QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Supporting a Data-Driven World through Data Integration and Data Cleaning

Mourad Ouzzani

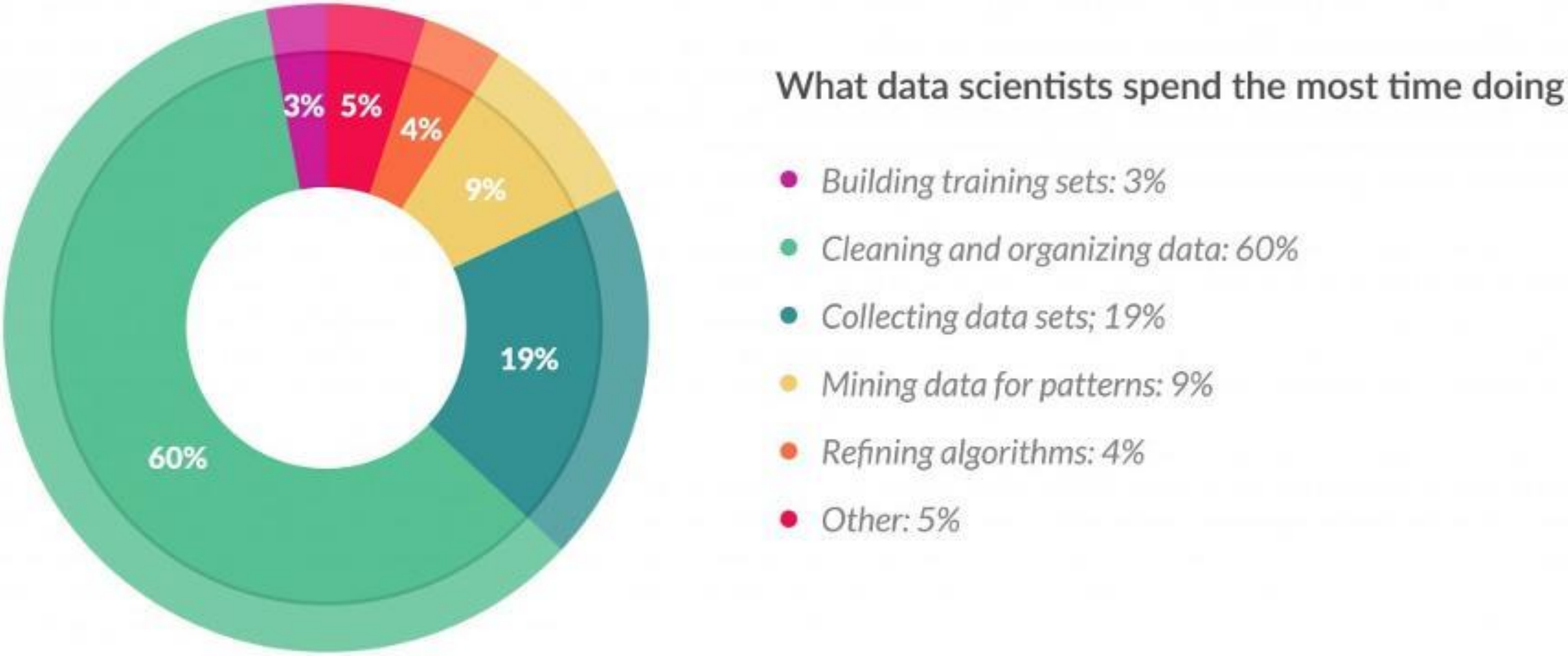
دولة قطر
وزارة التخطيط التنموي والإحصاء
Ministry of Development Planning and Statistics

ورشة عمل بشان تحديث الإحصاءات الرسمية في دولة قطر
Workshop on Modernization of Official Statistics in Qatar

الاثنين ١١ ديسمبر ٢٠١٧ – فندق روتانا, سيتي سنتر
Monday 11 December, 2017 - Rotana City Center Hotel - Doha

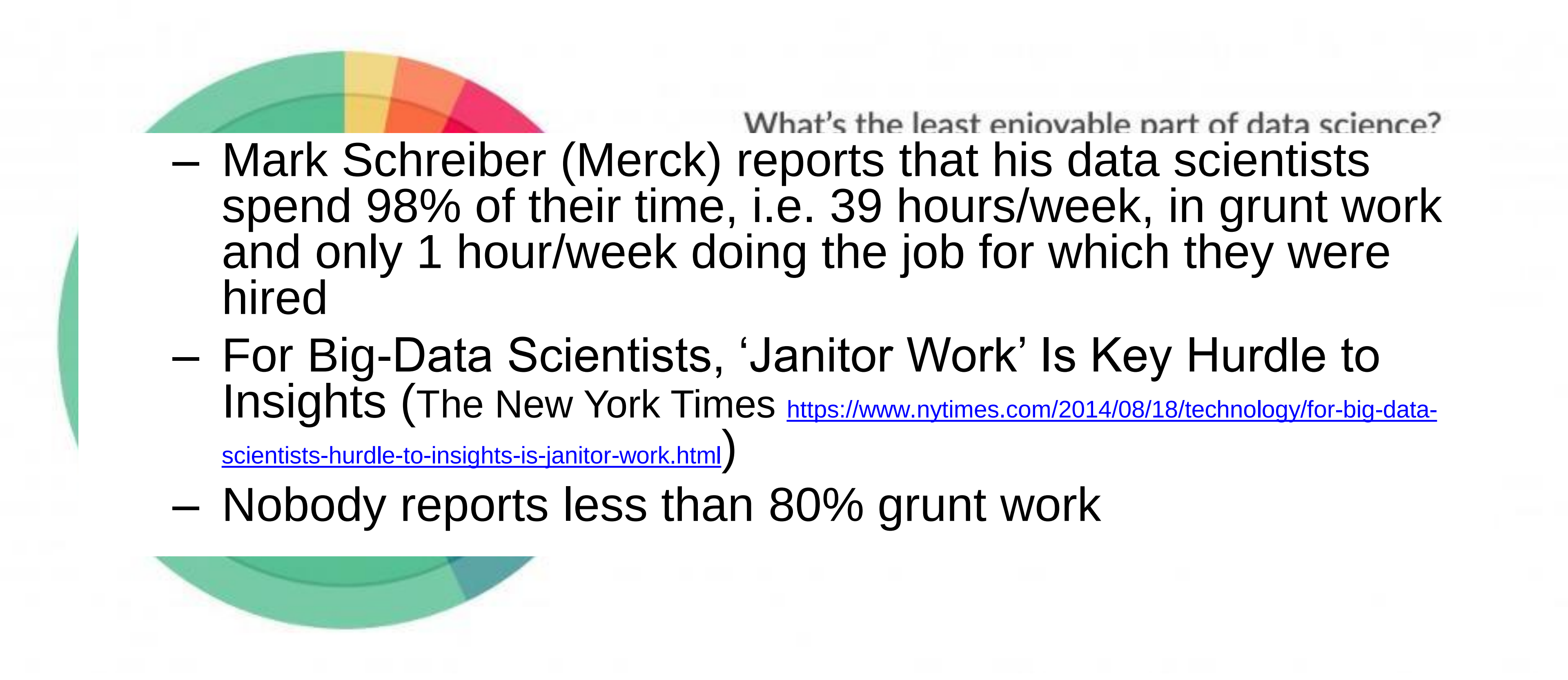
Agenda

- Why is this an important problem?
- Data Civilizer - An end-to-end system
- Overview of some key components



What data scientists spend the most time doing

http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf



- Mark Schreiber (Merck) reports that his data scientists spend 98% of their time, i.e. 39 hours/week, in grunt work and only 1 hour/week doing the job for which they were hired
- For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights (The New York Times https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html)
- Nobody reports less than 80% grunt work

http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

We’re building Data Civilizer to help …

✔**discover** data of interest from large numbers of data sets;

✔**link** and **enrich** relevant data sets;

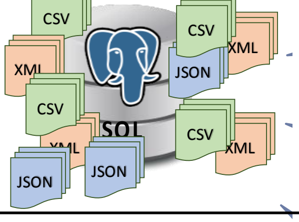
✔**deduplicate** and **consolidate** the data;

✔**clean** the data; and

✔**iterate** through these tasks using a workflow system.

Algorithms do the grunt work (80% of the pain) while data scientists can do what *they* are good at

A Typical Data Civilizer Pipeline



Profile and Index

Find relevant data

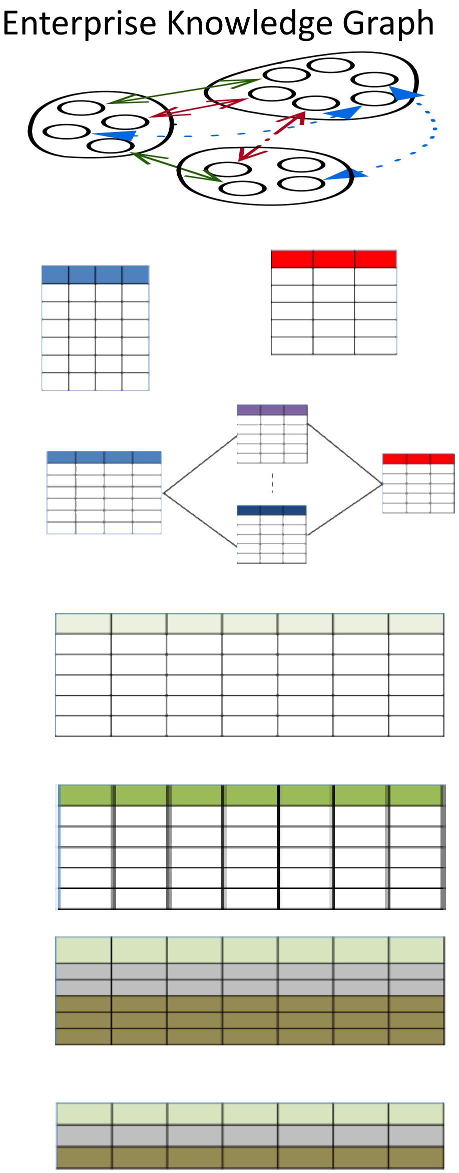
Find Join Paths

Union the Join Paths

Clean the data

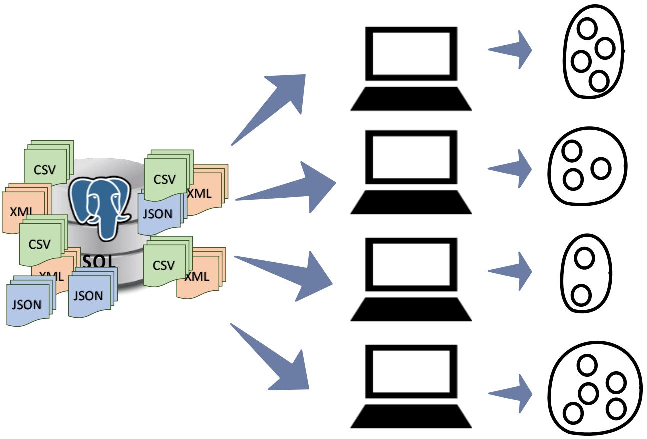
Entity Resolution

Entity Consolidation



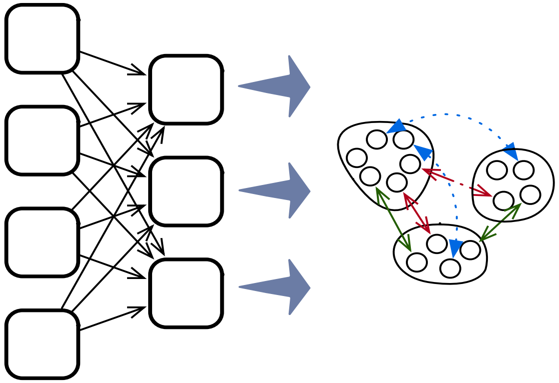
Data Discovery

Profiler
Create Summaries



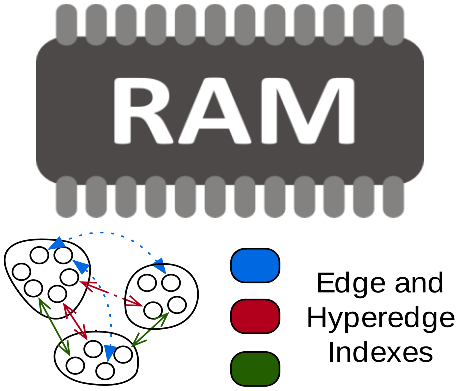
Distributed architecture to scale data summarization

Graph Builder
Connect Summaries



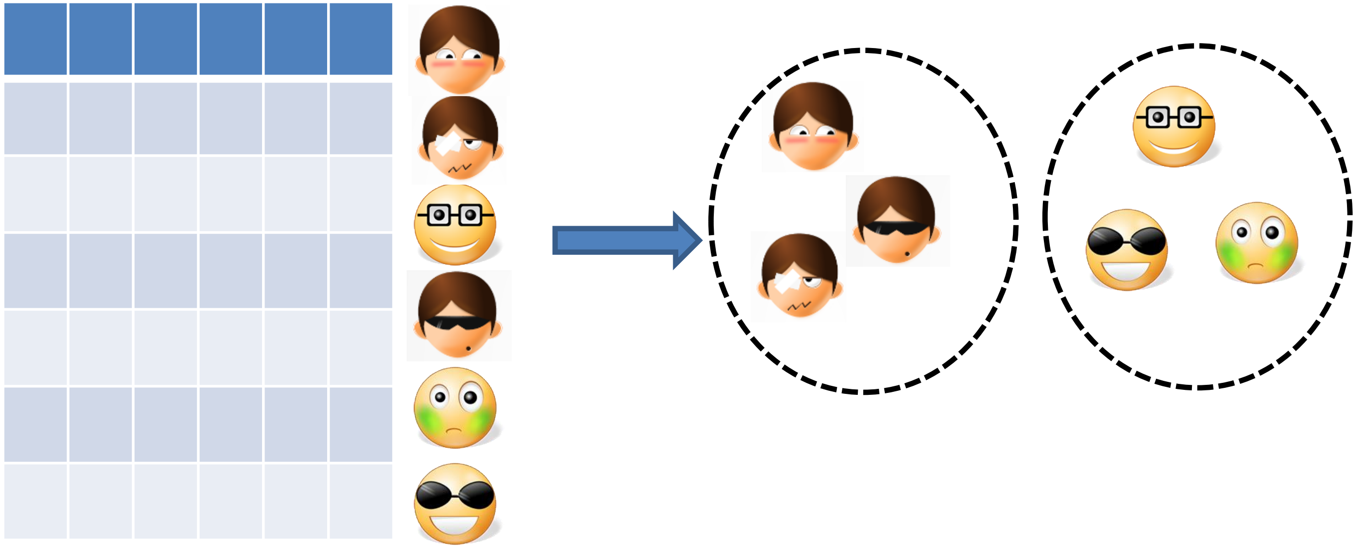
Scalable all-pairs comparison of multiple data types

SRQL Query Processing
Find relevant data



Concise in-memory indexes for interactive query answering

Entity Resolution using Deep Learning

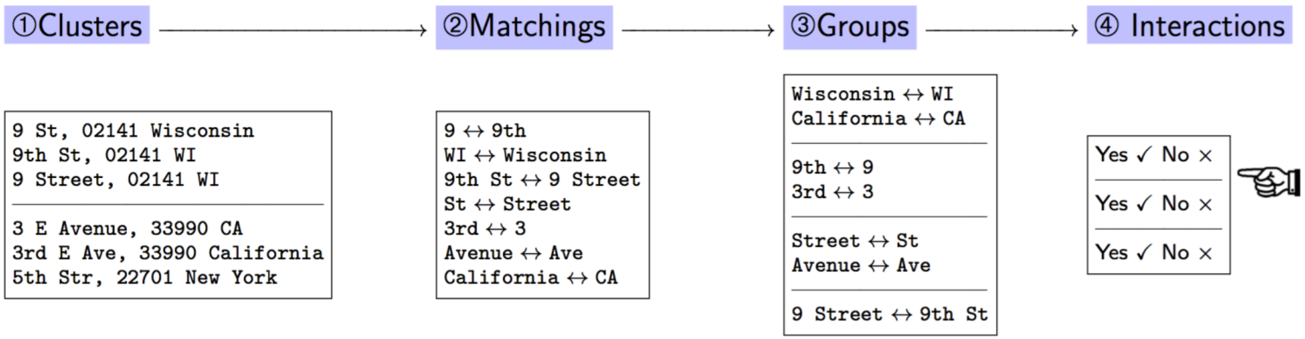


A turn key solution using distributed representation (DR) and deep learning (DL)

- Tuples → high dimensional vectors where (semantically) similar tuples have a high (cosine) similarity
- Using pre-trained DR dictionaries (e.g., GloVe which is trained on a corpus of 840B tokens) → no need for manual feature engineering
- Much less training data
- Competitive or superior results wrt prior state-of-the-art methods
- Locality Sensitive Hashing-based blocking
  - automated and semantic blocking based on the entire tuple
  - no need for blocking functions from domain experts

Entity Consolidation

From clusters of duplicate records to Golden Records



①Clusters → ②Matchings → ③Groups → ④ Interactions

①Clusters

9 St, 02141 Wisconsin
9th St, 02141 WI
9 Street, 02141 WI

3 E Avenue, 33990 CA
3rd E Ave, 33990 California
5th Str, 22701 New York

②Matchings

9 ↔ 9th
WI ↔ Wisconsin
9th St ↔ 9 Street
St ↔ Street
3rd ↔ 3
Avenue ↔ Ave
California ↔ CA

③Groups

Wisconsin ↔ WI
California ↔ CA

9th ↔ 9
3rd ↔ 3

Street ↔ St
Avenue ↔ Ave

9 Street ↔ 9th St

④ Interactions

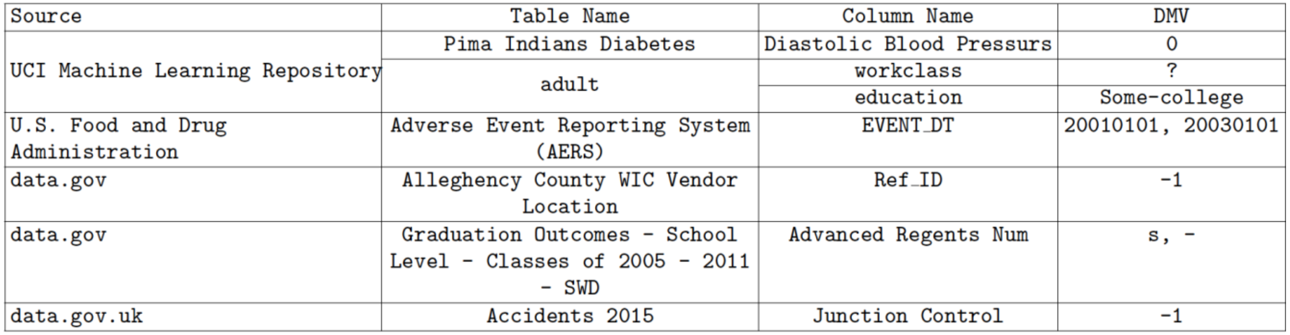
Yes ✓ No ×

Yes ✓ No ×

Yes ✓ No ×

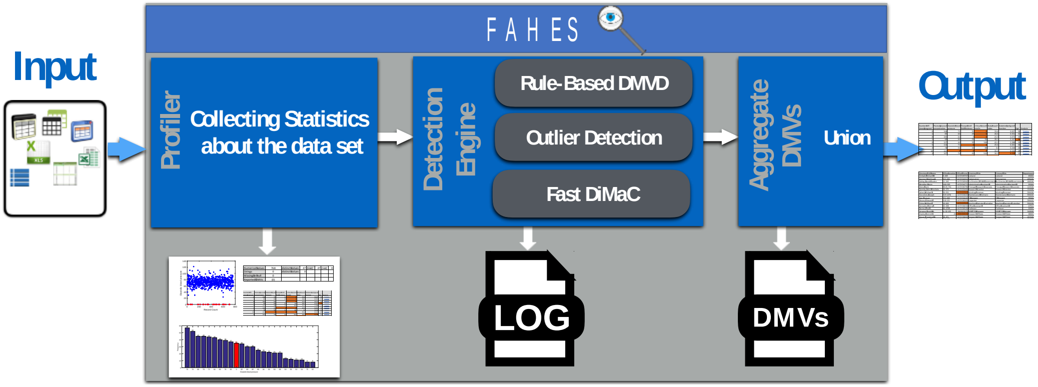
Cluster duplicates, detect matchings and group them, and ask a human

Detecting Disguised Missing Values



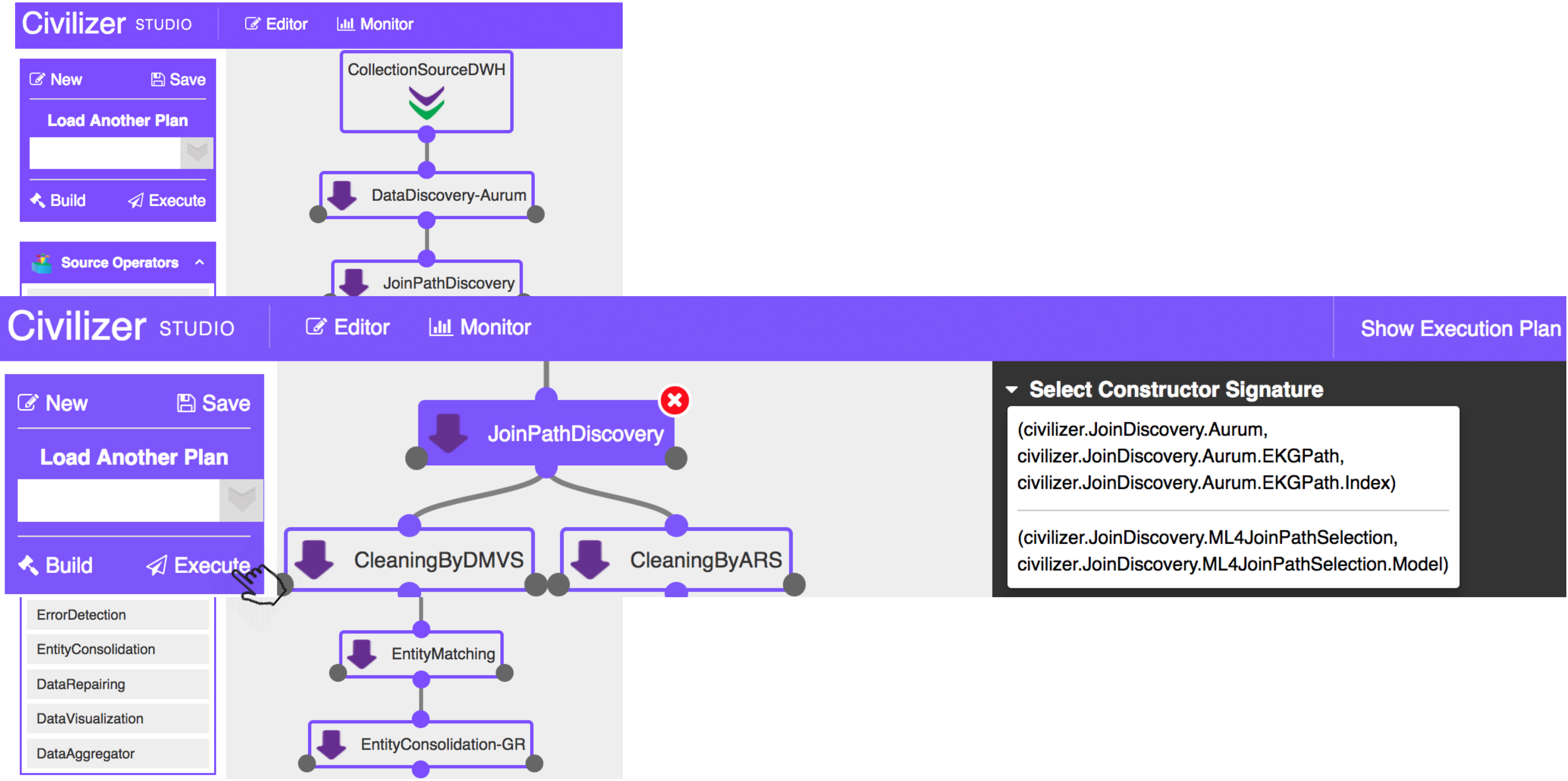
Source	Table Name	Column Name	DMV
UCI Machine Learning Repository	Pima Indians Diabetes	Diastolic Blood Pressurs	0
	adult	workclass	?
		education	Some-college
U.S. Food and Drug Administration	Adverse Event Reporting System (AERS)	EVENT_DT	20010101, 20030101
data.gov	Alleghency County WIC Vendor Location	Ref_ID	-1
data.gov	Graduation Outcomes - School Level - Classes of 2005 - 2011 - SWD	Advanced Regents Num	s, -
data.gov.uk	Accidents 2015	Junction Control	-1

DMV in different databases



- Rules to detect DMVs with special patterns, e.g., strings with repeated substrings
- Outlier detection algorithms
- A fast algorithm for detecting DMVs following a missing at random model

The Civilizer Studio – Gluing Things Together



Select Constructor Signature

(civilizer.JoinDiscovery.Aurum,
civilizer.JoinDiscovery.Aurum.EKGPath,
civilizer.JoinDiscovery.Aurum.EKGPath.Index)

(civilizer.JoinDiscovery.ML4JoinPathSelection,
civilizer.JoinDiscovery.ML4JoinPathSelection.Model)

Next Steps …

- Open-source release (ver 0.1)

- Get our technology in as many users' hands as possible

- Run tutorials in Spring 2018

شكراً

أسئلة؟

Thank You

Questions?