Civis Product Features:
Finding a True Match

Feb 15, 2024

Introduce yourself in the Chat!

Name
Pronouns
Organization
What brought you here today?

Webinar Norms

1. Please keep yourself muted, help us reduce background noise.
2. Raise your hand or type questions into the chat.
3. Feel free to ask other participants questions in the chat!
4. We all have varying levels of Platform knowledge on this call. Ask the questions you have–no question is too simple or too advanced!

Agenda

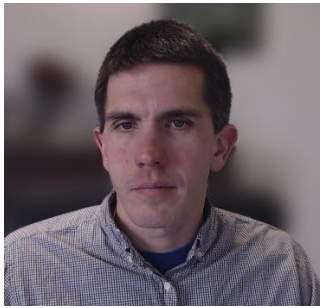
- Intros
- The Big Challenges
- Data Crosswalk
- IDR
- Questions

Intro

Our Presenters

Matt Fessler
Head of Product

Laura Cleary
Lead Data Scientist



Mike Heilman
Principal Machine Learning Engineer

Imaad Khan
Lead Pre-Sales Engineer

The Big Challenges

There are many challenges that organizations face when it comes to dealing with person-level data

- Varied data sources
- Incomplete data
- Messy Data
- Duplicates

Recent Civis Updates

Two notable releases for Platform can provide some help to you in these areas.

- Data Crosswalk (released 2/8)
  - Helps clean and categorize data
- Updated Identity Resolution (2/14)
  - Helps manages duplicates

Data Crosswalk Intro

Data Crosswalk Intro

- What is Data Crosswalk?
  - There are a lot of manual processes for string cleaning and reconciliation – standardizing all the different ways your text data shows up in a key category. We have created a way to streamline that process.

Key Benefits of Data Crosswalk

- With just a few clicks, you can now categorize your data in Platform, a process that used to involve a lot of setup with outside tools.
- Builds lasting value, by storing the output as a clear, highly reusable Platform Table that can easily be leveraged in future queries, workflows, and scripts.
- Allows more users to contribute to a very common data cleaning use case, without knowing R, Python, or even SQL.

Data Crosswalk Demo

IDR

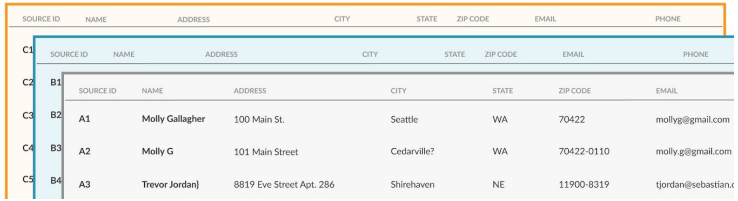
Why is resolving person-level records a difficult problem?

Some of the challenges:

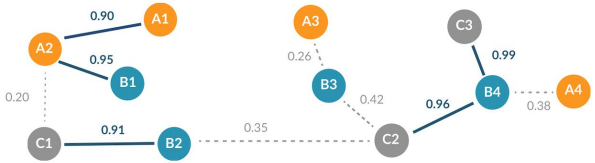
- Data may come from several sources. Both deduplication and finding links across tables can be part of the problem.
- There may be tens of millions of records to process.
- People have nicknames, move addresses, use multiple email addresses, etc.
- Some data may be incorrect (e.g., minor misspellings, incorrect January 1st dates of birth)

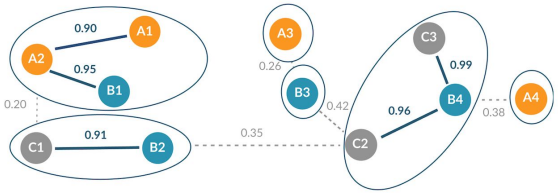
We've spent a lot of time making IDR address these challenges, so you don't have to.

What does IDR do?



SOURCE ID	NAME	ADDRESS	CITY	STATE	ZIP CODE	EMAIL	PHONE
A1	Molly Gallagher	100 Main St.	Seattle	WA	70422	mollyg@gmail.com	
A2	Molly G	101 Main Street	Cedarville?	WA	70422-0110	molly.g@gmail.com	
A3	Trevor Jordan)	8819 Eve Street Apt. 286	Shirehaven	NE	11900-8319	tjordan@sebastian.c	





1. First, IDR applies person data-specific preprocessing (e.g., to parse street addresses).
2. Then, a statistical model uses various knowledge (e.g., about the frequency of names) to compute scores between pairs of records, representing the likelihood that they represent the same person.
3. Next, a graph clustering algorithm then convert matches between pairs of records into groups of records representing the same individual, and assigns each group a "Resolved ID".
4. Finally, IDR produces outputs, including the Golden Table with one record per individual.

Interested in licensing IDR?
Email hello@civisanalytics.com