DATA 301 Introduction to Data Analytics Open Data

Dr. Ramon Lawrence
University of British Columbia Okanagan
ramon.lawrence@ubc.ca

What is Open Data?

Open Data is the movement to make data freely available to all with no restrictions on use or copyright.

Governments have been major supporters and providers of open data as data collected by governments is primarily done to benefit its citizens.

Corporations and other organizations are both producers and consumers of open data.

DATA 301: Data Analytics (3)

Open Data in Canada

Federal, provincial, and local governments have all been involved in the open data movement.

Canadian Federal government: http://open.canada.ca/en

- How to use: http://open.canada.ca/en/working-data
- Statistics Canada: http://www.statcan.gc.ca/eng/rdc/data

British Columbia government: http://www.data.gov.bc.ca/

City of Kelowna:

https://www.kelowna.ca/city-services/city-maps-open-data/open-data-catalogue

Open Data in Canada



DATA 301: Data Analytics (5)

Open Data in BC

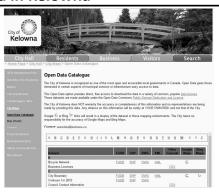


DATA 301: Data Analytics (6)

DATA 301: Data Analytics (2)

DATA 301: Data Analytics (4)

Open Data in Kelowna



DATA 301: Data Analytics (8)

DATA 301: Data Analytics (10)

DATA 301: Data Analytics (7)

Open Data in United States

United States government: https://www.data.gov/

Individual states have their own open data sites as well.

• Example: Washington state: https://data.wa.gov/

United States: Data.gov



DATA 301: Data Analytics (9)

Open Data Worldwide

UK: http://data.gov.uk

The World Bank: http://data.worldbank.org/

• Financial information and statistics

United Nations: http://data.un.org/

OECD: https://data.oecd.org/

Open Data Aggregators

There are many sites that aggregate open data sets (and some data sets for a cost). A Canadian based site is Quandl (http://www.guandl.com).

Kaggle provides many data sets and competitions and techniques for data analytics and machine learning. https://www.kaggle.com/datasets

DATA 301: Data Analytics (11)

Open Data from Companies

Many companies either have public data or application programming interfaces (APIs) that allow people to use their data.

- Google: https://www.google.com/publicdata/directory (public data explorer) and https://developers.google.com/maps/ (Google Maps API)
- Facebook: https://developers.facebook.com/ (API)
- reddit: https://www.reddit.com/dev/api (API)
- Twitter: https://dev.twitter.com/rest/public (API)
- Amazon: https://aws.amazon.com/public-data-sets/ (public data sets) and https://developer.amazon.com/ (API for developers)
- Best Buy: https://developer.bestbuy.com/ (API)

DATA 301: Data Analytics (12)

Try it: Open Data

Explore the federal, provincial, and City of Kelowna data sets to discover "something interesting". Report to your neighbors and to the class.

From any Canadian government open data site, retrieve a data set and analyze and visualize it using one of our tools: Excel, R, Python, Tableau.

DATA 301: Data Analytics (14)

Open Data for Researchers

Increasingly publicly funded researchers are responsible for making their data sets, procedures, and results available to the public (and other researchers).

- Canadian researchers funded by NSERC, SSHRC, CIHR must make their publications freely available within 12 months of publication.
- Researchers in bioinformatics and other fields must make their data sets publically available in a database or repository.

Researchers benefit by having access to public data sets and data sets of other researchers, but there is also a challenge as producing data sets (and perhaps commercializing results) may restrict open access.

Open Data Biology/Bioinformatics

Huge number of databases with most related to NCBI (http://www.ncbi.nlm.nih.gov/) but distributed world-wide.



DATA 301: Data Analytics (16)

DATA 301: Data Analytics (15)

DATA 301: Data Analytics (13)

Open Data in Chemistry

ChEMBL (https://www.ebi.ac.uk/chembl/) stores structures and properties of pharmacologically active molecules.

• Over 1.5 million compounds.

SureChEMBL (https://www.surechembl.org) is a database extracted automatically from patent applications.

 Growing at 80,000 compounds a month and has 16 million compounds from over 13 million annotated patents.

ChemSpider (http://www.chemspider.com/) is a free chemical structure database containing over 43 million structures.

· Supported and hosted by Royal Society of Chemistry.

Open Data in Computer Science

Computer scientists in various fields create standardized data sets for experimentation and research.

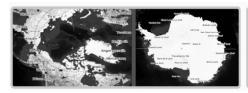
- Databases: Standard performance benchmarks such as TPC (www.tpc.org).
- Machine learning/data mining: UCI ML repository http://archive.ics.uci.edu/ml/
- Game path finding: http://www.movingai.com/benchmarks/



DATA 301: Data Analytics (17)

Open Data in Earth/Environment Science

- Climate Change Data Portal: http://sdwebx.worldbank.org/climateportal/
- National Climatic Data Center: https://www.ncdc.noaa.gov/cdo-web/
- National Geographic Data Center: http://www.nodc.noaa.gov/submit/
- Polar Data Catalog: https://www.polardata.ca/





DATA 301: Data Analytics (18)

Open Data in Physics

Modern physics produces a HUGE amount in data in experiments like astronomical observations and the Large Hadron Collider.

- New research systems developed to handle the large amount of data produced.
- CERN open data portal: http://opendata.cern.ch/
- Data produced is tens of petabytes/year. Large distributed computing of 170 facilities in 36 countries.

Astronomy:

- Canadian Astronomy Data Centre: http://www3.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/cadc/
- National Space Science Data Center: http://nssdc.gsfc.nasa.gov/

DATA 301: Data Analytics (20)

DATA 301: Data Analytics (19)

Open Data in Psychology and Social Sciences

Archaeology:

- Archaeology Data Service: http://archaeologydataservice.ac.uk/
- Many museums have online exhibits and open data.

Psychology:

- · Journals increasing requiring open data sets.
- List of open data sites at: http://guides.library.ucla.edu/c.php?g=180221&p=1188487

By Marie and another the Fertility of the St. Richards Boy Company on the Year 8200. Richards Boy Company on the St. Richards Boy Company on

History

• Digital Archive Database Project (UBC): http://dadp.ok.ubc.ca

Google Analytics

Google Analytics is an analysis service for tracking, optimizing, and understanding user interaction with a web site/service.

Using Google analytics is important for all business, but especially web companies, that rely on users interacting with their site to generate revenue and sales.

Google analytics helps identify and improve content to make it more accessible to potential customers.

• Very important skill set for business owners and managers.

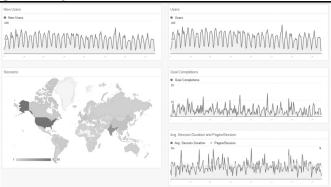
DATA 301: Data Analytics (21)

Google Analytics - Audience Overview



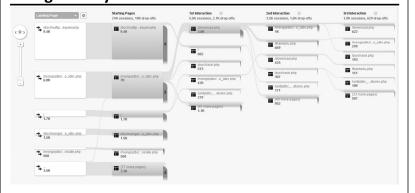
DATA 301: Data Analytics (22)

Google Analytics - Traffic Dashboard



DATA 301: Data Analytics (23)

Google Analytics - Behaviour Flow



Google Adwords

Google AdWords is a service to provide advertisements during searches and as display advertisements on web sites and in apps.

- Primary source of revenue for Google. https://www.google.ca/adwords/
- Companies bid on keywords and display opportunities that are presented by Google and affiliated sites.

Terminology:

- Ad Impression display of an advertisement. Pricing in cost-per-thousand impressions or cost per mille (CPM).
- Click through user clicks on an advertisement (and directly to new location)
- Click through rate fraction of impressions that are clicked on
- Pay-per-click (PPC) companies are billed on each click of an advertisement. The pricing depends on the bid amount and the desirability of the ad location.

DATA 301: Data Analytics (24)

DATA 301: Data Analytics (25)

DATA 301: Data Analytics (26)

Conclusion

Open Data is the movement to make data freely available to all with no restrictions on use or copyright.

Open data has been widely supported by governments and companies wishing to engage users (and developers) with their services.

Data analysts should use open data to help with their analysis whenever available.

Researchers are often responsible for making their publications and data available in an open fashion.

Google provides services for analytics and advertising that are valuable to understand as a business or site looking for user traffic.

Objectives

- Define open data and explain the motivations for making data "open".
- List some of the governments and organizations that provide data in an open fashion.
- Use open data sets when applicable when performing data analysis.
- Explain the role of Google Analytics and Google AdWords. Compare and contrast what these two services provide.