Data is Dead… Without What-If Models

Peter J. Haas Paul P. Maglio Patricia G. Selinger Wang-Chiew Tan
IBM Research — Almaden IBM Research—Almaden & UC Santa Cruz
{phaas, pmaglio, patseli, wangchiew}@us.ibm.com

ABSTRACT

Current database technology has raised the art of scalable *descriptive* analytics to a very high level. Unfortunately, what enterprises really need is *prescriptive* analytics to identify optimal business, policy, investment, and engineering decisions in the face of uncertainty. Such analytics, in turn, rest on deep *predictive* analytics that go beyond mere statistical forecasting and are imbued with an understanding of the fundamental mechanisms that govern a system's behavior, allowing what-if analyses. The database community needs to put what-if models and data on equal footing, developing systems that use both data and models to make sense of rich, real-world complexity and to support real-world decision-making. This model-and-data orientation requires significant extensions of many database technologies, such as data integration, query optimization and processing, and collaborative analytics. In this paper, we argue that data without what-if modeling may be the database community's past, but data with what-if modeling must be its future.

1. INTRODUCTION

In the beginning, there were data transactions and simple reports. Then came the relational model, SQL, and high-performance relational DBMSs to run transactions and generate simple reports in an elegant manner. This primeval form of descriptive analytics was enhanced with OLAP, data mining, and other business-intelligence technologies as enterprises realized that there was valuable information to be extracted from transactional data. Since then, DBMSs have expanded their capabilities to handle semi-structured data, unstructured text, web-based data, semantic data, uncertain data, and streaming data at scales approaching the Exabyte range. DBMS functionality has expanded to include simple programming within the database, various statistical analyses, and even more recently, machine learning techniques.

These are tremendous achievements, and we are justifiably proud of what we have accomplished as a database community. We can perform all kinds of descriptive analytics over all kinds of data at scale. But with a combined total of more than eight decades in the field of database research, the authors have come to a set of realizations, leading us to the conclusion that our focus on data is much too narrow and must be expanded dramatically.

Proceedings of the VLDB Endowment, Vol. 4, No. 12

2. DATA-CENTRISM IS WRONG

Enterprises need to make decisions, typically related to allocating scarce resources, such as money, equipment, food, medicine, people, or time, and usually in the face of uncertainty. The science of better decision-making has traditionally been the domain of Operations Research and related fields, where a wide variety of technologies for deterministic and stochastic optimization have been developed. The Institute for Operations Research and Management Science (INFORMS) recently promoted the term *prescriptive analytics* to describe such optimization methods [14]. Prescriptive analytics methods need to assess the consequences of alternative design, investment, or policy choices on the system of interest, and hence rest on "deep" *predictive analytics* that allow such what-if analyses. Arguably, most of the analytics technology developed by the database community—from simple querying to scalable machine learning—has been driven by the need to support enterprise decisions. But how far have we come in satisfying this need?

This brings us to our realizations about data.

Realization #1: Data is dead. By definition, data reflects facts or assertions of facts that are already in existence: the world as it is or was. Pat purchased red shoes for $50 on January 31. A newspaper article describes President Obama's speech to the United Nations on May 2. The sky telescope traversed a section of a galaxy and recorded a set of images. This is history, accomplished and recorded. Data just lies passively. Applications can look at this data, add it up, roll it up, cube it, summarize it, compare it, filter it, or join it with other data, but it is still a record of history, done deeds.

Being dead is not necessarily bad. Many very useful things are dead (or non-living, anyway) but extremely valuable—houses, books, soup, and so on. We learn from the past; we leverage what we have built and what we know. But that doesn't change the fact that data is a record, not a conclusion or an insight or a solution.

Realization #2: Descriptive analytics (including shallow predictive analytics) are last resorts for decision-making. Descriptive analytical techniques, including simple querying, OLAP, data mining, and machine learning, are powerful tools for finding important patterns and relationships in existing data. They help us build insight into the real world. We can use analytics to find clusters of disease, to identify possible drug side effects and interactions, or to determine which stores sold more produce in March. Analytic processing, both within and outside the database engine itself, can make businesses more efficient, medical treatments safer and more effective, and so on. As discussed, the database community has an excellent record of developing technology for descriptive analytics.

But data alone—even with very powerful descriptive analytics—tells us about the world as it is, and was, but cannot tell us much

about the world as it might be. To deal with an uncertain future, decision makers sometimes resort to "shallow" predictive analytics, by which we mean statistical techniques that simply extrapolate into the future the patterns and relationships observed in historical or training data. Examples include time-series forecasting models for home values or neural-network classifiers that predict whether a borrower will default on a home mortgage. Such predictions work well only if the future is fundamentally like the past, e.g., if historical home-value trends continue or if home-loan borrowers continue to behave as they have in the past. This approach is notoriously brittle: shallow predictive models do not allow reliable extrapolation to scenarios other than those that produced the historical or training data, and this lack of what-if capability can result in disastrous consequences when used as a basis for decision making. We resort to descriptions and extrapolations only when we lack access to domain experts.

Realization #3: We can understand so much more if we move from descriptive to deep predictive analytics that are model- and data-driven. Robust decisions are based on a thorough understanding of first principles, causes, and interconnections between system elements. This goal of understanding is the province of deep predictive analytics, and usually requires domain expertise. Take weather prediction. Descriptive analytics, such as time-series modeling, can tell us about historical weather patterns and perhaps be used to forecast the amount of rainfall over the next couple of days. Such techniques alone cannot help us predict the effects of, say, changes in air pollution laws upon climate. Analysis of such complex what-if questions calls for the sorts of weather models and simulations developed over decades by experts in hydrology, atmospheric science, and so on. These dynamic models are built from first principles (the equations of Newtonian mechanics, thermodynamics, radiative transfer, etc.), and embody deep knowledge and expertise. In general, good decision making rests on what-if models built from an understanding of causes and effects.

Many systems, such as sales, weather, transportation, biology, and healthcare, have been studied individually by domain experts using first-principles simulation models. This is all really a part of a bigger life cycle: data and information feed experts who form hypotheses, create models based on domain knowledge (science and engineering, economics, etc.) and data, validate the models against data, and use the models for prediction. Modeling exercises, in turn, provide guidance and direction in obtaining further data. Expert models developed via this cycle yield better accuracy than just statistical methods against plain data. We could not, for example, derive a weather prediction model by merely analyzing weather data, nor should we try, because weather experts already have tremendous knowledge that can be exploited.

Of course, data analysis can point the way for experts to look deeper; e.g., spot pancreatic cancer clusters so that experts can explore whether they are caused by groundwater chemicals, genetic components of disease, or other factors. But experts know so much more than can be discovered in the data or represented by even the best semantic descriptions and ontologies. Looking at data alone is a last resort, an admission that we have no other knowledge. Projecting sales of red shoes in Tucson in May via analytics may be made even more accurate and even more understandable if combined with the psychology of buying habits through agent-based modeling or a deep model of the impact of media and advertising.

So the data work we've all focused on for 30+ years is just half the story. As a community, we must incorporate models—not merely R scripts and user-defined functions—into our thinking, aiming ultimately to provide a more complete picture of how things work.

Realization #4: This is especially true when trying to solve complex problems involving systems of systems. No one expert understands the workings of large complex systems that emerge from the interactions of already complex systems. A prime example is the health system, which includes not only healthcare and treatments but also advertising, education, agriculture, transportation, economics, government policies, and so on [6]. Perhaps the best way to make sense of these is to combine models into larger composite system models (e.g., [4]), just as today we combine data in joins, subqueries, mashups, cubes, and so on.

To do this, the key challenge is to facilitate integration of datasets, along with existing heterogeneous simulation models, statistical models, and optimization models, for the type of what-if analysis that underlies prescriptive analytics. One research question is whether such models-and-data integration can be made feasible, practical, flexible, cost-effective, and usable. The Splash research prototype under development at IBM is our first attempt to address this question more broadly.

3. THE MODELS-AND-DATA APPROACH

Splash is a platform for integrating multiple deep domain models (especially dynamic simulation models) and data sources, making it easy to perform what-if analyses on a complex system of systems [2]. Splash differs from standard simulation inter-operability techniques [2][9], in which models are simultaneously executed and tightly coupled. For these, detailed knowledge of the participating models, and sometimes significant changes to the code, are required to achieve interoperability. In contrast, models in Splash are loosely coupled, meaning they can execute independently and communicate mainly via data, either through files, databases, or web-service calls.

Consider a hypothetical policymaking scenario in which policymakers must understand which combination of interventions is most cost-effective for reducing obesity in the population. Specifically, the goal is to reduce population body-mass index (BMI) by a certain percentage. The possible considerations to reduce BMI include:

1. What if the transportation infrastructure in a certain neighborhood is improved?
2. What if a supermarket that sells healthy and reasonably-priced food is built at a specific location?
3. What if more exercise facilities are built in certain neighborhoods?

Clearly, these what-if questions involve diverse factors—transportation, shopping, and exercise facilities—and they are difficult to answer without deep predictive analytics [11]. In fact, even with these three types of interventions, many combinations of parameters are possible, and they must be analyzed systematically to identify effective strategies for reducing obesity. Splash can be used to help determine the right combination—all the way from constructing a system of interacting models and data to analyzing what-if scenarios.

For the obesity scenario above, we created a composite model in Splash that takes account of individual food and activity choices, neighborhood traffic patterns, and the relationship between calorie intake and BMI (see Figure 1). Specifically, the composite model integrates (a) VISUM [13], a commercial traffic flow simulation software package that can simulate different modes of public and private transport to determine the impact of traffic demand; (b) an agent-based model that simulates the grocery-store shopping behavior of households over time [1], taking account of food preferences, travel times to the grocery stores, and social factors, such as where neighbors shop; (c) a stochastic discrete-event simulation model of the use of exercise facilities; and (d) a Hall-Chow differential equation model of changes in BMI [12] based on daily food consumption and physical activity. The output of the composite model is a time series of BMI values for each population member. These time series can be aggregated, visualized, and subjected to statistical analysis.

In addition, we incorporated disparate data sources to feed the models for our what-if analyses. Specifically, we used: (a) GIS data for a specific urban area, including road networks and zone configurations; (b) demographic data containing information about each person in a household, including age, weight, height, as well as nutritional characteristics of the food sold at each store; and (c) facility data containing information about types of exercise facilities that are used by the exercise model.

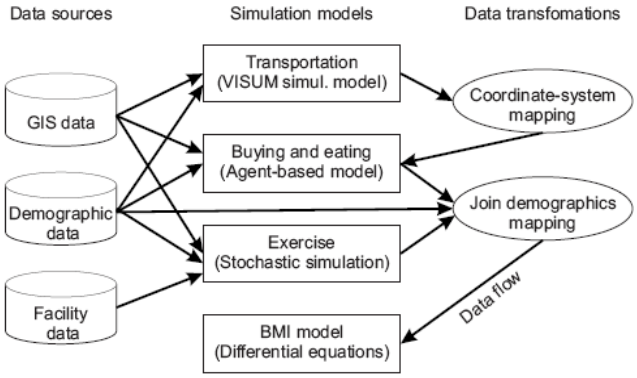

Figure 1. Our integration of four obesity-related models with three datasets in an effort to help policymakers understand how one should invest—in transportation, store incentives, exercise facilities, in order to reduce obesity.

In building the composite obesity model, we had to create a workflow of model execution and data flow, along with two specific data mappings to put the models and data sources together. (Simulation models can produce huge amounts of data, so scalability is an issue here—the prototype uses Hadoop to combine model outputs and demographic data.) To design data mappings in general, Splash currently relies on Clio++, an extension of the Clio [5] mapping-design tool. Clio was originally developed in the context of data integration and data exchange to generate a mapping between source and target schemas through a visual interface. The Clio enhancements were needed to generate *simulation-specific mappings* that can handle mismatches in time and space that commonly occur between different simulation models. For example, the BMI model expects a time series where each simulation tick means one day has passed, but in the output of the shopping model, each simulation tick means two days have passed. Clio++ therefore generates a mapping that performs a time interpolation of the shopping-model output. Similarly, a geospatial alignment is performed between the zones in the VISUM model and the grid coordinates of the shopping model.

To understand the semantics of models and to support both design-time model composition and run-time model execution, Splash uses metadata about models and datasets. Each model and data in the Splash repository is associated with a description. The description of a data source includes the schema, where the data is located, and other metadata such as ownership. The description of a model includes the type of simulation model, the input and output schemas of the model, the interpretation of time and space in the model, the invocation command, and so forth.

The Splash approach leverages the power of a community of experts and a range of individual modeling techniques, such as discrete-event modeling [10], agent-based modeling [3], or system dynamics modeling [15], to deal with complex systems of systems. There have actually been a couple of efforts in the database community to combine what-if models and data in a fundamental way [8][17], but these have focused on individual models and not composite modeling, and indeed could be integrated into the Splash framework.

4. RESEARCH CHALLENGES

Putting models and data on equal footing opens up many directions for research. Indeed, we see many basic research problems for creating models-and-data-oriented "databases" that natively support what-if analyses:

Searching for compatible models, data, and mappings. For different domain experts to collaborate on solving complex problems involving systems of systems, Splash must make it easy for each expert to search for models, data, and mappings that complement his or her area of expertise. Such a feature is crucial for the reuse of models, data, and mappings previously developed by the broad scientific and engineering communities. How can we help a domain expert easily find models, data, and mappings that are relevant and compatible, both functionally and semantically? The answers involve enhancements to (semantic) search technologies and repository management. Privacy and security issues are also relevant here.

Simulation-oriented mapping generation. Simulation models are usually dynamic, evolving over space and time, and are often stochastic as well. Integrating such models and data requires techniques that go well beyond traditional algorithms for data integration and mapping generation. We know how to describe ordinary datasets through schemas and how to semi-automatically generate mappings between schemas. Can we extend this technology to simulation-specific mappings? Such mapping systems would need to deal with issues such as time and space alignment, automatic matching of measurement units, hierarchical models with data at different resolutions, and complex data transformations—e.g., converting raw stochastic simulation outputs to histograms that represent a steady-state probability distribution of a system characteristic of interest.

Simulation-experiment optimization. The problem of efficiently performing a simulation experiment—executing models, transforming data, and analyzing results—can be viewed as a significant generalization of the query-optimization problem. Techniques are needed for automatically reconfiguring parts of a simulation-experiment workflow among (possibly distributed) data and models, factoring common operations across different mappings in the workflow, and avoiding redundant computations over different experimental runs. Dealing with statistical issues, such as management of pseudo-random numbers and Monte Carlo replications, adds additional complexity [18]. As discussed, scalability is a key issue as well.

Deep collaborative analytics. Different outcomes that are produced by integrating deep domain models and data must be visualized, analyzed, and discussed to build trust in the results. A key question is how to explain our visualizations to ourselves and others. For example, can we reasonably explain why average BMI for high-income households does not decrease over time? Can we extend technologies for collaborative data analytics such as ManyEyes [16] to handle collaborative modeling and analytics?

Causality. Another fundamental challenge is dealing with bidirectional causality between models. For instance, it may be feasible to approximate such causality by running models independently but periodically exchanging data, in the spirit of [17]. Providing (and justifying) such functionality poses both theoretical and system-design challenges.

5. CONCLUSION

Modern database technology effectively supports descriptive analytics. But deep predictive analytics—beyond statistical forecasting and based on understanding of mechanisms governing system behavior—are needed for complex decision-making, supporting prescriptive analytics and what-if analyses. Data plus models are what's needed today. And this requires significant extensions of database technologies. There is a clear opportunity now for the data community to redefine itself as the *models and data community*, reflecting the entire process of solution discovery and integration. Data is dead… without what-if models.

6. REFERENCES

[1] Auchincloss, A. H., Riolo, R. L., Brown, D. G., Cook, J. & Diez Roux, A. V., "An Agent-Based Model of Income Inequalities in Diet in the Context of Residential Segregation,"*Amer. J. Preventive Medicine*, 40(3), 303–311, 2011 .

[2] Cefkin, M., Glissmann, S., Haas, P. Jalali, L., Maglio, P. P.,Selinger, P., Tan, W. C., "Splash: A Progress Report on Building a Platform for a 360 Degree View of Health" in *5th INFORMS Workshop on Data Mining and Health Informatics*, Austin, TX, 2010. Available at https://informs.emeetingsonline.com/emeetings/formbuilder/clustersessiondtl.asp?csnno=14057&mmnno=201&ppnno=49296

[3] Chan, W.K.V., Son, Y.-J., and Macal, C.M., "Agent-based simulation tutorial: Simulation of emergent behavior and differences between agent-based simulation and discrete-event simulation," *Proc. Winter Simulation Conference*, pp. pp. 135–150, 2010.

[4] Godfray, H. C. J., Pretty, J., Thomas, S. M., Warham E. J. & Beddington, J. R., 2011, "Linking Policy on Climate and Food," *Science*, 331(6020), 1013–1014, 2011.

[5] Haas, L. M., Hernández, M.A., Ho, H., Popa, L., Roth, M., "Clio Grows Up: From Research Prototype to Industrial Tool," *Proc. ACM SIGMOD*, pp. 805–810, 2005.

[6] Huang, T. T, Drewnowski, A., Kumanyika, S. K., & Glass, T. A.,"A Systems-Oriented Multilevel Framework for Addressing Obesity in the 21st Century," *Prev. Chronic Disease*, 6(3), 2009.

[7] Jain S., and McLean C. R., "Integrated simulation and gaming architecture for incident management training," *Proc.Winter Simulation Conference*, pp. 904–913, 2005.

[8] Jampani, R., Perez, L., Wu, M., Xu, F., Jermaine, C., and Haas, P.J., "MCDB: A Monte Carlo approach to managing uncertain data," *Proc. ACM SIGMOD Intl. Conf. Management of Data,* pp. 687–700, 2008.

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

[10] Law, A.M., *Simulation Modeling and Analysis,* 4th Edition, McGraw-Hill, 2007.

[11] Levy, D. T., Mabry, P. L., Wang, Y. C., Gortmaker, S., Huang, T. T.-K., Marsh, T., Moodie, M., & Swinburn, B., 2010, "Simulation models of obesity: a review of the literature and implications for research and policy," *Obesity Rev.*, 12(5), 378–394, 2011.

[12] Navarro-Barrientos J. E., Rivera D. E., Collins L. M., "A dynamical systems model for understanding behavioral interventions for weight loss," *Proc. Int. Conf. Social Computing, Behavioral Modeling, and Prediction (SBP)*, Springer Lecture Notes in Computer Science 6007, pp. 170–279, 2010.

[13] Planung Transport Verkehr AG, VISUM. http://www.ptvag.com/software/transportation-planning-traffic-engineering/software-system-solutions/visum/.

[14] Robinson, A., Levis, J., and Bennett, G., 2010, "INFORMS to officially join analytics movement." *ORMS Today,* 37(5), October, 2010.

[15] Sterman, J.D., *Business Dynamics: Systems Thinking and Modeling for a Complex World,* McGraw-Hill/Irwin, Boston, 2000.

[16] Viégas, F., B. Wattenberg, M., van Ham, F., Kriss, J., and McKeon, M., "ManyEyes: A site for visualization at Internet scale," *IEEE Trans. Visualization and Computer Graphics*, 13(6), 1121–1128, 2007.

[17] Wang, G., Vaz Salles, M.A., Sowell, B., Wang, X., Cao, T., Demers, A.J., Gehrke, J., and White, W.M., 2010, "Behavioral Simulations in MapReduce," *PVLDB* 3(1), 952–963, 2010.

[18] Xu, F., Ercegovac, V., Haas, P.J., and Shekita, E., "$E = MC^3$: Managing uncertain enterprise data in a cluster-computing environment," *Proc. ACM SIGMOD Intl. Conf. Management of Data*, pp. 441–454, 2009.