Application of Machine Learning in Industries

CSAI3006

School of Computer Science
University of Petroleum and Energy Studies
Bidholi, Dehradun, Uttarakhand - 248007

ABOUT ME: DR. MOHAMMAD AHSAN

- PhD – National Institute of Technology Hamirpur (H.P.)
- M.Tech – National Institute of Technology Hamirpur (H.P.)
- Qualified UGC NET June-2015 and UGC NET Nov-2017 for Assistant Professor
- Qualified GATE 2012, GATE 2013 and GATE 2021.
- Before joining UPES, worked as Faculty in the CSE Department at NIT Andhra Pradesh and NIT Hamirpur.

COURSE OUTCOMES

- Understand the paradigm shift from classical AI to Modern AI and their applications in Industries.
- Understand need of data pre-processing and the importance of identifying features in data.
- Apply ML/Deep learning techniques in various sectors like banking, insurance, education, petroleum, healthcare, transportation etc.
- Understand some of the applications of NLP like Machine Translation, Chatbot etc. and their importance in various industries.

COURSE CONTENT

- **UNIT 1: Machine Learning in Banking and Securities**
  - Why machine learning in banking sector, Use of AI in banking and finance, Fraud detection, Tough competition in banking industry, Risk modeling and investment banks, Customer data management, Decreased customer experience and loyalty, Personalized marketing, Role of machine learning: Challenges of banking sector and securities, Widely used machine learning algorithm in banking and security, Fraud prevention and detection systems, Rule based and machine learning based approach in fraud detection, Anomaly detection: Ways to expose suspicious transactions in banks, Advanced fraud detection systems, Risk management systems, Case study: Application of machine learning for financial risk management, Credit risk analysis using machine learning classifier, Investment prediction systems, Portfolio management systems, Objectives of portfolio management, Algorithmic trading, Deep learning for customer services, Chatbot: Deep learning approach, AI powered marketing systems, Deep learning in cyber security, Types of cyber-attacks in banks, Deep learning methods used in cyber security, Deep learning v/s restricted Boltzmann machines, Convolution Neural Networks (CNNs), Recurrent neural networks, Machine learning techniques: Loan underwriting & sentiment/news analysis, Sentiment or news analysis, Current challenges and opportunities: Banking and security domain.

COURSE CONTENT

- **UNIT 2: Machine Learning in Communication, Media and Entertainment**
  - Machine learning in communication, media and entertainment, Usage of machine learning in media and entertainment industry, Machine learning techniques for customer sentiment analysis, World embedding's, Sentiment analysis with long short term memory networks, Real-time analytics in communication, media and entertainment industries, Real time analytics and social media, Deep learning for social media analytics, Recommendations engines, Collaborative filtering, Memory based collaborative filtering, Model based collaborative filtering, Content based filtering, Hybrid recommendation systems, Summary of recommendation systems, Deep learning techniques on recommender systems
- **UNIT3:** Machine Learning in Healthcare and Life Sciences.
- **UNIT4:** Machines Learning in Education
- **UNIT5:** Machine Learning in Manufacturing and Petroleum Industries

COURSE CONTENT

- **UNIT 6:** Applications of Machine Learning in Government Administration
- **UNIT 7:** Machine Learning in Insurance Industry
- **UNIT 8:** Applications of Machine Learning in Retail Industry and Supply Chain
- **UNIT 9:** Machine Learning in Transportation and Logistics
- **UNIT 10:** Machine Learning in Energy and Utilities

- **Text Books:**
  - Application of machine learning in Industries (IBM ICE Publication).

Welcome to:

Machines Learning in Banking and Securities

Unit objectives

After completing this unit, you should be able to:

- Learn about machine learning in banking sectors, challenges in banking sector and fraud detection system in banking sector
- Understand manage customer data and algorithms in banking and security
- Gain knowledge on deep learning technology for personalized marketing
- Learn about applications of machine learning classifiers in credit risk analysis
- Understand the importance of machine learning in portfolio management systems
- Learn about algorithmic trading and stages for implementing applications of ai in marketing
- Deep learning approach for sentiment analysis of customers and customer services
- Understand cyber security in banking sector and loan underwriting and sentiment/news analysis
- Understand current challenges and opportunities in implementing machine learning technologies in banking and security domain

Why machine learning in banking sector

- After financial crisis around the world, there has been a significant transformation in the banking services across the globe with intervention of Artificial Intelligence (AI) applications.

- With appropriate and robust machine learning algorithms, there is a lot of potential to address the key issues in the banking sector, leading to digital transformation and enhanced services.

- Major areas of banking with potential AI intervention include Anti Money Laundering(AML), Chabot's, fraud detection, algorithmic trading, and digitization.

Use of AI in banking and finance

- Efficient resource management, improvement in performance of operations and smart decision making are few of the promising advantages of machine learning technologies in the banking and finance sector.

- Fraud detection, automated customer support and services, banking security are few of the important use cases of ML in banking sector.

Challenges in banking sectors and securities

- Fraud detection.
- Though competitions.
- Risk modelling.
- Customer data management.
- Decreased customer experience and loyalty.
- Personalized marketing.

Fraud detection (1 of 2)

- One of the foremost requirements in banking sector is to provide enough security to the customers and employees in terms of operations and financial transactions.

- Ml techniques can help banks to detect fraud and can restrict account activities based on authentication and validation.

- Fraud detection includes the following initial steps:
  - Preliminary data testing and data sampling for model estimation.
  - Creating/estimating the appropriate model.
  - Deployment of the model and testing.

Fraud detection (2 of 2)

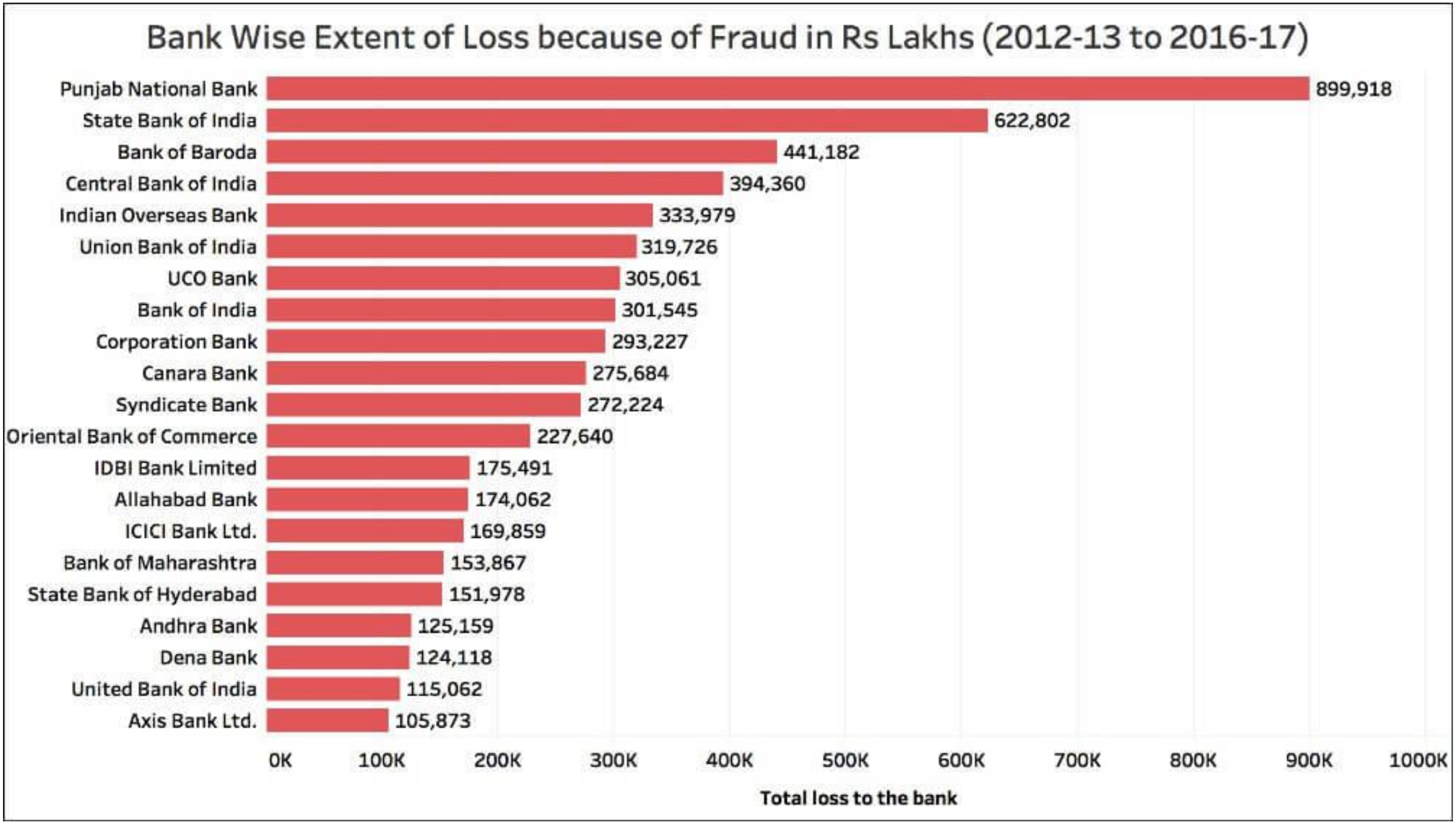


Figure: Fraud detection

Source: Factly

Tough competition in banking industry

- Indian banking services has evolved with many “pro-customer” applications leading to competition among major retail banks.

- User friendly online banking applications like Paytm, Phone pay, Google pay are redefining the banking services, especially in terms of its geographically limitless reachability.

- Personalized banking experience is another most important aspect of the afore mentioned applications.

Risk modeling and investment banks

- The foremost priorities for any investment banks is managing financial risks.

- With effective risk modelling, it becomes easier for banks to evaluate and strengthen the capabilities of companies in creating financial capital, sustain or restructure corporate operations, facilitate acquisitions and mergers whenever appropriate.

Customer data management

- Financial organizations have a challenging task of documenting and monitoring humongous amount of customer data.
- It is essential to establish cross data set linkages to develop unique insight of customer needs.
- Data science can help banks extract relevant information about their customer behavior, whereas appropriate ML algorithms can keep track of customer preferences.
- One such example for customer data management is “segmenting the customers”. It refers to clustering of customers based on their unique behavior.

Decreased customer experience and loyalty

- Ever increasing customer focused services provided by companies like phone pay, google pay etc., is redefining the customer experience.

- Now a day, customers expect totally hustle free and easy banking services and it is more likely that banks can lose customers if they can not provide quality services to its customers.

- Banks can make use of efficient emotion recognition algorithms to deeply analyze the customer feelings and sentiments.

Personalized marketing

- The key aspect that leads to success in marketing is, the extent to which customers receives offers that best suits their needs and preferences.

- In the banking sector, the financial products should be highly customized to the customers' needs and in addition, secured as well.

- As the customers are exposed to such personized banking experiences, there will be ever increasing expectation for improvement in the quality of services.

Role of machine learning: Challenges of banking sector and securities

Domain	Implications of ML based solutions
Fraud and Risk management	ML based solutions and predictive analytics are assisting in examination of real-time transactions to identify suspicious and fraudulent operations. Risk analysis experts are being guided by ML algorithms with appropriate recommendations to predict risk in the earlier stages of any banking operations.
Customer Services	With existing customer data along with ML powered AI applications are leading to effective personalized customer services by documenting and analyzing customer behavior and requirements. ML algorithms based cognitive machines are replacing humans in analyzing and responding to customer queries.
Financial trading and securities	ML based validation mechanisms are bridging the security and functional gap between front end trades and back end operations. AI based applications are assisting banks in effectively handling foreign exchange transactions and liquidity management operations.
Credit assessment	ML based applications along with big data analytics are viable solution to assess the credit worthiness of the customer in case of loan disbursement operations.
Portfolio management	AI and ML based technological ecosystems are helping banks in making real time, smarter decisions to ensure appropriate investment plans for their customers.

Widely used machine learning algorithm in banking and security

- Supervised machine learning algorithms.
- Unsupervised machine learning algorithms.
- Reinforcement algorithms.

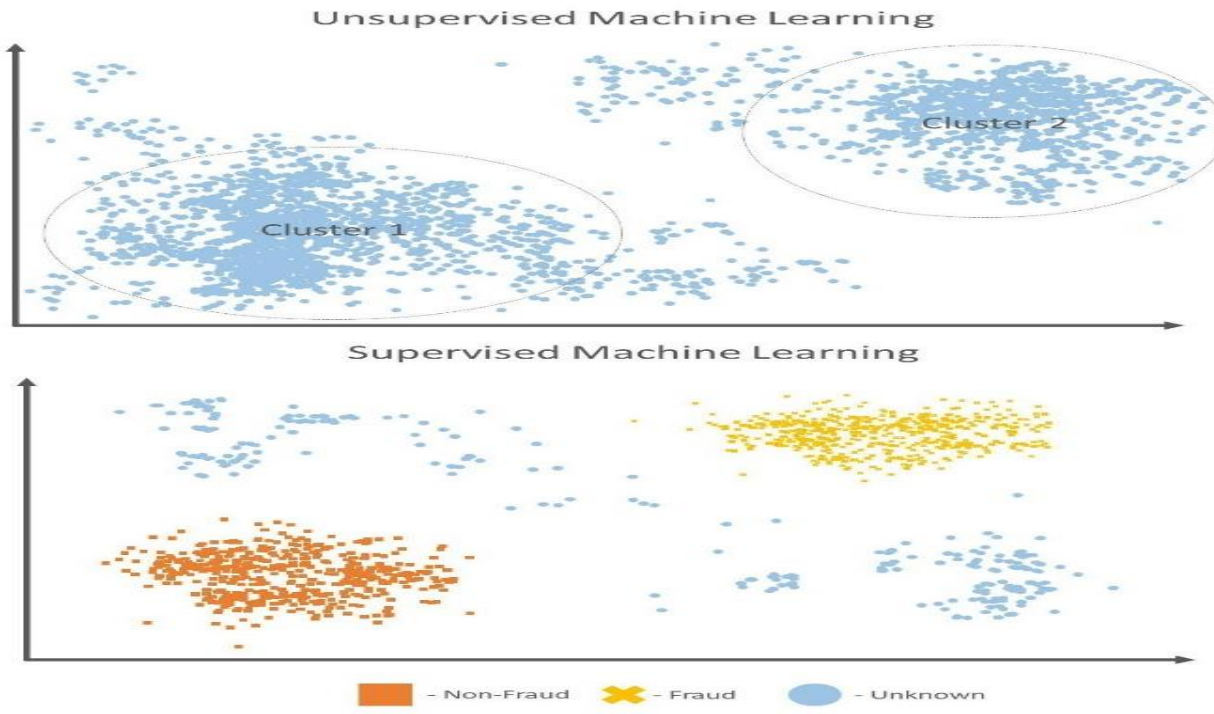


Figure: Machine learning algorithms in Banking and security

Source: Altexsoft

Fraud prevention and detection systems

- The most used algorithms in fraud prevention and detection systems are:
  - Bayesian algorithms.
  - K-Nearest neighbor.
  - Support Vector machines (SVM).
  - Bagging ensemble classifier based on decision tree.

Rule based and machine learning based approach in fraud detection

Rule-Based fraud detection	ML-Based fraud detection
Catching obvious fraudulent scenarios	Finding hidden and implicit correlations in data
Requires much manual work to enumerate all possible detection rules	Possible fraud scenario detection happens automatically
Multiple verification steps may become threat to user experience.	Reduction in the number of verification measures
Long term processing	Real time processing

Anomaly detection: Ways to expose suspicious transactions in banks

- The main feature of this approach is classification of all data objects into two major groups:
  - Normal distribution.
  - Outliers.

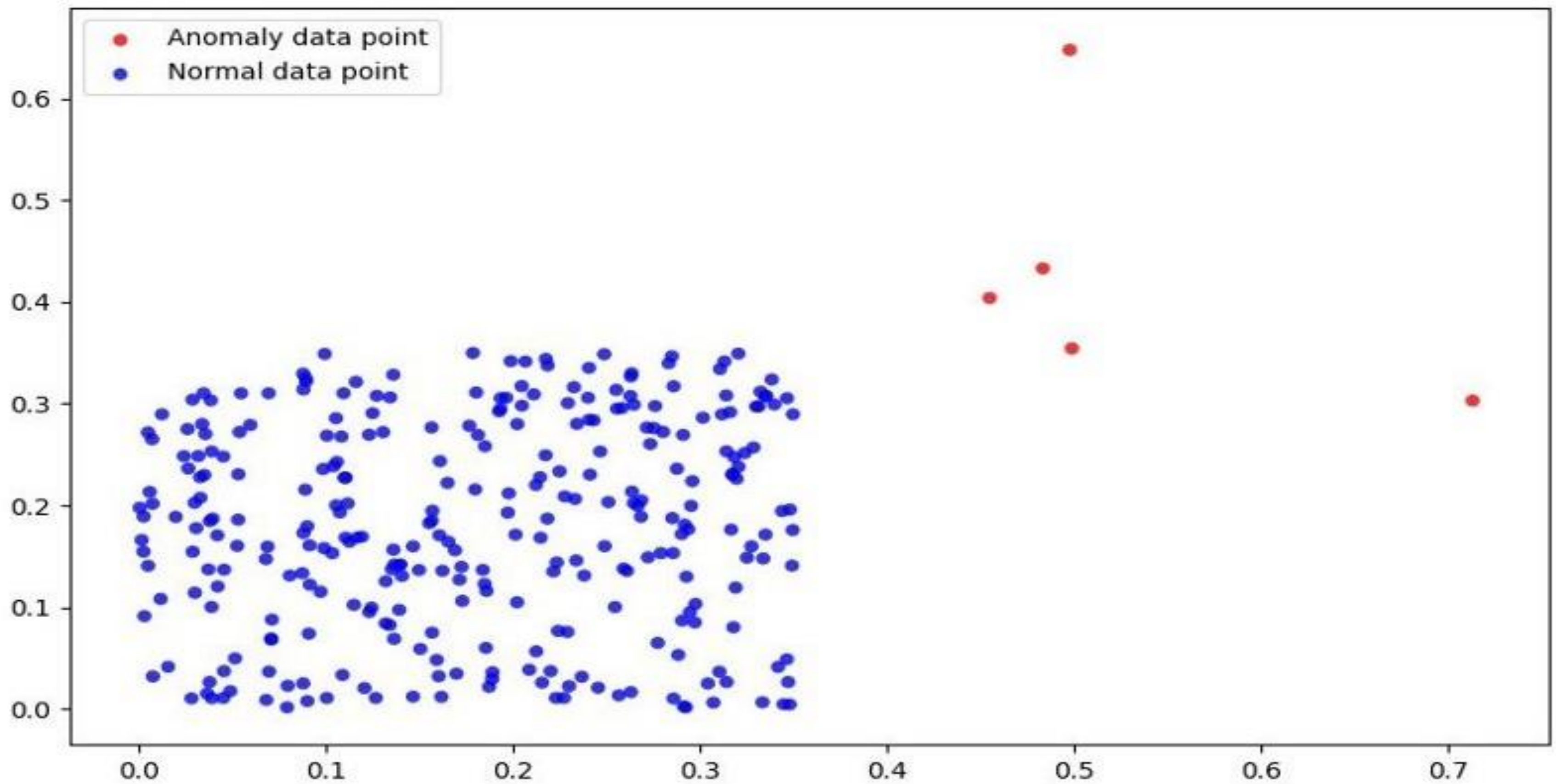


Figure: Ways to expose suspicious transactions in banks

Source: https://zindi.africa/blog/introduction-to-anomaly-detection-using-machine-learning-with-a-case-study

Advanced fraud detection systems

- One of the major limitations of the primitive fraud detection approaches is that they are limited to identification of anomalies.

- The two commonly used ML techniques to develop anti-fraud mechanisms are:
  - Supervised ML algorithms.
  - Unsupervised ML algorithms.
  - Deep neural network algorithms.

Risk management systems

IBM ICE (Innovation Centre for Education)

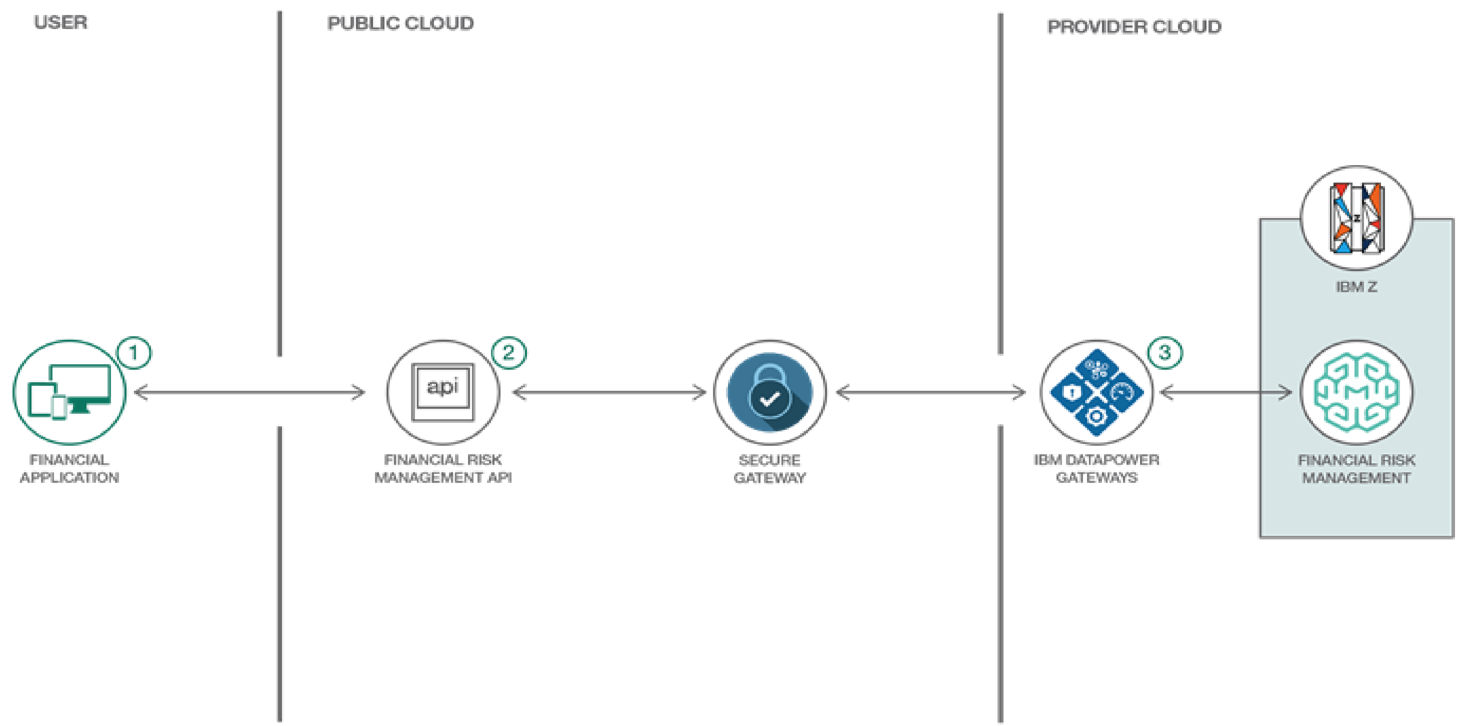
SL NO	Objectives of Portfolio Management	Key Features/ Objectives
1.	Security of Principal Investment	• To ensure safety investment and risk minimization • Along with keeping the investment unharmful, portfolio management also contribute in increase of its purchasing power over a period. • Only after ensuring safe investment, other factors like growth, income etc., are considered
2.	Consistency of returns	• In order to ensure stable return on investment by reinvesting the earned profit in good portfolios. • It helps the customers/ firms to obtain consistent returns.
3.	Capital growth	• Portfolio management helps the firms by increase in the rate of return by reinvesting securely or purchasing the profitable growth securities. • It takes care of any loss in purchasing power because of other economic factors, in order to appreciate values and to safeguard investors.
4.	Marketability	• In order to ensure flexibility in investment portfolio. • These portfolios consist of investment which concentrates only on the portfolios that can be marketed and traded. • Shifting from one investment to other will be a problem if the portfolio contains large number of inactive and unlisted shares. • It suggests investing only in secured stock exchanges.
5.	Liquidity	• It facilitates investors to make best use of available opportunities in future. • Based on the investor's requirement, it always ensure a enough fund availability in short notice
6.	Diversification of Portfolio	• To effectively manage the financial investment made by customer it is necessary to diversify the investment by considering benefits available across the industries.
7.	Favorable Tax status	• In order to minimize the tax burden and to increase the return in investors fund, the portfolio should assess by considering the applicable taxes.

Case study: Application of machine learning to financial risk management

- Summary: Machine learning affects all industry fields, like how banking companies and other sectors tackle stricter enforcement and threat mitigation criteria.

- This programmer path teaches you what to implement machine learning for a strategic threat system on IBM z/OS to assess the value of consumer debt. utilizing an API, you can know how to view the information, allowing you to integrate the information into company programs.

- Flow :



Credit risk analysis using machine learning classifier

- Promising business activity in banking industry is providing loan to the customers.

- Crediting risk analysis is evolving in the field of financial risk management.

- Based on the available database of customers many techniques were used in credit risk analysis in order to evaluate future credit risk that may occur.

- ML techniques can be used to analyze and evaluate the credit risk datasets.
  - Bayesian classifier.
  - Naive–Bayes classifier.
  - K-nearest neighbors.
  - K-means.
  - Multilayer perceptron.
  - Support vector machine.

Investment prediction systems

- The wealth management organizations are evolving with the potential AI based solutions for their investment decisions based on the historical data.

- The applications of collection of huge data about the assets are being recorded for digital assets or distributed industrial asset by making the assets ready for digitalization through AI.

Portfolio management systems

- Portfolio: It can be defined as financial assets like stocks, bonds, shares, mutual funds, cash equivalents, debt instruments etc., In order to maintain the risk in various asset pools of investment, a portfolio is planned.

- Management: In order to achieve its pre described objectives with well-defined policies, a management can be defined as a firm which coordinates the activities of that firm/enterprise.

No.	Investor's Portfolio	Investment	Percentage	Security	Returns
1.	Government Bonds	$ 25,000	25 %	High	Low
2.	Bank's Fixed Deposits	$ 15,000	15 %	High	Average
3.	Shares	$ 35,000	35 %	Low	High
4.	Mutual Funds	$ 25,000	25 %	Average	Average

- Note: This is just an example and may be taken as a slandered for the portfolio management.

Objectives of portfolio management

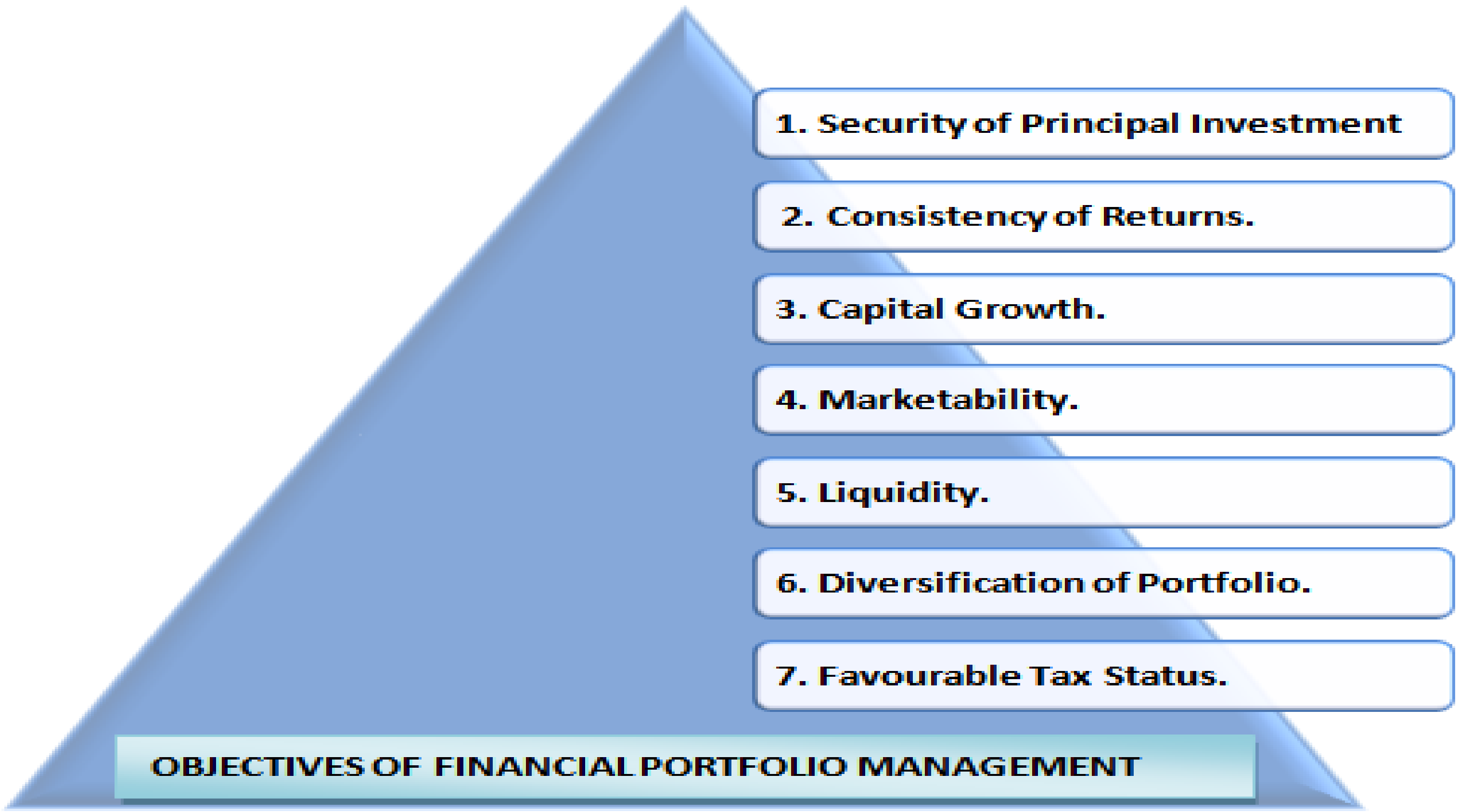


Figure: Objectives of portfolio management

Source: Image credits © moon rodriguez

Algorithmic trading

- Algorithm trading enables to produce profit at a very high intensity which is difficult by individual buyers.

- Algebraic equation is an example of algorithm, with prescribed rules of algebra.

- Complex formulas along with mathematical models and human inaccuracy will be used by algorithmic model in order to sell or buy the financial securities.

- The use of high-frequency algorithmic trading technology, enables algorithmic traders to make 10000 trades in fraction of time.

Deep learning for customer services

- Customer service center, are ready to adopt technologies like ML for their operations and these techniques is part of industries in upcoming days.

- In order to provide efficient customer service, it is always more flexible if we gather the data from their insights.

- Effective machine learning systems are implemented in environments where a lot of information is stored, as is important when the ultimate objective is to make an informed judgment.

Chatbot-deep learning approach

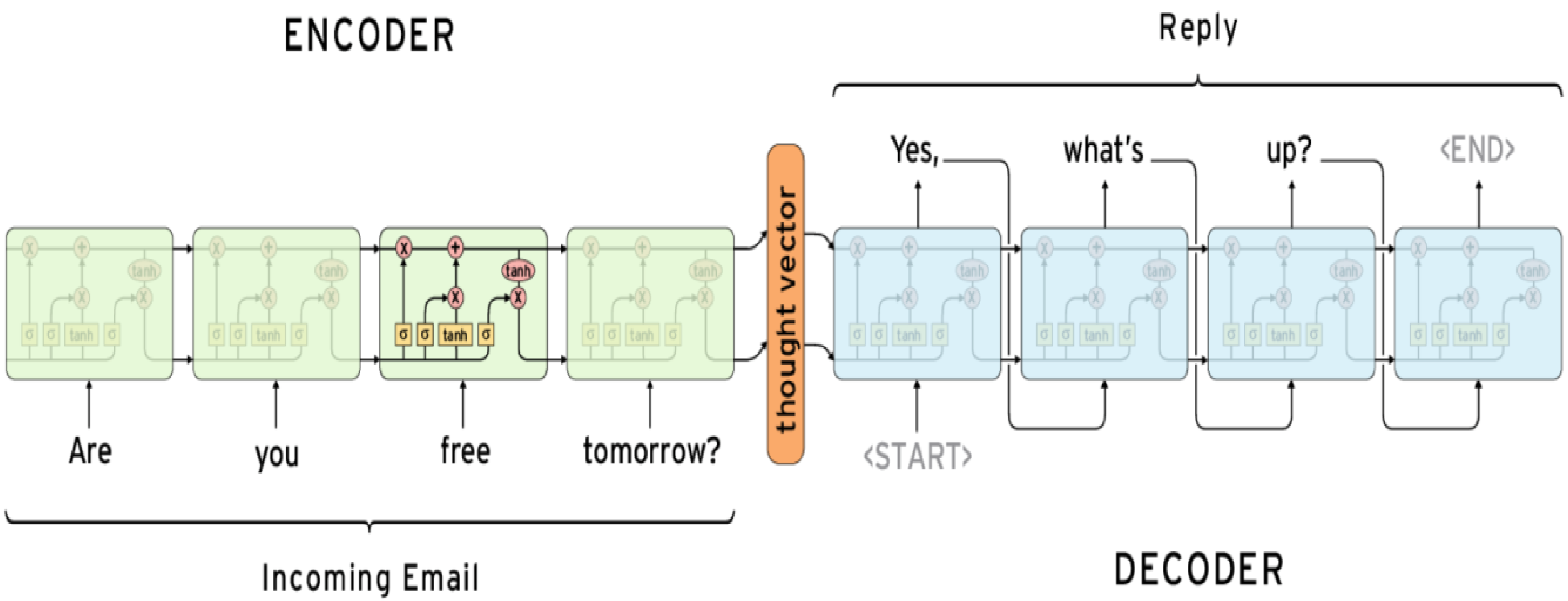


Figure: Chatbot-deep learning approach

Source: https://research.googleblog.com/2015/11/computer-respond-to-this-email.html

AI powered marketing systems

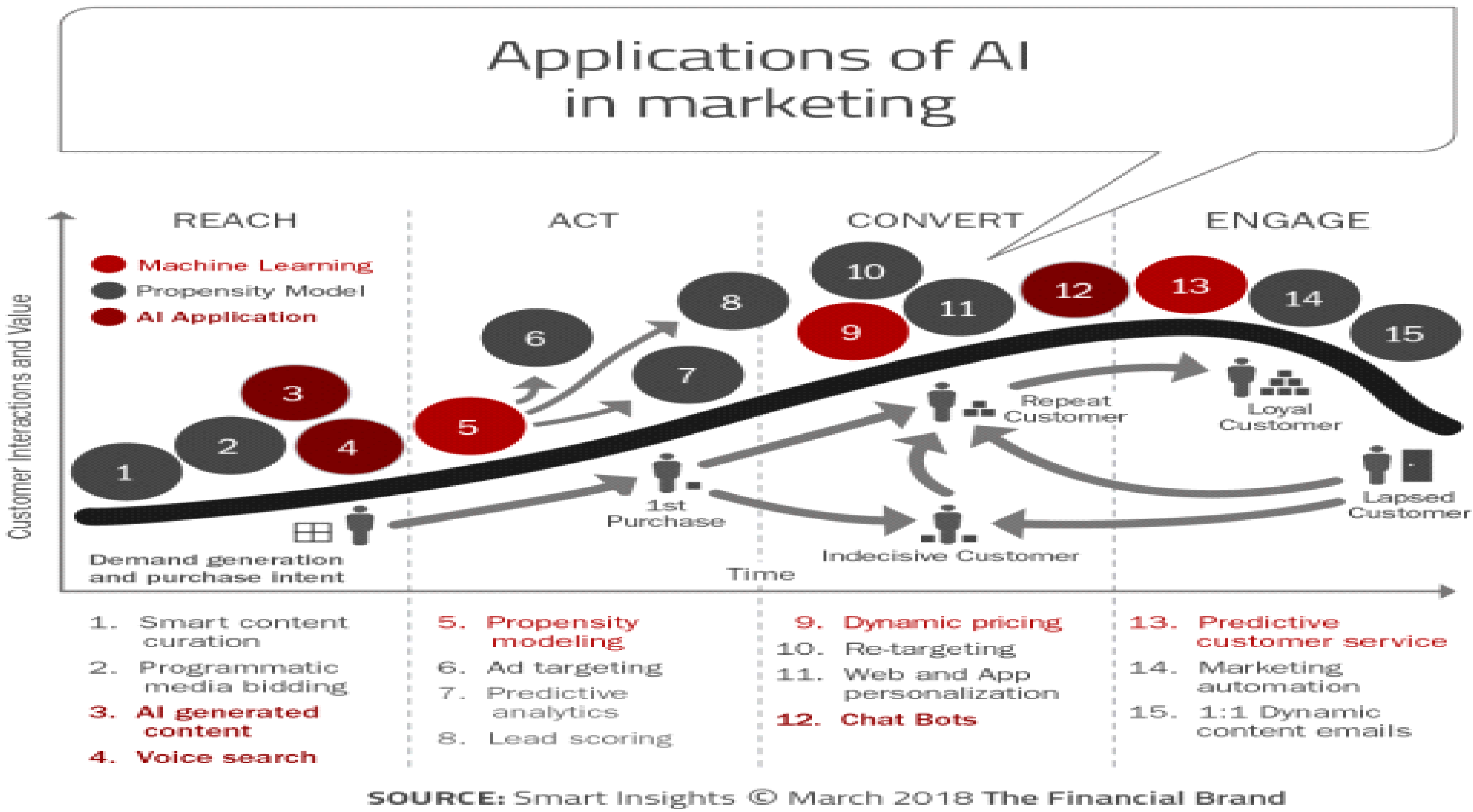


Figure: Stages for implementing applications of AI in marketing

Source: Smart insights © March 2018 the financial brand

Deep learning in cyber security

- Today cyber security uses smart technology such as machine learning and natural language analysis that can enable security analysts make smarter, quicker judgments.

- For the hackers banking sector is the hub where they can grab those important information of the customer.

- Security and its team are facing challenges to handle, interpret and prevent these type of mischievous events.

- Having a robust cyber security is an important requirement to maintain strong customer trust and credibility.

Types of cyber–attacks in banks

Denial of Services (DoS)	Phishing	Malware	Watering Hole	Zero-day exploits
Denial-of-service (DoS) attacks inundate systems with traffic to consume resources and bandwidth and make them unable to perform.	Phishing typically uses email that appears to be from a trusted or reputable source. Unsuspecting users open the email and take further actions like providing protected information or downloading malware.	Malware is malicious software. It's the chief weapon of a cyber-attack and includes viruses, worms, Trojans, ransomware, adware, spyware bots, bugs and rootkits. It installs when a user clicks a link or takes an action. When inside, malware can block access to data and programs, steal information and make systems inoperable.	In recent years' financial companies are most affected by watering hole, it is the most implemented cyber-attack. It mainly effects the IT system through online searches to figure out the behavioral pattern of employees to identify the website they frequently visit.	Zero-day exploits introduce malware through vulnerabilities unknown to the maker or user of software or systems. It is "zero-day" because developers have had zero-time to address or patch the vulnerability.[2]

Deep learning methods used in cyber security

- Deep learning is a machine learning approach that integrates natural systems in various levels to incrementally educate from information.

- The improvements in the deep learning technologies has increased the possibilities of utilizing machine learning approaches to address the problems in various domains.

- Deep learning has been applied towards number of use cases related to cyber security like identifying, malware detection, malware classification, android malware detection, phishing and spam detection.

- List of deep learning methods used in cyber security:
  – Deep belief networks.
  – Convolution neural networks (CNN).
  – Restricted Boltzmann machine.
  – Recurrent neural networks.

Deep learning v/s restricted Boltzmann machines

Deep Auto encoder	Restricted Boltzmann Machines
It is a type of unmonitored neural network that takes a vector reference and tries to fit the response to the same variable.	It is a two-layer, bipartite, undirected graphical models that from the building blocks of DBNs.
These are flexible due to their controlled learning of condensed data encryption.	RBM’s are unsupervised as Deep auto encoders and can be trained one layer at a moment.
These will reduce the computational resources to build an effective model by training one layer at a time.	In this type of network there is no interlayer connections but every nodes are fully connected as shown in Figure 4.
When the hidden layer has lower dimensionality than the feedback and production layer, the network is used to encrypt the information.	Restricted Boltzmann machines are deterministic, i.e. rather than definite values they give possibilities.
To gradually compress the information multi layers of auto encoders can be trained in series this is called stacked auto encoder	The design is conditioned by taking and supplying binary source information through the cycle. Then, the reconstruction of the source information is fed back through the design. The program's power will then be measured also utilized to adjust the scales. This method continues until the convergence of the system.
Sparse auto encoder comprises in the feedback and production surface of more secret points than there where only the secret level part is enabled at a given period.	Deep neural network can be created by stacking auto-encoders and RBMs, these are referred as stacked RBMs.

Convolution Neural Networks (CNNs)

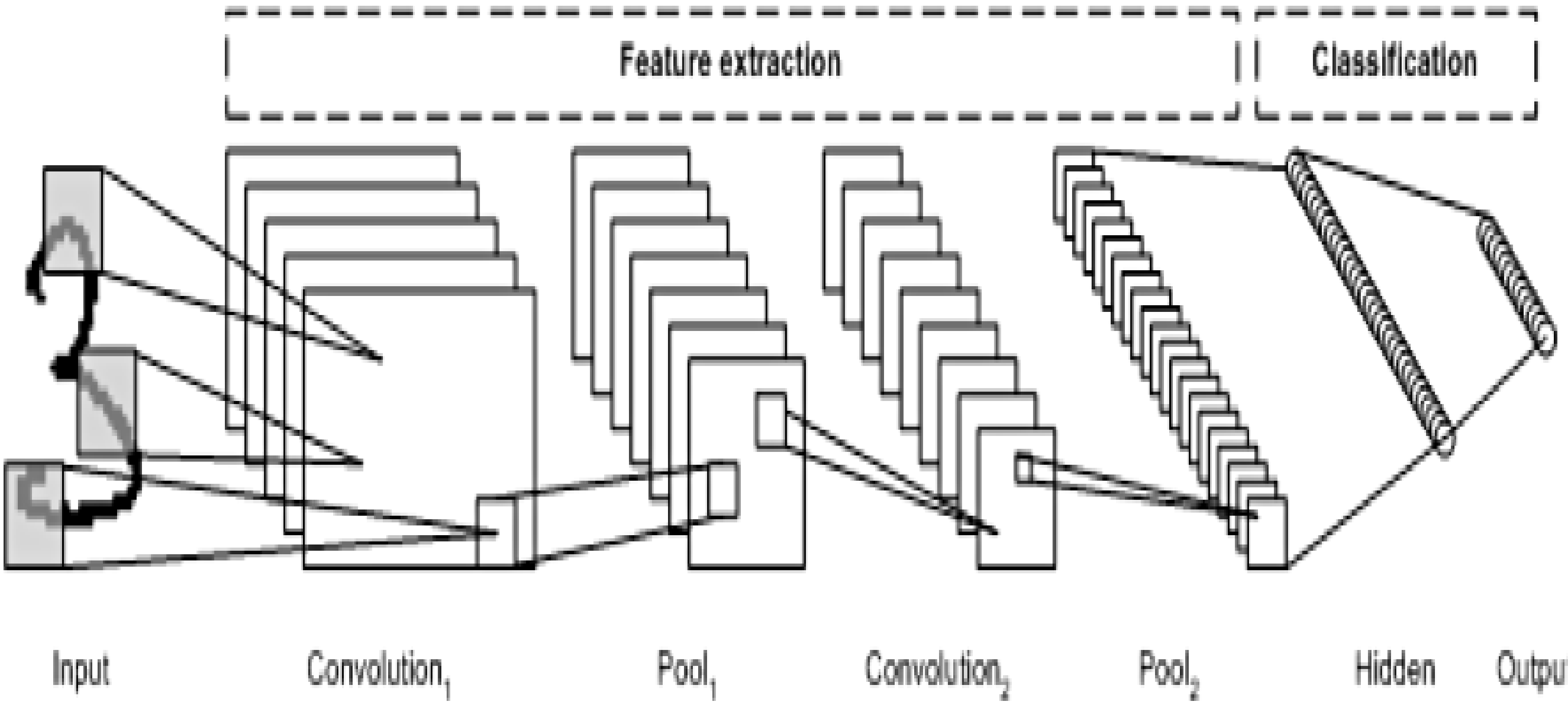


Figure: Convolution neural networks

Source: Smart insights © March 2018 the financial brand

Recurrent neural networks

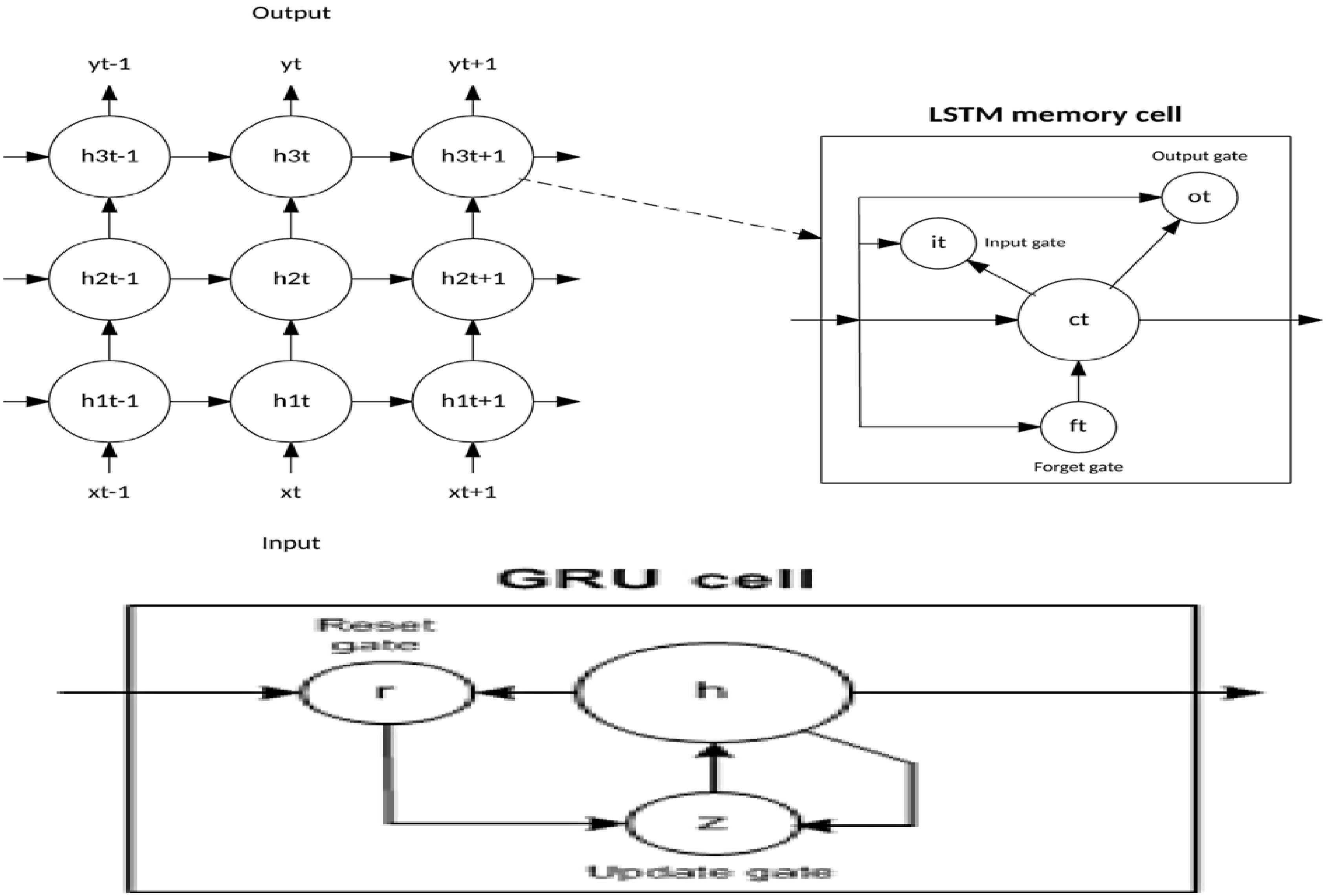


Figure: Recurrent neural networks

Source: Smart insights © March 2018 the financial brand

Machine learning techniques: Loan underwriting & sentiment/news analysis

IBM ICE (Innovation Centre for Education)

Challenges	Potential Solutions
Verification of discrepancies in the document **Variety of document formats to checked for its accuracy. If performed manually, increase the likelihood of errors and mistakes. Also, it is time consuming**	ML based Natural language processing, along with image analysis leads to effective digitization of the entire process reducing potential errors and reducing the time.
Issues in the credit analysis: Documenting and analyzing past financial transactions is a tedious task if performed manually.	AI based metrics for credit analysis may lead to unbiased and validated credit-worthiness. This involves deployment of appropriate ML algorithm to process the credit history of the customer along with data analytics.
Assessment of debt-to-income ratio: Manual assessment of borrower's overall debt to income along with evaluation of their ability to repay the loan is cumbersome task.	ML enabled applications to evaluate debt-to-income ratio can be authentic source of evidence to make appropriate decision.

Sentiment or news analysis

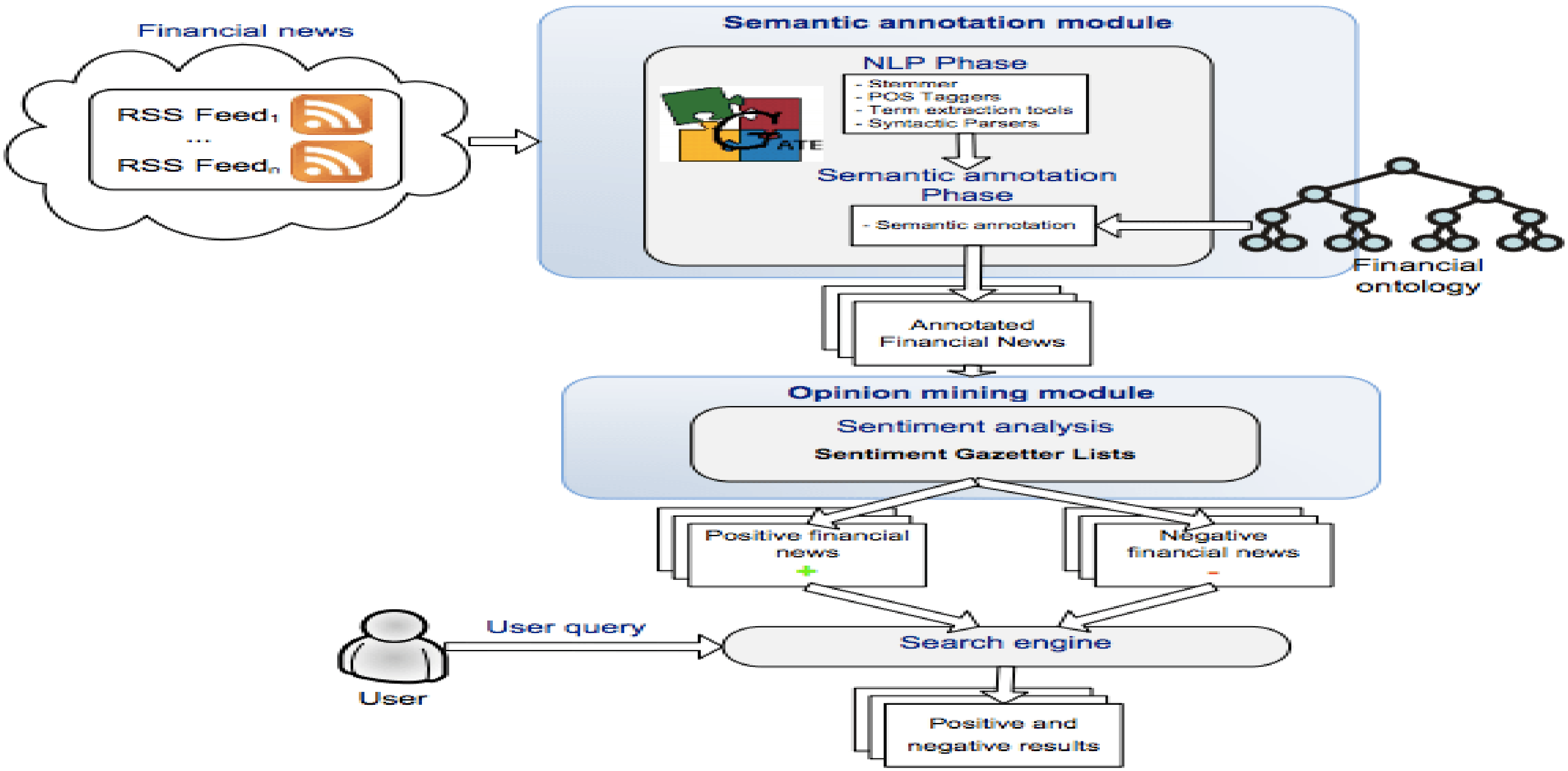


Figure: Flow of sentiment analysis application in the financial world.

Source: Semantic-Based Sentiment analysis in financial news, Vol-862

Current challenges and opportunities: Banking and security domain

- Challenge: Lack of skills and data.
- Opportunity: More usable AI is coming.
- Challenge: Adoption.
- Opportunity: Professionals recognize the wider value.
- Opportunity: AI can simplify transparency and explain ability.

Unit summary

Having completed this unit, you should be able to:

- Learn about machine learning in banking sectors, challenges in banking sector and fraud detection system in banking sector
- Understand manage customer data and algorithms in banking and security
- Gain knowledge on deep learning technology for personalized marketing
- Learn about applications of machine learning classifiers in credit risk analysis
- Understand the importance of machine learning in portfolio management systems
- Learn about algorithmic trading and stages for implementing applications of ai in marketing
- Deep learning approach for sentiment analysis of customers and customer services
- Understand cyber security in banking sector and loan underwriting and sentiment/news analysis
- Understand current challenges and opportunities in implementing machine learning technologies in banking and security domain

Welcome to:

Machine Learning in Communication, Media and Entertainment

Unit objectives

After completing this unit, you should be able to understand:

- Understand the purpose of using machine learning in communication, media and entertainment
- learn about the usage of machine learning in media and entertainment industry
- Learn about analyse the customer sentiment using machine learning models
- Understand need of real-time analytics in communication, media, entertainment industries
- Gain knowledge on deep learning for social media analytics
- Learn about different types of recommendation engines
- Understand the Restricted Boltzmann Machines (RBM) for collaborative filtering
- Gain knowledge on collaborative deep learning for recommender systems

Machine learning in communication, media and entertainment

- Identification and conviction of media content helps the customers to communicate easily with new media and materials from various sources successfully with the help of machine learning and data science techniques.

- AL and ML are the main technologies in the telecommunication industry that assist the firms to generate better income, build more trust from the customer end and have good customer relationships.

Usage of machine learning in media and entertainment Industry

- The contributions of analytics in the entertainment space are:
  - Helps us to understand consumer insights/psychology.
  - Tracking the customer's digital footprint to schedule ad campaigns accordingly.
  - Enhancing the product based on customer feedback.
  - To creating content based past available data.

- The following are the ways by which data science and machine learning are changing the entertainment and media industry:
  - Prediction of audience behaviour.
  - Analysing customer sentiment
  - Personalization of content

Machine learning techniques for customer sentiment analysis

- The concept of applying natural language processing and text analysis techniques to recognize and draw out subjective information from a piece of text is called sentiment analysis.

- Majority of the part, the feelings or opinions of a person are subjective and not actual facts.

- Another name given to sentiment Analysis is opinion mining, which is an area within Natural Language Processing (NLP) which builds up systems that can recognize and withdraw opinions within text.

World embedding's

- Word embedding is the collective name for a set of language modelling and feature learning techniques in Natural Language Processing (NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers.

- Conceptually it involves a mathematical embedding from a space with many dimensions per word to a continuous vector space with a much lower dimension.

- Methods to generate this mapping include neural networks, dimensionality reduction on the word co-occurrence matrix, probabilistic models, explainable knowledge base method, and explicit representation in terms of the context in which words appear.

Sentiment analysis with long short term memory networks

- LSTM networks are a type of RNN that uses special units in addition to standard units.
- LSTM units include a 'memory cell' that can maintain information in memory for long periods of time.
- A set of gates is used to control when information enters the memory, when it's output, and when it's forgotten. This architecture lets them learn longer-term dependencies.

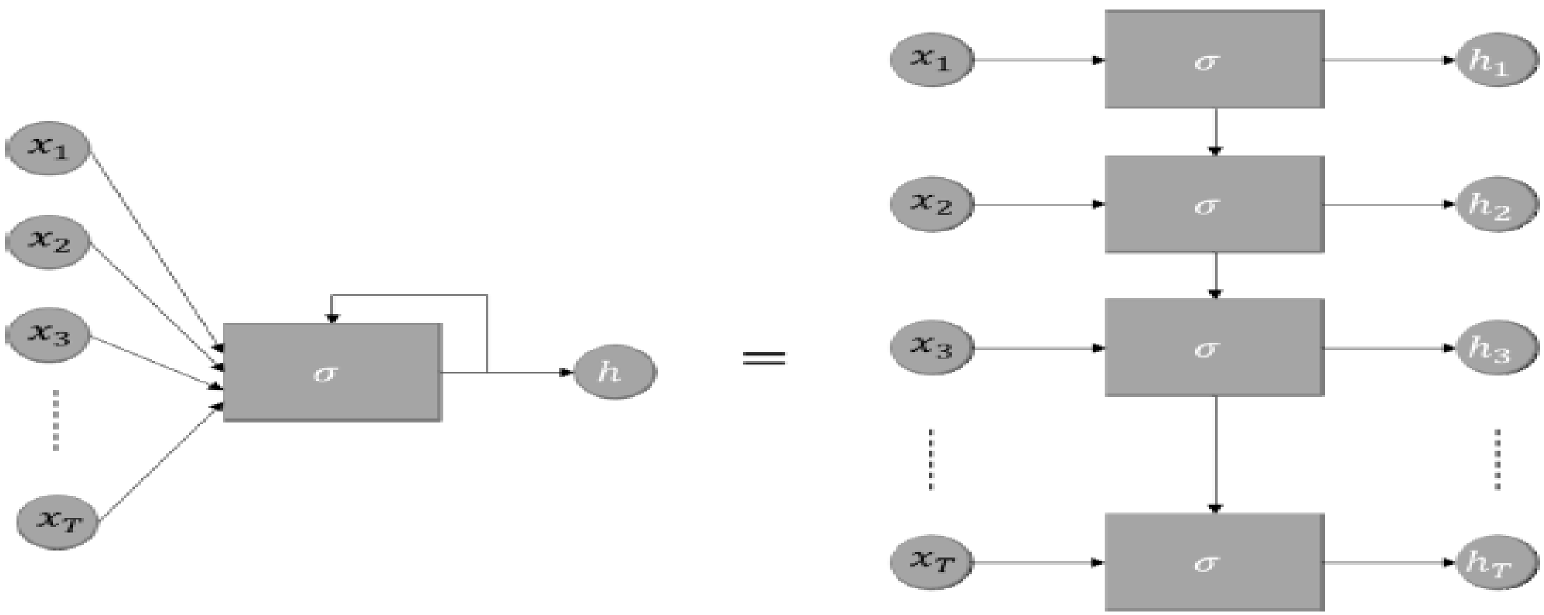


Figure: Recurrent neural network

Source: https://www.math.vu.nl/~sbhulai/papers/paper-miedema.pdf

real-time analytics in communication, media and entertainment industries

- Real time analytic is used in analysis of data as soon as the data become available.

- As soon as the data enters the system the customers were able to draw conclusions.

- Social sites like Facebook, Twitter and Instagram were used to advertise their products and services to brand it across the globe, media customers and entertainment customers had intensively high competition.

Real time analytics and social media

- The areas in which social media analytics make a big impact are as follows:
  - Innovation.
  - Marketing.
  - Sales.
  - Costumer services.
  - Competitive ontelligent.

Deep learning for social media analytics

- Social media analytics is the practice of gathering data from social media websites and analysing that data using social media analytics tools to make business decisions.

- The most common use of social media analytics is to mine customer sentiment to support marketing and customer service activities.

- The content generated by the user are composed by interactive web 2.0 internet-based applications such as:
  - Texts.
  - Posts.
  - Comments.
  - Videos.

- All the data is generated through online communications.

Recommendations engines

- The recommender system technology recommend users and help setting up in the environment for things like:
  - Products.
  - Movies.
  - Events.
  - Articles.

- Types of recommendation engines:
  - Collaborative filtering.
  - Content based filtering.
  - Hybrid recommendation systems.

Collaborative filtering

- This filtering technique is based on the resemblances of the users and its built on collecting and analyzing information of user's activities such as their:
  - Behaviors.
  - Choices.

- This predicts what the user may like or prefer.

- Types of collaborative filtering:
  - Memory based collaborative filtering.
  - Content based collaborative filtering.

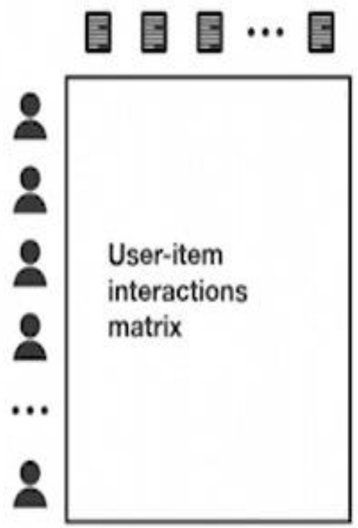
Collaborative Filtering Techniques

COLLABORATIVE FILTERING (CF)

MEMORY Based

Rely heavily on similarity measures to match similar people or items together

- Cosine similarity
- Pearson correlation
- Jaccard coefficient ...



MODEL Based

Use Machine Learning to find user ratings of unrated items

- Bayesian Networks
- Singular Value Decomposition (SVD)
- Probabilistic Latent Semantic Analysis
- Neural Nets
- Matrix Factorization ....

How much will user A like the new item B?

They try to build a model that can predict the user's rating for a NEW item

Memory Based Collaborative Filtering

- **How does it work?**

GIVEN

USERS $u \in \{1, \dots, U\}$

ITEMS $i \in \{1, \dots, M\}$

Training set T with observed, real-valued preferences rui for some user-item pairs (u, i)

rui = e.g. purchase indicator, item rating, click count . . .

GOAL

Predict unobserved preferences

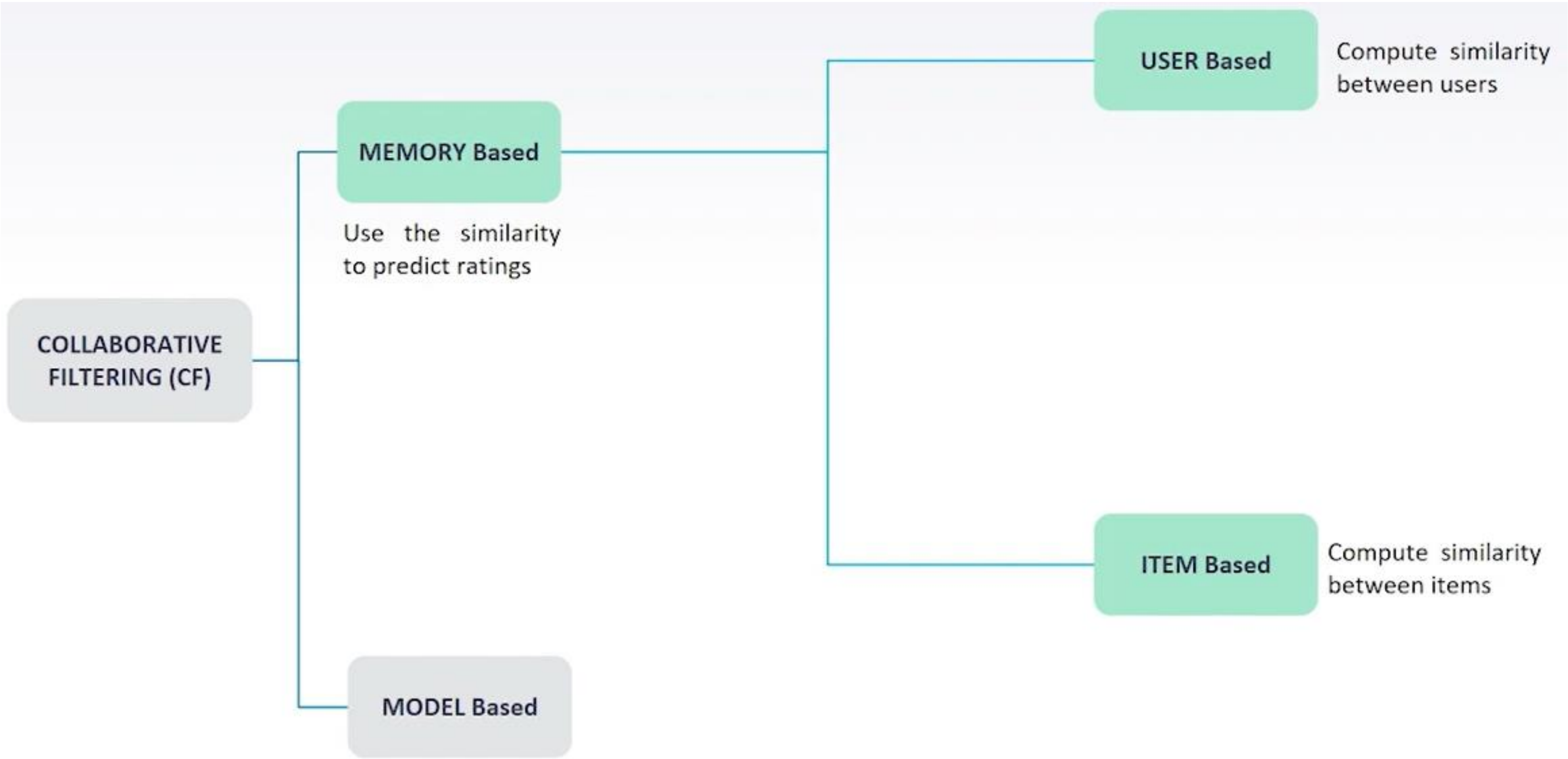
Test set Q with pairs (u, i) not in T

View as matrix completion problem

Fill in unknown entries of sparse preference matrix

$$\mathbf{R} = \underbrace{\left[\begin{array}{ccccc} ? & ? & 1 & \dots & 4 \\ 3 & ? & ? & \dots & ? \\ ? & 5 & ? & \dots & 5 \end{array}\right]}_{M \text{ items}} \Bigg\} \, U \text{ users}$$

Memory Based Techniques

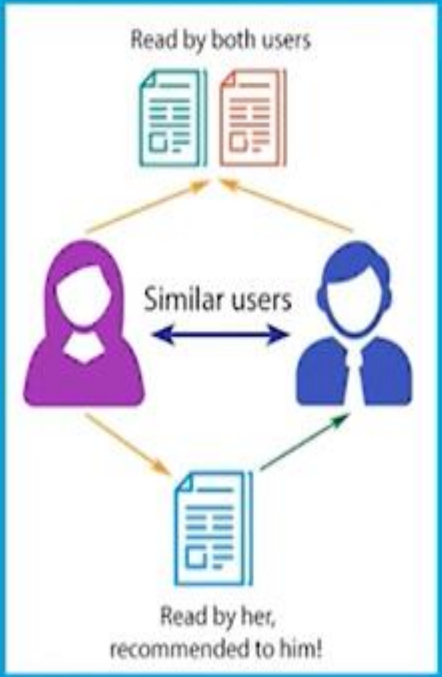


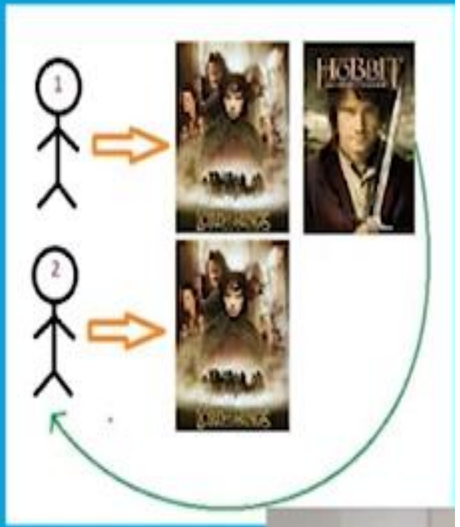
User-based Collaborative Filtering

- Similarity based on users

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

- Look for items that are similar to Item5
- Take Alice's ratings for these items to predict the rating for Item 5





Collaborative Filtering Algorithms

- User-based

ALGORITHM 1: Nearest Neighbours over users

Find K users whose taste in items (e.g. movies) is closest to you and who have expressed an opinion on item M. Combine their votes about M

Steps:

1. We need to make user vs item matrix
2. Compute similarity scores between users.
3. Find users who are similar to you based on past behaviours
4. Recommend items seen by users similar to you that you didn't see

ALGORITHM 2: Clustering users

U := user M := item (e.g. movie)

Divide users into clusters, based on their item preferences. Take vote on I over M's cluster.

	Book1	Book2	Book3	Book4	Book5	Book6
Customer A	X			X		
Customer B		X	X		X	
Customer C		X	X			
Customer D		X				X
Customer E	X				X	

Customers B, C and D are « clustered » together. Customers A and E are clustered into another separate group

- « Typical » preferences for CLUSTER are:
  - Book 2, very high
  - Book 3, high
  - Books 5 and 6, may be recommended
  - Books 1 and 4, not recommended at all

Item-based Collaborative Filtering

- Similarity based on items

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

- Look for items that are similar to Item5
- Take Alice's ratings for these items to predict the rating for Item 5

Collaborative Filtering Algorithms

- Item-based

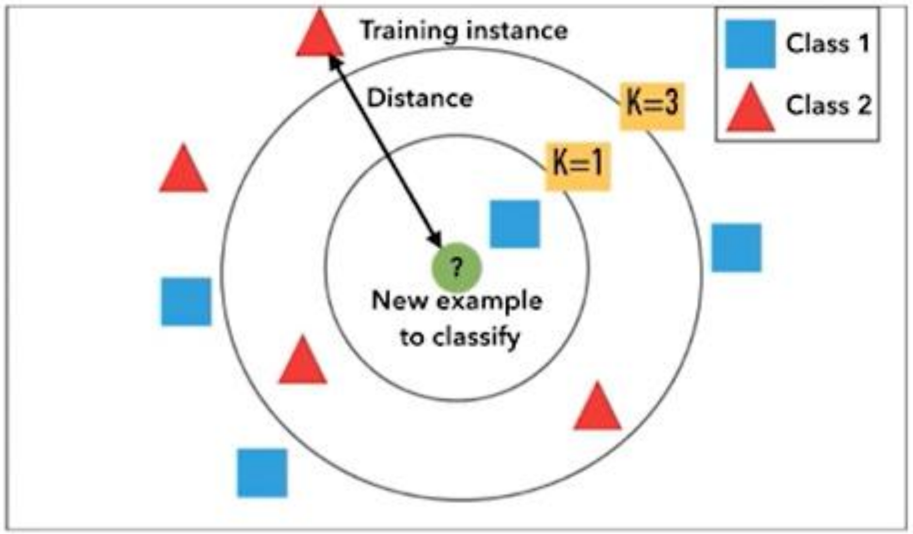
ALGORITHM 1: Nearest Neighbours over items

KNN will calculate the "distance" between the target movie and every other movie in its database, then it ranks its distances and returns the top K nearest neighbor movies as the most similar movie recommendations

ALGORITHM 2: Clustering users

U := user M := item (e.g. movie)

Divide items into clusters, based on their fan-list. Average U's votes over M's cluster.



Euclidean Distance

- Measuring User similarity

	Lady	Snake	Luck	Superman	Dupree	Night
Lisa	2.5	3.5	3.0	3.5	2.5	3.0
Gene	3.0	3.5	1.5	5.0	3.5	3.0
Michael	2.5	3.0		3.5		4.0
Claudia		3.5	3.0	4.0	2.5	4.5
Mike	3.0	4.0	2.0	3.0	2.0	3.0
Jack	3.0	4.0		5.0	3.5	3.0
Toby		4.5		4.0	1.0	

$$Similarity(X, Y) = \frac{1}{1 + EuclideanDistance(X, Y)}$$

(gives value between 0 and 1)

$$Similarity('Michael', 'Claudia') = \frac{1}{1 + \sqrt{(3.0 - 3.5)^2 + (3.5 - 4.0)^2 + (4.0 - 4.5)^2}} = 0.536$$

Cosine Similarity

- Measuring User Similarity

	Up	Wall-e	Finding Nemo	Toy Story	Tim
Joe	4	4	4	3	2
Beck	5	5	2	2	3

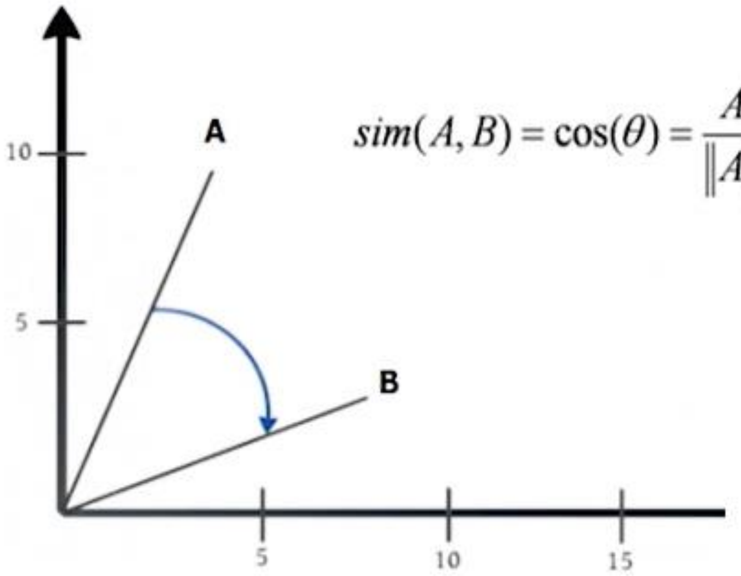
$$similarity = cos(\theta) = \frac{\boldsymbol{u} \cdot \boldsymbol{v}}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|} = \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}}$$

DOT PRODUCT

$$\boldsymbol{u} \cdot \boldsymbol{v} = [u_1 \ u_2 \ \ldots \ u_n] \cdot \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = u_1 v_1 + u_2 v_2 + \ldots + u_n v_n = \sum_{i=1}^{n} u_i v_i$$

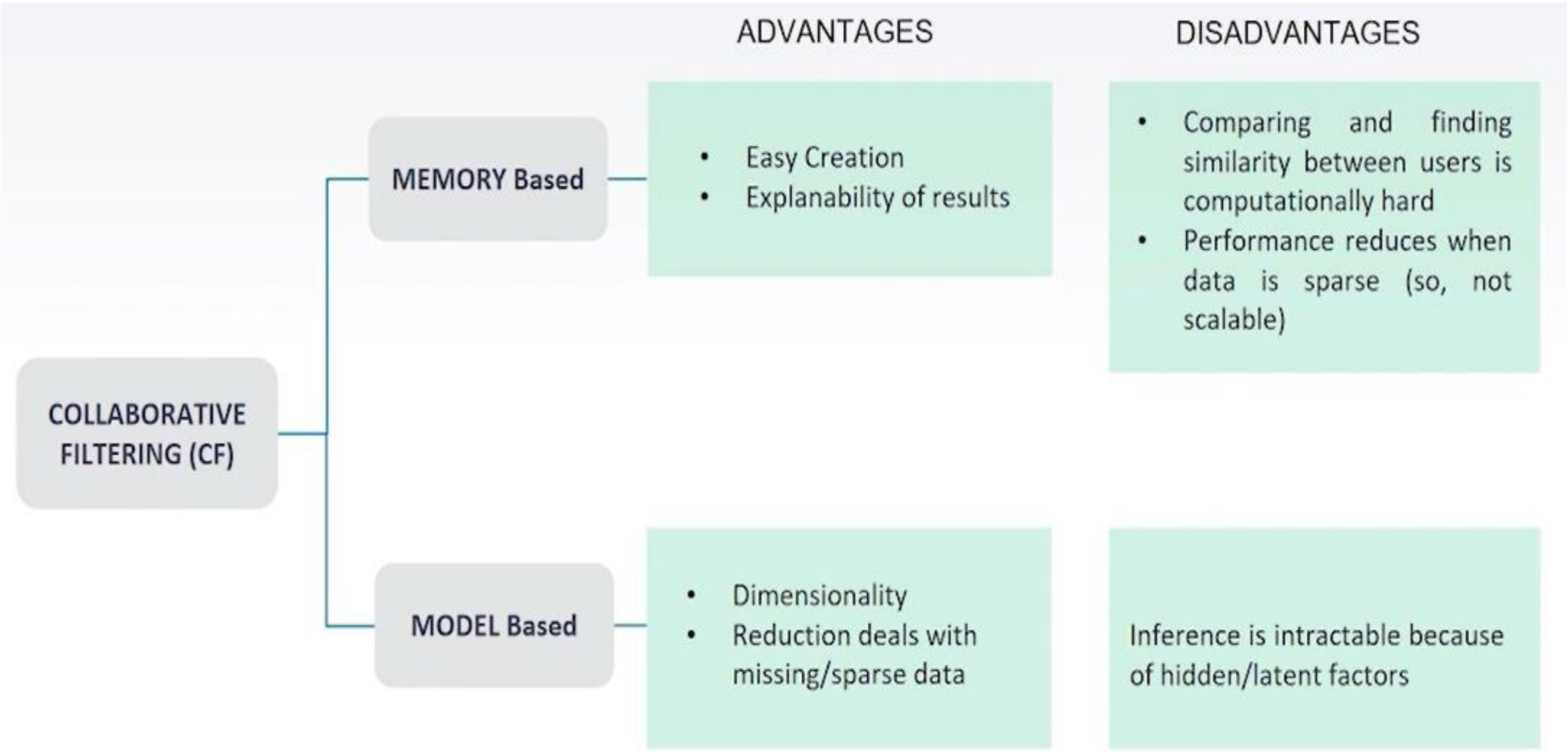
COSINE SIMILARITY BETWEEN JOE AND BECK :

```
import math
def cosine_similarity(v1,v2):
    "compute cosine similarity of v1 to v2: (v1 dot v2)/{||v1||*||v2||)"
    sumxx, sumxy, sumyy = 0, 0, 0
    for i in range(len(v1)):
        x = v1[i]; y = v2[i]
        sumxx += x*x
        sumyy += y*y
        sumxy += x*y
    return sumxy/math.sqrt(sumxx*sumyy)

v1,v2 = [4, 4, 4, 3, 2] [5, 5, 2, 2, 3]
print(v1, v2, cosine_similarity(v1,v2))

Output: [4, 4, 4, 3, 2] [5, 5, 2, 2, 3] 0.938531668711012
```

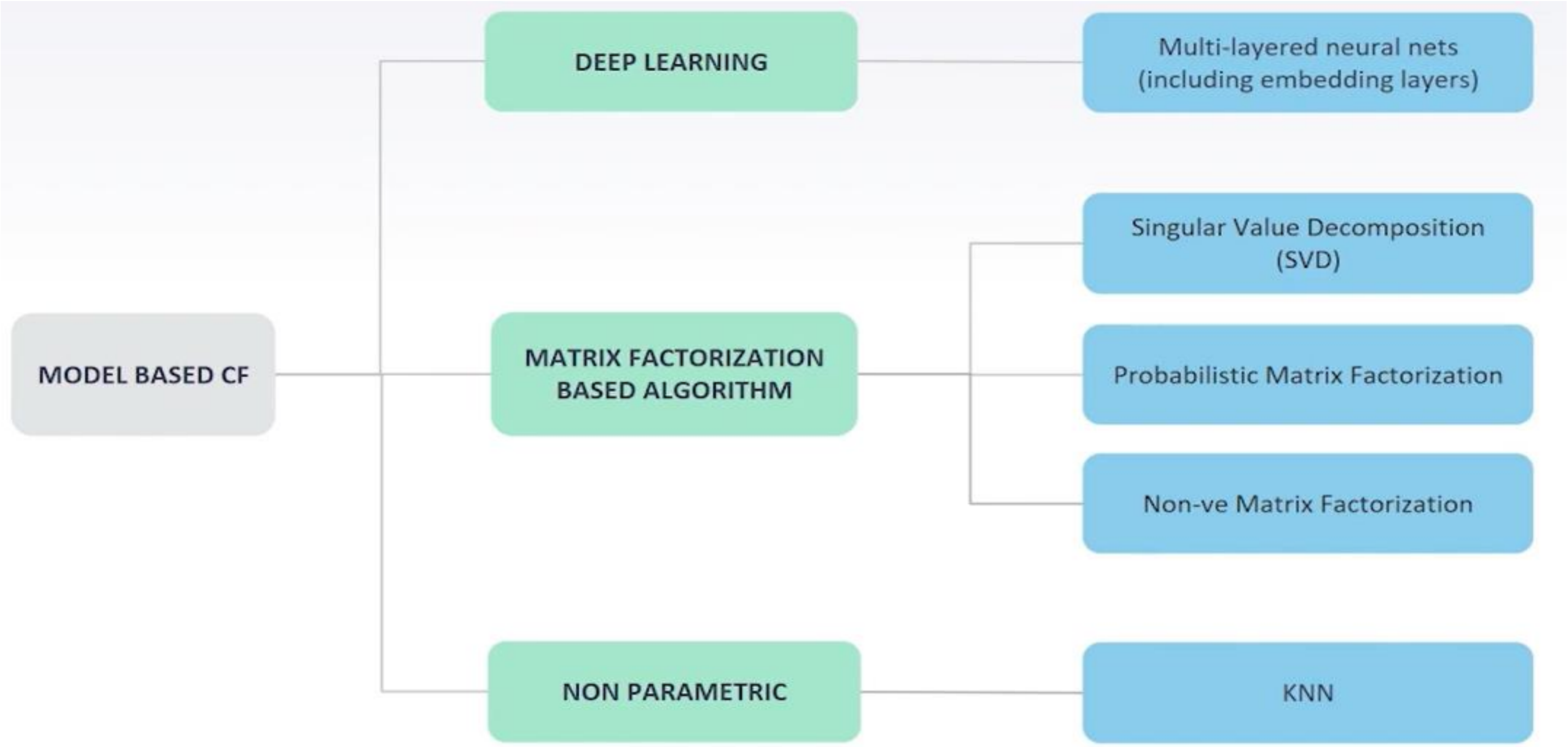


$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Collaborative Filtering Techniques



Model-based Collaborative Filtering



Matrix Factorization

Assume that both movies and users live in some low-dimensional space describing their properties

Recommend a movie based on its proximity to the user in the latent space

Given:

$A \in R^{m\,x\,n}$ Feedback Matrix m = # users n = # items

The model learns how to generate:

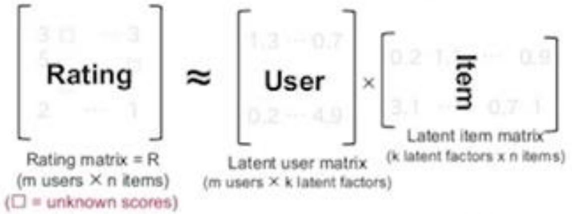
$U \in R^{m\,x\,d}$ User Matrix d = embedding dimension

$V \in R^{n\,x\,d}$ Item Matrix

The Feedback matrix A (or R) can be decomposed to latent factors of users and items

$A = U\ x\ V^T$ Rating Matrix

$A^* = U\ x\ V^T$ Prediction Matrix



Singular Value Decomposition (SVD)

- How to obtain a latent user/item matrix?

SVD is a famous linear algebra technique for matrix decomposition

m x m unitary matrix

n x n unitary matrix

$$\underset{\text{m x n matrix}}{X} = U \times \Sigma \times V^T$$

Rectangular diagonal matrix
(diagonal values are called as singular values)

$$X = U \times \Sigma \times V^T$$

$$\begin{bmatrix} 1 & 3 & 3 & 3 & 0 \\ 2 & 4 & 2 & 2 & 4 \\ 1 & 3 & 3 & 5 & 1 \\ 4 & 5 & 2 & 3 & 3 \\ 1 & 1 & 5 & 2 & 1 \end{bmatrix}$$

$$U = \begin{bmatrix} -0.369 & -0.325 & 0.282 & 0.343 & 0.749 \\ -0.459 & 0.448 & -0.339 & -0.584 & 0.364 \\ -0.468 & -0.359 & 0.544 & -0.459 & -0.381 \\ -0.563 & 0.491 & 0.07 & 0.562 & -0.348 \\ -0.341 & -0.569 & -0.71 & 0.118 & -0.202 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 13.368 & 0 & 0 & 0 & 0 \\ 0 & 4.708 & 0 & 0 & 0 \\ 0 & 0 & 2.792 & 0 & 0 \\ 0 & 0 & 0 & 1.586 & 0 \\ 0 & 0 & 0 & 0 & 0.904 \end{bmatrix}$$

$$V = \begin{bmatrix} -0.325 & 0.341 & -0.101 & 0.682 & -0.55 \\ -0.562 & 0.345 & 0.273 & 0.153 & 0.684 \\ -0.468 & -0.642 & -0.577 & 0.125 & 0.141 \\ -0.504 & -0.327 & 0.601 & -0.324 & -0.416 \\ -0.324 & 0.496 & -0.47 & -0.626 & -0.189 \end{bmatrix}$$

With SVD we can approximate a given matrix by focusing on the largest singular values, and the vectors of U and V which correspond to the singular values

Step 1 Run SVD

$$X = U \times \Sigma \times V^T$$

Step 2 Focus on important features

$$U_2 \times \Sigma_2 \times V_2^T$$

$$\parallel$$

Step 3 Multiply three matrix

$$\begin{bmatrix} 1.08 & 2.24 & 3.29 & 2.98 & \mathbf{0.84} \\ 2.72 & 4.17 & 1.52 & 2.41 & 3.04 \\ 1.46 & 2.93 & 4.02 & 3.71 & 1.19 \\ 3.24 & 5.03 & 2.04 & 3.04 & 3.59 \\ 0.57 & 1.64 & 3.86 & 3.18 & 0.15 \end{bmatrix}$$

Step 4 Check the values which were zero before running SVD

$$U = \begin{bmatrix} -0.369 & -0.325 & 0.282 & 0.343 & 0.749 \\ -0.459 & 0.448 & -0.339 & -0.584 & 0.364 \\ -0.468 & -0.359 & 0.544 & -0.459 & -0.381 \\ -0.563 & 0.491 & 0.07 & 0.562 & -0.348 \\ -0.341 & -0.569 & -0.71 & 0.118 & -0.202 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 13.368 & 0 & 0 & 0 & 0 \\ 0 & 4.708 & 0 & 0 & 0 \\ 0 & 0 & 2.792 & 0 & 0 \\ 0 & 0 & 0 & 1.586 & 0 \\ 0 & 0 & 0 & 0 & 0.904 \end{bmatrix}$$

Largest singular values

$$V = \begin{bmatrix} -0.325 & 0.341 & -0.101 & 0.682 & -0.55 \\ -0.562 & 0.345 & 0.273 & 0.153 & 0.684 \\ -0.468 & -0.642 & -0.577 & 0.125 & 0.141 \\ -0.504 & -0.327 & 0.601 & -0.324 & -0.416 \\ -0.324 & 0.496 & -0.47 & -0.626 & -0.189 \end{bmatrix}$$

$$U = \begin{bmatrix} -0.369 & -0.325 \\ -0.459 & 0.448 \\ -0.468 & -0.359 \\ -0.563 & 0.491 \\ -0.341 & -0.569 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 13.368 & 0 \\ 0 & 4.708 \end{bmatrix}$$

Ignore unimportant

$$V = \begin{bmatrix} -0.325 & 0.341 \\ -0.562 & 0.345 \\ -0.468 & -0.642 \\ -0.504 & -0.327 \\ -0.324 & 0.496 \end{bmatrix}$$

Limitations of Collaborative Filtering Algorithms

ABSTRACT ION

- A recommendation is calculated based on a mathematical model
- However, reality is far more complex than that

E.g.

User 1 watched Ice Ventura 1 and 2, which is a Comedy Movie. User 1 also watched Monthy Python

User 2 watched Ice Ventura 1 and 2, but not Monthy Python.

A User-Based CF system will recommend Monthy Python to User 2. But User 2 doesn't like English humor.

This recommendation is technically correct, but not really.

It's very hard to represent mathematically (as input to the system) british humor !

EVOLUTION

- Assumption: users with similar taste in past will have similar taste in future
- However, people change over time

Evaluation Metrics

- How good is the recommendation?

Interested in error on unseen test set Q, not on training set

For each (u, i) let

rui = true preference,

ˆrui = predicted preference

Precision

Root Mean Square Error

$$\mathrm{RMSE} = \sqrt{\frac{1}{|Q|} \sum_{(u,i) \in Q} (r_{ui} - \hat{r}_{ui})^2}$$

Mean Absolute Error

$$\mathrm{MAE} = \frac{1}{|Q|} \sum_{(u,i) \in Q} |r_{ui} - \hat{r}_{ui}|$$

Model based collaborative filtering

- **Advantages:**
  - Over fittings can be avoided easily and the dataset is scalable.
  - Improvement in prediction performance.

- **The limitations of model-based CF algorithms are:**
  - Due to inflexibility it becomes hard to attach information to model based systems for users who do not rate.
  - Since it's not able to generate reasonable recommendations it suffers from sparsity problems.

Content based filtering

- The profile of the user's choices and description of an item are the inputs for content-based filtering.

- The user profile is created to understand the sort of product an individual likes, keywords are also used to describe the item.

- If we like a product, we have the tendency to like a similar item and that's the principle behind content-based filtering.

Hybrid recommendation systems

- Hybrid methods can be implemented by independently creating and creating and integrating content based and collective predictions or in addition, by applying a cooperative approach to a content on the basis of capability and vice versa, or by integrating methods into a single model.

- The various ways of implementation of hybrid approaches are:
  - Implement content-based and collaborative methods separately and accumulate their observation.
  - Some of collaborative aspects are comprised into the content based approach.
  - Some content based features are integrated into collaborative approach.
  - General consolidate model is constructed which is integration of both collaborative and content based characteristics.
  - The problems faced in recommendation systems such as cold start and sparsity is resolved in hybrid recommender technique.

Summary of recommendation systems

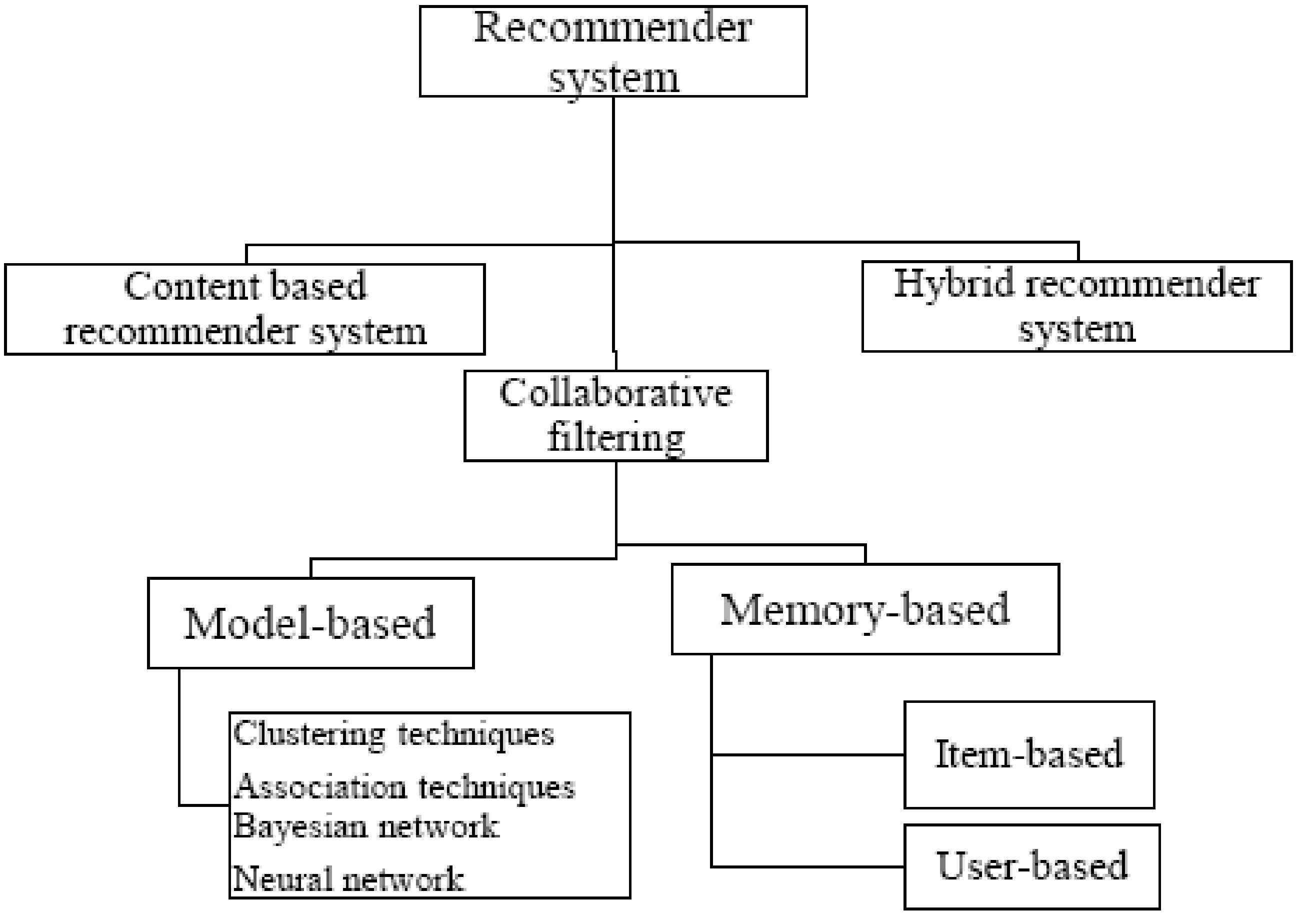


Figure: Summary of recommendation systems

Source: https://www.math.vu.nl/~sbhulai/papers/paper-miedema.pdf

Deep learning techniques on recommender systems

- Building recommender systems for collaborative and content-based approaches, deep learning has been suggested.

- Restricted Boltzmann Machines (RBM) for collaborative filtering:
  – The special version of Boltzmann Machine (BM) consists of layer hidden components and a layer of visible components with no hidden-hidden or visible-visible contacts is the Restricted Boltzmann Machine (RBM).

- Collaborative deep learning for recommender systems:
  – Collaborative deep learning for recommender systems was introduced to address the cold start problems and it utilizes review texts and ratings.

- Bayesian Stack De-Noise Auto Encoder (SDAE) and Collaborative Topic Regression (CTR) is integrated to collaborative deep learning.

Unit summary

Having completed this unit, you should be able to:

- Understand the purpose of using machine learning in communication, media and entertainment
- learn about the usage of machine learning in media and entertainment industry
- Learn about analyse the customer sentiment using machine learning models
- Understand need of real-time analytics in communication, media, entertainment industries
- Gain knowledge on deep learning for social media analytics
- Learn about different types of recommendation engines
- Understand the Restricted Boltzmann Machines (RBM) for collaborative filtering
- Gain knowledge on collaborative deep learning for recommender systems

Smarter education for smarter planet

IBM Innovation Centre for Education

Welcome to:

Machine Learning in Healthcare and Life Science

Unit objectives

After completing this unit, you should be able to:

- Learn about applications of machine learning in health care
- Gain knowledge on the role of machine learning in drug discovery
- Learn about machine learning approaches in drug discovery
- Understand the applications of machine learning in medical image analysis
- Learn about compare the architectures of different types of deep learning models
- Gain knowledge on the applications of machine learning in genetics and genomics
- Understand the ML applications in breast cancer diagnosis and prognosis

Applications of machine learning in health and life sciences

- Artificial Intelligence (AI), machine learning, and deep learning are storming the healthcare industry.

- The most promising fields of application are automated diagnosis.

- Almost all major healthcare firms have already begun to use the technology.

The most important applications of machine learning in healthcare

- Identifying diseases and diagnosis.
- Drug discovery and manufacturing.
- Medical imaging diagnosis.
- Personalized medicine.
- Machine learning based behavioural modification.
- Smart health records.
- Clinical trial and research.
- Crowd-sourced data collection.
- Better radiotherapy.
- Outbreak prediction.

Role of machine learning in drug discovery

- The drug development and production pipelines are large, complicated and rely on many considerations.
- Machine Learning (ML) methods have plentiful, high-quality data resources that can improve discovery and judgment-making.
- Examples include:
  - Target verification.
  - Prognostic biomarker recognition.
  - Clinical trial analysis of electronic pathology data.
- Applications varied from context to technique, with some methods making detailed forecasts and observations.

Machine learning approaches in drug discovery (1 of 6)

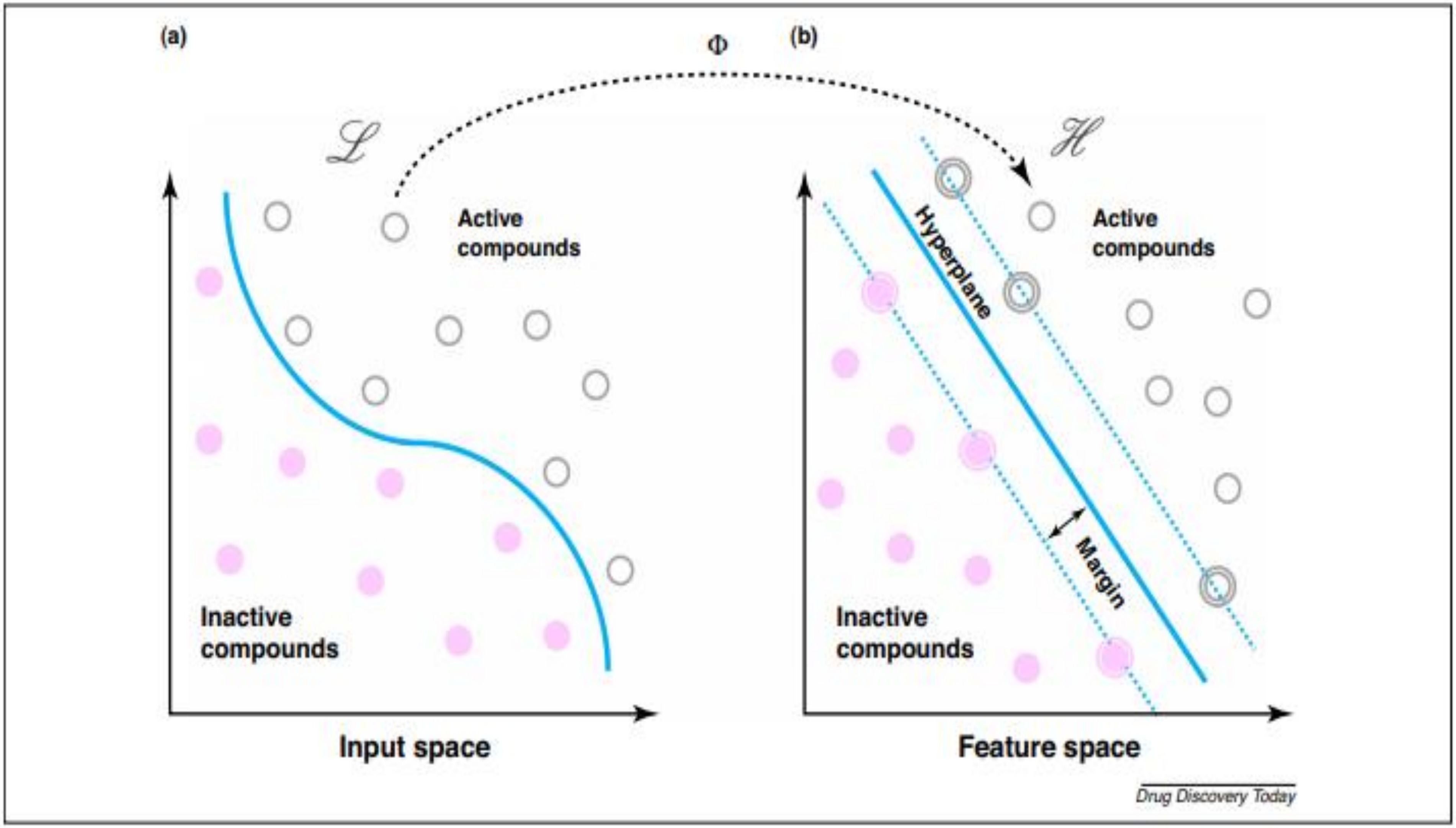


Figure: Support vector machines

Source: https://www.iris.unina.it/retrieve/handle/11588/593395/28708/Lavecchia.pdf

Machine learning approaches in drug discovery (2 of 6)

- Decision Tree (DT): DT includes a collection of regulations that include the process by which chemical characteristics and/or identifier attributes are correlated with interest behaviour or properties.

- Often known as nodes are the root and leaves.

- That leaf module is allocated a location estate, while a non-leaf module (core or inner node) is allocated to a genomic descriptive term the is an exam situation with departments divided into different classes of character.

Machine learning approaches in drug discovery (3 of 6)

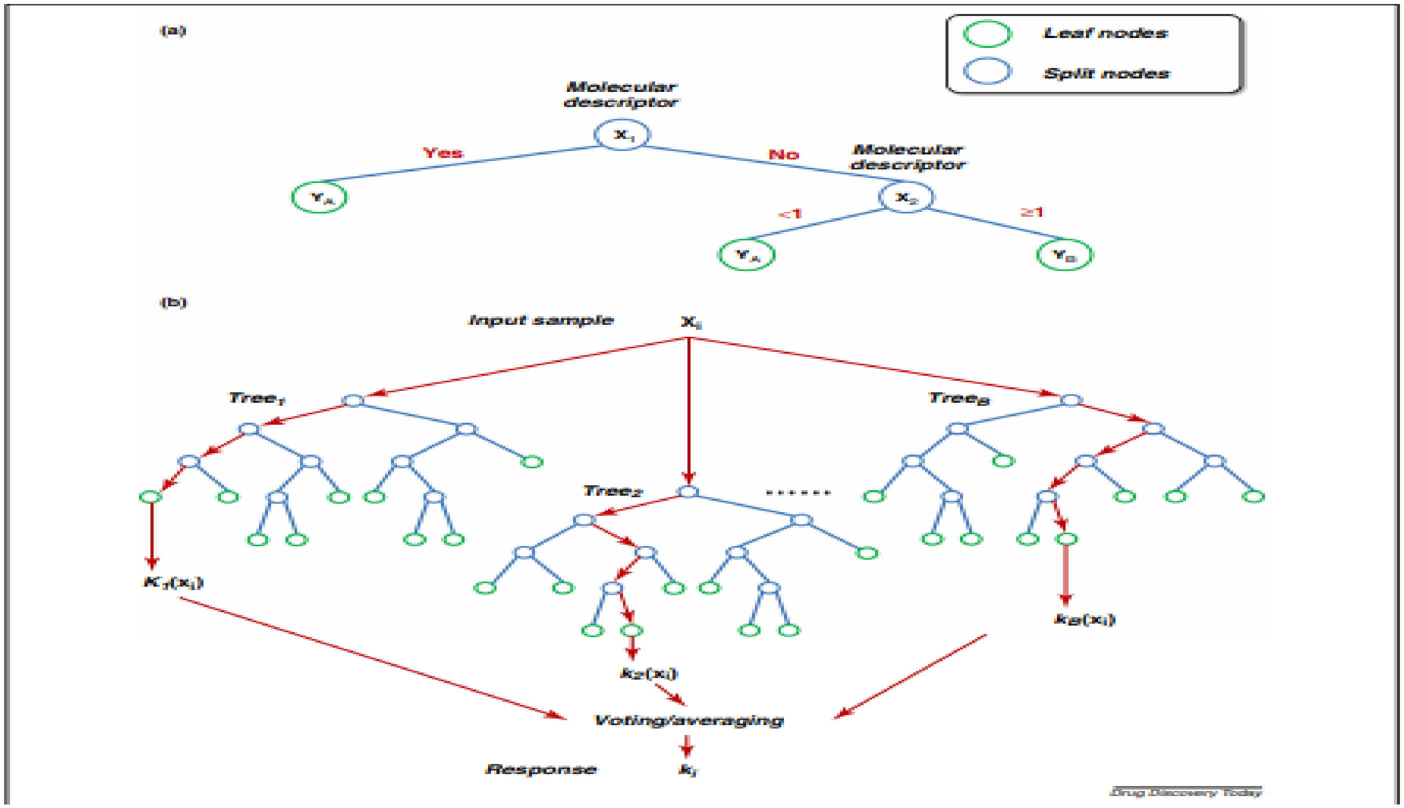


Figure: Ensemble methods

Source: https://www.iris.unina.it/retrieve/handle/11588/593395/28708/Lavecchia.pdf

Machine learning approaches in drug discovery (4 of 6)

IBM ICE (Innovation Centre for Education)

- Naïve Bayesian classifier: In chemo informatics, Naive Bayesian classifiers are commonly utilized to estimate biochemical rather than physicochemical characteristics alongside or compared to other classifiers.

- P(A/B) – $P\left(\frac{B}{A}\right)P(A)/P(B)$

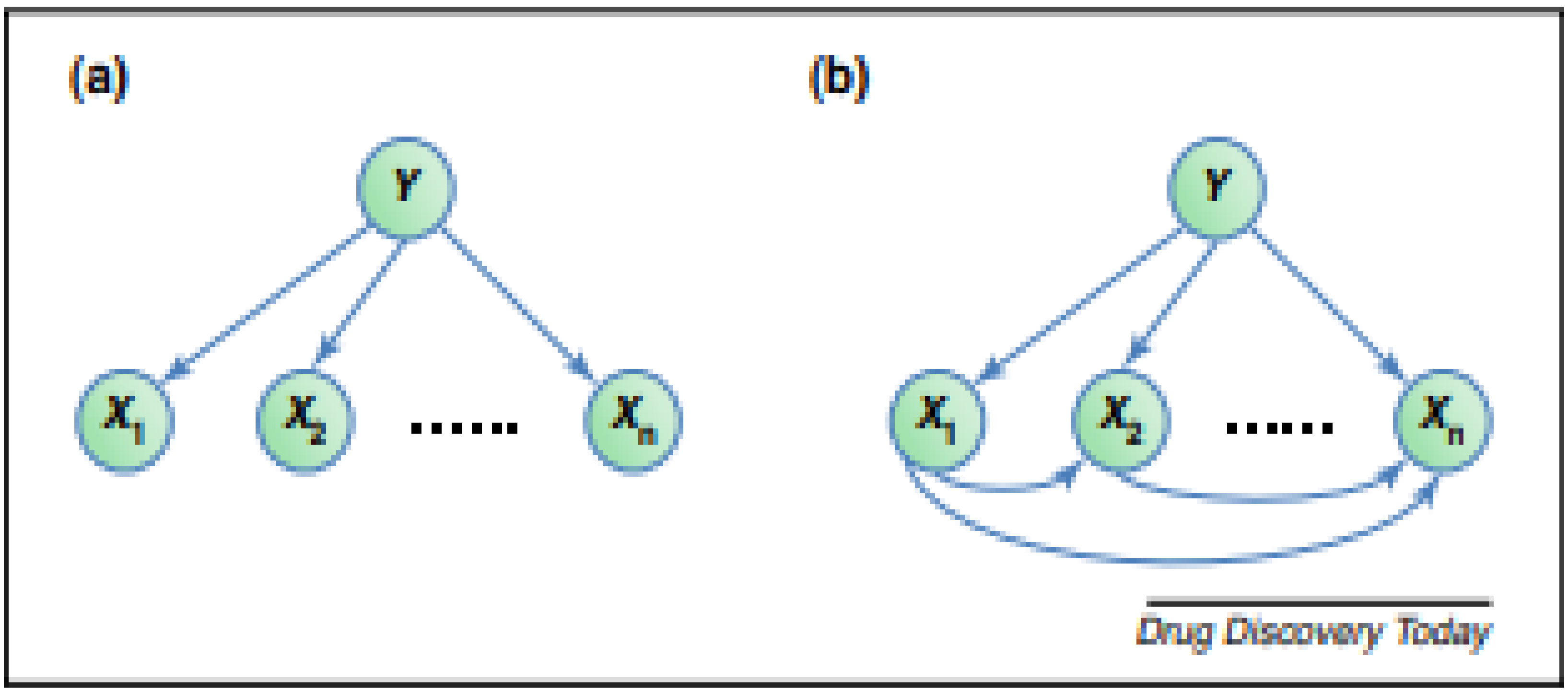


Figure: Drug discovery

Source: https://www.iris.unina.it/retrieve/handle/11588/593395/28708/Lavecchia.pdf

Machine learning approaches in drug discovery (5 of 6)

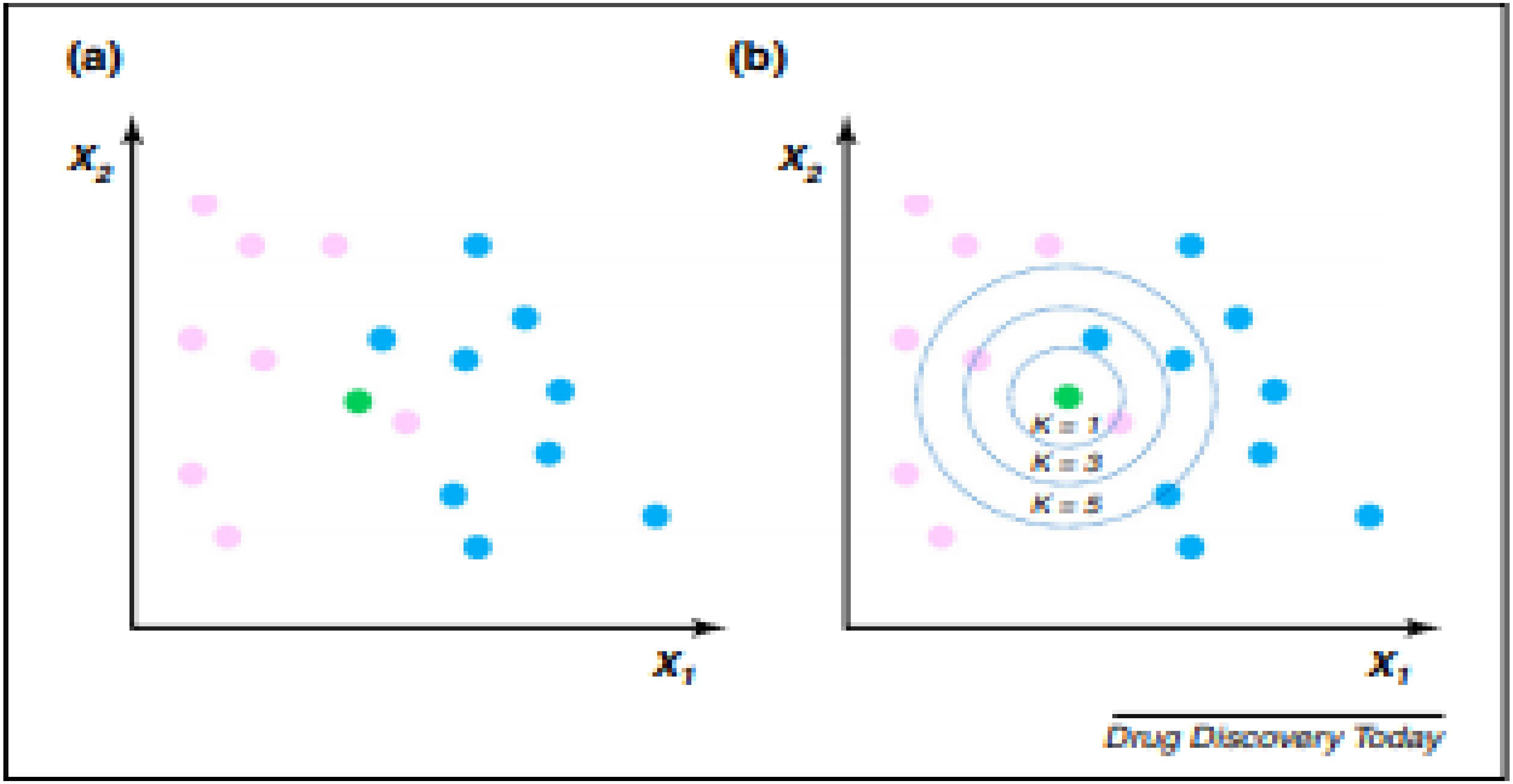


Figure: k-Nearest neighbours

Source: https://www.iris.unina.it/retrieve/handle/11588/593395/28708/Lavecchia.pdf

Machine learning approaches in drug discovery (6 of 6)

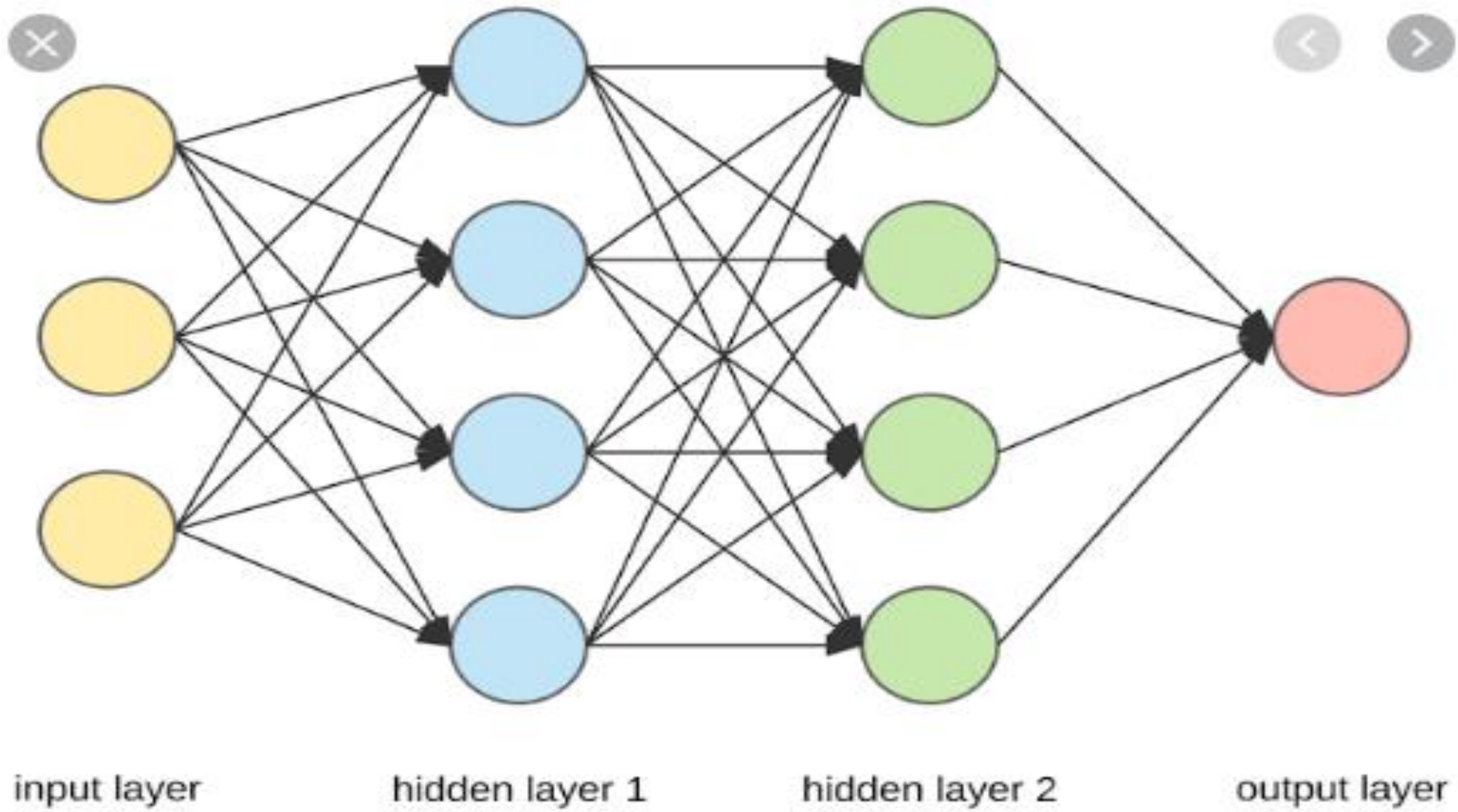


Figure: Artificial neural networks

Source: https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6

Medical image analysis

- The medical care industry is completely distinct from other sectors.
- It is a high preferential business and individuals allow the largest degree of care and facilities, regardless of cost.
- Limitations of human interpretation.

Why deep learning for medical image analysis

- Accurate diagnosis of disease.
- Improvements in image processing algorithms.
- Current training approaches are not reliable because of the wide variation between patient and medical outcomes.
- Deep learning now has a considerable potential.

Neural network and deep learning architecture

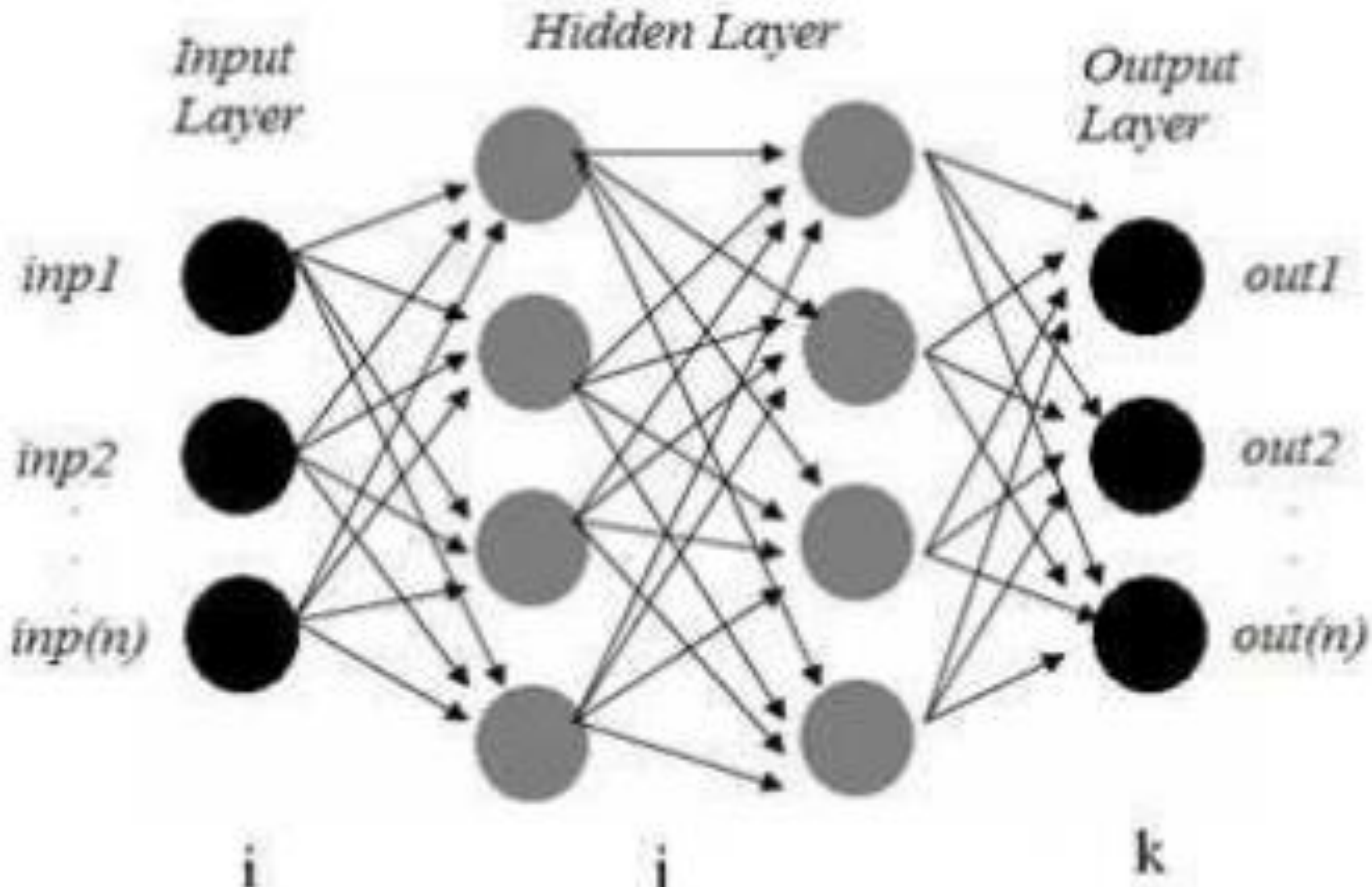


Figure:Neural network and deep learning architecture

Source: https://www.iris.unina.it/retrieve/handle/11588/593395/28708/Lavecchia.pdf

Comparisons between architecture of different types of deep learning models

- Deep Neural Network (DNN).
- Convolution Neural Network (CNN)
- Recurrent Neural Network (RNN).
- Deep Boltzmann Machine (DBM).
- Deep Belief Network (DBN).
- Deep Auto-encoder (DA).

Machine learning in genetics and genomics

- Genetics is a scientific study of the effects genes have on an organism, which are units of inheritance.

- Genes contain information in the DNA molecule, a sequence of chemicals called bases.

- Genomics All the genes taken together by an organism, and all the sequences and details found in it, are called the genome.

Genomics and AI background

- The ability to decode DNA allows scientists to "read" the biological code that directs a human organism's behaviours.
- The genome is also an organism's inherited number of genes.
- Genomics is closely associated with the medicine of accuracy.

Two category of genomics

- Genome sequencing (particularly as it applies to precision medicine).

- Direct-to-consumer genomics.

How to use deep learning effectively

- Data from genomics is often highly unbalanced.

- Effective implementation of profound learning, like all other aspects of machine learning, also involves domain knowledge.

Interpreting deep learning models

- There should be no ambiguity between the experimental methods addressed here and explanatory frameworks seeking to establish connections between cause and effect.

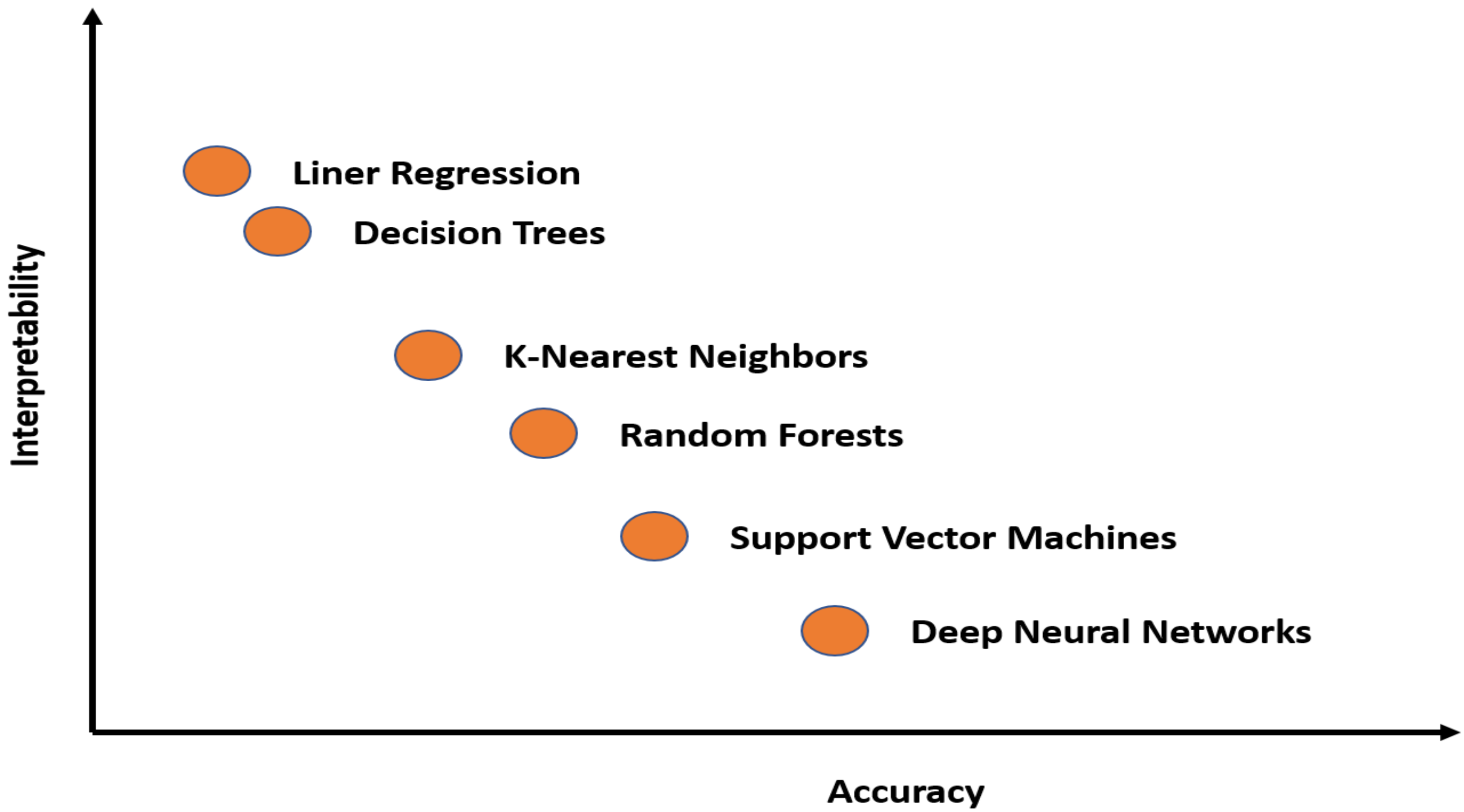


Figure: Interpreting deep learning models

Source: https://www.iris.unina.it/retrieve/handle/11588/593395/28708/Lavecchia.pdf

Predictive medicine: Prognosis and diagnostics accuracy

- Predictive medicine is a medicine area where the threat of infection is forecast and protective steps are applied to either fully avoid the infection or substantially reduce its effects on the person.

- The aim of preventive medication is to foresee the probability of potential illness.

Predictive medicine: Examples

- Carrier testing: Carrier testing is done to classify individuals with one version of a gene defect that, if contained in both versions, induces a genetic illness.

- Diagnostic testing: Diagnostic research is done to help diagnose and classify a particular illness.

- New-born screening: Shortly after birth, infants testing is carried out to recognize genetic illness that can be handled late in life.

- Prenatal testing: Prenatal screening is used to track fetal or embryo illnesses before conception.

ML applications in breast cancer diagnosis and prognosis

- Artificial Neural Networks (ANNs).
- Support Vector Machine (SVM).
- Decision Tree (DT).
- K Nearest Neighbours (KNN;x).

Unit summary

Having completing this unit, you should be able to:

- Learn about applications of machine learning in health care
- Gain knowledge on the role of machine learning in drug discovery
- Learn about machine learning approaches in drug discovery
- Understand the applications of machine learning in medical image analysis
- Learn about compare the architectures of different types of deep learning models
- Gain knowledge on the applications of machine learning in genetics and genomics
- Understand the ML applications in breast cancer diagnosis and prognosis