SFU

Automating Entity Matching Model Development

Jiannan Wang

Simon Fraser University

April 14, 2021, Thomson Reuters

Pei Wang, Weiling Zheng, Jiannan Wang, Jian Pei. Automating Entity Matching Model Development. ICDE 2021, Chania, Greece.

SFU in the cloud ☺

SFU Data Science Research Group

http://data.cs.sfu.ca

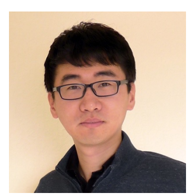
- Invented many famous data mining algorithms (e.g., **FP-Growth, DBScan**)
- **Research Strengh:** Cloud Databases, Data Preparation, Data Pricing, Data Security and Privacy, Recommender Systems
- **Ranked 13th** in databases and data mining in North America (source: csrankings.org)

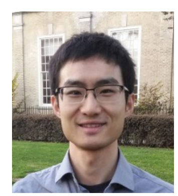
#	Institution	Count	Faculty
1	Carnegie Mellon University	17.7	33
2	Univ. of Illinois at Urbana-Champaign	14.9	11
3	Stanford University	13.0	15
4	Georgia Institute of Technology	11.5	23
4	University of Michigan	11.5	14
6	Massachusetts Institute of Technology	10.3	18
7	Cornell University	10.2	24
8	Purdue University	8.8	13
9	Pennsylvania State University	8.7	8
10	University of California - Los Angeles	8.6	10
10	University of Massachusetts Amherst	8.6	16
12	University of Illinois at Chicago	8.2	7
13	Simon Fraser University	8.1	7
13	University of Maryland - College Park	8.1	11
15	University of Waterloo	7.9	20
16	Duke University	7.6	8
16	University of California - Santa Barbara	7.6	8
18	University of California - Santa Cruz	7.5	12
19	University of Wisconsin - Madison	7.4	13
20	Ohio State University	7.2	11
21	University of California - Riverside	7.0	10
22	University of California - San Diego	6.7	14
23	University at Buffalo	6.4	12


Ke Wang (Joined in 2000)


Martin Ester (Joined in 2001)


Jian Pei (Joined in 2004)


Jiannan Wang (Joined in 2016)

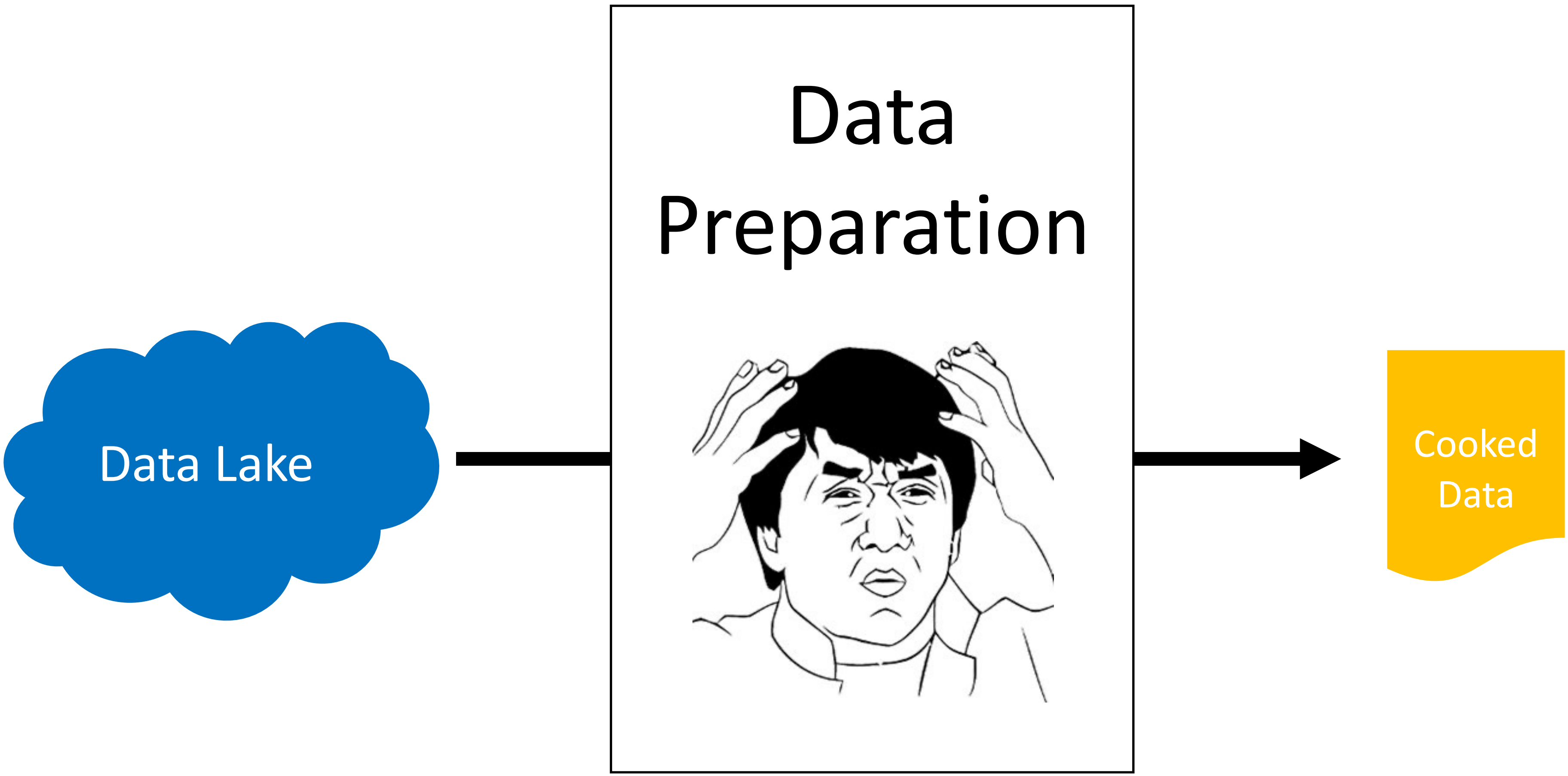

Tianzheng Wang (Joined Fall 2018)

Democratizing AI

- Computing
- Algorithms
- Data

Data Prep is the bottleneck

What is Data Prep?



Why is Data Prep hard?

Data Lake

- Data Discovery
- Data Profiling
- Data Extraction
- Data Normalization
- Data Enrichment
- Data Transformation
- Data Filtering
- Data Provenance
- Data Labeling
- Error Detection
- Schema Matching
- Deduplication
- Outlier Detection
- Imputation
- …

Cooked Data

Two Promising Directions

1. Using advanced ML technologies

- Automated Machine Learning (AutoML)
- Active Learning and Self-training

Today's Talk

2. Building open-source software

- Ease of Use
- Fast
- All-in-one

dataprep

http://dataprep.ai

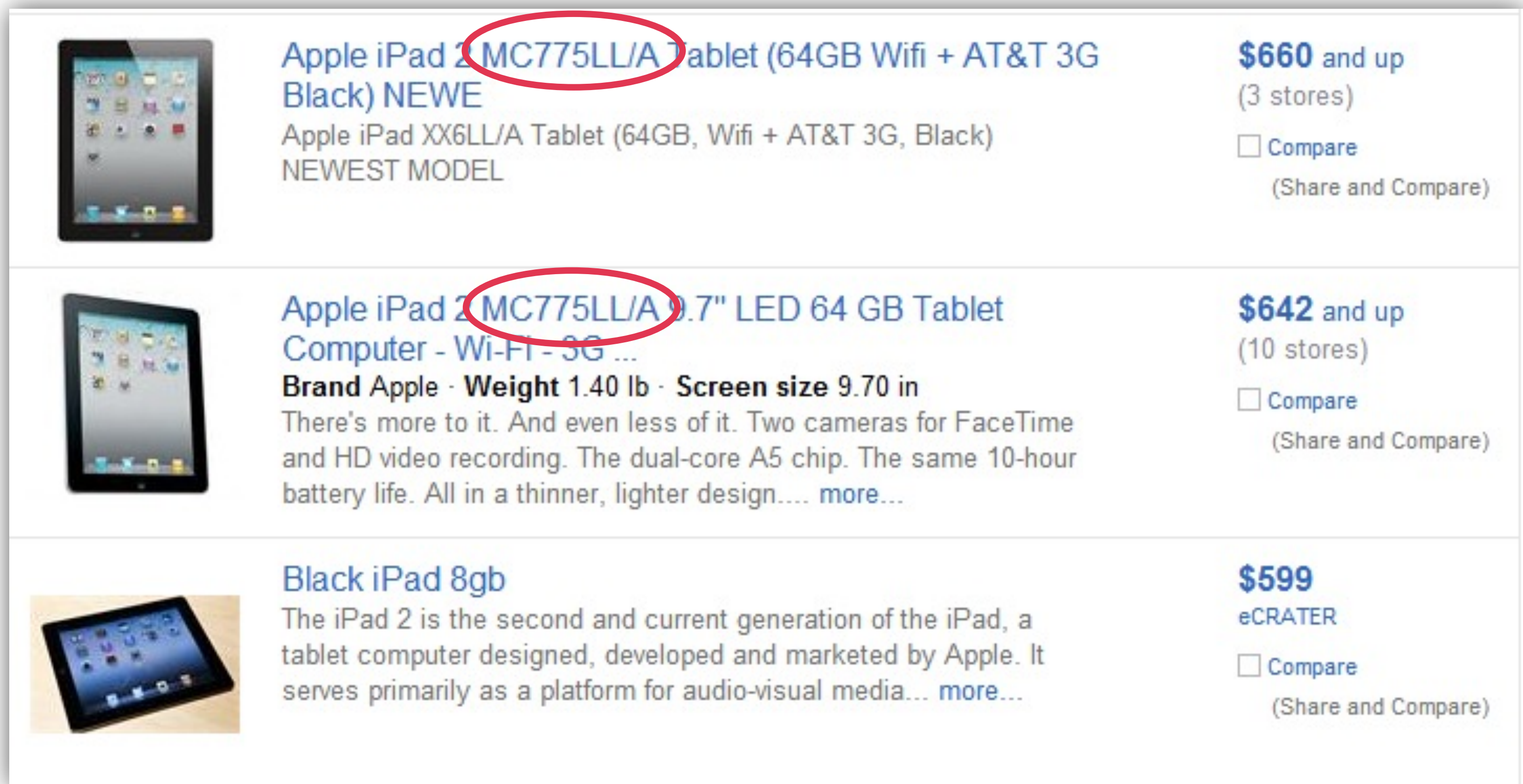
Next Wed's Talk

Talk Outline

1. Entity Matching (EM)
2. Automate Model Development
3. Automate Data Labeling
4. Future Direction

Entity Matching (EM)

EM is central to data integration and cleaning



Entity Matching (EM)

ID	Product Name	Price
r_1	iPad Eight 128GB WiFi White	$490
r_2	iPad 8th generation 128GB WiFi White	$469
r_3	iPhone 10th generation White 256GB	$545
r_4	Apple iPhone 11th generation Black 256GB	$375
r_5	Apple iPhone 10 256GB White	$520

Matching Pairs: (r_1, r_2) , (r_3, r_5)

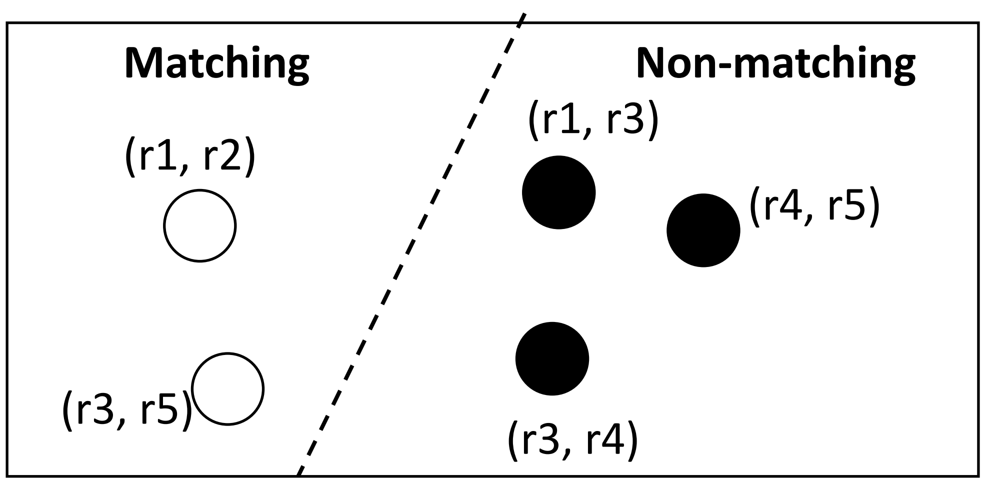
Entity Matching Techniques

1. Similarity-based

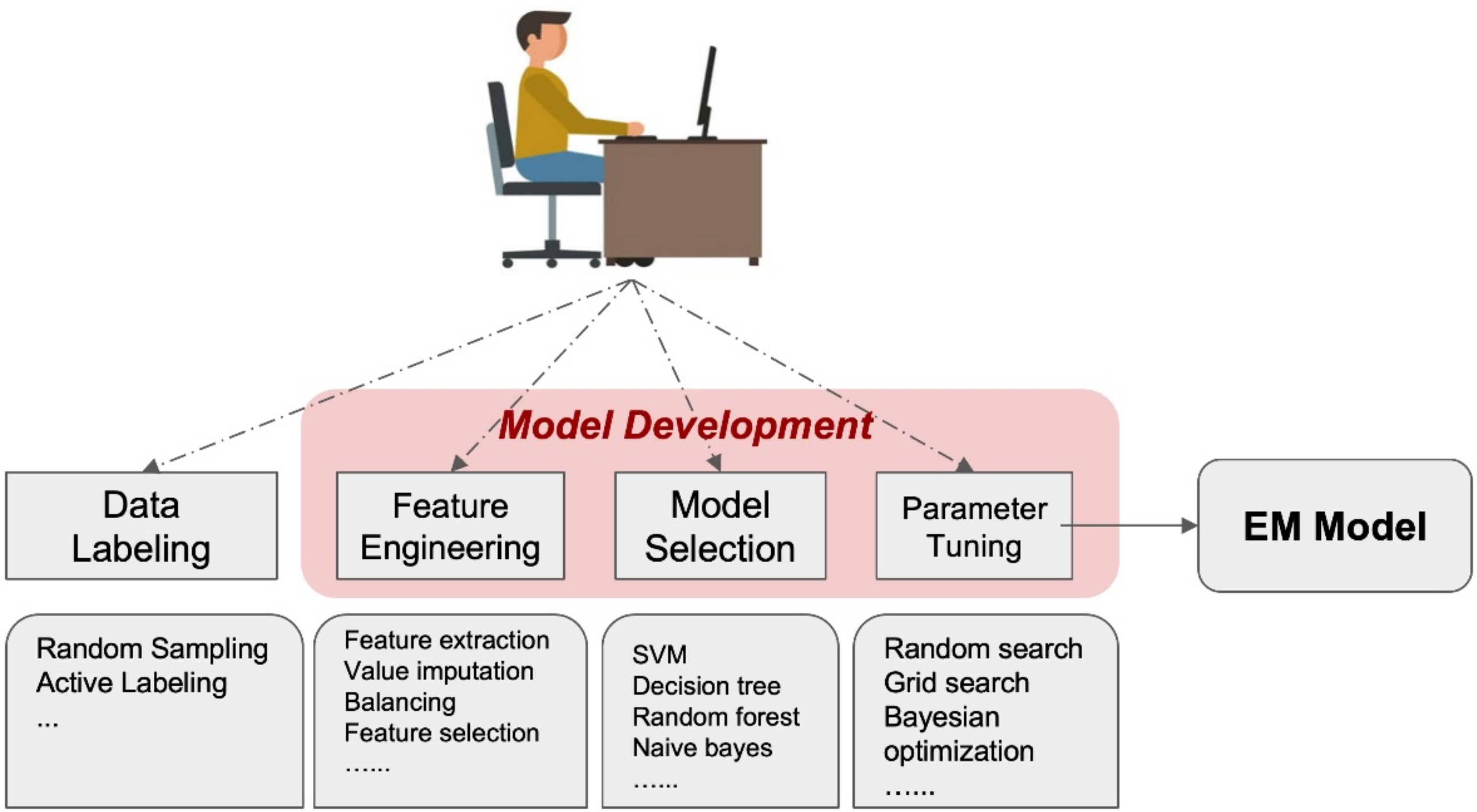
- Similarity function (e.g., Jaccard)
- Threshold (e.g., 0.8)

Jaccard(r1, r2) = 0.9 ≥ 0.8 √
Jaccard(r3, r4) = 0.4 < 0.8 ×

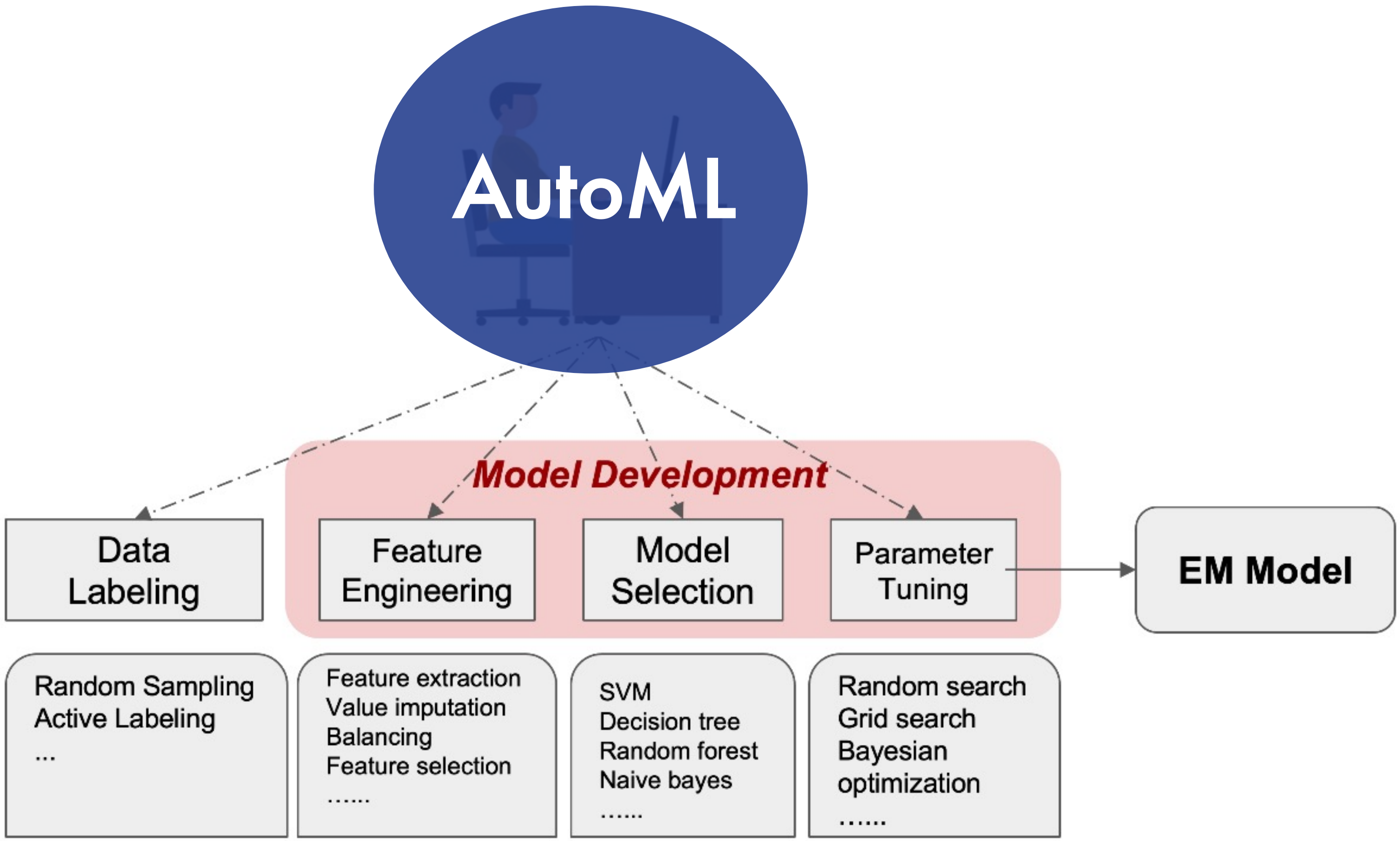
2. Learning-based



Manual Model Development

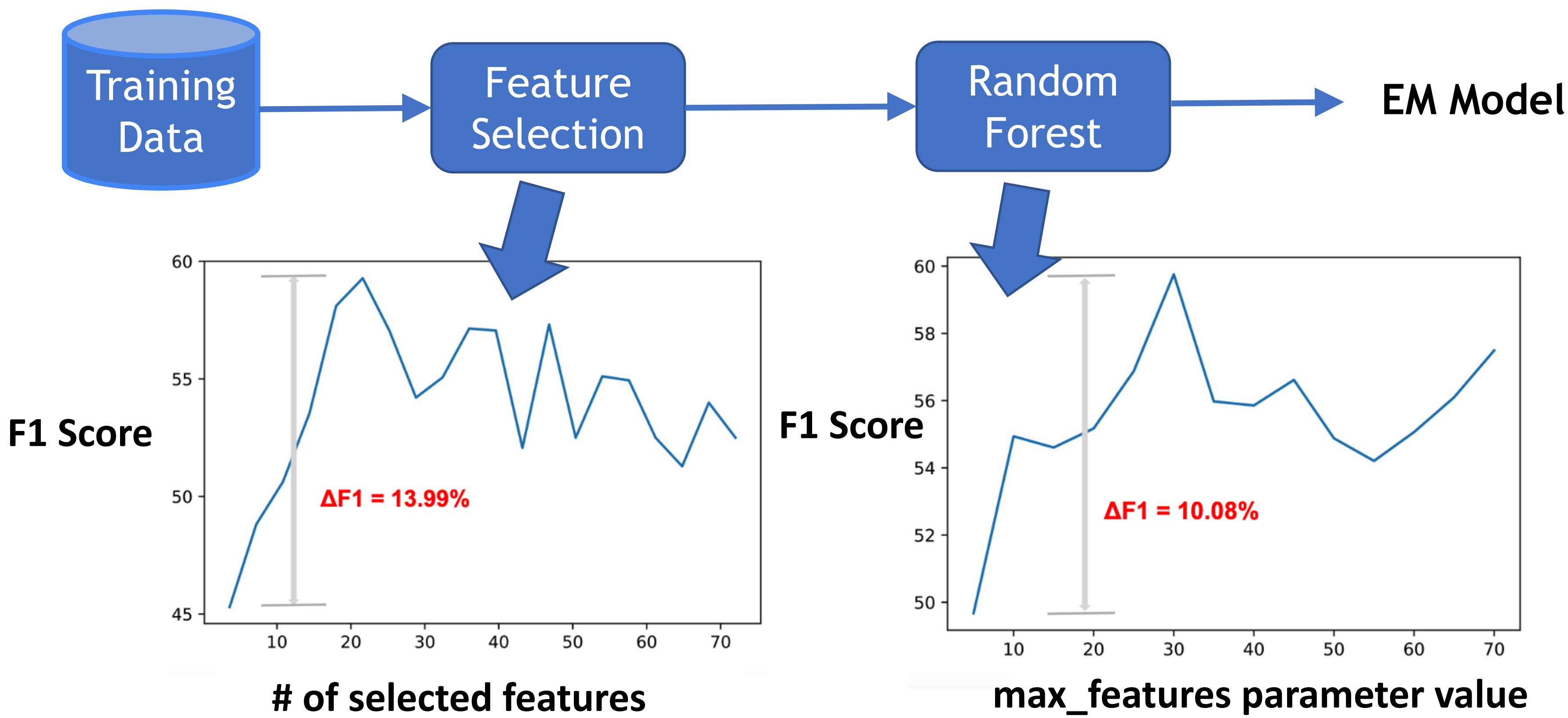


Our Goal


AutoML
Model Development
Data Labeling
Feature Engineering
Model Selection
Parameter Tuning
EM Model
Random Sampling
Active Labeling
...
Feature extraction
Value imputation
Balancing
Feature selection
......
SVM
Decision tree
Random forest
Naive bayes
......
Random search
Grid search
Bayesian optimization
......

Why AutoML?
Reason 1: Tuning Matters

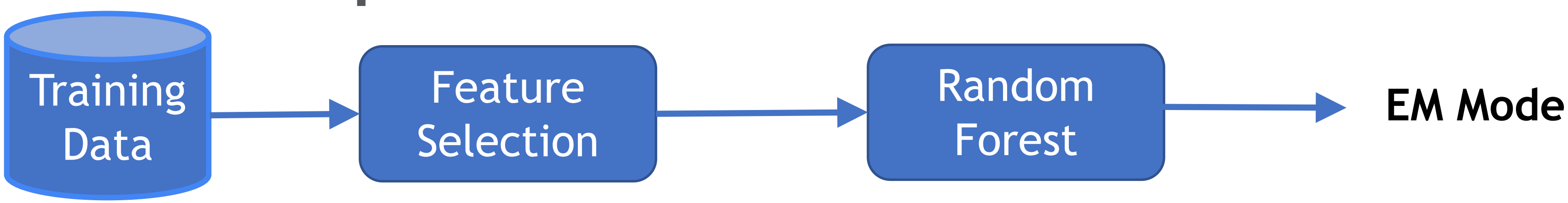
Abt-buy Dataset (70 features)



Why AutoML?
Reason 2: Huge Tuning Space

Abt-buy Dataset (70 features)

Space Size: 70 x 70 = 490



Huge Space

20+ components

Data Preprocessing
compute_class_weight(...)
SimpleImputer(...)
OneHotEncoder(...)
RobustScaler(...)
MinMaxScaler(...)
...

×

20+ components

Feature Preprocessing
SelectPercentile(...)
SelectRates(...)
ExtraTreePreprocessing(...)
pca(...)
FeatureAgglomeration(...)
...

×

10+ models

Model Selection
AdaBoost(...)
DecisionTree(...)
RandomForest(...)
GradientBoosting(...)
KNeighborsClassifier(...)
...

Talk Outline

1. Entity Matching (EM)
2. Automate Model Development
3. Automate Data Labeling
4. Future Direction

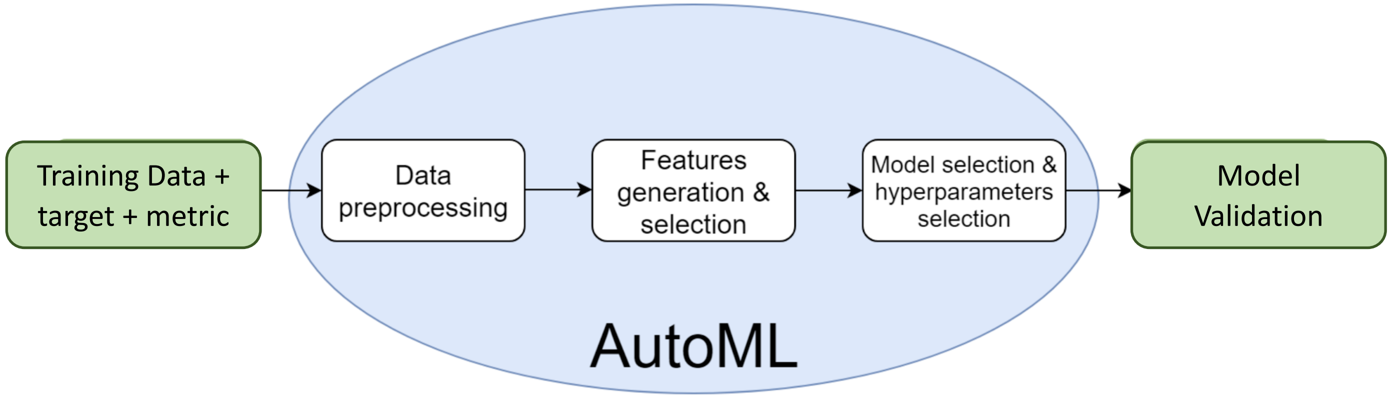
What is AutoML?

Vision

- AutoML allows non-experts to make use of machine learning models and techniques

Scope

- Automate *Data Preprocessing* → *Feature Engineering* → *Model Selection/Hyperparameter Tuning* for Supervised Learning



Will AutoML replace data scientists?

DataRobot

Amazon SageMaker

H2O AutoML

Google's AutoML

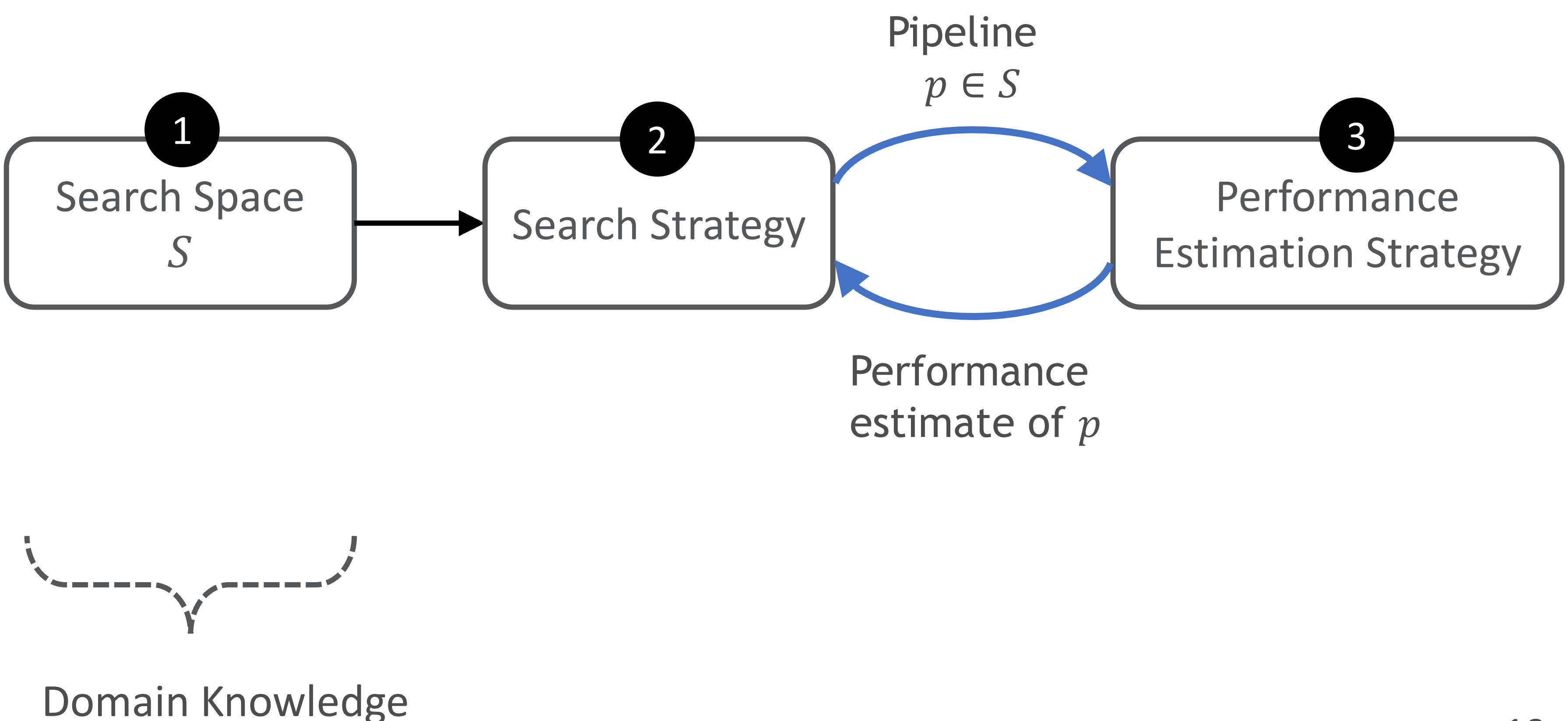
Auto-Sklearn

TPOT

NO! AutoML lacks domain knowledge

How does AutoML work?

Three Steps



How to adopt AutoML in EM?

SFU

Key Idea: Ingest domain knowledge through a careful search space design

Feature Generation:
Magellan Features [1] vs. AutoML-EM Features

Model Selection:
All Models vs. Random Forest

[1] Konda, Pradap, et al. "Magellan: Toward building entity matching management systems." Proceedings of the VLDB Endowment 9.12 (2016): 1197-1208.

Feature Generation

Magellan Features [1]

ID	Data Type	Similarity Function
1	Single-Word String	(Levenshtein Distance, N/A)
2		(Levenshtein Similarity, N/A)
3		(Jaro Distance, N/A)
4		(Exact Match, N/A)
5		(Jaro-Winkler Distance, N/A)
6		(Jaccard Similarity, 3-gram)
7	1-to-5-Word String	(Levenshtein Distance, N/A)
8		(Levenshtein Similarity, N/A)
9		(Needleman-Wunsch Algorithm, N/A)
10		(Smith-Waterman Algorithm, N/A)
11		(Monge-Elkan Algorithm, N/A)
12		(Cosine Similarity, Space)
13		(Jaccard Similarity, Space)
14		(Jaccard Similarity, 3-gram)
15	5-to-10-Word String	(Levenshtein Distance, N/A)
16		(Levenshtein Similarity, N/A)
17		(Monge-Elkan Algorithm, N/A)
18		(Cosine Similarity, Space)
19		(Jaccard Similarity, 3-gram)
20	Long String (>10 words)	(Cosine Similarity, Space)
21		(Jaccard Similarity, 3-gram)
22	Numeric	(Levenshtein Distance, N/A)
23		(Levenshtein Similarity, N/A)
24		(Exact Match, N/A)
25		(Absolute Norm, N/A)
26	Boolean	(ExactMatch, N/A)

V.S.

AutoML-EM Features

ID	Data Type	Similarity Function
1	String	(Levenshtein Distance, N/A)
2		(Levenshtein Similarity, N/A)
3		(Jaro Distance, N/A)
4		(Exact Match, N/A)
5		(Jaro-Winkler Distance, N/A)
6		(Needleman-Wunsch Algorithm, N/A)
7		(Smith-Waterman Algorithm, N/A)
8		(Monge-Elkan Algorithm, N/A)
9		(Overlap Coefficient, Space)
10		(Dice Similarity, Space)
11		(Cosine Similarity, Space)
12		(Jaccard Similarity, Space)
13		(Overlap Coefficient, 3-gram)
14		(Dice Similarity, 3-gram)
15		(Cosine Similarity, 3-gram)
16		(Jaccard Similarity, 3-gram)
17	Number	(Levenshtein Distance, N/A)
18		(Levenshtein Similarity, N/A)
19		(Exact Match, N/A)
20		(Absolute Norm, N/A)
21	Bool	(ExactMatch, N/A)

[1] Konda, Pradap, et al. "Magellan: Toward building entity matching management systems." Proceedings of the VLDB Endowment 9.12 (2016): 1197-1208.

Magellan Features

SFU

Record Pair

ID	Product Name	Price
r_1	iPad Eight 128GB WiFi White	$490

ID	Product Name	Price
r_2	iPad 8th generation 128GB WiFi White	$469

9 Features

ID	Data Type	Similarity Function
1	Single-Word String	(Levenshtein Distance, N/A)
2		(Levenshtein Similarity, N/A)
3		(Jaro Distance, N/A)
4		(Exact Match, N/A)
5		(Jaro-Winkler Distance, N/A)
6		(Jaccard Similarity, 3-gram)
7	1-to-5-Word String	(Levenshtein Distance, N/A)
8		(Levenshtein Similarity, N/A)
9		(Needleman-Wunsch Algorithm, N/A)
10		(Smith-Waterman Algorithm, N/A)
11		(Monge-Elkan Algorithm, N/A)
12		(Cosine Similarity, Space)
13		(Jaccard Similarity, Space)
14		(Jaccard Similarity, 3-gram)
15	5-to-10-Word String	(Levenshtein Distance, N/A)
16		(Levenshtein Similarity, N/A)
17		(Monge-Elkan Algorithm, N/A)
18		(Cosine Similarity, Space)
19		(Jaccard Similarity, 3-gram)
20	Long String (>10 words)	(Cosine Similarity, Space)
21		(Jaccard Similarity, 3-gram)
22	Numeric	(Levenshtein Distance, N/A)
23		(Levenshtein Similarity, N/A)
24		(Exact Match, N/A)
25		(Absolute Norm, N/A)

5 features for Product Name

4 features for Price

AutoML-EM Features

SFU

Record Pair

ID	Product Name	Price
r_1	iPad Eight 128GB WiFi White	$490

ID	Product Name	Price
r_2	iPad 8th generation 128GB WiFi White	$469

20 Features

ID	Data Type	Similarity Function
1	String	(Levenshtein Distance, N/A)
2		(Levenshtein Similarity, N/A)
3		(Jaro Distance, N/A)
4		(Exact Match, N/A)
5		(Jaro-Winkler Distance, N/A)
6		(Needleman-Wunsch Algorithm, N/A)
7		(Smith-Waterman Algorithm, N/A)
8		(Monge-Elkan Algorithm, N/A)
9		(Overlap Coefficient, Space)
10		(Dice Similarity, Space)
11		(Cosine Similarity, Space)
12		(Jaccard Similarity, Space)
13		(Overlap Coefficient, 3-gram)
14		(Dice Similarity, 3-gram)
15		(Cosine Similarity, 3-gram)
16		(Jaccard Similarity, 3-gram)
17	Number	(Levenshtein Distance, N/A)
18		(Levenshtein Similarity, N/A)
19		(Exact Match, N/A)
20		(Absolute Norm, N/A)

16 features for Product Name

4 features for Price

Experiments – setup & datasets

SFU

- **AutoML-EM**
  - Built on Auto-sklearn
- **Methods for comparison**
  - Magellan [1]:state-of-the-art library for EM model development
  - DeepMatcher [2]: state-of-the-art deep learning models for EM
- **Datasets**
  - Eight benchmark datasets

Type	Dataset	Training Size	Test Size	# Attr.
Easy & Small	BeerAdvo-RateBeer	359	91	4
	Fodors-Zagats	757	189	6
	iTunes-Amazon	430	109	8
Easy & Large	DBLP-ACM	9890	2473	4
	DBLP-Scholar	22965	5742	4
Hard & Large	Amazon-Google	9167	2293	3
	Walmart-Amazon	8193	2049	5
	Abt-Buy	7659	1916	3

[1]. Konda, Pradap, et al. "Magellan: Toward building entity matching management systems." VLDB 2016.
[2]. Mudgal, Sidharth, et al. "Deep learning for entity matching: A design space exploration." SIMGOD 2018.

Feature Generation

SFU

- Magellan Features vs. AutoML-EM Features

Dataset	Magellan		AutoML-EM		ΔF1 Score
	# Feature	Fscore	#Feature	Fscore	
BeerAdvo-RateBeer	36	81.3	87	82.3	+1.0
Fodors-Zagats	37	100	123	100	+0
iTunes-Amazon	30	88.1	155	96.3	+8.2
DBLP-ACM	18	98.3	89	98.4	+0.1
DBLP-Scholar	18	92.6	89	94.6	+2.0
Amazon-Google	21	62.9	72	66.4	+3.5
Walmart-Amazon	32	66.2	106	78.5	+2.3
Abt-Buy	15	48.1	72	59.2	+11.1

AutoML-EM features outperform Magellan Features by up to 11.1 %

Can AutoML-EM Beat Human?

- Human vs. AutoML-EM

Dataset	Human	AutoML-EM	△ F1 Score
BeerAdvo-RateBeer	78.8	82.3	+3.5
Fodors-Zagats	100	100	+0
iTunes-Amazon	91.2	96.3	+5.1
DBLP-ACM	98.4	98.4	+0
DBLP-Scholar	92.3	94.6	+2.3
Amazon-Google	49.1	66.4	+17.3
Walmart-Amazon	71.9	78.5	+6.6
Abt-Buy	43.6	59.2	+5.3
Average	**78.1**	**83.9**	**+5.8**

AutoML-EM beats human by an average of 5.8 % in F1 Score

Deep Learning Based EM

SFU

Deep learning for **entity matching**: A design space exploration

…, AH Doan, Y Park, G Krishnan, R **Deep**… - Proceedings of the …, 2018 - dl.acm.org

Entity matching (EM) finds data instances that refer to the same real-world entity. In this paper we examine applying deep learning (DL) to EM, to understand DL's benefits and limitations. We review many DL solutions that have been developed for related matching …

Cited by 197 Related articles All 6 versions

Distributed representations of tuples for entity resolution

…, S Joty, M Ouzzani, N **Tang** - Proceedings of the …, 2018 - dl.acm.org

Despite the efforts in 70+ years in all aspects of entity resolution (ER), there is still a high demand for democratizing ER-by reducing the heavy human involvement in labeling data, performing feature engineering, tuning parameters, and defining blocking functions. With the …

Cited by 107 Related articles All 5 versions

Deep entity matching with pre-trained language models

Y Li, J Li, Y Suhara, AH Doan, WC Tan - arXiv preprint arXiv:2004.00584, 2020 - arxiv.org

… Even though we can also train **deep learning** EM solutions to **learn** such knowledge, we … Types of Important Spans Publications, Movies, Music Persons (eg, Authors), Year, Publisher Organizations, Employers Last 4-digit of phone, Street number Products …

Cited by 19 Related articles All 7 versions

Can AutoML-EM beat deep learning?

AutoML-EM wins on structured data by up to 13%

Dataset	DeepMatcher	AutoML-EM	△ F1 Score
BeerAdvo-RateBee	72.7	80.9	+8.2
DBLP-ACM	98.4	98.1	-0.3
DBLP-Scholar	94.7	94.6	-0.1
Fodors-Zaqats	100.0	100.0	+ 0
Walmart-Amazon	66.9	79.9	+13
iTunes-Amazon	88.0	95.7	+7.7

Deep learning wins on textual data but NOT by a large margin

Dataset	DeepMatcher	AutoML-EM	△ F1 Score
Amazon-Google	69.3	63.8	-5.5
Abt-Buy	62.8	58.1	-4.7

Deep Learning vs. AutoML-EM

	Deep Learning	AutoML-EM
Interpretability	✗	✓
Time efficiency	✗	✓
Performance on structured data	✗	✓

Takeaways

Innovation

- The **first** work to apply AutoML to EM

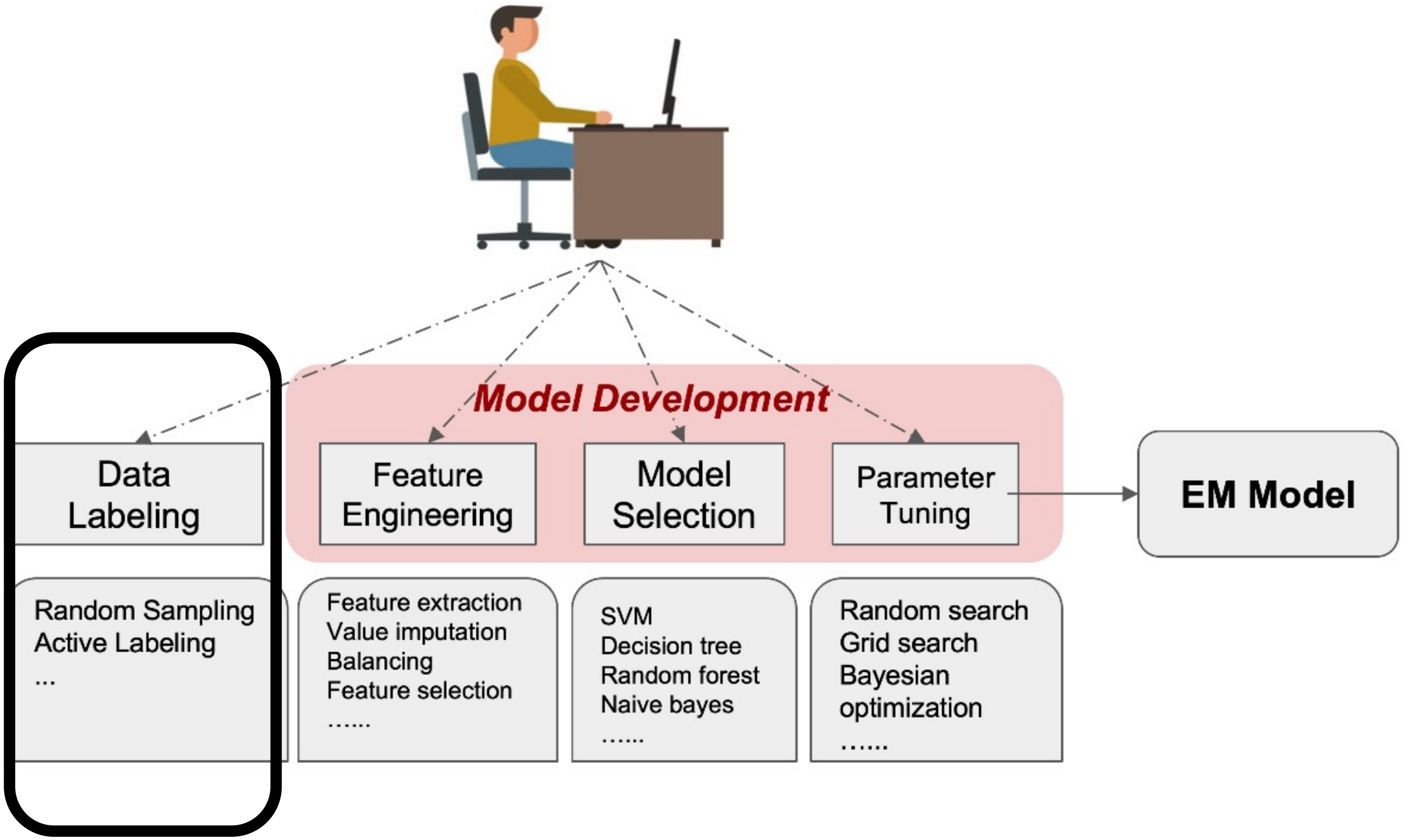
Key Findings

1. AutoML-EM beats human by a large margin
2. AutoML-EM outperforms deep learning on structured data
3. AutoML-EM is competitive to deep learning on textual data

Talk Outline

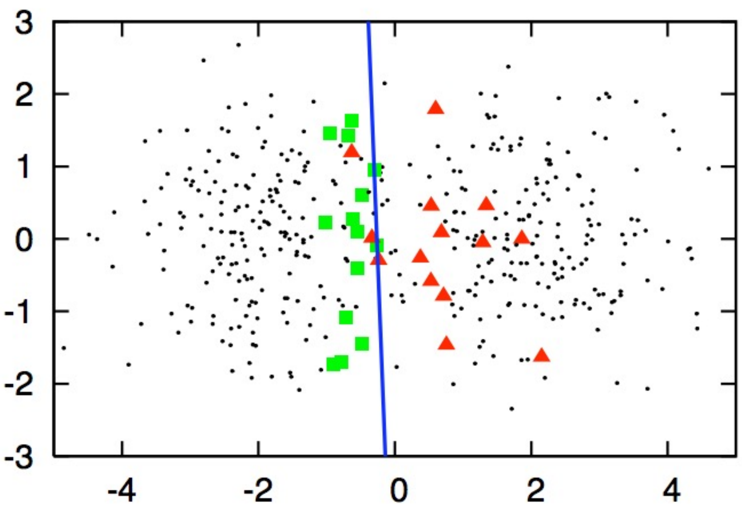
1. Entity Matching (EM)
2. Automate Model Development
3. Automate Data Labeling
4. Future Direction

Data Labeling

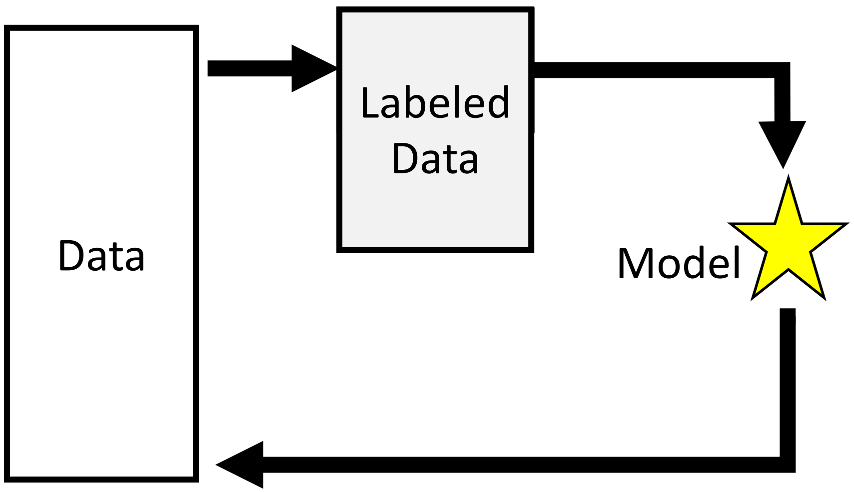


Active Learning

Illustration



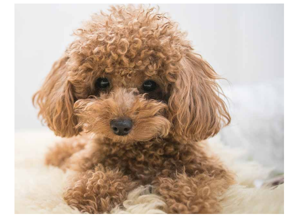
Workflow



SFU

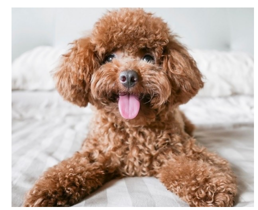
Self-training

1. Train model on labeled data



Labeled data

2. Use model to predict unlabeled data

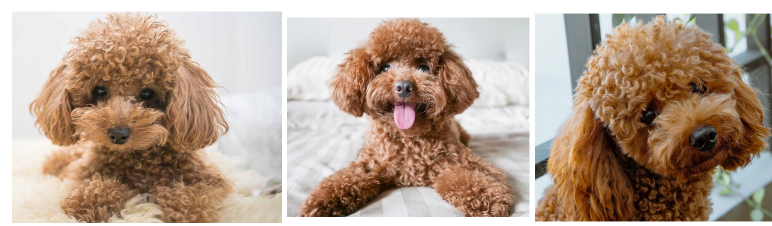


High Conf.

High Conf.

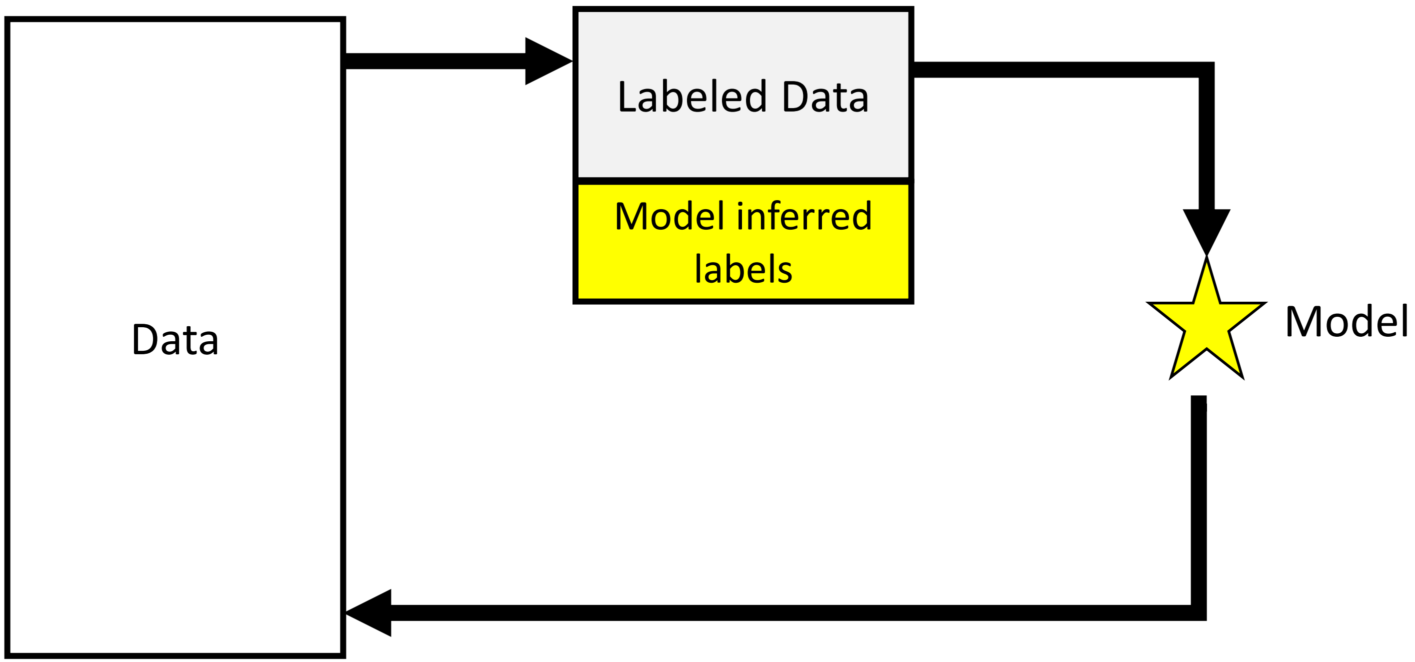
Low Conf.

3. Add predicted unlabeled with high confidence to training set



New labeled data

Active Learning + Self-training



Takeaways

Innovation

- The **first** work to combine active learning and self-training for EM

Key Findings

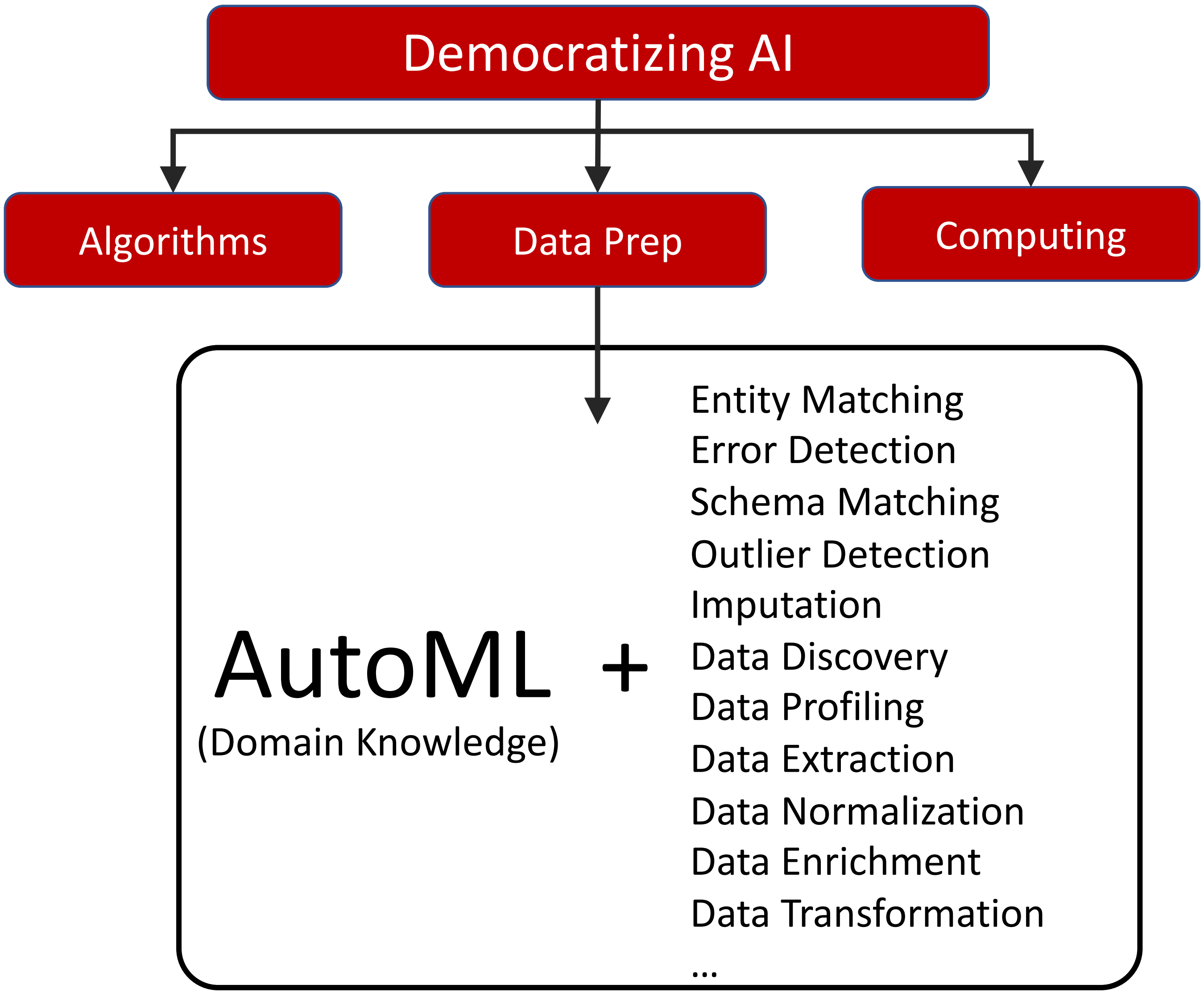
1. Our combined solution beats active learning only solution
2. Our combined solution beats self-training only solution

Talk Outline

1. Entity Matching (EM)
2. Automate Model Development
3. Automate Data Labeling
4. Future Direction

SFU

The journey has just begun



Thank you!