FOUNDATIONS OF MACHINE LEARNING

ANUSH SANKARAN

Most of the slide content are borrowed from multiple source online. Thanks to the original authors

OVERVIEW OF THE COURSE

Week	Topics
Week 1	Intro to ML Discovering ML Use Cases & ML in Business
Week 2	Python- Hands On Supervised Learning & Regression
Week 3	Neural Network - 1 Neural Network -2 (Bias, Variance) & Hands ON
Week 4	Kernel Learning & SVM Practical Advice for ML projects.
Week 5	Boosting Decision Trees, Random Forest, & xgBoost
Week 6	Unsupervised Learning Clustering & Dimensionality Reduction
Week 7	Time Series Data Analysis Imputation & Prediction Systems
Week 8	ML Use Cases from Products & Research

COURSE OUTCOMES

- Understand the fundamental concepts of different machine learning models
  - Supervised learning
  - Unsupervised learning
- Ability to formulate a business problem as machine learning task. Identify machine learning opportunities in businesses.
- Appreciate the challenging involved in data driven machine learning problems
- Ability to manage the building of tools and products that involves different aspects of machine learning

EASY LOGISTICS: GITHUB

- Github Repo: https://github.com/goodboyanush/isme-bangalore-Oct-Nov-2019
- Lectures slides, Hands-on code, Assignment solutions
- Have any doubt in my lectures or assignments?
  - Go ahead and create an issue in the repo!
  - I will try to answer them asap!
  - Everyone will be benefitted by the questions asked by one

WEEK 1:

INTRODUCTION TO MACHINE LEARNING

BUILDING THE ML MINDSET IN BUSINESS

WHAT IS MACHINE LEARNING?

- Learn a classifier: learn a mapping function

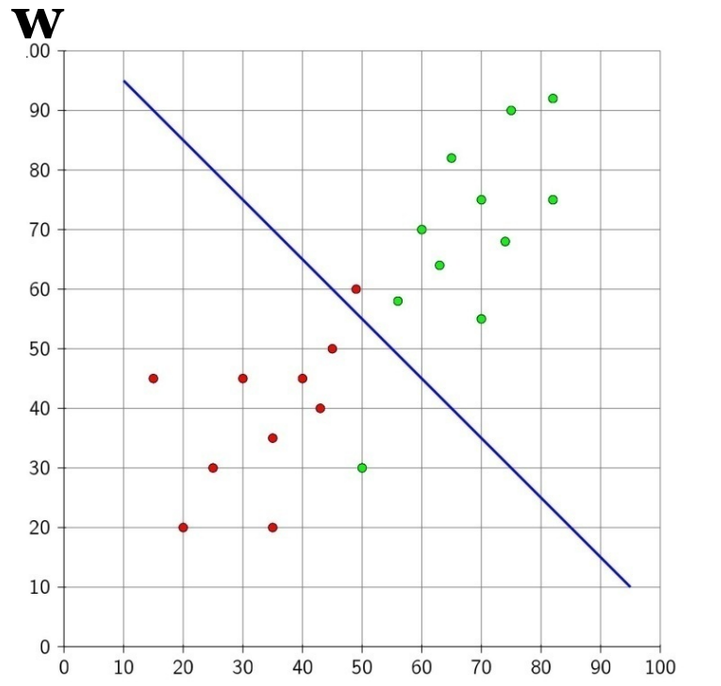
Instances/ Input Data Points

1. M: <h1, w1>
2. F: <h2, w2>
3. …

…

N. M: <hn, wn>

Labelled Features



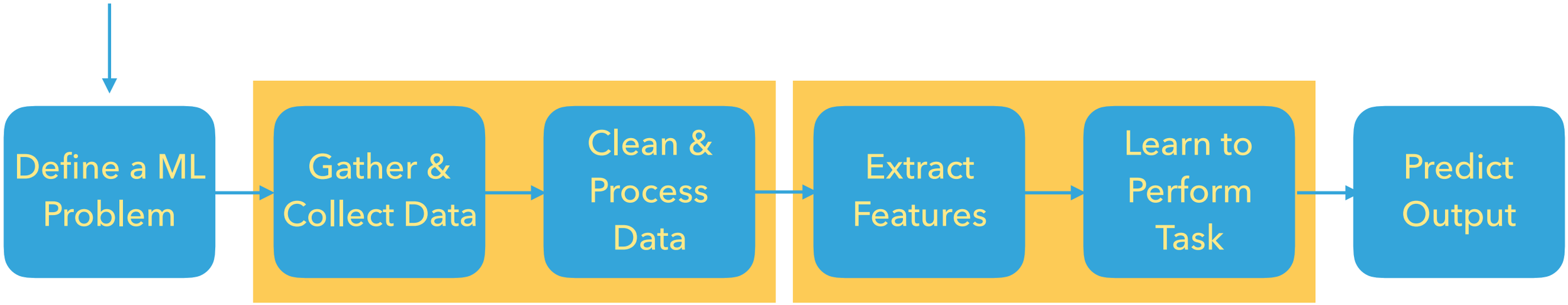
Classification

1. **Boundary**: Can be linear or non-linear boundary
2. **Method**: Can be a generative classifier or discriminative classifier

Examples: Naïve Bayes, Decision trees, Neural Network, Support Vector Machines etc.

MACHINE LEARNING PIPELINE

What are we focussing on today's lecture?



1. Articulate the problem (task)
2. Data Drive Strategy: Look for labelled data
3. Design your data for the task
4. Determine easily obtained inputs
5. Determine easily quantifiable outputs

COMMON LINGO

Output
Task Label

Input data/ Features

Instances

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y
LP001036	Female	No	0	Graduate	No	3510	0	76	360	0	Urban	N

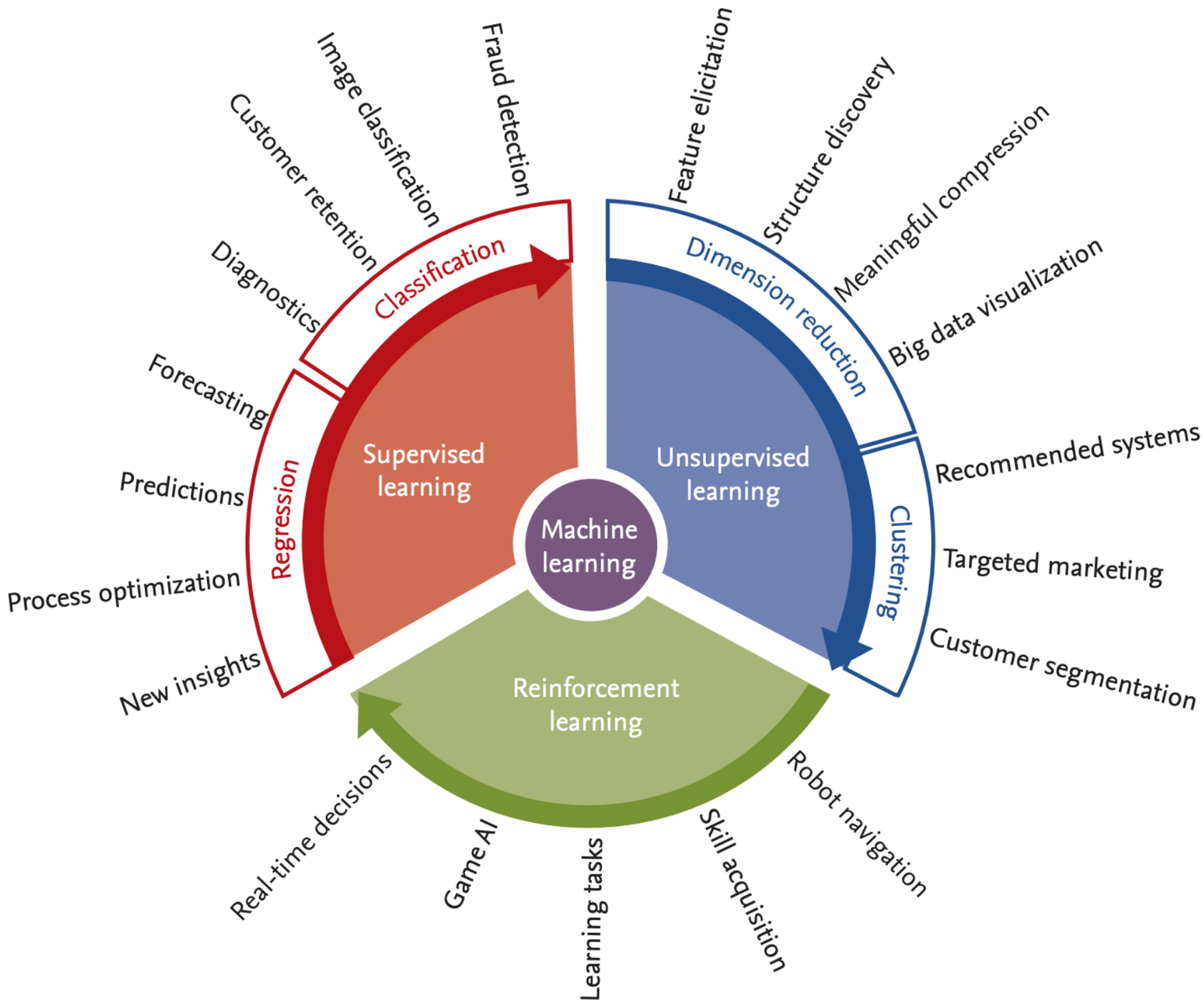
Seen data / Training data

Unseen data/ Test data

LP001038	Male	Yes	0	Not Graduate	No	4887	0	133	360	1	Rural	
LP001043	Male	Yes	0	Not Graduate	No	7660	0	104	360	0	Urban	
LP001046	Male	Yes	1	Graduate	No	5955	5625	315	360	1	Urban	
LP001047	Male	Yes	0	Not Graduate	No	2600	1911	116	360	0	Semiurban	

Machine Learning Glossary: https://developers.google.com/machine-learning/glossary

DIFFERENT TYPES OF ML ALGORITHMS



THE ML MINDSET

"Machine Learning changes the way you think about a problem.

The focus shifts from a mathematical science to a natural science, running experiments and using statistics, not logic, to analyse its results."

- **Peter Norvig**

THE ML MINDSET

Step	Example
1. Set the research goal.	I want to predict how heavy traffic will be on a given day.
2. Make a hypothesis.	I think the weather forecast is an informative signal.
3. Collect the data.	Collect historical traffic data and weather on each day.
4. Test your hypothesis.	Train a model using this data.
5. Analyze your results.	Is this model better than existing systems?
6. Reach a conclusion.	I should (not) use this model to make predictions, because of X, Y, and Z.
7. Refine hypothesis and repeat.	Time of year could be a helpful signal.

Get Comfortable with Some Uncertainty !

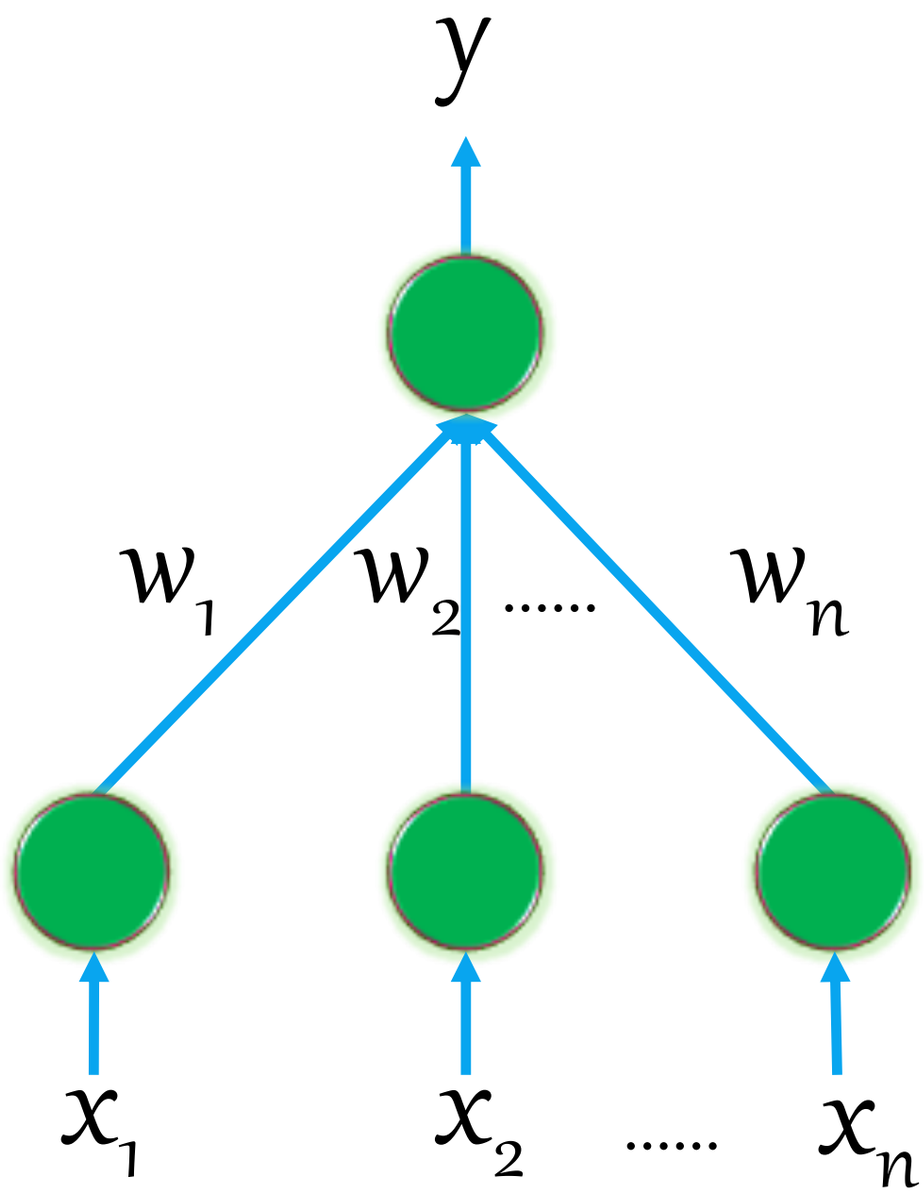
REGRESSION – LINGO

$$Y = X_1 + X_2 + X_3$$

Dependent Variable	Independent Variable
Outcome Variable	Predictor Variable
Response Variable	Explanatory Variable

LINEAR REGRESSION – ACTIVATION FUNCTION

$$y = f(\Sigma_i w_i x_i)$$



$$\text{Error} = (y - y')^2$$

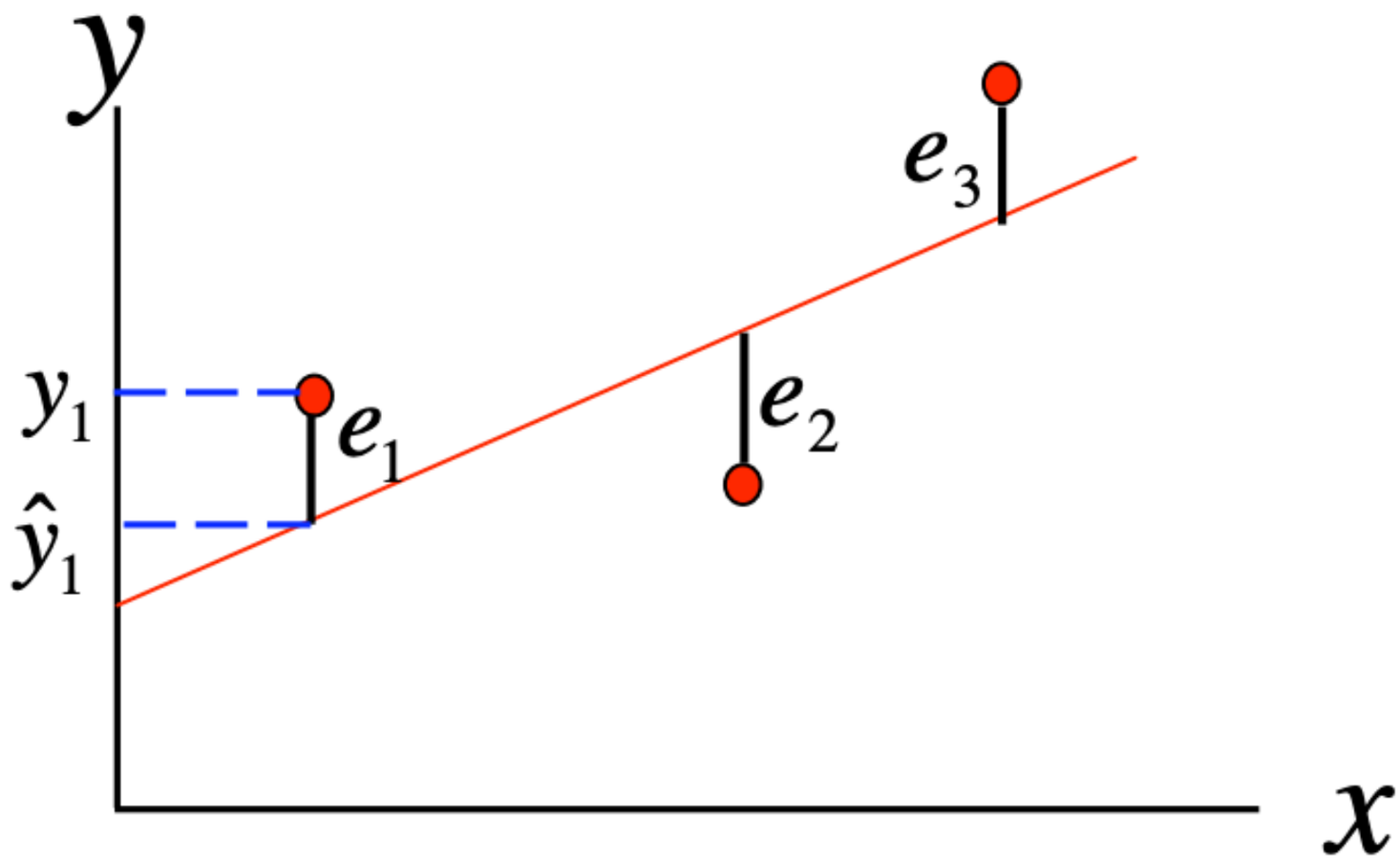
Residual Sum of Squares (RSS)

Update **w** in such a way that the error is minimized !

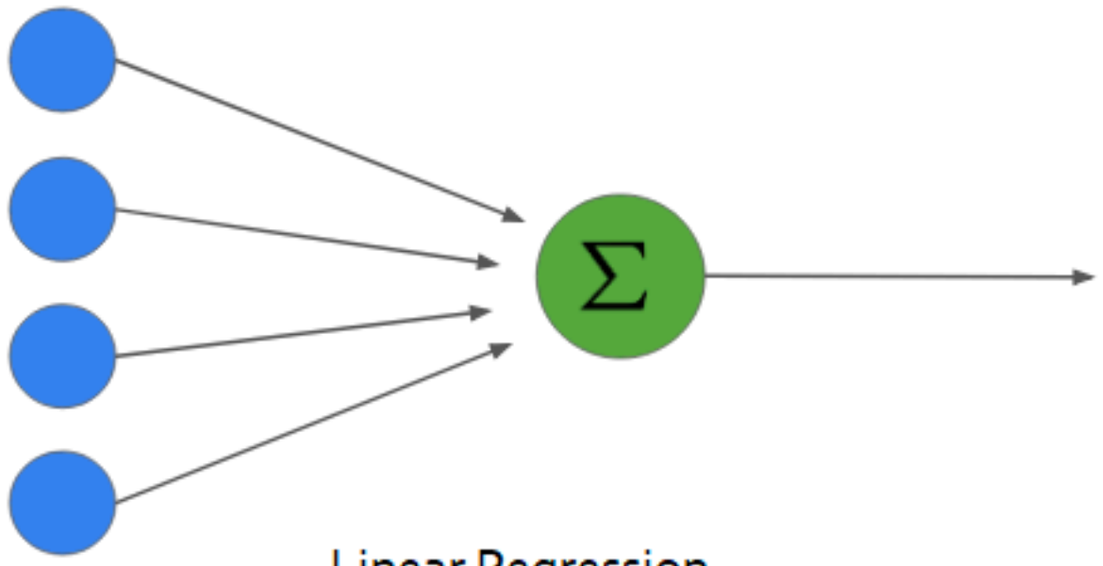
RESIDUAL SUM OF SQUARES

$$e_1 = y_1 - \hat{y}_1$$

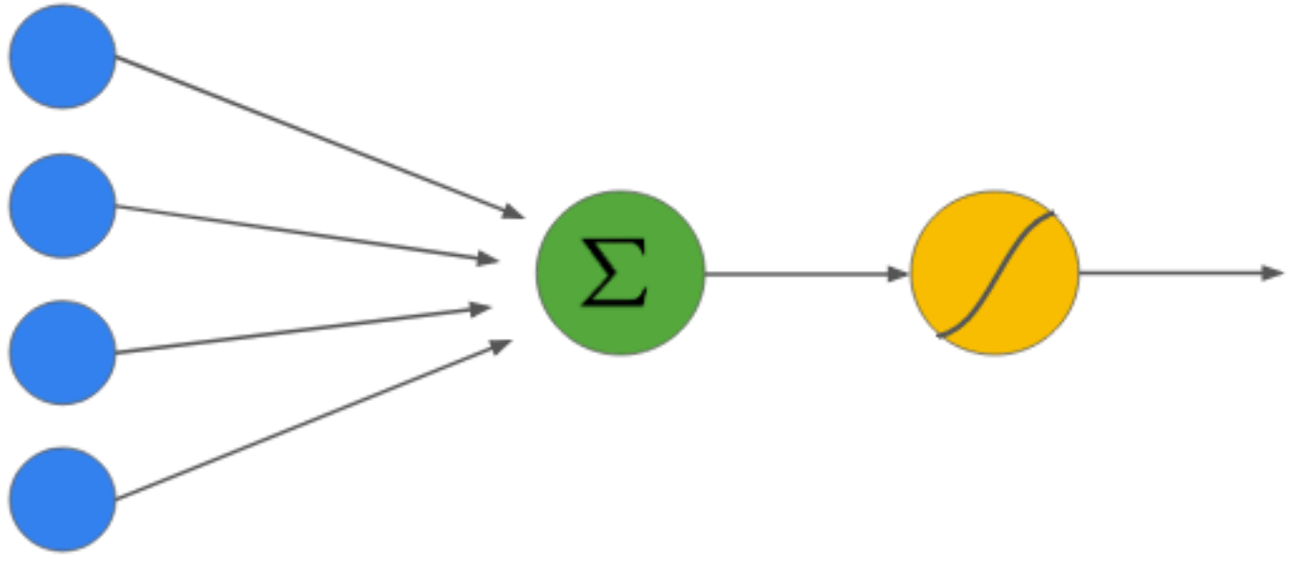
$$e_2 = y_2 - \hat{y}_2$$



LINEAR VS. LOGISTIC REGRESSION



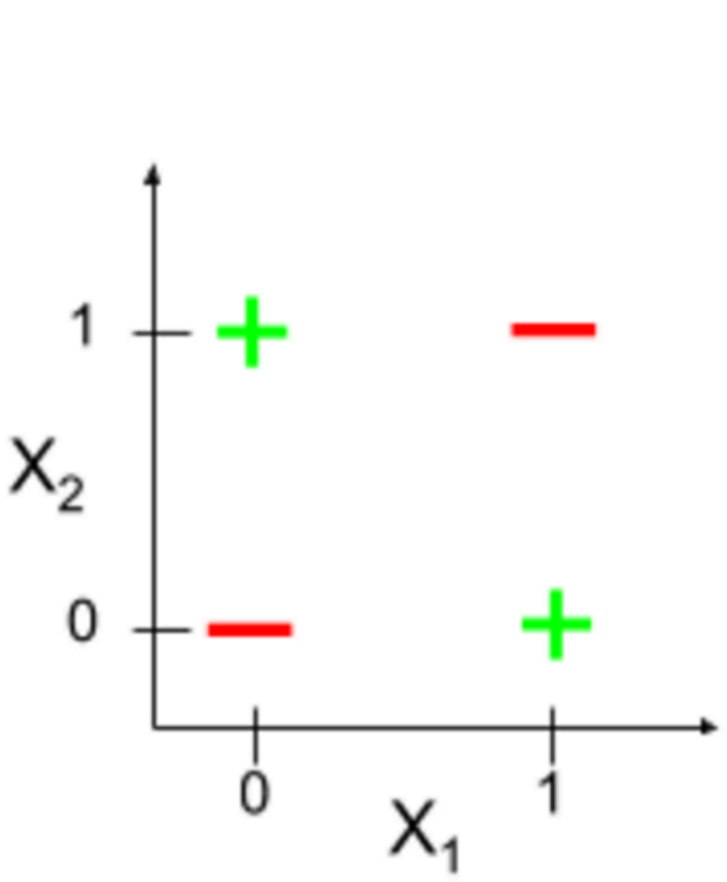
Linear Regression

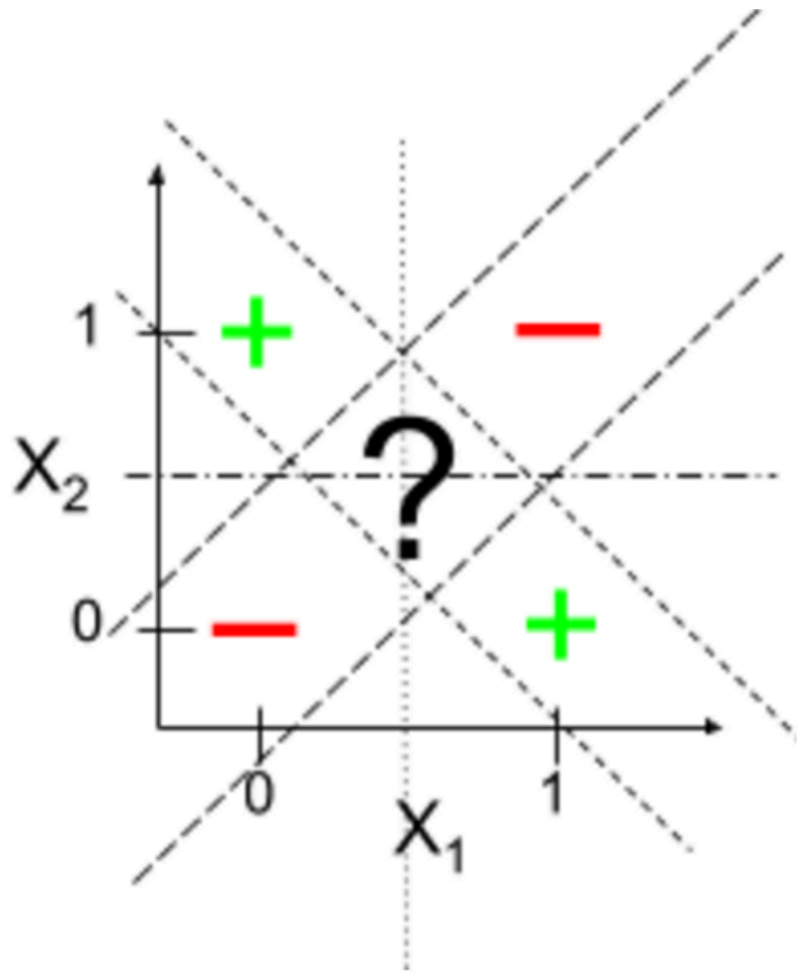


Logistic Regression

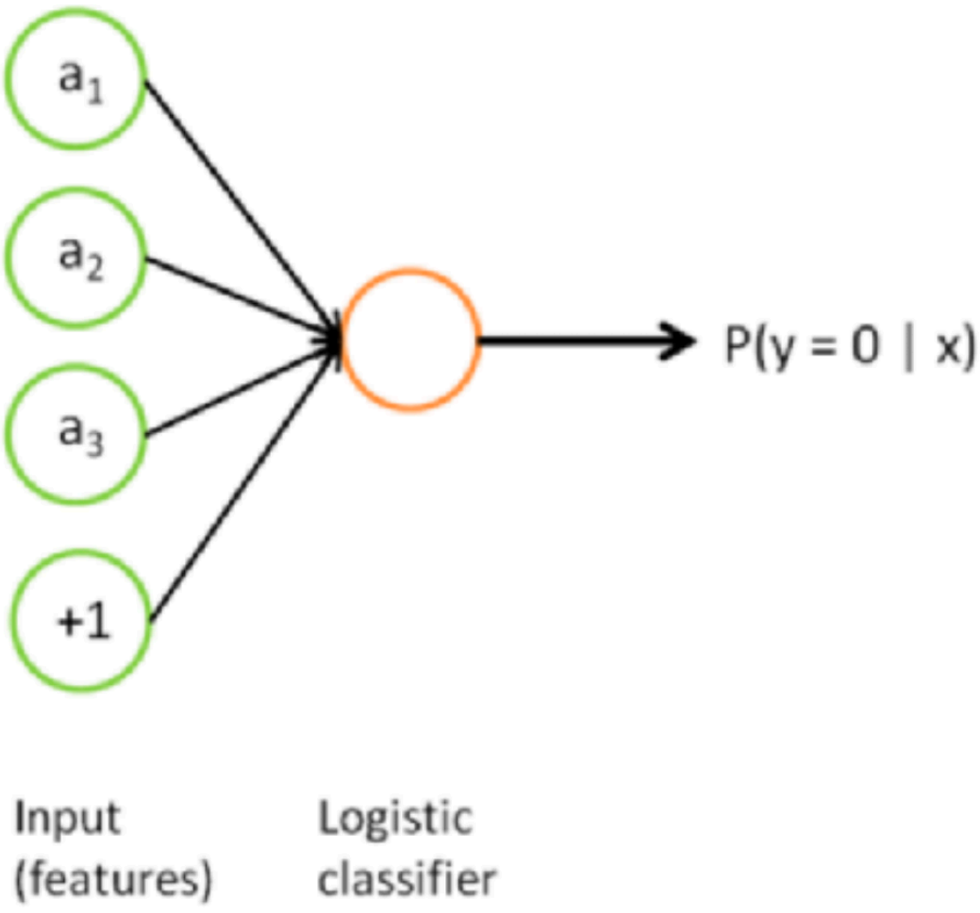
PROBLEM?

Produce only linear classification boundary (straight lines)

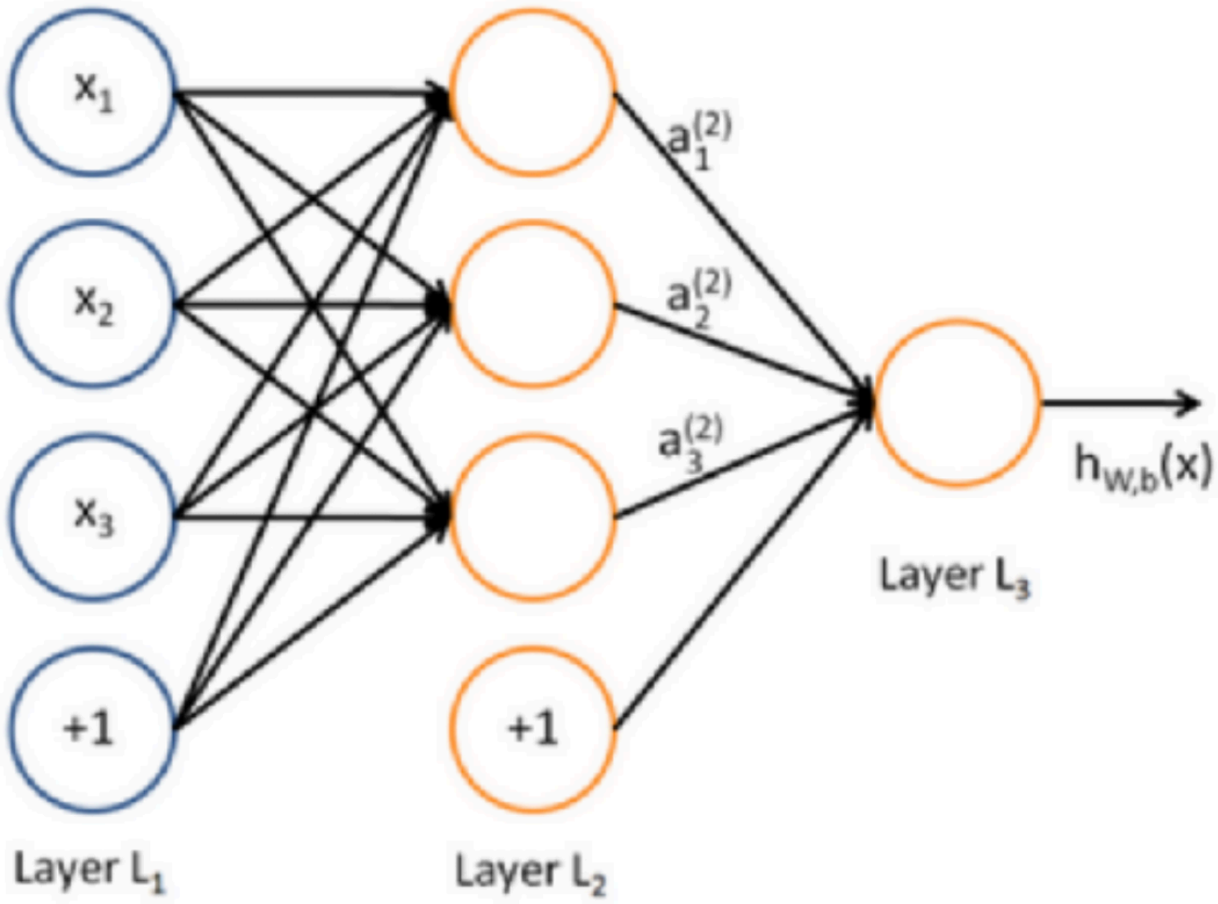




NEURAL NETWORKS

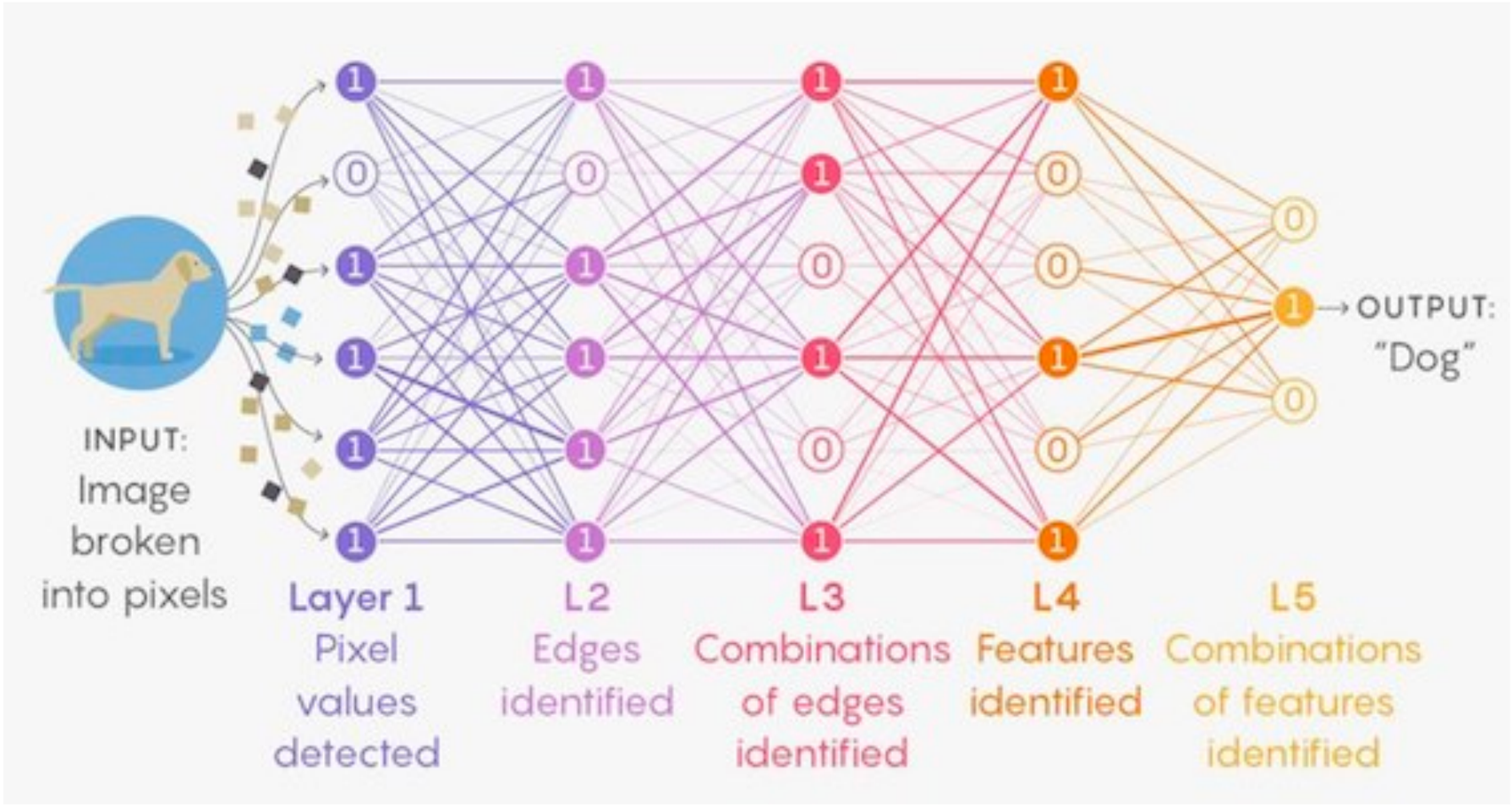


Logistic Regression

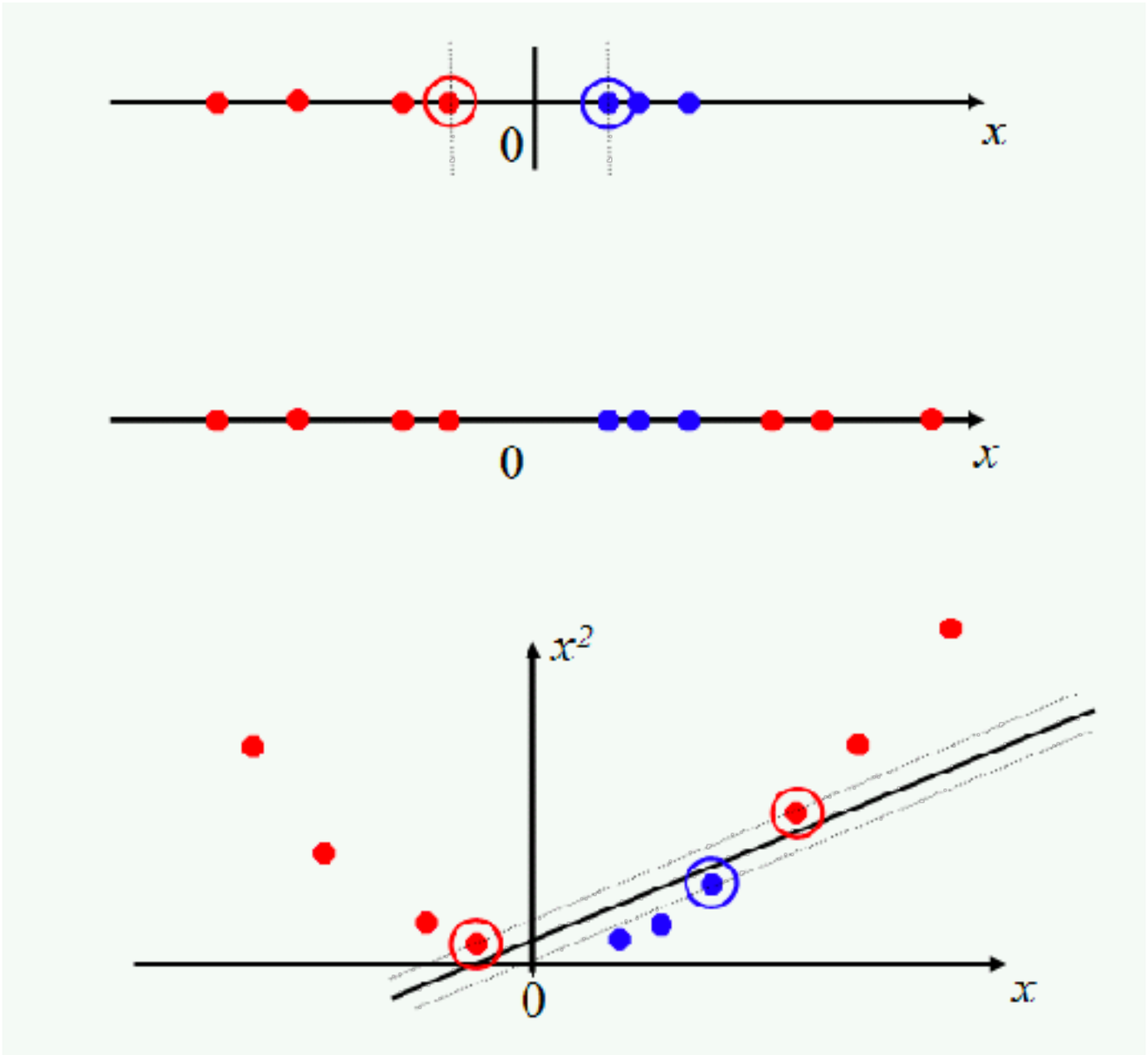


Neural Network

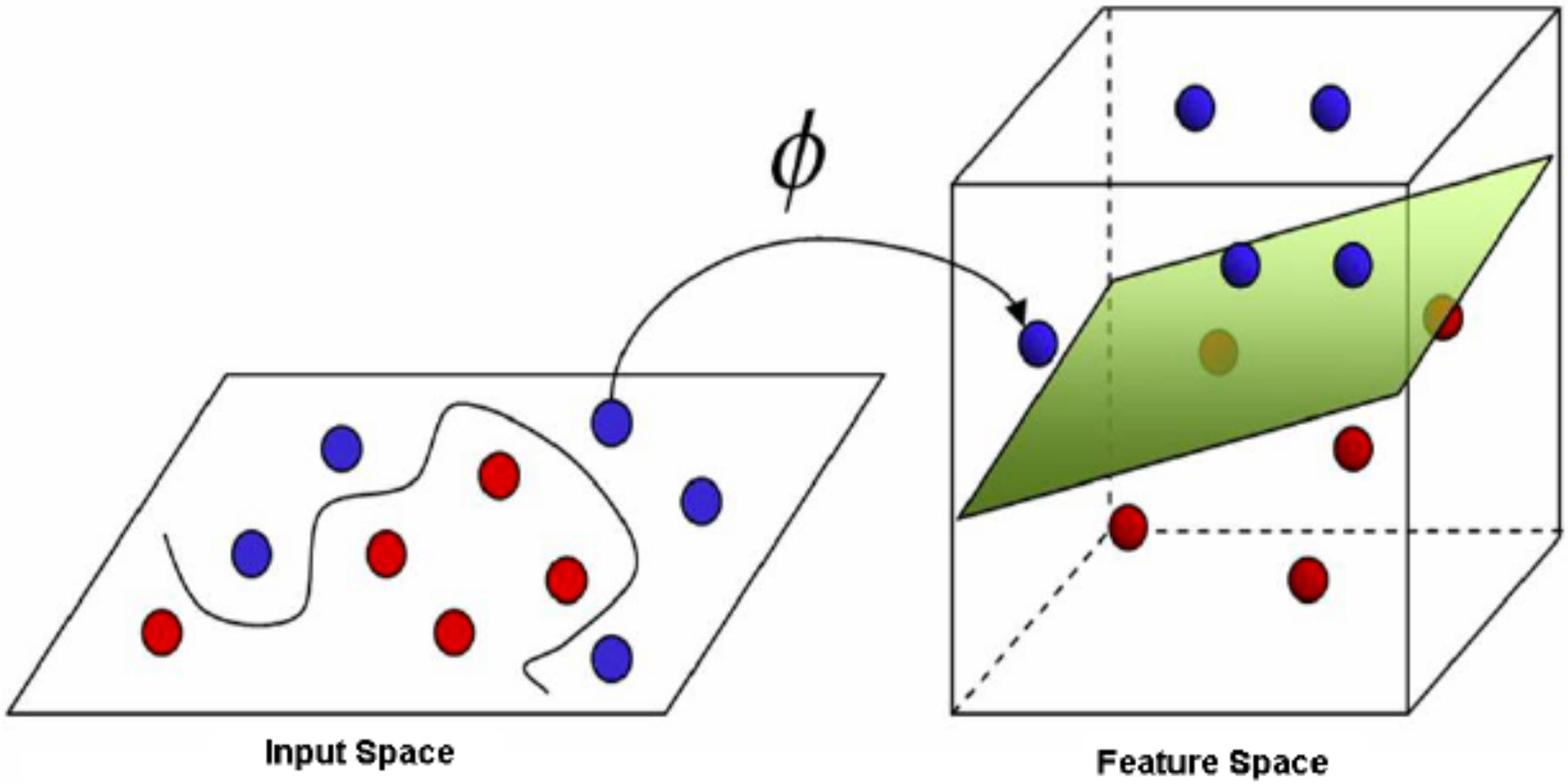
NEURAL NETWORKS



KERNEL TRICK



KERNEL LEARNING



THANK YOU – NEXT WEEK

Week	Topics
Week 1	Intro to ML Discovering ML Use Cases & ML in Business
Week 2	Python- Hands On Supervised Learning & Regression
Week 3	Neural Network - 1 Neural Network -2 (Bias, Variance) & Hands ON
Week 4	Kernel Learning & SVM Practical Advice for ML projects.
Week 5	Boosting Decision Trees, Random Forest, & xgBoost
Week 6	Unsupervised Learning Clustering & Dimensionality Reduction
Week 7	Time Series Data Analysis Imputation & Prediction Systems
Week 8	ML Use Cases from Products & Research

CLASSIFICATION TASKS

Predictors | Target

Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

ENTROPY

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Predictors				Target
Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Play Golf	
Yes	No
9	5

Entropy(PlayGolf) = Entropy (5,9)
= Entropy (0.36, 0.64)
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

		Play Golf		
		Yes	No	
Outlook	Sunny	3	2	5
	Overcast	4	0	4
	Rainy	2	3	5
				14

E(PlayGolf, Outlook) = **P**(Sunny)*__E__(3,2) + **P**(Overcast)*__E__(4,0) + **P**(Rainy)*__E__(2,3)
= (5/14)*0.971 + (4/14)*0.0 + (5/14)*0.971
= 0.693

INFORMATION GAIN

Predictors | Target

Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

		Play Golf	
		Yes	No
Outlook	Sunny	3	2
	Overcast	4	0
	Rainy	2	3
Gain = 0.247			

		Play Golf	
		Yes	No
Temp.	Hot	2	2
	Mild	4	2
	Cool	3	1
Gain = 0.029			

		Play Golf	
		Yes	No
Humidity	High	3	4
	Normal	6	1
Gain = 0.152			

		Play Golf	
		Yes	No
Windy	False	6	2
	True	3	3
Gain = 0.048			

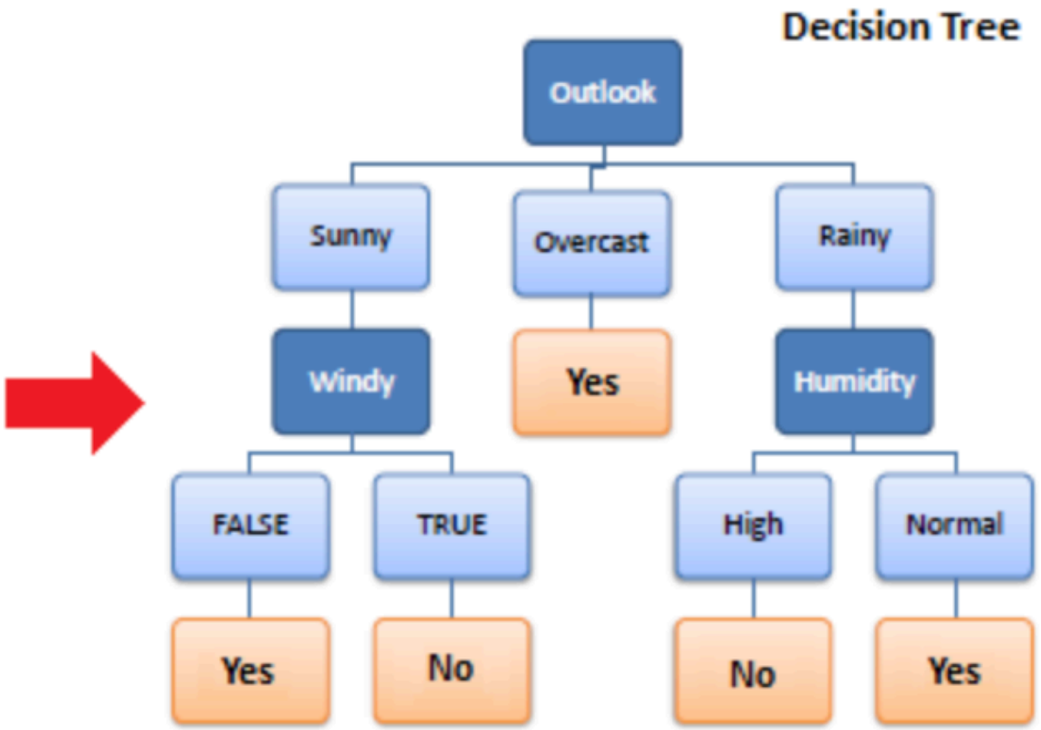
$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

G(PlayGolf, Outlook) = **E**(PlayGolf) – **E**(PlayGolf, Outlook)

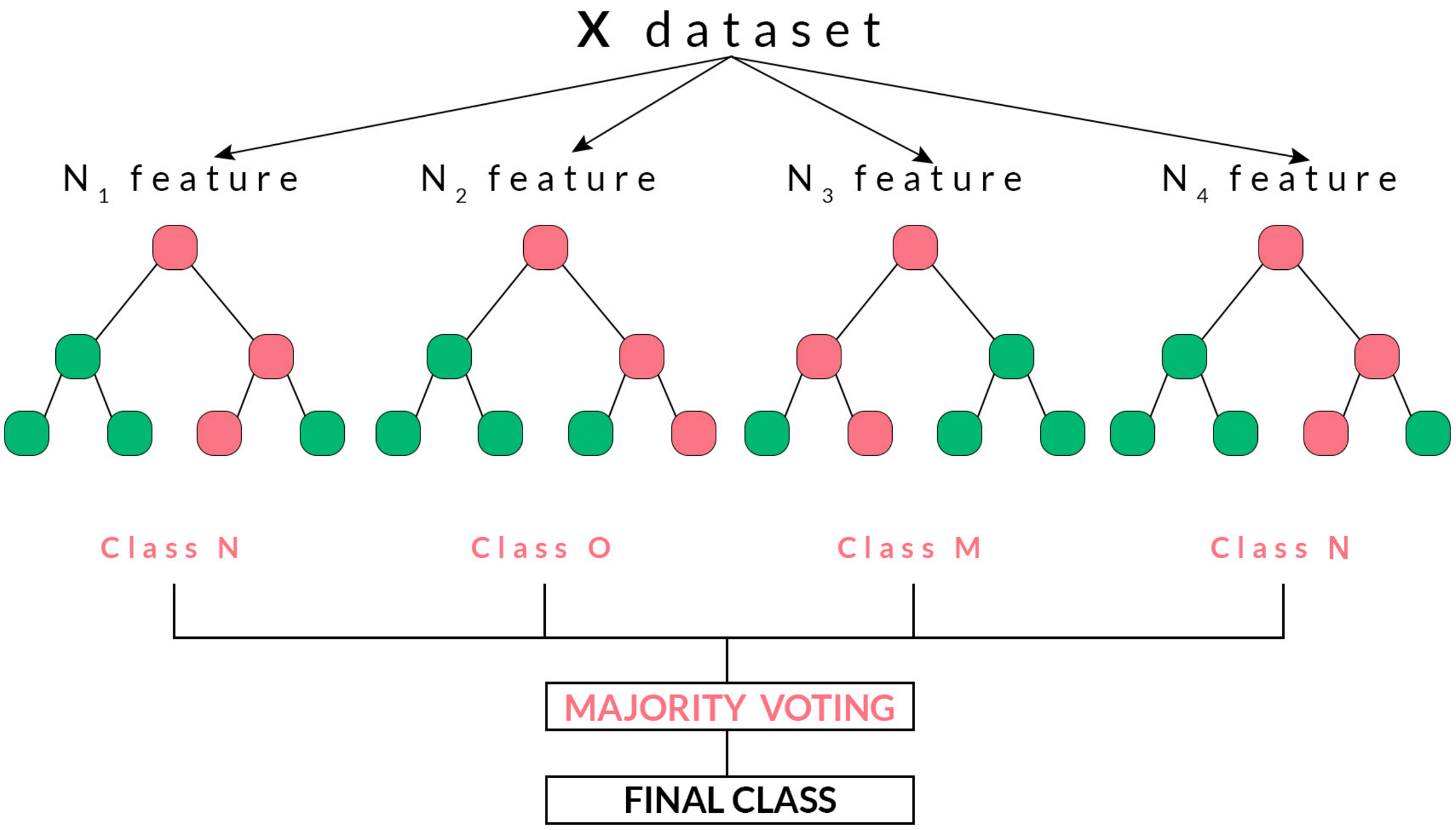
= 0.940 – 0.693 = 0.247

INFORMATION GAIN

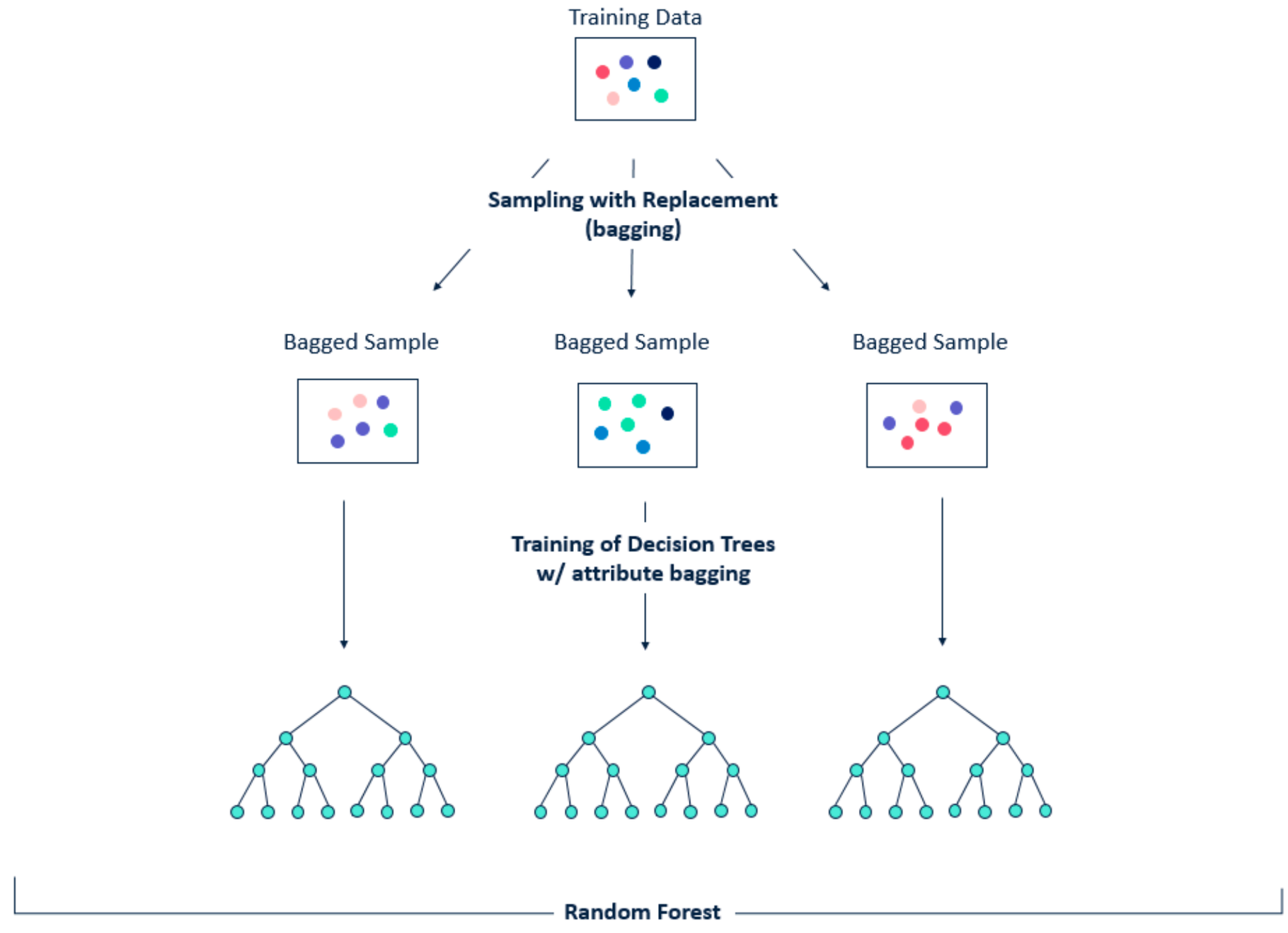
Predictors				Target
Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



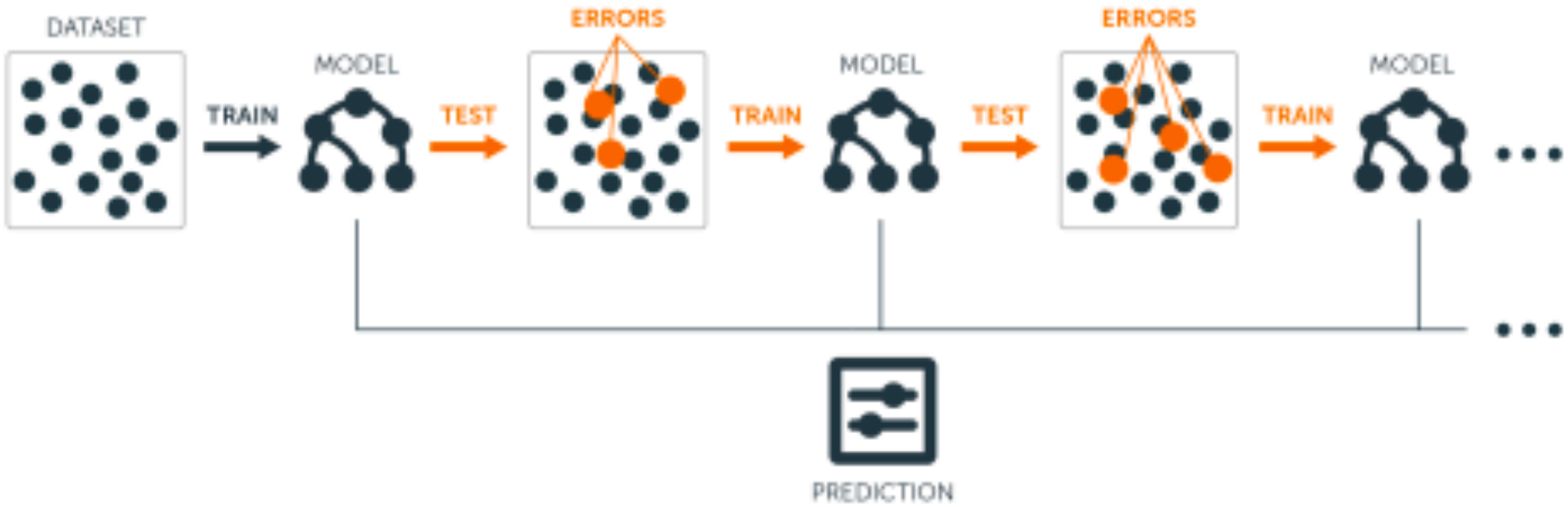
RANDOM FOREST



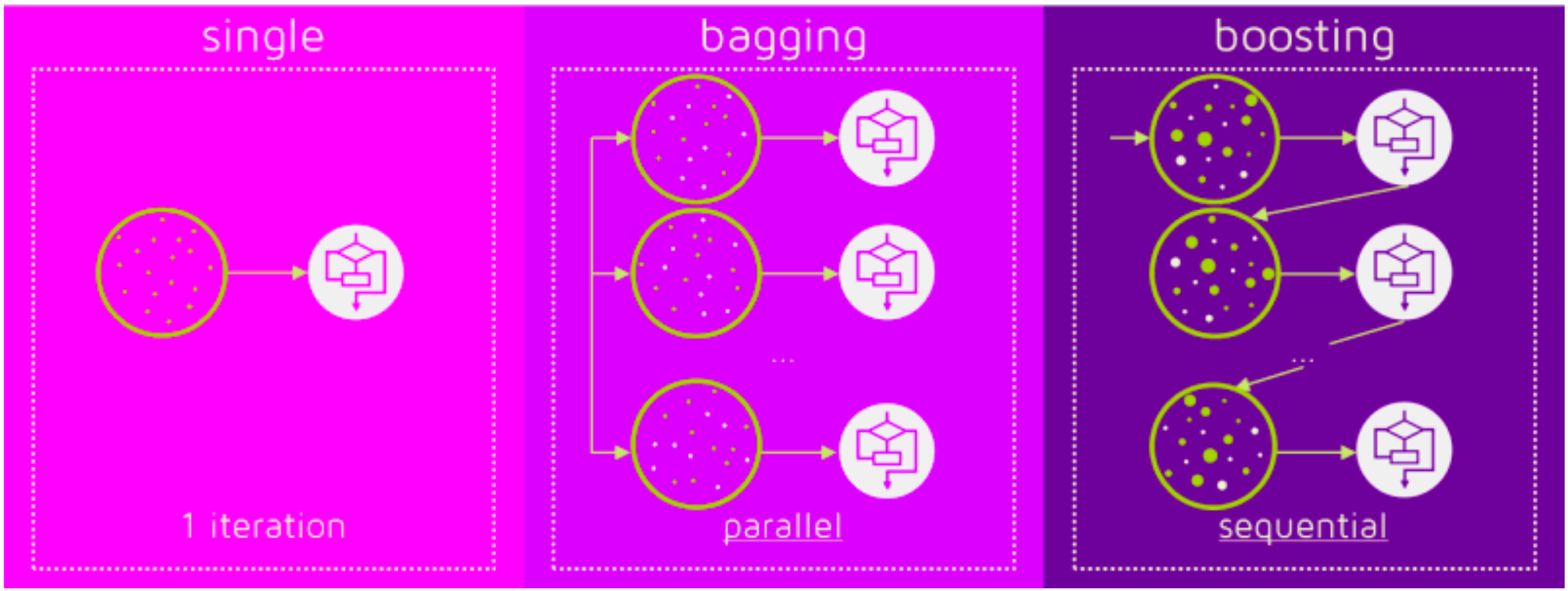
BAGGING


Training Data
Sampling with Replacement
(bagging)
Bagged Sample
Bagged Sample
Bagged Sample
Training of Decision Trees
w/ attribute bagging
Random Forest

BOOSTING


DATASET
TRAIN
MODEL
TEST
ERRORS
TRAIN
MODEL
TEST
ERRORS
TRAIN
MODEL
PREDICTION

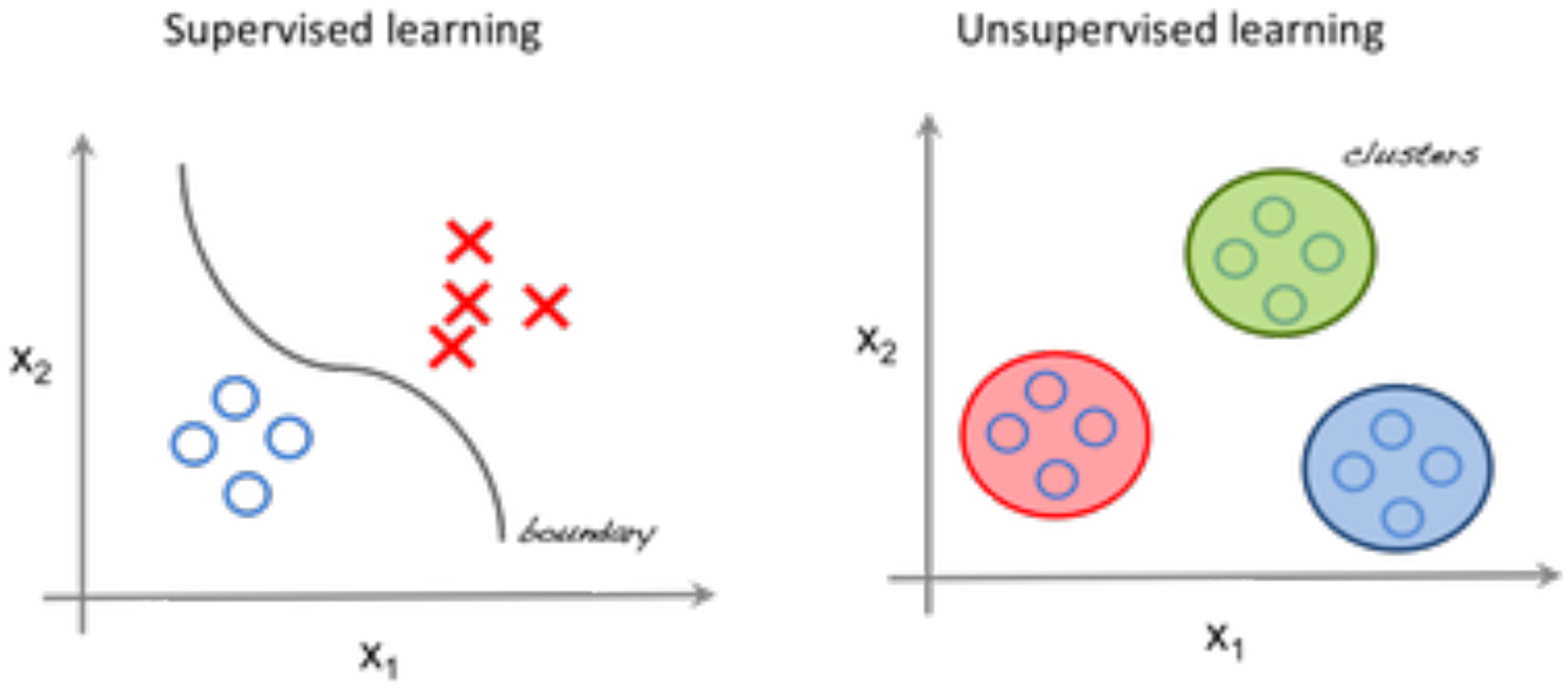
BAGGING AND BOOSTING



THANK YOU – NEXT WEEK

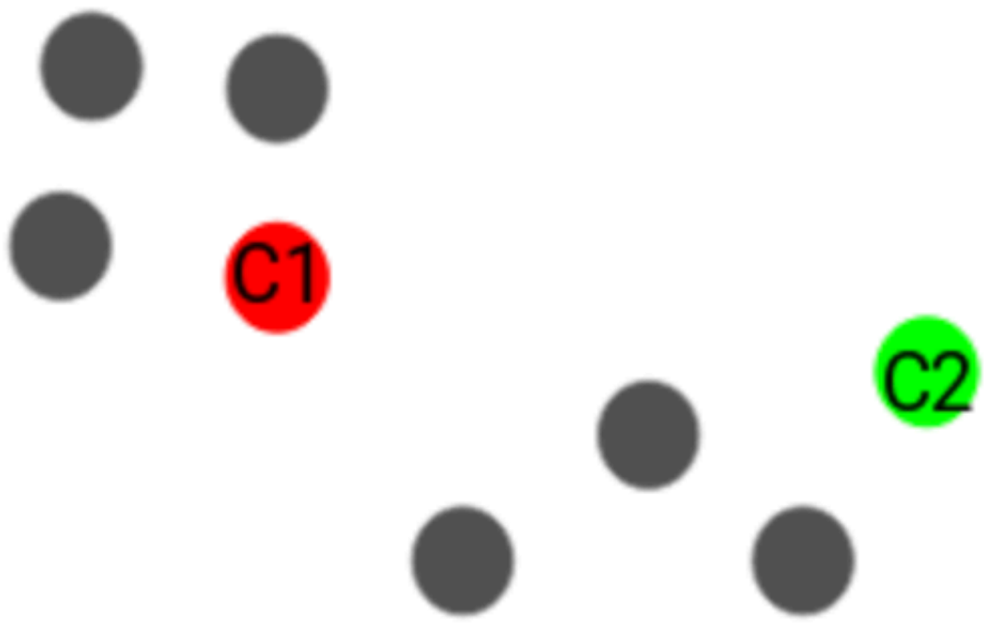
Week	Topics
Week 1	Intro to ML Discovering ML Use Cases & ML in Business
Week 2	Python- Hands On Supervised Learning & Regression
Week 3	Neural Network - 1 Neural Network -2 (Bias, Variance) & Hands ON
Week 4	Kernel Learning & SVM Practical Advice for ML projects.
Week 5	Boosting Decision Trees, Random Forest, & xgBoost
Week 6	Unsupervised Learning Clustering & Dimensionality Reduction
Week 7	Time Series Data Analysis Imputation & Prediction Systems
Week 8	ML Use Cases from Products & Research

UNSUPERVISED LEARNING

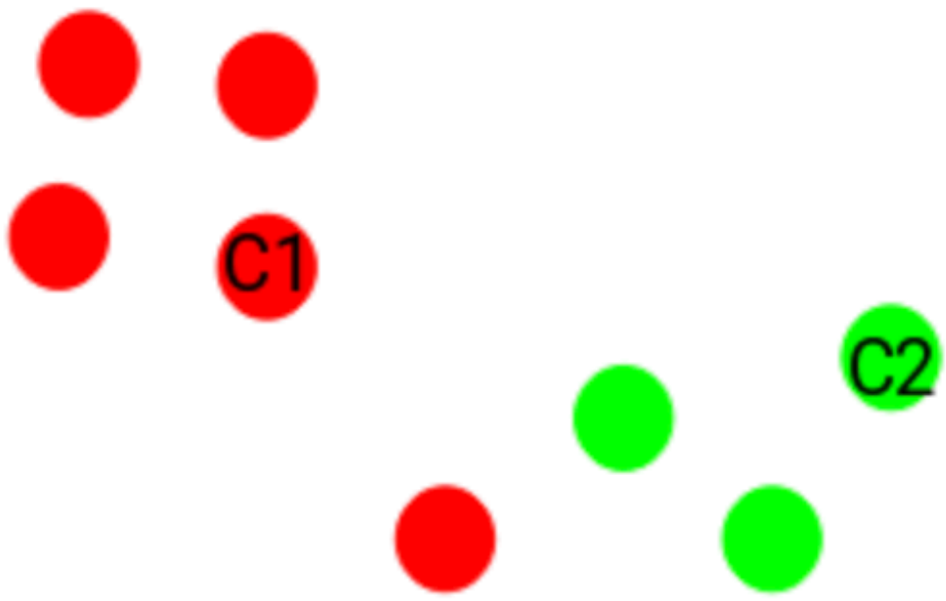


K-MEANS CLUSTERING

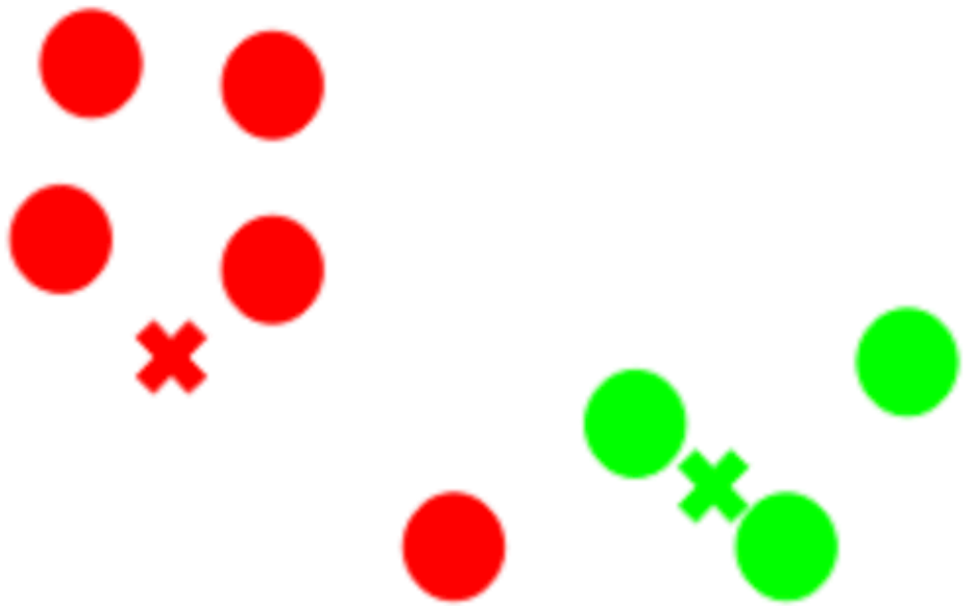
K-MEANS CLUSTERING



K-MEANS CLUSTERING

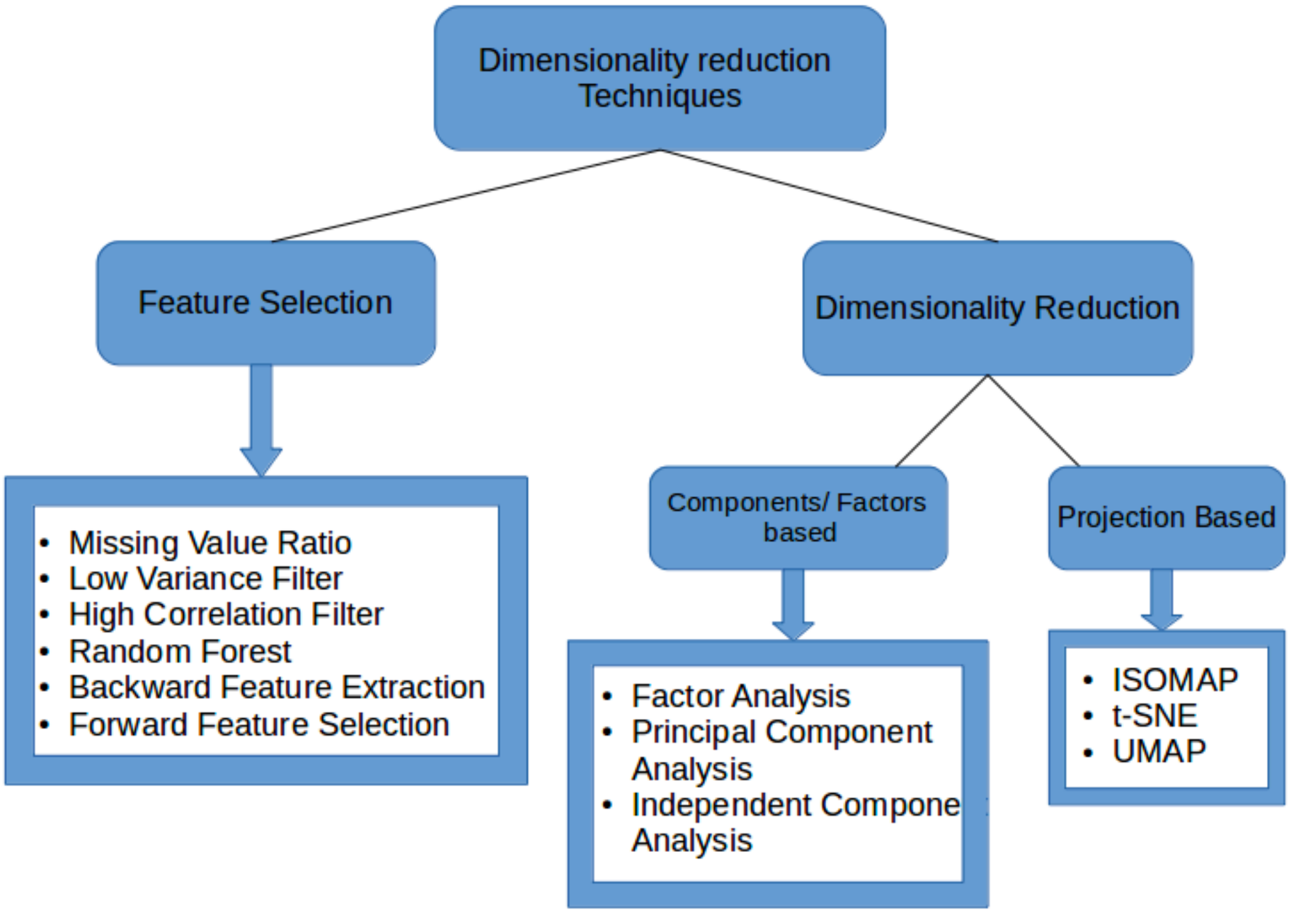


K-MEANS CLUSTERING

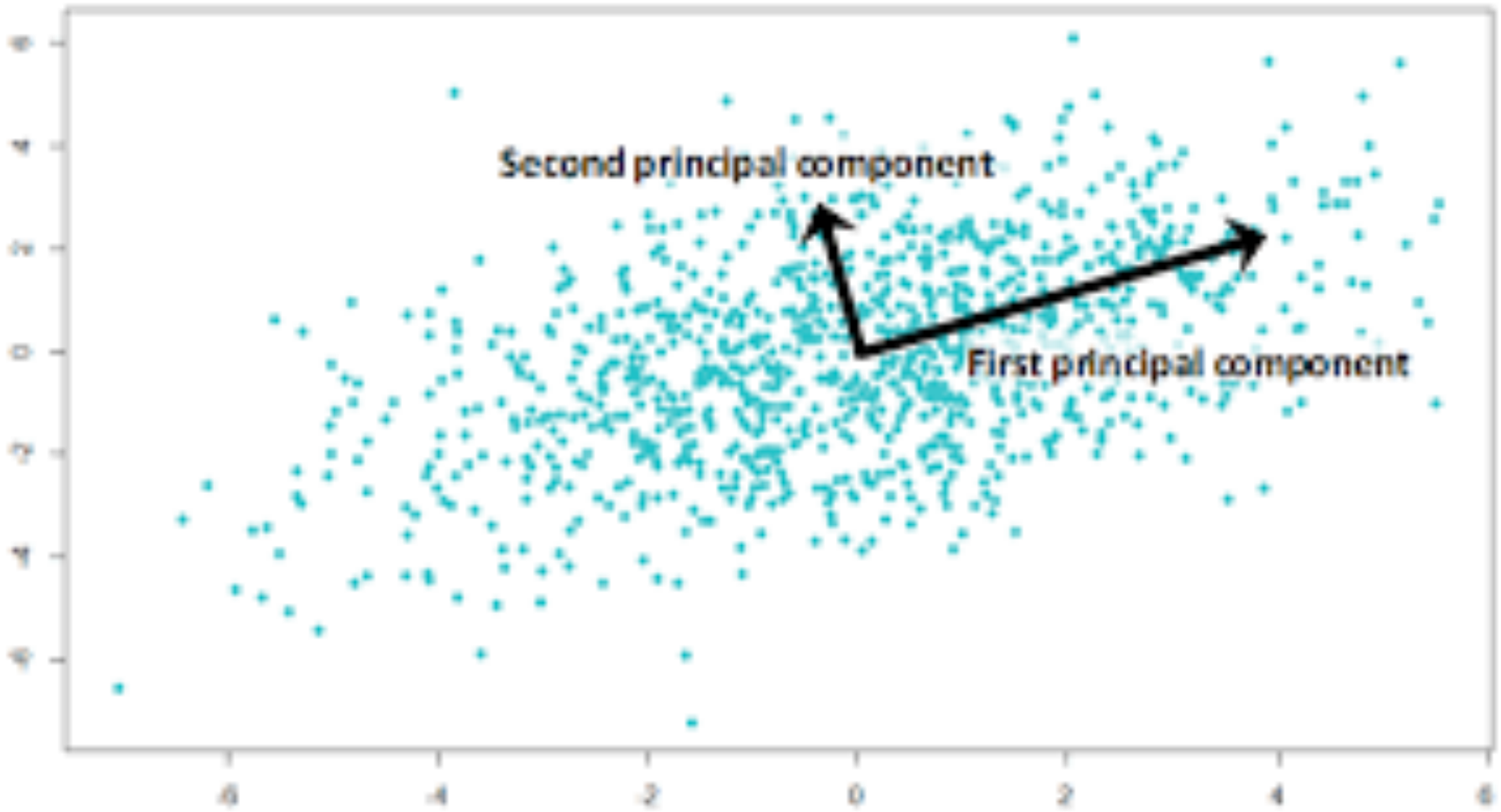


K-MEANS CLUSTERING

DIMENSIONALITY REDUCTION



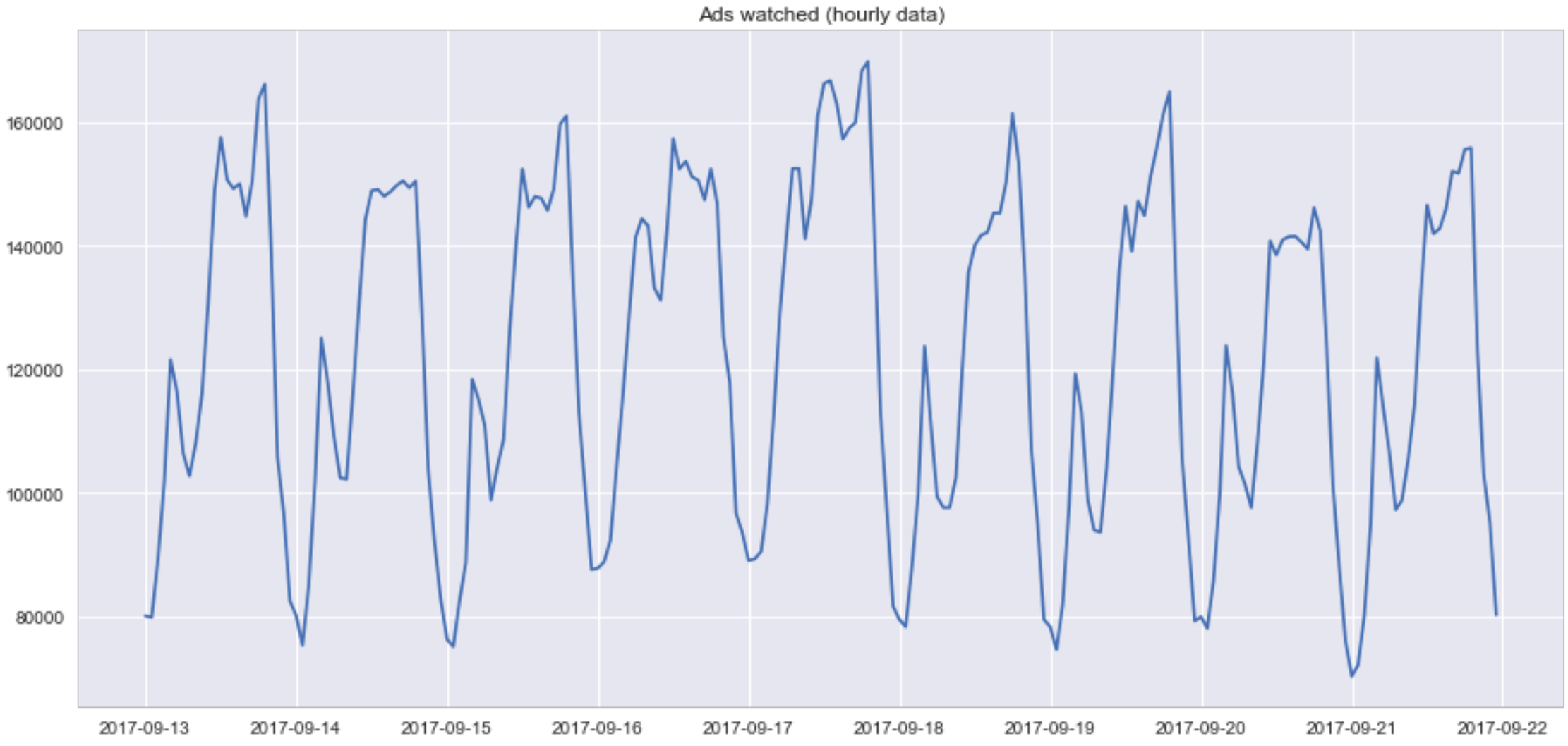
PRINCIPAL COMPONENT ANALYSIS (PCA)



THANK YOU – NEXT WEEK

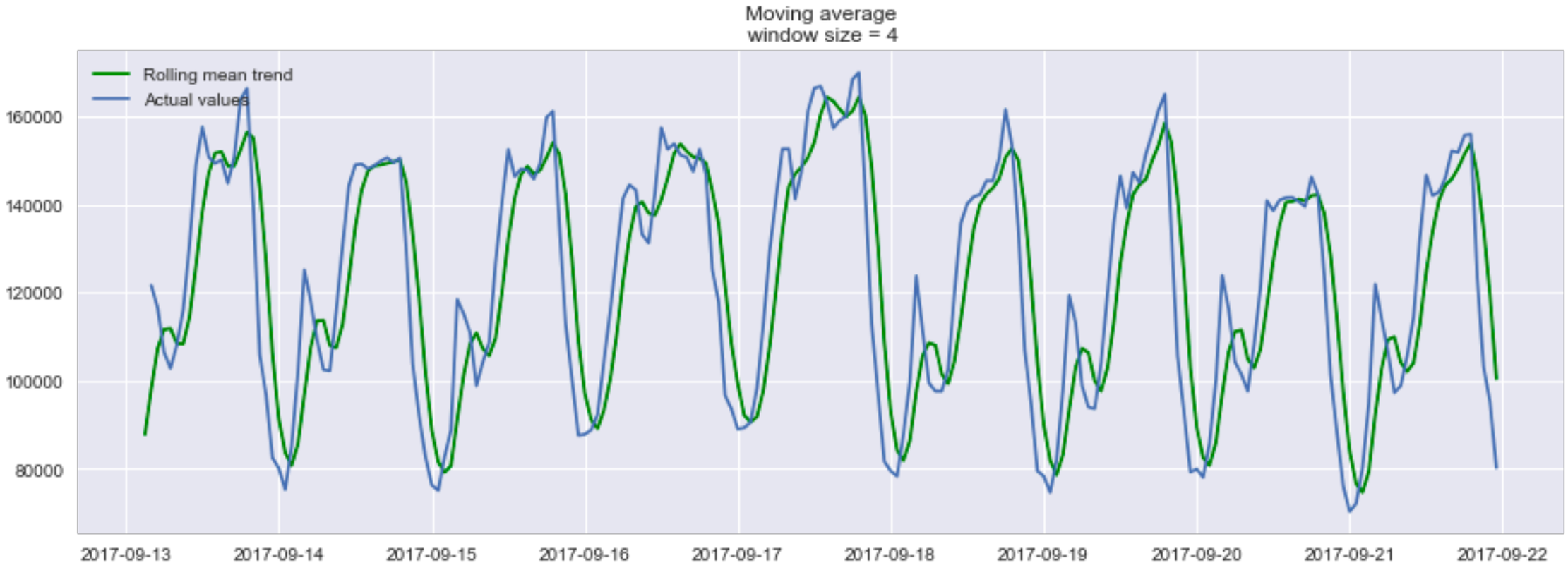
Week	Topics
Week 1	Intro to ML Discovering ML Use Cases & ML in Business
Week 2	Python- Hands On Supervised Learning & Regression
Week 3	Neural Network - 1 Neural Network -2 (Bias, Variance) & Hands ON
Week 4	Kernel Learning & SVM Practical Advice for ML projects.
Week 5	Boosting Decision Trees, Random Forest, & xgBoost
Week 6	Unsupervised Learning Clustering & Dimensionality Reduction
Week 7	Time Series Data Analysis Imputation & Prediction Systems
Week 8	ML Use Cases from Products & Research

TIME SERIES DATA



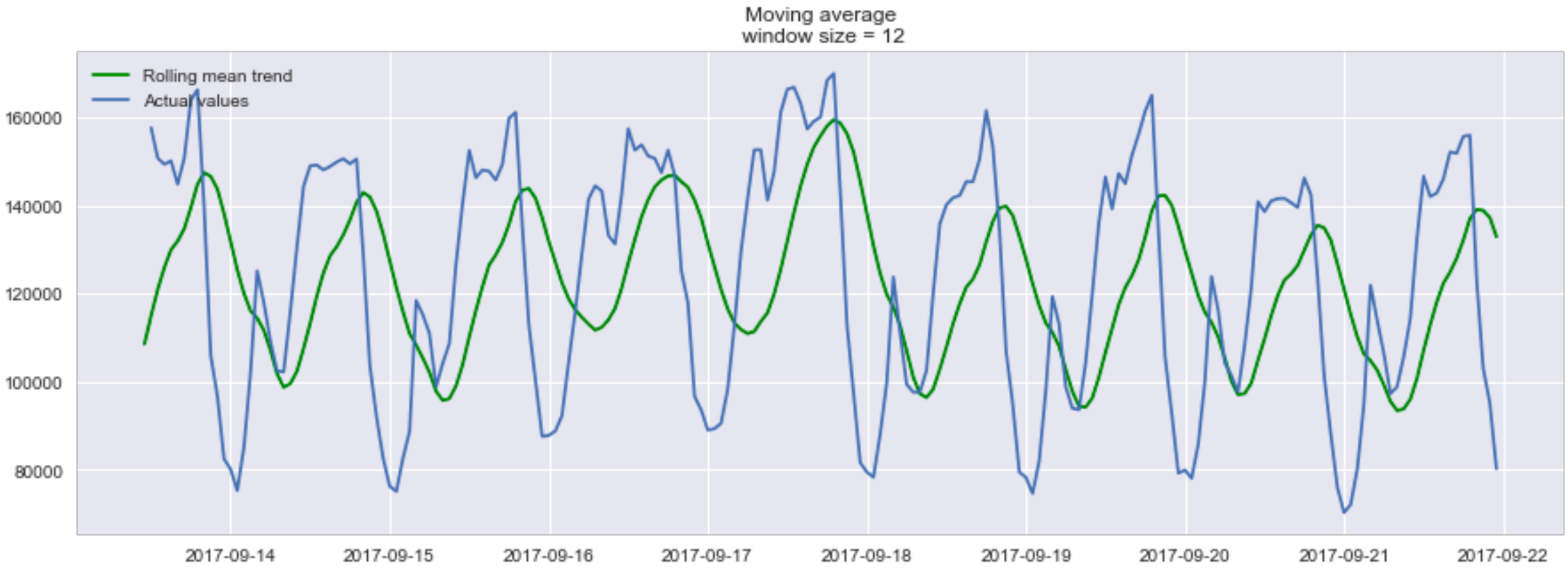
TIME SERIES DATA: MOVING AVERAGE WINDOW

$$\hat{y}_t = \sum_{n=1}^{k} \omega_n y_{t+1-n}$$



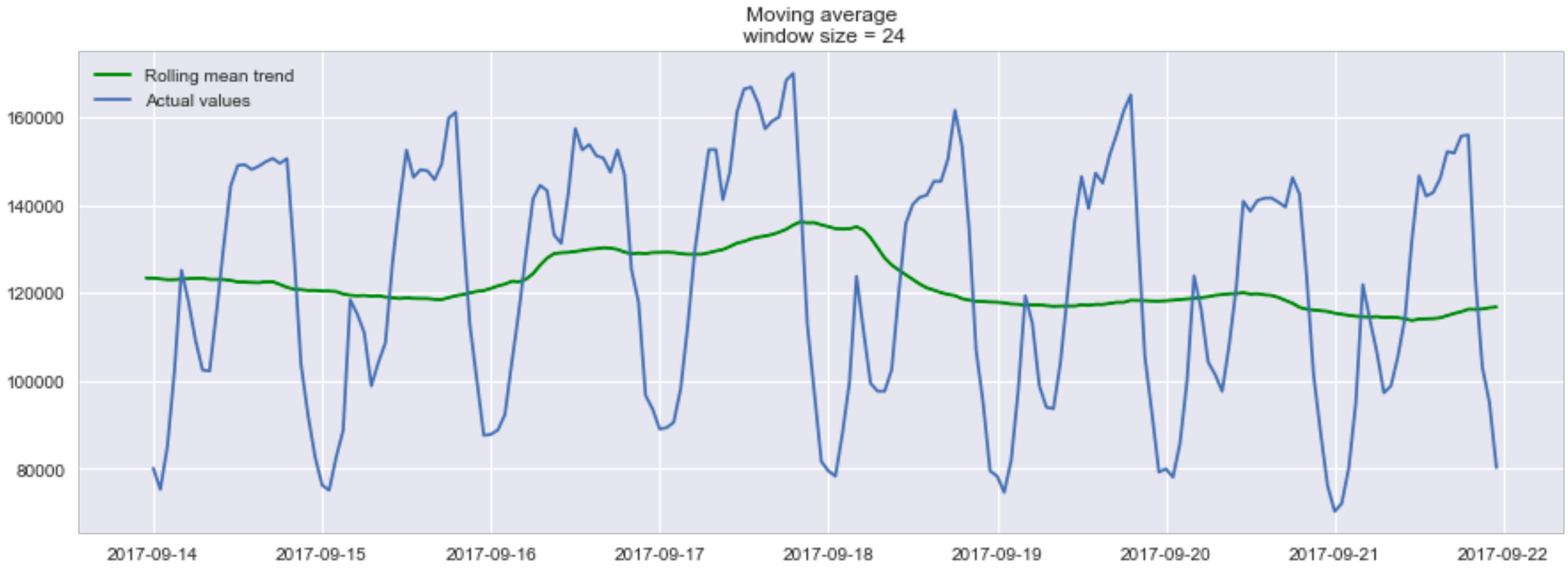
TIME SERIES DATA: MOVING AVERAGE WINDOW

$$\hat{y}_t = \sum_{n=1}^{k} \omega_n y_{t+1-n}$$

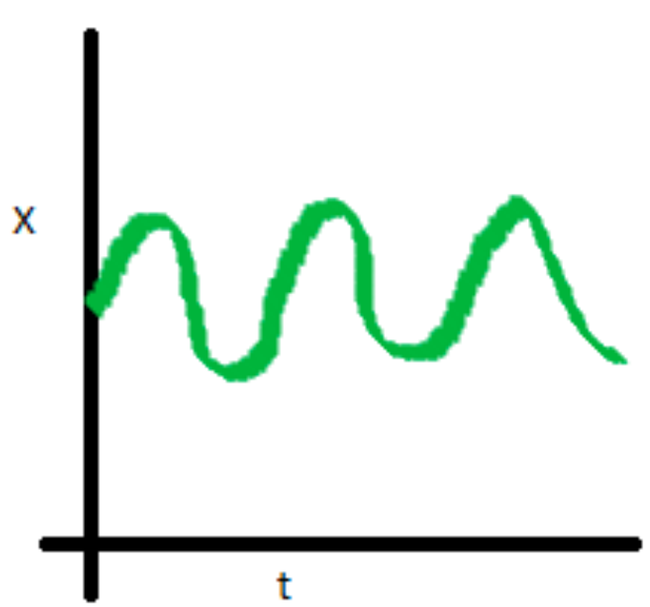


TIME SERIES DATA: MOVING AVERAGE WINDOW

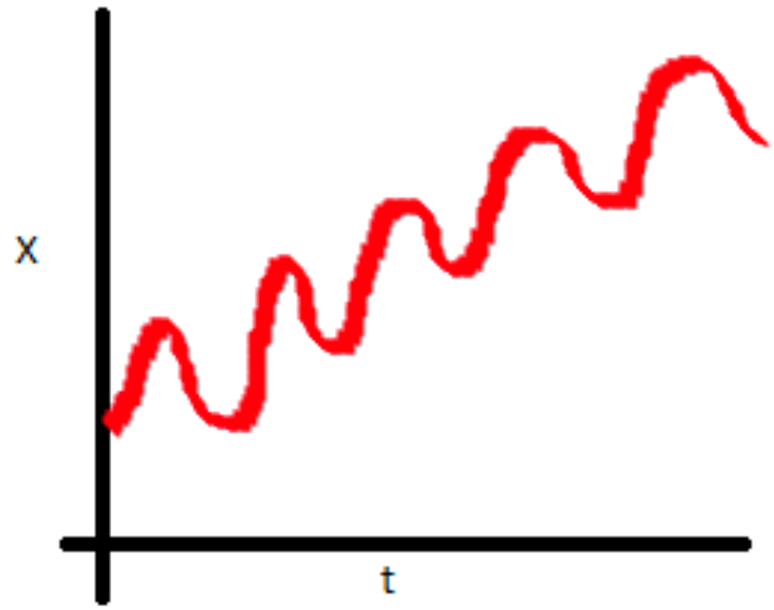
$$\hat{y}_t = \sum_{n=1}^{k} \omega_n y_{t+1-n}$$



STATIONARITY

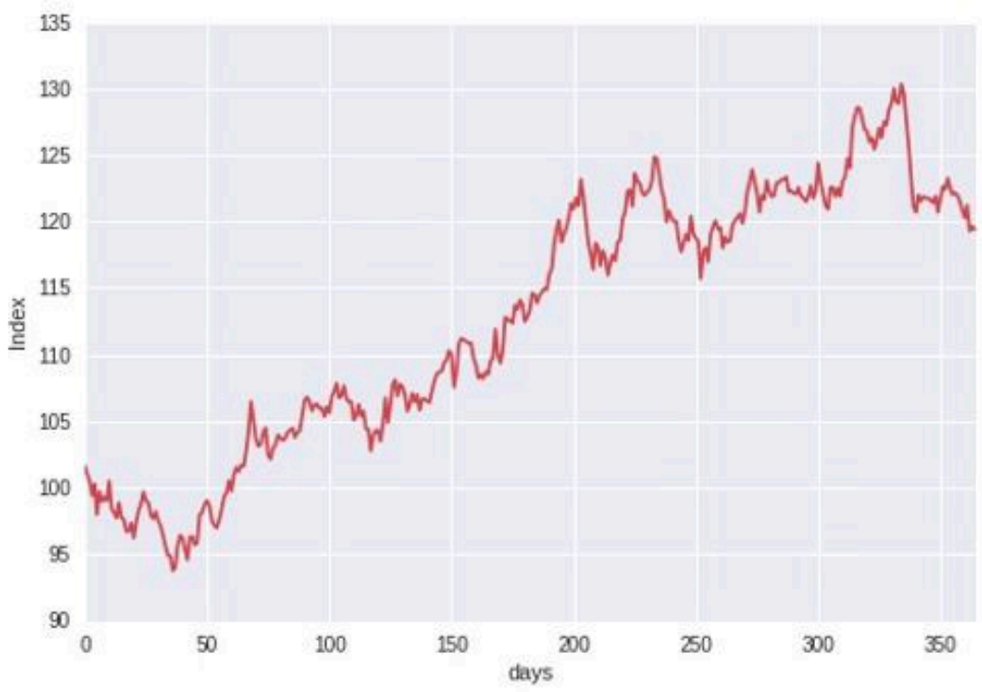


Stationary series

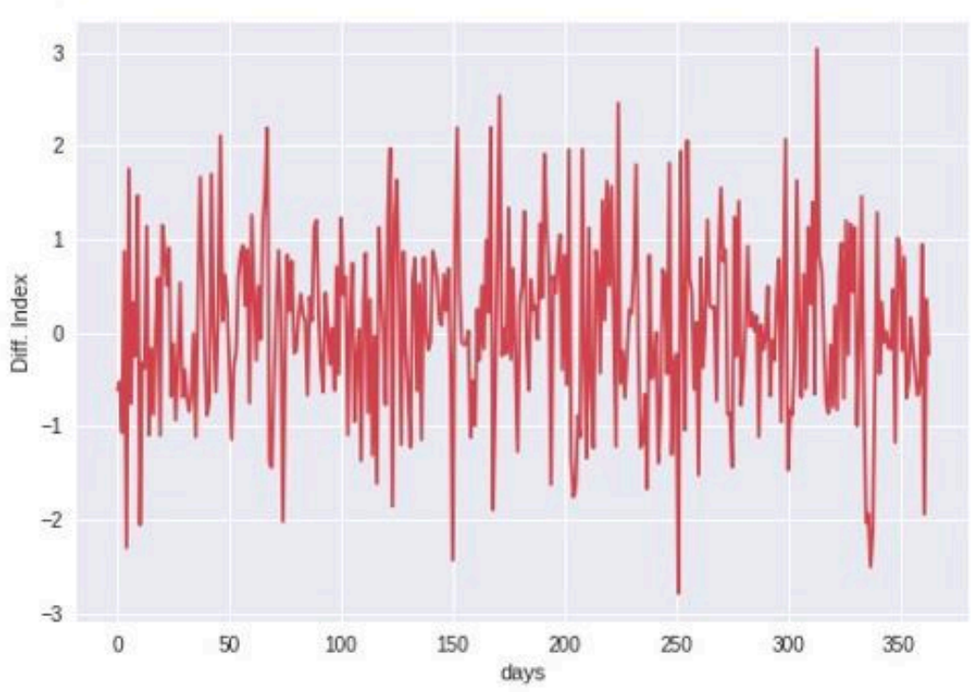


Non-Stationary series

STATIONARITY: TIME DIFFERENCING



Time differencing



ARIMA MODEL

- **Step 1 — Check stationarity:** If a time series has a trend or seasonality component, it must be made stationary before we can use ARIMA to forecast. .
- **Step 2 — Difference:** If the time series is not stationary, it needs to be stationarized through differencing. Take the first difference, then check for stationarity. Take as many differences as it takes. Make sure you check seasonal differencing as well.
- **Step 3 — Filter out a validation sample:** This will be used to validate how accurate our model is. Use train test validation split to achieve this
- **Step 4 — Select AR and MA terms:** Use the ACF and PACF to decide whether to include an AR term(s), MA term(s), or both.
- **Step 5 — Build the model:** Build the model and set the number of periods to forecast to N (depends on your needs).
- **Step 6 — Validate model:** Compare the predicted values to the actuals in the validation sample.

THANK YOU – THAT'S ALL FOLKS

Week	Topics
Week 1	Intro to ML Discovering ML Use Cases & ML in Business
Week 2	Python- Hands On Supervised Learning & Regression
Week 3	Neural Network - 1 Neural Network -2 (Bias, Variance) & Hands ON
Week 4	Kernel Learning & SVM Practical Advice for ML projects.
Week 5	Boosting Decision Trees, Random Forest, & xgBoost
Week 6	Unsupervised Learning Clustering & Dimensionality Reduction
Week 7	Time Series Data Analysis Imputation & Prediction Systems
Week 8	ML Use Cases from Products & Research