Elements of Machine Learning and Data Science

Part I: Data Science — Exam Notes (Living Document)

Emir Pisirici

January 29, 2026

Exam likelihood: High (overall Data Science part)

This document is structured to match the lecture topics exactly and is designed for adding **exam-style notes**, **common traps**, and **visual summaries**.

Contents

1 Introduction to Data Science

1.1 Introduction

1.2 Tabular Data

1.3 Data Science Process

Exam likelihood: High

Framework questions are easy to grade and strongly test “big picture” understanding.

Examiner favorite (what they love to ask)

Typical asks: **ETL vs ELT**, **CRISP-DM phases**, and mapping a scenario to the correct phase. Also: where data leakage/bias lives (data understanding + evaluation).

1.3.1 ETL vs ELT (Definitions + Differences)

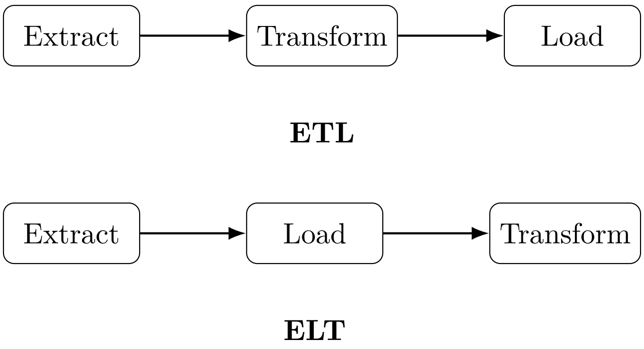
Cheat sheet / must-memorize

ETL: Extract → Transform → Load (transform before target).
ELT: Extract → Load → Transform (transform inside target platform).
Key contrast: where transformations happen; governance vs flexibility; raw history availability.

Common pitfall

People confuse “ELT = no cleaning”. Wrong. It means cleaning happens *after loading*, often in warehouse/lakehouse layers (staging → curated).

Visual

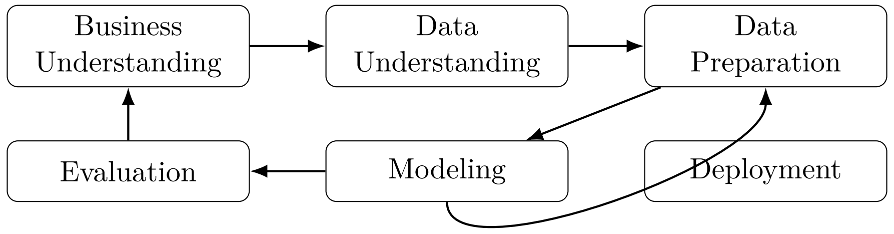


1.3.2 CRISP-DM

Cheat sheet / must-memorize

CRISP-DM: Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Deployment (iterative loops).

Visual



1.3.3 PDCA

Cheat sheet / must-memorize

PDCA: Plan → Do → Check → Act (continuous improvement loop).

1.3.4 DMAIC

Cheat sheet / must-memorize

DMAIC: Define → Measure → Analyze → Improve → Control. Often used for process/quality improvement + monitoring and part of the Six Sigma methodology.

1.4 Data Types

Exam likelihood: Medium

Often tested as a quick classification question or as a setup for choosing plots/models.

Examiner favorite (what they love to ask)

Identify the correct type for a variable and justify the appropriate summary or visualization.

Why (motivation): The data type tells you which summaries, plots, and models are valid.
What (definition): A data type is the measurement scale or structure of a variable.
How (procedure/usage): Classify each variable, then pick the right summary/plot/encoding. Also note data *structure* (structured vs semi vs unstructured) at the dataset level.

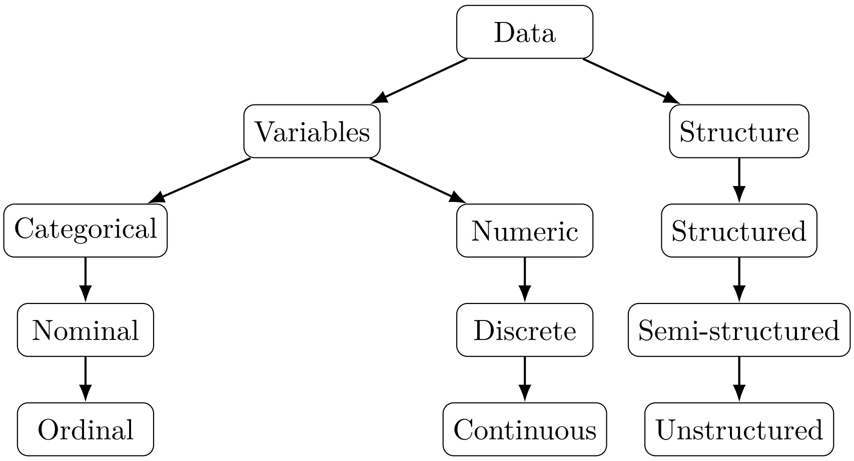
Cheat sheet / must-memorize

- **Categorical (nominal):** labels, no order (color, country, brand).
- **Categorical (ordinal):** ordered labels (low/med/high, Likert scale).
- **Numeric (discrete):** counts (number of emails, defects).
- **Numeric (continuous):** measurements (height, time, temperature).
- **Binary:** special nominal (0/1, yes/no).
- **Date/Time:** timestamps, durations (2024-01-01, 3.2 hours).
- **Text:** free-form strings (review text, comments).
- **Measurement scales:** interval vs ratio (ratio has a true zero).
- **Structure:** structured (tables), semi-structured (JSON/XML), unstructured (text, images, audio).

Common pitfall

Treating IDs or ordinal labels as numeric (mean of zip codes or Likert scores) can mislead. Also avoid mixing up variable types with dataset structure.

Visual



Key takeaways: Know the core families and give a quick example for each.

1.5 Descriptive Statistics

1.6 Basic Visualizations

1.7 Feature Transformations

1.8 “How to lie with statistics”

2 Decision Trees

2.1 Introduction to Decision Trees

2.2 Entropy and Information Gain

Exam likelihood: Very High

Almost guaranteed: compute entropy / information gain on a small dataset.

2.3 ID3 Algorithm

2.4 Pruning

2.5 Continuous Data (Threshold splits)

2.6 Ensembles (Bagging/Random Forest/Boosting)

3 Clustering

3.1 Introduction to Unsupervised Learning

3.2 Introduction to Clustering

3.3 Similarity and Dissimilarity

3.4 K-means and K-medoids

3.5 Agglomerative Clustering

3.6 DBSCAN

3.7 Closing

4 Frequent Itemsets

4.1 Introduction

4.2 Properties of Frequent Itemsets

4.3 Apriori Algorithm

4.4 FP-Growth Algorithm

5 Association Rules

5.1 Introduction

5.2 Generating Association Rules

5.3 Evaluation (support, confidence, lift, conviction)

5.4 Applications

5.5 Simpson's Paradox

6 Time Series

6.1 Temporal Data

6.2 Introduction to Time Series

6.3 Analysis

6.4 Forecasting