Beyond big data:
Identifying important information for real world challenges

Yaneer Bar-Yam and Maya Bialik

New England Complex Systems Institute

238 Main St., Suite 319, Cambridge, MA 02142, USA

(Dated: December 17, 2013)

Abstract

Much of human inquiry today is focused on collecting massive quantities of data about complex systems, with the underlying assumption that more data leads to more insight into how to solve the challenges facing humanity. However, the questions we wish to address require identifying the impact of interventions on the behavior of a system, and to do this we must know which pieces of information are important and how they fit together. Here we describe why complex systems require different methods than simple systems and provide an overview of the corresponding paradigm shift in physics. We then connect the core ideas of the paradigm shift to information theory and describe how a parallel shift could take place in the study of complex biological and social systems. Finally, we provide a general framework for characterizing the importance of information. Framing scientific inquiry as an effort to objectively determine what is important and unimportant rather than collecting as much information as possible is a means for advancing our understanding and addressing many practical biological and social challenges.

I. INTRODUCTION

How people combine to form social structures and global economic markets (and their crises) or how our proteins combine to form our functioning (or dysfunctioning) bodies are the kinds of questions we must answer to address the challenges facing us today. These questions are not just more complicated than questions about simple systems—they are qualitatively different. A key objective is knowing how to intervene in order to have a desired effect, which requires understanding the vulnerabilities, opportunities and levers of change of the system. For complex systems, this goal requires a deeper analysis than the traditional scientific methodologies, which separate the micro and the macro scales, and thus a more sophisticated strategy than collecting as much information as possible. The methodology described here that focuses on patterns rather than statistics is based on an advance in physics, but can be generalized and applied to various complex systems. Understanding complex systems through the new tools uniquely suited to address these patterns will enable us to design effective interventions for pressing societal questions.

There are two main reasons that traditional methods, which have been successful in our study of simple systems, begin to break down when applied to complex systems. The first is that parts are neither independent nor coherent; the second is that they form actions that occur across scales ranging from microscopic to macroscopic.

In complex systems [1], the units we are describing are acting neither totally independently nor totally coherently; rather, they are interdependent, both influencing each other and compelled by common causes. An example can be found in commodity markets. The traditional theory of markets assumes that people decide on investments independently and rationally, and therefore predicts a supply and demand equilibrium. Interestingly, it is not so much the assumption of rationality that does not hold up in complex systems analyses of markets today, but rather the independence. The breakdown of equilibrium due to trend-following has been well-established since 1990 [2], but the theory at that point, subject to the constraints of the concepts and mathematics of traditional economics, was not able to represent the dynamics after the breakdown. In a complex systems analysis of the commodity markets [3], actions of individuals are not fully independent; rather, due to trend-following, people make decisions that influence and are influenced by the decisions of others. These influences lead individual actions to combine into collective oscillations (Fig. 1).

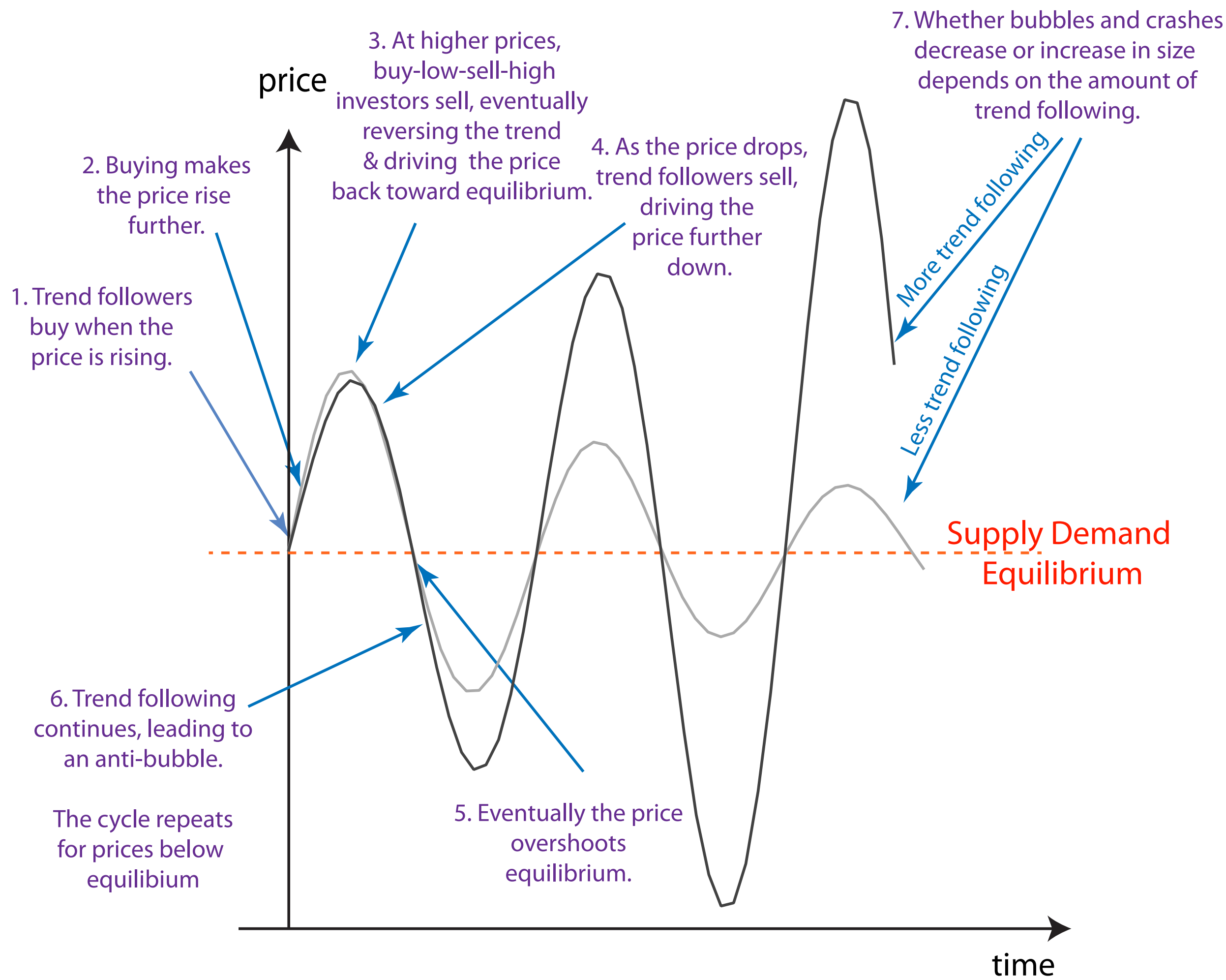

FIG. 1: Bubbles and crashes in market prices result from bandwagon effects when price changes themselves motivate traders, leading to mutual influence between traders, as buying can cause more buying and selling can cause more selling. More precisely, bubbles arise as an interplay between two different kinds of investors, trend-following speculators who buy when prices are increasing and sell when they decrease, and fundamental investors who buy-low-and-sell-high.

If the price of a commodity happens to go up, trend followers, who jump on the bandwagon, push the price further away from equilibrium. The further the price is from equilibrium, however, the more the conventional investors who “buy low and sell high” get involved, and their selling provides a force driving the price back towards equilibrium. Indeed, the further the price deviates from equilibrium, the more this “Walrasian force” strengthens, eventually reversing the upward trend. At this point, the bandwagon effect drives the price down (the crash!), eventually overshooting equilibrium, and restarting the cycle. This interplay results in an oscillation of prices deviating from equilibrium. Rather than calculating an

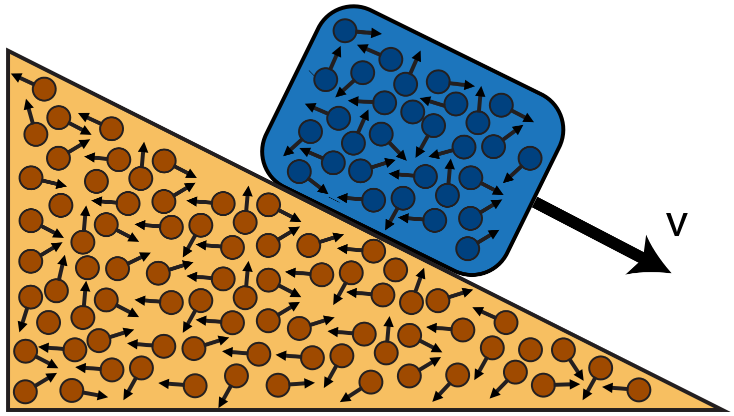

FIG. 2: Schematic diagram of a block (with a velocity at a particular moment, v) sliding down an inclined plane. The macroscopic motion subject to gravity and friction may be treated using Newton's laws of motion, while the microscopic behavior of the atoms may be treated using thermodynamics by considering the local oscillations of groups of atoms as random and independent; the statistical treatment of that movement leads to the determination of pressure and temperature of the block and the inclined plane.

unstable equilibrium price (which would make sense if people were acting entirely independently), this methodology identifies the large scale pattern of the system, and successfully maps it onto the bubble and crash dynamics, accurately describing global food prices [3].

Why didn't traditional methodologies have the ability to consider such interdependencies? A key to their limitation is that they are applicable only to systems in which there is a separation of behavior between the micro and macro scales. Interactions among the parts that cause large scale behavior, like the trend following induced market bubbles, violate this separation. This is not generally an issue in the simple systems that were the essence of academic science until recently.

Consider a block sliding down an inclined plane. To address the micro scale—the molecules—we average over them and, using thermodynamics, describe their temperature and pressure. To address the macro scale—the block and the inclined plane—we use Newtonian physics to talk about their large scale motion (see Fig. 2). In this case, the pieces can be considered to be acting either independently (the random relative motion on the micro scale) or coherently (the average motion on the macro scale) and since the scales are sufficiently distinct we do not encounter a problem in describing them separately.

But many systems, especially those we are interested in understanding and influencing, are not well described by two separate levels. Consider a flock of birds as it leaves a tree in response to a crack of thunder. If all of the birds flew independently in different directions, we would need to describe each one separately. If they instead all went in the same direction, we could simply describe their average motion. On the other hand, if we were interested in how the birds move from the initial mostly independent directions of movement, into groups and then came together as a flock, describing each bird's motion would be too much information and describing the average would be too little information. Understanding the complex behavior at the transition from independent to coherent behavior is best described across scales, and thus requires knowing which information is relevant: important at the scale of interest.

This example can be framed in a way that is generalizable to a wide range of complex systems. We will begin by examining an archetypal example from physics, in which the transition from independent to coherent actions happens across the entire system and the units we examine are at least conceptually simple. Most importantly, this example has attracted attention because of an undeniable contradiction between traditional theoretical results and empirical results, compelling a reevaluation of traditional theory. This provided inspiration for a profoundly important advance in science.

II. A PARADIGM SHIFT IN PHYSICS

The archetypal case of a system undergoing a transition from independent to dependent parts is when matter undergoes a phase transition, for example from liquid to gas. While many of the domains that the tools described here will be useful for (such as societal questions) are more complicated, we begin with a key simpler example to illustrate the idea, and then generalize to more complicated complex systems.

A. The Original Puzzle

This issue drew attention due to a mismatch between the theoretical expectation and the empirical observation of the characteristics of the critical point in the phase transitions of water [4].

It is well known that as the temperature and pressure of water change, so does its density (see Fig 3). There is a line of sharp transitions from the liquid to the vapor phase at a

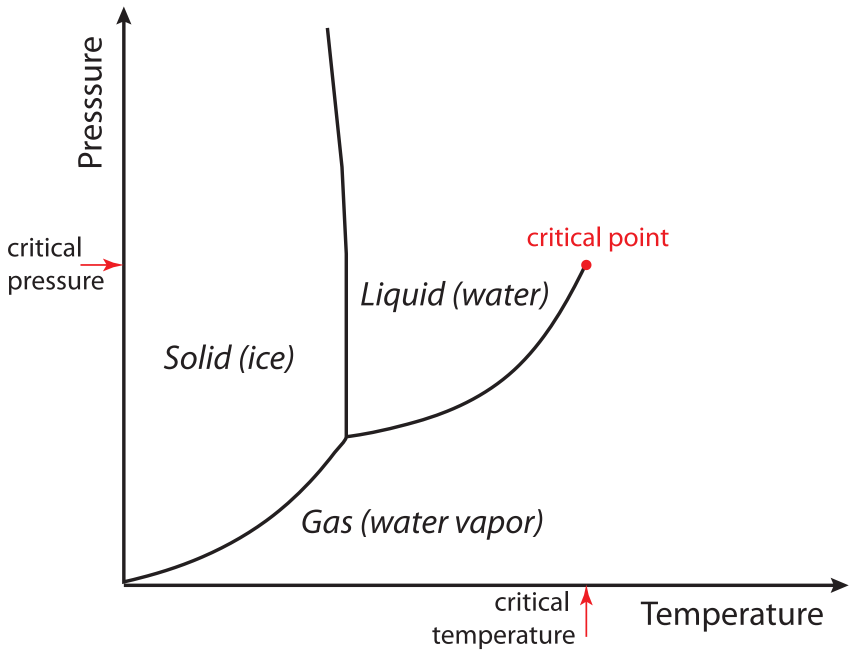

FIG. 3: The phase diagram of water. The line of transitions between liquid water and water vapor stops at the critical point (red dot). At that point the fluctuations between liquid-like and vapor-like densities extend across the system so that the system is not smooth (violating the assumptions of calculus) and averages are not well behaved (violating the assumptions of statistics). A new method that considers behaviors across scales, renormalization group, was developed to address this and similar questions.

temperature that increases with increasing pressure.

What happens at the critical point, where the line stops? The temperature is so high that the molecules have a lot of energy and are not able to cohere together into a liquid, but the pressure is so high that they are not able to fully separate and act independently as in a gas.

We therefore have a case of the scenario described above, with the units acting neither fully independently nor fully coherently, and the separation of scales begins to break down as we consider the behavior of the phase transition itself. Along the phase transition line, there is a discontinuity in the density—a clear distinction between the liquid phase and the gas phase. As conditions approach the critical point, the discontinuity disappears in the form of a 'power law.' A power law is a relationship in which one quantity does not grow in direct proportion to the other, but rather grows as the other to a power β, $y = ax^{\beta}$. This is precisely the point at which theory did not match the empirical findings. For phase transitions, the nature of the power law (its exponent, β) was supposed to be 0.5 according to traditional theory (Landau theory) [5], but was determined to be 0.33 empirically [6, 7].

B. The Solution

Our usual methods of calculus and statistics fail at this point because their assumptions no longer hold true. Calculus assumes that matter is smooth and statistics assumes that averages over large numbers of objects are well defined. Away from the critical point these assumptions are justified, since the microscopic behavior of atoms is well separated from the macroscopic behavior of the material as a whole. Different parts of the material appear essentially the same, making it smooth, and any average over atomic properties has a single well defined number. However, at the critical point, the density fluctuates—between water-like and vapor-like conditions—so that the material is not smooth and the average taken of the material as a whole is not representative of the density at any particular location or time. Near the critical point, the matter is composed of patches of lower and higher density, and this patchiness occurs on all scales.

How can we effectively characterize the behavior of the system to account for the mismatch between the theoretical prediction and the empirical observation? The answer to the puzzle lies in capturing more information than just the average, but less than describing each molecule. A new method called renormalization group [8, 9, 10, 11, 12] was developed in order to determine which information was "relevant," i.e. important at the largest scales. The renormalization group method showed that the relevant information that the traditional theory was missing by averaging was the variability of density across the patches. This variability can't be ignored because it is large enough that changes in densities at different locations interact with each other. The idea however, can be generalized to the analysis of many different complex systems by connecting it with information theory.

III. INFORMATION THEORY AND SCALES

As in physics, the main problem in trying to address many real world challenges in biological and social systems with big data is that describing every single possible variable is too much information, but describing only the average is not enough to capture how the system works. We need some more information, but only about the relevant variables. From an information theory perspective, a faithful representation must have the same number of states as the system it is representing. This enables the states of the representation to be mapped one to one to the states of the system. If a model has fewer states than the system, then it can't represent everything that is happening in the system. If a model has more states, then it is representing things that can't happen in the system.

Conventional models often do not take this into account and this results in a mismatch of the system and the model; they are unfaithful representations and do not properly identify vulnerabilities, opportunities, and levers of change in the system. Because we are interested in influencing the system, having too much information is counterproductive; we only want to know the most important factors.

To formalize these ideas, it is useful to understand information as related to scale. The complexity profile [1, 15] represents the amount of information as a function of scale. Typically, the finer the scale of inquiry about a system, the more information is needed to describe it (Fig. 4).

A sufficient representation, therefore, is one that has a set of possible states corresponding to the set of *distinguishable* states of the system at each level of resolution, down to the level we need to describe the properties we care about, the relevant parameters, and no further. Rather than accumulating details about the system, we should start with the largest scale pattern of behavior and add additional information only as needed. According to the complexity profile, each piece of information about a system has a size—the largest scale at which we can begin to detect that piece of information.

This relationship between the physics paradigm shift and information theory allows us to generalize these new ideas to important questions about all complex systems. Consider attempting to decide on policies about regulating the stock market based on a model with as much information as possible about as many variables as possible, as opposed to a sufficient model with only the relevant one or two parameters. When we have only a few parameters we can hope to validate and verify our understanding. Modeling systems according to this guideline allows one to disentangle the interdependencies, identify the structure of the system, and understand how to act upon it in order to achieve the desired large scale effect.

The methodology of renormalization group offers us more than just the idea. We gain formal guidance about how to construct models based upon the aggregation of components to identify the relevant parameters.

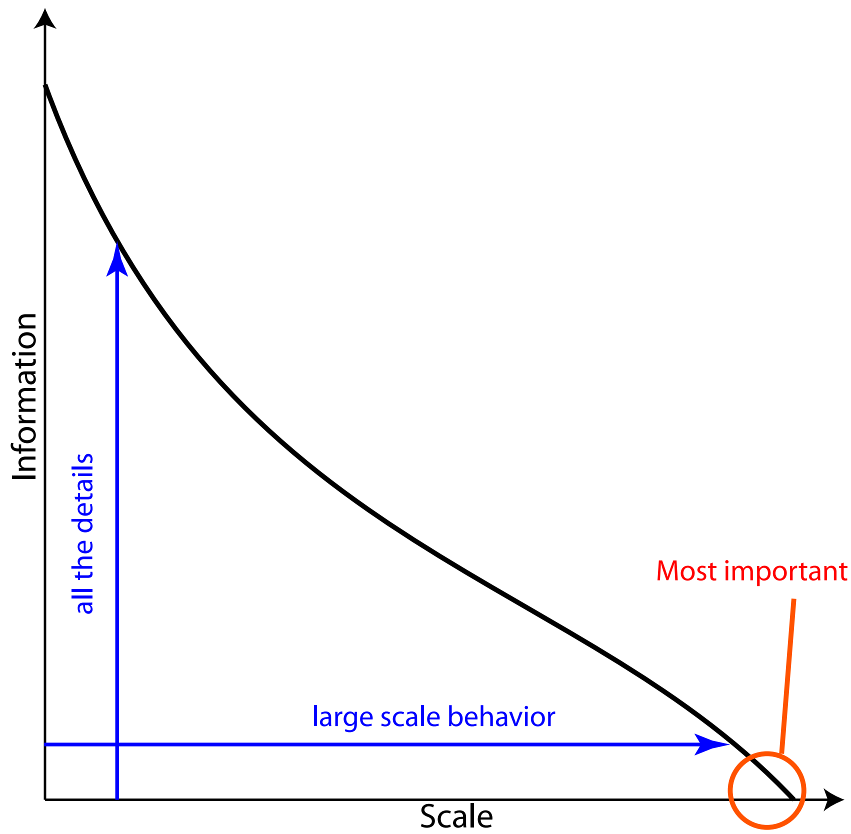

FIG. 4: The complexity profile is the amount of information that is required to describe a system as a function of the scale of description. Typically, larger scales require fewer details and therefore smaller amounts of information. The most important information about a system for informing action on that system is the behavior at the largest scale.

As we increase the scale we see fewer details. Small distinctions disappear and only larger distinctions that involve many parts of the system together remain. How properties of parts aggregate determines what is observed, i.e. what is important. By studying the way that properties aggregate we can identify the important larger scale system properties. Aggregation is determined by how the parts depend on each other. The simplest case is when they are independent; in this case, aggregation is what we know from statistics that gives rise to the average and the random deviations described by the normal (Gaussian) distribution. When there are other kinds of dependencies, different behaviors occur. These include behaviors such as dynamical oscillations, spatial patterns, and so on.

Understanding how to recognize these behaviors requires another lesson gained from the renormalization treatment of phase transitions: the surprising equivalence of water boiling to vapor and the loss of magnetization as a magnet is subjected to increasing temperatures.

IV. UNIVERSALITY

A. Other materials

It turns out that the relevant parameters of the phase transition of a liquid correspond perfectly to the relevant parameters of many magnets. This is neither coincidence nor mere analogy, but a direct mathematical relationship. Magnets have local magnetizations that fluctuate at a magnetic critical point just like the density at the water-vapor critical point. Just as molecules of water undergo a transition from order to disorder as the temperature increases, so do the little magnets undergo an order-to-disorder transition. The local magnetizations interact with each other, just like the density fluctuations in water at the critical point, and the result is that these two seemingly different types of systems map mathematically onto each other.

This idea has guided physicists in addressing diverse questions about the structure and dynamics of materials [13, 14]. We can consider these concepts even more generally for complex systems.

B. Universality in complex systems

When we go to the largest scale behaviors of a system, we simplify the mathematical description of the system, so that there are only a very limited set of possible behaviors that can happen. Just like with the mapping of water to vapor transitions onto magnetic transitions, one type of behavior can describe many possible systems.

In a particular sense, this idea is used in traditional theory as scientists use the normal distribution for many different biological and social systems. Why can the same distribution be used for all of these cases? The reason is that when a system has independent parts, the way they aggregate is the same, and the result is the normal distribution as the largest scale behavior of the system. When there are dependencies, the normal distribution no longer applies, but there are behaviors that are characteristic of other kinds of dependencies. To generalize, we have to determine the way kinds of dependencies give rise to kinds of large scale behavior. A different example is when scientists use wave equations to describe everything from music strings to water waves to light. Even though these are very different systems, the dependencies that give rise to their behaviors, and the behaviors themselves, are related

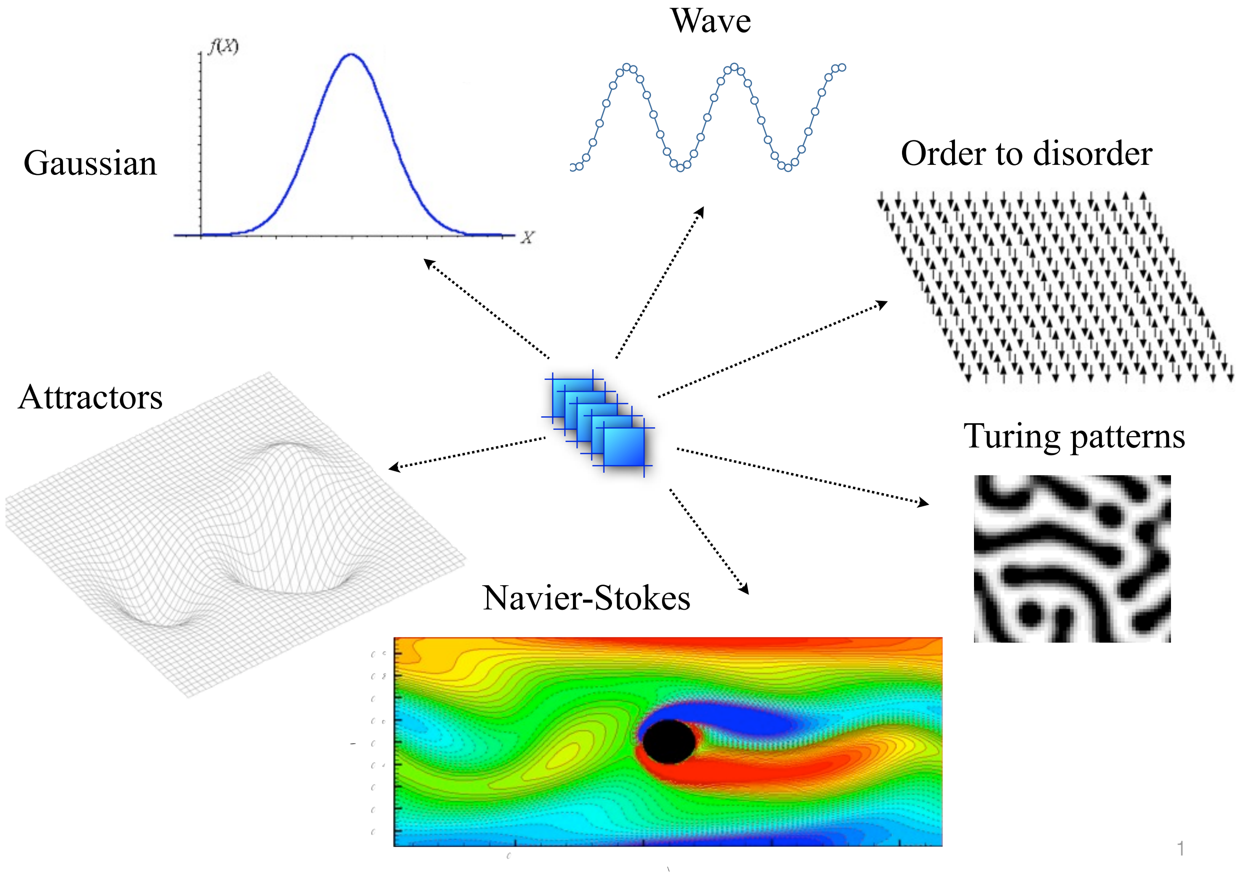

FIG. 5: The idea of universality recognizes that systems map onto a small set of large scale models, each of which applies to a large set of possible systems with widely different micro details. Examples shown in this figure: the normal (Gaussian) distribution, wave motion, order to disorder transitions (the subject of phase transitions discussed in the text), Turing spatial patterns, fluid flow described by Navier-Stokes equations, and attractor dynamics.

mathematically. This shows how many systems can have the same behavior even though they differ in detail, a concept called universality (Fig. 5).

Universality enables us to study classes of systems whose behaviors can be described the same way, and can be captured by a mathematical model. The best way to think about this is that the mathematical model is one member of the class. This is the principle of universality that is formalized by the analysis of renormalization group and generalized by the principles we described here of how to apply information theory to the scientific study of complex systems.

V. SENSITIVITY AND CHAOS

Applying the complexity profile to biological and social systems presents challenges that are important to recognize. Consider micro to macro connections in biology. A single genetic mutation in the β-globin gene, substituting valine for the glutamic acid at position 6, causes abnormal hemoglobin molecules. One such mutation cases sickle cell trait, which provides resistance to malaria, while two such mutations in an individual lead to sickle cell disease. Consider micro to macro connections in society. A single individual's idea, Steve Jobs' concept of an iPhone, changed the way hundreds of millions of people work and play. These large scale differences in the physiology of an organism or a society over time may seem to run counter to the crux of the point we are making and to differentiate biological and social systems from physical ones. But similar sensitivities to small scale events affect physical systems through the "butterfly effect" where small initial variations grow over time to have large scale impacts.

The large scale impacts of mutations and ideas arise from the possibility of informational replication over time that enables them to achieve large scale. Biological mutations can have large impacts because of the replication of DNA throughout the cells of the body, and across populations, and through the subsequent transcription of their information to many proteins that function in particular ways. Social systems have sensitivity to specific ideas as they are transmitted to others, embodied in machinery and organizational processes that mass produce and distribute them widely. The butterfly effect similarly arises when the information that is in a small scale motion is amplified over time by the available energy sources in heated oceans and embodied in the highly redundant large scale motion of a hurricane.

These processes are not counter to the framing of information importance. However, they do make its application more challenging as we need to understand the way that information is replicated over time. That there is sensitivity to micro information does not mean that all micro information can or will end up as large scale system behaviors. Not all molecular changes, or even genetic mutations, have large scale impact. Similarly, ideas and individuals that can change society as a whole are rare. Whether and to what extent a large scale behavior is sensitive over time to the impact of small scale events—which have increasing impact as they are replicated—is part of the analysis of the scale of information in the behavior of a complex system. In every case, understanding what is the large scale information is essential to the analysis and our eventual understanding of the system.

VI. CONCLUSIONS

While it is tempting to believe that the current focus on big data is bringing us closer to understanding the systems around us, we claim this is not the case. Describing in detail all the molecules that make up the people who are investing in a market will not enable us to answer policy questions. We have shown that considering the importance of information (scale) is crucial for understanding the vulnerabilities and levers of change of a system. By connecting the paradigm shift in phase transition theory to information theory, we have described how this idea may be applied to the study of all complex systems

Can we demonstrate that this approach can work for biological and social systems? In recent years we have tested this approach in addressing a number of important biological and social questions. We have developed theories that (1) accurately describe the locations of ethnic violence [16, 17] by looking at the universal properties of the separation of distinct types (in this case ethnic groups) similar to the separation of immiscible metals by spinodal decomposition, (2) accurately describe geographic distributions of biodiversity [18] by considering spontaneous spatial pattern formation in a high dimensional type of Turing patterns, (3) accurately describe the dynamics of global food prices [3, 19] by looking at the first order equations describing the dynamics of investor decision making. We have also contributed [20, 21, 22, 23, 24] to a raging controversy in evolution about the relevance of kin and group selection [25, 26, 27, 28], a key component of which is a direct translation of the failure of Landau theory onto biological evolution. Just as for the case of water boiling, averages across an evolving population don't describe evolutionary dynamics when there are fluctuations across space in its genetic composition. The breakdown of the approximations used in evolutionary biology result in mis-characterization of the evolution of altruism, which need not arise due to kin selection, but rather can robustly arise by association due to proximity in a spatially patchy system. In each of these questions, our analysis is guided by an understanding that is mostly not described in papers, as the motivation is hidden by assuming a hypothesis or model. But how can the right model be guessed? It can be inferred through an understanding of relevant variables.

The technical demands of determining relevance and irrelevance of information may seem to make modeling more difficult than it has to be. Why don't we just include more details? If we include enough, won't a model be correct? The answer is no, for two reasons. The first is sufficient but the second is more important. The first reason is that including many details is not sufficient since without attention to determining what is and is not important we cannot tell whether we have included the details that matter. The second reason is that including many details that don't matter actually prevents us from addressing the question we really want to answer: which levers are important. Determining the levers that are important is equivalent to determining what is important at the larger scales. Thus the questions we really want to answer about systems are exactly the same as determining which are the relevant variables.

Considering systems in this way, we should recognize that any mathematical model, and indeed any description or characterization, whether from theory or phenomenology, in words, pictures, movies, numbers or equations is "valid" only because of the irrelevance of details. Moreover, such information applies across different instances for the same reason: the sufficiency of the representation, having captured the important variables. Any two systems we look at (or the same system at different moments in time, or different cases of the same system) are different in detail. If we want to say anything meaningful about a system—meaningful in the sense of scientific replicability or in terms of utility of knowledge—the only description that is important is one that has universality, i.e. is independent of details. There is no utility to information that is only true in a particular instance. Thus, all of scientific inquiry should be understood as an inquiry into universality—the determination of the degree to which information is general or specific.

We thank Irving Epstein for comments on an earlier version and Karla Z. Bertrand, Blake Stacey, Casey Friedman and Dominic Albino for editorial comments.

---

[1] Y. Bar-Yam, Dynamics of Complex Systems (Perseus Press, Reading, 1997). http://necsi.edu/publications/dcs/

[2] J. B. De Long, A. Shleifer, L. H. Summers and R. J. Waldmann, Positive feedback investment strategies and destabilizing rational speculation, *The Journal of Finance* **45**, 379 (1990).

[3] Marco Lagi, Yavni Bar-Yam, Karla Bertrand and Yaneer Bar-Yam, The food crises: A quantitative model of food prices including speculators and ethanol conversion, *arXiv:1109.4859 [q-fin.GN]* (2011). http://necsi.edu/research/social/foodprices.html

[4] J. M. H. Levelt Sengers and S. C. Greer,Thermodynamic anomalies near the critical point of steam, *Intern. J. of Heat and Mass Transfer* **15**, 1865 (1972).

[5] L. D. Landau, *Phys. Zurn. Sowjetunion* **11**, 545 (1937); English translation: D. ter Haar, Men of Physics: L. D. Landau. Vol II (Pergamon, Oxford. 1969).

[6] P. Heller, Experimental investigations of critical phenomena *Repts. Prog. Phys.* **311**, 731 (1967).

[7] L. P. Kadanoff, W. Gtze, D. Hamblen, R. Hecht, E. A. S. Lewis, V. V. Palciauskas, M. Rayl. J. Swift, D. Aspres and J. Kane, Static phenomena near critical points: Theory and experiment, *Rev. Mod. Phys.* **39**, 395 (1967).

[8] K. G. Wilson, Renormalization group and critical phenomena, *Phys. Rev. B* **4**, 3174-3183 and 3184-3205 (1971).

[9] L. Kadanoff, Scaling, universality and operator algebras, in Phase Transitions and Critical Phenomena, C. Domb and M. S. Green. Eds. Vol. 5a (Academic. New York, 1978) p. 1.

[10] D. R. Nelson, Recent developments in phase transitions and critical phenomena, *Nature* **269**, 379 (1977).

[11] C. Domb, The critical point: A historical introduction to the modern theory of critical phenomena (Taylor & Francis, London, 1996).

[12] M. Kardar, Statistical physics of fields, (Cambridge University Press, Cambridge, England, 2007).

[13] P.-G. de Gennes, Scaling concepts in polymer physics (Cornell University Press, Ithaca, NY, 1979).

[14] M. Kardar, G. Parisi, and Y.-C. Zhang, Dynamic scaling of growing interfaces, *Phys. Rev. Lett.* **56**, 889 (1986).

[15] Y. Bar-Yam, Multiscale Variety in Complex Systems, *Complexity* **9**, 37-45 (2004).

[16] M. Lim, R. Metzler and Y. Bar-Yam, Global Pattern Formation and Ethnic/Cultural Violence, *Science* **317**, 1540 (2007). PMID: 17872443.

[17] A. Rutherford, D. Harmon, J. Werfel, S. Bar-Yam, A. Gard-Murray, A. Gros and Y. Bar-Yam, Good fences: The importance of setting boundaries for peaceful coexistence, *arXiv: 1110.1409*

(2011).

[18] M. A. M. de Aguiar, M. Baranger, E.M. Baptestini, L. Kaufman and Y. Bar-Yam, Global Patterns of Speciation and Diversity, *Nature* **460**, 384-387 (2009).

[19] Marco Lagi, Yavni Bar-Yam, Karla Bertrand and Yaneer Bar-Yam, UPDATE February 2012: The food crises: Predictive validation of a quantitative model of food prices including speculators and ethanol conversion, *arXiv:1203.1313 [physics.soc-ph]* (March 6, 2012). http://necsi.edu/research/social/foodprices/update/

[20] Y. Bar-Yam, Formalizing the gene-centered view of evolution, *Advances in Complex Systems* **2**, 277-281 (1999).

[21] H. Sayama, L. Kaufman and Y. Bar-Yam, Symmetry breaking and coarsening in spatially distributed evolutionary processes including sexual reproduction and disruptive selection, *Physical Review E* **62**, 7065-7069 (2000).

[22] E. Rauch, H. Sayama and Y. Bar-Yam, Relationship between measures of fitness and time scale in evolution, *Phys. Rev. Lett.* **88**, 228101 (2002).

[23] J. K. Werfel and Y. Bar-Yam, The evolution of reproductive restraint through social communication, *PNAS* **101**, 11019-11024, (2004).

[24] B. C. Stacey, A. Gros and Y. Bar-Yam, Beyond the mean field in host-pathogen spatial ecology, *arXiv:1110.3845 [nlin.CG]* (November 25, 2012). http://necsi.edu/research/evoeco/

[25] M. A. Nowak, C. E. Tarnitam and E. O. Wilson, The evolution of eusociality, *Nature* **466**, 1057-1062 (26 August 2010)

[26] P. Abbott, et al, Inclusive fitness theory and eusociality, *Nature* **471**, E1-E4 (24 March 2011)

[27] G. Wild, A. Gardner and S. A. West, Adaptation and the evolution of parasite virulence in a connected world. *Nature* **459**,983-986 (18 June 2009).

[28] M. Wade, D. S. Wilson, C. Goodnight, D. Taylor, Y. Bar-Yam, M. A. M. de Aguiar, B. Stacey, J. K. Werfel, G. Hoelzer, E. Brodie, P. Fields, F. Breden, T. Linksvayer, J. Fletcher, P. Richerson, J. Bever, J. D. Van Dyken and P. Zee, P. *Nature* **463**, E8-E9 (18 February 2010)