中国人民大学
RENMIN UNIVERSITY OF CHINA

From Big Data to Good Data

Data Preparation for AI Systems

Ju FAN (范举)

http://iir.ruc.edu.cn/~fanj/

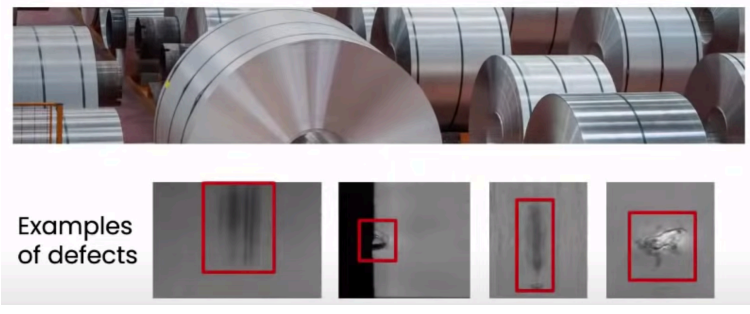
Renmin University of China

2022/06/27

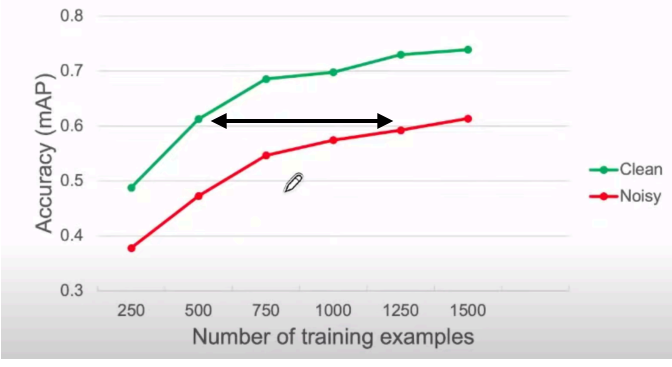

Keep Fixed
~99% of Research
AI System
(or Data Science)
=
Data
+
Code
(model/algorithm)

From Big Data to Good Data

- An interesting example from Andrew Ng (吴恩达)
  - Inspecting steel sheets for defects
  - The data is noisy, e.g., 12% mislabeled



	Steel Defect Detection	Solar Panel	Surface Inspection
Baseline	76.2%	75.68%	85.05%
Improving the Code	+0% (76.2%)	+0.04% (75.72%)	+0.00% (85.05%)
Improving the Data	+16.9% (93.1%)	+3.06% (78.74%)	+0.4% (85.45%)

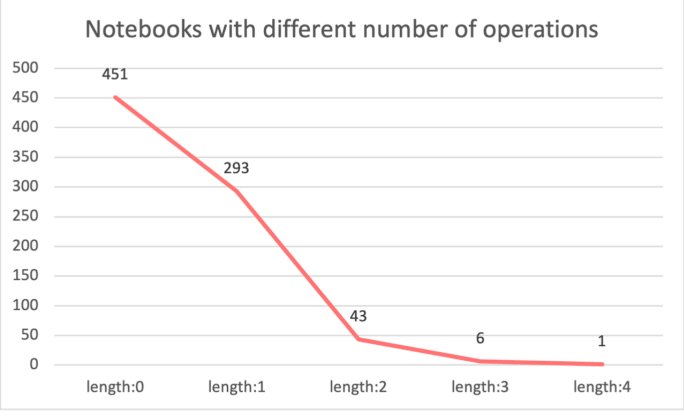


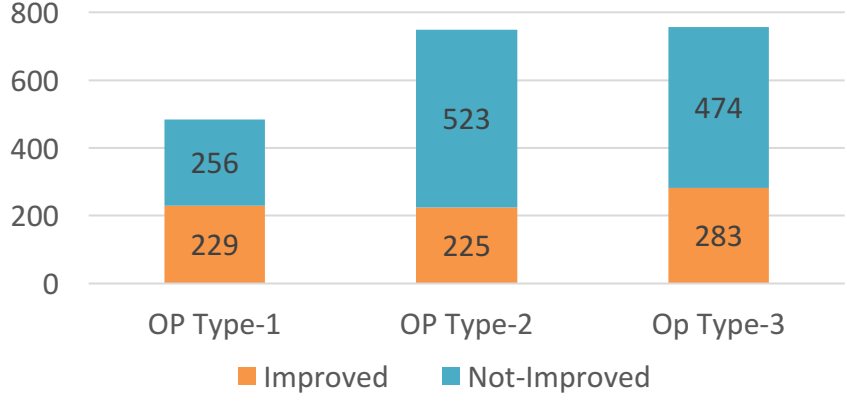
- The followings are about equal
  - 500 clean examples
  - 1250 noise examples

Andrew's talk:
https://www.youtube.com/watch?v=06-AZXmwHjo

From Big Data to Good Data

- We conduct a user study on 794 notebooks from Kaggle
- We examine whether they use various types of operations (OPs) for improving the data…
  - E.g., Missing value imputation, value standardization, etc.



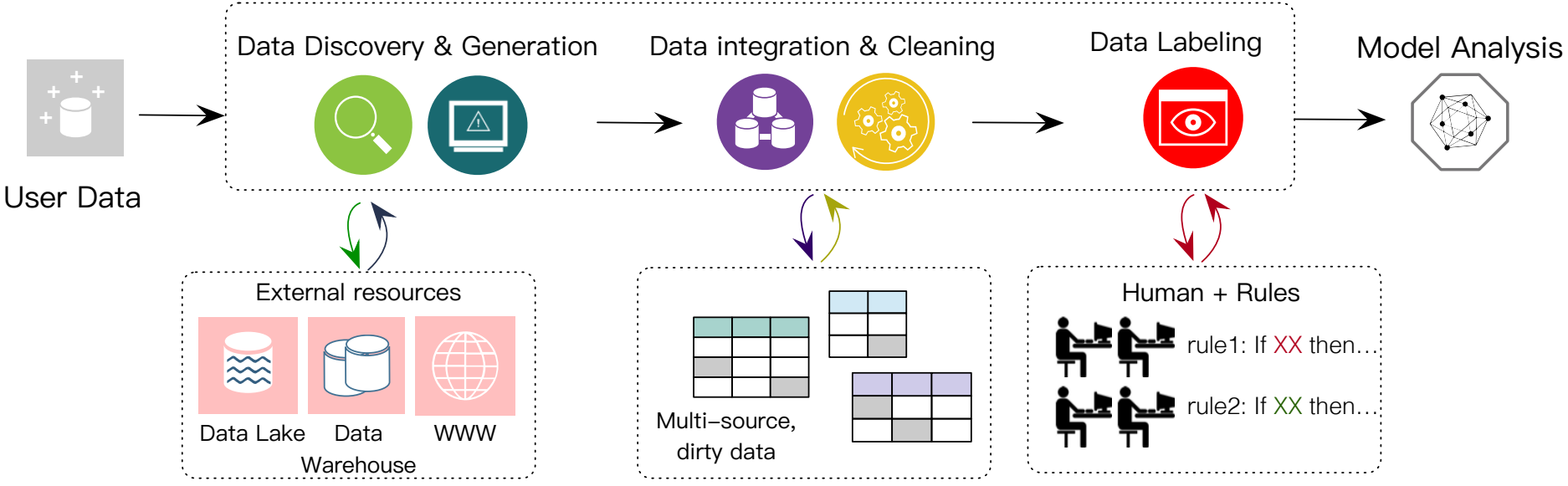


Most users lack comprehensive awareness of improving the data

When injecting OPs to improve their data, we improve 30% - 47% of the notebooks

Data Prep Meets AI Systems

- **Data Prep**aration, the process of turning big data into good data, is a crucial step of machine learning and AI systems.



Data Prep Meets AI Systems

2014

The New York Times

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

Yet far too much handcrafted work — what data scientists call "data wrangling," "data munging" and "data janitor work" — is still required. Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.

https://www.nytimes.com/2014/08/18/technology/for- big-data-scientists-hurdle-to-insights-is-janitor-work.html

2020

ANACONDA

The State of Data Science 2020

Moving from hype toward maturity

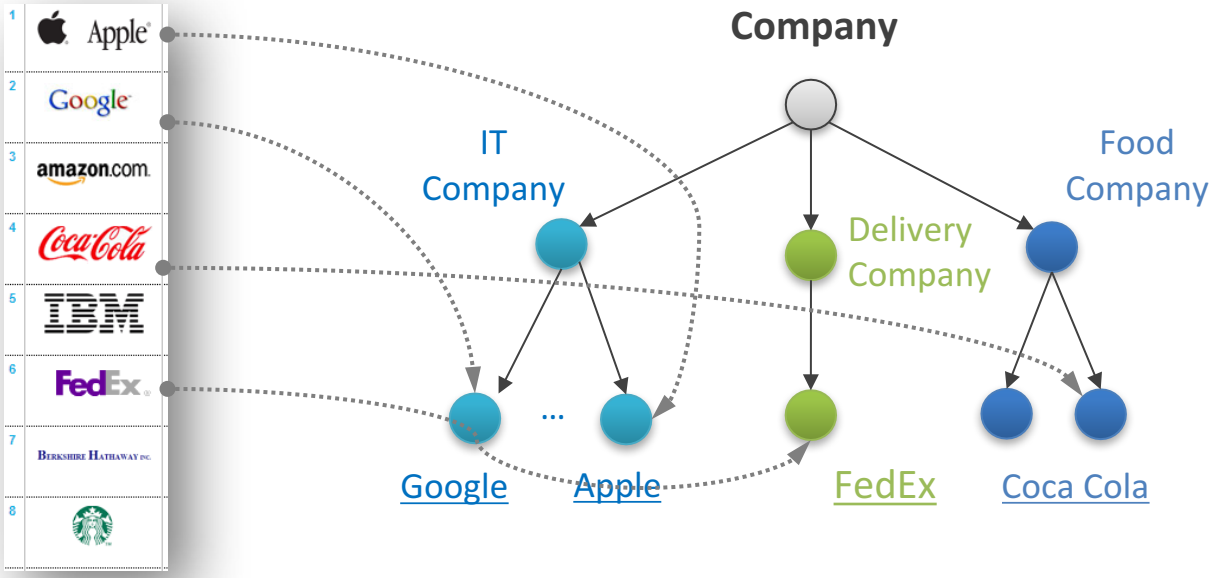
We were disappointed, if not surprised, to see that data wrangling still takes the lion's share of time in a typical data professional's day. Our respondents reported that almost half of their time is spent on the combined tasks of data loading and cleansing. Data

https://www.anaconda.com/state-of-data-science-2020

Data Prep is A Bottleneck in Practice

Large-Scale Data Labeling

- In many applications, it is indispensable to obtain large-scale data labels with high quality
  - E.g., Training an ML model for categorization



Large-Scale Data Labeling

- **Challenge**: Creating a high-quality training set can NOT be solely addressed by automated process

人类为人工智能“打工”！

Google Cloud & YouTube
Video Understanding

- 10 M labeled video

Visual QA

- 300K labeled images
- 1M aligned QA

从小作坊到

19-12-17 11:06:55 来源：企商网

百度：未来五年在山西培养5万名AI数据标注师

金融界 2020-07-02 15:07

7月2日消息，百度宣布将继续加大对新基建数据产业方面的投入，未来5年将在百度山西数据标注基地培养5万名AI数据标注师，并引入更多AI合作伙伴。百度与山西数据标注基地的合作模式，未来还将拓展到更多省市，提供更多的AI就业岗位，支持当地科技产业发展。

截止目前，百度（山西）人工智能基础数据产业基地AI数据标注师从业人员超过2000人，实现营业收入超亿元，企业入驻35家。此前6月6日，百度与山西省政府签约，将进一步深入合作打造山西综合改造示范区AI数据交易平台，促进数据资源的开放与共享，推动山西数字经济发展。

AI数据标注师是随着人工智能的发展出现的一个新兴就业岗位，主要工作是教会AI认识数据。有了足够多、足够好的数据，AI才能学会像人一样去感知、思考和决策，更好地为人类服务。例如，疫情期间山西数据标注中心完成的戴口罩的人脸图像标注，采集大量的戴口罩的人脸照片后，数据标注师对人脸的眉毛、眼镜、颧骨等人脸关键点进行精准的标注，标注的特征点越多，AI就越能精确地识别戴口罩场景下的人脸，让人们在不摘口罩的情况下也能实现精确的体温测量，或是通过人脸闸机。

Data Integration & Cleaning

新浪微博 weibo.com **Raw Data**

天边的一朵小马驹_ +关注

2月11日 00:04 来自 肺炎患者求助超话

肺炎患者求助超话 【姓名】 XYZ

【年龄】56

【所在城市】 武汉

【所在小区、社区】汉阳区麒麟路麒麟社区 X栋Y单元Z楼

【患病时间】1月26日

【联系方式】 1397xxx6027

【其他紧急联系电话】 1341xxx1935

【病情详细描述】病人有高血压，甲状腺癌，发烧十天，高烧39度，双肺感染，咳嗽，无力，2月8日做了试剂盒，结果还没出来，现在病人病危，靠氧气，血氧70。医院不收入院。

【周边家人及隔离情况】家里还有一个姐姐同住，姐姐有糖尿病重症，也昏睡了十几天，咳嗽，乏力，至今也没隔离，连去医院的劲都没有。家里还有一位父亲抗美援朝的离休老干部90岁，于2月5日在家离世（是不是这个病离世未知）

我已被隔离，妈妈目前情况非常严峻，只能躺床上靠吸氧根本不能动，连CT片都不能拍给我，我们报给社区说只能在家等，根本无力解决，现在情况紧急，生命危在旦夕，求求哪家医院能收治我的家人

More than 10,000 Microblogs
(29 Jan, 2020 – 17 Feb 2020)

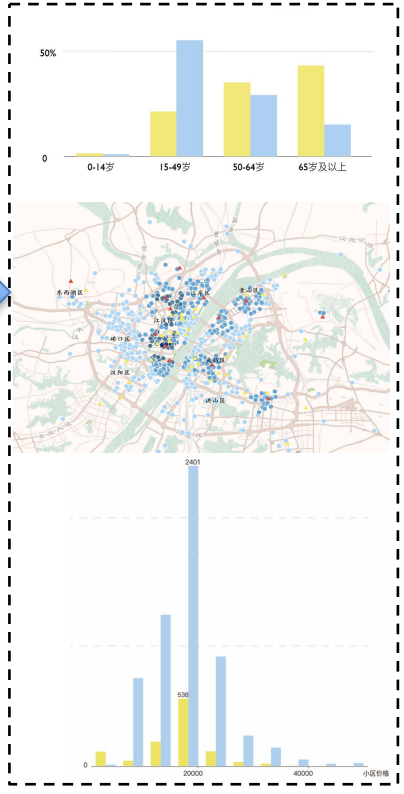
① Info Extraction

③ Clean

name	age	city	addr
ZBX	52	湖北武汉	长山社区万科嘉园小区
ZBX	52	湖北武汉	长山社区万科嘉园小区
CTC	75	武汉市洪山区	东湖御院 红星社区
DSX	51	武汉	云顶居，清水源社区
DSX		武汉	云顶居，清水源社区
DSX	51	武汉	云顶居，清水源社区
RC	36	武汉	洪山区
RC	36	武汉	洪山区
RC	36	武汉市	洪山区
RC	36	武汉市	洪山区
RC		湖北省武汉市	洪山区关山大道新竹路保利时代
RC	35	湖北省武汉市	保利时代
RC	36	武汉	洪山区

② Integrate

Analysis



External Resources

Geo-Location APIs

百度地图 高德地图

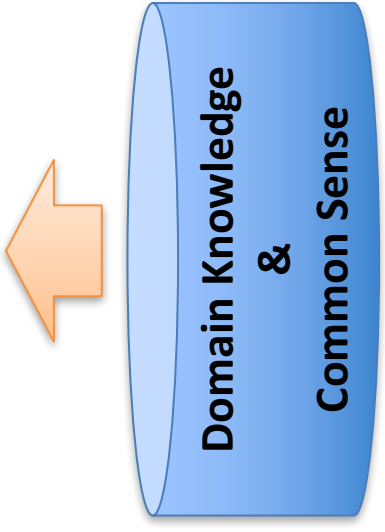
Apartment DB from 链家

小区名	区县	街道	年份	二手房价
洪山区 027社区	洪山	老南湖	2008	18869
黄陂区 08经典	黄陂	盘龙城	2010	12231
硚口区 2008城市花	硚口	长丰常码头	2007	13682
武昌区 2008新长江	武昌	积玉桥	2005	22585
青山区 20街坊	青山	青山	1987	17574
青山区 22街坊	青山	青山	1985	16752

Data Integration & Cleaning

- **Challenge**: Data Integration & Cleaning often rely on domain knowledge (e.g. the relationship between a city and its zip code for data cleaning) and common-sense reasoning, which are easy for human and hard for machines

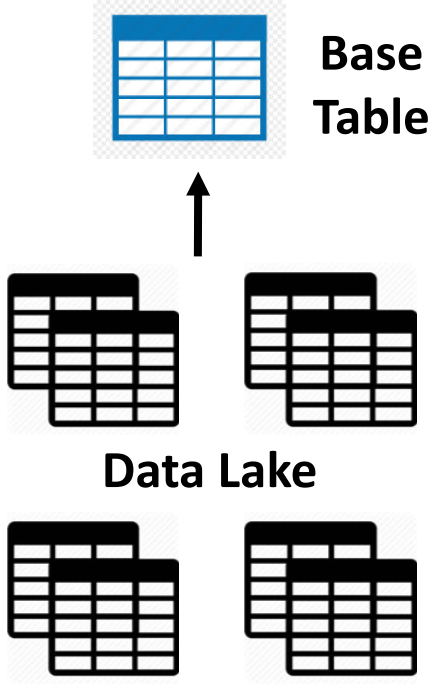
ID	Name	Expertise	City	Address
T1	Michael Jordan	Computer Science	Berkeley	9th Street
T2	Michael Jordan	Machine Learning	Berkeley ?	9 ST
T3	Michael Jordan ?	Basketball	New York City	3th Street



Data Discovery & Augmentation

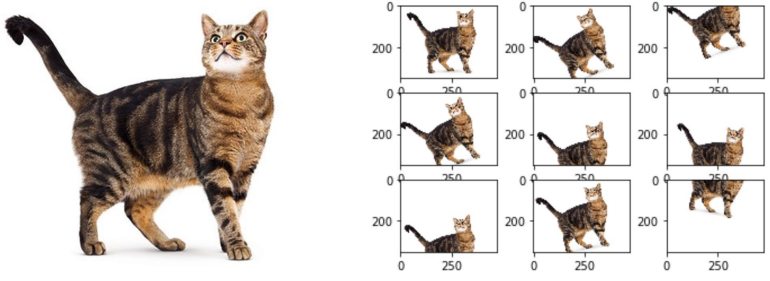
- When training ML models, it is essential to collect a large and representative dataset to sufficiently capture the variability observed in the real world

- **Data Discovery**
  - Given a base table, it finds the relevant tables from the data lake, to improve performance of model training
- **Challenges**
  - How to organize, index and rank datasets effectively and efficiently.

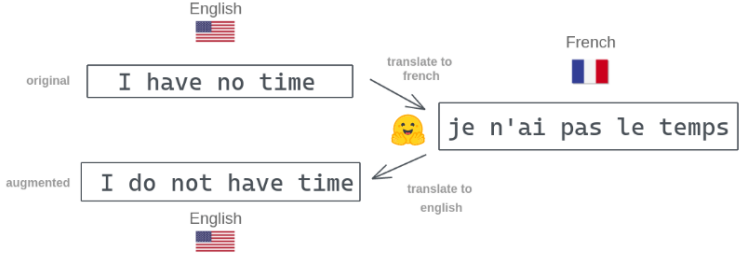


Data Discovery & Augmentation

- **Data Augmentation**
  - Given a base training dataset, it generates synthetic training examples from the training dataset
- **Challenges**
  - How to generate label invariant and diversified examples to make ML models generalizable to unseen data
- Please refer to an excellent VLDB21 Tutorial from Wang et al.



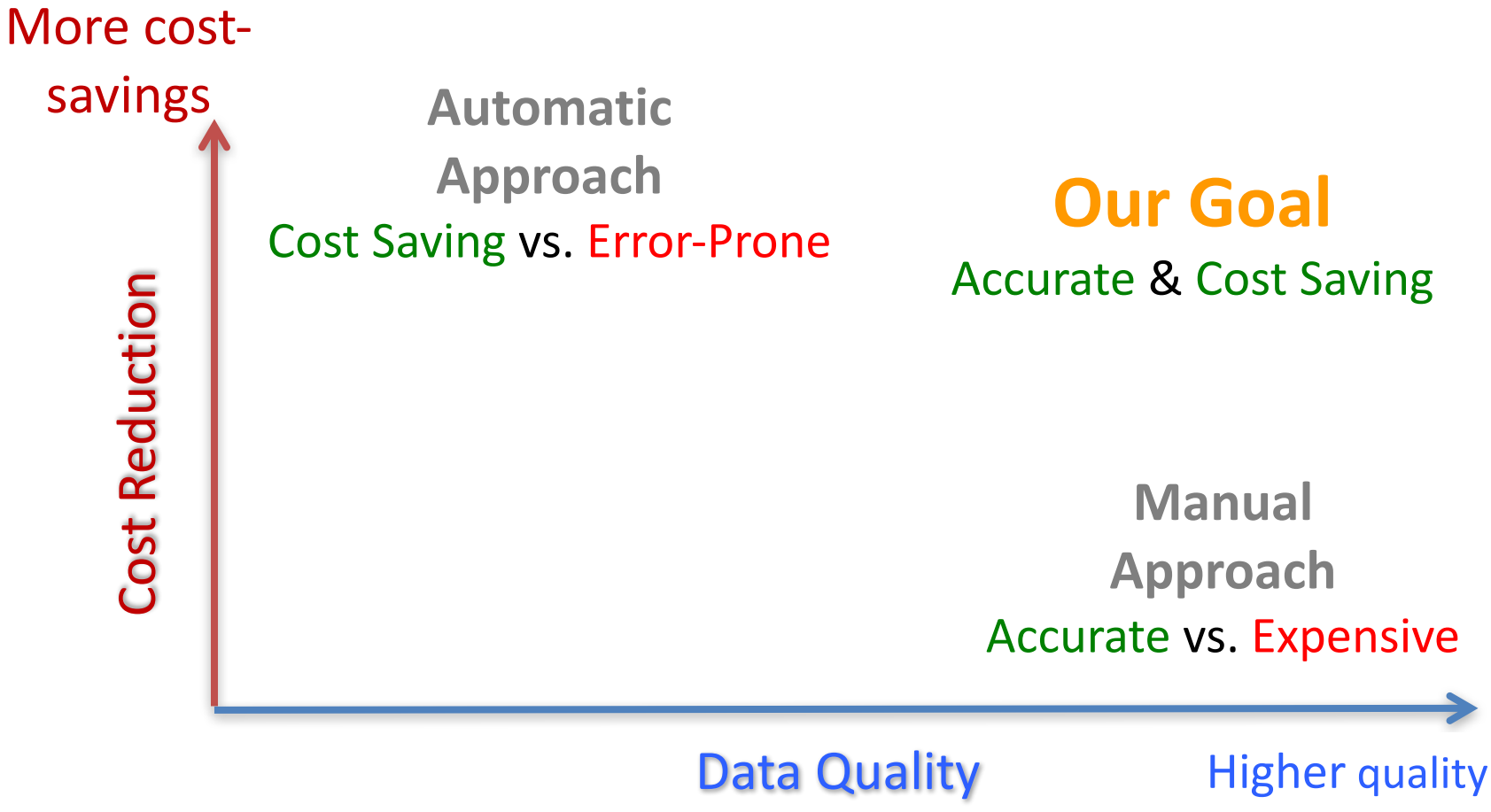
Data Augmentation In CV



Data Augmentation In NLP

Y. Li, X. Wang, Z. Miao, W. Tan: Data Augmentation for ML-driven Data Preparation and Integration. Proc. VLDB Endow. 14(12): 3182-3185 (2021)

Optimization Goal of Data Prep

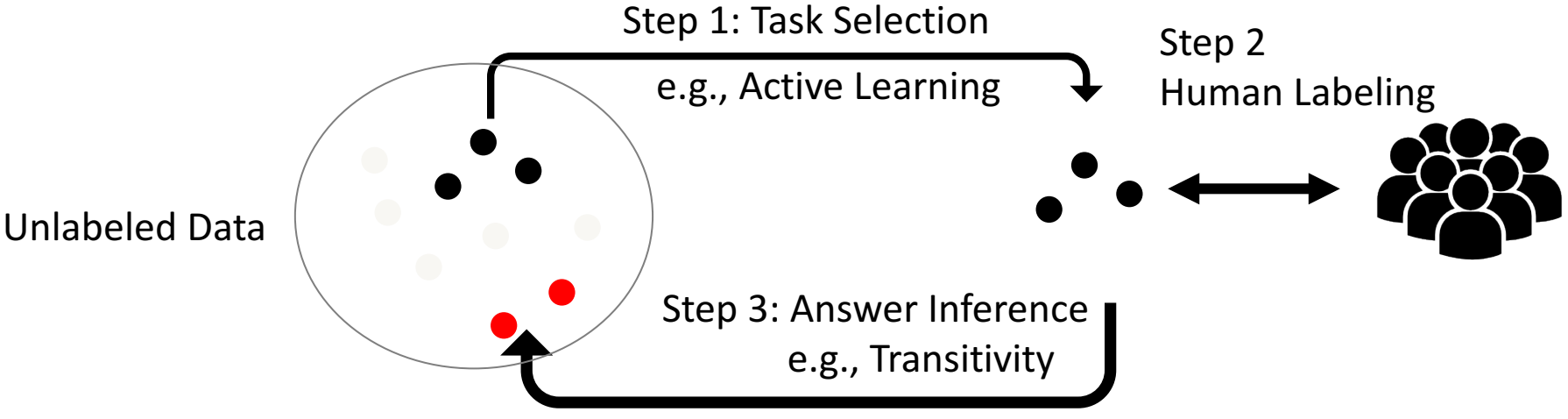


Talk Outline

- An Overview of Data Prep for AI
- **Weak-Supervision for Data Labeling**
- Pre-trained Models for Data Integration
- ML-Oriented Data Cleaning
- Model-Aware Data Discovery
- Summary and Future Directions

Traditional Approach to Data Labeling

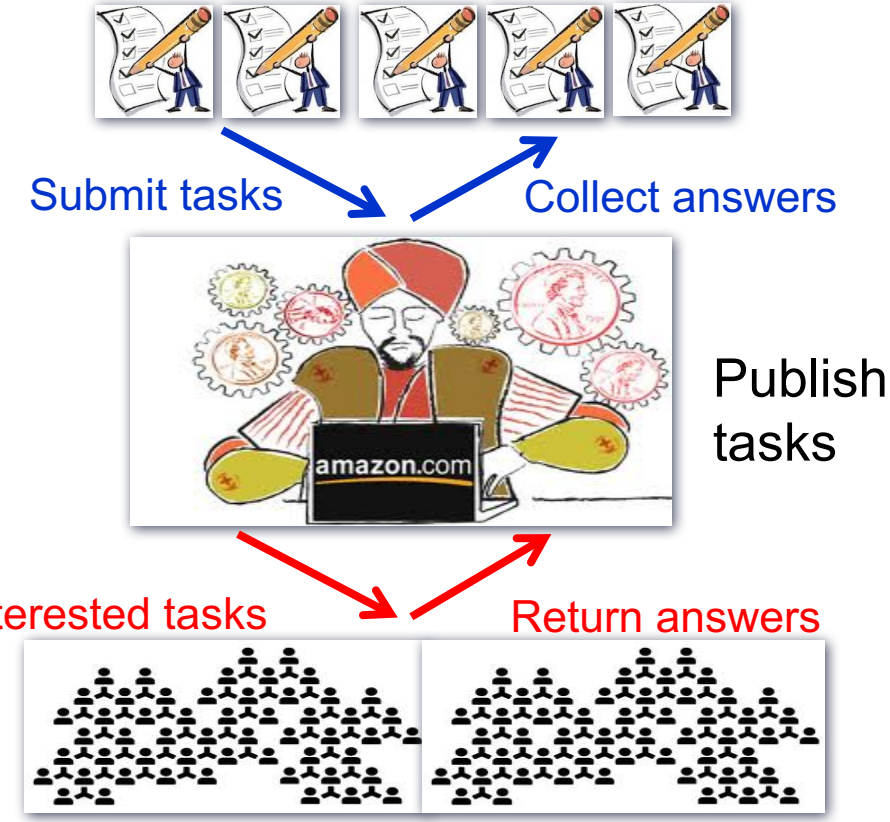
- Most ML approaches ask human for example-level labeling



- Limitation: large human labeling cost for big datasets
  - E.g., asking doctors for medical data labeling

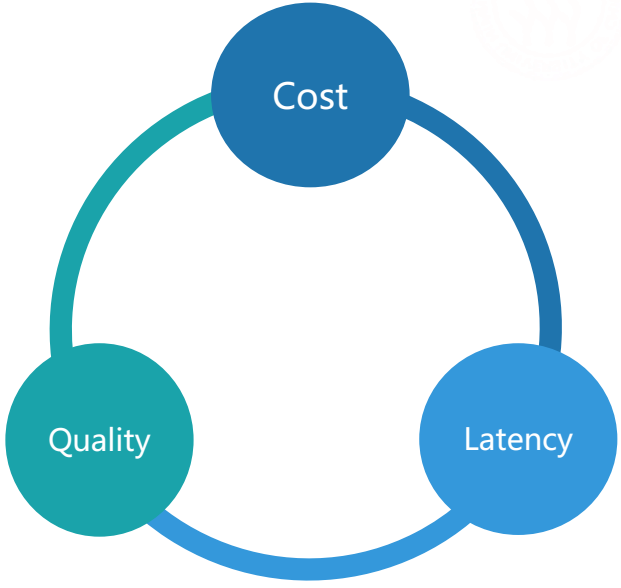
Data Labeling using Crowdsourcing

- Crowdsourcing can be utilized for relatively low cost per task
  - ➢ Requester
    - – Submit Tasks
  - ➢ Platforms
    - – Task Management
  - ➢ Workers
    - – Worker on Tasks



Data Labeling using Crowdsourcing

- Pros
  - A large number of crowd workers
  - Thanks to the crowdsourcing platforms, such as AMT, we can hire workers easily
- Cons
  - Crowd workers are error-prone
  - Crowd workers are not free



- The DB and ML communities have extensively studied how to improve quality and reduce cost & latency for crowdsourcing

G. Li, Y. Zheng, J. Fan, J. Wang, R. Cheng: Crowdsourced Data Management: Overview and Challenges. SIGMOD Conference 2017: 1711-1716

Comparisons of Data Labeling Approaches

Data Labeling Approaches	Label Quality	Cost	Latency
Expert Labeling	High	High	High
Crowdsourcing Labeling	Low	Medium	Medium

Data Labeling Approaches		Label Quality	Cost	Latency
Weak Supervision	Learning from imperfect labels	Medium	Medium	Medium
	Data Programming	High	Low	Low

Learning from Imperfect Labels

- Given observed training data D with N instances from R workers, the task is to
  - Estimate the weight vector w.
  - Estimate the true/false positive rate of each worker.
  - Infer the true label of each instance.

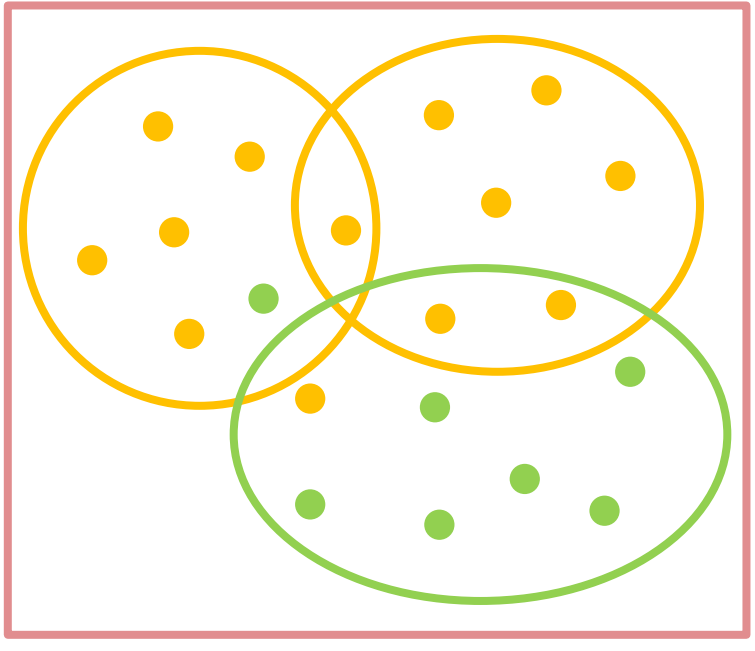
The true labels are treated as hidden variables, which can be solved by EM

$$\begin{aligned}\Pr[\mathcal{D}|\theta] &= \prod_{i=1}^{N}\{\Pr[y_i^1,\ldots,y_i^R|y_i=1,\boldsymbol{\alpha}]\Pr[y_i=1|\boldsymbol{x_i},\boldsymbol{w}] \\ &+ \Pr[y_i^1,\ldots,y_i^R|y_i=0,\boldsymbol{\beta}]\Pr[y_i=0|\boldsymbol{x_i},\boldsymbol{w}]\}.\end{aligned}$$

V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, L. Moy: Learning From Crowds. JMLR. 11: 1297-1322 (2010)

Data Programming: Leveraging Probabilistic Rules

- Data Programming
  - Leveraging probabilistic labeling rules to label examples.

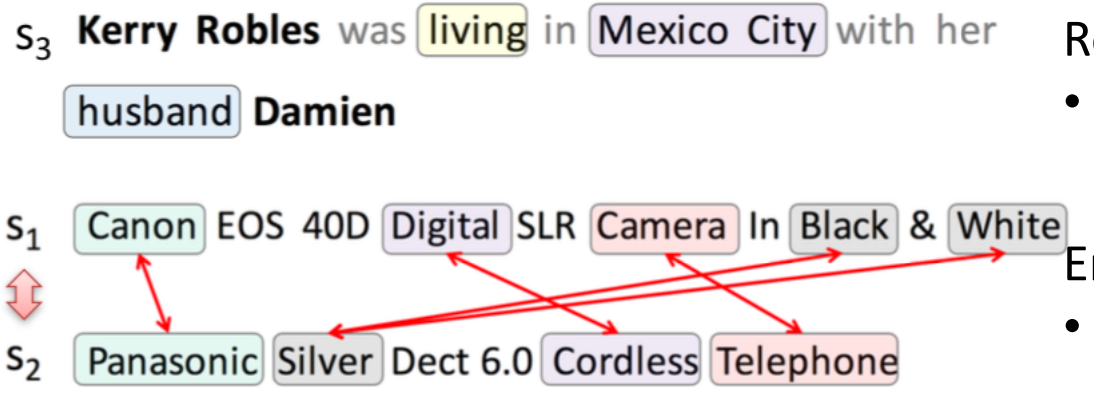


- What are probabilistic rules?
  - A rule covers multiple examples
  - A rule may make mistakes

A. Ratner, S. H. Bach, H. R. Ehrenberg, J. A. Fries, S. Wu, C. Ré: Snorkel: Rapid Training Data Creation with Weak Supervision. PVLDB 11(3): 269-282 (2017)

Data Programming: Leveraging Probabilistic Rules

- Data Programming
  - Leveraging probabilistic labeling rules to annotate examples.

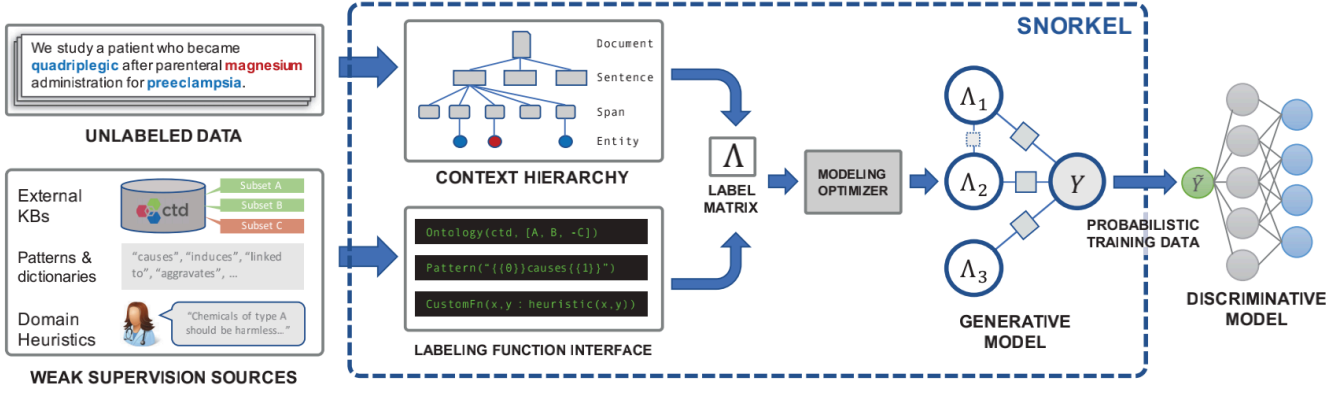


Relation Extraction
- Extracting spouse relation

Entity Matching
- Matching product records

Data Programming: Leveraging Probabilistic Rules

- Rule-generated labels are not perfect but at scale, where larger volume may compensate for lower label quality



A. Ratner, S. H. Bach, H. R. Ehrenberg, J. A. Fries, S. Wu, C. Ré: Snorkel: Rapid Training Data Creation with Weak Supervision. PVLDB 11(3): 269-282 (2017)

- Factor graph: L denotes labeling functions, Y denotes unknown labels

$$\phi^{cov}_{i,j}(\mathcal{L},Y) = \mathbb{I}\{\mathcal{L}^j_i \neq \emptyset\} \quad \phi^{cor}_{i,j,k}(\mathcal{L},Y) = \mathbb{I}\{\mathcal{L}^j_i = \mathcal{L}^k_i\} \quad \phi^{acc}_{i,j}(\mathcal{L},Y) = \mathbb{I}\{\mathcal{L}^j_i = y_i\}$$

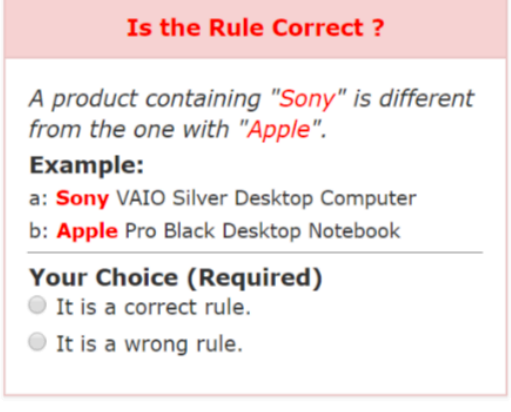
$$p_w(\mathcal{L},Y) = Z_w^{-1}\exp(\sum_{i=1}^{m} w^T\phi_i(\mathcal{L},y_i)),$$

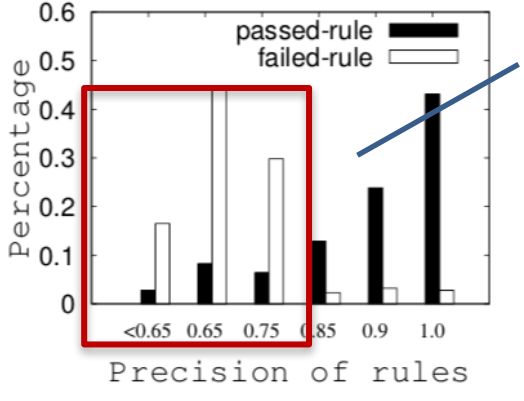
Weak Supervision for Data Labeling

- Challenge: hard to construct good labeling rules
  - Manual
    - Medium precision / Low coverage
  - ML-Generated
    - high Coverage / low quality

Methods	Coverage	Precision
Manual	0.372	0.645
ML-Generated	0.937	0.356

- Example: Human may provide imprecise rules

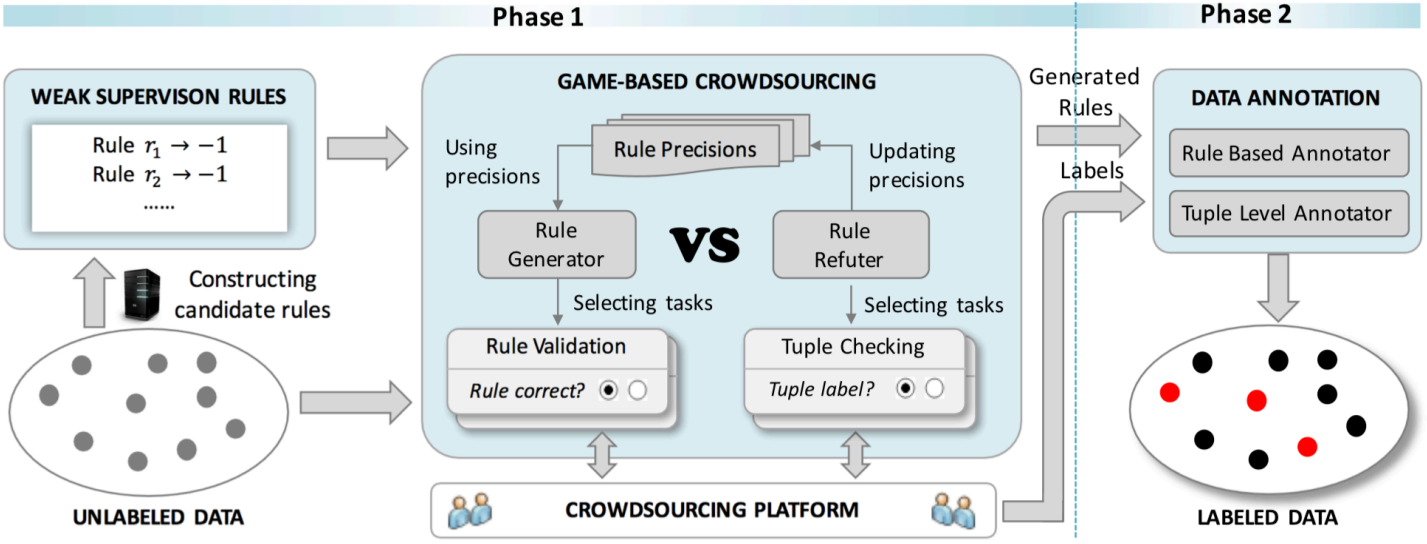




False Positive Rules
E.g., “Apple-Android”

Optimization for Weak Supervision

- CrowdGame: Generating high-quality labeling rules
  - Use ML to generate candidate labeling rules
  - Ask human to select good rules via a Game-based Strategy



J. Yang, J. Fan, Z. Wei, G. Li, T. Liu, X. Du: Cost-Effective Data Annotation using Game-Based Crowdsourcing. PVLDB 2018.
T., J. Yang, J. Fan, Z. Wei, G. Li, X. Du: CrowdGame: A Game-Based Crowdsourcing System for Cost-Effective Data Labeling. SIGMOD 2019, Demo
J. Yang, J. Fan, Z. Wei, G. Li, T. Liu, X. Du: A game-based framework for crowdsourced data labeling. VLDB Journal, 2020

Technical Challenges & Solutions

- Objective that balances among coverage and accuracy of rules
- Developing iterative task selection algorithms

Minimax Optimization

$$\mathcal{O}^{\mathcal{R}_q^*,\mathcal{E}_q^*} = \min_{\mathcal{R}_q} \max_{\mathcal{E}_q} \Phi(\mathcal{R}_q|\mathcal{E}_q)$$

$$\iff \max_{\mathcal{R}_q} \min_{\mathcal{E}_q} \sum_{e_i \in \mathcal{C}(\mathcal{R}_q)} \left\{ P(a_i|\hat{\Lambda}^{\mathcal{R}_i}(\mathcal{E}_q)) - \frac{1-2\gamma}{1-\gamma} \right\}$$

$$\iff \max_{\mathcal{R}_q} \min_{\mathcal{E}_q} \sum_{e_i \in \mathcal{C}(\mathcal{R}_q)} \left\{ \max_{r_j \in \mathcal{R}^i} \hat{\lambda}_j(\mathcal{E}_q) - \frac{1-2\gamma}{1-\gamma} \right\}$$

$\mathcal{C}(\mathcal{R})$, a set of tuples covered by rule set $\mathcal{R}$.

λ_j, precision of rule r_j

Algorithm 1: MinimaxSelect ($\mathcal{R}^c$, $\mathcal{E}$, k, b)

Input: $\mathcal{R}^c$: candidate rules; $\mathcal{E}$: tuples to be labeled; k: a budget; b: a crowdsourcing batch
Output: $\mathcal{R}_q$: a set of generated rules

Initialize $\mathcal{R}_q \leftarrow \emptyset$, $\mathcal{E}_q \leftarrow \emptyset$;
for *each iteration* t **do**
  /* Rule Generator Step */
  Select $\mathcal{R}_q^{(t)} \leftarrow \arg_{\mathcal{R}} \max_{\mathcal{R} \subseteq \mathcal{R}^c - \mathcal{R}_q, |\mathcal{R}|=b} \Delta g(\mathcal{R}|\mathcal{J}^{\mathcal{R}_q,\mathcal{E}_q})$;
  Crowdsource $\mathcal{R}_q^{(t)}$ as rule validation tasks ;
  Add the crowd validated rules in $\mathcal{R}_q^{(t)}$ into $\mathcal{R}_q$;
  Update objective $\mathcal{J}^{\mathcal{R}_q,\mathcal{E}_q}$;
  $\mathcal{R}^c \leftarrow \mathcal{R}^c - \mathcal{R}_q^{(t)}$;
  /* Rule Refuter Step */
  Select $\mathcal{E}_q^{(t)} \leftarrow \arg_{\mathcal{E}} \min_{\mathcal{E} \in \mathcal{E} - \mathcal{E}_q, |\mathcal{E}|=b} \Delta f(\mathcal{E}_q|\mathcal{J}^{\mathcal{R}_q,\mathcal{E}_q})$;
  Crowdsource $\mathcal{E}_q^{(t)}$ as tuple checking tasks ;
  Add the crowd-checked $\mathcal{E}_q^{(t)}$ into $\mathcal{E}_q$;
  Update precision $\hat{\Lambda}^{\mathcal{R}_q}(\mathcal{E}_q)$;
  Update budget $k \leftarrow k - 2b$;
  if $k = 0$ **then break** ;
Remove rules from $\mathcal{R}_q$ with $\hat{\lambda}_j \leq \frac{1-2\gamma}{1-\gamma}$;
Return $\mathcal{R}_q$;

Experimental Study

Data Labeling for Entity Resolution

Dataset	Method	$F1$ of EM	Total Crowd Cost
Abt-Buy	Trans	0.864	203,715
	PartOrder	0	**1,063**
	ACD	0.887	216,025
	Snorkel	0.909	26,381
	CrowdGame	**0.963**	26,381
Ama-Goo	Trans	0.896	158,525
	PartOrder	0.486	**763**
	ACD	0.919	167,958
	Snorkel	0.923	48,115
	CrowdGame	**0.982**	48,115
Ebay	Trans	0.971	50,163
	PartOrder	0.553	**170**
	ACD	**0.998**	57,637
	Snorkel	0.857	7,410
	CrowdGame	**0.998**	7,410

Data Labeling for Relation Extraction

Method	Precision	Recall	F_1
Snorkel (ManRule)	0.389	0.608	0.474
Snorkel (ManRule+Crowd)	0.519	**0.696**	0.595
CrowdGame	**0.81**	0.635	**0.712**

- CrowdGame achieves better result quality with low cost
- CrowdGame learns good rules from noisy candidates, and it outperforms the manual rules from domain experts.

Takeaways

- Data labeling usually introduces the trade-offs between quality, cost and latency.
- Crowdsourcing is cheap but with low-quality, while expert labeling is expensive but with high-quality.
- Weak supervision is a good choice to largely reduce cost while still ensuring labeling quality
- However, It is crucial to generate good labeling rules for weak supervision

Talk Outline

- An Overview of Data Prep for AI
- Weak-Supervision for Data Labeling
- **Pre-trained Models for Data Integration**
- ML-Oriented Data Cleaning
- Model-Aware Data Discovery
- Summary and Future Directions

Entity Resolution: A Core Task in Data Integration

- Entity resolution (ER) is to determine whether two data instances refer to the same real-world entity.

Table A

id	name	description	price
a_1	samsung 52 ' series 7 black flat ...	samsung 52 ' series 7 black flat panel lcd ...	NULL
a_2	sony 46 ' bravia ...	bravia z series ...	NULL
a_3	linksys wirelessn ...	security router ...	NULL

$(a_1, b_1, ?)$

$(a_2, b_2, ?)$

$(a_3, b_3, ?)$

Table B

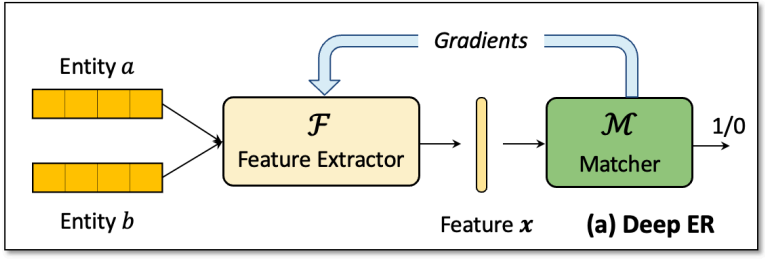
id	name	description	price
b_1	samsung ln52a750 ...	dynamic contrast ratio 120hz 6ms respons ...	2148.99
b_2	sony bravia ...	ntsc 16:9 1366 x 768 ...	597.72
b_3	linksys wirelessg ...	54mbps	NULL

- Traditional Approaches to ER
  - Rule-based: Disjunctive normal form, general Boolean formula, etc.
  - ML-based methods: SVM, Random Forests, etc.
  - Please refer to a VLDB12 tutorial from Getoor and Machanavajjhala

Lise Getoor, Ashwin Machanavajjhala: Entity Resolution: Theory, Practice & Open Challenges. Proc. VLDB Endow. 5(12): 2018-2019 (2012)

DL for Entity Resolution (Deep ER)

- Not Surprisingly, DL-based methods achieve the state-of-the-art (SOTA) results
- A Typical DL-based Framework for ER :
  - **Feature Extractor** converts entity pair (a, b) into a d-dimensional vector (*features*)
  - **Matcher** takes the feature of entity pair as input, and predicts whether they match or not.



$$\hat{y} = \mathcal{M}(\mathbf{x}) = \mathcal{M}(\mathcal{F}(a, b))$$

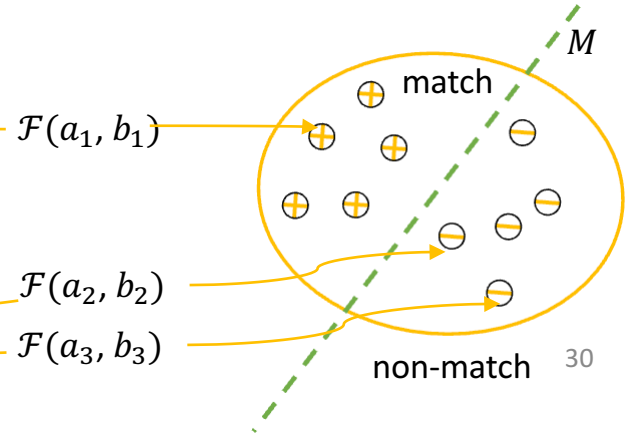
Table A: entity a

id	title	category	brand	price
a_1	balt wheasel ...	stationery ...	balt	239.88
a_2	kodak esp ...	printers	NULL	58.0
a_3	hp q3675a ...	printers	hp	194.84

$(a_1, b_1, 1)$
$(a_2, b_2, 0)$
$(a_3, b_3, 0)$

Table B: entity b

id	title	category	brand	price
b_1	balt inc. ...	laminating ...	mayline	134.45
b_2	kodak esp 7 ...	kodak	NULL	149.29
b_3	hewlett ...	cleaning repair	hp	NULL



DL for Entity Resolution (Deep ER)

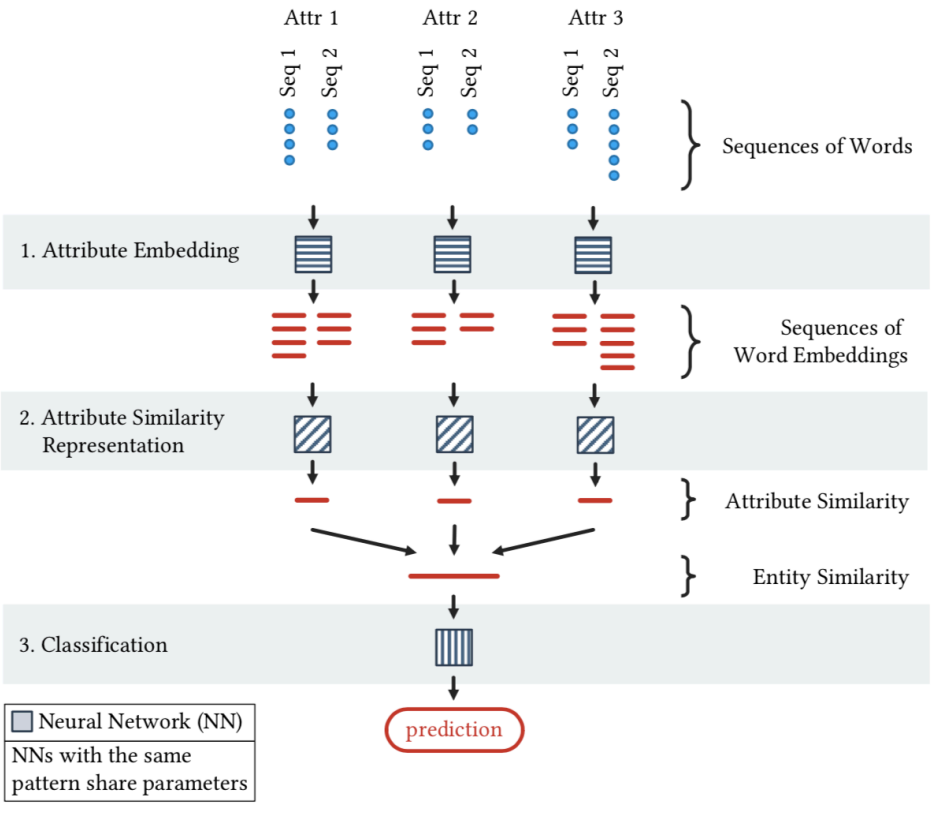
- DeepMatcher
  - **Attribute Embedding**

    Each attribute value is tokenized and converted into a sequence of embedding vectors
  - **Attribute Similarity Representation**

    One vector with the similarity per attribute is generated

    Entity Similarity: concatenate attribute similarities
  - **Binary classification**

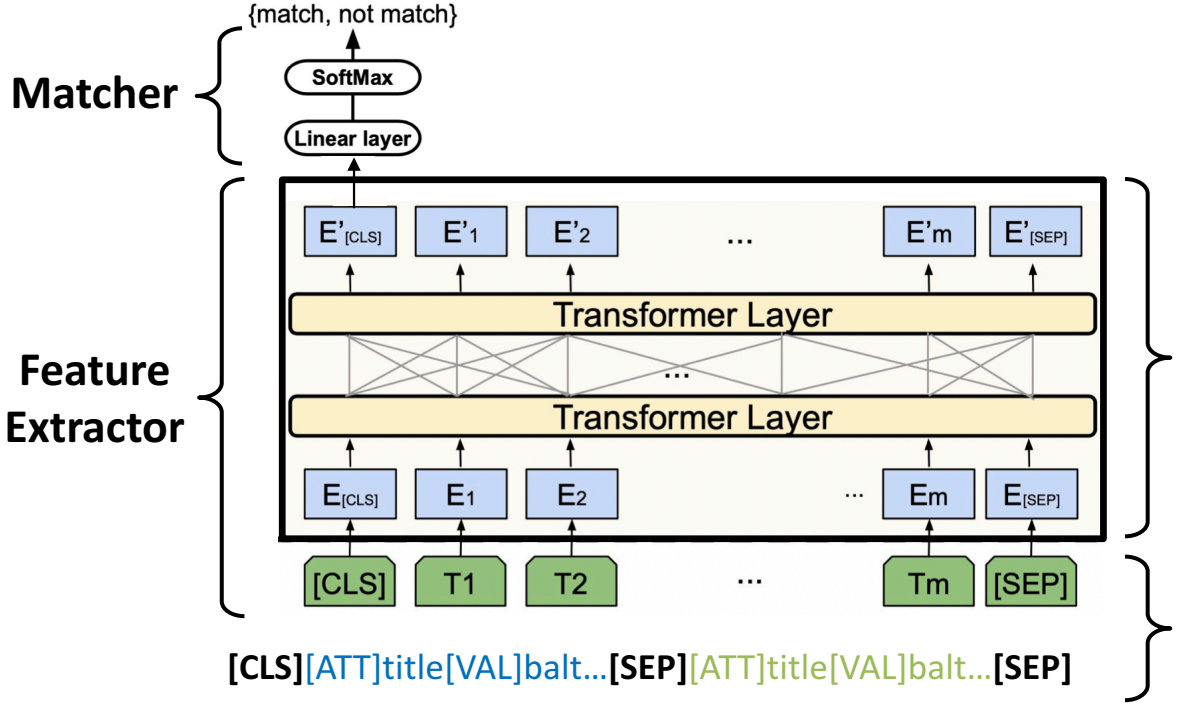
    A MLP-based solution is utilized



S. Mudgal, H. Li, T. Rekatsinas, A.H. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, V. Raghavendra. Deep Learning for Entity Matching: A Design Space Exploration. SIGMOD Conference 2018: 19-34

DL for Entity Resolution (Deep ER)

- Ditto: DL-based ER with pretrained Transformer Models



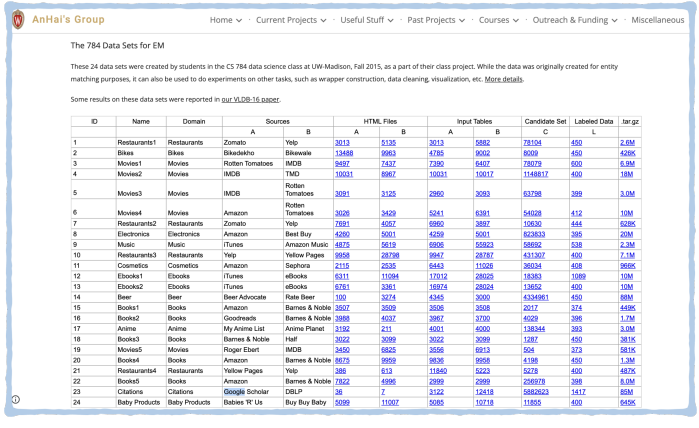
Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. Deep Entity Matching with Pre-Trained Language Models. PVLDB, 14(1): 50 - 60, 2021.

Problem: DL-based methods need a large amount of labeled training data.

Opportunity of Reusing Well-Labeled ER Datasets

- There are many well-labeled ER datasets, either public on the Web or available in enterprises
  - E.g., Magellan datasets and WDC datasets

Magellan datasets



AnHai's Group — Home · Current Projects · Useful Stuff · Past Projects · Courses · Outreach & Funding · Miscellaneous

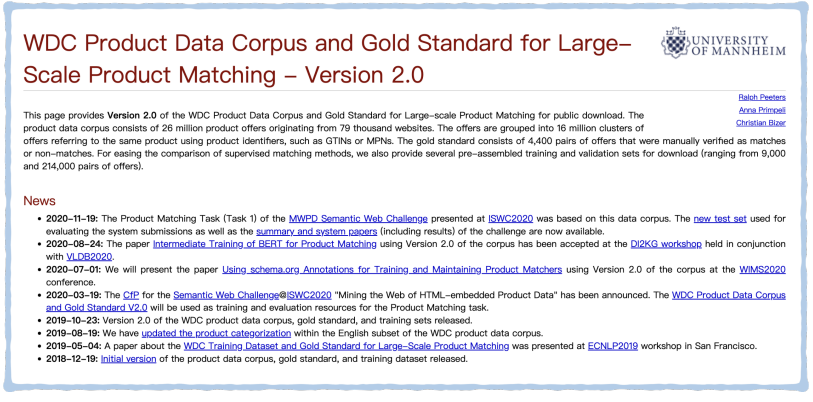
The 784 Data Sets for EM

These 24 data sets were created by students in the CS 784 data science class at UW-Madison, Fall 2015, as a part of their class project. While the data was originally created for entity matching purposes, it can also be used to do experiments on other tasks, such as wrapper construction, data cleaning, visualization, etc. More details.

Some results on these data sets were reported in our VLDB-16 paper.

ID	Name	Domain	Sources		HTML Files		Input Tables		Candidate Set	Labeled Data	.tar.gz
			A	B	A	B	A	B	C	L	
1	Restaurants1	Restaurants	Zomato	Yelp	3013	5135	3013	5882	78104	450	2.6M
2	Bikes	Bikes	Bikedekho	Bikewale	13488	9983	4785	9002	8009	450	428K
3	Movies1	Movies	Rotten Tomatoes	IMDB	9497	7437	7390	6407	78079	600	6.9M
4	Movies2	Movies	IMDB	TMD	10031	8967	10031	10017	1148817	400	18M
5	Movies3	Movies	IMDB	Rotten Tomatoes	3091	3125	2960	3093	63798	399	3.0M
6	Movies4	Movies	Amazon	Rotten Tomatoes	3026	3429	5241	6381	54028	412	10M
7	Restaurants2	Restaurants	Zomato	Yelp	7691	4957	6960	3897	10630	444	628K
8	Electronics	Electronics	Amazon	Best Buy	4260	5001	4259	5001	823833	395	20M
9	Music	Music	iTunes	Amazon Music	4875	5619	6906	55923	58692	538	2.3M
10	Restaurants3	Restaurants	Yelp	Yellow Pages	9958	28798	9947	26787	431307	400	7.1M
11	Cosmetics	Cosmetics	Amazon	Sephora	2115	2535	6443	11026	36034	408	966K
12	Ebooks1	Ebooks	iTunes	eBooks	6311	11094	17012	28025	18383	1089	10M
13	Ebooks2	Ebooks	iTunes	eBooks	6761	3361	16974	28024	13652	400	10M
14	Beer	Beer	Beer Advocate	Rate Beer	100	3274	4345	3000	4334981	450	68M
15	Books1	Books	Amazon	Barnes & Noble	3507	3509	3506	3508	2017	374	449K
16	Books2	Books	Goodreads	Barnes & Noble	3985	4037	3967	3700	4029	396	1.7M
17	Anime	Anime	My Anime List	Anime Planet	3192	211	4001	4000	138344	393	3.0M
18	Books3	Books	Barnes & Noble	Half	3022	3099	3022	3099	1287	450	381K
19	Movies5	Movies	Roger Ebert	IMDB	3450	6825	3656	6913	504	373	581K
20	Books4	Books	Amazon	Barnes & Noble	8675	9959	9836	9958	4198	450	1.3M
21	Restaurants4	Restaurants	Yellow Pages	Yelp	388	613	11840	5223	5278	400	487K
22	Books5	Books	Amazon	Barnes & Noble	7822	4996	2999	2999	256978	398	6.0M
23	Citations	Citations	Google Scholar	DBLP	36	7	3122	12418	5862623	1417	85M
24	Baby Products	Baby Products	Babies 'R' Us	Buy Buy Baby	5099	11007	5085	10718	11855	400	645K

WDC datasets



WDC Product Data Corpus and Gold Standard for Large–Scale Product Matching – Version 2.0

UNIVERSITY OF MANNHEIM

Ralph Peeters
Anna Primpeli
Christian Bizer

This page provides **Version 2.0** of the WDC Product Data Corpus and Gold Standard for Large–scale Product Matching for public download. The product data corpus consists of 26 million product offers originating from 79 thousand websites. The offers are grouped into 16 million clusters of offers referring to the same product using product identifiers, such as GTINs or MPNs. The gold standard consists of 4,400 pairs of offers that were manually verified as matches or non–matches. For easing the comparison of supervised matching methods, we also provide several pre–assembled training and validation sets for download (ranging from 9,000 and 214,000 pairs of offers).

News

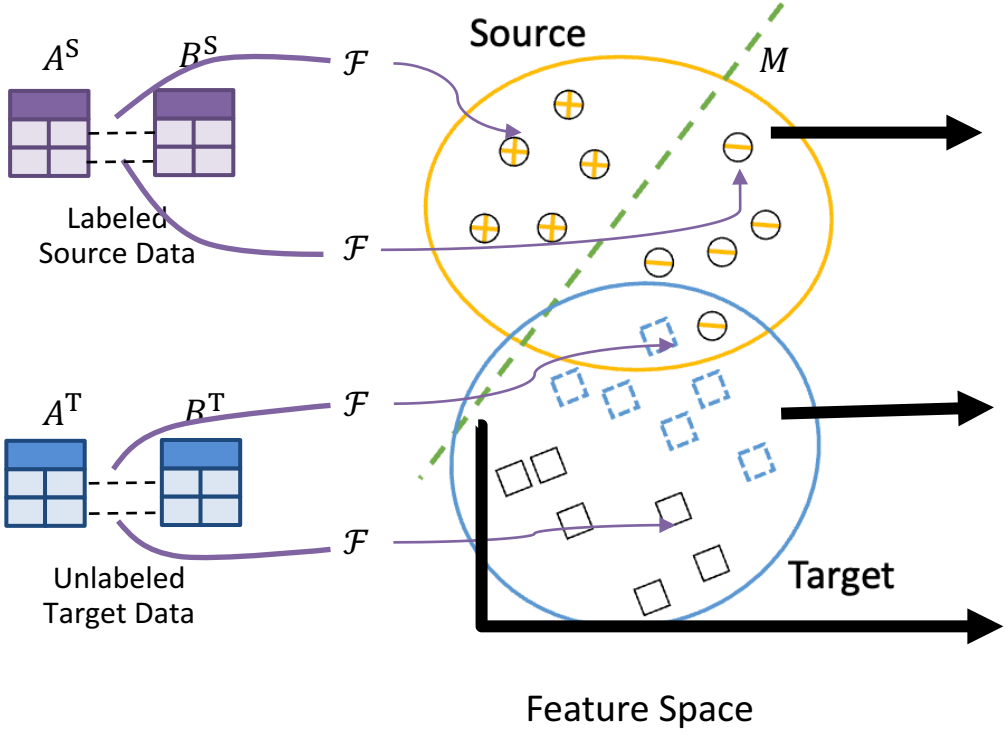
- **2020–11–19:** The Product Matching Task (Task 1) of the MWPD Semantic Web Challenge presented at ISWC2020 was based on this data corpus. The new test set used for evaluating the system submissions as well as the summary and system papers (including results) of the challenge are now available.
- **2020–08–24:** The paper Intermediate Training of BERT for Product Matching using Version 2.0 of the corpus has been accepted at the DI2KG workshop held in conjunction with VLDB2020.
- **2020–07–01:** We will present the paper Using schema.org Annotations for Training and Maintaining Product Matchers using Version 2.0 of the corpus at the WIMS2020 conference.
- **2020–03–19:** The CfP for the Semantic Web Challenge@ISWC2020 "Mining the Web of HTML–embedded Product Data" has been announced. The WDC Product Data Corpus and Gold Standard V2.0 will be used as training and evaluation resources for the Product Matching task.
- **2019–10–23:** Version 2.0 of the WDC product data corpus, gold standard, and training sets released.
- **2019–08–19:** We have updated the product categorization within the English subset of the WDC product data corpus.
- **2019–05–04:** A paper about the WDC Training Dataset and Gold Standard for Large–Scale Product Matching was presented at ECNLP2019 workshop in San Francisco.
- **2018–12–19:** Initial version of the product data corpus, gold standard, and training dataset released.

Can we reuse these labeled ER datasets for a new unlabeled ER dataset ?

Source

Target

Directly Reusing Feature Extractor and Matcher Trained on Labeled Source?



Step1: Directly using **Feature Extractor $\mathcal{F}$** and the **Matcher M** trained by labeled source data.

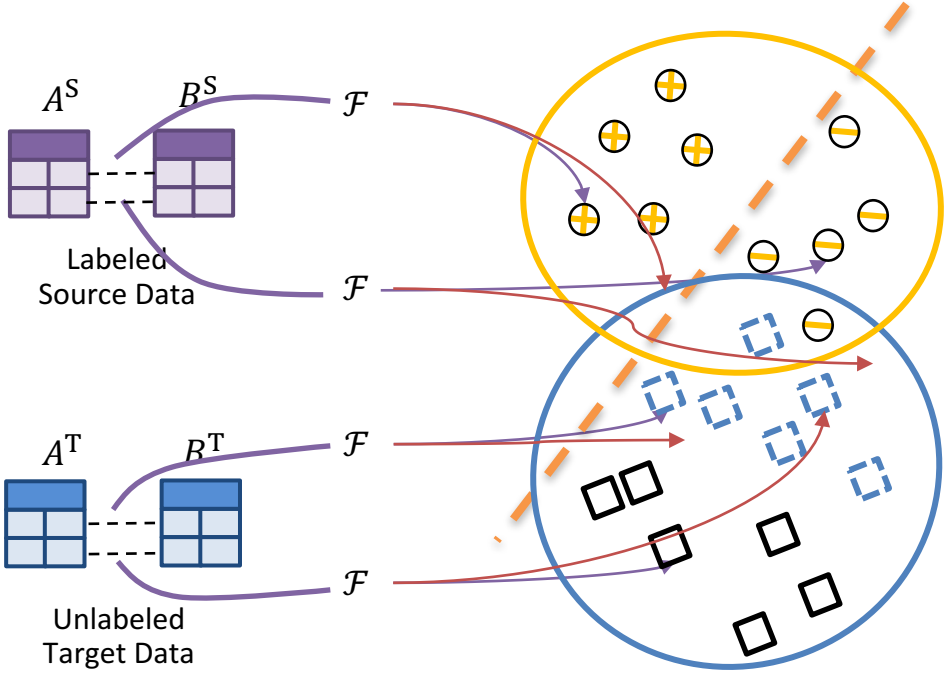
Step2: $\mathcal{F}$ Maps the unlabeled target data into the learned feature space.

Step3: M Predicts the target data

⚠ However, $\mathcal{F}$ and M will fail on the target due to the distribution change or domain shift between source and target

Domain Adaptation (DA) for Deep ER

- Learn domain-invariant and discriminative features.



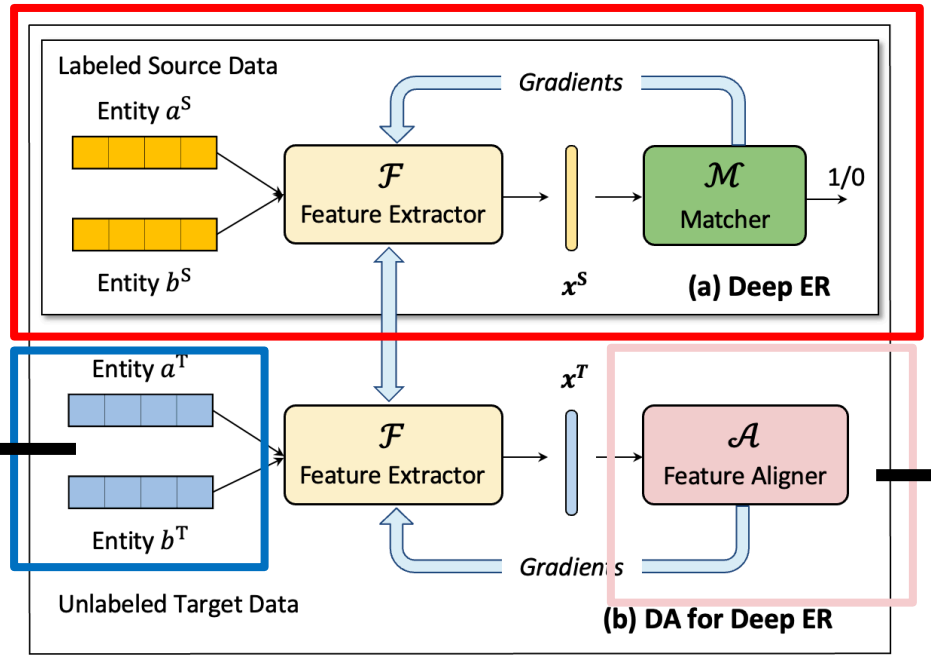
Domain-invariant : reducing distribution change.

Discriminative : extracting discriminative information.

Whether DA can be used for ER tasks?

DADER Framework

- Feature Extractor and Matcher
- Feature Aligner: the key module to realize domain adaptation.



① Training $\mathcal{F}$ and M with labeled source data.

② Utilizing unlabeled target data.

③ Adjusting $\mathcal{F}$ and M to align distributions of source and target, such that $\mathcal{F}$ and M can work well on target.

Jianhong Tu, Ju Fan, Nan Tang, Peng Wang, Chengliang Chai, Guoliang Li, Ruixue Fan, and Xiaoyong Du. 2022. Domain Adaptation for Deep Entity Resolution. In Proceedings of the 2022 International Conference on Management of Data (SIGMOD '22)

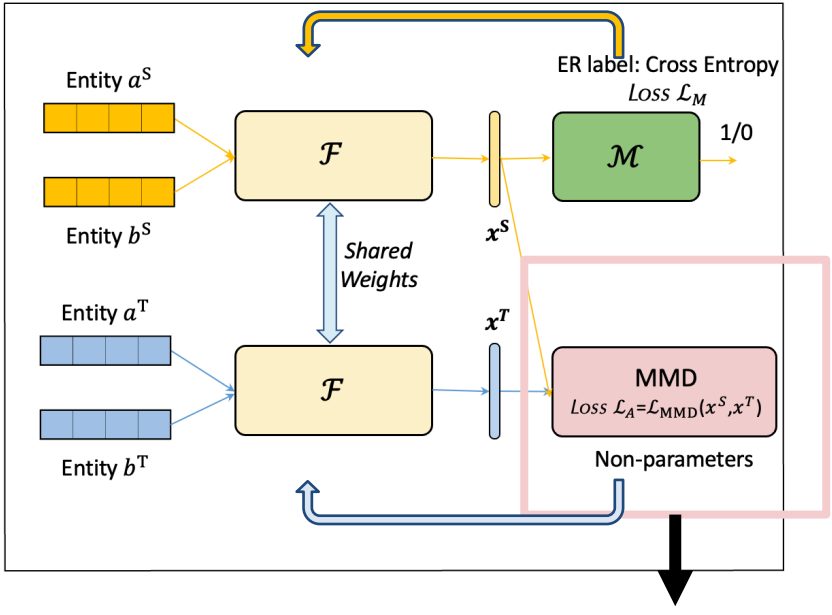
DADER Design Space

- Feature Extractor: RNN, LMs
- Matcher: MLP
- Feature Aligner: Discrepancy-based, Adversarial-based, Reconstruction-based

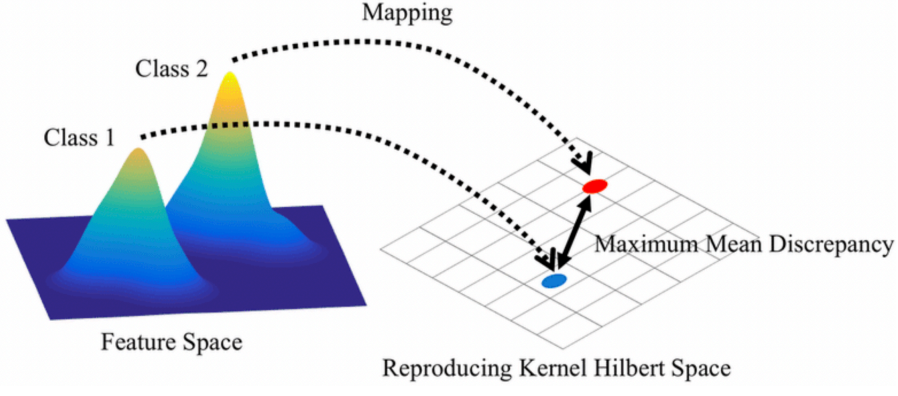
Modules	Categorization	
Feature Extractor ($\mathcal{F}$)	(I) Recurrent neural network (RNN) (II) Pre-trained language models (LMs)	
Matcher ($\mathcal{M}$)	Multi-layer Perceptron (MLP)	
Feature Aligner ($\mathcal{A}$)	(1) Discrepancy-based	(a) MMD (b) K-order
	(2) Adversarial-based	(c) GRL (d) InvGAN (e) InvGAN+KD
	(3) Reconstruction-based	(f) ED

Representative Method: MMD (Discrepancy-based)

- Feature Aligner is a function to measure **maximum mean discrepancy**.



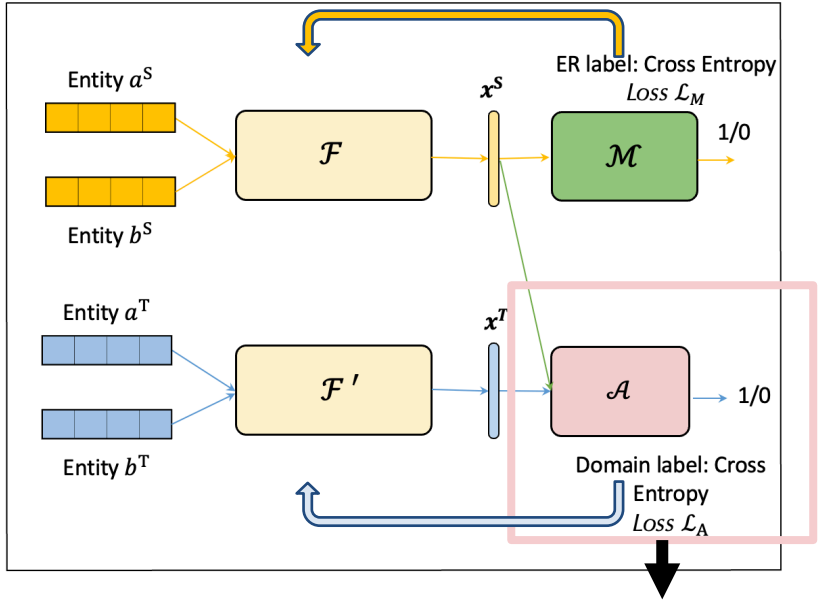
During training, the Maximum Mean Discrepancy of source and target feature spaces is **computed** and **reduced**. The smaller the MMD, the more similar the distributions.



$$\mathcal{L}_{\mathrm{MMD}} = \sup_{\|\phi\|_H \leq 1} \|E_{\mathbf{x}^S \sim p_S}[\phi(\mathbf{x}^S)] - E_{\mathbf{x}^T \sim p_T}[\phi(\mathbf{x}^T)]\|_H^2$$

Representative Method: InvGAN (Adversarial-based)

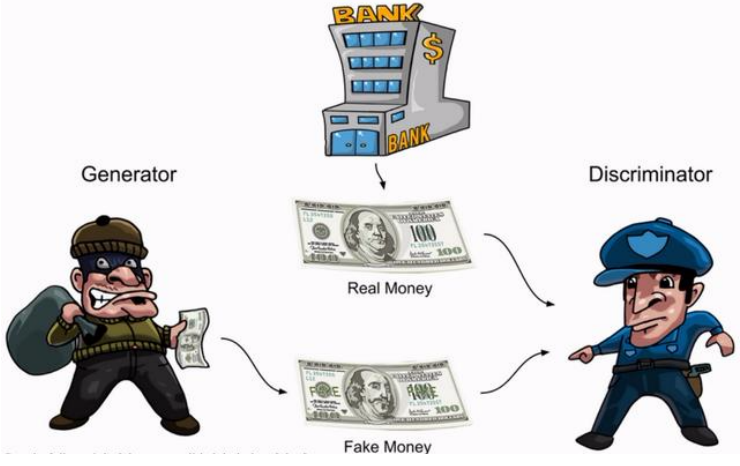
- Feature Aligner is a binary domain classifier to discriminate source/target dataset.



$$\min_{\mathcal{F},\mathcal{M}} \max_{\mathcal{A}} V(\mathcal{F}, \mathcal{M}, \mathcal{A}) = \mathcal{L}_M(\mathcal{F}, \mathcal{M}) + \beta \mathcal{L}_A(\mathcal{F}, \mathcal{A}),$$

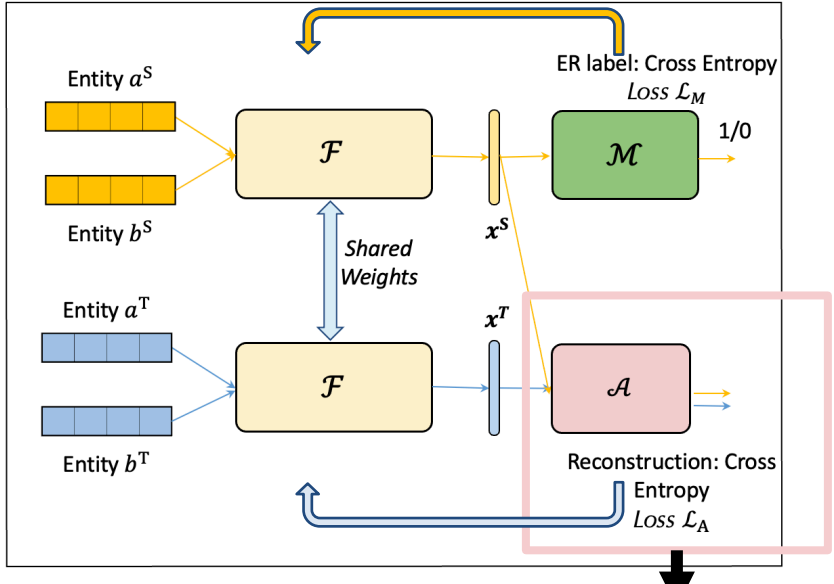
$$\mathcal{L}_A = E_{x^S \sim \mathcal{D}^S} \log \mathcal{A}(\mathcal{F}(x^S)) + E_{x^T \sim \mathcal{D}^T} \log(1 - \mathcal{A}(\mathcal{F}(x^T))),$$

During training, the optimization objective of Feature Aligner is to **minimize the domain classification loss**, while Feature Extractor is to generate the **indistinguishable features** that confuse Feature Aligner.



Representative Method: ED (Reconstruction-based)

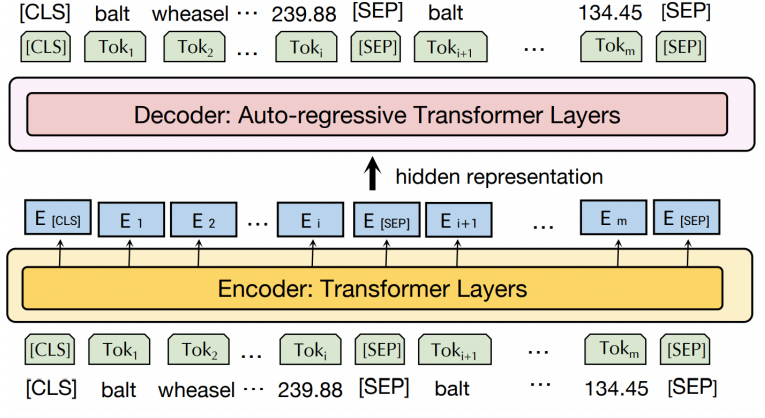
- Feature Aligner is a decoder to reconstruct the initial data for source and target.



$$\mathcal{L}_{\text{REC}} = E_{x \sim \mathcal{D}^S \cup \mathcal{D}^T} [\mathcal{L}_{CE}(\mathcal{A}(\mathcal{F}(x)), x)]$$

During training, the **auxiliary reconstruction task** can ensure the shared Feature Extractor (encoder) to extract important and shared information from both domains.

One example of Encoder-Decoder (ED) Architecture: Bart



Datasets (DeepMatcher, Magellan, and WDC)

Similar domains: Partially different attributes

Similar domains: Partially different textual styles

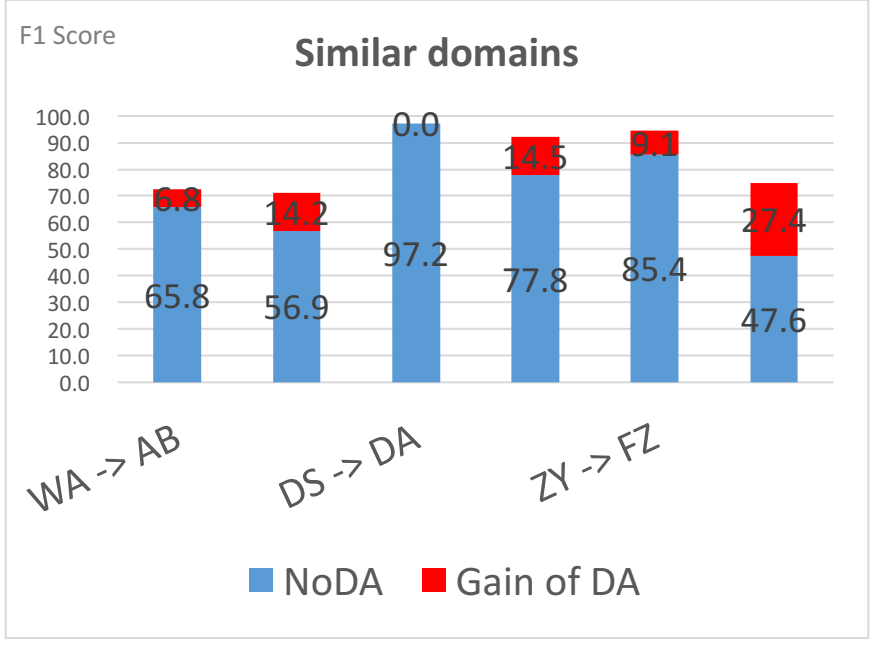
Different domains: Totally different attributes

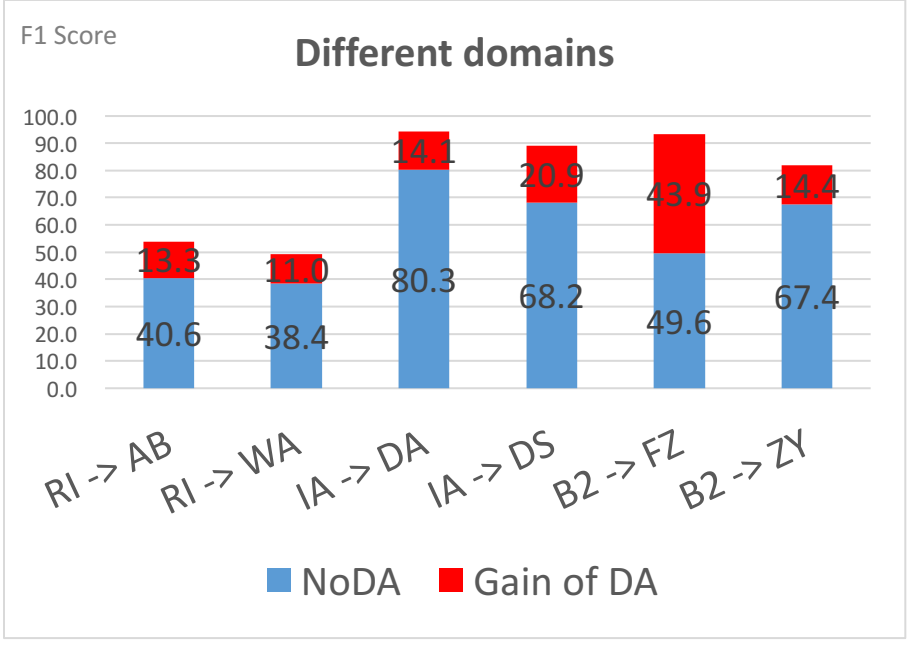
Similar domains: different categories within the same website.

Datasets	Domain	#Pairs	#Matches	#Attrs
Walmart-Amazon (WA)	Product	10,242	962	5
Abt-Buy (AB)	Product	9,575	1,028	3
DBLP-Scholar (DS)	Citation	28,707	5,347	4
DBLP-ACM (DA)	Citation	12,363	2,220	4
Fodors-Zagats (FZ)	Restaurant	946	110	6
Zomato-Yelp (ZY)	Restaurant	894	214	3
iTunes-Amazon (IA)	Music	532	132	8
RottenTomatoes-IMDB (RI)	Movies	600	190	3
Books2 (B2)	Books	394	92	9
WDC-Computers (CO)	Product	1,100	300	2
WDC-Cameras (CA)	Product	1,100	300	2
WDC-Watches(WT)	Product	1,100	300	2
WDC-Shoes (SH)	Product	1,100	300	2

Overall Results of DA for ER

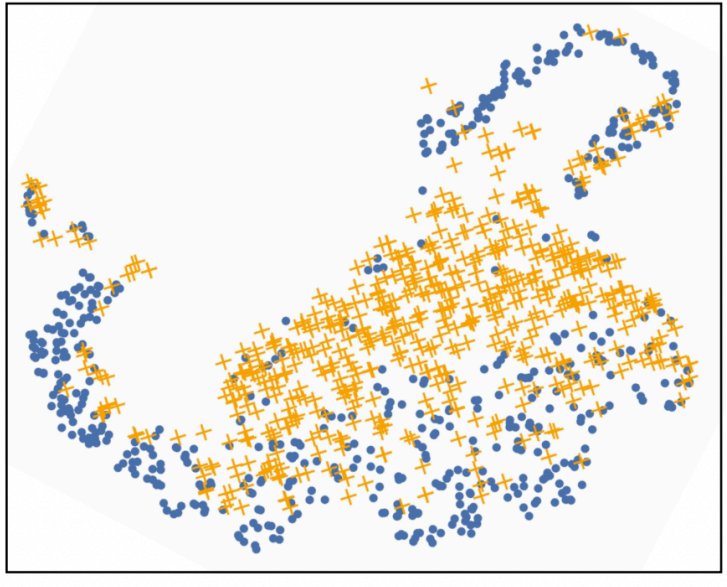
- DA works well on the datasets from:
  - Similar domains, e.g., WA (product)→ AB (product).
  - Different domains, e.g., RI (movie) → AB (product).



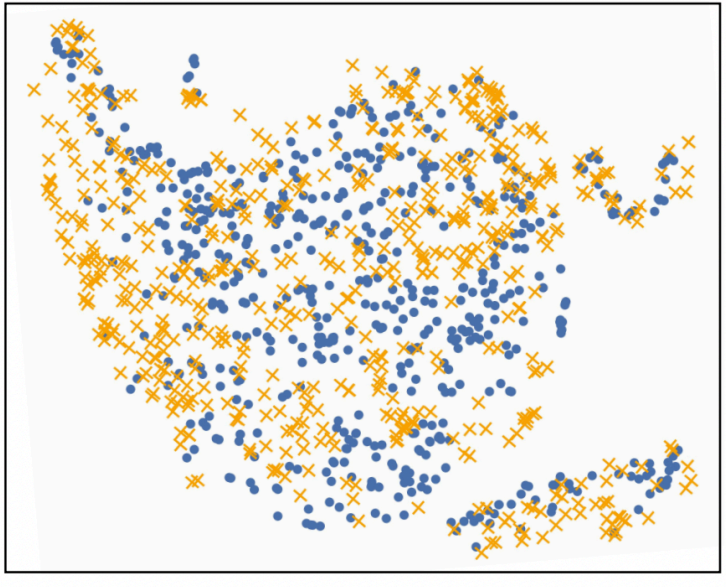


Effect of Domain Shift Reduction

- Datasets: Abt-Buy (Source) → Walmart-Amazon (Target).
- Distributions of source and target are much closer after DA (b) than without DA (a).



(a) NoDA

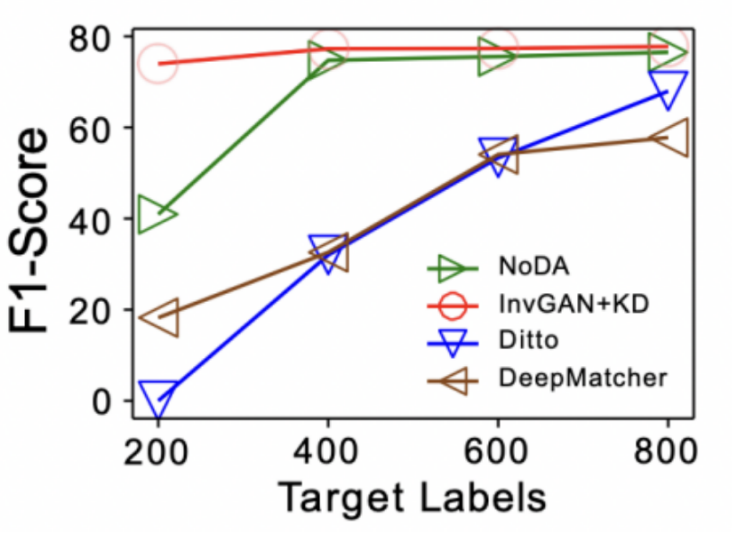


(b) DA (InvGAN+KD)

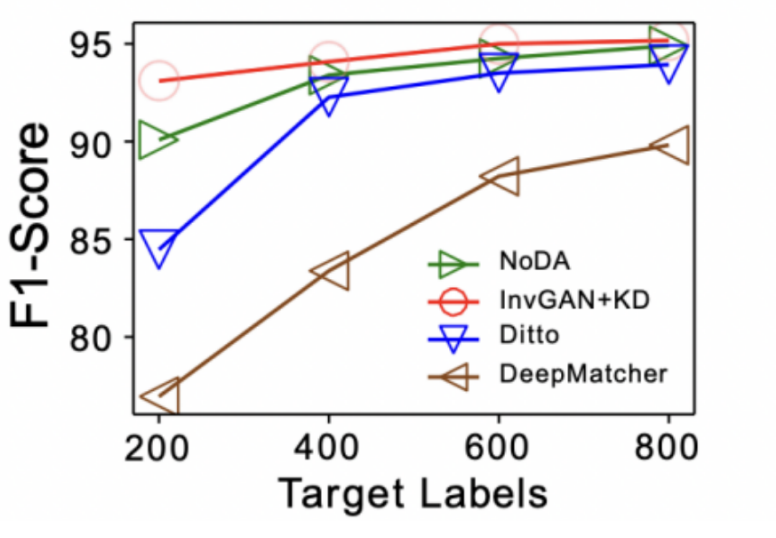
Comparison with a Few Labels.

- The performance of the model after DA can always **be maintained at a high level** with some labeled data.

Walmart-Amazon



DBLP-Scholar



Remaining Challenges: Data Matching Tasks

- Data matching generally refers to the process of deciding whether two data elements are the same (a.k.a. a "match")

Data Elements

String

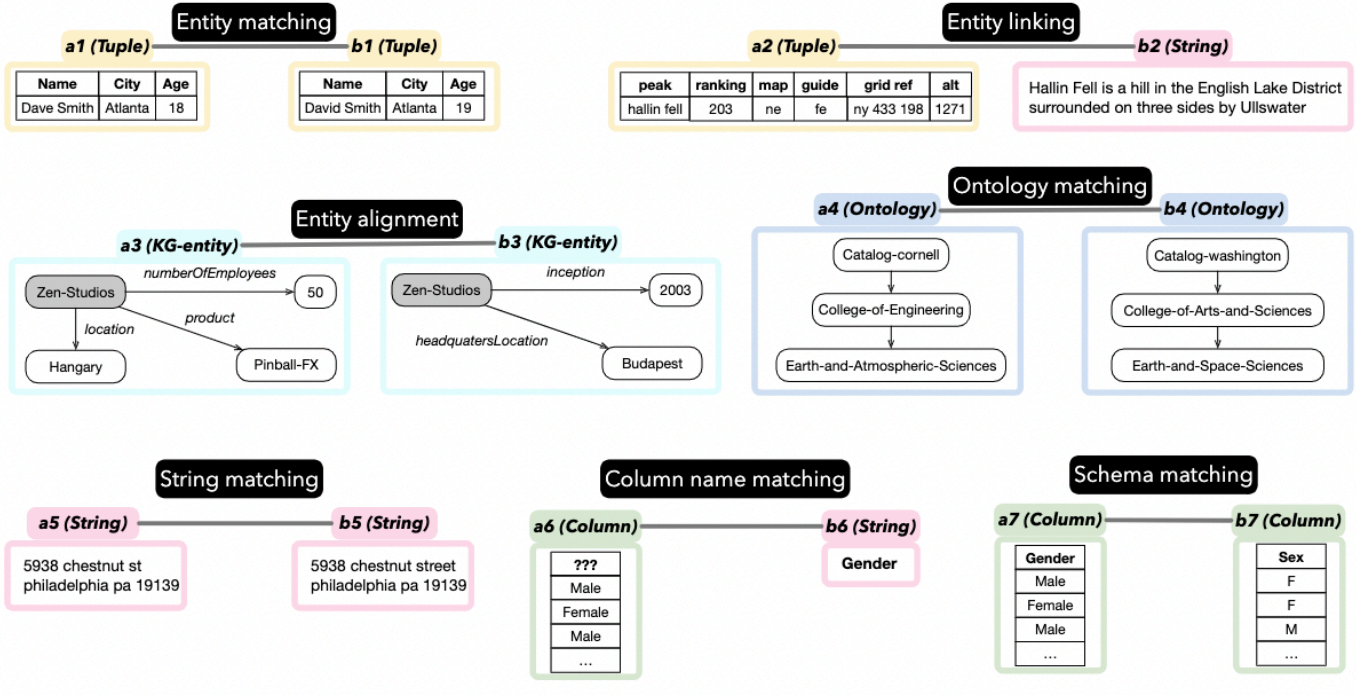
Tuple

Column

Ontology (tree)

Knowledge Graph Entity

Seven Common Data Matching Tasks



Limitations of Existing Solutions

- Due to their importance, almost all matching tasks have been studied for decades, and remain to be important research topics.
  - DeepMatcher and Ditto for entity matching, Sato for column name inference, TURL for entity linking, etc.
  - Current DL-based solutions are task-specific or even dataset-specific

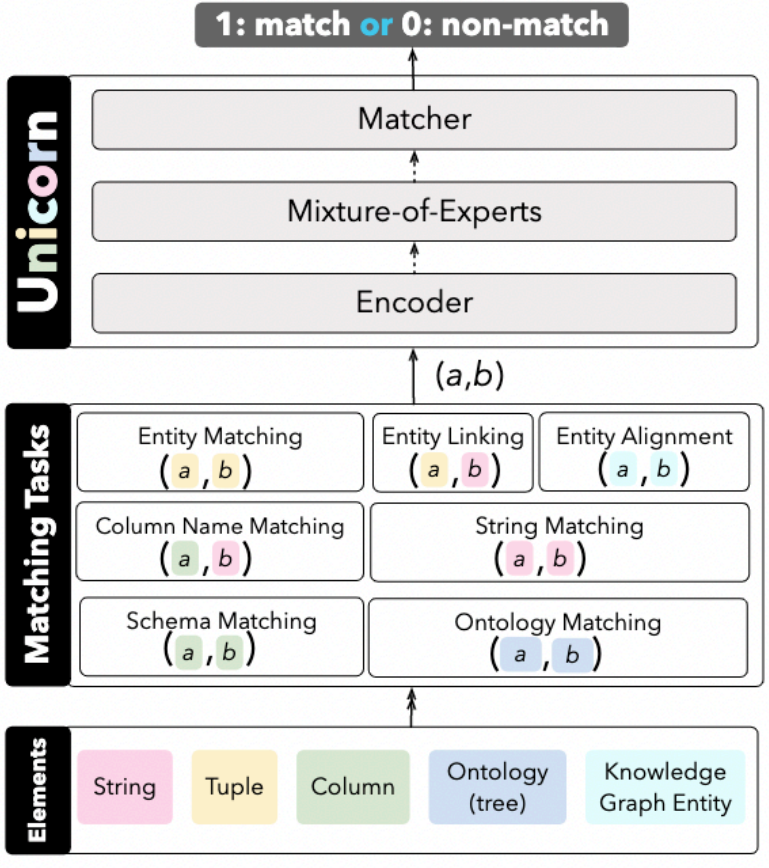
- Limitations of the specific models
  - The learned knowledge cannot be shared across different models
  - One model has to be trained (or fine-tuned) for each task or dataset, which is inefficient and has a high monetary cost

Can we build a unified model that learns from multiple tasks/datasets?

A Unified Model for Data Matching

- Can we develop a unified model to serve a variety of data matching tasks
- The model should enable knowledge sharing across multiple data matching, which may outperform specific models

- Building such a unified model is hard
  - Multiple Modalities: data items have different data format
  - Multiple tasks: tasks have different data matching semantics
  - Labeled datasets: a lot of labels for each task may be required



Can we use Foundation Models?

- Foundation Models, e.g., GPT-3, are models trained on broad data that can be adapted to many downstream tasks
- We can cast data integration/cleaning tasks as prompting tasks
- Example: Entity Matching

Dataset	Magellan	Ditto	GPT3-175B (k=0)	GPT3-175B (k=10)
Fodors-Zagats	100	100	87.2	**100**
Beer	78.8	94.37	78.6	**100**
iTunes-Amazon	91.2	97.06	65.9	**98.2**
Walmart-Amazon	71.9	86.76	60.6	**87.0**
DBLP-ACM	98.4	**98.99**	93.5	96.6
DBLP-Google	92.3	**95.60**	64.6	83.8
Amazon-Google	49.1	**75.58**	54.3	63.5

Table 1: Entity matching results measured by F1 score where k is the number of task demonstrations.

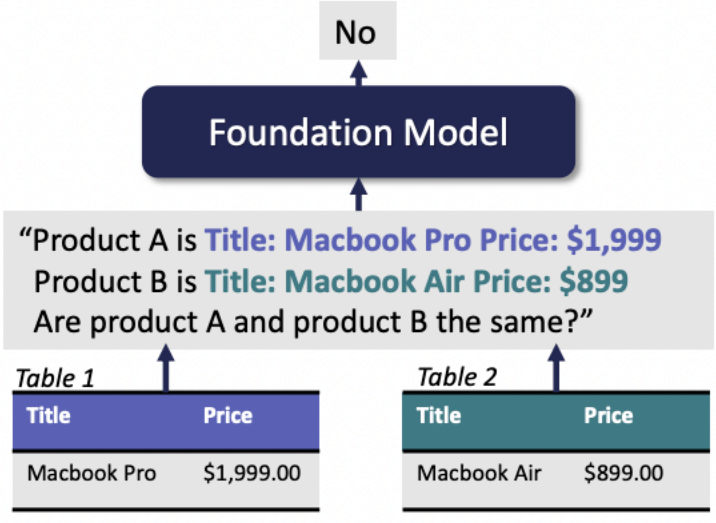


Figure 1: A large FM can address an entity matching task using prompting. Rows are serialized into text and passed to the FM with the question "Are products A and B the same?". The FM then generates a string "Yes" or "No" as the answer.

A. Narayan, I. Chami, L. J. Orr, C. Ré: Can Foundation Models Wrangle Your Data? CoRR abs/2205.09911 (2022).

Takeaways

- Data Integration tasks, such as schema matching, entity resolution, have been studied for decades, and remain to be important research topics
- DL-based methods with pretrained models achieve the state-of-the-art (SOTA) results, but they need a large amount of labeled training data
- Domain adaptation works for ER for reusing existing labels
- It calls a unified model (or foundation models) that learns from multiple tasks and multiple datasets

Talk Outline

- An Overview of Data Prep for AI
- Weak-Supervision for Data Labeling
- Pre-trained Models for Data Integration
- **ML-Oriented Data Cleaning**
- Model-Aware Data Discovery
- Summary and Future Directions

Data Cleaning

- The DB community has been focusing on the topic for decades
- Types of Dirty Data
  - Inconsistency
  - Duplicates
  - Outlier
  - Missing Values
  - Mislabels
  - ……

Question:

Do we need to clean all types of dirty data for ML?

No!!!

Impact of Data Cleaning on ML

- A case study on how data cleaning affects model training
  - **Positive**: whether cleaning certain error has positive effect for ML.
  - **Negative**: whether cleaning certain error has negative effect for ML.

Clean\ Impact	Positive	Negative
Inconsistency	—	—
Duplicates	—	✓
Outlier	✓	✓
Missing value	✓	—
Mislabels	✓	—

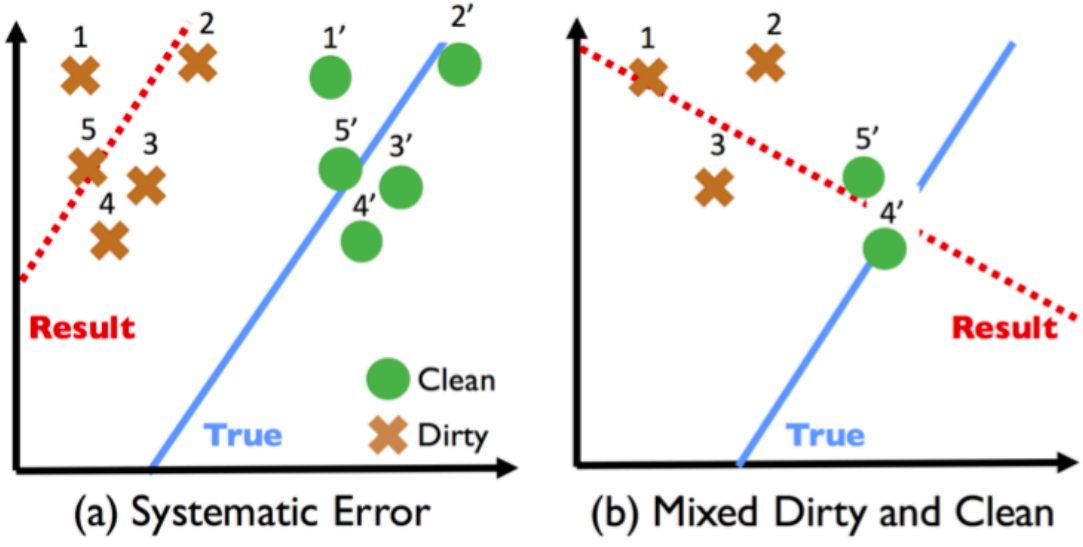
Takeaways:

- We may need to shift our mindset from model-agnostic data cleaning to model-aware data cleaning
- From pursuing ground-truth to improving ML performance

P. Li, X. Rao, J. Blase, Y. Zhang, X. Chu, C. Zhang: CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks. ICDE 2021.

Case Study: Active Data Cleaning

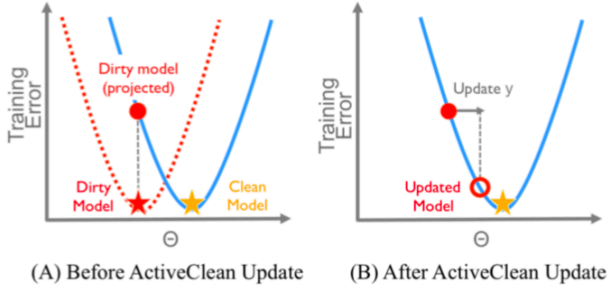
- ActiveClean: Prioritizing training examples for data cleaning
  - Given a dirty dataset and model, which data records are the most beneficial if we want to obtain a well-performed model?



Sanjay Krishnan, Jiannan Wang, Eugene Wu, Michael J. Franklin, Ken Goldberg: ActiveClean: Interactive Data Cleaning For Statistical Modeling. Proc. VLDB Endow. 9(12): 948-959 (2016)

Case Study: Active Data Cleaning

- ActiveClean
  - Do we need to clean all the data to obtain a good model?



(A) Before ActiveClean Update (B) After ActiveClean Update

Key Point:

Compute the Gradient Correctly!!!

  - Ideal scenario: Compute gradient on totally cleaned data.

$$g^*(\theta) = \nabla\phi(\theta) = \frac{1}{N}\sum_{i=1}^{N}\nabla\phi(x_i^{(c)}, y_i^{(c)}, \theta)$$

  - Sample a subset of dirty data, and estimate

$$g(\theta) = \frac{|R_{clean}|}{|R|}\cdot g_C(\theta) + \frac{|R_{dirty}|}{|R|}\cdot g_S(\theta)$$

Sanjay Krishnan, Jiannan Wang, Eugene Wu, Michael J. Franklin, Ken Goldberg: ActiveClean: Interactive Data Cleaning For Statistical Modeling. Proc. VLDB Endow. 9(12): 948-959 (2016)

BoostClean

- BoostClean
  - Input: a set of detectors and repair rules.
  - Output: an optimal sequence of repair operations.
  - Strategy: ensemble learning
    - View each repair operation as a weak classifier.
    - Generate the best current classifier.
    - Weight the dataset based on the mis-predictions

```
def repair(r, r_orig):
  if r.region in ('USWest', 'USWESTERN'):
    r.region = 'USW'
  return r
```

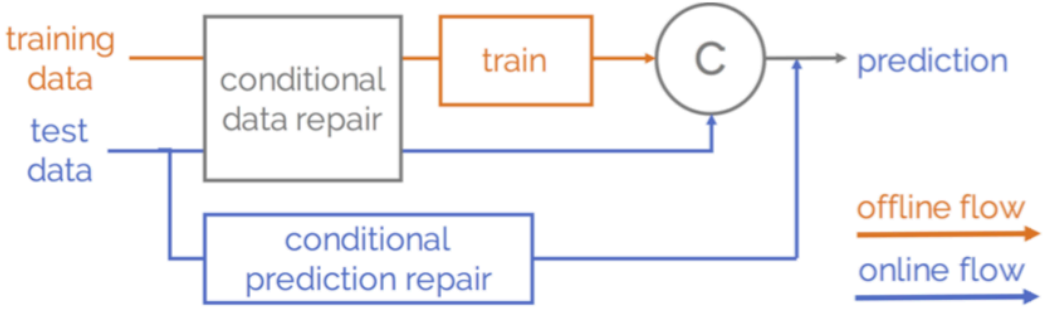
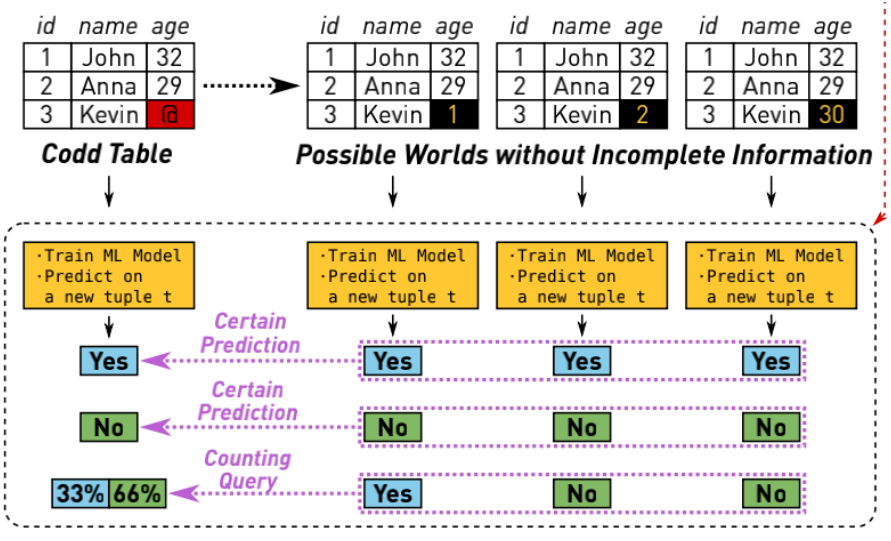


Figure 2: Offline (orange) and online (blue) workflows.

S. Krishnan, M. J. Franklin, K. Goldberg, E. Wu: BoostClean: Automated Error Detection and Repair for Machine Learning. CoRR abs/1711.01299 (2017)

CPClean

- CPClean: missing value imputation for training ML models
- Observations: Different possible worlds for missing value imputations may lead to consistent result for test examples
- Key ideas: Prioritize imputing the records that are likely to make test samples have consistent labels
- Challenges
  - Restrict to KNN classifier
  - Leverage the tuple similarities.
  - intractable training time



B. Karlas, P. Li, R. Wu, N. M. Gürel, X. Chu, W. Wu, C. Zhang: Nearest Neighbor Classifiers over Incomplete Information: From Certain Answers to Certain Predictions. Proc. VLDB Endow. 14(3): 255-267 (2020)

Adaptive Data Cleaning for Dataset Shift

- Adaptive Data Cleaning
  - Traditional Data Cleaning cares about recovering ground-truth
  - Data Cleaning for ML also needs to consider **dataset shift**!

(a) A labeled source data ($\mathcal{D}_s$) — *labels*

	age	chol	gluc	smoke	alcohol	cardio
s_1	30	2	3	0	0	no
s_2	35	2	1	1	0	no
s_3	50	3	3	NA	1	yes
s_4	65	2	3	1	NA	yes
s_5	70	3	1	1	0	no

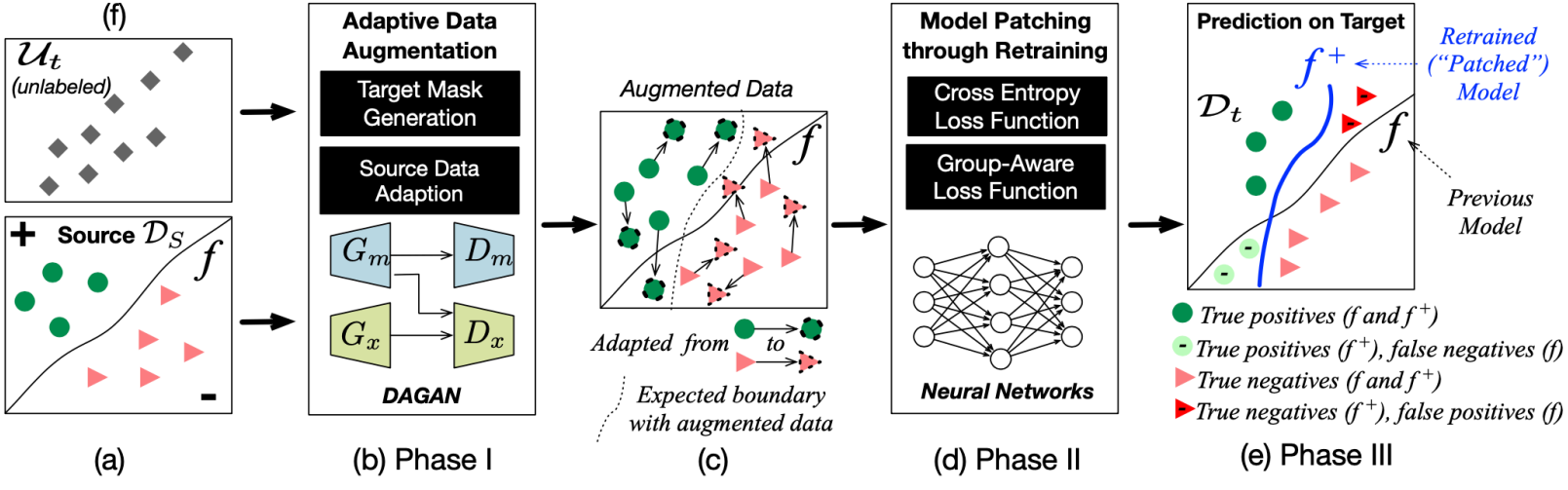
(b) An unlabeled target data ($\mathcal{D}_t$) — *no labels*

	age	chol	gluc	smoke	alcohol	cardio
t_1	25	1	NA	0	1	?
t_2	37	NA	3	0	0	?
t_3	40	3	NA	1	1	?
t_4	72	3	2	0	0	?

- Dataset shift in ML:
  - ML models are difficult to maintain in production environments
  - One common challenge is the noise shift between source (train) and target (test) datasets

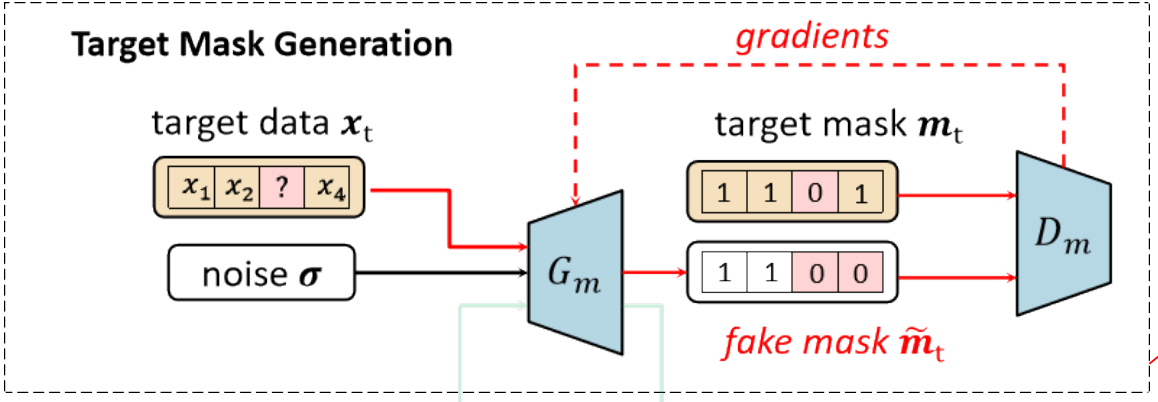
Adaptive Data Cleaning for Dataset Shift

- Adaptive Data Cleaning
  - Data Cleaning for ML also needs to consider **dataset shift**!
  - One solution: Adaptive Data Augmentation using GAN



T. Liu, J. Fan, Y. Luo, N. Tang, G. Li, and X. Du. Adaptive Data Augmentation for Supervised Learning over Missing Data. **PVLDB 2021**.

Target Mask Generation



We utilize Conditional GAN to learn the p(mask|observed data)

$$\mathcal{L}_m(D_m, G_m) = \mathbb{E}_{(x_t, m_t)\sim\mathcal{U}_t}[D_m(m_t, x_t)] - \mathbb{E}_{\sigma\sim p(\sigma),(x_t,m_t)\sim\mathcal{U}_t}[D_m(G_m(\sigma, x_t), x_t)]$$

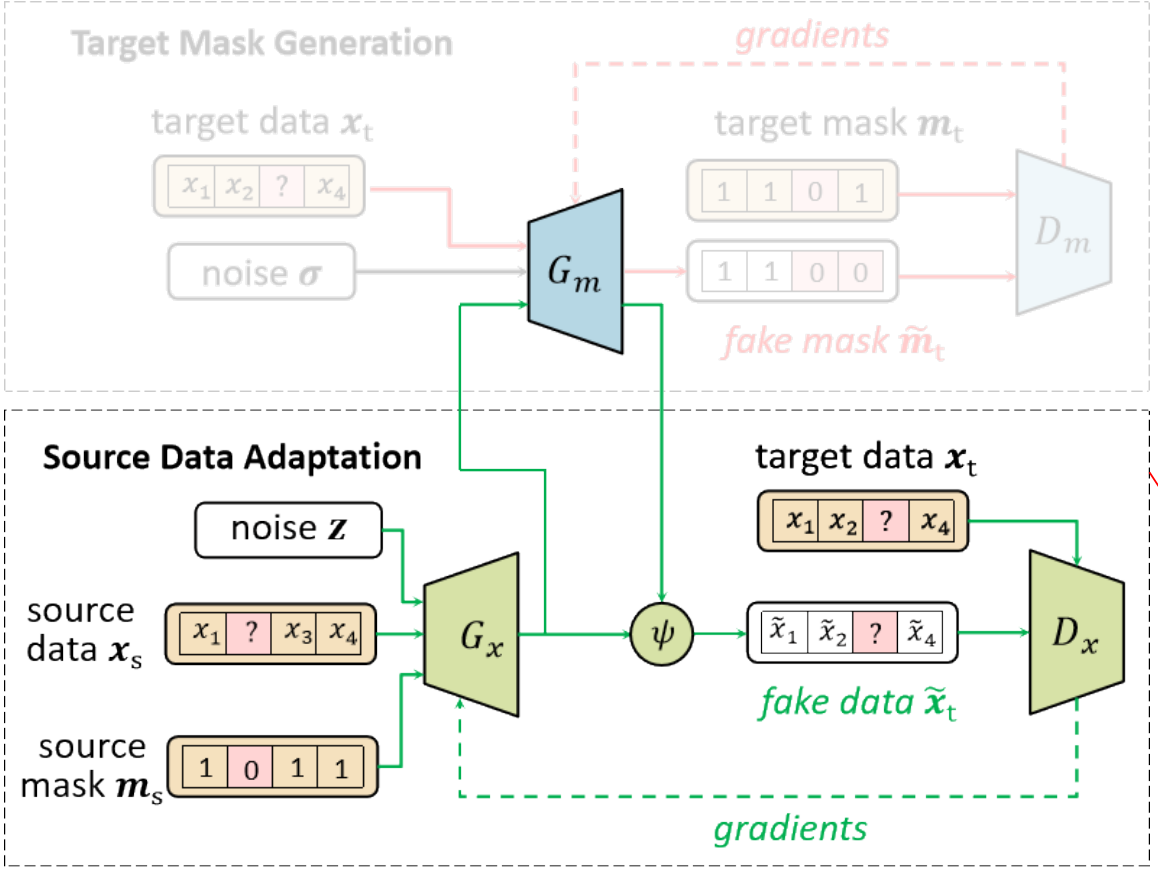
Target Data

age	cholesterol	glucose	smoke	alcohol
25	1	NA	0	1
37	NA	3	0	0
70	3	2	1	1

Mask Matrix

age	cholesterol	glucose	smoke	alcohol
1	1	0	1	1
1	0	1	1	1
1	1	1	1	1

Source Data Adaptation



Adapted Source Data

age	cholesterol	glucose	smoke	alcohol
35	2	NA	1	0
50	3	3	0	0
65	2	3	1	1

Be "Similar"

age	cholesterol	glucose	smoke	alcohol
25	1	NA	0	1
37	NA	3	0	0
70	3	2	1	1

Target Data

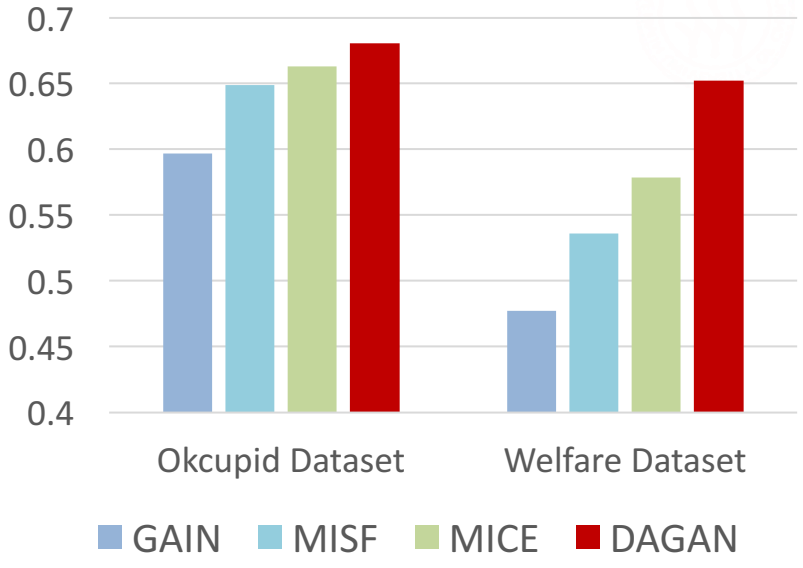
We then use another GAN to learn the p(target data)

$$\mathcal{L}_x(D_x, G_x, G_m) = \mathbb{E}_{(x_t, m_t)\sim\mathcal{U}_t}[D_x(x_t)] - \mathbb{E}_{\sigma\sim p(\sigma), z\sim p(z), (x_s, m_s)\sim\mathcal{D}_s}[D_x(\psi(\tilde{x}, G_m(\sigma, \tilde{x})))], \tilde{x} = G_x(z, x_s, m_s)$$

Adaptive Data Cleaning for Dataset Shift

- Adaptive Data Cleaning

Missing Rate	MCAR		MAR		MNAR	
	AdaSrc	ClnSrc	AdaSrc	ClnSrc	AdaSrc	ClnSrc
0.2	0.615	**0.616**	**0.621**	0.610	**0.664**	0.593
0.3	**0.629**	0.569	**0.621**	0.563	**0.623**	0.578
0.4	**0.632**	0.581	**0.643**	0.585	**0.617**	0.593
0.5	**0.623**	0.539	**0.603**	0.590	**0.576**	0.546
0.6	0.579	**0.582**	0.616	**0.645**	**0.600**	0.592
0.7	**0.589**	0.464	**0.593**	0.512	**0.595**	0.492
0.8	**0.596**	0.471	**0.614**	0.561	**0.588**	0.570



Takeaways:

- Only cleaning training set cannot achieve satisfactory results because the train and test data may have divergence on data distribution
- Adaptive data cleaning outperforms SOTA when handling noise shift

T. Liu, J. Fan, Y. Luo, N. Tang, G. Li, and X. Du. Adaptive Data Augmentation for Supervised Learning over Missing Data. **PVLDB 2021**.

Comparisons of ML-Oriented Data Cleaning

Method	Target	Method	Automatic	Cleaning Type	Model Type
ActiveClean	Model ↑	Gradient Estimation	×	Outlier, Normalization	Loss convex model
BoostClean	Model ↑	Boosting	✓	Outlier	All
CPClean	Model ↑	Incompletion analysis	✓	Missing values	KNN
DAGAN	Model ↑	GAN	✓	Missing values	All

Can we have a Unified Framework?

- There are many small tasks in data cleaning
- Can we develop relation-aware pre-trained transformer-based models, so as to solve all the data cleaning tasks?

Q1: *r1*[name, expertise, city] = (Michael Jordan, Machine Learning, **[M]**)
A1: Berkeley

Q2: *r3*[name, affiliation] = (Michael **[M]**, CSAIL MIT)
A2: Cafarella

Q3: *r2*[name, expertise, **[M]**] = (Michael Jordan, Basketball, New York City)
A3: city

(a) Sample Tasks for Value Filling ([M]: value to fill)

	product	company	year	memory	screen
e1	iPhone 10	Apple	2017	64GB	5.8 inchs
e2	iPhone X	Apple Inc	2017	256GB	5.8-inch
e3	iPhone 11	AAPL	2019	128GB	6.1 inches

(b) A Sample Entity Resolution Task

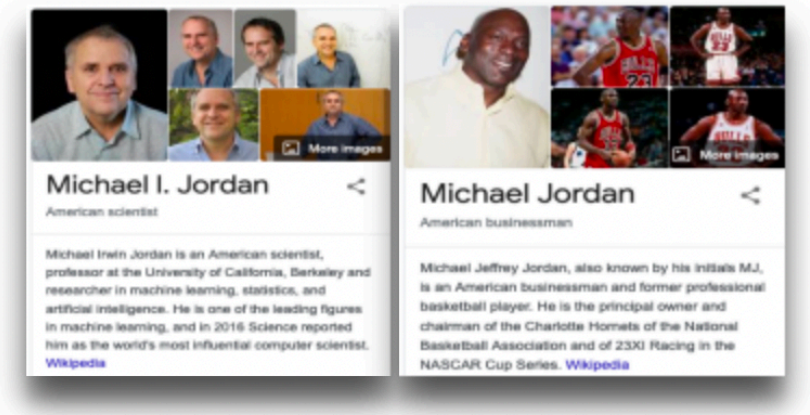
	type	description	label
s1	notebook	2.3GHz 8-Core, 1TB Storage, 8GB memory, 16-inch Retina display	8GB
t1	phone	6.10-inch touchscreen, a resolution of 828x1792 pixels, A14 Bionic processor, and come with 4GB of RAM	4GB

(c) A Sample Information Extraction Task (*s1*: example, *t1*: task)

N. Tang, **J. Fan**, F. Li, J. Tu, X. Du, G. Li, S, Madden, M. Ouzzani. RPT: Relational Pre-trained Transformer Is Almost All You Need towards Democratizing Data Preparation. **PVLDB 2021**, Vision

Existing Pre-trained LM

- **Sequence semantics**
  - Michael Jordan is good at ___ **(Basketball)**
  - Professor Michael Jordan at UC Berkeley is good at ___ **(Machine)**



Data: text corpora
Task: next token

We need: relation-aware model

Name	Expertise	City
Michael Jordan	Machine Learning	Berkeley
Michael Jordan	Basketball	New York City

- **<Name, Value> pairs**

<Name, Michael Jordan>, <Expertise,Machine Learning>, <City, Berkeley>

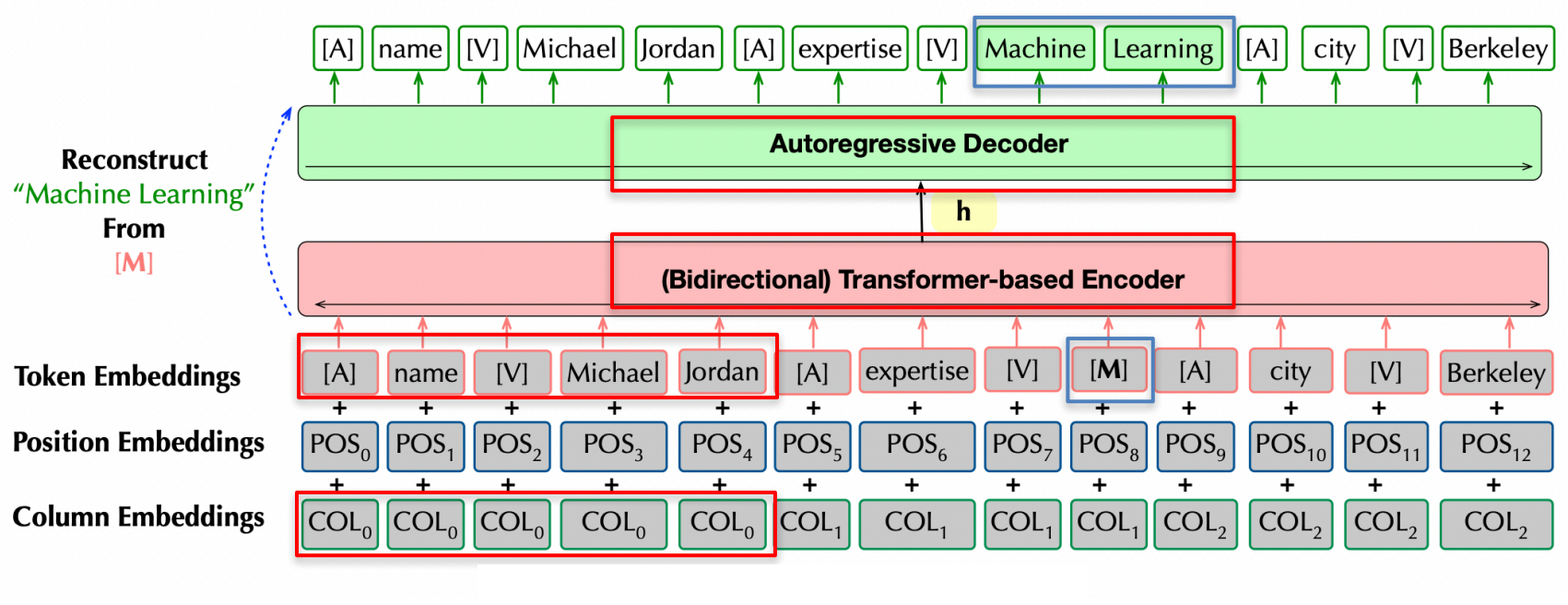
- **Set semantics** =

<Expertise,Machine Learning>, <Name, Michael Jordan>, <City, Berkeley>

Data: table corpora
Task: guessing a missing attribute name/value

RPT: Relational Pre-trained Transformer

- Transformer-based sequence-to-sequence architecture
  - BART: A generation of both **BERT** (bidirectional encoder) and **GPT-3** (left-to-right autoregressive decoder)



RPT: Preliminary Results

- RPT Settings:
  - Train/Valid datasets: Abt-Buy (3 attri, 2137 lines) and Walmart-Amazon (5 attri, 24627 lines).
  - Test dataset: Amazon-Google (3 attri, 4589 lines)
  - Pre-training: learning rate=3e-5, max-sequence-length=2048, epoch=1.
- Baseline: Bart (pre-trained with a large corpus of text)
- Results: RPT is more capable of predicting missing values.

title	manufacturer	price	BART	RPT	Ground truth
instant home design (jewel case)	topics entertainment	[M]	Topics	9	9.99
disney's 1st & 2nd grade ...	disney	[M]	Dis	19	14.99
adobe after effects pro-fessional 6.5 ...	[M]	499.99	$1.99	adobe	adobe
stomp inc recover lost data 2005	[M]	39.95	39.95	stomp	stomp inc
[M]	write brothers	269.99	1.99	write brothers	write brothers dramatica ...

Takeaways

- Data cleaning aims to detect/repair data errors
- Not all cleaning operations can directly improve ML models, while some may even have negative impacts
- Cleaning directly for ML always focuses on looking into the ML models and finding the impact of data errors on ML models
- Relational pre-trained transformers, with pre-training/fine-tuning, is promising to (semi)automate data cleaning tasks

Talk Outline

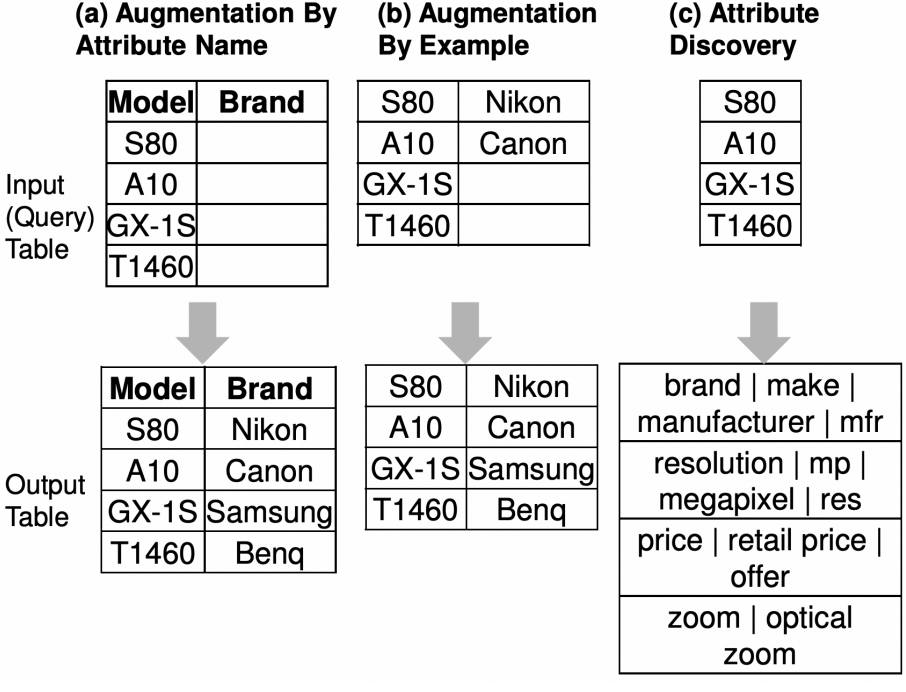
- An Overview of Data Prep for AI
- Weak-Supervision for Data Labeling
- Pre-trained Models for Data Integration
- ML-Oriented Data Cleaning
- **Model-Aware Data Discovery**
- Summary and Future Directions

Categorization of Data Discovery

Category	Granularity	Method	Sources	Goal
Model Agnostic	Attribute/Tuple Level	Infogather	Web Tables	Attribute/Value Filling
		SmartCrawl	Hidden Web	Attribute/Value Filling
		Crowd-Based	Human, KB	Attribute/Value Filling Tuple Collection
	Table Level	Goods	Data Lake	Table Search
		Aurum	Data Lake	Table Search
Model Aware	Tuple Level	AutoData	Data Lake	Tuple Enrichment
	Table Level	Halmet	Data Warehouse	Feature Augmentation
		Halmet+	Data Warehouse	Feature Augmentation
		ARDA	Data Lake	Feature Augmentation

Model-Agnostic Data Discovery

- **Attribute/Tuple Level Discovery**: Given a table, it finds missing attribute values or entire tuples from external data sources
  - Example Discovery Tasks



  - Finding from Web tables[1]
  - Finding from Hidden Web[2]
  - Finding via Crowdsourcing[3][4]

[1] M. Yakout, K. Ganjam, K. Chakrabarti, S. Chaudhuri: InfoGather: entity augmentation and attribute discovery by holistic matching with web tables. SIGMOD Conference 2012: 97-108

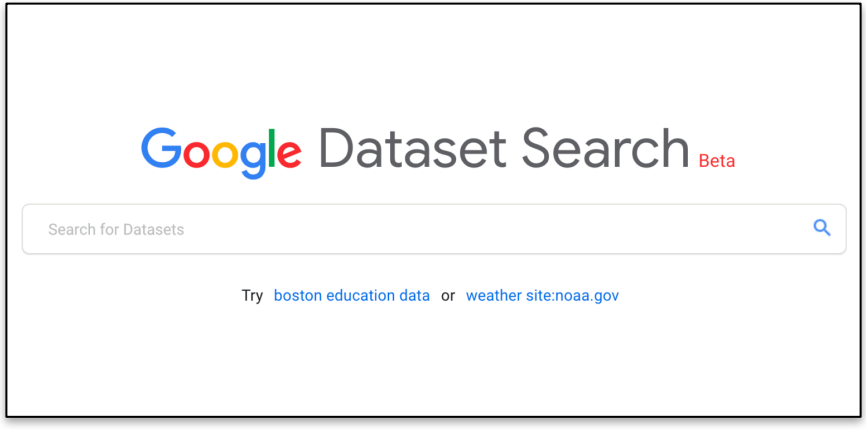
[2] P. Wang, R. Shea, J. Wang, E. Wu: Progressive Deep Web Crawling Through Keyword Queries For Data Enrichment. SIGMOD Conference 2019: 229-246

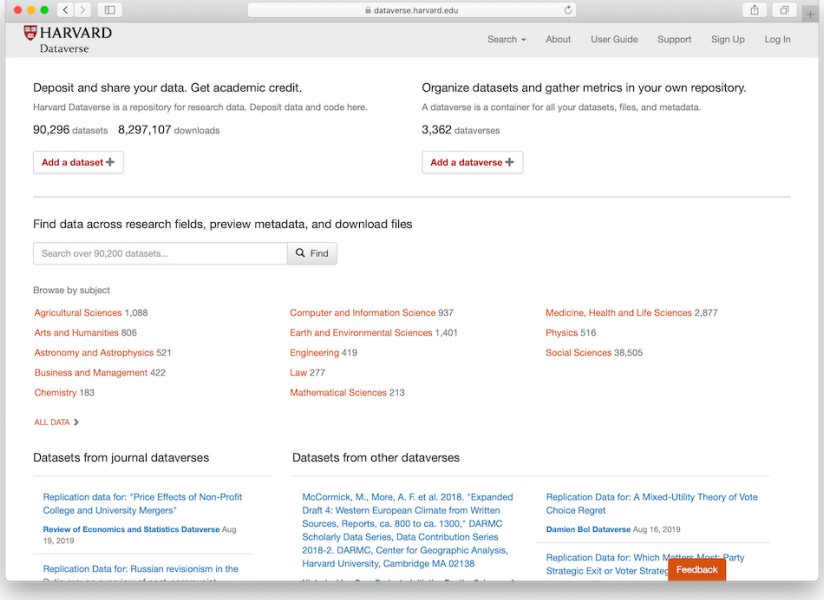
[3] H. Park, J. Widom: CrowdFill: collecting structured data from the crowd. SIGMOD Conference 2014: 577-588

[4] C. Chai, J. Fan, G. Li: Incentive-Based Entity Collection Using Crowdsourcing. ICDE 2018: 341-352

Model-Agnostic Data Discovery

- A number of efforts have been made for table-level discovery
  - Google Dataset Search[1], Socrata data platform[2]
  - Dataverse at Harvard University[3]
  - See details in a recent survey[4]

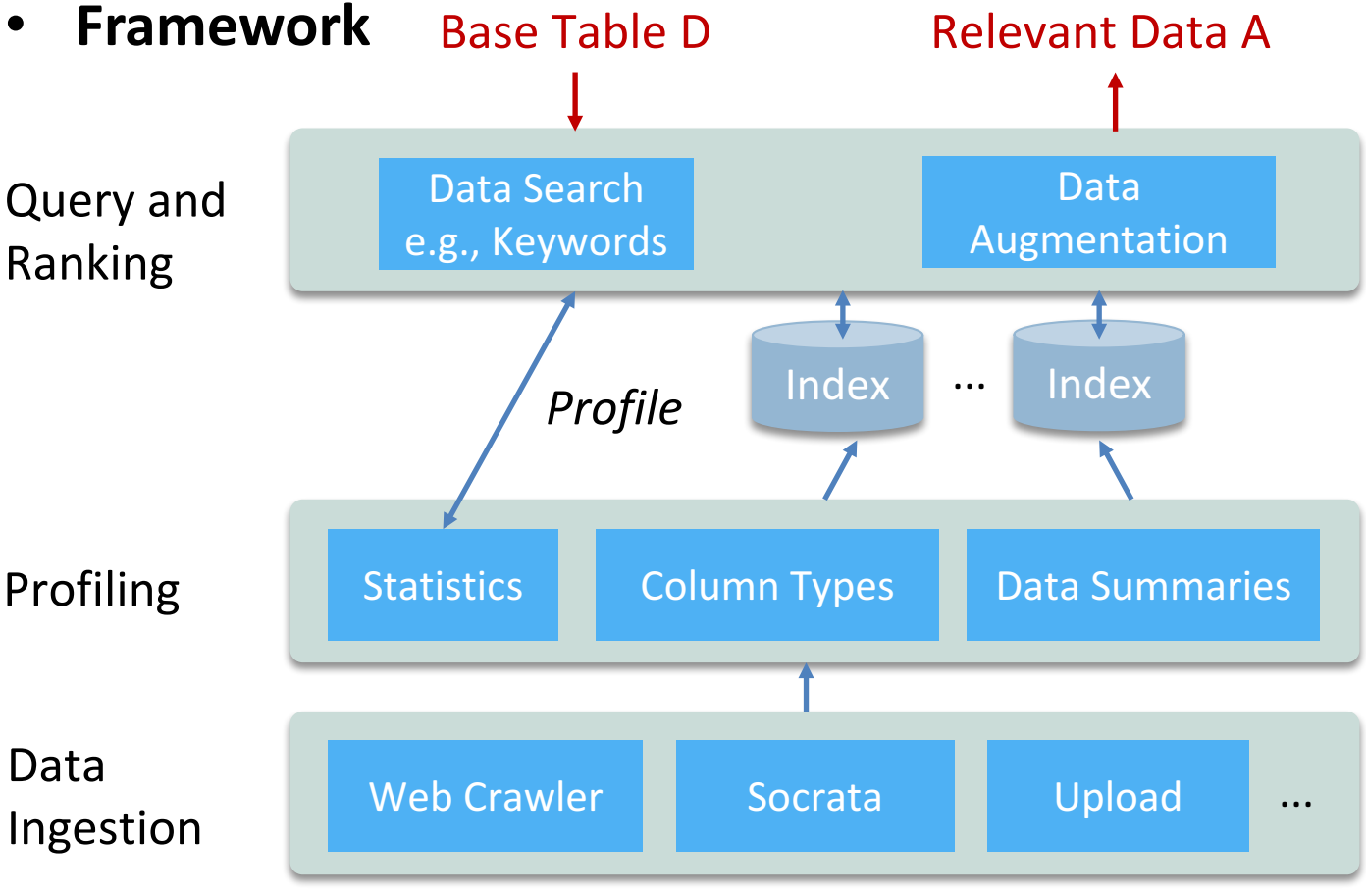




[1] https://datasetsearch.research.google.com
[2] https://www.tylertech.com/products/socrata
[3] https://data.harvard.edu/dataverse
[4] Adriane Chapman, Elena Simperl, Laura Koesten, George Konstantinidis, Luis Daniel Ibáñez-Gonzalez, Emilia Kacprzak, and Paul T. Groth. 2019. Dataset Search: A Survey. CoRR abs/1901.00735 (2019).

Model-Agnostic Data Discovery

- **Framework**

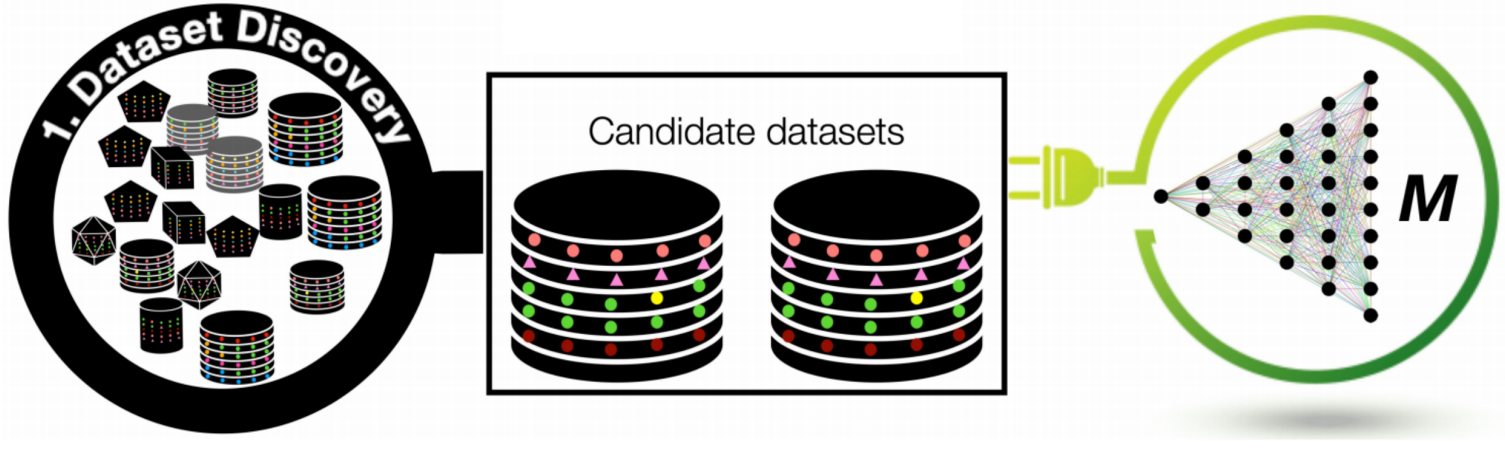


[1] A. Y. Halevy, F. Korn, N. F. Noy, C. Olston, N. Polyzotis, S. Roy, S. E. Whang: Goods: Organizing Google's Datasets. SIGMOD Conference 2016: 795-806

[2] R. C. Fernandez, Z. Abedjan, F. Koko, G. Yuan, S. Madden, M. Stonebraker: Aurum: A Data Discovery System. ICDE 2018: 1001-1012

Model-Aware Data Discovery

- **Goal:** Enriching the tuples or attributes (i.e., features) of the training data, to improve accuracy of the trained ML models
- Categorization
  - Tuple Enrichment: Finding representative tuples
  - Feature Augmentation: Augmenting relevant features



Model-Aware Tuple Enrichment

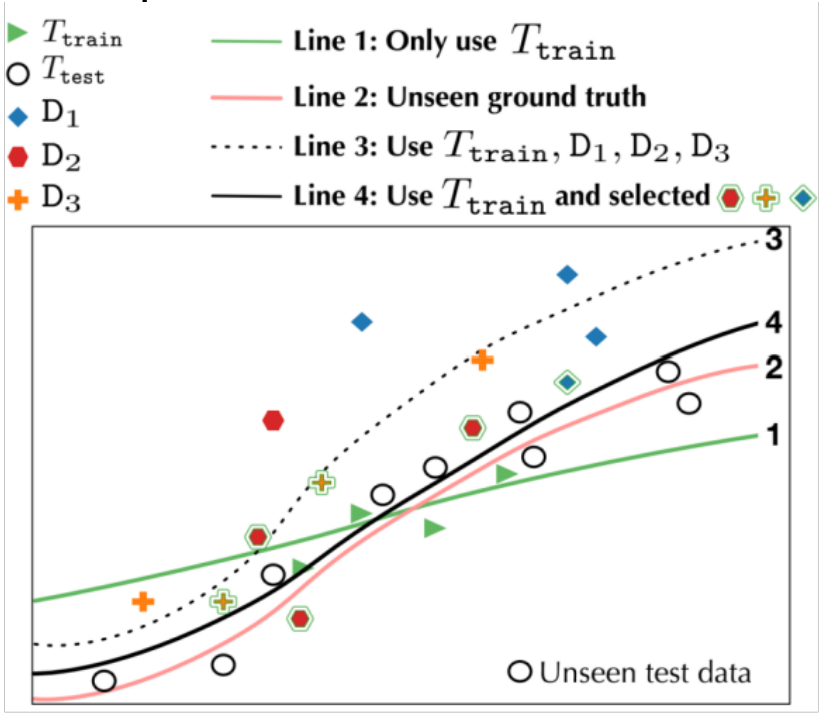
- Insufficient training data
  - E.g., using insufficient data to learn how to predict “Price”

City	Year	Area	Security	Price
Kolkata	2009	710	No	3,200,000
Kolkata	2013	770	No	3,850,000
Kolkata	2007	935	No	2,524,000
Kolkata	2006	973	Yes	3,611,000

(a) A sample training dataset

City	Year	Area	Security	Price	Ground Truth
Kolkata	2017	350	No	?	2,100,000
Kolkata	2019	465	Yes	?	4,365,000
Kolkata	2015	572	No	?	3,268,000
Kolkata	2012	655	Yes	?	2,599,000
Kolkata	2012	735	No	?	3,300,000
Kolkata	2017	881	Yes	?	4,698,000
Kolkata	2011	1123	Yes	?	3,324,000
Kolkata	2014	1210	Yes	?	5,000,000

(b) A sample test dataset



Model-Aware Tuple Enrichment

- Modeling heterogeneous datasets
  - Goal: finding useful data points → fine-grained modeling
  - Clustering → partition all data into distinct groups
  - Data pool P → clusters $\mathbf{C} = \{C_1, \ldots, C_n\}$
  - C_i has its own distribution $p_i = (\mu_i, \Sigma_i)$
- Iterative data selection
  - Select
  - Retrain & Evaluate
  - Update
  - *Repeat above process*

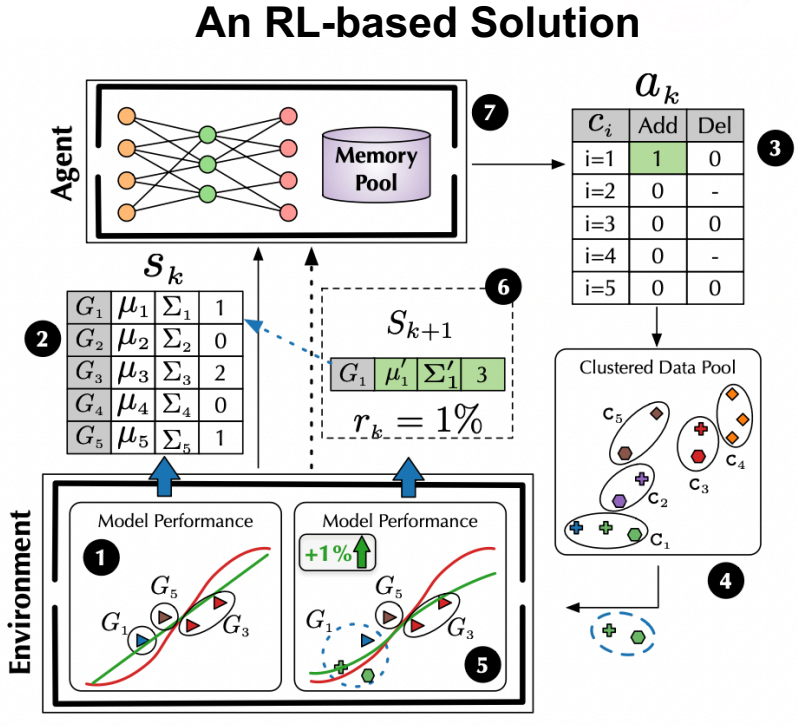


Figure 6: A running example of DQN-based RL solution.

Model-Aware Feature Augmentation

- Halmet: Discovery for ML—Join or not to Join (PKFK)

  Customers (CustomerID, Churn, Gender, Age, EmployerID)

  Employers (EmployerID, Country, Revenue)

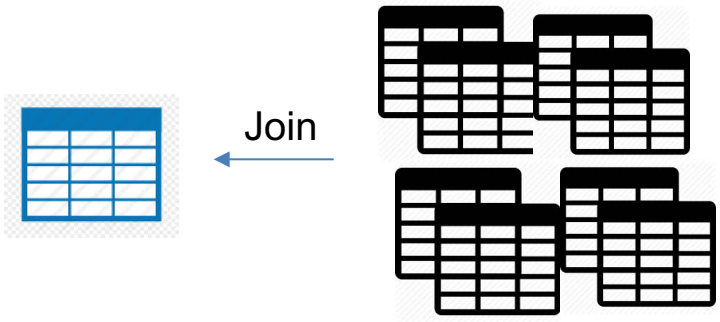
- Observations:
  - The foreign key has already encoded much information.
- Leverage ML Theory:

$$ROR = \frac{\sqrt{v_{Yes} log(\frac{2en}{v_{Yes}})} - \sqrt{v_{No} log(\frac{2en}{v_{No}})}}{\delta\sqrt{2n}} + \Delta bias$$

A. Kumar, J. F. Naughton, J. M. Patel, X. Zhu: To Join or Not to Join?: Thinking Twice about Joins before Feature Selection. SIGMOD Conference 2016: 19-34

Model-Aware Feature Augmentation

- ARDA: Automatically finding fuzzy tables to Join



- Brute-forcing approach: Joining the base table with all related tables, and then conducting feature selection.

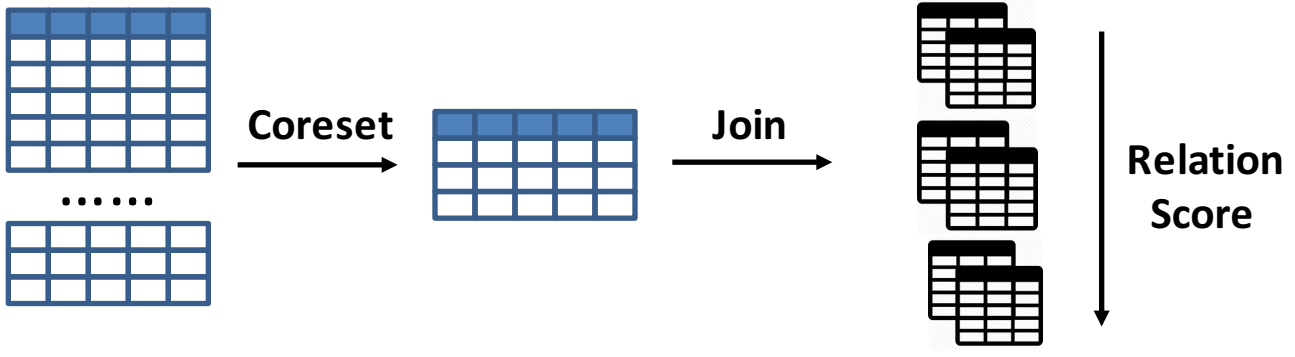
Join is always time-consuming

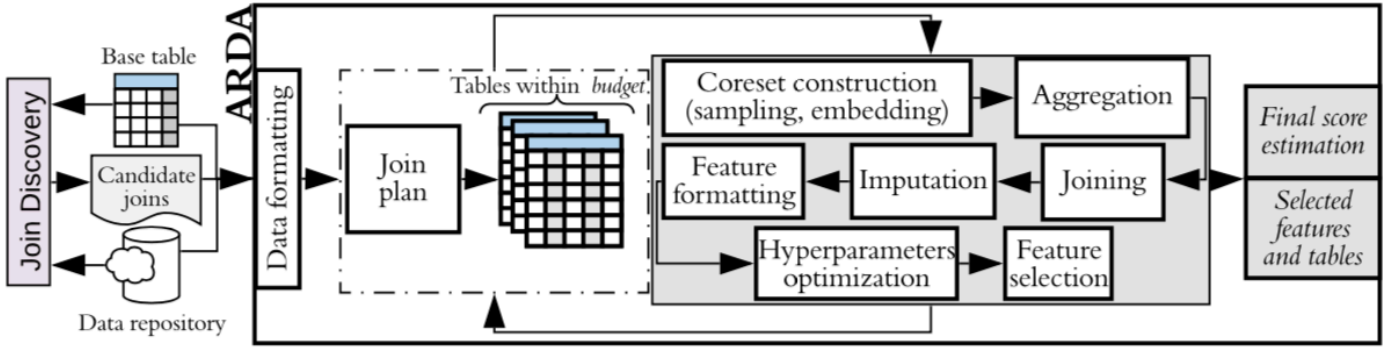
- Improving the efficiency by
  - Reducing the size of base table
  - Joining with fewer tables

N. Chepurko, R. Marcus, E. Zgraggen, R. C. Fernandez, T. Kraska, D. R. Karger: ARDA: Automatic Relational Data Augmentation for Machine Learning. Proc. VLDB Endow. 13(9): 1373-1387 (2020)

Model-Aware Feature Augmentation

- Coreset Selection: selecting a subset of records to join
- Joining the most related table first, then conducting feature selection and repeating the above process





Takeaways

- Data Profiling
  - Data lake is much less organized compared to data warehouse
  - A number of methods for type detection, domain discovery, semantic links identification, dataset embedding, etc.
- Query and Ranking
  - Effectiveness: many scoring functions are proposed
  - Efficiency: Coreset selection for Join
- Search Interface
  - Supporting search capabilities in Jupyter notebooks

Many Research Problems are Still Open!

Talk Outline

- An Overview of Data Prep for AI
- Weak-Supervision for Data Labeling
- Pre-trained Models for Data Integration
- ML-Oriented Data Cleaning
- Model-Aware Data Discovery
- **Summary and Future Directions**

Summary of Takeaways

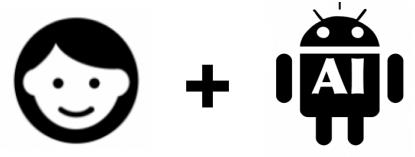
- From model-centric to data-centric

Call for more research

AI System = Data + Code
(or Data Science) (model/algorithm)

Summary of Takeaways

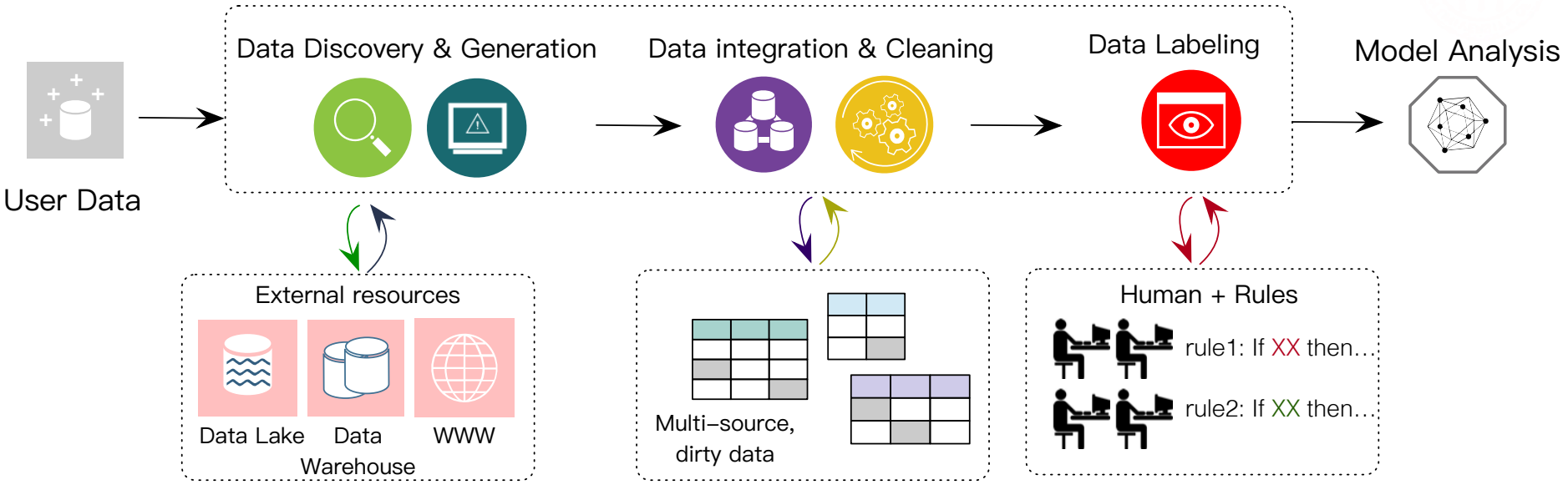
- It calls for arms to work on..
  - Weak-Supervision for Data Labeling
  - Pre-trained Models for Data Integration
  - ML-Oriented Data Cleaning
  - Model-Aware Data Discovery
  - ……
- Design Principles: Human + AI
  - Human-in-the-loop yet few-shot
  - Encoded & transferable domain knowledge
  - Piggybacking on advanced DL models
  - ……



Future Directions

- Benchmarks for Data Prep
  - Can we have “TPC” or “ImageNet” for Data Prep?
- Deep learning (DL) for Data Prep
  - Can we develop DL models to further reduce human cost?
- Data Prep for Deep Learning (DL)
  - Can we develop better Data Prep to improve performance of DL?

The Journey Just Starts…



Thanks