

# Towards Deep Table Understanding

Yunyao Li

Distinguished Research Staff Member  
Senior Manager

Scalable Knowledge Intelligence Department

IBM Research – Almaden

Twitter: @yunyao\_li  
Email: [yunyaoli@us.ibm.com](mailto:yunyaoli@us.ibm.com)

DIS @KDD'2021



# What is document conversion?

Source: **Business documents**  
(PDF/Word/Image)

Designed for document exchange, printing, human consumption and archival

**FOOD**

**Food, in particular grocery retailing, looks set to continue on its spectacular growth trajectory.**

**FOOD RETAIL INDUSTRY**

**FOOD CONSUMPTION - SIZE OF MARKET IN US DOLLARS (BILLIONS)**

Year	Value	CAGR
2019	1,146	3.2%
2020	1,141	3.2%
2021	1,147	3.1%
2022	1,153	3.1%



Target: **Content in an open text-based format**  
(JSON, HTML, etc.)

Amenable for NLP/AI driven understanding of its contents

**FOOD**

**Food, in particular grocery retailing, looks set to continue on its spectacular growth trajectory.**

**FOOD RETAIL INDUSTRY**

**FOOD CONSUMPTION - SIZE OF MARKET IN US DOLLARS (BILLIONS)**

Year	Value	CAGR
2019	1,146	3.2%
2020	1,141	3.2%
2021	1,147	3.1%
2022	1,153	3.1%

# What is Document Conversion ?

An informal view

## Preserving the reading order

Identify the correct sequencing of text, tabular and image content mimicking human reading order

**FOOD**

*Food, in particular grocery retailing, looks set to continue on its spectacular growth trajectory.*

...Over when income levels are normally stunted, a good part of this spending will go towards food consumption. This is especially applicable to the emerging market, where per capita will account for about 26% of household spending in 2015-2016.

...Retailers who are looking to take the market by storm should be aware that the market size of the US is around 100 million with the quarter of a billion people will account for 26% of the total. When you have a market size of 100 million, you are looking at a market size of 26 million to be being consumed. This is a large market, which is being followed by emerging markets and increasingly urban preferences. Food, in particular grocery retailing, looks set to continue on its spectacular growth trajectory in Asia. In 2015, the global retail spending in groceries grew to US\$180 billion, an increase of over 20% from 2014.

**RETAIL INDUSTRY**

Emerging markets are expected to be a major driver for the retail industry. Within Asia, the retail industry is largely fragmented, with a few players controlling the market. In Hong Kong and China, there are a handful of large players and well-established global players such as Wal-Mart and Carrefour. The Chinese food supply chain is also undergoing a major restructuring in comparison to developed economies.

Food retailers stand to gain a bigger piece of the pie by attracting their customer base through opening more outlets and offering their products online. The subscription is particularly important for the Chinese market, as regional sales are projected to all other ways across the country.

With its on their side, food retailers will be able to harness the benefits of sourcing power and operating scale that traditional retailers are unable to do.

Competition in this sector is somewhat less, as this continues to grow as retailers look along the value chain.

Higher costs in Asia, including labor and expenses, could drive profitability, especially for smaller retailers, who are less targeted by power 100, as consumer demand grows, to focus on the food market.

**FOOD CONSUMPTION - SIZE OF MARKET IN US DOLLARS-BILLION**

Region	2014	2015	2016
Asia 1	1,146	1,271	1,401
Asia 2	571	626	681
China	676	741	806
Indonesia	197	218	239
Philippines	86	99	112
Thailand	59	66	73
Vietnam	29	33	37
Singapore	6	7	8

Source: UN Comtrade, UN Statistics Division

## Identifying and extracting document structure

For example, extracting document elements such as sections and paragraphs, tables, and relationships among them.

## Extracting internal structure of complex document elements

For example, for tables, extracting table cells, column and row headers, individual cell values, table title, etc.



# Deep Document Understanding

Source: If applicable, describe source origin

## Understanding of complex documents

- Visual elements such as charts, tables, photos, images, sketches, etc.
- PDF manuals and technical guides, financial and scientific documents, scanned invoices and passports, etc.

## Focus on converting challenging documents

- in terms of structure, content and quality





# Table Extraction: Problem Definition

## Expenses

	Expense (\$ in thousands)			CASM (cents)		
	Three months ended Sept. 30			Three months ended Sept. 30		
	2015	2014	Change	2015	2014	Change
Aircraft fuel	206,924	206,817	(27.9%)	3.00	4.41	(32.0%)
Airport operations	142,176	128,381	10.7%	2.06	1.98	4.0%
Flight operations and navigational charges	126,821	112,886	12.3%	1.84	1.74	5.7%
Sales and distribution	98,351	94,764	3.8%	1.43	1.46	(2.1%)
Depreciation and amortization	69,739	56,620	23.2%	1.01	0.87	16.1%
Marketing, general and administration	58,604	50,647	15.7%	0.85	0.78	9.0%
Maintenance	62,347	48,062	29.7%	0.89	0.74	20.3%
Inflight	47,029	43,600	7.9%	0.68	0.67	1.5%
Aircraft leasing	40,572	43,062	(5.8%)	0.59	0.66	(10.6%)
Employee profit share	32,974	19,039	73.2%	0.48	0.29	65.5%
<b>Total operating expenses</b>	<b>885,537</b>	<b>883,898</b>	<b>0.2%</b>	<b>12.83</b>	<b>13.60</b>	<b>(5.7%)</b>
<b>Total, excluding fuel and profit share</b>	<b>645,639</b>	<b>578,042</b>	<b>11.7%</b>	<b>9.35</b>	<b>8.90</b>	<b>5.1%</b>

During the three months ended September 30, 2015, operating expenses remained flat compared to the same period in 2014. On an ASM basis, operating expenses decreased by 5.7 per cent to 12.83 cents from 13.60 cents in the same period in 2014. This decrease was driven largely by a decrease in aircraft fuel expense partially offset by the increases in our employee profit share expense, maintenance expense and depreciation and amortization expense.

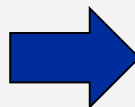
Our CASM, excluding fuel and profit share, increased by 5.1 per cent, which was outside of our previously disclosed guidance of up to 3.5 to 4.0 per cent, primarily as the result of lower than expected ASMs during the quarter in addition to a weaker Canadian dollar.

	Expense (\$ in thousands)			CASM (cents)		
	Nine months ended Sept. 30			Nine months ended Sept. 30		
	2015	2014	Change	2015	2014	Change
Aircraft fuel	632,317	846,514	(25.3%)	3.10	4.41	(29.7%)
Airport operations	412,510	377,378	9.4%	2.02	1.96	3.1%
Flight operations and navigational charges	372,560	339,616	9.7%	1.83	1.77	3.4%
Sales and distribution	281,304	282,598	1.0%	1.40	1.47	(4.8%)
Depreciation and amortization	189,684	172,044	10.3%	0.93	0.90	3.3%
Marketing, general and administration	175,108	164,030	6.8%	0.86	0.85	1.2%
Maintenance	171,599	151,612	13.2%	0.84	0.79	6.3%
Inflight	150,098	130,410	15.1%	0.74	0.68	8.8%
Aircraft leasing	132,208	136,904	(3.4%)	0.65	0.71	(8.5%)
Employee profit share	92,096	49,388	102.9%	0.46	0.24	93.7%
<b>Total operating expenses</b>	<b>2,613,484</b>	<b>2,646,294</b>	<b>(1.2%)</b>	<b>12.83</b>	<b>13.78</b>	<b>(6.9%)</b>
<b>Total, excluding fuel and profit share</b>	<b>1,889,071</b>	<b>1,754,392</b>	<b>7.7%</b>	<b>9.27</b>	<b>9.13</b>	<b>1.5%</b>

During the nine months ended September 30, 2015, operating expenses decreased by 1.2 per cent to \$2,613.5 million as compared to \$2,646.3 million in the same period in 2014, primarily driven by the year-over-year decrease in aircraft fuel expense, partially offset by the increase in our employee profit share expense.

**Input:** Document contents in native format

- PDF
- Image
- Office Docs
- ...



## Expenses

	Expense (\$ in thousands)			CASM (cents)		
	Three months ended Sept. 30			Three months ended Sept. 30		
	2015	2014	Change	2015	2014	Change
Aircraft fuel	206,924	206,817	(27.9%)	3.00	4.41	(32.0%)
Airport operations	142,176	128,381	10.7%	2.06	1.98	4.0%
Flight operations and navigational charges	126,821	112,886	12.3%	1.84	1.74	5.7%
Sales and distribution	98,351	94,764	3.8%	1.43	1.46	(2.1%)
Depreciation and amortization	69,739	56,620	23.2%	1.01	0.87	16.1%
Marketing, general and administration	58,604	50,647	15.7%	0.85	0.78	9.0%
Maintenance	62,347	48,062	29.7%	0.89	0.74	20.3%
Inflight	47,029	43,600	7.9%	0.68	0.67	1.5%
Aircraft leasing	40,572	43,062	(5.8%)	0.59	0.66	(10.6%)
Employee profit share	32,974	19,039	73.2%	0.48	0.29	65.5%
<b>Total operating expenses</b>	<b>885,537</b>	<b>883,898</b>	<b>0.2%</b>	<b>12.83</b>	<b>13.60</b>	<b>(5.7%)</b>
<b>Total, excluding fuel and profit share</b>	<b>645,639</b>	<b>578,042</b>	<b>11.7%</b>	<b>9.35</b>	<b>8.90</b>	<b>5.1%</b>

During the three months ended September 30, 2015, operating expenses remained flat compared to the same period in 2014. On an ASM basis, operating expenses decreased by 5.7 per cent to 12.83 cents from 13.60 cents in the same period in 2014. This decrease was driven largely by a decrease in aircraft fuel expense partially offset by the increases in our employee profit share expense, maintenance expense and depreciation and amortization expense.

Our CASM, excluding fuel and profit share, increased by 5.1 per cent, which was outside of our previously disclosed guidance of up to 3.5 to 4.0 per cent, primarily as the result of lower than expected ASMs during the quarter in addition to a weaker Canadian dollar.

	Expense (\$ in thousands)			CASM (cents)		
	Nine months ended Sept. 30			Nine months ended Sept. 30		
	2015	2014	Change	2015	2014	Change
Aircraft fuel	632,317	846,514	(25.3%)	3.10	4.41	(29.7%)
Airport operations	412,510	377,378	9.4%	2.02	1.96	3.1%
Flight operations and navigational charges	372,560	339,616	9.7%	1.83	1.77	3.4%
Sales and distribution	281,304	282,598	1.0%	1.40	1.47	(4.8%)
Depreciation and amortization	189,684	172,044	10.3%	0.93	0.90	3.3%
Marketing, general and administration	175,108	164,030	6.8%	0.86	0.85	1.2%
Maintenance	171,599	151,612	13.2%	0.84	0.79	6.3%
Inflight	150,098	130,410	15.1%	0.74	0.68	8.8%
Aircraft leasing	132,208	136,904	(3.4%)	0.65	0.71	(8.5%)
Employee profit share	92,096	49,388	102.9%	0.46	0.24	93.7%
<b>Total operating expenses</b>	<b>2,613,484</b>	<b>2,646,294</b>	<b>(1.2%)</b>	<b>12.83</b>	<b>13.78</b>	<b>(6.9%)</b>
<b>Total, excluding fuel and profit share</b>	<b>1,889,071</b>	<b>1,754,392</b>	<b>7.7%</b>	<b>9.27</b>	<b>9.13</b>	<b>1.5%</b>

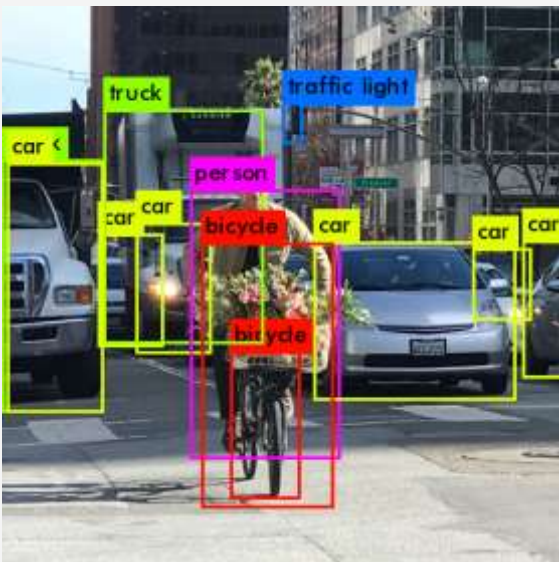
During the nine months ended September 30, 2015, operating expenses decreased by 1.2 per cent to \$2,613.5 million as compared to \$2,646.3 million in the same period in 2014, primarily driven by the year-over-year decrease in aircraft fuel expense, partially offset by the increase in our employee profit share expense.

**Output:** Document contents with tabular information:

- 1) Table border for each table
- 2) Partitioning table contents into cells
- 3) Both vertical and horizontal alignment of cells

# Tables and Cells as Objects

# Leverage object detection improvements with deep learning networks to detect tables and cells



## Notes to Consolidated Financial Statements (continued)

Dollars in millions except per share amounts

### Pending Acquisition

**Time Warner Inc.** On October 22, 2016, we entered into and announced a merger agreement (Merger Agreement) to acquire Time Warner Inc. (Time Warner) in a 50% cash and 50% stock transaction for \$107.50 per share of Time Warner common stock, or approximately \$85,400 at the date of the announcement (Merger). Combined with Time Warner's net debt at December 31, 2017, the total transaction value is approximately \$108.523. Each share of Time Warner common stock will be exchanged for 253.75 per share in cash and a number of shares of AT&T common stock equal to the exchange ratio. If the average stock price (as defined in the Merger Agreement) at the time of closing the Merger is between (or equal to) \$37.411 and \$41.349 per share, the exchange ratio will be the quotient of 253.75 divided by the average stock price. If the average stock price is greater than \$41.349, the exchange ratio will be 1.900. If the average stock price is less than \$37.411, the exchange ratio will be 1.437. Post-transaction, Time Warner shareholders will own between 14.4% and 15.7% of AT&T shares on a fully-diluted basis based on the number of AT&T shares outstanding.

Time Warner is a global leader in media and entertainment whose major businesses encompass an array of some of the most respected and successful media brands. The deal combines Time Warner's vast library of content and ability to create new premium content for audiences around the world with our extensive customer relationships and distribution, one of the world's largest pay-TV subscriber bases and leading scale in TV, media and broadband distribution. The Merger Agreement was approved by Time Warner shareholders on February 15, 2017. The transaction has been approved by all the relevant foreign jurisdictions. On November 20, 2017, the U.S. Department of Justice filed a civil antitrust lawsuit against AT&T, challenging our proposed acquisition of Time Warner. The case will be heard in the U.S. District Court for the District of Columbia on March 18, 2018. On December 21, 2017, an agreement was reached with Time Warner to extend the original termination date of April 22, 2018 to June 21, 2018 if the Merger is terminated as a result of reaching the extended termination date, and at that time one or more of the conditions relating to certain regulatory approvals have not been satisfied or there is a final, non-appealable order preventing the transaction relating to antitrust laws, communications laws, utilities laws or foreign regulatory laws, then under certain circumstances, we would be obligated to pay Time Warner \$500.

### Dispositions

**VP Holdings LLC** In June 2017, VP Holdings LLC was acquired by Dex Media. Our results include a gain of \$36 for our portion of the proceeds.

### NOTE 6. PROPERTY, PLANT AND EQUIPMENT

Property, plant and equipment is summarized as follows at December 31:

	2017	2016
Land	\$ 1,630	\$ 1,663
Buildings and improvements	2,44	36,319
Central office equipment	310	94,076
Cable, wiring and conduit	1510	87,895
Software	1417	2,987
Other equipment	3120	90,037
Software	31	16,750
Under construction	4,645	5,118
	313,499	319,648
Accumulated depreciation and amortization	188,277	194,738
Property, plant and equipment, net	\$125,222	\$124,899

Includes certain intangible assets.

Our depreciation expense was \$19,761 in 2017, \$20,661 in 2016 and \$19,012 in 2015. Depreciation expense included amortization of software totaling \$2,810 in 2017, \$2,362 in 2016 and \$1,660 in 2015.

During the fourth quarter of 2017, we announced that certain copper assets will not be necessary to support future network activity due to fiber deployment plans in particular markets. We recorded a nonrecurring pretax charge of \$2,883 to abandon these assets.

During the fourth quarter of 2016, we recorded a nonrecurring pretax charge of \$272 for the impairment of certain wireless assets that were under construction. These assets primarily related to capitalized costs for wireless sites that are no longer in our construction plans.

Certain facilities and equipment used in operations are leased under operating or capital leases. Rental expenses under operating leases were \$4,953 for 2017, \$4,482 for 2016 and \$5,029 for 2015. As of December 31, 2017, the future minimum rental payments under noncancelable operating leases for the years 2018 through 2022 were \$1,606, \$1,614, \$2,266 and \$2,026, with \$9,256 due thereafter. Certain real estate operating leases contain renewal options that may be exercised. At December 31, 2017, the present value of the future minimum rental payments under capital leases for the years 2018 through 2022 were \$142, 590, \$100, \$103 and \$115, with \$1,268 due thereafter.

During fiscal 2016, the Company acquired a direct interest in Vice for \$400 million of cash, and at September 29, 2018 owned an 11% interest. The Company accounts for its interest in Vice as an equity method investment.

During fiscal 2018, the Company recorded a \$157 million impairment of its interest in Vice.

### Hulu

At the end of fiscal 2015, the Company had a 13% interest in Hulu, a joint venture owned one-third each by the Company, 21CF and Comcast Corporation. Warner Media LLC (WML) acquired a 10% interest from Hulu for \$0.6 billion in August 2016, which diluted the Company's ownership interest to 30%. In addition, WML has made \$0.2 billion in subsequent capital contributions. For not more than 36 months from August 2016, WML has the right to sell its shares to Hulu and Hulu has the right to purchase the shares from WML under certain limited-circumstances arising from regulatory review. The Company and 21CF have agreed to make a capital contribution for up to approximately \$0.4 billion each if Hulu is required to repurchase WML shares. The August 2016 transaction resulted in a deemed sale by the Company of a portion of its interest in Hulu at a gain of approximately \$175 million. The Company expects to recognize the gain if and when the put and call options expire. Following completion of the 21CF acquisition the Company will consolidate Hulu's financial results and assume 21CF's capital contribution obligations.

The Company accounts for its interest in Hulu as an equity method investment.

### Goodwill

The changes in the carrying amount of goodwill for the years ended September 29, 2018 and September 30, 2017 are as follows:

	Media Networks	Parks and Resorts	Studio Entertainment	Consumer Products & Interactive Media	Unaffiliated <sup>(1)</sup>	Total
Balance at Oct. 1, 2016	\$ 16,145	\$ 291	\$ 6,830	\$ 4,344	\$ —	\$ 27,610
Acquisitions	—	—	—	—	3,600	3,600
Dispositions	—	—	—	—	—	—
Other, net	(20)	—	(13)	49	—	16
Balance at Sept. 30, 2017	\$ 16,125	\$ 291	\$ 6,817	\$ 4,393	\$ 3,600	\$ 31,426
Acquisitions	—	—	—	—	—	—
Dispositions	—	—	—	—	—	—
Other, net	3,063	—	347	33	(3,600)	(177)
Balance at Sept. 29, 2018	\$ 19,188	\$ 291	\$ 7,164	\$ 4,426	\$ —	\$ 31,269

<sup>(1)</sup> Other, net primarily represents the allocation of BAMTch goodwill to acquiree based on the final purchase price allocation and also includes the impact of updates to our initial estimated fair value of intangible assets related to BAMTch.

## 4 Other Income, net

Other income, net is as follows:

	2018	2017	2016
Gains on sales of real estate and property rights	\$ 560	\$ —	\$ —
Settlement of litigation	38	(177)	—
Gain related to the acquisition of BAMTch	3	255	—
Other income, net	\$ 601	\$ 78	\$ —

Gains from sales of real estate and property rights

In fiscal 2018, the Company recorded gains of \$560 million in connection with the sale of real estate and property rights in New York City.

### Settlement of litigation

In fiscal 2018, the Company recorded \$38 million in insurance recoveries in connection with the settlement of a litigation matter for which the Company recorded a charge of \$177 million, net of committed insurance recoveries in fiscal 2017.



# But where do we get the data???

ICDAR 2013 Dataset [1]: 256 table examples with table structure

ICDAR 2019 Dataset [2]: 80 page images with table structure

Marmot dataset [3]: 2000 total labelled PDF pages with 1349 tables

Too small for training

TableBank [4]: 145k labelled table examples ... but labelled tables structures only have logical coordinates

Missing critical information

[1] M. Göbel et al., *ICDAR 2013 Table Competition* ICDAR 2013

[2] L.Gao et. al., *ICDAR 2019 Competition on table detection and recognition (CTDAR)*. ICDAR 2019

[3] [http://www.icst.pku.edu.cn/cdpd/data/marmot\\_data.htm](http://www.icst.pku.edu.cn/cdpd/data/marmot_data.htm)

[4] Minghao Li et al. *Tablebank: Table benchmark for image-based table detection and recognition*. LREC 2020

# PubTabNet Dataset

Zhong et al. *Image-Based Table Recognition: Data, Model, and Evaluation* ECCV 2020

Enhanced scientific documents with cell bounding boxes with automated text matching.

568K Tables from PubMed

[ibm.biz/pubtabnet](http://ibm.biz/pubtabnet)

The image shows a page from a scientific paper. At the top, there is a hierarchical diagram of a model structure. Below it, there is a table with columns for 'Type of studies', 'Year', 'Number of studies', 'Abstracts', 'Full text', and 'Total'. The table lists various study types and their corresponding counts. Below the table, there is a bar chart showing the distribution of studies across different years.

The image shows a page from a scientific paper with a table of results. The table has columns for 'Year', 'Number of studies', 'Abstracts', 'Full text', and 'Total'. The table lists various study types and their corresponding counts. The text in the table is highlighted in red, indicating automated text matching.

The image shows a page from a scientific paper with a table of results. The table has columns for 'Year', 'Number of studies', 'Abstracts', 'Full text', and 'Total'. The table lists various study types and their corresponding counts. The text in the table is highlighted in red, indicating automated text matching.

# FinTabNet Dataset

New dataset with auto-generated table boundary, cell boundary and structure annotations.

>110k Tables from annual reports of S&P 500 companies.

[ibm.biz/fintabnet](https://ibm.biz/fintabnet)

**International Segment Results**

	2017	2016	2015	2017 (%)	2016 (%)
Revenue (including licenses)	\$6,674	\$6,631	\$7,312	61.3%	57.3%
Revenue adjustments	\$6,436	\$6,351	\$7,312	7.1	1.1
Revenue (excl.)	\$447	\$280	\$0	4.2	0.0
Revenue (excl.)(1)	788	571	241	86.7	51.6
Operating income	\$246	\$281	\$223	1.6	1.7
Operating income (excl.)(1)	\$1,404	\$1,338	\$1,319	9.4	11.0
Operating income (excl.)	\$1,214	\$1,057	\$828	4.4	6.3
Operating income (excl.)	\$622	\$726	\$335	7.8	7.4
Operating income (excl.)	\$1,911	\$1,133	\$253	26.5	17.0
Operating income (excl.)	\$7	\$1	\$1	0.1	0.1
Operating income (excl.)	\$1,244	\$1,041	\$1,462	16.8	13.3

The following tables highlight other key measures of performance for the international segment.

	2017	2016	2015	2017 (%)	2016 (%)
Revenue (including licenses)	\$6,674	\$6,631	\$7,312	61.3%	57.3%
Revenue adjustments	\$6,436	\$6,351	\$7,312	7.1	1.1
Revenue (excl.)	\$447	\$280	\$0	4.2	0.0
Revenue (excl.)(1)	\$4,499	\$4,761	\$4,214	37.4	40.1
Revenue (excl.)	\$254	\$281	\$25	2.2	0.3
Operating income	\$246	\$281	\$223	1.6	1.7
Operating income (excl.)(1)	\$1,404	\$1,338	\$1,319	9.4	11.0
Operating income (excl.)	\$1,214	\$1,057	\$828	4.4	6.3
Operating income (excl.)	\$622	\$726	\$335	7.8	7.4
Operating income (excl.)	\$1,911	\$1,133	\$253	26.5	17.0
Operating income (excl.)	\$7	\$1	\$1	0.1	0.1
Operating income (excl.)	\$1,244	\$1,041	\$1,462	16.8	13.3

**USE OF ESTIMATES AND ASSUMPTIONS**

The financial statements are prepared on the basis of estimates and assumptions. Management's estimates and assumptions are based on our assessment of existing facts, conditions, and circumstances.

**EXPLANATORY INFORMATION**

**DEFERRED TAX ASSETS AND LIABILITIES**

Deferred tax assets and liabilities are recognized for the future tax consequences of events that have already occurred. Deferred tax assets and liabilities are measured using enacted tax rates estimated to apply to taxable income in the years in which the deferred tax assets and liabilities are expected to be realized or settled.

**Table of Contents**

valuation of impairment. The Company reviews certain store assets for potential impairment. This fair value is determined based on the income approach, in which the Company utilizes internal cash flow projections over the life of the underlying lease agreements discounted based on the Company's cost-adjusted rate. Three measures of fair value, and related inputs, are considered a Level 3 approach under the fair value hierarchy. Refer to "Note 3 - Summary of Significant Accounting Policies" under the caption "Impairment of Long-Lived Assets and Long-Lived Assets to be Disposed-of" for information regarding the impairment charges recorded in fiscal 2014, 2017 and 2018.

The Company's liability to all single-class assets is recorded at carrying value, and, if impaired, an adjusted fair value using Level 3 inputs. See "Note 3 - Goodwill and Intangible Assets" for further information regarding the process of determining the fair value of these assets.

**Fair Value of Plans of Investment**

The carrying amounts of Cash and cash equivalents, Restricted cash and Accounts payable as reported in the accompanying consolidated balance sheets approximate fair value due to their short-term maturities.

The aggregate fair values and carrying values of the Company's long-term investments were as follows:

	February 2, 2018	February 3, 2017
in millions		
<b>Level 1:</b>		
Short-Term and Acquisition Value	\$ 4,196.0	\$ 4,275.3
<b>Level 2:</b>		
Equity Instruments	\$ 2.87.6	\$ 2.176.3

The fair value of the Company's 5.00% Senior Note due 2021 and the Note to Purchase the "Senior Note", and the fair value of the 3.75% Acquisition Note due 2022 and 3.75% Acquisition Note due 2023 (together, "the Acquisition Notes") that were acquired during the first quarter of 2016, were determined using Level 3 inputs as quoted prices in active markets for identical assets or liabilities are available. The fair values of the Term Loan A-1 and Term Loan B-1, which the Company prepaid in full during the first quarter of 2018, were determined using Level 3 inputs as quoted prices are readily available from pricing services, but the prices are not published. The carrying values of the Company's Revolving Credit Facility on February 2, 2018 and the Company's Tranche A Revolving Credit Facility on February 3, 2017, approximated their fair values because the interest rates were with market interest rates.

**NOTE 8 - SHAREHOLDERS' EQUITY**

**Preferred Stock**

The Company is authorized to issue 300,000 shares of Preferred Stock, \$100 per share per share. No preferred shares are issued and outstanding as of February 2, 2018 and February 3, 2017.

So, just finetune the existing object detectors, right?

# Motivation for a Specialized Table and Cell Detector

Tables and cells have very different sizes and aspect ratios, unlike the classic object detection classes

Cells are always inside tables and tables always contain cells

## Notes to Consolidated Financial Statements (continued)

Dollars in millions except per share amounts

### Pending Acquisition

**Time Warner Inc.** On October 22, 2016, we entered into and announced a merger agreement (Merger Agreement) to acquire Time Warner Inc. (Time Warner) in a 50% cash and 50% stock transaction for \$107.50 per share of Time Warner common stock, or approximately \$65.600 at the date of the announcement (Merger). Combined with Time Warner's net debt at December 31, 2011, the total transaction value is approximately \$106,523. Each share of Time Warner common stock will be exchanged for \$53.75 per share in cash and a number of shares of AT&T common stock equal to the exchange ratio. If the average stock price (as defined in the Merger Agreement) at the time of closing the Merger is between (or equal to) \$37.413 and \$41.349 per share, the exchange ratio will be the quotient of \$53.75 divided by the average stock price. If the average stock price is greater than \$41.349, the exchange ratio will be 1.300. If the average stock price is less than \$37.413, the exchange ratio will be 1.437. Post-transaction, Time Warner shareholders will own between 14.4% and 15.7% of AT&T shares on a fully-diluted basis based on the number of AT&T shares outstanding.

Time Warner is a global leader in media and entertainment whose major businesses encompasses an array of some of the most respected and successful media brands. The deal combines Time Warner's vast library of content and ability to create new premium content for audiences around the world with our extensive customer relationships and distribution, one of the world's largest pay-TV subscriber bases and leading scale in TV mobile and broadband distribution.

The Merger Agreement was approved by Time Warner shareholders on February 15, 2016. The transaction has been approved by all requisite foreign jurisdictions. On November 20, 2017, the U.S. Department of Justice filed a civil antitrust lawsuit against AT&T challenging our proposed acquisition of Time Warner. The case will be heard in the U.S. District Court for the District of Columbia on March 19, 2018. On December 21, 2017, an agreement was reached with Time Warner to extend the original termination date of April 22, 2018 to June 21, 2018. If the Merger is terminated as a result of reaching the extended termination date, none of that time one or more of the conditions relating to certain regulatory approvals have not been satisfied or there is a final non-appealable order preventing the transaction relating to antitrust laws, communications laws, utilities laws or foreign regulatory laws, then under certain circumstances, we would be obligated to pay Time Warner \$500.

### Dispositions

**YP Holdings LLC.** In June 2017, YP Holdings LLC was acquired by Dex Media. Our results include a gain of \$36.0 million of the proceeds.

### NOTE 6. PROPERTY, PLANT AND EQUIPMENT

Property, plant and equipment is summarized as follows at December 31:

	2017	2016
Land	\$ 2,830	\$ 1,843
Buildings and improvements	36,319	35,036
Central office equipment <sup>1</sup>	84,076	92,924
Cable, wiring and conduit	67,095	73,273
Satellite	2,887	7,110
Other equipment	40,017	86,436
Software	18,750	14,472
Under construction	4,045	5,138
<b>Accumulated depreciation and amortization</b>	<b>313,499</b>	<b>315,648</b>
Property, plant and equipment—net	188,277	194,149
Intangible assets	\$15,222	\$12,899
<b>Total assets under construction</b>		

Our depreciation expense was \$13,761 in 2017, \$20,861 in 2016 and \$15,389 in 2015. Depreciation expense included amortization of software totaling \$2,810 in 2017, \$2,362 in 2016 and \$1,066 in 2015.

During the fourth quarter of 2017, we determined that certain copper assets will not be necessary to support future network activity due to fiber deployment plans in particular markets. We recorded a non-cash pretax charge of \$2,883 to abandon these assets.

During the fourth quarter of 2016, we recorded a non-cash pretax charge of \$278 for the impairment of certain wireless assets that were under construction. These assets primarily related to construction costs for wireless sites that are no longer in our controlled plans.

Certain facilities and equipment used in operations are leased under operating or capital leases. Rental expenses under operating leases were \$4,953 for 2017, \$4,482 for 2016 and \$5,025 for 2015. At December 31, 2017, the future minimum rental payments under noncancelable operating leases for the years 2018 through 2022 were \$3,846, \$3,686, \$3,414, \$2,968 and \$2,656, with \$9,256 due thereafter. Certain real estate operating leases contain renewal options that may be exercised. At December 31, 2017, the present value of the future minimum rental payments under capital leases for the years 2018 through 2022 were \$142,590, \$100,510 and \$115, with \$1,268 due thereafter.

During fiscal 2016, the Company acquired a direct interest in Vice for \$400 million of cash, and at September 29, 2018 owned an 11% interest. The Company accounts for its interest in Vice as an equity method investment.

During fiscal 2018, the Company recorded a \$157 million impairment of its interest in Vice.

### Hulu

At the end of fiscal 2015, the Company had a 33% interest in Hulu, a joint venture owned one-third each by the Company, 21CF and Comcast Corporation. Warner Media LLC (WM) acquired a 10% interest from Hulu for \$0.6 billion in August 2016, which diluted the Company's ownership interest to 30%. In addition, WM has made \$0.2 billion in subsequent capital contributions. For not more than 36 months from August 2016, WM has the right to sell its shares to Hulu and Hulu has the right to purchase the shares from WM under certain limited circumstances arising from regulatory review. The Company and 21CF have agreed to make a capital contribution for up to approximately \$0.4 billion each if Hulu is required to repurchase WM's shares. The August 2016 transaction resulted in a deemed sale by the Company of a portion of its interest in Hulu at a gain of approximately \$175 million. The Company expects to recognize the gain in 2018 when the put and call options expire. Following completion of the 21CF acquisition, the Company will consolidate Hulu's financial results and assume 21CF's capital contribution obligations.

The Company accounts for its interest in Hulu as an equity method investment.

### Goodwill

The changes in the carrying amount of goodwill for the years ended September 29, 2018 and September 30, 2017 are as follows:

	Media Networks	Prints and Brands	Studio Entertainment	Publishing Intellectual Property	Unallocated <sup>(1)</sup>	Total
Balance at Oct. 1, 2016	\$ 16,345	\$ 291	\$ 6,830	\$ 4,344	\$ —	\$ 27,810
Acquisition	—	—	—	—	3,600	3,600
Dispositions	—	—	—	—	—	—
Other, net	(20)	—	(11)	99	—	16
Balance at Sept. 30, 2017	\$ 16,325	\$ 291	\$ 6,817	\$ 4,393	\$ 3,600	\$ 31,426
Acquisition	—	—	—	—	—	—
Dispositions	—	—	347	33	(3,600)	(177)
Other, net	1,063	—	—	—	—	1,063
Balance at Sept. 29, 2018	\$ 19,388	\$ 291	\$ 7,164	\$ 4,426	\$ —	\$ 31,269

<sup>(1)</sup> Other, net primarily represents the allocation of BMTCo goodwill to segments based on the final purchase price allocation and also includes the impact of updates to our initial estimated fair value of intangible assets related to BMTCo.

## 4 Other Income, net

Other income, net is as follows:

	2018	2017	2016
Gains on sales of real estate and property rights	\$ 560	\$ —	\$ —
Settlement of litigation	18	(177)	—
Gains related to the acquisition of BMTCo	3	255	—
Other income, net	\$ 601	\$ 78	\$ —

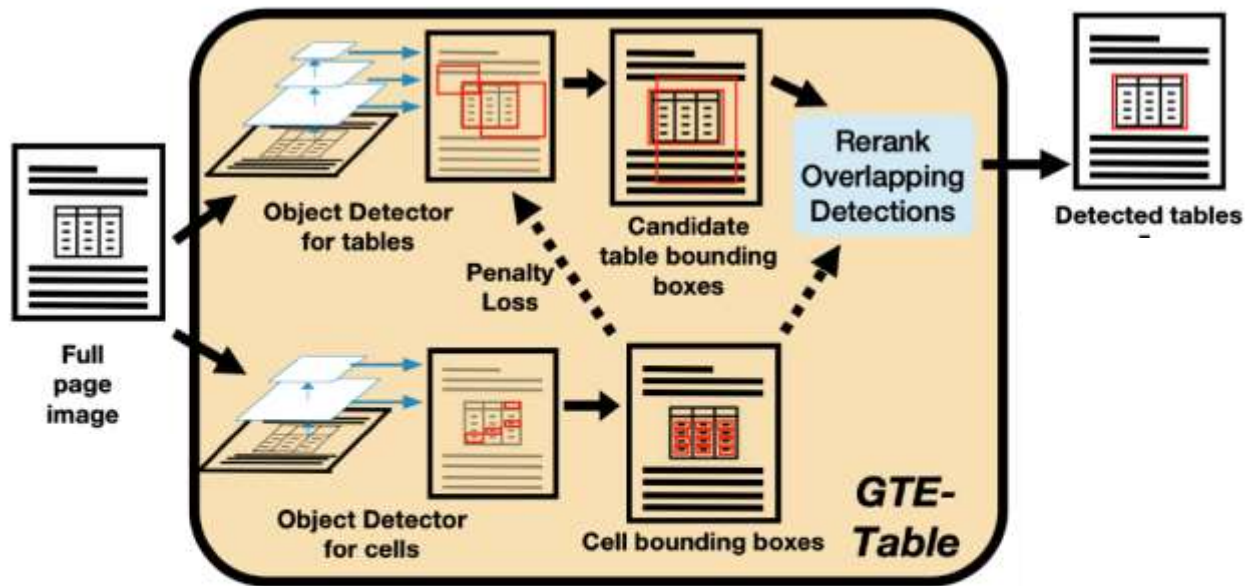
### Gains from sales of real estate and property rights

In fiscal 2018, the Company recorded gains of \$50 million in connection with the sale of real estate and property rights in New York City.

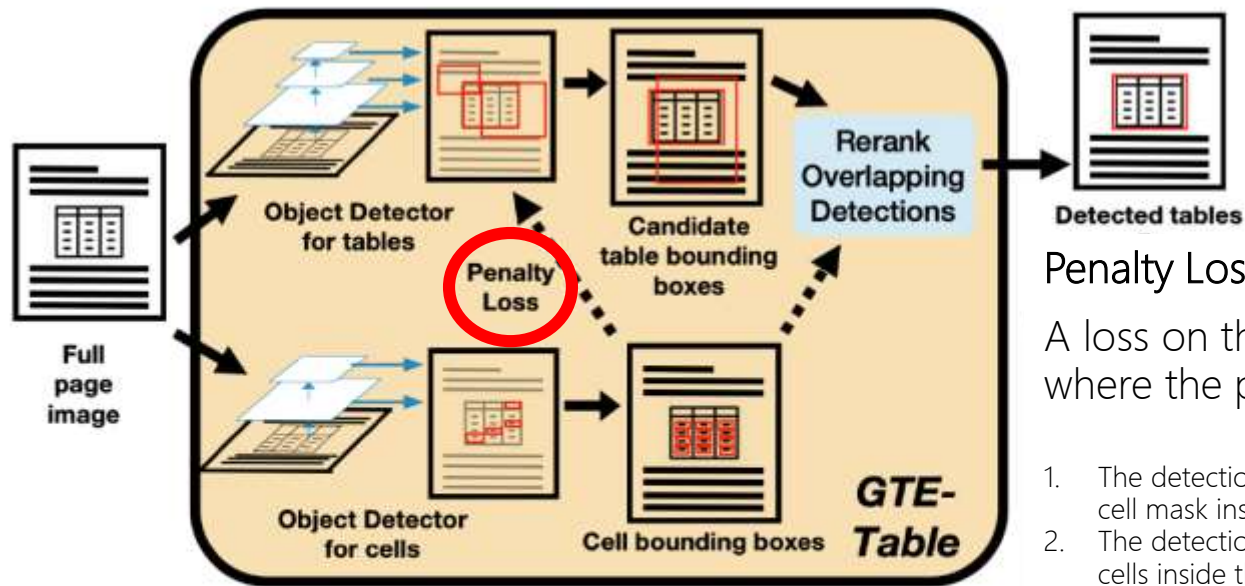
### Settlement of litigation

In fiscal 2018, the Company recorded \$18 million in insurance recoveries in connection with the settlement of a litigation matter for which the Company recorded a charge of \$177 million, net of committed insurance recoveries in fiscal 2017.

# Global Table Extractor (GTE)



# Penalty Loss

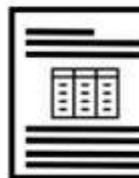


## Penalty Loss:

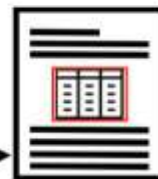
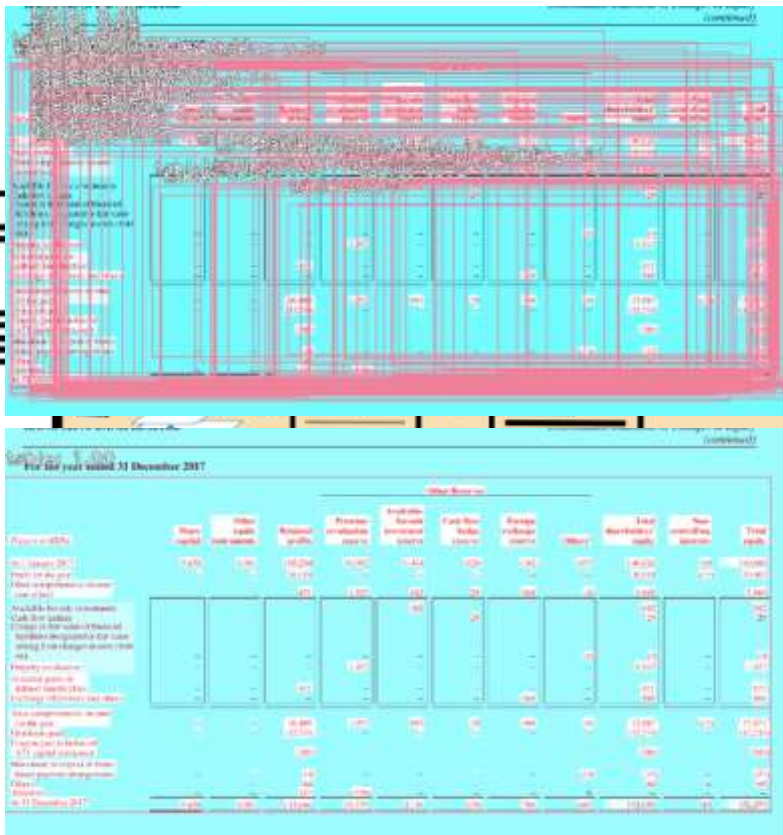
A loss on the table classification where the penalty is applied if:

1. The detection is a table but contains very little cell mask inside the bounding box.
2. The detection is not a table but contains a lot of cells inside the bounding box.

# Cell-Aware Table Reranking



Full page image



Detected tables

When tables with similar confidence are overlapping, we select the table based on its “table-ness”.

1. Have lots of cellular regions.
2. NOT have cellular regions just outside that is not being covered by a non-overlapping region.



# SoTA on Table Border Detection

## ICDAR 2013

Character level recall,  
precision and F1

Category	Method	Input type	Recall	Precision	F1
Commercial Softwares	<i>FineReader</i>	PDF	99.71	97.29	98.48
Non Deep Learning	<i>Nurminen</i> [8]	PDF	90.77	92.10	91.43
Deep Learning	<i>TableBank</i> [18]	Image	/	/	96.25
Ours	GTE	Image	<b>99.77</b>	<b>98.97</b>	<b>99.31</b>
Ablation	Detection-Base	Image	84.64	90.65	84.65
	GTE-Table-Sep	Image	95.71	98.18	95.71

## ICDAR 2019

Precision and Recall at different  
levels of Intersection Over Union  
(IOU)

Method	IOU = 0.8		IOU = 0.9		Weighted F1
	P	R	P	R	
<i>NLPR-PAL</i> [4]	93	93	86	86	93
<i>TableRadar</i> [4]	95	94	<b>90</b>	<b>89</b>	<b>94</b>
<i>GTE</i>	<b>96</b>	<b>95</b>	<b>90</b>	<b>89</b>	<b>94</b>

# Tables Come in Different Styles

Object detection models focus on local areas and cannot handle nuances of different global table styles.

Table 8.12 - Own brand shares (food only) for leading retailers, 1996

Names	Own brands market share	Number of items
Fransprix	28.0	n.a.
Casino	24.8	1800
Intermarché	24.7	2500
Géant	20.0	1800
Carrefour	18.9	1642
Micromark	18.7	1800
Système U	18.5	985
Continant	17.8	1440
Stoc	16.2	650
Auchan	15.7	1500
March	15.0	1100
Champion	15.1	1240
Leclerc	14.8	500
Cora	12.2	1224
Prismaic	11.7	550

Source: *Secodip-linéaires, 1997*

Table 8.13 illustrates the development of own brand market shares in supermarkets and hypermarkets:

Table 8.13 - National brands, Own brand and low price items shares for supermarkets and hypermarkets

	1991	1994	1995	1996
National Brands	80.6	75.0	73.3	76.0
Own Brands	14.7	17.1	17.4	17.1
Low price items	4.7	7.9	7.3	6.9

Source: *LSA, 1998*

Own brands growth is a major trend in the recent evolution of distribution. In 1995, own brands provided on average 20% of sales (25% of shelf space) compared to only 10% ten years before and the development goes on. Leclerc, for instance, which was initially opposed to own brand development (less than 7% of sales before 1997 were own brand), changed its strategy in 1997 and its goal is now to double its own brands turnover. Retailers try to reinforce the association between their name and their own products.

Table of Contents

Canada	Dollar Tree
Alberta	37
British Columbia	49
Manitoba	13
Ontario	110
Saskatchewan	16
Total	225

We lease the vast majority of our stores and expect to lease the majority of our new stores as we expand. Our leases typically provide for a short initial lease term, generally five years, with options to extend, however, in some cases we have initial lease terms of seven to fifteen years. We believe this leasing strategy enhances our flexibility to pursue various expansion opportunities resulting from changing market conditions. As current leases expire, we believe that we will be able to obtain lease renewals, if desired, for present store locations, or to obtain leases for equivalent or better locations in the same general area.

#### Distribution Centers

The following table includes information about the distribution centers that we operate in the United States. Except for 0.4 million square feet of our distribution center in San Bernardino, California, all of our distribution center capacity is owned. In 2018, we completed our Warrenburg, Missouri distribution center, which is 1.2 million square feet, automated and currently serves stores in our Dollar Tree segment. In 2016, we completed our 1.5 million square foot Cherokee County, South Carolina distribution center and expanded our Stockton, California distribution center by 0.3 million square feet. Our St. George, Utah distribution center services both Family Dollar and Dollar Tree stores. In addition, we ship select product from our Dollar Tree distribution centers to our Family Dollar distribution centers and in fiscal 2019, we expect to ship select product from our Dollar Tree distribution centers directly to certain of our Family Dollar stores. We believe our distribution center network is currently capable of supporting approximately \$28.0 billion in annual sales in the United States.

Dollar Tree Distribution Centers	Square Footage	Family Dollar Distribution Centers	Square Footage
Chesapeake, Virginia	400,000	Mathews, North Carolina	930,000
Olive Branch, Mississippi	425,000	West Memphis, Arkansas	850,000
Joliet, Illinois	1,470,000	Front Royal, Virginia	907,000
Stockton, California	854,000	Duncan, Oklahoma	907,000
Savannah, Georgia	1,014,000	Morehead, Kentucky	907,000
Briar Creek, Pennsylvania	1,003,000	Mapletoke, Iowa	907,000
Marietta, Oklahoma	1,004,000	Odesa, Texas	907,000
San Bernardino, California	802,000	Marianna, Florida	907,000
Ridgfield, Washington	665,000	Rome, New York	907,000
Windsor, Connecticut	1,001,000	Ashley, Indiana	814,000
Cherokee County, South Carolina	1,512,000	St. George, Utah*	814,000
Warrenburg, Missouri	1,200,000		

\*Services both Dollar Tree and Family Dollar stores

In 2018, we began construction on our Morrow County, Ohio distribution center, which will be 1.2 million square feet and automated, and will initially serve stores in our Dollar Tree segment. We expect this facility to be operational in the third quarter of 2019. In fiscal 2019, we announced plans to construct a new 1.2 million square foot distribution center in Rosenberg, Texas which is expected to provide service directly to Dollar Tree and Family Dollar stores and be operational by the summer of 2020. All future distribution centers will open with the capability to service both Dollar Tree and Family Dollar stores.

Each of our distribution centers contains advanced materials handling technologies, including radio-frequency inventory tracking equipment and specialized information systems. With the exception of our Ridgfield, Washington facility and our Mathews, North Carolina facility, each of our distribution centers in the United States also contains automated conveyor and sorting systems.

Distribution services in Canada are provided by a third party from facilities in British Columbia and Ontario.

42

Styrene-Acrylonitrile Trimer, NTP TR 573

TABLE 6  
Mean Body Weights of F<sub>1</sub> Pups to Postnatal Day 20  
in the 7-Week Postnatal and Postnatal Feed Study of Styrene-Acrylonitrile Trimer

Concentration (ppm)	Postnatal Day 1			Postnatal Day 7			Postnatal Day 14			Postnatal Day 20		
	Body Weight (g)	Weight Relative to Controls (%)	No.	Body Weight (g)	Weight Relative to Controls (%)	No.	Body Weight (g)	Weight Relative to Controls (%)	No.	Body Weight (g)	Weight Relative to Controls (%)	No.
<b>Male</b>												
0	39	5.8		10	8.8		11.5	25.7		35.0		
250	30	5.9	102	10	9.0	102	11.6	100	26.2	102	35.8	102
500	33	6.0	103	10	8.6	98	11.0	96	25.1	98	34.7	99
1,000	31	5.8	100	10	8.4	96	11.0	96	24.9	97	34.5	99
2,000	38	5.8	100	10	8.8	100	12.9	96	24.6	96	31.6**	99
4,000	27	5.3**	91	10	7.5**	85	10.4**	77	16.8**	65	19.8**	57
<b>Female</b>												
0	23	5.4		10	8.2		11.7	24.7		31.6		
250	34	5.6	104	10	8.5	104	12.9	102	24.7	100	33.9	101
500	32	5.4	100	10	8.2	100	12.7	100	25.0	101	33.6	100
1,000	40	5.5	102	10	8.0	98	12.4	98	24.2	98	32.7	97
2,000	49	5.3	98	10	8.2	100	12.4	98	23.7	98	30.3**	99
4,000	31	5.0**	93	10	7.3*	89	9.9**	78	16.1**	65	18.8**	56

\* Significantly different (P<0.05) from the control group by Dunnett's test

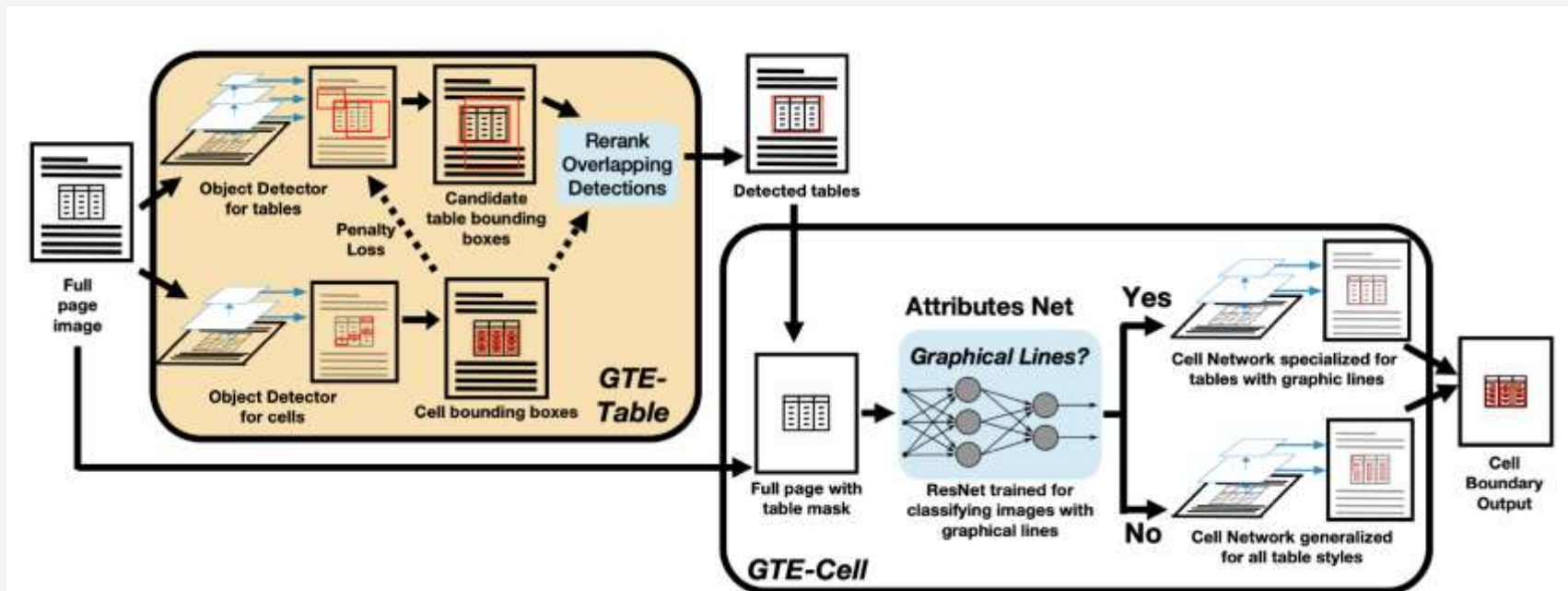
\*\* P<0.01

† Weights are given as group means.

\* Number of animals weighed on postnatal day 1

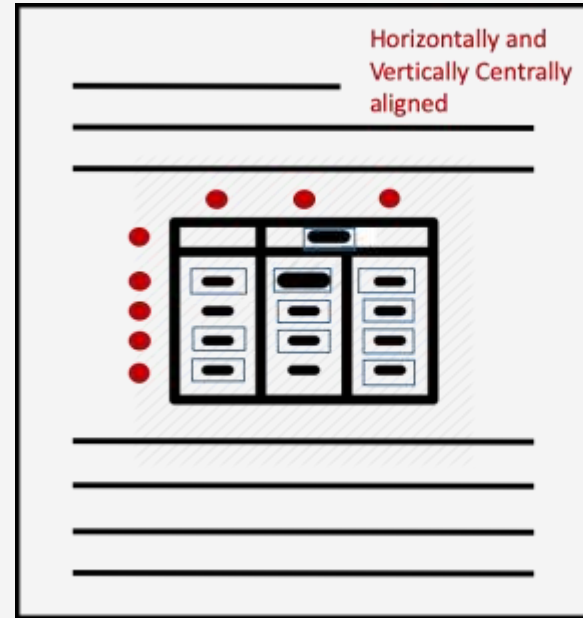
† Number of animals weighed on postnatal days 4, 7, 14, and 20

# Cell Detection Framework



# Cell Boundary to Structure with Position Clustering

1. Sample at each box center horizontally and vertically to determine **the number of rows and columns** in table
2. Determine **alignment** for columns and rows based on cell bounding box distance
3. Use **K-means clustering** with number of centers from step 3 for each direction with coordinates depending on alignment
4. **Assign cell boxes** to each row/column depending on alignment



# Detection to Structure with Location and Alignment Clustering

	All items		New items		Trend items	
	Number	Percent	Number	Percent	Number	Percent
<b>Content domain and process</b>						
Total items	135	100	60	100	75	100
<b>Purposes of reading</b>						
Literary experience	72	53	33	55	39	52
Acquire and use information	63	47	27	45	36	48
<b>Processes of comprehension</b>						
Focus on and retrieve explicitly stated information	33	24	14	23	19	25
Make straightforward inferences	46	34	20	33	26	35
Interpret and integrate ideas and information	38	28	18	30	20	27
Examine and evaluate content, language, and textual elements	18	13	8	13	10	13

	All items		New items		Trend items	
	Number	Percent	Number	Percent	Number	Percent
<b>Content domain and process</b>						
Total items	135	100	60	100	75	100
<b>Purposes of reading</b>						
Literary experience	72	53	33	55	39	52
Acquire and use information	63	47	27	45	36	48
<b>Processes of comprehension</b>						
Focus on and retrieve explicitly stated information	33	24	14	23	19	25
Make straightforward inferences	46	34	20	33	26	35
Interpret and integrate ideas and information	38	28	18	30	20	27
Examine and evaluate content, language, and textual elements	18	13	8	13	10	13

Faculty cluster	Population size	Sample size
Sciences	1269 (19.9%)	101(20.4%)
Social Sciences	3212 (50.6%)	247(50.0%)
Humanities	1168 (18.4%)	95(19.3%)
Civil Sciences	705 (11.1%)	51(10.3%)

Faculty cluster	Population size	Sample size
Sciences	1269 (19.9%)	101(20.4%)
Social Sciences	3212 (50.6%)	247(50.0%)
Humanities	1168 (18.4%)	95(19.3%)
Civil Sciences	705 (11.1%)	51(10.3%)

	THRESHOLD FOR RELEASES		
	to air kg/year	to water kg/year	to land kg/year
Asbestos	1	1	1
Chlorides (as total Cl)	-	2 million	2 million
Cyanides (as total CN)	-	50	50
Fluorides (as total F)	-	2 000	2 000
Particulate matter (PM10)	50 000	-	-
Total Nitrogen	-	50 000	50 000
Total Phosphorus	-	5 000	5 000

	THRESHOLD FOR RELEASES		
	to air kg/year	to water kg/year	to land kg/year
Asbestos	1	1	1
Chlorides (as total Cl)	-	2 million	2 million
Cyanides (as total CN)	-	50	50
Fluorides (as total F)	-	2 000	2 000
Particulate matter (PM10)	50 000	-	-
Total Nitrogen	-	50 000	50 000
Total Phosphorus	-	5 000	5 000

# Cell Structure Evaluation

Description	Initial balance	Increase	Decrease	Final balance
Accrued income	1 669	0	1 269	400
Deferred income	26 676	0	26 079	597
Accrued expenses	49 734	0	14 467	35 267

(a) Original table as in ground truth

Description	Initial balance	Increase	Decrease	Final balance
Accrued income	1 669	0	1 269	400
Deferred income	26 676	0	26 079	597
Accrued expenses	49 734	0	14 467	35 267

(b) Incorrectly recognized cell structure with split column

■ Correct adjacency relations   □ Incorrect adjacency relations

$$\text{Recall} = \frac{\text{correct adjacency relations}}{\text{total adjacency relations}} = \frac{24}{31} = 77.4\%$$

$$\text{Precision} = \frac{\text{correct adjacency relations}}{\text{detected adjacency relations}} = \frac{24}{28} = 85.7\%$$

## Cell Adjacency Relations

ICDAR2013: Match based on text

ICDAR2019: Match based on bounding box IOU

# SOTA for Cell Structure Recognition

## ICDAR 2013

Method	GT?	Rec.	Prec.	F1
<i>Nurminen</i> [8]	N	80.78	86.93	83.74
GTE	N	<b>92.72</b>	<b>94.41</b>	<b>93.50</b>
<i>Tensmeyer</i> [33]	Y	94.64	95.89	95.26
GTE	Y	<b>95.77</b>	<b>96.76</b>	<b>96.24</b>
Detection-Base	Y	76.66	80.63	78.10
GTE-Cell-Style-Mix -no-pt	Y	89.78	89.30	89.43
GTE-Cell-Style-Mix	Y	92.39	94.20	93.15
GTE-Cell-Border	Y	91.60	93.67	92.48

## ICDAR 2019

Method	IOU			Weighted F1
	0.1	0.5	0.6	
<i>NLPR-PAL</i> [4]	-	36.5	30.5	20.6
<i>CascadeTabNet</i> [23]	-	43.8	35.4	23.2
<i>GTE</i>	77.5	<b>54.8</b>	<b>38.5</b>	<b>24.8</b>

Is better really better?



# Let's Look an Example

Original Table:

	Revenue (\$Bn)	
	2008	2009
AAL	23.8	19.9
LUV	11.0	10.4
SAVE	1.10	0.70
ULCC	1.40	1.10
UAL	20.2	16.3
DAL	22.7	28.1
JBLUE	3.40	3.30
ALK	3.70	3.40



Ground Truth:

	Revenue (\$Bn)	
	2008	2009
AAL	23.8	19.9
LUV	11.0	10.4
SAVE	1.10	0.70
ULCC	1.40	1.10
UAL	20.2	16.3
DAL	22.7	28.1
JBLUE	3.40	3.30
ALK	3.70	3.40

# Which One is Better?

Candidate Table A

	Revenue (\$Bn)	
	2008	2009
AAL	23.8	19.9
LUV	11.0	10.4
SAVE	1.10	0.70
ULCC	1.40	1.10
UAL	20.2	16.3
DAL	22.7	28.1
JBLUE	3.40	3.30
ALK	3.70	3.40

Candidate Table B

	Revenue (\$Bn)	
	2008	2009
AAL	23.8	19.9
LUV	11.0	10.4
SAVE	1.10	0.70
ULCC	1.40	1.10
UAL	20.2	16.3
DAL	22.7	28.1
JBLU E	3.40	3.30
ALK	3.70	3.40

# Let's Compute the Metrics ... Wait ... Which One?

ICDAR 2013

	Candidate Table A			Candidate Table B		
	P	R	F1	P	R	F1
Table Area	1	0.79	0.88	1	0.81	0.90

ICDAR 2019

	Candidate Table A		Candidate Table B	
	IOU		IOU	
Table Area	0.8		0.8	

Candidate Table A

	Revenue (\$Bn)	
	2008	2009
AAL	23.8	19.9
LUV	11.0	10.4
SAVE	1.10	0.70
ULCC	1.40	1.10
UAL	20.2	16.3
DAL	22.7	28.1
JBLUE	3.40	3.30
ALK	3.70	3.40

Candidate Table B

	Revenue (\$Bn)	
	2008	2009
AAL	23.8	19.9
LUV	11.0	10.4
SAVE	1.10	0.70
ULCC	1.40	1.10
UAL	20.2	16.3
DAL	22.7	28.1
JBLUE	3.40	3.30
ALK	3.70	3.40

# Let's Compute the Metrics ... Wait ... Which One?

ICDAR 2013

	Candidate Table A			Candidate Table B		
	P	R	F1	P	R	F1
Table Area	1	0.79	0.88	1	0.81	0.90
Cell Adjacency	1	0.76	0.86	1	0.88	0.94

ICDAR 2019

	Candidate Table A		Candidate Table B	
	IOU		IOU	
Table Area	0.8		0.8	

Candidate Table A

	Revenue (\$Bn)	
	2008	2009
AAL	23.8	19.9
LUV	11.0	10.4
SAVE	1.10	0.70
ULCC	1.40	1.10
UAL	20.2	16.3
DAL	22.7	28.1
JBLUE	3.40	3.30
ALK	3.70	3.40

Candidate Table B

	Revenue (\$Bn)	
	2008	2009
AAL	23.8	19.9
LUV	11.0	10.4
SAVE	1.10	0.70
ULCC	1.40	1.10
UAL	20.2	16.3
DAL	22.7	28.1
JBLUE	3.40	3.30
ALK	3.70	3.40

# Let's Compute the Metrics ... Wait ... Which One?

## ICDAR 2013

	Candidate Table A			Candidate Table B		
	P	R	F1	P	R	F1
Table Area	1	0.79	0.88	1	0.81	0.90
Cell Adjacency	1	0.76	0.86	1	0.88	0.94

## ICDAR 2019

	Candidate Table A		Candidate Table B	
	IOU		IOU	
Table Area	0.8		0.8	

## Downstream Applications

	Candidate Table A			Candidate Table B		
	P	R	F1	P	R	F1
Functional metric <sup>[1]</sup>	1	0.75	0.86	1	0	0

Keep information to recover semantics for most data cells

Missing column headers → Cannot recover semantics for the data cells

## Candidate Table A

	Revenue (\$Bn)	
	2008	2009
AAL	23.8	19.9
LUV	11.0	10.4
SAVE	1.10	0.70
ULCC	1.40	1.10
UAL	20.2	16.3
DAL	22.7	28.1
JBLUE	3.40	3.30
ALK	3.70	3.40

## Candidate Table B

	Revenue (\$Bn)	
	2008	2009
AAL	23.8	19.9
LUV	11.0	10.4
SAVE	1.10	0.70
ULCC	1.40	1.10
UAL	20.2	16.3
DAL	22.7	28.1
JBLUE	3.40	3.30
ALK	3.70	3.40

[1] Max C. Göbel, Tamir Hassan, Ermelinda Oro, Giorgio Orsi: A methodology for evaluating algorithms for table understanding in PDF documents. Document Engineering 2012

# Our Goal

- Achieve SoTA on public benchmark datasets
- Significantly outperform our competitors on data and metrics that matter



How to make it works better?  
Retraining with interactive labeling

# TableLab: Interactive Labeling

The screenshot displays the TableLab interface, which is used for extracting and labeling tables from documents. The interface is divided into several sections:

- Left Panel:** Contains file selection and filtering options. It shows a list of extracted tables, with the first one selected. Below the list are buttons for 'Page 1', 'Template type 1', and 'Tables with confidence:'. At the bottom, there are buttons for '100', '100', '0', and '100'.
- Top Panel:** Shows the document title 'TableLab' and a 'Enlarge' button. Below this are tabs for 'MAGNIFY MODE' and 'OVERLAY MODE'.
- Main Document View:** Displays a page from 'GOLDMAN SACHS INTERNATIONAL (LIMITED COMPANY) Management Report'. The text discusses the company's exposure to credit risk, its debtors, and its cash at bank and in hand. Below the text are three tables:
  - Table 1:** 'Gross credit exposure and net credit exposure by financial asset class (in \$ millions)'. It shows data for 'As of August 2019' and 'As of December 2019' across categories like 'Gross credit exposure', 'Assets captured by market risk', 'Cash collateral', 'Security collateral retained', and 'Net credit exposure'.
  - Table 2:** 'Investment Grade'. It shows data for 'As of August 2019' and 'As of December 2019' across categories like 'Gross credit exposure', 'Assets captured by market risk', 'Cash collateral', 'Security collateral retained', and 'Net credit exposure'.
  - Table 3:** 'Non-Investment Grade'. It shows data for 'As of August 2019' and 'As of December 2019' across categories like 'Gross credit exposure', 'Assets captured by market risk', 'Cash collateral', 'Security collateral retained', and 'Net credit exposure'.
- Right Panel:** Contains a 'Submit Annotations' button and two interactive table views. The top view is 'Table 1 Confidence: 100' and the bottom view is 'Table 2 Confidence: 100'. Both views show the same data as the tables in the main document but with interactive elements like 'Merge +', 'Split +', 'Move +', 'Add New +', and 'Regenerate Table' buttons. The tables are color-coded to highlight specific cells.



# Preliminary Experimental Results for Retraining

## Table Boundary Detection

Method	CEDAR	EDGAR	Invoices	Appraisals	Health Docs
GTE	0.94	0.84	0.47	0.85	0.93
GTE with Retraining	0.96	0.91	0.92	0.96	0.98

## Cell Adjacency Detection

Method	CEDAR	EDGAR	Invoices	Appraisals	Health Docs
GTE	0.88	0.62	0.42	0.71	0.55
GTE with Retraining	0.90	0.82	0.68	0.90	0.77

**Summary:** Retraining is effective even with small amount of labeled data

## Dataset

20 pages with tables per category:  
10 for retraining, 10 for testing

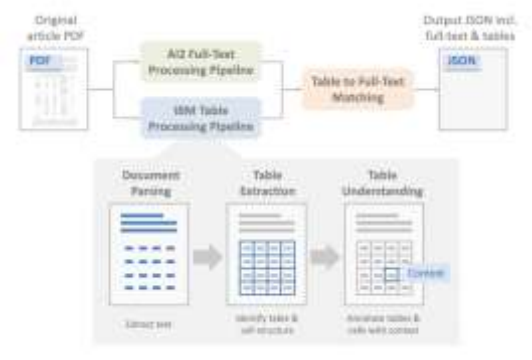
## Evaluation Metric

F1 metric for Table Boundary and  
Cell Adjacency as defined in [1]

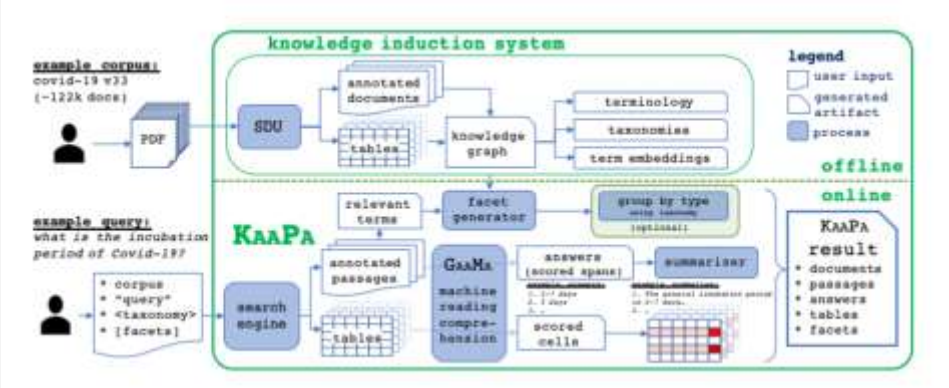
[1] Göbel et al. "A Methodology for Evaluating Algorithms for Table Understanding in PDF Documents". DocEng '12

# Example Use Cases – Research & Social Good

**CORD-19** [Wang et al, ACL-CORD-19'21]



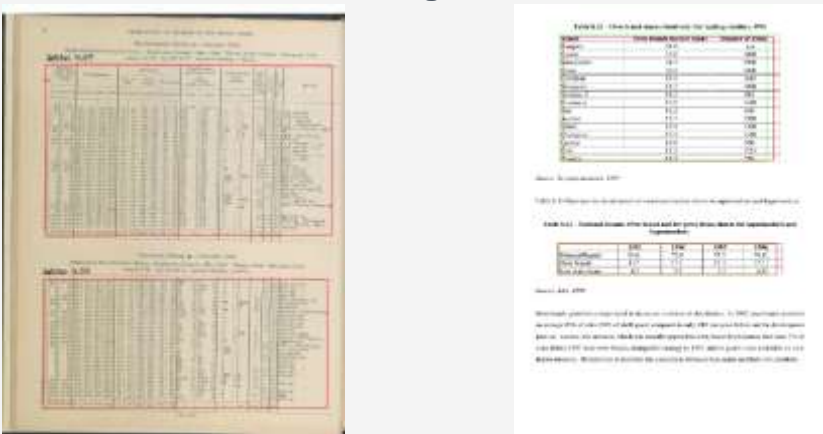
**Table QA** [Fauceglia et al, AAAI'21, Glass et al, NAACL'21]



**Weather.com COVID-19 Dashboard**



**Better Understand Climate Change via Historical Records**



# Example Use Cases - Business

## Invoice Understanding

**AUBRIE PICK** Invoice  
Invoice # 22324  
Total \$23,200.00

**Order Summary**

Item	Description	Qty	Unit	Amount
1	Standard	100	Units	1000.00
1	Advanced	100	Units	1000.00
1	Special	100	Units	1000.00
1	Elite	100	Units	1000.00
1	Ultimate	100	Units	1000.00
1	Platinum	100	Units	1000.00
1	Diamond	100	Units	1000.00
1	Emerald	100	Units	1000.00
1	Sapphire	100	Units	1000.00
1	Ruby	100	Units	1000.00
1	Peridot	100	Units	1000.00
1	Garnet	100	Units	1000.00
1	Amethyst	100	Units	1000.00
1	Opal	100	Units	1000.00
1	Malachite	100	Units	1000.00
1	Obsidian	100	Units	1000.00
1	Flint	100	Units	1000.00
1	Jet	100	Units	1000.00
1	Onyx	100	Units	1000.00
1	Spinel	100	Units	1000.00
1	Zircon	100	Units	1000.00
1	Topaz	100	Units	1000.00
1	Quartz	100	Units	1000.00
1	Amber	100	Units	1000.00
1	Agate	100	Units	1000.00
1	Jasper	100	Units	1000.00
1	Malachite	100	Units	1000.00
1	Obsidian	100	Units	1000.00
1	Flint	100	Units	1000.00
1	Jet	100	Units	1000.00
1	Onyx	100	Units	1000.00
1	Spinel	100	Units	1000.00
1	Zircon	100	Units	1000.00
1	Topaz	100	Units	1000.00
1	Quartz	100	Units	1000.00
1	Amber	100	Units	1000.00
1	Agate	100	Units	1000.00
1	Jasper	100	Units	1000.00

## Purchase Order Understanding

**PURCHASE ORDER**

ABC Corp  
123 Main St  
New York, NY 10001  
Phone: 212-555-1234

DEF Inc  
456 Elm St  
New York, NY 10002  
Phone: 212-555-5678

ITEM NO	DESCRIPTION	QTY	UNIT	PRICE
001	Standard	100	Units	1000.00
002	Advanced	100	Units	1000.00
003	Special	100	Units	1000.00
004	Elite	100	Units	1000.00
005	Ultimate	100	Units	1000.00
006	Platinum	100	Units	1000.00
007	Diamond	100	Units	1000.00
008	Emerald	100	Units	1000.00
009	Sapphire	100	Units	1000.00
010	Ruby	100	Units	1000.00
011	Peridot	100	Units	1000.00
012	Garnet	100	Units	1000.00
013	Amethyst	100	Units	1000.00
014	Opal	100	Units	1000.00
015	Malachite	100	Units	1000.00
016	Obsidian	100	Units	1000.00
017	Flint	100	Units	1000.00
018	Jet	100	Units	1000.00
019	Onyx	100	Units	1000.00
020	Spinel	100	Units	1000.00
021	Zircon	100	Units	1000.00
022	Topaz	100	Units	1000.00
023	Quartz	100	Units	1000.00
024	Amber	100	Units	1000.00
025	Agate	100	Units	1000.00
026	Jasper	100	Units	1000.00

## Contract Understanding

1. **Service Level** - 99.99% availability for the Service during the term of the Contract.

2. **Service Level** - 99.99% availability for the Service during the term of the Contract.

3. **Service Level** - 99.99% availability for the Service during the term of the Contract.

4. **Service Level** - 99.99% availability for the Service during the term of the Contract.

5. **Service Level** - 99.99% availability for the Service during the term of the Contract.

6. **Service Level** - 99.99% availability for the Service during the term of the Contract.

7. **Service Level** - 99.99% availability for the Service during the term of the Contract.

8. **Service Level** - 99.99% availability for the Service during the term of the Contract.

9. **Service Level** - 99.99% availability for the Service during the term of the Contract.

10. **Service Level** - 99.99% availability for the Service during the term of the Contract.

# Open Challenges

Quality

Runtime Performance

Customizability

Evaluation

Multi-task learning

