End-to-end machine learning

Saptarshi Purkayastha, Ph.D.

https://plhi.lab.Indianapolis.iu.edu

- Associate Professor, Health Informatics, Data Science at Indiana University Indianapolis
  - Director, Health Informatics, School of informatics & Computing
  - Director of Undergraduate Education & Research, BHI
- Affiliated Scientist, CBMI, Regenstrief Institute
- Faculty Associate, STEM Education Innovation and Research Institute (SEIRI)

Atika Paddo
PhD Student

- Research funded by:

BILL & MELINDA GATES foundation

NIH National Institutes of Health

NSF

NVIDIA

Google Cloud Platform

World Health Organization

AMD

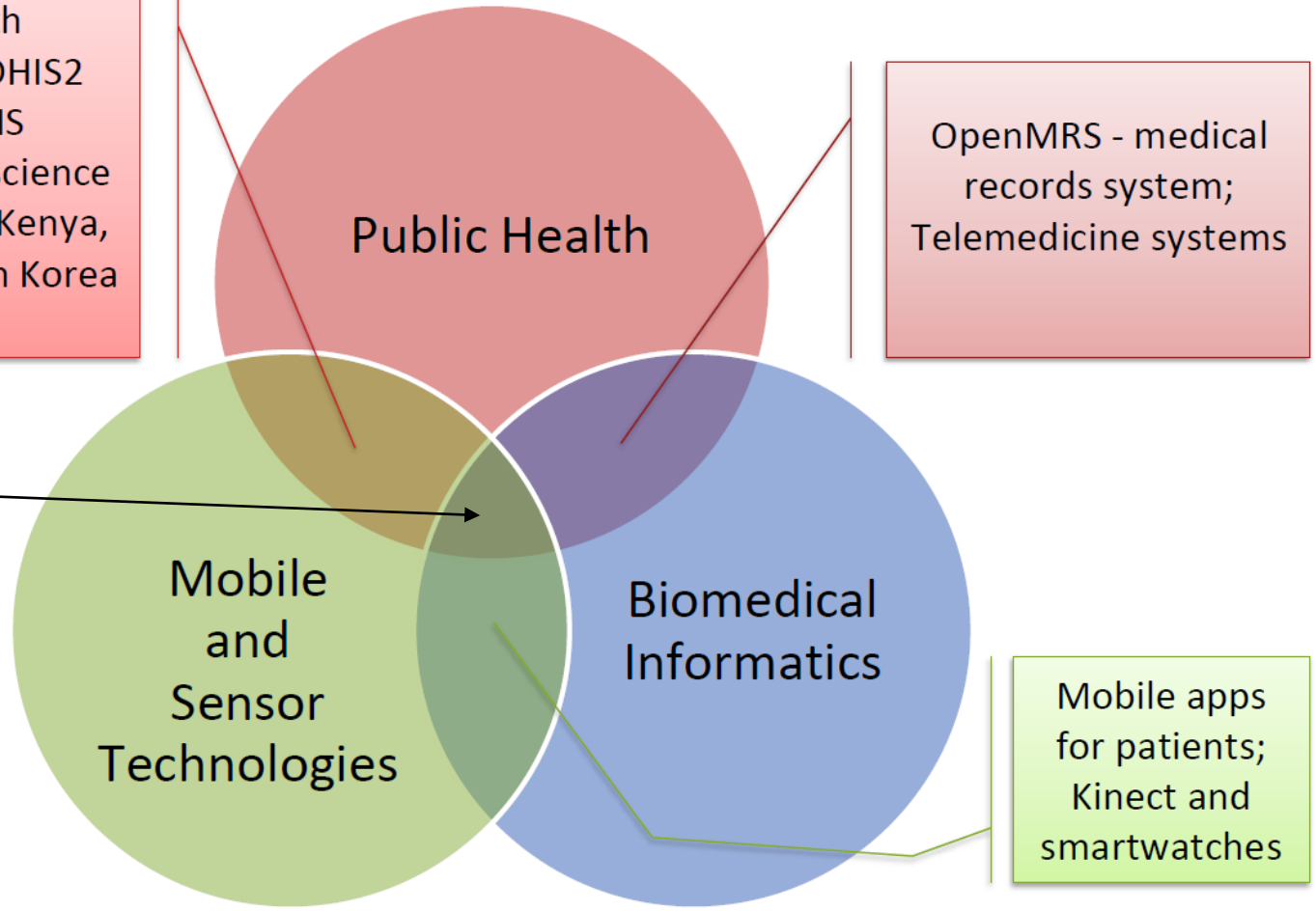
FDA U.S. FOOD & DRUG ADMINISTRATION

Work centered around building teaching-learning infrastructure

"Information is Care"

A world-wide recognized fact that clinicians provide best treatment to patients, when medical history of the patient is available

Global Health
Open-source - DHIS2
Country HMIS
implementation science
in India, Malawi, Kenya,
Bangladesh, North Korea



OpenMRS - medical records system; Telemedicine systems

Mobile apps for patients; Kinect and smartwatches

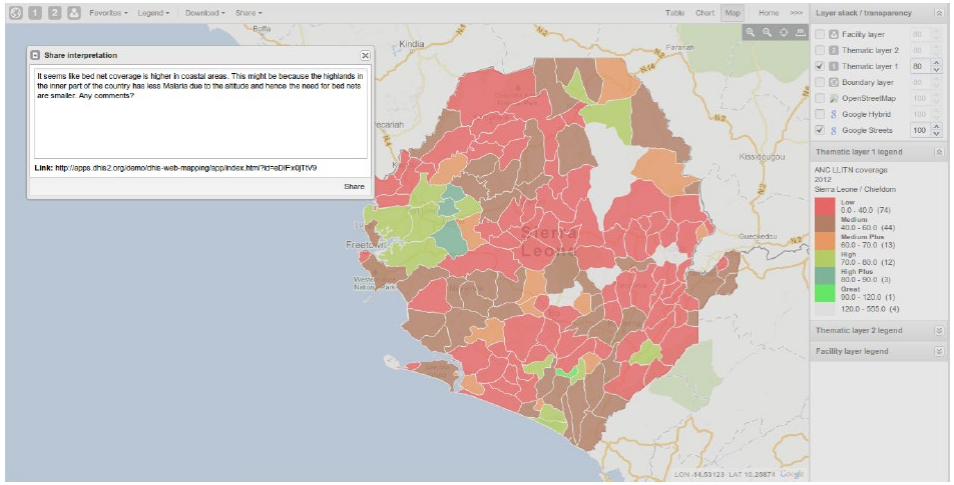
Teaching-Learning Infrastructure for humans and machines

Health Information Infrastructure

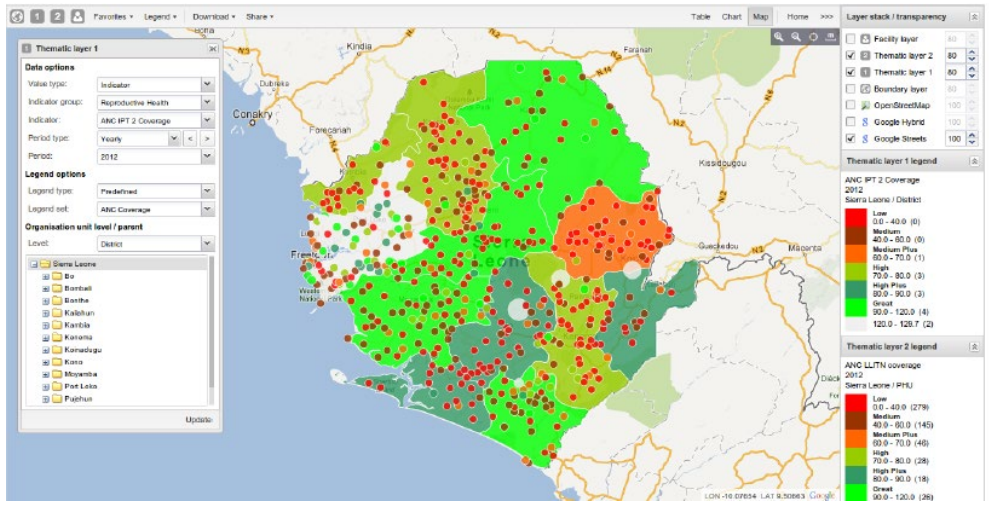
- National eHealth Architectures: With WHO SEARO, I developed the first country-wide telemedicine system in Bhutan (2015) based on web-telephony and OpenMRS EHR. This was followed by DHIS2 optimization for patient records in Bangladesh (2015), training and implementation of national HMIS in North Korea (2015). The most recent is the development of the eHealth Architecture for Myanmar (2017).

- Cybersecurity in the cloud for privacy preservation: Development and integration of novel cybersecurity technologies in the NSF first public cloud called JetStream. $600k NSF-funded project (co-PI)

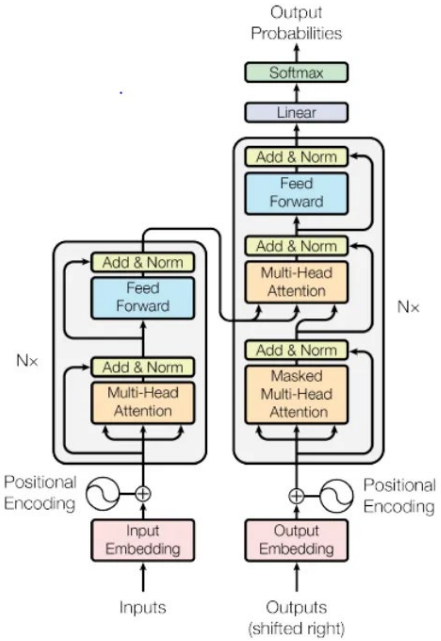
- Wound care platform: Using DHIS2, we've built a platform for Indiana Center for Regenerative Medicine and Engineering for program monitoring and evaluation about amputations after non-recoverable wounds.



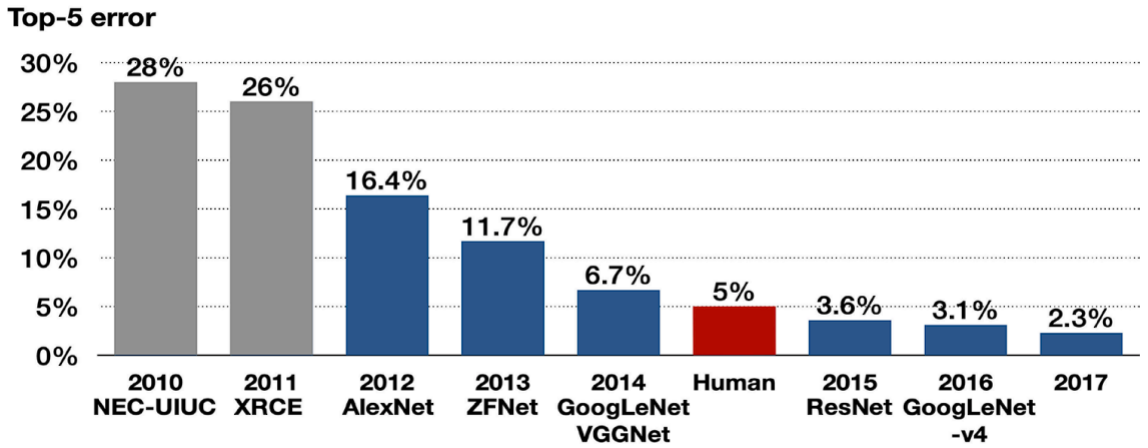
And building a social network to distribute the analysis of the data



Displaying disease outbreak in Sierra Leone

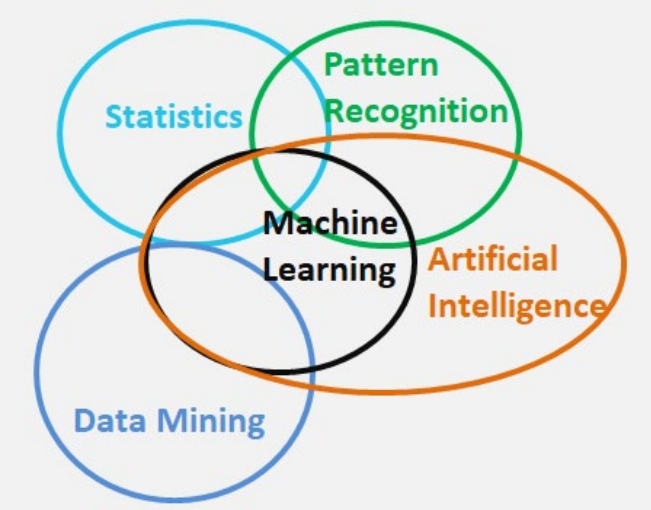


Why need Artificial Intelligence?

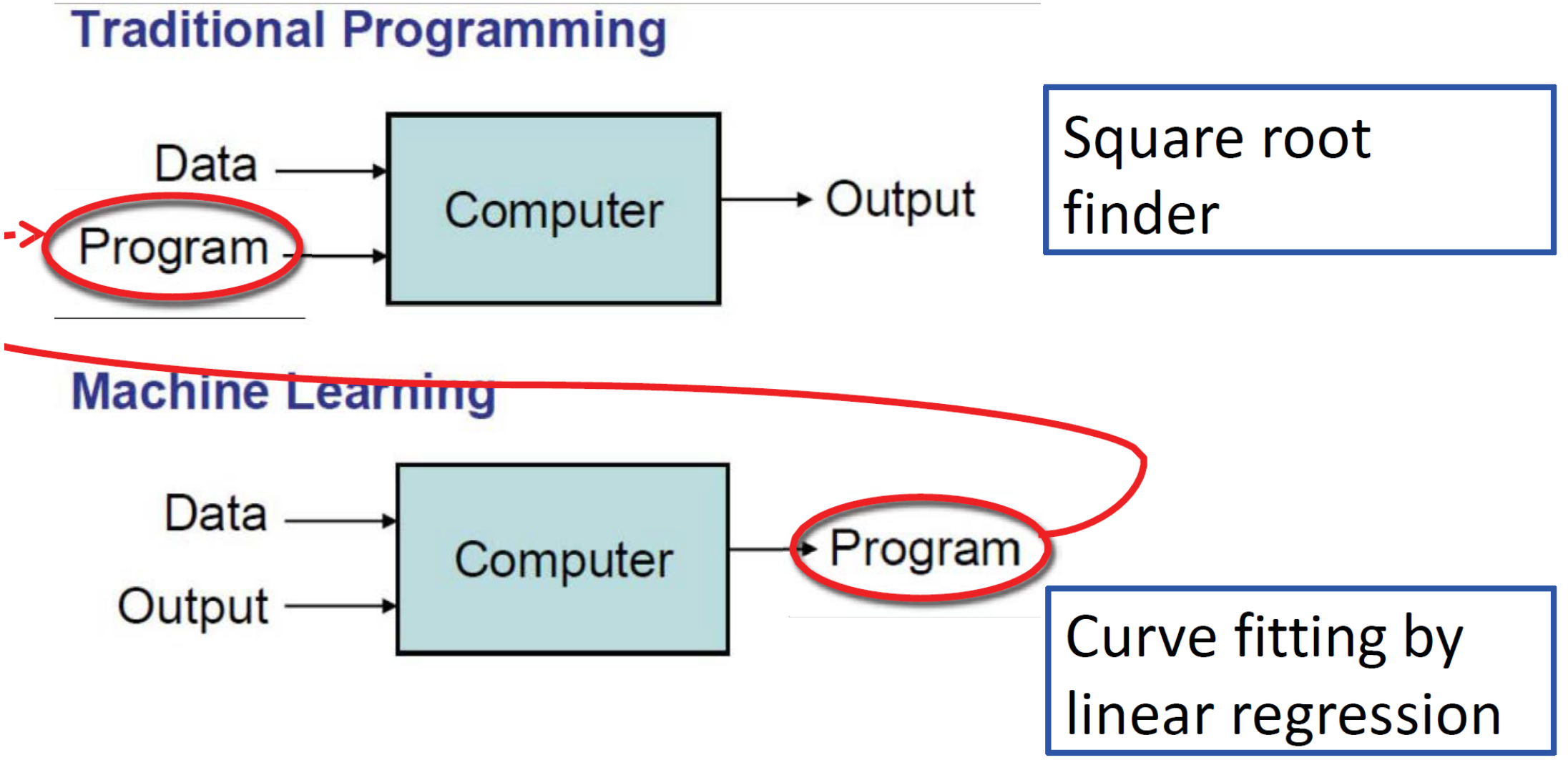


What is machine learning?

- All useful programs "learn" something.
- Suppose you wrote a program to calculate the square root of any number with some acceptable error, say 0.001. Is the program “learning” something?
- Machine learning as a field started with Arthur Samuel's 1959 paper which said, "Field of study that gives computers the ability to learn without being explicitly programmed.“
  - Computer pioneer who wrote first self-learning program, which played checkers - learned from "experience“.
  - Invented alpha-beta pruning - widely used in decision tree searching



Traditional vs machine learning programs



How are things learned?

- Memorization
  - Accumulation of individual facts

Declarative knowledge

- Limited by
  - Time to observe facts
  - Memory to store facts
- Generalization
  - Deduce new facts from old facts
  - Limited by accuracy of deduction process
    - Essentially a predictive activity
    - Assumes that the past predicts the future

Imperative knowledge

- Interested in extending to programs that can infer useful information from **implicit** patterns in data

Basic paradigm of machine learning

- Observe set of examples
  - *Training data*

Spatial deviations relative to mass displacement of spring

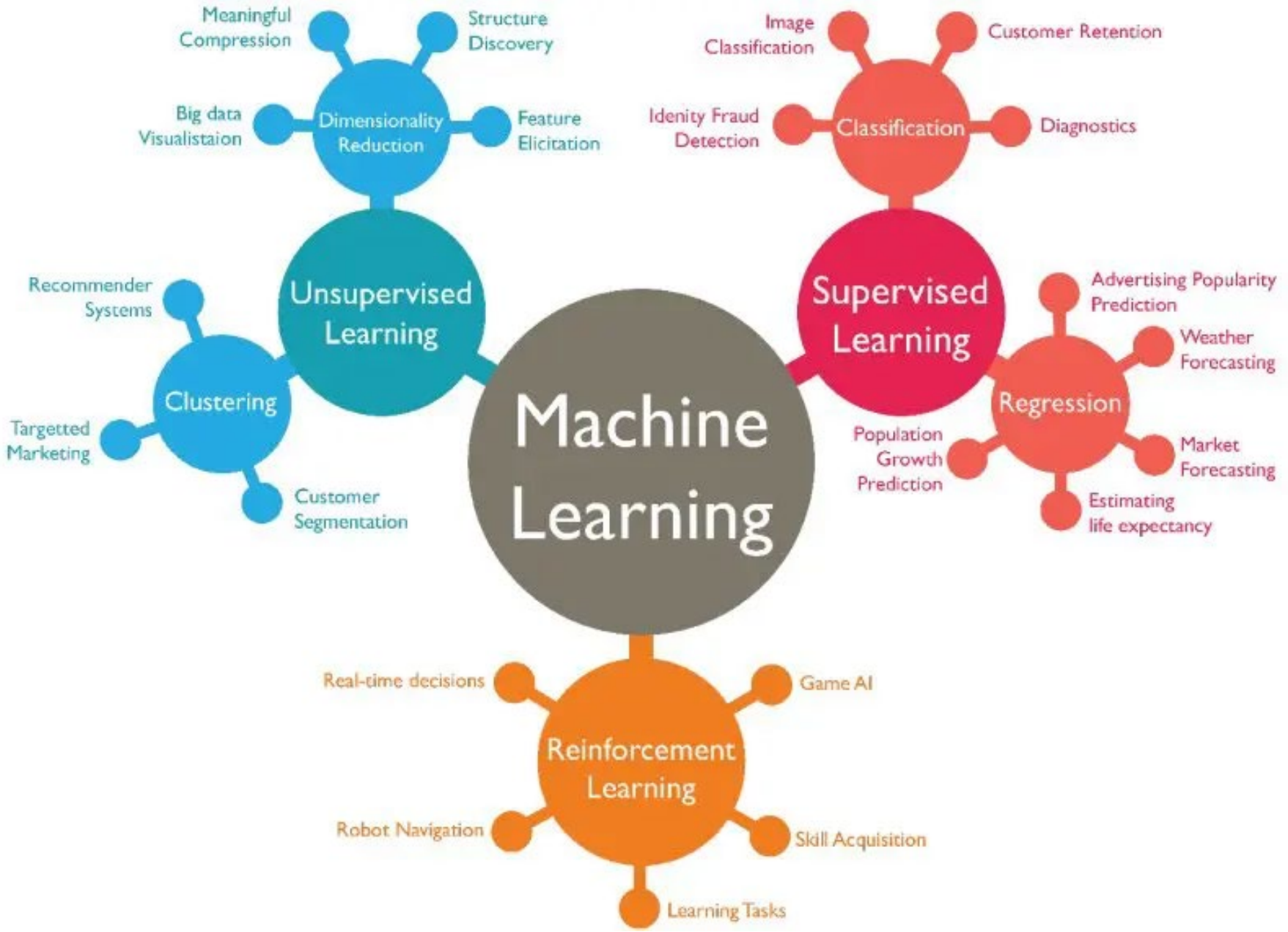
- Infer something about process that generated the data
  - Types of model
  - Create a model
  - *Parameters* to a model

Fit polynomial curve using linear regression

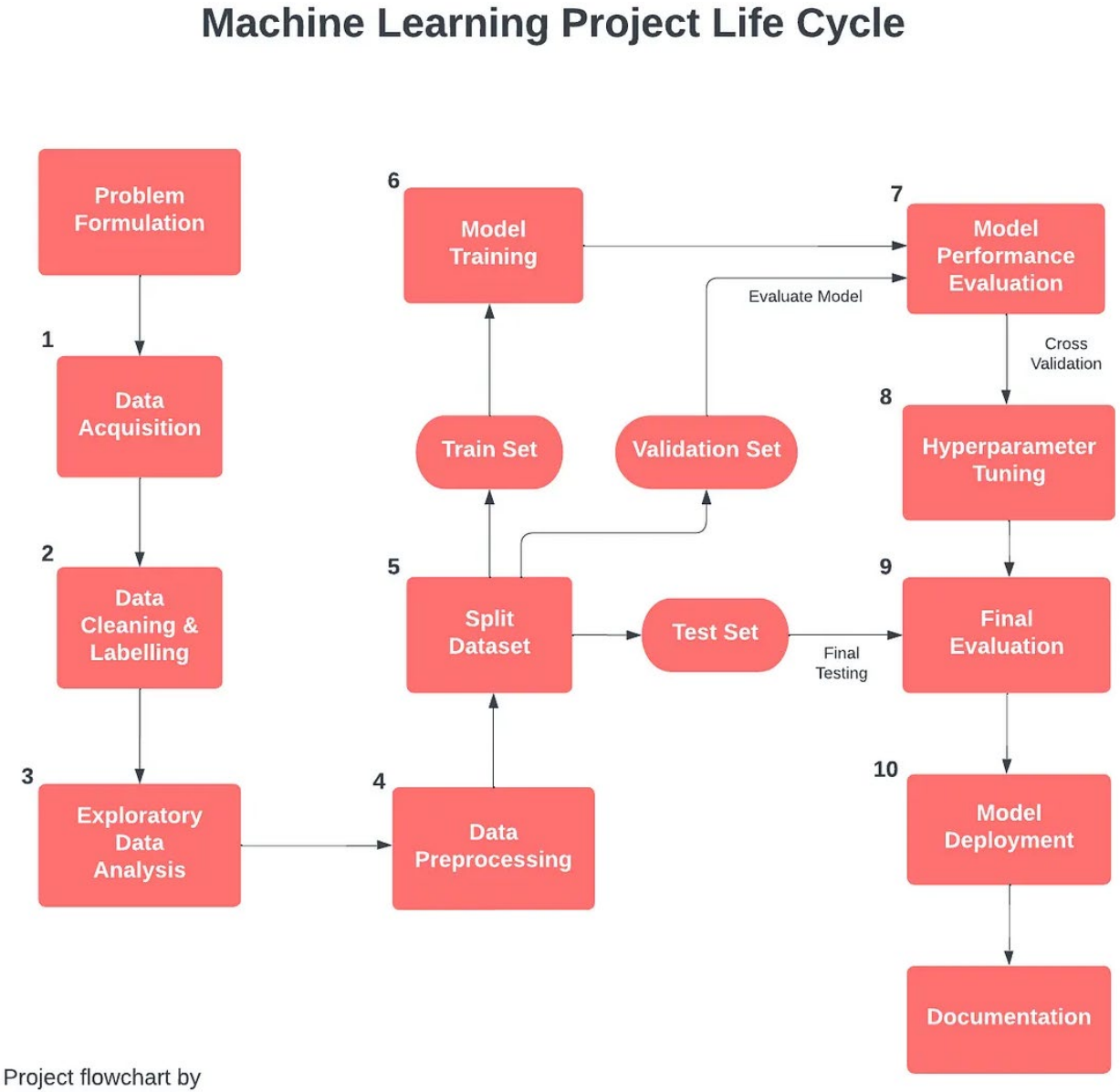
- Use inference to make predictions about previously unseen data
  - *Test data*

Predict displacements for other weights

- Two paradigms on how to learn:
  - **Supervised**: given a set of feature/label pairs, find a rule that predicts the label associated with a previously unseen input.
  - **Unsupervised**: given a set of feature vectors {without labels}, group them into "natural clusters" {or create labels for the groups}


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

What is the end-to-end process for machine learning?



ML Project flowchart by
Ihsanul Haque Asif
North South University

Emory ICU dataset

Problem formulation

- The length of stay (LOS) in MIMIC-IV is typically measured in days, not hours.
- Some important features often used for LOS prediction in MIMIC-IV include:

  1. Patient demographics: Age, Gender, Ethnicity
  2. Admission details: Admission type (emergency, elective, etc.), Admission location, Insurance type
  3. Clinical data: Primary diagnosis, Comorbidities, Severity scores (e.g. SOFA, SAPS II)
  4. Vital signs: Heart rate, Blood pressure, Respiratory rate, Temperature, Oxygen saturation
  5. Laboratory values: Complete blood count, Basic metabolic panel, Liver function tests
  6. Interventions: Mechanical ventilation, Vasopressor use, Dialysis
  7. Medication data
  8. Previous hospital visits/admissions

- The specific set of most important features can vary depending on the particular prediction model and patient population. Some studies have found factors like age, admission type, primary diagnosis, and certain lab values to be particularly predictive.

Data acquisition - Where is my data?

What is parquet?

Tables in parquet format

- Apache Parquet is a free and open-source column-oriented data storage format in the Apache Hadoop ecosystem.
- Pandas library native

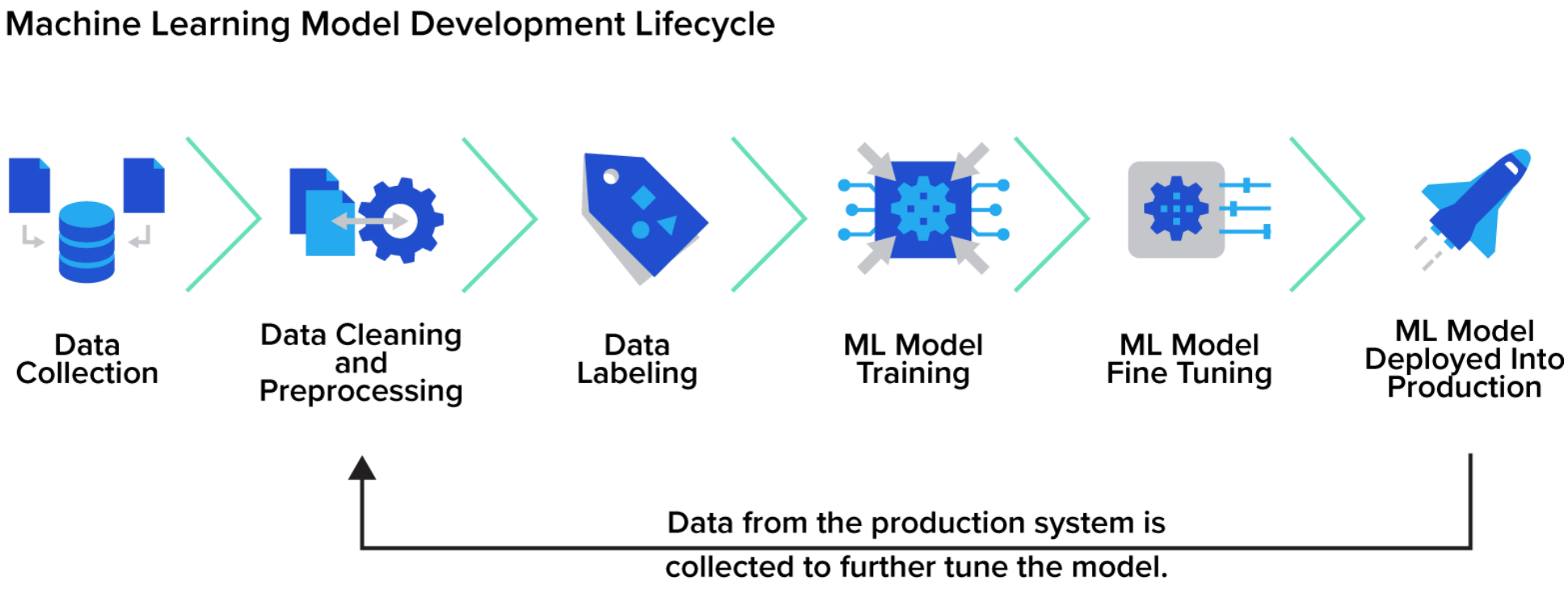
	Column 1	Column 2	Column 3	Column 4	Column 5
	Product	Customer	Country	Date	Sales Amount
Row Group 1	Ball	John Doe	USA	2023-01-01	100
	T-Shirt	John Doe	USA	2023-01-02	200
Row Group 2	Socks	Maria Adams	UK	2023-01-01	300
	Socks	John Doe	USA	2023-01-03	100
Row Group 3	T-Shirt	Maria Adams	UK	2023-01-02	500
	Socks	John Doe	USA	2023-01-05	200

The engine will not scan these records

Spark Format Showdown		File Format		
		CSV	JSON	Parquet
Attribute	Columnar	No	No	Yes
	Compressable	Yes	Yes	Yes
	Splittable	Yes*	Yes**	Yes
	Human Readable	Yes	Yes	No
	Nestable	No	Yes	Yes
	Complex Data Structures	No	Yes	Yes
	Default Schema: Named columns	Manual	Automatic (full read)	Automatic (instant)
	Default Schema: Data Types	Manual (full read)	Automatic (full read)	Automatic (instant)

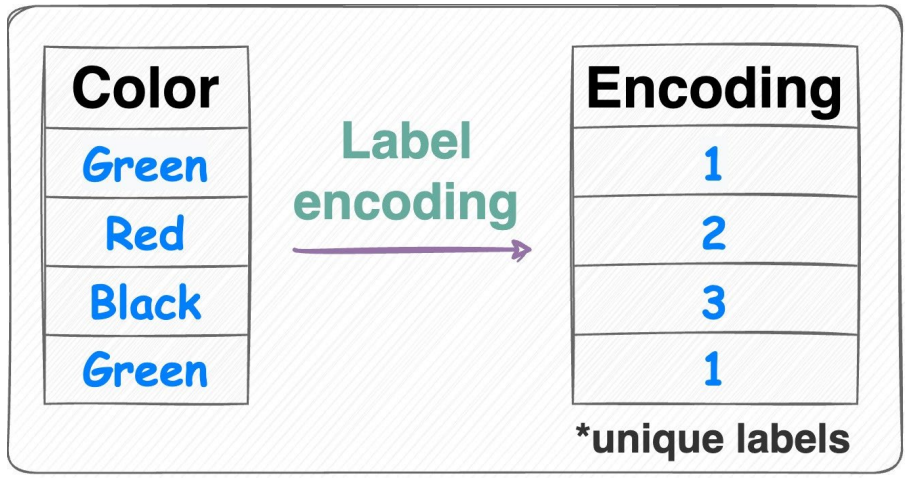
Data cleaning and labeling

- Remove duplicates.
- Data imputation – remove where an estimate is unrealistic, fill where distribution remains the same.
- Merging data



Exploratory data analysis (EDA)

- Distribution of the variables. Different types of visualization.
- Reviewing the Data types.
- Applying specific pre-processing based on continuous or discrete variables.
- Label Encoder vs One Hot Encoder



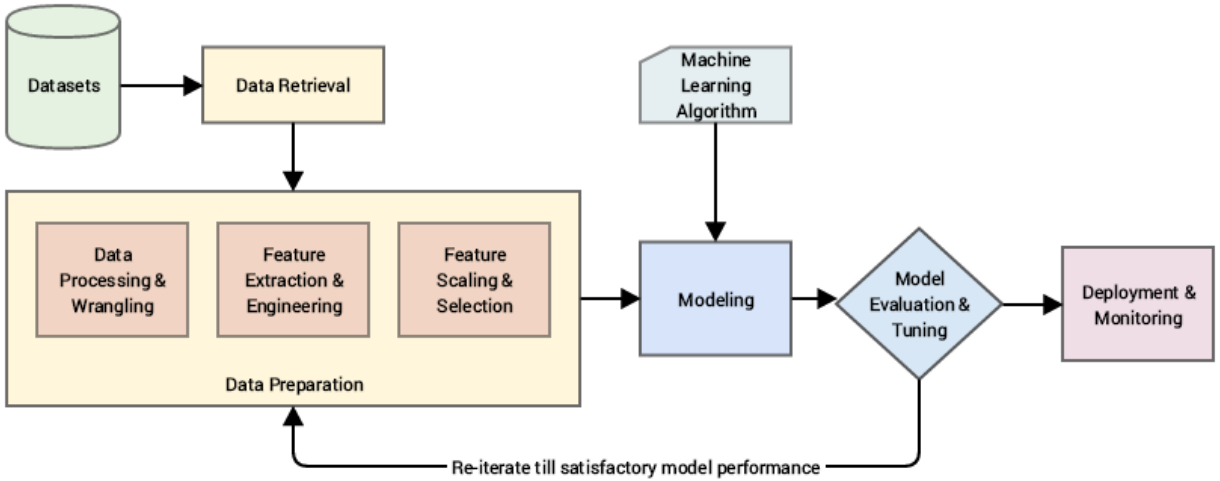
id	color
1	red
2	blue
3	green
4	blue

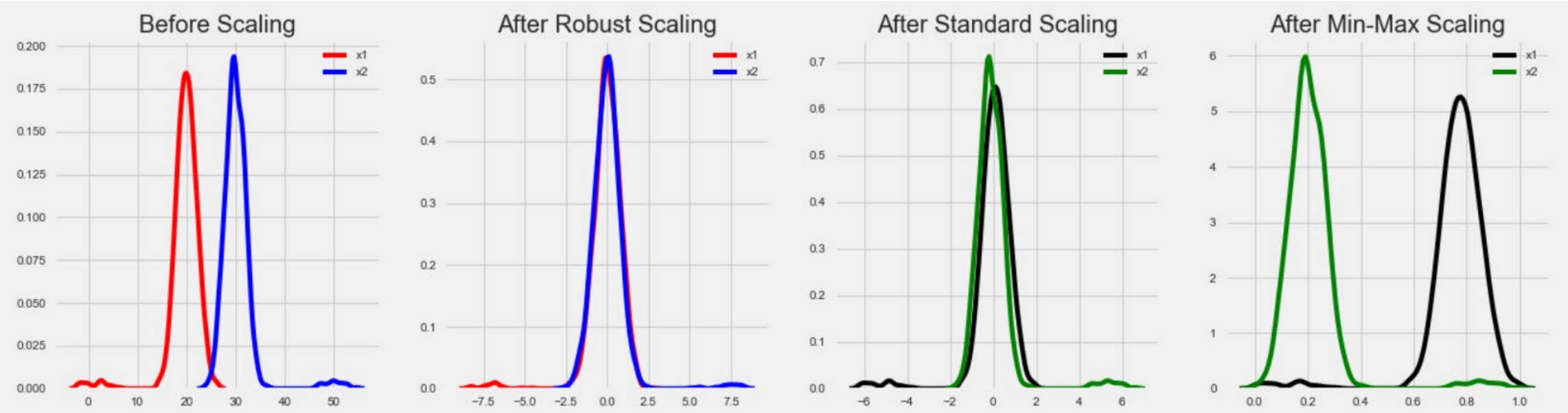
One Hot Encoding

id	color_red	color_blue	color_green
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

Feature engineering

- An iterative process with domain knowledge – knowing what to change and associate.



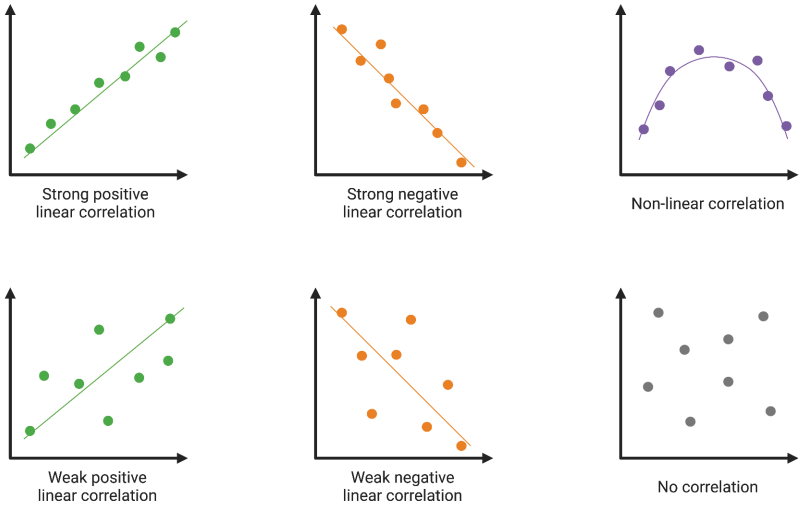


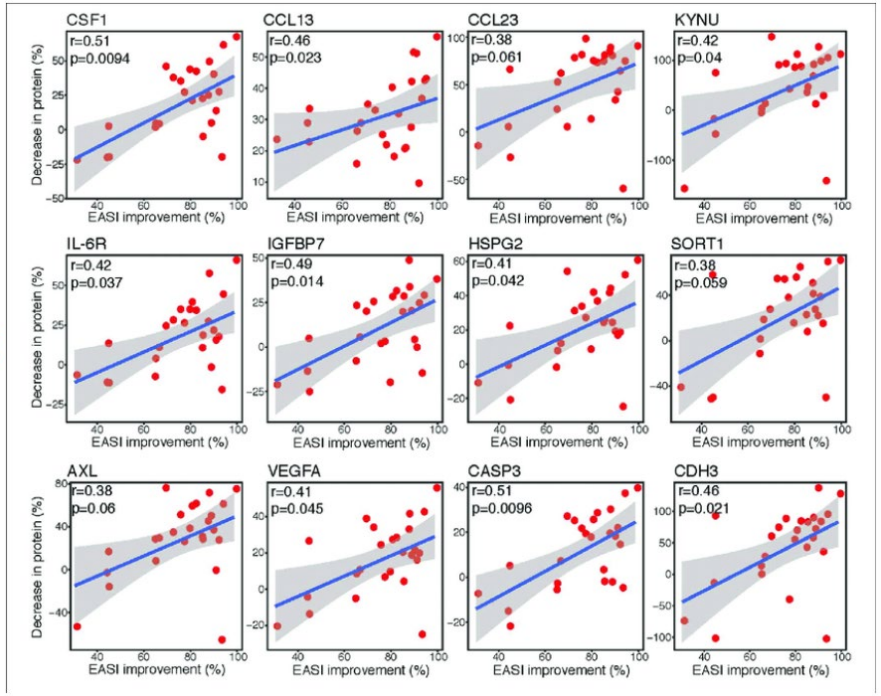
Correlation

- Keeping one out of strongly correlated features – so that it doesn't affect the model compared to other variables.
- Once you've understood the datatype, choice of correlation method will explain type
- Most people use heatmap, but corr plots are possibility, if you know what it means.

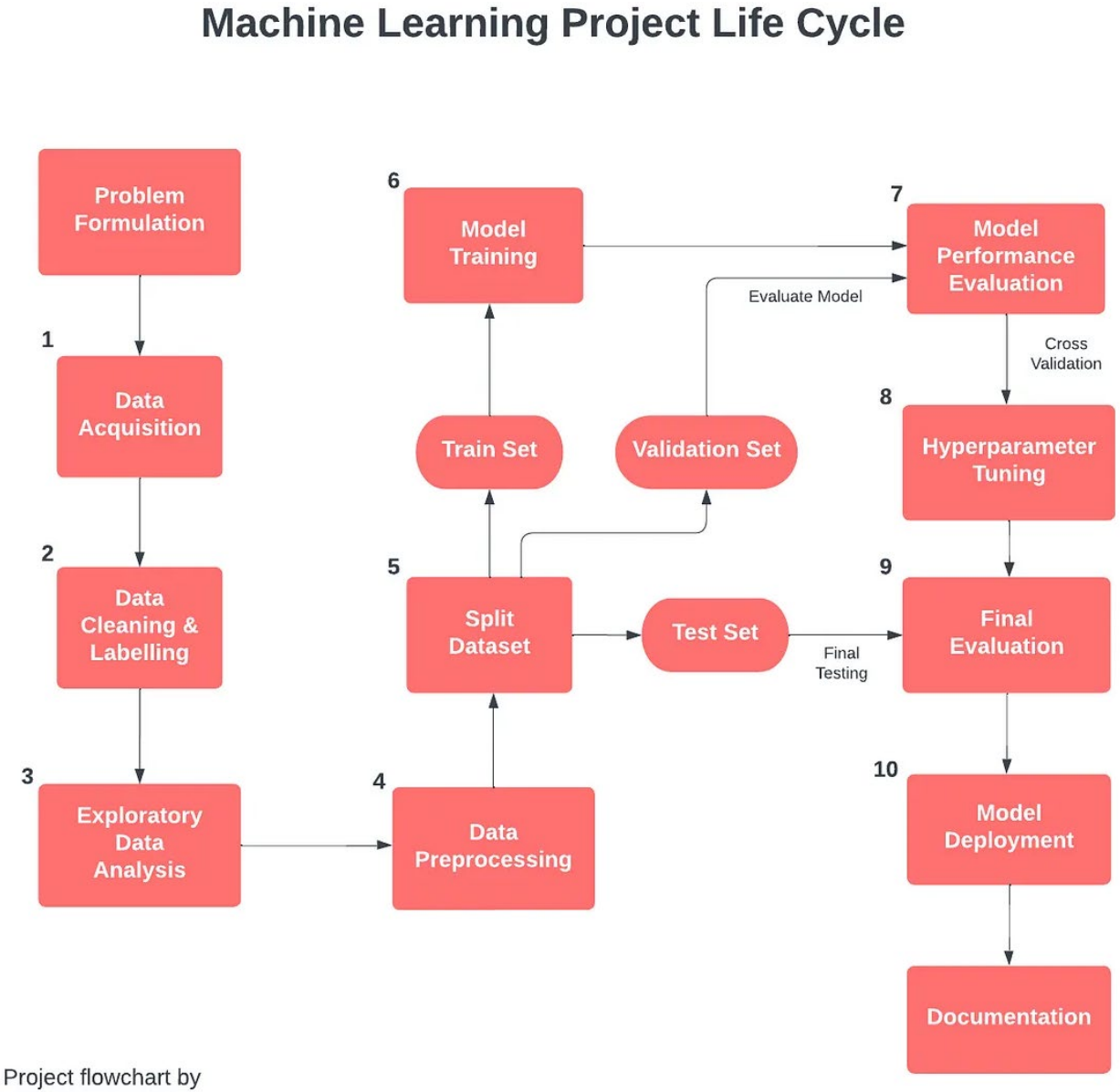
Variable Type	Continuous Variable	Ordinal Variable	Categorical Variable
Continuous variable	Pearson coefficient	Spearman's coefficient	Eta coefficient or difference analysis
Ordinal variable	Spearman's coefficient	Spearman's coefficient	Contingency coefficient, Phi (Φ) coefficient, or difference analysis
Categorical variable	Eta coefficient or difference analysis	Contingency coefficient, Phi (Φ) coefficient, or difference analysis	Contingency coefficient or Phi (Φ) coefficient

Types of Correlation





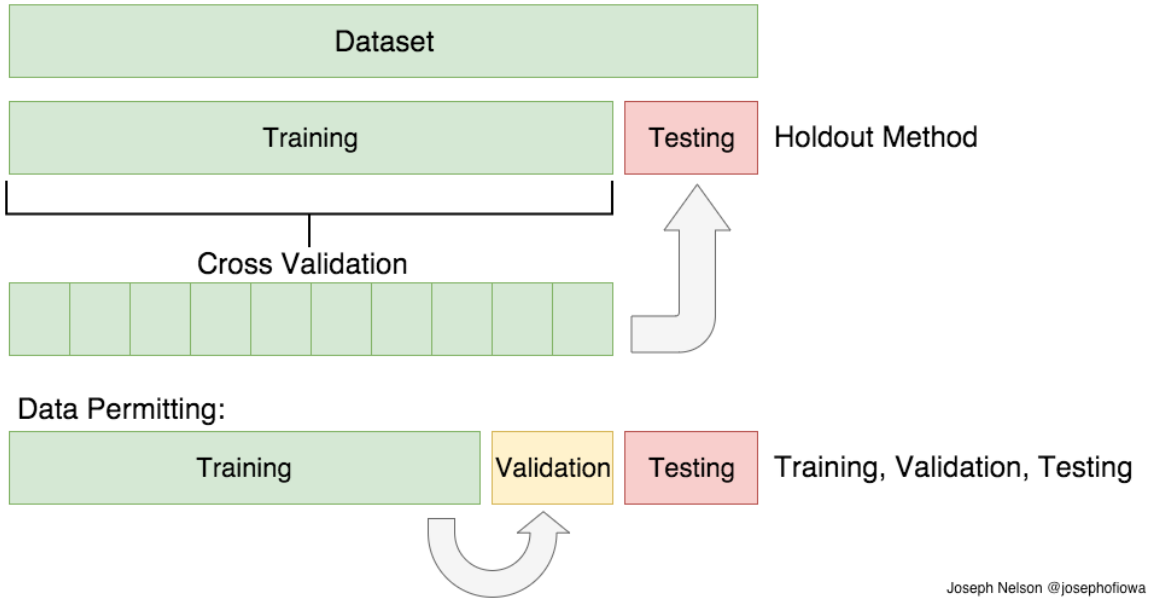
What is the end-to-end process for machine learning?

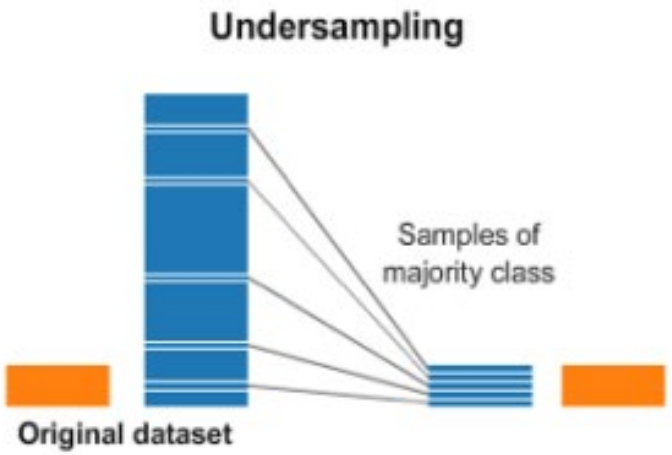


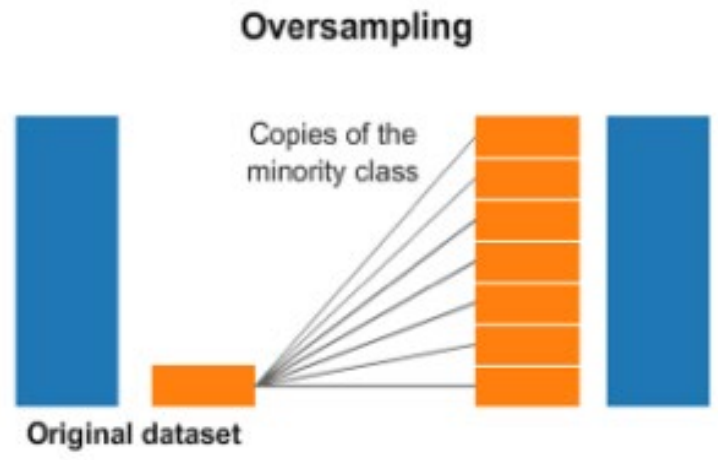
ML Project flowchart by
Ihsanul Haque Asif
North South University

Dataset split

- Hiding data from the model, to validate in the training process and also test once the model is built.

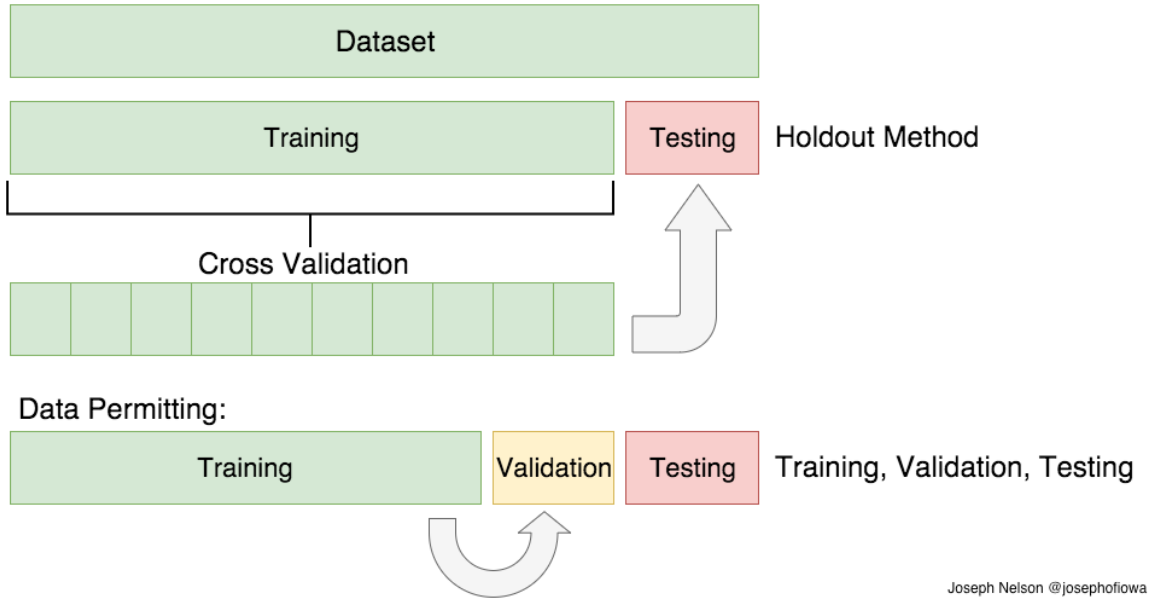


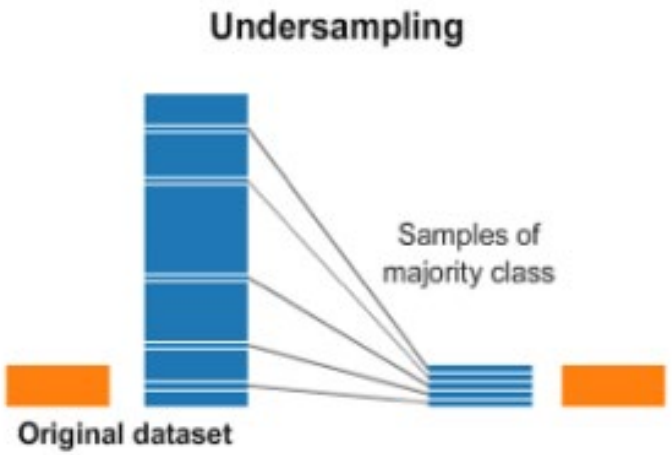


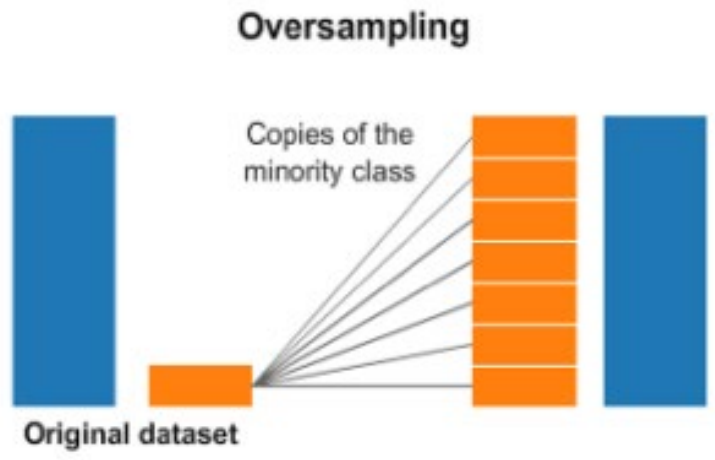


Decision Trees

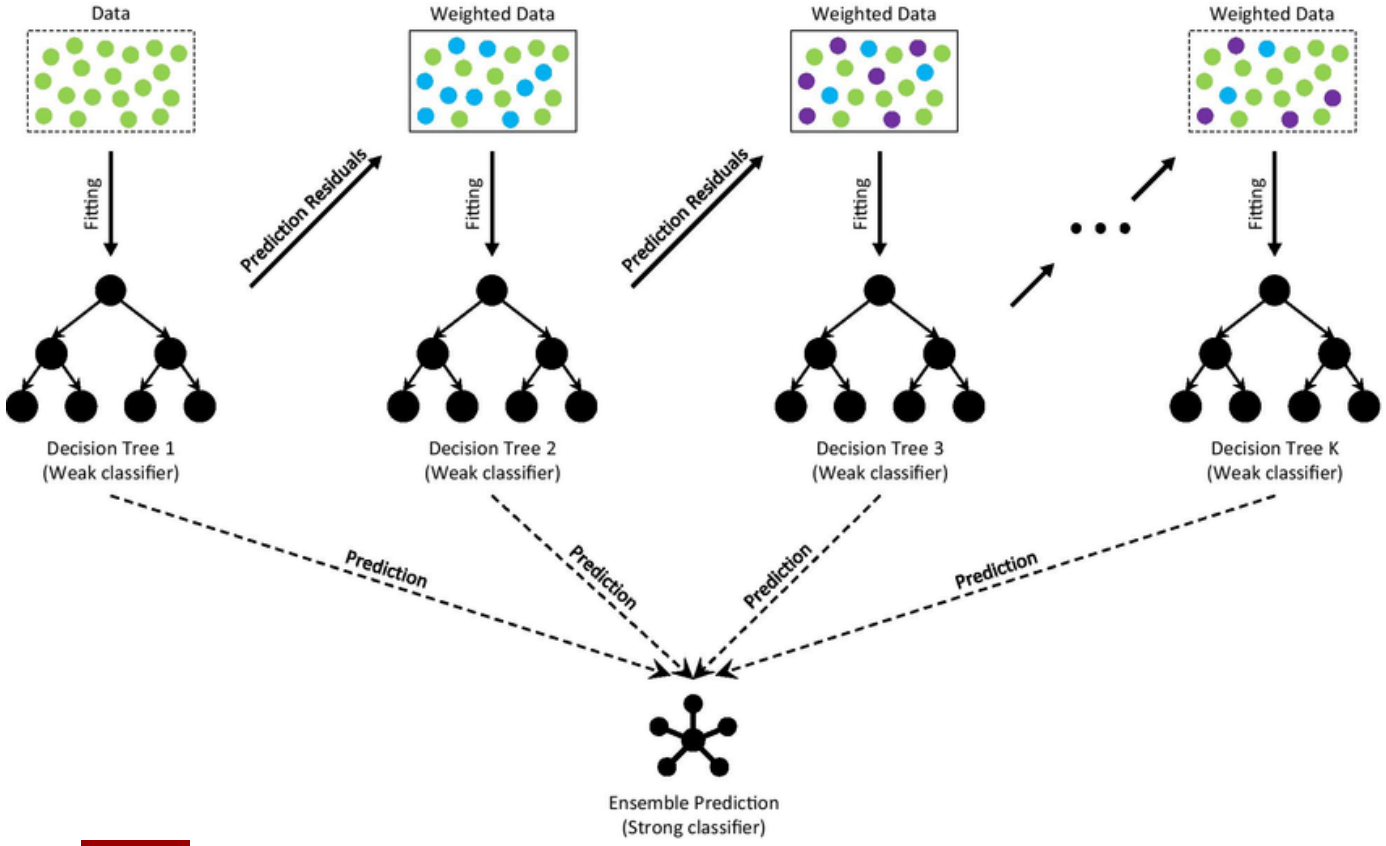
- Hiding data from the model, to validate in the training process and also test once the model is built.

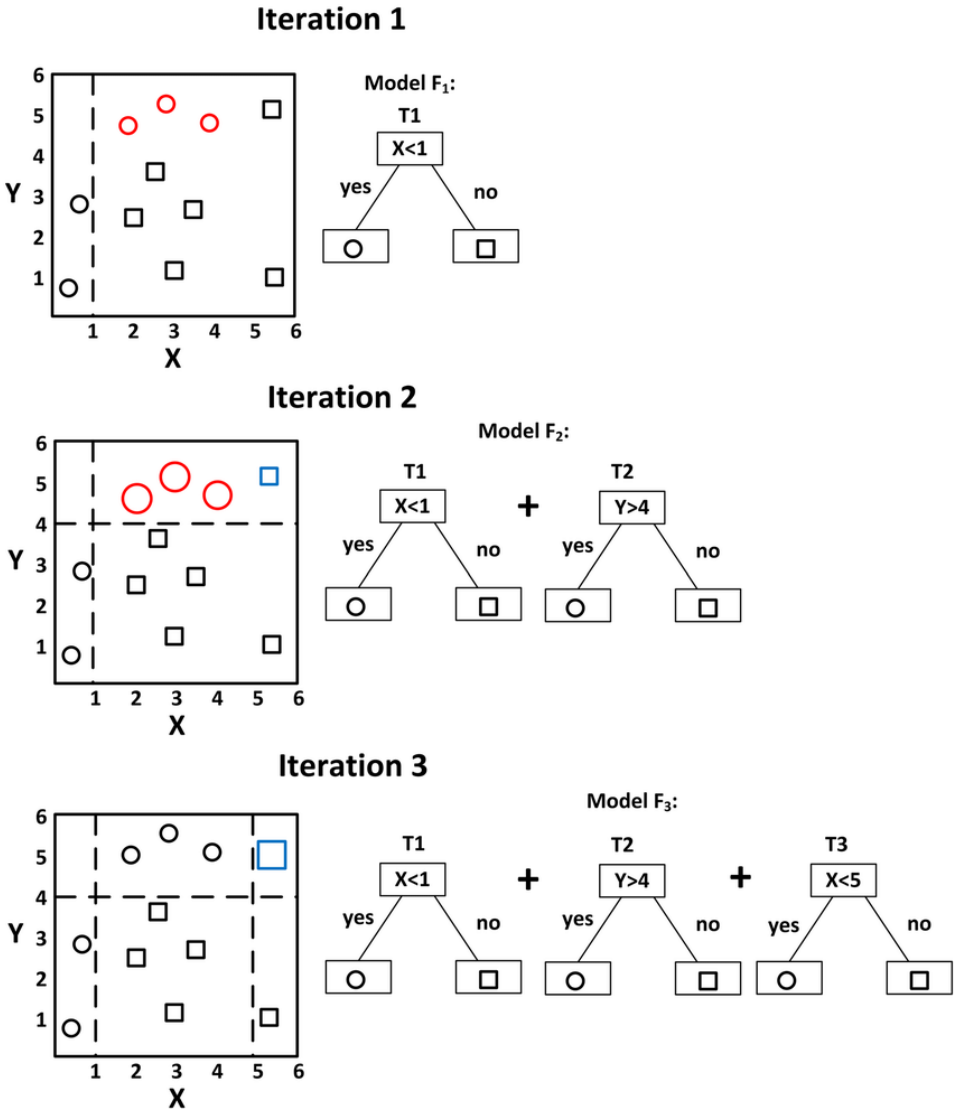




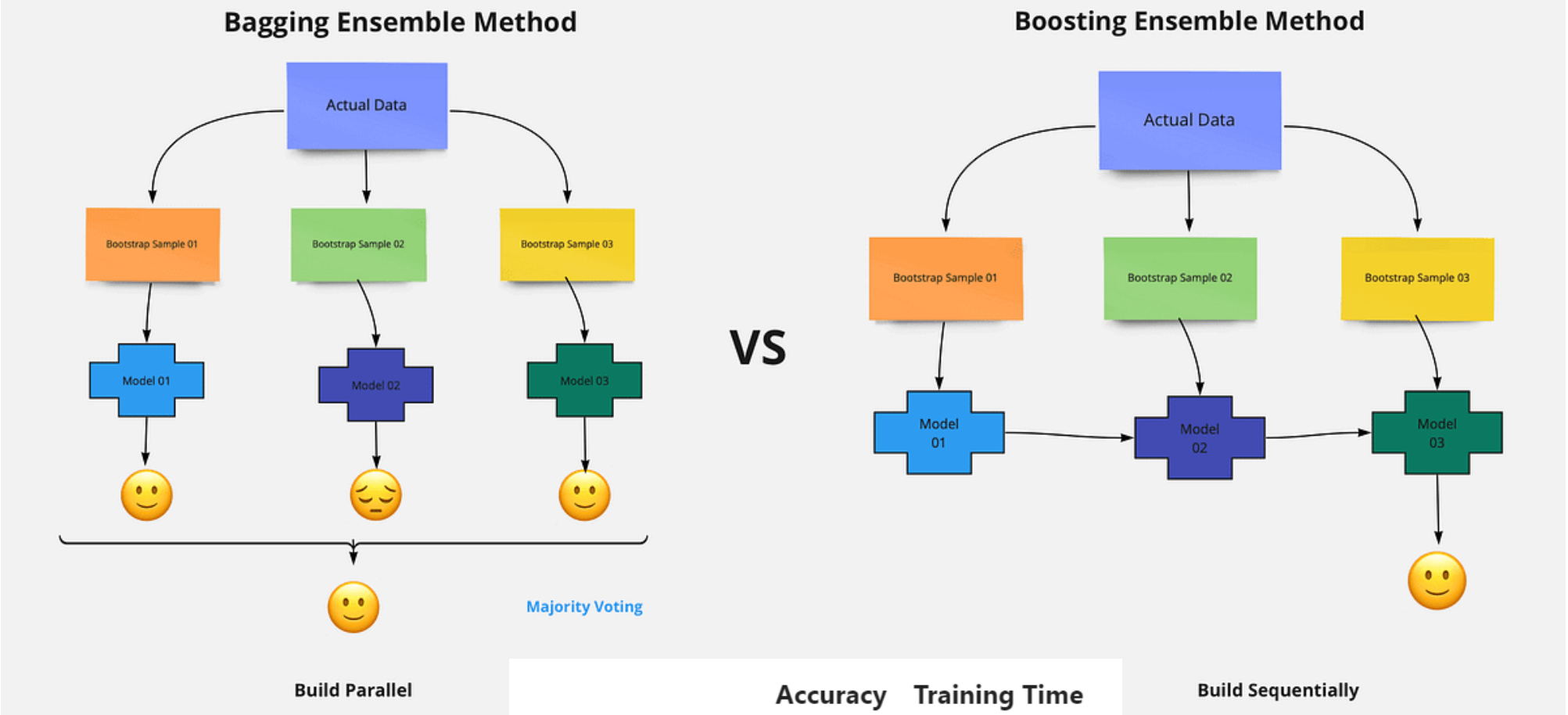


Gradient Boosting methods





Boosting vs Bagging (Random Forest)



	Accuracy	Training Time
LigthGBM	99.964176	23.403013
XGBoost	96.399400	310.672026
Random Forest	96.098400	364.416733
Gradient Boosting	96.346800	670.682759

Explaining with feature importance

Univariate

- F-statistic
- Maximal Information Coefficient
- Predictive Power Score

Multivariate

Fit time

- Impurity Reduction
- Split Count
- Coverage

Predict time

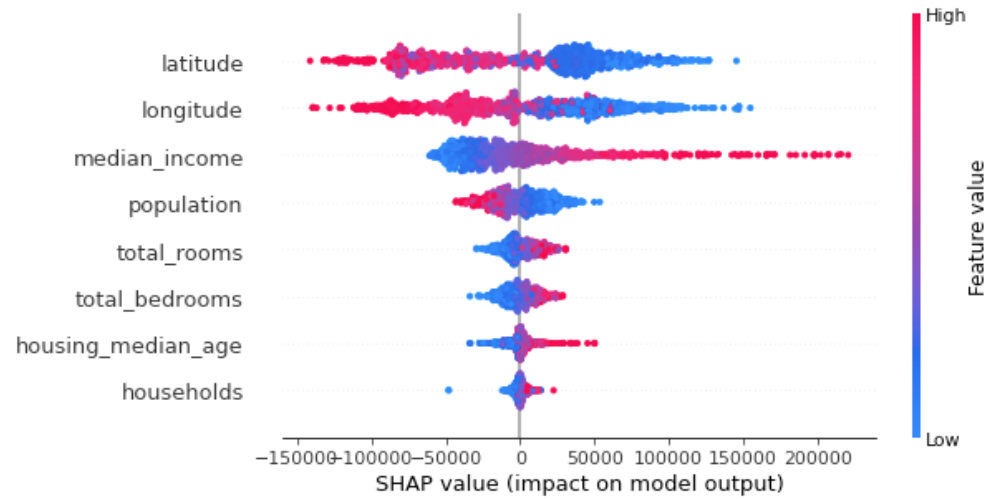
Target required

- Permutation I.

Target not required

- Absolute I.
- Main Factor
- Main Factor with Sign

```
shap.summary_plot(xg_shap_values, X_sh
```



Metrics to evaluate the regression model

$$
\begin{aligned}
R^2 &= 1 - \frac{\text{Residual variance}}{\text{Total variance}} \\
&= \frac{\text{Total variance} - \text{Residual variance}}{\text{Total variance}} \\
&= \frac{\text{Explained variance}}{\text{Total variance}} \\
&= \text{Fraction of total variance explained}
\end{aligned}
$$

number of samples: n

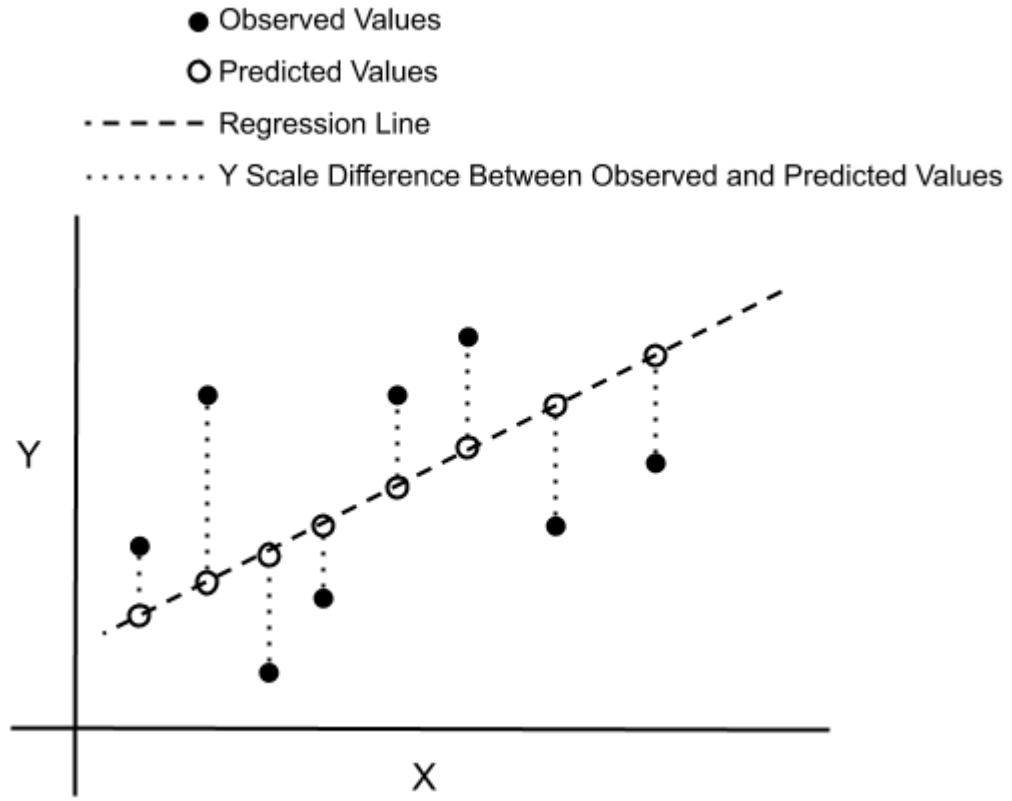
real value: Y_i

predicted value: $\hat{Y}_i$

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

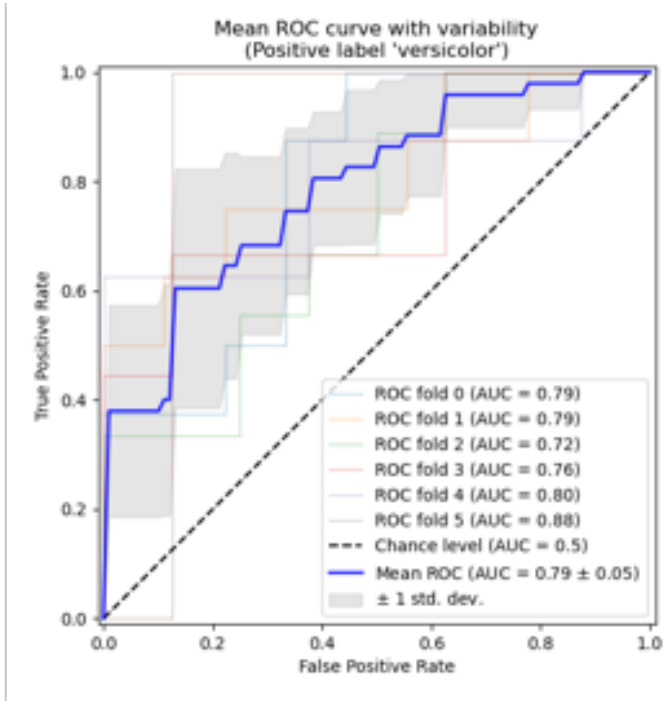
sum of the errors of all samples

R	Interpretation
0.00-0.199	Very weak
0.20-0.399	Weak
0.40-0.599	Medium
0.60-0.799	Strong
0.80-1.00	Very strong



Metrics to evaluate the classification model

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$



CALCULATE PRECISION, RECALL, F1-SCORE

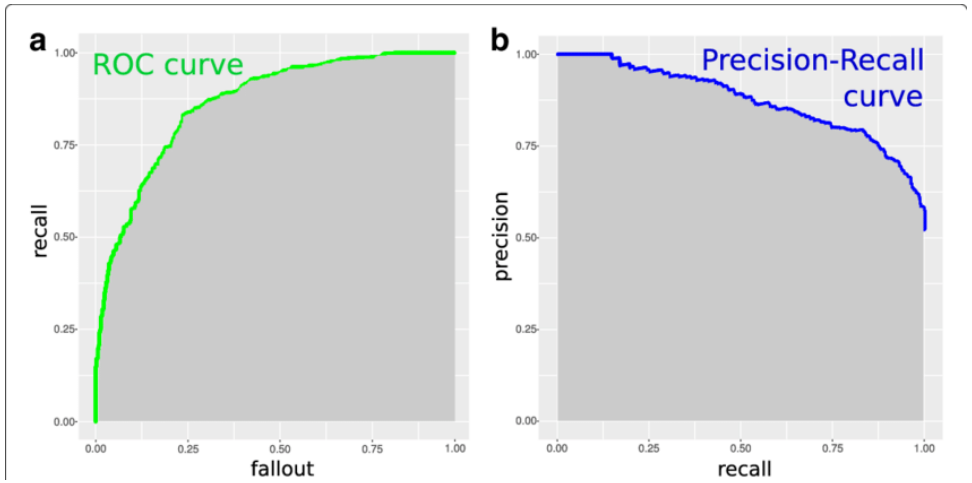
$$precision = \frac{TP}{TP + FP}$$

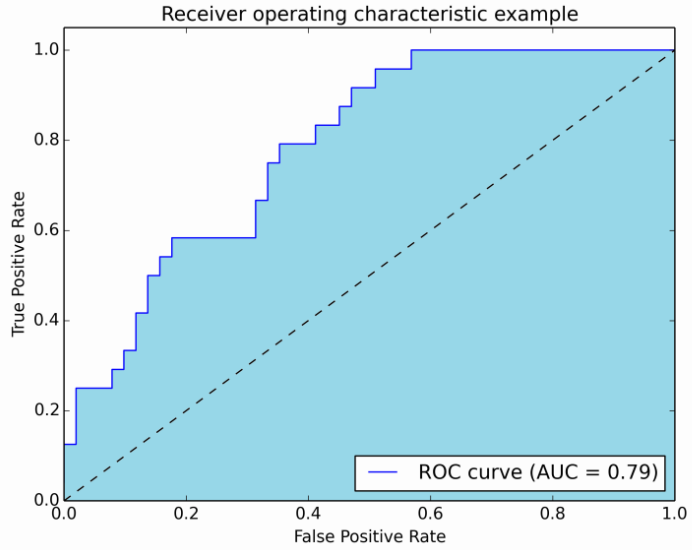
$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

$$specificity = \frac{TN}{TN + FP}$$





Thank you for your attention