Exploring the Data Wilderness through Examples

Davide Mottin
Aarhus University
davide@cs.au.dk

Matteo Lissandrini
Aalborg University
matteo@cs.aau.dk

Yannis Velegrakis
Utrecht University
i.velelgrakis@uu.nl

Themis Palpanas
Paris Descartes University
themis@mi.parisdescartes.fr

ABSTRACT

Exploration is one of the primordial ways to accrue knowledge about the world and its nature. As we accumulate, mostly automatically, data at unprecedented volumes and speed, our datasets have become complex and hard to understand. In this context *exploratory search* provides a handy tool for progressively gather the necessary knowledge by starting from a tentative query that hopefully leads to answers at least partially relevant and that can provide cues about the next queries to issue. Recently, we have witnessed a rediscovery of the so-called *example-based methods*, in which the user or the analyst circumvent query languages by using examples as input. This shift in semantics has led to a number of methods receiving as query a set of example members of the answer set. The search system then infers the entire answer set based on the given examples and any additional information provided by the underlying database. In this tutorial, we present an excursus over the main example-based methods for exploratory analysis, show techniques tailored to different data types, and provide a unifying view of the problem. We show how different data types require different techniques, and present algorithms that are specifically designed for relational, textual, and graph data.

ACM Reference Format:
Davide Mottin, Matteo Lissandrini, Yannis Velegrakis, and Themis Palpanas. 2019. Exploring the Data Wilderness through Examples. In *2019 International Conference on Management of Data (SIGMOD '19), June 30-July 5, 2019, Amsterdam, Netherlands.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3299869.3314031

SIGMOD '19, June 30-July 5, 2019, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5643-5/19/06…$15.00
https://doi.org/10.1145/3299869.3314031

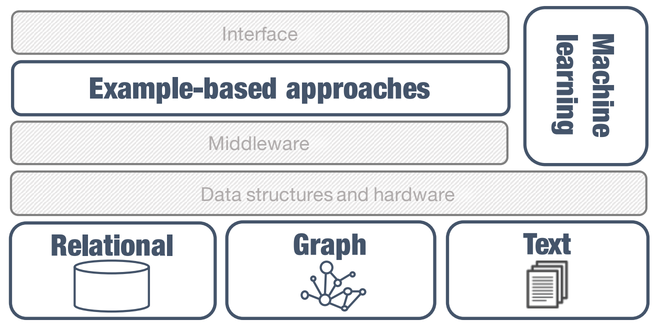

Figure 1: A view of example-based data exploration.

1 SCOPE OF THE TUTORIAL

Data exploration includes methods to efficiently extract knowledge from data, even if we do not know what exactly we are looking for, nor how to precisely describe our needs. In such exploration, the user progressively acquires the knowledge by issuing a sequence of generic queries to gather intelligence about the data. However, the existing body of work in data analysis, data visualization, and predictive models assumes the user is willing to pose several queries to the underlying database in order to progressively gather the required information. This assumption stems from the intuition that the user, being accustomed to data analysis, can more intuitively dig into the data.

Recently, *examples* became a popular proxy for data exploration. One of the earliest attempts to bring examples as a query method is query-by-example [51]. The main idea was to help the user in the query formulation, allowing them to specify the results in terms of templates for tuples, i.e., examples. Nowadays, examples are not anymore a mere template, but rather the representative of the intended results the user would like to have. These example-based approaches are fundamentally different from the initial query-by-example idea, and have are successfully applied to relational [10, 36, 43], textual [7, 48, 50], and graph [12, 17, 25] data.

We note that the flexibility of examples does not compromise the richness of the results, yet, it overcomes the ambiguity of simple keyword searches, which is traditionally studied in information retrieval. On the other hand, while data exploration techniques (Idreos et al.: Overview of data exploration techniques, tutorial in SIGMOD 2015) assume the user is willing to pose several exploratory queries, the use of examples requires almost no supervision from the user perspective, making example-based methods a more palatable choice for novice users, as well as for practitioners. This new functionality can empower existing data exploration methods with a complementary tool: whenever a query is too complex to be expressed with a query language, such as SQL, examples represent a natural alternative. In this respect (cf. Figure 1) example-based exploration is a middle ground between the user interface, and the middleware/hardware, enabling new functionalities for the former and allowing more natural exploitation of the latter. Moreover, the use of examples has been demonstrated to be very effective in data visualization [24, 39].

In this tutorial, we aim at describing the main developments of examples as an expressive and powerful method for exploratory data analysis.

The first part of the tutorial introduces the broad topic of data exploration, highlighting the hardness of query languages for simple users and advocating the need for different query methods. We will introduce the example-based methods as flexible delegates for more complex queries that would otherwise need to be expressed through a very complex traditional query. In this part, we will discuss various cases, where queries cannot be expressed in declarative languages without requiring complex constructs. We will also present an expressive formulation of example-based approaches as seeking a similarity among objects.

The second part of the tutorial discusses the current main techniques for relational, textual, and graph data. In this part, we will present the algorithms, show how they work, and demonstrate their ability to (conceptually) solve complex tasks (e.g., data integration, community detection) from simple examples. We will also highlight the differences among data types, focusing on the scalability perspective, presenting the motivations and drawing parallels among methods for different data types.

The third part of the tutorial focuses on the latest developments of machine learning to progressively discover user intention. We will introduce the general area of online learning, some early methods based on relevance feedback [16], and show some recent applications of multi-armed bandits theories, that include active search.

Challenges and open research questions. The last part of the tutorial is dedicated to the challenges and open research questions. Exploratory search based on examples is rapidly attracting attention and getting traction, though, the support for such techniques in modern data management systems is lagging behind. Some challenges have already been discussed in recent vision papers [45, 47]. We will discuss the following major challenges.

• *Adaptivity*: current data management technologies and systems do not take into account individual user preferences, and tend to optimize certain kind of queries and respond slowly to others.

• *Explanation*: Data management systems usually include little or no explanation for the results of a query. In example-based methods, in which the user query is only implicit, this requirement is even more prominent.

• *Interactivity*: Current prototypes show the advantage of example-based methods with regards to visualization techniques. However, in order to achieve the real-time, interactive performance needed by visualization tools, the algorithms should incorporate intelligent and efficient techniques for navigating through the search space.

Finally, we conclude the tutorial with remarks about the current state of affairs and engage the audience in a discussion about their experiences and challenges in this area.

2 TUTORIAL OUTLINE

In this tutorial, we provide a detailed overview of the new area of example-based methods for exploratory search, surveying the relevant state-of-the-art techniques. We also present future directions discussing various machine learning techniques used to infer user preferences interactively.
Next, we report the summary of the outline. We also provide an extended description of example-based approaches in Section 2.1, and machine learning approaches in Section 2.2.

I. **Introduction, motivation, and formulation**
  - Why example-based approaches are important
    - Usefulness of exploratory analysis
    - Main characteristics of exploratory analysis
    - Example-based methods for exploratory analysis
    - Use cases of failing keyword and declarative queries
    - Applications in current database systems and data analysis
  - Connection to data exploration
  - Problem formulation as similarity discovery

II. **Example-based approaches**
  - Query-by-example: [51]
  - Example methods in relational databases:
    - Reverse engineering of SQL queries [19, 28, 31, 36, 42, 43, 46, 49];
    - Schema mapping [1, 6, 13];
    - Data cleaning: entity matching [38], data repairing [15];
    - Example-based systems [8, 34, 35].
  - Example methods in textual data:
    - Exploring Web documents as examples [7, 50];
    - Example based Entity and Relation extraction [14, 37];
    - Web table search and augmentation [48];

- Goal oriented content discovery [29];

- Example methods in graphs:
  - Cluster and Community exploration by Example Nodes [12, 18, 30, 33];
  - Entity Search [23, 39];
  - Reverse Engineering Path Queries [4] and SPARQL queries [2, 9] from Examples;
  - Search by Example Structures [17, 21, 25].

III. **Learning methods based on examples**

- Passive similarity learning: MindReader [16]
- Active learning:
  - Multi-armed bandits and the Upper Confindence Bound algorithm [3]
  - Gaussian processes and GP-Select [44]
  - Relevance feedback learning [10] and for graphs [22, 40]

IV. **Challenges and Discussion**

- Can we *interactively* assist the user toward the retrieval of the correct answer?
- Can we provide *explanations* for the query results?
- How can machine learning help in an exploratory analysis?
- Can we easily integrate these techniques into existing data management systems?

2.1 Example-based approaches

We survey the main approaches for exploratory queries, highlighting the main differences among data models, and presenting in-depth insights into the current status of research in this area. We first introduce query-by-example [51] as a first attempt to simplify query formulation. In query-by-example, the user, instead of explicitly typing a query, specifies the shape of the results in a tabular fashion. We present the main body of work within relational, textual, and graph data, even though examples have been successfully employed also in learning syntactic program transformations [32], time series [11], and image search [41].

For *relational data* the tutorial gives an overview of techniques that solve various tasks using examples. We show how from examples we can infer fully specified SQL queries through reverse engineering [19, 28, 31, 36, 42, 43, 46, 49]. This very active area has reached maturity discovering both approximate and exact queries with different expressiveness and SQL operators. The use of examples is also beneficial in more complex tasks, such as data integration via schema mapping [1, 6, 13]. More recently, example-based approaches have been used for data cleaning by finding duplicate entities [38] or cleaning rules [15]. Last, we present prototype systems that build upon examples, such as Ziggy [35], Bleau [34], and QPlain [8].

For *textual data* the techniques include search approaches based on documents used as representatives for the set of results [48], and serendipitous search based on the current visited pages [7]. These approaches focus on documents as examples for retrieving related information. Recently, examples have been successfully employed in entity extraction [14, 37], in which the user provides either mention of entities in a text [14] or tuples and similarities among attributes [37], and the system automatically returns extraction rules that can be applied to the given dataset.

For *graph data* there are two prominent approaches: the first uses subgraphs, or partially specified structures as input examples [2, 9, 17, 25], while the second focuses on the vertices of the graph, which are used for making the selections [4, 18, 30]. Structures convey more precise information and therefore can be used to quickly prune the search space. Among the existing approaches exemplar queries [24, 25] and Graph Query by Example (GQBE) [17] use subgraph isomorphism or structural similarities to identify structures related to the one the user provided. A different approach is the reverse engineering of SPARQL queries [2, 9] in which the input is a set of positive and negative entity mentions in an RDF dataset. This approach is similar to those discussed for the relational case, and is related to learning path or join queries given positive and negative nodes [4, 5].

Instead of returning results of interest, examples can also be employed for targeted analysis of networks, in order to discover communities [18], dense regions [12, 33], or subspaces and outliers [30]. Such approaches ask the user to mark nodes belonging to a community and perform an analysis using the information in the nodes and their connections to discover regions of interest in the graph. These regions can then be used for targeted analyses or advertising campaigns.

2.2 Machine learning with examples

Current techniques use ad-hoc notions of similarity to retrieve results that are likely to be part of the solution of an unknown query. The current development in machine learning and active search [22, 27, 40] present a different perspective: user preferences can be learned from user interactions instead of manually crafted in the system. Current hardware capabilities allow to process large amount of data, and at the same time dynamically change the internal preference model. One of the earliest work in this direction is MindReader [16] in which the user specifies a set of tuples and optional relevance scores and the system infers a distance function on the objects in the database. The exploration of such *relevance learning* or *metric learning* approaches form the basis of interactive exploratory systems. Moreover, the study of Gaussian Processes as a mean of interactively learning any function given a set of points from the user has recently found applications in graphs [22, 27]. Therefore, we will present a body of work that takes the machine learning perspective into account. The research in this area is still at its infancy and forms a fertile ground for a new generation of data management systems.

3 TARGET AUDIENCE

This tutorial is intended for researchers and practitioners interested in big data analytics, graph analytics, and data exploration methods. No prior knowledge is required in order to understand the concepts in the tutorial, but we assume familiarity with database and graph concepts and basic machine learning terminology.

The tutorial aims at fostering collaborations between several disciplines, including data management, data mining, and machine learning. Researchers and students will find interesting ideas and challenges to start research in exploratory analysis, with a focus on example-based methods. Moreover, they will get an overview of the existing approaches for various data types. Addressed to practitioners, this tutorial will present a new generation of exploratory analysis techniques based on examples, which can be easily applied, and improve on a variety of existing data exploration tools for structured and non-structured data.

4 TUTORIAL MATERIAL

This tutorial builds upon our tutorial "New Trends on Exploratory Methods for Data Analytics" presented at VLDB 2017 [26] and expands it with the material from our book [20]. The current tutorial extends the previous material with

- **a unified perspective**, based on explicit and implicit similarity discovery (Part I)
- **recent developments** in data cleaning, graph queries, and entity matching (Part II)
- a more clear **connection** between online machine learning, statistics, and example-based approaches (Part III)
- **recent example-based approaches** published in the last two years (Part II and Part III)
- **a new taxonomy** of the works in each data model based on the user's question and task (Part II - beginning of model)

5 PRESENTERS

The proponents of the tutorial have several years of expertise in data management and organization of tutorials, workshops, university courses, projects, and conferences.

Davide Mottin is an assistant professor at the Aarhus University. Previously, he was a postdoctoral research at the Hasso Plattner Insitute. His research interests include graph mining, novel query paradigms, and interactive methods. He also presented exploratory techniques in KDD 2015, VLDB 2014, SIGMOD 2015, ICDE 2018, EDBT 2018 and is actively engaged in teaching database, big data analytics, and graph mining for Bachelor and Master courses. He is the proponent of exemplar queries paradigm for exploratory analysis. He received his Ph.D. in 2015 from the University of Trento.

Matteo Lissandrini is a postdoctoral researcher at the Aalborg University. He received his Ph.D. in Computer Science from the University of Trento. He also visited HP Labs at Palo Alto and the Cheriton School of Computer Science at the University of Waterloo. His scientific interests include novel query languages, information extraction with a focus on exploratory search on large graph data. He published the first Exemplar Query methods for Knowledge Graphs in VLDB and VLDBJ and presented their application at SIGMOD 2014. He served as a teacher assistant for several bachelor courses at the University of Trento.

Yannis Velegrakis is a faculty member at the Utrecht University and the University of Trento, with expertise in schema mapping, interoperability, information integration, data exchange, view management. He graduated from the University of Toronto, with a thesis on mapping management. Prior to joining the University of Trento, he held a researcher position at ATT Research Labs (USA). He also visited the IBM Almaden Research Center, the Center of Advanced Studies of the IBM Toronto Lab, and the University of California, Santa-Cruz. He served in program committees of many national and international conferences and was a Marie Curie Reintegration fellow between 2006 and 2008. He has been a general chair for the DESWeb 2010 and 2011 ICDE Workshops and was General Chair for VLDB 2013.

Themis Palpanas is Senior Member of the Institut Universitaire de France (IUF) and professor of computer science at Paris Descartes University, France. Before that, he was a professor at the University of Trento, Italy, and he has worked as a researcher at the IBM T.J. Watson Research Center and the University of California at Riverside, as well as Microsoft Research and IBM Almaden Research Center. He is the author of nine US patents, three of which are part of commercial products. He has received three best paper awards, the IBM Shared University Research (SUR) Award and was General Chair for VLDB 2013. Professor Palpanas has been working on the field of exploratory data analytics for both structured and non-structured data for the last several years, publishing relevant methods to major journals (TKDE, VLDBJ) and conferences (VLDB, SIGMOD).

REFERENCES

[1] Bogdan Alexe, Balder Ten Cate, Phokion G Kolaitis, and Wang-Chiew Tan. 2011. Designing and refining schema mappings via data examples. In *SIGMOD*. 133–144.

[2] Marcelo Arenas, Gonzalo I Diaz, and Egor V Kostylev. 2016. Reverse engineering SPARQL queries. In *WWW*. 239–249.

[3] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47, 2-3 (2002), 235–256.

[4] Angela Bonifati, Radu Ciucanu, and Aurélien Lemay. 2015. Learning path queries on graph databases. In *EDBT*.

[5] Angela Bonifati, Radu Ciucanu, and Sławek Staworko. 2016. Learning join queries from user examples. *TODS* 40, 4 (2016), 24.

[6] Angela Bonifati, Ugo Comignani, Emmanuel Coquery, and Romuald Thion. 2017. Interactive Mapping Specification with Exemplar Tuples. In *SIGMOD*. 667–682.

[7] Ilaria Bordino, Gianmarco De Francisci Morales, Ingmar Weber, and Francesco Bonchi. 2013. From machu_picchu to rafting the urubamba river: anticipating information needs via the entity-query graph. In *WSDM*. 275–284.

[8] D. Deutch and A. Gilad. 2016. QPlain: Query by explanation. In *ICDE*. 1358–1361.

[9] Gonzalo Diaz, Marcelo Arenas, and Michael Benedikt. 2016. SPARQLByE: Querying RDF data by example. *PVLDB* 9, 13 (2016), 1533–1536.

[10] Kyriaki Dimitriadou, Olga Papaemmanouil, and Yanlei Diao. 2014. Explore-by-example: An automatic query steering framework for interactive data exploration. In *SIGMOD*. ACM, 517–528.

[11] Bahaeddin Eravci and Hakan Ferhatosmanoglu. 2013. Diversity based relevance feedback for time series search. *PVLDB* 7, 2 (2013), 109–120.

[12] Aristides Gionis, Michael Mathioudakis, and Antti Ukkonen. 2015. Bump hunting in the dark: Local discrepancy maximization on graphs. In *ICDE*. 1155–1166.

[13] Georg Gottlob and Pierre Senellart. 2010. Schema mapping discovery from data instances. *JACM* 57, 2 (2010), 6.

[14] Maeda F Hanafi, Azza Abouzied, Laura Chiticariu, and Yunyao Li. 2017. Synthesizing Extraction Rules from User Examples with SEER. In *SIGMOD*. 1687–1690.

[15] Jian He, Enzo Veltri, Donatello Santoro, Guoliang Li, Giansalvatore Mecca, Paolo Papotti, and Nan Tang. 2016. Interactive and deterministic data cleaning. In *SIGMOD*. 893–907.

[16] Yoshiharu Ishikawa, Ravishankar Subramanya, and Christos Faloutsos. 1998. MindReader: Querying databases through multiple examples. In *VLDB*.

[17] Nandish Jayaram, Arijit Khan, Chengkai Li, Xifeng Yan, and Ramez Elmasri. 2015. Querying knowledge graphs by example entity tuples. *TKDE* 27, 10 (2015), 2797–2811.

[18] Isabel M Kloumann and Jon M Kleinberg. 2014. Community membership identification from small seed sets. In *KDD*.

[19] Hao Li, Chee-Yong Chan, and David Maier. 2015. Query from examples: An iterative, data-driven approach to query construction. *PVLDB* 8, 13 (2015), 2158–2169.

[20] Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis. 2018. Data Exploration Using Example-Based Methods. *Synthesis Lectures on Data Management* 10, 4 (2018), 1–164.

[21] Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis. 2018. Multi-Example Search in Rich Information Graphs. In *ICDE*.

[22] Yifei Ma, Tzu-Kuo Huang, and Jeff G Schneider. 2015. Active Search and Bandits on Graphs using Sigma-Optimality. In *UAI*. 542–551.

[23] Steffen Metzger, Ralf Schenkel, and Marcin Sydow. 2013. QBEES: query by entity examples. In *CIKM*. 1829–1832.

[24] Davide Mottin, Matteo Lissandrini, Yannis Velegrakis, and Themis Palpanas. 2014. Searching with xq: the exemplar query search engine. In *SIGMOD*. ACM, 901–904.

[25] Davide Mottin, Matteo Lissandrini, Yannis Velegrakis, and Themis Palpanas. 2016. Exemplar queries: a new way of searching. *VLDB J.* (2016), 1–25.

[26] Davide Mottin, Matteo Lissandrini, Yannis Velegrakis, and Themis Palpanas. 2017. New trends on exploratory methods for data analytics. *PVLDB* 10, 12 (2017), 1977–1980.

[27] Fabricio Murai, Diogo Rennó, Bruno Ribeiro, Gisele L Pappa, Don Towsley, and Krista Gile. 2017. Selective Harvesting over Networks. *arXiv preprint arXiv:1703.05082* (2017).

[28] Kiril Panev and Sebastian Michel. 2016. Reverse Engineering Top-k Database Queries with PALEO.. In *EDBT*. 113–124.

[29] Dimitra Papadimitriou, Georgia Koutrika, Yannis Velegrakis, and John Mylopoulos. 2017. Finding related forum posts through content similarity over intention-based segmentation. *TKDE* 29, 9 (2017), 1860–1873.

[30] Bryan Perozzi, Leman Akoglu, Patricia Iglesias Sánchez, and Emmanuel Müller. 2014. Focused clustering and outlier detection in large attributed graphs. In *KDD*. 1346–1355.

[31] Fotis Psallidas, Bolin Ding, Kaushik Chakrabarti, and Surajit Chaudhuri. 2015. S4: Top-k Spreadsheet-Style Search for Query Discovery. In *SIGMOD*. 2001–2016.

[32] Reudismam Rolim, Gustavo Soares, Loris D'Antoni, Oleksandr Polozov, Sumit Gulwani, Rohit Gheyi, Ryo Suzuki, and Björn Hartmann. 2017. Learning syntactic program transformations from examples. In *ICSE*. IEEE Press, 404–415.

[33] Natali Ruchansky, Francesco Bonchi, David García-Soriano, Francesco Gullo, and Nicolas Kourtellis. 2015. The Minimum Wiener Connector Problem. In *SIGMOD*. 1587–1602.

[34] Thibault Sellam and Martin Kersten. 2016. Cluster-driven navigation of the query space. *TKDE* 28, 5 (2016), 1118–1131.

[35] Thibault Sellam and Martin Kersten. 2016. Ziggy: Characterizing query results for data explorers. *PVLDB* 9, 13 (2016), 1473–1476.

[36] Yanyan Shen, Kaushik Chakrabarti, Surajit Chaudhuri, Bolin Ding, and Lev Novik. 2014. Discovering Queries Based on Example Tuples. In *SIGMOD*. 493–504.

[37] Rishabh Singh. 2016. Blinkfill: Semi-supervised programming by example for syntactic string transformations. *PVLDB* 9, 10 (2016), 816–827.

[38] Rohit Singh, Venkata Vamsikrishna Meduri, Ahmed Elmagarmid, Samuel Madden, Paolo Papotti, Jorge-Arnulfo Quiané-Ruiz, Armando Solar-Lezama, and Nan Tang. 2017. Synthesizing entity matching rules by examples. *PVLDB* 11, 2 (2017), 189–202.

[39] Grzegorz Sobczak, Mateusz Chochół, Ralf Schenkel, and Marcin Sydow. 2015. iqbees: Towards interactive semantic entity search based on maximal aspects. In *ISMIS*. 259–264.

[40] Yu Su, Shengqi Yang, Huan Sun, Mudhakar Srivatsa, Sue Kase, Michelle Vanni, and Xifeng Yan. 2015. Exploiting relevance feedback in knowledge graph search. In *KDD*. 1135–1144.

[41] Simon Tong and Edward Chang. 2001. Support vector machine active learning for image retrieval. In *MM*. 107–118.

[42] Quoc Trung Tran, Chee-Yong Chan, and Srinivasan Parthasarathy. 2009. Query by output. In *SIGMOD*. 535–548.

[43] Quoc Trung Tran, Chee-Yong Chan, and Srinivasan Parthasarathy. 2014. Query reverse engineering. *VLDB J.* 23, 5 (2014), 721–746.

[44] Hastagiri P. Vanchinathan, Andreas Marfurt, Charles-Antoine Robelin, Donald Kossmann, and Andreas Krause. 2015. Discovering Valuable Items from Massive Data. In *KDD*. 1195–1204.

[45] Abdul Wasay, Manos Athanassoulis, and Stratos Idreos. 2015. Queriosity: Automated Data Exploration. In *BigData*.

[46] Yaacov Y Weiss and Sara Cohen. 2017. Reverse Engineering SPJ-Queries from Examples. In *SIGMOD*. 151–166.

[47] Eugene Wu, Leilani Battle, and Samuel R. Madden. 2014. The Case for Data Visualization Management Systems: Vision Paper. *Proc. VLDB Endow.* 7, 10 (June 2014), 903–906.

[48] Mohamed Yakout, Kris Ganjam, Kaushik Chakrabarti, and Surajit Chaudhuri. 2012. InfoGather: Entity Augmentation and Attribute Discovery by Holistic Matching with Web Tables. In *SIGMOD*. 12.

[49] Meihui Zhang, Hazem Elmeleegy, Cecilia M Procopiuc, and Divesh Srivastava. 2013. Reverse engineering complex join queries. In *SIGMOD*. 809–820.

[50] Mingzhu Zhu and Yi-Fang Brook Wu. 2014. Search by Multiple Examples. In *WSDM*. 667–672.

[51] Moshé M. Zloof. 1975. Query by Example. In *AFIPS NCC*. 431–438.