Center for Science of Information

Grand Challenges –DRAFT!!! November 2013

This document summarizes the grand challenges of the Center for Science of Information (CSoI) in its three main thrusts: knowledge extraction, life sciences, and communication. We developed these challenges through our work in the Center, and we refined them during two special CSoI workshops in March 2012 at Princeton and in March 2013 at the CSoI "Big Data Workshop." We periodically update this working document.

Challenges in Knowledge Extraction

Information theory has enjoyed great success in establishing fundamental limits for problems related to fairly simple classes of random processes. Specifically, traditional analysis considers memoryless processes and, to a lesser extent, stationary ergodic sequences. The most celebrated of these successes includes the channel and source-coding theorems. However, many real-world domains exhibit complexities that violate the assumptions used to derive these fundamental results.

Particularly in the new area of "big data," many domains consist of data with one or more of the following characteristics:

- Very large scale (with possibly distributed storage)
- High dimensional
- Multiscale patterns
- Dynamic, temporal, or geographic effects
- Heterogeneous structure
- Complex dependencies
- Noisy records

Example Data

<insert intro text here?>

Sequence Data

With the advent of inexpensive sequencing technologies (next-generation sequencing and beyond), vast amounts of genomic data must be securely stored. This data is expected to increase in volume significantly in the foreseeable future. Sequencing data takes the form of 'reads'—typically sequences of 100–300 characters (nucleotides). The reads are mapped to genome sequences/locations. To provide a perspective, the human genome sequence had circa 3 billion characters. These sequences are also accompanied by confidence values (estimates of correctness of individual nucleotides). It is desirable to store all the reads (as opposed to mapped variants). Operations on this data include mapping of reads to

reference genomes (approximate matching to a large string), identification of variants (is there a variant at a specific location in the genome), identification of all variants, and computing partial assemblies. While traditional compression techniques can be developed for these reads (associated with quantization-based techniques for confidence values), these do not support required analyses.

Network Data

Network databases are increasingly used to store interrelationships in domains ranging from social networking, economic transactions, and biological processes. Nodes and edges in these networks may themselves be multi-attributed (vectors or tensors). Networks are typically large (hundreds of millions of nodes and beyond), noisy, and often dynamic (changing with time). Analyses on networks range from lookups of subnetworks to more comprehensive tasks of identifying modularity, motifs, divergence, pathways, and causality of interactions. While there is some recent work on network compression, the underlying techniques do not naturally lend themselves to efficient analysis. The noisy nature of network data necessitates statistical quantification of results (p-value, e-value, etc.). A closely related task is (network) model inference and quantification of network complexity. Extensions of methods to accommodate for fluxes through networks are important in chemical pathways and information flow.

Structural Data

Structure databases frequently arise in life sciences (such protein databases as PDB and small molecule databases) and manufacturing (parts databases). Queries in such databases relate to similar and complementary shapes (binding pockets and docking configurations), structural similarities (similar drug targets), and reconstruction from partial projections. Often, there is a dynamic aspect to these structures as well (conformational changes over time or course). Such molecular trajectories are extremely storage intensive, often making analysis infeasible.

Streaming Data

Tremendous data is generated from streams of events—ranging from business transactions to sensor data. Recent conservative studies estimate that in 2008, enterprise server systems in the world processed 9.57 × 1021 bytes of data. This number is expected to double every two years. Walmart servers handle more than one million customer transactions every hour, and this information is inserted into databases that store more than 2.5 PB of data. The Large Hadron Collider at CERN will produce roughly 15 PB of data annually. Each day Facebook operates on nearly 100 TB of user-log data and several hundreds of TBs of image data. Every minute 48 hours of video are uploaded to YouTube; 23,148 apps are downloaded from the iTunes App Store; 208,333 minutes of the Angry Birds video game is played via smartphones; and more than 28,000 MMS messages are sent every second. Twitter serves more than 200 million users, who produce more than 90 million tweets per day, or 800 per second. In the domain of real-world sensor systems, Boeing jet engines can produce 10 TB of operational information for every 30 minutes of operation. This corresponds to a few hundred TB of data for a single Atlantic crossing. Analyses on these streams take different forms— from correlation and anomaly detection to higher-order semantic analysis and association. At the scales at which this data is generated

are virtually no methods for analysis. A critical need exists for effective compressed analyses of streaming data.

High-Dimensional/Tensor Data

Challenges associated with sparse high-dimensional data sets are well recognized. As techniques for dealing with these data sets are developed, there is increasing recognition of the fact that the magnitude of this data poses significant challenges. Current approaches to this problem focus on distributing the data sets (virtually all major Internet infrastructure—Google, Facebook, Twitter use this) or relying on out-of-core computations to deal with the data sets. We posit that compressed analysis would yield one or two orders of magnitude improvement in storage and network cost in distributed storage. The challenge is to support analyses and query tasks, such as dimensionality reduction on compressed data, clustering, projections, search, tensor decompositions, and visualization.

Challenges

The largest challenge in modern times is the conversion of big data into context-specific knowledge. In the world view of Kant, data are representations of phenomena and our comprehension of the phenomena lies in the projection of this data into concepts, structure and categories (knowledge). Shannon laid the ground work for formulating this problem in an information theoretic frame and our goals are to extend the Shannon concepts.

Perhaps the greatest challenge in applying information-theoretic principles to a broader suite of problems—including biological systems, analytics for massive data sets, and social networks—is that of developing meaningful notions of "structural information" and establishing a set of corresponding fundamental results.
Specifically with respect to the analysis of big data, the data size and complexity leads to the following three main challenges with accuracy and efficiency tradeoffs.

1. *Discovering structure in data.* To identify, encode, and test the underlying structure of "big" data, a tradeoff exists between the accuracy of the data model and the amount of data needed to support or test the model. Specific instantiations of this include, deriving community structure from data, reconstructing networks that serve as a "connected map" of the discrete data elements, and obtaining causal inferences from correlations in time varying data.

2. *Managing and querying data.* To collect, store, and query "big" data in data management systems, a tradeoff exists between the efficiency (space) of the data structures and the efficiency (time) of the algorithms that query and access the data. The specific challenges include, interactive and computational querying of data, developing a dynamical "ontological structure" on the data and statistical learning from multidimensional heterogeneous data.

3. *Ranking findings and finding rankings in a very wide data set.* A very wide data set contains more features than records. By a finding, we mean here an association rule, a correlation, or a pattern. By a ranking, we mean a function that associates a measure of validity or interest with a

finding. Such ranking function may return a p-value combined with some measure of interest of a finding. Specific examples include, inferring patterns from complex data and associating a significance to this inference in a context-specific manner. Such statistical metrics will need to be computable along with the data for discovery of the inherent structure.

Since the size and complexity of "big" data make it difficult to visualize or explore manually, it is critical we have methods that recognize the structural information in the data. Fundamental results concerning limits in this area would help analysts and systems engineers optimize the tradeoff between efficiency and accuracy.

To exemplify the notion of structural information, consider a setting where we have two large data sets (say, for example, in the form of an undirected graph). How can we efficiently determine whether the data sets—modulo any additive noise—are generated by underlying mechanisms that are fundamentally similar or different? Tools that allow us to answer such questions would serve two objectives: (i) Knowledge of the underlying data structure would aid in our developing algorithms designed to extract useful information from the data set, and (ii) we could simulate and analyze complex data sets by assuming data is drawn from a statistical model exhibiting key structural properties.

In addition, we (as system designers) often have control over which data we collect, process, and store for analytics purposes. For example, if our ultimate goal is to perform queries on a database, how can we process and store the database so we can still query it efficiently and reliably? As another example, how can we collect and process online advertising data so that advertisers can maximize profit by serving relevant ads? The key point in this latter example is that we have some degree of control over which data is collected (e.g., clicks, cursor position, scroll speed, time spent on a page, etc.). Which data, and how much of it, should we collect to attain near-maximal profit?

The essential challenge here is one of determining data relevance with respect to a given objective. In order to accomplish a particular task (e.g., query, advertise, predict, etc.), which information is relevant? And what data-collection schemes permit us to infer the relevant information? To give a concrete example, suppose we want to determine the structure of friendships within a social network. Which sampling mechanisms will allow us to determine the structure reliably? Clearly, the definition of "structure" here is of great importance, and would be a topic of study in the above section.

Also related to the notion of data collection, organization, and processing is the concept of information security. Traditional information-theoretic security results assume adversaries with unbounded computational power, and therefore, give overly conservative results (e.g., any "good" encryption requires a one-time pad). On the other hand, currently employed security measures have little theoretical foundation and rely on our belief that certain problems (e.g., factoring large numbers) are computationally difficult. Under the assumption that adversaries are restricted in some way—for example, in the types of attacks they can perform—can we derive information-theoretical analyses that guarantee the security of certain protocols?

With respect to wide data sets, we quote here from the National Research Council's prepublished 2013 report, Frontiers in Massive Data Analysis, on big data,

Consider a database where the rows correspond to people and the columns correspond to "features" that are used to describe people. If the database contains data on only a thousand people, it may suffice to measure only a few dozen features (e.g., age, gender, years of education, city of residence) to make the kinds of distinctions that may be needed to support assertions of "knowledge." If the database contains data on several billion people, however, we are likely to have heightened expectations for the data, and we will want to measure many more features (e.g., latest magazine read, culinary preferences, genomic markers, travel patterns) to support the wider range of inferences that we wish to make on the basis of the data.

We might roughly imagine the number of features scaling linearly in the number of individuals. Now, the knowledge we wish to obtain from such data is often expressed in terms of combinations of the features. For example, if one lives in Memphis, is a male, enjoys reading about gardening, and often travels to Japan, what is the probability that the person will click on an ad about life insurance? The problem is that there are exponential numbers of such combinations of features and, in any given data set, a vast number of these combinations will appear to be highly predictive of any given outcome by chance alone.

As this scenario suggests, a naive appeal to a "law of large numbers" for massive data is unlikely to be justified. If anything, we should expect the perils associated with statistical fluctuations to increase as data sets grow in size. Of course, if we do not ask new questions as the data grow in size, but are content to more precisely answer old questions, then statistical error rates may not grow as the data scale. But that is not the perspective that underlies the current interest in massive data.

Possible Directions

<Insert intro text here?>

Topological Data Analysis

One promising avenue for discovering structure in data is topological data analysis. Roughly speaking, topological data analysis provides a set of tools that we can apply to infer topological structure in a data set. For example, it has recently been claimed that the graph representing the Internet can be embedded in a hyperbolic metric space. If data lives in a simple topological space, how do we exploit this to extract useful information from the data and develop relevant algorithms?

Inspired by the geometric nature of protein structures, it is reasonable to believe that many biological data sets exhibit well-behaved topological properties. Discovering this structure could pave the way for deeper understanding of the informational structure of biological data and what it reveals. For example, an interesting line of research would be to infer topological properties of neural data and draw conclusions based on structural discoveries.

Compressed Data Analytics

Conventional approaches for data storage and analysis decouple critical steps of compression for storage efficiency and for analysis and querying. While this may work for today's moderate-sized data sets, this approach does not scale, either in terms of computational overhead (for decompression or recompression) or in terms of space requirements. We have a critical need for compression techniques and succinct data structures that directly support data analyses and querying.

Integrating compression and analysis potentially compromises storage efficiency of compression techniques and computational efficiency of analyses. Toward the overall goal, the first challenge is in developing formal methods for quantifying the tradeoffs and deriving suitable metrics for compressed data analysis. These metrics must account for constraints of space and computation, overheads of distributed storage, and considerations of robustness and fault tolerance.

Causality and Generative Data Models

A critical component of many of the aforementioned applications is the need for statistical models for analyses. We must view the results from analyses in the context of our prior knowledge. This prior knowledge (or simply a prior) is typically modeled as a statistical process. These statistical processes are critical to the quality of results, as well as computational feasibility of analyses. Moreover, the ability to distinguish causal relationships from statistical associations and correlations is critical to scientific understanding, decision making, and planning.

We aim to derive suitable statistical and causal models for a variety of data structures and distributions. Where necessary, these models will be dynamic and derived to facilitate computational analyses.

Challenges in Life Sciences

Data in life sciences is represented in diverse data structures. Some of these are highlighted above. We aim to specifically focus on some of these data structures in the near term—due to the pressing need for solutions and the potential for significant applications advances resulting from them.

Sequence Analysis

The analysis of genomic sequences has become particularly important, due in part to the high-throughput methods for acquiring such sequences. We use patient- and phenotype-specific sequences to diagnose and classify diseases and to chart patient-specific therapies.

Analysis Techniques for Sequencers

Next-generation sequencing (NGS) has significantly reduced the cost and overhead of acquiring patient- and tissue-specific sequences. However, a number of challenges remain. In particular, the high error rate of NGS data (1–3%) is still a major consideration. While error-correction techniques have been designed and shown to work, empirically, we still need formally quantified and provably optimal error-correction techniques. Based on our work on the statistics of suffix trees, we will design such techniques, algorithms implementing these techniques, and release-quality software.

Reconstructing Sequences for Emerging Nanopore Sequencers

Emerging sequencing technologies (such as those from Oxford Nano) hold the promise of significantly larger read lengths (tens of thousands of bases). This would significantly simplify assembly and mapping problems; however, their higher error rates pose other significant challenges. We will design novel methods for reducing error rates through a process of staggered reads (the DNA strand translocates through the pore in a staggered fashion). This is a novel experimental approach, and its computational formulation does not exist at this point. We believe that such computational techniques are critical to successful use of nanopore sequencers.

Genome-Wide Associations

The role of genome-wide associations for phenotypes has been well documented. However, approaches to determining such associations are largely empirical—they do not rely on suitable priors or inadequately quantify the significance of the association for reasonable priors. We will develop rigorous methodology for deriving provably significant associations with respect to realistic priors. We will quantify the role of other forms of data—specifically, the use of network information in driving the search for suitable associations.

Repetitive Channel and Darwin Channel

Several information-theoretic tools should help us in modeling information flow in biological systems. For example, we still do not have a good information-theoretic model for Darwin's paradigm of "survival of the fittest." We should point out that the information-theory community has already introduced and analyzed such biologically inspired channels as an insertion–deletion channel to model mutation. The next step is to introduce the Darwin channel to model the effects of preferential Darwinian selection. One may argue that the noisy constrained channel—a well-studied object in information theory—is a special case of the Darwin channel. In general, we aim to model a situation in which information-bearing biomolecules (primarily DNA) are transferred through these channels to effect functional specialization, phenotypic diversity, and natural selection.

Another biologically inspired channel is the so-called repetitive channel, which models the DNA assembly process, using emerging nanopore sequencers, which are discussed above. In this channel, random substrings of the input sequence are repeated. This model resembles the insertion channel; however, the insertions in the repetitive channel are not independent, which makes analysis much harder. The channel capacity in this model informs us about the rate of reliable DNA reconstruction.

Network Data and Integration

<Add general intro text here?>

Network Reconstruction

One of the major developments in life sciences has been the emergence of systems biology. A central task in building system-level models is the construction of flux-based networks from sparse observations. Typically, these observations admit a large number of possible models. We will investigate the role of information theory in driving model selection from among these models that fit sparse data.

Semantic and Syntactic Integration of Data from Different Data Sets

Going forward, major advances in modeling will be at the interfaces of different data structures—e.g., integration of network and structure data in drug discovery and the integration of sequence and network data for genome-wide associations. In such studies, each data structure has its own associated statistical models. The challenge is to integrate these models into a single significant figure for the solution derived from an integration of the models.

Data from Imaging Techniques

We are rethinking interactions and signals from fMRI and want the capability to generalize from single fMRI signals to correlating across multiple fMRI signals or external signals. In seeking to develop an in-vivo system to monitor a three-dimensional volume of an animal's brain, we plan to play molecular and cellular biology tricks so that we can image cell-type specificity (i.e., differentiate neurons at different cortical layers) and see the activity of numerous cells—not only neurons but also glia, the "bouncers" of the brain.

This approach could allow us to see the interaction between different cells in the cortex and also see how the glia and extracellular environment and neurovasculature are all coupled. Monitoring with sub-millisecond time resolution—to see not only action potentials from neurons but also neurotransmitter release—would be a game-changer.

My research group is beginning to explore the use of flexible electronics that can transduce chemical, electrical, mechanical, and optical energy (from our Aug 2011 Science paper), along with genetic engineering tricks, to do some hardcore imaging along these lines. What we are doing now is a *very* simple first step along this path—this is a "long-term vision." From a "science of information" perspective, this would provide an unparalleled glimpse into the dynamics of brain function, from the cellular level to the systems level, and crucially important—how the cells signal to one another electrically, chemically, mechanically, and beyond.

Challenges in Communication Systems

The following topics are actively pursued within this thrust:

- Coordination over networks
- Game-theoretic solution concept for coordination over networks
- Applications in control
- Relearning information theory in nonasymptotic regimes
- Modeling energy, computational power, feedback, sampling, and delay
- Point-to-point communication with delay or finite-rate feedback
- Actions in networks
- Soft decisions and logarithmic loss in source coding

Information theory has traditionally been focused on the problem of communication. Properly viewed, communication is an aspect of the modification of the information state—prior to the communication, the information state of the receiver does not include the message that the transmitter intended to send, while after the communication, it does. Modern applications require a broader view of this process of information-state modification over networks. Developing this is one of the grand challenges in the area of communication systems.

As an example, consider a distributed game played between a network and another player (which might also be a network). To establish strategic equilibrium, the network player needs in general to play randomized strategies against its opponent. This requires the ability to generate the strategically relevant joint probability distributions at the nodes of the network player. What joint probability distributions can be generated over a network for a given set of communication rates? To date, no one knows the answer to this question.

To appreciate the subtlety of this question, consider the following counterintuitive fact that is known in the study of correlation distillation—if two nodes respectively receive the two marginals of independent and identically distributed sequences of correlated pairs of random variables with a correlation coefficient of absolute value less than 1, and are then required to separately generate a bit each without any communication, then there is a universal upper bound, strictly less than 1, to the probability with which they can make the two bits agree. This bound holds, irrespective of how many copies of the random variables they receive and irrespective of what the joint distribution is for the given correlation coefficient.

As another example, consider the problem of controlling a dynamical system over a communication network. The objectives of the communication should be measured in terms of the performance of the control objective. To date, the study of the tradeoff between control performance and communication rates is essentially a wide-open problem. The study of this class of problems is of increasing importance with the growth of remote-controlled applications, such as drone-based surveillance and robotic environmental cleanup.

Recent work by Center members has made significant strides in extending Shannon's theory to a finite blocklength regime. While understanding this regime is of great importance, it does not fully capture the reality of many communication and information-processing applications where the capability to process very large streams of data is needed. Delay rather than blocklength is the system-performance figure of merit. Characterization of the tradeoffs between delay and other performance metrics, such as distortion, bit-error rate, and complexity, is a challenge that Center members have begun to rise up to, and that promises to remain a focal point for the activity of several Center members in the years to come.

A major challenge is to understand complex networks, where each node in the network is not only involved in the transmission of information, but can also take actions that influence the structure or characteristics of the network. Examples of actions include selection of transmission power, frequency, or modulation; energy harvesting; hostile-user sensing; acquisition of side information or feedback; and probing of the channel quality. In the design of existing network architectures, the policy for selection of actions is determined separately and independently of the coding schemes. We thus intend to devote effort to understanding and quantifying the benefits (such as increased throughput and decreased latency) and potential hazards (such as information leakage) of action policies that are designed jointly, in cooperation and coordination with the network coding.

We shall strive to mathematically model and characterize the fundamental performance limits in these scenarios. Such characterizations should yield insight into the benefits, which should be considerable given our recently gained understanding of point-to-point scenarios, in schemes that allow actions to depend on the information to be communicated, compressed, or processed.

In addition to fundamental limits, we shall strive to develop guidelines for the construction of joint coding-action schemes for networks, with particular emphasis on wireless networks, which are typically very dynamic. A central component here that taps into some of the challenges mentioned above is the study of control-theoretic notions, such as action and actuation in complex networks, from an information-theoretic perspective.

Most multiterminal source-coding scenarios have, to date, defied our complete understanding of the achievable tradeoffs between rates and distortions. Perhaps the simplest case in point is lossy compression with a rate-constrained description of side information at the decoder (the "one-helper" problem). The problem remains open, despite being but a slight, seemingly innocuous extension of the unconstrained case, solved by Wyner and Ziv in a celebrated paper more than 35 years ago. This problem is a special case of what is widely considered the holy grail of multiterminal source coding—lossy reconstruction of two correlated sources based on limited-rate descriptions of each of the sources separately—a problem referred to generically as "multiterminal source coding."

Center researchers recently solved this problem for the case where distortion is measured by the logarithmic loss, i.e., rather than a particular reconstruction symbol, the decoder is required to give a "soft" reconstruction, which is a probability (belief) on the value of the original source symbol. The

distortion is the logarithm of the reciprocal of the probability assigned to that symbol. This formulation put a timely spotlight on the canonical idea of a soft reconstruction in the context of lossy compression.

The tools and the point of view we developed should be applicable to a much wider family of multiterminal source, as well as joint source-channel coding problems under logarithmic loss. It is a challenge of considerable potential dividends to theory and applications to progress in this direction. Indeed, from the theory standpoint, this new perspective on multiterminal source coding via logarithmic loss is likely to turn out key to understanding and ultimately solving the multiterminal source-coding problem under a general loss function.

From the applications viewpoint, an increasing number of scenarios involving online recommendation systems can naturally be broken into two parts: (i) a data-mining part that extracts relevant customer information from databases (which are typically distributed), and (ii) a recommendation part that selects the best recommendations based on the extracted customer information. These scenarios naturally fall under the umbrella of multiterminal source coding—with soft reconstruction under logarithmic loss.

Economic Systems

- Impact of information flow on the dynamics of economic systems
- Impact of finite-rate feedback and delay on system behavior
- Impact of agents with widely differing capabilities (information and computation) on overall system state and on the state of individual agents
- Choosing portfolios in the presence of information constraints

Economic systems share many common features with complex communication networks—they consist of multiple entities (agents) with vastly heterogeneous capabilities for acquiring, storing, sharing, and processing information and with differing degrees of authority for acting upon that information. Yet, for all the commonalities they share with communication networks, one crucial feature that distinguishes economic networks is that agents have objectives that extend beyond simply reliable communication. In other words, in an economic system, information has value. One major challenge in economics is to formalize the notion of information value, particularly in dynamic settings involving multiple agents with different capabilities.

To date, there is no universally agreed-upon definition of information value, although several reasonable alternatives have been proposed since at least the late 1960s. However, if we abstract away the details of these various proposals, we can distill one basic underlying idea: the value of information has to do with the change of the information state of one or many agents, where, broadly speaking, the information state encapsulates all payoff-relevant knowledge available to the agent(s). For example, according to one definition proposed by Stratonovich (year?) in the Soviet Union and independently by Howard (year?) in the United States, the value of one bit of information acquired about a random variable of interest is the largest difference between expected utilities achievable with and without that additional one bit of information. Thus, we can speak of the “best” observation channel that can deliver

one bit of information as the one that would provide the largest increase in expected utility relative to what one could achieve without any observation, based solely on prior knowledge.

Of course, now we have to specify (i) the utility function and (ii) the nature of the information constraint. The choice of the utility function is conceptually similar to the choice of the distortion function in communication problems, and so we can assume that the utility function is a fixed exogenous quantity.

On the other hand, the specification of the information constraint is a separate issue. We can require the observation channel to be implementable by a deterministic quantizer that partitions the state space into two disjoint regions, but then the problem of optimizing the choice of such a channel is fundamentally combinatorial. Alternatively, we may measure information in the sense of Shannon as the mutual information between the input and the output of the channel. If we adopt this mutual-information criterion, we end up with the rational inattention framework proposed by Christopher Sims (year?). In this case, the problem of optimizing the observation channel is exactly the Shannon rate-distortion problem, and the value of information is given by the Shannon distortion-rate function (DRF) with the negative utility playing the role of the distortion function.

Another appealing feature of the rate-distortion viewpoint is as follows—If we are allowed to optimize not only the observation channel but also the observation space, then we can show that the belief state induced by the channel that solves the rate-distortion problem is the best representation. This meshes well with the fundamental fact that the Bayesian posterior is the minimal sufficient statistic.

Of course, the Shannon DRF is only an asymptotic measure of performance, and it needs to be related to an operational criterion via an appropriate coding theorem. In a communication system, the relevant operational criteria pertain to the quality of signal reconstruction at the receiver, and the mutual information constraint is only an asymptotic abstraction of the channel's reliability by way of the law of large numbers. However, in economics, it is not at all clear which operational interpretation one should attach to mutual information constraints faced by rationally inattentive agents.

Finding an appropriate operational interpretation is a central challenge one faces when applying information theory to economics. Useful applications include such questions as investment decisions (portfolio selection) in the presence of information constraints. Another interesting and challenging problem related to information capacity constraints is a justification, from first principles, of the emergence of "coarse thinking" or quantization phenomena in situations involving strategically behaving agents. One vivid illustration is in the well-known work by Crawford and Sobel (year?) on strategic transmission of information—if one agent must communicate some signal to another agent who will in turn use that signal to select an action, if the action chosen by the second agent affects both agents' utilities, and if the agents' utility functions are different, then the optimal (equilibrium) strategy of the first agent is implemented by a quantizer. Moreover, the number of quantization bins increases as the disagreement between the two agents' utility functions vanishes. Thus, strategic interaction may

introduce hard information constraints, and this deserves further study within the rational inattention framework.

However, so far we have addressed only the simplest arrangement—static optimization by a single agent. What happens when we introduce dynamics into the picture? Indeed, any sufficiently complex economic system evolves in time. Agents take actions based on information they obtain, and these actions may, in turn, affect these and other agents' information states. They may even affect the quality of any information that may become available in the future. Thus, the choice of an observation channel can be viewed as a control action that alters the information state.

In general, the control law has a dual effect—it affects both the utility (or cost) at the current stage and the uncertainty about the information state at future stages. The presence of a mutual information constraint enhances this dual effect, because it prevents the agent from ever learning too much about the state. This, in turn, limits the agent's future ability to optimize expected utility or cost.

These issues need to be thoroughly understood both on a theoretical and on a practical level. We have done some preliminary work on designing stationary finite-rate feedback-control laws that operate in the presence of information constraints. The optimal control law turns out to be a solution to a rate-distortion problem, but the distortion function must incorporate a correction term that would account for the agent's forecast of the future information state. These results apply to steady-state (ergodic) dynamics in a single-agent setting. We will build upon them to investigate what happens in networked settings. Moreover, we will explore the effect of transients on the information flow and on the dynamics of an economic system.

Another type of information constraint that arises in economic systems has to do not with the amount of information, but with its timeliness. The above discussion focused on instantaneous information flow, and the only limitation was the amount of information available to the agent(s) per time step. However, we must also understand the effect of delayed feedback on the dynamics of the information state. In many cases, it is possible to quantify the impact of delay in terms of information-theoretic quantities, such as mutual information or directed information (the latter is more appropriate in situations involving feedback and causality). For example, we may consider the value of information as a function of the delay and optimize the observation channel subject to capacity and delay constraints. We hypothesize that, in many settings, there may be a diminishing returns effect—the net increase in information value is a monotonically decreasing function of the delay (when the delay is already small, decreasing it further may not yield appreciable information gain). These and related issues will be investigated in an information-theoretic framework.

Quantum Information Theory

We describe two grand challenges in quantum information theory. The first is the question of how best to do fault tolerance on a quantum computer. The second is to find the capacity of quantum Gaussian channels.

Quantum Fault-Tolerant Computation

When quantum computation was first proposed, one objection to it was that it could not be made fault-tolerant. The standard techniques for making classical computation fault-tolerant involve either redundancy (duplicating information) or error-correcting codes. It was believed that these techniques would not work for quantum computation because of the quantum no-cloning theorem (Wootters & Zurek, 1982). It appears at first that classical error-correcting codes require redundancy, so that quantum error-correcting codes would violate the no-cloning theorem. However, it was discovered that quantum error-correcting codes do exist (Shor, 1995), (Steane, 1996), (Calderbank & Shor, 1996). Furthermore, these error-correcting codes could be used to design fault-tolerant quantum computers (Preskill, 1998). The theoretical results on fault tolerance are usually given in threshold theorems; the exact details of these threshold theorems depend on the architecture chosen for the quantum computer. The threshold theorem says that if quantum gates have accuracy above some threshold, then any circuit for a quantum algorithm can be transformed into a quantum fault-tolerant circuit using a polylogarithmic factor OF? WITH? more gates.

While the theory of threshold theorems is very nice, we are still quite far from achieving a practical fault-tolerant quantum computer. The accuracy required for the first threshold theorems was error less than 10-6. More recent results have improved this to the order of 10-2, but with an enormous overhead and using current techniques, fault tolerance with reasonable overhead is believed to require errors less than 10-4 or 10-5. These accuracies are still reasonably far from what experimenters can achieve. In experimental quantum computation, there currently is, and likely will continue to be, a tradeoff between the accuracy achievable and the number of qubits that can be built. To make experimental quantum computers useful as soon as possible, what we would like to find is fault-tolerance schemes that work with low overhead requiring relatively low accuracy.

We know several completely different methods for achieving fault tolerance, and it appears that more methods could be discovered. The methods currently known all have fairly high accuracy-overhead tradeoffs. The grand challenge is to discover a new method, or improve on one or more existing methods, to yield fault-tolerance schemes that we can operate at reasonable accuracy and with reasonable overhead.

Quantum Capacity of Gaussian Channels

For classical channels, Shannon's theorem gives a formula for computing the capacity. This capacity is essentially the only capacity of a classical channel, and appears in many formulae describing the properties of a channel. Quantum channels have many capacities (Holevo & Giovannetti, 2012); in particular, there is one for the transmission of quantum information over the channel, and one for the transmission of classical information over the channel. What is more frustrating is that while we have some formulas for the capacity of quantum channels, for very many simple channels, we are unable to calculate the capacity.

Possibly the most embarrassing case of this is the classical capacity of a quantum Gaussian channel. Quantum Gaussian channels are essentially those that can be implemented with linear quantum optics elements. Finding the classical capacity of an arbitrary quantum Gaussian channel could be solved if we could show that the minimum entropy output of a symmetric single-mode Gaussian channel with thermal noise was achieved on the input of a coherent state (Guha, Shapiro, & Erkmen, 2008). This very simply described channel is a good approximation to many real-life channels. It can be implemented by putting the input into one port of a beam splitter and a thermal state of light into the other. While this seems to be a relatively straightforward problem, and there does not seem to be any better way of achieving a low-entropy output than a coherent state, it has eluded proof so far. It would follow from the proof of one of several entropy inequalities (Koenig & Smith, 2012; Guha, Shapiro, & Erkmen, 2008).