



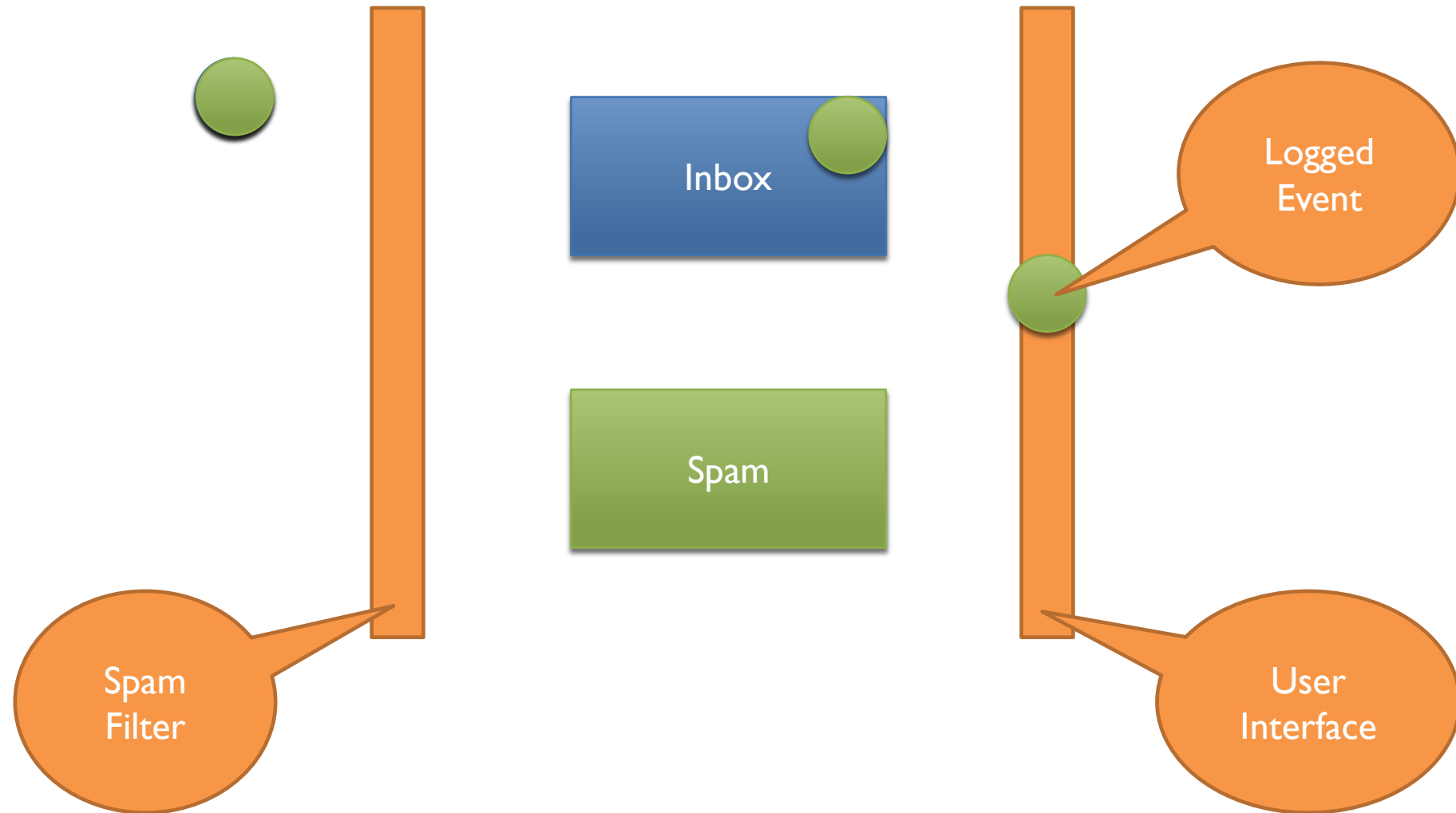
Practical machine learning



SESSION I: INTRODUCTION

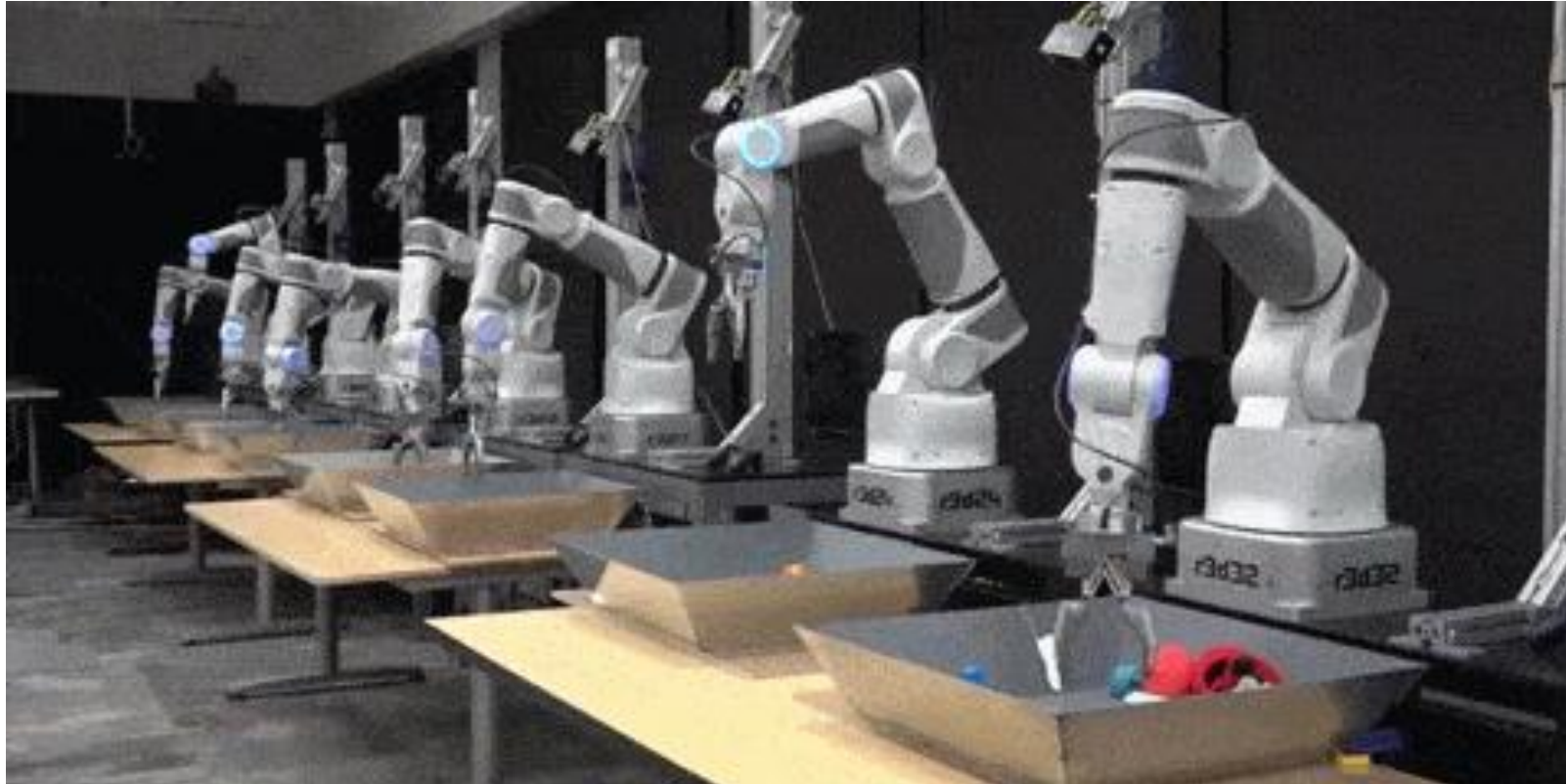


WHAT IS MACHINE LEARNING?



"MACHINE LEARNING"

Google Has a Room Full of Robot Arms Learning Hand-Eye Coordination



The **machine learning experiment** seeks to teach bots how to interact with their environment.

"MACHINE LEARNING"

Google's Self Driving Car



"MACHINE LEARNING"

Self Driving Car in Singapore



"MACHINE LEARNING..."

Are these Apples?



MACHINE LEARNING IN DEFENCE

US Army considers machine learning for unmanned ground vehicles (UGVs)

- Enable ground robots to learn from their surroundings and their mistakes
- Enhance the ability of unmanned ground vehicles (UGVs) to operate independently without human intervention



Big Data in Battle Field



- Provide intelligence preparation of the battle space, target development, and early warning of emerging threats
- Obtain security insights from data that is not intuitively security-related
- Analyzing insider threats

The **machine learning & Big Data** has the potential areas of applications in DoD....

"MACHINE LEARNING...PREDICTIVE ANALYTICS"

A global aerospace manufacturer uses analytics and modeling to predict and avoid program delays and cost overruns.

Challenge:

predict and prevent program risk, particularly project delays and cost overruns, and to better understand the factors that lead to risk.

Solution:

Uses **predictive analytics** on structured and unstructured project data to identify and predict where and when key causal factors may trigger program risk.



- **10x better** at predicting slippages of more than 100 days
- **50% increase in ability** to identify and predict overall schedule risk
- **\$10 Million saved** by avoiding missing a delivery deadline by even one month

"MACHINE LEARNING FOR MILITARY INTELLIGENCE"

How analytics is driving military intelligence ?

- predict the routes that are probably going to have roadside bombs planted on them
- who are the individuals in certain areas that we have to worry about.
- who is expressing the most anti-American sentiment.
- Where are the **soft targets** (the market places and other places where somebody may decide pull an event that would do harm to people).



Definition:

Recommender Systems are software tools that **elicit the interests and preferences** of individual consumers and **make recommendations accordingly**.

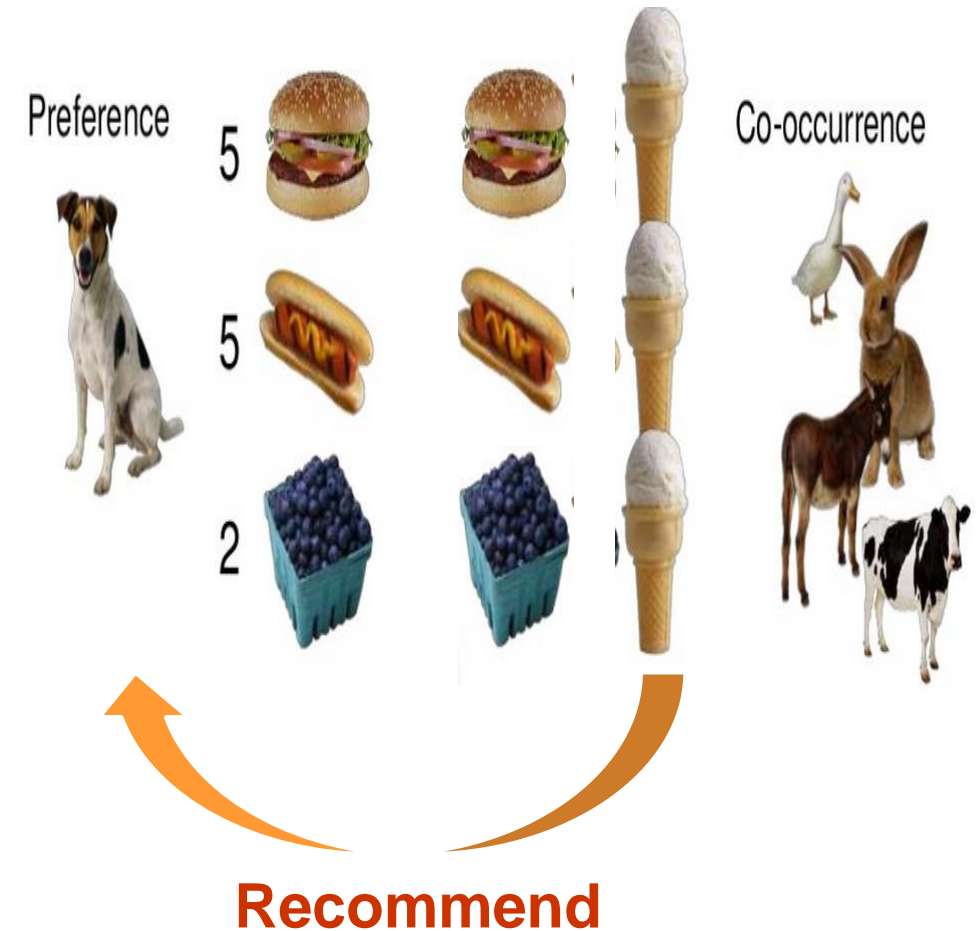
Types

- User based recommenders
- Item based recommender
- Matrix factorization based recommenders

RECOMMENDERS – USER BASED

(a) User based recommenders: (Collaborative filtering)

- **Predicting what users will like based on their similarity to other users.**
- **Assumption:** Users like the similar kinds of items they like in the past.
- Doesn't care whether the items are books, ice creams or mobile phones.
- Require no knowledge of the properties of the items
- **Example:** Facebook recommends friends



RECOMMENDERS – ITEM BASED

(b) Item based recommenders: (Content based filtering)

- Based on a description of the item and a profile of the user's preference
- A user profile is built to indicate the type of item this user likes
- Requires the knowledge of the properties of the items.



MACHINE LEARNING?

“**Learning** is any process by which system improves performance from experience”

“Machine Learning is concerned with computer programs that automatically improve their performance through experience”

- **Herbert Simon**



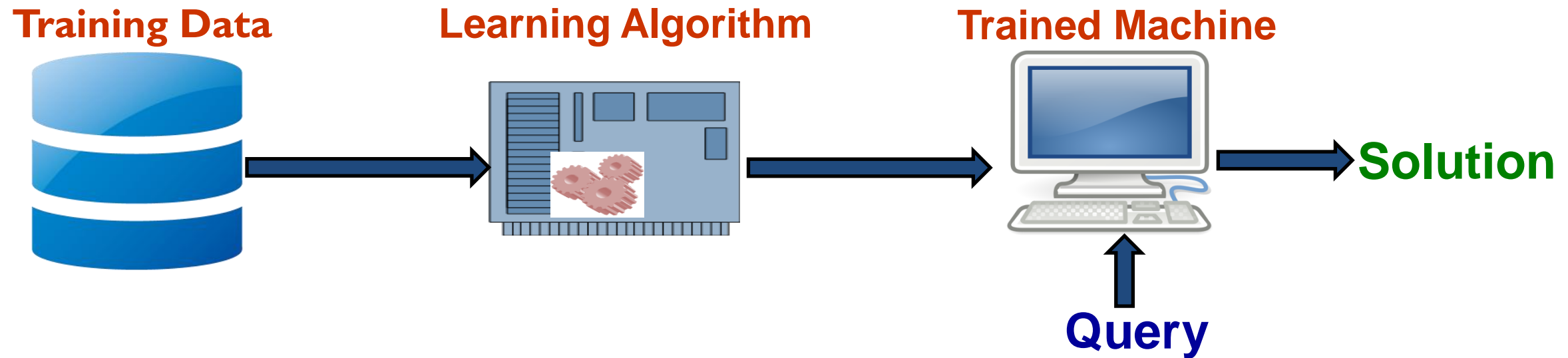
Herbert A. Simon

Turing Award - 1975

Nobel Prize in Economics - 1978

MACHINE LEARNING?

Machine Learning is the study of computer algorithms that **improve** automatically **through experience.**



Learning

For a class of Tasks ‘T’, Performance ‘P’ should improve with more Experience ‘E’.

More ‘E’ → More ‘P’ for a fixed ‘T’

EXAMPLE I – CLASSIFY A DISEASE

Preprocessed data of patient I	
Age	= 67
Sex	= 1
Chest pain type	= 4
Resting blood pressure	= 160
Serum cholesterol	= 286
Fasting blood sugar	= 0
...	

Classification

Presence = 1

EXAMPLE I – CLASSIFY A DISEASE

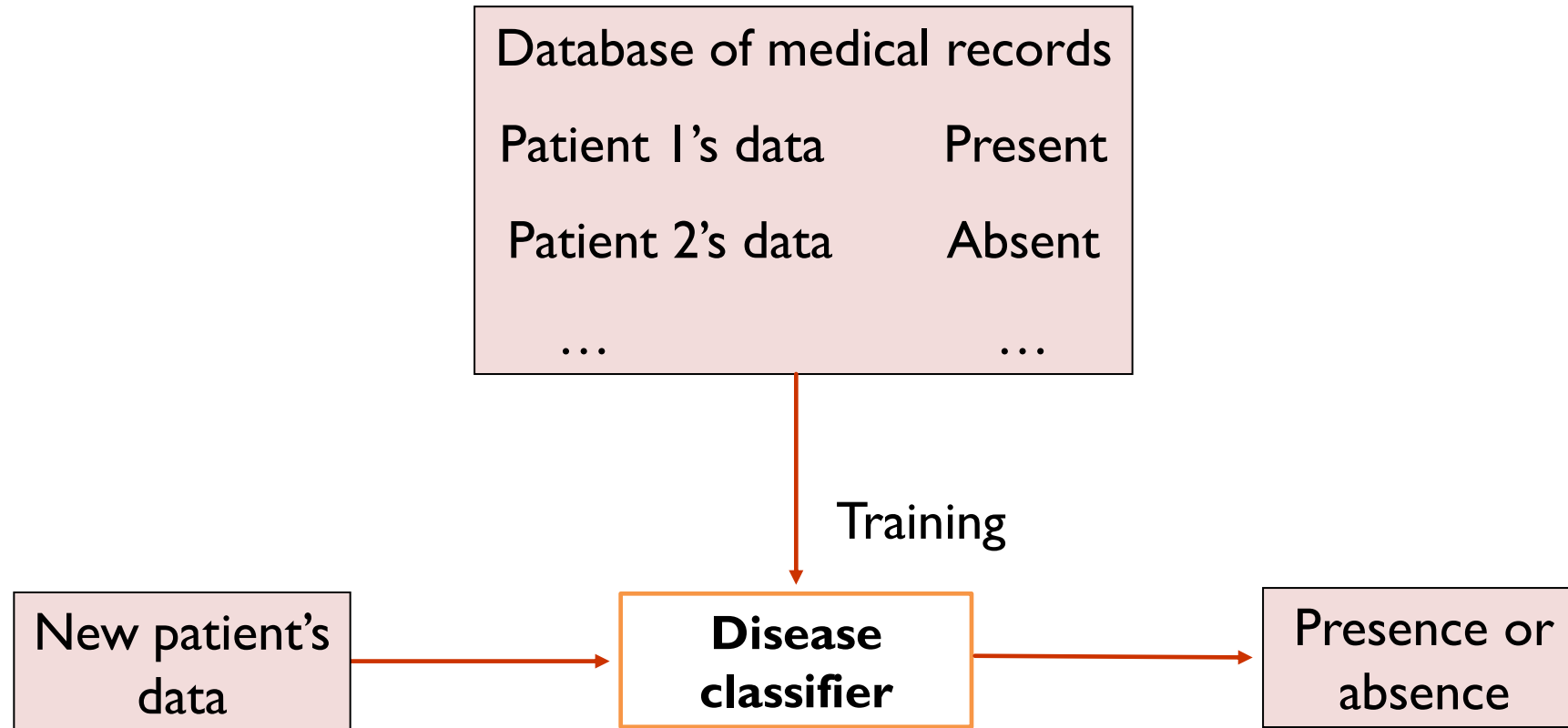
Preprocessed data of patient 2

Age	= 63
Sex	= 1
Chest pain type	= 1
Resting blood pressure	= 145
Serum cholesterol	= 233
Fasting blood sugar	= 1
...	
























Classification

Presence = 0

EXAMPLE I – CLASSIFY A DISEASE – MACHINE LEARNING



EXAMPLE 2 – MOVIE RECOMMENDER

	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	...
Ram						
Krish						
James						
Anu						
Akira						
...						

EXAMPLE 2 – MOVIE RECOMMENDER

	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	...
Ram	★★★★★	★★★★☆	★★★★☆	★★★★☆	★★★☆☆	
Krish	★★★★★	★★★★★	★★★★★	★★★★★	★★★★★	
James	★★★★☆	★★★★☆	★★★★☆	★★★★★	★★★★☆	
Anu	★★★★★	★★★★☆	★★★★☆	?	★★★☆☆	
Akira	?	★★★☆☆	★★★★☆	★★★★☆	★★★★☆	
...						

EXAMPLE 2 – MOVIE RECOMMENDER

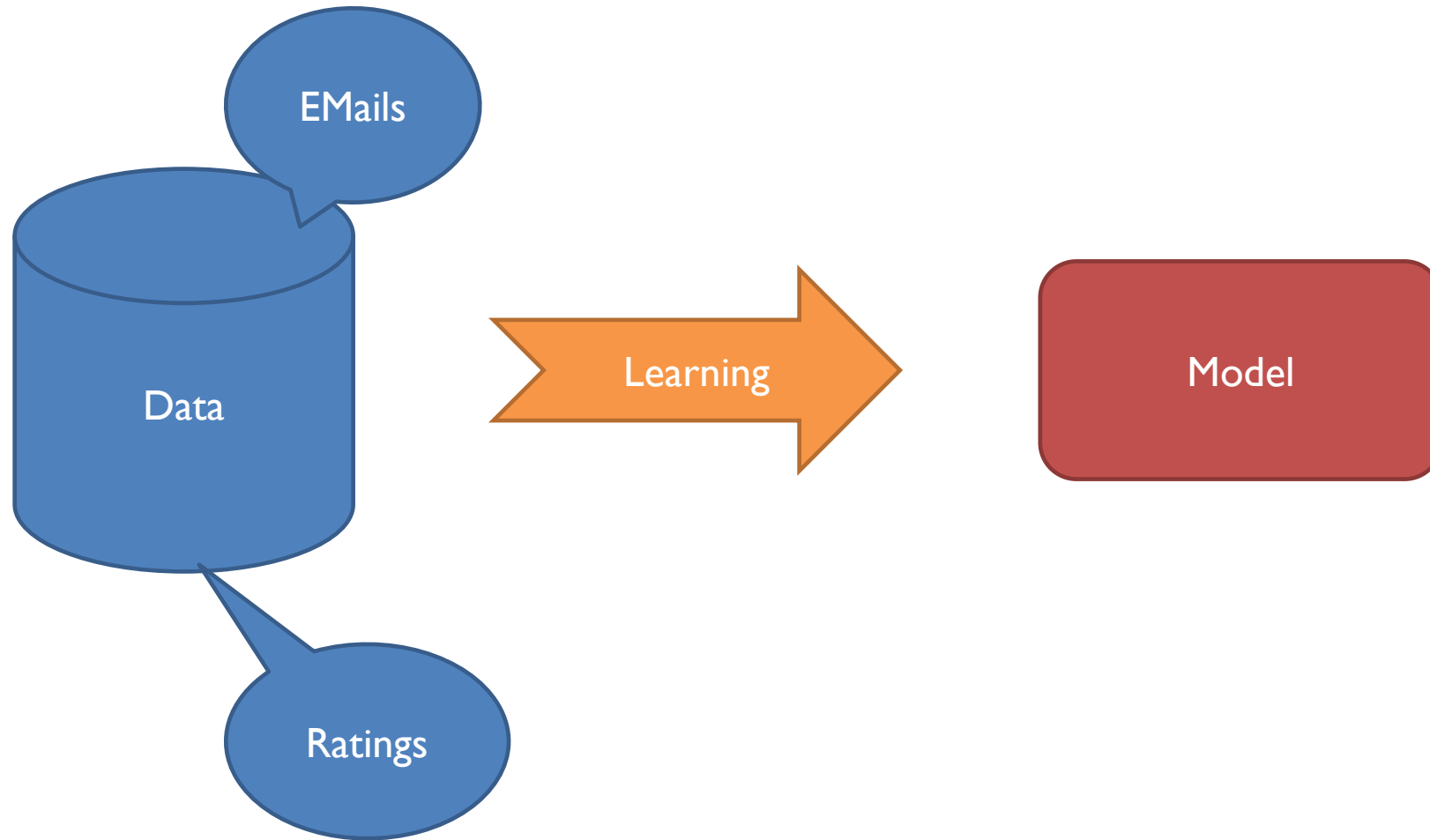
	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	...
Ram	★★★★★	★★★☆☆	★★★★☆	★★★★☆	★★★☆☆	
Krish	★★★★★	★★★★★	★★★★★	★★★★★	★★★★★	
James	★★★★☆	★★★★☆	★★★★☆	★★★★★	★★★★☆	
Anu	★★★★★	★★★☆☆	★★★★☆	?	★★★☆☆	
Akira		★★★☆☆	★★★★☆	★★★★☆	★★★★☆	
...						

EXAMPLE 2 – MOVIE RECOMMENDER

	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	...
Ram	★★★★★	★★★☆☆	★★★★☆	★★★★☆	★★★☆☆	
Krish	★★★★★	★★★★★	★★★★★	★★★★★	★★★★★	
James	★★★★☆	★★★★☆	★★★★☆	★★★★★	★★★★☆	
Anu	★★★★★	★★★☆☆	★★★★☆	★★★★☆	★★★☆☆	
Akira		★★★☆☆	★★★★☆	★★★★☆	★★★★☆	
...						

EXAMPLE 2 – MOVIE RECOMMENDER

	Movie 1	Movie 2	Movie 3	Movie 4	Movie 5	...
Ram	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★	
Krish	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★	
James	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★	
Anu	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★	
Akira	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★	
...						



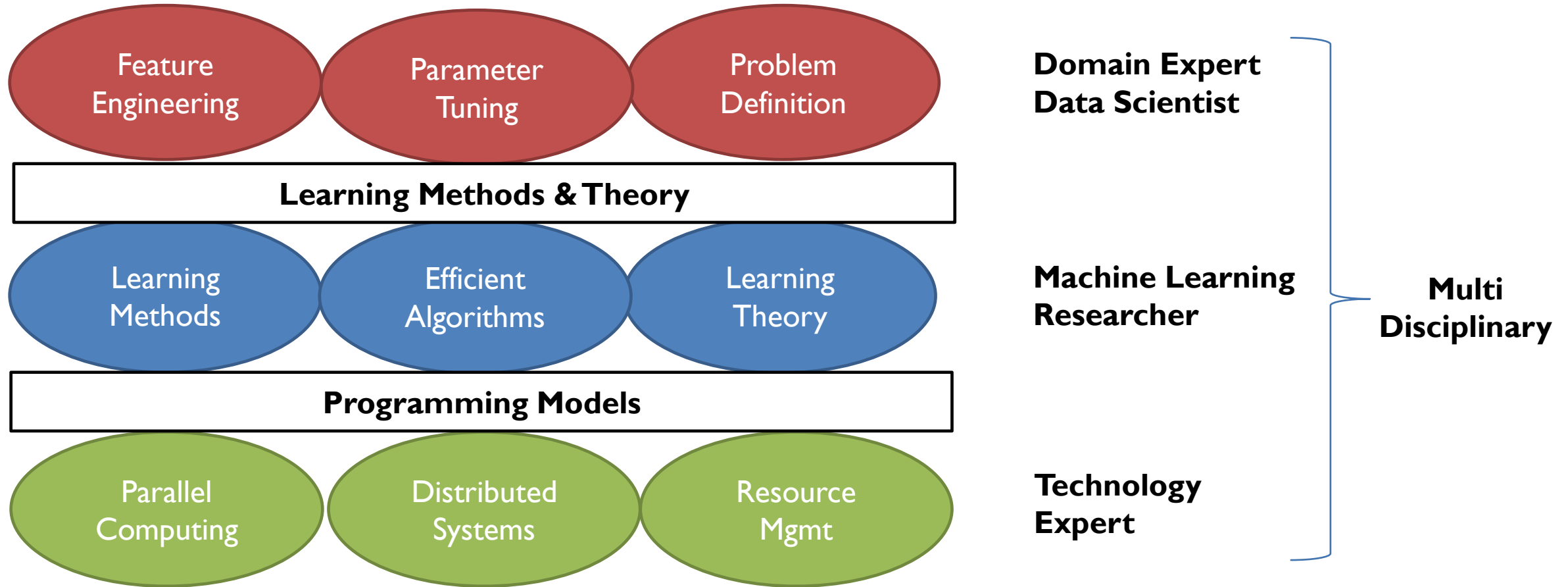
Supervised

- Classification
- Regression
- Recommender

Unsupervised

- Clustering
- Dimensionality reduction
- Topic modeling

A LAYERED VIEW



UNLABELED DATA

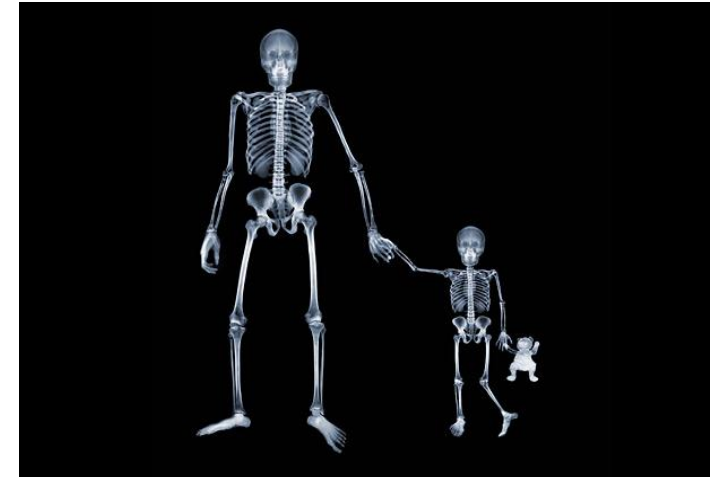
Unlabeled data:

Consists of natural or human-created artefacts that can obtain easily from the world.

Examples:

Photos, audio recordings, videos, news articles, tweets, x-rays etc.

No "explanation" for each piece of unlabelled data, just contains the data, nothing else.



LABELLED DATA

Labeled Data:

Take a set of unlabelled data and augments each piece of that unlabelled data with meaningful "tag," "label," or "class" that is somehow informative.

Example:

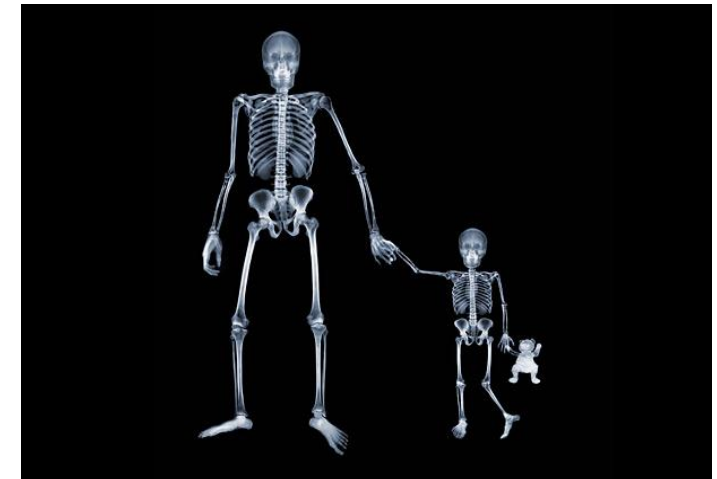
- (a) whether this **photo contains a horse or a cow**,
- (b) what the topic of this news article is,
- (c) Whose X-Ray is this, etc.



A Cat

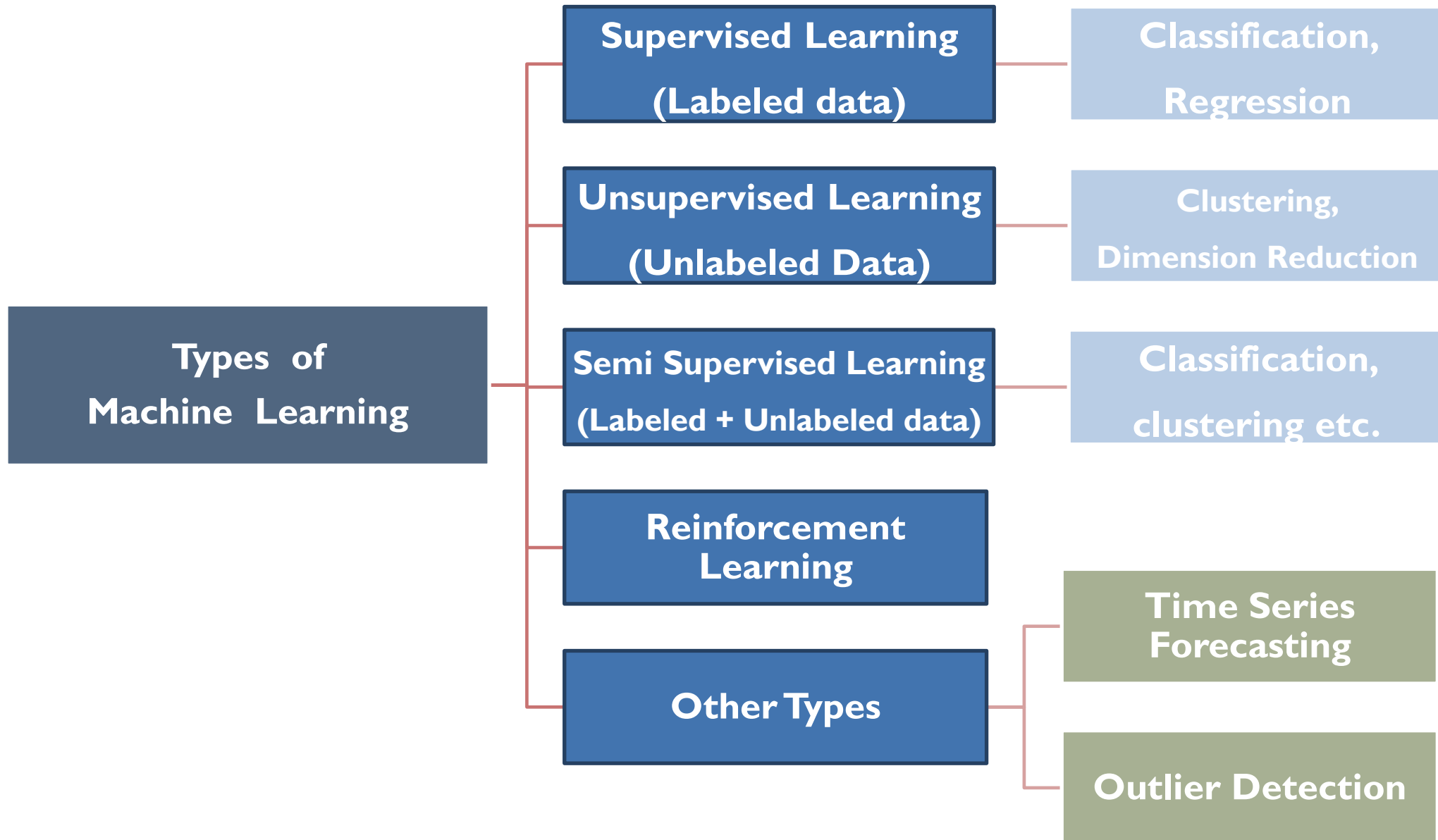


Exchange of old notes to coins



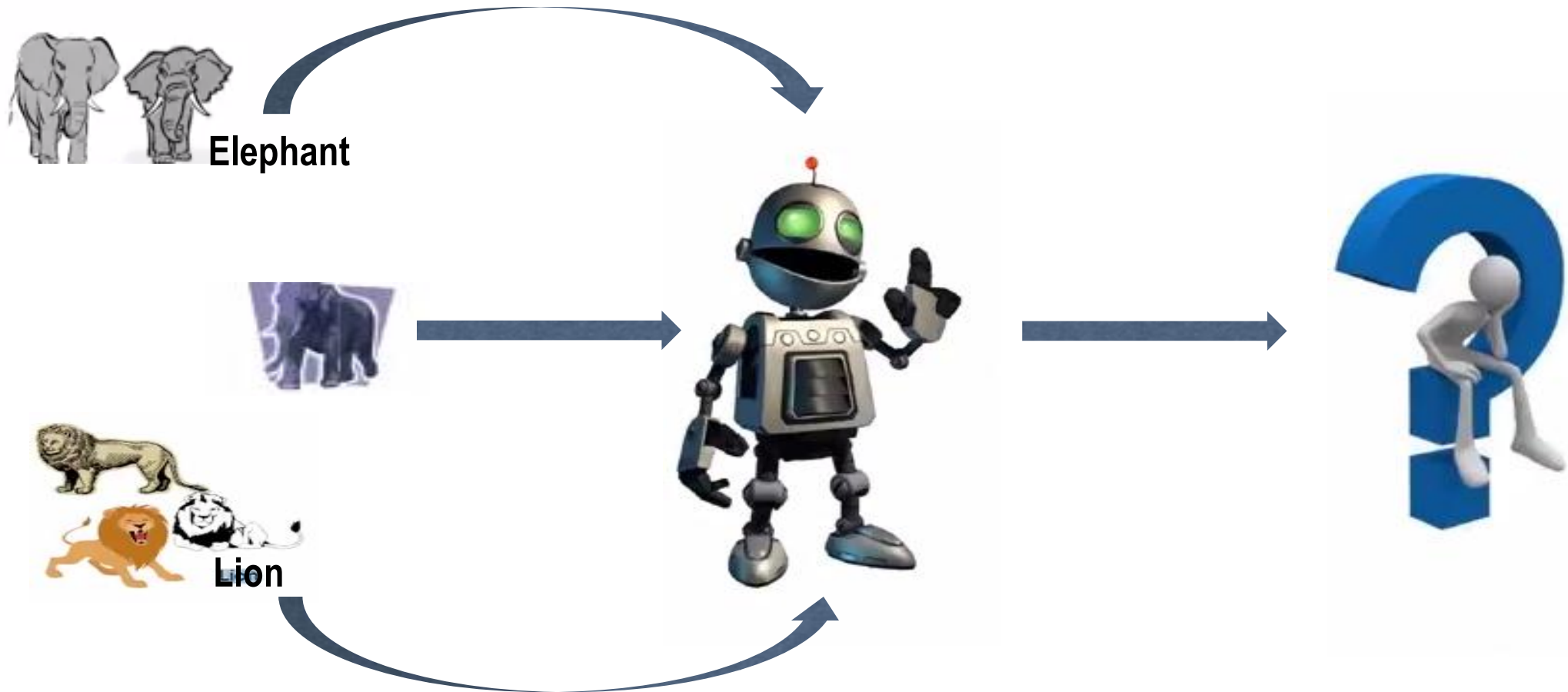
X-Ray of Mr. James & His Son

TYPES OF MACHINE LEARNING



SUPERVISED LEARNING

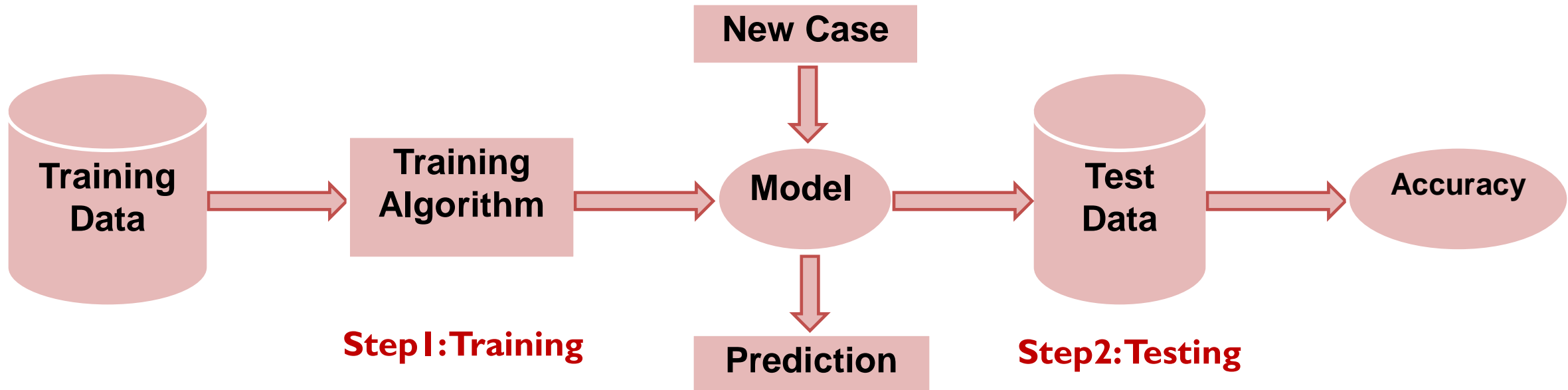
- Takes a known set of input data and known responses to the data
- Seeks to build a predictor model to the new data.
 - ✓ It will have **fully labeled** set of known data
 - ✓ In another form, Training data includes desired output



SUPERVISED LEARNING – TWO STEP PROCESS

- **Learning (Training):** Learn a model using the **training data**
- **Testing:** Test the model using **unseen test data** to assess the model accuracy

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}}$$



SUPERVISED LEARNING

Types: Classification and Regression

- I. **Classification** – Identifying to which of a set of categories (sub-populations) a new observation belongs.
 - Normally works on **discrete data**
 - **Data:** A set of data records (also called examples, instances or cases) described by
 - **k attributes:** $A_1, A_2, \dots A_k$.
 - **a class:** Each example is labelled with a pre-defined class.
 - **Goal:** To learn a **classification model** from the data that can be used to predict the classes of new (future, or test) cases/instances.
 - **Examples:**
 - Predict whether an E-mail is **SPAM or NOT?**
 - Whether to **approve Loan** to a Bank Customer **or Not?**

SUPERVISED LEARNING

2. Regression – Regression analysis helps to understand how the typical value of the **dependent variable** (or 'criterion variable') **changes** when any one of **the independent variables is varied**.

- Normally works on **continuous data**
- **Examples:**
 - **Stock market analysis** - How many units a consumer will purchase?
 - How **long** a patient “Mr.ABC” **will live?**

ALGORITHMS

- Decision tree induction
- Random Forest
- Classification using association rules
- Naïve Bayesian classification
- Support vector machines (SVM)
- K-nearest neighbor (KNN)

SUPERVISED LEARNING – EXAMPLE I – SPAM DETECTION

Task (T): Identify whether an e-mail is SPAM or non-SPAM → **Classification**

Training Data (E): Large collection of SPAM and non-SPAM messages → **Labelled Data**

- No. of Attributes: 57 (Frequency of words such as Money, Lottery etc.)
- Last attribute denotes whether the e-mail is spam (1) or not (0)
- Use this data to train and test the machine learning model.

0,0.64,0.64,0,0.32,0,0,0,0,0,0,0.64,0,0,0,0.32,0,1.29,1.93,0,0.96,0.778,0,0,
3.756,61,278, **1**

0.21,0.28,0.5,0,0.14,0.28,0.21,0.07,0,0.94,0.21,0.79,0.65,0.21,0.14,0.14,0.07,0.28,3.47,0,1.59,0,0.43,0.43,0,0,0,0,0,0,0,0,0,0,0,0,
.07,0,0,0,0,0,0,0,0,0,0,0,0.132,0,0.372,0.18,0.048,5.114,101,1028, **0**

0.06,0,0.71,0,1.23,0.19,0.19,0.12,0.64,0.25,0.38,0.45,0.12,0,1.75,0.06,0.06,1.03,1.36,0.32,0.51,0,1.16,0.06,0,0,0,0,0,0,0,0,0,0,
0,0,0,0.06,0,0,0.12,0,0.06,0.06,0,0,0.01,0.143,0,0.276,0.184,0.01,9.821,485,2259, **1**

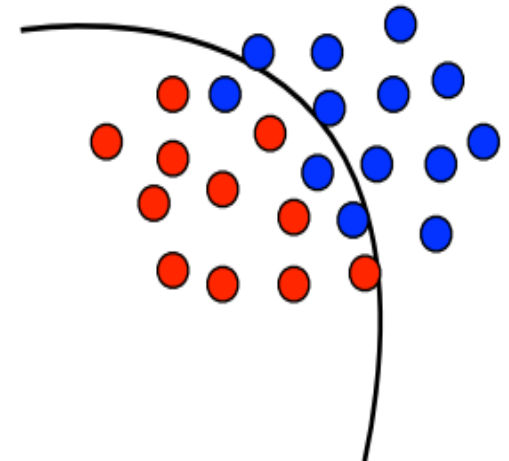
SUPERVISED LEARNING – EXAMPLE I – SPAM DETECTION

Learning Stages:

- ✓ **Divide** Labelled collection into **Train data (70%)** and **Test data (30%)**.
- ✓ Use training data and features to **Train ML Algorithm**.
- ✓ **Predict labels** in test data to Evaluate Algorithm.
- ✓ Algorithms may choose a Parameter:
e.g: Number of rounds.

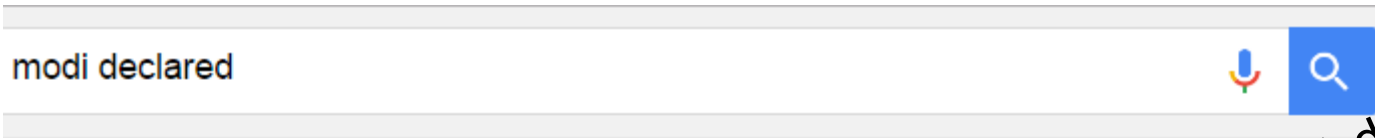
Performance (P):

- % of SPAM mails that were filtered correctly



CLUSTERING EXAMPLE – GROUPING OF NEWS ARTICLES

Clusters of News articles from Google



All **News** Videos Images Maps More ▾ Search tools

About 3,85,000 results (0.51 seconds)



In an attempt to curb black money, PM Narendra Modi declares Rs ...

Economic Times - 08-Nov-2016

In a move to curb the black money menace, PM Narendra Modi declared that from midnight currency notes of Rs 1000 and Rs 500 ...

Rs 500, Rs 1000 notes declared illegal from midnight: Narendra Modi

India.com - 08-Nov-2016

Rs 500, Rs 1000 notes declared illegal: Here is how politicians ...

The Indian Express - 08-Nov-2016

PM Modi's rupee revamp: Don't worry, your money stays your money

India Today - 08-Nov-2016

Massive googly bowled by PM Narendra Modi: Anil Kumble on ...

Financial Express - 08-Nov-2016

Rs 500, Rs 1000 notes currency ban: Check how Narendra Modi ...

In-Depth - Firstpost - 10-Nov-2016



India.com



The Indian Ex...



India Today



Financial Exp...



Northbridge T...



Moneycontrol...

News from India Today

The Economic Times



In an attempt to curb black money, PM Narendra Modi declares Rs 500, 1000 notes to be invalid

By ECONOMICTIMES.COM | Updated: Nov 09, 2016, 12:09 AM IST

Post a Comment

READ MORE ON » Rs 500 notes | Rs 1000 notes | PM Narendra Modi | Black Money

In a move to curb the black money menace, PM Narendra Modi declared that from midnight currency notes of Rs 1000 and Rs 500 denomination will not be legal tender. People can deposit notes of Rs 1000 and Rs 500 in their banks from November 10 till December 30, 2016.

In his 40-minute address, first in Hindi and later in English, the Prime Minister said the notes of Rs 500 and Rs 1000 "will not be legal tender from midnight tonight" and

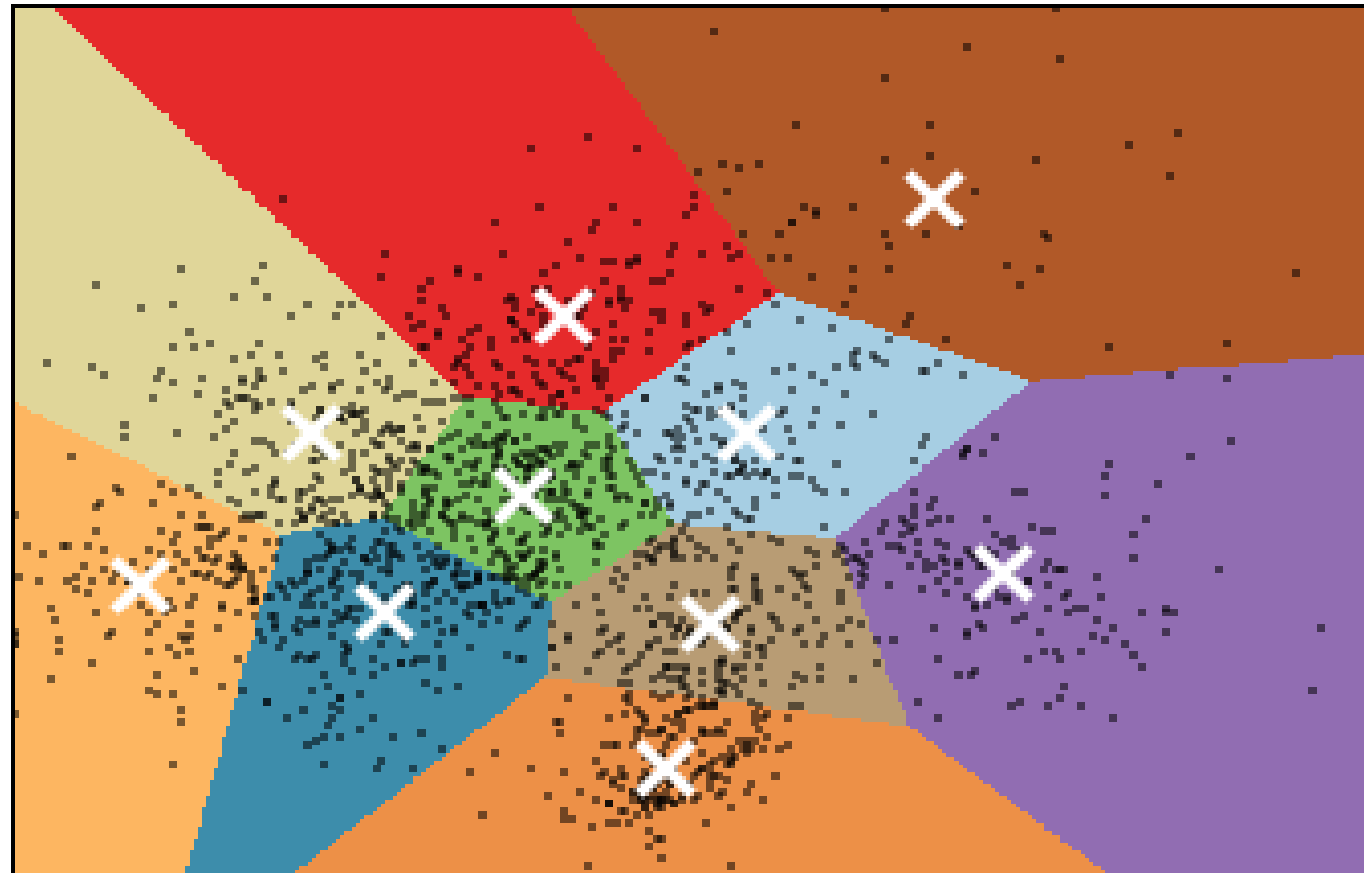
RELATED VIDEO



PM Modi declares Rs 500, 1000 notes to be void from midnight

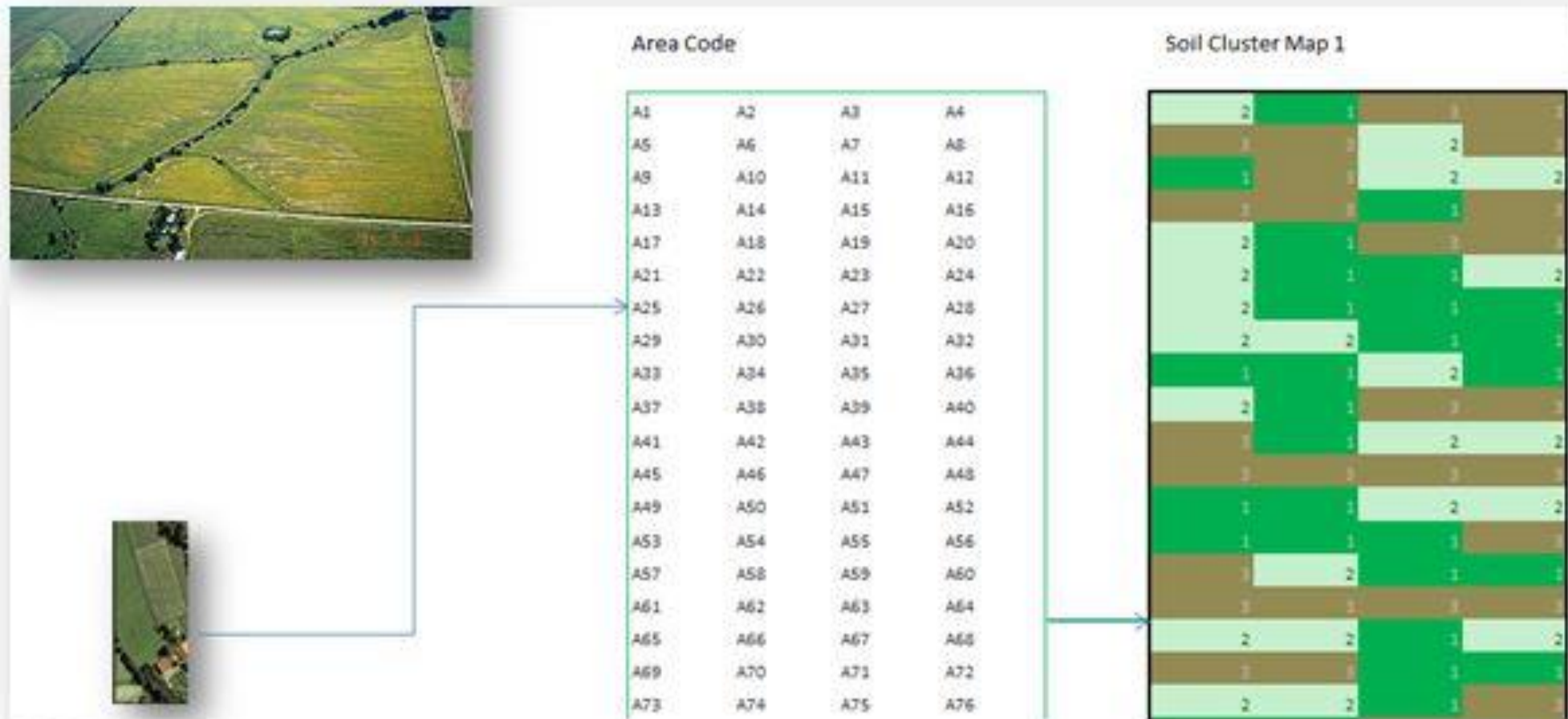
Digit Recognition

K-means clustering on the digits dataset (PCA-reduced data)
Centroids are marked with white cross



CLUSTERING – EXAMPLE2

Soil Map - Agriculture



SEMI-SUPERVISED LEARNING

Definition:

- Only a subset of the training data is labeled.
- Training Data = Labeled (very expensive) + Unlabeled (Cheap).
- Typically, **small amount of labelled data** + **a large amount of unlabelled data**.

Advantages:

- It is very useful when the acquisition of **fully labelled data is very expensive**
- Preparing **labeled data** requires **human expertise** and **so much time**
- For example, to analyze and label **1 hr speech requires 4000 hours** for an expert
- So, **combine the strengths of Supervised learning + Unsupervised learning**

Examples:

- Text processing
- Handwriting recognition

REINFORCEMENT LEARNING

Definition:

An agent (e.g., a robot or controller) seeks to learn the optimal actions based on the outcomes of the past actions. → **Learning By Doing**

Note:

- The agent, initially, only knows the set of possible states & the set of possible actions.
- The learner is not told which actions to take, but instead **must discover which actions yield the most reward** by trying them.
- An agent can act in a world and, **after each step**, it can observe the state of the world and observe **what reward or punishment** it obtained.

REINFORCEMENT LEARNING

- Learning By Doing

- **Examples:**

Robot: A robot that can act in a world, receiving rewards and punishments and determining from these what it should do.

Chess: A master chess player makes a move using reinforcement learning.

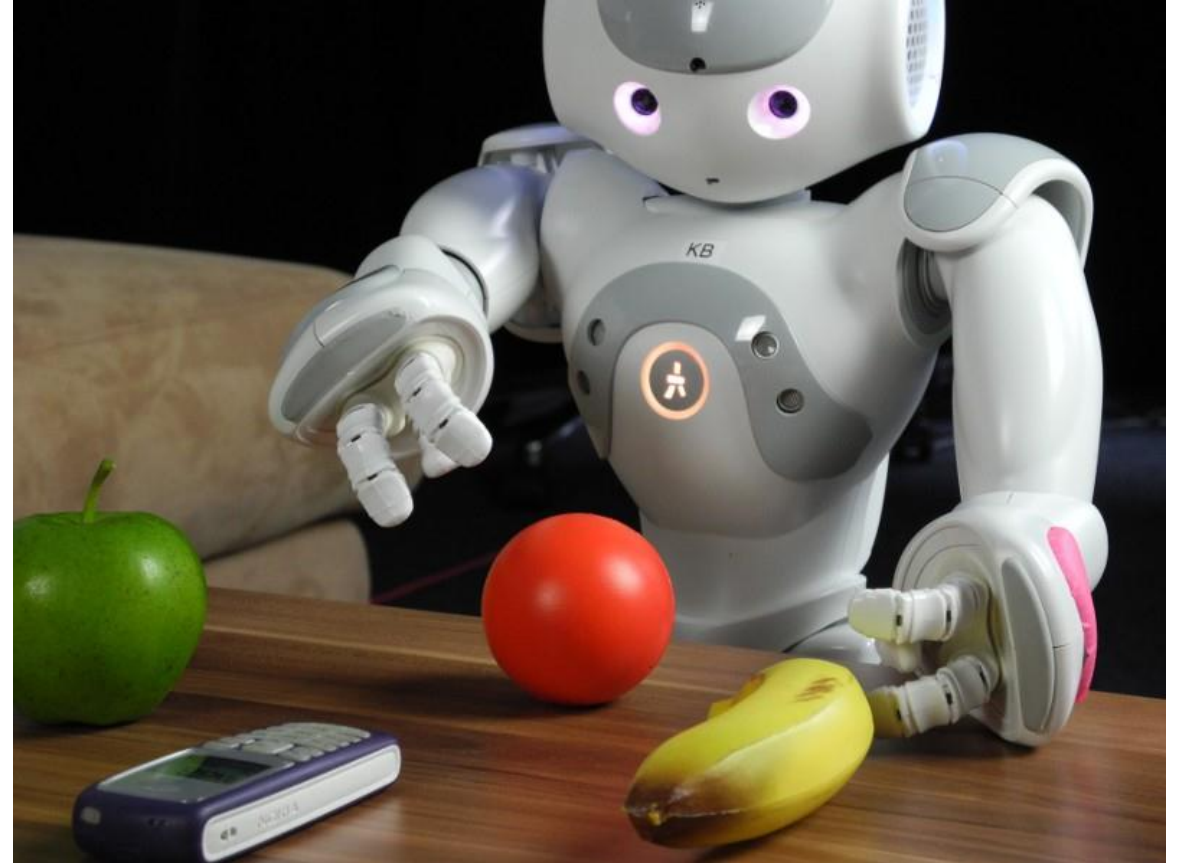


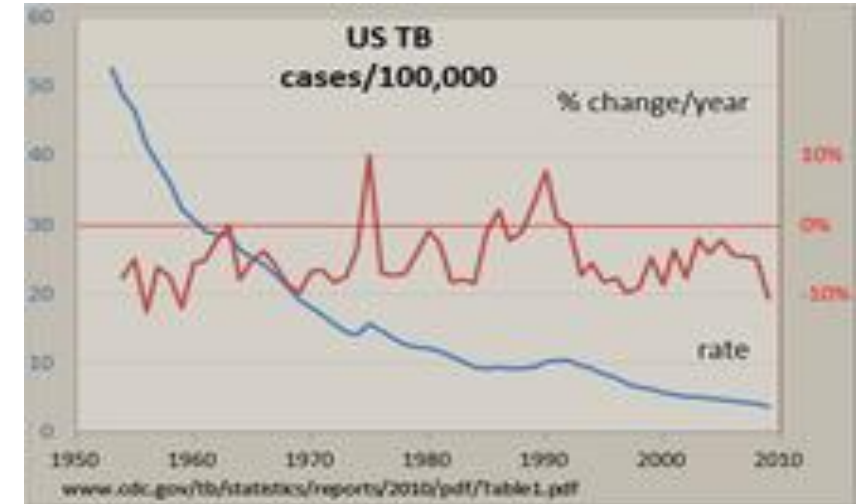
Image Source: <http://www2.informatik.uni-hamburg.de>

TIME SERIES FORECASTING & ANOMALY DETECTION

Time Series Forecasting: Model to **predict the future values** based on the previously observed values (**past values**).

Applications:

- ✓ Meteorology
- ✓ Financial markets



Tuberculosis incidence US 1953-2009

Anomaly detection: Identification of items, events or observations which do not conform to an expected pattern or other items in a dataset.

- Also called as **outlier detection**

Example: Used for **fault-detection** in factories

USES OF MACHINE LEARNING

- Spam detection
- Voice Recognition
- Stock Trading
- Robotics
- Medicine and Healthcare
- Advertising
- Retail and E - Commerce
- Gaming
- IoT

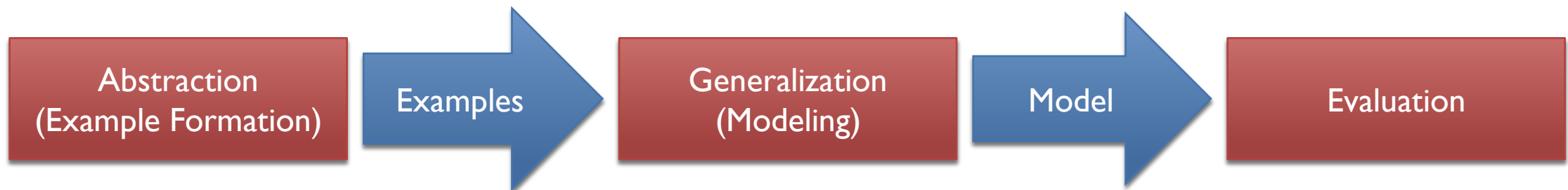
LANGUAGES FOR MACHINE LEARNING

- Python
- R
- Scala
- Clojure
- Ruby
- Java

THE LEARNING PROCESS - I

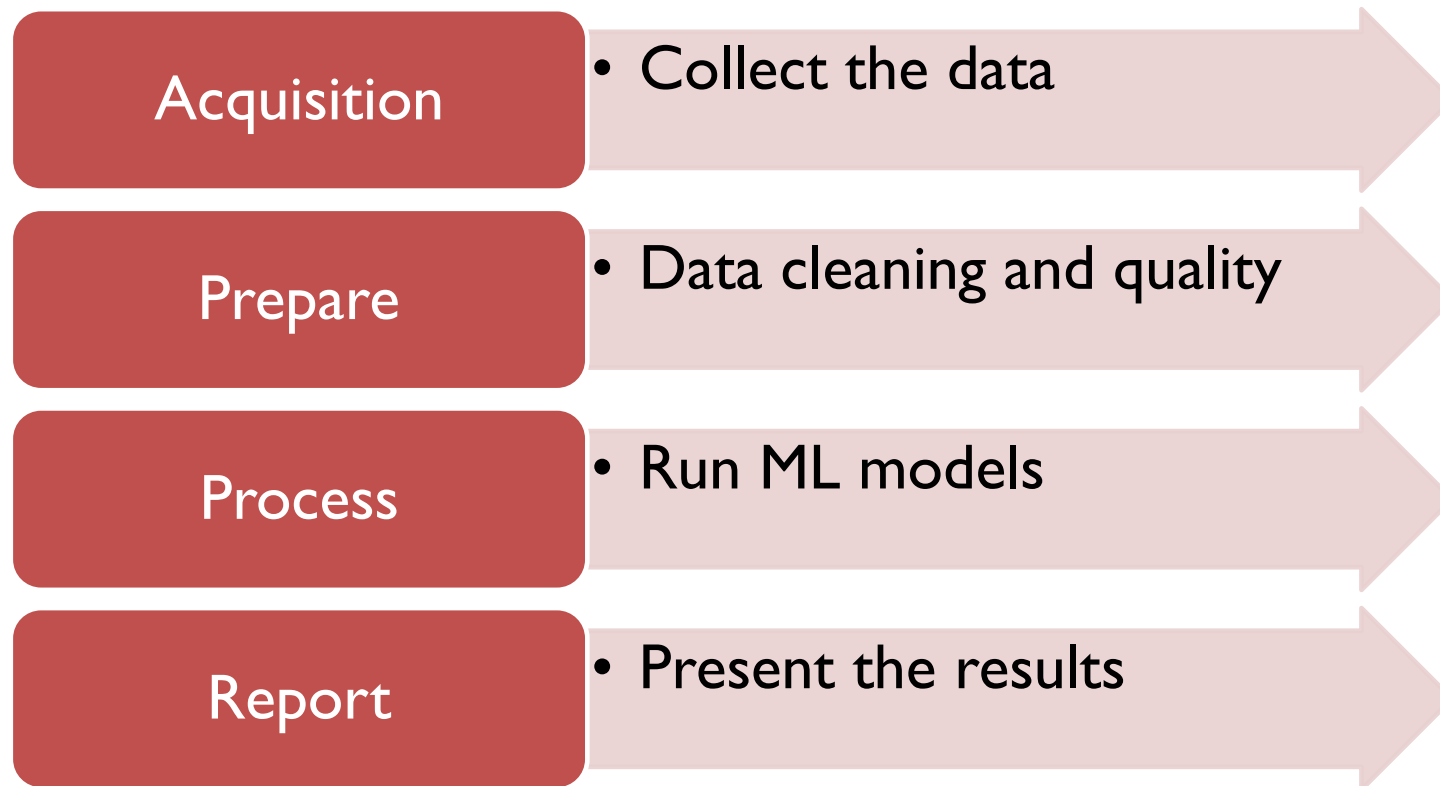
Regardless of whether the learner is a human or machine, the basic learning process is similar. It can be divided into four interrelated components:

- **Step 1: Data Storage**
 - provide a factual basis for further reasoning
- **Step 2: Abstraction**
 - Feature and Label Extraction
- **Step 3: Generalization**
 - Create knowledge and inferences that drive action in new contexts
- **Step 4: Evaluation**
 - Provide a feedback, measure the utility of the learned knowledge, and improve it.



ML PLANNING - THE MACHINE LEARNING CYCLE

Regardless of whether the learner is a human or machine, the basic learning process is similar. It can be divided into the following components:



ML PLANNING - (A) DEFINING THE PROCESSES

Planning might take into account:

- Where the data is coming from?
- How to clean?
- What learning methods to use
- what the output is going to look like.

The main point is that these things can be changed at any time—the earlier in the process they change, the better. So it's worth taking the time to sit around a table with and the team and figure out what you are trying achieve.

This process might involve algorithm development or code development.

- The more iterations you perform on the code the better it will be.
- Agile development processes work best; in agile development, you only work on what needs to be done without trying to future-proof the software as you go along.
- It's worth using some form of code repository site like Github or BitBucket to keep all your work private;
- It also means you can roll back to earlier versions if you're not happy with the way things are going.

ML PLANNING - (C) TESTING

Testing means testing with data.

- You might use a random sample of the data or the full set.
- The important thing is to remind yourself that you're testing the process, so it's okay for things to not go as planned.
- If you push things straight to production, then you won't really know what's going to happen.
- You might find:
 - data loading issues,
 - data-processing issues,
 - or answers that just don't make sense.
- When you test, you have time to change things.

Sit down with the stakeholders and discuss the test results.

- Do the results make sense?
- The developers and mathematicians might want to amend algorithms or the code.
- Stakeholders might have a new question to ask (this happens a lot),
- or perhaps you want to introduce some new data to get another angle on the answers.
- Regardless of the situation, make sure the original people from the planning phase are back around the table again.

When everyone is happy with the way the process is going it's time to refine code and, if possible, the algorithms.

- With huge volumes of data, squeeze every ounce of performance you can from your code and the quicker the overall processing time will be.
- Think of a bobsled run;
- A slower start converts to a much slower finish.

ML PLANNING - (F) PRODUCTION

When all is tested, reviewed, and refined by the team, moving to production shouldn't be a big job.

- Be sure to give consideration to when this project will be run—is it an hourly/daily/weekly/monthly job?
- Will the data change wildly between the project going in to production and the next run?
- Make sure the team reviews the first few production runs to ensure the results are as expected,
- and then look at the project as a whole and see if it's meeting the criteria of the stakeholders.
- Things might need to be refined.
- As you probably already know, software is rarely finished.

ML PLANNING - HOW TO BUILD A TEAM??

In a ML project, A *data scientist* is someone who can bring the facets of:

- Data processing,
- Simple Analytics,
- Mathematics and statistics,
- Programming,
- Visualization or Graphics Design

With so many skill sets in action, even for the smallest of projects, it's a lot to ask for one person to have all the necessary skills.

ML PLANNING - DATA QUALITY AND CLEANING

In the real world, data is messy, usually unclean, and error prone.

- Presence Checks
- Type Checks
- Length Checks
- Range Checks
- Format Checks
- Other Checks, if any.

EXAMPLE I - PRESENCE CHECKS

Check that data has been entered at all.

Example:

Registration – It usually involves at least an e-mail address, first name, and last name.

	FIRSTNAME	LASTNAME	E-MAIL	AGE
Correct	Jason	Bell	me@domain.com	42
Incorrect		Bell		42

EXAMPLE 2 - WHAT'S IN A COUNTRY NAME??

Consider the database of a hotel.

- Its data was gathered via a web-based enquiry form, but instead of offering a selection of countries from a drop-down list of countries, there was just an open text field.

Example:

- If you take a country like India. then you might have the following entries for country name:
 - Ireland
 - Republic of Ireland
 - Eire
 - EIR
 - Rep. of Ireland

What you have is a huge job to clean up the country field of a database.

EXAMPLE 2 - WHAT'S IN A COUNTRY NAME?? . . .

Solution:

- find all the distinct names in the country field and associate them with a two-letter country code.
- So, Ireland and all the other names that were associated with Ireland become IE.
- You would have to do this for all the countries.

In programming terms, you could make each of the distinct countries a key in a HashMap and add a method to get the value of the corresponding input name.

INPUT DATA FORMATS

Data comes in all sorts of forms.

Data Formats:

- Raw Text
- Comma Separated Values (CSV)
- JavaScript Object Notation (JSON)
- YAML Ain't Markup Language (YAML)
- Extensible Markup Language (XML)
- Spreadsheets
- Database/Videos
- Streams



MACHINE LEARNING APPLICATIONS





Case Study 1: Network Intrusion Detection

Network Intrusion Detection

Details about data:

- **Lincoln Labs** set up an environment to acquire **nine weeks of raw TCP dump** data for a local-area network (LAN) simulating a typical **U.S. Air Force LAN**.
- Each connection **record consists of about 100 bytes**.

Example Input feature vector of a network connection:

0,tcp,http,SF,233,2032,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,4,4,0.
00,0.00,0.00,0.00,1.00,0.00,0.00,15,15,1.00,0.00,0.07,0.00,
0.00,0.00,0.00,0.00, **normal**.

Network Intrusion Detection...

1. Problem & Objective:

- Finding out whether a network connection is **normal** or **intrusion**

2. Solution:

- Classification each network connection using its features with **Logistic Regression**

3. Features/Data Set

- KDD CUP 1999 Intrusion Detection Dataset
- Information about **4.9 Million Network connections**, each having **42 features**
- File Size: **~1GB** (CSV Format)

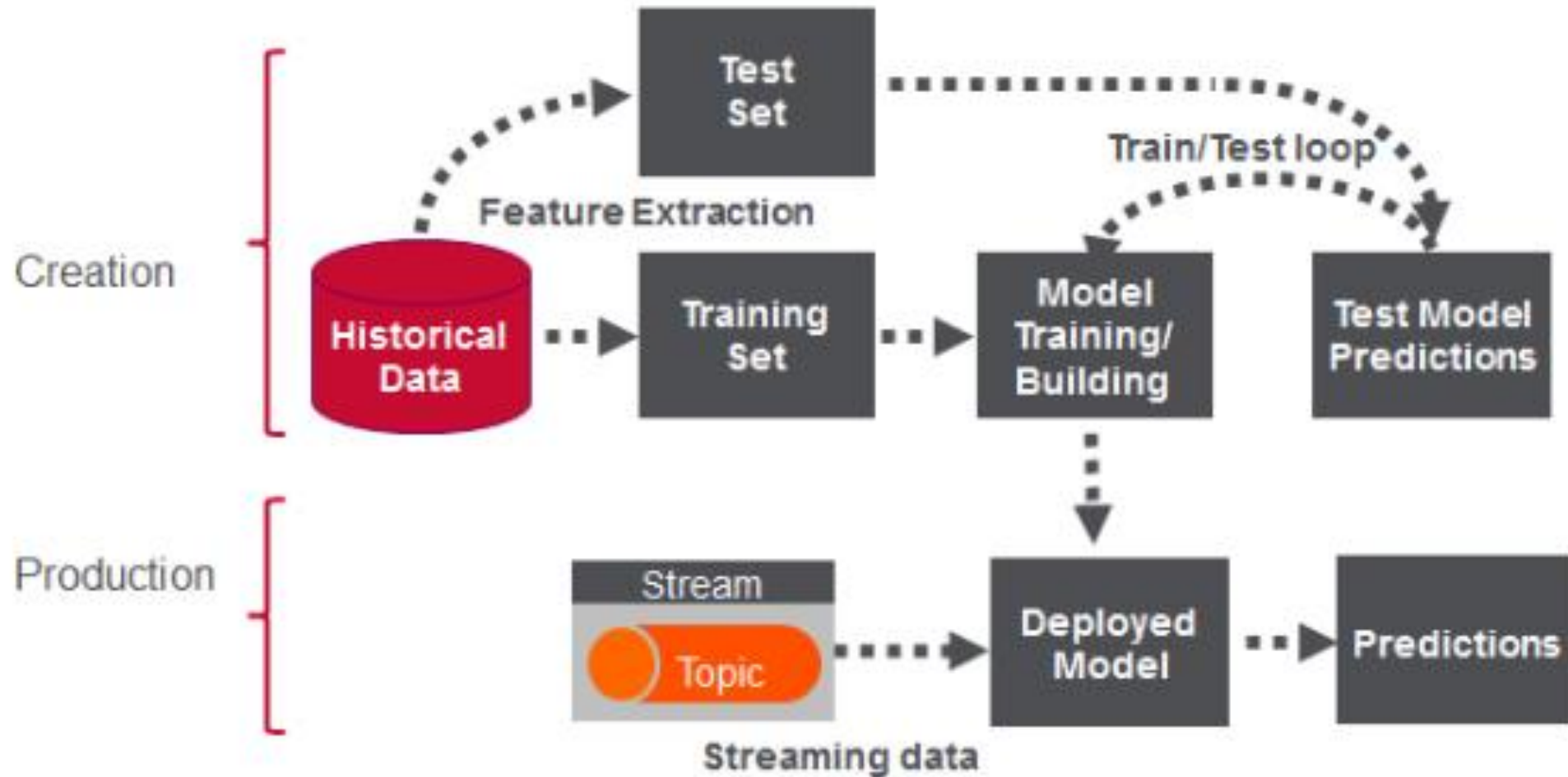
Network Intrusion Detection...

4. Tools

- Apache Spark (& HDFS)
- Python Programming
- MLlib

Intrusion Detection Phases

Two phases of fraud detection



Network Intrusion Detection...

Output:

(P, A) **P** – Predicted Label

(0, 0) **A** – Actual Label

(1, 0)

(0, 0)

(1, 1)

.....

.....

.....

Case Study 2: Recommender Systems

Flipkart



Search for Products, Brands and More



CART 0

[^ Back to top](#)

You may also be interested in

[All Categories](#) [Mobile Screen Guards](#) [Plain Cases & Covers](#) [Designer Cases & Covers](#)



Chevron Tempered Glass Guard for Google Pixel

2.4★ (27)

₹319 ₹599 46% off



Spigen Back Cover for Google Pixel

4.9★ (10) Assured

₹999 ₹1,099 9% off



Chevron Tempered Glass Guard for Google Pixel

2.8★ (17)

₹299 ₹599 50% off



Vatsin Screen Guard for Google Pixel

2.9★ (8)

₹442 ₹999 55% off



Dainty Tempered Glass Guard for Google Pixel

2.3★ (6)

₹361 ₹799 54% off



Aspir Back Cover for Google Pixel

5★ (1)

₹249 ₹999 75% off

Definition:

Recommender Systems are software tools that **elicit the interests and preferences** of individual consumers and **make recommendations accordingly**.

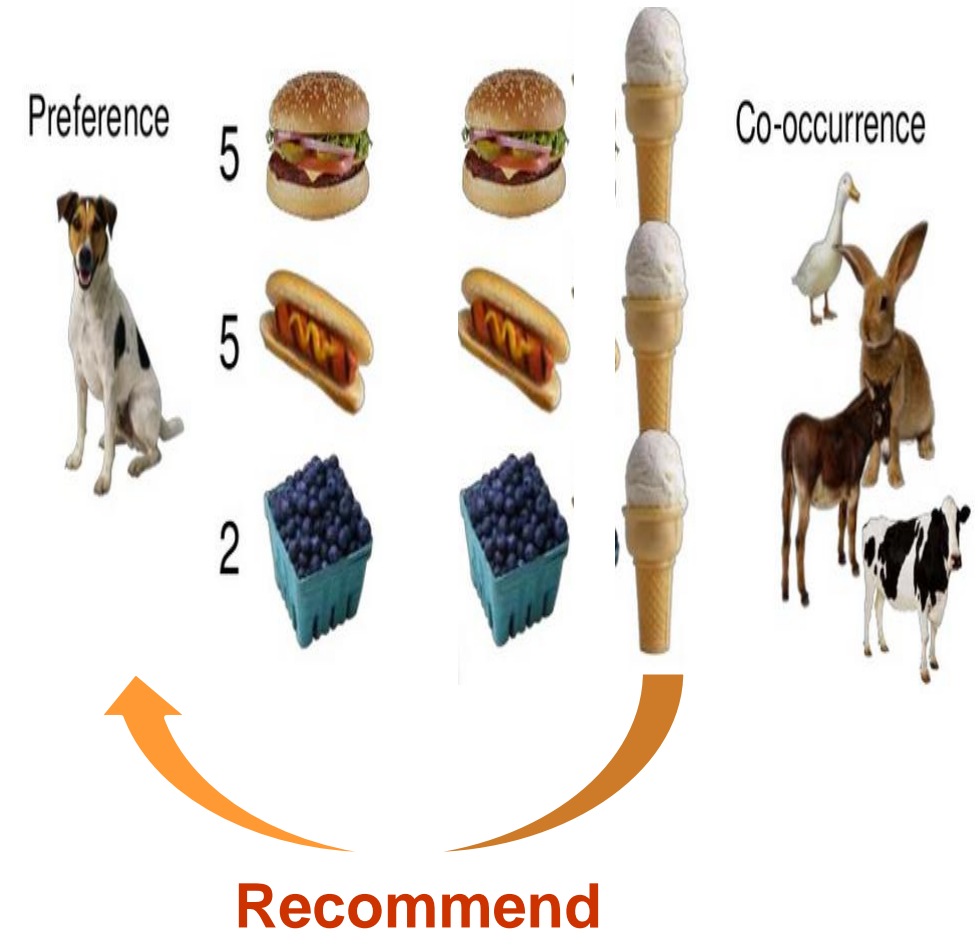
Types

- User based recommenders
- Item based recommender
- Matrix factorization based recommenders

RECOMMENDERS – USER BASED

(a) User based recommenders: (Collaborative filtering)

- **Predicting what users will like based on their similarity to other users.**
- **Assumption:** Users like the similar kinds of items they like in the past.
- Doesn't care whether the items are books, ice creams or mobile phones.
- Require no knowledge of the properties of the items
- **Example:** Facebook recommends friends



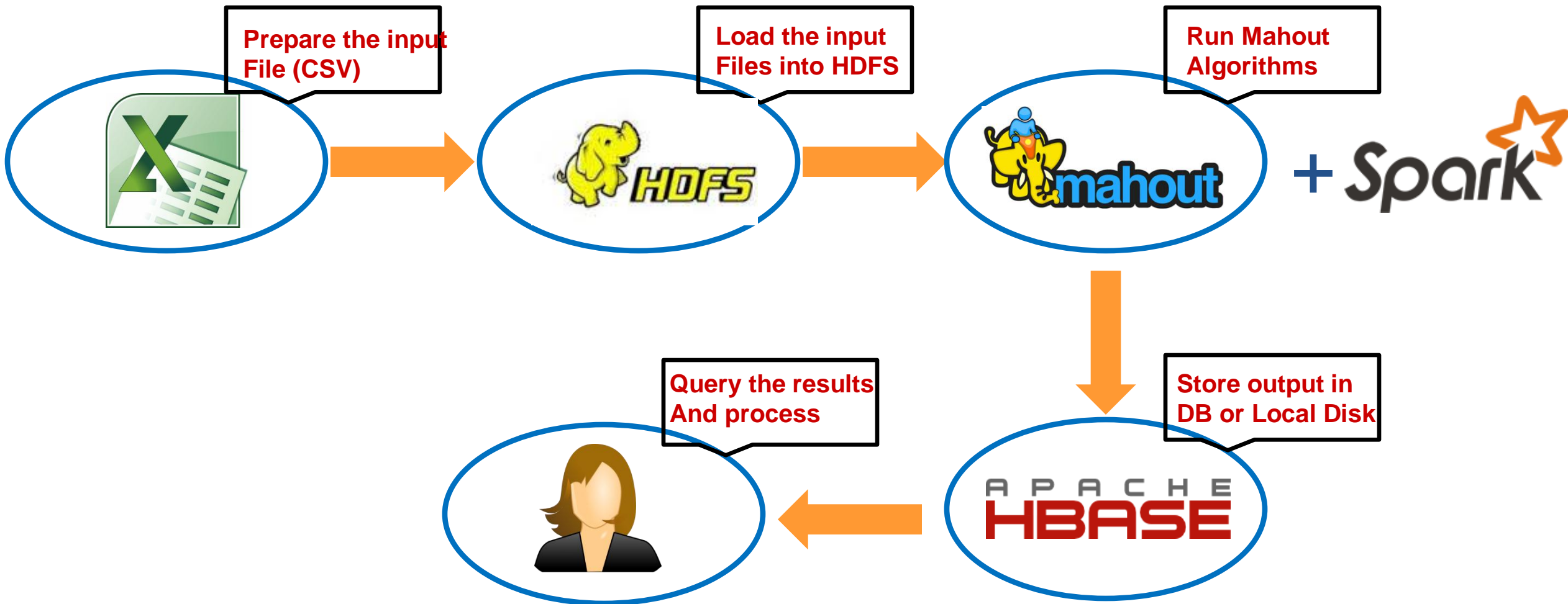
RECOMMENDERS – ITEM BASED

(b) Item based recommenders: (Content based filtering)

- Based on a description of the item and a profile of the user's preference
- A user profile is built to indicate the type of item this user likes
- Requires the knowledge of the properties of the items.



HOW IT WORKS?



I. OBTAINING DATA SETS AND PREPROCESSING

- **MovieLens** data set collected by **GroupLens** Project of **University of Minnesota**
- Provide the rating of movies from different users
- Each user has rated at least 20 movies
- Downloaded for generating user based and item based recommendations

Data Set	No. of Ratings	No. of Users	Ratings Range	No .of Movies rated	File Size
1) 100k	100,000 (1 Lakh)	943	1 to 5	1,682	1.8 MB
2) 1M	1,000,209 (~1 Million)	6040	1 to 5	3,900	21 MB
3) 10M	10,000054 (~ 10 Million)	71567	1 to 5	10,681	253 MB

2. RECOMMENDER ENGINE DETAILS

The details of this recommender are listed below:

- **Data Model:**
 - File Data Model
- **Algorithms:**
 - Pearson Correlation Similarity, Similarity Cooccurrence
- **Recommender:**
 - GenericItemBasedRecommender
- **Input:**
 - CSV file with input parameters such as rating, userID, Movie ID
- **Output:**
 - Text file with list of recommendations

3. OUTPUT & BENCHMARKS

Data

Data Set
1) 100k (~1 Lakh Ratings)
2) 1M (~1 Million Ratings)
3) 10M (~10 million Ratings)

- Each line represents the recommendation for a user.
- The first number is the user id and the 10 number pairs represents a movie id and a score.

Output.txt (From 100K dataset)

```
1
[2478:5.0,237:5.0,2133:5.0,368:5.0,4489:5.0,161:5.0,
6942:5.0,1016:5.0,6753:5.0,209:5.0]

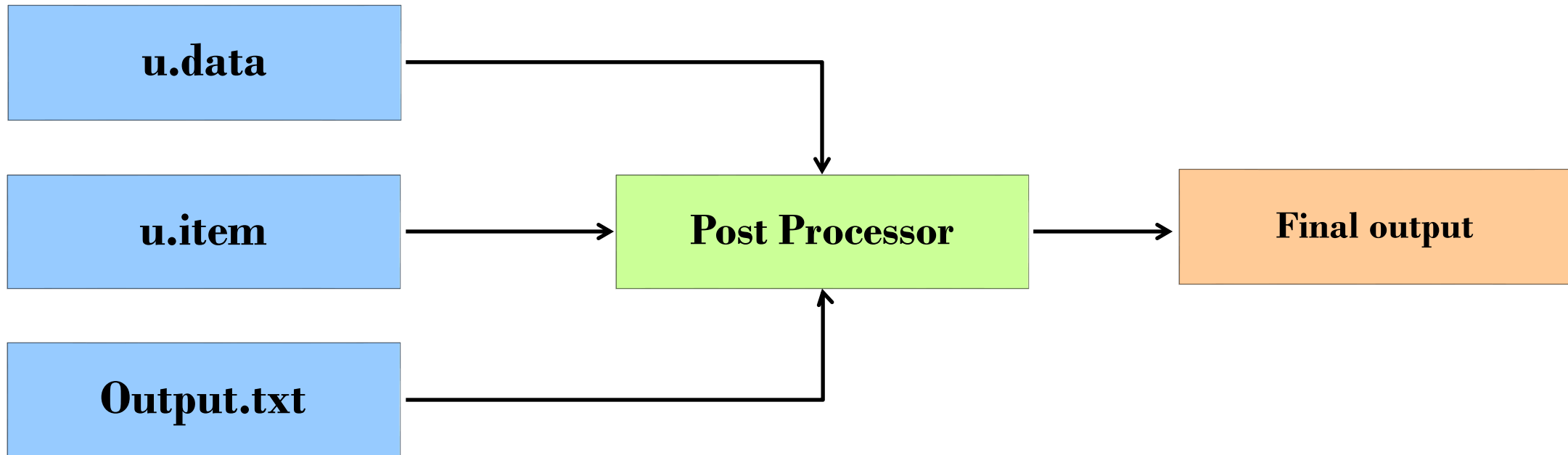
2
[1135:5.0,2022:5.0,1272:5.0,1079:5.0,279:5.0,485:5.0
,3763:5.0,7325:5.0,6059:5.0,3168:4.67]

.....
.....
.....

71567
[2018:5.0,2021:5.0,1203:5.0,3254:5.0,2871:5.0,6709:
5.0,2709:5.0,2134:5.0,2912:5.0,1304:5.0]
```

4. POST PROCESSING OUTPUT

- It's not easy to see what those recommendation means
- A simple script is developed to post process the output



```
$ recommendations.py 4 u.data u.item output.txt
```

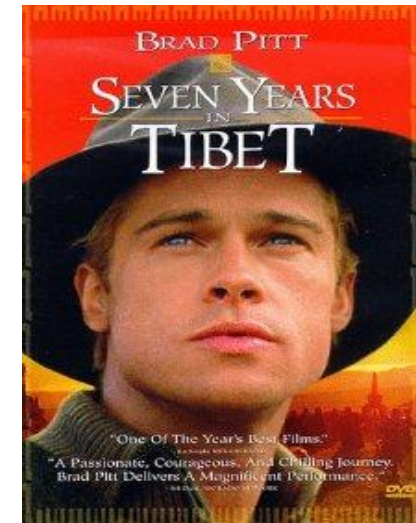
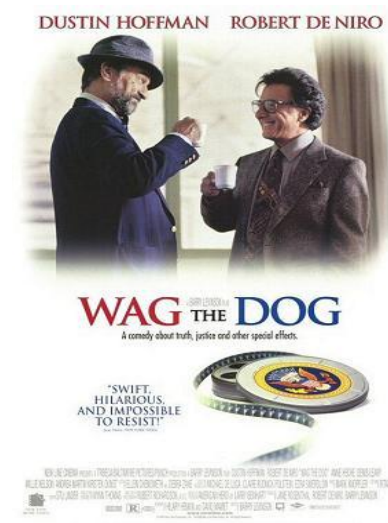
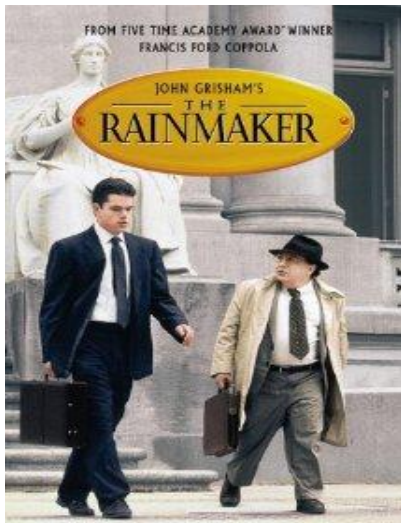
5. OUTPUT

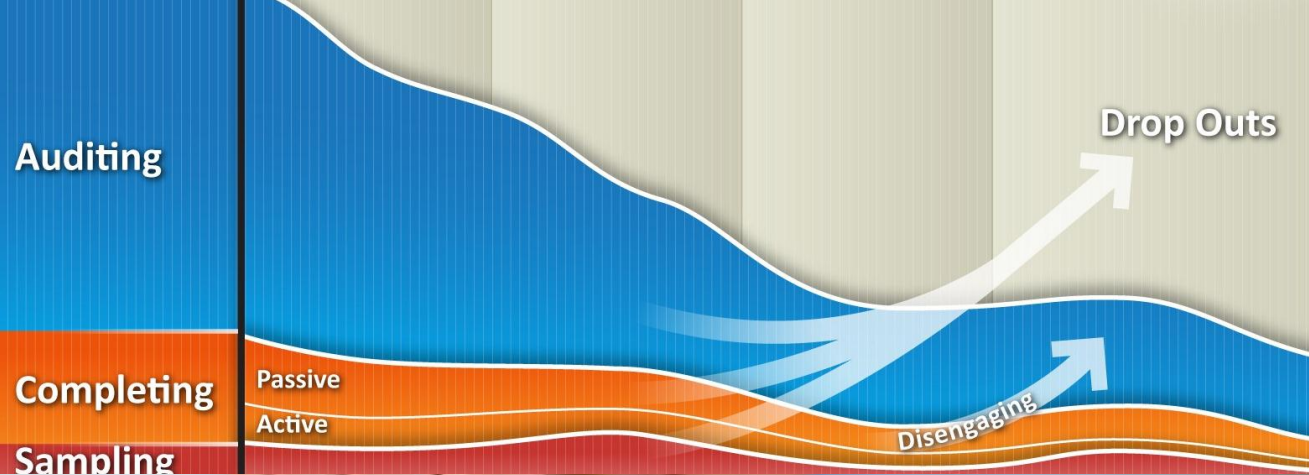
The recommended movies for the user4 are listed below:

Movie	Score
Rainmaker, The (1997)	Score=5.0
House of Yes, The (1997)	Score=5.0
Kull the Conqueror (1997)	Score=5.0
Wag the Dog (1997)	Score=5.0
Seven Years in Tibet (1997)	Score=5.0
Beautician and the Beast, The (1997)	Score=5.0
Cats Don't Dance (1997)	Score=5.0
Mighty Aphrodite (1995)	Score=5.0
I Know What You Did Last Summer (1997)	Score=5.0
Sense and Sensibility (1995),	Score=5.0

5. OUTPUT...

The top 5 movies recommended to “user4” are listed below:

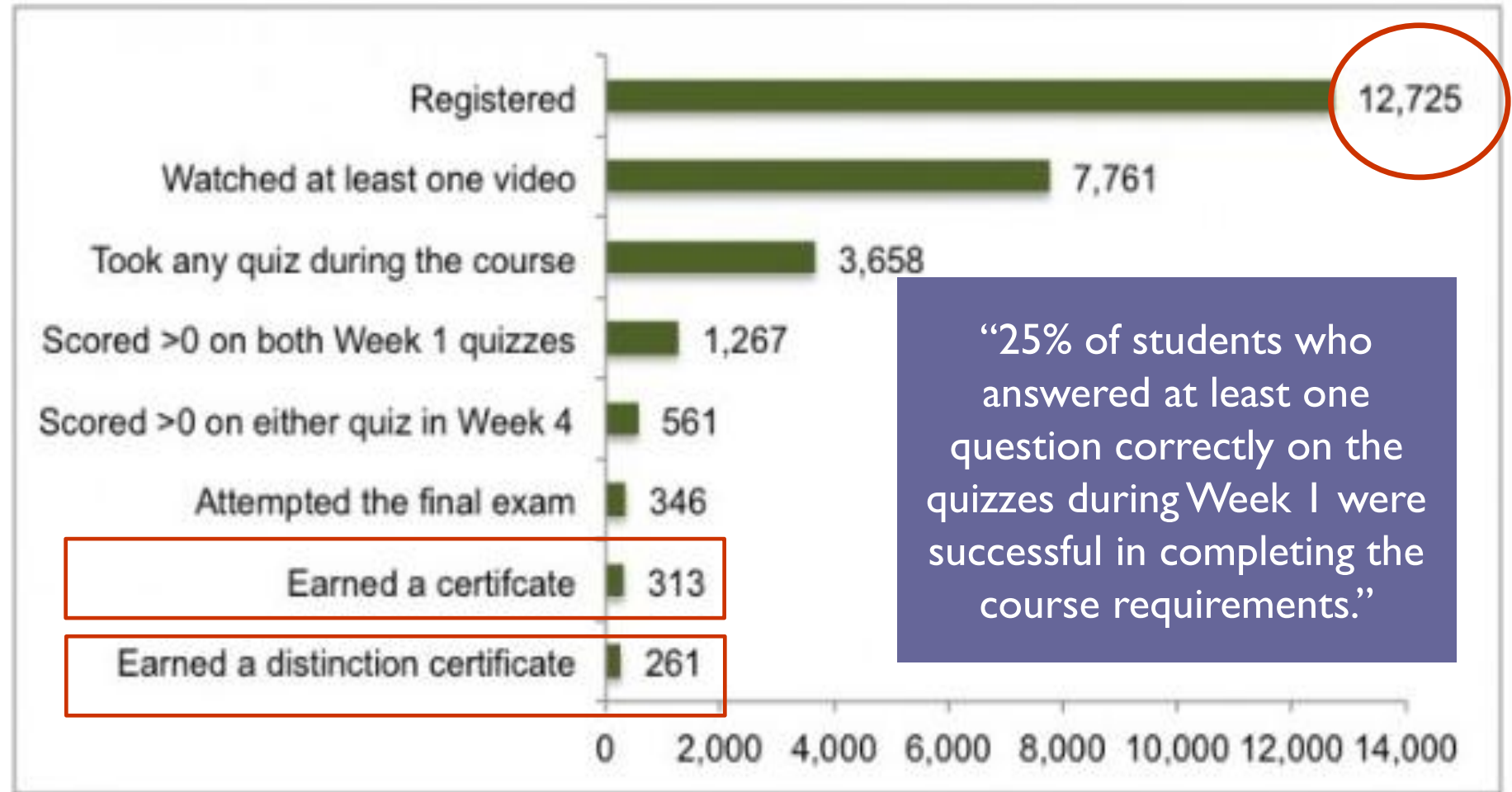




Case Study 3: Analytics in MOOCs

MOOCS - CHALLENGES

MOOC on Bioelectricity - From Duke University



Course completion
rates ????

Approx. 2 %

PROBLEM

Predict whether a user will drop a course within next 10 days based on his or her prior activities.

How to solve?

If a user 'U' leaves no records for course "C" in the log during the next 10 days.



- XuetangX, a Chinese MOOC learning platform, in collaboration with Edx

Details about datasets: 2 Lakh students, 1.35 Crore logged events

Enrollment

enrollment_id,
username,
course_id

Log

enrollment_id,
tstamp,
Source,
Logged_event,
Object

Object

course_id,
module_id,
category,
tstart

Truth Info

enrollment_id,
dropped_out

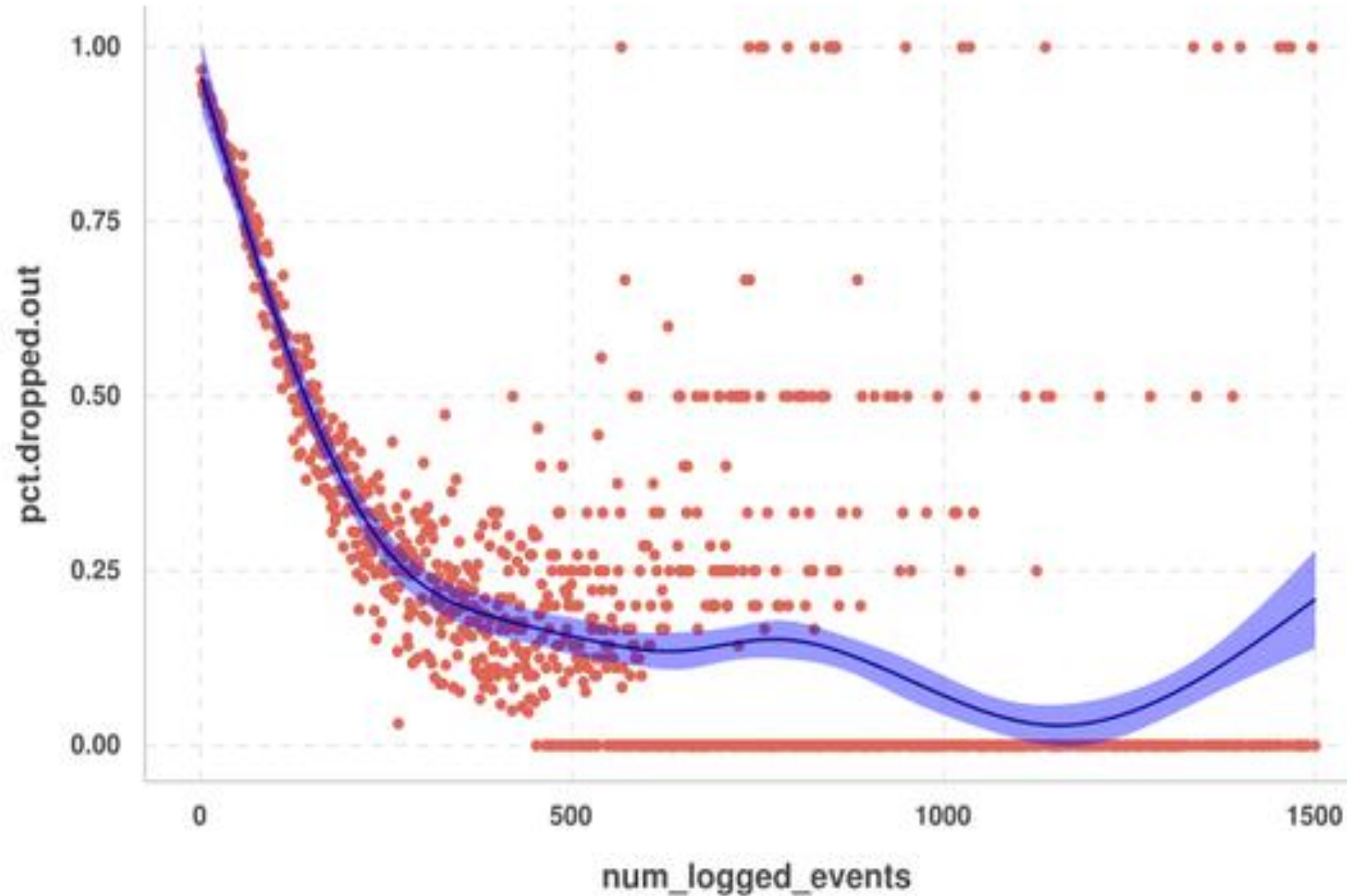
ANALYTICS

Less number of Logged Events

High probability of drop-out



Proportion of MOOC students who dropped out by # of logged online interactions with the website.





Case Study 4: Distracted Driver Detection & Alarming System

OBJECTIVE OF THE PROJECT

According to the CDC motor vehicle safety division, **one in five car accidents is caused by a distracted driver.**

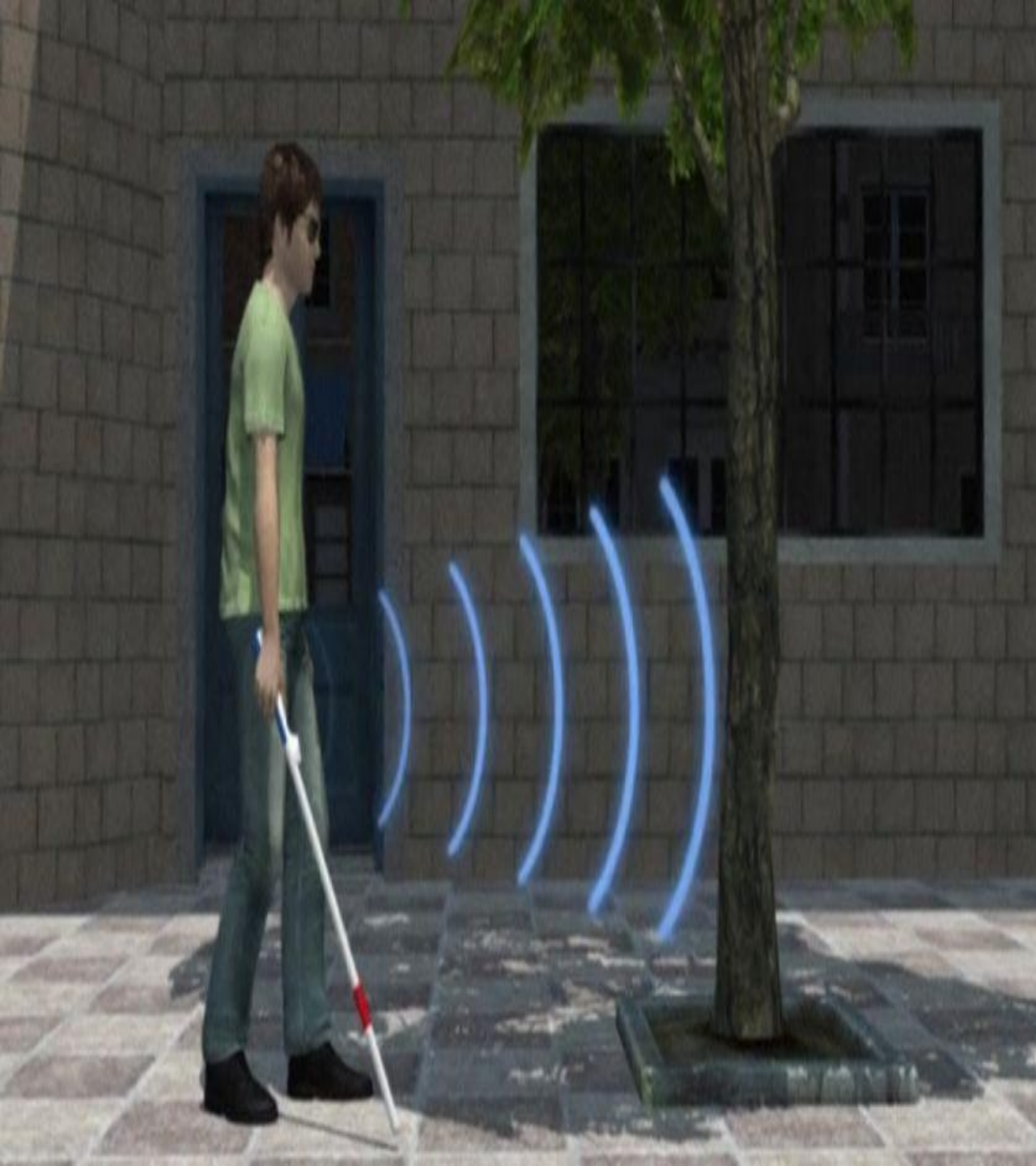


Predict the likelihood of what the driver is doing from the stream of video frames of a driver.

Deep Learning

10 classes to predict:

- c0: safe driving
- c1: texting - right
- c2: talking on the phone - right
- c3: texting - left
- c4: talking on the phone - left
- c5: operating the radio
- c6: drinking
- c7: reaching behind
- c8: hair and makeup
- c9: talking to passenger



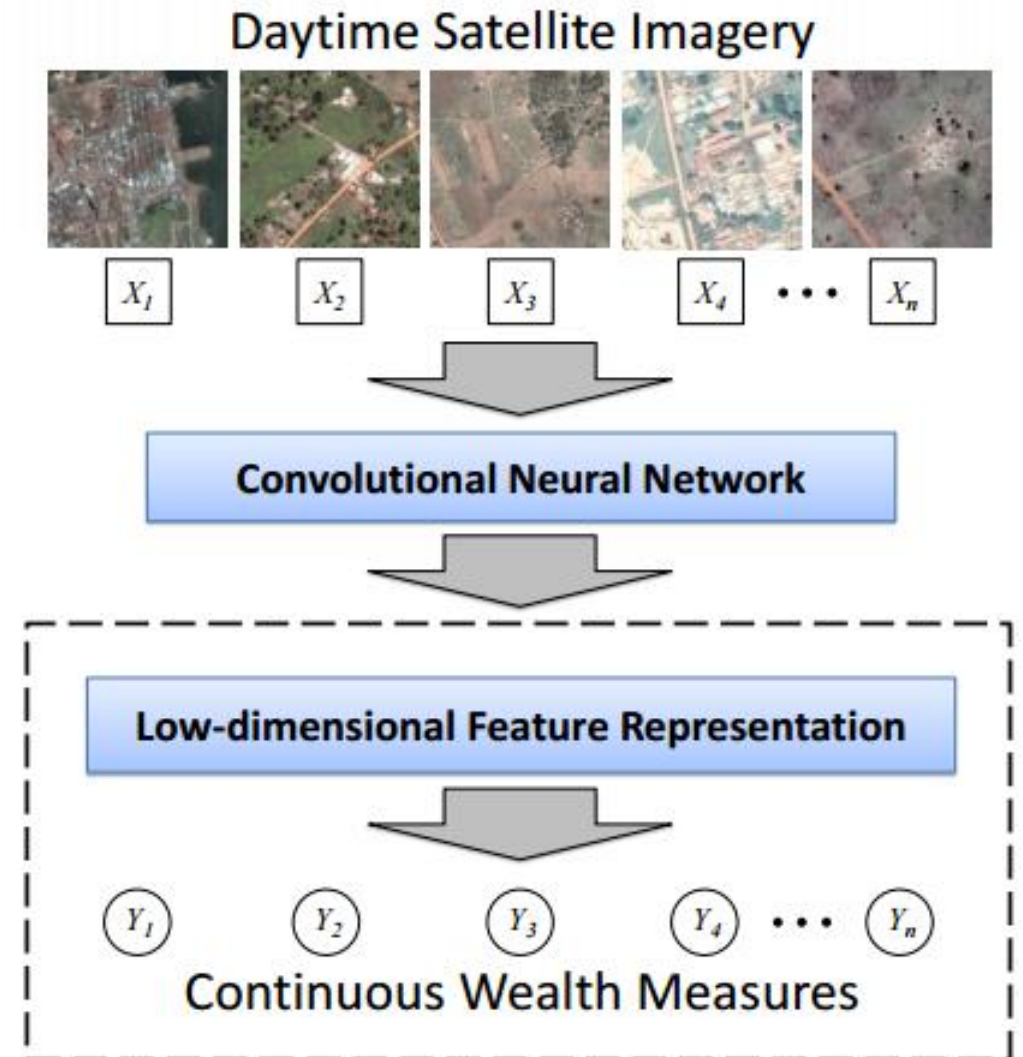
Case Study 5: Vision Based Blind Stick



What type of projects we can do using ML?

FINDING POVERTY USING SATELLITE IMAGES (STANFORD)

1. Lack of reliable poverty data in developing countries poses a major challenge for making informed policy decisions
2. Comprehensive surveys are often prohibitively expensive – a cheap, scalable method of producing detailed poverty maps would greatly facilitate economic progress



EMOTION CLASSIFICATION ON FACE IMAGES

1. It can also be useful for research on behaviors on social networks.



Surprise



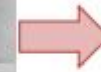
Happiness

2. Seven emotions:

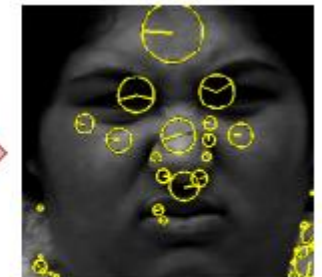
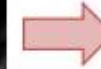
- anger
- contempt
- disgust
- disgust
- fear
- happiness
- sadness
- surprise



Raw image



Pre-processing



Local descriptors

OTHER PROJECTS

1. Predicting Future Employment, Productivity, and Income
2. Using Decision Tree to predict repeat customers
3. Personalized Company Recommender System for Job Seekers
4. Advanced machine learning techniques for thyroid cancer diagnosis
5. Estimating Medical Costs with Machine Learning
6. Classifying brain tumors

“BIG DATA” - “MACHINE LEARNING”

Amazon has billions of users, still it manages to recommend you based on your search history

Your Recently Viewed Items and Featured Recommendations

Inspired by your browsing history

Page 1 of 10

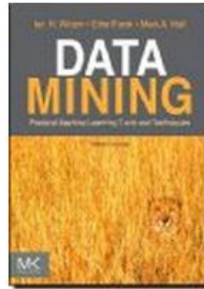


Learning From Data

› Hsuan-Tien Lin

★★★★☆ (67)

Hardcover



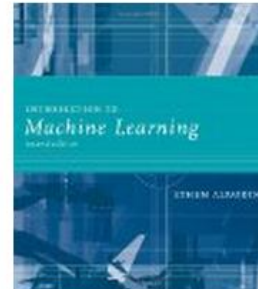
Data Mining: Practical Machine...

› Ian H. Witten

★★★★☆ (46)

Paperback

\$44 01 Prime



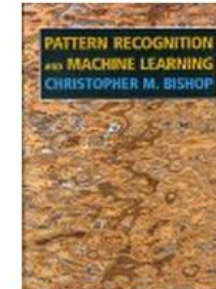
Introduction to Machine Learning

› Ethem Alpaydin

★★★★☆ (26)

Hardcover

\$45 14 Prime



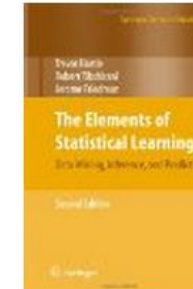
Pattern Recognition and Machine...

› Christopher M. Bishop

★★★★☆ (101)

Hardcover

\$61 92 Prime



The Elements of Statistical...

Trevor Hastie

★★★★☆ (40)

Hardcover

\$65.81 Prime



MACHINE LEARNING TOOLS

Python



Java



.net



We need scalable Machine Learning/Data Mining Algorithms



Java, Scala



Spark

MLlib

Python, Scala



Statistical Analysis

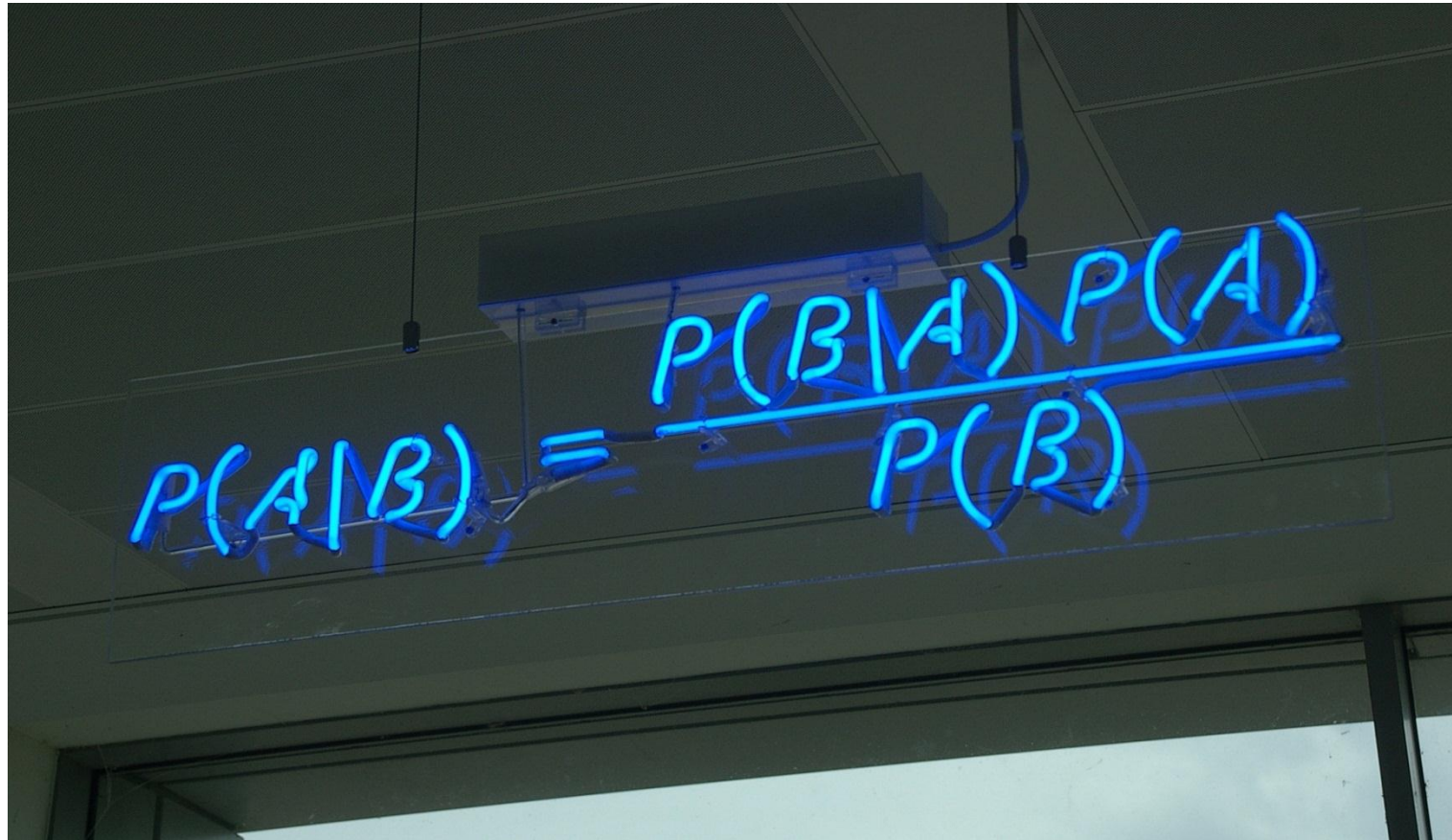
C, FORTRAN Languages

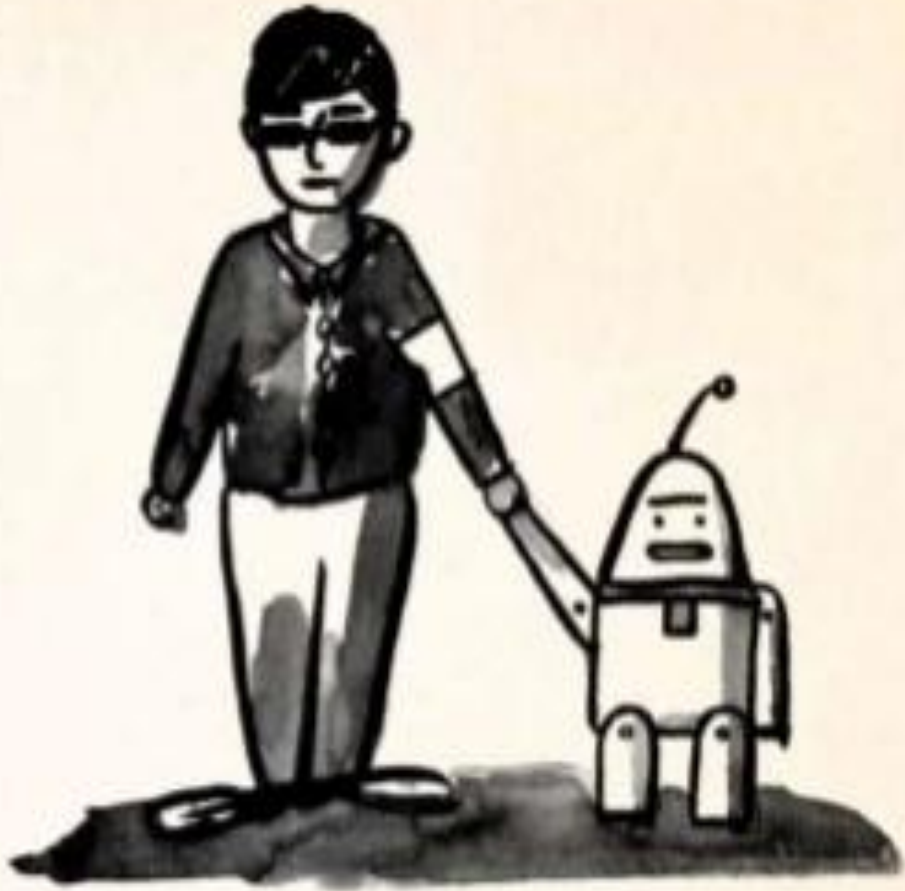
MACHINE LEARNING IN DEFENCE IS GAINING TRACTION...

Defense Advanced Research Projects Agency (DARPA) launches PPAML program to move machine learning forward.

Probabilistic
Programming for
Advancing
Machine
Learning

2013 - 2017





THANK YOU!

QUESTIONS?