CAREER: Evolving and Self-Managing Data Integration Systems
Project Summary

The problem of integrating data from multiple sources has been a long standing challenge for the database community. Many architectures for data integration have been proposed, with an important one being *virtual data integration systems*. Such systems provide a *global query interface* over a multitude of data sources, thereby freeing users from the tedious job of £nding and querying the relevant sources. Virtual data integration systems have the potential to revolutionize the way we access data on the Web and at enterprises. **Unfortunately, today such systems are still extremely hard to build and costly to maintain.** They must be taught in tedious detail how to interact with the data sources and understand their languages. In dynamic environments such as the Web, sources constantly change their presentation and data formats. Hence, once deployed the systems must still be under continuous supervision and told how to deal with the changes. The laborious teaching and supervision incur huge expenses, and severely limit the deployment of data integration systems in practice.

To break this limitation, a compelling solution is to build data integration systems that learn to *evolve and self manage over time*, with minimal human intervention. This vision £ts the emerging paradigm of self-tuning databases and autonomic computing: the growing complexity of computing systems is overwhelming our management capabilities, hence we must build systems that manage themselves.

The goal of this proposal is to make fundamental contributions toward realizing the above vision, drawing from the core idea that to evolve and self manage, a data integration system can learn from numerous types of information and entities in its environment, including itself, past system development activities, data in the domain, behaviors of systems and data sources, and even from the multitude of users. The central challenges that I attack are: *(a) how can we effectively automate key labor-intensive tasks, such as schema matching, global schema creation, and duplicate detection? (b) how can a system detect failures due to changes at the sources, with minimal human intervention? (c) can a system further reduce the tremendous data integration burden of the system admins by spreading the burden thinly over the mass of users?* I have carried out preliminary research on several of the above challenges, with promising results. I will build on my extensive background in data integration and machine learning to develop innovative solutions to these problems, then integrate them to build and evaluate evolving and self-managing data integration systems. I will evaluate the systems in the context of the Deep Web (book and real-estate data sources), Surface Web (CS department websites), and a real-world organization (the Illinois Fire Service Institute).

My education plan aims to broaden the current UIUC database curriculum to include data integration issues, and develop a new course on interdisciplinary paradigms for data management, which interacts closely with the proposed research. To break the perceived isolation of Midwest universities, I will initiate a program that enable graduate students to visit different universities and give talks, thereby facilitating knowledge sharing and communication skills. I plan also to leverage this research and work with the Illinois Fire Service Institute to tackle their acute data integration problems, and educate rural Illinois £re £ghters in using the resulting systems.

Intellectual Merit: The project will take a next logical step in data integration research, and make fundamental advances in the current state of the art. It combines database and AI techniques to attack critical data management problems. It brings conceptually novel solutions (e.g., mass collaboration) to fundamental issues underlying virtually any data integration or sharing efforts (e.g., semantic heterogeneity). The project results have the potential for application to the £eld of autonomic computing.

Broader Impacts: The project will facilitate the widespread deployment of data integration systems, thus resulting in more effective information management and access for society. It plays an integral part in educating next-generation professional workers and researchers. I have graduated 2 female Masters students and am working closely with 3 female students and 1 African-American student. This research will enable me to continue the vigorous training of students from underrepresented groups. The research will help integrate data for rural Illinois £re £ghters, and train them in access and use. Finally, data and system artifacts from the project will be disseminated broadly in the research community, to enhance the infrastructure for research and education.