USC Viterbi Information Sciences Institute

THE OHIO STATE UNIVERSITY COLLEGE OF ENGINEERING

From Tables to Knowledge
Recent advances in table understanding

Jay Pujara Pedro Szekely Huan Sun Muhao Chen

Meet the presenters

Jay Pujara

USC Viterbi Information Sciences Institute

Pedro Szekely

USC Viterbi Information Sciences Institute

Huan Sun

OHIO STATE

Muhao Chen

USC Viterbi Information Sciences Institute

Outline

Time (SGT/PDT)	**Presenter**	**Topic**
00:00 – 00:35 09:00 – 09:35	Jay Pujara	Table structures
00:40 – 01:15 09:40 – 10:15	Pedro Szekely	Semantic models of tables
	Break	
01:30 – 02:05 10:30 – 11:05	Huan Sun	Neural representation learning
02:10 – 02:45 11:10 – 11:45	Muhao Chen	Downstream tasks
	Questions & discussion	

Asking questions

Please wait until the end of each tutorial section to ask questions about the material

Tutors will be available after the end of the tutorial for more open-ended discussion

Why tables?

We use tables to convey important information

	Race/Ethnicity	Non-hospitalized	Non-fatal hospitalized	Confirmed deaths[1]	Probable deaths[2]	Total deaths
Count (% of known)	Black/African American	15927 (28.7%)	9432 (34.7%)	4239 (29.4%)	1330 (32.9%)	5569 (30.2%)
	Hispanic/Latino[3]	17036 (30.7%)	8780 (32.3%)	4513 (31.3%)	1183 (29.2%)	5696 (30.9%)
	White	18170 (32.8%)	6614 (24.3%)	3882 (26.9%)	1119 (27.6%)	5001 (27.1%)
	Asian/Pacific Islander	4088 (7.4%)	2128 (7.8%)	1143 (7.9%)	389 (9.6%)	1532 (8.3%)
	Other known[6]	237 (0.4%)	233 (0.9%)	630 (4.4%)	27 (0.7%)	657 (3.6%)
Count (% of total)	Total known	55458 (40.9%)	27187 (77.2%)	14407 (93.9%)	4048 (80.0%)	18455 (90.4%)
	Other/Unknown	80286 (59.1%)	8013 (22.8%)	942 (6.1%)	1009 (20.0%)	1951 (9.6%)
	Total	135744	35200	15349	5057	20406

Example: COVID cases

Humans can quickly understand these tables and extract knowledge.

Can AIs?

Race/Ethnicity	No. Cases	Percent Cases	No. Deaths	Percent Deaths	Percent CA population
Latino	887,580	55.6	11,575	47.4	38.9
White	319,136	20.0	7,679	31.4	36.6
Asian	100,569	6.3	2,818	11.5	15.4
African American	64,518	4.0	1,697	6.9	6.0
Multi-Race	20,372	1.3	252	1.0	2.2
American Indian or Alaska Native	5,090	0.3	81	0.3	0.5

Expressing that information is tedious otherwise

Race/Ethnicity	Non-hospitalized	Non-fatal hospitalized	Confirmed deaths[1]	Probable deaths[2]	Total deaths
Black/African American	15927 (28.7%)	9432 (34.7%)	4239 (29.4%)	1330 (32.9%)	5569 (30.2%)

Among African Americans there were 5,569 total deaths, representing 30.2% of total deaths. Of the 5,569 total deaths, 1,330 deaths were probable deaths (representing 32.9% of all probable deaths) while 4,239 were confirmed deaths (representing 29.4% of all confirmed deaths). Additionally, there were 9,432 non-fatal hospitalizations (representing 34.7% of all non-fatal hospitalizations) and 15,927 cases that did not require hospitalizations (representing 28.7% of all cases where hospitalization was unnecessary).

Tables (and surface forms) are diverse

	Race/Ethnicity	Non-hospitalized	Non-fatal hospitalized	Confirmed deaths[1]	Probable deaths[2]	Total deaths
Count (% of known)	Black/African American	15927 (28.7%)	9432 (34.7%)	4239 (29.4%)	1330 (32.9%)	5569 (30.2%)
	Hispanic/Latino[3]	17036 (30.[illegible]%)	8780 (32.3%)	4513 (31.3%)	1183 (29.2%)	5696 (30.9%)
	White	18170 (32.8%)	6614 (24.3%)	3882 (26.9%)	1119 (27.6%)	5001 (27.1%)
	Asian/Pacific Islander	4088 (7.4%)	2128 (7.8%)	1143 (7.9%)	389 (9.6%)	1532 (8.3%)
	Other known[6]	237 (0.4%)	233 (0.9%)	630 (4.4%)	27 (0.7%)	657 (3.6%)
Count (% of total)	Total known	55458 (40.9%)	27187 (77.2%)	14407 (93.9%)	4048 (80.0%)	18455 (90.4%)
	Other/Unknown	80286 (59.[illegible]%)	8013 (22.8%)	942 (6.1%)	1009 (20.0%)	1951 (9.6%)
	Total	135744	35200	15349	5057	20406

Example: COVID cases

- Labels for classes differ
- Percentage and units: data, metadata, or header?
- What percentage?
- Which CA?

Race/Ethnicity	No. Cases	Percent Cases	No. Deaths	Percent Deaths	Percent CA population
Latino	887,580	55.6	11,575	47.4	38.9
White	319,136	20.0	7,679	31.4	36.6
Asian	100,569	6.3	2,818	11.5	15.4
African American	64,518	4.0	1,697	6.9	6.0
Multi-Race	20,372	1.3	252	1.0	2.2
American Indian or Alaska Native	5,090	0.3	81	0.3	0.5

Tables structures can be complex

Often data in the real world differs from "neat" dataframes.

Data spec in a cell next to months

Months in Roman numerals

Part I Selected monthly macroeconomic indicators						
Data updating:	24.07.2018					
Specification A – corresponding period of the previous year=100 A_1 – from the beginning of the year to the end of the period (corresponding period of the previous year=100) B – the previous period=100 C – December of the previous year=100 I_2 – monthly average of 2005=100 I_3 – monthly average of 2010=100 I_4 – monthly average of 2010=100		2009				
		IV	V	VI	VII	VIII
Average paid employment in an enterprise sector[a]	in thous.	5,309	5,292	5,280	5,273	5,270
	A	98.6	98.3	98.1	97.8	97.8
	B	99.7	99.7	99.8	99.9	99.9
	I_2	111.0	110.7	110.5	110.4	110.3
	I_3					
	I_4					
Registered unemployed persons (end of the period)	in thous.	1,719.9	1,683.4	1,658.7	1,676.1	1,689.0
	A	107.1	110.3	114.0	117.8	120.3
	B	97.8	97.9	98.5	101.1	100.8

Lots of sheets, each one different

Spis tablic | 1.1.1 | 1.1.2 | 1.2 | 1.3 | 1.4 | 1.5 | 1.6 | 1.7 | 1.8 | 1.9.1 | 1.9.2 | 1.10.1

ATA EARS	Ludność *Population*		Małżeń-stwa *Marriages*	Rozwody *Divorces*	Urodzenia żywe *Live births*	Zgony *Deaths*		Przyrost naturalny *Natural increase*	Migracje wewnętrzne *Internal migration*			Migracje zagraniczne[c] *International migration*[c]			Ogólne saldo migracji *Total net migration*	Małżeń-stwa *Marriages*	Rozwody *Divorces*	Urodz żyw *Live bi*
	stan w dniu 30 VI *as of June 30*	stan w dniu 31 XII *as of De-cember 31*				ogółem *total*	w tym niemowląt *of which infant*		napływ *inflow*	odpływ *outflow*	saldo *net*	imigracja *immigra-tion*	emigracja *emigra-tion*	saldo *net*				
	w tysiącach *in thousands*															na 1000 ludnoś		

Holy dimensional nesting, Batman!

URBAN AREAS		
Age groups	1989	1990
TOTAL	**23384**	**23546**
0 - 2 years	964	920
3 -6	1571	1523
7 - 14	3090	3054
15 - 18	1415	1446
of which: 18	332	342
19 - 24	1829	1897
0 - 17	6708	6600
18 -59/64	14020	14209
of which:		
18 - 44	9787	9838
45 - 59/64	4233	4371
60/65 years and more	2656	2737
0 - 14	5626	5496
15 - 64	15689	15904
65 years and more	2070	2147
Women at childbearing 15 - 49 years	6182	6255
TOTAL	**x**	**162.0**
0-2 years	x	-44.0
3 -6	x	-48.0
7 - 14	x	-36.0
15 - 18	x	31.0
of which: 18	x	10.0
19 - 24	x	68.0
0 - 17	x	-108.0
18 -59/64	x	189.0
of which: 18		
18 - 44	x	51.0

Table segmentation

Core tasks in table understanding

structure

- table segmentation
- cell role prediction
- functional block detection
- join identification

Core tasks in table understanding

structure	table segmentation	cell role prediction	functional block detection	join identification
knowledge alignment	semantic typing	semantic modeling	schema mapping	entity linking
	representation	completion	knowledge extraction	augmentation

Core tasks in table understanding

structure	table segmentation	cell role prediction	functional block detection	join identification
knowledge alignment	semantic typing	semantic modeling	schema mapping	entity linking
	representation	completion	knowledge extraction	augmentation
downstream tasks	retrieval	summarization	question answering	verification

Core tasks in table understanding

structure	table segmentation	cell role prediction	functional block detection	join identification
knowledge alignment	semantic typing	semantic modeling	schema mapping	entity linking
	representation	completion	knowledge extraction	augmentation
downstream tasks	retrieval	summarization	question answering	verification

Table design principles

Principles of (human) table organization

- Keep it simple:
  - only provide necessary information for the use case
  - explicitly provide the information necessary – avoid reader computation
  - reduce the number of categories and subcategories to find data
  - avoid 2-dimensional tables
  - round numbers

- Use spatial cues:
  - reading often left-to-right, top-to-bottom – important information first
  - keep related items nearby
  - use spacing, lines/rules, font style and weight, and indentation to organize
  - easier to compare values in a column
  - sort items based on salient value
  - span redundant values

X. Wang. Tabular Abstraction, Editing, and Formatting. Thesis, 1996

Principles of (human) table organization

	Race/Ethnicity	Non-hospitalized	Non-fatal hospitalized	Confirmed deaths[1]	Probable deaths[2]	Total deaths
Count (% of known)	Black/African American	15927 (28.7%)	9432 (34.7%)	4239 (29.4%)	1330 (32.9%)	5569 (30.2%)
	Hispanic/Latino[3]	17036 (30.7%)	8780 (32.3%)	4513 (31.3%)	1183 (29.2%)	5696 (30.9%)
	White	18170 (32.8%)	6614 (24.3%)	3882 (26.9%)	1119 (27.6%)	5001 (27.1%)
	Asian/Pacific Islander	4088 (7.4%)	2128 (7.8%)	1143 (7.9%)	389 (9.6%)	1532 (8.3%)
	Other known[6]	237 (0.4%)	233 (0.9%)	630 (4.4%)	27 (0.7%)	657 (3.6%)
Count (% of total)	Total known	55458 (40.9%)	27187 (77.2%)	14407 (93.9%)	4048 (80.0%)	18455 (90.4%)
	Other/Unknown	80286 (59.1%)	8013 (22.8%)	942 (6.1%)	1009 (20.0%)	1951 (9.6%)
	Total	135744	35200	15349	5057	20406

- Keep it simple:
  - only provide necessary information for the use case
  - explicitly provide the information necessary – avoid reader computation
  - reduce the number of categories and subcategories to find data
  - avoid 2-dimensional tables
  - round numbers

Principles of (human) table organization

	Race/Ethnicity	Non-hospitalized	Non-fatal hospitalized	Confirmed deaths[1]	Probable deaths[2]	Total deaths
Count (% of known)	Black/African American	15927 (28.7%)	9432 (34.7%)	4239 (29.4%)	1330 (32.9%)	5569 (30.2%)
	Hispanic/Latino[3]	17036 (30.7%)	8780 (32.3%)	4513 (31.3%)	1183 (29.2%)	5696 (30.9%)
	White	18170 (32.8%)	6614 (24.3%)	3882 (26.9%)	1119 (27.6%)	5001 (27.1%)
	Asian/Pacific Islander	4088 (7.4%)	2128 (7.8%)	1143 (7.9%)	389 (9.6%)	1532 (8.3%)
	Other known[6]	237 (0.4%)	233 (0.9%)	630 (4.4%)	27 (0.7%)	657 (3.6%)
Count (% of total)	Total known	55458 (40.9%)	27187 (77.2%)	14407 (93.9%)	4048 (80.0%)	18455 (90.4%)
	Other/Unknown	80286 (59.1%)	8013 (22.8%)	942 (6.1%)	1009 (20.0%)	1951 (9.6%)
	Total	135744	35200	15349	5057	20406

- Use spatial cues:
  - reading often left-to-right, top-to-bottom – important information first
  - keep related items nearby
  - use spacing, lines/rules, font style and weight, and indentation to organize
  - easier to compare values in a column
  - sort items based on salient value, span redundant values

Guide to table structures

Parts of a table

Table 1: Number of acceptances by category of work in conferences[1]						
Year	**Conference**	**Papers**		**Workshops**	**Tutorials**	**Total**
		Research	*Industry*			
2021	KDD	800	500	25	40	1365
2021	WWW	700	400	25	40	1165
2020	KDD	600	300	25	35	960
2020	WWW	500	200	25	35	760
1. The source of this data is Jay's made-up data generation						

Parts of a table

Table 1: Number of acceptances by category of work in conferences[1]

Year	**Conference**	**Papers**		**Workshops**	**Tutorials**	**Total**
		Research	*Industry*			
2021	KDD	800	500	25	40	1365
2021	WWW	700	400	25	40	1165
2020	KDD	600	300	cell	35	960
2020	WWW	500	200	25	35	760

1. The source of this data is Jay's made-up data generation

Parts of a table

Table 1: Number of acceptances by category of work in conferences[1]						
Year	**Conference**	**Papers**		**Workshops**	**Tutorials**	**Total**
		Research	*Industry*			
2021	KDD	800	500	25	40	1365
row						
2020	KDD	600	300	25	35	960
2020	WWW	500	200	25	35	760
1. The source of this data is Jay's made-up data generation						

Parts of a table

Table 1: Number of acceptances by category of work in conferences[1]					
Year	**Conference**	**Papers**		**Workshops**	**Tutorials**
		Research	*Industry*		
2021	KDD	800	500	25	40
2021	WWW	700	400	25	40
2020	KDD	600	300	25	35
2020	WWW	500	200	25	35
1. The source of this data is Jay's made-up data generation					

column

Parts of a table

Table 1: Number of acceptances by category of work in conferences[1]

Year	Conference	Papers		Workshops	Tutorials	Total
		Research	*Industry*			
2021	KDD	800	500	25	40	1365
2021	WWW	700				1165
2020	KDD	600		block		960
2020	WWW	500				760

1. The source of this data is Jay's made-up data generation

Parts of a table in stylistic manuals

Stubhead

Boxhead

Table 1: Number of acceptances by category of work in conferences[1]

Year	**Conference**	**Papers**		**Workshops**	**Tutorials**	**Total**
		Research	*Industry*			
2021	KDD	800	500	25	40	1365
2021	WWW	700	400	25	40	1165
2020	KDD	600	300	25	35	960
2020	WWW	500	200	25	35	760

1. The source of this data is Jay's made-up data generation

Stub

Body

Chicago Manual of Style, 1993

Parts of a table in dataset annotations

Metadata

Table 1: Number of acceptances by category of work in conferences[1]

Header

Left Attr.

Derived

Year	Conference	Papers		Workshops	Tutorials	Total
		Research	*Industry*			
2021	KDD	800	500	25	40	1365
2021	WWW	700	400	25	40	1165
2020	KDD	600	300	25	35	960
2020	WWW	500	200	25	35	760

Notes

1. The source of this data is Jay's made-up data generation

Data

Chen, Z., Cafarella, M.: Integrating spreadsheet data via accurate and low-effort extraction. In: KDD 2014.

Koci, E., Thiele, M., Romero Moral, ́O., Lehner, W.: A machine learning approach for layout inference in spreadsheets. In: IC3K 2016.

Parts of a table from a knowledge perspective

The number of acceptances in the workshops category at the KDD conference for the year 2020 was 25 based on the source Jay's made-up data..

Table 1: Number of acceptances item by category of work property in conferences[1]

Year	Conference	Papers		Workshops	Tutorials	Total
properties		*Research*	*Industry*	attribute		
2021	KDD	800	500	25	40	1365
2021	WWW	700	400	25	40	1165
2020 attributes	KDD	600	300	25 value	35	960
2020	WWW	500	200	25	35	760

1. The source of this data property is Jay's made-up data generation attribute

Parts of a table from a knowledge perspective

metadata

Table 1: Number of acceptances by category of work in conferences[1]

headers

attributes

Year	Conference	Papers		Workshops	Tutorials	Total
		Research	*Industry*			
2021	KDD	800	500	25	40	1365
2021	WWW	700	400	25	40	1165
2020	KDD	600	300	25	35	960
2020	WWW	500	200	25	35	760

attributes

values

1. The source of this data is Jay's made-up data generation

metadata

Common relational table structures

Classical table structures - horizontal

iPhone	Released with	Release date	Final supported OS	Support ended
iPhone	iPhone OS 1.0	June 29, 2007	iPhone OS 3.1.3	June 20, 2010
iPhone 3G	iPhone OS 2.0	July 11, 2008	iOS 4.2.1	March 3, 2011
iPhone 3GS	iPhone OS 3.0	June 19, 2009	iOS 6.1.6	September 18, 2013
iPhone 4	iOS 4.0	June 21, 2010	iOS 7.1.2	September 17, 2014
iPhone 4S	iOS 5.0	October 14, 2011	iOS 9.3.5	September 12, 2016
iPhone 5	iOS 6.0	September 21, 2012	iOS 10.3.3	September 18, 2017
iPhone 5C	iOS 7.0	September 20, 2013	iOS 10.3.3	September 18, 2017
iPhone 5S	iOS 7.0	September 20, 2013	latest iOS	(current)
iPhone 6 (Plus)	iOS 8.0	September 19, 2014	latest iOS	(current)
iPhone 6S (Plus)	iOS 9.0	September 25, 2015	latest iOS	(current)
iPhone SE	iOS 9.3	March 31, 2016	latest iOS	(current)
iPhone 7 (Plus)	iOS 10.0	September 16, 2016	latest iOS	(current)
iPhone 8 (Plus)	iOS 11.0	September 22, 2017	latest iOS	(current)
iPhone X	iOS 11.0.1	November 3, 2017	latest iOS	(current)
iPhone XS (Max)	iOS 12	September 21, 2018	latest iOS	(current)
iPhone XR	iOS 12	October 26, 2018	latest iOS	(current)

Classical table structures - vertical

Model [hide]	iPhone 7	iPhone 7 Plus
Picture		
Initial release operating system	iOS 10.0	
Latest release operating system		
In development		
Display	4.7 in (120 mm), 4.1 in (100 mm) by 2.3 in (58 mm), 16:9 aspect ratio, aluminosilicate glass covered 16,777,216-color (24-bit), IPS LCD screen, 1,334 × 750 px screen resolution at 326 ppi, 1400:1 contrast ratio, 625 cd/m² max brightness, LED backlight and fingerprint-resistant oleophobic coating	5.5 in (140 mm), 4.8 in (120 mm) by 2.7 in (69 mm), 16:9 aspect ratio, aluminosilicate glass covered 16,777,216-color (24-bit), IPS LCD screen, 1,920 × 1,080 px (Full HD) screen resolution at 401 ppi, 1300:1 contrast ratio, 625 cd/m² max brightness, LED backlight and fingerprint-resistant oleophobic coating
Storage	32, 128, and 256 GB NAND Flash driven by NVM Express controller	
Processor	2.33 GHz quad-core Apple-designed 64-bit Apple A10 Fusion (4-cores: 2 Hurricane high-performance, 2 Zephyr high-efficiency) with embedded M10 motion coprocessor	
Bus width		
Graphics	Custom Apple PowerVR GT7600 Plus (hexa-core) GPU[7]	
RAM	2 GB LPDDR4 DRAM	3 GB LPDDR4 DRAM

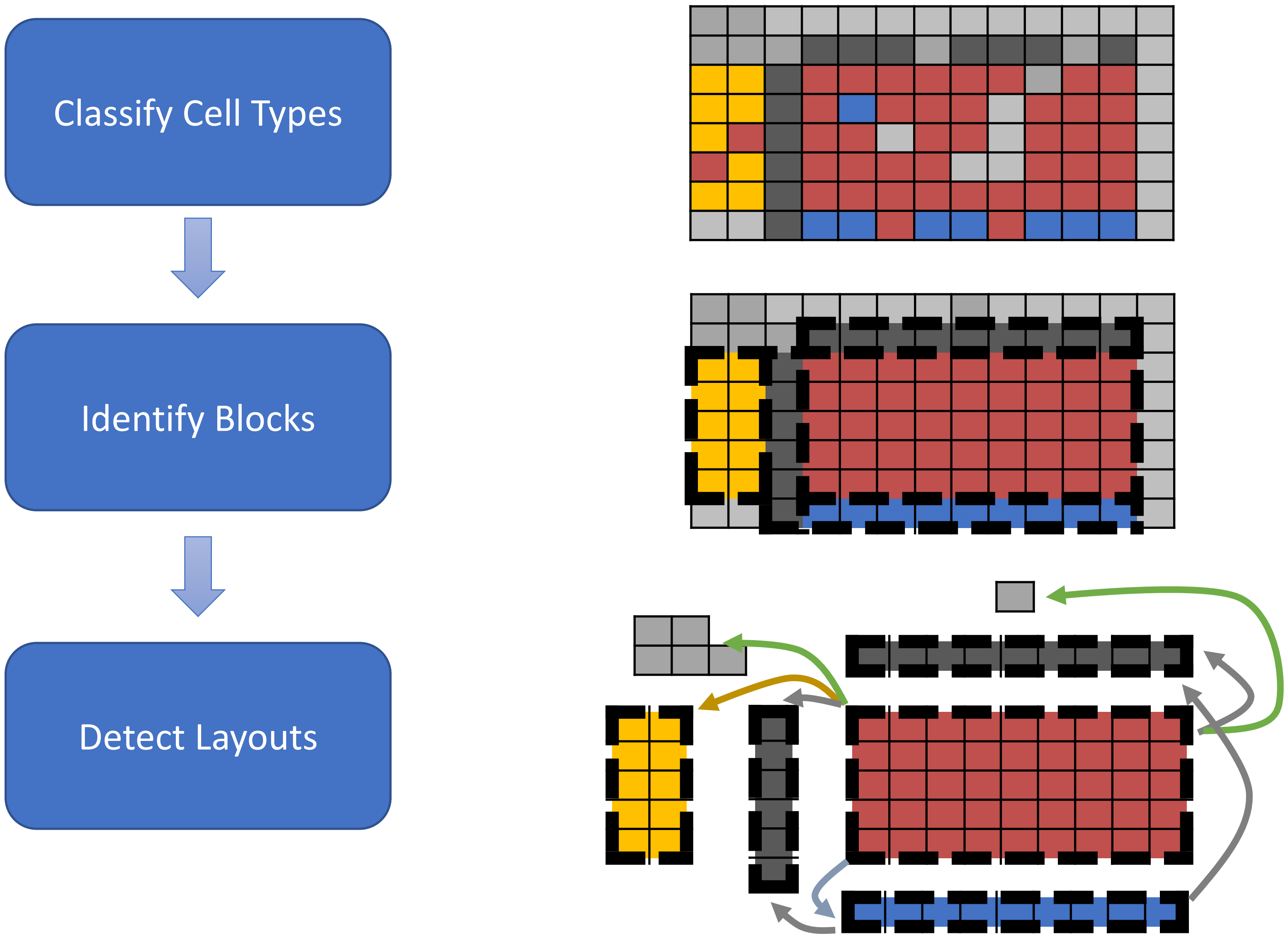
Classical table structures - matrix

U.S. Exports of Crude Oil (Thousand Barrels)

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1920	469	853	892	693	761	627	723	553	790	777	796	823
1921	743	794	750	748	874	586	538	885	881	747	869	525
1922	727	583	806	924	771	826	893	764	1,127	741	879	1,122
1923	762	666	1,028	1,511	1,324	2,598	1,547	1,542	1,592	1,315	1,397	2,103
1924	1,528	1,680	1,550	1,547	1,858	1,542	1,409	1,242	1,893	1,488	1,453	1,049
1925	1,149	1,122	1,058	798	1,356	1,255	1,302	1,465	923	1,292	740	877
1926	1,183	1,049	965	1,308	1,842	1,226	1,726	1,083	1,388	1,010	1,344	1,283
1927	1,204	1,165	1,199	1,171	1,390	1,411	1,089	1,382	1,297	1,539	1,280	1,717
1928	1,225	1,243	1,530	1,303	1,493	1,879	1,669	1,883	1,506	2,015	1,691	1,529
1929	1,972	1,678	1,600	1,726	1,932	2,615	3,117	2,236	1,988	2,869	2,579	2,089
1930	1,808	1,731	1,944	1,900	2,202	2,508	1,973	2,407	1,961	2,167	1,765	1,339
1931	1,919	1,710	1,586	1,826	2,268	2,544	2,621	2,856	2,296	2,389	2,449	1,071
1932	1,592	1,897	2,090	2,867	2,942	2,791	2,249	2,839	2,113	2,541	1,318	2,154
1933	1,913	1,886	2,137	2,939	2,679	4,355	4,523	3,141	3,182	3,888	3,305	2,636
1934	2,288	2,511	2,582	3,942	3,724	3,794	4,128	3,696	4,068	3,277	4,680	2,437
1935	2,369	2,804	3,281	3,776	4,613	5,589	5,832	4,946	4,971	4,810	4,289	4,098
1936	3,067	3,474	3,155	3,743	4,390	4,792	4,458	5,561	5,025	4,708	4,145	3,666
1937	3,596	3,777	3,196	4,899	6,796	6,181	6,363	7,423	6,602	6,692	6,645	5,116
1938	5,953	5,328	6,121	7,553	7,798	7,424	7,250	7,003	5,577	6,780	5,602	4,884
1939	4,477	4,810	4,966	6,222	8,643	5,831	7,304	5,969	6,925	6,947	5,323	4,656

Structural table understanding

Three tasks in table structure understanding



Pujara J, Rajendran A, Ghasemi-Gol M, Szekely PA. A Common Framework for Developing Table Understanding Models. ISWC Satellites 2019

Three tasks in table structure understanding

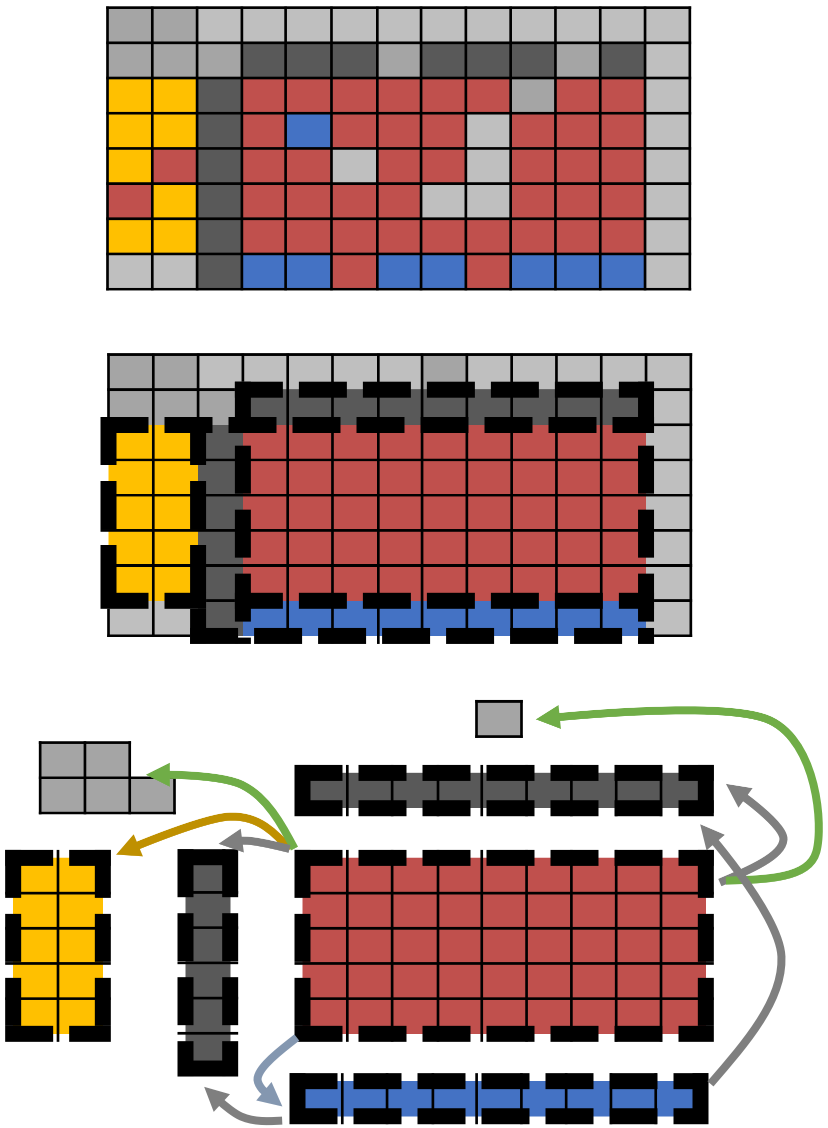
Classify Cell Types

↓

Identify Blocks

↓

Detect Layouts



Structural	Semantic
datatypes (number, year)	semantic type (Person, Time)
role-based (header)	ontology-based (Companies)
relational join (indexing, hierarchy)	semantic model (properties, isA)

Pujara J, Rajendran A, Ghasemi-Gol M, Szekely PA. A Common Framework for Developing Table Understanding Models. ISWC Satellites 2019

Cell Classification

Location

Alabama

Table 1.
Number of OASDI beneficiaries in current-payment status and total monthly benefits, December 2009

Congressional district		Number of beneficiaries						Total monthly benefits (thousands of dollars)			Number of beneficiaries aged 65 or older
		Total	Retired workers	Disabled workers	Widow(er)s a	Spouses b	Children c	All beneficiaries	Retired workers	Widow(er)s a	
	Alabama	983,341	543,725	204,573	91,034	42,103	101,906	995,047	614,516	91,971	603,628
1		145,362	81,226	27,438	13,794	7,107	15,797	149,513	93,418	14,491	90,598
2		140,628	79,232	29,378	12,185	5,274	14,559	136,363	85,772	11,611	86,405
3		143,959	78,531	33,571	11,885	4,740	15,232	140,850	85,325	11,189	84,876
4		160,325	86,682	34,781	15,745	7,592	15,525	159,152	95,311	15,558	97,489
5		137,871	81,153	24,876	12,984	7,013	11,845	142,963	92,810	13,430	90,068
6		128,444	76,814	21,578	12,262	6,367	11,423	146,182	96,707	14,230	85,707
7		126,752	60,087	32,951	12,179	4,010	17,525	120,024	65,173	11,462	68,485
All areas d		52,522,819	33,514,013	7,788,013	4,488,492	2,501,723	4,230,578	55,905,731	39,020,920	4,893,329	36,594,122

Number

SOURCE: Social Security Administration, Master Beneficiary Record, 100 percent data.

a. Includes nondisabled widow(er)s, disabled widow(er)s, widowed mothers and fathers, and parents receiving payment on the record of a worker who is deceased.

b. These beneficiaries receive payment on the record of a worker who is retired or disabled.

c. These beneficiaries receive payment on the record of a worker who is retired, deceased, or disabled.

d. Includes beneficiaries in the 50 states, District of Columbia, American Samoa, Guam, Northern Mariana Islands, Puerto Rico, U.S. Virgin Islands, and foreign countries.

File available from:
U.S. Social Security Administration, Office of Retirement and Disability Policy
Congressional Statistics, December 2009
http://www.socialsecurity.gov/policy/docs/factsheets/cong_stats/2009/

String

Block Detection

Metadata

Header

Data

Attribute

Alabama

Table 1.
Number of OASDI beneficiaries in current-payment status and total monthly benefits, December 2009

Congressional district		Number of beneficiaries						Total monthly benefits (thousands of dollars)			Number of beneficiaries aged 65 or older
		Total	Retired workers	Disabled workers	Widow(er)s a	Spouses b	Children c	All beneficiaries	Retired workers	Widow(er)s a	
	Alabama	983,341	543,725	204,573	91,034	42,103	101,906	995,047	614,516	91,971	603,628
1		145,362	81,226	27,438	13,794	7,107	15,797	149,513	93,418	14,491	90,598
2		140,628	79,232	29,378	12,185	5,274	14,559	136,363	85,772	11,611	86,405
3		143,959	78,531	33,571	11,885	4,740	15,232	140,850	85,325	11,189	84,876
4		160,325	86,682	34,781	15,745	7,592	15,525	159,152	95,311	15,558	97,489
5		137,871	81,153	24,876	12,984	7,013	11,845	142,963	92,810	13,430	90,068
6		128,444	76,814	21,578	12,262	6,367	11,423	146,182	96,707	14,230	85,707
7		126,752	60,087	32,951	12,179	4,010	17,525	120,024	65,173	11,462	68,485
All areas d		52,522,819	33,514,013	7,788,013	4,488,492	2,501,723	4,230,578	55,905,731	39,020,920	4,893,329	36,594,122

SOURCE: Social Security Administration, Master Beneficiary Record, 100 percent data.

a. Includes nondisabled widow(er)s, disabled widow(er)s, widowed mothers and fathers, and parents receiving payment on the record of a worker who is deceased.

b. These beneficiaries receive payment on the record of a worker who is retired or disabled.

c. These beneficiaries receive payment on the record of a worker who is retired, deceased, or disabled.

d. Includes beneficiaries in the 50 states, District of Columbia, American Samoa, Guam, Northern Mariana Islands, Puerto Rico, U.S. Virgin Islands, and foreign countries.

File available from:
U.S. Social Security Administration, Office of Retirement and Disability Policy
Congressional Statistics, December 2009
http://www.socialsecurity.gov/policy/docs/factsheets/cong_stats/2009/

Layout Prediction

Alabama											
Table 1. Number of OASDI beneficiaries in current-payment status and total monthly benefits, December 2009											
Congressional district		Number of beneficiaries						Total monthly benefits (thousands of dollars)			Number of beneficiaries aged 65 or older
		Total	Retired workers	Disabled workers	Widow(er)s a	Spouses b	Children c	All beneficiaries	Retired workers	Widow(er)s a	
	Alabama	983,341	543,725	204,573	91,034	42,103	101,906	995,047	614,516	91,971	603,628
1		145,362	81,226	27,438	13,794	7,107	15,797	149,513	93,418	14,491	90,598
2		140,628	79,232	29,378	12,185	5,274	14,559	136,363	85,772	11,611	86,405
3		143,959	78,531	33,571	11,885	4,740	15,232	140,850	85,325	11,189	84,876
4		160,325	86,682	34,781	15,745	7,592	15,525	159,152	95,311	15,558	97,489
5		137,871	81,153	24,876	12,984	7,013	11,845	142,963	92,810	13,430	90,068
6		128,444	76,814	21,578	12,262	6,367	11,423	146,182	96,707	14,230	85,707
7		126,752	60,087	32,951	12,179	4,010	17,525	120,024	65,173	11,462	68,485
All areas d		52,522,819	33,514,013	7,788,013	4,488,492	2,501,723	4,230,578	55,905,731	39,020,920	4,893,329	36,594,122

SOURCE: Social Security Administration, Master Beneficiary Record, 100 percent data.

a. Includes nondisabled widow(er)s, disabled widow(er)s, widowed mothers and fathers, and parents receiving payment on the record of a worker who is deceased.

b. These beneficiaries receive payment on the record of a worker who is retired or disabled.

c. These beneficiaries receive payment on the record of a worker who is retired, deceased, or disabled.

d. Includes beneficiaries in the 50 states, District of Columbia, American Samoa, Guam, Northern Mariana Islands, Puerto Rico, U.S. Virgin Islands, and foreign countries.

File available from:
U.S. Social Security Administration, Office of Retirement and Disability Policy
Congressional Statistics, December 2009
http://www.socialsecurity.gov/policy/docs/factsheets/cong_stats/2009/

Header of

Global information of

TL;DR of prior approaches to table structures

- conditional random fields
  - Sequence-based features of white space

    Pinto D, McCallum A, Wei X, Croft WB. Table extraction using conditional random fields.

- graphical models
  - Use style, type information, adjacency, orientation as potential functions

    Chen, Z., Cafarella, M.: Integrating spreadsheet data via accurate and low-effort extraction. In: KDD 2014.

- supervised machine learning
  - Collect content, stylistic, font, and spatial features, use SVM/RF to predict

    Koci, E., Thiele, M., Romero Moral, ´O., Lehner, W.: A machine learning approach for layout inference in spreadsheets. In: IC3K 2016.

- heuristic rules
  - Find domain specific patterns that identify structural elements

    Eberius J, Werner C, Thiele M, Braunschweig K, Dannecker L, Lehner W. Deexcelerator: A framework for extracting relational data from partially structured documents. CIKM 2013.

    Shigarov A, Khristyuk V, Mikhailov A, Paramonov V. TabbyXL: Rule-based spreadsheet data extraction and transformation. ICIST 2019.

Common feature sets in methods

Format
- indentation
- bold, italic, strike
- underline style
- sub/super
- font face, size
- font color
- fill color
- alignment
- merging
- borders

Spatial
- row position
- col position
- neighbors
- neighbor styles match
- neighbor types match
- neighbor types

Content
- data type
- semantic type
- link
- numeric format
- numeric range
- length
- tokens
- capitalization
- special characters
- punctuation
- keywords
- formulas

Koci et al., IC3K 2016

Cell embeddings for structural table understanding

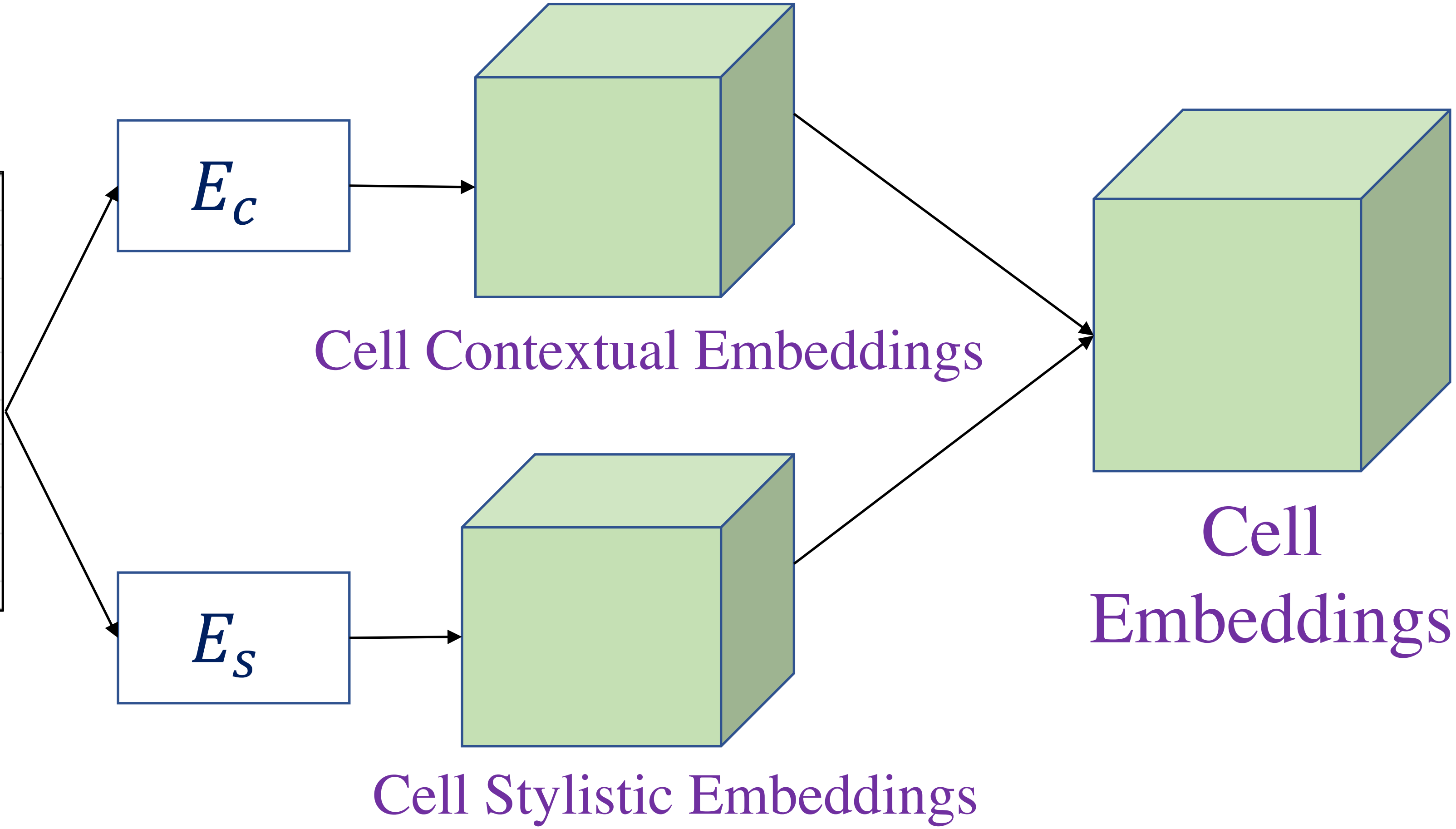
Cell Embeddings for Table Structure

Table1 - 2018-19 academic year course evaluations

Computer Science Department

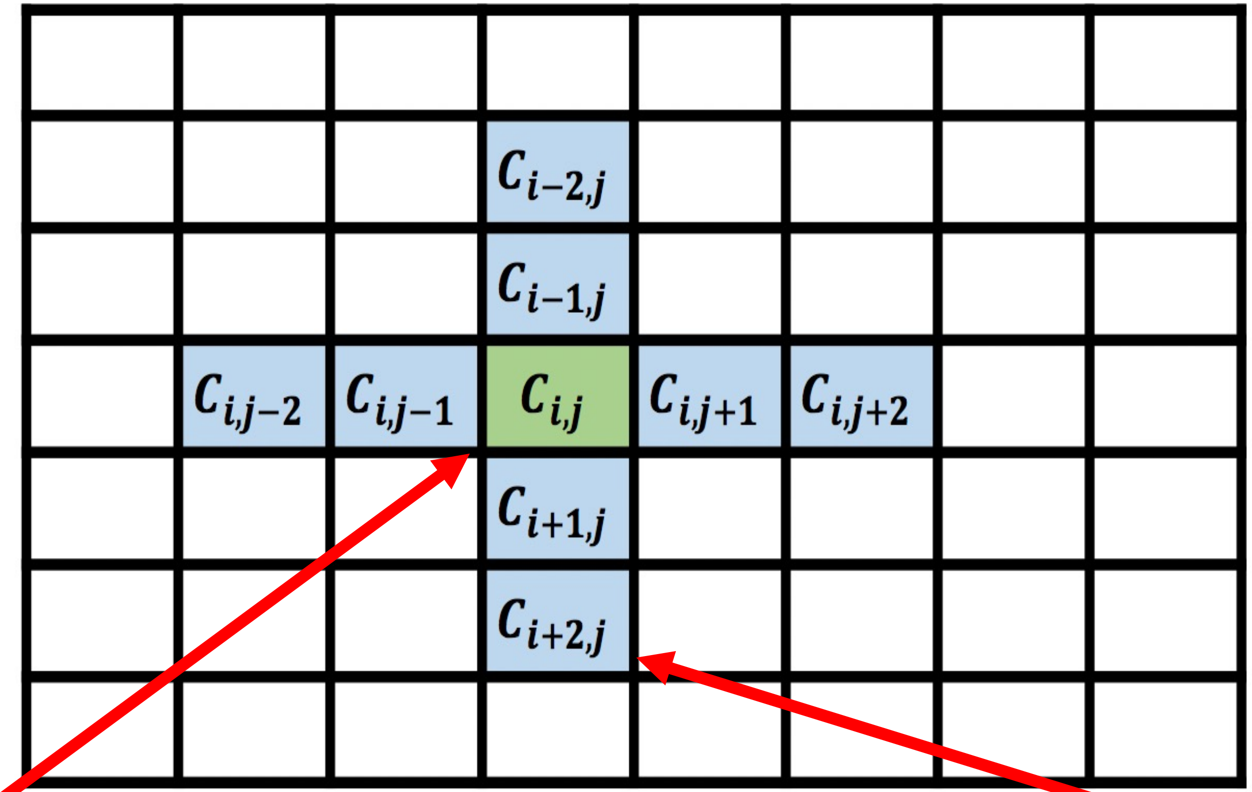
	Course Information						Course Evaluation*		Semester Average Evaluation Score
	Course Code	Course Name	Instructor Name	Number of Registered Students	Start Date	Exam Date	Number of Students Responded to course Evaluation	Evaluation Score	
Fall 2018	CSCI561	Foundations of Artificial Intelligence	Prof. Tejada	540	8/26/18	12/2/18	45	4.8	4.9
	CSCI567	Machine Learning	Prof. Sha	334	8/27/18	12/5/18	32	4.9	
	CSCI585	Database Systems	Prof. Shahabi	460	8/30/18	12/4/18	44	4.9	
Spring 2019	CSCI561	Foundations of Artificial Intelligence	Prof. Tejada	540	1/9/19	5/2/19	75	4.9	5.0
	CSCI670	Advanced Analysis of Algorithms	Prof. Kempe	70	1/11/19	5/7/19	12	5	

Evaluation surveys were completed before the final exam date. Survey links had been emailed to all students.



Gol MG, Pujara J, Szekely P. Tabular cell classification using pre-trained cell embeddings. ICDM 2019.

Unsupervised Learning of context: E_c



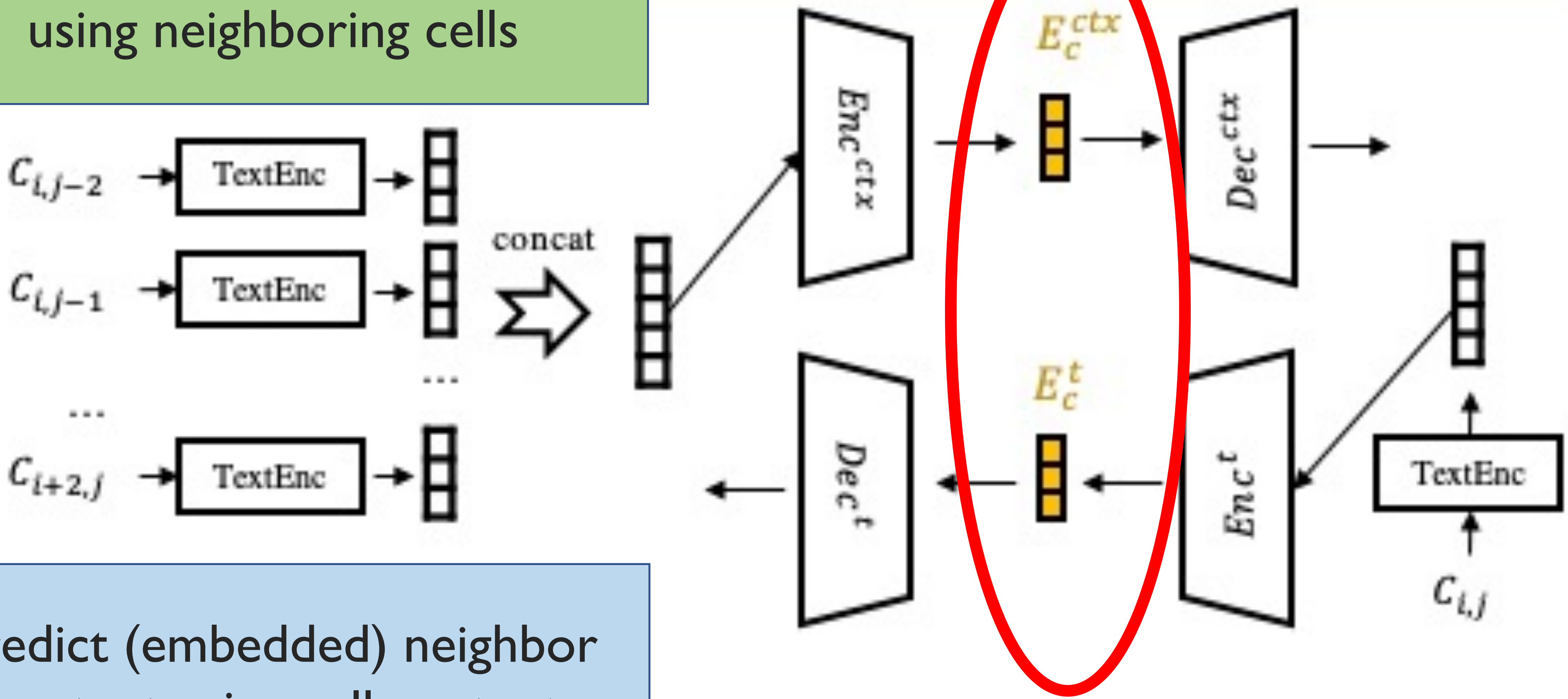
Predict (embedded) cell content using neighboring cells

Predict (embedded) neighbor content using cell content

Gol et. al, ICDM 2019

Unsupervised encoder-decoder architecture

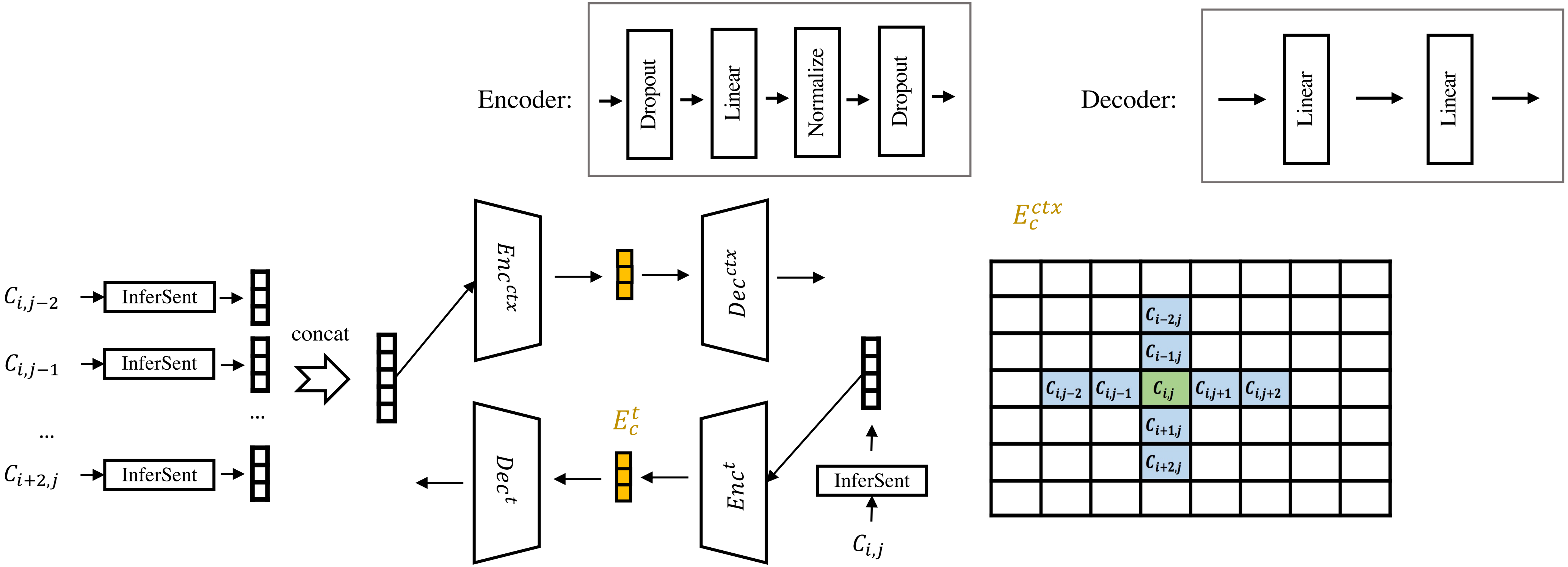
Predict (embedded) cell content using neighboring cells



Predict (embedded) neighbor content using cell content

Contextual cell embedding

Unsupervised Learning of context: E_c



$$l(\phi) = \sum_i \left| I(C_i) - Dec_{\phi_1}\left(Enc_{\phi_2}(X_{C_i})\right)\right|^2 +$$

$$\sum_i \sum_{C_j \in X_{C_i}} \left| I(C_i) - Dec_{\phi_3}\left(Enc_{\phi_4}(C_i)\right)\right|^2$$

$$E_c = \underset{\Phi}{\operatorname{argmin}}\, l(\varphi)$$

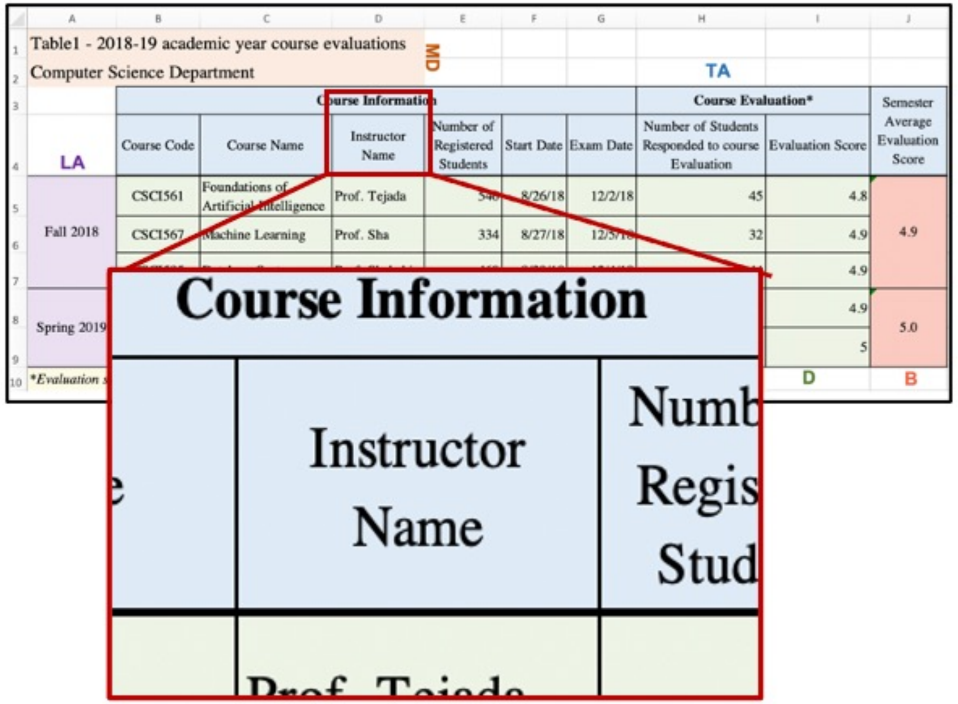
Loss Function

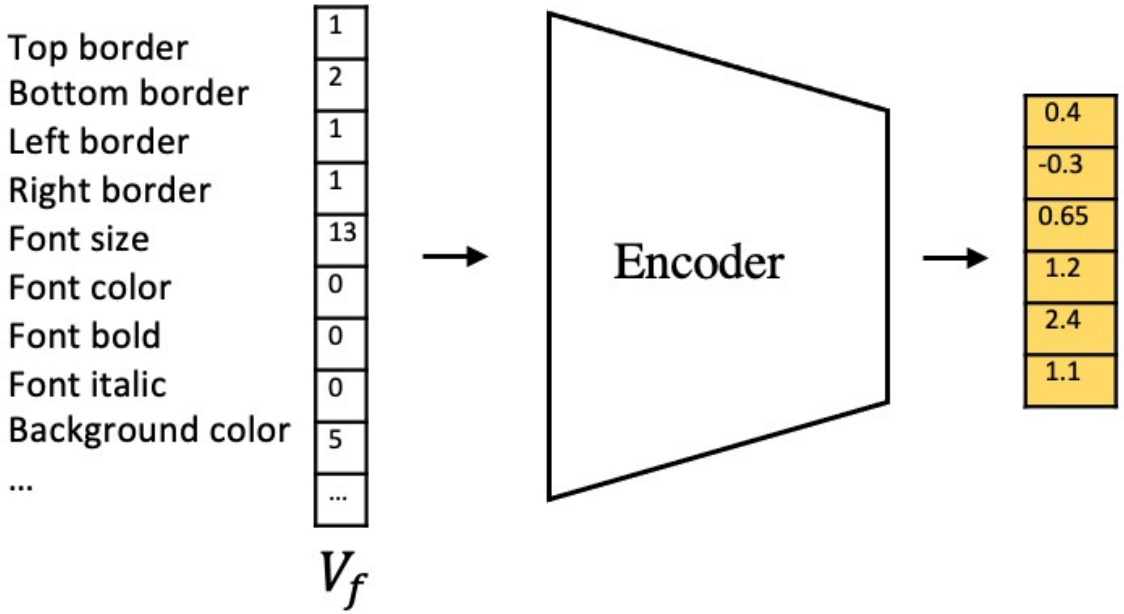
sum of mean squared error for
all the cells in all training documents

Model Parameters
Optimization

Gol et al., ICDM 2019

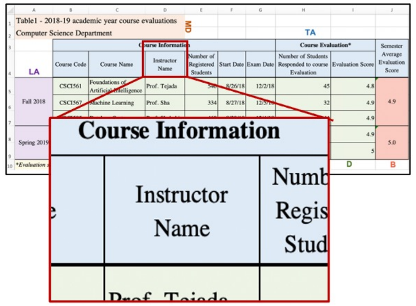
Unsupervised Learning from styles: E_s

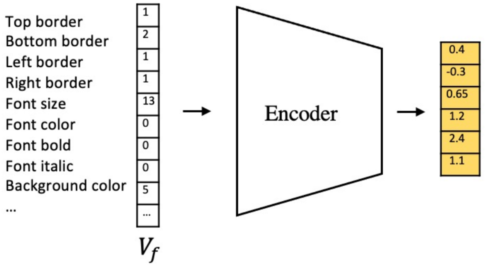




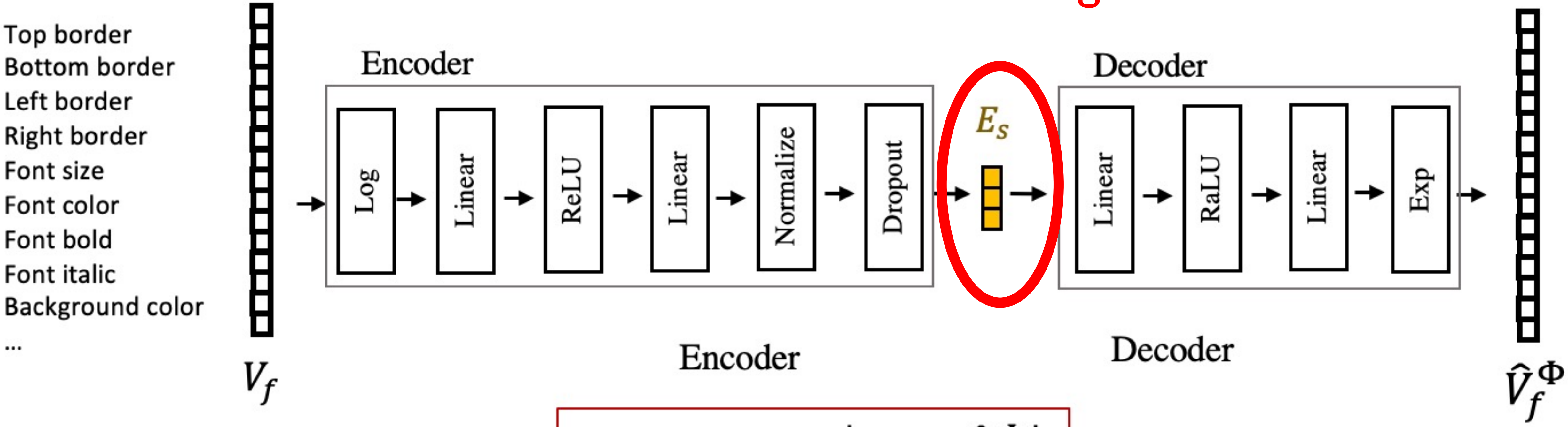
Gol MG, Pujara J, Szekely P. Tabular cell classification using pre-trained cell embeddings. ICDM 2019.

Autoencoder for learning from styles: E_s





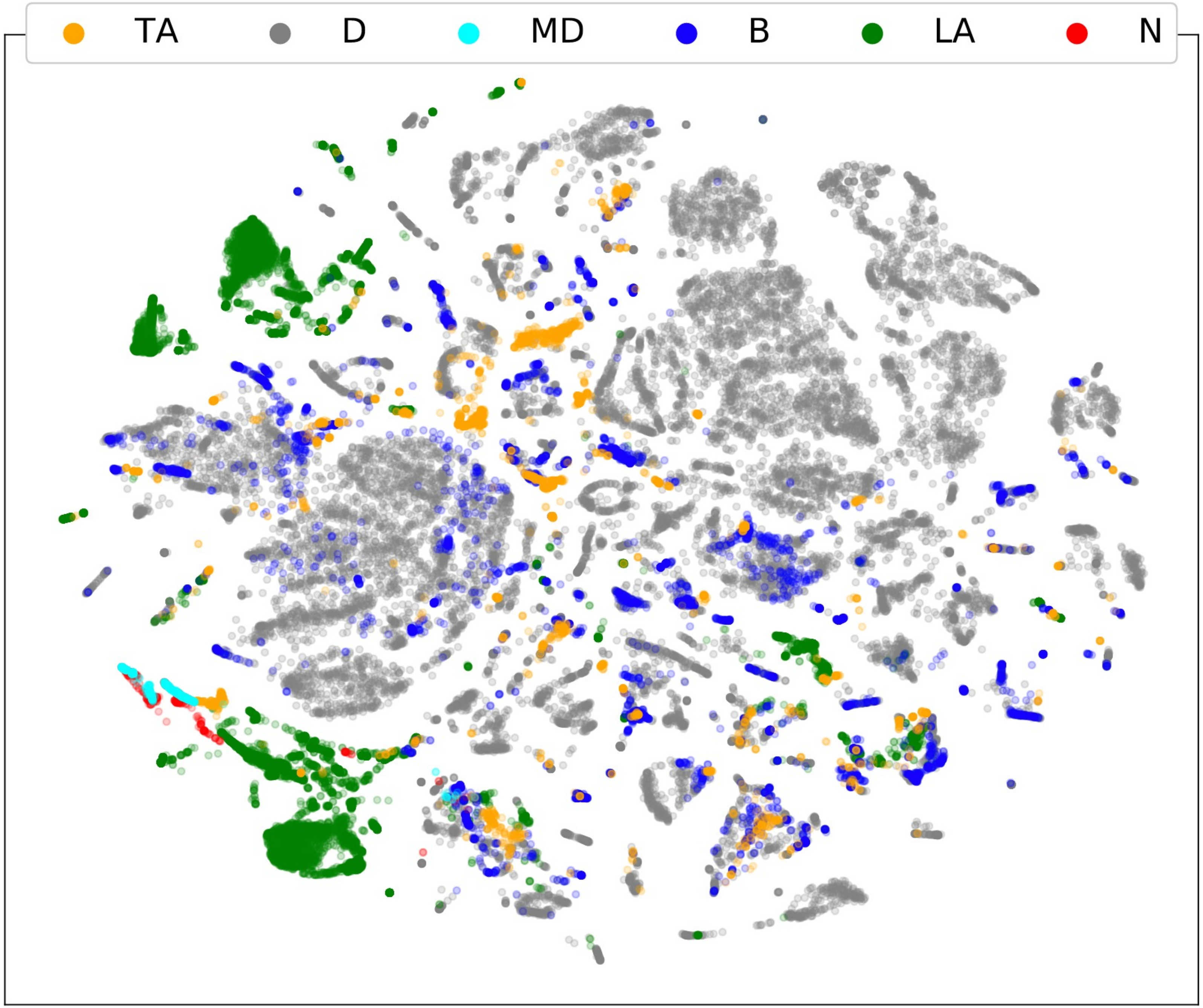
Stylistic cell embedding



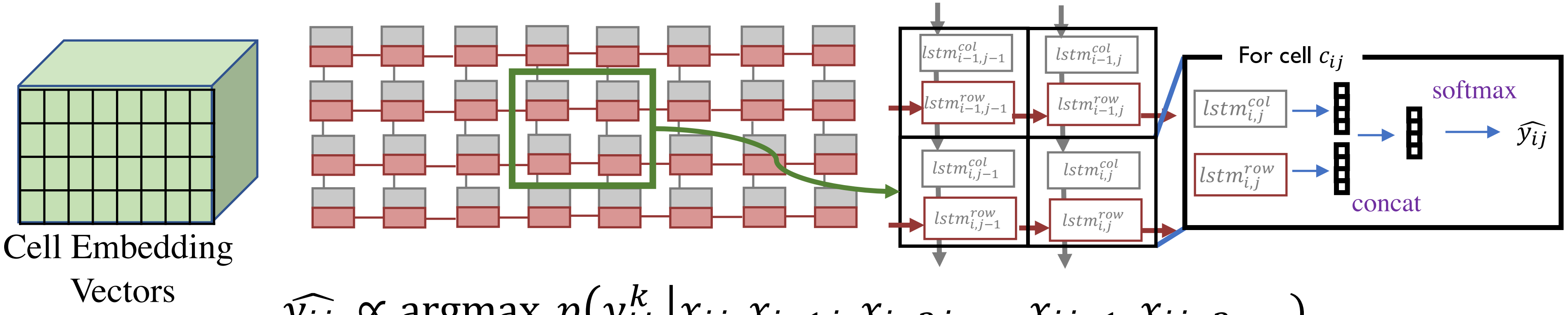
$$E_s = \underset{\Phi}{\operatorname{argmin}} \left| V_f - \hat{V}_f^{\Phi} \right|$$

Gol MG, Pujara J, Szekely P. Tabular cell classification using pre-trained cell embeddings. ICDM 2019.

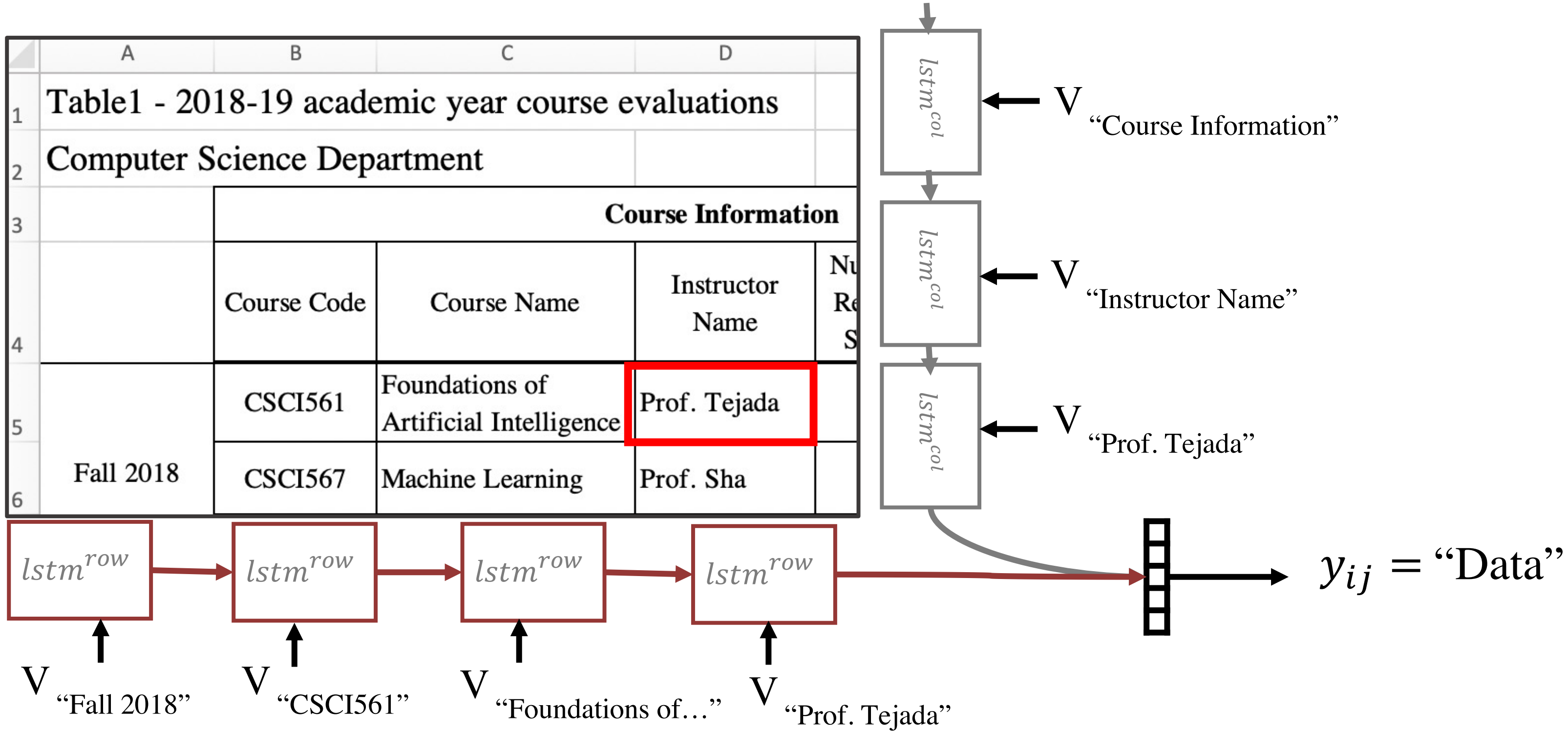
Spatial view of cell embeddings



RNN-Based Cell Classification Approach

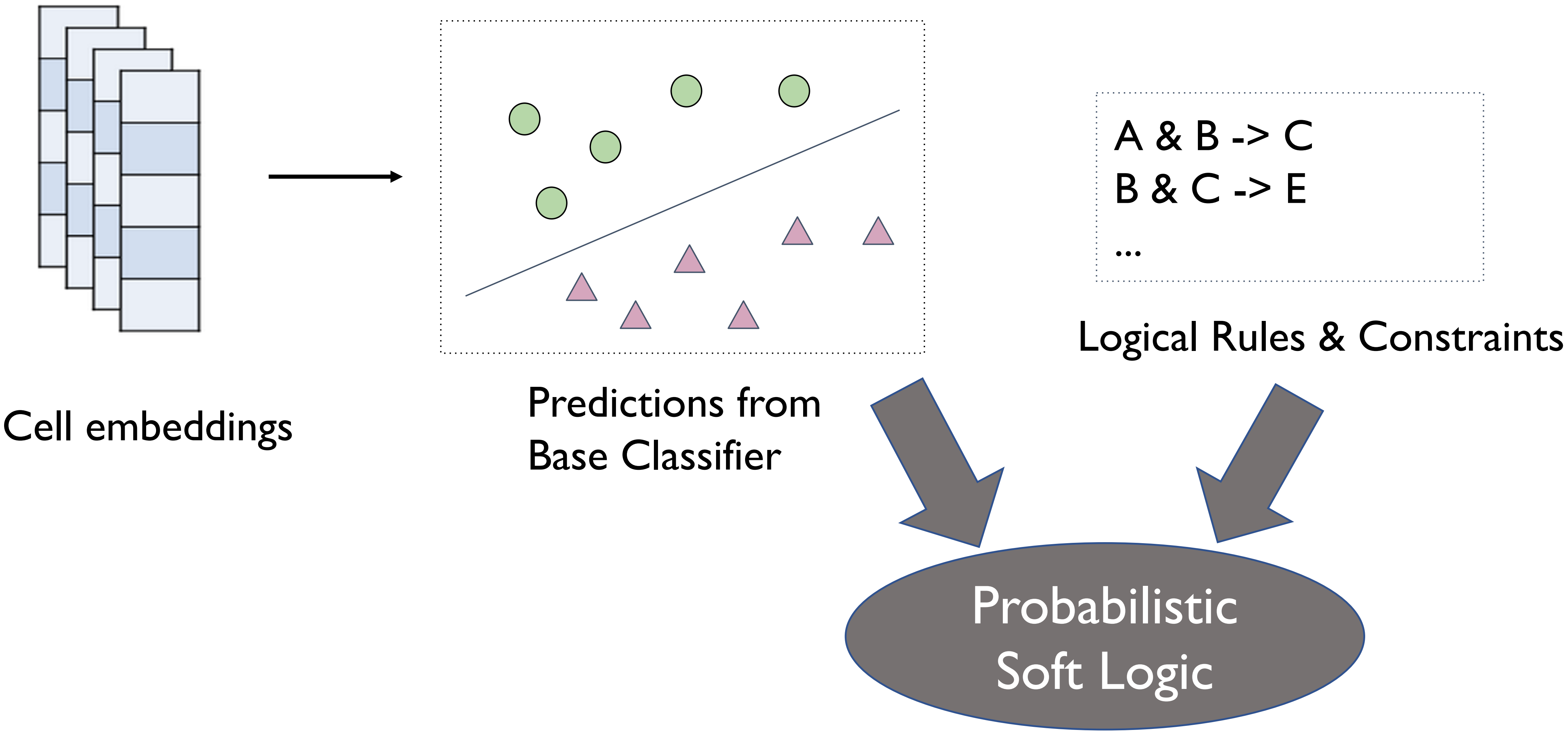


$$\widehat{y_{ij}} \propto \underset{k}{\operatorname{argmax}}\ p\left(y_{ij}^{k} \,\middle|\, x_{ij}, x_{i-1j}, x_{i-2j}, \dots, x_{ij-1}, x_{ij-2}, \dots\right)$$



Beyond embeddings: hybrid table understanding

Hybrid models for table understanding



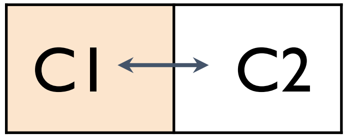
Sun K, Rayudu H, Pujara J. A Hybrid Probabilistic Approach for Table Understanding. AAAI 2021

Background: Probabilistic Soft Logic

Weight ← $w\colon BlockType(A,T) \wedge Neighbor(A,B) \Longrightarrow BlockType(B,T)$

Predicates

Variables



$$w\colon BlockType(C1,Data) \wedge Neighbor(C1,C2) \Longrightarrow BlockType(C2,Data)$$

Bach SH, Broecheler M, Huang B, Getoor L. Hinge-Loss Markov Random Fields and Probabilistic Soft Logic. JMLR 2017

Modeling probabilistic relationships (via PSL)

Collective Classification

```
Label(C1, L) & Adj(C1, C2)
        -> Label(C2, L)
```

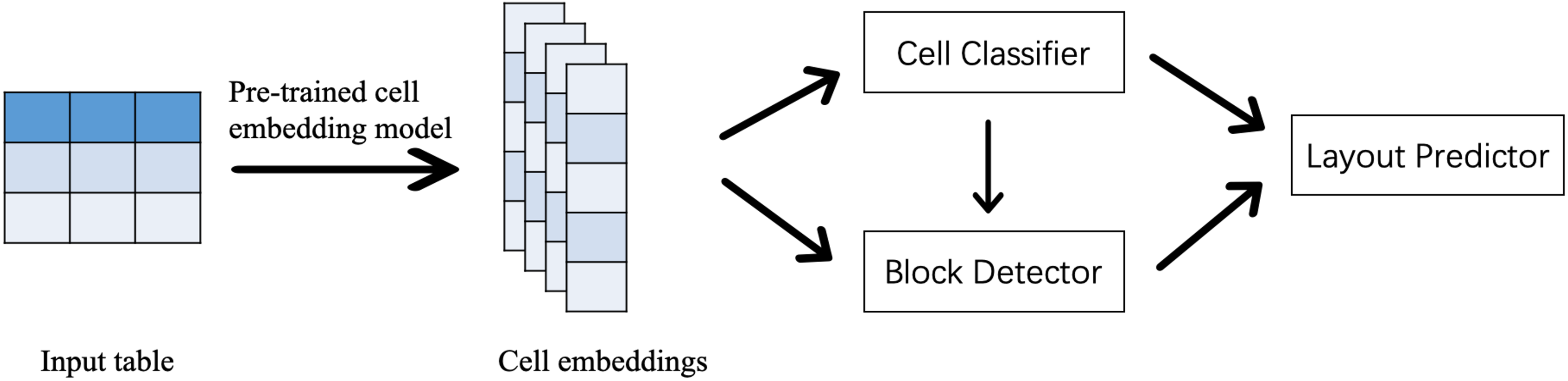
Entity Resolution

```
Label(C1, L) & Label(C2, L) & Adj(C1, C2)
        -> SameBlock(C1, C2)
```

Link Prediction

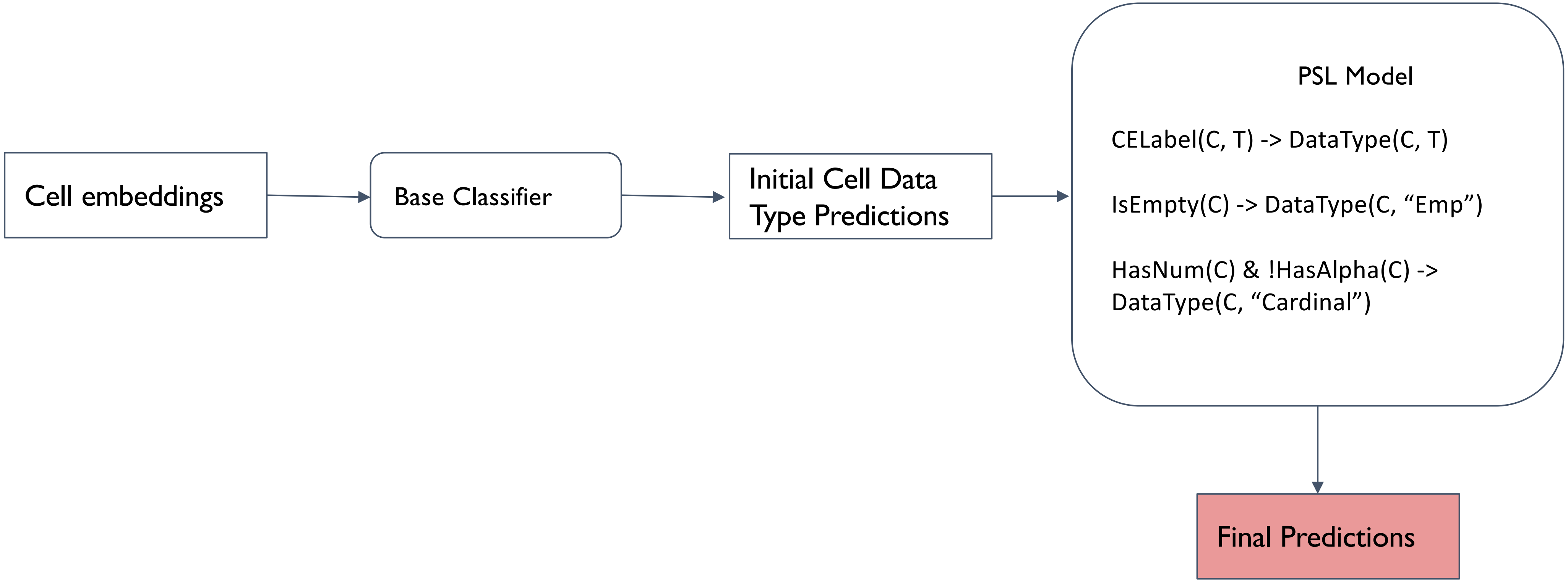
```
Label(B1, ‘Attr’) & Label(B2, ‘Val’)
        -> Rel(B1, B2, ‘Idx’)
```

Hybrid System Architecture



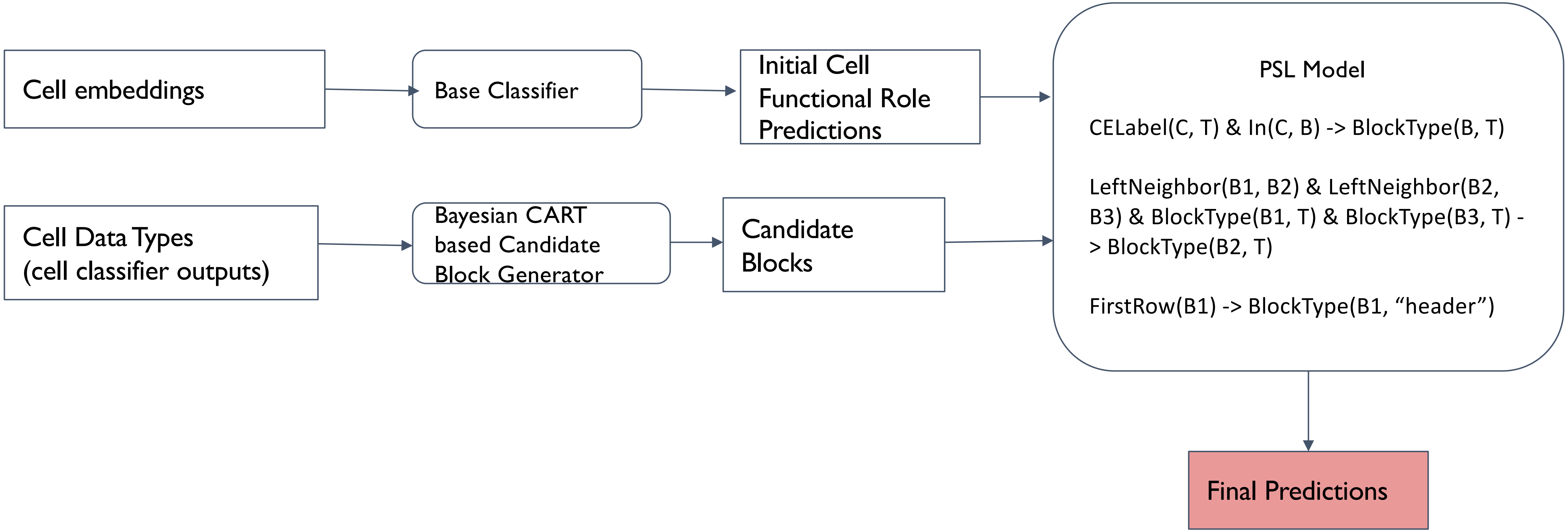
Sun et al., AAAI 2021

Cell Classifier



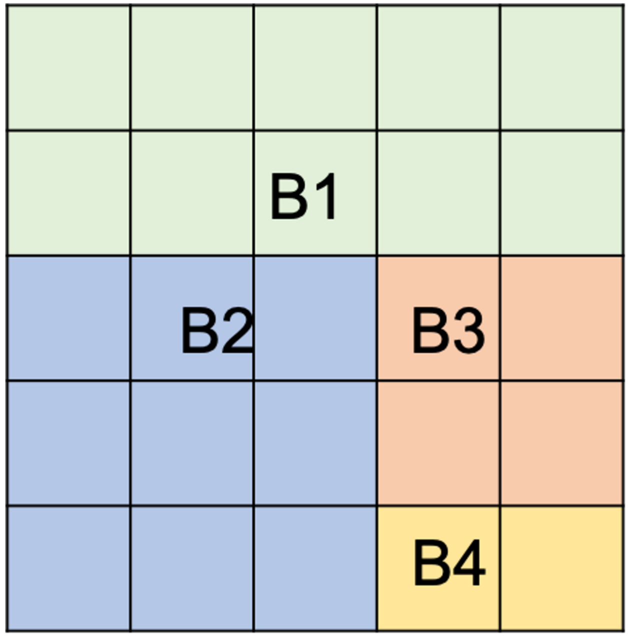
Sun et al., AAAI 2021

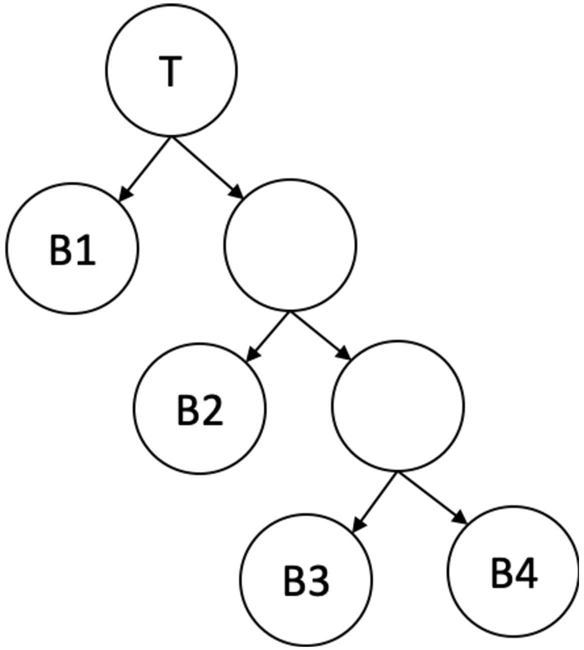
Block Detector



Sun et al., AAAI 2021

Block Detector - Bayesian CART-based Block Generator





- Data Type distribution
- Depth of the node

Algorithm 1: Candidate Block Generation

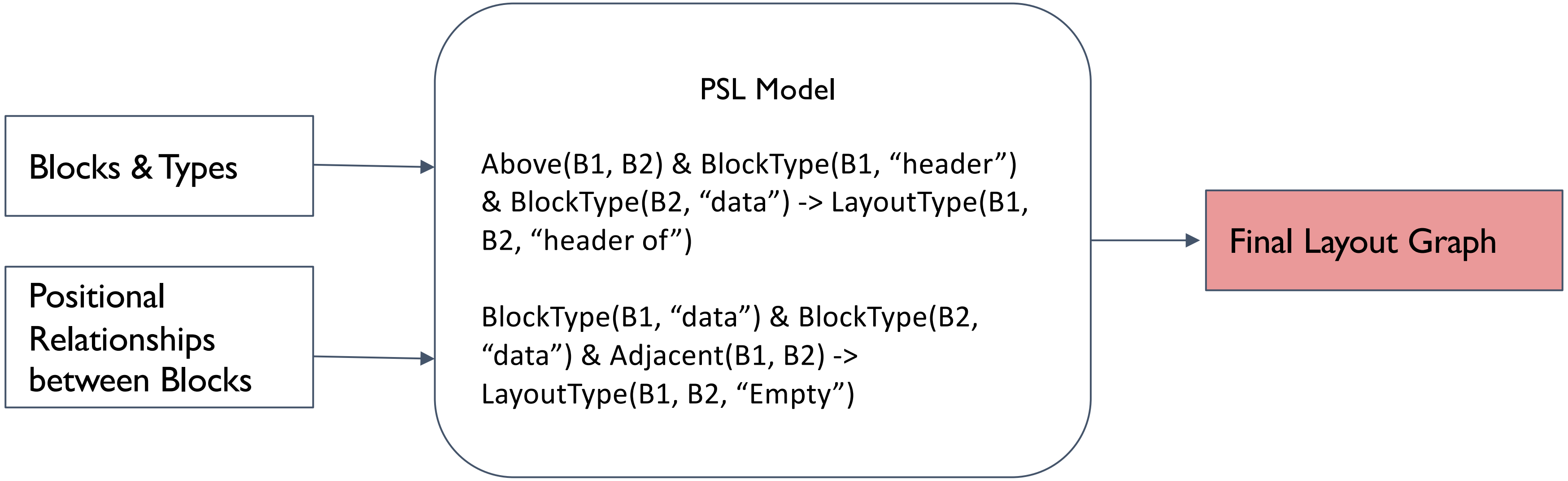
Function `Split` ($block$)**:**
  $queue \longleftarrow \{(block, 0)\}$; $blocks \longleftarrow \{\}$;
  while $queue \neq \emptyset$ **do**
    $(\langle t, b, l, r \rangle, d) \longleftarrow queue.get()$
    `// t, b, l, and r are indices.`
    Randomly select a number v within $[0, 1]$
    if $v < p_{split}(d)$ **then**
      Randomly split $\langle t, b, l, r \rangle$ into B_1 and B_2 using p_{rule}.
      $queue.push((B_1, d+1))$
      $queue.push((B_2, d+1))$
    else
      $blocks.add(\langle t, b, l, r \rangle)$
  return $blocks$;

Function `GenerateATree` (T, $types$)**:**
  $row\_blocks =$ `Split`(T); `// Row-wise`
  $blocks \longleftarrow \emptyset$;
  foreach B *in* row_blocks **do**
    $blocks.union($`Split`$(B))$ `// Column-wise`
  return $blocks$;

Function `SampleATree` (T, $types$, N)**:**
  $trees \longleftarrow \emptyset$
  foreach $1 \leq i \leq N$ **do**
    $trees.add($`GenerateATree`$(T, types))$
  return Sample a tree from $trees$ using W_{ent}.

Sun et al., AAAI 2021

Layout Predictor



Sun et al., AAAI 2021

Evaluation snapshot

	Emp	Crd	Str	Dat	Loc	Org	Ord	Nom	Per	Avg
CRF	81.9	82.5	42.4	56.2	34.4	16.8	0.0	36.0	1.3	39.1±1.8
MLP	84.5	85.6	69.1	59.3	54.9	46.8	0.0	52.0	1.2	50.4±5.4
RF	85.0	84.4	73.2	61.4	65.2	**55.5**	**0.3**	**53.4**	**39.3**	57.5±4.7
PSL (MLP)	96.5	**88.3**	70.2	77.8	55.8	43.3	**0.3**	52.4	1.0	54.0±3.1
PSL (RF)	**96.8**	87.8	**74.3**	**78.4**	**66.1**	52.5	0.2	53.0	31.7	**60.1±3.2**

Cell classification

Method	EP	HO	AO	GA	SC	Avg
RF	81.7	1.1	2.1	22.7	0.0	21.5±1.2
CRF	88.5	33.7	32.2	40.0	0.0	38.9±3.1
PSL	**89.6**	**70.3**	**32.8**	**43.0**	**25.6**	**52.3±3.4**

Layout Prediction

Sun et al., AAAI 2021

		MD	DT	HD	AT	Avg
CIUS	CRF	96.5	67.6	94.9	36.8	73.9±8.9
	RNN	**99.5**	99.3	97.4	90.5	96.7±4.1
	RF	95.9	**99.7**	88.9	97.0	95.4±0.6
	PSL(RNN)	94.8	99.2	**97.8**	89.3	95.3±4.1
	PSL(RF)	93.6	**99.7**	96.0	**97.6**	**96.7±1.1**
SAUS	CRF	80.7	82.2	**95.7**	38.2	74.2±5.8
	RNN	**94.3**	97.5	84.1	79.5	88.9±2.3
	RF	79.1	98.6	78.8	91.1	86.9±4.0
	PSL(RNN)	87.6	97.8	86.7	79.5	87.9±1.4
	PSL(RF)	80.6	**99.0**	85.4	**92.8**	**89.4±2.5**
DeEx	CRF	35.6	55.7	48.0	1.7	35.3±6.9
	RNN	33.8	96.1	47.2	39.5	54.2±5.9
	RF	53.4	98.4	51.0	26.5	57.3±2.0
	PSL(RNN)	38.5	97.2	53.5	**44.9**	58.5±8.0
	PSL(RF)	**65.4**	**98.8**	**60.5**	26.0	**62.7±3.9**
DG	CRF	41.3	53.1	**94.1**	34.8	55.9±9.3
	RNN	45.4	**95.9**	82.9	**78.8**	75.8±4.3
	RF	74.0	95.8	80.7	77.8	82.1±2.5
	PSL(RNN)	69.9	95.7	89.2	77.4	83.1±5.2
	PSL(RF)	**77.2**	95.7	91.4	77.4	**85.4±4.8**

Block Detection

Improving over cell-level structural models

Alabama											
Table 1. Number of OASDI beneficiaries in current-payment status and total monthly benefits, December 2009											
Congressional district		Number of beneficiaries						Total monthly benefits (thousands of dollars)			Number of beneficiaries
		Total	Retired workers	Disabled workers	Widow(er)s a	Spouses b	Children c	All beneficiaries	Retired workers	Widow(er)s a	aged 65 or older
	Alabama	983,341	543,725	204,573	91,034	42,103	101,906	995,047	614,516	91,971	603,628
1		145,362	81,226	27,438	13,794	7,107	15,797	149,513	93,418	14,491	90,598
2		140,628	79,232	29,378	12,185	5,274	14,559	136,363	85,772	11,611	86,405
3		143,959	78,531	33,571	11,885	4,740	15,232	140,850	85,325	11,189	84,876
4		160,325	86,682	34,781	15,745	7,592	15,525	159,152	95,311	15,558	97,489
5		137,871	81,153	24,876	12,984	7,013	11,845	142,963	92,810	13,430	90,068
6		128,444	76,814	21,578	12,262	6,367	11,423	146,182	96,707	14,230	85,707
7		126,752	60,087	32,951	12,179	4,010	17,525	120,024	65,173	11,462	68,485
All areas d		52,522,819	33,514,013	7,788,013	4,488,492	2,501,723	4,230,578	55,905,731	39,020,920	4,893,329	36,594,122

SOURCE: Social Security Administration, Master Beneficiary Record, 100 percent data.

a. Includes nondisabled widow(er)s, disabled widow(er)s, widowed mothers and fathers, and parents receiving payment on the record of a worker who is deceased.

b. These beneficiaries receive payment on the record of a worker who is retired or disabled.

c. These beneficiaries receive payment on the record of a worker who is retired, deceased, or disabled.

d. Includes beneficiaries in the 50 states, District of Columbia, American Samoa, Guam, Northern Mariana Islands, Puerto Rico, U.S. Virgin Islands, and foreign countries.

File available from:
U.S. Social Security Administration, Office of Retirement and Disability Policy
Congressional Statistics, December 2009
http://www.socialsecurity.gov/policy/docs/factsheets/cong_stats/2009/

Cell based functional role detection

Sun et al., AAAI 2021

Improving over cell-level structural models

Alabama

Table 1.
Number of OASDI beneficiaries in current-payment status and total monthly benefits, December 2009

Congressional district		Number of beneficiaries						Total monthly benefits (thousands of dollars)			Number of beneficiaries aged 65 or older
		Total	Retired workers	Disabled workers	Widow(er)s a	Spouses b	Children c	All beneficiaries	Retired workers	Widow(er)s a	
	Alabama	983,341	543,725	204,573	91,034	42,103	101,906	995,047	614,516	91,971	603,628
1		145,362	81,226	27,438	13,794	7,107	15,797	149,513	93,418	14,491	90,598
2		140,628	79,232	29,378	12,185	5,274	14,559	136,363	85,772	11,611	86,405
3		143,959	78,531	33,571	11,885	4,740	15,232	140,850	85,325	11,189	84,876
4		160,325	86,682	34,781	15,745	7,592	15,525	159,152	95,311	15,558	97,489
5		137,871	81,153	24,876	12,984	7,013	11,845	142,963	92,810	13,430	90,068
6		128,444	76,814	21,578	12,262	6,367	11,423	146,182	96,707	14,230	85,707
7		126,752	60,087	32,951	12,179	4,010	17,525	120,024	65,173	11,462	68,485
All areas d		52,522,819	33,514,013	7,788,013	4,488,492	2,501,723	4,230,578	55,905,731	39,020,920	4,893,329	36,594,122

SOURCE: Social Security Administration, Master Beneficiary Record, 100 percent data.

a. Includes nondisabled widow(er)s, disabled widow(er)s, widowed mothers and fathers, and parents receiving payment on the record of a worker who is deceased.

b. These beneficiaries receive payment on the record of a worker who is retired or disabled.

c. These beneficiaries receive payment on the record of a worker who is retired, deceased, or disabled.

d. Includes beneficiaries in the 50 states, District of Columbia, American Samoa, Guam, Northern Mariana Islands, Puerto Rico, U.S. Virgin Islands, and foreign countries.

File available from:
U.S. Social Security Administration, Office of Retirement and Disability Policy
Congressional Statistics, December 2009
http://www.socialsecurity.gov/policy/docs/factsheets/cong_stats/2009/

Collective block-level structural models

Sun et al., AAAI 2021

Synopsis

- Table understanding is an important problem with many constituent tasks ranging from structural understanding, knowledge alignment, to downstream applications
- Understanding table structure is a foundational task enabling knowledge-centric table understanding
- In the past decade, structural models have evolved from using many cell-level features to incorporating more contextual information and using deep learning
- Hybrid, neuro-symbolic models now have state-of-the-art performance by incorporating human expectations of table design into predictions from deep learning, correcting potential errors