COURSE GLOSSARY

Introduction to Data

Aggregation: The process of summarizing many individual data points into simpler metrics (for example sums, averages, minimums, maximums, or modes) to make large datasets interpretable

Correlation vs causation: The important distinction that correlation indicates a statistical relationship between variables while causation means one variable directly causes changes in another, and the two should not be assumed equivalent

Data anonymization: Techniques that remove, mask, or transform personally identifiable details so records cannot be traced back to individuals while preserving analytical value

Data bias: Systematic errors in data collection, sampling, or measurement that make a dataset unrepresentative and can lead to misleading or unfair conclusions

Data context: Information about the circumstances and conditions surrounding data (for example units, timing, or source) that determines its meaning and appropriate use

Data ethics: The set of moral principles guiding the responsible collection, storage, and use of data, including consent, transparency, fairness, and consideration of outcomes

Data governance: The policies, roles, and processes that ensure data is consistent, trustworthy, and used responsibly across an organization

GraphQL: A query language and runtime for APIs that lets clients request exactly the data they need and aggregate responses in a single request

Least-privilege access: A security principle that grants users the minimum permissions required to perform their job, reducing exposure of sensitive data and actions

Data quality: The degree to which data is accurate, complete, valid, and consistent for its intended purpose

Data-driven decision making: A repeatable process of asking questions, collecting and preparing the right data, performing analysis, and interpreting results to inform choices rather than relying on intuition alone

Data: Raw facts or observations in any form (numbers, text, images, sounds) that have little meaning until they are given context or processed

DIKW pyramid: A mental model (Data → Information → Knowledge → Wisdom) that describes how raw data is refined through context and understanding to support better decisions

Key Performance Indicator (KPI): A measurable metric derived from data that tracks performance against strategic goals or benchmarks

Metadata: Data that describes other data—such as when, where, how, or by whom it was collected—providing context needed to interpret the underlying data

PII (Personally Identifiable Information): Any information that can uniquely identify an individual, such as full name, birthdate, ID numbers, phone numbers, or addresses

Qualitative data: Non-numerical, descriptive data that captures categories, attributes, opinions, or labels used to identify patterns and meanings

Quantitative data: Numerical data that expresses amounts or measurements and is suitable for calculations and statistical analysis

Structured data: Data organized in a predefined format (like tables, rows, and columns) that is easy to filter, query, and analyze

Unstructured data: Data without a fixed schema or organization—such as videos, images, or free-text—that requires processing before traditional analysis