DS 203
COURSE OVERVIEW AND BACKGROUND

Semester 1
2024-2025

DS 203

Assignments	• 8 - 10 • Every submission will be reviewed for completeness and correctness • There will be penalties for late / no / fraudulent submissions
Evaluation scheme	• 10% : 2 surprise quiz • 30% : Mid-semester test • 30% : Project (Group Project, max 4 members) • 30% : End-semester
Penalties	• (-1) : Late / non submission of assignments (each event) • (-10) : Copying assignments / project (each event) • (-10) : Fraudulent assignment submissions (each event) • The penalty will be in addition to zero credit (where applicable) for that particular submission
Attendance	• As per IITB rules for attendance (self regulated) • No attendance will be taken in class, except during quiz / test / examination

SAFE Application

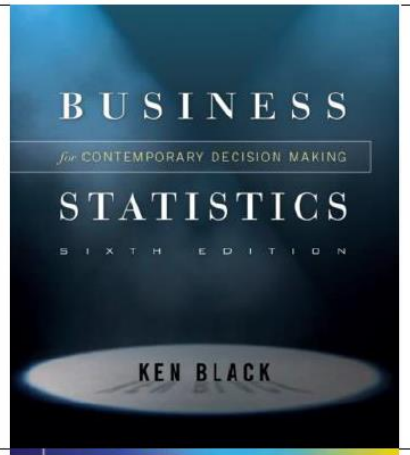
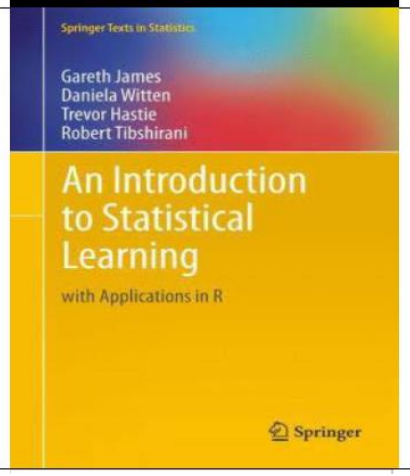
- SAFE app will be used for conducting quiz, and for marking attendance during the quiz
- All participants **should** compulsorily register themselves on SAFE app using the following registration code:

20FLXYSI

- You may mark your class attendance using the app (for your own records)
- In case of teething troubles, write to safe@iitb.ac.in and visit their office in the CC to get them resolved

Books and References

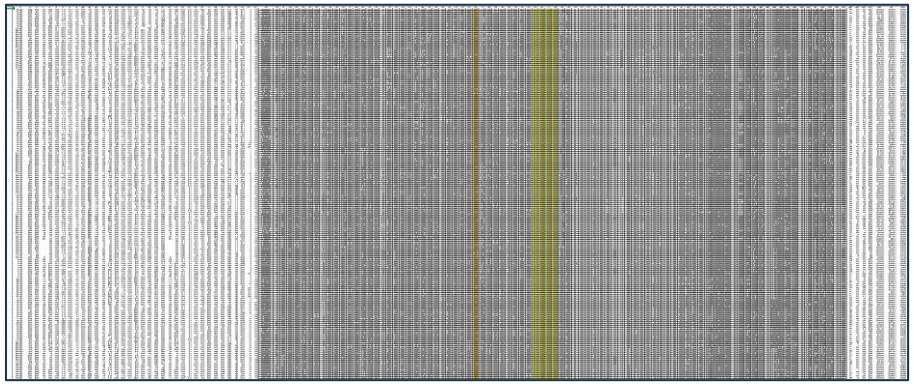
- Learning Data Science : https://learningds.org/intro.html
- Veridical Data Science : https://vdsbook.com/

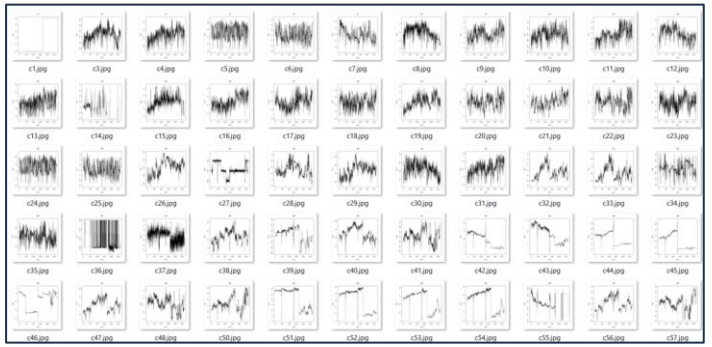
For topics related to Statistics	Business Statistics for Contemporary Decision Making Author: Ken Black (Available online – 6th Edition)	
For topics related to Machine Learning	An Introduction to Statistical Learning Authors: Gareth James and others (Available online)	

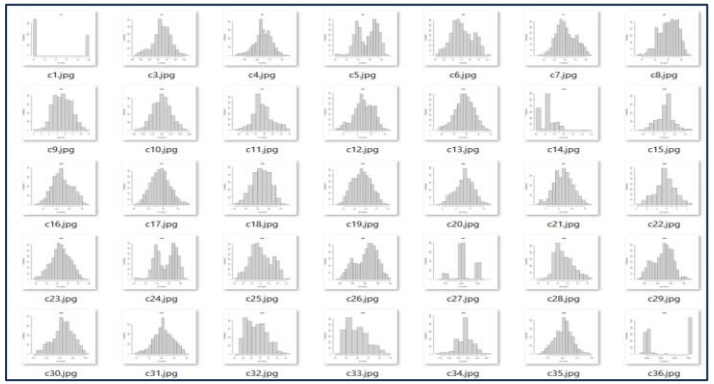
If you have questions … today

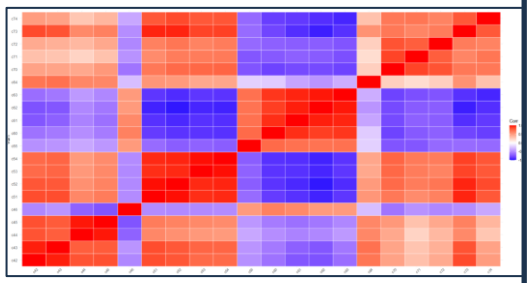
https://tinyurl.com/ds203-2024-q

Goal of Data Science : **Extract Insights from Data**





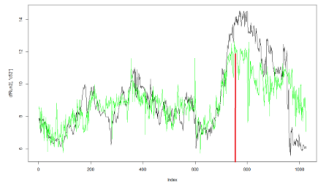




Training: First 750 data points

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.209642   5.592980  -2.898 0.003866 **
c161          0.011909   0.001154  10.317  < 2e-16 ***
c28           0.192089   0.031109   6.175 1.10e-09 ***
c7            1.762212   0.244322   7.213 1.38e-12 ***
c17          -0.099812   0.019082  -5.231 2.21e-07 ***
c158          0.123261   0.018093   6.813 2.01e-11 ***
c160          0.004787   0.001771   2.703 0.007041 **
c39           6.898476   1.100166   6.270 6.18e-10 ***
c22          -0.104352   0.034339  -3.039 0.002460 **
c11          -0.122940   0.036342  -3.383 0.000756 ***
c15          -0.371219   0.056582  -6.561 1.02e-10 ***
c30           1.702216   0.337681   5.041 5.85e-07 ***
c23          -0.274084   0.039142  -7.002 5.74e-12 ***
c35           5.033540   1.451099   3.469 0.000554 ***
c16          -0.381731   0.073873  -5.167 3.07e-07 ***
c139         -0.221037   0.036769  -6.011 2.91e-09 ***
c31           0.156248   0.022446   6.961 7.55e-12 ***
c143         -0.236297   0.034607  -6.828 1.82e-11 ***
c157          0.160972   0.038140   4.221 2.75e-05 ***
c163          0.009764   0.002779   3.513 0.000471 ***
c9           -0.267089   0.070883  -3.768 0.000178 ***
c8           -0.369612   0.122558  -3.016 0.002652 **
c10           3.208038   1.541097   2.082 0.037723 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8832 on 727 degrees of freedom
Multiple R-squared:  0.6309,	Adjusted R-squared:  0.6198
F-statistic: 56.49 on 22 and 727 DF,  p-value: < 2.2e-16
```

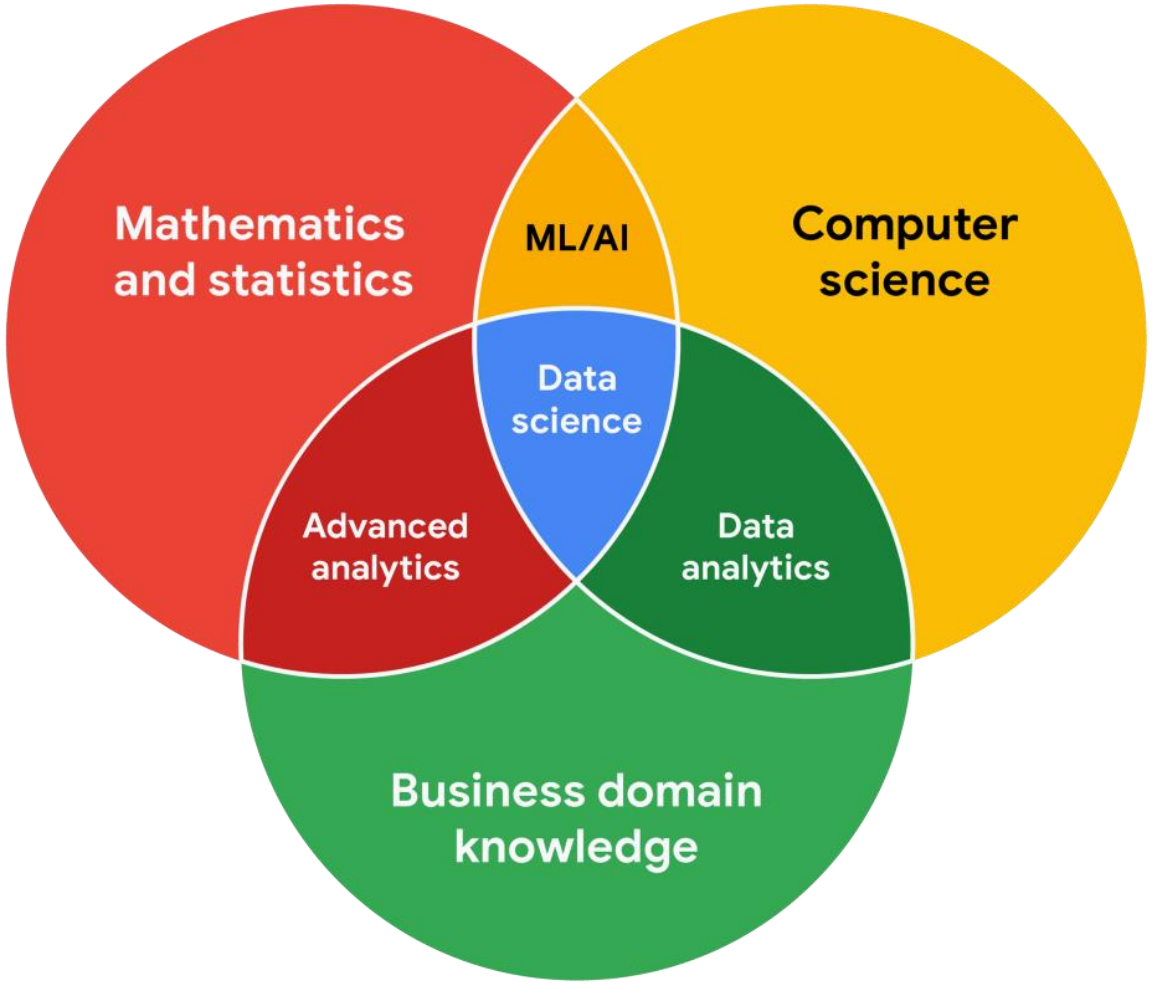


Positioning of DS 203

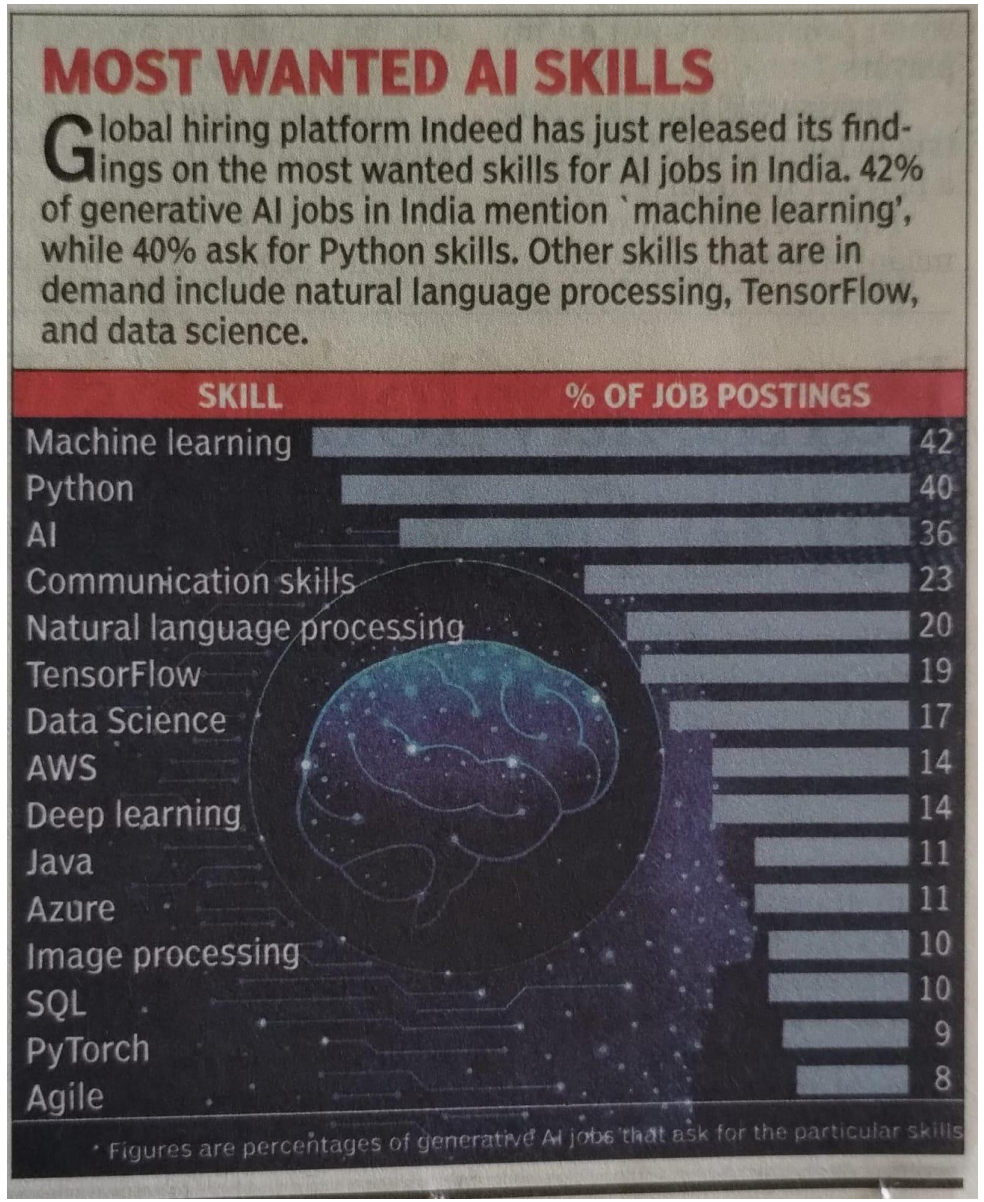
- Mastering Data Science
  - Statistical Foundations
  - Machine Learning Fundamentals (DS 303)
  - **Knowledge of Tools, and Programming (DS 203)**
  - **Domain Knowledge**
  - **Communication Skills**

- Important questions
  - Can you understand ML without Statistics?
  - ML and Programming – what is the connection?

Putting it all together !



The Essential Skills



MOST WANTED AI SKILLS

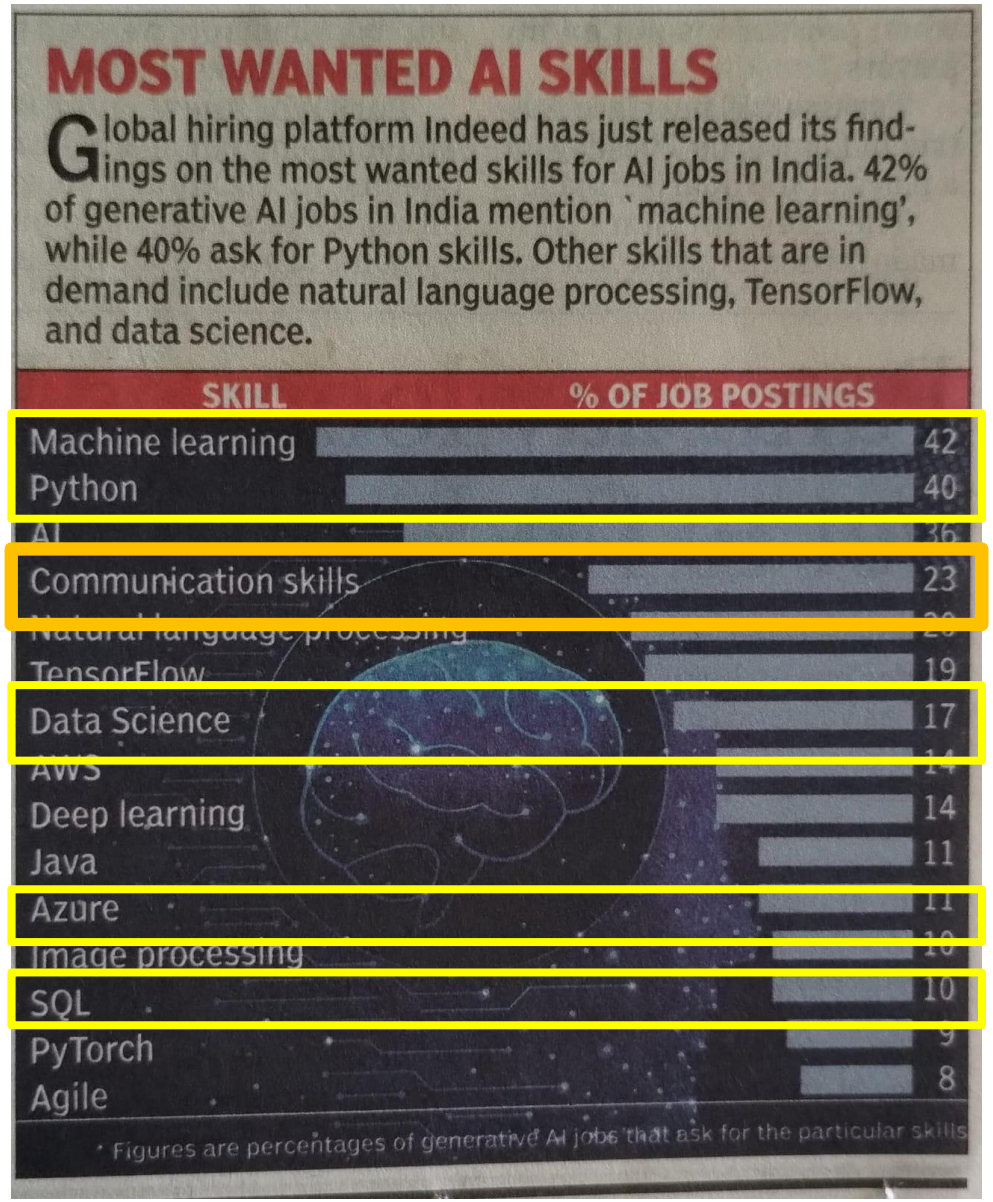
Global hiring platform Indeed has just released its findings on the most wanted skills for AI jobs in India. 42% of generative AI jobs in India mention `machine learning', while 40% ask for Python skills. Other skills that are in demand include natural language processing, TensorFlow, and data science.

SKILL	% OF JOB POSTINGS
Machine learning	42
Python	40
AI	36
Communication skills	23
Natural language processing	20
TensorFlow	19
Data Science	17
AWS	14
Deep learning	14
Java	11
Azure	11
Image processing	10
SQL	10
PyTorch	9
Agile	8

* Figures are percentages of generative AI jobs that ask for the particular skills

(TOI 29/06/2024)

The Essential Skills



MOST WANTED AI SKILLS

Global hiring platform Indeed has just released its findings on the most wanted skills for AI jobs in India. 42% of generative AI jobs in India mention `machine learning', while 40% ask for Python skills. Other skills that are in demand include natural language processing, TensorFlow, and data science.

SKILL	% OF JOB POSTINGS
Machine learning	42
Python	40
AI	36
Communication skills	23
Natural language processing	20
TensorFlow	19
Data Science	17
AWS	14
Deep learning	14
Java	11
Azure	11
Image processing	10
SQL	10
PyTorch	9
Agile	8

* Figures are percentages of generative AI jobs that ask for the particular skills

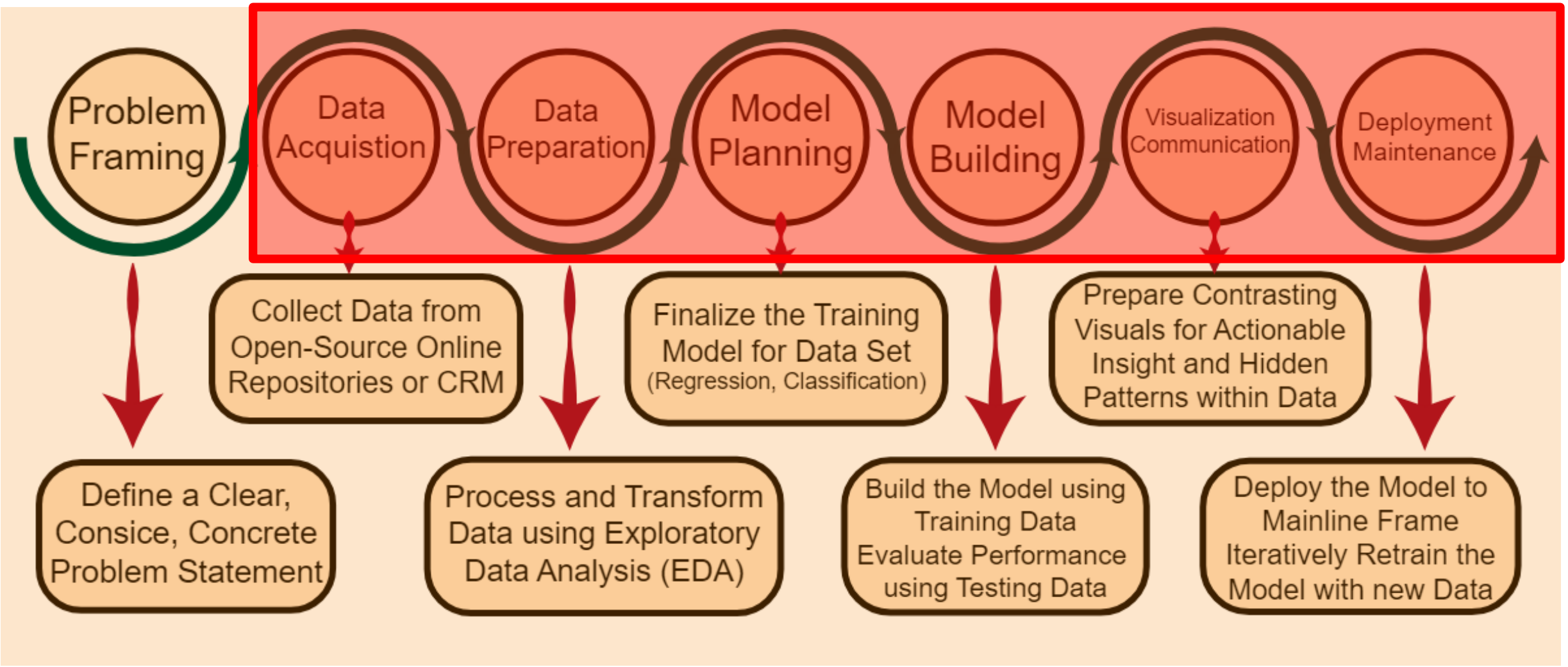
(TOI 29/06/2024)

Major Parts of DS 203

Part	Focus	Classes
1	Introduction to DS Basics of ML and Statistics Programming for ML and DS	8
2	Dealing with Data, visualization, pre-processing, transformation, Feature Engineering	8
3	Big Data tools and techniques Cloud Computing Database programming	8

Data Science Tools

Availability of tools



Programming for Data Science

- The nature of ‘programming’ has changed over generations
  - Machine, Assembly, C like, SQL like, AI powered
- Programming, conventionally
  - It is all about syntax, semantics, structure, optimization, etc.
- What does ‘programming’ mean, in the context of Data Science?
  - Understanding and effective use of the libraries / modules / packages
  - Understanding and correct use of various tools
  - Lately: Generating correct problem descriptions for LLMs
    - Prompt engineering

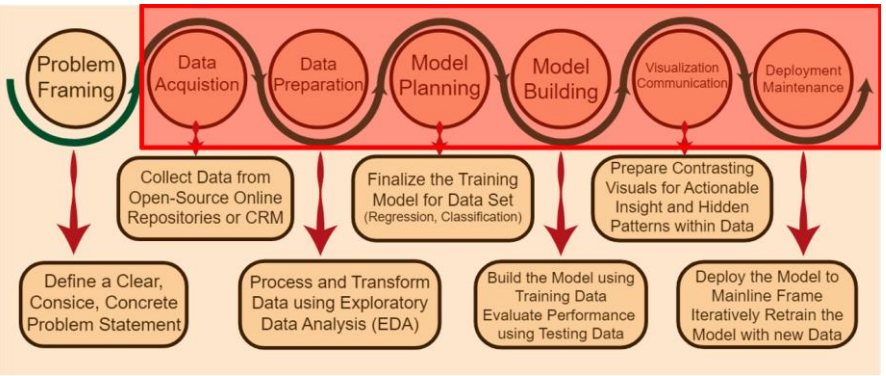
Introduction to Python for DS and ML

- 01-python-variables-types-and-basic-io.ipynb
- 02-more-basic-types.ipynb
- 03-program-flow-and-control.ipynb
- 04 -SLR-using-basic-python.ipynb
- 05-numpy.ipynb
- 06-matplotlib.ipynb
- 07-pandas.ipynb

… and many more

Category of tools

- Specific, standalone tools
  - Addressing specific parts of the entire process
- Integrated, on-premise tools
  - Supporting most steps of the entire cycle
- Cloud-based integrated solutions
  - Supporting most steps of the entire cycle
- **No-Code** solutions
  - For ML modelling
- Solutions for specific verticals & requirements
  - Fraud detection, sentiment analysis, etc.



Plethora of tools … (by no means complete!)

- Machine Learning
- Computer Vision
- Edge AI for Smart Devices and Machinery
- Predictive Analytics
- Data Visualization
- Big Data Engineering
- Open API
- Cloud Technologies
- Robotic Process Automation
- Sentiment Analysis
- Natural Language Processing
- DataOps, MLOps, and DevOps
- Blockchain and Decentralized Ledger Technologies

Plethora of tools … (by no means complete!)

- Machine Learning
- Computer Vision
- Edge AI for Smart Devices and Machinery
- Predictive Analytics
- Data Visualization
- Big Data Engineering
- Open API
- Cloud Technologies
- Robotic Process Automation
- Sentiment Analysis
- Natural Language Processing
- DataOps, MLOps, and DevOps
- Blockchain and Decentralized Ledger Technologies

Immediate Tasks

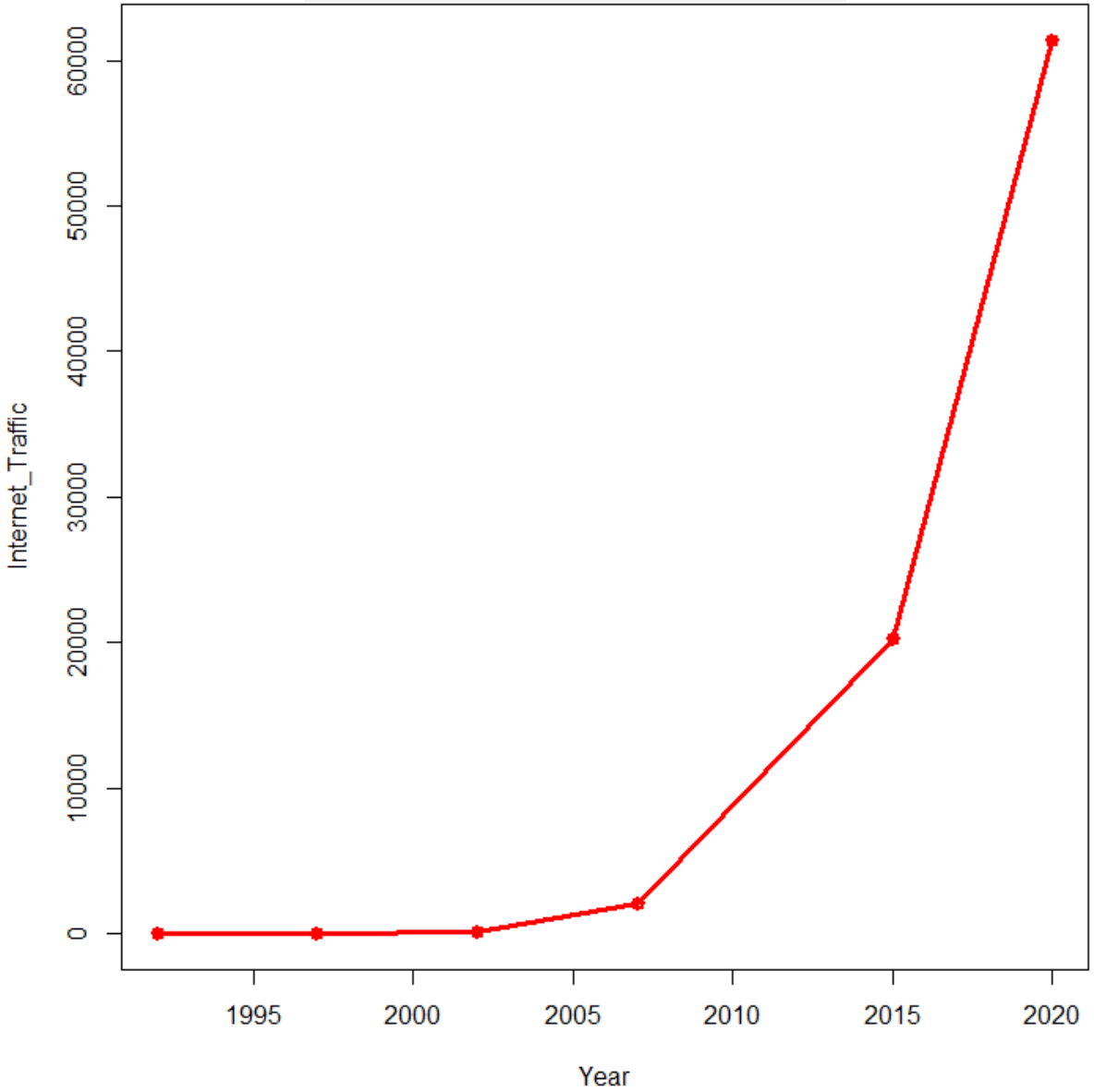
- Install Anaconda
  - With the latest released version of Python
  - Install Jupyter Notebook
- Install Visual Studio Code (VSC)
  - Install extensions: Jupyter Notebook and required extensions
- Get familiar with Google Colab
  - (assuming you already have a Google account!)

- **Get your laptop to the class – every class!**

Data Science – Why now?

- Data Science – Why now?
- De-mystifying Data Science
  - Analytics, ML, AI

5,30,56,179


Internet_Traffic
0
10000
20000
30000
40000
50000
60000
1995
2000
2005
2010
2015
2020
Year

- “Open” only when required …
- Prior knowledge of the content
- Predictable
- Manageable

- **Flows continuously**
- **NO prior knowledge of the content**
- **Unpredictable**
- **Unmanageable**

With Data - Increased capabilities, Insights … Better Decisions!

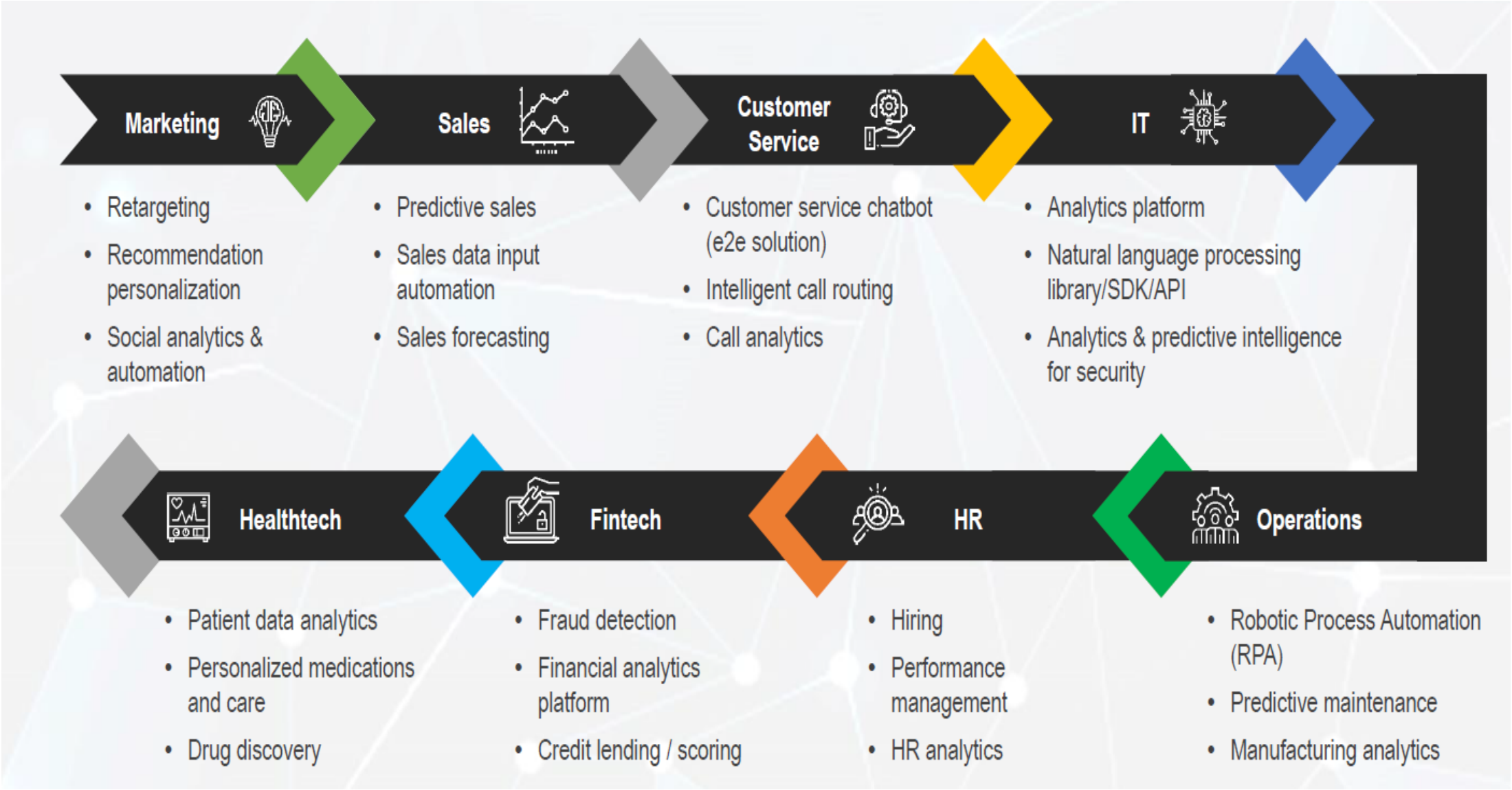
First step …

Subsequent steps …

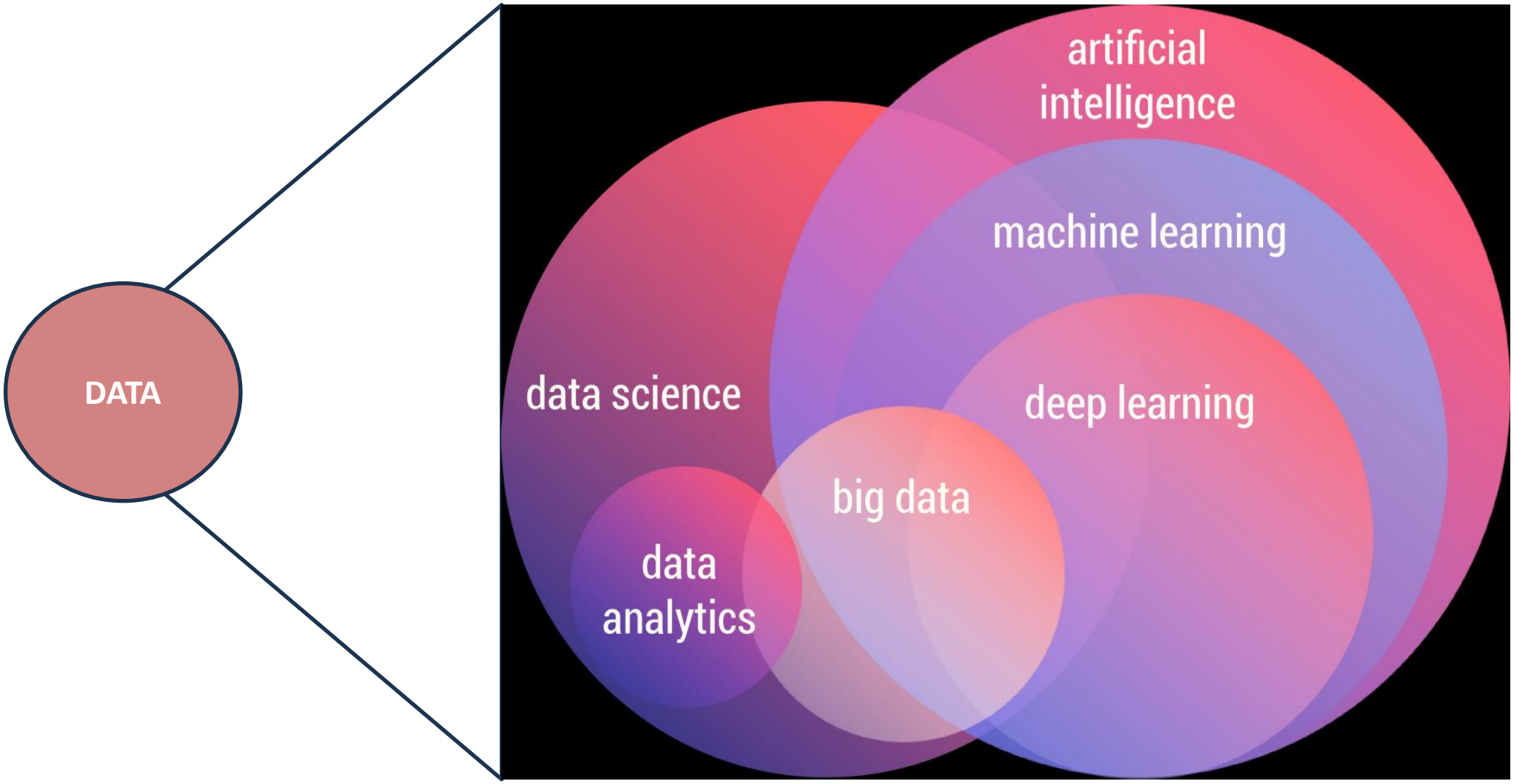
'Gems' of Information

Where are the ‘Gems’ useful?

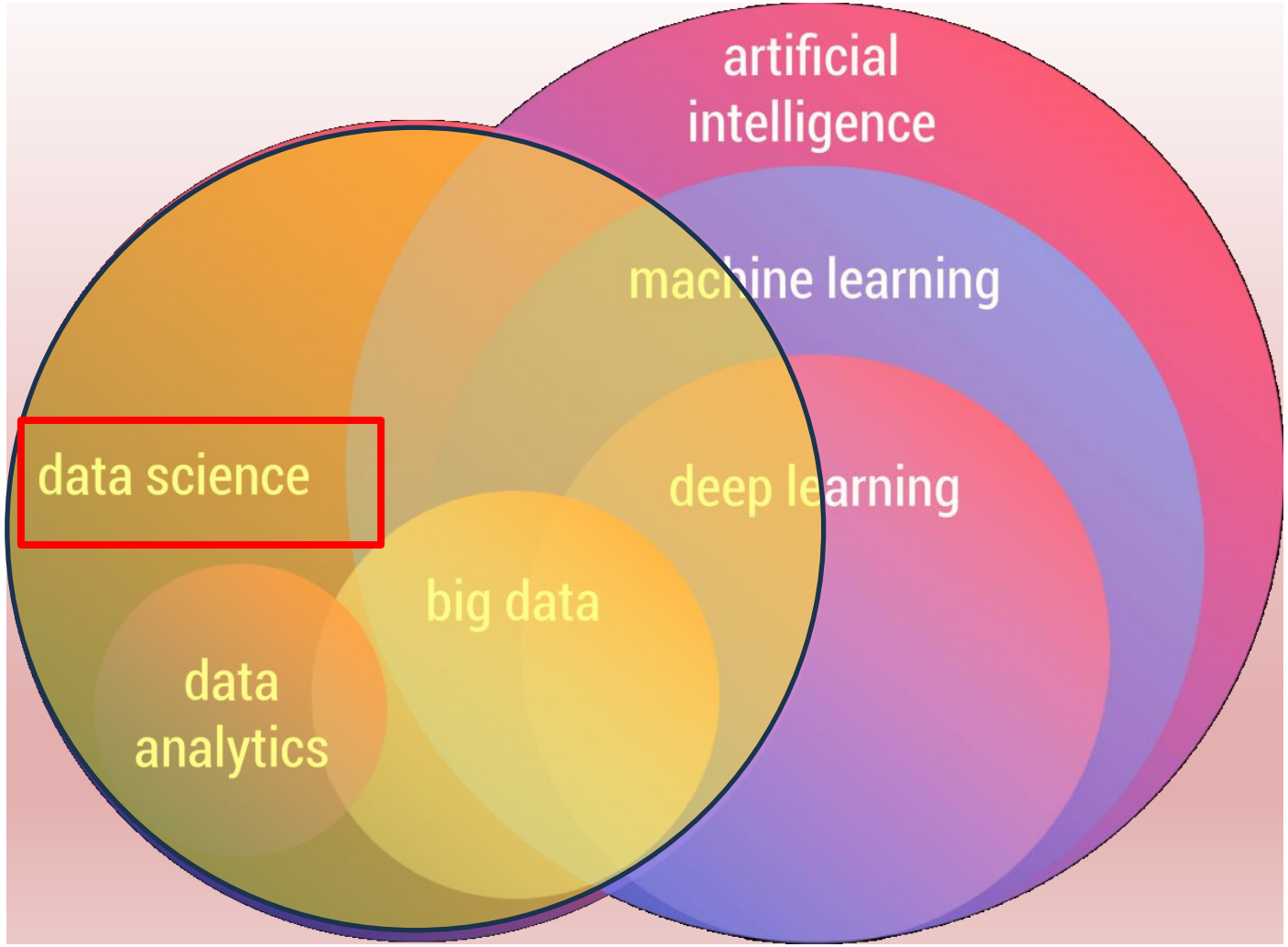
Where is Data making a difference?



“Where there is Data, There is a way …”



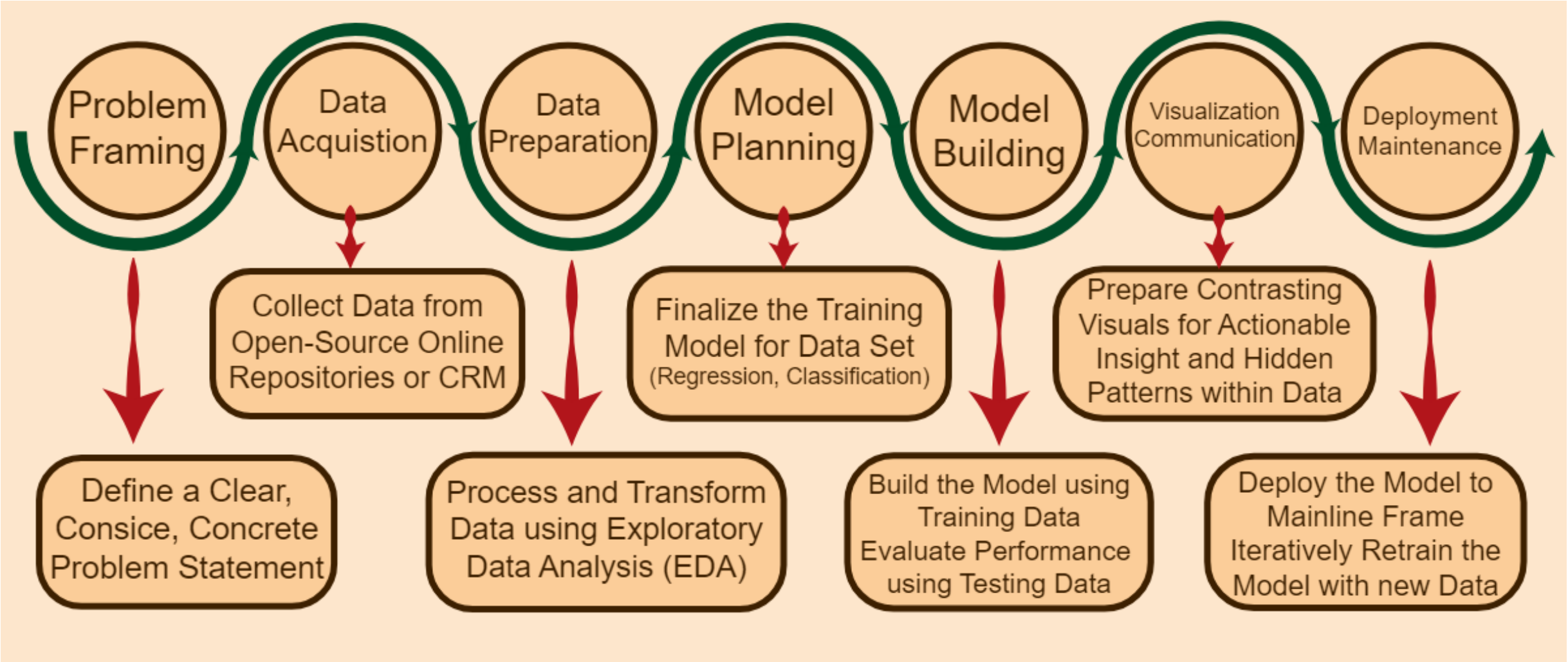
Data Science



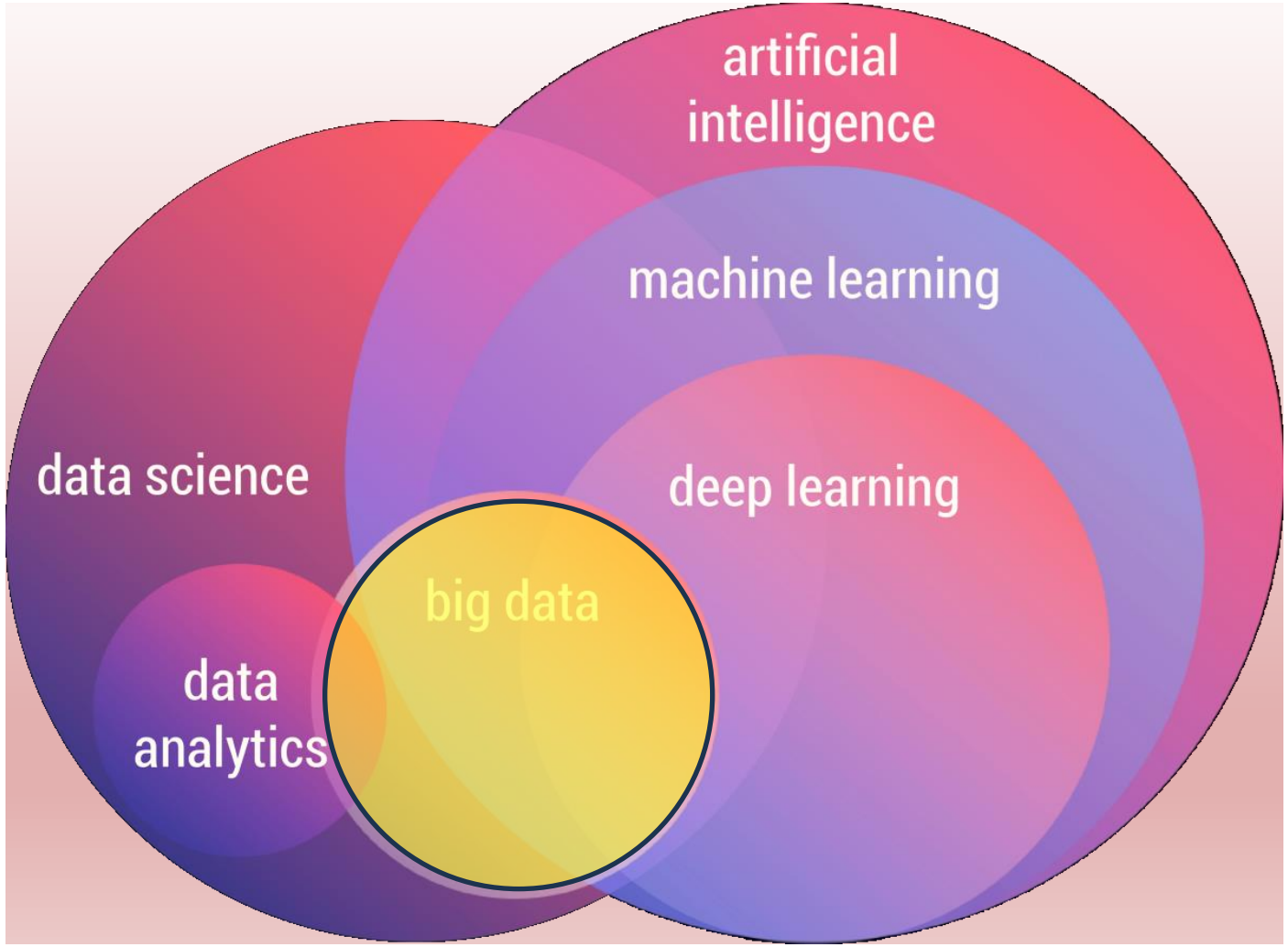
What is Data Science?

- Data science is a broad field that encompasses the overall process of **extracting insights** and **knowledge** from data. It involves **collecting, cleaning, organizing, analyzing, and interpreting data** to uncover **patterns, trends, and meaningful information**.
- Data science utilizes various techniques, methodologies, and tools to **extract valuable insights** from data.

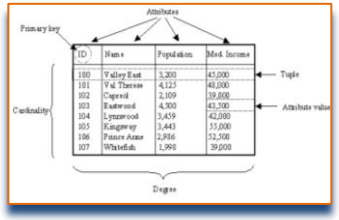
The Data Science Process



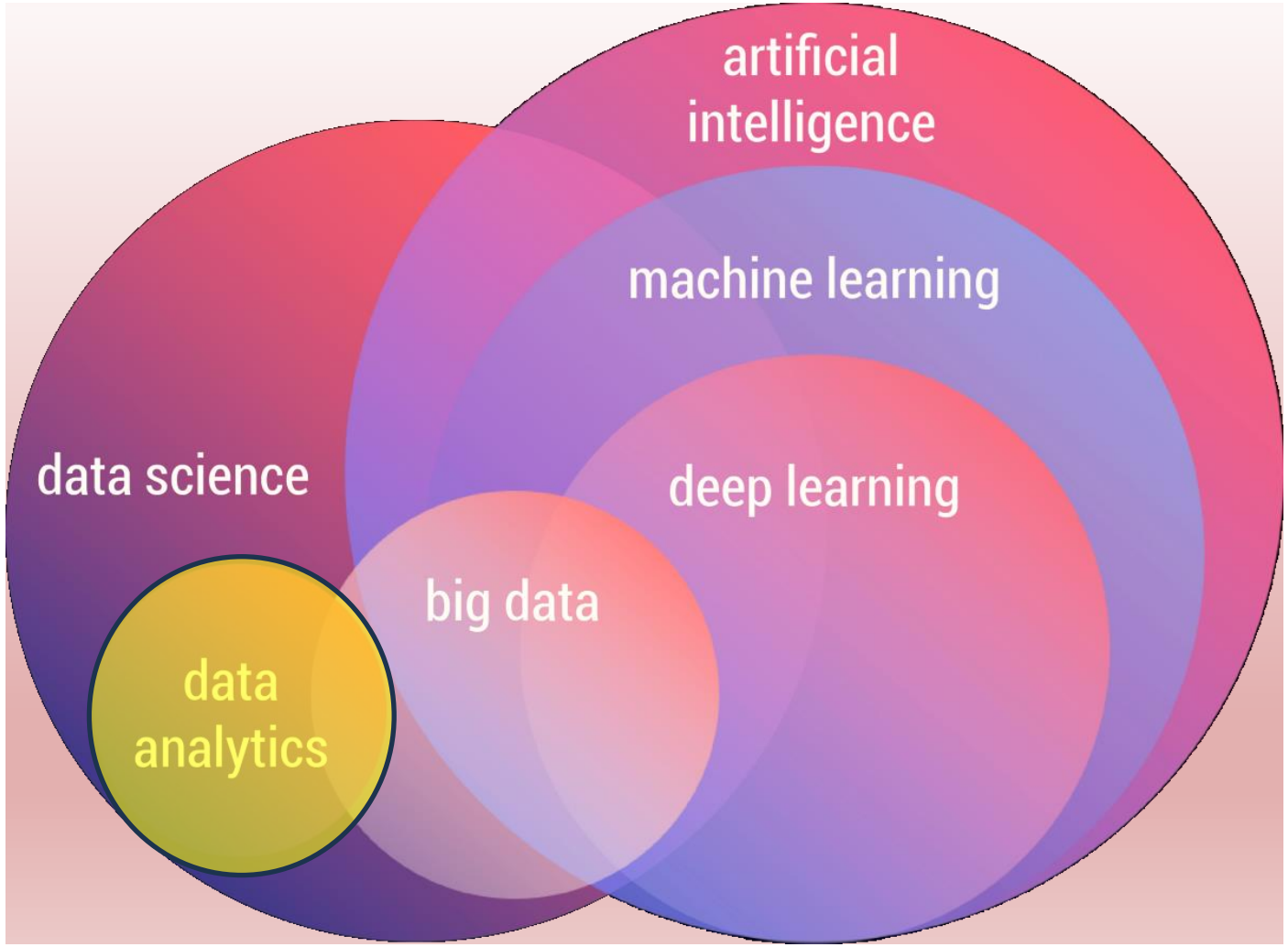
Big Data


artificial
intelligence
machine learning
data science
deep learning
big data
data
analytics

Big Data


Attributes
Primary key
ID
Name
Population
Med. Income
100 Valley East 3,200 45,000
101 Val Thereze 4,125 48,000
102 Cupwod 2,109 39,000
103 Eastwood 4,300 43,500
104 Lynnwood 3,459 42,000
105 Kingsway 3,443 55,000
106 Prince Anne 2,986 52,500
107 Whitefish 1,998 39,000
Tuple
Attribute value
Cardinality
Degree

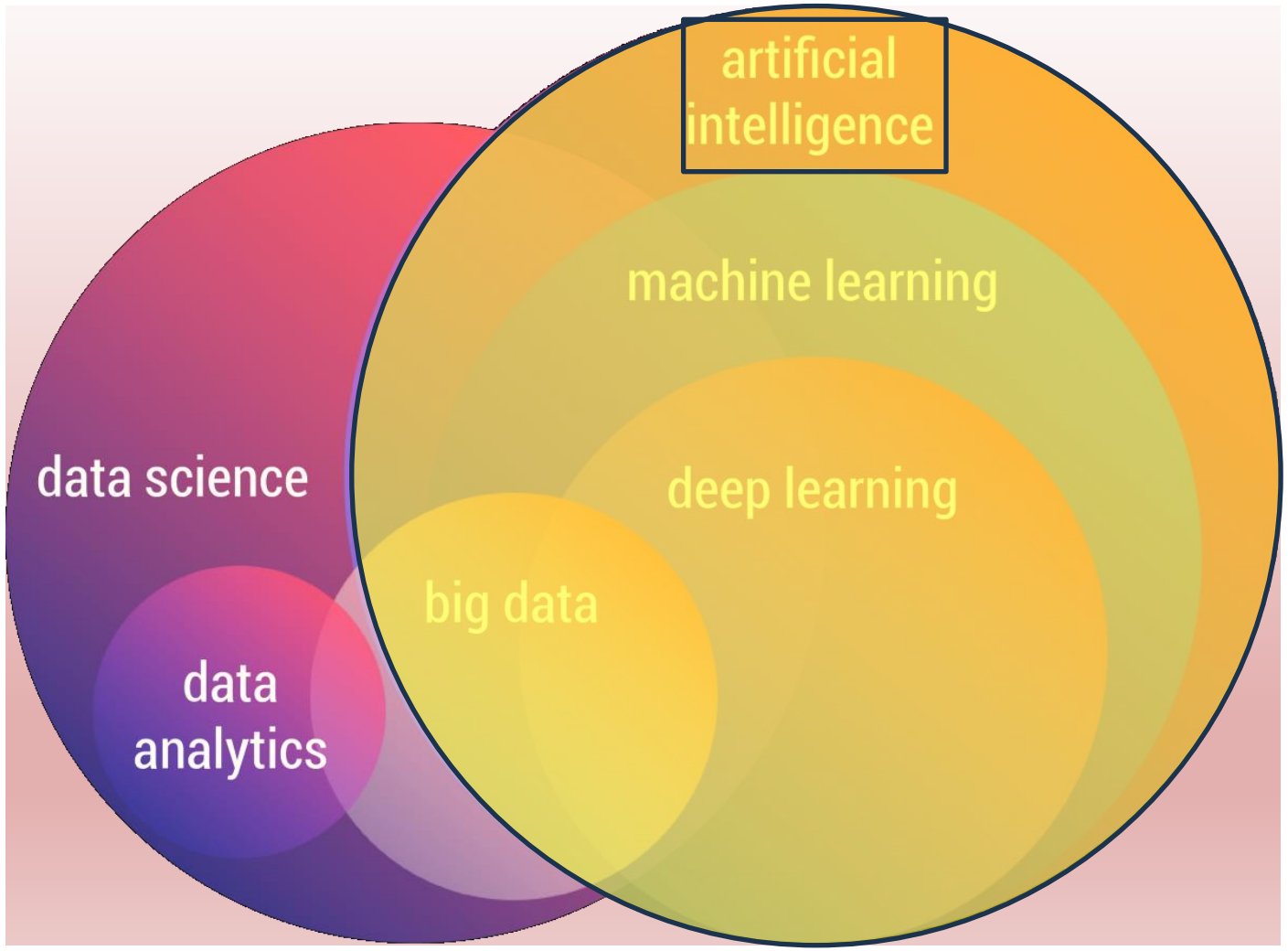
Data Analytics



What is *Analytics*

- The systematic **computational analysis** of data or statistics
- **Information resulting from** the systematic analysis of data or statistics
- **Discovery and communication** of meaningful patterns in data
- Science of examining raw data with the purpose of **drawing conclusions** about that information

AI, ML, DL


artificial intelligence
machine learning
deep learning
big data
data science
data analytics

What is AI?

AI is the ability of machines to perform functions similar to that of a human mind like perceiving, learning, problem solving, etc.

NASSCOM AI Primer

AI Business Categories

Advanced Analytics

Advanced Analytics majorly represent machine learning algorithms which works on large data sets to make predictions and recommendations like predictive modeling etc.

APPLICATIONS

- Predictive Modelling
- Demand Forecasting
- Customer Segmentation
- Cross Selling and Up Selling
- Dynamic Pricing
- Risk Analytics

Natural Language Processing

NLP allows the engagement of virtual agents that will interact with humans and provide a solution.

APPLICATIONS

- Machine Translation
- Paraphrasing
- Natural Language Generation
- Image Captioning
- Topic Modelling
- Speech Tagging

Computer Vision

Computer Vision allows generation of intelligent and insightful description of the visual image by using signals received like face recognition, automated medical image analysis, etc.

APPLICATIONS

- Face Detection
- Object Recognition
- Facial Recognition
- Eye Tracking
- Emotion Recognition
- Motion Detection

Robotics Process Automation

RPA enables a robot or software to collect, interpret, extract knowledge and information to make transactions and communicate with other systems like automated customer order processing, etc.

APPLICATIONS

- Manipulating Data
- Trigger Responses
- Communicating with Digital Systems
- Creating virtual workforce

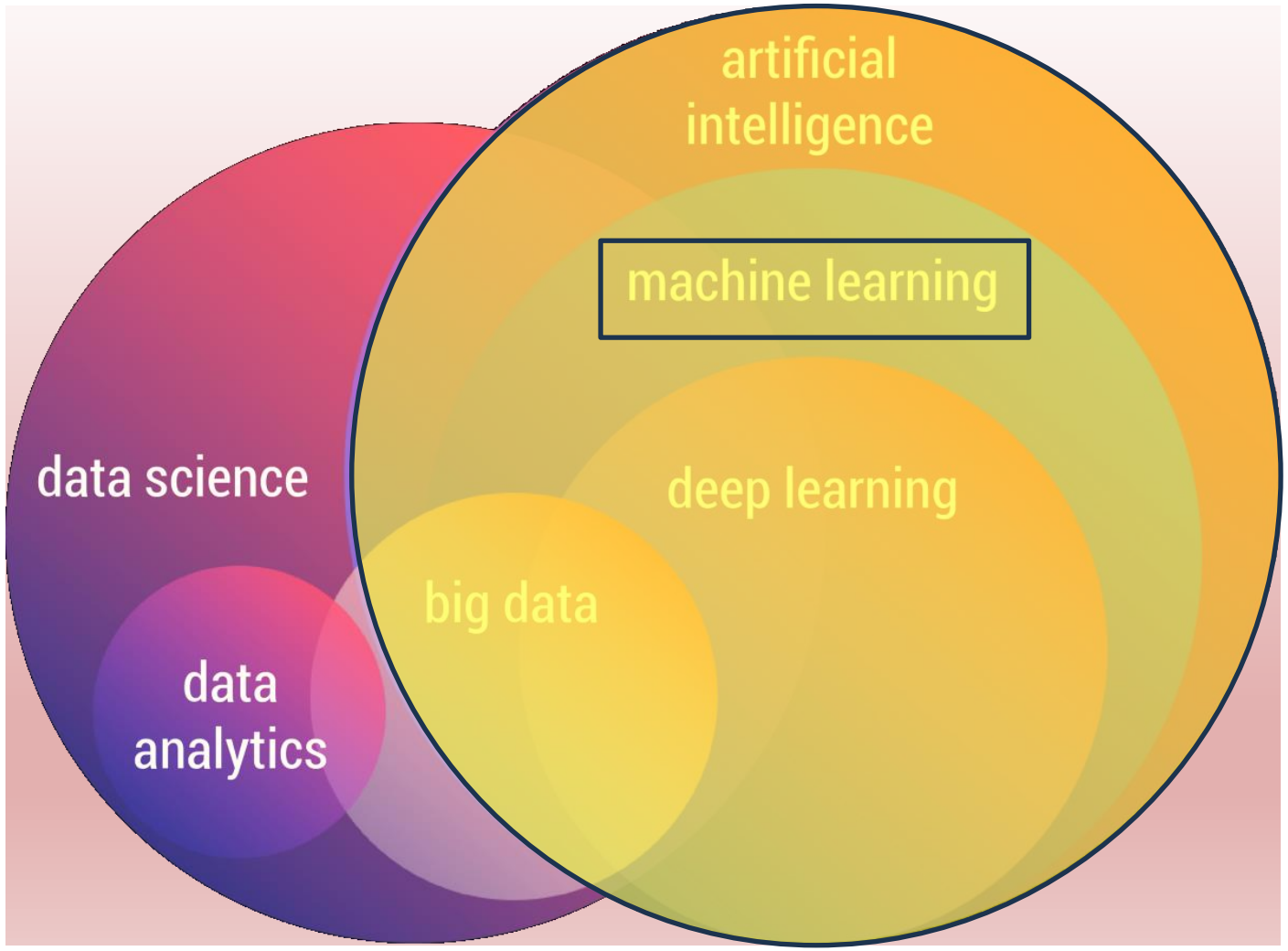
Speech Recognition

Speech recognition allows computers to interpret human speech and provide relevant solutions. It converts spoken words into text and navigational commands.

APPLICATIONS

- Speaker Identification
- Language Identification
- Speech to text processing
- Multimodal Interaction
- Voice Web Search

AI, ML, DL


artificial intelligence
machine learning
data science
deep learning
big data
data analytics

What is Machine Learning?

Machine Learning is most important AI technique to handle large and complex data

Machine Learning

	Algorithm	Use Case Example	Outcome
Supervised Learning Used when we know the classification of data and what to predict	Liner Regression Logistics Regression Linear / Quadratic Discriminant Analysis Decision Tree Naïve Bayes Support Vector Machine Random Forest AdaBoost	Estimating product price elasticity Classify customers on likeliness to repay a loan Classify customer on likeliness to repay a loan Find attributes in a product that make it likely for purchase Analyze sentiments to assess product perception Analyze sentiments to assess product perception Predict power usage in a distribution grid Detect fraudulent activity in a credit card	**Descriptive** What Happened?
Unsupervised Learning Used when we don't know the classification of data and want the algorithm to classify data	K Means Clustering Gaussian Mixture Model Hierarchical clustering Recommender System	Segment customers into groups by characteristics Segment customers based on less distinctive characteristics Inform product usage by grouping customers Recommend news article to a readers based on what they are currently reading	**Predictive** What Will Happen?
Reinforcement Learning Used when we don't have training data and only way to learn about the environment is to learn with it		Balance the load on electricity grids in varying demand cycles Optimize the driving behavior of self-driving cars Finding real time pricing during a product auction	**Prescriptive** What To Do?

What is Machine Learning?

Machine Learning is most important AI technique to handle large and complex data

Machine Learning

	Algorithm	Use Case Example	Outcome
Supervised Learning Used when we know the classification of data and what to predict	Liner Regression Logistics Regression Linear / Quadratic Discriminant Analysis Decision Tree Naïve Bayes Support Vector Machine Random Forest AdaBoost	Estimating product price elasticity Classify customers on likeliness to repay a loan Classify customer on likeliness to repay a loan Find attributes in a product that make it likely for purchase Analyze sentiments to assess product perception Analyze sentiments to assess product perception Predict power usage in a distribution grid Detect fraudulent activity in a credit card	**Descriptive** What Happened?
Unsupervised Learning Used when we don't know the classification of data and want the algorithm to classify data	K Means Clustering Gaussian Mixture Model Hierarchical clustering Recommender System	Segment customers into groups by characteristics Segment customers based on less distinctive characteristics Inform product usage by grouping customers Recommend news article to a readers based on what they are currently reading	**Predictive** What Will Happen?
Reinforcement Learning Used when we don't have training data and only way to learn about the environment is to learn with it		Balance the load on electricity grids in varying demand cycles Optimize the driving behavior of self-driving cars Finding real time pricing during a product auction	**Prescriptive** What To Do?

The ML algorithm toolkit

1. Supervised learning

- 1.1. Linear Models
- 1.2. Linear and Quadratic Discriminant Analysis
- 1.3. Kernel ridge regression
- 1.4. Support Vector Machines
- 1.5. Stochastic Gradient Descent
- 1.6. Nearest Neighbors
- 1.7. Gaussian Processes
- 1.8. Cross decomposition
- 1.9. Naive Bayes
- 1.10. Decision Trees
- 1.11. Ensemble methods
- 1.12. Multiclass and multioutput algorithms
- 1.13. Feature selection
- 1.14. Semi-supervised learning
- 1.15. Isotonic regression
- 1.16. Probability calibration
- 1.17. Neural network models (supervised)

2. Unsupervised learning

- 2.1. Gaussian mixture models
- 2.2. Manifold learning
- 2.3. Clustering
- 2.4. Biclustering
- 2.5. Decomposing signals in components (matrix factorization problems)
- 2.6. Covariance estimation
- 2.7. Novelty and Outlier Detection
- 2.8. Density Estimation

3. Model selection and evaluation

- 3.1. Cross-validation: evaluating estimator performance
- 3.2. Tuning the hyper-parameters of an estimator
- 3.3. Metrics and scoring: quantifying the quality of predictions
- 3.4. Validation curves: plotting scores to evaluate models

4. Inspection

- 4.1. Partial Dependence and Individual Conditional Expectation plots
- 4.2. Permutation feature importance

6. Dataset transformations

- 6.1. Pipelines and composite estimators
- 6.2. Feature extraction
- 6.3. Preprocessing data
- 6.4. Imputation of missing values
- 6.5. Unsupervised dimensionality reduction
- 6.6. Random Projection
- 6.7. Kernel Approximation
- 6.8. Pairwise metrics, Affinities and Kernels
- 6.9. Transforming the prediction target (y)

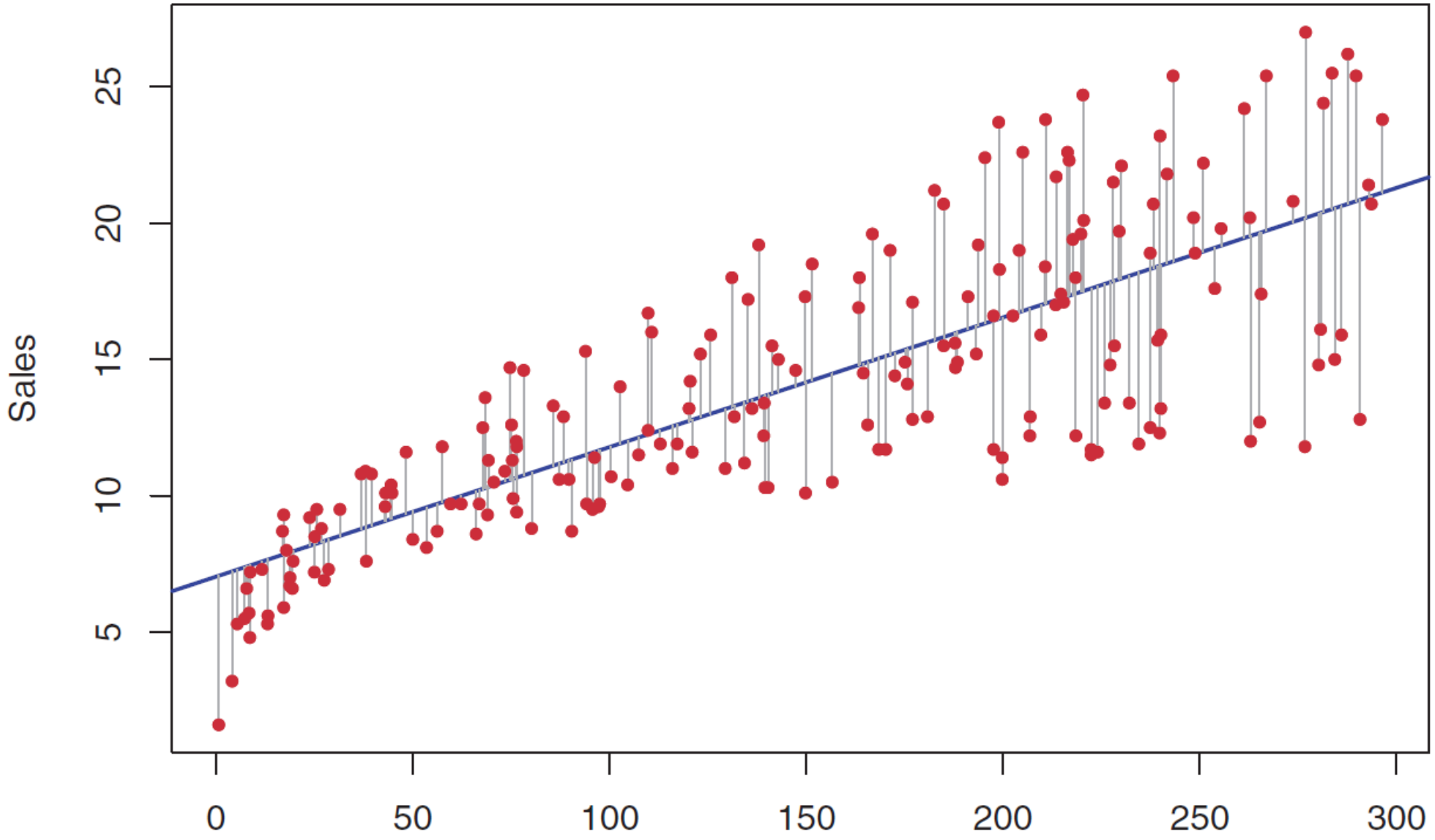
7. Dataset loading utilities

- 7.1. Toy datasets
- 7.2. Real world datasets
- 7.3. Generated datasets
- 7.4. Loading other datasets

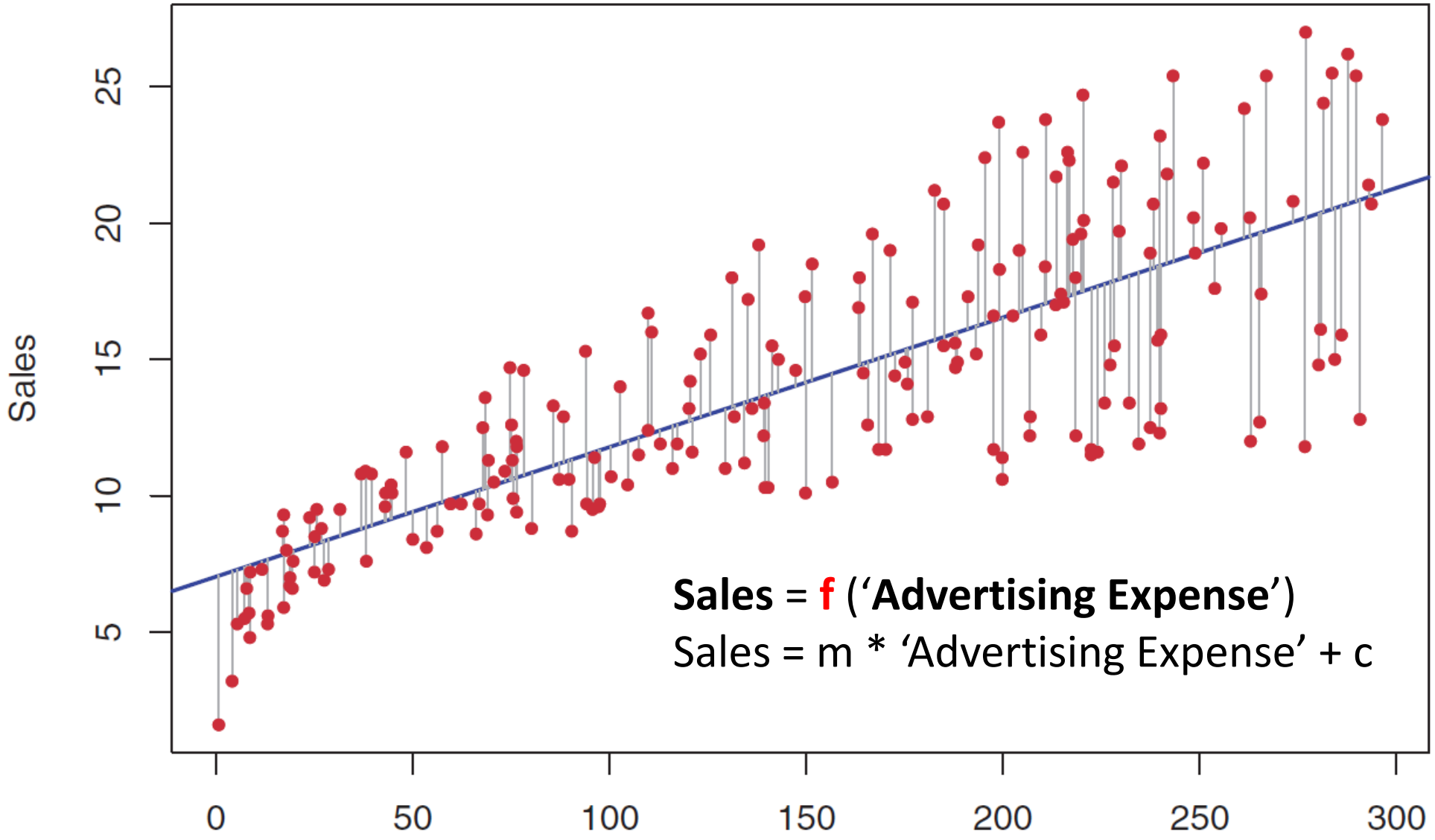
8. Computing with scikit-learn

- 8.1. Strategies to scale computationally: bigger data
- 8.2. Computational Performance
- 8.3. Parallelism, resource management, and configuration

Linear Regression



Linear Regression

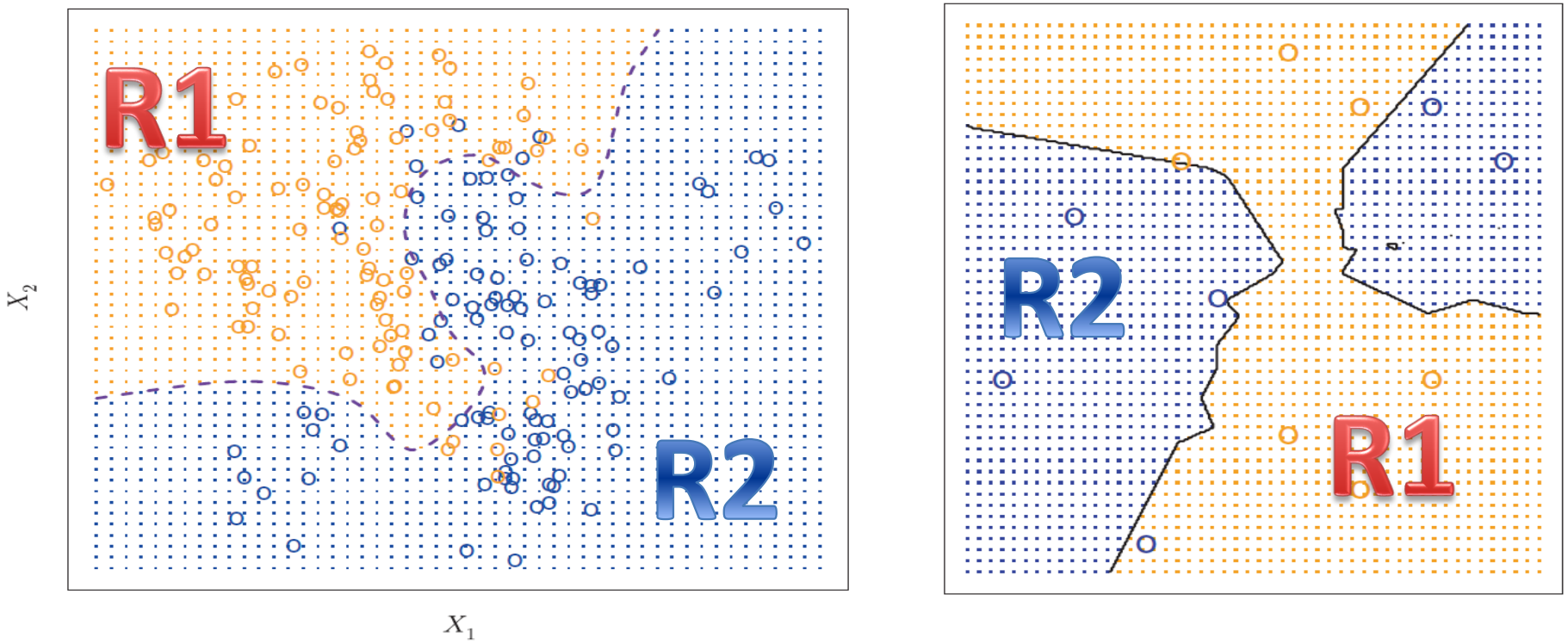


Some questions answered by Linear Regression

- If I spend “X” on advertising, how much will the sales be ?
  - Prediction

- If I change the advertising budget by “X”, by how much will the sales be impacted?
  - Sensitivity analysis

Logistic Regression – A Classification Method



Automatically derive the 'most logical' boundary between regions

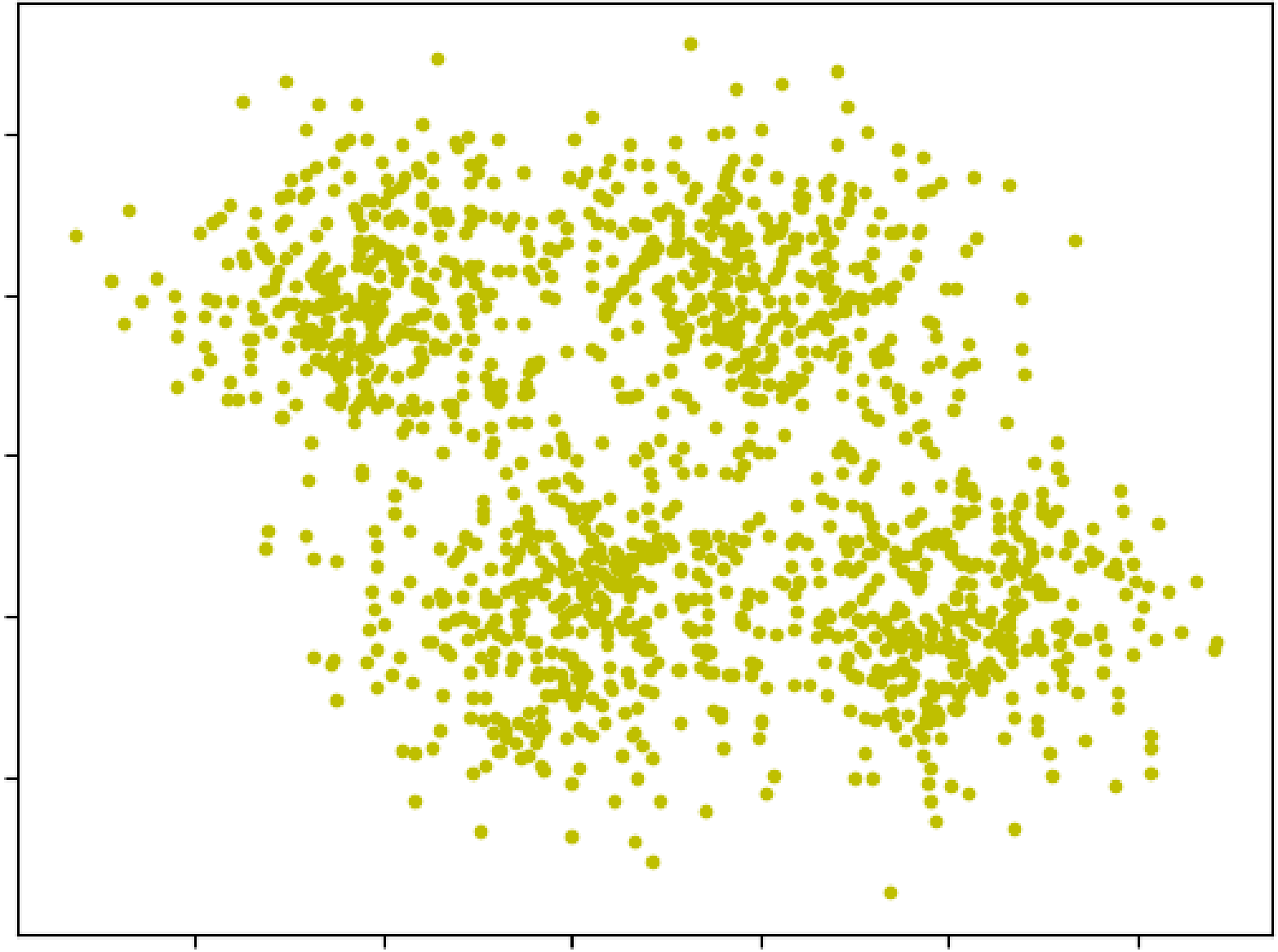
Some questions answered by Classification

- Is this transaction likely to be fraudulent?
- Is this customer likely to “leave”?
- Is this part / assembly defective?
- Is this machine ready for ‘maintenance’?
- What kind e-commerce customer is this person? Low / Moderate / Heavy spender?
- Based on the medical parameters, is this person potentially suffering a disease?

Clustering

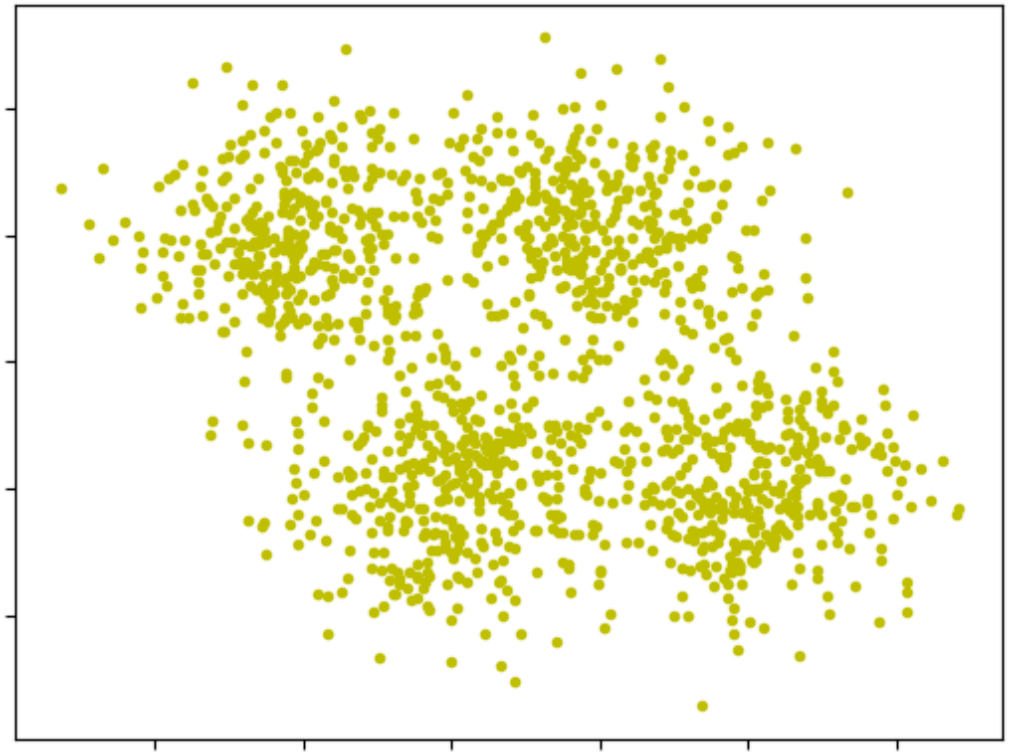
	x	y
0	-12.304702	3.499240
1	-21.302900	17.983794
2	-6.320254	29.639092
3	2.259775	26.227155
4	-14.777150	19.536615
5	-11.347139	-9.874762
6	-28.129312	15.026950
7	-7.662440	7.403947
8	-14.612828	30.144784
9	-25.559933	20.264730
10	2.186647	41.988614
11	-26.895259	14.034500
12	-8.596083	11.475732
13	-18.834186	23.573933
14	-21.165917	31.873053
15	-14.647640	14.019951
16	1.858298	30.085540
17	-26.188447	16.726300
18	-18.165225	33.418052
19	6.772593	19.605175
20	-10.144608	21.384122
21	-29.413260	6.499945
22	4.751169	9.118655
23	-27.313856	22.670176
24	-13.564653	13.222997
25	-27.203389	22.017289
26	-15.925783	41.299904
27	7.956781	0.279732
28	-1.692639	25.441604
29	-21.273197	24.386794
30	-22.428757	6.946292
31	-10.835374	36.036029
32	-27.600552	15.872282
33	-26.441987	20.834743

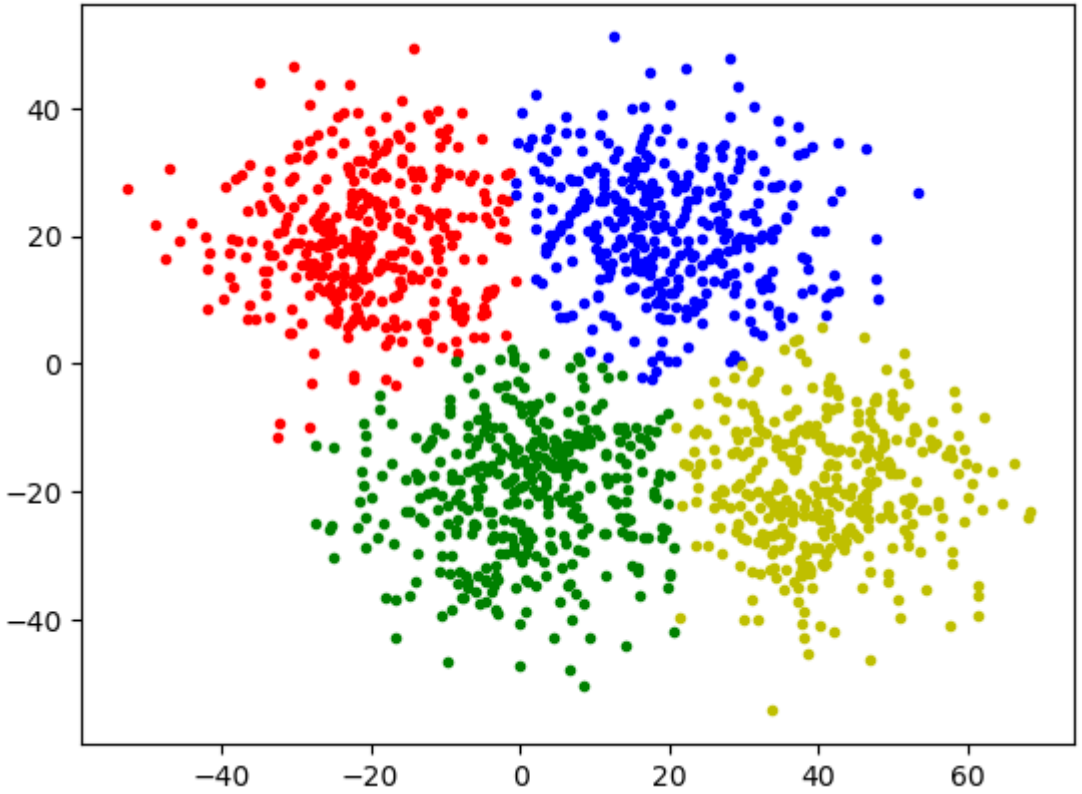
Clustering



	x	y
0	-12.304702	3.499240
1	-21.302900	17.983794
2	-6.320254	29.639092
3	2.259775	26.227155
4	-14.777150	19.536615
5	-11.347139	-9.874762
6	-28.129312	15.026950
7	-7.662440	7.403947
8	-14.612828	30.144784
9	-25.559933	20.264730
10	2.186647	41.988614
11	-26.895259	14.034500
12	-8.596083	11.475732
13	-18.834186	23.573933
14	-21.165917	31.873053
15	-14.647640	14.019951
16	1.858298	30.085540
17	-26.188447	16.726300
18	-18.165225	33.418052
19	6.772593	19.605175
20	-10.144608	21.384122
21	-29.413260	6.499945
22	4.751169	9.118655
23	-27.313856	22.670176
24	-13.564653	13.222997
25	-27.203389	22.017289
26	-15.925783	41.299904
27	7.956781	0.279732
28	-1.692639	25.441604
29	-21.273197	24.386794
30	-22.428757	6.946292
31	-10.835374	36.036029
32	-27.600552	15.872282
33	-26.441987	20.834743

Clustering

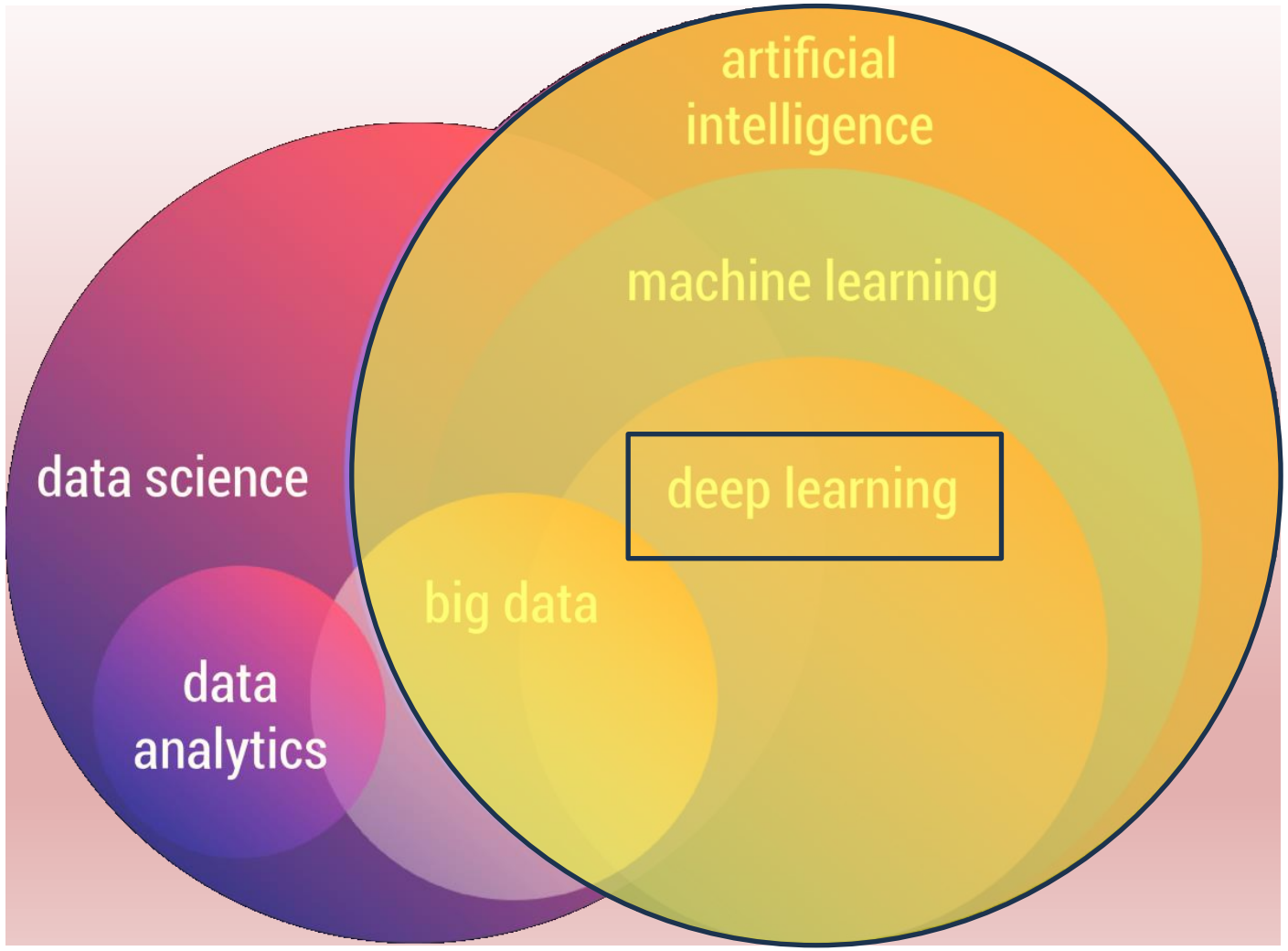




Clustering

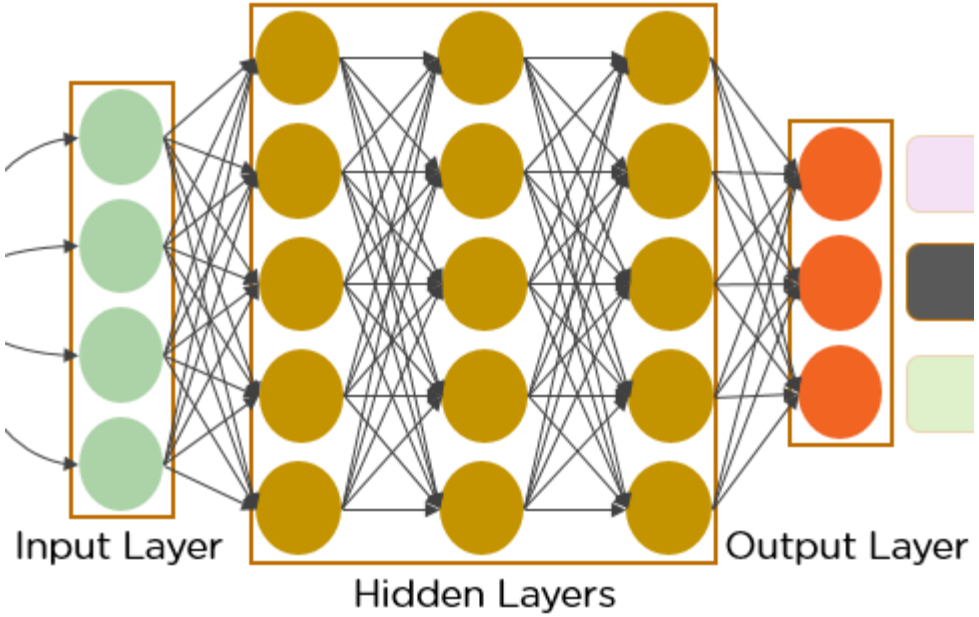
- Organizes Data without much prior knowledge
- Helps create Structured Knowledge from new Data
- Helps identify ‘outliers’ (eg. unexpected transactions) – thereby playing an important role in tasks like:
  - Anomaly detection
  - Fraud pattern discovery
  - Real-time monitoring / Fraud detection
- Helps group customers based on their transactions’ profile

AI, ML, DL

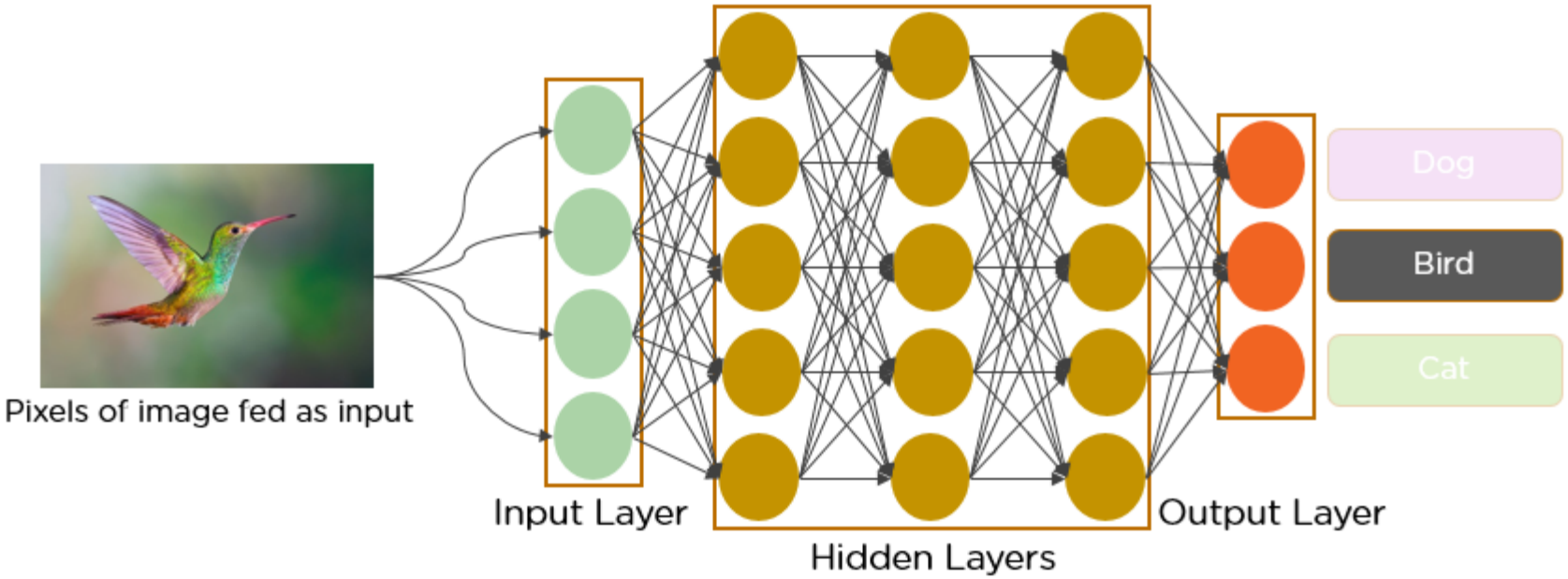

artificial intelligence
machine learning
deep learning
data science
big data
data analytics

What is Deep Learning?

A network that replicates the activities in the brain!



What is Deep Learning?



Applications of Deep Learning

- Image recognition
- Natural Language Understanding
- Speech Recognition
- Autonomous Vehicles
- Recommendation Systems
- Medical diagnosis
- Drug discovery
- Fraud detection
- Algorithmic trading
- Gaming
- Virtual reality / Augmented virtual reality