Name: _____ **Date:** _____

Answer Key: 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?

Answer: B) Encoding bias where geographical data hides systemic disparities

Using proxies like zip codes often introduces encoding bias, as the geographical data may inadvertently reflect historical redlining or systemic inequalities rather than individual health behaviors.

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.

Answer: A) True

Data attrition occurs when participants drop out of a study; if those who drop out share specific characteristics, the remaining data is no longer representative of the original population.

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

Answer: C) Cherry Picking

Cherry picking is a logical fallacy and a data literacy failure where individuals select a subset of data to confirm a bias, suppressing a more complete and accurate picture.

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?

Answer: B) The danger of ignoring lurking variables

Simpson's Paradox demonstrates that lurking variables (confounders) can drastically change results, requiring high-level critical analysis to ensure the context of the data is understood.

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

Name: _____ **Date:** _____

Answer: A) Data Governance

Data governance is the architectural and legal framework that manages the availability, usability, integrity, and security of data in an organization.

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

Answer: C) To protect individual identities by adding controlled noise to the data

Differential privacy is a sophisticated mathematical approach that allows researchers to share patterns in data without revealing specific information about any single individual in the set.

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

Answer: B) False

Even with an extremely high statistical correlation, one cannot assume causation. Spurious correlations can exist between unrelated variables due to coincidence or a third common factor.

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

Answer: C) Origin and history

Provenance is the record of ownership and custody of a dataset, which is critical for verifying its authenticity and identifying potential biases introduced during its lifecycle.

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

Answer: B) Undercoverage Bias

Undercoverage bias occurs when some members of a population are inadequately represented in the sampling frame, leading to insights that cannot be generalized to the whole population.

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

Answer: A) True

Metadata is a fundamental component of data management, allowing users to understand the structural and administrative context of a dataset without viewing the content itself.

Name: _____ Date: _____