Name: _____ **Date:** _____

Crack the Digital Code: An 8th Grade Data Synthesis Challenge

Students analyze algorithmic bias, evaluate longitudinal datasets, and verify metadata integrity to navigate complex information landscapes.

---

1. When examining a global health dataset spanning 50 years, you notice a sudden, extreme spike in cases during a single year that contradicts local news archives. What is the most rigorous next step in data validation?

A. Delete the outlier to maintain a smooth trend line for your report.
B. Cross-reference the metadata to check for changes in reporting methodology or diagnostic criteria.
C. Assume the digital dataset is more accurate than the physical news archives.
D. Average the spike with the previous year to minimize the impact on the graph.

2. If a machine learning algorithm used by a bank is trained on historical data from an era where certain groups were legally excluded from loans, the resulting AI will likely exhibit _____.

A. Algorithmic bias
B. Data redacting
C. Statistical insignificance
D. Maximum encryption

3. A dataset with a high 'p-value' (greater than 0.05) generally indicates that the observed data patterns are statistically significant and unlikely to have occurred by chance.

A. True
B. False

4. You are building a database of endangered species. Which strategy best ensures 'data integrity' during long-term storage?

A. Storing all records as uneditable physical printouts only.
B. Using checksums and regular redundancy audits to detect file corruption.
C. Sharing the admin password with all researchers to maximize access.
D. Compressing the data into a proprietary format that requires a subscription to open.

5. When a researcher only selects data points that support their preconceived theory while ignoring data that contradicts it, they are engaging in _____.

A. Data scraping
B. Cherry-picking
C. Data encryption
D. Metadata tagging

Name: _____ Date: _____

6. Synthetic data, which is artificially generated rather than collected from real-world events, can be used to protect privacy while still allowing for complex pattern analysis.

A. True
B. False

7. What is the primary risk of using 'Data Scraping' from social media platforms to predict public opinion on a new law?

A. The data is too encrypted to be read by computers.
B. The sample may be biased toward the most vocal or automated users rather than a representative population.
C. Social media data cannot be converted into quantitative metrics.
D. Public opinion is not considered 'data' in computer science.

8. The process of removing personally identifiable information (PII) from a dataset so that individuals cannot be recognized is known as _____.

A. Anonymization
B. Categorization
C. Data Mining
D. Web Crawling

9. Correlation between two variables in a dataset (such as ice cream sales and shark attacks) automatically proves that one variable causes the other to happen.

A. True
B. False

10. A city uses a 'Digital Twin' (a virtual real-time data model) to simulate traffic flow. If the model fails to predict a traffic jam, which data issue is the most likely culprit?

A. The computer used for the simulation was too small.
B. The sensors providing real-time telemetry inputs were miscalibrated or delayed.
C. The city has too many roads for a database to handle.
D. The data was stored in an alphabetized list rather than a numerical one.