CS 419: Introduction to Machine Learning

http://www.cse.iitb.ac.in/~cs419

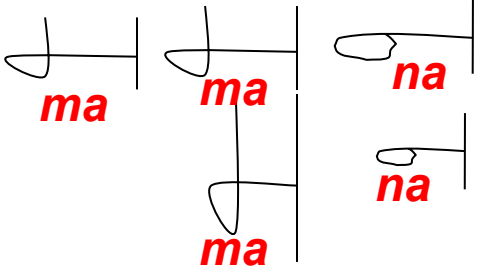
Machine Learning

- The design of algorithms to enhance the quality of specific tasks performed by a computer based on exemplar data.
  - Example: object recognition, forecasting
- Why learn, instead of creating perfect programs?
  - Difficult to understand perfect relationship between input and output in many cases

Character recognition

TRAINING

User inputs
(character-strokes, label)



Ma & Na examples

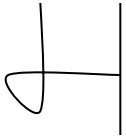
Machine Learns

Trained Model

ML component

TESTING

User draws
character-strokes



Trained Model

read model

Testing

"ma"

OCR

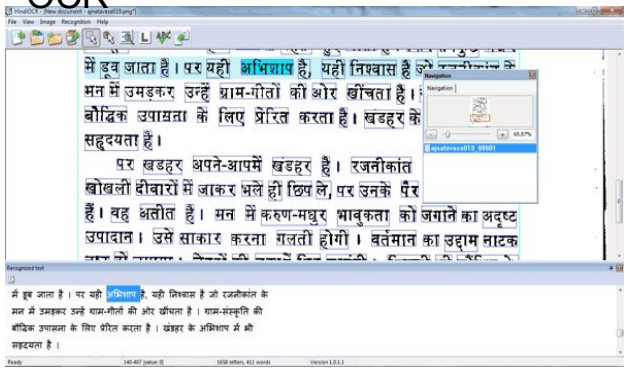


Image recognition

Image from "ImageNet classification with deep CNNs", Krizhevsky et al.

Translation

Input: **x**

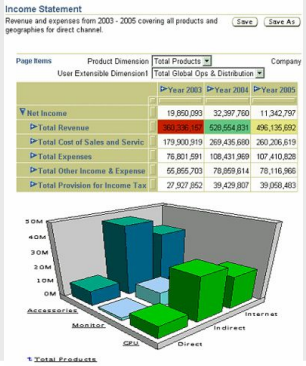
Output: **y**

Where can I find healthy and traditional Indian food? → स्वस्थ और पारंपरिक भारतीय भोजन कहां मिल सकता है?

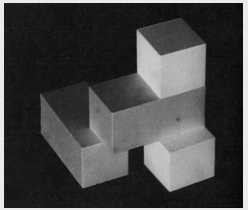
Applications of Machine learning

Unstructured information management: text mining, Web graph analysis, natural language processing, automatic translation

Retail, Banking
Commerce
Advertisements



Artificial intelligence



Knowledge representation, rule bases, logic, inference, planning, decision making

Speech and vision

Statistical models for spoken language and still and motion pictures, objects therein

GrabCut

- Roughly drag region around object you want to cut
- Paste into another image: how to blend?
- Segmenting into foreground and background
- Unsupervised learning problem
  - Only a few pixels define the boundary between foreground

Examples from Kaggle.com

isaggregation Competition
d energy consumption into individual appliances

Amazon.com - Employee Access Challenge
Predict an employee's access needs, given his/her job role

Multi-modal Gesture Recognition
Recognize gesture sequences in video and depth data from Kinect

Cause-effect pairs
Given samples from a pair of variables A, B, find whether A is a cause of B.

MLSP 2013 Bird Classification Challenge
Predict the set of bird species present in an audio recording, collected in field conditions.

AMS 2013-2014 Solar Energy Prediction Contest
Forecast daily solar energy with an ensemble of weather models

RecSys2013: Yelp Business Rating Prediction
RecSys Challenge 2013: Yelp business rating prediction

MasterCard - Data Cleansing Competition

Innumerable Other Applications

- Ad placement in search engines
- Inventory management: Predict sale of soft drinks in outlets based on weather, events (sport)
- Scheduling: predicting traffic, flight arrival times.
- Fraud detection: telecommunications, financial transactions
  - from an online stream of event identify fraudulent events
- Banking: loan/credit card approval
  - predict good customers based on old customers
- Customer relationship management:
  - identify those who are likely to leave for a competitor.
- Targeted marketing:
  - Recommendation of Movies, Books, Products on E-commerce sites

Applications (continued)

- Medicine: disease outcome, effectiveness of treatments
  - analyze patient disease history: find relation between diseases
- Molecular/Pharmaceutical: identify new drugs
- Scientific data analysis:
  - identify new galaxies by searching for sub clusters
- Image and vision:
  - Remove noise from images, Identifying scene breaks
- Education
  - Automatic grading of essays, selecting questions for exams

Types of tasks

- Predictive: Input-output functions
  - Output a real-value, given several input variables. E.g. forecasting, credit-card scoring **(Regression)**
  - Output a class label, given input variables. E.g. recognizing digits **(Classification)**
  - Output a structure, given an input object
    - E.g. Alignment between sentences in two languages
    - Extracting structured data from an address
- Discovery
  - Finding groups in data, patterns that hold in data, find abnormalities, projection, factorization

Types of tasks

- Supervised
  - Given supervision as examples of correct input to output mapping, learn a model
    - That fits the examples
    - Generalize to unseen examples
- Unsupervised
  - Given several examples but no expected output for each example, learn a model
- Combinations:
  - Semi-supervised, indirectly supervised, actively supervised

Relationship to AI and DL

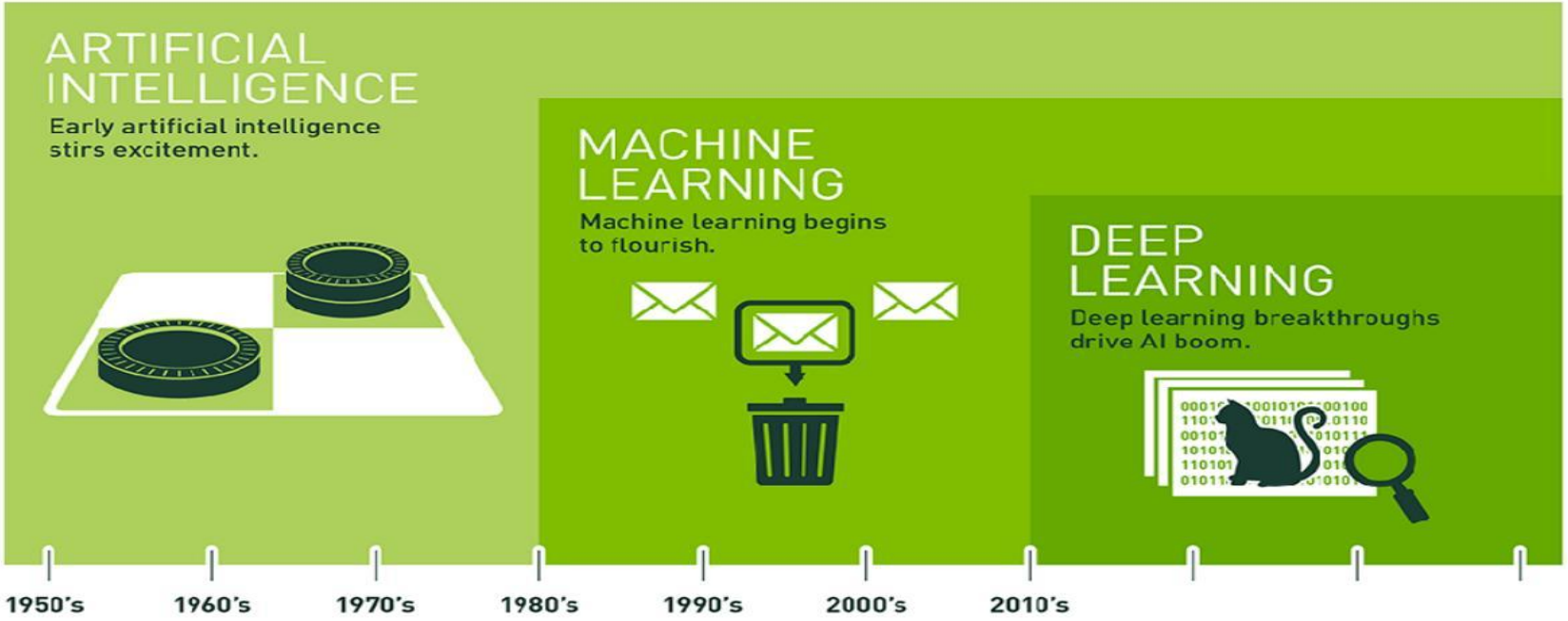


Image from: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

Course information

- Course Web page
  - http://www.cse.iitb.ac.in/~sunita/cs419

Course contents

- Machine learning tasks
  - Discrete classification:
    - decision trees, nearest neighbor
    - generative classifiers,
    - discriminative classifiers: logistic regression, SVMs
    - Deep Learning
  - Clustering: EM, hierarchical, k-means

The study process

- Pay attention to the class: don’t hesitate to ask questions
  - (Old Chinese proverb)
    - The one who asks a question is stupid only once, the one who does not ask questions is stupid forever.
- For every 1.5 hours of lectures, spend at least 1.5 hours in revising the lecture after class
- Do the homework yourself
- If you have difficulty, come to instructor’s office hours early on in the semester

Study material

- Unfortunately, no single text.
- Two of the most relevant text books mentioned on the course webpage
- Each topic will contain pointer to reading material on that topic
- My board work on the tablet will be available for reference.
- Exams will be open notes, but you cannot xerox someone else’s notes or my board work.

Data formats

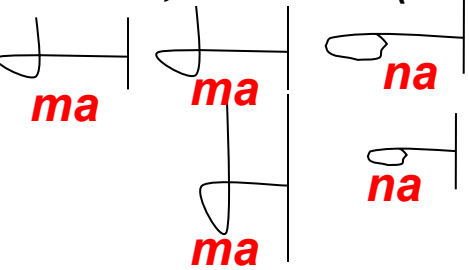
- Of various shapes and sizes. Eg. Document, speech signal, bitmap of images.
  - Set of instances/objects/cases/rows/points/examples
- An application-specific method of transforming data into this most common format.
  - fixed set of attributes/dimensions/columns
    - Continuous
    - Categorical

Basic notions (Classification)

ML component

TRAINING

Inputs, Examples, Trials,Instances (labelled)



Training/ Learning/ Model Building/

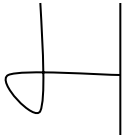
Classifier/ Function/ Predictor/ Model

TESTING

Trained Model

read model

Test example (unlabelled)



Testing

Predicted label

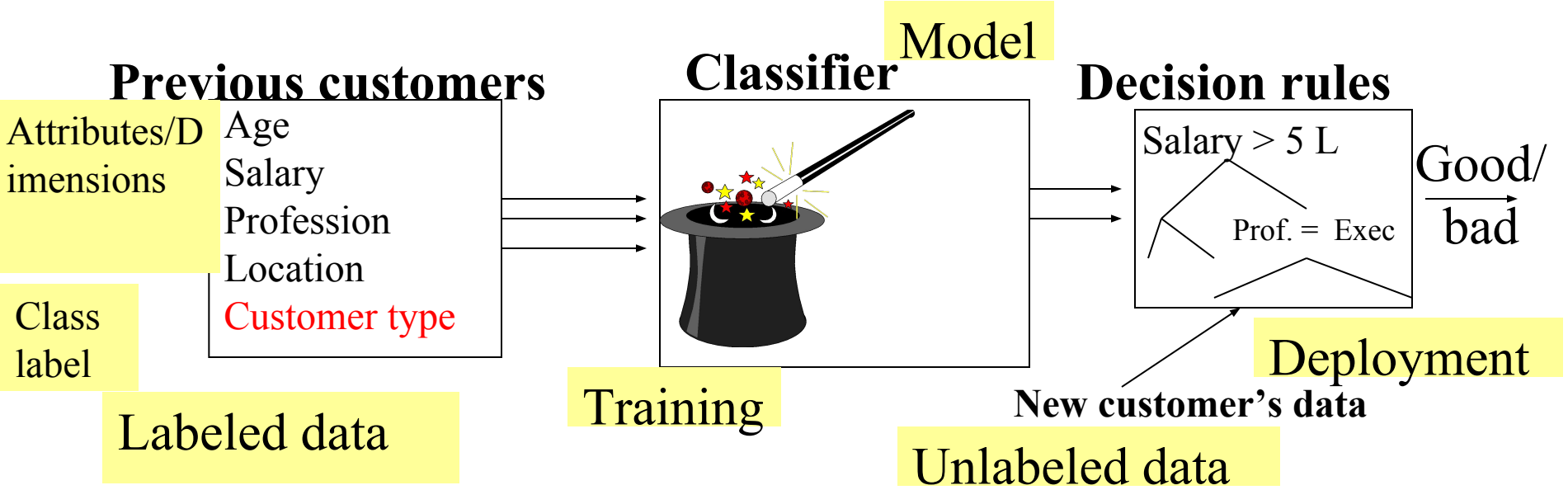
ma

Performance and reward/loss (0/1 loss, recall, precision, F1, ROC, AUC)

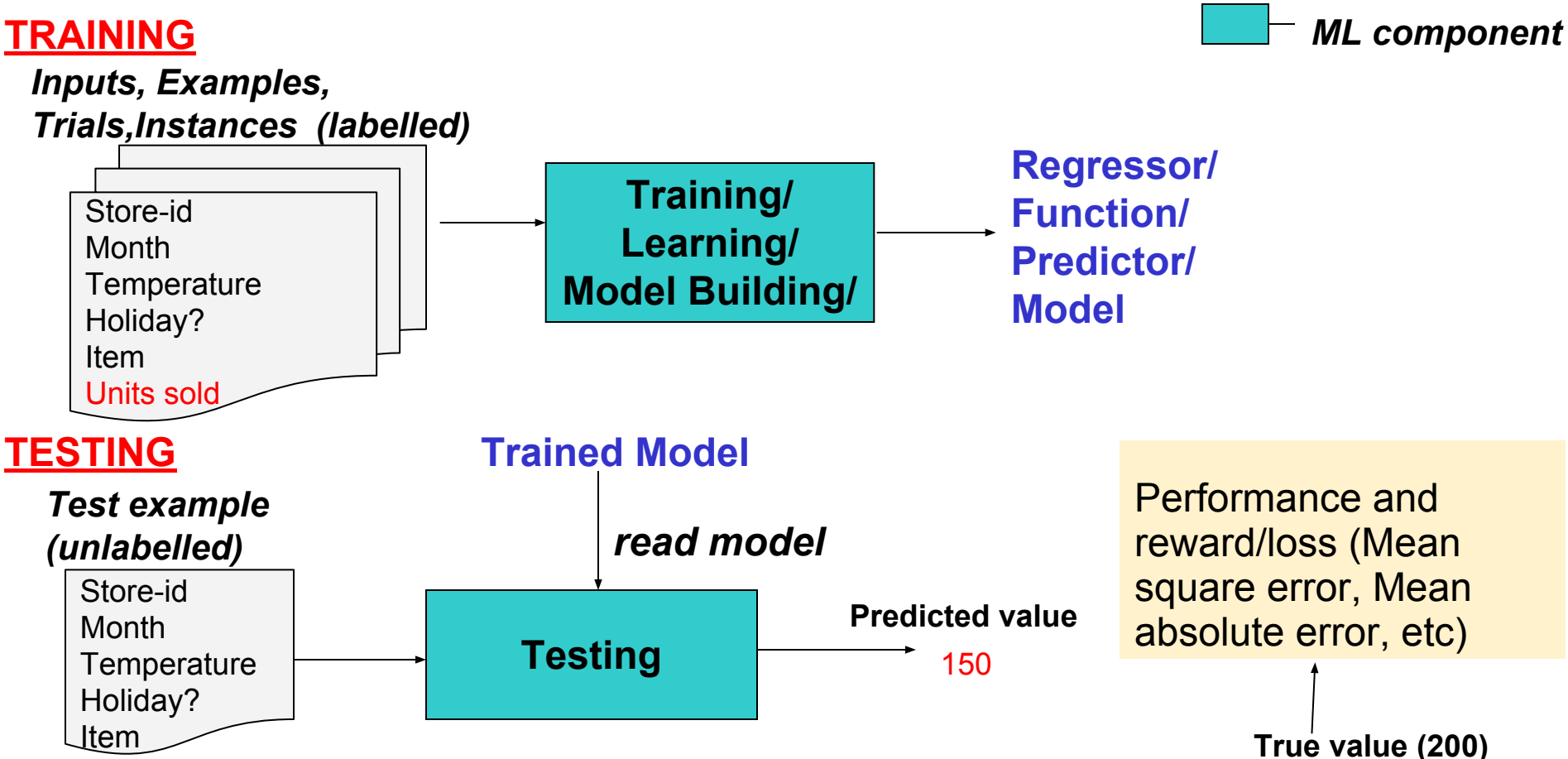
True label

Another example (Classification)

- Given old data about customers and payments, predict new applicant's loan eligibility.



Basic notions (Regression, Forecasting)



Some basic notions

- Trial, observation, instance, attributes
- Variable(s) to predict
  - Continuous, discrete, structured discrete
- Model, function, classifier, predictor
- Performance and reward/loss
  - Square loss, p-Norm loss
  - 0/1 loss, recall, precision, F1, ROC, AUC
- Hypothesis (class, space)
- Generalization power, model complexity

Discovery: clustering results of entity searches

- Input:
  - Several URLs that much a search query “Ashish Gupta”
- Output:
  - Clusters of urls, with each cluster hopefully referring to the same physical person.

Discovery: finding abnormal regions in a 2D space

- Input:
  - People and their 2D coordinate
  - Stores and their 2D coordinates
  - Number of purchases of gastroentities medicine in each store
- Discover:
  - Regions where the number of purchases is abnormally high.

The process of making learning models

- Problem formulation
- Data collection
  - subset data: sampling might hurt if highly skewed data
  - feature selection: principal component analysis, heuristic search
- Pre-processing: cleaning
  - name/address cleaning, different meanings (annual, yearly), duplicate removal, supplying missing values
- Transformation:
  - map complex objects e.g. time series data to features e.g. frequency