Introduction to Machine Learning

By

Dr. Prashant Singh Rana

Computer Science & Engineering Department,

Thapar Institute of Engineering and Technology,

Patiala, Punjab.

www.psrana.com | psrana@gmail.com

Lets Start

What is Machine Learning ?

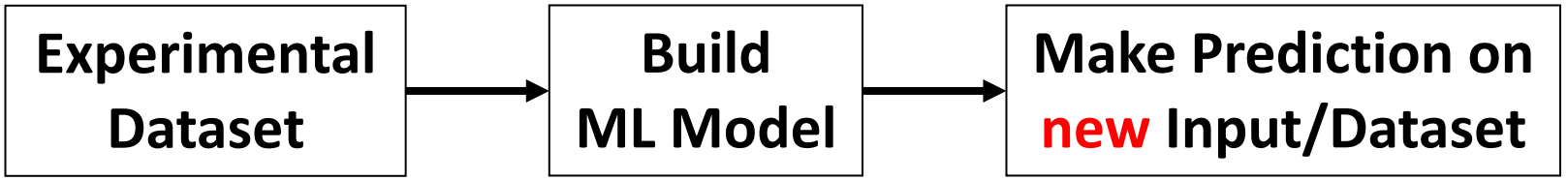
Explore

- Go to google and search for “**Teachable Machine**”

or

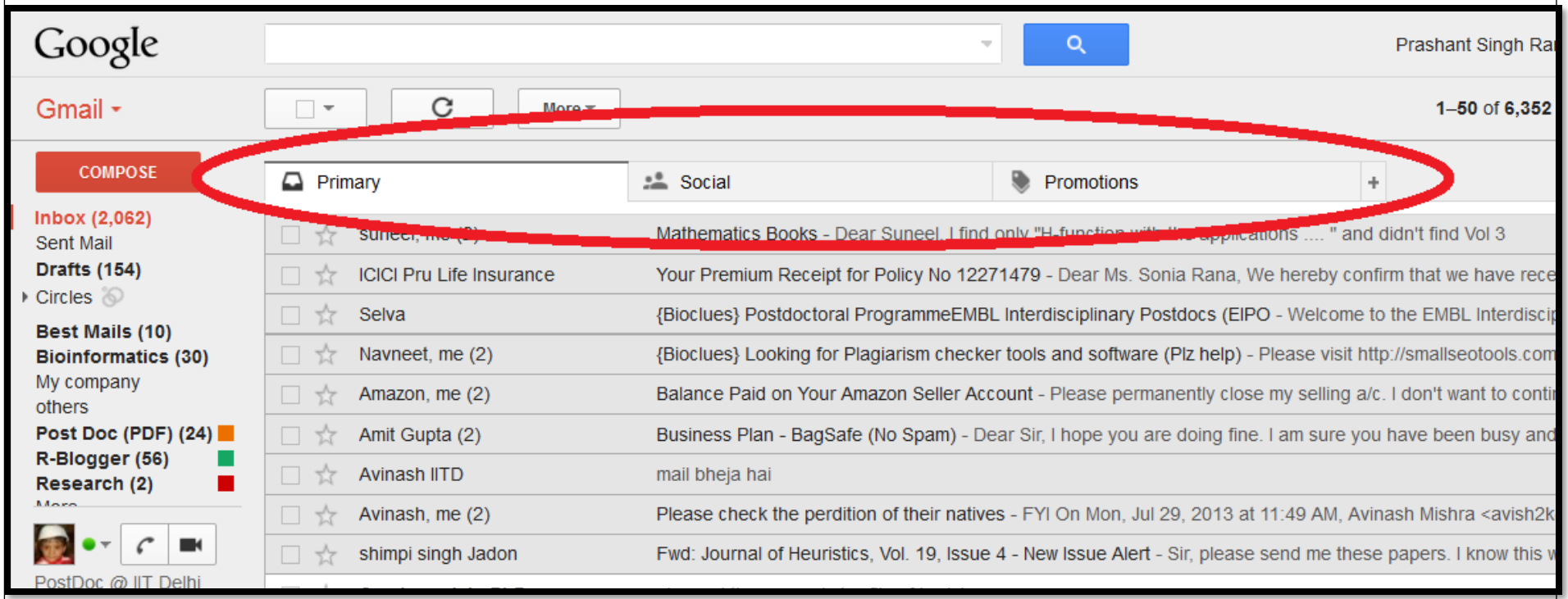
- Click on “https://teachablemachine.withgoogle.com”

Basic Idea for ML



What is Machine Learning ?

Example1: email classification in Gmail

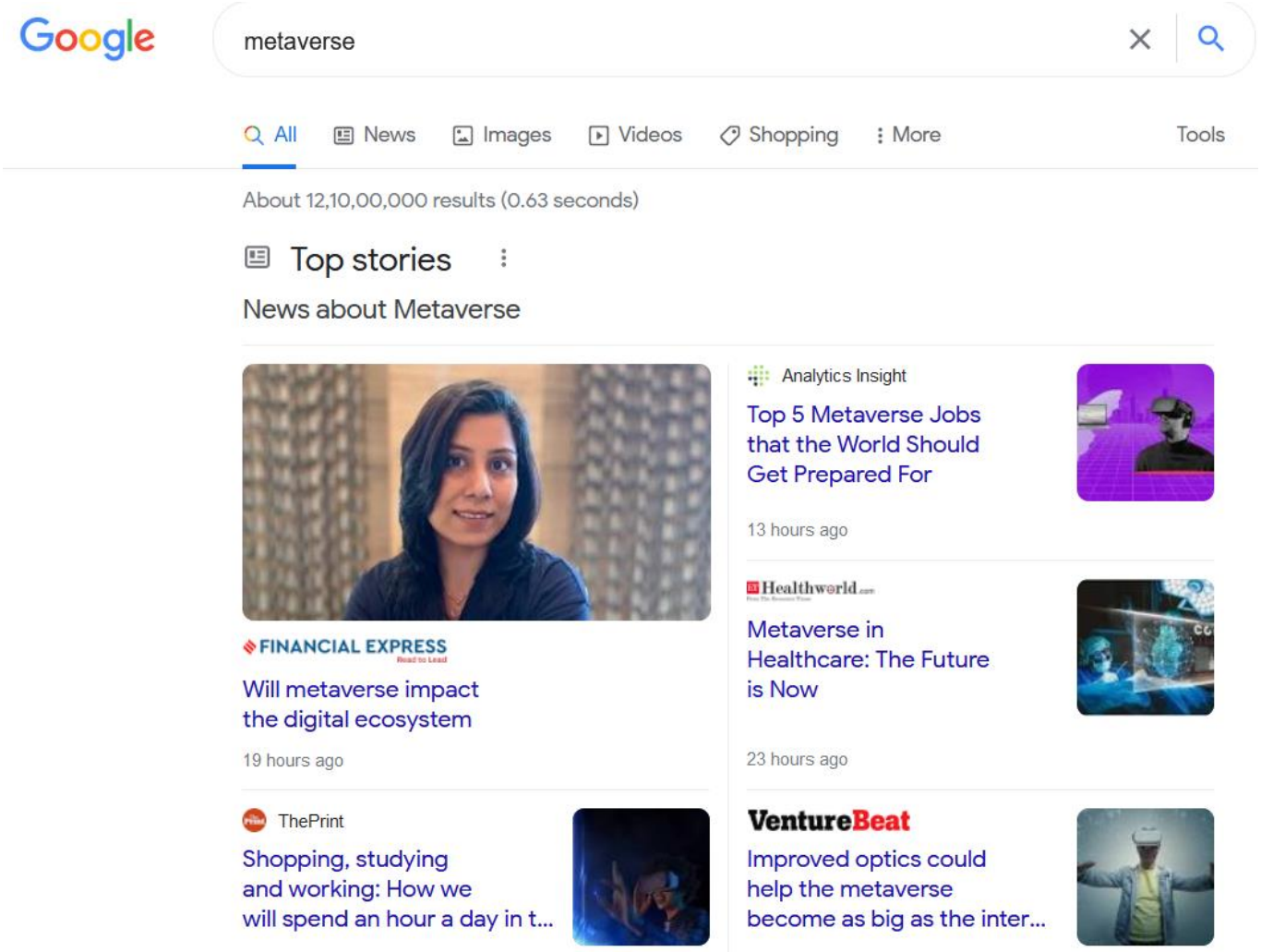


What is Machine Learning ?

Example2: suggestion in youtube

What is Machine Learning ?

Example3: Google News → Search for "Metaverse"



What is Machine Learning ?

Some more examples

1. Classification of e-mail into **SPAM or NON-SPAM**
2. Classification of transaction as **FRAUD or GENUINE**
3. Patients diagnosed as **DIABETIC or NON-DIABETIC**
4. Gene Classification into **CODING or NON-CODING**.

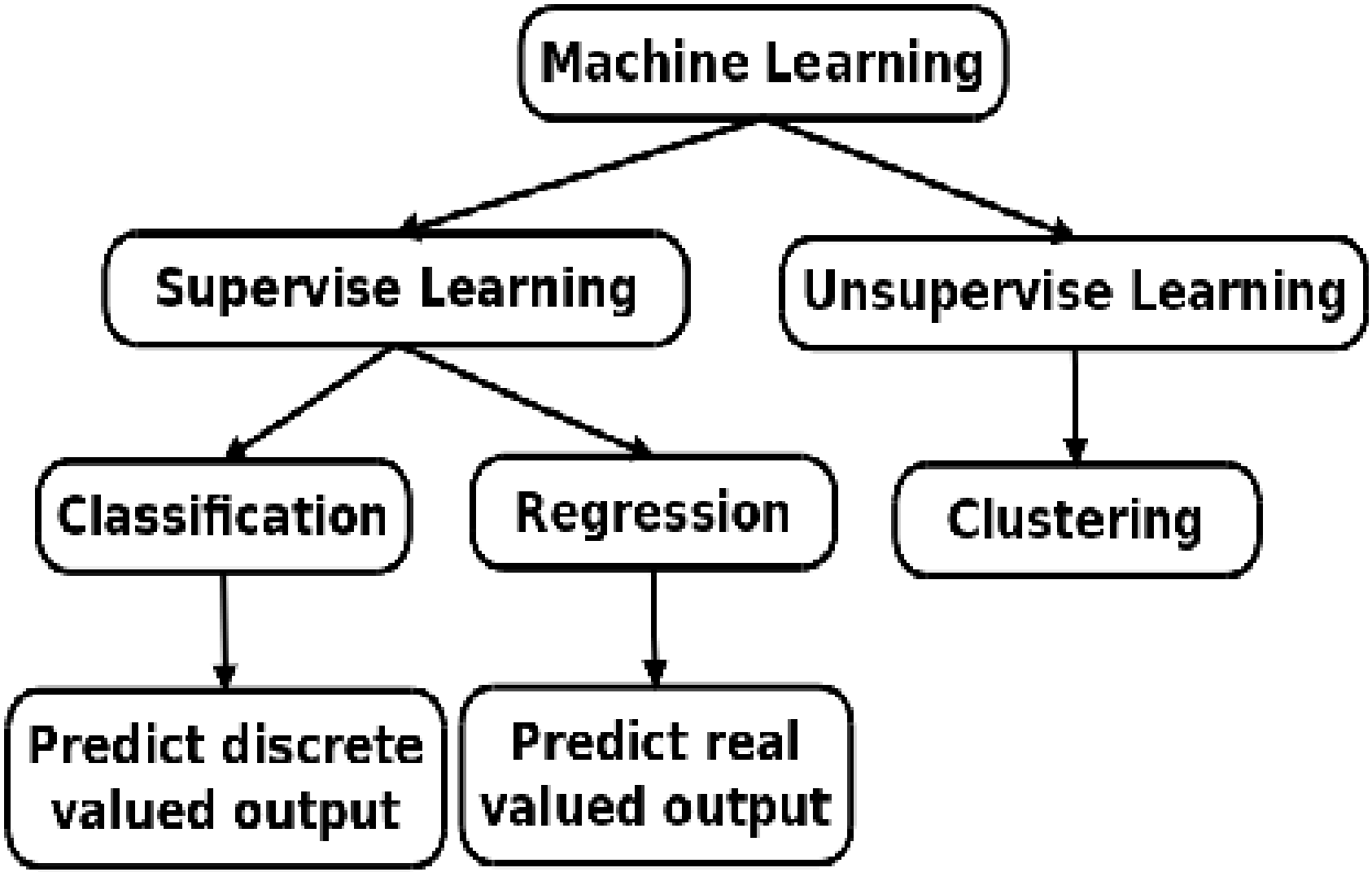
……. Many more.

What is Machine Learning ?

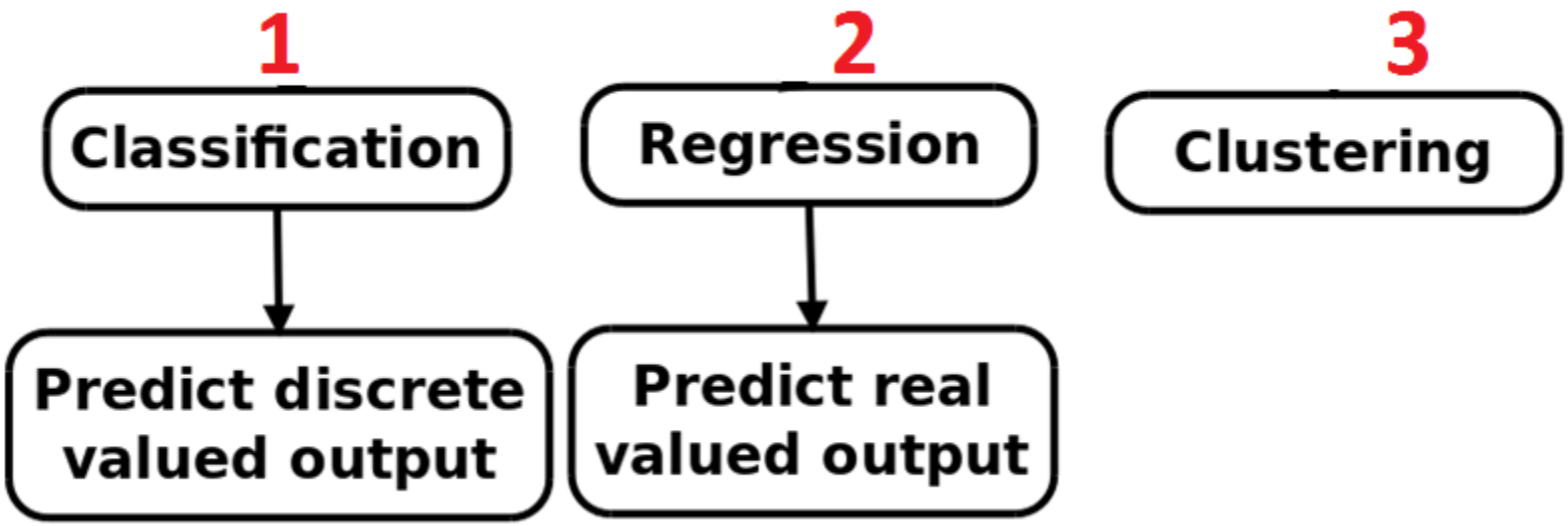
What is Machine Learning ?

- **Simple Definition I** - Branch of Artificial Intelligence that gives **computers to learn** without being explicitly programmed.
- **Simple Definition II** - Branch of Artificial Intelligence, about to construct a system that **learn from data**.
- **Actual Definition** - A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

Categories of Machine Learning



You have to only deals with .....



Examples

- **Regression Problem**
  - Prediction of wheat production.
  - Prediction of rainfall.
  - Point prediction of Stock Exchange.
- **Classification Problems**
  - Prediction of cancer.
  - Win prediction of Sheila Dixit.
  - Diabetic Prediction.
  - Classification of e-mail.

Examples

- **Clustering**
  - Grouping of NEWS.
  - Grouping the people on their similar hobbies/interests.
  - Grouping of animals.
  - Grouping of customers based on their performance. E.g. bank customers.

…….. Many more.

Data Set format for Machine Learning

Explore

- Sample Dataset

Understand the Data.....

Features / Properties

Class / Target

	A	B	C	D	E	F	G	H	I	J	K
1	Code	Clump_Th	Cell_Size	Cell_Shap	Marginal A	Single Epi	Bare Nucl	Bland Chr	Normal N	Mitoses	Class
2	1000025	5	1	1	1	2	1	3	1	1	2
3	1002945	5	4	4	5	7	10	3	2	1	2
4	1015425	3	1	1	1	2	2	3	1	1	2
5	1016277	6	8	8	1	3	4	3	7	1	2
6	1017023	4	1	1	3	2	1	3	1	1	2
7	1017122	8	10	10	8	7	10	9	7	1	4
8	1018099	1	1	1	1	2	10	3	1	1	2
9	1018561	2	1	2	1	2	1	3	1	1	2
10	1033078	2	1	1	1	2	1	1	1	5	2
11	1033078	4	2	1	1	2	1	2	1	1	2
12	1035283	1	1	1	1	1	1	3	1	1	2
13	1036172	2	1	1	1	2	1	2	1	1	2
14	1041801	5	3	3	3	2	3	4	4	1	4
15	1043999	1	1	1	1	2	3	3	1	1	2
16	1044572	8	7	5	10	7	9	5	5	4	4

Classification & Regression

Classification : Predict **discrete valued** output.

Regression : Predict **real valued** output.

Class	F1	F2	F3	F4	F5
5	5769.3	1634.9	0.3	57.0	76946
3	12962.3	3389.2	0.3	141.3	17618
13	5960.2	2230.7	0.4	64.3	84555
11	9926.8	3276.7	0.3	102.0	13869
15	6658.5	2590.6	0.4	62.2	95121
3	12272.7	2836.1	0.2	140.0	16656
2	12579.2	3473.6	0.3	129.4	17371
19	11969.7	4721.9	0.4	110.1	15700
19	21779.3	8269.9	0.4	250.2	29574
20	9020.8	2509.4	0.3	97.9	12392

1 Classification Data

Class	F1	F2	F3	F4
4.5	5769.3	1634.9	0.3	57.0
3.0	12962.3	3389.2	0.3	141.3
12.7	5960.2	2230.7	0.4	64.3
11.5	9926.8	3276.7	0.3	102.0
14.9	6658.5	2590.6	0.4	62.2
2.5	12272.7	2836.1	0.2	140.0
2.2	12579.2	3473.6	0.3	129.4
18.8	11969.7	4721.9	0.4	110.1
19.4	21779.3	8269.9	0.4	250.2
19.6	9020.8	2509.4	0.3	97.9

2 Regression Data

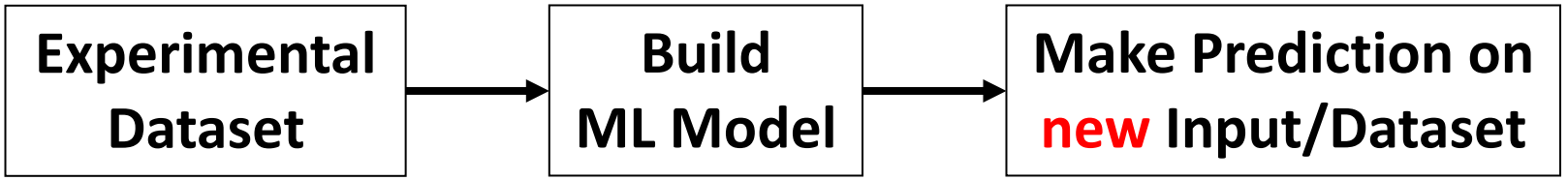
Clustering Data

Features / Properties

	A	B	C	D	E	F	G	H	I	J
1	Code	Clump_Th	Cell_Size	Cell_Shap	Marginal A	Single Epi	Bare Nucl	Bland Chr	Normal Nu	Mitoses
2	1000025	5	1	1	1	2	1	3	1	1
3	1002945	5	4	4	5	7	10	3	2	1
4	1015425	3	1	1	1	2	2	3	1	1
5	1016277	6	8	8	1	3	4	3	7	1
6	1017023	4	1	1	3	2	1	3	1	1
7	1017122	8	10	10	8	7	10	9	7	1
8	1018099	1	1	1	1	2	10	3	1	1
9	1018561	2	1	2	1	2	1	3	1	1
10	1033078	2	1	1	1	2	1	1	1	5
11	1033078	4	2	1	1	2	1	2	1	1
12	1035283	1	1	1	1	1	1	3	1	1
13	1036172	2	1	1	1	2	1	2	1	1
14	1041801	5	3	3	3	2	3	4	4	1
15	1043999	1	1	1	1	2	3	3	1	1
16	1044572	8	7	5	10	7	9	5	5	4

Only Features; No class/target/label

Basic Idea for ML



Example1: Applying Liner Model

1. Experimental Dataset

X1	X2	X3	X4	Output
5	1	1	3	25
1	5	5	1	20
2	3	1	3	27
4	2	5	4	29
1	4	5	5	30
4	5	3	2	21
5	5	1	1	30
3	2	1	4	29

Example1: Applying Liner Model

2. Build ML Model

Linear Model

$$Y = w1\,x1 + w2x2 + w3x3 + w4x4$$

$$\text{Output} = 9x1 + 6x2 - 10x3 + 5x4$$

Example1: Applying Liner Model

3. Make Prediction on new Input/Dataset

new Input →

X1	X2	X3	X4	Output
3	4	2	2	?

Apply Linear Model

Output = 9x1 + 6x2 - 10x3 + 5x4

= 9*3 + 6*4 – 10*2 + 5*2

= 41

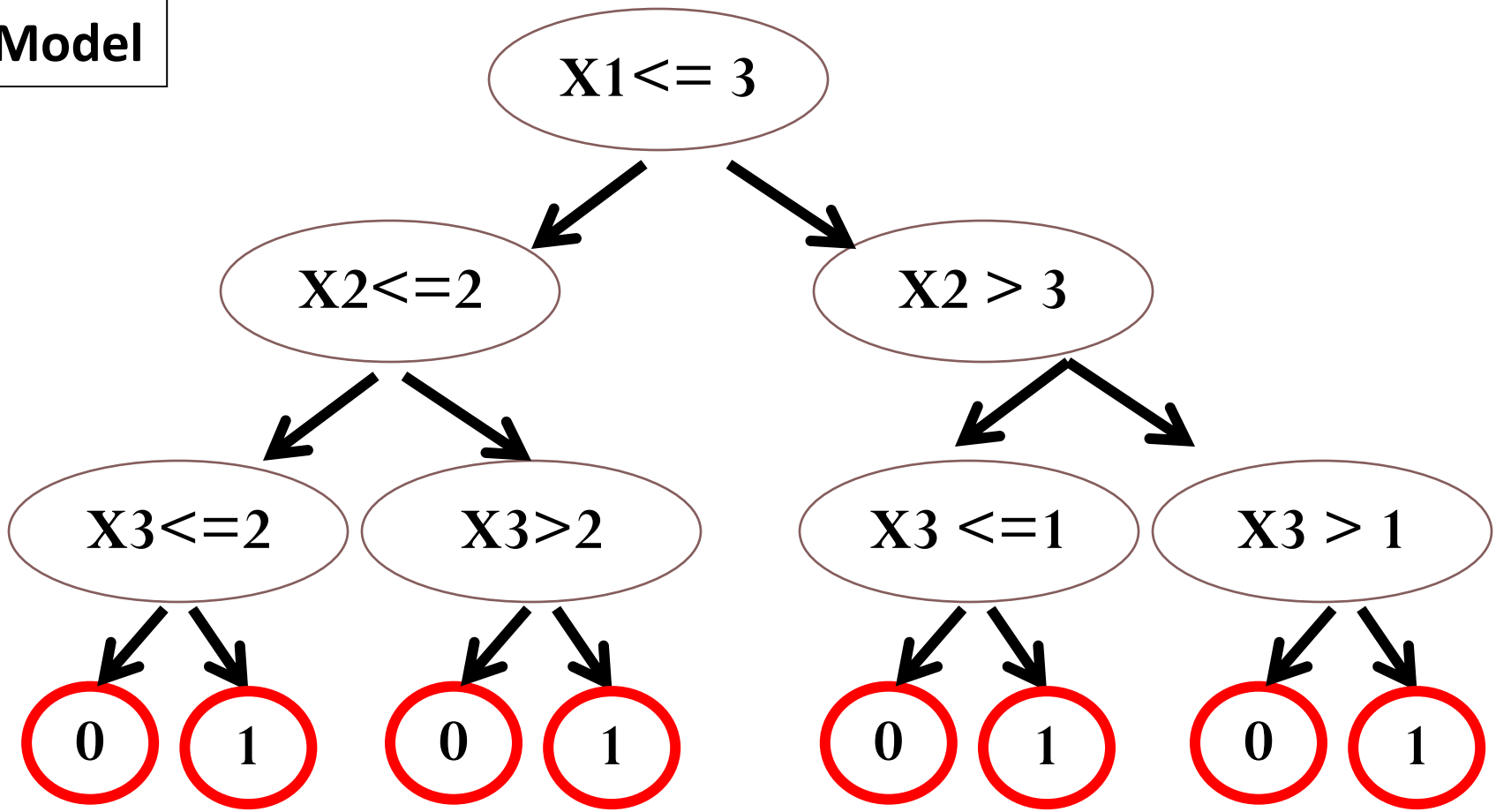
Example2: Applying Decision Tree

1. Experimental Dataset

X1	X2	X3	Output
5	1	1	0
1	5	5	0
2	3	1	1
4	2	5	0
1	4	5	1
4	5	3	1
5	5	1	0
3	2	1	1

Example2: Applying Decision Tree

2. Build ML Model

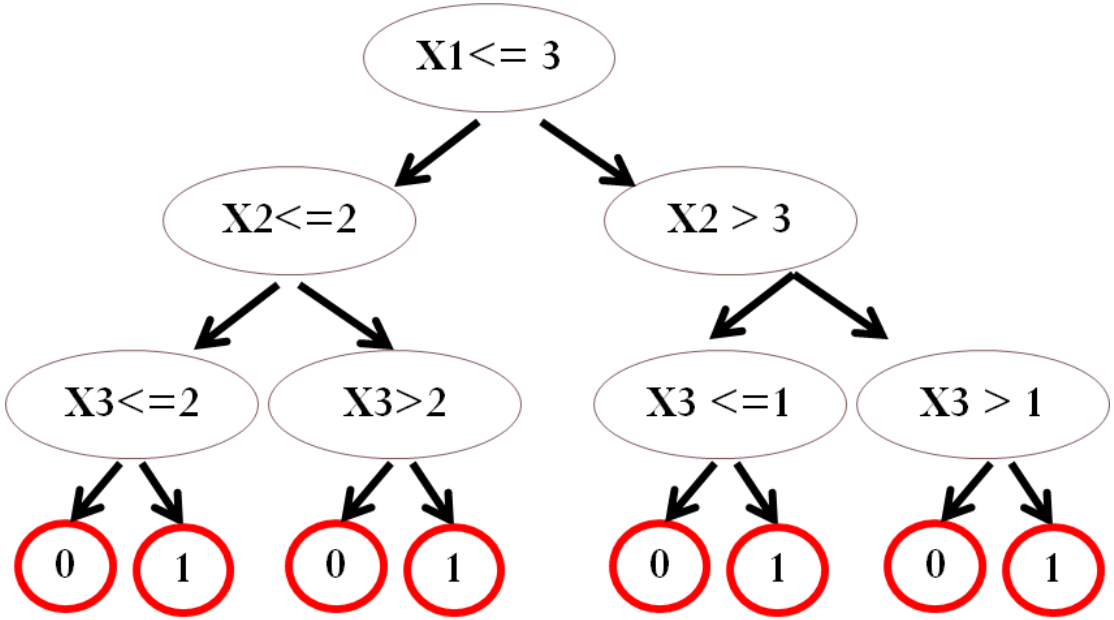


Example2: Applying Decision Tree

3. Make Prediction on new Input/Dataset

new Input →

X1	X2	X3	Output
3	4	2	?



Apply Decision Tree

Output = 0

Machine Learning models

Most Common models

- Decision tree model
- Random forest
- SVM (Support Vector Machine)
- Linear Model
- Neural Network

Machine Learning models

S.No	Model Name	Model Type	Method	Package	Tuning Parameter(s)
1	ada	Classification	ada	ada	maxdepth, iter, nu
2	avNNet	Dual Use	avNNet	caret	decay, size, bag
3	bag	Dual Use	bag	caret	vars
4	bdk	Dual Use	bdk	kohonen	xweight,topo,xdim,ydim
5	blackboost	Dual Use	blackboost	mboost	maxdepth, mstop
6	Boruta	Dual Use	Boruta	Boruta	mtry
7	bstTree	Dual Use	bstTree	bst	maxdepth, nu, mstop
8	C5.0	Classification	C5.0	C50	winnow, trials, model
9	cforest	Dual Use	cforest	party	mtry
10	ctree	Dual Use	ctree	party	mincriterion
11	cubist	Regression	cubist	Cubist	committees, neighbors
12	enet	Regression	enet	elasticnet	lambda, fraction
13	foba	Regression	foba	foba	lambda, k
14	GAMens	Classification	GAMens	GAMens	fusion, iter, rsm_size
15	gamLoess	Dual Use	gamLoess	gam	degree,span
16	gbm	Dual Use	gbm	gbm	trees, shrinkage,depth
17	gcvEarth	Dual Use	gcvEarth	earth	degree
18	glm	Dual Use	glm	stats	None
19	icr	Regression	icr	caret	n.comp
20	J48	Classification	J48	RWeka	C

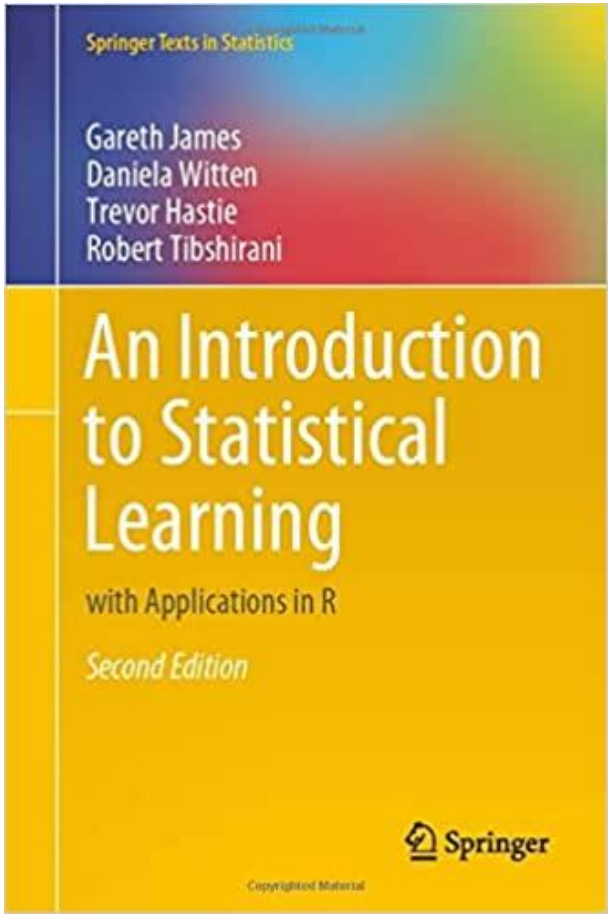
Machine Learning models

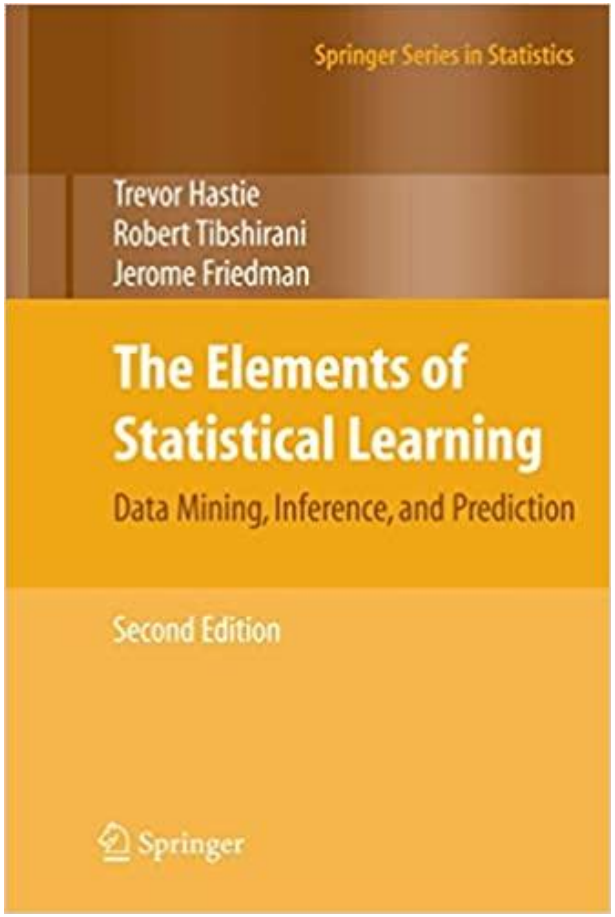
Already 500+ ML models are available for prediction

- Go for New Model Building
- Applied Modelling (Use existing models and solve real-world problem)

Books

Best Book for ML



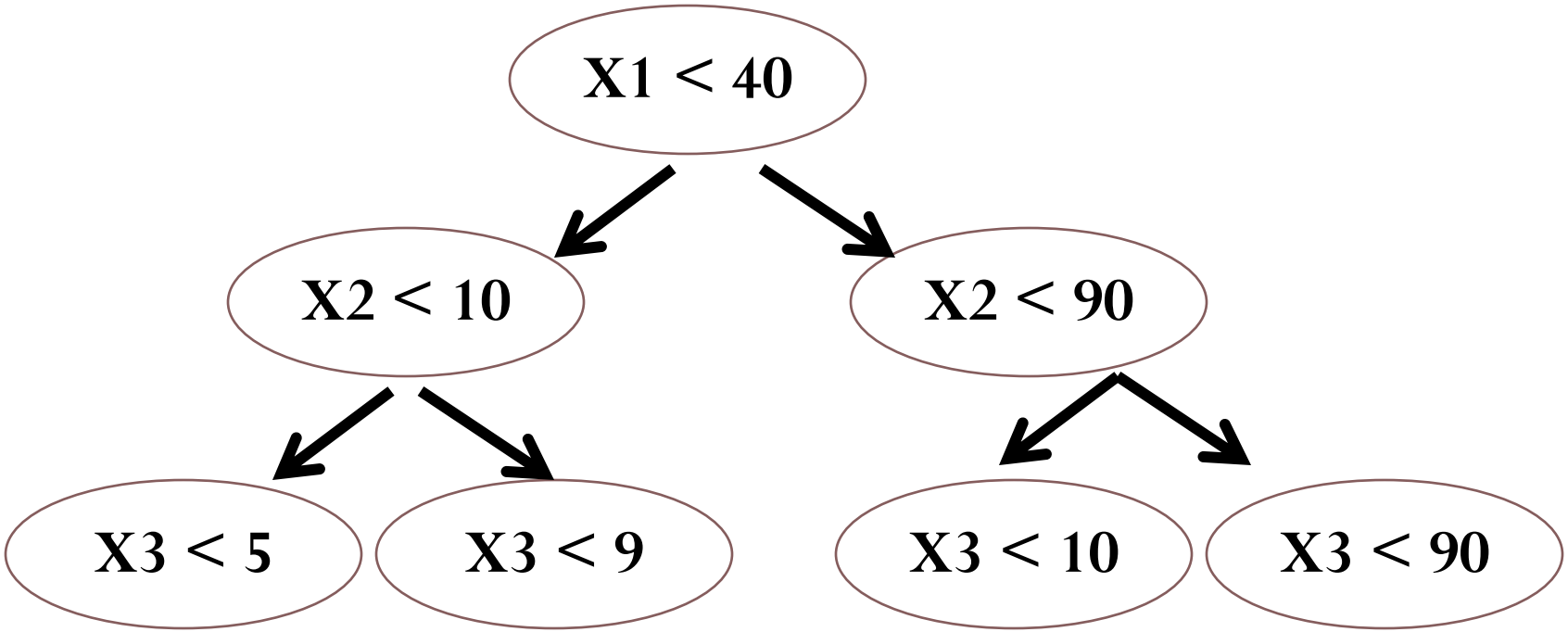


Types of Models

- **Linear Model**

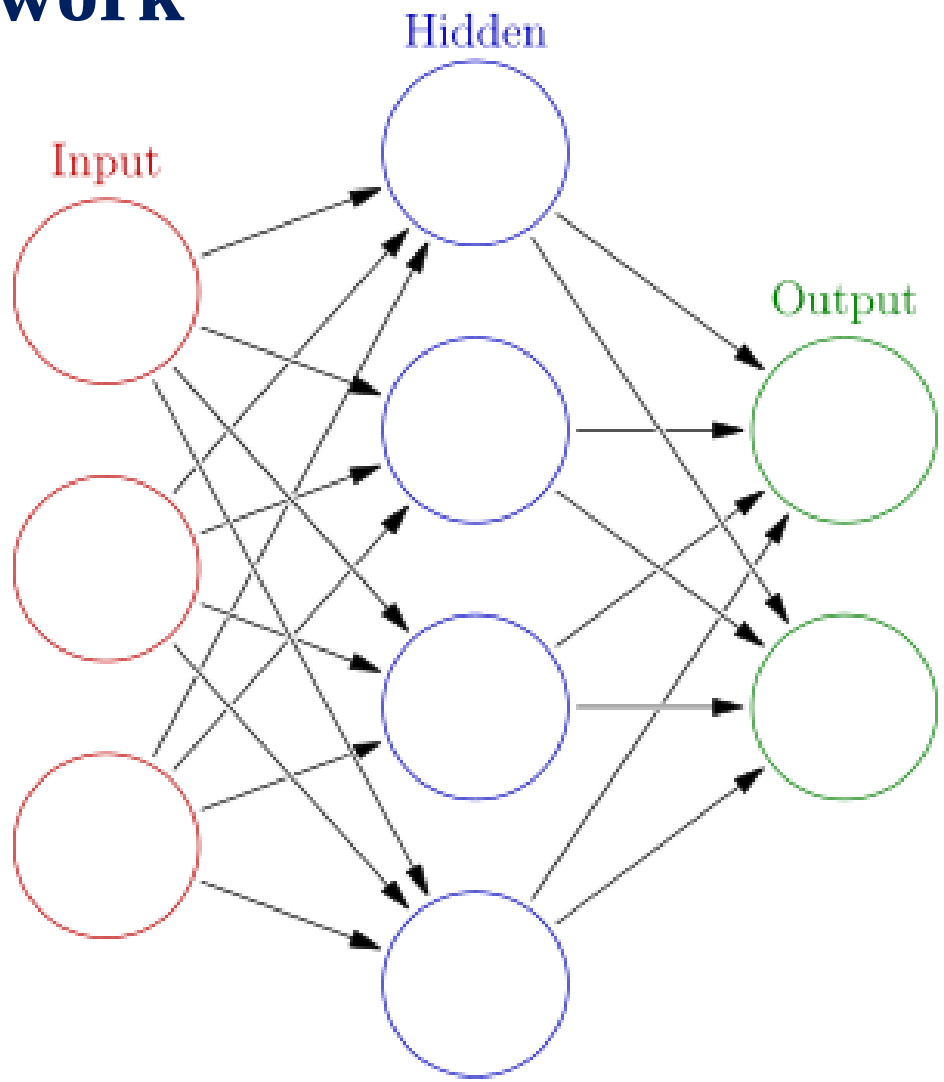
$$Y = w1\,x1 + w2x2 + w3x3 + w4x4 + w5x5$$

- **Tree Based model**



Types of Models.....

- **Neural Network**



Resources?

1. **www.psrana.com → Resource**
2. **leadingindia.ai → Resources**

Image Colorization

More Examples…..

More Examples…..

Can you think any APPLICATION that can be developed using Image Colorization??

Convert Black & White movies to Colored.

Methodology/Approach ??

Approach

1. Take any Colored Movie

↓

2. Extract Frames

↓

3. Convert Frames to B&W

↓

4. Train GAN on Colored and B&W images

GAN → Generative Adversarial Networks

(invented by Ian Goodfellow in 2014 in **Beer Bar**)

Approach…….

5. Take any new B&W Movie

↓

6. Extract Frames

↓

7. Apply GAN on B&W Frames

↓

8. Merge all the frames into a single movie

Outcome

Product (Web Service)

Example: Online Doc to PDF Converter

Web Interface

Product (Web Service)

B&W to Colour

Web Interface

How to create Web Applications??

- **Django**
- **Flask**

Hottest Topics ???

Hottest Topics Research / Projects

1. Deep Learning
2. Machine Learning
3. Transfer Learning
4. Reinforcement Learning
5. IoT
6. Big Data Analytics
7. Virtual Reality
8. Block Chain
9. NLP (Natural Language Processing)
10. Image Processing or Computer Vision
11. ……many many more

How to get the ideas for Project/Research

(A Million Dollar Question)

For Project Idea: Explore

It is not a job of single day.

For Project Idea: Explore

It is not a job of single day.

1. **Explore**
   - **Hackathons, Funding agency site, Project calls, etc**
2. **Study**
   - **Case study, Funding stories (YourStory.com)**
3. **Discuss**
   - **With Friend, Faculty, Teacher, Mentor, Students**
4. **Build Team and Implement**
5. **Present**

For Project Idea: Note it down

For any new idea just prepare 1-2 page writeup

1. Title
2. Methodology
3. Expected Outcome
4. Requirement (equipment)

For Project Idea

1. Leading India AI

www.LeadingIndia.ai

For Project Idea

1. Leading India AI

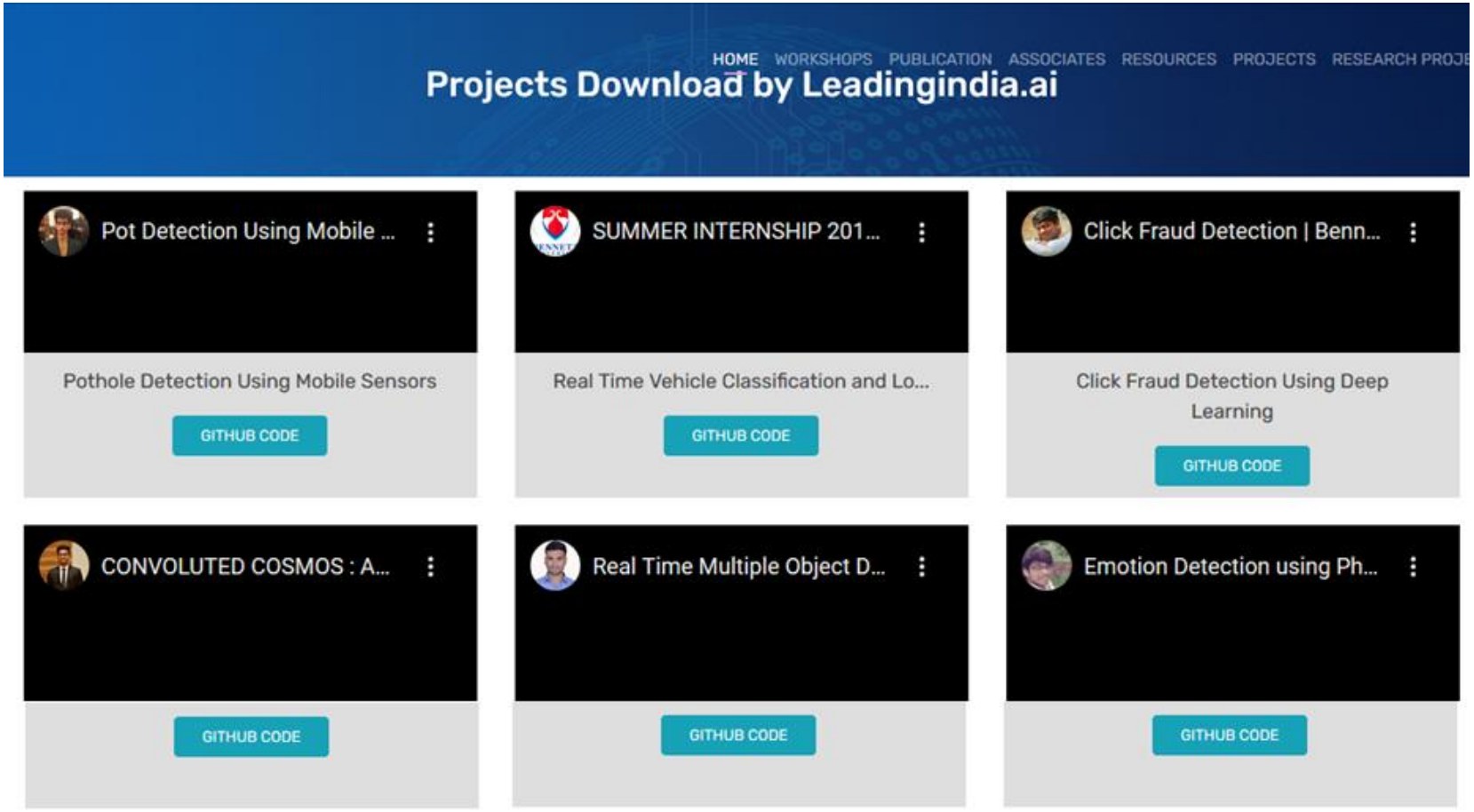
LEADINGINDIA.AI
NATIONWIDE AI SKILLING & RESEARCH INITIATIVE

- www.LeadingIndia.ai (AI/ML based problems)
- Lead by Prof. Deepak Garg
- Ready made AI/ML based problem statement.

Explore: Projects, Resources, Research Projects

Project Idea: LeadingIndia.ai

Go to → leadingindia.ai → Projects



Project Idea: LeadingIndia.ai

Go to → leadingindia.ai → Research Projects

Project Idea: LeadingIndia.ai

Go to → leadingindia.ai → Resources

- Deep Learning Online Certifications ⊞
- Machine Learning and Data Science Online Courses ⊞
- Machine Learning and Deep Learning Courses ⊞
- Free Online Books ⊞
- Videos and Lectures ⊞
- Github Tutorial Repositories ⊞
- Github Code Repositories ⊞
- Tutorials ⊞

For Project Idea: SIH

2. Smart India Hackathon (SIH)

www.sih.gov.in

For Project Idea: SIH

Winner in Smart India Hackathon (SIH) from Thapar

- **2017 → 01 Team (3rd position)**
- **2018 → 01 Team (2nd position)**
- **2019 → 05 Teams (1st Position (04) and 2nd (01))**
- **2020 → 02 Teams (1st Position (02))**

For Project Idea: SIH

2. Smart India Hackathon (SIH)

- www.sih.gov.in (Best resource for project idea)
- Explore problem statements of current and previous years. 353 problems in 2020 from different agencies and ministries.

For Project Idea

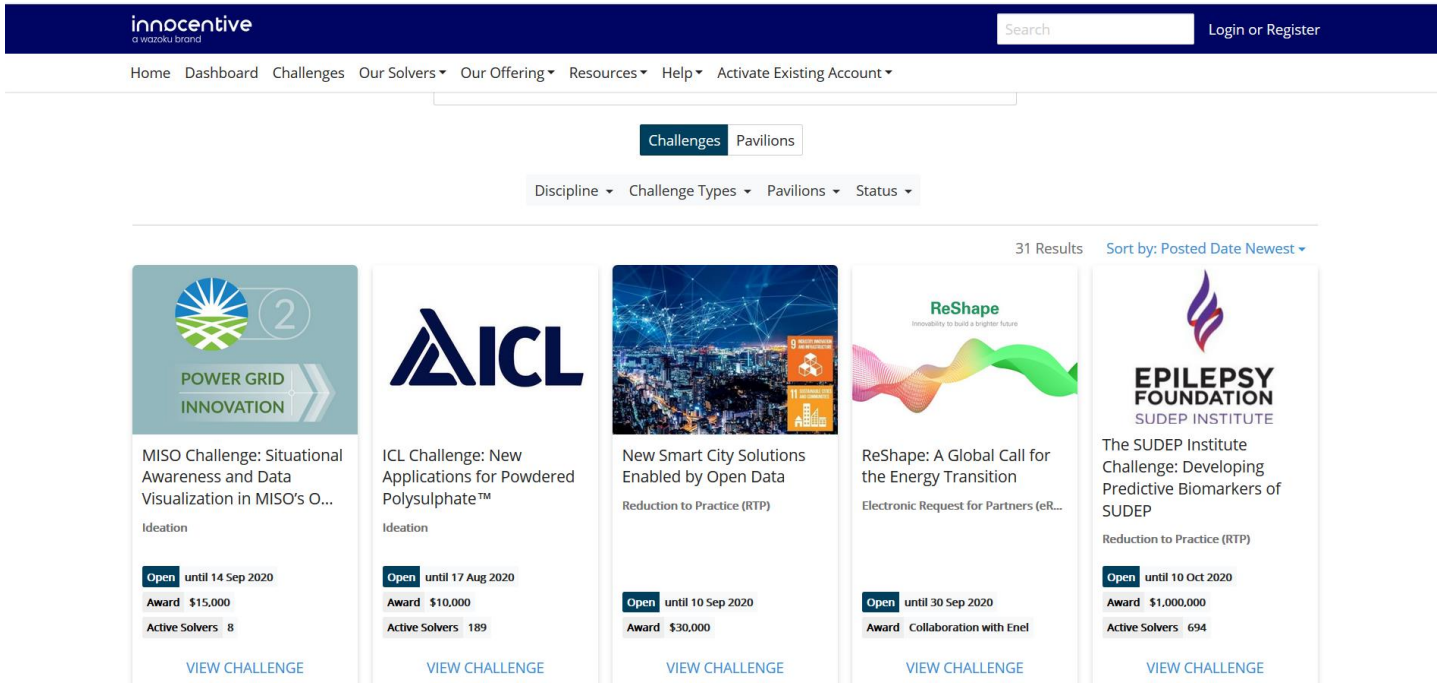
3. Innocentive

www.innocentive.com

For Project Idea

3. Innocentive

- www.innocentive.com (Real life problems)
- Go to Challenge Center → Challenges
- Ready made Problem statement, Literature, Expected outcome and price money.



For Project Idea

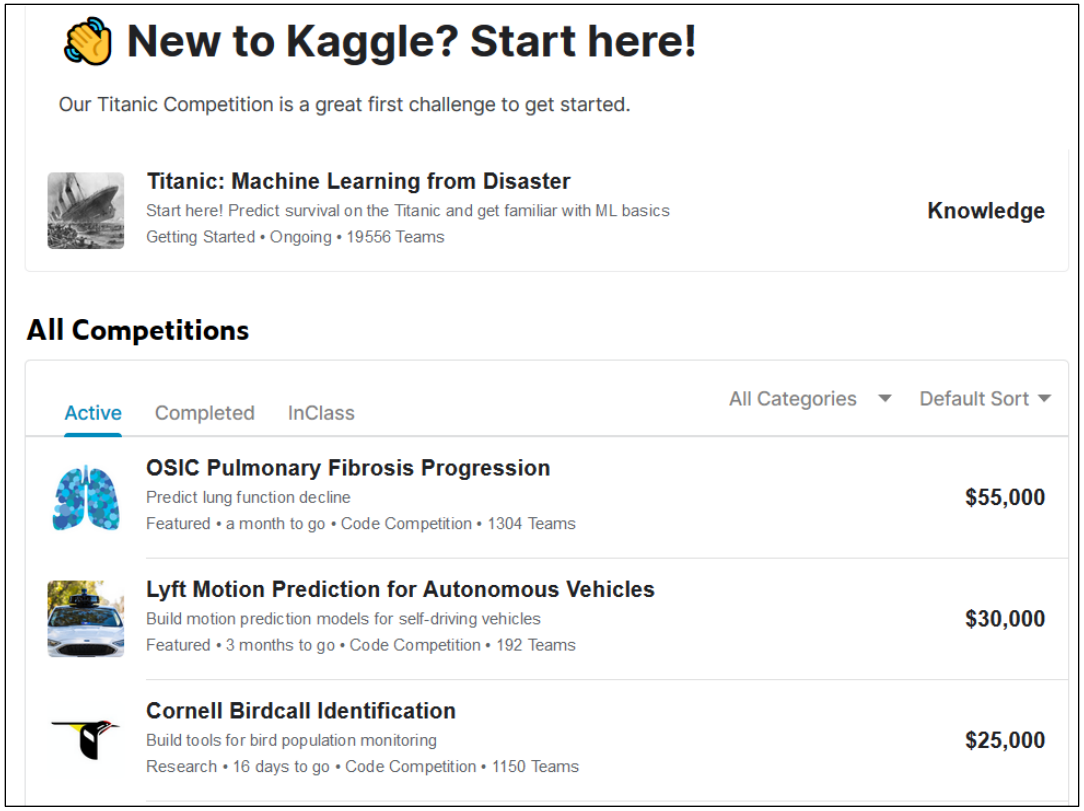
4. Kaggle

www.kaggle.com

For Project Idea

4. Kaggle

- www.kaggle.com (Data Science problems) → Compete
- Ready made Problem statement, Literature, Expected outcome and price money.



What is Structured, Semi-Structured and Unstructured Data??

Unstructured Data

- **No format is defined**
- **Example:**
  - **1000 movies in a folder**
  - **10k docs in a folder**
  - **500 images in a folder**
  - **…….. many more**

Structured Data

- **Stored in well defined format**
- **Example: Song**

Song ID	Language	Genre	Singers	Likes	Dislikes	.	.
1	En	Rock	2	10k	1k	.	.
2	En	Jazz	2	11k	1.5k	.	.
3	Hi	Pop	3	20k	2k	.	.
4	Hi	Jazz	1	15k	1.2k	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

Semi-structured Data

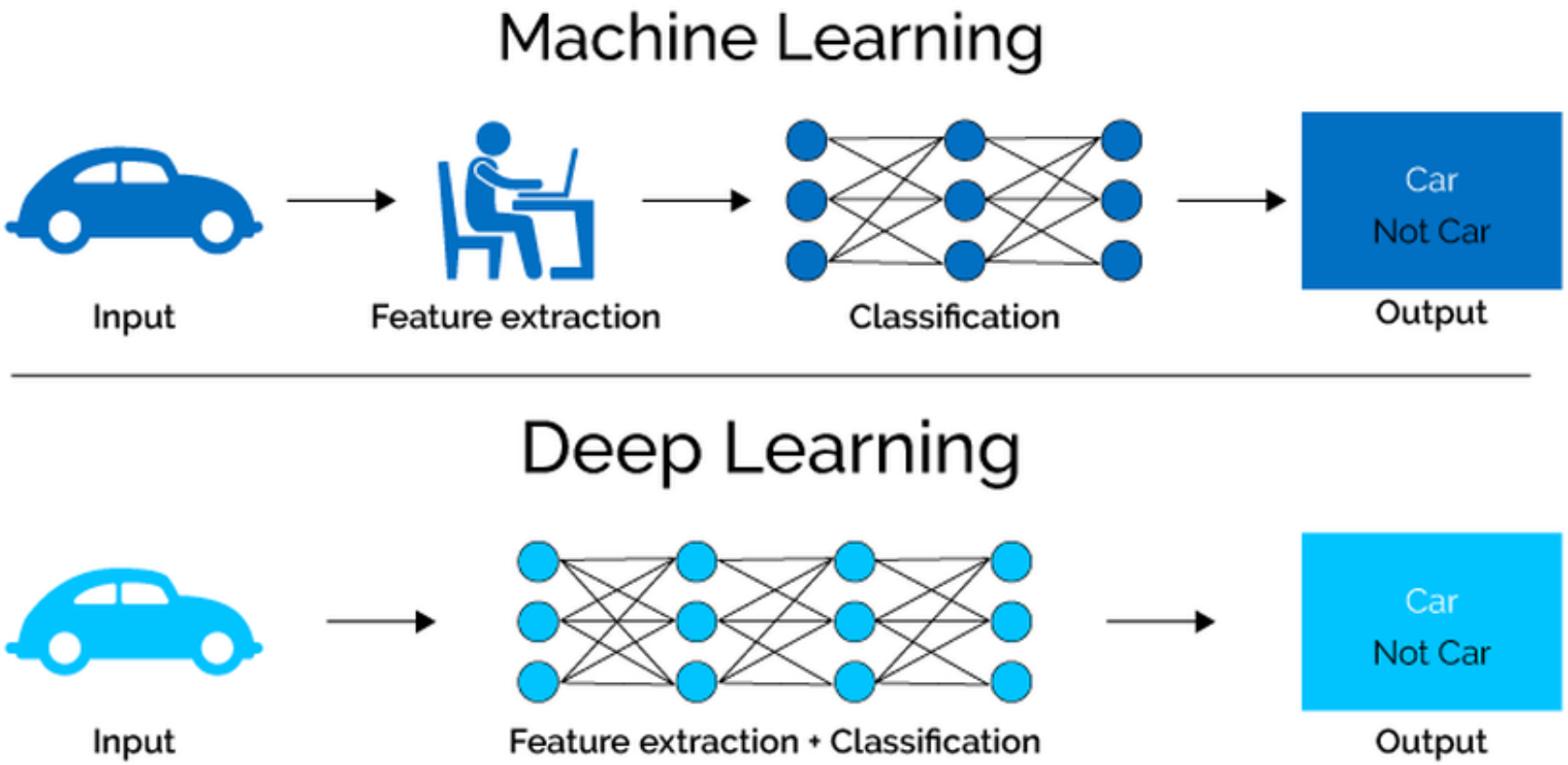
- Little format is defined
- Example:
  - 1000 movies in three folder (12+, 16+, 18+)
  - 10k docs in 4 folder (A, B, C, D)
  - 500 photos in two folder (Male, Female)
  - 10k tweets in two files (-ve, +ve)
  - ……. many more

Data Set format for Machine Learning

- **Multiple columns and one column is labelled (Strength)**

x1	x2	x3	x4	x5	Strength
17	0	-5	0.784245	37	26
12	0	-10	0.587296	25	27
18	0	-7	0.876622	40	25
11	0	-7	0.80826	24	23
18	0	-4	0.83215	37	28
10	1	-9	0.62842	27	28
19	0	7	0.522811	44	30
19	-1	4	0.548609	37	23
15	0	-6	0.177904	46	20

Difference between Machine Learning & Deep Learning ??


Machine Learning
Input
Feature extraction
Classification
Car
Not Car
Output
Deep Learning
Input
Feature extraction + Classification
Car
Not Car
Output

Additional Resources

- **Explore → psrana.com → Resources**

UNIT VI: Gold Mine for Researchers (Free Books, Papers, Thesis)

1. For Research Papers

- **Sci Hub**
  Search using **DOI** (Digital Object Identifier).
  Example (Search for): **10.1016/j.bbapap.2014.07.010**
- **BookSC**
  Search using title
  Example (Search for): **"Quality assessment of modeled protein structure using physicochemical properties"**

2. For Books

- **Library Genesis**
- **ZLibrary**

3. For Thesis | ProQuest

4. Paper with Code | Click Here | Most Important

5. Two Minutes Papers | Click Here

6. Explore

- Google Dataset Search | **Click Here**
- Explore Kaggle (Compete, Datasets, Notebooks, Jobs, more) | **Click Here**
- UCI dataset for Regression, Classification, Clustering, etc | **Click Here**

Where to get papers?

Where to get papers?

Sci Hub

www.sci-hub.tw

*** Domain name changes time to time.**

Where to get Books?

Where to get Books?

Library Genesis

http://gen.lib.rus.ec

*** Domain name changes time to time.**

Best Resource

Papers with Code

www.paperswithcode.com

- **Find recent paper in Machine Learning / Deep learning / Computer Vision / Natural Language Processing / etc**
- **Github code is available with every paper.**

Explore Hackathons

Explore Hackathons

Helps in selecting research topics.

www.kaggle.com
www.chalearn.org
www.mlwave.com
www.tunedit.org
www.codalab.org
www.gesture.chalearn.org
www.innocentive.com
www.dreamchallenges.org
www.crowdanalytix.com
www.datahack.analyticsvidhya.com
www.numer.ai
www.genomeinterpretation.org
http://grandchallenges.org/

Join Mailing Groups?

Join Mailing Groups?

Helps in selecting research topics, recent news and updates

http://bit.ly/MachineLearningBlogAndResource
http://feedburner.google.com/fb/a/mailverify?uri=analyticsvidhya
http://www.innocentive.com/blog
https://www.crowdanalytix.com/blog
http://www.kdnuggets.com/news/subscribe.html
http://www.r-bloggers.com/blogs-list/

Learn Fast

Learn Fast

1. To learn FAST Learn from slides.

2. Learn from Youtube / Videos.

Code Sharing

Code Sharing

1. **Share your code on GitHub**
   https://github.com

Present your work

Present your work

1. **Given Weekly / Monthly Presentation to you friends, boss, students, kids or to your wife.**

2. **Record it and improve it.**

Tools and Techniques

- **Python libraries**
  - **Numpy & Scipy** - Fundamental Scientific Computing
  - **Pandas** - Data Manipulation and Analysis
  - **Matplotlib** - Plotting and Visualization
  - **Scikit-learn** - Machine Learning and Data Mining
  - **StatsModels** - Statistical Modelling, Testing, and Analysis
  - **Seaborn** - For Statistical Data Visualization
  - **Plotly** - For sophisticated graphics
  - **Pydot** - For complex oriented and non-oriented graphs

* Install and explore the above libraries.

Tools and Techniques

- Tool: Anaconda
  - IDLE for Python

- Google Colab
  - Online Jupyter notebook for Python
  - Widely used platform for ML & DL

Data Set

- **Google Dataset Search**
  - **datasetsearch.research.google.com**
- **Kaggle**
  - **www.kaggle.com**
  - **Explore (Compete, Datasets, Notebooks, Jobs, more)**
- **UCI dataset**
  - **Google it “UCI Dataset”**
  - **Dataset available for Regression, Classification, Clustering**

Finally

Learning by Doing

Thanks

Q & A

Learning by Doing

Contact

www.psrana.com | psrana@gmail.com