# Splash: Smarter Planet Platform for Analysis and Simulation of Health

**Melissa Cefkin, Myron Flickner, Susanne Glissmann, Peter J. Haas, Leila Jalali, Paul P. Maglio, Pat Selinger, and Wang-Chiew Tan**

IBM Research - Almaden

July 27, 2010

## 1. What is Splash?

**Splash** is a research project aimed at building a framework that supports the integration of multiple existing models, simulations, and data that represent parts of the broader health ecosystem. Specifically, our goal is to create a platform that takes expert models of constituent real-world systems related to health, synthesizing and integrating those models, resulting in an interoperating complex composite system model with which policy-makers can try out alternatives in a low-cost, highly responsive way. The key research question is whether such integration of independently created, deep domain models can be made feasible, practical, flexible, cost-effective, attractive, and usable.

## 2. State of the Science Today

### 2.1 Health-Focused Policy Decision Making

Recent years have seen great innovations in both technologies and organizational mechanisms for promoting human health, but effective policy and investment decisions are needed to reap the benefits of these advances. The challenge for decision-makers is that the health ecosystem comprises a large set of complex, intricately connected subsystems. Consider, for instance, that major reductions in morbidity and mortality in the Western world in the early 1900s were achieved due to clean drinking water systems, modern sewerage systems, and modern hygiene [1]. The health system is not simply healthcare, but stands together in a complex relationship with many other real world systems: transportation, agriculture, housing, and education "have far-reaching health effects, but are not engaged or evaluated for those outcomes" [2]. For example, the treatment of chronic diseases presents multi-faceted issues that the healthcare sector alone cannot address. Indeed, it is generally recognized that chronic diseases such as obesity reflect cultural, social, educational, political, and economic conditions as well as policies, practices, costs, and pricing in industries such as advertising, transportation, agriculture and others [3]. Health-related investment and policy decisions by government agencies, healthcare providers, insurers, and others stakeholders may lead to complex interactions and have far-flung consequences, many of which may be difficult to foresee.

Policies and interventions that will have reliable and effective impact therefore require changes to a large set of interconnected systems. For instance, in the 1970's, Finland set out to make broad reforms aimed at substantially changing the smoking habits and diet of its population. The government instituted a program of community socialization, legal reforms, and regulatory changes, such as reducing the allowable amount of fat content in sausages and providing agricultural incentives to encourage growing more fruits and vegetables and less meat. Over about 20 years, this effort resulted in substantial improvements in morbidity and mortality, including a 50% reduction in coronary artery disease [1] [2]. The Finland case is a successful example of a broad intervention and multiple policy changes across a significant population. In general, however, such wide-scale, long-term, real-world experiments are difficult to do, may not be effective, and may have unsuspected and unintended consequences and costs [4]. For example, health plans that limit drug choices, intending to prevent unnecessary prescriptions for expensive drugs, can in fact raise overall medical costs [5], or a program meant to bolster small rural hospitals can actually result in a net loss of rural hospital beds as they downsize to qualify for Critical Access Hospital designation [6].

Science and technology are stepping up to the challenge of supporting the critical health decisions we face. Recent efforts include development of robust infrastructures for improved data collection, reporting, and analysis, such as the system being developed by the State of Vermont to support community health teams [7]. To help assess the comparative effectiveness of alternative health policies, numerous efforts are being devoted to developing ever more accurate "deep domain" models of various aspects of health. Under the aegis of the Models of Infectious Disease Agent Study (MIDAS), for example, networks of researchers with expertise in computational models related to the outbreak and spread of infectious diseases collaborate on an ongoing basis to advise decision makers [8].

Nevertheless, today there are no reliable and manageable means to explore the health-system consequences of many interacting subsystems. Although individual deep-domain models can be made more comprehensive [9], there will always be important factors that lie outside the expertise of a given group of modelers, and it can be difficult to nimbly adapt very large and complex models as new questions arise. **Splash** addresses this problem by creating a framework for integrating disparate individual models to create effective and useful composite models. But providing a framework alone will not address the problem adequately: We need to enable a community of disparate stakeholders – those with health-related data, deep domain models, and health policy issues or questions – to work together, using the platform to make progress and solve problems. The community must be set up as an effective service system, creating more value through interaction than through isolation [10].

### 2.2 Models of Large-Scale, Complex Health Ecosystems

Models are simplified representations of real-world systems that help us understand these systems and answer "what-if" questions about them. Mathematical models use equations to describe system components and the relationships between them, and express the output or behavior of a system given a set of input variables. These outputs can depend on the inputs in a deterministic or stochastic manner. Statistical models can often be viewed as functions that are learned from data, and relate input variables, such as age, exercise level, and caloric intake, to output variables, such as body-mass index (BMI) [11]. Such models are often useful for discovering existing trends, patterns, and relationships. Although statistical models can also be used for predictive purposes, the underlying assumption of such models is that past trends can predict future behavior, which may not be appropriate when trying to predict the effects of hypothetical interventions [12]. Deterministic or stochastic models derived from first principles, as those used in classical physics, are typically more useful for prediction. For first-principles models of complex systems, especially dynamic systems that evolve over time, the outputs usually cannot be computed analytically, and the model must be implemented in a computer program and solved using numerical or Monte Carlo techniques [13][14]. Examples include deterministic differential-equation models of disease progression [17], stochastic discrete-event simulation models of resource scheduling in emergency rooms [15], system dynamics models of patient flow through a regional healthcare system [16], and agent-based models of consumer eating behavior [19]. We refer to such models generically as "simulation models." To produce credible results, simulation models need to be parameterized with real world data, which is challenging because data are often inconsistent, heterogeneous, widely distributed, and embedded in text or semi-structured documents. In the health setting, data privacy is also a key concern [18].

## 3 The Splash Research Agenda

### 3.1 Principled Composition of Models

Understanding complex health issues will almost certainly require combining multiple deterministic and stochastic simulation models, as well as statistical models and data sources, to project the effects of policy or investment choices into the future. For example, a comprehensive understanding of obesity – one that will help inform the creation of effective policies to fight obesity – will require models of many real world systems, including food systems (which in turn include agriculture, processing, transportation, distribution, climate, etc.), and social, economic, and physical systems in which consumers and food services operate, among others [3]. In **Splash**, we focus primarily on integrating and composing, i.e., "mashing up," simulation and statistical models aimed at informing health policy decision making, but we expect our model-mashup technology and methods will be applicable to other complex problem domains as well.

There are many reasons why diverse models are rarely combined to create a comprehensive, detailed picture of a large-scale complex system. Different categories of models are constructed, maintained, and used by different people and organizations, each using different terms, conventions, and approaches. If models are not too diverse or too complex, it might be feasible to manually and tightly integrate them into a common framework, as in the STEM project [17] or the High Level Architecture (HLA) specification of the Department of Defense [20]. But if models are very heterogeneous and complex, it might not be feasible to combine them at all. We believe there are many scenarios that lie between the two extremes, in which models cannot be tightly integrated into a single common framework — because of their size, heterogeneity, and complexity — but in which it might be feasible to automatically or semi-automatically compose models via a looser coupling, perhaps mediated by sets of input and output data. We are building such a platform for model composition in **Splash.** This approach avoids the need for modelers to adhere to a rigid common interface, which can hinder leveraging prior work and impede future collaboration.

### 3.2 Major Research Challenges

We are currently addressing four main challenges to composing large-scale models for complex health ecosystems.

- *Not all models can be combined in a sensible way.* The assumptions, time scales, capabilities, level of detail, and indeed the selection of the key aspects to represent may be quite different: What factors characterize the models that are compatible with one another?
- *There exists no standard way to describe models in sufficient depth to determine compatibility*. Here, the challenge is to create mechanisms and methods for describing models so that it is easy to determine how to integrate them into larger, more complex models of larger, more complex systems. The goal is for the **Splash** platform to semi-automatically identify models that are potentially compatible with a specified model, perhaps after some transformation of model inputs and outputs.
- *There are no tools or platforms to support mashing up independently created models and datasets in a simple, flexible, and useful way.* This adds the challenge of providing efficient mechanisms for searching and identifying applicable models, for establishing an appropriate execution environment, for semi-automatically generating connectors between models and between models and datasets, and for enabling reuse, result pruning, data transformations, flexible model transformations, experiment management, visualization, simulation output analysis, and so on (see Figure 1).
- *There is no targeted technology and set of practices to facilitate collaboration between the varied people and organizations that develop and use deep-domain models.* We envision an active community of participants contributing models and data, combining models, discussing models, exploiting previous results, and optionally sharing their models and modeling results. Participants in such an open community must have the means to (a) combine their proprietary models and data securely without risking intellectual property or violating privacy, (b) evaluate the quality of models and transformations used and communicate their findings to others, and (c) assess the trustworthiness of the outcomes produced. These matters demand sensitivity to the ways the underlying assumptions, methods and goals of varying communities of experts inform their practices and expectations. The final challenge is to develop a deep understanding of what is required for such an open integrated community system to successfully enable cooperation among all

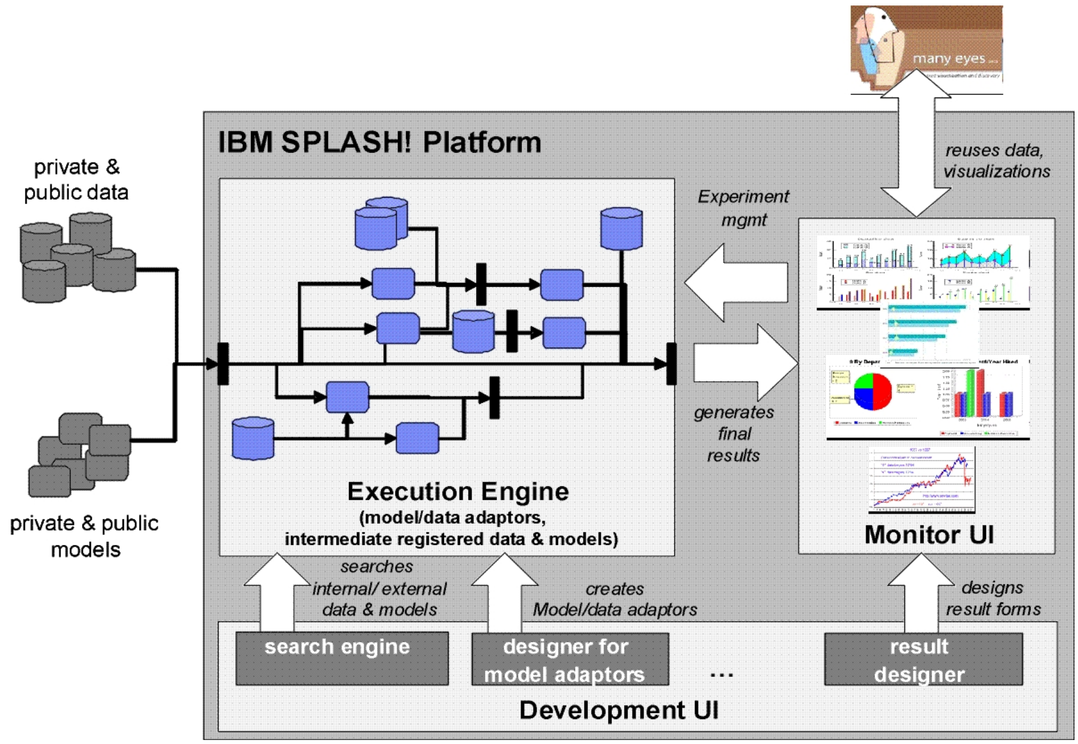


**Figure 1. Splash** will be an open community platform in which proprietary and public models, data, and outcomes can be searched, combined, executed, visualized, and shared.

stakeholders, and to incorporate our resulting insights into the design of our platform and methods.

In summary, there are conceptual challenges related to the nature and description of models, technical challenges related to the construction of a platform for automated model composition, and social and business challenges related to enabling diverse partners and stakeholders to work together effectively. We are addressing all of these challenges in **Splash.** Our technical approach builds on and extends a variety of existing technologies for web-service descriptions and mashups, data integration and warehousing, collaborative data analysis, and simulation model interoperability.

# 4 The Splash Roadmap

## 4.1 Phase One

In Phase One of **Splash**, we are creating and building specific model-mashup scenarios to gain insight into how to address the four challenges described previously. We have begun with a set of models related to a specific obesity scenario – public policy decisions regarding food availability's impact on a community's BMI.

To develop a deep understanding of model compatibility, we have identified a diverse set of models related to this problem (see Figure 2):

> **Household model**: An agent-based model in which each household has specific characteristics (income, food preferences, young married couple with 2 children, available transportation, etc.), and decides where to shop, which foods to purchase, and how much to exercise.
>
> **Transportation model**: A model providing travel time required between two input locations (such as household to nearby stores).
>
> **Store Location model**: An optimization model used by retail stores (in this case a food store) to select optimal locations for establishing a new store.
>
> **Recreation Model**: A queuing model for usage of parks and recreation facilities.
>
> **Calories to BMI model**: A statistical model that converts daily calories, fat, etc. consumption and physical activity to BMI change for a given individual (e.g. 6 foot tall 45 yr old man weighing 250 pounds).
>
> **Community Incentives model**: A decision tree of incentives that a community might offer to a food store to locate at a specific site.

Integrating these models provides the opportunity to have scientifically sound answers to questions such as:

- If a large chain grocery with low-cost fruits and vegetables opens at a specific location, what is the impact on the community obesity rates as measured by BMI?
- In terms of BMI impact, would it be better to locate a recreation facility at a certain location rather than a grocery store?
- Would a given community have lower obesity rates if it attracted a grocery store with many prepared food options versus a traditional restaurant to a certain location?

In mashing up these models we expect to leverage and extend results from the simulation interoperability literature [21], [22] — which is concerned with tightly coupled models — to our setting, in which models are most likely communicate via file or database I/O, or via web-service calls. One approach is to view a simulation model as a (very complex) data-transformation service, and exploit prior work on web-service description languages [23] and on data transformation [24]. It may also be possible to adapt constructs from existing model description languages such as the Unifed Modeling Lanaguge (UML) [25] used in software engineering and Systems Modeling Langauge (SysML) [26] used to specify embedded systems.

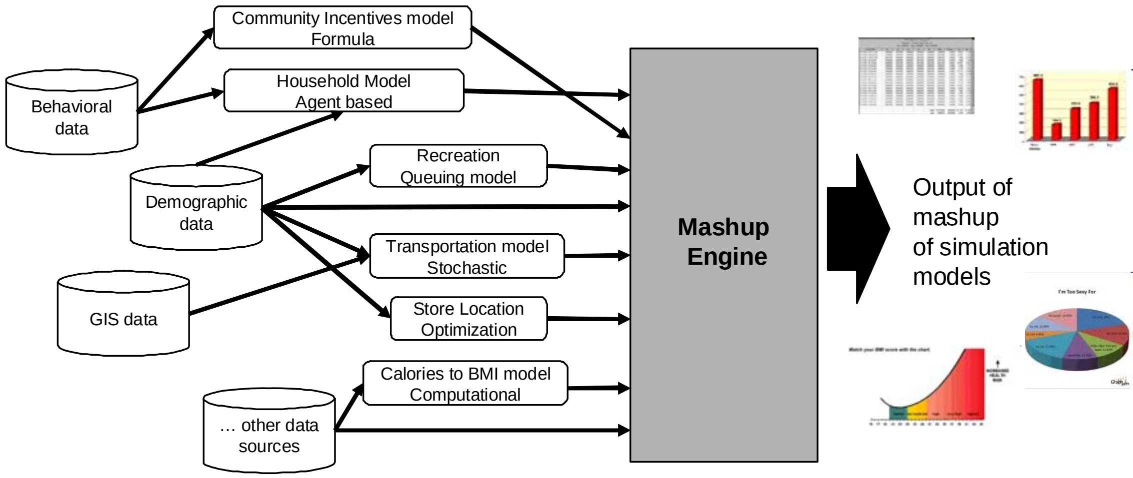


**Figure 2:** Phase One: Obesity model mashup. A variety of types of models and sources of data are combined to inform community planning: If a locality aimed to decrease obesity, how should it invest -- in recreation facilities, in store incentives, in advertising?

As illustrated in Figure 2, these obesity-related models require data from a variety of sources such as geographical information (e.g. street maps, location of current grocery stores and parks), demographic information (e.g. income, current physical profiles of households), food information (e.g. food choices, calories consumed), and community information (e.g. available locations for stores, available recreation facilities and capacities).

Beyond initial data sources required by one or more models, intermediate data transformations and connectors between models are essential to performing this model mashup. To do this, we are exploiting and extending technology developed for mashing up data sources, such as IBM Mashup Center [27]. Extensions include intermediation for differences in time (e.g. households may do grocery shopping once per week while the BMI model expects daily calorie input.). Besides time conversions, GIS conversions may also be needed between, say, the store locator model and the transportation model or the household model. Some models may produce a time series of data, while another expects only single event data. In this case, the "connector" is a sequence of multiple calls to the second model, one for each time series element. A language such as BPEL may be suitable for designing such process-type connectors [27].

We expect to complete this example model mashup in Phase One of the **Splash** project by the end of 2010, giving us a prototype platform (see Figure 1) and allowing us to support additional models and connectors for other stakeholders in Phase Two during 2011.

### 4.2 Phase Two

Phase Two of the **Splash** project will begin in 2011 and will aim to incorporate more models and more data from a community of stakeholders. Specifically, there are many research questions surrounding the community aspects of integrating disparate models to understand health and health policy. One surrounds the worldviews and social practices that inform the underlying assumptions of the models and the expectations of their application. In practical terms, we must understand how to bring together contributors in ways that will support fruitful collaboration across diverse communities of experts and that lead to the production of meaningful and useful outputs to create an effective service system [10]. What enabling technologies and sets of practices will support the kinds of knowledge-production and decision-making requirements of users? What operational considerations of the environment are needed to encourage experimentation and derive novel theories to help address the complex challenges of health and health policy?

Furthermore, to build a system in which a community of stakeholder can work together effectively, we will extend the Phase One prototype platform with a number of additional key features, including:

- Formal model description language for community participants to use to describe their models
- Catalog of models and data with search capabilities for identifying required data and single or composite models of interest. One example of building such a community of participants is illustrated by the ManyEyes web site for data sets and visualizations [29]. In ManyEyes, participants may register, comment, blog, and provide ratings on data sets. Provenance of data is explicitly described in bylines associated with data. One path of exploration for the cataloging work **Splash** will provide is to apply similar concepts as we build our community of data, simulation and statistical models
- Semi-automatic model compatibility assistant
- Library of connector mechanisms, including tooling for adding connectors or customizing an existing connector
- Mechanisms for describing how models can be put together to answer more complex questions and the implementation of those constructs in the platform
- Security and privacy mechanisms for protecting individual privacy and for protecting proprietary models and data

The success of the Splash platform and community depends on harnessing collaboration and expertise of researchers, health services professionals, decision makers, and other stakeholders in public and private sectors. We welcome participation of potential partners, collaborators, service-providers, and consumers in Splash.

## 5 Summary

We are embarking on an effort to provide scientific support for policy and investment choices affecting health by enabling stakeholders to try out possible alternatives via simulation. Because health involves a large set of intricately related real-world systems, adequately capturing the complex dynamics of the health ecosystem requires the composition of many simulation and statistical models created by a diverse set of deep-domain experts. In **Splash,** we are developing an open community platform, method, and service to enable integration of independently created, deep-domain simulation models, statistical models, and datasets in a manner that is practical, flexible, cost-effective, attractive, and usable. Our approach exploits and extends current technologies for web-services description and mashups, data integration and warehousing, collaborative data analysis, and simulation model interoperability, as well as insights from the emerging field of service science.

We expect **Splash** will ultimately have impact on health at multiple levels, ranging from understanding comparative effectiveness of treatments and preventions, to determining return on investment at an ecosystem level, to understanding global health consequences of policy and investment decisions. Once completed, we expect the approach to be applicable to decision-making in other complex domains as well.

## 6 Notes


[1] Gordon, T. Mortality in the United States, 1900-1950. *Public Health Rep* 68(4):441-444, 1953.

[2] Collins JL, Marks JS, Koplan JP. Chronic disease prevention and control: coming of age at the Centers for Disease Control and Prevention. *Prev Chronic Dis* 6(3), 2009. Available at http://www.cdc.gov/pcd/issues/2009/jul/08_0171.htm.

[3] Huang TT, Drewnowski A, Kumanyika SK, Glass TA. A systems-oriented multilevel framework for addressing obesity in the 21st century. *Prev Chronic Dis* 6(3), 2009. Available at http://www.cdc.gov/pcd/issues/2009/jul/09_0013.htm.

[4] Institute of Medicine and Robert Wood Johnson Foundation – *Workshop on Unintended Consequences of Health Policy Programs and Policies* 2001. Available at http://books.nap.edu/openbook.php?record_id=10192&page=15#p200049679970015001.

[5] Sterman, JD. Learning from Evidence in a Complex World, *Am J Public Health*, Jan. 2006.

[6] Rural Assistance Center, U.S. Department of Health and Human Services' Rural Initiative. Critical Access Hospitals. Available at http://www.raconline.org/info_guides/hospitals/cah.php.

[7] Vermont Blueprint for Health: Smart Choices, Powerful Tools. 2009 Annual Report. January, 2010. Vermont Department of Health. Available at http://healthvermont.gov/prevent/blueprint/documents/Blueprint_AnnualReport_2009_0110rev.pdf .

[8] Models of Infectious Disease Agent Study. National Institute of General Sciences, National Institutes of Health. Available at http://www.nigms.nih.gov/Initiatives/MIDAS/.

[9] Envision Project. National Collaborative on Childhood Obesity Research. Available at http://www.nccor.org/projects_envision.html.

[10] Spohrer J, Maglio PP, Service science: Toward a smarter planet. In W Karwowski & G Salvendy (Eds.), *Introduction to Service Engineering.* 2009. New York: Wiley & Sons.

[11] Navarro-Barrientos JE, Rivera DE, Collins LM. A dynamical systems model for understanding behavioral interventions for weight loss, *Proc Int Conf Social Computing, Behavioral Modeling, and Prediction (SBP)*, 2010.

[12] Pearl J. *Causality: Models, Reasoning, and Inference.* 2000. Cambridge University Press.

[13] Law AM. *Simulation Modeling and Analysis*. 2007. 4th Edition. McGraw-Hill.

[14] Robert CP, Casella G. *Monte Carlo Statistical Methods.* 2004. 2nd Edition. Springer.

[15] Alexopoulos C, Goldsman D, Fontanesi J, Sawyer M, Guire MD, Kopald D, Holcomb K. A discrete-event simulation application for clinics serving the poor, *Proc Winter Simul Conf*, pp. 1386–1391, 2001.

[16] Lane DC, Husemann E. System dynamics mapping of acute patient flows. *J Operational Res Soc*, 59(2):213-224, 2008.

[17] Ford DF, Kaufman JK, Eiron I. An extensible spatial and temporal epidemiological modeling system. *Int J Health Geographics* 5(4), 2006.

[18] Rindfleish TC. Privacy, information technology, and health care. *Comm ACM* 40(8) 92–100, 1997.

[19] Hammond, RA. *A Complex Systems Approach to Understanding and Combating the Obesity Epidemic.* Washington, DC: Brookings Institution, 2008.

[20] IEEE 1516-2000 - *Standard for Modeling and Simulation High Level Architecture - Framework and Rules.*

[21] Jain S, McLean CR: Integrated simulation and gaming architecture for incident management training, *Proc Winter Simul Conf,* pp. 904–913, 2005.

[22] Kuhl F, Weatherly R, Dahmann J. *Creating Computer Simulation Systems: An Introduction to the High Level Architecture*. 1999. New Jersey:Prentice Hall.

[23] Chinnich R, Moreau J, Ryman A, Weerawerana S. *Web Services Description Language (WSDL) Version 2.0 Part 1: Core Language.* W3C Recommendation, 2007. Available at http://www.w3.org/TR/wsdl20 .

[24] Fagin R, Kolaitis PG, Miller RJ, Popa L. Data exchange: semantics and query answering. *Theoret Comput Sci* 336(1), 2005, 89-124.

[25] Unified Modeling Language – UML. Available at http://www.uml.org/.

[26] Systems Modeling Langauge – SysML. Available at http://www.omgsysml.org/.

[27] IBM Mashup Center. Available at http://www-01.ibm.com/software/info/mashup-center/.

[28] Farahbod R, Glasser U, Vajihollahi M. Specification and Validation of the Business Process Execution Language for Web Services. *Proc ASM 2004*. LNCS, vol. 3052, pp. 78–94. Springer, Heidelberg, 2004.

[29] Viégas FB Wattenberg M, van Ham F, Kriss F, McKeon M: ManyEyes: A site for visualization at Internet scale. *IEEE Trans Visualization Computer Graphics* 13(6):1121-1128, 2007.