(http://csmatters.org) 4 - 1

0b100 - 0b1

Data Acquisition and Analysis

Unit 4. Data Acquisition

Revision Date: Jan 16, 2020
Duration: 1 50-minute session

Lesson Summary

Summary

In this lesson, students will learn how to acquire and analyze data to find answers to questions and solutions to problems. Students will consider whether or not the data they are presented with is necessarily valid, and research some of the various data sources online.

Outcome

- Students will explore how computation can be employed to help people process data and information to gain insight and knowledge.
- Students will learn how computation can be used to facilitate exploration and discovery when working with data.
- Students will consider what considerations and trade-offs arise in the computational manipulation of data.
- Students will identify evidence that digitally processed data can show a correlation between variables and that a correlation found in data does not necessarily indicate that a causal relationship exists.
- Students will explain how a single source does not contain the data needed to draw a conclusion. It may be necessary to combine data from a variety of sources to formulate a conclusion.
- Students will explore opportunities that large data sets provide for solving problems and creating knowledge.

Overview

Session 1

1. Getting Started (5 min) - Students journal on the importance of validating data
2. Discuss journal prompt (5 min)
3. Brainstorm types of online data (5 min)
4. Explore how meaning is created from data (10 min)
5. Work with data online (20 min)
6. Assign homework (5 min)

Session 2

1. Getting Started (5 min) - Students journal about what they learned the previous day
2. Analyzing Data (30 min) - Discuss correlation and causation; discuss and explore different types of data and analysis
3. Present homework findings (10 min)
4. Wrap Up - Journal (5 min)

Learning Objectives

CSP Objectives

- *EU DAT-1 - The way a computer represents data internally is different from the way the data is interpreted and displayed for the user. Programs are used to translate data into a representation more easily understood by people.*
  - LO DAT-1.A - Explain how data can be represented using bits.
  - LO DAT-1.D - Compare data compression algorithms to determine which is best in a particular context.
- *EU DAT-2 - Programs can be used to process data, which allows users to discover information and create new knowledge.*
  - LO DAT-2.A - Describe what information can be extracted from data.
  - LO DAT-2.C: - Identify the challenges associated with processing data.
  - LO DAT-2.D - Extract information from data using a program.
  - LO DAT-2.E - Explain how programs can be used to gain insight and knowledge from data.
- *EU AAP-2 - The way statements are sequenced and combined in a program determines the computed result. Programs incorporate iteration and selection constructs to represent repetition and make decisions to handle varied input values.*
  - LO AAP-2.A - Express an algorithm that uses sequencing without using a programming language.
  - LO AAP-2.L - Compare multiple algorithms to determine if they yield the same side effect or result.

Key Concepts

Students will be able to acquire data and analyze it to find answers to a specific question or solutions for a specific problem.

Essential Questions

- What opportunities do large data sets provide for solving problems and creating knowledge?

Teacher Resources

Student computer usage for this lesson is: **required**

Student computer usage for second lesson is: **optional**

Teacher Resources

In the Lesson Resources Folder

- PowerPoints: "Finding Data" and "Finding and Analyzing Data"
- Session 1 Homework: "Homework Unit 4 Lesson 1"

Webpages Session 1

- Rebecca Dovi's Blog: http://supercomputerscience.blogspot.com/2013/03/data-unit-day-one.html (http://supercomputerscience.blogspot.com/2013/03/data-unit-day-one.html)
- Bytes Review: http://highscalability.com/blog/2012/9/11/how-big-is-a-petabyte-exabyte-zettabyte-or-a-yottabyte.html (http://highscalability.com/blog/2012/9/11/how-big-is-a-petabyte-exabyte-zettabyte-or-a-yottabyte.html)
- TV Resolutions: http://www.rtings.com/info/what-is-the-resolution (http://www.rtings.com/info/what-is-the-resolution)
- New 3D Projector: https://www.amctheatres.com/sony4k (https://www.amctheatres.com/sony4k)
- Wolfram Alpha: https://www.wolframalpha.com/tour/what-is-wolframalpha.html (https://www.wolframalpha.com/tour/what-is-wolframalpha.html)

Webpages Session 2

- Strange Corrolations: http://www.tylervigen.com/spurious-correlations (http://www.tylervigen.com/spurious-correlations)
- What is a Data Scientist? https://www.youtube.com/watch?v=9PIqjaXJo7M (https://www.youtube.com/watch?v=9PIqjaXJo7M)
- What does a Data Scientist Do? https://www.youtube.com/watch?v=vowXaEDh1uk

(https://www.youtube.com/watch?v=vowXaEDh1uk)

Lesson Plan

Session 1

For this session, use the presentation "Finding Data" in the Lesson Resources Folder.

Getting Started (5 min)

Given this data: [slide 1]

A blood drive at the local high school reveals that 20% of the students were HIV positive.

Journal on these questions:

- What is your immediate reaction?
- What questions do you have?

Activities (40 minutes)

Activity 1 (5 min) - Discuss the journal prompt

Lead the students in discussion using the bullets below and slide 2 of the PowerPoint as guidance. Students should talk about WHY they assumed the data was true or were uncomfortable questioning the truth of the data.

- Two common reasons:
    1. Because their teacher told them.
    2. It had a % sign, so it looked authoritative. We all do that sometimes.
- From Rebecca Dovi's Blog: http://supercomputerscience.blogspot.com/2013/03/data-unit-day-one.html (http://supercomputerscience.blogspot.com/2013/03/data-unit-day-one.html)
- This discussion should reinforce that data online is not necessarily complete, up to date, or even valid.

Activity 2 (5 min) - Brainstorm: What kinds of data can be found online?

Part 1 - Discussion

Data comes from many places and takes many forms [slide 3]

- Have students discuss: How do business, personal, government and devices create and use data?

- Do computers perceive and store data in the same way that humans do?
- Additional research is needed to understand the exact nature of the relationship.

Part 2 - Brainstorm

Brainstorm as a class: what kinds of data are generated? Possible answers:

- video: movies, webcam images, CCTV, youTube, Netflix, Facebook, etc.
- pictures: maps, Instagram, photos, cartoons, drawings, …. everything!
- words: books, articles, news, stories, blogs, Facebook
- numbers: facts, financial transactions, scientific data
- sound: music, speech
- behavior tracking: GPS, click behavior, search history
- IMPORTANT POINT: Compters see and record digital data which is only an approximation of the real world. The sample rate determines the accuracy of the digital approximation. The real world is analog. Analog data have values that change smoothly, rather than in discrete intervals, over time. Some examples of analog data include pitch and volume of music, colors of a painting, or position of a sprinter during a race.

Activity 3: How is meaning created from data? (10 minutes)

1. Look at some data gathered about selfies from different cities around the world. [slide 4]
    - Main ideas:
        - You have to gather the data and analyze it to create meaning.
        - Creating meaning from pictures still takes some human interpretation.
        - Digitally processed data can show a correlation between variables and a correlation found in data does not necessarily indicate that a causal relationship exists.
        - A single source does not contain the data needed to draw a conclusion. It may be necessary to combine data from a variety of sources to formulate a conclusion.
    - Prompt students to come to a conclusion about the graphed data on the page.
    - Question for discussion: How large of a sample is needed to draw a conclusion?
2. Quick review: Make the point that there is a LOT of data even in a single picture. [slide 5]
    1. Define these and put them in order. Use this webpage to review bytes: http://highscalability.com/blog/2012/9/11/how-big-is-a-petabyte-exabyte-zettabyte-or-a-yottabyte.html (http://highscalability.com/blog/2012/9/11/how-big-is-a-petabyte-exabyte-zettabyte-or-a-yottabyte.html)
        - MB, bit, TB, ZB, byte, GB, pixel (one dot of color on the screen), KB, PB
    2. Look at the photo on slide 5.
        1. 365 gigapixels is 365 billion pixels, if the picture is a square, then it is 604,152 pixels on each side (too big to fit on any HDTV screen)
        2. http://www.rtings.com/info/what-is-the-resolution (http://www.rtings.com/info/what-is-the-resolution) A 4K super high

resolution TV is only about 3,000 X 2,000 pixels. Even a movie screen can't show all of the detail!

3. https://www.amctheatres.com/sony4k (https://www.amctheatres.com/sony4k), you can only look at it one part at a time.

3. Preview Wolfram Alpha, an engine for providing knowledge from data.
   - Show the introductory video: https://www.wolframalpha.com/tour/what-is-wolframalpha.html (https://www.wolframalpha.com/tour/what-is-wolframalpha.html) (1:18) [slide 6]
   - Identify ways that patterns can emerge when data are transformed using programs.
   - Experiment to demonstrate how insight and knowledge can be obtained from translating and transforming digitally represented information.
   - Students will explore these sites in the next activity.
4. Point out that there are processes that can be used to extract or modify information from data in both bebeficial and harmful ways. These processes include the following:
   - machine learning and data mining
   - transforming every element of a data set, such as doubling every element in a list, or extracting the parent's email from every student record
   - filtering a data set, such as keeping only the positive numbers from a list, or keeping only students who signed up for band from a record of all the student
   - combining or comparing data in some way, such as adding up a list of numbers, or finding the student who has the highest GPA
   - visualizing a data set through a chart, graph, or other visual representation

Activity 4 (20 min) - Work with some data online

1. Students should complete the Data Search and Analysis Handout. [slide 7]
   - Depending on how much time you have, you can pair students and assign even/odd questions or chunks of questions to different groups, or have each student research on their own.
2. If there's time in class, try to go over results and compare (especially the first 5) to see if people got similar answers. Why or why not? [slide 8]

Assign Homework (5 minutes)

Give students the worksheet: Homework Unit 4 Lesson 1.

There are 10 videos to choose from, each 10-15 minutes long. Either allow students to self-select, or assign them a particular video. Students should watch the video and answer the questions on the worksheet. This is an opportunity to discuss plagiarism: students are expected to watch the video and write from their own experience.

Session 2

For this session, use the presentation: Finding and Analyzing Data from the Lesson Resources Folder

Getting Started (5 min)

Students should journal on the following: Describe at least 2 ways that we create meaning out of data. [slide 1]

- Possible answers: graph it, total it, average it, find min and max, map it, compare it to other data, find trends, generate predictions (like weather), draw conclusions (facial recognition, emotions, voice inflection), diagnose diseases, discover new stars, etc.

Activities (40 min)

Activity 1 (35 min): Analyzing Data

Part 1: Correlation vs. Causation

1. Look at slide 2 from the PowerPoint. Creating meaning from data can be misleading.
2. Point out that the graph shows a direct relationship between the number of divorces in Maine and the amount of margarine that is purchased. When one goes up, the other does too, and vice versa. Is this a causal relationship?
    - Show some examples from the Tyler Vigen website http://www.tylervigen.com/spurious-correlations (http://www.tylervigen.com/spurious-correlations) . It has many examples of data connections that may be statistically valid but don't make sense. The site was created to point out how comparisons due to data correlation are often not valid.

Part 2: Data Science

1. What does a data scientist do? [slide 3] Show the two videos and discuss.
    - What is a data scientist? (0:54) https://www.youtube.com/watch?v=9PIqjaXJo7M (https://www.youtube.com/watch?v=9PIqjaXJo7M)
    - What does a data scientist do? (0:39) https://www.youtube.com/watch?v=vowXaEDh1uk (https://www.youtube.com/watch?v=vowXaEDh1uk)
    - Data science is a relatively new field combining computer science with statistical data analysis and processing data to create meaning.
2. Say: Tricks to analyzing big data:
    a. Knowing what data to use, and what to disregard.
    b. Dealing with non uniform data - or data enterred in a variety of formats
    c. Knowing how to clean data - to make the data conform to a format without changing its meaning.
    d. Knowing how to use data filtering tools to find information, recognize patterns and predict trends.

3. Look at 3 false assumptions about big data [slide 4]:
    a. It's complete and accurate
    b. It tells the whole story
    c. Bigger is better
4. What considerations and tradeoffs arise in the computational manipulation of data? [slide 4]
    a. How do you account for missing data?
    b. How do you certify your sources?
    c. How do you decide which data to include and which to exclude?
    d. How much data is enough? The size of a data set affects the amount of information that can be extracted from it.
    e. Are your processing algorithms accurate?
5. What is some of the data needed to successfully fly a space mission? (Possible answer: Knowing all about the spacecraft: speed, direction, amount of fuel/oxygen left.) The same problems that applied to early space missions are some of the same problems faced in dealing with big data.
    a. You need to decide which factors to include in your calculations, and which to exclude.
    b. You need to decide when to make an assumption for missing data or when to estimate.
    c. In writing a program for an early space flight there are many unknown factors using a space craft that has never flown before.
    d. It's usually impossible to create a perfect algorithm that can take into account every possibility, so how do you allow for errors and changes?
6. What are some of the calculations needed? (Possible answers: how much fuel to release and with which engines.)
    - They had to run many simulations first to see what would happen under various circumstances.
7. See if anybody knows how NetFlix, movie makers, or Amazon use data about their customers to be more successful. [slide 5] http://www.smartdatacollective.com/bernardmarr/312146/big-data-how-netflix-uses-it-drive-business-success (http://www.smartdatacollective.com/bernardmarr/312146/big-data-how-netflix-uses-it-drive-business-success) and http://www.fastcompany.com/3024655/pitch-perfect-and-how-analytics-are-transforming-movie-marketing (http://www.fastcompany.com/3024655/pitch-perfect-and-how-analytics-are-transforming-movie-marketing)

Businesses like Amazon and NetFlix learn the habits of different customers and make recommendations based on their previous choices and others who share similar characteristics (like Google ads).

See if anybody knows the story of Moneyball (based on a true story) of how a baseball team made decisions based on data analysis to become winners, https://en.wikipedia.org/wiki/Moneyball_(film) (https://en.wikipedia.org/wiki/Moneyball_(film)) and how Vivek Ranadivé--who knew little about basketball but owned a multi-million dollar computer processing company and knew how to choose and analyze data--coached his then twelve-year-old daughter's (http://www.newyorker.com/reporting/2009/05/11/090511fa_fact_gladwell) National Junior Championship basketball team to the national championship game. He relied upon his sporting knowledge of soccer and cricket paired with his analytic mindset, to create a system of play which allowed his relatively un-athletic team to excel. From the moment that he used intellect and his business experience to coach an inexperienced team to the championship game, the man who once thought basketball was "mindless" was hooked on the sport. http://www.newyorker.com/magazine/2009/05/11/how-david-beats-goliath (http://www.newyorker.com/magazine/2009/05/11/how-david-beats-goliath)

1. How is data analyzed? **Data analysis requires an algorithm, a plan to collect and process data. [slide 6]**
    1. Generate a discussion about what data is collected and how it is analyzed. What is a possible algorithm for making a decision about choosing what movies NetFlix might suggest for a customer?
    2. http://www.smartdatacollective.com/bernardmarr/312146/big-data-how-netflix-uses-it-drive-business-success (http://www.smartdatacollective.com/bernardmarr/312146/big-data-how-netflix-uses-it-drive-business-success)
       Brainstorm: what other data might they collect? (what's currently popular in that age group, demographic, etc.)
2. Choose one of the options and write an outline of an algorithm: choosing a movie to produce or a sports player to hire. [slide 7]
    1. Describe at least two calculations needed
    2. Describe some of the data you'd need to collect.
    3. Describe how the data sets needed could pose challenges regardless of size, such as:
        - the need to clean data
        - incomplete data
        - invalid data
        - the need to combine data sources

    Share and discuss.

Activity 2 (5 min):

Present homework from the previous day after watching TED talks on data. [slide 8]

Summarize all of the questions from the homework to be presented to the class and collect the written summaries to grade.

Journal (5 min)

In your writing journal, map out the steps to answer a specific question or find a solution to solve a specific problem using data.

Options for Differentiated Instruction

Extension Activities:

Data analysis activities from NOAA, NASA, and more! - http://climate-expeditions.org/educators/activities.html (http://climate-expeditions.org/educators/activities.html)

Differentiation Instruction:

What is data acquisition? - http://www.ni.com/data-acquisition/what-is/ (http://www.ni.com/data-acquisition/what-is/)

Data analysis and graphs (with Excel sample) - http://www.sciencebuddies.org/science-fair-projects/project_data_analysis.shtml (http://www.sciencebuddies.org/science-fair-projects/project_data_analysis.shtml)

Collecting and analyzing data - http://ctb.ku.edu/en/table-of-contents/evaluate/evaluate-community-interventions/collect-analyze-data/main (http://ctb.ku.edu/en/table-of-contents/evaluate/evaluate-community-interventions/collect-analyze-data/main)

Using Excel for Handling, Graphing, and Analyzing Scientific Data: A Resource for Science and Mathematics Students - http://academic.pgcc.edu/psc/Excel_booklet.pdf (http://academic.pgcc.edu/psc/Excel_booklet.pdf)

Evidence of Learning

Formative Assessment

Journal day 1:

Given this fictitious data:

A blood drive at the local high school reveals that 20% of the students were HIV positive.

- What is your immediate reaction?
- What questions do you have?

Journal day 2: Describe at least 2 ways that we create meaning out of data.

Homework: Feedback from a TED video on big data

Summative Assessment

Students complete the Data Search and Analysis student activity.

Write an outline of an algorithm to make a data-based decision about what movie to produce or what sports team member to hire.

(http://www.umbc.edu/)


(http://www.umd.edu/)


(http://www.nsf.gov/)

Authored by: CS Matters in Maryland
Website: csmatters.org (http://csmatters.org)
Email: csmattersinmaryland@gmail.com (mailto:csmattersinmaryland@gmail.com)