Big Data, Machine Learning, and Bias Quiz (Practice, Version B)

Name: _____ Period: _____ Date: _____

Practice version. 10 questions, 10 points, 15 minutes. Choose the single best answer unless a question says to select two.

1. A weather app lets users report hail, snow, or heavy rain at their current location. Thousands of reports come in during each storm. Collecting data this way is an example of which of the following?

(A) Data cleaning
(B) Crowdsourcing
(C) Machine learning
(D) Metadata

2. Astronomers post millions of telescope images online and ask volunteers to classify the shape of each galaxy. Which of the following is the most likely benefit of this approach?

(A) Far more images can be classified than the astronomers could classify on their own.
(B) Every classification is guaranteed to be correct.
(C) The astronomers no longer need to store the images.
(D) The volunteers' classifications cannot contain mistakes.

3. A state government publishes data on air quality at hundreds of monitoring stations as open data. Which of the following is a likely benefit?

(A) The air quality will improve automatically.
(B) The data can no longer contain errors.
(C) The monitoring stations no longer need to collect data.
(D) Researchers, journalists, and the public can analyze the data and look for patterns.

4. A program that analyzes 1,000 rows of data in a second takes far too long to analyze 1 billion rows on the same computer. Which of the following is the best approach?

(A) Only analyze the first 1,000 rows.
(B) Print the data and analyze it by hand.
(C) Split the processing across many computers that work at the same time.
(D) Rename the columns to shorter names.

 CC BY-NC 4.0

5. An e-mail service uses machine learning to filter spam. Which of the following best describes how the spam filter learns?

(A) A programmer writes a rule for every possible spam message.
(B) It finds patterns in many example e-mails that are already labeled “spam” or “not spam” (training data).
(C) It deletes every e-mail from an unknown sender.
(D) It can only recognize the exact e-mails it was trained on.

6. An app that identifies skin conditions from photos was trained mostly on photos of people with light skin. It is much less accurate for people with darker skin. What is the most likely cause?

(A) The training data did not represent the full range of people who use the app.
(B) The phone cameras were not high enough resolution.
(C) The app contains a syntax error.
(D) The users took the photos at the wrong time of day.

7. A bank trains a program on its past loan decisions. In the past, the bank often denied loans to people from certain neighborhoods. The program now gives lower scores to applicants from those neighborhoods. Which of the following best explains this?

(A) The applicants filled out the form incorrectly.
(B) The programmers wrote code to reject those neighborhoods.
(C) The program cannot read addresses.
(D) The program learned patterns from past decisions, and those decisions contained bias.

8. A team is building a program that screens job applications. Which of the following actions would help reduce bias in the program?

(Select two answers.)

(A) Use only data from the company’s most successful past employees.
(B) Include training data from all of the groups of people who will apply.
(C) Hide the program’s results from the hiring team so they cannot question them.
(D) Check whether the program’s results differ unfairly between groups before using it.

9. A face-detection model works poorly for some groups because its training photos came mostly from one website. A team member suggests adding a million more photos from that same website. Which statement is most accurate?

(A) This may not reduce the bias, because the new photos are likely to have the same skew as the old ones.
(B) This will remove the bias, because more data always averages out errors.
(C) Bias in a model can only come from the programmer, not from data.
(D) Adding any data always makes a model more biased.

CC BY-NC 4.0

10. A traffic app relies on drivers to report accidents and slowdowns. Many people use the app in cities, but very few use it in rural areas. Which of the following is a limitation of this crowdsourced data?

(A) Traffic data cannot be stored on a computer.

(B) Reports from cities are always false.

(C) The app may miss or be slow to show problems on rural roads, where there are few people to report them.

(D) The app can only work if every driver reports every trip.

CC BY-NC 4.0