

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	3
1.1	Introduction	3
1.2	Tabular Data	3
1.3	Data Science Process	3
1.3.1	ETL vs ELT (Definitions + Differences)	3
1.3.2	CRISP-DM	3
1.3.3	PDCA	4
1.3.4	DMAIC	4
1.4	Data Types	4
1.5	Descriptive Statistics	4
1.6	Basic Visualizations	4
1.7	Feature Transformations	4
1.8	“How to lie with statistics”	4
2	Decision Trees	5
2.1	Introduction to Decision Trees	5
2.2	Entropy and Information Gain	5
2.3	ID3 Algorithm	5
2.4	Pruning	5
2.5	Continuous Data (Threshold splits)	5
2.6	Ensembles (Bagging/Random Forest/Boosting)	5
3	Clustering	6
3.1	Introduction to Unsupervised Learning	6
3.2	Introduction to Clustering	6
3.3	Similarity and Dissimilarity	6
3.4	K-means and K-medoids	6
3.5	Agglomerative Clustering	6
3.6	DBSCAN	6
3.7	Closing	6

4	Frequent Itemsets	7
4.1	Introduction	7
4.2	Properties of Frequent Itemsets	7
4.3	Apriori Algorithm	7
4.4	FP-Growth Algorithm	7
5	Association Rules	8
5.1	Introduction	8
5.2	Generating Association Rules	8
5.3	Evaluation (support, confidence, lift, conviction)	8
5.4	Applications	8
5.5	Simpson's Paradox	8
6	Time Series	9
6.1	Temporal Data	9
6.2	Introduction to Time Series	9
6.3	Analysis	9
6.4	Forecasting	9

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).

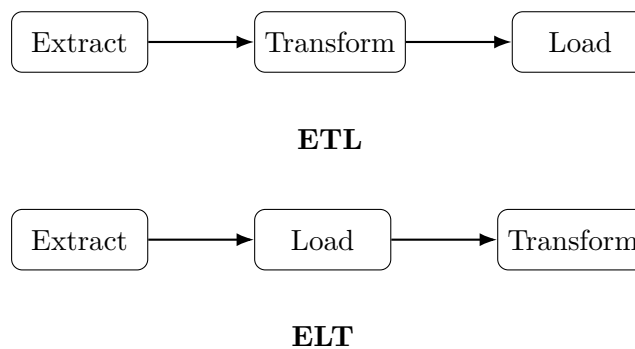
ELT: Extract → Load → Transform (transform inside target platform).

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 (for your cortex)

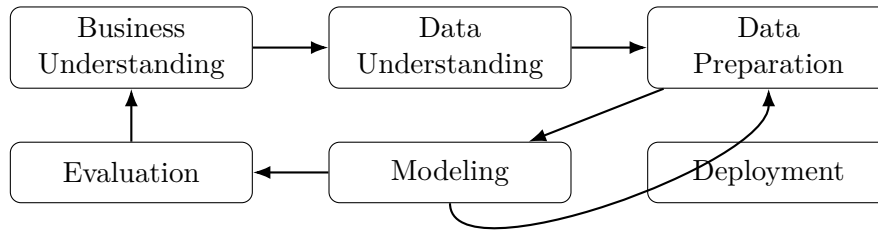


1.3.2 CRISP-DM

Cheat sheet / must-memorize

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

Visual (for your cortex)



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

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