Data Analytics Research in Public Universities

Alan F. Smeaton
Insight Centre for Data Analytics
Dublin City University
Glasnevin, Dublin 9, Ireland
Alan.smeaton@dcu.ie

There are now so many definitions of *big data* that its like a box of candy, you can pick out whichever one takes your fancy. This choice of definitions points to the fact that *big data* is an embracing term, a catch-all, more of a philosophy or a way of doing business than a technology that can be adopted. We define big data or data analytics not in terms of the usual 3 (or is it 4, or even 5) "V"s (velocity, volume, variety, veracity, visualization, etc.) but as the discovery and the exploitation of patterns, links and connections between and among different data sources. What that means is that it's not all about large volume, or even about data which captures changes in some underlying system, its more about being able to link, connect and infer or mine in order to learn new things, to garner new insights.

Research into *big data* in publicly-funded Universities and research centres has major disadvantages compared to the private sector, and not just in the obvious areas of funding and access to data. In this abstract we highlight some of these differences around the area of ethics and privacy, and two specific examples of our work are used to illustrate this.

The first example is in the area of personal big data[1], also known as lifelogging or the quantified-self. This involves automatic and ambient self-logging of everyday behaviour, whereby a subject gathers sensed information on him/herself in order to create a lifelog, a personal black box of your own daily activities. Many of us know this as using a Fitbit or a sleep monitor, but for over 7 years we have been working in this area, mostly using wearable cameras taking images and video from a first-person viewpoint but also using other wearable and environmental sensor data. We have gathered several dozens of person-years of (visual) lifelog data, equating to close to 50 million images. The applications we have been using this for include as a memory prosthesis/aid (specifically for people with early-stage dementia), for diet monitoring, for analysis of the exposure children have to fast-food advertising, for lifestyle analysis of professions known to have poor diet (specifically jockeys), for brand name exposure in marketing and shopping, and others. In all cases we perform content-based analysis on the logged images/video in order to infer activities and other semantic features, and we augment this information with data from other wearable sensors (GPS, acccelerometers, etc), pulling in publicly-available information to semantically enrich the person's lifelog. For example GPS coordinates -> placenname -> Wikipedia entry -> enriched description of a visit to the zoo which can help trigger recall for a person with dementia when reviewing their lifelog

[1] Gurrin, Cathal, Smeaton, Alan F. and Doherty, Aiden R. (2014) *LifeLogging: personal big data.* Foundations and Trends in Information Retrieval, 8 (1). pp. 1-125. ISSN 1554-0677

of a day in their life.

All our work on lifelogging has to be approved by our University's Research Ethics Committee and this has involved informed consent, plain language statements, ensuring there are opt-out options at all stages, deletion and removal of some lifelog data at the subject's request, as well as security of the data, preservation of anonymity, and complete deletion of all data a fixed time after the research has been completed. This has been restrictive, but workable partly because our lifelogging is regarded as *sous*veillance rather than a form of surveillance and the data is gathered by the subject and used exclusively by the subject and for the benefit of the subject.

The second example of our data analytics research which is relevant here is our predictive analytics on student exam performance, using demographic and behaviour information. Specifically we use some aspects of student demographics (age, gender, commuting distance to University, etc.) combined with their behaviour when accessing the University's Virtual Learning Environment (VLE) to generate a weekly prediction, on a module-by-module basis, of whether each student is likely to pass or fail. This uses the data warehouse of past student exam performances and past student VLE behaviours as training data for an ensemble of classifiers, one for each module/week combination.

There is much scope to leverage the wealth of student data available in Universities for applications like predicting student needs or alerting students of their likely exam performance based on their current behaviours, and this can be used as a force for good.[2] University records of student activities include access to library facilities and borrowings, access to sports facilities, exam performance, course preferences, as well as demographic information. Using information the University logs on internet access via *eduroam*, we can use the kinds of big data analytics techniques used in industry to infer things like social groupings which would allow us to explore whether who you have as friends in College will influence examination performance. We could track students' public social media postings to analyse their extra-curricular social activities and moods. Finally, we could use public information such as census or socio-economic distributions to calculate, for example, the value of a student's family home, family income, the types of car(s) the family drive, etc.

While such data mining may be acceptable in industry settings behind closed doors, in public institutions it is not. Our work on predictive analytics for education has required advance approval from the University's research ethics committee who have struggled with this issue since there is no best practice in ethics (yet) that we can follow. Without such best practice being available, our research ethics committee has steered us towards informed consent, plain language statements, ensuring there are opt-out options at all stages, deletion and removal of some student data at the subject's request, as well as preservation of anonymity, and complete deletion of all data a fixed time after the research has been completed. I

[2] University Data Can Be A Force For Good, The Guardian Newsppaer, November 2013 http://www.theguardian.com/higher-education-network/blog/2013/nov/27/university-data-student-engagement-retention

f that last sentence sounds familiar then it is indeed déjà-vu, because these are precisely the same conditions under which our work on lifelogging is carried out. It seems that University research committees have only one tool at their disposal, and they use that same tool all the time, and that is not a good place for us all to be.

What makes all this so frustrating and self-defeating is that by informing students that their online behaviour is being used to make predictions about their exam performance, we are thus corrupting that precise behavior which we want to measure. Facebook and Google can get away with doing that, we can not. It seems that the observer behavior, a phenomenon we know to be present in thermodynamics, particle physics and electronics, is also present in University research when we try to work with people as subjects.