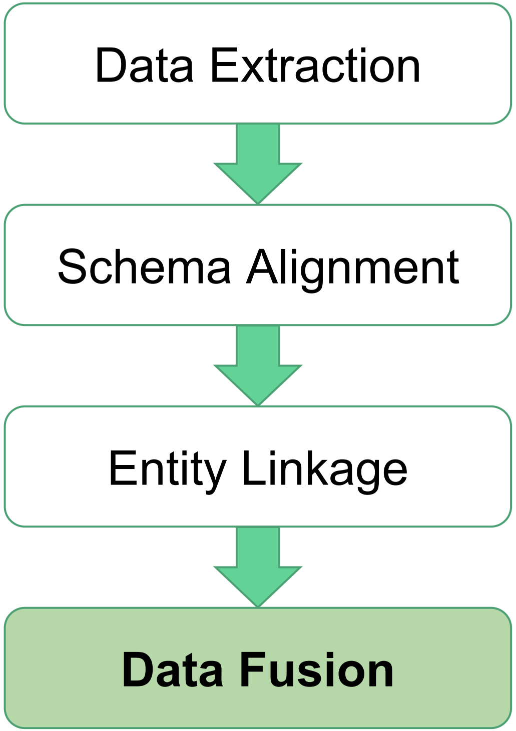
Outline

- 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 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**.

#	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

Conflicting value

Source reports
a value for a fact

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.

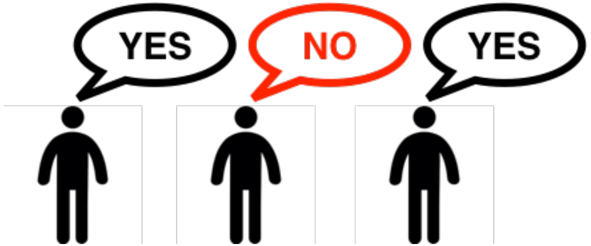
Majority Voting for Data Fusion

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



Majority voting can be limited. What if sources are correlated (e.g., copying)?
Idea: Model source quality for accurate results.

MV's assumptions
1. Sources report values independently
2. Sources are better than chance.

40 Years of Data Fusion (beyond Majority Voting)

1979 (Statistical learning)

Dawid-Skene model

- Model the error-rate of sources
- Expectation-maximization

~1996 (Rule-based)

Domain-specific Strategies

- Keep all values
- Pick a random value
- Take the average value
- Take the most recent value
- ...

2007 (Probabilistic)

Probabilistic Graphical Models

- Use of generative models
- Focus on unsupervised learning

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.

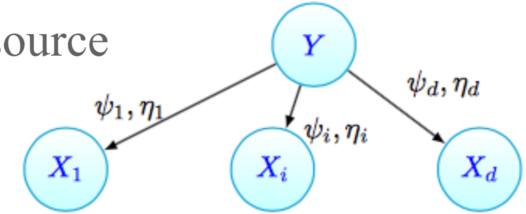
But the number of sources can be in the thousands or millions
and training data is limited!

Idea 1: Leverage redundancy and use unsupervised learning.

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

Iterative process to estimate data source error rates

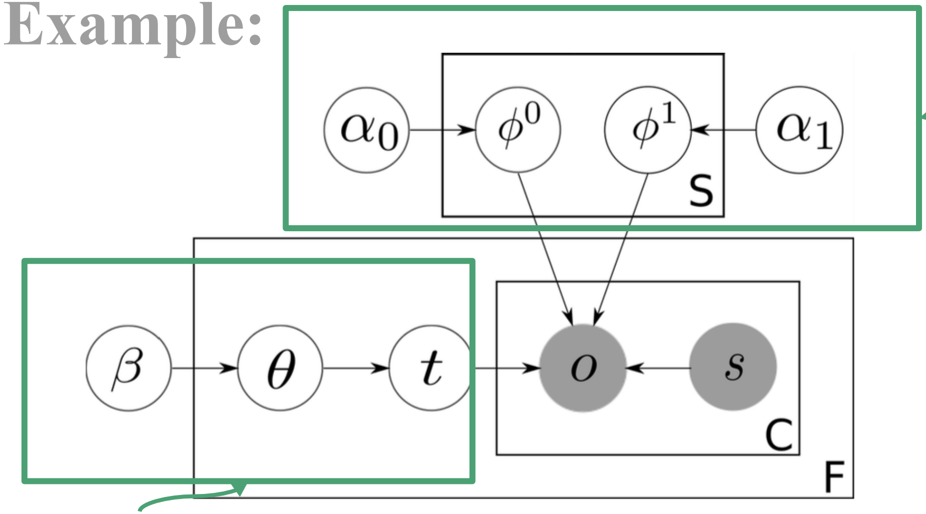
1. Initialize "inferred" true value for each fact (e.g., use majority vote)
2. Estimate **error rates** for workers (using "inferred" true values)
3. Estimate **"inferred" true values** (using error rates, weight source votes according to quality)
4. Go to Step 2 and iterate until convergence



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 for Data Fusion

Example:



Source Quality

Prior truth probability

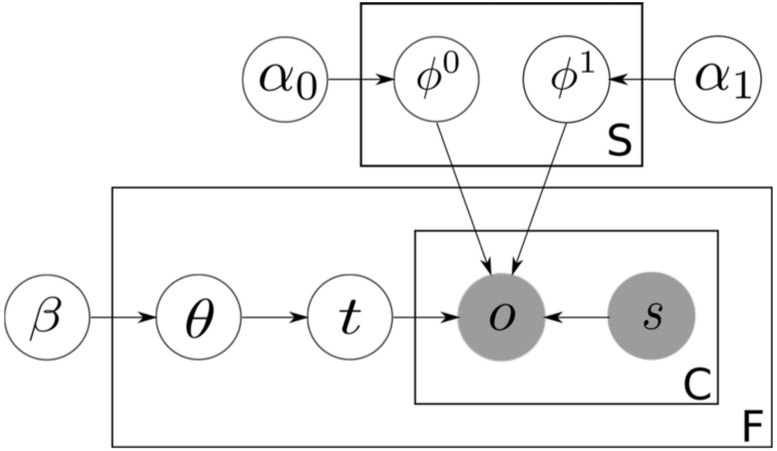
[Zhao et al., VLDB 2012]

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
...	...	...

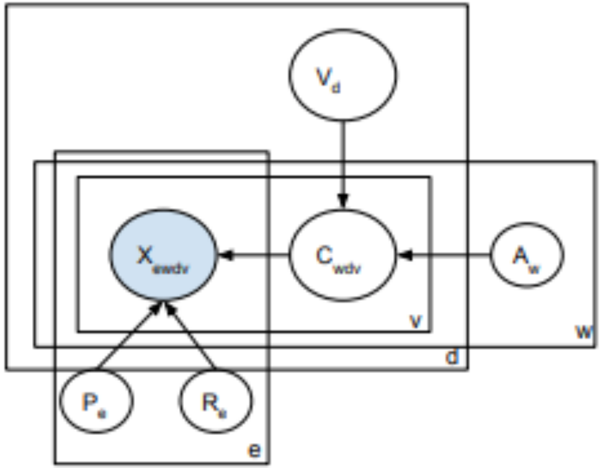
Extensive work on modeling source observations and source interactions to address limitations of basic Dawid-Skene.

Probabilistic Graphical Models for Data Fusion



[Zhao et al., VLDB 2012]

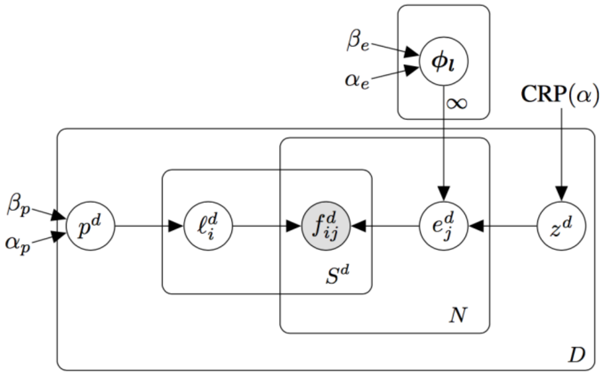
Modeling both source quality and extractor accuracy



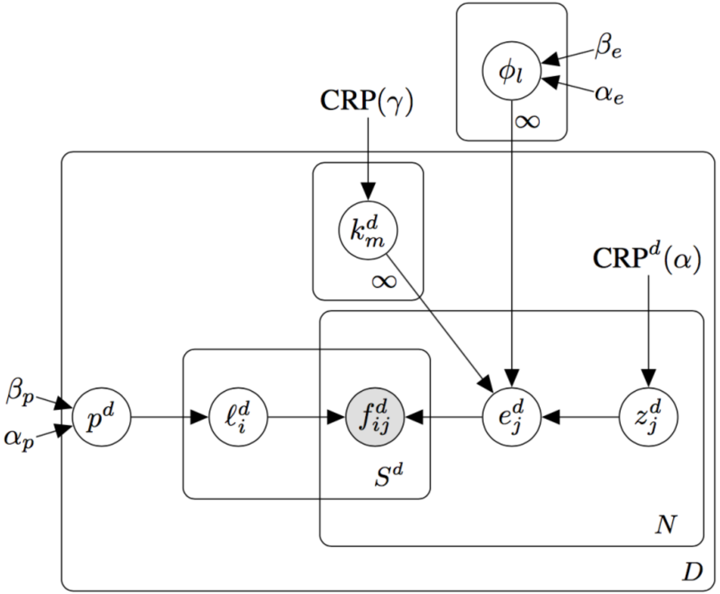
[Dong et al., VLDB 2015]

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.

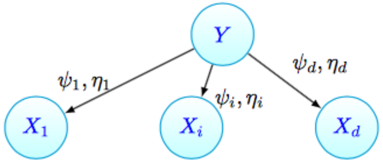
PGMs in Data Fusion [Li et al., VLDB'14]

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

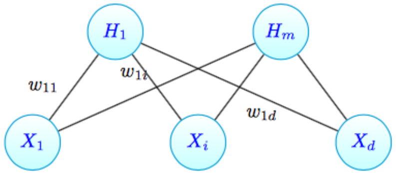
Modeling the quality of data sources leads to improved accuracy.

Dawid-Skene and Deep Learning [Shaham et al., ICML'16]

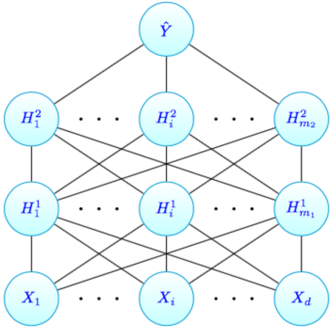
Theorem: The Dawid and Skene model is *equivalent* to a Restricted Boltzmann Machine (RBM) with a single hidden node.



Dawid and Skene model.



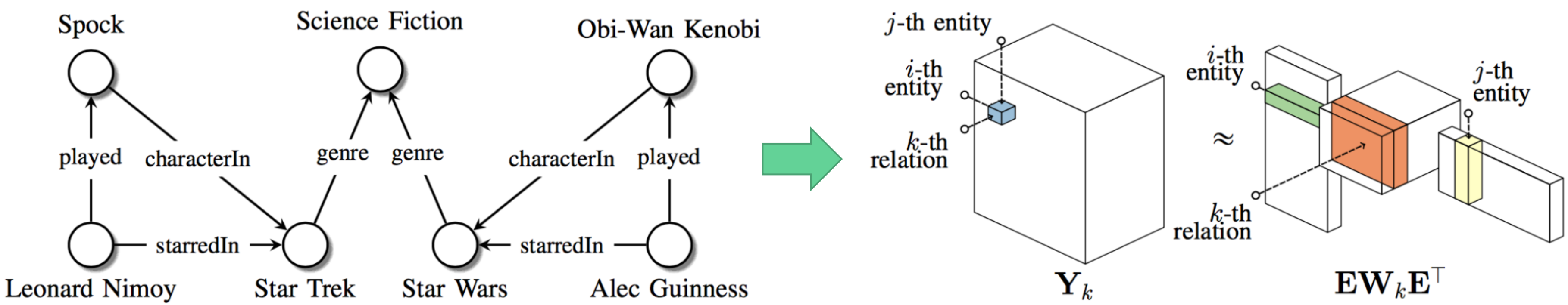
A RBM with *d* visible and *m* hidden units.



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

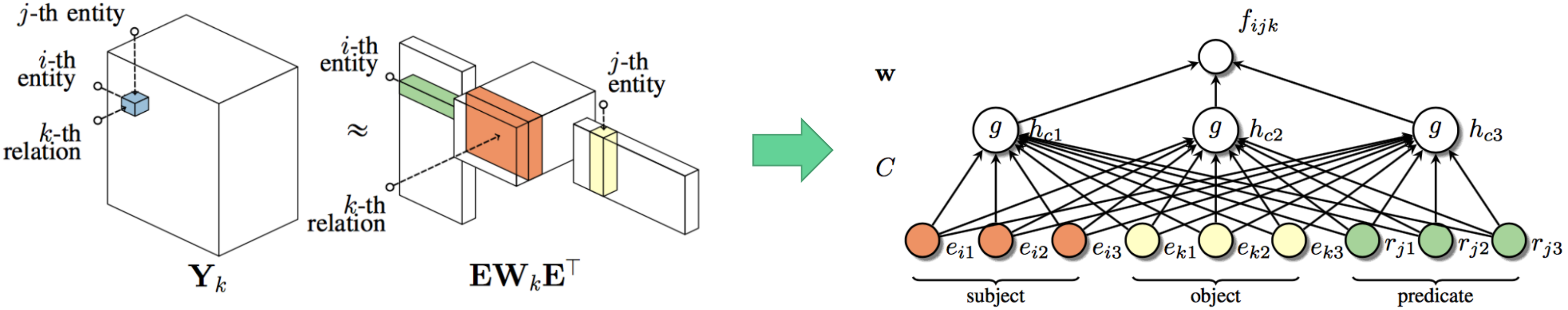
When the conditional independence assumption of Dawid-Skene does not hold, a better approximation may be obtained from a deeper network.

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



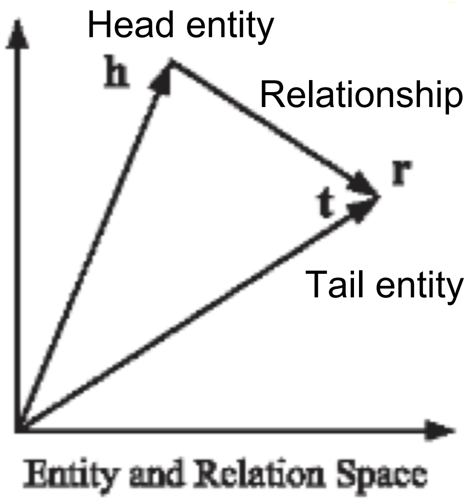
A knowledge graph can be encoded as a tensor.

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



Neural networks can be used to obtain richer representations.

Knowledge Graph Embeddings



- 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
...	...	...	...

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 1: Leverage redundancy and used unsupervised learning.
Idea 2: Limit model parameters and use a small number of training data.

SLiMFast: Discriminative Data Fusion [Rekatsinas et al., SIGMOD'17]

Limit the informative parameters of the model by using domain knowledge

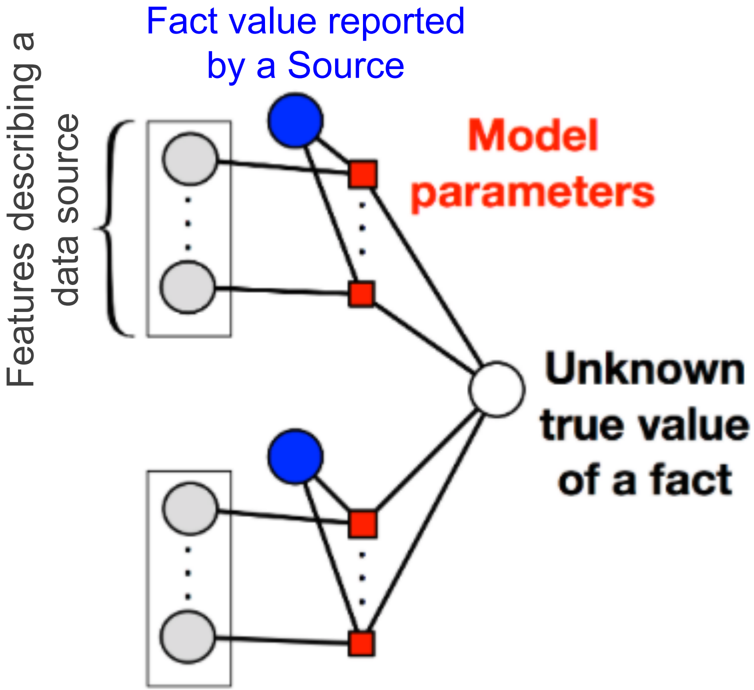
Key Idea: Sources have (domain specific) features that are indicative of error rates

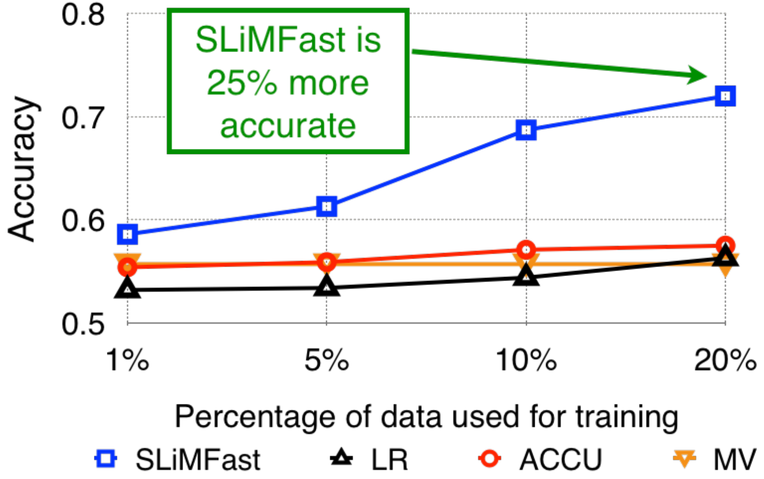
Example:

- 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

SLiMFast: Discriminative Data Fusion [Rekatsinas et al., SIGMOD'17]





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

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.**