# Graded Exercise 3

CSC230 Database Technologies for Analytics

20 November 2017

1. Find as many technical terms as you can in the conversation between Susan Athey and Russ Roberts.

2. In her conversation with Russ Roberts, Susan Athey discusses some of the challenges of understanding the likely and actual effects of changes in a public policy. To accomplish this, she uses which public policy as an example?

3. In which cities might people continue buying lunches at restaurants, even as those restaurants increase their prices?

4. Russ Roberts concedes that people who have earned college degrees enjoy higher incomes than do people who have not earned college degees, but argues that it does not follow that increasing the number of people with college degrees will not necessarily increase incomes. Why not?

5. Russ Roberts talks about using statistical software for "fishing expeditions." What does he mean by that phrase?

6. Russ Roberts speaks about spurious correlations. He offers as an example an imagined correlation that builds on the fact that San Francisco and Seattle share something in common that could not possibly explain wages, rates of employment, or other variables that interest us very much. Which common characteristic of the two cities does Russ Roberts use in his unserious example?

7. Susan Athey to label questions that ask: "What would have happened if a policy had not been changed?" What word does she use?

8. Sample splitting solves what problem?

9. Susan Athey tells us that *high dimensional statistics* and *machine learning* are the same thing. The two phrases name a set of methods that are particularly useful in which circumstances?

10. Russ Roberts was once asked how many jobs NAFTA (the North American Free Trade Agreement) had created. How did he answer and how did his interrogator respond to his answer?

11. Susan Athey points out that progress in the technology industry is most often not the consequence of brilliant insights, revolutionary inventions, or great leaps forward. Most progress is instead incremental—technology advances in small steps.

    For example, Google runs 10,000 A/B tests each year. What does that mean? How does it work?

12. Cathy O'Neil suggests that the way courts predict recidivism and use the predictions could increase the phenonmenon (recidivism) that we want to reduce. How?

13. Cathy O'Neil talks about proxy variables. In particular, she mentions proxies to ethnic origin. People who will not use ethnic origin as a factor in their models are using variables whose values strongly correspond to ethnic origin. In this way, they avoid using ethnic origin directly but wind up using it indirectly.

    What are some examples of these proxy variables?

14. A computer program that evaluates candidates for employment at an engineering firm might discriminate against a group if the authors of that program use a knowledge of who performed best for the company in the past to build their system.

    Why is this so?

15. Susan Athey argues that the frequency and number of experiments in the technology industry leads to a bias. What kind of bias is the A/B testing producing?

16. In his conversation with Susan Athey, Russ Roberts compares the methods of economists in universities with the A/B testing in industry. What are some of the differences that he sees?

17. Cathy O'Neil criticizes current methods of working with big data for assuming that "the truth is embedded in historical practices."

    What is her concern?

18. Cathy O'Neil explains why evaluating the performance of teachers by looking at the performance of their students is a poor idea. She goes on to highlight shortcomings of value-added models. In what way are these measures of the performance of teachers flawed?

19. Cathy O'Neil criticizes what she calls "predatory advertising." What meaning is she giving to that phrase? What are some of the examples does she offer?

20. Russ Roberts asked Cathy O'Neil what she would like data scientists to do to avoid the kinds of problems that she described in her interview. What steps did she recommend?

21. What are three reasons an organization might consider alternatives to a relational database?

22. The document model organizes data in a way that resembles which popular format for the exchange of data?

23. Developers who use the relational model will typically define many tables. Developers who use the document model will not. What advantages follow from the use of the document model?

24. What does it mean to say that schema in the document model are dynamic?

25. How might the promises that a non-relational system makes with respect to consistency of data differ from the promises that relational systems make?

26. How do the query languages of NoSQL databases compare to the query languages of relational databases?

27. What is MongoDB's stance on the usefulness of relational methods?

28. MongoDB offers a variety of products. Distinguish between MongoDB Enterprise Advanced and MongoDB Atlas.

29. What is an advantage of a RESTful API? a disadvantage?

30. The white paper from MongoDB describes drivers. In this context, a "driver" is software that makes a connection between a client's application and the MongoDB database management program. MongoDB's drivers are "idiomatic." What does that mean?