Data Science and Data Analytics (WS 2025/26)

International Business Management (B. A.)

© Benjamin Gross

September 9, 2025

This document provides the course material for Data Science and Data Analytics (B. A. – International Business Management). Upon successful completion of the course, students will be able to: recognize important technological and methodological advancements in data science and distinguish between descriptive, predictive, and prescriptive analytics; demonstrate proficiency in classifying data and variables, collecting and managing data, and conducting comprehensive data evaluations; utilize R for effective data manipulation, cleaning, visualization, outlier detection, and dimensionality reduction; conduct sophisticated data exploration and mining techniques (including PCA, Factor Analysis, and Regression Analysis) to discover underlying patterns and inform decision-making; analyze and interpret causal relationships in data using regression analysis; evaluate and organize the implementation of a data analysis project in a business environment; and communicate the results and effects of a data analysis project in a structured way.

Table of contents

1 Scope and Nature of Data Science

Let's start this course with some definitions and context.

Definition of Data Science:

> The field of Data Science concerns techniques for extracting knowledge from diverse data, with a particular focus on 'big' data exhibiting 'V' attributes such as volume, velocity, variety, value and veracity.

Maneth & Poulovassilis (2016)

Definition of Data Analytics:

Data analytics is the systematic process of examining data using statistical, computational, and domain-specific methods to extract insights, identify patterns, and support decision-making. It combines competencies in data handling, analysis techniques, and domain knowledge to generate actionable outcomes in organizational contexts (Cuadrado-Gallego et al., 2023).

Definition of Business Analytics:

> Business analytics is the science of posing and answering data questions related to business. Business analytics has rapidly expanded in the last few years to include tools drawn from statistics, data management, data visualization, and machine learning. There is increasing emphasis on big data handling to assimilate the advances made in data sciences. As is often the case with applied methodologies, business analytics has to be soundly grounded in applications in various disciplines and business verticals to be valuable. The bridge between the tools and the applications are the modeling methods used by managers and researchers in disciplines such as finance, marketing, and operations.

Pochiraju & Seshadri (2019)

For skills and competencies required for data science, see Skills Landscape.

1.1 Defining Data Science as an Academic Discipline

Data science emerges as an interdisciplinary field that synthesizes methodologies and insights from multiple academic domains to extract knowledge and actionable insights from data. As an academic discipline, data science represents a convergence of computational, statistical, and domain-specific expertise that addresses the growing need for data-driven decision-making in various sectors.

Data science draws from and interacts with multiple foundational disciplines:

- **Informatics / Information Systems:**

  Informatics provides the foundational understanding of information processing, storage, and retrieval systems that underpin data science infrastructure. It encompasses database design, data modeling, information architecture, and system integration principles essential for managing large-scale data ecosystems. Information systems contribute knowledge about organizational data flows, enterprise architectures, and the sociotechnical aspects of data utilization in business contexts.

- **Computer Science (algorithms, data structures, systems design):**

  Computer science provides the computational foundation for data science through algorithm design, complexity analysis, and efficient data structures. Core contributions include machine learning algorithms, distributed computing paradigms, database systems, and software engineering practices. System design principles enable scalable data processing architectures, while computational thinking frameworks guide algorithmic problem-solving approaches essential for data-driven solutions.

- **Mathematics (linear algebra, calculus, optimization):**

  Mathematics provides the theoretical backbone for data science through linear algebra (matrix operations, eigenvalues, vector spaces), calculus (derivatives, gradients, optimization), and discrete mathematics (graph theory, combinatorics). These mathematical foundations enable dimensionality reduction techniques, gradient-based optimization algorithms, statistical modeling, and the rigorous formulation of machine learning problems. Mathematical rigor ensures the validity and interpretability of analytical results.

- **Statistics & Econometrics (inference, modeling, causal analysis):**

  Statistics provides the methodological framework for data analysis through hypothesis testing, confidence intervals, regression analysis, and experimental design. Econometrics contributes advanced techniques for causal inference, time series analysis, and handling observational data challenges such as endogeneity and selection bias. These disciplines ensure rigorous uncertainty quantification, model validation, and the ability to draw reliable conclusions from data while understanding limitations and assumptions.

- **Social Science & Behavioral Sciences (contextual interpretation, experimental design):**

  Social and behavioral sciences contribute essential understanding of human behavior, organizational dynamics, and contextual factors that influence data generation and interpretation. These disciplines provide expertise in experimental design, survey methodology, ethical considerations, and the social implications of data-driven decisions. They ensure that data science applications consider human factors, cultural context, and societal impact while maintaining ethical standards in data collection and analysis.

The interdisciplinary nature of data science requires practitioners to develop competencies across these domains while maintaining awareness of how different methodological traditions complement and inform each other. This multidisciplinary foundation enables data scientists to approach complex problems with both technical rigor and contextual understanding, ensuring that analytical solutions are both technically sound and practically relevant.

For further reading on the academic foundations of data science, see the comprehensive analysis in Defining Data Science as an Academic Discipline.

1.2 Significance of Business Data Analysis for Decision-Making

- Supports evidence-based strategic, tactical, and operational decisions.
- Reduces uncertainty in forecasting, pricing, resource allocation, and risk management.
- Enables performance measurement and continuous improvement.
- Facilitates customer understanding, personalization, and retention strategies.

1.3 Emerging Trends

Key technological and methodological developments shaping the data landscape:

- Evolution of computing and data processing architectures.
- Digitalization of processes and platforms.
- Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL).
- Big Data ecosystems (volume, velocity, variety, veracity, value).
- Internet of Things (IoT) and sensor-driven data generation.
- Cloud computing and elastic infrastructure.
- Blockchain for distributed trust and data integrity.
- Industry 4.0: cyber-physical systems and automation.
- Remote and hybrid working environments: collaboration, distributed analytics, governance.

1.4 Types of Analytics

- Descriptive Analytics: What happened?
- Predictive Analytics: What is likely to happen?
- Prescriptive Analytics: What should we do?

2 Data Analytic Competencies

xy

2.1 Types of Data

- Cross-sectional data
- Panel (longitudinal) data
- Time-series data
- Geo-referenced / spatial data
- (Potentially) streaming / real-time data

2.2 Types of Variables

- Continuous (interval/ratio)
- Count
- Ordinal
- Categorical (nominal / binary)
- (Possibly) compositional or hierarchical structures

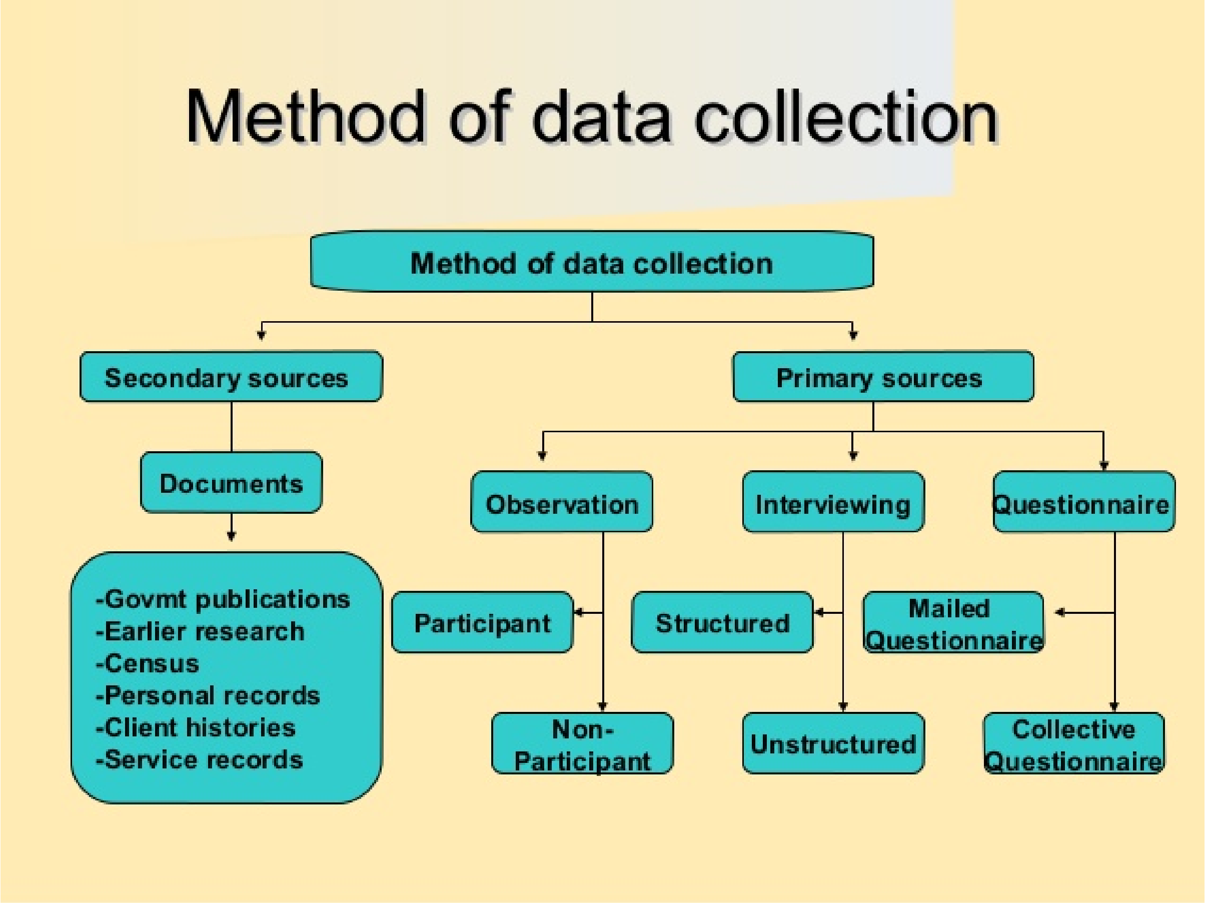
2.3 Conceptual Framework: Knowledge & Understanding of Data

- Clarify analytical purpose and domain context.
- Define entities, observational units, and identifiers.
- Align business concepts with data structures.

2.4 Data Collection

Data collection forms the foundational stage of any data science project, requiring systematic approaches to gather information that aligns with research objectives and analytical requirements. As outlined in modern statistical frameworks, effective data collection strategies must balance methodological rigor with practical constraints (M. & Hardin, 2021).

Figure 1: Methods of Data Collection



2.4.1 Core Data Collection Competencies

The competencies required for effective data collection encompass both technical proficiency and methodological understanding (see Data Collection Competencies.pdf):

- **Source Identification and Assessment**: Systematically identify internal and external data sources, evaluating their relevance, quality, and accessibility for the analytical objectives.
- **Data Acquisition Methods**: Implement appropriate collection techniques including APIs, database queries, survey instruments, sensor networks, web scraping, and third-party vendor partnerships, ensuring methodological alignment with research design.
- **Quality and Governance Framework**: Establish protocols for assessing data provenance, licensing agreements, ethical compliance, and regulatory requirements (GDPR, industry-specific standards).
- **Methodological Considerations**: Apply principles from research methodology to ensure data collection approaches support valid statistical inference and minimize bias introduction during the acquisition process.

2.4.2 Contemporary Data Collection Landscape

Modern data collection operates within an increasingly complex ecosystem characterized by diverse data types, real-time requirements, and distributed sources. The integration of traditional survey methods with emerging IoT sensors, social media APIs, and automated data pipelines requires comprehensive competency frameworks that address both technical implementation and methodological validity.

For comprehensive coverage of data collection methodologies and best practices, refer to: Research Methodology - Data Collection

2.5 Data Management

- Organize: schema design, naming conventions.
- Clean: resolve duplicates, inconsistencies, missingness.
- Convert: type casting, normalization, encoding.
- Curate: maintain lineage, documentation, metadata.
- Preserve: backups, versioning, retention policies.

2.6 Data Evaluation

- Plan analyses aligned with objectives and stakeholders.
- Conduct exploratory, inferential, and predictive procedures appropriately.
- Evaluate robustness, reliability, and validity.
- Assess limitations, bias, and ethical impact.

3 Applications in the Programming Language R

xy

3.1 Core tidyverse Tooling

Explore fundamental packages: * `dplyr` for data manipulation (filter, mutate, summarise, joins). * `tidyr` for data reshaping (pivoting, nesting, separating, unnesting). * `ggplot2` for layered grammar-based visualization. * (Optionally) `readr`, `purrr`, `stringr`, `forcats` for ingestion, functional iteration, text, and factor handling.

3.2 Data Visualization Principles

- Choose encodings appropriate to variable types.
- Emphasize clarity: reduce chart junk; apply perceptual best practices.
- Support comparison, trend detection, and anomaly spotting.

3.3 Detecting Outliers and Anomalies

- Rule-based methods (IQR, z-scores).
- Robust statistics (median, MAD).
- Model-based or multivariate detection (e.g., Mahalanobis distance, clustering residuals).
- Distinguish errors vs. novel but valid observations.

3.4 Dimensionality Reduction

- Motivation: mitigate multicollinearity, noise, and curse of dimensionality.
- Techniques: Principal Component Analysis (PCA), Factor Analysis, (optionally) t-SNE / UMAP (for exploration).
- Interpretability vs. compression trade-offs.

3.5 Data Exploration and Mining

- Structured EDA workflow: question → visualize → quantify → refine.
- PCA for variance structure.
- Factor Analysis for latent constructs.
- Regression Analysis for relationships and predictive structure.
- Clustering (k-means, hierarchical) for pattern discovery (if included).

3.6 Causal Inference with Regression Analysis

- Distinguish association vs. causation.
- Model specification and confounding control.
- Assumptions: linearity, independence, homoskedasticity, exogeneity.
- Interpretation of coefficients and marginal effects.
- Sensitivity and robustness checks.

4 Literature

All references for this course.

4.1 Essential Readings

Békés, G. and G. Kézdi (2021). *Resources for Data Analysis for Business, Economics, and Policy.* https://www-cambridge-org.eux.idm.oclc.org/highereducation/books/data-analysis-for-business-economics-and-policy/D67A1B0B56176D6D6A92E27F3F82AA20/resources/instructor-resources/F57C5762D1593E72250729668A08A53B. https://github.com/DrBenjamin/Analytical-Skills-for-Business/blob/c2ec1b2061c7dc36200977cfd58daf6020c1c774/literature/B%C3%A9k%C3%A9s_Data%20Analysis%20for%20Business%2C%20Economics%2C%20and%20Policy_2021_First%20Day%20of%20Class%20Slides.pdf/?raw=true.}

Evans, J. R. (2020). “Evans, J. R. (2020). Business analytics’’.

Huntington-Klein, N. (2025). *The Effect: An Introduction to Research Design and Causality.* https://theeffectbook.net/.

Wickham, H., M. Çetinkaya-Rundel, and G. Grolemund (2023). *R for Data Science (2e).* https://r4ds.hadley.nz/.

4.2 Further Readings

Cuadrado-Gallego, J. J., Y. Demchenko, J. G. Pérez, et al. (2023). “Data Analytics: A Theoretical and Practical View from the EDISON Project’’. In: *Data Analytics: A Theoretical and Practical View from the EDISON Project*, pp. 1-477. DOI: 10.1007/978-3-031-39129-3/COVER.

Irizarry, R. A. (2024). *Introduction to Data Science: Data Wrangling and Visualization with R.* https://rafalab.dfci.harvard.edu/dsbook-part-1/.

Kumar, U. D. (2017). “Business analytics: The science of data-driven decision making.’’

M., Ç. and J. Hardin (2021). *Introduction to Modern Statistics.* https://www.openintro.org/book/ims/. https://github.com/DrBenjamin/Analytical-Skills-for-Business/blob/491a9a84dd0227aab44e0a6db7e6330830a
literature/Introduction_to_Modern_Statistics_2e.pdf/?raw=true.

Maneth, S. and A. Poulovassilis (2016). “Data Science’’. In: *Advance Access publication on* 12. DOI: 10.1093/comjnl/bxw073. https://academic.oup.com/comjnl/article/60/3/285/2608072.

Pochiraju, B. and S. Seshadri (2019). *Essentials of Business Analytics.* Ed. by B. Pochiraju and S. Seshadri. Vol. 264. https://link.springer.com/10.1007/978-3-319-68837-4. Springer International Publishing. ISBN: 978-3-319-68836-7. DOI: 10.1007/978-3-319-68837-4. https://github.com/DrBenjamin/Analytical-Skills-for-Business/blob/c2ec1b2061c7dc36200977cfd58daf6020c1c774/literature/Essentials%20of%20Business%20Analytics.pdf/?raw=true.}

Vaughan, D. (2020). “Analytical skills for AI and data science’’.

https://learning.oreilly.com/library/view/analytical-skills-for/9781492060932/preface01.html#idm46388898852872.