Scientific Text Mining and Knowledge Graphs

Chapter 1
Part 4: Scientific Statements

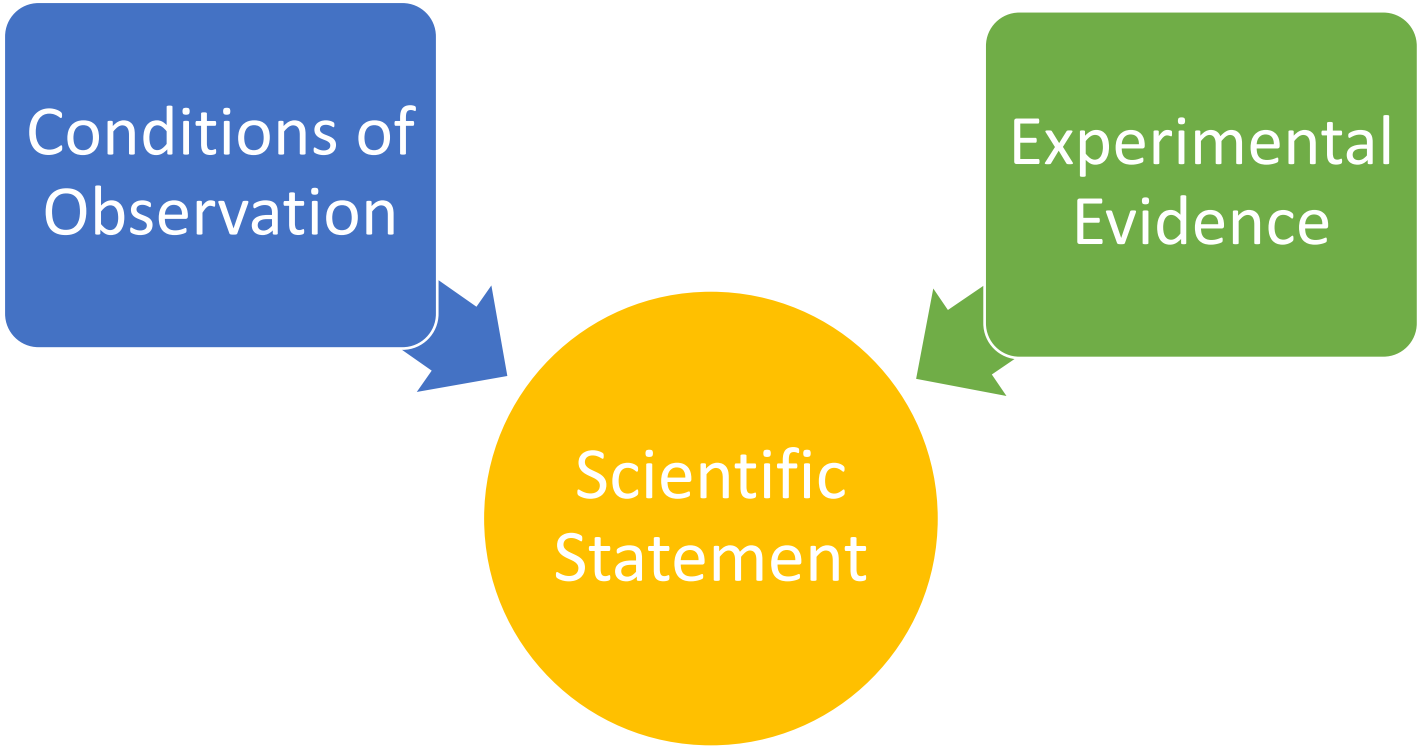
Presenter: Meng Jiang

University of Notre Dame

mjiang2@nd.edu

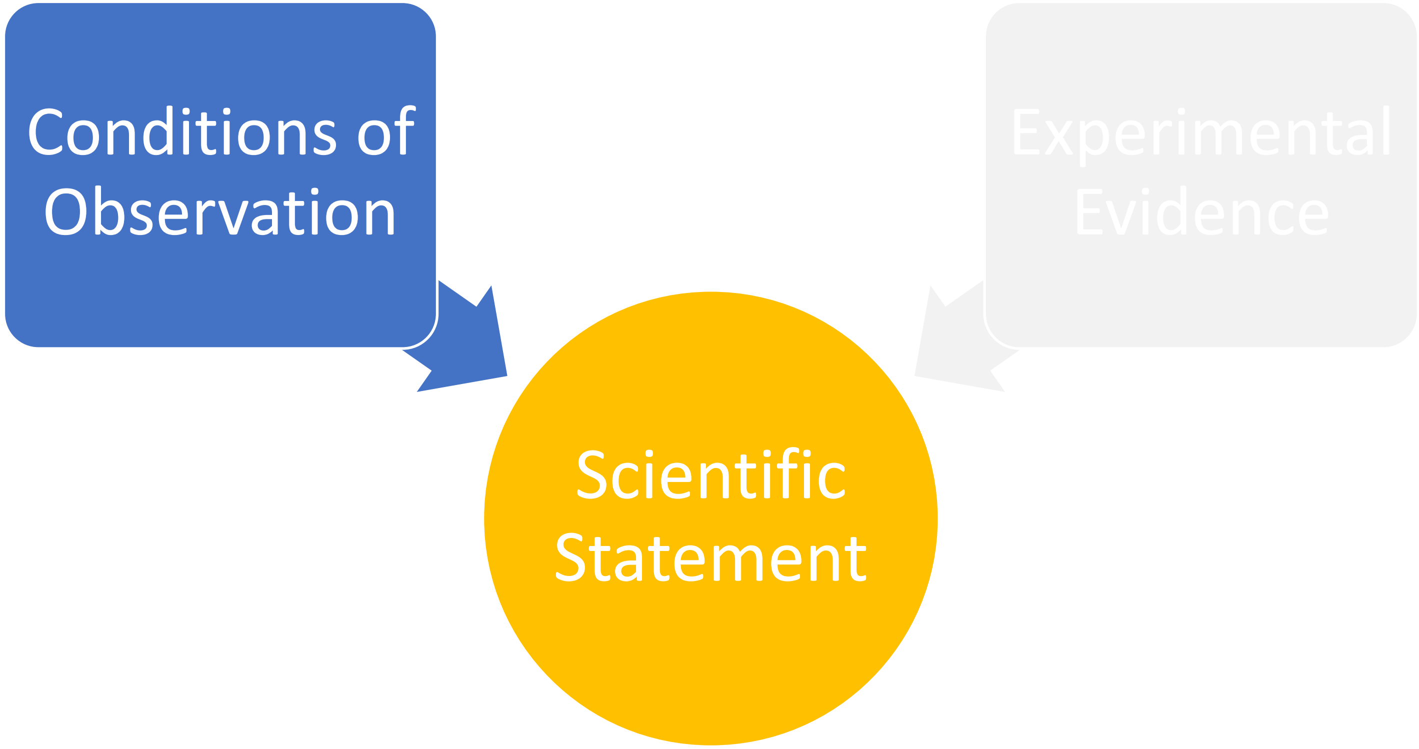
Two Works

- Extracting **conditional statements** for biomedical literature (KDD'19, EMNLP'19, TCBB)
- Extracting **experimental evidence** for data science (WWW'20)



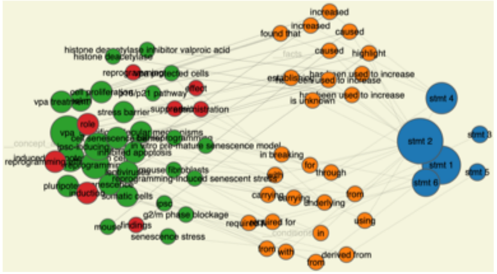
Two Works

- Extracting **conditional statements** for biomedical literature (KDD’19, EMNLP’19, TCBB)
- Extracting **experimental evidence** for data science (WWW’20)



Structuring Text into Knowledge Graph

ML



Scientific literature

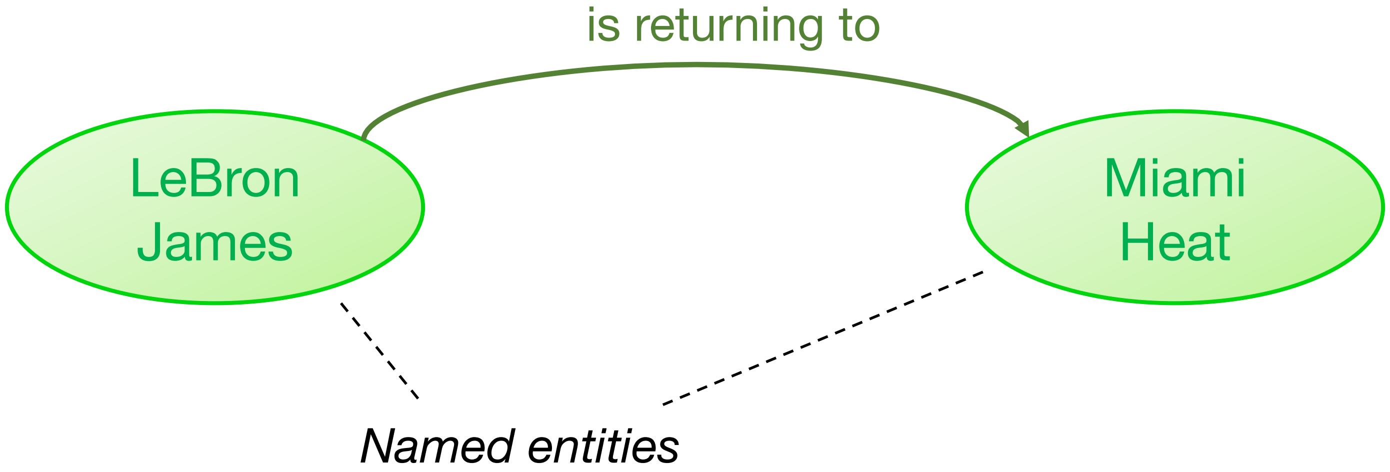
Unstructured

Scientific knowledge graph

Structured

Given "*LeBron James is returning to Miami Heat…*"

Find fact tuple: (LeBron James, is returning to, Miami Heat)



Science IE: Conditional Statements

"*We showed that extracellular acidic pH reduces the activity of TRPV5/V6 channels, whereas alkaline pH increases the activity of TRPV5/V6 channels in Jurkat T cells.*"

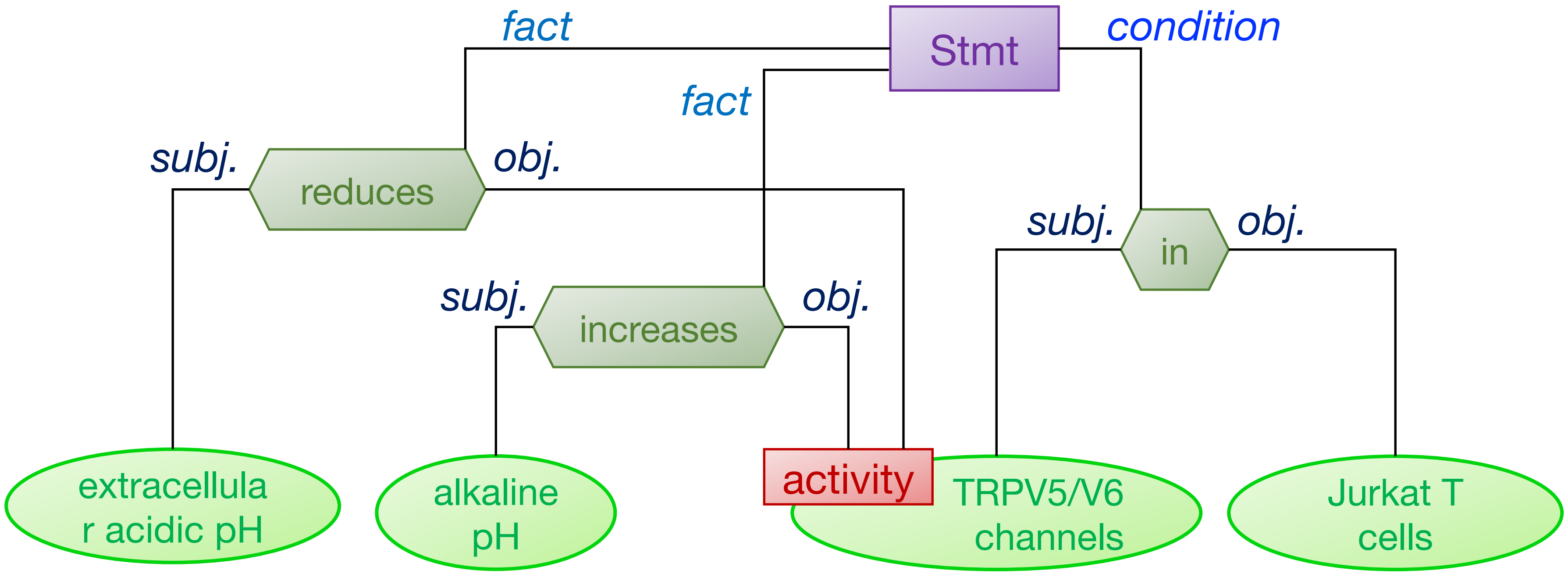
Fact tuple 1: (extracellular acidic pH, reduces, {TRPV5/V6 channels: activity})
Fact tuple 2: (alkaline pH, increases, {TRPV5/V6 channels: activity})
Condition tuple: (TRPV5/V6 channels, in, Jurkat T cells)

"*During T lymphocyte activation as well as production of cytokines, …*"

Condition tuple 1: (-, during, {T lymphocyte: activation})
Condition tuple 2: (-, during, {cytokines: production})

Three-Level Scientific KGs

Fact tuple 1: (extracellular acidic pH, reduces, {TRPV5/V6 channels: activity})
Fact tuple 2: (alkaline pH, increases, {TRPV5/V6 channels: activity})
Condition tuple: (TRPV5/V6 channels, in, Jurkat T cells)

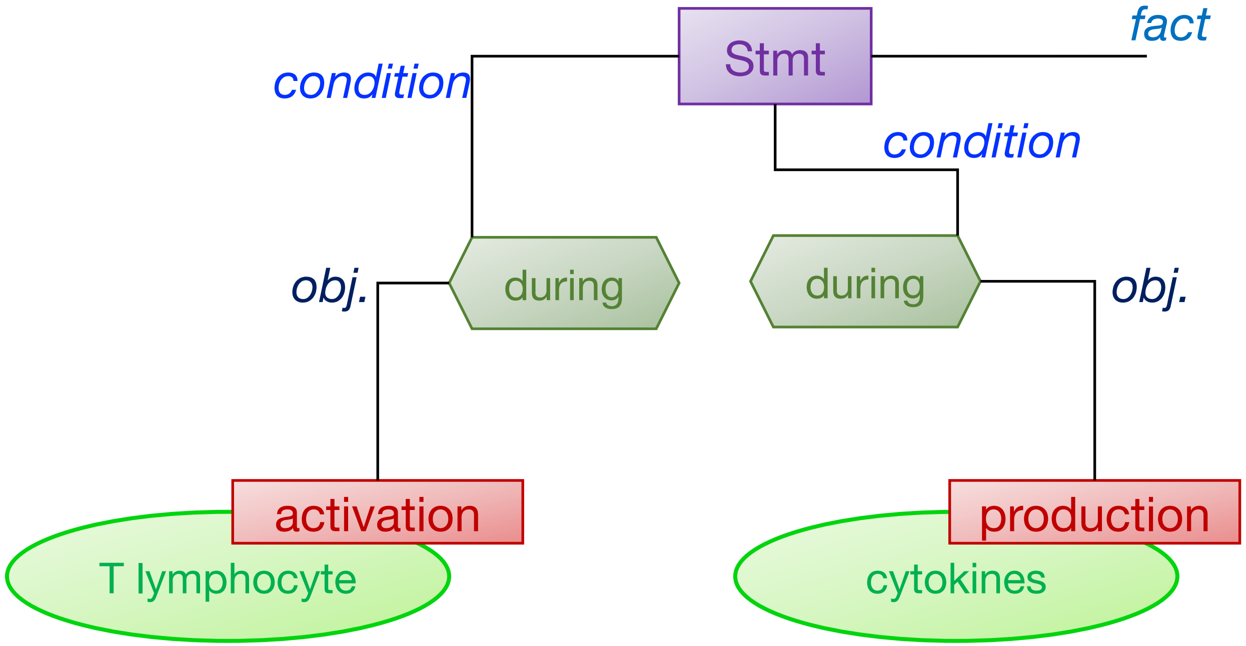


Three-Level Scientific KGs (cont'd)

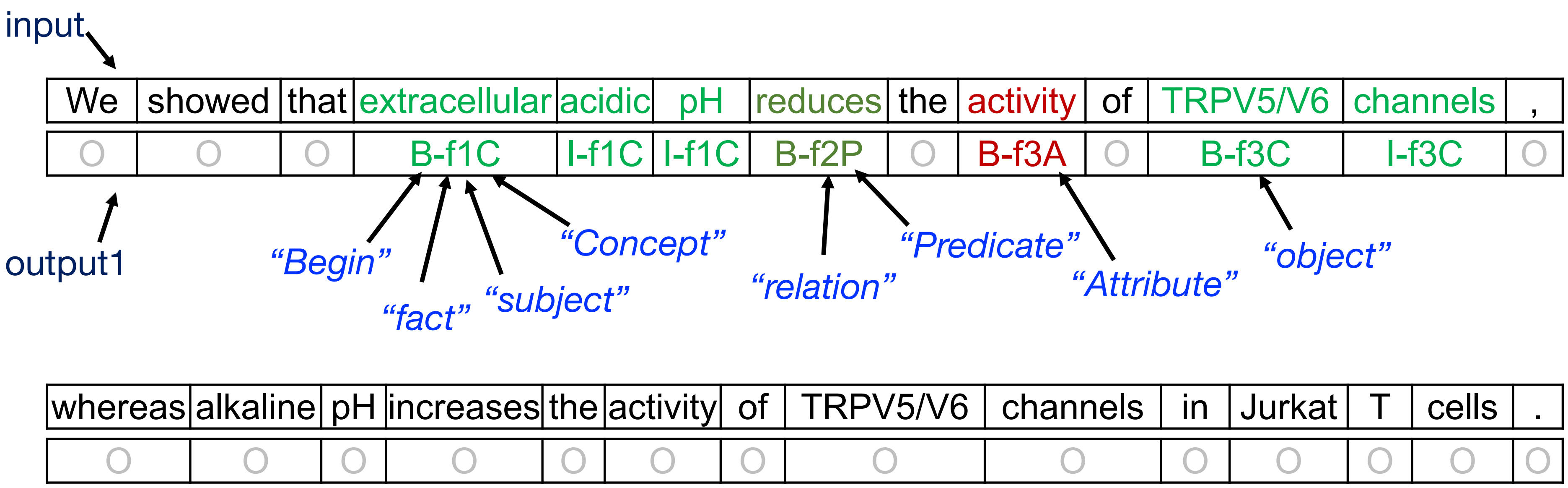
"*During T lymphocyte activation as well as production of cytokines, …*"

Condition tuple 1: (-, during, {T lymphocyte: activation})

Condition tuple 2: (-, during, {cytokines: production})



Sequence Labeling for IE



Fact tuple 1: (extracellular acidic pH, reduces, {TRPV5/V6 channels: activity})

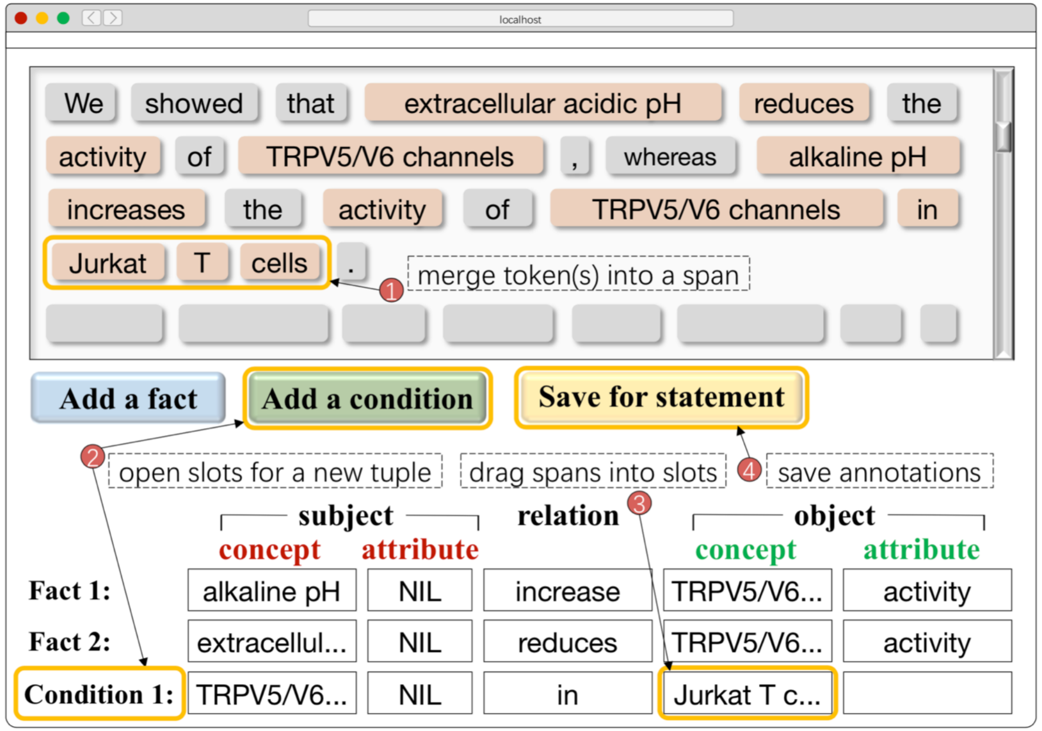
Multi-Output Sequence Labeling

We	showed	that	extracellular	acidic	pH	reduces	the	activity	of	TRPV5/V6	channels	,
O	O	O	B-f1C	I-f1C	I-f1C	B-f2P	O	B-f3A	O	B-f3C	I-f3C	O
O	O	O	O	O	O	O	O	O	O	O	O	O
O	O	O	O	O	O	O	O	O	O	O	O	O

whereas	alkaline	pH	increases	the	activity	of	TRPV5/V6	channels	in	Jurkat	T	cells	.
O	O	O	O	O	O	O	O	O	O	O	O	O	O
O	B-f1C	I-f1C	B-f2P	O	B-f3A	O	B-f3C	I-f3C	O	O	O	O	O
O	O	O	O	O	O	O	B-c1C	I-c1C	B-c2P	B-c3C	I-c3C	I-c3C	O

Fact tuple 1: (extracellular acidic pH, reduces, {TRPV5/V6 channels: activity})
Fact tuple 2: (alkaline pH, increases, {TRPV5/V6 channels: activity})
Condition tuple: (TRPV5/V6 channels, in, Jurkat T cells)

Sequence Labels



B-f1C	I-f1C	B-c1C	I-c1C
B-f1A	I-f1A	B-c1A	I-c1A
B-f2P	I-f2P	B-c2P	I-c2P
B-f3C	I-f3C	B-c3C	I-c3C
B-f3A	I-f3A	B-c3A	I-c3A
O			

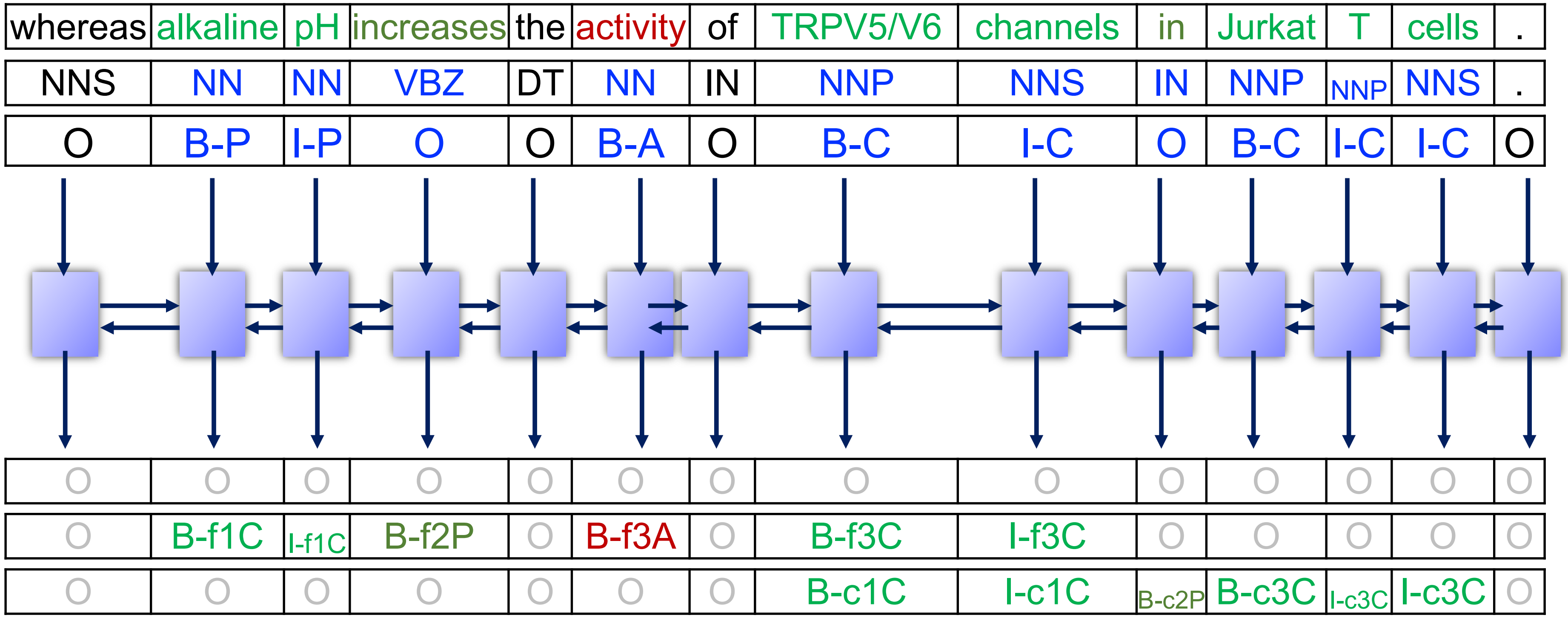
- 3 expert annotators
- 31 PubMed paper abstracts (docs)
- > 30 minutes per anno. per doc
- 336 statement sentences
- 756 fact tuples
- 654 condition tuples

More Signals from Massive Data

- Unlabeled data
  - 15,544,338 documents
  - 140,949,399 statement sentences
- Feature extraction
  - Tokenization
  - Part-of-speech tagging
  - Phrase mining
  - Concept detection
  - Attribute discovery
  - …

Structured Output	Publications ('17-)
P_{hrase}	TKDE'18
C_{oncept}	ACL'20sub
$H(c_{\text{oncept}} \times r_{\text{elation}})$	KDD'18c, TextGraphs'19
$D(e_{\text{ntity}} \times a_{\text{ttribute}} \times v_{\text{alue}})$	KDD'17, KDD'18b, EYRE'19
$D(e_{\text{ntity}} \times a_{\text{ttribute}} \times v_{\text{alue}} \times t_{\text{start}} \times t_{\text{end}})$	WWW'19a, FEVER'20

Multi-Input Multi-Output Sequence Labeling

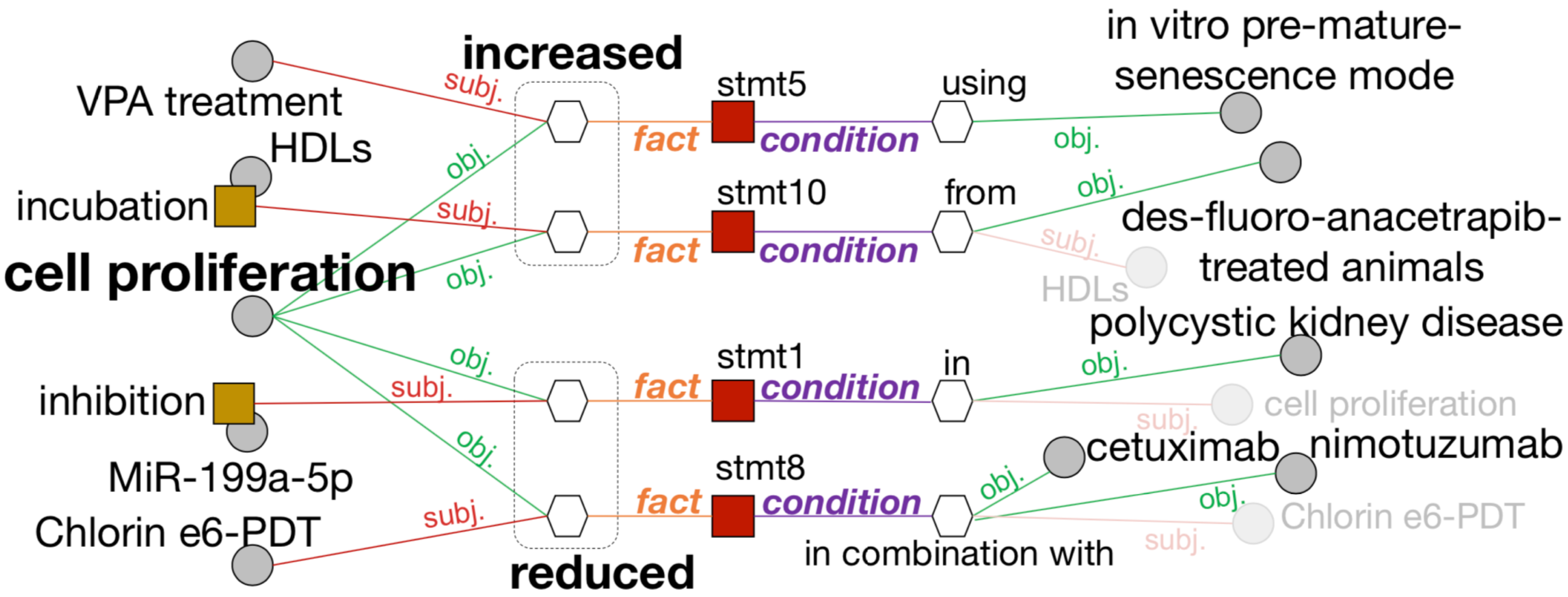


MIMO (BiLSTM/BERT)

Evaluation

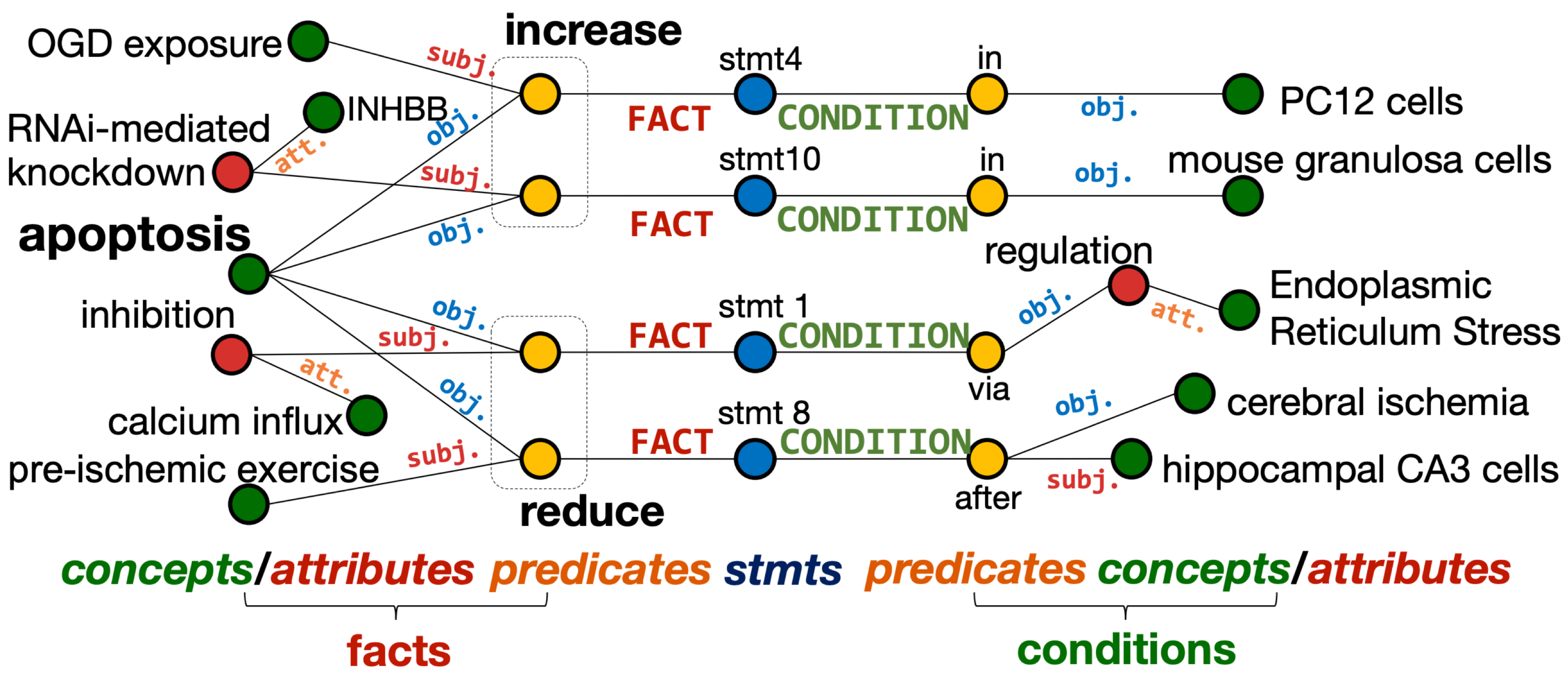
	Token Label Prediction (%)			Tuple Extraction (%)		
	P	R	F1 / fact, cond.	P	R	F1 / fact, cond.
Allennlp OpenIE (Stanovsky et al. 2018)	-	-	-	42.60	38.22	40.29 / -, -
Stanford OpenIE (Angeli et al. 2015)	-	-	-	47.11	41.62	44.19 / -, -
Structured SVM (Tsochantaridis et al. 2015)	32.68	25.80	28.83 / 32.76, 24.71	47.62	46.15	46.87 / 45.01, 48.72
CRF (Lafferty et al. 2001)	60.07	41.92	49.37 / 56.23, 41.87	65.19	62.44	63.78 / 64/07, 63.44
BiLSTM-LSTMd (Zheng et al. 2017)	61.00	56.26	58.53 / 65.16, 51.78	71.57	66.55	68.97 / 69.51, 68.41
MO (BiLSTM based)	-	-	-	71.80	72.34	72.07 / 72.39, 71.73
MIMO (BiLSTM based)	67.80	58.24	62.66 / 66.67, 58.58	75.35	74.67	75.01 / 74.91, 75.10
BERT-BiLSTM	70.07	70.19	70.13 / 74.30, 65.88	78.64	73.67	76.08 / 76.14, 75.99
MO (BERT based)	-	-	-	77.38	79.19	78.27 / 76.74, 79.89
MIMO (BERT based)	**75.91**	**71.08**	**73.41 / 76.01, 70.75**	**81.06**	**80.53**	**80.79 / 79.94, 81.64**

Case Study



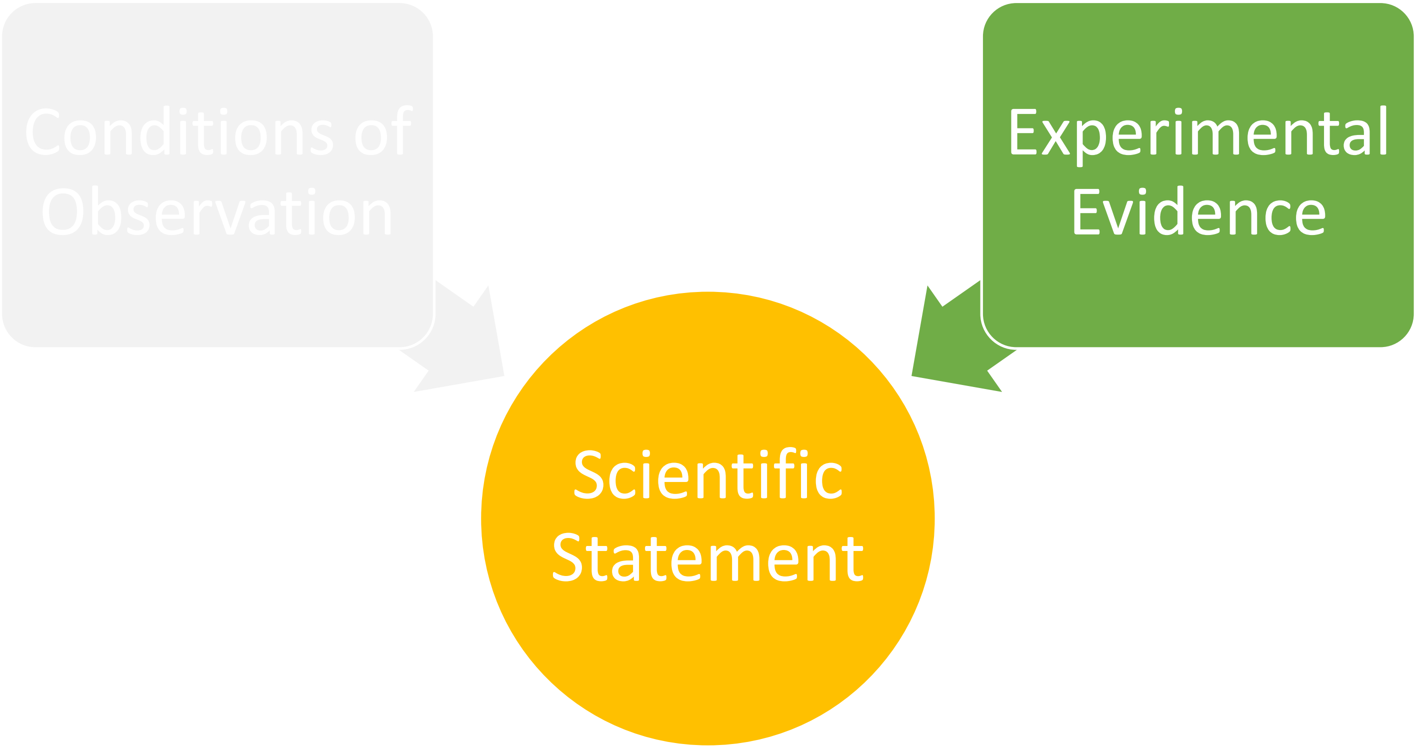
increased
reduced
VPA treatment
HDLs
incubation
cell proliferation
inhibition
MiR-199a-5p
Chlorin e6-PDT
subj.
obj.
fact
stmt5
stmt10
stmt1
stmt8
condition
using
from
in
in combination with
in vitro pre-mature-senescence mode
des-fluoro-anacetrapib-treated animals
HDLs
polycystic kidney disease
cell proliferation
cetuximab
nimotuzumab
Chlorin e6-PDT

Case Study (cont’d)



Two Works

- Extracting **conditional statements** for biomedical literature (KDD’19, EMNLP’19, TCBB)
- Extracting **experimental evidence** for data science (WWW’20)



Motivation

PPDSparse: A Parallel Primal-Dual Sparse Method for Extreme Classification
by CMU, UT Austin, Pentuum [**KDD 2017**]

Data	Metrics	FastXML	PfastreXML	SLEEC	PDSparse	DiSMEC	PPDSparse
Amazon-670K	T_{train}	**5624s**	6559s	20904s		174135s	**921.9s**
N_{train}=490449	P@1 (%)	33.12	32.87	35.62		43.00	**43.04**
N_{test}=153025	P@3 (%)	28.98	29.52	31.65	MLE	38.23	**38.24**
D=135909	P@5 (%)	26.11	26.82	28.85		34.93	**34.94**
K=670091	model size	**4.0G**	6.3G	6.6G		8.1G	5.3G
	T_{test}/N_{test}	**1.41ms**	1.98ms	6.94ms		148ms	20ms
WikiLSHTC-325K	T_{train}	**19160s**	20070s	39000s	94343s	271407s	**353s**
N_{train}=1778351	P@1 (%)	50.01	57.17	58.34	60.70	64.00	**64.13**
N_{test}=587084	P@3 (%)	32.83	37.03	36.7	39.62	**42.31**	42.10
D=1617899	P@5 (%)	24.13	27.19	26.45	29.20	**31.40**	31.14
K=325056	model size	14G	16G	650M	**547M**	8.1G	4.9G
	T_{test}/N_{test}	**1.02ms**	1.47ms	4.85ms	3.89ms	65ms	290ms
Delicious-200K	T_{train}	8832.46s	8807.51s	**4838.7s**	5137.4s	38814s	**2869s**
N_{train}=196606	P@1 (%)	**48.85**	26.66	47.78	37.69	44.71	45.05
N_{test}=100095	P@3 (%)	**42.84**	23.56	42.05	30.16	38.08	38.34
D=782585	P@5 (%)	**39.83**	23.21	39.29	27.01	34.7	34.90
K=205443	model size	1.3G	20G	2.1G	**3.8M**	18G	9.4G
	T_{test}/N_{test}	1.28ms	7.40ms	2.685ms	**0.432ms**	311.4ms	275ms
AmazonCat-13K	T_{train}	11535s	13985s	119840s	**2789s**	11828s	**122.8s**
N_{train}=1186239	P@1 (%)	**94.02**	86.06	90.56	87.43	92.72	92.72
N_{test}=306782	P@3 (%)	**79.93**	76.24	76.96	70.48	78.11	78.14
D=203882	P@5 (%)	**64.90**	63.65	62.63	56.70	63.40	63.41
K=13330	model size	9.7G	11G	12G	**15M**	2.1G	355M
	T_{test}/N_{test}	1.21ms	1.34ms	13.36ms	0.87ms	**0.20ms**	1.82ms

Motivation (cont'd)

AnnexML: Approximate Nearest Neighbor Search for Extreme Multi-label Classification by Yahoo Japan Corporation [**KDD 2017**]

Dataset		AnnexML	SLEEC	FastXML	PfastreXML	PLT	PD-Sparse	Most common
AmazonCat-13K	P@1	**0.9355**	0.8919	0.9310	0.8994	0.9147	0.8931	0.2988
	P@3	**0.7838**	0.7517	0.7818	0.7724	0.7584	0.7403	0.1878
	P@5	0.6332	0.6109	0.6338	**0.6353**	0.6102	0.6011	0.1486
Wiki10-31K	P@1	**0.8650**	0.8554	0.8295	0.8263	0.8434	0.7771	0.8079
	P@3	**0.7428**	0.7359	0.6756	0.6874	0.7234	0.6573	0.5050
	P@5	**0.6419**	0.6310	0.5770	0.6006	0.6272	0.5539	0.3675
Delicious-200K	P@1	0.4666	**0.4703**	0.4320	0.3762	0.4537	0.3437	0.3873
	P@3	0.4079	**0.4167**	0.3868	0.3562	0.3894	0.2948	0.3675
	P@5	0.3764	**0.3888**	0.3621	0.3403	0.3588	0.2704	0.3552
WikiLSHTC-325K	P@1	**0.6336**	0.5557	0.4975	0.5810	0.4567	0.6126	0.1588
	P@3	**0.4066**	0.3306	0.3310	0.3761	0.2913	0.3948	0.0603
	P@5	**0.2979**	0.2407	0.2445	0.2769	0.2195	0.2879	0.0380
Wikipedia-500K	P@1	**0.6386**	0.5839	0.4934	0.5891	–	–	0.1529
	P@3	**0.4269**	0.3788	0.3351	0.3937	–	–	0.0583
	P@5	**0.3237**	0.2821	0.2586	0.3005	–	–	0.0368
Amazon-670K	P@1	**0.4208**	0.3505	0.3697	0.3919	0.3665	0.3370	0.0028
	P@3	**0.3665**	0.3125	0.3332	0.3584	0.3212	0.2962	0.0027
	P@5	0.3276	0.2856	0.3053	**0.3321**	0.2885	0.2684	0.0023

Motivation (cont'd)

Dataset	(%)	SLEEC	FastXML	PfastreXML	PDSparse
AmazonCat	P@1	90.56/89.19	94.02/93.10	86.06/89.94	87.43/89.31
-13K	P@3	76.96/75.17	79.93/78.18	86.06/77.24	87.43/74.03
	P@5	62.63/61.09	64.90/63.38	63.65/63.53	56.70/60.11
Delicious	P@1	47.78/47.03	48.85/43.20	26.66/37.62	37.69/34.37
-200K	P@3	42.05/41.67	42.84/38.68	23.56/35.62	30.16/29.48
	P@5	39.29/38.88	39.83/36.21	23.21/34.03	27.01/27.04
WikiLSHTC	P@1	58.34/55.57	50.01/49.75	57.17/58.10	60.70/61.26
-325K	P@3	36.70/33.06	32.83/33.10	37.03/37.61	39.62/39.48
	P@5	26.45/24.07	24.13/24.45	27.19/27.69	29.20/28.79

Motivation

On *dataset*, *method* makes a *metric* of *XXX (score)* on the task.

Definition

Experimental evidence

Component

Literature survey
Method design
…

Usage

What are the datasets we can use?
How do people evaluate the methods?
What is the "state-of-the-art"?
…

Research study

Extreme (multi-label) classification

Recommender systems

……

Goal

Research problems

Experimental Evidence Extraction System in Data Science with **Hybrid Table Features and Ensemble Learning**

Develop a computational method to build the system

- *Feature extraction*
- *Learning strategies*

System Pipeline

PDFs in
Digital Libraries

Tables in PDF

Experimental Result
Database (**ERD**)

(ACM TIST 2011)

Table III. Performance Comparisons (A Smaller MAE or RMSE Value Means a Better Performance)

Training Data	Metrics	Dimensionality = 5							
		UserMean	ItemMean	NMF	PMF	TCF	Trust	SoRec	RSTE
90%	MAE	0.9134	0.9768	0.8738	0.8676	0.9005	0.9054	0.8442	**0.8377**
	RMSE	1.1688	1.2375	1.1649	1.1575	1.1697	1.1959	1.1333	**1.1109**
80%	MAE	0.9285	0.9913	0.8975	0.8951	0.9044	0.9221	0.8638	**0.8594**
	RMSE	1.1817	1.2584	1.1861	1.1826	1.1761	1.2140	1.1530	**1.1346**

Training Data	Metrics	Dimensionality = 10							
		UserMean	ItemMean	NMF	PMF	TCF	Trust	SoRec	RSTE
90%	MAE	0.9134	0.9768	0.8712	0.8651	0.9005	0.9039	0.8404	**0.8367**
	RMSE	1.1688	1.2375	1.1621	1.1544	1.1697	1.1917	1.1293	**1.1094**
80%	MAE	0.9285	0.9913	0.8951	0.8886	0.9044	0.9215	0.8580	**0.8537**
	RMSE	1.1817	1.2584	1.1832	1.1760	1.1761	1.2132	1.1492	**1.1256**

(WSDM 2011)

Table 5: Performance Comparisons (Dimensionality = 10)

Dataset	Training	Metrics	UserMean	ItemMean	NMF	PMF	RSTE	$SR1_{vss}$	$SR1_{pcc}$	$SR2_{vss}$	$SR2_{pcc}$
Douban	80%	MAE	0.6809	0.6288	0.5732	0.5693	0.5643	0.5579	0.5576	0.5548	**0.5543**
		Improve	18.59%	11.85%	3.30%	2.63%	1.77%				
		RMSE	0.8480	0.7898	0.7225	0.7200	0.7144	0.7026	0.7022	0.6992	**0.6988**
		Improve	17.59%	11.52%	3.28%	2.94%	2.18%				
	60%	MAE	0.6823	0.6300	0.5768	0.5737	0.5698	0.5627	0.5623	0.5597	**0.5593**
		Improve	18.02%	11.22%	3.03%	2.51%	1.84%				
		RMSE	0.8505	0.7926	0.7351	0.7290	0.7207	0.7081	0.7078	0.7046	**0.7042**
		Improve	17.20%	11.15%	4.20%	3.40%	2.29%				
	40%	MAE	0.6854	0.6317	0.5899	0.5868	0.5767	0.5706	0.5702	0.5690	**0.5685**
		Improve	17.06%	10.00%	3.63%	3.12%	1.42%				
		RMSE	0.8567	0.7971	0.7482	0.7411	0.7295	0.7172	0.7169	0.7129	**0.7125**
		Improve	16.83%	10.61%	4.77%	3.86%	2.33%				
Epinions	90%	MAE	0.9134	0.9768	0.8712	0.8651	0.8367	0.8290	0.8287	0.8258	**0.8256**
		Improve	9.61%	15.48%	5.23%	4.57%	1.33%				
		RMSE	1.1688	1.2375	1.1621	1.1544	1.1094	1.0792	1.0790	1.0744	**1.0739**
		Improve	8.12%	13.22%	7.59%	6.97%	3.20%				
	80%	MAE	0.9285	0.9913	0.8951	0.8886	0.8537	0.8493	0.8491	0.8447	**0.8443**
		Improve	9.07%	14.83%	5.68%	4.99%	1.10%				
		RMSE	1.1817	1.2584	1.1832	1.1760	1.1256	1.1016	1.1013	1.0958	**1.0954**
		Improve	7.30%	12.95%	7.42%	6.85%	2.68%				

MAE on Epinions (80% Training) **Best baseline vs the proposed**
RMSE on Epinions (80% Training) **Conflicting between papers**

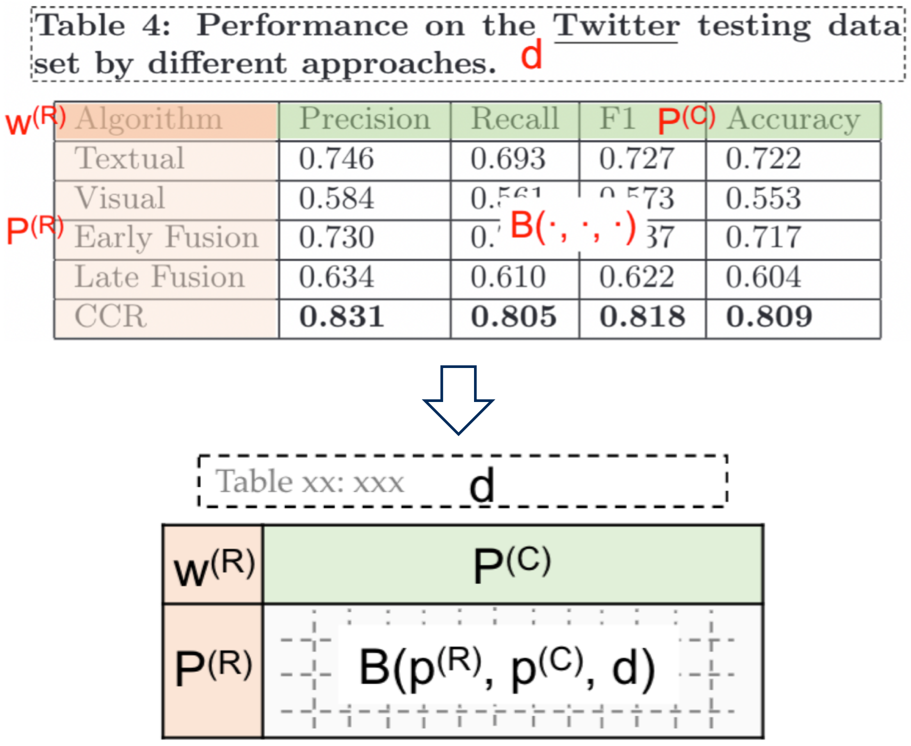
	A	B	C	D	E
1	**Method**	**Dataset**	**Metric**	**Score**	**Source**
10	UserMean	Epinions	MAE	0.9319	TOIS11-paper7-table3
11	UserMean	Epinions	MAE	0.9285	TIST11-paper3-table3
12	UserMean	Epinions	MAE	0.9285	WSDM11-paper12-table5
109	ItemMean	Epinions	RMSE	1.1973	TOIS11-paper7-table4
110	ItemMean	Epinions	RMSE	1.2584	TIST11-paper3-table3
111	ItemMean	Epinions	RMSE	1.2584	WSDM11-paper12-table5
112	Trust	Epinions	RMSE	1.2132	TIST11-paper3-table3
113	NMF	Epinions	RMSE	1.1832	TOIS11-paper7-table4
114	NMF	Epinions	RMSE	1.1832	TIST11-paper3-table3
115	NMF	Epinions	RMSE	1.1832	WSDM11-paper12-table5
116	SVD	Epinions	RMSE	1.1812	TOIS11-paper7-table4
117	TCF	Epinions	RMSE	1.1761	TIST11-paper3-table3
118	PMF	Epinions	RMSE	1.1760	TOIS11-paper7-table4
119	PMF	Epinions	RMSE	1.1760	TIST11-paper3-table3
120	PMF	Epinions	RMSE	1.1760	WSDM11-paper12-table5
121	SoRec	Epinions	RMSE	1.1492	TOIS11-paper7-table4
122	RSTE	Epinions	RMSE	1.1256	TIST11-paper3-table3
123	RSTE	Epinions	RMSE	1.1256	WSDM11-paper12-table5
124	SR1VSS	Epinions	RMSE	1.1016	WSDM11-paper12-table5
125	SR1PCC	Epinions	RMSE	1.1013	WSDM11-paper12-table5
126	SR2VSS	Epinions	RMSE	1.0958	WSDM11-paper12-table5
127	SR2PCC	Epinions	RMSE	1.0954	WSDM11-paper12-table5
169	SoRec	MovieLens	RMSE	...	...

SQL

This is the most challenging task!

Table Components

- Caption: d
- Row names: $P^{(R)}$
- Column names: $P^{(C)}$
- Name indicator: $W^{(R)}$
- Table body: $B(P^{(R)}, P^{(C)}, d)$



(a) 1×1, 1 row indicator, caption

Table Templates

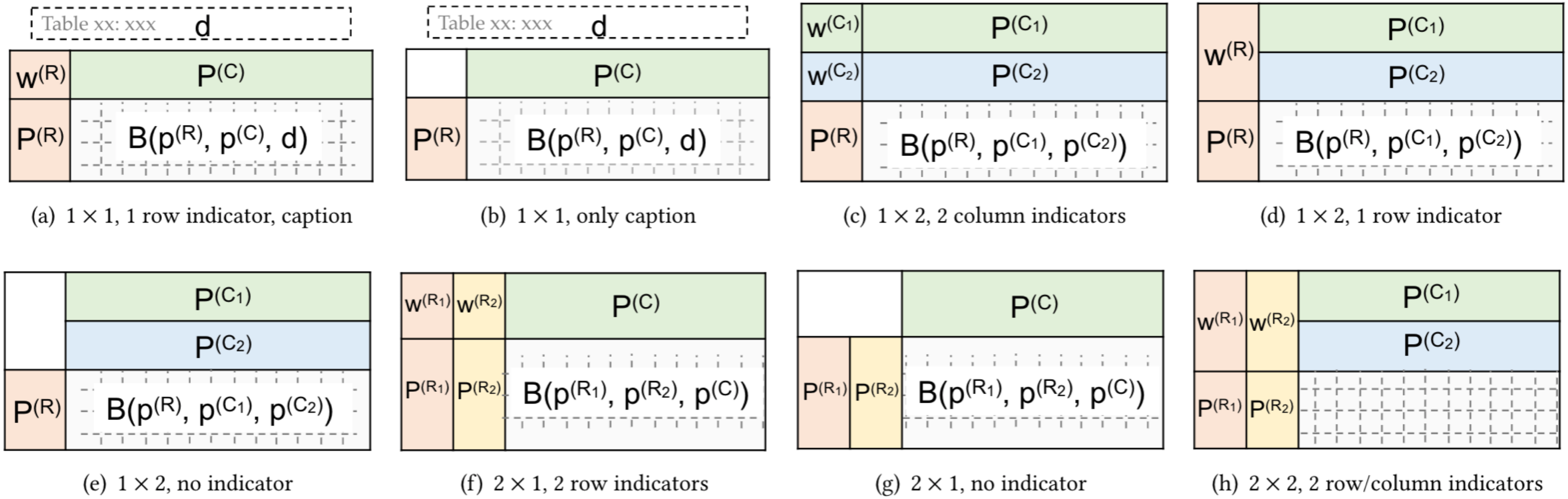


Figure 3: Eight major table templates: We will use the first seven templates which cover more than 95% of the tables in our dataset. The cells in the table's body are triplets based on rows/columns/caption. (Best viewed in color)

Table Templates (cont'd)

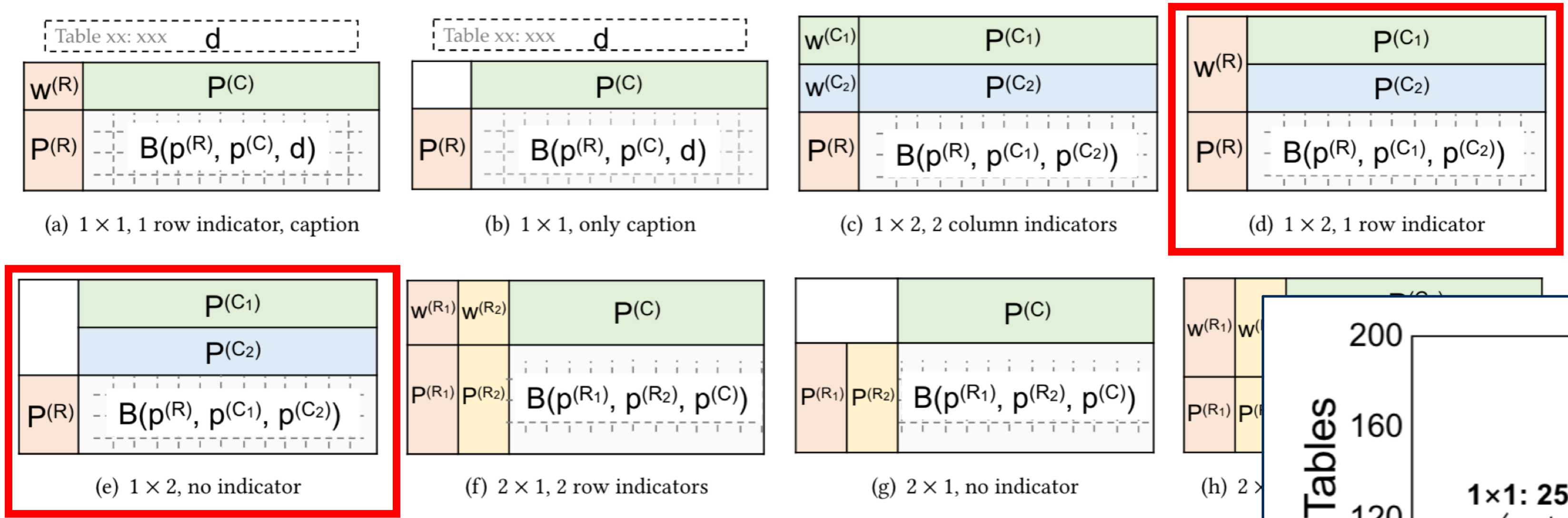


Figure 3: Eight major table templates: We will use the first seven templates which cover more tha dataset. The cells in the table's body are triplets based on rows/columns/caption. (Best viewed in co

Top 2 popular templates

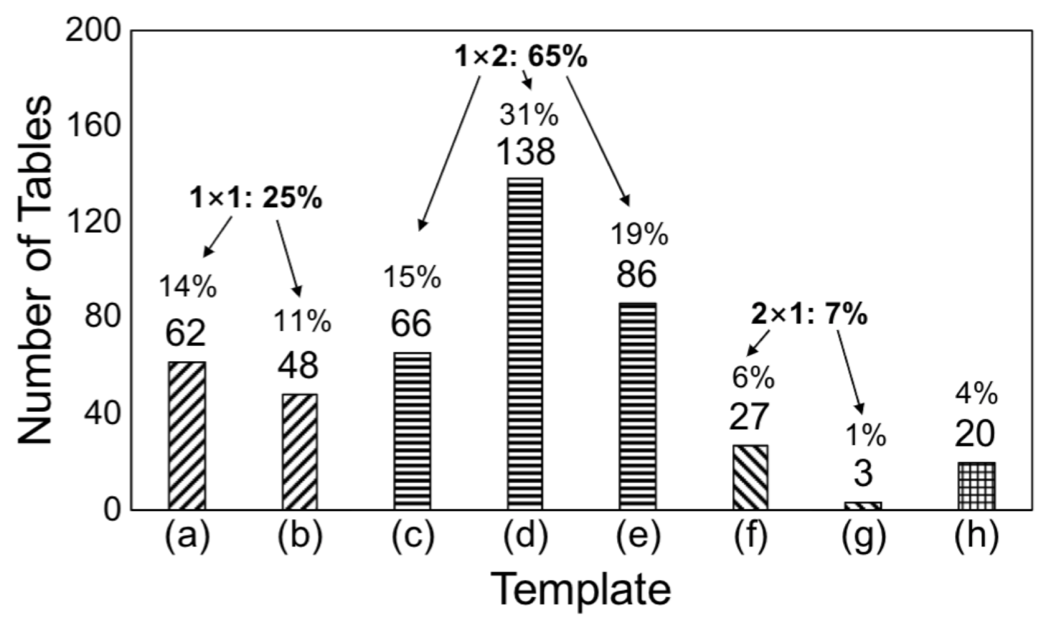


Figure 5: The distribution of table templates.

Problem Definition

Table 4: Performance on the Twitter testing data set by different approaches.

d

$w^{(R)}$ $P^{(C)}$ $P^{(R)}$

Algorithm	Precision	Recall	F1	Accuracy
Textual	0.746	0.693	0.727	0.722
Visual	0.584	0.561	0.573	0.553
Early Fusion	0.730	[illegible]	[illegible]	0.717
Late Fusion	0.634	0.610	0.622	0.604
CCR	**0.831**	**0.805**	**0.818**	**0.809**

B(·, ·, ·)

Dataset	Method	Metric	Score
Twitter	Textual	Precision	0.746
Twitter	Textual	Recall	0.693
…	…	…	…
Twitter	CCR	F1	0.818
Twitter	CCR	Accuracy	0.809

$$\mathcal{P} = \cup_{T=[\mathcal{R}, \mathcal{C}, d, \mathcal{B}]} P^{(R_{(:)})} \cup P^{(C_{(:)})}, \Rightarrow \mathcal{L} = \{\text{“}method\text{”}, \text{“}dataset\text{”}, \text{“}metric\text{”}\}.$$

Problem: Given a set of tables extracted from PDFs $\{T\}$,
(1) classify the concepts into three categories $f: \mathcal{P} \rightarrow \mathcal{L}$
(2) unify the cells into (method, dataset, metric, score)-tuples.

Ensemble Learning

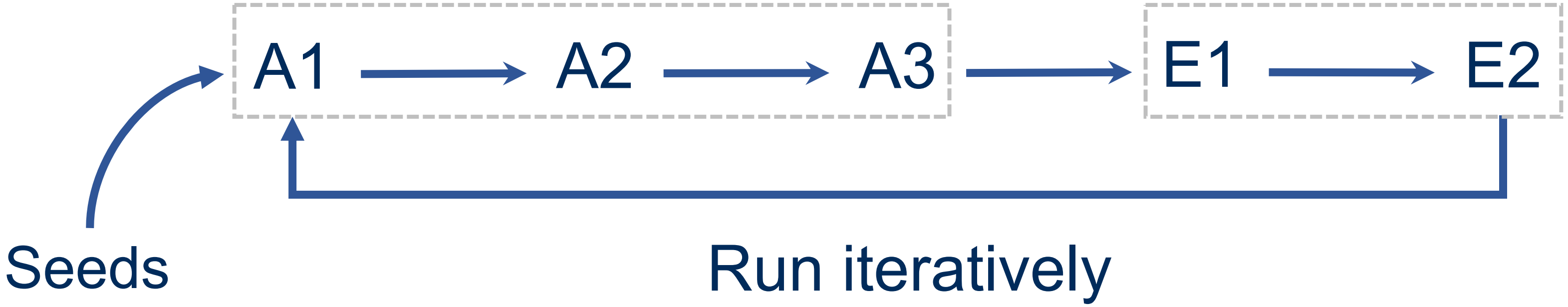
Concept-to-Label $f : \mathcal{P} \rightarrow \mathcal{L}$

Rule-based classifiers

- Three Assumptions

Learning-based classifiers

- Semantic concept Embeddings
- Structural concept Embeddings



Assumption 1

Row/column header indication. If the upper-leftmost cell of the table has a specific word (e.g., “Methods”, “Algorithm”), the names on the corresponding columns/rows are more likely to have the label as the word indicates.

Table 4: Performance on the Twitter testing data set by different approaches. d

$w^{(R)}$ Algorithm	Precision	Recall	F1 $P^{(C)}$	Accuracy
Textual	0.746	0.693	0.727	0.722
Visual	0.584	0.561	0.573	0.553
$P^{(R)}$ Early Fusion	0.730	0.[illegible] $B(\cdot,\cdot,\cdot)$	[illegible]37	0.717
Late Fusion	0.634	0.610	0.622	0.604
CCR	**0.831**	**0.805**	**0.818**	**0.809**

$$\min_{\phi,\psi} J_1(\phi,\psi) = \sum_{T=[\mathcal{R},\mathcal{C},\dots]} \sum_{(w,P)\in\mathcal{R}\cup\mathcal{C}} \sum_{l\in\mathcal{L}} \left(\sum_{p\in P} \phi(p \in P^{(l)}) - |P| \cdot \psi(w \in W^{(l)}) \right)^2, \quad (6)$$

label prediction ϕ word indication ψ

Assumption 2

Row/column type consistency. Concepts on the same column/row are likely to have the same type of label. For example, if we know "Precision" is a "metric", then "Recall" is likely to be a "metric".

Table 4: Performance on the Twitter testing data set by different approaches. d

$w^{(R)}$ $P^{(C)}$ $P^{(R)}$ $B(\cdot, \cdot, \cdot)$

Algorithm	Precision	Recall	F1	Accuracy
Textual	0.746	0.693	0.727	0.722
Visual	0.584	0.561	0.573	0.553
Early Fusion	0.730	[illegible]	[illegible]	0.717
Late Fusion	0.634	0.610	0.622	0.604
CCR	**0.831**	**0.805**	**0.818**	**0.809**

$$\max_{\phi} J_2(\phi) = \sum_{T=[\mathcal{R}, C, \dots]} \sum_{P \in \mathcal{R} \cup C} \sum_{p \in P} \phi(p \in P^{(l^*(P))}), \qquad (8)$$

majority of the concepts

Assumption 3

Cell context completeness. A table often **covers all the three types** of labels on its columns, rows, and caption, in order to provide complete contexts to explain the values in the cells. For example, if the caption has a dataset name and row names are methods, then the column names are likely to be metric.

Table 4: Performance on the Twitter testing data set by different approaches. a

Algorithm	Precision	Recall	F1	Accuracy
Textual	0.746	0.693	0.727	0.722
Visual	0.584	0.561	0.573	0.553
Early Fusion	0.730	0.7[illegible]	[illegible]37	0.717
Late Fusion	0.634	0.610	0.622	0.604
CCR	**0.831**	**0.805**	**0.818**	**0.809**

$W^{(R)}$ $P^{(R)}$ $P^{(C)}$ $B(\cdot,\cdot,\cdot)$

$$\max_{\phi} J_3(\phi) = \sum_{T=[\ldots,\mathcal{B}(B_1,B_2,B_3)]} |\cup_{k\in\{1,2,3\}} l_k^*|. \qquad (10)$$

Learning-based Classifier

Semantic concept embeddings (BERT[1])

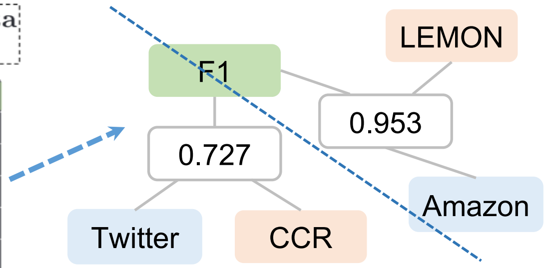
[Paper text] On the other hand, the proposed CCR model can improve the performance of both precision and recall than the two single models. Meanwhile, CCR performs best among all the methods in terms of both F1 and accuracy score.

Structural concept embeddings (HEBE[2])

Table 4: Performance on the Twitter testing data set by different approaches. d

$w^{(R)}$ $p^{(R)}$ $p^{(C)}$ $B(\cdot,\cdot,\cdot)$

Algorithm	Precision	Recall	F1	Accuracy
Textual	0.746	0.693	0.727	0.722
Visual	0.584	0.561	0.573	0.553
Early Fusion	0.730	[illegible]	[illegible]	0.717
Late Fusion	0.634	0.610	0.622	0.604
CCR	**0.831**	**0.805**	**0.818**	**0.809**



Seen Concepts

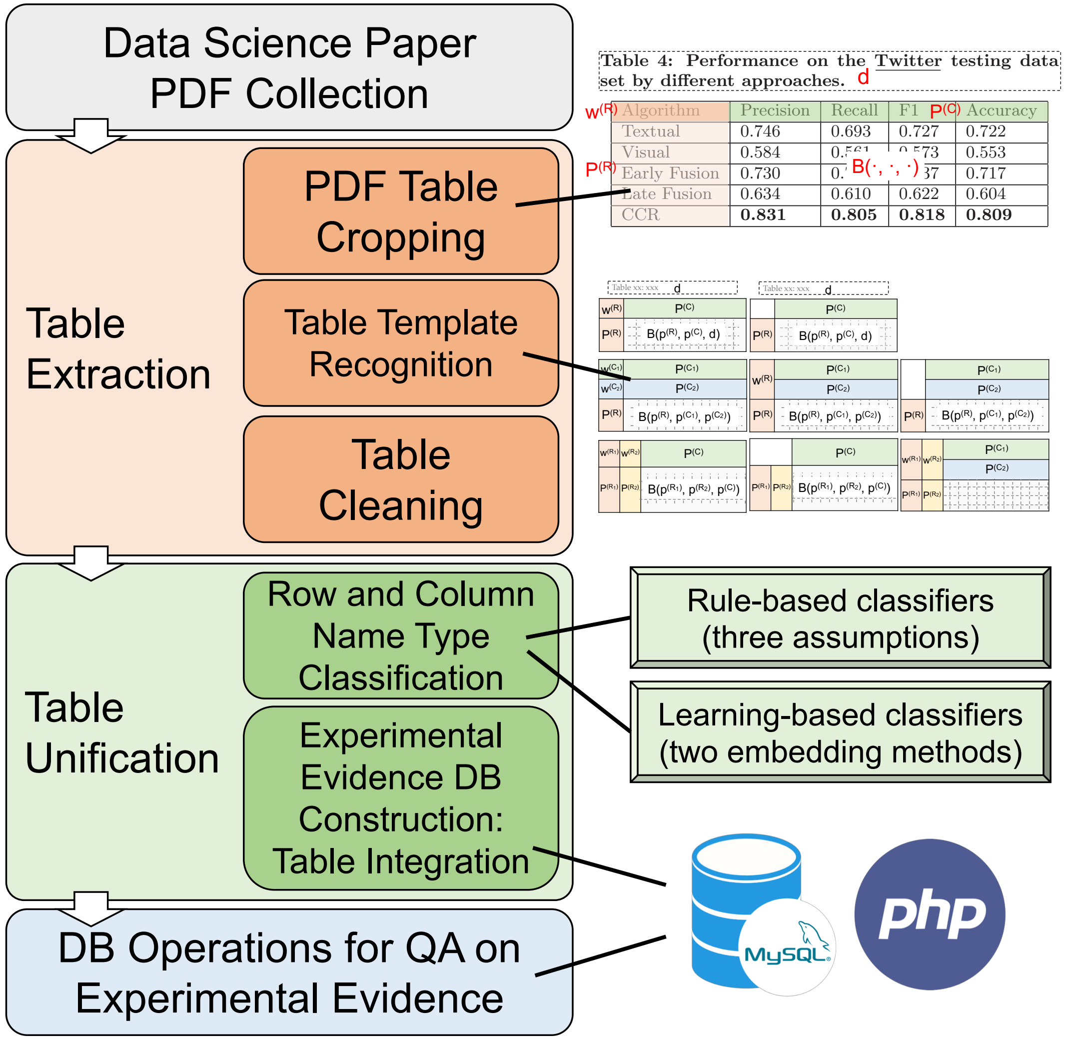
LEMON ⟶ Method
Amazon ⟶ Dataset
Precision ⟶ Metric
… …

Unseen Concepts

CCR ⟶ ?
Twitter ⟶ ?
… …

[1] Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL* 2019.
[2] Gui et al., Embedding learning with events in heterogeneous information networks. In *TKDE* 2017.

Review: Tablepedia System



Results

	Rule-based (Assumptions:)			Learning-based (Embeddings:)		Ensembled
	A1: Header indication	**A2**: Type consistency	**A3**: Completeness	**E1**: Structural	**E2**: Semantic	
TableUni-R	✔	✔	✔	✗	✗	✗
TableUni-L	✗	✗	✗	✔	✔	✗
TableUni-(R+E1)	✔	✔	✔	✔	✗	✔
TableUni-(R+E2)	✔	✔	✔	✗	✔	✔
TableUni-(A1+L)	✔	✗	✗	✔	✔	✔
TableUni-(A2+L)	✗	✔	✗	✔	✔	✔
TableUni-(A3+L)	✗	✗	✔	✔	✔	✔
TableUni-(R+L)	✔	✔	✔	✔	✔	✔

Method	**Micro F1**	**Macro F1**
TableUni-R	0.6908 (0.0040)	0.6542 (0.0047)
TableUni-L	0.6333 (0.0024)	0.6072 (0.0021)
TableUni-(R+E1)	0.7505 (0.0039)	0.7115 (0.0053)
TableUni-(R+E2)	0.8175 (0.0021)	0.7798 (0.0029)
TableUni-(A1+L)	0.6980 (0.0024)	0.6612 (0.0026)
TableUni-(A2+L)	0.7567 (0.0037)	0.7179 (0.0046)
TableUni-(A3+L)	0.6474 (0.0032)	0.6129 (0.0038)
TableUni-(R+L)	**0.8307** (0.0022)	**0.8104** (0.0023)

R > L

Rule is better than Learning.

A2 > A1 > A3

Type consistency is the most effective.

Semantic embedding is more effective than structural.

E1 > E2

R+L is the best!

Using all the Five (Three plus Two) is the best!

Results (cont'd)

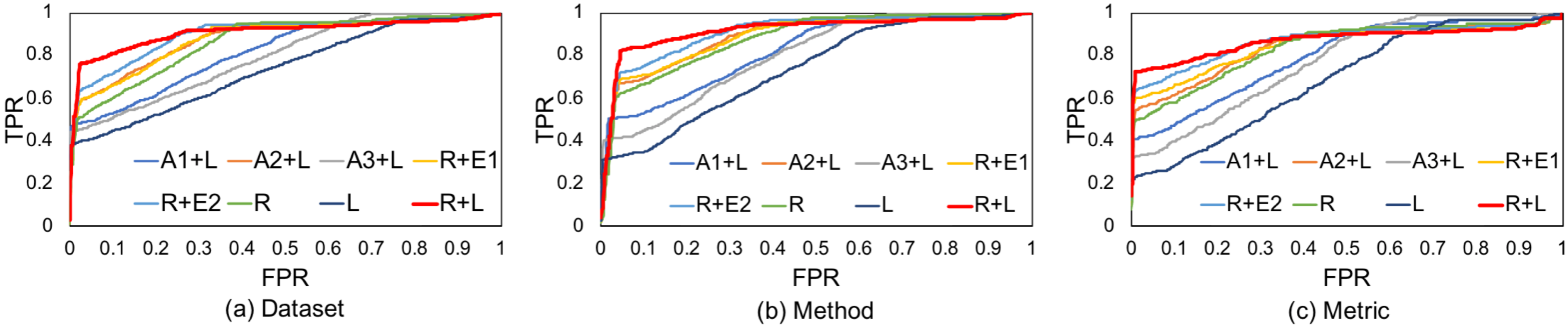


Figure 6: ROC curves comparing the variants of our proposed TableUni methods with respect to the type of classes.

- Rule is better than Learning.
- Type consistency (Rule 2) is the most effective.
- Semantic embedding is more effective than structrual embedding.
- Rule + Learning is the best!

Results (RecSys)

(ACM TOIS 2011)

Table III. MAE Comparison with Other Approaches on Epinions Dataset

Methods		90% Training	80% Training	70% Training	60% Training
User Mean		0.9294	0.9319	0.9353	0.9384
Item Mean		0.8936	0.9115	0.9316	0.9528
Trust		0.9005	0.9044	0.9082	0.9153
5D	NMF	0.8938	0.8975	0.9229	0.9430
	SVD	0.8739	0.8946	0.9214	0.9421
	PMF	0.8678	0.8946	0.9127	0.9350
	SoRec	**0.8442**	**0.8638**	**0.8751**	**0.8948**
10D	NMF	0.8712	0.8951	0.9211	0.9408
	SVD	0.8702	0.8921	0.9189	0.9382
	PMF	0.8651	0.8886	0.9092	0.9328
	SoRec	**0.8404**	**0.8580**	**0.8722**	**0.8921**

(ACM TOIS 2011)

Table IV. RMSE Comparison with Other Approaches on Epinions Dataset

Methods		90% Training	80% Training	70% Training	60% Training
User Mean		1.1927	1.1968	1.2014	1.2082
Item Mean		1.1678	1.1973	1.2276	1.2505
Trust		1.1697	1.1761	1.1797	1.1894
5D	NMF	1.1649	1.1861	1.2090	1.2311
	SVD	1.1635	1.1845	1.2067	1.2298
	PMF	1.1583	1.1798	1.2008	1.2271
	SoRec	**1.1333**	**1.1530**	**1.1690**	**1.1892**
10D	NMF	1.1621	1.1832	1.2073	1.2294
	SVD	1.1600	1.1812	1.2011	1.2268
	PMF	1.1544	1.1760	1.1968	1.2230
	SoRec	**1.1293**	**1.1492**	**1.1660**	**1.1852**

(ACM TIST 2011)

Table III. Performance Comparisons (A Smaller MAE or RMSE Value Means a Better Performance)

Training Data	Metrics	Dimensionality = 5							
		UserMean	ItemMean	NMF	PMF	TCF	Trust	SoRec	RSTE
90%	MAE	0.9134	0.9768	0.8738	0.8676	0.9005	0.9054	0.8442	**0.8377**
	RMSE	1.1688	1.2375	1.1649	1.1575	1.1697	1.1959	1.1333	**1.1109**
80%	MAE	0.9285	0.9913	0.8975	0.8951	0.9044	0.9221	0.8638	**0.8594**
	RMSE	1.1817	1.2584	1.1861	1.1826	1.1761	1.2140	1.1530	**1.1346**
Training Data	**Metrics**	**Dimensionality = 10**							
		UserMean	ItemMean	NMF	PMF	TCF	Trust	SoRec	RSTE
90%	MAE	0.9134	0.9768	0.8712	0.8651	0.9005	0.9039	0.8404	**0.8367**
	RMSE	1.1688	1.2375	1.1621	1.1544	1.1697	1.1917	1.1293	**1.1094**
80%	MAE	0.9285	0.9913	0.8951	0.8886	0.9044	0.9215	0.8580	**0.8537**
	RMSE	1.1817	1.2584	1.1832	1.1760	1.1761	1.2132	1.1492	**1.1256**

(WSDM 2011)

Table 5: Performance Comparisons (Dimensionality = 10)

Dataset	Training	Metrics	UserMean	ItemMean	NMF	PMF	RSTE	$SR1_{vss}$	$SR1_{pcc}$	$SR2_{vss}$	$SR2_{pcc}$
Douban	80%	MAE	0.6809	0.6288	0.5732	0.5693	0.5643	0.5579	0.5576	0.5548	**0.5543**
		Improve	18.59%	11.85%	3.30%	2.63%	1.77%				
		RMSE	0.8480	0.7898	0.7225	0.7200	0.7144	0.7026	0.7022	0.6992	**0.6988**
		Improve	17.59%	11.52%	3.28%	2.94%	2.18%				
	60%	MAE	0.6823	0.6300	0.5768	0.5737	0.5698	0.5627	0.5623	0.5597	**0.5593**
		Improve	18.02%	11.22%	3.03%	2.51%	1.84%				
		RMSE	0.8505	0.7926	0.7351	0.7290	0.7207	0.7081	0.7078	0.7046	**0.7042**
		Improve	17.20%	11.15%	4.20%	3.40%	2.29%				
	40%	MAE	0.6854	0.6317	0.5899	0.5868	0.5767	0.5706	0.5702	0.5690	**0.5685**
		Improve	17.06%	10.00%	3.63%	3.12%	1.42%				
		RMSE	0.8567	0.7971	0.7482	0.7411	0.7295	0.7172	0.7169	0.7129	**0.7125**
		Improve	16.83%	10.61%	4.77%	3.86%	2.33%				
Epinions	90%	MAE	0.9134	0.9768	0.8712	0.8651	0.8367	0.8290	0.8287	0.8258	**0.8256**
		Improve	9.61%	15.48%	5.23%	4.57%	1.33%				
		RMSE	1.1688	1.2375	1.1621	1.1544	1.1094	1.0792	1.0790	1.0744	**1.0739**
		Improve	8.12%	13.22%	7.59%	6.97%	3.20%				
	80%	MAE	0.9285	0.9913	0.8951	0.8886	0.8537	0.8493	0.8491	0.8447	**0.8443**
		Improve	9.07%	14.83%	5.68%	4.99%	1.10%				
		RMSE	1.1817	1.2584	1.1832	1.1760	1.1256	1.1016	1.1013	1.0958	**1.0954**
		Improve	7.30%	12.95%	7.42%	6.85%	2.68%				

MAE on Epinions (80% Training)

RMSE on Epinions (80% Training)

Best baseline vs the proposed

Conflicting between papers

	A	B	C	D	E
1	**Method**	**Dataset**	**Metric**	**Score**	**Source**
10	UserMean	Epinions	MAE	0.9319	TOIS11-paper7-table3
11	UserMean	Epinions	MAE	0.9285	TIST11-paper3-table3
12	UserMean	Epinions	MAE	0.9285	WSDM11-paper12-table5
109	ItemMean	Epinions	RMSE	1.1973	TOIS11-paper7-table4
110	ItemMean	Epinions	RMSE	1.2584	TIST11-paper3-table3
111	ItemMean	Epinions	RMSE	1.2584	WSDM11-paper12-table5
112	Trust	Epinions	RMSE	1.2132	TIST11-paper3-table3
113	NMF	Epinions	RMSE	1.1832	TOIS11-paper7-table4
114	NMF	Epinions	RMSE	1.1832	TIST11-paper3-table3
115	NMF	Epinions	RMSE	1.1832	WSDM11-paper12-table5
116	SVD	Epinions	RMSE	1.1812	TOIS11-paper7-table4
117	TCF	Epinions	RMSE	1.1761	TIST11-paper3-table3
118	PMF	Epinions	RMSE	1.1760	TOIS11-paper7-table4
119	PMF	Epinions	RMSE	1.1760	TIST11-paper3-table3
120	PMF	Epinions	RMSE	1.1760	WSDM11-paper12-table5
121	SoRec	Epinions	RMSE	1.1492	TOIS11-paper7-table4
122	RSTE	Epinions	RMSE	1.1256	TIST11-paper3-table3
123	RSTE	Epinions	RMSE	1.1256	WSDM11-paper12-table5
124	SR1VSS	Epinions	RMSE	1.1016	WSDM11-paper12-table5
125	SR1PCC	Epinions	RMSE	1.1013	WSDM11-paper12-table5
126	SRCVSS	Epinions	RMSE	1.0958	WSDM11-paper12-table5
127	SR2PCC	Epinions	RMSE	1.0954	WSDM11-paper12-table5
169	SoRec	MovieLens	RMSE	...	...

Results: Asking ERD

Question 1: Find related methods, metrics, and datasets.

Query: How many methods were used for the Epinions dataset?

select count(distinct Method) from ERD where Dataset=“Epinions”

36. (“UserMean”, “ItemMean”, “Trust”, “NMF”, “SVD”, “TCF” …)

Query: How many metrics were used to evaluate Amazon dataset?

select count(distinct Metric) from ERD where Dataset=“Amazon”

15. (“Precision”, “Recall”, “F1”, “Accuracy”, etc …)

Query: How many datasets used with Amazon in the same table?

select count(distinct Dataset) from ERD where Source=(select (distinct Source) from ERD where Dataset= “Amazon”);

53. (“DBLP”, “Wikipedia”, “Delicious”, “Epinions”, etc …)

Question 2: Find top-performing methods on a dataset.

Query: What are the top 3 methods on Amazon in terms of F1?

select Method, Score from ERD where Dataset = “Amazon” and Metric = “F1” order by Score limit 3;

“LEMON” (0.953), “LEMON-auto” (0.91), “LC” (0.815).

Question 2: Find top-performing methods on a dataset.

Query: What are top 3 methods on Epinions in terms of RMSE?

select Method, Score from ERD where Dataset = “Epinion” and Metric = “RMSE” order by Score limit 3;

“SR2pcc” (1.0954), “SR2vss” (1.0958), “SR1pcc” (1.1013).

Question 3: Find conflicting reported numbers.

Dataset	(%)	SLEEC	FastXML	PfastreXML	PDSparse
AmazonCat	P@1	90.56/89.19	94.02/93.10	86.06/89.94	87.43/89.31
-13K	P@3	76.96/75.17	79.93/78.18	86.06/77.24	87.43/74.03
	P@5	62.63/61.09	64.90/63.38	63.65/63.53	56.70/60.11
Delicious	P@1	47.78/47.03	48.85/43.20	26.66/37.62	37.69/34.37
-200K	P@3	42.05/41.67	42.84/38.68	23.56/35.62	30.16/29.48
	P@5	39.29/38.88	39.83/36.21	23.21/34.03	27.01/27.04
WikiLSHTC	P@1	58.34/55.57	50.01/49.75	57.17/58.10	60.70/61.26
-325K	P@3	36.70/33.06	32.83/33.10	37.03/37.61	39.62/39.48
	P@5	26.45/24.07	24.13/24.45	27.19/27.69	29.20/28.79

Table 1: Our system found inconsistent precision scores reported by two papers [42] (left numbers) and [36] (right numbers) in ACM SIGKDD 2017 Research Track for multi-label classification. Precision differences of bigger than 3% are underlined, which has been able to be claimed as significant improvement on the well-accepted benchmarks.