Name: _____ **Date:** _____

Answer Key: Raw Facts, Fine Print: Senior Data Ethics & Literacy Quiz

Can you spot the algorithmic bias in a dataset? Deconstruct complex data provenance and evaluate the socio-technical implications of information architecture.

---

1. A researcher examines a dataset of urban mobility patterns where data was only collected from users with high-end smartphones. This is an example of which data literacy concern?

Answer: B) Sampling Bias

Sampling bias occurs when certain members of a population are systematically more likely to be included in a dataset than others, such as excluding those who cannot afford expensive hardware.

2. The concept of ___ refers to the chronological record of the origin, movement, and transformations of a dataset, essential for verifying its integrity.

Answer: C) Data Provenance

Data provenance (or lineage) allows researchers to trace data back to its source to evaluate the methodology and ethical standards of its collection.

3. True or False: In a high-stakes predictive model, a high correlation coefficient (r) between two variables is sufficient evidence to establish a direct causal mechanism for policy-making.

Answer: B) False

Correlation does not imply causation; establishing a causal link requires experimental design or rigorous causal inference modeling to rule out confounding variables.

4. When evaluating the 'Veracity' of Big Data in a corporate audit, which factor is most critical to investigate?

Answer: C) The consistency and trustworthiness of the data points

Veracity refers to the quality, accuracy, and uncertainty of data. In an audit, ensuring data hasn't been tampered with or poorly cleaned is a high-level literacy skill.

5. To protect individual privacy in large public datasets, organizations often use ___, which adds 'mathematical noise' to the data to prevent de-identification.

Answer: A) Differential Privacy

Name: _____ **Date:** _____

Differential privacy is a sophisticated technique used by entities like the Census Bureau to share trends without compromising individual identity.

6. Simpson's Paradox is a data phenomenon where a trend appears in several groups of data but ___ when these groups are combined.

Answer: B) Disappears or reverses

Simpson's Paradox is a critical high-level data literacy concept where aggregate data can mislead if the underlying variables are not properly weighed or partitioned.

7. True or False: Using an 'unsupervised learning' algorithm for data analysis eliminates the risk of human bias being integrated into the final output.

Answer: B) False

Internal biases in the training data (historical prejudice, exclusion) are learned by the algorithm regardless of whether the learning is supervised or unsupervised.

8. An analyst uses a ___ to identify outliers in a dataset that might indicate sensor failure or fraudulent activity rather than genuine trends.

Answer: A) Standard Deviation Test

Understanding variance and standard deviation allows analysts to evaluate whether a data point is a significant finding or an error (noise).

9. Which of these is a primary ethical implication of 'Data Persistence' in the context of the Internet of Things (IoT)?

Answer: A) The difficulty of correcting inaccurate historical data

Data persistence means digital footprints remain indefinitely; ethically, this makes 'the right to be forgotten' or correcting past errors very difficult for individuals.

10. True or False: Metadata (data about data) can often reveal more sensitive personal information in aggregate than the actual content of the primary data itself.

Answer: A) True

Metadata like timestamps, geolocation, and duration of communication can be used to map a person's entire life and associations, often bypassing privacy protections on content.