Introduction to Data Science

Kamal Al Nasr, Matthew Hayes and Jean-Claude Pedjeu

Computer Science and Mathematical Sciences
College of Engineering
Tennessee State University

Outline

- Data, Big Data and Challenges
- Data Science
  - Introduction
  - Why Data Science
- Data Scientists
  - What do they do?
- Major/Concentration in Data Science
  - What courses to take.

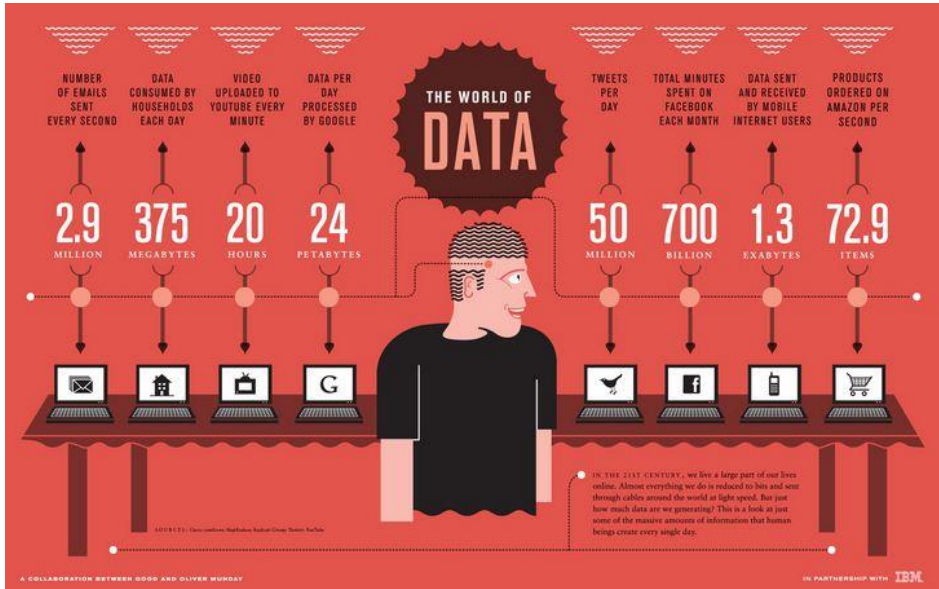
Data All Around

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Financial transactions, bank/credit transactions
  - Online trading and purchasing
  - Social Network

How Much Data Do We have?

- Google processes 20 PB a day (2008)
- Facebook has 60 TB of daily logs
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- 1000 genomes project: 200 TB

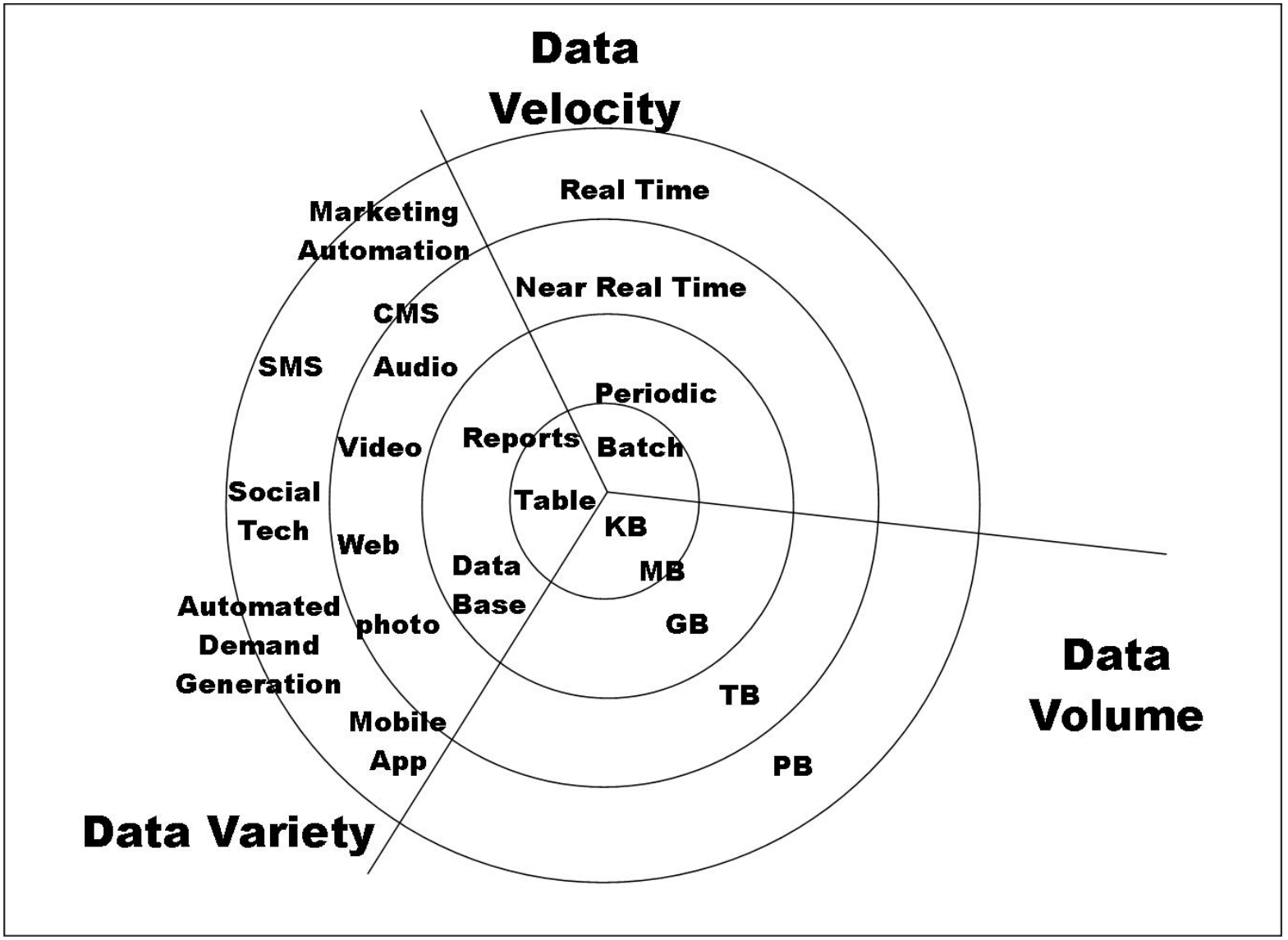
- Cost of 1 TB of disk: $35
- Time to read 1 TB disk: 3 hrs (100 MB/s)



Big Data

- Big Data is any data that is expensive to manage and hard to extract value from
  - Volume
    - The size of the data
  - Velocity
    - The latency of data processing relative to the growing demand for interactivity
  - Variety and Complexity
    - the diversity of sources, formats, quality, structures.

Big Data


Data
Velocity
Real Time
Near Real Time
Periodic
Batch
Marketing
Automation
CMS
SMS
Audio
Video
Reports
Table
Social
Tech
Web
KB
MB
GB
TB
PB
Data
Base
Automated
Demand
Generation
photo
Mobile
App
Data Variety
Data
Volume

Types of Data We Have

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data
- You can afford to scan the data once

What To Do With These Data?

- Aggregation and Statistics
  - Data warehousing and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling

Big Data and Data Science

- "… the most demanding job in the next 10 years will be statisticians," Hal Varian, Google Chief Economist
- The U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018. McKinsey Global Institute's June 2011
- New Data Science institutes being created or repurposed – NYU, Columbia, Washington, UCB,...
- New degree programs, courses, boot-camps:
  - e.g., at Berkeley: Stats, I-School, CS, Astronomy…
  - One proposal (elsewhere) for an MS in "Big Data Science"

What is Data Science?

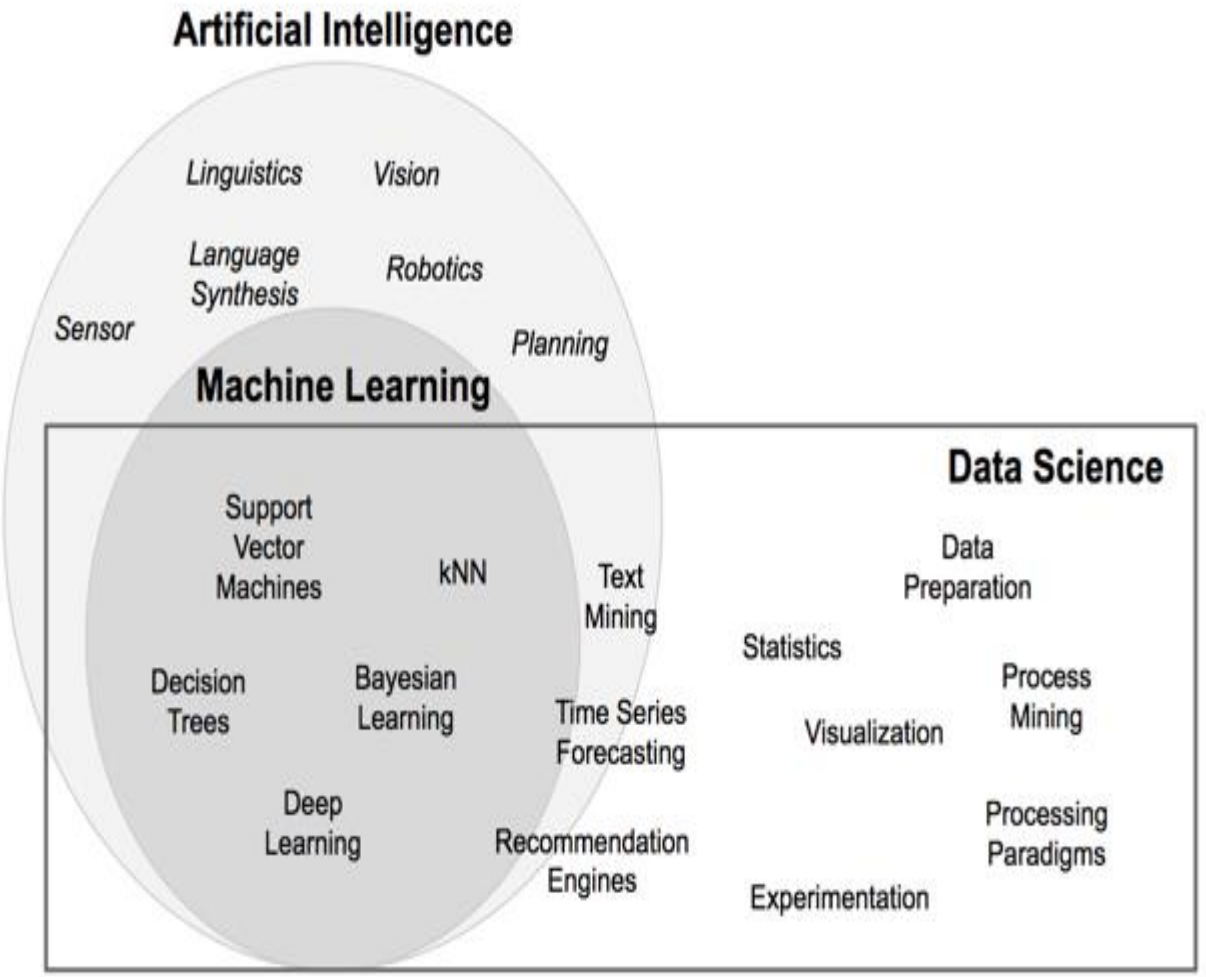
- An area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data
- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data
- Data science principles apply to all data – big and small

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

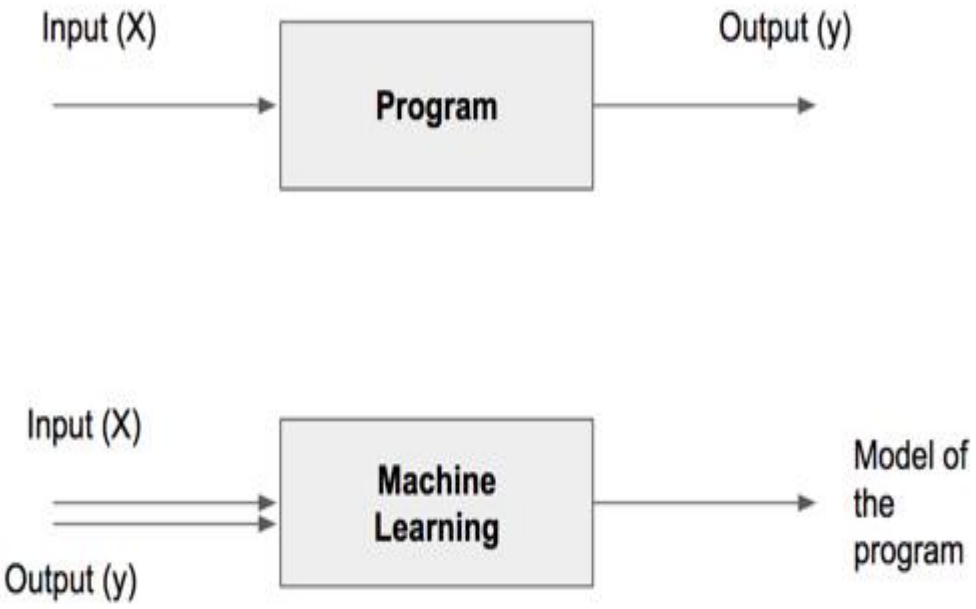
What is Data Science?

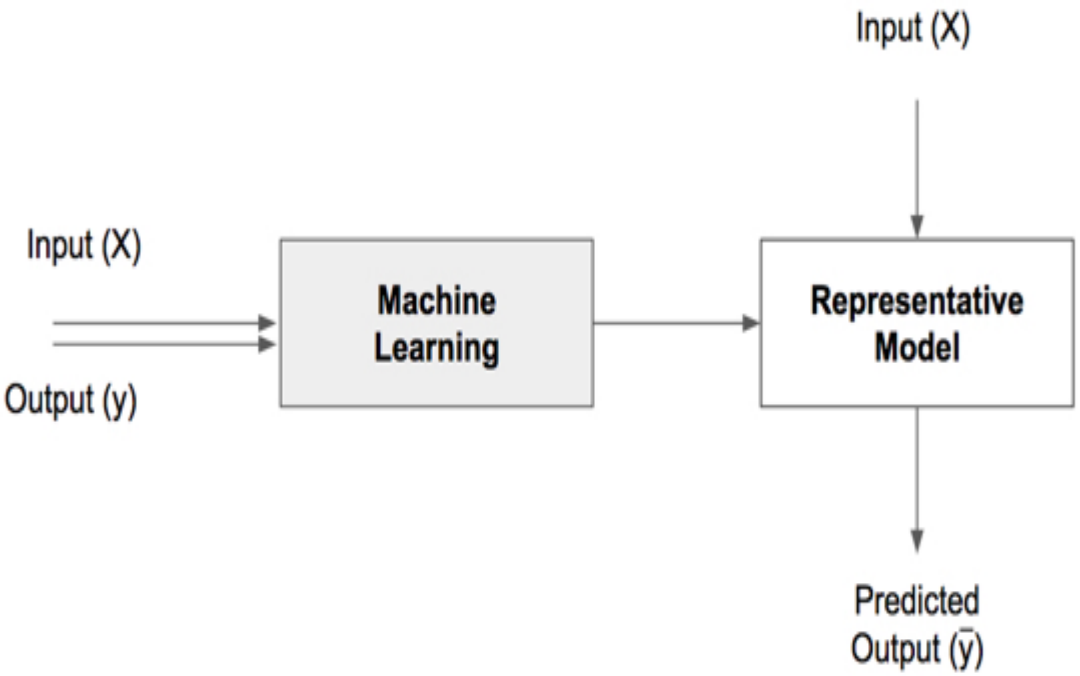
- Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education
  - Computer Science
    - Pattern recognition, visualization, data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

What is Data Science

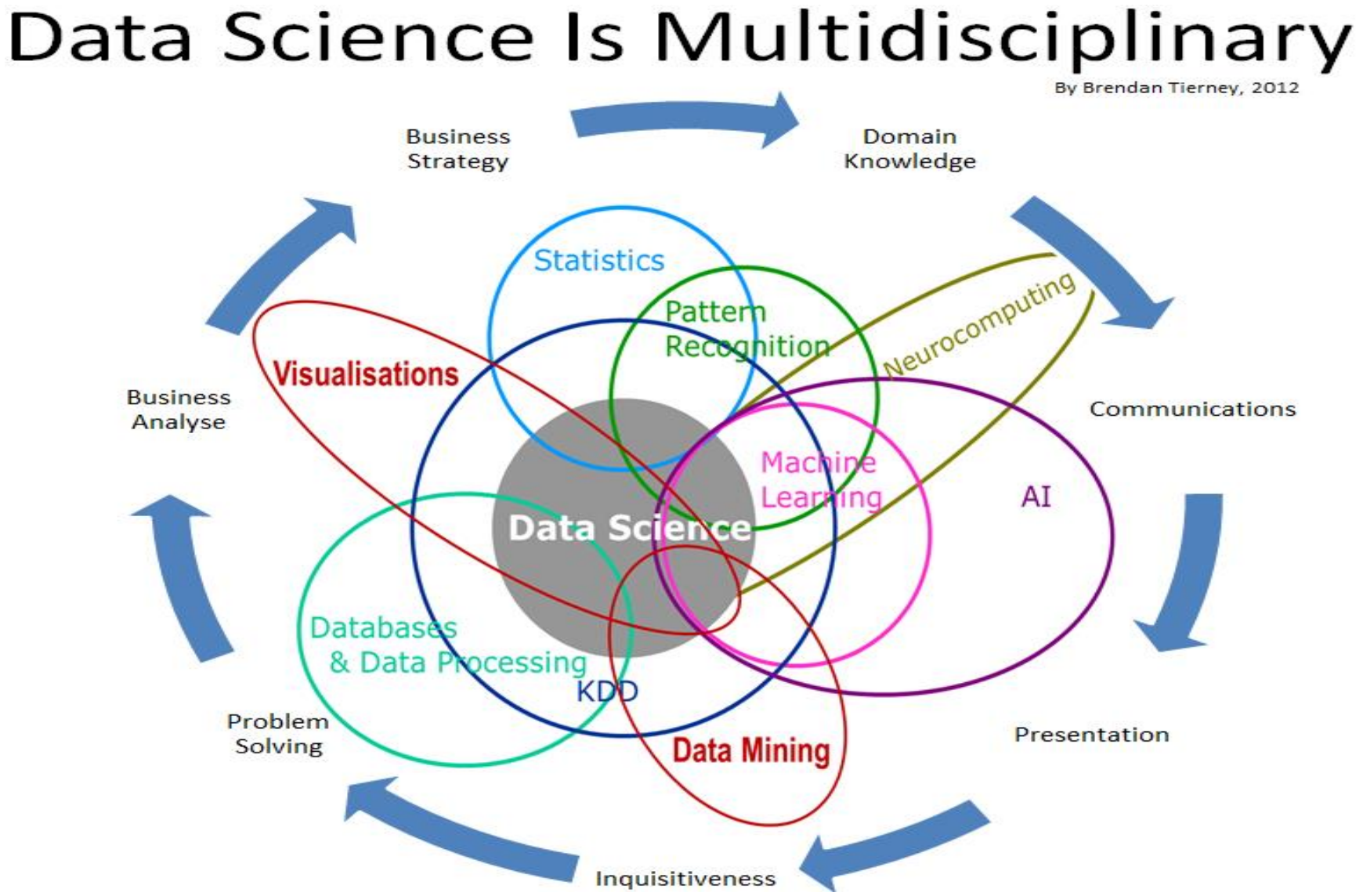


Models

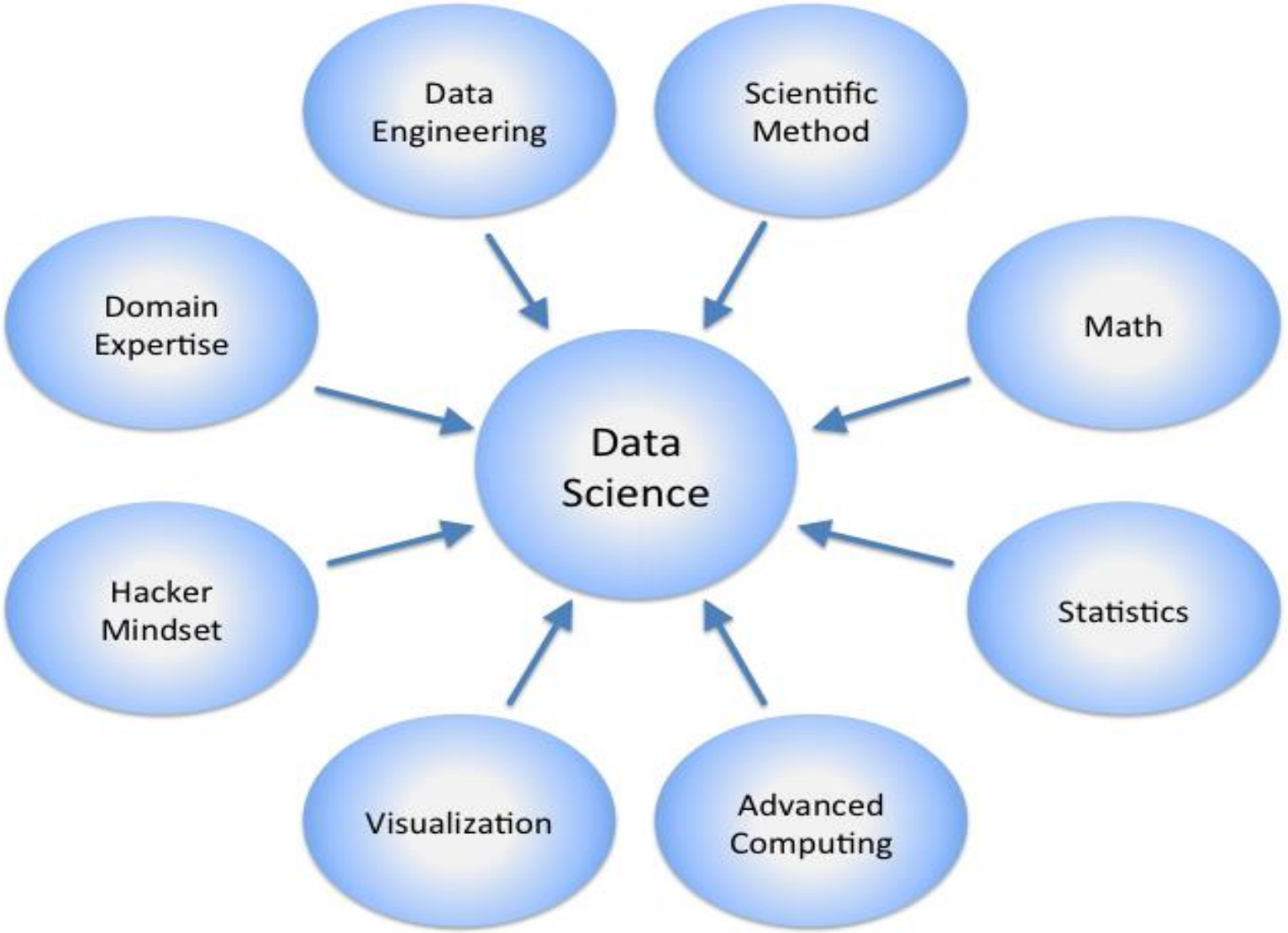

Input (X)
Program
Output (y)
Input (X)
Output (y)
Machine Learning
Model of the program


Input (X)
Input (X)
Machine Learning
Representative Model
Output (y)
Predicted Output (ȳ)

Data Science

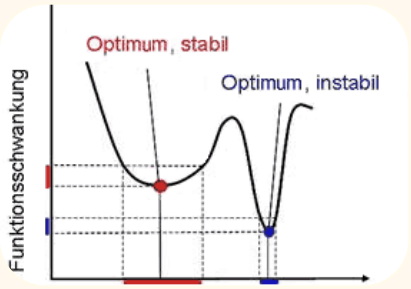


Data Science



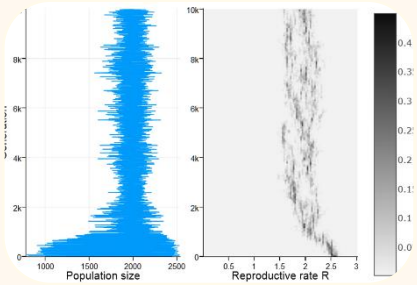
Mathematical Aspects

Computational Geometry



Optimization

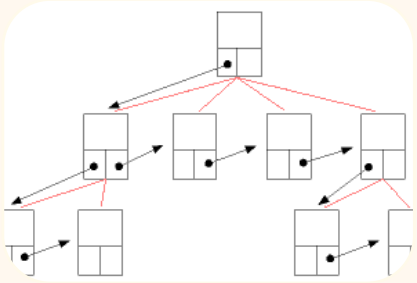
Stochastics



Scientific Computing

Machine Learning

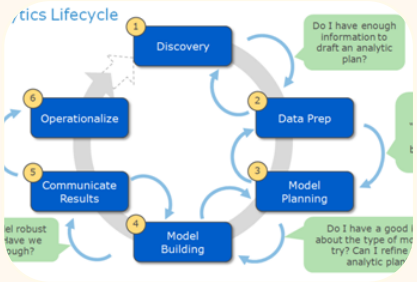
Computer Science Aspects



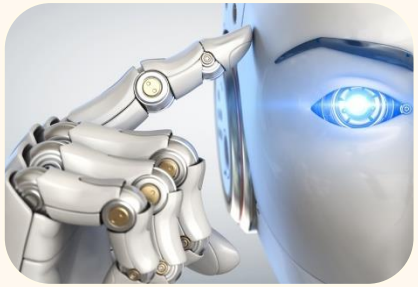
Data Structures and Algorithms

Databases

Distributed Computing



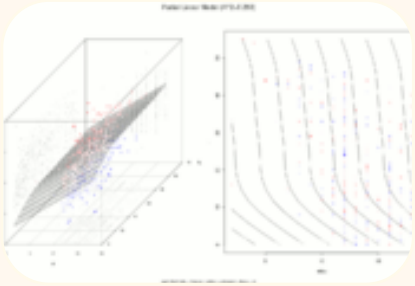
Software Engineering



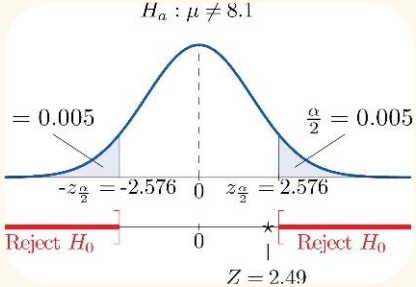
Artificial Intelligence

Machine Learning

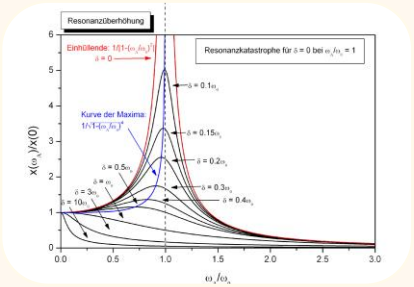
Statistical Aspects



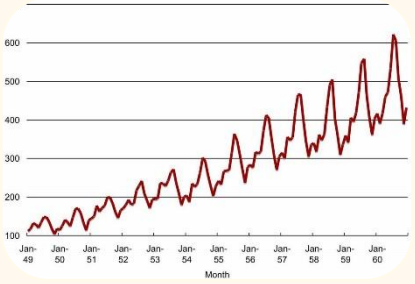
Linear Models



Statistical Tests



Inference



Time Series Analysis

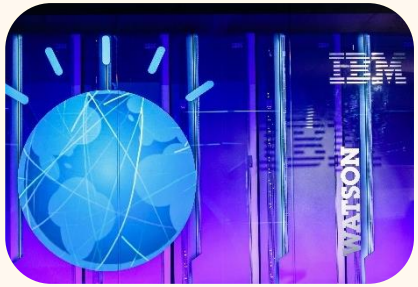
Machine Learning

Applications

Intelligent Systems

Robotics

Marketing

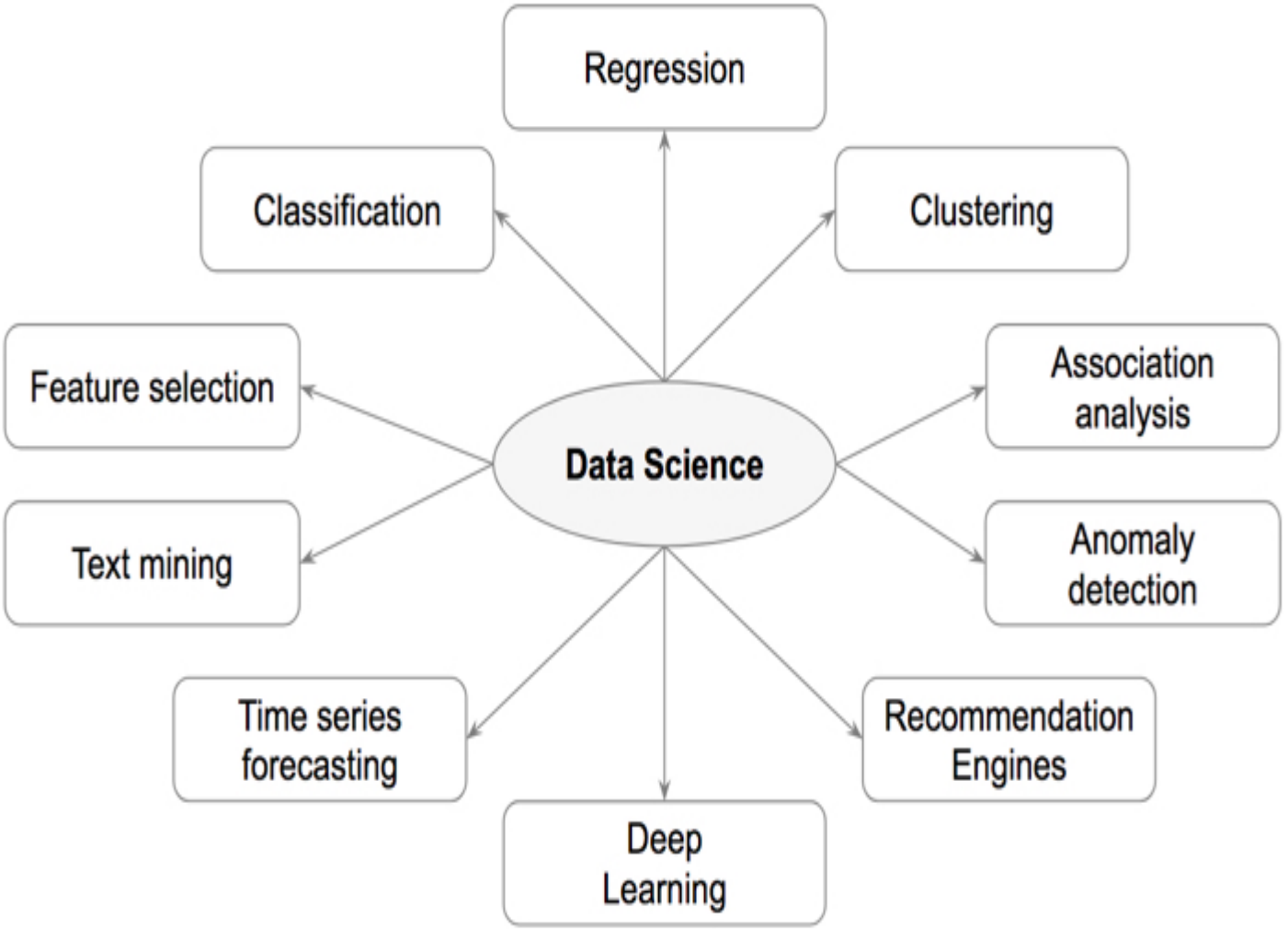


Medicine

Autonomous Driving

Social Networks

Types of Data Science



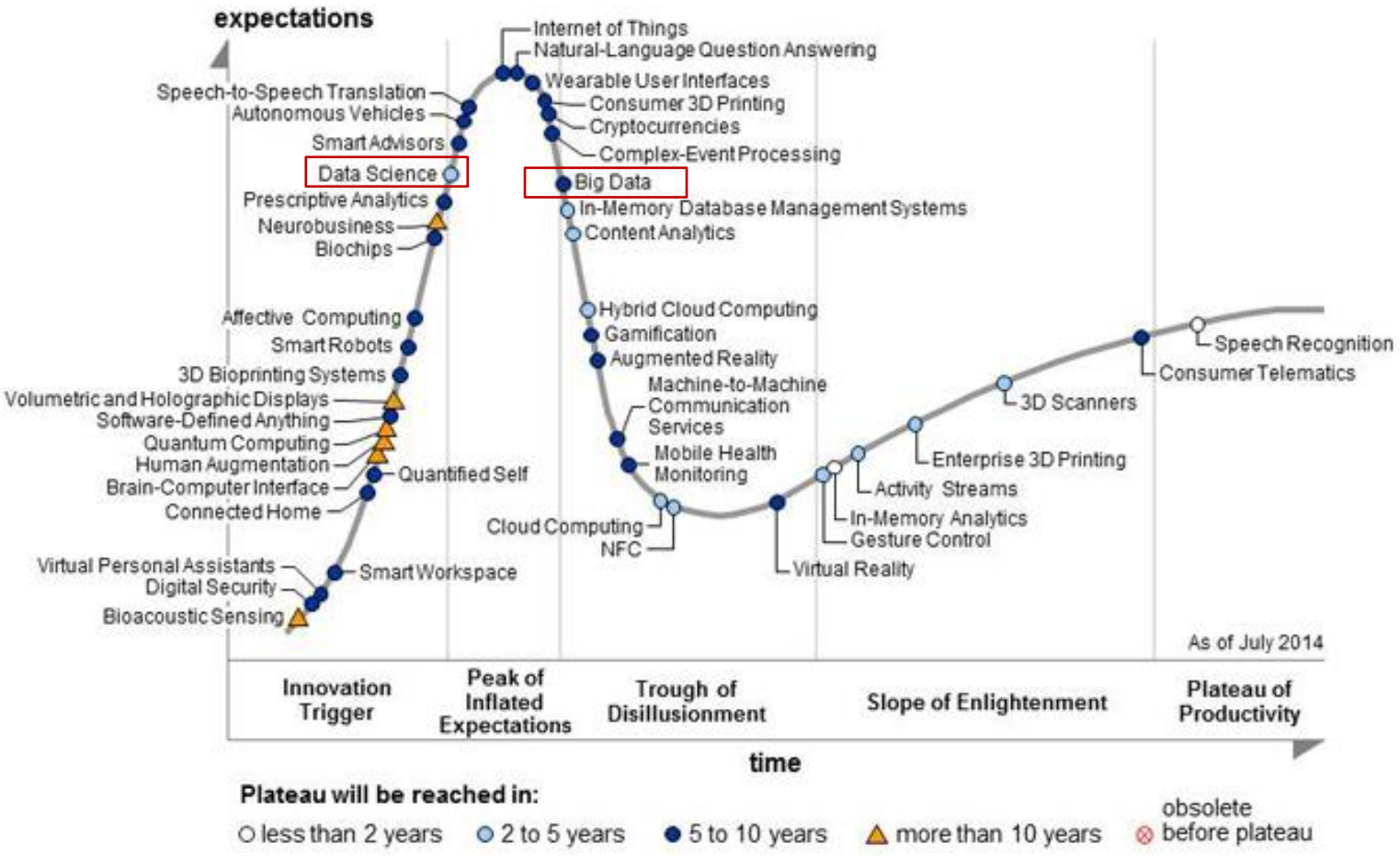
Tasks	Description	Algorithms	Examples
Classification	Predict if a data point belongs to one of predefined classes. The prediction will be based on learning from known data set.	Decision Trees, Neural networks, Bayesian models, Induction rules, K nearest neighbors	Assigning voters into known buckets by political partie. Bucketing new customers into one of known customer groups.
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from known data set.	Linear regression, Logistic regression	Predicting unemployment rate for next year. Estimating insurance premium.
Anomaly detection	Predict if a data point is an outlier compared to other data points in the data set.	Distance based, Density based, LOF	Fraud transaction detection in credit cards. Network intrusion detection.
Time series	Predict if the value of the target variable for future time frame based on history values.	Exponential smoothing, ARIMA, regression	Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated
Clustering	Identify natural clusters within the data set based on inherit properties within the data set.	K means, density based clustering - DBSCAN	Finding customer segments in a company based on transaction, web and customer call data.
Association analysis	Identify relationships within an itemset based on transaction data.	FP Growth, Apriori	Find cross selling opportunities for a retailor based on transaction purchase history.

Real Life Examples

- Companies learn your secrets, shopping patterns, and preferences
  - For example, can we know if a woman is pregnant, even if she doesn't want us to know? Target case study
- Data Science and election (2008, 2012)
  - 1 million people installed the Obama Facebook app that gave access to info on "friends"

Why is it so demanding?

- Gartner's 2014 Hype Cycle



Data Scientists

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

Data Scientists

- Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions

What do Data Scientists do?

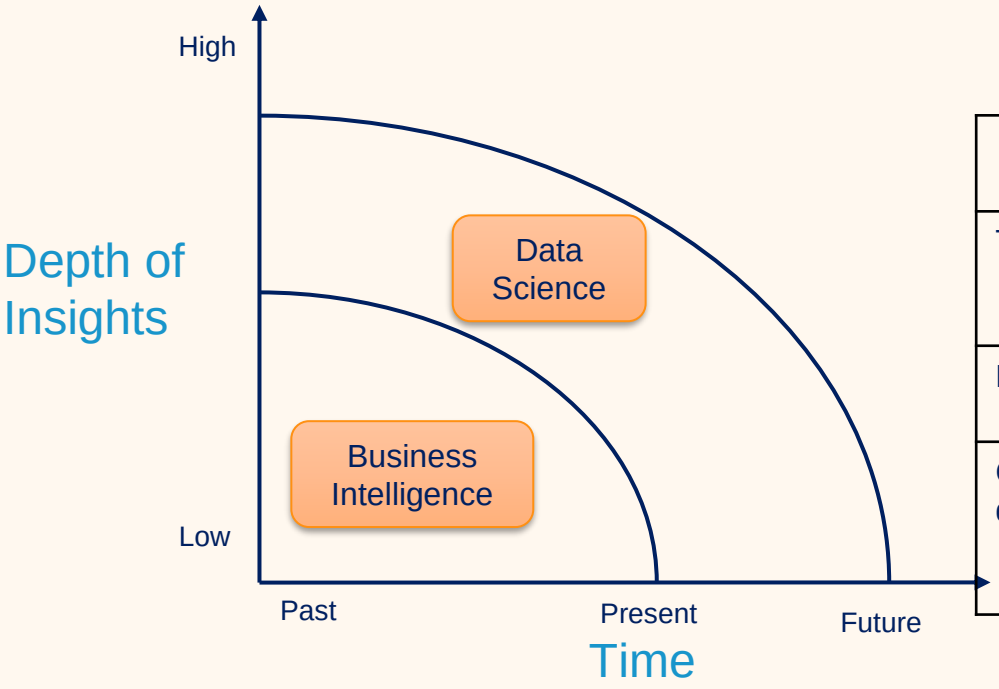
- National Security
- Cyber Security
- Business Analytics
- Engineering
- Healthcare
- And more ….

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 Science vs. Business Intelligence

- Business Intelligence (Gartner IT Glossary)
  - […] best practices that enable access to and analysis of information to improve and optimize decisions and performance.



	Business Intelligence	Data Science
Techniques	Dashboards, alerts, queries	Optimization, predictive modelling, forecasting
Data Types	Structured, data warehouses	Any kind, often unstructured
Common questions	What happened…? How much did…? When did…?	What if…? What will…? How can we…?

More Data More Opportunities

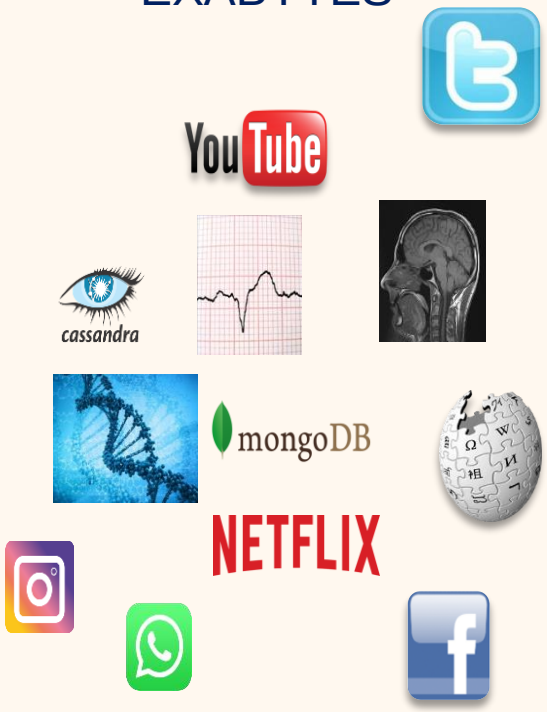
TERABYTES

PETABYTES

EXABYTES

1990's
Relational Databases
& Data Warehouses

2000's
Content Management

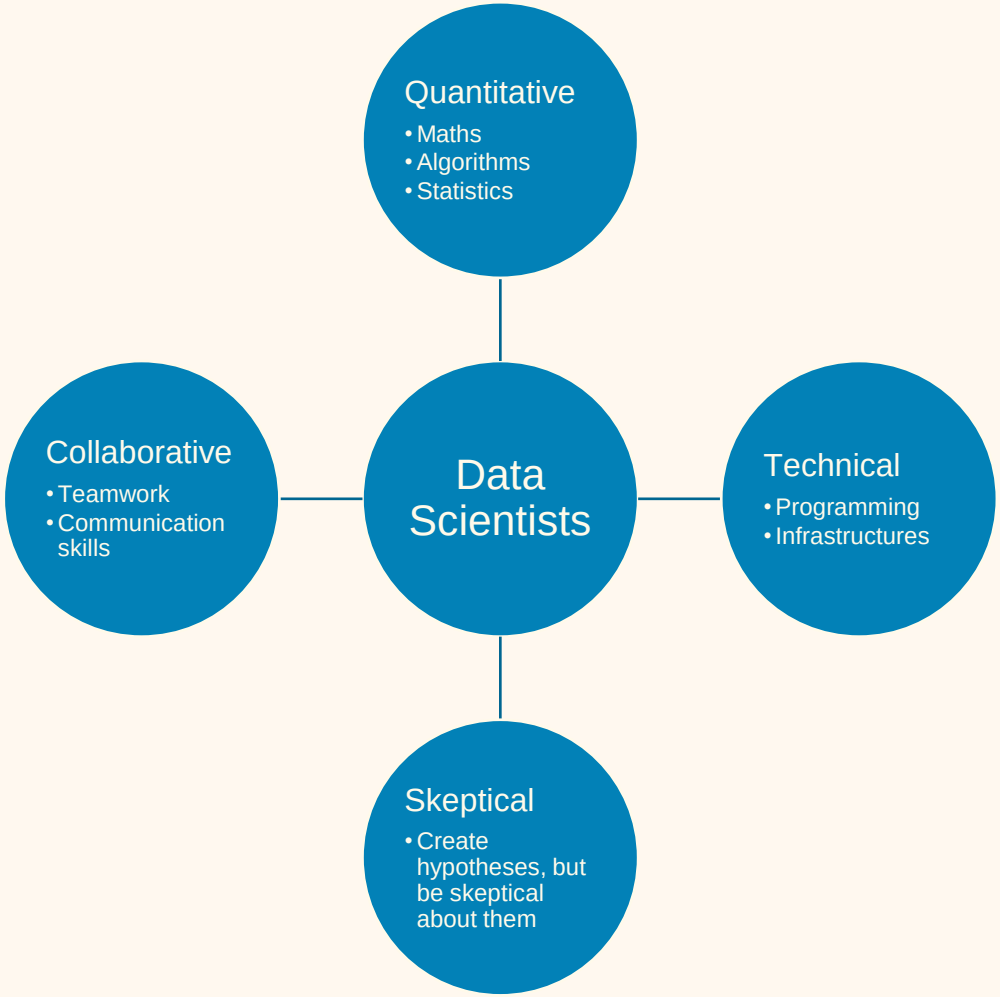


2010's
Key-Value Storages
& Unstructured Data

What are Data Scientists?

- Not computer scientists
  - But should know about databases, data structures, algorithms, etc.
- Not mathematicians
  - But should know about optimization, stochastics, etc.
- Not statisticians
  - But should know about regression, statistical tests, etc.
- Not domain experts
  - But must work together with them

Skills of Data Scientists



A bit of everything

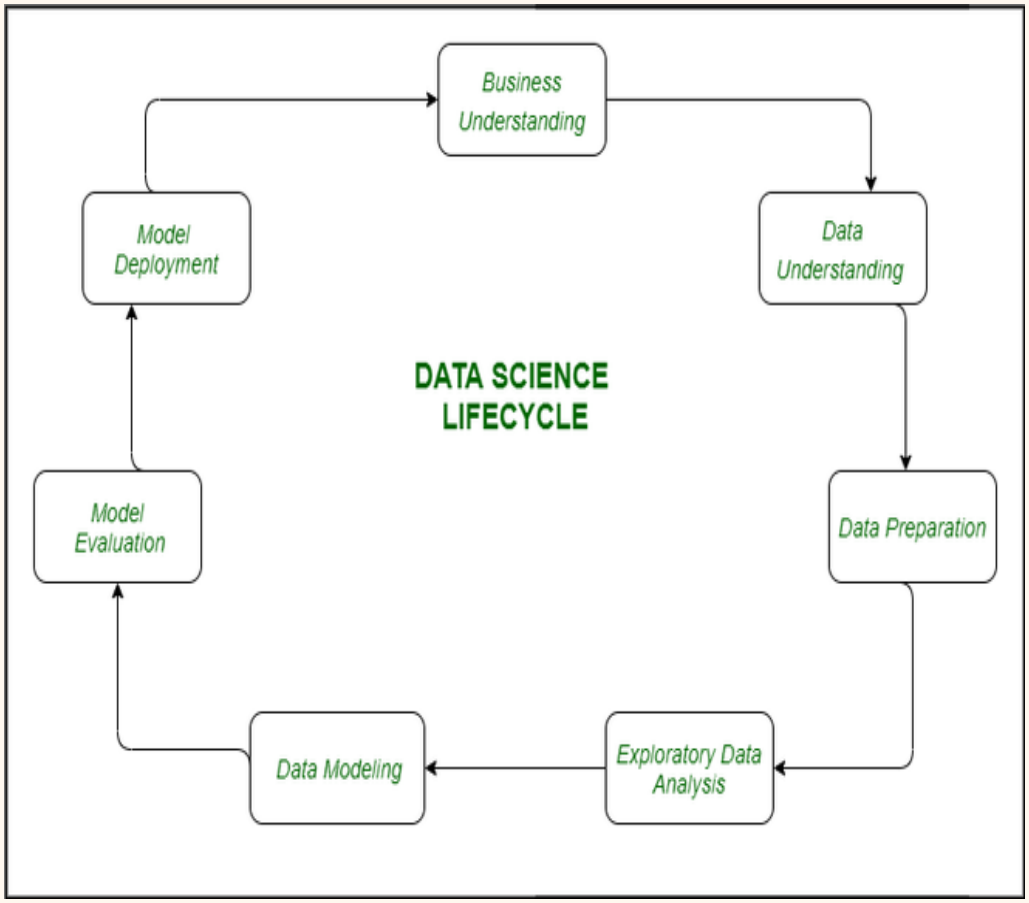
… but actually as much as possible of everything

Summary

- Big data has a high volume, velocity, and variety
- Different data structures
  - Structured, semi-structured, quasi-structured, unstructured
- Data science is a very diverse discipline
  - Maths, computer science, statistics, applications

 Data scientists require a diverse skillset

Data science life cycle



Data science life cycle

1. Business Understanding: Define objectives and requirements from a business perspective.
2. Data Understanding: Collect initial data and assess its quality and characteristics.
3. Data Preparation: Clean, normalize, and transform data for analysis.
4. Exploratory Data Analysis: Analyze data to discover patterns and insights.
5. Data Modeling: Develop models to predict or classify data based on identified patterns.
6. Model Evaluation: Assess the model’s performance and validate its accuracy.
7. Model Deployment: Implement the model in a real-world environment for practical use.

KAGGLE DATASETS

https://www.kaggle.com/datasets?fileType=csv