INTRODUCTION TO MACHINE LEARNING

PRASAN YAPA

Tech Lead, Independent Researcher, Lecturer

AGENDA

- Why Machine Learning?
- Introduction to Machine Learning
- Types of Machine learning
- Resources
- Common Issues and Challenges

ABOUT

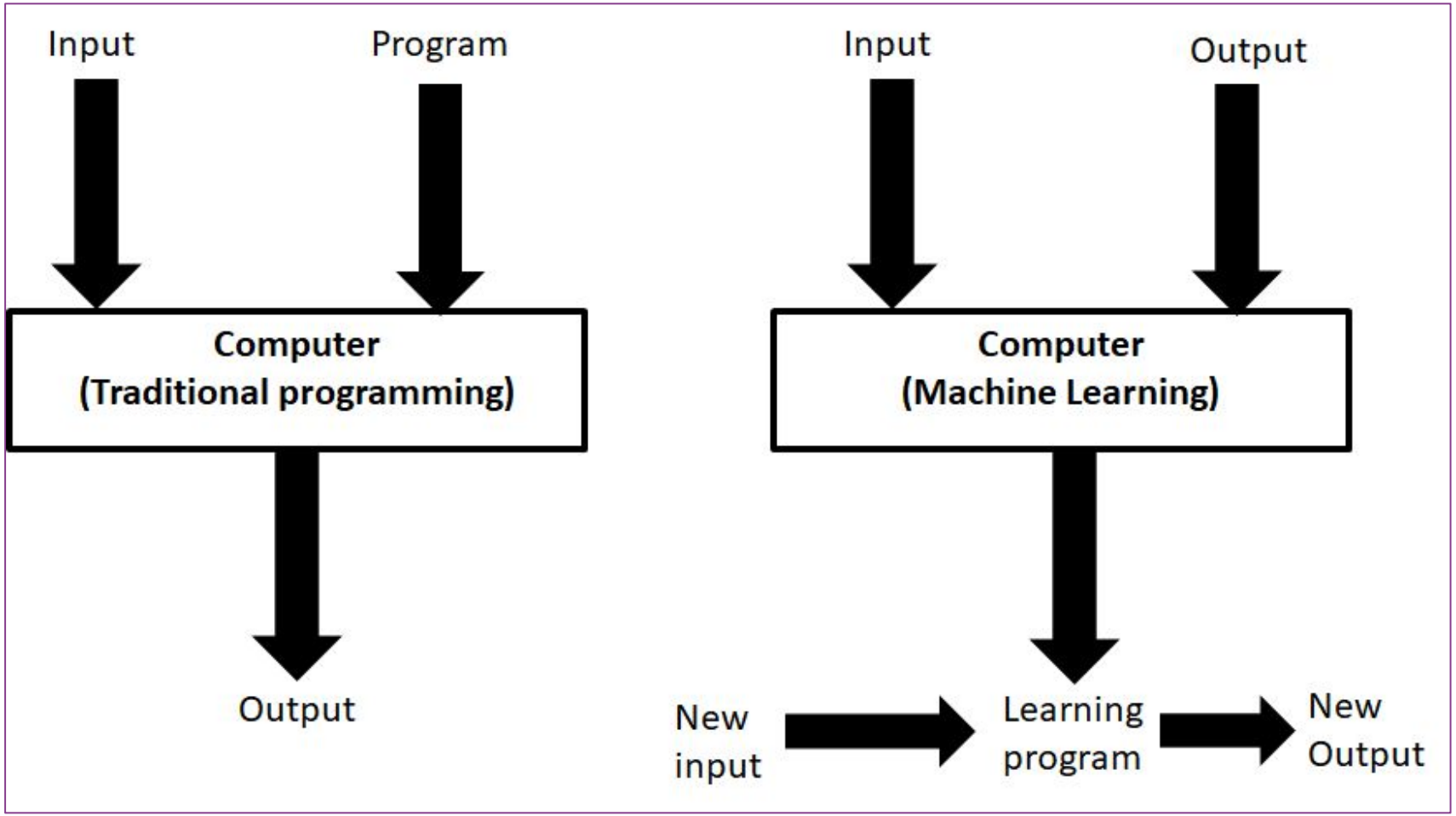
- Subfield of Artificial Intelligence (AI)/Application of optimization
- Name is derived from the concept that it deals with “construction and study of systems that can learn from data”
- Can be seen as building blocks to make computers learn to behave more intelligently
- It is an applied field of study. There are various techniques with various implementations

WHY

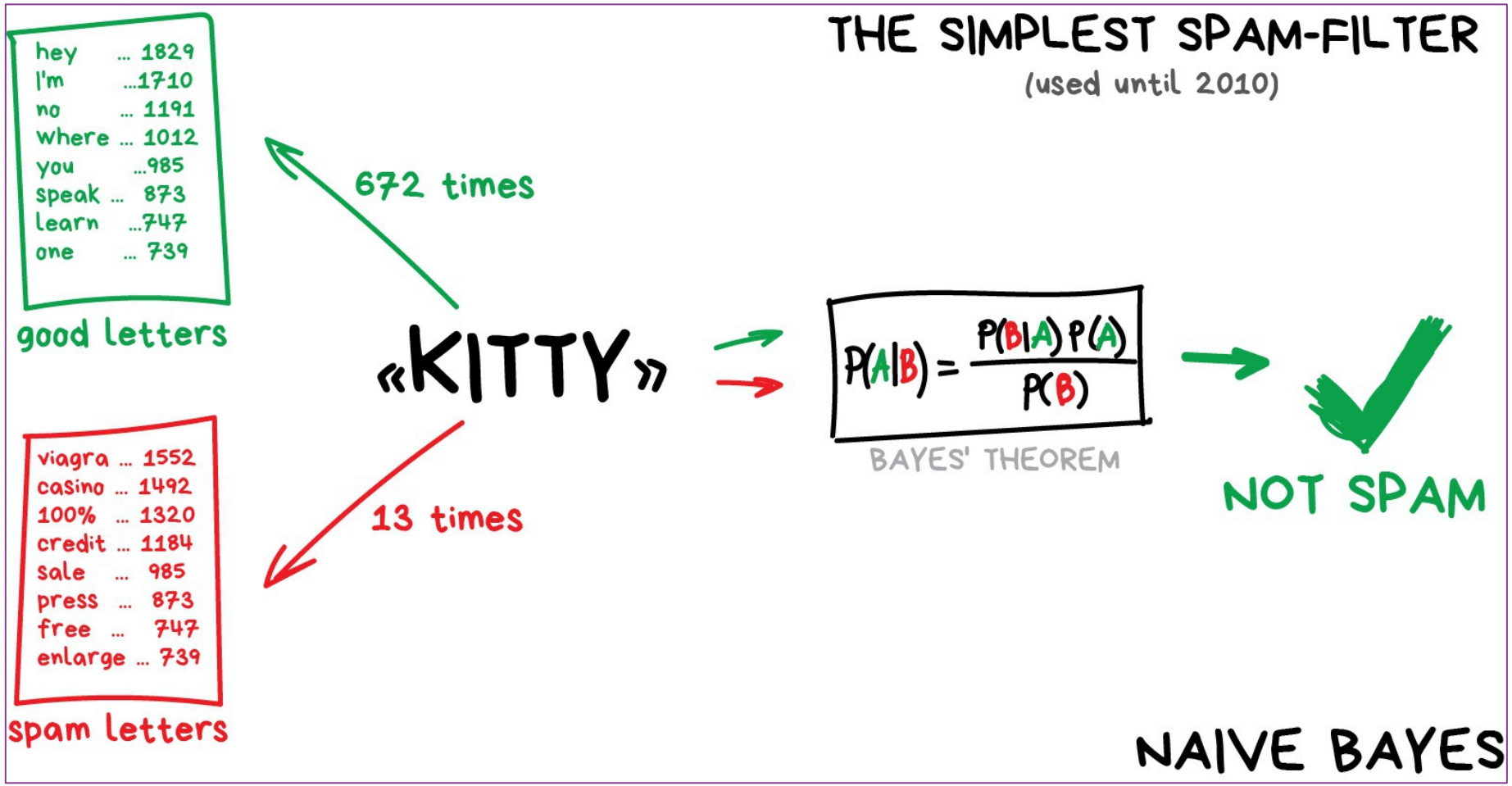
- Flood of available data (especially with the advent of the Internet)
- Increasing computational power (e.g. Multi-core)
- Growing progress in available algorithms and theory developed by researchers
- Increasing support from industries
- Cloud computing

“A COMPUTER PROGRAM IS SAID TO LEARN FROM EXPERIENCE (E) WITH SOME CLASS OF TASKS (T) AND A PERFORMANCE MEASURE (P) IF ITS PERFORMANCE AT TASKS IN T AS MEASURED BY P IMPROVES WITH E”

ML VS TRADITIONAL PROGRAMMING



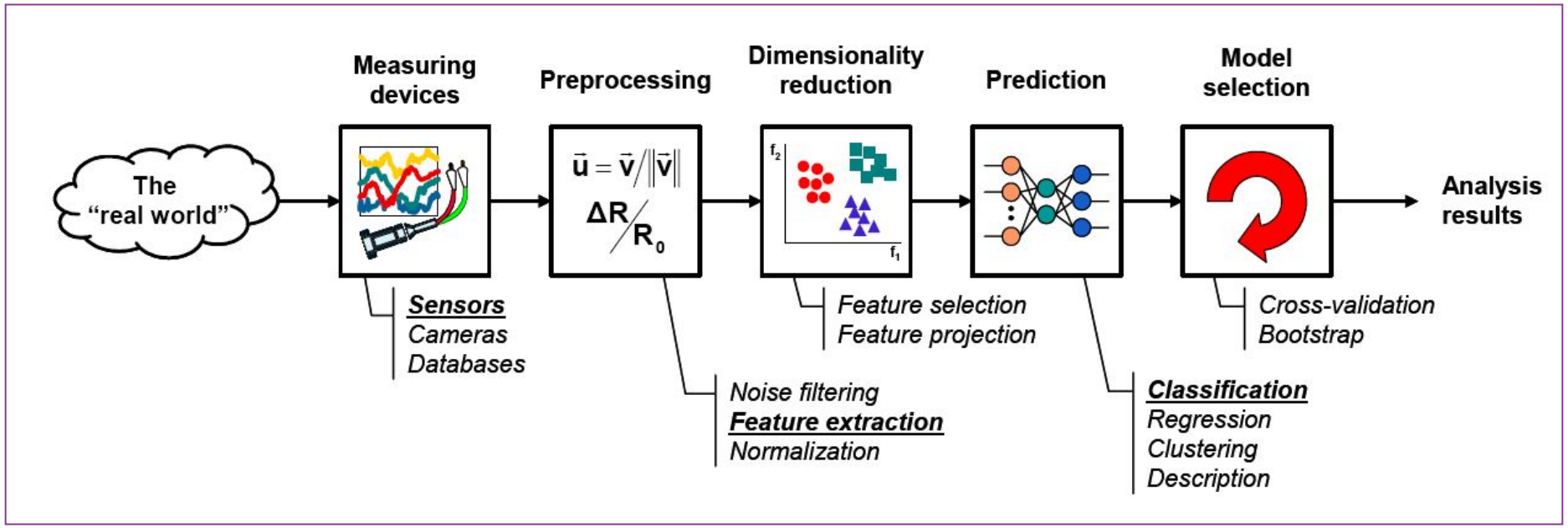
SPAM FILTERING



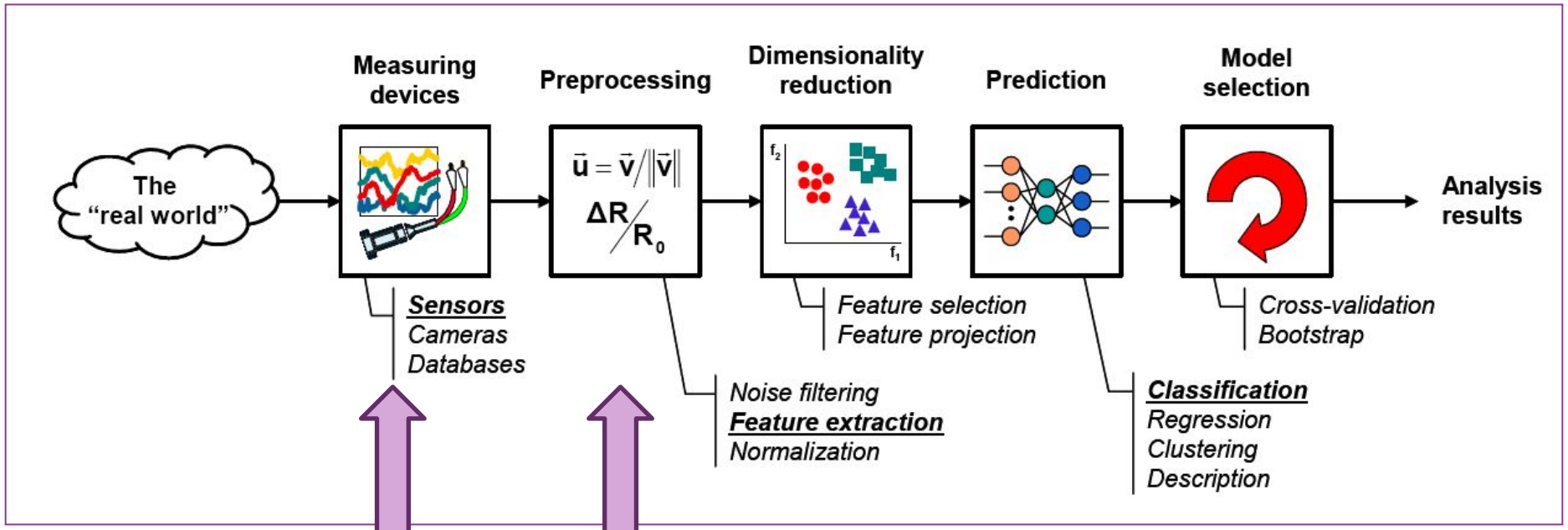
TERMINOLOGY

- Features
  - number of features or distinct traits that can be used to describe each item in a quantitative manner.
- Samples
  - sample is an item to process (e.g. classify). It can be a document, a picture, a sound, a video, a row in database or CSV file, or whatever you can describe with a fixed set of quantitative traits.
- Feature vector
  - n-dimensional vector of numerical features that represent some object.
- Feature extraction
  - preparation of feature vector
  - transforms the data in the high-dimensional space to a space of fewer dimensions.
- Training/Evolution set
  - set of data to discover potentially predictive relationships.

PROCESS


The "real world"
Measuring devices
Preprocessing
Dimensionality reduction
Prediction
Model selection
Analysis results
Sensors
Cameras
Databases
Noise filtering
Feature extraction
Normalization
Feature selection
Feature projection
Classification
Regression
Clustering
Description
Cross-validation
Bootstrap

PROCESS IN SPAM FILTERING



- Number of recipients
- Size of message
- Number of attachments
- Number of "re's" in the subject line

…

DATA SET

Input Attributes

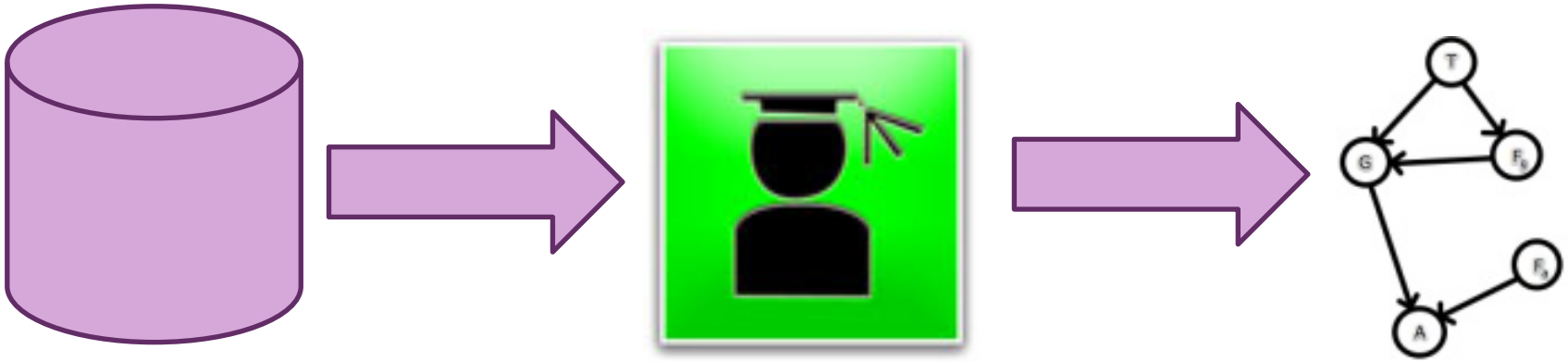
Target Attribute

Instances

Email Type	Customer Type	Country ((IP	Email (Length (K	Number of new Recipients
Ham	Gold	Germany	2	0
Ham	Silver	Germany	4	1
Spam	Bronze	Nigeria	2	5
Spam	Bronze	Russia	4	2
Ham	Bronze	Germany	4	3
Ham	Silver	USA	1	0
Spam	Silver	USA	2	4

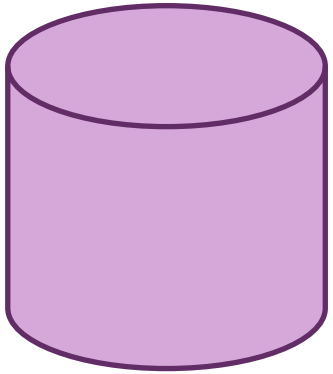
Numeric

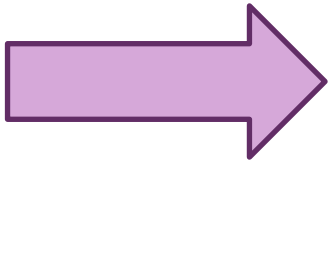
MODEL LEARNING

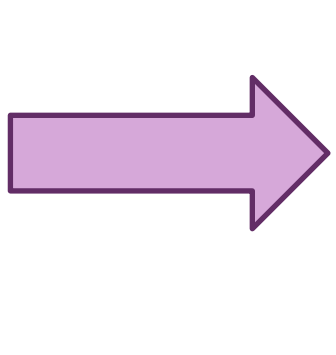


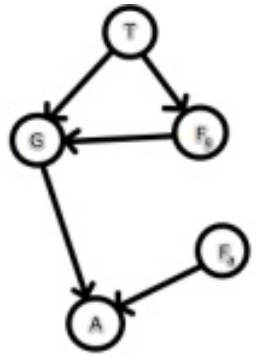
Learner
Inducer
Induction Algorithm

MODEL TESTING







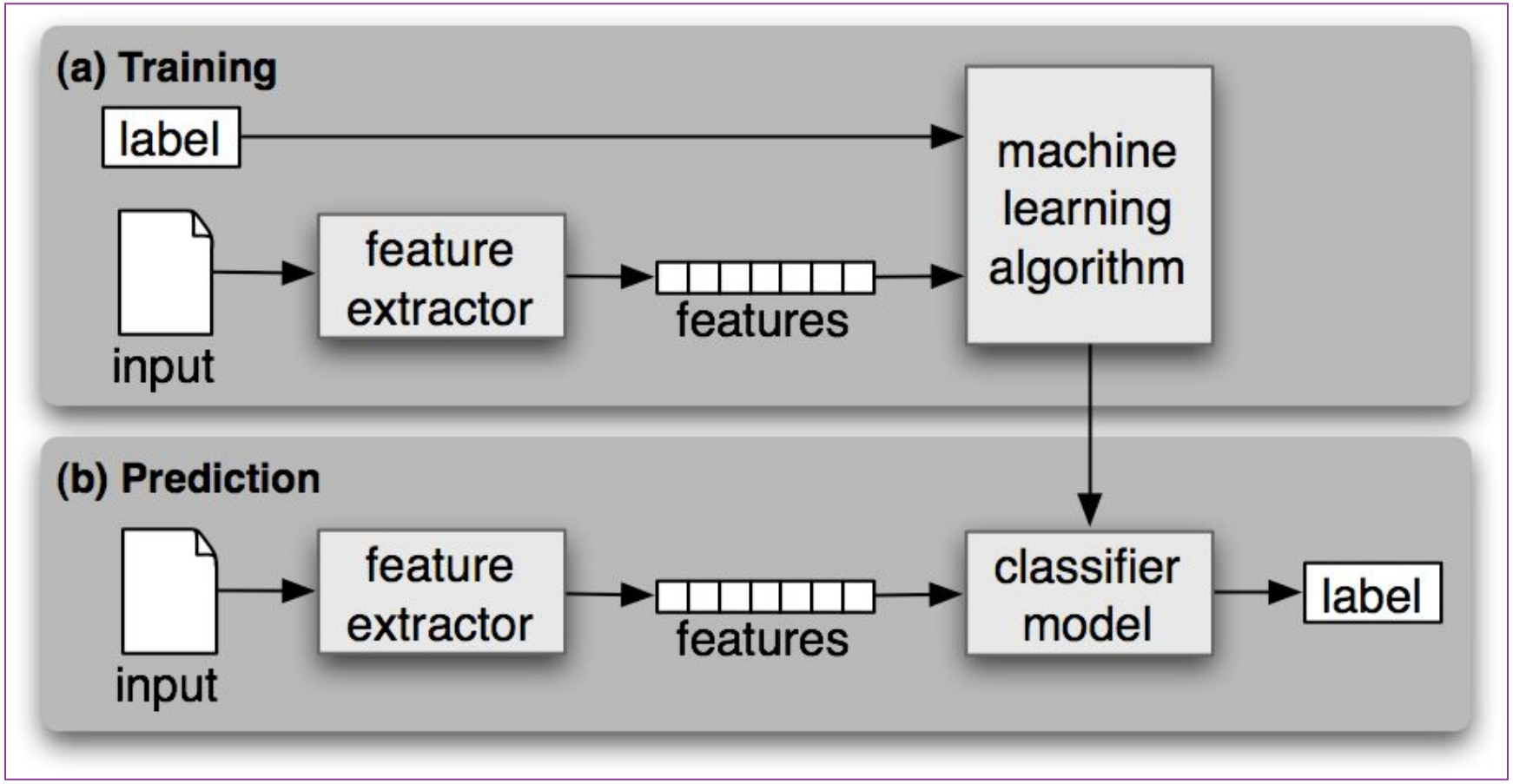


Database
Training Set

Learner
Inducer
Induction Algorithm

Classifier
Classification Model

WORKFLOW


(a) Training
label
input
feature extractor
features
machine learning algorithm
(b) Prediction
input
feature extractor
features
classifier model
label

CATEGORIES

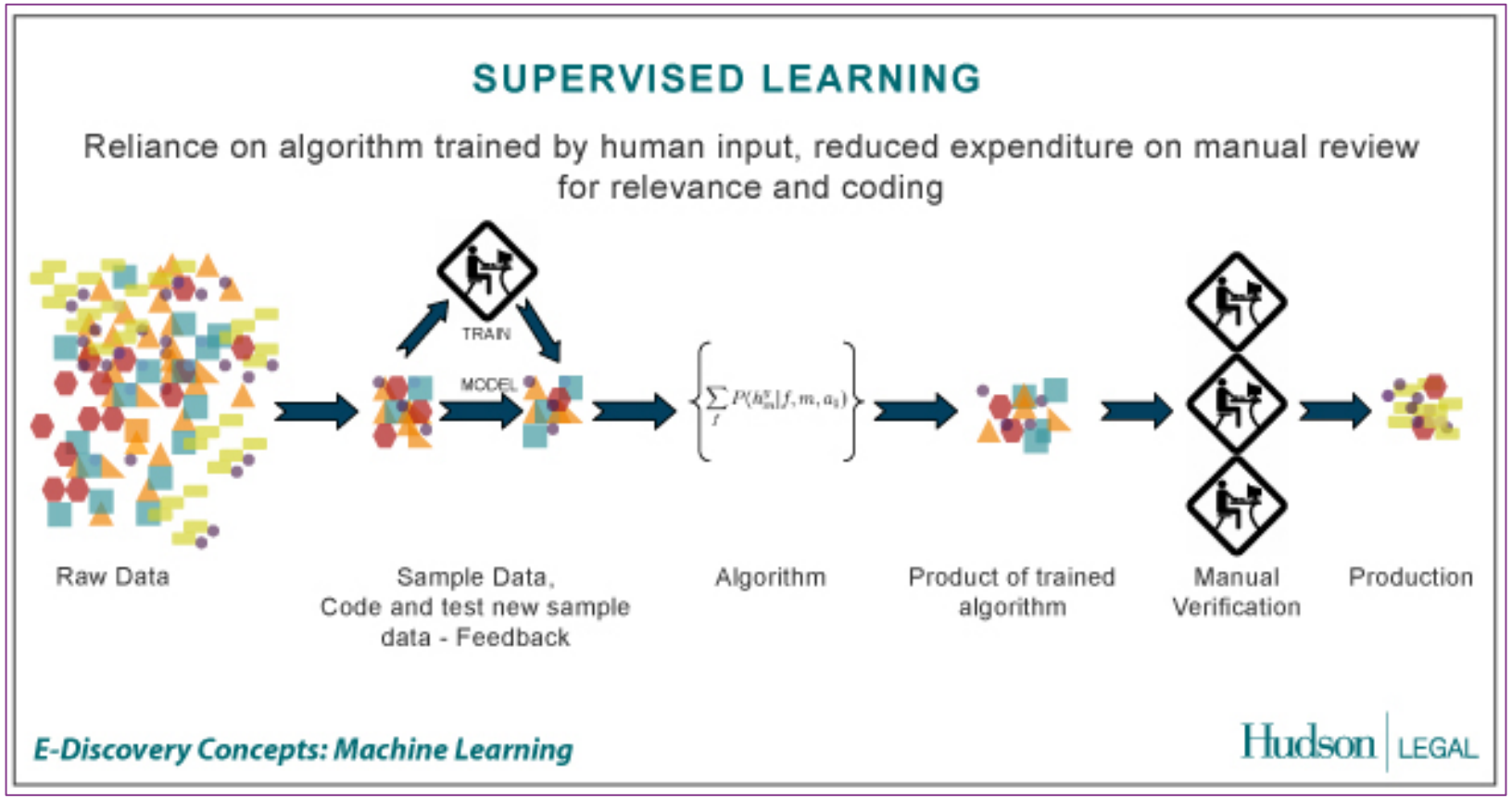
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Semi-Supervised Learning
- Bayesian learning

USE-CASES

- Spam Email Detection (Classification)
- Image Search (Similarity/Classification)
- Clustering (KMeans) : Amazon Recommendations
- Autonomous driving/flying : Reinforcement learning

CATEGORIES

SUPERVISED LEARNING (CLASSIFICATION)



SUPERVISED LEARNING EXAMPLES

- A Bank may have borrower details (age, income, gender, etc.) of the past (features)
- Also it may have details of the borrowers who defaulted in the past (labels)
- Based on the above, can train a classifier to learn the patterns of borrowers who are likely to default on their payments

SUPERVISED LEARNING

- Used when the dataset has classes/labels
- Includes a ‘training’ phase with the dataset and a ‘testing’ phase to validate the accuracy of the classifier
- Algorithms – Regression, Support Vector Machines, Neural Networks, Convolutional Neural Networks, Decision Trees, Logistic Regression, Random Forest, Naïve Bayesian, etc.

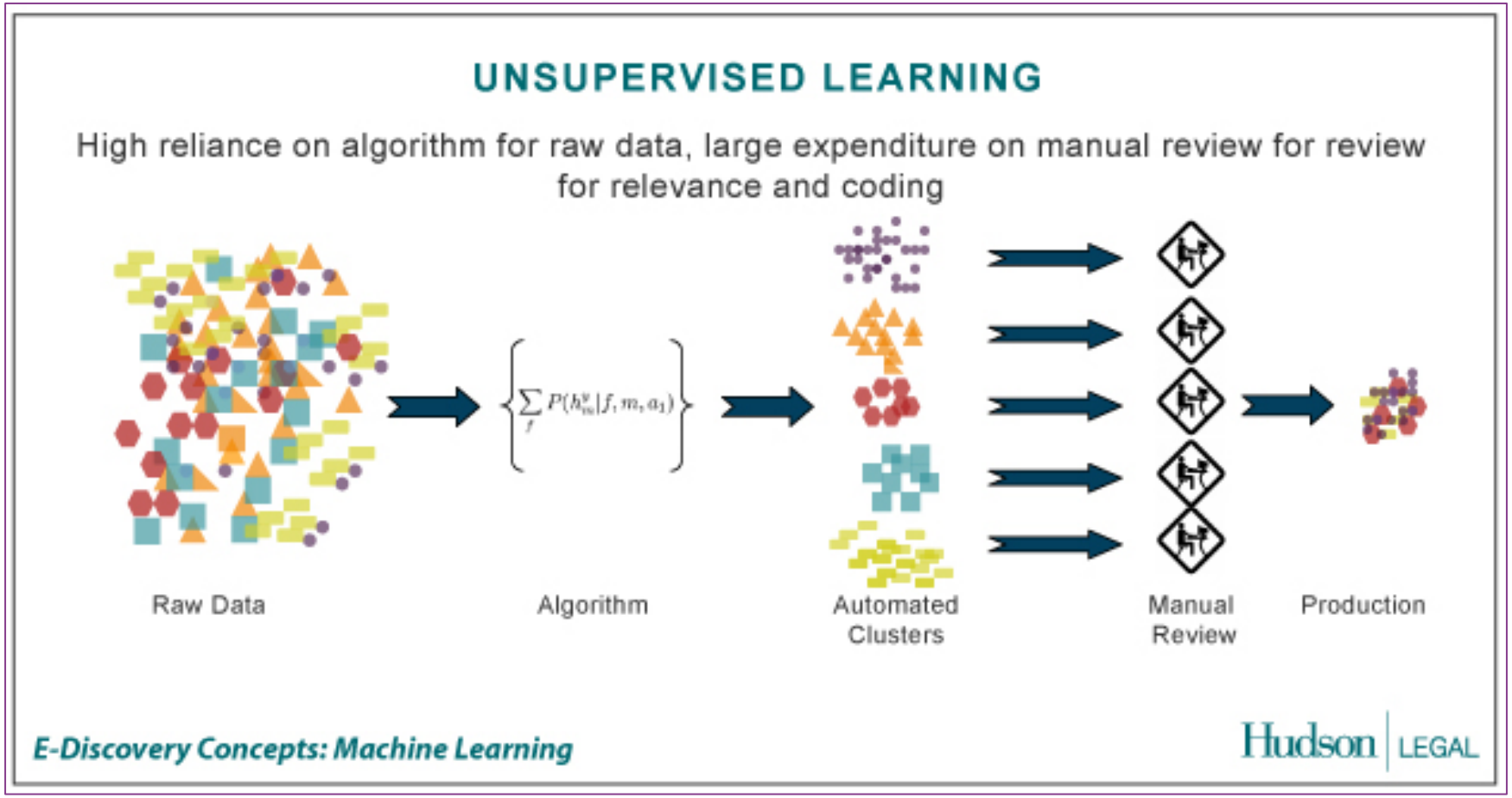
SUPERVISED LEARNING

- **Regression** – Predict continuous variables (salary, rent)
- **Binary classification** (facial recognition, whether a tumor is benign or malignant)
- **Multi-class classification** (the type of a vehicle, the stage of progression of a cancer – level 1,2,3)

UNSUPERVISED LEARNING

- Used when the dataset does not have the labels (classes)
- Used to group/cluster the data into clusters, which may then be used for decision making, making recommendations, classification, etc.
- Algorithms – K-means, Self Organizing Maps, Deep belief Networks, etc.

UNSUPERVISED LEARNING/CLUSTERING



UNSUPERVISED LEARNING EXAMPLES

- A Supermarket may store each buyer's basket content details (features)
- There are NO grouping (labels)
- Need to group the buyers based on their buying patterns in order to best use the shelf space (recommendation)

UNSUPERVISED LEARNING/CLUSTERING

- K-means clustering

- The algorithm will categorize the items into k groups of similarity. The Euclidean distance as the measurement will be used to calculate that similarity.

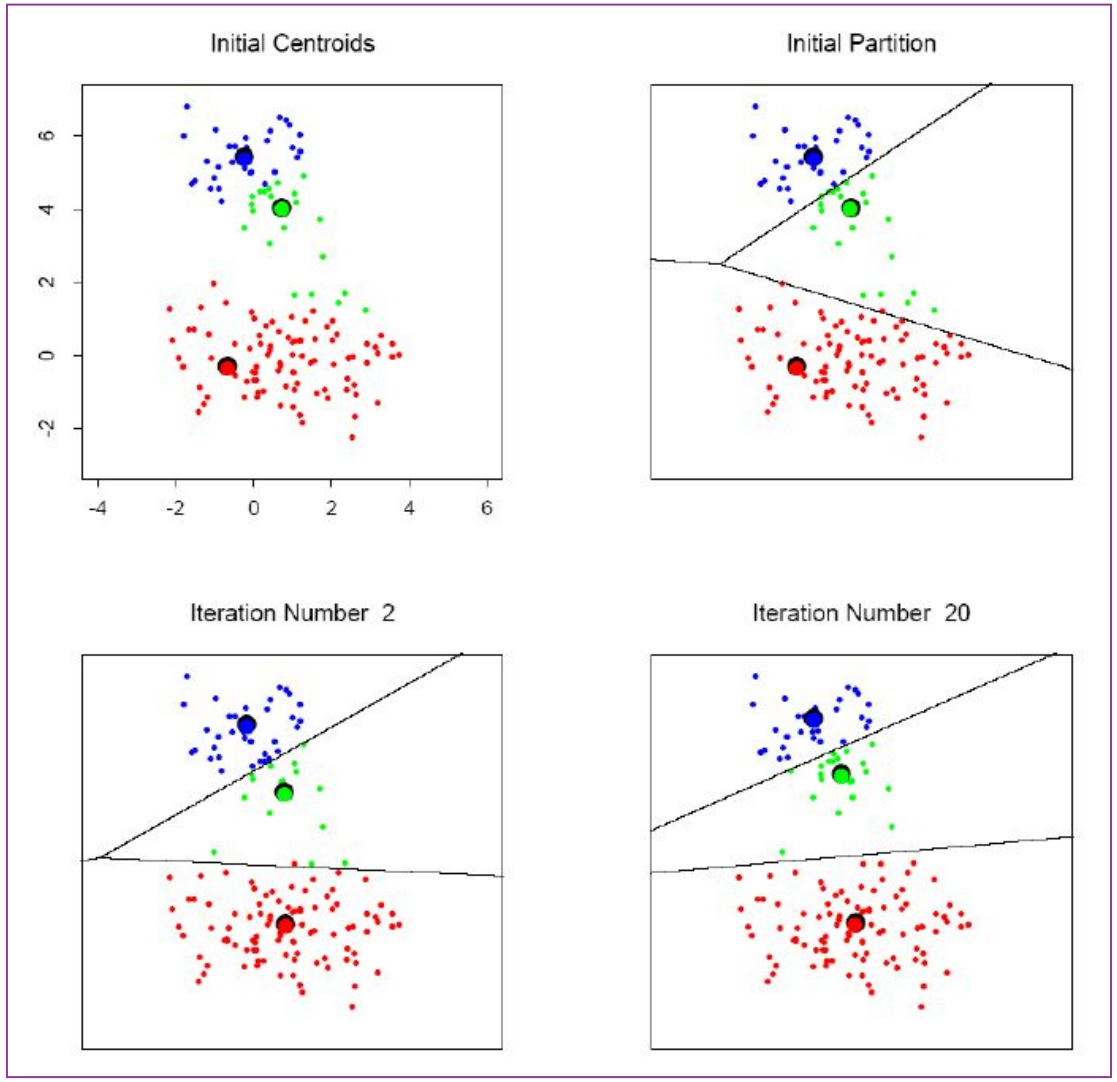
- Self organizing maps

- A type of artificial neural network (ANN) that is trained using unsupervised learning to produce a low-dimensional, discretized representation of the input space of the training samples, called a map, and is therefore a method to do dimensionality reduction.

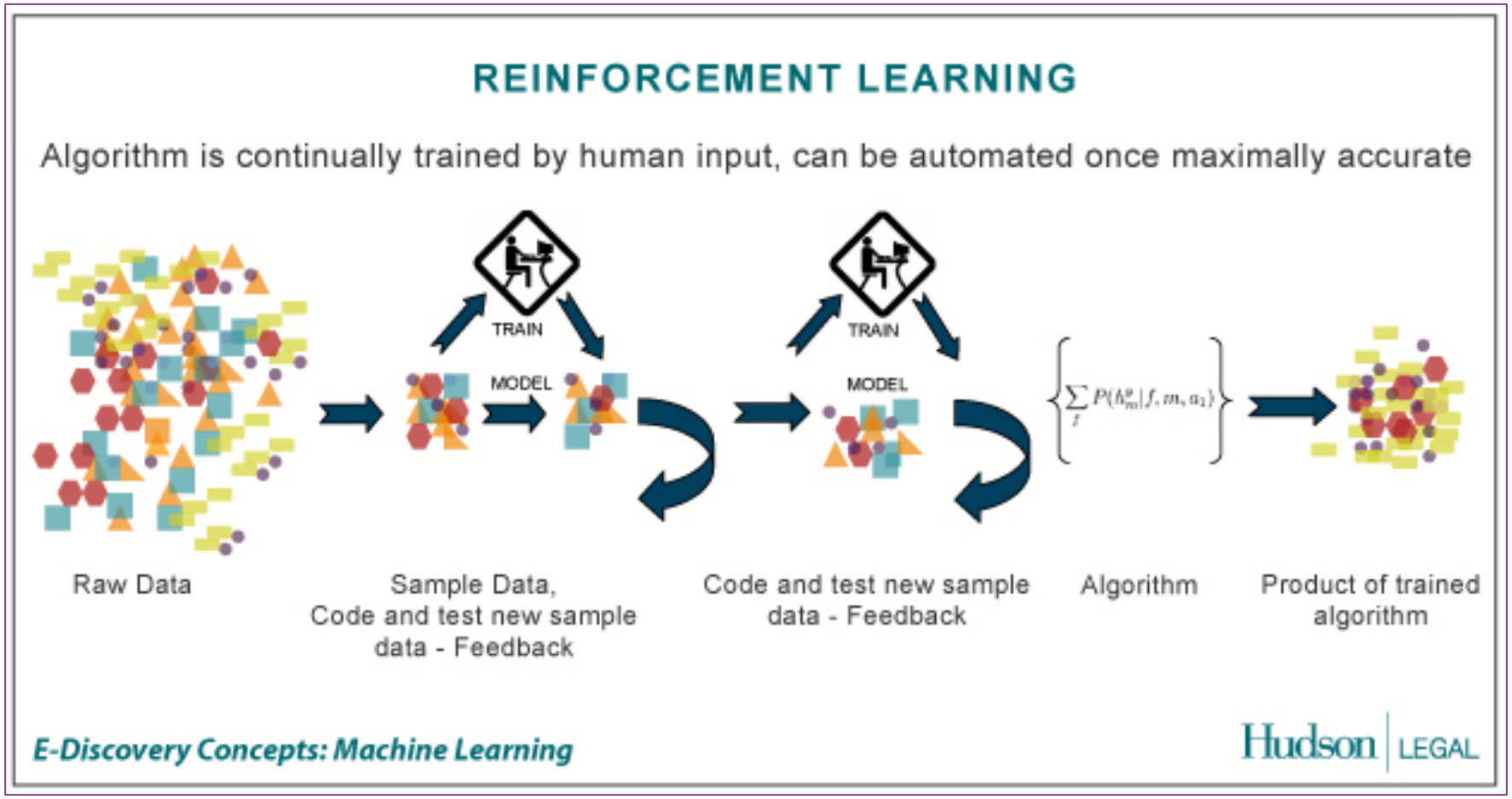
- Deep Belief Networks

- A graphical representation which are essentially generative in nature i.e. it produces all possible values which can be generated for the case at hand.

K-MEANS CLUSTERING EXAMPLE



REINFORCEMENT LEARNING



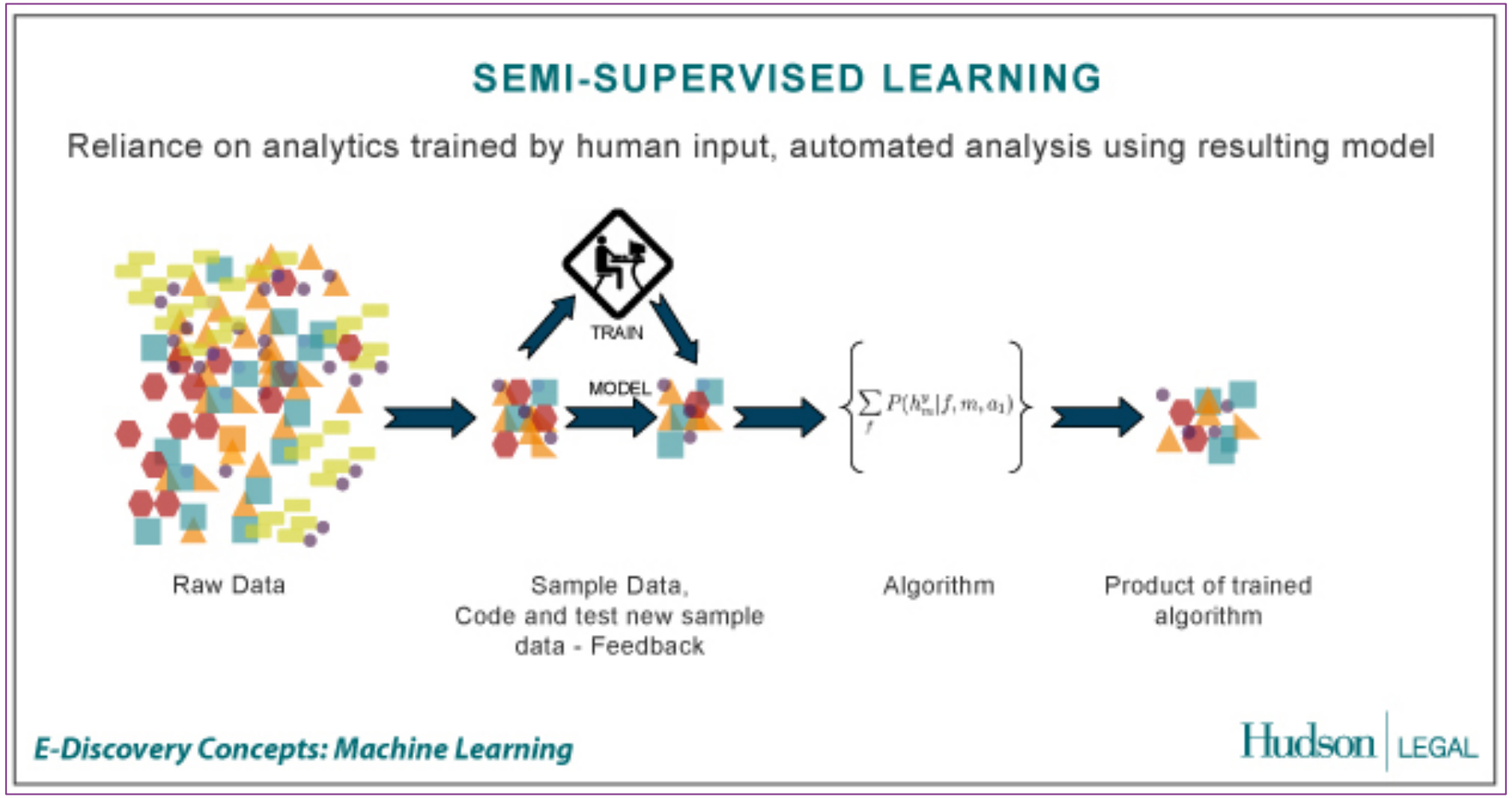
REINFORCEMENT LEARNING

- Can be used when there's no data available
- A reward function is used to measure the reward for a given action
- Based on the reward values, a probability distribution can be obtained for a given set of functions
- This can be continued over time and also can be deployed in both single/multi-agent systems
- Algorithms – Actor Critic learning, Q learning, Monte-carlo methods, etc.

REINFORCEMENT LEARNING EXAMPLES

- A group of robots have been deployed in an unknown territory
- The objective is for them to collaboratively find the navigation path to reach a particular destination/goal
- Can use reinforcement learning where achieving the goal/getting closer to the goal gives a positive reward. Negative reward otherwise
- Can share the information among robots (multi-agent system)

SEMI-SUPERVISED LEARNING



SEMI-SUPERVISED LEARNING

- Labeled data is expensive/difficult to get
- Unlabeled data is cheap/easier to get
- The idea is to use smaller amount of labelled data with larger amount of unlabeled data to creating the training/testing datasets
- Algorithms - Self Training, Generative models, Semi-Supervised Support Vector Machines, etc.

SEMI-SUPERVISED LEARNING APPLICATIONS

- Web page classification

- Limited amount of labeled data can at most train models of lower complexity well, the addition of unlabeled data makes the updated models of higher complexity much improved.

- Speech to text conversion

- Producing a large amount of annotated speech data for training ASR systems remains difficult for more than 95% of languages all over the world which are low-resourced.

- Video/image generation

- A Generative Adversarial Networks (GAN) with a classifying discriminator would be able to exploit both the unlabeled as well as the labeled data.

BAYESIAN LEARNING

- **Bayes Theorem:**

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- Naïve bayesian, Multinomial bayesian, Bayesian networks, Hidden markov model
- Applications: Sentiment analysis, medical diagnosis
- Needs some initial knowledge

MACHINE LEARNING ON BIG DATA

- Use large unstructured data sets for learning (Call records, Social media data, etc.)
- Two main approaches
- Use a Big Data Platform (e.g. Apache Hadoop, Apache Spark)
- Use a Cloud based Big Data Analytics platform (Amazon AWS Services, Microsoft Azure ML)
- GPUs to speed up the learning (particularly in Deep learning)

THINGS TO CONSIDER

- If there's an algorithmic way instead of ML, use it!!! (ML is messy)
- Refer the literature
- Try different ML algorithms
- Check the dataset against the usage/strength of each algorithm (e.g. RNNs, ARIMA is good in time-series predictions)
- Be mindful of 'external factors' (e.g. seasonal effects, RL if you don't have data, Clustering if you have unlabeled data, etc.)
- Test your algorithm(s) with test data and select the best performing one for production
- No algorithm will be perfect!

POPULAR FRAMEWORKS/TOOLS

- Scikit-learn - Python (Anaconda Python Distribution)
- R (R studio)
- MATLAB/Octave (can export DLLs)
- Weka (Java based)
- Java OpenNLP/Python NLTK (Natural language processing + ML)
- Apache Spark (part of the Apache Hadoop platform)
- Google TensorFlow (Python library for Deep neural networks)
- Apache Keras (Python library of neural networks)
- Theano (Python library for Multicore processing of DNNs)
- Amazon AWS Services/Microsoft Azure ML (Cloud based ML)

COMMONLY USED PYTHON LIBRARIES

- NumPy
  - Matrix algebra
- Pandas
  - Data Frames, Series
- Matplotlib
  - Visualization

RESOURCES

- Coursera – Andrew Ng. Machine Learning
- Udacity – Introduction to Machine Learning, Reinforcement Learning
- Python Machine Learning – Sebastian Raschka
- Advance Machine Learning with Python – John Hearty
- Machine Learning – Tom Mitchell

QUESTIONS

HANDS ON SESSION