SAND2011-7029 C
SAND2011-7029C

# Monsoons, Movies, Memes, and Genes: Combining KD and M&S for Prediction

**Rich Colbaugh**

Sandia National Laboratories
New Mexico Institute of Mining and Technology

August 2011

## Introduction

**Background**

- Knowledge discovery (KD) and modeling and simulation (M&S) have each made profound contributions to human knowledge and offer complementary perspectives – it's natural to try to combine them.
- However, the KD and M&S communities have evolved essentially independently, so potential benefits remain largely unexplored.

**Objective**

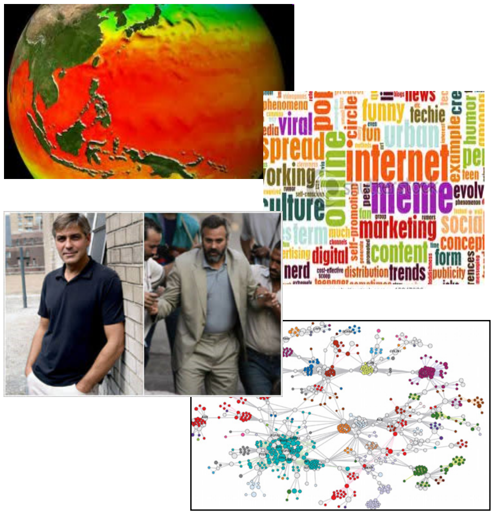
Illustrate, through a series of examples taken from the *predictive analysis* domain, that combining the two approaches is actually a good idea!

*Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.*

# Introduction

**Outline**

- "Standard" approaches to combining KD and M&S:
  - KD on *data*;
  - climate examples.
- Another perspective on combining KD and M&S:
  - KD on *model*;
  - social and biological network examples.



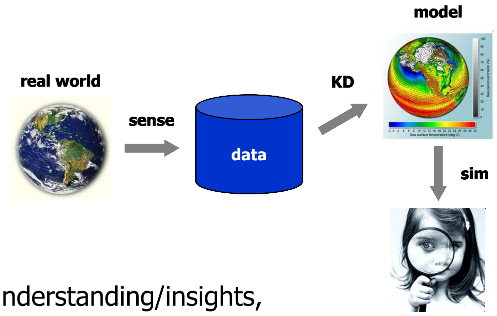
# Standard Approaches

**Basic idea**

The idea is simple and natural: apply KD to *data*, then use the results to better construct and/or exploit M&S.

**Approach one**

- Collect data on real-world phenomenon of interest.
- Apply KD to uncover patterns in data, which can be incorporated into computational model.



- Simulate model to develop understanding/insights, make predictions, etc.

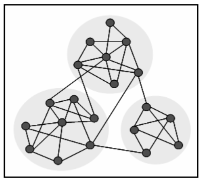
# Standard Approaches

**Example [Steinhaeuser/Chawla/Ganguly 2011]**

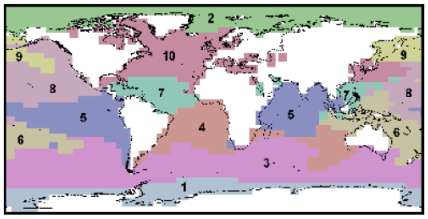
Goal: Leverage high-resolution climate data to build predictive models.

Approach:

- build "climate network" – vertices are spatial grid cells, edges connect cells that are (significantly) correlated in climatic variability;
- cluster cells according to network communities;
- form predictions using linear regression with *cell community averages* as predictive features.



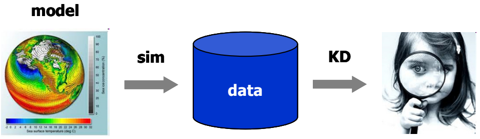
Results: improved prediction accuracy, discovery of meaningful new climate indices.



# Standard Approaches

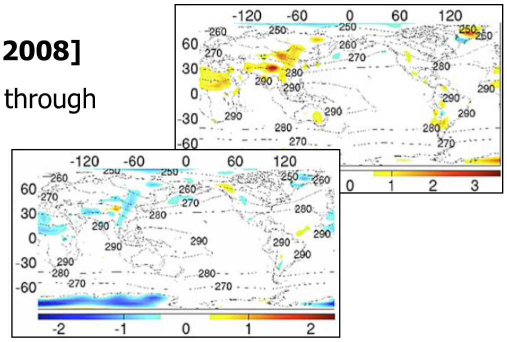
**Approach two**

- Generate data via simulation of a computational model.
- Apply KD to simulation data, enabling rigorous analysis and new insights.



**Example [Danforth/Kalnay 2008]**

- Goal: improve M&S accuracy through online bias correction.
- Results: improved model forecasts compared with standard offline schemes.



# New Approach

**Basic idea**

We now introduce an alternative approach, in which KD is applied *directly to the model*, for instance for improved data analysis or enhanced M&S.

**Two interesting problems**

The discussion is organized around two important, challenging problems:

- predicting the outcomes of social dynamics processes (cultural and other markets, political and social movements, emerging contentious situations, etc.);
- predicting the presence of vulnerabilities, especially those associated with rare events, in complex systems.

In each case we first quickly describe why the problem is hard, then give the proposed analytic approach, and finally demonstrate the efficacy of the methodology with a real-world case study.

# New Approach: Social Prediction

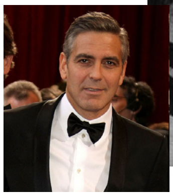
**Why is predicting social dynamics so hard?**

Vast resources are devoted to predicting outcomes of social processes, but prediction quality is often poor. One difficulty is *social influence* [Salganik et al. 2006, Colbaugh/Glass 2009]:

- people are influenced by what others do;
- consequently, “intrinsics” (like George) usually matter less than social influence;
- standard prediction is based upon intrinsics.

**Proposal**

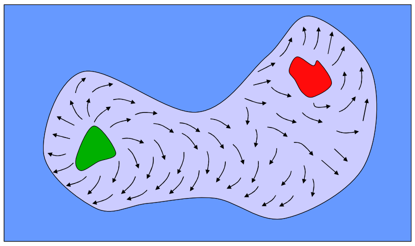
Conduct predictability assessment via combined KD/M&S analysis to identify aspects of social influence which possess predictive power.

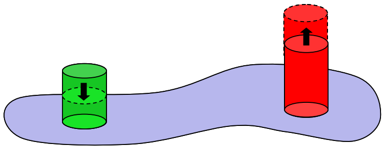


# New Approach: Social Prediction

**Predictability assessment: elements**

- Predictability: a phenomenon is predictable if there is adequate difference in probabilities of qualitatively distinct outcomes.
- Predictive features: measurables/ patterns that boost predictability.
- Approach: formulate predictability in terms of reachability, and evaluate reachability by applying KD to social network dynamics model via "altitude" functions.
- Technical details in [Colbaugh/ Glass 2009], but for basic idea …





# New Approach: Social Prediction

**Predictability assessment: illustrative example**

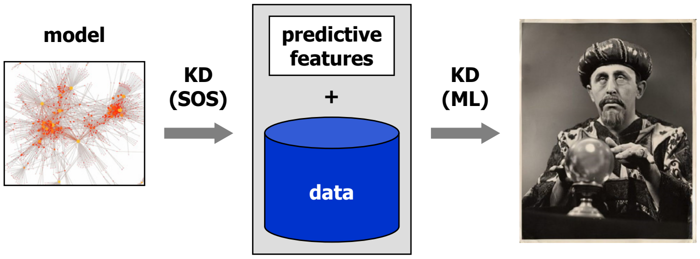
- System $\Sigma_{sc}$: $dx = f(x)\,dt + g(x)\,dw$, where $w(t)$ is a Wiener process.
- Theorem: $\gamma$ is an upper bound on the probability of reaching $X_r \subseteq X$ from $X_0 \subseteq X$ while remaining in X if there exists A(x) such that
  - $A(x) \leq \gamma \;\forall x \in X_0$;
  - $A(x) \geq 1 \;\forall x \in X_r$;
  - $A(x) \geq 0 \;\forall x \in X$;
  - $(\partial A/\partial x)\, f + (1/2)\, \mathrm{tr}\, [g^T (\partial^2 A/\partial x^2)\, g] \leq 0 \;\forall x \in X$.
- Computation: existence of function A(x) satisfying theorem criteria can be verified, efficiently and constructively, through semidefinite programming (via convex relaxation and SOS programming, for instance using SOSTOOLS [Prajna et al. 2001]).

# New Approach: Social Prediction

**Predictive analysis process**

Procedure:

- Construct model and perform SOS reachability analysis to assess predictability and identify predictive features (if any).
- Form predictions through KD-based data analysis (e.g., machine learning)



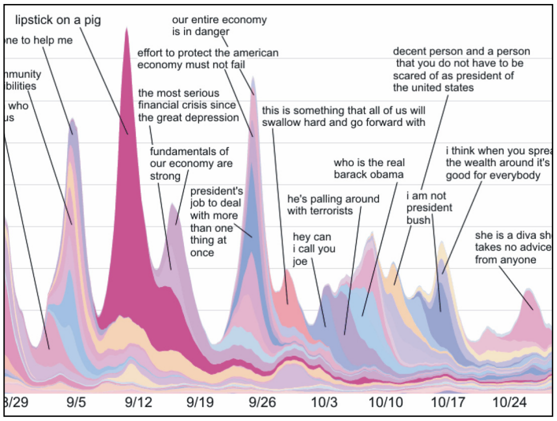
# New Approach: Social Prediction

**Case study: "meme" prediction**

Problem: distinguish memes which will "go viral" from those that will not early in meme lifecycle.

Data:

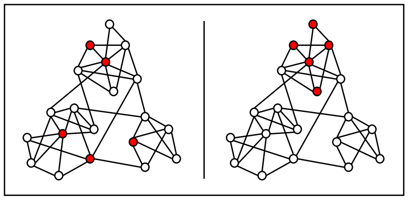
- time series for 70K U.S. political memes collected in last half of 2008 [Leskovec et al. 2009];
- associated Web graph (550K sites, 1.4M hyperlinks) and blog content.

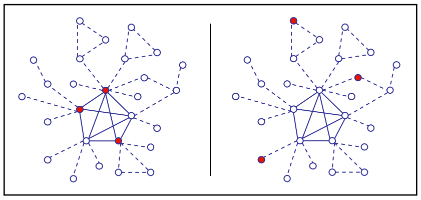


# New Approach: Social Prediction

**Case study: "meme" prediction (cont'd)**

Reachability-based predictability assessment suggests the following two features should be predictive of large social diffusion events:





**Community Structure**

- Graph analysis: fast modularity-based vertex partitioning.
- Predictive feature: early entropy of activity across network communities.

**Core-periphery Structure**

- Graph analysis: fast (decentralized) k-core decomposition.
- Predictive feature: early density ratio of k-core v. periphery activity.

# New Approach: Social Prediction

**Case study: "meme" prediction (cont'd)**

Method

- Learn classifier [AVATAR 2010] which takes candidate features as input and predicts whether given meme will be successful (≥1000 posts) or unsuccessful (≤100 posts).
- Estimate prediction accuracy via ten-fold cross-validation with data set of 100 successful memes and 100 unsuccessful memes.

Results

**Predictive Features**

1. Early network community dispersion.
2. Early network k-core activity.
3. Early number of posts, post rate.
4. Language features (sentiment, emotion).

**Prediction Performance**

| Time window | Accuracy |
|---|---|
| language-only | ~66% |
| first 12hr | ~84% |
| first 24hr | ~92% |
| first 36hr | ~94% |

# New Approach: Vulnerability

**Why is identifying complex system vulnerabilities so hard?**

- The usual things: scale, complexity, “rare event” issues, … .
- More subtle things: for example, the robust yet fragile nature of evolving systems [Carlson/Doyle 2002, Colbaugh/Glass 2009].

**Proposal**

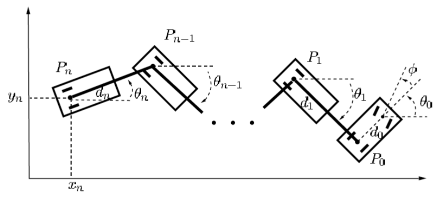
Exploit system structure by applying KD directly to system model, thereby allowing provably-correct vulnerability assessments. We illustrate with an interesting/important class of systems.

**“I’m sure it’s all right – it’s a horse you have to worry about.”**

# New Approach: Vulnerability

**Flat systems**

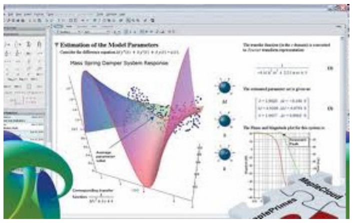
- Many complex networks are *differentially flat* [Martin et al. 2003].
- Such systems possess flat outputs:
  - which can realize any specified trajectory;
  - whose trajectory completely define the system evolution.
- Consequently, vulnerability analysis of flat systems can be conducted in flat output space *and without simulation*.
- Example: car with n trailers.



# New Approach: Vulnerability

**Flat systems: some details**

- Definition: $dx/dt = f(x,u)$ is *differentially flat* if there exist flat outputs z, equal in number to the number of inputs u, such that $z = H(x)$, $x(t) = F_1(z, dz/dt, \ldots, d^rz/dt^r)$, and $u(t) = F_2(z, dz/dt, \ldots, d^rz/dt^r)$.
- Deciding flatness/finding flat outputs (this is the trick!):
  - one option: exhibit flat outputs by exploiting knowledge of system;
  - more systematically, algorithm of [Antritter/Levine 2008]:
    - models system with differential forms rather than vector fields;
    - uses computer algebra (CA) to check necessary and sufficient conditions for flatness and, if satisfied, generate flat outputs.

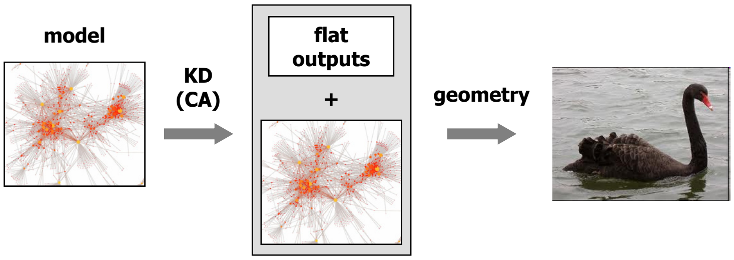


# New Approach: Vulnerability

**Vulnerability analysis process**

Procedure:

- Construct model and perform CA flatness analysis to decide whether the system is flat and, if it is, identify the flat outputs.
- Discover the vulnerabilities through simple (e.g., geometric) analysis of flat output trajectories.



model

KD (CA)

flat outputs

+

geometry

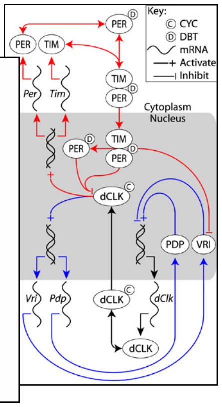
# New Approach: Vulnerability

**Case study: circadian rhythm gene network**

Problem: find vulnerabilities (therapeutic control targets) for resetting circadian rhythm (CR) gene networks (surprisingly important!).

Example: model for *drosophila* CR network.

$$
\begin{aligned}
\frac{dM_P}{dt} &= v_{sP}\frac{K_{IP}^n}{K_{IP}^n + C_N^n} - v_{mP}\frac{M_P}{K_{mP}+M_P} - k_d M_P \\
\frac{dP_0}{dt} &= k_{sP}M_P - V_{1P}\frac{P_0}{K_{1P}+P_0} + V_{2P}\frac{P_1}{K_{2P}+P_1} - k_d P_0 \\
\frac{dP_1}{dt} &= V_{1P}\frac{P_0}{K_{1P}+P_0} - V_{2P}\frac{P_1}{K_{2P}+P_1} - V_{3P}\frac{P_1}{K_{3P}+P_1} + V_{4P}\frac{P_2}{K_{4P}+P_2} - k_d P_1 \\
\frac{dP_2}{dt} &= V_{3P}\frac{P_1}{K_{3P}+P_1} - V_{4P}\frac{P_2}{K_{4P}+P_2} - k_3P_2^2 + k_4C - v_{dP}\frac{P_2}{K_{dP}+P_2} - k_d P_2 \\
\frac{dC}{dt} &= k_3P_2^2 - k_4C - k_1C + k_2C_N - k_{dC}C \\
\frac{dC_N}{dt} &= k_1C - k_2C_N - K_{dC}C
\end{aligned}
$$



# New Approach: Vulnerability

**Case study: circadian rhythm gene network (cont'd)**

- Solution One: quantify sensitivity of all parameters of CR gene network; most sensitive are then candidate control targets [Baghari et al. 2008].
- Solution Two: because CR gene network is flat [Colbaugh/Glass 2010], simply read-off flat inputs associated with flat outputs – this process enables top four candidate control targets to be identified directly from the model.
- Remarks: flatness also enables
  - same sort of control target analysis for CR gene networks for neurospora and mouse;
  - extremely efficient characterization of reachability properties of CR networks.