# Data Augmentation for ML-driven Data Preparation and Integration

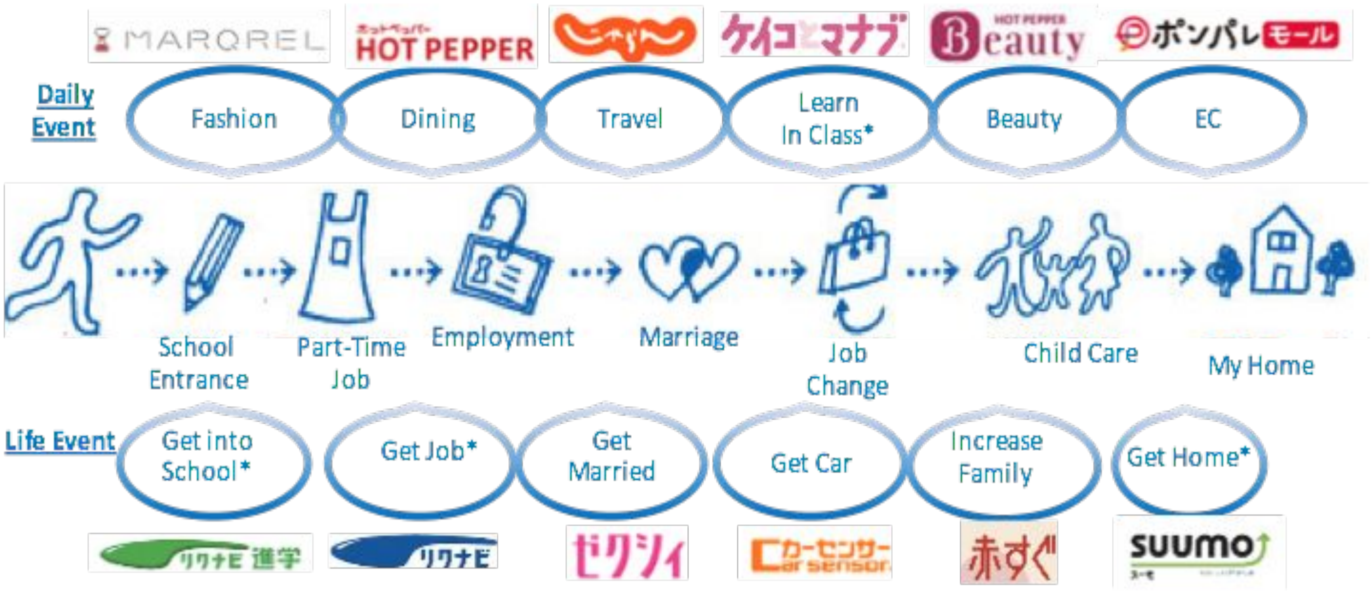
Yuliang Li, Xiaolan Wang, Zhengjie Miao, Wang-Chiew Tan

Megagon Labs

DUKE COMPUTER SCIENCE

FACEBOOK AI

# Megagon Labs - Research arm of Recruit Holdings




indeed

glassdoor

TRUSTYOU

(Many) Research challenges in Data Integration, KB, and NLP

# Research at Megagon Labs

**Subjective Data**

**Experiences or Opinions**

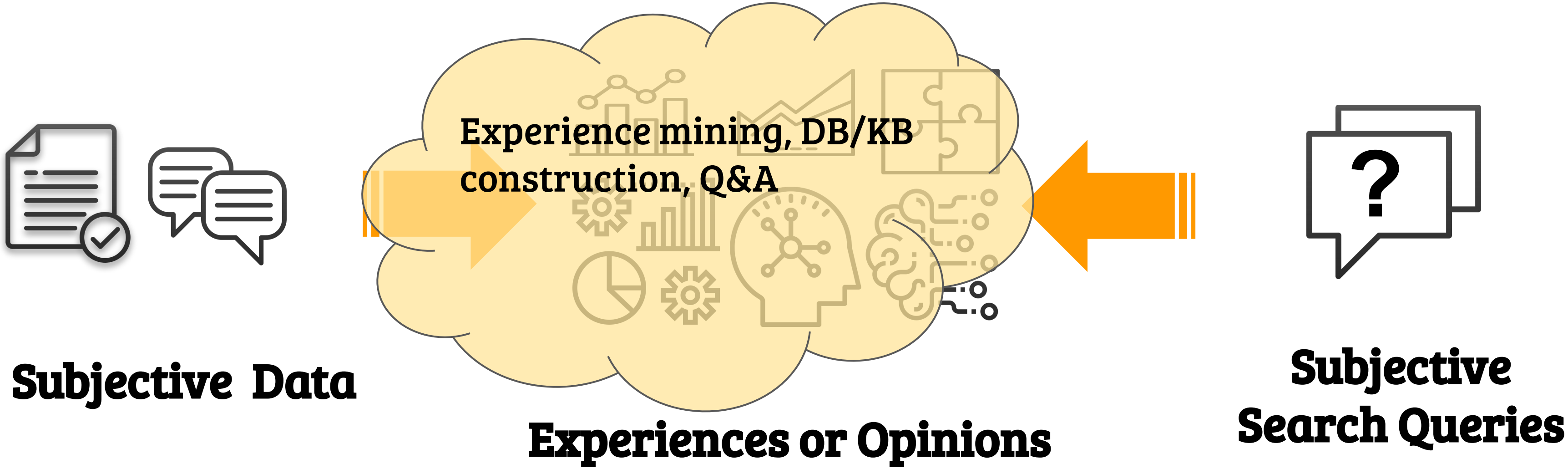
# Research at Megagon Labs

**Subjective Data**

I reached the hotel by cab. **Checkin was very smooth** but mostly I was surprised by the **very spacious bathroom** with **lots of provided toiletries.**

**Experiences or Opinions**

# Research at Megagon Labs



# Example Research Projects

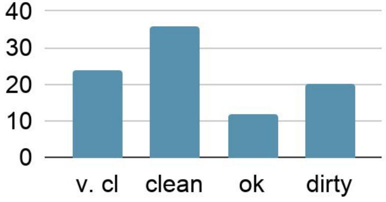
**Subjective DB (OpineDB)**

[VLDB19, CHI20, theWebConf20]

Review Corpus

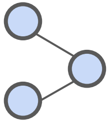
“very clean room”,
“close to attraction”
...

Aggregate




Correlate

KB



**Entity Matching (Ditto)**

[VLDB21, JDIQ21] + Ongoing

*“Google LLC” ==? “Alphabet Inc”*

Resumes ⋈ Job postings

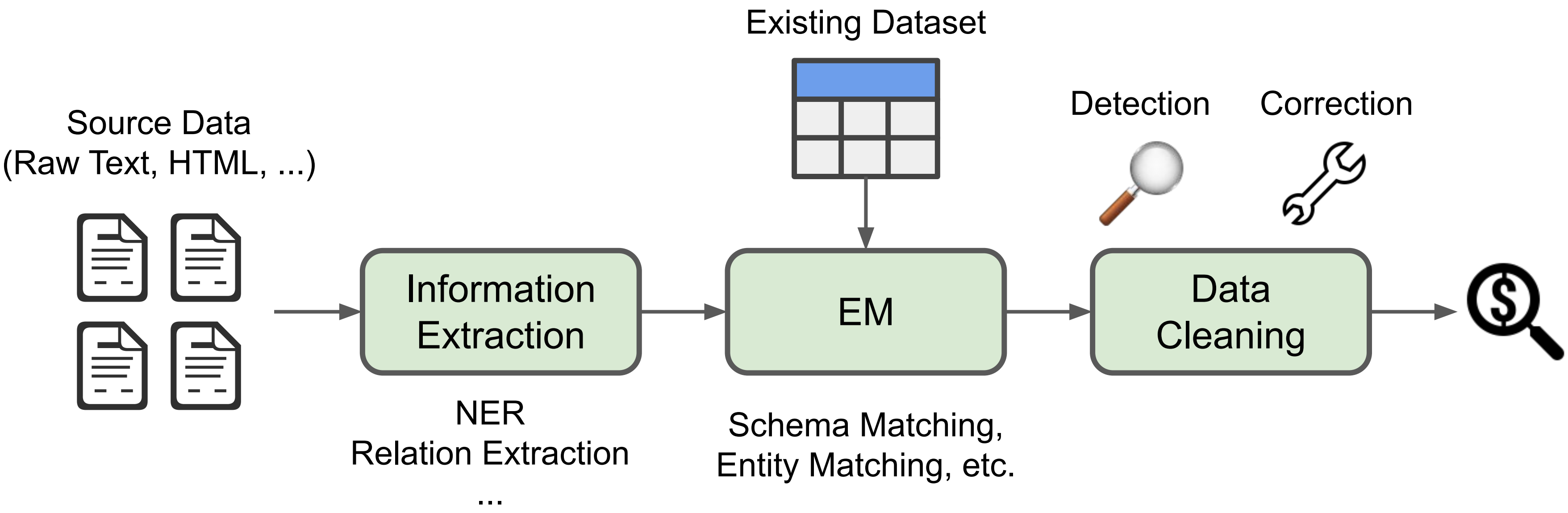
Explanation

**KBs / QA**

*Q: What are some common benefits for Cashiers of Walmart?*

*A: 401K, Health Insurance, product discount, ...*

# An end-to-end workflow of Data Preparation & Integration

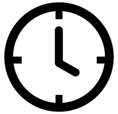



ML, especially DL, achieves promising results in every step

Machine (Deep) Learning contributes a lot to the success of these projects

*… when there is sufficient training data*

*(e.g., ImageNet: 14M; SquAD2.0: 150K)*



# Challenge: need more labels

- Datasets for training an opinion extraction pipeline:

| | Tagging | Pairing | Attribute |
|---|---|---|---|
| Hotel | (800, 112) | (1000, 1000) | (4000, 1000) |
| Restaurant | (3041, 800) | | (4000, 1000) |

Took two of us ~5 hours of labeling

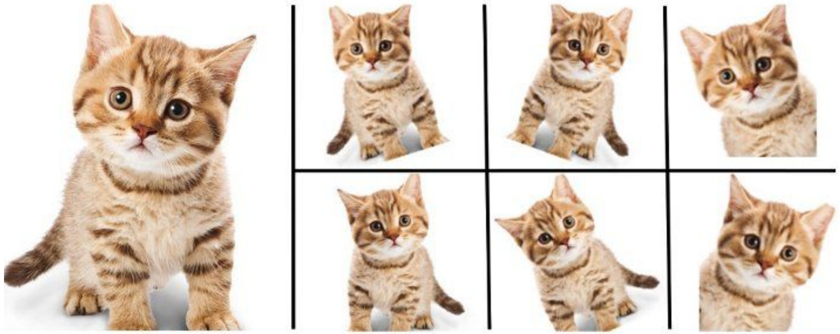
| | F1 Tagging | F1 Pairing | Acc Attribute |
|---|---|---|---|
| Hotel | 74.71% | 83.87% | 86.63% |
| Restaurant | 85.53% | | 88.29% |

More data => higher-quality

- How to collect more high-quality training examples?
  - **Domain Expertise:** not scalable
  - **Crowd-sourcing:** expensive (thousands of $), noisy labels, etc.

# Data Augmentation (DA)

- Data transformation for generating additional training example
- **DA In computer vision:**



Invariance property

- **DA In NLP**: word deletion, synonym replacement etc.

We have seen success stories of applying DA to Data Management

# In this tutorial:

- Part I: DA for Data Management (Xiaolan Wang)
  - EM, Cleaning, schema matching, Information extraction (sequence tagging)
  - Deep learning for Data Preparation and Integration
  - Data augmentation operators
- Part II: Advanced DA (Yuliang Li)
  - Interpolation (MixUp, MixDA, and follow-up)
  - Generation (Conditional generation, GAN, InvDA)
  - Learned DA policy (AutoDA, HoloDetect, meta-learning e.g., Rotom,)
- Part III: Connection with other learning paradigms (Zhengjie Miao)
  - SSL (DA used as consistency regularization)
  - active learning (used together with DA to get more labels)
  - Weak-supervision (e.g., present Snorkel and discuss how to combine Snorkel with DA)
  - Pre-training for relational data

# Part I: DA for Data Management

Which data management tasks can be benefited from data augmentation?

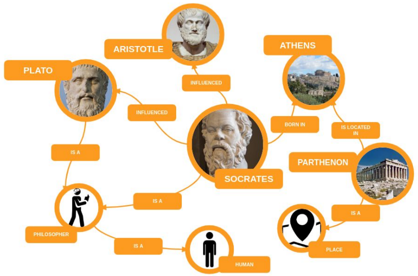
What are the basic data augmentation operators?

# Part I: DA for Data Management

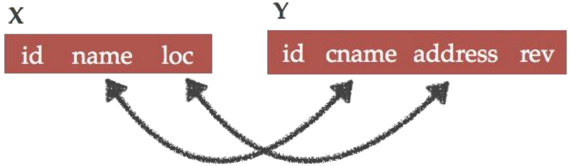
Which data management tasks can be benefited from data augmentation?

What are the basic data augmentation operators?

# Data management tasks

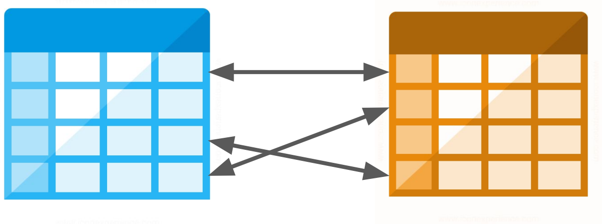



**Information Extraction**




**Schema Matching**

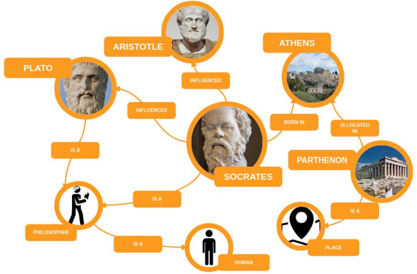
**Data Cleaning**



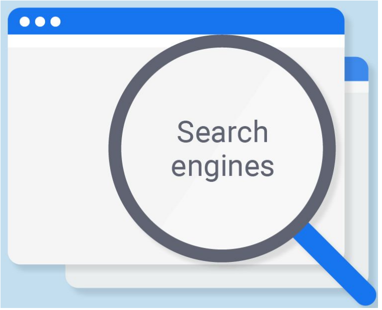
**Entity Matching**

# Information Extraction

- Extracting structured information from unstructured or semi-structured data sources.
  - Named entity recognition
  - Relation extraction
  - … …




**Information Extraction**




**Enrich Search Results**


**Improve Reading Comprehension**

alexa WolframAlpha

**Support Various Products**

# Problem definition (Named Entity Recognition)

For example:

“VLDB conference is an annual conference held by the non-profit Very Large Data Base Endowment Inc.”

# Problem definition (Named Entity Recognition)

- Input: a sequence of text.
- Output: B/I/O tags for every token (or word) in the input sequence.
    - B: Begin;
    - I: Inside;
    - O: Outside.

For example:

"VLDB conference is an annual conference held by the non-profit Very Large Data Base Endowment Inc."

B I O O

VLDB conference

# Problem definition (Named Entity Recognition)

- Input: a sequence of text.
- Output: B/I/O tags for every token (or word) in the input sequence.
  - B: Begin of a sequence;
  - I: Inside a sequence;
  - O: Outside of a sequence.

For example:

"VLDB conference is an annual conference held by the non-profit Very Large Data Base Endowment Inc."

B I O O B I O O O O B I I I I I

VLDB conference

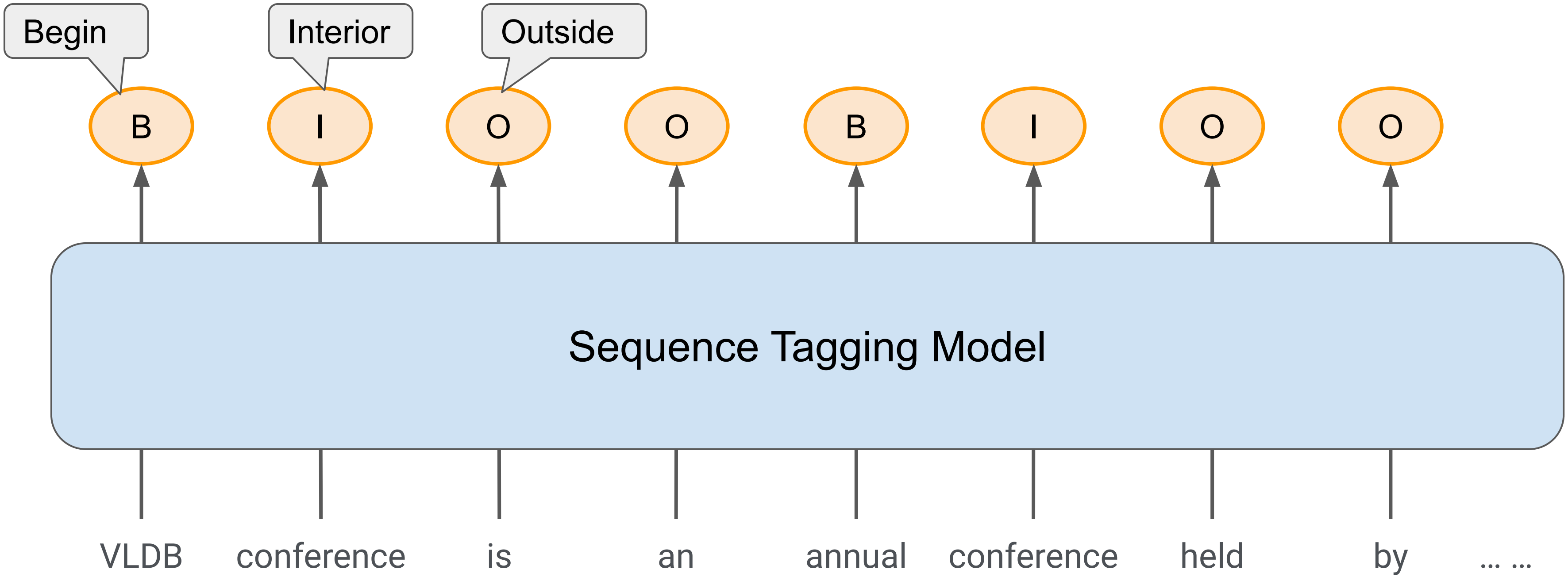
annual conference

Very Large Data Base Endowment Inc.

Named entities

# Machine learning-based solutions

- Machine learning can be used to solve information extraction problem.



# Data Cleaning


- Identify & correct erroneous data.
  - Error detection
  - Error correction

## Data Error Incorrectly Estimated 500K More Vaccine Shots In PA

Posted By: Tyler Friel on: July 13, 2021 In: Featured News Print Email

The Pennsylvania Department of Health says a data collection error wrongly added about 500,000 total COVID vaccine shots to the state's reporting system.

It came after DOH staff were trying to link first and second doses to individuals in each county.

At a local level, the number of people with at least one shot of the vaccine in Butler County nearly dropped by 50 percent.

Last week the data showed that over 10,000 county residents were at least partially vaccinated, but now that number sits at just shy of 6,000. The number of fully vaccinated individuals in the county is still over 91,000.

# Problem definition

- Input: a cell in a database
- Output: whether the cell is correct or not.

| Year | City | Country | Link |
|---|---|---|---|
| 2021 | Copenhagen | United States | http://vldb.org/2021/ |
| 2020 | Tokyo | Japan | https://vldb2020.org/ |
| 2019 | Los Angeles, California | United States | https://vldb.org/2019/ |
| 2018 | Rio de Janeiro | Brazil | http://vldb2018.lncc.br |
| 2017 | Munich | Germany | http://www.vldb.org/2017/ |

VLDB Venues

# Problem definition

- Input: a cell in a database
- Output: whether the cell is correct or not.

Identify erroneous values in data

| Year | City | Country | Link |
|---|---|---|---|
| 2021 | Copenhagen | **United States** | http://vldb.org/2021/ |
| 2020 | Tokyo | Japan | https://vldb2020.org/ |
| 2019 | Los Angeles, California | United States | https://vldb.org/2019/ |
| 2018 | Rio de Janeiro | Brazil | http://vldb2018.lncc.br |
| 2017 | Munich | Germany | http://www.vldb.org/2017/ |

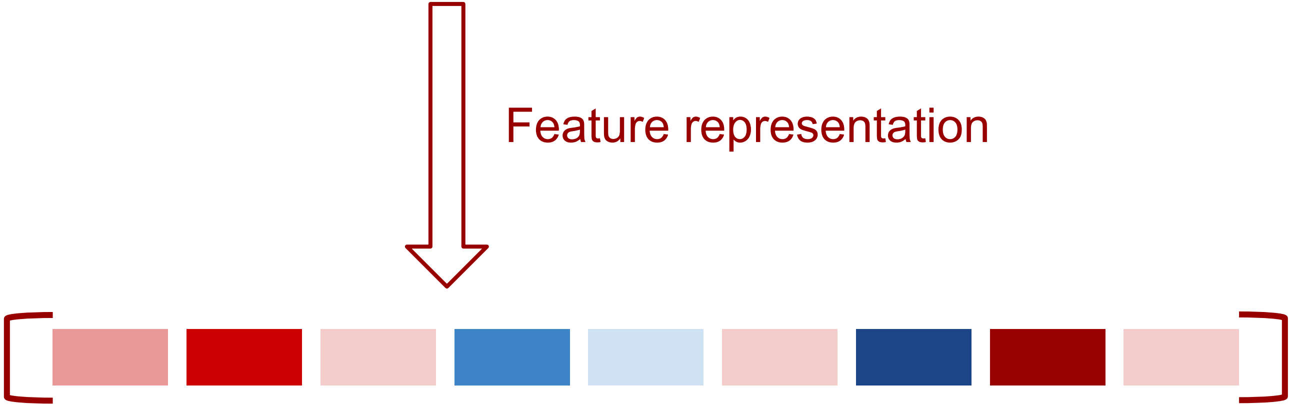
VLDB Venues

**Typos** | **Duplicated values** | **Outliers** | **Missing values** | **Bogus values**

# Machine learning-based solutions

- Machine learning can be used to solve error detection problem.

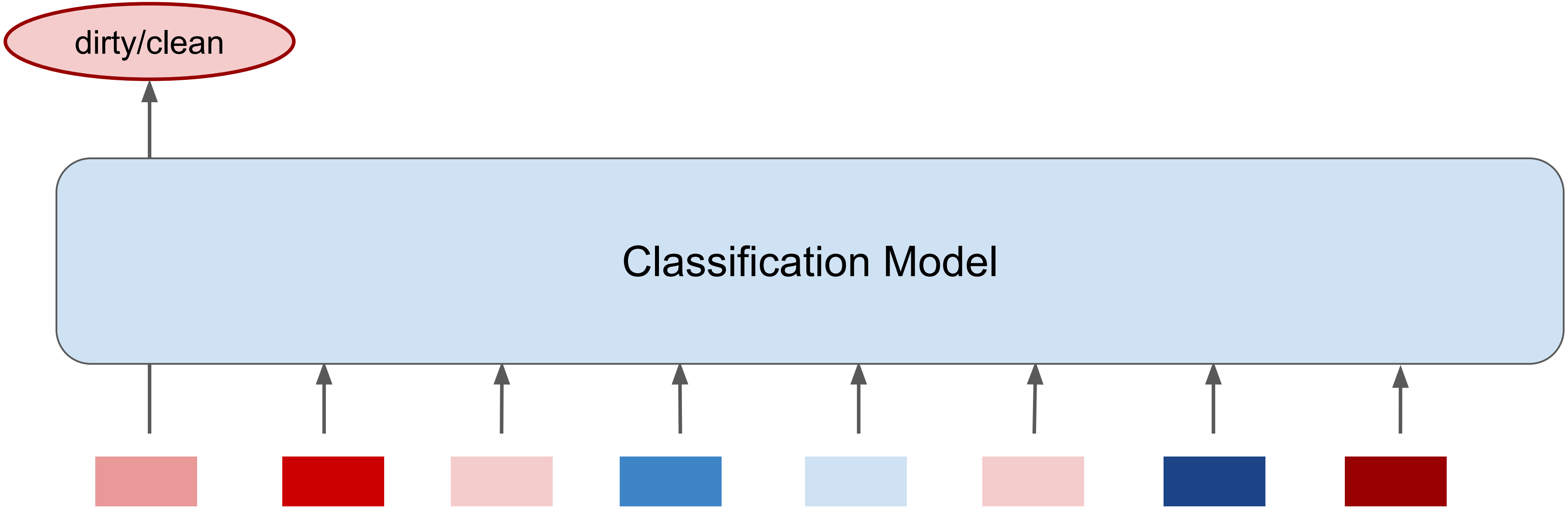
| Year | City | Country | Link |
|---|---|---|---|
| 2021 | Copenhagen | United States | http://vldb.org/2021/ |



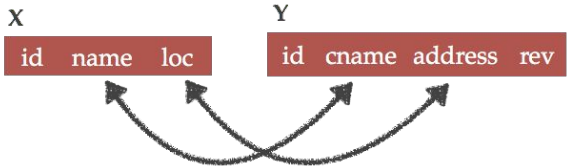

**Feature vector:**

# Machine learning-based solutions

- Machine learning can be used to solve error detection problem.



# Schema Matching




**Schema Matching**

- Schema Matching focuses on finding the correspondence among schema elements in two semantically correlated schema.
  - Input: a column A from the source db and a column B from the target db.
    Output: whether column A matches to column B

| Year | City | Country | Link |
|---|---|---|---|
| 2021 | Copenhagen | Danmark | http://vldb.org/2021/ |
| 2020 | Tokyo | Japan | https://vldb2020.org/ |
| 2019 | Los Angeles, California | United States | https://vldb.org/2019/ |
| 2018 | Rio de Janeiro | Brazil | http://vldb2018.lncc.br |
| 2017 | Munich | Germany | http://www.vldb.org/2017/ |

VLDB Venues

| Time | Location | Website |
|---|---|---|
| 2019 | Amsterdam | [1] |
| 2018 | Houston | Page |
| 2017 | Chicago | [2] |
| 2016 | San Francisco | [3] |
| 2015 | Melbourne | [4] |
| 2014 | Snowbird | [5] |

SIGMOD Venues

# Machine learning-based solutions

- Machine learning can be used to solve schema matching problem.

| Year | City | Country | Link |
|---|---|---|---|
| 2021 | Copenhagen | Danmark | http://vldb.org/2021/ |
| 2020 | Tokyo | Japan | https://vldb2020.org/ |
| 2019 | Los Angeles, California | United States | https://vldb.org/2019/ |
| 2018 | Rio de Janeiro | Brazil | http://vldb2018.lncc.br |
| 2017 | Munich | Germany | http://www.vldb.org/2017/ |

VLDB Venues

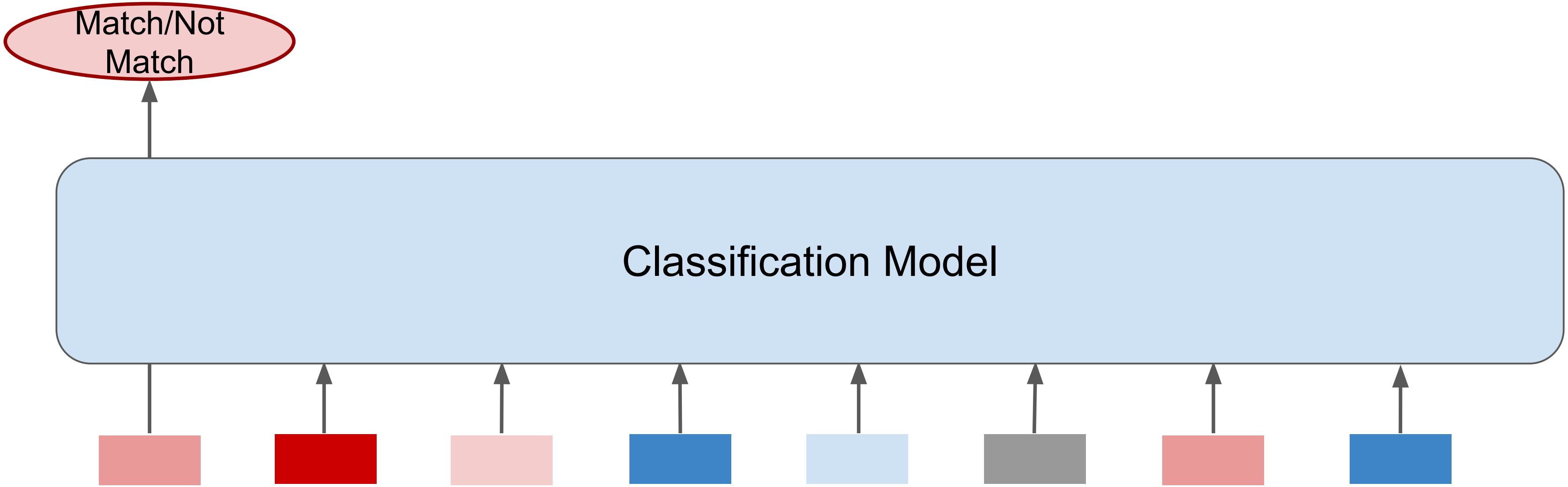
| Time | Location | Website |
|---|---|---|
| 2019 | Amsterdam | [1] |
| 2018 | Houston | Page |
| 2017 | Chicago | [2] |
| 2016 | San Francisco | [3] |
| 2015 | Melbourne | [4] |
| 2014 | Snowbird | [5] |

SIGMOD Venues

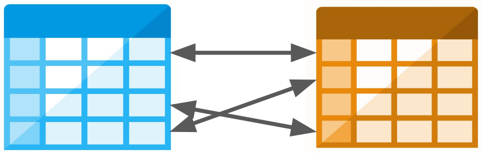
Feature representation

# Machine learning-based solutions

- Machine learning can be used to solve schema matching problem.



# Entity Matching



**Entity Matching**

- Entity Matching, a.k.a., record linkage and entity resolution, is the problem of identifying records that refer to the same real-world entity.
    - Input: a pair of tuples
      Output: whether they refer to the same entity or not.

| Name | Impact |
|---|---|
| SIGMOD: ACM SIGMOD Conf on Management of Data | 0.99 |
| VLDB: Very Large Data Bases | 0.99 |
| KDD: Knowledge Discovery and Data Mining | 0.99 |
| ICDE: Intl Conf on Data Engineering | 0.98 |
| ICDT: Intl Conf on Database Theory | 0.97 |
| S&P: IEEE Symposium on Security and Privacy | 0.97 |
| SIGIR: ACM SIGIR Conf on Information Retrieval | 0.96 |
| PODS: ACM SIGMOD Conf on Principles of DB Systems | 0.95 |

| Avg citations | Conference |
|---|---|
| 31.6 | VLDB—Very Large Data Bases |
| 30.9 | BioMED—Biomedical Engineering |
| 30.9 | IEEE TRANS ROBOTICS AUTOMAT—IEEE Transactions on Robotics and Automation |
| 30.6 | CRYPTO—International Crytology Conference |
| 30.1 | PAMI—IEEE Transactions on Pattern Analysis and Machine Intelligence |

# Machine learning-based solutions

- Machine learning can be used to solve entity matching problem.

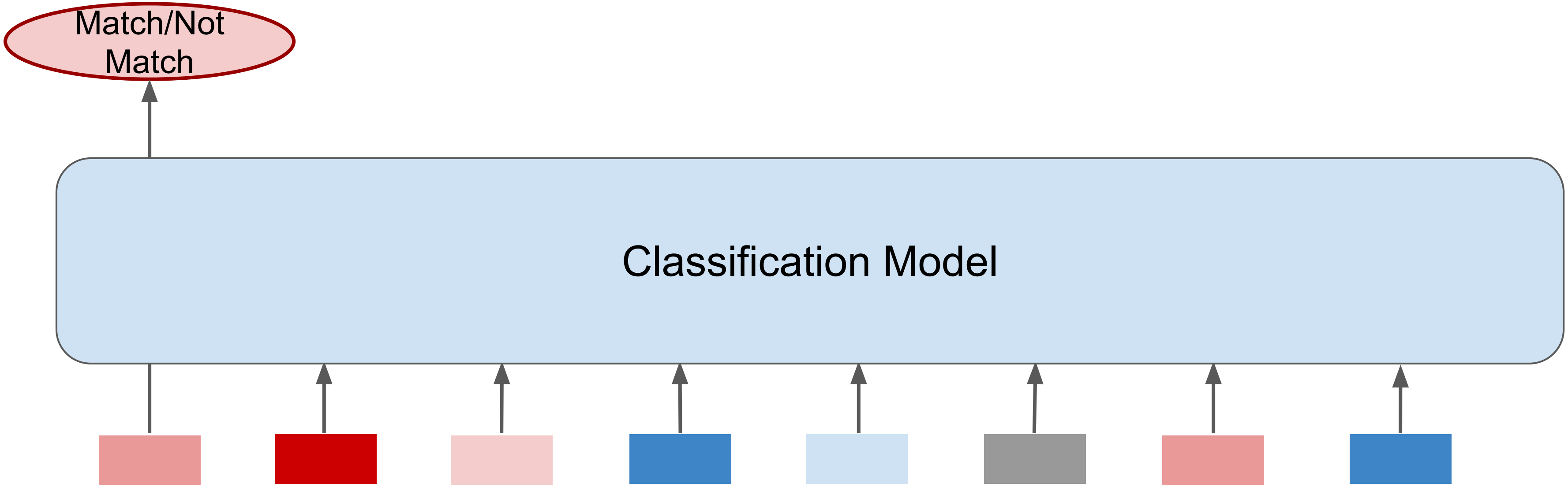
| Name | Impact |
|---|---|
| SIGMOD: ACM SIGMOD Conf on Management of Data | 0.99 |
| VLDB: Very Large Data Bases | 0.99 |
| KDD: Knowledge Discovery and Data Mining | 0.99 |

| Avg citations | Conference |
|---|---|
| 31.6 | VLDB—Very Large Data Bases |
| 30.9 | BioMED—Biomedical Engineering |

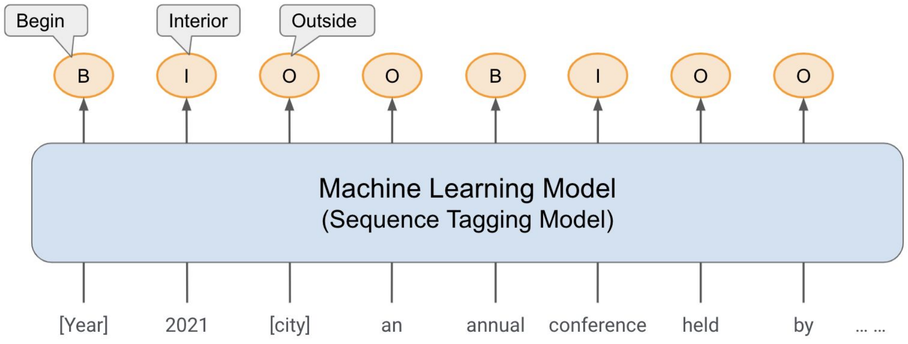
Feature representation

# Machine learning-based solutions

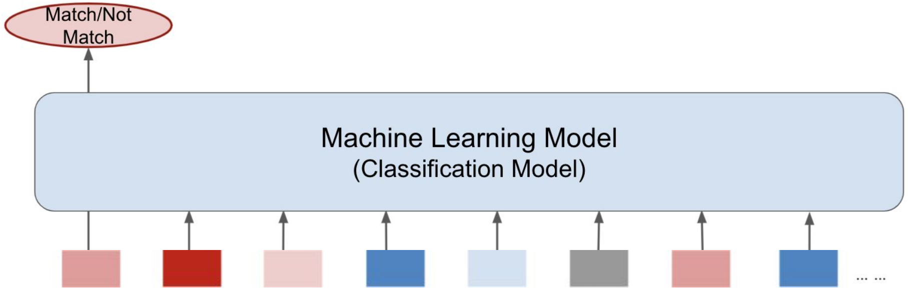
- Machine learning can be used to solve entity matching problem.



# Machine Learning for Data Management tasks

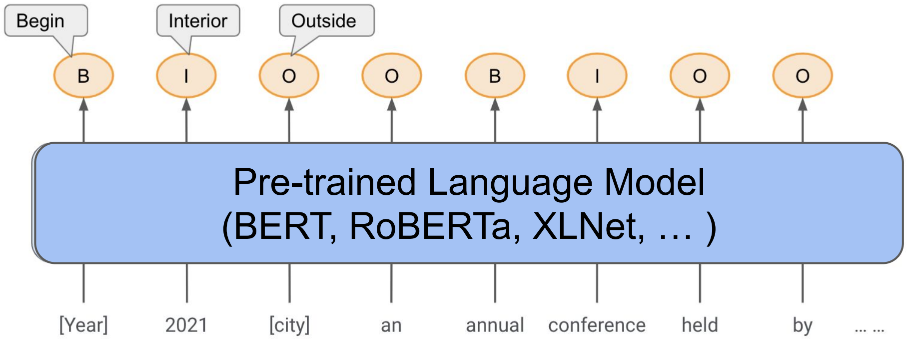



**Sequence Tagging Task**

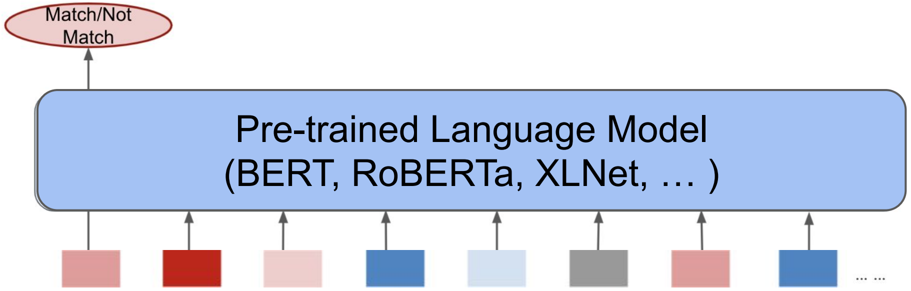



**Sequence (pair) Classification Task**

# Pre-trained LM for Data Management tasks

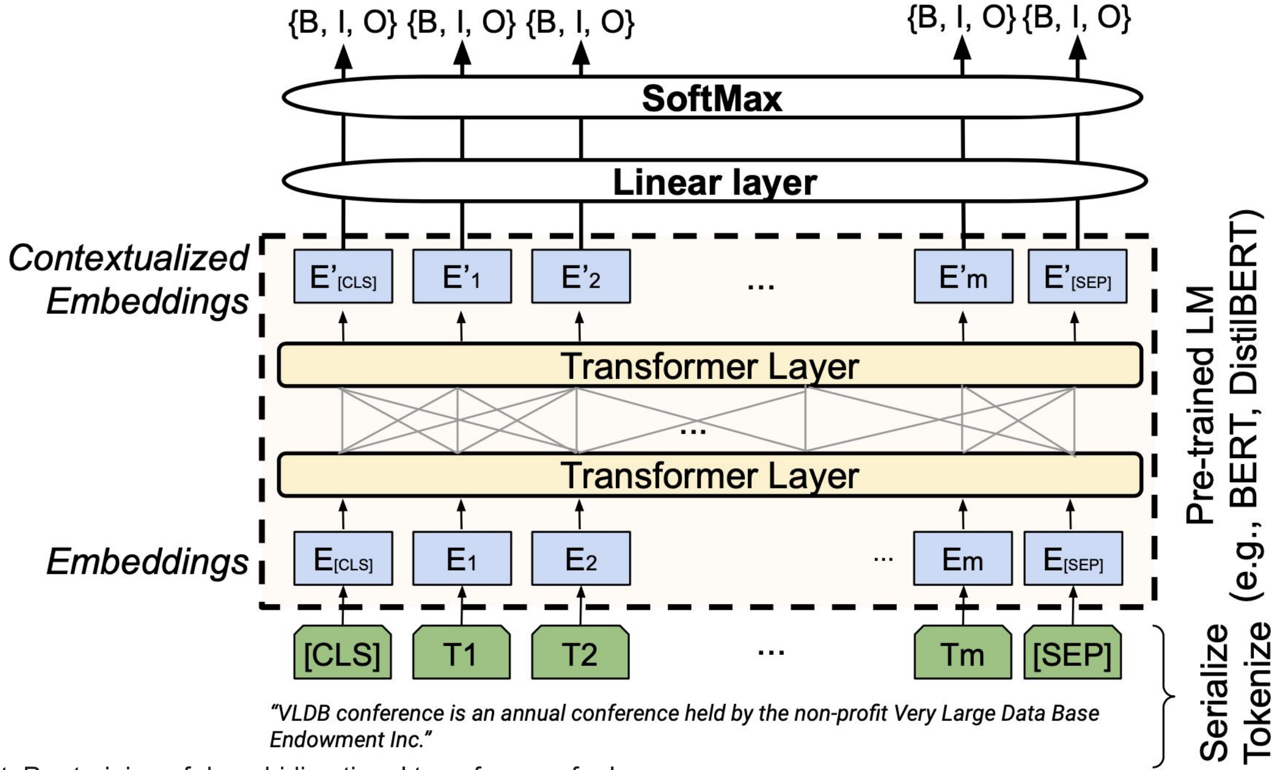



**Sequence Tagging Task**



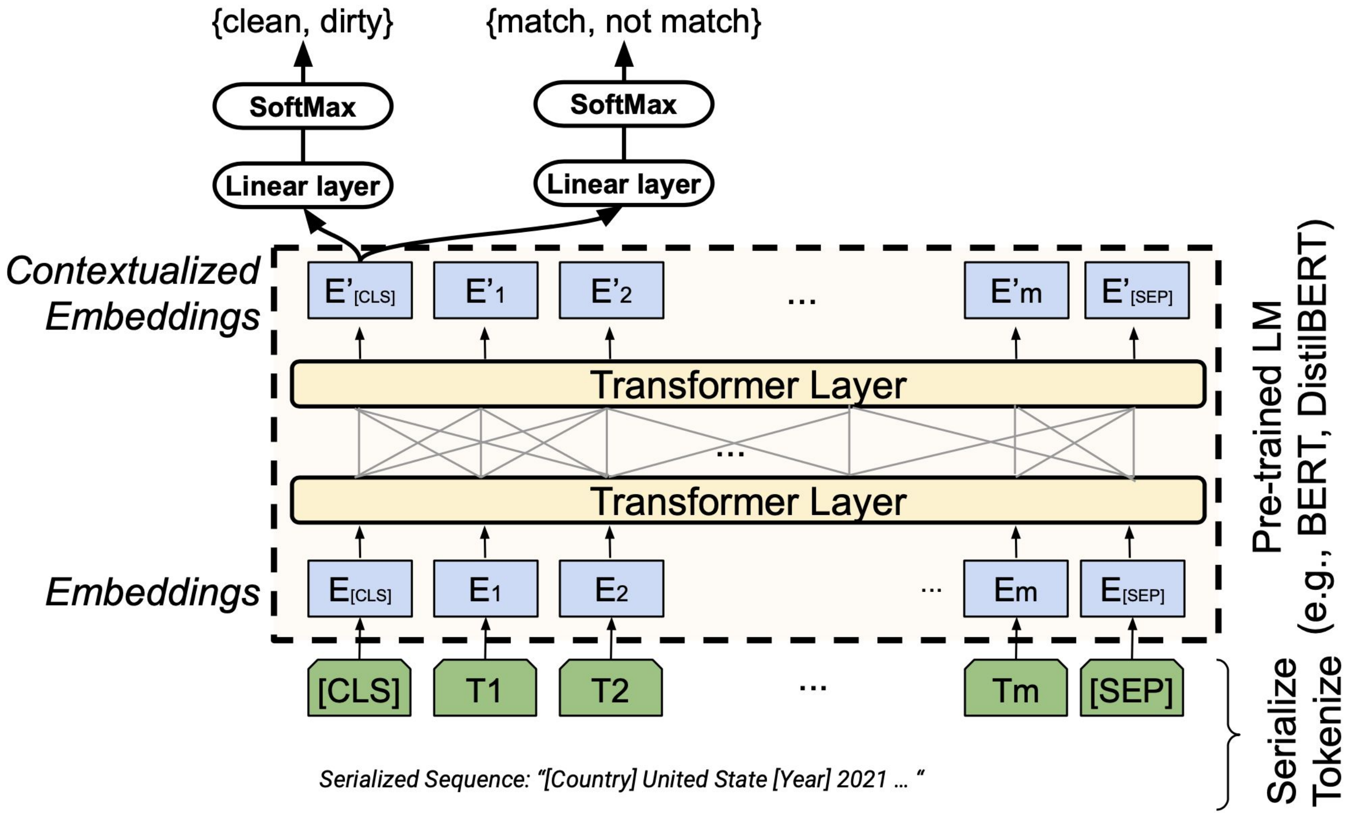

**Sequence (pair) Classification Task**

# Architecture: fine-tuning LMs (Sequence Tagging)




Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

# Architecture: fine-tuning LMs (Sequence Classification)



# Part I: DA for Data Management

Which data management tasks can be benefited from data augmentation?

What are the basic data augmentation operators?

# Basic DA operators for NLP

- Word replacement (**TR**)

"VLDB conference is an ~~annual~~ conference held by the non-profit Very Large Data Base Endowment Inc."

yearly

- Word insertion (**INS**)

"VLDB conference is an annual conference held by the non-profit Very Large Data Base Endowment Inc."

international

- Word deletion (**DEL**)

"VLDB conference is an annual conference held by ~~the~~ non-profit Very Large Data Base Endowment Inc."

Wei, Jason, et al. "Eda: Easy data augmentation techniques for boosting performance on text classification tasks." *arXiv preprint arXiv:1901.11196* (2019)

# Basic DA operators [Snippext]

| | |
|---|---|
| TR | Replace non-target token with a new token. |
| INS | Insert before or after a non-target token with a new token. |
| DEL | Delete a non-target token. |
| SW | Swap two non-target tokens. |
| **SPR** | Replace a target span with a new span. |

Token-level

Span-level

“Snippext achieves SOTA results in multiple opinion mining tasks with **only half the amount of training data** used by SOTA techniques.” -- Snippext

Miao, Zhengjie, et al. "Snippext: Semi-supervised opinion mining with augmented data." *Proceedings of The Web Conference 2020*. 2020.

# Basic DA operators [HoloDetect]

Tries to learn three different DA transformations (operators):

Remove Character: $2 \longmapsto \varnothing$

Exchange Character: $\mathrm{Denmark} \longmapsto \mathrm{United\ States}$

| Year | City | Country | Link |
|---|---|---|---|
| **201** | Copenhagen | **United States** | http://vldb.org/2021/ |
| 2020 | Tokyo | Japan | https://vldb2020.org/ |
| 2019 | Los Angeles, California | United States | https://vldb.org/2019/ |
| 2018 | Rio de Janeiro | Brazil | http://vldb2018.lncc.br |
| 2017 | **Munichm** | Germany | http://www.vldb.org/2017/ |

Add Character: $\varnothing \longmapsto m$

"We show that data augmentation yields an average improvement of **20 F1 points** while it requires access to **3× fewer labeled examples** compared to other ML approaches." -- HoloDetect

# Basic DA operators [Ditto]

| Avg citations | Conference |
|---|---|
| 31.6 | VLDB—Very Large Data Bases |
| 30.9 | BioMED—Biomedical Engineering |
| 30.9 | IEEE TRANS ROBOTICS AUTOMAT—IEEE Transactions on Robotics and Automation |
| 30.6 | CRYPTO—International Crytology Conference |
| 30.1 | PAMI—IEEE Transactions on Pattern Analysis and Machine Intelligence |

**serialize**

[COL] avg citation [VAL] 31.6 [COL] conference [VAL] VLDB—Very Large Data Bases

**span deletion**

[COL] ~~avg citation~~ [VAL] 31.6 [COL] conference [VAL] VLDB—Very Large Data Bases

**span shuffle**

[COL] **citation avg** [VAL] 31.6 [COL] conference [VAL] VLDB—Very Large Data Bases

Li, Yuliang, et al. "Deep entity matching with pre-trained language models." *Proceedings of the VLDB Endowment* 14.1 (2020): 50-60.

# Basic DA operators [Ditto]

| Avg citations | Conference |
|---|---|
| 31.6 | VLDB—Very Large Data Bases |
| 30.9 | BioMED—Biomedical Engineering |
| 30.9 | IEEE TRANS ROBOTICS AUTOMAT—IEEE Transactions on Robotics and Automation |
| 30.6 | CRYPTO—International Crytology Conference |
| 30.1 | PAMI—IEEE Transactions on Pattern Analysis and Machine Intelligence |

**serialize**

[COL] avg citation [VAL] 31.6 [COL] conference [VAL] VLDB_Very Large Data Bases

**attribute deletion**

~~[COL] avg citation [VAL] 31.6~~ [COL] conference [VAL] VLDB_Very Large Data Bases

**attributes shuffle**

[COL] conference [VAL] VLDB_Very Large Data Bases
[COL] citation avg [VAL] 31.6

Li, Yuliang, et al. "Deep entity matching with pre-trained language models." *Proceedings of the VLDB Endowment* 14.1 (2020): 50-60.

# Basic DA operators [Ditto]

| Avg citations | Conference |
|---|---|
| 31.6 | VLDB—Very Large Data Bases |
| 30.9 | BioMED—Biomedical Engineering |
| 30.9 | IEEE TRANS ROBOTICS AUTOMAT—IEEE Transactions on Robotics and Automation |

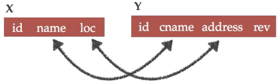
| Name | Impact |
|---|---|
| SIGMOD: ACM SIGMOD Conf on Management of Data | 0.99 |
| VLDB: Very Large Data Bases | 0.99 |
| KDD: Knowledge Discovery and Data Mining | 0.99 |

**swap entries**

| Name | Impact |
|---|---|
| SIGMOD: ACM SIGMOD Conf on Management of Data | 0.99 |
| VLDB: Very Large Data Bases | 0.99 |
| KDD: Knowledge Discovery and Data Mining | 0.99 |

| Avg citations | Conference |
|---|---|
| 31.6 | VLDB—Very Large Data Bases |
| 30.9 | BioMED—Biomedical Engineering |
| 30.9 | IEEE TRANS ROBOTICS AUTOMAT—IEEE Transactions on Robotics and Automation |

Li, Yuliang, et al. "Deep entity matching with pre-trained language models." *Proceedings of the VLDB Endowment* 14.1 (2020): 50-60.




# Basic DA operators [ADnEV]

| Year | City | Country | Link |
| --- | --- | --- | --- |
| 2021 | Copenhagen | Danmark | http://vldb.org/2021/ |
| ... | ... | ... | ... |

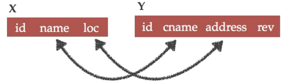
| Time | Location | Website |
| --- | --- | --- |
| 2019 | Amsterdam | [1] |
| ... | ... | ... |

S1 S2

$$
\begin{array}{c|ccc}
 & S1.time & S1.location & S1.website \\
\hline
S1.year & \mathbf{0.73} & \mathbf{0.1} & \mathbf{0.08} \\
S1.city & 0.12 & 0.67 & 0.12 \\
S1.country & 0.2 & 0.59 & 0.07 \\
S1.link & \mathbf{0.13} & \mathbf{0.15} & \mathbf{0.77}
\end{array}
$$

**Augment the continuous similarity matrix**

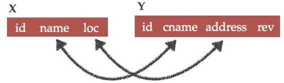
Shraga, Roee, Avigdor Gal, and Haggai Roitman. "Adnev: Cross-domain schema matching using deep similarity matrix adjustment and evaluation." *Proceedings of the VLDB Endowment* 13.9 (2020): 1401-1415.




# Basic DA operators [ADnEV]

$$
\begin{array}{l c}
 & \begin{array}{ccc} \textit{S1.time} & \textit{S1.location} & \textit{S1.website} \end{array} \\
\begin{array}{l} \textit{S1.year} \\ \textit{S1.city} \\ \textit{S1.country} \\ \textit{S1.link} \end{array} & \begin{pmatrix} 0.73 & 0.1 & 0.08 \\ 0.12 & 0.67 & 0.12 \\ 0.2 & 0.59 & 0.07 \\ 0.13 & 0.15 & 0.77 \end{pmatrix}
\end{array}
$$

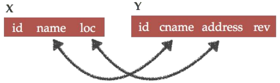
Shraga, Roee, et al. "Adnev: Cross-domain schema matching using deep similarity matrix adjustment and evaluation." Proceedings of the VLDB Endowment 13.9 (2020): 1401-1415.




# Basic DA operators [ADnEV]

$$\begin{array}{rc} & \begin{array}{ccc} \mathit{S1.time} & \mathit{S1.location} & \mathit{S1.website} \end{array} \\ \begin{array}{r} \mathit{S1.year} \\ \mathit{S1.city} \\ \mathit{S1.country} \\ \mathit{S1.link} \end{array} & \begin{pmatrix} 0.73 & 0.1 & 0.08 \\ 0.12 & 0.67 & 0.12 \\ 0.2 & 0.59 & 0.07 \\ 0.13 & 0.15 & 0.77 \end{pmatrix} \end{array} \xrightarrow{\textbf{swap rows}} \begin{array}{rc} & \begin{array}{ccc} \mathit{S1.time} & \mathit{S1.location} & \mathit{S1.website} \end{array} \\ \begin{array}{r} \mathit{S1.city} \\ \mathit{S1.year} \\ \mathit{S1.country} \\ \mathit{S1.link} \end{array} & \begin{pmatrix} 0.12 & 0.67 & 0.12 \\ 0.73 & 0.1 & 0.08 \\ 0.2 & 0.59 & 0.07 \\ 0.13 & 0.15 & 0.77 \end{pmatrix} \end{array}$$

Shraga, Roee, et al. "Adnev: Cross-domain schema matching using deep similarity matrix adjustment and evaluation." Proceedings of the VLDB Endowment 13.9 (2020): 1401-1415.



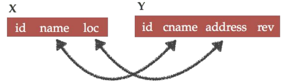

# Basic DA operators [ADnEV]

$$
\begin{array}{cc}
 & \begin{array}{ccc} \mathit{S1.time} & \mathit{S1.location} & \mathit{S1.website} \end{array} \\
\begin{array}{c} \mathit{S1.year} \\ \mathit{S1.city} \\ \mathit{S1.country} \\ \mathit{S1.link} \end{array} &
\begin{pmatrix} 0.73 & 0.1 & 0.08 \\ 0.12 & 0.67 & 0.12 \\ 0.2 & 0.59 & 0.07 \\ 0.13 & 0.15 & 0.77 \end{pmatrix}
\end{array}
\xrightarrow{\textbf{swap rows}}
\begin{array}{cc}
 & \begin{array}{ccc} \mathit{S1.time} & \mathit{S1.location} & \mathit{S1.website} \end{array} \\
\begin{array}{c} \mathit{S1.city} \\ \mathit{S1.year} \\ \mathit{S1.country} \\ \mathit{S1.link} \end{array} &
\begin{pmatrix} 0.12 & 0.67 & 0.12 \\ 0.73 & 0.1 & 0.08 \\ 0.2 & 0.59 & 0.07 \\ 0.13 & 0.15 & 0.77 \end{pmatrix}
\end{array}
$$

$$
\Downarrow
$$

$$
\begin{array}{cc}
 & \begin{array}{ccc} \mathit{S1.time} & \mathit{S1.location} & \mathit{S1.website} \end{array} \\
\begin{array}{c} \mathit{S1.city} \\ \mathit{S1.year} \\ \mathit{S1.country} \\ \mathit{S1.link} \end{array} &
\begin{pmatrix} 0.12 & 0.67 & 0.12 \\ 0.73 & 0.1 & 0.08 \\ 0.2 & 0.59 & 0.07 \\ 0.13 & 0.15 & 0.77 \end{pmatrix}
\end{array}
$$

Shraga, Roee, et al. "Adnev: Cross-domain schema matching using deep similarity matrix adjustment and evaluation." Proceedings of the VLDB Endowment 13.9 (2020): 1401-1415.




# Basic DA operators [ADnEV]

| | *S1.time* | *S1.location* | *S1.website* |
|---|---|---|---|
| *S1.year* | 0.73 | 0.1 | 0.08 |
| *S1.city* | 0.12 | 0.67 | 0.12 |
| *S1.country* | 0.2 | 0.59 | 0.07 |
| *S1.link* | 0.13 | 0.15 | 0.77 |

**swap rows**

| | *S1.time* | *S1.location* | *S1.website* |
|---|---|---|---|
| *S1.city* | 0.12 | 0.67 | 0.12 |
| *S1.year* | 0.73 | 0.1 | 0.08 |
| *S1.country* | 0.2 | 0.59 | 0.07 |
| *S1.link* | 0.13 | 0.15 | 0.77 |

| | *S1.time* | *S1.location* | *S1.website* |
|---|---|---|---|
| *S1.city* | 0.12 | 0.67 | 0.12 |
| *S1.year* | 0.73 | 0.1 | 0.08 |
| *S1.country* | 0.2 | 0.59 | 0.07 |
| *S1.link* | 0.13 | 0.15 | 0.77 |

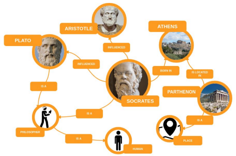
**swap columns**

| | *S1.time* | *S1.website* | *S1.location* |
|---|---|---|---|
| *S1.city* | 0.12 | 0.12 | 0.67 |
| *S1.year* | | | |
| *S1.country* | | | |
| *S1.link* | 0.13 | 0.77 | 0.15 |

**Augmented similarity matrix**

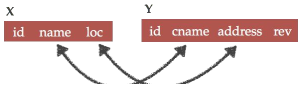
Shraga, Roee, et al. "Adnev: Cross-domain schema matching using deep similarity matrix adjustment and evaluation." Proceedings of the VLDB Endowment 13.9 (2020): 1401-1415.

# Discussion and Q&A

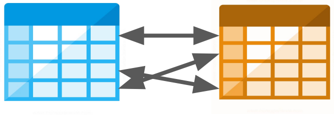



**Information Extraction**

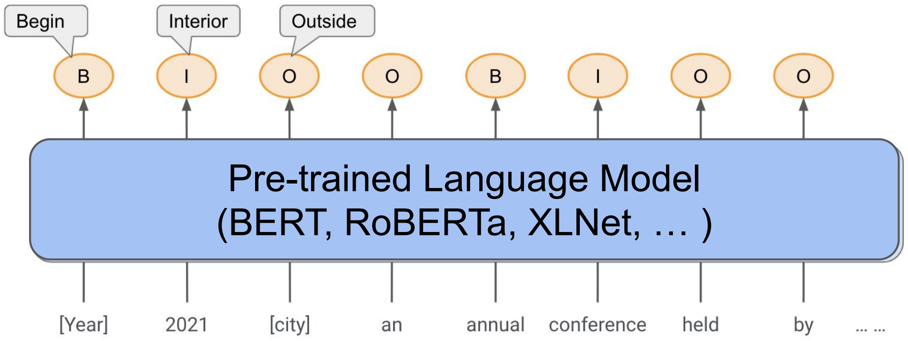
**Data Cleaning**

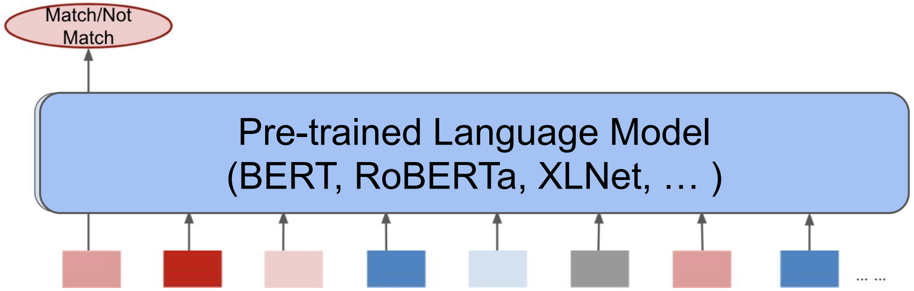



**Schema Matching**



**Entity Matching**



# Discussion and Q&A

**Operators:**
- Addition
- Replacement
- Deletion
- Swapping

**Target:**
- Character
- Token
- Span

**Maintain Semantics?**
- Yes
- No

**Operators:**
- Swapping
- … …

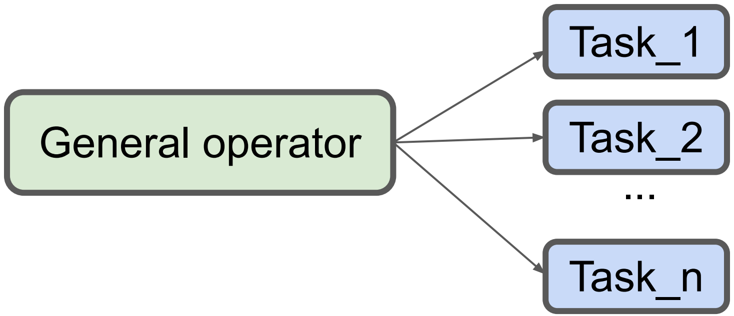
**Target:**
- Rows/Columns
- … …

# Outline

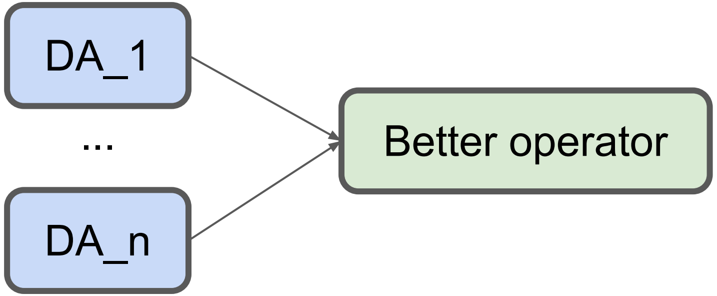
- Part I: DA for Data Management (Xiaolan Wang)
  - EM, Cleaning, schema matching, Information extraction (sequence tagging)
  - Deep learning for Data Preparation and Integration
  - Data augmentation operators
- Part II: Advanced DA (Yuliang Li)
  - Interpolation (MixUp, MixDA, and follow-up)
  - Generation (Conditional generation, GAN, InvDA)
  - Learned DA policy (AutoDA, HoloDetect, meta-learning e.g., Rotom,)
- Part III: Connection with other learning paradigms (Zhengjie Miao)
  - SSL (DA used as consistency regularization)
  - active learning (used together with DA to get more labels)
  - Weak-supervision (e.g., present Snorkel and discuss how to combine Snorkel with DA)
  - Pre-training for relational data

# Advanced DA

*Operators applicable to multiple tasks:*




*Optimize task-specific DA operators:*




*We will present:*

Interpolation-based DA

Generation-based DA

Learned DA

# Interpolation-based DA

- Instead of transforming a single example, combine two (or more) into a new one
- In computer vision:


**Intuition:**

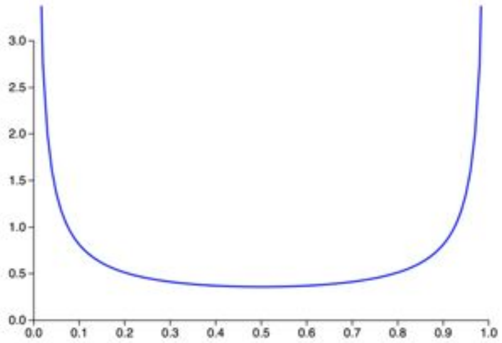
The model should make smooth transition between the two classes

Image taken from https://amitness.com/2020/05/data-augmentation-for-nlp/

# The Mixup Operator

- Randomly sample two examples: `(x1, y1)` and `(x2, y2)`:
  - Sample λ from a Beta distribution, e.g., **Beta**`(0.2, 0.2)`

```
λ ~ Beta(0.2, 0.2)
x' = λ x1 + (1 - λ) x2
y' = λ y1 + (1 - λ) y2
```



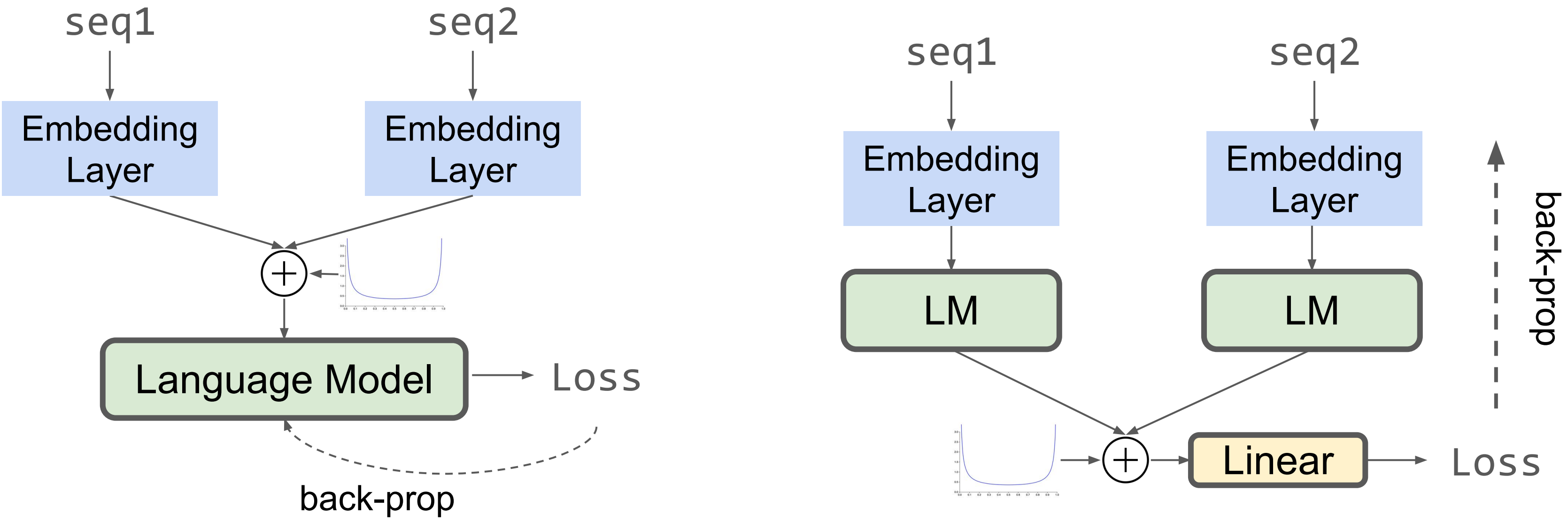

Train the model with `(x',y')`

**In CV:** improve generalization, robustness against noisy or adversarial examples

Zhang, Hongyi, et al. "mixup: Beyond Empirical Risk Minimization." *ICLR* 2018.

# How about sequence data?

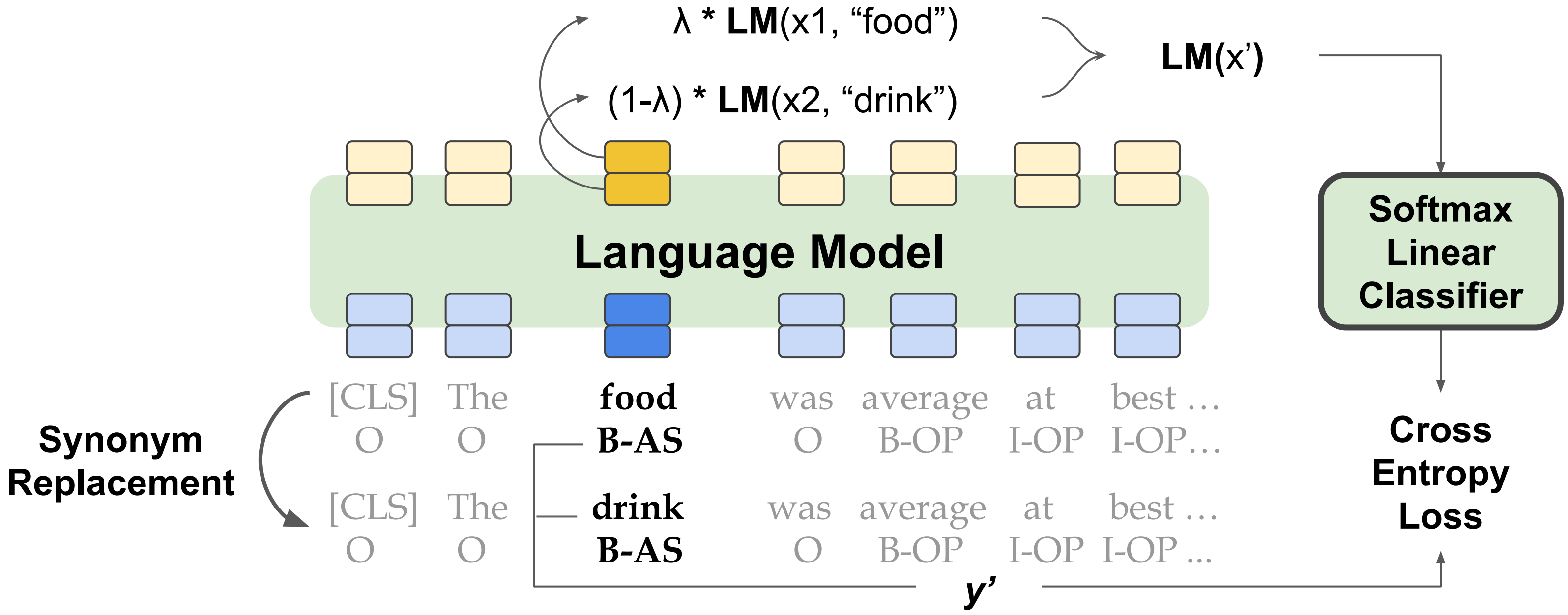
- Unlike images, textual / tabular data cannot be interpolated directly
- **Idea:** interpolate sequence representations instead




Guo, Hongyu et al.. "Augmenting data with mixup for sentence classification: An empirical study." *arXiv preprint arXiv:1905.08941* (2019).

# MixDA: Augment and Interpolate

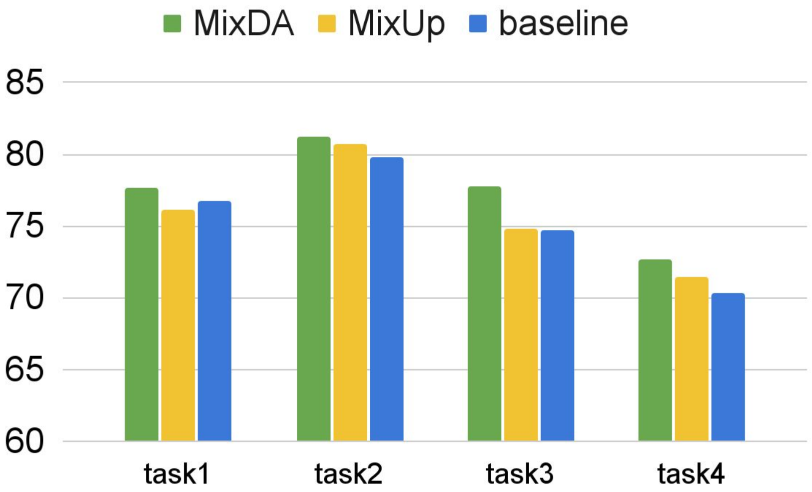
- Interpolate the original sequence with a transformed sequence
- The interpolation is in-between => less likely to have a corrupted label




Miao, Zhengjie, et al. "Snippext: Semi-supervised opinion mining with augmented data." *theWebConf* 2020.

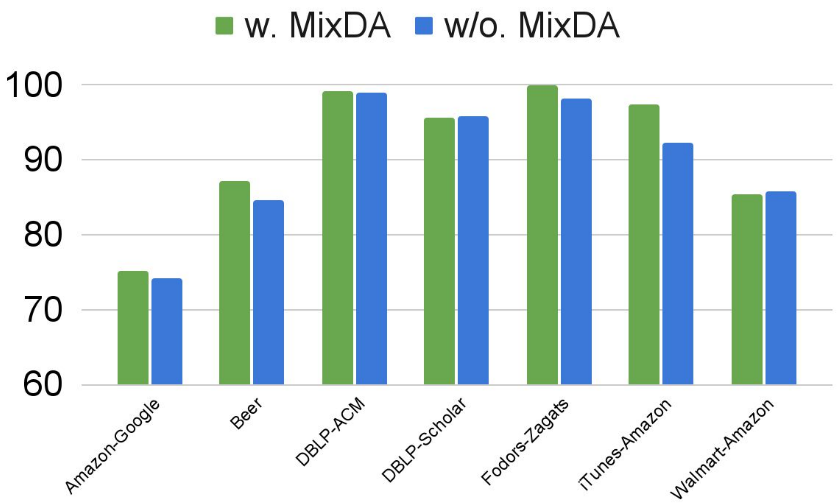
# Results of MixUp and MixDA

**IE**: both MixDA and MixUp are effective




Aspect and sentiment (restaurant, laptop)

**EM**: MixDA leads to 1.5% F1 improvement




7 EM benchmark datasets

Miao, Zhengjie, et al. "Snippext: Semi-supervised opinion mining with augmented data." *theWebConf* 2020.
Li, Yuliang, et al. "Deep entity matching with pre-trained language models.", VLDB 2021

# Generation-based DA

- Simple DA operators can generate: (1) examples with corrupted labels or (2) examples that are not diverse enough

| | | | |
|---|---|---|---|
| **Original:** | Where is the Orange Bowl? | [Intent: Location] | |
| **Replace:** | Where is the Orangish Bowl? | | Okay |
| | What is the Orange Bowl? | [Intent: Info]? | Wrong label |
| **Delete:** | Where is ~~the~~ Orange Bowl? | | Not diverse |

- Composing multiple operators:

| | | |
|---|---|---|
| **Multiple:** | the Bowl ? orangish arena | Diverse but wrong label |

# Generation-based DA

- Pre-trained generative LM can generate natural, human-like sequences

Generative LM →

GPT-2, BART, T5, etc.

prompt

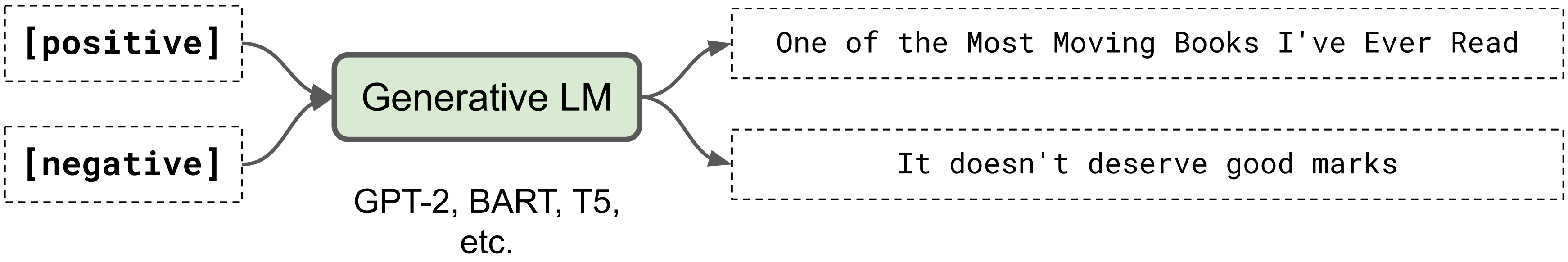
**Hello, I'm a language model,** a system model. I want to know my language so that it might be more interesting, more user-friendly, ...

- **Idea:** leverage generative LMs to generate synthetic training data
- **Challenge:** how to guide the LM to generate the examples of a target class?

Radford, Alec, et al. "Language models are unsupervised multitask learners." *OpenAI blog* 1.8 (2019): 9.

# Conditional Generation

- Add special tokens as the prompt to guide what the LM generates
- Requires labeled data for fine-tuning

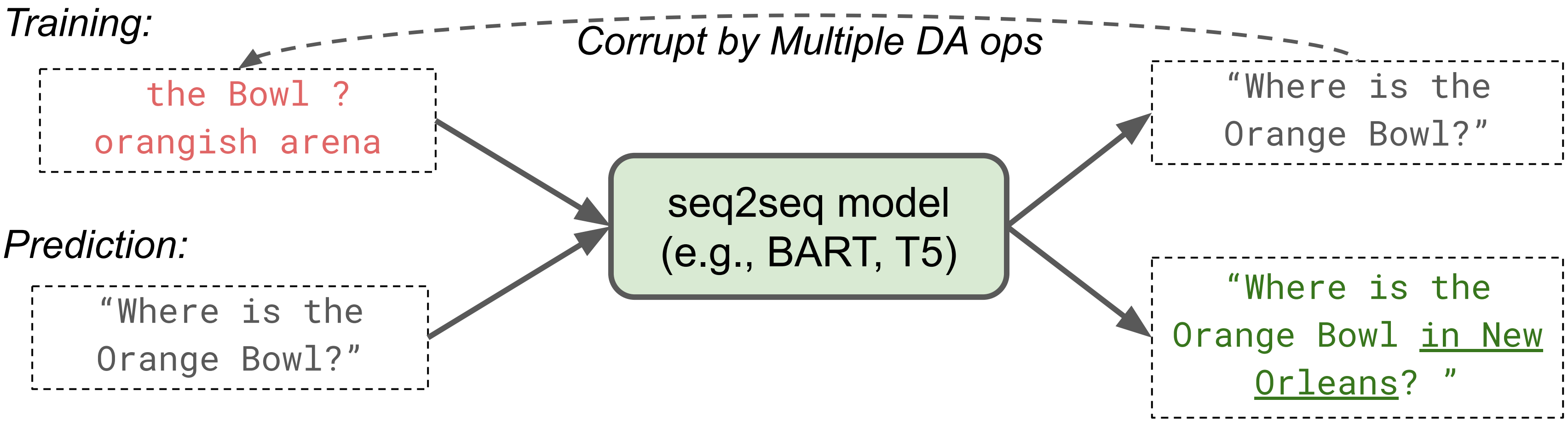



**In their experiment:** Pre-trained models (GPT-2 and BART) improve classification performance in the low-data setting (e.g., 50 - 100 labeled examples)

Kumar, Varun et al. "Data Augmentation using Pre-trained Transformer Models." *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*. 2020.

# InvDA: Seq2seq-based DA via weak-supervision

- Train a seq2seq model to augment sequences (no labels required)




By fine-tuning, the LM learns how to add information in a natural way

Miao, Zhengjie et al.. "Rotom: A Meta-Learned Data Augmentation Framework for Entity Matching, Data Cleaning, Text Classification, and Beyond." *SIGMOD* 2021.

# Examples

- Error Detection - cleaning movie data

| | |
|---|---|
| original | [COL] Name [VAL] The DUFF |
| InvDA | [COL] Name [VAL] The DUFF (The Wrestling Wizard) |
| | [COL] Name [VAL] The DUFF: The Adventures of Lena Green |
| | [COL] Name [VAL] The Duff Boy With The Devil |

InvDA generates natural “fake” movie names

# Examples (cont.)

- EM - DBLP-ACM paper matching

| | |
|---|---|
| original | [COL] title [VAL] effective timestamping in relational databases |
| InvDA | [COL] title [VAL] effective timestamping in databases |
| | [COL] title [VAL] effective timestamping in database systems |
| | [COL] title [VAL] effective timestamping in open-source databases |

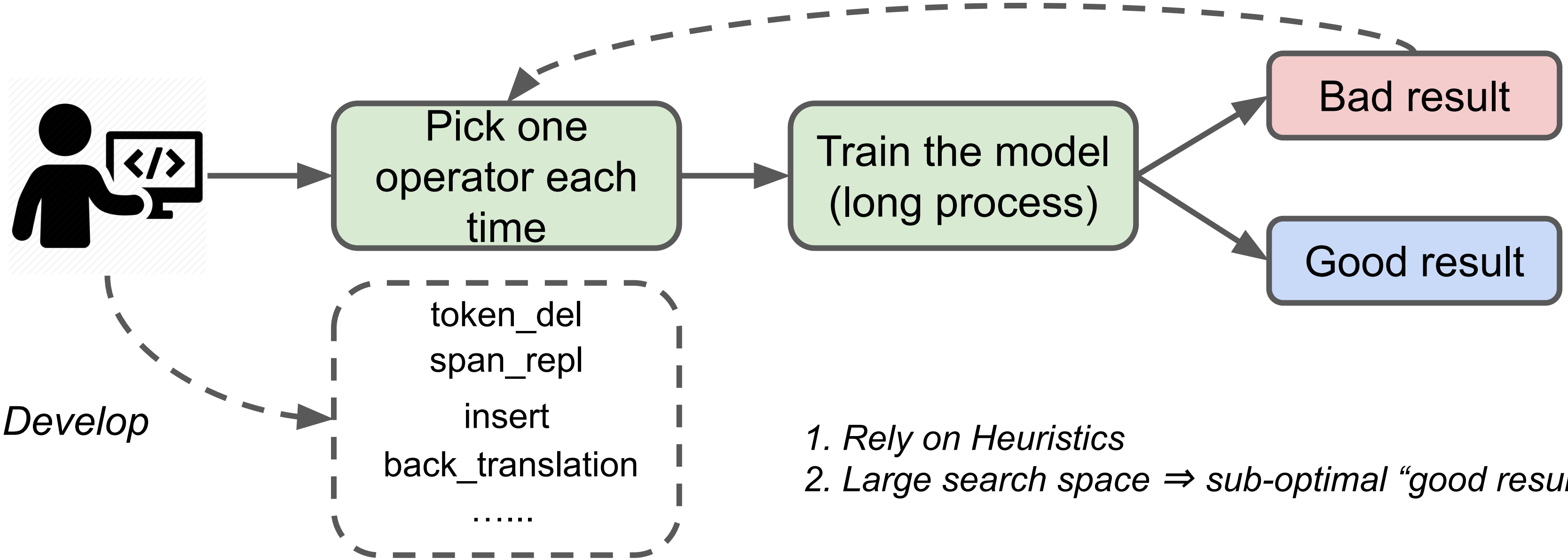
InvDA generates meaningful terms from “relational databases”

# Maybe 1 more paper?

- GAN or VAE

# Learning-based DA: Motivation

- DA introduces a whole new set of hyperparameters for tuning
- Especially when combining multiple DA operators



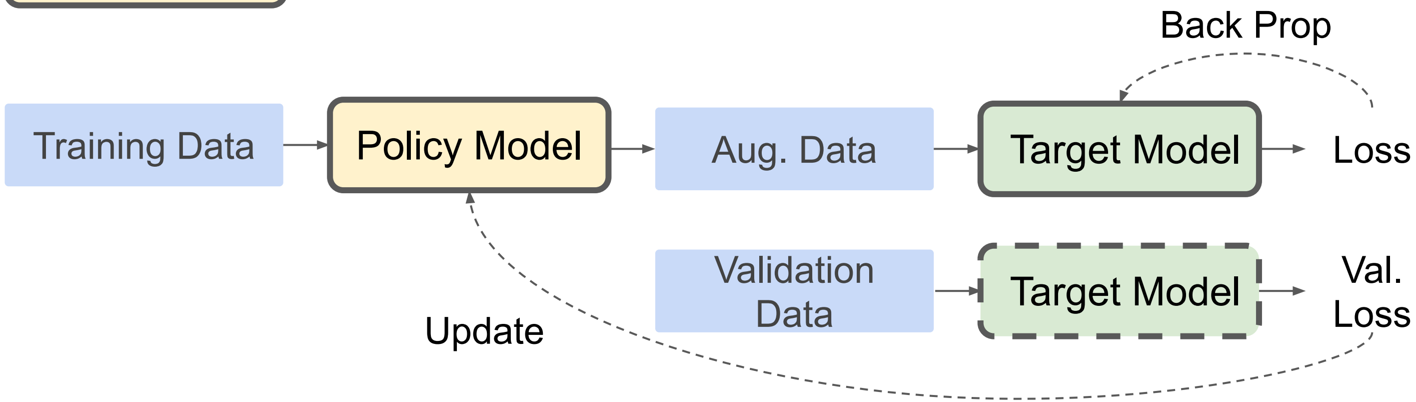

*1. Rely on Heuristics*
*2. Large search space ⇒ sub-optimal “good result”*

# Along this line of work

- **Goal:** automate the process of developing DA operators and/or combinations
- Formulate as a policy training task:

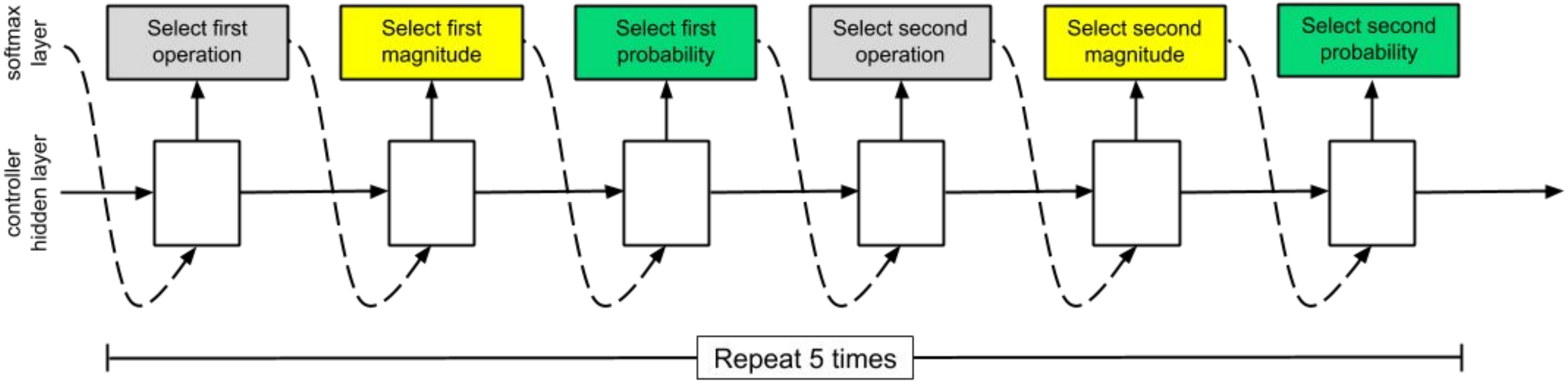
Target Model — The classification model that we want to optimize

Policy Model — A model that outputs the “optimal” DA operator or combination



# Autoaugment: Learning augmentation policies

- DA policy as a RNN; optimized the RNN via reinforcement learning to improve the target model's performance

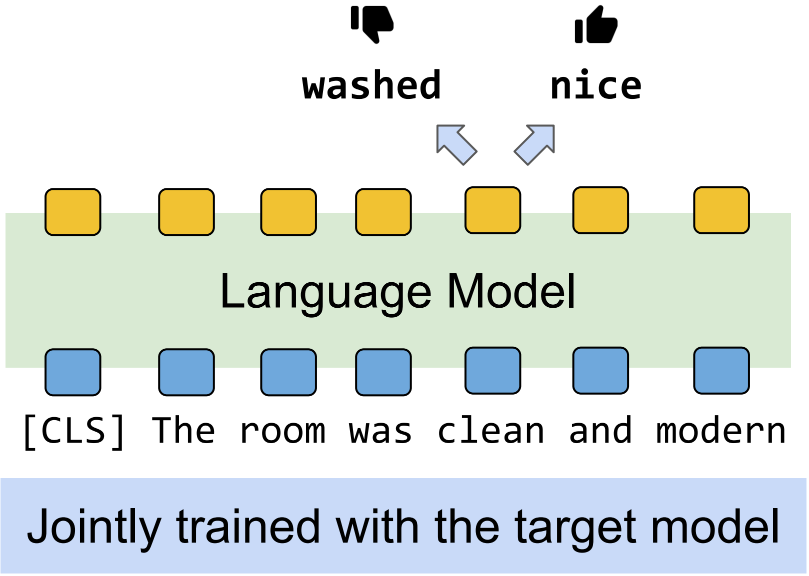


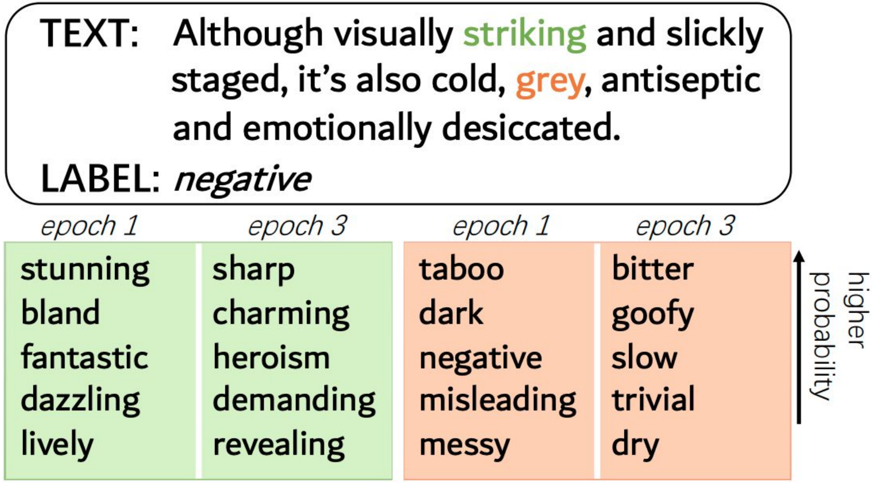

**Main challenge:** the RNN model is very expensive to train (100x of GPU hours)


Cubuk, Ekin D., et al. "Autoaugment: Learning augmentation policies from data." *CVPR 2019*
Zoph, Barret, et al. "Learning transferable architectures for scalable image recognition." CVPR 2018
Lim, Sungbin, et al. "Fast autoaugment." *NeurIPS 2019*

# Learning task-specific transformations

- Learn replacement / insertion by jointly fine-tuning a LM with the target model




Hu, Zhiting, et al. "Learning Data Manipulation for Augmentation and Weighting." *NeurIPS 2019*

# HoloDetect: learn data transformation for data cleaning

- **Goal:** learn how to inject noise to clean data to obtain more training examples
- Learn how to transform a cell value (represented by a string)
- Learn the distribution of 3 types of transformations:

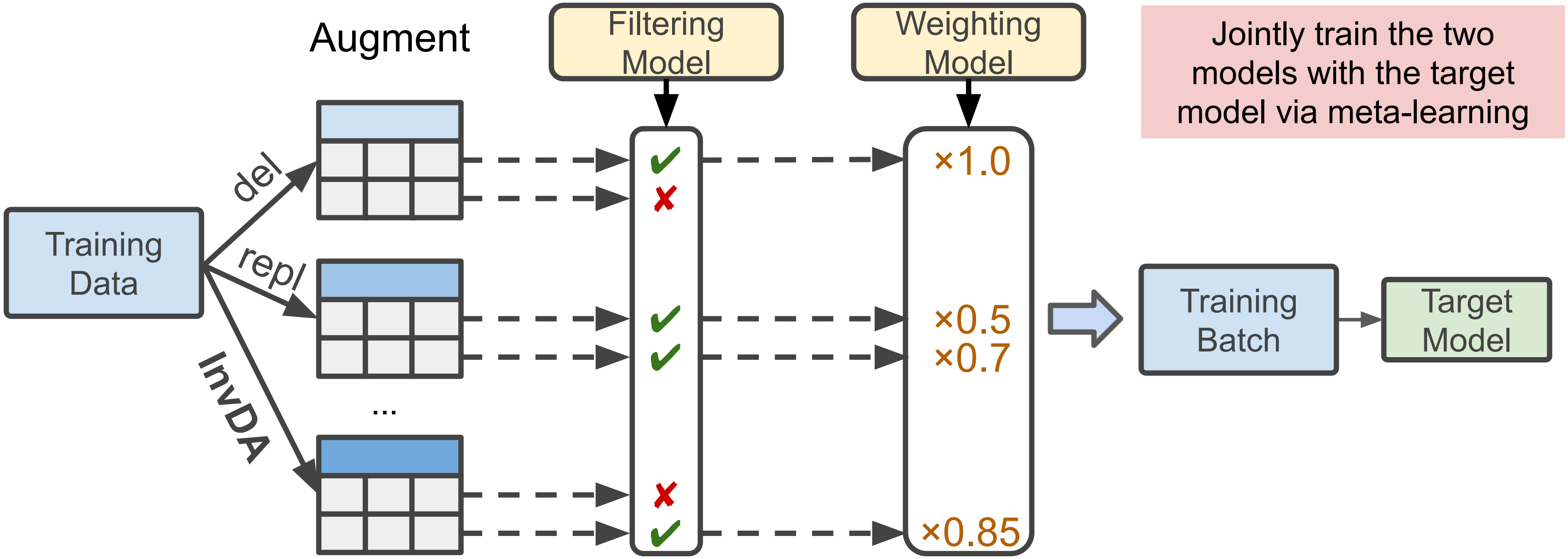
Add character | Remove character | Exchange characters

**In the paper:** DA yields an average improvement of **20 F1 points**; requires **3x** fewer labeled examples compared to other ML approaches

*Heidari, Alireza, et al. "Holodetect: Few-shot learning for error detection." SIGMOD 2019*

# Rotom: combine examples from multiple operators

- Leverage meta-learning to select and combine examples from any operators

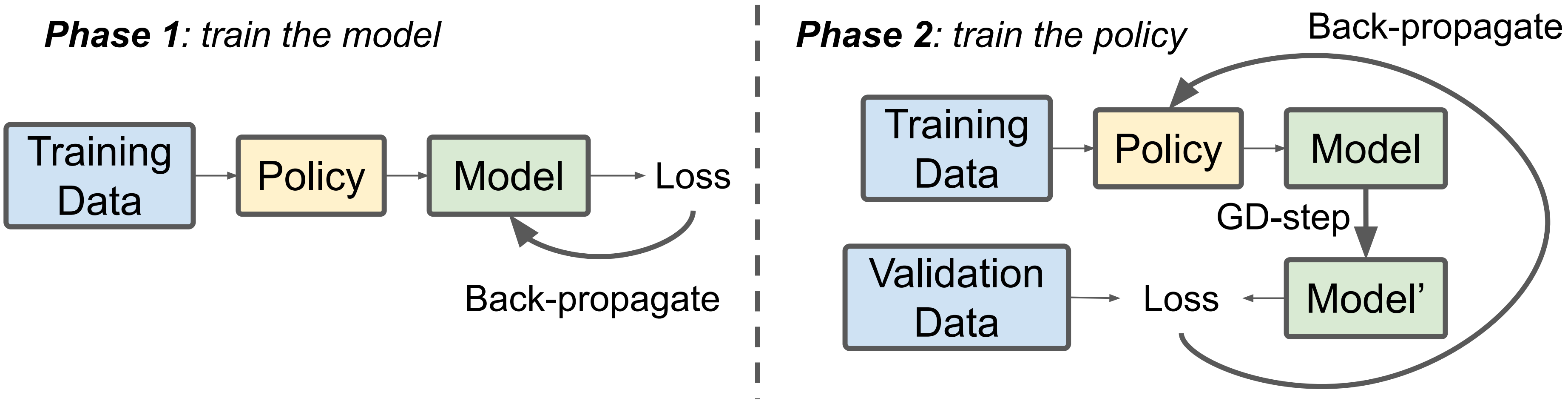



Miao, Zhengjie, et al. "Rotom: A Meta-Learned Data Augmentation Framework for Entity Matching, Data Cleaning, Text Classification, and Beyond." *SIGMOD* 2021.

# A two-phase training algorithm

Need to estimate the 2nd-order gradient

- Jointly trains the policy models and the target model

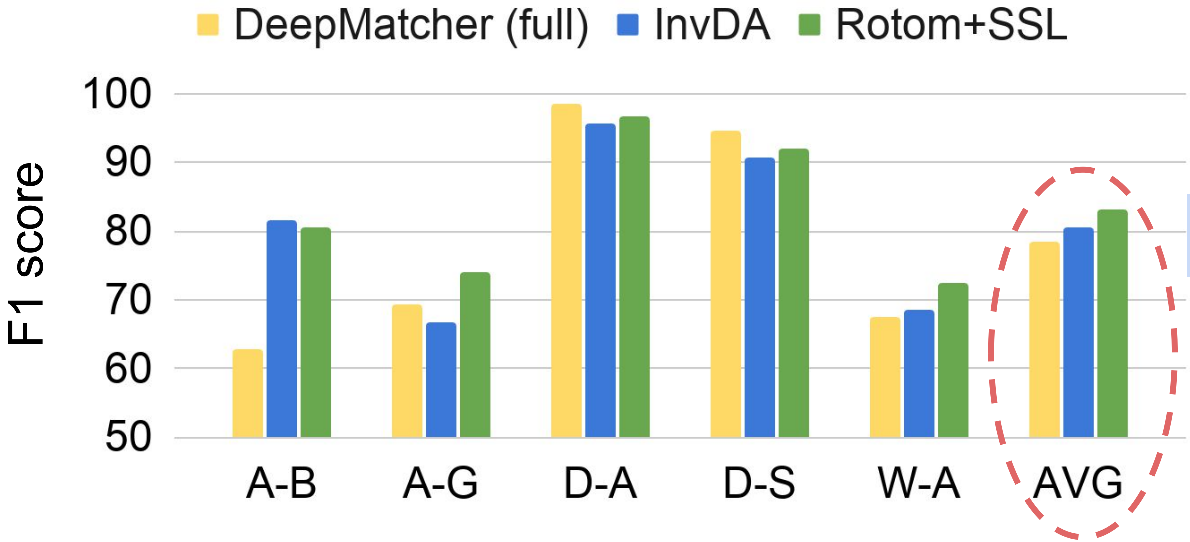



Train the DA policy such that the target model performs well on the v. set

Li, Yonggang, et al. "DADA: Differentiable Automatic Data Augmentation." arXiv preprint arXiv:2003.03780 (2020).

# Results on Entity Matching (EM)

- **DeepMatcher [SIGMOD18]** is a previous SOTA trained on the full datasets




78.56 vs. 83.16

Rotom outperforms the previous SOTA by 4.6 F1 with only 6.4% of labels

# Outline

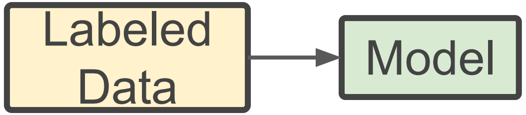
- Part I: DA for Data Management (Xiaolan Wang)
  - EM, Cleaning, schema matching, Information extraction (sequence tagging)
  - Deep learning for Data Preparation and Integration
  - Data augmentation operators
- Part II: Advanced DA (Yuliang Li)
  - Interpolation (MixUp, MixDA, and follow-up)
  - Generation (Conditional generation, GAN, InvDA)
  - Learned DA policy (AutoDA, HoloDetect, meta-learning e.g., Rotom,)
- Part III: Connection with other learning paradigms (Zhengjie Miao)
  - Semi-supervised learning (DA used as consistency regularization)
  - active learning (used together with DA for better sampling)
  - Weak-supervision (use weak-supervision for DA)
  - Representation learning for relational data (DA in pre-training tasks)

# Beyond Supervised-learning

- DA for Semi-supervised learning
- DA for Active learning
- Weak-supervision for DA
- DA for Representation learning / Representation learning for DA

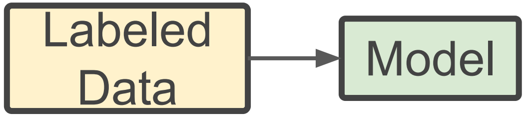
# Semi-supervised Learning

- Fully Supervised
  - Training data: (data, label), predict label for unseen data points

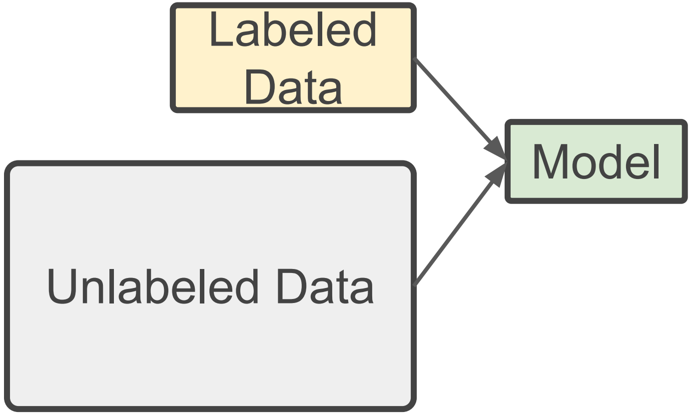


# Semi-supervised Learning

- Fully Supervised
    - Training data: (data, label), predict label for unseen data points



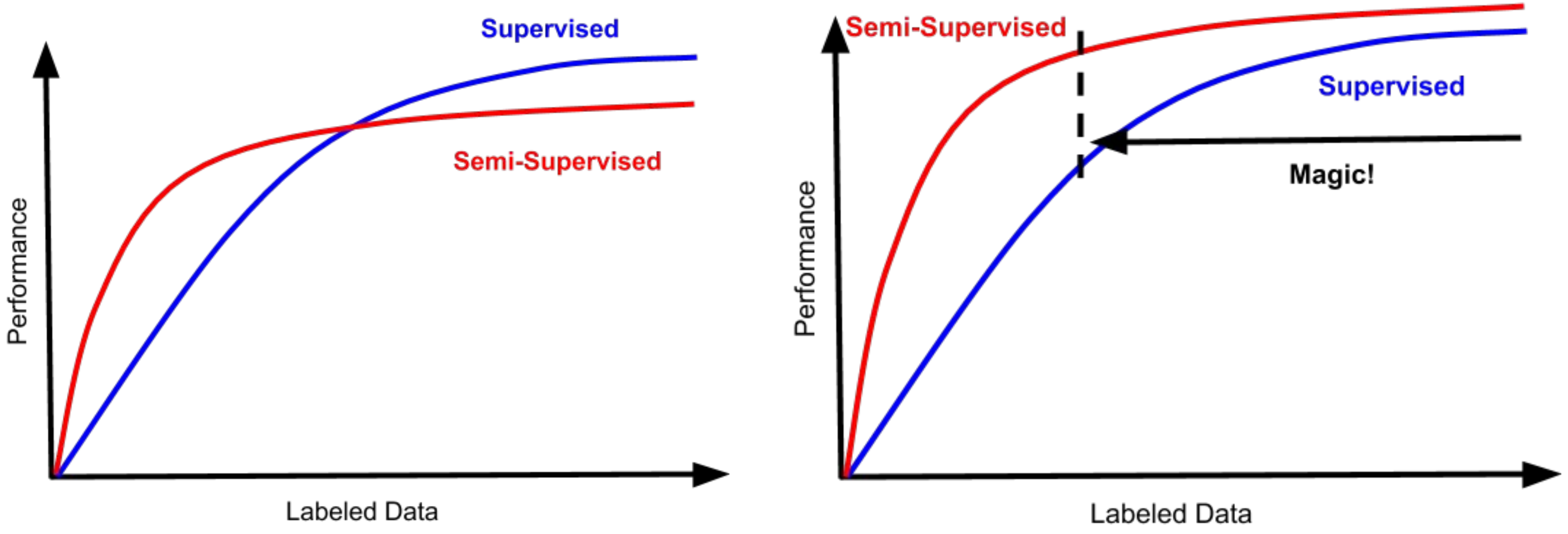

- Semi-Supervised
    - Training data: Labeled (data, label) + unlabeled (data)
    - Leverage the large collection of unlabeled data
    - Reduce the labelling cost (same goal as DA)



# Semi-Supervised Learning

- Examples in data management tasks
  - Relation extraction
    - [Carlson et al. 2010] adapts semi-supervised multitask learning and injects domain knowledge;
    - [Hu et al. 2020] uses pseudo-labeling
  - Entity Matching
    - [Kejriwal and Miranker 2015] uses ensemble learning to predict confidence scores

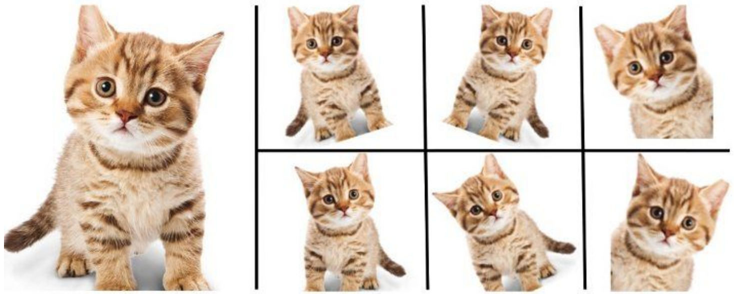
# Semi-Supervised Learning



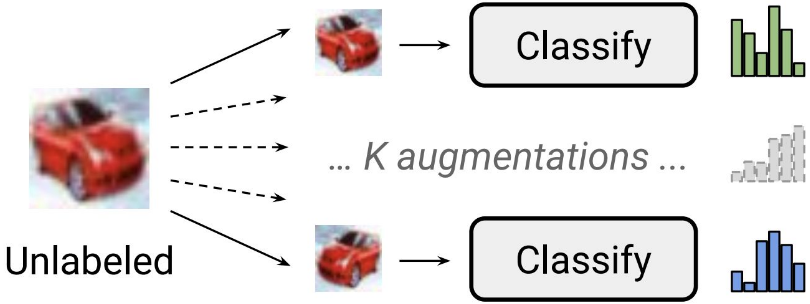

Vincent Vanhoucke, https://towardsdatascience.com/the-quiet-semi-supervised-revolution-edec1e9ad8c

# Consistency Training

- In supervised learning: label should be invariant to small noise/transformations
  - Often use DA to achieve it

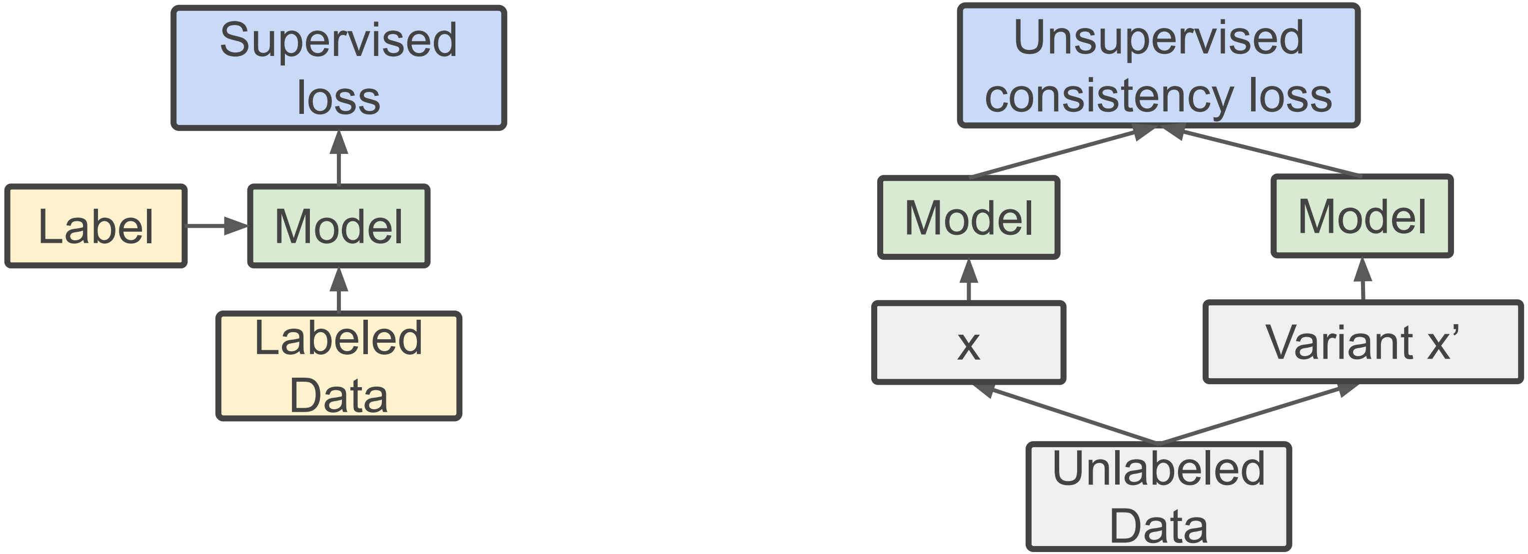


- In semi-supervised learning: model outputs for all augmentations are similar



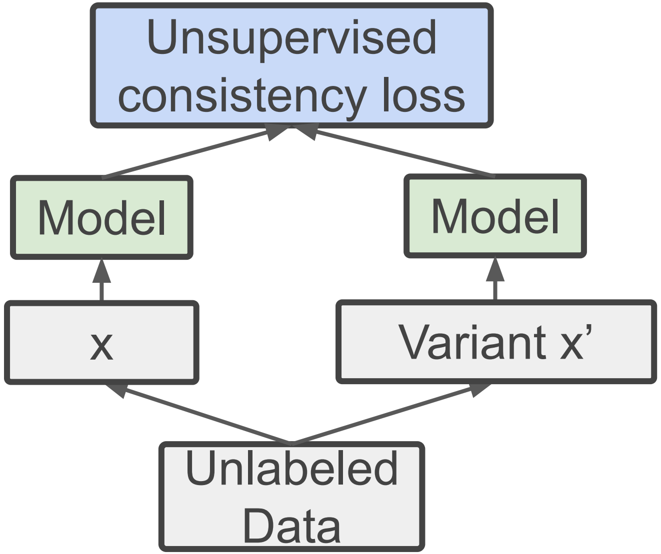
# Consistency Training

- In semi-supervised learning: model outputs for all augmentations are similar



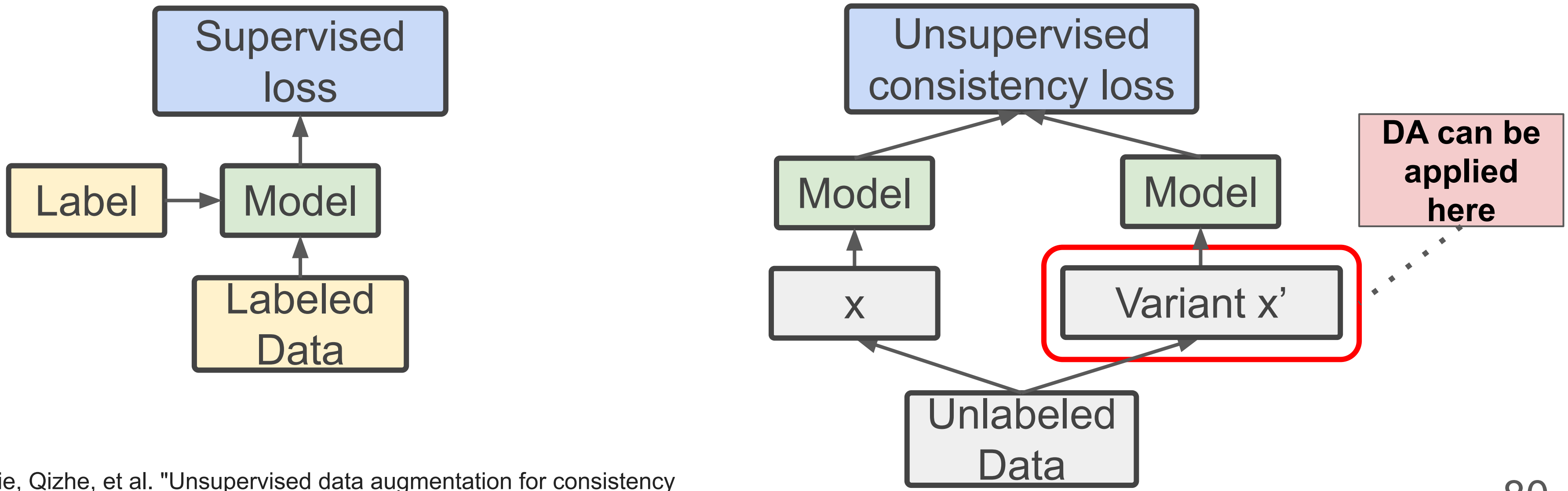
# Consistency Training

- In semi-supervised learning: model outputs for all augmentations are similar
- Minimizing the divergence between the model output distributions
  - $D(p_\theta(y|x)||p_\theta(y|aug(x)))$
  - E.g. Cross-entropy loss



# DA in Semi-Supervised Learning

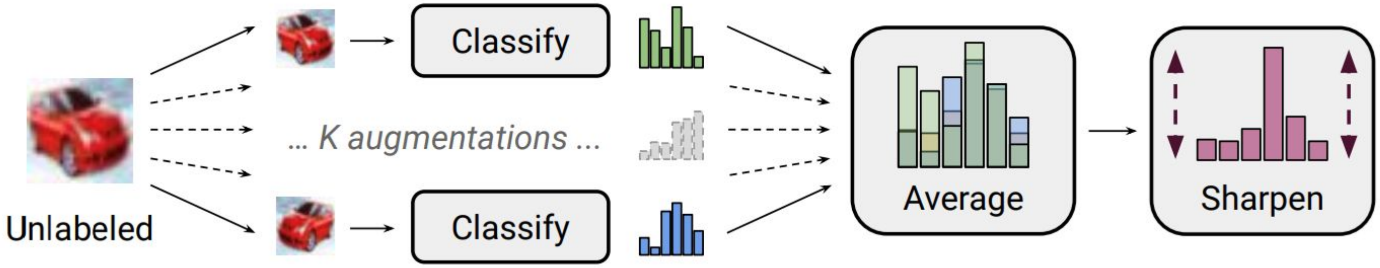
- Unsupervised data augmentation (UDA)




Xie, Qizhe, et al. "Unsupervised data augmentation for consistency training." *arXiv preprint arXiv:1904.12848* (2019).

# DA for Semi-Supervised Learning

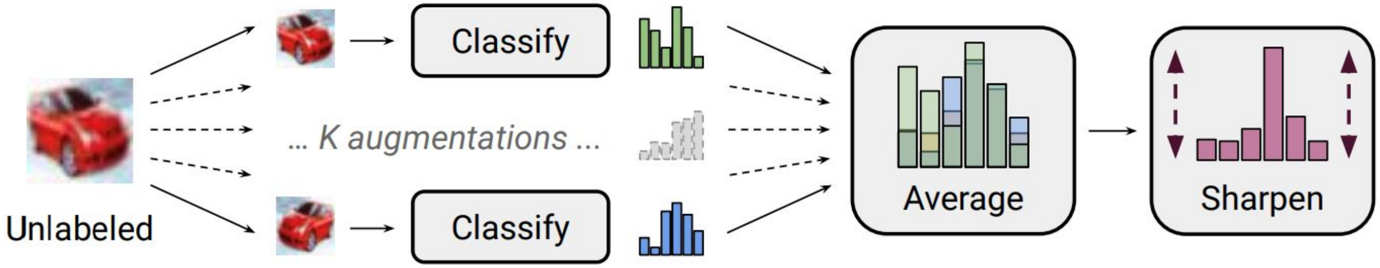
- MixMatch [Berthelot et al. 2019]
  - L, U → L*, U*  (augment both labeled and unlabeled examples)
  - Guess labels for U* by computing the average classed distributions across different U*




Berthelot, David, et al. "MixMatch: A Holistic Approach to Semi-Supervised Learning." *Advances in Neural Information Processing Systems* 32 (2019).

# DA for Semi-Supervised Learning

- MixMatch [Berthelot et al. 2019]
  - Apply MixUp to pair of examples from either (L*, L*), unlabeled (L*, U*), or (U*, U*)
  - Ask the model to make consistent predictions on different augmented U*

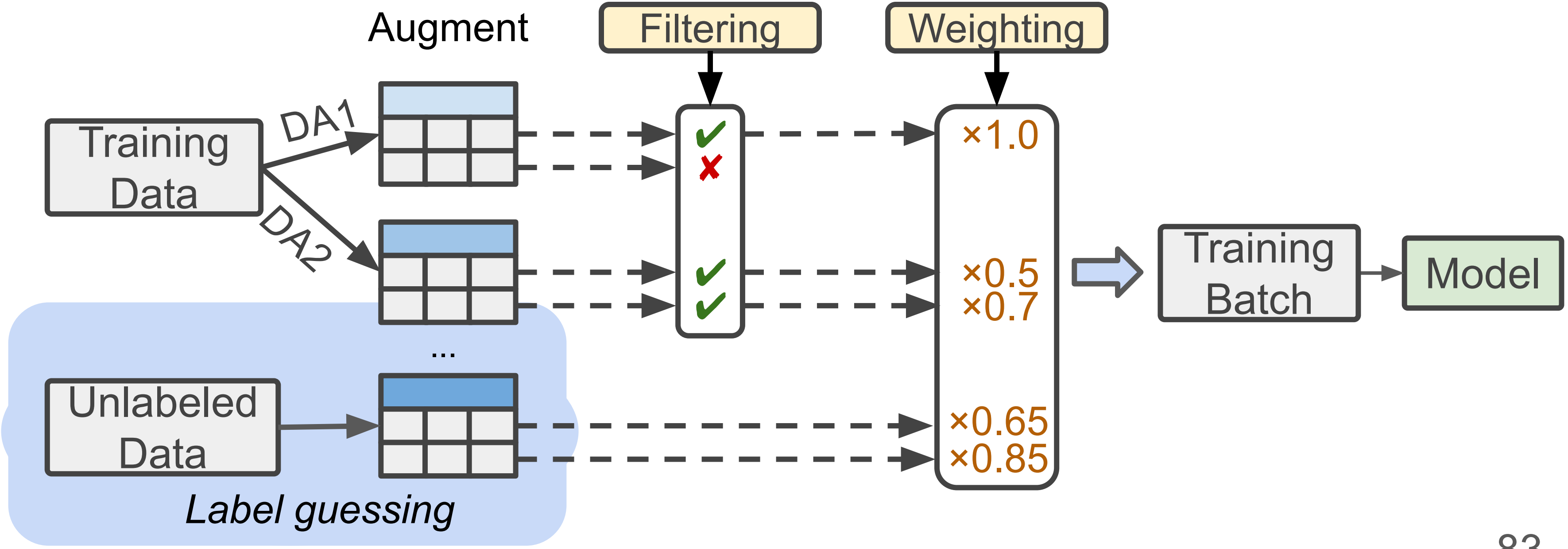



**In their experiment**: MixMatch achieves 90% accuracy on CIFAR-10 with only 250 examples (66% for the next-best-performing model)


Berthelot, David, et al. "MixMatch: A Holistic Approach to Semi-Supervised Learning." *Advances in Neural Information Processing Systems* 32 (2019).

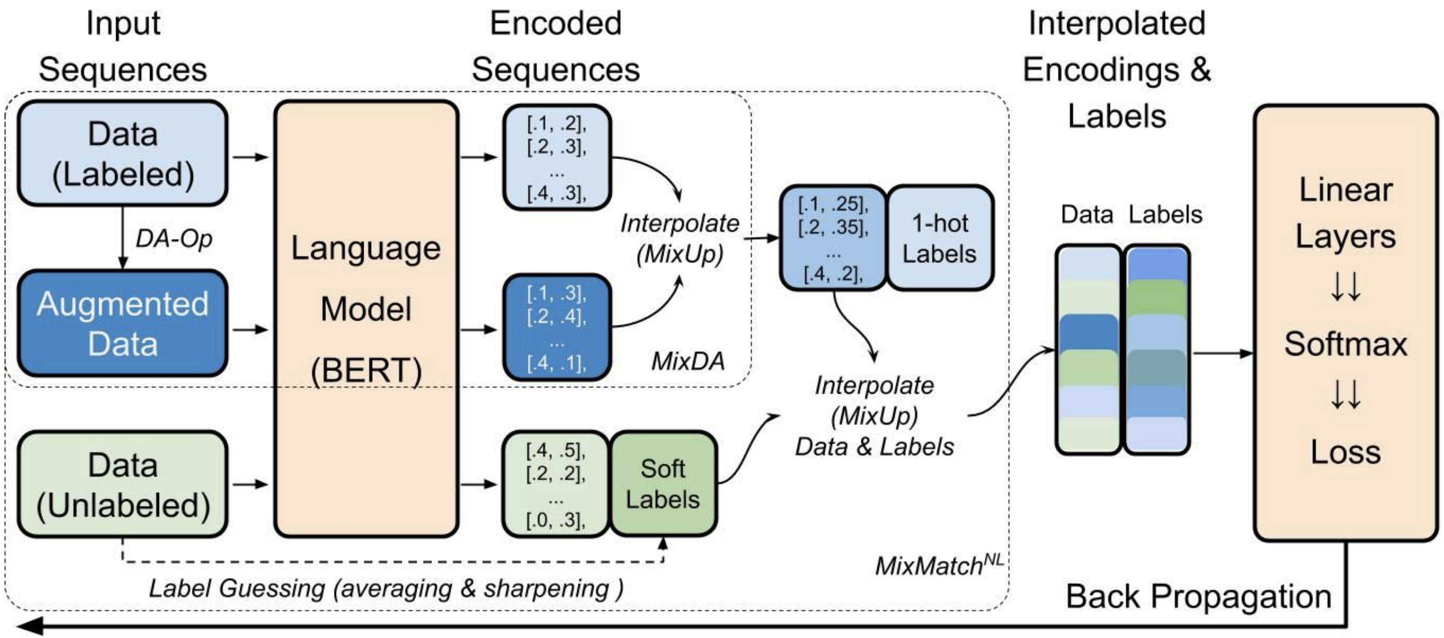
# DA in Semi-Supervised Learning

- Rotom [Miao et al. 2021]
  - Selects and combines examples generated by multiple DA ops



# DA for Semi-Supervised Learning

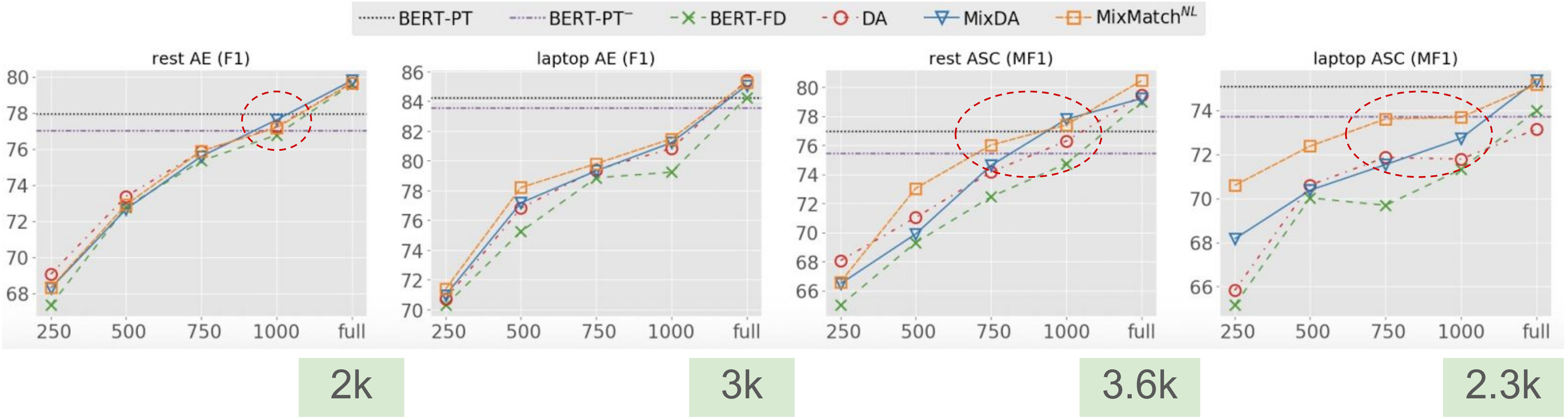
- MixMatch for information extraction [Miao et al. 2020]




Miao, Zhengjie, et al. "Snippext: Semi-supervised opinion mining with augmented data." *Proceedings of The Web Conference 2020*. 2020.

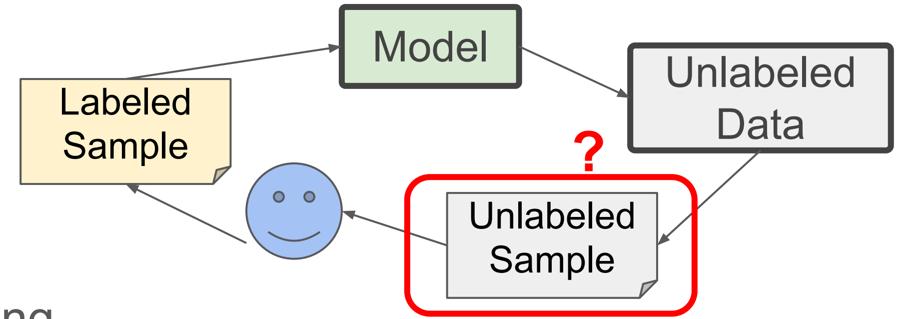
# DA for Semi-Supervised Learning

- Reach/outperform previous results with only 1/2 to 1/3 of data on opinion mining tasks



# Active Learning

- A special case of semi-supervised learning
  - Iteratively asks the user to label a few unlabeled examples



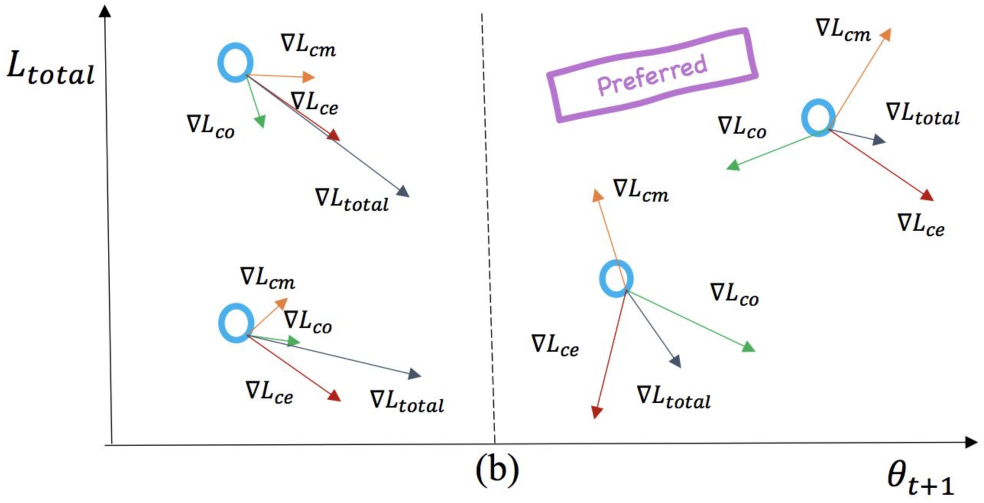

- Uncertainty sampling
  - Identify most informative unlabeled examples for users to label
  - Common approaches: least confidence / smallest margin of confidence / entropy

# Active Learning

- Popular in data management tasks to reduce the labelling budgets
  - Entity matching
    - [Kasai et al. 2019] combines transfer learning and active learning; samples both uncertain examples and high-confidence examples
    - [Jain et al. 2021] leverages LM for both blocker and matcher for more effective sampling
  - Error detection
    - Raha [Mahdavi et al. 2019] ensembles multiple error detection models to obtain feature vectors for clustering-based representative value selection

# DA for Active Learning

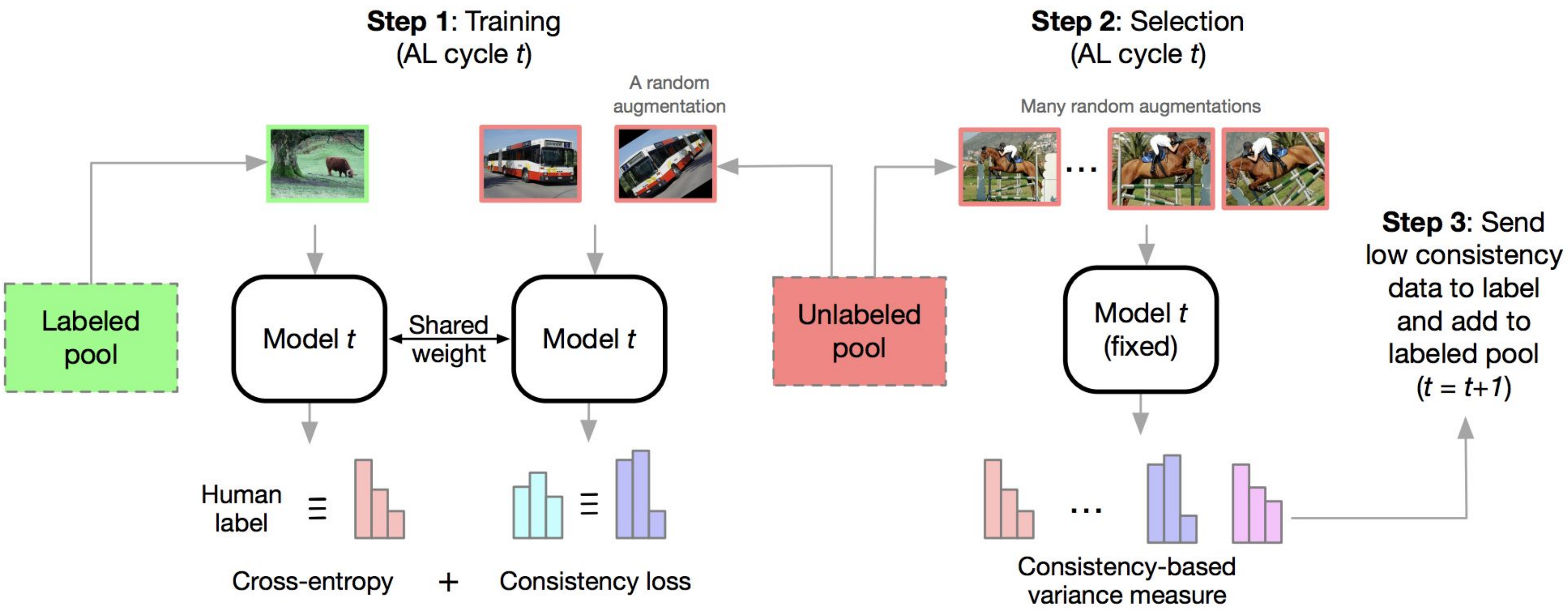
- DA for uncertainty sampling [Hong et al. 2020]
    - For an unlabeled example x, generate K augmented examples x_1, …, x_K
    - Uncertainty = mean distance between or variance of the model's inference results of every pair of (x_i, x_j)




Hong, SeulGi, et al. "Deep Active Learning with Augmentation-based Consistency Estimation." *arXiv preprint arXiv:2011.02666* (2020).

# DA for Active Learning

- DA for uncertainty sampling [Gao et al. 2020]




Gao, Mingfei, et al. "Consistency-based semi-supervised active learning: Towards minimizing labeling cost." ECCV 2020

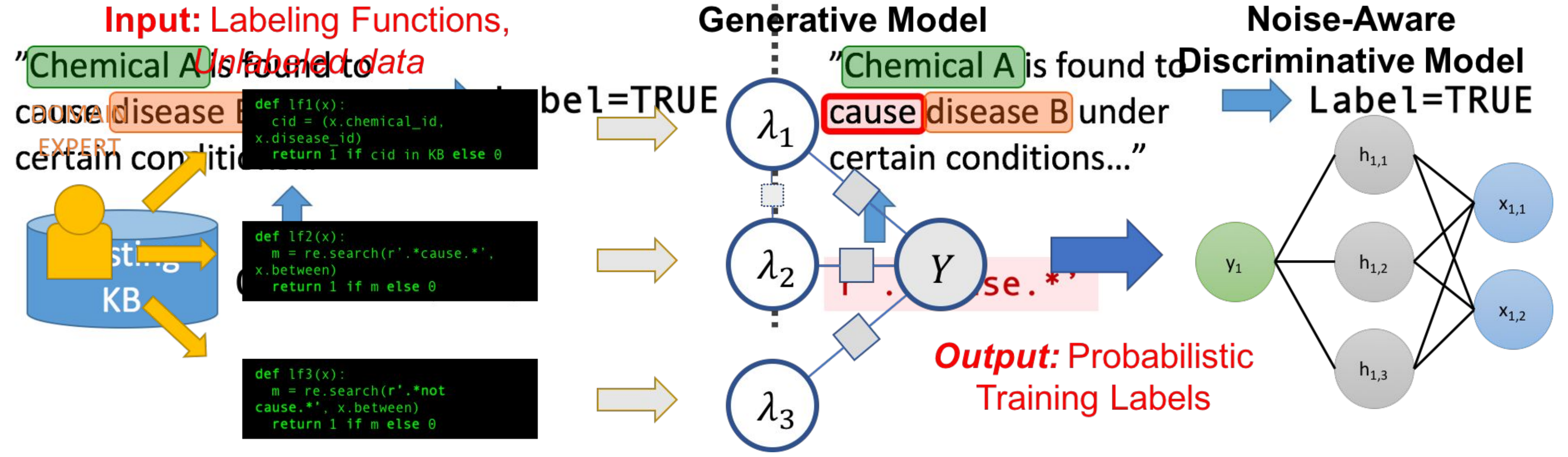
# Weak-supervision

- Use noisy sources to provide supervision signal for unlabeled data
    - External knowledge bases, crowd-sourcing, user-defined heuristics, etc.
    - Data programming: users provide programs (labeling functions) that labels a subset of the unlabeled data

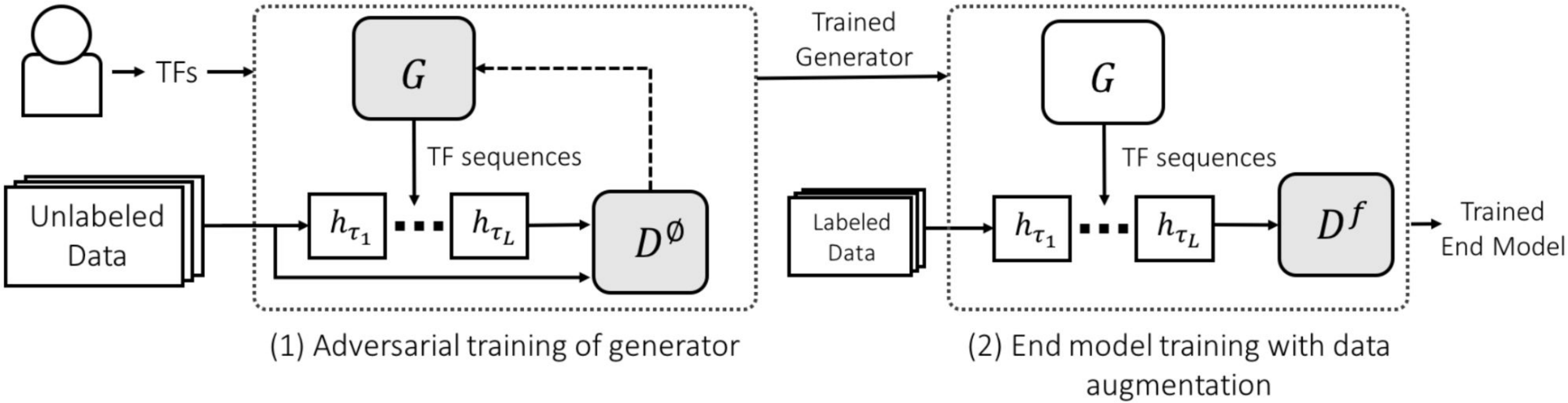
# Weak-supervision

```
def lf1(x):
  cid = (x.chemical_id, x.disease_id)
  return 1 if cid in KB else 0
```

```
def lf2(x):
  m = re.search(r'.*cause.*', x.between)
  return 1 if m else 0
```



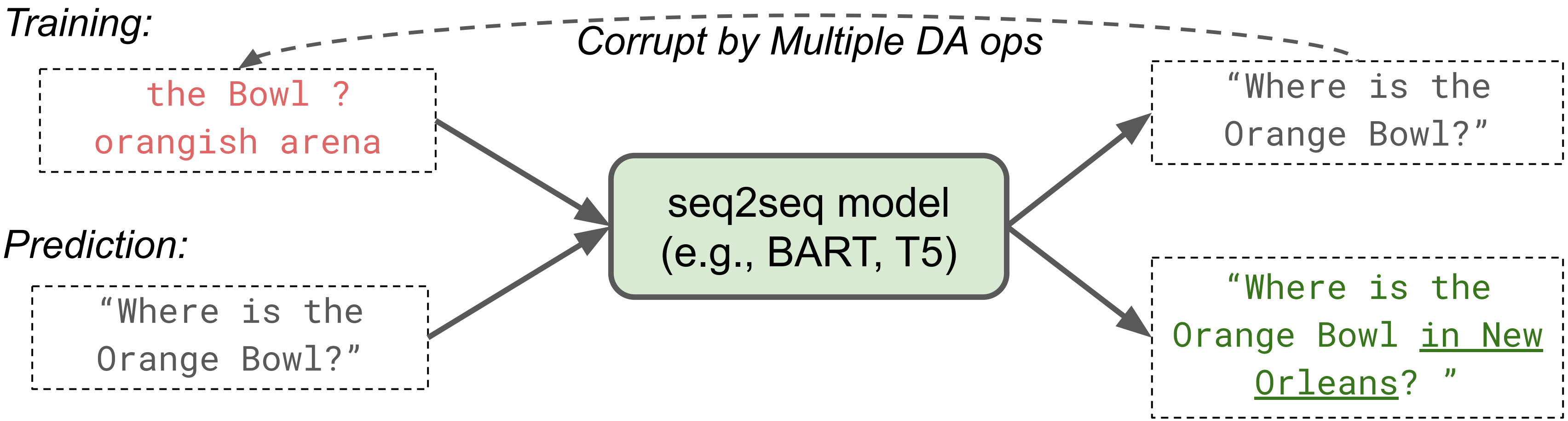
# Weak-supervision for DA: Snorkel




Ratner, Alexander J., et al. "Learning to compose domain-specific transformations for data augmentation." *Advances in neural information processing systems* 30 (2017): 3239.

# Weak-supervision for DA: InvDA

- Train a seq2seq model to augment sequences (no labels required)




By fine-tuning, the LM learns how to add information in a natural way

Miao, Zhengjie, et al. "Rotom: A Meta-Learned Data Augmentation Framework for Entity Matching, Data Cleaning, Text Classification, and Beyond." *SIGMOD* 2021.

# Representation Learning for Relational Data

- Pre-trained language representations are shown to be effective in data integration/preparation tasks (Word2Vec, fastText, BERT...)
- However, it fails to encode
  - Structure information in relational data
  - Semantics about entities

How to further improve representation learning for relational data?

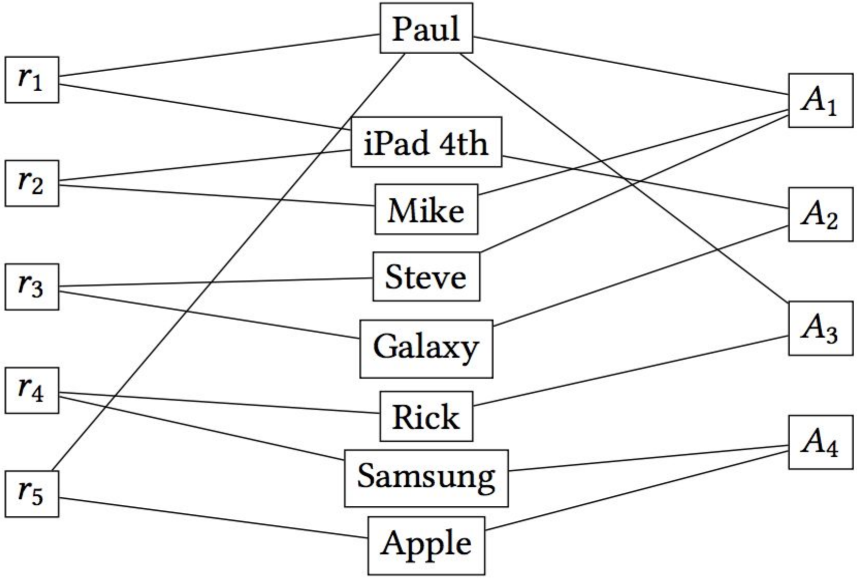
# Representation Learning for Relational Data

- Through graph embedding
    - EmbDI

**Datasets**

| | $A_1$ | $A_2$ |
|---|---|---|
| $r_1$ | *Paul* | iPad 4th |
| $r_2$ | *Mike* | iPad 4th |
| $r_3$ | Steve | Galaxy |

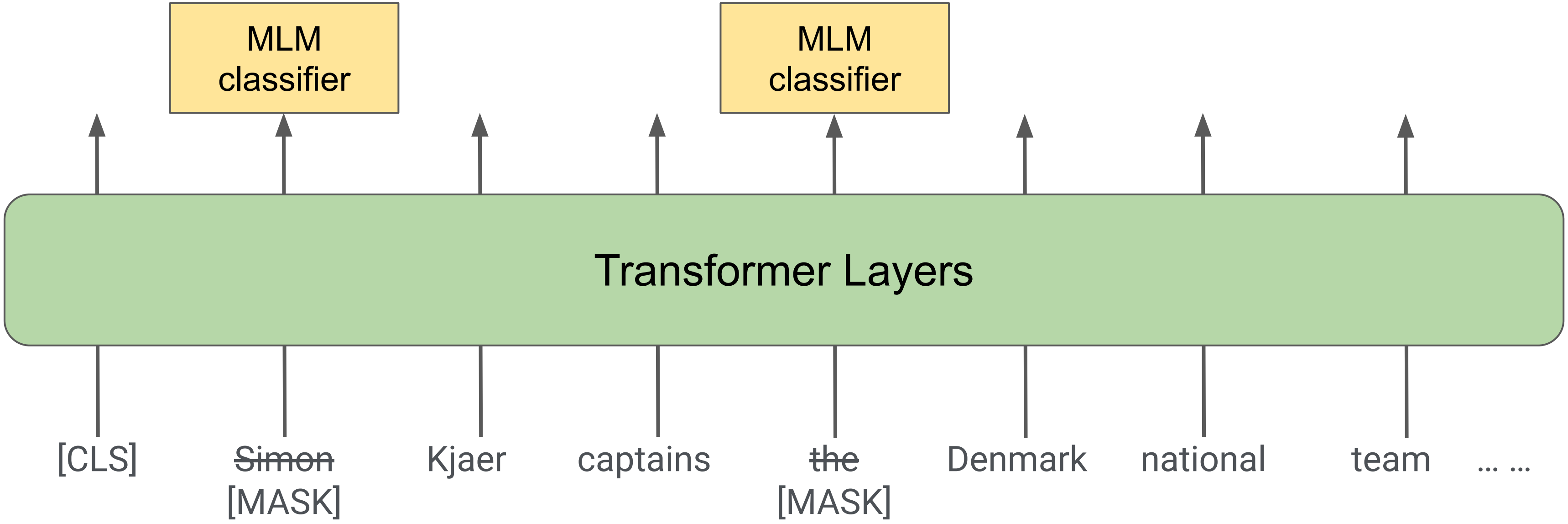
| | $A_3$ | $A_4$ |
|---|---|---|
| $r_4$ | **Rick** | Samsung |
| $r_5$ | Paul | Apple |




Cappuzzo, Riccardo, Paolo Papotti, and Saravanan Thirumuruganathan. "Creating embeddings of heterogeneous relational datasets for data integration tasks." *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 2020.

# Representation Learning for Relational Data

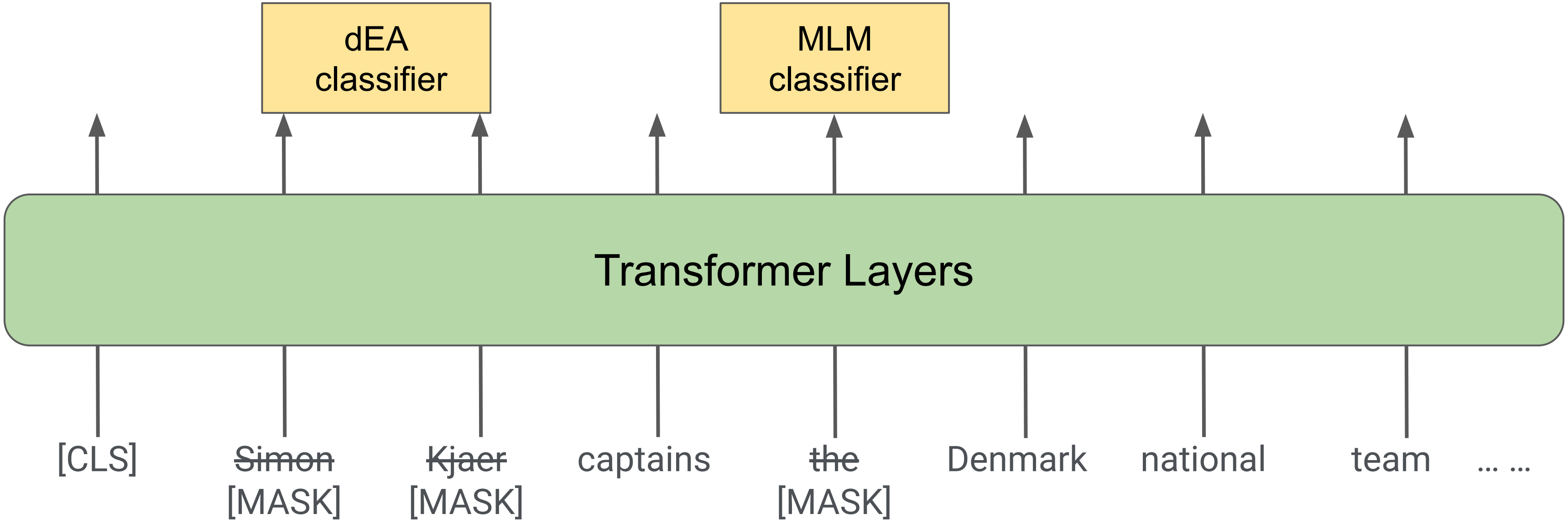
- Knowledge-enhanced Pre-trained LM
  - Recap: Masked language model in BERT




Kenton, Jacob Devlin Ming-Wei Chang, and Lee Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of NAACL-HLT*. 2019.

# Representation Learning for Relational Data

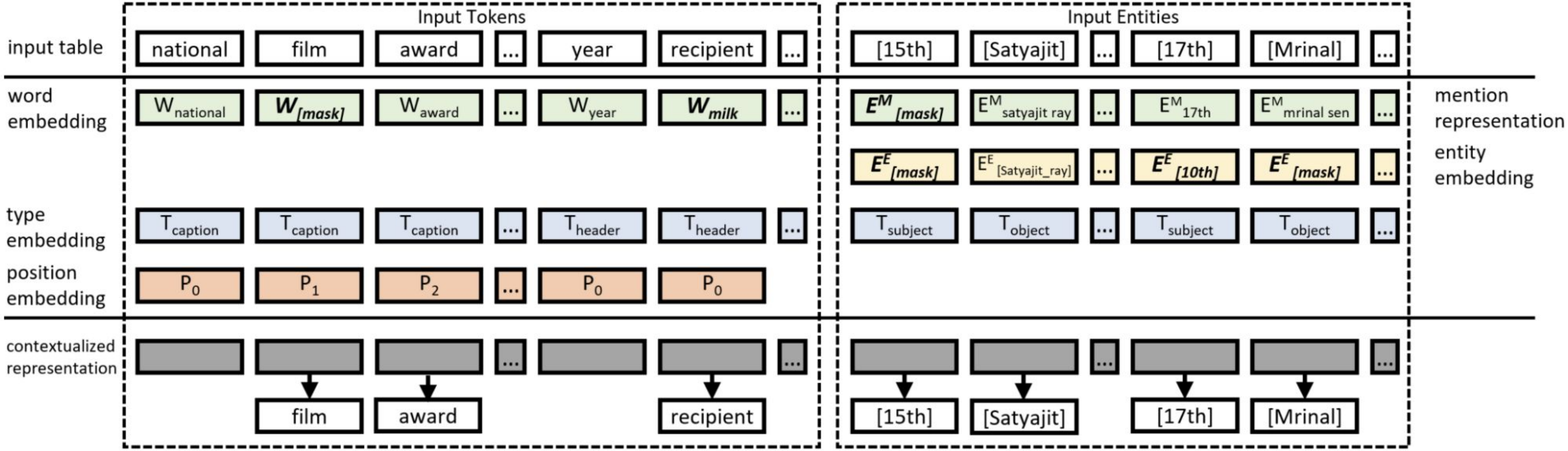
- Knowledge-enhanced Pre-trained LM
  - Denoising Entity Autoencoder [ERNIE, Zhang et al. 2019]




Zhang, Zhengyan, et al. "ERNIE: Enhanced Language Representation with Informative Entities." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019.

# Representation Learning for Relational Data

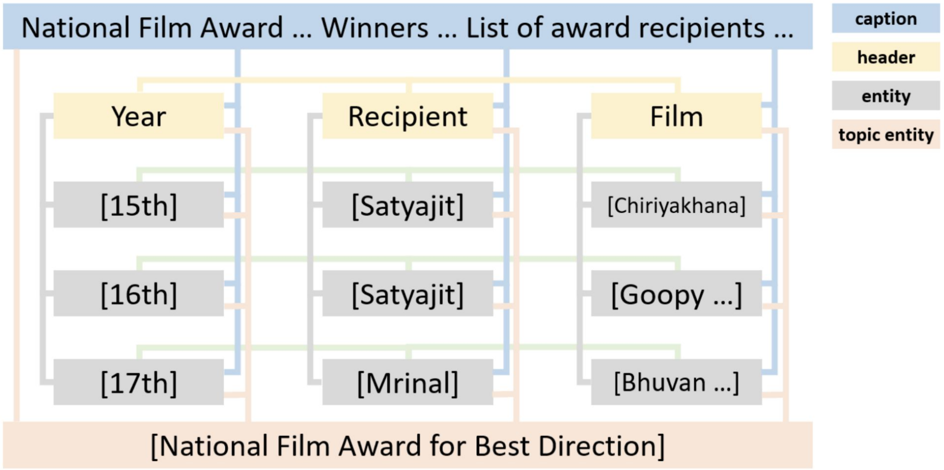
- Joint representation for NL and relational data
  - TURL [Deng et al. 2020]
    - Separate input embedding for table metadata and table content
    - Similar to dEA --- Masked Entity Recovery




Deng, Xiang, et al. "TURL: table understanding through representation learning." *Proceedings of the VLDB Endowment* 14.3 (2020): 307-319

# Representation Learning for Relational Data

- Joint representation for NL and relational data
  - TURL [Deng et al. 2020]
    1. Separate input embedding for table metadata and table content
    2. Similar to dEA --- Masked Entity Recovery
    3. Masked self-attention: token/entity can only attend to its directly connected neighbors

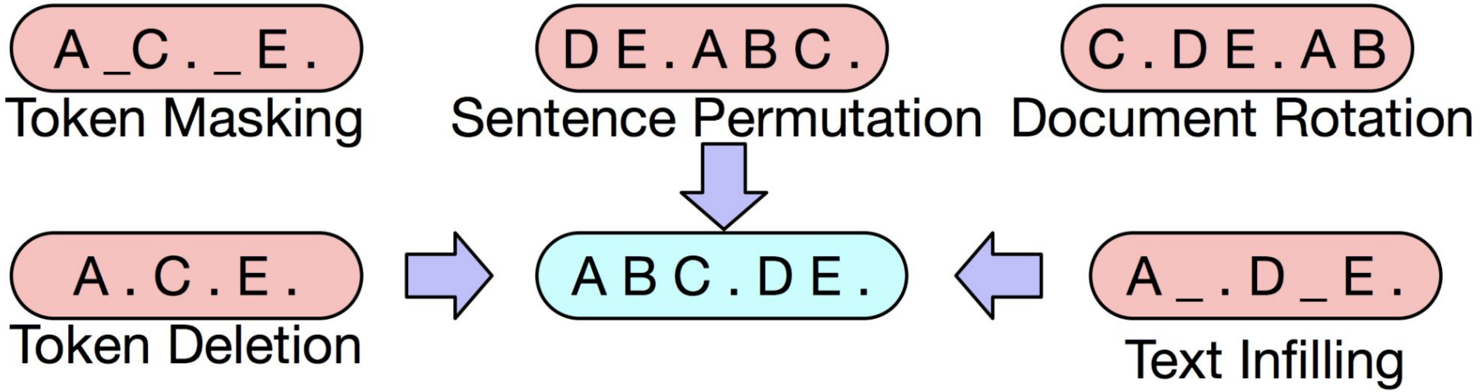


# Representation Learning for Relational Data

- Joint representation for NL and relational data
  - TURL [Deng et al. 2020]
    - Separate input embedding for table metadata and table content
    - Similar to dEA --- Masked Entity Recovery
    - Generalizes well to 6 table understanding tasks and substantially outperforms existing methods
- Heated topic!
  - TURL [Deng et al. 2020], TaBERT [Yin et al. 2020], TAPAS [Herzig et al. 2020]
  - RPT [Tang et al. 2021], TUTA [Wang et al. 2021], TABBIE [Iida et al. 2021] ...

# DA for Representation Learning

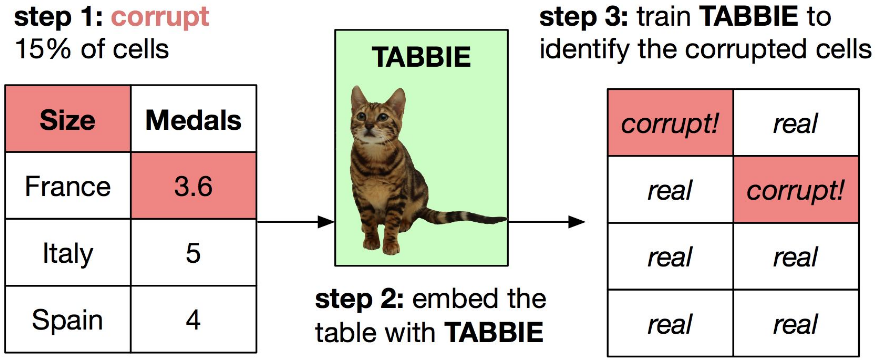
- Denoising [BART, Lewis et al. 2020]
    - Corrupt the text with arbitrary transformations
    - The model learns to reconstruct the original text




Lewis, Mike, et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.

# DA for Representation Learning

- Design space for relational data
  - TABBIE [Iida et al. 2021]: corrupt cell detection




(a) original table

| Rank | Country | Gold |
|---|---|---|
| 1 | France | 9 |
| 2 | Italy | 5 |
| 3 | Spain | 4 |

(b) sample cells from other tables

| Rank | Size | Gold |
|---|---|---|
| 1 | France | 3.6 |
| 2 | Italy | 5 |
| 3 | Spain | 4 |

(c) swap cells on the same row

| Rank | Country | Gold |
|---|---|---|
| 1 | France | 9 |
| 2 | 5 | Italy |
| 3 | Spain | 4 |

(d) swap cells on the same column

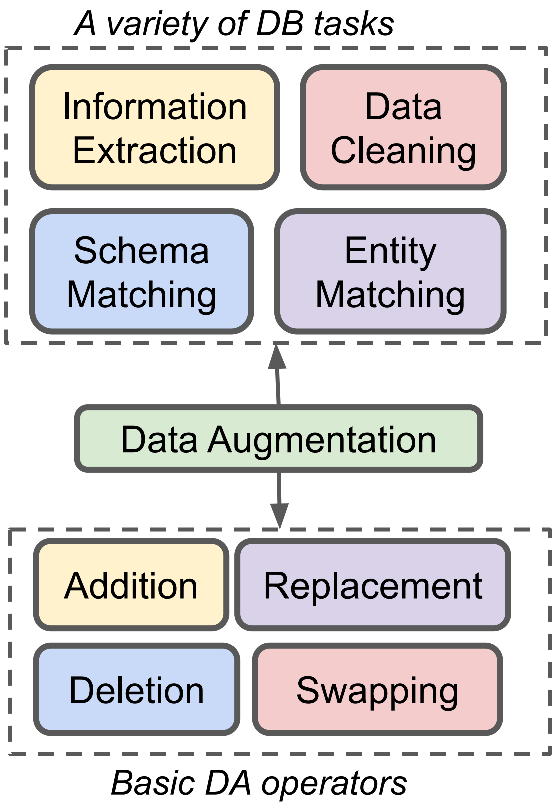
| Rank | Country | Gold |
|---|---|---|
| 1 | France | 9 |
| 3 | Italy | 5 |
| 2 | Spain | 4 |

Iida, Hiroshi, et al. "TABBIE: Pretrained Representations of Tabular Data." *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021.

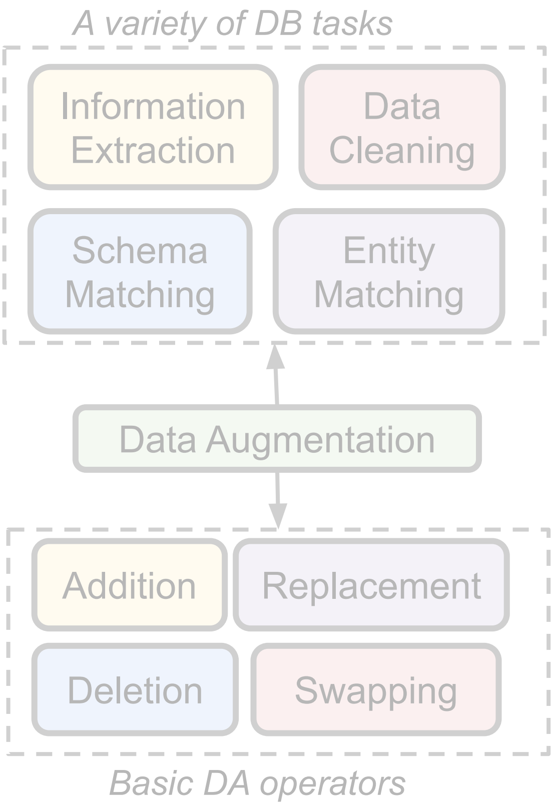
# Representation Learning for DA

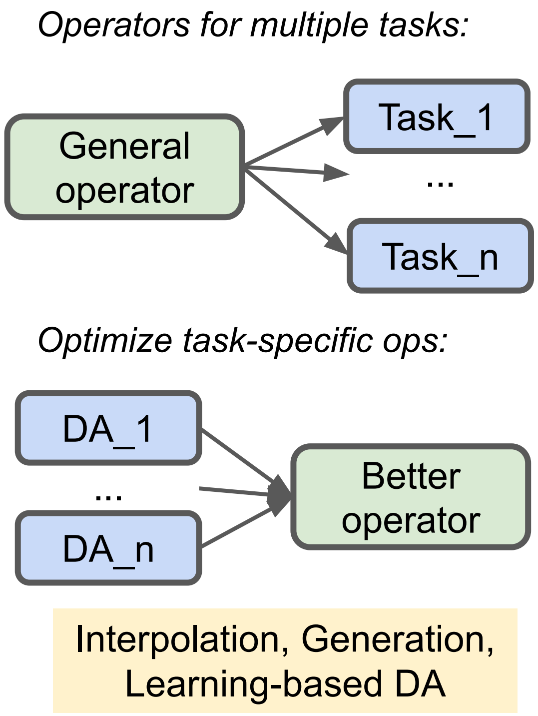
- Pre-trained LMs provides effective text data augmentation
  - Learned token replacement [Hu et al. 2019]
  - Conditional generation [Kumar et al. 2020]
  - InvDA [Miao et al. 2021]
- Opportunities with DA using pre-trained models for relational data
  - E.g. cell value prediction, row population

# Conclusion

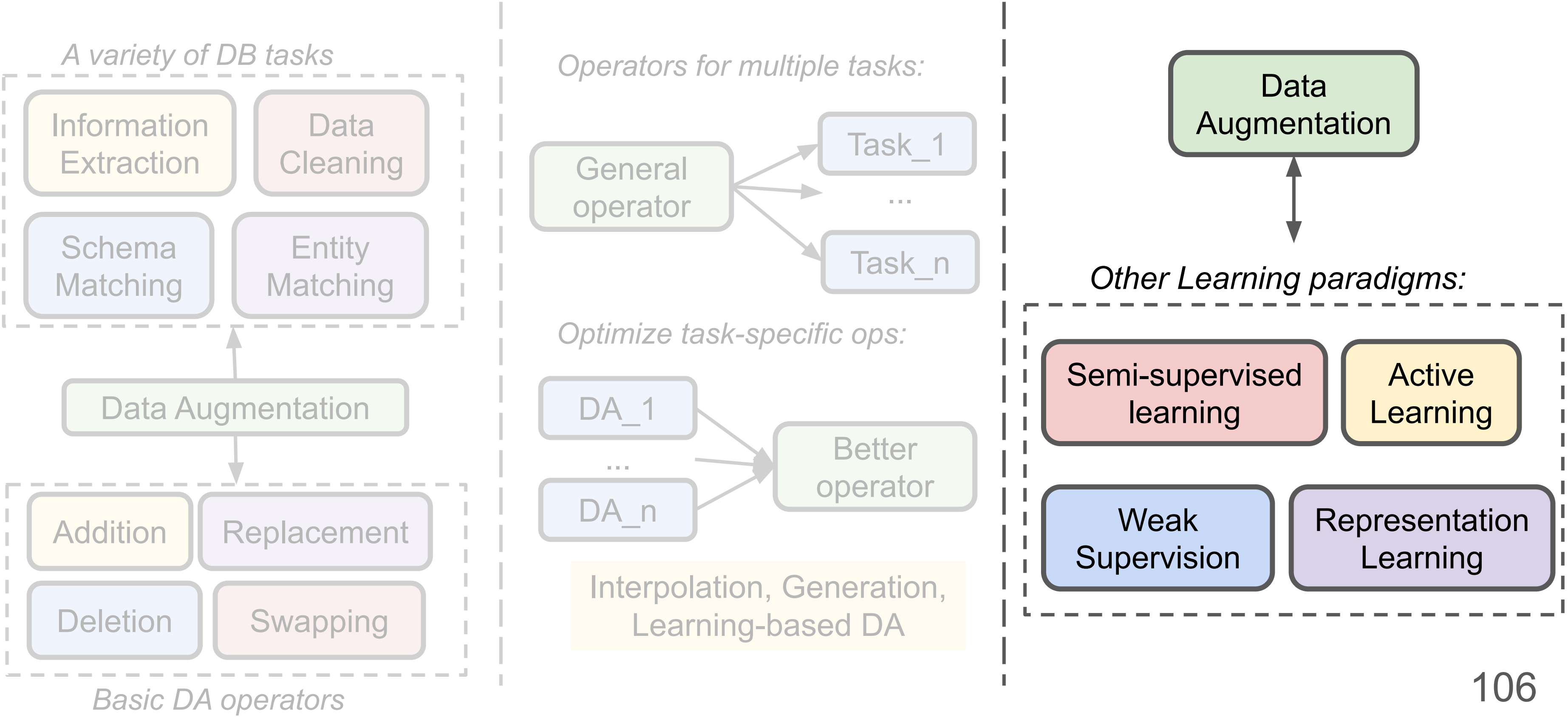

A variety of DB tasks
Information Extraction
Data Cleaning
Schema Matching
Entity Matching
Data Augmentation
Addition
Replacement
Deletion
Swapping
Basic DA operators

# Conclusion


A variety of DB tasks
Information Extraction
Data Cleaning
Schema Matching
Entity Matching
Data Augmentation
Addition
Replacement
Deletion
Swapping
Basic DA operators


Operators for multiple tasks:
General operator
Task_1
...
Task_n
Optimize task-specific ops:
DA_1
...
DA_n
Better operator
Interpolation, Generation, Learning-based DA

# Conclusion


A variety of DB tasks
Information Extraction
Data Cleaning
Schema Matching
Entity Matching
Data Augmentation
Addition
Replacement
Deletion
Swapping
Basic DA operators
Operators for multiple tasks:
General operator
Task_1
...
Task_n
Optimize task-specific ops:
DA_1
...
DA_n
Better operator
Interpolation, Generation, Learning-based DA
Data Augmentation
Other Learning paradigms:
Semi-supervised learning
Active Learning
Weak Supervision
Representation Learning

# References


Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

Wei, Jason, et al. "Eda: Easy data augmentation techniques for boosting performance on text classification tasks." *arXiv preprint arXiv:1901.11196* (2019)

Miao, Zhengjie, et al. "Snippext: Semi-supervised opinion mining with augmented data." *Proceedings of The Web Conference 2020*. 2020.

Heidari, Alireza, et al. "Holodetect: Few-shot learning for error detection." *Proceedings of the 2019 International Conference on Management of Data*. 2019.

Li, Yuliang, et al. "Deep entity matching with pre-trained language models." *Proceedings of the VLDB Endowment* 14.1 (2020): 50-60.

Shraga, Roee, et al. "Adnev: Cross-domain schema matching using deep similarity matrix adjustment and evaluation." *Proceedings of the VLDB Endowment* 13.9 (2020): 1401-1415.

# References

Zhang, Hongyi, et al. "mixup: Beyond Empirical Risk Minimization." ICLR 2018.

Guo, Hongyu et al.. "Augmenting data with mixup for sentence classification: An empirical study." arXiv preprint arXiv:1905.08941 (2019).

Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9.

Kumar, Varun et al. "Data Augmentation using Pre-trained Transformer Models." Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems. 2020.

Miao, Zhengjie et al.. "Rotom: A Meta-Learned Data Augmentation Framework for Entity Matching, Data Cleaning, Text Classification, and Beyond." SIGMOD 2021.

Cubuk, Ekin D., et al. "Autoaugment: Learning augmentation policies from data." CVPR 2019

Zoph, Barret, et al. "Learning transferable architectures for scalable image recognition." CVPR 2018

Lim, Sungbin, et al. "Fast autoaugment." NeurIPS 2019

# References

Hu, Zhiting, et al. "Learning Data Manipulation for Augmentation and Weighting." NeurIPS 2019

Carlson, Andrew, et al. "Coupled semi-supervised learning for information extraction." Proceedings of the third ACM international conference on Web search and data mining. 2010.

Hu, Xuming, et al. "Semi-supervised relation extraction via incremental meta self-training." arXiv preprint arXiv:2010.16410 (2020).

Kejriwal, Mayank, and Daniel P. Miranker. "Semi-supervised instance matching using boosted classifiers." European semantic web conference. Springer, Cham, 2015.

Xie, Qizhe, et al. "Unsupervised data augmentation for consistency training." arXiv preprint arXiv:1904.12848 (2019).

Berthelot, David, et al. "MixMatch: A Holistic Approach to Semi-Supervised Learning." Advances in Neural Information Processing Systems 32 (2019).

Kasai, Jungo, et al. "Low-resource Deep Entity Resolution with Transfer and Active Learning." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

# References

Jain, Arjit, Sunita Sarawagi, and Prithviraj Sen. "Deep Indexed Active Learning for Matching Heterogeneous Entity Representations." arXiv preprint arXiv:2104.03986 (2021).

Hong, SeulGi, et al. "Deep Active Learning with Augmentation-based Consistency Estimation." arXiv preprint arXiv:2011.02666 (2020).

Gao, Mingfei, et al. "Consistency-based semi-supervised active learning: Towards minimizing labeling cost." European Conference on Computer Vision. Springer, Cham, 2020

Ratner, Alexander J., et al. "Learning to compose domain-specific transformations for data augmentation." Advances in neural information processing systems 30 (2017): 3239.

Cappuzzo, Riccardo, Paolo Papotti, and Saravanan Thirumuruganathan. "Creating embeddings of heterogeneous relational datasets for data integration tasks." Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. 2020.

Zhang, Zhengyan, et al. "ERNIE: Enhanced Language Representation with Informative Entities." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.

# References


Deng, Xiang, et al. "TURL: table understanding through representation learning." Proceedings of the VLDB Endowment 14.3 (2020): 307-319

Lewis, Mike, et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

Tang, Nan, et al. "RPT: Relational Pre-trained Transformer Is Almost All You Need towards Democratizing Data Preparation." Proceedings of the VLDB Endowment 14.8 (2021): 1254-1261

Iida, Hiroshi, et al. "TABBIE: Pretrained Representations of Tabular Data." Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021.

Yin, Pengcheng, et al. "TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

Herzig, Jonathan, et al. "TaPas: Weakly Supervised Table Parsing via Pre-training." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

Wang, Zhiruo, et al. "Structure-aware pre-training for table understanding with tree-based transformers." arXiv preprint arXiv:2010.12537 (2020).