

Title: Dealing with the Data Deluge: What can the Robotics Community Teach us?

Authors: London, Darin - Institute for Genome Science and Policy, Duke University Medical Center (darin.london@duke.edu); Furey, Terry - Institute for Genome Science and Policy, Duke University Medical Center; Boyle, Alan - Institute for Genome Science and Policy, Duke University Medical Center

URL for Overall Project Website and Access to the Code: <http://search.cpan.org/~dmlond/Google-Spreadsheet-Agent-0.01/lib/Google/Spreadsheet/Agent.pm>

License: This program is free software; you can redistribute it and/or modify it under the terms of either: the GNU General Public License as published by the Free Software Foundation; or the Artistic License.

In 1986, Rodney Brooks published a paper which revolutionized the robotics industry (Brooks, 1986). In it, he presented work on a robot based on a new architecture, called the Subsumption architecture, which is composed of layers of very simple autonomous modules each designed with a very tight sense-response circuit, with modules in one layer able to mask/amplify (subsume) the inputs and outputs of modules in other layers. The critical advantage of this architecture is that it places many layers of intelligence between the 'data' of the world, and the overall action of the robot, without the need for any single internal representation of the world. This architecture was radically different from previous approaches which attempted to use a single, central processing system to 1. store a map or representation of the outside world based on the combined inputs from all sensors, 2. evaluate the state of the world, and decide which steps to take towards reaching the overall goals, 3. instruct all effectors how to act to achieve the current goal, 4. repeat. Though controversial, and highly debated within the robotics community at the time, no one now disputes that this architecture has single-handedly moved robotics from a theoretical academic adventure to a multi-billion dollar industry.

In many ways, we as bioinformaticians are now reaching a similar point. Most of us older folk grew up in a world of 1-server:1-dataset, whereby we wrote our scripts to store all of the tasks required to conduct our analyses along with all of the logic for performing those tasks, and very little intelligence for when (best) to conduct these tasks, or how to (re)prioritize tasks in different ways. All of us are now entering into a world of potentially infinite servers, and potentially infinite amounts of data to analyze. In order to make full use of the servers that will be available to us to analyze this data, our simple, dumb scripts are going to need to become more organic, responsive, and adaptive. We must also begin to move past the view that NextGen sequencing presents a problem of too much data, and begin to view it as a problem of not having enough layers of intelligence between the data and us to filter and amplify the signals that we are interested in from the noise that we wish to ignore.

This talk presents a module facilitating the creation of a subsumption architecture, called `Google::Spreadsheet::Agent`, available from the Comprehensive Perl Archive Network under the same terms as Perl itself. This module allows many highly focused intelligent agents, running on many different servers, to act autonomously to perform tasks on many datasets stored in the same, central Google Spreadsheet. Furthermore, by allowing concurrent access and editing to both humans and computer agents, the system presents a natural way of integrating the activities of human agents into the overall function of the system. This architecture has been implemented successfully into a pipeline to analyze Next Generation (Solexa and Illumina) Sequence Data associated with the Encyclopedia of DNA Elements (EncODE, <http://www.genome.gov/10005107>), and is currently being applied to a similar pipeline designed to research cross-species comparisons of Chimp, Rhesus, Mouse, and Human cell lines data.

References:

Brooks, Rodney, A., 1986. 'A Robust Layered Control System for a Mobile Robot'. IEEE Journal of Robotics and Automation, VOL RA-2, No. 1