4th QS MAPLE, Abu Dhabi, UAE, May 6-8, 2014

National and International Economic and Social Data in Teaching and Learning

Repositories, web interfaces, analytics

Chris Leowski
University of Toronto

1. Abstract of presentation

Statistical data – whether economic, social, financial, industrial, ecological, cultural, etc. – collected by national statistical bodies and by international organizations, are frequently made available to academic institutions on special terms, and form the basis for exciting new teaching curricula, as well as provide stimulus for academic research and international academic collaboration. This presentation will show how such repositories were built at the University of Toronto, in cooperation with Statistics Canada and other institutions, and how their use has expanded to over 60 universities in Canada and in the U.S. Other similar on line repositories and analytical portals will be described, and a case will be made for academic collaboration in making real data available to students and researchers, and for building on line analytical tool chests to teach students advanced analytical methods and prepare them for their future careers through immersion in real, dynamic, and changing daily flow of data.

2. Remarks

The proposed presentation is aimed at either Track 2 or Track 5 of the Conference. The objective is to highlight the need for building academic curricula, especially in social sciences and finance, around real data, as compiled by national statistical bodies and by international organizations (e.g. OECD, World Bank) – and enhance teaching and learning experience, as well as academic research, by providing students and researchers with large easily accessible on line data repositories and analytical tools to mine those repositories. Through this approach, students get hands-on experience in retrieving and analyzing real data, and are better prepared for their future professional careers.

3. Summary/biography

Dr. Chris Leowski is Director of CHASS, an information and instructional technologies computing centre in the **Faculty of Arts & Science, University of Toronto, Canada**, providing teaching,

learning, and research IIT technologies to over 30 academic departments, and to 40,000 students. Formerly Acting Assistant Dean responsible for the IIT portfolio. Architect and co-designer of on-line search and retrieval systems for economic and financial data repositories.

Dr. Leowski graduated from the Warsaw School of Economics, Poland, where he taught economics for 8 years, before becoming research professor in the Graduate Centre for Administrative Systems, Institute of Technology, Mexico. He joined the University of Toronto in 1984 and became director of CHASS in 1991.

Dr. Leowski is a specialist in areas of management of instructional and information technologies, with emphasis on best practices frameworks; data architecture; information storage, search and retrieval; knowledge repositories; collaborative portals, remote collaboration, and course management solutions.

TRANSCRIPT

Please note that this is merely a transcript of a brief presentation at a conference, and not an academic paper. As such, it follows different rules, and makes certain allowances, regarding the flow of argument, the style, the grammar, repetitions, and/or emphasis. Due to time restrictions, I attempted to present a coherent line of thought as per supporting slides and live online examples.

The slides themselves serve merely as a "teleprompter" to allow me to expand on topics that they outline. The live online examples are more complex and interesting, but they can be meaningless to the unprepared listener without the accompanying description and explanation – which you can find in the transcript. By the same token, reading only the transcript – at least parts of it - can be meaningless without being able to look at on line interfaces and follow live examples of data retrieval and analysis.

I have added a few online links, used by me during the presentation, and a few other references, for the reader's perusal.

Hello, and welcome.

As you can see on the title slide, this presentation is about national and international databases in teaching, learning, and in academic research, and I will be talking about data repositories, about web interfaces, and about data analysis.

A couple of decades ago I was put in touch with one of the Senior Directors at Statistics Canada. His division was responsible for distributing data compiled by Statistics Canada, and we – in Canadian universities - were looking for up-to-date data on every aspect of Canadian life – be it economic performance, health statistics, information on education, tourism, crime, housing, you name it. We wanted live data feeds – daily, monthly, etc., (depending on the frequency of data), and all information that was not suppressed for privacy or national security reasons. Why, what were the reasons?

The first reason is obvious: economists and other researchers, who develop econometric models and conduct various research activities, require access to the latest information. Back in those days – 1990s - when Statistics Canada pre-released (to a very limited audience) quarterly national accounts data, my team was working after hours to update the relevant databases on the University's servers.

But even more important, in my view, are the students. These young people graduate, get their postgraduate degrees, and assume various positions in business and the government, where – sooner or later – they are responsible for making many decisions, and for arguing their case *vis-à-vis* other agencies,

ministries, corporations, individuals in position of power, based on how they can use, understand, and process all the information at their disposal.

Data on health care and health trends, labor markets, criminology, tourism, immigration, infrastructure, communication, and many others, are critical to daily decisions of those who are in charge of our public life, and those who are in charge of business organizations.

Before we get to the topic of **how** we should teach students to use real data, let us briefly go over the data "life cycle". Use of data is only one stage in that sequence.

<next slide>

Data "life cycle"

The full data "life cycle" comprises data collection, storage, search and retrieval, then data presentation, and finally statistical analysis of data. I will cover briefly all of them, but will focus on search, retrieval, presentation, and analysis – that is, the phases of the data "life cycle" after the data have already been compiled and made available.

Data collection / compilation:

Most, if not all, social and economic data are collected and compiled either as time series (data points over time) or cross-sectional and/or longitudinal studies (surveys). A population census is, in essence, a survey – a

questionnaire with a number of questions; the results of which form a huge micro-data file, from which all aggregate statistics are compiled.

<drop from live presentation if time restricted>

Data can be collected "at source" (e.g. point of sale); can be collected by individual government agencies and ministries in respect to their areas of authority (e.g. data on education; extraction of oil; exports by product category and destination; etc.); data can be collected by well established reporting procedures (e.g. hospital data, and generally most of – though not all - health data). Data can be collected in a very exact manner, and can be very reliable, or they can be estimated or inferred from smaller sampling. The latter can be statistically significant, or less so. Data can be collected in a mandatory, even if often imperfect, way (e.g. population census), or *via* random (or not-so-random) sampling of respondents (think about various surveys of CEOs of large corporations regarding the state and future directions of the economy).

Data storage / repositories:

Most economic and social data, in fact – most of any type of data, are stored in relational database management systems. Why is that so? – first and foremost, a well designed RDBMS allows us to maintain data consistency and data integrity. RDBMSs have matured over the last two decades, and offer extremely efficient and functional storage, search, and retrieval of even very large bodies of data: hundreds of millions and billions of records. They accommodate all known data types. They offer many tools for data

manipulation, and even data analysis, though there are many specialized tools for the latter task.

<next slide>

Data presentation, search and retrieval:

Historically, economic and social data have been most often presented as time series, that is, collections of data points over time - the points themselves expressing values per unit of time, or at a certain point in time. We have soon realized that most data have multiple dimensions, for example data on car manufacturing can be by type and model, geography, intended use (passenger, commercial), type of engine (diesel, petrol, electric, hybrid), and probably a few more. We started developing simple multi-dimensional views, and then more complex cubes and hyper-cubes based on OLAP techniques, in order to be able to look at data from different perspectives, through those individual dimensions. And, finally, some data do not form time series at all, but are snapshots compiled from surveys conducted at a specific point in time. We use different techniques to present and analyze such data.

<next slide>

Let us move now to some on line examples of data search, data retrieval, and – especially – data presentation. There are hundreds, or even thousands, of WEB-based data repositories, from which I have listed 5 on this slide. They are kind of "representative" – some of them allow access to, and retrieval, of just time series, with rudimentary or more sophisticated search engines, and with or without data visualization. Others offer quite advanced view of data

dimensions, even manipulation of those dimensions ("rotating the data cube"); and a few have built-in analytics.

[Web page links from the slide:]
http://dc.chass.utoronto.ca
http://wrds-web.wharton.upenn.edu/wrds/
http://odesi.ca
http://oecd.org
http://data.worldbank.com

Let us go through the slide examples in reverse order – and we shall do it quickly, because we need time for more interesting examples later. As you can see, the World Bank databases allow simple, though quite elegant, search and retrieval of time series: by country, by topics, etc. OECD databases are somewhat similar. The OECD site offers a wealth of statistical information, and much of it is easily accessible and often quite advanced in data presentation.

[OECD Home -> Topics -> Agriculture and Fisheries -> FIND: OECD-FAO Agricultural Outlook -> Database -> By Commodity (English) -> and there we have it.]

Here is another good example of data comparison using interactive data retrieval and presentation, that is, the user is in the driver's seat: life expectancy in OECD countries: oecd.org -> Statistics -> Go to: Data Lab -> Health: Health Data.

We can compare life expectancy between the entire group and individual countries; or between any two individual countries.

Switching to the Health Risks tab – health risks expressed as consumption of alcohol and smoking – let us compare, for example, Canada to France, and then Canada to Turkey (very small alcohol consumption, but heavy smoking).

The OECD Data Lab is also a nice example of data visualization. Visualization can be very attractive, especially for younger audiences, and useful in preliminary data exploration, but is rarely a serious research tool.

Moving on in our quick survey of data sites: here is <ODESI>, a Scholars Portal for Ontario universities. It offers standard data search and extraction, but also includes an analytics engine Nesstar, developed in Norway. Since I will be showing you a similar analytics engine running at my computing centre, and developed at the University of California, Berkeley, we will skip this one.

We are now moving towards more specialized data sites offering very detailed, and often either monothematic statistical information, or information on a single country. WRDS (Wharton Research Data Services) is an example of the former – mostly financial data. CANSIM is an example of the latter.

Here is the Canadian National Database CANSIM – around 50 million times series organized in 3,500 multi-dimensional hyper-cubes. I'll show you a couple of those cubes, and we will be doing real time flipping of dimensions and selecting subsets of data through an OLAP-based retrieval and

presentation engine. I hope to be able to show you how useful such tools, and such methods of working with data, are in teaching and learning. And how much fun they offer to students studying specific economic or social topics and completing their assignments. For illustration purposes, I will use data on sales of natural gas in Canada.

I'll need to authenticate, since I am not on the University of Toronto's network – here we are, and now we go to table search: Locate tables by numbers - type our table number (1290003), and select OLAP multidimensional view. As you can see, there are other views available to those who got used to earlier interfaces, but we won't deal with those today.

Let me backtrack a little and show you how else you can find this information – after all, you rarely know the relevant table's number. Well, you have at least two other options available: one is "Search by subject". We are looking for Sales of natural gas, so logically it should be within the "Energy" subject, and – from the choices available to me "Energy consumption and disposition" seems to be the best fit. Our table is in position 36 – but the important fact here is that we can see what else is available that might potentially be of interest.

Finally, we can go straight to "Text search", type "Sales of natural gas" – and find our table in position 9. Look at what else is available. Look also at "More search options".

Presentation of multi-dimensional data:

[Links from the data presentation slide:]

Sales of natural gas:

http://clouddc.chass.utoronto.ca.myaccess.library.utoronto.ca/ds/cansim/olap/displayCube.do?action=browse&a=1290003&lang=
Postsecondary enrolment:

http://clouddc.chass.utoronto.ca.myaccess.library.utoronto.ca/ds/cansim/olap/displayCube.do?action=browse&a=4770031&lang=
Teaching staff at Canadian universities:

http://clouddc.chass.utoronto.ca.myaccess.library.utoronto.ca/ds/cansim/olap/displayCube.do?action=browse&a=4770018&lang=

We are now looking at a dataset on Monthly Sales of Natural Gas in Canada. There are only 4 dimensions here: **TIME** is obviously one of them; then **GEOGRAPHY** (Canada and all provinces and territories); then **SECTOR** (which shows whether the data refer to commercial, residential, or industrial customers, or all of them); and finally a dimension called here **ESTIMATES** (which refers to the number of customers, or sales in natural units (e.g. cubic metres), or sales / revenue in dollars).

Since a flat surface can display two dimensions, we choose **Geography** in columns and **Time** dimension in rows, and filter data by one and only one member of all remaining dimensions. Let's see: from the **ESTIMATES** dimension we have selected Revenues from sales of natural gas; and from the **SECTOR** dimension we have selected Total Sales (as opposed to, say, Residential Sales). But we could have selected any other member of those dimensions.

Let me now show you on-the-fly operation on data. Let us assume that I am interested only in data for Canada and four of its provinces: Ontario, Quebec,

British Columbia, and Alberta. Easy: in the **GEOGRAPHY** dimension I de-select all those that I am currently not interested in. Let me now assume, that in each of these geographical regions I want to compare residential to industrial sales of natural gas. That is, I want to see data for two members of the **SECTOR** dimension for each of the selected provinces (and for the entire country), and over time. First, I nest the **SECTOR** dimension within the **GEOGRAPHY** dimension – like this. Then, within the SECTOR dimension I select Residential and Industrial sales, and de-select total sales. Done. Data are recalculated on the fly. Finally, I want to download data to Excel on my laptop, but I'd like to have GEOGRAPHY displayed in rows, rather than columns. I swap dimensions, and then save my data by clicking on the Excel icon. Done.

[The next data table is an example of a hypercube with 6 dimensions: Post-secondary enrolment by immigration status, country of citizenship, and sex. GEOGRAPHY and TIME are our standard dimensions on the axes. We might as well limit the GEOGRAPHY dimension's members to Ontario, Quebec, British Columbia, and Canada as a whole – these are the most significant ones. In the Geography dimension we de-select all other provinces, and data are recalculated on the fly. The other 3 dimensions are: INSTITUTION TYPE (with three members: university, college, total); and let us assume that we are interested only in university enrolment. IMMIGRATION STATUS dimension differentiates between Canadian and international students. Let us select international students only. As you can see, there were over 100,000 international students in Canada in the last two years for which the data have been published. Incidentally, you can also see that the numbers have doubled over the past decade. The bulk of students are enrolled in universities in Ontario, although Quebec and British Columbia have also a very significant

international enrolment. Finally, we would like to compare numbers for students coming to Canada from Europe vs. those coming from Asia. Since we want to display data on two members of a dimension that is on the filter, we need to nest that dimension in either one that is on the axes. It is more logical to nest it within the GEOGRAPHY dimension. Now we can de-select totals, and select Europe and Asia as the regions from which students came to study in Canada. [Code 2* is for Europe, code 4* is for Asia]. Not much of a surprise that while in Quebec many more students come from Europe than from Asia (French is probably the major factor), in British Columbia the proportions are completely different: there are over 6 times as many international students from Asia as those from Europe. The surprising finding is that for Ontario the proportion of Asian to European students is even higher – almost 8 times as many come from Asia than from Europe.

A link to one more CANSIM hyper-cube, with 7 dimensions, is provided on the slide. The data refer to various characteristics of teaching staff at Canadian universities – but let me leave it to users to explore this one.]

Analytics:

Let us move now to on line data analysis. As an example, I have chosen the SDA software, developed by the University of California at Berkeley and built into the on line data search and analysis system at the University of Toronto, with approximately 15 Canadian universities subscribing to this offering. SDA is particularly useful in analyzing survey data, and in teaching students statistical methods for data analysis. Our time today allows us to cover only a

tiny fraction of its features, and we will focus on one or two surveys, and on frequency analysis, cross-tabs, and correlations.

First, let us have a look at typical features that we would like to see in this kind of an analytical system:

1. It should be web-based, or at least accessible *via* the Internet. By Web-based I mean systems that are accessible through standard Web browsers. There are other ways of accessing applications on the Internet, but this one is by far the most common.
2. It should allow users to discover data, search for data, explore them, and to study the documentation.
3. It should offer many built-in statistical analysis methods, like measures of significance and direction, frequencies, cross-tabulations, comparison of means, regression analysis, etc.
4. It should allow users to graph results of the analysis.
5. It should allow users to download and/or export data to other analytical packages (e.g. SPSS, SAS, STATA, etc.), and to download results of the analysis to include them in student assignments and in research papers.

Here is the SDA interface at the University of Toronto. There are around 900 surveys in the collection. Let's pick up Canadian Community Health Survey. We are interested in looking at how people perceive their own health while they are getting older. This has been quite an extensive survey, and you can see all variables in the left-hand window. The top menu includes the Codebooks, and this is usually a good source of information about the survey

itself. You can see the sequential and alphabetical listing of all variables, you can also study all other documentation on this survey, including the actual questionnaire.

[Canadian Community Health Survey; cycle 3.1; common and optional content; documentation; data. NOTE: login required – interested users should contact Chris Leowski as per contact information at the end of this transcript and/or in the last slide.]

http://sda.chass.utoronto.ca.myaccess.library.utoronto.ca/sdaweb/html/cchs.htm

Let us do a simple cross-tabulation of self-rated health by various age categories. First, we select the self-rated health variable (General Health (GEN) -> Self-rated health (genedhdi)) and put it into rows, then the age variable (Demographics and household (DHH) -> age, sex, marital status -> age (dhhegage)), which we put into columns, click the Run button, and here is our cross-tabulation of self-rated health by age. As could be expected, the older people get, the higher percentage of them rates their health as poor, and smaller percentage declare to be in excellent or very good health. This is clearly visible once we graph this relationship.

Back to the interface – let us look at what analytical tools are at our disposal. We could have asked for summary statistics in the previous step, and get the cross-tabulation with all summary statistics, and the graph. We have a choice of various graphing options. Clicking on the Analysis menu button, we can

select frequencies or cross-tabulations – just what we have done – but also comparison of means, correlation, regression, and so on.

I find SDA@CHASS to be an excellent research tool, but also a wonderful teaching tool. Students work with real data obtained through surveys – imagine courses in public health, community services, police services, criminology, labor markets, immigration management, etc. They can apply their theoretical knowledge of analytical methods right into analysis of these real data.

<next slide>

Conclusions:

So, now to conclusions.

I started this presentation with two questions:

(1) why we should teach students to use real data in completing their assignments, writing essays, diagnosing economic and social issues, and discussing solutions, during the course of their studies?; and

(2) how we should be doing it?

The answer to the first question focused on the fact that many of these young people, after graduating and completing their post-graduate degrees, assume important positions in business, public administration, and government, and need to be trained in finding information, assessing trends, grasping issues, and evaluating alternate solutions in their respective areas of expertise.

Now – how we should be doing it? The primary condition for using up-to-date statistical information is to provide access to as many academic and commercial data repositories as possible, and to train faculty how to use them. This does not mean duplicating costly repositories at every institution of higher education. Academic consortia, and other methods of sharing databases, are not only more cost-effective, but also offer a wider selection of case studies, and a body of expertise that can be shared between institutions. [In some countries, notably in the U.K., there exist national consortia and other government funded national data repositories, but whether this is the best model, or whether a somewhat less centralized approach, closer to teaching programs at individual schools, would be better, is still debatable.]

And finally, teaching students how to navigate those databases, how to search for and extract data, needs to be accompanied by a thorough preparation in analytical methods, by which I mean not only statistical methods – though they are a must - but also essential rules of logical inference.

<next slide>

This concludes my presentation, but no presentation is complete without information on where to find the slides, the transcript, and the recording, and this is what the last slide provides.

Many thanks for your attention.