Nan Tang

Associate Professor
Data Science and Analytics Thrust
Information Hub
HKUST (GZ)

☎ *+86 20-88330888*
✉ *nantang@hkust-gz.edu.cn*
https://nantang.github.io/

Research

- Synergizing large language models and multi-modal data lakes
- AI-powered data preparation and Data preparation for AI (data-centric AI)
- AI-powered data visualization

Education

2004/07–2007/12 **Ph.D.**, Systems Engineering & Engineering Management
The Chinese University of Hong Kong, Hong Kong
- Thesis: Efficient XPath Query Processing in Native XML Databases
- Co-Supervisors: Jeffrey Xu Yu and Kam-Fai Wong

2001/09–2004/01 **M.Sc.**, Computer Science
Northeastern University, China
- Thesis: Parallel XML Databases
- Supervisors: Guoren Wang and Ge Yu

1997/09–2001/01 **B.S.**, Computer Science
Northeastern University, China

Professional Experience

2023/07–now *Associate Professor*, **HKUST (GZ)**, China

2015/04–2023/06 *Senior Scientist*, **Qatar Center for Artificial Intelligence, QCRI**, Qatar

2017/07–08 *Visiting Scientist*, **MIT**, US. Worked on the DATA CIVILIZER project that provides data preparation as a service for data science, with Michael Stonebraker, Samuel Madden, and Armando Solar-Lezama

2011/12–2015/03 *Scientist*, **Data Analytics, QCRI**, Qatar

2010/02-2012/01 *Research Fellow*, **University of Edinburgh**, UK. Worked on data cleaning and graph algorithms, with Wenfei Fan

2008/02-2010/01 *Scientific Staff Member*, **CWI** (the national research institute for mathematics and computer science), the Netherlands. Worked on column-store database MonetDB and distributed XQuery processing, with Peter Boncz

2007/03–08 *Visiting Scholar*, **University of Waterloo**, Canada. Worked on XML indexing and query rewriting, with Tamer Özsu

Awards

2023 *ICDE 2023 Distinguished Reviewer Award*

2021 *VLDB 2021 Distinguished Reviewer Award*

2020 *SIGMOD 2020 Reproducibility Award*: Raha: A Configuration-Free Error Detection System

2018 *Best papers of ICDE 2018*: Discovering Mis-Categorized Entities

2015 *Best papers of VLDB 2015*: Lightning Fast and Space Efficient Inequality Joins

2012 *Best papers of ICDE 2012*: Incremental Detection of Inconsistencies in Distributed Data

2010 *The Best Paper Award of VLDB 2010*: Towards Certain Fixes with Editing Rules and Master Data

2009 *Best papers of ICDE 2009*: Projective Distribution of Full-Fledged XQuery

Research at QCRI (Dec 2011–June 2023)

— Data preparation with human intelligence for data science

[DP4DS] • A commodity system for declarative data cleaning: NADEEF (SIGMOD'13, SIGMOD'15).

[DP4DS] • Analyzing real-world data errors that were provided by various organizations (PVLDB'16).

[DP4DS] • Statistical error detection: a robust disguised missing values detector (KDD'18).

[DP4DS] • Logical data cleaning: FIXING rules (SIGMOD'14), and pattern functional dependencies (PVLDB'20).

[DP4DS] • Logical data cleaning with master data: SHERLOCK rules (ICDE'15).

[DP4DS] • Logical data cleaning with knowledge bases: KATARA (SIGMOD'15) and DETECTIVE rules (ICDE'17).

[DP4DS] • Data Preparation as a Service (with **MIT**): a suite of pre-built tools and pipeline orchestration (CIDR'17).

[DP4DS] • Data storage and query co-optimization: CASTOR (OOPSLA'20).

[HI4DP] • Human-in-the-loop data repairing with SQL Update queries: FALCON (SIGMOD'16).

[HI4DP] • Human-in-the-loop error detection with functional dependencies: UGUIDE (SIGMOD'17).

[HI4DP] • Cleaning mis-categorized Google scholar entries (ICDE'18).

[HI4DP] • A data (not code) debugger: DAGGER (CIDR'20).

[HI4DP] • An Overleaf-like platform for collaborative data cleaning (with **UW-Madison**) (SIGMOD'20 demo).

— Data preparation meets artificial intelligence

[AI4DP] • Distributed representations of tuples for "matching" in entity resolution: DEEPER (PVLDB'18).

[AI4DP] • An ML-powered configuration-free error detector with a few examples: RAHA (SIGMOD'19).

[AI4DP] • Deep learning technologies and architectures for different data preparation tasks (EDBT'20 vision).

[AI4DP] • A design space exploration of deep learning for "blocking" in entity resolution (PVLDB'21).

[AI4DP] • Relational pre-trained Transformers for data preparation (PVLDB'21) – a GPT-3 like tool for data prep.

[AI4DP] • Domain adaptation for deep entity resolution (SIGMOD'22).

[AI4DP] • ymphony: Towards Natural Language Query Answering over Multi-modal Data Lakes. (CIDR'23).

[AI4DP] • Few-shot NL2SQL Translation using Structure and Content Prompt Learning. (SIGMOD'23).

[AI4DP] • Unicorn: A Unified Multi-tasking Model for Supporting Matching Tasks in Data Integration. (SIGMOD'23).

[DP4AI] • Adaptive data augmentation for noise shift in supervised ML using generative models (PVLDB'21).

[DP4AI] • Selective data acquisition for supervised ML (PVLDB'21 demo, PVLDB'22).

[DP4AI] • HybridPipe: Combining Human-generated and Machine-generated Pipelines for Data Preparation. (SIGMOD'23).

[DP4AI] • GoodCore: Coreset Selection over Incomplete Data for Data-effective and Data-efficient Machine Learning. (SIGMOD'23).

— Self-automatic data visualization for interpretable data science

[AI4DV] • DEEPEYE: ML-based automatic data visualization (ICDE'18), used by Tencent and ByteDance.

[AI4DV] • Steerable self-driving data visualization (TKDE'20).

[AI4DV] • Produced the 1st natural language to visualization benchmark for machine translation (SIGMOD'21).

[AI4DV] • Proposed the 1st natural language to visualization neural machine translation model (IEEE VIS'21).

[AI4DV] • Learned Data-aware Image Representations of Line Charts for Similarity Search (SIGMOD'23).

[DV4DS] • DEEPEYE$^+$: supporting Google-like queries for visualization recommendation (EDBT'18 vision).

[DV4DS] • Towards democratizing relational data visualization (SIGMOD'19 tutorial).

[DV4DS] • Making data visualization more efficient and effective: a survey (VLDBJ'20).

[DV4DS] • A data science system for exploring COVID-19 data (IEEE Data Eng. Bulletin'20, invited).

[DV4DS] • Qatar COVID-19 situation dashboard: used by MOI Qatar, showcased on Al Jazeera (2020).

[DV4DS] • COVID-19 mobility analysis: used by MOPH Qatar and Kuwait Health Ministry (2020).

[HI4DV] • Interactive cleaning for progressive visualization (ICDE'20).

Research at University of Edinburgh (Feb 2010–Dec 2011)

[DP4DS] • Worked on data cleaning using master data (PVLDB'10, the Best Paper Award), interacting different types of data quality rules (SIGMOD'11), incrementally detecting errors in distributed data (ICDE'12), and inferring data currency and consistency for conflict resolution (ICDE'13).

Research at CWI (Feb 2008–Jan 2010)

[MonetDB] • Worked on efficiently supporting updates in column-stores using packed memory arrays. building space-economical Q-gram index for exact string matching over a 400+GB data set (CIKM'09), and enabling efficient distribution of full-fledged XQuery on top of MonetDB/XQuery (ICDE'09), for supporting the use cases from the Netherlands Forensic Institute.

Teaching and Mentoring Experience

Mentored Interns and Postdocs, QCRI

Hakim Qahtan	Ph.D., KAUST, Saudi Arabia (now assistant professor at Utrecht)	2017/09-2020/08
Jinsong Guo	Ph.D., University of Oxford, UK	2017/03-2017/09
Sibo Wang	Ph.D., NTU, Singapore (now assistant professor at CUHK)	2016/06-2016/11
Dong Deng	Ph.D., Tsinghua University, China (now assistant professor at Rutgers)	2016/06-2016/08
Sourav Medya	Ph.D., UC Santa Barbara, US (now research assistant professor at Northwestern)	2016/06-2016/08
Qing Chen	Master, Fudan University, China (now Ph.D. at Zurich University)	2015/07-2016/04
Jian He	Master, Tsinghua University, China (now at Google)	2014/11-2015/02
Matteo Interlandi	Ph.D., University of Modena, Italy (now at MSR)	2014/03-2014/05
Chu Xu	Ph.D., University of Waterloo, Canada (now assistant professor at Georgia Tech)	2013/05-2014/07
Jiannan Wang	Ph.D., Tsinghua University, China (now associate professor at Simon Fraser)	2012/12-2013/02
Yu Tang	Master, Hong Kong University, HK	2012/11-2013/01
Amr Ebaid	Ph.D., Purdue University, US (now at Google)	2012/04-2013/01
Ahmed Eldawy	Ph.D., University of Minnesota, US (now assistant professor at UC Riverside)	2012/01-2012/05
Michele Dallachiesa	Ph.D., University of Trento, Italy	2012/01-2012/05

Teaching, University of Edinburgh, UK (Tutorials)

Applied Databases	2010/09-11

Teaching, The Chinese University of Hong Kong, Hong Kong (Tutorials)

Digital Logical and Systems	2006/09-12, 2007/09-12
Fundamentals of Information Systems	2004/09-12, 2006/01-05
Information Systems Design & Analysis	2005/01-05

Selected Professional Activities and Services

PC Member	SIGMOD Exhibition Chair (2021), SIGMOD (2015, 2017–2020, 2022), PVLDB (2015, 2019–2021), KDD (2019–2021), CHI (2021), IEEE VIS (2021), ICDE (2013, 2018), EDBT (2017), SDM (2017), CIKM (2011, 2012)
Journal Reviewer	VLDB Journal (2009 – 2011, 2017, 2020 – 2021), TKDE (2007, 2011, 2012, 2016, 2018, 2020 – 2021), TKDD (2012), TODS (2013), TWEB (2012, 2015)

Selected Publications, Tutorials, Patents, and Grants

Data preparation with human intelligence: big data to good data for data science

[1] Abdulhakim Qahtan, **Nan Tang**, Mourad Ouzzani, Yang Cao, and Michael Stonebraker. *Pattern Functional Dependencies for Data Cleaning*. PVLDB 2020.

[2] John K. Feser, Samuel Madden, **Nan Tang**, and Armando Solar-Lezama. *Deductive Optimization of Relational Data Storage*. OOPSLA 2020.

[3] El Kindi Rezig, Lei Cao, Giovanni Simonini, Maxime Schoemans, Samuel Madden, Mourad Ouzzani, **Nan Tang**, and Michael Stonebraker. *Dagger: A Data (not code) Debugger*. CIDR 2020.

[4] Mashaal Musleh, Mourad Ouzzani, **Nan Tang**, and AnHai Doan. *CoClean: Collaborative Data Cleaning*. SIGMOD demo, 2020.

[5] Abdulhakim A. Qahtan, Ahmed Elmagarmid, Raul Castro Fernandez, Mourad Ouzzani, and **Nan Tang**. *FAHES: A Robust Disguised Missing Values Detector*. KDD, 2018.

[6] Rohit Singh, Vamsi Meduri, Ahmed Elmagarmid, Samuel Madden, Paolo Papotti, Jorge-Arnulfo Quiané-Ruiz, Armando Solar-Lezama, and **Nan Tang**. *Synthesizing Entity Matching Rules by Examples*. PVLDB, 2018.

[7] Shuang Hao, **Nan Tang**, Guoliang Li, and Jianhua Feng. *Discovering Mis-Categorized Entities*. ICDE, 2018.

[8] Saravanan Thirumuruganathan, Laure Berti-Equille, Mourad Ouzzani, Jorge-Arnulfo Quiane-Ruiz, and **Nan Tang**. *UGuide – User-Guided Discovery of FD-Detectable Errors*. SIGMOD, 2017.

[9] Shuang Hao, **Nan Tang**, Guoliang Li, Jian Li, and Jianhua Feng. *Cleaning Relations using Knowledge Bases*. ICDE, 2017.

[10] Dong Deng, Raul Castro Fernandez, Ziawasch Abedjan, Sibo Wang, Michael Stonebraker, Ahmed Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, and **Nan Tang**. *The Data Civilizer System*. CIDR, 2017.

[11] Jian He, Enzo Veltri, Donatello Santoro, Guoliang Li, Giansalvatore Mecca, Paolo Papotti, and **Nan Tang**. *Interactive and Deterministic Data Cleaning: A Tossed Stone Raises a Thousand Ripples*. SIGMOD, 2016.

[12] Ziawasch Abedjan, Xu Chu, Dong Deng, Raul Castro Fernandez, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker, and **Nan Tang**. *Detecting Data Errors: Where are we and what needs to be done?* PVLDB, 2016.

[13] Zuhair Khayyat, Ihab F. Ilyas, Alekh Jindal, Samuel Madden, Mourad Ouzzani, Paolo Papotti, Jorge-Arnulfo Quiané-Ruiz, **Nan Tang**, and Si Yin. *BigDansing: A System for Big Data Cleansing*. SIGMOD, 2015.

[14] Xu Chu, John Morcos, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, **Nan Tang**, and Yin Ye. *KATARA: Reliable Data Cleaning with Knowledge Bases and Crowdsourcing*. SIGMOD, 2015.

[15] Matteo Interlandi, and **Nan Tang**. *Proof Positive and Negative in Data Cleaning*. ICDE, 2015.

[16] Jiannan Wang, and **Nan Tang**. *Towards Dependable Data Repairing with Fixing Rules*. SIGMOD, 2014.

[17] Michele Dallachiesa, Amr Ebaid, Ahmed Eldawy, Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, and **Nan Tang**. *NADEEF: A Commodity Data Cleaning System*. SIGMOD, 2013.

[18] Wenfei Fan, Jianzhong Li, Shuai Ma, **Nan Tang**, and Wenyuan Yu. *Interaction Between Record Matching and Data Repairing*. SIGMOD, 2011.

[19] Wenfei Fan, Jianzhong Li, Shuai Ma, **Nan Tang**, and Wenyuan Yu. *Towards Certain Fixes with Editing Rules and Master Data*. PVLDB, 2010. (**The best paper award**)

— Data preparation meets artificial intelligence

[20] **Nan Tang**, Ju Fan, Fangyi Li, Jianhong Tu, Xiaoyong Du, Guoliang Li, Sam Madden, and Mourad Ouzzani. *RPT: Relational Pre-trained Transformer Is Almost All You Need for Democratizing Data Preparation*. PVLDB, 2021.

[21] Saravanan Thirumuruganathan, Han Li, **Nan Tang**, Mourad Ouzzani, Yash Govind, Derek Paulsen, Glenn Fung, and AnHai Doan. *Deep Learning for Blocking in Entity Matching: A Design Space Exploration*. PVLDB, 2021.

[22] Tongyu Liu, Yinqing Luo, Ju Fan, **Nan Tang**, Guoliang Li, and Xiaoyong Du. *Adaptive Data Augmentation for Supervised Learning over Missing Data*. PVLDB, 2021.

[23] Jianbin Liu, Fu Zhu, Chengliang Chai, Yuyu Luo, and **Nan Tang**. *Automatic Data Acquisition for Deep Learning*. VLDB demo, 2021.

[24] Saravanan Thirumuruganathan, **Nan Tang**, Mourad Ouzzani, and AnHai Doan. *Data Curation with Deep Learning*. EDBT, 2020.

[25] Muhammad Ebraheem, Saravanan Thirumuruganathan, Shafiq Joty, Mourad Ouzzani, and **Nan Tang**. *Distributed Representations of Tuples for Entity Resolution*. PVLDB, 2018.

— **Semi-automatic data visualization for interpretable data science**

[26] Yuyu Luo, **Nan Tang**, Guoliang Li, Jiawei Tang, Chengliang Chai, and Xuedi Qin. *Natural Language to Visualization by Neural Machine Translation*. IEEE VIS, 2021.

[27] Yuyu Luo, **Nan Tang**, Guoliang Li, Chengliang Chai, Wenbo Li, and Xuedi Qin. *Synthesizing Natural Language to Visualization (NL2VIS) Benchmarks from NL2SQL Benchmarks*. SIGMOD, 2021.

[28] Yuyu Luo, **Nan Tang**, Guoliang Li, Tianyu Zhao, Wenbo Li, and Xiang Yu. *DEEPEYE: A Data Science System for Monitoring and Exploring COVID-19 Data*. IEEE Data Engineering Bulletin, 2020. (Invited)

[29] Yuyu Luo, Chengliang Chai, Xuedi Qin, **Nan Tang**, and Guoliang Li. *Interactive Cleaning for Progressive Visualization through Composite Questions*. ICDE, 2020.

[30] Yuyu Luo, Wenbo Li, Tianyu Zhao, Xiang Yu, Lixi Zhang, Guoliang Li, and **Nan Tang**. *DeepTrack: Monitoring and Exploring Spatio-Temporal Data (A Case of Tracking COVID-19)*. VLDB demo, 2020.

[31] Xuedi Qin, Yuyu Luo, **Nan Tang**, and Guoliang Li. *Making Data Visualization More Efficient and Effective: A Survey*. VLDBJ, 2020.

[32] Xuedi Qin, Yuyu Luo, **Nan Tang**, and Guoliang Li. *DeepEye: Towards Automatic Data Visualization*. ICDE, 2018.

— **Tutorials**

[33] **Nan Tang**, Eugene Wu, and Guoliang Li. *Towards Democratizing Relational Data Visualization*. SIGMOD tutorial, 2019.

— **Patents**

[34] *Dependable Data Repairing with Fixing Rules*. QCRI, HBKU (PCT/EP2013/052476).

[35] *Towards Dependable Data Repairing with Fixing Rules*. QCRI, HBKU (PCT/EP2014/052494).

[36] *KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing*. QCRI, HBKU (PCT/GB2014/051670).

[37] *NADEEF: A Holistic and Extensible Data Cleaning Platform*. QCRI, HBKU (PCT/EP2012/062446).

[38] *Generalized Data Cleaning using SAT-Solvers*. QCRI, HBKU (PCT/EP2012/062445).

— ***Grants***

[39] *Credible Open Knowledge Network* (NSF grant #1937143). Start date: September 1, 2019. End Date: May 31, 2021. Prof. Chengkai Li from **`University of Texas at Arlington`** is the PI and I serve as a strategic partner.

[40] *Effective and Efficient Data Quality Management for Data Lakes* (Australian Research Council: DP210103593). From 2021 to present. Professor Wei Wang from **`University of New South Wales`** is the PI and I serve as a co-PI.

Invited Talks

2020/05 *Data Visualization and Exploration of COVID-19 data*, QCRI lectures on the use of AI techniques for COVID-19, Qatar. (Reported by Gulf Times.)

2019/10 *Data Preparation meets Data Visualization*, at Northeastern University, US.

2016/10 *Mind Your Analytics, Clean Your Data*, at Harvard University, US.

2016/03 *Graph Stream Summarization*, at MIT, US.

2015/12 *Trusted Data Cleaning*, at KAUST, Saudi Arabia.

2014/09 *Big Data Cleaning*, at Asia-Pacific Web Conference 2014, Distinguished Lecturer series.