# Icarus: Minimizing Human Effort in Iterative Data Completion

Protiva Rahman, Courtney Hebert, Arnab Nandi • The Ohio State University • {rahmanp, arnab}@cse.ohio-state.edu, courtney.hebert@osumc.edu

*Icarus* – A data completion system that allows users to quickly fill in large datasets by iteratively editing digestible subsets. Each edit is generalized to a rule, applicable to multiple cells, by leveraging foreign-key relations.

## Data Completion

- Addressing missing data is one of the first steps in data analysis pipelines
- Data unreported due to domain characteristics cannot be filled by imputation
- Require domain expert input in form of rules

## User Effort

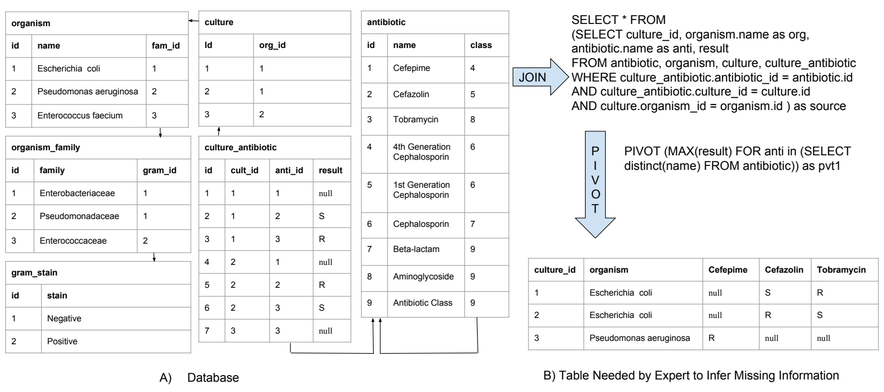
- Large dataset – in motivating example 10,000 rows, over 50 columns, 83,000 missing cells
- Manually filling in cells is infeasible
- Manually specifying rules is inefficient
  - Unclear which rules will be most effective
  - Possibility of redundant rules
  - Granularity of rule
- Navigating dataset is overwhelming
- Need interactive rule application

## Challenges

- Showing user digestible subsets of data is computationally expensive
- Subset should show relevant attributes together
- Editing subset should lead to high impact rules
- Generalize edits into rules
- Immediate application of rules
- Maintain interactive latency

## Motivating Example

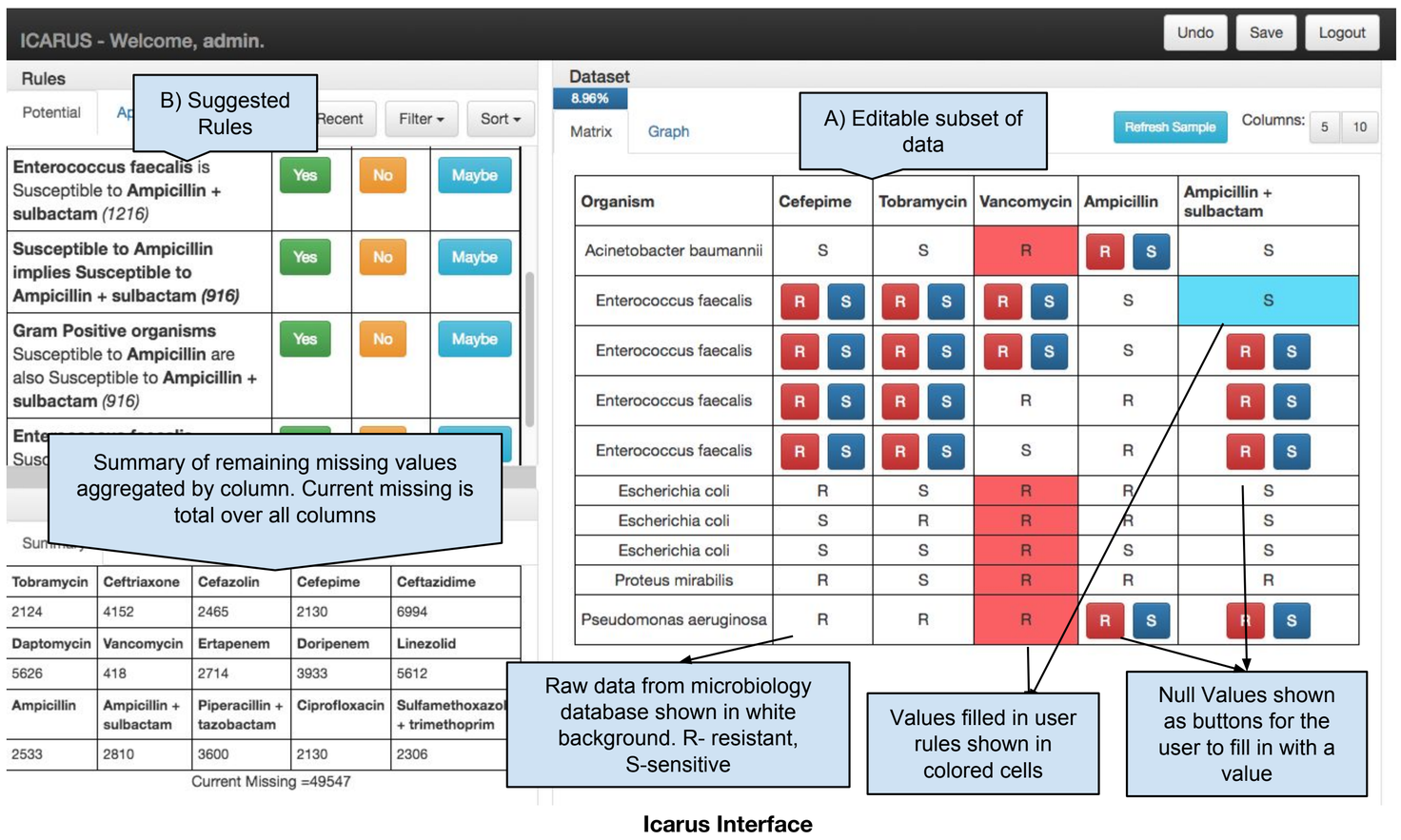
Microbiology labs contain an organism and its sensitivities to certain antibiotics. Physicians are able to infer the remaining sensitivities based on their domain knowledge. When using this data for predictive modeling, domain experts such as physicians are needed to fill in unreported data.




**Schema of Microbiology culture reports – 4 tables need to be joined and then pivoted for a domain expert to understand it. The pivoted table is long and wide, making it difficult for the expert to navigate it.**

## Repair with Rules

- Rules correspond to update queries:
  UPDATE culture_antibiotic SET result = *Resistant*
  WHERE cult_id IN (SELECT id FROM culture
  WHERE org = *E. coli*) AND anti = Vancomycin
- Rules are represented to the user as statements:
  - *E. coli is resistant to Vancomycin*
- Can have complex queries/conditional rules
  - *Staphylococcus sensitive to Cefazolin are sensitive to Cefepime*




**Icarus Interface**

## Subset Selection

- Two-stage sampling – first sample rows then columns -- to select subset in form of matrix c = x × y w.r.t to following optimization function:

$$\sum_{m_{ij} \in \{c \cap M\}} \sum_{n_k \in \{x_i \cap N\}} sim[y_j][y_k].impact[y_j][y_k] + H(c).temperature$$

- M is set of missing cells, N is set of filled in cells
- $sim[y_j][y_k]$ : similarity between columns j and k, to ensure relevant attributes are shown together
- $impact[y_j][y_k]$ : no. of tuples where column k is filled and j is missing, to ensure high impact rules can be generated
- H(c) is information entropy, temperature is no. of subsets user has seen – increasing entropy to ensure diverse subsets are shown so user has options

## Rule Generation

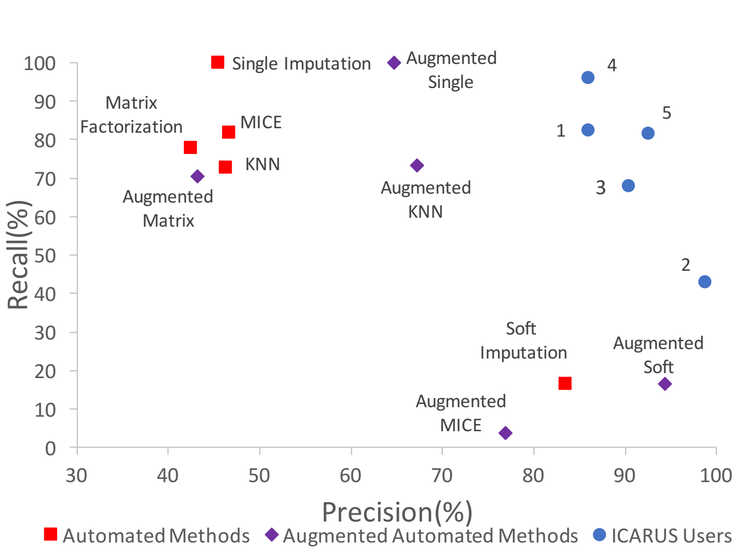
- When a user updates a cell, independent and dependent rules are suggested
- Independent rules: Rules generated between identifying attributes of join relations, i.e., update based on organism and antibiotic
  - *E. coli is resistant to Cefepime*
- Dependent rules: Rules conditioned on other tuples in join relation, i.e., update based on sensitivity to a related antibioitic
  - S. Aureus sensitive to Cefazolin implies sensitive to Cefepime
- Suggested rules are generalized up the hierarchy of the normalized table
  - *E. coli is resistant to Cefepime*
  - *Enterobacteriaceae are resistant to Cefepime*
  - *Enterobacteriaceae are resistant to Cephalosporins*

## Experimental Evaluation

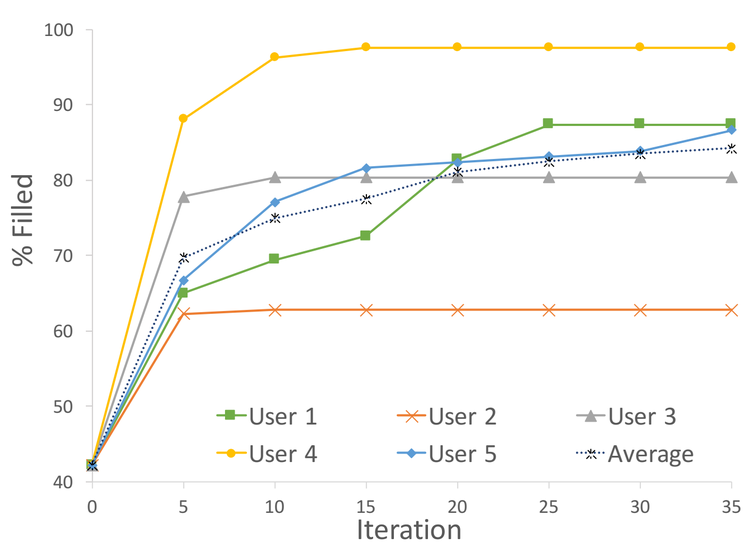
- User study with 5 domain experts
- 15 min training, 45 minutes to complete task, followed by usability survey.
- Icarus has avg. improvement of 50% across three datasets over other systems
- Users can fill in 68% of the missing values in an hour, while manual process took weeks.

| # Cells Filled | Edits |
|---|---|
| 58,672 | 246 |
| 29,299 | 46 |
| 57,104 | 155 |
| 79,480 | 126 |
| 57,104 | 147 |

**Users, on average, filled in 320 cells per edit using Icarus**




**Icarus users perform better than imputation**




**Data Completion per Iteration of 5 users**

THE OHIO STATE UNIVERSITY
http://interact.osu.edu

We acknowledge the support of NSF under awards IIS-1422977, IIS-1527779, and CAREER IIS-1453582 and NIAID of NIH under award R01AI116975.