Unit 9: Data, Data, Everywhere

Learning Objectives

- 1: The student can use computing tools and techniques to create artifacts.
- 2: The student can collaborate in the creation of computational artifacts.
- 10: The student can use models and simulations to raise and answer questions.
- The student can use computers to process information to gain insight and knowledge.
- The student can communicate how computer programs are used to process information to gain insight and knowledge.
- The student can use computing to facilitate exploration and the discovery of connections in information.
- The student can use large datasets to explore and discover information and knowledge.
- The student can analyze the considerations involved in the computational manipulation of information.

Readings/Lectures

External Resources

- The Beauty of Data Visualization (http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization.html) - TED Talk
- Data Explosion creates Revolution (http://www.sfgate.com/technology/dotcommentary/article/Web-2-0-Summit-Data-explosion-creates-revolution-2326463.php) - SFGate
- Big data and society (http://www.guardian.co.uk/media-network/media-network-blog/2013/apr/12/big-data-privacy-economist) - - interview with Kenneth Cukier, The Economist
- Mike2.0 definition of Big Data (http://mike2.openmethodology.org/wiki/Big_Data_Definition)
- IBM’s definition of Big Data (http://www-01.ibm.com/software/data/bigdata/)
- The Pandora Music Genome Project (http://www.pandora.com/about/mgp)
- GCF Excel Resources (http://www.gcflearnfree.org/excel2010)

Labs/Exercises

- Lab 9.01: Top Songs (/bjc-course/curriculum/09-data/labs/01-top-songs)
- Portfolio Project 9.02: Data Portfolio Project (/bjc-course/curriculum/09-data/labs/02-data-portfolio-project)

Resources for Project

- Data Sets (http://archive.ics.uci.edu/ml/datasets.html)
- Amazon Data (https://aws.amazon.com/datasets)
- Census Data (https://www.census.gov/main/www/cen2000.html)
- KDnuggets (http://www.kdnuggets.com/datasets/)

Big data and society - interview with Kenneth Cukier, The Economist

Big data will transform the world, but issues around privacy and propensity need to be resolved, says Kenneth Cukier

- Share
- Tweet this
- ✉Email

Big data is something very new and will touch all aspects of society. Photograph: KC

Can you tell us a little bit about your role as data editor for The Economist?

The position is a new one, but it stems from the idea that the new wealth of data and new tools to process and visualise it means that we as journalists can tell stories in new ways. Instead of basing stories on a string of anecdotes with a single statistic dropped in, we can invert the form and make the data the story, while just using a judicious anecdote to illustrate the information. In this respect, The Economist can be said to have been practising data journalism for 170 years; we're known for our data-driven content like the Big Mac index to compare currencies.

How do you characterise big data?

There is no concrete definition and that is probably a good thing since to define is also to limit. But it's not woolly either. We can understand big data by its features, and the central one is this: we can do things with a huge corpus of data that we are unable to do with smaller amounts, to extract new insights and create new sources of value. This encompasses things like machine learning, in which we have self-driving cars and decent language translation. This is not because we have faster chips or cleverer algorithms, but because we have more data (and the tools to process it at a vast and affordable scale).

In what ways is big data transforming the world?

It is on track to touch all aspects of society. We will go from a world that we understand by experiencing it on an individual level, to one we comprehend on a more universal level. By that, I mean that we tend to base decisions on small amounts of data that are usually just a simulacrum of the complex reality we are trying to deal with, and tailored to our cognitive limitations to make sense of it. Tomorrow, we will use big data to surpass our faith in our individual powers and instead place trust in the data (though not blind trust).

Take medicine. Today, doctors make diagnoses based on their judgement. Sounds reasonable? In time, this will probably be considered as barbaric as bloodletting. Why not use big data? We could enshrine the experience of all doctors, and of hundreds of millions of patients over decades, to identify the best treatments to achieve the best outcomes and spot hidden adverse drug side effects. After all, the sum of all medical knowledge isn't in the possession of any single physician. But if we aggregate vast troves of healthcare information, we may learn what works best, just as Amazon recommends books not based on the inklings of a literary critic but from correlating sales data. This will mark a revolution in how society uses information.

Concerns have been raised regarding the privacy implications of big data. What is your view of this?

Privacy is a big problem today and it will be a bigger problem tomorrow. We need to improve the legal regime to govern privacy to move beyond the system of notice-and-consent (that is, companies inform users what data they collect and how it's used, and people give their okay). In reality it means that people tick a box agreeing to 60 pages of legal jargon with nary a glance. Instead, we need to consider the use and misuse of the data, not just the collection. We need to focus on the area of harm and not just on the inert, potential harm.

Are there any other challenges associated with big data?

While privacy is a problem, a newer issue is 'propensity'. This refers to the idea that algorithms may be making predictions about what we are likely to do, and we may find that we're penalised before we've actually committed the infraction. So big data may assign a 95% likelihood that a certain person will shoplift, or default on a loan, or fail to survive a surgical operation. We'll need to sanctify human agency and freewill. At the same time, we'll need a new class of professional the "algorithmists" to review big data analyses and provide society the same transparency and accountability that we have today, to ensure that big data is not a black box that obviates the public interests.

Finally – what will be your message to the audience at Big Data Week?

Big data is something very new and will touch all aspects of society. Whilst it helps to have the technical skills, the key to success will be in applying one's imagination, creativity, intellectual ambition and risk-taking – characteristics that are intrinsically human and cannot be reduced to number crunching.

Kenneth Cukier will be speaking at Big Data Week London – Putting Data to Work, 25 April, London — visit their website for further details

Get more articles like this sent direct to your inbox by signing up for free membership to the Guardian Media Network – this content is brought to you by Guardian Professional.

Curriculum (/bjc-course/curriculum) / Unit 9 (/bjc-course/curriculum/09-data) /
Lab 1 (/bjc-course/curriculum/09-data/labs/01-top-songs) /

Lab: Top Songs

Learning Objective 13: The student can use large datasets to explore and discover information and knowledge. P3

Evidence for Learning Objective 13: Student work is characterized by:

- Use of large datasets to extract information and knowledge.
- Explanation of how large datasets can facilitate exploration and discovery.

Rolling Stone created a list of the top 500 songs of all time. We've created a .csv file of these files. This file, as well as others you will need, is attached.

Task 1

Determine the top 10 artists based on how many songs the artist has in the top 100. You have a file of the top 100 artists, associated ranking, and song that received that ranking. The artist with the most songs in the top 100 is the top artist. Look at the top 100 artists and songs. Attempt to identify the top artist by looking over the document.

Next, brainstorm how you would approach this computationally.

Task 2

Using a computer, identify the top 10 artists from the top 500. The .csv file (top500.csv) is also available for you. You can do this anyway you want. If it helps we've uploaded the top 500 songs to the IBM Many Eyes website (but you absolutely don't have to to use this particular tool). That website lives here:

Top 500 songs in many eyes:

http://www-958.ibm.com/software/data/cognos/manyeyes/datasets/top-500-songs/versions/1

You can also find these songs by searching data sets on the Many Eyes site for 'song'.

Task 3

Does the approach you developed in the last task work with the data that's the top 1000 rock and roll songs (e.g., as determined by KZOK 102.5)? This data is also attached for you. Spend a couple of minutes looking at it.

Consider the problem of treating all songs equally: the first song and the 500th song count equally in weighting the "top" group by the number of songs in the top 500 (or 1000). What alternatives can you develop to this ranking in determining the "top" group? How would your method of finding the top group change given the alternative you develop?

To Turn In Your Work

Add all of your answers and comments to a document to upload. Make sure to tell how you determined your answers.

This is a test/project grade.

Learning Objectives

The Data Portfolio Task addresses the following CS Principles Learning Objectives (LOs):

- 1: The student can use computing tools and techniques to create artifacts.
- 2: The student can collaborate in the creation of computational artifacts.
- 4: The student can use computing tools and techniques for creative expression.
- 10: The student can use models and simulations to raise and answer questions.

Task

This portfolio entry includes work on a team (usually 2 members) and as an individual.You may work with another student in class if possible.

Collaborative portion:

Identify and describe a significant area in which you will conduct an investigation to gain insight and knowledge from publicly available data. * The area can be an academic topic you are studying or an area of interest.

Develop a set of 3-5 questions that will be the focus of the investigation, find one or more large data sets that will allow you to obtain answers to your questions, and then apply computational tools and techniques to answer your questions, e.g., by finding patterns in the data, by transforming or translating the data, or by finding connections between the data and other sources of knowledge.

- For example, you could find data in Excel spreadsheet format to analyze by using a pivot table or graph/chart.
- Data sets can be tables or spreadsheets. Professional organizations (associations) and government organizations are good places to look for data.

Find data (quantitative data) on the Internet to answer your questions * For example, a data set that you can pull into Excel (or another program).

Produce an artifact (e.g., a report, video, presentation, visualization, or combinations of these) that will allow someone else to understand your investigation, including both the set of questions you pose and the answers you obtain.

The artifact should

- describe the computational tools and techniques you use,
- explanations of why you chose them, and
- why they are appropriate for your investigation.

Individual portion

Create a document in which you

- make it clear that you can use the computational tools and techniques described in your collaborative artifact to verify the answers that are provided there.
- provide enough details about the tools and techniques and their use with the data and other sources of knowledge so that a skilled reader could verify and reproduce your answers using the same data set and other

sources of knowledge.In other words, you should tell me how you analyzed your data so that I could re-create it. This is similar to a lab report in science.

You must include a reflection that describes and analyzes the collaborative aspects of the investigation.

Prepare and submit the following

A collaboratively developed artifact that communicates a detailed description of your group's investigation, the questions, and your collective findings.

- You may use any form of digital artifact (e.g., a report, video, presentation, visualization, or combinations of these) that allows you to best communicate your investigation and findings.
- You and your partner will each submit the same artifact.
- Submit a Works Cited listing all information sources. (MLA format)
- Submit your data file/s.

An individually written document that addresses the investigation.

- Each group member must write her/his own individual document.
- In writing the individual document you must adhere to the Task description above and the Requirements description below in supplying details of your investigation.

Suggestions

- If you are interested in researching different occupations try http://www.bls.gov/oes/current/oes_stru.htm
- If you are interested in researching different colleges try http://www.cfnc.org/index.jsp
- There are associations associated with every field, look those up for available data.
- Need help with creating graphs in Excel, try http://www.gcflearnfree.org/excel2010/17

Requirements

- You must work with a partner on this project.
- You will collaboratively conduct the investigation: formulating the area to be investigated, developing the questions, finding data set(s) to address those questions, and creating the artifact that allows someone else to understand your investigation.

Rubric

Preparation (20 points)

- Area identified
- 3-5 questions
- Appropriate data set
- Computational tools/techniques chosen

Artifact (40 points)

- Communicates detailed description of investigation
- Questions & answers explained
- Describes tools/techniques used and why chosen/appropriate

Individual (30 points)

- Clear you can use tools/techniques
- Details so that investigation could be duplicated

- Reflection

Works Cited (10 points)

- Appropriate sources
- MLA format