Generic Entity Resolution Models

Jiawei Tang
American School of Doha, Qatar

Yifei Zuo
USTC, China

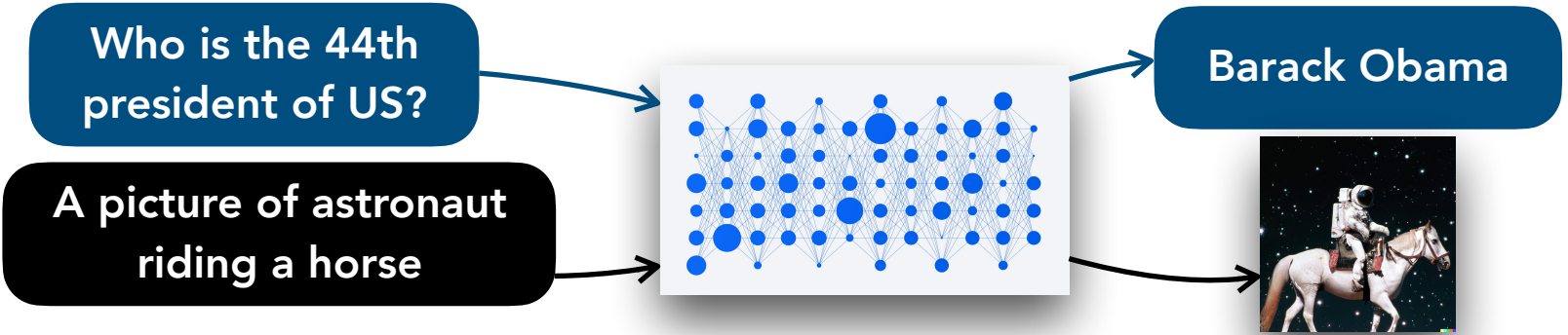**Lei Cao, Samuel Madden**
CSAIL, MIT, USA

The Entity Resolution Problem

Given two entities **a** and **b**, decide if (**a**, **b**) match or not?
Answer: 1 (match)

a

Company	Address	City
P Sherman Orthodontics	32 Wallaby Way	Sydney, NSW

b

Company	Address	City
Sherman Orthodontics Ltd.	Sky Tower, 32 Wallaby Way	Sydney

Specific Models vs. Generic Solutions

Specific Solutions
- traditional wisdom -

D1 → Train → 0/1
D2 → 0/1
... ...
Dm → 0/1

Generic Solutions
- goal of this project -

D1 D2 ... ... Dm → 0/1

Limitations
1. Need a lot of train data for each dataset
2. Lack of generalizability
3. Large sizes
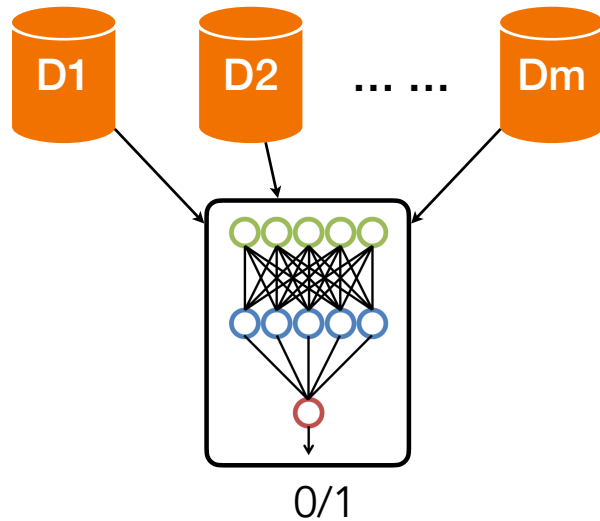
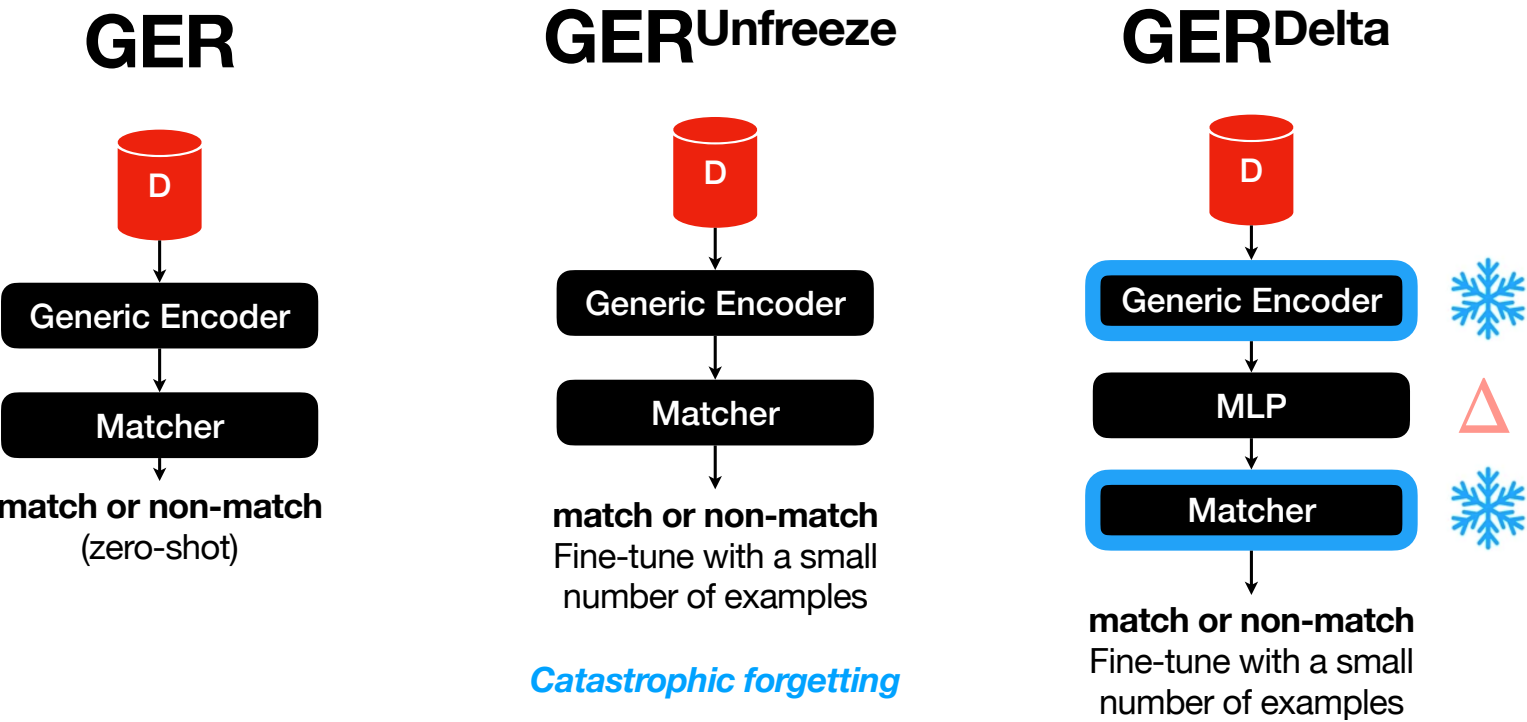**Opportunities**
1. Foundation models
2. A trained generic model

Foundation Models

Foundation models are giant artificial intelligence models trained on a large corpus of data and can perform many tasks
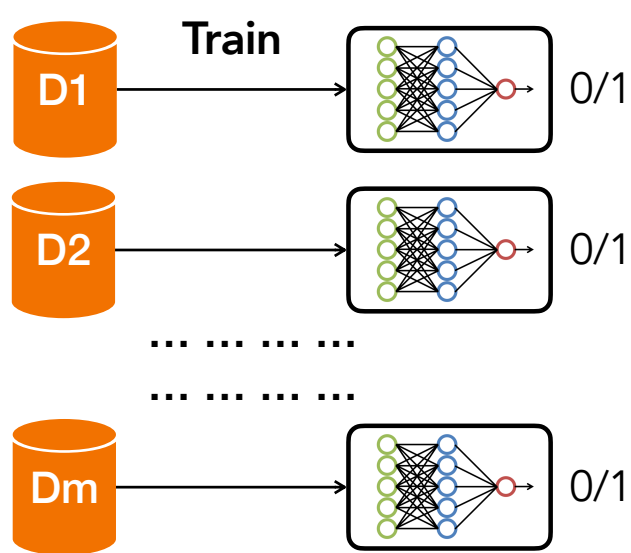
Who is the 44th president of US? → Barack Obama

A picture of astronaut riding a horse →

Foundation Models for Entity Resolution

Case 1: Zero-shot no example

0 (not a match)

GPT3

Case 2: Few-shot few examples

Restaurant A is Name: 456 Shanghai Cuisine Phone: 1261 Address: 69 Mott Street, New York, NY
Restaurant B is Name: Shanghai Asian Manor Phone Number: (212)766-6311 Address: 21 Mott Street, New York
Are A and B the same?

Tokenization

Name	Phone	Address
456 Shanghai Cuisine	1261	69 Mott Street, New York, NY

Name	Phone	Address
Shanghai Asian Manor	(212)766-6311	21 Mott St, New York

Experiments with Foundation Models

Zero-shot results

Dataset	Precision	Recall	F-measure
Restaurant	0.991	0.919	0.954
Bike	1	0.1	0.182
Movie	0.714	0.368	0.486
Book	0.632	0.598	0.615

Few-shot results

Restaurant   Music

1. Under zero-shot, GPT3 works, but underperforms specific solutions
2. Few-shot gives very biases results (either all 0's or all 1's)
3. **Conclusion**: Zero-shot should be used on GPT3 for entity resolution

A Trained Generic Model

Designed three solutions for building generic models

GER

D → Generic Encoder → Matcher → match or non-match (zero-shot)

GER^Unfreeze

D → Generic Encoder → Matcher → match or non-match
Fine-tune with a small number of examples
Catastrophic forgetting

GER^Delta

D → Generic Encoder → MLP → Matcher → match or non-match
Fine-tune with a small number of examples

Experiments with Generic Models

	Specific Model			GPT-3			GER		
Dataset	P	R	F1	P	R	F1	P	R	F1
Restaurant	0.941	0.941	0.941	0.991	0.919	0.954	0.978	0.936	0.956
Bike	0.773	0.548	0.642	1	0.1	0.182	0.405	0.81	0.54
Movie	0.978	0.957	0.968	0.714	0.368	0.486	0.873	0.984	0.925
Book	1	1	1	0.632	0.598	0.615	0.837	1	0.911

	GER^Unfreeze			GER^Delta		
Dataset	P	R	F1	P	R	F1
Restaurant	0.979	0.957	**0.968**	0.939	0.979	0.958
Bike	0.554	0.857	0.673	0.56	0.881	**0.685**
Movie	1	0.984	0.992	1	0.984	**0.992**
Book	0.973	1	**0.986**	0.935	1	0.966

1. Both GER^Unfreeze and GER^Delta can be adapted to a new task with a few train examples
2. If one still wants the entity resolution model to perform well on previously tra...

Unfreezing all layers can help, but it may cause catastrophic forgetting, which is not generic

A small delta network can help quickly adapt the generic model to a specific task