Big Data – Big Deal

Nick Cercone, F'IEEE

[1] Department of Electrical Engineering and Computer Science
Lassonde School of Engineering
York University
Toronto, Ontario, Canada M3J 1P3
nick.cercone@lassonde.yorku.ca

With significant contributions from Igor Jurisica (University of Toronto), Ming Li (Waterloo), Sara Diamond (OCAD University), Fred Popowich (Simon Fraser University), Marin Litoiu (York University), Jimmy Huang (York) and Aijun An (York University)

Abstract. This paper position paper is based on a major cooperative research and development proposal to Canada's Natural Science and Engineering Research Council for a Big Data Research, Analytics, and Information Network (BRAIN). Challenges presented by Big Data research are introduced and several projects are sketched in four theme areas of important Big Data research, the solutions of which will further decision making in these areas of investigation. The four themes are large-scale data analytics and cloud computing, computational biology, health informatics, and interactive content analytics. The importance of training highly qualified personnel, knowledge mobilization and novelty are discussed.

Keywords: big data, large-scale data analytics, computational biology, health informatics, interactive content analytics, visualization.

1 Introduction

The old saw goes something like this: "Be careful what you ask for, you might get it." Never has this old adage been more meaningful than now. Big Data is a big deal!

"Big Data" present challenges to circumvent and offer incredible opportunities. Analysis tools must evolve to meet both. People with data analysis skills are in demand and demand is growing. By 2018 there will be a "talent gap" of between 140,000-190,000 people, says the McKinsey Global Institute (in the U.S.). Lacking intelligent data analysis, planning and successful partnerships, and big data creates huge problems for individuals and companies as well as regulatory challenges for government.

Digital data generation is exploding, with a staggering 1.8 zettabytes (1 zetta=1 trillion gigabytes) in 2011, up from 1.2 zettabytes in 2010 (the *Economist*, Nov. 17, 2011). 90% of the world's data today has been created in the last two years alone. Why? Consider biology: huge data are attainable since new automation is available: robotics, new chips, sequencing, imaging, etc. Sensors measure everything from speed to smell. Smart phone "apps" generate vast data

quantities. Social media adds to the deluge. Software to lever “big data” is improving but not rapidly enough.

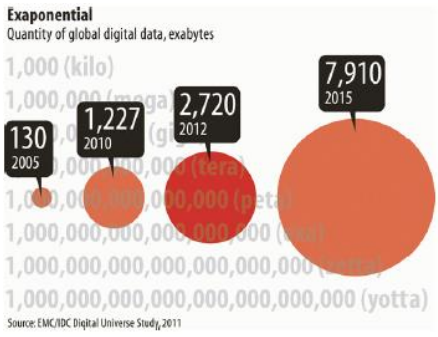
Big Data is big business. Market research firm IDC forecasts that the market for Big Data is expected to grow from $3.2 billion in 2010 to $16.9 billion in 2015. To illustrate the need for additional research, consider current unbounded, high speed and continuous characteristics of streaming (machine generated) data that render ineffective traditional data mining analytics.

The data analytics vision includes the scientific, engineering, and design aspects of medical, ethical, legal, social and cultural implication (MELSCI) issues surrounding information based research and development (R&D) and their usefulness when considering real world information analysis challenges. The need of private and public sector organizations for solutions in a wide range of topics involving big data is evident by the large number of such organizations spending research funds to tackle big dat opportunities.

The research proposed is organized into four nonexclusive strategic themes exploiting the notion of "from data to knowledge to action": 1) Large Scale Data Analytics and Cloud Computing, 2) Computational Biology, 3) Interactive Content Analytics, and 4) Health Informatics. Choice of themes reflects the interests of industry and government and the strong capabilities and interests of researchers. Several examples can be cited, using the Canadian context, IBM Canada’s recent announcement of new $42 million IBM Compute Cloud Centre in Ontario will permit researchers to exploit all four themes especially large-scale data analytics. Another example is that Genome Canada, in partnership with the Canadian Institutes of Health Research, is seeking proposals for research to address any aspect of bioinformatics and computational biology; advancing these fields will be a key to enabling the development of novel translational research applications in health related areas. The massive and ongoing influx of data from large-scale sequencing projects underscores the need for new computational and theoretical tools in modern biology. The lack of efficient business-to-business and business to consumer tools and methodologies available to analyze these data sets is a major bottleneck faced by the genomics research community.

1.1 Why the proposed research is strategic

The core of this proposed research is developing tools and methodologies in order to address the challenges posed by Big Data. Solutions to these challenges should be developed in conjunction with public and private sector partners to bring economic benefits. The application of these tools and methodologies will permit the processing of the exponentially growing volume of data. An example is the projects designed to take advantage of the massively increasing pool of “omics” data and using those data to optimize clinical decision-making.



Our knowledge-based economy increasingly demands the mastery of Big Data in order to function effectively. Over 2.7 zettabytes of data exist today, up from 1.2 zettabytes in 2010.[1] Over 571 new websites are being created per minute. IDC estimates that by 2020, business transactions on the Internet will reach 450 billion/day. This volume of activity and the increasingly huge data such activity creates demand new algorithms, systems and methods to tackle the challenges these data produce.

[1] *The Economist*, “Welcome to the yotta world: Big Data will flood the planet,” Nov 17th 2011 | from The World In 2012 print edition.

Big data is big business. Market research firm IDC forecasts the market for big data is expected to grow from $3.2 billion in 2010 to $16.9 billion in 2015 in its report, *Worldwide Big Data Technology and Services 2012-2015*. A report developed by leading researchers in the United States, states, "The promise of data driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of 'Big Data.'" While the promise of big data is real – for example, it is estimated Google alone contributed over 4 billion dollars to the U.S. economy in 2009 – there is currently a wide gap between its potential and its realization."[2]

Substantial opportunities exist for Canadian industry for commercialization: for example, to be more responsive to user behavior in the services sector and to manage complex supply chains. Nonetheless, Big Data presents challenges (volume, structure and data complexity) to circumvent, yet offers incredible opportunities. Analysis tools must evolve to meet both. People with data analysis skills are in demand, and that demand is growing. Privacy and security, and evolving understandings of intellectual property, remain problems. Without intelligent data analysis, planning, successful partnerships, and cooperative networks of resources, big data creates problems for individuals and companies and for government regulatory challenges.

Information analysis advances as society's needs become more complex. New applications for "big data" are continually evolving. To respond to these advances, we constantly evaluate our research methods and advance new research paradigms. Valuable analysis of big data requires new techniques (e.g., analyzing machine generated streaming data). Interpretation of *data* leads to *information* which contextualized becomes *knowledge* that can be applied as *wisdom* for decision-making, the "DIKW" paradigm. Figure 1 illustrates the increments in the DIKW paradigm leading to decision making.

Information analysis is multidisciplinary and serves as a starting point for other specializations, such as Business Intelligence, Health Informatics, Computational Biology, Interactive Content Analytics, Cloud Computing, etc.

Big data with appropriate *information analysis* and *knowledge mobilization* presents an opportunity to find insights in new and emerging types of data and content, to make Canadian industry more agile and responsive, and to answer questions that were previously considered beyond our reach. Until now, there was no practical way to harvest this opportunity. Today, advanced analytics open the door to a world of possibilities.

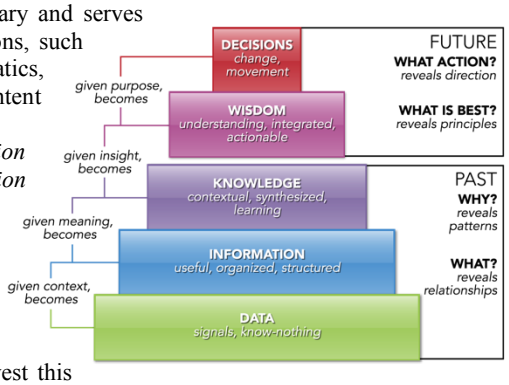


Figure 1 The DIKW Paradigm.

In a January 21, 2013 press release "Building the Knowledge Economy through Partnerships," Canada's Hon. Gary Goodyear, Minister of State (Science and Technology) said: "Collaboration between business and academia is essential to leveraging Canada's research strengths and seizing market-driven opportunities." NSERC also recently announced two programs in its Discovery Frontiers Call for Proposals: Exploring Big Data and Advancing Big Data Science in Genomics Research. The latter fits nicely with our Computational Biology Theme. An internationally funded competition Digging into Data, to address how "big data" changes the research landscape for the humanities and social sciences, fits well with Interactive Content Analytics.

2 http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf

In May 2012 Doug Cutting[3] stated

"The combination of the affordability of commodity hardware and software that allows effective processing of big data is propelling the trend, but there is a skills issue. A lot of smart folk are not being leveraged to the degree they could be because of the way things are structured and how information is siloed. Big data technology removes silos by using fewer but bigger clusters which are more economical and create data sets for better analysis. This allows people to experiment and explore data with their ideas and test them, but we are in a new cycle and it takes time for people to come up to speed. There's a skills deficit; demand is great."

1.2 Vision of the proposed research

Ninety percent of the data in the world today has been created in the last two years alone. These data come from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals, to name a few. We need to distinguish useful data from "garbage" in order for it to become information, then knowledge and subsequently wisdom.

We have the opportunity to find insights in new and emerging types of data and content by utilizing advanced research, knowledge transfer, and creative training program for highly qualified persons (HQP), and private and public sector partnerships.

Our vision includes scientific, engineering, and design aspects of information analysis and knowledge mobilization. This vision includes medical, ethical, legal, social and cultural implications issues surrounding information-based research and development and their usefulness when considering real world information analysis challenges. We must offer an integrative platform for private- and public-sector partners and to collaborate and partner with academic experts in four areas of large-scale data analysis.

1) *Large-scale Data Analytics and Cloud Computing*: This is the overarching theme for the research. There is growing interest in *large-scale analytics* using cloud computing. The challenges for cloud data analytics (merging structured and unstructured data, big data with clinical; ensuring privacy and confidentiality) still require research. Responding to the growth of raw data, data analytics must marshal significant computing power and robust storage to conduct analytics using on-demand, scalable infrastructure. This infrastructure is analytics-aware and is driven by the requirements of large-scale data analytics, optionally including analytics support platform.
2) *Computational Biology* (Bioinformatics, Molecular Medicine): With an emphasis on cancer treatment, novel tools development and application integrate, analyze, and interpret complex biomedical data. This will help identify testable hypotheses and construction of useful models. Existing algorithms either do not scale with the size and complexity of realistic problems or provide only partial answers. New applications and visual analytics are required to provide effective tools for scientists, clinicians, and patients.
3) *Health Informatics*: Free-form text is a common form of valuable health-care data, ranging from electronic record information, images, doctors' notes, patient histories, to health-care messages patients post on social media. Such narrative text data contain information for physicians to use in their practice and inform public/government agencies' healthcare-related decisions. Data are continuously generated daily in large volumes, too vast to read manually and analyze. Automatic text analysis tools are needed to discover hidden information trapped inside free-form texts. Tools that identify and analyze health-care text posts detect public opinions and activities/preferences in health-care issues. Language and image analysis, data mining and information retrieval are key techniques used to build such tools.

[3] Doug Cutting is a big name in big data. He helped create the open-source Hadoop framework that allows applications based on MapReduce to be run on large clusters of commodity hardware.

4) *Interactive Content Analytics* (ICA): ICA combines research approaches from digital humanities, visual analytics, Human Computer Interaction (HCI), design thinking and practice, and computational linguistics to provide new insights into the creation, exploration and analysis of online data. Our perspective on interactive content analytics embraces topics ranging from the management of entertainment and advertising brand data in digital media, to tracking digital rights through data mining visualization, to data mining large cultural datasets that integrate both digitized (e.g., non-electronic data sources such as print documents subsequently scanned) and digital materials, to creating novel dashboards to analyze and compare online learners, to building tools to manage complex supply chains. Comparative and other methodologies from the traditional humanities disciplines, the arts and social sciences are combined with computing tools such as data visualization, information retrieval, data mining, statistics, and computational linguistics analysis and digital publishing to create interpretive applications.

Beyond these four theme areas, related applications include the management of entertainment and advertising brand data in digital media, building interfaces for data management using inclusive design principles, tracking digital rights through visualization, producing and managing open data, security (e.g., response to government queries; facial recognition, scanning e-mail traffic; or defence queries in interpreting satellite data), text/data mining of text-based data for Knowledge Mobilization purposes, institutional memory crystallization, and self-managing computing.

1.3 Objectives

Big Data provides opportunities for business users to ask questions they were not able to ask before. How can a financial organization find better ways to detect fraud? How can an insurance company gain a deeper insight into its customers to see who may be the least economical to insure? How does a software company find its most at-risk customers—those who are about to deploy a competitive product? They need to integrate Big Data techniques with their current enterprise data to gain a competitive advantage.

Data analytics are not always effectively used. Misuse may harm important decision-making processes and business intelligence of Canadian industry. Big Data analysis should become one of the fastest growing elements of Canadian business intelligence spending.[4] Data analytics play an important role in Canadian industry information technology and planning systems.

A primary long-term goal of this proposal is to develop Big Data analytics software solutions (e.g., decision support systems for evidence-based management, quality control and best practices) in the areas important to our Canadian industry partners, broadly conceived. Results of our research will improve accessibility, management and manipulation of information in a cloud environment by:

1. Building, verifying, validating and evaluating application prototypes for private and public sector partners which incorporate best research and development practices and drive new research into data analytics to serve as *a solutions base for specific Canadian industry partner problems;* and
2. Employing and developing the newest and best research methods to provide a base for *solutions for projects across and within our four theme areas.*

[4] Big Data will drive $232 billion in spending through 2016. It will directly or indirectly drive $96 billion of worldwide IT spending in 2012, and is forecast to drive $120 billion of IT spending in 2013. Big Data's influx will force a change in products, practices and solutions. The change is so rapid that companies may have to retire early existing solutions that are not up to par. In 2012, "IT spending driven by big data functional demands totaled $28 billion." Most of that went toward adapting existing solutions to new demands driven by machine data, social data and the unpredictable velocity that comes with it. Gartner Research, http://techcrunch.com/2012/10/17/big-data-to-drive-232-billion-in-it-spending-through-2016/

1.4 Anticipated Value of the research and benefits expected

We produce research results broadly in information analysis, more particularly we develop innovative multidisciplinary applied research centering on all aspects of *Information Access, Extraction, Analysis and Discovery*, an area of critical need. Commercialization of research results is attained through collaborative projects with private and public sector partners and significant innovative HQP training.

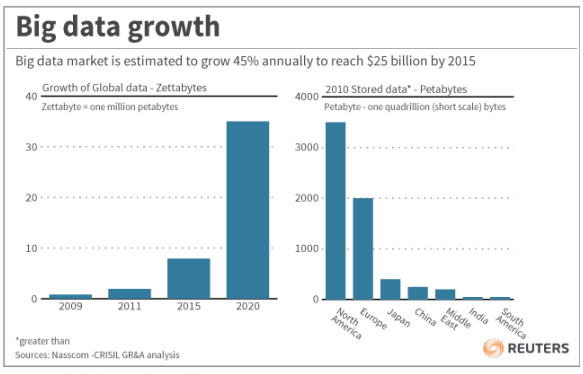
The scientific or technical advances research engenders include technical disclosure; new intellectual property; patents of innovative techniques (such as methods for early analysis of machine-generated (streaming) data for which effective techniques do not presently exist); training HQP to cultivate much needed Big Data expertise for partners; and novel processes and products that are developed and commercialized through our partners or by spin-off companies.

We expect many benefits to accrue, most importantly (1) HQP training, (2) knowledge mobilization and technology transfer, and (3) innovative research and development solutions to information analysis of big data. All benefits should have direct economic benefit to industry. The HQP trained in data analytics will help to address the skills gap in this area. Through project research and development, careful intern deployment, and academic/industry exchanges, technology transfer will be facilitated. New algorithms and methods that provide solutions to opportunities presented by the challenges of Big Data should spawn patents and spinoff companies for economic benefit.

HQP are trained in research on the state-of-the-art techniques in information analysis, and in knowledge mobilization via internships and "soft skills" training. Training focusses on applications of techniques in the theme areas. The trainees will become the next generation of researchers to bridge the gap between information analysis and deploying the analysis to applications.

1.4 Information analysis and big data

Digital data generation is exploding! Why? Huge amounts of data derive from new automation in robotics, new chips, sequencing, imaging, etc. Although the storage cost is dropping rapidly,[5] storage capacity is not meeting demand. Software to handle "big data" is improving.[6] Sensors measure everything from speed to smell. Smart phone "apps" generate vast data quantities. Social media adds to the deluge: Twitter messages will exceed 500M daily by 2013. And governments are opening their vast data vaults for analysis.



Big data presents challenges to meet and offers opportunities. Analysis tools must evolve to meet both. Privacy, security and intellectual property remain problems. Without intelligent data analysis, successful partnerships and planning, big data creates complex problems for individuals, companies and government

Big data, with appropriate

5 A petabyte will cost a mere $4 by 2020, predicts Forrester, a market-research firm.

6 Hadoop "database" can sift through big data streams in real time, including "unstructured" data (any kind of text).

information analysis and knowledge mobilization, provides an opportunity to find insights in new and emerging types of data and content, to make sectors more agile, and to answer questions that were previously considered beyond reach. Until now, there was no practical way to harvest this opportunity. Today, advanced analytics open the door to a world of possibilities.

Traditional database management tools and data processing applications face challenges. Capture, curation, storage, search, sharing, analysis, and visualization are problematic when applied to extremely large and complex datasets [1]. In part, the trend to larger datasets is due to the additional information derived from analysis of a related single large dataset. Separate smaller datasets with the same total amount of data allow correlations to be found to spot business trends, determine research quality, prevent diseases, combat crime, determine real-time traffic conditions, and so on [2-3].

In 2012, the size of datasets size that could be processed in reasonable time was on the order of exabytes of data [4]. Scientists regularly face large data set limits in many areas, including meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research [5-6]. These complexities also affect Internet search, finance and business informatics. Thus, the high capacity requirements of big data render it difficult to analyze when one uses relational databases and desktop statistics and visualization packages. Hence there is a need for the *next generation* of information analysis research and development.

One data-mining exemplar illustrates the need for additional research in the wake of big data analytics. Currently unbounded, high-speed and continuous characteristics of streaming (machine-generated) data render ineffective traditional data mining and analytics methods. For example, frequent itemset mining from data streams is an important data-mining problem. Inherent challenges for data stream mining exist [7, 177, 179, 180], including: 1) each data element can be examined at most once; 2) although data elements are continuously generated, memory space consumption should be limited; 3) every incoming data element should be processed as fast as possible; and 4) the analytical result of a data stream of an acceptable quality should be available when users request results.

Traditional frequent pattern-mining algorithms cannot be directly applied due to data stream characteristics. In general, the mining-result set includes a large number of frequent itemsets. Therefore, closed or maximally frequent itemsets are often used to represent them in a more compact notation. Finding such itemsets over (online transactional) data streams is difficult because of the requirements of a data stream [8].

While it depends on the stream processing model [9], the research of frequent patterns mining over data streams can be divided into three categories: 1) the landmark window model, 2) the damped window model, and 3) the sliding window model. In the landmark window model, the range of mining includes all the data between a specific timestamp, called the landmark, and the current time. In the damped window model (also referred to as the time fading window approach), each transaction is associated with a weight depending on the order of its appearance. In other words, the new transactions of data receive higher weights than the older ones. In the sliding-window model, the range of mining is the length of the most recent transactions within a window. The basic processing unit of window sliding is either a time unit or an expired transaction. Processing the recent data is usually important for the applications that handle stream-oriented data. Therefore, the sliding window model is widely used to find recent frequent patterns in data streams.

Because of space and time constraints, we cannot afford to keep all itemsets or even frequent itemsets in the streaming environment that would otherwise allow the discovery of frequent itemsets on a data stream. On the other hand, any deletion may prevent us from discovering future frequent itemsets [10]. Therefore, the challenge is to organize a compact data structure that does not miss any frequent itemset information over a sliding window and is built with only one scan over the stream.

Given these requirements [11-12], an efficient algorithm was developed for mining recent maximal frequent itemsets from a high-speed stream of transactions within a sliding window.

Whenever a new transaction is inserted in the current window, only its maximum itemset should be inserted into a prefix tree-based summary data structure for maintaining the number of independent appearances of each transaction in the current window. Finally, we obtain set of recent maximal frequent itemsets.

Comprehensive experimental results for both real and synthetic datasets show the significance of their methods for streaming or machine generated data [11-12]. Broad applications such as retail market data analysis, network monitoring, web usage, traffic signals analysis, web click-stream mining, ATM transactions analysis, and sensor network data analysis, etc. can be investigated.

We integrate information analysis and knowledge mobilization across the four broad themes. The overarching Theme is Large-Scale Data Analytics (and Cloud Computing). Research results from this theme should unceasingly prove useful to the research of all themes by serving as a framework for information analysis of big data across computational biology, health informatics, and interactive content analytics. All the Themes are progressively affected by the availability of big data and the uninterrupted streaming tide of big data at rates not contemplated a scant few years ago. Thus methods developed under Large-Scale Data Analytics (and Cloud Computing) take this into account and will inform research across themes.

2 Large-scale data analytics and cloud computing

Cloud computing is a new computation model in which hardware and software are offered on-demand as services over the web. Enabled by new Internet technologies and standards and driven by new business models, cloud computing targets three main areas of computing: the infrastructure, (Infrastructure as a Service- IaaS), the programming and runtime environments (Platform as a Service- PaaS) and the end user software (Software as a Service-SaaS). While the categorization as IaaS, PaaS and SaaS is useful in identifying and partitioning many distinct cloud research challenges, an orthogonal set of issues arises when looking at the requirements of different classes of applications that we aim to run in the cloud. For example, a cloud that runs real-time and safety-critical applications may have different services and may offer different qualities-of-service than a cloud for e-commerce applications.

The Large Scale Data Analytics (LSDA) theme focuses on addressing some of the challenges in creating a cloud for data analytics. Analytics, in general terms, refers to extracting information from raw data, often in the context of business decision-making. In response to the growth of this raw data, cloud analytics marshals significant amounts of computing power and robust storage to conduct analytics using on-demand, scalable infrastructure. Gartner describes six key elements that may exist in a cloud analytics solution: data sources, data models, processing applications, computing power, analytic models, and sharing or storing of results [14]. This theme focuses on *platform as a service for analytics:* an infrastructure that is analytics-aware and is driven by the requirements of large-scale data analytics. The platform includes: reusable services, patterns / algorithms that can be assembled for analytic applications development and runtime services for application scalability in clouds.

2.1 Problem

There is growing academic interest in large-scale analytics using cloud computing. Position papers have suggested a need for substantial work in this area (e.g., [14-16]). Individual efforts have made piecemeal advances, for example by migrating analytic applications to the cloud for time series data [17] and by devising data warehouses for the cloud [18]. There is interest in

using the MapReduce paradigm for analytics (e.g., [19, 20]); for example, IBM Research reported efforts – primarily focused on extracting and analyzing large-scale unstructured data – to do search-driven analytics using the Hadoop platform [21]. To our knowledge, there is little or no reported research at the infrastructure or platform level; the focus is on applications that run on stock infrastructure and platforms. Recent efforts have investigated scheduling analytics tasks among heterogeneous resources without addressing the platform needed to achieve interoperability [22] and sharing segments of Big Data with analytics users using ordinary Hadoop jobs [23].

Commercial vendors have focused on providing analytics-as-a-service (for example, SAP and Opera Solutions[7]). Google's BigQuery[8] service queries terabytes of data, runs on Google's cloud infra-structure, scales as needed, and is accessed through RESTful APIs or a web interface. HP through Vertica[9] offers a data warehouse run on Amazon EC2. Microsoft SQL Azure Reporting[10] is a business intelligence system that runs only on Windows Azure. All solutions require structured data.

2.2 Objectives

1. To design, implement, and evaluate a cloud development platform for large scale data analytics based on software development methodologies and design patterns, including runtime algorithms and services that can be discovered and reused by a range of analytic applications;
2. To design, implement, and evaluate algorithms for performing common data analytic tasks, including data mining, that are well adapted to the Large Scale Data Analytics framework;
3. To design, implement, and evaluate optimization algorithms for computing and storage resource scaling; and
4. To design, implement, and evaluate secure storage and data services for analytic applications.

2.2 Background and methodology

2.3.1 Hybrid Clouds for Large Scale Data Analytics

Hybrid clouds are made of private and public sub-clouds working together to mitigate privacy and security concerns while addressing the need for large computation and storage capacity. Academic research into hybrid clouds has focused on the middleware / abstraction layers for creating, managing, and using hybrid clouds (e.g., [24-26]). For example, Zhang et al. used the MapReduce paradigm to split a data-intensive workload into mapping tasks sorted by the sensitivity of the data, with the most sensitive data being processed locally and the least sensitive processed in a public cloud [27]. Abraham et al. provision private clouds from multiple collaborating entities as well as public clouds from Amazon, automatically and semi-transparently [28].

Commercial support for hybrid clouds is growing in response to the business case for cloud federation. For example, HP offers software to manage a private cloud, but also provides infrastructure as a service and the ability to have the two communicate (interoperability with

---

7 Respectively, SAP BusinessObjects BI OnDemand, http://www.ondemand.com/businessintelligence; Opera Solutions LLC, http://operasolutions.com/.

8 http://code.google.com/apis/bigquery/

9 http://www.vertica.com/

10 http://www.microsoft.com/windowsazure/features/reporting/

other vendors is not discussed) [29]. Fujitsu and Microsoft partnered to produce a hybrid of the Windows Azure public cloud with Windows Server running in a private cloud or on a private server [30]. IBM offers both SmartCloud Provisioning (to manage a private cloud) [31] and SmartCloud Enterprise (a public cloud offering) and software to bridge the two at the Software-as-a-Service (SaaS) level [32]. These examples do not enable the federation of heterogeneous clouds; that is, the private and public cloud must run specific software to enable linking.

In the open source realm, support for cloud federation and hybrid clouds is emerging. Apache Deltacloud is an abstraction layer that allows developers to work with various public IaaS providers (e.g., Amazon EC2, Rackspace) and internal private clouds (supporting various commercial and open-source solutions) using a unified RESTful API. This open-source project recently transitioned from incubation to a top-level Apache project [33]. Apache Libcloud [34] and jClouds [35] are Python and Java libraries (respectively) with similar abstraction goals. All three offer stable support for working with compute instances and cloud storage, but also have beta support for other abstractions.

Objective 1 Design*, implement, and evaluate architecture for analytics on combinations of private and public clouds that is elastic, scalable, and tailored for the analytics application class.*

We propose an architecture that optimizes the execution of analytics jobs on a combination of private and public clouds. Performing an analytics task across heterogeneous, hybrid clouds can combine the scalable, on-demand computation resources of the public cloud with the privacy and control provided by a private cloud. Supporting these tasks at the infrastructure and platform level allows out-of-the-box analytics operations to leverage these capabilities. In the typical case, the large data stores will be local to the enterprise; the research challenge is enabling the use of public cloud infrastructure while supporting data locality, privacy, and bandwidth limitations. An analytics cloud may include highly optimized data warehouse hardware in addition to the commodity machines commonly used. The architecture of the hybrid cloud for analytics must be scalable and elastic, and must systematize support for non-functional properties such as privacy, data security, and quality of service. Contributions could include data partitioning to minimize data transfer among clouds, computation task modeling to identify candidate modules for migration to a public cloud, and migrating tasks among heterogeneous cloud infrastructures at run-time.

Objective 2 To design, evaluate and implement an automated solution (i.e., a broker) capable of matching the requirements of an analytics application with the available resources in a heterogeneous, hybrid cloud.

With the growing numbers of available providers, and the mix of public and private resources, it is not always clear which resources would best meet the needs of a given task. This is the Resource Acquisition Decision (RAD) problem [36], which in general must be addressed in a federation of heterogeneous resources [37, 38]. Existing approaches make static decisions at design-time or deployment-time [39, 40]; to enable efficient use of resources, run-time decisions are preferable.

We propose the Resource Broker as an additional layer in the cloud architecture that serves as an intermediary between an application or a task and a pool of resources. Typically an application developer must statically decide on a deployment (identify cloud provider, number and type of instances, etc.). The Broker accumulates knowledge about the services offered by providers (public and private) and provides a unified interface to the developer, allowing them to instead specify requirements. The heterogeneous, hybrid cloud is abstracted through a set of APIs and the application may never be aware of the exact underlying infrastructure. The broker can be used to acquire infrastructure-, platform-, or software-as-a-service. Contributions include design and architecture of this broker, algorithms for matching requirements with resources, and understanding what level and quality of information the application should provide about its requirements to achieve best matching.

2.3.2 Analytic Algorithms for LSDA

The *Analytic Algorithms for LSDA* project addresses the problem of creating or adapting data analysis methods to the LSDA platform. These methods are useful themselves and provide natural test cases for the effectiveness of the LSDA platform. Three avenues to investigate are: application of data analysis to data stored in hybrid clouds, provision of data analytics as a service resident in a hybrid cloud, and techniques to make such services discoverable and utilizable on the LSDA platform. Highly useful forms of analytics include statistical analysis, data mining, and data visualization. In this project, we emphasize data mining because it is more challenging computationally than statistical analysis and lends itself more to being distributed in a hybrid cloud platform than visualization. *Data mining* provides ways to automatically or semi-automatically explore large quantities of data while searching for useful patterns. Although data mining techniques are specifically designed for large data sets, few existing methods have been tested in or adapted to hybrid clouds.

Objective 1 To design, implement, and evaluate data mining software for application to data sources located in hybrid clouds; and to design, implement, and evaluate data mining software as services located in hybrid clouds.

When mining data from hybrid clouds the main issues to be faced are: the scale of data, the lack of consistent structure in data sources, inconsistency of data from diverse data sources, and privacy concerns on behalf of owners/managers about the data potentially available from data sources in the hybrid cloud. Spatio-temporal data sets, which include the location and time of events as crucial fields for analysis, are particularly appropriate for this research project because much analysis can be usefully be performed on them in a domain-independent manner. Time series data sets are well suited to performing multiple independent analyses, e.g., looking for patterns at different amounts of lag [41] and different levels of granularity, e.g., hour, day, week, month, etc. [42], and such analyses can be performed on an extremely wide variety of data sets. Spatial data sets are often geographically distributed and in varying formats; they provide a good challenge for data mining and will require the development of new techniques. Data concerning recommendations, e.g., customer recommendations for shopping choices or ingredients for recipes [43] are also appropriate for this research because they have high utility and are particularly likely to accumulate in a hybrid cloud comprising company-run sites and independent ones.

A particular challenge for many data mining techniques is the enormous number of patterns that can be found in data. This problem can be addressed by applying interestingness measures [44] to constrain the generation of patterns or to filter the results [44]. In the context of hybrid clouds, the issues are: which interestingness measures should be employed, how should thresholds be set and adjusted in an environment where computing performance can adapt dynamically to the current workload, and how should results be integrated. Appropriate initial data mining problems are finding frequent itemsets [45] and finding effective data aggregations to produce interesting summaries [46].

Objective 2 To design, implement, and evaluate techniques for encapsulating data mining techniques as cloud-resident services that can be discovered and engaged by the LSDA broker being created by Project 1.

To provide a data mining service that can be easily discovered in the hybrid cloud by appropriate users, one approach would be to provide suitable meta-information that describes its requirements and capabilities (input format, output format, maximum sizes, speed, privacy guarantees). Another approach would be to some software induce these characteristics from the service or its operation. The latter is easier to use but more complex to implement. A spectrum of approaches between these choices will be evaluated.

In providing data mining techniques as discoverable services in cloud, additional challenges include locating suitable computation and I/O resources for the service to run on (and thereby deal with the scaling problem), preserving the privacy of input data, output data, and intermediate data while the service runs [15], and dealing with inconsistencies. Initially, we devise a data service that performs itemset mining using existing algorithms. Then, we will

devise a data service that filters data mining results by applying selected interestingness measures to these results. Using these two services, we will also address the problem of connecting them effectively and efficiency while preserving privacy. The implemented services can be profitably used as test cases for Project 4.1.3.

2.3.3 Scaling for Analytics

The *Scaling for Analytics* project addresses the problem of maintaining performance as the scale of the data sets analyzed increases. Previous research has emphasized the concepts of virtualization and workload balancing. By their nature, cloud architectures are well suited for scaling since the layer of virtualization and abstraction typically employed in a cloud allows for fine-tuned and responsive resource management not possible with traditional infrastructure. The most common approach is to add or remove machines dynamically in order to scale the infrastructure to meet service level objectives, which may be explicit or implicit (e.g., [47-50]). These techniques for managing a dynamic workload typically focus on e-commerce applications that are CPU-bound rather than analytics tasks where CPU, memory, and I/O each play important roles.

This research opportunity moves beyond the usual approach of adding or removing resources (typically focused on computation power, namely number of cores) and instead identifies the key characteristics of a Large Scale Data Analytics architecture that will enable scaling of the analytic work performed. This research will both enable and require extending the state-of-the-art for scaling of analytics.

Objective 1 To design, implement, and evaluate novel approaches for adaptively scaling analytic algorithms and services in response to a changing workload; in particular, where more resources are needed, design, implement, and evaluate strategies for deciding where to add these resources.

When applying analytics to increasingly large data sets, adding computation power may not resolve performance issues. This project will consider the cost-benefit trade-offs of a variety of actions other than adding computation resources: allocating more bandwidth, moving a running instance closer to a remote data source, moving an instance to a cloud with better I/O performance, factoring analysis tasks, and other remedial actions not typically considered in existing approaches.

When a task that is analyzing data across several private and public clouds requires more resources, a question not well answered by existing research is where to add resources. The most promising approaches are to determine where by considering constraints (capacity limits, privacy / security policies, data locality) or by optimizing criteria (available capacity, financial cost). Work may be migrated from a private to a public cloud with a different API and a different set of governing policies.

Objective 2 To design, implement, and evaluate strategies for changing the functionality offered by the cloud analytics service for cases where more resources cannot be added.

In general, adaptive management can change the functionality of a managed system. This under-explored research area offers high-potential opportunities for cloud analytics. By adjusting the data model, the analytic model, or even the analytics application, we can manage the demand for computing resources without adding resources. Adjustments could include introducing or modifying sampling, reducing data granularity (e.g., from seconds to days), switching from an exact algorithm to a more efficient heuristic, etc. This approach is valuable for private clouds, where there is typically a fixed cap on the amount of resources available. Novel contributions are expected to result from analyzing the data quality versus cost relationship and from identifying and evaluating the adjustments just mentioned as well as devising new adjustments.

2.3.4. Storage and data services for large data analytics

Large-scale data analytics typically relies on large-scale data sets. As long as processing capacity currently exceeds network capacity, the storage, retrieval, and management of this data

must be supported in a location proximal to processing power. *Data-as-a-service* makes available data and data sets accessible through a services interface, retrievable on-demand for use in analytics tasks. This project's purpose is to design, implement, and evaluate a Data Service to support large-scale data analytics, ensuring that data can be easily accessed, processed, visualized, used, and re-used.

Data may be stored to and retrieved from the Data Service from the local cloud environment, but also from producers and consumers on various networks. Not all producers/consumers will have the same experience; data access will have varying performance (throughput, response time). Rather than pushing high-granularity, complete data over a network link not sufficient for the task, differentiated Quality of Data may be used to provide abstracted or filtered data [10]. Data may be supplied in real-time, or with a delay, or in batch transfers as possible given network capacity.

There are also open questions about how to collect, store, and distribute large-scale information. Such a data service should be capable of acquiring data from multiple providers and federating that data for analysis. Storage may involve NoSQL, SQL, or both; given the costs and benefits of collecting and processing data via either approach, best practices on which approach to use for a given data set are not yet established. Finally, streaming data is of growing importance; the introduction of streaming data to Hadoop, and the use of dedicated stream processors like S4 (Yahoo) and Storm (Twitter), suggests the importance of supporting data streams, particular for real-time data.

As with any system storing and managing data, security and privacy are prime considerations. There are concerns based in technology: the adoption of clouds (and their basis in virtualization on physically shared infrastructure) and mobile devices (portable, operating on different networks) to data analytics presents privacy and security concerns that must be addressed [51-53]. There are also basic issues of how data is correctly segmented at scale to be shared only with those entitled to see it, how data is placed on private and public clouds, and how data can be secured without deleterious effects on performance [23]. Existing work has addressed this with restrictive policies that required data labeling and limiting the analytics that could be performed (e.g., [44]).

Objective 1 To draw from analytics use cases to iteratively establish, implement, and evaluate an architecture for data-as-a-service, considering proximal access and remote access; various collection, storage, and aggregation mechanisms; and streaming real-time and batched data.

The Data Service will play a key role in collecting, storing, and distributing data to stakeholders for performing tasks including analysis and visualization. Requirements will be established in consultation with other Projects and Themes. It will allow for the creation and curation of large-scale data-sets, optionally federated from multiple sources, that can be provided to analytics tasks with fine-grained control over access (see Objective 4.2). Storage design decisions will be made after systematic examination of NoSQL and SQL approaches to data management. Performance considerations like throughput and response time will be key requirements. The advantages of streaming data will be evaluated empirically in realistic scenarios for possible inclusion in the architecture.

Data delivery will be examined in the context of adaptive management of available resources. Potential adaptations include adjusting data quality in response to the quality of network connections, switching among real-time, delayed, and batched data transfers in response to requirements and capabilities, and multiplexing data streams to preserve bandwidth.

The architecture will be iteratively developed and refined as other stakeholders to support analytics tasks use it and strengths and weaknesses are developed. The result, in addition to an architecture meeting the specified use cases and requirements, will be a general approach to the management of large-scale data in the cloud.

Objective 2 To investigate end-to-end security and privacy concerns for large-scale analytics on cloud infrastructure and develop, implement, and evaluate mechanisms, protocols, and authentication mechanisms to support data-as-a-service

Security and data privacy are key issues in the deployment and usage of any cloud application or infrastructure. The advent of smartphones and the widespread use of other mobile devices such as tablets have introduced a new dimension to cloud computing, offering a higher degree of flexibility in data access, and with it has increased the need for security standards. We propose to look at a range of security and privacy issues that arise in large-scale analytics applications deployed in hybrid. They include permitting fine-grained segmentation of existing data-sets into subsets that can be shared with different groups, approaches to assuring and ensuring data security on cloud infrastructure, and correctly placing data on infrastructure best-suited to the sensitivity of the information (e.g., public versus private cloud). Technology considerations include examining data transfer protocols and security mechanisms on all platforms (clouds, PCs, mobile devices, etc.).

We propose an architecture to manage security and privacy end-to-end in the data service, protecting providers and consumers. This architecture will be evaluated on an experimental test bed, using applications for key areas in the analytics space and central to Canada's knowledge economy (e.g., mobile commerce, healthcare, digital humanities, bioinformatics). The proposed architecture and the evaluation will align a generic framework and a unified model for security that will enable industrial prototypes and a set of security solutions applicable to other cloud environments and applications.

3 Computational Biology

We emphasize cancer treatment, development and application of novel tools that integrate, analyse and interpret complex biomedical data to help identify testable hypotheses and construction of useful models. Existing algorithms either do not scale with the size and complexity of realistic problems or provide only partial answers. Visual analytics are required to provide tools for scientists, clinicians, and patients.

3.1 Problem

Merely coping with the deluge of data is no longer an option; their systematic analysis is a necessity in biomedical research. Big data in biomedicine derives from two sources: genomics-driven (genotyping, gene expression, and now next-generation sequencing data), and the payer-provider (electronic medical records, pharmacy prescription information, insurance records).

From genomics, with next-generation sequencing (a process that greatly simplifies the DNA sequencing), it is now possible to generate whole genome sequences for large numbers of people at low cost. Raw data-wise, it is 4 terabytes of data from one person. Imagine sequencing for thousands of people in the course of a month. Petabyte scales of raw data are generated. How do you manage and organize that scale of information in ways that facilitate downstream analyses?

Computational biology is concerned with developing and using techniques from computer science, informatics, mathematics, and statistics to solve biological problems. Analyzing biomedical data requires robust approaches that deal with (ultra) high dimensionality, multimodal and rapidly evolving representations, missing information, ambiguity and uncertainty, noise, and incompleteness of domain theories. However, advancing computational tools alone is insufficient to have an impact on computational biology and related biomedical fields. Many theoretically excellent approaches are inadequate for the high-throughput (HT)

biological domains, because of the scale or complexity of the problem, or because of the unrealistic assumptions on which they are based. Developing new algorithms for real problems provides added value to the interdisciplinary approach, and ensures access to data and diverse expertise for validation during algorithm performance evaluation. *We focus on developing and applying machine learning and visualization algorithms in cancer informatics.*

3.2 Objectives

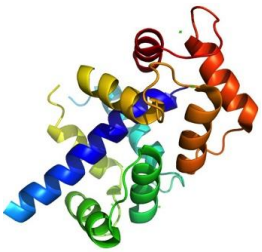
We will develop and apply novel tools to integrate, analyze and interpret complex biomedical data to help identify testable hypotheses and construct useful models.

3.3 Background and methodology

To achieve these objectives, we investigate the following lines of research:

3.3.1. Network analysis/visualization/integration [synergizing with the interactive content analytics theme in terms of user interface/interaction and with health informatics and large-scale data analytics in terms of ontologies/ representation].

Biology offers a diversity of problems, leading to many computational biology workflows, including tasks where network visualization is helpful during data interpretation and analysis. Individual network visualization systems differ greatly in terms of features and standards they support, and consequently analyses they enable. Importantly, users have a broad range of skills and expectations, and thus, network visualization tools must satisfy diverse requirements and offer different user interface and features.



Network visualization and analysis tools can be compared using diverse metrics, and clearly, none of the existing tools supports all workflows. Several good reviews of important features and system comparison have been published [55-57], and several papers reviewed useful workflows/scenarios in network visualization, integration and analysis [55, 58, 192], guiding both users and tool developers.

Diverse data can be represented as a graph – physical protein-protein interactions, metabolic networks, genetic interactions, microRNA-to-target associations, gene regulatory network, correlation, similarity relationships, etc.

While many papers incorporate some form of network visualization, the potential of existing tools is neither fully realized, nor is the existing use optimal in many cases. Several critical challenges have to be addressed by useful visualization tools: 1) large and complex networks need to be handled; 2) interactive manipulation of the network should be supported; 3) biological insight from the analysis and visualization is required; 4) diverse biological datasets and annotations must be easily integrated with the network; and 5) besides data visualization and interpretation, it may be necessary to support visual data mining. The most basic features of network visualization tools include layout, analysis, import and export formats.

While many existing visualization tools are effective and widely used, there are several critical areas where these applications require improvement, as reviewed [57]. Scalability is essential to visualize the tens of thousands of known protein-interaction networks, which is a challenge for current layout algorithms and software. Biological graph drawing software must also be able to handle richly annotated data, including genomic and proteomic profiles, pathways, GeneOntology annotations [59], data in PSI-MI [60] and BioPAX [61-62] formats (http://www.biopax.org/), in addition to the vast quantity of microarray data that are currently available.

As the data get more complex, performance of layout algorithms will need to improve, and new options of differentiating multiple attributes will be required. As certain workflows become more mainstream, they may be turned into *analysis patterns* and implemented as plug-ins. Standardizing file formats, API and plug-ins will further intertwine existing tools, enabling their easier integration and specialization. Importantly, none of these algorithms would make a broad difference unless a user interface appropriate for biologists is available.

3.3.2 Prognostic signatures [integrating network-based approaches with advanced machine learning to comprehensively identify all clinically useful prognostic signatures. Focus on lung and ovarian cancer].

Canadian cancer statistics (2012) indicate that lung cancer accounts for almost 14% of all cancer cases in Canada, leading to the highest number of deaths (20,100, or 27% of all cancers). Recent studies rely on the development of customized therapeutic solutions for the treatment of individual patients (personalized medicine) that require the analysis of biomarkers to form a signature, specific to the disease and patient. The discovery and validation of signatures are complex and computationally intensive. It involves analyzing a large number of parameters (clinical variables, gene, protein, microRNA, activity, etc.) to identify subsets that best describe patients, their prognosis and response to treatment. Selecting the best subset represents a challenging computational optimization task.

Existing prognostic and predictive biomarkers overlap only partially, and usually do not validate on external cohorts or by other biological assays. Several technical reasons may explain this: 1) patient diversity and tumour heterogeneity, 2) range of HT profiling platforms with different biases and noise, and 3) application of diverse statistical and bioinformatics approaches to marker identification. Further evidence shows several clinically identical gene signatures exist [63], and the best signature may not comprise the most differential genes [64]. It has been suggested that we need thousands of samples to find overlapping signatures [65]. To address these challenges, we need to integrate multiple datasets and diverse algorithms for their analysis (such as http://ophid.utoronto.ca/cdip), and conduct multi-tier validation of putative markers.

To address these challenges, we propose to conduct an integrative analysis and comprehensive computational and biological validation of putative markers. To date, some of the most successful network-based methods of gene group identification for class prediction have been the score-based sub-network markers. Sub-networks identified using these approaches were recently shown to be highly conserved across studies and to perform better than individual genes or pre-defined gene groups at predicting breast cancer metastasis.

In addition, these findings will provide insights into how we can optimize the computational tasks as a whole for other cancers and for other data sets (e.g., proteomic, genomic, epigenetic). Further, this data will be used to re-purpose existing FDA-approved drugs for combination treatment, using our drug-related resources: SCRIPDB [67], NetwoRx [68] and CMapBatch.

Combining molecular cancer profiling and computational analysis will enable "personalized medicine," where treatment strategies are individually tailored based on combinations of single nucleotide polymorphisms, gene, protein or microRNA expression levels in biological samples [69-73]. Integrating genomic and proteomic profiles with protein interactions will enable objective target selection for identification and validation [74-81]. Combining binary interactions into networks will implement systems-level understanding of cancer biology, and will expand our knowledge of individual cancer pathways and their cross-talk [82-83].

Comprehensive analysis of signatures from CDIP (http://ophid.utoronto.ca/cdip) in lung cancer resulted in 3- and 6-gene prognostic signatures [64, 66] and 15-gene predictive signature [77], currently being commercially pursued by Pittsburgh-based Precision Therapeutics. Similar analysis, integrated with protein interactions from I2D [84], identified early disease biomarkers in ovarian cancer, leukemia, head & neck cancer, pancreas cancer, prostate cancer [75, 76, 85, 86].

Despite inherent noise present in interaction datasets, systematic analysis of resulting networks can uncover biologically relevant information, such as lethality [87, 88], functional organization [89-91], hierarchical structure [92, 93], modularity [94, 95] and network-building

motifs [96-98]. This suggests that networks have a strong structure-function relationship [98], which can be used to help interpret integrated cancer profile data. Many interactions are transient, and the networks change in different tissue or under different stimuli [99, 100]. Studying the dynamics of these networks is an exponentially more complex task. Many stable complexes show strong co-expression of corresponding genes, whereas transient complexes lack this support [101, 84]. These contextual network dynamics must be considered when linking interactions to phenotypes, and when studying cancer networks topology. Adding this to different biases of single detection methods, the simple intersection of results achieves high precision at low recall cost. Systematic graph theory analysis of dynamic changes in interaction networks, combined with gene/protein cancer profiles, enables integrated cancer analysis [102-104, 80, 66]. Developing algorithms using heuristics tuned for interaction networks [105, 106] ensures scalability.

Some of the most successful network-based methods of gene group identification for class prediction have been the score-based sub-network markers [107-110]. Sub-networks identified using these approaches were recently shown to be highly conserved across studies and to perform better than individual genes or pre-defined gene groups at predicting breast cancer metastasis [107]. Considering network modularity improves these methods, and results in better prediction of aging [95]. Combining existing known and predicted interactions from I2D with novel local co-expression annotation of existing edges will elucidate disease-specific dynamics and identify local network structures (graphlets, [106-111]) that are the most aberrant components in the cancer network, as compared to normal.

The goal is to prioritize putative biomarkers, validating first those that are over-expressed, are on amplified cytobands, are secreted, and are functionally linked to other biomarkers. While gene expression-based classifiers may predict outcome, they can also be used to identify the most important pathways in cancer using systematic biological annotation and analysis of the corresponding protein interaction network. These pathways may represent the core functional milieu of prognostic classifier genes, and thus may represent the processes underlying the specific cancer. Identifying tissue-specific and age-related markers will help to predict early events in cancer development [112].

3.3.3 High-resolution structure prediction

The goals of this proposal are to combine innovative bioinformatics approaches with emerging experimental techniques for high-resolution structure predictions, and to develop a new avenue for understanding protein folding. [Knowing the structure of a protein is key to understanding how it works and to targeting it with drugs. A small protein can consist of 100 amino acids, while some human proteins can be huge (1000 amino acids). The number of different ways even a small protein can fold is astronomical because there are so many degrees of freedom. Figuring out which of the many, many possible structures is the best one is regarded as one of the hardest problems in biology today.] Konermann's group recently devised a ground-breaking biophysical method for tracking protein structural changes during folding. This technique is based on rapid mixing of denatured protein with refolding buffer. After various time points (milliseconds to minutes), the protein is briefly exposed to hydroxyl radical (·OH) via a pulsed excimer laser that introduces oxidative modifications at solvent accessible side chains, whereas buried regions are protected. The protein is then cleaved with a suitable protease, and the resulting peptides are analyzed by liquid chromatography and tandem MS (LC-MS/MS).The oxidative labeling pattern undergoes dramatic changes during folding, reflecting the transition from a disordered structure to the compact biologically active state. Analysis of the protein oxidation pattern provides insights into the nature of short-lived intermediates that become populated during folding, and that are unobservable by traditional high resolution (X-ray or NMR) techniques. In a complementary fashion, hydrogen/deuterium exchange (HDX)/MS experiments report on the evolution of protein secondary structure during folding. Our idea is to use these experimental data as constraints for computer-based structure prediction algorithms. We face and will overcome major bioinformatics challenges below:

1. The oxidative labeling pattern imprinted onto the protein carries detailed information regarding the solvent-accessible surface area of individual side chains. Unfortunately, current data analysis strategies only permit reliable quantitation of oxidation data at the peptide level, "smearing out" the available information over 10 or more residues. A new approach will be developed to calculate the oxidation level of single amino acids from the LC-MS/MS data. Differently labeled versions of the same peptide are often isobaric and co-elute, resulting in overlapping MS/MS spectra. Existing algorithms cannot tackle such a scenario. For this purpose, Ma will combine his spectrum prediction algorithm with large scale optimization to compute the relative quantity of each differently labeled peptide in the same overlapping spectrum. This will help to map structural features of both native proteins and folding intermediates with unsurpassed spatial resolution.

2. Existing structure prediction algorithms, such as FALCON by Ming Li's group, use an energy function as a guide for making gradual changes to the protein structure during folding *in-silico*. At present, these simulations have no relationship to the actual folding pathways of "real" proteins. As a result, the algorithms can suffer from poor convergence, and they often get trapped in local minima (there are billions of different pathways by which a protein chain might fold). A new algorithm and scoring function will be developed that takes experimental accessibility and HDX data to guide the simulations along to discover the actual biophysical pathway and guide the simulations along it. This approach will dramatically improve the reliability of the final predicted structure. In addition, it will for the first time elucidate the 3D structures of solution-phase folding intermediates. These transient species act as kinetic branching points that can lead to misfolded aggregates (e.g., in Alzheimer's and Parkinson's disease). Development of potent folding visualization software represents an added focus of our plan.

3. Less than 0.1% of all known bio-molecular structures represent membrane proteins, despite their enormous medical importance. We extend approaches outlined above to membrane proteins, thereby working towards 3D structural models of key drug targets that have been intractable by other means.

Our unique approach makes it possible to generate reliable protein structure predictions in a matter of hours, whereas complete X-ray or NMR investigations require weeks or months. Hence, our initiative will help propel the scope of large-scale structural genomics initiatives to an entirely new level.

3.3.4. Crystallization pattern discovery

The NIH Protein Structure Initiative consortia enabled many technological breakthroughs. However, the benefits of high-throughput (HT) approaches to structure determination must be considered in the context of their drawbacks [113-132]. Based on TargetDB [69] data, of 177,330 proteins cloned, 8.3% of the total crystallized, 4.1% of the total diffract, and 3.8% of the total resulted in structures. While all steps in the structure determination pipeline, from protein production, through screening, to crystallization and structure determination can be improved, we focus on computationally-driven optimization of HT crystallization strategy.

There are three main approaches to structure determination: 1) *in silico* prediction [133-141, 127], 2) NMR (Nuclear Magnetic Resonance) [142, 128-130, 143-147], and 3) X-ray crystallography [148-154]. While NMR approaches are growing in their importance [155-157, 130, 142], and *in silico* methods are becoming more accurate [133, 137], single crystal X-ray crystallography remains the most generic and powerful method for protein structure determination.

Technical difficulties in protein crystallization are due mainly to two reasons: 1) many parameters affect the crystallization outcome, e.g., purity of proteins, super-saturation, temperature, pH, time, ionic strength and purity of chemicals, volume and geometry of samples, and 2) we only partially understand correlations amid parameter variation and the propensity for a given protein to crystallize.

Conceptually, protein crystal growth can be divided into two phases: search and optimization. Search phase determines a subset of all possible crystallization conditions that yield a promising crystallization outcome [158]. These conditions are varied during the optimization phase to produce diffraction-quality crystals [159]. Neither of the two phases is trivial to execute. If we consider only 15 possible conditions, each having 15 possible values, the result would be 4.3789e+017 possible experiments; impossible to test exhaustively. Even a broad search phase may not produce any promising conditions, and many of the promising leads may elude optimization strategies. This large search space of conditions can be systematically screened with robotic liquid handling techniques, which turns a protein chemistry problem into big data analytics.

Systematic HT crystallization screen data analyzed by diverse computational algorithms will lead to an increased number of structures being determined. HT screening can speed up the search phase, and may increase process quality [158, 160, 129, 161-162, 152, 163-167, 159]. The computationally-optimized approach will increase the number of structures by combining and further improving successful image classification that will enable comprehensive mining of 19 million crystallization experiments. Comprehensive data mining will optimize crystallization screens and identify patterns in data that will lead to improved understanding of crystallization process and in turn more structures.

The current image classifiers both benefit and suffer from a wealth of image analysis data. The analysis software, describing a range of textural and statistical features in the image, measures a total of 14,908 features per image. The same image analysis algorithm computes groups of features across a range of parameter settings, in an effort to capture optimal parameter settings.

Current feature selection is made in two stages using the Random Forest (RF) algorithm. In the first pass, an initial RF is trained on all 14,908 features of the training data. The variable importance measures of each feature (a product of RF training) are used to select the 10% most informative features. Only the selected features are fed to the second pass of the RF algorithm, resulting in a higher-quality classifier. Although the performance of this classifier is promising, one drawback to the feature-pruning method has been identified: RF variable importance can assign high importance to multiple, correlated features. A preliminary look into the 10% of features selected indicates that multiple, correlated members of certain feature families are included in the 10%, and thereby exclude less informative but uncorrelated features from the classifier.

A new feature-selection approach that we are developing combines RF variable-importance criteria with the correlation information from the feature set. Information from general feature-feature correlation and per-image-class correlation matrices will affect the selection process.

Further improvements in scoring accuracy will be obtained from incorporating time-series information. Each trial is photographed at multiple time-points, and a probabilistic relationship links the state of a crystallization trial at any one point with its past and future states. Whereas the image classifier calculates a belief $P(st|ft)$, i.e., belief in the state st of the trial at time t given the image features ft, a model incorporating the time-series would also incorporate relationships $P(st|st-1)$ and $P(st+1|st)$ for a complete a posteriori assessment of the state (i.e., score) st.

Discovering and illuminating the relationship between proteins and cocktails in the crystallization process will increase the success of protein crystallization trials and improve the process of crystal optimization. Chemically-similar cocktails will react similarly with a given protein [168-169]. Thus, a screen of 1,536 cocktails represents a sampling of points of chemical space in a protein's crystallization landscape. Equally, similar proteins should react similarly with a given cocktail, and thus any one cocktail screened against HWI's 12,000 protein samples to-date represents a sampling of points of protein space in a cocktail's crystallization landscape. Neighbouring proteins in the crystallization landscape are those whose crystallization reactions are correlated across a range of cocktails. Neighbouring

cocktails are those whose crystallization reactions are correlated across a range of proteins. (The chemical constituents of the cocktails may also be used to infer cocktail-space neighbours.) Knowledge of the crystallization reaction of a given protein with a given cocktail implies partial knowledge of the reaction at neighbouring points, in protein space or cocktail space.

The correlations required to derive these probabilistic relationships must be computed from the scores of multiple crystallization trials. HWI's archive of 120 million crystallization-trial images comprises 12,000 proteins against 1,536 cocktails. Detailed scores of 165,000 images have been scored by experts at HWI [168, 169]. The remainder will be scored by automated classification and image analysis. Computation-intensive image analysis is performed on the WorldCommunityGrid GPU/CPU.

We will optimize the existing HT crystallization screen by applying systematic data mining. Discovering the relationship between proteins and cocktails in the crystallization process will increase the success of protein crystallization trials and improve the process of crystal optimization. Chemically-similar cocktails will react similarly with a given protein [168-169]. Analogously, similar proteins should react similarly with a given cocktail. Knowledge of the crystallization reaction of a given protein with a given cocktail implies partial knowledge of the reaction at neighbouring points, in protein space or cocktail space. Systematically identifying patterns in the protein/cocktail space will lead to an optimized set of informative cocktails. Importantly, clustering of the image data may lead us to identify more optimal, data-driven crystallization image classes.

We will apply association mining to identify important patterns in the large crystallization data warehouse – proteins and their properties × crystallization conditions × crystallization result × 6 time points. Structural-chemical properties will be computed from sequence by PSIPRED [170] and the pepstats application from the European Molecular Biology Open Software Suite. This information will be combined with the detailed description of 1,536 crystallization conditions, 10 classes from image analysis, and analyzed across six time-points to identify both global and local patterns.

It has been shown that systematic mining of the crystallization data provides useful leads to improve crystallization success rate, and to increase our understanding of the protein chemistry and crystallization process [171-179, 158, 180-189]. However, these results contained only successful crystallizations, many from smaller, partially biased protein classes and families. To learn the process and principles, we also need the failed experiments. HWI screen provides us with a broad range of proteins and diverse outcomes

Association mining identifies frequently occurring patterns among variables and their values [190, 187, 189]. Our FPClass algorithm will segregate proteins into groups based on how they react in a broad range of conditions and group cocktails to reflect their potential to achieve crystallization. These results may lead to optimizing the crystallization screen, and reveal associations among protein properties and the crystallization screen, in relation to time and composition of individual cocktails. The database of crystallization results is considered a transaction database. Transactions represent sets of facts about proteins, cocktails, or particular experiments, e.g., {57, pH < 6.0, contains: PEG 20000, CAPS, crystallizes proteins: Z1257, Z1258, Z1875}. An itemset is a subset of a transaction. Frequent-itemset discovery finds all itemsets in a database whose frequency (support) exceeds some threshold. An association rule describes the co-occurrence of two disjoint itemsets in transactions, such as: {protein concentration >10 mg/mL, medium molecular weight, low pI, organism: A. aeolicus, cocktail contains CaCl2*2H2O} → {crystal}. The right side of this rule is supported by 20 of the 25 transactions that fit the left side of the rule in our database, and thus the rule has support 20 and confidence 0.8 (20/25). Association-rule discovery finds all association rules in a database whose support and confidence exceed minimum values [187, 189, 190].

With the protein at hand, and a small list of the most highly similar proteins encountered in the crystallization repository, a new crystallization plan can be formulated. At first the

formulation will be done by manual interpretation of the successful crystallization conditions identified in the most highly similar proteins treated in the database. Results from data mining may be used to further optimize the screen, by identifying probable favorable crystallization conditions for a given protein. A list of the chemical components, their concentrations ranges, and ranges of pH identified in successful crystallizations of similar proteins will be used to generate a new screen of crystallization cocktails.

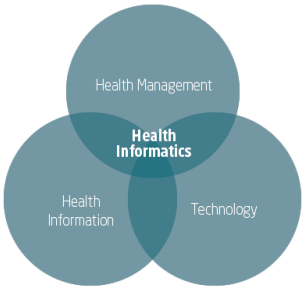
The proposed comprehensive analysis of HT crystallization screen data is possible both due to WorldCommunityGrid [191] resources with over 1.5 million computers, of which we use 300K a day in average (http://www.worldcommunitygrid.org/research/hcc1/overview.do), and our own computing facility that enables archival storage as well as pre- and post-processing.

4 Health Informatics

Our perspective on health informatics is a study between computer and information science and healthcare. Health informatics encompasses the resources, devices, and methods required to optimize the acquisition, storage, retrieval, and use of information in health and biomedicine. The tools we build include clinical guidelines, formal medical terminologies, and information and communication systems.

4.1 Problem

Free-form texts are the most common form of valuable data in healthcare. They range from doctors' notes, descriptions of patient histories, to healthcare-related messages posted by patients on social media such as blogs, bulletin boards, and discussion forums. Such narrative text data contain the most valuable information for physicians to use in their practice and for public and government agencies to make their healthcare-related decisions. Recently, the New York Times reported on a study by MIT researchers, showing companies included in their study that adopted data-driven decision making achieved 5-6% higher productivity than those that did not. However, since data are continuously generated every day in large volumes, the sheer amount of data is too overwhelming for humans to read and analyze manually. Automatic text analysis tools are in great demand to discover the hidden information trapped inside the free-form texts. For example, a tool that identifies and analyzes the healthcare-related posts in social media can detect the public opinions, activities and preferences in healthcare-related issues. Natural language analysis, data mining and information retrieval are key techniques that can be used to build such text analysis tools.



4.2 Objectives

With the general objective of discovering hidden and valuable information from healthcare-related text data, we propose to develop innovative text analysis techniques for analyzing healthcare-related data on social media. Our ultimate goal is two-fold. First, we will develop tools that provide policymakers and program managers with the information needed to plan and implement successful health care initiatives. Second, we will develop tools that can potentially identify health patterns in socio-demographic and geographic data to help consumers assess

health risks and educate the public for active participation in preventative healthcare. We choose to focus on social media data because they are publically available, are sensors of the real world and there are high demands from industry and government to discover and keep track of public activities and opinions expressed in social media. We will also use the socio-demographic and psychographic data collected by our partners (e.g., Manifold Data Mining) to find out how population health is impacted by external factors in addition to their biological and physiological conditions in order to develop tools for preventative health-care.

4.3 Background and methodology

To achieve these objectives, we propose the following lines of research:

4.3.1. (part a) Resource selection for seeking information about healthcare issues

There are numerous forums and networks on social media where people share their opinions and insights. Not all of them are relevant to either general or specific health-care issues that are of concern. Appropriate and wise selections of forums and relevant posts in social networks are fundamental for the successful discovery of useful hidden information on relevant health-care issues.

In the field of information retrieval, many algorithms [e.g., 193-195, 198] have been proposed for resource selection in federated text search. Many of them can be viewed conceptually as treating each information source as a "big document", and using variations of traditional document ranking algorithms to rank available information sources with respect to a user query. However, recent research [196] has demonstrated that the "big document" approach does not work well when there are wide variations in the sizes of available information sources because the "big document" approach does not consider the individual documents in the information sources. The GlOSS algorithm [197] turns away from the "big document" approach by calculating the veracity of information sources to a query as the number of relevant documents, but no satisfactory method has been provided for estimating the veracity. The hierarchical database sampling resource selection algorithm [198] and the classification-aware selection algorithm [199] use base resource selection algorithms such as "CORI" and the GlOSS approach, and still suffer from the weakness of the base algorithms. A set of more robust algorithms [e.g., 196, 200] has been designed to estimate explicitly the veracity/usefulness of individual documents that each information source contains. However, such measures were not designed for evaluating documents (i.e., posts) on social media. We will work on designing goodness measures for evaluating a discussion forum and individual posts in a forum in social media. Resource retrieval systems will be developed to collect the relevant sources of information based on the designed veracity measures.

4.3.1. (part b) Adaptive information extraction and topic/event detection and tracking

Free-form discussions of medical or health-care issues on social media do not follow the standards for classifying and coding medical information such as the Systematized Nomenclature of Medicine (SNOMED). A useful first step will be to design a system that is able to extract key elements from narrative texts that are relevant to the task(s) at hand.

We will work on two directions for identifying the key information embedded in the text: 1) adaptive information extraction that automatically learns extraction knowledge from data to extract structured information such as entities, attributes and relationships between entities from unstructured text [201, 202], and 2) advanced topic and event detection that considers the content as well as the temporal and social dimensions of the data [213-215].

For information extraction, we focus on improving the robustness of current IE systems (e.g., WHISK [206], RAPIER [207], SRV [208] and LP2 [209]) which can automatically learn extraction knowledge from data, and on easing the difficulty of adapting an IE system to different extraction tasks. We will fully examine the naive Bayes IE model as a purely adaptive IE model, in which the formulation problem existing in previous naive Bayes IE systems is

corrected. We also investigate the effect of smoothing techniques in this context (essentially a general issue associated with any probabilistic model), and we design our own smoothing strategy [210] to obtain more stable probability estimation in statistical IE learning. Our initial experimental results show that a good smoothing method is critical to the robustness of naive Bayes IE systems. In most existing probabilistic systems, a natural evolution from the naïve Bayes' models is to more advanced Hidden Markov models (HMMs) [211]. Our work on HMM IE will solve the extraction redundancy issue in current HMM IE modeling on an entire document. To this end, we propose a segment- based HMM IE approach, in which a segment retrieval step is included to identify extraction related segments from the entire document. Note that our segment-based IE modeling is actually a general framework. In addition to HMM IE applicability, the same segment-based IE framework applies to other IE models in which the extraction is performed by sequential state labeling. To improve the system's adaptability to the situation when the labeled texts are limited, we will extend our segment-based HMM IE modeling to semi-supervised learning using a modified version of the multi-view Co-EM learning strategy.

Motivated by the need to choose term weighting related system design choices in segment retrieval in our segment-based HMM IE, we also investigate the use of information theoretic principles as tools to analyze the term vector models employed in IE and information retrieval (IR). Thus far, a series of theoretical analyses [210] show that the information theoretic principles provide a good framework to help make related design decisions in term vector models with sound theoretical justifications. We advocate an integrated IE system in which different learners are selected not only according to a particular IE domain, but also based on the characteristics of different extraction tasks (i.e., different slots in an extraction template). Inter-slot relationship utilization in the integrated IE framework will improve extraction performance. Our concept of *redundant extraction* will adapt an existing IE system to perform extraction with redundancy. By introducing some redundancy, the IE system can identify more extraction related information from documents thus providing a solution that bridges the gap between performance limits of current IE systems and performance needs stipulated by applications.

For topic and event detection, most of the existing methods use clustering techniques or the generative Latent Dirichlet Allocation (LDA) probabilistic model [212] to group the documents according to the content of the text. A problem with most existing topic detection techniques is that the topics are represented by a set of words, which together may not be meaningful. We will investigate how to integrate computational linguistics techniques with probabilistic topic modeling to extract meaningful topics. We will also extend these techniques by considering both temporal and social factors so that topics or events are related to the context in which they are discussed or occur.

4.3.1. (part c) Development and testing of indicators of public interest in healthcare activities

We need to create indicators or indices to measure the activity of consumer participation with the goal of finding factors that influence and initiate consumer involvement in health-care choices. While formal surveys exist to measure patient participation, there are no universal indicators that track ongoing real-world consumer activities in or opinions about health care. To measure consumers' interest and participation in healthcare decision-making from social media, we will create a set of broad indicators of healthcare-related activities that operates independently of specific conditions and treatments or providers. For example, an indicator can be related to choice-making: are consumers actively trying to make choices, or are they passively waiting for advice?

We will create and test indicators along the following dimensions:

- *Proactive orientation*. To what extent are consumers asking questions or seeking information before healthcare-seeking events, versus commenting afterwards. For

example, how often are consumers seeking different options or opinions after being recommended for a medical procedure?

- *Comparison-making*. To what extent are consumers comparing one health-care choice/source of care to another, rather than open-ended requests for options. Is this proactive or retrospective? For example, how often are consumers comparing price/cost?
- *Retrospective assessments of process/outcome/price*. To what extent are consumers making assessments informally or formally about health-care process (service, waiting time, appointment-making), outcomes (feeling better, recovering, healing) and price?
- *Sentiment regarding health-care services*. To what extent are retrospective assessments positive, negative, or neutral?

We will create and test these indicators using both supervised and unsurprised machine learning techniques on selected social media data.

4.3.2. (part a) Sentiment and emotion detection

The rise of social media has provided various forms for public to express their opinions on various issues. Businesses have tremendous interests in sentiment analysis over social media data in order to analyze markets and identify new opportunities. Most of existing sentiment analysis techniques determine the overall sentiment orientation for a single object as positive, negative or neutral [213] but they do not specify exactly what reviewers like or dislike. However, quite often, a person likes certain aspects of an object but dislike others. Such mixed feelings cannot be extracted by sentiment analysis techniques that only identify the polarity (i.e., positive/negative/neutral) of the sentiment.

Some agreement on categories exists, yet debates remain as to what are fundamental emotions. Emotions have variable intensity and emotions mix; hence emotion may be best represented through overlapping three-dimensional diagrams. Picard [214] notes the lack of emotion definition and whether the lack of agreement on basic emotions or continuous spaces of emotion is an obstacle to computer recognition and synthesis. There is little proof that emotion is uniform across cultures. Ekman [215, 216] insists that some emotions are universal, yet remains inconclusive regarding others, in particular contempt, shame and interest. Ekman [215] acknowledges that differences arise in the ways that cultures teach their members how to manage emotions and in what triggers or causes specific emotions. Lazarus' theory of emotion [217] includes interdependent mediations through environment, personality, process and outcome expectations – context! He provided a comprehensive analysis of the ways that culture influences emotional expression by forming 'shared and divergent meanings' acquired over the course of psychological development. Emotion plays out in how people make sense of life's events, 'how a person perceives, understands, and appraises what is happening socially'. He noted, 'culture could have a major influence in both a constitutive and regulative sense on the goals we acquire–and on other appraisal components'. Lazarus' approach aligns with Ortony, Clore and Collins' Cognitive Structure of Emotions (OCC) model that represents emotional process or 'valence' according to cognitive eliciting conditions. The OCC model has been pre-eminent in designing affective computing systems as rules can represent if/then [245].

Despite these limits, analyses of digital media using emotional categories have proven to have value with the growth of social media and their effect as tools for communication or collaboration. For example, Shami et al. [218] argue that it is necessary to measure group emotions, not only individual, in the context of social media, and acknowledge that measuring affect "is a challenging and complex issue." They propose that analysis of emotion in online contexts should use "implicit measures such as analysis of linguistic cues" rather than self-reporting which is unreliable. Discovery in the structure of affective extraction that tests different assumptions and compares these is of ongoing necessity.

We propose a novel sentiment analysis approach that employs four indicators in four linguistic levels, containing clause, phrase, word, and feature levels, to determine sentiment polarity. We also identify object features for which sentiment is expressed, and propose to apply feature weights to determine object sentiment polarity. Features permit us to know on

what object aspects a consumer is positive or negative, providing refined sentiment analysis where mixed sentiment is expressed.

We will extend our current emotion detection method (the initial approach was developed in the CIVDDD project) to consider wider context, and relationships and factors that affect the emotion conveyed by the text. We will also extend emotion categories being considered.

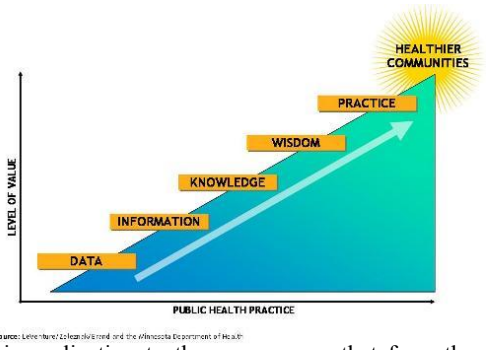
4.3.2. (part b) Discovering and tracking patterns of information seeking, comparisons and sentiment about health care.

After identifying indicators of public activities and opinions in health-care, we would like to find patterns in which consumers seek information, make comparisons and express sentiments, and also track changes in these patterns over time and in response to significant public events or stories. To this end, we will conduct the following lines of research:

a) Compare levels of activity across geographical locations, demography, conditions or other factors. Using the identified indicators or indices of public activities and opinions, we will compare different groups of consumers in terms of these indices;
b) Conduct root-cause analysis to identify the factors or events that result in, for example, positive or negative sentiments. We will combine our findings on event detection and sentiment detection, and find the relationships between them by using data mining techniques; and
c) Conduct trend analysis to track changes in the identified patterns over time.

4.3.2. (part c) Data Visualization Strategies and methodologies for social media content and free form texts

Lee [219] argues that user interface design must integrate form and function, rather than accept a time-worn polarization that privileges function. Hence affective computing has a significant potential application within visualization design. Engaging the body and mind in emotion and feeling allows for an immersive and embodied relationship to knowing – suggesting performative ways of navigating data sets, but also underscoring the subjective nature of image experience. Tractinsky et al. [220] propose that interfaces are more engaging when presenting emotionally compelling content. Their analyses of users' response to various metaphors and navigation strategies include the analysis of affect and related aesthetics. The study of multiple approaches to visualization of datasets, user engagement and protocol emergence for data visualization design remains critical.



We will study variable aesthetics and their application to the user groups that form the collaboration, using participatory design, design and data-driven design methods as well as usability testing.

Tufte [221] argues that data visualization requires choosing data sets that are of research value, mining the data, creating a structure for data, analyzing that data to find meaningful representations, analyzing patterns, translating analysis through aesthetic representation, refining the representation to communicate better, and creating means of manipulating data. In Tufte's view, data enunciate their structures. There is no base case with data; inductive reasoning elicits knowledge. Through this process data find form, and sometimes also metaphor or narrative. We view such data naturalism, structuralism, or, with large datasets representing phenomena not viewable, as data-driven design.

Reas et al. [222] argue that the designer must start not with the data but with the empirical question asked by the researcher, then work back to data. Reas considers the nature of the data to be obtained, finds data to fit the question and parses them to provide a structural fit for their

meaning, then orders them into categories and filters out all but the data of interest. This approach maintains the role of the scientist in producing theory, illustrating, testing and deducing. It also offers an opportunity for metaphor, design variation and the recognition of multiple interpretations of the same data by different disciplines. Both approaches need comparative testing. Each of these approaches requires the adaptation of participatory design approaches in which the user is engaged in the development of the visualization. We will test methods for this engagement.

4.3.3. Developing online health risk assessment systems for preventative healthcare

Preventative health-care is a core component of a sustainable public health-care system. Our goal here is to develop predictive models for assessing risks of chronic diseases. Together with our partners (such as Manifold), we will collect and analyze health-related data (including socio-demographic and psychographic data), and build an online predictive tool to provide information and knowledge on health management of chronic diseases. Our tool allows:

1) Consumer-centric and data-driven health assessment and self-management;
2) Interactive and dynamic understanding and management of health and wellness;
3) Constant updates to reflect new knowledge discovered from data.

Various data mining techniques, such as classification, regression, clustering and correlation analysis, will be used in finding the relationships among data

5 Interactive content analytics

We bring together approaches from digital humanities, visual analytics, HCI, design thinking and practice and computational linguistics to provide new insights into the creation, exploration and analysis of online data. Our perspective on interactive content analytics embraces topics ranging from curating online collections to data mining large cultural data sets integrating both digitized and digital materials. The methodologies from the traditional humanities disciplines (such as history, philosophy, linguistics, literature, art, archaeology, music, and cultural studies) and social sciences are combined with tools provided by computing (such as data visualization, information retrieval, data mining, statistics, computational analysis) and digital publishing.

5.1 Problem

Quantities of data are continually produced through the creation of online resources and through the wide range of different interactions that users have with various types of online data. Quantities of structured data have been growing for many decades such as email, Web analytics, or customer feedback. There is a recent growth of unstructured data and an explosion of data through social media that demand text analysis, and collaborative cloud data intelligence well beyond existing CRM (customer-relationship management) systems. The amount of these data is growing exponentially in Canada and at the global level.

Consumers and organizations are increasingly interested in tools and techniques that allow them to manage and visualize their own data collections, online paths and social networks. These tools and their associated analytic approaches have already demonstrated value, as illustrated with the following examples.

- Media industries (including companies such as *the Globe and Mail* or gaming companies such as StatDrive and the advertising agencies and data analytics companies such as Infersystems that support them) use tools to understand, analyze and support the dramatic and disintermediating shift from print or linear media to multimedia formats with intensive user interaction. They need to build sophisticated databases,

repurpose content, track digital rights, analyze user responses, often in real-time, and provide bespoke experiences and tools for subscribers across different screen platforms.

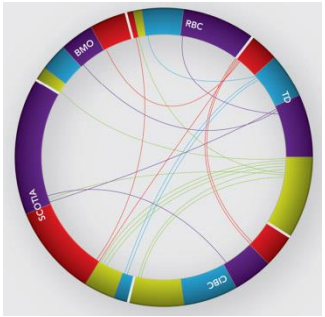
- Learners, employers and educational institutions require online learning tools (such as those created by Desire2Learn) for training, learning, knowledge making, dispersal and collection, text/data mining for knowledge management purposes, digital rights management, data security as well as data collection, and interactive analysis.
- Companies that provide search technologies, open innovation platforms and open source tools contribute to the growing interest in open data and tools to manage open data. They are increasingly looking towards visual tools to facilitate online experience.
- Software standards such as WebGL and Canvas2D provide developers and their users with a platform. The current two-dimensional tools facilitate content for different applications such as visual analytics (Processing.js), and online learning and data collaboration (Desire2Learn). Given the need to represent, navigate and analyze multivariate data sets online, in future we will see the emergence of 3D/Stereoscopic Graphics on the Web (which will require a standard to enable 3D visualization and run simulations on the Web).

5.2 Objectives

1. An open framework for integrating and analyzing online data
2. Design and Evaluation of interactive integrated applications and interfaces (including natural language processing, design, and visual analytics)

5.3 Background and methodology

Our overall approach to interactive content analytics involves the integration of computer science, engineering, and digital humanities research and development methods. Computer science and engineering provide methodologies to build and test technologies through the application of the scientific method and rigorous testing. Digital humanities research seeks an understanding of the ways that humans interact with technologies, applied to fields such as linguistics, cultural studies, communications theory, education, visual art and design, to develop interfaces that facilitate human interaction with data. Digital humanities study the use of online text, visual language and sound, providing both analysis and new tools. Digital humanities facilitate knowledge making, dispersal and collection while always seeking to keep the human "in the loop."



We will adopt a project-based approach with projects being associated with the two objectives listed above. The projects will draw upon research in machine learning, text analytics design, human computer interaction, visualization, visual analytics and digital humanities as elaborated in the research plan below. Project 1 will involve fundamental tool building. Our tools will provide a platform for the project 2 (*Applications Design, Usability and evaluation studies in test bed environments)* component of Interactive Content Analytics. We will equally provide tools as a resource for other projects within the broader initiative, in particular the projects on "Sentiment and emotion detection" and "Adaptive information extraction and topic/event detection and tracking" within Health Informatics.

4.4.1. Fundamental tool building

We will build upon open source frameworks/workbenches for data mining, text analysis and visualization. A natural starting-point would be the Unstructured Information Management Architecture (UIMA) framework, which was first developed by IBM and became open source in 2005 [223]. As noted at http://uima-framework.sourceforge.net/

The Unstructured Information Management Architecture (UIMA) framework is an open, industrial-strength, scalable and extensible platform for building analytic applications or search solutions that process text or other unstructured information to find the latent meaning, relationships and relevant facts buried within. It enables developers to build analytic modules and to compose analytic applications from multiple analytic providers, encouraging collaboration and facilitating value extraction for unstructured information.

Use of a common (open-source) framework will make it easier for researchers to collaborate with different groups working on different components in parallel. Furthermore, building on a well-established framework will facilitate integration of text analytics algorithms, machine learning algorithms, and the development of applications.

Given the large amount of textual (unstructured) data used by online services and applications, it will be important to have a principled means to incorporate and evaluate different text processing algorithms that can provide a deeper understanding of the natural (human) language contained within the various on-line resources, drawing not only on computational techniques but also linguistic knowledge. For fundamental natural language processing tool development, the open source NLTK environment will also be used [224]. Aside from NLTK supporting training of students from a wide range of different backgrounds with its associated course material, it is also widely used in the development of prototypes, and work has already been done on an interface between NLTK and UIMA (http://code.google.com/p/uima-nltk/). As noted at http://nltk.org/,

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning. Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike.

We will contribute to the next generation development of WebGL [225] and related software, which will provide developers and their users with a platform. WebGL is a JavaScript API for rendering interactive 2D and 3D graphics within any compatible Web browser without the use of plug-ins. WebGL is integrated completely into all web standards of the browser, and allows GPU accelerated usage of physics and image processing effects as part of the Web page. 2D tools facilitate content for different applications such as visual analytics. Canvas2D [226] is part of HTML5 which can be used to draw graphics using scripting. Processing.js [227] is an open source programming language and environment that enables the creation of images, animations, and interactions. Our significant contribution will be to extend the WebGL platform into 3D online graphics capabilities.

Our approach to processing data and unstructured (textual) information and providing interactive visual interfaces will draw on developments from the field of visual analytics. This also addresses the growth of big data, a paradigm shift in which instruments and systems produce large quantities of data that are meaningless without extraction and analysis. Visual analytics is a multidisciplinary practice that includes scientific and social science investigation, cognitive science, and visual and design practices and knowledge. Visual analytics supports a mix of cognitive and perceptual reasoning. Whitelaw [228] points out that data become information only when they are placed into an interpretative context. In order to analyze data, they must first be pre-processed, transformed, cleaned, selected and integrated. Users need tools for data-driven design and analysis (developing hypotheses from data) or to illustrate and test their theories [229]. Hence visual analytics provides automated analysis methods or

visualization methods in order to reveal patterns. The structured data are then visualized. Visualizations are designed with users in mind.

Data visualization (information and scientific) images create a bridge between the empirical world and the viewer. Data visualizations reveal patterns of data and unexpected relationships, hence evoking interpretation. Cleveland [230] states that data visualization makes the invisible visible, "providing a front line of attack, revealing intricate structures…we discover unimagined effects, and we challenge imagined ones" [230, p. 1] Effective representations can allow the monitoring of simultaneous data sources. Visual analytics can support comparative analysis, the recognition of anomalies or problems as these emerge, and can support prediction.

Ware and Bobrow [231] express the visual analytics proposal well: understanding data requires an interaction between human cognition, computer memory and its related algorithms, and the physical actions of the user. The results of successful visual analytics are: 1) the ability to make available and visible diverse data in order to support decision-making processes; 2) the ability to communicate data and its analysis to diverse users or receptors; and 3) effective use of human visual perception in order to navigate large data quantities, compressing it into screen space.

4.4.2 Applications Design, Usability and evaluation studies in test bed environments

We bring a unique approach to interactive content analytics, by integrating engineering, design and usability testing and following with evaluation studies. This project will bring together research capacities in design research, HCI and digital humanities research, building secure applications on top of the fundamental tool framework that Popowich is leading. New applications must respond to the increasing need for individuals and industry to manage their data effectively and securely. There is an increased demand for accessible visual analytics, forecasting tools, dash boards and multi-screen and mobile applications that will mainstream data analytics. The principles of visual analytics discussed earlier will be applied and compared across diverse contexts and platforms allowing for the expression and refinement of methodologies, interfaces and tools. Interfaces for use in the field can take advantage of Natural User Interfaces (NUIs) and wearable technologies.

The applications will be driven by interactions with industry partners, participating from key application areas that include:

- Healthcare - Hospitals requiring secure analytics of clinician/patient interaction;
- Education and Training - Educational software companies requiring data analysis, translation tools, evaluation tools, design, and HCI research and visual analytics or simulation capabilities;
- Technology/Media/Telecommunications (TMT) Industries - TMT companies requiring database search capacities, customer analytics and customer-facing visual analytics tools;
- Financial Services - Financial institutions developing secure customer-facing analytics for mobile devices;
- Wholesalers - require sophisticated tools for e-commerce, supply train management, data tracking, ownership transfer, fraud prevention in the international marketplace;
- Software companies requiring sophisticated tools for data analysis and visual representation on the Web.

As well as providing a test bed for these specific verticals, we will provide support for other project areas that require visual analytics, such as project 6 on "Data Visualization Strategies and methodologies for social media content and free form texts" within the Health Informatics theme, and "Network analysis/visualization/integration" within the Computational Biology Theme. Support will also be provided to industry partners as appropriate, and to new projects that they bring into the network in the later years of the strategic network operations.

Greenberg and Buxton [232] emphasize the importance of integrating different methodologies such as those of design and humanities to facilitate paradigmatic breakthroughs characterized by a 'vision' of a new interface, or the ability to foresee "the creation of a new

culture of use" [232, p. 116]. Bardzell and Bardzell [233] agree that researching, designing and testing interactive tools require cross-disciplinary engagement with disciplinary knowledge derived from digital humanities in order to better understand the impacts of technologies once these have been designed. Design methods will include an engagement with visual aesthetics, linguistic analysis and cognitive assessment.

Design research methods will incorporate designers, engineers, HCI researchers and end users into the technology-creation process. An agile software development process [234-236] will lead to incrementally re-engineered prototypes. Design typically includes four iterative stages [237]:

- *Analysis*. This phase aims at understanding the context and collecting information in order to define a problem as well as the collection of data regarding the users and the identification of datasets;
- *Design*. In this phase, the team brainstorms ideas and draws sketches based on their prior analysis;
- *Prototype*. This phase consists of selecting the most effective sketches, refining and implementing them as low fidelity prototypes. Engineering prototyping and interface sketching occur in an integrated and iterative fashion; and
- *Evaluation*. The final phase consists of presenting the prototypes to the end users and usability testers in order to identify strengths and weaknesses of the concepts. Fully-engineered functional prototypes are then built with interfaces refined throughout the process. These are used, and their usability is tested.

We will apply 'agile usability processes' [236] to continually test interfaces. We will exercise three modalities of usability testing: 1) whether the technology could be improved [238]; 2) whether the aesthetics could be improved [238-240]; and 3) whether the tool was useful [232]. A common practice is to run heuristic evaluations with usability experts on a series of goals that the design aims at targeting as well as working closely with end users.

Hence our approach will centre not only on usability but equally usefulness: does the imagined new tool or application bring value to the end users? Can they imagine new ways that the applications we tackle can be applied? Researchers will study the impacts of the new technologies in their application environments, and conduct longitudinal research to evaluate usefulness and identify any transformative impacts for further cycles of technology design and development.

6 Training

Worldwide IT spending is set to increase by 3.8% over the next year, reaching a total of $3.7 trillion, with the excitement surrounding big data driving much of that growth, according to the latest predictions from Gartner.[11] Gartner's Peter Sondergaard asserts that by 2015, big data would help to create around 4.4 million jobs worldwide, with 1.9 million of them being in the US. Student marketability and employment prospects in data analytics appear bright.

Training of students and post-docs is an important component of any research endeavor and we aim to do so in a novel way. Students will be trained on the most advanced technologies at participating universities and hospital/medical research institutes to apply these technologies to information-rich problems, for example, in healthcare. We adopt a rotation mechanism, which is commonly used in medical training. Each student will spend time at a partner university and/or company for research training. "Rotation" training exposes students to more projects/ environments, and rotation out of academia into industry and government provides exchange of research ideas and results. Most private-sector partners have expressed sincere interest to take

[11] Gartner: Big Data = Big Job Prospects, http://servicesangle.com/blog/2012/10/25/gartner-big-data-big-job-prospects/

on interns for work on projects and becoming involved in training of HQP. The key participants identified for each private-sector partner will provide the link to this aspect of student mobility and training.

In addition to, and similar to, internship training, in research training the HQP will develop technical and professional skills. We will provide training in:

1. Research skills. Training in research skills is a key element in researcher development. We will train students to obtain knowledge of recent advances in intelligent information analysis through directed literature survey, seminars, invited talks, researcher exchange, and by sending trainees to conferences. We will guide trainees to develop their own original, independent and critical thinking ability through the coordinated effort of supervision and project work. Trainees will be asked to analyze critically the work of others by providing them with opportunities to review papers (thanks to our positions as journal editors and conference committee chairs). We will provide them with opportunities to develop abilities to recognize and validate research problems through industry collaborator interactions. Trainees are expected to make original contributions to information analysis.
2. Communication skills. We will take advantage of Mitacs Inc. support to organize regular research seminars and teleconferences at participating institutions and our industry partners for the trainees to present their work. Trainees will be encouraged to present conference papers and communicate with our industry collaborators regularly. Students will develop their writing skills by writing progress reports, research papers, project documents and a thesis.
3. Networking and team-working skills. Again, we take advantage of Mitacs Inc. support, and our "rotation" training method allows students to work in different research/work environments. Thus trainees will develop cooperative networks and relationships. In rotations, they will work in teams.
4. Leadership skills. We will recruit trainees at all levels from undergraduates to postdocs. For each research project, we will place trainees in teams so that senior trainees (PhD students or postdocs) supervise junior trainees (undergraduate or MSc students) on research tasks. This arrangement will teach them to supervise research, and develop their leadership skills. York's national lead in knowledge mobilization will deliver IP, commercialization and knowledge mobilization capacity building sessions for trainees.

7 Partnerships and knowledge mobilization

Information analytics makes an impact on productivity through three main channels:

- *Efficiencies* are realized through rapid *technological progress in the production of information analysis goods and services in producing industries*. Thus, information analysis is a driver of productivity growth for the entire economy. Efficiency gains in the deployment of information analysis results are also reflected in the fast price declines of affected products.
- *Investments in information analytics* provide more capital for workers, which raise their productivity.
- *Greater use of information analysis* in all sectors in the economy helps firms to increase their efficiency.

These three effects do not occur simultaneously. Investments translate into efficiency gains only after a time lag, as the results of information analysis are used to reorganize the production process. Therefore, the impact of information analysis on the wider economy can be expected in two time-frames: in the *short term*, reductions in the relative prices of products increase investment; in the *longer term*, as the new technologies are adopted throughout the economy, new goods are developed, and new modes of business organization come into use.

Information analytics is part of the ICT sector, but affects and enables all other sectors. In Canada, the ICT industry is strongly concentrated in *Ontario*. About 50% of Canadian ICT GDP, revenue and employment are in Ontario, home to more than 16,000 ICT firms including country headquarters of many of the world's largest industry leaders.

Involvement of the partner organizations and research priorities of the participating partners:

Research priorities of the participating partners were incorporated through many discussions between co-applicants and partners, and by requesting significant initial project descriptions from the partners, as reflected in their letters of support, which describe the project, situate the research in a larger context within the four research themes, and include references.

In addition, partners were requested in their letters to confirm their support for the Network in cash and in-kind contributions; what they anticipated their project outcomes and timelines would be; the overall benefits to their organization and the Canadian economy; and how they expected to integrate the research results into their operations.

Linkages existent and planned:

We have interacted with over 50 potential private-sector partners, selecting a dozen of them for participation. This interaction is ongoing, and we anticipate that it will continue throughout the life of the proposed research.

Communications involved constant e-mail contact, telephone conversations, the two planning workshops, which were well attended by the co-applicants, and, at times, individual meetings of those in close proximity.

At least five private-sector partners were exposed to the research via recent NSERC ENGAGE grants (six month initial grants to solve a small technical problem. Exposure to that program was a definite plus for several partners.

For private-sector partners in Ontario, the Ontario Centres of Excellence has discussed with us and promoted the use of their OCE Technical Problem Solving Projects grants for short-term exploitation and their Collaborative Research program for intermediate, longer-term bigger projects.

Mechanisms proposed for internal communications between the participants:

Part of the role of the PI is to maintain close liaison with partners. In addition, communications is planned with several specific primary activities: 1) production of a bi-monthly four page newsletter featuring a research project write-up in plain English (1.5 pps), news about the research alliance, calendar of events, etc.; 2) constant liaison with partners, existing and potential; and 3) liaison with public-sector partners and government.

7.1 Mechanisms for knowledge and/or technology transfer

The collaborating partners have varying intellectual property policies. Because the collaborating partners, along with their respective researchers, resolve to develop and commercialize the intellectual property resulting from the research, these organizations are working together under inter-institutional agreements and the appropriate research agreement in order to ensure that the pathways and goals for commercialization are shared. A three-tiered modified Canadian Network Centres of Excellence model of Network intellectual property management is employed amongst co-applicants and partners of this researh. In order to manage this process effectively, a Knowledge Mobilization Advisory Council (KMAC) will be established and work in close collaboration with the participating university technology transfer offices (TTOs) and co-applicants. The KMAC will be comprised of representatives from the partners, commercialization experts, and government representatives.

York's knowledge mobilization unit (KMb) provides services and funding for faculty, graduate students, and community organizations seeking to maximize the impact of academic research and expertise on public policy, social programming, and professional practice. York University's Knowledge Mobilization Unit is a national leader in connecting research and

people for social innovation. The Unit has supported researchers since 2006, and has helped facilitate 260 collaborations resulting in over 75 projects. York is also part of Research Impact, which is Canada's largest knowledge mobilization network. The KMb Unit will support this research by contributing services including (1) training on how to write clear language output summaries (lay language); (2) social media training and support; and (3) advice and consultation for the KMAC from the KM experts at York U.

The present application proposes to create a multi-disciplinary research consortium[12] that enables the development of cross-functional projects and the creation of unique intellectual property. We have discussed individual projects we propose. In addition to the individual projects and the resultant project-based intellectual property, this research has the potential to enable the development of a service-based research consortium that provides multi-disciplinary research services to external parties. The value proposition of this project is the cross-functional nature of the research teams through which third parties will be able to leverage knowledge and expertise that would not otherwise be available.

New project-based inventions or technologies with commercial potential are disclosed to the KMAC in advance of public dissemination. In the event that patent protection will be sought, then the participating organizations will agree on external patent counsel for filing and prosecuting patents. The KMAC subscribes to a 'just-in-time' patent strategy, whereby the filing of a patent application is delayed until such time that the intellectual property is to be disclosed, thus allowing researchers to gather additional data and strengthen the basis on which the patent application will be filed. In addition, the parties will agree not to file patent applications other than a US provisional & International PCT, unless a commercial party has licensed the rights for the technology. The Network agreement will minimize any adverse impact on publication or thesis work.

The KMAC will engage the intellectual property protection process to determine whether patent protection or copyright registration is deemed a most appropriate commercialization path.

Collaborating partners are responsible for facilitating the transfer of knowledge developed by researchers at their respective institutions, into products and programs that result in social and economic benefit. This is evident through the creation of many startup companies based on IP developed at York: for example, Total Synthesis Inc. (combinational chemistry); Thoth Technolo-gies, Inc. (space engineering); Geo-Tango Inc. (GIS technology, sold to Microsoft); Dalton Pharma Services (pharmaceutical manufacturing); and SiRaCoR Inc. (cancer biomarker diagnostics).

Upon disclosure of an invention to the KM, participating universities and partners will be the principal funders of intellectual property protection. The participating organizations will supplement patent expenses that are above the allowable patent expenses provided under this Network, and may seek help from organizations engaged in such commercialization, e.g., Innovation York, MaRS, Venture Lab and so on. For intellectual property jointly-owned between multiple parties, the costs of intellectual property protection shall be borne by each party in accordance with the percentage of ownership.

---

12 The multi-disciplinary research service consortium would adopt a similar model to existing contract research organizations (CROs). The MDRSC would provide research services spanning across key research areas: applications-oriented information analytics, frontiers of information analysis, and information management as well as knowledge mobilization. Currently contract research services can be obtained through traditional CROs, which focus on pharmacology and biotechnology, or through universities, which have not typically had a culture that supports the milestone-based research required in contract research.

Our KM strategy involves:

- partners – public- and private-sector partners support of the Network, and their reasons for joining as indicated in their letters;
- expertise – led by the KMAC and aided by the university TTOs;
- audiences – projects are targeted to meet partner's requests and co-applicant expertise;
- goals – goals align with partner's projects outcomes (research and development) and the continuous incremental transfer of technology from the projects;
- methods – both research methods and KM tools and activities are outlined in proposal;
- impact – measured by new IP and patents, HQP trained into successful professionals, when wealth is generated for Canada.

8 Examples of novelty, difference, improvement, and significance

Big data research should be new, different, better and significant. Following are some examples to illustrate this point

One example of *novel research* is a project to find best evidence for evidence based medicine and evidence based best practice recommendations. This research will find support from multiple evidence sources (doctors' written notes, clinical trial data, published reports, and so on) using: (1) state-of-the art techniques of adaptive information extraction; and (2) validation methods based on multiply sectioned Bayesian networks to provide a probabilistic interpretation of aggregated evidence.

An example of *research that makes a difference* is the project for evaluating technologies with an interactive content analytics (ICA) methodology. Thus we explore and experiment with media and art technologies as they create possibilities and venues for inquiry into classic, yet still pressing metaphysical, questions regarding subjectivity, and the relation of mind to body, the physical and material to the digital. Moreover, ICA offers a broad range of theoretical approaches and grounded concepts with which we can excavate technological assumptions and expectation, and pursue, critically, experimental technology practices in both the studio environment and the real-world environment, that, importantly, merge the analogue with the digital.

An example of *better research in training* combines soft-skills training via the partnership with Mitacs (*Networking and Technical Training* and the Mitacs *Step* program) with student mobility opportunities (students will spend time in more than one academic laboratory) and internships with private- and public-sector partners for the benefit of students. Lack of emphasis on these skills has led to a 'soft skills' deficit in the workforce.

A research example of a *significant project* is analyzing and controlling the spread of an infectious disease that involves characteristics of the agent, the host and the environment in which transmissions take place. The purpose of modeling infectious diseases, in relation to public health, is to evaluate the agent-host-environment interface and efforts to alter the interface through intervention to our advantage, be they preventative or therapeutic in nature.

Mathematical epidemiology has a long history. In recent years more complex and biologically relevant models have been developed, and these models and their analysis have become important for influencing the design of control programs. Some of these models have been developed for emerging and re-emerging diseases; some include new medical treatments; some involve evolutionary aspects; and some consider new patterns of social behavior and travel. Computer simulations that use demographic and disease incidence data often complement theoretical analysis of the models. Nevertheless, there remain many challenging problems in the understanding of disease transmission and spread, and an interdisciplinary approach is required.

Acknowledgments. The support of Canada's Natural Sciences and Engineering Research Council (NSERC) is gratefully acknowledged for their encouragement and funding. We also thank the theme contributors, and the many academic researchers and private and public sector organizations that have contributed to the ideas contained herein.

References

1. White, Tom (2012). *Hadoop: The Definitive Guide*. O'Reilly Media. p. 3. ISBN 978-1-4493-3877-0.
2. Kenneth Cukier "Data, data everywhere". (2010) *The Economist*. p 30.
3. Amanda Jones & Ben Kerschberg, (2012) "What Technology-Assisted Electronic Discovery Teaches Us About The Role Of Humans In Technology – Re-Humanizing Technology-Assisted Review", Forbes.
4. Watters, Audrey (2010). "The Age of Exabytes: Tools and Approaches for Managing Big Data". Hewlett-Packard Development Company.
5. Editorial, "Community cleverness required". *Nature* 455 (7209): 1. 4 September 2008. doi:10.1038/455001a.
6. Reichman, O.J., Jones, M.B., Schildhauer, M.P. (2011). "Challenges and Opportunities of Open Data in Ecology". *Science* 331 (6018): 703–5. doi:10.1126/science.1197962.
7. M.N. Garofalakis, J. Gehrke, R. Rastogi, (2002). Querying and mining data streams: you only get one look a tutorial, in: Proc. 2002 ACM SIGMOD Int. Conf. On Management-of Data, SIGMOD'02, Madison, WI, p. 635.
8. H.J. Woo, W.S. Lee, (2007). EstMax: tracing maximal frequent itemsets over online data streams, in: Proc. ICDM, 2007, pp. 709–714.
9. Y. Zhu, D. Shasha, Statstream: statistical monitoring of thousands of data streams in real time, in: Proc. VLDB, pp. 358–69.
10. H. Li, S. Lee, (2009). Mining frequent itemsets over data streams using efficient window sliding techniques, presented at Expert Syst. Appl., pp. 1466–1477.
11. Farzanyar, Z., Kangavari, M., and Cercone, N. (2012) Max-FISM: Mining (Recently) Maximal Frequent Itemsets over Data Streams using the Sliding Window Model, Computers & Mathematics with Applications 64, Elsevier, 1706-1718.
12. Farzanyar, Z., Kangavari, M., and Cercone, N. (2012, submitted) P2P-FISM: mining (recently) frequent itemsets from distributed data streams over P2P network, Information Processing Letters, Elsevier.
13. Anand Rajaraman, Jeffrey David Ullman (2011) Mining of Massive Datasets, Cambridge University Press, ISBN:9781107015357, 326 pages.

Large-Scale Data Analytics (and Cloud Computing)

14. B. Gassman and R. Knox, "'Cloud analytics' means many different kinds of opportunity," Gartner Research, Stamford, CT, Tech. Rep. 1386527, 2010.
15. A. Cuzzocrea, I. Song and K. C. Davis, "Analytics over large-scale multidimensional data: The big data revolution!" in *Proceedings of the ACM 14th International Workshop on Data Warehousing and OLAP,* Glasgow, Scotland, UK, 2011, 101-104.
16. S. Frischbier and I. Petrov, "Aspects of data-intensive cloud computing," in Anonymous 2010, 57-77.
17. H. Vashishtha, M. Smit and E. Stroulia, "Moving text analysis tools to the cloud," in *IEEE Congress on Services,* 2010, 107-114.
18. K. Doka, D. Tsoumakos and N. Koziris, "Efficient updates for a shared nothing analytics platform," in *Proceedings of the 2010 Workshop on Massive Data Analytics on the Cloud,* Raleigh, North Carolina, 2010, 7:1-7:6.
19. I. Konstantinou, E. Angelou, D. Tsoumakos and N. Koziris, "Distributed indexing of web scale datasets for the cloud," in *Proceedings of the 2010 Workshop on Massive Data Analytics on the Cloud,* Raleigh, North Carolina, 2010, 1:1-1:6.
20. M. Shmueli-Scheuer, H. Roitman, D. Carmel, Y. Mass and D. Konopnicki, "Extracting user profiles from large scale data," in *Proceedings of the 2010 Workshop on Massive Data Analytics on the Cloud,* Raleigh, North Carolina, 2010, 4:1-4:6.
21. K. S. Beyer, V. Ercegovac, R. Krishnamurthy, S. Raghavan, J. Rao, F. Reiss, E. J. Shekita, D. E. Simmen, S. Tata, S. Vaithyanathan and H. Zhu, "Towards a Scalable Enterprise Content Analytics Platform," *IEEE Data Eng.Bull.,* vol. 32, 28-35, 2009.
22. G. Lee, B. Chun, and H. Katz. Heterogeneity-aware resource allocation and scheduling in the cloud.

In Proceedings of the 3rd USENIX conference on Hot topics in cloud computing (HotCloud'11). USENIX Association, Berkeley, CA, USA. 2011.
23. M. Shtern, B. Simmons, M. Smit, and M. Litoiu. Toward an Ecosystem for Precision Sharing of Segmented Big Data. Submitted to the IEEE Conference on Cloud Computing, 2013.
24. H. Li and J. Jeng, "CCMarketplace: A marketplace model for a hybrid cloud," in *Proceedings of the 2010 Conference of the Center for Advanced Studies on Collaborative Research,* Toronto, Ontario, Canada, 2010, pp. 174-183.
25. M. Hajjat, X. Sun, Y. E. Sung, D. Maltz, S. Rao, K. Sripanidkulchai and M. Tawarmalani, "Cloudward bound: planning for beneficial migration of enterprise applications to the cloud," *SIGCOMM Comput.Commun.Rev.,* vol. 41, pp. 243-254, August, 2010.
26. C. Baun and M. Kunze, "The KOALA cloud management service: A modern approach for cloud infrastructure management," in *Proceedings of the First International Workshop on Cloud Computing Platforms,* Salzburg, Austria, 2011, pp. 1:1-1:6.
27. K. Zhang, X. Zhou, Y. Chen, X. Wang and Y. Ruan, "Sedic: Privacy-aware data intensive computing on hybrid clouds," in *Proceedings of the 18th ACM Conference on Computer and Communications Security,* Chicago, Illinois, USA, 2011, pp. 515-526.
28. L. Abraham, M. A. Murphy, M. Fenn and S. Goasguen, "Self-provisioned hybrid clouds," in *Proceeding of the 7th International Conference on Autonomic Computing,* Washington, DC, USA, 2010, pp. 161-168.
29. Hewlett-Packard Company. Enterprise cloud services - compute | HP services. *2011(Nov 15),* Available: http://www.hp.com/enterprise/cloud.
30. Fujitsu. FACT SHEET: Fujitsu hybrid cloud services for windows azure. *2011(Nov 16),* 2011. Available: http://solutions.us.fujitsu.com/pdf/services/Services-Cloud-Hybrid-Windows-Azure-factsheet.pdf.
31. IBM - A highly scalable, low-touch private cloud which offers near zero downtime, rapid image deployment and automated recovery across heterogeneous platforms - IBM SmartCloud provisioning software. 2011 Available: http://www-01.ibm.com/software/tivoli/products/smartcloud-provisioning/.
32. IBM. IBM infrastructure as a service. 2011.Available: http://www-935.ibm.com/services/us/en/cloud-enterprise/index.html.
33. Apache Software Foundation. Deltacloud | many clouds. one API. no problem. *2011(Nov 14),* 2011. Available: http://incubator.apache.org/deltacloud/.
34. Apache Software Foundation. Apache libcloud | a unified interface to the cloud. *2011(Nov 14),* Available: http://libcloud.apache.org/.
35. jClouds Inc. jClouds. *2011(Nov 14),* Available: http://www.jclouds.org/.
36. D. Bernstein, E. Ludvigson, K. Sankar, S. Diamond, and M. Morrow, “Blueprint for the intercloud - protocols and formats for cloud computing interoperability,” in Proceedings of the 2009 Fourth International Conference on Internet and Web Applications and Services. Washington, DC, USA: IEEE Computer Society, 2009, pp. 328–336.
37. R. Buyya, R. Ranjan, and R. N. Calheiros, “Intercloud: Utility-oriented federation of cloud computing environments for scaling of application services,” in ICA3PP (1), 2010, pp. 13–31.
38. A. Khajeh-Hosseini, I. Sommerville, J. Bogaerts, and P. B. Teregowda, “Decision support tools for cloud migration in the enterprise.” in Cloud Computing (CLOUD), 2011 IEEE International Conference on, L. Liu and M. Parashar, Eds. IEEE, 2011, pp. 541–548.
39. S.-M. Han, M. M. Hassan, C.-W. Yoon, and E.-N. Huh, “Efficient service recommendation system for cloud computing market,” in Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, ser. ICIS ’09. New York, NY, USA: ACM, 2009, pp. 839–845.
40. Mathisen, Eystein , “*Security Challenges and Solutions in Cloud Computing,*” in the proceedings of 5th IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2011), 31 May -3 June 2011, Daejeon, Korea.
41. Tang, L. and Li, T., “Discovering Lag Intervals for Temporal Dependencies,” 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-2012), Beijing, China.
42. Zhong, L. and Wu, J., “On Modeling Data Mining with Time Granularity,” 2009 International Conference on Computational Intelligence and Natural Computing (CINC '09). Wuhan, China, June 2009.
43. Viappiani, P., Zilles, S., Hamilton, H.J., and Boutilier, C., “Learning Complex Concepts using Crowdsourcing: A Bayesian Approach.” In *Second International Conference on Algorithmic Decision Theory (ADT 2011)* Rutgers University, October, 2011.
44. L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey. ACM Comput.

Surv. 38, 3, Article 9. 2006.
45. Li, L.J. and Zhang, M., "The Strategy of Mining Association Rule Based on Cloud Computing," International Conference on Business Computing and Global Informatization (BCGIN) 2011, July 2011, 475-478.
46. Hilderman, R.J., Hamilton, H.J., and Cercone, N. (1999) Data Mining in Large Databases Using Domain Generalization Graphs, J. of Intelligent Information Systems 13, 195-234.
47. E. Mancini, U. Villano, M. Rak, and R. Torella. A simulation-based framework for autonomic web services. In 11th International Conference on Parallel and Distributed Systems, 2005. pp. 433–437.
48. W. Iqbal, M. Dailey, and D. Carrera, "SLA-driven adaptive resource management for web applications on a heterogeneous compute cloud," Cloud Computing, pp. 243–253, 2009.
49. B. Simmons, M. Litoiu, D. Ionescu, and G. Iszlai, "Towards a cloud optimization architecture using strategy-trees," in Proceedings of The 9th International Information and Telecommunication Technologies Symposium (I2TS 2010), 13-15, December, Rio de Janeiro, Brazil, 2010.
50. P. Pawluk, B. Simmons, M. Smit, M. Litoiu, and S. Mankovski, "Introducing STRATOS: A cloud broker service," in IEEE 5th International Conference on Cloud Computing (CLOUD), pp. 891–898, 2012.
51. X. Lin, "*Survey on Cloud Based Mobile Security and A New Framework for Improvement,*" in the proceeding of the IEEE International Conference on Information and Automation, Shenzhen, China, June 2011.
52. Hong, J. W., "*Monitoring and detecting abnormal behavior in mobile cloud infrastructure,*" published in 2012 IEEE Network Operations and Management Symposium, pp. 1303-1310.
53. I. Roy, S. T. Setty, A. Kilzer, V. Shmatikov, and E. Witchel, "Airavat: Security and privacy for MapReduce," in Proceedings of the 7th USENIX conference on Networked systems design and implementation. USENIX Association, 2010.
54. S Saklikar. "Embedding Security and Trust Primitives in Map Reduce". Presented at RSA Conference China, Trusted Computing session. August 29, 2012.

Computational Biology

55. Gehlenborg, N., S. I. O'Donoghue, et al. (2010). "Visualization of omics data for systems biology." Nat Methods 7(3 Suppl): S56-68.
56. Pavlopoulos, G. A., A. L. Wegener, et al. (2008). "A survey of visualization tools for biological network analysis." BioData Min 1: 12.
57. Suderman, M. and M. Hallett (2007). "Tools for visually exploring biological networks." Bioinformatics 23(20): 2651-2659.
58. Merico, D., D. Gfeller, et al. (2009). "How to visually interpret biological data using networks." Nat Biotechnol 27(10): 921-924.
59. Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet 25(1): 25-29.
60. Hermjakob, H., L. Montecchi-Palazzi, et al. (2004). "The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data." Nat Biotechnol 22(2): 177-183.
61. Jeong, E., M. Nagasaki, et al. (2007). "Conversion from BioPAX to CSO for system dynamics and visualization of biological pathway." Genome Inform 18: 225-236.
62. Cerami, E. G., G. D. Bader, et al. (2006). "cPath: open source software for collecting, storing, and querying biological pathways." BMC Bioinformatics 7: 497.
63. Ein-Dor, L., I. Kela, et al. (2005). "Outcome signature genes in breast cancer: is there a unique set?" Bioinformatics 21(2): 171-178.
64. Boutros, P. C., S. K. Lau, et al. (2009). "Prognostic gene signatures for non-small-cell lung cancer." Proc Natl Acad Sci U S A 106(8): 2824-2828.
65. Ein-Dor, L., O. Zuk, et al. (2006). "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer." Proc Natl Acad Sci U S A 103(15): 5923-5928.
66. Lau, S. K., P. C. Boutros, et al. (2007). "Three-gene prognostic classifier for early-stage non small-cell lung cancer." J Clin Oncol 25(35): 5562-5569.
67. Heifets, A. and I. Jurisica (2012). "SCRIPDB: a portal for easy access to syntheses, chemicals and reactions in patents." Nucleic acids research 40(Database issue): D428-433.
68. Fortney, K., W. Xie, et al. (2013). "NetwoRx: connecting drugs to networks and phenotypes in Saccharomyces cerevisiae." Nucleic Acids Research 41(D1): D720-727.
69. Chu, L. H. and B. S. Chen (2008). "Construction of a cancer-perturbed protein-protein interaction network for discovery of apoptosis drug targets." BMC Syst Biol 2: 56.

70. Tsao, M. S., S. Aviel-Ronen, et al. (2007). "Prognostic and predictive importance of p53 and RAS for adjuvant chemotherapy in non small-cell lung cancer." J Clin Oncol 25(33): 5240-5247.
71. Foekens, J. A., D. Atkins, et al. (2006). "Multicenter validation of a gene expression-based prognostic signature in lymph node-negative primary breast cancer." J Clin Oncol 24(11): 1665-1671.
72. Buyse, M., S. Loi, et al. (2006). "Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer." J Natl Cancer Inst 98(17): 1183-1192.
73. Spentzos, D., D. A. Levine, et al. (2004). "Gene expression signature with independent prognostic significance in epithelial ovarian cancer." J Clin Oncol 22(23): 4700-4710.
74. McKee, C. M., D. Xu, et al. (2012). "Protease nexin 1 inhibits hedgehog signaling in prostate adenocarcinoma." The Journal of clinical investigation 122(11): 4025-4036.
75. Reis, P. P., L. Waldron, et al. (2011). "A gene signature in histologically normal surgical margins is predictive of oral carcinoma recurrence." BMC Cancer 11(1): 437.
76. Eppert, K., K. Takenaka, et al. (2011). "Stem cell gene expression programs influence clinical outcome in human leukemia." Nature medicine 17(9): 1086-1093.
77. Zhu, C. Q., K. Ding, et al. (2010). "Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer." J Clin Oncol 28(29): 4417-4424.
78. Cox, B., M. Kotlyar, et al. (2009). "Comparative systems biology of human and mouse as a tool to guide the modeling of human placental pathology." Mol Syst Biol 5: 279.
79. Tomasini, R., K. Tsuchihara, et al. (2008). "TAp73 knockout shows genomic instability with infertility and tumor suppressor functions." Genes Dev 22(19): 2677-2691.
80. Gortzak Gortzak Sodek, K. L., A. I. Evangelou, et al. (2008). "Identification of pathways associated with invasive behavior by ovarian cancer cells using multidimensional protein identification technology (MudPIT)." Mol Biosyst 4(7): 762-773.
81. Gortzak-Uzan, L., A. Ignatchenko, et al. (2008). "A proteome resource of ovarian cancer ascites: integrated proteomic and bioinformatic analyses to identify putative biomarkers." J Proteome Res 7(1): 339-351.
82. Mills, G. B., I. Jurisica, et al. (2009). "Genomic amplicons target vesicle recycling in breast cancer." J Clin Invest 119(8): 2123-2127.
83. Agarwal, R., I. Jurisica, et al. (2009). "The emerging role of the RAB25 small GTPase in cancer." Traffic 10(11): 1561-1568.
84. Brown, K. R. and I. Jurisica (2007). "Unequal evolutionary conservation of human protein interactions in interologous networks." Genome Biol 8(5): R95.
85. Elschenbroich, S., V. Ignatchenko, et al. (2011). "In-depth proteomics of ovarian cancer ascites: combining shotgun proteomics and selected reaction monitoring mass spectrometry." Journal of proteome research 10(5): 2286-2299.
86. Chang, Q., I. Jurisica, et al. (2011). "Hypoxia predicts aggressive growth and spontaneous metastasis formation from orthotopically grown primary xenografts of human pancreatic cancer." Cancer research 71(8): 3110-3120.
87. Jeong, H., S. P. Mason, et al. (2001). "Lethality and centrality in protein networks." Nature 411(6833): 41-42.
88. Hahn, M. W. and A. D. Kern (2005). "Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks." Mol Biol Evol 22(4): 803-806.
89. Maslov, S. and K. Sneppen (2002). "Specificity and stability in topology of protein networks." Science 296(5569): 910-913.
90. Gavin, A. C., M. Bosche, et al. (2002). "Functional organization of the yeast proteome by systematic analysis of protein complexes." Nature 415(6868): 141-147.
91. Wuchty, S. (2006). "Topology and weights in a protein domain interaction network--a novel way to predict protein interactions." BMC Genomics 7: 122.
92. Ravasz, E., A. L. Somera, et al. (2002). "Hierarchical organization of modularity in metabolic networks." Science 297(5586): 1551-1555.
93. Yu, H. and M. Gerstein (2006). "Genomic analysis of the hierarchical structure of regulatory networks." Proc Natl Acad Sci U S A.
94. Han, J. D., N. Bertin, et al. (2004). "Evidence for dynamically organized modularity in the yeast protein-protein interaction network." Nature 430(6995): 88-93.
95. Fortney, K., M. Kotlyar, et al. (2010). "Inferring the functions of longevity genes with modular subnetwork biomarkers of Caenorhabditis elegans aging." Genome Biol 11(2): R13.
96. Milo, R., S. Shen-Orr, et al. (2002). "Network motifs: simple building blocks of complex networks." Science 298: 824-827.
97. Rice, J. J., A. Kershenbaum, et al. (2005). "Lasting impressions: motifs in protein-protein maps may

provide footprints of evolutionary events." Proc Natl Acad Sci U S A 102(9): 3173-3174.
98. Przulj, N., D. G. Corneil, et al. (2004). "Modeling interactome: scale-free or geometric?" Bioinformatics 20(18): 3508-3515.
99. Barrios-Rodiles, M., K. R. Brown, et al. (2005). "High-throughput mapping of a dynamic signaling network in mammalian cells." Science 307(5715): 1621-1625.
100. Hu, W., Z. Feng, et al. (2007). "A single nucleotide polymorphism in the MDM2 gene disrupts the oscillation of p53 and MDM2 levels in cells." Cancer Res 67(6): 2757-2765.
101. Jansen, R., D. Greenbaum, et al. (2002). "Relating whole-genome expression data with protein-protein interactions." Genome Res 12(1): 37-46.
102. Kato, T., Y. Murata, et al. (2006). "Network-based de-noising improves prediction from microarray data." BMC Bioinformatics 7 Suppl 1: S4.
103. Jonsson, P. F. and P. A. Bates (2006). "Global topological features of cancer proteins in the human interactome." Bioinformatics 22(18): 2291-2297.
104. Wachi, S., K. Yoneda, et al. (2005). "Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues." Bioinformatics 21(23): 4205-4208.
105. King, A. D., N. Przulj, et al. (2012). "Protein complex prediction with RNSC." Methods in molecular biology 804: 297-312.
106. Przulj, N., D. G. Corneil, et al. (2006). "Efficient estimation of graphlet frequency distributions in protein-protein interaction networks." Bioinformatics 22(8): 974-980.
107. Ideker, T., O. Ozier, et al. (2002). "Discovering regulatory and signalling circuits in molecular interaction networks." Bioinformatics 18 Suppl 1: S233-240.
108. Chuang, H. Y., E. Lee, et al. (2007). "Network-based classification of breast cancer metastasis." Mol Syst Biol 3: 140.
109. Nacu, S., R. Critchley-Thorne, et al. (2007). "Gene expression network analysis and applications to immunology." Bioinformatics 23(7): 850-858.
110. Hwang, D., J. J. Smith, et al. (2005). "A data integration methodology for systems biology: experimental verification." Proc Natl Acad Sci U S A 102(48): 17302-17307.
111. Przulj, N., D. A. Wigle, et al. (2004). "Functional topology in a network of protein interactions." Bioinformatics 20(3): 340-348.
112. Dong, J., T. Kislinger, et al. (2009). "Lung cancer: developmental networks gone awry?" Cancer Biol Ther 8(4): 312-318.
113. Burley, S.K., Joachimiak, A., Montelione, G.T. & Wilson, I.A. Contributions to the NIH-NIGMS Protein Structure Initiative from the PSI Production Centers. Structure 16, 5-11 (2008).
114. Chazin, W.J. Evolution of the NIGMS Protein Structure Initiative. Structure 16, 12-4 (2008).
115. Edwards, A.M. & Edwards, E.A. A future for the protein structure initiative. Structure 15, 1525-6 (2007).
116. Gerlt, J.A. A Protein Structure (or Function ?) Initiative. Structure 15, 1353-6 (2007).
117. Harrison, S.C. Comments on the NIGMS PSI. Structure 15, 1344-6 (2007).
118. Hendrickson, W.A. Impact of structures from the protein structure initiative. Structure 15, 1528-9 (2007).
119. Mayo, C.J. et al. Benefits of automated crystallization plate tracking, imaging, and analysis. Structure 13, 175-82 (2005).
120. McPherson, A. Some ill considered comments on the protein structure initiative. Structure 15, 1526-7 (2007).
121. Moore, P.B. Let's call the whole thing off: some thoughts on the protein structure initiative. Structure 15, 1350-2 (2007).
122. Norvell, J.C. & Berg, J.M. Update on the protein structure initiative. Structure 15, 1519-22 (2007).
123. Steitz, T.A. Collecting butterflies and the protein structure initiative: the right questions? Structure 15, 1523-4 (2007).
124. Vakser, I.A. PSI has to live and become PCI: Protein Complex Initiative. Structure 16, 1-3 (2008).
125. Adams, M., Joachimiak, A., Kim, R., Montelione, G.T. & Norvell, J. Meeting review: 2003 NIH Protein Structure Initiative Workshop in Protein Production and Crystallization for Structural and Functional Genomics. J Struct Funct Genomics 5, 1-2 (2004).
126. Service, R. Structural biology. Structural genomics, round 2. Science 307, 1554-8 (2005).
127. Moult, J. Comparative modeling in structural genomics. Structure, 14-16 (2008).
128. Yee, A., Gutmanas, A. & Arrowsmith, C.H. Solution NMR in structural genomics. Curr Opin Struct Biol 16, 611-7 (2006).
129. Peti, W. et al. Towards miniaturization of a structural genomics pipeline using micro-expression and microcoil NMR. J Struct Funct Genomics 6, 259-67 (2005).

130. Page, R., Peti, W., Wilson, I.A., Stevens, R.C. & Wuthrich, K. NMR screening and crystal quality of bacterially expressed prokaryotic and eukaryotic proteins in a structural genomics pipeline. Proc Natl Acad Sci U S A 102, 1901-5 (2005).
131. Acton, T.B. et al. Robotic cloning and protein production platform of the northeast structural genomics consortium. Methods Enzymol 394, 210-43 (2005).
132. Segelke, B.W. et al. Laboratory scale structural genomics. J Struct Funct Genomics 5, 147-57 (2004).
133. Floudas, C.A. Computational methods in protein structure prediction. Biotechnol Bioeng 97, 207-13 (2007).
134. Guo, J.T. et al. PROSPECT-PSPP: an automatic computational pipeline for protein structure prediction. Nucleic Acids Res 32, W522-5 (2004).
135. Lambert, C., Leonard, N., De Bolle, X. & Depiereux, E. ESyPred3D: Prediction of proteins 3D structures. Bioinformatics 18, 1250-6 (2002).
136. Srinivasan, R. & Rose, G.D. Ab initio prediction of protein structure using LINUS. Proteins 47, 489-95 (2002).
137. Helles, G. A comparative study of the reported performance of ab initio protein structure prediction algorithms. J R Soc Interface 5, 387-96 (2008).
138. Oganov, A.R. & Glass, C.W. Crystal structure prediction using ab initio evolutionary techniques: principles and applications. J Chem Phys 124, 244704 (2006).
139. Zhang, Y., Kolinski, A. & Skolnick, J. TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J 85, 1145-64 (2003).
140. Hardin, C., Pogorelov, T.V. & Luthey-Schulten, Z. Ab initio protein structure prediction. Curr Opin Struct Biol 12, 176-81 (2002).
141. Bonneau, R. et al. Rosetta in CASP4: progress in ab initio protein structure prediction. Proteins Suppl 5, 119-26 (2001).
142. Aramini, J.M., Rossi, P., Anklin, C., Xiao, R. & Montelione, G.T. Microgram-scale protein structure determination by NMR. Nat Methods 4, 491-3 (2007).
143. Parsons, L. & Orban, J. Structural genomics and the metabolome: combining computational and NMR methods to identify target ligands. Curr Opin Drug Discov Devel 7, 62-8 (2004).
144. Staunton, D., Owen, J. & Campbell, I.D. NMR and structural genomics. Acc Chem Res 36, 207-14 (2003).
145. Powers, R. Applications of NMR to structure-based drug design in structural genomics. J Struct Funct Genomics 2, 113-23 (2002).
146. Kennedy, M.A., Montelione, G.T., Arrowsmith, C.H. & Markley, J.L. Role for NMR in structural genomics. J Struct Funct Genomics 2, 155-69 (2002).
147. Montelione, G.T., Zheng, D., Huang, Y.J., Gunsalus, K.C. & Szyperski, T. Protein NMR spectroscopy in structural genomics. Nat Struct Biol 7 Suppl, 982-5 (2000)
148. Radaev, S., Li, S. & Sun, P.D. A survey of protein-protein complex crystallizations. Acta Crystallogr D Biol Crystallogr 62, 605-12 (2006).
149. Schlichting, I. X-ray crystallography of protein-ligand interactions. Methods Mol Biol 305, 155-66 (2005).
150. Ohishi, H. et al. The development of crystallization and X-ray crystallography method of long stem DNA. Nucleic Acids Symp Ser (Oxf), 67-8 (2005).
151. Schotte, F., Soman, J., Olson, J.S., Wulff, M. & Anfinrud, P.A. Picosecond time-resolved X-ray crystallography: probing protein function in real time. J Struct Biol 147, 235-46 (2004).
152. Blundell, T.L. & Patel, S. High-throughput X-ray crystallography for drug discovery. Curr Opin Pharmacol 4, 490-6 (2004).
153. Hui, R. & Edwards, A. High-throughput protein crystallization. J Struct Biol 142, 154-61 (2003).
154. Luft, J.R. & DeTitta, G.T. A method to produce microseed stock for use in the crystallization of biological macromolecules. Acta Crystallogr D Biol Crystallogr 55, 988-93 (1999).
155. Andrec, M. et al. A large data set comparison of protein structures determined by crystallography and NMR: statistical test for structural differences and the effect of crystal packing. Proteins 69, 449-65 (2007).
156. Kim, S. & Szyperski, T. GFT NMR, a new approach to rapidly obtain precise high-dimensional NMR spectral information. J Am Chem Soc 125, 1385-93 (2003).
157. Snyder, D.A. et al. Comparisons of NMR spectral quality and success in crystallization demonstrate that NMR and X-ray crystallography are complementary methods for small protein structure determination. J Am Chem Soc 127, 16505-11 (2005).
158. Luft, J. et al. Macromolecular crystallization in a high throughput laboratory - the search phase. Journal of Crystal Growth 232, 591-595 (2001).

159. Chayen, N.E. & Saridakis, E. Protein crystallization for genomics: towards high-throughput optimization techniques. Acta Crystallogr D Biol Crystallogr 58, 921-7 (2002).
160. Hiraki, M. et al. Development of an automated large-scale protein-crystallization and monitoring system for high-throughput protein-structure analyses. Acta Crystallogr D Biol Crystallogr 62, 1058-65 (2006).
161. Morissette, S.L. et al. High-throughput crystallization: polymorphs, salts, co-crystals and solvates of pharmaceutical solids. Adv Drug Deliv Rev 56, 275-300 (2004).
162. D'Arcy, A., Sweeney, A.M. & Haber, A. Practical aspects of using the microbatch method in screening conditions for protein crystallization. Methods 34, 323-8 (2004).
163. Rupp, B. Maximum-likelihood crystallization. J Struct Biol 142, 162-9 (2003).
164. Page, R. et al. Shotgun crystallization strategy for structural genomics: an optimized two-tiered crystallization screen against the Thermotoga maritima proteome. Acta Crystallogr D Biol Crystallogr 59, 1028-37 (2003).
165. Klyushnichenko, V. Protein crystallization: from HTS to kilogram-scale. Curr Opin Drug Discov Devel 6, 848-54 (2003).
166. Grunlan, J.C., Mehrabi, A.R., Chavira, A.T., Nugent, A.B. & Saunders, D.L. Method for combinatorial screening of moisture vapor transmission rate. J Comb Chem 5, 362-8 (2003).
167. Juarez-Martinez, G., Steinmann, P., Roszak, A.W., Isaacs, N.W. & Cooper, J.M. High-throughput screens for postgenomics: studies of protein crystallization using microsystems technology. Anal Chefm 74, 3505-10 (2002).
168. Snell, E. H., A. M. Lauricella, et al. (2008). "Establishing a training set through the visual analysis of crystallization trials. Part II: crystal examples." Acta Crystallogr D Biol Crystallogr 64(Pt 11): 1131-1137.
169. Snell, E. H., J. R. Luft, et al. (2008). "Establishing a training set through the visual analysis of crystallization trials. Part I: approximately 150,000 images." Acta Crystallogr D Biol Crystallogr 64(Pt 11): 1123-1130.
170. McGuffin, M. J. and I. Jurisica (2009). "Interaction techniques for selecting and manipulating subgraphs in network visualizations." IEEE Trans Vis Comput Graph 15(6): 937-944.
171. D'Arcy, A. Crystallizing proteins - a rational approach? Acta Crystallogr D Biol Crystallogr 50, 469-71 (1994).
172. Hassell, A.M., Harrocks, T.J., Dashman, E.H. & Mistry, A. Two distinct approaches to crystallization results-recording databases. Acta Crystallogr D Biol Crystallogr 50, 459-65 (1994).
173. Hennessy, D., Gopalakrishnan, V., Buchanan, B.G., Rosenberg, J.M. & Subramanian, D. Induction of rules for biological macromolecule crystallization. Proc Int Conf Intell Syst Mol Biol 2, 179-87 (1994).
174. Hennessy, D., Buchanan, B., Subramanian, D., Wilkosz, P.A. & Rosenberg, J.M. Statistical methods for the objective design of screening procedures for macromolecular crystallization. Acta Crystallogr D Biol Crystallogr 56, 817-27 (2000).
175. Jurisica, I. et al. High Throughput Macromolecular Crystallization: An Application of Case-Based Reasoning and Data Mining. in Methods in Macromolecular Crystallography (eds. Johnson, L. & Turk, D.) (Kluwer Academic Press, 2000).
176. Jurisica, I. & Glasgow, J. Extending case-based reasoning by discovering and using image features in IVF. in ACM Symposium on Applied Computing (SAC 2000) 52-59 (ACM Press, Villa Olmo, Como, Italy, 2000).
177. Jurisica, I. et al. Image-Feature Extraction for Protein Crystallization: Integrating Image Analysis and Case-Based Reasoning. in The Thirteenth Innovative Applications of Artificial Intelligence Conference on Artificial Intelligence (IAAI-2001) (AAAI Press, Seattle, WA, 2001).
178. Jurisica, I. et al. Improving objectivity and scalability in protein crystallization: Integrating image analysis with knowledge discovery. IEEE Intelligent Systems Journal, Special Issue on Intelligent Systems in Biology, 26-34 (2001).
179. Jurisica, I. et al. Intelligent Decision Support for Protein Crystal Growth. IBM Systems Journal, Special Issue on Deep Computing for Life Sciences 40, 394-409 (2001).
180. Spraggon, G., Lesley, S.A., Kreusch, A. & Priestle, J.P. Computational analysis of crystallization trials. Acta Crystallogr D Biol Crystallogr 58, 1915-23 (2002).
181. Goh, C.S. et al. SPINE 2: a system for collaborative structural proteomics within a federated database framework. Nucleic Acids Res 31, 2833-8 (2003).
182. Kimber, M.S. et al. Data mining crystallization databases: knowledge-based approaches to optimize protein crystal screens. Proteins 51, 562-8 (2003).
183. Rupp, B. & Wang, J. Predictive models for protein crystallization. Methods 34, 390-407 (2004).

184. Goh, C.S. et al. Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. J Mol Biol 336, 115-30 (2004).
185. Jurisica, I. & Glasgow, J. Application of case-based reasoning in molecular biology. Artificial Intelligence Magazine, Special issue on Bioinformatics 25, 85-95 (2004).
186. Page, R. & Stevens, R.C. Crystallization data mining in structural genomics: using positive and negative results to optimize protein crystallization screens. Methods 34, 373-89 (2004).
187. Cumbaa, C. and I. Jurisica (2005). "Automatic classification and pattern discovery in high-throughput protein crystallization trials." J Struct Funct Genomics 6(2-3): 195-202.
188. Asur, S., Raman, P., Otey, M.E. & Parthasarathy, S. A model-based approach for mining membrane protein crystallization trials. Bioinformatics 22, e40-8 (2006).
189. Jurisica, I. & Wigle, D.A. Knowledge Discovery in Proteomics, (Chapman & Hall/CRC Press, 2006).
190. Kotlyar, M. & Jurisica, I. Predicting protein-protein interactions by association mining. Information Systems Frontiers 8, 37-47 (2006).
191. Cumbaa, C. A. and I. Jurisica (2010). "Protein crystallization analysis on the World Community Grid." J Struct Funct Genomics 11(1): 61-69.
192. Pastrello, C., Otasek, D., Fortney, K., Agapito, G., Cannataro, M., Shirdel, E., **Jurisica, I.** Visual data mining of biological networks: one size does not fit all, *PLOS Comp Biol*, **9**(1): e1002833. doi:10.1371/journal.pcbi.1002833, 2013

Health Informatics

193. A. Chakravarthy and K. Hasse. Netserf: Using semantic knowledge to find internet information archives. Proceedings of the 18th ACM SIGIR Conference, 1995.
194. J. Callan. Distributed information retrieval. In Croft W. B., editor, Advances in Information Retrieval. Kluwer Academic Publishers, 2000.
195. J. G. Conrad, X. S. Guo, P. jackson, and M. Meziou. Database selection using actual physical and acquired logical collection resources in a massive domain-specific operational environment, Proc. of the 28th VLDB Conference, 2002.
196. L. Si and J. Callan. Relevant document distribution estimation method for resource selection. In Proc. of SIGIR, 2003.
197. L Gravano, H. Garcia-Molina, and A. Tomasic. Gloss: Text-source discovery over the internet. ACM Transactions on Database Systems, 24(2), 1999.
198. P. Ipeirotis and L. Gravano. Distributed search over the hidden web: Hierarchical database sampling and selection, Proc. of VLDB, 2002.
199. P. Ipeirotis and L. Gravano. Classification-aware hidden-web text database selection. Technical Report CeDER-06-04, CeDER Working Papers Series, NYU, 2006.
200. K. L. Liu, C. Yu, W. Meng, W. S. Wu, and N. Rishe. A statistical method for estimating the usefulness of text databases. IEEE Transactions on Knowledge and Data Engineering, 14(6), 2002.
201. Cercone, N., An, X., Li, J., Gu, Z. and An, A. (2011) Finding best-evidence for evidence-based best practice recommendations in health care, Knowledge and Information Systems, 29, Springer, 159-201.
202. Pattaraintakorn, P. and Cercone, N. (2008) Integrating rough set theory and medical applications, Applied Mathematics Letters Volume 21, Issue 4, Pages 400-403.
203. Wan, Q. and An, A. (2007). Transitional Patterns and Their Significant Milestones, Proceedings of the 2007 IEEE International Conference on Data Mining (ICDM'07), 691-696.
204. Liu, Y., Huang, X., An, A. and Yu, X. (2007). A Sentiment-Aware Model for Predicting Sales Performance Using Blogs, Proceedings of the 30th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR'07), Amsterdam, 607-614.
205. Huang, X., Wen, M., An, A. and Huang, Y. (2006). A Platform for Okapi-Based Contextual Information Retrieval, Proceedings of the 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR'06), Seattle.
206. S. Soderland. Learning information extraction rules for semi-structured and free text. Machine Learning, 34, 1999.
207. M. Califf and R. J. Mooney, Relational learning of pattern-match rules for information extraction. In Proceedings of the 16th National Conference on AI (AAAI-99), Orlando, FL, 1999.
208. D. Freitag. Multistrategy learning for information extraction. In Proceedings of the Fifteenth International Conference on Machine Learning, 1998.
209. F. Ciravegna. Adaptive information extraction from text by rule induction and generalization. In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, 2001.
210. Z. Gu, Adaptive Information Extraction from Online Documents, PhD thesis, U Waterloo, Canada,

2006.
211. A. McCallum, D. Freitag, and F. Pereira. Maximum entropy Markov models for information extraction and segmentation. In Proceedings of ICML, 2000.
212. David M. Blei, Andrew Y. Ng , Michael I. Jordan (January 2003). Lafferty, John. ed. "Latent Dirichlet allocation". Journal of Machine Learning Research **3** (4–5): *pp.* 993–1022.
213. B. Liu, (2010). Sentiment Analysis and Subjectivity. (N. Indurkhya & F. Damerau, Eds.) Handbook of Language Processing, (1), 1-38. CRC Press, Taylor and Francis Group.
214. Picard, R. (1998) Human-Computer Coupling, *Proceedings of the IEEE*, 86(8), 1803-1807.
215. Ekman, P. & Fridlund, A. J. (1987) *Assesment Of Facial Behvior In Affective Disorders*. In J. D. Maser (Ed.) Depression and Expressive Behavior. Hillsdale, N.J.: Lawrence Erlbaum, 37-56.
216. Ekman, P. (1972) *Universals And Cultural Differences In Facial Expressions Of Emotions.* In J. Cole (ed.), Nebraska Symposium on Motivatioin, University of Nebraska Press, 207- 283.
217. Lazarus, R.S. (1991) Emotion and Adaptation, Oxford University Press.
218. Shami, N. S., and Ehrlich K. (2009) Making sense of strangers' expertise from signals in digital artifacts, CHI 2009, Boston.
219. Lee, C. HCI (2007) Aesthetics The Future Of User Interface Design. www.carrielee.net - 1.
220 Tractinsky, N., and Zmiri, D. (2005) Exploring Attributes of Skins as Potential Antecedents of Emotion in HCI, in Fishwick, P. (ed.) Aesthetic Computing, MIT Press.
221. Tufte, E. (2001) The Visual Display of Quantitative Information, 2nd ed. Cheshire: Graphics Press.
222. Reas C., and Fry, B. (2007) Processing: A Programming Handbook for Visual Designers and Artists Cambridge, Massachusetts: MIT Press.

Interactive Content Analytics

223. Ferrucci, David, and Adam Lally. "UIMA: an architectural approach to unstructured information processing in the corporate research environment." Natural Language Engineering 10.3-4 (2004): 327-348.
224. Bird, Steven. "NLTK: the natural language toolkit." Proceedings of the COLING/ACL on Interactive presentation sessions. Association for Computational Linguistics, 2006.
225. Marrin, Chris. "Webgl specification." Khronos WebGL Working Group (2011).
226. Hickson, Ian. "HTML Canvas 2D Context. W3C Working Draft." World Wide Web Consortium (2012).
227. Salga, Andor, Daniel Hodgin, Anna Sobiepanek, Scott Downe, Mickael Medel, and Catherine Leung. "Processing. js: Sketching with< canvas>." In ACM SIGGRAPH 2011 Talks, p. 15. ACM, 2011.
228. Whitelaw, M. (2006) Art Against Information: Case Studies in Data Practice In: Murphy, A., ed. *Proceedings, Fibreculture. Digital Arts and Culture Conference*, Perth. [Internet] Available from <http://journal.fibreculture.org/issue11/issue11whitelaw.html> [AccessedAugust, 2008] pp. 1 – 12.
229. Diamond, S. (2012) Many Points of Light: Visualization and the Converging Aesthetics of Art, Design and Science, Reprinted 2012, in *Critical Digital Studies Reader***.** ed. Kroker, A. & M. L.. Toronto: UT Press.
230. Cleveland, W.S. (1993) *Visualizing Data*. Summit, NJ: Hobart Press.
231. Ware, C. & Bobrow, R. (2005) Supporting Visual Queries on Medium-size Node-link Diagrams. *Information Visualization*, 2005, 4, London: Palgrave, pp. 49-58.
232. Greenberg, S. & Buxton, B. (2008) Usability Evaluation Considered Harmful (Some of the Time). *CHI 2008 Proceedings, April 5-10, Florence*. Florence: ACM, pp. 111-120.
233. Bardzell, J. & Bardzell, J. (2008) Interaction Criticism: A Proposal for a Framework for a New Discipline of HCI. *CHI 2008 Proceedings*, April 5 – 10, 2008, Florence, Italy: ACM, pp. 2463-2465.
234. Beck et al. 2001.
235. Spagnuolo, C. (2013) *Social Media*. [Internet] <http://agilecommons.org/pages/home> Boulder, Colorado: Rally Software [Accessed, January, 2013].
236. Wolkerstorfer, P., Tscheligi, M., Sefelin, R., Milchrahm,H., Hussain, Z., Lechner, M. & Shahzad, S. (2008) Probing an Agile Usability Process, *CHI 2008 Proceedings,* April 5 – 10, Florence, Italy: ACM, pp. 2151-2157.
237. Szigeti,S., Chevalier, F., & Diamond, S. (2013) A Faster Horse : Data Driven versus User-Centred Design for Data Visualization. Submitted to *CHI'13*, April 27 – May 2, 2013, Paris, France.
238. Gould, J.D. and Lewis, C.H. Designing for Usability: key principles and what designers think. *Communication of the ACM*, 28, 3 (1985) 300-311.
239. Cockton, G. (2008) Revisiting Usability's Three Key Principles *CHI 2008 Proceedings, April 5-10, Florence*, Florence: ACM, pp. 2473 – 2484.
240. Lee, 2007.

241. Tractinsky, Noam (2004) A Few Notes on the Study of Beauty in HCI. *Human-Computer Interaction*, 19 (4) pp. 351-357.
242. Guia, G., Oliver, S., Diamond, S., Chevalier, F. (2012) Visualizing Sentiments in Business Customer Relationships. *ACM, CHI'12***,** May 5–10, 2012, Austin, Texas, USA.
243. Oliver, S.; Gali, Guia, Chevalier, F. & Diamond, S. (2012) Fracturing Media Paradigms : Online Navigation of Online News Media. *ACM. DIS 2012,* June 11-15, 2012, Newcastle, UK.
244. Sedlmair, M., Isenberg, P., Baur, D. and Butz, A. Evaluating Information Visualization in Large Companies: Challenges, Experiences and Recommendations. In *Proc. BELIV'10* (2010) 79-86.
245. Diamond, S. (2012) Many Points of Light: Visualization and the Converging Aesthetics of Art, Design and Science, Reprinted 2012, in **Critical Digital Studies Reader.** ed. Kroker, A. & M. L.. Toronto: UT Press.