1.3 — Data Analytics Lifecycle, Quality and Quantity of Data, Measurement

---

Overview

This session comprehensively covers **Unit 1.3** as per your syllabus, connecting critical concepts to real-world scenarios, structured for clarity and depth. The focus is on understanding the data analytics lifecycle, the importance of data quality and quantity, and the various types of data measurement.

---

1. The Data Analytics Lifecycle

Definition

The data analytics lifecycle is a systematic, step-by-step process that guides how data is gathered, prepared, analyzed, modeled, and used for business decisions. It ensures that analytics projects are structured and repeatable.

Main Phases

Phase	Description	Real-Life Example/Case Scenario
1. Discovery	Identify the problem, objectives, and data sources.	A hospital wants to reduce patient readmission rates.
2. Data Preparation	Cleaning, filtering, and organizing data for analysis.	Collating patient records, removing duplicates, and correcting errors.
3. Model Planning	Selecting statistical/machine learning models to solve the given business problem.	Deciding between logistic regression or decision tree for risk prediction.
4. Model Building	Creating and training the chosen models on prepared data.	Building the predictive model to identify high-risk patients.
5. Operationalize	Deploying the model for practical use—often in live or pilot environments.	Integrating the model into hospital workflow for daily checks.
6. Communicate	Visualizing and reporting results to stakeholders for decision-making.	Dashboard showing reduction in readmissions to hospital staff.

Origin & Research

- Lifecycle models were popularized by methodologies like **CRISP-DM** (Cross-Industry Standard Process for Data Mining, 1999), which structured analytics projects into clear, repeatable steps.
- Emphasized in Thomas & Cook's work on visual analytics and by industry in Big Data portfolios.

Why Follow a Lifecycle?

- Ensures reliability, repeatability, and alignment with business goals, leading to more effective decision-making and resource allocation.

---

2. Models in Data Analytics

Definition

A **model** is a mathematical or computational representation of real-world relationships within data. Models are used to predict, classify, or discover patterns.

Types of Models

- **Statistical Models:** Examples include linear regression and normal distribution analysis.
- **Machine Learning Models:** Examples include decision trees and neural networks.
- **Simulation Models:** Examples include queuing simulations in hospitals to streamline patient flow.

Purpose

- To predict outcomes, detect patterns, and generate actionable insights that can inform business strategies.

Example

A bank builds a credit scoring model to evaluate loan risks by analyzing repayment history and income, allowing for better risk management.

---

3. Data Quality

Definition

Data quality measures how well data is suited to its intended purpose, based on factors such as **accuracy, completeness, reliability, timeliness, consistency, and validity**.

Key Dimensions

- **Accuracy:** Is the data correct?
    - *Real-life Example:* Customer addresses updated after verification.
- **Completeness:** Is any data missing?
    - *Scenario:* A dataset missing patient age may skew healthcare analysis.
- **Timeliness:** How current is the data?
    - *Example:* Stock market data needs real-time updates.
- **Consistency:** Are data values uniform across datasets?
    - *Scenario:* “NY” vs. “New York” can cause duplication.
- **Validity:** Does data conform to required formats/rules?
    - *Example:* Phone number fields contain only valid numbers.

Impact

- High-quality data leads to valid, reliable results. Poor quality can mislead analysts and decision-makers, resulting in poor business outcomes.

---

4. Data Quantity

Definition

Data quantity refers to the **volume and depth** of data collected, which can significantly impact analysis outcomes.

- **Granularity:** The level of detail collected (e.g., every hospital visit vs. monthly summary).
- **Scope:** How broadly data represents the phenomena (e.g., all patients vs. a single department).

Balancing Quantity

- **Too Little Data:** Results in weak inference and unreliable conclusions.
- **Too Much Data:** Leads to higher storage and computation costs.
- **Best Practice:** Strive for just enough data detail and coverage to provide robust insights efficiently.

Example

An airline analyzes millions of boarding passes yearly but may sample only 10% to study queue bottlenecks, ensuring efficient resource use while still gaining valuable insights.

---

5. Data Measurement Types

Levels of Data Measurement

Level	Description	Example
Nominal	Categories without order	Gender, blood groups
Ordinal	Categories with order, but not precise differences	Satisfaction rating (low–high)
Interval	Ordered, constant difference, no true zero	Temperature in Celsius
Ratio	Ordered, constant difference, true zero	Height, weight, age, income

Origin

- Proposed by S.S. Stevens in “On the Theory of Scales of Measurement” (Science, 1946).

Application

- Correct type informs analysis (e.g., mean for ratio data, mode for nominal data).

---

6. Case Scenario — Healthcare Data Analytics

A hospital uses the analytics lifecycle to reduce readmissions:

- **Discovery:** Sets a goal to reduce readmissions.
- **Data Preparation:** Gathers patient medical records and cleans inconsistencies.
- **Model Planning/Building:** Develops a predictive model to flag high-risk patients.
- **Operationalize:** Assigns follow-up calls to flagged patients.
- **Results:** Achieves a 15% reduction in readmissions in 6 months and improves patient satisfaction.

---

7. Visuals for Core Concepts

- **Lifecycle Diagram:** Illustrating the phases of the data analytics lifecycle.
- **Data Quality and Quantity Matrix:** Showing the relationship between data quality dimensions and their impact.
- **Levels of Measurement Chart:** Visual representation of the different levels of data measurement.

8. References to Foundational Concepts

- **CRISP-DM Standard:** “CRISP-DM 1.0 Step-by-step Data Mining Guide,” Chapman et al., 1999.
- **S.S. Stevens:** “On the Theory of Scales of Measurement,” Science, 1946.
- **Thomas & Cook:** “Illuminating the Path: The Research and Development Agenda for Visual Analytics,” 2005.
- Your attached PPT and syllabus for context.

Key Takeaways

- **Lifecycle Approach:** Improves efficiency and reliability in analytics projects.
- **Data Quality and Quantity:** Must be actively managed for valid results.
- **Measurement Levels:** Determine suitable analytical techniques.
- Real-world cases showcase how structured analytics deliver tangible organizational value.