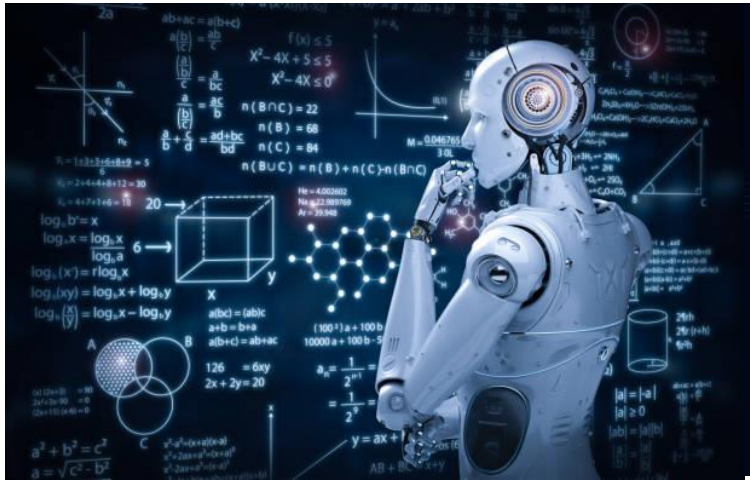
Welcome…

Introduction to Data Science

CS 797Q
Fall 2024

Aug 21,2024



Outline

- Data, Big Data and Challenges
- Data Science
  - Introduction
  - Why Data Science
- Data Scientists
  - What do they do?
- Data Science Process
- Cross Industry Standard Process for Data Mining
- Data Science Application
- Data science Challenges

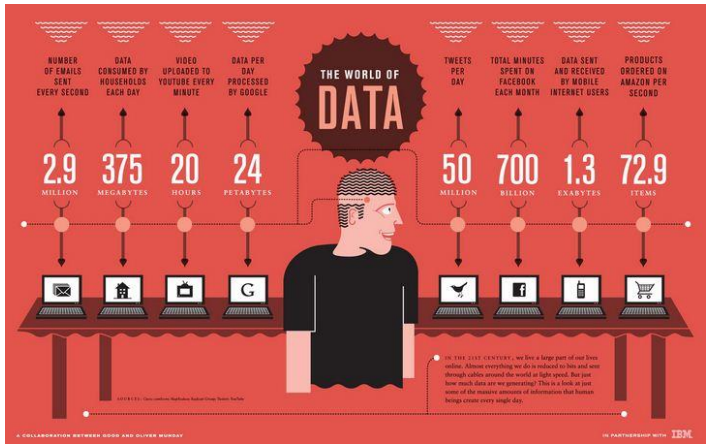
Data All Around

- 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?

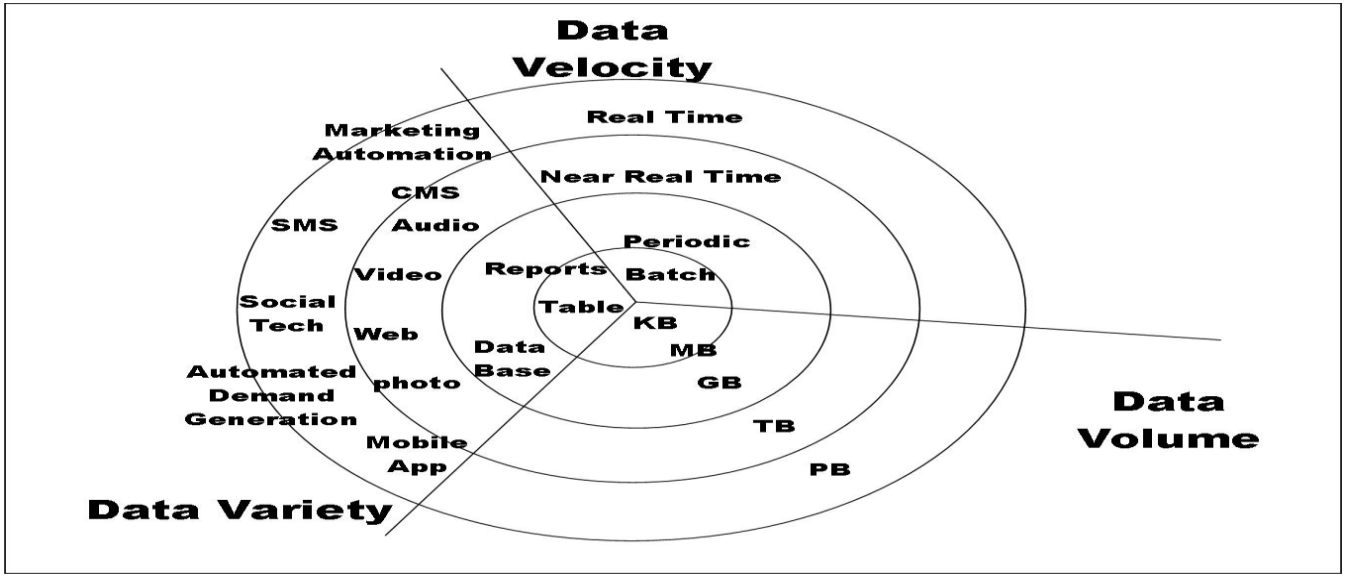
- Google processes 20 PB a day (2008)
- Facebook has 60 TB of daily logs
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- 1000 genomes project: 200 TB

- 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.



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

What is Data Science?

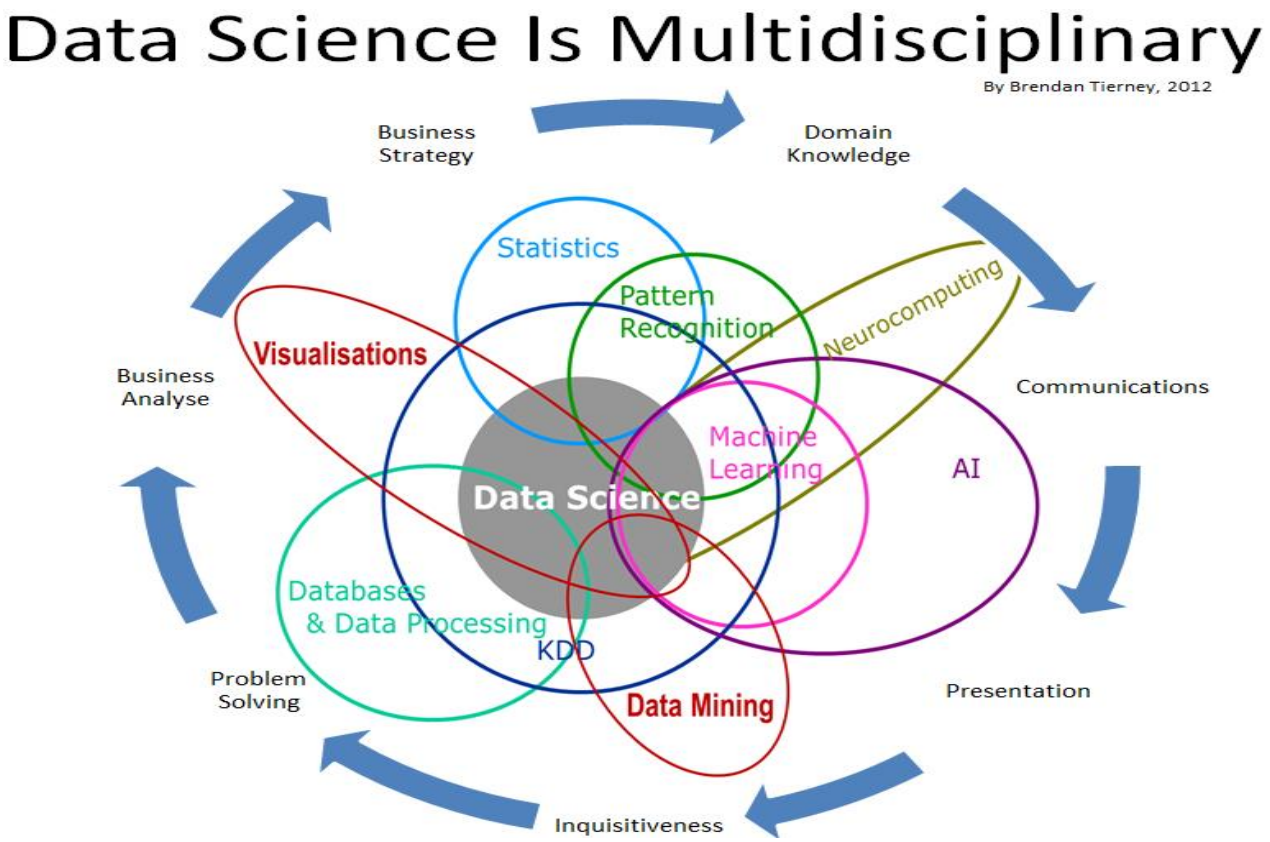
- An area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data
- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data
- Data science principles apply to all data – big and small

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

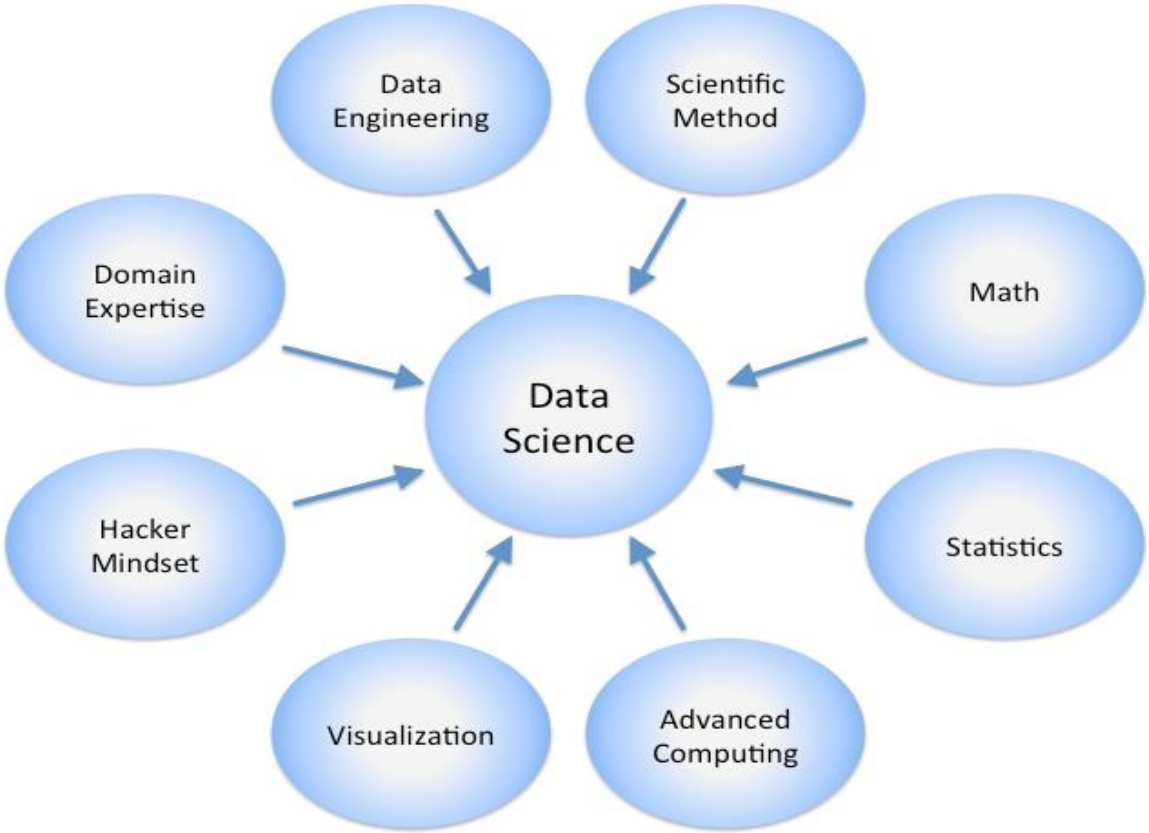
What is Data Science?

- 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.

Data Science



Data Science



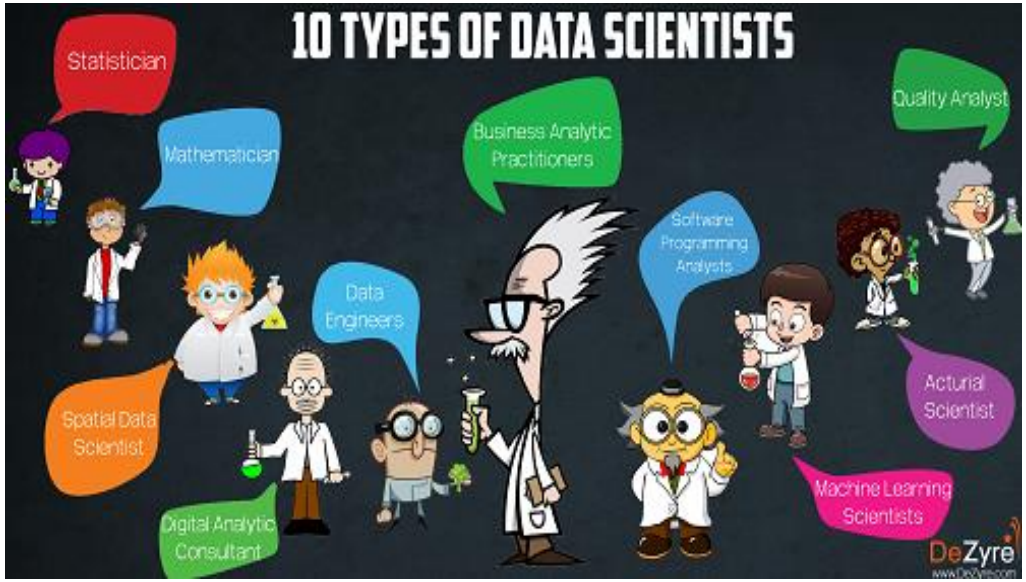
Data Scientists

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

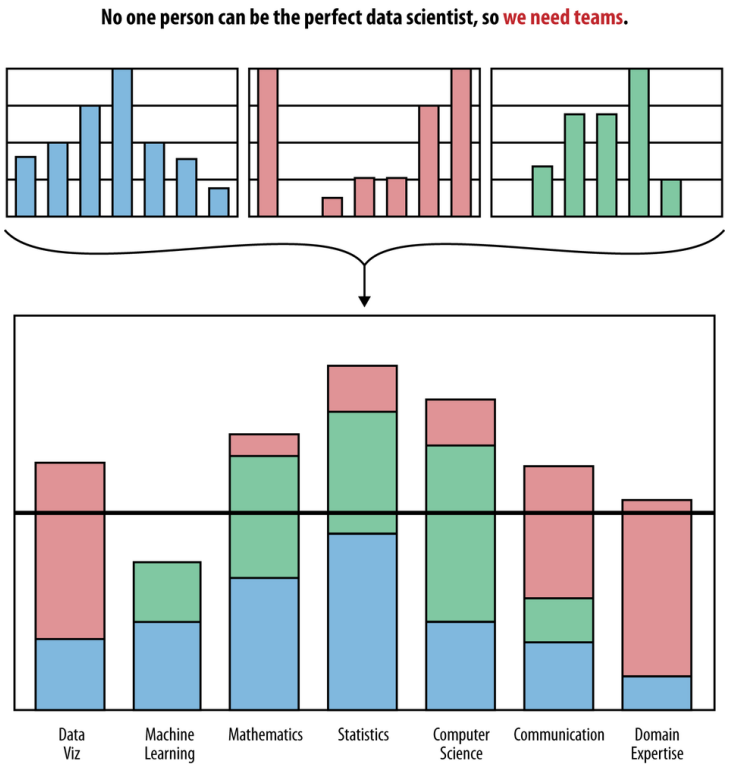
Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions.

Types of Data Scientists

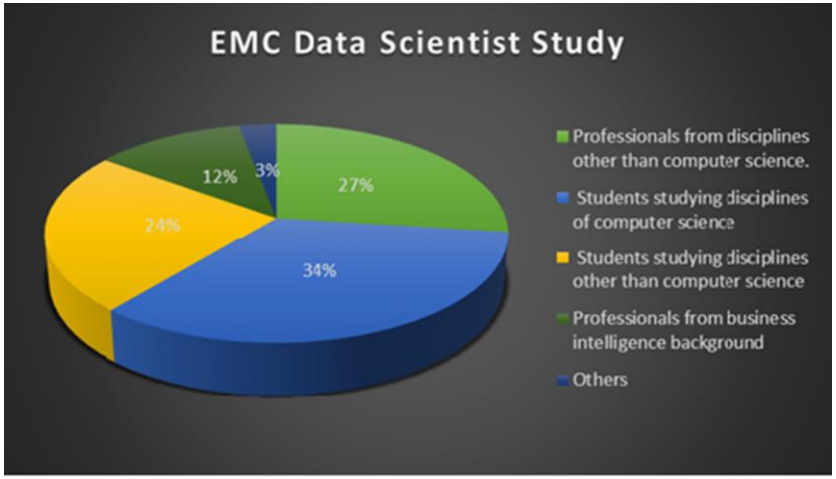
- Machine Learning Scientist
- Statistician
- Software Programming Analyst
- Data Engineer
- Actuarial Scientist
- Business Analytic Practitioner
- Quality Analyst
- Spatial Data Scientist
- Mathematician
- Digital Analytic Consultant



Data Science team

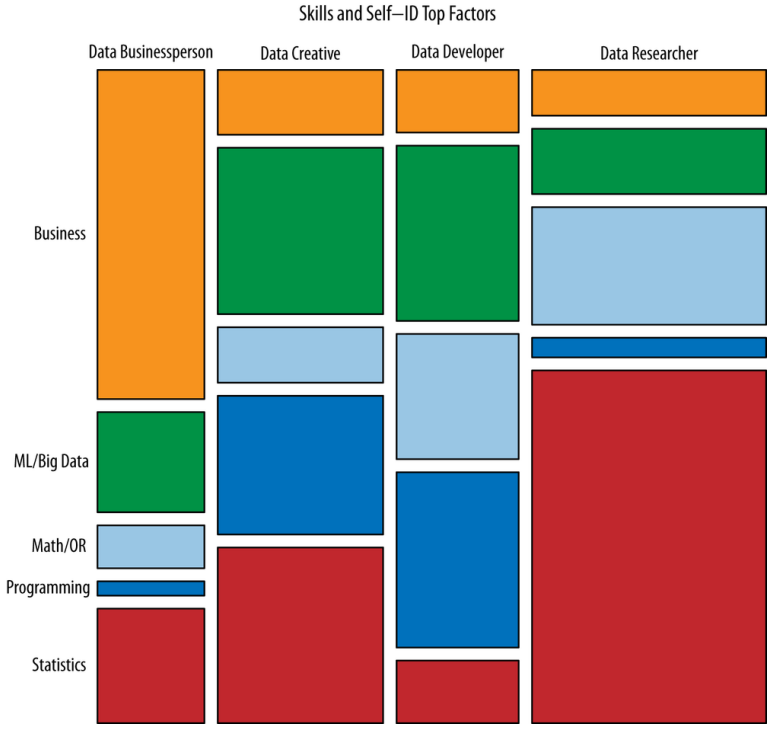


- individual data scientist profiles are merged to make a Data science team
- team profile should align with the profile of the data problems to tackle



Data science: skills and actors

Clustering and visualization of data science subfields based on a survey of data science practitioners (Analyzing the Analyzers by Harlan Harris, Sean Murphy, and Marck Vaisman, 2012)



- Data Businesspeople are the product and profit-focused data scientists. They're leaders, managers, and entrepreneurs, but with a technical bent. A common educational path is an engineering degree paired with an MBA.
- Data Creatives are eclectic jacks-of-all-trades, able to work with a broad range of data and tools. They may think of themselves as artists or hackers, and excel at visualization and open source technologies.
- Data Developers are focused on writing software to do analytic, statistical, and machine learning tasks, often in production environments. They often have computer science degrees, and often work with so-called "big data".
- Data Researchers apply their scientific training, and the tools and techniques they learned in academia, to organizational data. They may have PhDs, and their creative applications of mathematical tools yields valuable insights and products.

What do Data Scientists do?

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

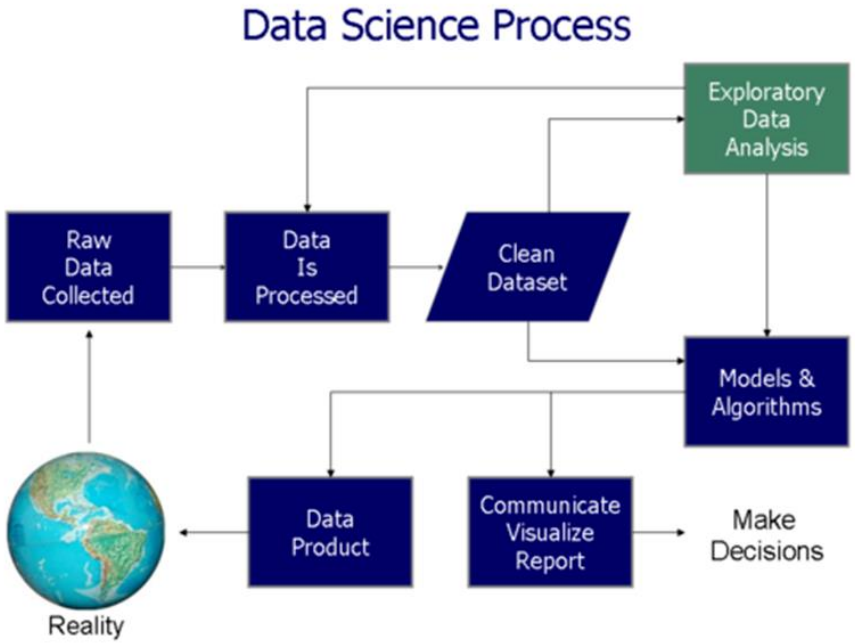
What do data scientists do?

- In academia, a data scientist is trained in some discipline, works with large amounts of data, grapples with computational problems posed by the structure, size, messiness, and the complexity and nature of the data, and solves real-world problems.

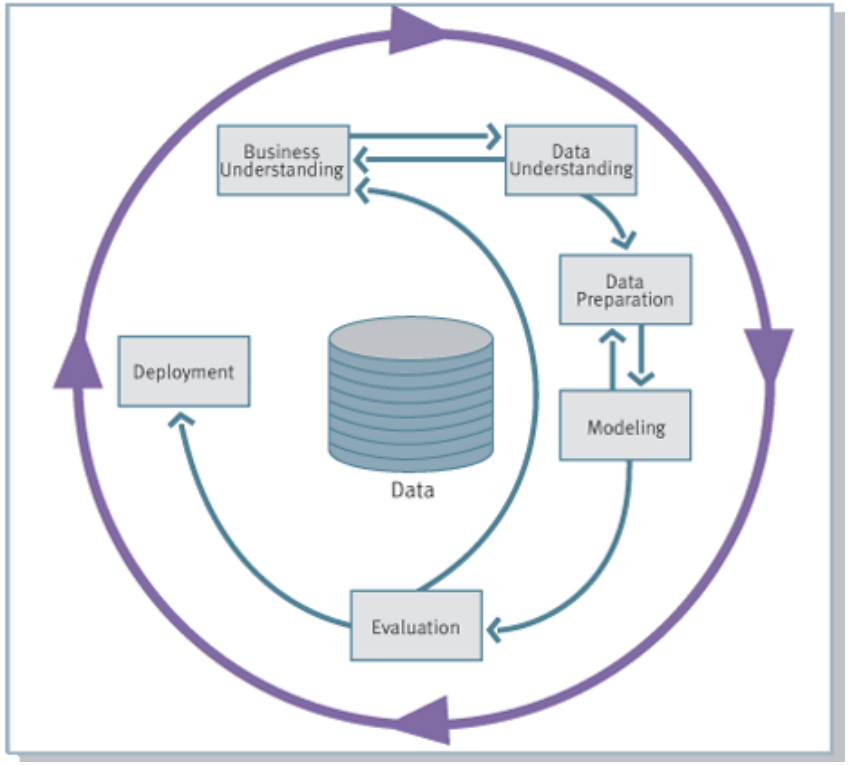
- In industry, a data scientist
  - knows how to extract meaning from and interpret data, which requires both tools and methods from statistics and machine learning, as well as being human.
  - spends a lots of effort in collecting, cleaning, and munging data utilizing statistics and software engineering skills.
  - performs exploratory data analysis, finds patterns, builds models, and algorithms.
  - communicates the findings in clear language and with data visualizations so that even if her/his colleagues unfamiliar with the data can understand the implications.

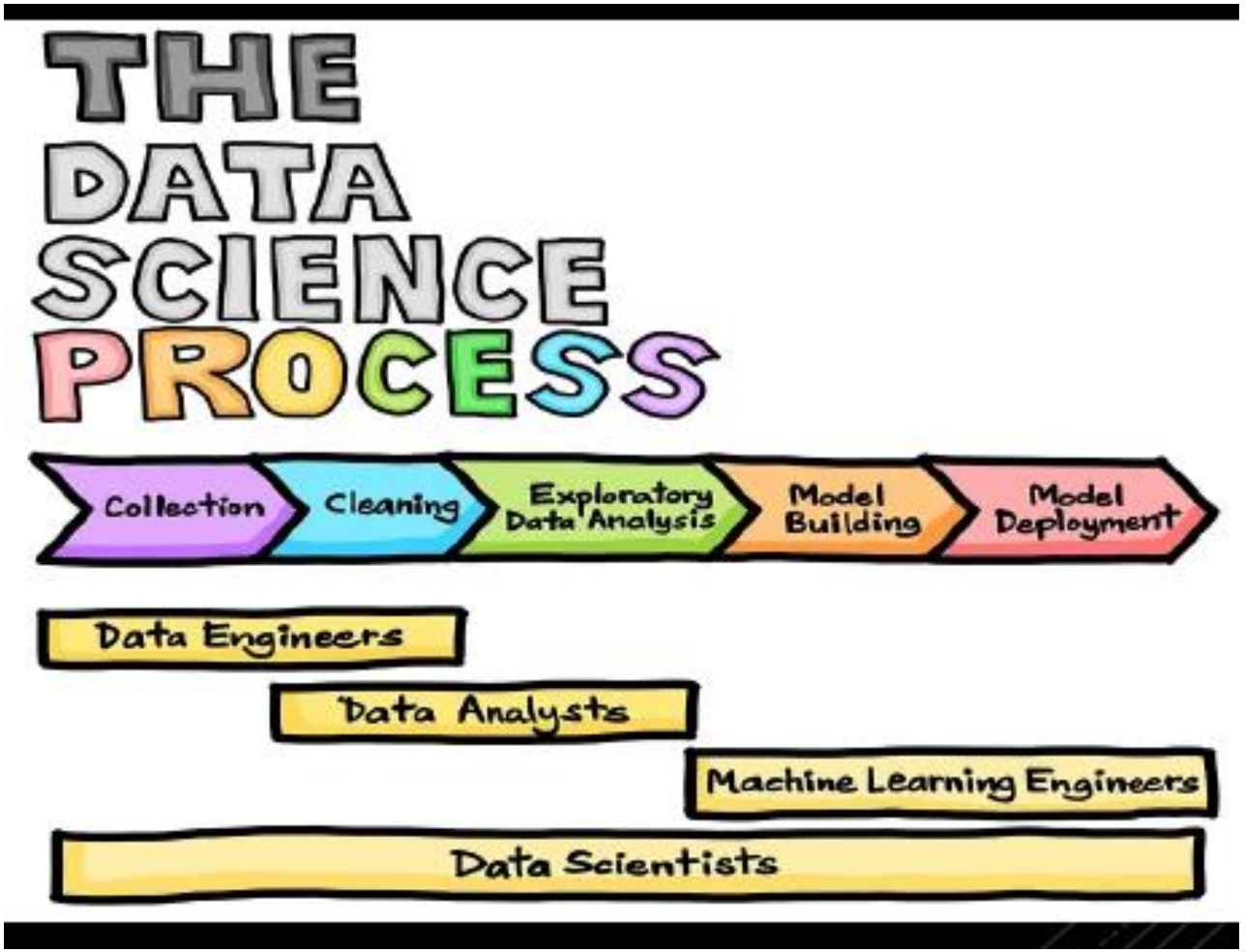
Data Science Process

Data science process flowchart (O'Neil and Schutt)



CRISP-DM (Cross Industry Standard Process for Data Mining)

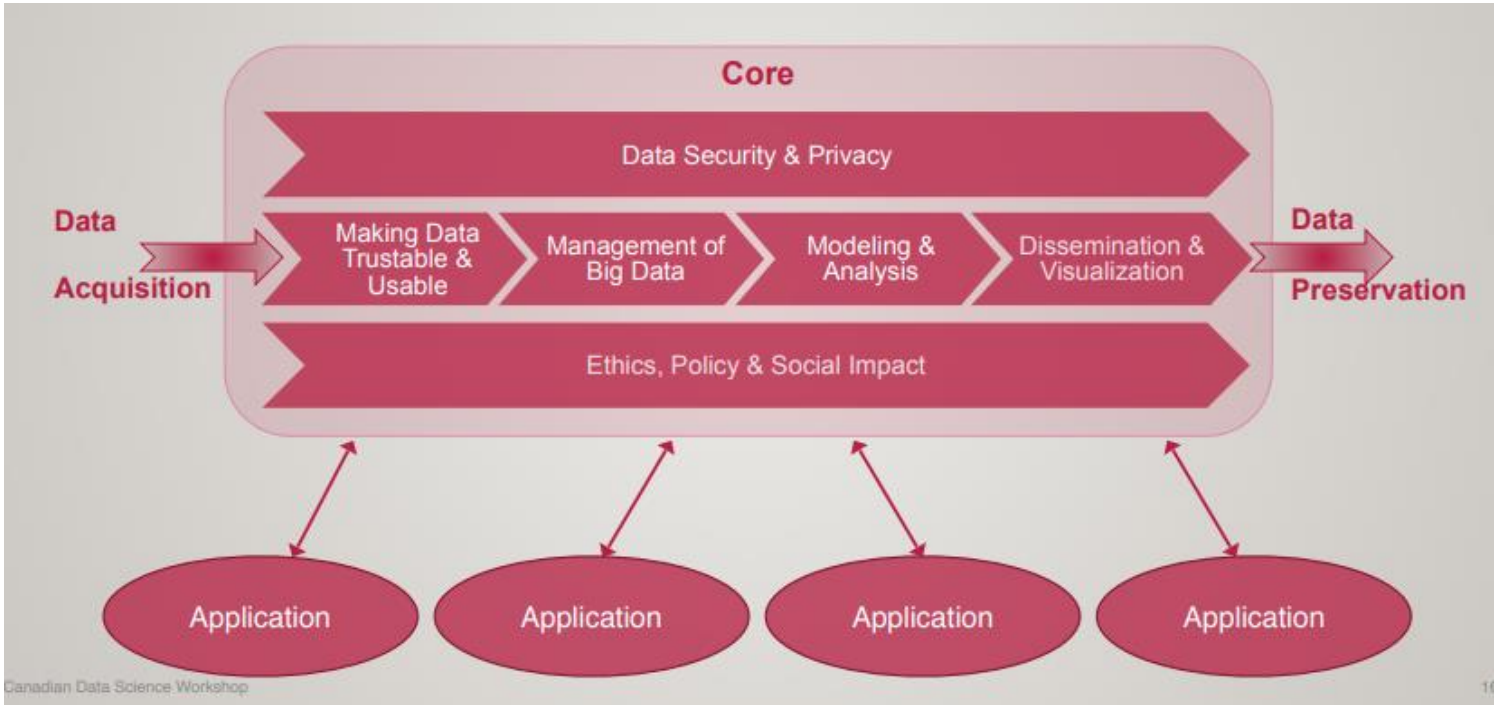



THE DATA SCIENCE PROCESS
Collection
Cleaning
Exploratory Data Analysis
Model Building
Model Deployment
Data Engineers
Data Analysts
Machine Learning Engineers
Data Scientists

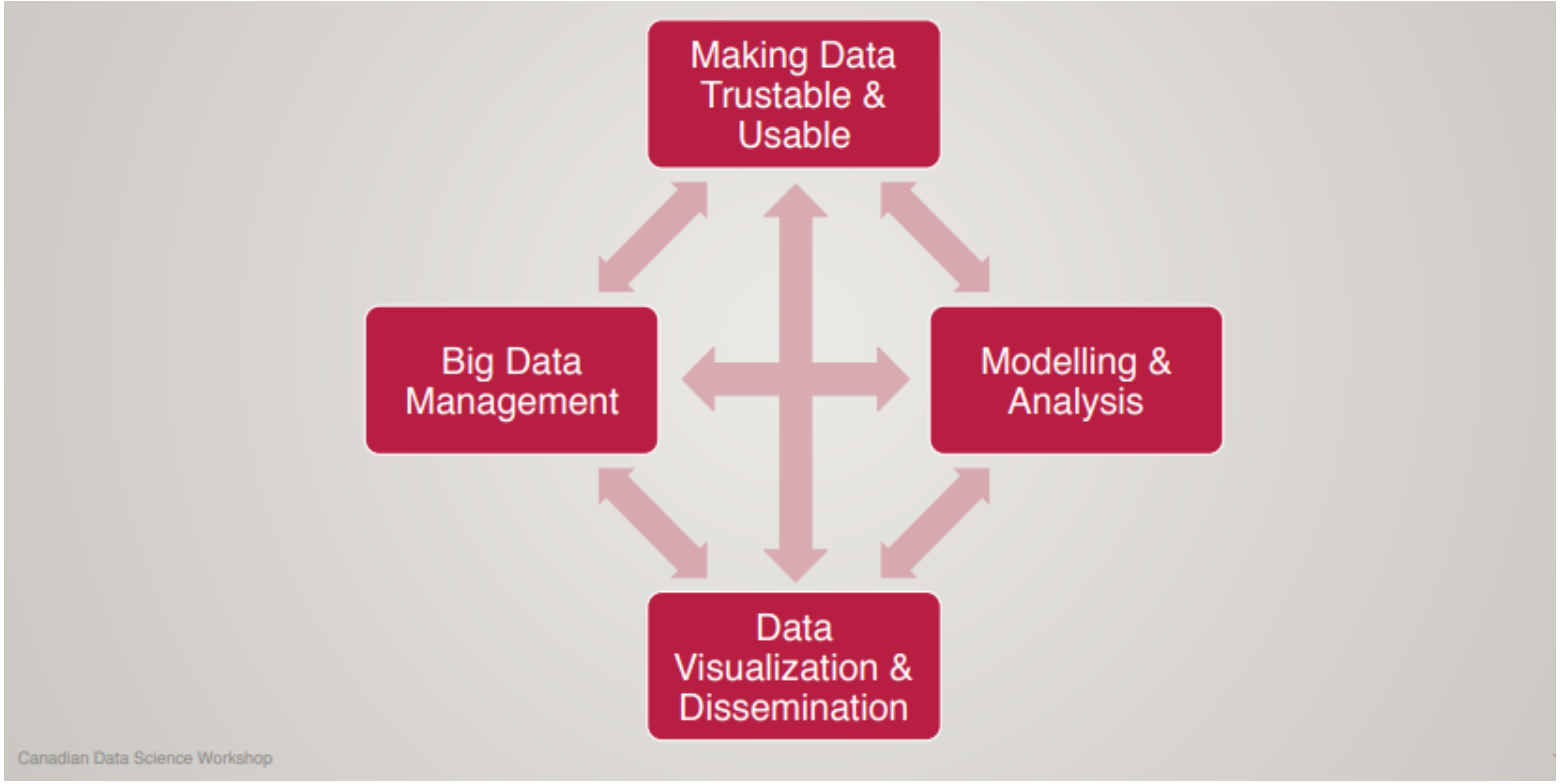
CRISP-DM Phases, tasks, *outputs*

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria*	**Collect Initial Data** *Initial Data Collection Report*	***Data Set*** *Data Set Description*	**Select Modeling Technique** *Modeling Technique* *Modeling Assumptions*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models*	**Plan Deployment** *Deployment Plan*
Situation Assessment *Inventory of Resources Requirements,Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits*	**Describe Data** *Data Description Report*	**Select Data** *Rationale for Inclusion / Exclusion*	**Generate Test Design** *Test Design*	**Review Process** *Review of Process*	**Plan Monitoring and Maintenance** *Monitoring & Maintenance Plan*
Determine Data Mining Goal *Data Mining Goals* *Data Mining Success Criteria*	**Explore Data** *Data Exploration Report*	**Clean Data** *Data Cleaning Report*	**Build Model** *Parameter Settings* *Models* *Model Description*	**Determine Next Steps** *List of Possible Actions* *Decision*	**Produce Final Report** *Final Report* *Final Presentation*
Produce Project Plan *Project Plan* *Initial Asessment of Tools and Techniques*	**Verify Data Quality** *Data Quality Report*	**Construct Data** *Derived Attributes* *Generated Records*	**Assess Model** *Model Assessment* *Revised Parameter Settings*		**Review Project** *Experience Documentation*
		Integrate Data *Merged Data*			
		Format Data *Reformatted Data*			

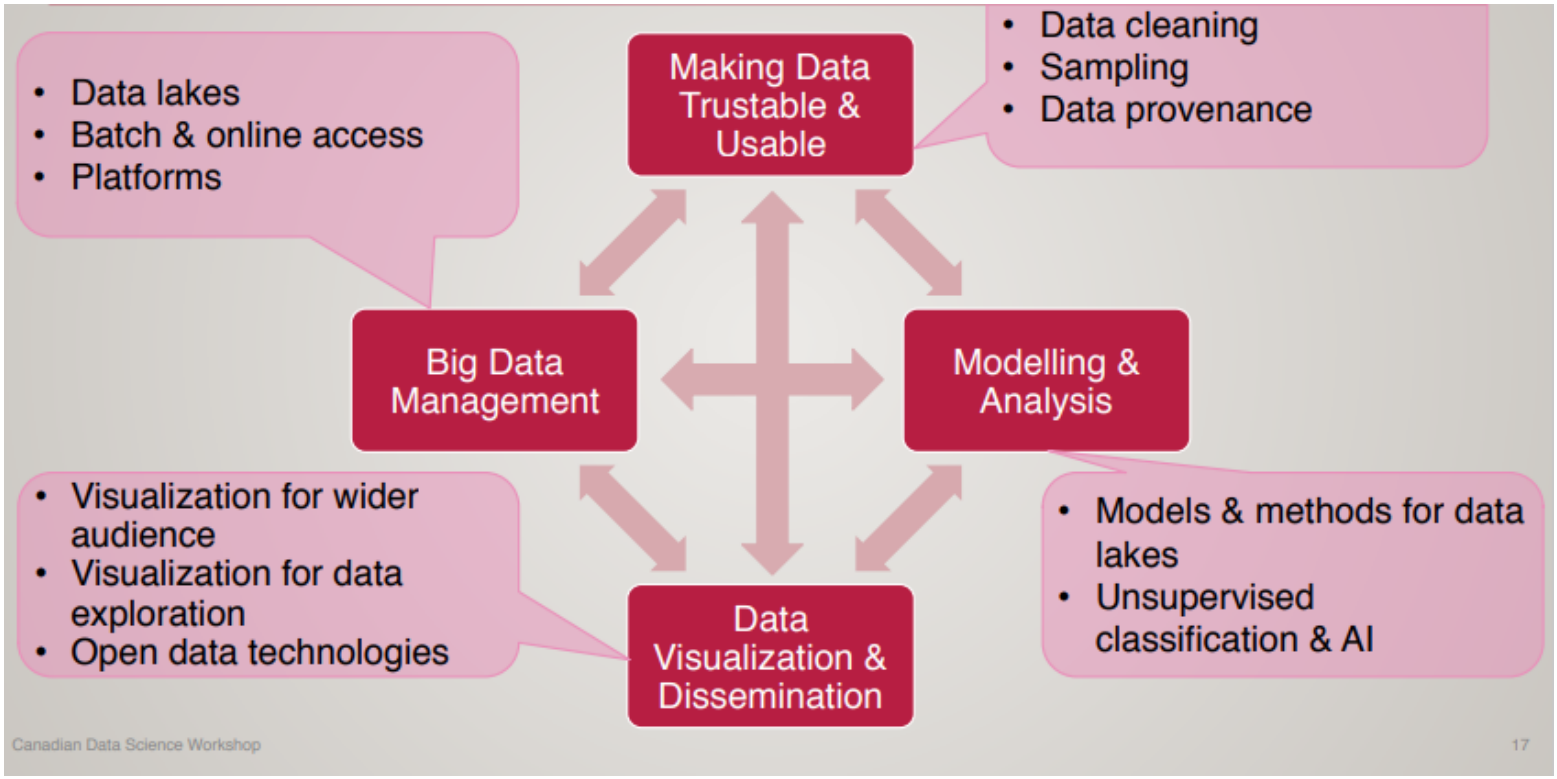
HOLISTIC APPROACH TO DATA SCIENCE



CORE RESEARCH ISSUES & INTERACTIONS



CORE RESEARCH ISSUES & INTERACTIONS



DATA SCIENCE APPLICATION EXAMPLES

- Fraud detection
  - Investigate fraud patterns in past data
  - Early detection is important
    - Before damage propagates
    - Harder than late detection
  - Precision is important
    - False positive and false negative are both bad
  - Real-time analytics

- Recommender systems
  - The ability to offer unique personalized service
  - Increase sales, click-through rates, conversions, …
    - Netflix recommender system valued at $1B per year
    - Amazon recommender system drives a 20-35% lift in sales annually
  - Collaborative filtering at scale

- Predicting why patients are being readmitted
  - Reduce costs
  - Improve population health
  - Find the "why" behind specific populations being readmitted
  - Data lakes of multiple data sources
  - Investigate ties between readmission ar socioeconomic data points, patient history, genetics, …

Challenges in data science

- Accessing data
- Cleanliness of data
- Ever changing technical landscape – you can't know just one language
- Communication between technical and non-technical players
- Resources: humans, software, hardware, and time
- Representing complex concepts with data
- Creating production-ready environments that solve problems