Boosting a Signal

Using Trees and Forests for Classification Tasks

Evan Simpson

p.evansimpson@gmail.com

EducatorsRLearners

Agenda

- Introduction
- Review of Machine Learning Types
- Decision Trees
- Random Forest
- XGBoost
- Future Directions


HELLO,
My Name Is..


18th Annual MwALT Conference Purdue University
October 1st 2016
Iowa State University
English Department: Applied Linguistics and Technology
Shifting Baselines: Changing from NRT to CRT for Students' Placement Tests
Strength of NRT
NRT Requires
Evan Simpson


USC
Lisa Deng


Model 3: Computer Vision
Can you judge a book by its cover?


HELLO,
My Name Is..

Activity Time!

Match the terms with their definitions

Check your Answers

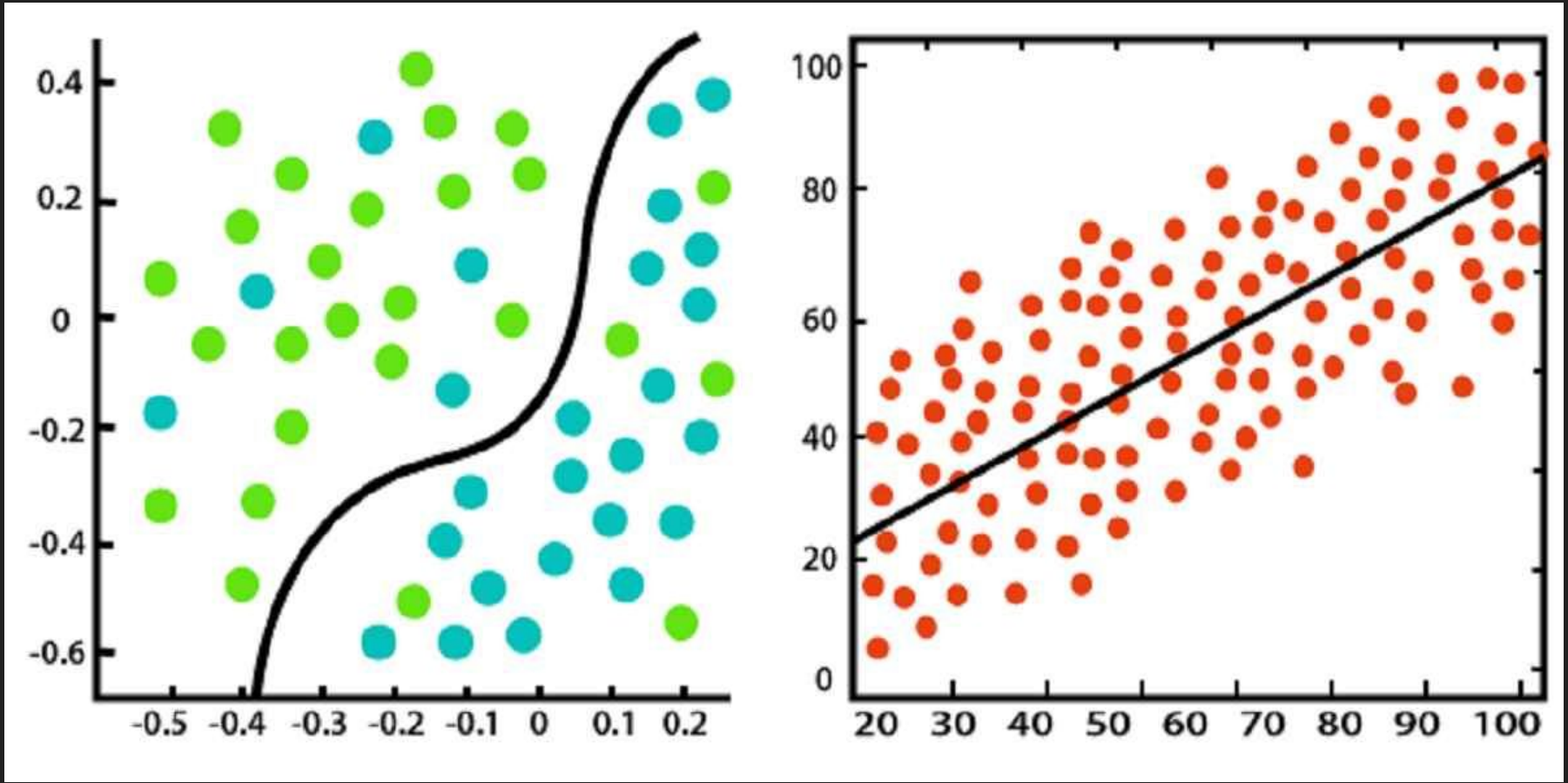
F, D, C, B, E, A

Types of Machine Learning

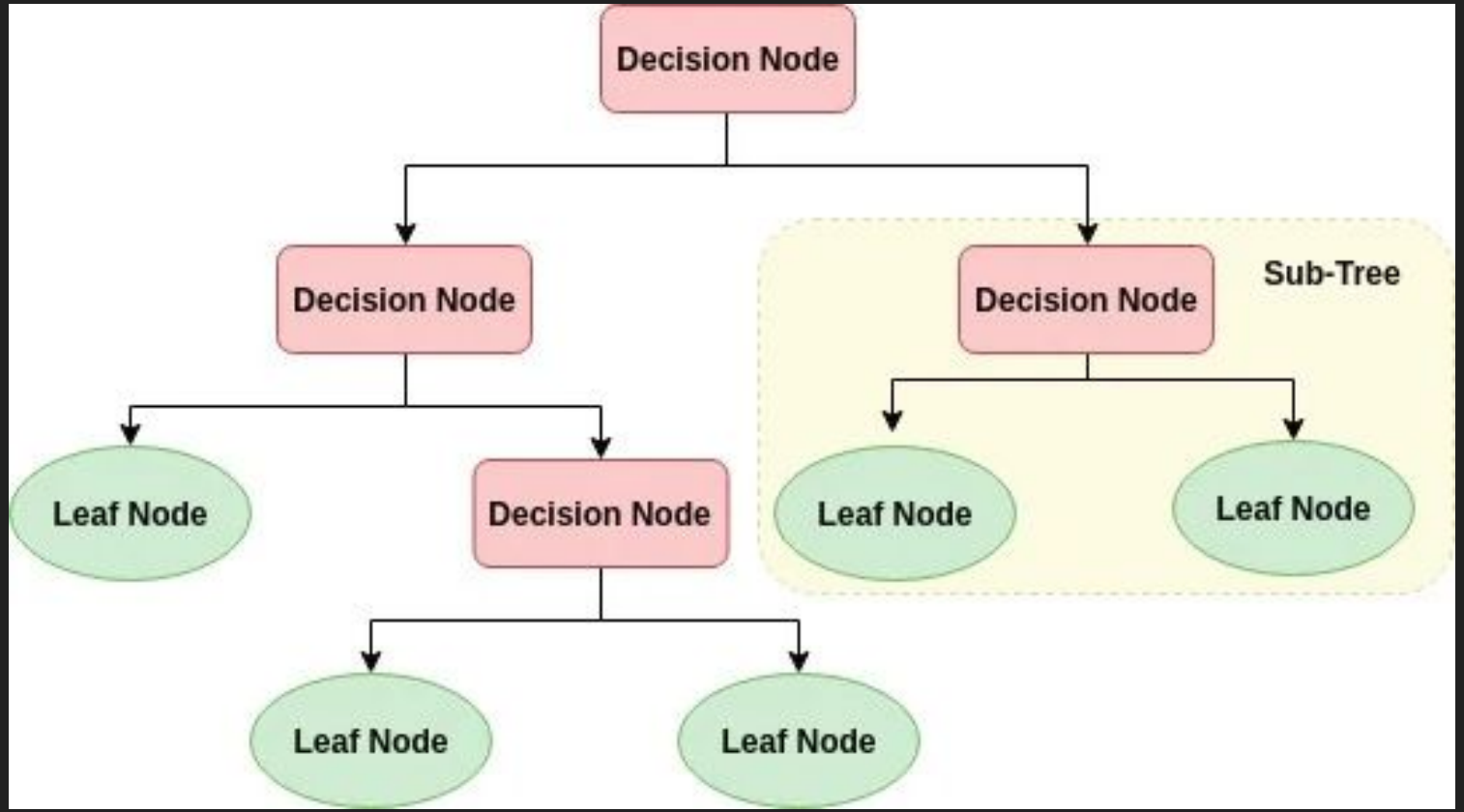
- Unsupervised?
- Supervised?

Why?

Types of Machine Learning

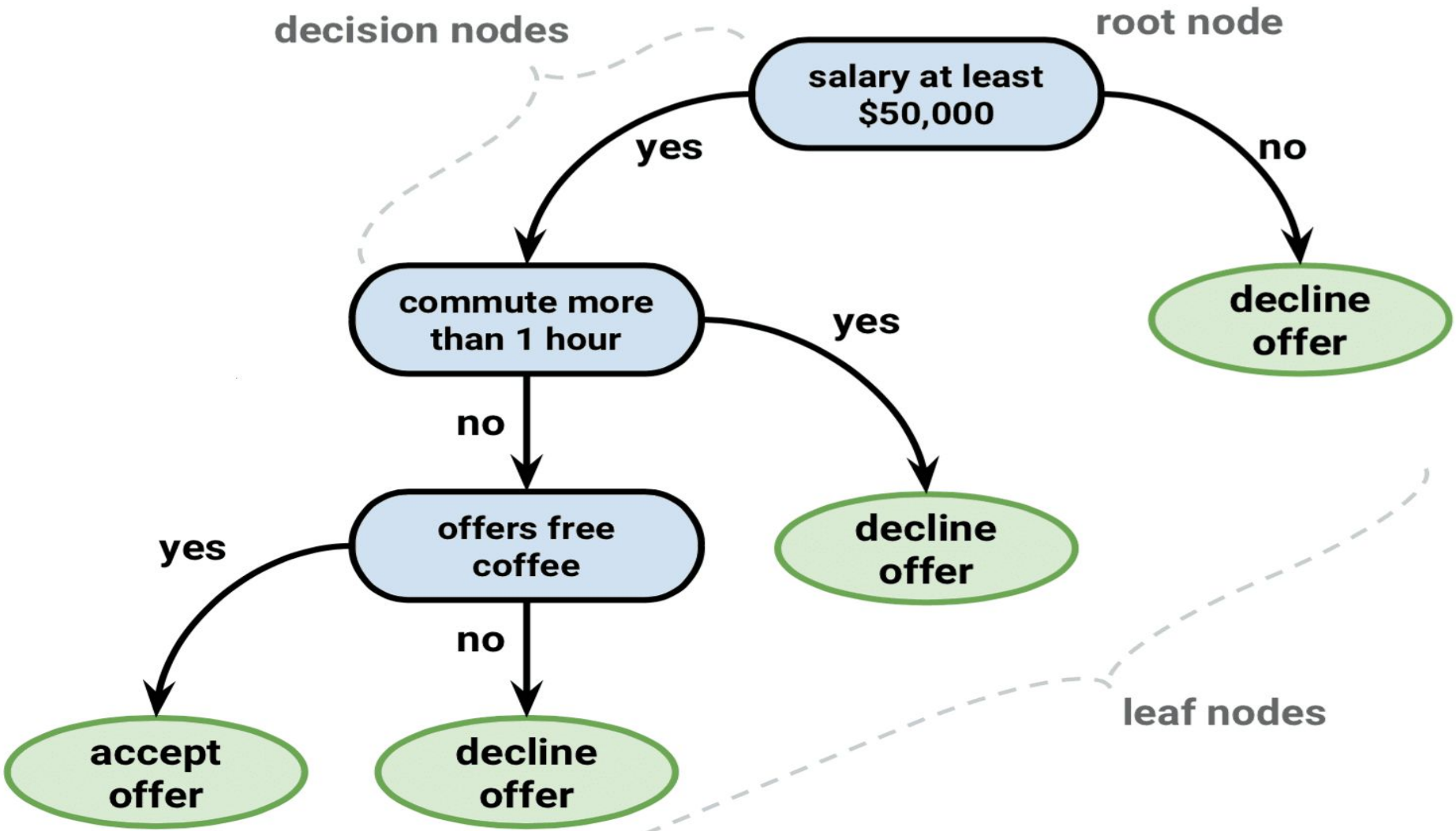


Decision Trees



Decision Trees

Example: Job Offer


decision nodes
root node
salary at least $50,000
yes
no
commute more than 1 hour
yes
decline offer
no
offers free coffee
decline offer
yes
no
accept offer
decline offer
leaf nodes

Decision Trees

Advantage

-Easy to interpret and use

Disadvantage

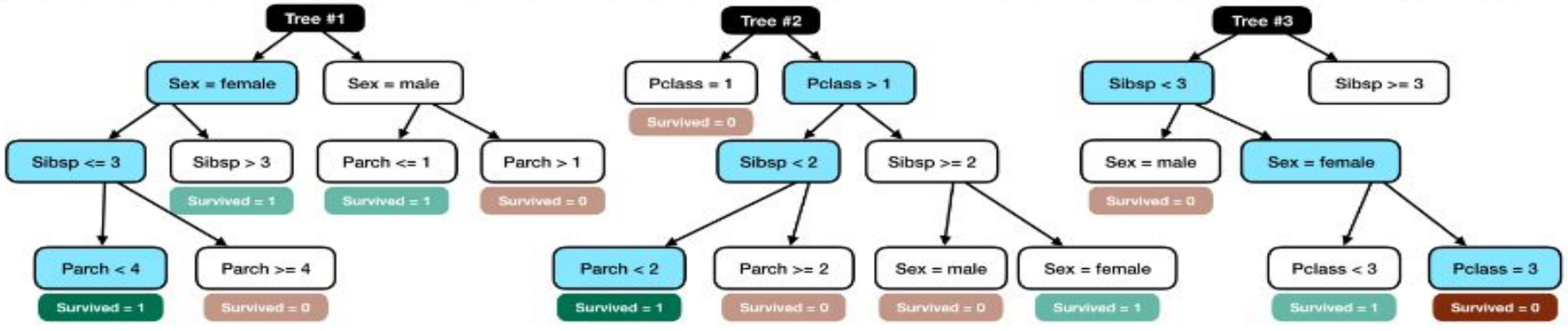
- Easy to overfit

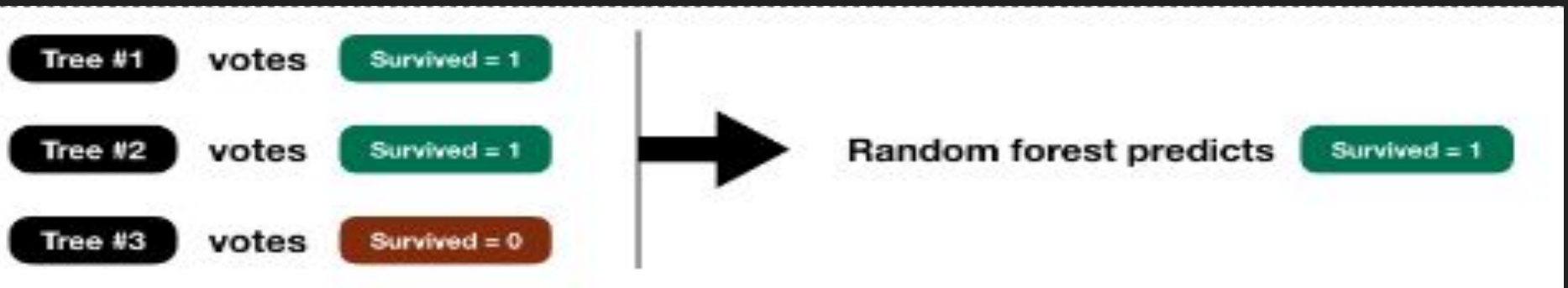
Ensemble Methods

Random Forest

Did the passenger survive?

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
893	3	Wilkes, Mrs. James (Ellen Needs)	female	47	1	0	363272	7		S



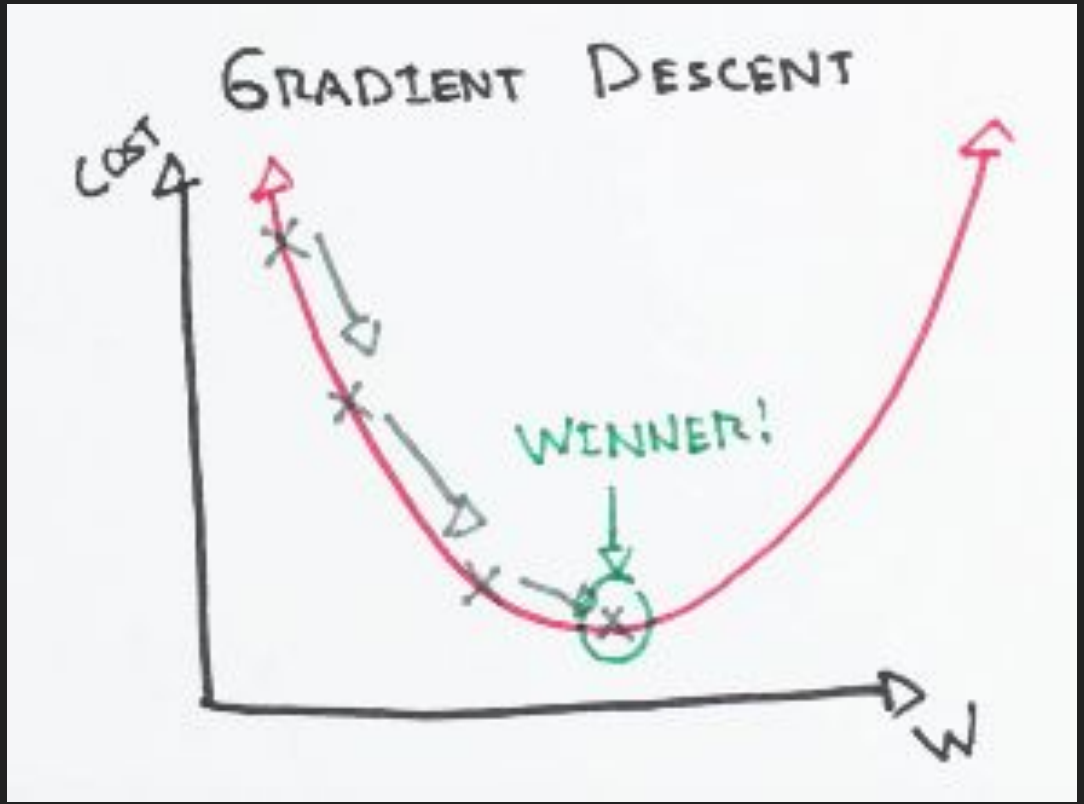


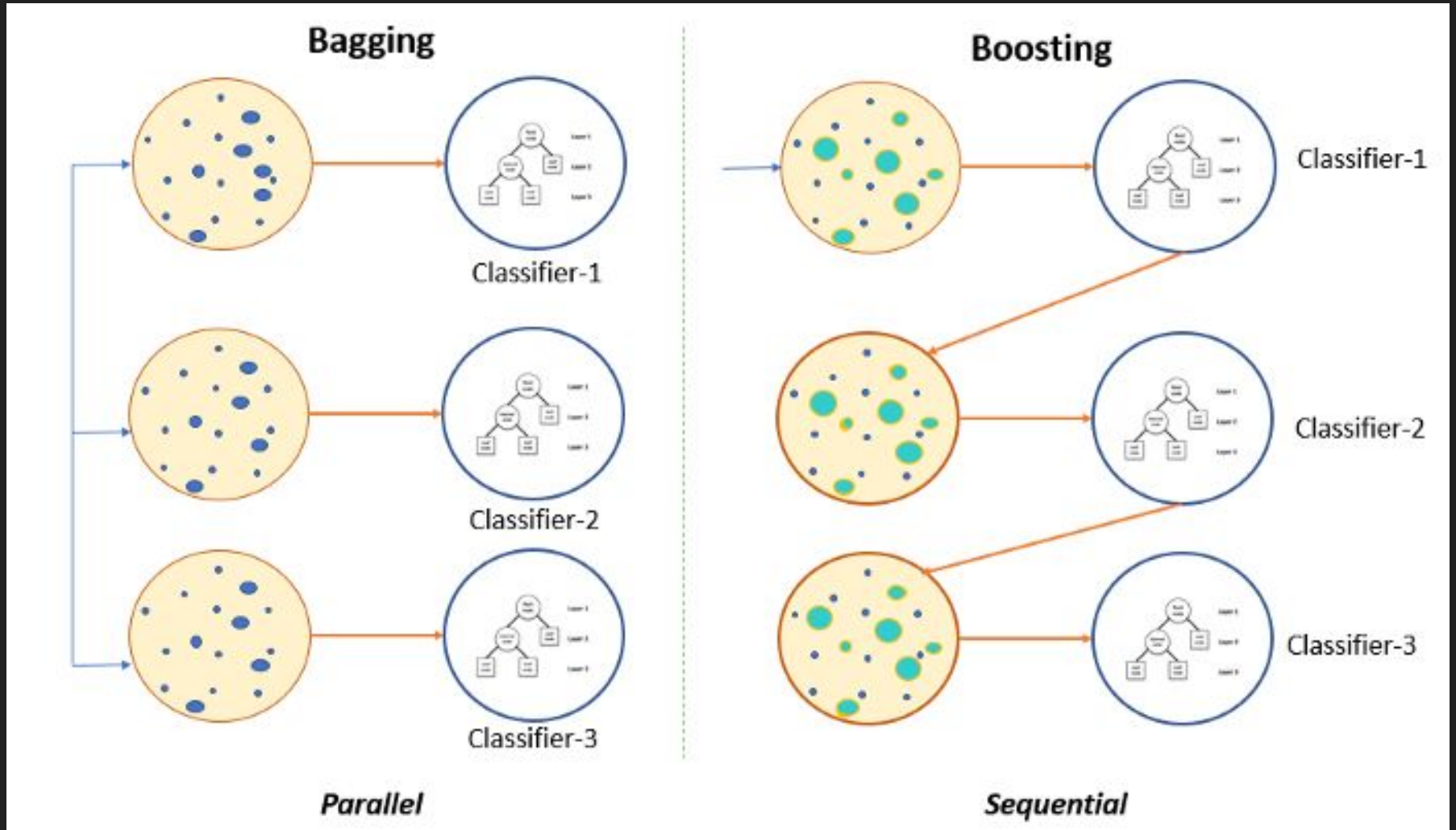
Random Forest

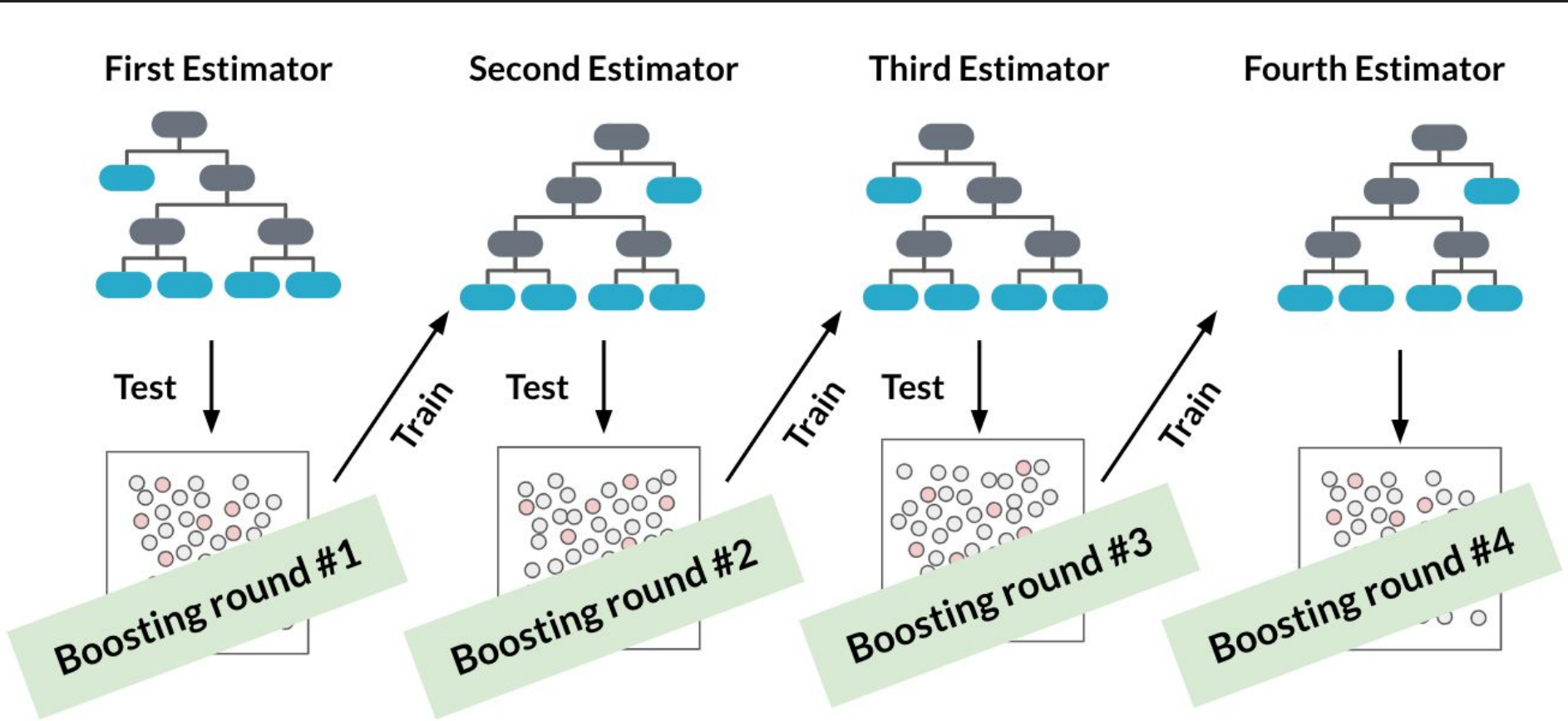
+ Incredibly Powerful

- Trees are built randomly


dmlc
XGBoost


GRADIENT DESCENT
COST
WINNER!
W


Bagging
Classifier-1
Classifier-2
Classifier-3
Parallel
Boosting
Classifier-1
Classifier-2
Classifier-3
Sequential


First Estimator
Second Estimator
Third Estimator
Fourth Estimator
Test
Train
Test
Train
Test
Train
Boosting round #1
Boosting round #2
Boosting round #3
Boosting round #4

Activity Time Part 2:

survived	0,1
pclass	1, 2, 3
sibsp	# of siblings
parch	# of parents
fare	price of ticket
is_female	yes or no

Sample Workflow

Frame the problem

Install and Load the Libraries

Collect/Load the Data

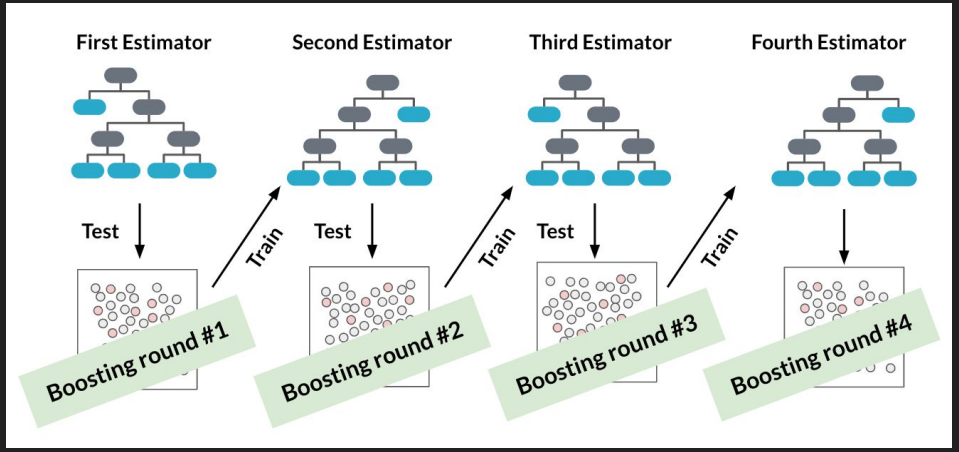
Conduct Exploratory Data Analysis (EDA)

Prepare the Data

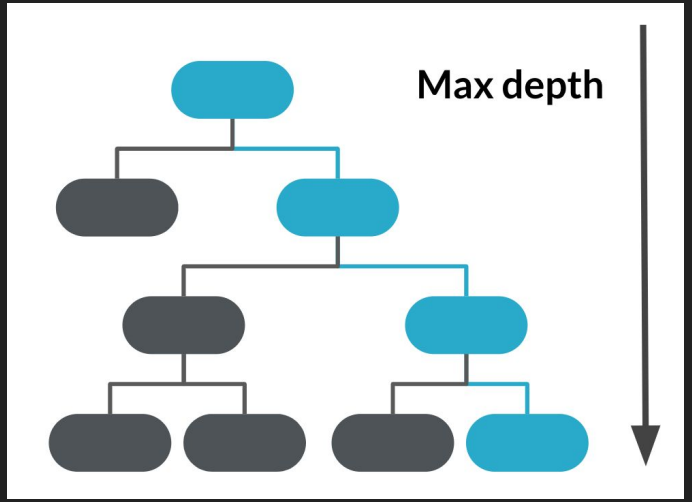
Build and Evaluate the Model

Inspect the Feature Importance

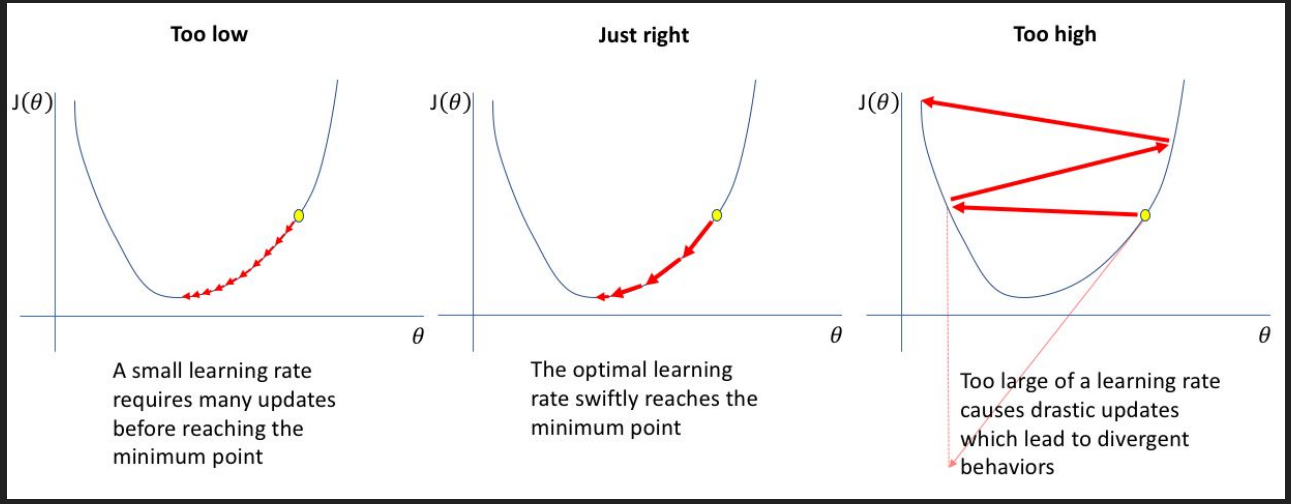
n_estimators



max_depth



eta (aka, “learning rate”)



early_stopping

Next Steps

OPTUNA

LightGBM

CatBoost

Resources

Kaggle Tutorial using Random Forest
DataCamp Decision Tree Tutorial
DataCamp XGBoost Tutorial
Hands On Machine Learning Chapters 6 and 7
Numsense! Data Science for the Layman
Medium Reading list based on this topic
YouTube Playlist
Repo for this Lesson