Bisque:Advances in Bioimage Databases

Kristian Kvilekval, Dmitry Fedorov, Utkarsh Gaur,
Steve Goff, Nirav Merchant, B.S. Manjunath, Ambuj Singh

Abstract

Biological image databases have quickly replaced the personal media collections of individual scientists. Such databases permit objective comparisons, benchmarking, and data-driven science. As these collections have grown using advanced (and automated) imaging tools and microscopes, scientists need high-throughput large-scale statistical analysis of the data.

Traditional databases and standalone analysis tools are not suited for image-based scientific endeavors due to subjectivity, non-uniformity and uncertainty of the primary data and their analyses. This paper describes our image-database platform Bisque, which combines flexible data structuring, uncertain data management and high-throughput analysis. In particular, we examine: (i) Management of scientific images and metadata for experimental science where the data model may change from experiment to experiment; (ii) Providing easy provisioning for high-throughput and large-scale image analysis using cluster/cloud resources; (iii) Strategies for managing uncertainty in measurement and analysis so that important aspects of the data are not prematurely filtered.

1 Challenges for Bioimage Researchers

Current research in biology is increasingly dependent on conceptual and quantitative approaches from information sciences, ranging from theory through models to computational tools [6]. Ready availability of new microscopes and imaging techniques has produced vast amounts of multi-dimensional images and metadata. The introduction of new models, measurements, and methods has produced a wealth of data using image-based evidence [24]. Two notable examples of image-based studies are cellular Alzheimer's studies and plant genetics.

In a recent Alzheimer's study, the ability to reliably detect nuclei in three dimensions was critical to quantitative analysis [25]. The use of nuclei detection also finds use in a wide range of applications such as the accurate determination of how an organism is perturbed by genetic mutations, treatment with drugs, or by injury. Additionally, nuclei centroid locations can be used for further analysis, such as cell membrane segmentation or to initialize and validate computational models of cells and their patterns.

In the plant domain, technologies for quantifying plant development are underdeveloped relative to technologies for studying and altering genomes. As a result, information about plant gene function inherent in mutant phenotypes or natural genetic variation remains hidden. For example, plant scientists are trying to uncover gene function by studying seedling growth and development using high throughput image analysis [19].

In both cases, researchers dependent on images as experimental evidence face the daunting task of managing, analyzing and sharing images in addition to gaining and providing access to analysis methods and results [1]. In

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

the following paragraphs, we enumerate several challenges commonly faced by scientific researchers working with large-scale image data.

Growth of data: automated imaging and large scale processing Image-based experiments can produce hundreds to millions of images per experiment. For example, automated image-based phenotyping [19] produces several terabytes of image data. In addition to the data management problem, researchers are increasingly dependent on automated or semi-automated image analysis [32] due to large amounts of images involved in modern biological studies. Researchers working with very large datasets must thus take advantage of scalable computational resources. The results of analyses also pose a data management problem requiring careful management of initial (raw) and processed data while maintaining relationships to analyzed results.

Dealing with novelty and reproducibility in scientific experiments Biological image data models require a flexibility not usually needed by traditional database applications. In fact, the key to properly interpreting biological images is the experimental and image related metadata captured during experimental preparation and imaging. For example, a sample's biology including genetic mutations, or imaging techniques such as antibody staining are not discernible from the available pixel data. Furthermore, image data is often multi-dimensional including volume and time as well as being multi-channel (i.e. antibody labels rendered as false color images). Biology labs employ diverse experimental procedures [36] and continually invent new procedures and preparations resulting in unique local workflows [28]. New measurements, analysis, and statistics have also become increasingly complex and challenging [8]. Novel analysis techniques and results may require changes to the underlying data model [33] in order to be made available along with original data. While several laboratory image database systems have been developed, database schema rigidity has often become problematic as requirements evolve due to changes in experimental protocols and required analyses.

As computational image analysis is further integrated into the scientific process, accurate tracking of experimental results has become a primary concern [20]. Maintenance of original data while ensuring accurate tracking of results is fast becoming a requirement. In order to ensure accurate provenance, result data needs to be reliably marked by a tamper-resistant system in which the analysis and the resultant data become fixed once added to the system. Provenance provides security, confidence and traceability for resulting analysis.

Summarizing, comparing, disseminating, re-evaluating, mining Image analysis workflows can create large amounts of data that must be summarized, and compared for human understanding. Object identification image-analyses (i.e., segmentation and classification) can produce millions of spatial annotations per image (e.g., cells in tissue sample, microtubules within a cell). Furthermore, dissemination of large datasets can in itself be challenging. While some funding agencies do require access to published data for open science requirements [20], many researchers also realize the importance of data sharing and collaboration that can lead to better data validation, increased knowledge discovery, and cross-domain collaborations [37]. Even though researchers are usually willing to share data once published, strict security must be in place for works in progress. Achieving the goals of strict security and ease of sharing has proved challenging, and in many cases sharing of data has suffered.

Sources and management of uncertainty In an extensive number of scientific investigations, there is a growing dependence on large volumes of image and video data, and subsequently their analysis. Moreover, most image-based measurements are inherently uncertain and when used in extensive workflows accumulate errors if uncertainty estimation is not properly propagated. While it is possible to record and propagate uncertainty of measurements throughout processing pipeline, most systems rely on some sort of thresholding of results at each processing step. Even if the uncertainty measurement is recorded, the data produced (containing inherent ambiguities and uncertainties) can pose significant challenges to information processing systems.

2 Bisque: a Scientific-Image Database for Image Analysis

Bisque is a novel image database and analysis system built for biologists working with digital artifacts (such as images and video) as a fundamental tool for gathering evidence. Bisque allows users to store, access, share and process biological images, files and other resources from any web-connected device. The system is built to manage and analyze both data artifacts and metadata for large scale datasets by utilizing emerging large-scale computing cloud computing principles. The Bisque system [17] has been in development for several years and is being used for large scale image management and analysis. We highlight some of the unique features and services provided by Bisque for scientific-image management and analysis.

2.1 Interpreting scientific images: pixels to understanding

Biological digital artifacts including microscopic images and sensor data are only interpretable with the appropriate context. For example, an RGB image of cells does not explain what tissue the cells are embedded in, what the sample came from, nor how it was treated. It is vitally important to preserve this contextual metadata if we are to have any hope of re-utilizing (through analysis or mining) scientific data. In some cases, the context is well understood or previously defined. For example, the human genome project required researchers to contribute identified sequences to a known corpus in well defined formats such that the results could be easily shared and understood by all [15]. The majority of scientific experiments are not carried out with such a degree of coordination between members of the scientific community. Nor it is desirable as the time and effort needed to make them compatible must be paid in advance before the results are known. Instead, we believe, it is better to capture the experimental context as the researchers describe it and provide tools to make it available and searchable later.

2.2 Context and metadata : flexible modeling for scientific context

At the core of a Bisque site is the Bisque data service allowing Bisque resources to be stored and queried. Resources allow flexible modeling of context and experimental data permitting experimenters and practitioners to model their metadata as their experiments require. Each resource is comprised of a metadata record (a nested tree of key-value pairs) with an optional binary component. Resources can be easily linked together to provide common context. For example, in Fig. 1) an image may be linked to both a project and sample resources allowing each to be described independently and completely. Bisque resources form graphs which are used for data modeling, and data retrieval. The system is schemaless in that scientists define the data model per lab, experiment etc.

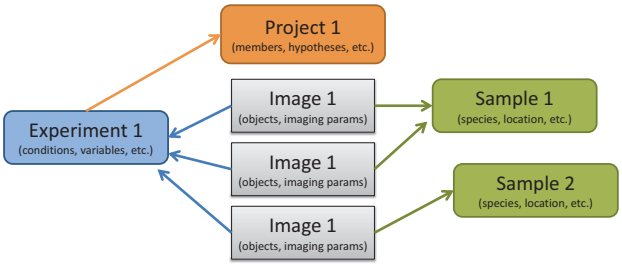

Figure 1: Bisque resources are document graphs linking contextual metadata. Here, images are linked to both a experiment and sample resources allowing each to be described independently, while an experiment is linked to a project. These graphs are also used for data retrieval. Annotations in each of the nodes are schemaless in that scientists are free to define complex trees of name-value-type tuples on the fly.

Resources are easily represented by linked XML documents. Every Bisque resource has a unique URL and may be accessed using common web protocols. The Bisque data service permits queries of top level resources based on elements of the Bisque resource graph.

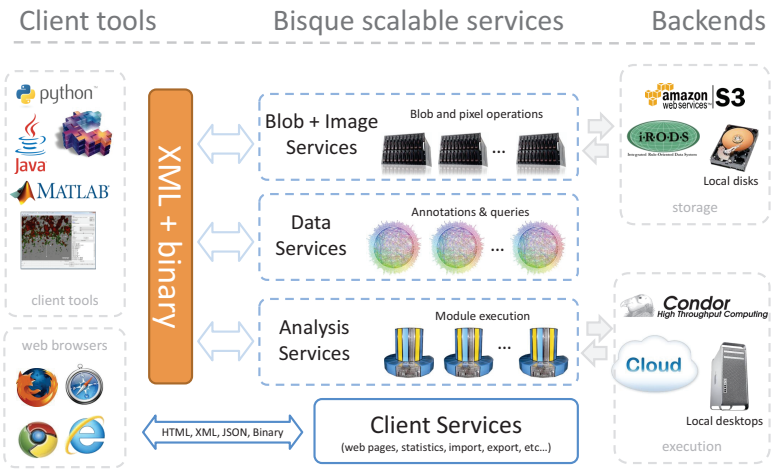

Figure 2: Bisque is implemented as a scalable and modular web-service architecture. *Image Servers* store and manipulate images. Data servers provide flexible metadata storage. Execution servers house executable modules. The client-server architecture seamlessly integrates services across storage and processing hardware. Communication between various components is performed in a RESTful manner through HTTP requests carrying XML and JSON. Web clients use dynamic web components built with AJAX. Backends seamlessly provide access to local resources as well as large scalable resources like Amazon S3 or iRODS for data storage or Condor grids for computation.

2.3 Large-scale storage and processing

The Bisque system has been designed for scalability from top to bottom. When installed, a Bisque site (Fig. 2) is comprised of many micro-services, each performing a specialized task. The majority of Bisque services simply manipulate and interpret the resource metadata. Other specialized micro-servers work with particular binary resources. For example, the Bisque image server is capable of performing common image operations (channel mapping, slicing, format conversion, etc.) on many image formats (both proprietary and open).

Scalability requires that the system can grow to support very large datasets, large numbers of users, and large analysis workflows. User scalability is provided by the fact that Bisque servers may be replicated and may be scaled using well-understood web-server scaling techniques (reverse proxies, replicated servers, etc). Furthermore, a Bisque site can take advantage of computational scaling for image analysis using cluster and grid resources in order to process very large datasets. Two internal components directly take advantage of these services: the blob server for large-scale storage and the module engine for parallel and scalable analysis. The blob service allows binary objects to be stored on a variety of storage systems from local file-systems to large distributed file-stores such as iRods [16], HDFS [14], or Amazon S3 [3]. Users are freed from storage issues and can access their resource from any Internet-accessible site. The module engine allows developers to integrate analysis algorithms with one or more Bisque sites. The engine performs both Bisque interfacing and job management for analysis routines. The engine can also be configured to use local resource or to use external job schedulers such as Condor [35] permitting access to cluster and grid resources. A key feature is that developers can often develop locally and deploy on grid resources with no changes to the analysis code. This is due to the fact that analysis modules using Bisque services need only make HTTP accesses.

2.4 Extensibility and availability

Bisque is extensible through both service development and analysis development. Example extended services include the statistics and summarizing service, and a search-by-content service. Bisque services are RESTful [22] allowing access from any programming environment that can make HTTP request.

Bisque views analysis modules simply as specialized websites (providing specific http endpoints) offering

analysis services. Many example analysis modules are included with Bisque and can be used as a guide for development. User modules can then be linked to an existing Bisque site (see Fig. 3). Analysis developers can install a small package which wraps common developed code to provide the needed protocol elements.

2.5 Provenance and trust

A core feature of Bisque system is that resources retain their provenance. When a resource is created or modified, it is marked by the analysis or session of the action. The Module Execution Record (MEX) can be linked to form the complete provenance of any item in the database from original upload or creation to the resulting statistics from the analysis of a dataset. Since provenance maintenance is managed by the system itself, users (and reviewers) can be assured that items are tamper-free, thus providing a level of trust in the data. Users can user provenance data to follow faulty data or simply manage chains of analysis.

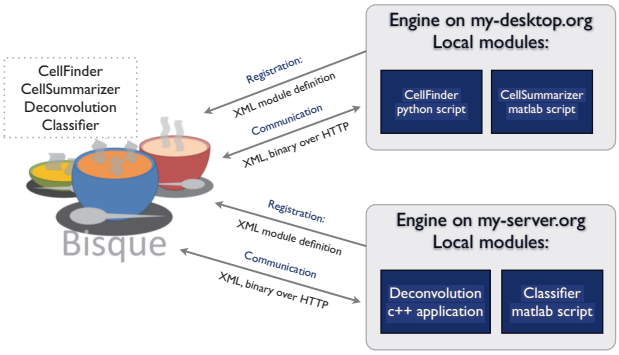

Figure 3: Bisque image analysis capabilities can be augmented by adding new services available on external computational resources. These external resources can be arbitrarily complex and harness available local power seamlessly from the main system. A Bisque module is simply a web-enabled internet end-point which adheres to a specific Bisque API for communication. The API is based on simple XML and HTTP/HTTPS transfers. Each end point can be made available by running a Bisque Engine Server that wraps all the communication and can enable simple Matlab or Python scripts to become Bisque modules. Each module is registered to the system by providing an XML description that includes required inputs and outputs.

3 Scientific Data Services

Scientific labs can often produce more data and analytical problems than they can handle. Each lab can search for or buy computational resources but these tasks can incur high overheads both in terms of people, money and time. With the advent of virtualized commodity computing, labs have new choices: utilizing free computational resources or renting large-scale resources.

Bisque users and developers have many choices on how to use the system: use a public installation such as the ones at Center for Bioimage Informatics [9] or iPlant [13], install a private instance on local hardware, or install an instance on a public cloud.

3.1 Available infrastructure and Cloud computing for e-sciences

The iPlant Collaborative [13] is an NSF-funded cyberinfrastructure (CI) effort directed towards the plant sciences community. The inherent interdisciplinary nature of plant sciences research produces diverse and complex data products that range from molecular sequences to satellite imagery as part of the discovery life cycle. With the constant creation of novel analysis algorithms, the advent and spread of large data repositories, and the need for collaborative data analysis, marshaling resources to effectively utilize these capabilities necessitates a highly flexible and scalable approach for implementing the underlying CI.

The iPlant infrastructure simultaneously supports multiple interdisciplinary projects providing essential features found in traditional science gateways as well as highly customized direct access to its underlying frameworks through use of APIs (Application Programming Interfaces). This allows the community to develop de novo applications. This approach allows us to serve broad community needs while providing flexible, secure, and creative utilization of our platform that is based on best practices and that leverages established computational resources.

Bisque is available for use by qualified plant science projects utilizing iPlant resources including the Data Store and High Performance Computing Infrastructure.

3.2 Large-scale processing using temporary resources

Commoditized computing services have become available from commercial providers such as Amazon, Google, Rackspace, and RightScale. In many cases, these are provided at low cost or freely for qualified projects. Bisque has been designed to work with such services allowing large-scale processing to be available to small labs.

For example, the Bisque project provides templates to deploy a Bisque server with a parallel Condor grid on Amazon EC2/S3. A scientist faced with analyzing a very large image dataset can, without much effort, create a "temporary" cluster-system for processing the data. Scientists can utilize these on-demand services to process very large datasets without the need to build large-infrastructure.

4 Future: Coping with Uncertainty

Imaging is at the core of many scientific discoveries, at scales varying from nano to astronomical observations. The information is captured in terms of raw pixel intensities and in multiple channels for color or hyperspectral imagery. These pixel values by themselves have very little meaning in most cases, and it is their spatial and temporal variations that are of much interest. Thus, most image analysis methods implicitly address the *uncertainty* in the observed values, whether it is for edge detection or segmentation.

The Bisque system is being extended to support *uncertainty* for confidence measures and spatial annotations which results from measurements and analysis. Two key challenges being addressed are: modeling uncertainty and computing with uncertain data.

4.1 Modeling uncertainty

Uncertainty in database systems can be modeled at the level of tuples or attributes. In the former case, the existence of a tuple is uncertain and a probability value is assigned to each tuple representing the confidence with which that particular tuple is present in the database. In the case of attribute uncertainty, there is certainty about the existence of the tuple but uncertainty about its specific attributes. Combinations of tuple and attribute uncertainties are also possible. Furthermore, dependence between attributes and tuples in a given relation or across relations can be modeled by probabilistic graphical models [29] in which random variables represent tuples or attributes. However, inference (the basis for answering queries) is difficult due to NP-completeness, and various approximations need to be employed.

Existing database literature in the area of uncertainty can be broadly divided into two categories: models and semantics, and query processing and index structures. The usually accepted model for evaluation of probabilistic queries is the "possible-worlds" model [4, 11, 12, 18] in which a set of possible certain worlds is defined for the database objects, the query is evaluated in these certain worlds, and the result is aggregated to return the answer. In general, such evaluation is computationally challenging and makes a straightforward implementation all but infeasible for large databases. A number of strategies have been developed to find subclasses of queries that can be answered efficiently. This includes the isolation of safe query plans [4, 11]. Other approaches combine uncertainty with lineage [5] and consider the decomposition of possible worlds into a manageable set of relations [2]. Aspects of completeness under various models of uncertainty have been considered [27].

Bisque is being extended to work with confidence tags for existential annotations and probability masks to map spatial uncertainty. We are also investigating vertex models for spatial uncertainty including uncertain points, lines and polygons. Initially, probabilistic data will be available to analysis methods and visualization systems in order to allow researchers to gain experience utilizing probabilistic data. For example, annotation

visualization will allow filtering results based on confidence and permit visualizing zones of uncertainty for spatial objects.

4.2 Computing and searching

Answering queries on probability density functions (*pdf*) has been well studied [7,10,34] under the Gaussian and Uniform *pdf*s. These assumptions allow for interesting and efficient index structures, and can be appropriate for the uncertainty of many of the individual measurements. They are too restrictive for *pdf*s that occur as a result of summarization, however, the observations being summarized may be generated by different mechanisms. For instance, a summary of the density of bipolar cells in a detached cat retina will have two peaks, corresponding to parts of the retina that are injured and healthy, respectively. Fitting a model distribution, such as a Gaussian, to the data is only appropriate when the data are well understood, and in a scientific database, the most interesting data to query are precisely the ones that are not well understood. Others have considered the indexing of uncertain categorical data [30], the use of Monte Carlo simulations and state space search to answer top-k queries [26, 31], and skyline queries [23]. The TRIO system [38] supports data uncertainty at different levels and lineage. The ORION project [21] is a recent work aimed at developing an advanced database system with direct support for uncertain data.

Analysis methods such as segmentation result in different shapes and sizes based on the method or the initial conditions and subsequent data mining (classification, proximity analysis) can produce different outcomes. Since the possible worlds are too numerous to examine or visualize, one challenge is to produce a set of "diverse" outcomes by sampling from the underlying segmentation uncertainty and providing them to a user or another analysis tool. Another interesting possibility is to extract features of a given image and use these to suggest an analysis method (e.g., segmentation, initial seed) that is likely to work well based on past experience.

We are currently working to extend Bisque with spatial join and filtering including probabilistic objects with spatial uncertainty. Initially we will focus on generating spatial indexes for probabilistic objects and visualization tools. These tools will allow researchers to filter for spatial properties. For example, a researcher might be interested in determining characteristics of labeled astrocyte cells lying "near" a blood vessel in the retina. Determining the extent of astrocyte is ambiguous due to dendrite extensions connecting cells. The uncertainty of the shape of final astrocyte cell needs to be factored into the spatial join at the last possible moment.

5 Conclusion

Scientists working with images as fundamental evidence face many challenges. Gathering, documenting, organizing, analyzing and disseminating original data are at the core scientific of the process. Large scale imaging presents issues at each of the activities. We have outlined the challenges that we have discovered while working with collaborators from the biological sciences and have outlined our system Bisque to assist researchers working with image data. Key amongst those challenges are: accurate collection and organization of metadata, and analysis and dissemination of original images, analysis results, and provenance. We have also outlined our efforts to support uncertain data within the Bisque system allowing researchers to manage the uncertainty of results throughout a scientific workflow.

References

[1] P. Andrews, I. Harper, and J. Swedlow. To 5d and beyond: Quantitative fluorescence microscopy in the postgenomic era. *Traffic*, 3(1):29–36, 2002.

[2] L. Antova, C. Koch, and D. Olteanu. 10106worlds and beyond: Efficient representation and processing of incomplete information. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 606 –615, april 2007.

[3] Amazon EC2 and S3. http://aws.amazon.com/.

[4] D. Barbará, H. Garcia-Molina, and D. Porter. The management of probabilistic data. *IEEE Tansaction on Knowledge Engineering*, 4(5), 1992.

[5] O. Benjelloun, A. D. Sarma, A. Halevy, and J. Widom. Uldbs: Databases with uncertainty and lineage. In *IN VLDB*, pages 953–964, 2006.

[6] R. Brent. A partnership between biology and engineering. *Nature Biotechnology*, 22:1211–1214, 2004.

[7] C. Bhm, A. Pryakhin, and M. Schubert. The gauss-tree: Efficient object identification in databases of probabilistic feature vectors. In *In Proc. ICDE*, 2006.

[8] A. Carpenter. Software opens the door to quantitative imaging. *Traffic*, 4(2):120–121, 2007.

[9] Center for bioimage informatics. http://bioimage.ucsb.edu/.

[10] R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J. S. Vitter. Efficient indexing methods for probabilistic threshold queries over uncertain data. In *Proc. 30th VLDB*, 2004.

[11] N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. *The VLDB Journal*, 16(4):523–544, Oct 2007.

[12] D. Dey and S. Sarkar. A probabilistic relational model and algebra. *ACM Trans. on Database Systems*, 21(3):339369, 1996.

[13] S. A. Goff et al. The iplant collaborative: Cyberinfrastructure for plant biology. *Frontiers in Plant Science*, 2(00034), 2011.

[14] Hadoop and HDFS. http://hadoop.org/.

[15] H. G. S. C. International. Finishing the euchromatic sequence of the human genome. *Nature*, 431, October 2004.

[16] iRods: Data grids, digital libraries, persistent archives, and real-time data. http://irods.org/.

[17] K. Kvilekval, D. Fedorov, B. Obara, A. Singh, and B. Manjunath. Bisque: A platform for bioimage analysis and management. *Bioinformatics*, 26(4):544–552, Feb 2010.

[18] L. V. S. Lakshmanan, N. Leone, R. B. Ross, and V. S. Subrahmanian. Probview: A flexible probabilistic database system. *ACM Transaction on Database Systems*, 22(3):419469, 1997.

[19] N. Miller, T. D. Brooks, A. Assadi, and E. Spalding. Detection of a gravitropism phenotype in glutamate receptor-like 3.3 mutants of arabidopsis thaliana using machine vision and computation. *Genetics*, 186:585–593, 2010.

[20] NIH data sharing policy. http://grants1.nih.gov/grants/policy/data_sharing/.

[21] Orion project. http://orion.cs.purdue.edu/.

[22] C. Pautasso, O. Zimmermann, and F. Leymann. Restful web services vs. "big"' web services: making the right architectural decision. In *Proc. of 17th Int. Conf. on World Wide Web*, pages 805–814. ACM, 2008.

[23] J. Pei, B. Jiang, X. Lin, and Y. Yuan. Probabilistic skylines on uncertain data. In *Proc. 33rd Int. Conf. on VLDB*, 2007.

[24] H. Peng. Bioimage informatics: a new area of engineering biology. *Bioinformatics*, 24(17):1827–1836, 2008.

[25] D. Piedrahita et al. Silencing of cdk5 reduces neurofibrillary tangles in transgenic alzheimer's mice. *Journal of Neuroscience*, 30(42):13966–13976, 2010.

[26] C. Re, N. Dalvi, and D. Suciu. Efficient top-k query evaluation on probabilistic data. In *ICDE'07*, pages 886–895, 2007.

[27] A. D. Sarma, O. Benjelloun, A. Y. Halevy, and J. Widom. Working models for uncertain data. In *ICDE*, page 7, 2006.

[28] M. Schilling, A. C. Pfeifer, S. Bohl, and U. Klingmuller. Sharing images. *Standardizing experimental protocols*, 19(4):354–359, 2008.

[29] P. Sen and A. Deshpande. Representing and querying correlated tuples in probabilistic databases. In *ICDE*, pages 596–605, 2007.

[30] J. Shi and C. Tomasi. Good features to track. In *CVPR*, pages 593–600, jun 1994.

[31] M. Soliman, I. Ilyas, and K. Chang. Top-k query processing in uncertain databases. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 896 –905, april 2007.

[32] J. Swedlow, I. Goldberg, E. Brauner, and P. Sorger. Informatics and quantitative analysis in biological imaging. *Science*, 300(5616):100–102, 2003.

[33] J. Swedlow, S. Lewis, and I. Goldberg. Modelling data across labs, genomes, space and time. *Nature Cell Biology*, 8(11):1190–1194, 2006.

[34] Y. Tao, R. Cheng, X. Xiao, W. K. Ngai, B. Kao, and S. Prabhakar. Indexing multi-dimensional uncertain data with arbitrary probability density functions. In *Proc. of 31st Int. Conf. on VLDB*, pages 922–933, 2005.

[35] D. Thain, T. Tannenbaum, and M. Livny. Distributed computing in practice: the condor experience. *Concurrency - Practice and Experience*, 17(2-4):323–356, 2005.

[36] R. Tuan and C. Lo, editors. *Developmental Biology Protocols, Volume I*, volume 137 of *Methods in Molecular Biology*. Humana Press, November 1999.

[37] M. W. Vannier and R. M. Summers. Sharing images. *Radiology*, 228:23–25, 2003.

[38] J. Widom. Trio: A system for integrated management of data, accuracy, and lineage. In *CIDR'05*, pages 262–276, 2005.