CS4221
Cloud Databases IV. Data Integration

Yao LU

2024 Semester 2

National University of Singapore
School of Computing

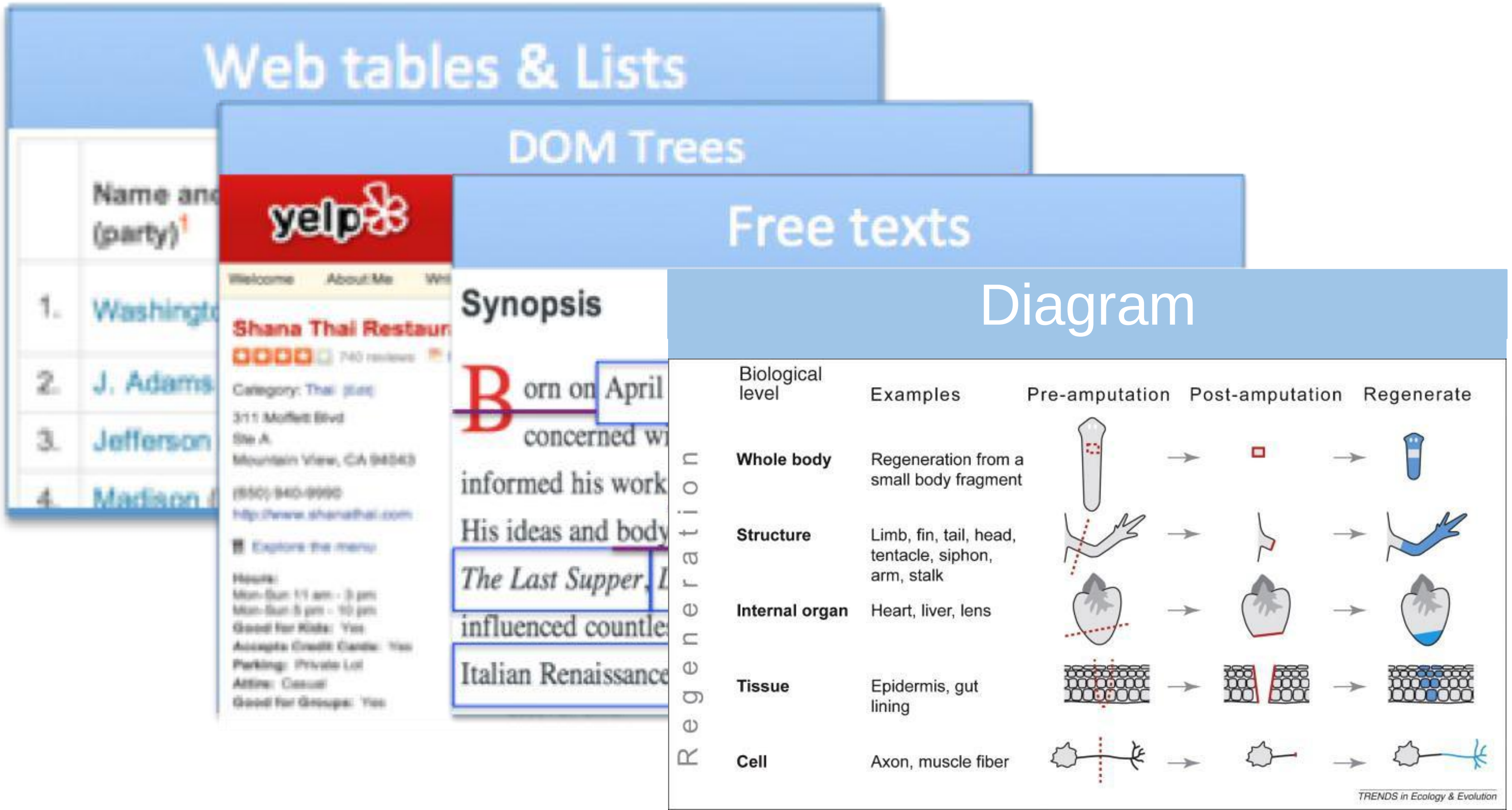
What is data integration?

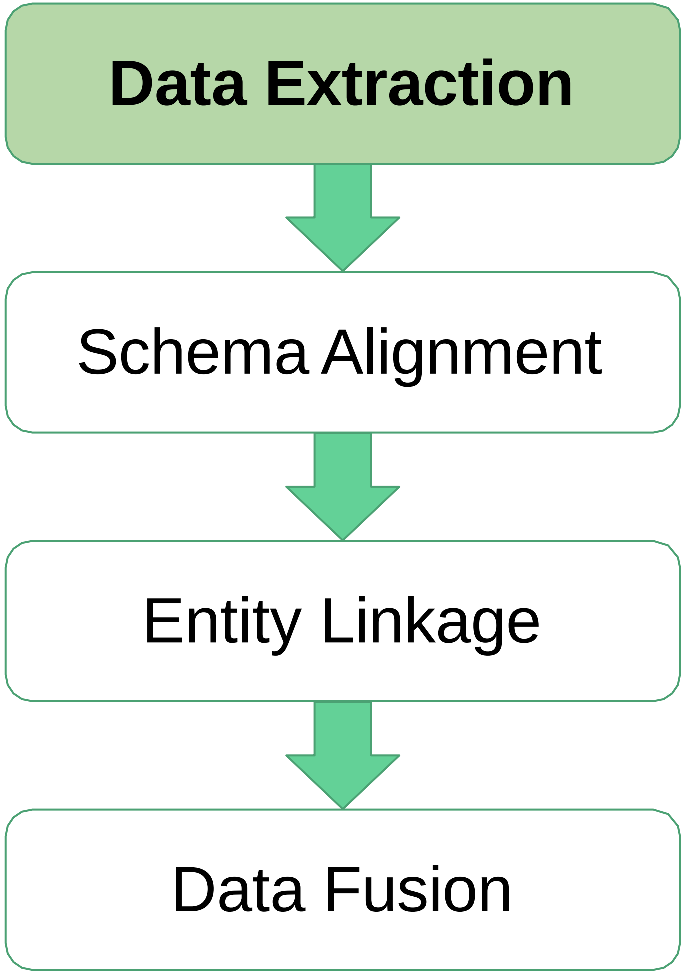
- **Data integration**: to provide unified access to data residing in multiple, autonomous data sources
  - **Data warehouse**: create a single store (materialized view) of data from different sources offline. Multi-billion dollar business.
  - **Virtual integration**: support query over a mediated schema by applying online query reformulation. E.g., Kayak.com.

- In the Resource Description Framework: different names for similar concepts
  - **Knowledge graph** is equivalent to a data warehouse. Has been widely used in Search and Voice
  - **Linked data** is equivalent to virtual integration

What is data integration?

- Heterogeneity everywhere
  - Different data formats



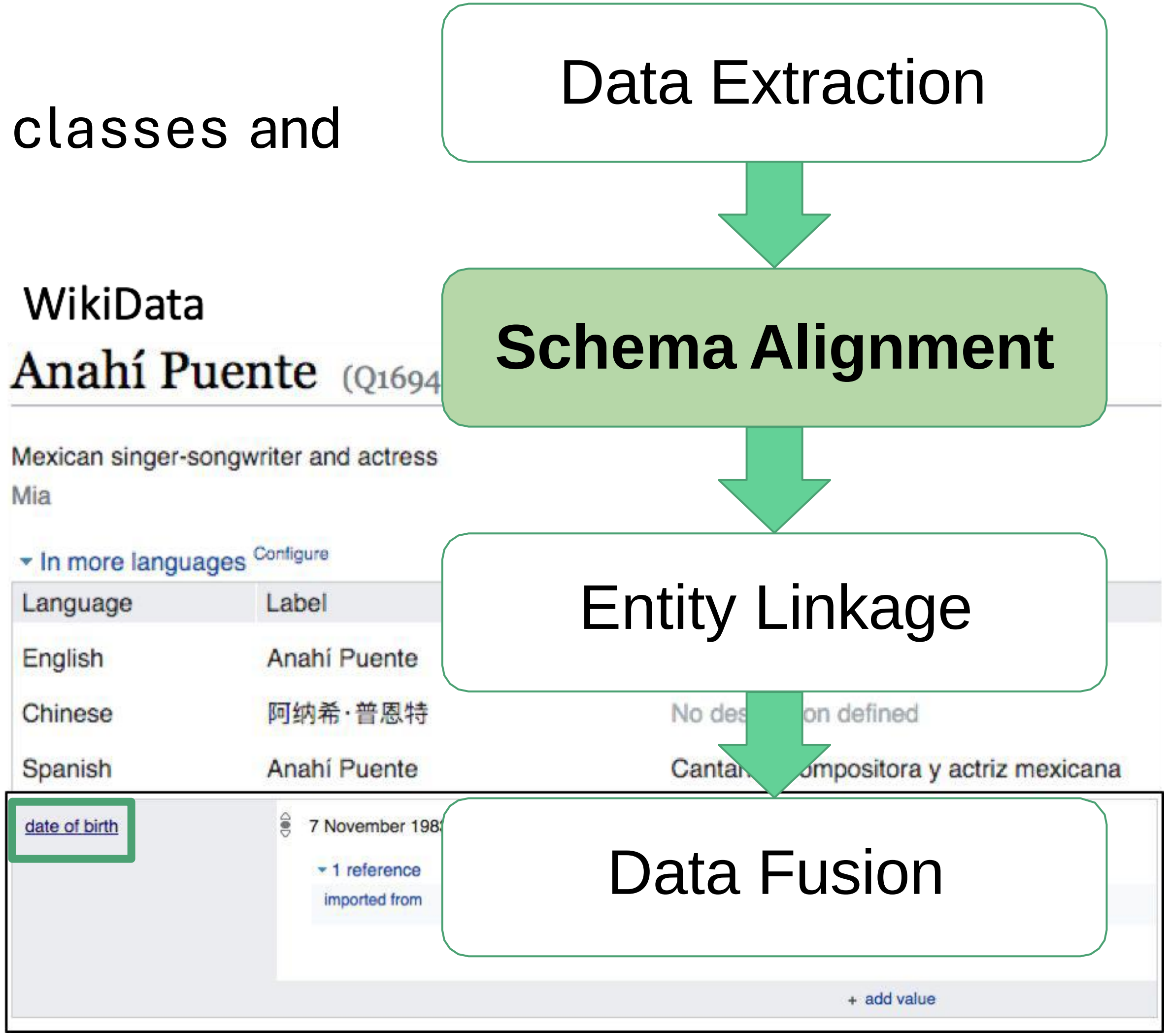


Why is data integration hard?

- Heterogeneity everywhere
  - Different ways to express the same classes and attributes

IMDB

WikiData

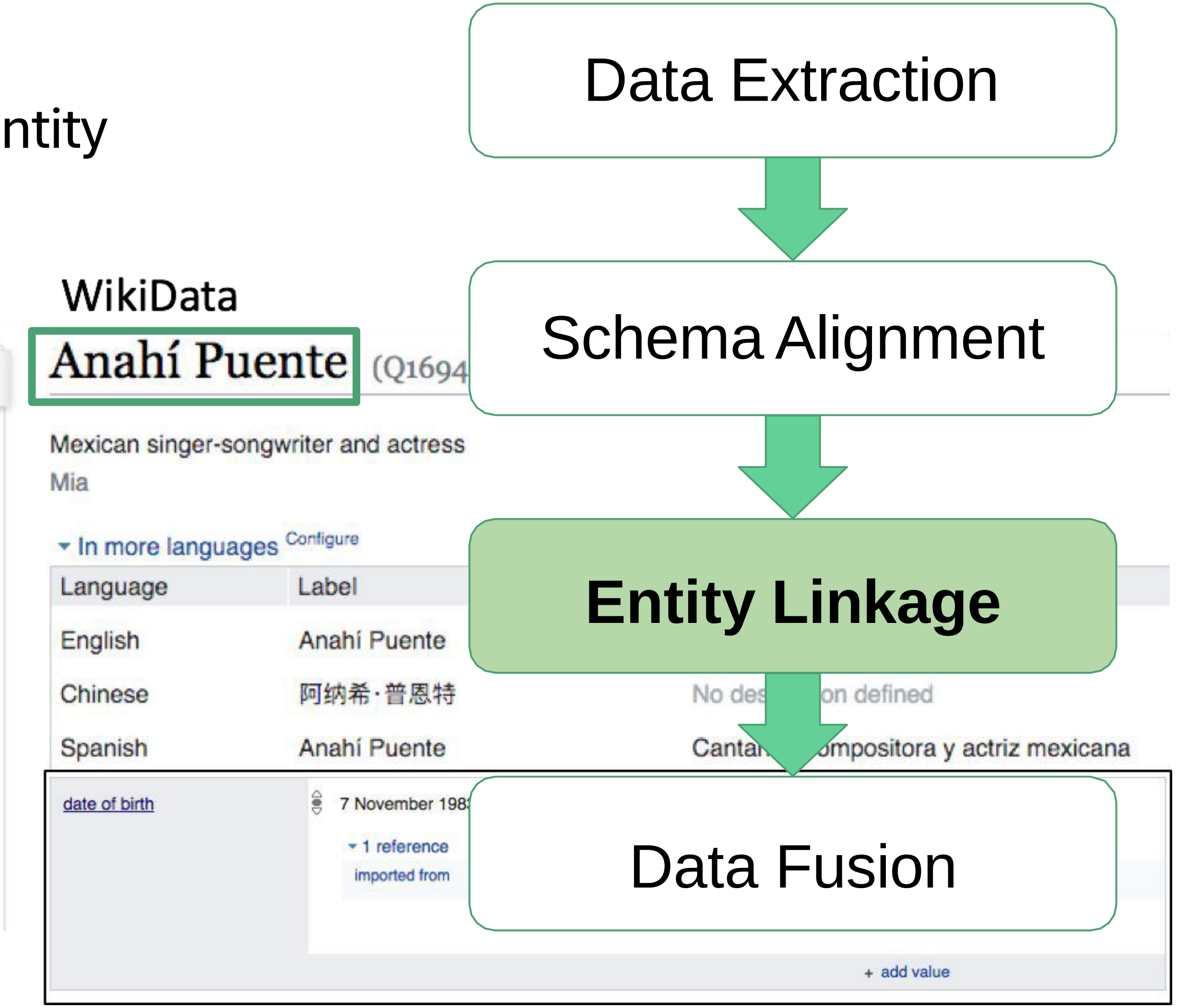


Why is data integration hard?

- Heterogeneity everywhere
  - Different references to the same entity

IMDB

WikiData

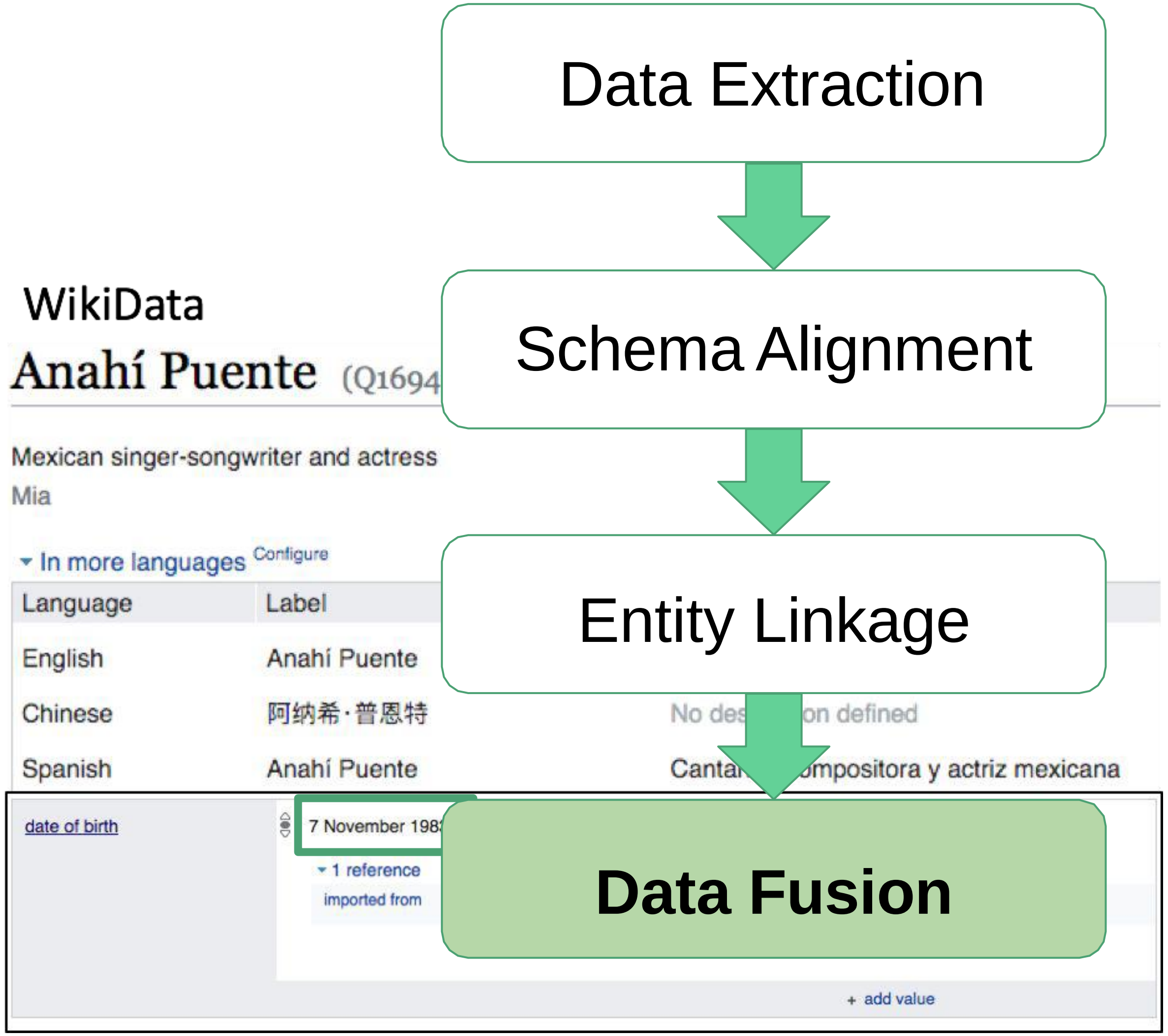


Why is data integration hard?

- Heterogeneity everywhere
  - Conflicting values

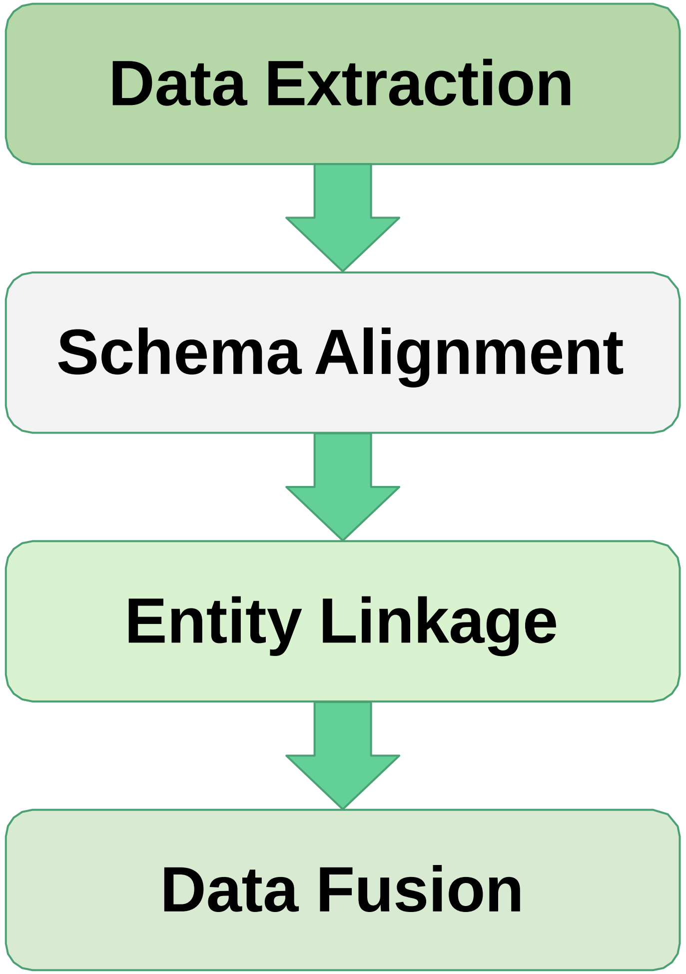
IMDB

WikiData



Importance from a practitioner's point of view

- Entity linkage is indispensable whenever integrating data from different sources
- Data extraction is important for integrating non-relational data
- Data fusion is necessary in presence of erroneous data
- Schema alignment is helpful when integrating relational data, but not affordable for manual work if we integrate many sources



Two main types of Machine Learning

- Supervised learning: learn by examples
- Unsupervised learning: find structure w/o examples

	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

DI & ML as synergy

- **ML for effective DI: AUTOMATION**
  - Automating DI tasks with training data
  - Better understanding of semantics by neural network

- **DI for effective ML: DATA**
  - Create large-scale training datasets from different sources
  - Cleaning of data used for training
  - Refer to the Data Curation lecture earlier

Many systems where DI & ML leverage each other

NELL

MacroBase

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Magellan

HoloClean

snorkel

Dedupe.io

KNOWLEDGE VAULT

BigGorilla

tamr

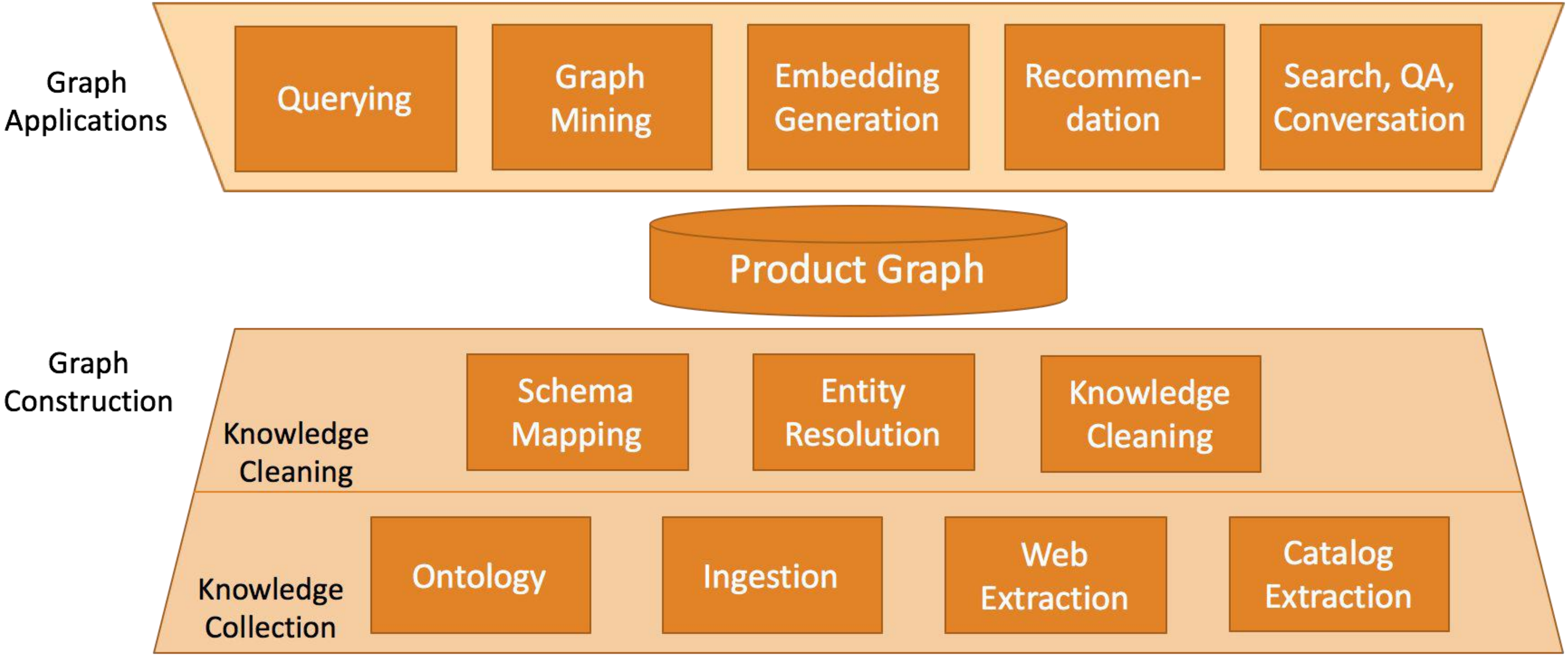
amperity

product graph

TRIFACTA

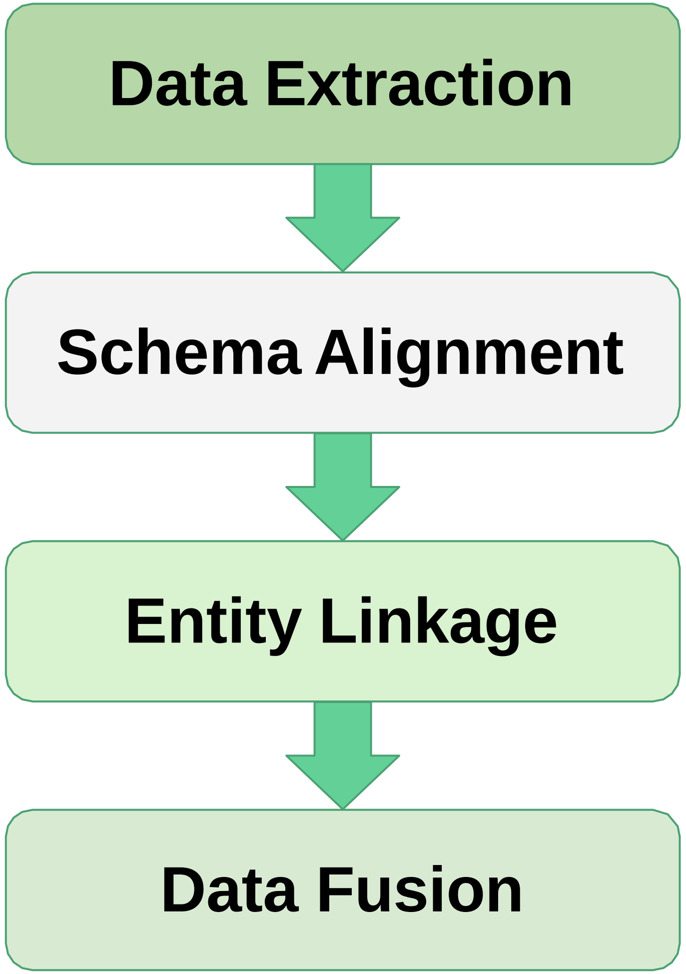
Increasing number of systems in industry and academia.

Example system: Product Graph [Dong, KDD'18]



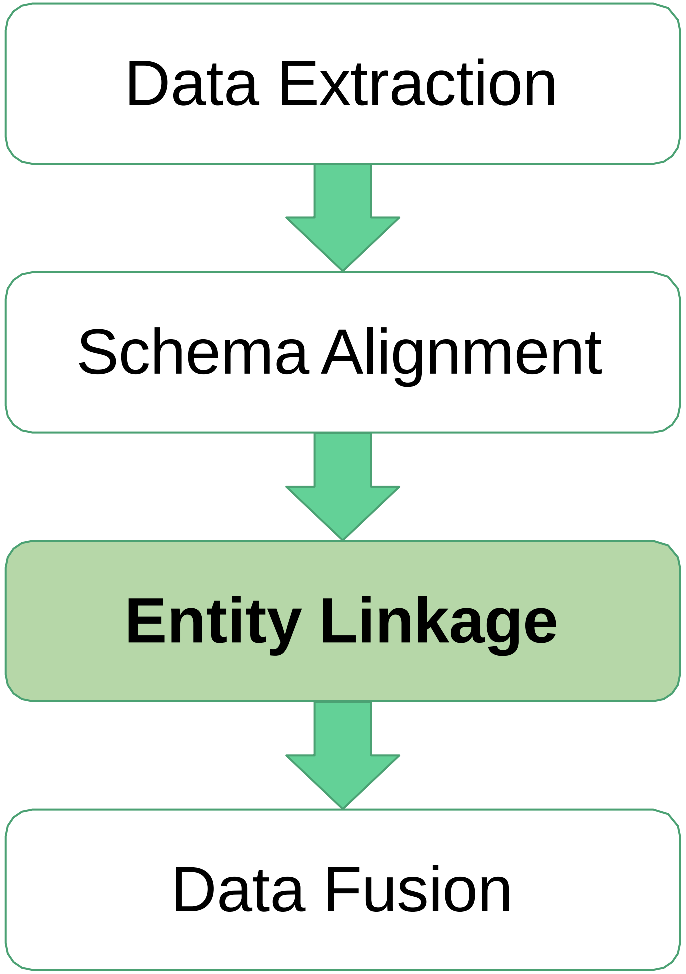
Data integration overview

- Entity linkage: linking records to entities; indispensable when different sources exist
- Data extraction: extracting structured data; important when non-relational data exist
- Data fusion: resolving conflicts; necessary in presence of erroneous data
- Schema alignment: aligning types and attributes; helpful when different relational schemas exist



Today's agenda

- Part I. Introduction
- Part II. ML for DI
    - ML for entity linkage
    - ML for data extraction
    - ML for schema alignment
    - ML for data fusion



What is entity linkage?

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

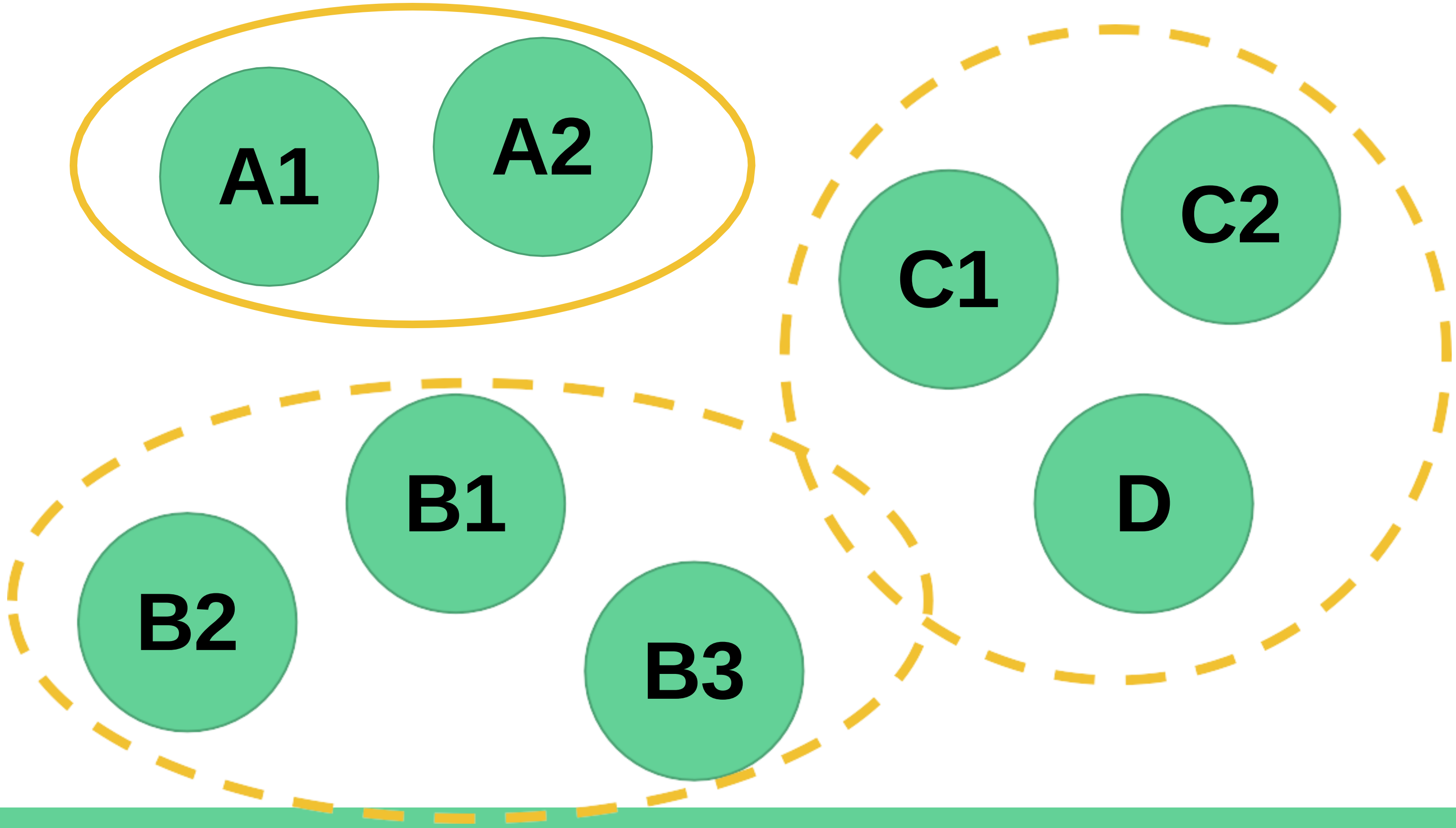
Are they the same entity?

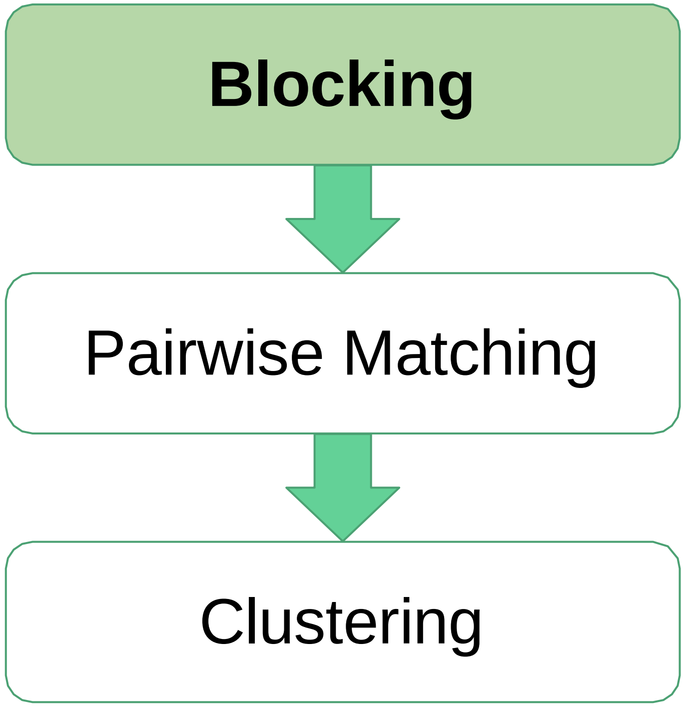
IMDB

WikiData

Quick tour for entity linkage

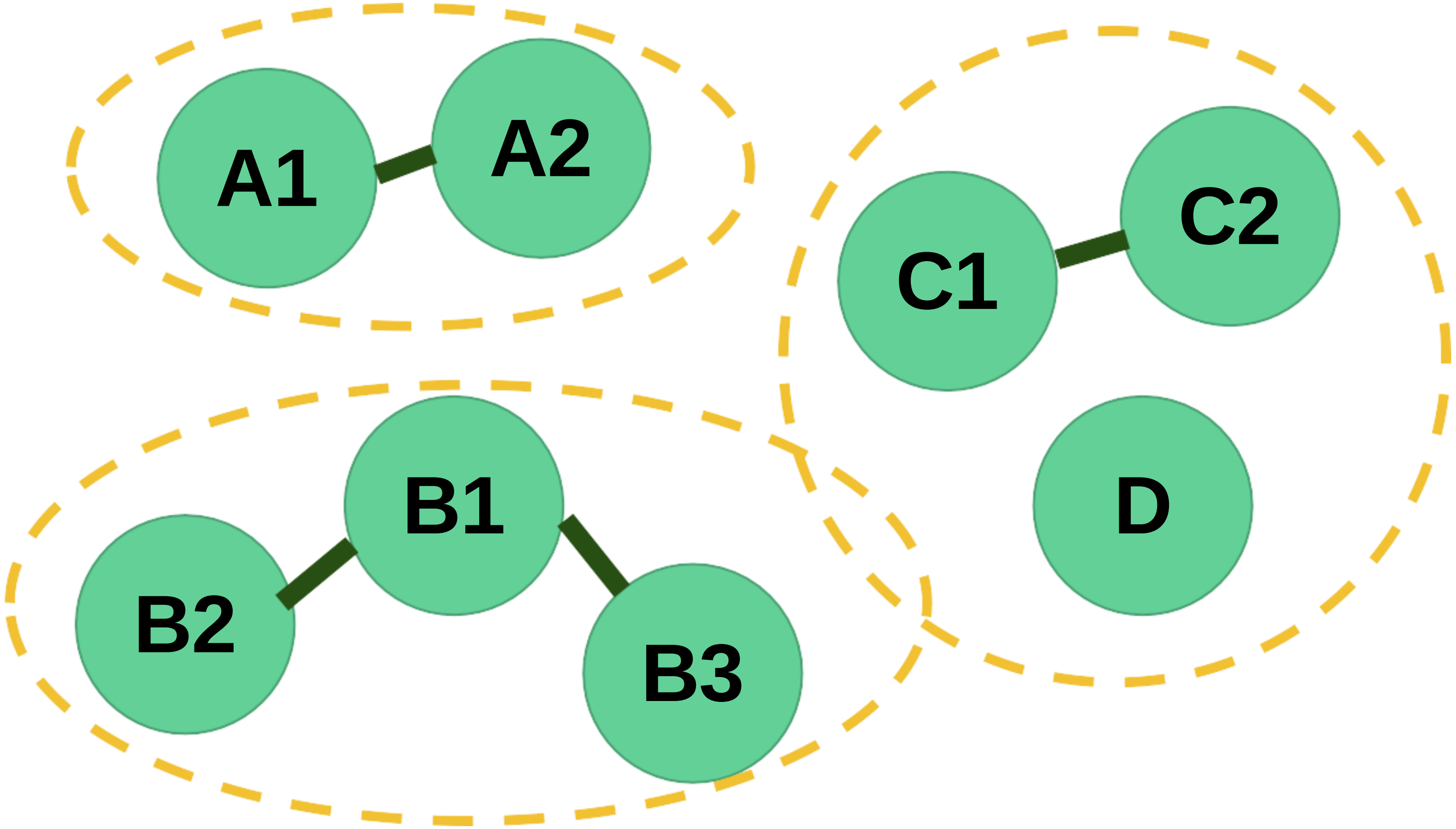
- **Blocking**: efficiently create small blocks of similar records

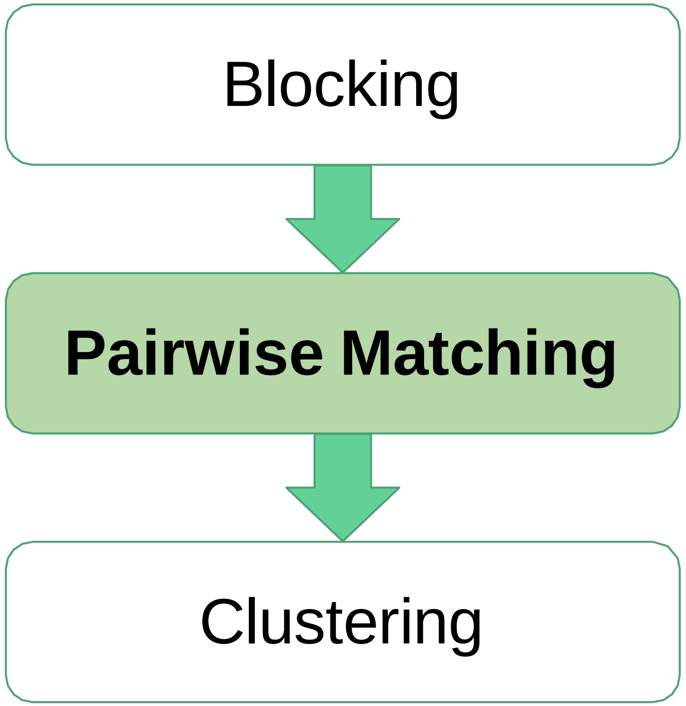




Quick tour for entity linkage

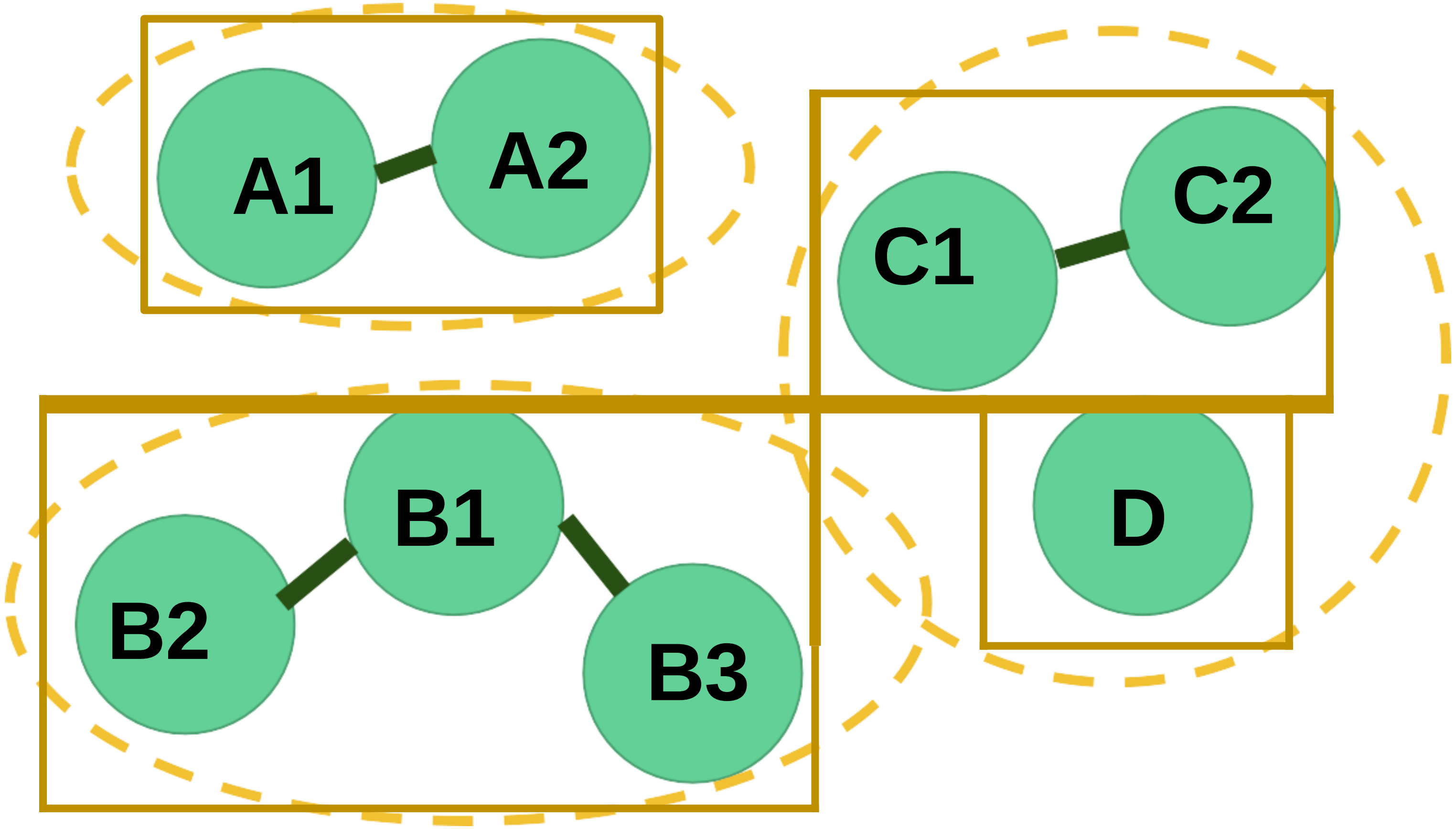
- **Pairwise matching**: compare all record pairs in a block

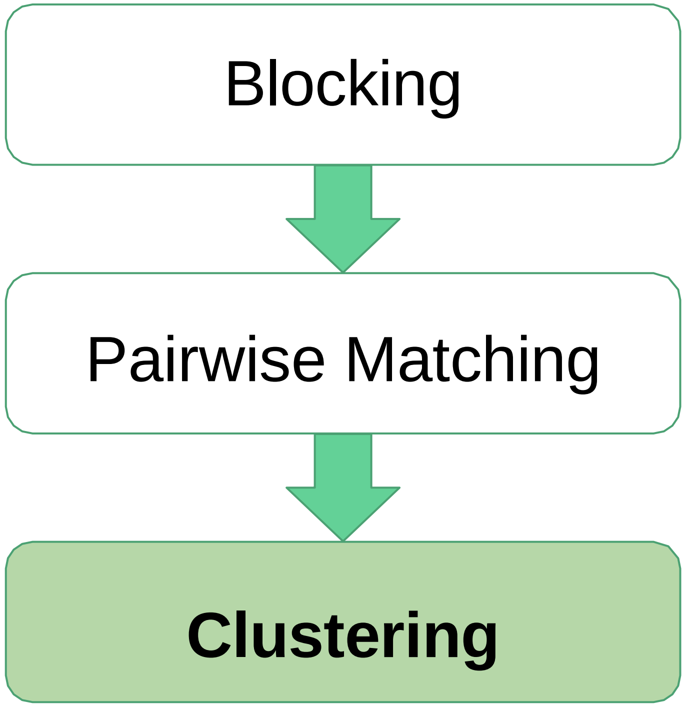




Quick tour for entity linkage

- **Clustering**: group records into entities





50 years of entity linkage

1969 (Pre-ML)

Rule-based and stats-based

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

~2000 (Early ML)

Sup / Unsup learning

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

~2015 (ML)

Supervised learning

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

2018 (Deep ML)

Deep learning

- Deep learning
- Entity embedding

Rule-based solution

Rule-based and stats-based

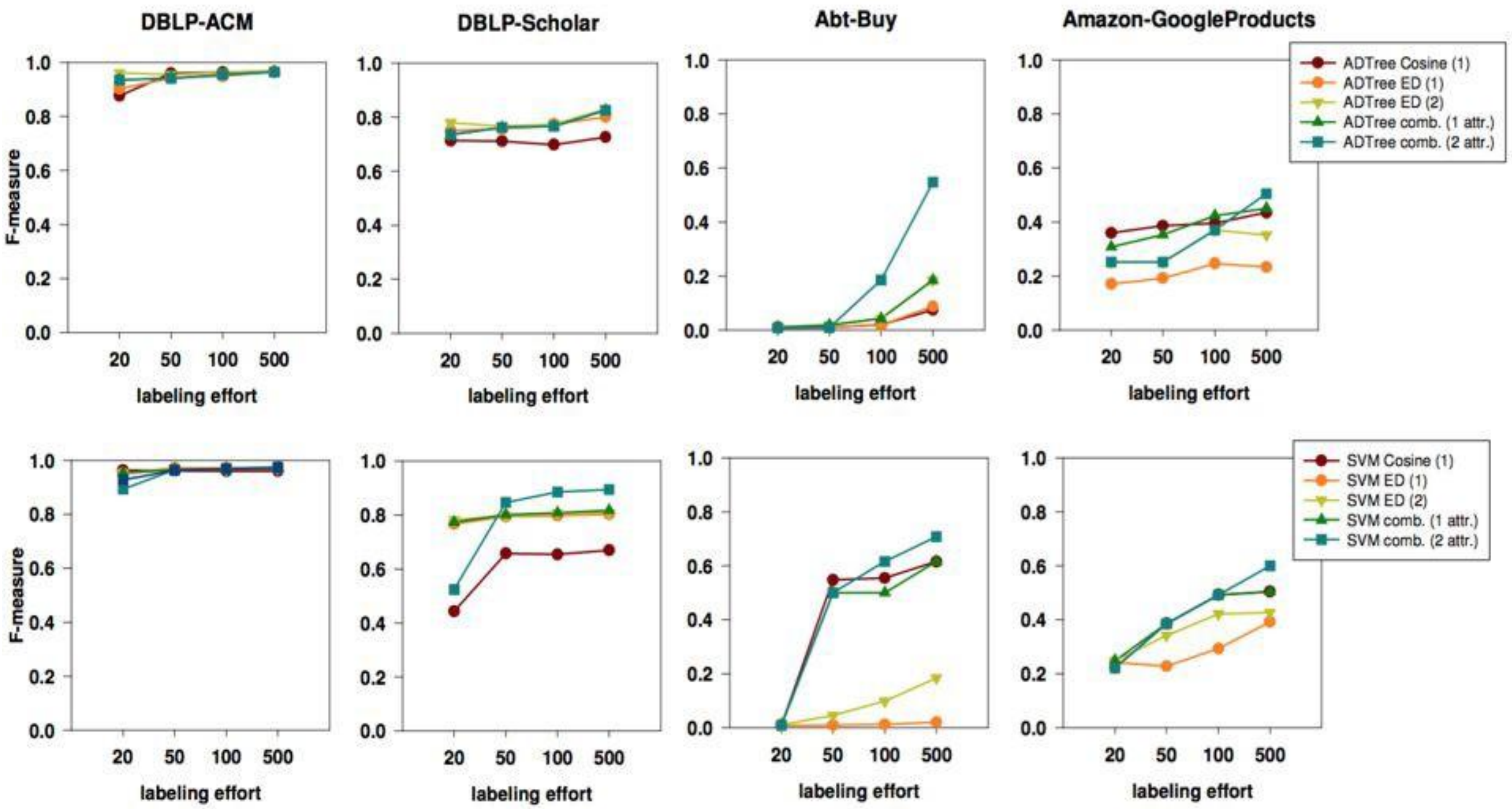
- Blocking: e.g., same name
- Matching: e.g., avg similarity of attribute values
- Clustering: e.g., transitive closure, etc.

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]



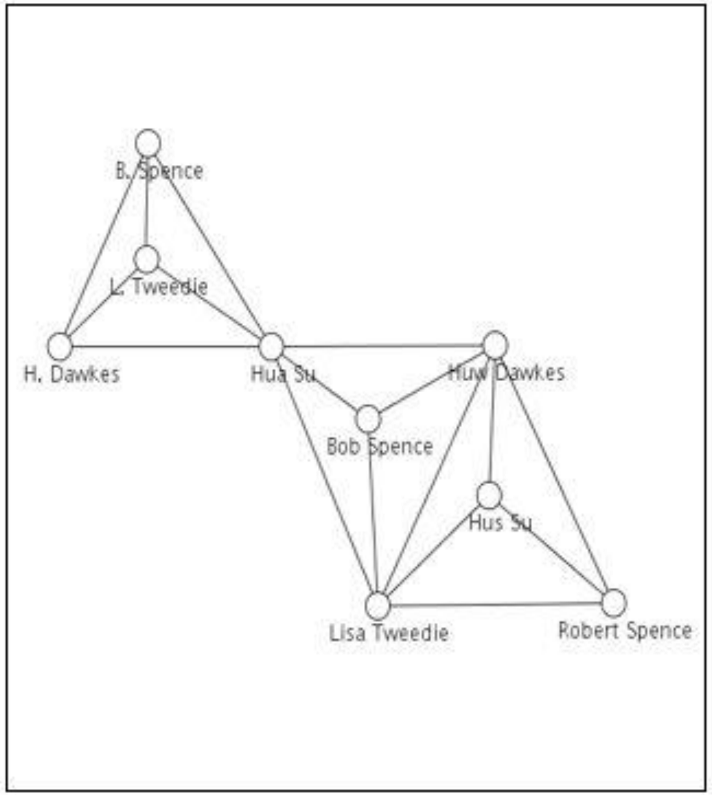
~2000 (Early ML)

Sup / Unsup learning

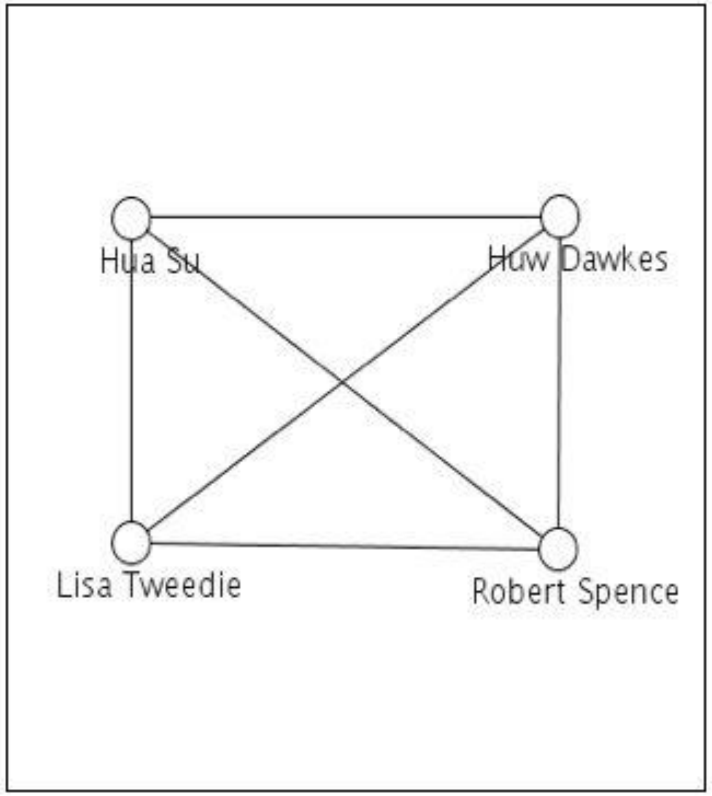
- Matching: Decision tree, SVM
  F-msr: 70%-90% w. 500 labels
- Clustering: Correlation clustering, Markov clustering

Collective entity resolution: beyond pairs

- Collective reasoning across entities.
- Constraints across entities:
  - Aggregate constraints
  - Transitivity, Exclusivity
  - Functional dependencies
- Use of probabilistic graphical models, etc., to capture such domain knowledge



before



after

[Example by Getoor and Machanavajjhala]

Classic 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

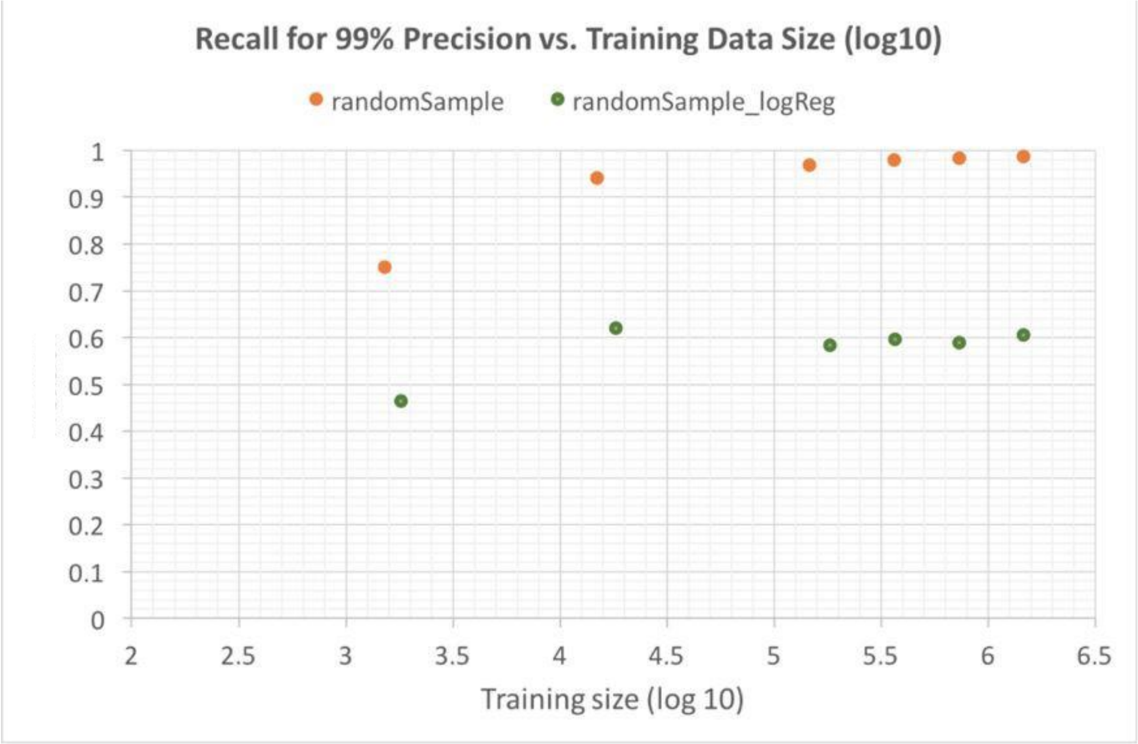
Classic ML models [Dong, KDD'18]

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



Classic 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

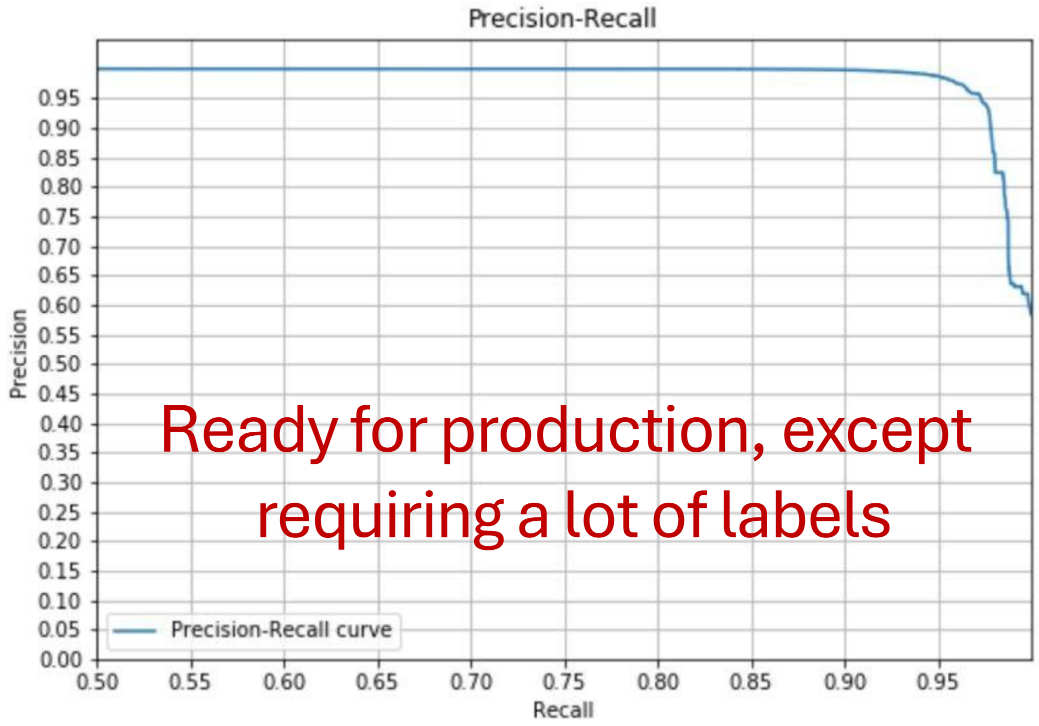
Classic ML models [Dong, KDD’18]

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



Ready for production, except requiring a lot of labels

Classic ML models [Dong, KDD'18]

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)

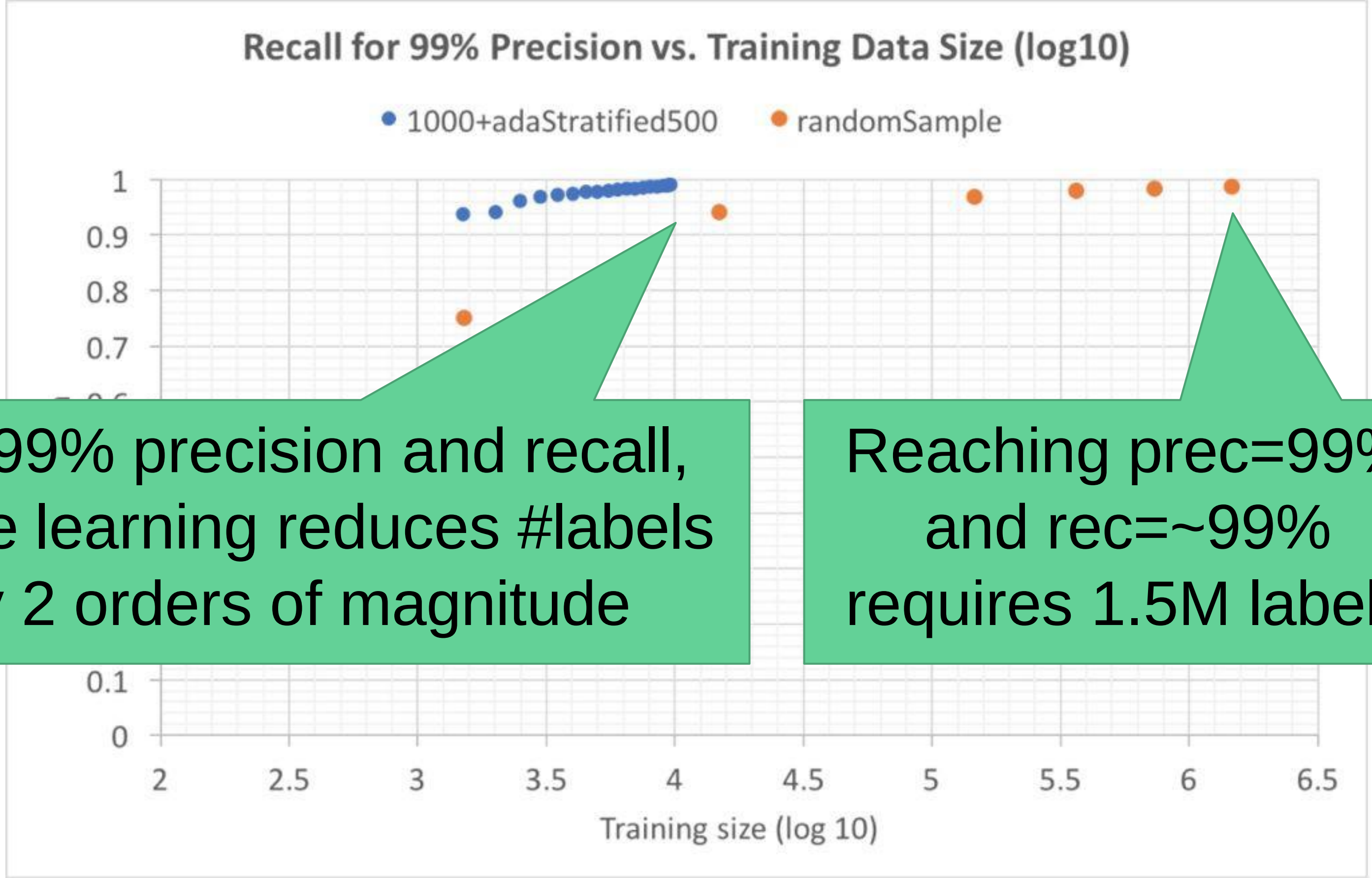
Classic ML models [Dong, KDD'18]

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)

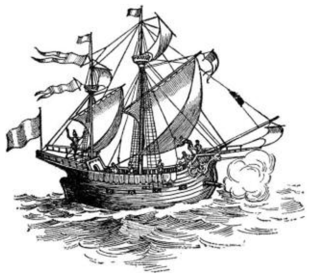
- Apply active learning to minimize #labels



For 99% precision and recall, active learning reduces #labels by 2 orders of magnitude

Reaching prec=99% and rec=~99% requires 1.5M labels

Deep learning models [Mudgal et al., SIGMOD'18]



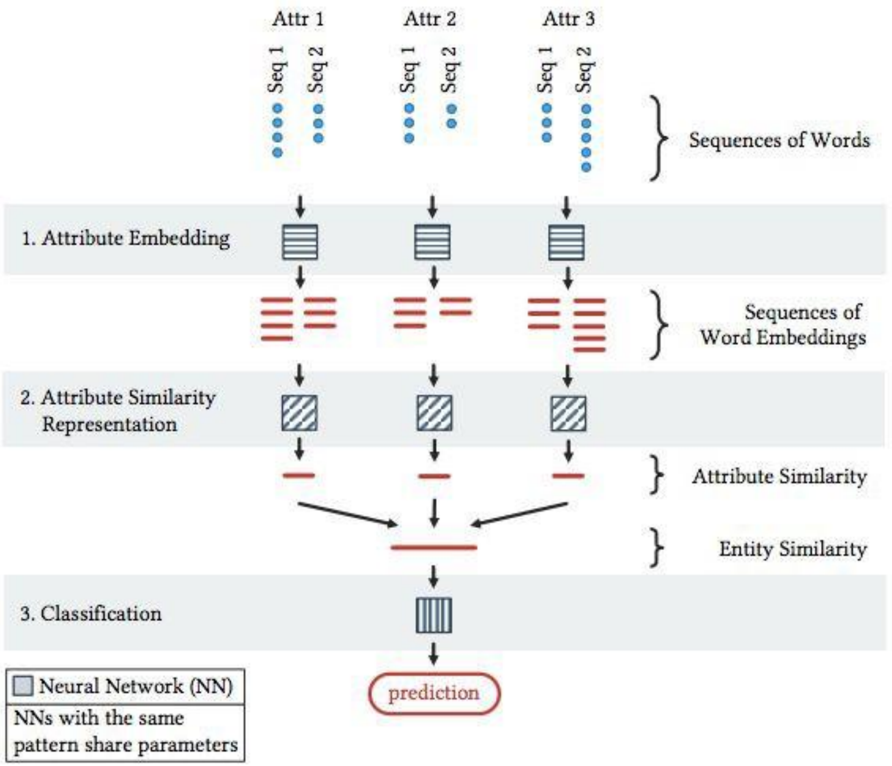
Magellan

- Embedding on similarities
- Similar performance for structured data; Significant improvement on texts and dirty data

2018 (Deep ML)

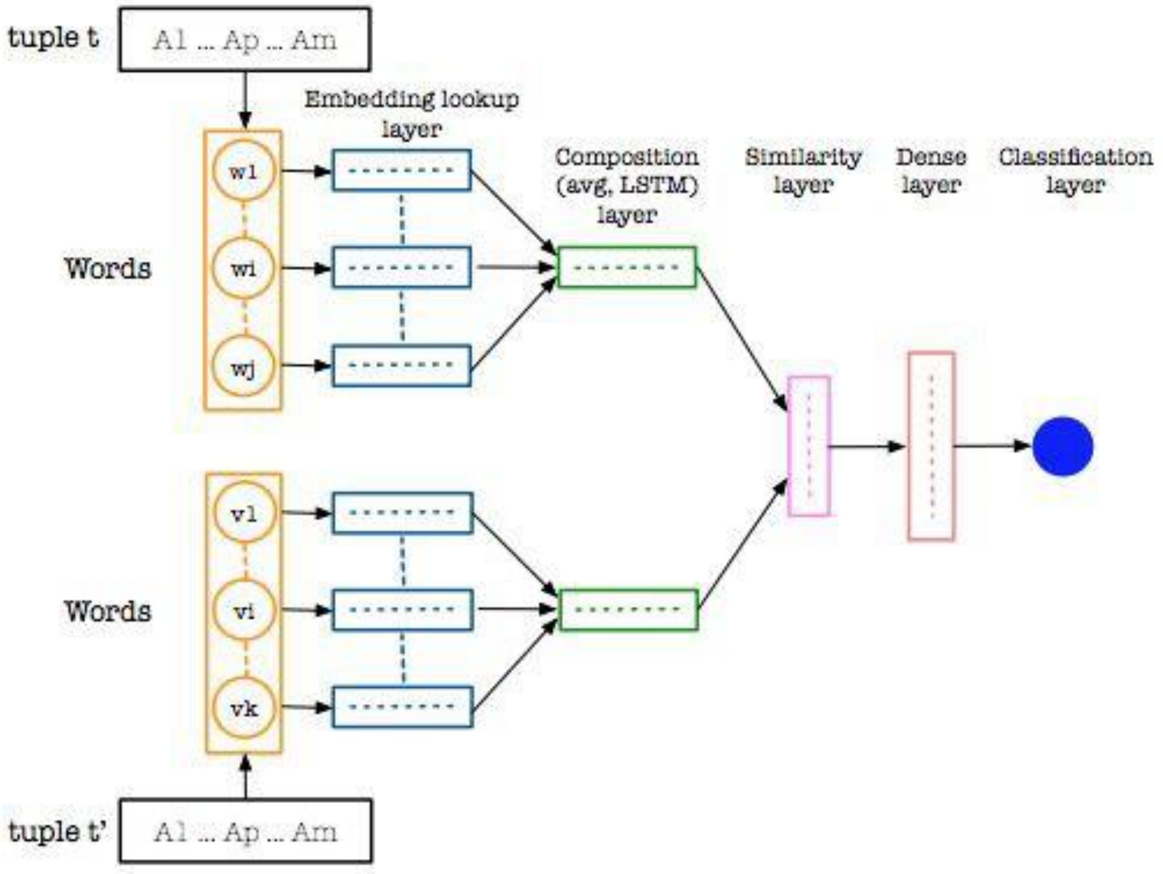
Deep learning

- Deep learning
- Entity embedding



Deep learning models [Ebraheem et al., VLDB’18]

- Embedding on entities
- Outperforming existing solution



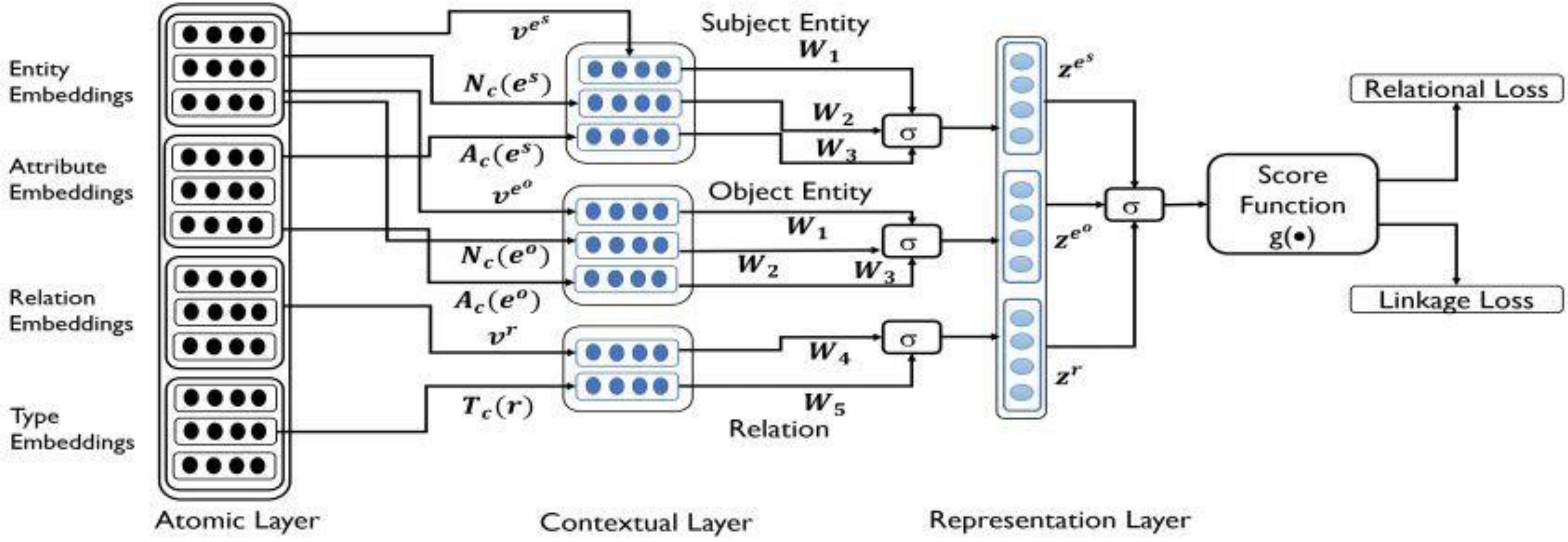
2018 (Deep ML)

Deep learning

- Deep learning
- Entity embedding

Deep learning models [Trivedi et al., ACL'18]

- LinkNBed: Embeddings for entities as in knowledge embedding



Deep learning models [Trivedi et al., ACL’18]

2018 (Deep ML)

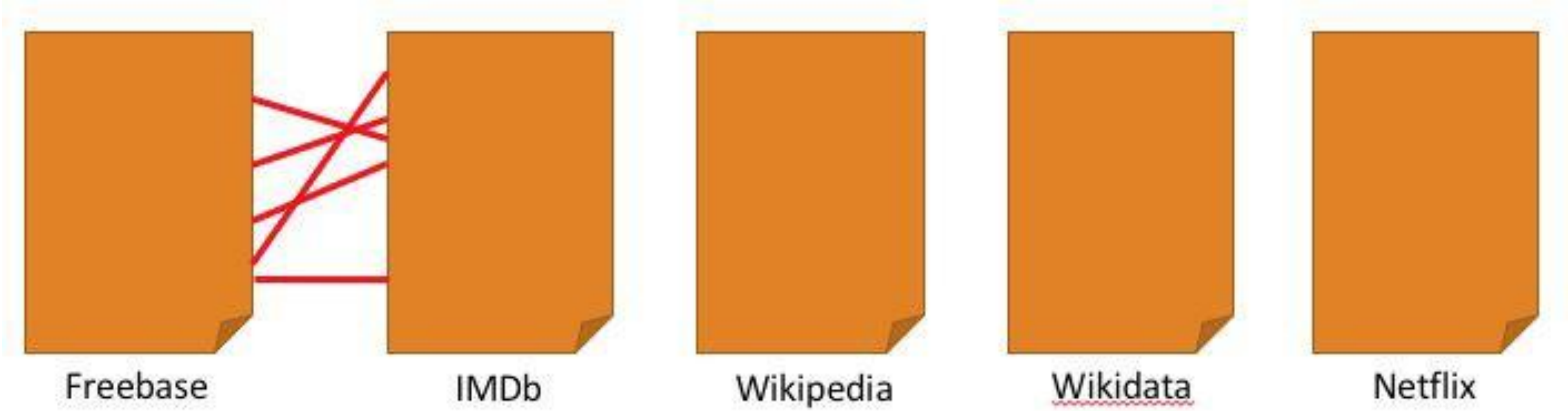
Deep learning

- Deep learning
- Entity embedding

- LinkNBed: 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

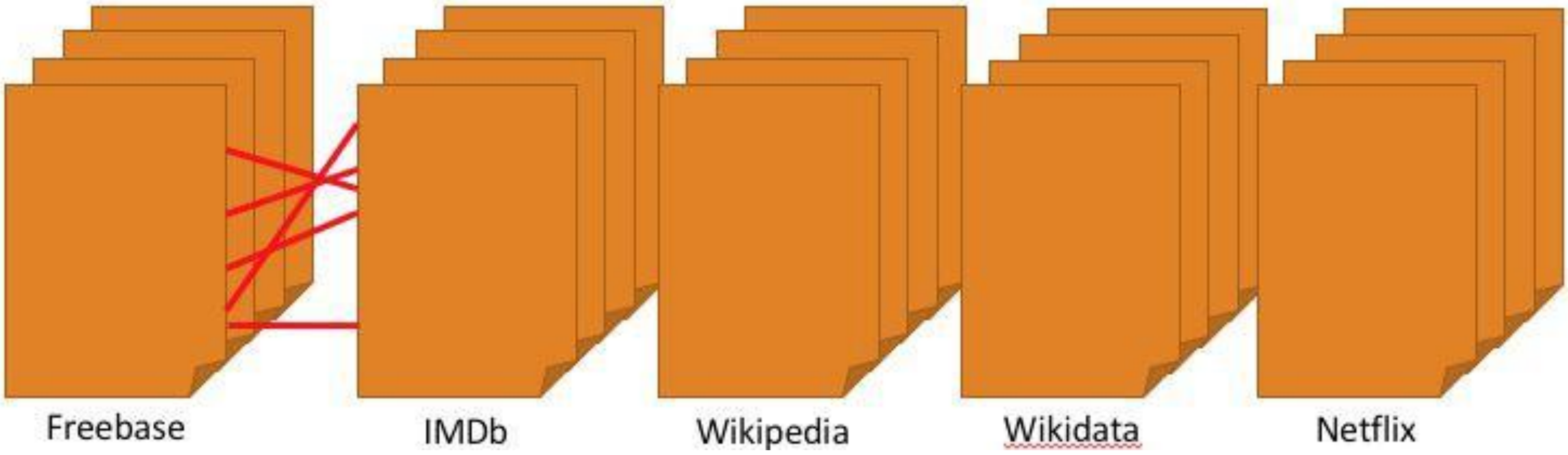
Challenges in applying ML on EL

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



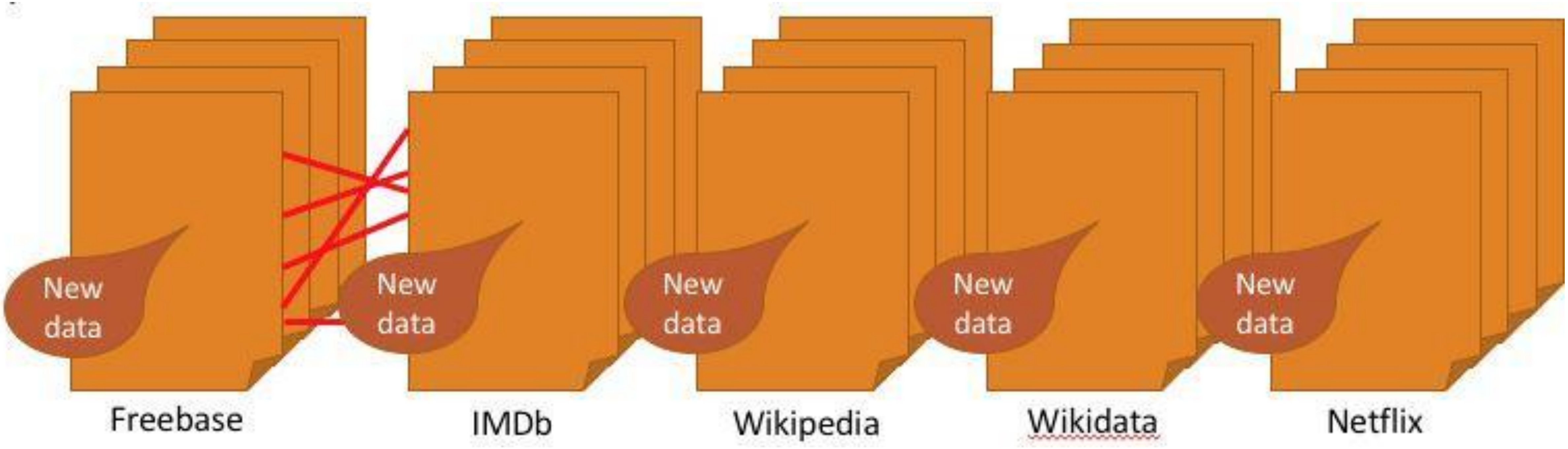
Challenges in applying ML on EL

- **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



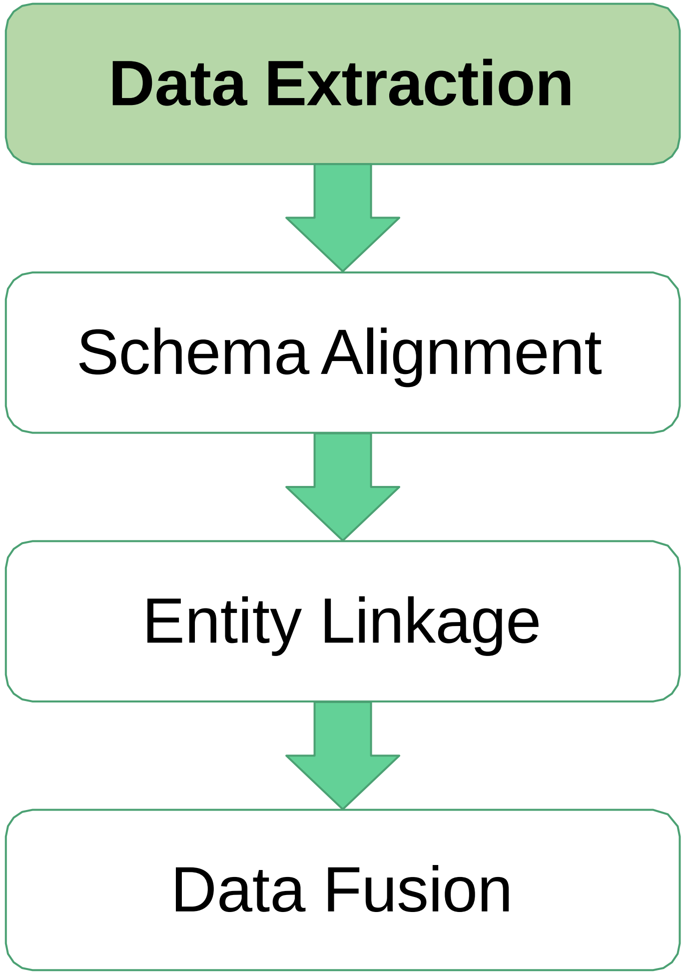
Recipe for entity linkage

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

Data Extraction
↓
Schema Alignment
↓
Entity Linkage
↓
Data Fusion

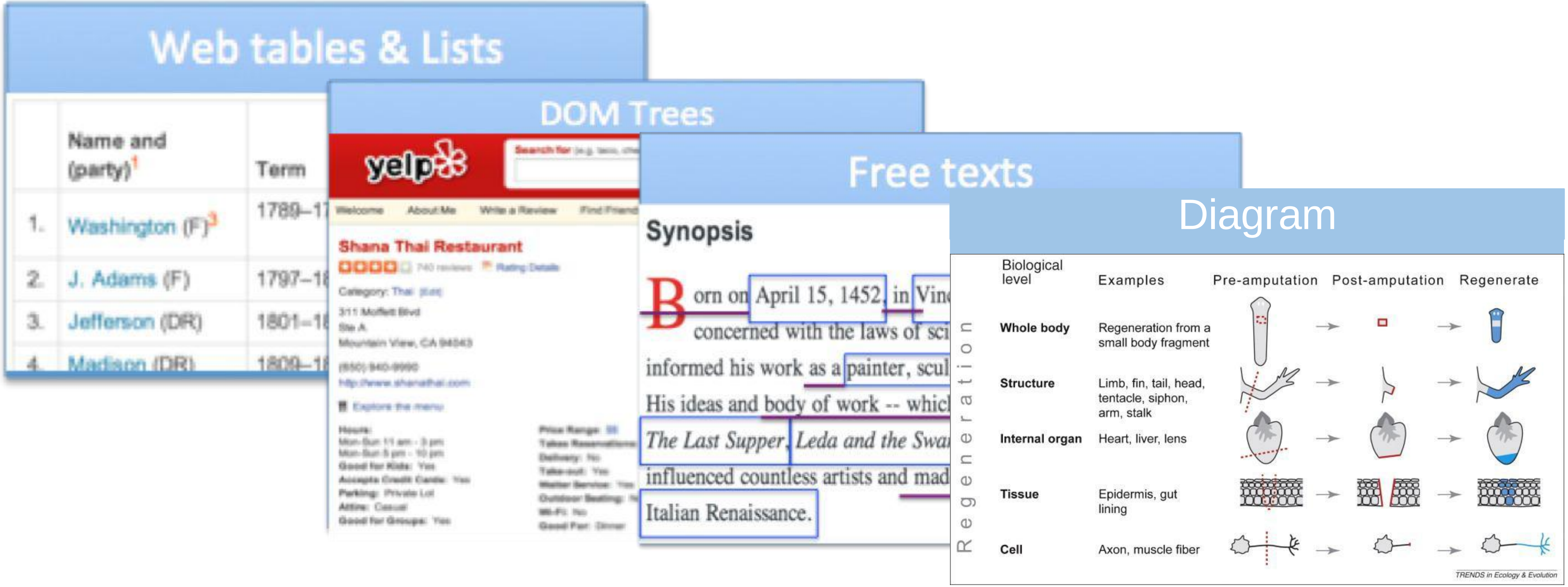
Today's agenda

- Part I. Introduction
- Part II. ML for DI
  - ML for entity linkage
  - ML for data extraction
  - ML for schema alignment
  - ML for data fusion



What is data extraction?

- Definition: Extract structured information, e.g., (entity, attribute, value) triples, from semi-structured data or unstructured data.



Three types of data extraction

- **Closed-world extraction**: align to existing entities and attributes; e.g., (ID_Obama, place_of_birth, ID_USA)
- **ClosedIE**: align to existing attributes, but extract new entities; e.g., (“Xin Luna Dong”, place_of_birth, “China”)
- **OpenIE**: not limited by existing entities or attributes; e.g.,
(“Xin Luna Dong”, “was born in”, “China”),
(“Luna”, “is originally from”, “China”)

35 years of data extraction

1992 (Rule-based)

Early Extraction

- Rule-based: Hearst pattern, IBM System T
- Tasks: IS-A, events

~2005 (Rel. Ex.)

Relation extraction from texts

- NER→EL→RE
  - Feature based: LR, SVM
  - Kernel based: SVM
- Distant supervision
- OpenIE

2008 (Semi-stru)

Extraction from semi-structured data

- WebTables: search, extraction
- DOM tree: wrapper induction

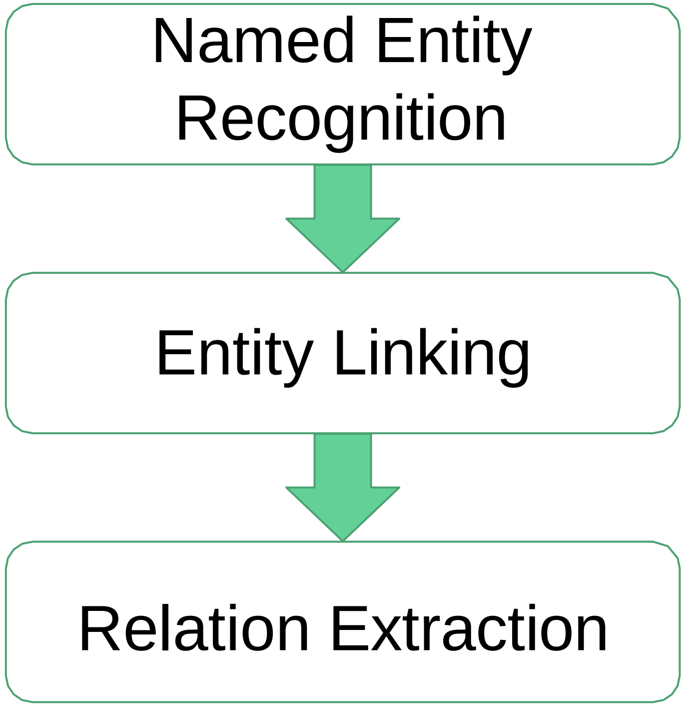
2013 (Deep ML)

Deep learning

- Use RNN, CNN, attention for RE
- Data programming / Heterogeneous learning
- Revisit DOM extraction

Extraction from texts: quick example

Bill Gates founded Microsoft in 1975.



Extraction from texts: quick example

Bill Gates founded Microsoft in 1975.

Person

Company

Named Entity Recognition

Entity Linking

Relation Extraction

Extraction from texts: quick example

Bill Gates founded Microsoft in 1975.

Named Entity Recognition

↓

Entity Linking

↓

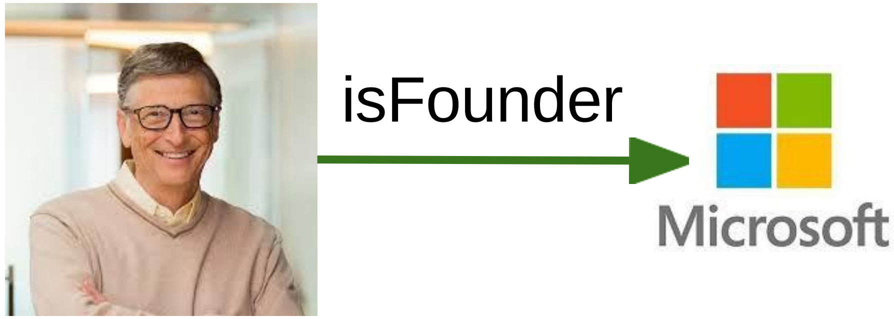
Relation Extraction

Entity **linkage**: linking two structured records

Entity **linking**: linking a phrase in texts to an entity in a reference list (e.g., knowledge graph)

Extraction from texts: quick example

Bill Gates founded Microsoft in 1975.



We focus on Relation Extraction.

Named Entity Recognition

Entity Linking

Relation Extraction

Extraction from texts: feature based [Zhou et al., ACL'05]

~2005 (Rel. Ex.)

Relation extraction from texts

- NER→EL→RE
  - Feature based: LR, SVM
  - Kernel based: SVM
- Distant supervision
- OpenIE

- **Models**
  - Logistic regression
  - SVM (Support Vector Machine)
- **Features**
  - Lexical: entity, part-of-speech, neighbor
  - Syntactic: **chunking**, parse tree
  - Semantic: concept hierarchy, entity class
- **Results**
  - Prec=~60%, Rec=~50%

Extraction from texts: feature based [Zhou et al., ACL'05]

~2005 (Rel. Ex.)

Relation extraction from texts

- NER→EL→RE
  - Feature based: LR, SVM
  - Kernel based: SVM
- Distant supervision
- OpenIE

Features	P	R	F
Words	69.2	23.7	35.3
+Entity Type	67.1	32.1	43.4
+Mention Level	67.1	33.0	44.2
+Overlap	57.4	40.9	47.8
+Chunking	61.5	46.5	53.0
+Dependency Tree	62.1	47.2	53.6
+Parse Tree	62.3	47.6	54.0
+Semantic Resources	63.1	49.5	55.5

Table 2: Contribution of different features over 43 relation subtypes in the test data

Major Lift

Extraction from texts: kernel based [Mengqiu Wang, IJCNLP'08]

~2005 (Rel. Ex.)

Relation extraction from texts

- NER→EL→RE
  - Feature based: LR, SVM
  - Kernel based: SVM
- Distant supervision
- OpenIE

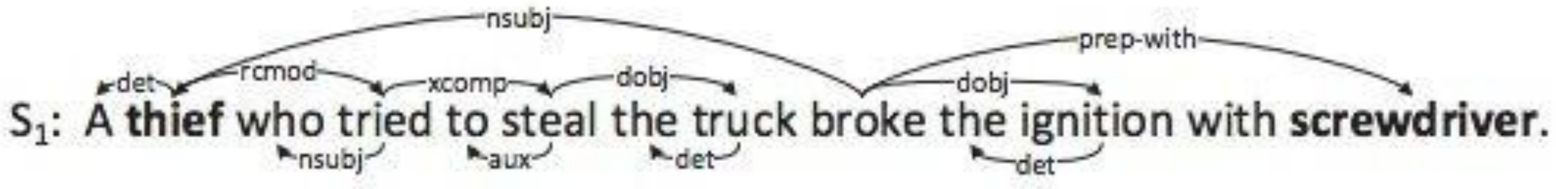
- **Models**
  - SVM (Support Vector Machine)
- **Kernels**
  - Subsequence
  - Dependency tree
  - **Shortest dependency path**
  - Convolution dependency

Extraction from texts: kernel based [Mengqiu Wang, IJCNLP'08]

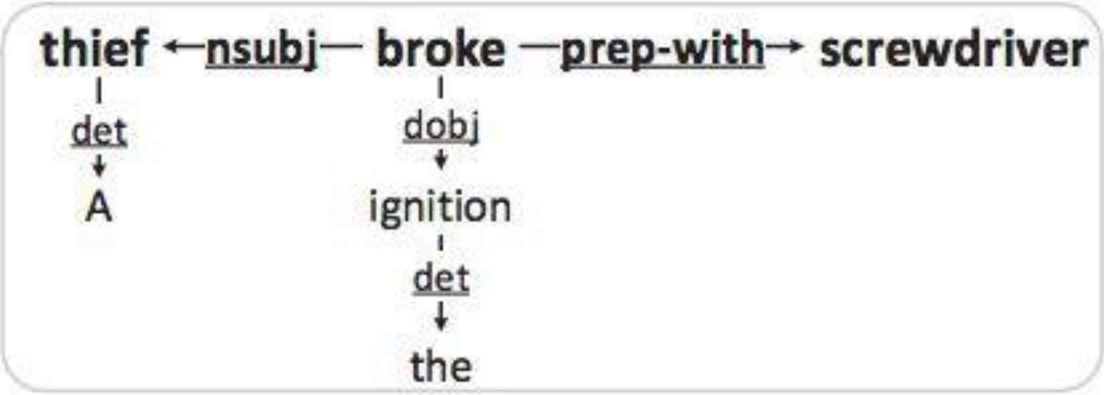
~2005 (Rel. Ex.)

Relation extraction from texts

- NER→EL→RE
  - Feature based: LR, SVM
  - Kernel based: SVM
- Distant supervision
- OpenIE



Dependency tree



Shortest dependency path

Extraction from texts: kernel based [Mengqiu Wang, IJCNLP'08]

~2005 (Rel. Ex.)

Relation extraction from texts

- NER→EL→RE
  - Feature based: LR, SVM
  - Kernel based: SVM
- Distant supervision
- OpenIE

- **Models**
  - SVM (Support Vector Machine)
- **Kernels**
  - Subsequence
  - Dependency tree
  - **Shortest dependency path**
  - Convolution dependency
- **Results**
  - Prec=~70%, Rec=~40%

Extraction from texts: kernel based [Mengqiu Wang, IJCNLP'08]

~2005 (Rel. Ex.)

Relation extraction from texts

- NER→EL→RE
  - Feature based: LR, SVM
  - Kernel based: SVM
- Distant supervision
- OpenIE

kernel method	5-fold CV on ACE 2003 Precision	Recall	F1
subsequence	0.703	**0.389**	0.546
dependency tree	0.681	0.290	0.485
shortest path	**0.747**	0.376	**0.562**

Table 1: Results of different kernels on ACE 2003 training set using 5-fold cross-validation.

Extraction from Texts: deep learning

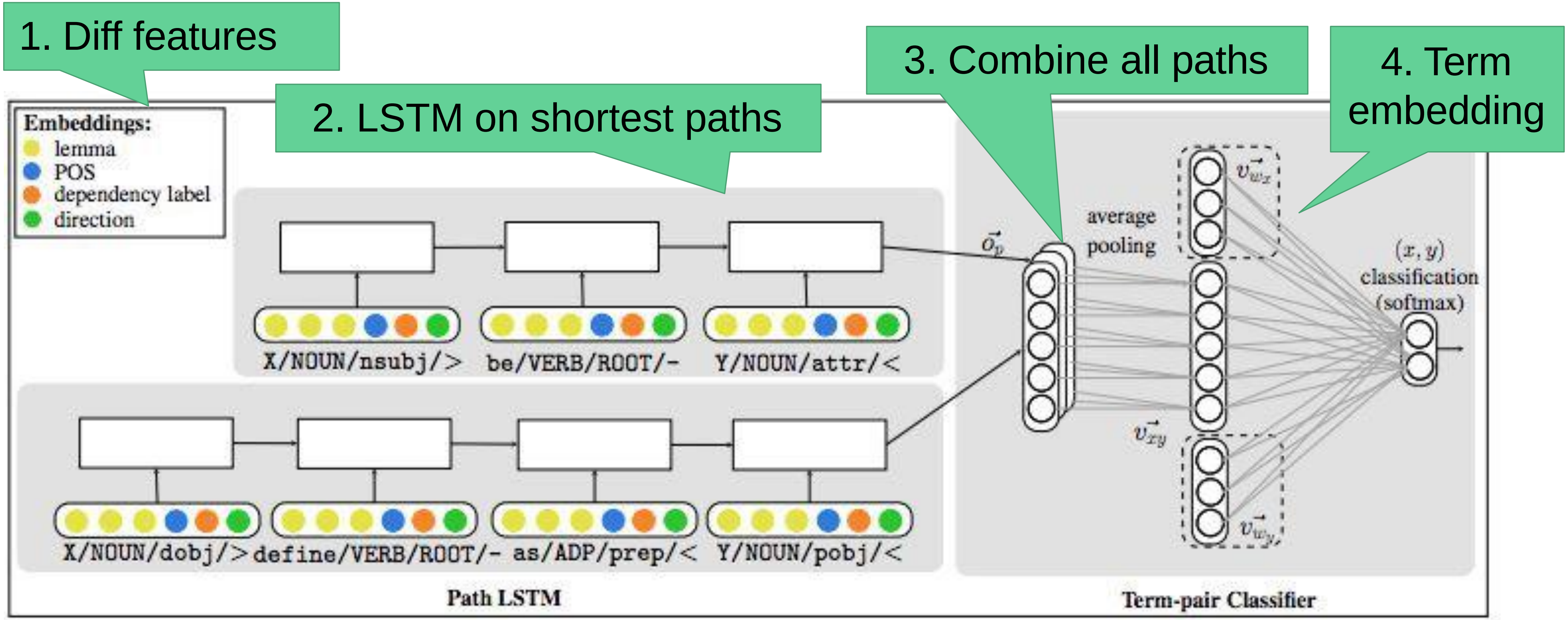
2013 (Deep ML)

Deep learning

- Use RNN, CNN, attention for RE
- Data programming / Heterogeneous learning
- Revisit DOM extraction

- **Same intuitions, different models**
  - (2012-13) Recursive NN: dependency tree
    [Socher et al., EMNLP'12] [Hashimoto et al., EMNLP'13]
  - (2014-15) CNN: shortest dependency path
    [Zeng et al., COLING'14][Liu et al., ACL'15]
  - (2015+) LSTM: shortest dependency path, lexical/syntactic/semantic features
    [Xu et al., EMNLP'15][Shwartz et al., ACL'16]
    [Nguyen, NAACL'16]

Example system: HyperNET [Shwartz et al., ACL'16]



Quality in identifying hypernyms: Prec = 0.9, Rec = 0.9

Label generation for extraction training

Where are training labels from?

~2005 (Rel. Ex.)

Relation extraction from texts

- NER→EL→RE
  - Feature based: LR, SVM
  - Kernel based: SVM
- Distant supervision
- OpenIE

- **Semi-supervised learning**
  - Iterative extraction [Carlson et al., AAAI'10]
    Use new extractions to retrain models
    E.g., NELL

Iterations	Estimated Precision (%)	# Promotions
1–22	90	88,502
23–44	71	77,835
45–66	57	76,116

Label generation for extraction training

~2005 (Rel. Ex.)

Relation extraction from texts

- NER→EL→RE
  - Feature based: LR, SVM
  - Kernel based: SVM
- Distant supervision
- OpenIE

Where are training labels from?

- **Semi-supervised learning**
  - Iterative extraction [Carlson et al., AAAI'10]
    Use new extractions to retrain models
    E.g., NELL
- **Weak learning**
  - Distant supervision [Mintz et al., ACL'09]
    Rule-based annotation with seed data
    E.g., DeepDive, Knowledge Vault

Will cover in "DI for ML"

Distant Supervision [Mintz et al., ACL'09]

Corpus Text

- Bill Gates founded Microsoft in 1975. Bill Gates, founder of Microsoft, … Bill Gates attended Harvard from …
- Google was founded by Larry Page …

Training Data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y

Freebase

- (Bill Gates, Founder, Microsoft) (Larry Page, Founder, Google)
- (Bill Gates, CollegeAttended, Harvard)

[Adapted example from Luke Zettlemoyer]

Distant Supervision [Mintz et al., ACL'09]

Corpus Text

- Bill Gates founded Microsoft in 1975. Bill Gates, founder of Microsoft, … Bill Gates attended Harvard from …
- Google was founded by Larry Page …

Freebase

- (Bill Gates, Founder, Microsoft) (Larry Page, Founder, Google)
- (Bill Gates, CollegeAttended, Harvard)

Training Data

(Bill Gates, Microsoft)
Label: Founder
Feature: X founded Y
Feature: X, founder of Y

(Bill Gates, Harvard)
Label: CollegeAttended
Feature: X attended Y

For negative examples, sample unrelated pairs of entities.

[Adapted example from Luke Zettlemoyer]

Label generation for extraction training

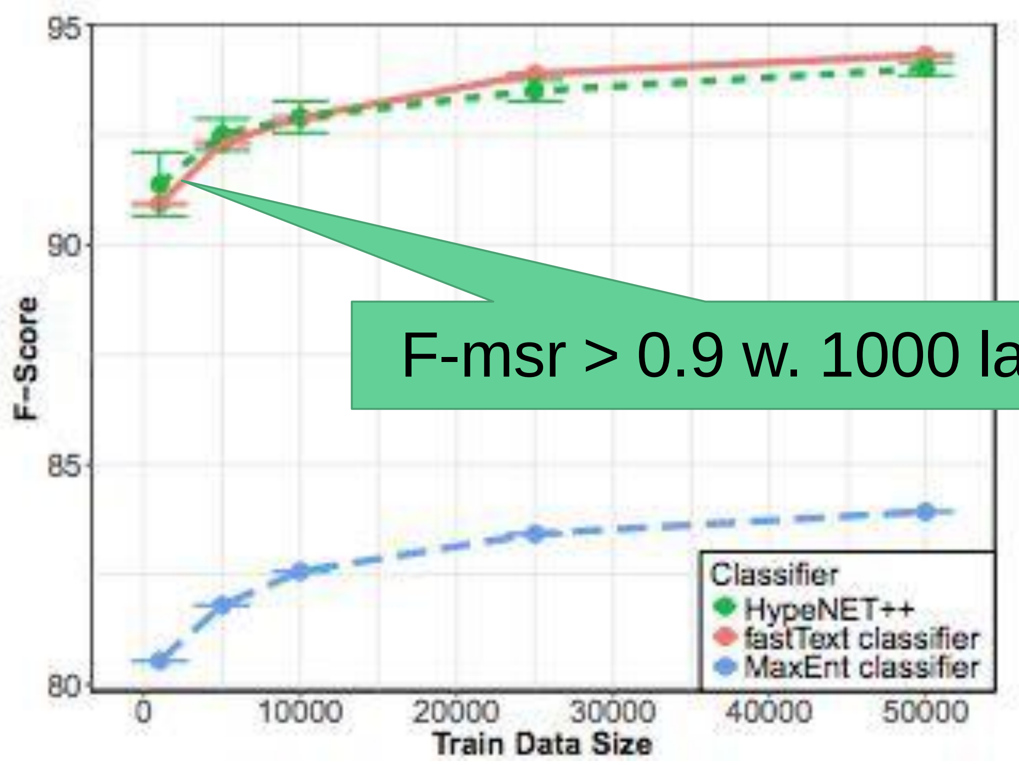
Where are training labels from?

~2005 (Rel. Ex.)

Relation extraction from texts

- NER→EL→RE
  - Feature based: LR, SVM
  - Kernel based: SVM
- Distant supervision
- OpenIE

- Distant supervision: HyperNet++
  [Christodoulopoulos & Mittal, 18]



F-msr > 0.9 w. 1000 labels

Label generation for extraction training

2013 (Deep ML)

Deep learning

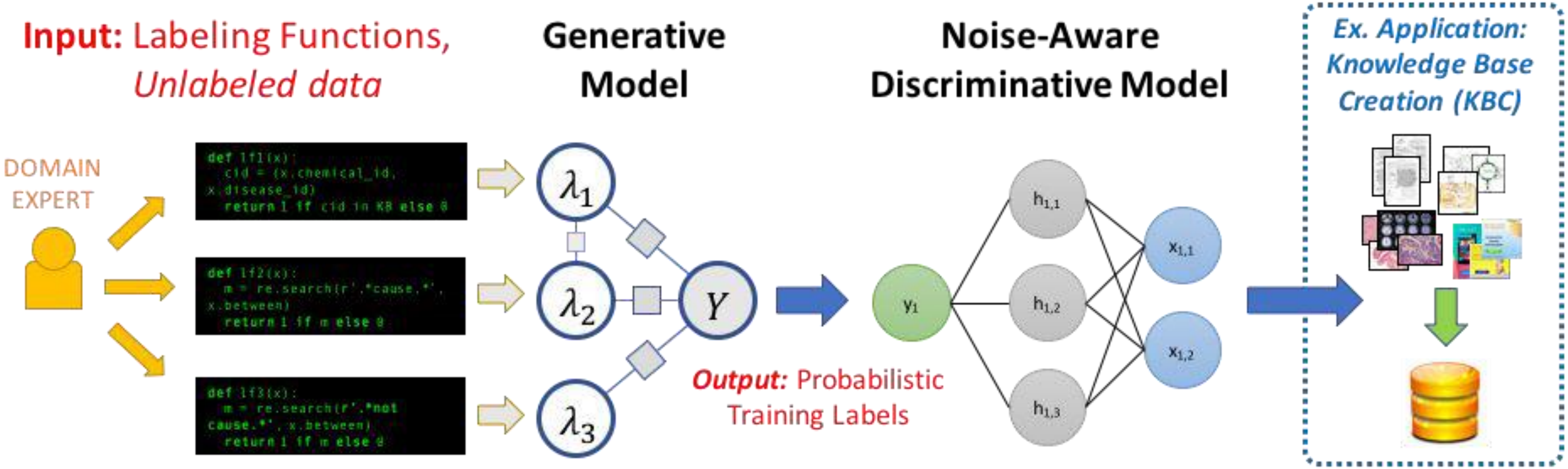
- Use RNN, CNN, attention for RE
- Data programming / Heterogeneous learning
- Revisit DOM extraction

Will cover in “DI for ML”

Where are training labels from?

- **Semi-supervised learning**
  - Iterative extraction [Carlson et al., AAAI’10]
    Use new extractions to retrain models
    E.g., NELL
- **Weak learning**
  - Distant supervision [Mintz et al., ACL’09]
    Rule-based annotation with seed data
    E.g., DeepDive, Knowledge Vault
  - Data programming [Ratner et al., NIPS’16]
    Manually write labelling functions
    E.g., Snorkle, Fouduer

Snorkel: code as supervision [Ratner et al., NIPS'16, VLDB'18]

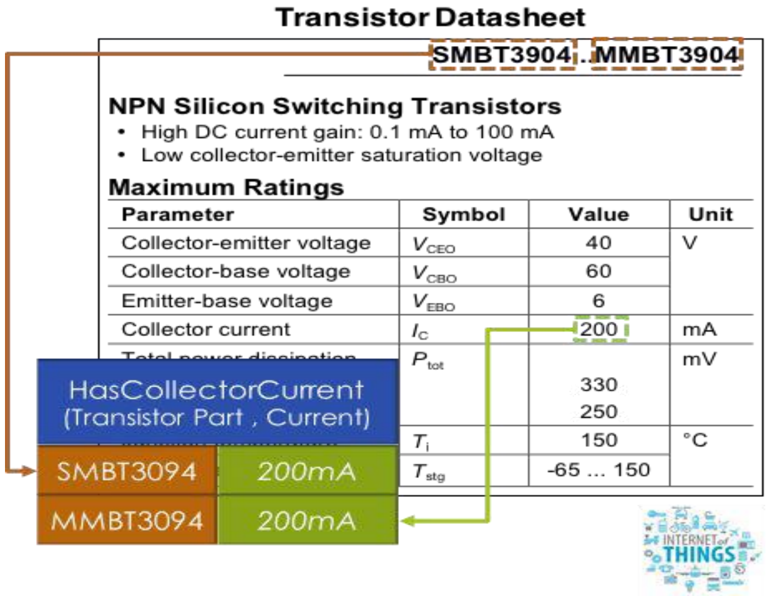


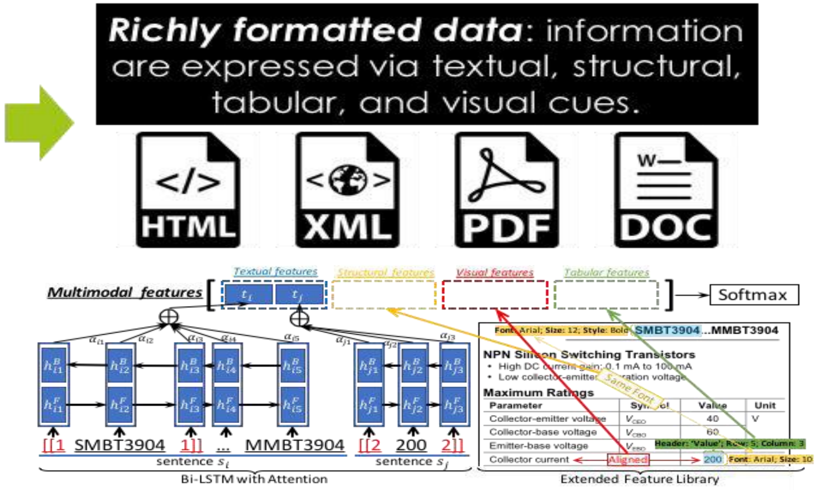
1. Users write *labeling functions* to generate noisy labels

2. We model the labeling functions' behavior to de-noise them

3. We use the resulting prob. labels to train a model

Example system: Fonduer [Wu et al., SIGMOD'18]





Fonduer combines a new **biLSTM with multimodal features** and **data programming**.

System	ELEC.	GEN.	
Knowledge Base	Digi-Key	GWAS Central	GWAS Catalog
# Entries in KB	376	3,008	4,023
# Entries in Fonduer	447	6,420	6,420
Coverage	0.99	0.82	0.80
Accuracy	0.87	0.87	0.89
# New Correct Entries	17	3,154	2,486
Increase in Correct Entries	1.05×	1.87×	1.42×

Code: https://github.com/HazyResearch/fonduer

Extraction from semi-structured data

Extraction from semi-structured data

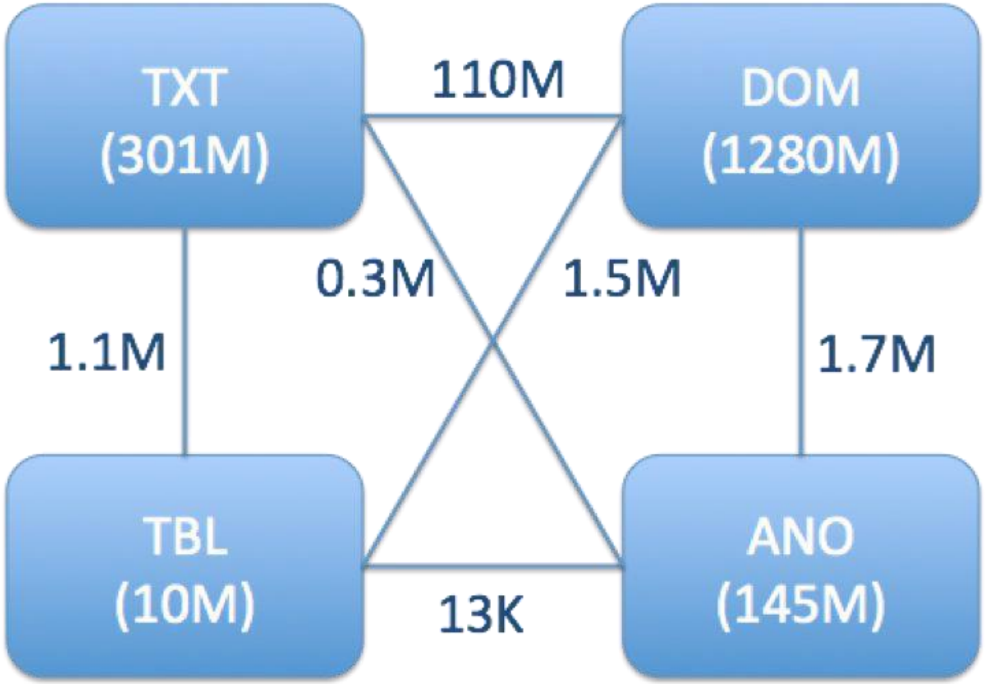
- WebTables: search, extraction
- DOM tree: wrapper induction

2008 (Semi-stru)

Why semi-structured data?

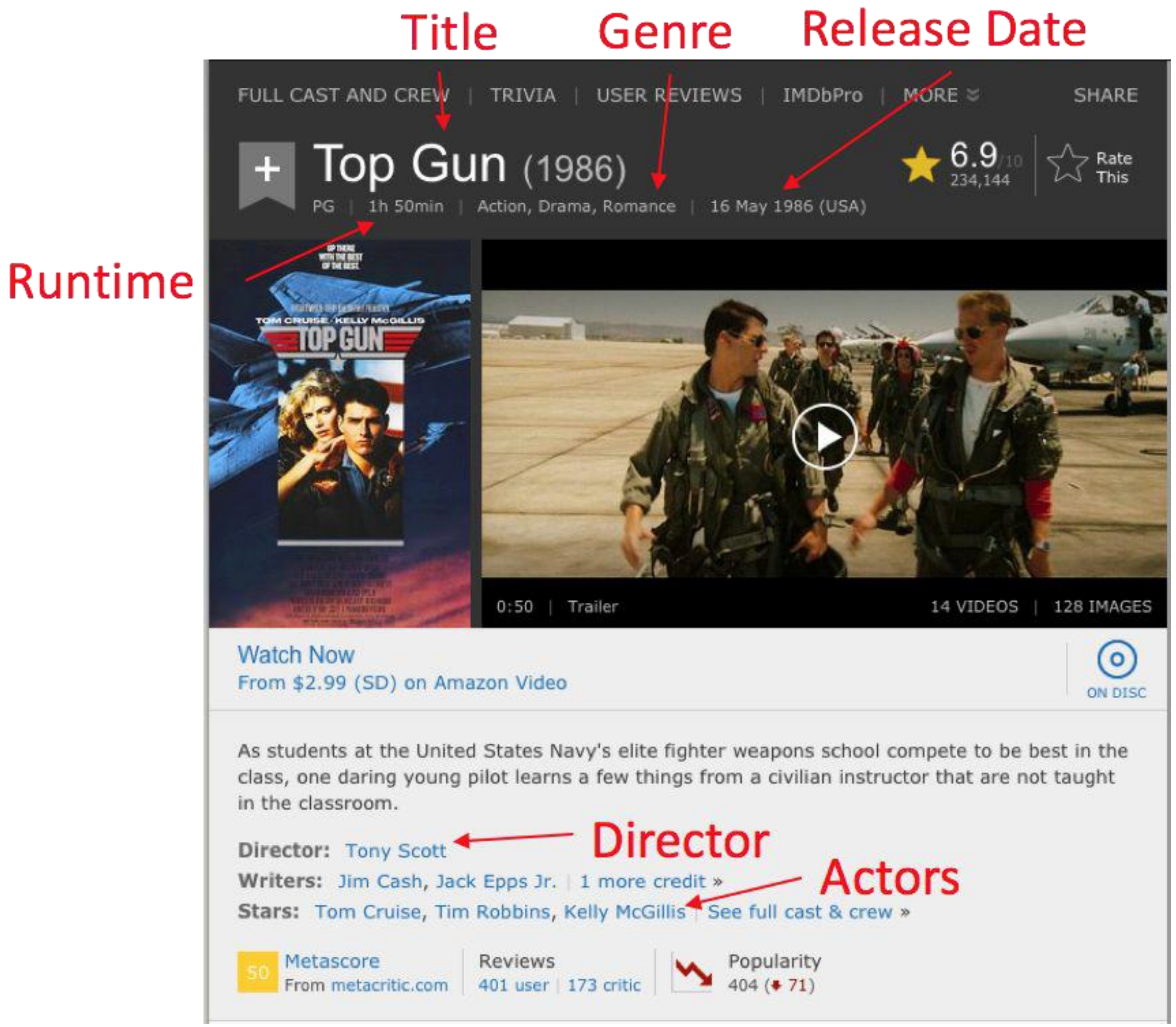
- Knowledge Vault @Google showed big potential from DOM-tree extraction [Dong et al., KDD'14][Dong et al., VLDB'14]

Accu	Accu (conf $\geq$ .7)
0.36	0.52



Accu	Accu (conf $\geq$ .7)
0.43	0.63
0.09	0.62

Wrapper Induction--Vertex [Gulhane et al., ICDE'11]



Extracted relationships

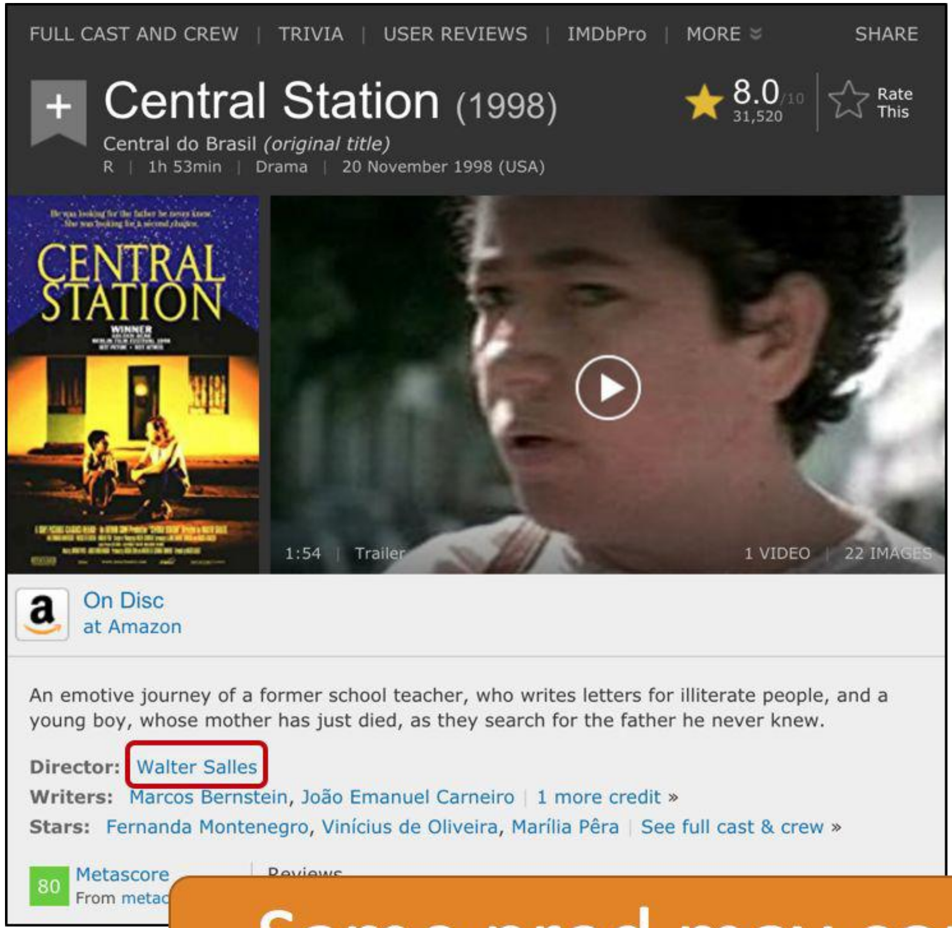
- (Top Gun, type.object.name, "Top Gun")
- (Top Gun, film.film.genre, Action)
- (Top Gun, film.film.directed_by, Tony Scott)
- (Top Gun, film.film.starring, Tom Cruise)
- (Top Gun, film.film.runtime, "1h 50min")
- (Top Gun, film.film.release_Date_s, "16 May 1986")

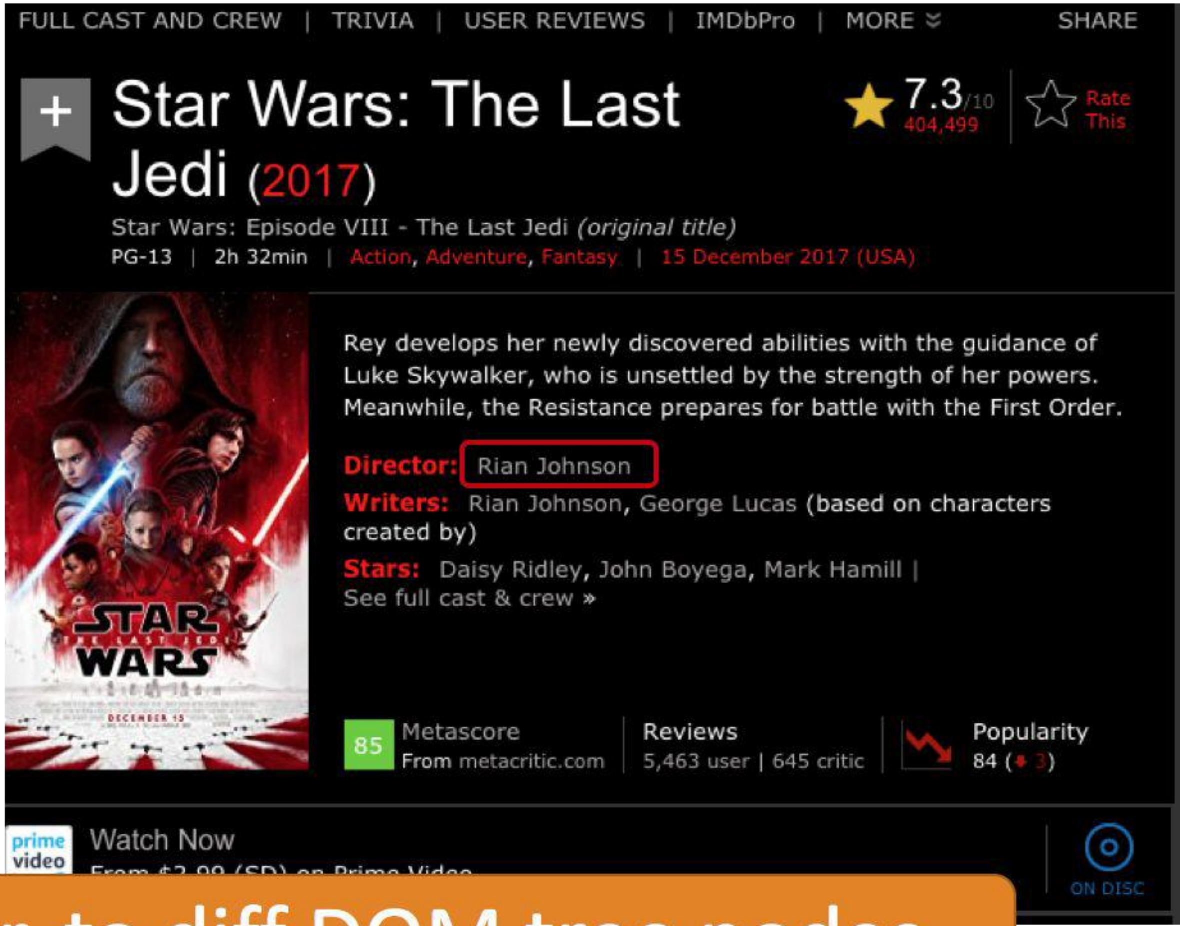
Wrapper Induction--Vertex [Gulhane et al., ICDE'11]

- Solution: find XPaths from DOM Trees

Wrapper Induction--Vertex [Gulhane et al., ICDE'11]

- Challenge: slight variations from page to page

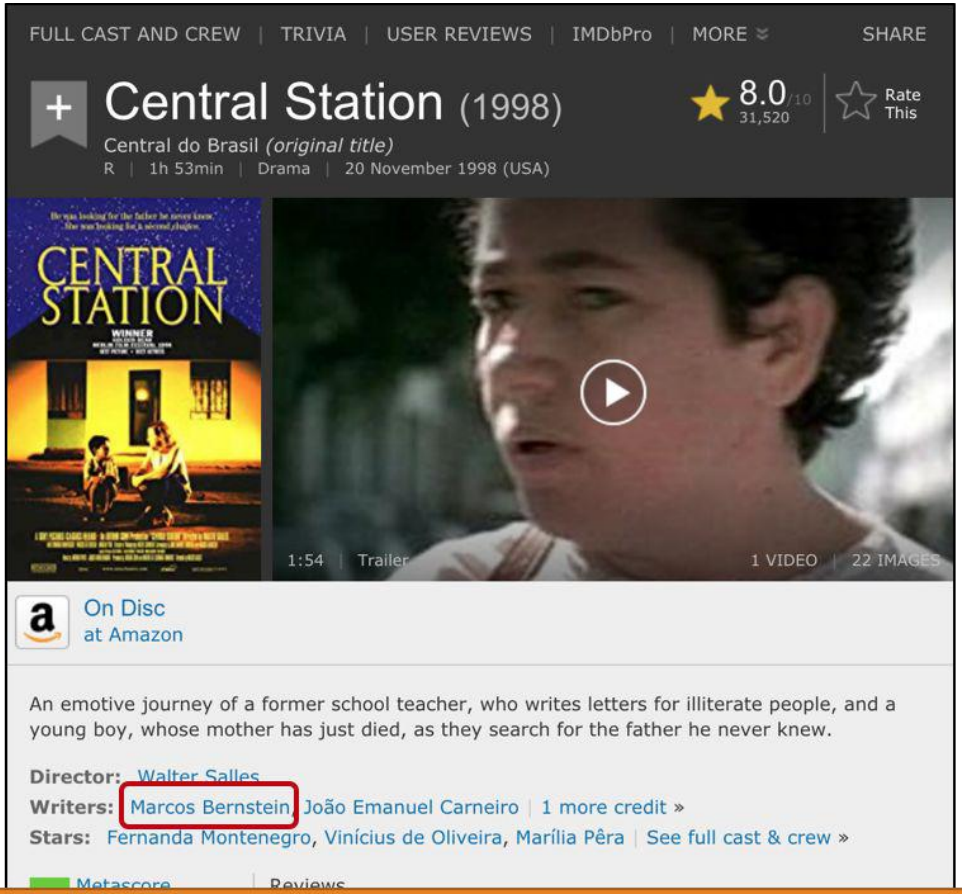


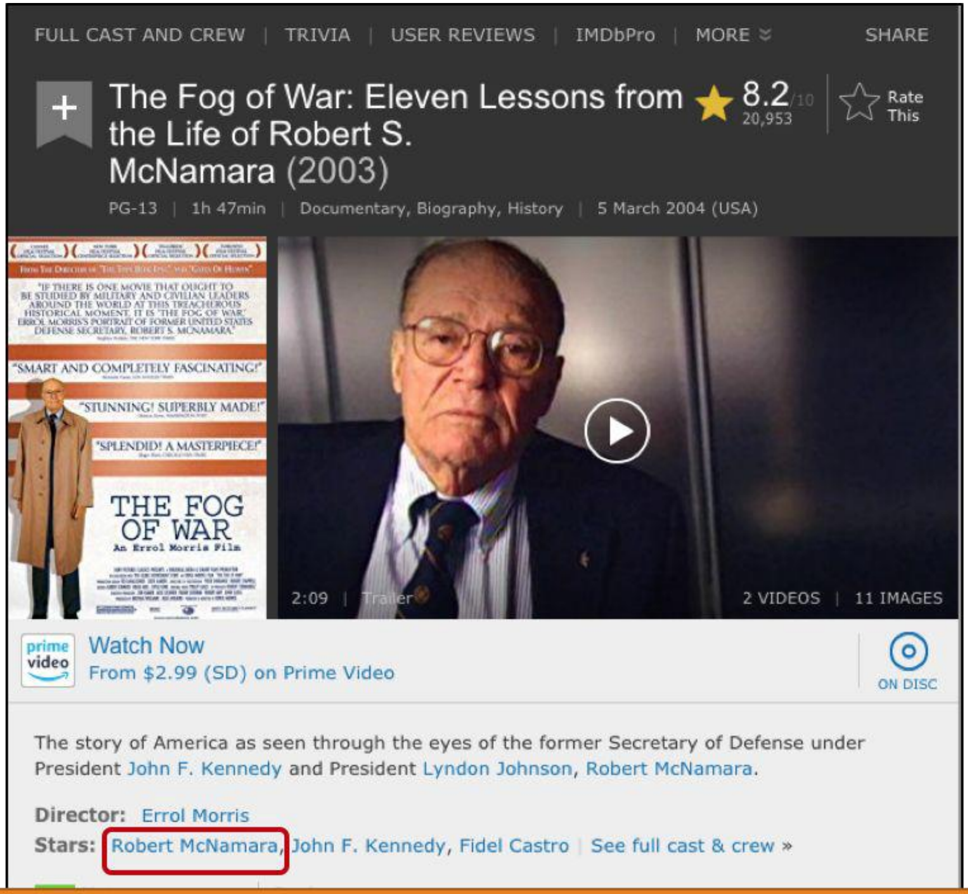


Same pred may corr. to diff DOM tree nodes

Wrapper Induction--Vertex [Gulhane et al., ICDE'11]

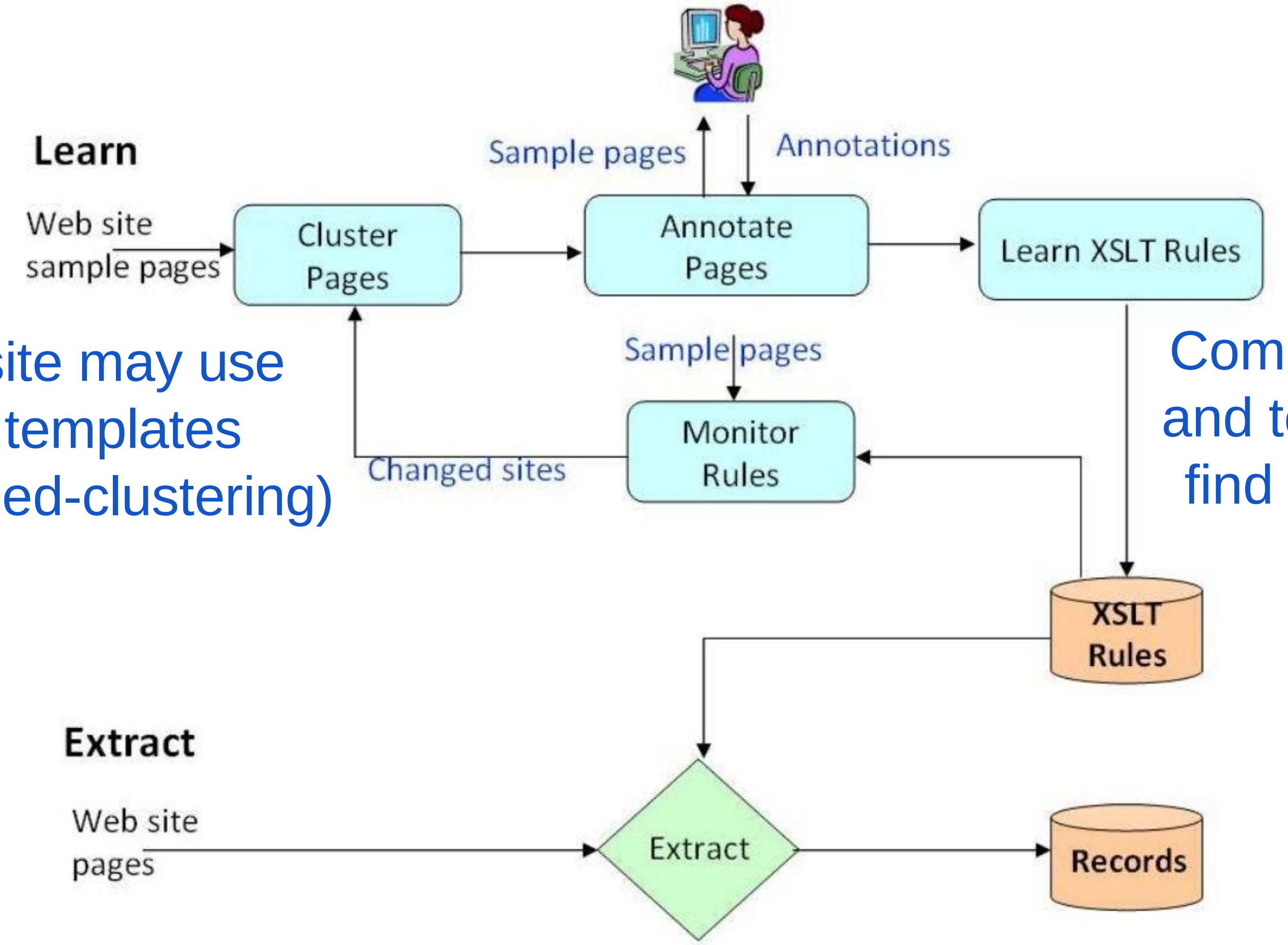
- Challenge: slight variations from page to page





Same DOM tree node may correspond to diff preds

Wrapper Induction--Vertex [Gulhane et al., ICDE'11]



One website may use multiple templates (Unsupervised-clustering)

Combine attr features and textual features to find a general XPath (LR)

Wrapper Induction--Vertex [Gulhane et al., ICDE’11]

- **Sample learned XPaths on IMDb**
  - //*[@itemprop="name"]
  - //*[@class="bp_item bp_text_only"]/*/*/*[@class="bp_heading"]
  - //*[following-sibling::*[position()=3][@class="subheading"]]/*[followin g-sibling::*[position()=1][@class="attribute"]]
  - //*[preceding-sibling::node()[normalize-space(.)!=""][text()="Languag e:"]

Ensure high recall

Ensure high precision

Distantly supervised extraction

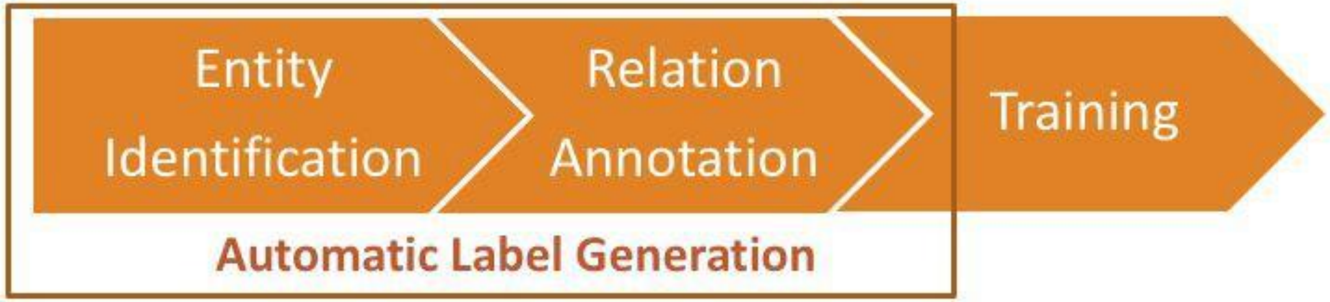
2013 (Deep ML)

Deep learning

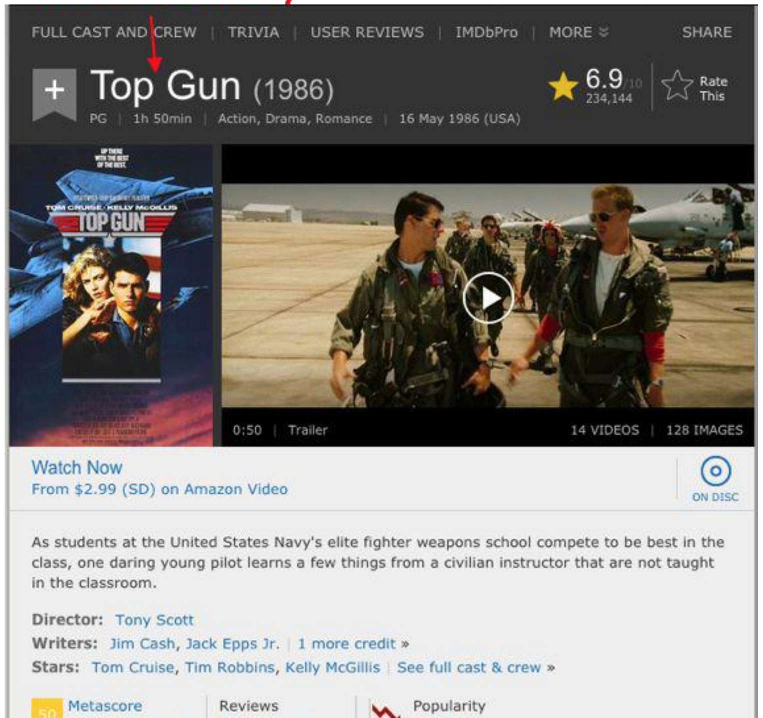
- Use RNN, CNN, attention for RE
- Data programming / Heterogeneous learning
- Revisit DOM extraction

- **Annotation-based extraction**
  - Pros: high precision and recall
  - Cons: does not scale--annotation per cluster per website
- **Distantly-supervised extraction**
  - Step 1. Use seed data to automatically annotate
  - Step 2. Use the (noisy) annotations for training
  - E.g., DeepDive, Knowledge Vault

Distantly supervised extraction--Ceres [Lockard et al., VLDB'18]

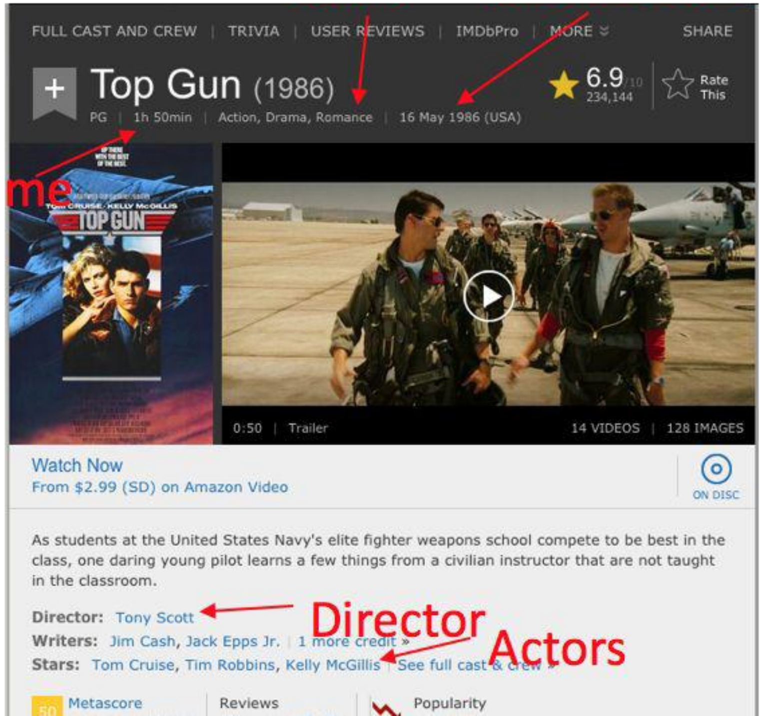


Movie entity



Genre Release Date

Runtime



Extracted triples

- (Top Gun, type.object.name, "Top Gun")
- (Top Gun, film.film.genre, Action)
- (Top Gun, film.film.directed_by, Tony Scott)
- (Top Gun, film.film.starring, Tom Cruise)
- (Top Gun, film.film.runtime, "1h 50min")
- (Top Gun, film.film.release_Date_s, "16 May 1986")

Distantly supervised extraction--Ceres [Lockard et al., VLDB'18]

- Annotation-based extraction
- Distantly-supervised extraction

2013 (Deep ML)

Deep learning

- Use RNN, CNN, attention for RE
- Data programming / Heterogeneous learning
- Revisit DOM extraction

	Vertex (Gulhane et al, 2011)				Ceres			
	Prec	Rec	F1	#Pred	Prec	Rec	F1	#Pred
Movie	0.97	0.97	0.97	4	0.97	0.99	0.98	4
NBAPlayer	1.00	1.00	1.00	4	0.98	0.98	0.98	4
University	0.99	0.98	0.99	4	0.87	0.94	0.90	4
Book	0.93	0.93	0.93	5	0.94	0.63	0.70	

Very high precision

Competent w. rule-based wrapper induction

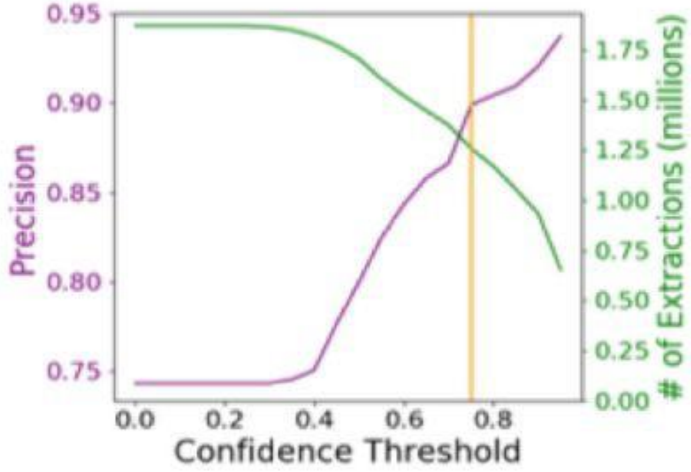
Distantly supervised extraction--Ceres [Lockard et al., VLDB'18]

- Extraction on long-tail movie websites

#Websites / #Webpages	33 / 434K
Language	English and 6 other languages
Domains	Animated films, Documentary films, Financial performance, etc.
# Annotated pages	70K (16%)
Annotated : Extracted #entities	1 : **2.6**
Annotated : Extracted #triples	1 : **3.0**
# Extractions	1.25 M
Precision	90%

Distantly supervised extraction--Ceres [Lockard et al., VLDB'18]

- Extraction on long-tail movie websites



Distantly supervised extraction

2013 (Deep ML)

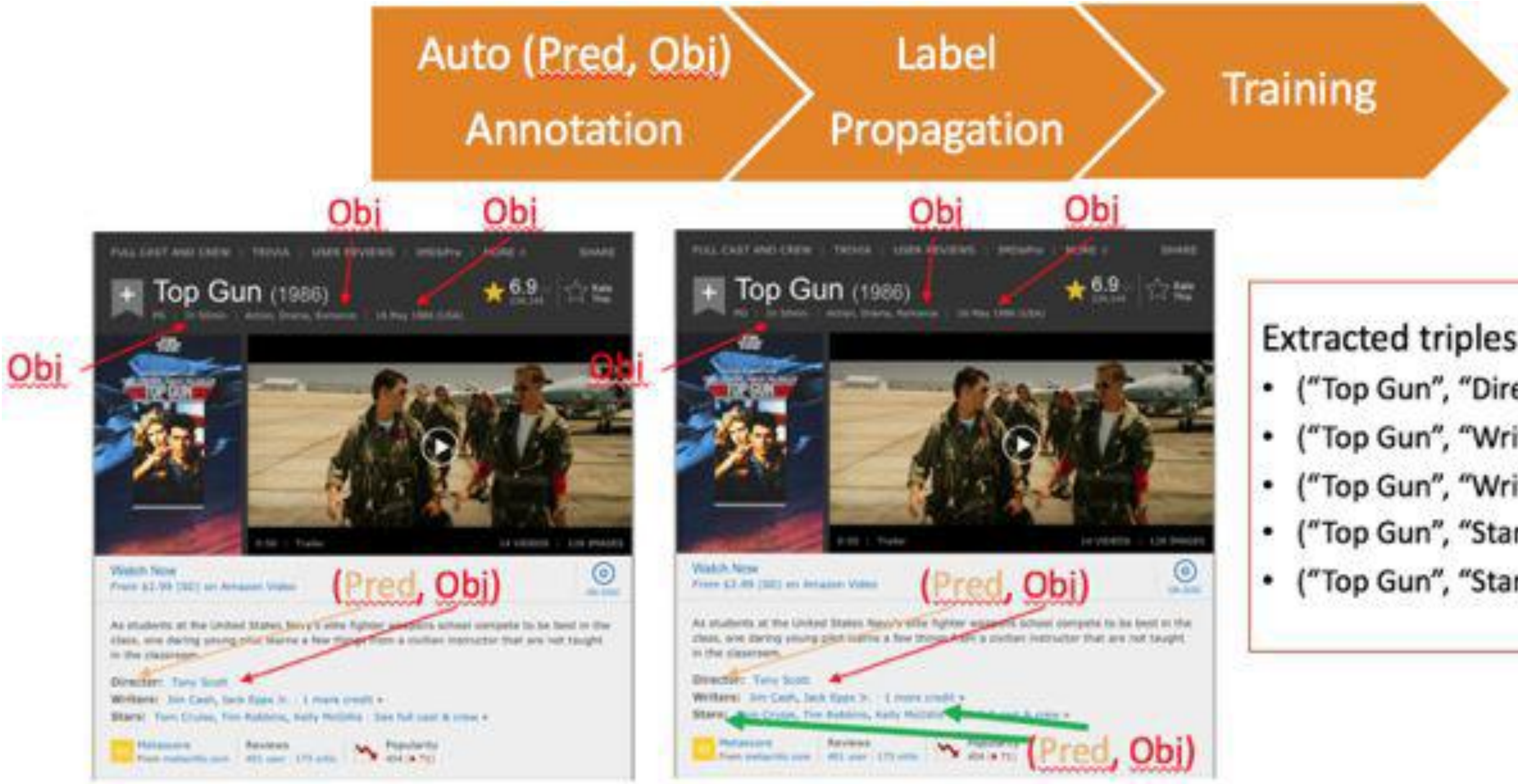
Deep learning

- Use RNN, CNN, attention for RE
- Data programming / Heterogeneous learning
- Revisit DOM extraction

- **Annotation-based extraction**
  - Pros: high precision and recall
  - Cons: does not scale--annotation per cluster per website
- **Distantly-supervised extraction**
  - Step 1. Use seed data to automatically annotate
  - Step 2. Use the (noisy) annotations for training
  - E.g., DeepDive, Knowledge Vault
- **OpenIE extraction**

OpenIE on semi-structured data--OpenCeres

[Lockard et al., NAACL’19]



Extracted triples

- (“Top Gun”, “Director”, “Tony Scott”)
- (“Top Gun”, “Writers”, “Jim Cash”)
- (“Top Gun”, “Writers”, “Jack Epps Jr.”)
- (“Top Gun”, “Stars”, “Tom Cruise”)
- (“Top Gun”, “Stars”, “Tim Robbins”)

OpenIE on semi-structured data--OpenCeres

[Lockard et al., NAACL'19]

- Annotation-based extraction
- Distantly-supervised extraction
- OpenIE extraction

	Vertex (Gulhane et al, 2011)				Ceres				OpenCeres			
	Prec	Rec	F1	#Pred	Prec	Rec	F1	#Pred	Prec	Rec	F1	#Pred
Movie	0.97	0.97	0.97	4	0.97	0.99	0.98	4	0.77	0.68	0.72	18
NBAPlayer	1.00	1.00	1.00	4	0.98	0.98	0.98	4	0.74	0.48	0.58	17
University	0.99	0.98	0.99	4	0.87	0.94	0.90	4	0.65	0.29	0.40	92
Book	0.93	0.93	0.93	5	0.94	0.63	0.70	5	-	-	-	-

Precision much lower

Much more predicates

OpenIE on semi-structured data--OpenCeres

[Lockard et al., NAACL'19]

Movie

- Seed: Director, Writer, Producer, Actor, Release Date, Genre, Alternate Title
- New: Country, Filmed In, Language, MPAA Rating, Set In, Reviewed by, Studio, Metascore, Box Office, Distributor, Tagline, Budget, Sound Mix

NBA Player

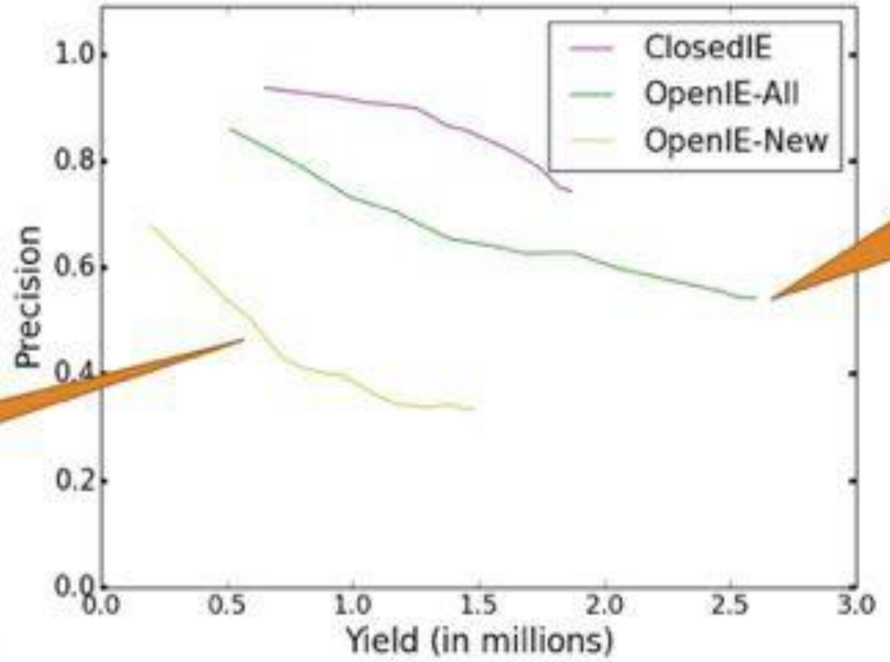
- Seed: Height, Weight, Team
- New: Birth Date, Birth Place, Salary, Age, Experience, Position, College, Year Drafted

University

- Seed: Phone Number, Web address, Type (public/private)
- New: Calendar System, Enrollment, Highest Degree, Local Area, Student Services, President

OpenIE on semi-structured data--OpenCeres

[Lockard et al., NAACL'19]



OpenIE added significant amount of knowledge

Still need prec improvement on new relations

Extraction from semi-structured websites

2013 (Deep ML)

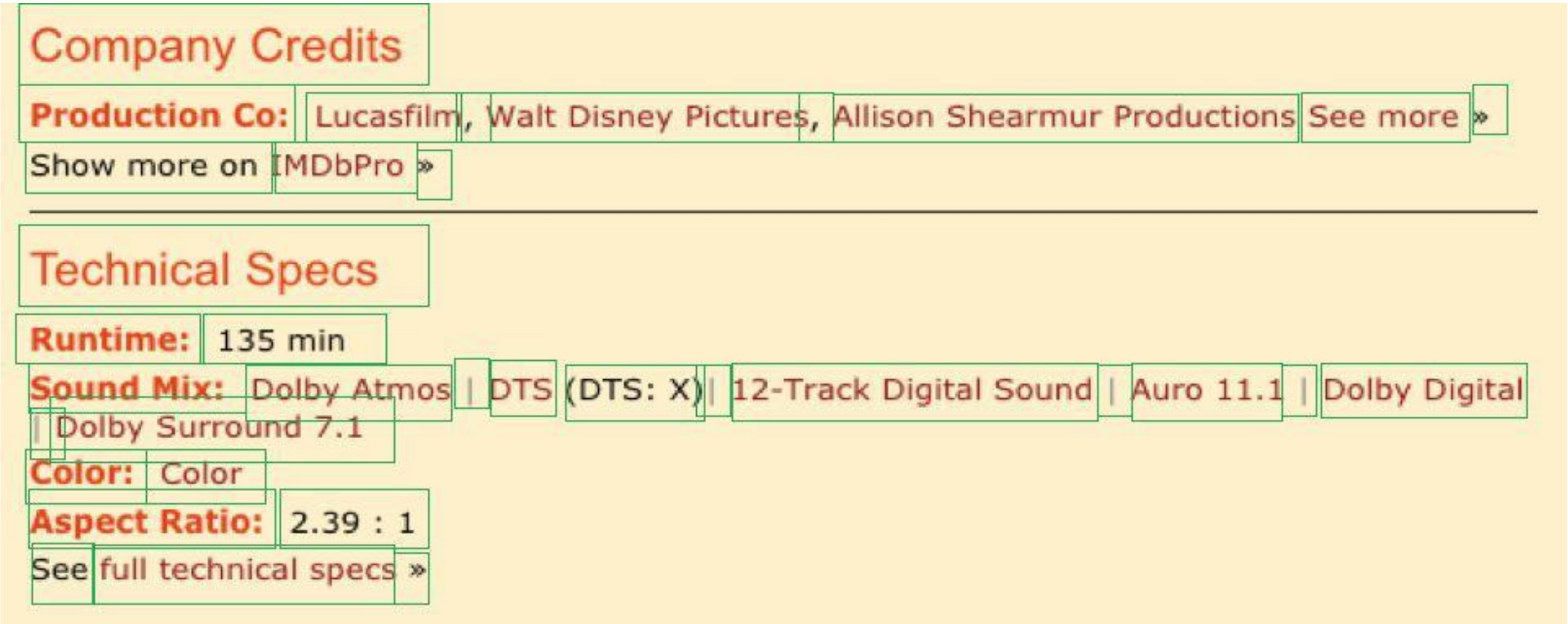
Deep learning

- Use RNN, CNN, attention for RE
- Data programming / Heterogeneous learning
- Revisit DOM extraction

- **Which model is the best?**
  - Logistic regression: best results (20K features on one website)
  - Random forest: lower precision and recall
  - Deep learning??

Challenges in applying deep learning on extracting semi-structured data

- Web layout is neither 1D sequence nor regular 2D grid, so CNN or RNN does not directly apply



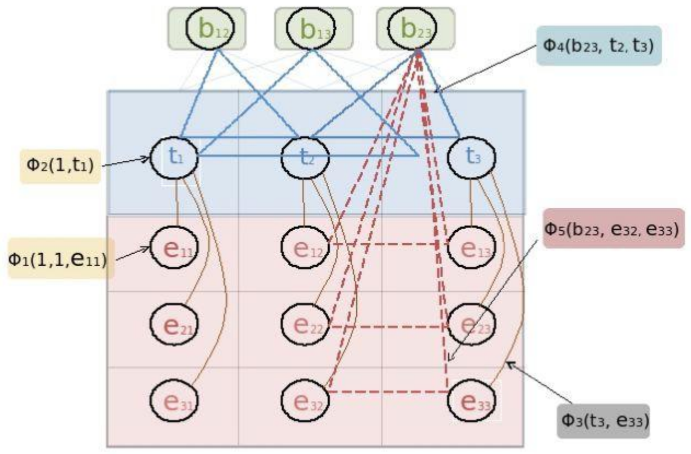
WebTable Extraction [Limaye et al., VLDB'10]

- Model table annotation using interrelated random variables, represented by a probabilistic graphical model
  - Cell text (in Web table) and entity label (in catalog)
  - Column header (in Web table) and type label (in catalog)
  - Column type and cell entity (in Web table)

Extraction from semi-structured data

- WebTables: search, extraction
- DOM tree: wrapper induction

2008 (Semi-stru)



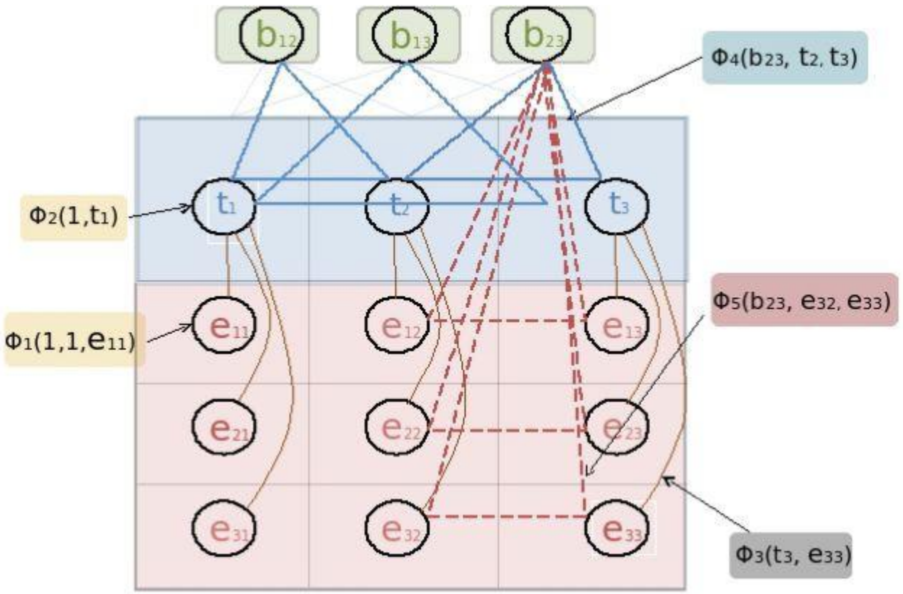
WebTable Extraction [Limaye et al., VLDB'10]

- Model table annotation using interrelated random variables, represented by a probabilistic graphical model
  - Pair of column types (in Web table) and relation (in catalog)
  - Entity pairs (in Web table) and relation (in catalog)

Extraction from semi-structured data

- WebTables: search, extraction
- DOM tree: wrapper induction

2008 (Semi-stru)



Challenges in applying ML on DX

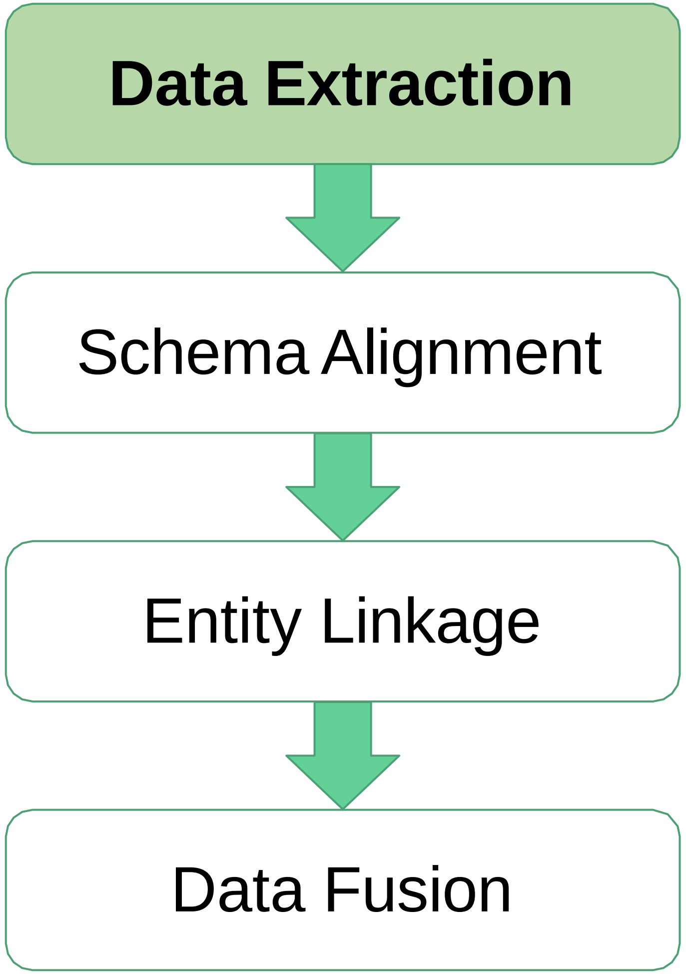
- Automatic data extraction cannot reach production quality requirement. How to improve precision?

- Every web designer has her own whim, but there are underlying patterns across websites. How to learn extraction patterns on different websites, especially for semi-structured sources?

- ClosedIE throws away too much data. How to apply OpenIE on all kinds of data?

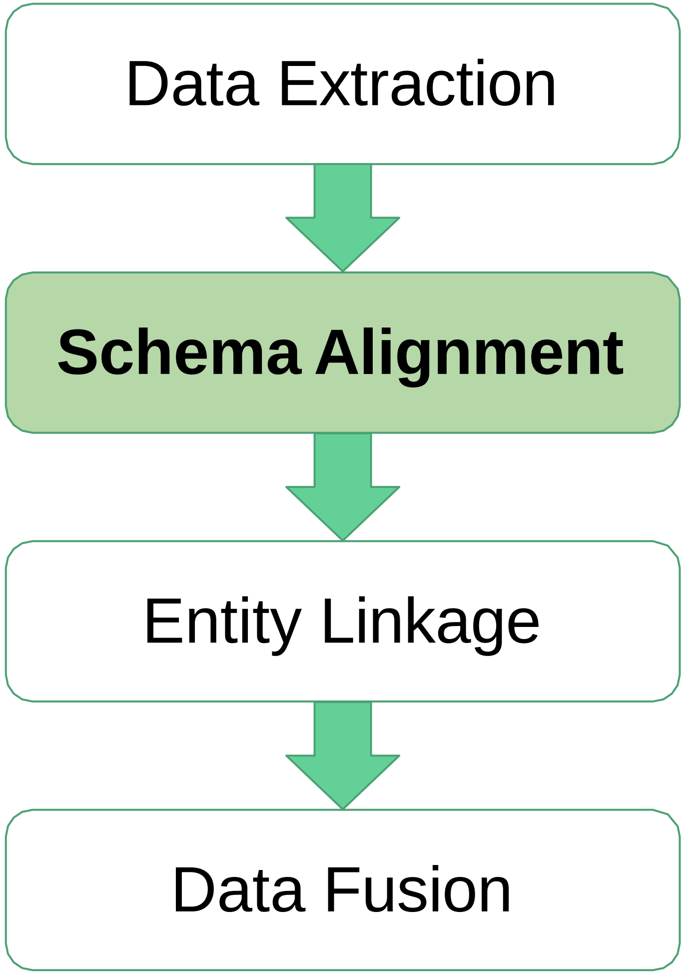
Recipe for data extraction

- Problem definition: Extract structure from semi- or un-structured data
- Short answers
  - Wrapper induction has high prec/rec
  - Distant supervision is critical for collecting training data
  - DL effective for texts and LR is often effective for semi-stru data



Today's agenda

- Part I. Introduction
- Part II. ML for DI
  - ML for entity linkage
  - ML for data extraction
  - ML for schema alignment
  - ML for data fusion

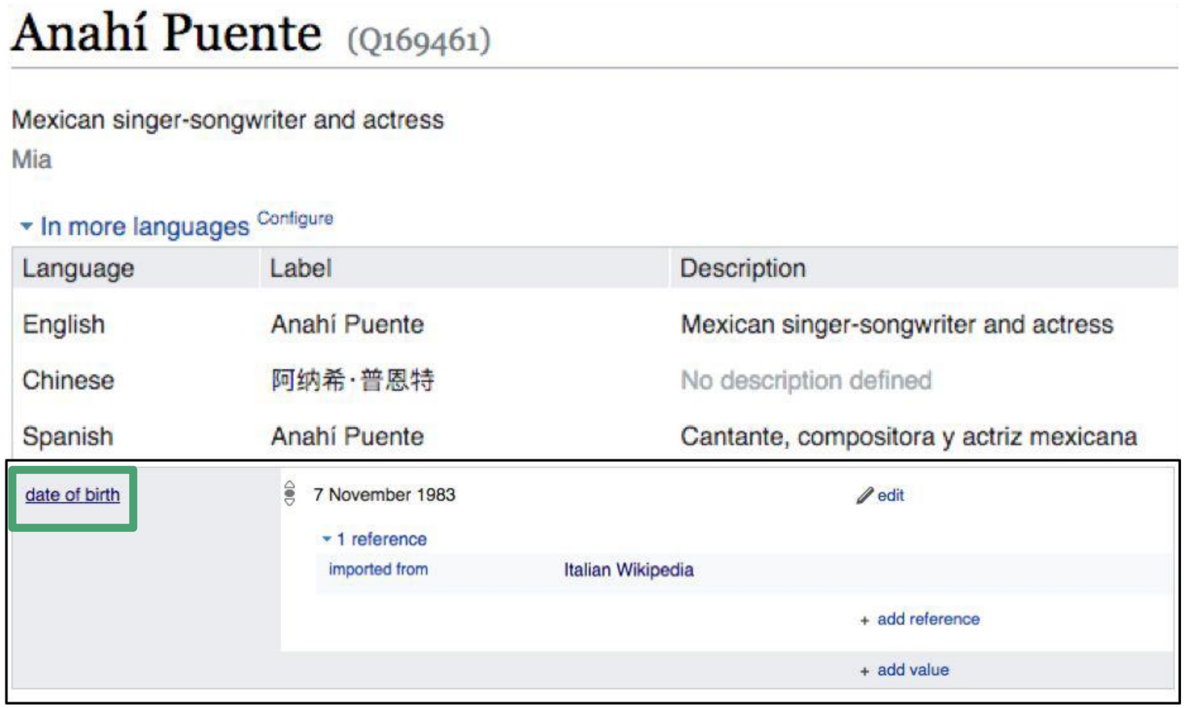


What is schema alignment?

- Definition: Align schemas and understand which attributes have the same semantics.

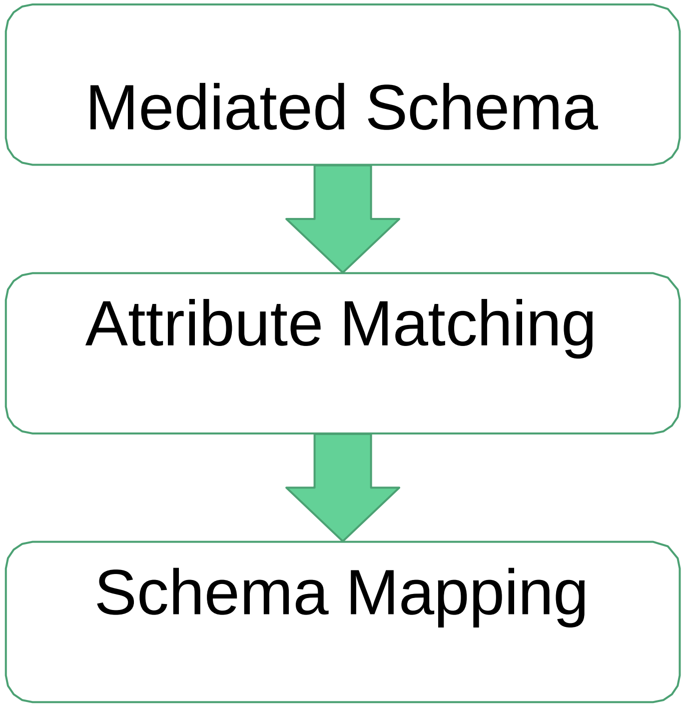
IMDB

WikiData



Quick tour for schema alignment

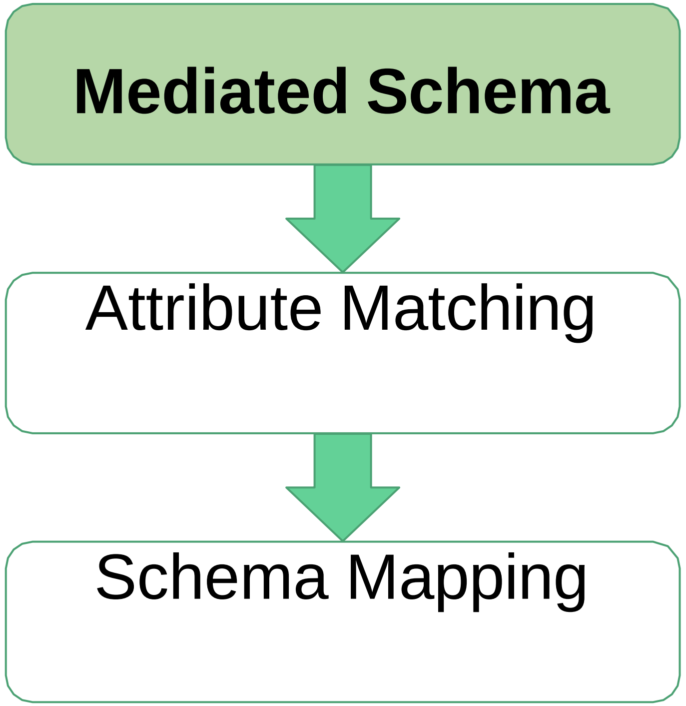
S1	(name, hPhone, hAddr, oPhone, oAddr)
S2	(name, phone, addr, email)
S3	a: (id, name); b: (id, resPh, workPh)
S4	(name, pPh, pAddr)
S5	(name, wPh, wAddr)



Quick tour for schema alignment

- **Mediated schema:** a unified and virtual view of the salient aspects of the domain

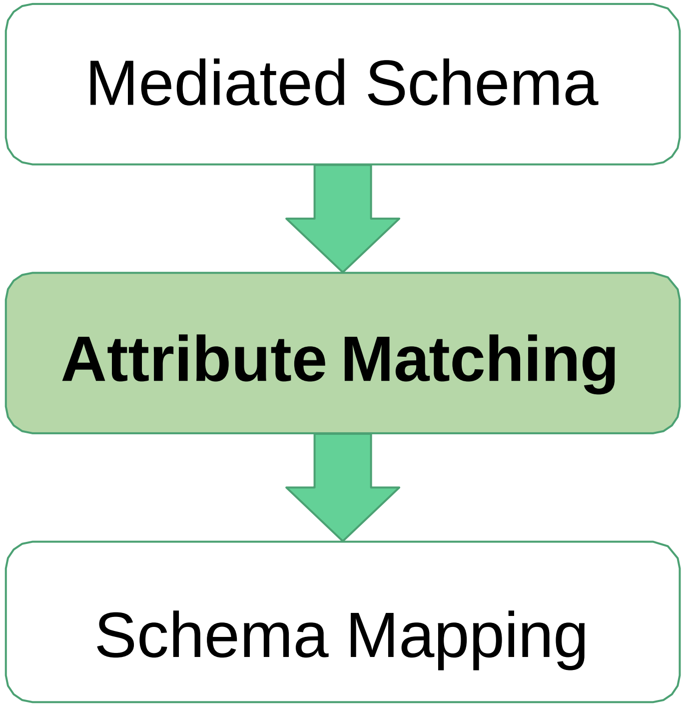
S1	(name, hPhone, hAddr, oPhone, oAddr)
S2	(name, phone, addr, email)
S3	a: (id, name); b: (id, resPh, workPh)
S4	(name, pPh, pAddr)
S5	(name, wPh, wAddr)
MS	(n, pP, pA, wP, wA)



Quick tour for schema alignment

- **Attribute matching:** correspondences between schema attributes

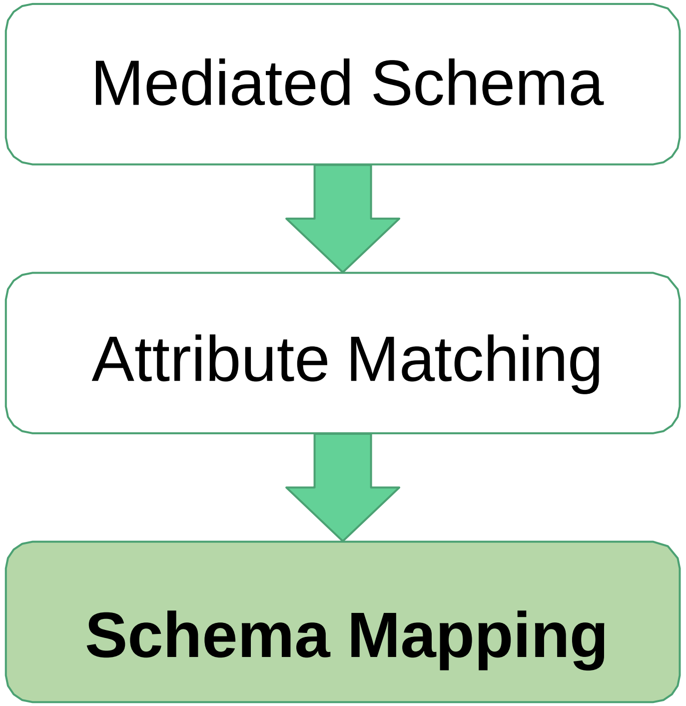
S1	(name, hPhone, hAddr, oPhone, oAddr)
S2	(name, phone, addr, email)
S3	a: (id, name); b: (id, resPh, workPh)
S4	(name, pPh, pAddr)
S5	(name, wPh, wAddr)
MS	**(n, pP, pA, wP, wA)**
MSAM	**MS.n**: S1.name, S2.name, S3a.name, … **MS.pP**: S1.hPhone, S3b.resPh, S4.pPh **MS.pA**: S1.hAddr, S4.pAddr **MS.wP**: S1.oPhone, S2.phone, … **MS.wA**: S1.oAddr, S2.addr, S5.wAddr



Quick tour for schema alignment

- **Schema mapping**: transformation between records in different schemas

S1	(name, hPhone, hAddr, oPhone, oAddr)
S2	(name, phone, addr, email)
S3	a: (id, name); b: (id, resPh, workPh)
S4	(name, pPh, pAddr)
S5	(name, wPh, wAddr)
MS	**(n, pP, pA, wP, wA)**
MSSM (GAV)	MS(n, pP, pA, wP, wA) :- S1(n, pP, pA, wP, wA) MS(n, _, _, wP, wA) :- S2(n, wP, wA, e) MS(n, pP, _, wP, _) :- S3a(i, n), S3b(i, pP, wP) MS(n, pP, pA, _, _) :- S4(n, pP, pA) MS(n, _, _, wP, wA) :- S5(n, wP, wA)



30 years of schema alignment

1990 (Desc Logics)

Description Logics
- Gav vs. Lav. vs. Glav
- Answering queries using views
- Warehouse vs. EII

1994 (Early ML)

Semi-Auto mapping
- Learning to match
- Schema mapping: Clio
- Data exchange

2005 (Dataspaces)

Pay-as-you-go dataspaces
- Probabilistic schema alignment

2013 (Deep ML)

Logic & Deep learning
- Collective disc. by PSL
- Universal schema

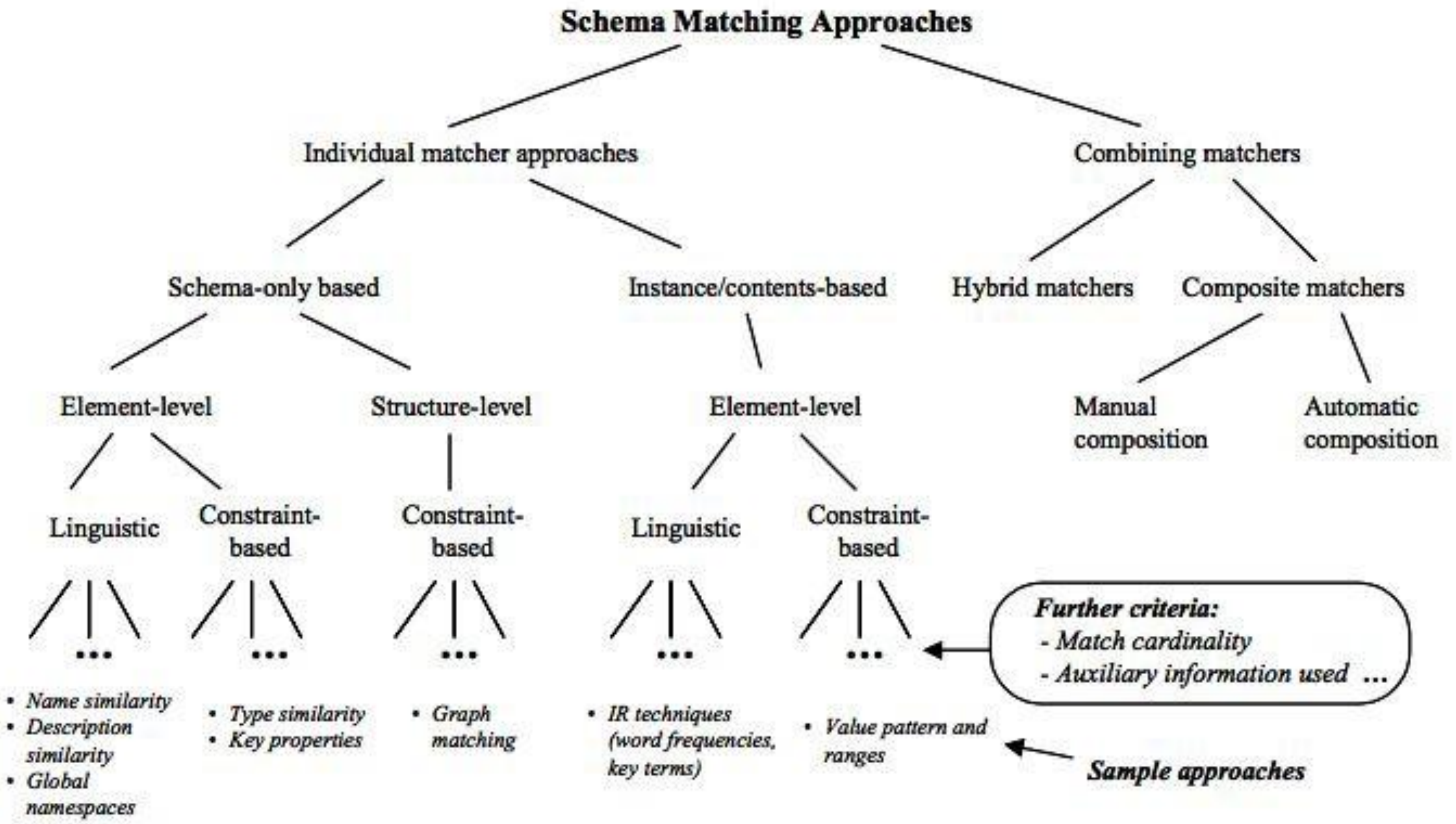
Early ML models

[Rahm and Bernstein, VLDBJ'2001]

~2000 (Early ML)

Semi-Auto mapping

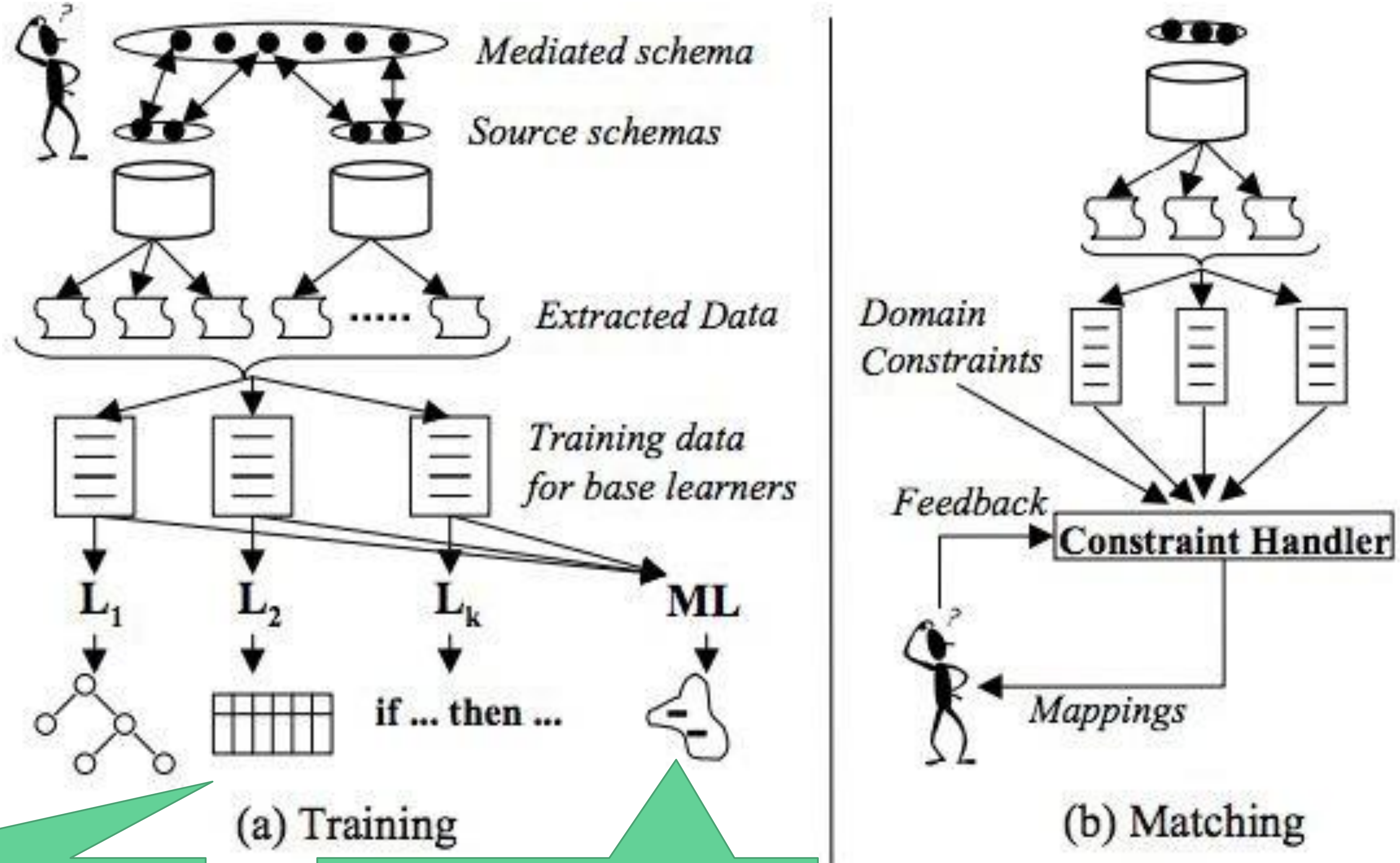
- Learning to match
- Schema mapping: Clio
- Data exchange



Signals: name, description, type, key, graph structure, values

Early ML models

[Doan et al., Sigmod'01]



~2000 (Early ML)

Semi-Auto mapping

- Learning to match
- Schema mapping: Clio
- Data exchange

Base learners: kNN, naive Bayes, etc.

Meta learner--Stacking

Early ML models

[Doan et al., Sigmod'01]

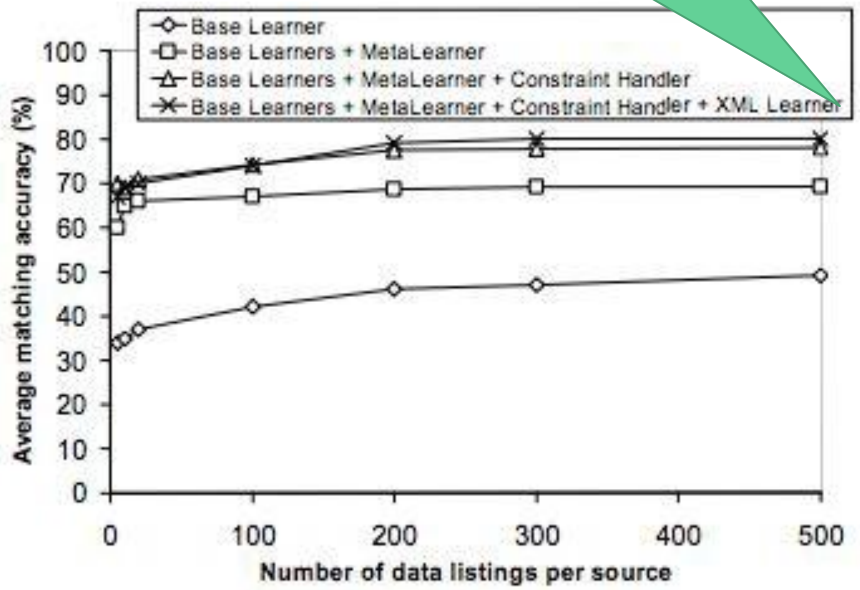
Avg Accuracy: 71-92%

Meta learning and constraints help

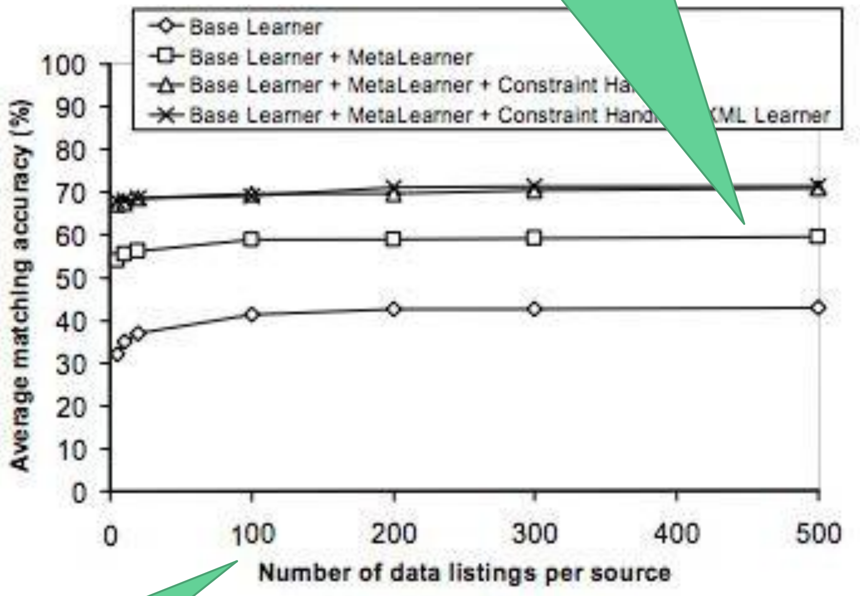
~2000 (Early ML)

Semi-Auto mapping

- Learning to match
- Schema mapping: Clio
- Data exchange



(b) Matching accuracy for Real Estate I



(c) Matching accuracy for Time Schedule

More data instances help

Collective mapping discovery by PSL

[Kimmig et al, ICDE'17]

2013 (Deep ML)

Logic & Deep learning

- Collective disc. by PSL
- Universal schema

Step 1. Generate candidate mappings

E.g.,

$$\theta_0 : \mathsf{proj}(\mathbf{t}, \mathrm{m}, \mathrm{l}) \wedge \mathsf{emp}(\mathrm{m}, \mathbf{n}, \mathrm{c}) \rightarrow \exists\, \mathrm{o}.\, \mathsf{task}(\mathbf{t}, \mathbf{n}, \mathrm{o})$$
$$\theta_1 : \mathsf{proj}(\mathbf{t}, \mathrm{m}, \mathrm{l}) \wedge \mathsf{emp}(\mathrm{l}, \mathbf{n}, \mathrm{c}) \rightarrow \exists\, \mathrm{o}.\, \mathsf{task}(\mathbf{t}, \mathbf{n}, \mathrm{o})$$
$$\theta_2 : \mathsf{proj}(\mathbf{t}, \mathrm{m}, \mathrm{l}) \wedge \mathsf{emp}(\mathrm{m}, \mathbf{n}, \mathbf{c}) \rightarrow \exists\, \mathrm{o}.\, \mathsf{task}(\mathbf{t}, \mathbf{n}, \mathrm{o}) \wedge \mathsf{org}(\mathrm{o}, \mathbf{c})$$
$$\theta_3 : \mathsf{proj}(\mathbf{t}, \mathrm{m}, \mathrm{l}) \wedge \mathsf{emp}(\mathrm{l}, \mathbf{n}, \mathbf{c}) \rightarrow \exists\, \mathrm{o}.\, \mathsf{task}(\mathbf{t}, \mathbf{n}, \mathrm{o}) \wedge \mathsf{org}(\mathrm{o}, \mathbf{c})$$

Step 2. Solve PSL

1.Prefer fewer mappings: penalty=#atoms

$$\begin{aligned} \mathsf{size}_m(F) :&\quad in(F) \rightarrow \perp \\ 1 :&\quad J(T) \rightarrow \exists F.\ \mathsf{covers}(F, T) \wedge in(F) \\ 1 :&\quad in(F) \wedge \mathsf{creates}(F, T) \rightarrow J(T) \end{aligned}$$

3. Tuples inferred from the mapping should exist

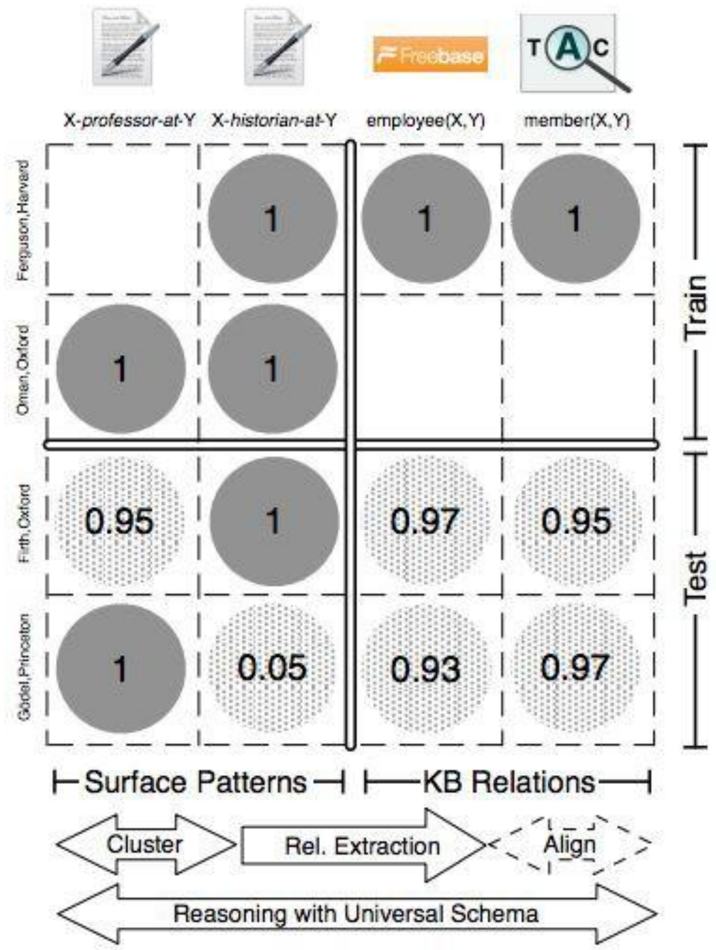
2. An existing tuple can be inferred from the mappings

Universal Schema [Riedel et al., NAACL'13][Yao et al., AKBC'13]

- Attribute matching → Instance inference

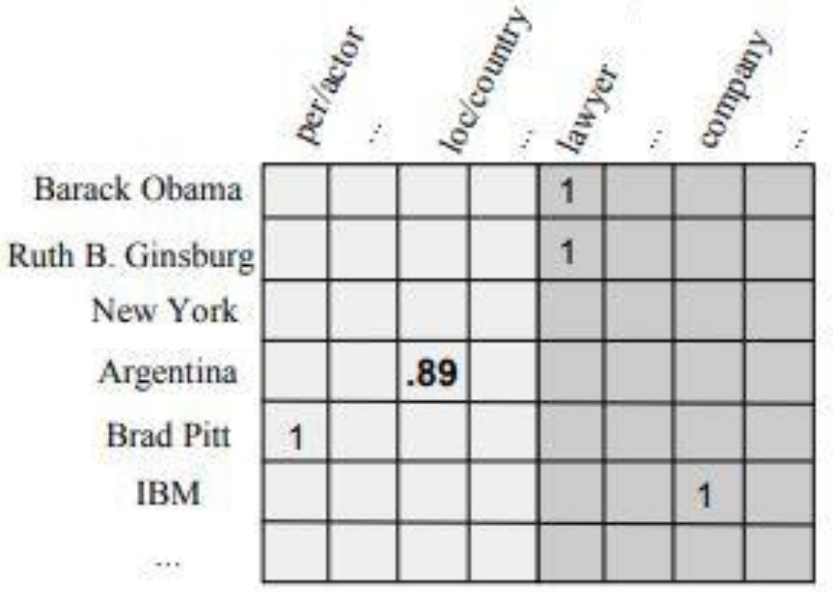
2013 (Deep ML)

Logic & Deep learning
- Collective disc. by PSL
- Universal schema



Relation prediction

Matrix factorization



Type prediction

Universal Schema [Riedel et al., NAACL'13]

2013 (Deep ML)

Logic & Deep learning
- Collective disc. by PSL
- Universal schema

- Attribute matching → Instance inference
- $f(e_s, r, e_o)$ is computed using embeddings; the higher, the more likely to be true
- DistMult is a relation embedding model

Limitation: Cannot apply to new entities or relations

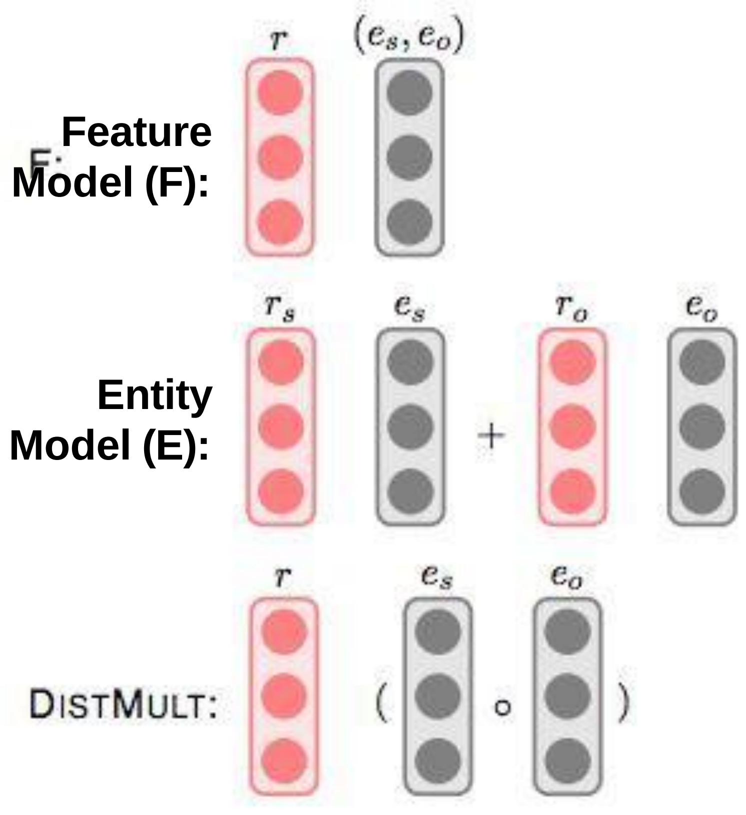


Figure 3: The continuous representations for model F, E and DistMult. [Toutanova et al., EMNLP'15]

Columnless univ. schema w. CNN

[Toutanova et al., EMNLP’15]

2013 (Deep ML)

Logic & Deep learning

- Collective disc. by PSL
- Universal schema

- Relation: organizationFoundedBy

Textual Pattern	Count
SUBJECT $\xrightarrow{\text{appos}}$ founder $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	12
SUBJECT $\xleftarrow{\text{nsubj}}$ co-founded $\xrightarrow{\text{dobj}}$ OBJECT	3
SUBJECT $\xrightarrow{\text{appos}}$ co-founder $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	[illegible]
SUBJECT $\xrightarrow{\text{conj}}$ co-founder $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	[illegible]
SUBJECT $\xleftarrow{\text{pobj}}$ with $\xleftarrow{\text{prep}}$ co-founded $\xrightarrow{\text{dobj}}$ OBJECT	[illegible]
SUBJECT $\xleftarrow{\text{nsubj}}$ signed $\xrightarrow{\text{xcomp}}$ establishing $\xrightarrow{\text{dobj}}$ OBJECT	2
SUBJECT $\xleftarrow{\text{pobj}}$ with $\xleftarrow{\text{prep}}$ founders $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	2
SUBJECT $\xrightarrow{\text{appos}}$ founders $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	2
SUBJECT $\xleftarrow{\text{nsubj}}$ one $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ founders $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	2
SUBJECT $\xleftarrow{\text{nsubj}}$ founded $\xrightarrow{\text{dobj}}$ production $\xrightarrow{\text{conj}}$ OBJECT	2
SUBJECT $\xleftarrow{\text{appos}}$ partner $\xleftarrow{\text{pobj}}$ with $\xleftarrow{\text{prep}}$ founded $\xrightarrow{\text{dobj}}$ production $\xrightarrow{\text{conj}}$ OBJECT	2
SUBJECT $\xleftarrow{\text{pobj}}$ by $\xleftarrow{\text{prep}}$ co-founded $\xleftarrow{\text{rcmod}}$ OBJECT	1
SUBJECT $\xleftarrow{\text{nn}}$ co-founder $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	1
SUBJECT $\xrightarrow{\text{dep}}$ co-founder $\xrightarrow{\text{prep}}$ of $\xrightarrow{\text{pobj}}$ OBJECT	1
SUBJECT $\xleftarrow{\text{nsubj}}$ helped $\xrightarrow{\text{xcomp}}$ establish $\xrightarrow{\text{dobj}}$ OBJECT	1
SUBJECT $\xleftarrow{\text{nsubj}}$ signed $\xrightarrow{\text{xcomp}}$ creating $\xrightarrow{\text{dobj}}$ OBJECT	1

Similarity of phrases → CNN

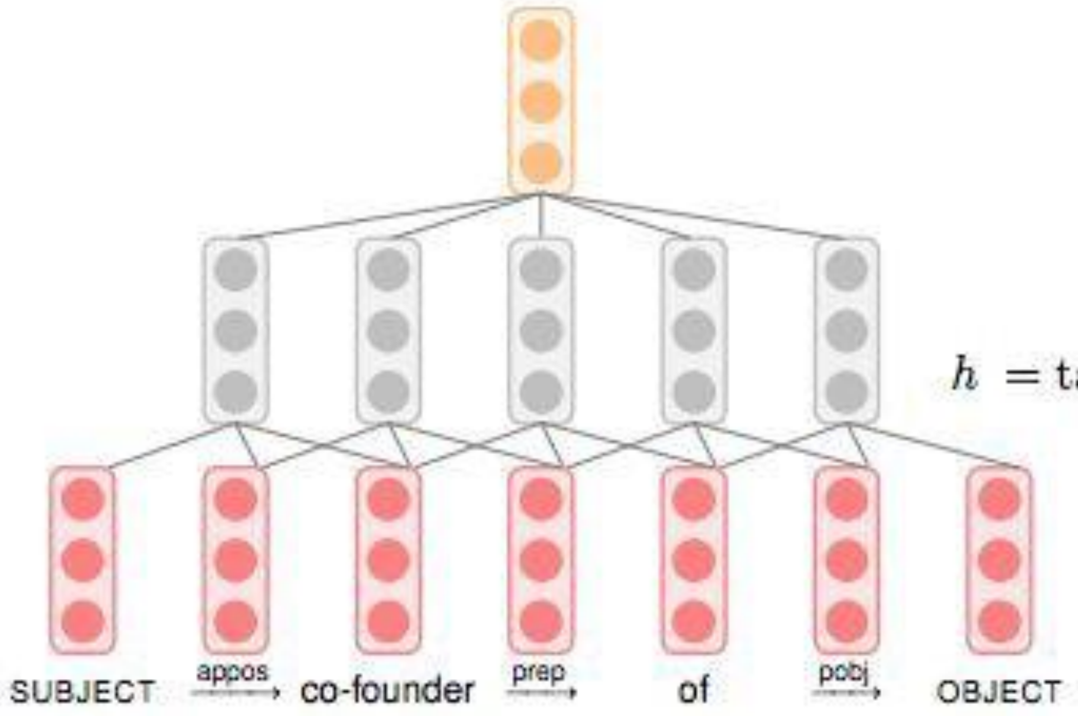
Columnless univ. schema w. CNN

[Toutanova et al., EMNLP’15]

2013 (Deep ML)

Logic & Deep learning

- Collective disc. by PSL
- Universal schema



$$r = \max\{h\}$$

$$h = \tanh(W^{-1}v_{-1} + W^{0}v + W^{1}v_{+1} + b)$$

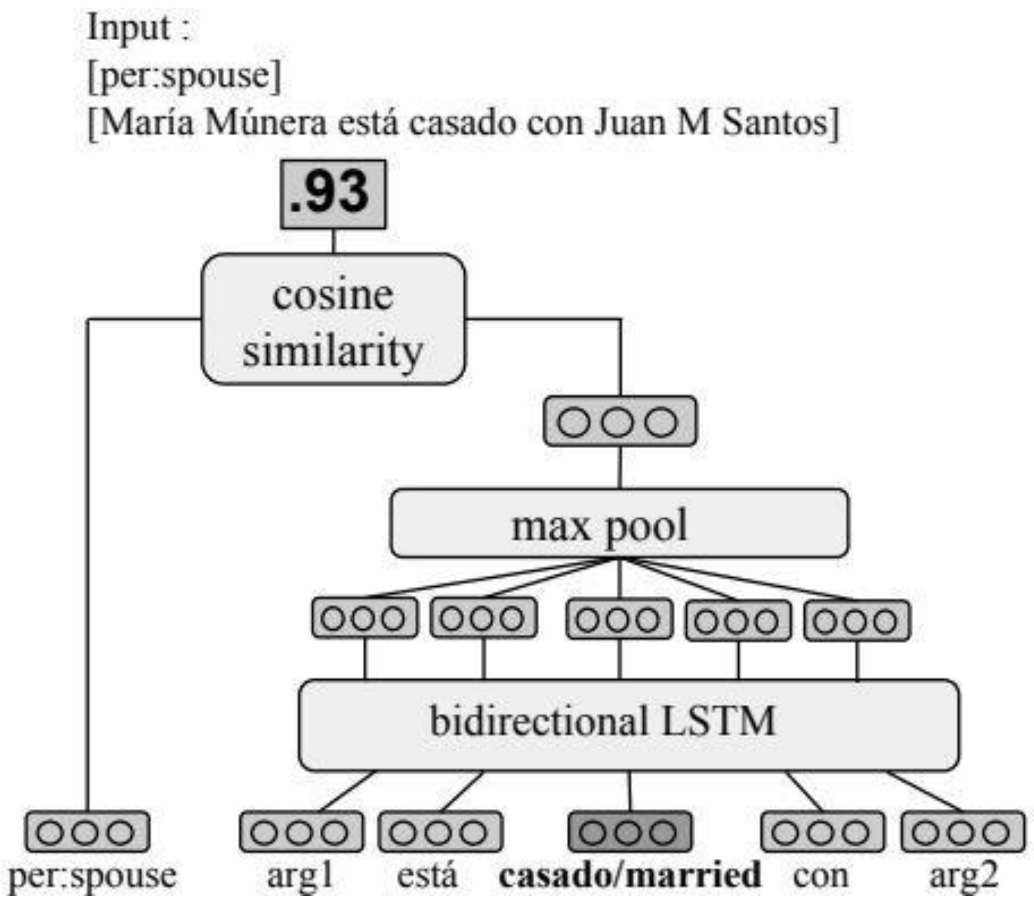
$$v = Ve$$

Figure 4: The convolutional neural network architecture for representing textual relations.

Columnless univ. schema w. RNN

[Verga et al., ACL'16]

- Similar sequences of context tokens should be embedded similarly



2013 (Deep ML)

Logic & Deep learning
- Collective disc. by PSL
- Universal schema

Rowless Univ. Schema

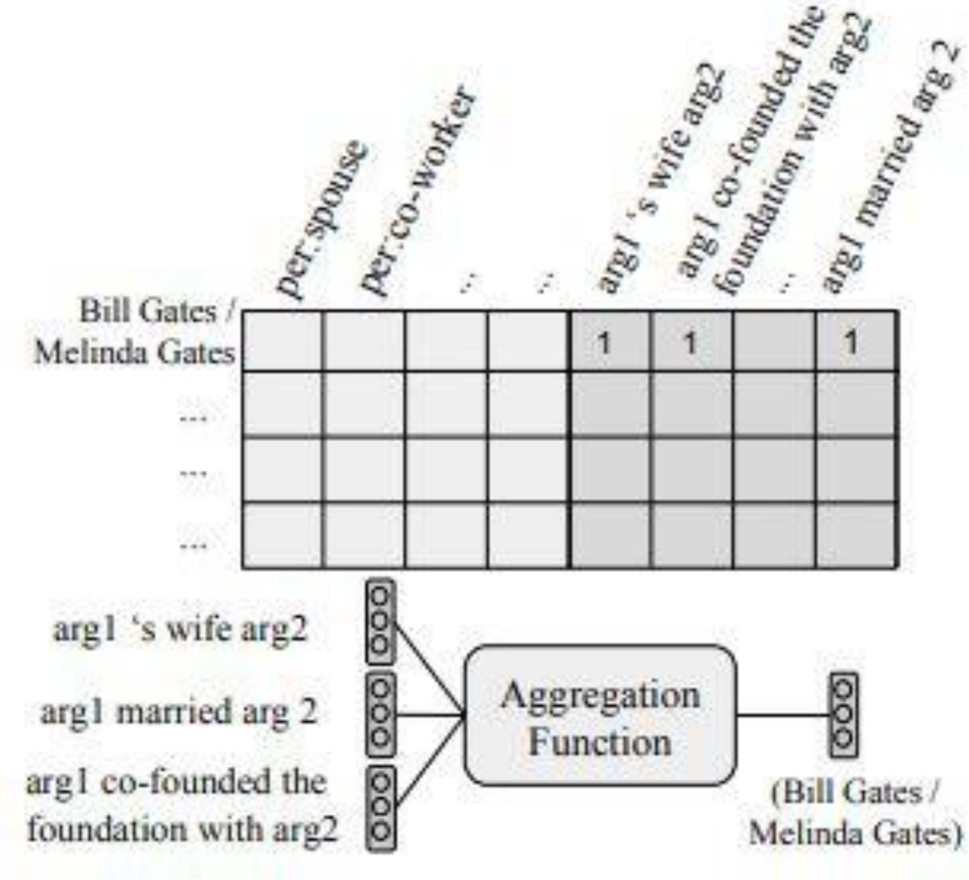
[Verga et al., ACL'16]

- Infer relation from a set of observed relations
- Similar to schema mapping w. signals from values

2013 (Deep ML)

Logic & Deep learning

- Collective disc. by PSL
- Universal schema



Rowless univ. schema

[Verga et al., ACL'16]

Rowless & Columnless

2013 (Deep ML)

Logic & Deep learning

- Collective disc. by PSL
- Universal schema

Model	MRR	Hits@10
Entity-pair Embeddings	31.85	51.72
Entity-pair Embeddings-LSTM	**33.37**	54.39
Attention	31.92	51.67
Attention-LSTM	30.00	53.35
Max Relation	31.71	51.94
Max Relation-LSTM	30.77	**54.80**

(a)

Recall still low

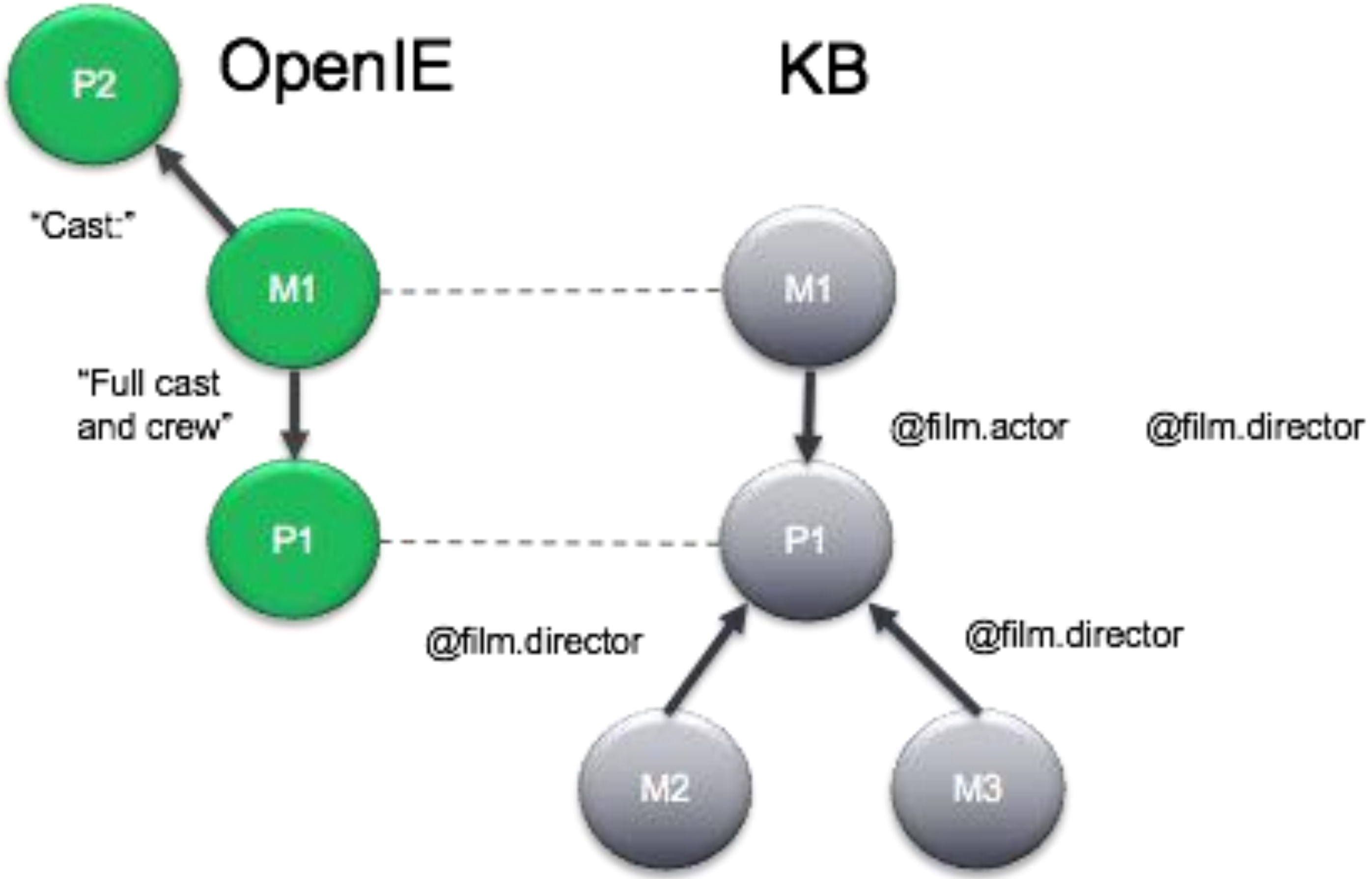
Model	MRR	Hits@10
Entity-pair Embeddings	5.23	11.94
Attention	**29.75**	49.69
Attention-LSTM	27.95	51.05
Max Relation	28.46	48.15
Max Relation-LSTM	29.61	**54.19**

(b)

Similar for new entity pairs

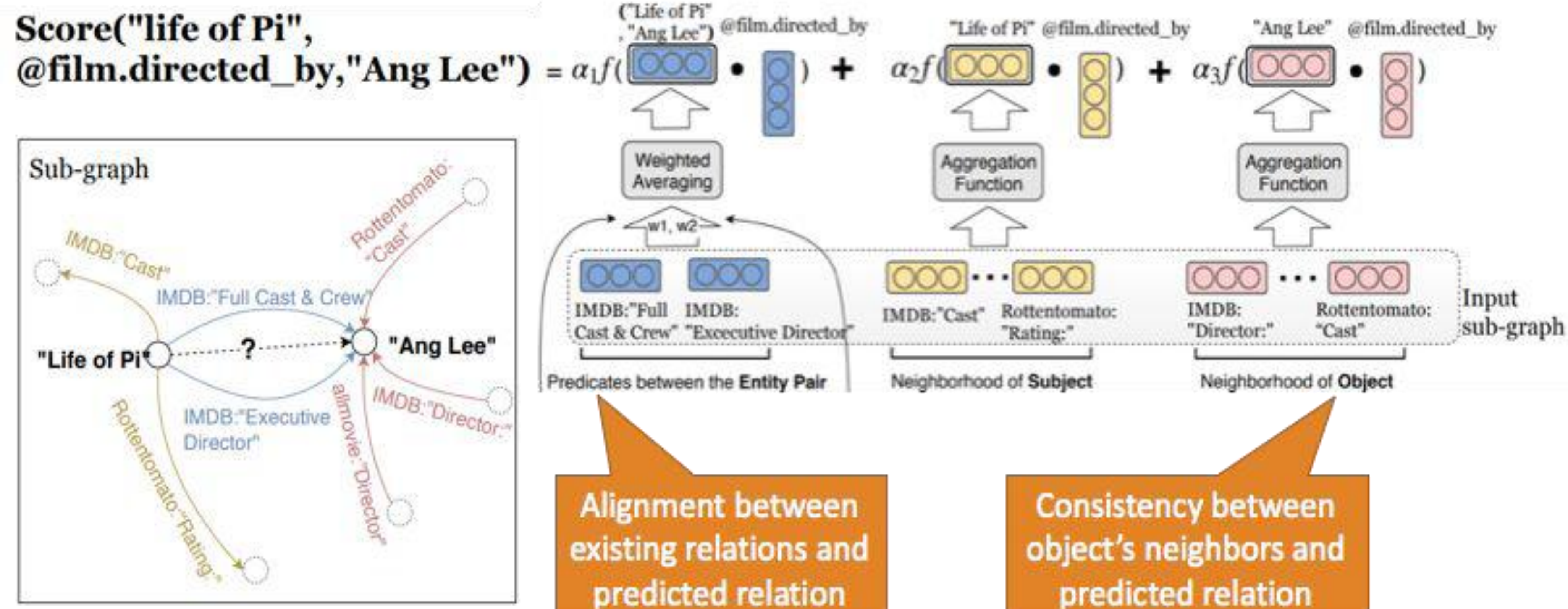
OpenKI: relation inference for OpenIE

[Zhang et al., NAACL'19]



OpenKI: relation inference for OpenIE

[Zhang et al., NAACL'19]



Alignment between existing relations and predicted relation

Consistency between object's neighbors and predicted relation

OpenKI: relation inference for OpenIE

[Zhang et al., NAACL'19]

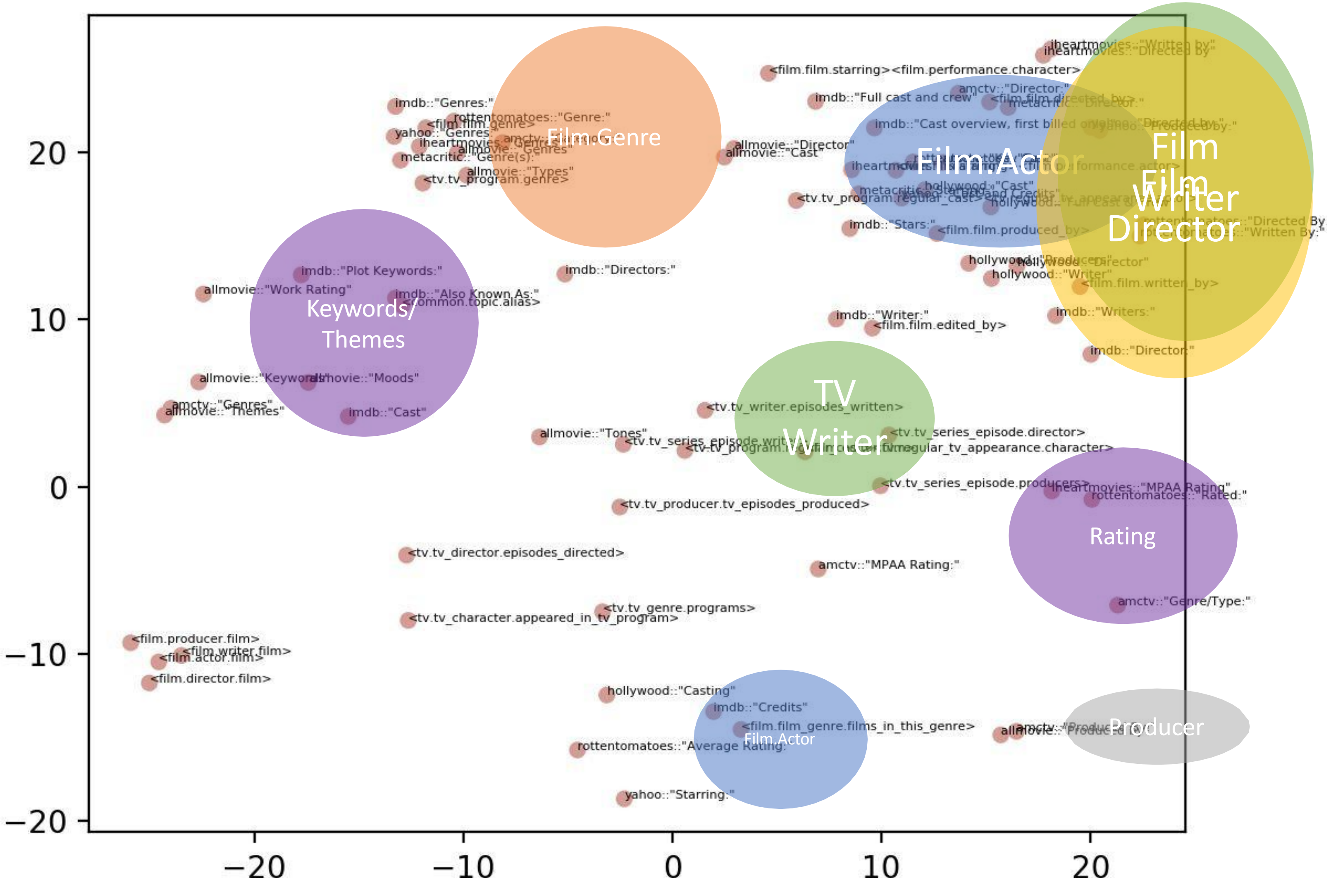
Models	All data	At least one seen
Rowless Model	0.278	0.282
OpenKI with Dual Att.	0.365	0.419

Table 5: Mean average precision (MAP) of Rowless and OpenKI on ReVerb + Freebase (/film) dataset.

Consider neighbors help

OpenKI: relation inference for OpenIE

[Zhang et al., NAACL’19]

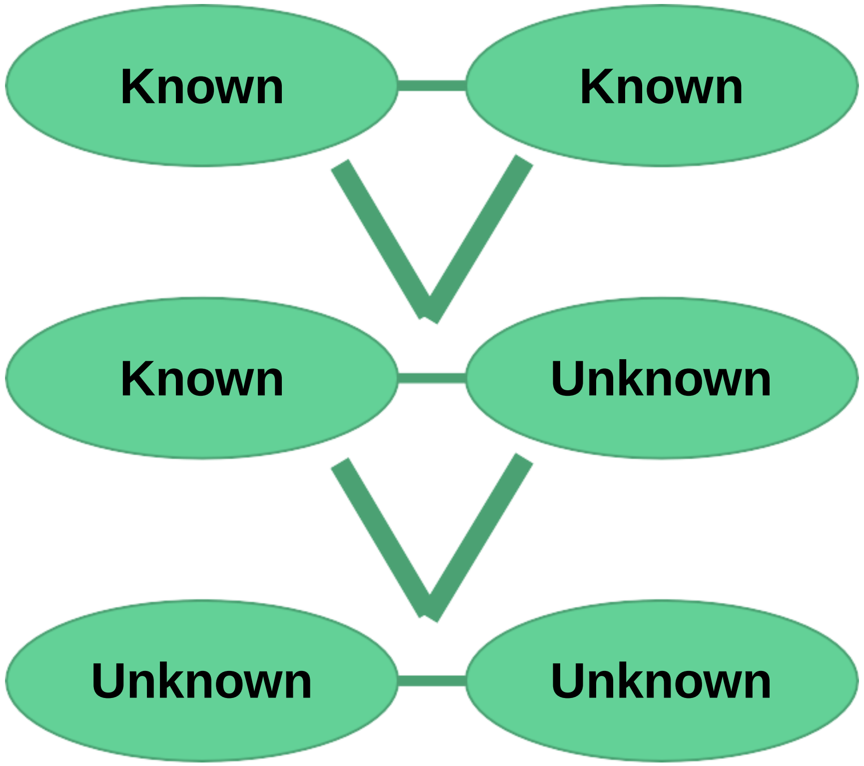


Schema mapping vs. universal schema

	Schema matching	**Universal schema**
Granularity	Column-level decision	Cell-level decision
Expressiveness	Mainly 1:1 mapping	Allow overlap, subset/superset, etc.
Signals	Name, description, type, key, graph structure, values	Values
Results	Accu: 70-90%	MRR=~0.3, Hits@10=~0.5
Community	Database	NLP

Challenges in applying deep learning on SM

- How can we combine techs from schema matching and universal schema?



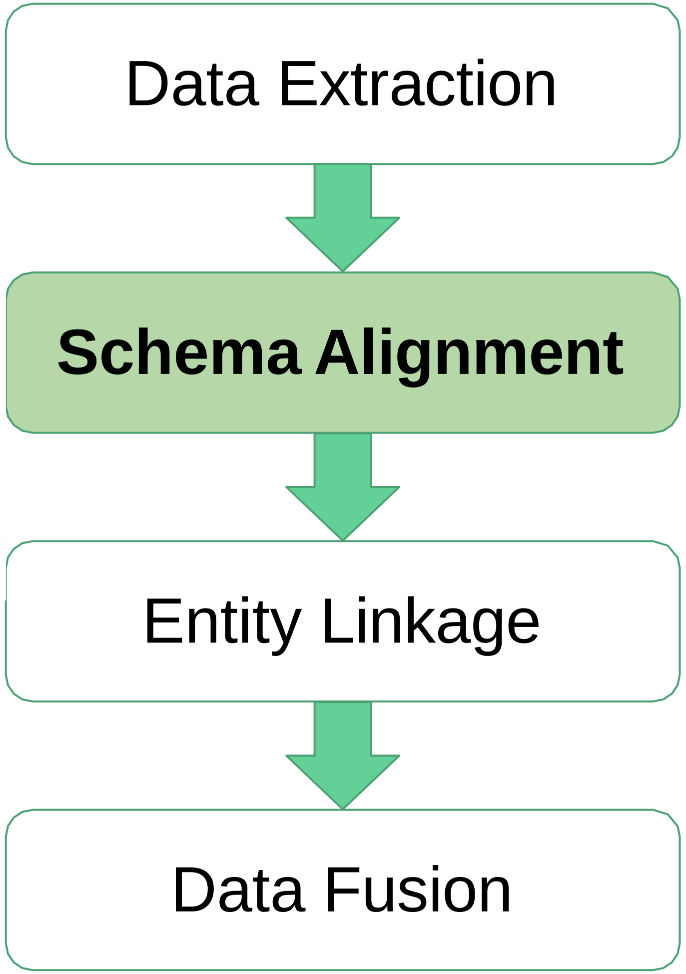
Leverage knowledge by inference

Leverage knowledge on types

Rowless

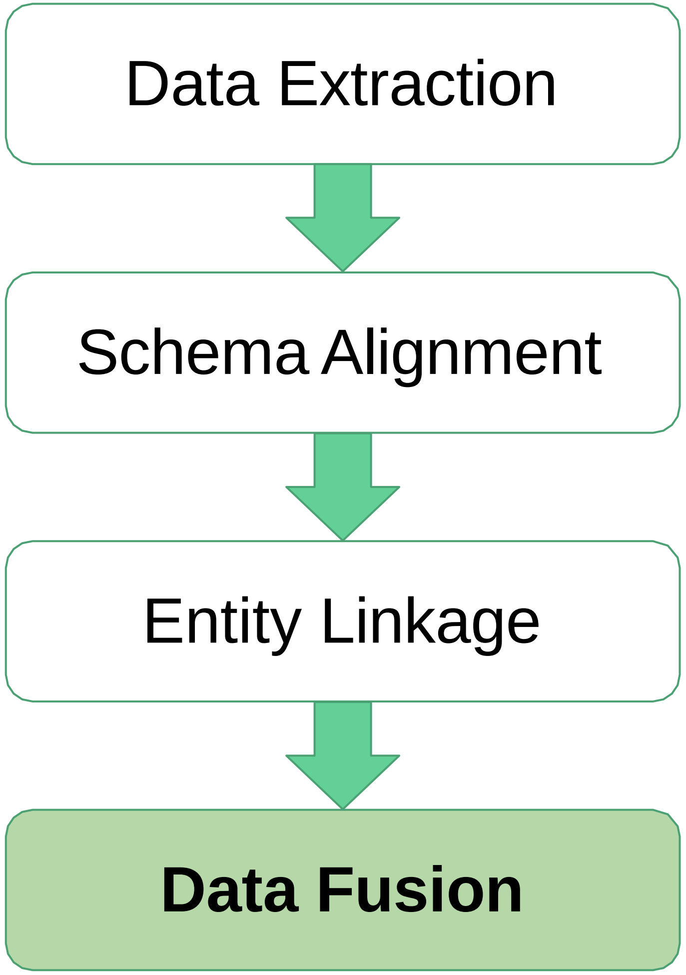
Recipe for schema alignment

- Problem definition: **Align attributes with the same semantics**
- Short answers
  - Interactive semi-automatic mapping
  - DL-based universal schema revived the field
  - Combine schema matching and universal schema for future



Today's agenda

- Part I. Introduction
- Part II. ML for DI
  - ML for entity linkage
  - ML for data extraction
  - ML for schema alignment
  - ML for data fusion



What is data fusion?

- **Definition:** Resolving conflicting data and verifying facts.
- **Example:** "*OK Google,How long is the Mississippi River?*"

Mississippi River / Length

2,320 mi

People also search for


Missouri River
2.341K mi


Nile
4.258K mi

Mississippi River

River in the United States of America

4.2 ★★★★ 400 Google reviews

The Mississippi River is the chief river of the second-largest drainage system on the North American continent, second only to the Hudson Bay drainage system. Wikipedia

Discharge: 593,000 cubic feet per second

Basin area: 1.151 million mi²

Source: Lake Itasca

Mouth: Gulf of Mexico

Country: United States of America

Did you know: The Mississippi River is the second-longest river in the US (2,202 mi). wikipedia.org

Mississippi River Facts - Mississippi National River and Recreation ...

https://www.nps.gov/miss/riverfacts.htm

Nov 14, 2017 - The staff of Itasca State Park at the Mississippi's headwaters suggest the main stem of the river is **2,552 miles** long. The US Geologic Survey has published a number of **2,300 miles**, the EPA says it is **2,320 miles** long, and the Mississippi National River and Recreation Area suggests the river's length is **2,350 miles**.

Longest main-stem rivers of the United States

#	Name	Mouth[5]	Length	Source coordinates[11]	Mouth coordinates[11]	Watershed area[12]	Discharge[12]	States, provinces, and image[5][11]
1	Missouri River	Mississippi River	2,341 mi 3,768 km[13]	45°55′39″N 111°30′29″W[14]	38°48′49″N 90°07′11″W	529,353 mi² 1,371,017 km²[15] ‡[n 2]	69,100 ft³/s 1,956 m³/s [n 3]	Montana[s], North Dakota, South Dakota, Nebraska, Iowa, Kansas, Missouri[m] 
2	Mississippi River	Gulf of Mexico	2,202 mi 3,544 km[17] [n 4]	47°14′22″N 95°12′29″W[18]	29°09′04″N 89°15′12″W	1,260,000 mi² 3,270,000 km²[19] ‡[n 5]	650,000 ft³/s 18,400 m³/s	Minnesota[s], Wisconsin, Iowa, Illinois, Missouri, Kentucky, Tennessee, Arkansas, Mississippi, Louisiana[m] 

The basic setup of data fusion

Source Observations

Source	River	Attribute	Value
KG	Mississippi River	Length	2,320 mi
KG	Missouri River	Length	2,341 mi
Wikipedia	Mississippi River	Length	2,202 mi
Wikipedia	Missouri River	Length	2,341 mi
USGS	Mississippi River	Length	2,340 mi
USGS	Missouri River	Length	2,540 mi

Fact

Conflicting value

Source reports a value for a fact

True Facts

River	Attribute	Value
Mississippi River	Length	?
Missouri River	Length	?

Fact's true value

Goal: Find the **latent** true value of facts.

The basic setup of data fusion

Source Observations

Source	River	Attribute	Value
KG	Mississippi River	Length	2,320 mi
KG	Missouri River	Length	2,341 mi
Wikipedia	Mississippi River	Length	2,202 mi
Wikipedia	Missouri River	Length	2,341 mi
USGS	Mississippi River	Length	2,340 mi
USGS	Missouri River	Length	2,540 mi

Fact

Source reports a value for a fact

Conflicting value

True Facts

River	Attribute	Value
Mississippi River	Length	?
Missouri River	Length	?

Fact's true value

Idea: Use *redundancy* to infer the true value of each fact.

Majority voting for data fusion

Source Observations

Source	River	Attribute	Value
KG	Mississippi River	Length	2,320 mi
KG	Missouri River	Length	2,341 mi
Wikipedia	Mississippi River	Length	2,202 mi
Wikipedia	Missouri River	Length	2,341 mi
USGS	Mississippi River	Length	2,340 mi
USGS	Missouri River	Length	2,540 mi

True Facts

River	Attribute	Value
Mississippi River	Length	?
Missouri River	Length	2,341

MV's assumptions

1. Sources report values independently
2. Sources are better than chance.

Majority voting can be limited. What if sources are correlated (e.g., copying)?

Idea: Model source quality for accurate results.

40 years of data fusion (beyond majority voting)

Dawid-Skene model
- Model the error-rate of sources
- Expectation-maximization

1979

(Statistical learning)

~1996 (Rule-based)

Domain-specific Strategies
- Keep all values
- Pick a random value
- Take the average value
- Take the most recent value
- …

Probabilistic Graphical Models
- Use of generative models
- Focus on unsupervised learning

2007 (Probabilistic)

2016 (Deep ML)

Deep learning
- Use Restricted Boltzmann Machine; one layer version is equivalent with Dawid-Skene model
- Knowledge graph embeddings

A probabilistic model for data fusion

- **Random variables:** Introduce a *latent random variable* to represent the true value of each fact.
- **Features:** Source observations become features associated with different random variables.
- **Model parameters:** Weights related to the error-rates of each data source.

error-rate scores
(model parameters)

$$P(\text{Fact} = v|\text{data}) = \frac{1}{Z} \exp \sum_{s \,\in\, \text{Sources}} \sum_{v' \,\in\, \text{Values}} \sigma_S^{v,v'} \cdot 1[S \text{ reports Fact} = v']$$

Normalizing constant

$$\sigma_S^{v,v'} = \log\left(\frac{\text{Error-rate of Source } S}{1 - \text{Error-rate of Source } S}\right)$$

Error-rate = probability that a source provides value v' instead of value v

The challenge of training data

- How much data do we need to train the data fusion model?
- **Theorem:** We need a number of labeled examples proportional to the number of sources [Ng and Jordan, NIPS'01]
- **Model parameters:** Weights related to the error-rates of each data source.

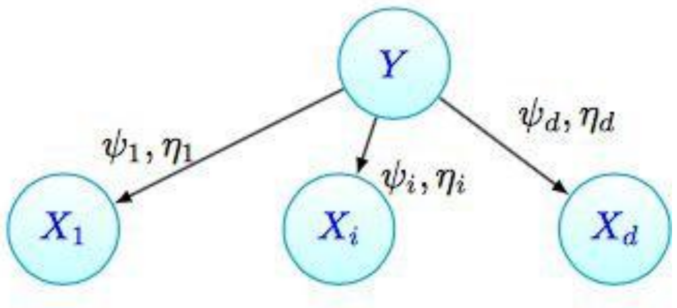
But the number of sources can be in the thousands or
millions and training data is limited!

Idea: Leverage redundancy and use unsupervised learning.

The Dawid-Skene Algorithm [Dawid and Skene, 1979]

Iterative process to estimate data source error rates

1. Initialize "inferred" true value for each fact (e.g., use majority vote)
2. Estimate **error rates** for workers (using "inferred" true values)
3. Estimate **"inferred" true values** (using error rates, weight source votes according to quality)
4. Go to Step 2 and iterate until convergence



Assumptions: (1) average source error rate < 0.5, (2) dense source observations, (3) conditional independence of sources, (4) errors are uniformly distributed across all instances.

Probabilistic Graphical Models

- Bayesian Networks (BNs)

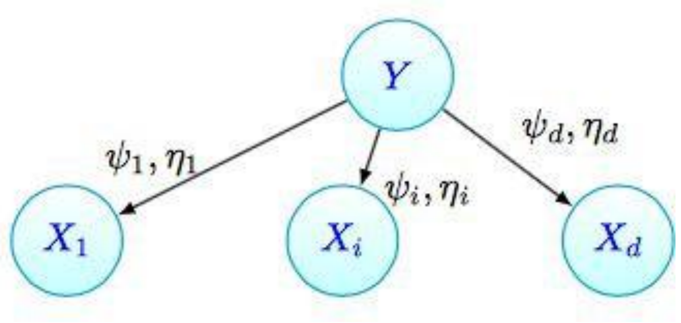
  Local Markov Assumption: A variable X is independent of its non-descendants given its parents (and *only* its parents).

- Recipe for BNs

  Set of random variables X

  Directed acyclic graph (each X[i] is a vertex)

  Conditional probability tables P(X |Parents(X))



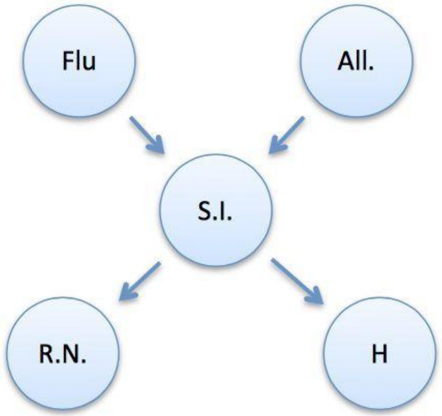
- Joint distribution: Factorizes over conditional probability tables

Probabilistic Graphical Models

- Where do independence assumptions come from?

  Causal structure captures domain knowledge

  - The flu causes sinus inflammation
  - Allergies *also* cause sinus inflammation
  - Sinus inflammation causes a runny nose
  - Sinus inflammation causes headaches

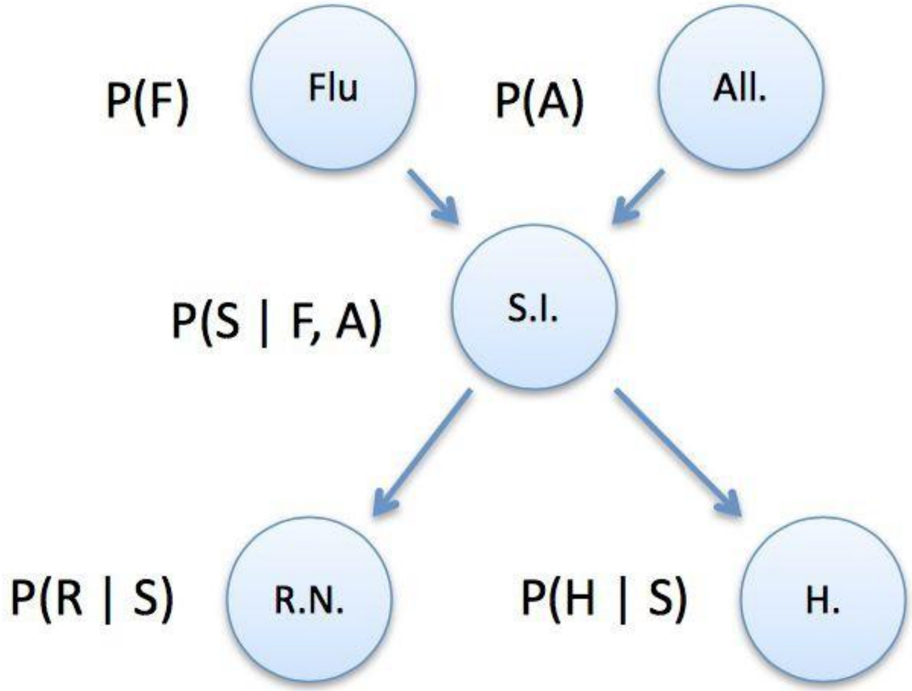


[Example by Andrew McCallum]

Probabilistic Graphical Models

Factored joint distribution

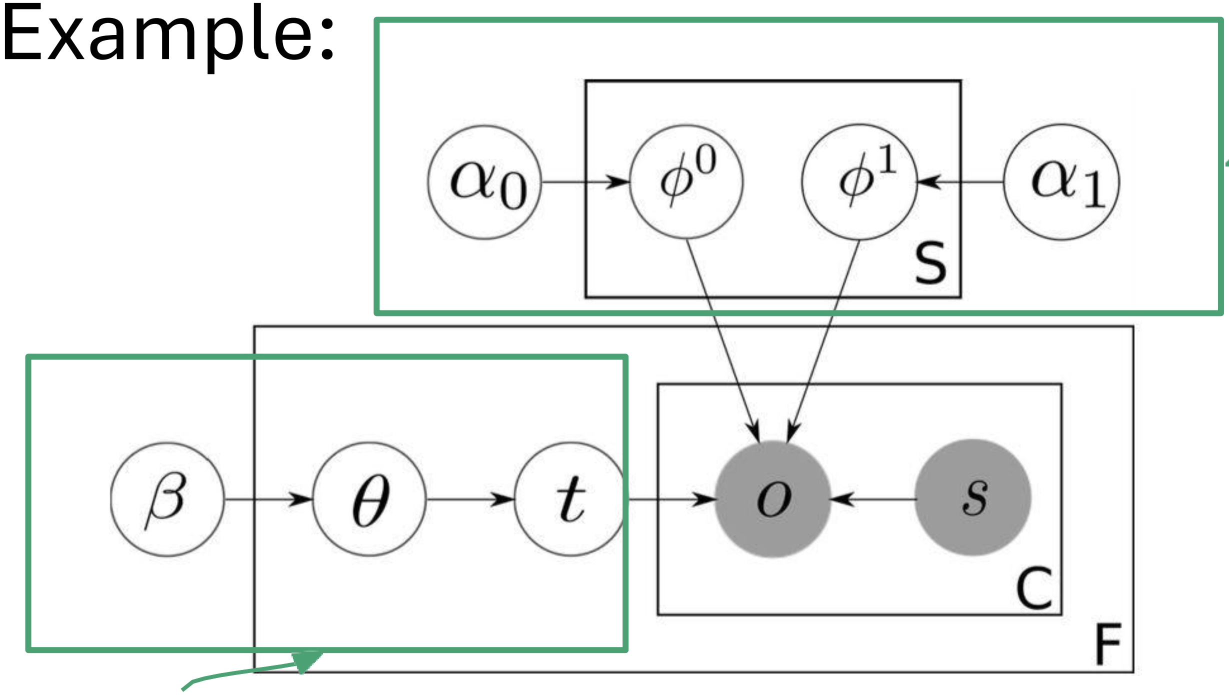
$$
\begin{aligned}
&P(F, A, S, R, H) \\
&= P(F) \\
&\quad P(A) \\
&\quad P(S \mid F, A) \\
&\quad P(R \mid S) \\
&\quad P(H \mid S)
\end{aligned}
$$



[Example by Andrew McCallum]

Probabilistic Graphical Models for data fusion

Example:



Source Quality

Prior truth probability

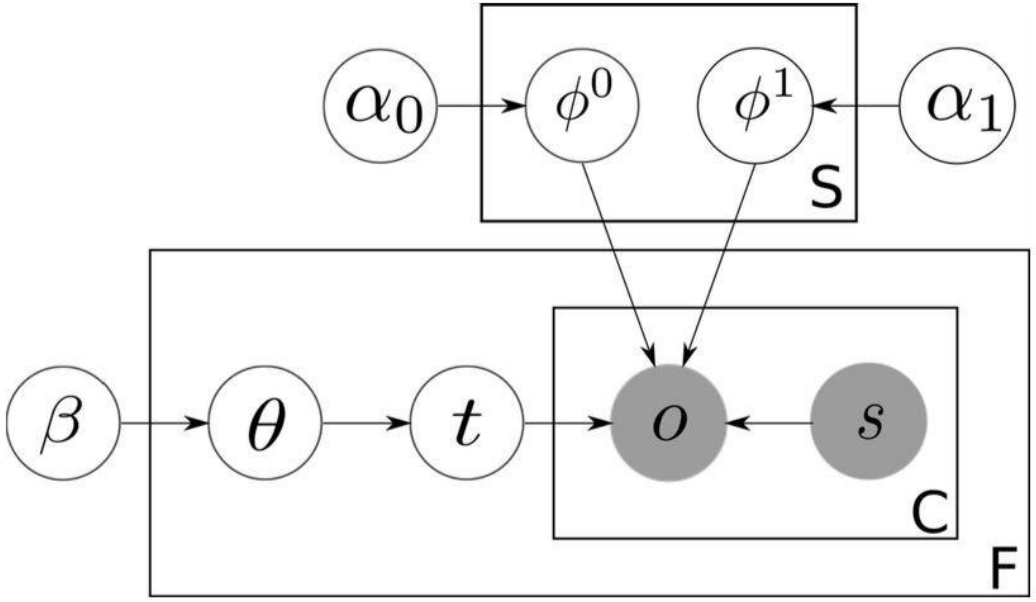
[Zhao et al., VLDB 2012]

Setup: Identify true source claims

Entity (Movie)	*Attribute (Cast)*	*Source*
Harry Potter	Daniel Radcliffe	IMDB
Harry Potter	Emma Waston	IMDB
Harry Potter	Rupert Grint	IMDB
Harry Potter	Daniel Radcliffe	Netflix
Harry Potter	Daniel Radcliffe	BadSource.com
Harry Potter	Emma Waston	BadSource.com
Harry Potter	Johnny Depp	BadSource.com
Pirates 4	Johnny Depp	Hulu.com
...	...	...

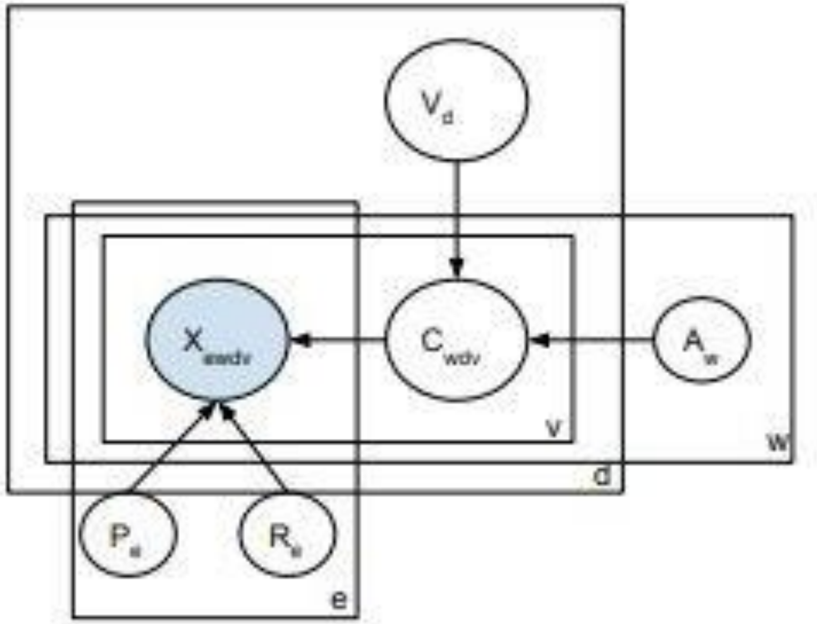
Extensive work on modeling source observations and source interactions to address limitations of basic Dawid-Skene.

Probabilistic Graphical Models for data fusion



[Zhao et al., VLDB 2012]

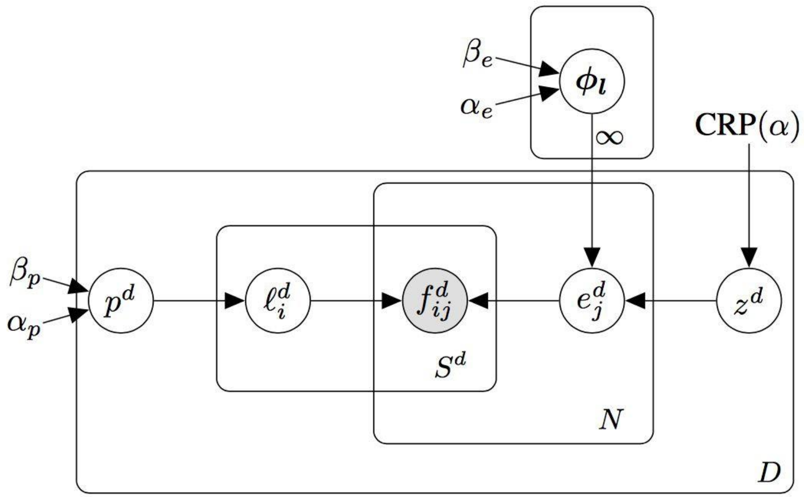
Modeling both source quality and extractor accuracy



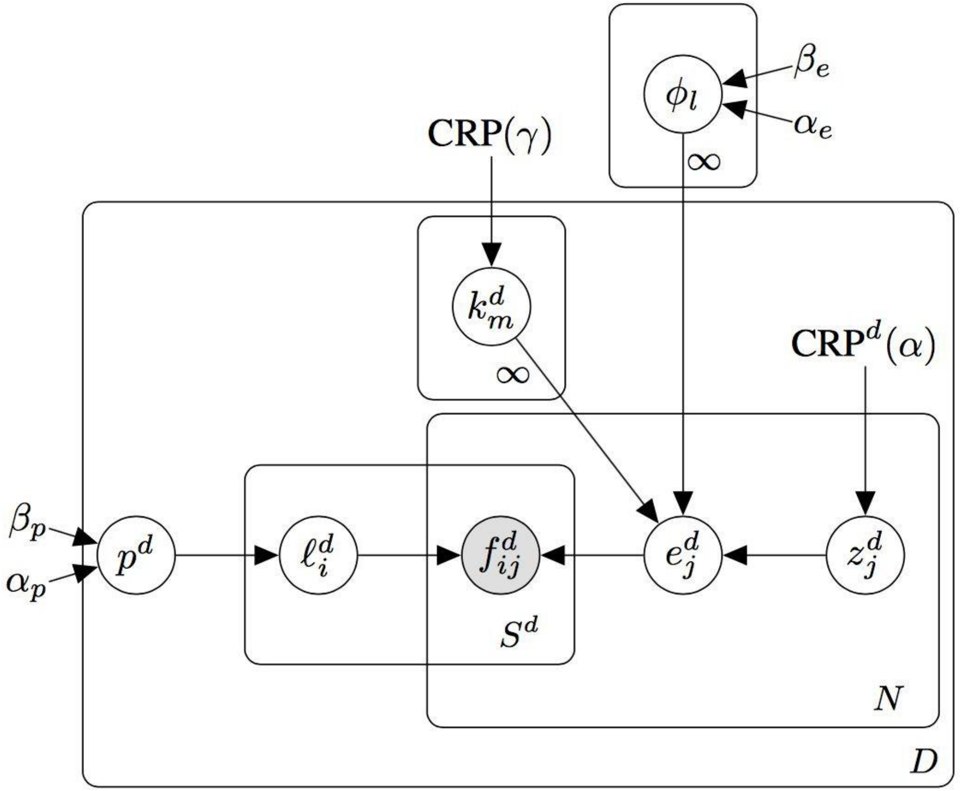
[Dong et al., VLDB 2015]

Extensive work on modeling source observations and source interactions to address limitations of basic Dawid-Skene.

Probabilistic Graphical Models for data fusion



Modeling source dependencies



[Platanios et al., ICML 2016]

Extensive work on modeling source observations and source interactions to address limitations of basic Dawid-Skene.

PGMs in data fusion [Li et al., VLDB'14]

Table 6: Summary of data-fusion methods. X indicates that the method considers the particular evidence.

Category	Method	#Providers	Source trustworthiness	Item trustworthiness	Value Popularity	Value similarity	Value formatting	Copying
Baseline	Vote	X						
Web-link based	HUB	X	X					
	AVGLOG	X	X					
	INVEST	X	X					
	POOLEDINVEST	X	X					
IR based	2-ESTIMATES	X	X					
	3-ESTIMATES	X	X	X				
	COSINE	X	X					
Bayesian based	TRUTHFINDER	X	X			X		
	ACCUPR	X	X					
	POPACCU	X	X		X			
	ACCUSIM	X	X			X		
	ACCUFORMAT	X	X			X	X	
Copying affected	ACCUCOPY	X	X			X	X	X

Bayesian models capture source observations and source interactions.

PGMs in data fusion [Li et al., VLDB'14]

Category	Method	*Stock* prec w. trust	*Stock* prec w/o. trust	*Stock* Trust dev	*Stock* Trust diff	*Flight* prec w. trust	*Flight* prec w/o. trust	*Flight* Trust dev	*Flight* Trust diff
Baseline	Vote	-	.908	-	-	-	.864	-	-
Web-link based	HUB	.913	.907	.11	.08	.939	.857	.2	.14
	AVGLOG	.910	.899	.17	-.13	.919	.839	.24	.001
	INVEST	.924	.764	.39	-.31	.945	.754	.29	-.12
	POOLEDINVEST	.924	.856	1.29	0.29	.945	***.921***	17.26	7.45
IR based	2-ESTIMATES	.910	.903	.15	-.14	.87	.754	.46	-.35
	3-ESTIMATES	.910	.905	.16	-.15	.87	.708	.95	-.94
	COSINE	.910	.900	.21	-.17	.87	.791	.48	-.41
Bayesian based	TRUTHFINDER	.923	.911	.15	.12	***.957***	.793	.25	.16
	ACCUPR	.910	.899	.14	-.11	.91	.868	.16	-.06
	POPACCU	.909	.892	.14	-.11	***.958***	***.925***	.17	-.11
	ACCUSIM	.918	***.913***	.17	-.16	.903	.844	.2	-.09
	ACCUFORMAT	.918	.911	.17	-.16	.903	.844	.2	-.09
	ACCUSIMATTR	***.950***	***.929***	.17	-.16	.952	.833	.19	-.08
	ACCUFORMATATTR	***.948***	**.930**	.17	-.16	.952	.833	.19	-.08
Copying affected	ACCUCOPY	**.958**	.892	.28	-.11	**.960**	**.943**	.16	-.14

Modeling the quality of data sources leads to improved accuracy.

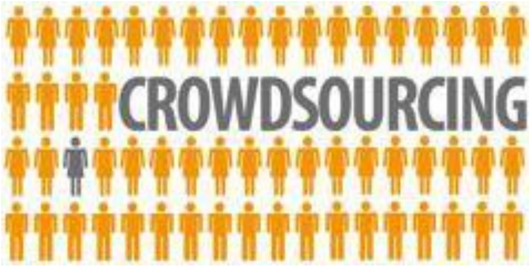
Discriminative data fusion [SLiMFast Rekatsinas et al., SIGMOD’17]

Limit the informative parameters of the model by using domain knowledge and use semi-supervised learning

Key Idea: Sources have (domain specific) features that are indicative of error rates

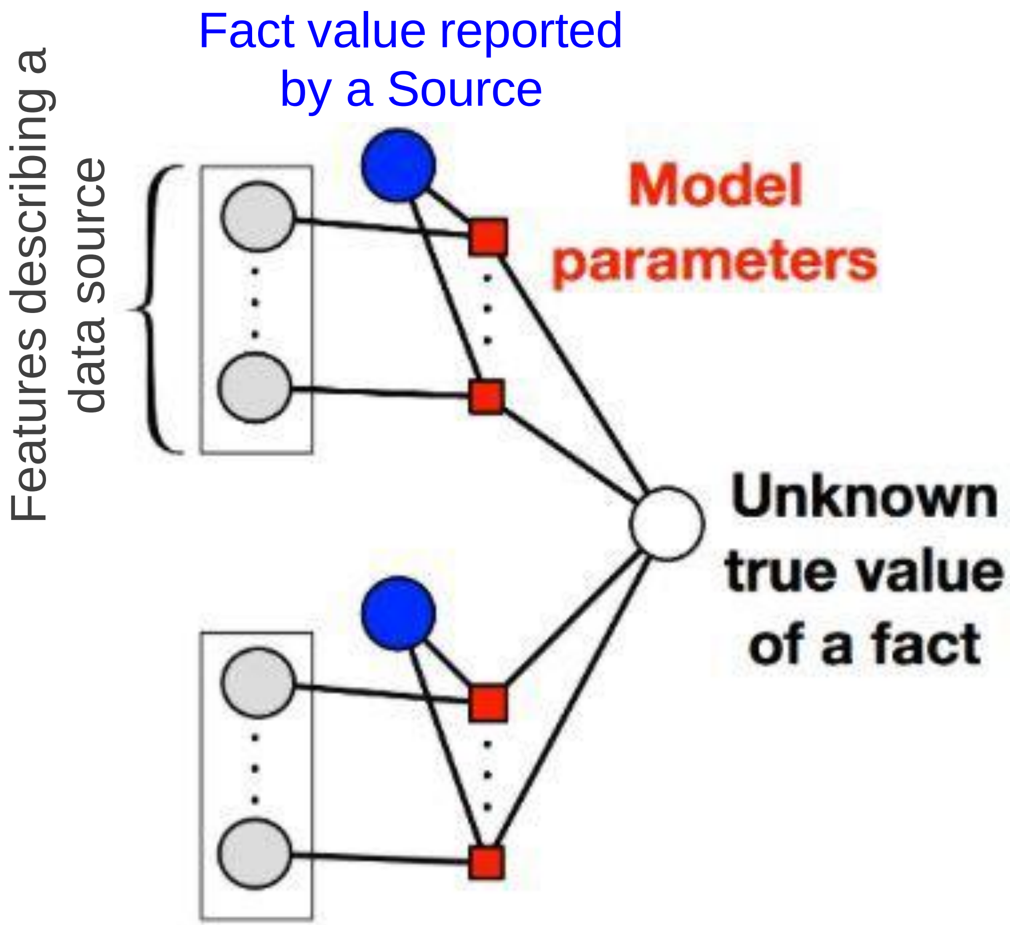
Example:

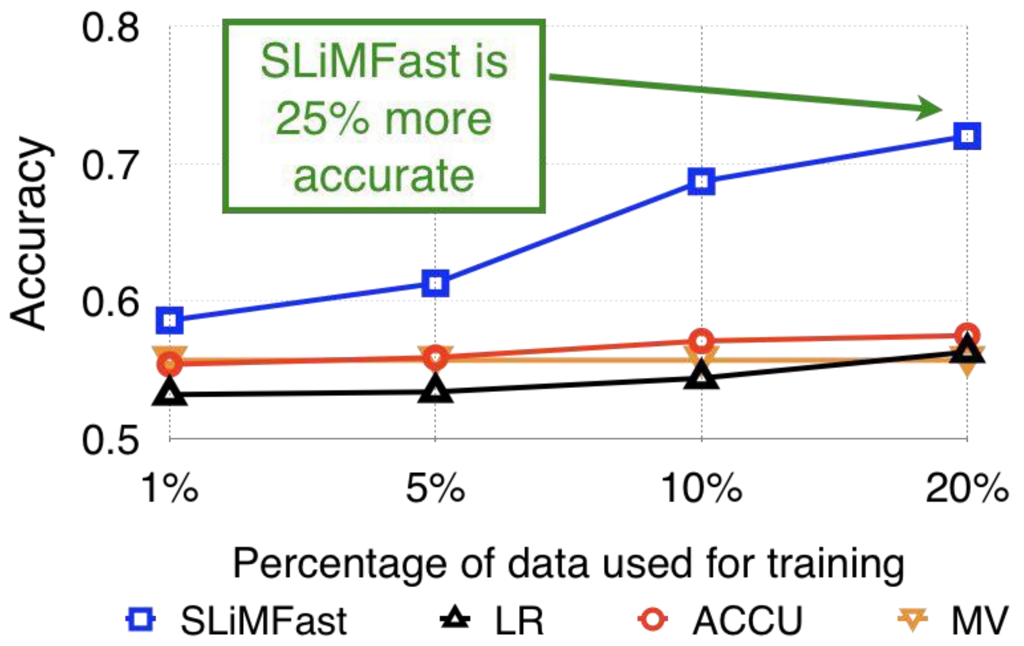
- newly registered similar to existing domain
- traffic statistics
- text quality (e.g., misspelled words, grammatical errors)
- sentiment analysis



- avg. time per task
- number of tasks
- market used

Discriminative data fusion [SLiMFast Rekatsinas et al., SIGMOD'17]

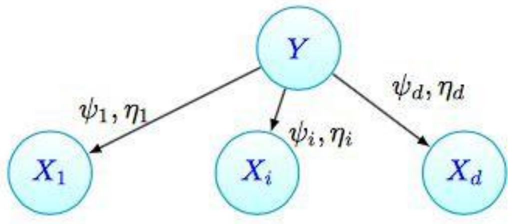




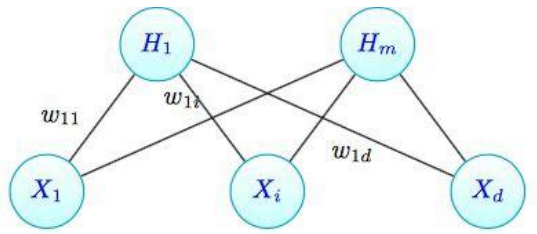
Genomics data: 2.7k sources (articles), 571 objects (gene-disease), 4 domain features (year, citation, author, journal)

Data fusion and Deep Learning [Shaham et al., ICML'16]

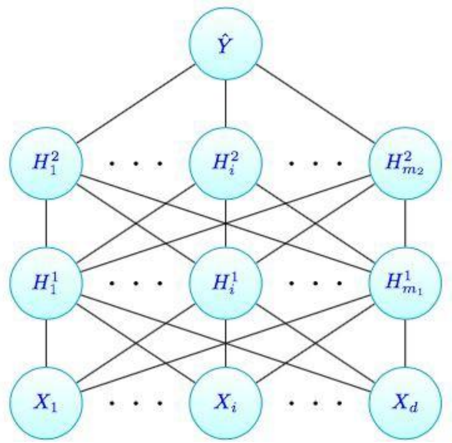
Theorem: The Dawid and Skene model is *equivalent* to a Restricted Boltzmann Machine (RBM) with a single hidden node.



Dawid and Skene model.



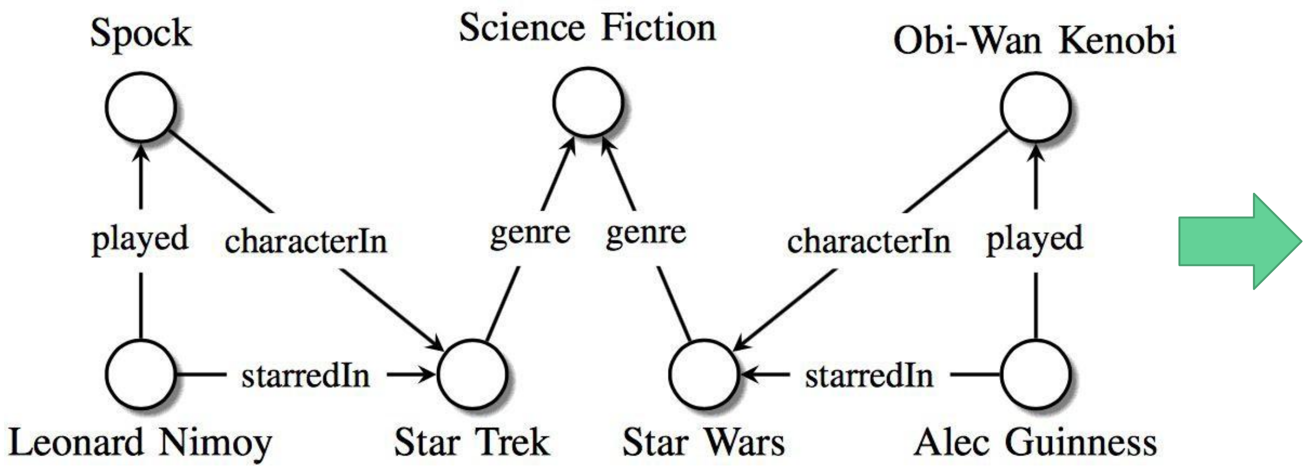
A RBM with *d* visible and *m* hidden units.

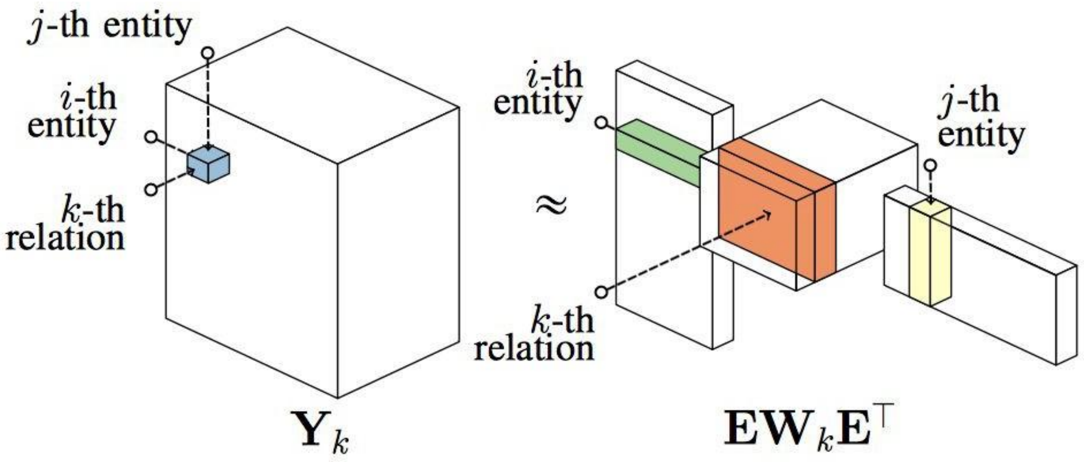


Sketch of a two-hidden-layer RBM-based DNN.

When the conditional independence assumption of Dawid-Skene does not hold, a better approximation may be obtained from a deeper network.

Data fusion for complex data

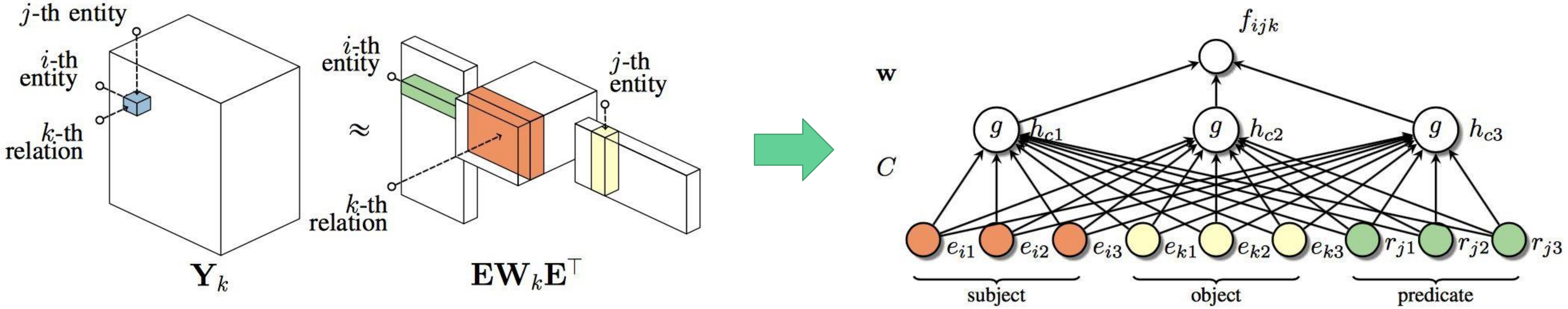




Knowledge Graph Embeddings [Survey: Nicket et al., 2015]

A knowledge graph can be encoded as a tensor.

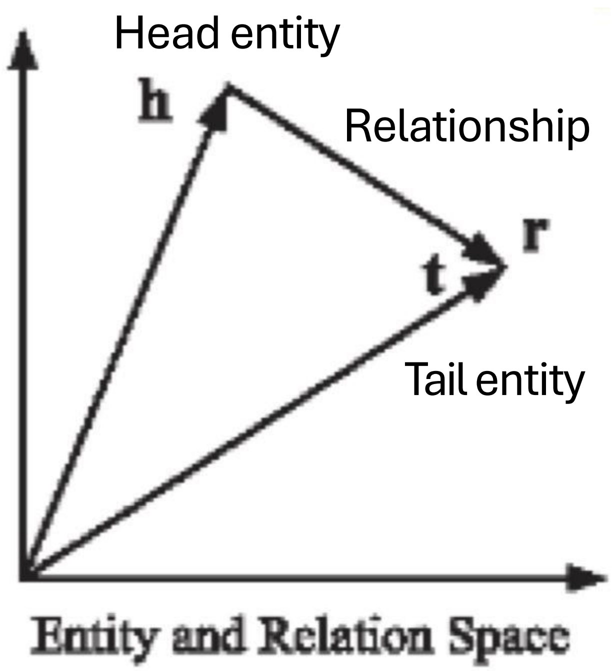
Data fusion for complex data



Knowledge Graph Embeddings [Survey: Nicket et al., 2015]

Neural networks can be used to obtain richer representations.

Data fusion for complex data



- TransE: $score(h,r,t) = -\|h+r-t\|_{1/2}$
- Hot field with increasing interest [Survey by Wang et al., TKDE 2017]

Example: Learn embeddings from IMDb data and identify various types of errors in WikiData [Dong et al., KDD'18]

Subject	Relation	Target	Reason
The Moises Padilla Story	writtenBy	César Ámigo Aguilar	Linkage error
Bajrangi Bhaijaan	writtenBy	Yo Yo Honey Singh	Wrong relationship
Piste noire	writtenBy	Jalil Naciri	Wrong relationship
Enter the Ninja	musicComposedBy	Michael Lewis	Linkage error
The Secret Life of Words	musicComposedBy	Hal Hartley	Cannot confirm
...	...	...	...

Challenges in data fusion

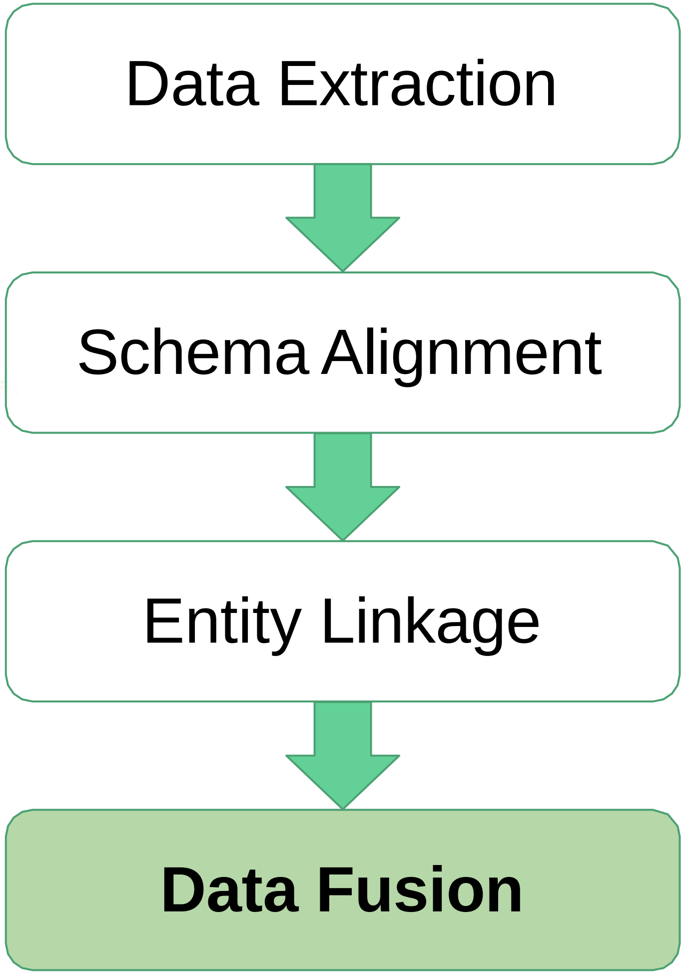
- There are few solutions for unstructured data. Mostly work on fact verification [Tutorial by Dong et al., KDD`2018]. Most data Fusion solutions assume data extraction. Can state-of-the art DL help?

- Using training data is key and semi-supervised learning can significantly improve the quality of Data Fusion results. How can one collect training data effectively without manual annotation?

- We have only scratched the surface of what representation learning and deep learning methods can offer. Can deep learning streamline data fusion? What are its limitations?

Recipe for data fusion

- Problem definition: Resolve conflicts and obtain correct values
- Short answers
  - Reasoning about source quality is key and works for easy cases
  - Semi-supervised learning has shown BIG potential
  - Representation learning provides positive evidence for streamlining data fusion.



DI & ML as Synergy

- **ML for effective DI: AUTOMATION, AUTOMATION, AUTOMATION**
  - Automating DI tasks with training data
  - Ensemble learning and deep learning provide promising solutions
  - Better understanding of semantics by neural network

- **DI for effective ML: DATA, DATA, DATA**
  - The software 2.0 stack is data hungry
  - Create large-scale training datasets from different sources
  - Cleaning of data used for training

DI and ML: A natural synergy

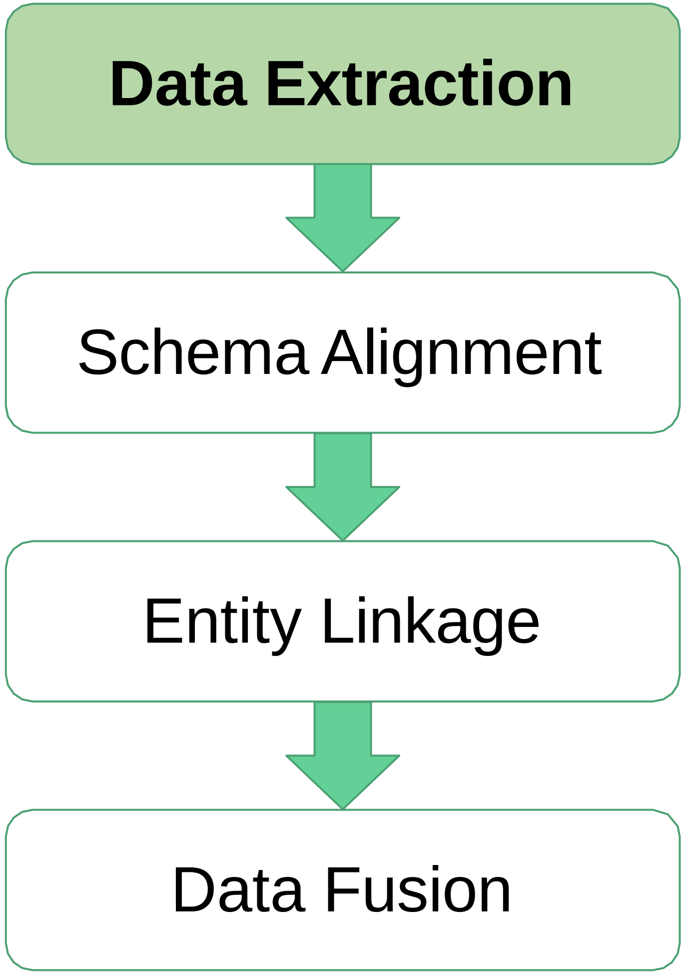
- Data integration is one of the oldest problems in data management

- Transition from logic to probabilities revolutionized data integration
    - Probabilities allow us to reason about inherently noisy data
    - Similar to the AI-revolution in the 80s [https://vimeo.com/48195434]

- Modern machine learning and deep learning have the power to streamline DI

Revisit: recipe for data extraction

- Problem definition: Extract structure from semi- or un-structured data
- Short answers
  - Wrapper induction has high prec/rec
  - Distant supervision is critical for collecting training data
  - DL effective for texts and LR is often effective for semi-stru data



Revisit: recipe for schema alignment

- Problem definition: Align attributes with the same semantics
- Short answers
  - Interactive semi-automatic mapping
  - DL-based universal schema revived the field
  - Combine schema matching and universal schema for future

Data Extraction

↓

Schema Alignment

↓

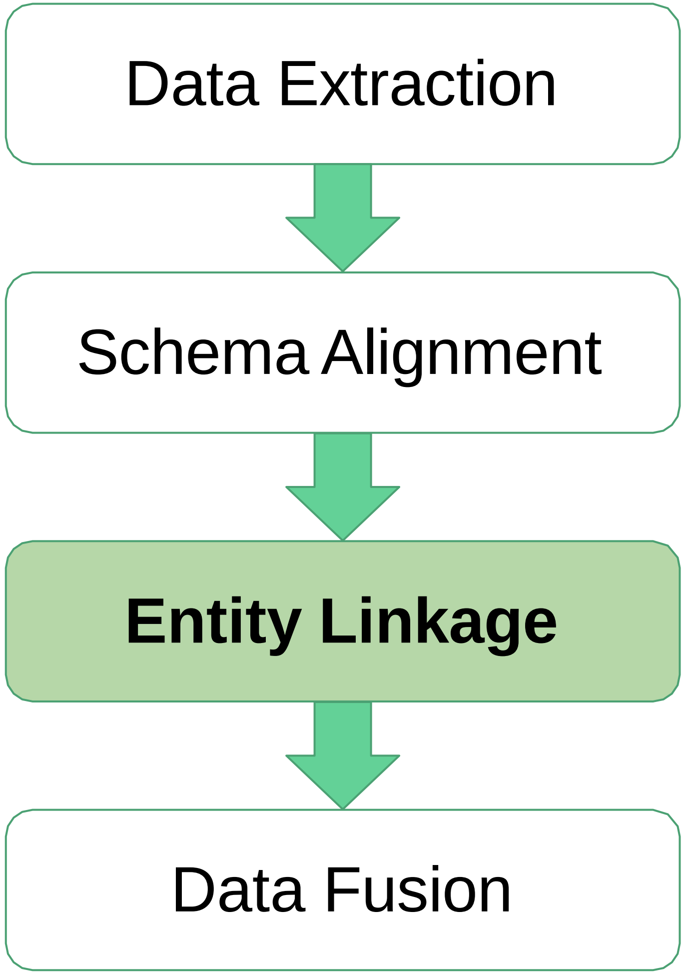
Entity Linkage

↓

Data Fusion

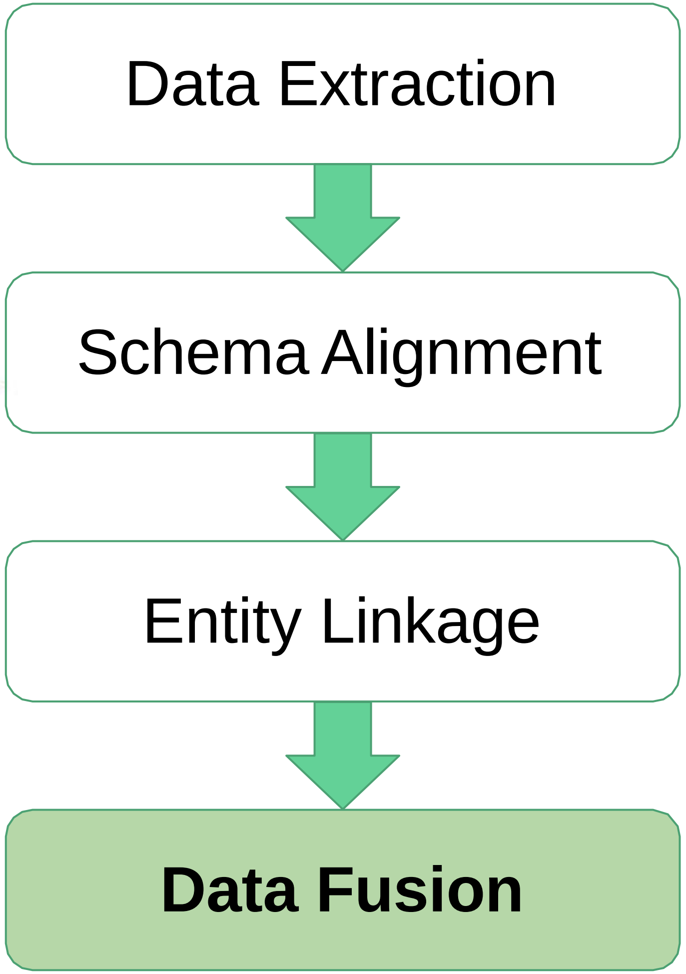
Revisit: recipe for entity linkage

- 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



Recipe for data fusion

- Problem definition: Resolve conflicts and obtain correct values
- Short answers
  - Reasoning about source quality is key and works for easy cases
  - Semi-supervised learning has shown BIG potential
  - Representation learning provides positive evidence for streamlining data fusion.



Credits

- Luna Dong Xin
- Theo Rekatsinas