Getting Started

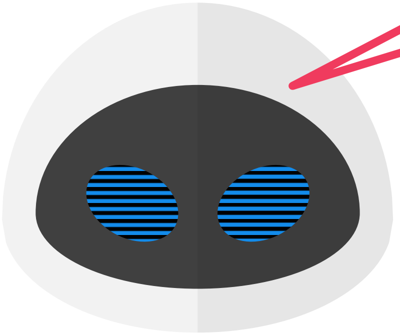
CS771: Introduction to Machine Learning

Purushottam Kar

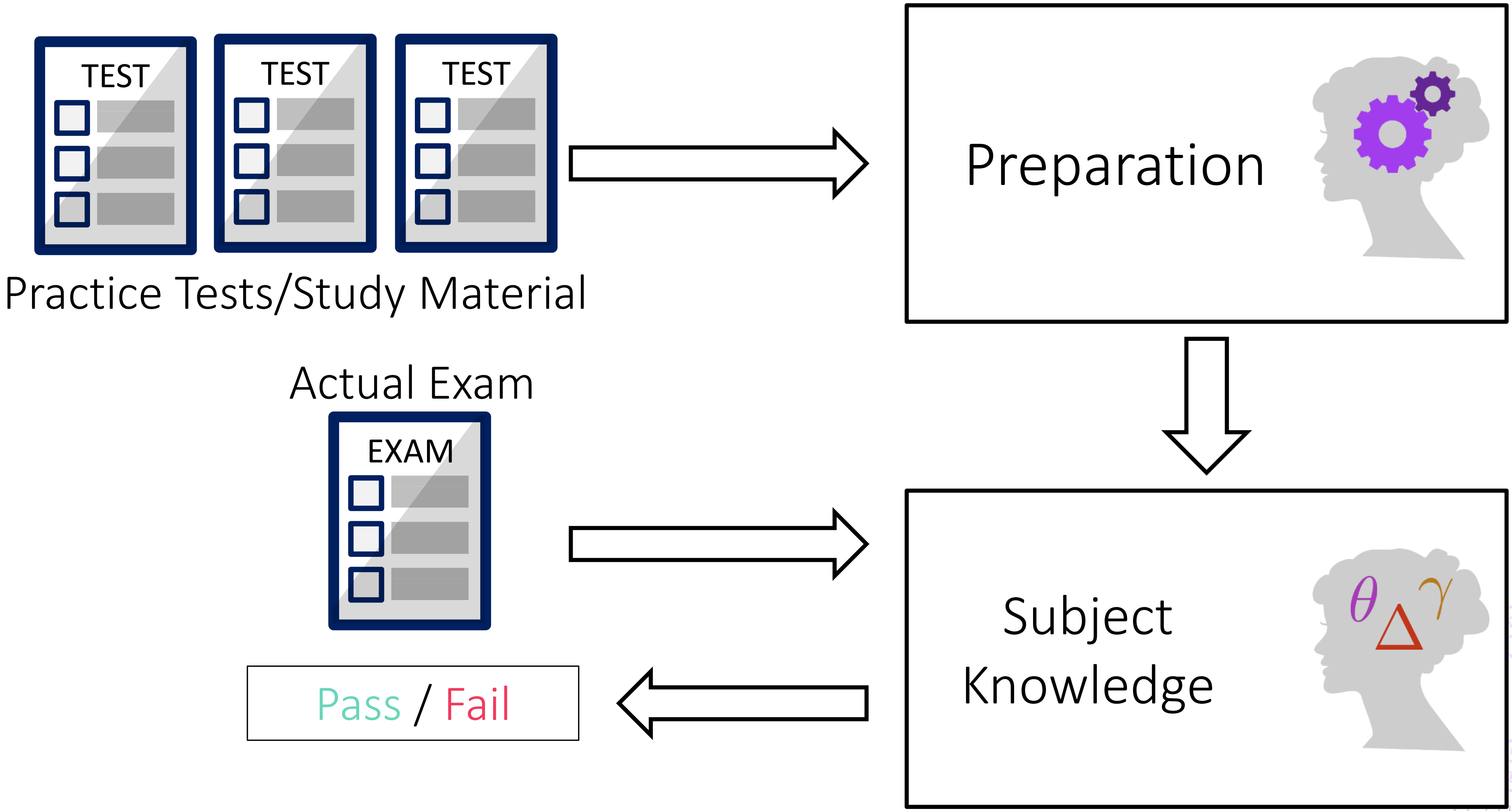
An overview of ML

- Study similarity of ML with a student preparing for an exam
- Look at a toy ML problem
- Learn what is training data, test data?
- Learn what is a model?

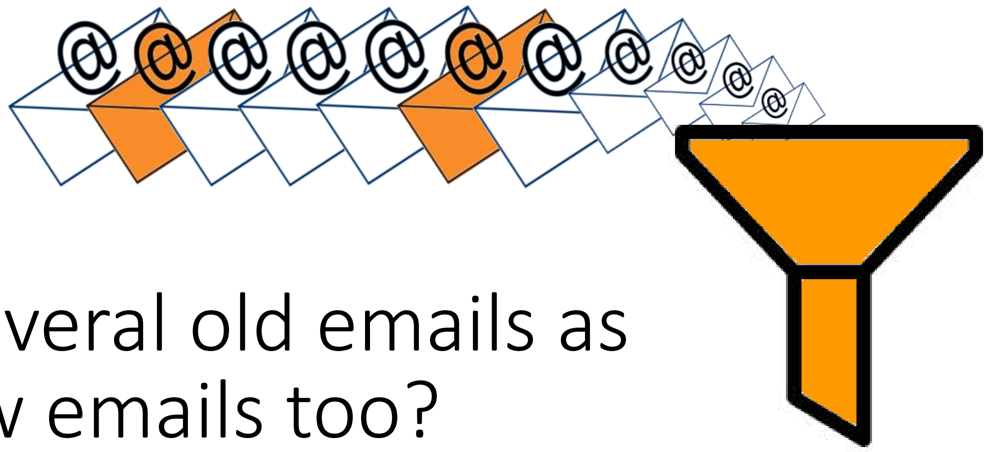
Warning: lots of oversimplifications ahead!



A typical study cycle (e.g. in a course)

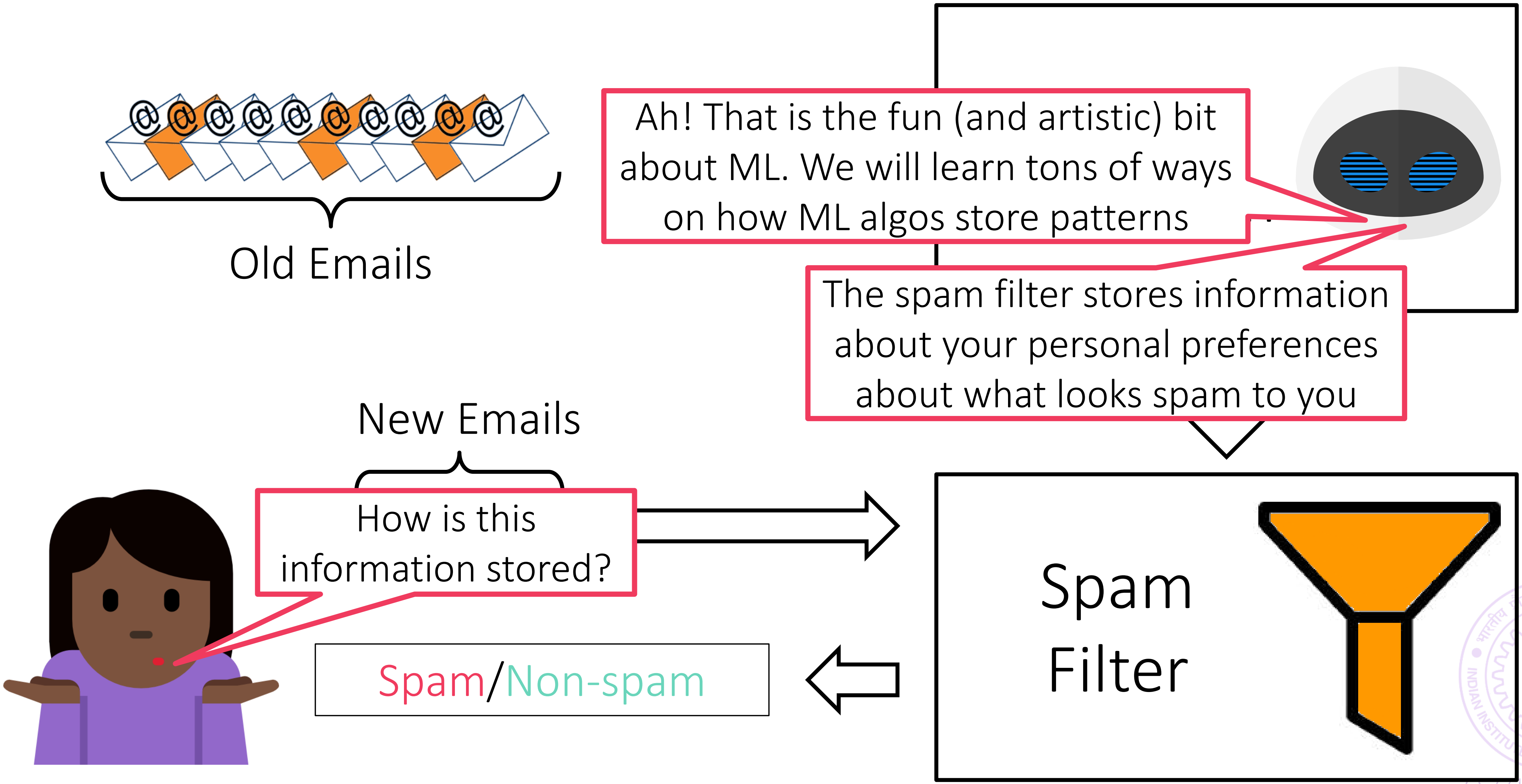


Spam Filtering

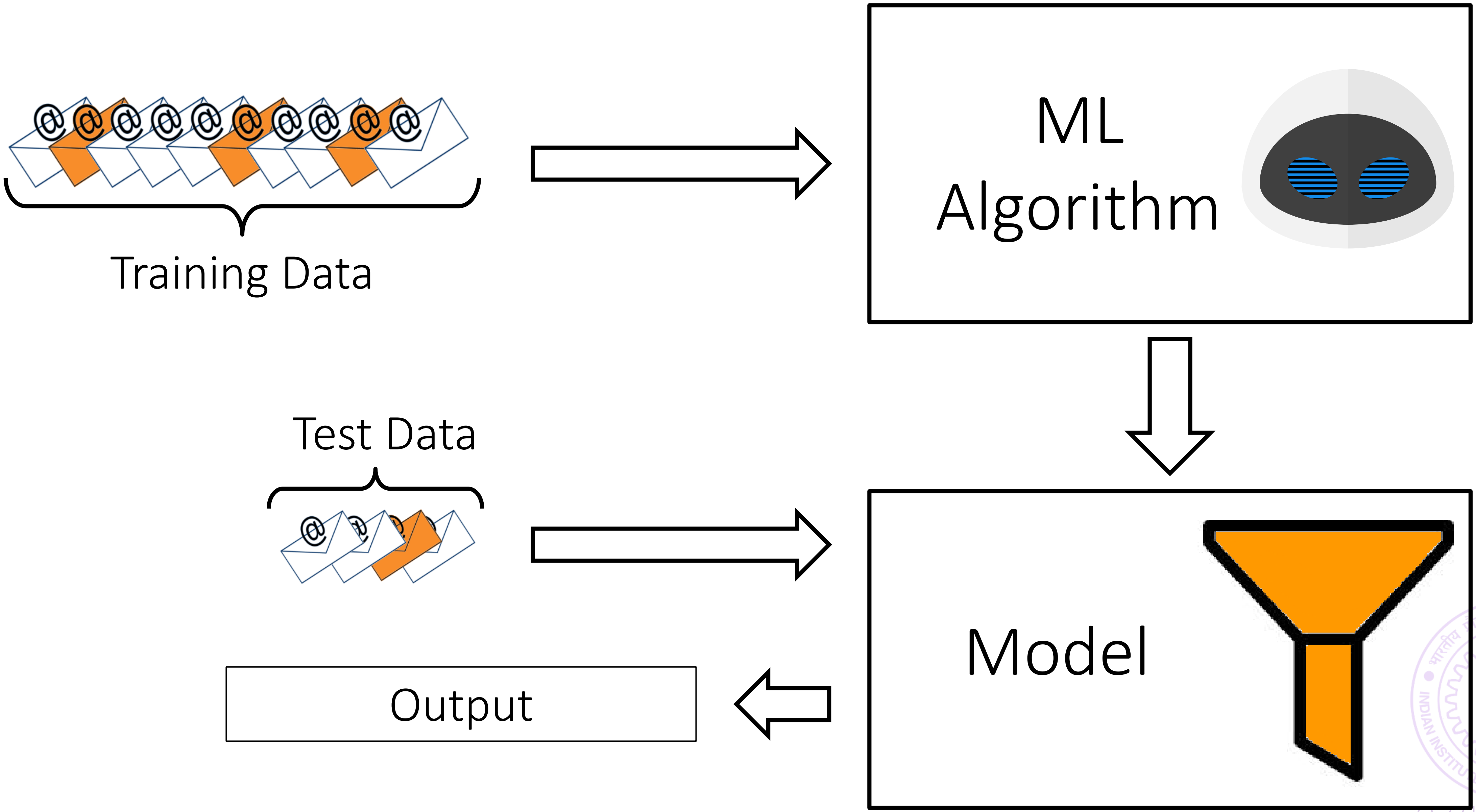


- Suppose Mary has already tagged several old emails as spam/non-spam, can we tag her new emails too?
- **Trick**: use the old tagged emails to try and understand what sort of emails does Mary think of as spam and which as non-spam!
- E.g. may find that emails about shopping always tagged as spam
- E.g. may find that emails from Jill are never tagged as spam
- These insights/patterns are what are stored in the spam filter
- Our spam filter helps us make predictions on new emails

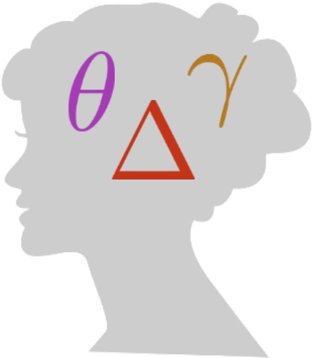
A typical ML workflow

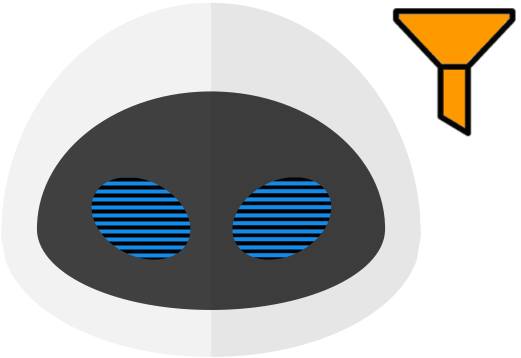


A typical ML workflow



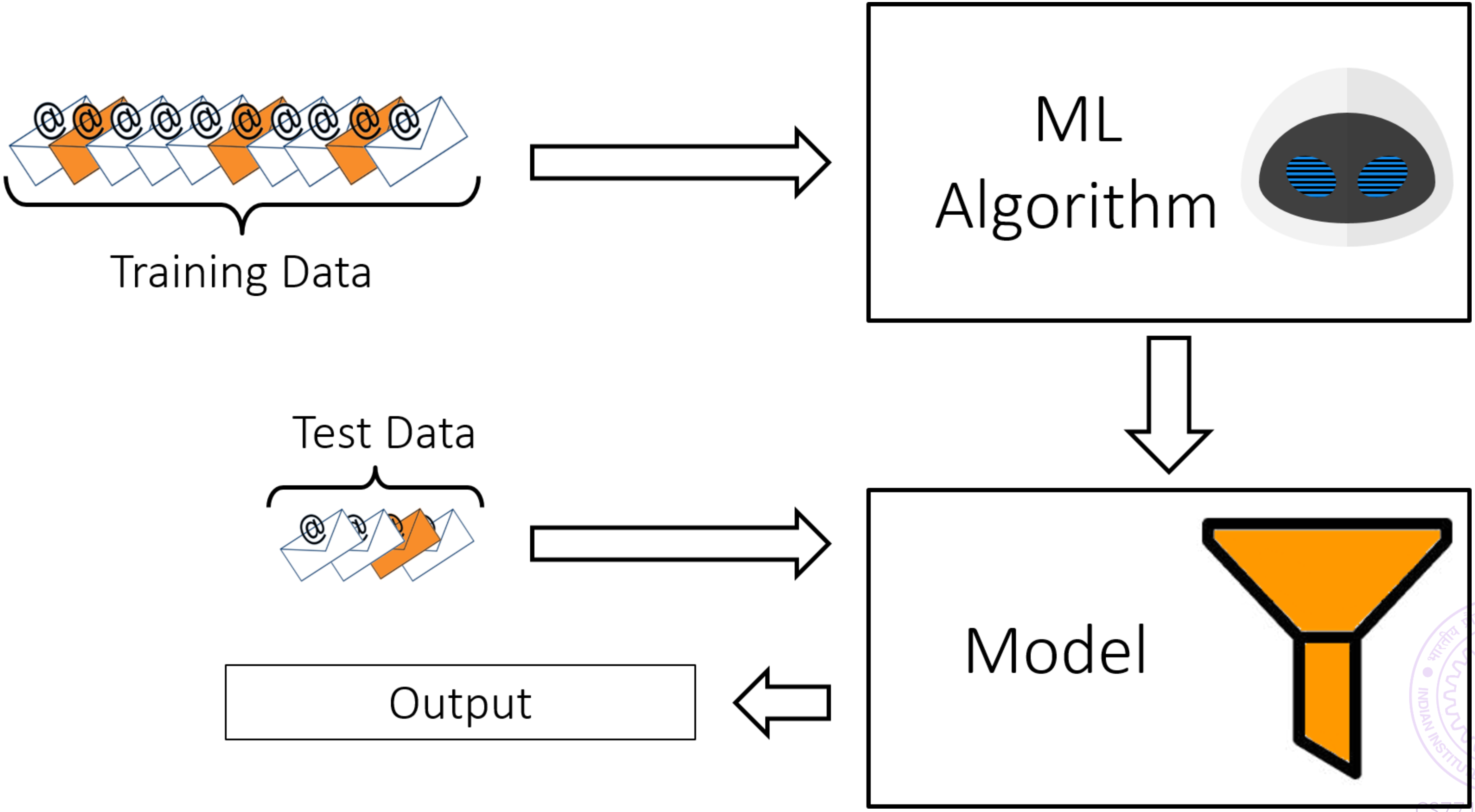
ML as an “examination”





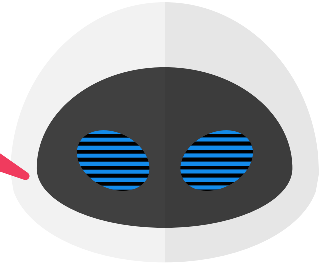
Our brain stores subject matter	The model stores data patterns
Use subject matter to solve exam	Use model to predict on test data
Critical to do well on exam-day	Critical to do well on test data
Mock test results indicative	Training accuracies indicative
No out-of syllabus questions	Training/test data are similar
Should not leak exam paper before exam	Should not look at test data while training

A typical ML workflow



ML can do lots of cool things with test data

In this course, we will learn how to do most of these operations with test data

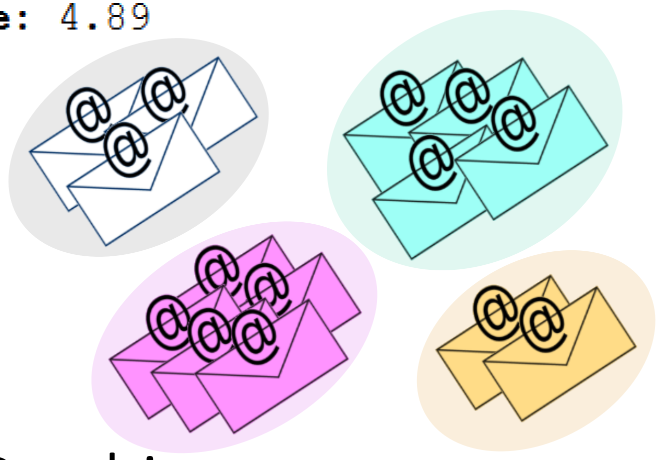


Regression

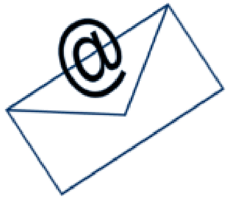
Subject: [**SPAM**] Free movie tickets every month
X-Barracuda-Spam-Score: 4.89

Multi-classification

Clustering

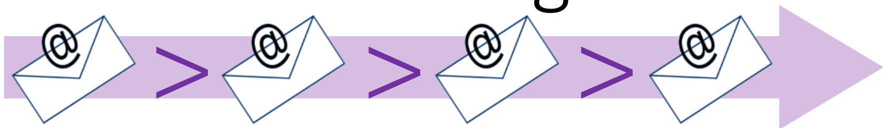


Tagging



URGENT,
OFFICIAL,
TAX

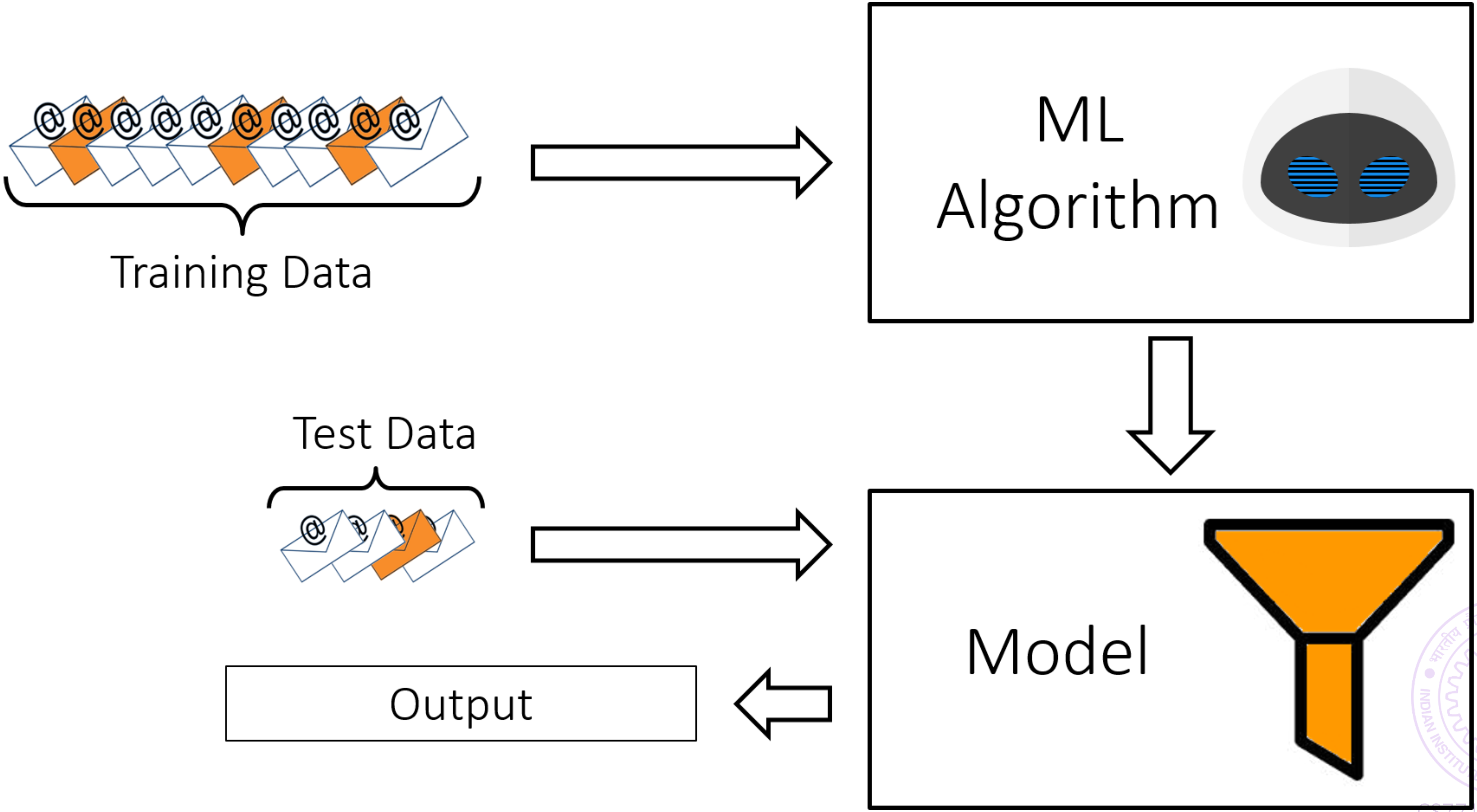
Ranking



Binary Classification

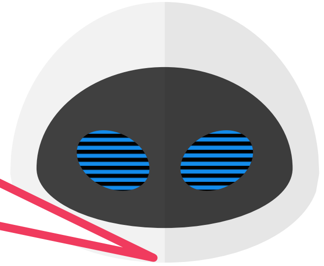
Subject: [**SPAM**] Free movie tickets every month
X-Barracuda-Spam-Status: Yes

A typical ML workflow



ML can take in lots of kinds of training data

We wont be able to cover all these training settings in this course – there are entire courses devoted to specific training settings e.g. CS773 (Online Learning)



Training Data

Batch Learning

Active Learning

Robust Learning

Semi-supervised Learning

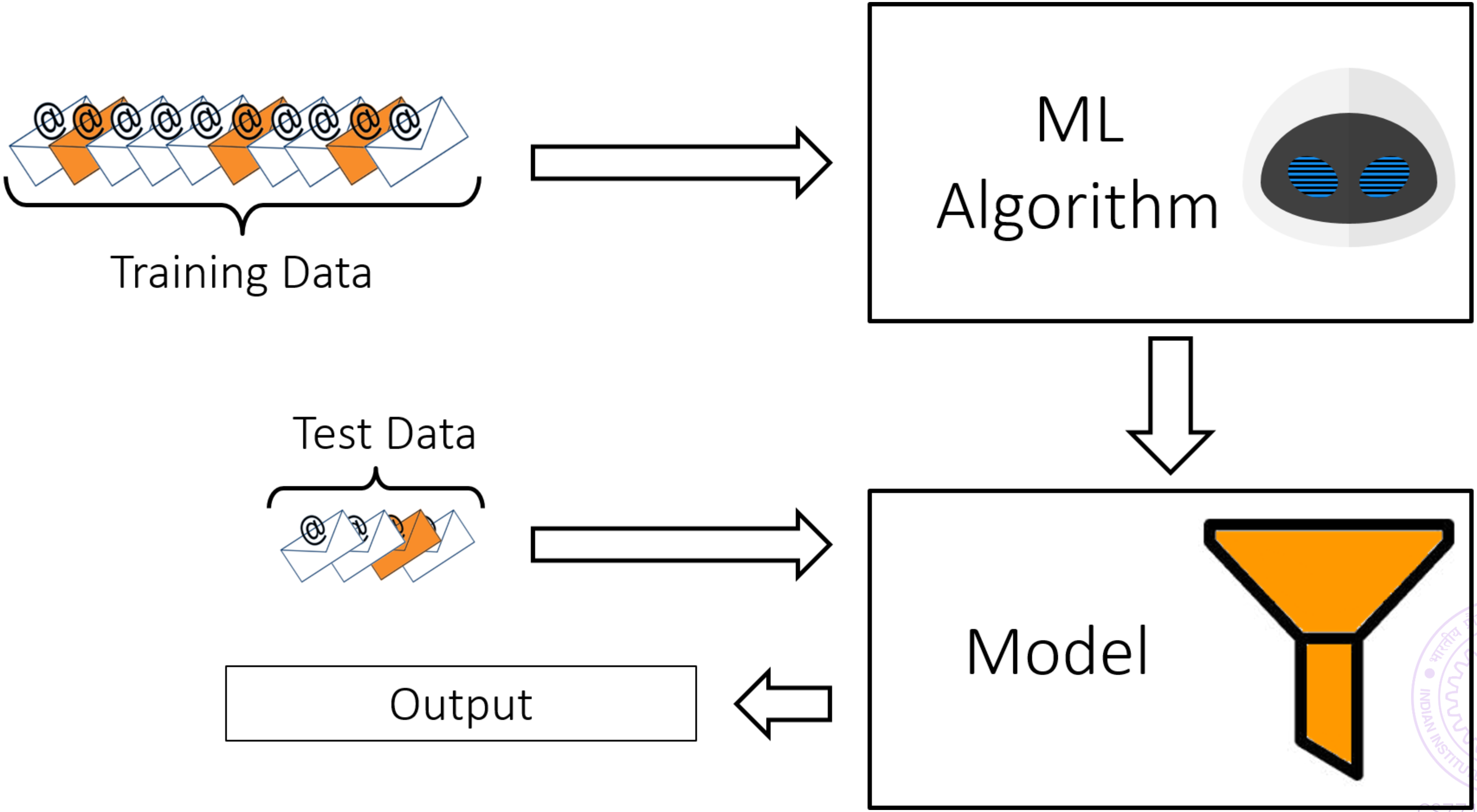
Supervised Learning

Online Learning

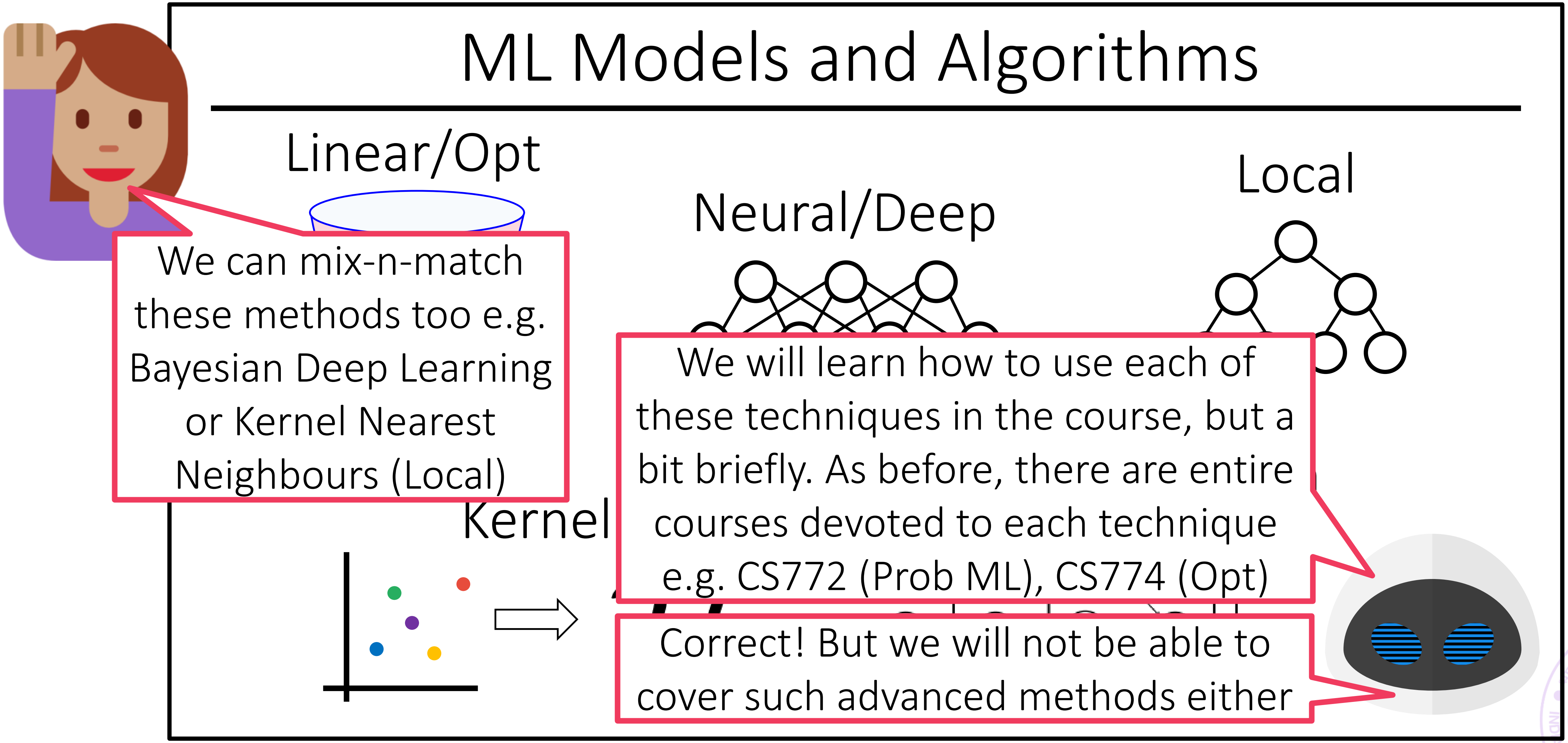
Unsupervised Learning

Reinforcement Learning

A typical ML workflow



ML can store info in lots of innovative ways



Fantastic Features

- … and how to find them
- What are vectors
- How are vectors used in ML
- Useful operations on vectors

What are features

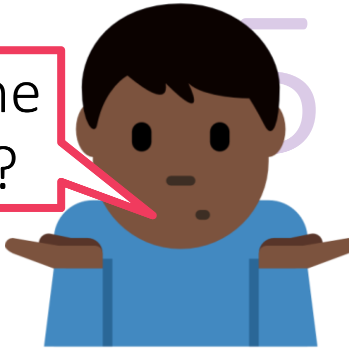
Feature … t to ML algorithms

ML al… …ectors to represent features

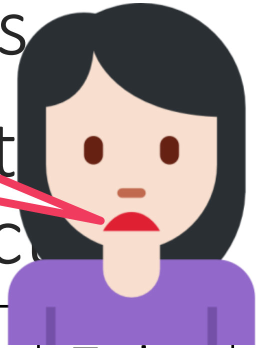
…ample, in spam c… …dat…

every new email (test data) must be converted into a vec…

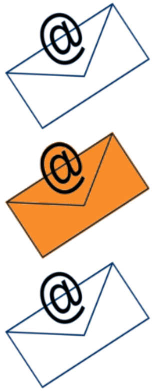
We could have – but it does not carry much information about spam/non-spam since it is such a common word!



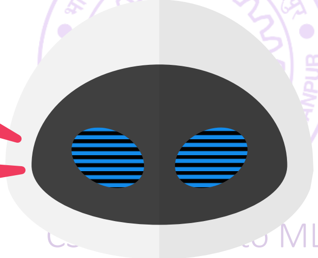
Why did we not keep the word "to" as a feature?



Guys, something is wrong with our feature. The word "do" means different things in two emails



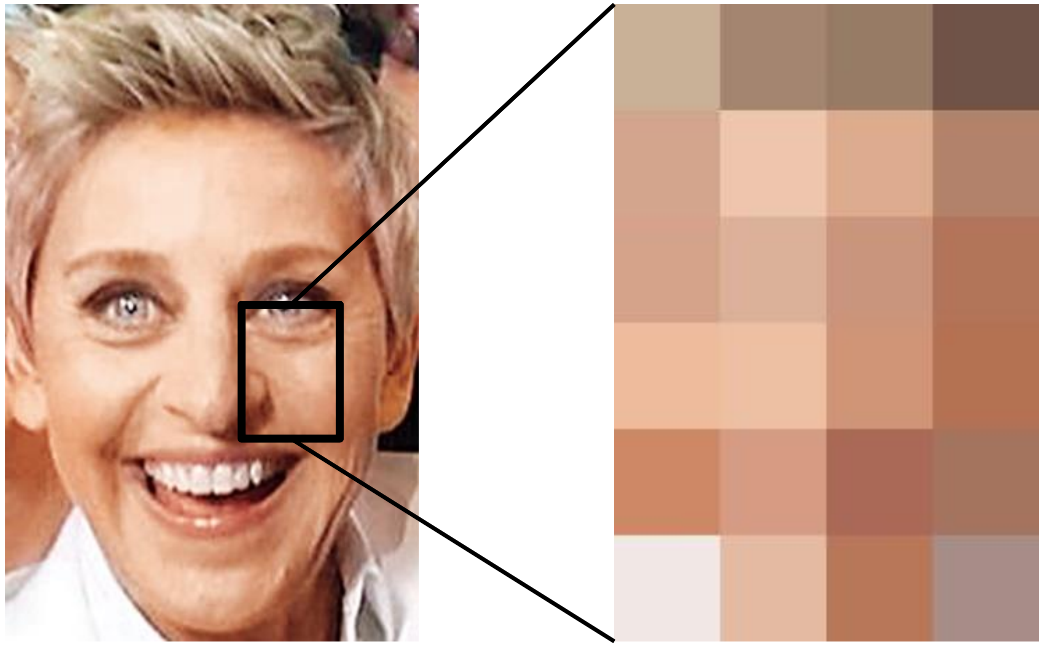
Email	Do	You	Want	Go	Million	Dollars	Dinner	Today
Do you want to go for dinner?	1	1	1	1	0	0	1	0
Do you want to win a million dollars?	1	1	1	0	1	1	0	0
I have a million things to do today!	1	0	0	0	1	0	0	1

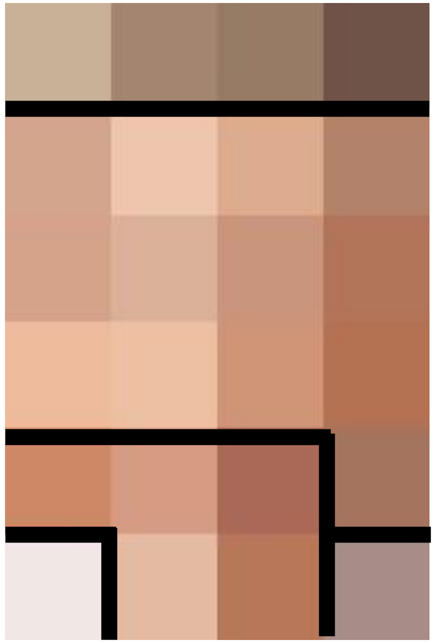


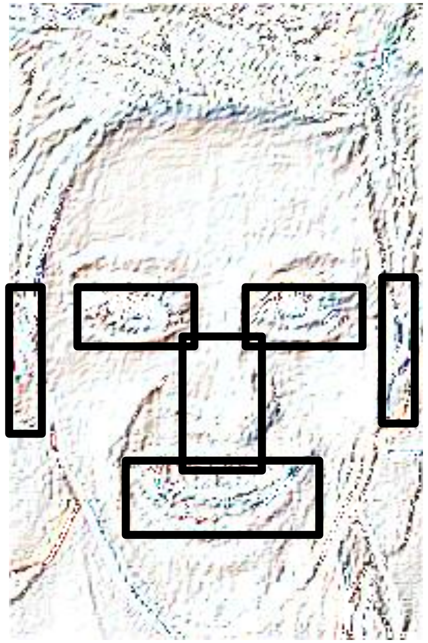
Since I am basically a computer program, I need you to convert your data into a nice set of numbers

Good catch – this may not be the best feature representation!

What are features



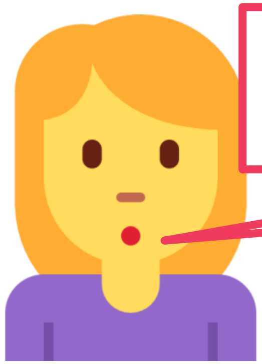




May represent images as a vector/matrix of pixel RGB values

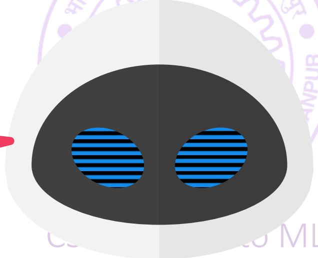
May instead encode images using edges present in them

May instead encode images using presence of eyes/noses/ears



Makes sense. Features are also a way for us to store what we know about data. If we throw away data, naturally doing anything becomes hard!

If your features are good, my job becomes twice as easy. If not, then doing any ML may become impossible!



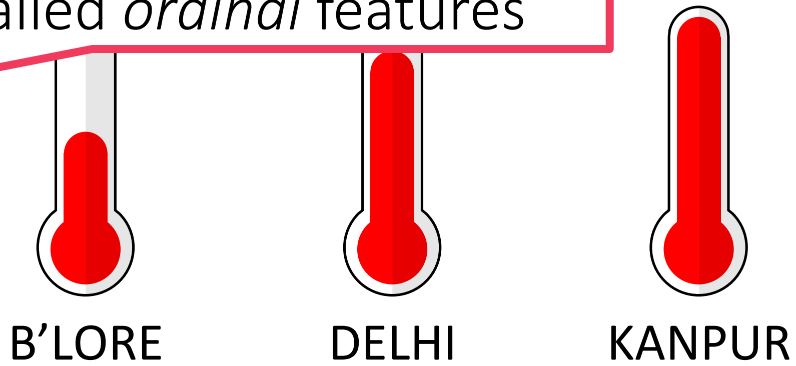
Types of Features

Discrete numerical features often called *ordinal* features

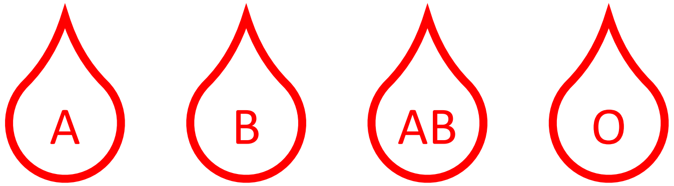
Numerical Features

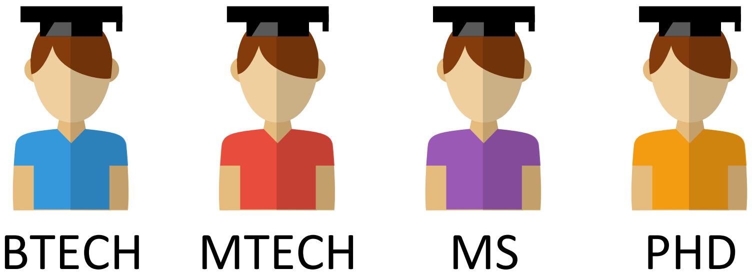
Dangal ★★★★★

Dhoom 3 ★★★

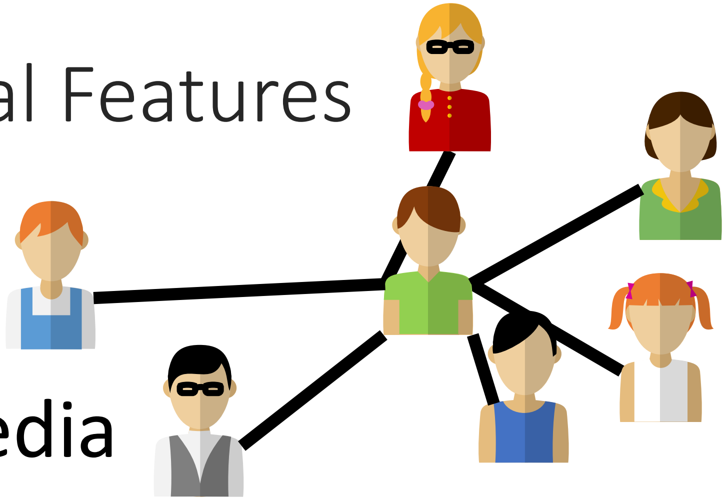


Categorical Features

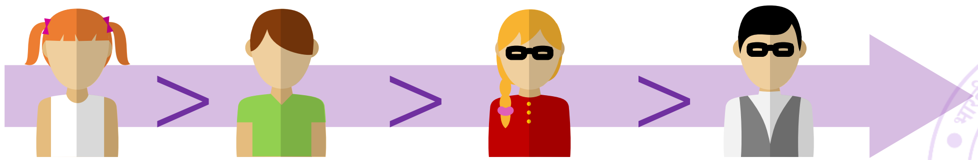




Relational Features



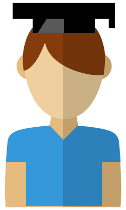
Social Media



Ranking in Class

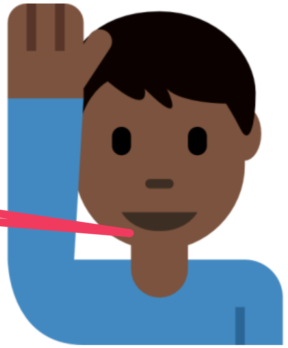
Derived Features

Bagged/binned features

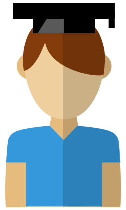


ESC101 (A), ESO207 (B), CS220 (B)
(C), MSO201 (A), CS771 (A

As we saw, the bag-of-words (also called BoW) is not always a very good feature (it confuses polysemous words). However, it is still extremely popular due to its simplicity



Pooled/aggregated features



ESC101 (A), ESO207 (B), CS220 (B), CS340
(C), MSO201 (A), CS771 (A)

⇒ **10 (max), 8.67 (avg)**

	Do	You	Want	Go	Million	Dollars	Dinner	Today
Do you want to go for dinner?	1	1	1	1	0	0	1	0
Do you want to win a million dollars?	1	1	1	0	1	1	0	0
I have a million things to do today!	1	0	0	0	1	0	0	1

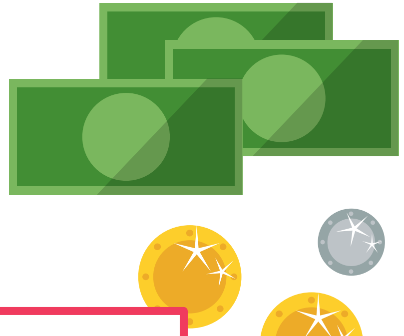
The way we represented emails in the spam problem is called the *bag-of-words* feature

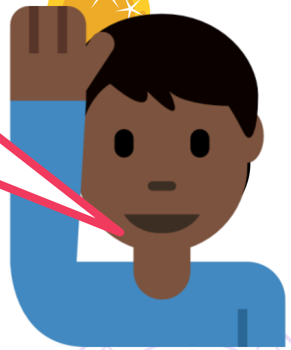
Feature Selection??

- DIET (VEG/NON-VEG)
- EYE COLOR
- TOTAL INCOME
- EDUCATION LEVEL
- FAMILY SIZE
- HOUSE NO (ODD/EVEN)
- NO OF CHIL
- RENTED/SE
- SURNAME

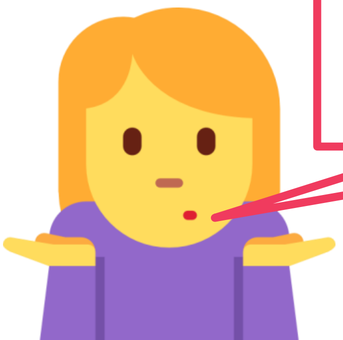
Useful for predicting expenditure

Not useful for predicting expenditure





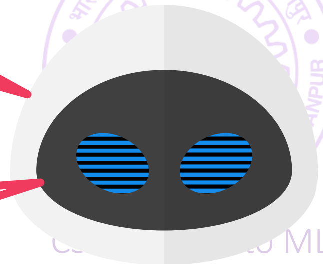
In fact, one reason for the success of deep learning is that it learns good features themselves!



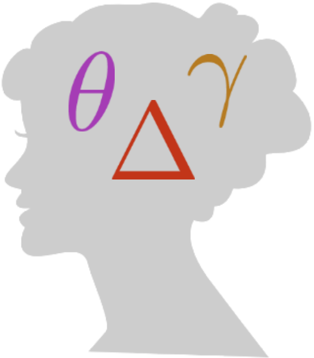
Can we perform automatic feature selection?

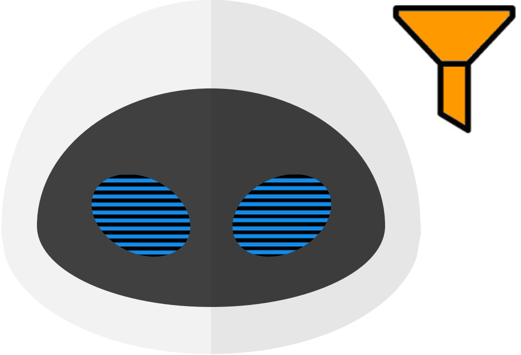
Yes indeed, but those methods are a bit advanced for now!

True, but more on deep learning later! For now, back to basics.



A bit of caution with features





Tricks, mnemonics lessen cognitive load, increase speed

Easy questions can be solved in one step with a mnemonic

Too many mnemonics ca confuse you at time of ex

Derived features make learning easier, faster at test

What you are trying to predict is just another feature of the data!

can

However, not to worry. For most of this course, we will give you pre-made feature vectors ☺

In fact, one of the main challenges in deep learning is that it learns way too many features