

Data Science



Jnaneswar Bohara

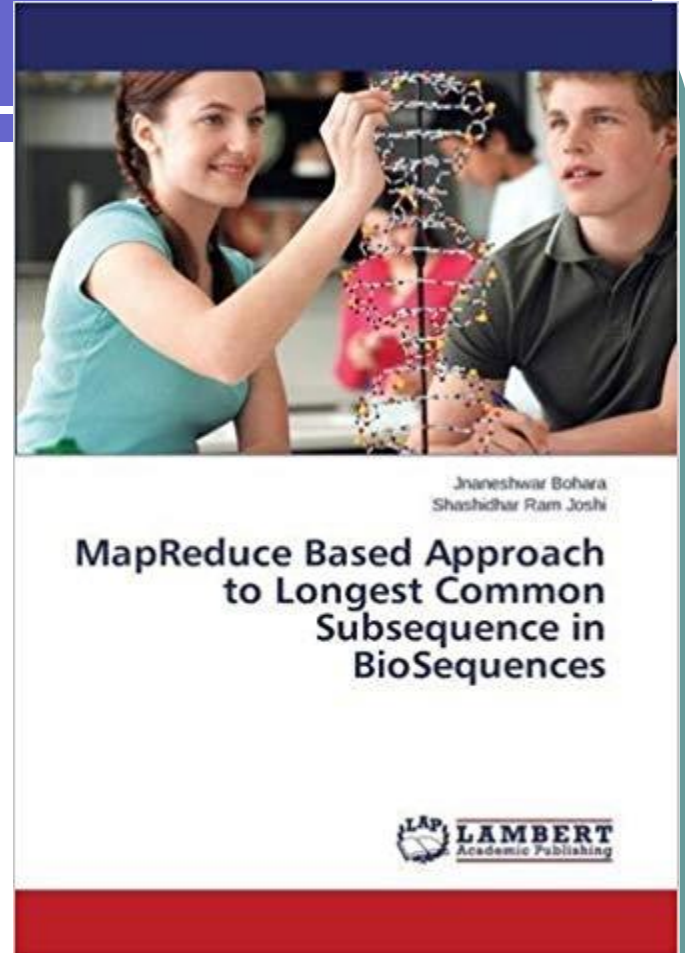
Know Your Instructor

- Jnaneshwar Bohara
 - M. Sc. Computer System and Knowledge Engineering, IOE, TU (Gold Medal)
 - Certified Scrum Master
 - Senior Java Programmer
 - Big Data Analyst



Know Your Instructor

- Jnaneshwar Bohara
 - Researcher on Big Data and Bioinformatics



<https://www.amazon.com/MapReduce-Approach-Longest-Subsequence-BioSequences/dp/3659680508>

Team Members

- Mr. Dinesh Saud
- Mr. Sunil Ghimre

What is Data Science?

- An **interdisciplinary** field that uses scientific methods, processes, algorithms and systems to extract **knowledge** and insights from structured and **unstructured data**, and apply knowledge and actionable insights from data across a broad range of application domains
- Is related to **data mining**, **machine learning** and **big data**

What is Data Science?

- An area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data
- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data
- Data science principles apply to all data – big and small

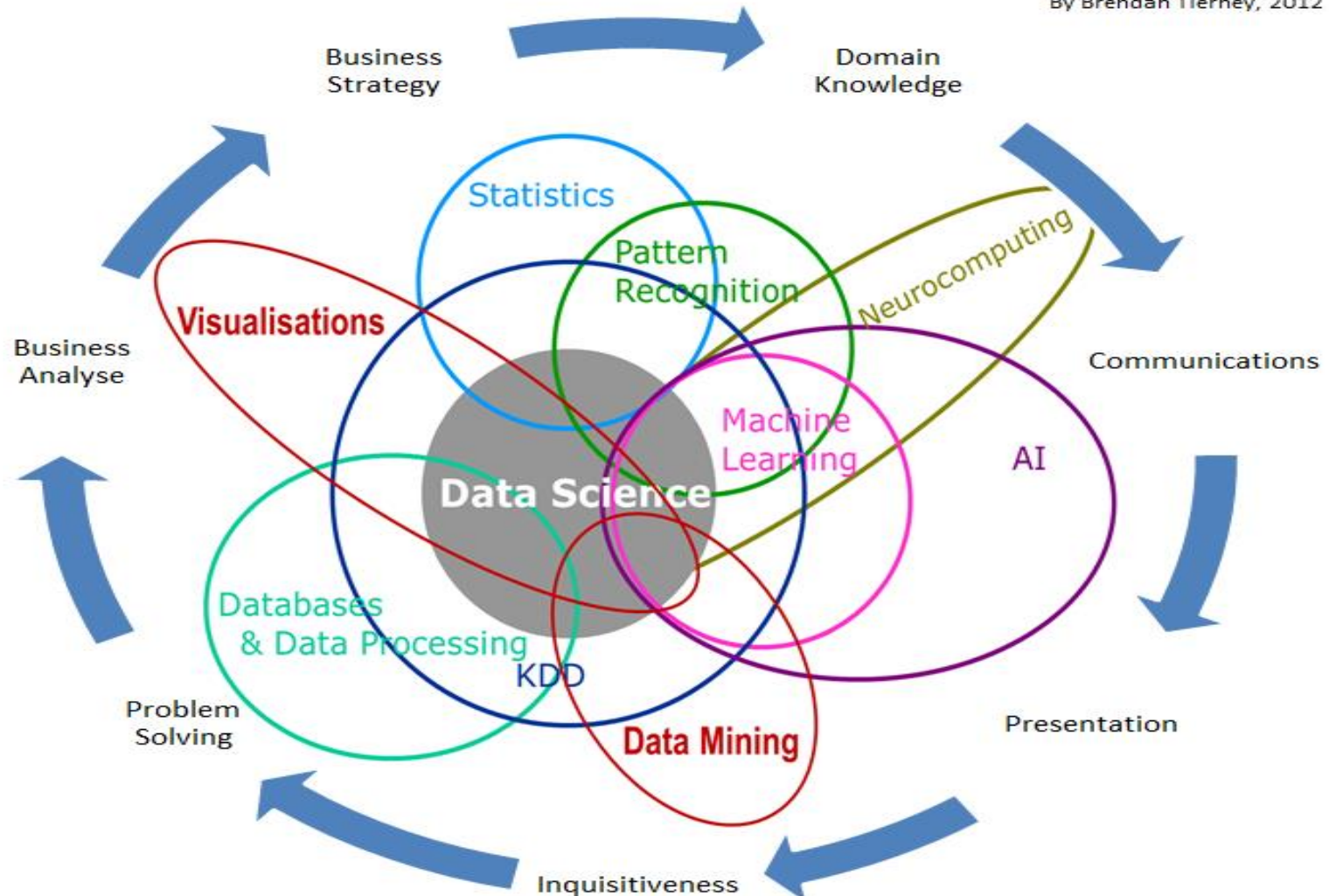
What is Data Science?

- Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education
 - Computer Science
 - Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI
 - Mathematics
 - Mathematical Modeling
 - Statistics
 - Statistical and Stochastic modeling, Probability.

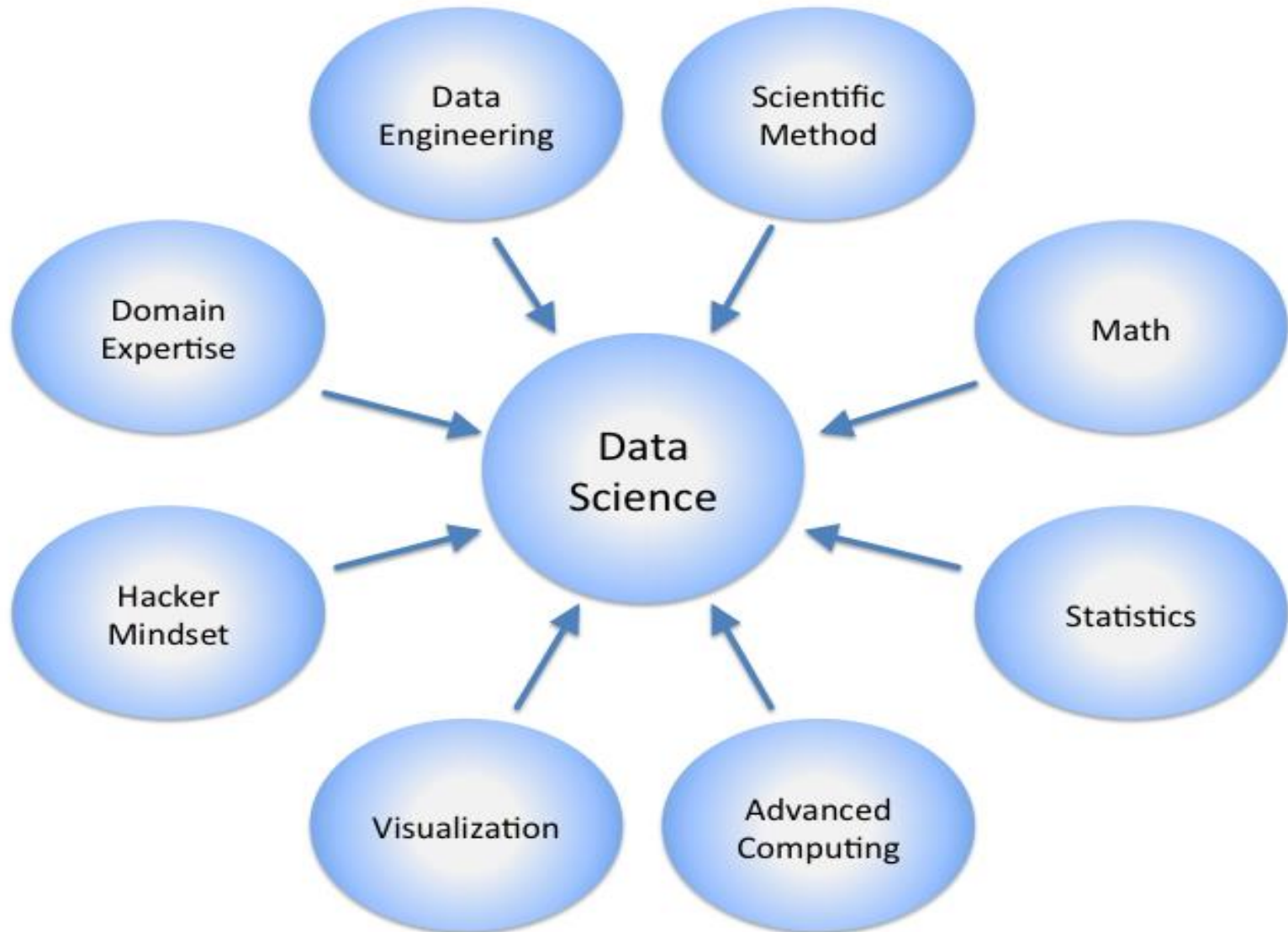
What is Data Science?

Data Science Is Multidisciplinary

By Brendan Tierney, 2012



What is Data Science?



Real Life Examples

- Identifying and predicting disease
- Personalized healthcare recommendations
- Optimizing shipping routes in real-time
- Getting the most value out of soccer rosters
- Stamping out tax fraud
- Automating digital ad placement

<https://builtin.com/data-science/data-science-applications-examples>

Data Scientists

- Data Scientist
 - The Sexiest Job of the 21st Century
- They find stories, extract knowledge.
They are not reporters



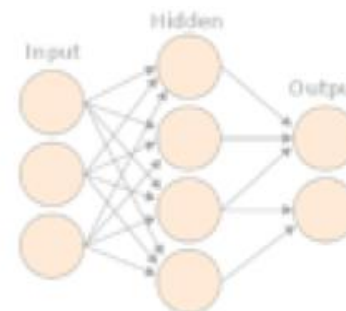
Data Scientist

- High ranking professional with training and curiosities to make discovery in the world of big data.
- The people who understand how to fish out answers to important business questions from today's tsunami of unstructured information.
- Newly coined term , in 2008 by D.J Patil and Jeff Hammerbacher
- A hybrid of data hacker, analyst, communicator, and trusted adviser. The combination is extremely powerful—and rare

Data Scientist

The Data Scientist

- A New Role Exists – the **Data Scientist**
 - One Part Scientist/Statistician
 - Two Parts Sleuth/Artist
 - One Part Programmer
 - Focused on *data* not models
- Working with **analysts** to create business value



Confidential Think Big Analytics

Data scientist: a brand new profession

- Data Scientist: The Sexiest Job of the 21st Century [Harvard Business Review 2013]
- Data scientist? A guide to 2015's hottest profession [Mashable 2015]
- “It’s official – data scientist is the best job in America” [Forbes, 2016]
- "This hot new field promises to revolutionize industries from business to government, health care to academia."
— *The New York Times*

Successful Data Scientist Characteristics

- **Intellectual curiosity, Intuition**
 - Find needle in a haystack(something that is difficult to locate in a much larger space)
 - Ask the right questions – value to the business
- **Communication and engagements**
- **Presentation skills**
 - Let the data speak but tell a story
 - Story teller – drive business value not just data insights
- **Creativity**
 - Guide further investigation
- **Business Savvy**
 - Discovering patterns that identify risks and opportunities
 - Measure

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants



PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing package e.g. R
- ☆ Databases SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants



- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Concentration in Data Science

- Mathematics and Applied Mathematics
- Applied Statistics/Data Analysis
- Solid Programming Skills (R, Python, Julia, SQL)
- Data Mining
- Data Base Storage and Management
- Machine Learning and discovery

Data Mining

- The process of discovering meaningful patterns and trends often previously unknown by using some mathematical algorithm on huge amount of stored data.
- Extraction of interesting, non-trivial, implicit, previously unknown and potentially useful information or patterns from data in large database.
- Data mining is basically concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data.

Data Mining

- Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.

Valid: The patterns hold in general.

Novel: We did not know the pattern beforehand.

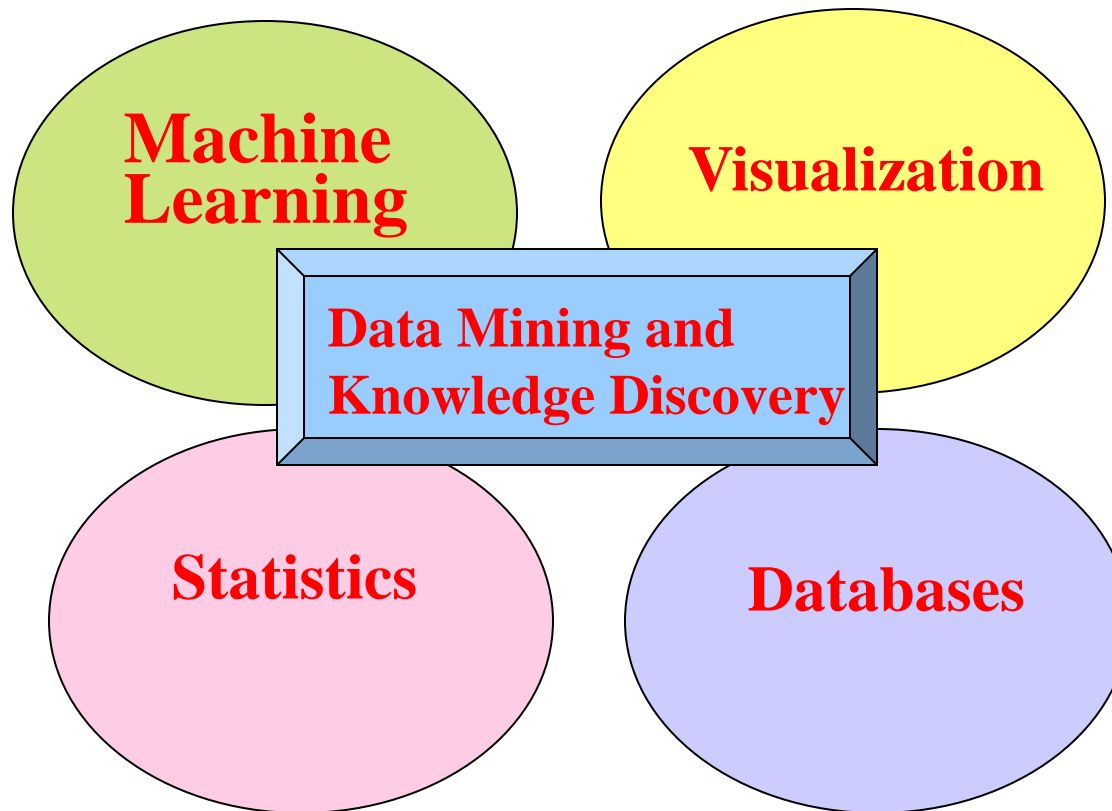
Useful: We can devise actions from the patterns.

Understandable: We can interpret and comprehend the patterns.

Data Mining

- Finding interesting structure in data
- *Structure*: refers to statistical patterns, predictive models, hidden relationships
- Examples of tasks addressed by Data Mining
 - Predictive Modeling (classification, regression)
 - Segmentation (Data Clustering)
 - Summarization
 - Visualization

Related Fields in Data Mining



Data Science Vs. Data Mining

Basis for comparison	Data Science	Data Mining
What is it?	An area	A technique
Focus	Scientific study	Business process
Goal	Building Data-centric products for an organization	Make data more usable
Output	Varied	Patterns
Purpose	Social analysis, building predictive models, unearthing unknown facts, and more	Finding trends previously not known
Deals with (the type of data)	All forms of data – structured, semi-structured and unstructured	Mostly structured
	Data Science Jnaneshwar Bohara	23

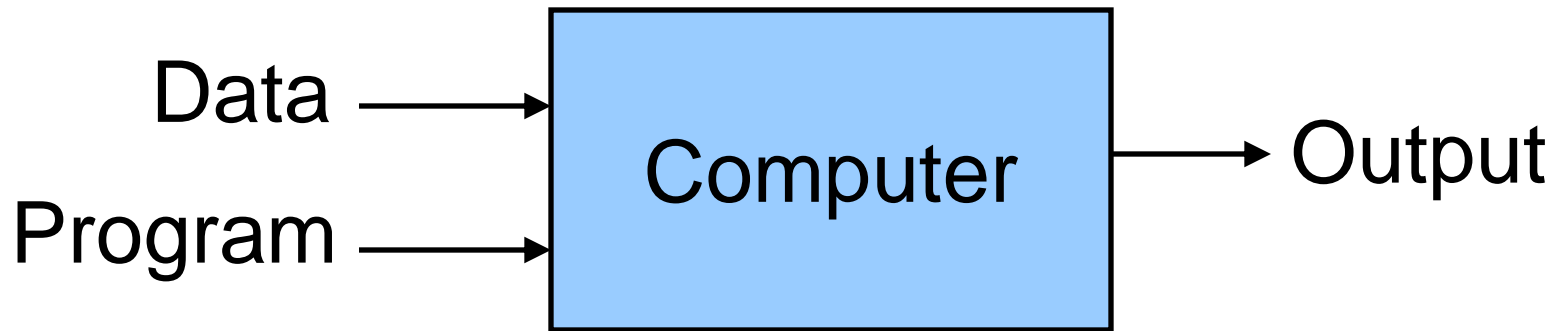
Data Science Vs Data Mining

Basis for comparison	Data Science	Data Mining
Vocational Perspective	A person needs to understand Machine Learning, Programming, info-graphic techniques and have the domain knowledge to become a data scientist	Someone with a knowledge of navigating across data and statistical understanding can conduct data mining
Extent	Multidisciplinary – Data Science consists of Data Visualizations, Computational Social Sciences, Statistics, Data Mining, Natural Language Processing, et cetera	Data mining can be a subset of Data Science as Mining activities are part of the Data Science pipeline
	Data Science Jnaneshwar Bohara	

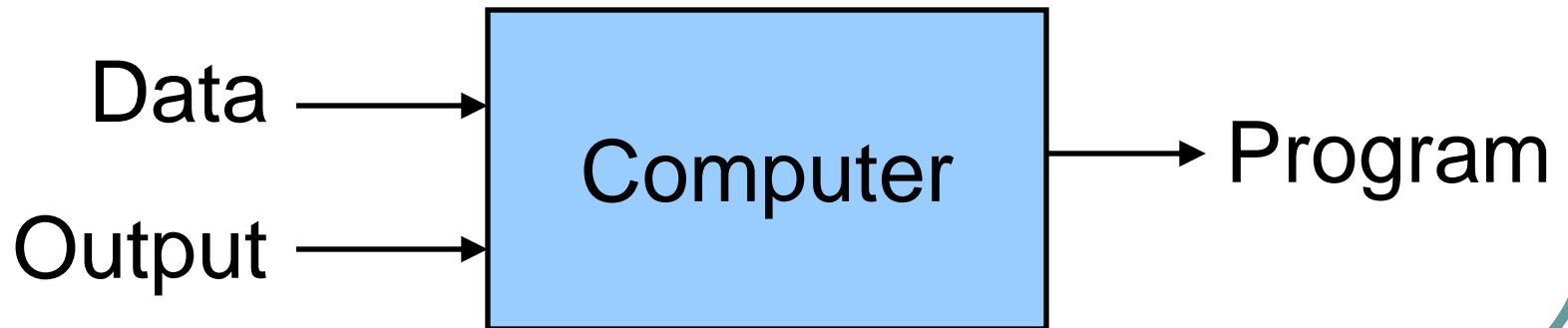
Machine Learning

- Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
- **Machine learning focuses on the development of computer programs** that can access data and use it to learn for themselves.

Traditional Programming



Machine Learning



How does machine learning work

- Select and prepare a training data set
- Choose an algorithm to run on the training data set
- Training the algorithm to create the model
- Using and improving the model

Machine Learning Methods

- **Supervised (inductive) learning**
 - Training data includes desired outputs
- **Unsupervised learning**
 - Training data does not include desired outputs
- **Semi-supervised learning**
 - Training data includes a few desired outputs
- **Reinforcement learning**
 - Rewards from sequence of actions

Deep learning

- Deep learning is a subset of machine learning (all deep learning is machine learning, but not all machine learning is deep learning).
- Deep learning algorithms define an artificial neural network that is designed to learn the way the human brain learns.
- Deep learning models require large amounts of data that pass through multiple layers of calculations, applying weights and biases in each successive layer to continually adjust and improve the outcomes.

Deep Learning

- Deep learning models are typically unsupervised or semi-supervised.
- Reinforcement learning models can also be deep learning models.
- Certain types of deep learning models—including convolutional neural networks (CNNs) and recurrent neural networks (RNNs)—are driving progress in areas such as computer vision, natural language processing (including speech recognition), and self-driving cars.

Machine Learning Example

- Facebook's machine learning algorithms gather behavioral information for every user on the social platform.
- Based on one's past behavior, the algorithm predicts interests and recommends articles and notifications on the news feed.
- When Amazon recommends products, or when Netflix recommends movies based on past behaviors, machine learning is at work.

Few widely publicized examples of machine learning applications

- The heavily hyped, self-driving Google car
- Online recommendation offers such as those from Amazon and Netflix
- Knowing what customers are saying about you on Twitter
- Fraud detection

Data Science Vs. Machine Learning

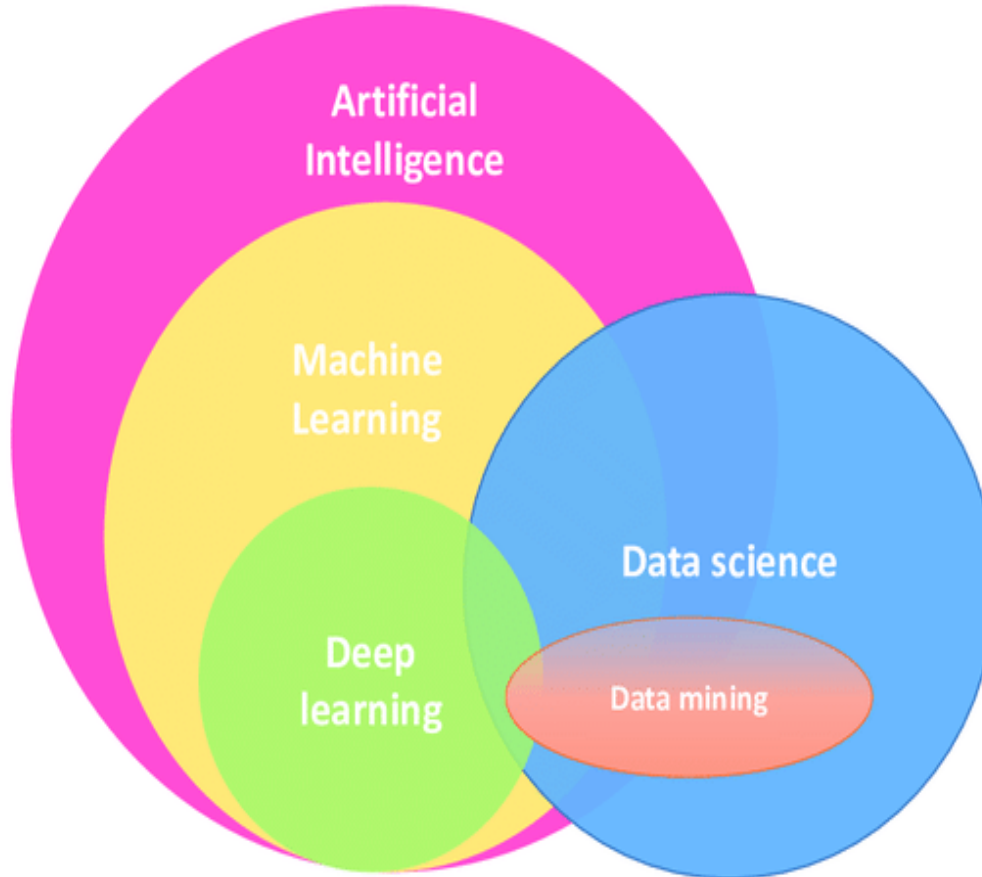
Data Science	Machine Learning
It is an interdisciplinary field where unstructured data is cleaned, filtered, analyzed and business innovations are churned out of the result.	It is a part of data science where tools and techniques are used to create algorithms so that the machine can learn from data via experience.
It has a vast scope	It comes only in the data modeling stage of data science.
Data science can work with manual methods as well though they are not as efficient as machine algorithms	Machine learning cannot exist without data science as data has to be first prepared to create, train and test the model.
Data Science as a broader term not only focuses on algorithms statistics but also takes care of the data processing.	But it is only focused on algorithm statistics.

Data Science Vs. Machine Learning

Data Science	Machine Learning
It deals with understanding and finding hidden patterns or useful insights from the data, which helps to take smarter business decisions.	It is a subfield of data science that enables the machine to learn from the past data and experiences automatically.
It is used for discovering insights from the data.	It is used for making predictions and classifying the result for new data points.
It is a broad term that includes various steps to create a model for a given problem and deploy the model.	It is used in the data modeling step of the data science as a complete process.
Data scientists spent lots of time in handling the data, cleansing the data, and understanding its patterns.	ML engineers spend a lot of time for managing the complexities that occur during the implementation of algorithms and mathematical concepts behind that.
Data Science Jnaneshwar Bohara	

Data Science Vs. Machine Learning

context of wireless networks.

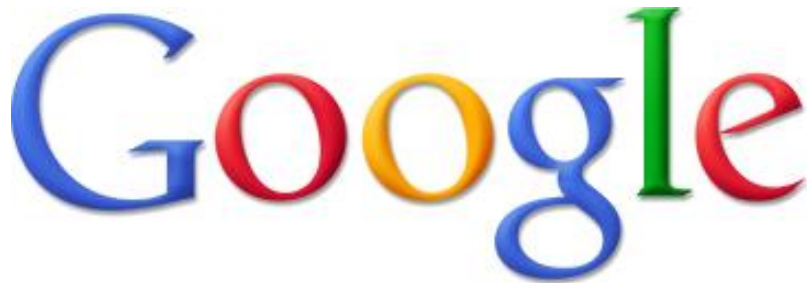
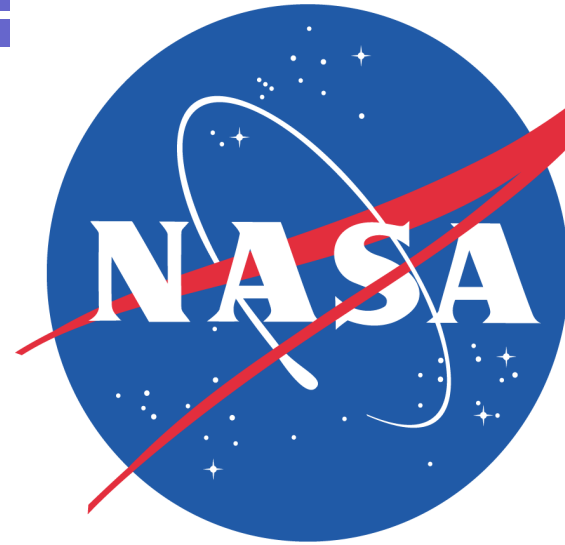


Python!

- Created in 1991 by Guido van Rossum (now at Google)
 - Named for Monty Python
- Useful as a **scripting language**
 - **script**: A small program meant for one-time use
 - Targeted towards small to medium sized projects
- Used by:
 - Google, Yahoo!, Youtube
 - Many Linux distributions
 - Games and apps (e.g. Eve Online)



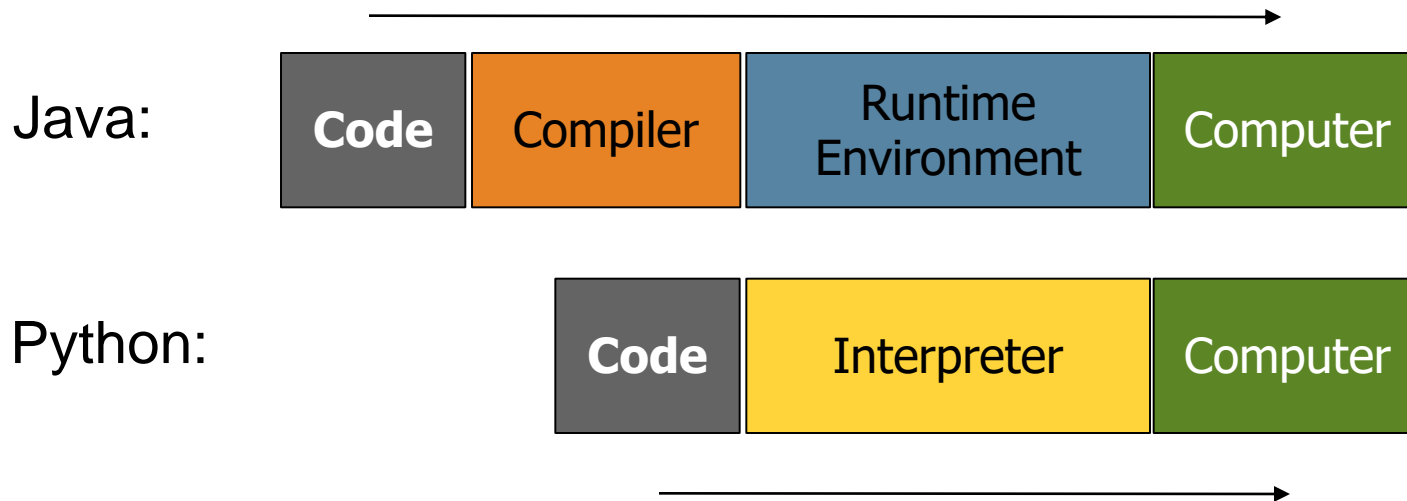
Python is used everywhere!



Interpreted Languages

- **Interpreted**

- Not compiled like Java
- Code is written and then directly executed by an **interpreter**
- Type commands into interpreter and see immediate results



Installing Python

Windows:

- Download Python from <http://www.python.org>
- Install Python.
- Run **Idle** from the Start Menu.

Mac OS X:

- Python is already installed.
- Open a terminal and run `python` or run Idle from Finder.

Linux:

- Chances are you already have Python installed. To check, run `python` from the terminal.
- If not, install from your distribution's package system.

Installing and Setting Up Python

- Although there are many Python installations available, one of the easiest way to install Python on your machine is by using a pre-packaged distribution such as Anaconda from <https://www.anaconda.com/>
- **Jupyter Notebook**
 - You can also use a Web-based user-friendly environment called Jupyter notebook to write and execute your Python program

Some Useful Libraries

- **Numpy**

- stands for numerical Python
- is a Python library package to support numerical computations
- the basic data structure in numpy is a multidimensional array object called ndarray
- Numpy provides a suite of functions that can efficiently manipulate elements of the ndarray.

Some Useful Libraries

● Pandas

- Pandas is a Python library used for working with data sets
- It has functions for analyzing, cleaning, exploring, and manipulating data
- The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis"
- Built on top of Numpy

Some Useful Libraries

- **Matplotlib**

- is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy
- offers a viable open source alternative to MATLAB
- is mostly written in python, a few segments are written in C, Objective-C and Javascript for Platform compatibility

Thank You !