Mary Kane – Exercise 5.1 Big Data

1. What's the difference between structured and unstructured data? Can you give examples that you've encountered for both types?

Structured data is information that's organized in a predictable format (usually rows and columns) with clearly defined fields.

Example: The orders table of the Instacart project, which provided information in each row, organized by columns "order_id", "user_id", "order_number", etc.

Unstructured data is information that doesn't follow a fixed schema and is harder to fix into a table.

Example: Free-form text, images, audio, video, emails, chat logs.

2. Given that much of big data is produced by machines and sensors, how trustworthy do you think that big data is? What characteristic of big data relates to the question of trustworthiness?

Big data can help us reduce human errors, but its trustworthiness depends on how the data was captured, transmitted, and interpreted. If monitored and validated, data can be highly trustworthy for patterns and trends. "Veracity" is the characteristic that relates to data quality, accuracy, uncertainty and reliability.

3. Assume that you receive a table containing customer data. You notice that some values are missing or incomplete, and the formatting is inconsistent in some columns. Based on what you've learned so far, how would you go about cleaning this table? Think about what you would do first, second, third, etc.

First, I would inspect and profile the data (shape, column names, data types, get missing and duplicate value-counts, spelling check). Second, I would apply necessary fixes, such as remove duplicates and fix typos. For inconsistent data, applying standardized formatting to ensure uniformity across the dataset improves overall quality and analysis readability.

4. Can you describe tools such as Hadoop and Apache Spark and their role in big data? What do they do and how do they work?

Hadoop and Apache Spark are big data frameworks used to store and process large datasets efficiently. Hadoop uses distributed storage (HFDS) and a processing model known as MapReduce to break down data into smaller part processed across multiple computers. Apache Spark utilizes HFDS while running an in-memory processing allowing it to perform faster and support real-time analytics. Together, they make it possible to handle massive amounts of structured and unstructured data quickly and cost-effectively.

5. How has the application of analytics to big data led to new discoveries and innovation? Can you give some examples?

Analytics applied to big data has shown significant discoveries in many fields because it spots patterns that are invisible in small samples, combines many data sources, and tests ideas quickly and continuously. This creates a feedback loop for learning and innovation. Some examples:

In healthcare*, it helps predict disease outbreaks and improve* early diagnosis.

In e-commerce*, it helps personalize customer promotions based on purchase history.*

In transportation, *companies can optimize their delivery routes and reduce fuel costs.*