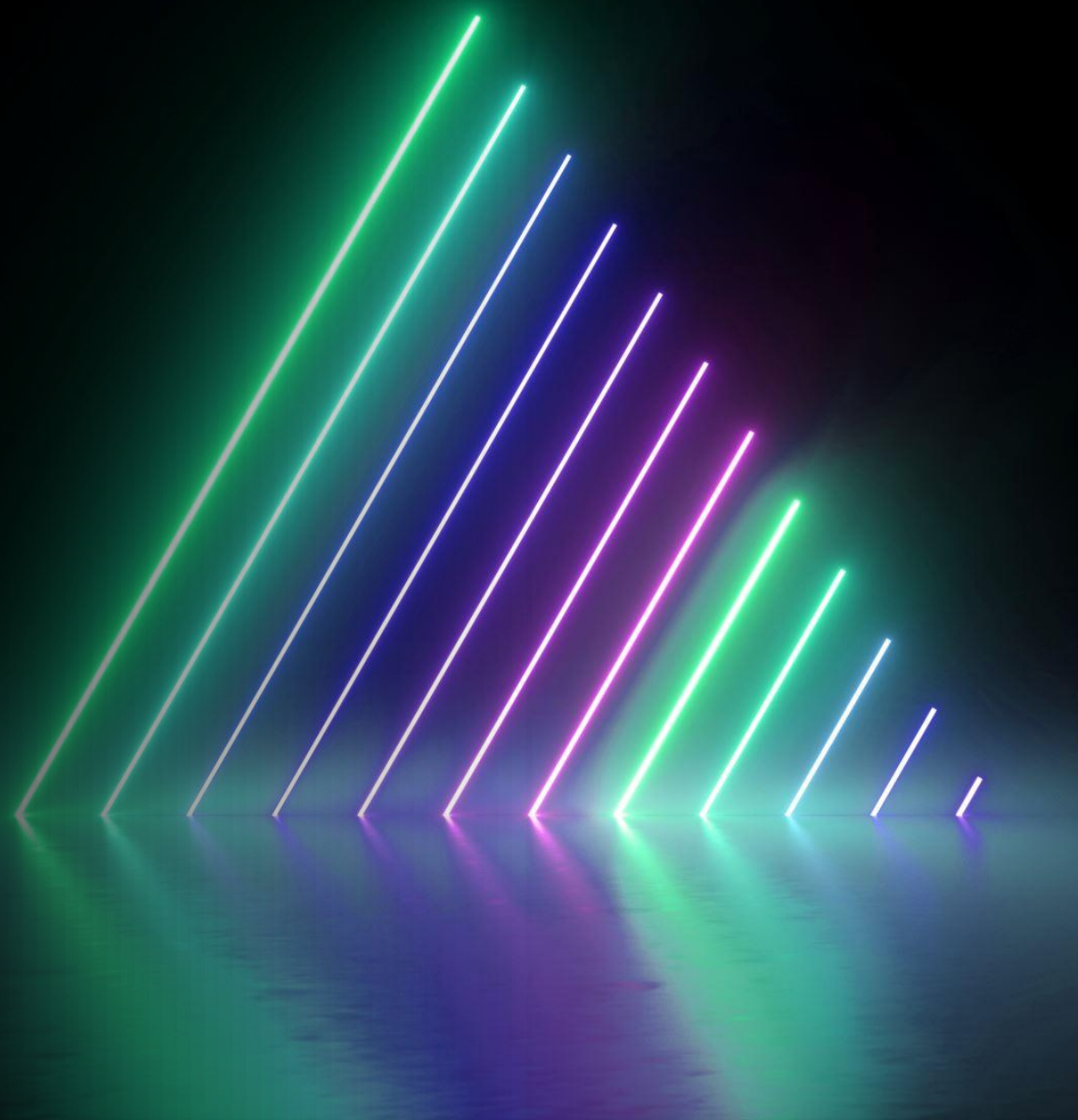
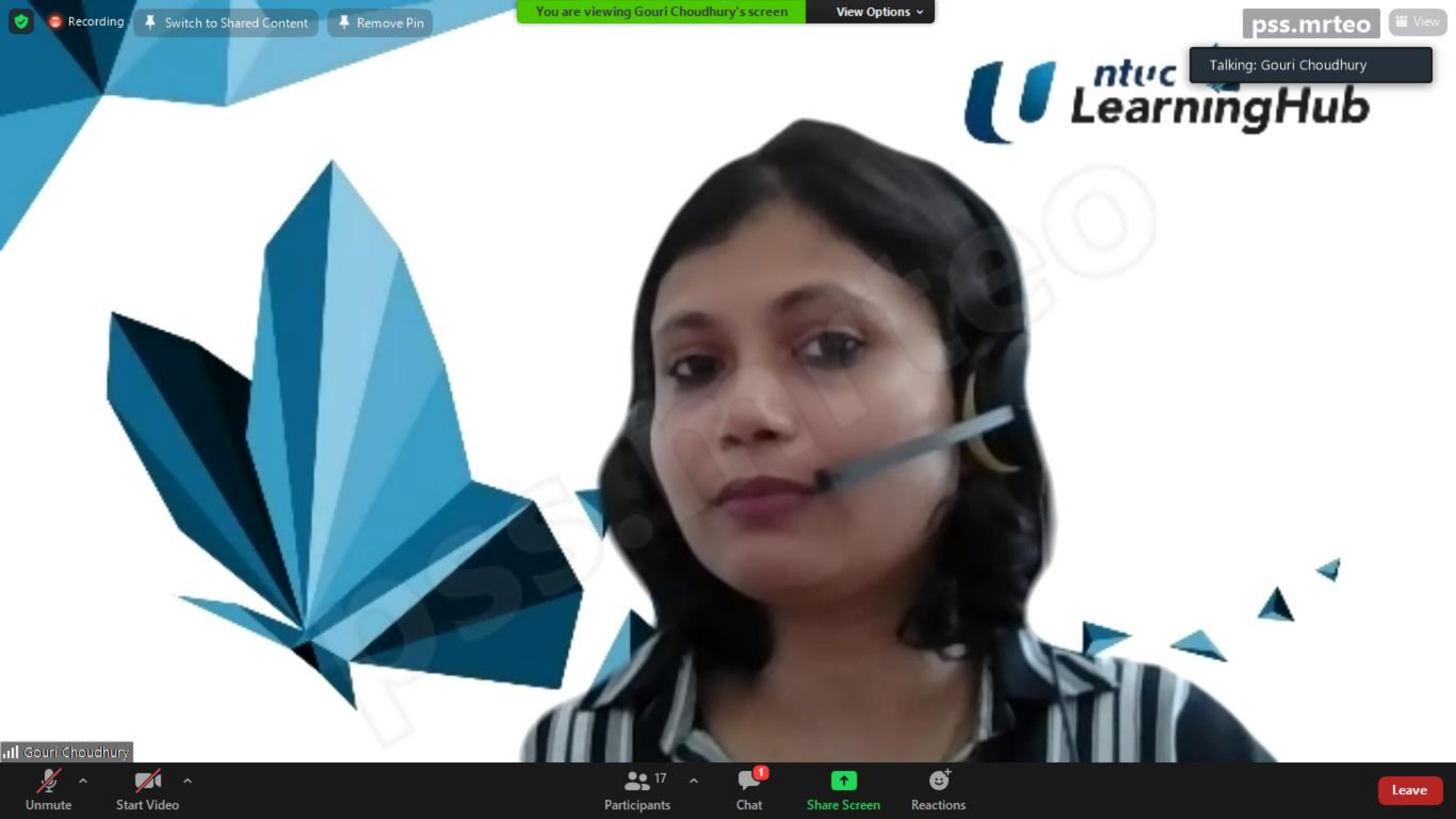
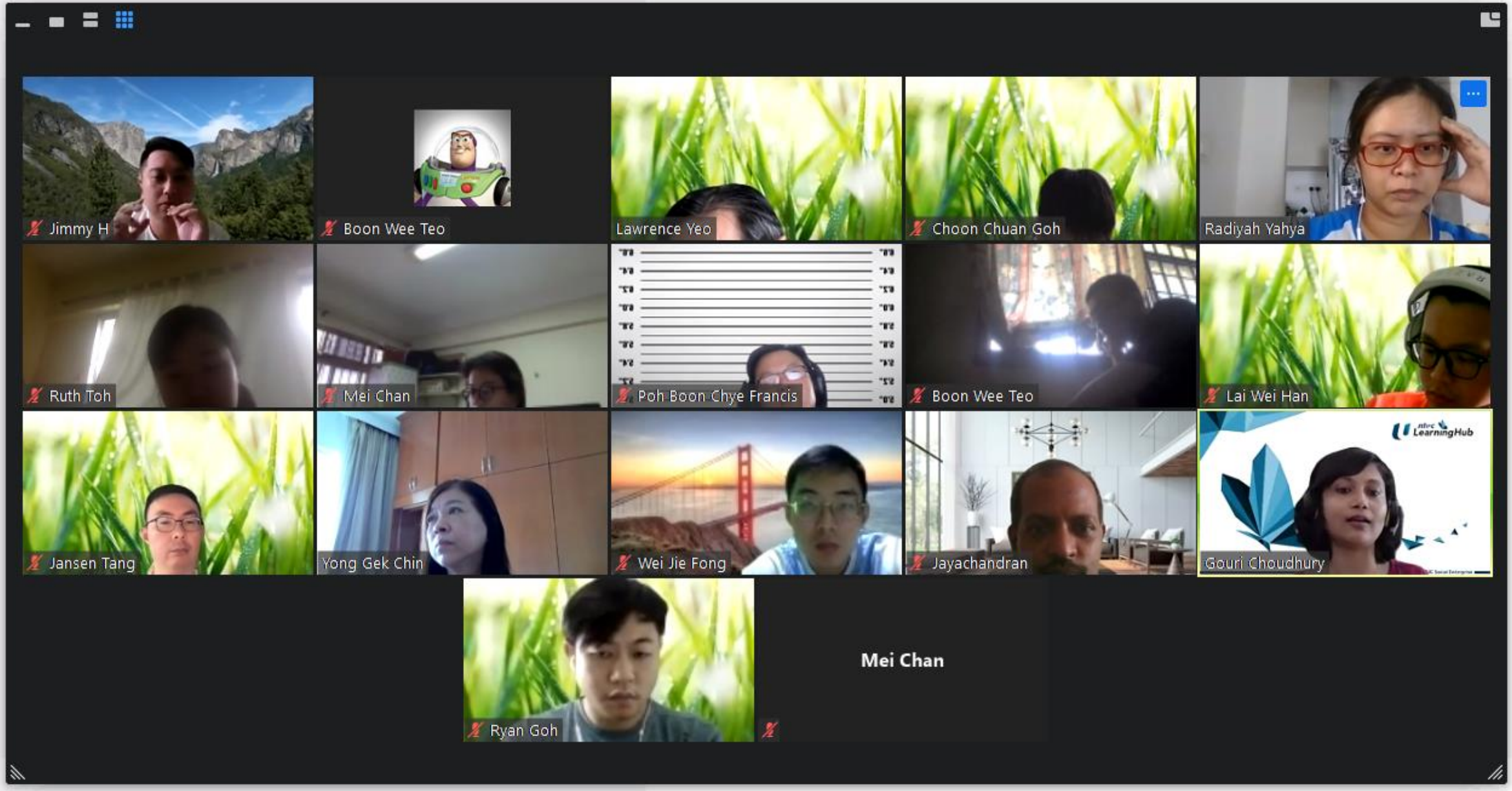
Machine Learning and Advanced Analytics Using Python



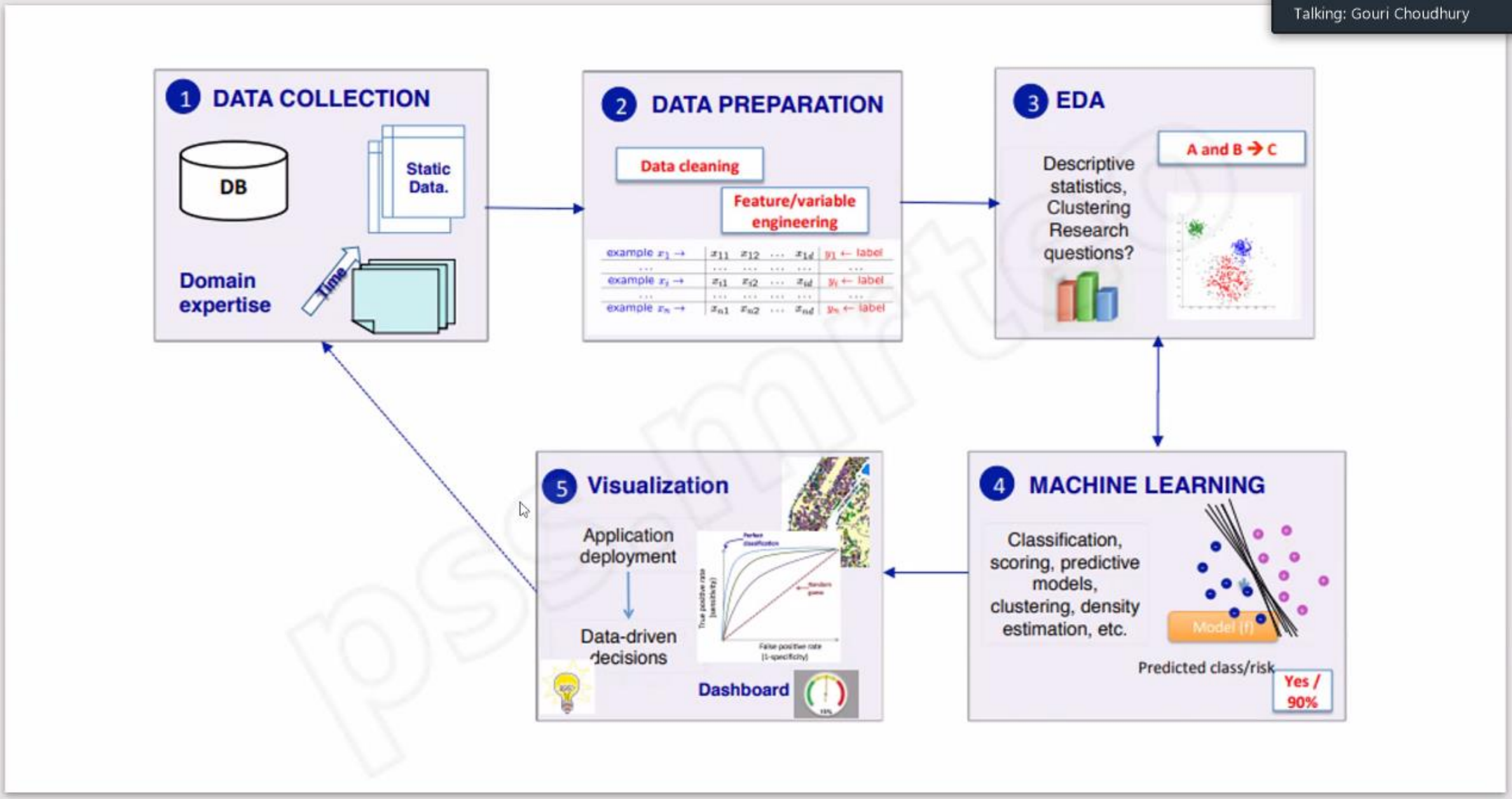

Recording
Switch to Shared Content
Remove Pin
You are viewing Gouri Choudhury's screen
View Options
pss.mrteo
View
Talking: Gouri Choudhury
ntuc
LearningHub
Gouri Choudhury
Unmute
Start Video
17
Participants
Chat
Share Screen
Reactions
Leave


Recording
pss.mrteo
ntuc LearningHub
Talking:
Gouri Choudhury
Enterprise


Jimmy H
Boon Wee Teo
Lawrence Yeo
Choon Chuan Goh
Radiyah Yahya
Ruth Toh
Mei Chan
Poh Boon Chye Francis
Boon Wee Teo
Lai Wei Han
Jansen Tang
Yong Gek Chin
Wei Jie Fong
Jayachandran
Gouri Choudhury
Ryan Goh
Mei Chan


Remove Pin
pss.mrteo
View
Talking:
Ruth Toh
Unmute
Start Video
17
Participants
Chat
Share Screen
Reactions
Leave

Recording
pss.mrteo
Talking: Gouri Choudhury



pss.mrteo
Talking: Gouri Choudhury

What is Learning?

1

"Learning denotes changes in a system that ... enable a system to do the same task ... more efficiently the next time." - Herbert Simon

2

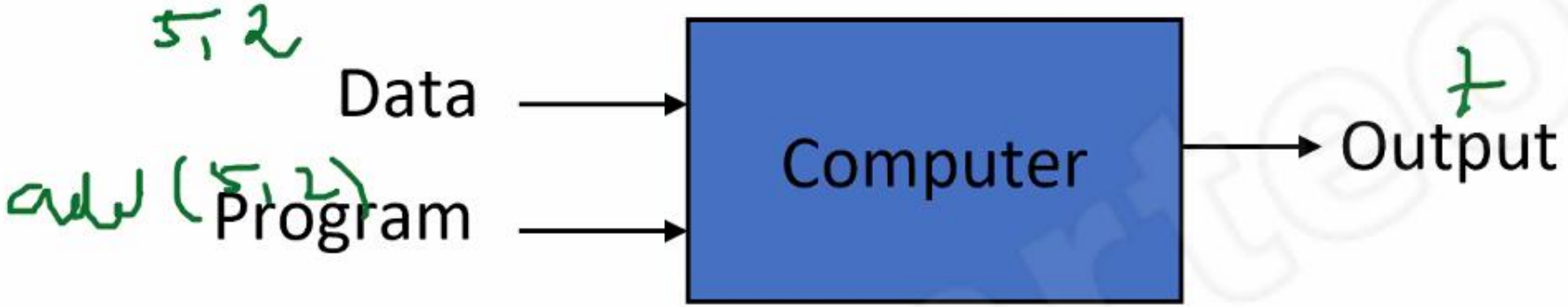
"Learning is constructing or modifying representations of what is being experienced." - Ryszard Michalski

3

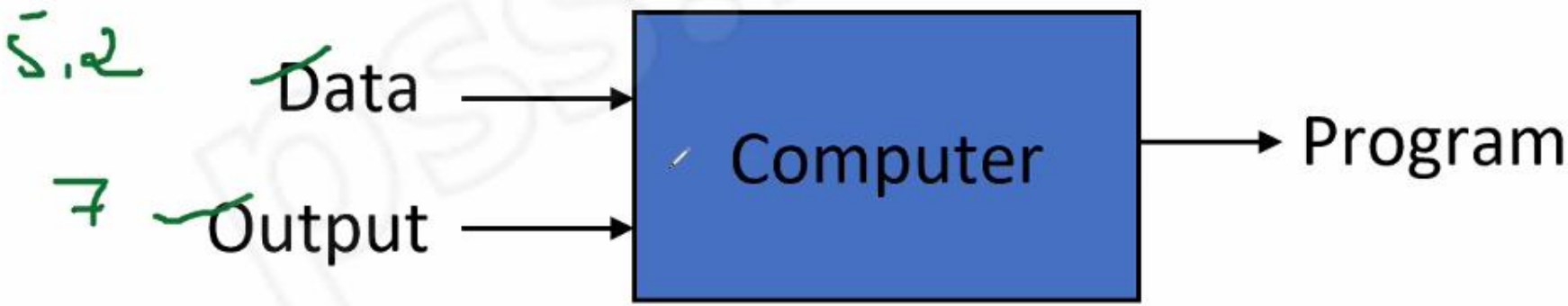
"Machine learning refers to a system capable of the autonomous acquisition and integration of knowledge."

pss.mrteo
Talking: Gouri Choudhury

Traditional Programming



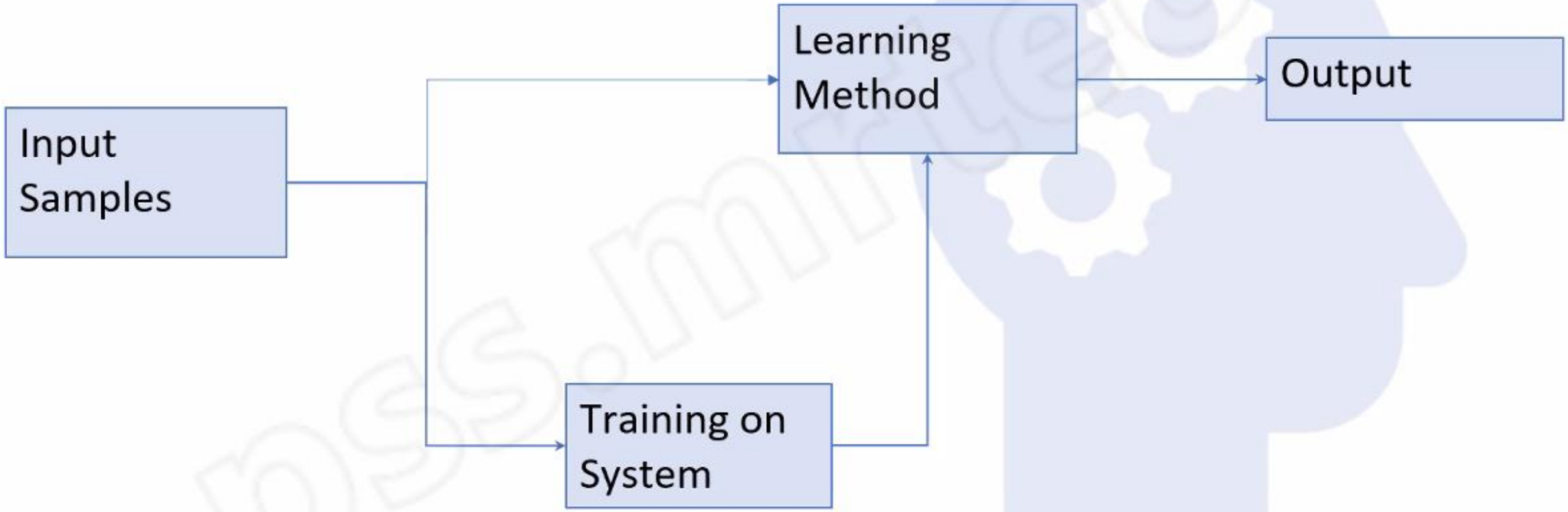
Machine Learning



pss.mrteo

Talking: Gouri Choudhury

Learning System Model



Recording

pss.mrteo

Talking: Gouri Choudhury

Why is machine learning required?

pssmrteo

pss.mrteo

Talking: Gouri Choudhury

Some examples that machine learning solves

Recognizing patterns:

- Facial identities or facial expressions
- Handwritten or spoken words
- Medical images

Generating patterns:

- Generating images or motion sequences

Recognizing anomalies:

- Unusual credit card transactions
- Unusual patterns of sensor readings in a nuclear power plant

Prediction:

- Future stock prices or currency exchange rates

Recording

pss.mrteo

Talking: Gouri Choudhury

3 vital things to define

Task: Recognizing hand-written words

Performance Metric: Percentage of words correctly classified

Experience: Database of human-labeled images of handwritten words

pss.mrteo

Talking: Gouri Choudhury

Types of Learning

Supervised (inductive) learning –

- Given: training data + desired outputs (labels)

Unsupervised learning –

- Given: training data (without desired outputs)

Semi-supervised learning –

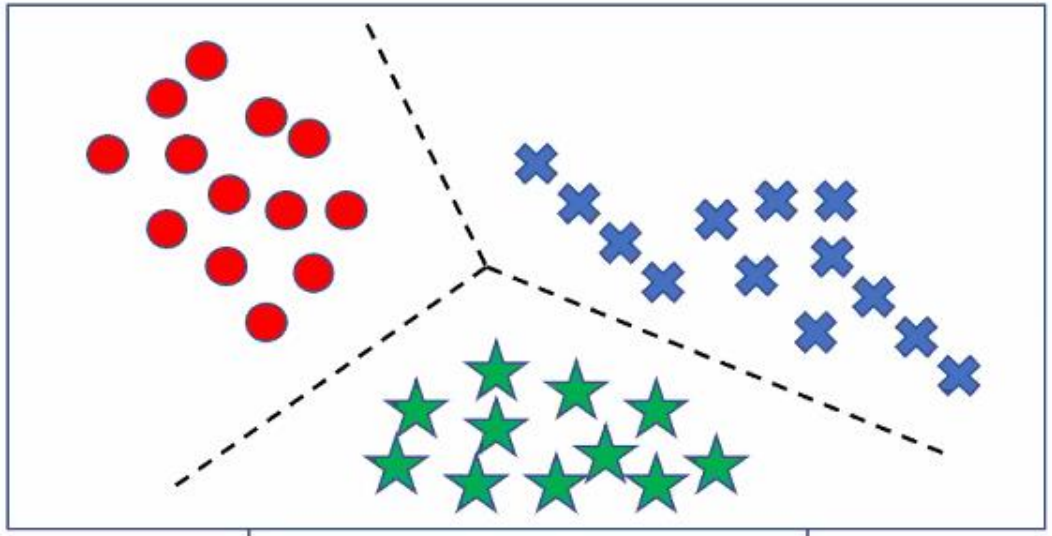
- Given: training data + a few desired outputs

Reinforcement learning –

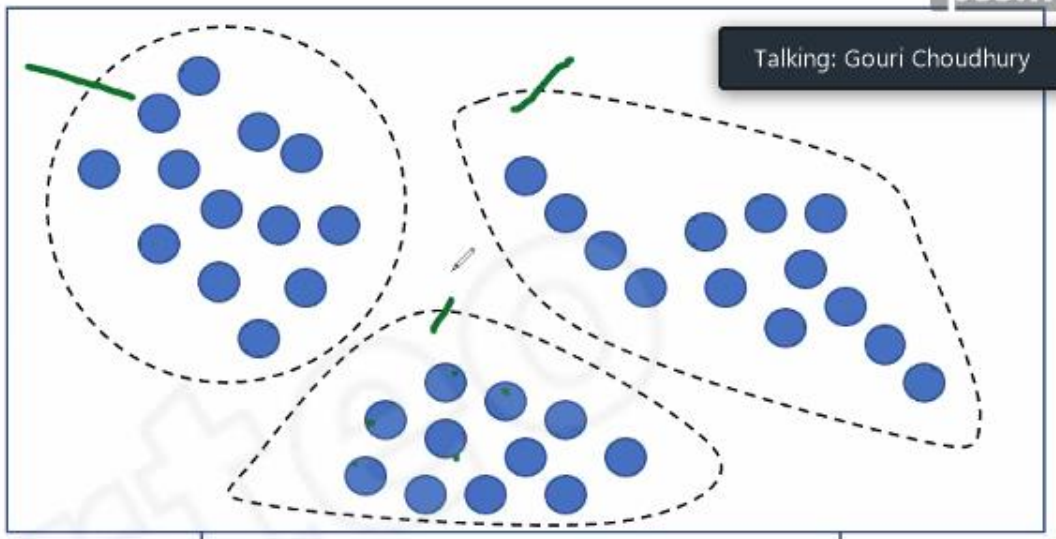
- Rewards from sequence of actions

pss.mrteo

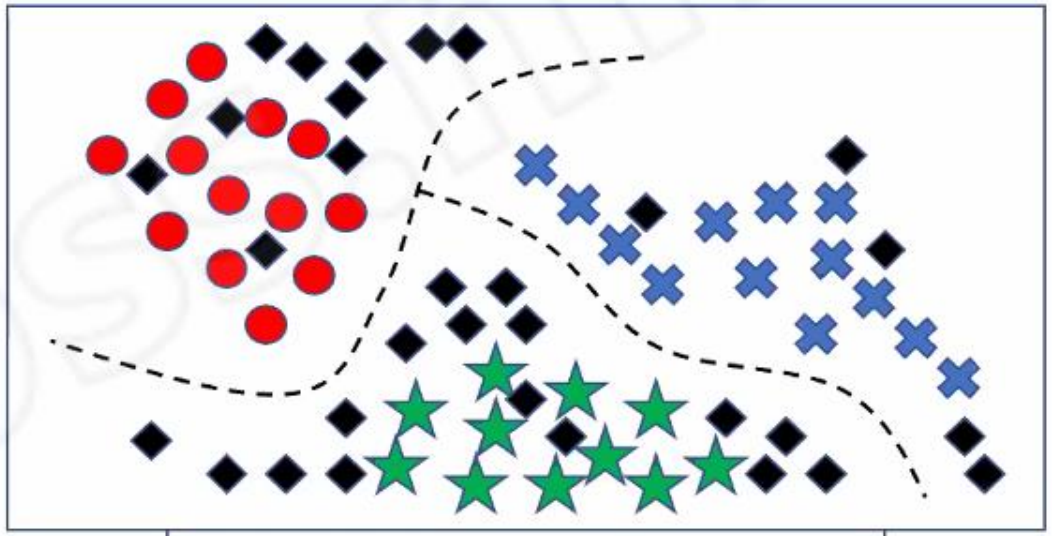
Talking: Gouri Choudhury



Supervised learning



Unsupervised learning



Semi-supervised learning

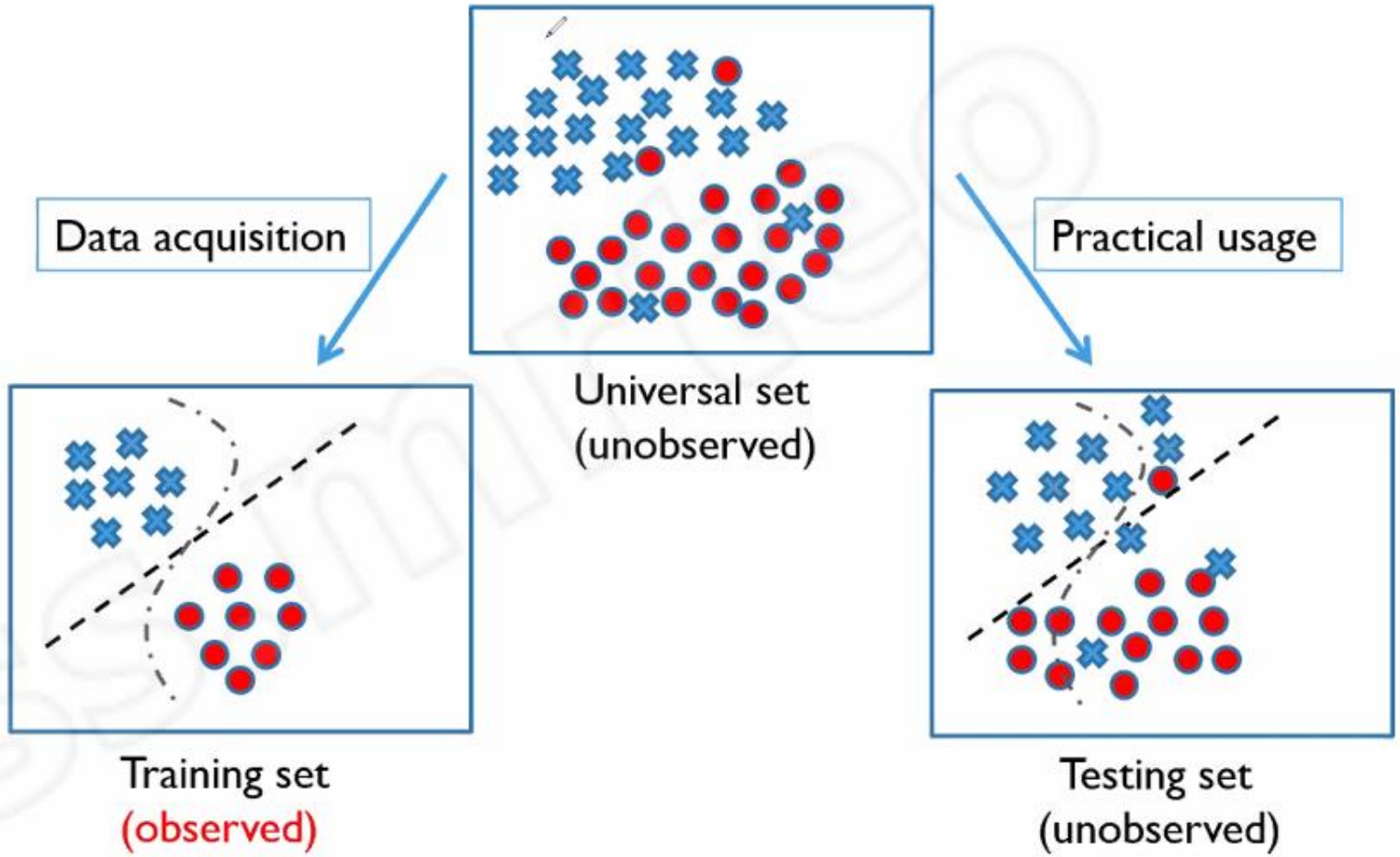
Recording

pss.mrteo

Talking: Gouri Choudhury

M - L

Training and Test Sets



Recording

pss.mrteo

Talking: Gouri Choudhury

Unsupervised Learning

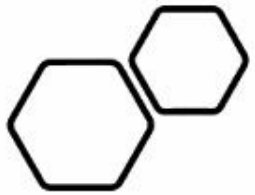
The data has no target attribute.

We want to explore the data to find some intrinsic structures in them.

pss.mrteo

pss.mrteo

Talking: Gouri Choudhury



What is Clustering?

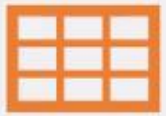
Clustering

Recording

pss.mrteo

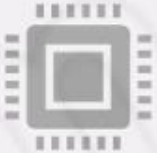
Talking: Gouri Choudhury

What's a cluster?



Intuitive definition:

Grouping of data points that are close to each other



To make this computer friendly, need a mathematical definition of "close."

Closeness (most common definitions):

based on distance or density

pss.mrteo

Talking: Gouri Choudhury

Clustering as unsupervised learning

Unlabeled data —algorithm→ Structured data

New data (unlabeled) —assignment→ New data included in structure

pss.mrteo

pss.mrteo

Talking: Gouri Choudhury

K-means is a partitional clustering algorithm

The *k*-means algorithm partitions the given data into *k* clusters.

Each cluster has a cluster **center**, called **centroid**.

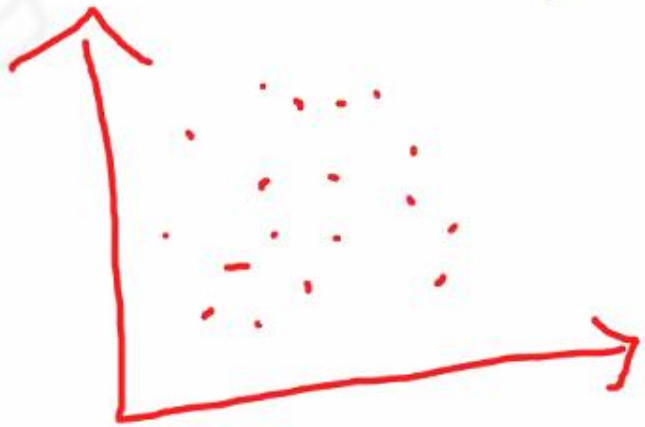
k is specified by the user

pss.mrteo

Talking: Gouri Choudhury

k-means clustering: the **algorithm**

- Choose *k* centroids
- Assign points to cluster based on nearest centroid
- Recompute centroids
- Repeat steps (2) and (3) until there is no more change to the centroids



pss.mrteo

Recording

pss.mrteo

Talking: Gouri Choudhury

k-means performance

Recording

pss.mrteo

Talking: Gouri Choudhury

k-means: adding new data

Add — Add new data to nearest cluster

Treat — Treat clusters as labeled data

Use — Use this data to train a classifier

Apply — Apply classifier to new data

Recording

pss.mrteo

Talking: Gouri Choudhury

k-means: strengths and weaknesses

Strengths:

- Simple—one parameter (*k* clusters)
- Typically fast
- Easy to implement

Weaknesses:

- Optimal *k* is often not obvious
- Sensitive to outliers
- Scaling affects results

Recording

pss.mrteo

Talking: Gouri Choudhury

Clustering - Real life Examples

Example 1: groups people of similar sizes together to make “small”, “medium” and “large” T-Shirts.

Tailor-made for each person: too expensive	One-size-fits-all: does not fit all.

Example 2: In marketing, segment customers according to their similarities

To do targeted marketing.

pss.mrteo

Talking: Gouri Choudhury

Other clustering algorithms

Self Organizing Maps (SOM)

Agglomerative Hierarchical Clustering

pss.mrteo

pss.mrteo

Talking: Gouri Choudhury

Supervised Learning

Data includes both the input and the desired results.

pss.mrteo

pss.mrteo

Talking: Gouri Choudhury

Think of the following examples.

- An emergency room in a hospital measures 17 variables (e.g., blood pressure, age, etc) of newly admitted patients.
- **A decision is needed: whether to put a new patient in an intensive-care unit.**
- Due to the high cost of ICU, those patients who may survive less than a month are given higher priority.
- **Problem**: to predict high-risk patients and discriminate them from low-risk patients.

pss.mrteo

pss.mrteo

Talking: Gouri Choudhury

Another example..

- A credit card company receives lots of applications for new cards. Each application contains information about the applicant for the card,
  - age
  - Marital status
  - annual salary
  - location
  - outstanding debts
  - credit rating
  - Family information etc
- **Problem**: to decide whether an application should be approved or not approved.

pss.mrteo

pss.mrteo

Talking: Gouri Choudhury

Types of Data vs Algorithm

Supervised Learning

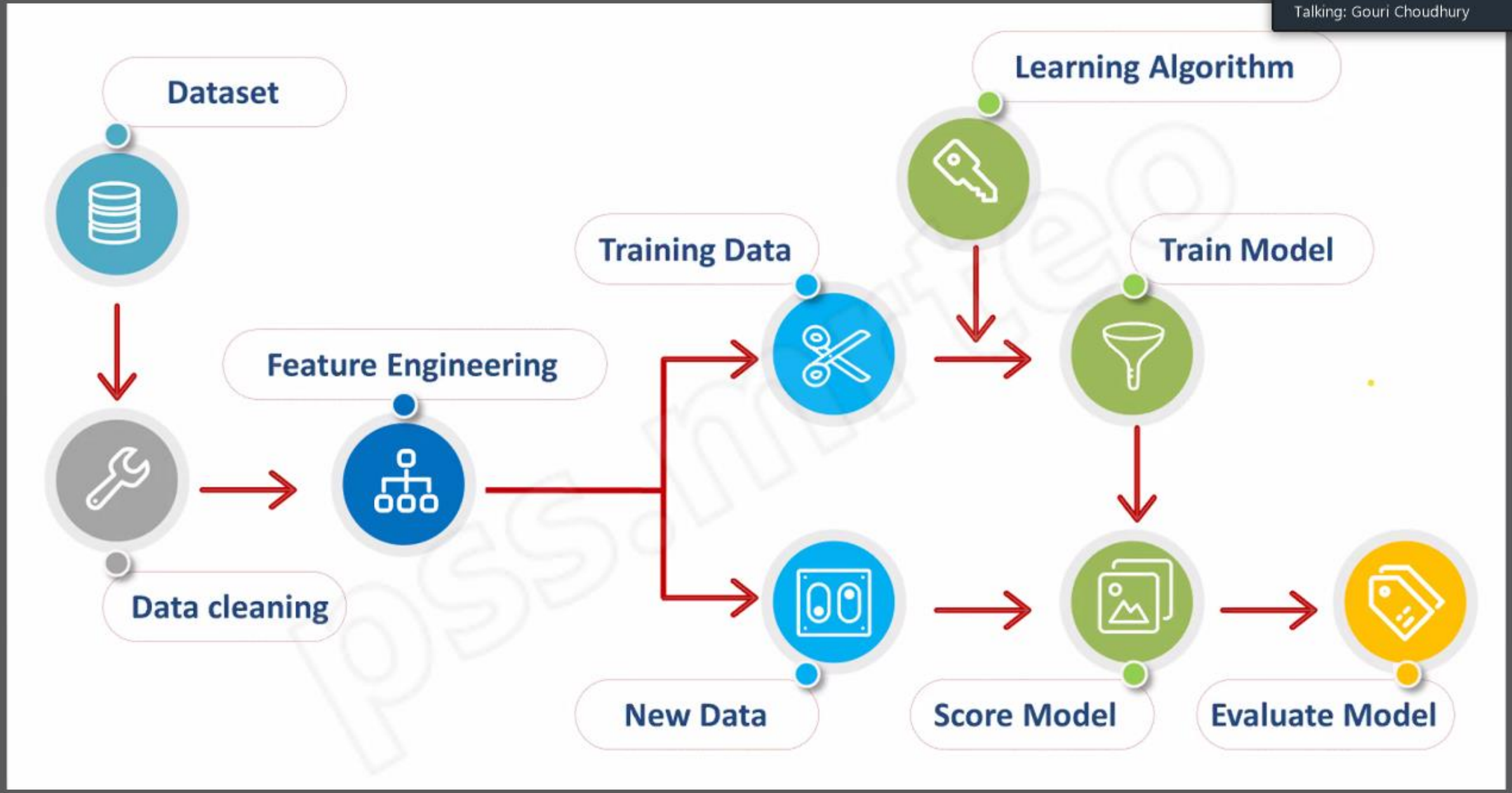
Continuous

- Regression

Categorical

- Classification

Recording
pss.mrteo
Talking: Gouri Choudhury



Recording

pss.mrteo

Talking:

Jargons to be aware of!

Model Inputs: Features, Attributes, Predictors, Inputs, Independent Variables, Dimensions, probably more.

Model Outputs (what we're trying to predict): Target, Response, Output, Dependent Variable, Labels

Row of Data (Inputs + Outputs): Observation, Datapoint, Record, Row

Labels: The values on the target variables in Supervised Learning

Recording

pss.mrteo

Talking: Gouri Choudhury

Linear Regression

Getting our line straight!

pss.mrteo

pss.mrteo

Talking:

Introduction to Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

- Dependent variable:

The variable we wish to predict or explain

- Independent variable:

The variable used to explain the dependent variable

pss.mrteo

Recording

pss.mrteo

Talking: Gouri Choudhury

Simple Linear Regression Model

Only **one** independent variable, X

Relationship between X and Y is described by a linear function.

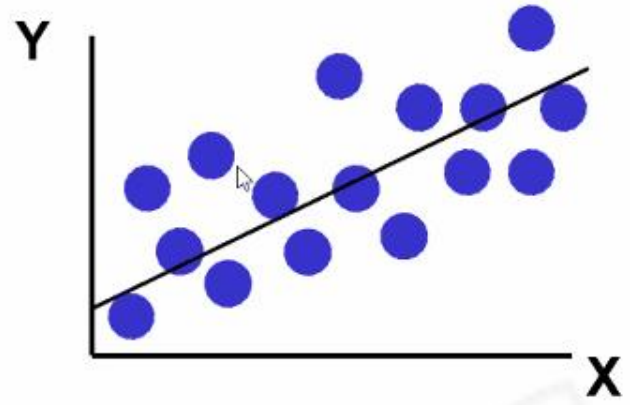
Changes in Y are assumed to be caused by changes in X

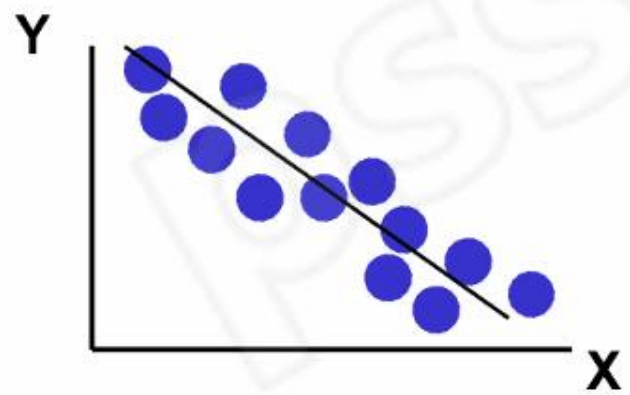
pss.mrteo

pss.mrteo
Talking: Gouri Choudhury

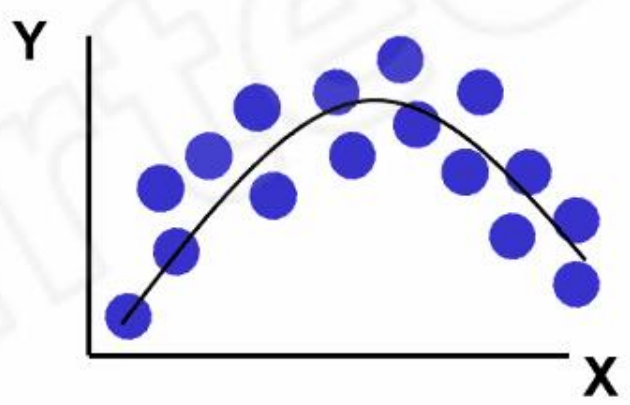
Types of Relationships

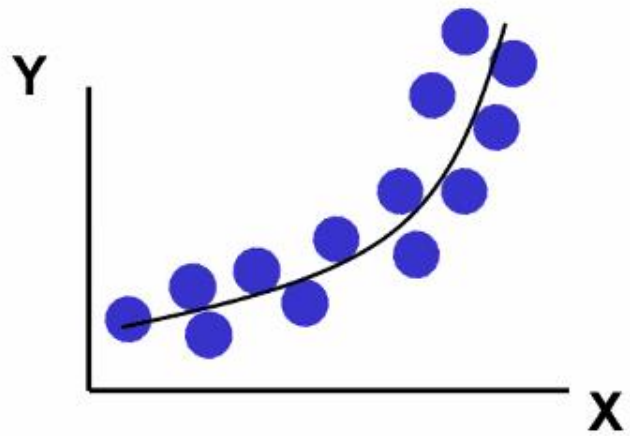
Linear relationships





Curvilinear relationships





Recording

pss.mrteo

Talking: Gouri Choudhury

Feature Engineering

What is it all about?

pss.mrteo

pss.mrteo

Feature engineering

Talking: Gouri Choudhury

- The first thing we need to do when creating a machine learning model is to decide what to use as features.
- **Features** are key to a model, like a person's name or favorite color. pieces of information that we take from the text and give to the algorithm so it can work its magic.
- E.g, if we do classification on health, some features could be a person's height, weight, gender, and so on.
  - We would exclude things that maybe are known but aren't useful

pss.mrteo

Benefits of Feature Engineering

Talking: Gouri Choudhury

- **Reduces Overfitting** : Less redundant data means less opportunity to make decisions based on noise.
- **Improves Accuracy :** Less misleading data means modeling accuracy improves.
- **Reduces Training Time :** Fewer data points reduce algorithm complexity and algorithms train faster.

pss.mrteo

pss.mrteo

What is logistic regression?

- **Logistic regression** is the appropriate regression analysis to conduct when the dependent variable is **binary**.
- Like all regression analyses, the logistic regression is a predictive analysis.
- Logistic regression is used to describe data and to explain the relationship between one dependent **binary variable** and **one or more nominal, ordinal, interval or ratio-level independent variables.**

Recording

pss.mrteo

Good to know!

Nominal

- Nominal scales are used for labeling variables, without any quantitative value. “Nominal” scales could simply be called “labels.”
  - E.g Male/Female, Red/Green/Yellow

Ordinal

- With ordinal scales, the order of the values is what’s important and significant, but the differences between each one is not really known.
  - E.g Good, Very good, Excellent, Fantastic – 1#, 2#, 3#, 4#

Interval

- Interval scales are numeric scales in which we know both the order and the exact differences between the values.
  - E.g Temp Celsius - because the difference between each value is the same.

Recording

pss.mrteo

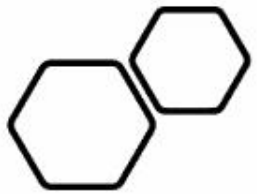
Model Evaluation

Model Evaluation is an integral part of the model development process.

It helps to find the best model that represents our data and how well the chosen model will work in the future.

pss.mrteo

pss.mrteo



Performance Metrics (Classification)

Confusion Matrix

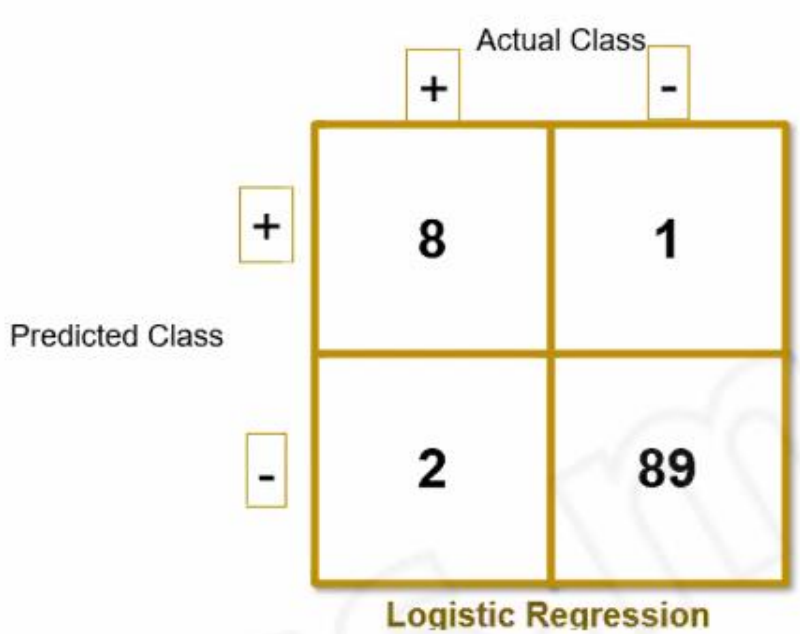
Accuracy

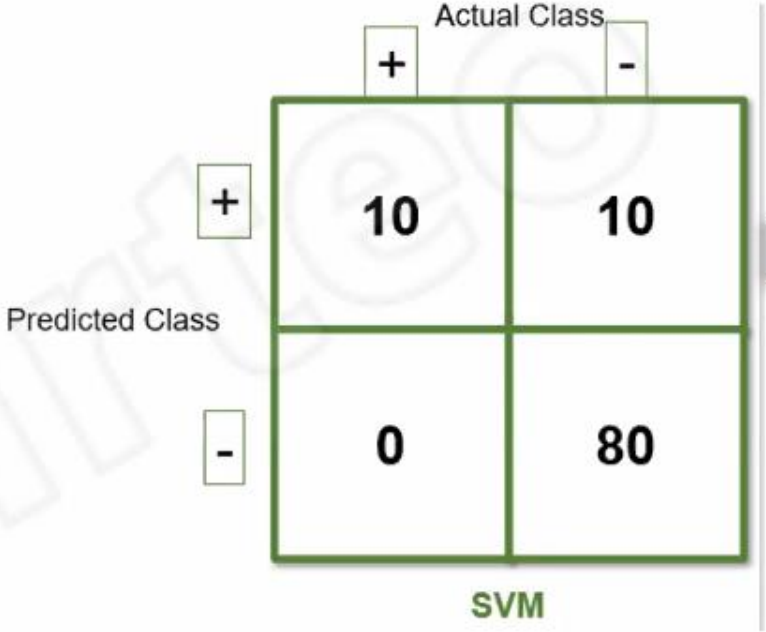
Precision and Recall

pss.mrteo

You are viewing Gouri Choudhury's screen
View Options
View
pss.mrteo

We can compare them!





Talking:

Accuracy: (TP+TN)/(TP+TN+FP+FN)	97%	90%
Precision: TP/(TP+FP)	89%	50%
Recall:	80%	100%

Unmute
Stop Video
17
Participants
Chat
Share Screen
Reactions
Leave

pss.mrteo

	Predicted class POSITIVE (spam ✉)	Predicted class NEGATIVE (normal 📧)
Actual class POSITIVE (spam ✉)	TRUE POSITIVE (TP) ✉ ✉ **320**	FALSE NEGATIVE (FN) ✉ 📧 **43**
Actual class NEGATIVE (normal ✉)	FALSE POSITIVE (FP) ✉ ✉ **20**	TRUE NEGATIVE (TN) ✉ 📧 **538**

Recall

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

$$= \frac{320}{320 + 43} = 0.882$$

Talking: Gouri Choudhury

Precision

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

$$= \frac{320}{320 + 20} = 0.941$$

pss.mrteo

Precision and Recall

Precision attempts to answer the following question:

What proportion of positive identifications was corre

Talking: Gouri Choudhury

Recall attempts to answer the following question:

What proportion of actual positives was identified correctly?

pss.mrteo

Recording

pss.mrteo

Talking: Gouri Choudhury

Support Vector Machines

pss.mrteo

pss.mrteo

Talking: Gouri Choudhury

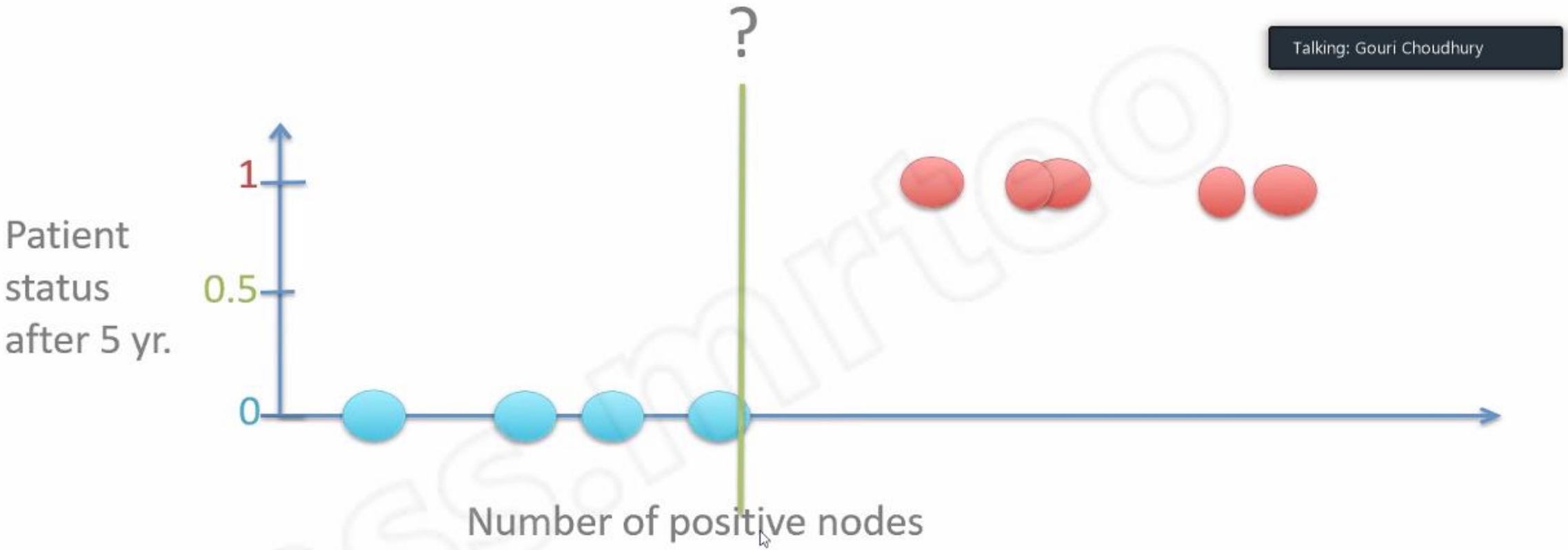
What are SVMs?

- SVMs are linear or non-linear classifiers that find a hyperplane to separate two class of data, positive and negative.
- SVM not only has a rigorous theoretical foundation, but also performs classification more accurately than most other methods in applications, especially for high dimensional data

pss.mrteo

pss.mrteo

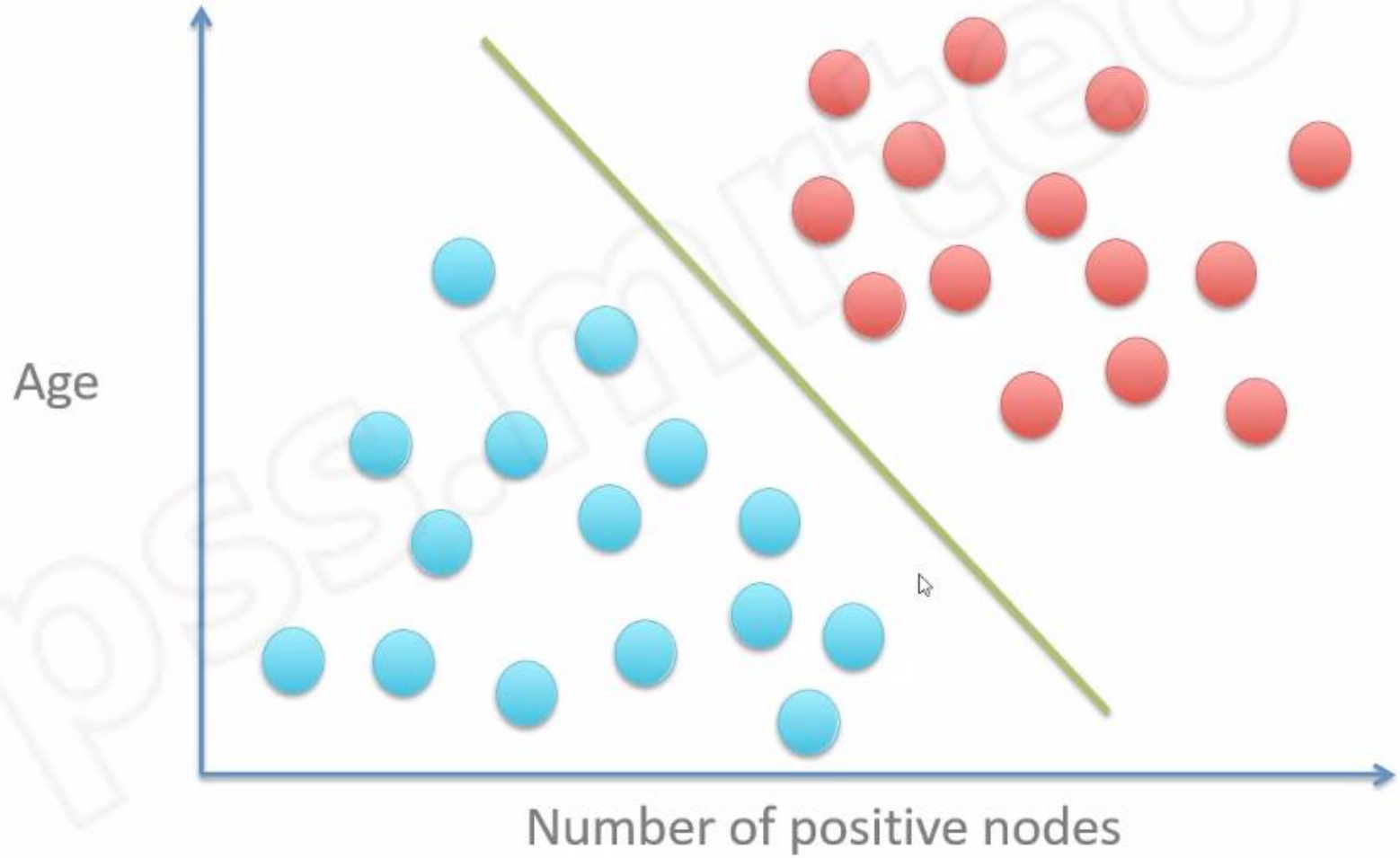
Support Vector Machine (SVM)



Accuracy: 100%

pss.mrteo

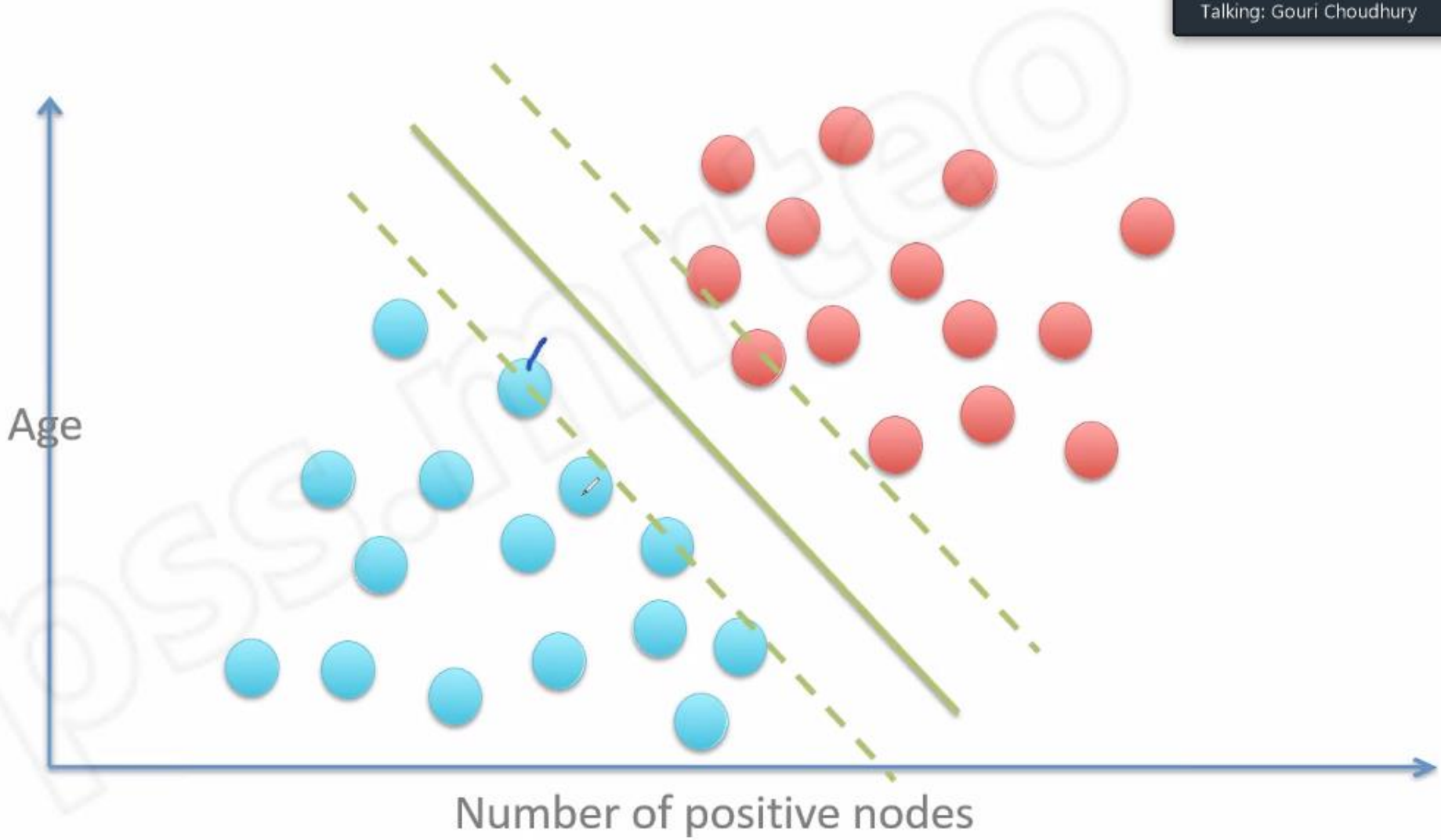
Talking: Gouri Choudhury



pss.mrteo

Best boundary

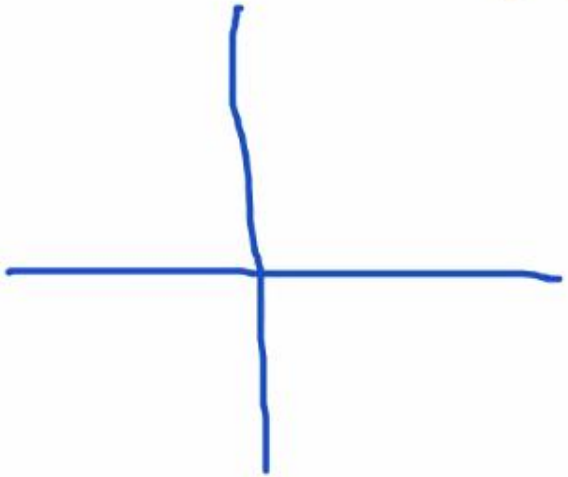
Talking: Gouri Choudhury



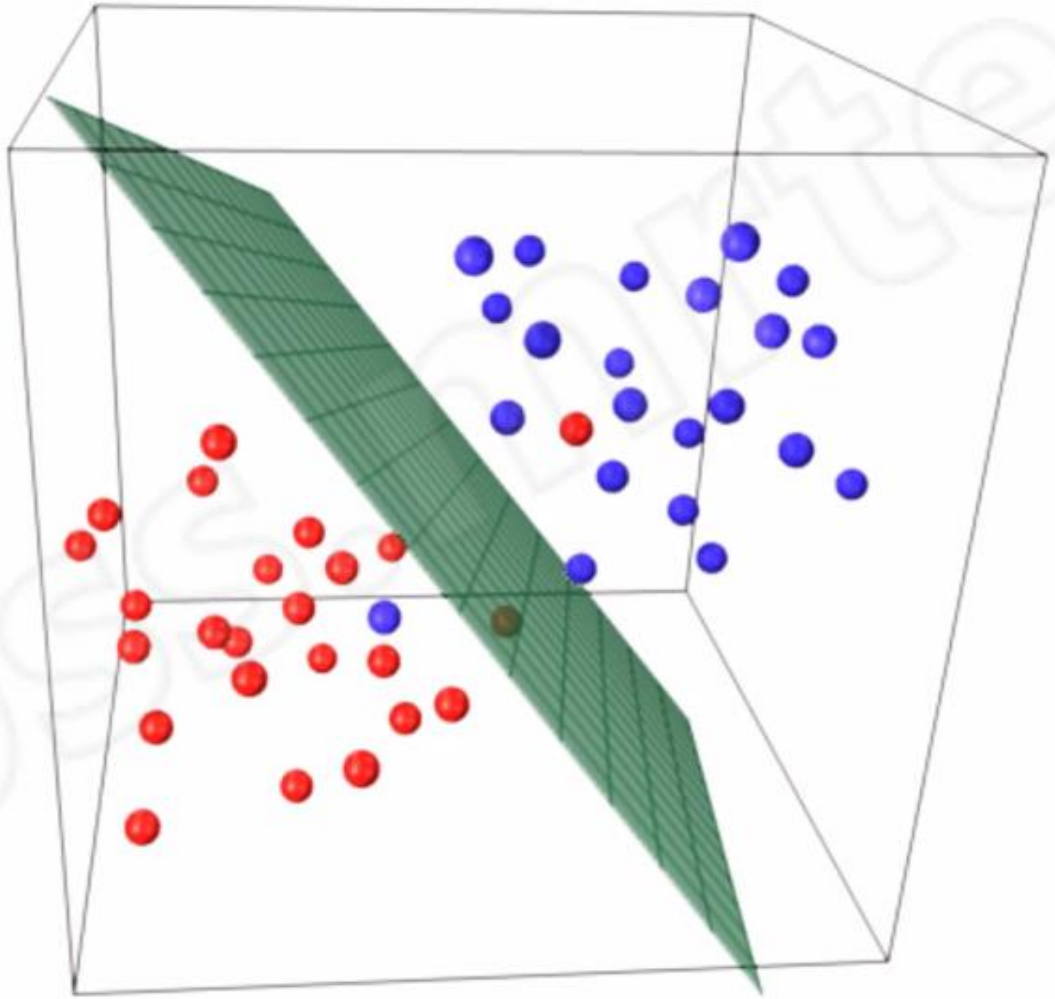
pss.mrteo

3 features: Find the best boundary plane

(More features: hyperplane)



Talking: Gouri Choudhury



Recording

pss.mrteo

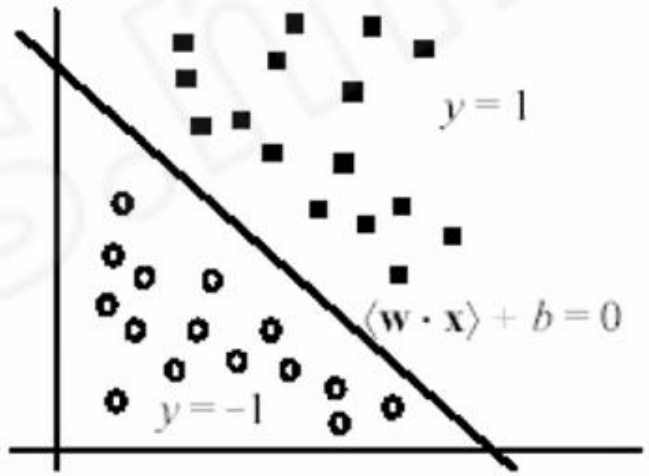
What is a hyperplane?

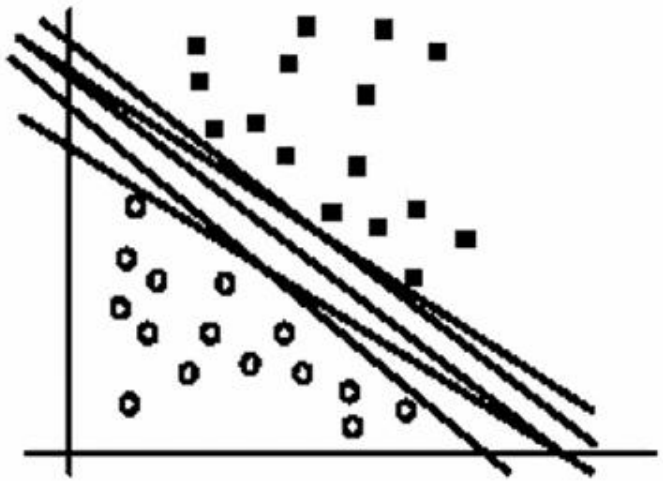
Talking: Gouri Choudhury

- The hyperplane that separates positive and negative training data is

$$\langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 0$$

- It is also called the decision boundary (surface).





pss.mrteo

Talking: Gouri Choudhury

Pros

- Accuracy
- Works well on smaller cleaner datasets
- It can be more efficient because it uses a subset of training points

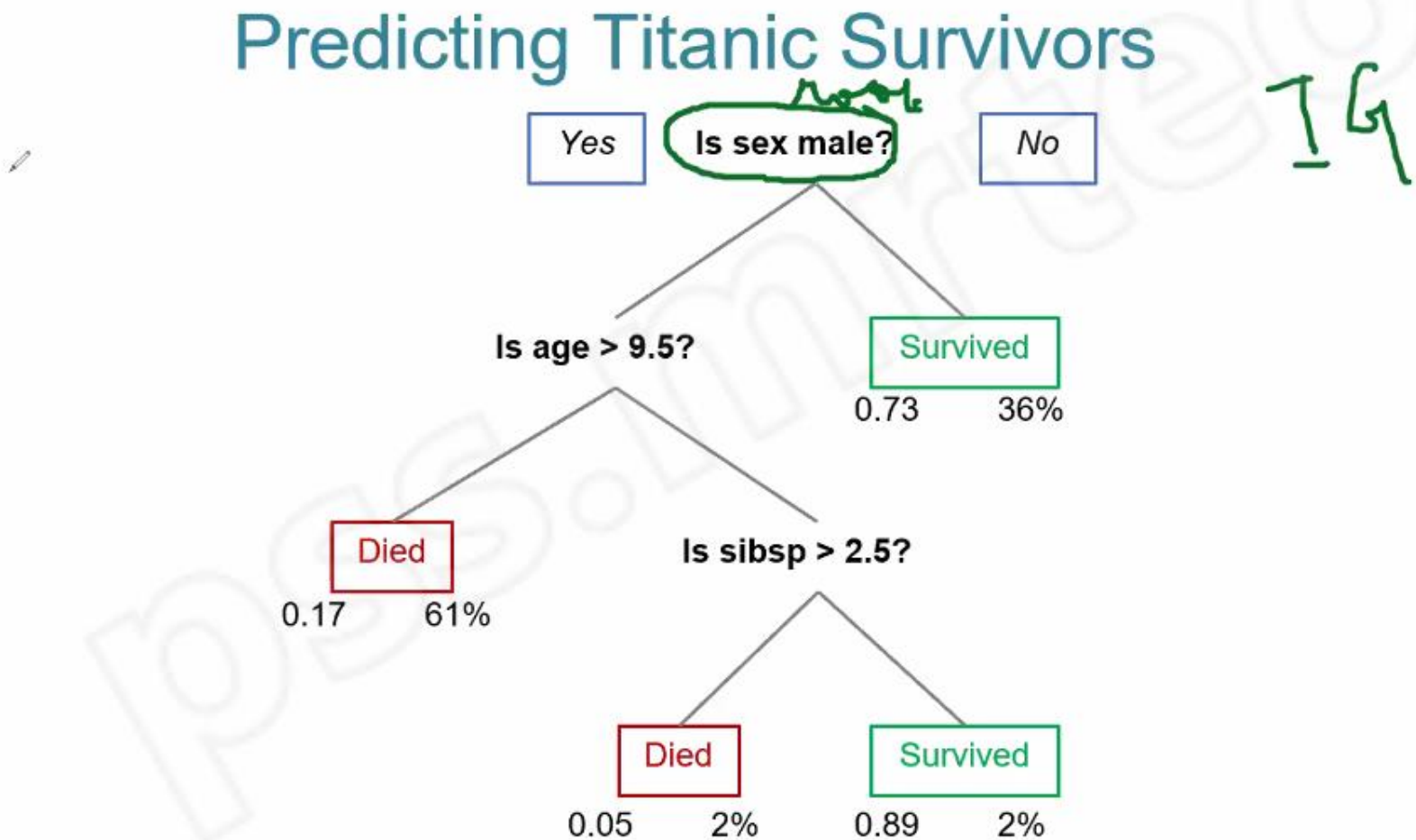
Cons

- Isn't suited to larger datasets as the training time with SVMs can be high
- Less effective on noisier datasets with overlapping classes

pss.mrteo

Each inner node is a decision based on a feature
Each leaf node is a **class label**

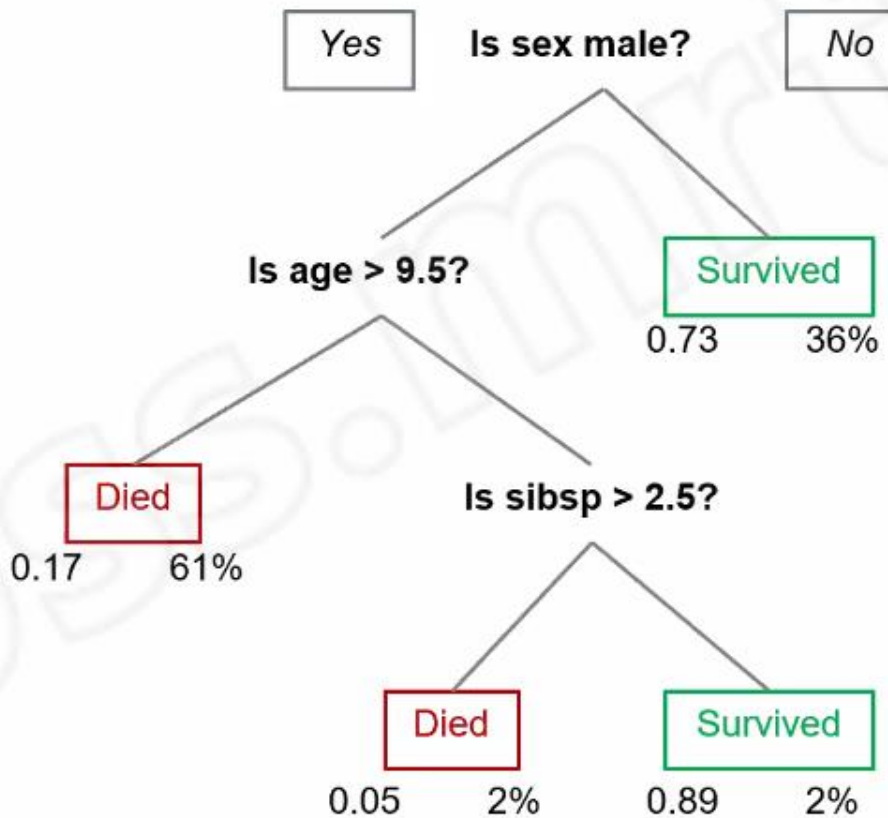
Talking: Gouri Choudhury



pss.mrteo

Build tree split by split,
Find the best split you can at each step

Talking: Gouri Choudhury



Recording

pss.mrteo

Talking:

Strengths of decision tree methods

- Generates understandable rules.
- Perform classification without requiring much computation.
- able to handle both continuous and categorical variables.
- Provides a clear indication of which fields are most important for prediction or classification.
- Natural multiclass classifier.

pss.mrteo

pss.mrteo
Talking: Gouri Choudhury

Weaknesses of decision tree

- It is less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
- Prone to errors in classification problems with many class and relatively small number of training examples.
- Computationally expensive to train.
  - Growing a decision tree is computationally expensive.
  - At each node, each candidate splitting field must be sorted before its best split can be found.
- Small changes in input data can result in totally different trees.
- Can make mistakes with unbalanced classes.

pss.mrteo

Financial Services

- Customer targeting/engagement
- Improved risk management
- Fraud detection in real-time

Retail & CPG

- Multi-channel sales analysis & optimization
- Customer behaviour modeling
- Real-time recommendation engines

Talking: Gouri Choudhury

Transportation

- Consumers choose time of home deliveries
- Fleet vehicle maintenance optimization
- Making logistics and fuel consumption less dependent on weather and traffic

E-commerce

- Analyze internet behavior and buying patterns
- Digital asset piracy

Telecommunications

- Customer churn & experience analysis
- Network service quality/predictive maintenance via sensor data

Utilities

- Service Quality Optimization
- Weather impact analysis on power generation
- Smart meter data analysis

Call Centers

- On-the-fly offer prompting
- Improved consumer experience
- Compliance verification

IT

- Network analysis & optimization
- Application log analysis (performance, threats, optimization

Healthcare

- E-Prescriptions
- Remote Patient Monitoring

Source: bigdata4analytics

Recording

pss.mrteo

Talking:

A Self Organizing Map (SOM) is an example of which type of learning algorithm?

- Unsupervised Learning
- Supervised Learning

pss.mrteo

Recording

pss.mrteo

Talking: Gouri Choudhury

Imagine, you are solving a classification problems with highly imbalanced class.

The majority class is observed 99% of times in the training data. Which of the following is a suitable metric to look at?

- Accuracy
- Precision
- Mean Absolute Error
- None of the above

pss.mrteo

Recording

pss.mrteo

A feature F can take certain value: A, B, C, D, E, & F and represents grade of students from a college.

Which of the following statement is true?

Talking: Gouri Choudhury

- Feature F is an example of nominal variable
- Feature F is an example of ordinal variable
- Both the above
- None of the ab

pss.mrteo

Recording

pss.mrteo

At least eighty percent of the time spent on a Web-based data mining project is devoted to this

Talking: Gouri Choudhury

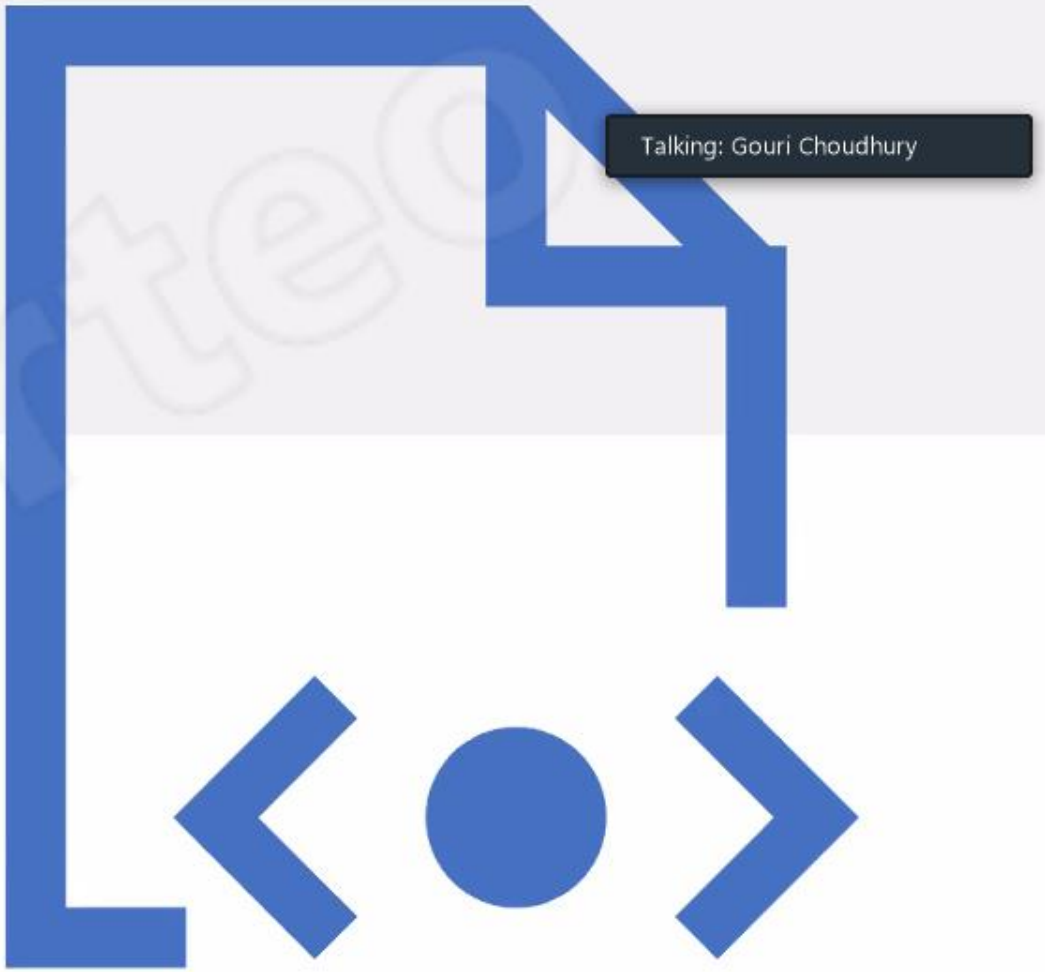
- interpretation of results
- data mining
- goal identification
- data preparation

pss.mrteo

Recording

pss.mrteo

Which statement is true about the K-Means algorithm?



- All attribute values must be categorical
- The output attribute must be cateogrical
- Attribute values may be either categorical or numeric
- All attributes must be numeric

pss.mrteo

Recording

pss.mrteo

The correlation between the number of years an employee has worked for a company and the salary of the employee is 0.75.

What can be said about employee salary and years worked?

- Individuals that have worked for the company the longest have higher salaries
- There is no relationship between salary and years worked
- Individuals that have worked for the company the longest have lower salaries.
- The majority of employees have been with the company a long time.

pss.mrteo

Simple regression assumes a

_____

relationship between the input attribute and output attribute

Talking: Gouri Choudhury

- quadratic
- reciprocal
- inverse
- linear

pss.mrteo

Recording

pss.mrteo

Talking: Gouri Choudhury

A correlation coefficient enables you to:

- establish whether the data is telling you what you think it should tell you.
- quantify the strength of the linear relationship between two ranked or quantifiable variables.
- assess whether two variables measure the same phenomenon.
- measure the difference between two variables.

pss.mrteo

Recording

pss.mrteo

Talking:

Exploratory Data Analysis (EDA) is:

- A set of statistical methods specially designed for exploring a small, unruly data set and identifying any abnormalities in distribution or highly unusual scores
- The stage at which the data are described by the traditional measures of central tendency, spread and distribution shape
- Especially appropriate for nominal data
- Of limited value because no formal statistical tests are made

pss.mrteo

Recording

pss.mrteo

Talking: Gouri Choudhury

The average squared difference between classifier predicted output and actual output

- mean squared error
- root mean squared error
- mean absolute error
- mean relative error

pss.mrteo