How to Kick-Start your Career in Data Science

Prashant Sahu

B.Tech., PhD (IIT Bombay, ongoing)

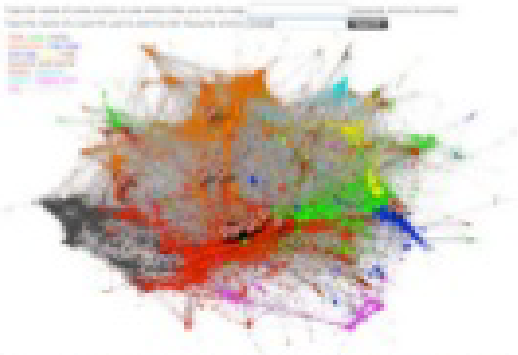
prashant.sahu@iitb.ac.in; prashant9501@gmail.com

https://www.linkedin.com/in/prashantksahu

What is Artificial Intelligence (AI)?

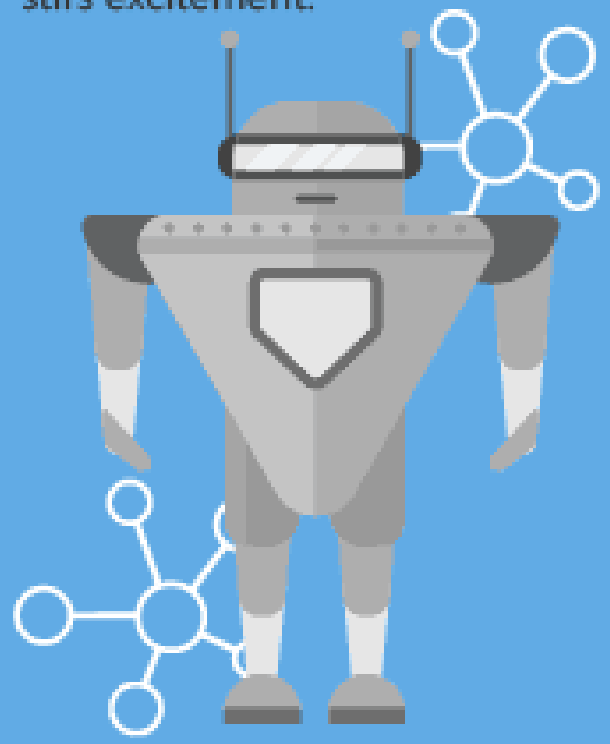
- John McCarthy coined AI term in 1956 as '**the science and engineering of making intelligent machines**' at a conference at Dartmouth College. Intelligent machine terms refer to the capability of performing intelligent human processes as:
  - Learning
  - Reasoning
  - Problem solving
  - Perception
  - Language understanding

- AI has become **an essential part of the technology industry**, providing the heavy lifting for many of the most difficult problems in computer science.
  - Prediction
  - Classification
  - Regression
  - Clustering
  - Function optimization



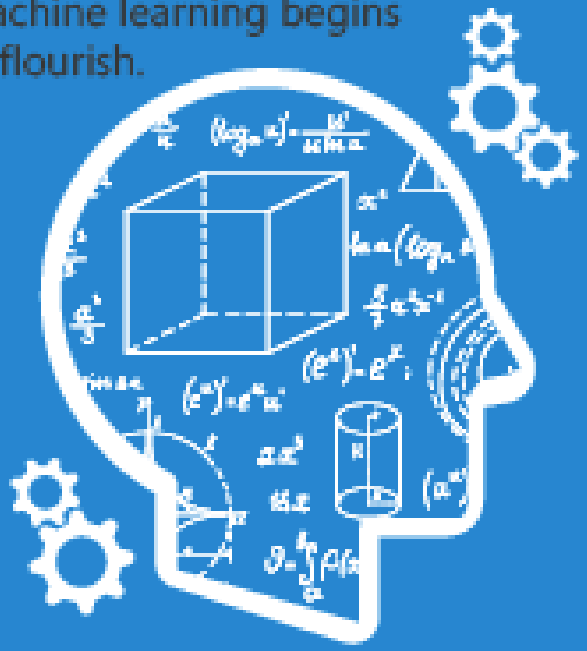
ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.



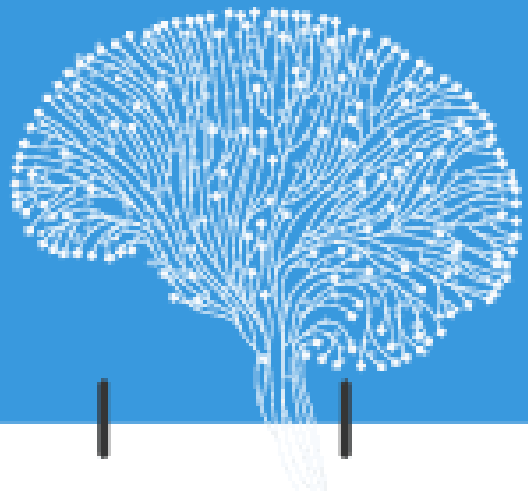
MACHINE LEARNING

Machine learning begins to flourish.



DEEP LEARNING

Deep learning breakthroughs drive AI boom.



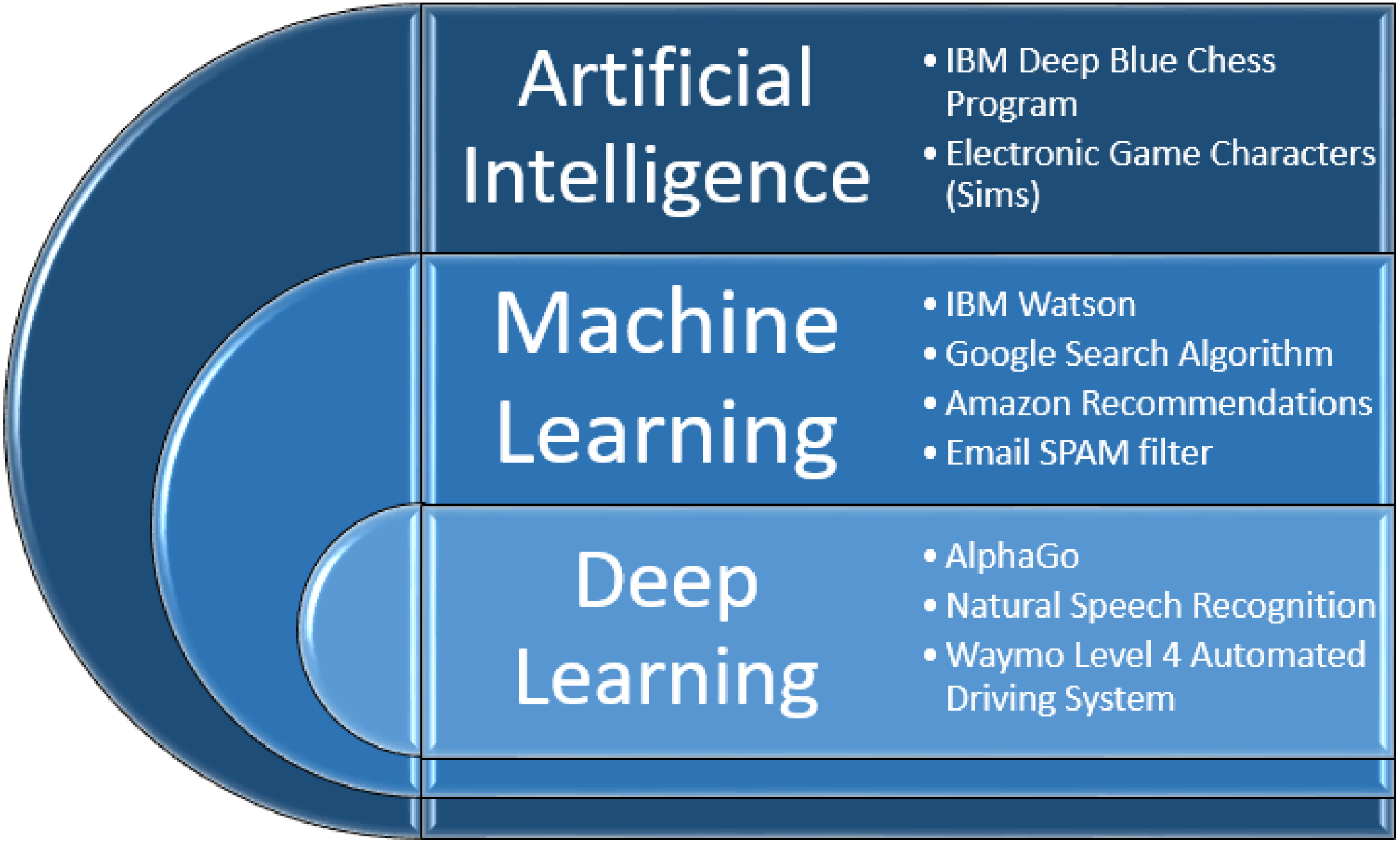
1950's 1960's 1970's 1980's 1990's 2000's 2010's

Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

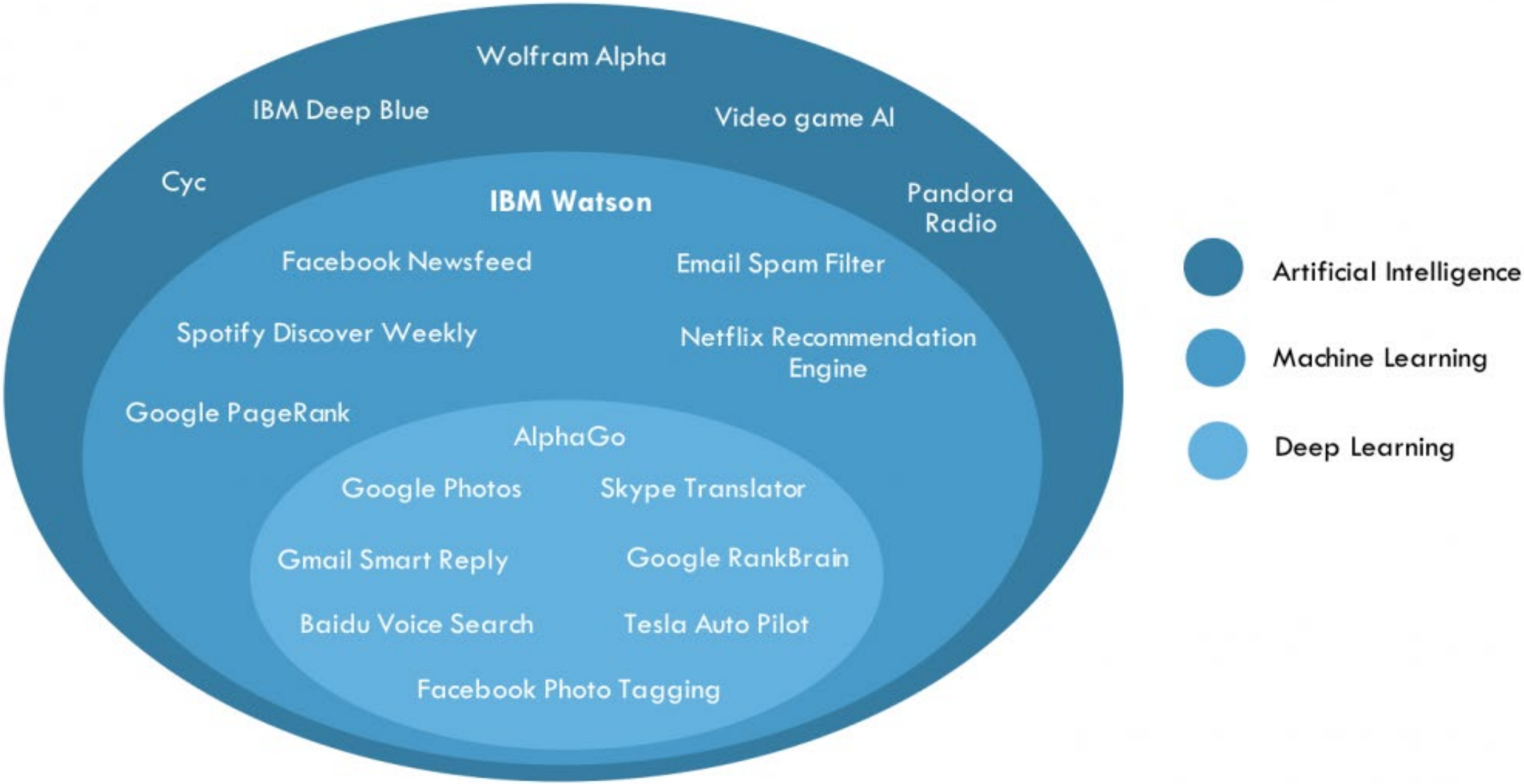
ARTIFICIAL INTELLIGENCE

- MACHINE LEARNING
  - PREDICTIVE ANALYTICS
  - DEEP LEARNING
- SPEECH
  - TEXT TO SPEECH
  - SPEECH TO TEXT
- VISION
  - IMAGE RECOGNITION
  - MACHINE VISION
- LANGUAGE PROCESSING (NLP)
  - CLASSIFICATION
  - TRANSLATION
  - DATA EXTRACTION
- EXPERT SYSTEMS
- PLANNING & OPTIMIZATION
- ROBOTICS

G2 CROWD


Artificial Intelligence
• IBM Deep Blue Chess Program
• Electronic Game Characters (Sims)
Machine Learning
• IBM Watson
• Google Search Algorithm
• Amazon Recommendations
• Email SPAM filter
Deep Learning
• AlphaGo
• Natural Speech Recognition
• Waymo Level 4 Automated Driving System

Artificial Intelligence Categories



Enter Data Scientists

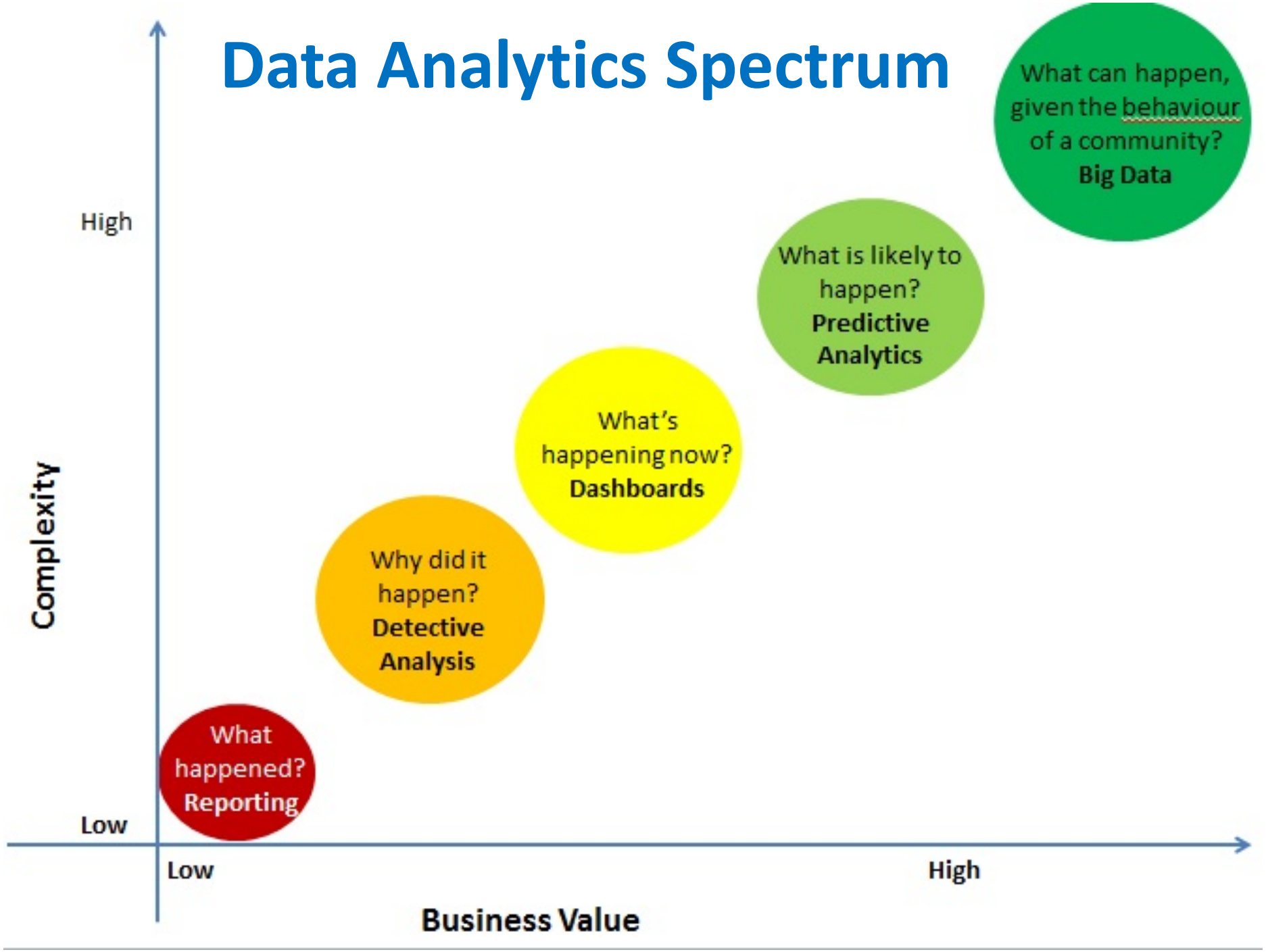
A Business analyst is not able to discover insights from huge sets of data of different domains.

Data scientists can work in co-ordination with different verticals of an organization and find useful patterns/insights for a company to make tangible business decisions.

15,000%

INCREASE IN JOB POSTINGS FOR DATA SCIENTISTS IN THE US BETWEEN 2011-12

Data Analytics Spectrum

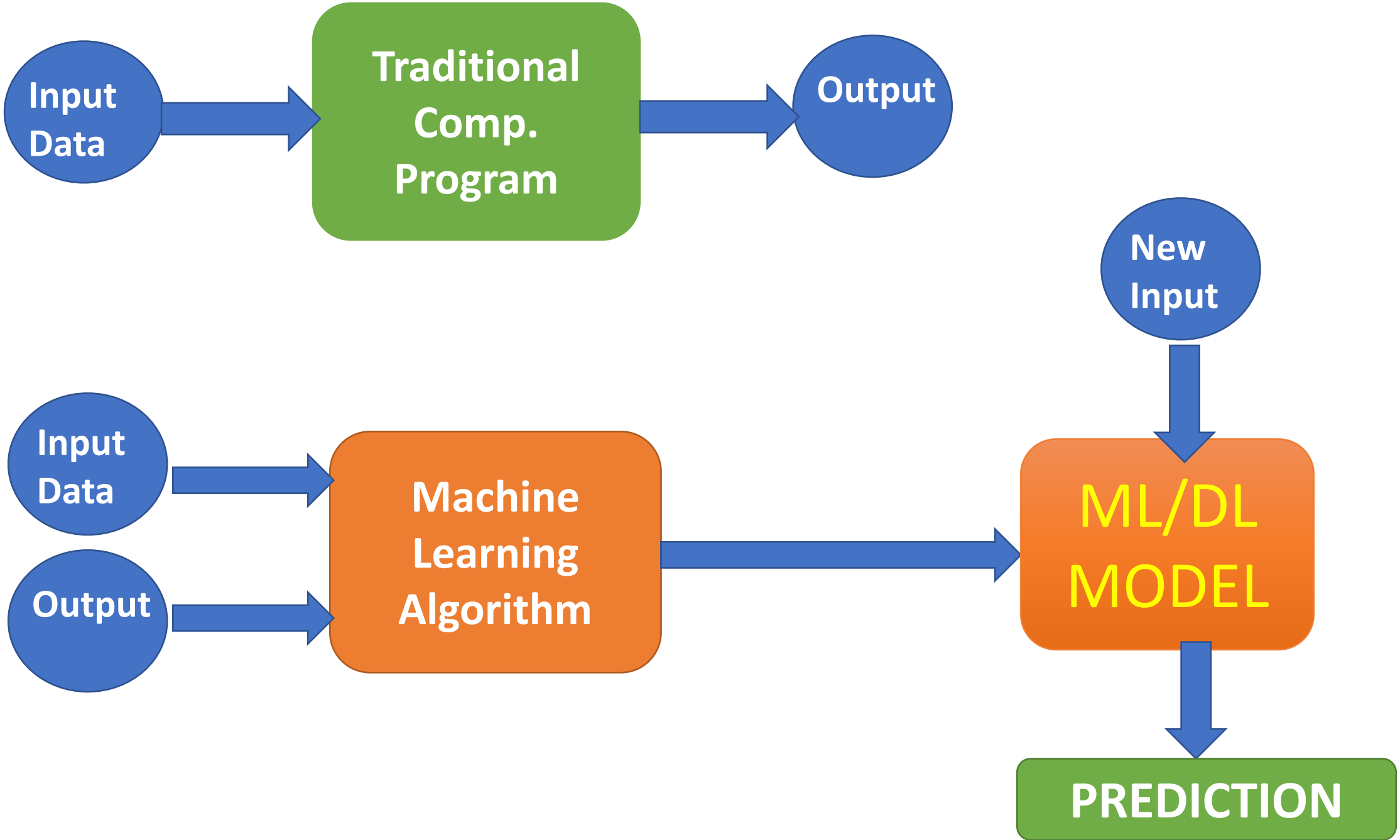


So what exactly is machine learning?

- ***Machine learning*** teaches computers to do what comes naturally to humans: **learn from experience.**

Machine learning algorithms use computational methods to "learn" information directly from **data** without relying on a predetermined equation as a model.

The algorithms adaptively improve their performance as the number of samples available for learning increases.


Input Data
Traditional Comp. Program
Output
New Input
Input Data
Output
Machine Learning Algorithm
ML/DL MODEL
PREDICTION

What Problem are we Solving ???

Classification: Is this A or B?

Regression: How much or how many?

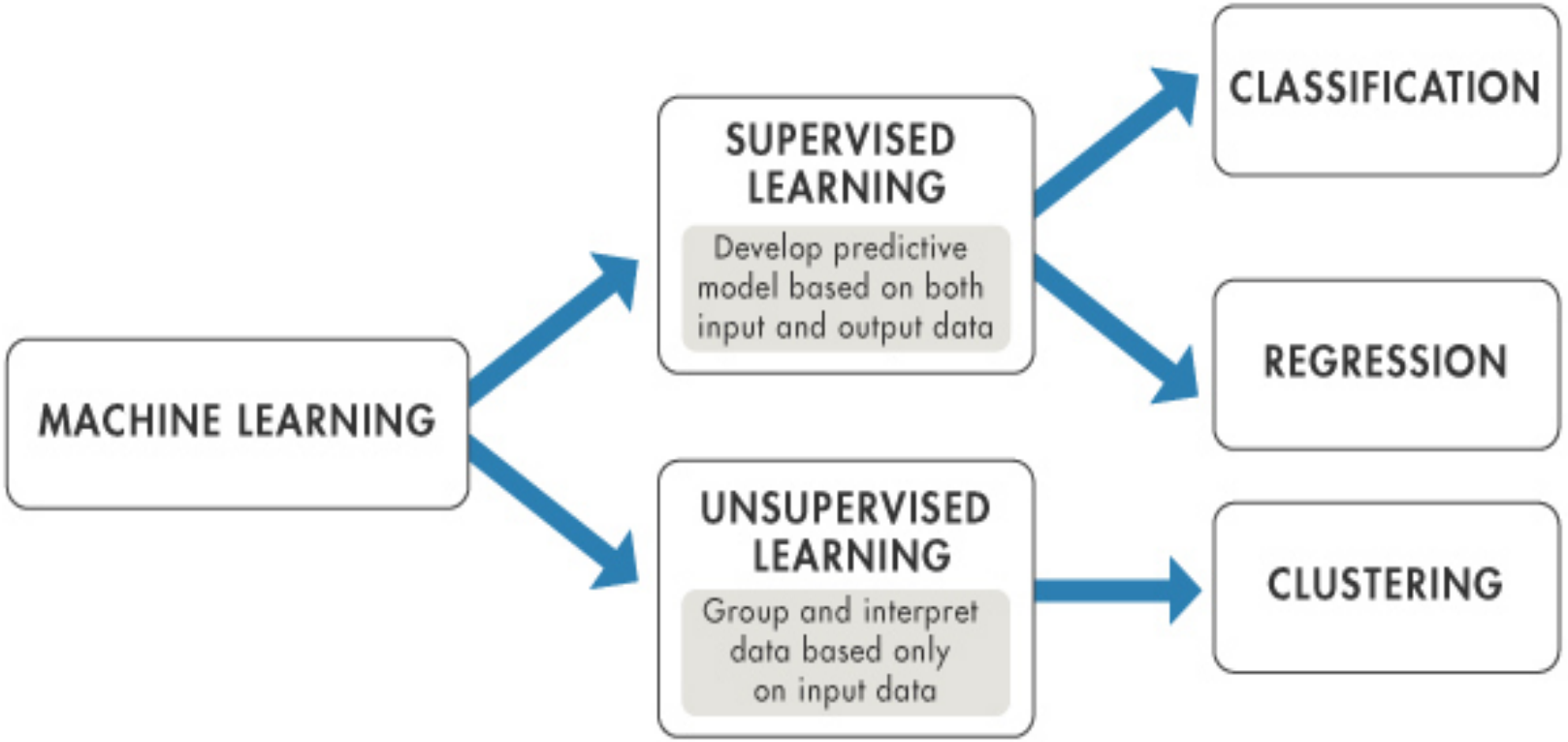
Clustering: How is this organized?

Machine learning types

Machine learning uses two types of techniques:

supervised learning, which trains a model on known input and output data so that it can predict future outputs, and

unsupervised learning, which finds hidden patterns or intrinsic structures in input data.



Classification

In machine learning, *classification* is identifying to which set of categories a new observation belongs based on the set of training data containing in the observed categories. Here are some examples of classification problems:

- Predicting the winner for the U.S. 2020 Presidential Election
- Predicting if a tumor is cancerous
- Classifying the different types of flowers

A classification problem with two classes is known as a *two-class classification* problem. Those with more than two classes are known as *multi-class classification* problems.

CLASSIFICATION EXAMPLES

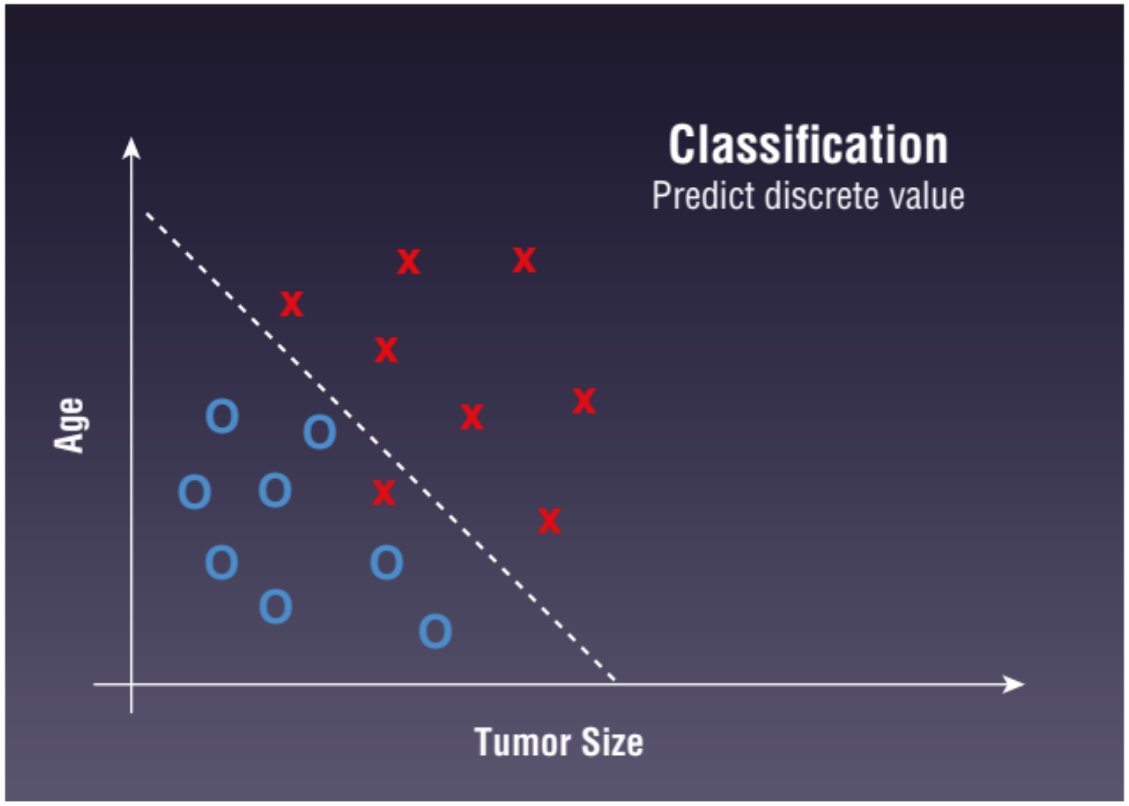
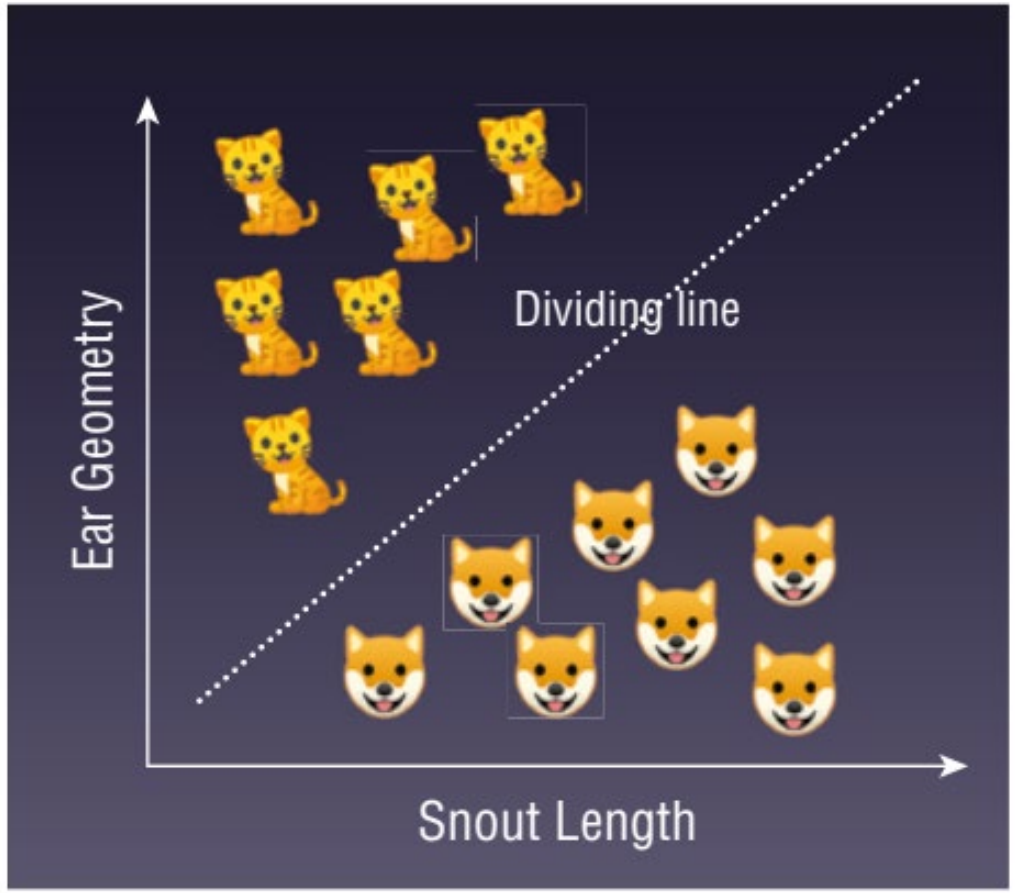


Figure 1.4: Using classification to categorize data into distinct classes



Cat – Dog Classifier

TITANIC DATASET

X (Predictor Variables/Features/Attributes)							Target Class
X1	X2	X3	X4	X5	X6	X7	Y
Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Survived
3	male	22	1	0	7.25	S	0
1	female	38	1	0	71.2833	C	1
3	female	26	0	0	7.925	S	1
1	female	35	1	0	53.1	S	1
3	male	35	0	0	8.05	S	0
3	male	25.4	0	0	8.4583	Q	0
1	male	54	0	0	51.8625	S	0
3	male	2	3	1	21.075	S	0
3	female	27	0	2	11.1333	S	1
2	female	14	1	0	30.0708	C	1
3	female	4	1	1	16.7	S	1

Regression

Regression helps in forecasting the future by estimating the relationship between variables. Unlike classification (which predicts the class to which an observation belongs), regression returns a continuous output variable. Here are some examples of regression problems:

- Predicting the sales number for a particular item for the next quarter
- Predicting the temperatures for next week
- Predicting the lifespan of a particular model of tire

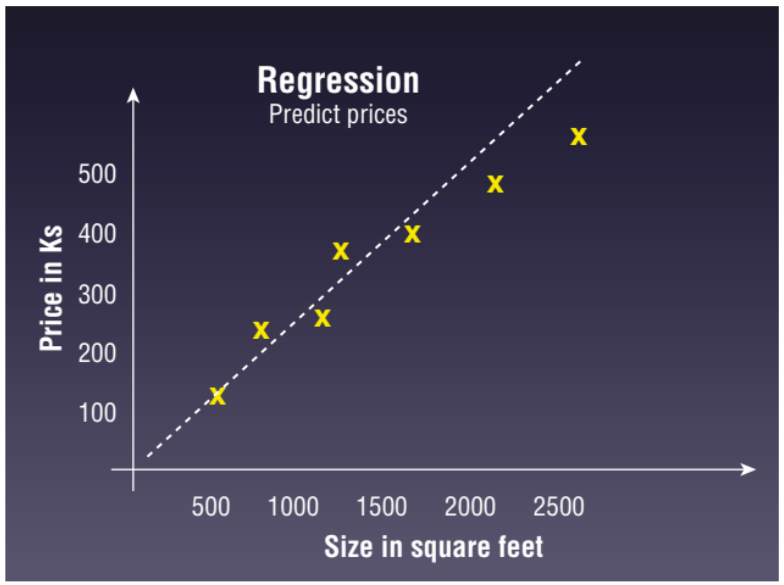


Figure 1.3: Using regression to predict the expected selling price of a house

Boston House Prices Dataset

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2

Clustering

Clustering helps in grouping similar data points into intuitive groups. Given a set of data, clustering helps you discover how they are organized by grouping them into natural clumps.

Examples of clustering problems are as follows:

- Which viewers like the same genre of movies
- Which models of hard drives fail in the same way

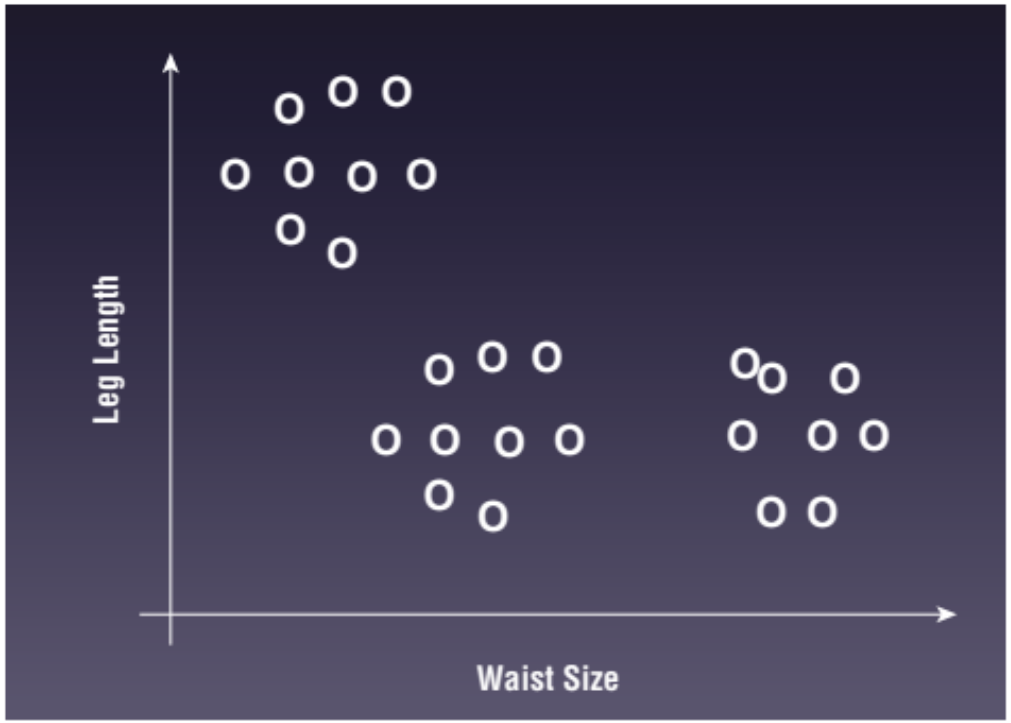


Figure 1.5: Plotting the unlabeled data

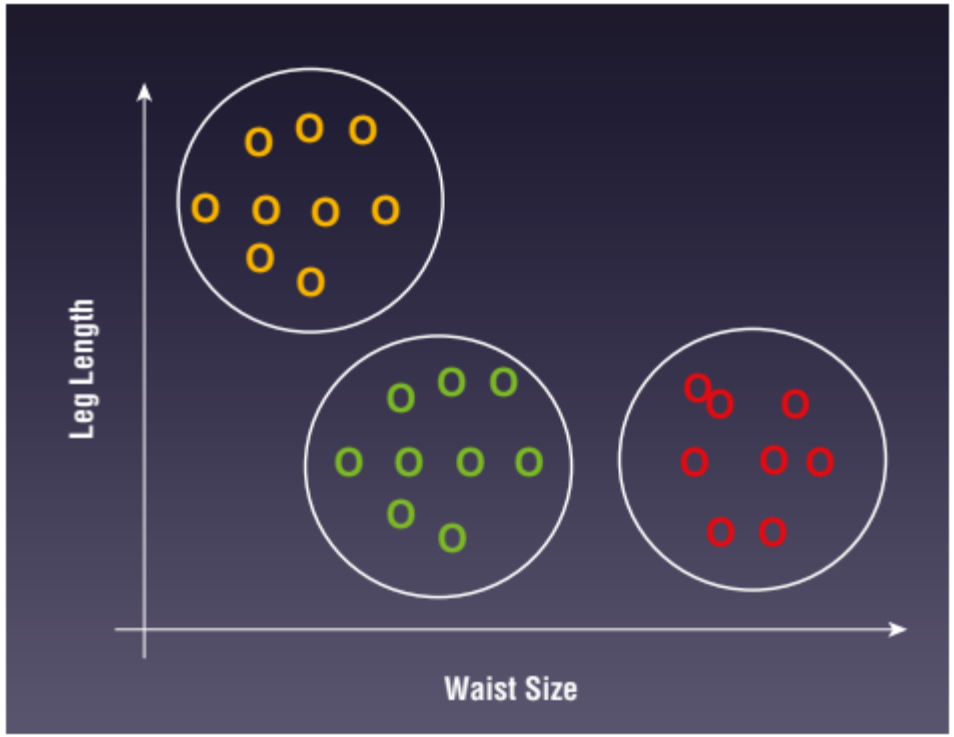
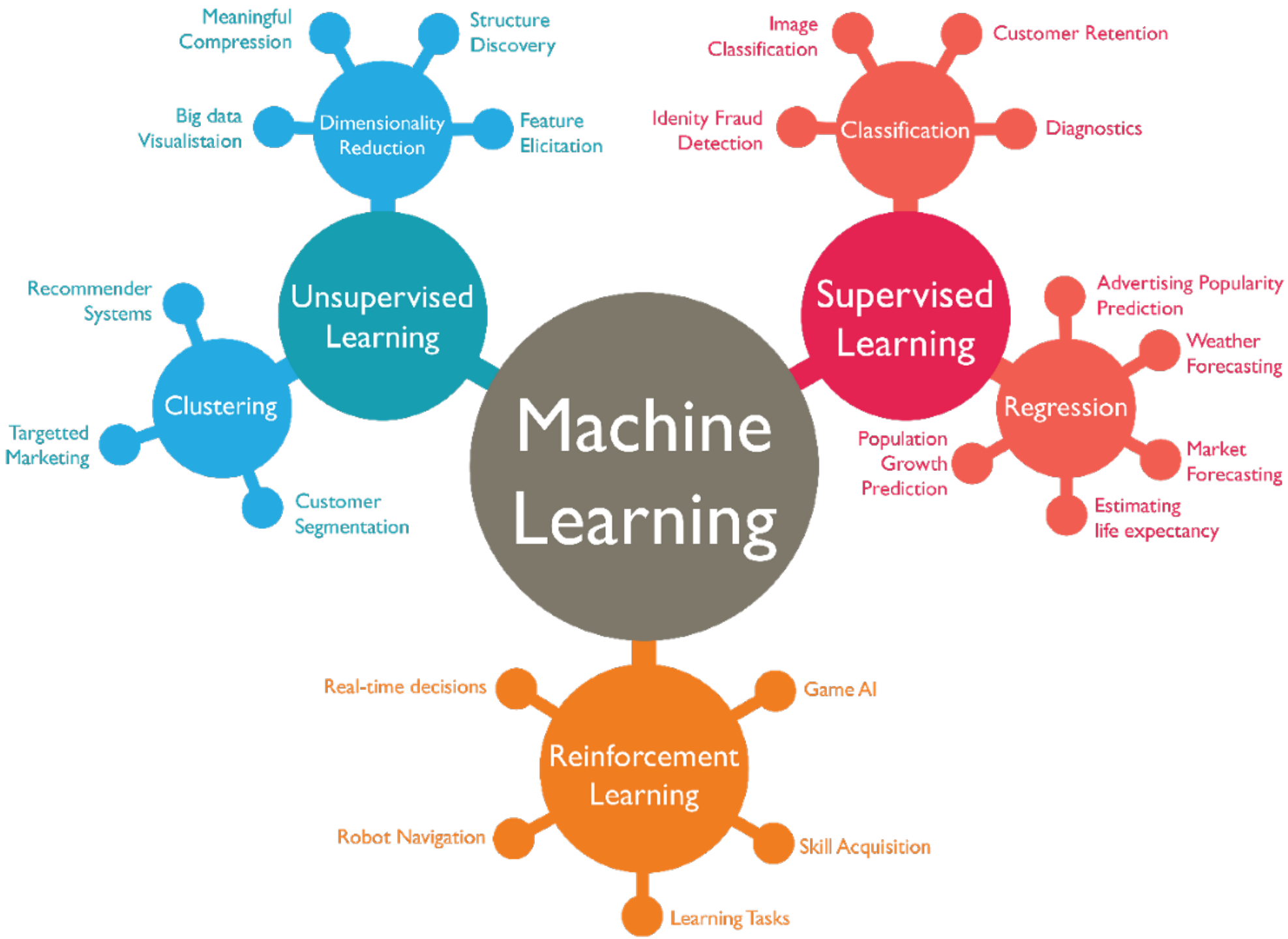


Figure 1.6: Clustering the points into distinct groups

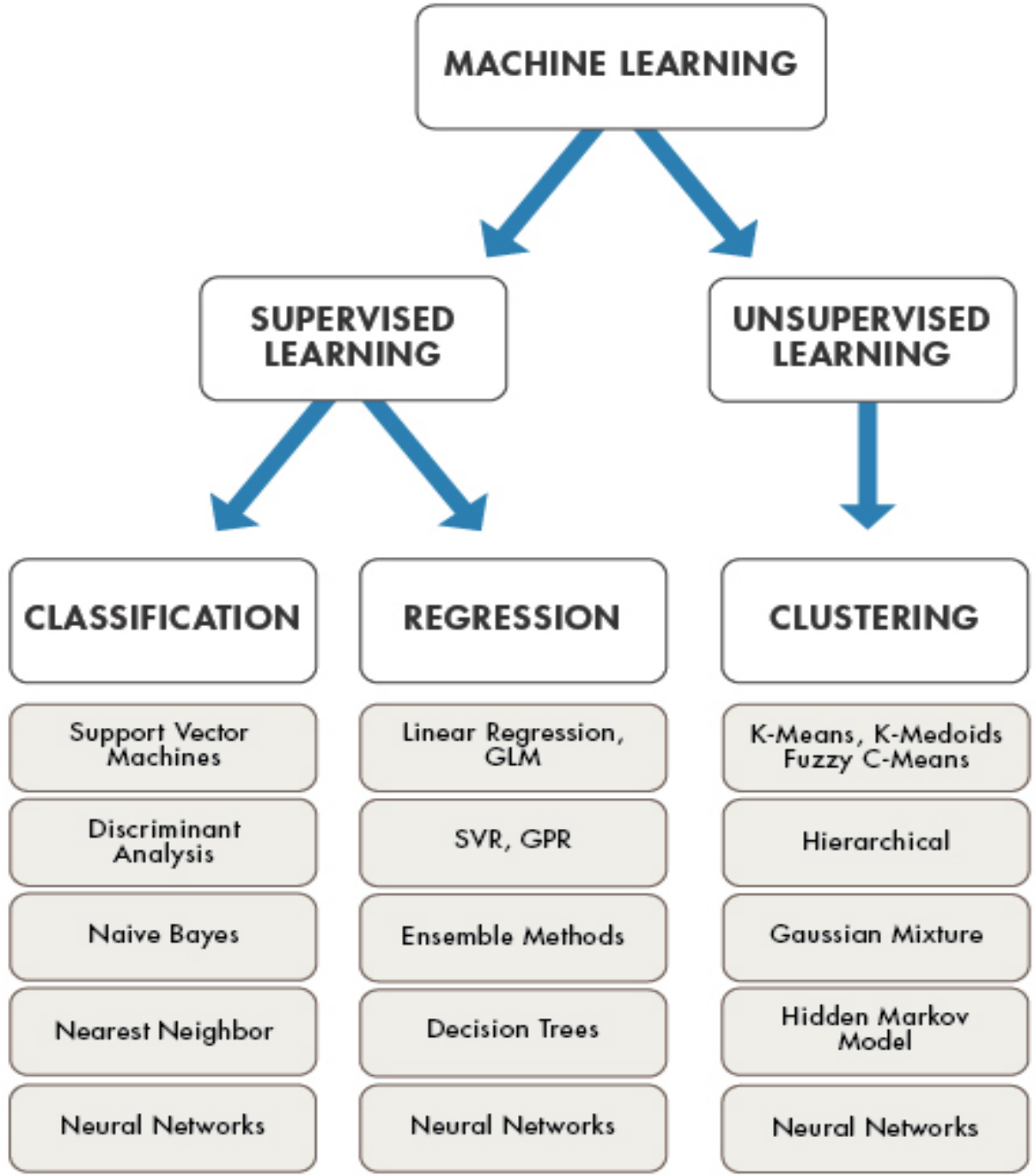

Machine Learning
Unsupervised Learning
Supervised Learning
Reinforcement Learning
Dimensionality Reduction
Clustering
Classification
Regression
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Recommender Systems
Targetted Marketing
Customer Segmentation
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Supervised machine learning

- ***Classification*** techniques predict categorical responses, for example, whether an email is genuine or spam, or whether a tumor is cancerous or benign. Classification models classify input data into categories.
  - Typical applications include medical imaging, image and speech recognition, and credit scoring.
- ***Regression*** techniques predict continuous responses, for example, changes in temperature or fluctuations in power demand.
  - Typical applications include electricity load forecasting and algorithmic trading.

Unsupervised learning

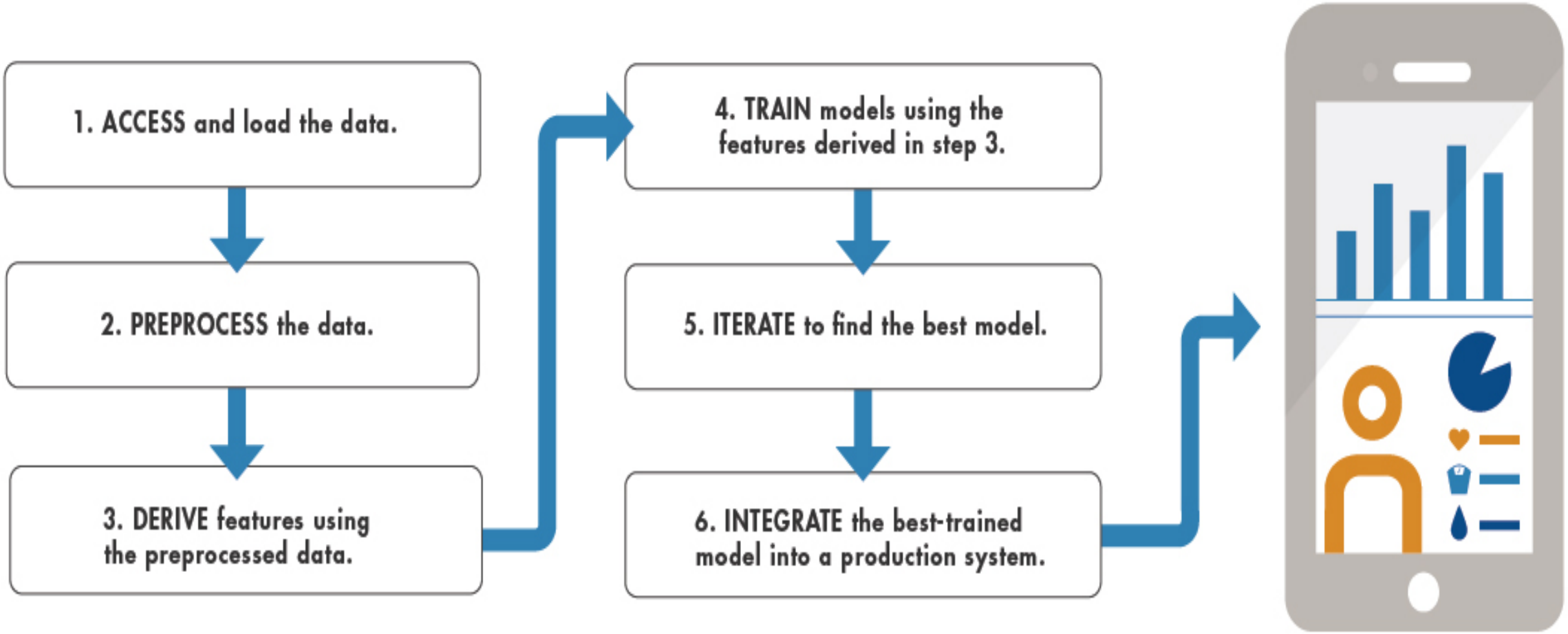
- **Unsupervised learning finds hidden patterns or intrinsic structures in data.**
- It is used to draw inferences from datasets consisting of input data without labelled responses.
- ***Clustering*** is the most common unsupervised learning technique.
- It is used for exploratory data analysis to find hidden patterns or groupings in data.
- Applications for clustering include gene sequence analysis, market research, and object recognition.


MACHINE LEARNING
SUPERVISED LEARNING
UNSUPERVISED LEARNING
CLASSIFICATION
REGRESSION
CLUSTERING
Support Vector Machines
Discriminant Analysis
Naive Bayes
Nearest Neighbor
Neural Networks
Linear Regression, GLM
SVR, GPR
Ensemble Methods
Decision Trees
Neural Networks
K-Means, K-Medoids Fuzzy C-Means
Hierarchical
Gaussian Mixture
Hidden Markov Model
Neural Networks

Selecting the Right Algorithm

- There are dozens of supervised and unsupervised machine learning algorithms, and each takes a different approach to learning.
- **There is no best method or one size fits all.** Finding the right algorithm is partly based on trial and error—even highly experienced data scientists cannot tell whether an algorithm will work without trying it out.
- **Highly flexible models** tend to overfit data by modeling minor variations that could be noise.
- **Simple models** are easier to interpret but might have lower accuracy.
- Therefore, choosing the right algorithm requires trading off one benefit against another, including model speed, accuracy, and complexity.
- **Trial and error** is at the core of machine learning—if one approach or algorithm does not work, you try another.

Systematic Machine Learning Workflow



Tech stack requirements

- Tools & programming languages

- Python with Anaconda, Enthought CanoPy, Pycharm >>> IDEs

Python Packages for ML & DL

- NumPy, SciPy (for Optimization)
- Pandas for Data Analysis, pandas-profiling
- Data Visualization: matplotlib, Pandas, Seaborn~~, Cufflinks, Plotly~~
- ~~Web Scraping: BeautifulSoup, Scrapy~~
- ~~Advanced Statistics: Statsmodels~~
- Machine Learning Algorithms: **sklearn**, ~~xgboost~~
- ~~Natural Language Processing (NLP): nltk, Spacy, Gensim~~
- ~~Deep Learning: TensorFlow, Keras, Theano~~
- ~~Deployment: Flask, Django~~

Python Packages for ML & DL

- NumPy, SciPy (for Optimization)
- Pandas for Data Analysis, pandas-profiling
- **Data Visualization:** matplotlib, Pandas, Seaborn, Cufflinks, Plotly
- **Web Scraping:** BeautifulSoup, Scrapy
- **Advanced Statistics:** Statsmodels
- **Machine Learning Algorithms: sklearn**, xgboost
- **Natural Language Processing (NLP): nltk**, **Spacy**, **Gensim**, Textblob
- **Deep Learning: TensorFlow**, **Keras**, Theano, **PyTorch**,
- **Deployment: Flask**, Django

Application Areas

Financial :

- Financial data, collected in the banking and financial industry are often relatively complete, reliable, and of high quality, which facilitates systematic data analysis and data mining.
- Loan payment prediction and customer credit policy analysis.
- Classification and clustering of customers for targeted marketing.
- Detection of money laundering and other financial crimes.
- Automated trading systems
- Credit Scoring
- Risk management
- Underwriting
- Fraud detection
- Collection analytics

Application Areas

Retail Industry:

The retail industry is a major application area for data mining, since it collects huge amounts of data on sales, customer shopping history, goods transportation, consumption and service. Retail data mining can help identify customer buying behaviors, discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios, design more effective goods transportation and distribution policies, and reduce the cost of business.

- Multidimensional analysis of sales, customers, products, time, and region.
- Analysis of the effectiveness of sales campaigns.
- Customer retention—analysis of customer loyalty.
- Product recommendation and cross-referencing of items.
- Market basket analysis

Application Areas

Telecommunication Industry:

The telecommunication industry provides local and long distance telephone services and many other comprehensive communication services, including fax, pager, cellular phone, Internet messenger, images, e-mail, computer and Web data transmission, and other data traffic. Data Mining Applications:

- Multidimensional analysis of telecommunication data.
- Fraudulent pattern analysis and the identification of unusual patterns.
- Multidimensional association and sequential pattern analysis.
- Mobile telecommunication services.
- Use of visualization tools in telecommunication data analysis .
- Tools for OLAP visualization, linkage visualization, association visualization, clustering, and outlier visualization have been shown to be very useful for telecommunication data analysis.

Application Areas

Biological Data Analysis:

•Explosive growth in genomics, proteomics, functional genomics, and biomedical research. Examples are the identification and comparative analysis of the genomes of human and other species (by discovering sequencing patterns, gene functions, and evolution paths), the investigation of genetic networks and protein pathways and the development of new pharmaceuticals and advances in cancer therapies. This is a new research field called bioinformatics.

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search, and comparative analysis of multiple nucleotide/protein sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis:
- identifying co-occurring gene sequences and linking genes to different stages of disease development.
- Visualization tools in genetic data analysis: Alignments among genomic or proteomic sequences and the interactions among complex biological structures are most effectively presented in graphic forms, transformed into various kinds of easy-to-understand visual displays. Such visually appealing structures and patterns facilitate pattern understanding, knowledge discovery, and interactive data exploration.

Application Areas

Intrusion Detection:

•Development of data mining algorithms for intrusion detection

•Association and correlation analysis, and aggregation to help select and build discriminating attributes:

•Analysis of stream data:

•Distributed data mining:

•Visualization and querying tools: Visualization tools should be available for viewing any anomalous patterns detected. Such tools may include features for viewing associations, clusters, and outliers. Intrusion detection systems should also have a graphical user interface that allows security analysts to pose queries regarding the network data or intrusion detection results.

Application Areas

Classification by Object Recognition in Satellite Images:

•Remotely sensed images have a lot of geographical information inside. Geographical information can be useful for different sectors like government, business, science, engineering and research institutes. An automatic mechanism is provided based on object recognition by Data Mining to extract objects from the image and then do classification of the image.

•Geographical information can be used :

- •for city and regional planning,
- •analysis of natural resources in an area and
- • improvement in vegetation of an area.
- •Agricultural Monitoring /Modeling e.g., cultivated vs. non-cultivated areas, types of crops, varieties, ground deformations
- •mining satellite images may help detect forest fire,
- • find unusual phenomena on earth,
- •predict hurricane landing site, discover weather patterns, and
- •outline global warming trends.

Application Areas

Business:

Most business processes in most organizations have the potential to benefit from predictive modeling. That said, there are certain situations where predictive models can be especially beneficial in delivering a great deal of value:

- Processes that require a large number of similar decisions
- Where the outcomes have a significant impact, i.e., where there's a lot at stake in terms of money or lives
- Where there's abundant information in electronic data form available on which to base decisions and measure outcomes
- Where it's possible to insert a model calculation into the actual business process, either to automate decisions or to support human decision makers
- CRM related: Analytical customer relationship management (CRM) ,Clinical decision support systems ,Cross-sell, Customer retention ,Direct marketing.
- Portfolio, product or economy-level prediction

Application Areas

Automated controls and recognition include :

- the system identification and control (vehicle control, process control, natural resources management),
- quantum chemistry,
- game-playing and decision making (backgammon, chess, poker),
- pattern recognition (radar systems, face identification, object recognition and more),
- sequence recognition (gesture, speech, handwritten text recognition),
- Natural Language processing, text mining and e-mail spam filtering.

Application Areas

Web searching/ mapping/ social networking:

•Web Content Mining, Web Structure Mining, and Web Usage Mining.

•Change in our lives in the decade following the turn of the century was the availability of efficient and accurate Web search, through search engines such as Google.

•Much information is gained by analyzing the large-scale data derived from social networks. It is important in a social network is how to identify communities, i.e., subsets of the nodes (people or other entities that form the network) with unusually strong and usually overlapping connections.

•Social networks are naturally modeled as graphs, which are called social graph. Types: Telephone, Email , Collaboration , information (documents, web graphs, patents), infrastructure (roads, planes, water pipes, power grids), biological (genes, proteins, food-webs of animals eating each other), as well as other types, like product co-purchasing(e.g., Group on) networks.

Application Areas

Social Network Analytics

•When we think of a social network, we think of Facebook, Twitter, Google+, or another website that is called a "social network." Characteristics are:

- •There its a collection of entities that participate in the network.
- •There is at least one relationship between entities of the network.
- •There is an assumption of non randomness or locality.
- •broad applications, are
  - • social network analysis, web community discovery,
  - •terrorist network mining,
  - •computer network analysis, and
  - • network intrusion detection.

What does a Data Scientist do?

- Data collection
- Data preparation
- Exploratory data analysis (EDA)
- Evaluating and interpreting EDA results
- Model building
- Model testing
- Model deployment
- Model optimization

The above is iterative, meaning a data scientist will be in "evaluation mode" throughout the entire process.

Data Scientist job description -1

- Collecting massive amounts of data and converting it to an analysis-friendly
- Problem-solving business-related challenges while using data-driven techniques and tools.
- Using a variety of programming languages, as well as programs, for data collection and analysis.
- Having a wealth of knowledge with analytical techniques and tools.
- Communicating findings and offering advice through effective data visualizations and comprehensive reports.

Data Scientist job description -2

- Identifying patterns and trends in data; providing a plan to implement improvements.
- Predictive analytics; anticipate future demands, events, etc.
- Contribute to data mining architectures, modeling standards, reporting and data analysis methodologies.
- Invent new algorithms to solve problems and build analytical tools.
- Recommend cost-effective changes to existing procedures and strategies.

Data Scientist Skill Set -1

- ***Experience and Fluency in many of these computer/coding programs*:** SAS, SPSS, MATLAB R, Python, Java, C/C++, Hadoop Platform, SQL/NoSQL Databases.
- ***Business Savviness*:** Data scientists need to understand the business sector they are working in and create solutions to complex problems that align with business logic/objectives.
- ***Communication skills*:** A data scientist can clearly and fluently translate their technical and analytical findings to a non-technical department.

Data Scientist Skill Set -2

- ***Expert Technical skills*** in the following: Math (g., linear algebra, calculus, and probability)
  - Statistics
  - Machine learning tools and techniques
  - Data mining
  - Data cleaning and munging
  - Data visualization and reporting techniques
  - Unstructured data techniques

Six steps to become a Data Scientist - 1

Step 0: Find out if it's really for you.

Step 1: Early Preparation: Programming languages, Python / R / MATLAB and basic machine learning stuff.

Step 2: Complete undergraduate studies: Add Database architecture, and SQL/MySQL to your skills set.

Step 3: Obtain an entry-level job: Junior Data Analyst or Junior Data Scientist, data engineer etc..

Spend time in excelling your ML/DL knowledge.

Six steps to become a Data Scientist - 2

- **Step 4: Earn a Master's Degree or Ph.D.: Learn** how to use enterprise-grade data management programs and how distributed storage and computation operate (e.g., Hadoop, MapReduce, and Spark) in relation to model building and predictive analytics.
- **Step 5: Get promoted (or actually hired as a Data Scientist):** Coupling strong technical skills with project management and leadership experience.
  - Learn Deployment of large scale projects, Business Analytics (which will require domain expertise as well).
- **Step 6: Never Stop Learning**

Top Blogs to Follow

- Medium.com
- AnalyticsVidhya
- KDNuggets
- TowardsDataScience
- DataScienceCentral
- MachineLearningMastery
- PyImageSearch (Adrian)

Books

- - ESL, ISL >> Trevor Hastie, Tibshirani, J. Firedman
- - ML by Tom Mitchell
- - Pattern Recognition in Machine Learning >> Chistropher M. Bishop
- - Machine Learning: A Probabilistic Perspective >> Kevin Murphy

Books

- Deeplearningbook.com > Ian GoodFellow
- **Pandas for Everyone - Daniel Chen (Pearson)**
- Deep Learning With Applications Using Python >> Navin Manaswi (Apress)
- **Sebastian Raschka >> Python Machine Learning >> Packt**
- - scikit-learn cookbook >> Trent Hauck >> Packt
- Machine Learning with Python Cookbook: Chris Albon >> - O'Reilley
- - Introduction to Machine Learning with Python: A Guide for Data Scientists- by Andreas Muller - O'Reilley
- - Statistics for Machine Learning – Pratap Dangeti (Packt)
- - Business Analytics – Prof. U. Dinesh Kumar (Wiley) >> ML in Python

Prashant Sahu

Freelance Corporate Trainer for
Data Analytics | Machine Learning...

Thank You

https://www.linkedin.com/in/prashantksahu