El Kindi Rezig

Contact Information	32 Vassar St., 32-G930 Cambridge, MA 02139, USA	*Phone:* (765) 464-4152 *E-mail:* elkindi@mit.edu *Website:* elkindi.github.io

Research Interests Data scientists spend most of their time doing "grunt" work, i.e., discovery, preparing and cleaning the data. The goal of my research is to build systems that target the main pain points in data science development: Data discovery; Data preparation and Data debugging. In collaboration with industrial parties (e.g., Intel, Massachusetts General Hospital, US Air Force), my systems are motivated by real-world use cases.

Education

- **Massachusetts Institute of Technology**, Cambridge, MA, USA

  Postdoc, Computer Science and Artificial Intelligence Laboratory (CSAIL), June 2022 (expected)
  - Advisors: Prof. Michael Stonebraker and Prof. Samuel Madden
- **Purdue University**, West Lafeyette, IN, USA

  Ph.D., Computer Science, August 2018
  - Advisors: Prof. Walid Aref and Prof. Mourad Ouzzani
- **University of Science and Technology Houari Boumediene (USTHB)**, Algiers, Algeria.

  M.S., Computer Science, June 2010
  - *Summa cum Laude*
  - Advisor: Prof. Youcef Aklouf
- **University of Science and Technology Houari Boumediene (USTHB)**, Algiers, Algeria.

  B.S., Computer Science, June 2008
  - *Magna cum Laude*

Publications

CIDR'22 **El Kindi Rezig**, Anshul Bhandari, Anna Fariha, Benjamin Price, Allan Vanterpool, Andrew Bowne, Lindsay McEvoy, Vijay Gadepally: Examples are All You Need: Iterative Data Discovery by Example in Data Lakes [Extended Abstract]. *Conference on Innovative Data Systems Research, 2022*

VLDB'21 **El Kindi Rezig**, Mourad Ouzzani, Walid Aref, Ahmed Elmagarmid and Ahmed R. Mahmood, Michael Stonebraker: Horizon: Scalable Dependency-driven Data Cleaning. *International Conference on Very Large Databases, 2021*

VLDB'21 **El Kindi Rezig**, Anshul Bhandari, Anna Fariha, Benjamin Price, Allan Vanterpool, Vijay Gadepally, Michael Stonebraker: DICE: Data Discovery by Example [Demo]. *International Conference on Very Large Databases, 2021*

CIDR'21 **El Kindi Rezig**: Data Cleaning in the Era of Data Science: Challenges and Opportunities [Extended Abstract]. *Conference on Innovative Data Systems Research, 2021*

arXiv **El Kindi Rezig**, Michael J. Cafarella, Vijay Gadepally: Technical Report on Data Integration and Preparation. CoRR abs/2103.01986 (2021)

Poly@VLDB'20 **El Kindi Rezig**, Allan Vanterpool, Vijay Gadepally, Benjamin Price, Michael J. Cafarella, Michael Stonebraker: Towards Data Discovery by Example. *Workshop on Polystore systems for heterogeneous data in multiple databases (co-located with VLDB), 2020*

VLDB'20 **El Kindi Rezig**, Ashrita Brahmaroutu, Nesime Tatbul, Mourad Ouzzani, Nan Tang, Timothy G. Mattson, Samuel Madden, Michael Stonebraker: Debugging Large-Scale Data Science Pipelines using Dagger [Demo]. *International Conference on Very Large Databases, 2020*

CIDR'20 **El Kindi Rezig**, Lei Cao, Giovanni Simonini, Maxime Schoemans, Samuel Madden, Nan Tang, Mourad Ouzzani, Michael Stonebraker: Dagger: A Data (not code) Debugger. *Conference on Innovative Data Systems Research, 2020*

IEEE Data Eng. Bull.'19 Michael Stonebraker, **El Kindi Rezig**: Machine Learning and Big Data: What is Important? *IEEE Data Engineering Bulletin, 2019*

VLDB'19 **El Kindi Rezig**, Lei Cao, Michael Stonebraker, Giovanni Simonini, Wenbo Tao, Samuel Madden, Mourad Ouzzani, Nan Tang and Ahmed Elmagarmid: Data Civilizer 2.0: A Holistic Framework for Data Preparation and Analytics. *International Conference on Very Large Databases, 2019*

HILDA@SIGMOD'19 **El Kindi Rezig**, Mourad Ouzzani, Ahmed Elmagarmid, Walid Aref, Michael Stonebraker: Towards an End-to-End Human-Centric Data Cleaning Framework. *Workshop on Human-In-the-Loop Data Analytics (co-located with SIGMOD), 2019*

arXiv **El Kindi Rezig**, Mourad Ouzzani, Walid Aref, Ahmed Elmagarmid and Ahmed R. Mahmood: Pattern-Driven Data Cleaning. CoRR abs/1712.09437 (2017)

ICDE'16 **El Kindi Rezig**, Eduard Dragut, Mourad Ouzzani, Ahmed Elmagarmid and Walid Aref: ORLF: A Flexible Framework for Online Record Linkage and Fusion [Demo]. *IEEE International Conference on Data Engineering, 2016*

VLDB'15 Ahmed R. Mahmood, Ahmed M. Aly, Thamir Qadah, **El Kindi Rezig**, Anas Daghistani, Amgad Madkour, Ahmed S. Abdelhamid, Mohamed S. Hassan, Walid G. Aref, Saleh Basalamah: Tornado: A Distributed Spatio-Textual Stream Processing [Demo]. *International Conference on Very Large Databases, 2015*

ICDE'15 **El Kindi Rezig**, Eduard C. Dragut, Mourad Ouzzani, Ahmed K. Elmagarmid: Query-Time Record Linkage and Fusion over Web Databases. *IEEE International Conference on Data Engineering, 2015*

IJBIS'13 **El Kindi Rezig**, Youcef Aklouf, Hadj Madani Meghazi: Leveraging human experts' knowledge to detect and publish compositions of Semantic Web services in a repository. *International Journal of Business Information Systems, 2013*

SIGMOD'11 Hazem Elmeleegy, Jaewoo Lee, **El Kindi Rezig**, Mourad Ouzzani, Ahmed Elmagarmid: UMAP: A System for Usage-Based Schema Matching and Mapping [Demo]. *ACM Special Interest Group on Management of Data Conference, 2011*

ICEIS'10 **El Kindi Rezig**, Valerie Monfort: Towards Location-Based Services standardization: An Application based on Mobility and Geo-Location. *International Conference on Enterprise Information Systems, 2010*

IGI Global (book chapter) Youcef AKLOUF , **El Kindi Rezig**: Rule-based approach for a better B2B discovery. *Organizational Advancements through Enterprise Information Systems: Emerging Applications, Chapter 19, IGI Global, 2009, USA*

Patents

Paolo Papotti, Felix Naumann, Sebastian Kruse, **El Kindi Rezig**: Systems and Methods for Data Integration. Patent No 20160154830. June 2016.

Edited Books

El Kindi Rezig , Vijay Gadepally, Timothy G. Mattson, Michael Stonebraker, Tim Kraska, Fusheng Wang, Gang Luo, Jun Kong, Alevtina Dubovitskaya: Heterogeneous Data Management, Polystores, and Analytics for Healthcare - VLDB Workshops, Poly 2021 and DMAH 2021, Revised Selected Papers. ISSN 0302-9743. Springer Lecture Notes in Computer Science.

Awards and honors

- Raymond Boyce teaching award (highest teaching award in the Purdue CS dept.), 2017.
- Purdue Computer Science Teaching Fellowship, 2015 - 2016.
- Purdue Graduate School Summer Research Grant, 2015.
- Purdue CS travel award, 2016.
- ICDE Student Travel Award, 2015.
- Purdue College of Science Graduate Student International Travel Award, 2015.
- Ranked 1^{st} at the Microsoft Imagine Cup competition, Software Design category, Algeria, 2009.
- Microsoft Student Partner, USTHB, 2009 - 2010.
- Top computer science student award (most prestigious university-wide student award given to the first-ranking student in the M.S program), USTHB, Algeria, 2010.

Teaching Experience

- (Summer 2017 and 2018) CS50011: Introduction To Systems For Information Security. Role: **Co-developer and instructor (2 terms)**.
- (8 terms) CS252: Systems Programming. Role: TA (6 terms) and **instructor (2 terms)**.
- (Spring 2018) CS251: Algorithms and Data Structures. Role: TA (1 term).
- (Spring 2017) CS240: Programming in C. Role: TA (1 term).

Mentorship

- Mentored 6 students (4 MIT Master of Engineering/graduate, 1 undergraduate and 1 high school), MIT, 2018 - now
- Mentored Master Theses
  - Peter Griggs: Database updates using interactive Pan and Zoom visualizations. MIT EECS, 2020
  - Zhaoyuan Zhang: A New Authoring System for Diverse Data Visualization At Scale. MIT EECS, 2021
  - Jim Peraino: Architectural epidemiology: a computational framework. MIT EECS, 2020
  - Erica Zhou: Interactive visualization and discovery of possible transmission routes of Clostridioides difficile. MIT EECS, 2020

Professional Experience

Massachusetts Institute of Technology - CSAIL, Cambridge, MA, USA.

Postdoctoral Associate **November 2018 to now**

Purdue University, West Lafayette, IN, USA.

Graduate Teaching Assistant / Graduate Instructor **June 2014 to August 2018**

Graduate Research Assistant **August 2010 to May 2014**

Qatar Computing Research Institute, Doha, Qatar.

Research Associate Intern **Jan 2012 to Jul 2012 and May 2013 to Aug 2013**

INEODEV Ltd, Algiers.

Software developer **July 2008 to July 2010**

Service

- Reviewer of the SIGMOD Record 2021
- Session chair at SIGKDD 2021 (research track)
- PC member of SIGMOD 2022 (demo track)
- Co-chair of the POLY'21 workshop (co-located with VLDB 2021)
- Reviewer of: SIGMOD 2023, HILDA@SIGMOD 2022, SIGMOD 2022 (Demo), VLDB 2022 (Demo), HILDA@SIGMOD 2022, SIGKDD 2021, SIGKDD 2022, DASFAA 2021, the VLDB Journal 2021, ACM Journal of Data and Information Quality 2021, SIAM International Conference on Data Mining 2020, IEEE Transactions on Knowledge and Data Engineering. 2018 - 2020
- Demo session chair at VLDB 2020
- Volunteer student organizer at the ICDE conference 2014 (Chicago) and 2016 (Helsinki)
- External reviewer: CIKM (2015), SSTD (2015), TKDE (2014), EDBT (2013, 2014), IEEE Transactions on Services Computing (2011), WISE (2011)
- After-school Math and English tutoring for at-risk youth at the Lafayette Transitional Housing Center, Lafayette, IN, USA, 2013
- Vice president of the .NET Club at the USHTB computer science department, Algeria, 2010

Recent talks

- Dagger: a Provenance-Based Data Debugging System. Intel Labs. March 2021
- DICE: Data Discovery by Example. AI Accelerator annual meeting. April 2021
- Dagger: a Provenance-Based Data Debugging System. Data Systems and Artificial Intelligence (DSAIL) convention at MIT. April 2021
- Data Cleaning in the Era of Data Science: Challenges and Opportunities. CIDR January 2020
- Debugging Large-Scale Data Science Pipelines using Dagger. VLDB. August 2020
- Dagger: a Data (not code) Debugger. CIDR. January 2020
- Data debugging for Data Science. Qatar Computing Research Institute. March 2019