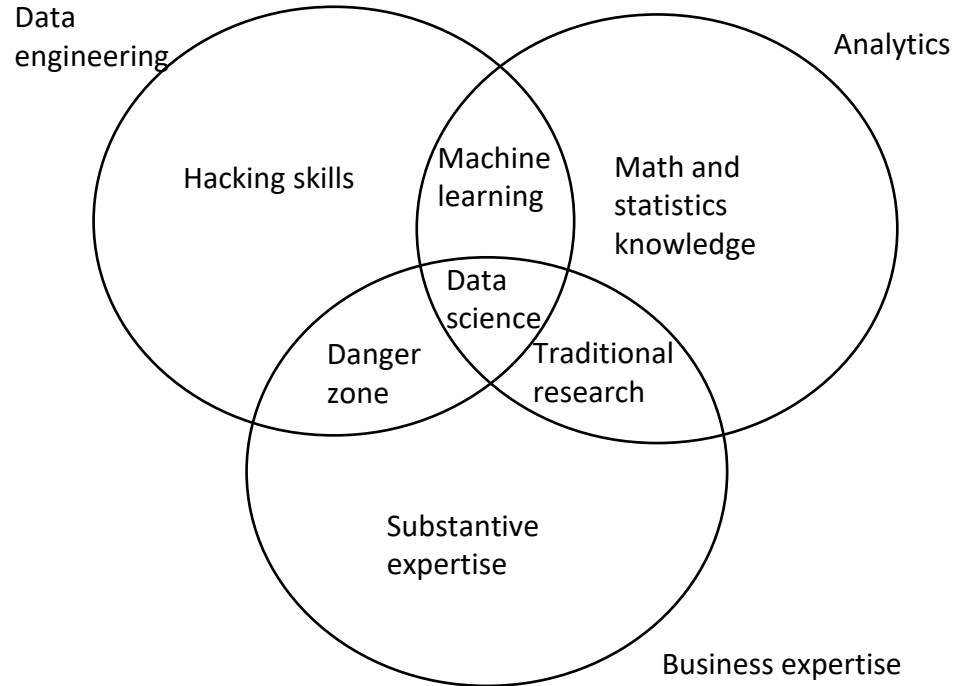


Overview – Data Science at IMD

Jung PARK, PhD
Research Fellow in 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.



Source: Anton Ovchinnikov

Data Science: analogy to a flying vehicle



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

Five processes in Data Science

Big Data Challenges and Solutions

Process	Challenges	Solutions	Key references
Data access and collection	<ul style="list-style-type: none"> • 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 	<ul style="list-style-type: none"> • Sensors • Web scraping • Web traffic and communications monitoring 	<ul style="list-style-type: none"> • Chaffin et al. (2015) • Sismeiro and Bucklin (2004)
Data storage	<ul style="list-style-type: none"> • Tools for data storage, matching and integration of different big datasets • Data reliability • Warehousing 	<ul style="list-style-type: none"> • SQL, NoSQL, Apache Hadoop • Save essential information only and update in real time 	<ul style="list-style-type: none"> • Varian (2014) • Prajapati (2013)
Data processing	<ul style="list-style-type: none"> • Use non-numeric data for quantitative analyses 	<ul style="list-style-type: none"> • Text mining tools to transform text into numbers • Emotion recognition 	<ul style="list-style-type: none"> • Manning, Raghavan, and Schütze (2009) • Teixeira, Wedel, and Pieters (2012)
Data analysis	<ul style="list-style-type: none"> • Large number of variables • Causality • Find latent topics and attach meaning • Data too large to process 	<ul style="list-style-type: none"> • 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 	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Facilitate interpretation, representation with external partners and knowledge users • Difficult to understand complex patterns 	<ul style="list-style-type: none"> • Describe data sources • Describe methods and specifications • Bayesian analysis • Visualization and graphic interpretations 	<ul style="list-style-type: none"> • Loughran and McDonald (2011) • Simonsohn, Simmons, and Nelson (2015)

- 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

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

- Other relevant information
 - SQL and NoSQL
 - Cloud technology

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

- Contents analysis
 - NPL (Natural Language Processing)

- 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
 - Robustness check
 - Information for reproducibility

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