Data Tamer: A Scalable Data Curation System

by

Michael Stonebraker

DBg Database Group
MIT Computer Science and Artificial Intelligence Lab

How Does This Fit Into "Big Data"?

- I have too much of it
  - Volume problem
- It's coming at me too fast
  - Velocity problem
- It's coming at me from too many places
  - Variety problem

Database Group is Working on all Three Problems

- Volume problem
  - Working on SciDB; a DBMS for complex analytics
- Velocity problem
  - Working on integrating stream processing and high performance OLTP
- Variety problem
  - Working on Data Tamer

One Slide on SciDB

- Complex analytics is:
  - Machine learning, data clustering, data mining, predictive modeling
- Used in:
  - Recommendation engines, mass personalization of insurance products, predictive maintenance of complex machinery, genomics, …
- Defined on arrays, poorly served by RDBMSs
  - Write an array DBMS

One Slide on Velocity

- We wrote Aurora (StreamBase)
  - A stream processing engine
  - Good for "big pattern – little state" problems
- We wrote H-Store (VoltDB)
  - A high performance main-memory OLTP SQL DBMS
  - Good for "little pattern – big state" problems
- A lot of common machinery
  - Do both adding minimal complexity
  - By adding streams to H-Store

Data Curation

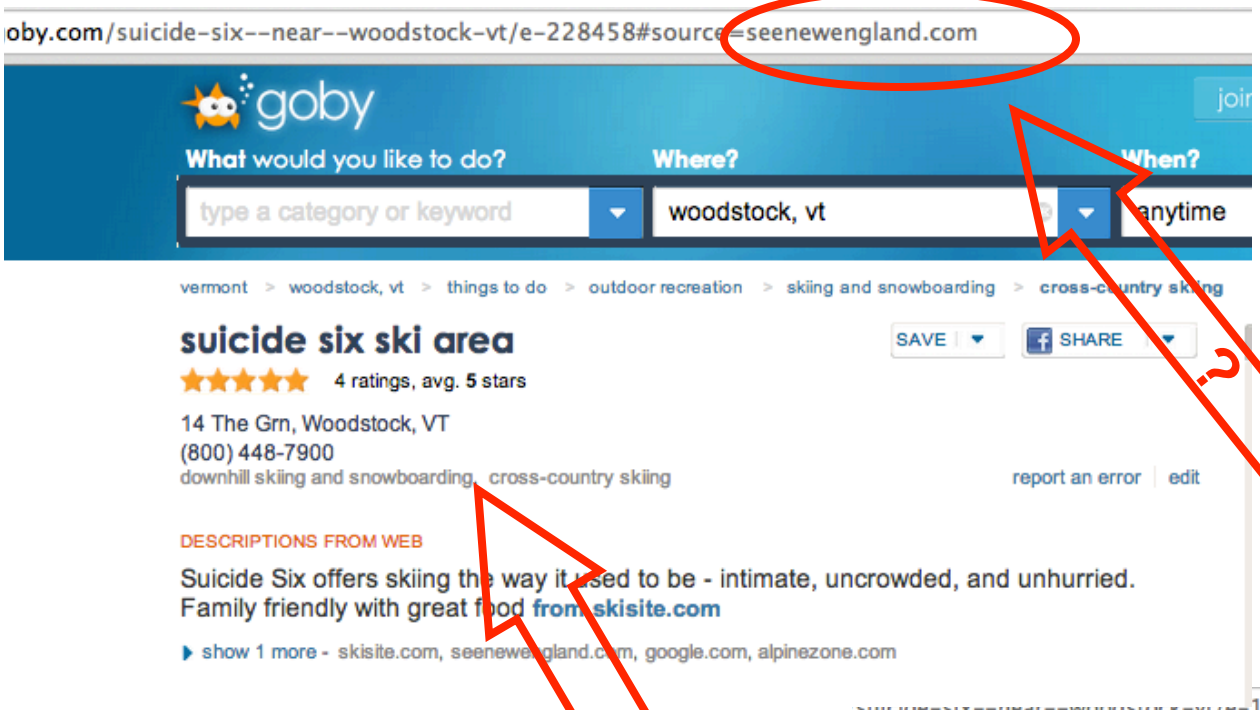
- Ingest
- Validate
- Transform
- Correct
- Consolidate (dedup)
- And visualize information to be integrated

Example Problem – Goby.com

- Web aggregator -- currently 80,000 URLs
- For “events” and “things to do”

Goby Problem

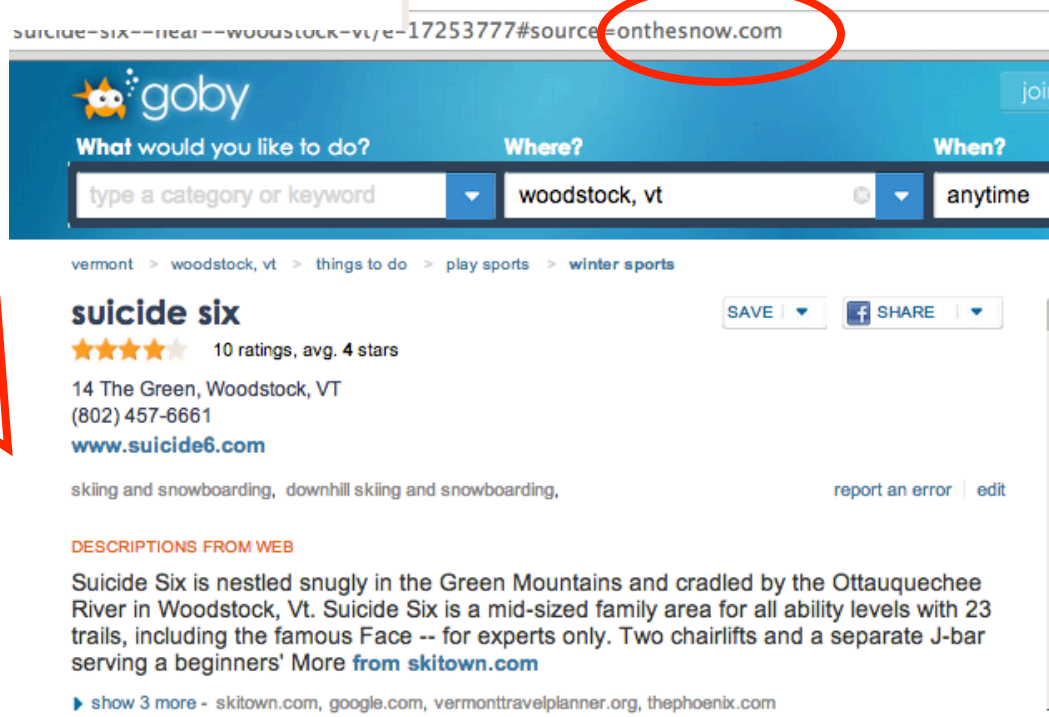
The Same ?



Different sources

Same address,

Different phone.



DETAILS FROM WEB

dining rating: 4.00/5
helicopter: No
peak elevation: 650 Ft / 198 M
service rating: 5.00/5
snow conditions: Yes
snowcast: No
terrain rating: 5.00/5
trail map: Yes
value rating: 5.00/5
vertical (ft): 650'
webcam: No

advanced trails %: 30%
beginner trails %: 30%
expert trails %: 0%
intermediate trails %: 40%
lift count: 3 - 2 Doubles; 1 Surface Lift
snowphone: (802) 457-1622
snowmaking: 50%
trail acreage: 100 skiable acres
trail count: 22
vertical drop (feets): 650 feet

DETAILS FROM WEB

lift tickets: Adult Junior Senior Weekday Full Day: $36 $30 $30 Weekday Half Day: $30 $23 $23 Weekend Full Day: $55 $40 $40 Weekend Half Day:...
state: Vermont
advanced runs: 30
average annual snowfall: 90"
base elevation: 550ft
beginner runs: 30
expert runs: 0
intermediate runs: 40
number of double chairs: 2
number of high speed quads: 0
number of high speed sixes: 0

number of regular quads: 0
number of surface lifts: 1
total lifts: 3
number of gondolas and trams: 0
number of triple chairs: 0
visitor recommendations: Single/Newlyweds: 101%, Beginner: 101%, Families: 126%, Intermediate: 126%, Empty Nesters: 101%, Advanced: 76%
summit elevation: 1200ft
vertical drop: 650ft
years open: 79
event category: Kids Activities Recreation Ski Area

Traditional Wisdom - ETL

- Human defines a global schema
- Assign a programmer to each data source to
  - Understand it
  - Write local to global mapping (in a scripting language)
  - Write cleaning routine
  - Run the ETL
- Scales to (maybe) 25 data sources

What About the Rest of Your Data Sources?

- Typical enterprise has 5K data sources inside the firewall
- And wants additional public ones
- Web aggregators (like Goby) have a lot more

Other "Long Tail" Application

- Novartis
  - Integrates 8000 "lab notebooks"
- Verisk Health
  - Integrates 300 medical insurance sources

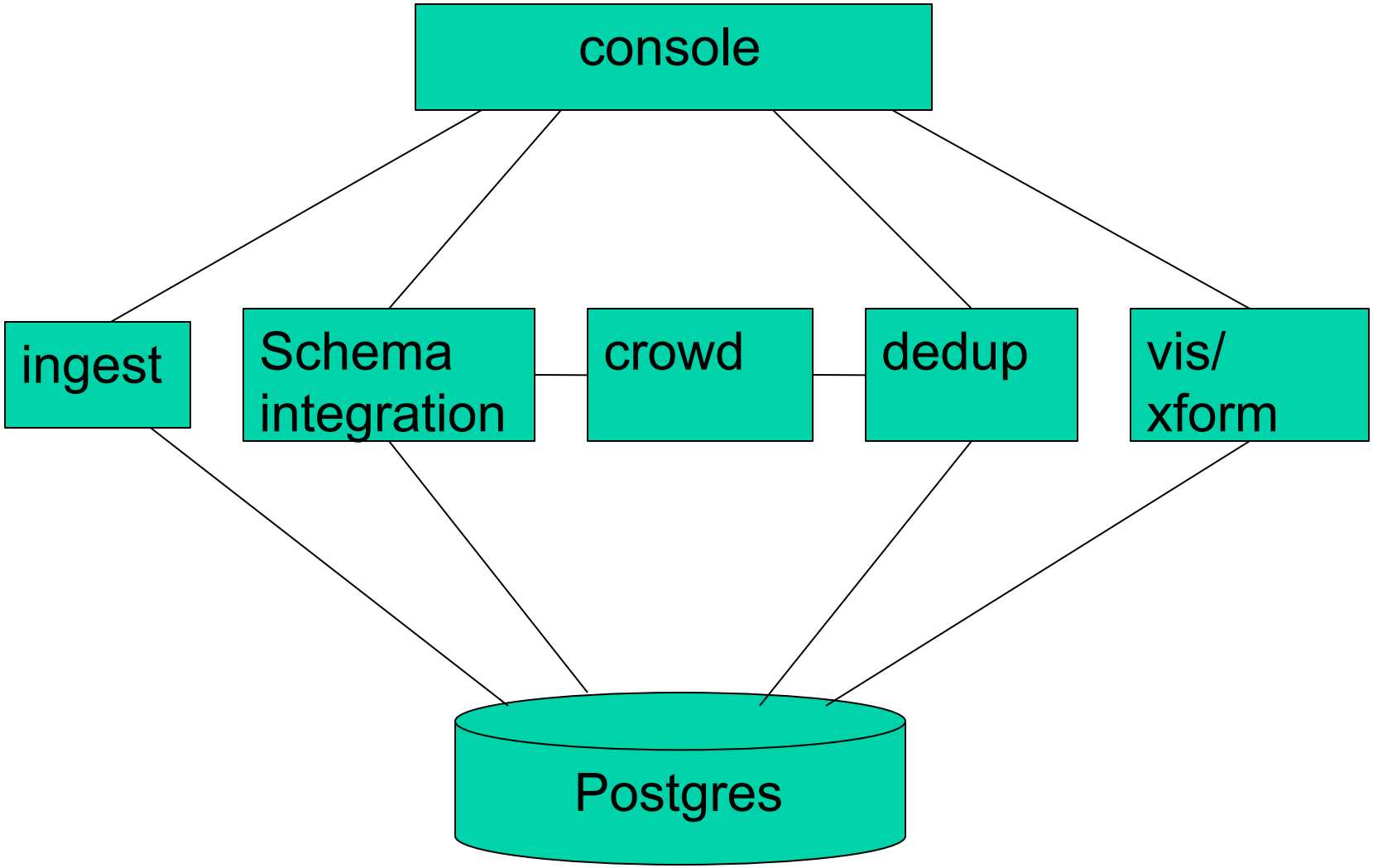
In All Cases….

- Integration is
  - Ad-hoc
  - Application-specific
- But everybody is
  - Doing roughly the same thing

Data Tamer Goals

- Do the "long tail"
  - Better/cheaper/faster than the ad-hoc techniques being used currently
- By inverting the normal ETL architecture
  - Machine learning and statistics
  - Ask for human help only when automatic algorithms are unsure

Data Tamer Architecture



Data Tamer -- Ingest

- Assumes (for now) a data source is a collection of records, each a collection of (attribute-name, value) pairs.
- Loaded into Postgres
  - We believe in databases!

Data Tamer – Schema Integration

- Must be told whether there is a predefined partial or complete global schema or nothing
- Starts integrating data sources
  - Using synonyms, templates, and authoritative tables for help
  - 1st couple of sources require asking the crowd for answers
  - System gets better and better over time

Data Tamer – Schema Integration

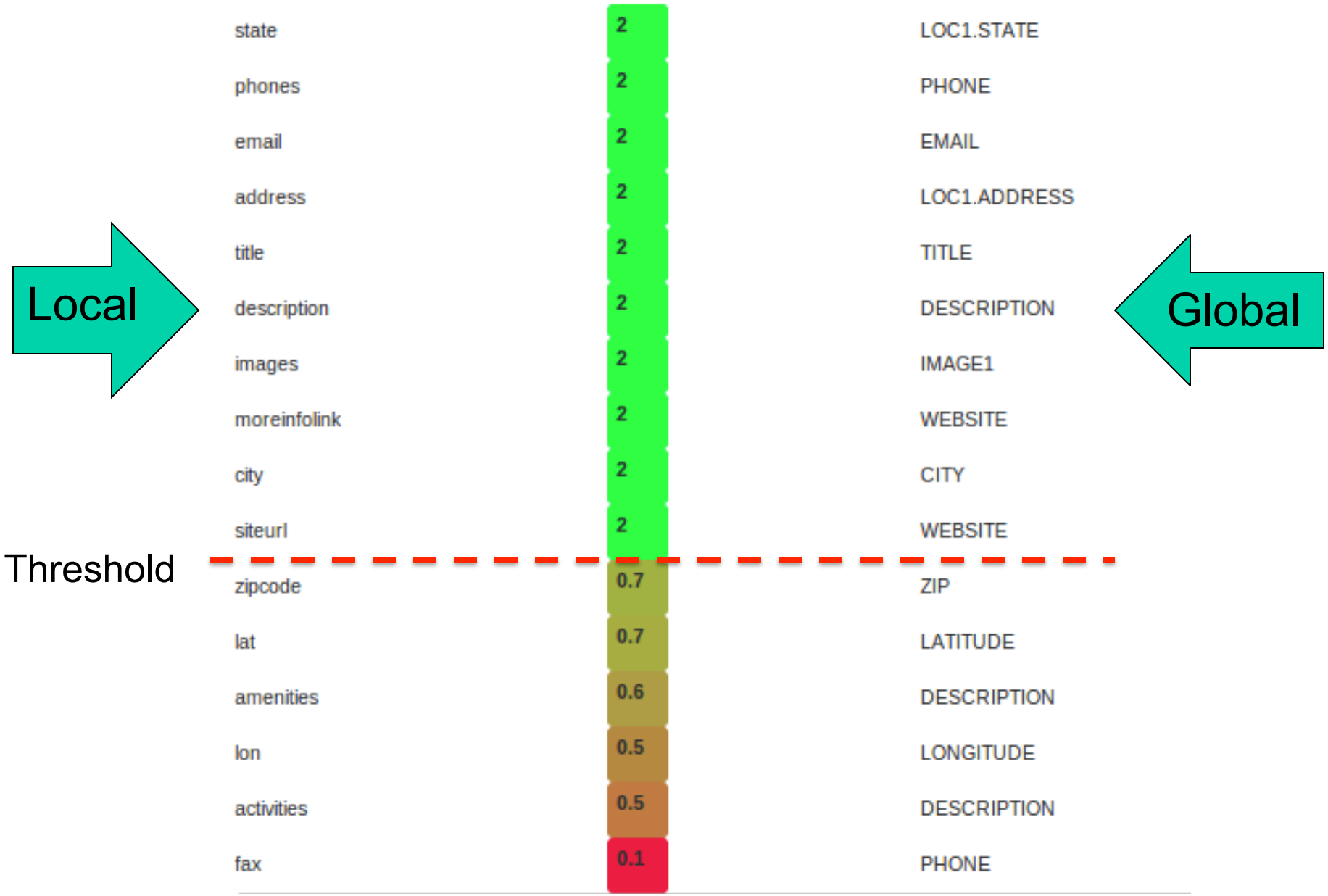
- Inner loop is a collection of experts
  - T-test on the data
  - Cosine similarity on attribute names
  - Cosine similarity on the data
- Scores combined heuristically
- After modest training, get 90% of the matching attributes on Goby and Novartis automatically
  - Cuts human cost dramatically

Data Tamer – Crowd Sourcing

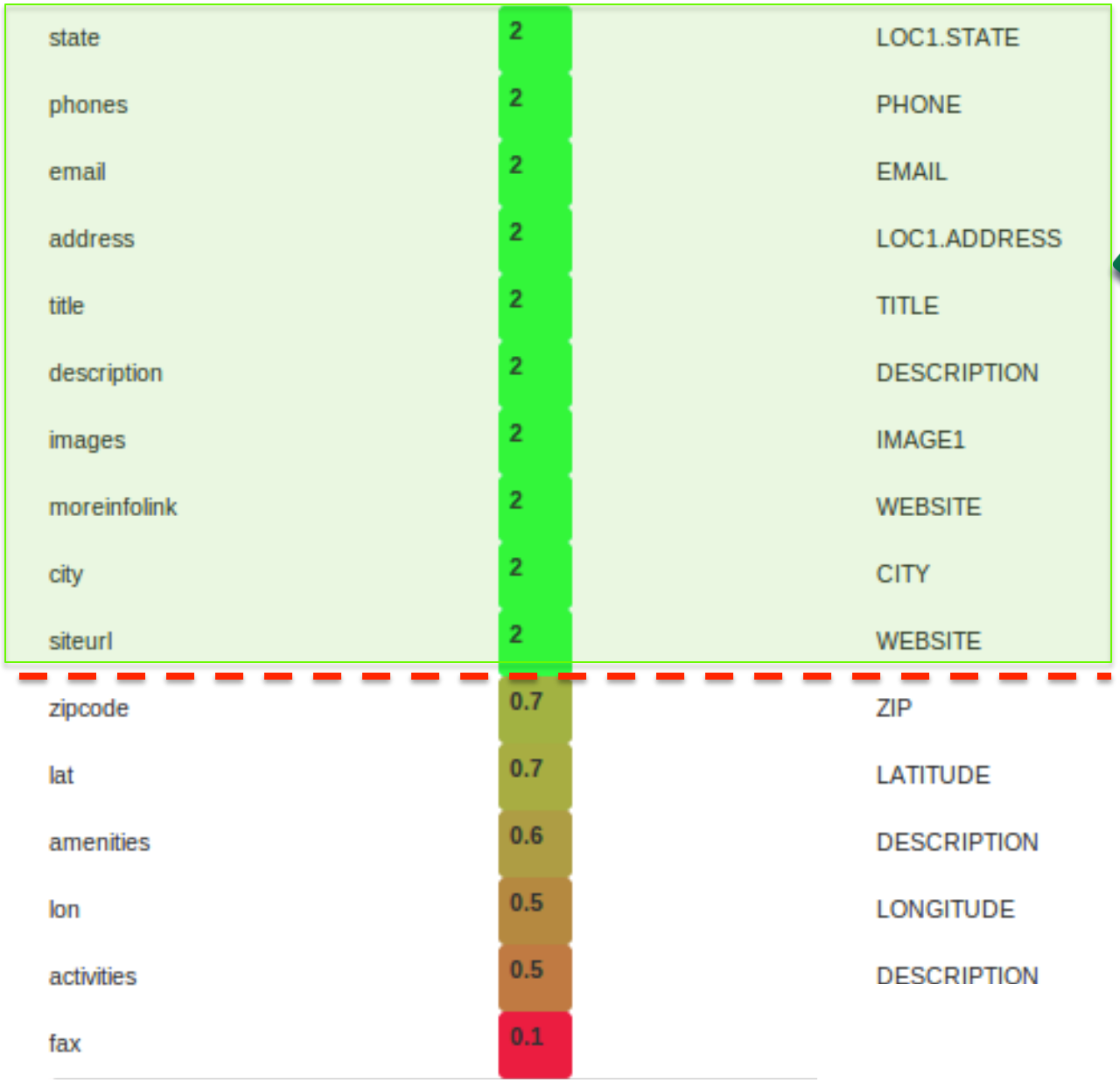
- Hierarchy of experts
- With specializations
- With algorithms to adjust the "expertness" of experts
- And a marketplace to perform load balancing
- Currently doing a large scale evaluation at Novartis
  - Late flash: it works!!!!

Schema Mapping Suggestions

Schema Mapping Suggestions

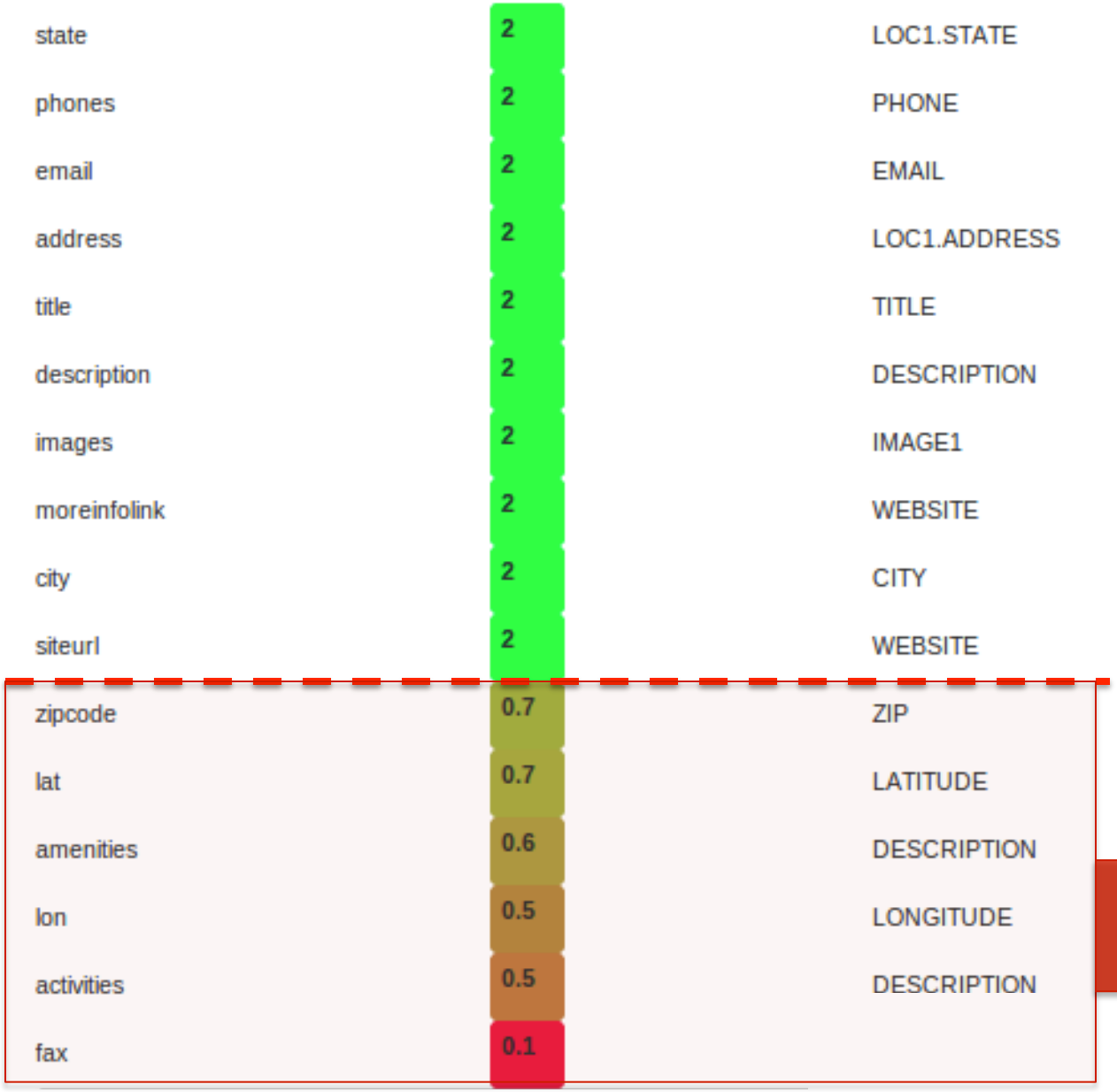


Schema Mapping Suggestions



Threshold

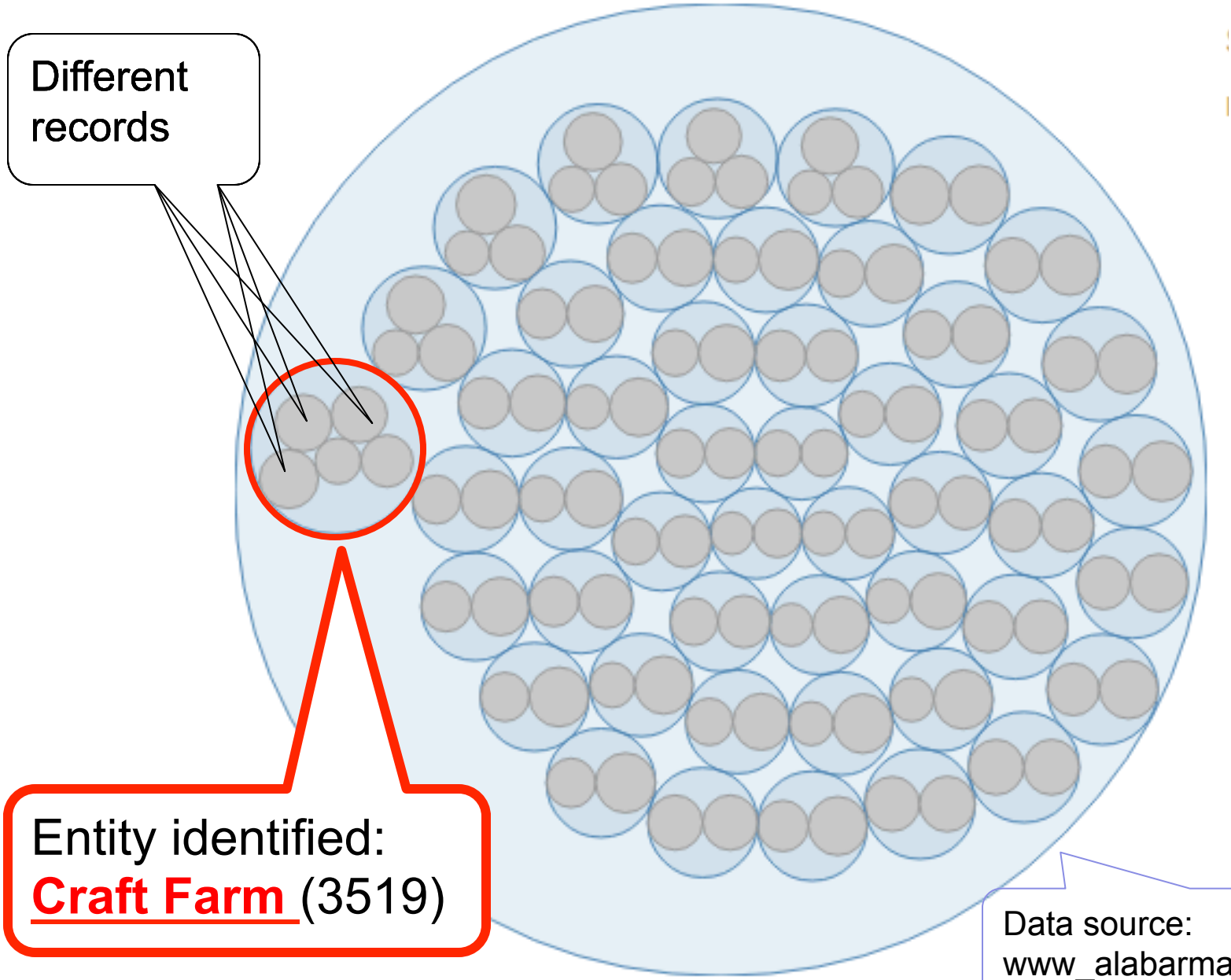
Schema Mapping Suggestions



Data Tamer – Entity Consolidation

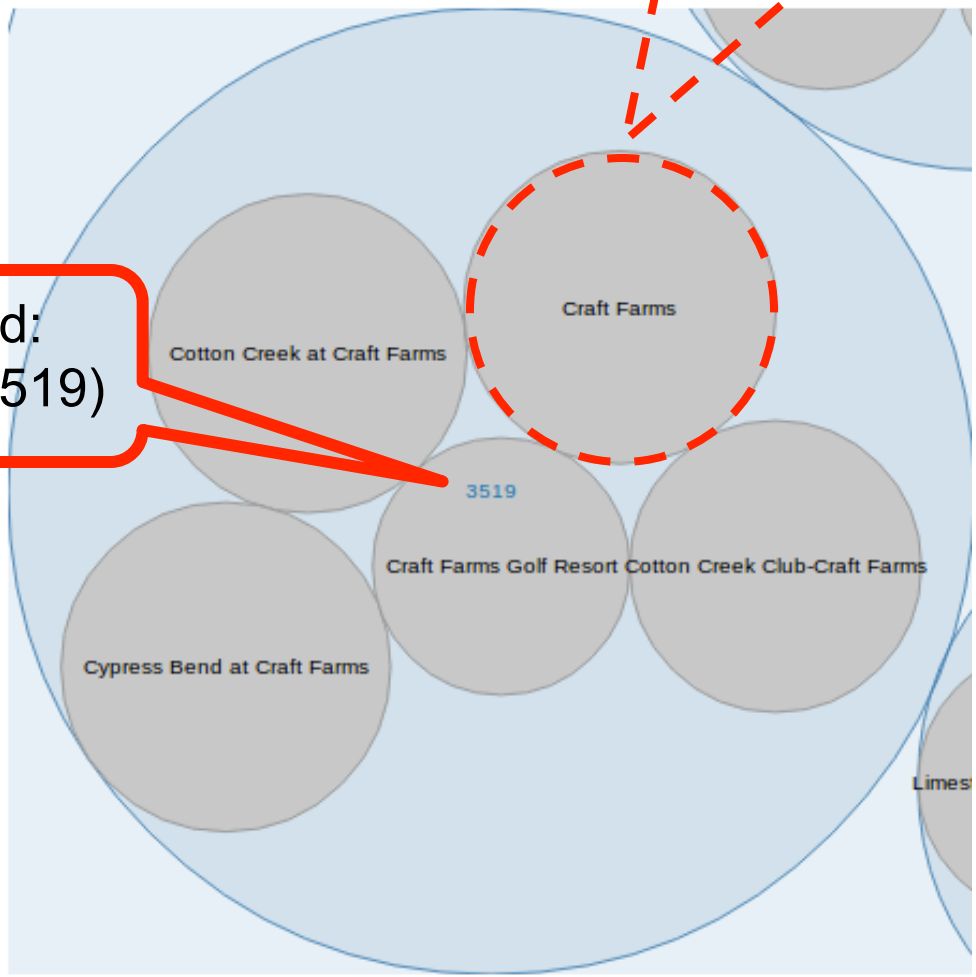
- On tables defined by schema integration module
- Entity matching on all attributes, weighted by value presence and distribution
- Basically a data clustering problem
- With a first pass to try to identify "blocks" of records
  - Otherwise N ** 2 in the number of records
- Wildly better than Goby; a bit better than domain-specific Verisk module

Entity Clusters



Select a specific record

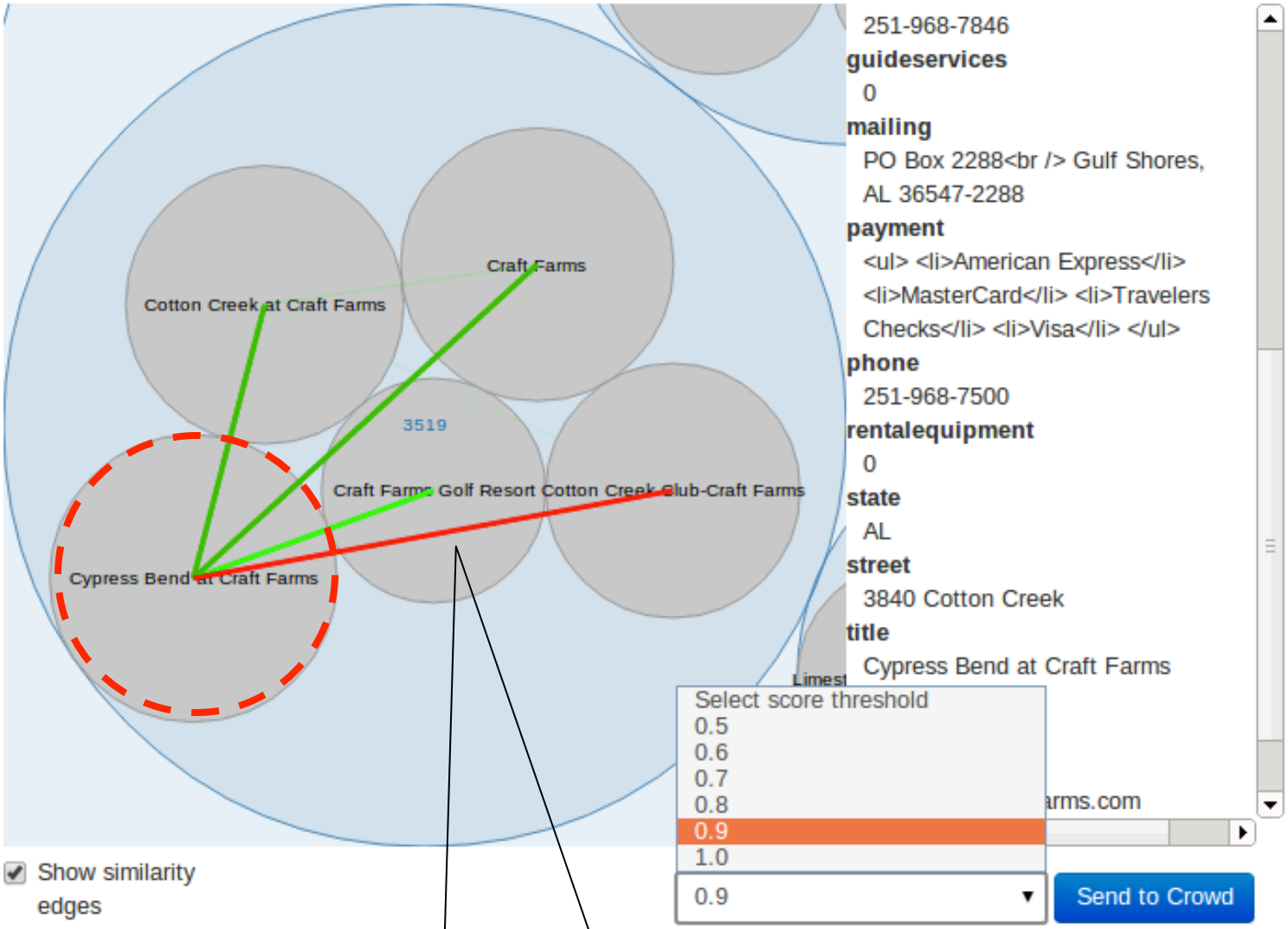
Entity identified:
Craft Farm (3519)



Show similarity edges

operationschedule
Open daily, 6:30 until sunset.
Closed Christmas Day.
payment
 Travelers Checks
Visa MasterCard
American Express
phone
251-968-7500
rentalequipment
0
state
AL
street
3840 Cotton Creek Cir.
title
Craft Farms
tollfree
877-482-2895
website
http://www.craftfarms.com
zipcode
36542

Record details



Green line: above threshold
Red line: below threshold

Data Tamer Status

- Schema mapping and entity consolidation work well
- Crowd sourcing seems to work
- Vis (yet to begin)
- Cleaning (will integrate DBWipes)
- Transformations (will adapt Data Wrangler)

Data Tamer Future

- Commercial company
  - Solicit/run pilots
  - Do serious scalability (heavy lifting)
  - adaptors
- MIT/QCRI
  - Algorithms
  - Text
  - Automatic transformations
  - Crowd experiments