

# El Kindi Rezig

CONTACT INFORMATION	32 Vassar St. Cambridge, MA 02139 USA	Phone: (765) 464-4152 E-mail: elkindi@mit.edu
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 make building data science pipelines “fun” by developing systems that target the main pain points in pipeline development: (1) Data discovery; (2) Data preparation and (3) Data debugging. In collaboration with industrial parties (e.g., Intel, Massachusetts General Hospital), my systems target real-world use cases.	
EDUCATION	<ul style="list-style-type: none"><li>• <b>Purdue University</b>, West Lafayette, IN, USA Ph.D., Computer Science, August 2018<ul style="list-style-type: none"><li>• Advisors: Prof. Walid Aref and Mourad Ouzzani.</li></ul></li><li>• <b>University of Science and Technology Houari Boumediene</b>, BP 32, El Alia, Algiers, Algeria. M.S., Computer Science, June 2010<ul style="list-style-type: none"><li>• <i>Summa cum Laude</i></li><li>• Advisor: Prof. Youcef Aklouf</li></ul></li><li>• <b>University of Science and Technology Houari Boumediene</b>, BP 32, El Alia, Algiers, Algeria. B.S., Computer Science, June 2008<ul style="list-style-type: none"><li>• <i>Magna cum Laude</i></li></ul></li></ul>	
PUBLICATIONS	<p><b>El Kindi Rezig</b>, Anshul Bhandari, Anna Fariha, Benjamin Price, Allan Vanterpool, Vijay Gadepally, Michael Stonebraker: “DICE: Data Discovery by Example”. VLDB 2021 [Demo]</p> <p><b>El Kindi Rezig</b>: “Data Cleaning in the Era of Data Science: Challenges and Opportunities”. Conference on Innovative Data Systems Research. [Extended Abstract] CIDR 2021</p> <p><b>El Kindi Rezig</b>, Michael J. Cafarella, Vijay Gadepally: “Technical Report on Data Integration and Preparation”. CoRR abs/2103.01986 (2021)</p> <p><b>El Kindi Rezig</b>, Allan Vanterpool, Vijay Gadepally, Benjamin Price, Michael J. Cafarella, Michael Stonebraker: “Towards Data Discovery by Example”. Poly@VLDB 2020: 66-71</p> <p><b>El Kindi Rezig</b>, Ashrita Brahmaroutu, Nesime Tatbul, Mourad Ouzzani, Nan Tang, Timothy G. Mattson, Samuel Madden, Michael Stonebraker: “Debugging Large-Scale Data Science Pipelines using Dagger”. VLDB 2020 [Demo]</p> <p><b>El Kindi Rezig</b>, 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. CIDR 2020.</p> <p>Michael Stonebraker, <b>El Kindi Rezig</b>: “Machine Learning and Big Data: What is Important?” IEEE Data Engineering Bulletin, Vol. 42(4), December 2019.</p> <p><b>El Kindi Rezig</b>, 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”. [Demo] VLDB 2019.</p> <p><b>El Kindi Rezig</b>, 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 2019 (co-located with SIGMOD)</p> <p><b>El Kindi Rezig</b>, Mourad Ouzzani, Walid Aref, Ahmed Elmagarmid and Ahmed R. Mahmood: “Pattern-Driven Data Cleaning”. 2017 (preprint arXiv:1712.09437).</p> <p><b>El Kindi Rezig</b>, Eduard Dragut, Mourad Ouzzani, Ahmed Elmagarmid and Walid Aref: “ORLF: A Flexible Framework for Online Record Linkage and Fusion”. 32nd International Conference on Data Engineering. [Demo] ICDE 2016.</p> <p>Ahmed R. Mahmood, Ahmed M. Aly, Thamir Qadah, <b>El Kindi Rezig</b>, Anas Daghistani, Amgad Madkour, Ahmed S. Abdelhamid, Mohamed S. Hassan, Walid G. Aref, Saleh Basalamah: “Tornado: A Distributed Spatio-Textual Stream Processing”. [Demo] VLDB 2015.</p> <p><b>El Kindi Rezig</b>, Eduard C. Dragut, Mourad Ouzzani, Ahmed K. Elmagarmid: “Query-Time Record Linkage and Fusion over Web Databases”. ICDE 2015.</p>	

	<p><b>El Kindi Rezig</b>, 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 Vol.14, No.1, pp.83 - 95.</p> <p>Hazem Elmeleegy, Jaewoo Lee, <b>El Kindi Rezig</b>, Mourad Ouzzani, Ahmed Elmagarmid: “UMAP: A System for Usage-Based Schema Matching and Mapping”. [Demo] SIGMOD 2011.</p> <p><b>El Kindi Rezig</b>, Valerie Monfort: “Towards Location-Based Services standardization: An Application based on Mobility and Geo-Location”, 12th International Conference on Enterprise Information Systems, 2010.</p> <p>Youcef AKLOUF , <b>El Kindi Rezig</b>: “Rule-based approach for a better B2B discovery”. Organizational Advancements through Enterprise Information Systems: Emerging Applications, Chapter 19, IGI Global, 2009, USA.</p>
	<p><b>Under review</b></p> <p><b>El Kindi Rezig</b>, Mourad Ouzzani, Walid Aref, Ahmed Elmagarmid and Ahmed R. Mahmood. [under review]. 2021</p> <p>Anshul Bhandari, <b>El Kindi Rezig</b>: Text Indexing in Data Discovery by Example. [under review]. 2021</p>
PATENTS	<p>Paolo Papotti, Felix Naumann, Sebastian Kruse, <b>El Kindi Rezig</b>, “Systems and Methods for Data Integration”, Granted. Patent No 20160154830. June 2016.</p>
AWARDS AND HONORS	<ul style="list-style-type: none"> <li>• Raymond Boyce teaching award (highest teaching award in the CS dept.), 2017.</li> <li>• Purdue Computer Science Teaching Fellowship, 2015 - 2016.</li> <li>• Purdue Graduate School Summer Research Grant, 2015.</li> <li>• ICDE Student Travel Award, 2015.</li> <li>• Purdue College of Science Graduate Student International Travel Award, 2015.</li> <li>• Ranked 1<sup>st</sup> at the Microsoft Imagine Cup competition, Software Design category, Algeria, 2009.</li> <li>• Microsoft Student Partner, 2009 - 2010.</li> <li>• Top computer science student award, Algeria, 2010.</li> </ul>
TEACHING EXPERIENCE	<ul style="list-style-type: none"> <li>• (Spring 2018) CS50011: Introduction To Systems For Information Security (30+ students). Role: Co-developer and instructor (2 terms). [Summer 2017, 2018]</li> <li>• (8 terms) CS252: Systems Programming (200+ students). Role: TA (6 terms) and instructor (2 terms).</li> <li>• (Spring 2018) CS251: Algorithms and Data Structures (200+ students). Role: TA (1 term).</li> <li>• (Spring and Fall 2017) CS240: Programming in C (200+ students). Role: TA (2 terms).</li> </ul>
PROFESSIONAL EXPERIENCE	<p><b>Massachusetts Institute of Technology - CSAIL</b>, Cambridge, MA, USA.</p> <p><i>Postdoctoral Associate</i> <span style="float: right;"><b>November 2018 to now</b></span></p> <p><b>Purdue University</b>, West Lafayette, IN, USA.</p> <p><i>Graduate Teaching Assistant / Graduate Lecturer</i> <span style="float: right;"><b>June 2014 to August 2018</b></span></p> <p><i>Graduate Research Assistant</i> <span style="float: right;"><b>August 2010 to May 2014</b></span></p> <p><b>Qatar Computing Research Institute</b>, Doha, Qatar.</p> <p><i>Research Associate Intern</i> <span style="float: right;"><b>Jan 2012 to Jul 2012 and May 2013 to Aug 2013</b></span></p> <p><b>INEODEV Ltd</b>, Algiers.</p> <p><i>Software developer</i> <span style="float: right;"><b>July 2008 to July 2010</b></span></p>
SERVICE	<ul style="list-style-type: none"> <li>• Co-chair of the POLY’21 workshop (co-located with VLDB 2021)</li> <li>• Reviewer of: SIGKDD 2021, SIGMOD 2021 (Demo), 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</li> <li>• Demo session chair at VLDB 2020</li> <li>• Volunteer student organizer at the ICDE conference 2014 (Chicago) and 2016 (Helsinki).</li> <li>• External reviewer: CIKM (2015), SSTD (2015), TKDE (2014), EDBT (2013, 2014), IEEE Transactions on Services Computing (2011), WISE (2011).</li> </ul>

RECENT TALKS	<ul style="list-style-type: none"> <li>• Dagger: a Provenance-Based Data Debugging System. Intel Labs. March 2021</li> <li>• DICE: Data Discovery by Example. AI Accelerator annual meeting. April 2021</li> <li>• Dagger: a Provenance-Based Data Debugging System. Data Systems and Artificial Intelligence (DSAIL) convention at MIT. April 2021</li> <li>• Data Cleaning in the Era of Data Science: Challenges and Opportunities. CIDR January 2020</li> <li>• Debugging Large-Scale Data Science Pipelines using Dagger. VLDB. August 2020</li> <li>• Dagger: a Data (not code) Debugger. CIDR. January 2020</li> </ul>
TECHNICAL SKILLS	<ul style="list-style-type: none"> <li>• Programming languages: Java (proficient), C (proficient), C++ (proficient), SQL (proficient), C# (fair), Assembly (fair), Python (fair), PHP (fair), Bash (good), JESS (fair), CAML (fair).</li> <li>• Operating Systems administration (Windows 2003 Server, Linux).</li> <li>• Operating systems development: implemented device drivers, schedulers, garbage collection, etc. for the XINU OS as part of CS503 (graduate OS class at Purdue).</li> </ul>
EXTRA CURRICULAR	<ul style="list-style-type: none"> <li>• Active member of the Purdue Cycling Club, USA, 2013.</li> <li>• After-school Math and English tutoring for at-risk youth at the Lafayette Transitional Housing Center, Lafayette, IN, USA, 2013.</li> <li>• Vice president of the .NET Club at the USHTB computer science department, Algeria, 2010.</li> </ul>