# Auto-Join: Join Tables by Leveraging Transformations

Erkang (Eric) Zhu* ekzhu@cs.toronto.edu
Yeye He† yeyehe@microsoft.com
Surajit Chaudhuri† surajitc@microsoft.com
*University of Toronto †Microsoft Research

## 1. A Bird-Eye View

• **Problem:** How to join these pairs of tables automatically **without any human inputs** (including rows/columns to join)?

| President | Popular Vote |
|---|---|
| Barack Obama | 52.93% |
| George W. Bush | 47.87% |
| Bill Clinton | 43.01% |
| George H. W. Bush | 53.37% |
| Ronald Reagan | 50.75% |

| President | Approval Rating |
|---|---|
| Obama, Barack(1961-) | 47.0 |
| Bush, George W.(1946-) | 49.4 |
| Clinton, Bill(1946-) | 55.1 |
| Bush, George H. W.(1924-) | 60.9 |
| Reagan, Ronald(1911- 2004) | 52.8 |

| Name | Title |
|---|---|
| Suhela Chowdhury | Principal |
| Maureen Paluzzi | Instructor |
| Missy Payne | Instructor |
| Carolyn Craddock | Admin |
| Kelly Moore | Instructor |

| Email | School |
|---|---|
| schowdhury@forsyth.k12.ga.us | Big Creek |
| mpaluzzi@forsyth.k12.ga.us | Brookwood |
| mipayne@forsyth.k12.ga.us | Chattahoo |
| ccraddock@forsyth.k12.ga.us | Chestatee |
| kmoore@forsyth.k12.ga.us | Princeville |

| ATU | Manager Alias |
|---|---|
| France.01 | V-JOHH |
| France.03 | JOFORD |
| United States.01 | RICHT |
| United States.02 | MICHM |
| United States.03 | ANDYW |

| Sub-ATU | Segment |
|---|---|
| France.01.MIX | SMB |
| United States.01.Government | Major |
| United States.01.Education | AM EPG |
| United States.03.PS-LRG | TM SMS&P |
| United States.04.Retail | AM SMS&P |

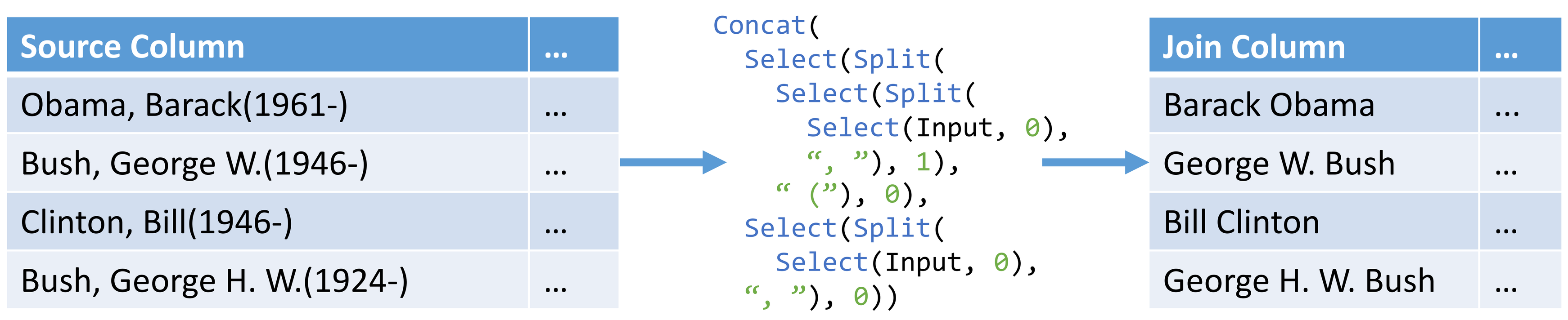
Each pairs of tables have a clear syntactic transformation between the matching rows, such as *Split*, *Concatenation*, *Substring* and *Constant*. Applying the transformation to one table creates a *join column* that can be used to equi-join with a *key column* of the other table.

• **Scenarios:** One-off, *ad hoc* data analysis often requires joining data from different sources whose data values are formatted differently. An automated solutions saves time and money on ETL.

• **Fuzzy join?** Manual parameter tuning is required otherwise likely to produce unsatisfactory result.

• **Our Solution:**

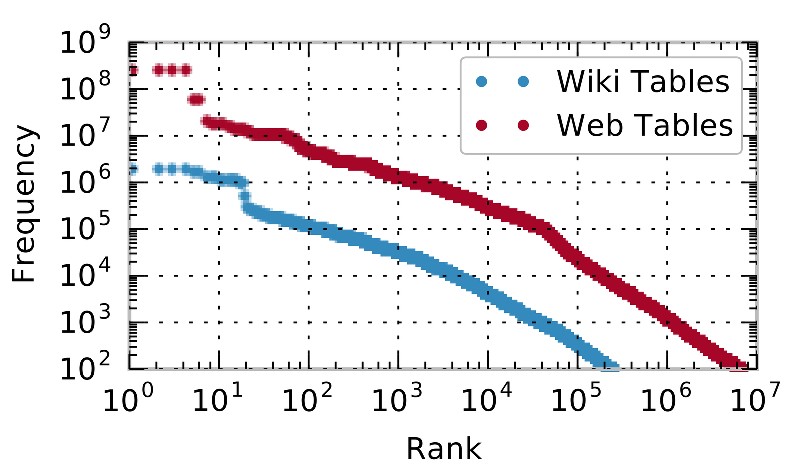
1. Efficiently identifies promising row pairs that can potentially join using substring indexes
2. Using the row pairs as examples to learn a *minimum-complexity* transformation whose execution can lead to equi-joins.



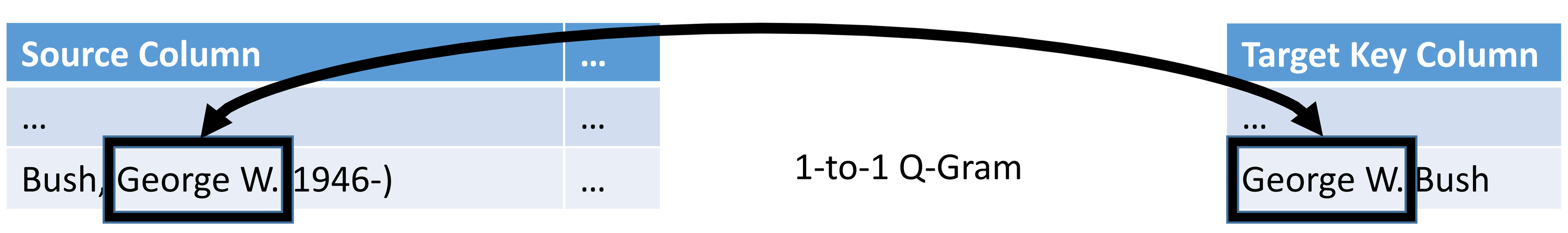

3. Maintains interactive speed even on large tables (10K rows) with a novel sampling scheme.
4. Precision 98% and recall 93% on a benchmark of 73 real-world cases.

## 2. Identify Promising Row Pairs

• **Q-Gram Distribution** in real-world tables is Zipfian, this makes the probability that a Q-Gram appears **exactly once** in each of two columns **by chance** is very small.




• **1-to-1 Q-Grams** can be used to identify promising row pairs.

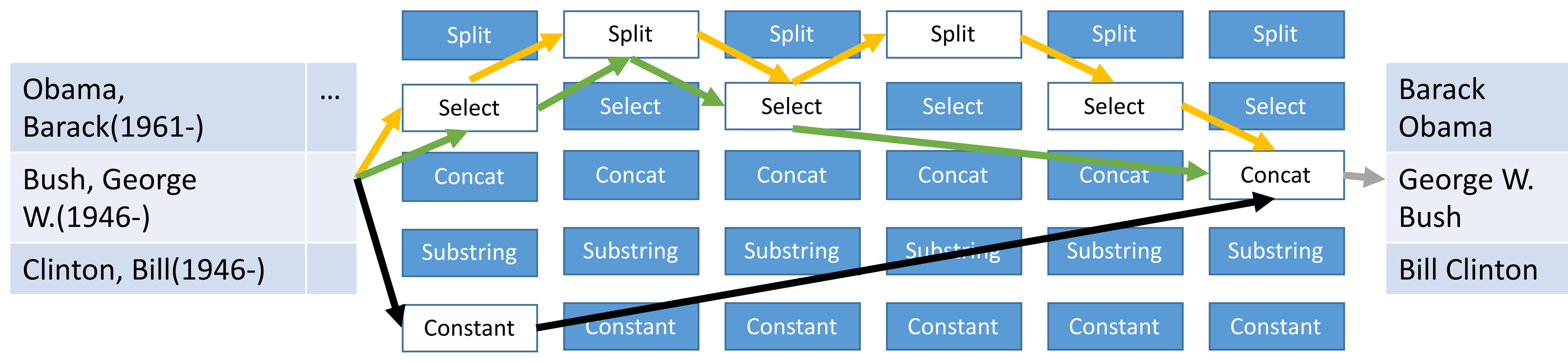



• **Q-Gram Scoring:** Since 1-to-1 Q-Grams may not be sampled, target key column may contains a few duplicates, and a transformation may not result in 1-to-1 Q-Grams (e.g., N:1). We also use *n-to-m* Q-Grams, and quantify their "goodness" as $\frac{1}{nm}$.

• **Q-Gram Search:** Our Q-Gram search algorithm uses a combination of suffix indexes and binary search to efficiently identify the optimal Q for every sampled data values and produce a ranked list of n-to-m Q-Grams for learning transformation.

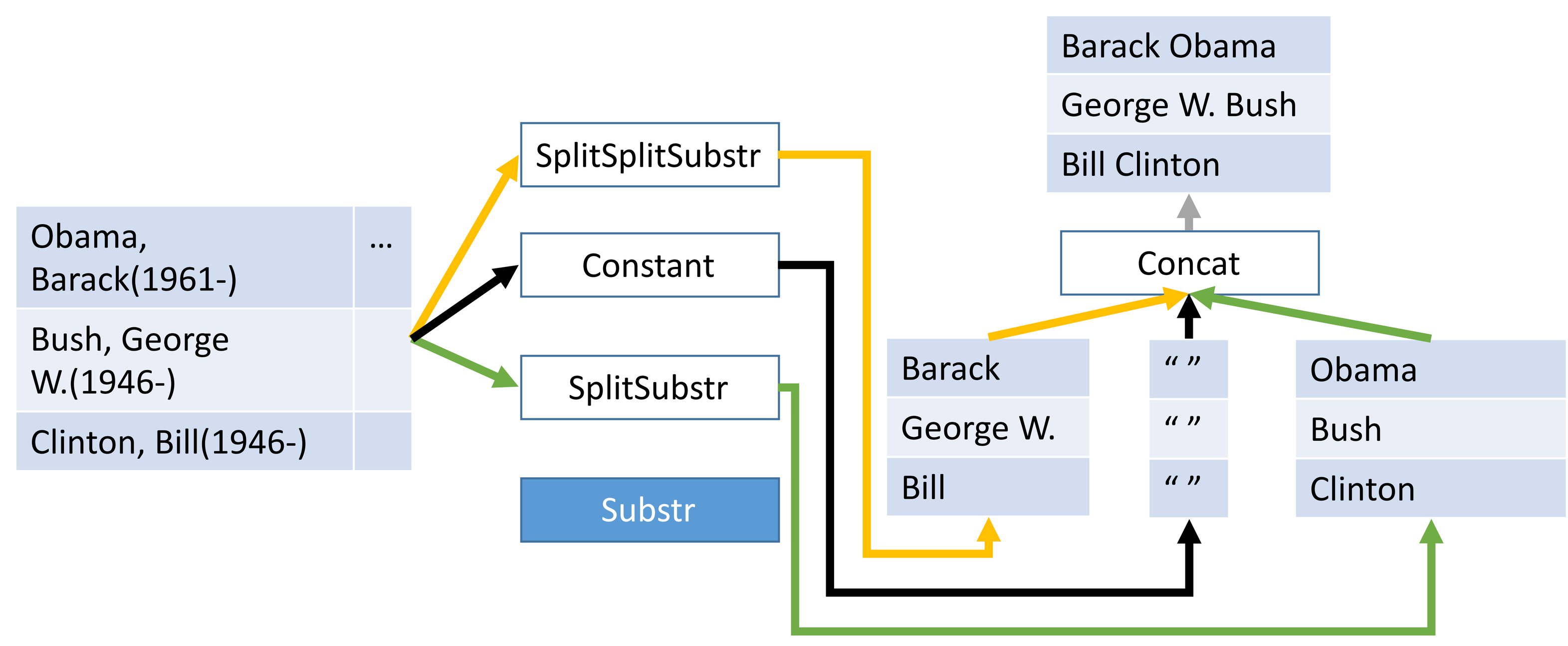
## 3. Learning Transformation

• **Learning as a search** over a graph of all possible syntactic operators and their parameters, and the transformation is the set of paths from the input to output.




• **Search space shrinks** exponentially with respect to the number of examples.

• **Logical operators** are easier for human to rationalize (e.g., "extract the first component") and reduces search space.



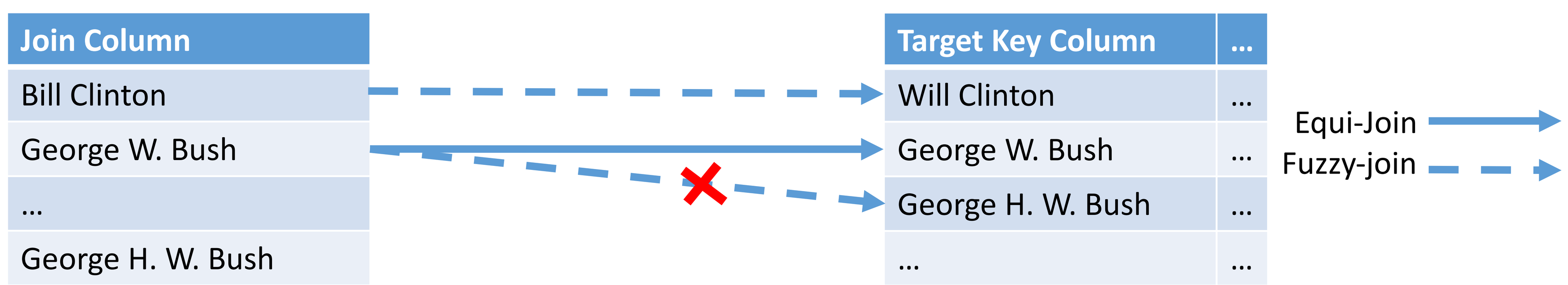

• **Algorithm:** Greedily construct a *minimum-complexity* transformation (the one with the least number of operators) by iteratively expanding an existing partial transformation with the most progress-yielding logical operator.

## 4. Optimized Fuzzy Join

• **Dirty Data:** Data may contain typos and errors, and different sources may have different namings. Applying transformation and equi-join may miss row pairs that are joinable.

• **Fuzzy join constraints** for archiving high-quality join result:

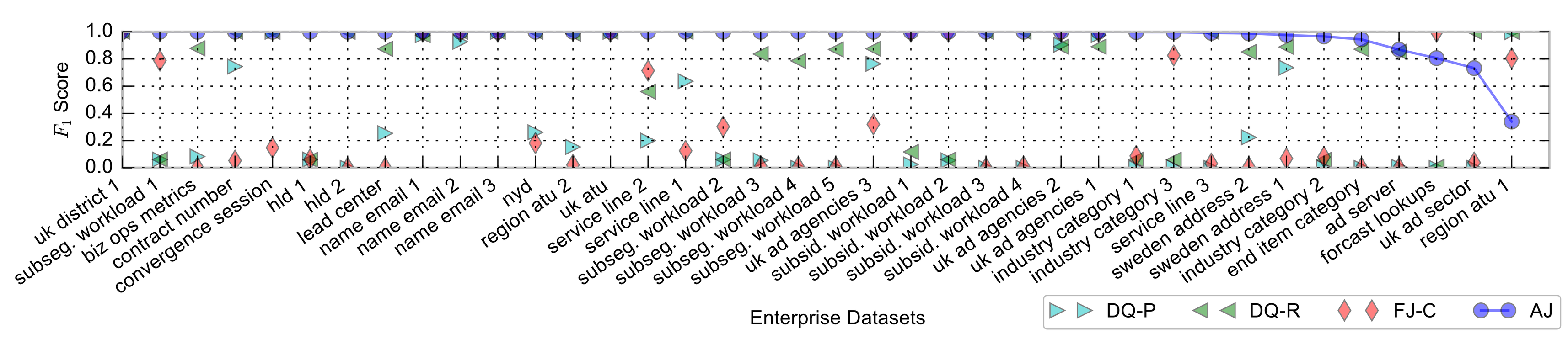
1. Every row in the join column cannot be joined with more than one distinct row in target key column – similar to key-foreign-key constraint
2. Every row in the target key column cannot be joined with more than one distinct row in the join column – assume consistency within one column

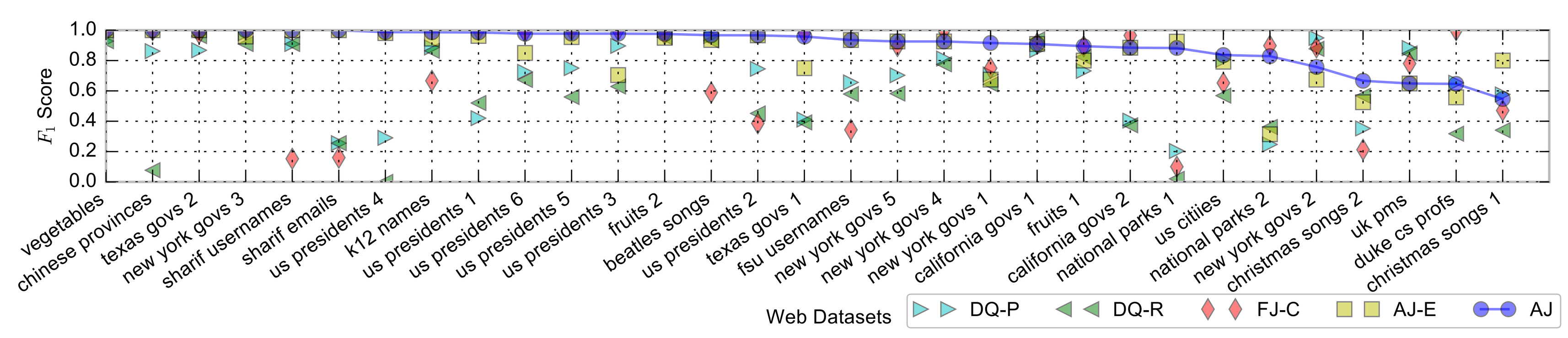



• **Optimization:** We apply binary search and the above constraints to efficiently find the optimal tuning in the fuzzy join parameter space.

## 5. Evaluation

• **Quality evaluation** uses tables from Microsoft enterprise spreadsheets and tables from the Web.




• **Performance evaluation** uses DBLP dataset; Auto-Join runs less than 5 seconds at 10K rows and 14 seconds at 100K rows.