Collecting Big Data

Core Concepts:

- Big data health research typically involves collecting large amounts of data that include many potential variables without identifying which might be used as dependent or independent variables.
- Big data health research involves using electronic sources of information to create a huge data base containing many different types of data.
- The science of big data research is expanding to include new technologies to collect data and new ways to link existing data bases.
- With such a huge amount of data potentially available, it is critical that consideration be given to protecting individuals' rights to privacy.

Class Time Required:

1 forty-minute class period

Teacher Provides:

- 1 copy of student handout entitled **Collecting Big Data** for each student
- Approximately 40 sticky notes (3"X 3") for each team. *Note: It is best to get the "super sticky" notes. Collect and save unused sticky notes at the end of the activity.*
- 4 large sheets of poster paper, each labeled at the top as shown below:
    A. Electronic health records from doctors, hospitals, and insurance records
    B. Computer files, search engines used, and Apps on digital devices
    C. Online purchases, store reward cards, credit card payments
    D. Wearable devices, home security systems, or home control devices
- 4 different colors of water soluble markers—a different color for each team. At least one marker per team but one marker per student is ideal.

Teacher Resources:

- **20 Big Data Repositories You Should Check Out**
  http://www.datasciencecentral.com/profiles/blogs/20-free-big-data-sources-everyone-should-check-out

This lesson was developed with support from the National Institutes of Health under Award Number R25OD010494-03S1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Suggested Class Procedure:

*Some students need processing time to prepare for team brainstorming. Consider assigning a data source (A-D) to individual students. For homework, students should make a list of 10 different **kinds of information (data)** about people that might be available from that source. Encourage students to ask adults for ideas to include in the list.*

1. Consider using the introductory lesson entitled **Big Data: A Different Kind of Science** to provide background information.

2. Distribute a copy of **Collecting Big Data** to each student.

3. Divide class into 4 teams (A-D). Each team will brainstorm the types of data that could be collected from the data sources assigned to their team.
    A. Electronic health records from doctors, hospitals, and insurance records
    B. Computer files, search engines used, and Apps on digital devices
    C. Online purchases, store reward cards, credit card payments
    D. Wearable devices, home security systems, or home control devices

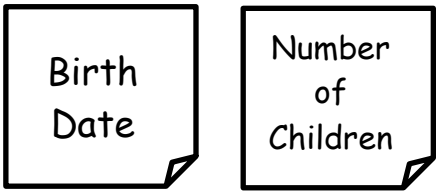
    Note: For large classes, you should have 2 of each team.

4. Distribute colored markers. Assign each team to use a different marker color. Each team should have at least one marker, but having more markers per team is desirable.

5. Distribute one poster (A-D) and approximately 40 sticky notes to each team.

6. Read the instructions for 1, 2, and 3 (on page 1) aloud to the class.

7. Allow 10 minutes for students work with their team to brainstorm data that could be obtained about people from the sources assigned to their team. Students should write their ideas on the sticky notes that you distributed to their team. They should write large and legibly. *Hint: Check as students work to be sure that they are writing types of data, not devices for obtaining data.*



8. Students attach their sticky notes to their poster paper.

9. Students observe the three posters done by other teams. Pass the posters around to other teams or post them on the wall and have each of the teams visit the poster "gallery." Each team should read what the previous teams have written and add **new** ideas written on sticky notes to the poster. Repeat passing the posters until all teams have had an opportunity to read and add new sticky notes to the four posters. *Note: You can tell if teams have added new sticky notes because there should be sticky notes with four different colors of ink on each poster at the end.*

10. Students should share and discuss the answers to questions 6-11.

11. *Be certain to SAVE THE POSTERS for use in the **Mining Big Data** lesson!*

Extension Activities (Optional):

- Students select the sticky notes for some of the types of data on their poster and add examples of more specific information. For example, gaming could include types of games, rate of win/loss, time spent on games, when games are played.

- Show one or both of the following videos to illustrate that big data can reveal surprising patterns/trends/connections:
    - **Data Scientists Find Connections Between Birth Month and Health** - http://newsroom.cumc.columbia.edu/blog/2015/06/08/data-scientists-find-connections-between-birth-month-and-health/
    - **See What Diseases You're at Risk For Based on Your Birth Month** – http://time.com/3913118/birth-month-disease-risk/

- Students read and discuss one (or both) of these articles:
    - **The Secretive World of Selling Data About You** – http://www.newsweek.com/secretive-world-selling-data-about-you-464789
    - **How Data Brokers Make Money Off Your Medical Records** – https://www.scientificamerican.com/article/how-data-brokers-make-money-off-your-medical-records/

- Students research and prepare a 10 minute lesson that explains how correlation and causation are different.

Collecting Big Data

Have you ever hit the "accept" button without carefully reading the terms or conditions for a program on your computer or an App on your cell phone or tablet? If so, you may already be a participant in a big data study that collects current data and future data related to your life and your health.

You may have granted big data researchers access to **electronic data sources** such as:

A. Health information from doctors, hospitals, and insurance records

B. Computer files, search engines used, and Apps on digital devices

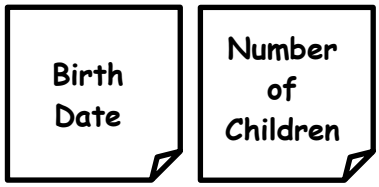
C. Online purchases, store reward cards, credit card payments

D. Wearable devices, home security systems, or home control devices

1. Your teacher provided a sheet of poster paper with the **electronic data sources** assigned to your team.

2. **What kinds of data (information about you or other people) might be stored in these electronic data sources?** You will have 10 minutes to brainstorm with your team to make a set of sticky notes. Each sticky note should list <u>one</u> possible type of data that could be obtained from the electronic data sources assigned to your team.

   Use the colored marker and write neatly.



3. Attach your sticky notes to your team's poster.

4. You will have 5 minutes to observe and add sticky notes to one of the posters made by another team. Read the title of the poster and read what previous teams have written on the sticky notes. Use your marker and sticky notes to add **<u>new</u>** ideas for types of data to the poster.

5. Repeat observing the posters until you have read and added **<u>new</u>** ideas for types of data to each of the posters.

Big data health research involves using electronic sources of information and new technologies to create huge data sets containing many different types of data. The data collected includes many potential **variables** without identifying which might be used as dependent or independent variables.

Not all data collected will be relevant to health issues. However, it is possible that big data analysis may result in the discovery of unanticipated connections, patterns or trends. Sometimes new variables are discovered that show trends or patterns related to health issues.

6. Give three examples of variables from the class sticky notes that are likely to be linked to human disease.

A **variable** is any factor, trait, or condition that can exist in differing amounts or types.

Student answers will vary.

7. Give three examples of variables from the class sticky notes that are unlikely to be related to human diseases.

Student answers will vary.

8. Should big data scientists limit the data they collect to variables (factors) that are clearly identified as related to health or disease? Explain why or why not.

No, because it is possible that apparently irrelevant data may in the discovery of unanticipated connections, patterns or trends.

The **right to privacy** refers to the concept that one's personal information (including health information) may be kept confidential and may be protected from becoming public knowledge. Some experts are concerned that big data research may allow unregulated access to personal information, and this can threaten individuals' rights to privacy. Other experts think that the potential benefits of big data research far outweigh the risks to individuals' privacy.

9. With your team, look at the posters and select and list four types of data that people might want to keep private.

Student answers will vary.

10. Do you think the benefits from big data research outweigh the risks to individuals' privacy?

Student answers will vary. This question is likely to trigger discussions with a wide range of perspectives.

11. What actions might be taken to reduce the risks that big data research will interfere with individuals' rights to privacy?

Student answers will vary. This question is likely to trigger an awareness of how much electronic information is available for people.