MOHAMMAD RAZA

Email: m4raza@yahoo.com
Tel: +1 (425) 503-0096
Web: mohammadraza4.github.io

RESEARCH INTERESTS

My research lies at the intersection of program synthesis and data management, with a blended research and engineering focus on developing novel techniques that have been shipped in mass-market products. Writing programs correctly is a difficult and time-consuming task, especially in data-manipulation domains where users often have limited programming expertise. The goal of my research is to help users to create such programs through natural interaction paradigms such as examples, natural language, multi-modal inputs, or even fully automatic suggestions. This has been applied in various data-manipulation domains such as data science, spreadsheets, business intelligence, and data ingestion from unstructured sources such as log files, web extraction, etc. My work builds on techniques from AI and Formal methods, which recently has included neuro-symbolic approaches using large language models such as GPT, as well as purely formal symbolic methods based on programming language semantics.

EMPLOYMENT

Microsoft (2010 - present)

Principal Researcher (2022 – present) & Senior Researcher (2015–2022), Microsoft Redmond

Research and development of program synthesis technologies for data manipulation tasks, applied across different products (Visual Studio Code, Power BI, Excel, SSMS, Office, ADF, …) for automatic synthesis in various languages (Python Pandas, PowerQuery M, Regular Expressions, CSS, XML, …). Notable novel techniques (see also publications below) and features I have developed and shipped:

Combining Large Language Models and Program Synthesis. A novel approach that combines the powerful search of Large Language Models with the semantic robustness of Program Synthesis techniques. This approach uses the initial programs suggested by LLMs (which may be noisy, imprecise or semantically incorrect), to mine components from them to synthesize new programs that satisfy formal specifications, to rerank programs semantically, as well as techniques to infer predicted outputs and using the LLM itself to infer *repaired* programs. These techniques have been used to develop features such as NL to M inference for **Power BI**, NL to Pandas for **Visual Studio Code**, as well as **Regex** and **CSS** synthesis.

***Predictive Program Synthesis*.** Programming-by-example approaches such as Flash Fill in Excel allowed the inference of programs from input-output examples given by the user. I extended this to a new paradigm of inference from *input-only* examples, where the user need not explicitly provide any examples and the system can suggest useful programs based only on the input data. This allowed fully automatic inference of data extraction/transformation programs from various formats (text, web, etc.) and forms the basis of features such as automatic data cleaning suggestions in **Visual Studio Code** and **Power BI**, automatic web table suggestions from webpages in **Power BI** (and soon to appear in **Excel**), and automatic reading of text files in **SQL Server Management Studio** and **Azure Data Factory**.

***Programming by Example using Hybrid Synthesis*.** A novel approach that improves the robustness of program synthesis by combining the different approaches of top-down deductive synthesis and bottom-up enumerative synthesis. These techniques have been used to develop the *Web By Example* feature shipped in **Power BI** (and soon to appear in **Excel**) that was voted top 5 best features of 2018.

Post-doctoral Researcher/Consultant, Microsoft Research Cambridge (2010–2015)

Research in program synthesis techniques for end-user programming.

***Flash Format*.** Developed a programming-by-example technique for XML transformations based on the notion of least general generalization from inductive inference. Implemented as a feature in Microsoft **PowerPoint** for automating rich formatting transformations.

Text Transformations using Natural Language and Examples*.** Developed a novel program synthesis paradigm based on a combination of natural language and examples. Implemented as an application for performing complex text manipulation beyond the scope of existing systems such as ***Excel Flash Fill.

***Structured Information Retrieval*.** Developed a retrieval system for extracting entity-relationship information distributed across multiple web documents. Based on a SPARQL-like query language for performing structured keyword-based queries in open web domain (without ontology restrictions or predefined entity or

relationship types).

Research intern, Microsoft Research Cambridge, Aug 2009 – Oct 2009

Developed static analysis techniques for concurrent programs manipulating shared data structures.

Imperial College London, UK (2003 - 2006)

Research Associate (Department of Computing and Royal School of Mines)

Agent-based simulation modelling of plankton ecosystems and shipping logistics.

EDUCATION

PhD in Computer Science, Imperial College London, 2010

Thesis: "Resource Reasoning and Labelled Separation Logic". Developed program logics and algorithms for analysis and optimization of heap-manipulating programs operating on shared mutable data structures.
Supervisors: Philippa Gardner and Cristiano Calcagno

Masters in Mathematics (Mathematical Tripos Part III), University of Cambridge, 2003

Certificate of Advanced Study in Mathematics awarded with merit (applied statistics and probability theory)

BSc in Mathematics and Computer Science, Imperial College London, 2002

Graduate with First Class Honours

RECOGNITIONS / AWARDS

- Microsoft CTO Codex Challenge Winner (Code-first data wrangling), 2022
- Microsoft Special Performance Award (2019)
- Web by example in Power BI: Top 5 Features of 2018
- PhD research grant from Engineering and Physical Sciences Research Council (EPSRC) 2006 - 2010
- Overseas Research Studentship (ORS) Award for PhD research, 2006 - 2010
- NATO Advanced Study Institute Summer School on Software Engineering Methods Funding, 2008
- Associateship of the Royal College of Science (ARCS), 2002
- Imperial College Department of Computing Prize for First in Class (BSc Joint Mathematics and Computing), 2001

SELECTED PUBLICATIONS

- *CoWrangler: Recommender System for Data-Wrangling Scripts*. SIGMOD (Special Interest Group on Management of Data), 2023
- *CORNET: A neurosymbolic approach to learning conditional table formatting rules by example*. *In submission*, 2023
- *Overwatch: Learning Patterns in Code Edit Sequences*. OOPSLA (Object-Oriented Programming, Systems, Languages & Applications), 2022
- *Landmarks and Regions: A Robust Approach to Data Extraction*. PLDI (Programming Language Design and Implementation), 2022
- *Multi-modal Program Inference: a Marriage of Pre-trained Language Models and Component-based Synthesis*. OOPSLA (Object-Oriented Programming, Systems, Languages & Applications), 2021
- *Web data extraction using hybrid program synthesis: a combination of top-down and bottom-up inference*. SIGMOD (Special Interest Group on Management of Data), 2020
- *Structure interpretation of text formats*. OOPSLA (Object-Oriented Programming, Systems, Languages & Applications), 2020
- *Disjunctive Program Synthesis: A Robust Approach to Programming-by-Example*. AAAI (Association for the Advancement of Artificial Intelligence), 2018
- *Automated Data Extraction using Predictive Program Synthesis*. AAAI (Association for the Advancement of

Artificial Intelligence), 2017

- *Compositional Program Synthesis from Natural Language and Examples*. IJCAI (International Joint Conference on Artificial Intelligence), 2015
- *Mixed-Initiative Approaches to Global Editing in Slideware*. CHI (Computer Human Interaction), 2015
- *Programming by Example using Least General Generalizations*. AAAI (Association for the Advancement of Artificial Intelligence), 2014
- *Leveraging Human Intelligence: Semi-automated Processing in Assuring Access to Digital Content*. ORC-iPres (Open Research Challenges in Digital Preservation), 2013
- *A new level of social search: discovering the user's opinion before he can make one*, Microsoft technical report MSR-TR-2011-148, 2011
- *Resource Reasoning and Labelled Separation Logic*, PhD thesis, Imperial College London, 2010
- *Footprints in Local Reasoning*. LMCS (Journal of Logical Methods in Computer Science), 2009
- *Automatic Parallelization with Separation Logic*. ESOP (European Symposium on Programming), 2009
- *Footprints in Local Reasoning*. FOSSACS (Foundations of Software Science and Computational Structures), 2008

PATENTS

- Learning conditional formatting rules from natural language and examples
- Semi-automated Editing of Documents in XML Based Formats
- Automatic splitting of a column into multiple columns
- Interactive splitting of a column into multiple columns
- Improving robustness of programming-by-example systems by synthesizing disjunctive programs
- Multi-modal Program Inference: a Marriage of Pre-trained Language Models and Component-based Synthesis
- Universal Tabular Data Movement using Smart Paste
- Code templates based on examples of user’s code base
- Conditional Formatting by Example
- Predicting User Preferences

PROGRAMMING EXPERIENCE

C#, Python, Java, C++, SQL, ASP.NET, PHP, OCaml, Haskell, PROLOG

REFERENCES

- **Dr. Sumit Gulwani**
  Partner Research Manager, Microsoft Corporation
  One Microsoft Way, Redmond, WA 98052, United States
  Tel: +1 (425) 7067709
  sumitg@microsoft.com
- **Mr. Sid Jayadevan**
  Partner Director of Engineering, Azure Data Integration, Microsoft Corporation
  One Microsoft Way, Redmond, WA 98052, United States
  Tel: +1 (425) 7057266
  sidjay@microsoft.com
- **Prof. Dr. Philippa Gardner**
  Professor of Computer Science, Imperial College London
  Department of Computing, 180 Queen's Gate, South Kensington, SW7 2AZ, United Kingdom
  Tel: +44 (0) 207 594 8292

p.gardner@imperial.ac.uk

- **Prof. Dr. Natasa Milic-Frayling**
  Professor and Chair in Data Science, University of Nottingham
  Room B30 Computer Science, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB, UK
  Tel: +44 1159514212
  natasa.milic-frayling@nottingham.ac.uk
- **Dr. Darren Edge**
  Director, RaR, Microsoft Research Cambridge
  21 Station Rd, Cambridge CB1 2FB, UK
  Tel: +44 (1223) 745256 X256
  daedge@microsoft.com