21st International Congress on Modelling and Simulation, Gold Coast, Australia, 29 Nov to 4 Dec 2015
www.mssanz.org.au/modsim2015

EXTENDED ABSTRACT ONLY

Data farming: what it is, and why you need it!

Sanchez, S.M.

Naval Postgraduate School (NPS), Monterey, California, USA
Email: ssanchez@nps.edu

Abstract: Simulation models are integral to modern scientific research, national defense, industry and manufacturing, and public policy debates. These models tend to be extremely complicated, often with large numbers of factors and many sources of uncertainty, but recent breakthroughs help analysts deal with this complexity. Data farming is a descriptive metaphor that captures the notion of generating data purposefully in order to maximize the information "yield" from simulation models. Large-scale designed experiments let us grow the simulation output efficiently and effectively. We can explore massive input spaces, uncover interesting features of complex simulation response surfaces, and explicitly identify cause-and-effect relationships. Data farming has been used in the defense community for over a decade, and has resulted in quantum leaps in the breadth, depth, and timeliness of the insights yielded by simulation models.

Data farming draws on tools and techniques from data mining and "big data" analytics—but goes a step further. Numerous big data success stories are touted regarding the interesting patterns that are found by sifting through massive volumes of data, and then treated as actionable information. However, a key drawback to the big data paradigm is its reliance on observational data—causality remains unprovable. With data farming, we can leverage techniques developed for big data while retaining the ability to determine cause and effect. Expanding on the "3 V's" of observational big data (*volume*, *velocity*, and *variety*), the "3 F's" of inferential big data are *factors*, *features*, and *flexibility*. A "big factor" view embraces a broad exploration of the inputs (or functions of inputs) that, when varied, increase our understanding of the simulation responses. A "big feature" view refers to the simulation responses—we are typically interested in many, and they may be of different types. A "big flexibility" view captures the need to answer a wide variety questions from our experiments, even if we don't know *a priori* all the questions that might be asked—so our simulation study must be designed accordingly.

I will give an overview of the principles of data farming, and describe some recent applications in defense and homeland security. The bottom line: once you have invested the time and effort to develop a simulation model, it's time to let that model work for you!

Keywords: *Data farm, experiment design, data analytics*