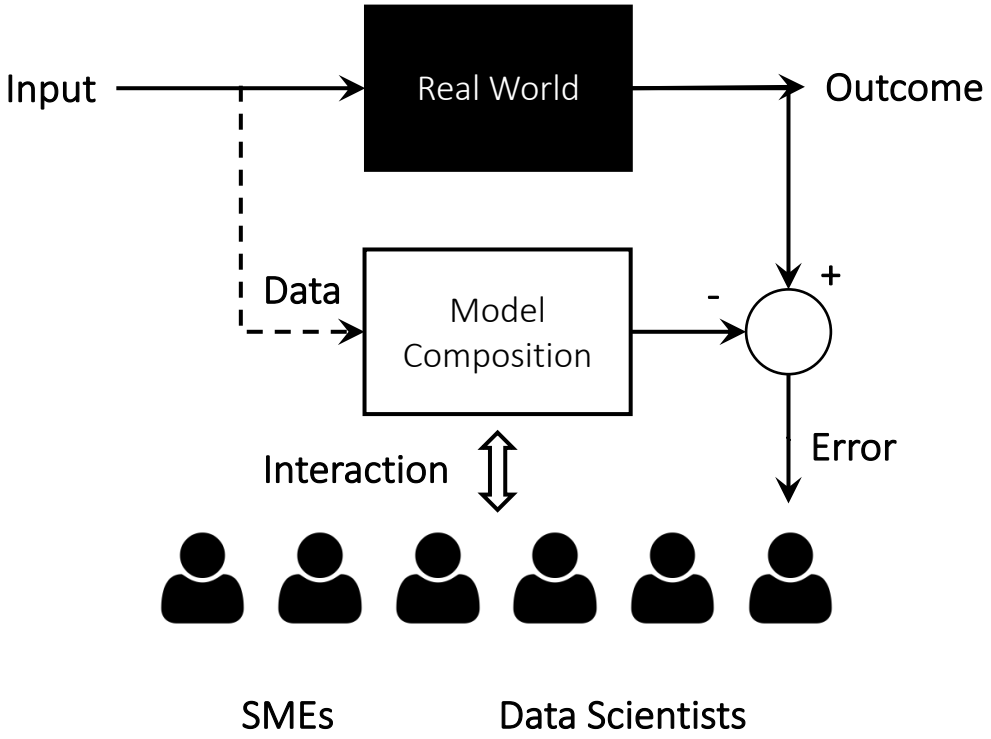
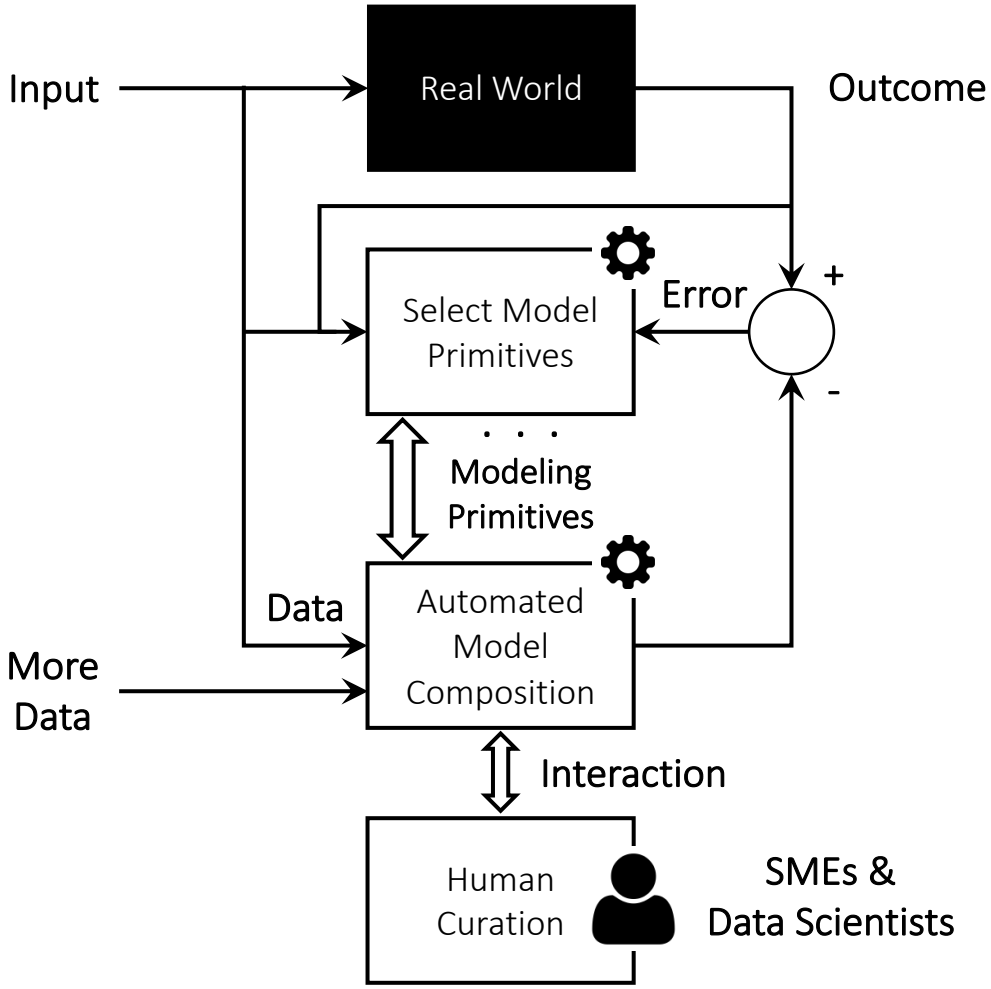
Auctus: A Dataset Search Engine for Data Augmentation

Fernando Chirigati, Rémi Rampin, Aécio Santos, and Juliana Freire


Input
Real World
Outcome
Data
Model
Composition
-
+
Error
Interaction
SMEs
Data Scientists

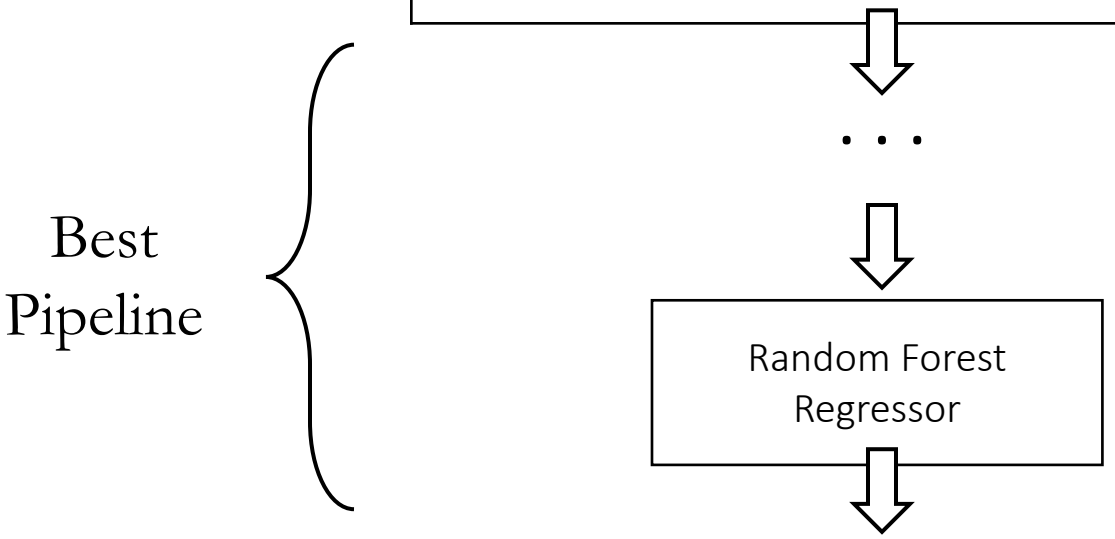
Data-Driven Discovery of Models (D3M)



https://www.darpa.mil/program/data-driven-discovery-of-models

NYC Yellow Taxi Data (Jan-Jun 2017)

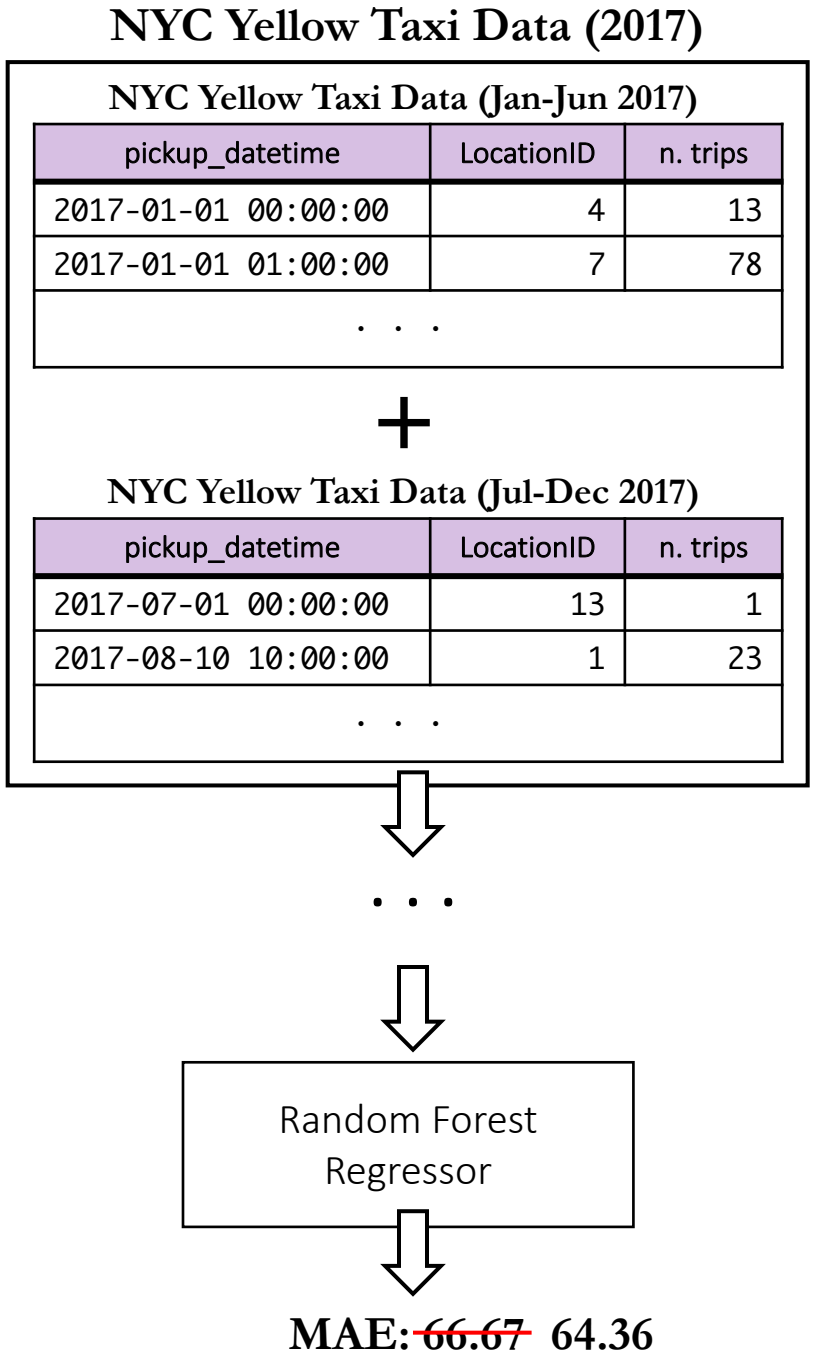
pickup_datetime	LocationID	n. trips
2017-01-01 00:00:00	4	13
2017-01-01 01:00:00	7	78
2017-02-01 10:00:00	12	39
2017-01-10 13:00:00	13	104
2017-03-03 10:00:00	14	128
. . .		



MAE: 66.67

How to further reduce the error of the model?

Data Augmentation



More Records

NYC Yellow Taxi Data & Weather

NYC Yellow Taxi Data (Jan-Jun 2017)

pickup_datetime	LocationID	n. trips
2017-01-01 00:00:00	4	13
2017-01-01 01:00:00	7	78
2017-02-01 10:00:00	12	39
2017-01-10 13:00:00	13	104
2017-03-03 10:00:00	14	128
. . .		

+

NYC Weather Data

time	temp	precip
2017-01-01 00:00:00	7.2	0.0
2017-01-01 01:00:00	7.2	0.0
2017-02-01 10:00:00	5.0	1.0
2017-01-10 13:00:00	10.0	0.0
2017-03-03 10:00:00	9.6	2.0
. . .		

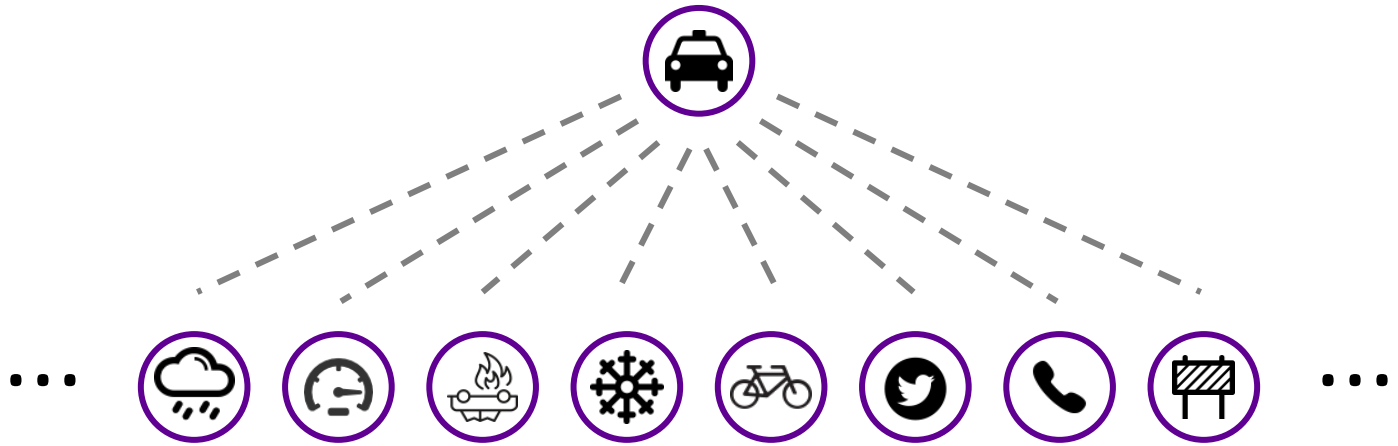
More Features

. . .

Random Forest Regressor

MAE: ~~66.67~~ 39.30

NYC Yellow Taxi Data



Datasets on the Web

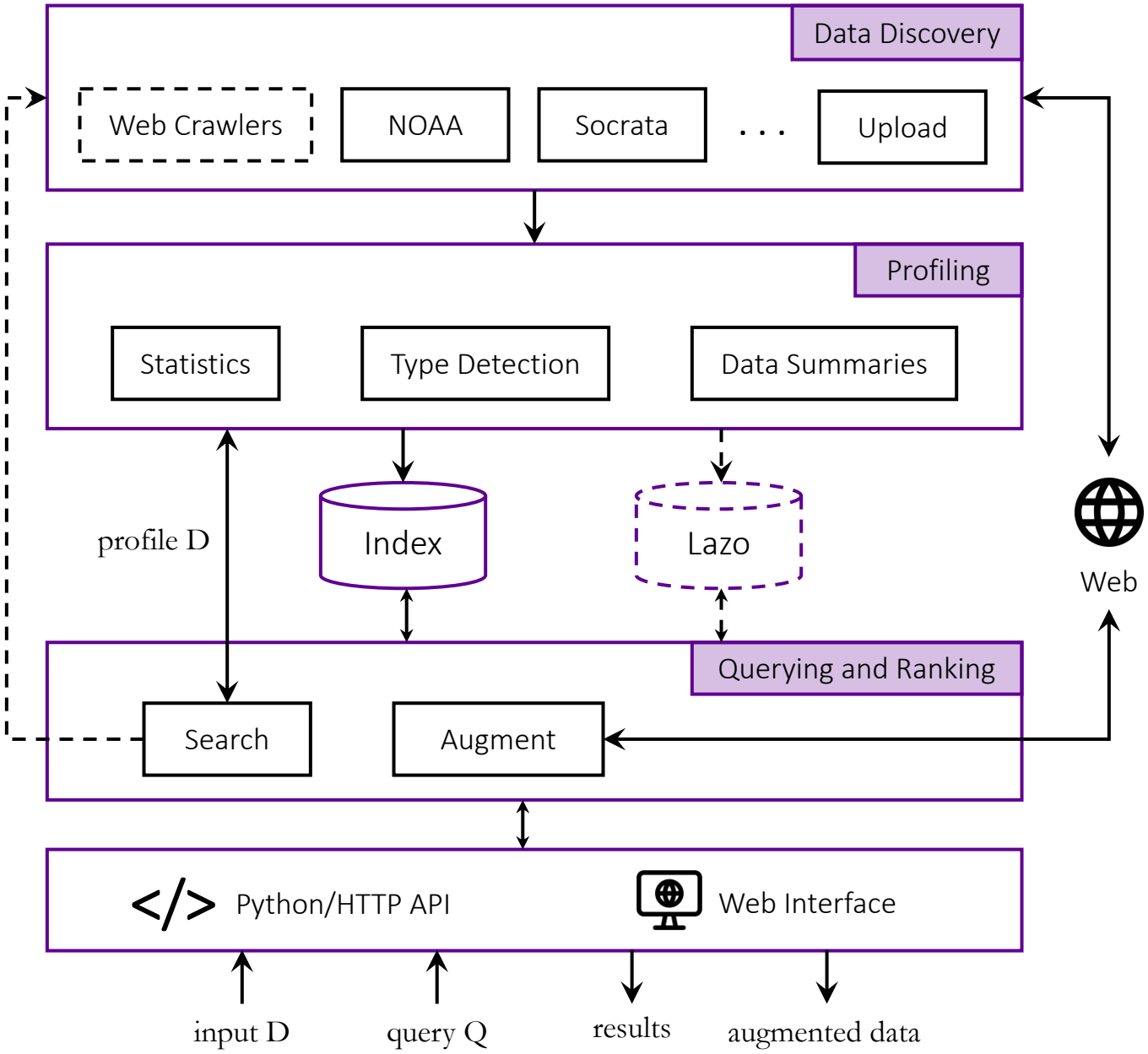
Auctus

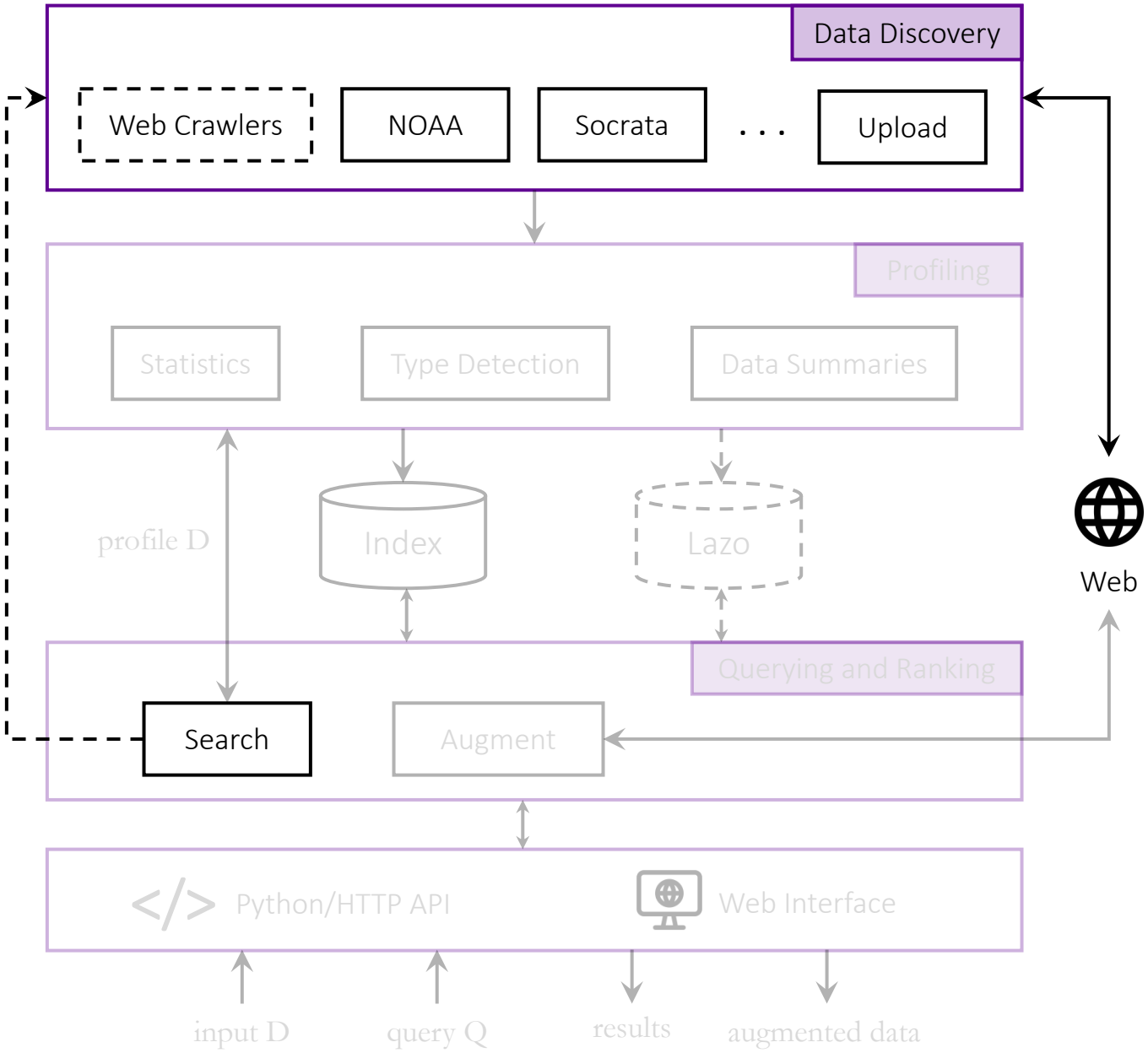
A dataset search engine tailored for data augmentation

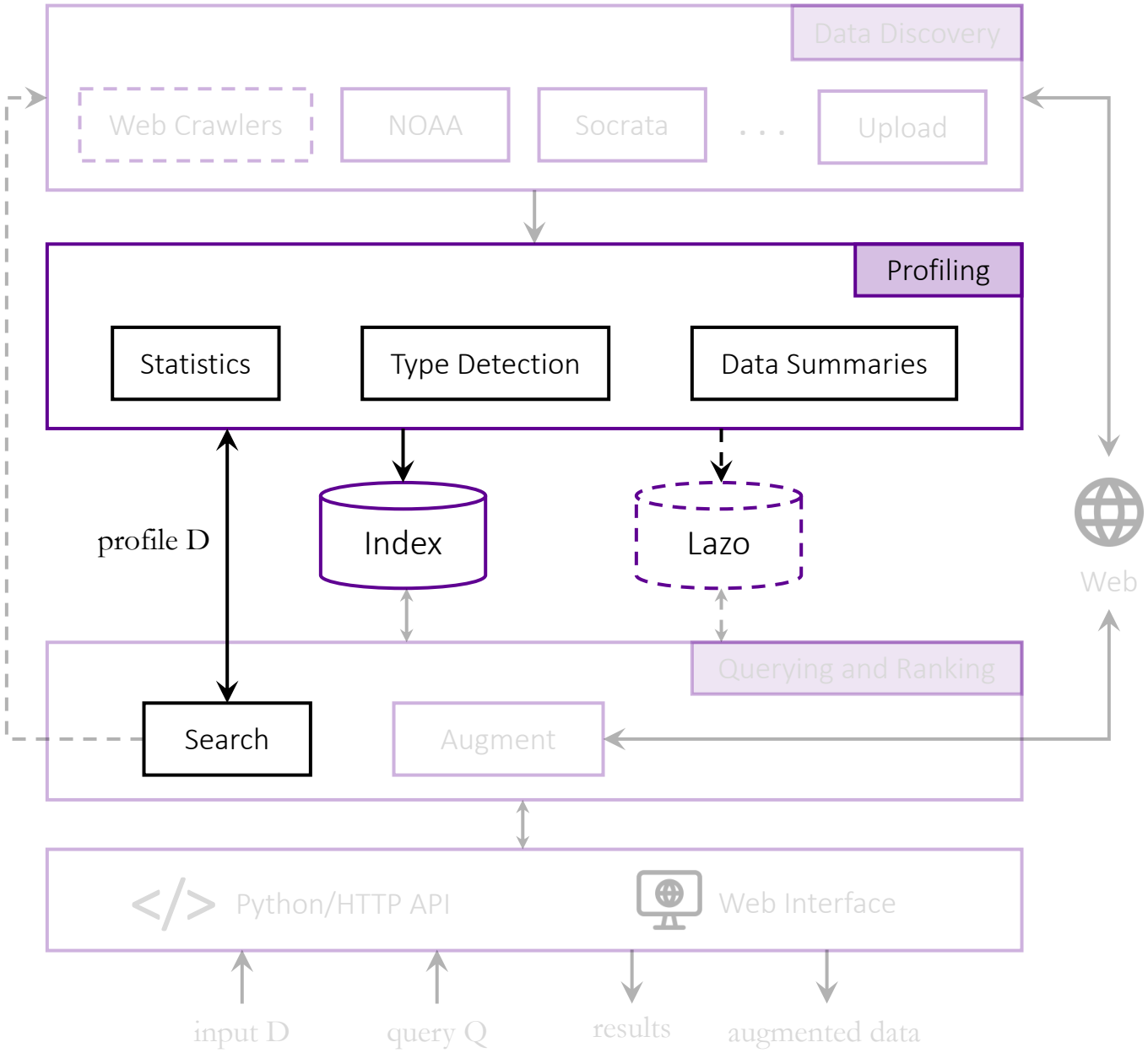
Given an input dataset D and an optional query Q, efficiently and effectively discover and rank a set of datasets from the Web that can be used to augment D

Types of augmentation

More features	↔	Joinable datasets
More records	↔	“Unionable” datasets


Data Discovery
Web Crawlers
NOAA
Socrata
. . .
Upload
Profiling
Statistics
Type Detection
Data Summaries
profile D
Index
Lazo
Web
Querying and Ranking
Search
Augment
</> Python/HTTP API
Web Interface
input D
query Q
results
augmented data


Data Discovery
Web Crawlers
NOAA
Socrata
. . .
Upload
Profiling
Statistics
Type Detection
Data Summaries
profile D
Index
Lazo
Web
Querying and Ranking
Search
Augment
</> Python/HTTP API
Web Interface
input D
query Q
results
augmented data



2018 Green Taxi Trip Data

Transportation

View Data | Visualize | Export | ...

The yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached...

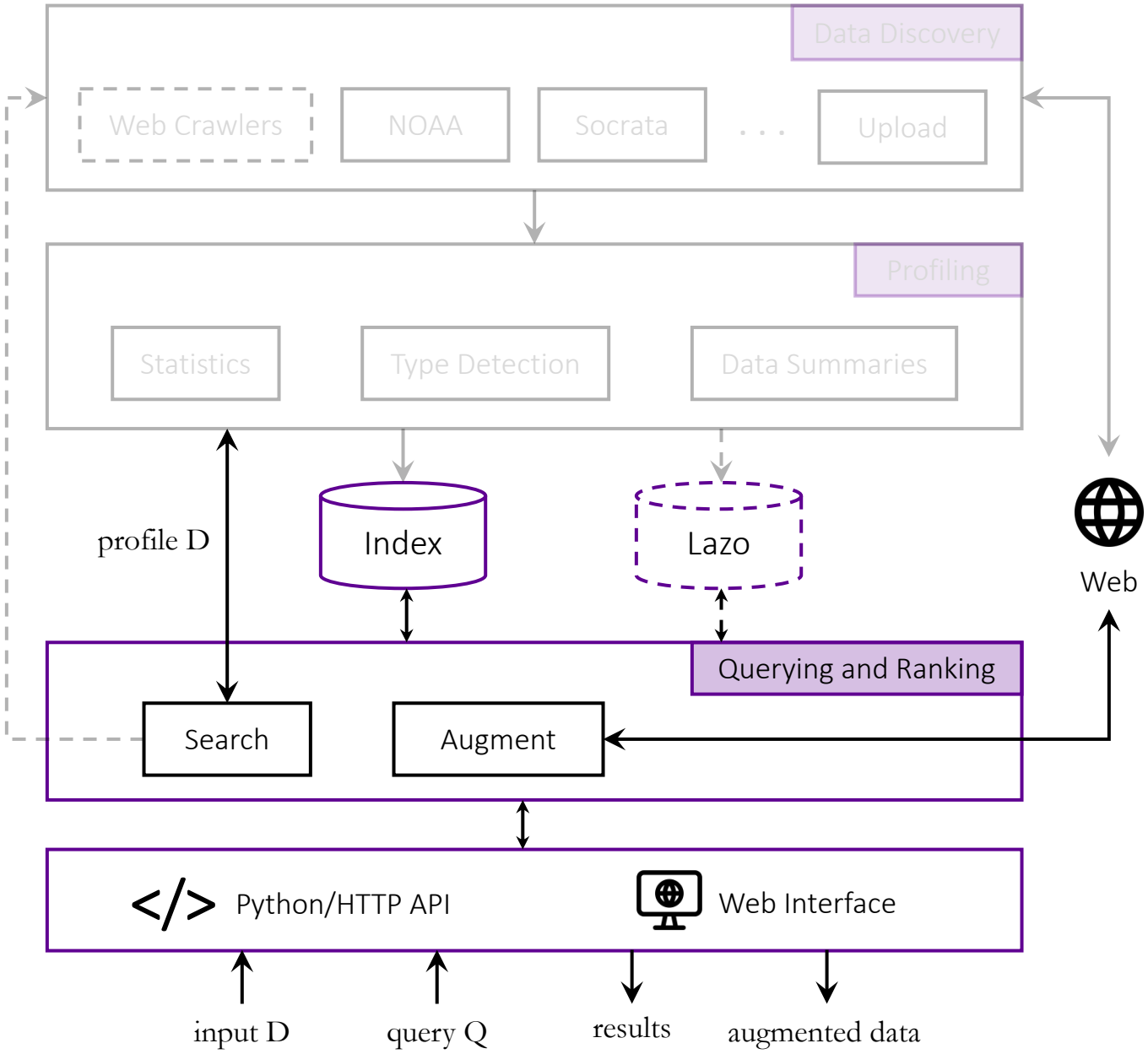
More

Updated April 2, 2019

Data Provided by Taxi and Limousine Commission (TLC)

https://data.cityofnewyork.us/Transportation/2018-Green-Taxi-Trip-Data/w7fs-fd9i

Lazo: A Cardinality-Based Method for Coupled Estimation of Jaccard Similarity and Containment. Raul Castro Fernandez, Jisoo Min, Demitri Devada, Samuel Madden. **ICDE'19**



Join and Union Search

Discussion and Challenges

Auctus is under active development

Deployed to improve ML performance, but useful in many scenarios

Open research questions

- Data noise
- Semantic matching [1]
- Dataset ranking [2]

[1] *InfoGather+: semantic matching and annotation of numeric and time-varying attributes in web tables. Meihui Zhang and Kaushik Chakrabarti. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (SIGMOD '13)*

[2] *Are key-foreign key joins safe to avoid when learning high-capacity classifiers?. Vraj Shah, Arun Kumar, and Xiaojin Zhu. Proc. VLDB Endow. 11, 3 (November 2017), 366-379.*

Thanks! Questions?

Web Interface: https://auctus.vida-nyu.org

Demo Video: http://bit.ly/auctus-video

Source Code: https://gitlab.com/ViDA-NYU/datamart/datamart

This work was funded by
the Defense Advanced Research Projects Agency (DARPA)

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER