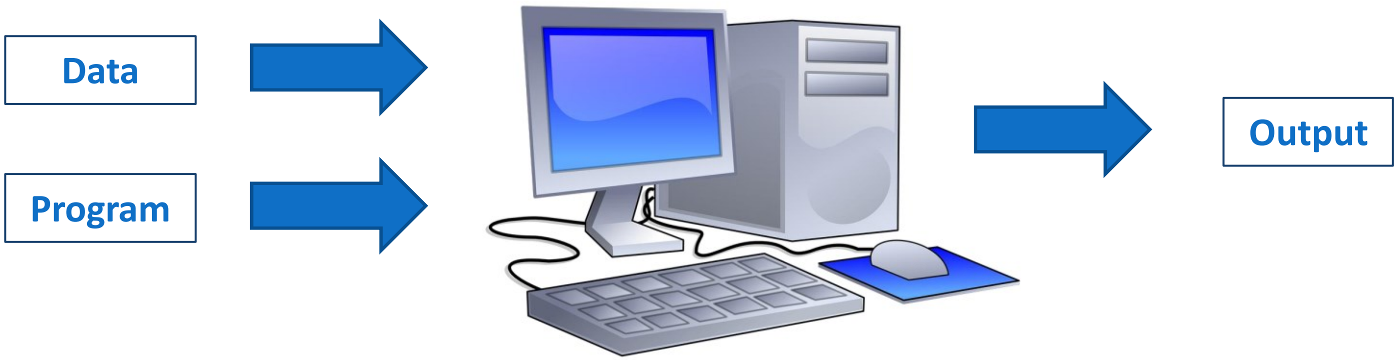
Review

MUHAMMAD HAROON SHAKEEL

Traditional Computer Science

- Tasks like:
  - Play an audio/video file
  - Display a text file on screen
  - Perform a mathematical operation on two numbers
  - Sort an array of numbers using *Insertion Sort*
  - Search for a string in a text file
  - …

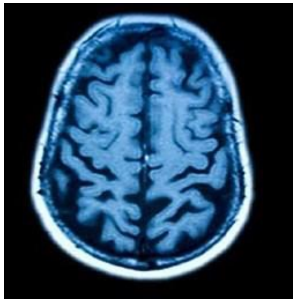


Data

Program

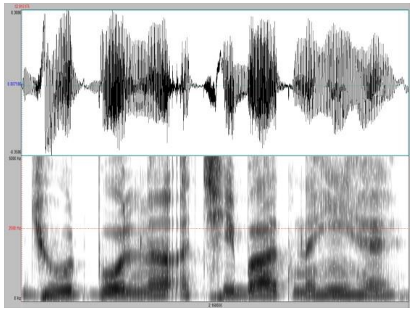
Output

Problems that Traditional CS Can’t Handle

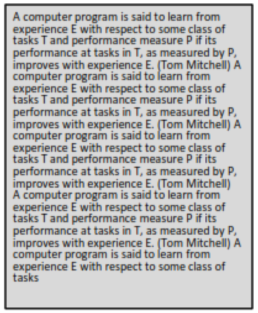


Tumor? Y/N

Price?



What was said?

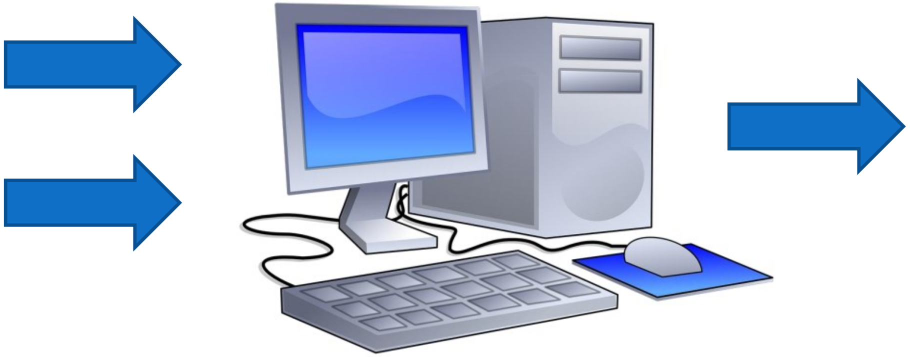


A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E. (Tom Mitchell) A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E. (Tom Mitchell) A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E. (Tom Mitchell) A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E. (Tom Mitchell) A computer program is said to learn from experience E with respect to some class of tasks

Summarize text

Data

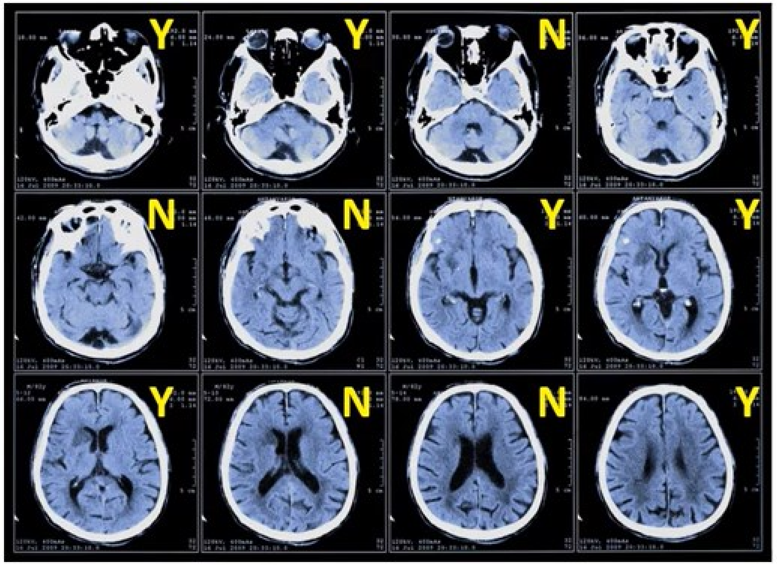
Program



Output

Machine Learning

Classification



Regression

$100,000

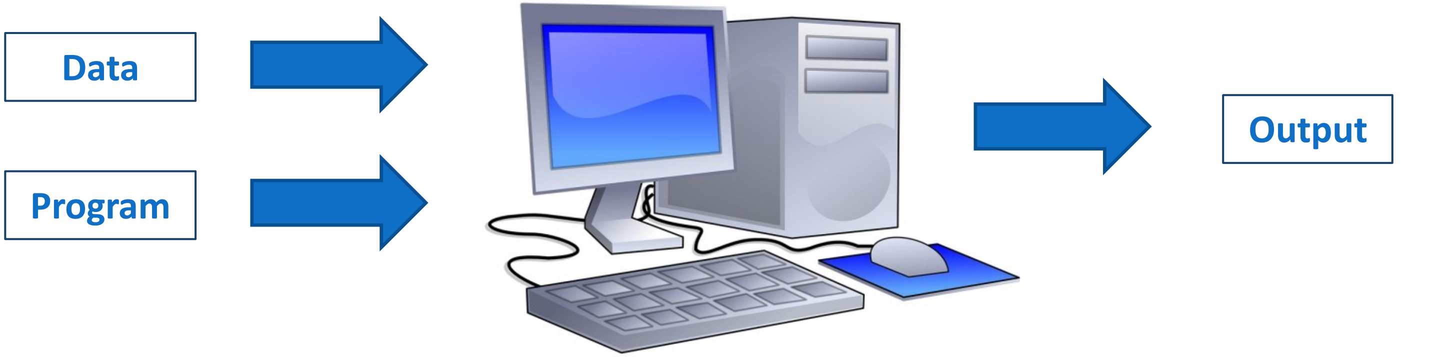
$140,000

$400,000

$250,000

$190,000

Traditional CS

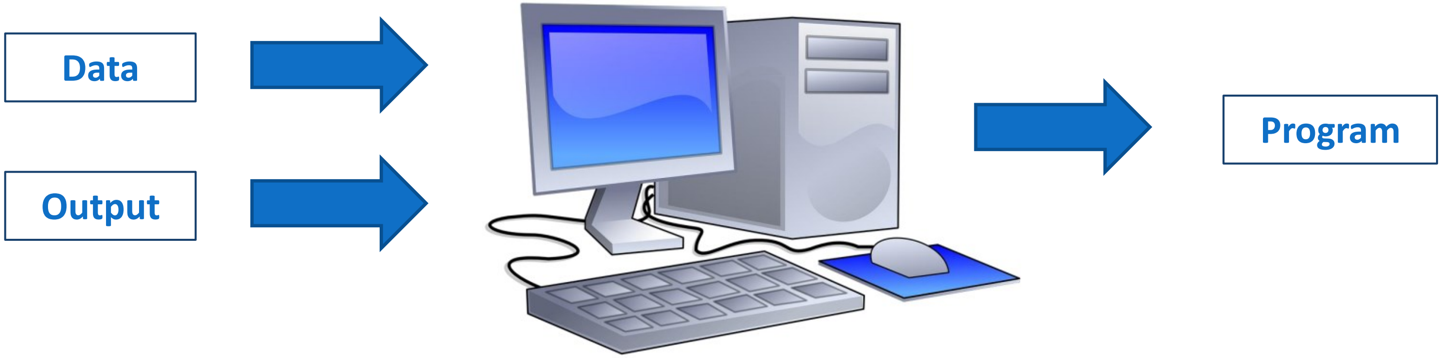


Data

Program

Output

Machine Learning

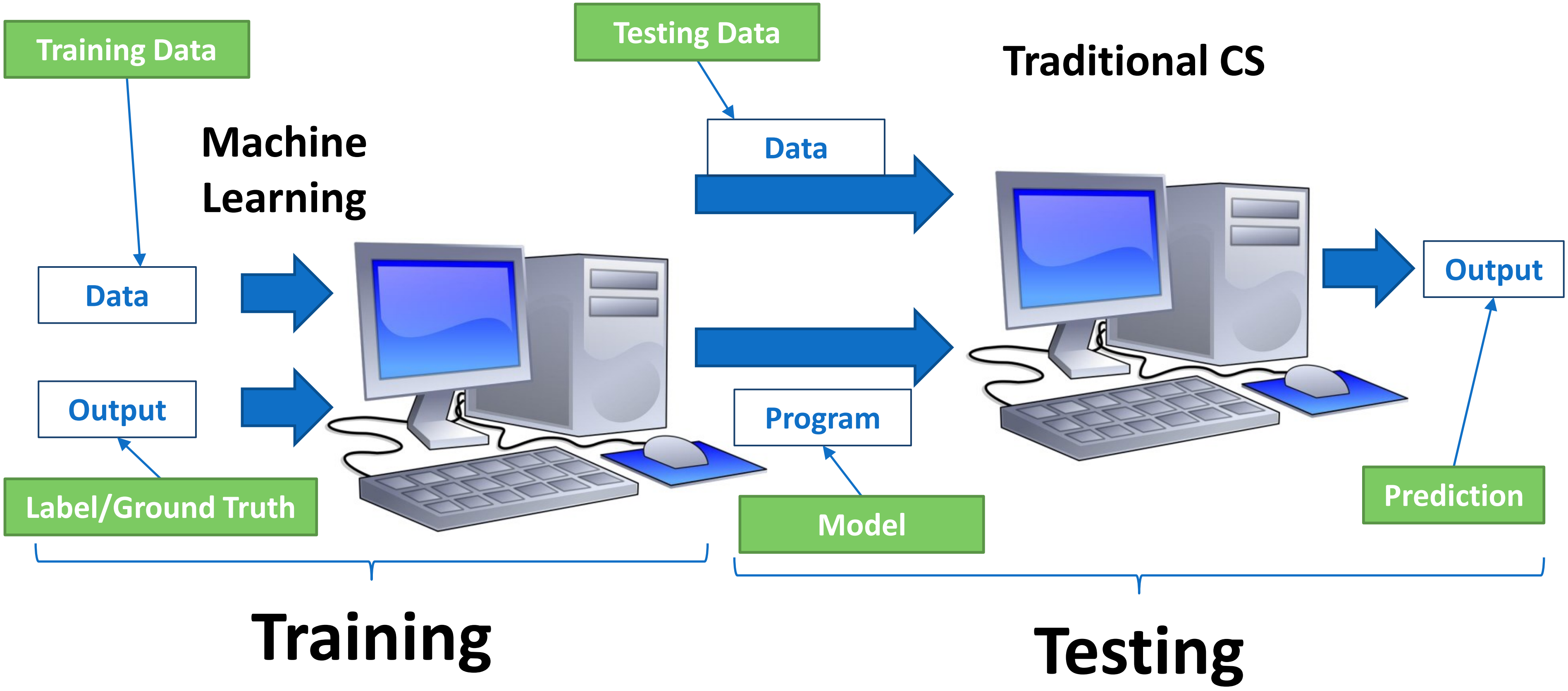


Data

Output

Program

Machine Learning Pipeline



What is Machine Learning?

❑ **Formally:**

- A computer program A is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E. (Tom Mitchell, 1997)

❑ **Informally:**

- Algorithms that improve on some task with experience.

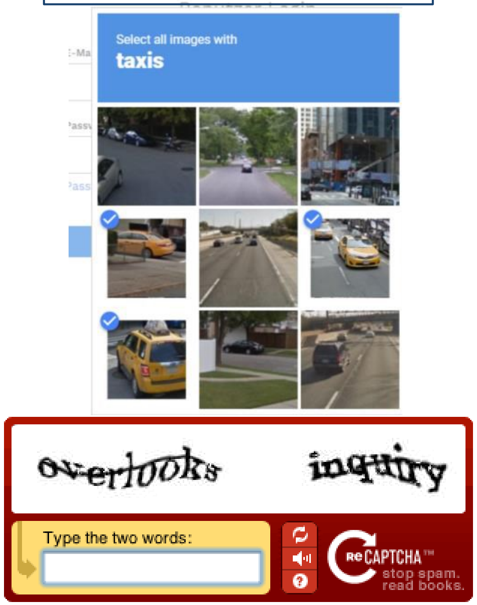
To train a classifier, we need data with labels! (called dataset)

Data – Big, Big,… data!

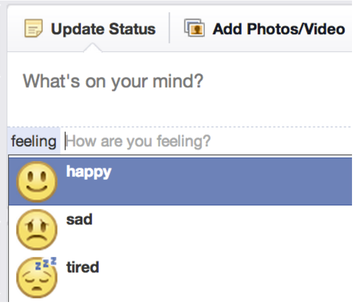
❑ **How do we obtain these massive datasets to train our Machine Learning models?**

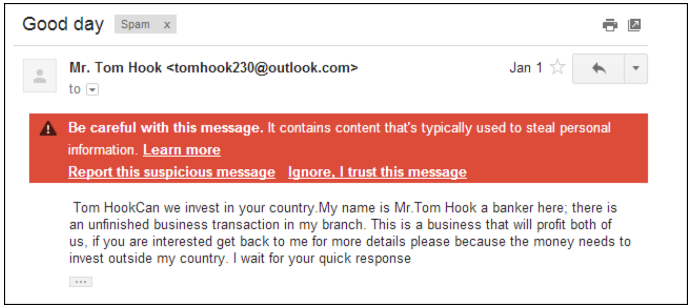
- From real interactions e.g., call centers
- Expert annotators e.g., hired tams of annotators
- Crowd sourcing

Recaptcha



Tagging





Good day Spam x

Mr. Tom Hook <tomhook230@outlook.com> Jan 1

to

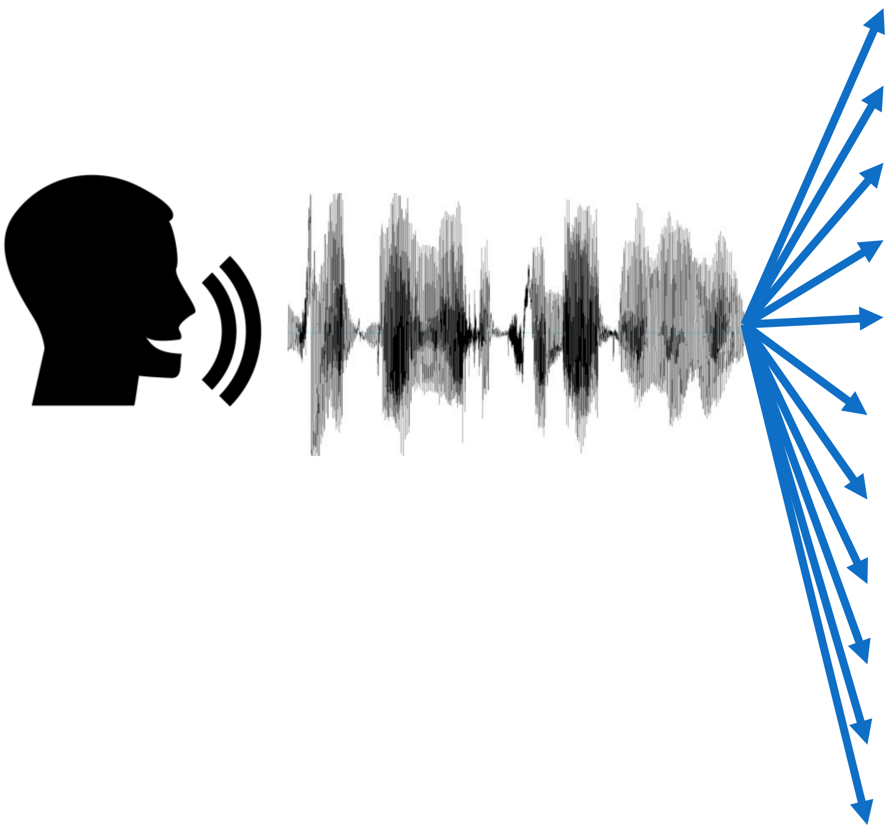
Be careful with this message. It contains content that's typically used to steal personal information. Learn more

Report this suspicious message Ignore, I trust this message

Tom HookCan we invest in your country.My name is Mr.Tom Hook a banker here; there is an unfinished business transaction in my branch. This is a business that will profit both of us, if you are interested get back to me for more details please because the money needs to invest outside my country. I wait for your quick response

Task-Label Relationship

- Labels are dictated by the task to be performed.
- **Example:** Speech Technologies



What was said? — Speech Recognition

Who said it? — Speaker Recognition

Was it John Doe? — Speaker Verification

Did it mention “hey Google”? — Keyword Detection

What’s the language? — Language Identification

Is the language native for the speaker?

What is their height?

What is the age of the speaker?

What is emotional state?

What was the sentiment?

Is the voice fake?

Task-Label Relationship

- **Example:** Text Technologies

Who wrote it?

Summary of what was written?

Was it plagiarized?

What was the intent?

What language is this?

Is the language native for the speaker?

What is author’s literacy level?

What is the topic of this document?

What is emotional state?

What was the sentiment?

Can we fake this writing style?

Challenges of ML - Explainability

- A classifier can potentially learn to classify on the basis of features not desirable for humans
  - All dogs waring a collar in the training data while no cat is wearing it – ML just learns to separate based on collar
  - All horse images have a copyrights notice – ML just learns to recognize horses based on the copyrights notice

- **Explainable ML:** The results should be understandable by humans
  - As opposed to a black-box system

Challenges of ML – Fairness

- ❑AI tends to reflect the biases of the society
  - Human taggers who mark a recording as misinformation based on accent or gender
  - Court decisions in country that make a rich person’s acquittal more likely
  - Automated standardized testing in the US could yield unfavorable results for certain demographic groups
  - AI plays a decision role in hiring decisions, with up to 72% of resumes in the US never being viewed by a human **(Automation Bias)**
  - Decision on immigration, bank loans, credit history checks, criminal profiling

ML in Low-resource settings

- ❑ Problems where large datasets and tools are not available
- ❑ Natural Language Processing and Speech
  - Pakistan has 71 languages
  - We barely have speech recognition capabilities for Urdu!

Why is it important?

Offline and Verbal Communicators!

The Offline Ones

- ❑ 3.6 billion people worldwide are offline
  - That is 46.6% of the world population
  - 13.4% of the develop world, 53% of the developing world, and 80.9% of the Least Developed Countries are offline*
- ❑ Offline Populations
  - Too poor to afford internet-enabled devices
  - Too remote to access the internet
  - Too low-literate to navigate the mostly text-driven internet
- ❑ 285 million visually impaired individuals

Speech-based services are becoming a necessity.

*International Telecommunication Union (ITU), https://itu.foleon.com/itu/measuring-digital-development/offline-population/ (accessed: Feb 2023)

Developing Datasets

How to store the labeled dataset?

- ❑The best way (arguably) is **Tabular Format**
- ❑But real-world data does not exist in Tables.
- ❑Needs some preprocessing step to make the data useable for ML algorithms.
- ❑If data is in the form of Tables (Rows and Columns)
  - **What should be the names of columns and rows?**

Developing Dataset

It's a good idea to use generic terms for column names.

setosa

versicolor

virginica

	X				Y
	x_1	x_2	x_3	x_4	y
	~~Sepal Length~~	~~Sepal Width~~	~~Petal Length~~	~~Petal Width~~	~~Species/Label/ Class~~
x^1 or $\overrightarrow{x_1}$	5.1	3.5	1.4	0.2	0
x^2	4.9	3.0	1.4	0.2	0
x^3	7.0	3.2	4.7	1.4	1
x^4	6.4	3.2	4.5	1.5	1
x^5	6.3	3.3	6.0	2.5	2
x^6	5.8	3.3	6.0	2.5	2

Labels are mapped to numerical IDs (most of the time)

```
Labels2Idx = {"Setosa":0, "Versicolor":1, "Virginica":2}
Idx2Labels = {0:"Setosa", 1:"Versicolor", 2:"Virginica"}
```

What if data is not tabular?

❑ **We need to preprocess it and convert into tabular format (but not always!)**

❑ Deep learning-based classifiers/models can work on 2D, 3D or 4D arrays directly

- Note: Array for each record/instance can be of size 2D, 3D or 4D

❑ For now, we will limit ourselves to converting each record to 1D array (aka a vector)

Implementation Tip:
It’s a good idea to
keep X and Y in
separate variables.

Feature Space: Tabular Data

Features/Dimensions

Label/Class/Category

Height (inches)	Weight (kgs)	B.P.Sys	B.P.Dia	Heart disease
62	70	120	80	No
72	90	110	70	No
74	80	130	70	No
65	120	150	90	Yes
67	100	140	85	Yes
64	110	130	90	No
69	150	170	100	Yes
66	125	145	90	?
74	67	110	60	?

Record is 4-dimensional Feature Vector

Training Data/Training Split

Testing Data/Testing Split

As labels are discrete, this is a classification task and a classifier would be trained for predictions.

Feature Space: Tabular Data

Features/Dimensions

Label

Height (inches)	Weight (kgs)	B.P.Sys	B.P.Dia	Cholesterol Level
62	70	120	80	150
72	90	110	70	160
74	80	130	70	130
65	120	150	90	200
67	100	140	85	190
64	110	130	90	130
69	150	170	100	250
66	125	145	90	?
74	67	110	60	?

A Record is 4-dimensional Feature Vector

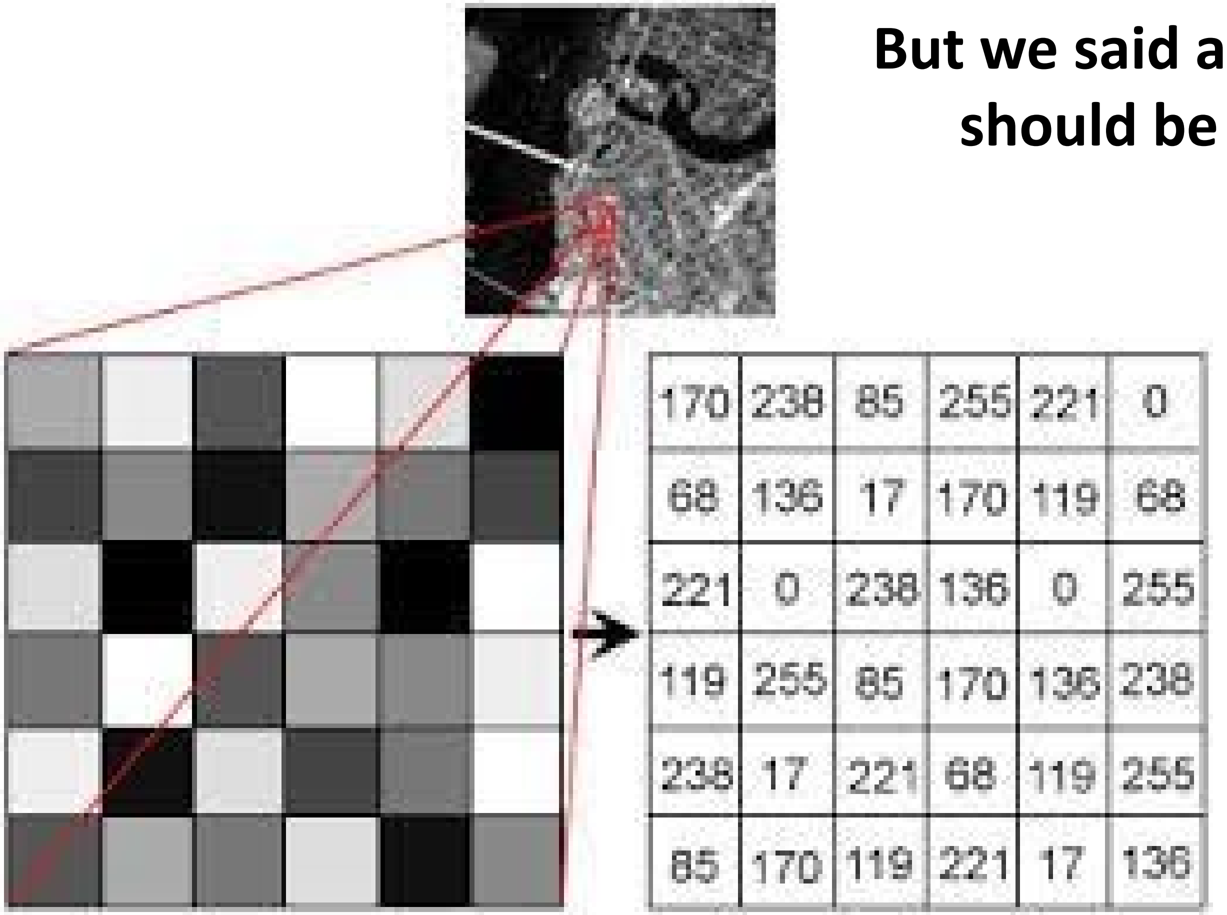
Training Data/Training Split

Testing Data/Testing Split

As labels are continuous, this is a regression task and a regressor would be trained for predictions.

Feature Space: Image Data

❑Images are nothing but a **2D/3D arrays** with values of color intensities, typically ranging $\mathbf{0-255}$



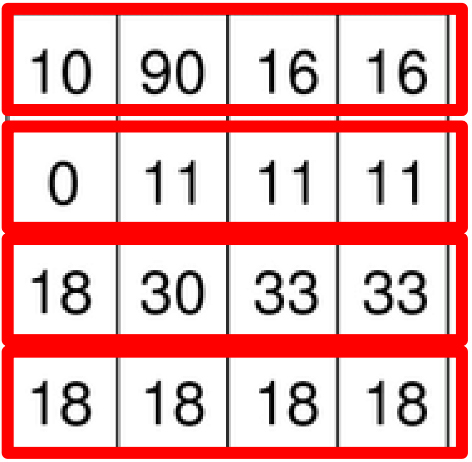
170	238	85	255	221	0
68	136	17	170	119	68
221	0	238	136	0	255
119	255	85	170	136	238
238	17	221	68	119	255
85	170	119	221	17	136

But we said a record should be 1D!

Feature Space: Image Data

- Images are nothing but a **2D/3D arrays** with values of color intensities, typically ranging $\mathbf{0-255}$

Do this for all of your images and now each record is a vector!



10	90	16	16
0	11	11	11
18	30	33	33
18	18	18	18

10	90	16	16	0	11	11	11	18	30	33	33	18	18	18	18

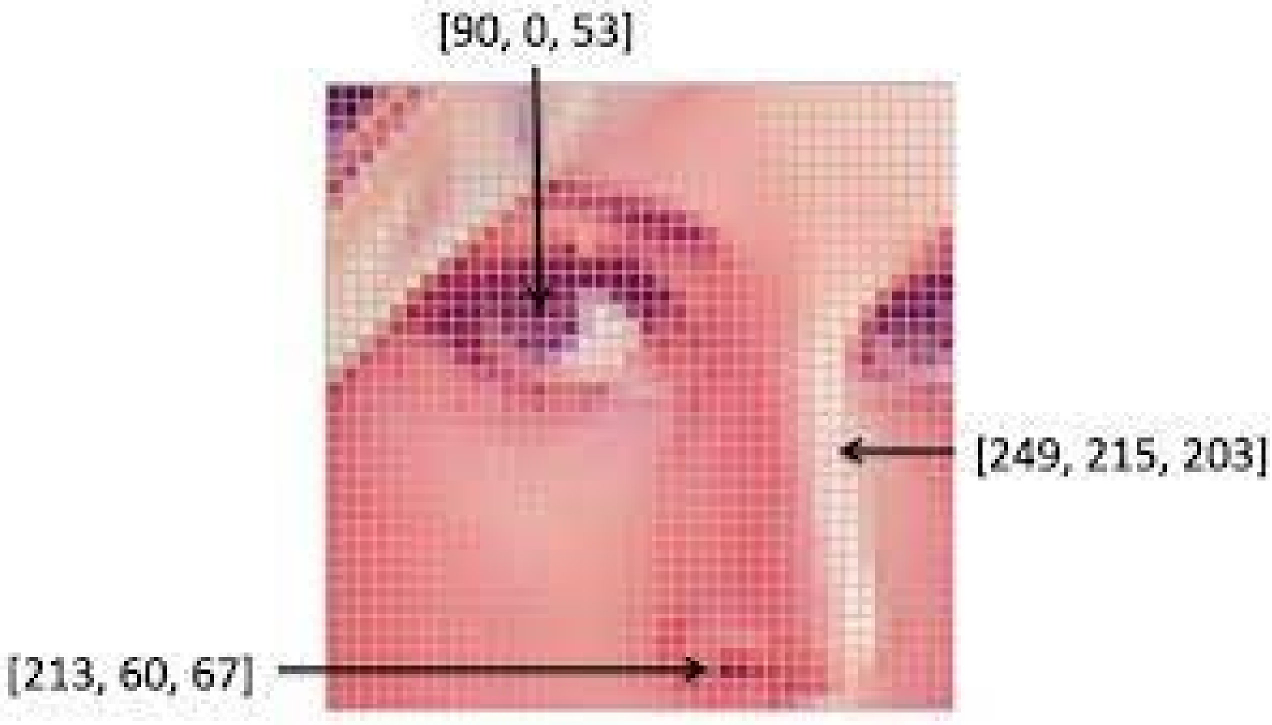
The label is stored separately for corresponding record.

If label is "Me" and "Not Me", it's classification.
If label is "Age", its regression.

Implementation Tip:
Use numpy reshape.

Feature Space: Image Data

- ❑The color Image is 3D array $(Width \times Height \times Channels)$
- ❑Color image has 3 channels while grayscale image has 1 channel.



How would you convert color image to 1D array?

Feature Space: Text Data

- □ Suppose you are given labeled textual data in excel sheet

	Document#	Text	Class
Training	1	The Best movie best	Pos
	2	The Best best ever	Pos
	3	The Best film	Pos
	4	The Worst cast ever	Neg
Testing	5	The Best best best worst ever	?

the	best	movie	ever	film	worst	cast
1	1	1	0	0	0	0
1	1	0	1	0	0	0
1	1	0	0	1	0	0
1	0	0	1	0	1	1

label
1
1
1
0

1	1	0	1	0	1	0

?

Assignment 1 – Task 0

- ❑Make a group of 2 people; try to diversify.

- ❑Install Anaconda Python Distribution on your computer

- ❑Or get familiar with Google Colab.

Assignment 1 – Task 1

- The task is to train a classifier that classifies your images into “Your name” and “Unknown”.

- For this purpose, gather your photos – more the better.

- Gather photos of your friends/colleagues/celebrities etc. separately.

- Label your photos. Split them into train and test data.

- How are you going to do that?
    - There are many ways!
    - Make train and test folders and then two subfolders in each, called “Your name” and “Unknown”.
    - Or….
    - Make two excel files for training and testing and place the paths of the images in one column, and label in the other. **(Should be preferred as it will come in handy for defining multiple labels).**
    - …

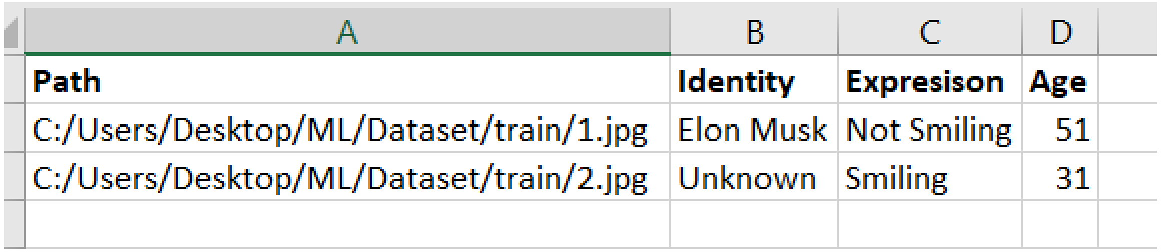
Assignment 1 – Task 1

□ Now define two more columns for labels

- Age (In years)
- Expression (Smiling, not smiling)

□ This way, for each image, you have three labels for each split

- Age (Age of person in years)
- Expression (Smiling, not smiling)
- Identity (Your name, unknown)



A	B	C	D
Path	**Identity**	**Expresison**	**Age**
C:/Users/Desktop/ML/Dataset/train/1.jpg	Elon Musk	Not Smiling	51
C:/Users/Desktop/ML/Dataset/train/2.jpg	Unknown	Smiling	31

□ Read those images and convert both train and test splits into array of feature vectors along with their corresponding labels stored separately.

- (X_train, X_test, y1_train, y2_train, y3_train, y1_test, y2_test, y3_test) – This is just an example of variable names. Use variable names as per your convenience.

Book Reading

❑ Murphy – Chapter 1