Getting Started With Data Science

Data Science

MACHINE LEARNING

DEEP LEARNING

Getting Started in Data Science

Introduce YourSelf

1. Name
2. Experience
3. Exposure in Data Science
4. City
5. Organization

DATA

Machine/Deep Learning Consultant Profile

Utkarsh Sharma

Data Scientist | Business Consultant | Project Manager | Entrepreneur

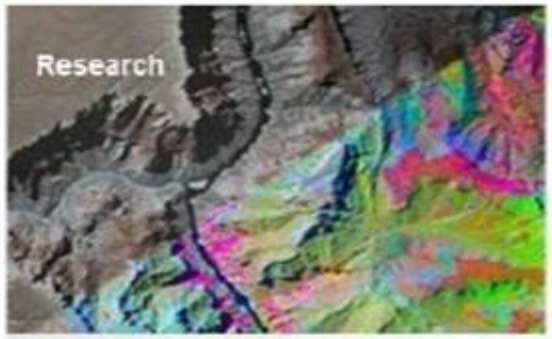
Bangalore | New Delhi

utkarsh.sharma012@gmail.com

https://www.linkedin.com/in/utkarsh-sharma-6a1b8024

- ✓ 10+years of professional Experience. Worked for companies incl. HCL, MAKINO AUTO, CGI, Aon, Ericsson, HARMAN Connected Services etc.
- ✓ Currently working as Data Science lead Consultant and Trainer
- ✓ Delivered 100+ Training Batches on ML/AI/Python/R/ Data Science.

Getting Started With Data Science



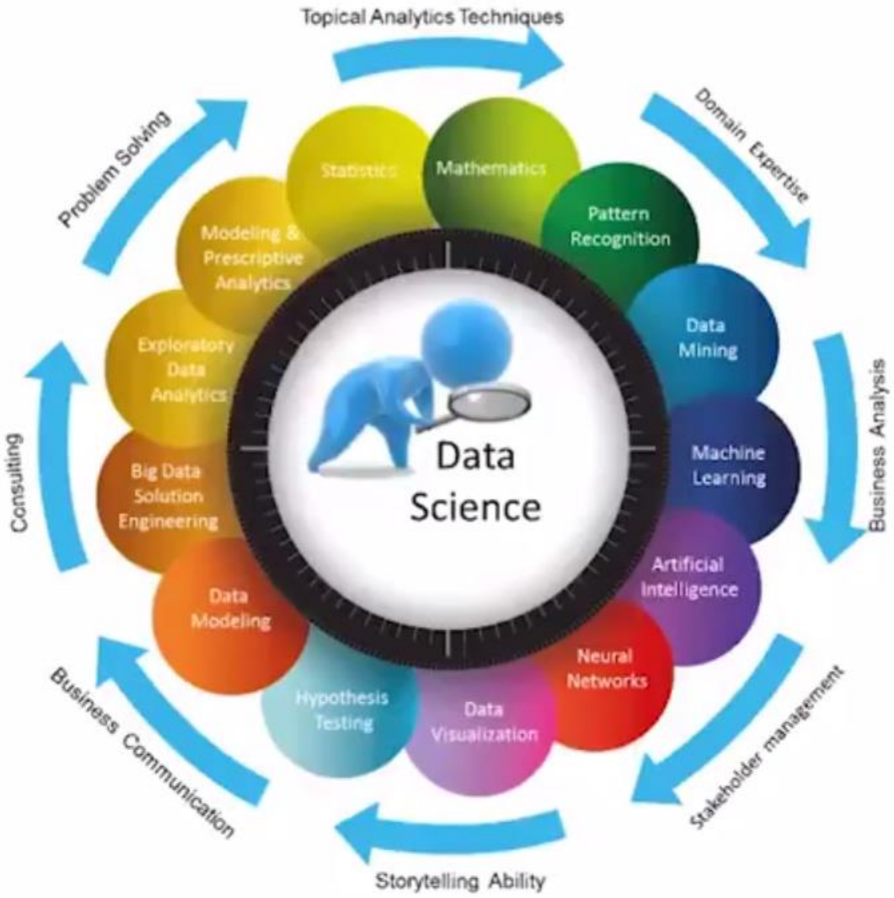
With
UTKARSH SHARMA
Data Science Expert

Inc.

Agenda

1. What is Data Science
2. Who Are Data Scientists
3. Data Science at LinkedIn
4. Steps for Success in Data Science Projects
5. How to Become a Data Scientist
6. The Top Ten Machine Learning Algorithms
7. Jobs in Data Science
8. Scope & Opportunities
9. The Data Science Dictionary
10. Conclusion

What is Data Science ?



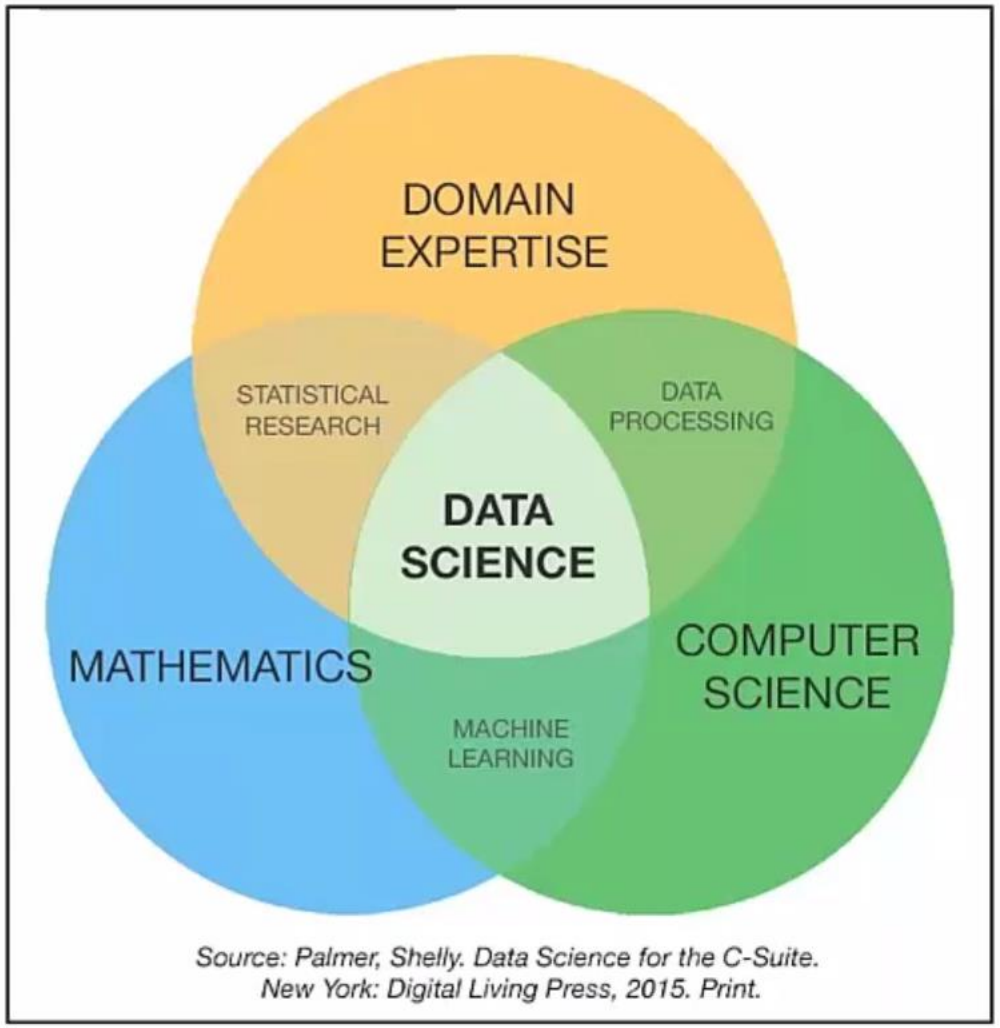
Data Science refers to processes and methods that help make sense of large volumes of data for organizational purposes. It is an amalgamation of many disciplines, however, it does not draw from each of them equally or in fixed proportions.

Back

What is Data Science ?

- Data science is the process of extracting meaning full insights from raw data which might be in structure or un-structured format.

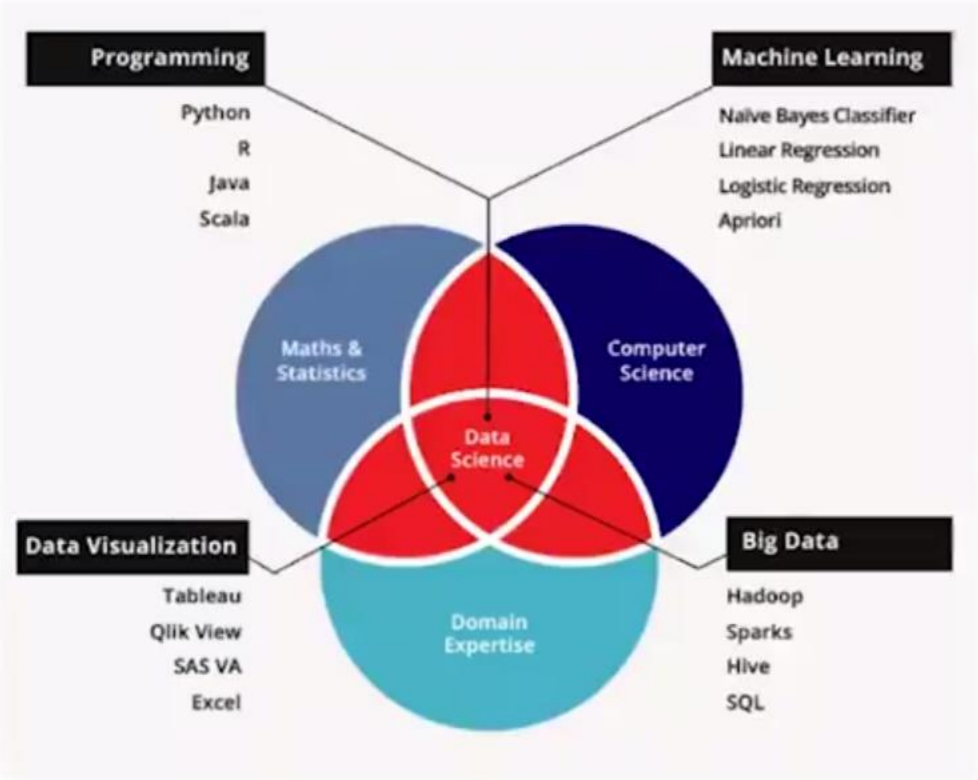
- What Data Science is not ?
  - No a piece of software – end to end App
  - Not about Visualization
  - Academic research – its more of providing business solution



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

Back

How to become a Data scientist ?

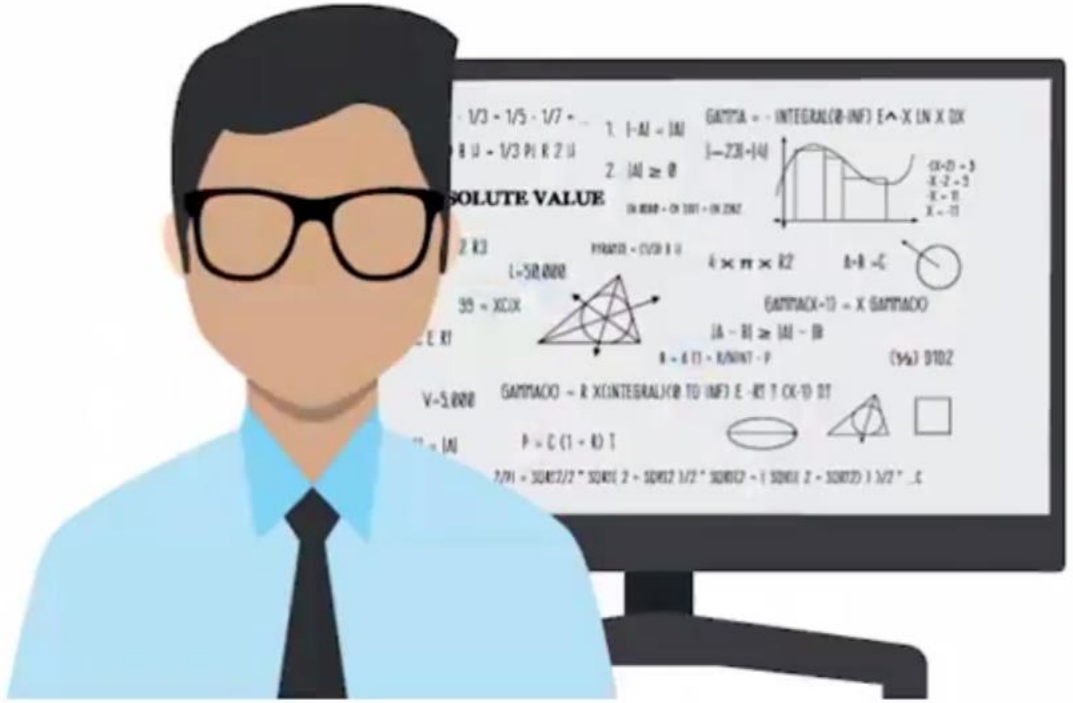


To become a proficient data scientist one must master in the following six essential and broad components of data science.

They are - Statistics, Programming, Big Data, Data Visualization, Machine Learning, and Domain Expertise

Back

Who are Data scientists ?

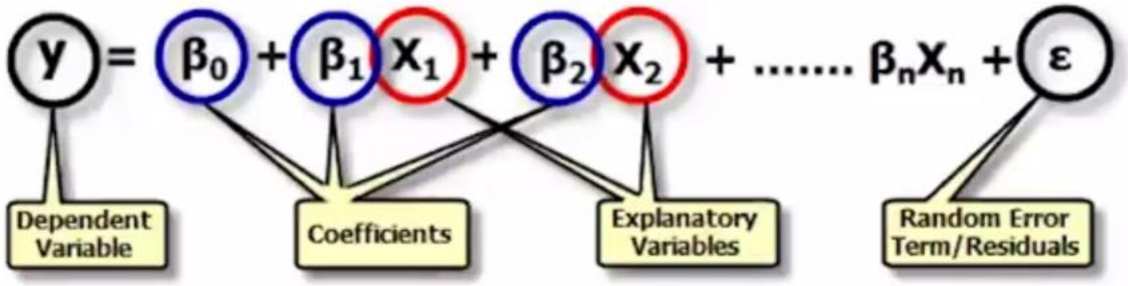


A Data scientist is a hacker, an analyst, a communicator, and an adviser, all wrapped in one. They help decision makers interact and interpret data for specific purposes and also contribute greatly towards the development of products and businesses.

Back

Statistics

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots \beta_n X_n + \varepsilon$$



Example: suppose you want to predict residential burglary (RES_BURG) for the census tracts in your community. You've identified median income (MED_INC), the number of vandalism incidents (VAND), and the number of households (HH_UNITS) to be key explanatory variables. The regression equation would have the elements below:

$$RES\_BURG = \beta_0 + \beta_1 * (MED\_INC) + \beta_2 * (VAND) + \beta_3 * (HH\_UNITS) + \varepsilon$$

Image Source: http://resources.esri.com

Statistics is a broad field that deals with collection, analysis, interpretation, presentation, and organization of data. All data analytics algorithms use statistical principles for data analysis. The process requires a basic understanding of descriptive statistics, and probability theory.

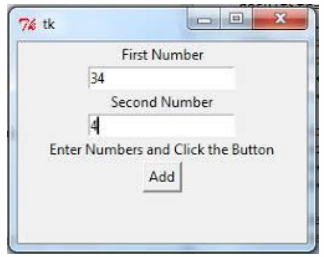
Back

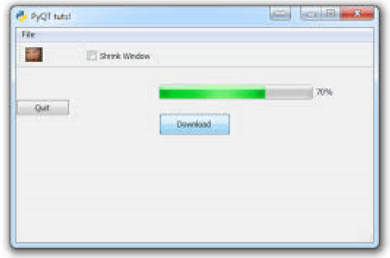
Programming Language

Programming languages help data scientists design tools for data analysis. **Python and R** are two programming languages that data scientists use widely.

Back

Python and R Programming Language





Python Programming

This language is becoming popular amongst data scientists. Python lets you work fast, is flexible, and uses elegant syntax that is easy to learn.

R Programming

R is a language and environment for statistical computing and statistical graphics. The open-source platform offers many features which are useful for statistical analysis and representation.

Back

Python and R Programming Language

- R Studio
  - Majorly used for statistical analysis (regression Time series)
  - Good Community support
  - ggplot, ggplot2 packages for visualization

- Python
  - Majorly used for Text mining, sentiment analysis, neural nets, Deep Learning
  - Pretty good community with huge packages
  - Ntlk, TextBlob and other various packages for Text mining
  - Tensorflow and Keras for Deep learning

Back

Utilities & basic Setup

R installation

- https://cran.r-project.org/bin/windows/base/
- https://www.rstudio.com/products/rstudio/download/

Python installation

- https://www.python.org/downloads/
- Download Anaconda 3: https://www.anaconda.com/download/
- IDE: Jupyter Notebook or Spyder

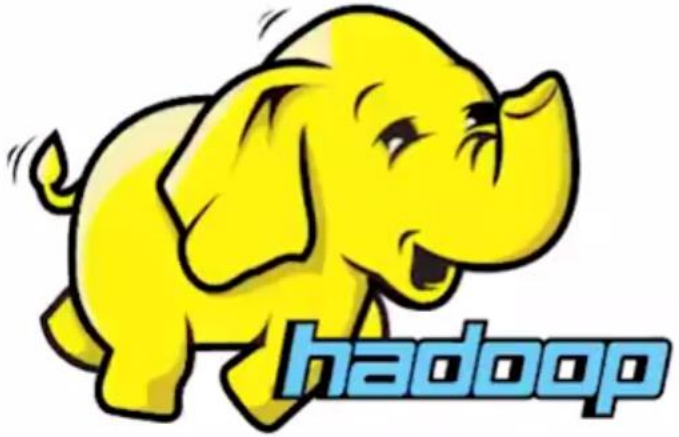
Back

Big Data Technologies

Data scientists are expected to have decent understanding of big data technologies to make use of big data. Hadoop and Spark are two technologies that are mostly used big data technologies.

Back

Hadoop and Spark



Hadoop

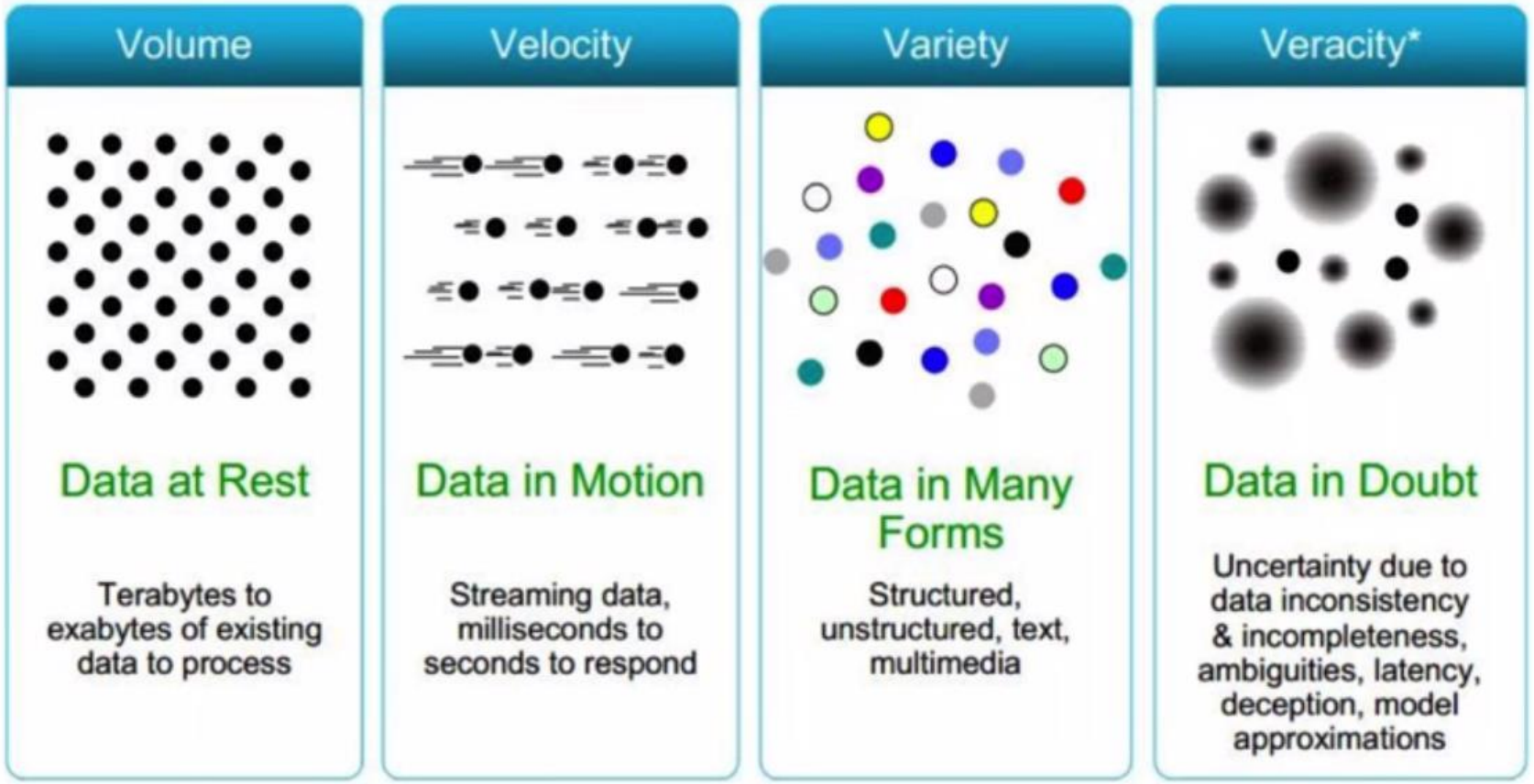
Apache Hadoop allows data storage and easily process large amounts of data. It uses distributed file system to speed up computing and eliminate the risk of failure.

Spark

It is easy to use and comes with high-level libraries that include support for SQL queries, machine learning and graph processing..

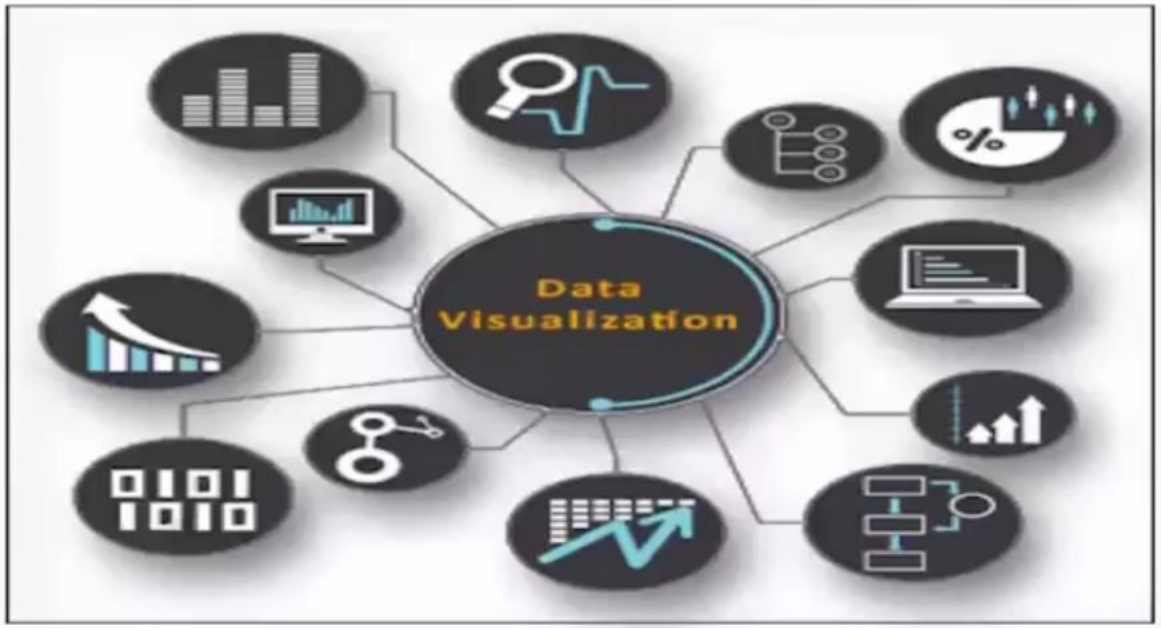
Back

What Is Big Data ?



Back

Data Visualization Tools



Data visualization tools help data scientists communicate what statistics show and what data reveals in a attractive and efficient way. An understanding of tools like Tableau, QlikView and Microsoft's Power BI enhance a data scientist's ability to explain key findings. Tableau is one of the most popular visualization tools in data science circles.

Back

Data Science Jargons

Jargons	Year	Description
Machine Learning	1980's	Focus was on algorithms & the amount of data was limited
Predictive Analytics / Data mining	1990's	Used Algorithms that are developed & applied on Large amount of data
Big data Analytics	2000's	Focus was on computing on big volume of data in distributed fashion
Data Science	2010's	Filed where complex Algorithm works on large volume of data to solve business problem Lot of emphasis on visualization & story telling

Back

Machine Learning

It is the computer's ability to learn from a set of data and adapt itself without being explicitly programmed. Machine learning uses algorithms to analyze input data and predict an output within an acceptable range. The learning is either supervised or unsupervised.

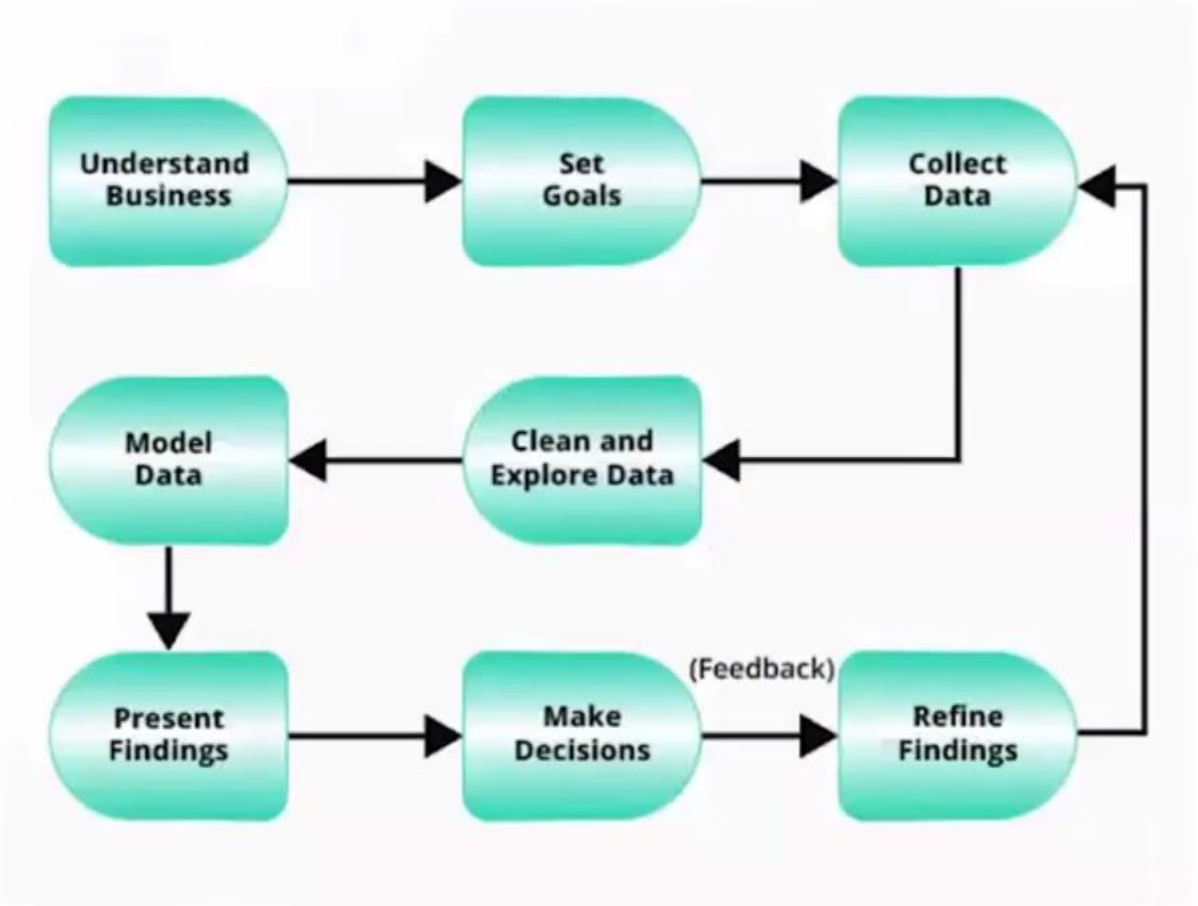
Back

Domain Expertise

Domain expertise is a key component of data science because it provides the context for all data science endeavors. Without an understanding of how businesses- and, more specifically, domains - function, the data scientists would not know how to generate key insights, process the data.

Back

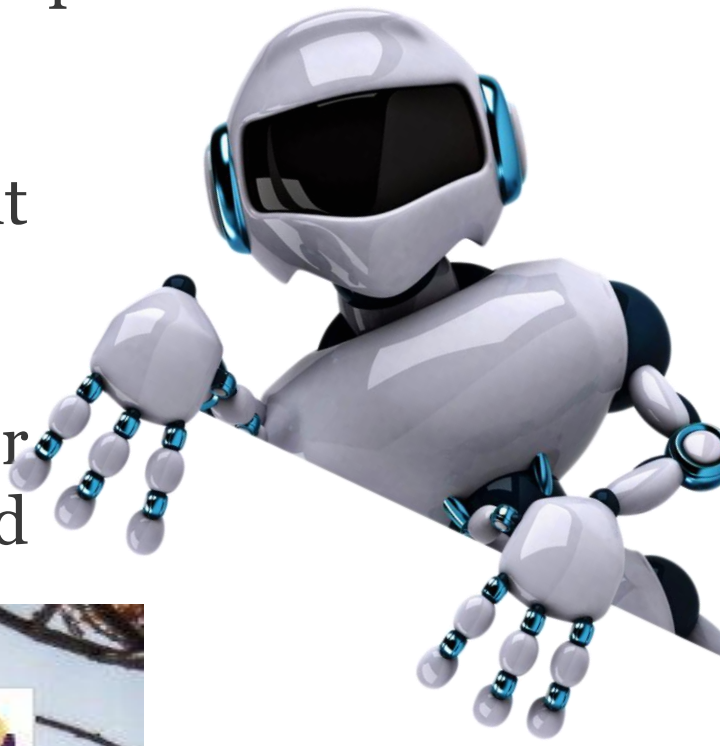
Steps for success in Data science projects



Back

Machine Learning vs Robotic Process Automation

This topic is attracting the interest of many professionals for quite some time.

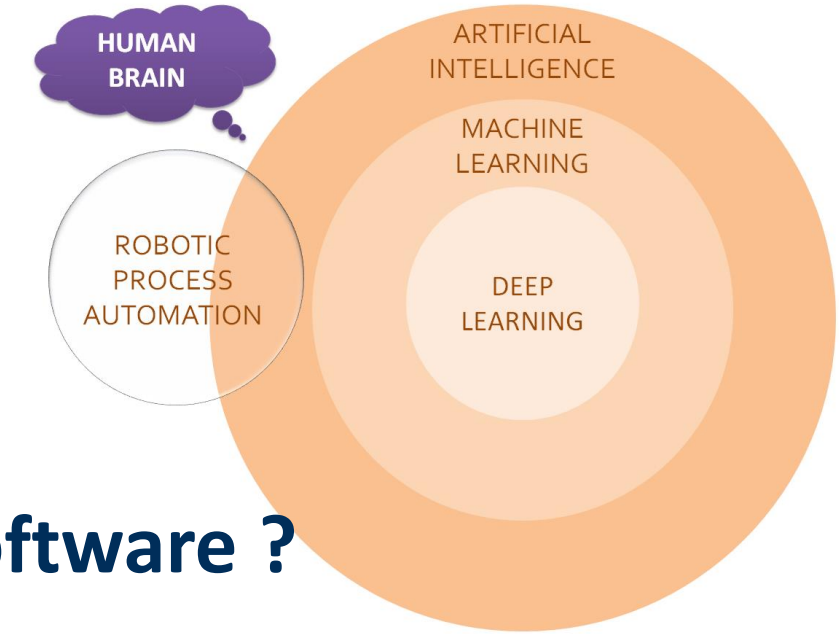


Mainly, to what is it and how it is going to impact the current tasks performed by professionals working in various sectors.

In this article, will try to make it simple to understand for professionals from a Technical or Non-Technical background with a comparison.

Back

Machine Learning vs Robotic Process Automation

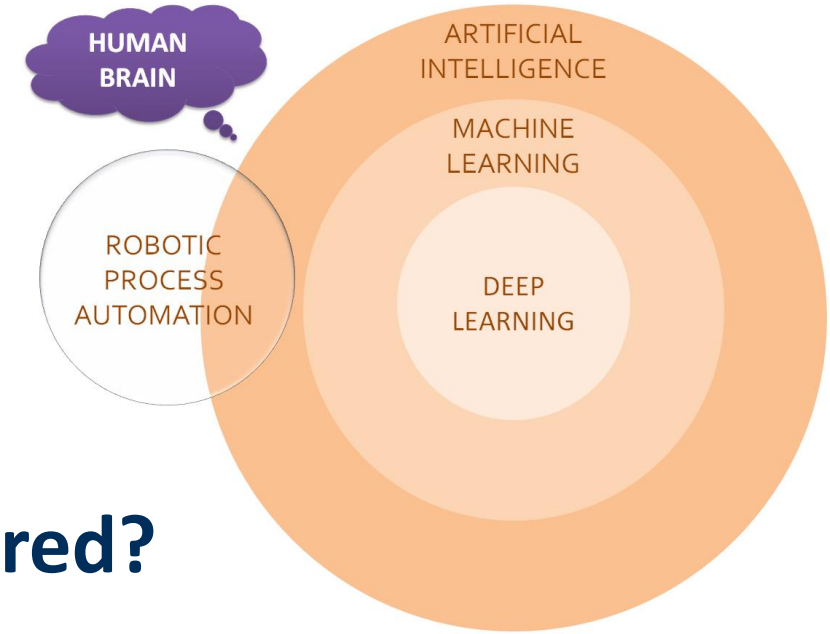


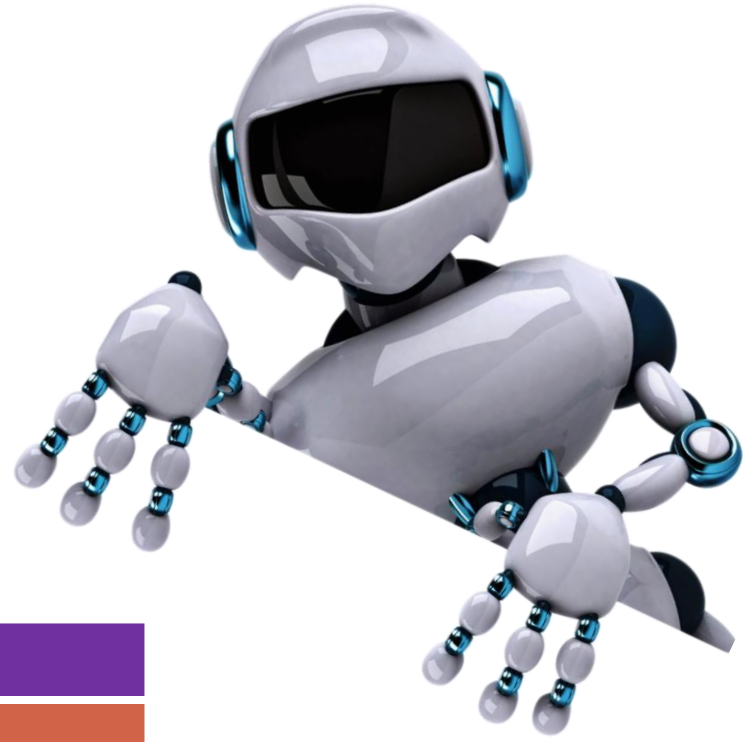
Does it have GUI based software ?

Robotic Process Automation	Machine Learning
Yes, RPA can be performed with help GUI based applications like UiPath, Blue Prism, Automation Anywhere and other.	Yes, Machine Learning can be performed with the help of application like Azure ML,Aws,SAP,Knime, Alteryx.

Back

Machine Learning vs Robotic Process Automation



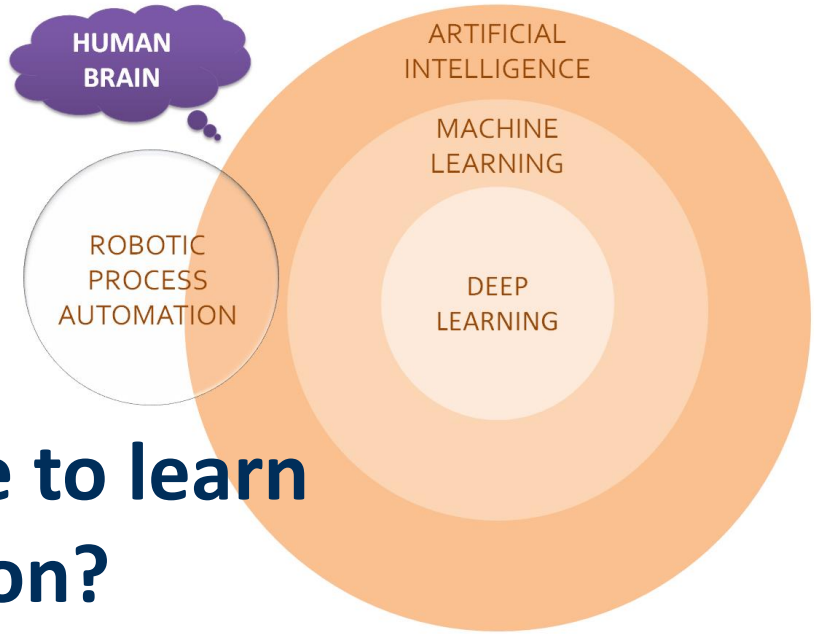


Is coding knowledge required?

Robotic Process Automation	Machine Learning
Basic tasks can be automated in RPA without coding knowledge. For advanced RPA users, programming skills are required.	Yes, R and Python programming skills are required for Machine Learning.

Back

Machine Learning vs Robotic Process Automation

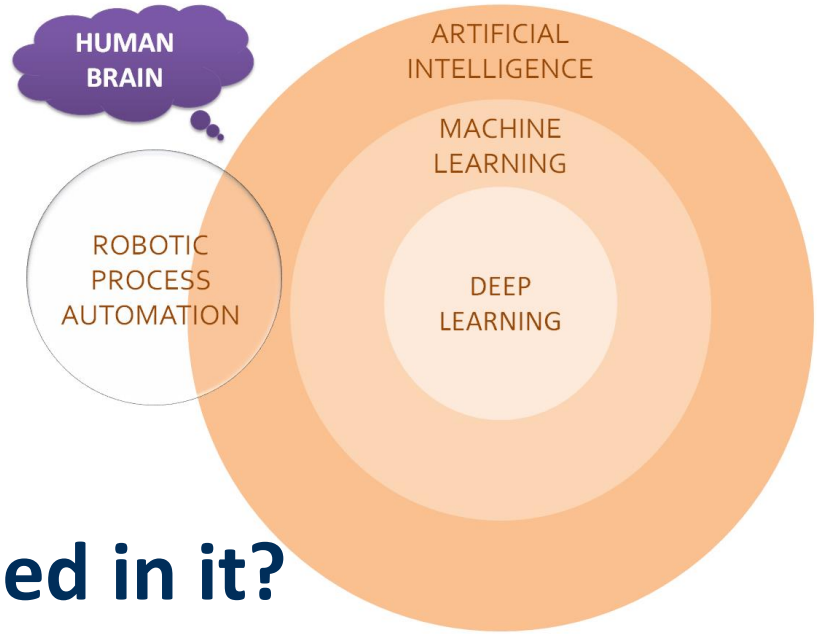


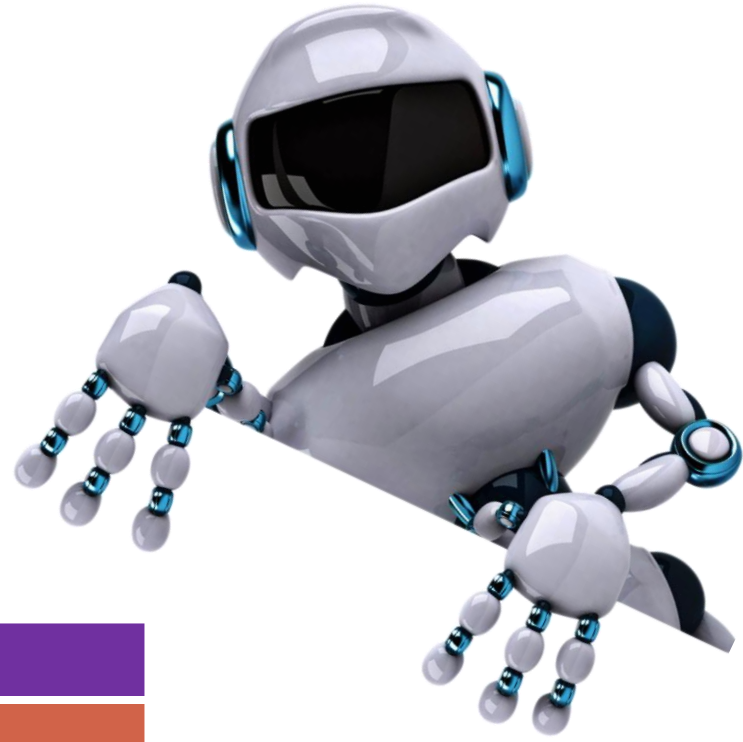
Does it enable the machine to learn without manual intervention?

Robotic Process Automation	Machine Learning
No. It is a combination of Man and Machine.	Yes. Once programmed, it learns from each output and applies it as input in the next stage.

Back

Machine Learning vs Robotic Process Automation



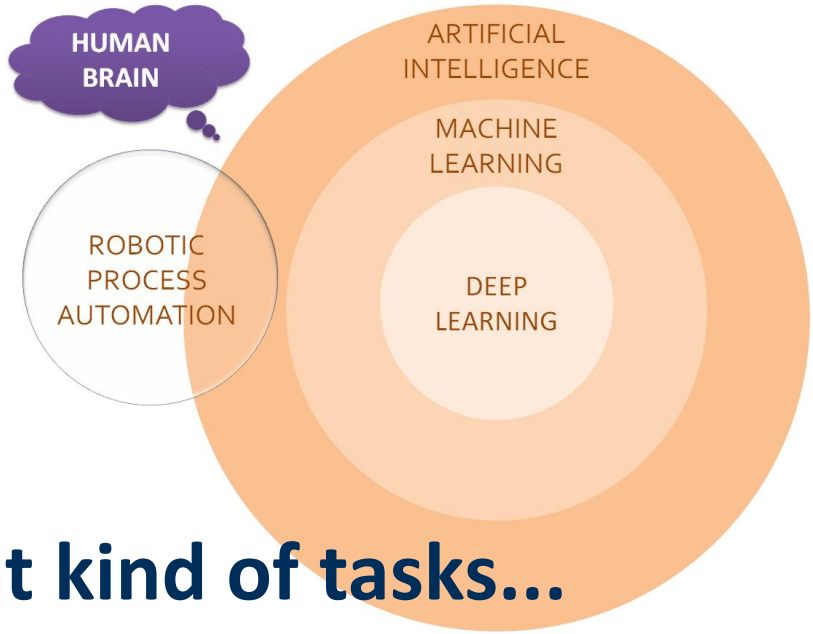


Which technologies are used in it?

Robotic Process Automation	Machine Learning
RPA software like UiPath, Automation Anywhere, Blue Prism and more	Softwares like Sklearn,Tensorflow, Azure ML and many other using open source library of R programming & Python

Back

Machine Learning vs Robotic Process Automation



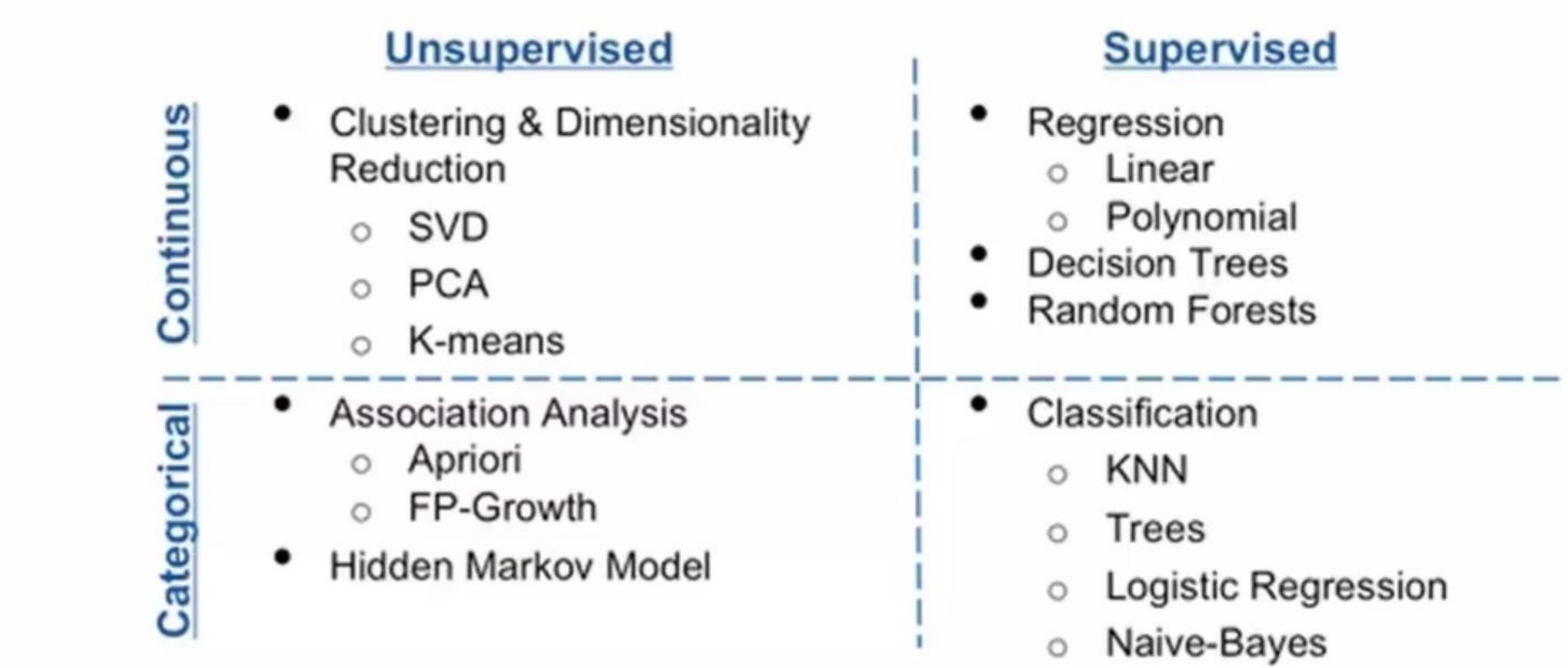
It helps in performing what kind of tasks...

Robotic Process Automation	Machine Learning
RPA helps in automation of repetitive tasks in business processes like HR, Finance, Shared Services, etc.	Machine Learning helps in automation of complex tasks like Data security, Financial Trading, Fraud Detection, Marketing Personalization, Healthcare, Smart Cars, Natural language Processing, Online Search, Recommendation and more.

Back

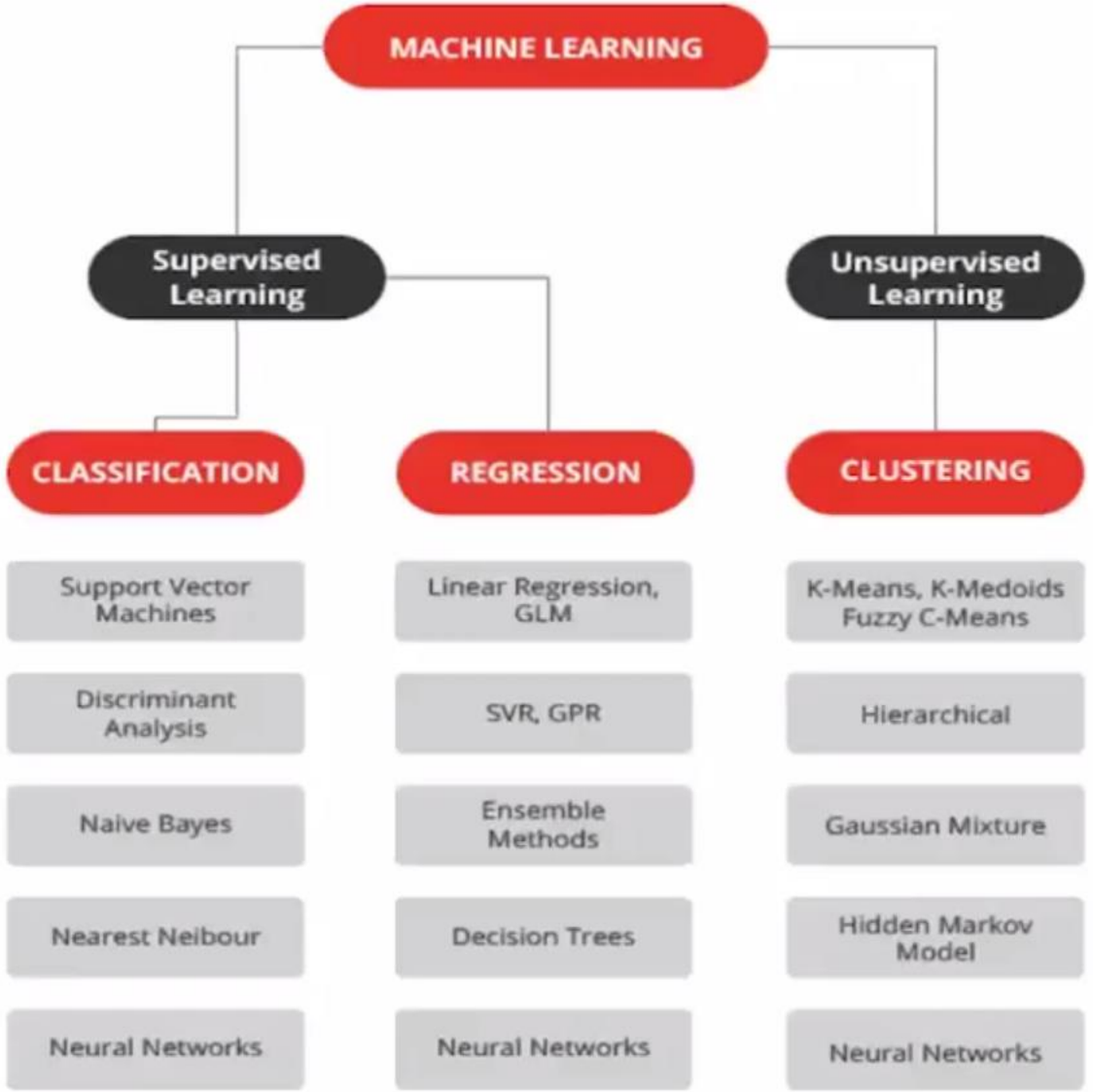
Machine Learning

Machine Learning Algorithms



Back

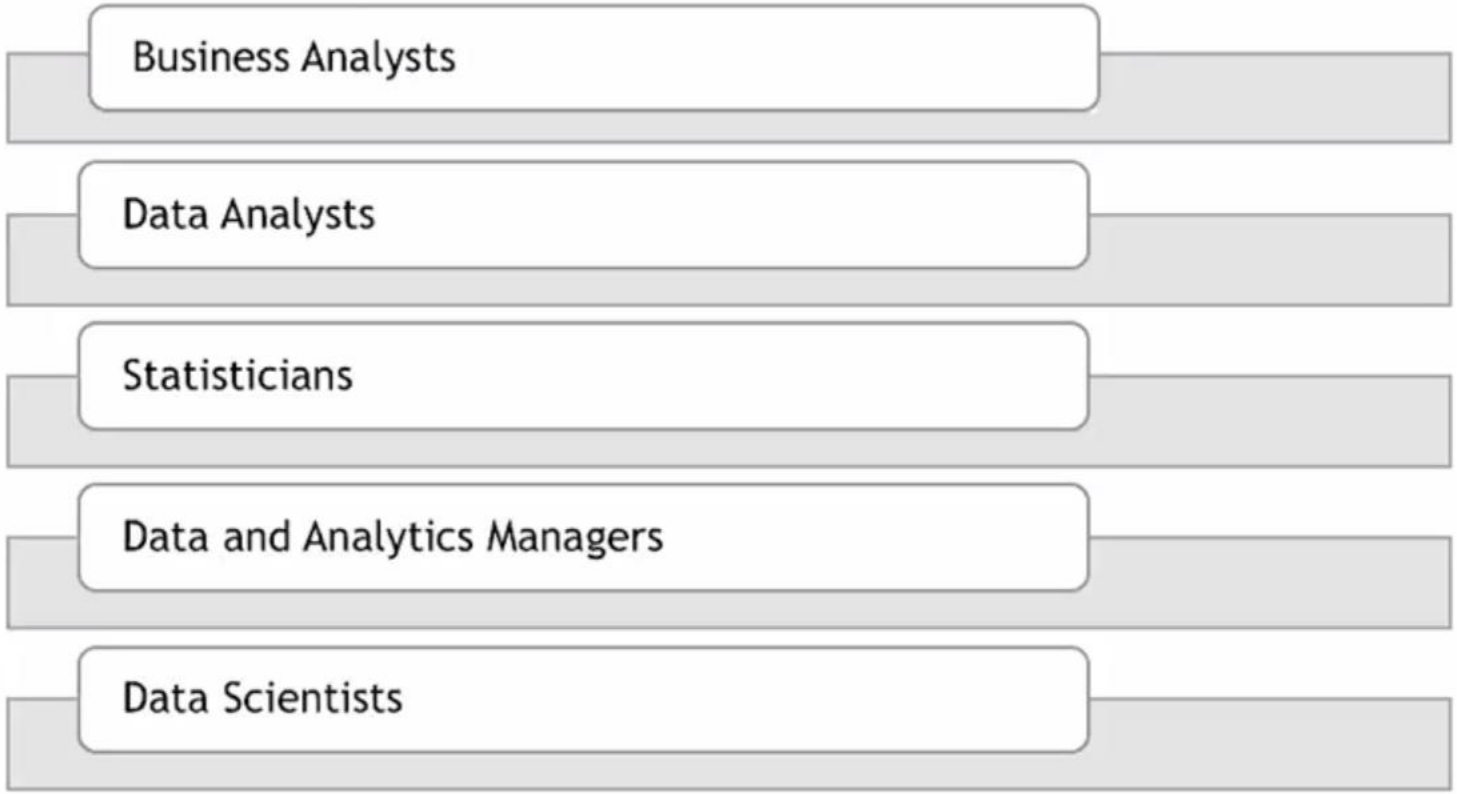
Top 10 Machine Learning Algorithms



Back

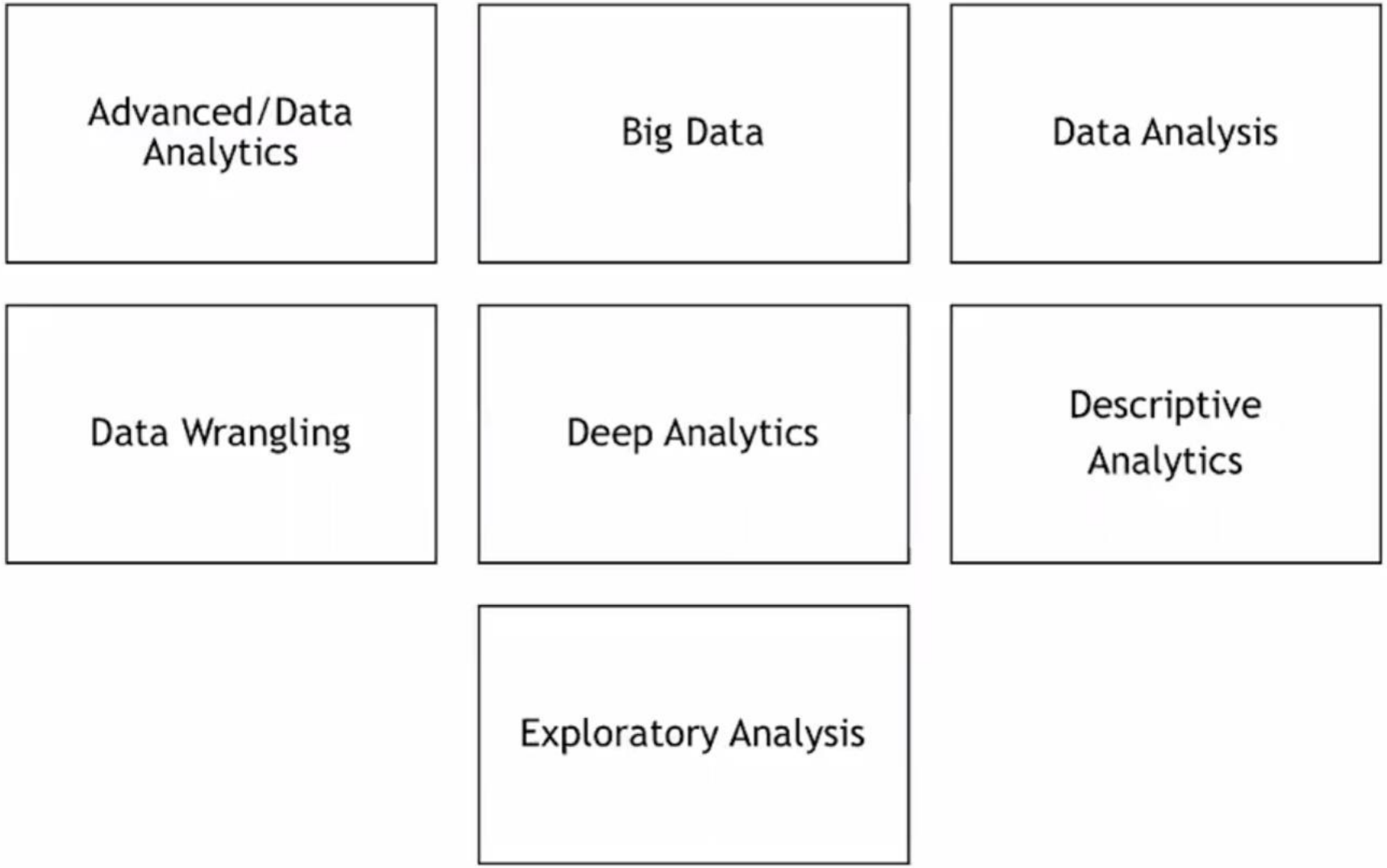
Jobs in Data Science..

Data science is inter-disciplinary and draws from many fields like statistics, mathematics, computer science, and business management to collect, organize, analyze, and interpret data. The most popular jobs in data science are as follows:



Back

The Data Science Dictionary



Back

The Data Science Dictionary

Business Analytics Vs Data Science

Back

Types of Predictive Analytics

Forecasting

- Used to forecast outcomes that are of a continuous nature
- How much will customer Y spend in the next month?
- What is the forecasted movement of our portfolio on a weekly basis for the next 12 weeks?

Segmentation

- Used to bucket or cluster things
- Customers in a particular segment have similar behaviors

Logistic Regression

- Used to predict a binary outcome
- Will Customer X respond or not to y marketing offer?
- What is the chance Customer Y will un-enroll in the next 12 months?

Back

Data Science Applications-01

University of Nottingham researchers created an Al system that scanned routine medical data to predict which patients would have strokes or heart attacks within 10 years.

https://futurism.com/neoscope/confirmed-ai-can-predict-heart-attacks-and-strokes-more-accurately-than-doctors

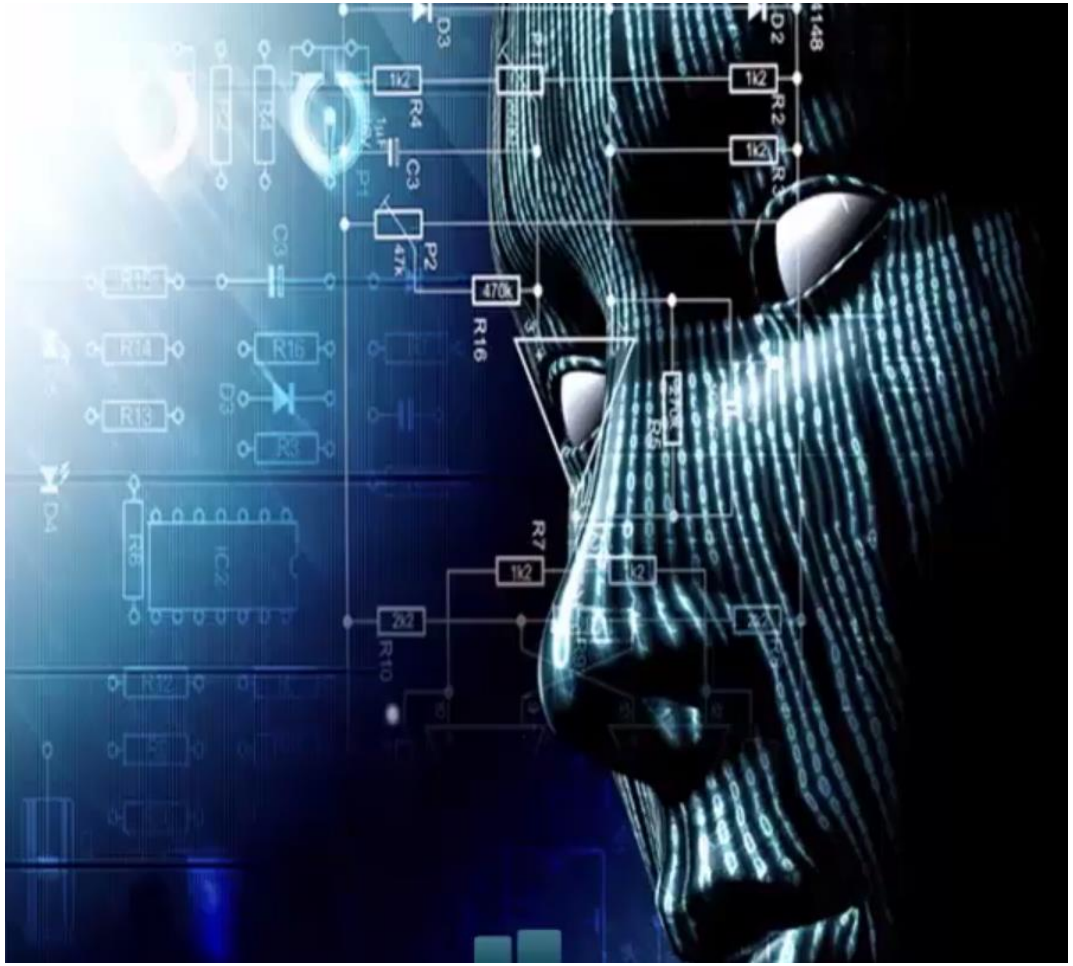
https://hbr.org/2017/05/how-machine-learning-is-helping-us-predict-heart-disease-and-diabetes

Back

Data Science Applications-02

Finding Criminals

Police officials in Durham, U.K. are slated to roll out an artificial intelligence system designed to help the authorities determine whether or not a suspect should be kept in police custody. The system, known as the Harm Assessment Risk Tool (HART) will classify suspects as low, medium, or high risk offenders.



https://www.business-standard.com/article/news-ians/alwar-police-testing-ai-based-app-to-register-criminal-offences-117052901171_1.html

https://www2.deloitte.com/content/dam/Deloitte/ie/Documents/PublicSector/deloitte-uk-future-of-policing.pdf

Back

Data Science Applications-03

Machine Learning In Retail

what to stock, how much to buy, what products to suggest to repeat customers. But doing more with that data using machine learning is just what retailers need to really succeed in the current market.

https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world

https://www.forbes.com/sites/bobevans1/2017/06/20/how-google-and-amazon-are-torpedoing-the-retail-industry-with-data-ai-and-advertising/#4327ff5f5c66

Back

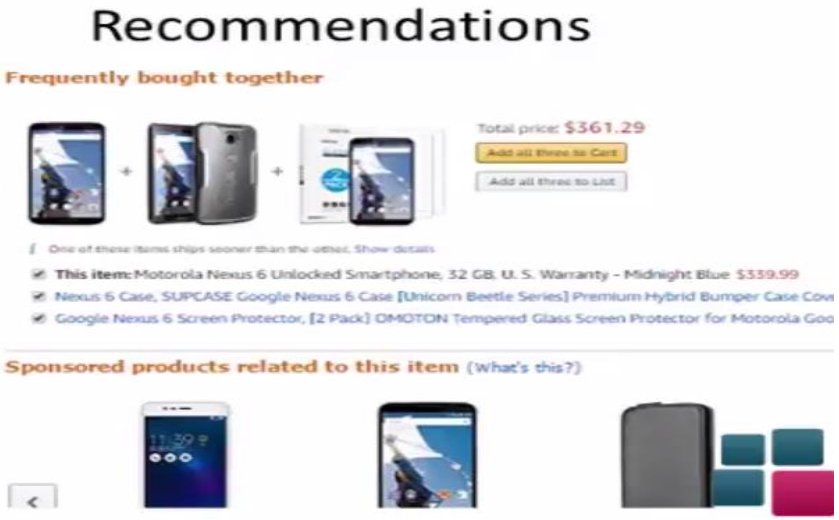
Data Science Applications-04

HR Analytics

Facial Recognition

Skin Cancer Deduction

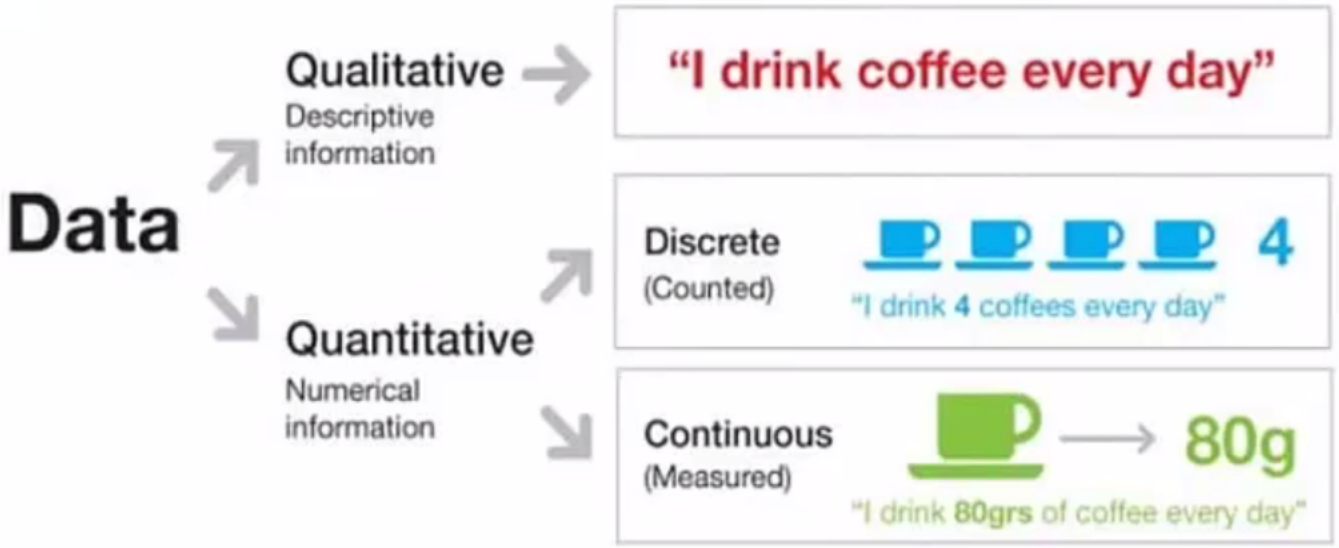
Recommendations

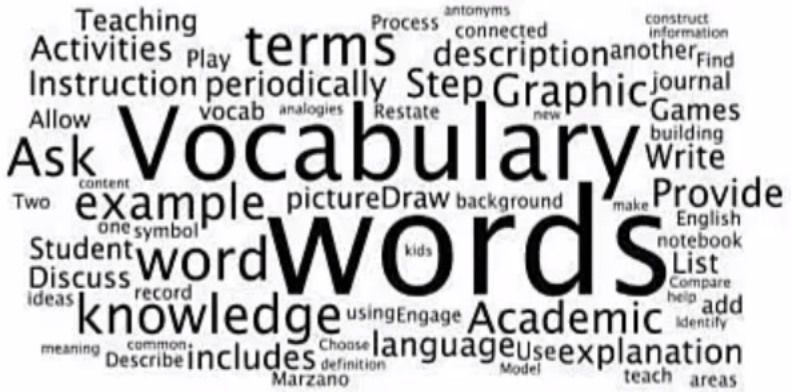


Back

Data

Data is a collection of facts, such as numbers, words, measurements, observations or even discretion of things.





Participant Number	Age	Gender	Place of Origin	Average years of residence	Years of education	Number of activities attended per year
1	32	Male	Minneapolis	9.7	6	2
2	48	Female	Saint Paul	11.6	10	5
3	40	Male	Chaska	7.5	12	3

How can I use this data – Analytics

Qualitative data: Descriptive information

- "Your friends house is pretty good"
- "Amazon Echo is the best AI Assistant"
- "Computer Vision is the new area of research"

Usage: Text mining, NLP, Sentiment analysis.

Sentiment analysis

Quantitative data: Data represents some quantity (numerical value).

It is of two types

- Discreate data: can take certain value (whole number)
- Continues data: can take any value (range of values)

Usage: Predictive analytics, Classification, Regression

Sales Predictions

Data, Data Analytics

Various types of Data (Storage):

1. *Structured data:* DB, ERP systems, CRM
2. *Semi- structured data:* Log files, XMLs
3. *Unstructured data:* Facebook, twitter

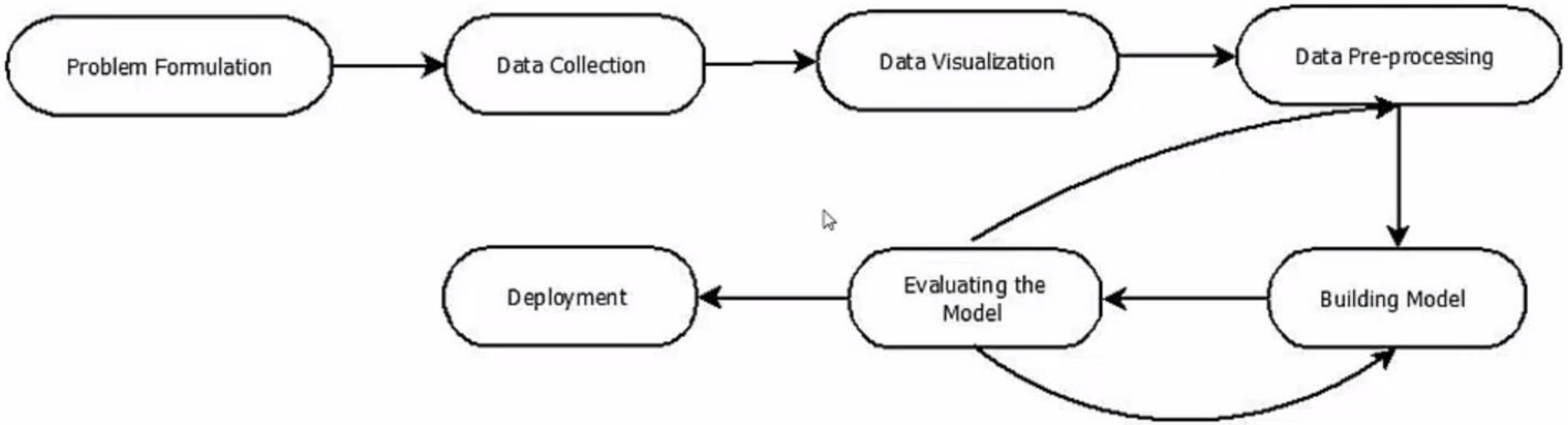
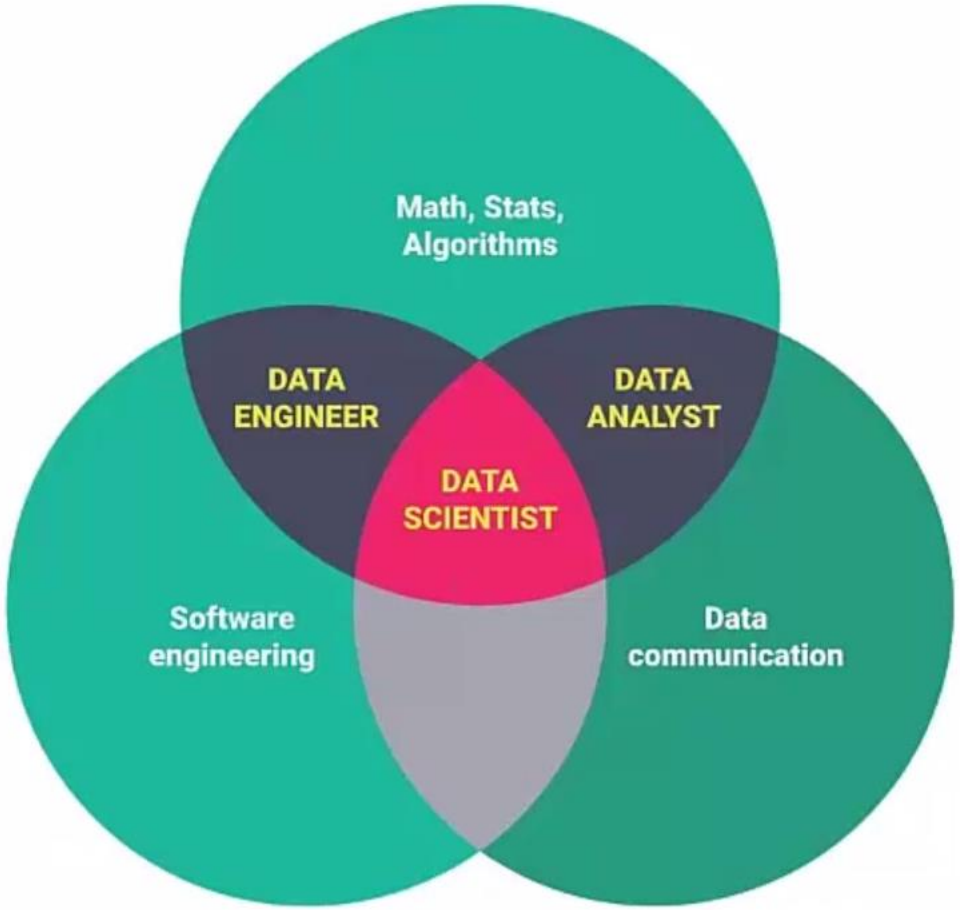


Fig: Data Analytics Process

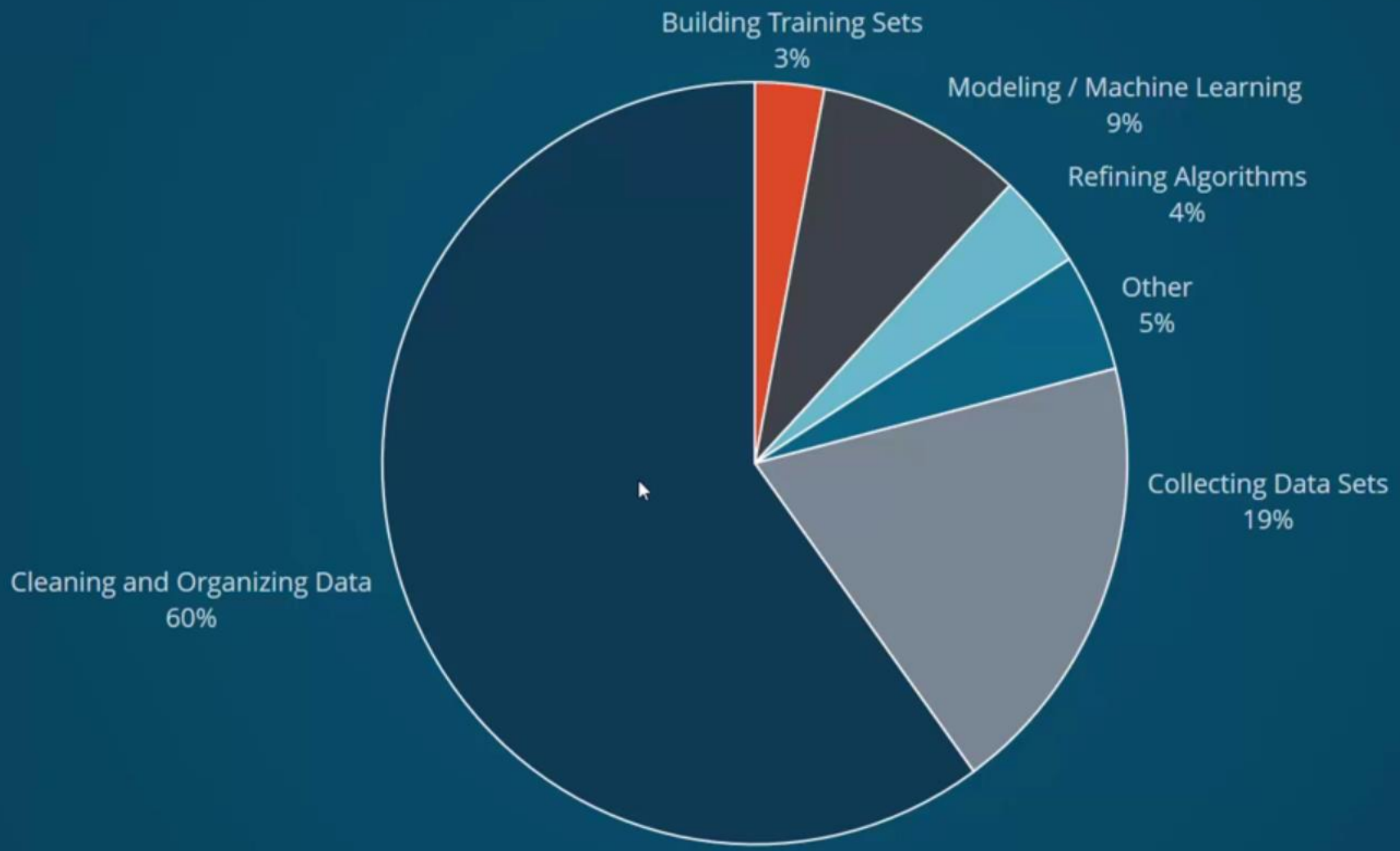
Career in Data Science

- Data Scientist
  - Uses Various Algorithms to solve complex business problems
- Data Engineer
  - Tries to handle huge data with proper constancy and availability
- Data Analysts
  - Analysts draws the insights from raw data and understands/study the data by visualizing and grouping it

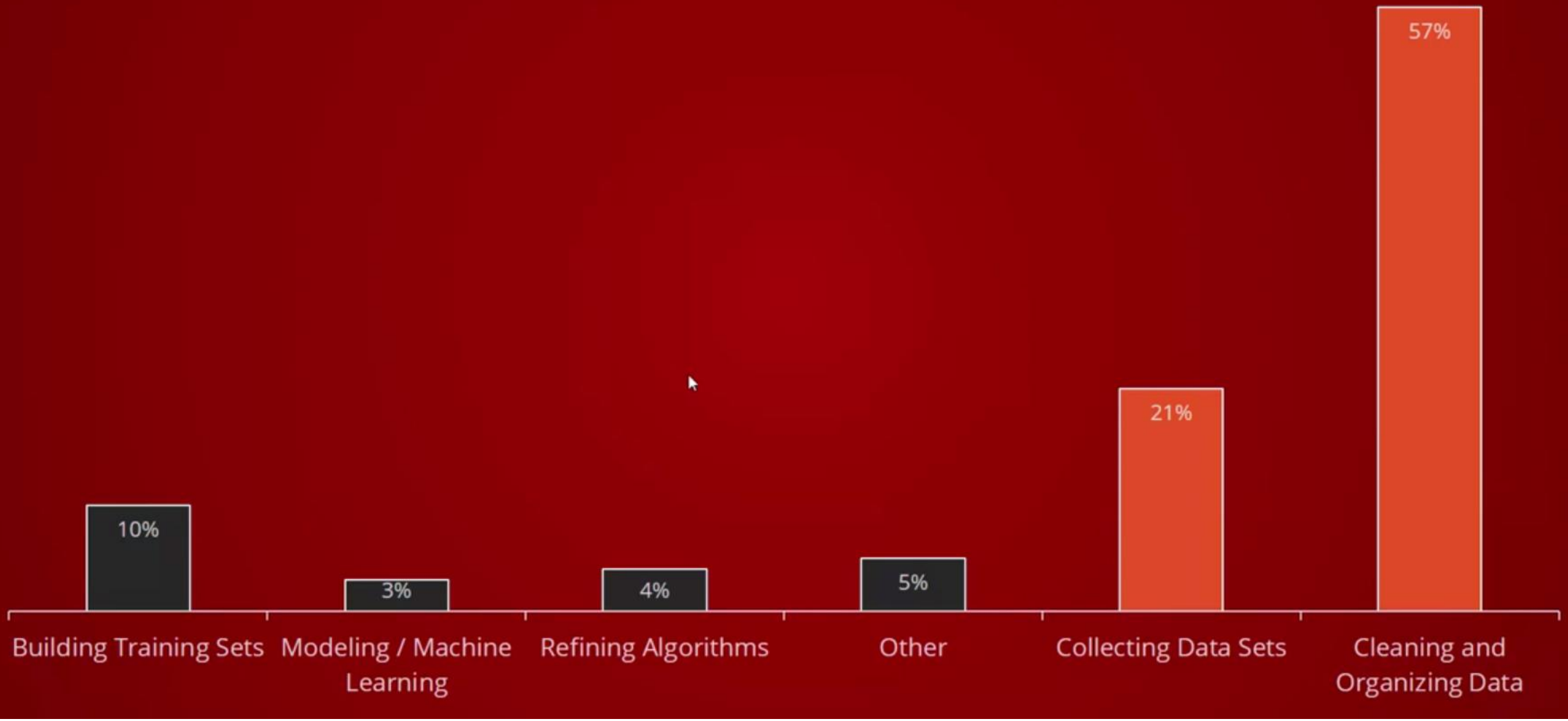


Back

What data scientists spend the most time doing



What's the least enjoyable part of data science?



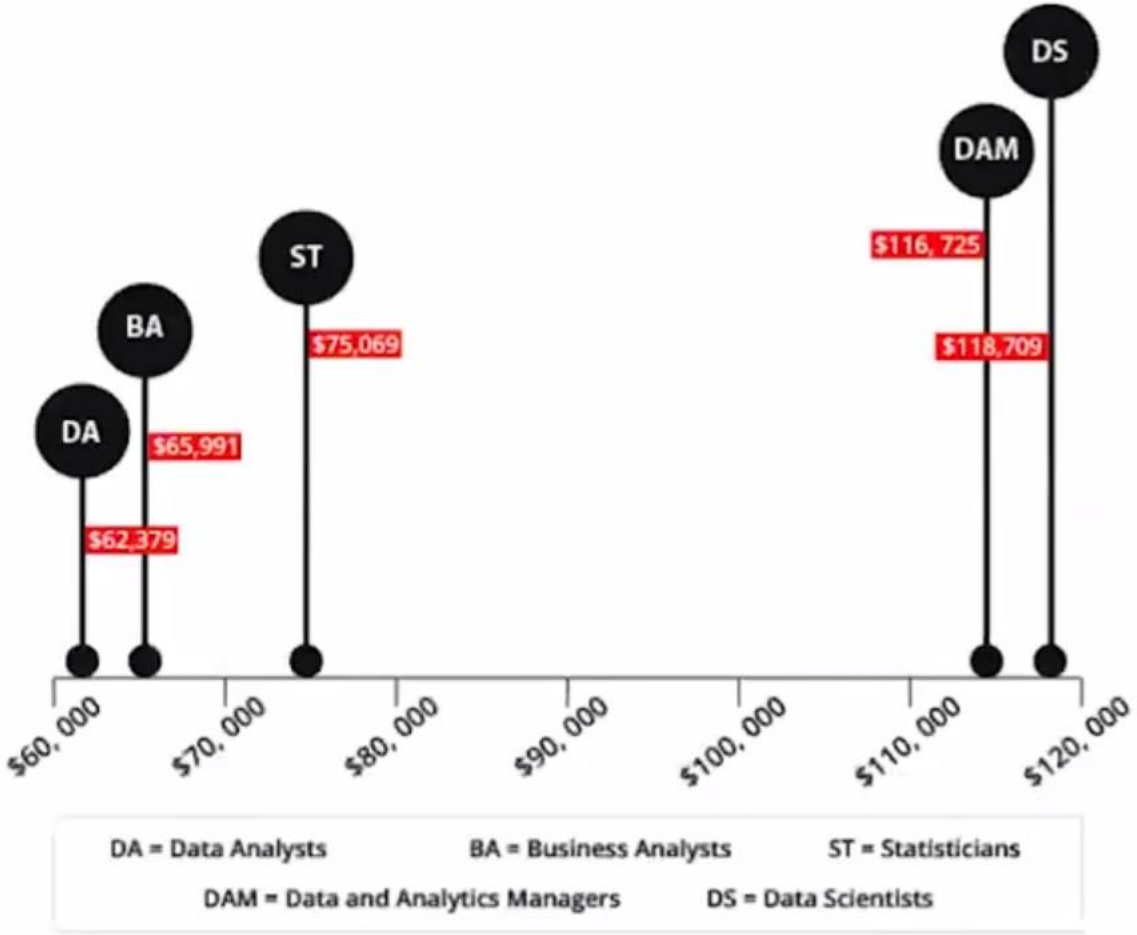
Job Trends

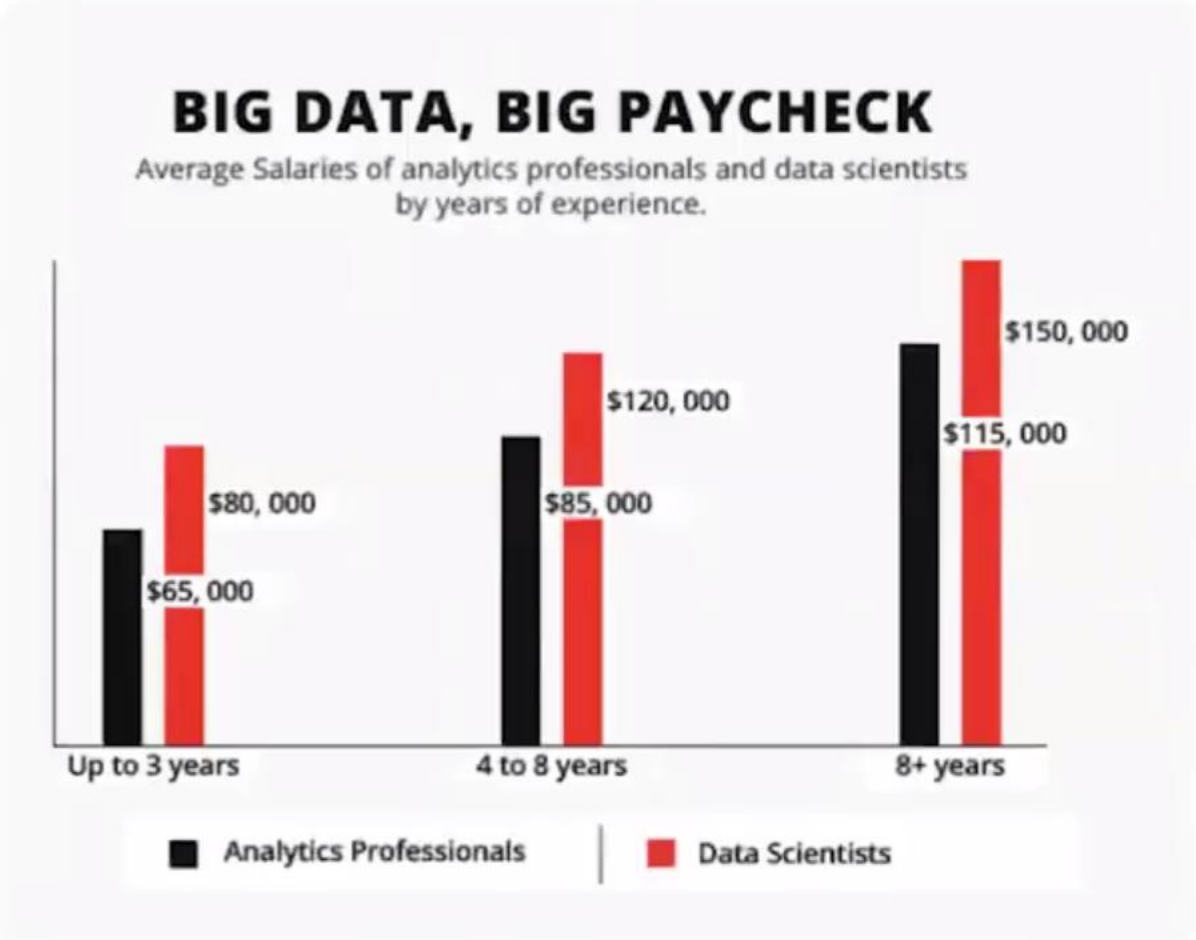
Data Scientist salaries by company in the United States

Company	Average salary
Cisco Data Scientist 12 salaries Data Scientist Job available	**$185,853** per year
Harnham Data Scientist 560 salaries Data Scientist Job available	**$155,747** per year
Apple Data Scientist 11 salaries Data Scientist Job available	**$147,169** per year
Airbnb Data Scientist 27 salaries Data Scientist Job available	**$138,019** per year
Facebook Data Scientist 109 salaries Data Scientist Job available	**$136,384** per year

Source – Indeed.com

Salary of Data Science Professionals





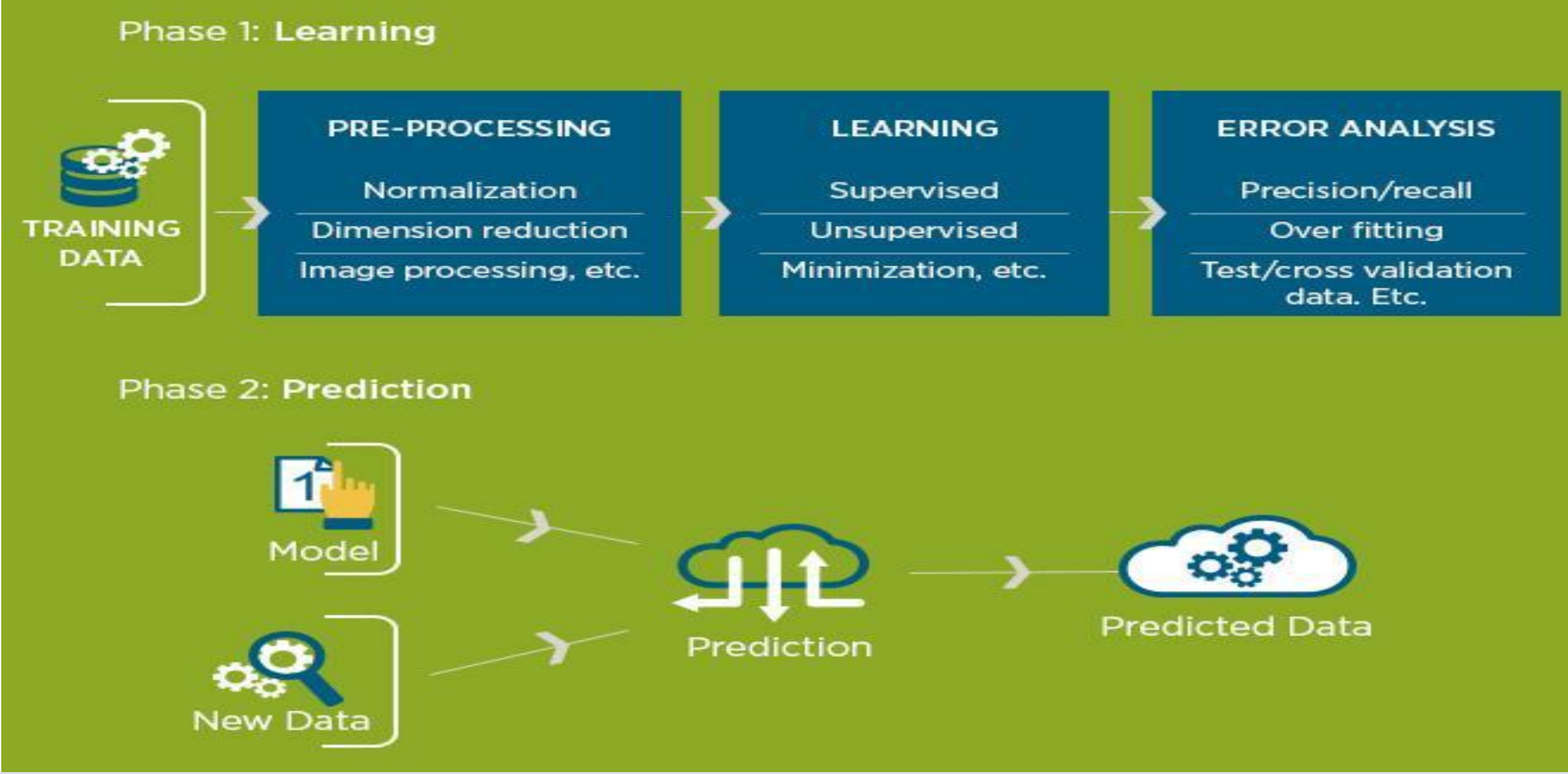
Back

The Data Science Dictionary

Feature	Predictive Analytics	Prescriptive Analysis
Production Code	Product Requirements Document (PRD)	Statement of Work (SoW)
	Target Variable	

Back

MACHINE LEARNING PROCESS



Data Science & Machine Learning Case Study

Chat Bots

Sentiment analysis

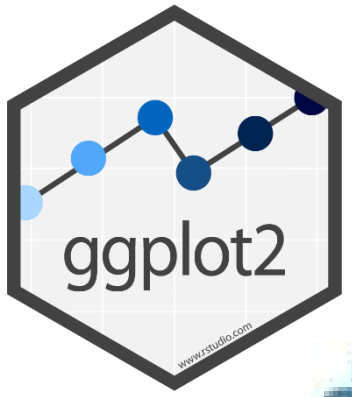
Sales Predictions

Self Driven Cars

Facial Exp recognition

Image Tagging

Tools,Programing Platforms



Power BI

kaggle

ANACONDA
Powered by Continuum Analytics®

APACHE Spark™

Back

Roles & Career

Data Science Developers

Big Data Specialists

Infra Engineers

Data Scientist

Researchers

Analysts

Business Person

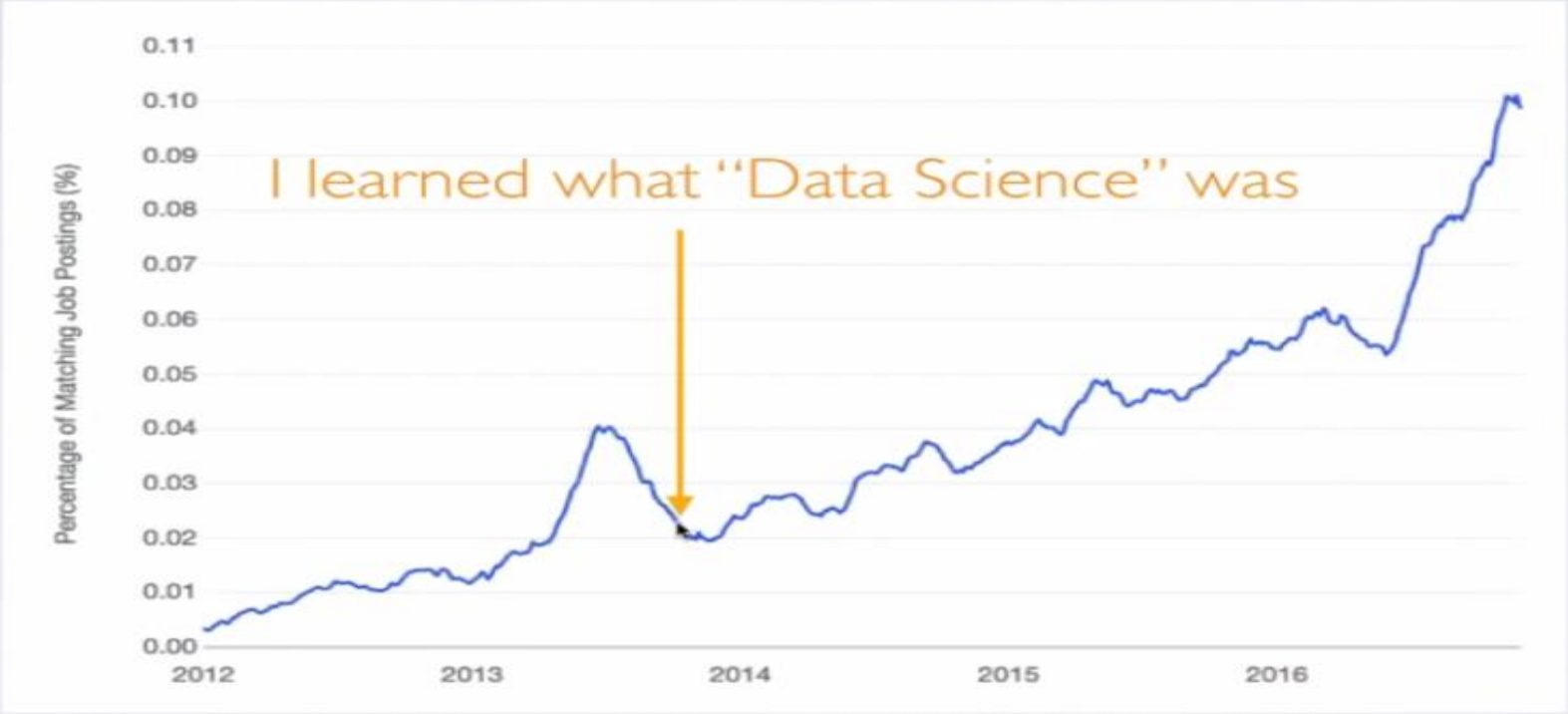
Back

Conclusion

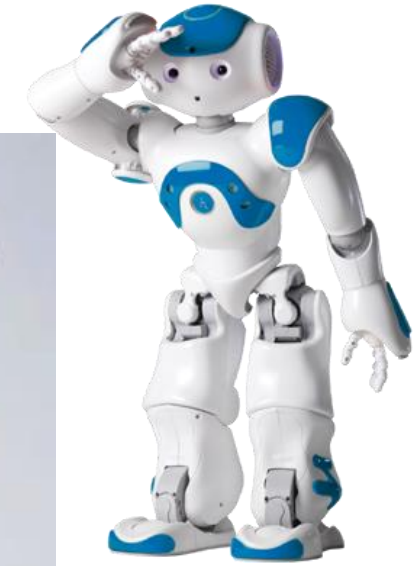
Data science refers to the processes and methods that help make sense of large volumes of data for organizational purposes. They are in high demand as their skills are extensively required across various sectors. This book was put together to set aspiring data scientists on a novel, exciting and fruitful journey in data science.

Back

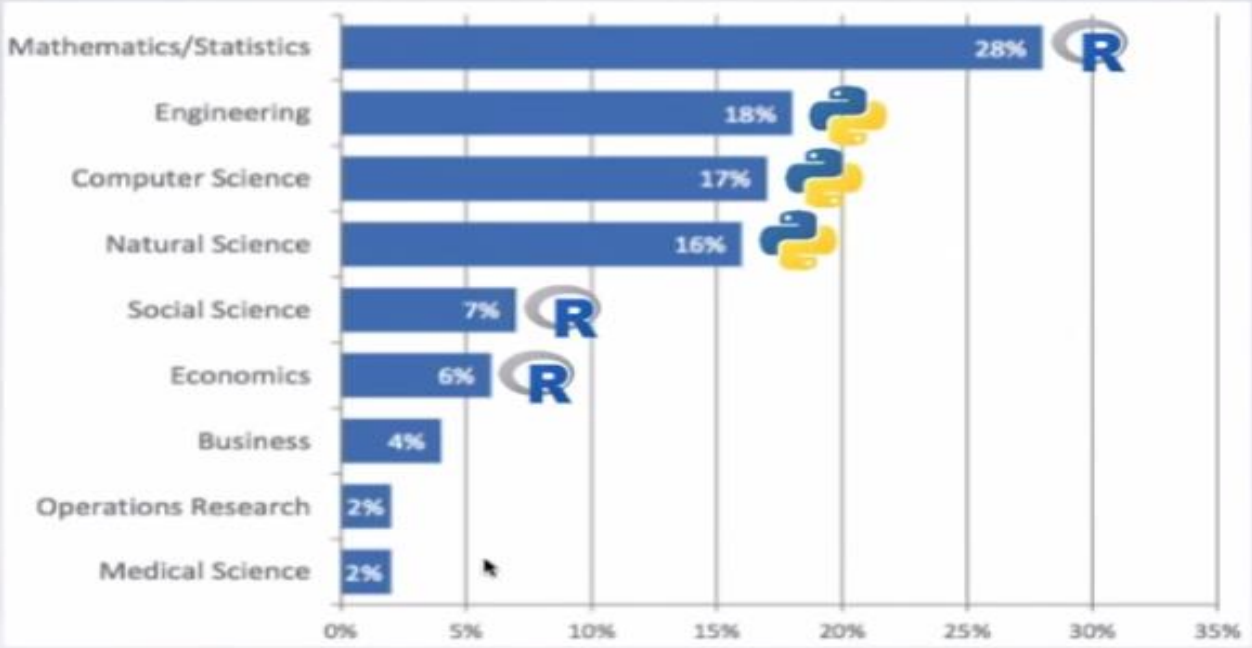
INDEED JOB TRENDS FOR "DATA SCIENTIST"



https://www.indeed.com/jobtrends/q-%22Data-Scientist%22.html



POPULAR DATA SCIENCE BACKGROUNDS



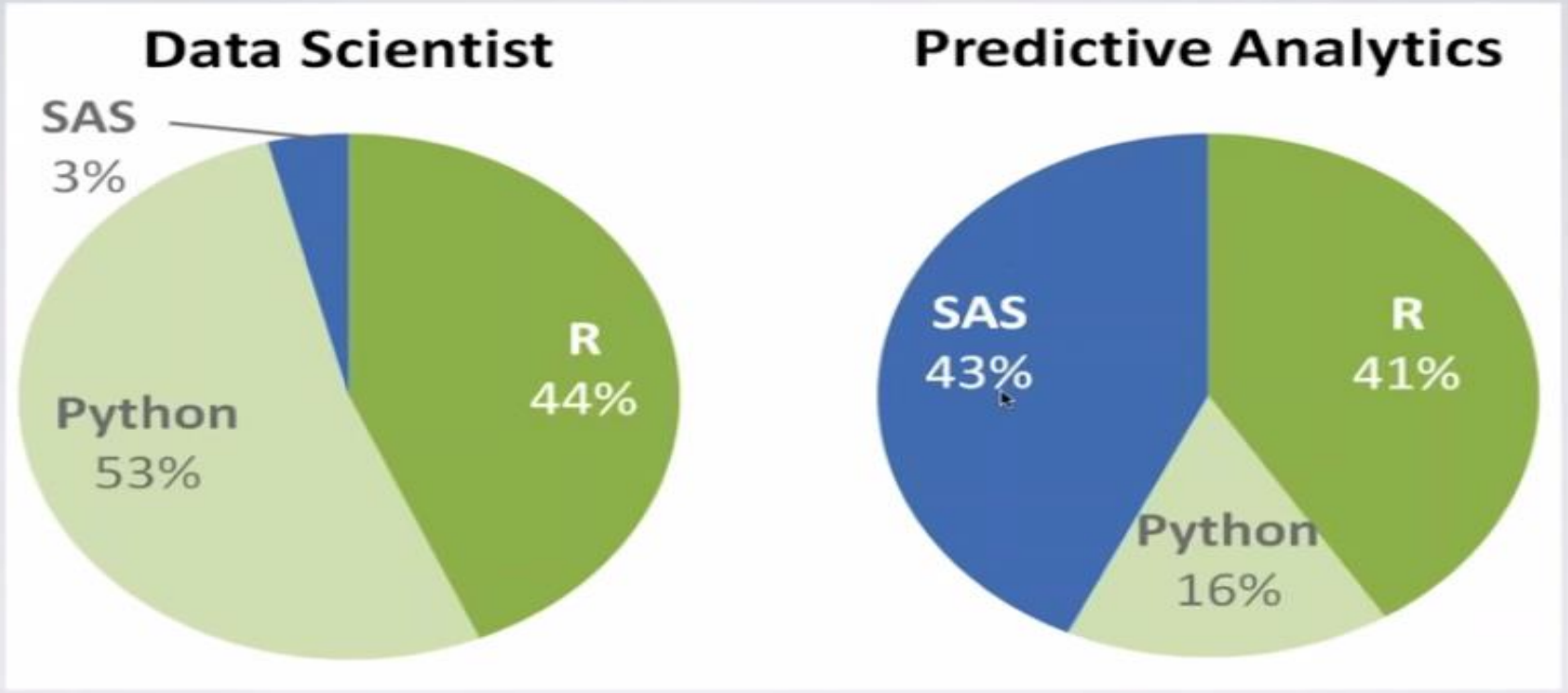
Burtch Works 2016 Study, Data Scientist Backgrounds

http://www.burtchworks.com/files/2016/04/Burtch-Works-Study_DS-2016-final.pdf

= R more common

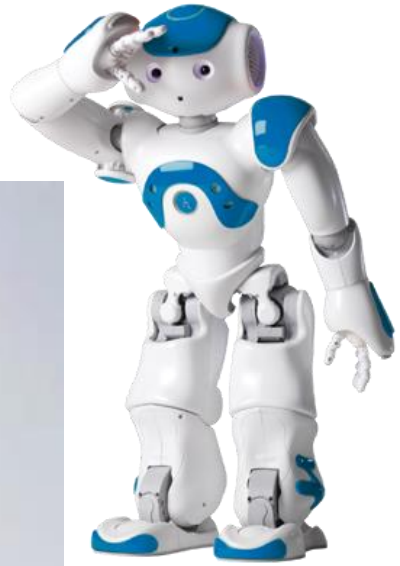
= Python more common

TOOL USE BY POSITION



Burtch Works 2016 Tool Survey

http://www.burtchworks.com/2016/07/13/sas-r-python-survey-2016-tool-analytics-pros-prefer/



PYTHON DATA STACK

NumPy

SciPy

matplotlib

IP[y]: IPython
Interactive Computing

pandas

scikit learn

Learn these libraries well!

ALTERNATIVE FACT #1

Most of Data Science is fine-tuning models to get the highest performance possible

REALITY:

You are going to spend most of your time cleaning/merging data

ALTERNATIVE FACT #2

Big Data is EVERYWHERE! You will need Hadoop and Spark all the time to solve every problem!

REALITY:

With exceptions, most problems can be handled on a single machine

ALTERNATIVE FACT #3

Deep Learning solves EVERYTHING! Other methods are obsolete.

REALITY:

You probably don't need it, unless you are working with images and want to maximize performance

AUDIENCE QUESTIONS

How can a college fresher (say, studying in sophomore or final year) become a data scientist? What projects can they do? What skills should they focus on? How to start applying for jobs?

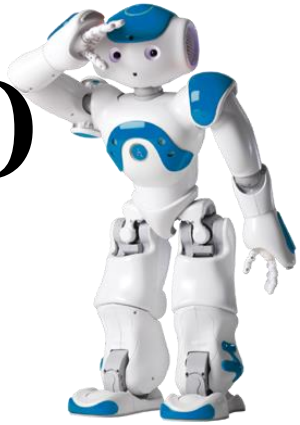
AUDIENCE QUESTIONS

How can an experienced professional make a career shift into data science? Let's say, someone has 3 years of experience in Java, and they now want to become a data scientist. Or, let's say, someone knows Hive, Pig, Flume, Hadoop, what could be a natural career progression for them?

AUDIENCE QUESTIONS

What are the most commonly used ML algorithms in industry today, so that students can master them first ?

Fresher Salaries In India(INR)



2017

1 **Data Scientist**

4.8 / 5
Job Score

4.4 / 5
Job Satisfaction

$110,000
Median Base Salary

4,184
Job Openings

View Jobs

https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

MACHINE LEARNING

References

[1] Python-course.eu, 'Python3 Tutorial: Python Online Course', 2015. [Online]. Available: http://www.python-course.eu/python3_course.php.

[2] www.tutorialspoint.com, 'Python tutorial', 2015. [Online]. Available: http://www.tutorialspoint.com/python/index.htm.

[3] Wikipedia, 'History of Python', 2015. [Online]. Available: https://en.wikipedia.org/wiki/History_of_Python#Version_release_dates.

[4] Docs.python.org, 'What's New In Python 3.0', 2015. [Online]. Available: https://docs.python.org/3/whatsnew/3.0.html.

[5] Sebastianraschka.com, 'Python 2.7.x and Python 3.x key differences', 2015. [Online]. Available: http://sebastianraschka.com/Articles/2014_python_2_3_key_diff.html.

[6] Programcreek.com, 'Java vs. Python: Why Python can be more productive?', 2015. [Online]. Available: http://www.programcreek.com/2012/04/java-vs-python-why-python-can-be-more-productive/.

[7] Stsdas.stsci.edu, 'A Quick Tour of Python', 2015. [Online]. Available: http://stsdas.stsci.edu/pyraf/python_quick_tour.html.

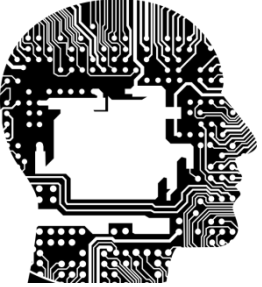
[8] Lynda.com - A LinkedIn Company, 'Python 3 Essential Training | Lynda.com Training', 2015. [Online]. Available: http://www.lynda.com/Python-3-tutorials/essential-training/62226-2.html.

[9] Pymbook.readthedocs.org, 'Welcome to Python for you and me — Python for you and me 0.3.alpha1 documentation', 2015. [Online]. Available: http://pymbook.readthedocs.org/en/latest/index.html.

[10] Code Geekz, '10 Best Python IDE for Developers | Code Geekz', 2014. [Online]. Available: https://codegeekz.com/best-python-ide-for-developers/.

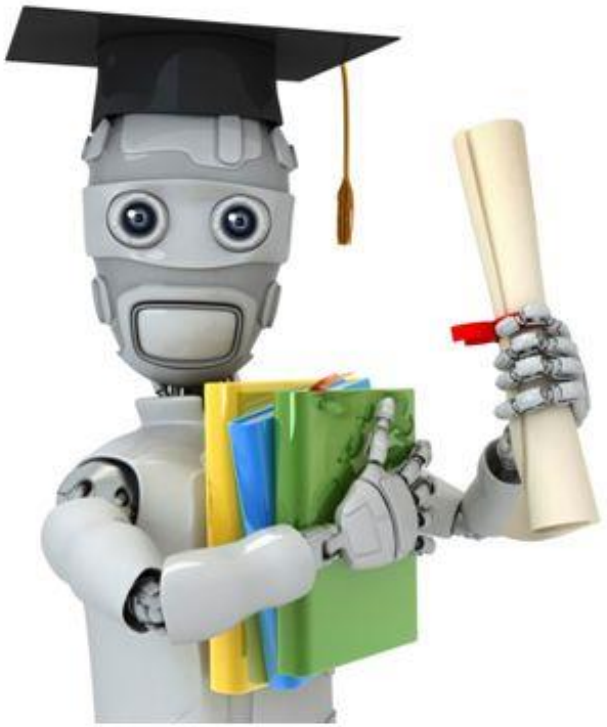
[11] K. Radhakrishnan, 'Top 10 Python Powered Web Frameworks For Developers', *Toppersworld.com*, 2014. [Online]. Available: http://toppersworld.com/top-10-python-powered-web-frameworks-for-developers/.

[12] Wiki.python.org, 'OrganizationsUsingPython - Python Wiki', 2015. [Online]. Available: https://wiki.python.org/moin/OrganizationsUsingPython.



MACHINE LEARNING

Spark



BIG DATA

hadoop

Thank You

theano

Pandas

Machine Learning

AI

TensorFlow

python™


To be Continued……