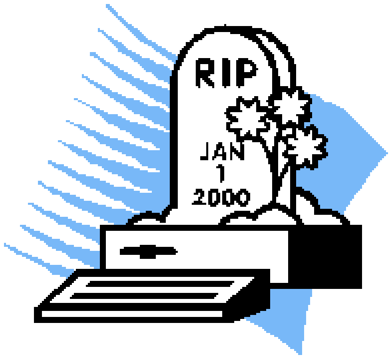
DATA IS DEAD … WITHOUT "WHAT-IF" MODELS

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



Congratulations, Database Community!

Statistical analysis

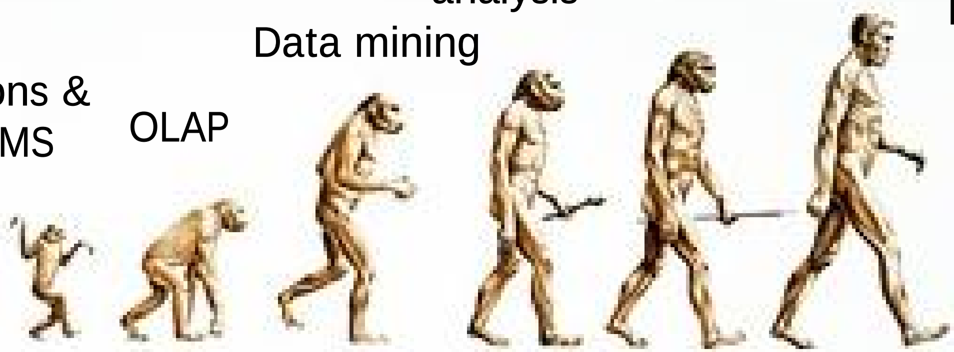
Text analytics

Machine learning

Data mining

Transactions & Reports, IMS

OLAP



Massive Data/Cloud DB

Uncertain data

Relational model & SQL

Semi-structured & Unstructured text

Web data

Streaming data

Semantic data

Congratulations, Database Community!

Statistical analysis

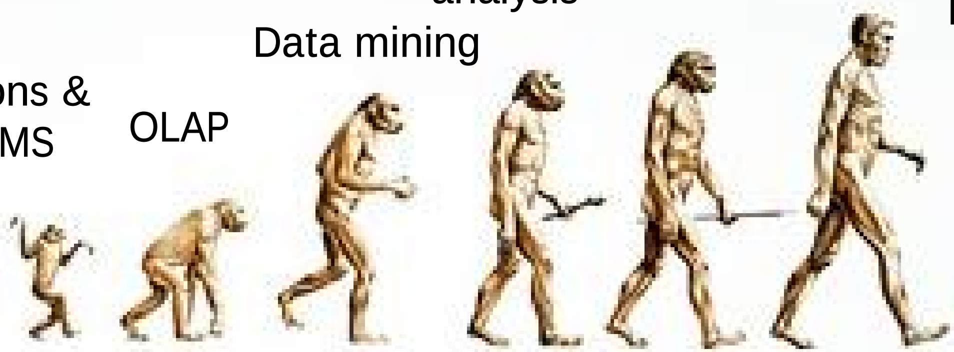
Text analytics

Machine learning

Data mining

Transactions & Reports, IMS

OLAP



Massive Data/Cloud DB

Uncertain data

Relational model & SQL

Semi-structured & Unstructured text

Web data

Streaming data

Semantic data

BUT: Why do enterprises care about data in the first place?

Because enterprises need to make DECISIONS

Allocation of scarce resources

informs online

Analytics Section

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

informs online

Analytics Section

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

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

Allocation of scarce resources

Because enterprises need to make DECISIONS

Allocation of
scarce resources

informs online

Analytics Section

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

"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

Allocation of scarce resources

informs online

Analytics Section

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

"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

Because enterprises need to make DECISIONS

informs online

Analytics Section

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

Allocation of
scarce resources

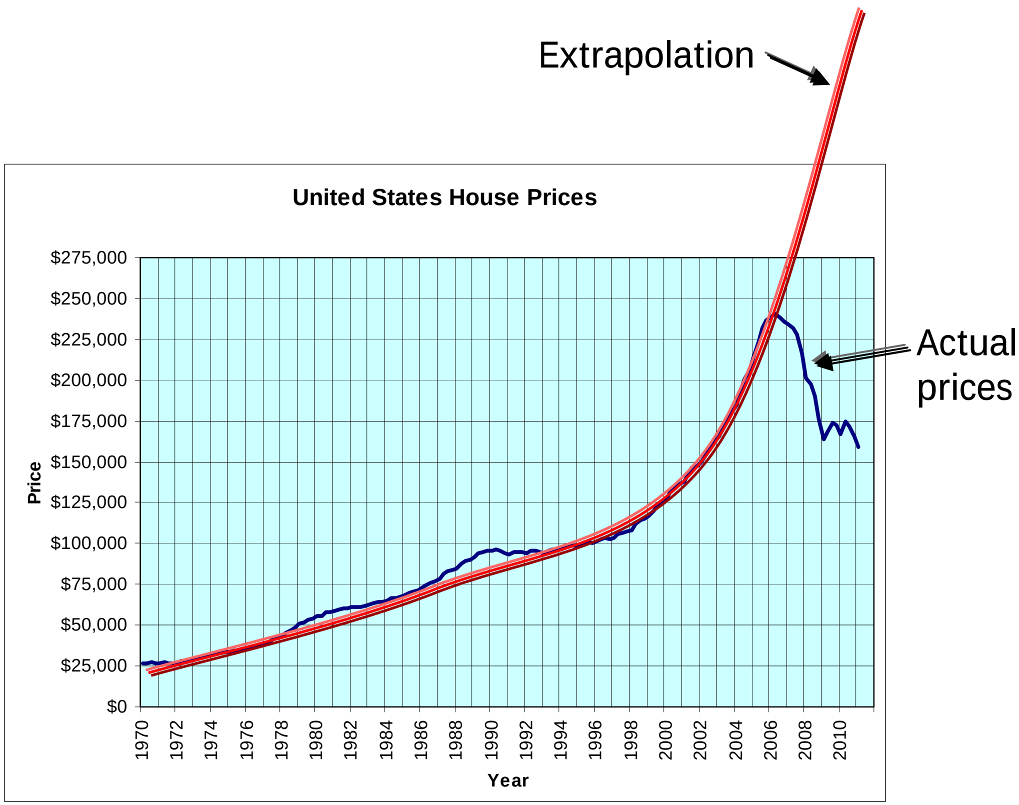
"Analytics is...a complete business problem solving and decision making process"

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

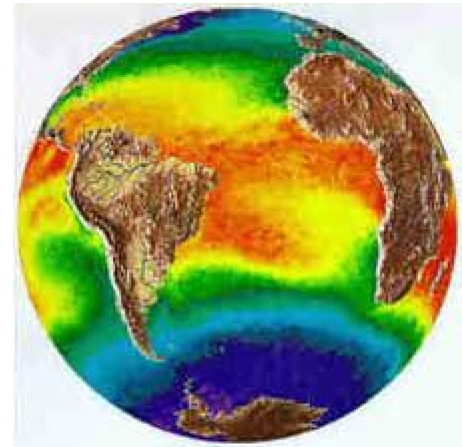
Predictive analytics: predict future probabilities and trends to allow what-if analysis

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

$$\frac{\partial \zeta}{\partial t} = \boldsymbol{k} \cdot \nabla \times (\boldsymbol{n}/\cos\phi) + F_{\zeta_H},$$

$$\frac{\partial \delta}{\partial t} = \nabla \cdot (\boldsymbol{n}/\cos\phi) - \nabla^2 (E + \Phi) + F_{\delta_H},$$

$$\frac{\partial T}{\partial t} = \frac{-1}{a\cos^2\phi}\left[\frac{\partial}{\partial \lambda}(UT) + \cos\phi \frac{\partial}{\partial \phi}(VT)\right] + T\delta - \dot{\eta}\frac{\partial T}{\partial \eta} + \frac{R}{c_p^*} T_v \frac{\omega}{p} + Q + F_{T_H} + F_{F_H},$$

$$\frac{\partial q}{\partial t} = \frac{-1}{a\cos^2\phi}\left[\frac{\partial}{\partial \lambda}(Uq) + \cos\phi \frac{\partial}{\partial \phi}(Vq)\right] + q\delta - \dot{\eta}\frac{\partial q}{\partial \eta} + S,$$

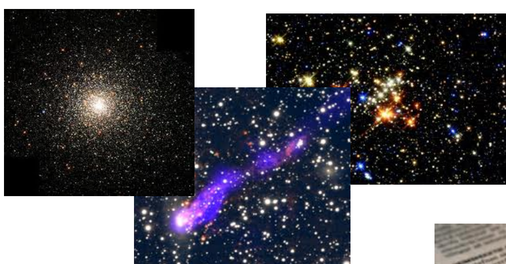
$$\frac{\partial \pi}{\partial t} = \int_1^{\eta_t} \boldsymbol{\nabla} \cdot \left(\frac{\partial p}{\partial \eta}\boldsymbol{V}\right) d\eta.$$

Is the DB community truly helping decision makers?

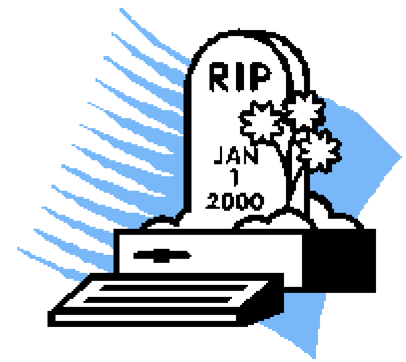
Some realizations...

Data is **dead**

Name	Item	Price	Date
Pat	Red shoes	$50	1/23/11



=



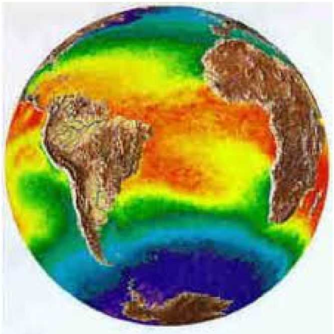
...a record of history that says nothing about future or hypothetical worlds

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

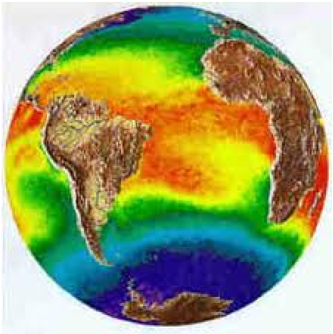
We can understand much more by moving to deep predictive analytics based on **models and data**



3.3 Eulerian Dynamical Core

$$
\begin{aligned}
\frac{\partial \zeta}{\partial t} &= \boldsymbol{k} \cdot \nabla \times (\boldsymbol{n}/\cos\phi) + F_{\zeta_H}, \\
\frac{\partial \delta}{\partial t} &= \nabla \cdot (\boldsymbol{n}/\cos\phi) - \nabla^2 (E + \Phi) + F_{\delta_H}, \\
\frac{\partial T}{\partial t} &= \frac{-1}{a \cos^2\phi} \left[\frac{\partial}{\partial \lambda}(UT) + \cos\phi \frac{\partial}{\partial \phi}(VT) \right] + T\delta - \dot{\eta}\frac{\partial T}{\partial \eta} + \frac{R}{c_p^*} T_v \frac{\omega}{p} \\
&\quad + Q + F_{T_H} + F_{F_H}, \\
\frac{\partial q}{\partial t} &= \frac{-1}{a \cos^2\phi} \left[\frac{\partial}{\partial \lambda}(Uq) + \cos\phi \frac{\partial}{\partial \phi}(Vq) \right] + q\delta - \dot{\eta}\frac{\partial q}{\partial \eta} + S, \\
\frac{\partial \pi}{\partial t} &= \int_1^{\eta_t} \boldsymbol{\nabla} \cdot \left(\frac{\partial p}{\partial \eta} \boldsymbol{V} \right) d\eta.
\end{aligned}
$$

We can understand much more by moving to deep predictive analytics based on **models and data**

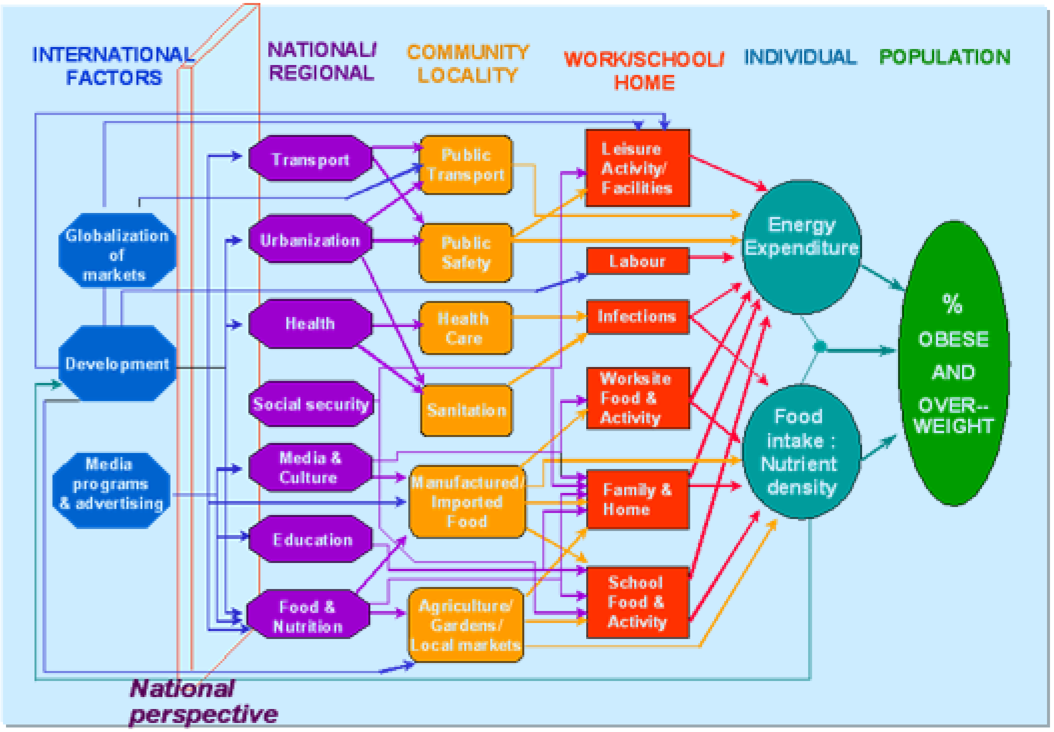


3.3 Eulerian Dynamical Core

$$
\begin{aligned}
\frac{\partial \zeta}{\partial t} &= \boldsymbol{k} \cdot \nabla \times (\boldsymbol{n}/\cos\phi) + F_{\zeta_H}, \\
\frac{\partial \delta}{\partial t} &= \nabla \cdot (\boldsymbol{n}/\cos\phi) - \nabla^2 (E + \Phi) + F_{\delta_H}, \\
\frac{\partial T}{\partial t} &= \frac{-1}{a\cos^2\phi}\left[\frac{\partial}{\partial\lambda}(UT) + \cos\phi\frac{\partial}{\partial\phi}(VT)\right] + T\delta - \dot{\eta}\frac{\partial T}{\partial\eta} + \frac{R}{c_p^*}T_v\frac{\omega}{p} \\
&\quad + Q + F_{T_H} + F_{F_H}, \\
\frac{\partial q}{\partial t} &= \frac{-1}{a\cos^2\phi}\left[\frac{\partial}{\partial\lambda}(Uq) + \cos\phi\frac{\partial}{\partial\phi}(Vq)\right] + q\delta - \dot{\eta}\frac{\partial q}{\partial\eta} + S, \\
\frac{\partial \pi}{\partial t} &= \int_1^{\eta_t} \boldsymbol{\nabla} \cdot \left(\frac{\partial p}{\partial\eta}\boldsymbol{V}\right) d\eta.
\end{aligned}
$$

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

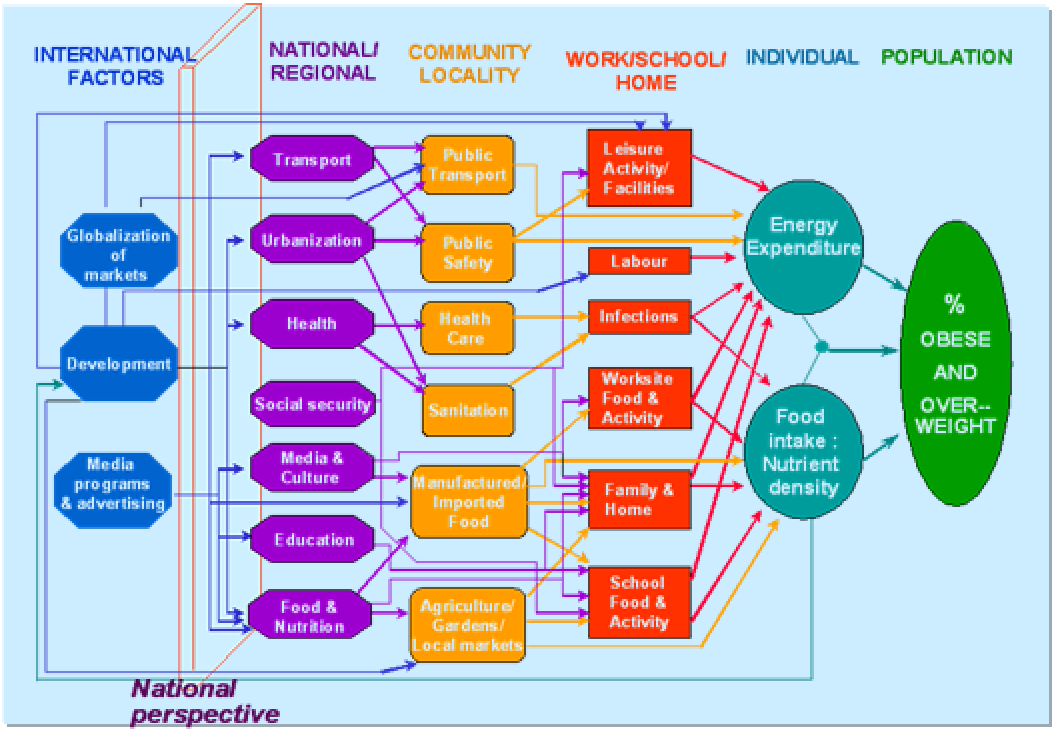
Especially true for complex systems-of-systems



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

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

Especially true for complex systems-of-systems



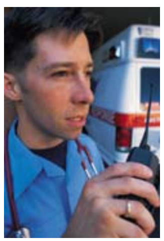
Huang, T. T, Drewnowski, A., Kumanyika, S. K., & Glass, T. A., 2009, "A Systems-Oriented Multilevel Framework for Addressing Obesity in the 21st Century, " Preventing Chronic Disease, 6(3).

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



Provide models and data

Use models and data

Multi-disciplinary users

SPLASH REPOSITORY contains

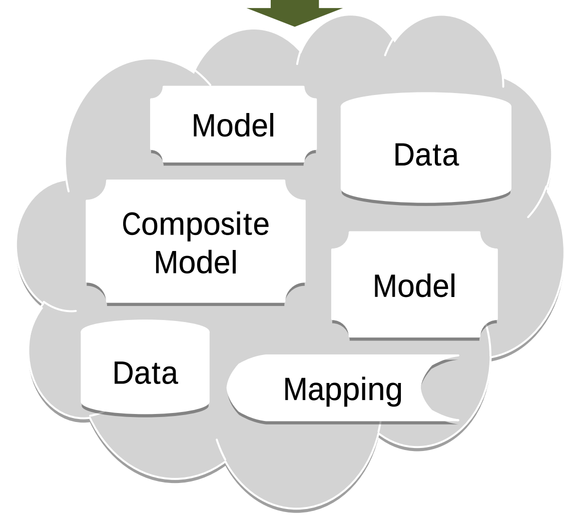
Metadata

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

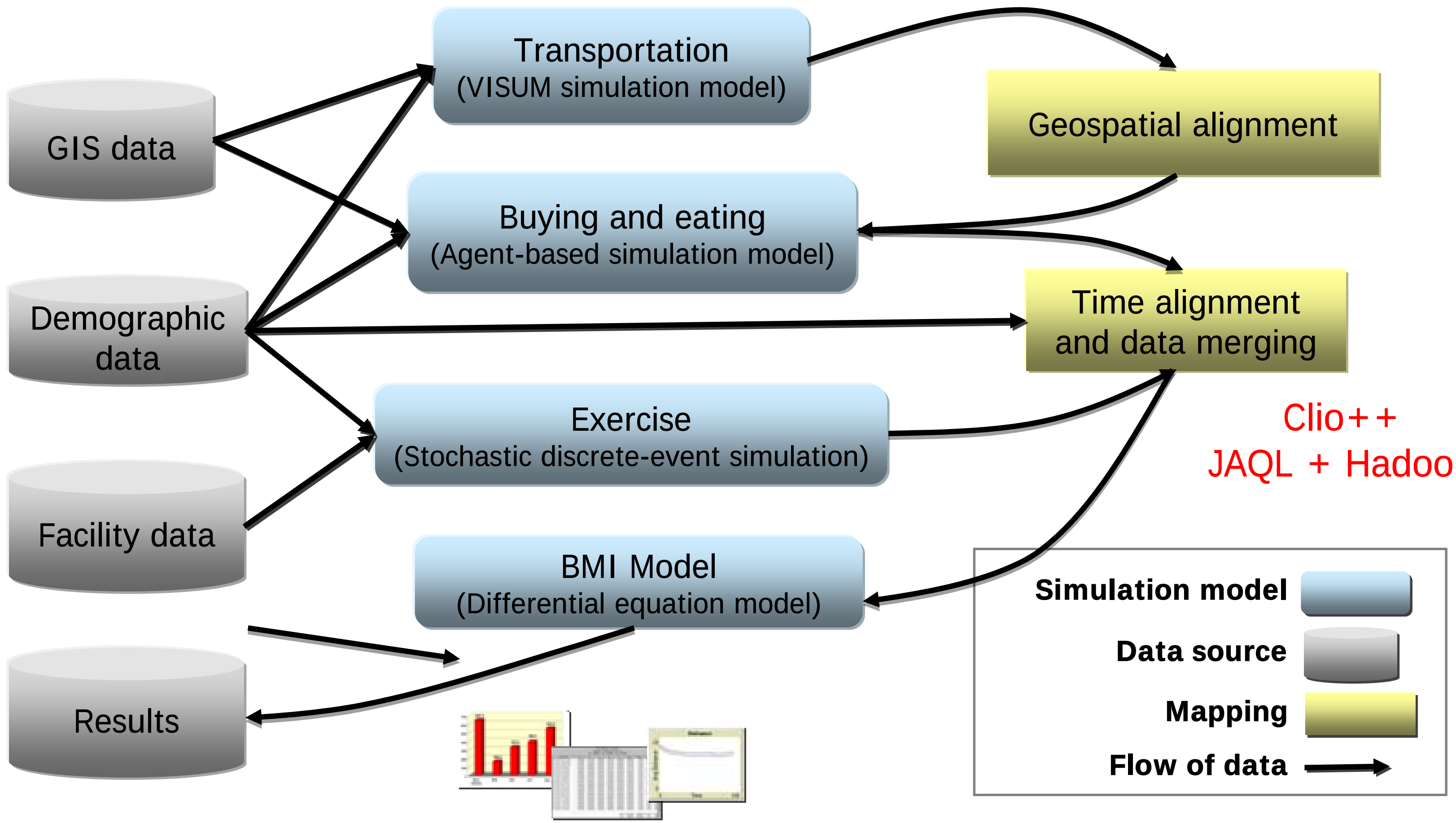
SPLASH MODULES

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

describes



Splash composite obesity model (proof of concept)

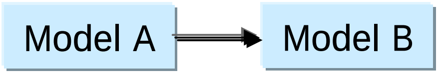


Database research++

eHarmony

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

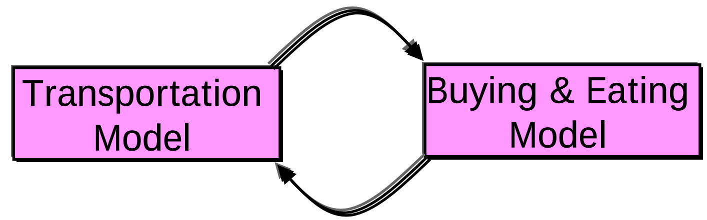
- **Data integration → model integration**
  - Simulation-oriented data mapping
  - Time, space, unit alignment [e.g., Howe & Maier 2005]
  - Hierarchical models with different resolutions
  - Complex data transformations (e.g., raw simulation output to histogram)



- **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)

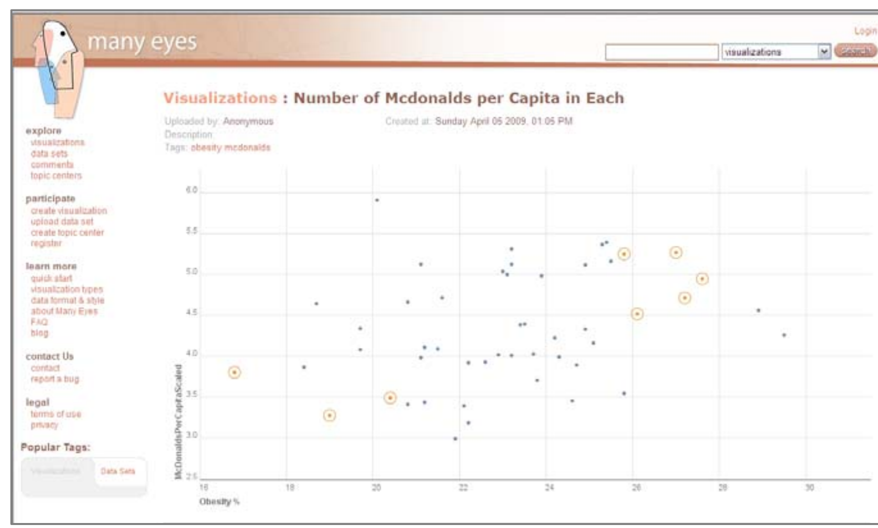
- **Causality approximation**
  - Fixed-point + perturbation approaches
  - System support
  - Theoretical support



$$\dot{f}_n(t) = \Lambda_1\left(f_n(t), g_{n-1}(t)\right)$$
$$\dot{g}_n(t) = \Lambda_2\left(f_{n-1}(t), g_n(t)\right)$$

$$\left.\begin{aligned}\dot{f}(t) &= \Lambda_1\left(f(t), g(n\Delta t)\right)\\ \dot{g}(t) &= \Lambda_2\left(f(n\Delta t), g(t)\right)\end{aligned}\right\} \text{ for } t \in \left[n\Delta t, (n+1)\Delta t\right)$$

- **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++



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

Conclusion

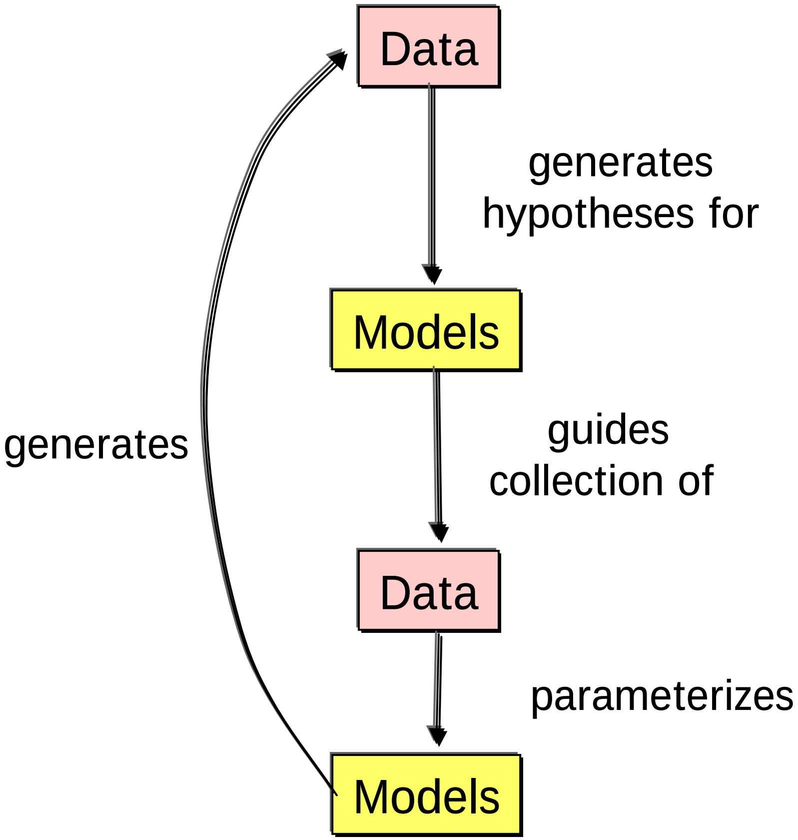
- DB community has focused on descriptive analytics, but enterprises need deep predictive analytics for what-if analysis, based on expert understanding of underlying mechanisms

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

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

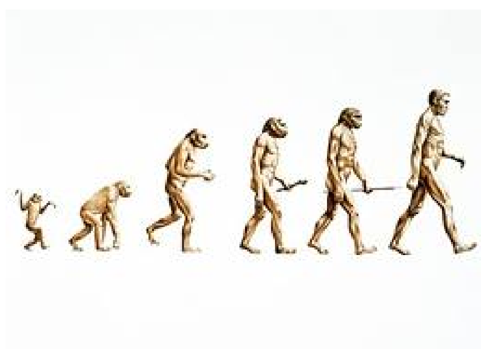
- Opportunity to redefine ourselves as the model-and-data community

- 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