

- EDUCATION** *University at Buffalo, The State University of New York* May 2017 (expected)
Ph.D. in Computer Science Engineering, GPA: 3.8/4.0
Advisors: Oliver Kennedy, Jan Chomicki.
- University at Buffalo, The State University of New York* June 2012
Master, Computer Science Engineering, GPA: 3.78/4.0
- PROFESSIONAL EXPERIENCE** *Internship – Oracle labs (HQ)* May 2016 – Aug. 2016
- Developed prototypes of data guide and data cleaning algorithms (Schema Inference, Schema Validation, Domain Constraint Repair, Schema Matching, Automatic cleaning) in JavaScript as table functions within SQL query.
 - Proposed a general solution for automatic generation of normalized schemas for nested key-value data.
 - Conducted performance experiments against Oracle built-ins and across V8, GraalJS, SVM and Walnut platforms.
- Internship – National Center for Atmospheric Research* May 2014 – Aug. 2014
- Designed database schemas, shredded log data of super computers and developed metrics of job size and CPU consumptions for open source XDMoD and contributed back to open source.
 - Conducted exploratory data analysis on the log data using R.
 - Analyzed time-series log data using exponential smoothing and ARIMA for predicting CPU consumptions.
- Software Engineer, SONY* Jan. 2010 – July 2010
Developed and maintained the internal system of Sony Group-Offshore System using *JavaEE* and *Oracle 10g* by communicating with customers and understanding their requirements.
- SIGMOD Reproducibility Committee* 2016
Served as one of Reproducibility Committee in ACM SIGMOD 2016 conference.
- TEACHING** *Course Project Assistant* 2011-2012
- Frequency analysis search engine: Built a word frequency analysis search engine for large data using Hadoop MapReduce framework running on Amazon EC2.
 - Project: Name2Face: Developed a cloud application which consumed the Microsoft cloud services (blob, queue and table) using C#, .Net
- Teaching Assistant*
- CSE 405, Programming Language
 - CSE 505, Programming Language
 - CSE 587, Data Intensive Computing
 - CSE 421/521, Operating System
 - CSE 462 Database
- PROJECTS** *ETL, Data Cleaning, Probabilistic database, Query Optimization, Data Mining* Sept. 2012 - May 2016
[Mimir: An On-Demand Approach to ETL \(Funded by Oracle\)](http://mimirdb.info) (<http://mimirdb.info>)
Design and implementation of a general, extensible infrastructure for on-demand data curation tasks based on probabilistic query processing. My contribution involves:
- Designed a modular framework called lens for handling data cleaning task in a probabilistic fashion.
 - Designed several optimization techniques for probabilistic query processing.
 - Provided analysis of several heuristics for estimating the quality and the costs of data cleaning processes.
 - Proposed an algorithm of ranking user feedback in a cost-efficient fashion.
 - Developed a Java prototype system.
 - Guided the development and research of Mimir as a team leader.

- This work is published in VLDB 2015 in the research track.

Query Processing, Machine Learning

May 2015 – Sept. 2016

Convergent Inference Algorithms in Probabilistic Graphical Model

Proposed a new family of inference algorithms called *Convergent Inference algorithms (CIAs)*, which enjoy the benefits of both exact and approximate inference algorithms by providing approximate results over the course of inference, and eventually converging to an exact inference result. My contribution involves:

- Proposed a base CIA that performs pseudo-random sampling without replacement using Linear Congruential Generators.
- Proposed a new online aggregation algorithm called *Leaky Joins* that produces samples of a query's result in the course of normally evaluating the query.
- Provided analysis of time complexity and confidence bounds for CIAs.
- Generalized the algorithms for any aggregate queries over small but dense tables.
- Implemented Variable Elimination, Belief Propagation and Gibbs Sampling algorithms as comparisons.

Information retrieval, Java, OpenMP, C++, Mysql.

Jan. 2011 – Jan. 2012

- Developed a search engine by tokenization, building inverted index, query processing and ranking using tf-idf scores and PageRank.
- *iRec* [[pdf](#)]. Designed and implemented a recommendation system developed using Semantic Analysis and Web Mining concepts.

Machine Learning, Matlab.

2014 - 2015

- Performed handwriting recognition using logistic regression, SVM, neural network.
- Conducted Bayesian Network structure learning and missing value imputation using expectation-maximization principle for handwriting recognition applications.

Data integration, RDF, SPARQL, Jena.

Jan. 2012 – May 2013

- Translated relational database to RDF database and SQL to SPARQL.
- Conducted SPARQL query optimization.

Distributed system, Java, Mysql, Google Maps API.

Sept. 2011 - Feb. 2012

- *Weather Forecast Station*: Designed and implemented an Amazon EC2 deployable multi-tier distributed system using SOAP protocol and REST architectures.
- Frequency analysis search engine: Built a word frequency analysis search engine for large data using Hadoop MapReduce framework running on Amazon EC2.
- Name2Face: Developed a cloud application which consumed the Microsoft cloud services using C#, .Net

Database, Java.

Sept.2011 – Jan. 2012

- Implemented SQL parsing, schema management, query plan generation and query evaluator with support for general database operations.
- Optimized query plan using pushing down selections, external sort, hash and merge join.
- Built a B-tree index.

Software Development, MVC, JavaEE, MySQL, C#.NET, XML, SQL Server.

- Developed CampusLootr (an ebay-like website for university students) as a server-side developer. 2013
- Developed Game called Memorize and implemented real-time interaction with users. 2009

SKILLS

- Java (proficient), C#, C++, SQL, PL/SQL, Matlab, JavaScript, Node.js, R, MVC, JSP, Java Servlet, RESTful Web Services, XML, RDF, SPARQL, XQUERY, ASP.NET, Hadoop MapReduce, OpenMP, MPI.
- Oracle, MySQL, SQL Server.
- Eclipse, WebStorm, Visual Studio, Git, Tomcat, Amazon EC2, S3, Windows Azure

PUBLICATIONS

- [1] **Ying Yang**, Oliver Kennedy. Convergent Inference with Leaky Joins. In: EDBT 2017
- [2] W. Spoth, B. S. Arab, E. S. Chan, D. Gawlick, A. Ghoneimy, B. Glavic, B. Hammerschmidt, O. Kennedy, S. Lee, Z.H.Liu, X. Niu, **Y. Yang**. Adaptive Schema Databases. In: CIDR 2017
- [3] **Ying Yang**, Niccolo Meneghetti, Ronny Fehling, Zhen Hua Liu, Oliver Kennedy. Lenses: An On-Demand Approach to ETL. In: Proceedings of the VLDB 2015
- [4] **Ying Yang**, Oliver Kennedy, Jan Chomicki. On-Demand Query Result Cleaning. In: VLDB 2014
- [5] Oliver Kennedy, **Ying Yang**, Jan Chomicki, Ronny Fehling, Zhen Hua Liu, and Dieter Gawlick. Detecting the Temporal Context of Queries. In: VLDB BIRTE 2014

AWARDS

- UB Poster Awards (2016)**
- National Scholarship of China (2006-2007, 2007-2008)**
- Award for excellence in Java Programming Competition (2007)**