

The NYU Stern Master of Science in Business Analytics program works at the intersection of business and technology to mold business leaders who are capable of leveraging data as a strategic asset. The most important asset of a Data Scientist is the data, but the challenge is in finding the right data and in a form that meets your strategic needs.

As part of the capstone team project, you will work on a yearlong team project and your biggest challenge will be how to approach the data problem.

There are two primary approaches to choosing a dataset to work on:

- 1) Think of an idea / problem that interests you and try to find the data for it.
- 2) Explore readily available datasets and think of interesting ways to analyze the data and come up with business solutions / strategic analysis.

The first approach is very attractive in that you start with a clean slate and think of a business problem that might have a great solution through the strategic analysis of data. However, it poses certain significant challenges as well as the data might not be available publically or quite simply might not even exist. If you go with the first approach, we recommend that you consider the following:

- Does the availability of more data help in the business problem you are trying to solve?
- What type of data would be useful in addressing the business problem—structured or unstructured data?
- Who owns the data? The people who own the data might have great insights into it but might not necessarily realize the strategic value of the data and might not be collecting it.
- Does data exist but is just not being collected? How feasible would it be to collect the data to address the problem?
- Has this problem been tackled before? And if so, what approach was used to gather data and how was it used?

These questions would help you in your approach to solving the data problem when all you have is a problem you are looking to solve or analyze.

For the second approach, we have compiled a list of publically available datasets / sources which should provide you with a good starting point to get a sense of the types of data that are available in the public domain.

Important points to consider before starting your search:

- Who cares about this information? Collecting data is expensive and data providers are big businesses these days. Most likely collectors include governments, marketers, trade groups.
- Time of Data Collection: since collecting data is expensive, the data available might not be the most recent. Given the various capstone deadlines you will have throughout the year, it is important to be realistic about your timeline for data procurement.
- If your data is non-traditional, evaluate the source you are using. Quite often certain data sources might seem very appealing, but the quality of the data might be questionable.
- If you can't find the data you looking for, ask a forum: You can search existing answers or ask a question at GetTheData (getthedata.org) or on Quora(quora.com). Quite often, they yield great results. GetTheData is a great Q&A site where you can ask your data related questions, including where to find data relating to a particular issue, how to query or retrieve a particular data source, what tools to use to explore a data set in a visual way
- Ask an expert: The NYU Bobst library has subject mater librarians who are experts in their fields of locating and sourcing data related to specific subject matter. Leverage this wonderful resource that the school provides.

Streamlining your internet searches:

When searching for data on the web, you need to make sure that you include search terms relating to the content of the data you're trying to find as well as some information on the format or source that you expect it to be in:

- Searching by file type: Append your search with 'filetype:XLS filetype:CSV', geodata ('filetype:shp'), or database extracts ('filetype:MDB, filetype:SQL, filetype:DB').
- Search by part of a URL. Google: 'inurl:downloads filetype:xls' will find all Excel files that have "downloads" in their web address. You can also limit your search to only those results on a single domain name, by searching for, e.g. 'site: agency.gov'.

Public data sources:

Cross-disciplinary data repositories, data collections and data search engines:

1. <http://usgovxml.com>
2. <http://aws.amazon.com/datasets>
3. <http://databib.org>
4. <http://datacite.org>
5. <http://figshare.com>
6. <http://linkeddata.org>
7. <http://reddit.com/r/datasets>
8. <http://thedatahub.org> alias <http://ckan.net>
9. <http://quandl.com>
10. [Social Network Analysis Interactive Dataset Library \(Social Network Datasets\)](#)
11. [Datasets for Data Mining](#)
12. <http://enigma.io>

Single datasets and data repositories

1. <http://archive.ics.uci.edu/ml/>
2. <http://crawdad.org/>
3. <http://data.austintexas.gov>
4. <http://data.cityofchicago.org>
5. <http://data.govloop.com>
6. <http://data.gov.uk/>
7. data.gov.in
8. <http://data.medicare.gov>
9. <http://data.seattle.gov>
10. <http://data.sfgov.org>
11. <http://data.sunlightlabs.com>
12. <https://datamarket.azure.com/>
13. <http://developer.yahoo.com/geo/g...>
14. <http://econ.worldbank.org/datasets>
15. <http://en.wikipedia.org/wiki/Wik...>
16. <http://factfinder.census.gov/ser...>
17. <http://ftp.ncbi.nih.gov/>
18. <http://gettingpastgo.socrata.com>
19. <http://googleresearch.blogspot.c...>
20. <http://books.google.com/ngrams/>
21. <http://medihal.archives-ouvertes.fr>
22. <http://public.resource.org/>
23. <http://rechercheisidore.fr>
24. <http://snap.stanford.edu/data/in...>
25. <http://timetric.com/public-data/>
26. <https://wist.echo.nasa.gov/~wist...>
27. <http://www2.jpl.nasa.gov/srtm>
28. <http://www.archives.gov/research...>
29. <http://www.bls.gov/>

30. <http://www.crunchbase.com/>
31. <http://www.dartmouthatlas.org/>
32. <http://www.data.gov/>
33. <http://www.datakc.org>
34. <http://dbpedia.org>
35. <http://www.delicious.com/jbaldwi...>
36. http://www.faa.gov/data_research/
37. <http://www.factual.com/>
38. <http://research.stlouisfed.org/f...>
39. <http://www.freebase.com/>
40. <http://www.google.com/publicdata...>
41. <http://www.guardian.co.uk/news/d...>
42. <http://www.infochimps.com>
43. <http://www.kaggle.com/>
44. <http://build.kiva.org/>
45. <http://www.nationalarchives.gov....>
46. <http://www.nyc.gov/html/datamine...>
47. <http://www.ordnancesurvey.co.uk/...>
48. <http://www.philwhln.com/how-to-g...>
49. <http://www.imdb.com/interfaces>
50. <http://imat-relpred.yandex.ru/en...>
51. <http://www.dados.gov.pt/pt/catal...>
52. <http://knoema.com>
53. <http://daten.berlin.de/>
54. <http://www.qunb.com>
55. <http://databib.org/>
56. <http://datacite.org/>
57. <http://data.reegle.info/>
58. <http://data.wien.gv.at/>
59. <http://data.gov.bc.ca>
60. <https://pslcdatashop.web.cmu.edu/> (interaction data in learning environments)
61. <http://www.icpsr.umich.edu/icpsrweb/CPES/> - Collaborative Psychiatric Epidemiology Surveys: (A collection of three national surveys focused on each of the major ethnic groups to study psychiatric illnesses and health services use)
62. <http://www.dati.gov.it>
63. <http://dati.trentino.it>
64. <http://www.databagg.com/>
65. <http://networkrepository.com> - Network/ML data repository w/ visual interactive analytics

Economics:

- 1) American Economic Assn (AEA): <http://www.aeaweb.org/RFE/toc.php?show=complete>
- 2) Gapminder: <http://www.gapminder.org/data/>
- 3) UMD: <http://inforumweb.umd.edu/econdata/econdata.html>
- 4) World bank: <http://data.worldbank.org/indicator>
- 5) FRED database: St Louis Federal Reserve : <http://research.stlouisfed.org/fred2/>

Data Science:

- 1) This section contains data sets used in the book "Doing Data Science" by Rachel Schutt and Cathy O'Neil (O'Reilly 2014) Datasets on the book site: https://github.com/oreillymedia/doing_data_science
- 2) Enron Email Dataset: <http://www.cs.cmu.edu/~enron/>
- 3) GetGlue (time stamped events: users rating TV shows): <http://bit.ly/1aL8XS0>
- 4) Titanic Survival Data Set: <http://bit.ly/1kJ4pkF>
- 5) Half a million Hubway rides: <http://hubwaydatachallenge.org/trip-history-data/>

Finance:

- 1) CBOE Futures Exchange: <http://cfe.cboe.com/Data/>
- 2) Google Finance: <https://www.google.com/finance>
- 3) GoogleTrends: <http://www.google.com/trends?q=google&ctab=0&geo=all&date=all&sort=0>
- 4) NASDAQ: <https://data.nasdaq.com/>
- 5) OANDA: <http://www.oanda.com/>
- 6) Quandl: <http://www.quandl.com/>
- 7) Yahoo Finance: <http://finance.yahoo.com/> (R)

Government:

- 1) Archived national government statistics: <http://www.archive-it.org/>
- 2) Australia: <http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/3301.02009?OpenDocument>
- 3) Canada: <http://www.data.gc.ca/default.asp?lang=En&n=5BCD274E-1>
- 4) DataMarket: <http://datamarket.com/>
- 5) FDA: <https://open.fda.gov/index.html>
- 6) Fed Stats: <http://www.fedstats.gov/cgi-bin/A2Z.cgi>
- 7) Guardian world governments: <http://www.guardian.co.uk/world-government-data>
- 8) HUD: <http://www.huduser.org/portal/datasets/pdrdatas.html>
- 9) London, U.K. data: <http://data.london.gov.uk/catalogue>
- 10) New Zealand: http://www.stats.govt.nz/tools_and_services/tools/TableBuilder/tables-by...
- 11) NYC data: <http://nycplatform.socrata.com/>
- 12) OECD: http://www.oecd.org/document/0,3746,en_2649_201185_46462759_1_1_1_1,00.html
- 13) RITA: http://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp
- 14) San Francisco Data sets: <http://datasf.org/>
- 15) U.K. Government Data: <http://data.gov.uk/data>
- 16) United Nations: <http://data.un.org/>
- 17) U.S. Federal Government Data Catalog: <http://catalog.data.gov/dataset>
- 18) U.S. Federal Government Agencies: <http://www.data.gov/metric>
- 19) US CDC Public Health datasets: http://www.cdc.gov/nchs/data_access/ftp_data.htm

- 20) The World Bank: <http://wdronline.worldbank.org/>
UK 2011 Census Open Atlas Project: <http://www.alex-singleton.com/2011-census-open-atlas-project/>

Health Care:

- 1) Gapminder: <http://www.gapminder.org/data/>
- 2) US department of Health and Services: (Largest collection of longitudinal hospital care data in the United States:
<http://www.ahrq.gov/research/data/dataresources/index.html>
- 3) Open Payments Data [http://healthdata.gov/data/dataset/ open-payments-data-0](http://healthdata.gov/data/dataset/open-payments-data-0)
- 4) Physician Referral Data [http://healthdata.gov/data/dataset/ physician-referral-patterns](http://healthdata.gov/data/dataset/physician-referral-patterns)

Assorted (interesting)

- 1) Peer to Peer lending data : lendingClub:
<https://www.lendingclub.com/info/download-data.action>
- 2) Facebook Data: Complete set of Friends of various School networks:
<http://masonporter.blogspot.ae/2011/02/facebook100-data-set.html>
- 3) Public Dataset on AWS: <http://aws.amazon.com/public-data-sets/>
- 4) Youtube Networks Dataset: <http://netsg.cs.sfu.ca/youtubedata/>
- 5) IMDB Dataset: <http://www.imdb.com/interfaces>
- 6) UCI – Machine Learning Dataset: <http://archive.ics.uci.edu/ml/>
- 7) Google Books n-gram dataset:
<http://aws.amazon.com/datasets/8172056142375670>
- 8) Million song database: <http://labrosa.ee.columbia.edu/millionsong/>