Database Creation and Information Extraction from ETDs

Lamont Banks

Joseph Luke

Client: Venkat Srinivasan

Virginia Tech, Blacksburg, VA

CS 4624, Spring 2013

Project Description

Create a database of data extracted from graduate dissertations to be used for determining why students pursue research.

CRA – Computing Research Association

ETD – Electronic Thesis and Dissertation

XML Tags

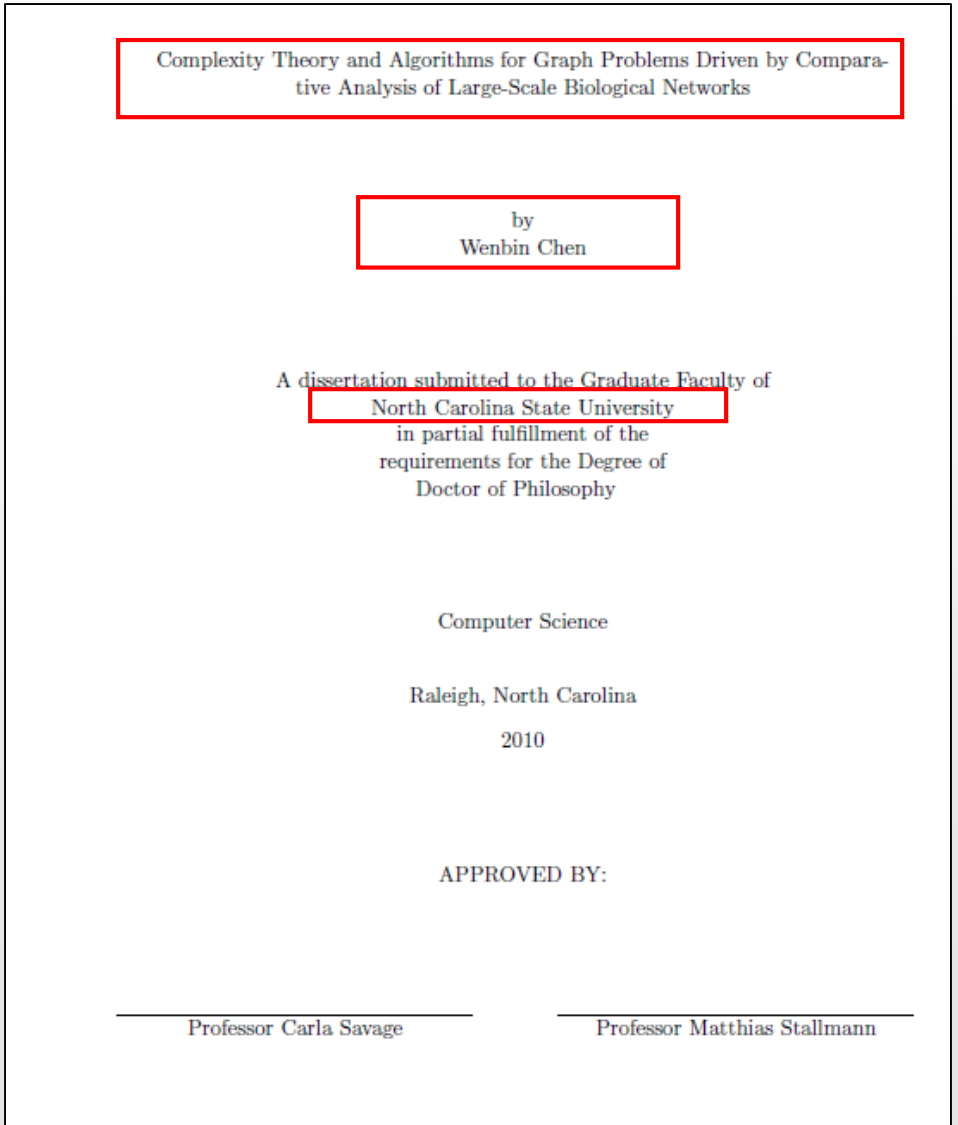
```
<?xml version="1.0"?>

<!-- Basic Information -->
<title />
<research_category />
<author />
<institution />
<submission_year />
<advisor />

<!-- Undergraduate Information -->
<undergraduate_institute_name />
<undergraduate_institute_degree />

<!-- Social -->
<num_acknowedgements_colleagues />
<num_acknowledged_friends />

<!-- Research funding -->
<acknowledged_funding_org_1 />
<acknowledged_funding_org_2 />
<acknowledged_funding_org_3 />

<!-- Research History -->
<prior_research_area_1 />
<prior_research_area_2 />

<!-- Work History -->
<prior_workplace_1 />
<prior_workplace_2 />
<prior_worlplace_3 />

<!-- Origins -->
<country />
<city />
<province_state />
```

Tagging Documents

- ETD Resources
  - University repositories
    - Virginia Tech
    - North Carolina State University
    - Wake Forest University, Florida State University, etc.
- Manual data extraction
- Challenges
  - Missing data
    - No resume, acknowledgements, biography
  - Insufficient information

General ETD Structure

1. Abstract
2. Title Page
3. Biography
4. Acknowledgements
5. Resume



Complexity Theory and Algorithms for Graph Problems Driven by Comparative Analysis of Large-Scale Biological Networks

by
Wenbin Chen

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Computer Science

Raleigh, North Carolina

2010

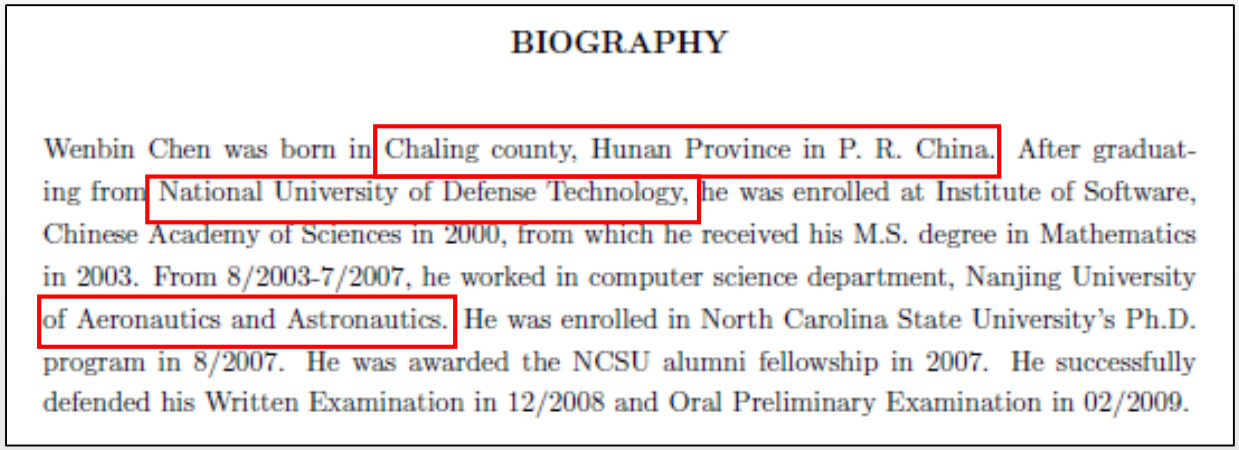
APPROVED BY:

Professor Carla Savage

Professor Matthias Stallmann

General ETD Structure (cont'd)

1. Abstract
2. Title Page
3. Biography

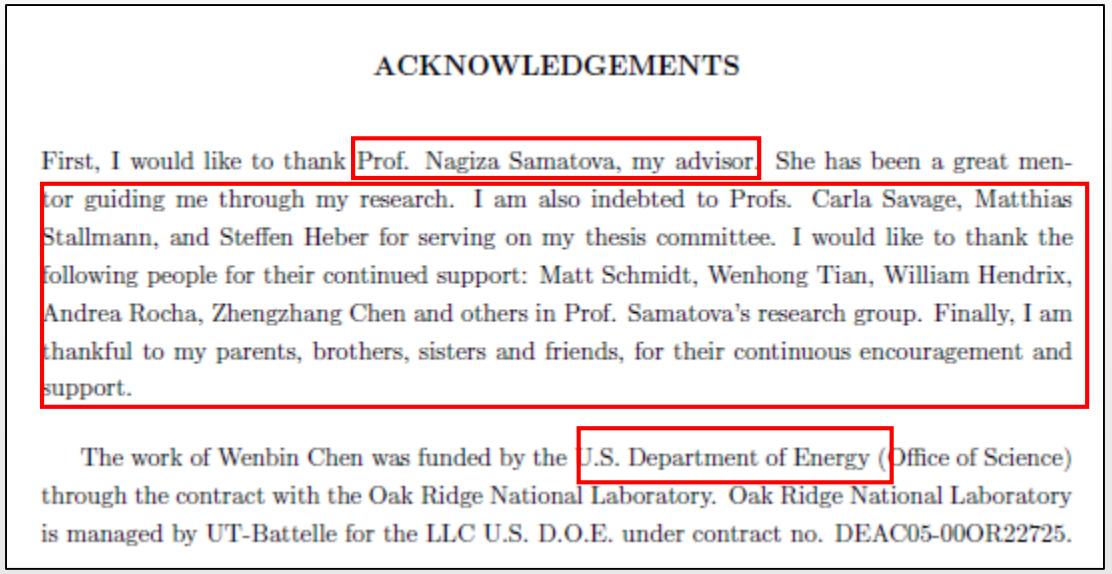


BIOGRAPHY

Wenbin Chen was born in Chaling county, Hunan Province in P. R. China. After graduating from National University of Defense Technology, he was enrolled at Institute of Software, Chinese Academy of Sciences in 2000, from which he received his M.S. degree in Mathematics in 2003. From 8/2003-7/2007, he worked in computer science department, Nanjing University of Aeronautics and Astronautics. He was enrolled in North Carolina State University's Ph.D. program in 8/2007. He was awarded the NCSU alumni fellowship in 2007. He successfully defended his Written Examination in 12/2008 and Oral Preliminary Examination in 02/2009.

General ETD Structure (cont'd)

1. Abstract
2. Title Page
3. Biography
4. Acknowledgements
5. Resume (not often)



ACKNOWLEDGEMENTS

First, I would like to thank Prof. Nagiza Samatova, my advisor. She has been a great mentor guiding me through my research. I am also indebted to Profs. Carla Savage, Matthias Stallmann, and Steffen Heber for serving on my thesis committee. I would like to thank the following people for their continued support: Matt Schmidt, Wenhong Tian, William Hendrix, Andrea Rocha, Zhengzhang Chen and others in Prof. Samatova's research group. Finally, I am thankful to my parents, brothers, sisters and friends, for their continuous encouragement and support.

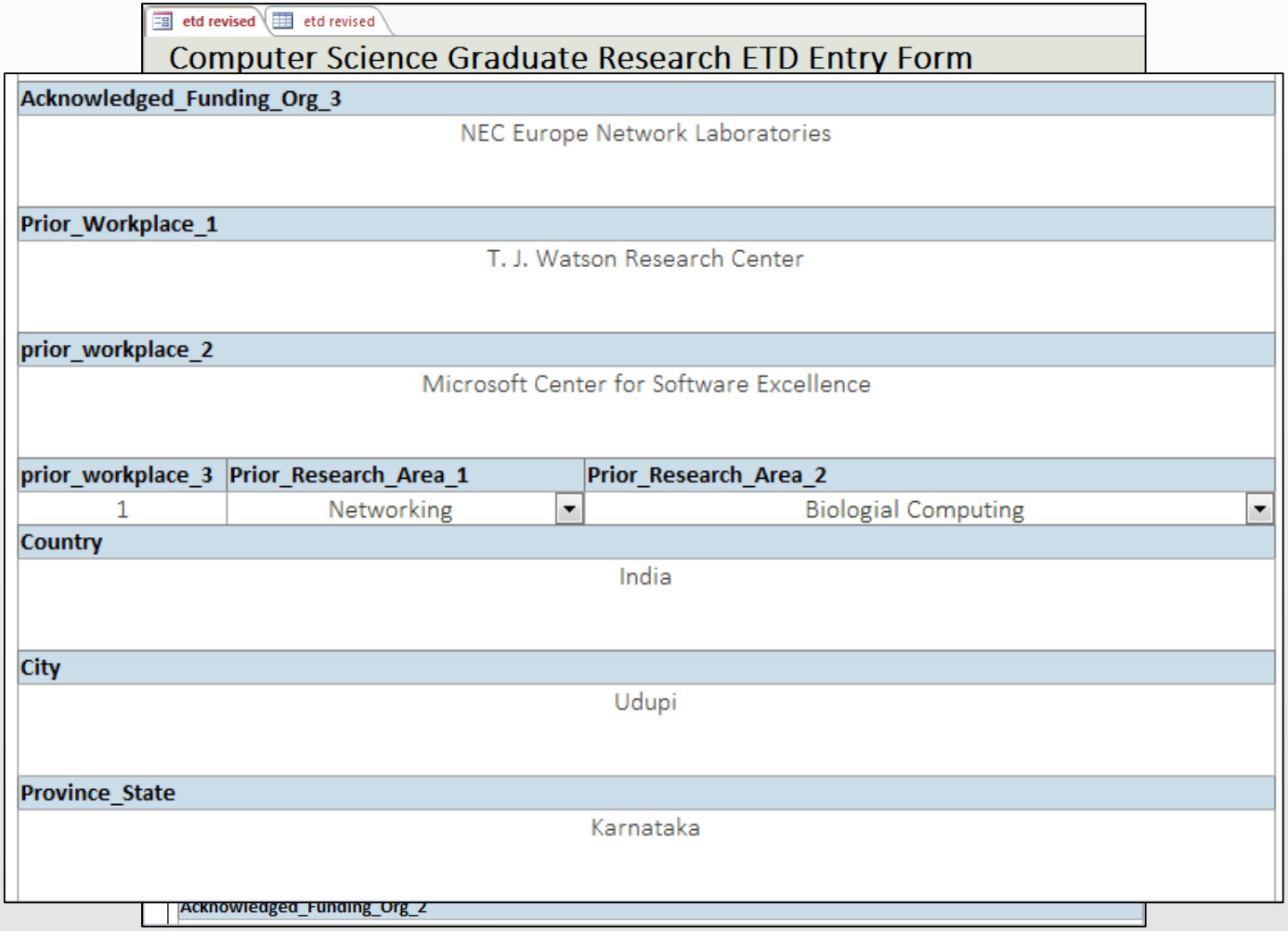
The work of Wenbin Chen was funded by the U.S. Department of Energy (Office of Science) through the contract with the Oak Ridge National Laboratory. Oak Ridge National Laboratory is managed by UT-Battelle for the LLC U.S. D.O.E. under contract no. DEAC05-00OR22725.

Database

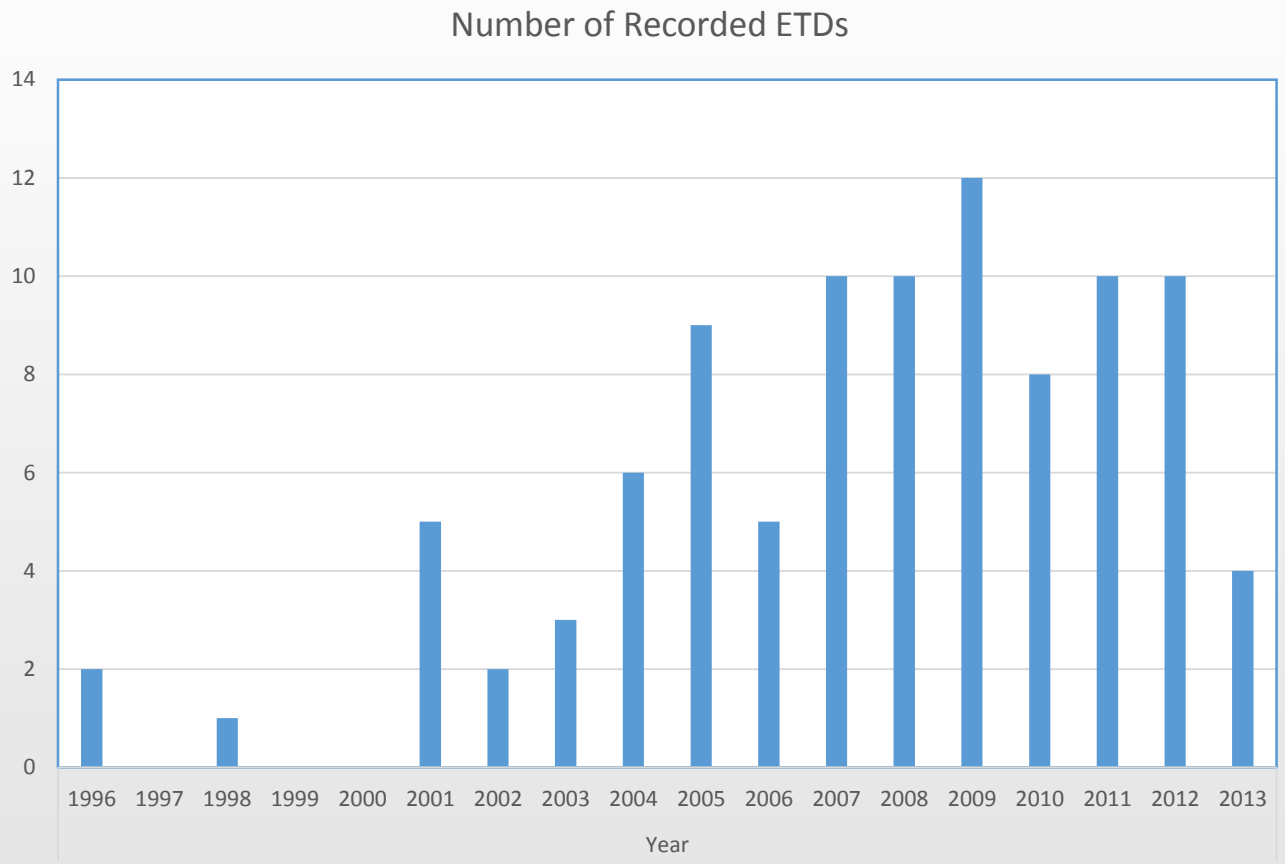
- Microsoft Access 2013
  - Convenience
  - Not limited to Access
- ~100 ETDs currently stored

g API Specifications from Source Code for I	Data Managem	Mithun Puthige Acharya	North Carolina State University	2009	Tao Xie
ing Efficient SQL Sequences Via Prefetching	Data Managem	Ahmet S. Bilgin	North Carolina State University	2007	Dr. Munindar P. Singh
ATIVE GENERATION:	Human Compu	MARK OWEN RIEDL	North Carolina State University	2004	Michael Capps
agent Referral Systems: Maintaining and Ap	Software Engin	Narendran Ranjit	North Carolina State University	2007	Dr. Munindar Singh
limination of Overheads due to Type Annot	Data Managem	Cohan Sujay Carlos	North Carolina State University	2002	Dr. Purushothaman Iyer
aging Multiple Mechanisms for Informatior		Andrew W. Wicker	North Carolina State University	2012	Jon Doyle
formance Review Strategy for Regulating S	Security	Raoul Jetley	North Carolina State University	2006	Dr. Purush Iyer
Centric Scheduling Strategies for Workflow	Software Engin	Yang Zhang	Rice University	2009	Ken Kennedy
THODOLOGY USING ASSISTIVE SKETCH RECO	Human Compu	DANIEL MEYER DIXON	Texas A&M University	2009	
PARALLELIZATION STUDIES IN BIO-MOLECUL	Biologial Comp	LEI JI	THE FLORIDA STATE UNIVERSITY	2006	
HICAL VISUALIZATION OF ARCHITECTURAL S	Media and Visu	KELLEY C. JONES	THE FLORIDA STATE UNIVERSITY	2007	Dr. Gary Tyson
ration of Graphical User Interface and Data	Human Compu	Dhananjay Mishra	Virginia Polytechnic Institute and State University	2004	Dr. Cli♪ord Sha♪er
tural Design Using Cellular Automata	Biologial Comp	Douglas J. Slotta	Virginia Polytechnic Institute and State University	2001	
ceptual Framework for Specification of	Networking	DZMITRY CHURBANAU	Virginia Polytechnic Institute and State University	2010	Dr. Osman Balci
: A Java-Based Framework for	Software Engin	Philip L. Isenhour	Virginia Polytechnic Institute and State University	1998	Professor Cliff Shaffer
rds the Development of User Interface Desi	Human Compu	Khaled Hussein	Virginia Polytechnic Institute and State University	2007	Dr. McCrickard
al Issues in the Processing of cDNA	Biologial Comp	Vincent Jouenne	Virginia Polytechnic Institute and State University	2001	Dr. Craig A. Struble
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

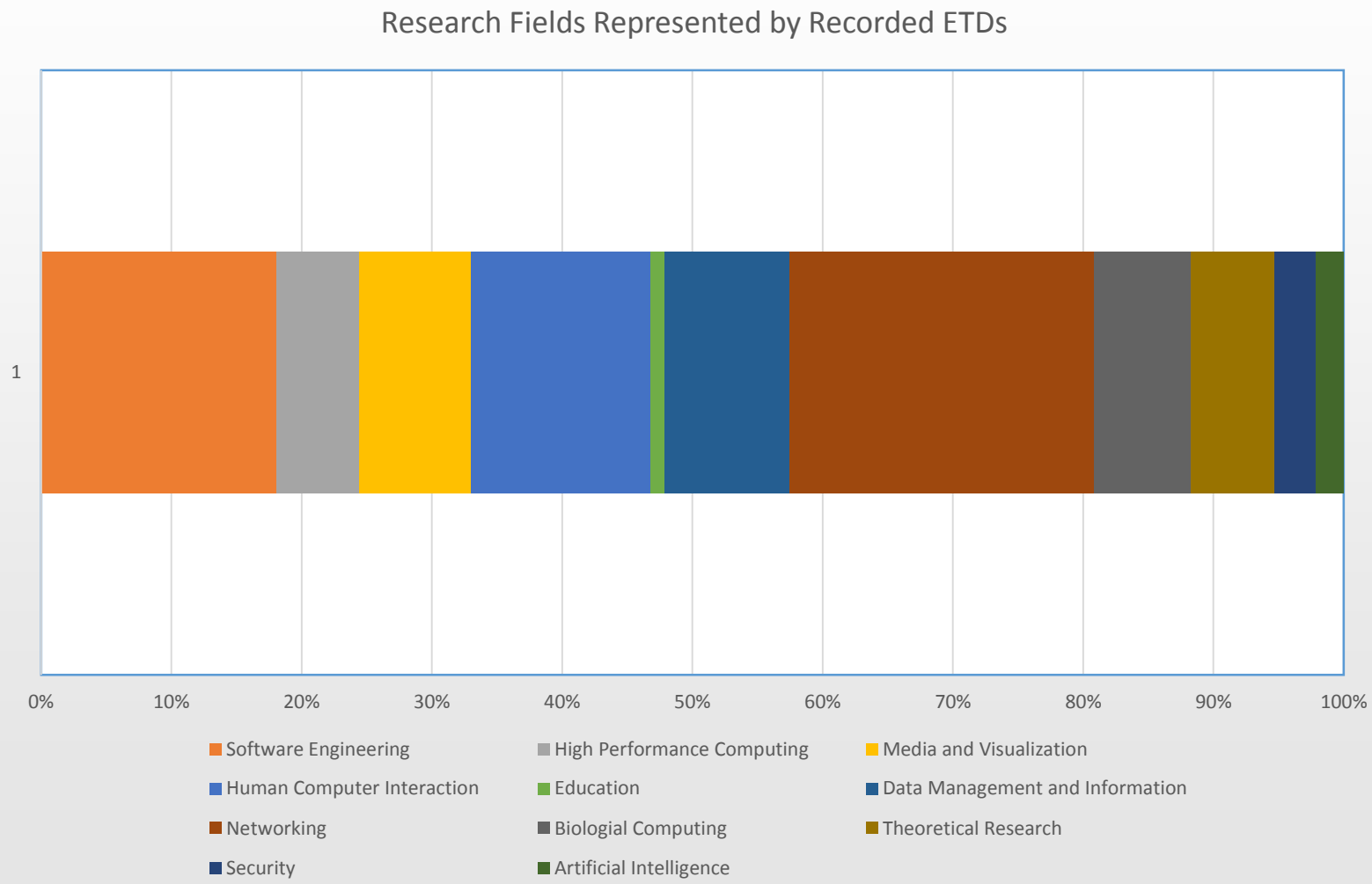
Database (cont'd)



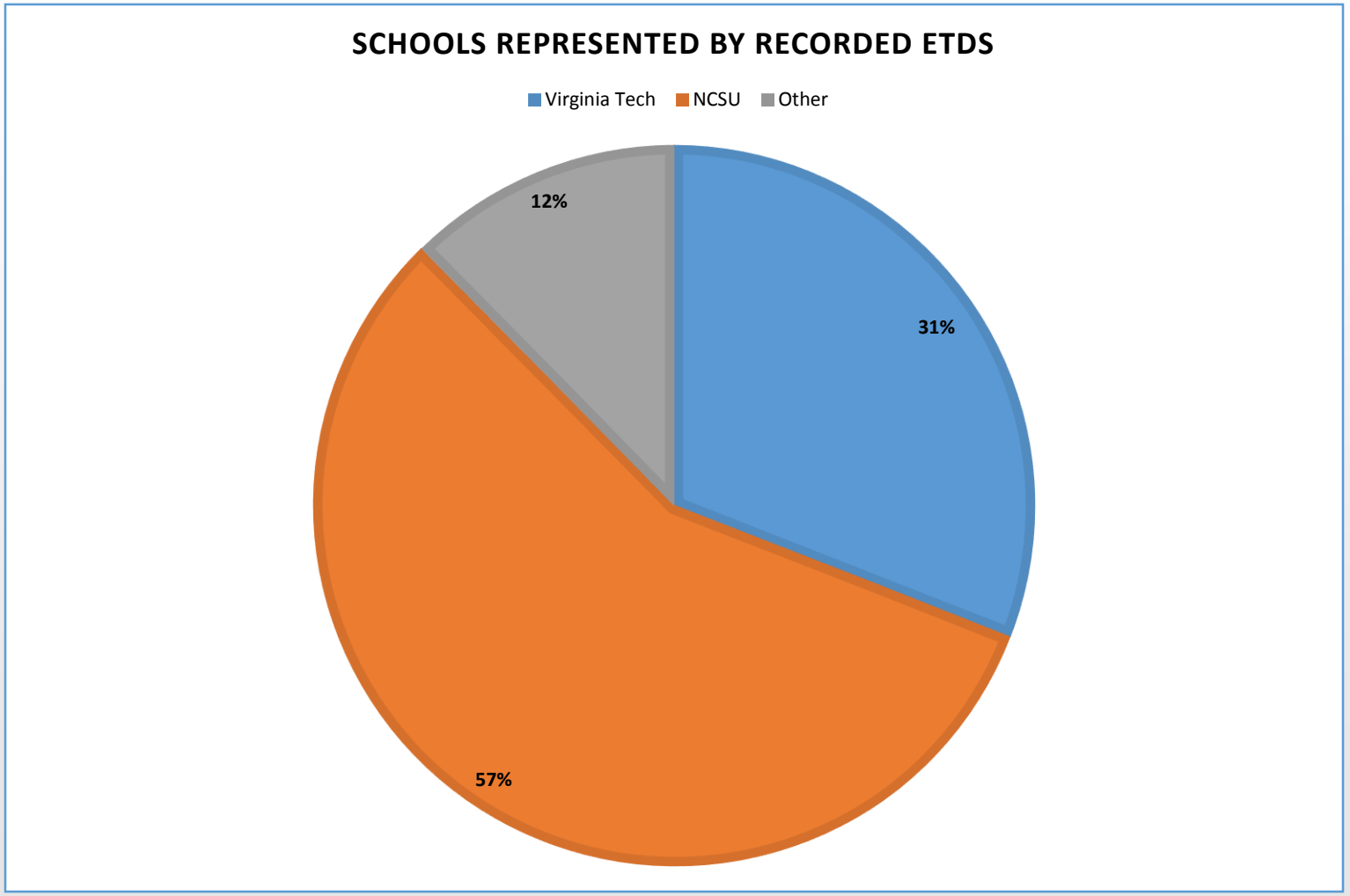
Database Statistics (cont'd)



Database Statistics (cont'd)



Database Statistics (cont'd)



Future

- More entries
- Expand database structure
- Expand XML tags
- Process using WEKA

Questions?