CIFellows 2020-2021

Computing Innovation Fellows

Metam: Goal-oriented data discovery

Sainyam Galhotra, University of Chicago

Introduction

- Availability of large amounts of data
- Explosion of data sources
  - Open data
  - Web tables
  - Cloud repositories
  - Knowledge Graphs

Goal: Leverage available information for better data-driven decision making

Research questions:

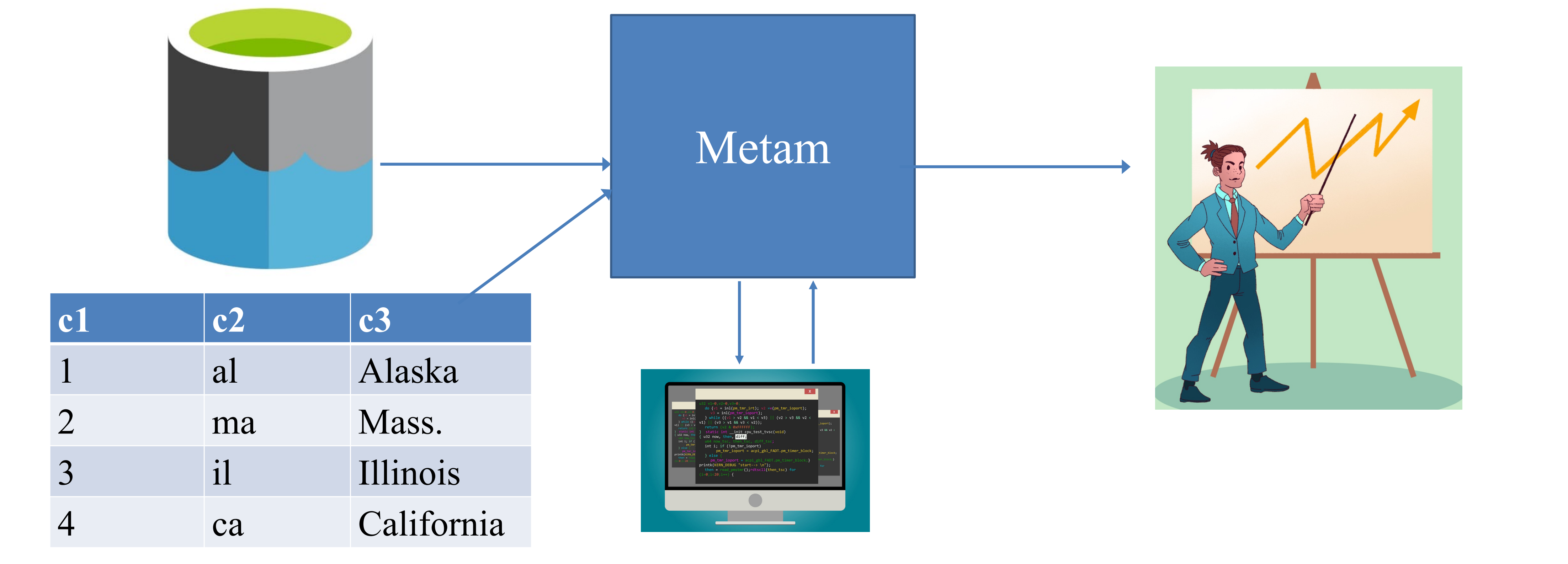
- Data Discovery: How to search for useful datasets?
- Data sharing: How can we share and trade useful data?

Challenges: Data Discovery

- Heterogeneity of representation across sources
  - -Varying data format
  - - Presence of contradictory and missing values
- Lack of join-path information
- Exponential search space

Main Insight

- Use downstream application to guide data discovery
- Automatic identification of useful datasets
- What does the data scientist do?
  - Implement downstream task
  - Define its utility metric



c1	c2	c3
1	al	Alaska
2	ma	Mass.
3	il	Illinois
4	ca	California

Problem: Given a dataset D, a data repository and downstream task t, identify join-paths to augment D such that task utility > θ

Assumption: Task outputs a utility score

Applications Studied

- Classification
- Regression
- Causal inference: What-if and how-to analysis
- Clustering
- Fairness

Our Results

- Greedy-algorithm provides (1-1/e) approximation of the optimal solution
  - Evaluation metrics are monotonic and sub-modular
- Metam tests $O(1/\epsilon^d)$ join-paths
- Empirically
  - Metam requires less than 50 iterations to identify useful datasets
  - Query datasets contain more than 10K options

CRA Computing Research Association
CCC Computing Community Consortium Catalyst
NSF