

Name: \_\_\_\_\_ Date: \_\_\_\_\_

## Data Dynasty: The 12th Grade Literacy Legacy Challenge

Can you spot the bias in a billion-point dataset? Critique real-world sampling errors and ethical dilemmas in high-stakes algorithmic decision-making.

---

**1. A public health researcher uses 'Proxy Variables' like zip codes to predict health outcomes because direct socio-economic data is unavailable. What is the primary data literacy concern here?**

- A. Selection bias causing overrepresentation of urban areas
- B. Encoding bias where geographical data hides systemic disparities
- C. Lack of data sovereignty for the participants
- D. Data decay due to outdated census records

**2. In a longitudinal study, 'Data Attrition' refers to the systematic loss of participants over time, which can threaten the validity of the final analysis.**

- A. True
- B. False

**3. When an analyst chooses only the data points that support their preconceived hypothesis while ignoring contradictory evidence, they are performing \_\_\_\_.**

- A. Data Normalization
- B. Data Scrubbing
- C. Cherry Picking
- D. Linear Regression

**4. The 'Simpson's Paradox' occurs when a trend appears in several different groups of data but disappears or reverses when these groups are combined. What does this highlight about data interpretation?**

- A. The importance of visualizing data via scatter plots
- B. The danger of ignoring lurking variables
- C. The necessity of using a larger sample size
- D. The superiority of qualitative data over quantitative data

**5. To ensure data integrity and prevent unauthorized 'data silos,' organizations implement \_\_\_\_ frameworks to define who has authority over data assets.**

- A. Data Governance
- B. Machine Learning
- C. Web Scraping
- D. Sentiment Analysis

**6. A tech company releases a dataset but applies 'Differential Privacy' techniques before publication. What is the primary goal of this action?**

**Name:** \_\_\_\_\_ **Date:** \_\_\_\_\_

- A. To increase the speed of data processing
- B. To compress the data for easier storage
- C. To protect individual identities by adding controlled noise to the data
- D. To ensure the data is compatible with open-source software

**7. Correlation always implies causation if the R-squared value of a dataset is higher than 0.95.**

- A. True
- B. False

**8. Evaluating a source's 'Provenance' in data literacy refers specifically to investigating the \_\_\_\_\_ of the data.**

- A. Storage capacity
- B. Visual aesthetics
- C. Origin and history
- D. Encryption level

**9. When analyzing the 'Digital Divide,' a researcher notes that data collected via smartphone apps excludes elderly populations. This is an example of what?**

- A. Measurement Error
- B. Undercoverage Bias
- C. Confirmation Bias
- D. Data Redundancy

**10. Metadata is essentially 'data about data,' providing context such as when, where, and how the primary data was collected.**

- A. True
- B. False