What Are The Grand Challenges for Data Mining?
KDD-2006 Panel Report

Gregory Piatetsky-Shapiro
KDnuggets
gps at acm.org

Chabane Djeraba
U. of Lille
Chabane.Djeraba at lifl.fr

Lise Getoor
U. of Maryland
getoor at cs.umd.edu

Robert Grossman
UIC & Open Data Group
rlg at opendatagroup.com

Ronen Feldman
U. of Bar-Ilan & ClearForest
ronenf at gmail.com

Mohammed Zaki
RPI
zaki at cs.rpi.edu

ABSTRACT

We discuss what makes exciting and motivating Grand Challenge problems for Data Mining, and propose criteria for a good Grand Challenge. We then consider possible GC problems from multimedia mining, link mining, large-scale modeling, text mining, and proteomics. This report is the result of a panel held at KDD-2006 conference.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining

General Terms

Measurement, Performance, Experimentation.

Keywords

Data mining, bioinformatics, multimedia mining, image mining, video mining, link mining, text mining, web mining, grand challenge, X-prize.

1. INTRODUCTION

Recently we saw several major scientific and engineering advances that were stimulated by a grand challenge/prize [CSM06, WSJ06]. The DARPA Grand Challenge produced great advances in robotic car navigation in 2005; X-prize led to the first successful commercial spaceflight; and RoboCup, (www.robocup.org) whose goal is to develop a team of humanoid robots that can win against the human world soccer champion team by 2050, has greatly advanced robotic performance and created many enthusiasts.

Looking further back, the first transatlantic flight by Charles Lindbergh in 1927 was also stimulated by a competition for a prize.

We have seen several examples where a Grand Challenge problem can get researchers, press, funding agencies, venture capitalists, and public interested, greatly stimulate research, and produce dramatic advances in science and technology.

What are the grand challenge problems for data mining ?

This question is timely - X-prize foundation is looking for additional fields where the prize can be created.

Fig. 1 Robotic Grand Challenge

??? Data Mining Grand Challenge

We propose the following criteria for a good grand challenge problem for data mining.

1) The **problem is hard** -- very difficult to solve given the current state of the art
2) **Involves data mining**: data mining plays an important role in solving the problem.
3) Based on a **large, publicly available data set**
4) There is a **specific goal**: it is clear when the problem is solved
5) Problem is **interesting to researchers** and **understandable to the public**, and preferably stated in one sentence.
6) There **is significant public benefit** if it is solved.

Some potential ideas for a grand challenge include:

- Automatic tagging and classification of 1 billion digital photos on the web. A company called Riya (www.riya.com) is already working on a smaller scale project.

- Identifying all genes and potential therapeutic targets for some specific types of cancer.
- A text-mining and understanding system that can use the web to pass standard tests, e.g. SAT in World History.
- Literature-based discovery of drug X side effects ([Swan86] is one of the earliest examples)
- Fraud detection based on company financial statements – can we find another Enron before it collapses?

Perhaps the KDD-06 Panel already had some effect – on October 2, 2006, Netflix announced $1 million prize for a program that substantially increases the accuracy of predictions about how much someone is going to love a movie based on their movie preferences (see www.netflixprize.com/). The Netflix Prize satisfies the first 5 of our proposed criteria.

Another related prize is Genomics X prize, recently announced by the X-prize foundation for technology that sequences the human genome quickly and cheaply [GenomeX06]. Here data mining plays some role, but is not central to the solution.

In the rest of this report we examine possible Grand Challenge problems in several hot research areas: multimedia mining, link mining, large scale data mining, text mining, and proteomics.

2. MULTIMEDIA MINING (Djeraba)

The rapid progress of data acquisition and storage technology has led to a tremendous amount of multimedia data stored in databases and files, and the amount of this data continues to grow very fast. Although valuable information is contained in multimedia data, the great majority of this data is non-structured or semi-structured, which makes it difficult (if not impossible) for human beings to extract the information without powerful tools.

Multimedia data is of no use unless we can actually access and mine it [Zai03]. How will the users explore the vast and growing multimedia information, including images, video and audio, at their disposal? There is a need for making sense out of the multimedia data and to use the multimedia content effectively and efficiently.

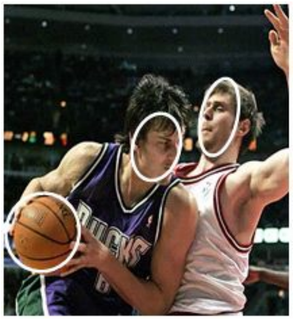
2.1 Grand Challenges

Let us consider the following grand challenge:

Annotate 1000 hours of digital video in one hour.

1000 hours is the approximate amount of “Rush” daily video produced by top news agencies. This currently needs thousands of man-hours to do manually. The annotation of one image in National Geographic Society takes about 20 minutes.

The challenge is to automate the entire annotation process. Advances in pattern recognition, or automatically extracting text from the speech accompanying video (when available) or recognition of text written on images may be a pragmatic way to bridge the semantic gap.



Extracting low level features such as colour distribution, texture and shape from pixels is easy. Extracting medium level features such as human faces, red ball, white, blue and green clothes, and people in the background is possible. However, extracting high level features such as Handball game, attack, defence, actions, is very difficult without user annotations

Fig 2: Low, medium and high level features

Another pragmatic way to bridge the semantic gap is to:

- Extract automatically primitive (low level) features, including key frames, shots, and other classical primitive features (e.g., colour distribution, Fourier transforms, wavelet, texture histograms, colour histograms, shape primitives, filter primitives) of large video databases.
- Annotate a subset of video database (e.g., presence of human faces, red ball, white, blue and green clothes, people in the background, Handball game, attack, defence, actions). In certain situations, medium-level features (e.g., faces, clothes) may be extracted automatically.
- Then, on the basis of the frequent patterns between primitive features and annotations (semantic features) of the subset of video, we generalize the annotations to the remaining video database. The complexity of the process turns around this last point.

Other grand challenges, may be considered, including:

- Predicting user interest on video lectures of a particular video web site on the basis of the first 5 minutes of browsing.
- Scanning an archive of video broadcasts to find similar interviews with a particular individual, e.g. a person running for a political office.
- Extracting from the football (soccer) game patterns that characterize the actions during 1 minute before the goal.

2.2 Great Research Areas

The grand challenges belong to two great research areas that involving usage and data.

1) *Mining user behaviours in interactions with multimedia data* and use the knowledge extracted in this process to anticipate future behaviours or to diagnose medical or psychological conditions of the users. The difficulty is to mine not only explicit actions (interactions, navigation), but also implicit reactions such as eye/gaze fixation, emotions (70% of people is emotion), heartbeat, respiration rate, stress, etc. The difficulty is also to use non-intrusive sensors (e.g., cameras), rather than intrusive sensors. The difficulty concerns also making possible the tracking of actions on multimedia data. This means that the tools for images, videos and audio should record user operations (e.g., play, pause, visualize, eye fixation) and multimedia pieces concerned by these operations - mining user actions, considering for example, intra/inter video actions.

2) *Crossing the semantic gap between multi-media data and semantics*. The difficulty is to extract automatically the meaning of multimedia content so that exploitation (e.g., retrievals) using semantic information can be tailored to individual applications (security, marketing, business, etc.). Multimedia data is the most natural information-conveying vehicle but also the most complex to index and mine. It is a very difficult process considering: the high volume (rapid explosion of available multimedia information), the complexity (videos, 3D models, audio, images), and the heterogeneity of data (streams, several sources). The difficulty is to generate metadata that describes the content and that may be exploitable in applications. Document semantics has been studied for quite some time. What is now needed is to develop approaches to extract semantics [Gro05] from multimedia documents so that retrievals using concept-based queries can be tailored to individual users. The semantic gap, or, as others put it, the semantic chasm, must be crossed. Multimedia usage mining coupled with domain ontology may be a revolutionary way to deal with the lack of semantics in multimedia information, and will certainly contribute to the hot domain of multimedia semantics.

3. LINK MINING GRAND CHALLENGE PROBLEM (Getoor)

There is an increasing need to both learn and extract structured data. Much of the input to today's data mining and machine learning algorithms is structured, often in the form of a graph or network. Examples include social networks, biological networks and communication networks. At the same time, in many cases there is a desire to learn structured outputs, for example extracting graphs describing entities and relationships from unstructured data.

Link mining [Get05] refers to both making use of the observed network's structure during learning and inference and inferring the (unobserved) link structure from other observations. Examples include using links for ranking nodes, using links for collective classification of nodes, and discovering links by predicting missing links or inferring new kinds of links and relationships.

Link mining tasks can be broken down into the following categories:

- Node Centric
  - Labeling/ranking nodes (aka Collective Classification/ PageRank)
  - Consolidating nodes (aka Entity Resolution)
  - Discovering hidden nodes (aka Group Discovery)
- Edge Centric
  - Labeling/ranking edges
  - Predicting the existence of edges
  - Predicting the number of edges
  - Discovering new relations/paths
- Graph/Subgraph Centric
  - Discovering frequent sub-patterns
  - Generative models
  - Metadata discovery, extraction, and reformulation

Current research mostly focuses on a single task such as node ranking or link prediction. In real data analysis scenarios, and particularly for a Grand Challenge, we need a mix of all of these capabilities.

The requirements for a Grand Challenge problem are discussed in section 1. While there is much structured data available, and even more unstructured data, finding a problem that meets the requirements is non-trivial. There are many problems which match some of the criteria such as social relevance, but for which the data is not publicly available, or for which the required domain knowledge is quite specialized.

One domain for which the data is available, the data mining tasks are difficult yet compelling and socially relevant, the required knowledge is accessible and there are not a great number of research groups working is Wikipedia. Wikipedia has generated a lot of interest in recent years, ranging from its founder and foremost evangelist, Jimmy Wales who describes Wikipedia as a project whose "goal [is] to distribute a free encyclopedia to every single person on the planet in their own language" to its detractors, such as Larry Sanger, Wikipedia co-founder who says, "Wikipedia has gone from a nearly perfect anarchy to an anarchy with gang rule" [Schiff06]. Other commentary includes that of Eric Raymond, Open-source movement figure, who opines "Disaster is not too strong a word for Wikipedia… the site is infested with moonbats" [Schiff06].

Regardless of one's opinion of Wikipedia, it is a great testbed for link mining algorithms. There are interesting studies involving building descriptive models of Wikipedia's growth, see for example en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia's_growth

Another useful task is predicting whether a contributor is a "wiki gnome" (a benevolent contributor who makes lots of edits, fixing typos and grammar mistakes) or a "wiki troll" (a destructive user who's edits are malicious). Text classification is also important, for example checking whether an article maintains the tenet of Wikipedia that a contribution must maintain a neutral point of view (NPOV).

Link prediction is also relevant, e.g. identifying where links should exist. This becomes even more compelling, because as Wikipedia grows, it becomes harder for any given author to know about other relevant information to which they should link. A link prediction method could help with this by doing link suggestion or automatic link construction. Evaluation can be done by generating a dataset of Wikipedia pages, removing some of the existing links, and then seeing if a system can identify those places and suggest appropriate links.

Other link mining tasks abound, including trust/reputation analysis, social network analysis and identification of communities, evaluation of accuracy, identifying misuse including vandalism and self-promotion and evaluation of coverage (which areas are not covered, or are poorly covered/linked?).

However, while each of these are interesting research topics, none of these really serve as a *Grand Challenge.* Instead, we propose the following

> **Wikipedia Test**: Given a collection of entries constructed via **participatory journalism** (such as the entries in Wikipedia) versus via automatic **link mining** tools, can you distinguish between the real Wikipedia entries and the automatically generated ones? Furthermore, which is better?

Evaluation could be done via a panel of human experts. Or, one can even automate the evaluation by leaving the entries up on Wikipedia and checking on their eventual page rank!

One compelling aspect of this challenge problem is that its solution will require a variety of integrated link mining capabilities. Another is that funding may already available: The Hutter prize, http://prize.hutter1.net/, provides 50,000 EUR for being able to being able to compress 100MB of Wikipedia to less than 18MB.

4. THE GRAND CHALLENGE OF ESTIMATING ONE BILLION PREDICTIVE MODELS (Grossman)

Large data sets can present challenges for data mining for a variety of reasons. One reason is that the data may be a mixture from several different sub-populations, each of which could benefit from a separate statistical or data mining model that is estimated using data just from that sub-population. For some applications, the sub-populations themselves may be unknown, with part of the challenge is to estimate these.

It has been a common practice for some time to build several different models in data mining. Manually segmenting populations and building a separate statistical model for each segment is a standard technique in statistics. For example, dividing a potential target audience into several different segments and estimating the parameters of a separate logistic model for each segment is a very common methodology for building response models in marketing.

Another example is provided by ensemble-based modeling techniques. Over the past two decades, a variety of ensemble based techniques have been used that estimate different statistical models either by re-sampling a small data set or by partitioning a large data set.

The challenge we address here is the challenge of automatically estimating the parameters in thousands or millions of individual statistical or data mining models, which can be required for very large or very complex data sets.

Here is an example from [Grossman06]. The data from this example comes from 833 traffic sensors in the Chicago metropolitan region and the goal is identify anomalous traffic patterns. In addition to the traffic sensor data, there is also semi-structured data about the weather and text data about any special events that can affect traffic, such as sports events. The goal is to decide whether traffic is unusual or anomalous. It is important to note that the goal is not to detect whether the traffic is congested, which is quite simple.

The approach taken was to segment the data into a separate segment, one for each hour of the day (24 hours), for each day of the week (7 days), and for each small segment of the highway (about 250 highway segments). This produced about 24x7x250 or 42,000 different segments. For each segment, the parameters of a separate change detection model were estimated using data belonging to that segment. In this way, over 42,000 separate statistical models were automatically created, updated, and used for detecting anomalous changes in the traffic. Due to the size of the data, its complexity, and its heterogeneity, this approach proved to be preferable to building fewer models.

Today, there are a variety of applications emerging in which makes sense to consider estimating a billion separate statistical models. Here are some examples:

- In online marketing, one could a build separate statistical model for each consumer. For large online companies in the near future, this could produce over one billion separate models.
- In detecting network anomalies, one could build separate statistical models for each IPv4 or IPv6 address.
- As a final example, for modern approaches to therapeutics, one could build a separate model predicting the efficacy of a new drug or treatment based upon the person's genotype. In other words, each genotype could be used to build a separate model. Again, over time, this could yield over a billion different models (cf. [Church05]).

We close with two remarks.

First, this challenge is not concerned with estimating a single model with the property that the model scales to a large number of different features vectors. Today, there are a variety of techniques that can be used to estimate the parameters of a model that will work for a large number of different feature vectors. The challenge addressed here is to estimate the parameters of a very large number of *different* models, each of which can work with a large number of different feature vectors.

Second, if we think of multiple models as arising from segmenting a large data set into segments $D_1, \dots, D_m$, some care is needed when stating the challenge to exclude using so many segments that the overall accuracy is harmed rather than helped by the segmentation. Here is one way to proceed. Given a data set D and a fixed class of possible models F, we can define the optimal partition number $m_{optimal} > 0$ as follows:

1. For m segments, where m = 1, 2, 3, …
2. Consider all ways P of partitioning the data set D into m segments: $D_1, \dots, D_m$
3. Let L_P denote the minimum total misclassification rate for all models $f_1, \dots, f_m \varepsilon F$ built on the data sets $D_1, \dots, D_m$. Note that L_p is a function of m.
4. Define $m_{optimal}$ to be the smallest m that minimizes $L_P(m)$ i.e. $m_{optimal} = \min \{ \text{argmin } L_P(m) \}$.

As the size and complexity of the data set D grows, so does the optimal partition number $m_{optimal}$. One way of stating the challenge is develop algorithms and an associated infrastructure that scales to large optimal partition numbers, in particular to optimal partition numbers > 1,000,000,000.

5. GRAND CHALLENGES FOR TEXT MINING (Feldman)

Text Mining is an exciting research area that tries to solve the information overload problem by using techniques from data mining, machine learning, NLP, IR and knowledge management. Text Mining involves the preprocessing of document collections (text categorization, information extraction, term extraction), the storage of the intermediate representations, the techniques to analyze these intermediate representations (distribution analysis, clustering, trend analysis, association rules etc) and visualization of the results.

Here are some of the challenges that are facing the text mining research area:

Challenge 1: Entity Extraction. Most text analytics systems rely on accurate extraction of entities and relations from the documents. However, the accuracy of the entity extraction systems in some of the domains reaches only 70-80% and creates a noise level which prohibits the adaptation of text mining systems by a wider audience. We are seeking domain independent and language independent NER (named entity recognition) systems that will be able to reach an accuracy of 99-100%. Based on such system, we are seeking domain independent and language independent relation extraction systems that will be able to reach precision of 98-100% and recall of 95-100%. Since the systems should work in any domain they must be totally autonomous and require no human intervention.

Challenge 2: Autonomous Text Analysis. Text Analytics systems today are pretty much user guided, and they enable users to view various aspects of the corpus. We would like to have a text analytics system which is totally autonomous and will analyze huge corpuses and come up with truly interesting findings that are not captured by any single document in the corpus and are not known before. The system can utilize the internet to filter findings that are already known. The "interest" measure which is totally subjective will be defined by a committee of experts in each domain. Such systems can then be used for alerting purposes in the financial domain, the anti-terror domain, the biomedical domain and many other commercial domains. The system will get streams of documents from a variety of sources and send emails to relevant people if an "interesting" finding is detected.

Based on systems developed in step 1 & 2, we would like to have (this is our text mining grand challenge)

> **Text mining systems that will be able to pass standard reading comprehension tests such as SAT, GRE, GMAT etc.**

Systems that will be able to pass the average scores will win the grand challenge. The systems can utilize the web when answering the test questions. We view this grand challenge as an extension of the classic Turing test. This grand challenge satisfies most of the criteria that were set for the various challenges. First, there are no systems today that are able to get above average score in any of the standard tests. Second, the criterion for success is very well defined. Then, we believe that within 5 years researchers will be able to build such systems based on technologies that are developed for annual competitions such as ACE, TREC and TIDES. Finally, having such systems will contribute to the advance of humankind as the underlying technologies deployed by these systems can be utilized by children and adults to more rapidly acquire knowledge about various topics.

6. MINING THE PROTEOME (Zaki)

Large-scale databases from sequencing projects, microarray studies, gene-function studies, protein-protein interactions, comparative genomics, structural biology, and open source journal articles, are growing at rapid rates. The challenge in systems biology is to connect all the dots from the diverse molecular, cellular, organism and environmental data sources to deduce how sub-systems and whole organisms work. We need to decipher the language of life – the language of the genome, protein folding, developmental pathways, and much more. There are numerous computational challenges in collecting, indexing, searching and mining these vast data sources. Mining the diverse sources of public data will be a crucial component in piecing together the bigger picture.

In particular, mining the proteome, especially mining new protein interactions and functionally enriching the proteome, is emerging as a grand challenge for data mining.

6.1 Protein Functions

Proteins are the fundamental molecules of life. A protein has a three-dimensional (3D) fold that determines its roles. Proteins play diverse roles in the cells, such as:

- Proteins as Molecular Machines: proteins can change their shape, allowing opening/closing movements, as well as twists and turns. Thus proteins can function as molecular switches.
- Proteins as Catalysts: enzymes are proteins that act as catalysts, speeding up biochemical reactions by several orders of magnitude, making life possible.
- Proteins in Pathways: proteins take part in "sequences" of biochemical reactions or pathways to enable a wide variety of functions.

Some of the common functions of proteins include: a) *Metabolism*: proteins mediate chemical reactions, b) *Signaling*: proteins are involved in signaling within and between cells, c) *Regulation*: proteins act as gateways in cellular membranes, d) *Cellular structure*: proteins can help define cell shape and form, e) *Transportation*: proteins are involved in moving other proteins, oxygen, sugar, nutrients and wastes into and around cells, f) *Movement*: proteins play a role in muscle contraction and cell movement, g) *DNA Transcription*: transcription factors are proteins that turn genes on/off, h) *Immunity*: proteins identify germs and other foreign substances and mark them for destruction.

6.2 Genomics to Proteomics

The genes, which are contiguous stretches of DNA, encode the information to manufacture proteins. Often there is a very complex regulatory network involving several genes that controls the production of proteins. Protein formation happens via two main steps: *transcription* of the gene into a mRNA molecule, and *translation* of the mRNA molecule into a protein, as illustrated in **Figure 3**.

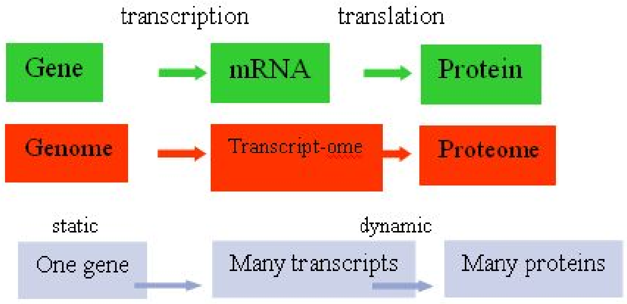


Figure 3: From Genes to Proteins

In the traditional view, it was thought that one gene gave rise to one mRNA transcript, which in turn produced one protein. However, the current view is that we have to consider all the genes (the genome), all the mRNA transcripts (the transcriptome) and all the proteins (the proteome) in totality. Thus a single gene can produce many transcripts (via alternate splicing), and these transcripts can produce many proteins. The proteins undergo many post-translational modifications that further increase the protein diversity. For example, in humans, there are around 30,000 genes, yet there are over a million proteins, when one accounts for post-translational modifications. Also note that whereas the genome is the static information repository, both the transcriptome and proteome are dynamic, since they change in response to the cellular environment.

6.3 Data Mining Challenge: Functional Annotation & Mining of the Proteome

The **Proteome** is the complete set of proteins in the cell under a set of conditions. It is dynamic and complex, and characterized in terms of:

- Structure – shape, electrostatic properties, etc.
- Abundance – protein expression level, i.e., the quantity of protein present.
- Localization – sub-cellular location.
- Modifications – post-translational modifications.
- Interactions – protein-protein interactions (PPI; also called the interactome).

The goal of functional annotation of the proteome is to comprehensively catalog the following information:

- **Why** is a given protein produced (biological process)?
- **What** kind of molecule is it (molecular function)?
- **Where** is it found (cellular localization)?

Functional annotation can help characterize unknown or hypothetical proteins via “guilt by association”.

The data mining grand challenge is to determine **which proteins are present**, **in what quantities**, **where are they localized**, **and whom they interact with** (binary/complex interactions).

If we could predict the 3D protein shape from sequence alone, we could then infer the protein-protein interactions and other interactions involving proteins directly. However, protein structure prediction is a grand challenge in its own right.

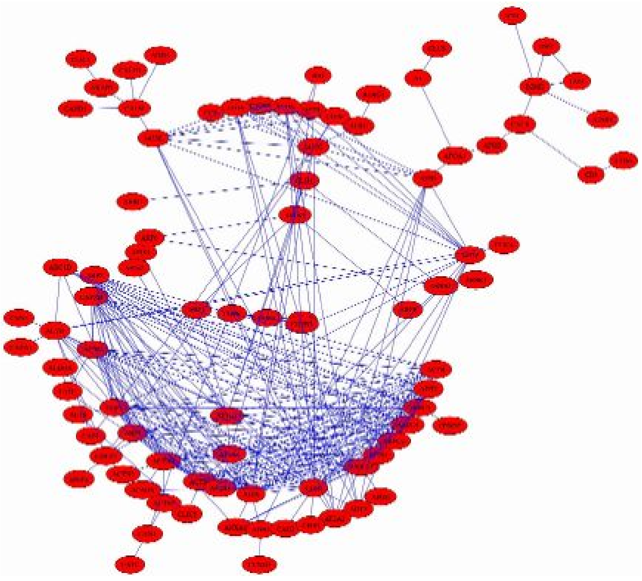


Figure 4. A part of the PPI network involving 90 human proteins, with 266 interactions

What we do have is a growing amount of publicly available data from protein mass spectrometry, protein arrays, PPI datasets across organisms, PubMed journal articles (requiring text/literature mining), and transcriptomics data (e.g., microarray datasets). The challenge is to integrate all these sources, to mine new protein interactions and to create a complete functional categorization of all proteins. For example **Figure 4** shows part of the PPI network involving 90 human proteins, with 266 interactions.

6.4 Public Data Sources

There is a wealth of data available that has to be integrated and mined to help solve the above grand challenge problem. The sources of data include:

- *Protein Expression and Raw PPI Databases:* These data come from 2D Gel Electrophoresis, Affinity Chromatography, Mass Spectrometry, and Protein Chips/Arrays.
- *Literature Curated PPI Databases*: There are many publicly available datasets cataloging the PPI across species. Examples include HPRD (Human Protein Reference Database), which has 18,284 proteins, and 33,710 interactions; DIP (Database of Interacting Proteins), which contains 19,075 proteins, and 55,757 interactions; MINT (Molecular Interaction DB), which catalogs 26,055 proteins, and 72,436 interactions; BIND (Biomolecular Interaction Network Database), which notes 205,846 interactions; IntAct (45,888 proteins, and 68,269 interactions); BioGrid (305,683 proteins, and 158,134 interactions); and many others (MIPS, CYGD - yeast, HPID, etc).
- *Potential interactions (Predictome):* There are databases the list inferred interactions, such as: InterDom (30,037 putative interactions), and OPHID - Online Predicted Human Interaction Database (49,008 predicted interactions).
- *Orthologs databases*: Orthologs are proteins that are homologs in other organisms. Using orthologs it is possible to infer new interactions, say in humans, by checking if the orthologous proteins interact in other organisms like yeast, fly, worm, etc. For example the Inparanoid database has orthologs information from 26 organisms, spanning 463,242 sequences. Orthologs (called Interologs) can also be predicted from sequence searches in model organisms, via BlastP (protein-protein BLAST) searches.
- *Post-translational Modifications*: Databases like RESID list hundreds of known protein modifications (such as glycosylation, phosporylation, etc.).
- *Literature Mining*: Mining PubMed journal articles, looking for protein interactions, is extremely useful in inferring new relationships among proteins. Databases like iHOP (Information Hyperlinked over Proteins) list such mined data.
- *Transcriptomics Data*: These are databases that catalog the gene expression and knockout information. Examples include cDNA libraries, DNA microarrays, mutagenesis and gene knockout experiments, and

RNAi interference databases. These data also provide clues as to the interacting proteins and the functional modules.

- *Gene Ontology (GO):* The three categories of the GO hierarchy span the biological process, molecular function, and sub-cellular location for many genes. The GO data can be integrated in proteomics studies to check the validity of mined modules.

7. SUMMARY

This is an opportune time to consider and propose Grand Challenge Problems for data mining. A good Grand Challenge problem should be hard, involve data mining, rely on a large, public dataset, have a specific goal, be interesting to researchers and the public, and promise significant public benefit if solved. We offer this discussion of possible grand challenge problems as a first step to creating such Data Mining Grand Challenges.

8. REFERENCES

[Church05] G. M. Church, The Personal Genome Project, Molecular Systems Biology, 2005, doi:10.1038/msb4100040.

[CSM06] “Grand challenges spur grand results - Private groups are offering big cash prizes to anyone who can solve a range of daunting problems”. The Christian Science Monitor, January 12, 2006 www.csmonitor.com/2006/0112/p13s01-stss.html

[GenomeX06] Genomics X Prize home, www.xprize.org/xprizes/genomics_x_prize.html

[Get05] SIGKDD Explorations Special Issue on Link Mining, Lise Getoor and Chris Diehl, December 2005

[Gro05] William I. Grosky, Nilesh Patel, Xin Li, Farshad Fotouhi: Dynamically Emerging Semantics in an MPEG-7 Image Database. Comput. J. 48(5): 536-544, 2005

[Grossman06] Robert L. Grossman, et al, Real Time Change Detection and Alerts from Highway Traffic Data, ACM/IEEE International Conference for High Performance Computing and Communications (SC '05).

[Schiff06] *Know It All: Can Wikipedia Conquer Expertise?* Stacy Schiff, New Yorker, July 31, 2006

[Swan86] Don R. Swanson, Fish Oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30, 7-18, 1986.

[WSJ06] Prize for DNA Decoding Aims to Fuel Innovation, Wall Street Journal, Jan 27, 2006

[Zai03] Osmar Zaiane, Simeon Smirof, Chabane Djeraba, *Knowledge Discovery from Multimedia and Complex data*, LNAI 2797, ISBN 3-540-20305-2, Springer Verlag, 2003.