Question 1. Given the fact that traditional surveys can already characterize important indicators related to first person alcohol consumption, why should we analyze other sources of data such as online health reports in social media?

A. Traditional surveys are time consuming and have limited coverage in terms of the number of participants in such studies. Therefore, the results might be only limited to cohort studied but not the public.

B. Mining the vast quantities of online health reports in social communications has the potential to deliver low-cost and high-resolution views into public health phenomena to complement traditional systems.

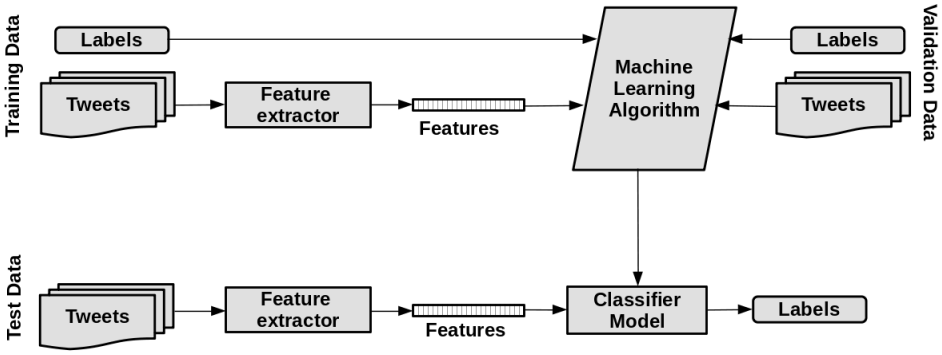
C. Mining the vast quantities of online health reports in social communications can reduce bias from “interviewer effect” which can inversely affect the results of traditional studies.

D. Resorting to social media enables faster access to comprehensive results, significantly better reach, and potentially better targeting. It is also more convenient to participant in online platforms.

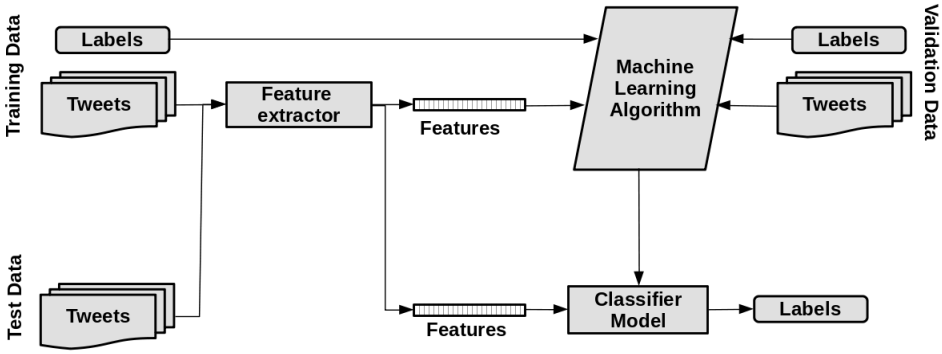
Question 1.
Answer: The most correct answer is **B**.
Rationale: Although traditional surveys are time consuming and have limited coverage in terms of the number people who participate in such studies, the results of these studies are not limited to their cohorts only. Furthermore, although computational models are free of interviewer effects, online health reports are subjective in nature and could carry biased information. Finally, although social media analysis can lead to faster results, significantly better reach, and potentially better targeting, traditional surveys are more comprehensive in terms of the type and number of questions that can be asked by interviewers. In fact, computational models are limited in terms of the type and number of questions that can be computationally modeled. However, mining online health reports in social communications can provide complementary and large scale information that can be used in conjunction with traditional systems to characterize important indicators related to public health issues.

Question 2. Assume that you would like to develop a classifier to address an important public health issue, e.g. classifying tweets as relevant or irrelevant to first-person alcohol consumption. Which of the following diagrams shows a correct design setting for such a classifier?

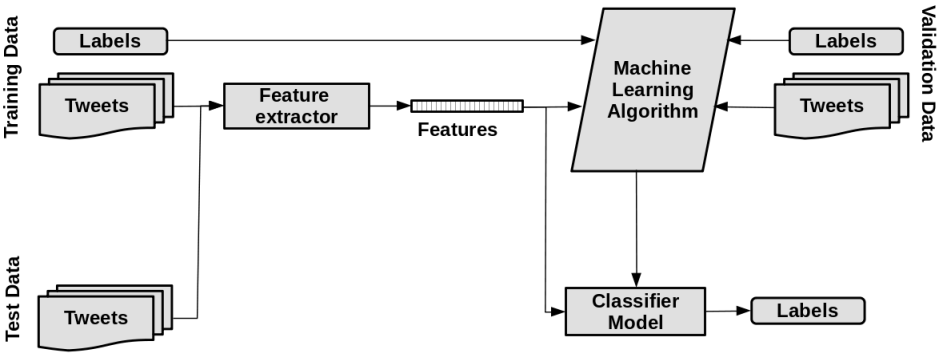
A.



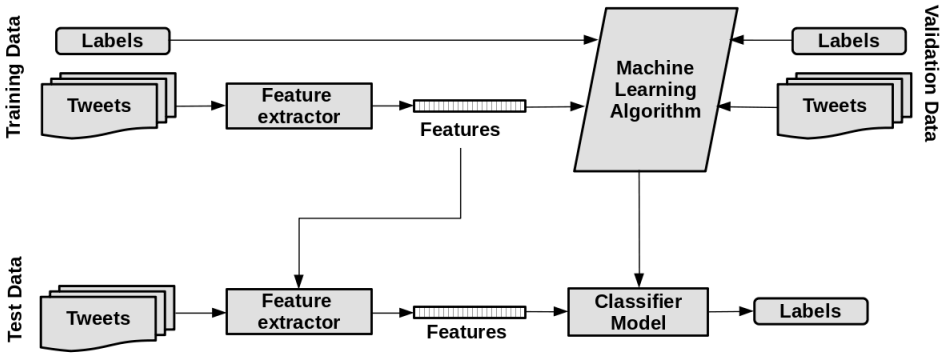
B.



C.



D.



Question 2.
Answer: The correct answer is **A**.
Rationale: In supervised classification (the case where we have both input data and the corresponding label for each input) a feature extractor is used to convert each input data to a feature representation or feature vector, which captures the basic information about the input that will be used to classify it. Pairs of feature vectors and labels are fed into the machine learning algorithm to generate a classification model. During this training process, validation data will be used to tune model parameters for best performance. At test time, the same feature extractor is used to convert test inputs to feature vectors. These feature sets are then fed into the classification model, which generates predicted labels. In this question, the issue with B and C is that the machine learning algorithm should not be given information about the test instances. We can think of training data as books that a learner (here the classifier) reads for taking an exam, validation data as sample exam questions from past years, and test data as the actual exam questions. A learner must not be given any of the actual exam questions while preparing for the exam (i.e. during training). The issue with the design in D is that it unnecessarily uses training feature vectors when testing the model. For more information, please see the following lecture: The Task of Text Classification by Dan Jurafsky at https://web.stanford.edu/class/cs124/lec/naivebayes.pdf (retrieved on 8/01/2018).