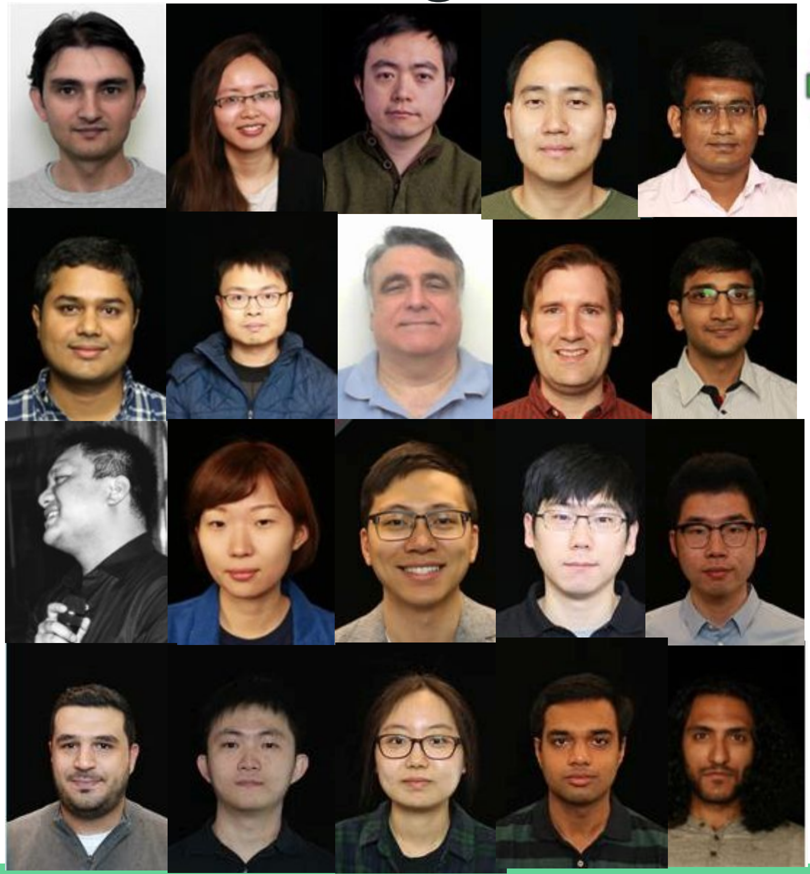
Data Integration and Machine Learning: A Natural Synergy

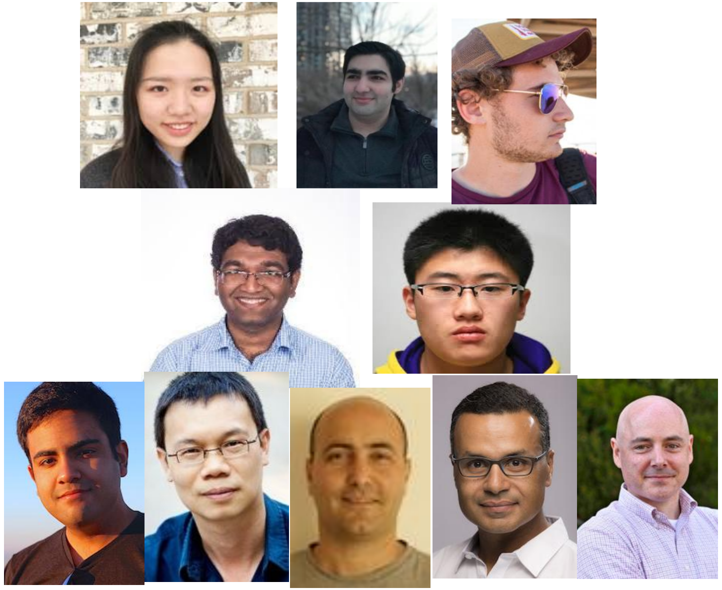
Xin Luna Dong @ Amazon.com
Theo Rekatsinas @ UW-Madison
http://dataintegration.ml

Acknowledgement



product graph

snorkel



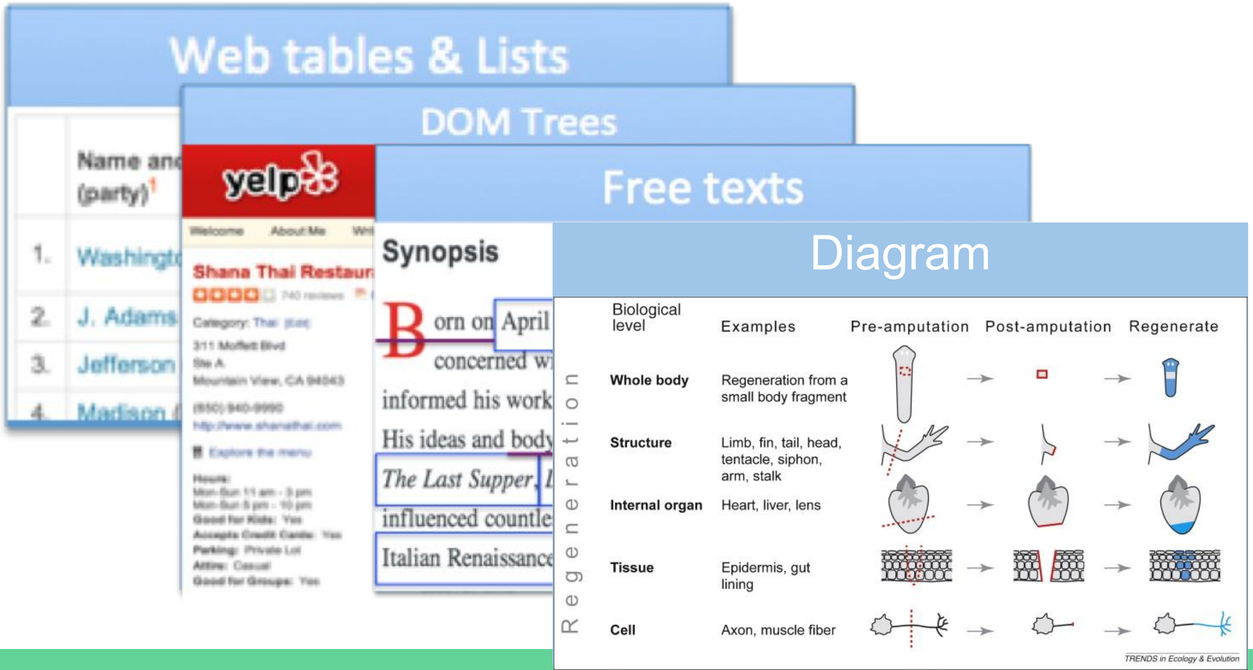
What is Data Integration?

- **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 RDF world: 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

Why is Data Integration Hard?

- Heterogeneity everywhere
  - Different data formats



Data Extraction

↓

Schema Alignment

↓

Entity Linkage

↓

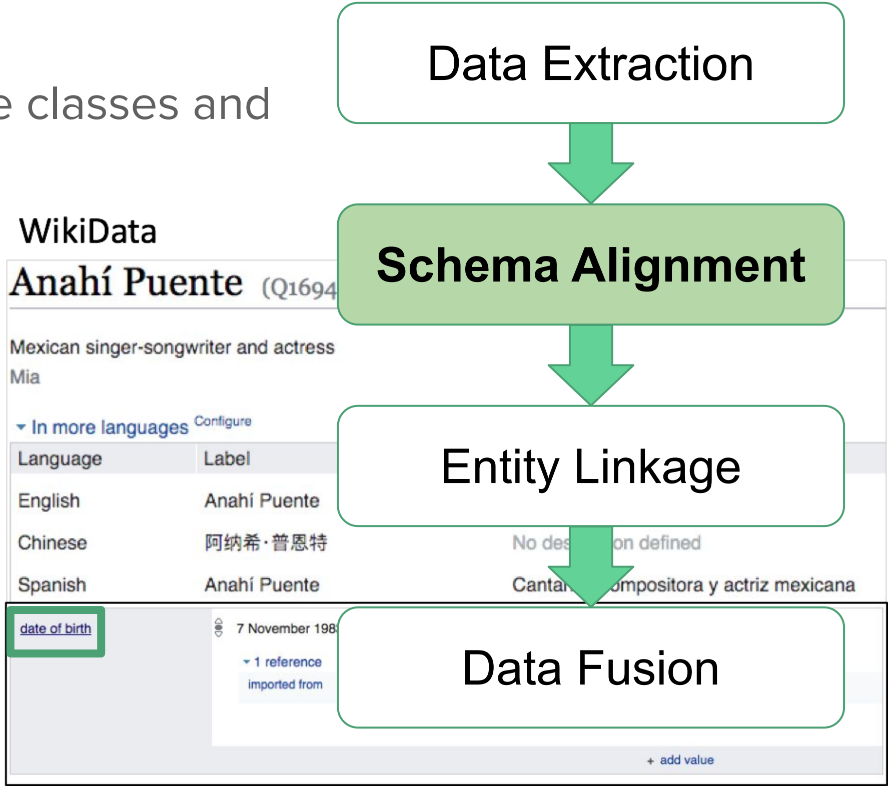
Data Fusion

Why is Data Integration Hard?

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

IMDB

WikiData

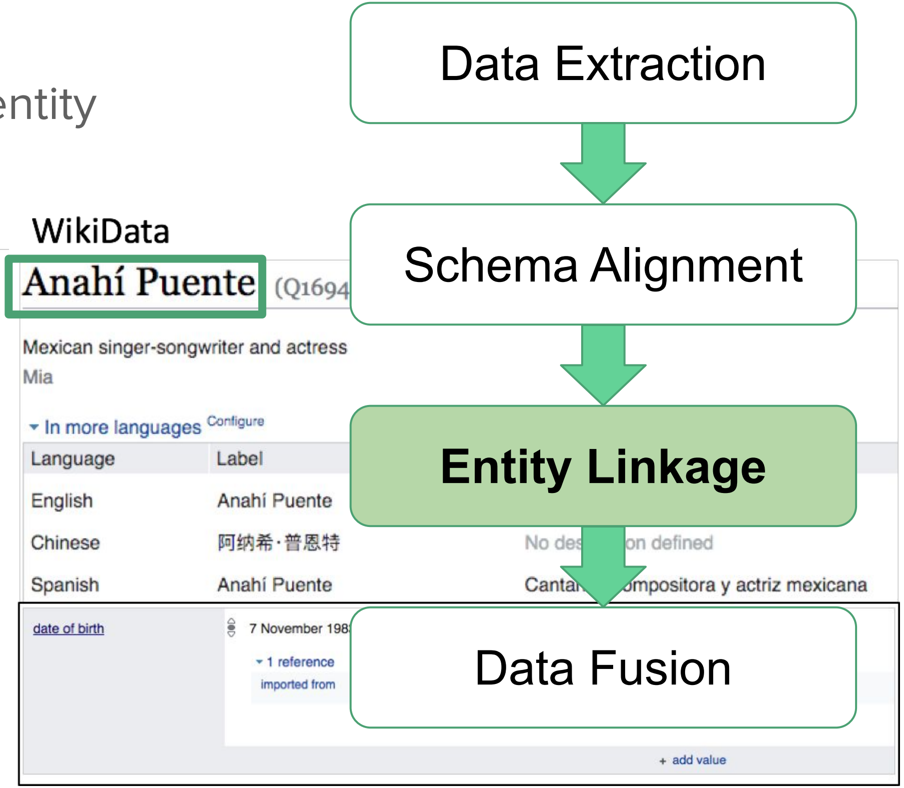


Why is Data Integration Hard?

- Heterogeneity everywhere
  - Different references to the same entity

IMDB

WikiData

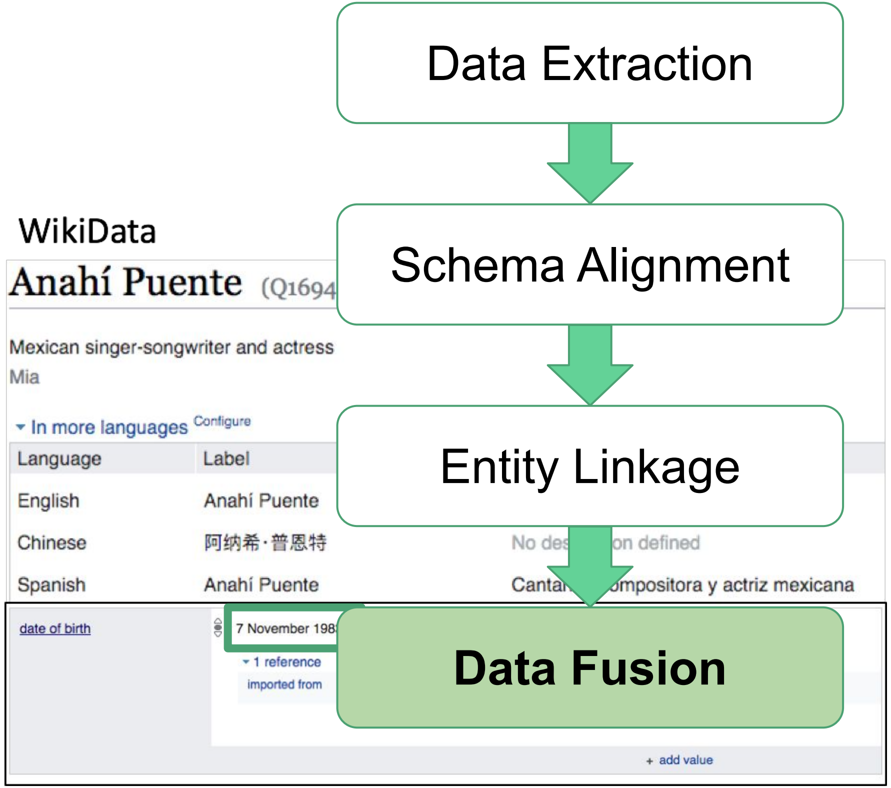


Why is Data Integration Hard?

- Heterogeneity everywhere
  - Conflicting values

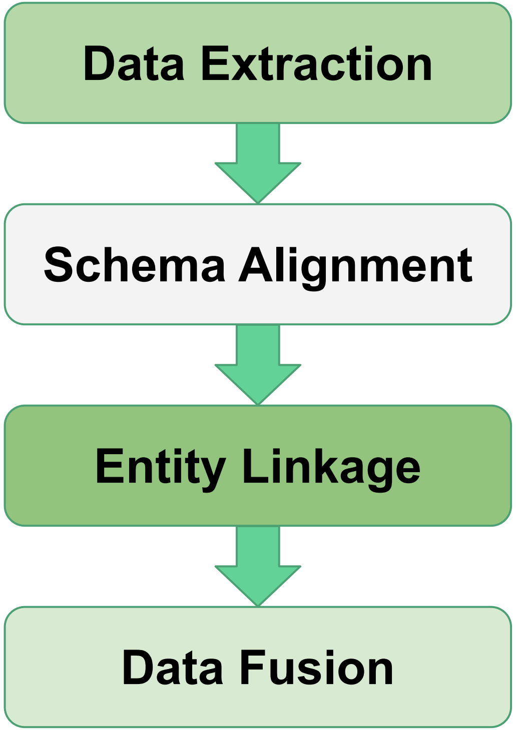
IMDB

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



What is Machine Learning?

- **Machine learning:** teach computers to ***learn*** with data, not by programming

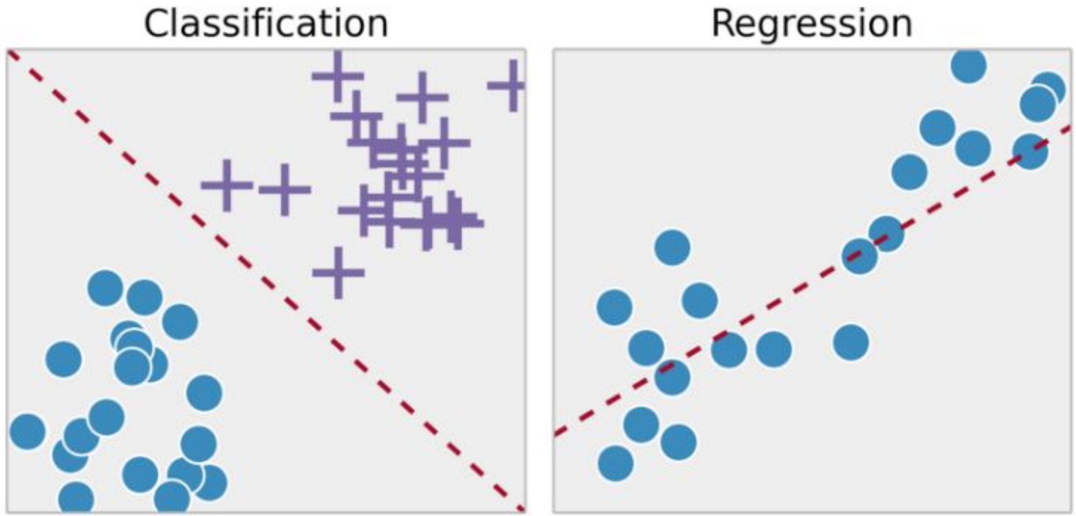
- **More Formal definition**

  A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, **improves with experience E**.

  -- Tom Mitchell

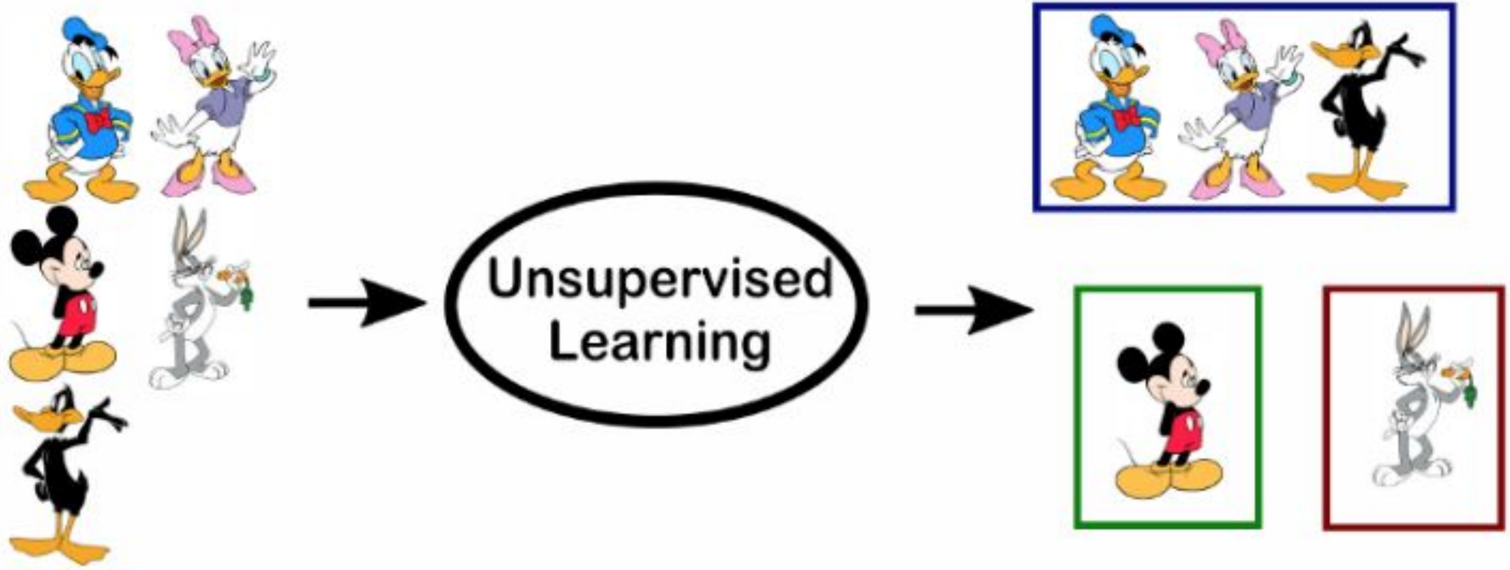
Two Main Types of Machine Learning

- Supervised learning: learn by examples



Two Main Types of Machine Learning

- Unsupervised learning: find structure w/o examples



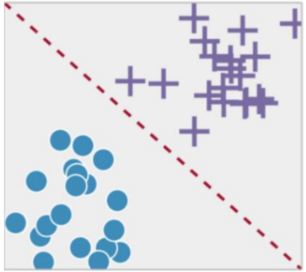
Two Main Types of Machine Learning

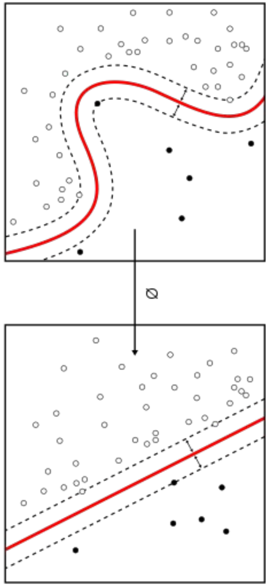
- Supervised learning: learn by examples
- Unsupervised learning: find structure w/o examples

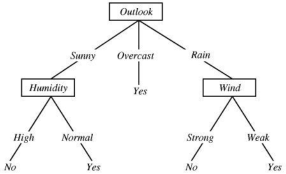
	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

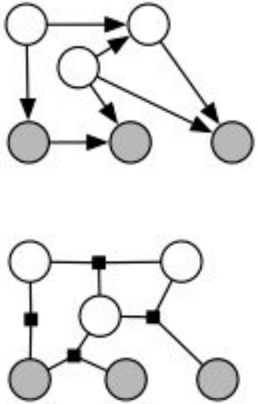
Techniques for Supervised ML

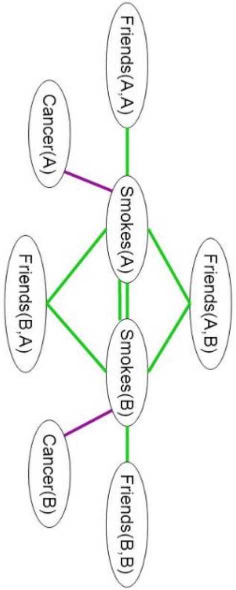
Hyperplanes	Kernel	Tree-based	Graphical Mdl	Logic Prog	Neural Netw
Linear/Logistic regression	SVM	Decision tree, Random forest	Bayes net, CRF	Pr soft logic, Markov logic net	ANN, CNN, RNN

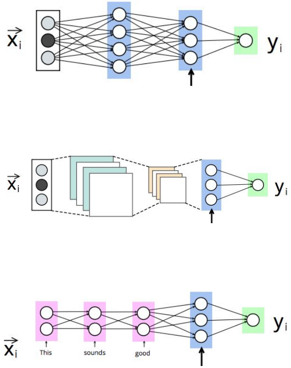












Key Lessons for ML [Domingos, 2012]

- Learning = Representation + Evaluation + Optimization
- **It's generalization that counts: generalize beyond training examples**
- Data alone is not enough: "no free lunch" theorem--No learner can beat random guessing over all possible functions to be learned
- Intuition fails in high dimensions: "curse of dimensionality"
- **More data beats a cleverer algorithm**: Google showed that after providing 300M images for DL image recognition, no flattening of the learning curve was observed.

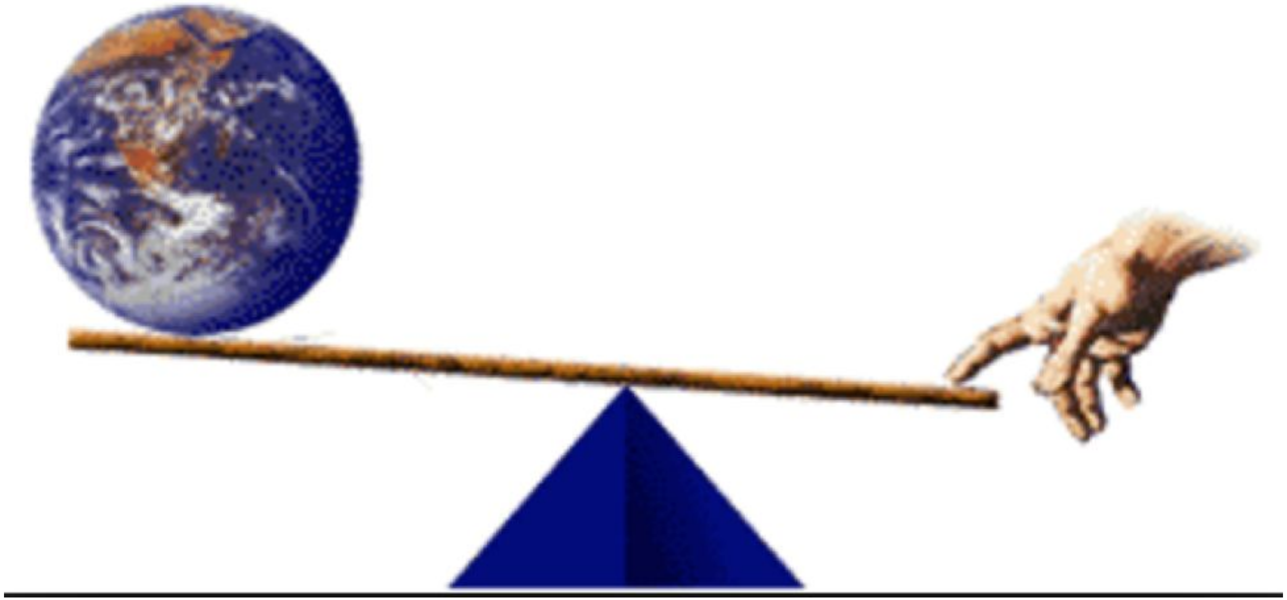
DI & ML as Synergy

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

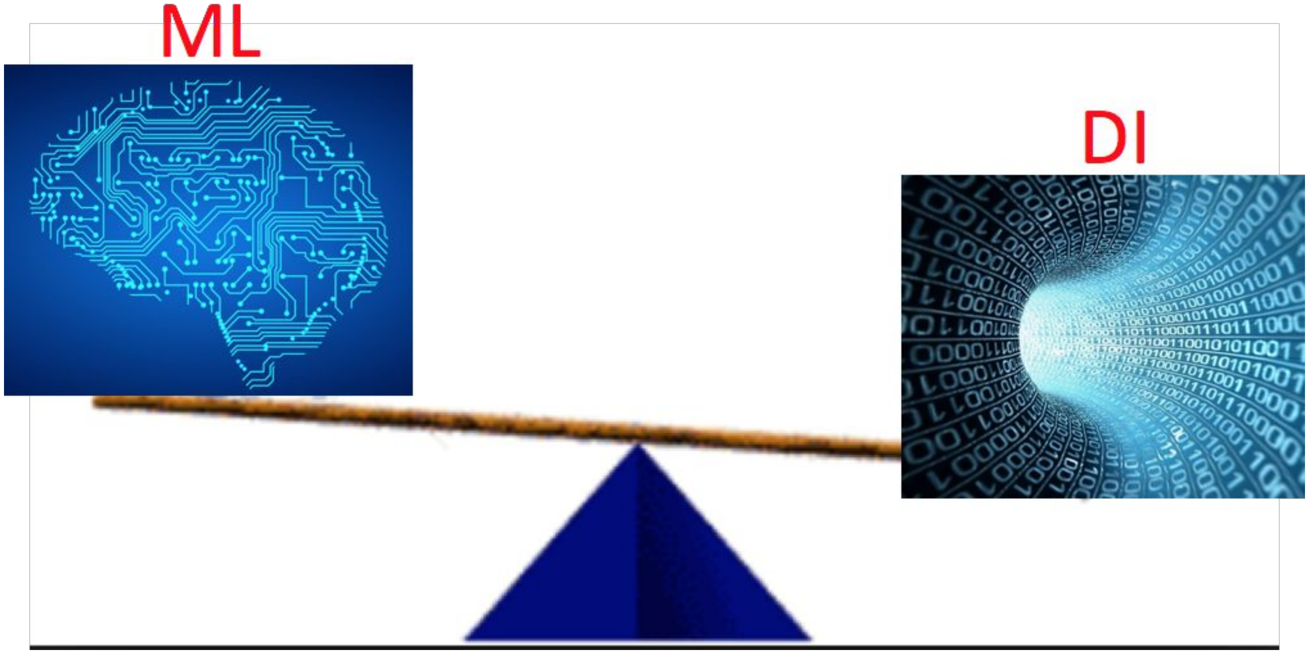
- **DI for effective ML: DATA, DATA, DATA**
  - Create large-scale training datasets from different sources
  - Cleaning of data used for training

Give me a Fulscrum, I will Move the Earth

-- Archimedes



Give me a DI funnel, I will Move ML



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

BigGorilla

snorkel

Dedupe.io

KNOWLEDGE VAULT

amperity

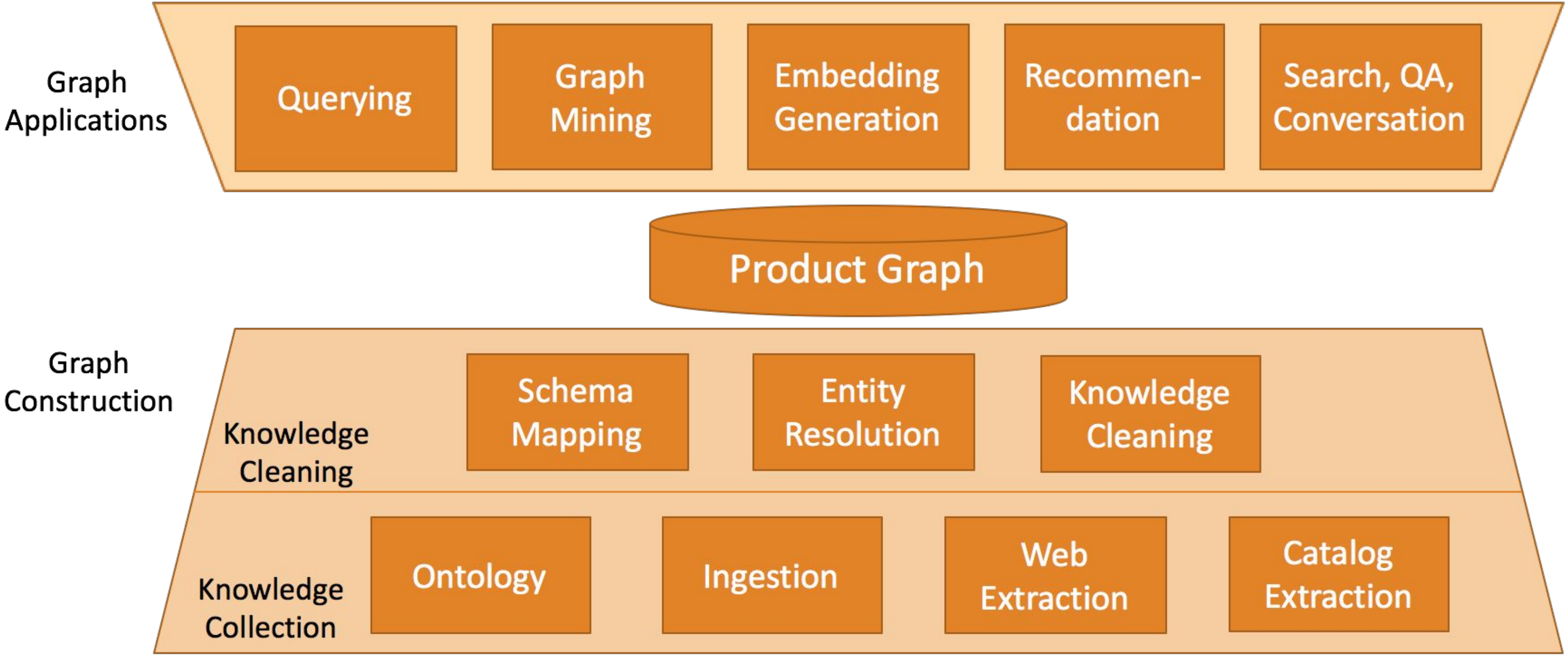
product graph

tamr

TRIFACTA

Increasing number of systems both in industry and academia.

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



Goal of This Tutorial

- **NO-GOALS**
  - Present a comprehensive literature review for all topics we are covering

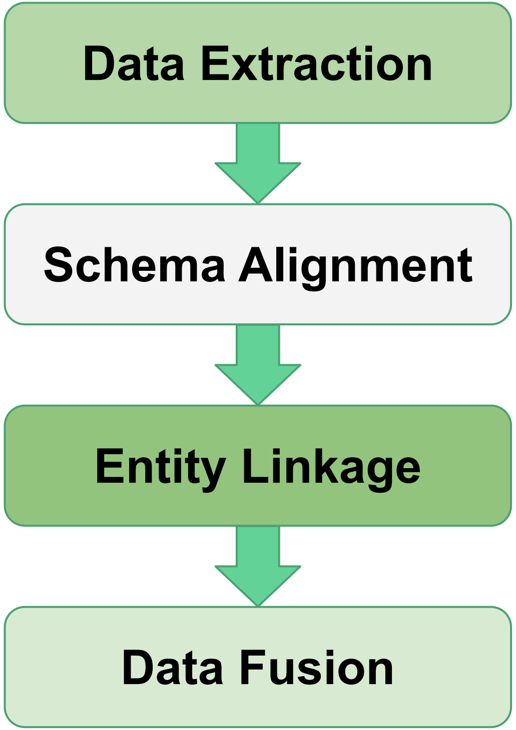
- **GOALS**
  - Present state-of-the-art for DI & ML synergy
  - Show how ML has been transforming DI and vice versa
  - Give some taste on which tool is working best for which tasks
  - Discuss what remains challenging

Outline

- Part I. Introduction
- Part II. ML for DI
- Part III. DI for ML
- Part IV. Conclusions and research directions

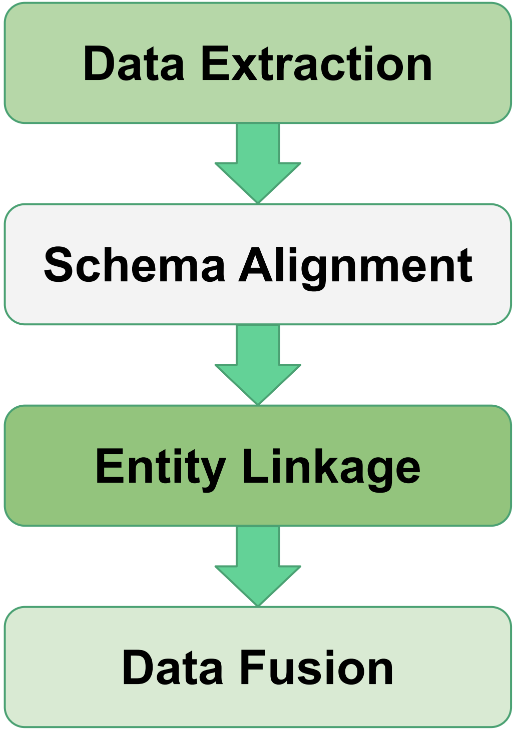
Data Integration Overview

- 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



Recipe

- Problem definition
- Brief history
- State-of-the-art ML solutions
- Summary w. a short answer

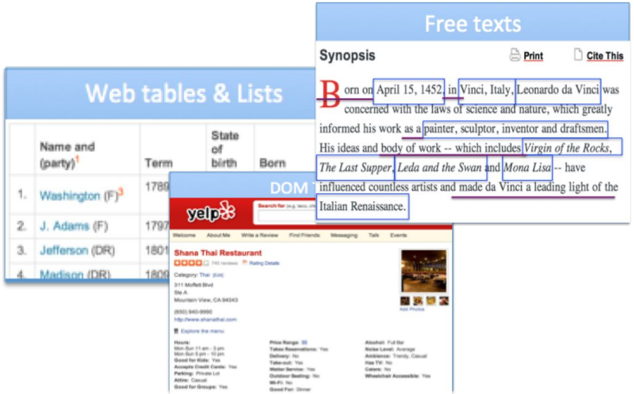


Theme I. Which ML Model Works Best?

Which ML Model Works Best?

ID	NAME	CLASS	MARK	SEX
1	John Deo	Four	75	female
2	Max Ruin	Three	85	male
3	Arnold	Three	55	male
4	Krish Star	Four	60	female
5	John Mike	Four	60	female
6	Alex John	Four	55	male
7	My John Rob	Fifth	78	male
8	Asruid	Five	85	male
9	Tes Qry	Six	78	male
10	Big John	Four	55	female

Tree-based models



??

SCENE FROM "DAN'L DRUCE."

This interesting domestic drama, by Mr. W. S. Gilbert, has continued to engage the sympathies of a nightly sufficient audience at the Haymarket Theatre, where it has now been represented more than sixty times. Its subject and character were described by us, in the ordinary report of theatrical novelties, about two months ago. Our readers will probably not need to be reminded that the hero of the story, Dan'l Druce, the blacksmith, is a solitary recluse dwelling on the coast of Norfolk, where his lone cottage is visited by fugitives from party vengeance during the civil wars of the Commonwealth. His hoard of money is stolen; but a different sort of treasure, a helpless female infant, is left by some mysterious agency, and may be accepted, as in George Eliot's tale of "Silas Marner," for a Divine gift to the sad-hearted misanthrope, far better than riches. In this spirit, at least, he is content to receive the precious human charge; and so to those who would remove it from his home, Dan'l Druce here makes answer with the solemn exclamation, "Touch not the Lord's gift!" This character is well acted by Mr. Hermann Vezin.

Neural network

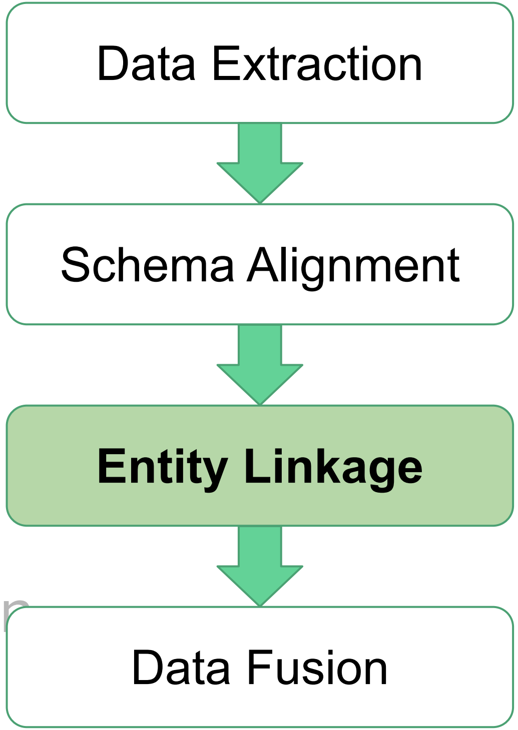
Theme II. Does Supervised Learning Apply to DI?

- **Supervised learning has made a big splash recently in many fields**

- **However, it is hard to bluntly apply supervised learning to DI tasks**
  - Our goal is to integrate data from many different data sources in different domains
  - The different sources present different data features and distributions
  - Collecting training labels for each source is a huge cost

Outline

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



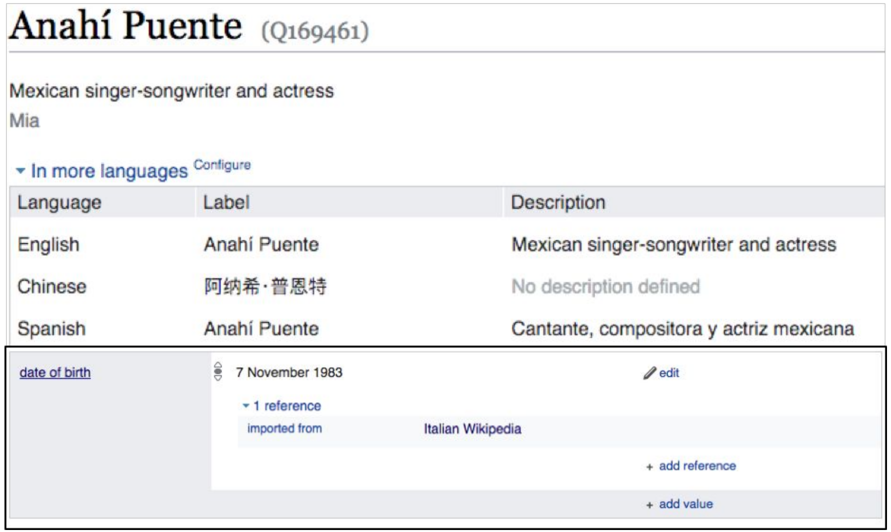
What is Entity Linkage?

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

Are they the same entity?

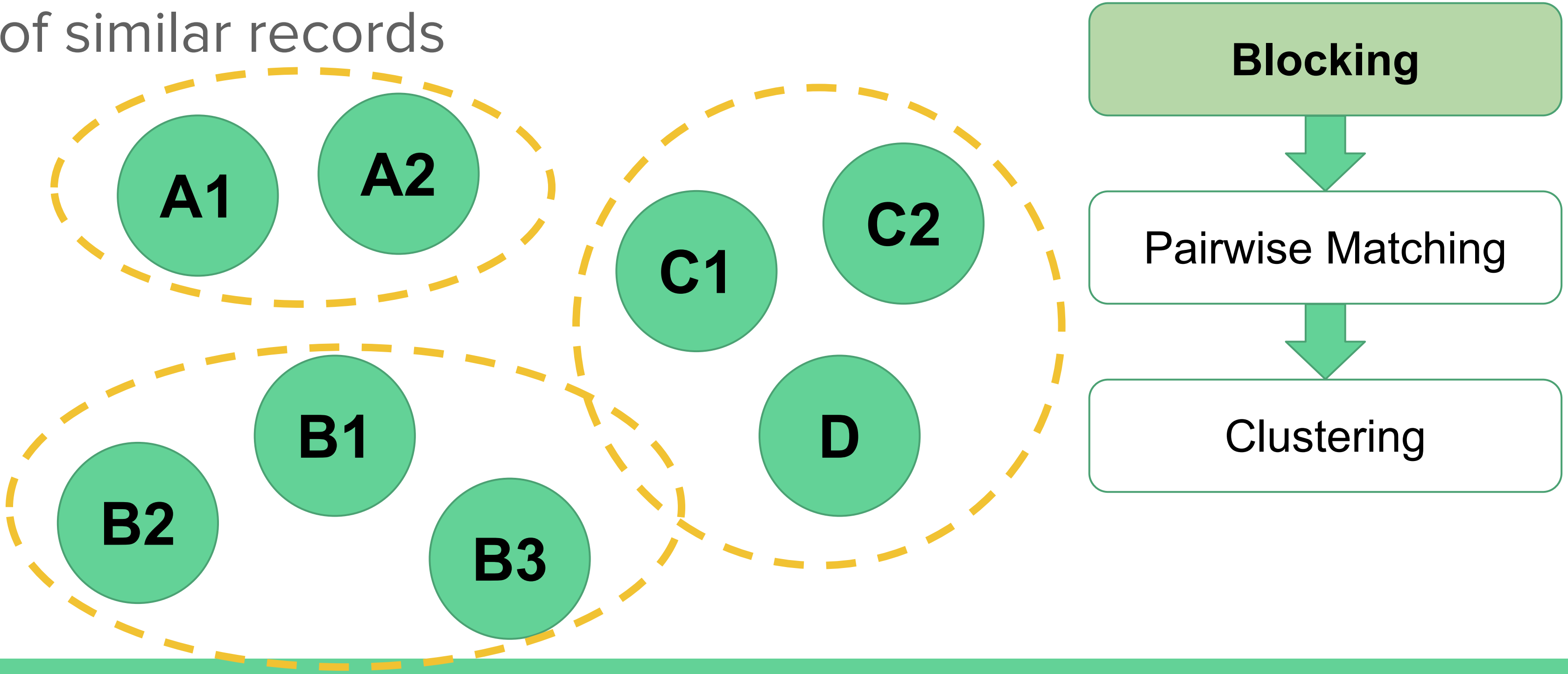
IMDB

WikiData



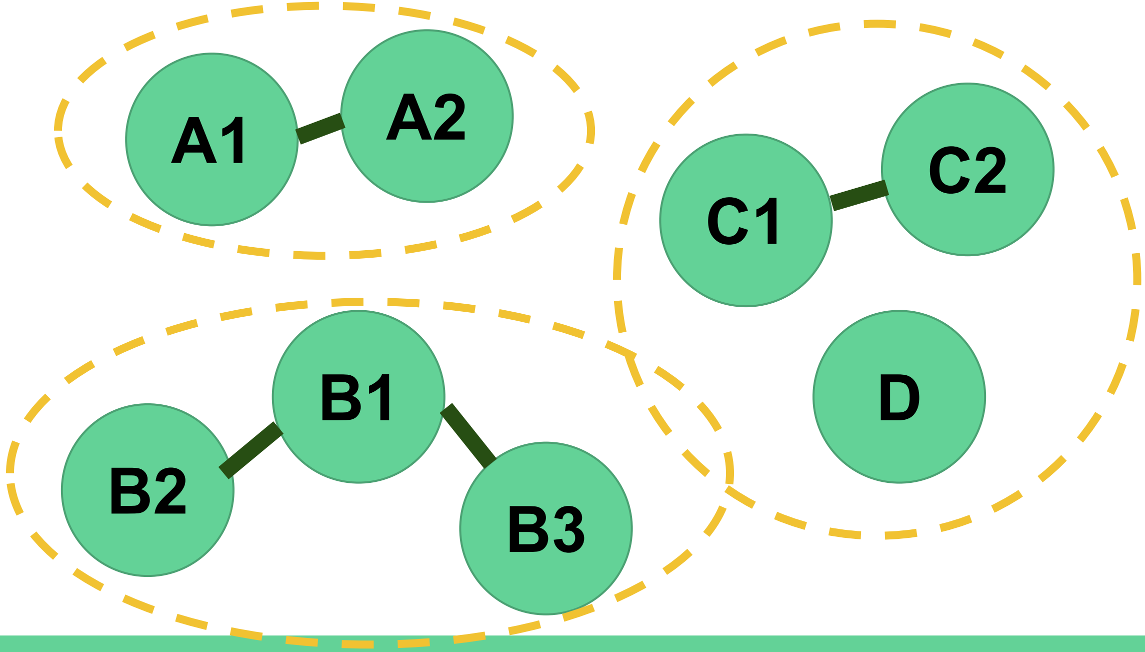
Quick Tour for Entity Linkage

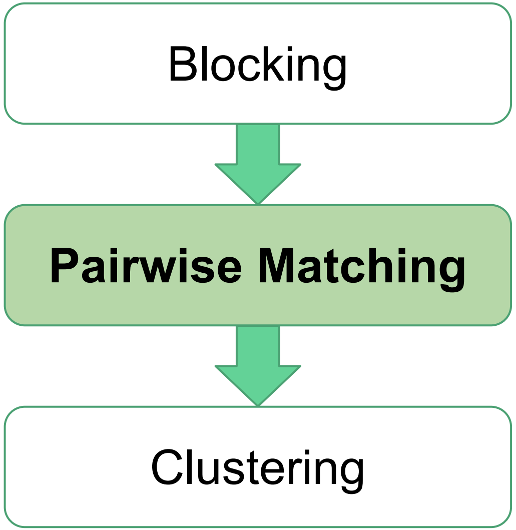
- **Blocking**: efficiently create small blocks of similar records



Quick Tour for Entity Linkage

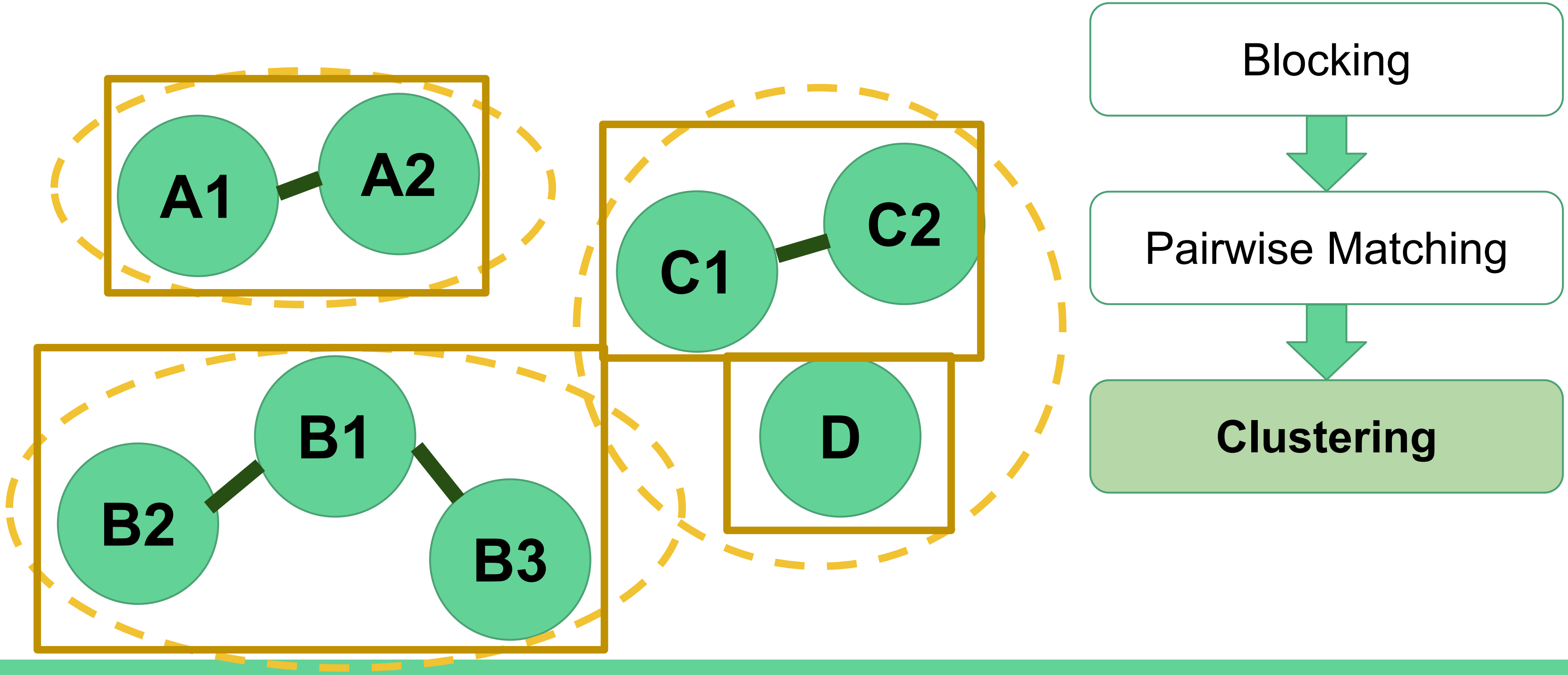
- **Pairwise matching**: compare all record pairs in a block





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') > θ_h
  - Unmatch: sim(r, r') < θ_l
  - Possible match: θ_l < sim(r, r') < θ_h

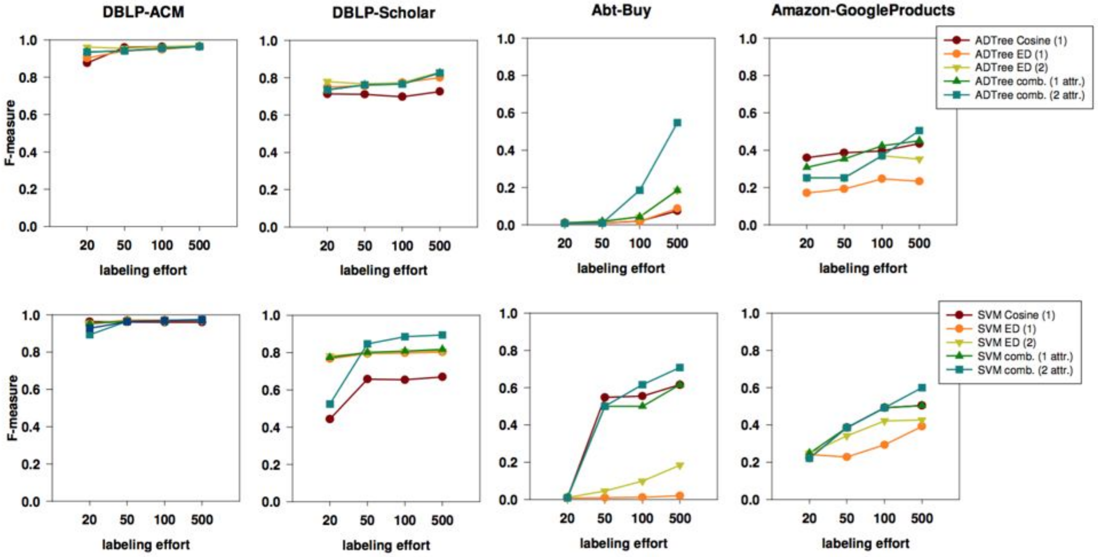
Early ML Models

- [Köpcke et al, VLDB'10]

~2000 (Early ML)

Sup / Unsup learning

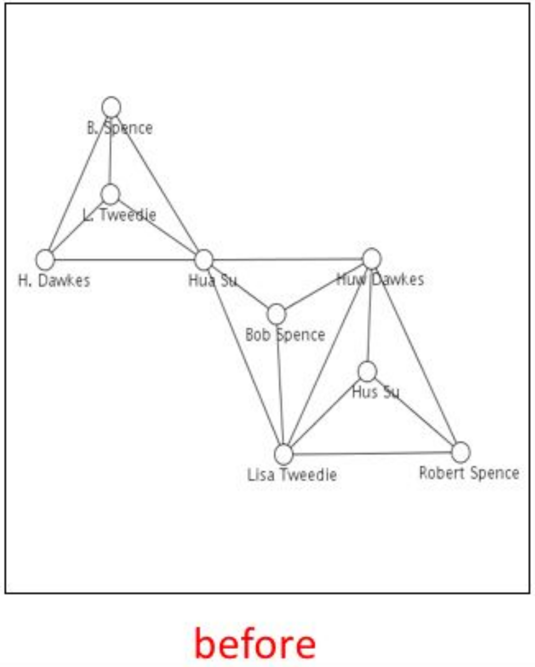
- Matching: Decision tree, SVM
  F-msr: 70%-90% w. 500 labels
- Clustering: Correlation clustering, Markov clustering

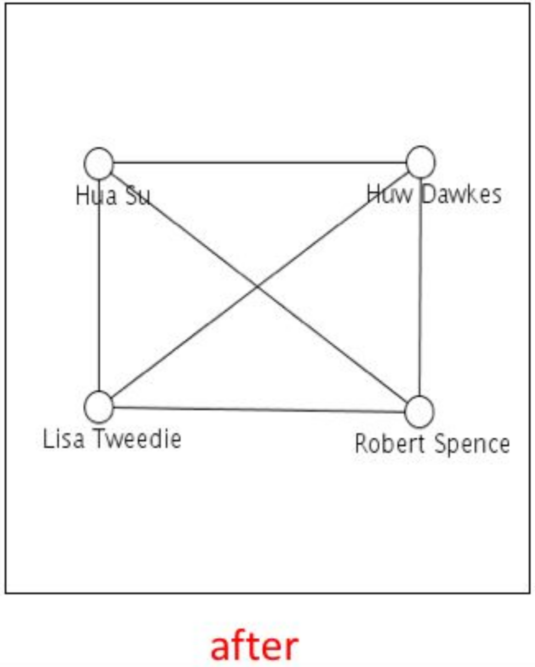


Collective Entity Resolution: Beyond Pairs

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

Out of the scope of this tutorial. For details: See tutorial by Getoor and Machanavajjhala, KDD, 2013.





[Example by Getoor and Machanavajjhala]

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

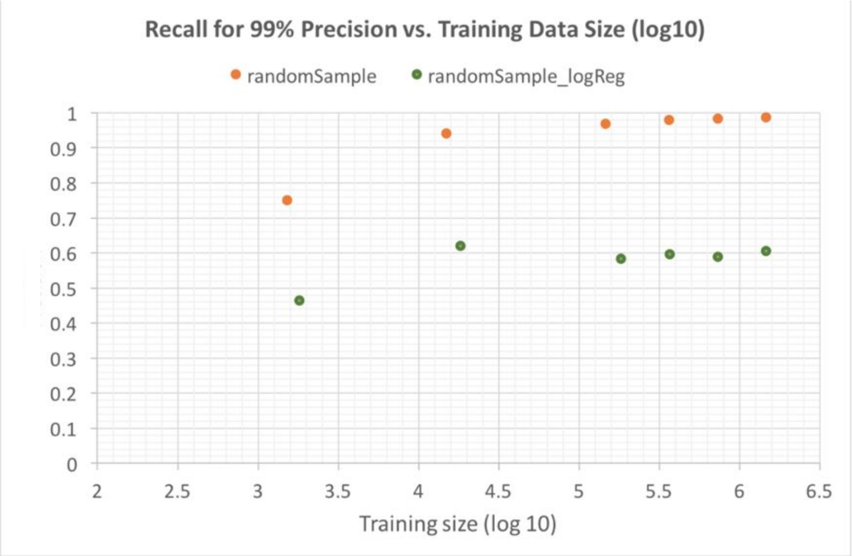
State-of-the-Art ML Models [Dong, KDD'18]

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

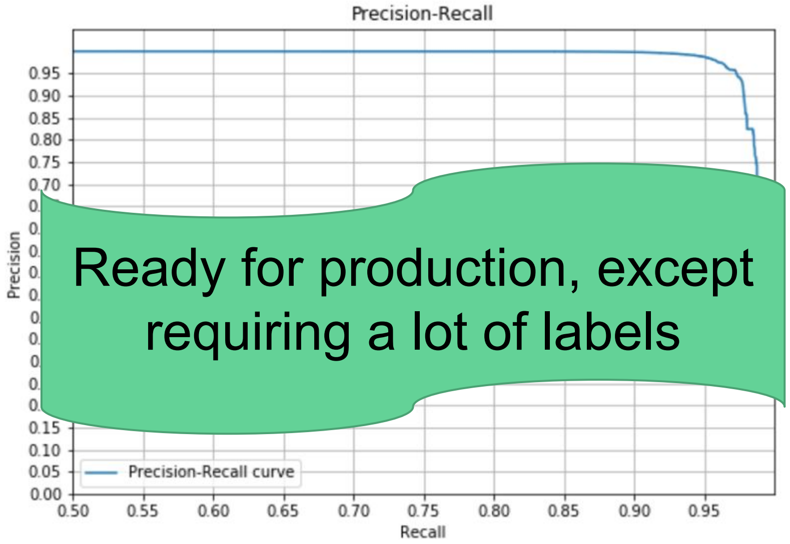
State-of-the-Art ML Models [Dong, KDD’18]

Supervised learning

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

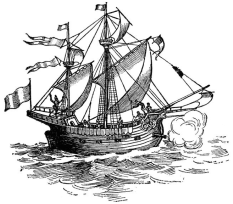
~2015 (ML)

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



Ready for production, except requiring a lot of labels

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)

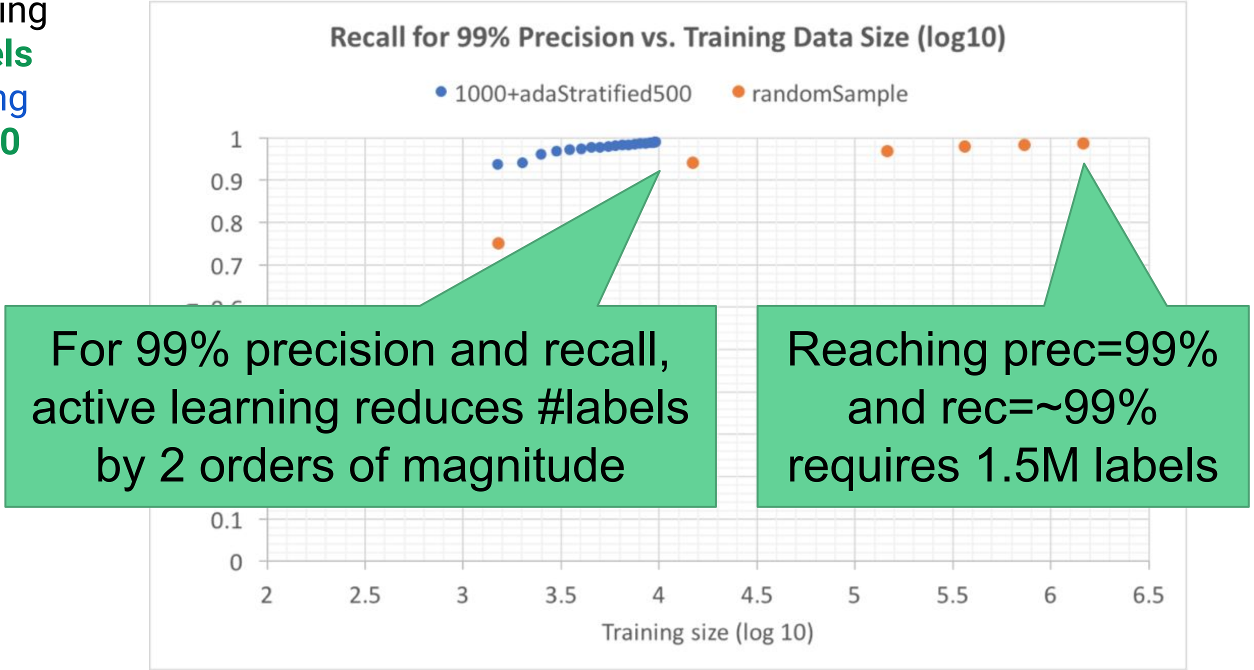
State-of-the-Art ML Models [Dong, KDD'18]

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]

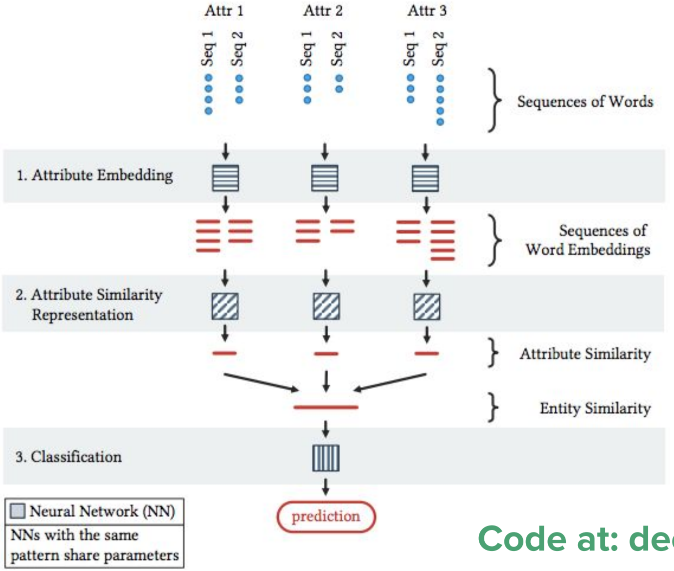
Magellan

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

2018 (Deep ML)

Deep learning

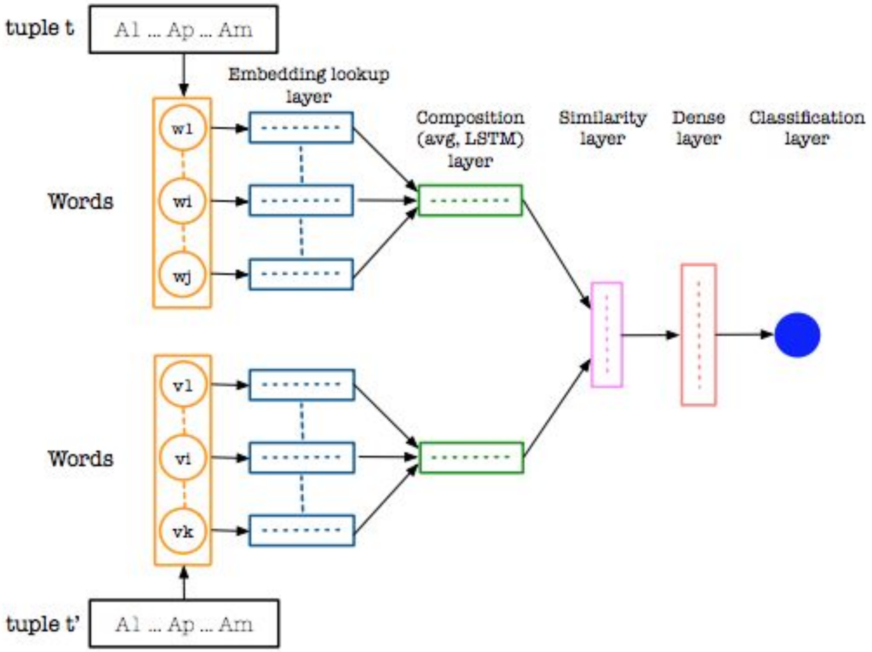
- Deep learning
- Entity embedding



Code at: deepmatcher.ml

Deep Learning Models [Ebraheem et al., VLDB'18]

- Embedding on entities
- Outperforming existing solution



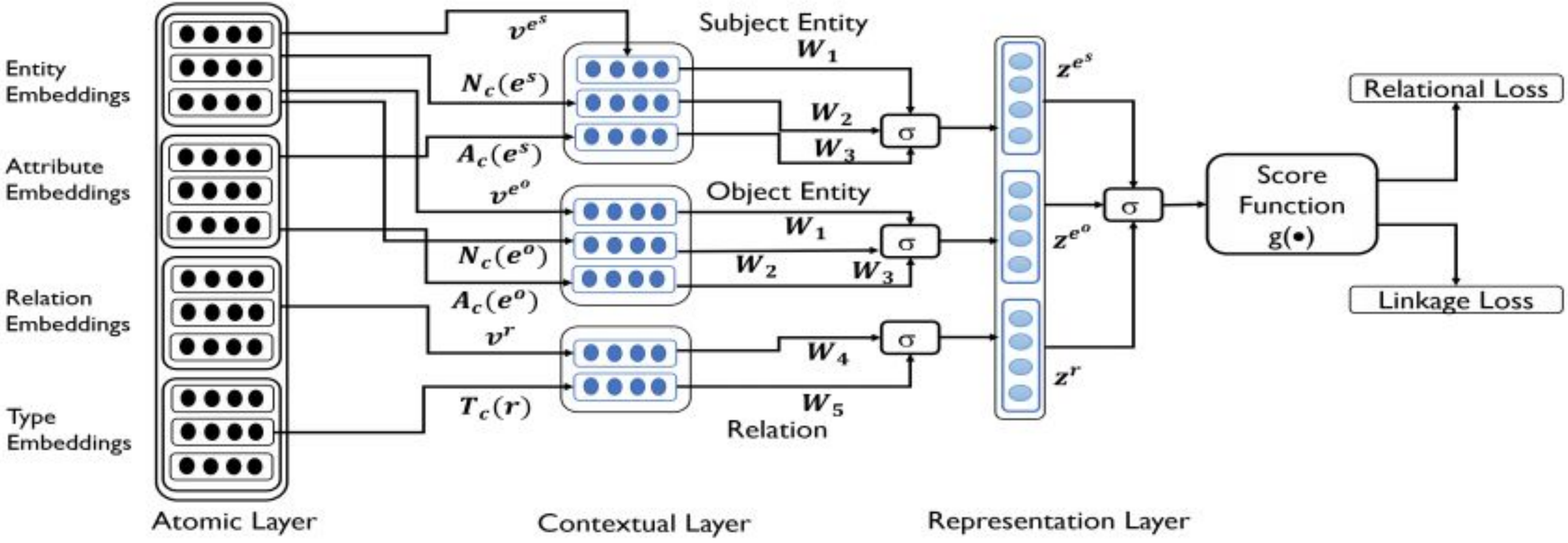
2018 (Deep ML)

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)

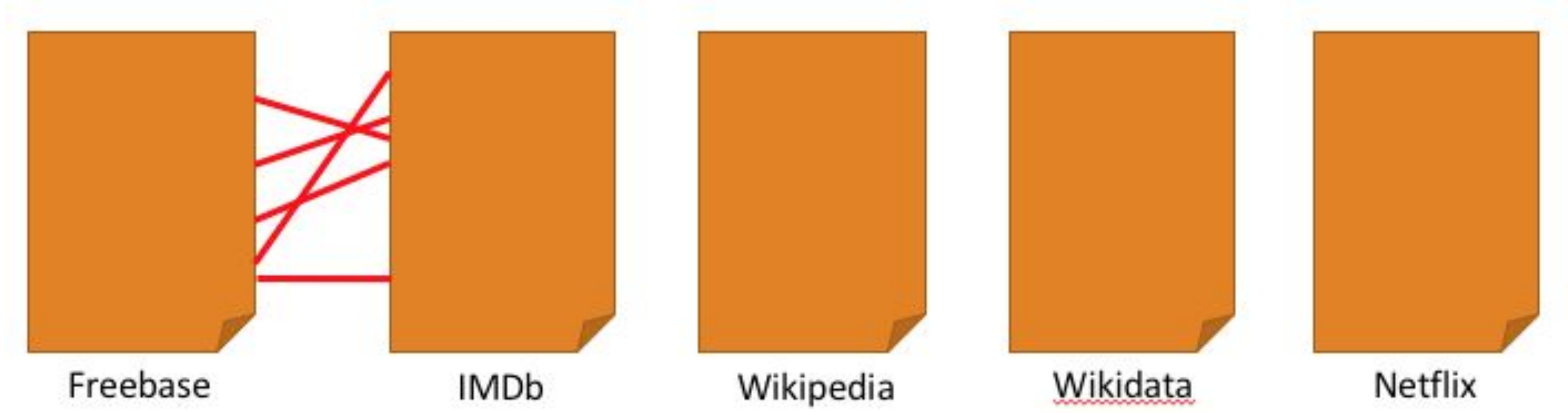
Deep learning

- 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

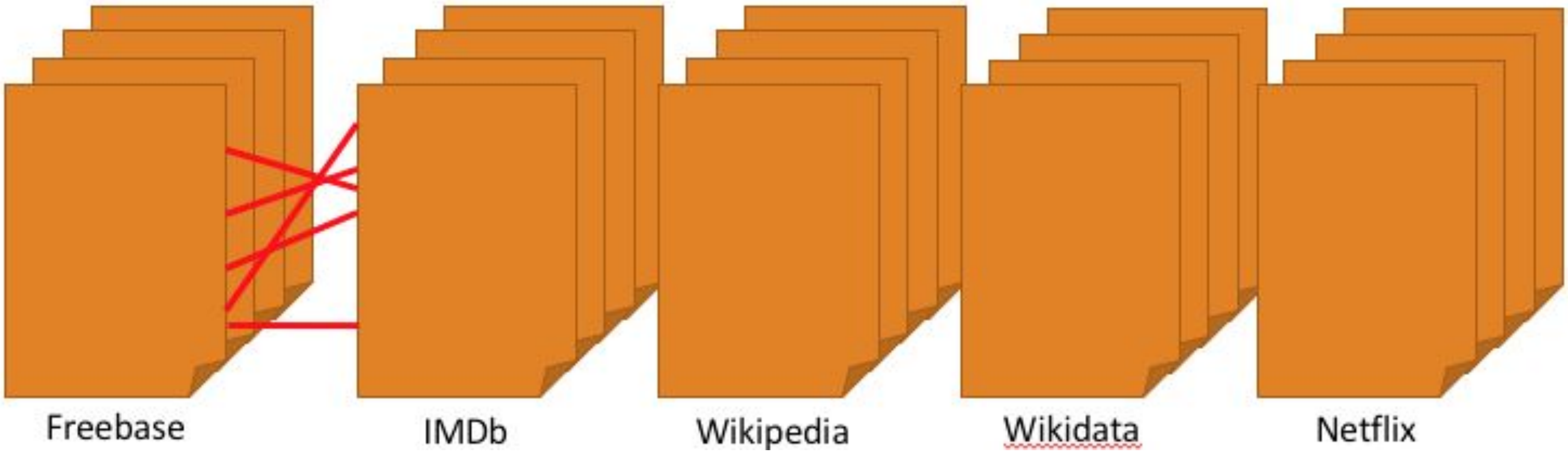
Challenges in Applying ML on EL

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



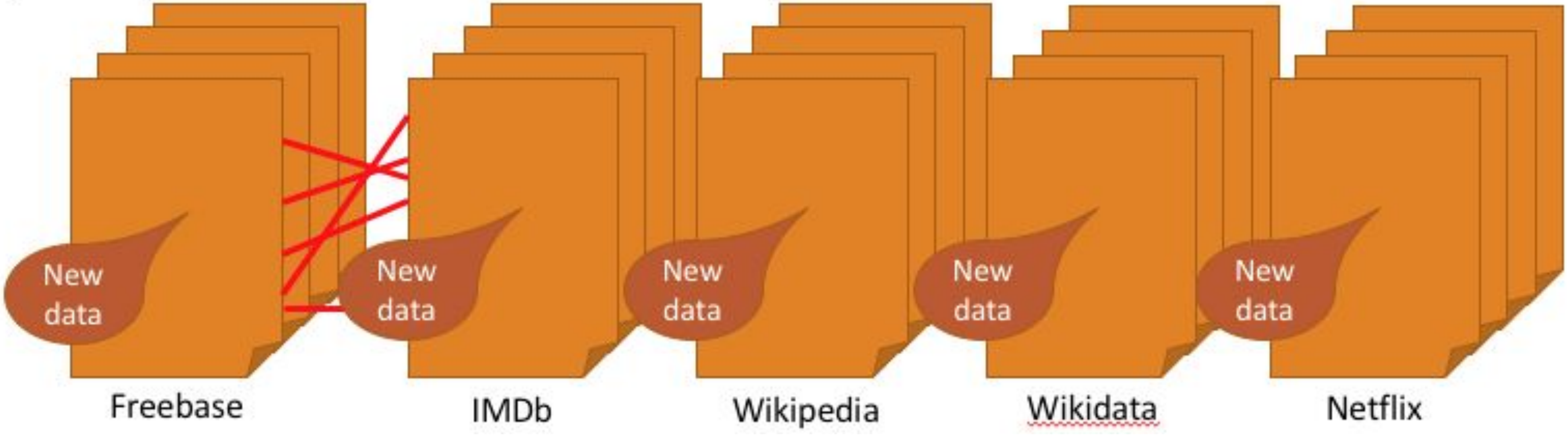
Challenges in Applying ML on EL

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



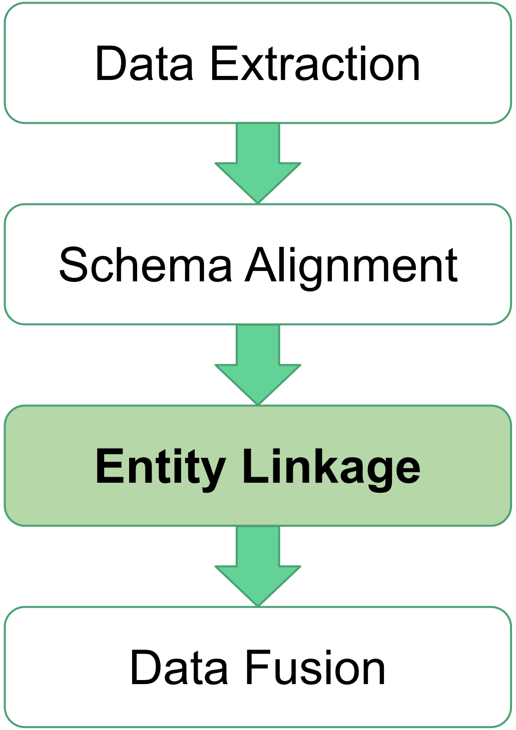
Challenges in Applying ML on EL

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



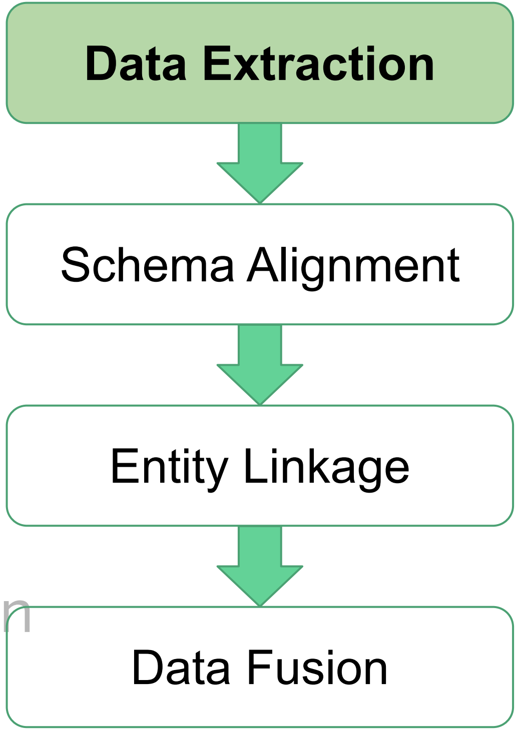
Recipe for Entity Linkage

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



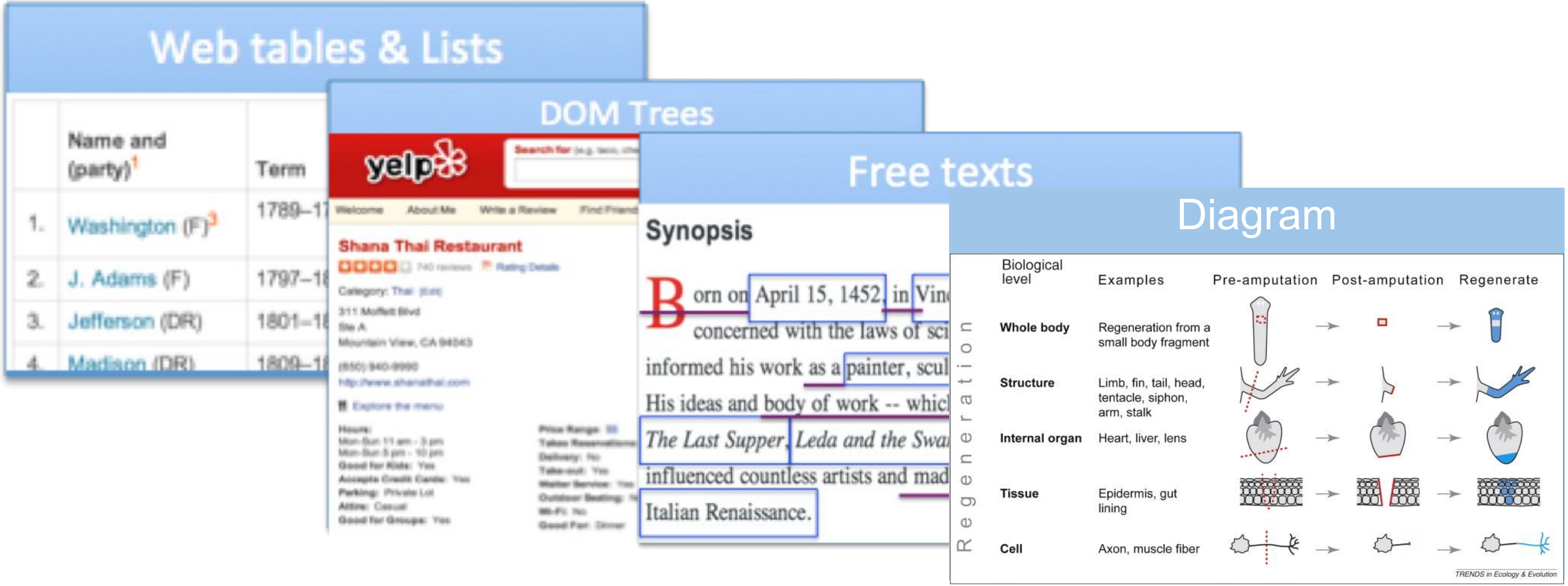
Outline

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



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

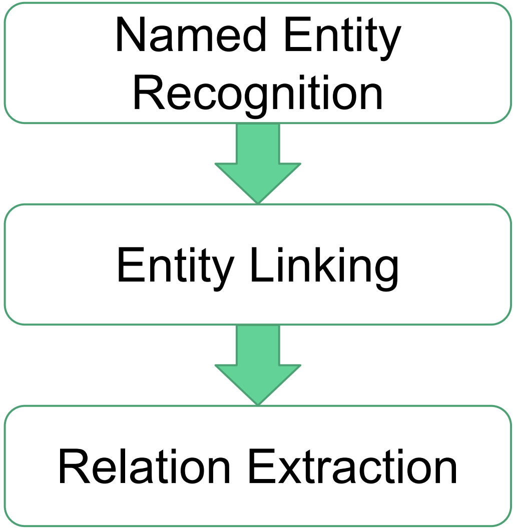
2013 (Deep ML)

Deep learning

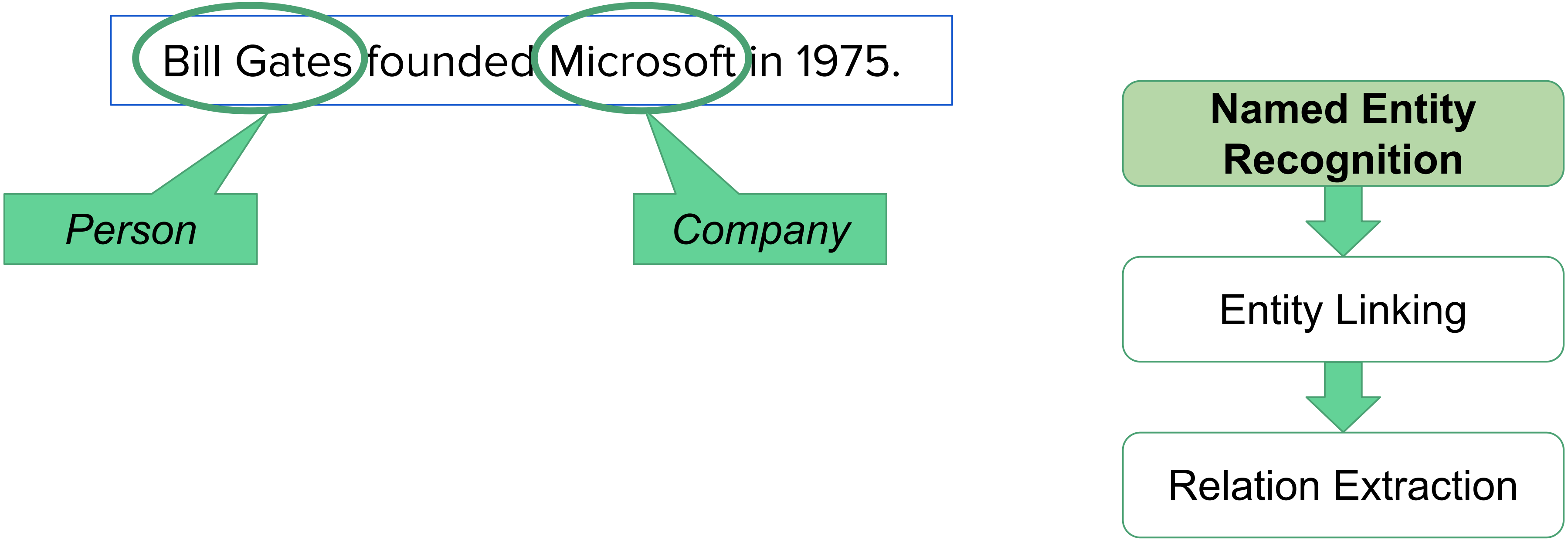
- Use RNN, CNN, attention for RE
- Data programming / Heterogeneous learning
- Revisit DOM extraction

Extraction from Texts: Quick Tour

Bill Gates founded Microsoft in 1975.



Extraction from Texts: Quick Tour



Extraction from Texts: Quick Tour

Bill Gates founded Microsoft in 1975.

Named Entity Recognition

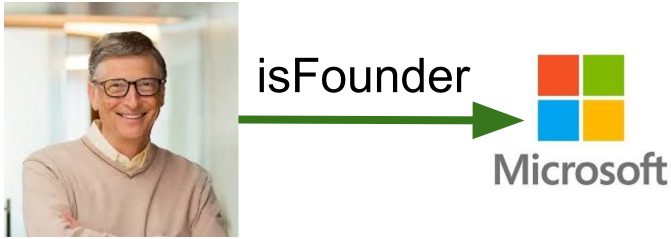
Entity Linking

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 Tour

Bill Gates founded Microsoft in 1975.



We focus on Relation Extraction in the rest of the tutorial.

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

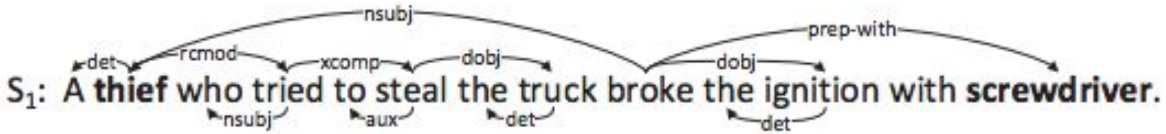
- **Models**
  - SVM (Support Vector Machine)
- **Kernels**
  - Subsequence
  - Dependency tree
  - **Shortest dependency path**
  - Convolution dependency

Extraction from Texts: Kernel Based [Mengqiu Wang, IJCNLP'08]

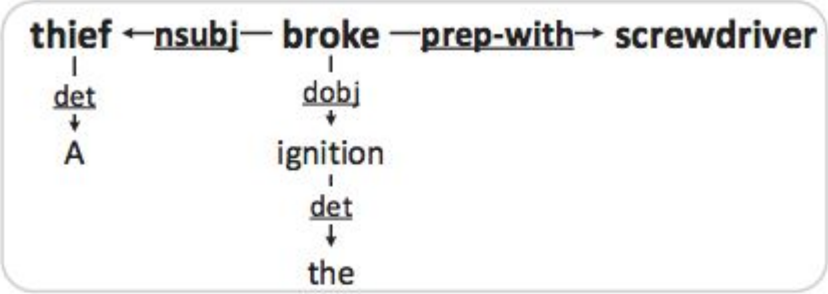
~2005 (Rel. Ex.)

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

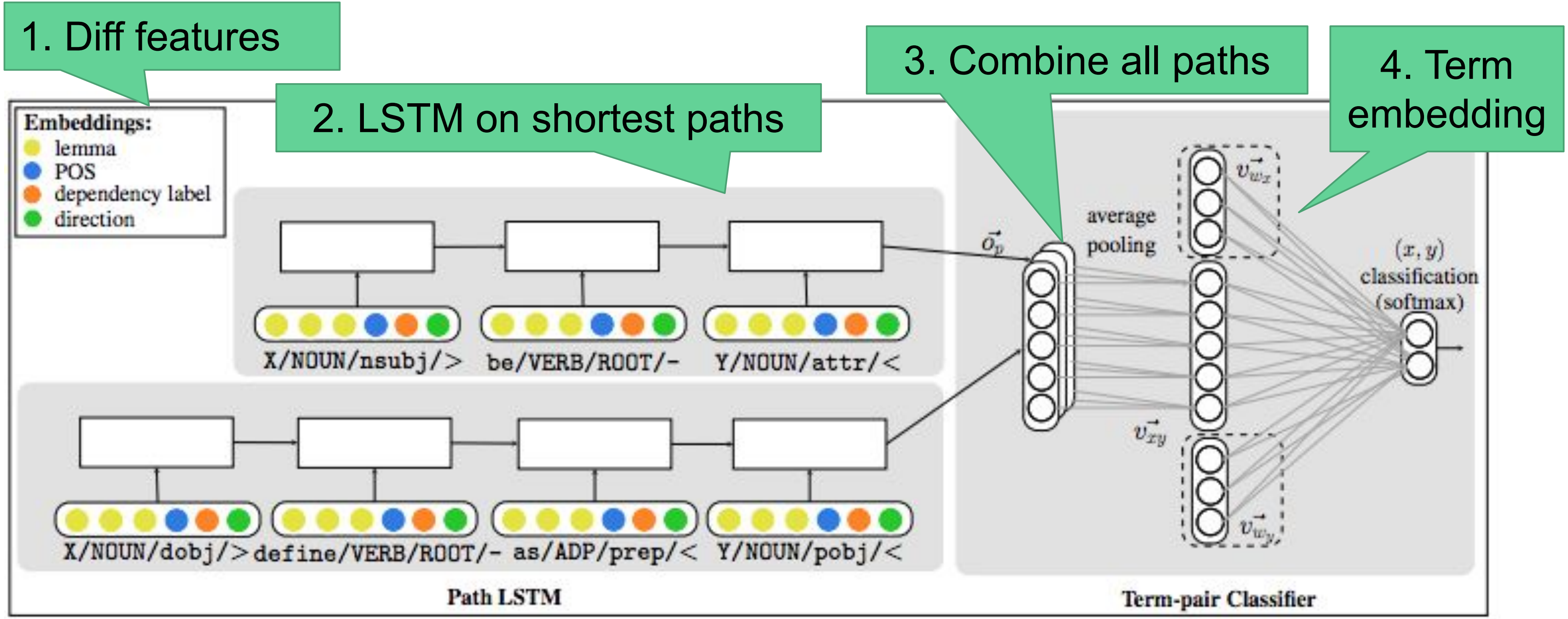
2013 (Deep ML)

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

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
- **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

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

[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

[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

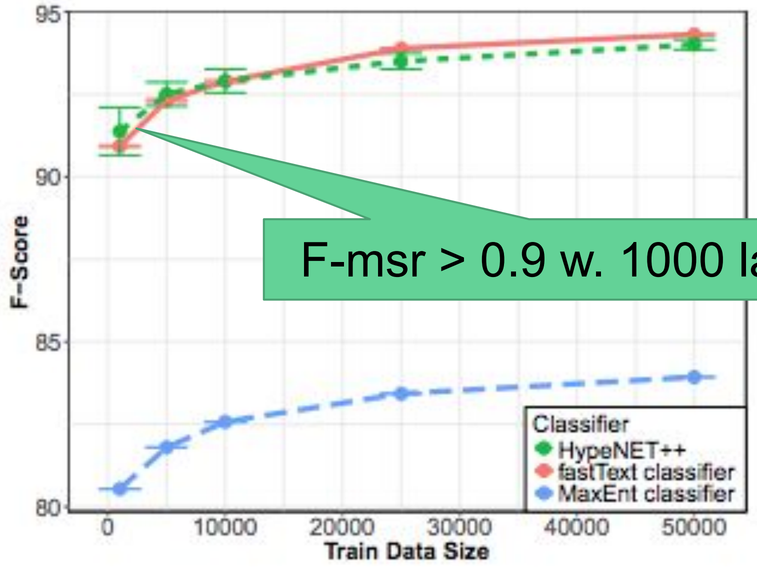
Where are training labels from?

~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

Where are training labels from?

2013 (Deep ML)

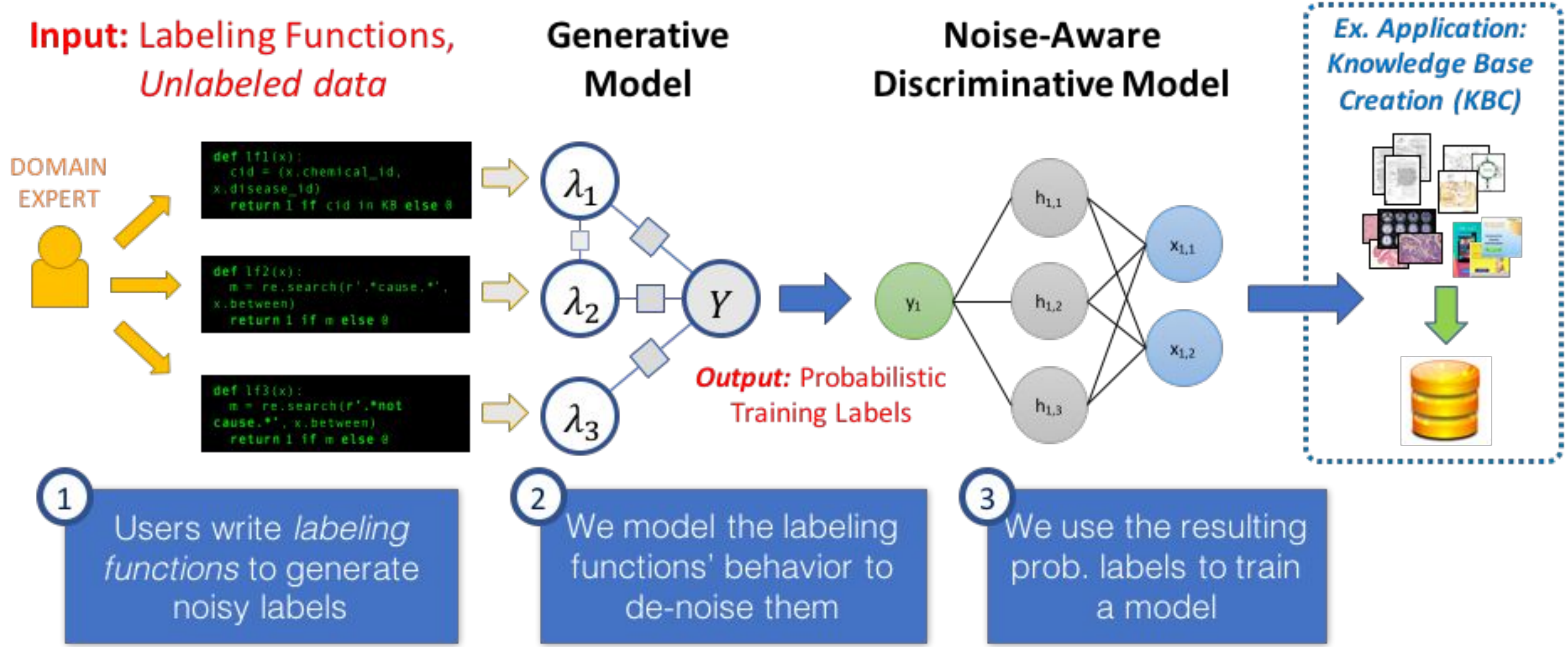
Deep learning

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

Will cover in “DI for ML”

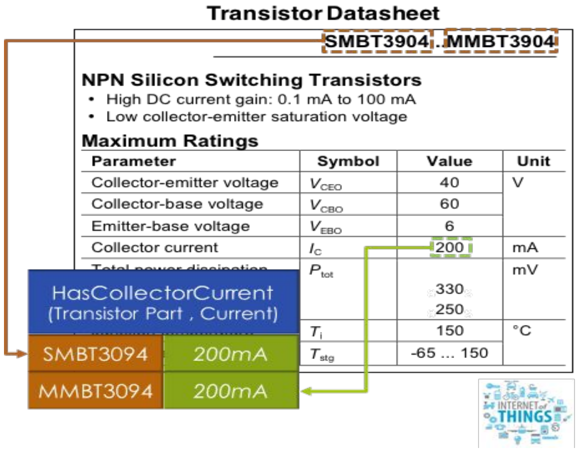
- **Semi-supervised learning**
  - Iterative extraction [Carlson et al., AAAI’10]
    Use new extractions to retrain models
    E.g., NELL
- **Weak learning**
  - Distant supervision [Mintz et al., ACL’09]
    Rule-based annotation with seed data
    E.g., DeepDive, Knowledge Vault
  - Data programming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

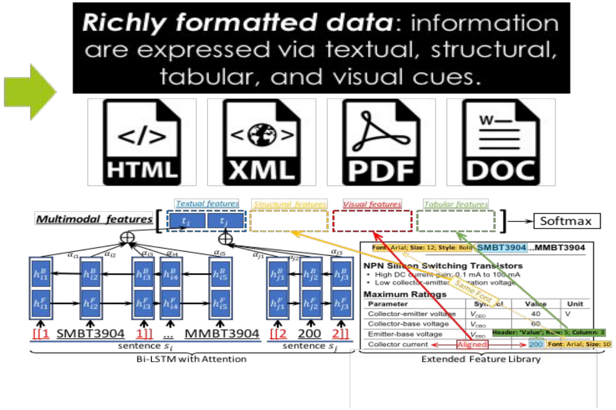
Snorkel: Code as Supervision [Ratner et al., NIPS'16, VLDB'18]



[Slide by Alex Ratner]

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

OpenIE from Texts

Where are predicates from?

- **ClosedIE**
  - Only extracting facts corresponding to ontology
  - Normalize predicates by ontology
  - E.g., (Bill Gates, /person/isFounder, Microsoft)

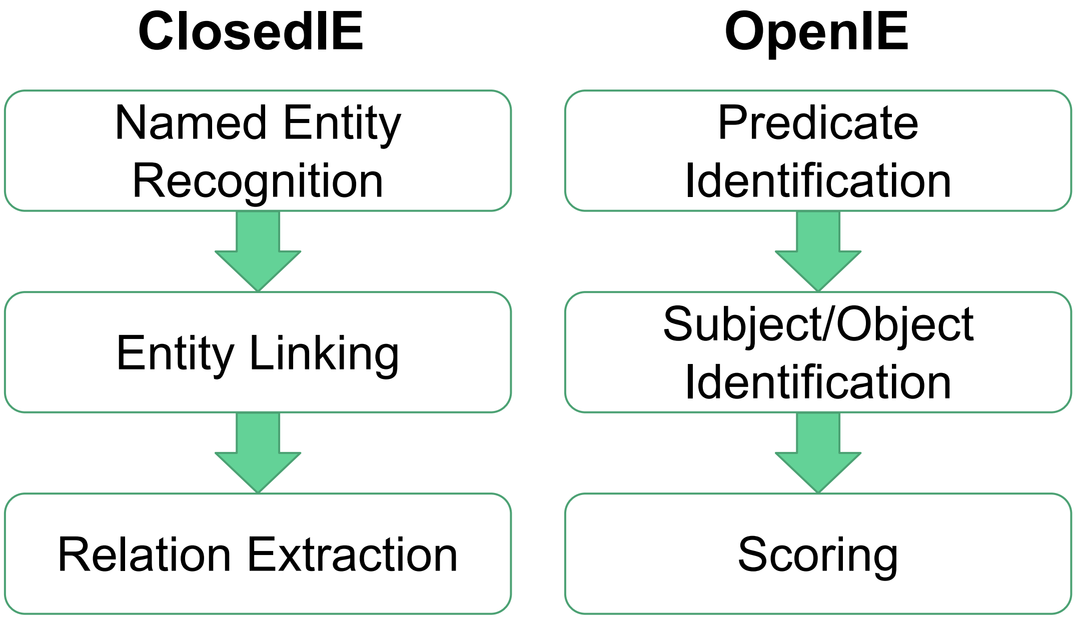
~2005 (Rel. Ex.)

Relation extraction from texts
- NER→EL→RE
  - Feature based: LR, SVM
  - Kernel based: SVM
- Distant supervision
- OpenIE

Bill Gates founded Microsoft in 1975.

- **OpenIE** [Banko et al., IJCAI'07]
  - Extract all relations expressed in texts
  - Predicates are unnormalized strings
  - E.g., ("Bill Gates", "founded", "Microsoft")

OpenIE from Texts [Etzioni et al., IJCAI'11]



Bill Gates founded Microsoft in 1975.

OpenIE from Texts [Etzioni et al., IJCAI'11]

ClosedIE

Named Entity Recognition

↓

Entity Linking

↓

Relation Extraction

OpenIE

Predicate Identification

↓

Subject/Object Identification

↓

Scoring

Bill Gates founded Microsoft in 1975.

- Predicate: longest sequence of words as light verb construction

OpenIE from Texts [Etzioni et al., IJCAI'11]

ClosedIE

Named Entity Recognition

Entity Linking

Relation Extraction

OpenIE

Predicate Identification

Subject/Object Identification

Scoring

Bill Gates founded Microsoft in 1975.

- Predicate: longest sequence of words as light verb construction
- Subject: learn left and right boundary
- Object: learn right boundary

OpenIE from Texts [Etzioni et al., IJCAI'11]

ClosedIE

Named Entity Recognition

↓

Entity Linking

↓

Relation Extraction

OpenIE

Predicate Identification

↓

Subject/Object Identification

↓

Scoring

Bill Gates founded Microsoft in 1975.

- Predicate: longest sequence of words as light verb construction
- Subject: learn left and right boundary
- Object: learn right boundary
- LR for triple confidence

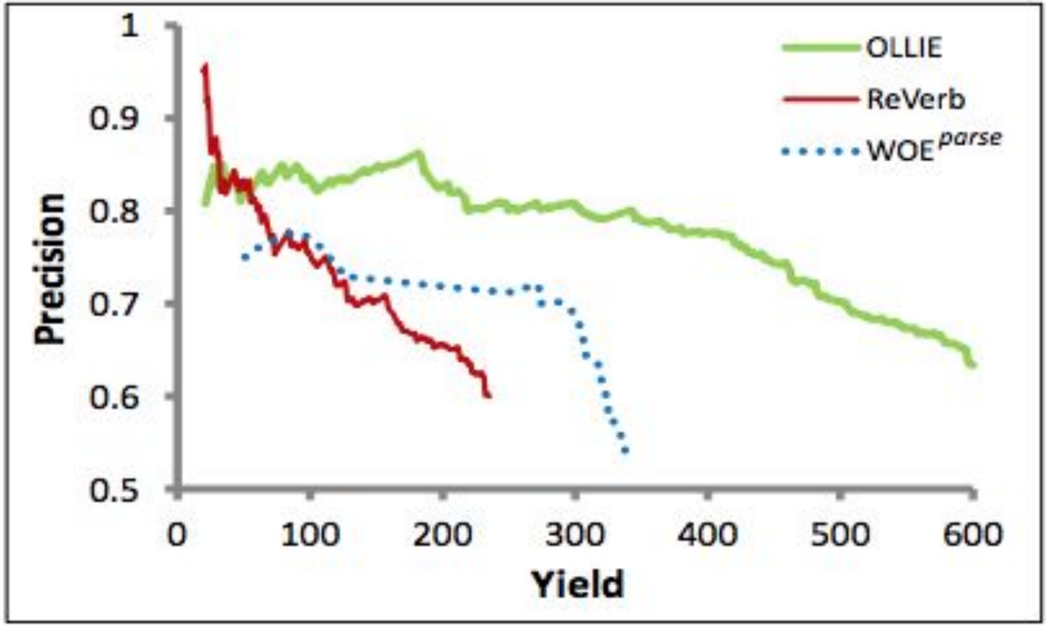
OpenIE from Texts [Mausam et al., EMNLP’12]

Where are predicates from?

~2005 (Rel. Ex.)

Relation extraction from texts

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



Extraction from Semi-Structured Data

Extraction from semi-structured data

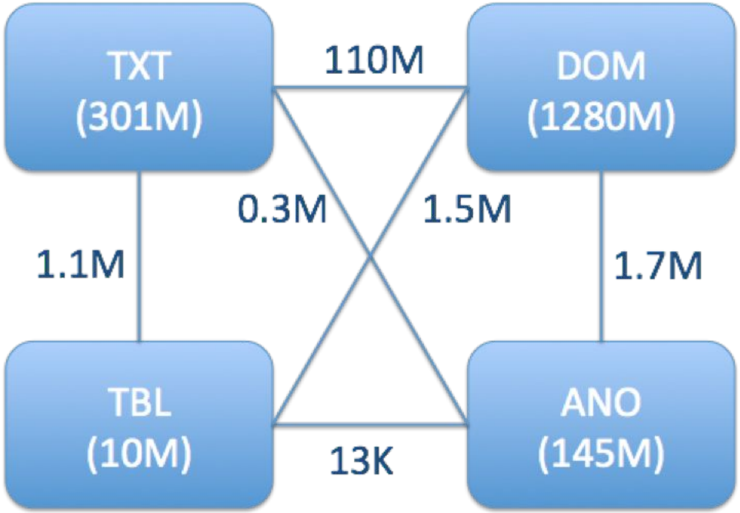
- WebTables: search, extraction
- DOM tree: wrapper induction

2008 (Semi-stru)

Why Semi-Structured Data?

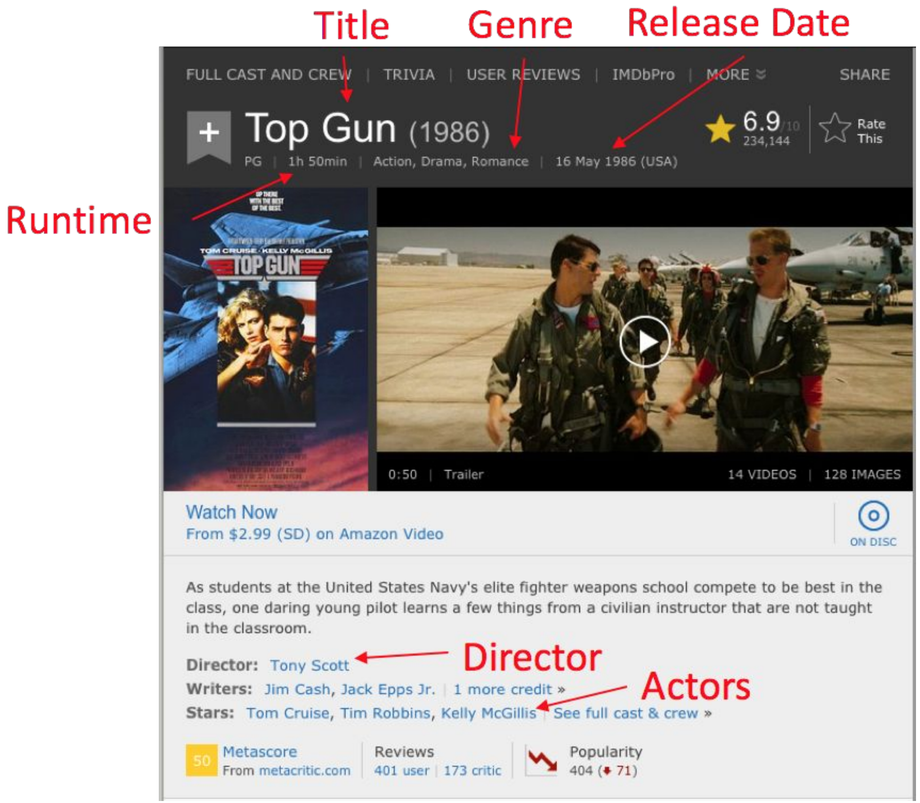
- Knowledge Vault @ Google showed big potential from DOM-tree extraction [Dong et al., KDD'14][Dong et al., VLDB'14]

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

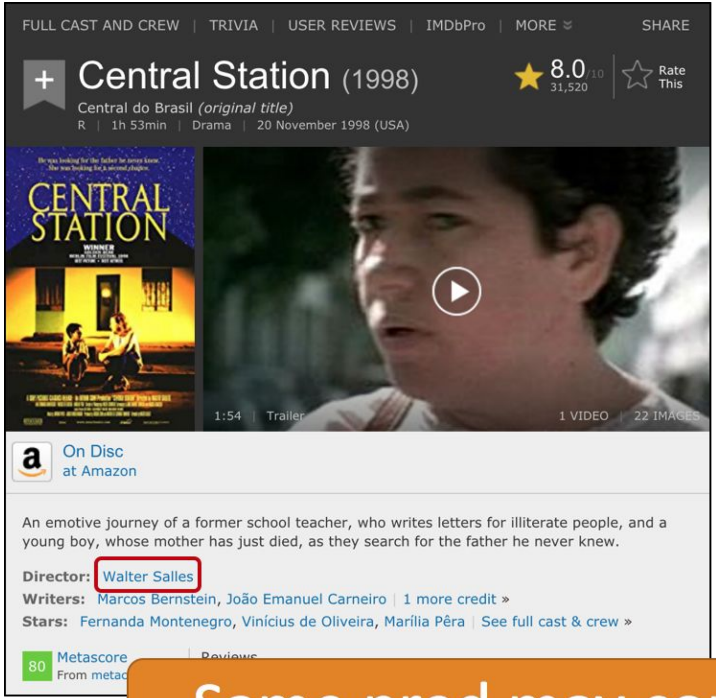
- (Top Gun, type.object.name, "Top Gun")
- (Top Gun, film.film.genre, Action)
- (Top Gun, film.film.directed_by, Tony Scott)
- (Top Gun, film.film.starring, Tom Cruise)
- (Top Gun, film.film.runtime, "1h 50min")
- (Top Gun, film.film.release_Date_s, "16 May 1986")

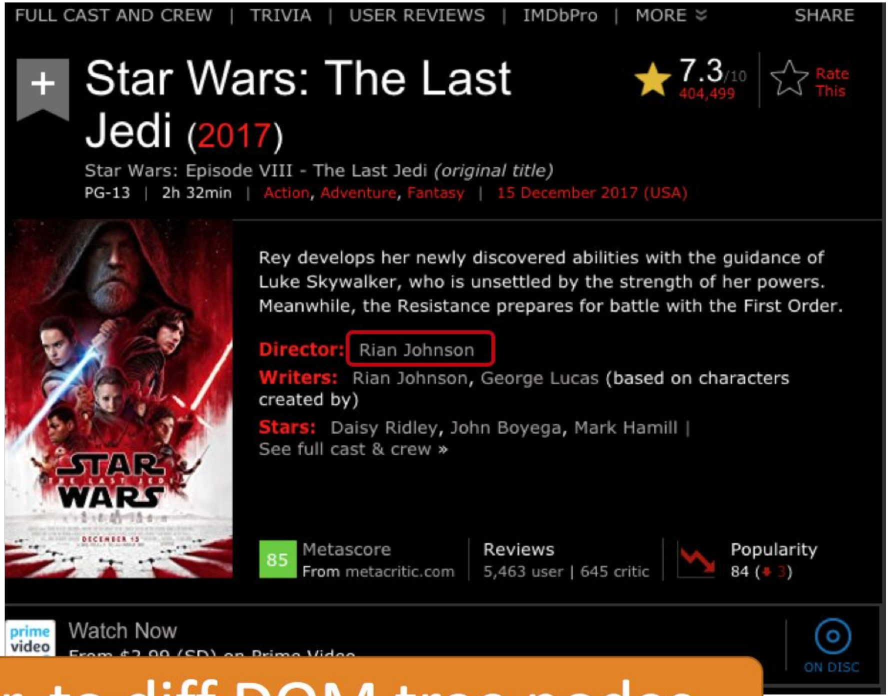
Wrapper Induction--Vertex [Gulhane et al., ICDE'11]

- Solution: find XPaths from DOM Trees

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

- Challenge: slight variations from page to page

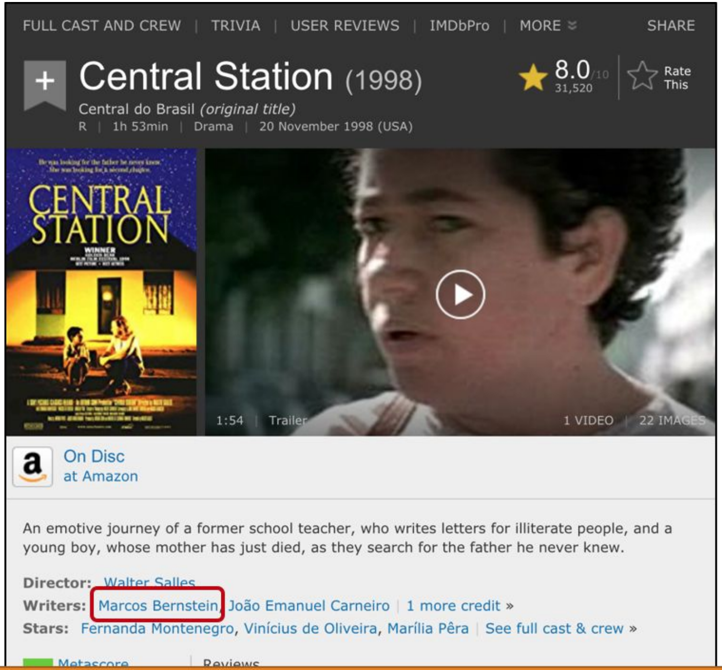


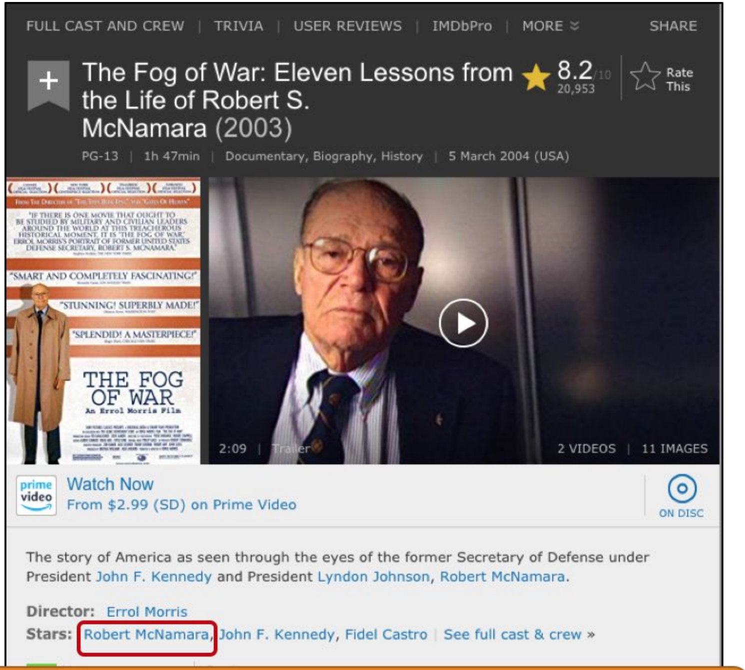


Same pred may corr. to diff DOM tree nodes

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

- Challenge: slight variations from page to page

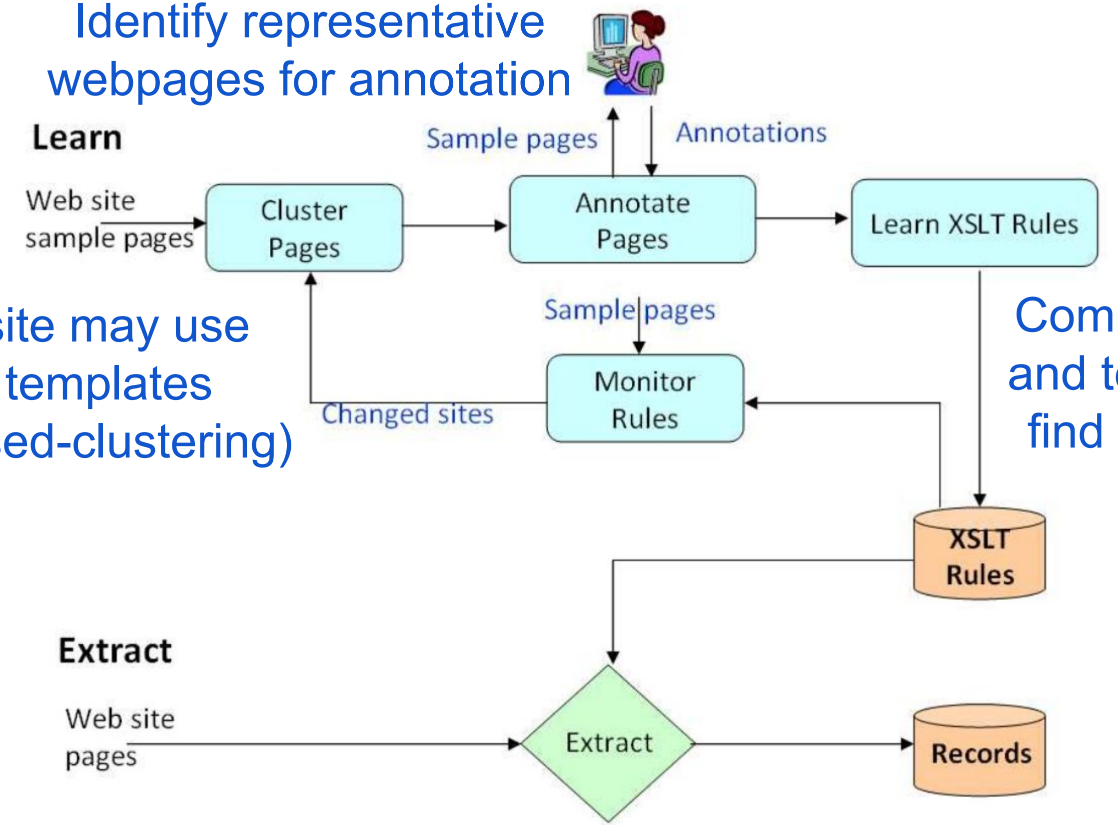




Same DOM tree node may correspond to diff preds

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

Identify representative webpages for annotation



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"]]/*[following-sibling::*[position()=1][@class="attribute"]]
  - //*[preceding-sibling::node()[normalize-space(.)!=""][text()="Language:"]

Ensure high recall

Ensure high precision

Distantly Supervised Extraction

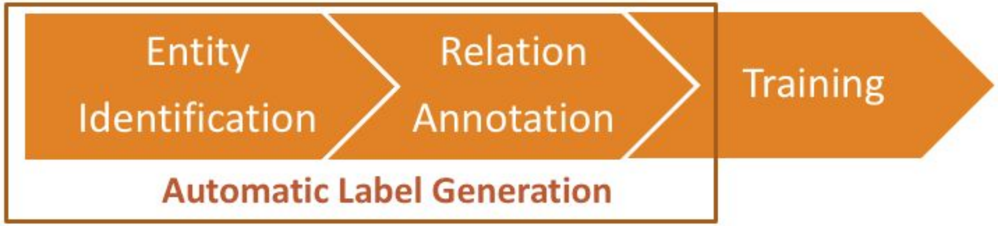
2013 (Deep ML)

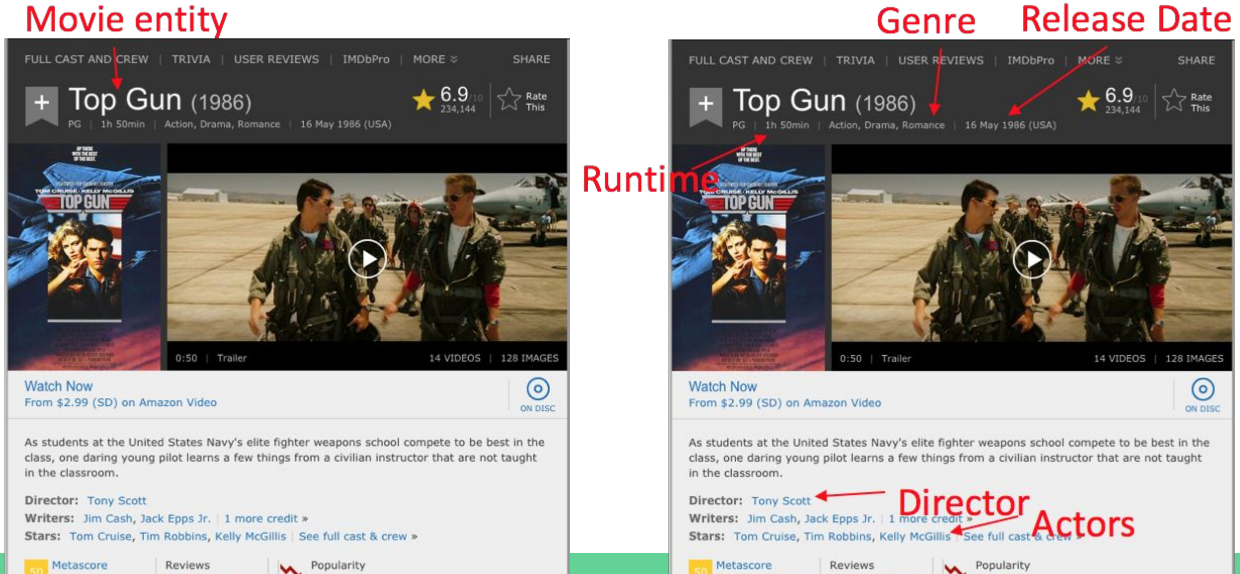
Deep learning

- 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]





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

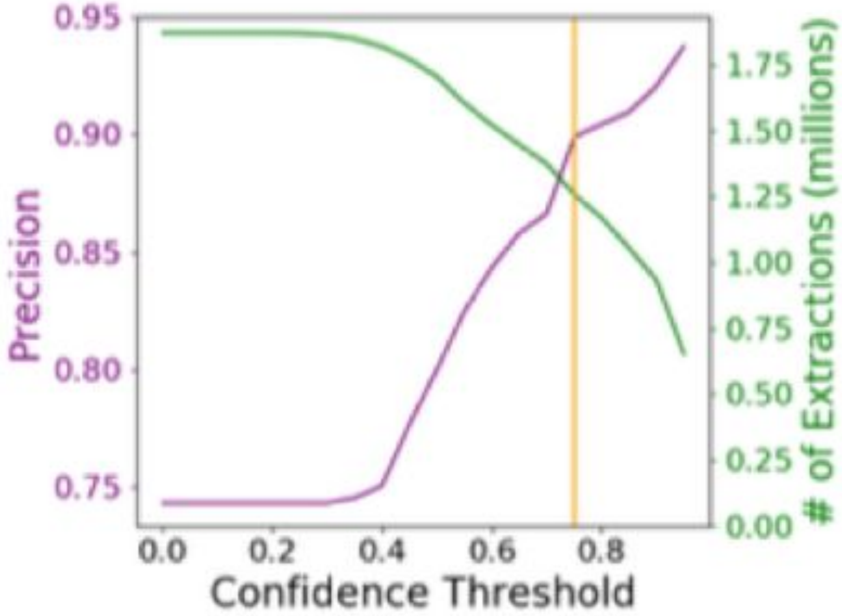
Distantly Supervised Extraction--Ceres [Lockard et al., VLDB'18]

- 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

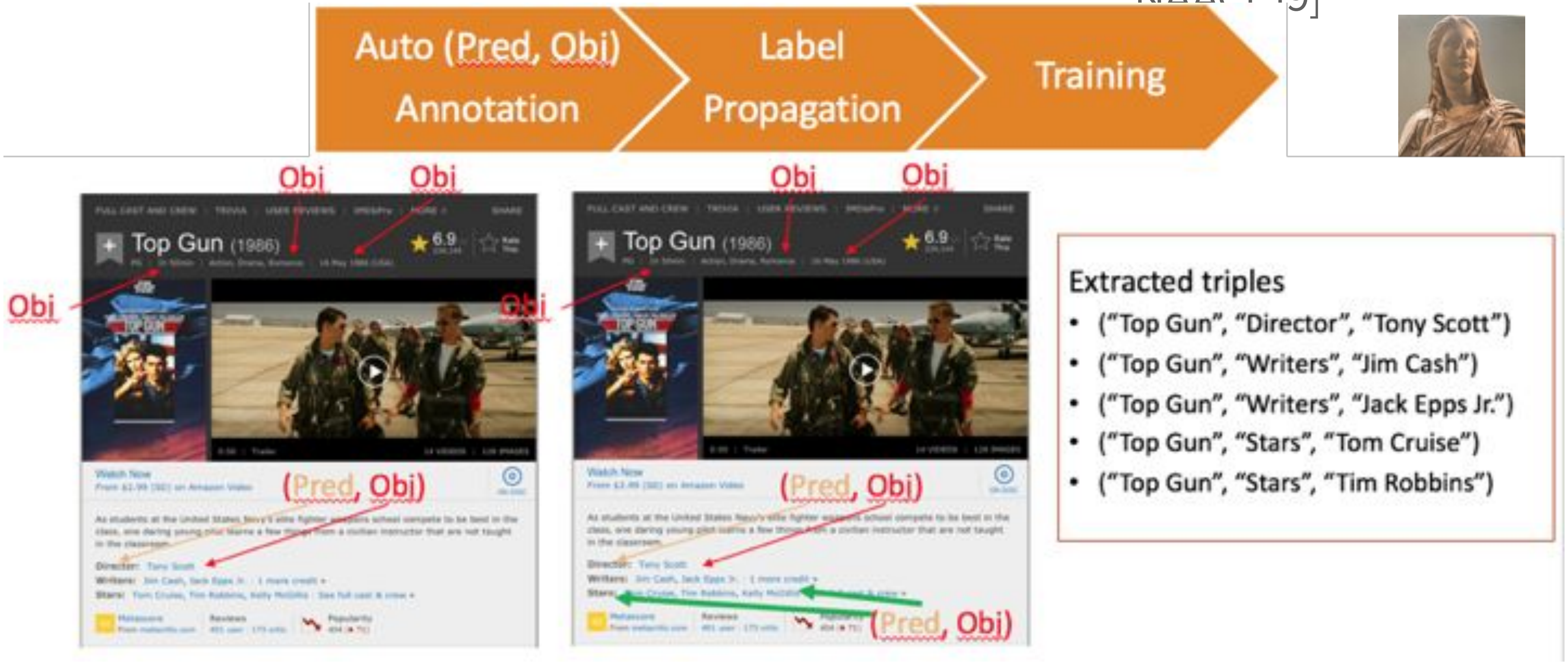
- **Annotation-based extraction**
  - Pros: high precision and recall
  - Cons: does not scale--annotation per cluster per website

- **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]



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

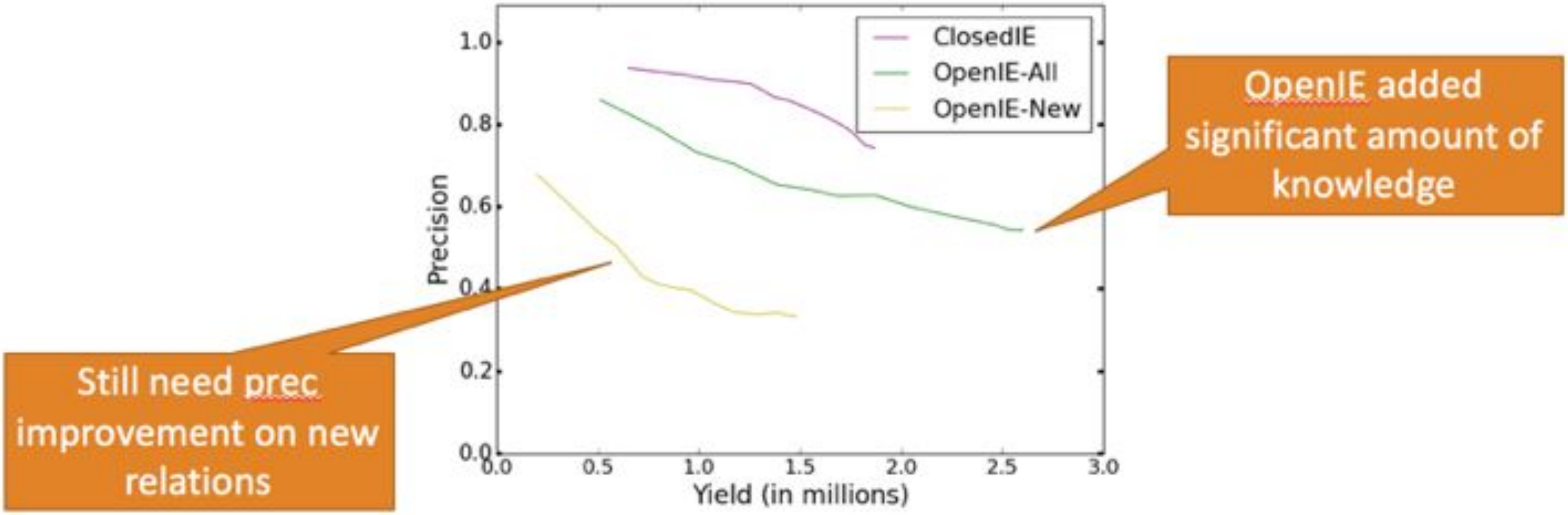
- Seed: Height, Weight, Team
- New: Birth Date, Birth Place, Salary, Age, Experience, Position, College, Year Drafted

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]



Extraction from Semi-structured Websites

2013 (Deep ML)

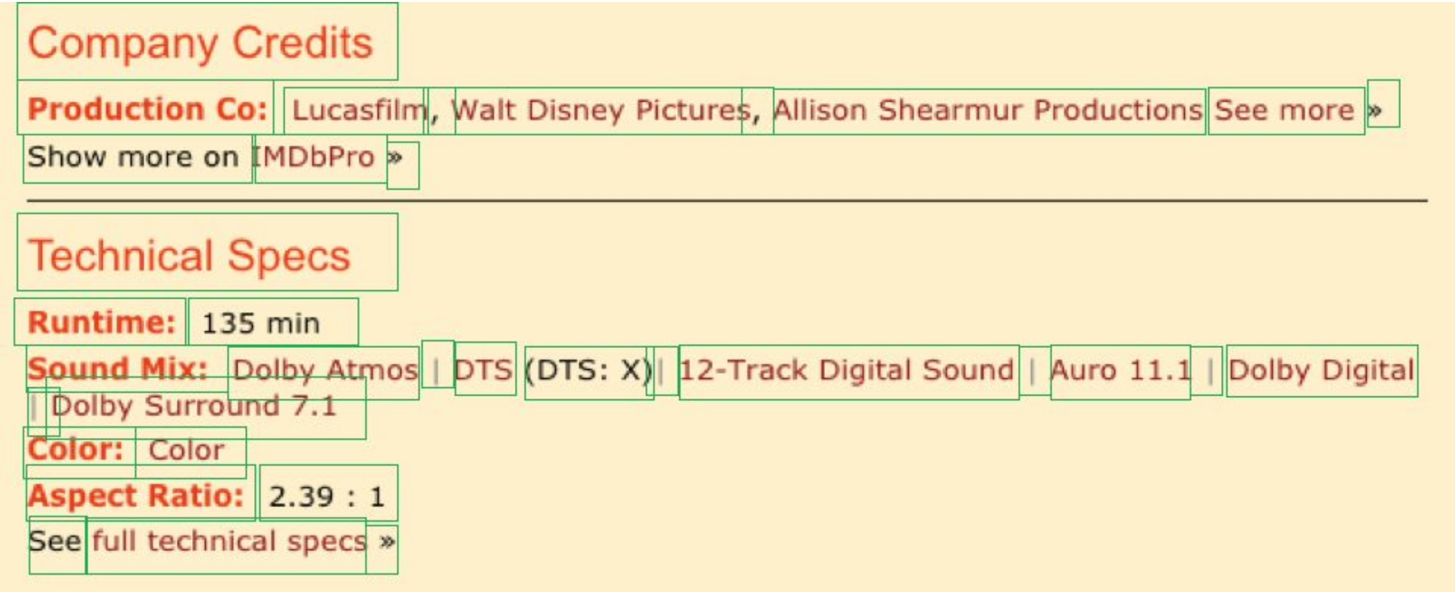
Deep learning

- 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]

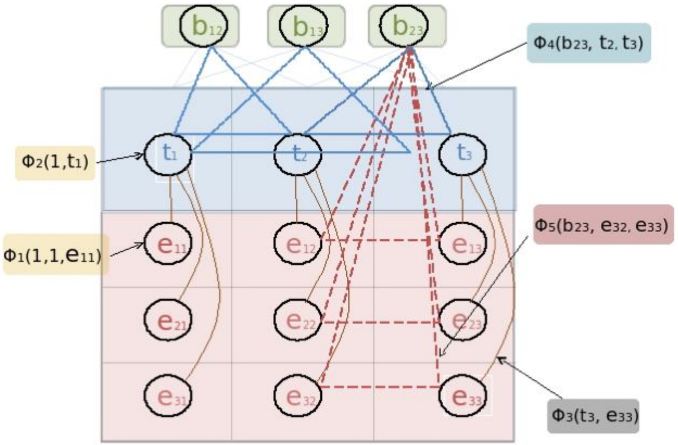
- Model table annotation using interrelated random variables, represented by a probabilistic graphical model
  - Cell text (in Web table) and entity label (in catalog)
  - Column header (in Web table) and type label (in catalog)
  - Column type and cell entity (in Web table)

Extraction from semi-structured data

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

2008 (Semi-stru)

Check-out 10-Year Best Paper Award for WebTable *Search* on Thursday!



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

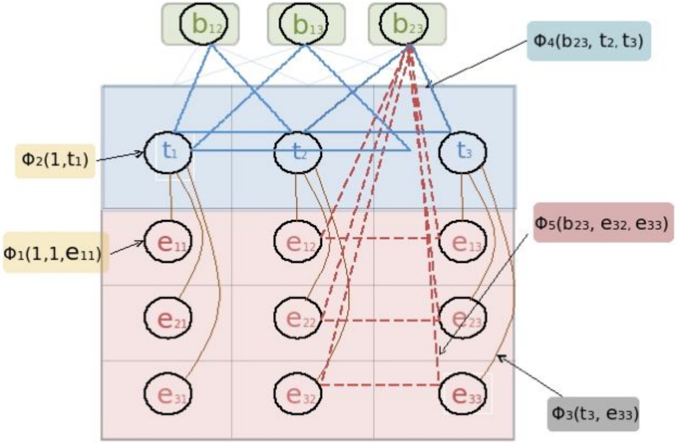
- Model table annotation using interrelated random variables, represented by a probabilistic graphical model
  - Pair of column types (in Web table) and relation (in catalog)
  - Entity pairs (in Web table) and relation (in catalog)

Extraction from semi-structured data

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

2008 (Semi-stru)

Check-out 10-Year Best Paper Award for WebTable *Search* on Thursday!



Challenges in Applying ML on DX

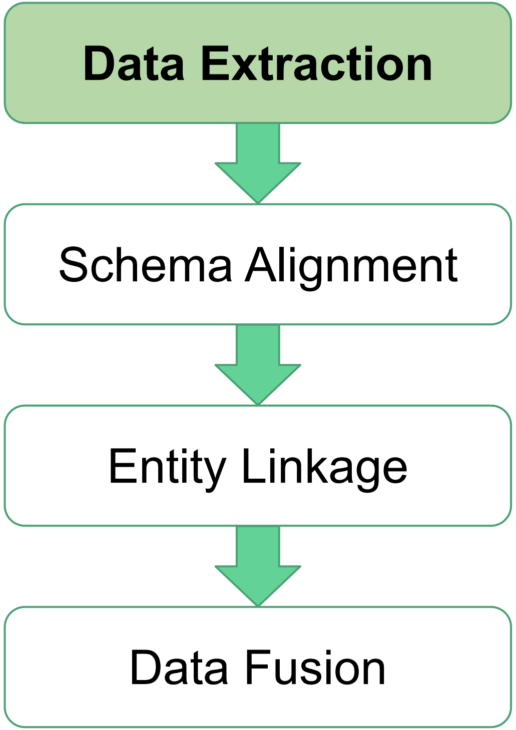
- Automatic data extraction cannot reach production quality requirement. How to improve precision?

- 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?

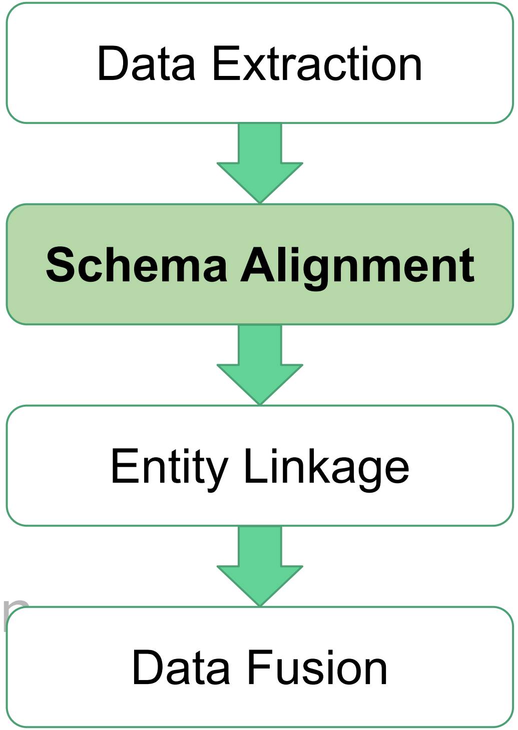
Recipe for Data Extraction

- 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**



Outline

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

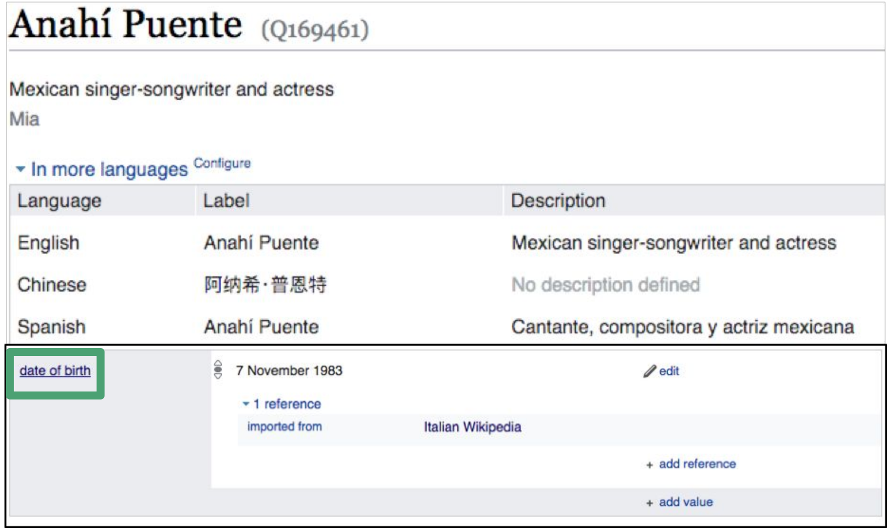


What is Schema Alignment?

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

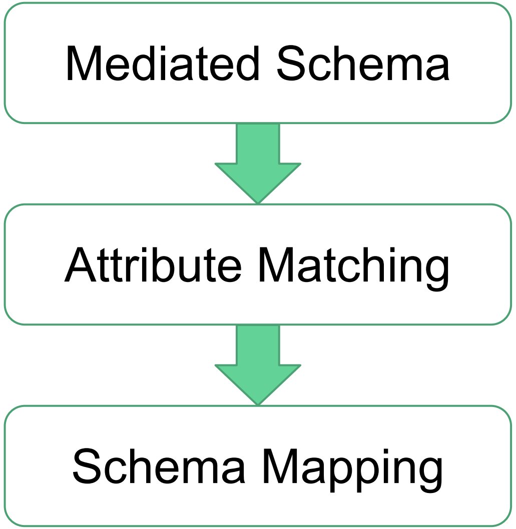
IMDB

WikiData



Quick Tour for Schema Alignment

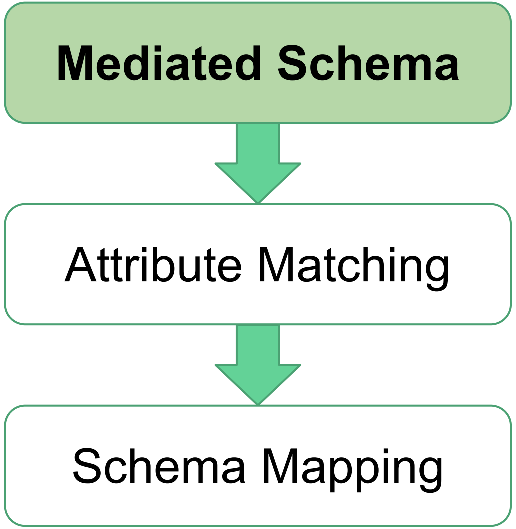
S1	(name, hPhone, hAddr, oPhone, oAddr)
S2	(name, phone, addr, email)
S3	a: (id, name); b: (id, resPh, workPh)
S4	(name, pPh, pAddr)
S5	(name, wPh, wAddr)



Quick Tour for Schema Alignment

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

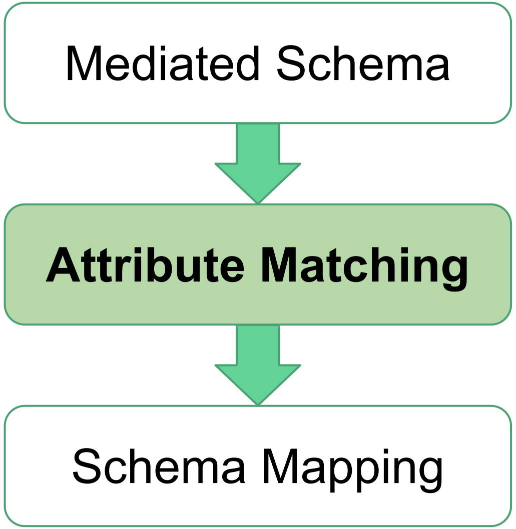
S1	(name, hPhone, hAddr, oPhone, oAddr)
S2	(name, phone, addr, email)
S3	a: (id, name); b: (id, resPh, workPh)
S4	(name, pPh, pAddr)
S5	(name, wPh, wAddr)
MS	(n, pP, pA, wP, wA)



Quick Tour for Schema Alignment

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

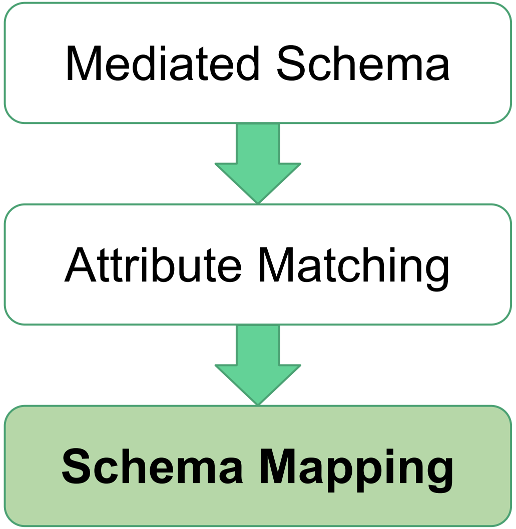
S1	(name, hPhone, hAddr, oPhone, oAddr)
S2	(name, phone, addr, email)
S3	a: (id, name); b: (id, resPh, workPh)
S4	(name, pPh, pAddr)
S5	(name, wPh, wAddr)
MS	**(n, pP, pA, wP, wA)**
MSAM	**MS.n**: S1.name, S2.name, S3a.name, ... **MS.pP**: S1.hPhone, S3b.resPh, S4.pPh **MS.pA**: S1.hAddr, S4.pAddr **MS.wP**: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

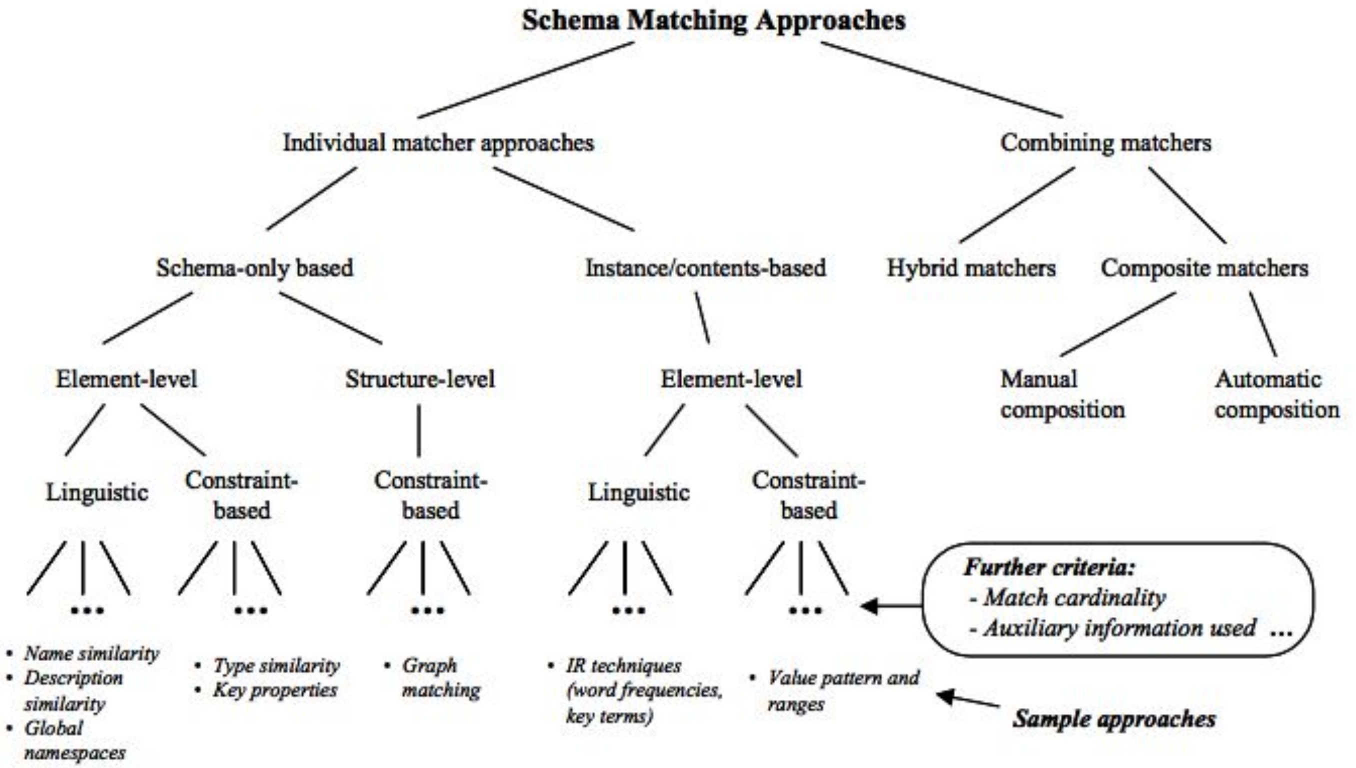
- Probabilistic schema alignment

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

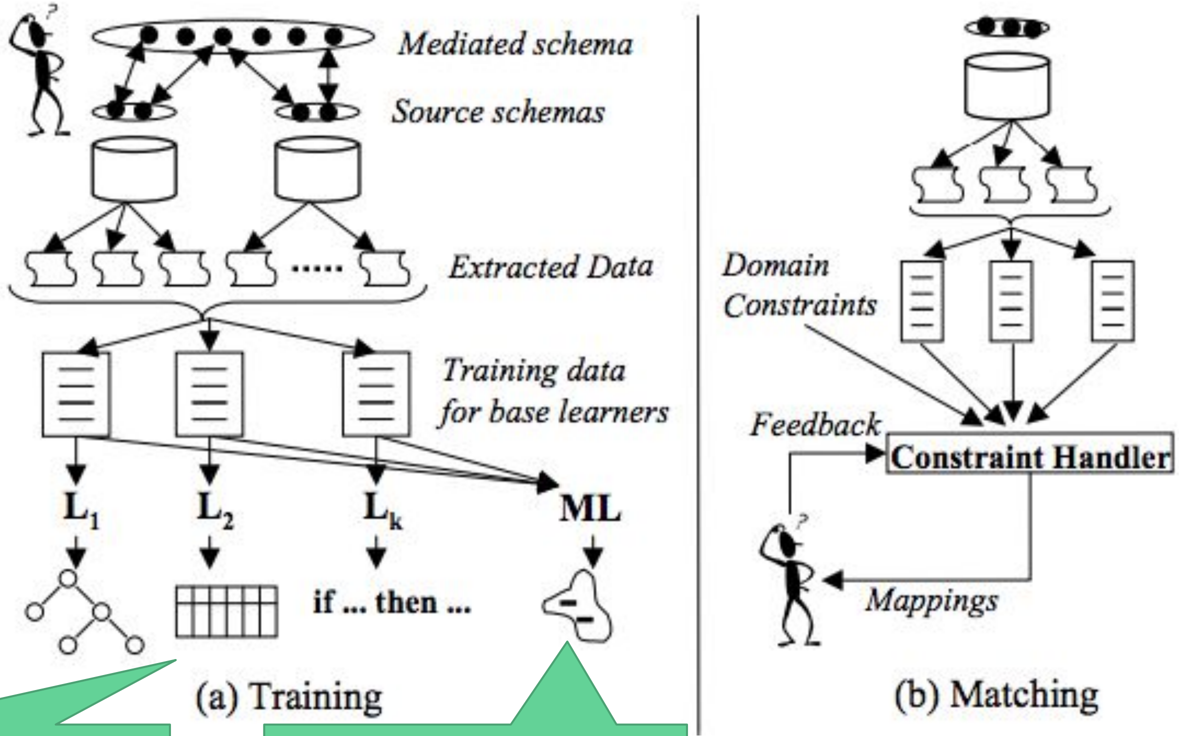
Signals: name, description, type, key, graph structure, values

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]

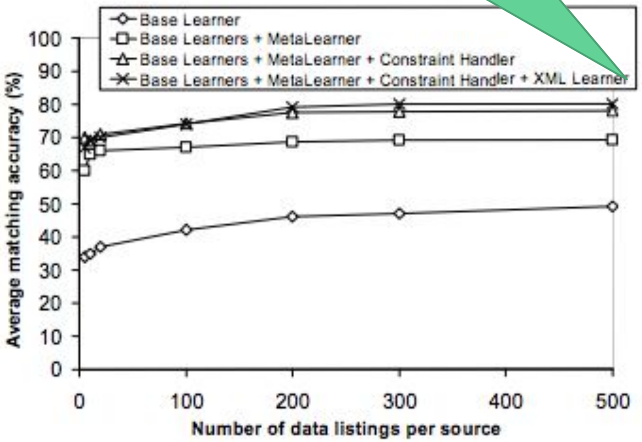
Avg Accuracy: 71-92%

Meta learning and constraints help

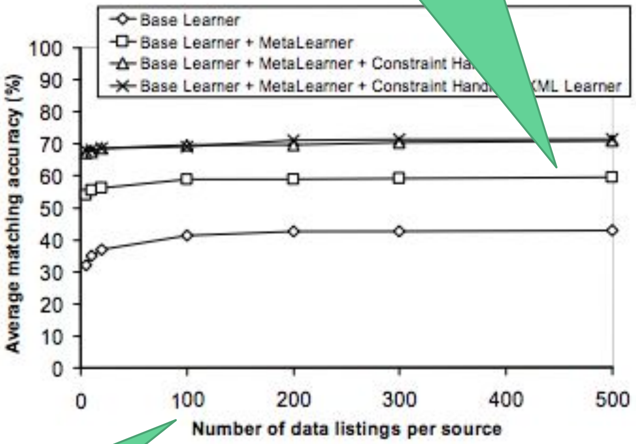
~2000 (Early ML)

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]

Step 1. Generate candidate mappings

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

2013 (Deep ML)

Step 2. Solve PSL

1.Prefer fewer mappings: penalty=#atoms

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

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

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]

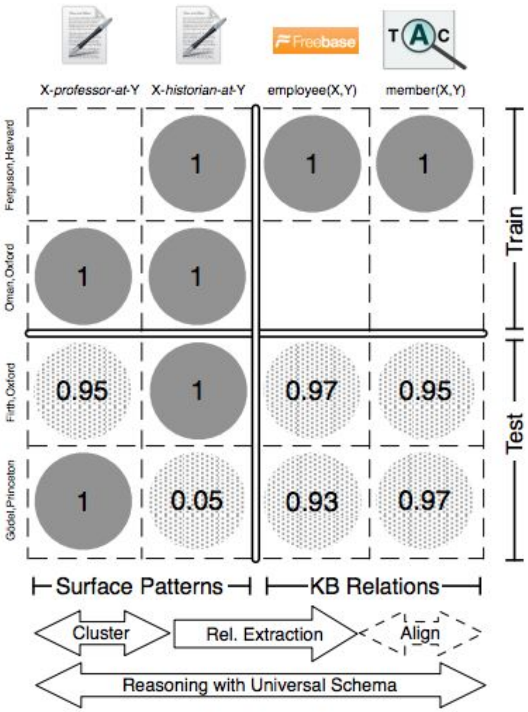
- Attribute matching ➔ Instance inference

Matrix factorization

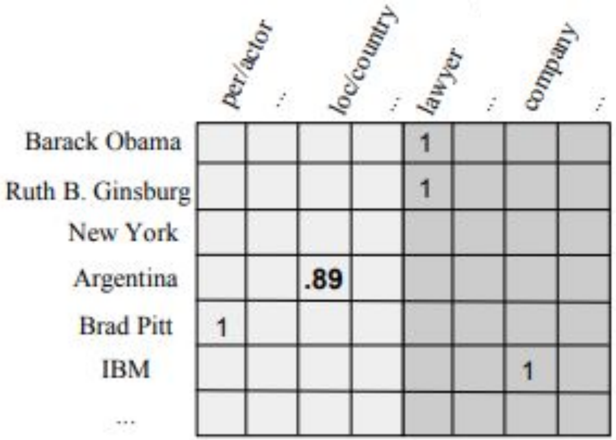
2013 (Deep ML)

Logic & Deep learning

- Collective disc. by PSL
- Universal schema



Relation prediction



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

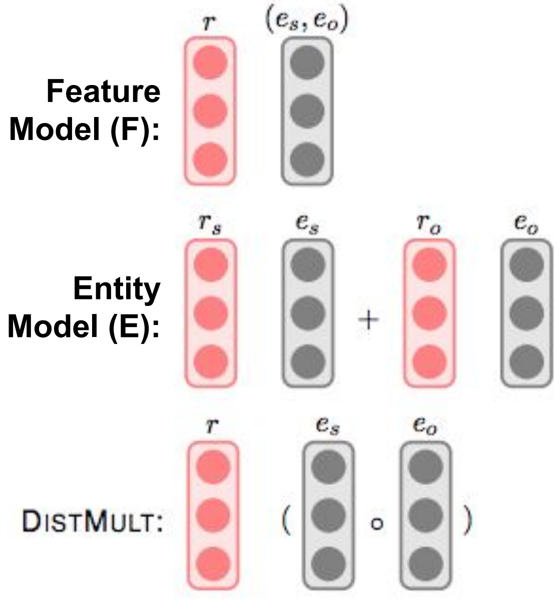


Figure 3: The continuous representations for model F, E and DISTMULT. [Toutanova et al., EMNLP’15]

Columnless Univ. Schema w. CNN [Toutanova et al., EMNLP'15]

- 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	[illegible]
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

2013 (Deep ML)

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

Columnless Univ. Schema w. CNN [Toutanova et al., EMNLP’15]

2013 (Deep ML)

Logic & Deep learning

- Collective disc. by PSL
- Universal schema

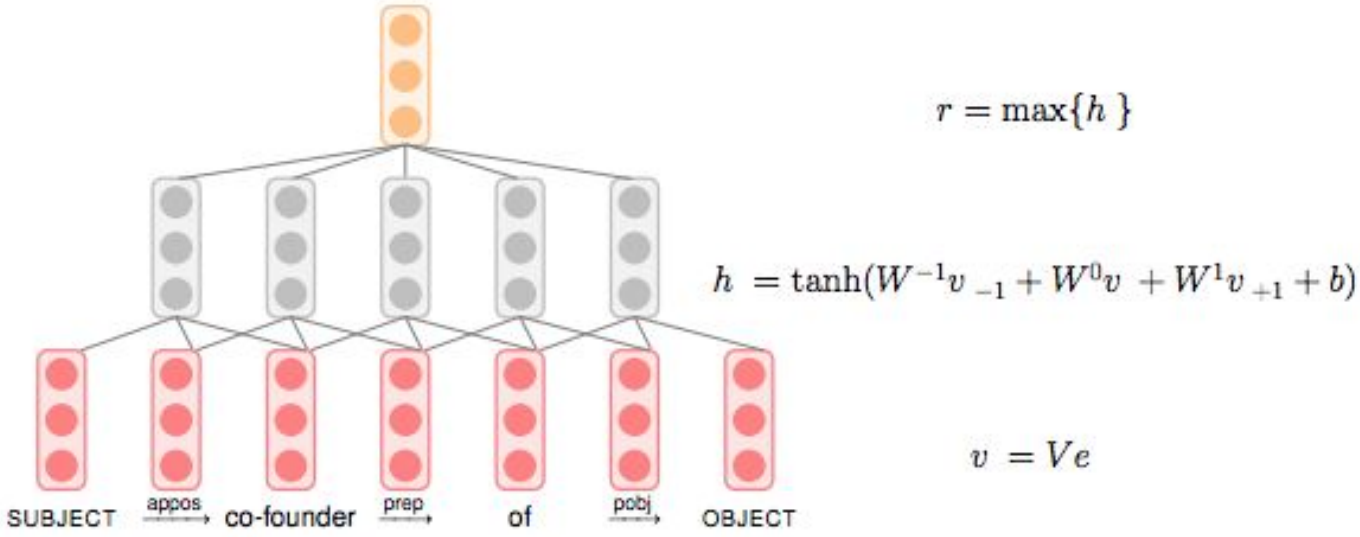
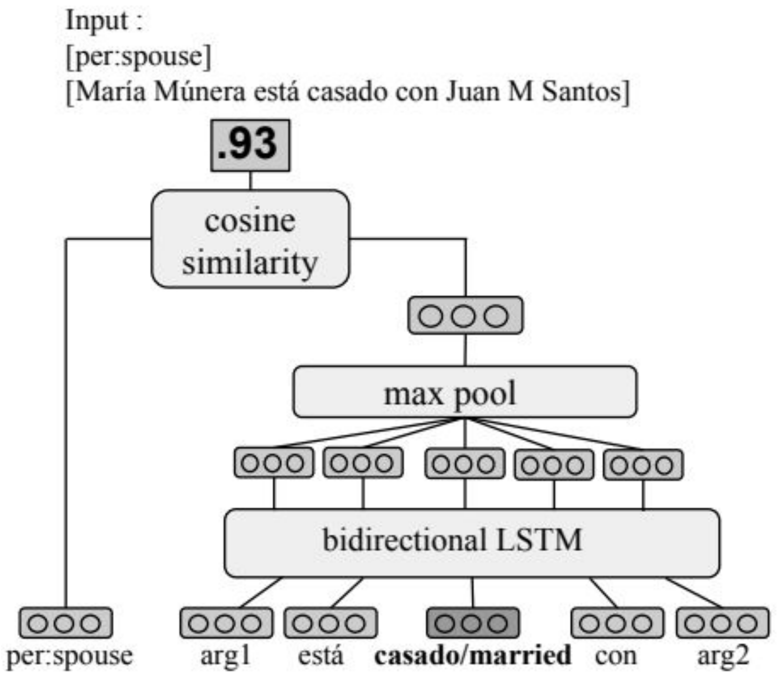


Figure 4: The convolutional neural network architecture for representing textual relations.

Columnless Univ. Schema w. RNN [Verga et al., ACL'16]

- Similar sequences of context tokens should be embedded similarly



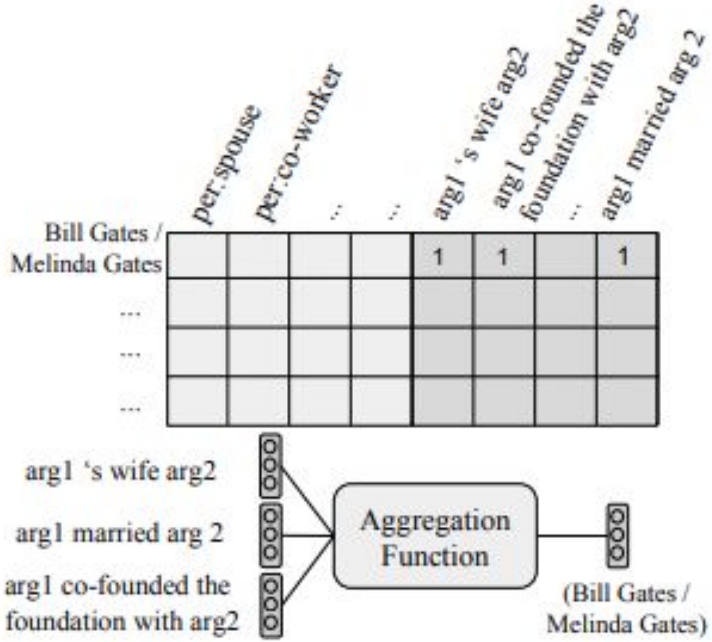
2013 (Deep ML)

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)

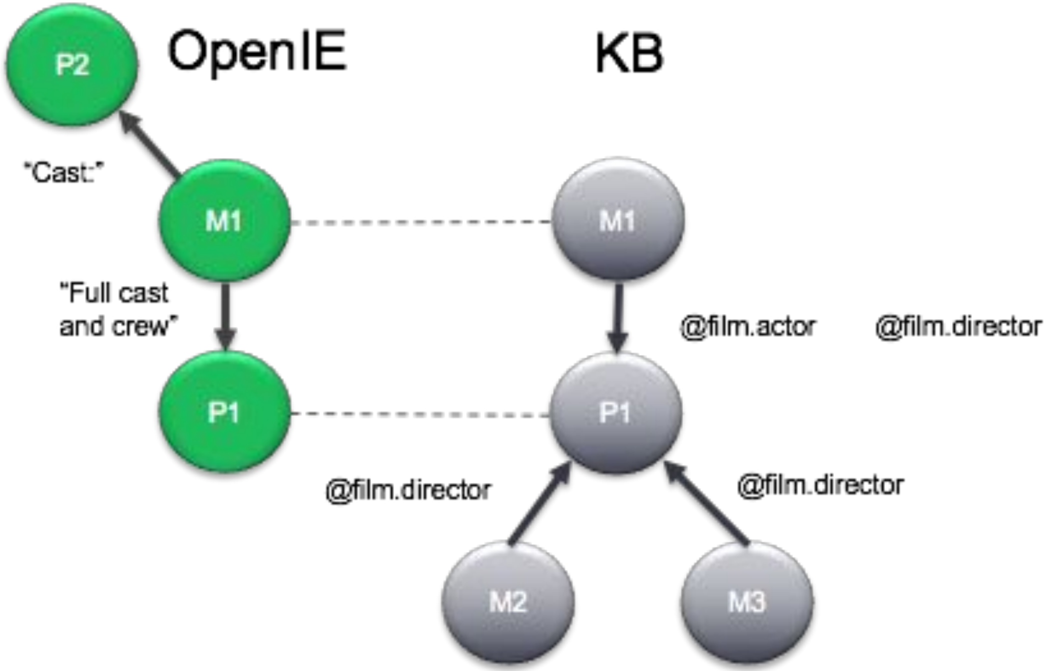
Recall still low

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**

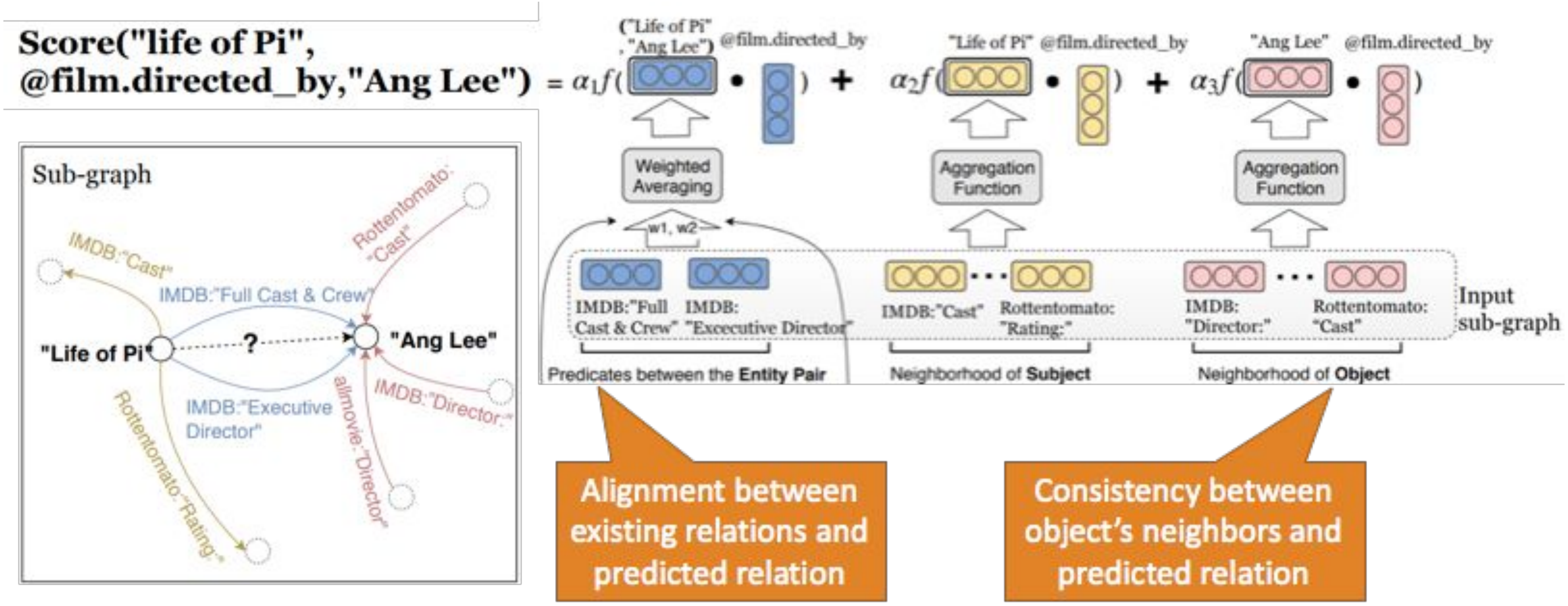
(b)

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]



OpenKI: Relation Inference for OpenIE [Zhang et al., NAACL'19]

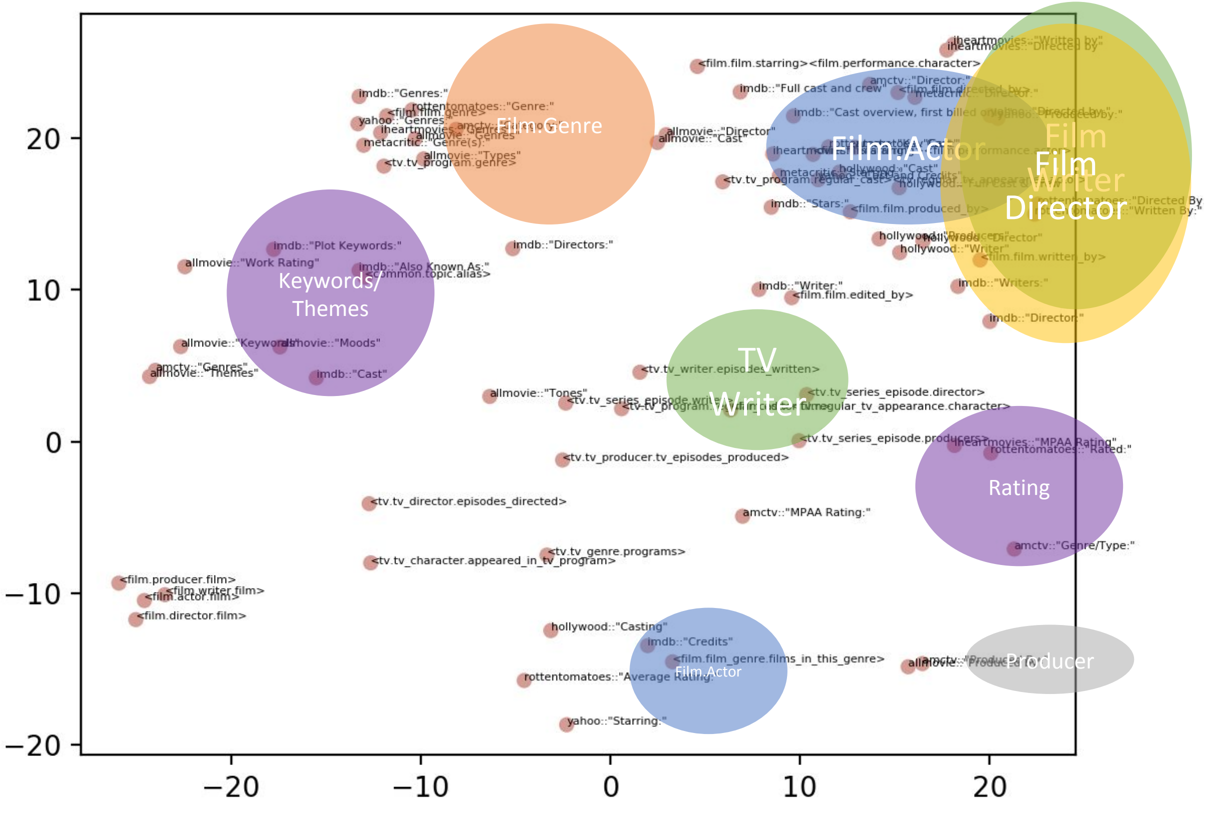
Models	All data	At least one seen
Rowless Model	0.278	0.282
OpenKI with Dual Att.	0.365	0.419

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]

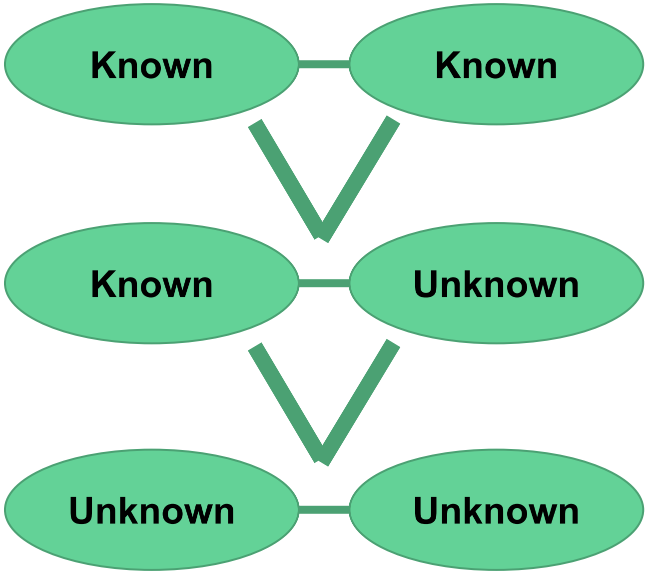


Schema Mapping vs. Universal Schema

	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??



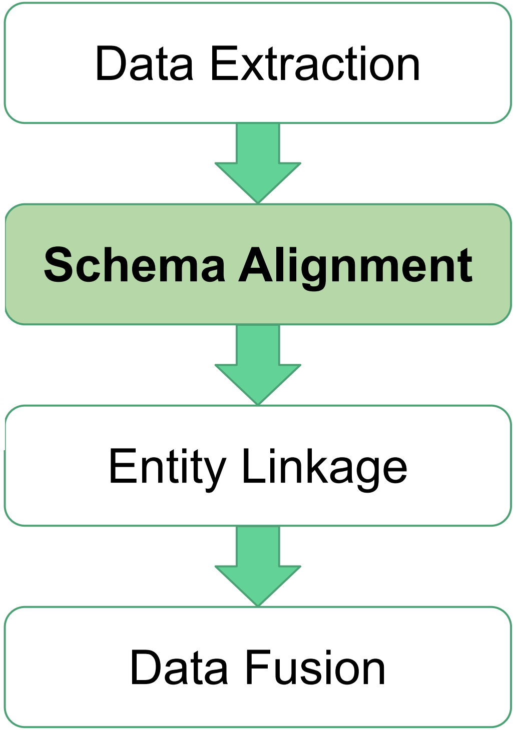
Leverage knowledge by inference

Leverage knowledge on types

Rowless

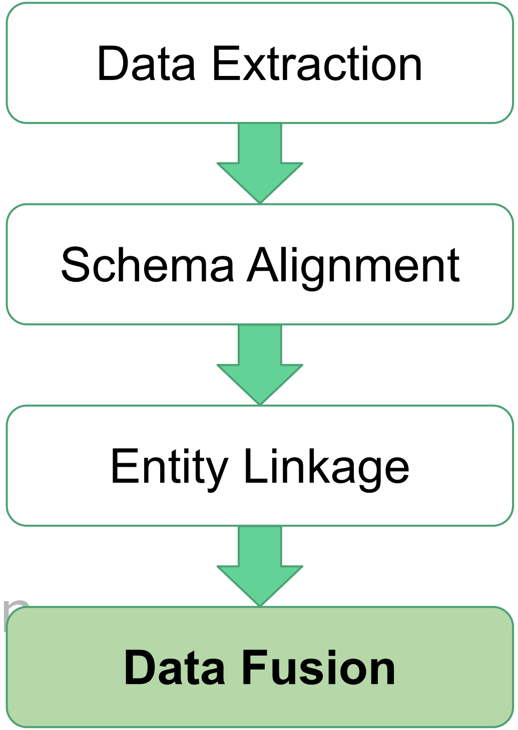
Recipe for Schema Alignment

- 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**



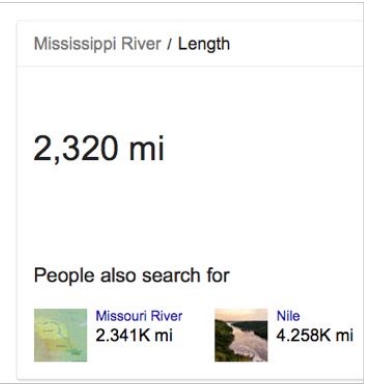
Outline

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



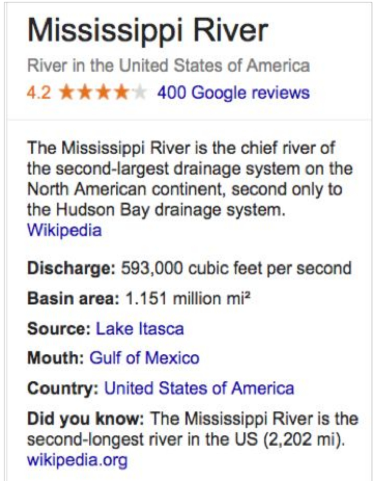
Mississippi River / Length

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

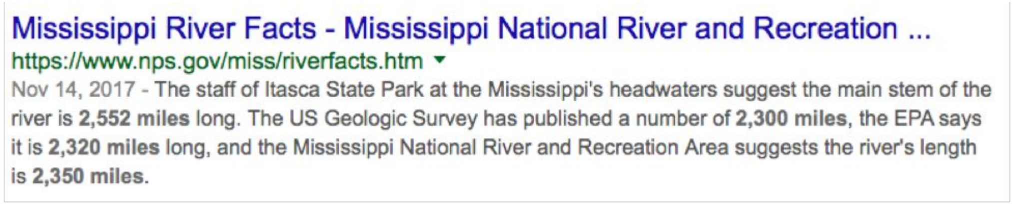
Basin area: 1.151 million mi²

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**.

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

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.

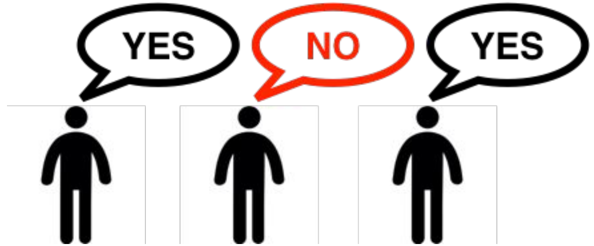
Majority Voting for Data Fusion

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.

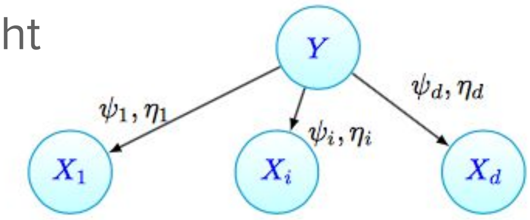
But the number of sources can be in the thousands or millions and training data is limited!

Idea: 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.

An Intro in 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).

An Intro in Probabilistic Graphical Models

Bayesian Networks (BNs)

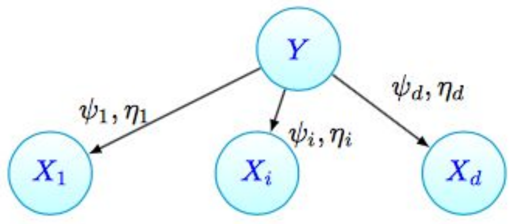
Local Markov Assumption: A variable X is independent of its non-descendants given its parents (and *only* its parents).

Recipe for BNs

Set of random variables X

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

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



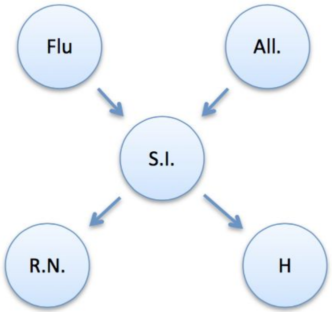
Joint distribution: Factorizes over conditional probability tables

An Intro in 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

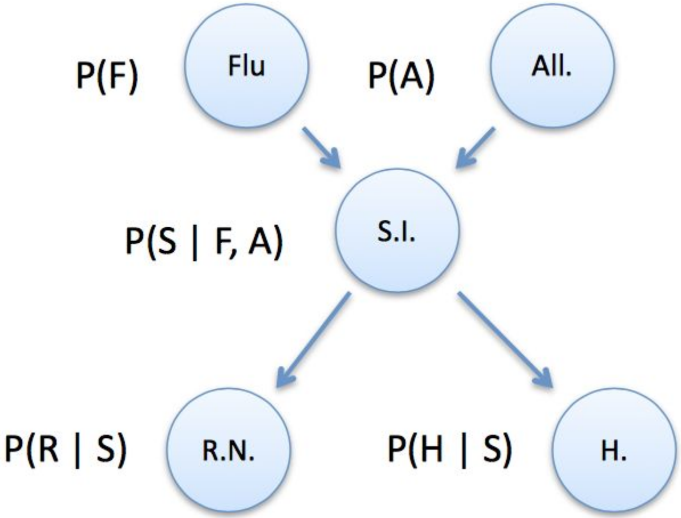


[Example by Andrew McCallum]

An Intro in Probabilistic Graphical Models

Factored joint distribution

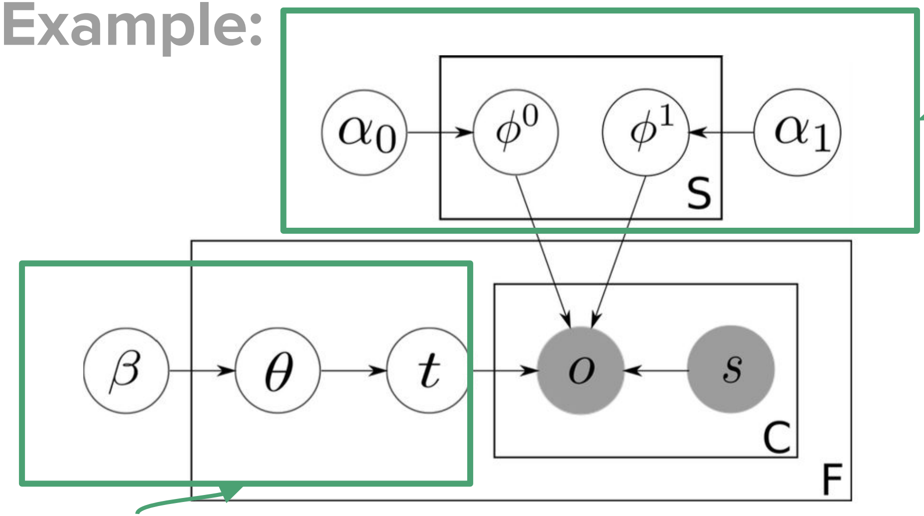
$$
\begin{aligned}
&P(F, A, S, R, H) \\
&= P(F) \\
&\quad P(A) \\
&\quad P(S \mid F, A) \\
&\quad P(R \mid S) \\
&\quad P(H \mid S)
\end{aligned}
$$



[Example by Andrew McCallum]

Probabilistic Graphical Models for Data Fusion

Example:



Source Quality

Prior truth probability

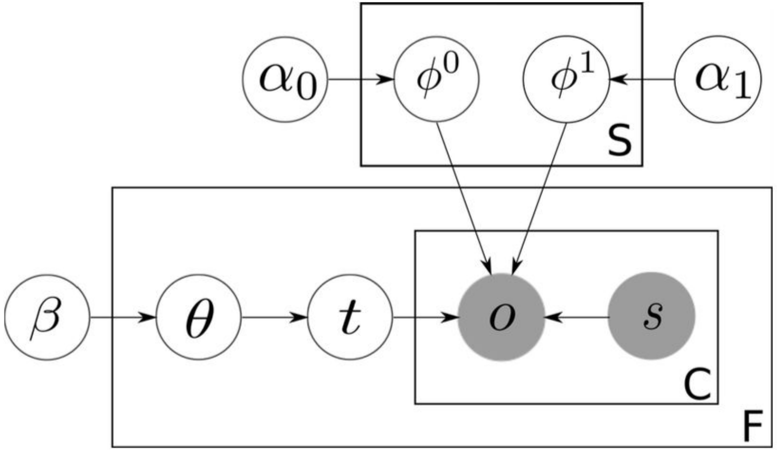
[Zhao et al., VLDB 2012]

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
...	...	...

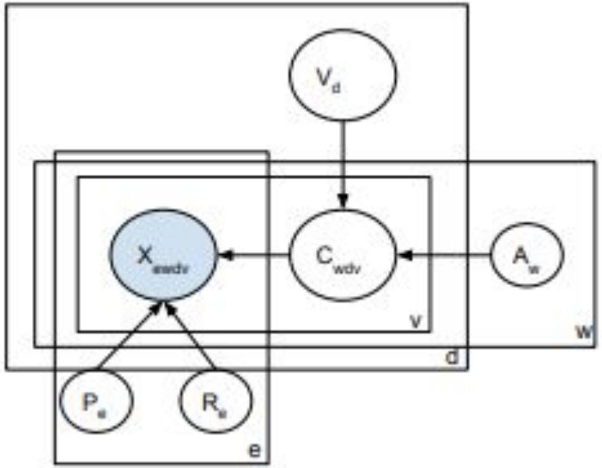
Extensive work on modeling source observations and source interactions to address limitations of basic Dawid-Skene.

Probabilistic Graphical Models for Data Fusion



[Zhao et al., VLDB 2012]

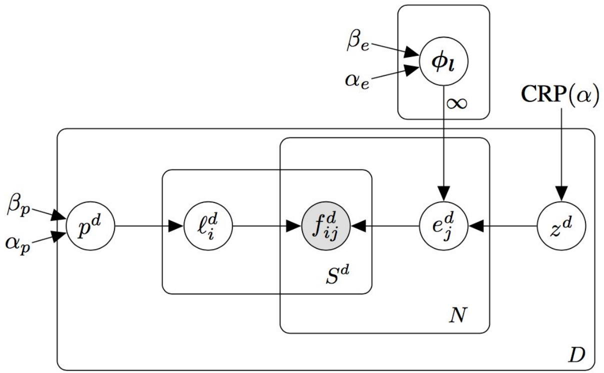
Modeling both source quality and extractor accuracy



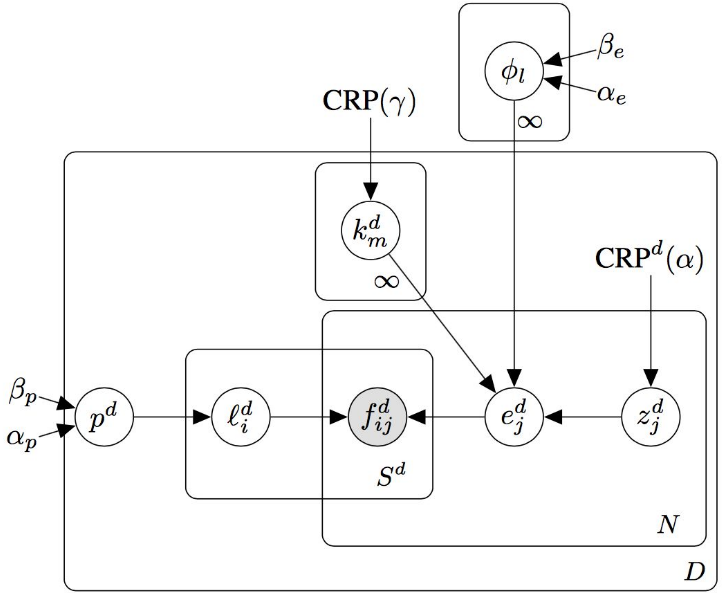
[Dong et al., VLDB 2015]

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				Flight			
		prec w. trust	prec w/o. trust	Trust dev	Trust diff	prec w. trust	prec w/o. trust	Trust dev	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.

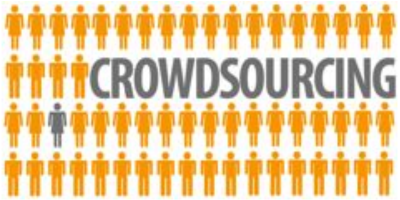
Discriminative Data Fusion [SLiMFast Rekatsinas et al., SIGMOD'17]

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

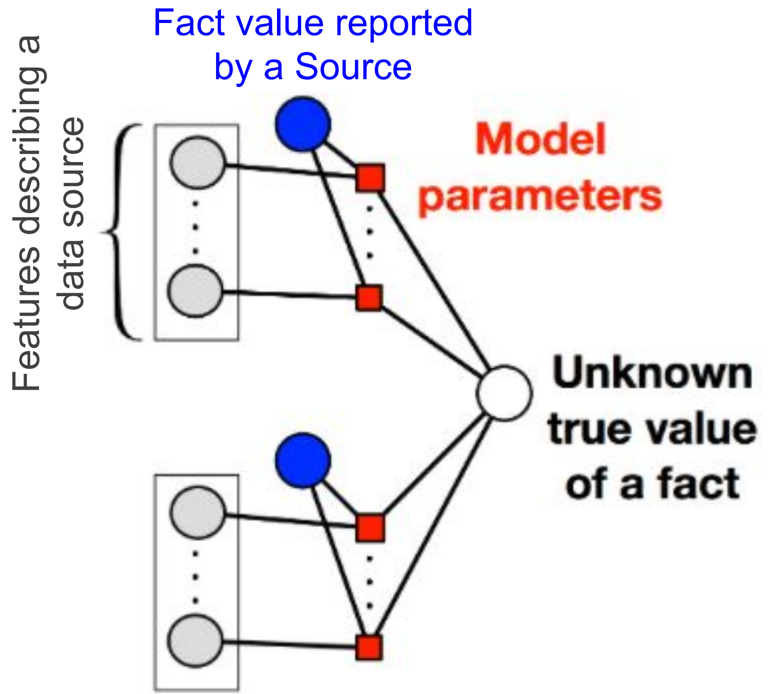
Example:

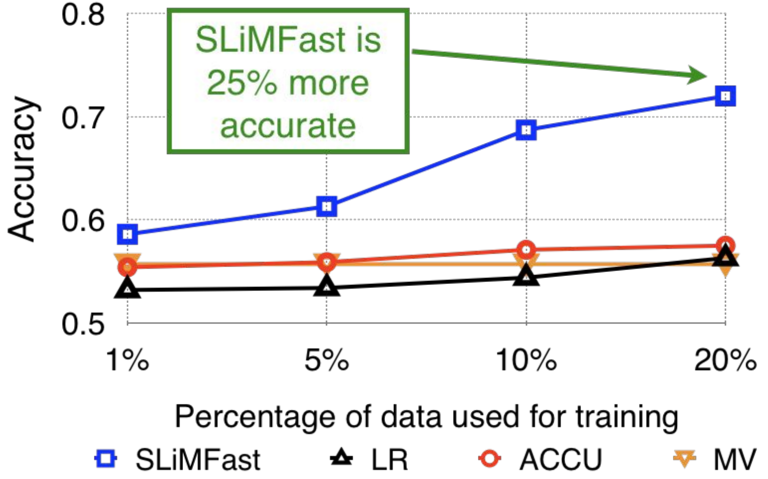
- newly registered similar to existing domain
- traffic statistics
- text quality (e.g., misspelled words, grammatical errors)
- sentiment analysis



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

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

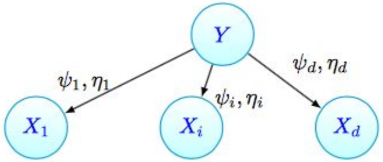




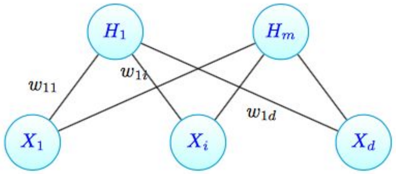
Genomics data: 2.7k sources (articles), 571 objects (gene-disease), 4 domain features (year, citation, author, journal)

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

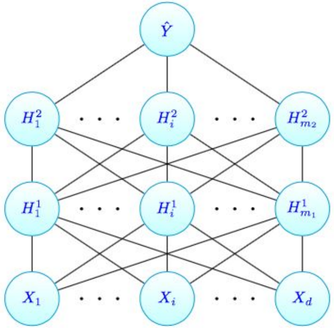
Theorem: The Dawid and Skene model is *equivalent* to a Restricted Boltzmann Machine (RBM) with a single hidden node.



Dawid and Skene model.



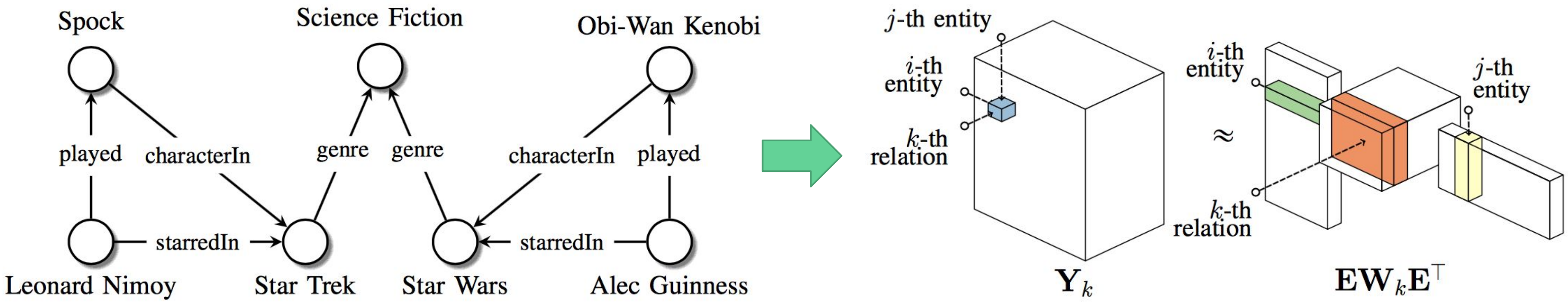
A RBM with *d* visible and *m* hidden units.



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.

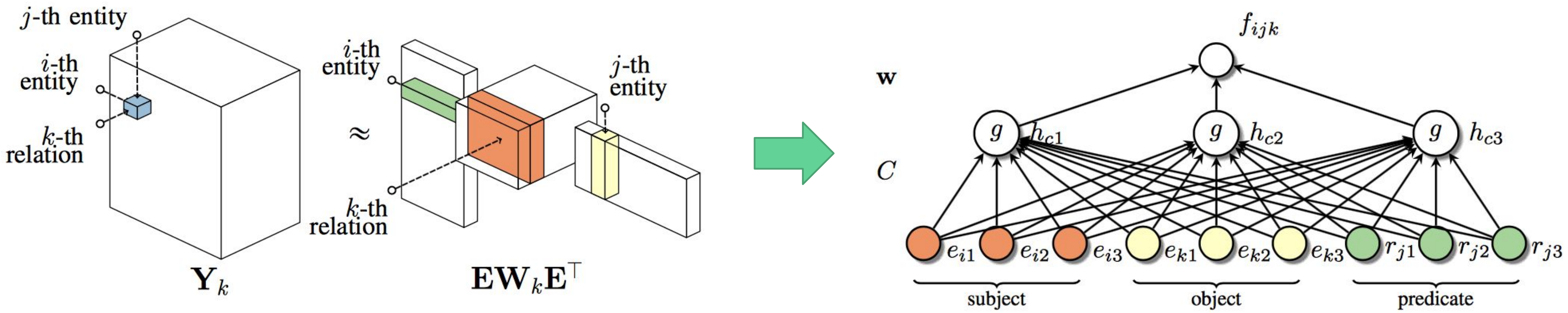
Data Fusion For Complex Data



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

A knowledge graph can be encoded as a tensor.

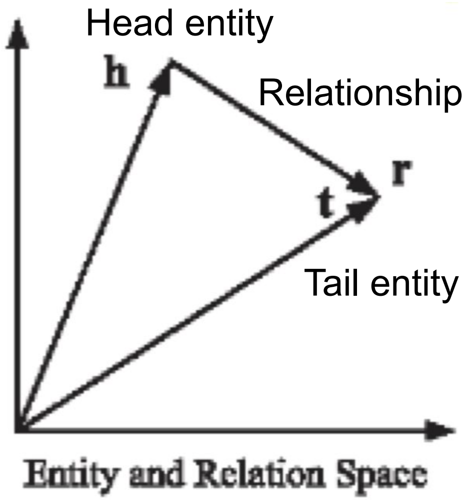
Data Fusion For Complex Data



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

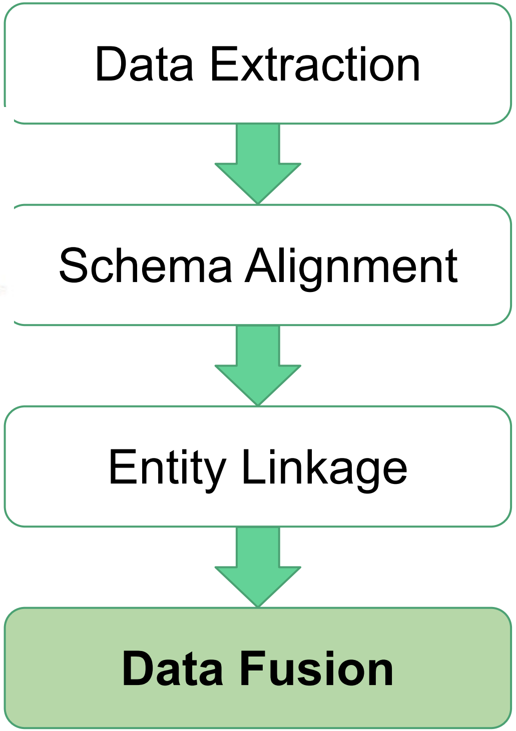
- There are few solutions for unstructured data. Mostly work on fact verification [Tutorial by Dong et al., KDD`2018]. Most data Fusion solutions assume data extraction. Can state-of-the art DL help?

- 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.**



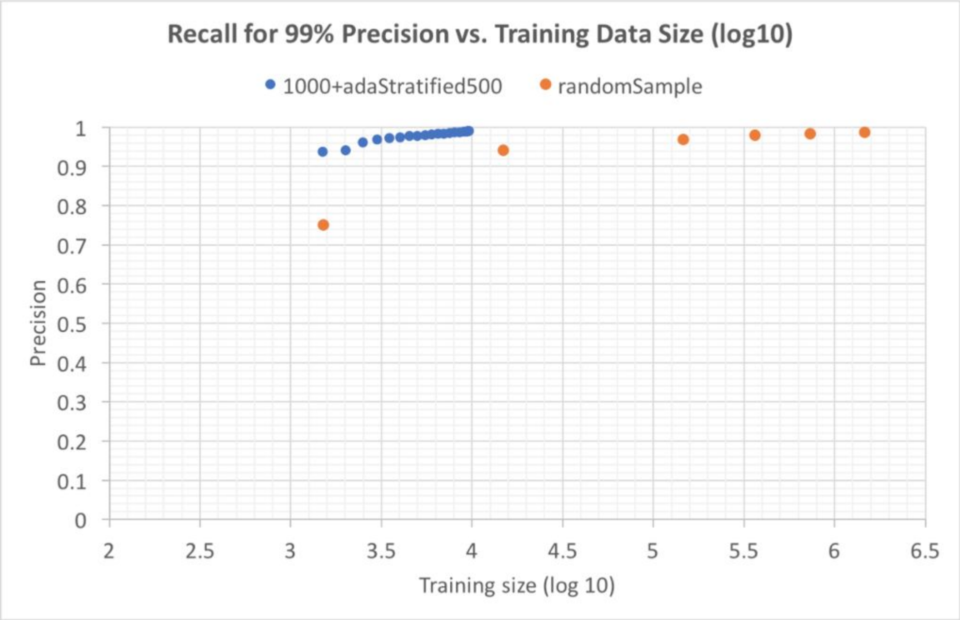
Takeaways

Revisit Theme I. Which ML Model Works Best?

DI tasks	Hyperplanes (e.g., Log Reg)	Kernal (e.g., SVM)	Tree-based (e.g., Random forest)	Graphical models (e.g., CRF)	Logic programs (e.g, soft logic)	Neural networks (e.g., RNN)
Entity resolution	X	X	X		X	X
Data fusion	X			X		
DOM extraction	X					
Text extraction	X	X		X		X
Schema alignment	X		X	X	X	X

For structured data, RF works well, and LR is often effective
For texts and semantics, deep learning shows big promise

Revisit Theme II. Does Supervised Learning Apply to DI?



Active learning, semi-supervised learning, and weak supervision lead to dramatically more efficient solutions.

Outline

- Part I. Introduction
- Part II. ML for DI
- Part III. DI for ML
  - Data Cleaning
  - Training Data Creation
- Part IV. Conclusions and research directions

50 Years of Artificial Intelligence

1970s (Rules)

Expert systems
- Manually curated knowledge bases of facts and rules
- Use of inference engines
- No support for high-dimensional data

1990s (Features)

Classical ML
- Low complexity models
- Strong priors that capture domain knowledge (feature engineering)
- Small amounts of training data

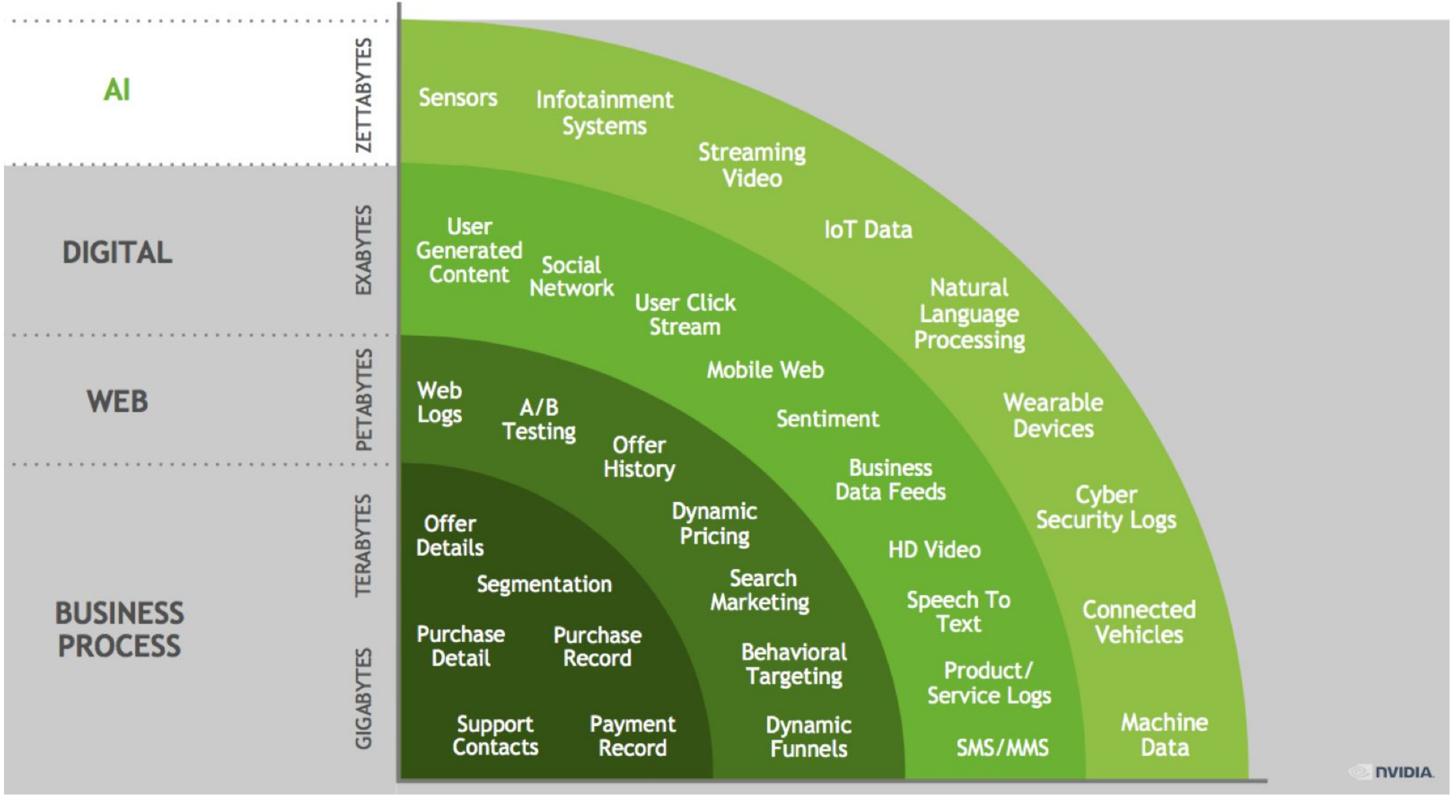
2009 (PGMs)

Graphical models and logic
- Relational statistical learning
- Markov logic network

2010s (Representation Learning)

Deep learning
- Automatically learn representations
- Impressive with high-dimensional data
- Data hungry!

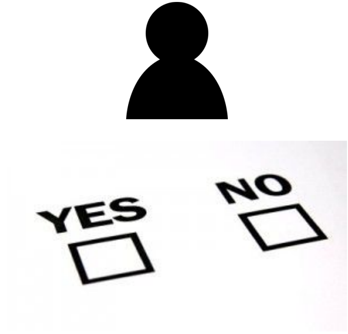
Modern ML is data-hungry



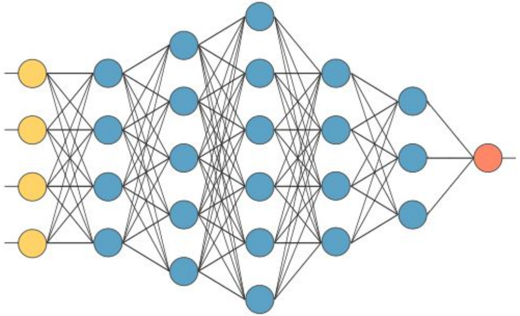
The ML Pipeline in the Deep Learning Era

Data Collection

Data Labeling



Representation Learning and Training



The ML Pipeline in the Deep Learning Era

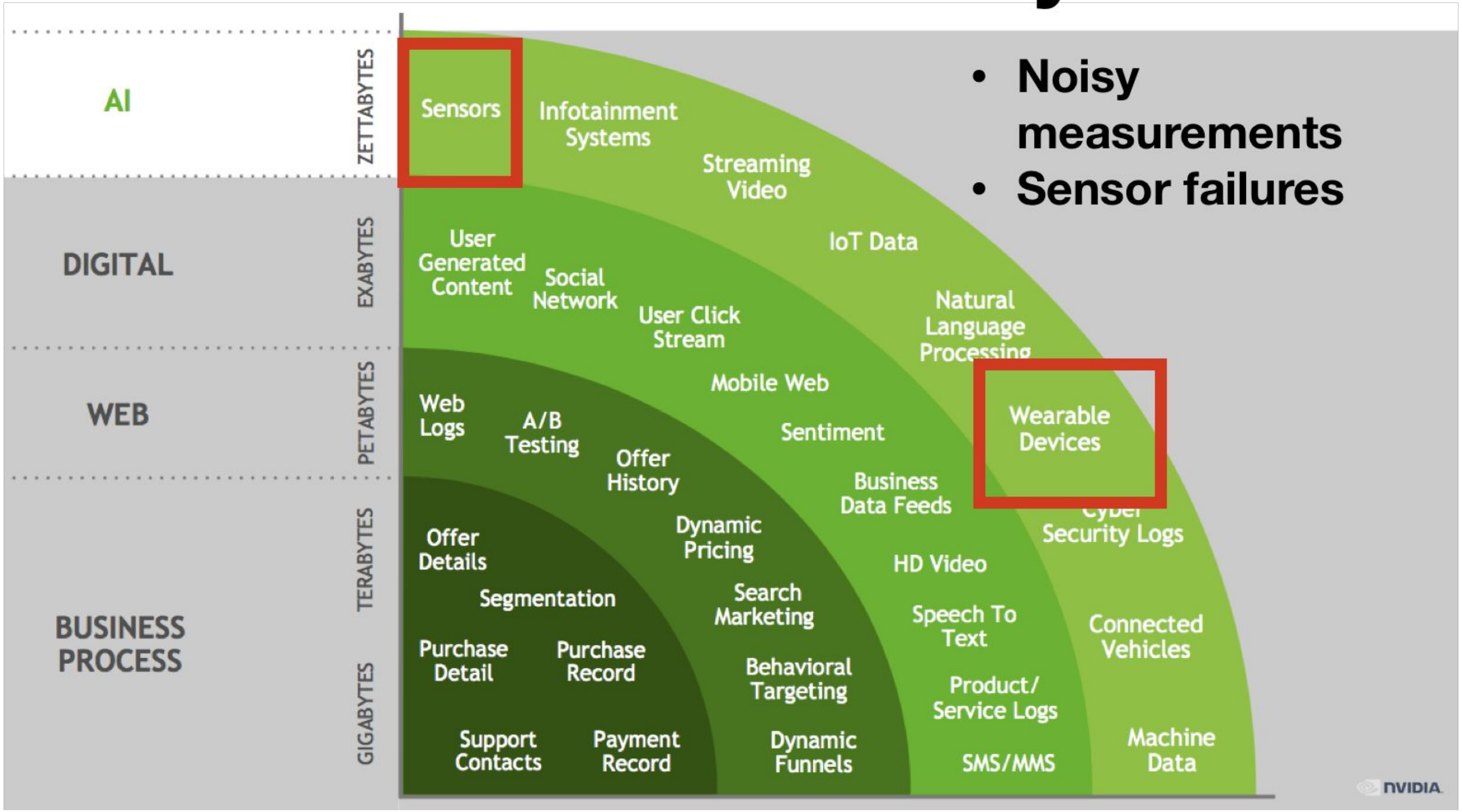
Large collections of curated training data are necessary for progress in ML. We need:

1. **Ensure correctness of the available data**
2. **Generate large volumes of training data**

Outline

- Part I. Introduction
- Part II. ML for DI
- Part III. DI for ML
    - Data cleaning
    - Training data creation
- Part IV. Conclusions and research directions

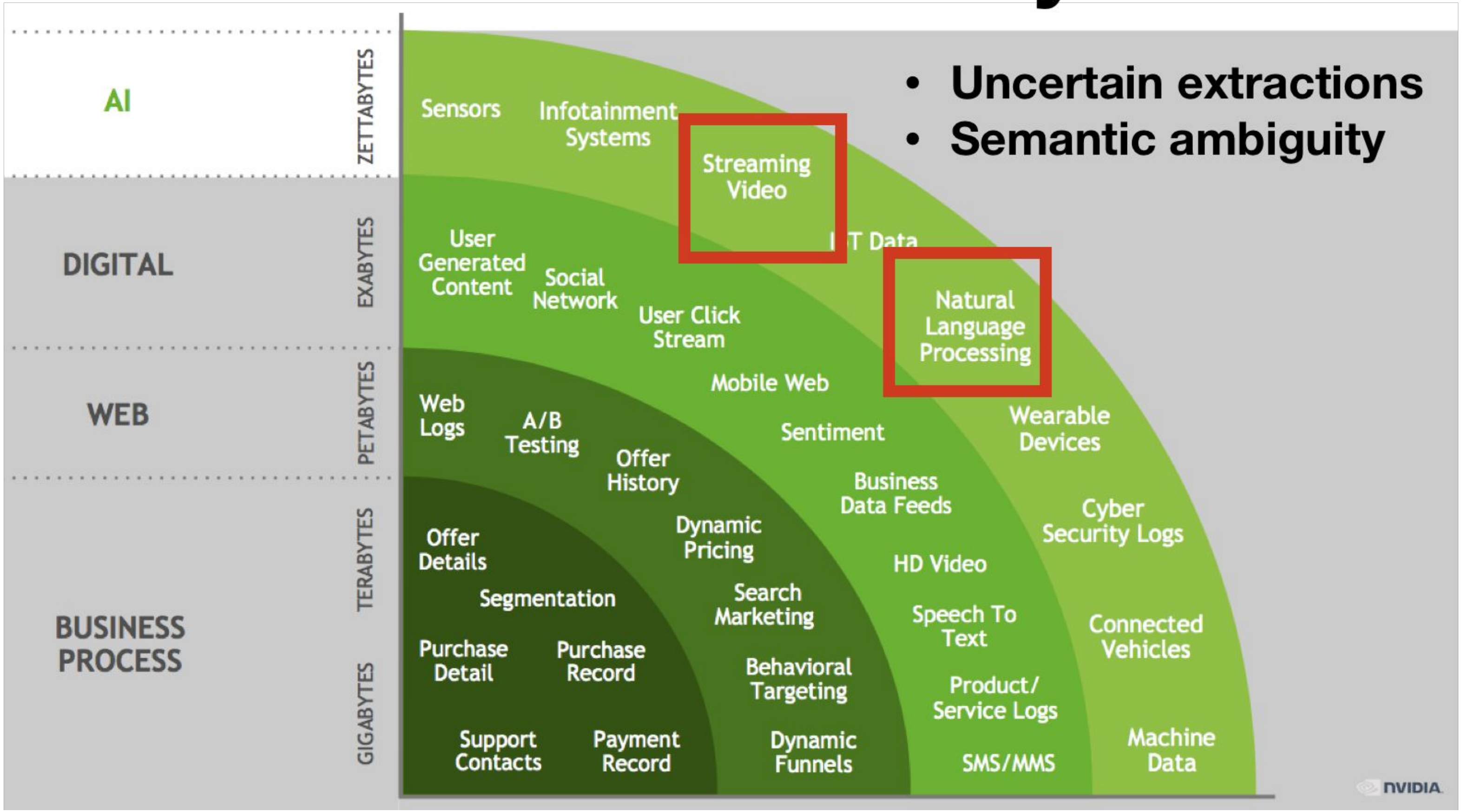
Data errors are everywhere



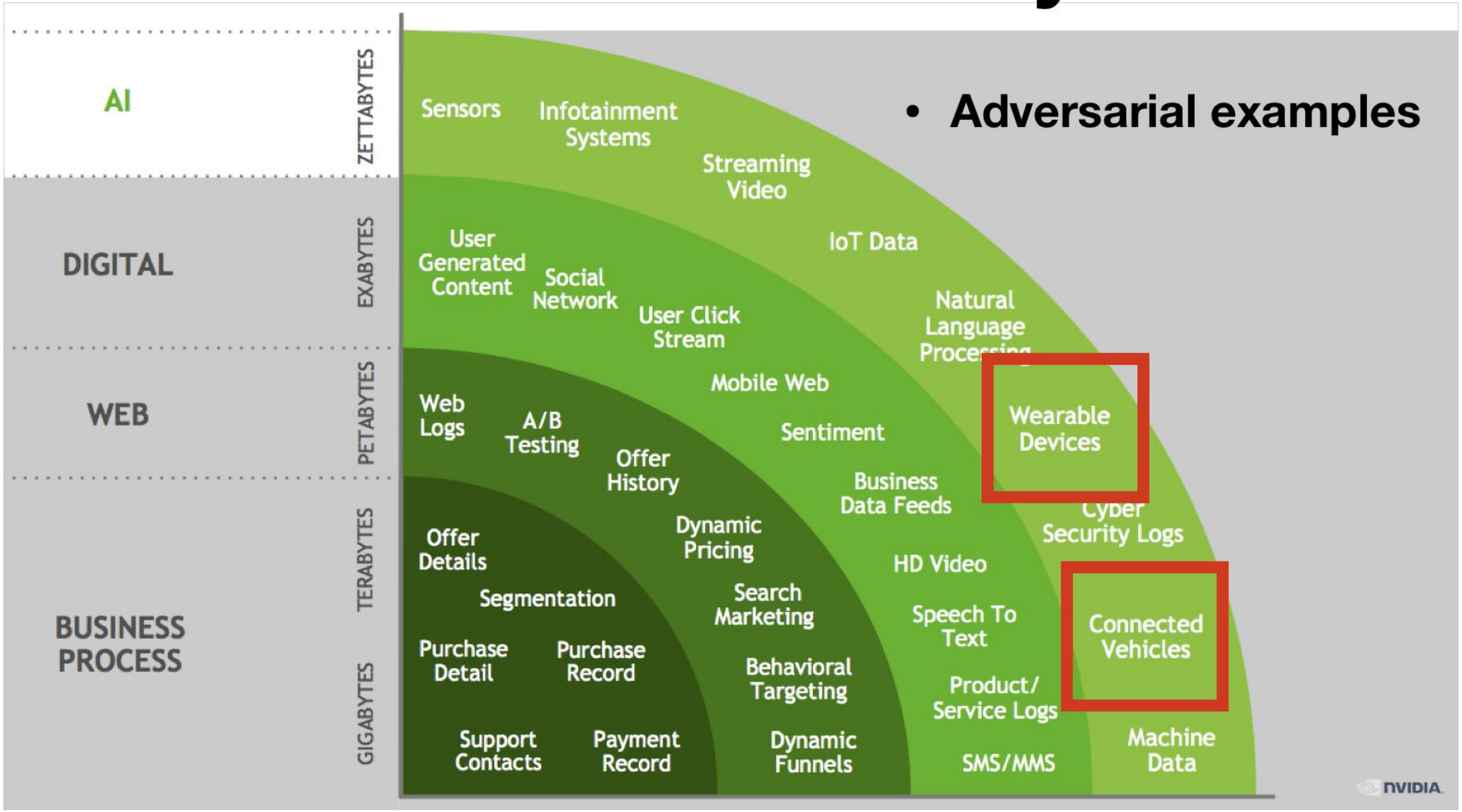
- **Noisy measurements**
- **Sensor failures**

Data errors are everywhere

- **Uncertain extractions**
- **Semantic ambiguity**

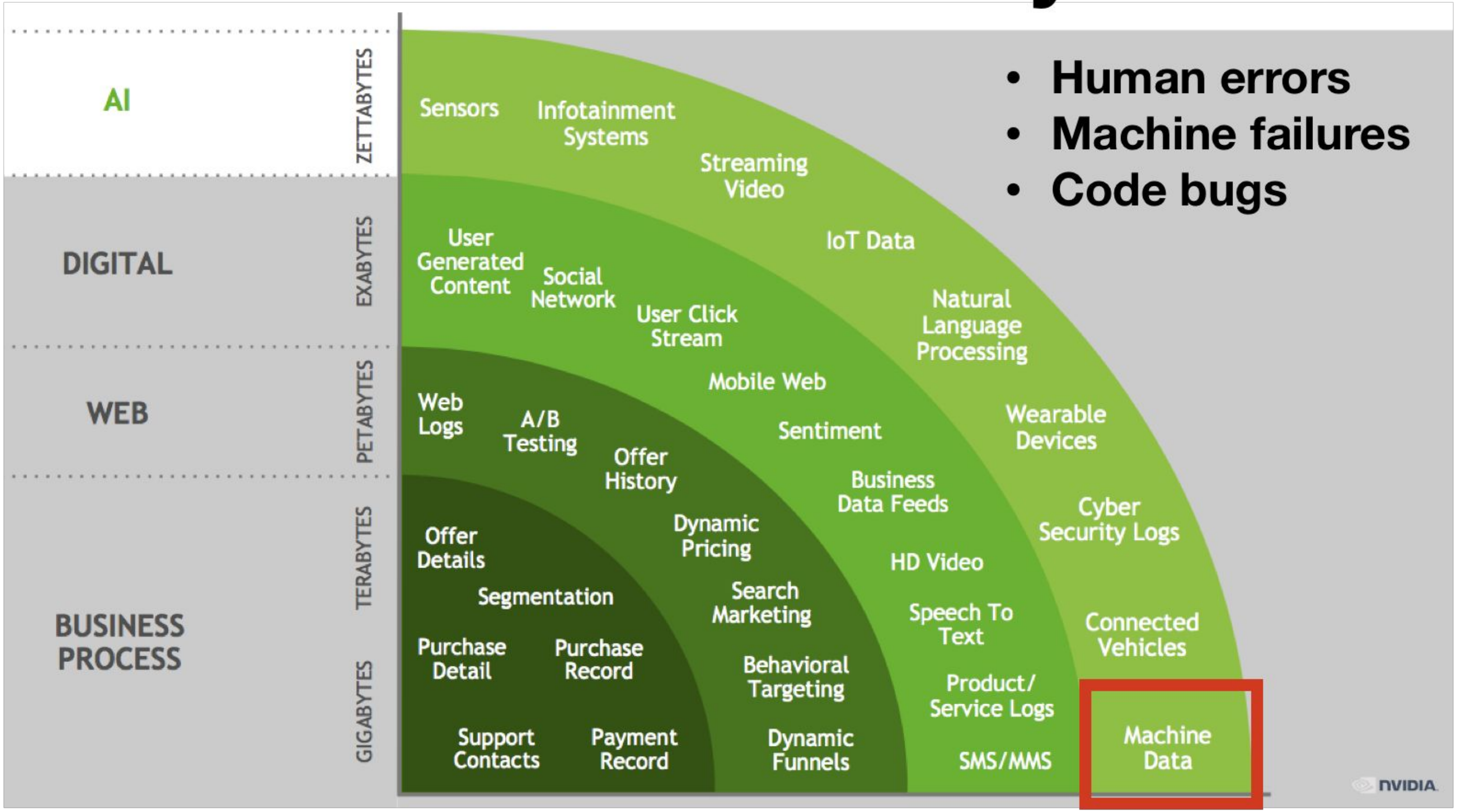


Data errors are everywhere



- **Adversarial examples**

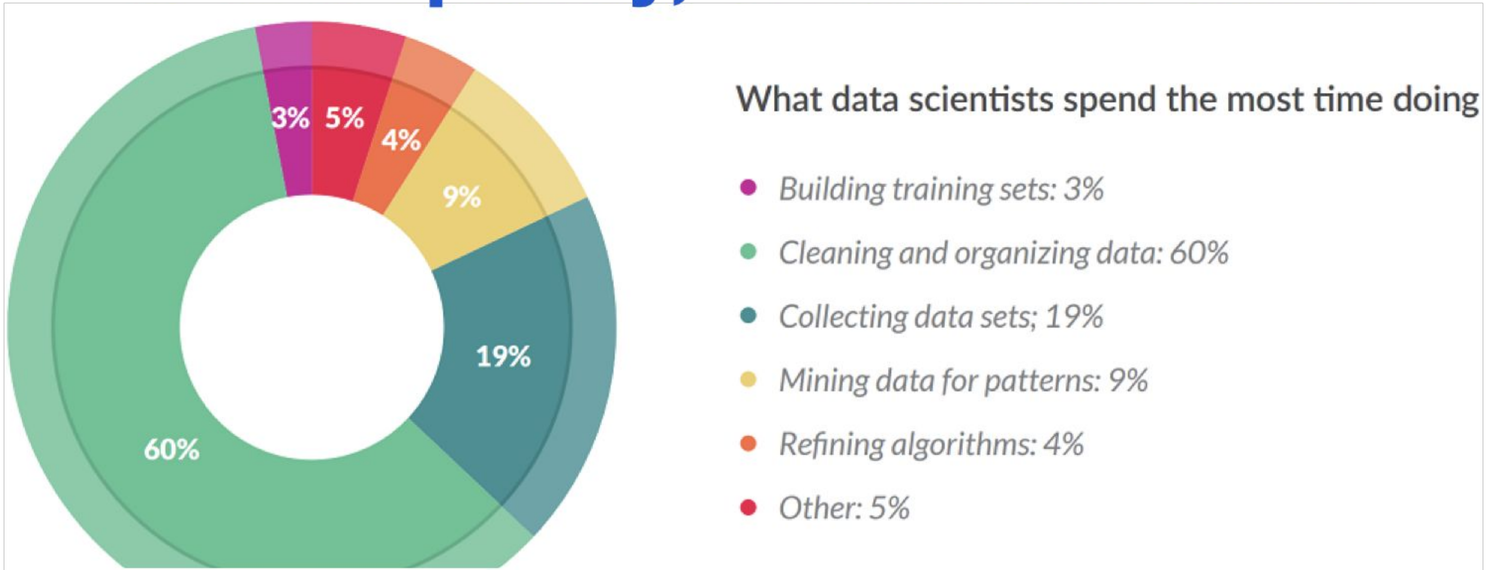
Data errors are everywhere



- Human errors
- Machine failures
- Code bugs

The Achilles' Heel of Modern Analytics

is low quality, erroneous data



Cleaning and organizing the data comprises 60% of the time spent on an analytics or AI project.

A simple example of noisy data

	DBAName	AKAName	Address	City	State	Zip
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	***Chicago***	IL	***60608***
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t4	***Johnnyo's***	Johnnyo's	3465 S Morgan ST	***Cicago***	IL	60608

Conflicts

$c1:\ \text{DBAName} \rightarrow \text{Zip}$

$c2:\ \text{Zip} \rightarrow \text{City, State}$

$c3:\ \text{City, State, Address} \rightarrow \text{Zip}$

Does not obey data distribution

Conflict

Computational problems: ***Detect*** errors, ***repair*** errors, compute "***consistent***" ***query answers***.

50 Years of Data Cleaning

1970s (Nulls)

E. F. Codd

- Understanding relations (installment #7). *FDT - Bulletin of ACM SIGMOD*, 7(3):23–28, 1975.
- Null-related features of DBs

1980s (Normalization)

Integrity Constraints

- Normal forms to reduce redundancy and integrity
- FDs, MVDs etc.

1990s (Warehouses)

Data transforms

- Part of ETL
- Errors within a source and across sources
- Transformation workflows and mapping rules; domain-knowledge is crucial

2000s (Data Repairs)

Constraints and Probabilities

- Dichotomies for consistent query answering
- Minimality-based repairs to obtain consistent instances
- Statistical repairs
- Anomaly detection

The case for inconsistent data

	DBAName	AKAName	Address	City	State	Zip
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	***Chicago***	IL	***60608***
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t4	***Johnnyo's***	Johnnyo's	3465 S Morgan ST	***Cicago***	IL	60608

$$\text{c1: DBAName} \rightarrow \text{Zip}$$
$$\text{c2: Zip} \rightarrow \text{City, State}$$
$$\text{c3: City, State, Address} \rightarrow \text{Zip}$$

An example unclean database *J*

- Errors correspond to tuples/cells that introduce inconsistencies (violations of integrity constraints).
- Inconsistencies are typical in data integration, extract-load-transform workloads, etc.
- **Data repairs:** A theoretical framework for coping with inconsistent databases [Arenas et al. 1999]

Minimal data repairs

Database Repairs

Definition (Arenas, Bertossi, Chomicki – 1999)

Σ a set of integrity constraints and I an inconsistent database.
A database J is a *repair* of I w.r.t. Σ if

- J is a consistent database (i.e., $J \models \Sigma$);
- J differs from I in a minimal way.

Fact

Several different types of repairs have been considered:

- Set-based repairs (subset, superset, $\oplus$-repairs).
- Cardinality-based repairs
- Attribute-based repairs
- Preferred repairs

Slide by Phokion Kolaitis
[SAT 2016]

Plethora of fundamental results on tractability of repair-checking and consistent query answering.

Limited adoption in practice.

Minimal data repairs

	DBAName	AKAName	Address	City	State	Zip
~~t1~~	~~John Veliotis Sr.~~	~~Johnnyo's~~	~~3465 S Morgan ST~~	~~***Chicago***~~	~~IL~~	~~***60608***~~
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t4	***Johnnyo's***	Johnnyo's	3465 S Morgan ST	***Cicago***	IL	60608

$c1: \text{DBAName} \rightarrow \text{Zip}$

$c2: \text{Zip} \rightarrow \text{City, State}$

$c3: \text{City, State, Address} \rightarrow \text{Zip}$

Minimal subset repair:
We remove t1

Several variations of Minimal repairs. E.g., update the minimum number of cells.

Errors remain:

(1) Cicago should clearly be Chicago

(2) Non-obvious errors: 60609 is the wrong Zip

Minimality can be used as an operational principle to prioritize repairs but these repairs are not necessarily correct with respect to the ground truth.

The case for most probable data [Gribkoff et al., 14]

	DBAName	AKAName	Address	City	State	Zip	p
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	***Chicago***	IL	***60608***	0.9
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***	0.4
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***	0.4
t4	***Johnnyo's***	Johnnyo's	3465 S Morgan ST	***Cicago***	IL	60608	0.8

$$\text{c1: DBAName} \rightarrow \text{Zip}$$
$$\text{c2: Zip} \rightarrow \text{City, State}$$
$$\text{c3: City, State, Address} \rightarrow \text{Zip}$$

Most probable world, conditioned on integrity constraint satisfaction

The case for **most probable** data [Gribkoff et al., 14]

	DBAName	AKAName	Address	City	State	Zip	*p*	*Factor (f)*
~~t1~~	~~John Veliotis Sr.~~	~~Johnnyo's~~	3465 S Morgan ST	~~***Chicago***~~	~~IL~~	~~***60608***~~	0.9	1 - 0.9
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***	0.4	0.4
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***	0.4	0.4
t4	***Johnnyo's***	Johnnyo's	3465 S Morgan ST	***Cicago***	IL	60608	0.8	0.8

$\text{c1: DBAName} \rightarrow \text{Zip}$

$\text{c2: Zip} \rightarrow \text{City, State}$

$\text{c3: City, State, Address} \rightarrow \text{Zip}$

Optimization Objective

$$\max_{I} \left(\prod_{t \in I} p(t) \prod_{t \notin I} (1 - p(t)) \right)$$

Most probable repairs

	DBAName	AKAName	Address	City	State	Zip	*p*	*Factor (f)*
~~t1~~	~~John Veliotis Sr.~~	~~Johnnyo's~~	~~3465 S Morgan ST~~	~~***Chicago***~~	~~IL~~	~~***60608***~~	0.9	1 - 0.9
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***	0.4	0.4
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***	0.4	0.4
t4	***Johnnyo's***	Johnnyo's	3465 S Morgan ST	***Cicago***	IL	60608	0.8	0.8

Optimization Objective $$\max_{I} \left(\prod_{t \in I} p(t) \prod_{t \notin I} (1 - p(t)) \right)$$

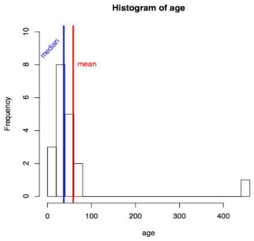
Probabilities offer clear semantics than minimality.
Fundamental question: How do we know p?

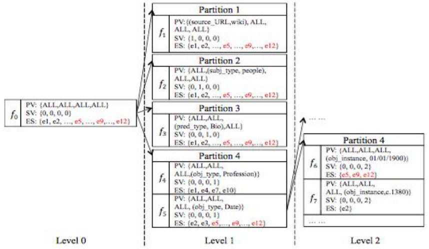
Where are we today?

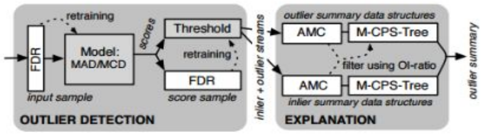
Machine learning and statistical analysis are becoming more prevalent.

Error detection (Diagnosis)

- Anomaly detection [Chandola et al., ACM CSUR, 2009]
- Bayesian analysis (Data X-Ray) [Wang et al., SIGMOD'15]
- Outlier detection over streams (Macrobase) [Bailis et al., SIMGOD'17]
- HoloDetect: Few-shot Learning for Error Detection [Heidari et al., SIGMOD'19]





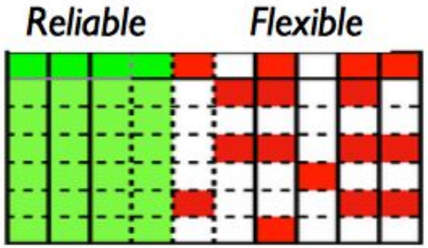


Where are we today?

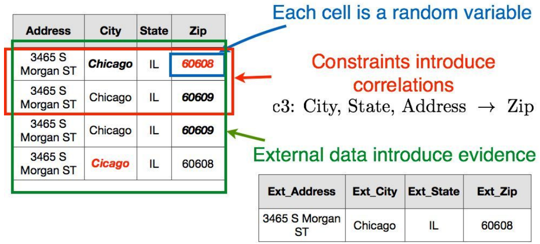
Machine learning and statistical analysis are becoming more prevalent.

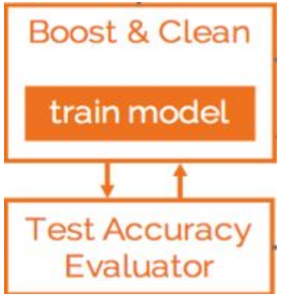
Data Repairing (Treatment)

- Classical ML (SCARE, ERACER) [Yakout et al., VLDB’11, SIGMOD’13, Mayfield et al., SIGMOD’10]
- Boosting [Krishan et al., 2017]
- Weakly-supervised ML (HoloClean) [Rekatsinas et al., VLDB’17]



HoloClean





Question:

What is an appropriate (formal) framework for managing noisy data?

Things to consider:

Simplicity and generality

The case of a noisy channel for data

Noisy Channel Model

1. We see an observation *x* in the noisy world
2. Find the correct world *w*

$$\hat{w} = \arg\max_{w \in W} P(w \mid x)$$

Applications: *Speech, OCR, Spelling correction, Part of speech tagging, machine translations, etc…*

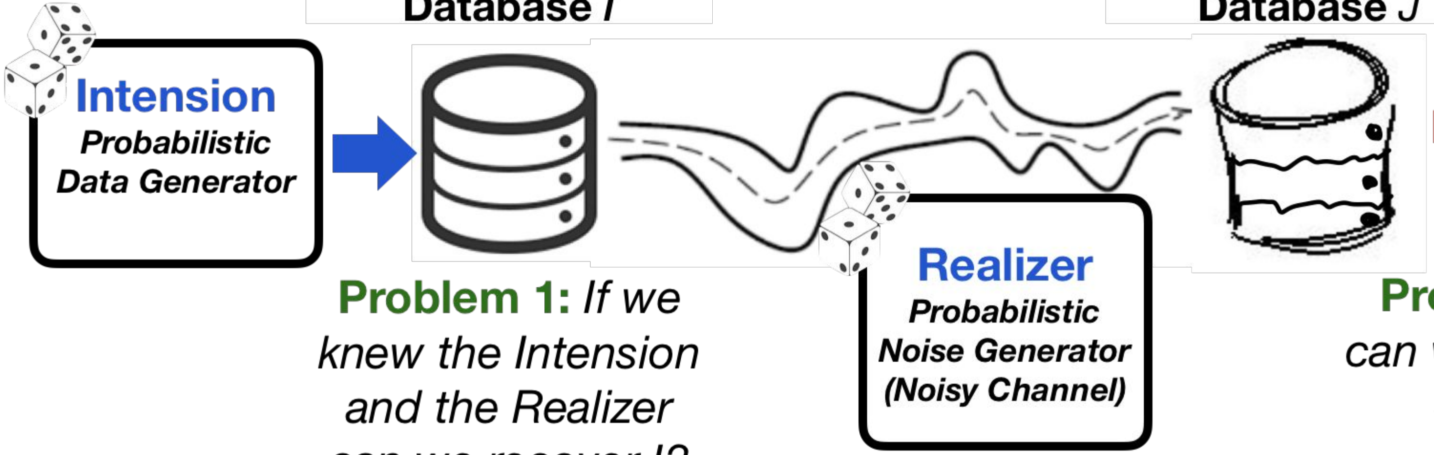
The Probabilistic Unclean Database Model

Problem 3: *Can we learn the Intension and the Realizer? Can we do that from J (i.e., **without any training data**)?*

Output: *An estimate for the Intension and the Realizer*

Clean Intended Database *I*

Observed Unclean Database *J*



Input: *We only observe this*

Problem 1: *If we knew the Intension and the Realizer can we recover I?*

Output: *An estimate of the most probable I*

Problem 2: *Given J can we answer a query on I correctly?*

Output: $Pr(a \in Q(I) \mid J)$

A Series of Theoretical Results

Complexity Results: When is data cleaning efficient? [De Sa et al., ICDT 2019]

Statistical Recovery Results: New theoretical results on the hardness of structured prediction under noisy data and new structured prediction methods for automated data cleaning with low-error guarantees [Heidari, Ilyas, Rekatsinas UAI, 2019]

Learnability Results: Learning the intended data distribution without any training data [De Sa et al., ICDT 2019]

From Theory to Systems

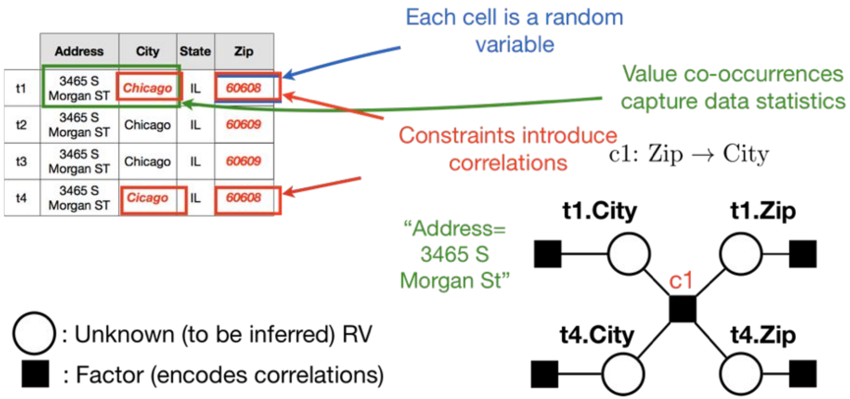
Is the PUDs framework useful in practice?

HoloClean: Probabilistic Data Repairs

HoloClean is the first practical probabilistic data repairing engine and a state-of-the-art data repairing system

HoloClean's factor-graph model is an instantiation of the PUDs Intention model.

HoloClean uses clean cells as training data to learn its PUD Intention model and uses the learned model to approximate MLI repairs.



Reference: "HoloClean: Holistic Data Repairs with Probabilistic Inference"
Rekatsinas, Chu, Ilyas, Ré, VLDB 2017

HoloClean: Probabilistic Data Repairs

Challenge: Inference under constraints is #P-complete

Applying probabilistic inference naively does not scale to data cleaning instances with millions of tuples

Idea 1: Prune domain of random variables.

Idea 2: Relax constraints over sets of random variables to features over independent random variables.

Relaxing constraints

Tuple ID	University	State
t1	U of Chicago	IL
t2	U of Chicago	IL
t3	U of Chicago	CA

Functional dependency:

$$\text{University} \rightarrow \text{State}$$

"The same University must be in the same State"

Relax constraints to features over independent RVs (corresponds to a voting model)

Example:

$$\text{t1.University} = \text{U of Chicago} \implies \text{IL} = \text{CA}$$

$$\text{U of Chicago} = \text{t3.University} \implies \text{IL} = \text{CA}$$

$$\text{U of Chicago} = \text{U of Chicago} \implies \text{t1.State} = \text{CA}$$

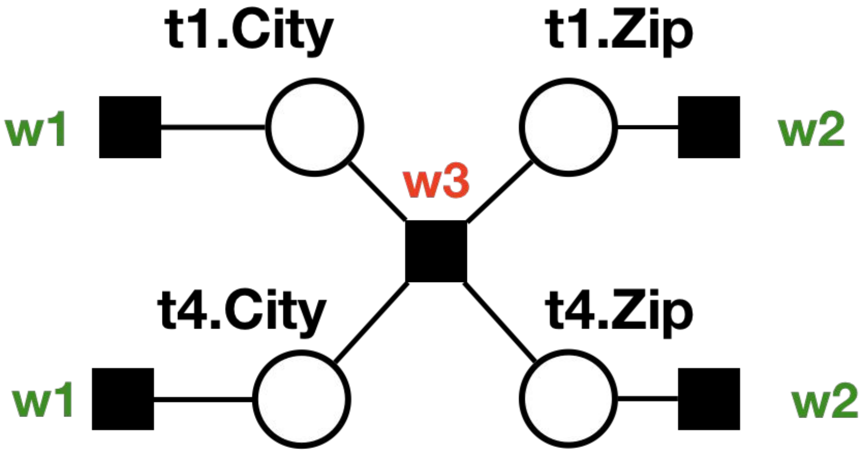
$$\text{U of Chicago} = \text{U of Chicago} \implies \text{IL} = \text{t3.State}$$

Only $4D$ possible worlds considered

HoloCleans' locally consistent model introduces features over independent random variables.

Relaxing constraints

	Address	City	State	Zip
t1	3465 S Morgan ST	*Chicago*	IL	*60608*
t2	3465 S Morgan ST	Chicago	IL	*60609*
t3	3465 S Morgan ST	Chicago	IL	*60609*
t4	3465 S Morgan ST	*Cicago*	IL	*60608*



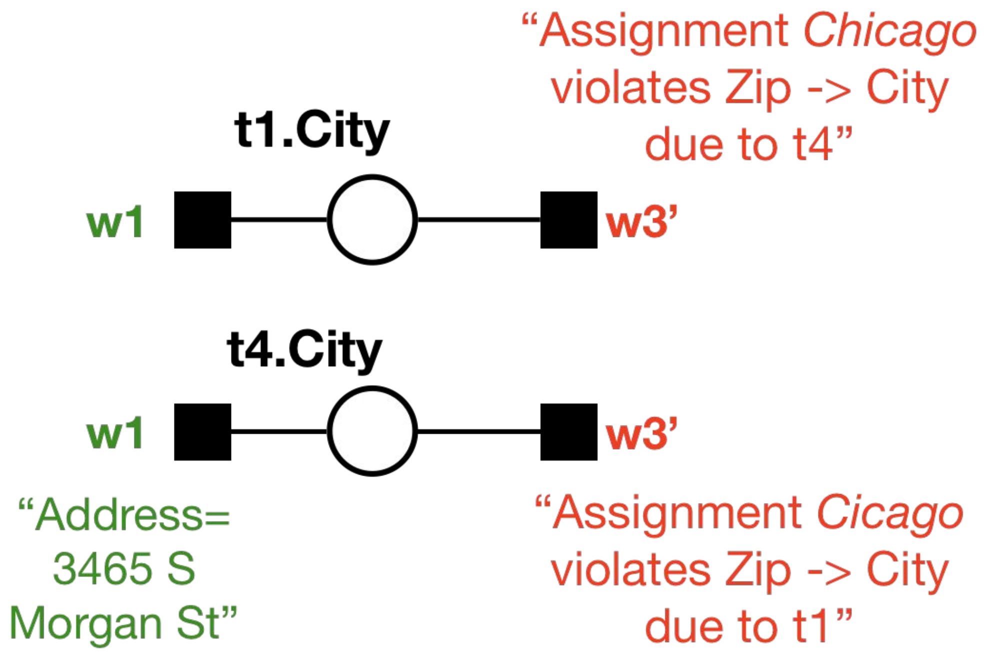
"Address= 3465 S Morgan St"

"Zip -> City"

"Address= 3465 S Morgan St"

Relaxing constraints

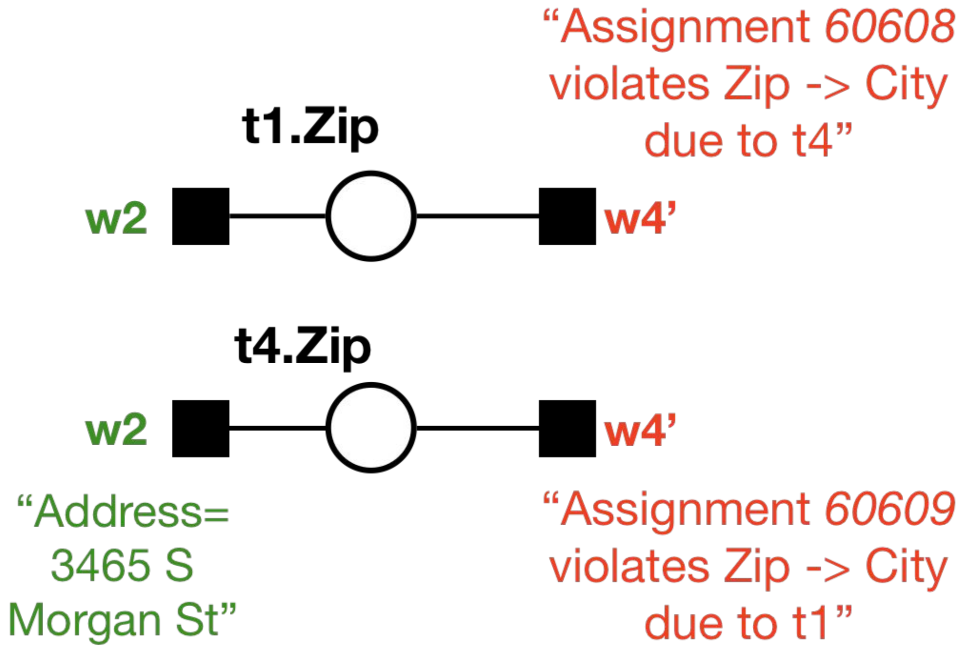
	Address	City	State	Zip
t1	3465 S Morgan ST	***Chicago***	IL	*60608*
t2	3465 S Morgan ST	Chicago	IL	*60609*
t3	3465 S Morgan ST	Chicago	IL	*60609*
t4	3465 S Morgan ST	***Cicago***	IL	*60608*



We have one ***relaxed factor*** for each value in the domain of the RV

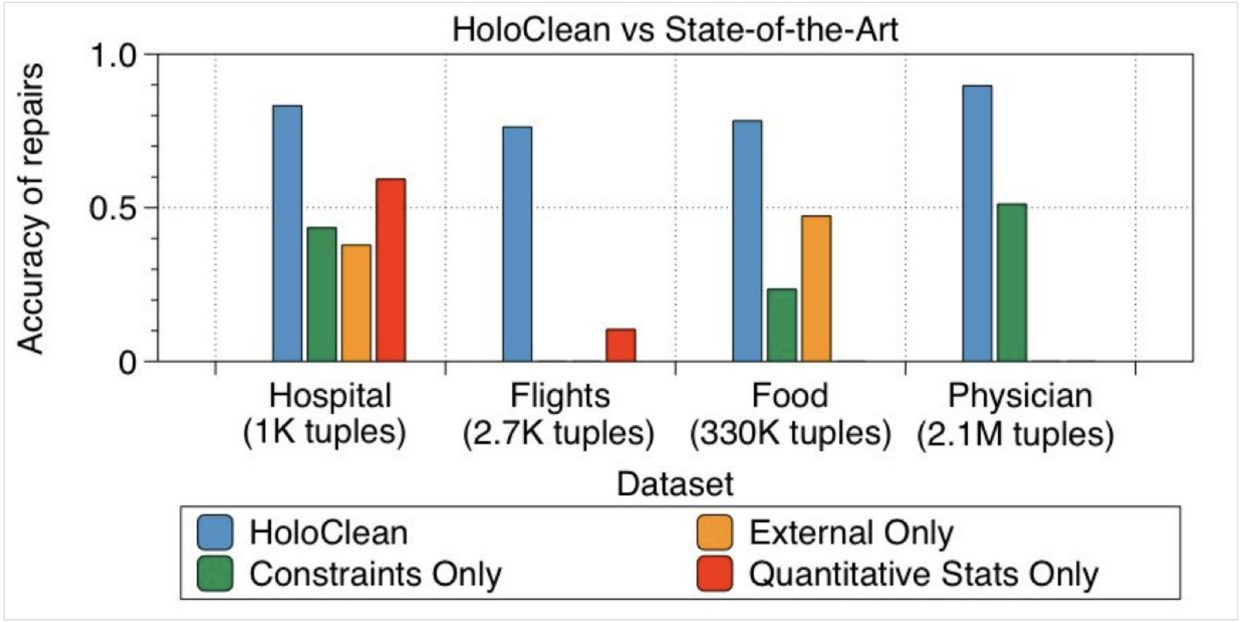
Relaxing constraints

	Address	City	State	Zip
t1	3465 S Morgan ST	***Chicago***	IL	***60608***
t2	3465 S Morgan ST	Chicago	IL	***60609***
t3	3465 S Morgan ST	Chicago	IL	***60609***
t4	3465 S Morgan ST	***Cicago***	IL	***60608***



We have one ***relaxed factor*** for each value in the domain of the RV

HoloClean in practice



Competing methods do not scale or perform correct repairs.

HoloClean: our approach combining all signals and using inference

Holistic[Chu,2013]: state-of-the-art for constraints & minimality

KATARA[Chu,2015]: state-of-the-art for external data

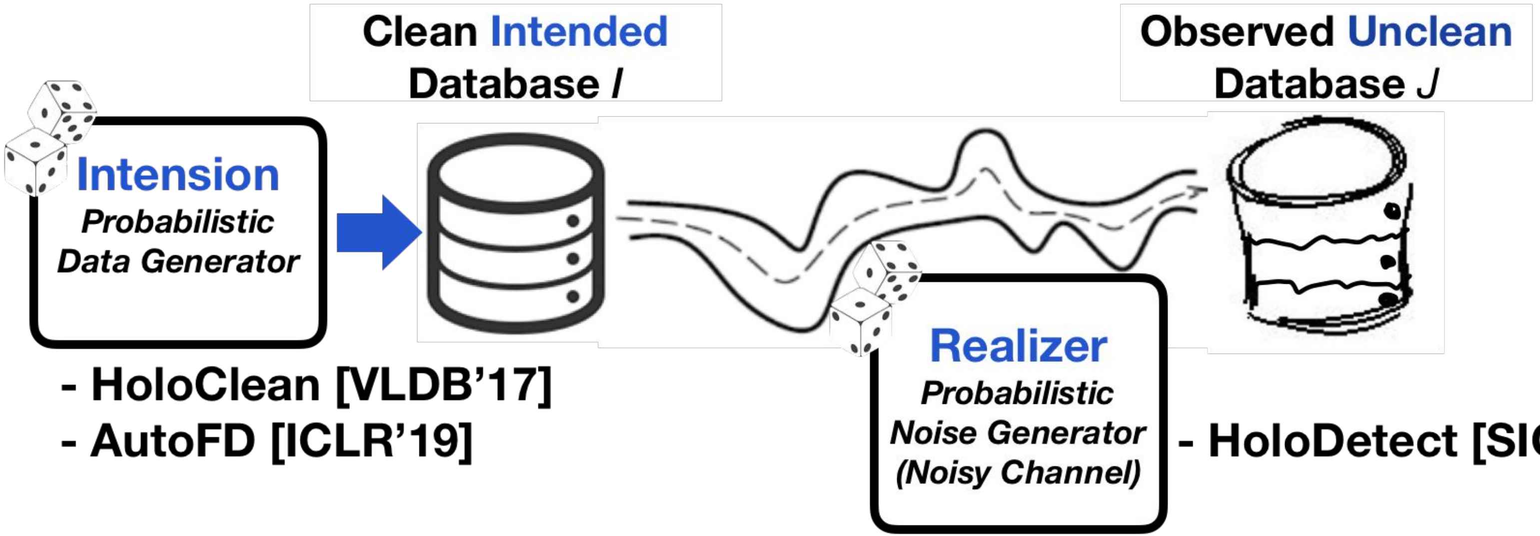
SCARE[Yakout,2013]: state-of-the-art ML & qualitative statistics

HoloClean

Code available at:

http://www.holoclean.io

The Probabilistic Unclean Database Model

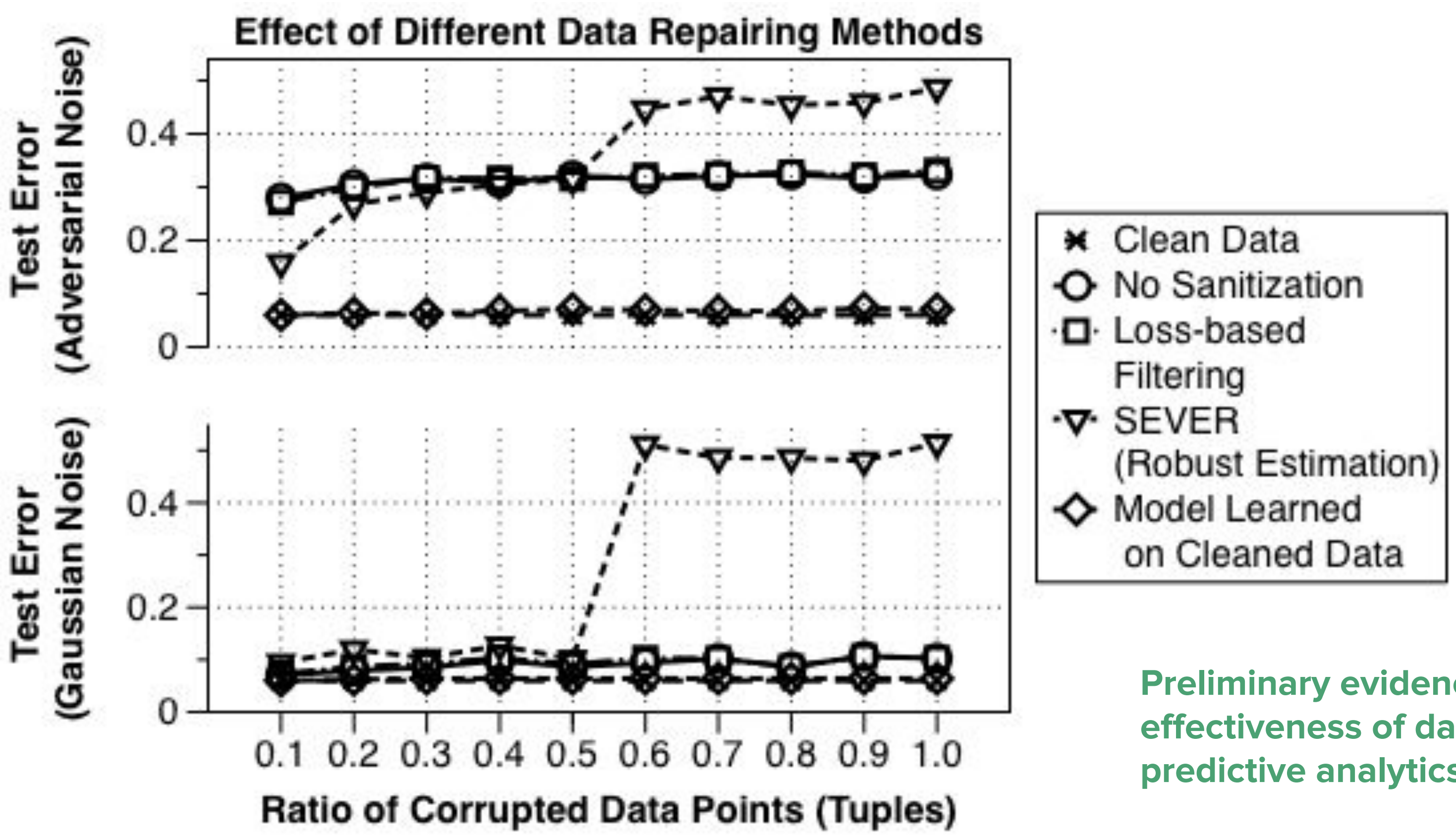


- HoloClean [VLDB'17]
- AutoFD [ICLR'19]

- HoloDetect [SIGMOD'19]

A formal noisy channel model that leads to new insights for managing noisy data and has immediate practical applications to data cleaning systems.

On the Interplay of Cleaning and ML



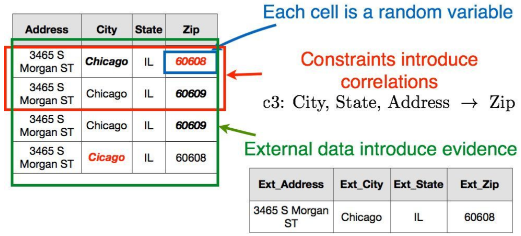
Preliminary evidence on the the effectiveness of data cleaning for predictive analytics.

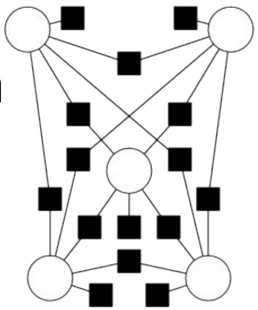
Challenges in Data Cleaning

- More research is needed on understanding when automated solutions are possible and what is the most effective way to bring humans in the loop.
- We need to study the interplay between data cleaning and machine learning closer. Especially in the presence of robust optimization methods.
- We need interpretable data cleaning solutions. Why should I trust the repairs?
- Few end-to-end solutions. Data cleaning workloads (mixed relational and statistical workloads) pose unique scalability challenges.

Recipe for Data Cleaning

- Problem definition: **Detect and repair erroneous data.**
- Short answers
  - **ML can help partly-automate cleaning. Domain-expertise is still required.**
  - **Scalability of ML-based data cleaning methods is a pressing challenge. Exciting systems research!**
  - **We need more end-to-end systems (interpretable, human-in-the-loop, optimized for analytical tasks)!**





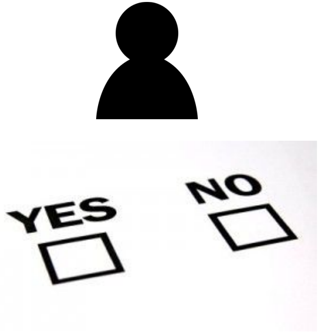
Outline

- Part I. Introduction
- Part II. ML for DI
- Part III. DI for ML
  - Data cleaning
  - Training data creation
- Part IV. Conclusions and research directions

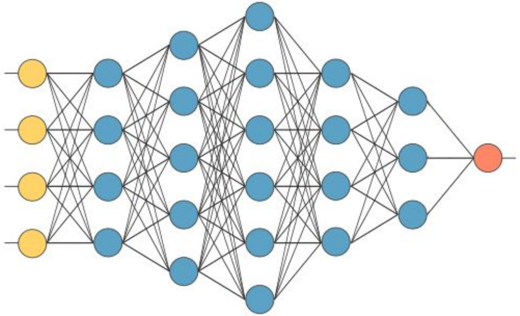
The ML Pipeline in the Deep Learning Era

Data Collection

Data Labeling



Representation Learning and Training



A core pain point today, lots of time spent in labeling data.

Training Data: Challenges and Opportunities

- Collecting training data is **expensive** and **slow.**
- We are overfitting to our training data. [Recht et al., 2018]
  - Hand-labeled training data does not change
- Training data is the point to inject domain knowledge
  - Modern ML is too complex to hand-tune features and priors

Training Data: Challenges and Opportunities

- Collecting training data is **expensive** and **slow.**
- We are overfitting to our training data. [Recht et al., 2018]
  - Hand-labeled training data does not change
- Training data is the point to inject domain knowledge
  - Modern ML is too complex to hand-tune features and priors

How do we get training data more effectively?

The Rise of Weak Supervision

Definition: Supervision with noisy (much easier to collect) labels; prediction on a larger set, and then training of a model.

Semi-supervised learning and ensemble learning

Examples:

- use of non-expert labelers (crowdsourcing),
- use of curated catalogs (distant supervision)
- use of heuristic rules (labeling functions)

The Rise of Weak Supervision

- Alexa – Customer embrace of Alexa continues, with Alexa-enabled devices among the best-selling items across all of Amazon. We're seeing extremely strong adoption by other companies and developers that want to create their own experiences with Alexa. There are now more than 30,000 skills for Alexa from outside developers, and customers can control more than 4,000 smart home devices from 1,200 unique brands with Alexa. The foundations of Alexa continue to get smarter every day too. We've developed and implemented an on-device fingerprinting technique, which keeps your device from waking up when it hears an Alexa commercial on TV. (This technology ensured that our Alexa Super Bowl commercial didn't wake up millions of devices.) Far-field speech recognition (already very good) has improved by 15% over the last year; and in the U.S., U.K., and Germany, we've improved Alexa's spoken language understanding by more than 2[illegible]% over the last 12 months through enhancements in Alexa's machine learning components and the use of semi-supervised learning techniques. (These semi-supervised learning techniques reduced the amount of labeled data needed to achieve the same accuracy improvement by 40 times!) Finally, we've dramatically reduced the amount of time required to teach Alexa new languag[illegible] by using machine translation and transfer learning techniques, which allows us to serve customers in more countries (like India and Japan).

The Rise of Weak Supervision

Definition: Supervision with noisy (much easier to collect) labels; prediction on a larger set, and then training of a model.

Related to semi-supervised learning and ensemble learning

Examples: use of non-expert labelers (crowdsourcing), use of curated catalogs (distant supervision), use of heuristic rules (labeling functions)

Methods developed to tackle data integration problems are closely related to weak supervision.

Learning from Crowds [Raykar et al., JMLR'10]

Setup: Supervised learning but instead of gold groundtruth one has access to multiple annotators providing (possibly noisy) labels (no absolute gold standard).

Task: Learn a classifier from multiple noisy labels.

Closely related to Dawid-Skene!

Difference: Estimating the ground truth and the annotator performance is a byproduct here. Goal is to learn a classifier.

Learning from Crowds [Raykar et al., JMLR'10]

$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$

Example Task: Binary classification

N examples, with labels $\mathbf{y}_i = y_i^1, \ldots, y_I^R$ provided by R different annotators

Learning from Crowds [Raykar et al., JMLR'10]

$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$

N examples, with labels $\mathbf{y}_i = y_i^1, \ldots, y_I^R$

provided by R different annotators

Example Task: Binary classification

Annotator performance:

Sensitivity (true positive rate)

$$\alpha^j = \Pr[y^j = 1 | y = 1]$$

Specificity (1 - false positive rate)

$$\beta^j = \Pr[y^j = 0 | y = 0]$$

Learning from Crowds [Raykar et al., JMLR'10]

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$$

N examples, with labels $\mathbf{y}_i = y_i^1, \ldots, y_I^R$

provided by R different annotators

Example Task: Binary classification

Annotator performance:

Sensitivity (true positive rate)

$$\alpha^j = \Pr[y^j = 1 | y = 1]$$

Specificity (1 - false positive rate)

$$\beta^j = \Pr[y^j = 0 | y = 0]$$

Learning: $\Pr[\mathcal{D}|\theta] = \prod_{i=1}^{N} [a_i p_i + b_i(1 - p_i)]$

$$\boxed{p_i := \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i).}$$

$$a_i := \prod_{j=1}^{R} [\alpha^j]^{y_i^j} [1 - \alpha^j]^{1 - y_i^j}.$$

$$b_i := \prod_{j=1}^{R} [\beta^j]^{1 - y_i^j} [1 - \beta^j]^{y_i^j}.$$

Model parameters {**w**, ***α***, ***β***}

EM algorithm to obtain maximum-likelihood estimates.

Difference with Dawid-Skene is the estimation of *w*.

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

Goal: Extracting structured knowledge from text.

Hypothesis: If two entities belong to a certain relation, any sentence containing those two entities is likely to express that relation.

Idea: Use a *database* of relations to gets lots of *noisy* training examples

- Instead of hand-creating seed tuples (bootstrapping)
- Instead of using hand-labeled corpus (supervised)

Benefits: has the advantages of supervised learning (leverage reliable hand-created knowledge), has the advantages of unsupervised learning (leverage unlimited amounts of text data).

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

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 …

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

[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]

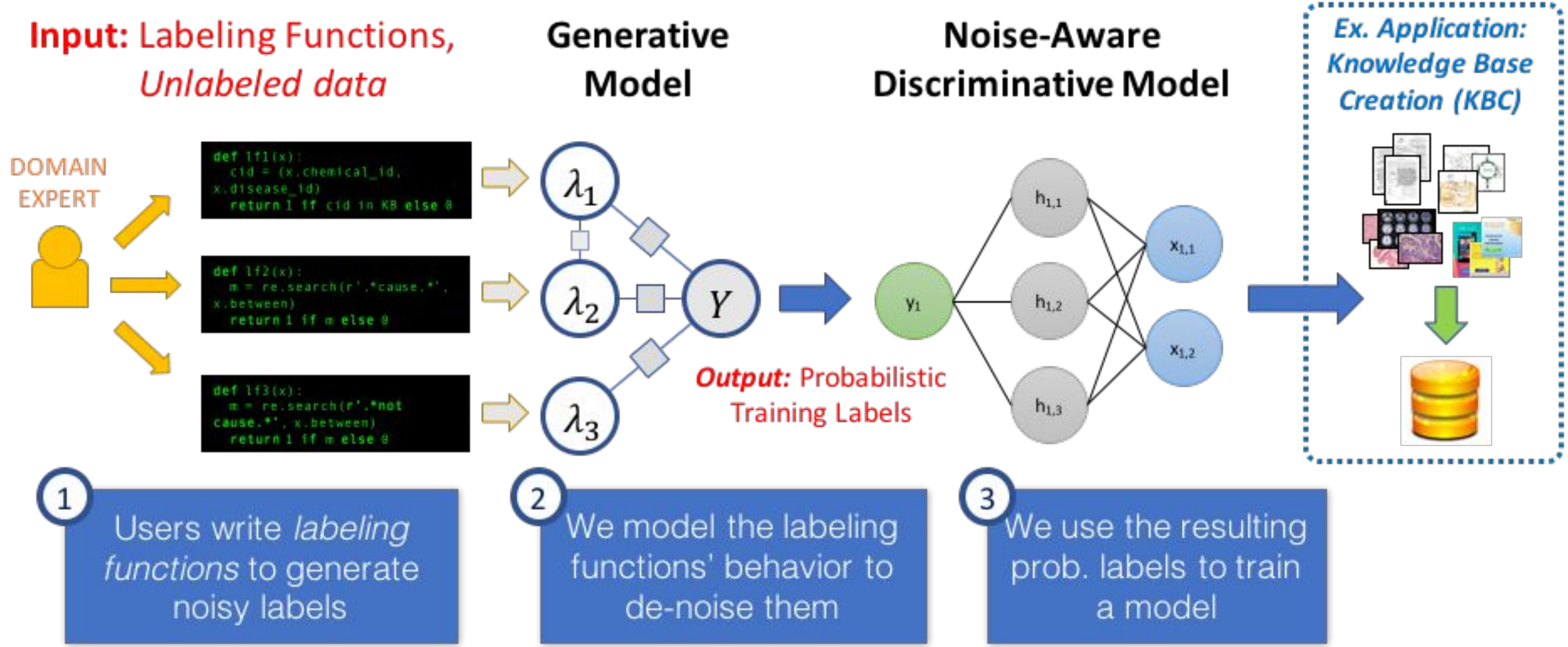
Distant Supervision [Mintz et al., ACL'09]

Entity Linking is an inherent problem in Distant Supervision.

The quality of matches can vary significantly and has a direct effect on extraction quality.

Relation	Freebase Matches	
	#sents	% true
/business/person/company	302	89.0
/people/person/place_lived	450	60.0
/location/location/contains	2793	51.0
/business/company/founders	95	48.4
/people/person/nationality	723	41.0
/location/neighborhood/neighborhood_of	68	39.7
/people/person/children	30	80.0
/people/deceased_person/place_of_death	68	22.1
/people/person/place_of_birth	162	12.0
/location/country/administrative_divisions	424	0.2

Snorkel: Code as Supervision [Ratner et al., NIPS'16, VLDB'18]



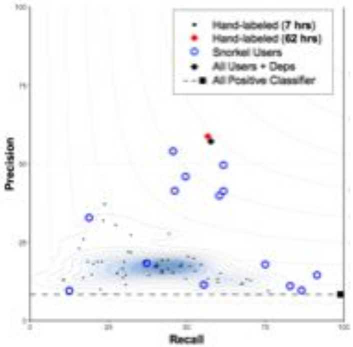
[Slide by Alex Ratner]

Snorkel: Code as Supervision [Ratner et al., NIPS'16, VLDB'18]

Snorkel biomedical workshop in collaboration with the NIH Mobilize Center

15 companies and research groups attended

How well did these new Snorkel users do?



71% New Snorkel users matched or beat 7 hours of hand-labeling

2.8x Faster than hand-labeling data

45.5% Average improvement in model performance

3rd Place Score

No machine learning experience

Beginner-level Python

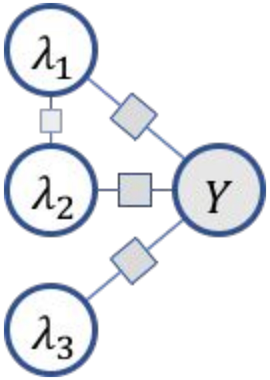
[Slide by Alex Ratner]

Challenges in Creating Training Data

- Richly-formatted data is still a challenge. How can attack weak supervision when data includes images, text, tables, video, etc.?
- Combining weak supervision with other data enrichment techniques such as data augmentation is an exciting direction. How can reinforcement learning help here (http://goo.gl/K2qopQ)?
- How can we combine weak supervision with techniques from semi-supervised?

Recipe for Creating Training Data

- Problem definition: **Go beyond gold labels to noisy training data.**
- Short answers
  - **Transition from “gold” labels to “high-confidence” labels.**
  - **Modeling error rates is key. The notion of *data source* is different.**
  - **Need for debugging tools, bias detection, and recommendations of weak supervision signals.**



Outline

- Part I. Introduction
- Part II. ML for DI
- Part III. DI for ML
  - Creating training data
  - Data cleaning
- Part IV. Conclusions and research direction

DI and ML: A Natural Synergy

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

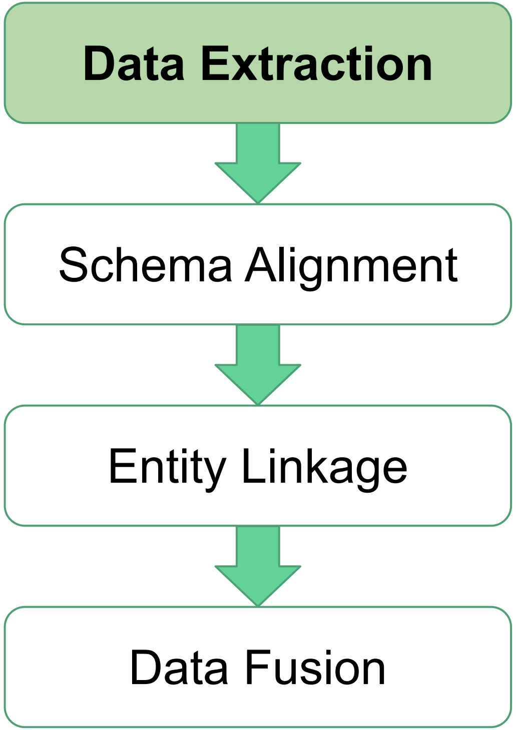
- Transition from logic to probabilities revolutionized data integration
  - Probabilities allow us to reason about inherently noisy data
  - Similar to the AI-revolution in the 80s [https://vimeo.com/48195434]

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

Revisit: Recipe for Data Extraction

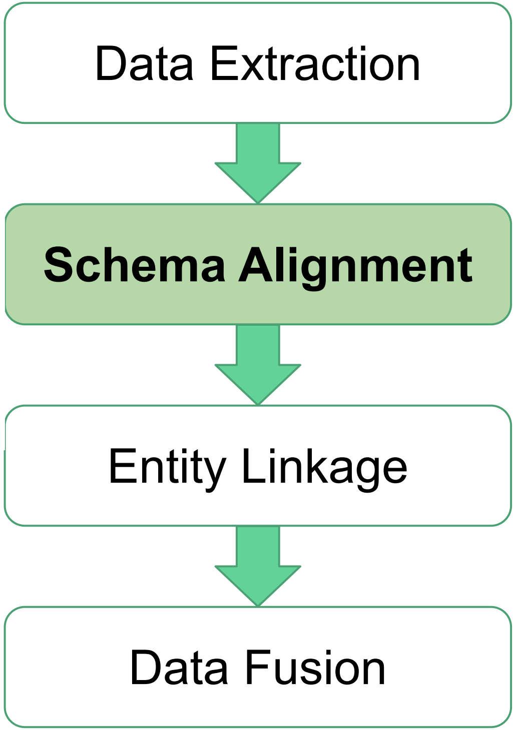
- Problem definition: **Extract structure from semi- or un-structured data**
- Short answers
  - **Wrapper induction has high prec/rec**

  - **Distant supervision is critical for collecting training data**
  - **DL effective for texts and LR is often effective for semi-stru data**



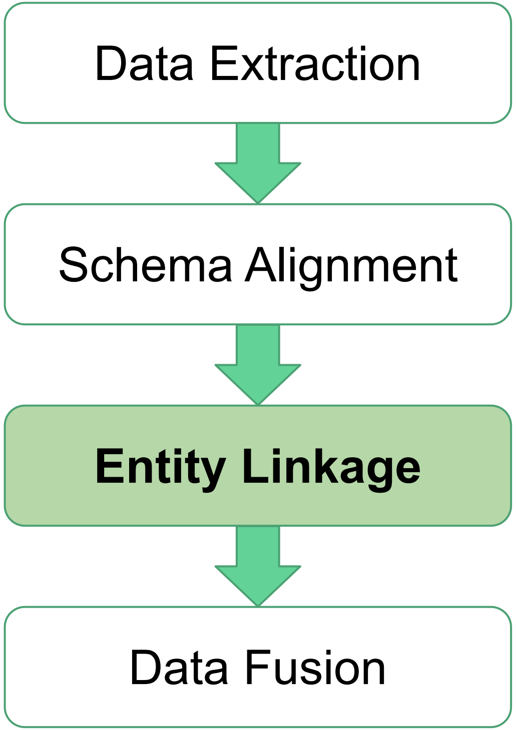
Revisit: Recipe for Schema Alignment

- 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**



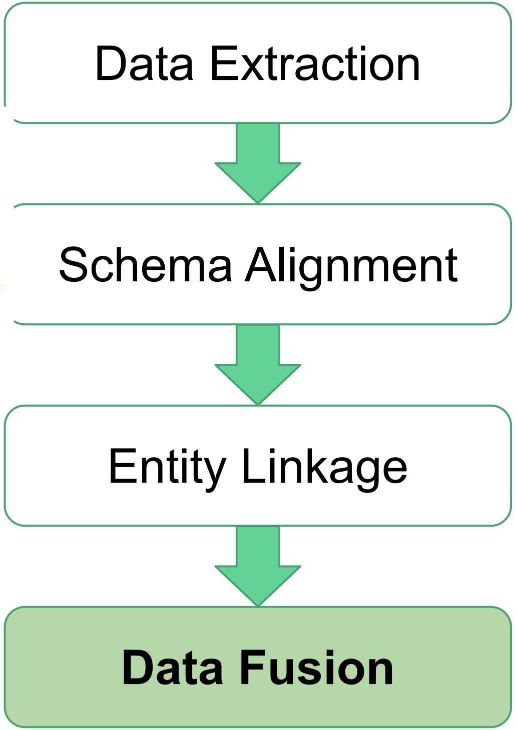
Revisit: Recipe for Entity Linkage

- 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.**



DI and ML: A Natural Synergy

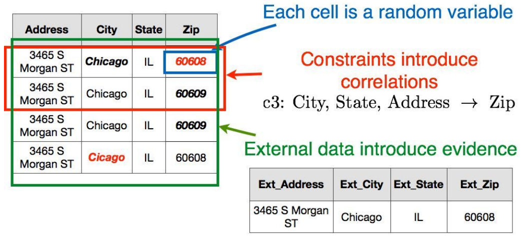
- Data is bottleneck of modern ML and AI applications

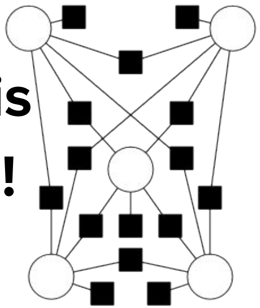
- DI-related methods and algorithms have revolutionized the way supervision is performed.
    - Weak supervision signals are integrated into training datasets

- Data integration solutions (e.g., data cataloging solutions) can lead to cheaper collection of training data and more effective data enrichment

Recipe for Data Cleaning

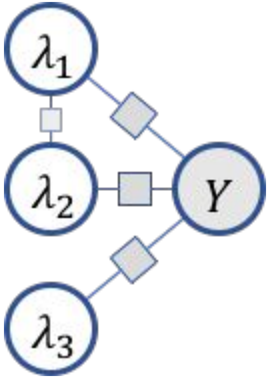
- Problem definition: **Detect and repair erroneous data.**
- Short answers
    - **ML can help partly-automate cleaning. Domain-expertise is still required.**
    - **Scalability of ML-based data cleaning methods is a pressing challenge. Exciting systems research!**
    - **We need more end-to-end systems!**





Revisit: Recipe for Creating Training Data

- Problem definition: **Go beyond gold labels to noisy training data.**
- Short answers
    - **Transition from “gold” labels to “high-confidence” labels.**
    - **Modeling error rates is key. The notion of *data source* is different.**
    - **Need for debugging tools, bias detection, and recommendations of weak supervision signals.**



Opportunities for DI

One System vs. An Ecosystem: Every RBMS is a monolithic system. This paradigm has failed for DI. Tools for different DI tasks are prevalent. We need abstractions and execution frameworks for such ecosystems.

Humans-in-the-loop: DI tasks can be very complex. Is weak supervision the right approach to inject domain knowledge? What about quality evaluation?

Multi-modal DI: ML-based DI has focused on structured data with the exception of DI over images using crowdsourcing and some recent efforts that target textual data. DL is the de facto solution to reasoning about high dimensional data. Can is help develop unified DI solutions for visual, textual, and structured data?

Efficient Model Serving: This means efficient model serving. Many compute-intensive operations such as normalization and blocking are required. Featurization may also rely on compute-heavy tasks (e.g., computing string similarity). What is the role of pipelining and RDBMS-style optimizations?

Opportunities for ML

Data Catalogs: Data augmentation relies on data transformations performed on data records in a single dataset. How can we leverage data catalogs and data hubs to enable data augmentation go beyond a single dataset?

Robust/Valuable Data for ML applications: The DB community has focused on assessing the value of data [Dong et al., VLDB'12, Koutris et al., JACM 2015]. These ideas are not pervasive to ML but if ML is to become a commodity [Jordan, 2018] we need methods to reason about the value of data.

DI for Benchmarks: Increasing efforts on creating manually curated benchmarks for ML. Current efforts rely on manual collection and curation. How can we leverage meta-data and existing DI solutions to automate such efforts?

"How reliable are our current measures of progress in machine learning?"
Do CIFAR-10 Classifiers Generalize to CIFAR-10?, Ben Recht et al., 2018

MLPerf

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 & 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

Thank you!

References Part I: Introduction

Bengio, Y., Goodfellow, I.J. & Courville, A., 2015. Deep learning. *Nature*, 521(7553), pp.436–444.

Bishop, C.M., 2016. *Pattern Recognition and Machine Learning*, Springer New York.

Doan, A., Halevy, A.Y. & Ives, Z.G., 2012. *Principles of Data Integration*, Morgan Kaufmann.

Domingos, P., 2012. A Few Useful Things to Know About Machine Learning. *Communications of the ACM*, 55(10), pp.78–87.

Dong, X. et al., 2014. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '14. New York, NY, USA: ACM, pp. 601–610.

Dong, X.L. & Srivastava, D., 2015. Big data integration. *Synthesis Lectures on Data Management*, 7(1), pp.1–198.

Dong, X.L. & Srivastava, D., 2013. Big Data Integration. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 6(11), pp.1188–1189.

Getoor, L. & Machanavajjhala, A., 2012. Entity resolution: theory, practice & open challenges. *PVLDB*, 5(12), pp.2018–2019.

Goodfellow, I. et al., 2016. *Deep learning*, MIT press Cambridge.

Halevy, A., Norvig, P. & Pereira, F., 2009. The Unreasonable Effectiveness of Data. *IEEE intelligent systems*, 24(2), pp.8–12.

Konda, P. et al., 2016. Magellan: Toward Building Entity Matching Management Systems. *PVLDB*, 9(12), pp.1197–1208.

References Part I: Introduction

Kumar, A., Boehm, M. & Yang, J., 2017. Data Management in Machine Learning: Challenges, Techniques, and Systems. In *Proceedings of the 2017 ACM International Conference on Management of Data*. SIGMOD '17. New York, NY, USA: ACM, pp. 1717–1722.

Lockard, C. et al., 2018. CERES: Distantly Supervised Relation Extraction from the Semi-Structured Web. *arXiv [cs.AI]*. Available at: http://arxiv.org/abs/1804.04635.

Mohri, M., Rostamizadeh, A. & Talwalkar, A., 2012. *Foundations of Machine Learning*, MIT Press.

Polyzotis, N. et al., 2017. Data Management Challenges in Production Machine Learning. In *Proceedings of the 2017 ACM International Conference on Management of Data*. SIGMOD '17. New York, NY, USA: ACM, pp. 1723–1726.

Ratner, A. et al., 2017. Snorkel: Rapid Training Data Creation with Weak Supervision. *PVLDB*, 11(3), pp.269–282.

Rekatsinas, T. et al., 2017. HoloClean: Holistic Data Repairs with Probabilistic Inference. *PVLDB*, 10(11), pp.1190–1201.

Wu, S. et al., 2018. Fonduer: Knowledge Base Construction from Richly Formatted Data. In *Proceedings of the 2018 International Conference on Management of Data*. ACM, pp. 1301–1316.

Zheng, G. et al., 2018. OpenTag: Open Attribute Value Extraction from Product Profiles. In *KDD*. Available at: https://people.mpi-inf.mpg.de/~smukherjee/research/OpenTag-KDD18.pdf.

References Part II: Entity Linkage

Bhattacharya, I. & Getoor, L., 2006. A latent dirichlet model for unsupervised entity resolution. In *SDM*. SIAM, pp. 47–58.

Das, S. et al., 2017. Falcon: Scaling Up Hands-Off Crowdsourced Entity Matching to Build Cloud Services. In *Sigmod*. pp. 1431–1446.

Doan, A. et al., 2017. Human-in-the-Loop Challenges for Entity Matching: A Midterm Report. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics, HILDA@SIGMOD 2017, Chicago, IL, USA, May 14, 2017*. pp. 12:1–12:6.

Fellegi, I.P. & Sunter, A.B., 1969. A Theory for Record Linkage. *Journal of the Americal Statistical Association*, 64(328), pp.1183–1210.

Getoor, L. & Machanavajjhala, A., 2012. Entity resolution: theory, practice & open challenges. *PVLDB*, 5(12), pp.2018–2019.

Gokhale, C. et al., 2014. Corleone: Hands-off Crowdsourcing for Entity Matching. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. SIGMOD '14. New York, NY, USA: ACM, pp. 601–612.

Hassanzadeh, O. et al., 2009. Framework for Evaluating Clustering Algorithms in Duplicate Detection. *PVLDB*, 2(1), pp.1282–1293.

Ji, H., 2014. Entity Linking and Wikification Reading List. Available at: http://nlp.cs.rpi.edu/kbp/2014/elreading.html.

Konda, P. et al., 2016. Magellan: Toward Building Entity Matching Management Systems. *PVLDB*, 9(12), pp.1197–1208.

Kopcke, H., Thor, A. & Rahm, E., 2010. Evaluation of entity resolution approaches on real-world match problems. *PVLDB*, 3(1), pp.484–493.

References Part II: Entity Linkage

Mudgal, S. et al., 2018. Deep Learning for Entity Matching: A Design Space Exploration. In *Proceedings of the 2018 International Conference on Management of Data*. ACM, pp. 19–34.

Pujara, J. & Getoor, L., 2016. Generic Statistical Relational Entity Resolution in Knowledge Graphs. In *AAAI*.

Rakshit Trivedi, Bunyamin Sisman, Xin Luna Dong, Christos Faloutsos, Jun Ma and Hongyuan Zha., LinkNBed: Multi-Graph Representation Learning with Entity Linkage. In *56th Annual Meeting of the Association for Computational Linguistics*. ACL.

Sarawagi, S. & Bhamidipaty, A., 2002. Interactive deduplication using active learning. In *SIGKDD*.

Singla, P. & Domingos, P., 2006. Entity Resolution with Markov Logic. In *ICDM*. Washington, DC, USA: IEEE Computer Society, pp. 572–582.

Stonebraker, M. et al., 2013. Data Curation at Scale: The Data Tamer System. In *CIDR*.

Verroios, V., Garcia-Molina, H. & Papakonstantinou, Y., 2017. Waldo: An Adaptive Human Interface for Crowd Entity Resolution. In *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017*. pp. 1133–1148.

References Part II: Data Extraction

Das, R. et al., 2017. Chains of reasoning over entities, relations, and text using recurrent neural networks. In *EACL*.

Dong, X. et al., 2014. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '14. New York, NY, USA: ACM, pp. 601–610.

Dong, X.L., 2017. Challenges and Innovations in Building a Product Knowledge Graph. In *AKBC*.

Gulhane, P. et al., 2011. Web-scale information extraction with vertex. In *2011 IEEE 27th International Conference on Data Engineering*. pp. 1209–1220.

He, R. et al., 2017. An Unsupervised Neural Attention Model for Aspect Extraction. In *ACL*.

Hoffmann, R. et al., 2011. Knowledge-based Weak Supervision for Information Extraction of Overlapping Relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. HLT '11. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 541–550.

Limaye, G., Sarawagi, S. & Chakrabarti, S., 2010. Annotating and Searching Web Tables Using Entities, Types and Relationships. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 3(1-2), pp.1338–1347.

Lockard, C. et al., 2018. CERES: Distantly Supervised Relation Extraction from the Semi-Structured Web. *arXiv [cs.AI]*. Available at: http://arxiv.org/abs/1804.04635.

References Part II: Data Extraction

Mintz, M. et al., 2009. Distant supervision for relation extraction without labeled data. In *ACL*.

Mitchell, T. et al., 2018. Never-ending Learning. *Communications of the ACM*, 61(5), pp.103–115.

Neelakantan, A., Roth, B. & McCallum, A., 2015. Compositional vector space models for knowledge base completion. In *ACL*.

Riedel, S. et al., 2013. Relation Extraction with Matrix Factorization and Universal Schemas. In *HLT-NAACL*.

Shin, J. et al., 2015. Incremental Knowledge Base Construction Using DeepDive. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 8(11), pp.1310–1321.

Wu, S. et al., 2018. Fonduer: Knowledge Base Construction from Richly Formatted Data. In *Proceedings of the 2018 International Conference on Management of Data*. ACM, pp. 1301–1316.

Zhang, C. et al., 2017. DeepDive: Declarative Knowledge Base Construction. *CACM*, 60(5), pp.93–102.

References Part II: Data Fusion

Dawid, A.P. & Skene, A.M., 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 28(1), pp.20–28.

Dong, X.L. et al., 2014. From Data Fusion to Knowledge Fusion. *PVLDB*.

Dong, X.L. et al., 2015. Knowledge-based Trust: Estimating the Trustworthiness of Web Sources. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 8(9), pp.938–949.

Dong, X.L. & Naumann, F., 2009. Data Fusion: Resolving Data Conflicts for Integration. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 2(2), pp.1654–1655.

Gao, J. et al., 2016. Mining Reliable Information from Passively and Actively Crowdsourced Data. In *KDD*. pp. 2121–2122.

Jaffe, A., Nadler, B. & Kluger, Y., 2015. Estimating the accuracies of multiple classifiers without labeled data. In *Artificial Intelligence and Statistics*. Artificial Intelligence and Statistics. pp. 407–415.

Li, H., Yu, B. & Zhou, D., 2013. Error rate analysis of labeling by crowdsourcing. In *ICML Workshop: Machine Learning Meets Crowdsourcing. Atlanta, Georgia, USA*.

Li, Q. et al., 2014. A Confidence-aware Approach for Truth Discovery on Long-tail Data. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 8(4), pp.425–436.

References Part II: Data Fusion

Li, X. et al., 2013. Truth Finding on the Deep Web: Is the Problem Solved? *PVLDB*, 6(2).

Li, Y. et al., 2016. A Survey on Truth Discovery. *SIGKDD Explor. Newsl.*, 17(2), pp.1–16.

Nickel, M. et al., 2016. A Review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IEEE*, 104(1), pp.11–33.

Pasternack, J. & Roth, D., 2010. Knowing what to believe (when you already know something). In *COLING*. pp. 877–885.

Platanios, E. A., Dubey, A., & Mitchell, T. (2016, June). Estimating accuracy from unlabeled data: A bayesian approach. In *International Conference on Machine Learning*(pp. 1416-1425).

Rekatsinas, T. et al., 2017. SLiMFast: Guaranteed Results for Data Fusion and Source Reliability. In *Proceedings of the 2017 ACM International Conference on Management of Data*. SIGMOD '17. New York, NY, USA: ACM, pp. 1399–1414.

Shaham, U. et al., 2016. A Deep Learning Approach to Unsupervised Ensemble Learning. In *International Conference on Machine Learning*. International Conference on Machine Learning. pp. 30–39.

Wang, Q. et al., 2017. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE transactions on knowledge and data engineering*, 29(12), pp.2724–2743.

Yin, X., Han, J. & Yu, P.S., 2007. Truth discovery with multiple conflicting information providers on the web. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 1048–1052.

References Part II: Data Fusion

Zhang, Y. et al., 2014. Spectral Methods meet EM: A Provably Optimal Algorithm for Crowdsourcing. In Z. Ghahramani et al., eds. *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., pp. 1260–1268.

Zhao, B. et al., 2012. A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 5(6), pp.550–561.

References Part III: Training Data Creation

Chapelle, O., Scholkopf, B. & Eds., A.Z., 2009. Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006) [Book reviews]. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 20(3), pp.542–542.

Dawid, A.P. & Skene, A.M., 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 28(1), pp.20–28.

Mintz, M. et al., 2009. Distant supervision for relation extraction without labeled data. In *ACL*.

Mitchell, T., 2017. Learning from Limited Labeled Data (But a Lot of Unlabeled Data). Available at: https://lld-workshop.github.io/slides/tom_mitchell_lld.pdf.

Platanios, E.A., Dubey, A. & Mitchell, T., 2016. Estimating Accuracy from Unlabeled Data: A Bayesian Approach. In *International Conference on Machine Learning*. International Conference on Machine Learning. pp. 1416–1425.

Ratner, A. et al., 2017. Snorkel: Rapid Training Data Creation with Weak Supervision. *PVLDB*, 11(3), pp.269–282.

Ratner, A.J. et al., 2016. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems*. pp. 3567–3575.

Raykar, V.C. et al., 2010. Learning From Crowds. *Journal of machine learning research: JMLR*, 11, pp.1297–1322.

Recht, B. et al., 2018. Do CIFAR-10 Classifiers Generalize to CIFAR-10? *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/1806.00451.

References Part III: Training Data Creation

Roth, B. & Klakow, D., 2013. Combining generative and discriminative model scores for distant supervision. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. pp. 24–29.

Russell, S. & Stefano, E., 2017. Label-free supervision of neural networks with physics and domain knowledge. *Proceedings of AAAI*.

Salimans, T. et al., 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*. pp. 2234–2242.

Schapire, R.E. & Freund, Y., 2012. Boosting: Foundations and Algorithms. Adaptive computation and machine learning.

References Part III: Data Cleaning

Bailis, P. et al., 2017. MacroBase: Prioritizing Attention in Fast Data. In *Proceedings of the 2017 ACM International Conference on Management of Data*. SIGMOD '17. New York, NY, USA: ACM, pp. 541–556.

Chu, X. et al., 2016. Data Cleaning: Overview and Emerging Challenges. In *Proceedings of the 2016 International Conference on Management of Data*. SIGMOD '16. New York, NY, USA: ACM, pp. 2201–2206.

Chandola, V., Banerjee, A. & Kumar, V., 2009. Anomaly Detection: A Survey. *ACM Comput. Surv.*, 41(3), pp.15:1–15:58.

Galhardas, H. et al., 2001. Declarative data cleaning: Language, model, and algorithms. In *VLDB*. pp. 371–380.

Hellerstein, J.M., 2008. Quantitative data cleaning for large databases. *Statistical journal of the United Nations Economic Commission for Europe*. Available at: http://db.cs.berkeley.edu/jmh/papers/cleaning-unece.pdf.

Ilyas, I.F., 2016. Effective Data Cleaning with Continuous Evaluation. *IEEE Data Eng. Bull.*, 39, pp.38–46.

Krishnan, S. et al., 2016. ActiveClean: Interactive Data Cleaning for Statistical Modeling. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 9(12), pp.948–959.

Krishnan, S. et al., 2017. BoostClean: Automated Error Detection and Repair for Machine Learning. *arXiv [cs.DB]*. Available at: http://arxiv.org/abs/1711.01299.

References Part III: Data Cleaning

Mayfield, C., Neville, J. & Prabhakar, S., 2010. ERACER: A Database Approach for Statistical Inference and Data Cleaning. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*. SIGMOD '10. New York, NY, USA: ACM, pp. 75–86.

Rekatsinas, T. et al., 2017. HoloClean: Holistic Data Repairs with Probabilistic Inference. *PVLDB*, 10(11), pp.1190–1201.

Wang, X., Dong, X.L. & Meliou, A., 2015. Data X-Ray: A Diagnostic Tool for Data Errors. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD '15. New York, NY, USA: ACM, pp. 1231–1245.

Yakout, M., Berti-Équille, L. & Elmagarmid, A.K., 2013. Don'T Be SCAREd: Use SCalable Automatic REpairing with Maximal Likelihood and Bounded Changes. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. SIGMOD '13. New York, NY, USA: ACM, pp. 553–564.