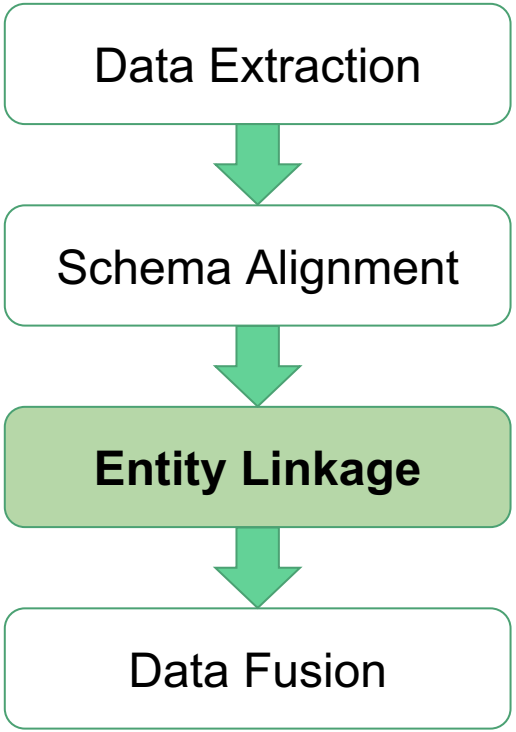
Outline

- Part I. Introduction
- Part II. ML for DI
  - ML for entity linkage
  - ML for data extraction
  - ML for data fusion
  - ML for schema alignment
- Part III. DI for ML
- Part IV. Conclusions and research direction



What is Entity Linkage?

- Definition: Partition a given set R of records, such that each partition corresponds to a distinct real-world entity.

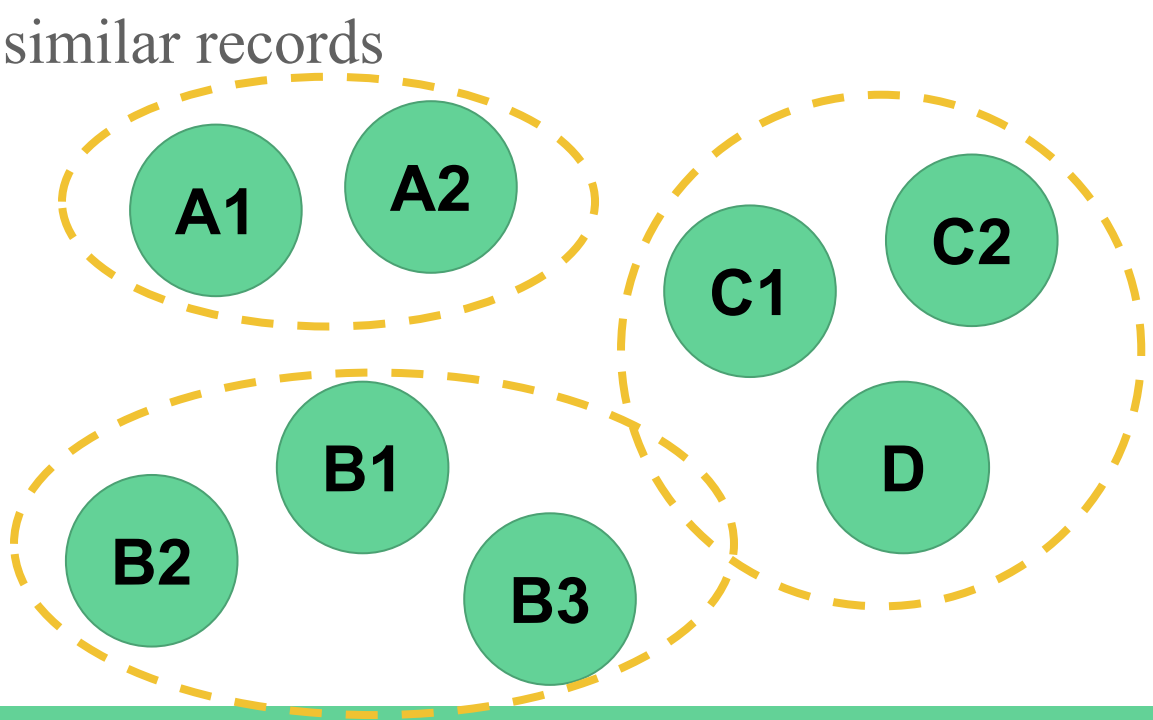
Are they the same entity?

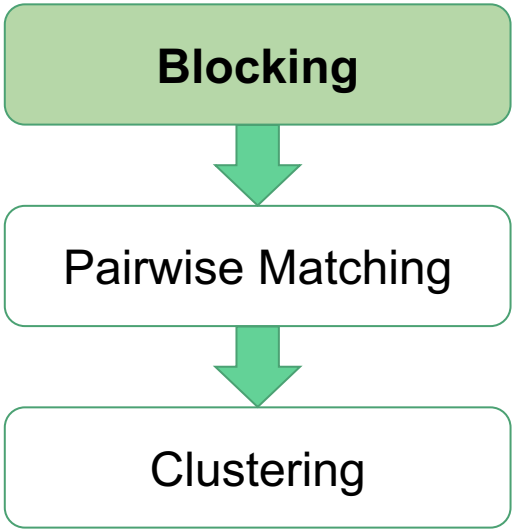
IMDB

WikiData

Three Steps in Entity Linkage

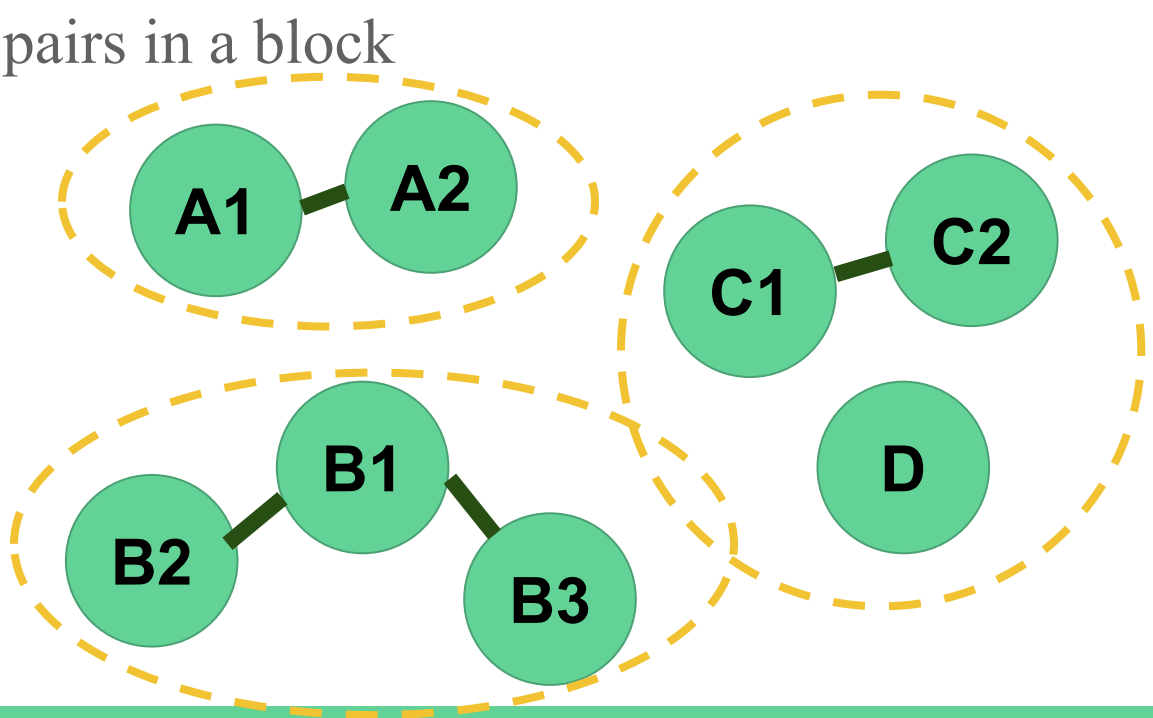
- **Blocking**: efficiently create small blocks of similar records

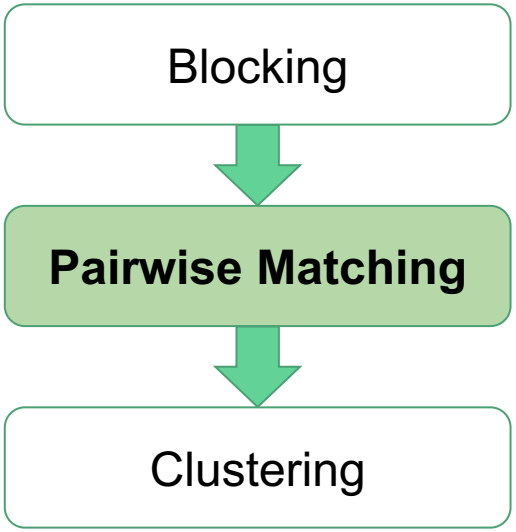




Three Steps in Entity Linkage

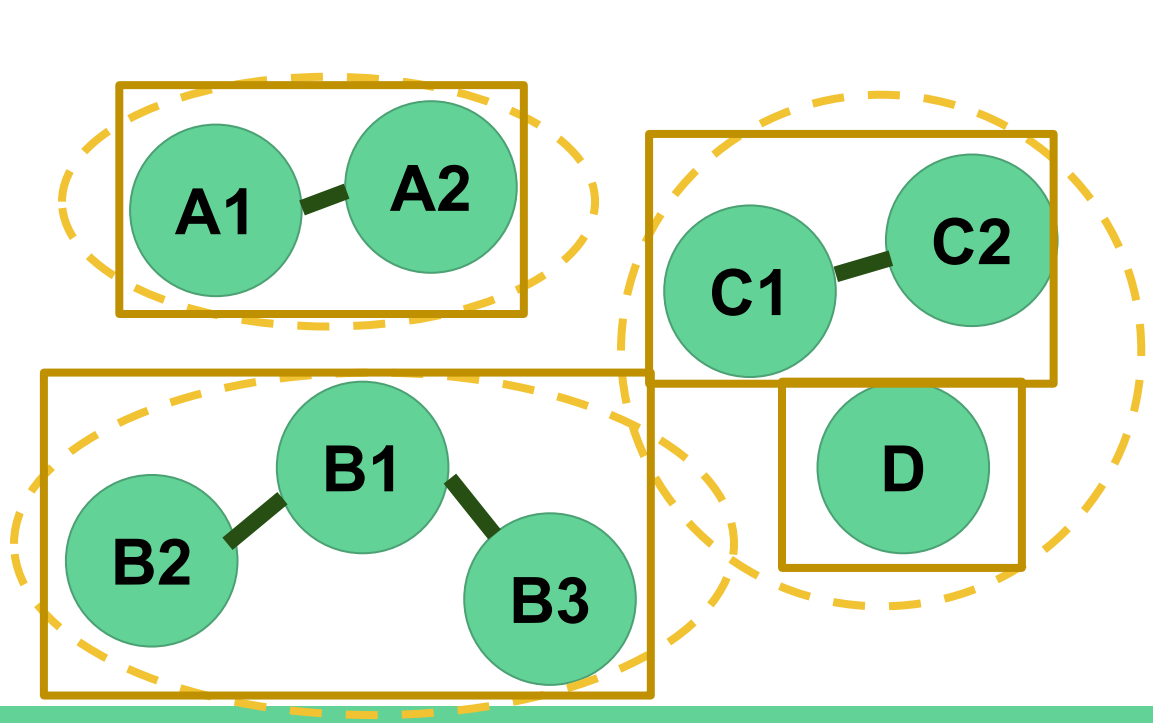
- **Pairwise matching**: compare all record pairs in a block

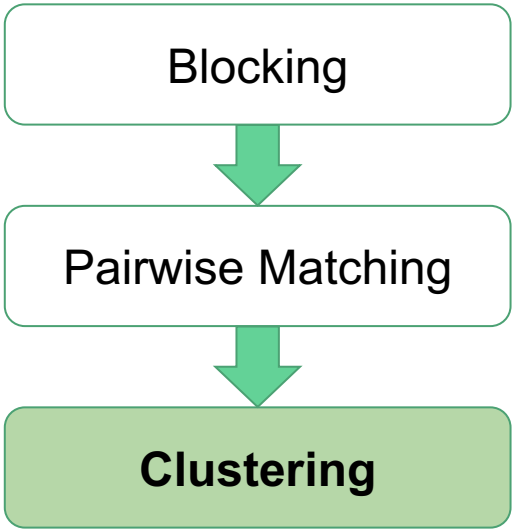




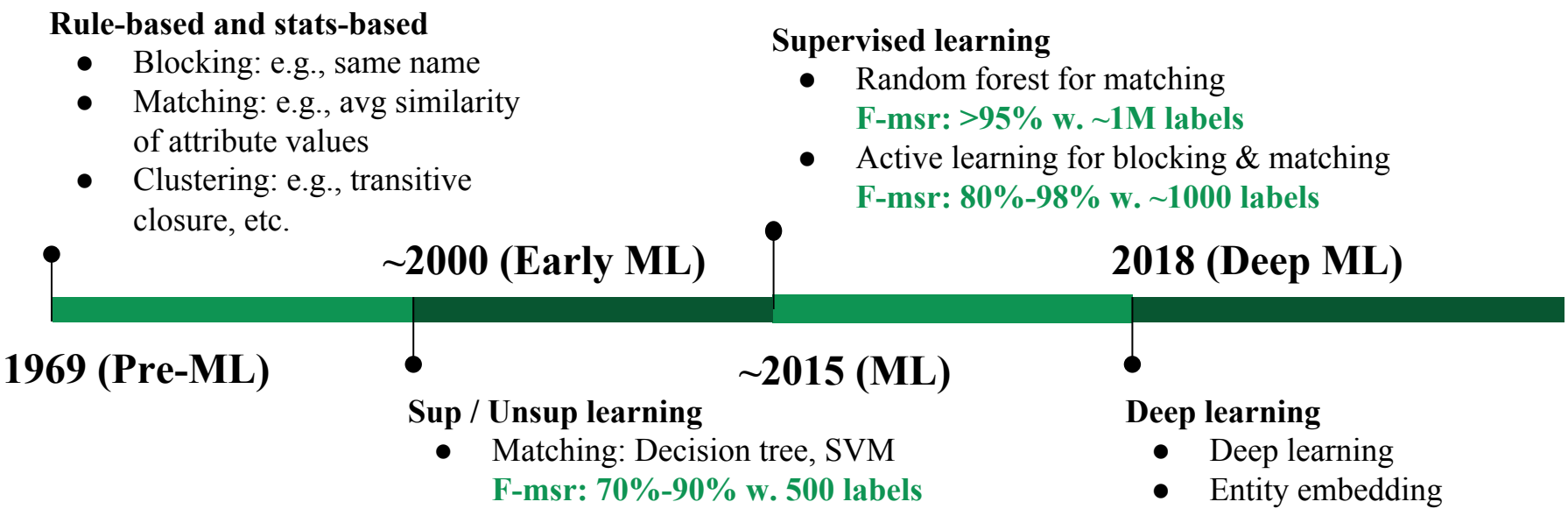
Three Steps in Entity Linkage

- **Clustering**: group records into entities





50 Years of Entity Linkage



Rule-Based Solution

Rule-based and stats-based

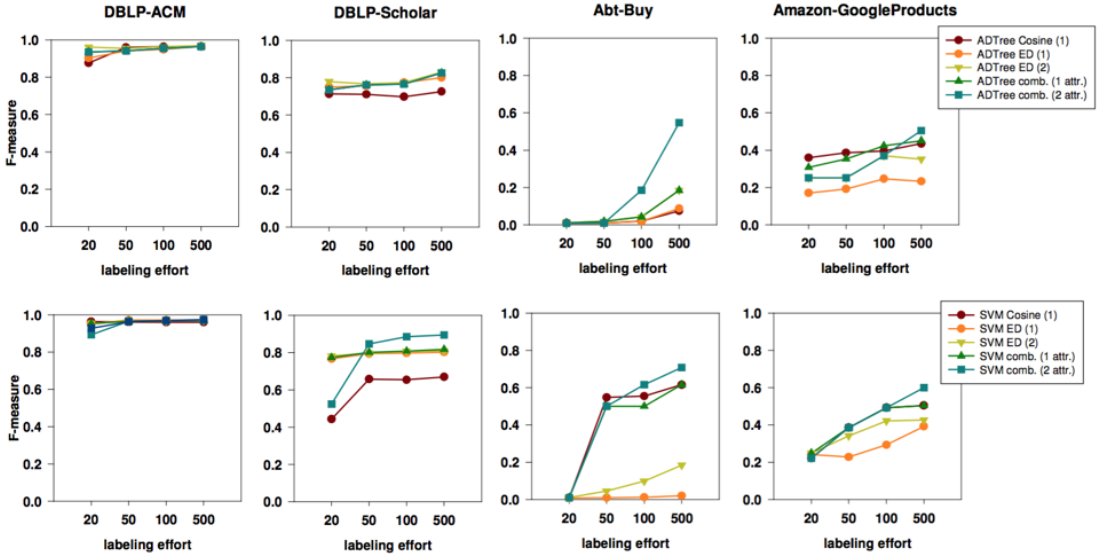
- Blocking: e.g., same name
- Matching: e.g., avg similarity of attribute values
- Clustering: e.g., transitive closure, etc.

1969 (Pre-ML)

- [Fellegi and Sunter, 1969]
  - Match: $sim(r, r') > \boldsymbol{\theta}_h$
  - Unmatch: $sim(r, r') < \boldsymbol{\theta}_l$
  - Possible match: $\boldsymbol{\theta}_l < sim(r, r') < \boldsymbol{\theta}_h$

Early ML Models

- [Köpcke et al, VLDB'10]



~2000 (Early ML)

Sup / Unsup learning

- Matching: Decision tree, SVM
  F-msr: 70%-90% w. 500 labels
- Clustering: Correlation clustering, Markov clustering

State-of-the-Art ML Models [Dong, KDD'18]

Supervised learning

- Random forest for matching
  F-msr: >95% w. ~1M labels
- AL for blocking & matching
  F-msr: 80%-98% w. ~1000 labels

~2015 (ML)

- Features: attribute similarity measured in various ways. E.g.,
  - string sim: Jaccard, Levenshtein
  - number sim: absolute diff, relative diff
- ML models on Freebase vs. IMDb
  - Logistic regression: Prec=0.99, Rec=0.6
  - Random forest: Prec=0.99, Rec=0.99

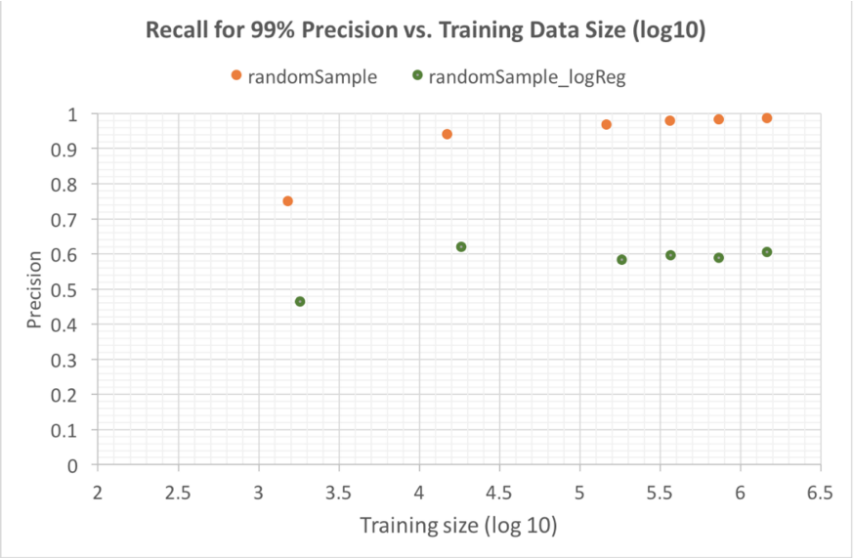
State-of-the-Art ML Models [Dong, KDD'18]

Supervised learning

- Random forest for matching
  F-msr: >95% w. ~1M labels
- AL for blocking & matching
  F-msr: 80%-98% w. ~1000 labels

~2015 (ML)

- Expt 1. IMDb vs. Freebase
  - Logistic regression: Prec=0.99, Rec=0.6
  - Random forest: Prec=0.99, Rec=0.99



State-of-the-Art ML Models [Dong, KDD'18]

Supervised learning

- Random forest for matching
  F-msr: >95% w. ~1M labels
- AL for blocking & matching
  F-msr: 80%-98% w. ~1000 labels

~2015 (ML)

- Features: attribute similarity measured in various ways. E.g.,
  - name sim: Jaccard, Levenshtein
  - age sim: absolute diff, relative diff
- ML models on Freebase vs. IMDb
  - Logistic regression: Prec=0.99, Rec=0.6
  - Random forest: Prec=0.99, Rec=0.99
  - XGBoost: marginally better, but sensitive to hyper-parameters

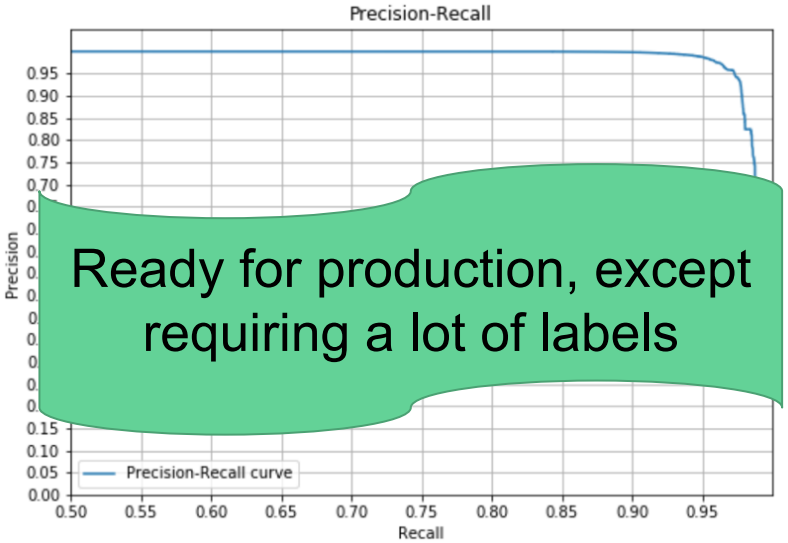
State-of-the-Art ML Models [Dong, KDD'18]

Supervised learning

- Random forest for matching
  F-msr: >95% w. ~1M labels
- AL for blocking & matching
  F-msr: 80%-98% w. ~1000 labels

~2015 (ML)

- Expt 2. IMDb vs. Amazon movies
  - 200K labels, ~150 features
  - Random forest: Prec=0.98, Rec=0.95



State-of-the-Art ML Models [Das et al., SIGMOD'17]

Magellan

Supervised learning

- Random forest for matching
  F-msr: >95% w. ~1M labels
- AL for blocking & matching
  F-msr: 80%-98% w. ~1000 labels

~2015 (ML)

- Falcon: apply active learning both for blocking and for matching; ~1000 labels

Dataset	Accuracy (%)			Cost (# Questions)
	P	R	F_1	
Products	90.9	74.5	81.9	$57.6 (960)
Songs	96.0	99.3	97.6	$54.0 (900)
Citations	92.0	98.5	95.2	$65.5 (1087)

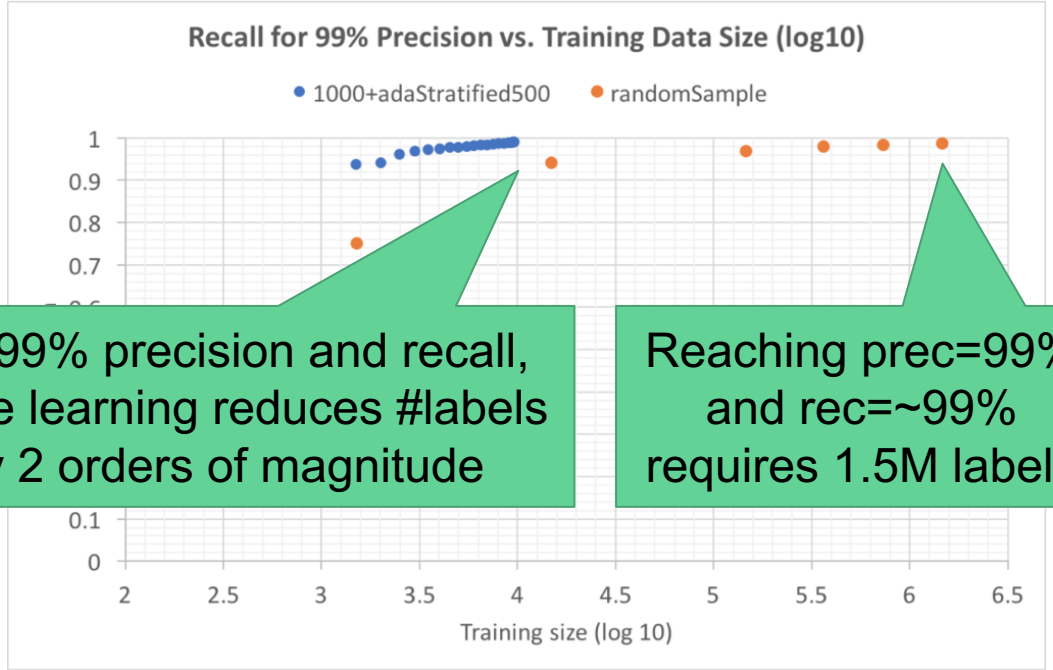
State-of-the-Art ML Models [Dong, KDD'18]

Supervised learning

- Random forest for matching
  F-msr: >95% w. ~1M labels
- AL for blocking & matching
  F-msr: 80%-98% w. ~1000 labels

~2015 (ML)

- Apply active learning to minimize #labels



Deep Learning Models [Mudgal et al., SIGMOD'18]

Magellan

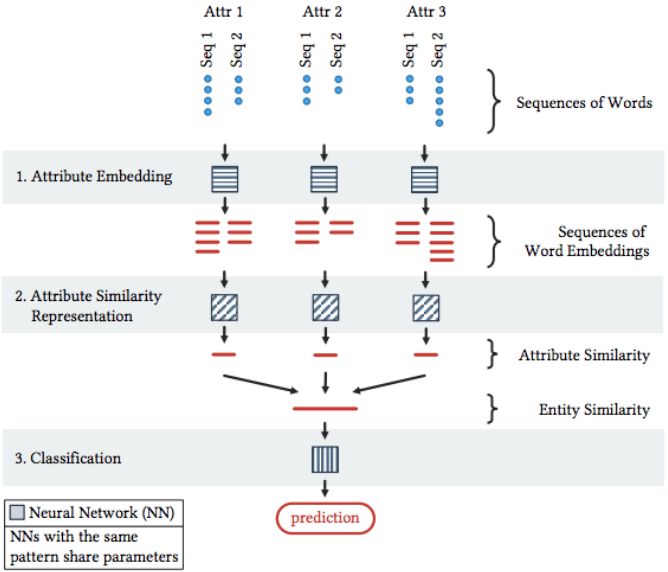
Check-out at poster session on Wednesday!
Code at: deepmatcher.ml

- Bi-RNN w. attention
- Similar performance for structured data;
  Significant improvement on texts and dirty data

2018 (Deep ML)

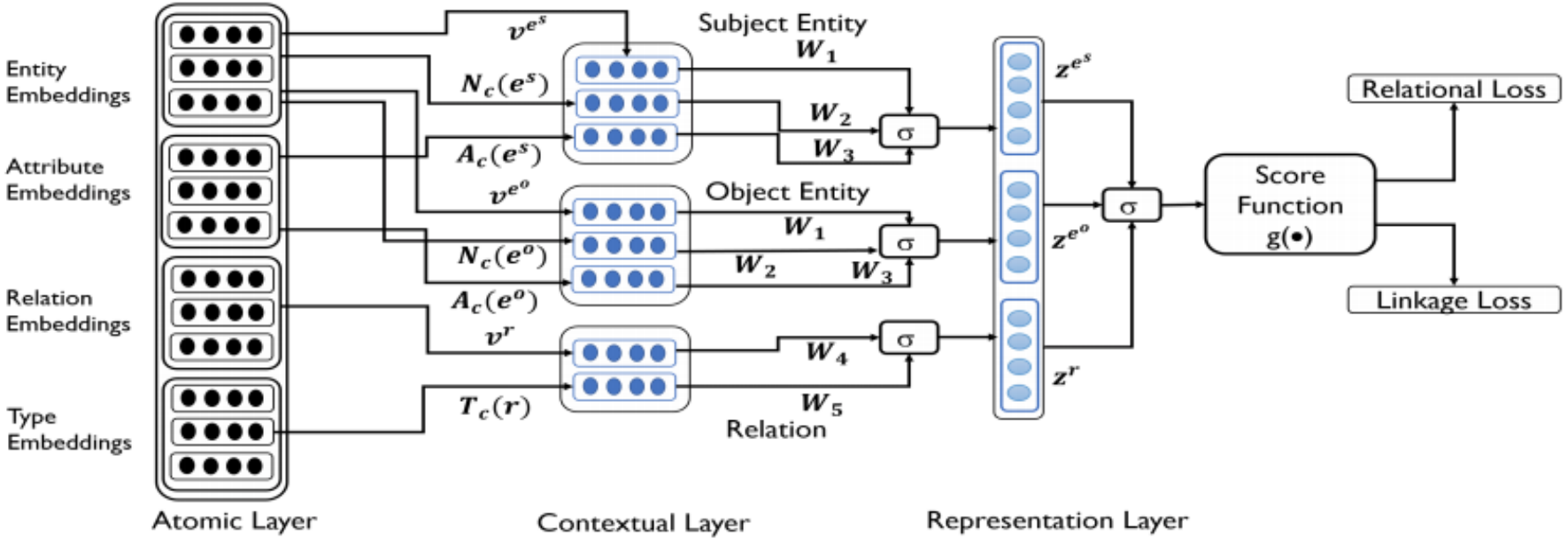
Deep learning

- Deep learning
- Entity embedding



Deep Learning Models [Trivedi et al., ACL'18]

- LinkNBed: Generate embeddings for entities as in knowledge embedding



Deep Learning Models [Trivedi et al., ACL'18]

2018 (Deep ML)

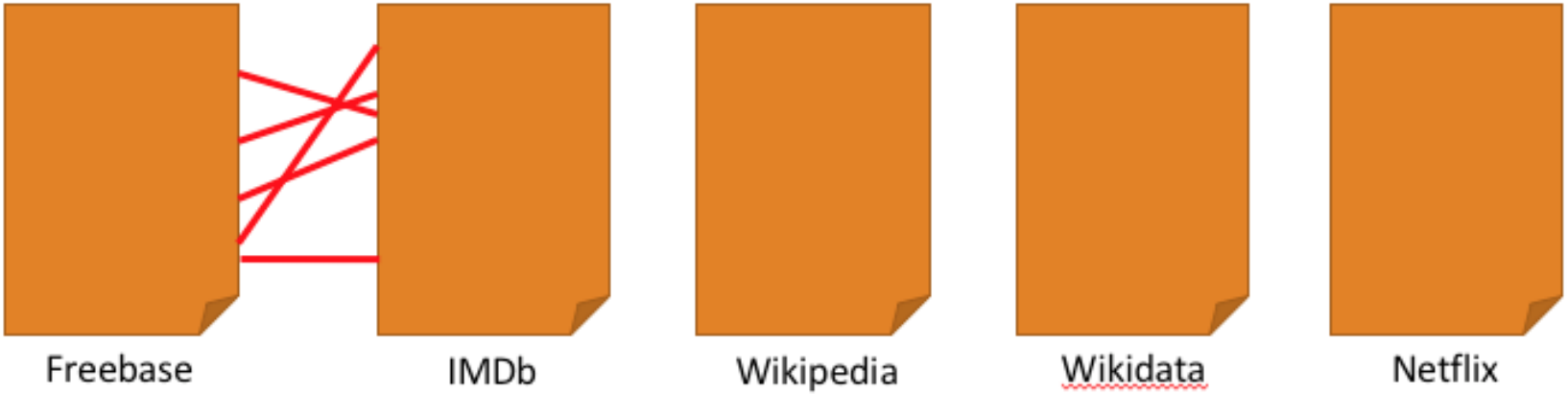
Deep learning

- Deep learning
- Entity embedding

- LinkNBed: Generate embeddings for entities as in knowledge embedding
- Performance better than previous knowledge embedding methods, but not comparable to random forest
- Enable linking different types of entities

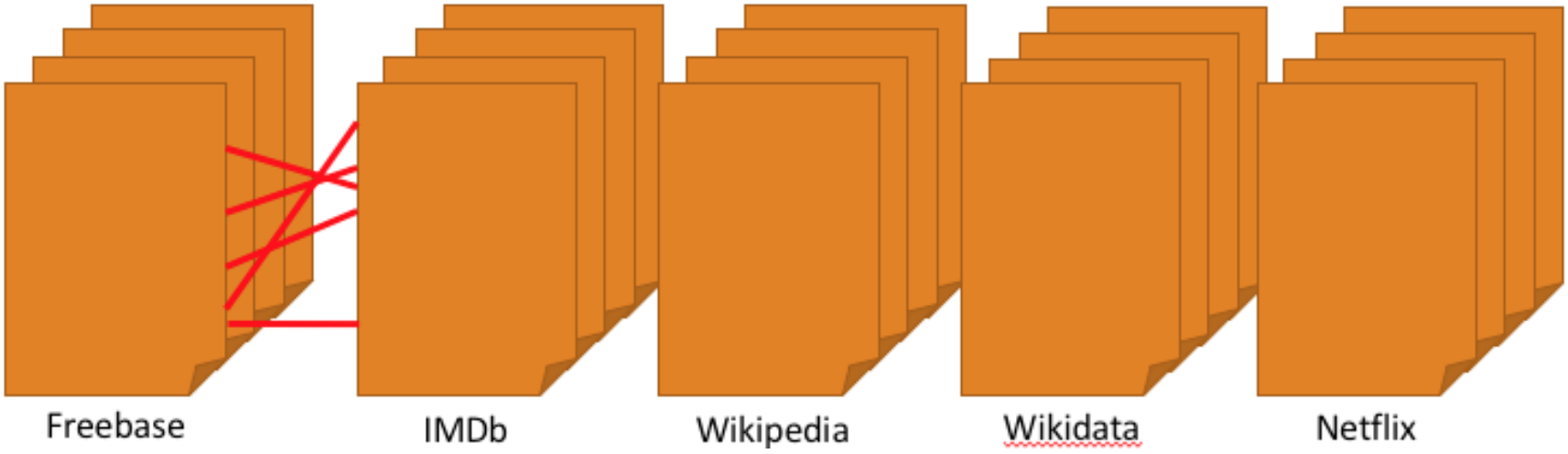
Challenges in Applying ML on EL

- **How can we obtain abundant training data for many types, many sources, and dynamically evolving data??**
- From two sources to multiple sources



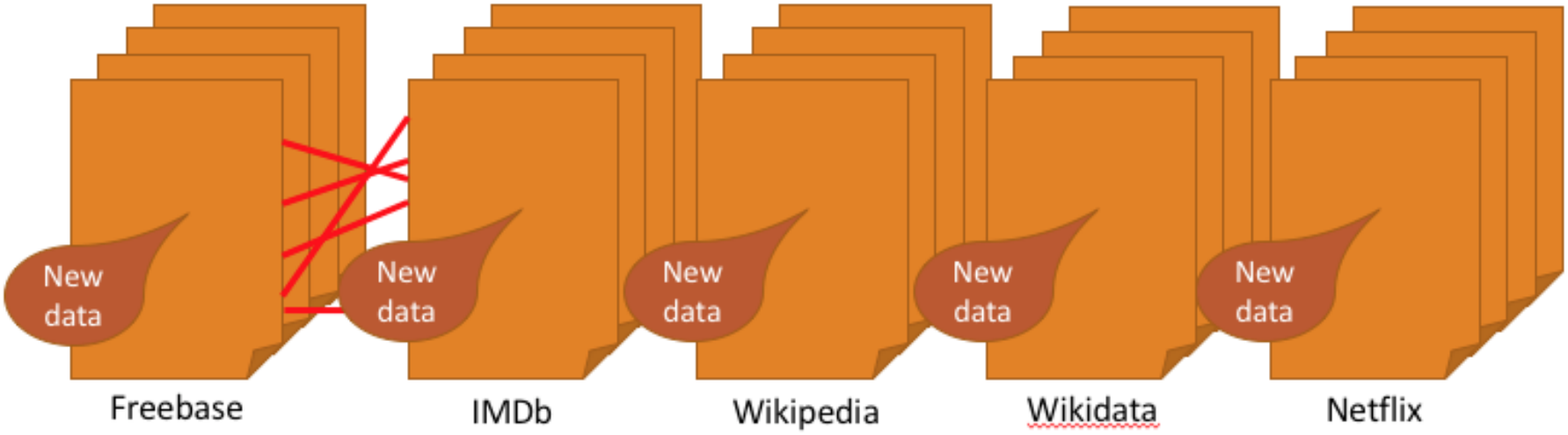
Challenges in Applying ML on EL

- **How can we obtain abundant training data for many types, many sources, and dynamically evolving data??**
- From one entity type to multiple types



Challenges in Applying ML on EL

- **How can we obtain abundant training data for many types, many sources, and dynamically evolving data??**
- From static data to dynamic data



Recipe for Entity Linkage

- Problem definition: **Link references to the same entity**
- Short answers
  - **RF w. attribute-similarity features**
  - **DL to handle texts and noises**
  - **End-to-end solution is future work**