Overview – Data Science at IMD

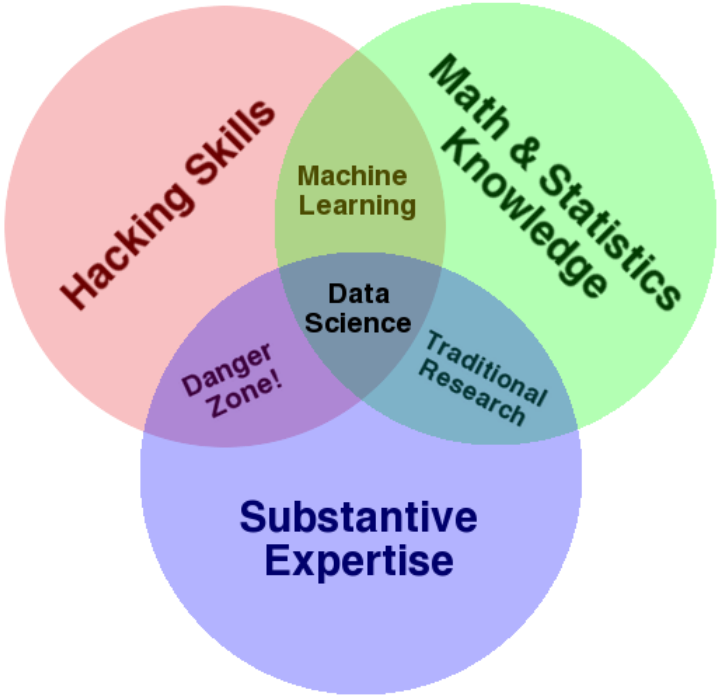
Jung PARK, PhD
Research Fellow in Data Science

First uploaded 2019 July 29
Last modified 2019 Aug 15

What is Data Science

- Data Science is a multidisciplinary field that combines information technology (IT), statistics and management study.
- Due to the rapid advancement of IT, much more data and new analytic techniques became available.
- We need to balance it with sound statistical knowledge and business expertise to create useful insights.

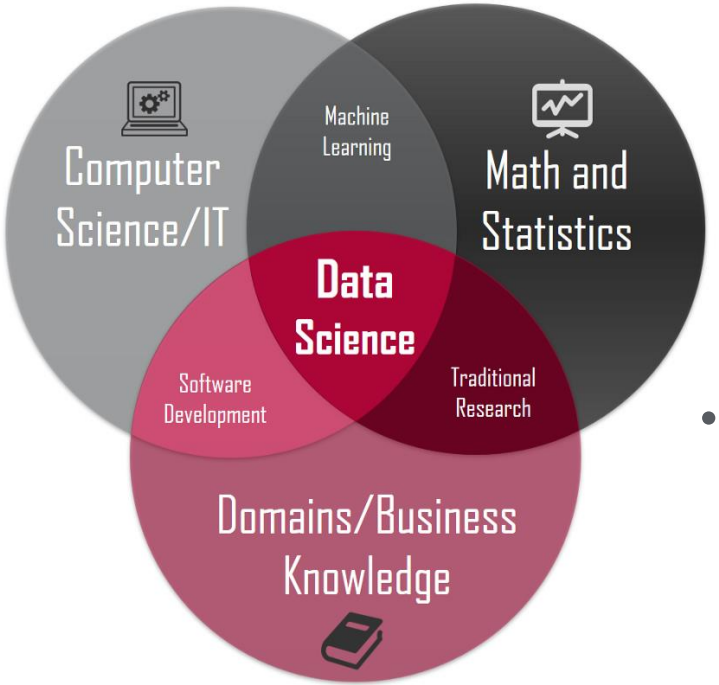
Multidisciplinary field



Data science venn diagram
Source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

- This diagram is often refered to show which knowledge is needed for data science
- “Hacking skills” is eye-catching but too narrow

Multidisciplinary field – alternative description



- This alternative diagram is less popular but more acceptable description

https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

Data Science: analogy to a flying vehicle

- The rapid advancement of IT forces us to adjust other areas accordingly

Source: https://sourceable.net/degree-in-flying-cars-coming-soon/

Five processes in Data Science

We will document knowledge using these five processes as a category

Big Data Challenges and Solutions

Process	Challenges	Solutions	Key references
Data access and collection	• Easy access to data offered in standardized formats. No practical limit to the size of these data offering unlimited scalability • Efficiently obtain detailed data for a large number of agents • Protocols on security, privacy, and data rights	• Sensors • Web scraping • Web traffic and communications monitoring	• Chaffin et al. (2015) • Sismeiro and Bucklin (2004)
Data storage	• Tools for data storage, matching and integration of different big datasets • Data reliability • Warehousing	• SQL, NoSQL, Apache Hadoop • Save essential information only and update in real time	• Varian (2014) • Prajapati (2013)
Data processing	• Use non-numeric data for quantitative analyses	• Text mining tools to transform text into numbers • Emotion recognition	• Manning, Raghavan, and Schütze (2009) • Teixeira, Wedel, and Pieters (2012)
Data analysis	• Large number of variables • Causality • Find latent topics and attach meaning • Data too large to process	• Ridge, lasso, principal components regression, partial least squares, regression trees • Topic modeling, latent Dirichlet allocation, entropy-based measures, and deep learning • Cross-validation and holdout samples • Field experiments • Parallelization, bags of little bootstrap, sequential analysis	• Hastie, Tibshirani, and Friedman (2009) • George and McCulloch (1993) • Archak, Ghose, and Ipeirotis (2011) • Tirunillai and Tellis (2012) • Blei, Ng, and Jordan (2003) • LeCun, Bengio, and Hinton (2015) • Lambrecht and Tucker (2013) • Wang, Chen, Schifano, Wu, and Yan (2015) • Wedel and Kannan (2016)
Reporting and visualization	• Facilitate interpretation, representation with external partners and knowledge users • Difficult to understand complex patterns	• Describe data sources • Describe methods and specifications • Bayesian analysis • Visualization and graphic interpretations	• Loughran and McDonald (2011) • Simonsohn, Simmons, and Nelson (2015)

Data Collecting

- Web scraping from public websites
  - How to collect information from websites
  - 10K reports from SEC.gov
  - EU Open Data
  - Open data Swiss
- Collecting data using API
  - Tweeter
  - Facebook
- Database in IMD library
  - Bloomberg
  - Thomson One
  - Datastream
  - Factiva: news media
- IoT (Internet of Things)
  - Smart building
  - Wearable devices
  - Web traffic

Data Storing

- Distributed computing
  - Do I need it?
  - What is Apache Hadoop
  - What is Apache Spark

- Other relevant information
  - SQL and NoSQL
  - Cloud technolgy

Data Processing

- Data wrangling
  - Merging the unstructured data
  - Text data
  - Images
  - Audios

- Contents analysis
  - NPL (Natural Language Processing)

Data Analysing

- Big data specific (list from McKinsey Global Institute 2011)
  - A/B testing
  - Cluster analysis
  - Data fusion and integration
  - Data mining, genetic algorithms, machine learning
  - Natural language processing
  - Neural networks
  - Network analysis
  - Signal processing and spatial analysis, simulation
  - Time series analysis
  - Visualisation
- Econometrics
  - Time series analysis
  - Score matching techniques
  - Two-stage models for endogeneity problems
- Artificial Intelligence and Machine Learning
  - History of Artificial Intelligence
  - Supervised Machine Learning
  - Unsupervised Machine Learning/ Deep learning/ Neural networks

Data Reporting

- Data Reporting
  - Robustness check
  - Information for reproducibility

- Related information
  - Open science
  - Ethics and privacy
  - GDPR (General Data Protection Regulation)

References

- George, G., Haas, M. R., & Pentland, A. (2014). FROM THE EDITORS BIG DATA AND MANAGEMENT. *Academy of Management Journal*, *57*(2), 321-326.
- Drew Conway's data science venn diagram, accessed 5 Aug 2019 at http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram
- Michael Barber, accessed 5 Aug 2019 at https://towardsdatascience.com/introduction-to-statistics-e9d72d818745