
# UNDERSTANDING DATA USING HUMAN COMPUTATION AND MACHINE LEARNING

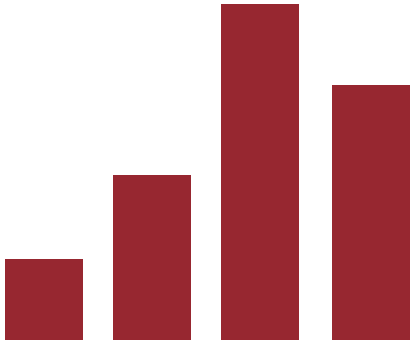
Deep Dhillon, Chief Technology Officer, Socrata

deep dhillon| cto @socrata | @zang0

CUSTOMER 2014 SUMMIT

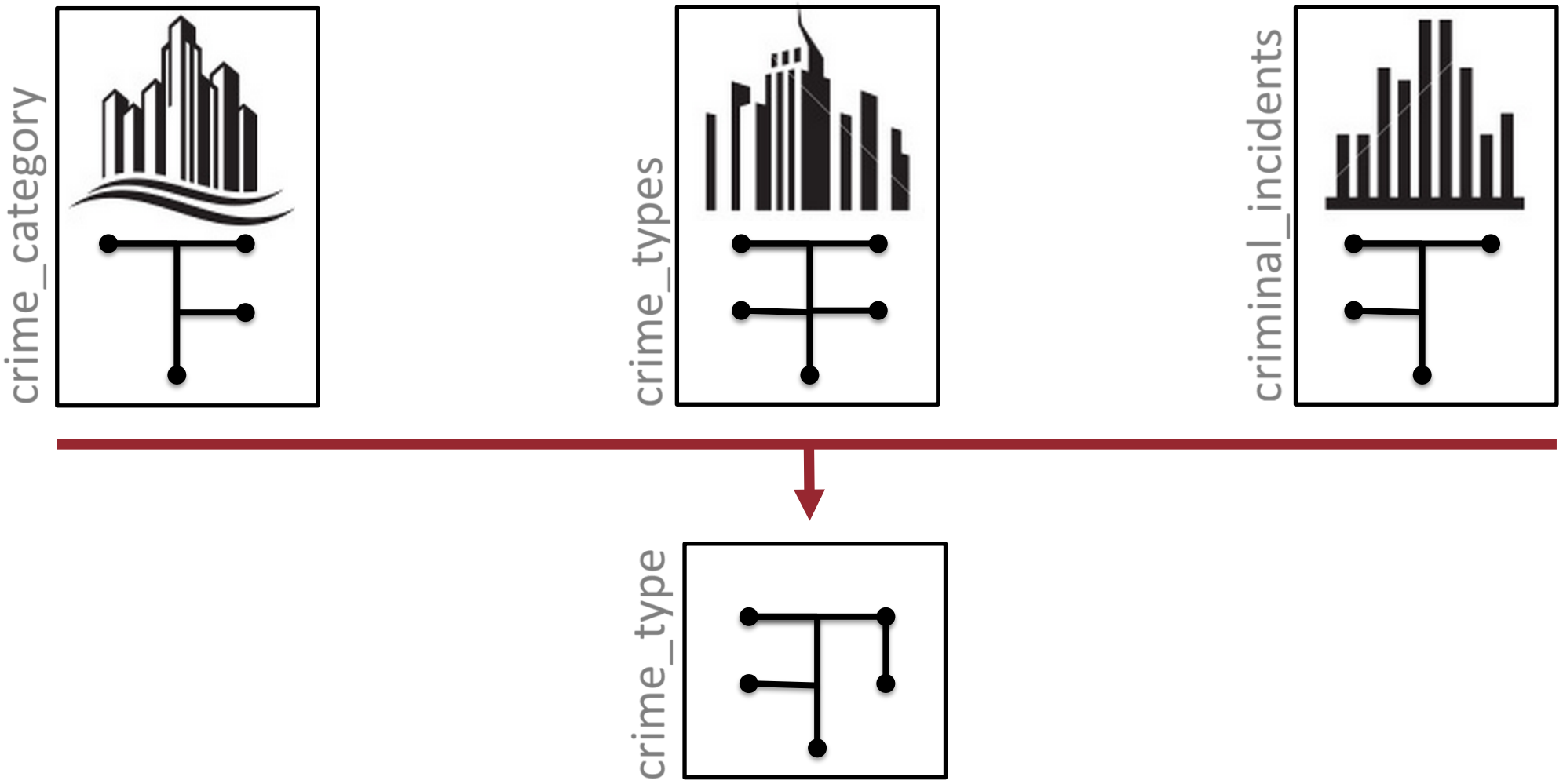

CUSTOMER
2014 SUMMIT

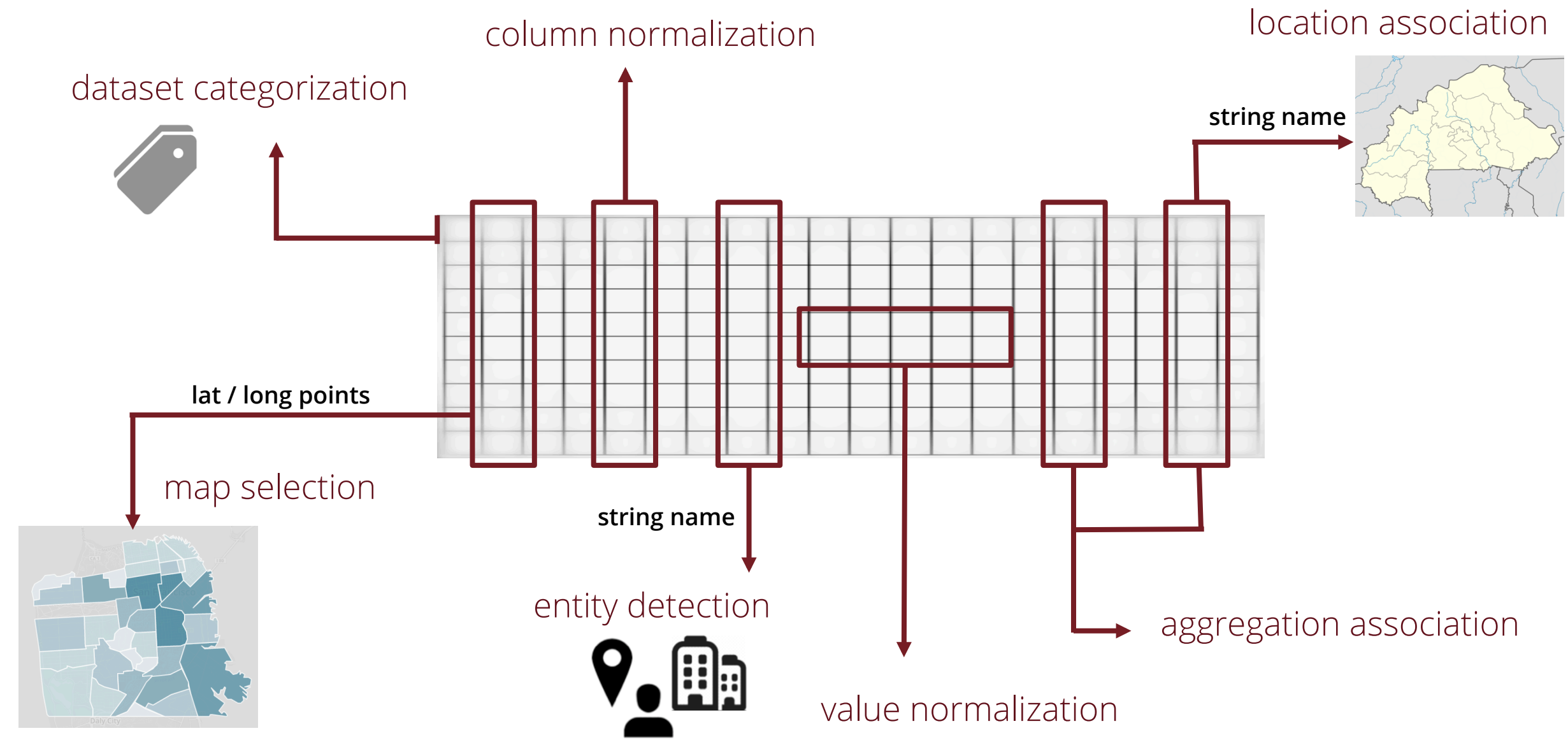
# why understand the data?



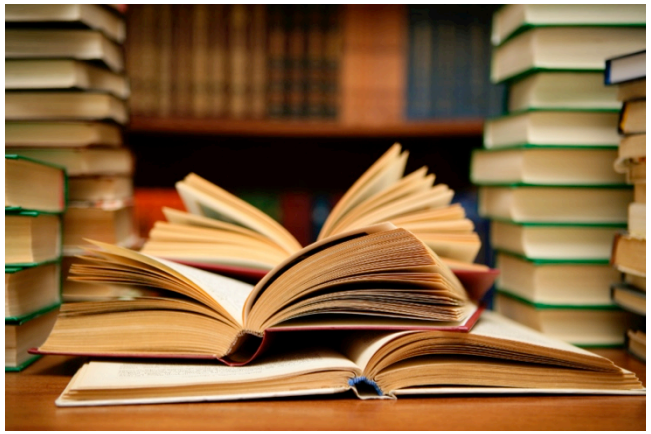
- improved ux
- comparisons
- recommendations
- search
- correlations/predictions

# normalizing data structure and values



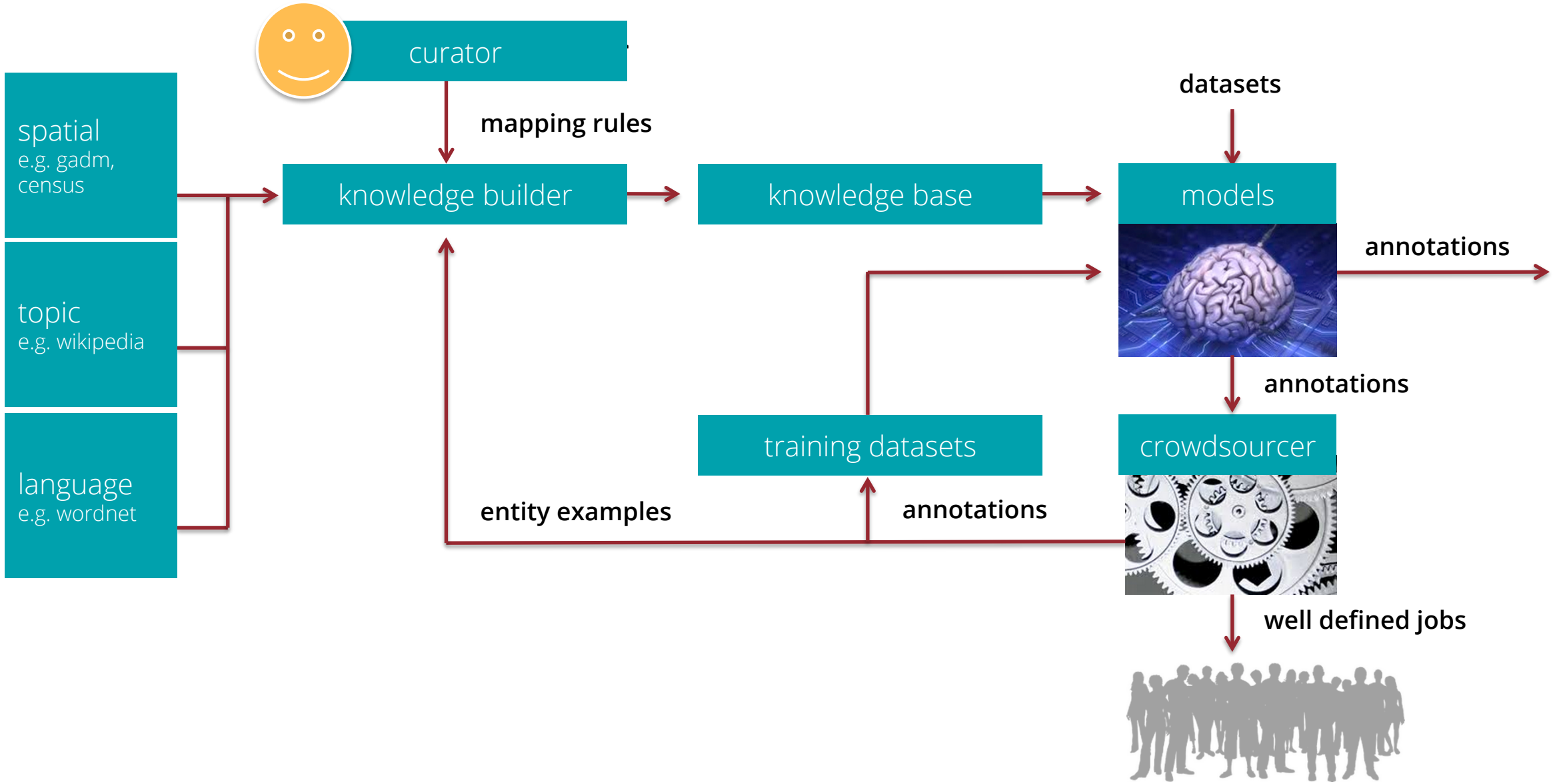

dataset categorization
column normalization
location association
string name
lat / long points
map selection
string name
entity detection
value normalization
aggregation association

knowledge



humans

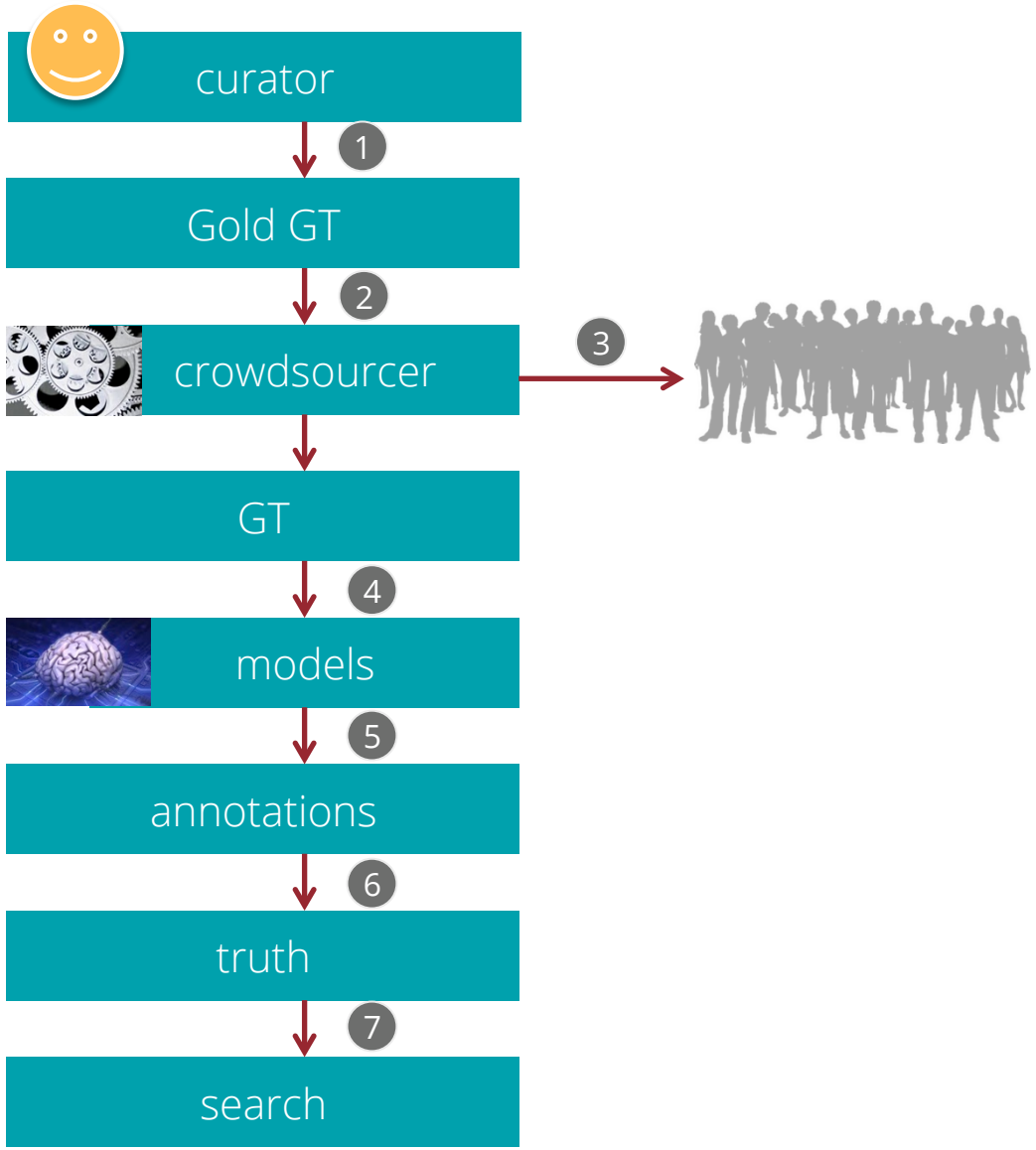
machines



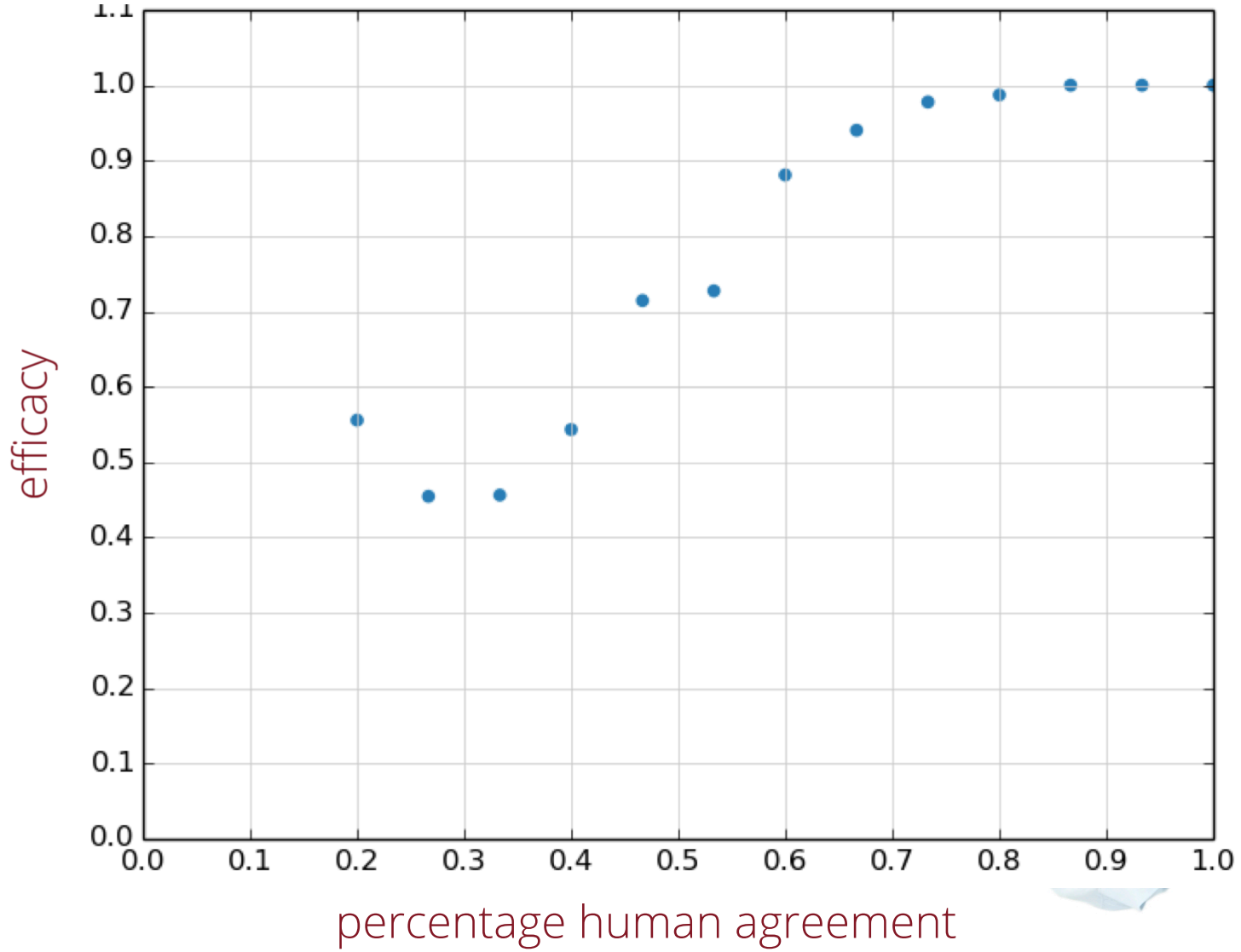

# model construction feedback loop

# annotations

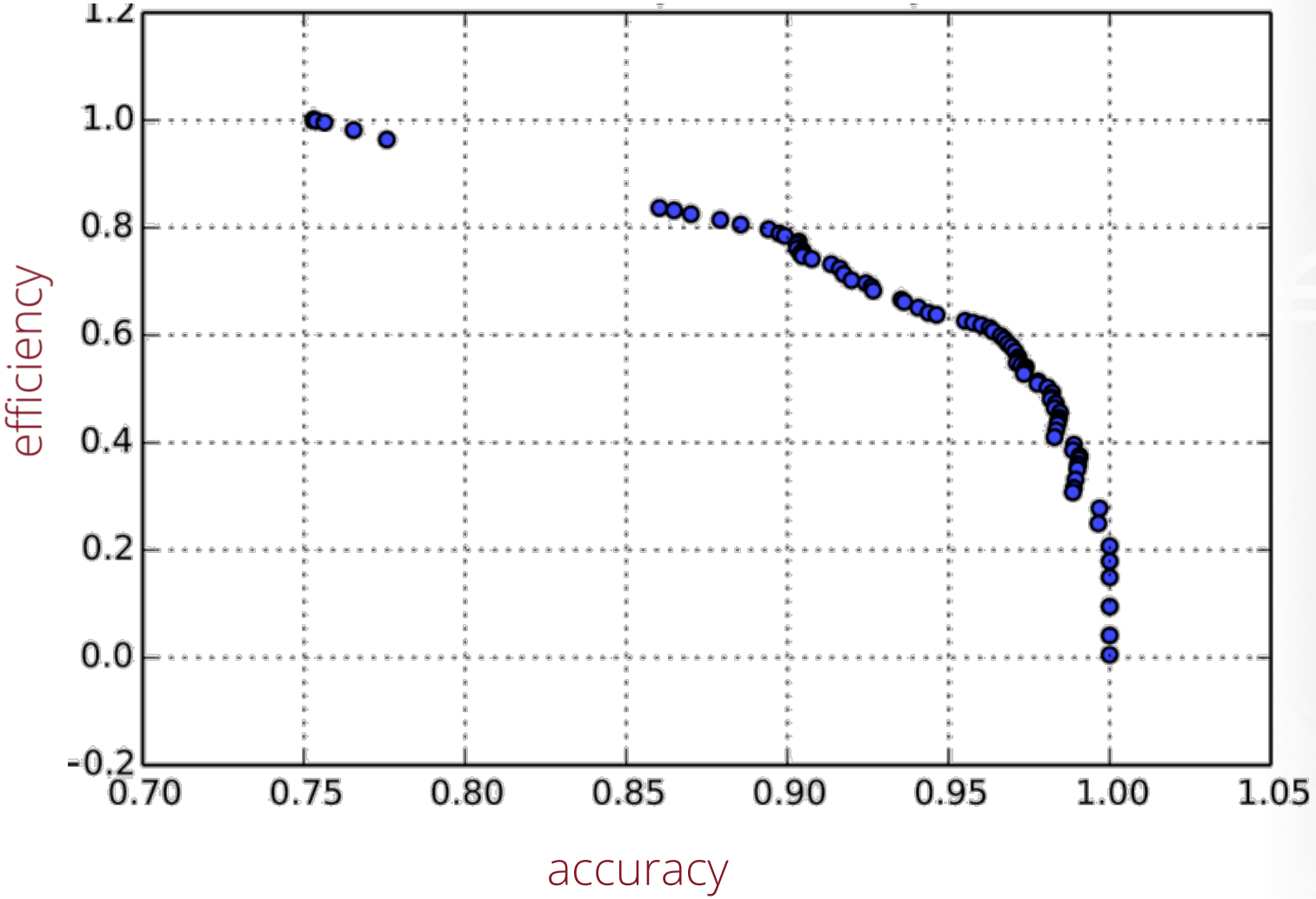
1. A trusted curator prepares a gold ground truth (gt) by manually labeling datasets.
2. The gold gt is used by a crowdsourcer system which coordinates jobs across untrusted distributed mechanical turk workers to annotate the gold gt set. Annotation quality is assessed via the gold gt.
3. The crowdsourcer leverages the distributed humans to annotate much larger sets of datasets.
4. Machine learning models are trained against the gt and applied against a larger set of datasets. In addition, trained models are applied in the synchronous workflow described above.
5. Model based and crowd sourced annotations are stored in an annotation service.
6. The truth system periodically queries the knowledge system for the latest annotation mappings and applies them to its datasets.
7. Secondary services like search are notified of changes. They pick them up and are now available for query.



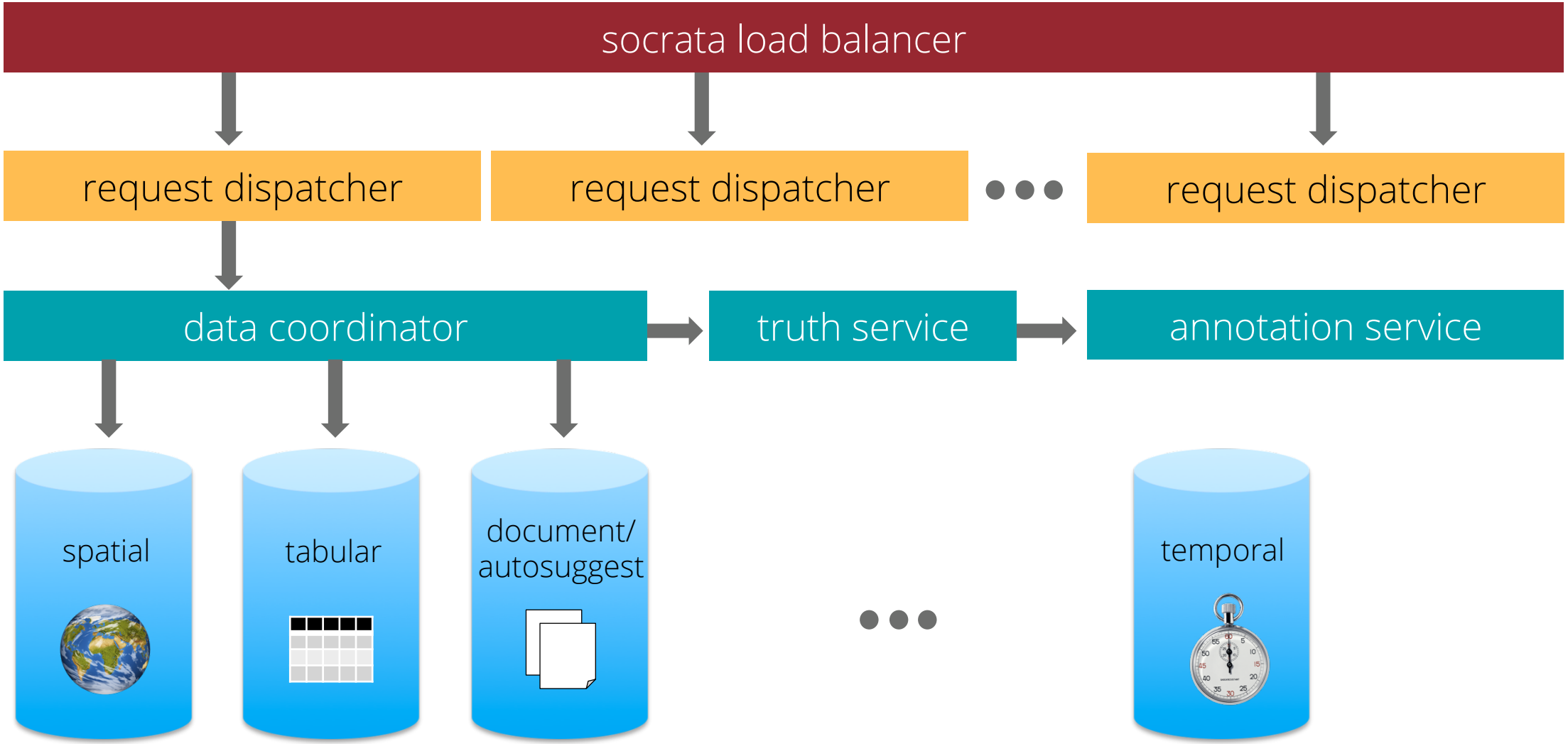
# demo




how do the humans do ?




how do the machines do?




dataset additions/updates

# what can i do?

## 1. use the product

| publisher usage | citizen usage |
| --- | --- |
| • spatial boundaries<br>• curatorial validations<br>  – accept / reject dataset cleanup<br>  – validate metadata recommendations<br>  – define app schema mappings to general sets | • spatial boundary importance<br>• view consumptions inform automated view guesses |

## 2. provide more data to the system

questions? | deep dhillon | cto @socrata | @zang0