 Datasets in science

This document is a resource to support the Stage 6 Science Extension course[[1]](#footnote-1).

[Preface 2](#_Toc2798094)

[Acknowledgements 2](#_Toc2798095)

[Contact 2](#_Toc2798096)

[The need for data in scientific investigations 3](#_Toc2798097)

[Large datasets 3](#_Toc2798098)

[Obtaining data 4](#_Toc2798099)

[Student-generated primary datasets 4](#_Toc2798100)

[Accessing data through government and scientific organisations 4](#_Toc2798101)

[Other data repositories 5](#_Toc2798102)

[Universities and research institutes 6](#_Toc2798103)

[Evaluating secondary datasets 6](#_Toc2798104)

[New perspectives in scientific data collection 6](#_Toc2798105)

[Why do scientific organisations share their data? 7](#_Toc2798106)

[Learning about large data in science 8](#_Toc2798107)

[Discussion 9](#_Toc2798108)

[Web Resource 9](#_Toc2798109)

Preface

This resource explores the use of datasets. A dataset is a collection of observations or measurements. Modern scientific research often involves the generation and analysis of datasets. In the Science Extension course[[2]](#footnote-2), students are required to explore the impact of large datasets (module 2). Scientific datasets may be produced by researchers, research organisations (for example the CSIRO and the Australian Bureau of Statistics) and citizen science groups. By analysing datasets, students can deepen their understanding of how scientists obtain evidence for explaining natural phenomena, as well as trends and patterns inherent in complex systems. Teachers can use the resources described in this document to access and analyse datasets in the classroom.

Acknowledgements

The Learning and Teaching Directorate at the NSW Department of Education developed this resource for use by science teachers. The department acknowledges the contributions of the following teachers in preparing this document:

* Carl Masens, Girraween High School
* Andrew Burns, Arthur Philip High School
* Sham Nair, Learning and Teaching Directorate

The department also acknowledges the following staff for critically reviewing the document:

* Chris Bormann, Science Curriculum Support Officer, Coffs Harbour
* Marc Keats, Science Curriculum Support Officer, Warilla
* Penelope Gill, Science Project Officer, Sydney

Contact

For more information, contact

Sham Nair
Science Advisor, 7-12
NSW Department of Education
sham.nair@det.nsw.edu.au

February 2019

The need for data in scientific investigations

Science is data-driven. Hypotheses are accepted or rejected depending on whether there is data support them. Therefore, when planning investigations, scientists must ensure that they can generate the necessary data or have access to previously-generated data to test their hypotheses.

Large datasets

Technological developments have allowed scientists to generate large volumes of data quickly. For example, the traditional method of sequencing DNA using radioisotopes would produce several hundred base pairs of DNA sequence information in a week. However, automated sequencing methods have reduced the time required to obtain similar information to a few hours. Modern DNA sequencing machines can generate 10-20 gigabase pairs of sequence information in a single experiment. Technological advances have reduced the cost of generating data. Traditional DNA sequencing methods cost about US$2,400 to produce 1 million base pairs of DNA sequence information, while modern methods cost between 5-8 cents to generate the same information[[3]](#footnote-3),[[4]](#footnote-4). According to the [National Human Genome Research Institute](http://www.genome.gov/sequencingcostsdata), sequencing the human genome in the first Human Genome Project cost about US$300 million and took 13 years to complete. In comparison, a similar project today costs less than US$1,500 and takes one to two days to complete.

Electronic sensors and data collectors enable scientists to take measurements very frequently or over a large time scale. Data from individual sensors can contain thousands or millions of discrete measurements. The Large Hadron Collider detectors take [40 million electronic photographs per second](https://www.elsevier.com/connect/how-big-data-advances-physics) to document the trajectory of subatomic particles after high energy collisions. This translates to about 40 terabytes of data per second. High-powered computer networks use statistical and other analytical algorithms to analyse such data.

The size of large datasets improves the statistical reliability of the data contained in them. Analytical tools can identify anomalies in large datasets and quantify errors accurately.

Large datasets allow scientists to investigate complex questions. For example, classical methods in genetics allowed scientists to investigate only one or a few genes at a time. Although those were powerful approaches, they were limited in the types of questions that could be answered. Whole genome sequence data allows scientists to examine the effects of perturbations on many genes simultaneously. [Genome-wide association studies](https://www.genome.gov/20019523/genomewide-association-studies-fact-sheet/) can be used to identify possible markers of diseases that cannot be identified using single-gene experiments.

Obtaining data

Students can generate data through well-designed, controlled experiments. Such data is referred to as primary data. Secondary data is produced by researchers other than the primary investigator or user. Secondary data is usually stored in databases and other repositories. Some of these are publically-accessible, while others are restricted to members of the research community or other organisations.

Student-generated primary datasets

The availability of data loggers and sensors means that students may be able to use equipment available at the school to generate large datasets. Using data from primary investigations is undoubtedly within the scope of the Science Extension course, as well as in the other Stage 6 science courses. However, before choosing to generate data, students need to consider the following:

* Will the dataset be large enough to be statistically reliable? Will it be large enough so that computational analysis can reveal patterns, trends and associations?
* Will the sensors and data logging equipment available at the school allow appropriate data to be collected to test a hypothesis? This may limit the types of data that students can collect.
* Will students be able to conduct a statistical analysis of the data? Students will need to export the data in a form that will be recognised by the software used for the statistical analysis.

Accessing data through government and scientific organisations

Australian Bureau of Statistics (ABS)

The ABS [Table Builder](http://www.abs.gov.au/websitedbs/D3310114.nsf/home/About%2BTableBuilder) provides free access to a wide range of datasets, including those gathered from several Australian censuses. To download data, do the following:

1. Select ‘Census TableBuilder Guest’ and click ‘continue’.
2. Accept the terms and conditions.
3. Select a dataset (click on the arrows to open the subfolders), double-click a pre-defined table or create a new table.
4. Once the table is displayed, select the ‘download table go’ button (at the top of the webpage).
5. Note that the Guest account only provides access to limited tables (for example - census of population and housing). To access other datasets, users have to create an account (free).

Another option is to access other datasets via the statistics portal of the ABS website. To do this:

1. Go to [ABS](http://www.abs.gov.au/) and click ‘statistics’.
2. Select a relevant dataset (multiple clicks may be needed).
3. Click on the ‘downloads’ tab and then on the relevant Excel file icon.

The Bureau of Meteorology (BoM)

Their [Climate Data Online](http://www.bom.gov.au/climate/data/?ref=ftr) allows access to a range of measurements over time from sites across Australia.

1. Fill in the dialogue box to specify the data for downloading.
2. Choose the type of data from the ‘Data About’ drop-down menu (weather & climate, rainfall, temperature, solar exposure).
3. Select ‘Type of Data’ (for example - daily observations).
4. Enter the information about weather station in the area of interest (type in the name of a city or centre (for example – Sydney, Sydney NSW, 066062 Sydney, NSW), or select an area from the map).
5. Type in the station number (for example - 066062), then click ‘Get Data’.
6. A new window opens up, and the data files can be downloaded as a CSV file (may be opened in Excel).

Other data repositories

* [United States Geological Survey (USGS)](https://www.usgs.gov/): Their [Science Data Catalog](https://data.usgs.gov/datacatalog/#fq=dataType%3A(collection%20OR%20non-collection)&q=*%3A*) allows students to select and download monitoring data, including seismic data. Some of their datasets include latitude and longitude data for events, thus enabling the data to be viewed using [Google MyMaps](https://www.google.com/maps/d/).
* [GitHub](https://github.com/awesomedata/awesome-public-datasets) has links to organisations that provide publicly- accessible datasets. Users can access a wide variety of datasets from scientific and non-scientific disciplines.
* [Australian National Data Service](https://www.ands.org.au/) – a repository of research data generated by Australian researchers that is maintained by a partnership of various Australian institutions. It contains links to many different types of research data (from scientific and non-scientific disciplines.
* [Data.gov.au](https://data.gov.au/) contains Australian open government data (federal, state and local government agencies). Contains both scientific and non-scientific data sets.
* [The Bureau of Meteorology](http://www.bom.gov.au/climate/data-services/) contains datasets on various meteorology-related topics.

Universities and research institutes

Researchers may be able to provide relevant data for analysis by students. Such researchers may be identified through internet searches or through the papers they have published. The contributions of the researchers must be acknowledged in students’ research reports.

Evaluating secondary datasets

Before analysing secondary datasets, students should evaluate the data in them. The following points may be useful for students when evaluating datasets:

* Is the data relevant to the area of the investigation?
* Does the data contain measurements of the independent and the dependent variables?
* Is the data range broad enough to make meaningful generalisations?
* Is the uncertainty in the data sufficiently low to make meaningful conclusions?
* Is the data file in a format that can be imported into an accessible data analysis software?
* Are appropriate tools available for analysing the data?
* Has permission been sought (from those who generated/own the data)?

New perspectives in scientific data collection

Technological advances (automation, electronic sensors and computing ability) have led to an enormous increase in data generated by scientists. This allows scientists to gather data in ways and at locations that were impossible before. Such advances include:

* Increased frequency of data collection. Using computers to gather and store data has made it possible to increase the sample rate of measurements and increase the number of instruments that can be used to collect data. This is particularly useful in experiments where change occurs rapidly, such as collisions in particle accelerators and automobile crash tests.
* Collect live data. Wireless connection of sensors through Bluetooth and Wi-fi has meant that data can be directly streamed and monitored. The data may be sent in smaller files that need to be combined and processed so the whole data set can be analysed. Uses of streamed data include tracking wildlife using GPS sensors and measuring climatic data from different locations.
* Remote data collection. Remote sensing uses sensors to detect electromagnetic radiation emitted or reflected off an object. This allows scientists to make real-time observations over large areas. The surface of the Earth is monitored continuously from satellite-based sensors across a range of wavelengths. Applications of remote sensing include measuring ocean temperatures and tracking cyclones and bushfires.

Why do scientific organisations share their data?

Many researchers and scientific organisations share their data. Some scientific journals require scientists to upload their research data to publically-accessible repositories before publishing their manuscripts. Some research organisations are committed to making their data publically-accessible. For example, the European Council for Nuclear Research (CERN) has made over one petabyte of data available online through their [Open Data portal](http://opendata.cern.ch/) for others to analyse. [Genbank](https://www.ncbi.nlm.nih.gov/genbank/statistics/), which is a repository for DNA and RNA sequences, contains more than 200 million sequences from various organisms. The Human Genome Project, which was an international research effort to sequence the entire human genome, committed to making all genome sequencing data publicly available within 24 hours of the data being generated.

The reasons for making data publically available include transparency (providing the evidence for scientific conclusions), synergy (different groups working in similar research areas can combine their findings), collaboration (different groups can analyse the same data) and evidence-based decision making (to provide an evidence base for public policy development).

However, when data is released to the public, it should be appropriately curated. Additional information about the data (metadata) should be provided so that it may be contextualised (for example, a brief description of the experiment that generated the data, how the data was collected and analysed, any software that was used, including the settings). Improperly curated data can be misleading or may lead to incorrect conclusions.

By releasing primary research data, anyone with a suitable internet connection can download the information for further analyses and, perhaps, discover new findings. For example, a university student in Montreal, Canada [discovered a new exoplanet](https://www.eurekalert.org/pub_releases/2018-09/uom-ane090518.php) using the data from the Kepler Mission. Table 1 includes examples of large datasets described in the Science Extension syllabus.

Table 1: Examples of scientific experiments that have produced large datasets. The table also indicates some significant discoveries arising from those investigations. Teachers and students can access relevant publications and datasets using the links provided.

| Source of the data set | Discoveries | Relevant publication (s) | More information |
| --- | --- | --- | --- |
| LHC (Large Hadron Collider) | * Higgs boson
* New subatomic particles
* Supersymmetry
 | [The ATLAS Experiment at the CERN Large Hadron Collider](http://cds.cern.ch/record/1129811) | * [Information about the ATLAS experiment.](https://home.cern/science/experiments/atlas)
* [The ATLAS experiment datasets](http://opendata.cern.ch/search?page=1&size=20&collections=ATLAS-Derived-Datasets).
 |
| Kepler Mission | * Exoplanets
 | [Kepler Mission design, realized photometric performance, and early science](http://iopscience.iop.org/article/10.1088/2041-8205/713/2/L79/meta) | * Information about the Kepler Mission may be found at these websites:
	+ [Kepler Mission (Harvard)](https://www.cfa.harvard.edu/kepler/)
	+ [Kepler Mission (NASA)](https://www.nasa.gov/mission_pages/kepler/main/index.html)
* [The Kepler Mission datasets.](https://exoplanetarchive.ipac.caltech.edu/docs/data.html)
 |
| Human Genome Project | * Identifying the genetic basis of some diseases.
* A genetic blueprint for constructing every human cell.
* The genetic relationship between humans and other organisms.
 | [The Sequence of the Human Genome](http://science.sciencemag.org/content/291/5507/1304?casa_token=UaYfwyRKqTkAAAAA:NRcUDhJp_A0RDS2ofjJhKhZ6jdzuCQBjocwzi2gmOduDckUSJPB65FAczFza2otGtNxPFvlTSDpqG7fv)[New Goals for the U.S. Human Genome Project: 1998-2003](http://science.sciencemag.org/content/282/5389/682?casa_token=TSuGM1AeEecAAAAA:tkOJWgMedv9YJOO8iSGKSR8D905odNLdjTWb00wyCqxrjxEQDCcC4vkU6n09hyhgUJ-4uQpHxnxV_Lxk) | * [Information on the human genome sequence.](https://www.genome.gov/)
* [The](http://www2.sl.nsw.gov.au/eresources/) assembled human genome sequence may be found at these websites:
	+ [National Center for Biotechnology Information](https://www.ncbi.nlm.nih.gov/genome/guide/human/)
	+ [Genome Reference Consortium Human Build 38 patch release 12](https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.38).
 |

Learning about large data in science

The Science Extension syllabus requires students to explore publicly-available datasets so that their impacts may be evaluated. Students should access and read one or more of the articles listed in Table 1, or other academic articles where datasets from one of these projects were used. They may also have to read other articles to understand the impacts of those datasets.

Discussion

After students have explored the articles (Table 1 or others), teachers should direct classroom discussions about the impact of large datasets on scientific understanding of complex systems. Some of the questions to lead those discussions may include:

* What do we now know that we would not have known if large datasets were not available?
* Could the same discoveries have been made without access to the large datasets?
* What are the benefits of sharing datasets with the broader scientific community (for example, discoveries that the original research groups did not find; the rate of discovery; alternative methods of analysing the same data; combine data from different research groups)?
* How has the availability of publicly-accessible scientific datasets benefitted society (for example, citizen science discovery; private ventures, such as ancestry search and personal genomics)?

Web Resource

The [Australian National Data Service](https://www.ands.org.au/working-with-data/publishing-and-reusing-data/data-in-education) lists several websites that teachers may use for teaching with data.

1. This document references the Stage 6 Science Extension syllabus © 2019 [NSW Education Standards Authority (NESA) for and on behalf of the Crown in right of the State of New South Wales.](http://syllabus.nesa.nsw.edu.au/copyright/) [↑](#footnote-ref-1)
2. This resource references the Science Extension Syllabus © 2017 [Copyright Board of Studies NSW](http://syllabus.nesa.nsw.edu.au/copyright/) for and on behalf of the Crown in right of the State of New South Wales. [↑](#footnote-ref-2)
3. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC genomics. 2012. 13(1):341. [↑](#footnote-ref-3)
4. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. Comparison of next-generation sequencing systems. BioMed Research International. 2012. 25:1365. [↑](#footnote-ref-4)