Dong Deng

Postdoctoral Associate, CSAIL
Massachusetts Institute of Technology
32-G904B, 32 Vassar Street
Cambridge, MA 02139

Email: dongdeng@csail.mit.edu
http://people.csail.mit.edu/dongdeng

Research Interests

Data integration, data management, data science, data cleaning, information retrieval, program synthesis, and database usability with a focus on developing end-to-end systems and proposing scalable algorithms to manage real-world data, including but not limited to: relational data, open data, web tables, enterprise data, data warehouse, logs, and knowledge bases.

Current Position

Massachusetts Institute of Technology — Cambridge, MA, USA
Postdoctoral Associate, Database Group at CSAIL — *Jul 2016 - Present*
Project: DataCivilizer: An End-to-End Data Integration System
Mentors: Michael Stonebraker and Samuel Madden

Education

Tsinghua University — Beijing, China
Ph.D., Computer Science — *Sep 2011 - Jun 2016*
Thesis: Error-Tolerant Big Data Processing
Advisor: Guoliang Li

Beihang University — Beijing, China
B.Sc., Computer Science — *Sep 2007 - Jun 2011*
Rank: 1/144

Professional Experience

Qatar Computing Research Institute — Doha, Qatar
Research Associate, Data Analytics Group — *Dec 2015 - Mar 2016*
Mentor: Mourad Ouzzani

University of Michigan — Ann Arbor, MI, USA
Research Assistant, Database Group at EECS — *Jan 2014 - Jun 2014*
Mentor: H. V. Jagadish

Selected Honors and Awards

2016 The Highest Doctoral Dissertation Award in Tsinghua University (only 26 in all disciplines)
2016 Beijing Municipal Government Award for "Outstanding PhD Graduates"
2016 Tsinghua University Excellent PhD Graduate
2015 Microsoft Research Asia PhD Fellowship (Only 13 PhD students in Asia)
2014 Google PhD Fellowship (Only 4 PhD students in China)
2014 Boeing Scholarship (Only 12 Students in Tsinghua University)
2014 Baidu Scholarship Finalist (Only 20 PhD students worldwide)
2014 Tsinghua University Academic Rising Star (only 21 in all disciplines)
2014 IEEE ICDE Student Travel Scholarship
2013 First Place in the String Similarity Join Competition held by EDBT/ICDT
2013 Siebel Scholar (Only 85 from the best universities, including MIT, Stanford, and UC Berkeley)
2013 TCDE Student Travel Award (Only 1 Recipient)
2013 ACM SIGMOD 2013 Programming Contest Finalist
2012 National Scholarship
2012 Intel Fellowship (Rank 1st in Tsinghua Computer Science Department)
2011 Beijing Municipal Government Award for "Outstanding Graduates"
2011 & 2014 ACM SIGMOD Student Travel Award

Teaching Experience

Course: Open Data Science — University of Toronto
Guest Lecturer — *Fall 2017*

Course: Database Systems — Tsinghua University
Teaching Assistant — *Fall 2015*

Course: Advanced Topics in Database Systems — Tsinghua University
Teaching Assistant — *Fall 2012-2014*

Publications

Full Research Papers

1. **Dong Deng**, Wenbo Tao, Ziawasch Abedjan, Ahmed Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, Michael Stonebraker, Nan Tang: *Entity Consolidation: The Golden Record Problem.* Computing Research Repository (**CoRR**), abs/1709.10436, Under Review in **SIGMOD'18**

2. **Dong Deng**, Yufei Tao, Guoliang Li: *Overlap Set Similarity Joins with Theoretical Guarantees.* Proc. 2018 ACM SIGMOD Int. Conf. on Management of Data (**SIGMOD'18**), Houston, TX, USA, June 2018: Accepted.

3. Wenbo Tao, **Dong Deng**, Michael Stonebraker: *Approximate String Joins with Abbreviations.* PVLDB 11(1): 53-65, 2018. Also, 44th Int. Conf. on Very Large Data Bases (**VLDB'18/PVLDB**), Rio de Janeiro, Brazil, Aug. 2018.

4. **Dong Deng**, Albert Kim, Samuel Madden, Michael Stonebraker: *SilkMoth: An Efficient Method for Finding Related Sets with Maximum Matching Constraints*. PVLDB 10(10): 1082-1093, 2017. Also, 43rd Int. Conf. on Very Large Data Bases (**VLDB'17/PVLDB**), Munich, Germany, Aug. 2017.

5. Minghe Yu, Jin Wang, Guoliang Li, Yong Zhang, **Dong Deng**, Jianhua Feng: *A Unified Framework for String Similarity Search with Edit-Distance Constraint*. The International Journal on Very Large Data Bases (**VLDB Journal**), 26(2): 249-274 (2017).

6. **Dong Deng**, Raul Castro Fernandez, Ziawasch Abedjan, Sibo Wang, Michael Stonebraker, Ahmed K. Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, Nan Tang: *The Data Civilizer System*. 2017, 8th Biennial Conference on Innovative Data Systems Research (**CIDR'17**), Chaminade, CA, USA, Jan. 2017: Online Proceedings.

7. **Dong Deng**, Guoliang Li, He Wen, H. V. Jagadish, Jianhua Feng: *META: An Efficient Matching-Based Method for Error-Tolerant Autocompletion*. PVLDB 9(10): 828-839, 2016. Also, 42nd Int. Conf. on Very Large Data Bases (**VLDB'16/PVLDB**), New Delhi, India, Sep. 2016.

8. Ziawasch Abedjan, Xu Chu, **Dong Deng**, Raul Castro Fernandez, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker, Nan Tang: *Detecting Data Errors: Where Are We and What Needs to Be Done?*. PVLDB 9(12): 993-1004, 2016. Also, 42nd Int. Conf. on Very Large Data Bases (**VLDB'16/PVLDB**), New Delhi, India, Sep. 2016.

9. Michael Stonebraker, **Dong Deng**, Michael L. Brodie: *Database Decay and How to Avoid It*. 2016 IEEE Int. Conf. on Big Data (**BigData'16**), Washington DC, USA, Dec. 2016: 7-16.

10. Chengliang Chai, Guoliang Li, Jian Li, **Dong Deng**, Jianhua Feng: *Cost-Effective Crowdsourced Entity Resolution: A Partial-Order Approach*. Proc. 2016 ACM SIGMOD Int. Conf. on Management of Data (**SIGMOD'16**), San Francisco, CA, USA, June, 2016: 969-984.

11. **Dong Deng**, Guoliang Li, He Wen, Jianhua Feng: *An Efficient Partition Based Method for Exact Set Similarity Joins*. PVLDB 9(4): 360-371, 2015. Also, 42nd Int. Conf. on Very Large Data Bases (**VLDB'16/PVLDB**), New Delhi, India, Sep. 2016.

12. **Dong Deng**, Guoliang Li, Jianhua Feng, Yi Duan, Zhiguo Gong: *A Unified Framework for Approximate Dictionary-based Entity Extraction*. The International Journal on Very Large Data Bases (**VLDB Journal**), 24(1): 143-167 (2015).

13. Jin Wang, Guoliang Li, **Dong Deng**, Yong Zhang, Jianhua Feng: *Two Birds with One Stone: An Efficient Hierarchical Framework for Top-k and Threshold-based String Similarity Search*. 2015 IEEE Int. Conf. on Data Engineering (**ICDE'15**), Seoul, South Korea, Apr. 2015: 519-530.

14. Guoliang Li, Jian He, **Dong Deng**, Jian Li: *Efficient Similarity Join and Search on Multi-Attribute Data*. Proc. 2015 ACM SIGMOD Int. Conf. on Management of Data (**SIGMOD'15**), Melbourne, Victoria, Australia, May 2015: 1137-1151.

15. Wenfei Fan, Xin Wang, Yinghui Wu, **Dong Deng**: *Distributed Graph Simulation: Impossibility and Possibility*. PVLDB 7(12), 1083-1094, 2014. Also, 40th Int. Conf. on Very Large Data Bases, (**VLDB'14/PVLDB**), Hangzhou, China, Sep. 2014.

16. **Dong Deng**, Guoliang Li, Shuang Hao, Jiannan Wang, Jianhua Feng: *MassJoin: A MapReduce-based Method for Scalable String Similarity Joins*. 2014 IEEE Int. Conf. on Data Engineering (**ICDE'14**), Chicago, IL, Mar. 2014: 340-351.

17. **Dong Deng**, Guoliang Li, Jianhua Feng: *A Pivotal Prefix based Filtering Algorithm for String Similarity Search*. Proc. 2014 ACM SIGMOD Int. Conf. on Management of Data (**SIGMOD'14**), Snowbird, UT, USA, June 2014: 673-684.

18. **Dong Deng**, Yu Jiang, Guoliang Li, Jian Li, Cong Yu: *Scalable Column Concept Determination for Web Tables Using Large Knowledge Bases*. PVLDB 6(13), 1606-1617, 2013. Also, 40th Int. Conf. on Very Large Data Bases, (**VLDB'14/PVLDB**), Hangzhou, China, Sep. 2014.

19. Guoliang Li, **Dong Deng**, Jianhua Feng: *A Partition-based Method for String Similarity Joins with Edit-Distance Constraints*. ACM Transactions on Database Systems (**TODS**), 38(2) 9:1-9:33 (2013).

20. **Dong Deng**, Guoliang Li, Jianhua Feng, Wen-Syan Li: *Top-k String Similarity Search with Edit-Distance Constraints*. 2013 IEEE Int. Conf. on Data Engineering (**ICDE'13**), Brisbane, Australia, Apr. 2013: 925-936.

21. **Dong Deng**, Guoliang Li, Jianhua Feng: *An Efficient Trie-based Method for Approximate Entity Extraction with Edit-Distance Constraints*. 2012 IEEE Int. Conf. on Data Engineering (**ICDE'12**), Arlington, VA, Apr. 2012: 762-773.

22. Guoliang Li, **Dong Deng**, Jiannan Wang, Jianhua Feng: *PassJoin: A Partition-based Method for Similarity Joins*. PVLDB 5(3), 253-264, 2011. Also, 38th Int. Conf. on Very Large Data Bases (**VLDB'12/PVLDB**), Istanbul, Turkey, Aug. 2012.

23. Guoliang Li, **Dong Deng**, Jianhua Feng: *Faerie: Efficient Filtering Algorithms for Approximate Dictionary-based Entity Extraction*. Proc. 2011 ACM SIGMOD Int. Conf. on Management of Data (**SIGMOD'11**), Athens, Greece, June, 2011: 529-540.

System Demos

24. Raul Castro Fernandez, **Dong Deng**, Essam Mansour, Abdulhakim Ali Qahtan, Wenbo Tao, Ziawasch Abedjan, Ahmed K. Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, Michael Stonebraker, Nan Tang: *A Demo of the Data Civilizer System*. Proc. 2017 ACM SIGMOD Int. Conf. on Management of Data (**SIGMOD'17**), Chicago, IL, USA, May 2017: 1639-1642.

25. Ji Sun, Zeyuan Shang, Guoliang Li, **Dong Deng**, Zhifeng Bao: *Dima: A Distributed In-Memory Similarity-Based Query Processing System*. PVLDB 10(12): 1925-1928, 2017. Also, 43rd Int. Conf. on Very Large Data Bases (**VLDB'17/PVLDB**), Munich, Germany, Aug. 2017.

Short, Workshop, and Survey Papers

26. Michael Stonebraker, Raul Castro Fernandez, **Dong Deng**, Michael L. Brodie: *What To Do About Database Decay*. Communications of the ACM (**CACM**), 60(1):11 (2017).

27. Sebastian Wandelt, **Dong Deng**, Stefan Gerdjikov, Shashwat Mishra, Petar Mitankin, Manish Patil, Enrico Siragusa, Alexander Tiskin, Wei Wang, Jiaying Wang, Ulf Leser: *State-of-the-art in String Similarity Search and Join*. **ACM SIGMOD Record**: 43(1): 64-76 (2014).

28. Minghe Yu, Guoliang Li, **Dong Deng**, Jianhua Feng: *String Similarity Search and Join: A Survey*. Frontiers of Computer Science (**FCS**), 10(3): 399-417 (2016).

29. Guoliang Li, **Dong Deng**, Jianhua Feng: *Extending Dictionary-based Entity Extraction to Tolerate Errors*. Proc. 2010 ACM Int. Conf. on Information and Knowledge Management (**CIKM'10**), Toronto, Canada, Oct. 2010: 1341-1344.

30. Yu Jiang, **Dong Deng**, Jiannan Wang, Guoliang Li, Jianhua Feng: *Efficient Parallel Partition based Algorithms for Similarity Search and Join with Edit Distance Constraints*. Joint 2013 EDBT/ICDT Conferences (**EDBT/ICDT'13**), Genoa, Italy, Mar. 2013, Workshop Proc.: 341-348.

Professional Services

- Program Committee Member:

  CIKM 2017, International Workshop on Similarity Search and Its Application (SISAP) 2014, 2015

- Invited Reviewer:

  The International Journal on Very Large Data Bases (The VLDB Journal)
  IEEE Transaction on Knowledge and Data Engineering (TKDE)
  ACM Journal of Data and Information Quality (JDIQ)
  ACM Transactions on Intelligent Systems and Technology (TIST)
  IEEE Transactions on Systems, Man and Cybernetics: Systems (TMC)
  Journal of Computer Science and Technology (JCST)

Invited Talks

1. *The Data Civilizer System*, Microsoft Research Asia PhD Forum 2017, Beijing, China, Sep 2017
2. *PassJoin: Partition-based String Similarity Joins*, Chinese Academy of Sciences, Beijing China, May 2017
3. *PassJoin: Partition-based String Similarity Joins*, Qatar Computing Research Institute, Doha Qatar, Apr 2017
4. *The Data Civilizer System*, Conference Talk at CIDR 2017, Chaminade, California USA, Jan 2017
5. *An Efficient Partition Based Method for Exact Set Similarity Joins.*, Conference Talk at VLDB 2016, New Delhi, India, Sep 2016
6. *META: An Efficient Matching-Based Method for Error-Tolerant Autocompletion*, Conference Talk at VLDB 2016, New Delhi, India, Sep 2016
7. *Scalable Column Concept Determination for Web Tables Using Large Knowledge Bases*, Conference Talk at VLDB 2014, Hangzhou China, Aug 2014
8. *A Pivotal Prefix based Filtering Algorithm for String Similarity Search*, Conference Talk at SIGMOD 2014, Snowbird Utah USA, Jun 2014
9. *MassJoin: A MapReduce-based Method for Scalable String Similarity Joins*, Conference Talk at ICDE 2014, Chicago USA, Apr 2014
10. *Top-k String Similarity Search with Edit-Distance Constraints*, Conference Talk at ICDE 2013, Brisbane Australia, Apr 2013
11. *Efficient Parallel Partition-based Algorithms for Similarity Search and Join with Edit Distance Constraints*, Conference Talk, EDBT/ICDT 2013, Genoa Italy, Mar 2013
12. *PassJoin: Partition-based String Similarity Joins*, Conference Talk at VLDB 2012, Istanbul Turkey, Aug 2012
13. *An Efficient Trie-based Method for Approximate Entity Extraction with Edit-Distance Constraints*, Conference Talk at ICDE 2012, Washington DC USA, Apr 2012
14. *Faerie: Efficient Filtering Algorithms for Approximate Dictionary-based Entity Extraction*, Conference Talk at SIGMOD 2012, Athens Greece, Jun 2011

References

Michael Stonebraker
Adjunct Professor
CSAIL
MIT
stonebraker@csail.mit.edu

Samuel Madden
Professor
CSAIL
MIT
madden@csail.mit.edu

H. V. Jagadish
Bernard A Galler Collegiate Professor
Elec. Engg. and Computer Science
University of Michigan
jag@eecs.umich.edu

Yufei Tao
Professor
Computer Science and Engineering
Chinese University of Hong Kong
taoyf@cse.cuhk.edu.hk

Guoliang Li
Associate Professor
Computer Science
Tsinghua University
liguoliang@tsinghua.edu.cn