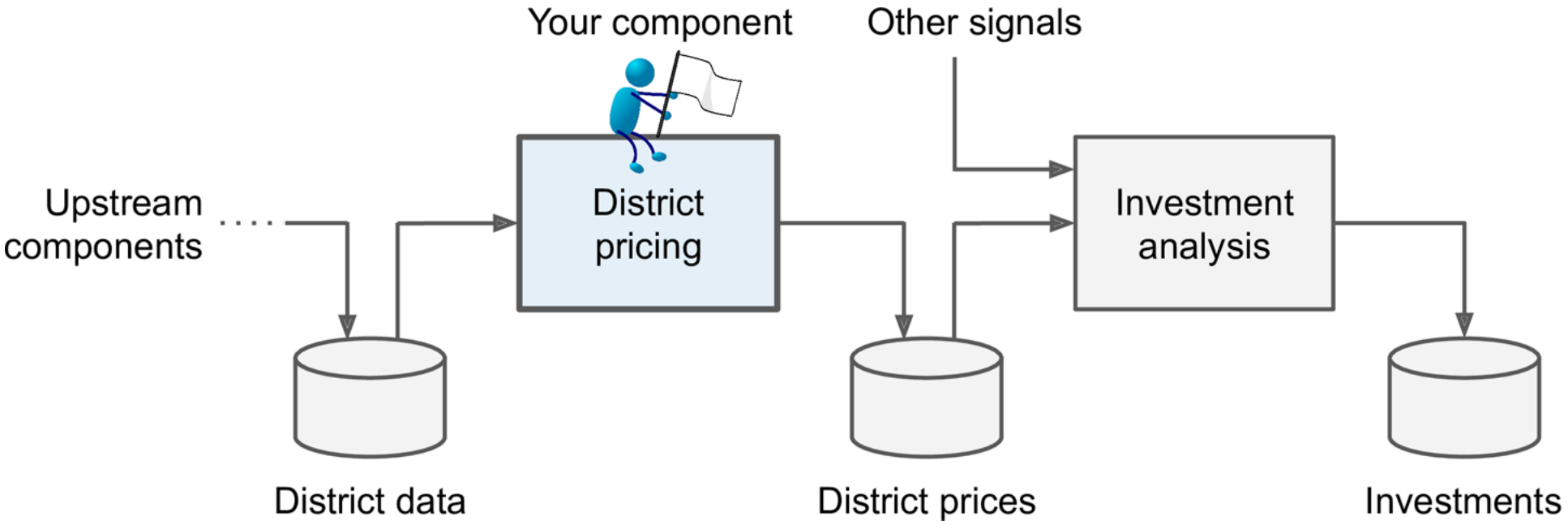


Deepak Subramani, deepakns@iisc.ac.in

Step 1: Business Need and ML Problem

- Business Need - Make real-estate investment decision for a large corporation.
- Sub-Problem - Predict the median house price in a block (smallest census group in US) given other demographic and geographic information.
- For use in - Other downstream models that need median house price in a block. For example, they may consider a greenfield project. So they need to know what a new block would be.
- Data Science Problem - Predict a value (the median house price in a block) given several features (other demographic and geographic information.)
- It is a case of regression - supervised learning problem.

ML Pipeline


Your component
Other signals
Upstream components
District pricing
Investment analysis
District data
District prices
Investments

Second Example: Predict bike-sharing counts per hour

- Business Case – Step 1
  - Plan Capacity Expansion
  - Know your customer behavior
  - Where to incentivize sharing
    - Price discounts?
    - Membership drives?
- Data Science Problem
  - Predict count (a continuous variable)
  - Given other features
- Explore data to understand features

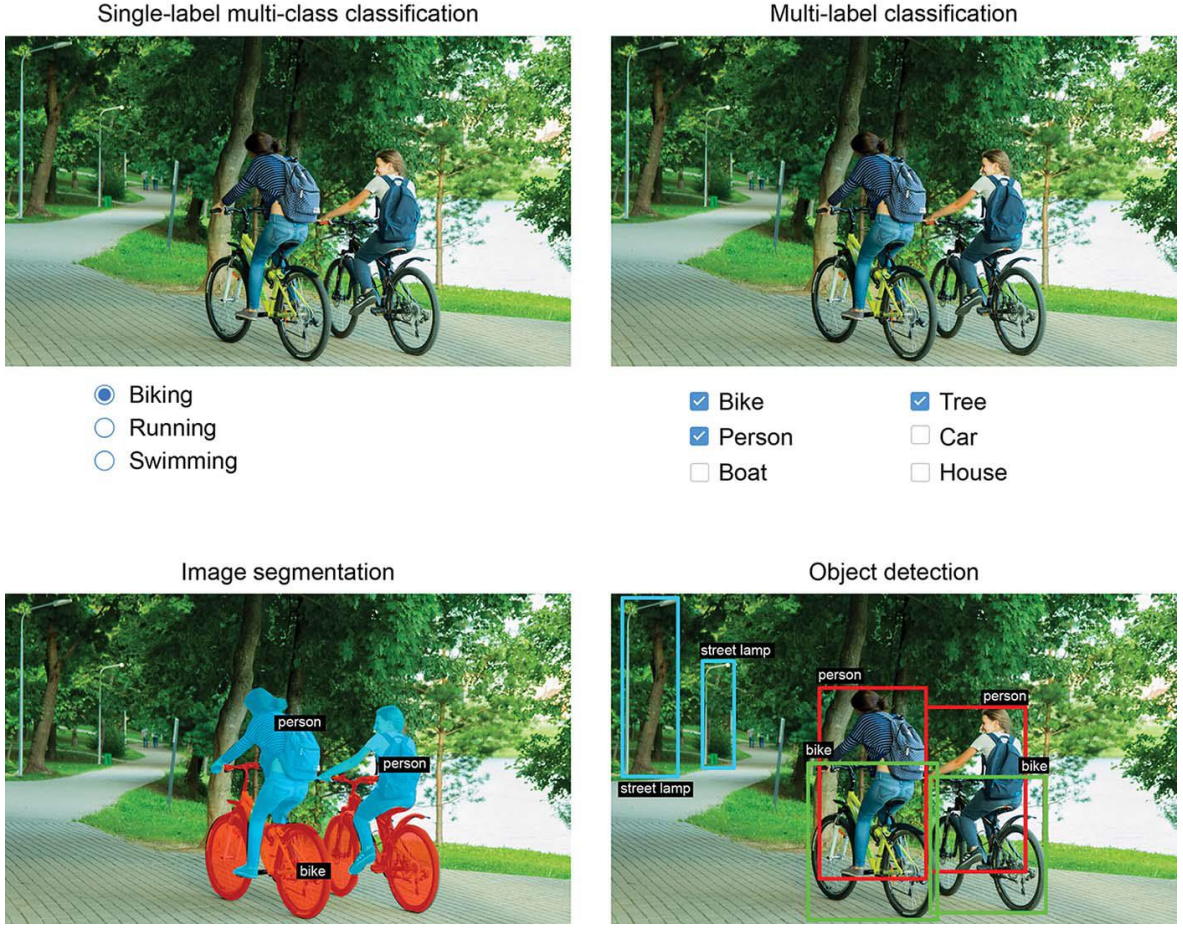
Deepak Subramani, deepakns@iisc.ac.in

Third Example: Triage a patient on presentation to ER

- Business Case – Step 1
  - ER Doctors and Nurses are over-worked
  - Especially in Covid
  - Can semi-automation with semi-skilled workforce handle triage?
- Data Science Problem
  - Predict a class (a discrete variable)
  - Given other features

Three Essential Tasks in Computer Vision

- Image Classification
  - Single Label
    - Binary
    - Multiclass
  - Multi Label
- Image Segmentation
  - Pixel wise identify the class
  - Example: Zoom background replacement
- Object Detection
  - Bounding box around objects
  - Self-driving cars, face detection in cameras



Deepak Subramani, deepakns@iisc.ac.in

AI System for Factory Workforce Attendance

Deepak Subramani, deepakns@iisc.ac.in

NLP: Major Tasks

- Modern NLP – Goal is not to understand language, but to ingest a piece of language as input and return useful quantities
  - Text Classification –
    - “What is the topic of this text?” – Topic Modelling
    - “Is this text inappropriate?” – Content Filtering
    - “Is this text, positive, neutral or negative?” – Sentiment Analysis
  - Text Regression
    - “What is the next word or character?” – Language Modeling, Sentence Completion
    - “What is “AI” in tamil?” – Machine Translation
    - “What is the crux of this paragraph?” – Text Summarization
    - Answer to “Where is the nearest hair salon?” – Question Answering

ML Workflow Step 2: Data Munging

- Needs practice
- There are some guidelines – which are abstract ideas
- In all mini-projects and course project, spend time to do the data munging
- Main software skill to learn
  - Pandas
  - Sklearn pre-processing
  - Database

Data Munging

1. Data Cleaning → In this step the primary focus is on
    1. Handling missing data
    2. Handling noisy data
    3. Detection and removal of outliers
2. Data Integration → Gather from various data sources and combine them before cleaning
3. Data Transformation → Convert the raw data into a specified format for feeding to downstream ML models.
    1. Normalization
    2. Aggregation
    3. Standardization
4. Data Reduction → remove redundancy and organize data efficiently.

Tabular Data

- In most applications, we perform data engineering operations and bring data into a tabular format.
- All types of data can be though of to be in a tabular format
- The column headers are either features or targets
- Housing data – Straightforward; Naturally Tabular
- Image Classification data – Not straightforward to think of tabular
  - Option 1: Locations where each image is located (feature column); Target is simply the true class
  - Option 2: Each pixel is a column (a feature)

Step 3: Explore different ML Models

1. Linear Regression with Regularization
   - For regression task – predicting a continuous variable
2. Logistic Regression
   - For classification tasks alone
3. K-Nearest Neighbours
   - Instance-Based Method
4. Naïve Bayes
   - Simplest Bayesian Network Model
5. Decision Trees, Random Forests
6. XGBoost, CatBoost
7. Neural Networks

Steps 4 to 6

- Step 4: Fine-tune (Ensemble models)
  - Voting mechanisms
  - Cross-Validation
- Step 5:
  - Present your solution
  - Key skill to convince your boss/team/clients
- Step 6:
  - Deploy – Internal tool? Cloud-based? UI/UX for Clients?
  - Monitor for drift
  - Monitor for change in data/assumptions

Explain your work to stakeholders and set expectations

- Success and customer trust are about consistently meeting or exceeding expectations
- The actual model is only half the picture; the level of expectation about system performance matters a lot
- Non-specialists expect AI to punch above its weight
  - They expect the system to “understand” and meet or exceed capability of a human doing the task
- Clearly setting the expectation is important
- Some guidelines
  - Don’t talk in easily mis-understood terminology – Accuracy is 98%
  - Show examples of what misclassification looks like
  - Understand if customer cares about False Positive or False Negative more
  - Discuss key parameters – the probability above which a fraud has to be detected
  - Explain how many cases on average we expect the system to be falsely labelled as positive [False Positive, False Negative, Explain in simple language]