GeoFlux: Hands-Off Data Integration Leveraging Join Key Knowledge

Jie Song
University of Michigan
jiesongk@umich.edu

Danai Koutra
University of Michigan
dkoutra@umich.edu

Murali Mani
University of Michigan, Flint
mmani@umflint.edu

H. V. Jagadish
University of Michigan
jag@umich.edu

ABSTRACT

Data integration is frequently required to obtain the full value of data from multiple sources. In spite of extensive research on tools to assist users, data integration remains hard, particularly for users with limited technical proficiency. To address this barrier, we study how much we can do with **no user guidance**. Our vision is that the user should merely specify two input datasets to be joined and get a meaningful integrated result. It turns out that our vision can be realized if the system can correctly determine the join key, for example based on domain knowledge.

We demonstrate this notion by considering a broad domain: socioeconomic data aggregated by geography, a widespread category that accounts for 80% of the data published by government agencies [5]. Intuitively two such datasets can be integrated by joining on the geographic unit column. Although it sounds easy, this task has many challenges: How can we automatically identify columns corresponding to geographic units, other dimension variables and measure variables, respectively? If multiple geographic types exist, which one should be chosen for the join? How to join tables with idiosyncratic schema, different geographic units of aggregation or no aggregation at all?

We have developed GeoFlux, a data integration system that handles all these challenges and joins tabular data by automatically aggregating geographic information with a new, advanced crosswalk algorithm. In this demo paper, we overview the architecture of the system and its user-friendly interfaces, and then demonstrate via a real-world example that it is general, fully automatic and easy-to-use. In the demonstration, we invite users to interact with GeoFlux to integrate more sample socioeconomic data from data.ny.gov.

CCS CONCEPTS

• **Information systems → Spatial-temporal systems**; *Mediators and data integration*; *Extraction, transformation and loading*; **Geographic information systems**;

KEYWORDS

automatic data integration; socioeconomic data; geographic data; crosswalk; multi-dimensional aggregate interpolation

ACM Reference Format:
Jie Song, Danai Koutra, Murali Mani, and H. V. Jagadish. 2018. GeoFlux: Hands-Off Data Integration Leveraging Join Key Knowledge. In *SIGMOD'18: 2018 International Conference on Management of Data, June 10–15, 2018, Houston, TX, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3183713.3193546

1 INTRODUCTION

Social scientists, policy makers, activists, and ordinary citizens all have unprecedented access to a variety of socioeconomic data, with the potential to derive valuable insights. However, in order to discover interesting results, users typically have to integrate data from multiple sources before analysis. Although many tools are available [1, 4, 7], data integration remains hard as the tools require more technical training than the typical target users have. As a result, the majority of the data integration work is still conducted manually by hand [6].

Our goal in this paper is to see how well we can integrate data **without any user guidance** at all. Ideally, we want the user only to identify two data sets, and leave it for the system to compute a meaningful integrated result.

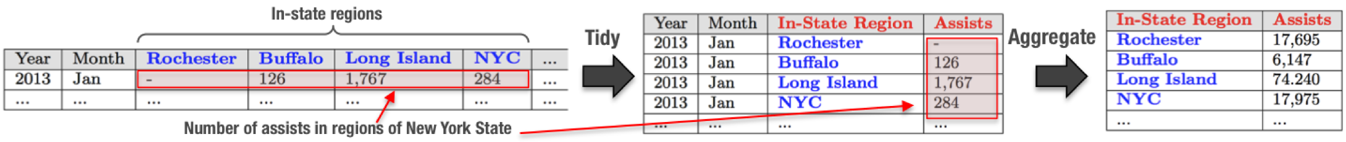
To make this problem tractable, we focus on a specific class that is both large and of practical importance: joining two tables leveraging geographic information. Though some of such data are referenced directly by some coordinate referencing system, many others, especially those in socio-economic fields are indirectly referenced without explicit coordinates. For instance, government data are often aggregated data reported in the form of tables, of which 80% include geographic information [5] organized in granularities (zip codes, counties, etc.) driven by administrative requirements. This geographic information has strategic importance as a link between datasets, and has high administrative and statistical value for governments and data scientists [3].

In the simplest case, we may imagine two tables with two columns each: the first table recording per capita income for each county and the second table recording number of reported crimes per county. An intelligent system could join these two tables on the county name column to get a three-column table that provides insight about the relationship of crime and income.

SIGMOD'18, June 10–15, 2018, Houston, TX, USA

ACM ISBN 978-1-4503-4703-7/18/06…$15.00
https://doi.org/10.1145/3183713.3193546

Table 1: Monthly HELP (Highway Emergency Local Patrol) Assists: Beginning 2010

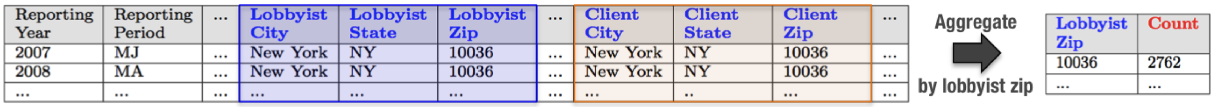

In-state regions
Tidy
Aggregate
Number of assists in regions of New York State

Year	Month	Rochester	Buffalo	Long Island	NYC	...
2013	Jan	-	126	1,767	284	...
...	...	...	...	...	...	...

Year	Month	In-State Region	Assists
2013	Jan	Rochester	-
2013	Jan	Buffalo	126
2013	Jan	Long Island	1,767
2013	Jan	NYC	284
...	...	...	...

In-State Region	Assists
Rochester	17,695
Buffalo	6,147
Long Island	74.240
NYC	17,975
...	...

Table 2: Registered Lobbyist Disclosures: Beginning 2007


Aggregate
by lobbyist zip

Reporting Year	Reporting Period	...	Lobbyist City	Lobbyist State	Lobbyist Zip	...	Client City	Client State	Client Zip	...
2007	MJ	...	New York	NY	10036	...	New York	NY	10036	...
2008	MA	...	New York	NY	10036	...	New York	NY	10036	...
...	...	...	...	...	...	...	...	..	...	...

Lobbyist Zip	Count
10036	2762
...	...

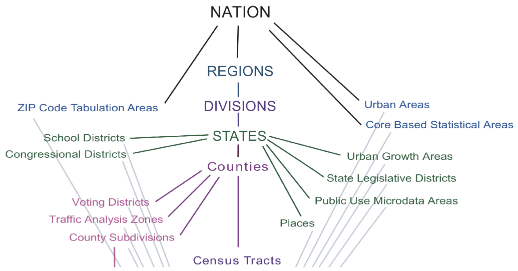


Figure 1: (Part of) Standard Hierarchy Diagram of Census Geographic Types from US Census Bureau [9].

In practice, the tables to be joined are much more complex. Among the heterogeneous integration cases, we identify three major challenges: data collection level discrepancy, data structure incompatibility and data aggregation level misalignment. In the next section, we describe two real data tables selected at random from data.ny.gov, the official website for open government data for New York State, and present the challenges we face in detail.

In this demo, we show that GeoFlux, a prototype data integration system, is able to tackle these challenges and automatically join complex tables leveraging geographic information. The system accomplishes this in three major steps. First, it identifies the messiness of the input data and transforms the given data tables into *canonical form.* Next, it automatically identifies the geographic join column(s) that maximize(s) the statistical values of the potential integration. Finally, it "aligns" the type of geographic aggregation in the join column(s) of the two tables, and then performs the required join. GeoFlux is now a working prototype joining indirect geo-referencing tables, available as a desktop Web application.

In the rest of the paper, we motivate the problem with three major challenges (§2), give an overview of the back-end integration components of GeoFlux (§3), describe how the end-user can interact with its interface with a real-world example (§4), and conclude with a summary of the problems the system addresses (§5). In the actual demonstration, our audience will be invited to interact with GeoFlux using sample datasets collected from data.ny.gov.

2 MOTIVATING EXAMPLE & CHALLENGES

Data integration is a messy process, even when restricted only to data in tabular form. In this section, we motivate the problem with a real-world integration example and present three major challenges during the integration process as the emphasis of GeoFlux, namely data collection level discrepancy, data structure incompatibility and data aggregation level misalignment.

Motivating Example. *The Monthly HELP (Highway Emergency Local Patrol) Assists data provides the number of motorists assisted by year, month and region, in vehicles on highways since 2010. Table 1 is a truncated version of its first row. In Table 2, we give a simplified version of a Registered Lobbyist Disclosures table, which contains information about biennial registration and bi-monthly filings by lobbyists to the New York State Joint Commission on Public Ethics since 2007. Let's suppose that a social scientist has some interesting hypothesis relating highway assists to lobbyists, and joining these two tables is central to understanding it.*

*Table 1 has three dimension variables (*Year*,* Month *and* In-State Region*) that describe the granularity of the data, and one measure variable (number of motorists assisted) summarized with respect to these three dimensions. The* Region *values are spread across columns such that each row is a combination of four assist number observations from four regions. Table 2 is well formatted with each column representing a dimension, describing attributes of filings including their lobbyist and client geographic information for city, state and zip code levels; and each row representing one filing as an observation. Unlike Table 1, which is reported at population level, Table 2 is reported at individual level with each record corresponding to one filing.*

While socioeconomic data are frequently reported as aggregated data at population level to protect the privacy of individual citizens or survey participants, it can also be collected at individual level with each record describing one observation surveyed. This collection level discrepancy is not rare for general data integration due to heterogeneity of data sources. Such data collected at different levels cannot simply be integrated.

As we also see in the example above, data values are often not organized in a standard way. Some tables are structured such that each column corresponds to one variable and each row is an observation; some tables use a partial two-way format in which values of variables are spread as column names; yet other tables may use some other structure. The structural incompatibility makes it hard

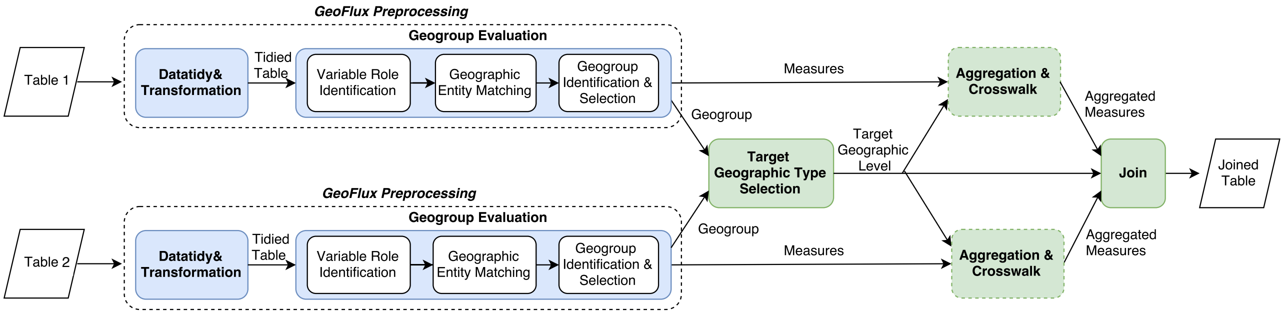


Figure 2: GeoFlux Data Integration Workflow Diagram

to understand the meaning of data in a systematic way, not to mention the subsequent integration.

Finally, data may be reported in numerous levels of aggregation with no easy alignment available. In the motivating example where geographic information can be used as join key after aggregation, there are multiple types of standard geography and there does not always exist a straightforward relationship between any two types. The United States Census Bureau defines legal, administrative and statistical boundaries and their hierarchy as in Figure 1 such that only geographies on the same line track have a fully inclusive relationship. For instance, a census tract is defined as a finer granularity region within a single county; however, a zip code may sit in several counties and a county may contain multiple zip codes.

3 SYSTEM OVERVIEW

In this section, we describe our proposed framework, GeoFlux, as a pipelined modular architecture comprising five modules, as shown in Figure 2. We walk through each module and illustrate their functionalities by the motivating example. These five modules are, in the order of processing:

(1) *Datatidy & Transformation.* In this module, we have developed an automatic messiness detection and transformation scheme that converts data into canonical form in database. For the five common types of messiness due to unalignment of data meaning and data structure defined in [10], the module tidies data so that there is exactly one variable per column, one observational unit per row and one type of observational unit per table [2, 10].

For Table 1, GeoFlux melds the four region columns into one In-State Region column and treat region names as its values. The number of assists for regions will span the new Assists column so that each row is an observation after transformation. The table is now tidied as in Table 1(middle).

(2) *Geogroup Evaluation.* This module is composed of three submodules: *Variable Role Identification*, *Geographic Entity Matching* and *Geogroup Identification & Selection*. It first analyzes the statistical types of variables as dimensions or measures by learning from variable metadata and its value distribution. Entries of geographic dimensions are then matched with a standard geographic library to identify the real-world geographic entity they represent. Lastly, it clusters geographic dimensions into geogroups, which are sets of geographic dimensions describing the same real world location, before the selection of the geographic dimension(s) in the geogroup with the highest composite data quality as the candidate join key(s).

Since Table 2 is already tidy, no tidying is performed by the previous module. This module then (i) identifies all columns in the table as dimensions and the six columns highlighted in Table 2 (left) as geographic dimensions, (ii) maps entries of geographic dimensions with corresponding geographic lookup tables (of, e.g., city, state and zip code levels), and (iii) groups City, State and Zip dimensions for Lobbyist and Client respectively as two geogroups. Since the data quality of the Lobbyist geogroup is higher, its member dimensions are considered as join key candidates.

(3) *Target Geographic Type Selection.* This module determines the best target geography, along with the source geography of each table, for the join of the two tables based on a selection heuristic that minimizes the crosswalk effort between types in the geographic type hierarchy defined by United States Census Bureau [9].

The target geography is determined by ranking all possible pair combinations of the candidate join key types, one from each table. Among in-state region level vs. city, state or zip code levels, the last pair is selected such that in-state region is the target geography and the source geography for Table 1, while zip code is the source geography for Table 2.

(4) *Aggregation & Crosswalk.* In this optional module, each table is aggregated by the source geography and crosswalked to the target geography if necessary. Apart from the available single-reference crosswalk using USPS population data, we developed GeoAlign, a multi-reference crosswalk algorithm that adaptively converts aggregates from one level to another. More details of GeoAlign are available in [8]. GeoAlign is currently available to crosswalk from zip code level to county level.

Since the source and target geographies for Table 1 (middle) are the same, it is aggregated by In-state Region without crosswalk as in Table 1 (right), while Table 2 (middle) is crosswalked to the target geography after aggregation in Table 2 (right).

(5) *Join.* The two tables are joined in this module by the target geography. The integrated result includes the join key and the measures of interest from two input tables.

Now in the motivating example, the number of assists of Table 1 and the number of filings in Table 2 are aggregated and joined by In-state Region in the integrated result.

System Execution. For efficiency of processing, all steps are carried out on individual data tables as far as possible. The first two modules depend only on the data being processed. They jointly form the GeoFlux *Preprocessing* step where the system prepares and understands the data. The third module depends on the pair of tables involved, but the computational complexity is minimal. Though the fourth module, *Aggregation & Crosswalk*, depends on the output of the third module, the actual computations are performed individually, on one table at a time. The very last module involves the actual join, requiring simultaneous access to both tables.

The integration process is invoked through a simple user interface. The user merely identifies, or provides, the two input tables, and obtains a visualization of the joined table as the result. The user optionally has access to intermediate results after each module. She can use this capability to make corrections, and to store the intermediate results, if desired, for later reuse (e.g., if the same table is later joined with something else).

4 DEMONSTRATION

The demonstration is organized into two phases: (1) an end-to-end demonstration of `GeoFlux` to automatically integrate two real-world tables to introduce the main system features; (2) a "hands-on" phase in which the public is invited to interact with the integration functionalities with more socioeconomic data examples collected from data.ny.gov. We will emphasize the first phase here.

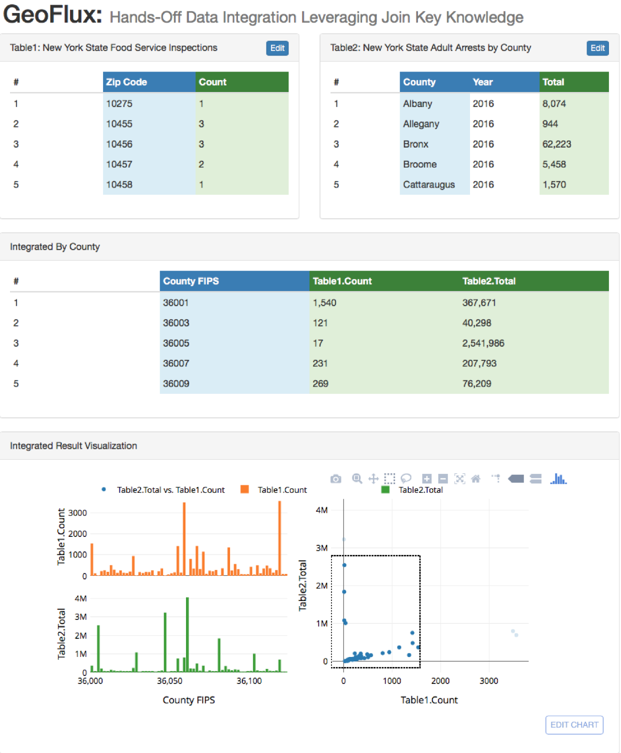
We will perform the demo on two data tables chosen from data.ny.gov: the food service inspection (by zip code) and the adult arrests by county, both in New York State. We refer to these two tables as Table1 and Table2.

`GeoFlux` is composed of two major interfaces: (1) `Upload` and (2) `Integrated Result Visualization`. `GeoFlux` first asks the user to upload multiple tables with geographic information for integration in .csv or .txt formats. As the integration process is fully automatic, the execution procedure is hidden from the user to eliminate the tedious work and avoid possible human errors.

After the two tables are integrated, the `Integrated Result Visualization Interface` shows the input tables in canonical form, the integrated table and the default visualizations (scatter plot and bar plot) as in Figure 3. Here we show part of these tables due to space limitation. For each input table, the system marks the background color of the headers of dimensions and measures in dark blue and dark green respectively, the source geographic join column in light blue and the measures in light green. The coloring of variables then propagate to their converted correspondence in the integrated result. For this example, the Count of Table1 is crosswalked from zip code to county and the Total of Table2 is grouped by County before their join by county. County FIPS is a five-digit Federal Information Processing Standards codes uniquely identifying counties in the United States. The interface also visualizes the integrated result: a bar plot for every measure of interest (Table1.Count and Table2.Total) by the target geographic type (County FIPS) and a scatter plot for every pair of measures of interest from different inputs (Table1.Count vs. Table2.Total). The user may customize the visualization as more types of plots and analysis features are available.

5 CONCLUSIONS

In this paper, we introduced `GeoFlux` as an automatic data integration system that joins government data tables based on geographic information. The demonstration will highlight the main functionalities of this system, and allow the audience to interact with `GeoFlux` and become familiar with its functionalities.



GeoFlux: Hands-Off Data Integration Leveraging Join Key Knowledge

Table1: New York State Food Service Inspections Edit

#	Zip Code	Count
1	10275	1
2	10455	3
3	10456	3
4	10457	2
5	10458	1

Table2: New York State Adult Arrests by County Edit

#	County	Year	Total
1	Albany	2016	8,074
2	Allegany	2016	944
3	Bronx	2016	62,223
4	Broome	2016	5,458
5	Cattaraugus	2016	1,570

Integrated By County

#	County FIPS	Table1.Count	Table2.Total
1	36001	1,540	367,671
2	36003	121	40,298
3	36005	17	2,541,986
4	36007	231	207,793
5	36009	269	76,209

Integrated Result Visualization

Figure 3: Integrated Result Visualization Interface

ACKNOWLEDGMENTS

This work is supported, in part, by the National Science Foundation under Grants ACI-1640575, IIS-1250880 and IIS-1743088.

REFERENCES

[1] Michael J. Cafarella, Alon Halevy, and Nodira Khoussainova. 2009. Data Integration for the Relational Web. *Proc. VLDB Endow.* 2, 1 (Aug. 2009), 1090–1101.

[2] E. F. Codd. 1990. *The Relational Model for Database Management: Version 2.* Addison-Wesley Longman Publishing Co., Inc.

[3] M. Craglia and H. Onsrud. 2004. *Geographic Information Research: Transatlantic Perspectives.* Taylor & Francis. https://books.google.com/books?id=kwTSupXTvwYC

[4] Hong-Hai Do and Erhard Rahm. 2002. COMA: A System for Flexible Combination of Schema Matching Approaches. In *VLDB*. VLDB Endowment, 610–621. http://dl.acm.org/citation.cfm?id=1287369.1287422

[5] Carl Franklin. 1992. An Introduction to Geographic Information Systems: Linking Maps to Databases. *Database* 15, 2 (1992), 12–21.

[6] Eduard Hovy, José Luis Ambite, and Andrew Philpot. [n. d.]. Addressing a Bottleneck in Data Integration using Automated Learning Techniques. ([n. d.]).

[7] Li Qian, Michael J. Cafarella, and H. V. Jagadish. 2012. Sample-driven Schema Mapping. In *SIGMOD*. ACM, New York, NY, USA, 73–84.

[8] Jie Song, Danai Koutra, Murali Mani, and H. V. Jagadish. 2018. GeoAlign: Interpolating Aggregates over Unaligned Partitions. In *Proceedings of the 21th International Conference on Extending Database Technology, EDBT 2018, Vienna, Austria, March 26-29, 2018.* 361–372. https://doi.org/10.5441/002/edbt.2018.32

[9] United States Census Bureau. 2010. Standard Hierarchy of Census Geographic Entities. Available from https://census.gov/. (2010).

[10] Hadley Wickham. 2014. Tidy data. *The Journal of Statistical Software* 59 (2014). Issue 10. http://www.jstatsoft.org/v59/i10/