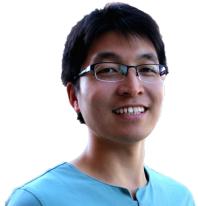


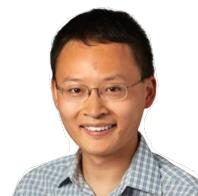
AAAI 2023 Tutorial: Economics of Data and ML



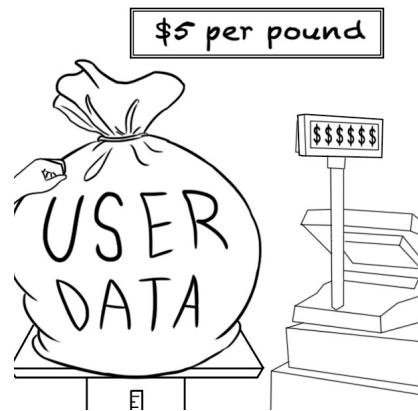
Haifeng Xu (Chicago)



Shuran Zheng (CMU)



James Zou (Stanford)



2/8/2023

Tutorial outline: economics of data and ML

Part I: 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

Microsoft

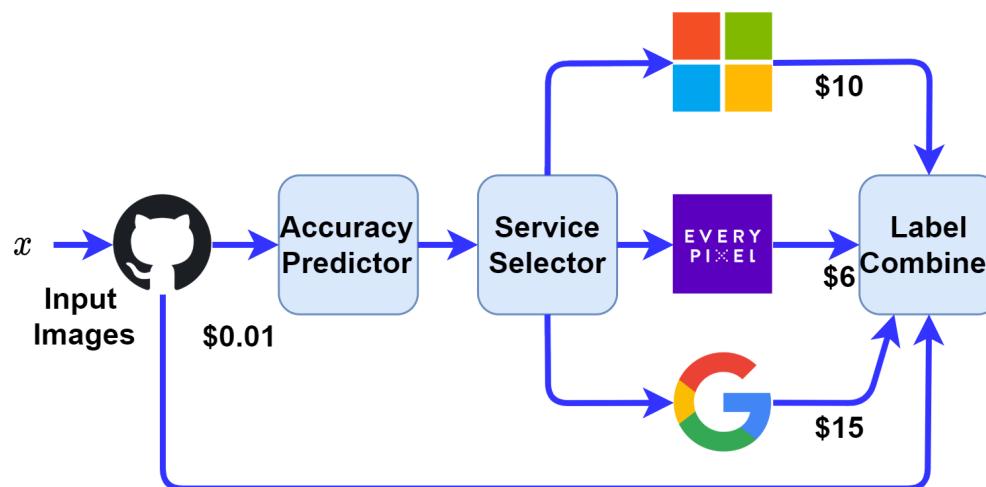


amazon

FACE+
sightcorp

FrugalML

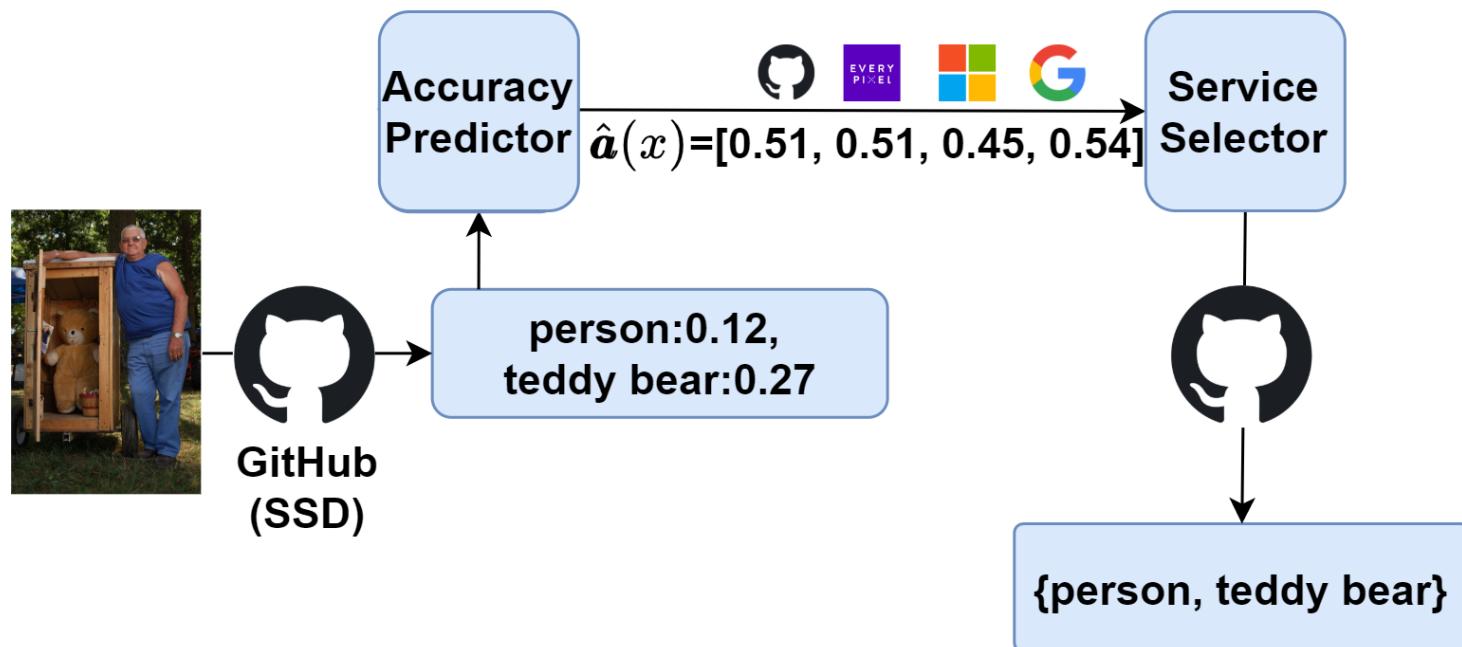
Idea: learn which API is the most cost-effective choice for each data.



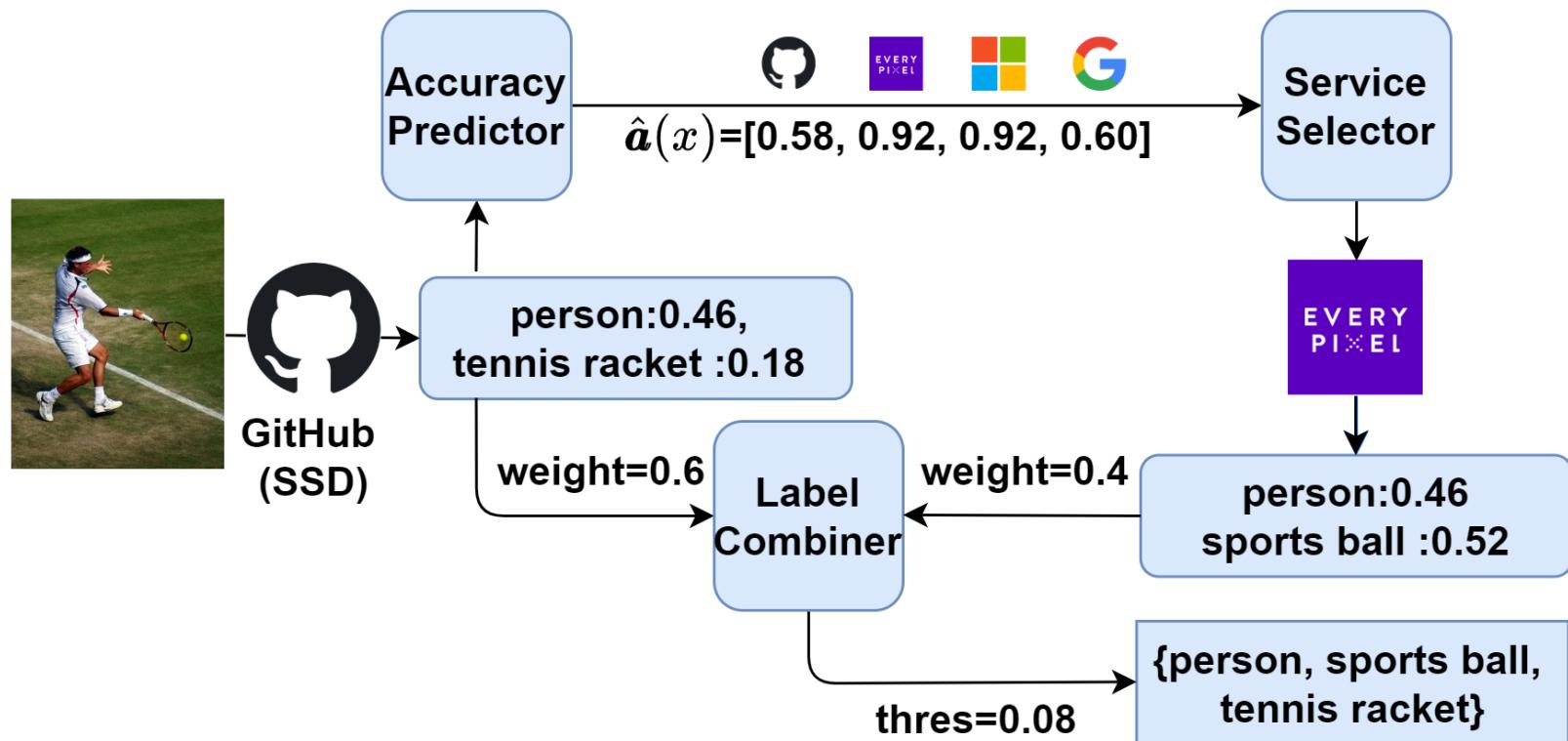
Benefit: up to 95% cost savings or 8% better accuracy with same cost across all tasks and datasets evaluated

Chen, Zaharia and Zou ICML 2022; NeurIPS 2020

FrugalML case study



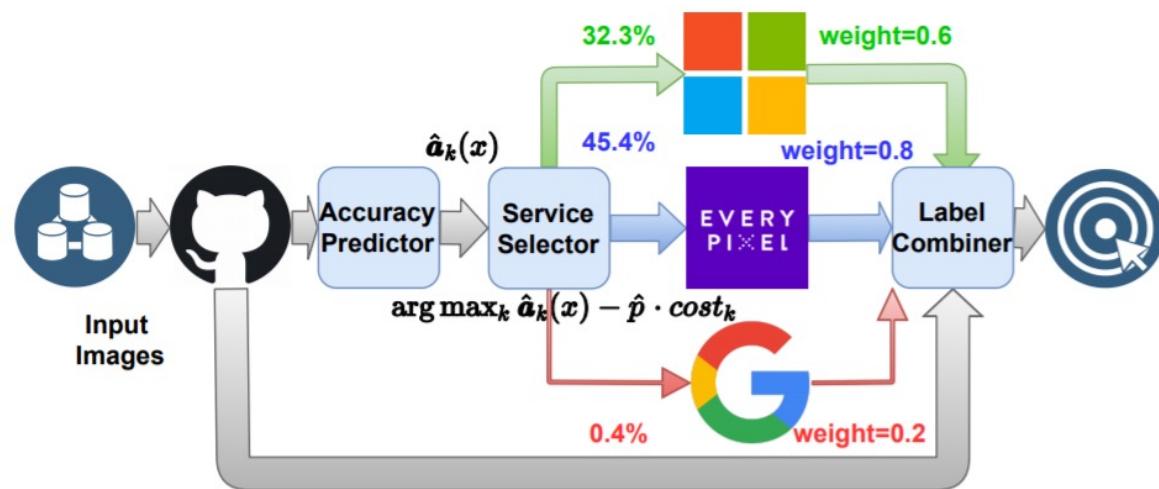
FrugalML case study



FrugalML case study

Case Study on a multi-label dataset, COCO

Budget: \$5



Learned FrugalML Strategy

FrugalML: service selector

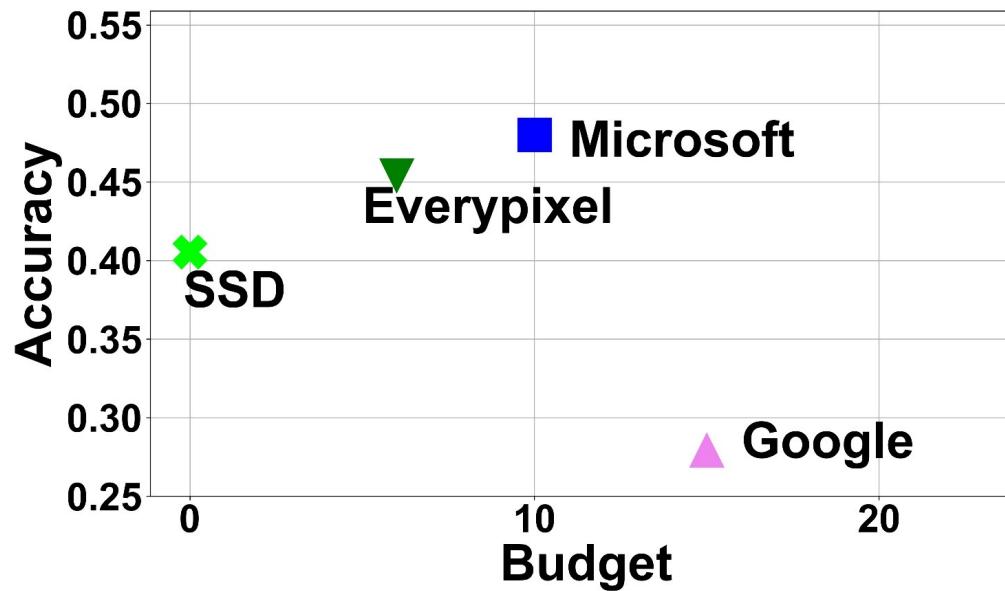
$$\begin{aligned} & \max_{\mathbf{Z} \in \mathbb{R}^{N \times K}} \frac{1}{N} \sum_{n=1}^N \mathbf{Z}_{n,k} \hat{\mathbf{a}}_k(x_n) \\ \text{s.t. } & \frac{1}{N} \sum_{n=1}^N \sum_{\substack{k=1, k \neq \text{base}}}^K \mathbf{Z}_{n,k} \mathbf{c}_k + \mathbf{c}_{\text{base}} \leq b \\ & \sum_{k=1}^K \mathbf{Z}_{n,k} = 1, \mathbf{Z}_{n,k} \in \{0, 1\}, \forall n, k \end{aligned}$$

data point
estimated accuracy
budget
cost

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

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

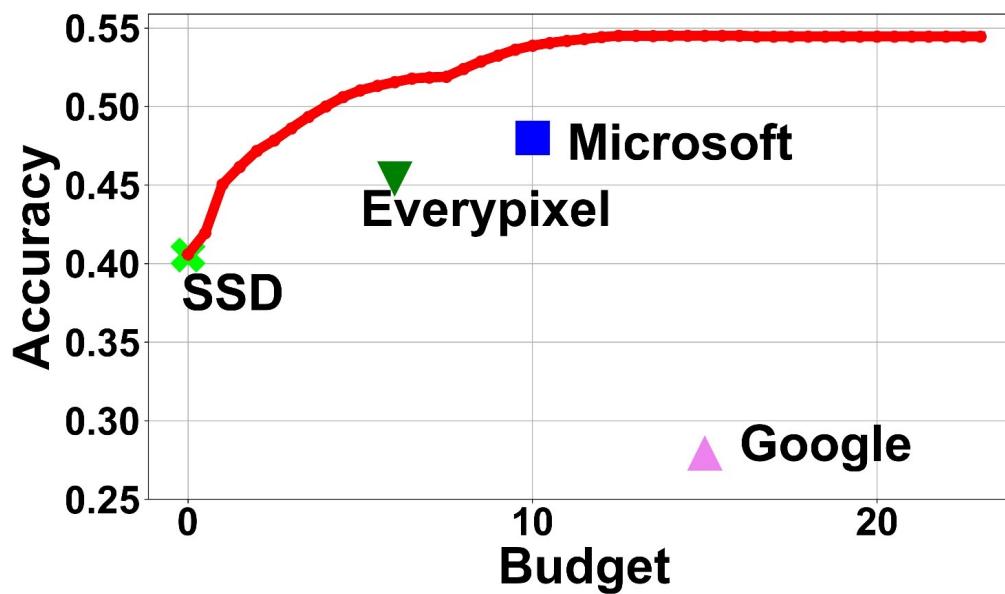
FrugalML saves cost and improves performance



Different APIs' Performance

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

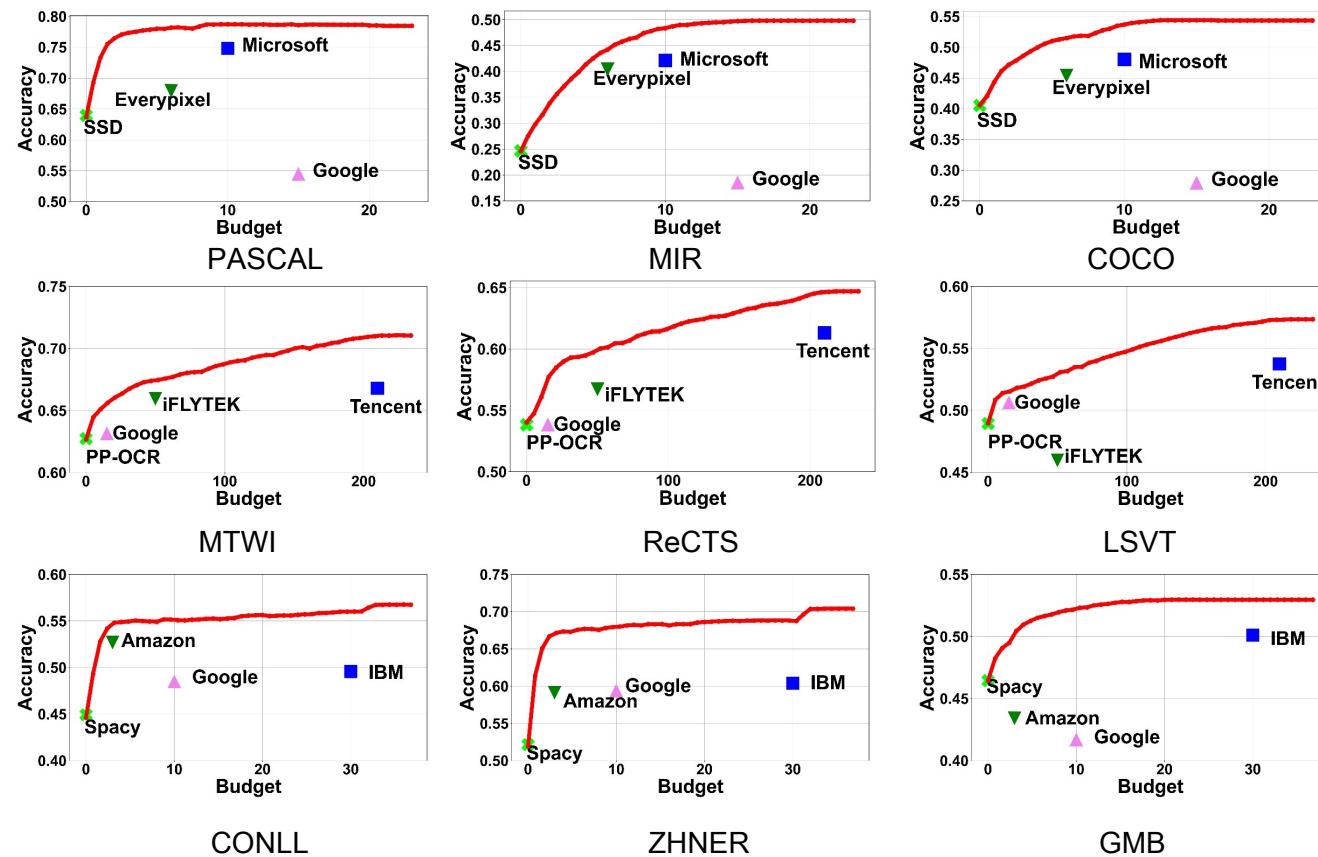
FrugalML saves cost and improves performance



FrugalML's performance (red line)

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

FrugalML saves cost and improves performance



Sentiment Analysis by Amazon API

Insights Info

Entities | Key phrases | Language | PII | **Sentiment** | Syntax

Analyzed text

We went to Contoso Steakhouse located at midtown NYC last week for a dinner party, and we adore the spot! They provide marvelous food and they have a great menu. The chief cook happens to be the owner (I think his name is John Doe) and he is super nice, coming out of the kitchen and greeted us all. We enjoyed very much dining in the place! The Sirloin steak I ordered was tender and juicy, and the place was impeccably clean. You can even pre-order from their online menu at www.contososteakhouse.com, call 312-555-0176 or send email to order@contososteakhouse.com! The only complaint I have is the food didn't come fast enough. Overall I highly recommend it!

▼ Results

Sentiment

Neutral 0.00 confidence	Positive 0.99 confidence	Negative 0.00 confidence	Mixed 0.00 confidence
----------------------------	-----------------------------	-----------------------------	--------------------------

ML API's predictions often shift over time

2020

Insights Info

Entities | Key phrases | Language | PII | **Sentiment** | Syntax

Analyzed text

We went to Contoso Steakhouse located at midtown NYC last week for a dinner party, and we adore the spot! They provide marvelous food and they have a great menu. The chief cook happens to be the owner (I think his name is John Doe) and he is super nice, coming out of the kitchen and greeted us all. We enjoyed very much dining in the place! The Sirloin steak I ordered was tender and juicy, and the place was impeccably clean. You can even pre-order from their online menu at www.contososteakhouse.com, call 312-555-0176 or send email to order@contososteakhouse.com! The only complaint I have is the food didn't come fast enough. Overall I highly recommend it!



predictions

“positive”

2021

Insights Info

Entities | Key phrases | Language | PII | **Sentiment** | Syntax

Analyzed text

We went to Contoso Steakhouse located at midtown NYC last week for a dinner party, and we adore the spot! They provide marvelous food and they have a great menu. The chief cook happens to be the owner (I think his name is John Doe) and he is super nice, coming out of the kitchen and greeted us all. We enjoyed very much dining in the place! The Sirloin steak I ordered was tender and juicy, and the place was impeccably clean. You can even pre-order from their online menu at www.contososteakhouse.com, call 312-555-0176 or send email to order@contososteakhouse.com! The only complaint I have is the food didn't come fast enough. Overall I highly recommend it!



“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.

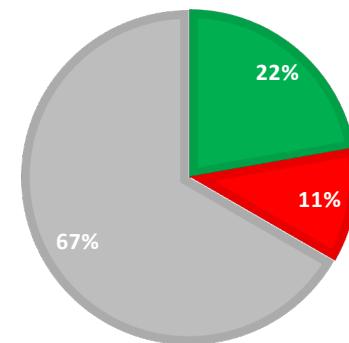
Commercial ML API shifts 2020-2021

		Sentiment Analysis			
		YELP	IMDB	WAIMAI	SHOP
ML API	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	0.0

		Speech Recognition			
		DIGIT	AMNIST	CMD	FLUENT
ML API	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	

		Facial Emotion			
		FER+	RAFDB	EXPW	AFNET
ML API	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	0.0

■ Shifts (+) ■ Shifts (-) ■ No Shifts



Chen, Zaharia and Zou /ICLR 2022.

Case Study: IBM on AMNIST

Overall accuracy:98.3											"\"
True label	0	1	2	3	4	5	6	7	8	9	"\"
0	9.9	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0
1	0.0	9.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	9.7	0.0	0.0	0.2	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	9.3	0.7	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	9.9	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	9.7	0.2	0.1
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	9.9	0.0
""	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

2020

Overall accuracy:91.2											"\"
True label	0	1	2	3	4	5	6	7	8	9	"\"
0	8.6	0.0	0.0	0.2	0.3	0.0	0.0	0.1	0.0	0.0	0.7
1	0.0	9.7	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1
2	0.0	0.0	9.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5
3	0.0	0.0	0.1	9.4	0.0	0.0	0.2	0.0	0.1	0.0	0.1
4	0.1	0.0	0.0	0.0	7.1	2.6	0.0	0.0	0.0	0.0	0.3
5	0.0	0.0	0.0	0.0	0.1	8.9	0.1	0.0	0.0	0.3	0.6
6	0.0	0.0	0.0	0.0	0.0	0.0	9.7	0.2	0.0	0.0	0.1
7	0.0	0.1	0.0	0.0	0.0	0.0	0.1	9.7	0.0	0.0	0.1
8	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	9.2	0.1	0.4
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	9.6	0.4
""	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

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

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)