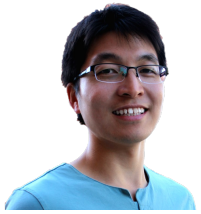
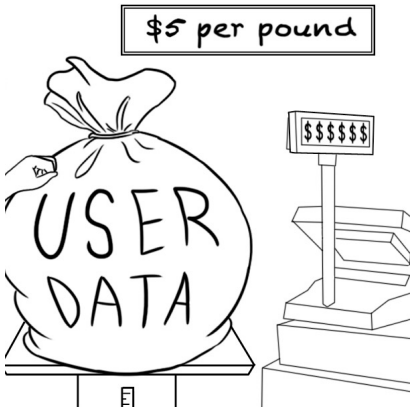
AAAI 2023 Tutorial: Economics of Data and ML



Haifeng Xu (Chicago)

Shuran Zheng (CMU)

James Zou (Stanford)



2/8/2023

Tutorial outline: economics of data and ML

Part 1: Data buyer's perspective.

- What data is the most useful? Statistical data valuation
- How to quantify the value of information.

Short break

Part II: Data seller's perspective.

- How to price information.
- How to collect truthful data.

Short break

Part III: economics of ML

- Market for ML-as-a-service

Machine learning as a service

Competing companies offer API access to ML models for fee.

Input query

ML vendors for text extraction

Google $15

Microsoft $10

EVERY PIXEL $6

Multi-label ML API

ML Prediction APIs: a data point → a label set (for a cost)

Google

[(person, 0.86), (car, 0.41), (table, 0.72)]

Google

[(75% off books, 0.46), (shops, 0.81), (factory, 0.92)]

Cost: $15/10K image

ML-as-service is a rapidly growing market

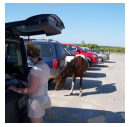
- Providers/Sellers:

- Buyers/Users:

Challenge: which ML APIs to use

Many APIs for the same tasks

Heterogeneity in those APIs' performance and cost

...     ...

Google

FACE+

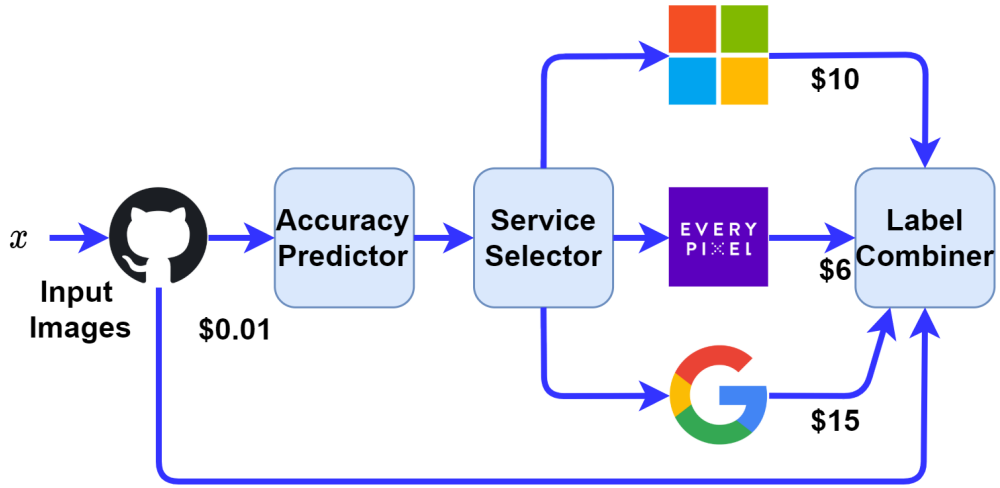
Microsoft

amazon

sightcorp

FrugalML

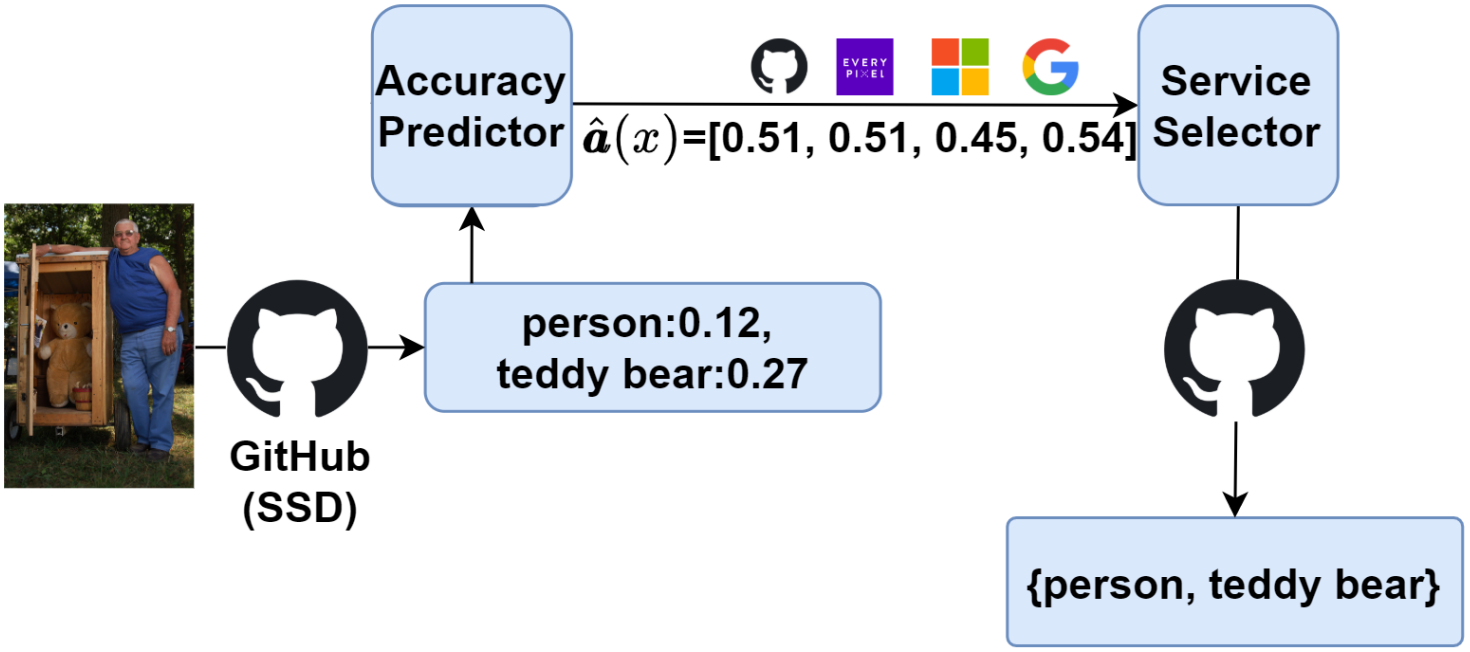
Idea: learn which API is the most cost-effective choice for each data.



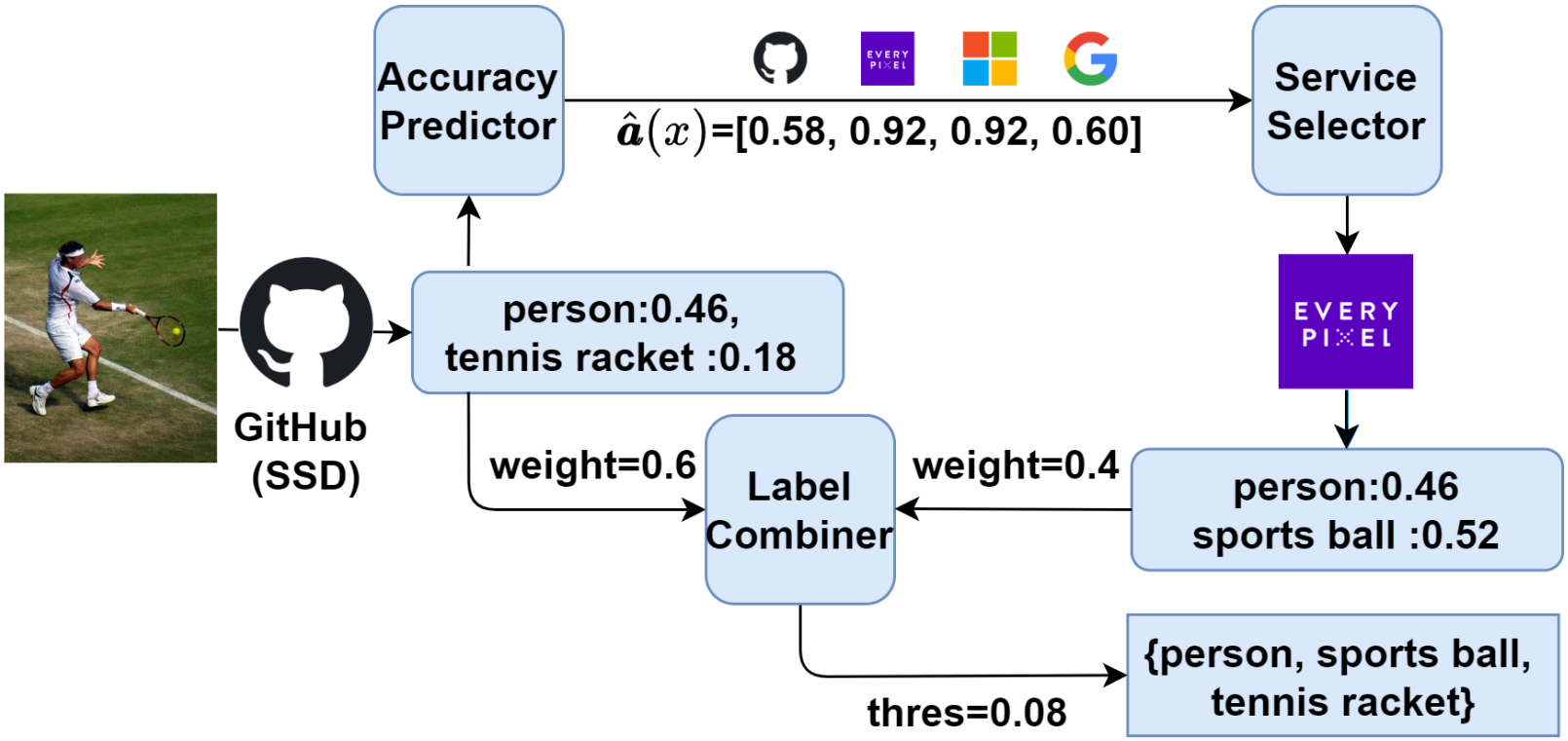
Benefit: up to 95% cost savings or 8% better accuracy with same cost across all tasks and datasets evaluated

Chen, Zaharia and Zou *ICML* 2022; *NeurIPS* 2020

FrugalML case study



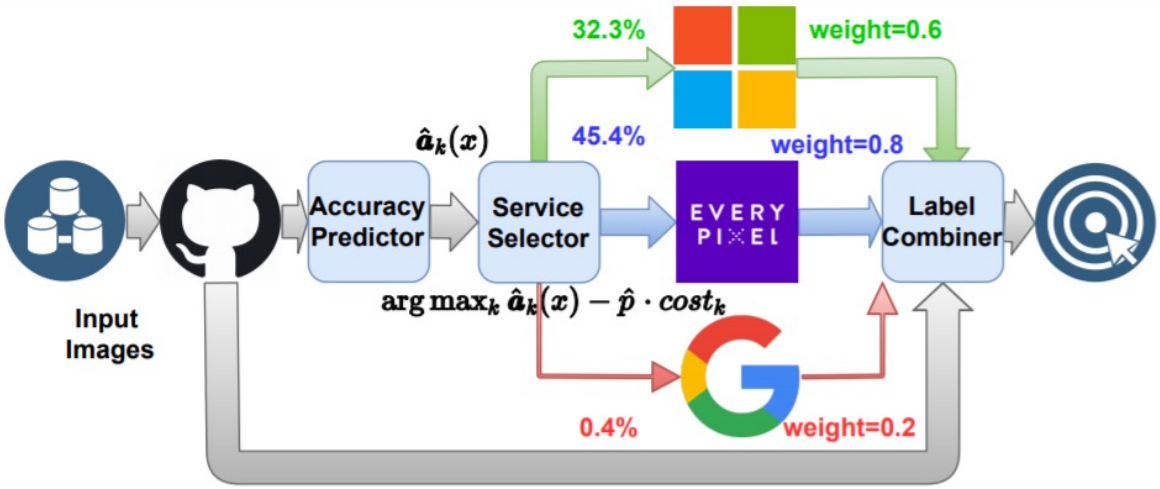
FrugalML case study



FrugalML case study

Case Study on a multi-label dataset, COCO

Budget: $5



Learned FrugalML Strategy

FrugalML: service selector

$$\max_{\boldsymbol{Z}\in\mathbb{R}^{N\times K}:} \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{Z}_{n,k}\hat{\boldsymbol{a}}_k(x_n)$$

$$s.t.\ \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1,k\neq base}^{K} \boldsymbol{Z}_{n,k}\boldsymbol{c}_k + \boldsymbol{c}_{base} \leq b$$

$$\sum_{k=1}^{K} \boldsymbol{Z}_{n,k} = 1, \boldsymbol{Z}_{n,k} \in \{0,1\}, \forall n,k$$

data point

estimated accuracy

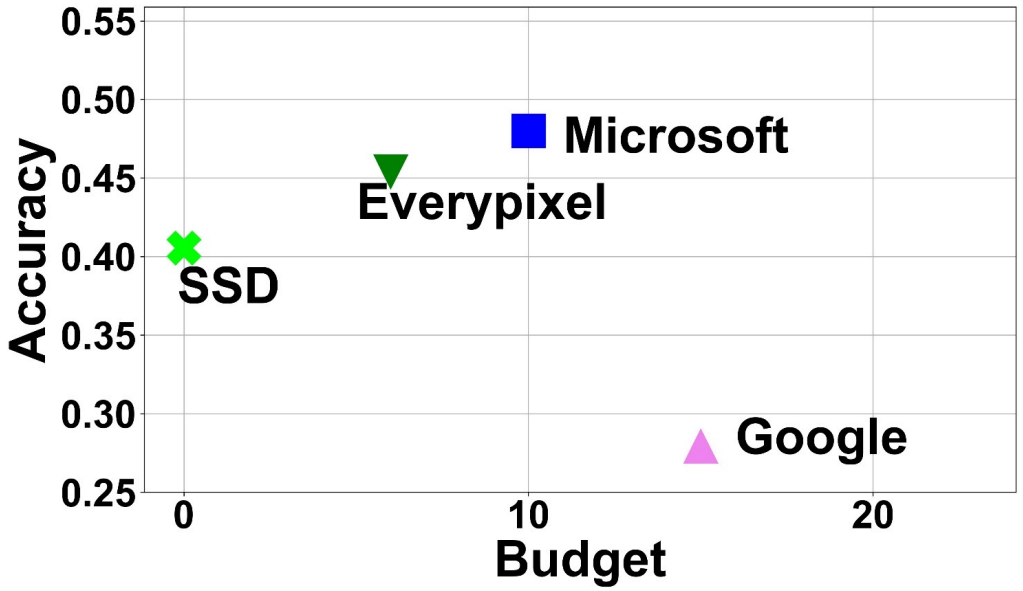
budget

cost

service selector $s^*(x_n) \triangleq \arg\max_k \boldsymbol{Z}^*_{n,k}$

- An integer linear programming problem
- Efficient solver: relaxation + rounding

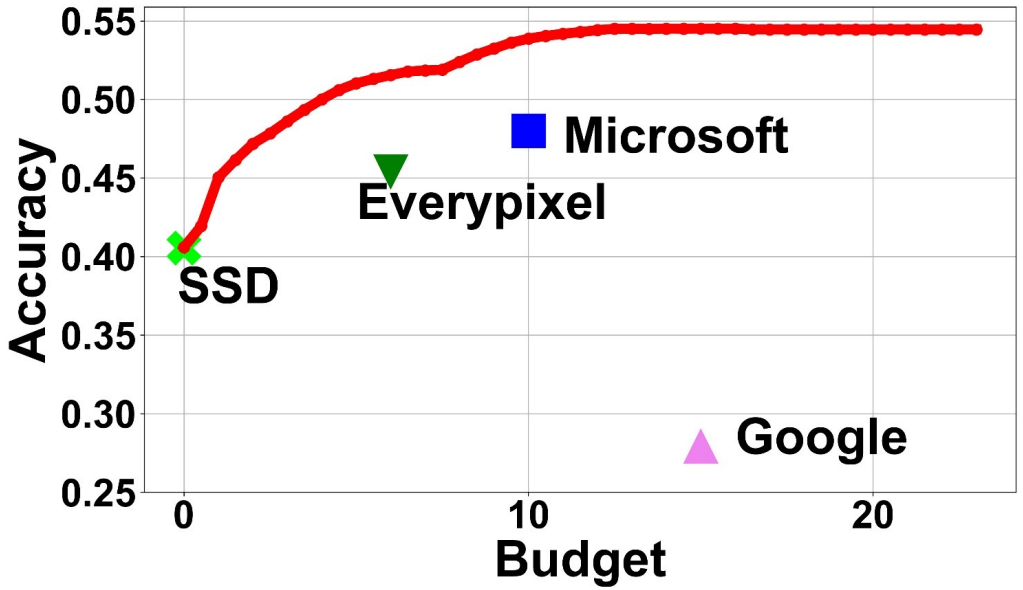
FrugalML saves cost and improves performance



Different APIs' Performance

Chen, Zaharia and Zou *ICML* 2022; *NeurIPS* 2020

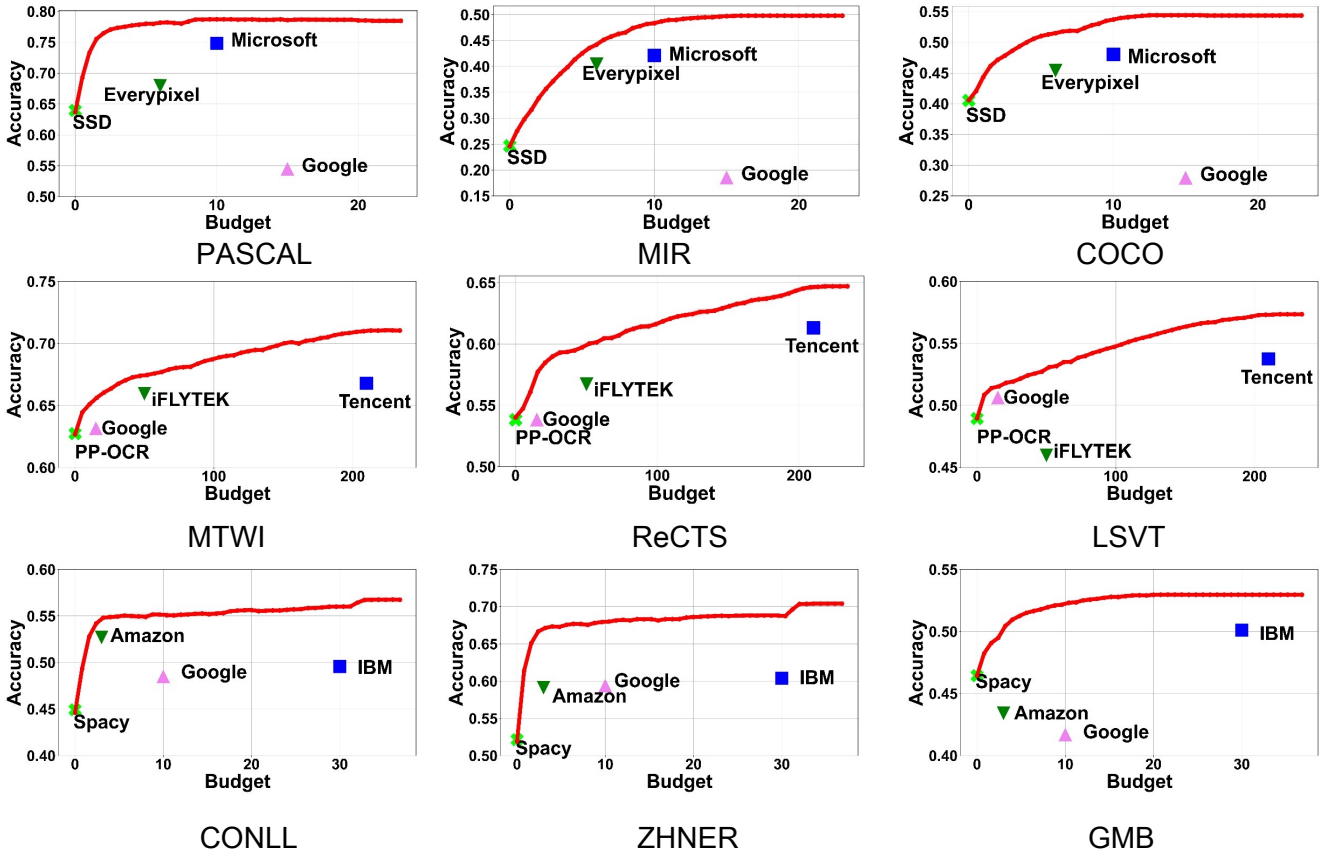
FrugalML saves cost and improves performance



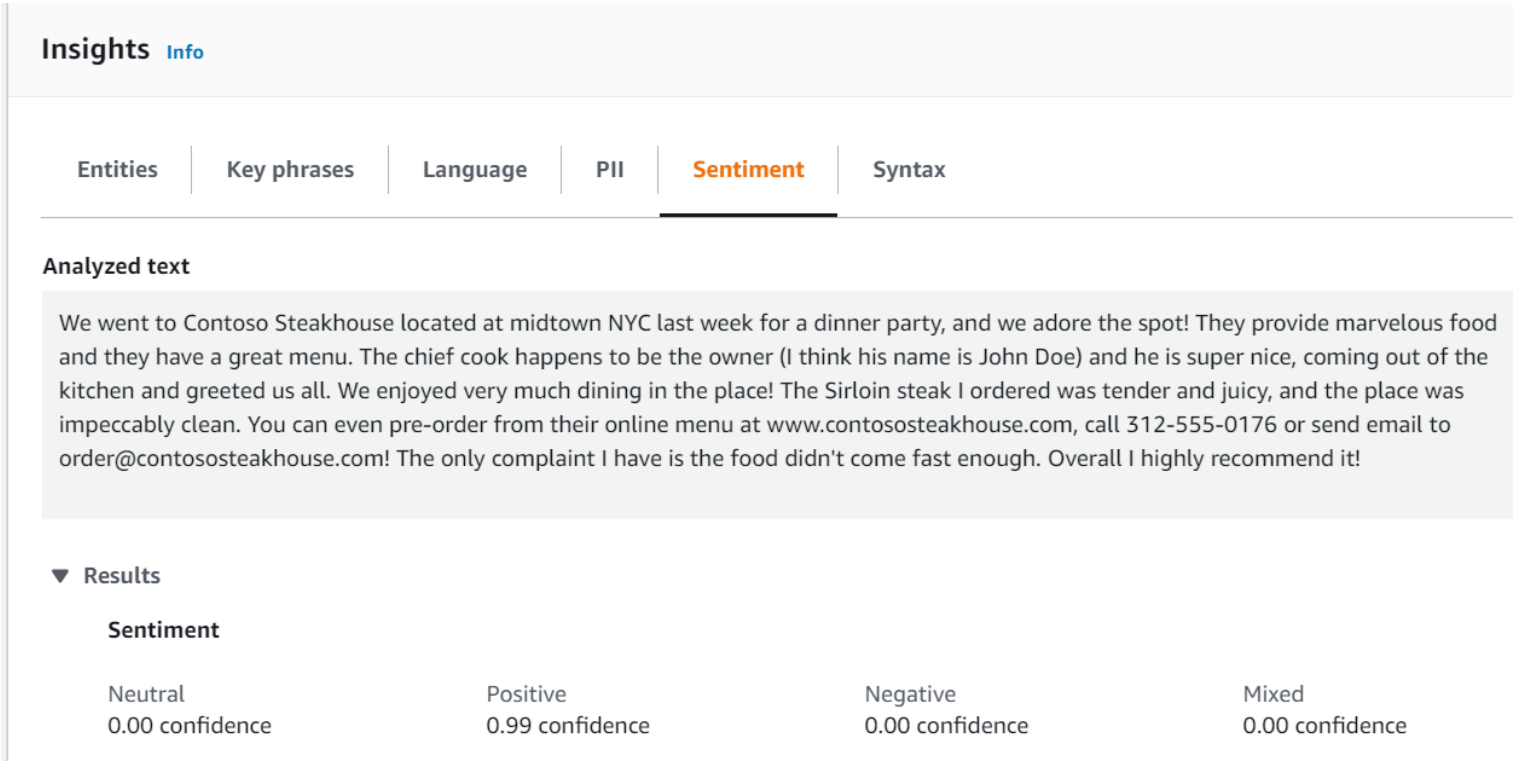
FrugalML's performance (red line)

Chen, Zaharia and Zou *ICML* 2022; *NeurIPS* 2020

FrugalML saves cost and improves performance

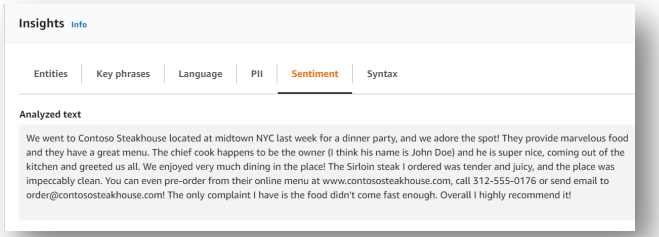


Sentiment Analysis by Amazon API

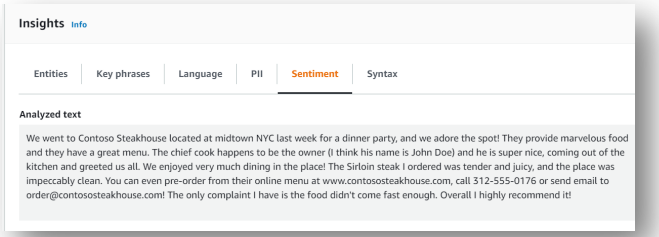


ML API's predictions often shift over time

2020



2021



predictions

"positive"

"negative"

Model shift: changes (often silent) in AI model's behavior and prediction when applied on the same data.

Chen, Zaharia and Zou *ICLR* 2022.

Commercial ML API shifts 2020-2021

Sentiment Analysis

ML API	YELP	IMDB	WAIMAI	SHOP
Amazon	+2.0	-1.1	+2.5	+1.5
Google	0.0	0.0	0.0	0.0
Baidu	0.0	0.0	0.0	0.0

Dataset

Speech Recognition

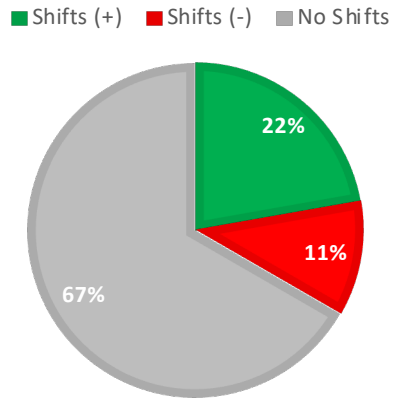
ML API	DIGIT	AMNIST	CMD	FLUENT
IBM	+3.3	-7.1	0.0	0.0
Google	+24.1	+7.5	+5.0	+0.6
MS	-1.6	+0.2	0.0	+0.3

Dataset

Facial Emotion

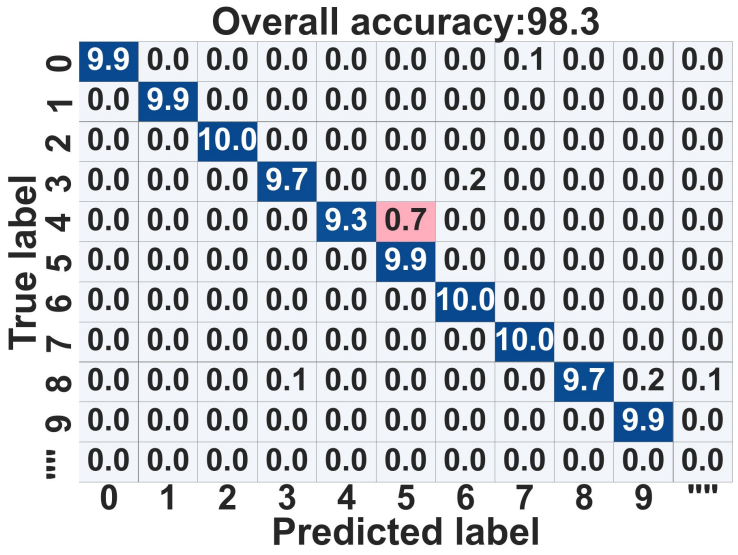
ML API	FER+	RAFDB	EXPW	AFNET
MS	+3.0	0.0	0.0	0.0
Google	0.0	0.0	-1.2	0.0
Face++	0.0	+0.7	0.0	0.0

Dataset

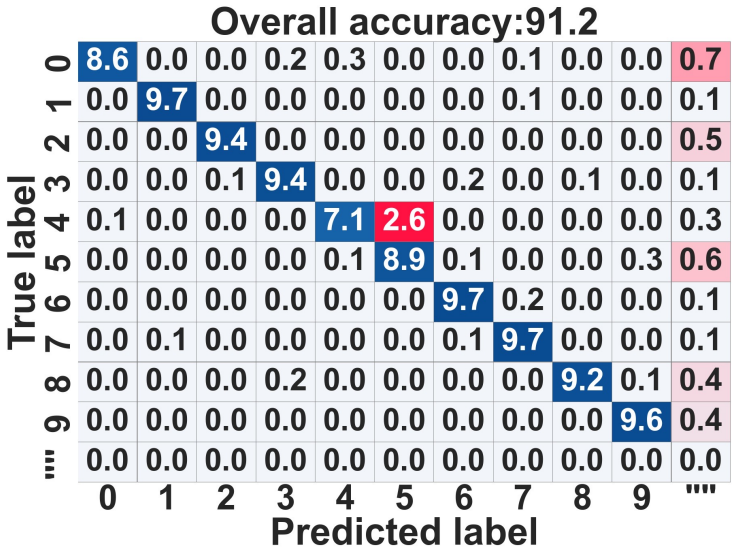


Chen, Zaharia and Zou *ICLR* 2022.

Case Study: IBM on AMNIST



2020



2021

IBM voice recognition's performance decreased on AMNIST over time.

Chen, Zaharia and Zou *ICLR* 2022.

Takeaways: ML-as-service market

- ML vendors have heterogeneous quality and cost (changes over time!)
- Opportunity to adaptively triage data to different vendors.
- Saves 95% of cost while doing better than any one vendor.
- Open challenge: FrugalML for large language models.

References

FrugalML: Chen, Zaharia and Zou. *NeurIPS* 2020.

FrugalML for more complex models: Chen, Zaharia and Zou. *ICML* 2020.

ML API shifts over time: Chen, Zaharia and Zou. *ICLR* 2022; *NeurIPS* 2022

jamesz@stanford.edu

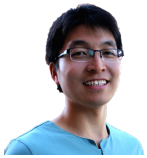
Discussion

Statistical data valuation identifies informative data points.

Economic models for the value of information.

Pricing based on log scoring rule prevents data manipulation. Easy to compute for many Bayesian ML models.

ML vendors have heterogeneous quality + cost; FrugalML gains by triaging different data to appropriate vendors.



Haifeng Xu (Chicago)

Shuran Zheng (CMU)

James Zou (Stanford)