DATA IS DEAD …
WITHOUT “WHAT-IF” MODELS

Peter J. Haas, Paul P. Maglio, Patricia G. Selinger, and Wang-Chiew Tan
IBM Almaden Research Center
Congratulations, Database Community!

Transactions & Reports, IMS

OLAP

Relational model & SQL

Semi-structured & Unstructured text

Statistical analysis

Data mining

Text analytics

Web data

Streaming data

Machine learning

Massive Data/Cloud DB

Uncertain data

Semantic data
Congratulations, Database Community!

BUT: Why do enterprises care about data in the first place?
Because enterprises need to make DECISIONS

Allocation of scarce resources

Overview

The Analytics Section of INFORMS is focused on promoting the use of data-driven analytics and fact-based decision making in practice. The Section recognizes that analytics is seen as both (i) a complete business problem solving and decision making process, and (ii) a broad set of analytical methodologies that enable the creation of business value. To this purpose, the Section promotes the integration of a wide range of analytical techniques and the end-to-end analytics process. It will support activities that illuminate significant innovations and achievements in specific steps and/or in the execution of the process as a whole, where success is defined by the impact on the business.

We recognize that analytics is defined by three categories:

Descriptive analytics
- Prepares and analyzes historical data
- Identifies patterns from samples for reporting of trends

Predictive analytics
- Predicts future probabilities and trends
- Finds relationships in data that may not be readily apparent with descriptive analysis

Prescriptive analytics
- Evaluates and determines new ways to operate
- Targets business objectives
- Balances all constraints
Because enterprises need to make decisions

Allocation of scarce resources

"Analytics is...a complete business problem solving and decision making process"

Overview

The Analytics Section of INFORMS is focused on promoting the use of data-driven analytics and fact-based decision making in practice. The Section recognizes that analytics is seen as both (i) a colligate business problem solving and decision making process, and (ii) a broad set of analytical methodologies that enable the creation of business value. To this purpose, the Section promotes the integration of a wide range of analytical techniques and the end-to-end analytics process. It will support activities that illuminate significant innovations and achievements in specific steps and/or in the execution of the process as a whole, where success is defined by the impact on the business.

We recognize that analytics is defined by three categories:

Descriptive analytics
- Prepares and analyzes historical data
- Identifies patterns from samples for reporting of trends

Predictive analytics
- Predicts future probabilities and trends
- Finds relationships in data that may not be readily apparent with descriptive analysis

Prescriptive analytics
- Evaluates and determines new ways to operate
- Targets business objectives
- Balances all constraints
Because enterprises need to make decisions


"Analytics is...a complete business problem solving and decision making process"

Descriptive Analytics: Finding patterns and relationships in historical and existing data
Because enterprises need to make decisions

“Analytics is...a complete business problem solving and decision making process”

Descriptive Analytics: Finding patterns and relationships in historical and existing data

Predictive analytics: predict future probabilities and trends to allow what-if analysis

Allocation of scarce resources

Overview
The Analytics Section of INFORMS is focused on promoting the use of data-driven analytics and fact-based decision making in practice. The Section recognizes that analytics is seen as both (i) a cognitively business problem solving and decision making process, and (ii) a broad set of analytical methodologies that enable the creation of business value. To this purpose, the Section promotes the integration of a wide range of analytical techniques and the end-to-end analytics process. It will support activities that illuminate significant innovations and achievements in specific steps and/or in the execution of the process as a whole, where success is defined by the impact on the business.

We recognize that analytics is defined by three categories:

Descriptive analytics
- Prepares and analyzes historical data
- Identifies patterns from samples for reporting of trends

Predictive analytics
- Predicts future probabilities and trends
- Finds relationships in data that may not be readily apparent with descriptive analysis

Prescriptive analytics
- Evaluates and determines new ways to operate
- Targets business objectives
- Balances all constraints
Because enterprises need to make decisions. Allocation of scarce resources.

"Analytics is...a complete business problem solving and decision making process."

**Descriptive Analytics**: Finding patterns and relationships in historical and existing data.

**Predictive analytics**: Predict future probabilities and trends to allow what-if analysis.

**Prescriptive analytics**: Deterministic and stochastic optimization to support better decision making.
Shallow versus deep predictive analytics

Extrapolation of 1970-2006 median U.S. housing prices

NCAR Community Atmosphere Model (CAM)

3.3 Eulerian Dynamical Core
\[
\begin{align*}
\frac{\partial C}{\partial t} &= k \cdot \nabla \times (n/\cos \phi) + F_{\text{HI}}, \\
\frac{\partial \delta}{\partial t} &= \nabla \cdot (n/\cos \phi) - \nabla^2 (E + \Phi) + F_{\text{HI}}, \\
\frac{\partial T}{\partial t} &= -\frac{1}{a \cos^2 \phi} \left[ \frac{\partial}{\partial \lambda} \left( UT \right) + \cos \phi \frac{\partial}{\partial \phi} \left( VT \right) \right] + T \delta - \frac{\partial^2 \delta}{\partial \eta^2} + \frac{R_{\text{HI}} \omega}{c_p} + Q + F_{\text{HI}} + F_{\text{HI}}, \\
\frac{\partial q}{\partial t} &= -\frac{1}{a \cos^2 \phi} \left[ \frac{\partial}{\partial \lambda} (U q) + \cos \phi \frac{\partial}{\partial \phi} (V q) \right] + q \delta - \frac{\partial q}{\partial \eta} + S, \\
\frac{\partial n}{\partial t} &= \int_1^n \nabla \cdot \left( \frac{\partial n}{\partial \eta} \right) d\eta.
\end{align*}
\]
Is the DB community truly helping decision makers?

Some realizations...
Data is **dead**

<table>
<thead>
<tr>
<th>Name</th>
<th>Item</th>
<th>Price</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pat</td>
<td>Red shoes</td>
<td>$50</td>
<td>1/23/11</td>
</tr>
</tbody>
</table>

...a record of history that says nothing about future or hypothetical worlds
Descriptive analytics & shallow predictive analytics are *last resorts* for decision making

(When you can’t find the domain experts)

...but are the main focus of most database and IM technology
We can understand much more by moving to deep predictive analytics based on **models and data**
We can understand much more by moving to deep predictive analytics based on **models and data**

Data-centrism is **WRONG**: Exploit expert knowledge of fundamental **structure, causal relationships, and dynamics** of system constituents to create first-principles simulation models
Especially true for complex systems-of-systems


Challenge: Facilitating integration of existing simulation models, statistical models, optimization models, and datasets for what-if analysis
Especially true for complex systems-of-systems


Challenge: Facilitating integration of existing simulation models, statistical models, optimization models, and datasets for what-if analysis

Making such integration feasible, practical, flexible, attractive, cost-effective, and usable
An example of a models-and-data approach: Splash

Loose model coupling through data exchange

Metadata
- Model inputs and outputs
- Access and execution
- Data schemas
- Model and data locations
- Model and data semantics

SPASH REPOSITORY contains

SPASH MODULES
- Model and Data Registration
- Model, Data, and Mapping Discovery
- Model and Data Composition
- Experiment Manager
- Model Execution
- Collaborative Reporting and Visualization

Provide models and data
Use models and data

Multi-disciplinary users

© 2011 IBM Corporation
Splash composite obesity model (proof of concept)

- GIS data
- Demographic data
- Facility data
- Results

Simulation model:
- Data source
- Mapping
- Flow of data

Transportation (VISUM simulation model)
Buying and eating (Agent-based simulation model)
Exercise (Stochastic discrete-event simulation)
BMI Model (Differential equation model)

Geospatial alignment
Time alignment and data merging

Clio++
JAQL + Hadoop
Data search → model-and-data search
  - Find compatible models, data, and mappings (using metadata)
  - Involves semantic search technologies, repository management, privacy and security

Data integration → model integration
  - Simulation-oriented data mapping
  - Time, space, unit alignment [e.g., Howe & Maier 2005]
  - Hierarchical models with different resolutions
  - Complex data transformations (e.g., raw simulation output to histogram)

Query optimization → simulation-experiment optimization
  - Optimally configure workflow among distributed data and models
  - Factoring common operations across different mappings in the workflow
  - Avoiding redundant computations across experiments
  - Statistical issues: managing pseudorandom numbers and Monte Carlo replications
Database research++ (continued)

- Causality approximation
  - Fixed-point + perturbation approaches
  - System support
  - Theoretical support

- Deep collaborative analytics
  - Visualizing and mining the results
  - Understanding and explaining results:
    - Dashboarding of parameters
    - Provenance [e.g., J. Friere et al.]
    - Root-cause analysis
    - Sensitivity analysis
  - Trusting results
    - Model validation
    - ManyEyes++, Swivel++

\[
\begin{align*}
\dot{f}_n(t) &= \Lambda_1(f_n(t), g_{n-1}(t)) \\
\dot{g}_n(t) &= \Lambda_2(f_{n-1}(t), g_n(t)) \\
\dot{f}(t) &= \Lambda_1(f(t), g(n\Delta t)) \\
\dot{g}(t) &= \Lambda_2(f(n\Delta t), g(t)) \\
\end{align*}
\]

for \( t \in [n\Delta t, (n+1)\Delta t) \)

Other models-and-data research: MCDB [Jermaine et al.], BRASIL [Gehrke et al.]
Conclusion

- DB community has focused on descriptive analytics, but enterprises need deep predictive analytics for what-if analysis, based on expert understanding of underlying mechanisms.

- Models and data need to be brought together on an equal footing.

- Requires significant extensions of database technology (exciting research opportunities!)

- Opportunity to redefine ourselves as the model-and-data community.

- In short:
DATA IS DEAD ... WITHOUT “WHAT-IF” MODELS

Peter J. Haas, Paul P. Maglio, Patricia G. Selinger, and Wang-Chiew Tan
IBM Almaden Research Center

Thanks to the Splash team: Melissa Cefkin, Susanne M. Glissmann, Cheryl A. Kieliszewski, Yinan Li, and Ronald Mak

www.almaden.ibm.com/asr/projects/splash