2) On *Methodology* (Practice) *of Research*

Jiming Liu

Dean of Science & Chair Professor in Computer Science

Hong Kong Baptist University

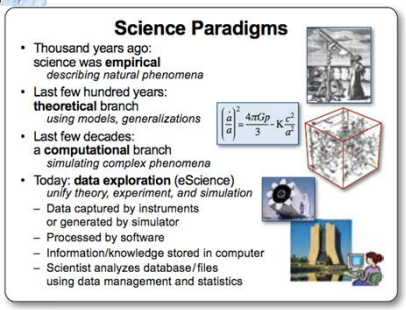
PARADIGMS

Paradigms

- Mathematics:
  - 17th century: Descartes, Hobbes, Spinoza, Leibnitz, and Pascal
- Psychology:
  - 18th century: Berkeley, Hume, Condillac, and Kant
- Synthesis/biology/nature:
  - 19th century: Schelling (construct a program which covers both nature and the intellectual life in a single system and method), Schopenhauer (world as representation), Spencer (application of evolution to every field), Nietzsche (creative powers of the individual),
  - 20th century: Bergson (rationalism)
  - ...

Jim Gray's Four Scientific Paradigms

Jim Gray (1944-2007)
Turing Award Winner 1998



1. empiricism
observe phenomenon and attempt to classify
Ptolemy's universe of concentric spheres

2. theory
describe above classifications with mathematical models
Newtonian/Einsteinian gravity

3. computation
build `virtual' physical systems via solution of math models
Cosmic structure formation

4. data-driven synthesis
unite empirical, theoretical and computational branches with data (X-info and Comp-X)
Matter/energy content of the universe

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

☆ Illustration: Marian Bantjes ☆ **"All models are wrong**, but some are useful."

So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all.

3 SEPTEMBER 2008

11 FEBRUARY 2011

DATA ☜ SCIENCE

DATA ☜☞ SCIENCE

A few years ago, Chris Anderson, former editor in chief of *Wired* magazine, published a provocative and thought-provoking article: "The end of theory: the data deluge makes the scientific method obsolete" (http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory/). As the title indicates, Anderson asserted that in the era of petabyte information and supercomputing, the traditional, hypothesis-driven scientific method would become obsolete. No more theories or hypotheses, no more discussions whether the experimental results refute or support the original hypotheses. In this new era, what counts are sophisticated algorithms and statistical tools to sift through a massive amount of data to find information that could be turned into knowledge.

> *"… [an] imagined future in which the long-established way of doing scientific research is replaced by computers that divulge knowledge from data at the press of a button…"*

press of a button deserves some inquiry from an epistemological point of view. Is data-driven research a genuine mode of knowledge production, or is it above all a tool to identify potentially useful information? Given the amount of scientific data available, is it now possible to dismiss the role of theoretical assumptions and hypotheses? Should this new mode of gathering information supersede the old way of doing research?

The scientific method encompasses an ongoing process of formulate a hypothesis–test with an experiment–analyze the results–reformulate the hypothesis. Such a way of proceeding has been in use for centuries and is basically accepted in our Western society as the most reliable way to produce robust knowledge.

However, Anderson is not the first to want to relegate hypotheses to a subordinate role. Francis Bacon, the "father of the scientific method" himself, in his Novum Organum (1620), argued that scientific knowledge should not be based on preconceived notions but on experimental data. Deductive reasoning, he argued, is eventually limited because setting a premise in advance of an experiment would constrain Johannes Kepler. In 1609 and 1619, Kepler, who was the assistant of Tycho Brahe, published the three laws of planetary motion based on his analysis of Brahe's observational data. These would be later verified by the laws of motion and universal gravitation in Isaac Newton's Principia. Newton was another follower of empiricism. *Hypotheses non fingo*—I frame no hypotheses—he asserted. Like Bacon, he advised a bottom-up approach, assuming the primacy of experiments, which provide empirical evidence on which to base induction.

> *"Deductive reasoning […] is eventually limited because setting a premise in advance of an experiment would constrain the reasoning so as to match that premise."*

Big Data science renews the primacy of inductive reasoning in the form of technology-based empiricism and has inspired a view of the future in which automated data mining will lead directly to new discoveries. According to this view, the new "hypothe-

Institute for Complex Systems of the National Research Council, Monterotondo, Italy. E-mail: fulvio.mazzocchi@isc.cnr.it
doi:10.1093/nsr/nwv054 | Published online 10 September 2015

> *"… correlations play an important role as heuristic devices [but] have to be further analyzed […] to assign them a meaning"*

The most relevant outcome from ENCODE is the finding that most of the human genome (about 80%) could be assigned a "biochemical function," meaning that it participates in at least one biochemical event in at least one cell type. This result, which has received much attention in the press, contrasts the notion of junk DNA—that is, DNA sequences with no apparent function—which were believed to make up more than 90 percent of the human genome. But is it really true that this concept has been debunked by the ENCODE project?

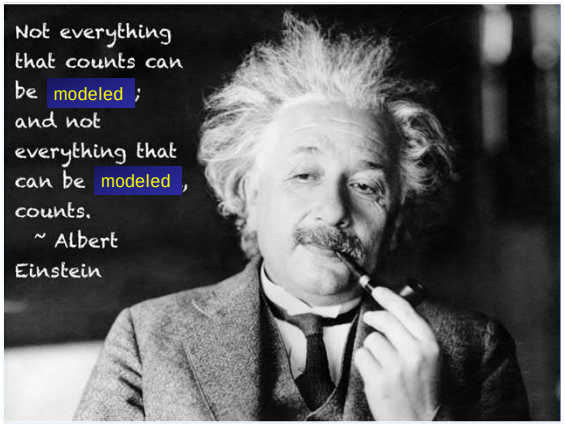
One argument concerns the notion of "function" by ENCODE: "Operationally, we define a functional element as a discrete genome segment that encodes a defined product (for example, protein or non-coding RNA) or displays a reproducible biochemical signature (for example, protein binding, or a specific chromatin structure)" [3]. In light of this definition, it is possible to assign function to 80 percent of the human genome. But the ENCODE definition is clearly very loose. The American biologist Michael White and his team randomly generated 1,300 DNA sequences and found that most of these can be regarded as functional along with the biochemical criteria used by particular region of the genome actually does "something useful for us" (http://www.huffingtonpost.com/michael-white/media-genome-science_b_1881788.html). Much more work is required to understand whether a certain part of the genome does have a biological function and how this works—and this requires, above all, smaller-scale, hypothesis-driven research.

More data do not necessarily generate more knowledge. Data by themselves are meaningless. The idea that "with enough data, the numbers speak for themselves" hardly makes sense.

The "no theory" thesis contrasts with the fact that the collection of data is not a merely empirical activity. Science does not collect data randomly. Experiments are designed and carried out within theoretical, methodological and instrumental limitations. Instruments are designed based on prior theories and knowledge, which determine what these instruments indicate with respect to the object under investigation. Research does not examine each possible manipulation that could occur, but selects what is relevant in light of a given perspective, sometimes in order to match theoretical predictions with experience.

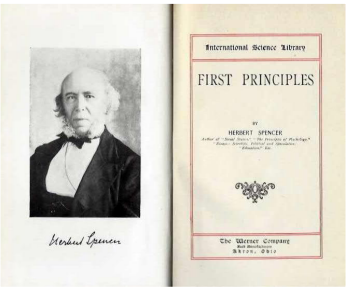
The collider experiments in high-energy physics illustrate this selective mode of conducting research. After the discovery of the W and Z bosons in 1983, the Standard Model of elementary particles—quarks, leptons and forces—was considered as basically proven; the only particle not yet discovered was the Higgs boson. In 2011 to generate enough raw data about decay products. The LHC generates up to 600 million collisions per second and produces 15 petabytes (15 million gigabytes) of data per year. Finding the traces of elementary particles requires sifting through this deluge of data to look for specific patterns. To handle this enormous task, the Worldwide LHC Computing Grid (WLCG) that links hundreds of data processing centers around the world was created in 2002. The performance of the Grid is essential for supporting LCH experiments and releasing results quickly. Big Data, distributed computing and sophisticated data analysis all played a crucial role in the discovery of the Higgs boson—and perhaps in finding new "patterns," they might also generate new hypotheses in this field. But the discovery of the Higgs boson was not data-driven. The collider experiments were mostly driven by theoretical predictions: It is because scientists were attempting to confirm the Standard Model of elementary particles that the discovery of the Higgs boson—the only missing piece—could occur.

> *"Big Data, distributed computing and sophisticated data analysis all played a crucial role in the discovery of the Higgs boson […] But the discovery of the Higgs boson was not data-driven."*



www.exkalibur.com

Q1: A *Single* Paradigm?

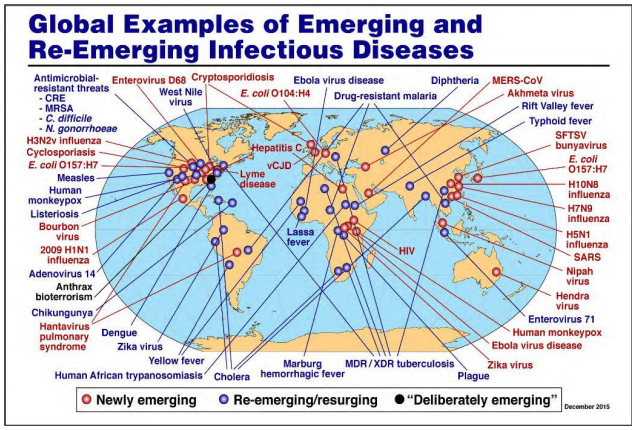


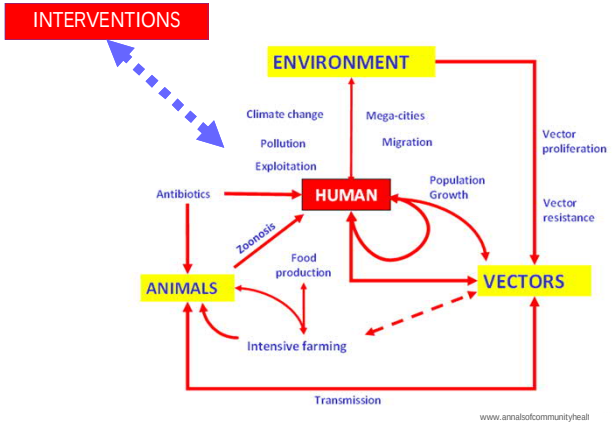
Herbert Spencer
(1820 – 1903)

CONTENTS

www.bythewaybooks.com/pages/books/19660/herbert-spencer/first-principles

The Real World…





"Anticipating and responding to disease risk requires interpreting disease events [. . .] as emergent properties of a complex system from which to gather infectious disease intelligence."

We think that such a system for classifying infectious disease risk would help to guide the development of infectious disease intelligence and to identify best courses of action. In the following sections, we consider data needs and modeling technologies that would serve such activities at each threat level.

The predictive capacity of infectious disease intelligence is not limited by technology. Machine learning methods have already been shown to be effective at harnessing data from multiple sources to characterize the zoonotic potential of particular wildlife species [5]. Instead, our capacity to predict spillover events depends on environmental and ecological data, such as the distribution of zoonoses and their vectors and reservoir species, knowledge about pathogens that are not yet known to infect humans, and the assimilation of data from multiple sources to quantify risk and identify trigger conditions early enough for timely intervention. Creating a data infrastructure that would enable real-time risk quantification would empower the health community to better evaluate the most reasonable preventative investments—such as disrupting plausible transmission may be carried out over the course of months or years to treat chronic infection and prevent transmission [7]. Importantly, the downstream consequences of an outbreak could exacerbate the effects of another disease. These complex interactions can be nonlinear and occur at dueling timescales whose dynamical consequences can again be explored using computer models [8].

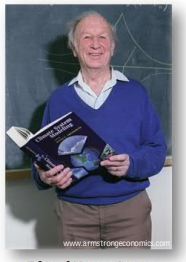
"The hard limits to forecasting are set by the volume and quality of basic scientific information."

In contrast, the goal of phylodynamic modeling is to provide a better understanding

By Barbara Han & John Drake

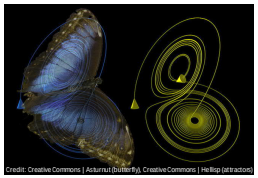
COMPLEXITY

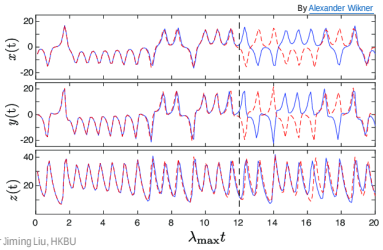
Butterfly Effect

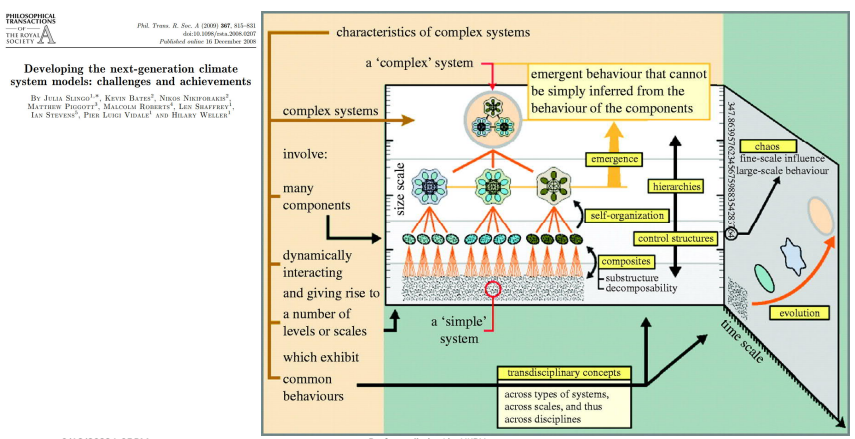


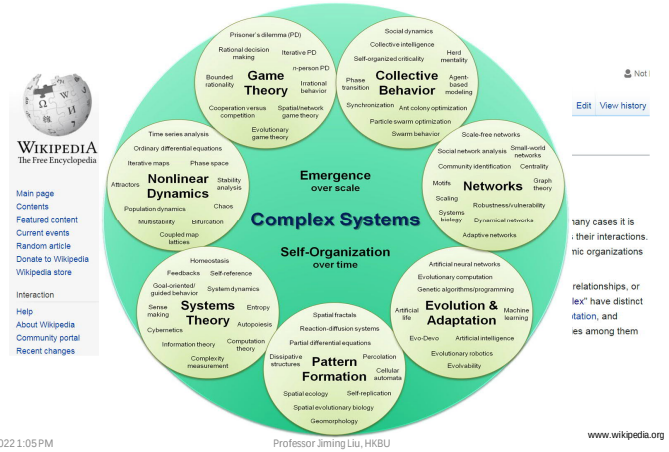
Edward Norton Lorenz
(1917–2008)

$$\frac{dx}{dt} = a(y - x)$$
$$\frac{dy}{dt} = x(b - z) - y$$
$$\frac{dz}{dt} = xy - cz$$



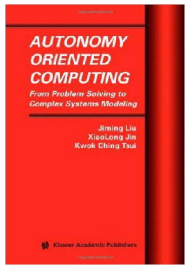






Complex Systems: Autonomy-Oriented Computing

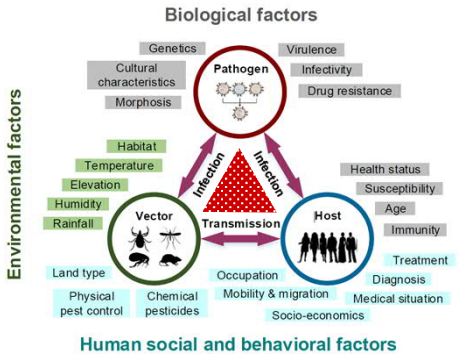
- Goal 1: *modeling* (of autonomous entities of) complex systems (e.g., cyber-physical-social systems)
- Goal 2: *computing* with autonomous entities (e.g., for tackling complex computational problems)

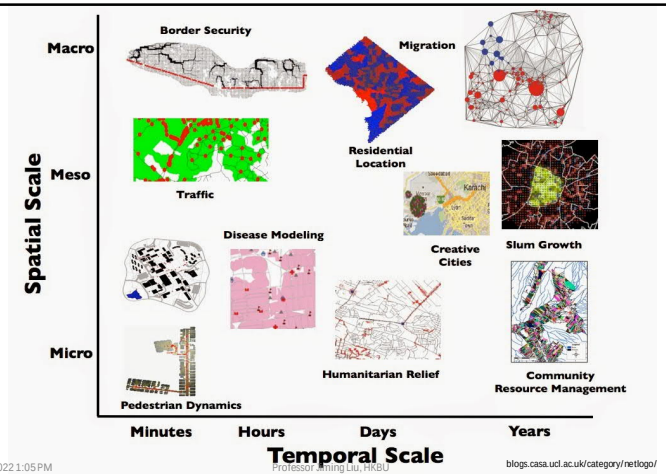


back to the Reality…

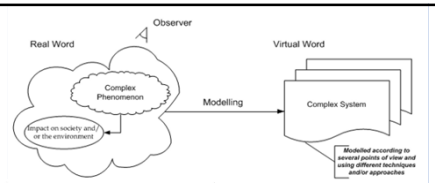
Complexity (“*Mysterious Triangle*”)

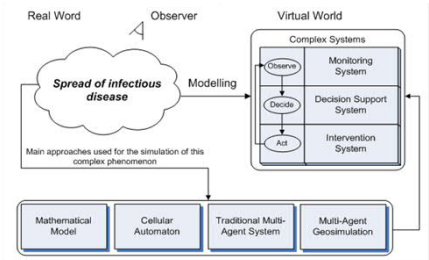
Some interacting components (in circles) and associated factors that can affect the transmission of diseases.





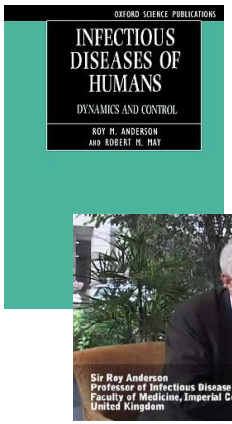
blogs.casa.ucl.ac.uk/category/netlogo/





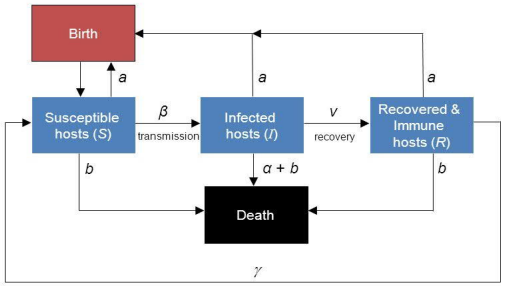
Mondher Bouden and Bernard Moulin, "Multi-level geosimulation of zoonosis propagation: A multi-agent and climate sensitive tool for risk management in public health," in Advances in Risk Management, *Giancarlo Nota (Ed), ISBN 978-953-307-138-1, Published: August 17, 2010.*

NATURE OF INQUIRY (AIM OF MODELING)



Modeling Microparasite Disease Dynamics

Coupled differential equations; one for each host **compartment**:
$dS/dt = fxn$; $dI/dt = fxn$; $dR/dt = fxn$
S, I & R are numbers of individuals; the other variables are rates



After Anderson & May (1979ab) & May (1983)

$$\frac{dX}{dt} = a(X+Y+Z) - bX - \beta XY + \gamma Z$$
$$\frac{dY}{dt} = \beta XY - (\alpha + b + \upsilon)Y$$
$$\frac{dZ}{dt} = \upsilon Y - (b+\gamma)Z$$

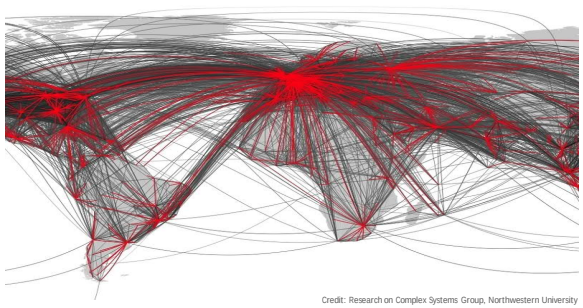
Table 1 The influence of various types of directly transmitted microparasites on host population growth

Type of disease	Growth characteristic (disease regulates host population if expression is negative)	Threshold host population, for successful introduction of the disease
Horizontal transmission		
No immunity ($\gamma \to \infty$)	$r - \alpha$	$(\alpha + b + \upsilon)/\beta$
Life-long immunity ($\gamma = 0$)	$r[1 + (\upsilon/b)] - \alpha$	$(\alpha + b + \upsilon)/\beta$
Transient immunity (duration $1/\gamma$)	$r[1 + \upsilon/(b+\gamma)] - \alpha$	$(\alpha + b + \upsilon)/\beta$
Transient immunity and an incubation (latent) period of duration $1/\sigma$	$r\left[1 + \frac{\upsilon}{(b+\gamma)} + \frac{(\alpha + b + \upsilon)}{\sigma}\right] - \alpha$	$\frac{(\alpha + b + \upsilon)(b+\sigma)}{\beta\sigma}$
Transient immunity and disease eliminates reproduction of infected class	$r\upsilon/(b+\gamma) - (b+\alpha)$	$(\alpha + b + \upsilon)/\beta$
Transient immunity and disease reduces birth rate of infected class to fa	$r\left[\frac{fa - b}{r} + \frac{\upsilon}{(b+\gamma)}\right] - \alpha$	$(\alpha + b + \upsilon)/\beta$
Vertical (and horizontal) transmission		
Transient immunity and all births from infected class are also infected	$r[1 + \upsilon/(b+\gamma)] - \alpha$	$(\alpha + b + \upsilon - a)/\beta$; threshold is zero if $a > \alpha + b + \upsilon$
Transient immunity and a fraction f of births from infected class are also infected	$r[1 + \upsilon/(b+\gamma)] - \alpha$	$(\alpha + b + \upsilon - fa)/\beta$; threshold is zero if $fa > \alpha + b + \upsilon$

Reproduced from Anderson and May (1979)

Nature of Inquiry
(Aim of Modeling)?

International Air Travel



Credit: Research on Complex Systems Group, Northwestern University

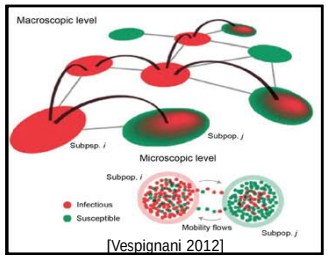
Dynamics on Networks: Super-Spreaders

By Madotto & Liu

Key to Unveiling *Meta-Population* Transmission

- To understand, predict, and control epidemic dynamics by characterizing age-specific or spatial *sub-populations*

$$I_{t+1} = \mathbb{K}_t I_t = g(S_t B C A) I_t$$



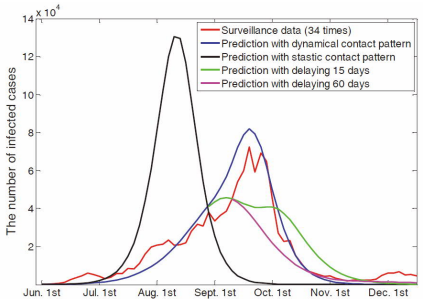
[Vespignani 2012]

S_t: Susceptible population
B: Infection acquiring rate
C: Contact matrix
A: Infection transmission rate

Contact: Individuals' mutual exposure in the same physical environment

By Yang, Pei, Xia, & Liu, *et al.*

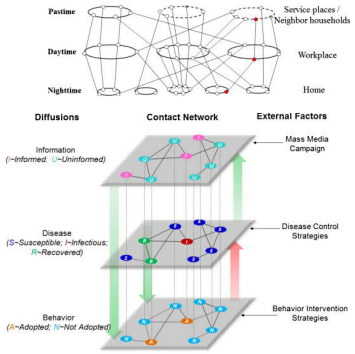
2009 HK **H1N1** Influenza



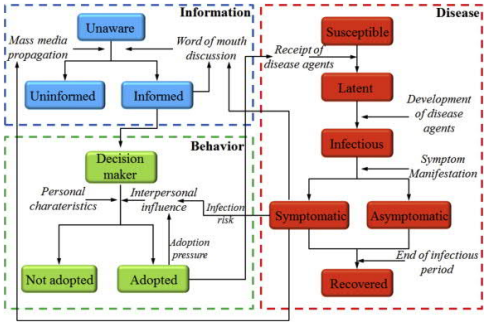
Predictions and quantitative evaluation of different strategies

Contact Networks

- Simultaneous diffusion of disease, information, and behavior (simulating "triple-diffusion" in metropolitan areas)
- "Results *reasonably* replicate observed influenza spread and information propagation."



Liang Mao, Modeling triple-diffusions of infectious diseases, information, and preventive behaviors through a metropolitan social network—An agent-based simulation, Applied Geography, Volume 50, June 2014, Pages 31-39



Nature of Inquiry
(Aim of Modeling)?

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Using data-driven agent-based models for forecasting emerging infectious diseases

Srinivasan Venkatramanan[a,*], Bryan Lewis[a], Jiangzhuo Chen[a], Dave Higdon[b,c], Anil Vullikanti[a,d], Madhav Marathe[a,d]

[a] Network Dynamics and Simulation Science Laboratory, Biocomplexity Institute of Virginia Tech, United States
[b] Social and Decision Analytics Laboratory, Biocomplexity Institute of Virginia Tech, United States
[c] Department of Statistics, Virginia Tech, United States
[d] Department of Computer Science, Virginia Tech, United States

ARTICLE INFO

Article history:
Received 2 July 2016
Received in revised form 30 January 2017
Accepted 17 February 2017
Available online xxx

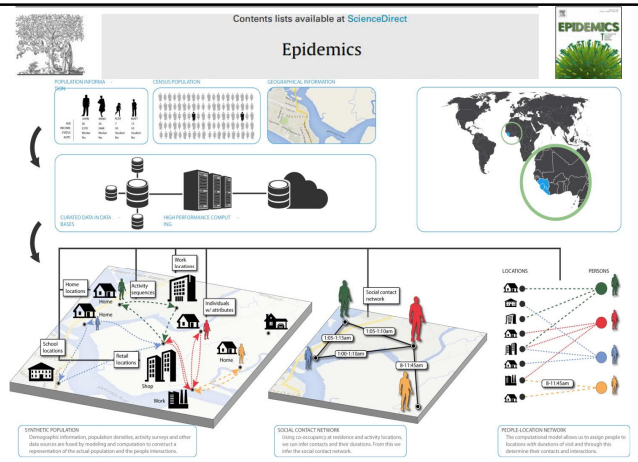
Keywords:
Emerging infectious diseases
Agent-based models
Simulation optimization
Bayesian calibration
Ebola

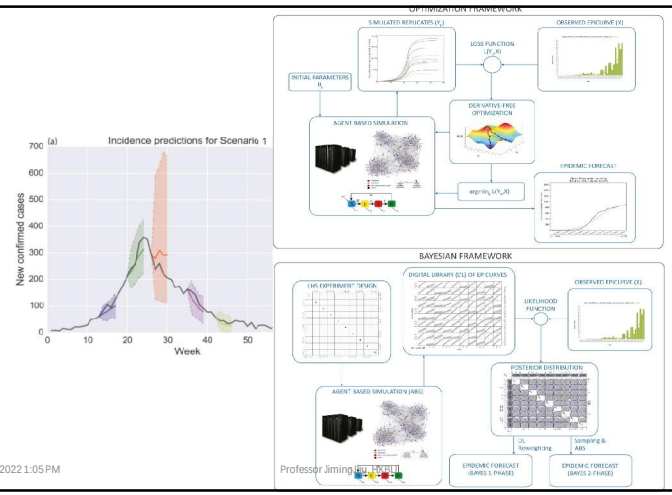
ABSTRACT

Producing timely, well-informed and reliable forecasts for an ongoing epidemic of an emerging infectious disease is a huge challenge. Epidemiologists and policy makers have to deal with poor data quality, limited understanding of the disease dynamics, rapidly changing social environment and the uncertainty on effects of various interventions in place. Under this setting, detailed computational models provide a comprehensive framework for integrating diverse data sources into a well-defined model of disease dynamics and social behavior, potentially leading to better understanding and actions. In this paper, we describe one such agent-based model framework developed for forecasting the 2014–2015 Ebola epidemic in Liberia, and subsequently used during the Ebola forecasting challenge. We describe the various components of the model, the calibration process and summarize the forecast performance across scenarios of the challenge. We conclude by highlighting how such a data-driven approach can be refined and adapted for future epidemics, and share the lessons learned over the course of the challenge.

Contents lists available at ScienceDirect

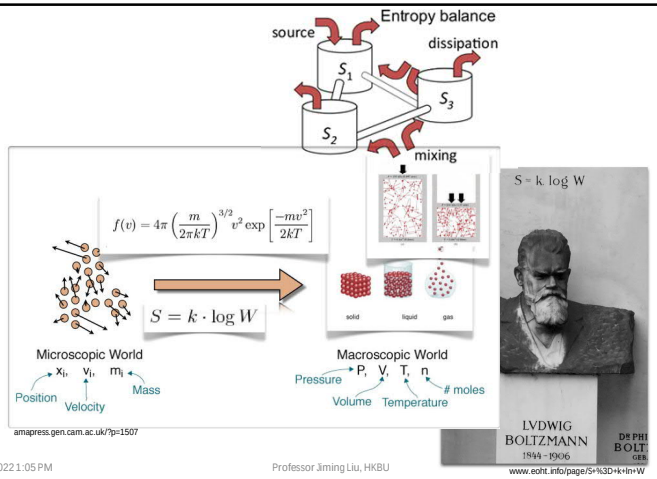
Epidemics





Q2: ...*Right Model*,
at *Right Scale*,
for *Right Inquiry*?

Q3: Multiple *Scales* are
Inter-Related... How?



Recommended Readings