Table Extraction and Understanding for Scientific and Enterprise Applications

Doug Burdick

Alexandre V Evfimievski

Nancy Wang

Yannis Katsis

Marina Danilevsky

IBM Research - Almaden

Outline

- Introduction
  - Problem Definition
  - Challenges
  - Takeaways
- Table Identification
- Table Understanding
- Conclusion

Introduction

Introduction Outline

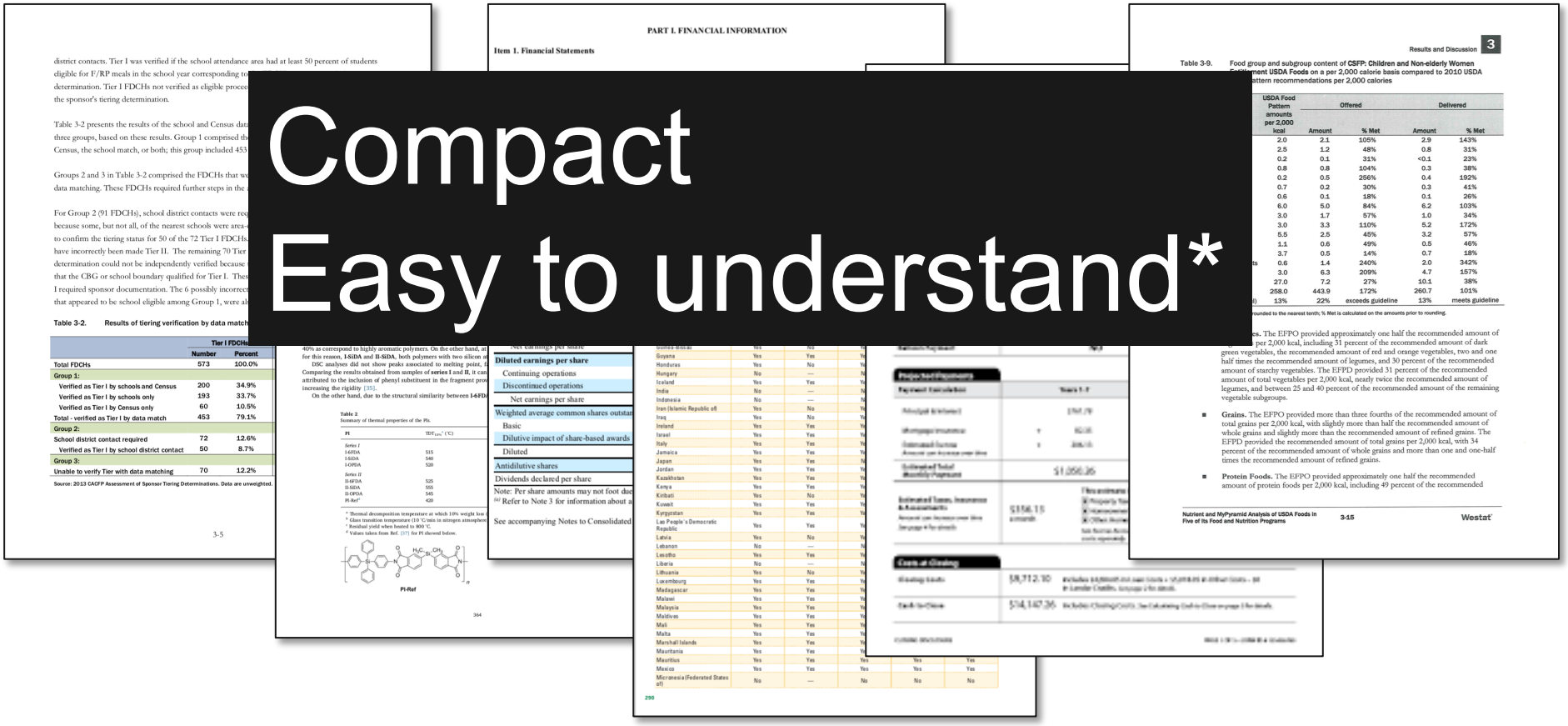
- Problem definition
  - Table Extraction
  - Table Understanding
- Challenges
  - Limited document format support for table structure
  - Table variety
- Takeaways

Tables are popular data representation

Invoices

Financial Reports

Contracts



Compact
Easy to understand*

Government Reports

Scientific Papers

Loan A

(*) For humans

End-to-end example

- What does the value 672 in the following table mean?
- **Answer:** Net earnings for three months ended July 29th, 2017 was $672 million USD

Table Extraction: Identify table location and structure

Table Understanding: Provide semantic context to table values

Steps:

1) Find location of table on page
2) Find cells in column containing ”672”
3) Find cells in row corresponding to “672”
4) Identify aligned row / column header cells
5) Normalize using additional context from table

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	286,817	(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	43,082	(5.8%)	0.59	0.66	(10.6%)
Employee profit share	32,974	19,039	73.2%	0.48	0.29	65.5%
Total operating expenses	**885,537**	**883,898**	**0.2%**	**12.83**	**13.60**	**(5.7%)**
Total, excluding fuel and profit share	**645,639**	**578,042**	**11.7%**	**9.35**	**8.90**	**5.1%**

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 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	377,178	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	285,304	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	45,388	102.9%	0.46	0.24	91.7%
Total operating expenses	**2,613,484**	**2,646,294**	**(1.2%)**	**12.83**	**13.78**	**(6.9%)**
Total, excluding fuel and profit share	**1,889,071**	**1,754,392**	**7.7%**	**9.27**	**9.13**	**1.5%**

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.

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	286,817	(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	43,082	(5.8%)	0.59	0.66	(10.6%)
Employee profit share	32,974	19,039	73.2%	0.48	0.29	65.5%
Total operating expenses	**885,537**	**883,898**	**0.2%**	**12.83**	**13.60**	**(5.7%)**
Total, excluding fuel and profit share	**645,639**	**578,042**	**11.7%**	**9.35**	**8.90**	**5.1%**

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 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	377,178	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	285,304	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	45,388	102.9%	0.46	0.24	91.7%
Total operating expenses	**2,613,484**	**2,646,294**	**(1.2%)**	**12.83**	**13.78**	**(6.9%)**
Total, excluding fuel and profit share	**1,889,071**	**1,754,392**	**7.7%**	**9.27**	**9.13**	**1.5%**

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
- …

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

Table Understanding: Problem Definition

	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	286,817	(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	43,082	(5.8%)	0.59	0.66	(10.6%)
Employee profit share	32,974	19,039	73.2%	0.48	0.29	65.5%
Total operating expenses	**885,537**	**883,898**	**0.2%**	**12.83**	**13.60**	**(5.7%)**
Total, excluding fuel and profit share	**645,639**	**578,042**	**11.7%**	**9.35**	**8.90**	**5.1%**

Input: 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

Output: Table content representation:

1) Captures semantic information
2) Amenable to post-processing

Table Understanding: Example

	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	286,817	(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	43,082	(5.8%)	0.59	0.66	(10.6%)
Employee profit share	32,974	19,039	73.2%	0.48	0.29	65.5%
Total operating expenses	**885,537**	**883,898**	**0.2%**	**12.83**	**13.60**	**(5.7%)**
Total, excluding fuel and profit share	**645,639**	**578,042**	**11.7%**	**9.35**	**8.90**	**5.1%**


```
{
  "tables": [
    {
      "column_headers": [
        {
          "cell_id": "colHeader-1050-1082",
          "text": "Expenses ($ in thousands)",
          ...
        },
        {
          "cell_id": "colHeader-1270-1301",
          "text": "Three months ended Sept. 30",
          ...
        },
        {
          "cell_id": "colHeader-1544-1548",
          "text": "2015"
        },
        ...
      ],
      "row_headers": [
        {
          "cell_id": "rowHeader-2244-2262",
          "text": "Aircraft fuel"
        },
        {
          "cell_id": "rowHeader-3197-3217",
          "text": "Airport operations"
        },
        {
          "cell_id": "rowHeader-4148-4176",
          "text": "Flight operations and navigational changes"
        },
        ...
      ],
      "body_cells": [
        {
          "cell_id": "bodyCell-2450-2455",
          "text": "206,924",
          "row_header_ids": [
            "rowHeader-2244-2262"
          ],
          "column_header_ids": [
            "colHeader-1050-1082",
            "colHeader-1270-1301",
            "colHeader-1544-1548"
          ],
        },
        {
          "cell_id": "bodyCell-5415-8945",
          "text": "142,176",
          "row_header_ids": [
            "rowHeader-3197-3217"
          ],
          "column_header_ids": [
            "colHeader-1050-1082",
            "colHeader-1270-1301",
            "colHeader-1544-1548"
          ],
```


...

Input: 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

Output: Table content representation:

1) Captures semantic information
2) Amenable to post-processing

Table Understanding: Example

	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	286,817	(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	43,082	(5.8%)	0.59	0.66	(10.6%)
Employee profit share	32,974	19,039	73.2%	0.48	0.29	65.5%
Total operating expenses	**885,537**	**883,898**	**0.2%**	**12.83**	**13.60**	**(5.7%)**
Total, excluding fuel and profit share	**645,639**	**578,042**	**11.7%**	**9.35**	**8.90**	**5.1%**

Value	Norm Value	Year	Time Period	Type	LineItem
206,924	$206,924,000	2015	Q3	Expense	Aircraft Fuel
286,817	$286,817,000	2014	Q3	Expense	Airport Fuel
142,176	$142,176,000	2015	Q3	Expense	Airport operations
...	...	...	...	...	...

Change	Change Normalized	Begin Time Period	End Time Period
(27.9%)	-27.9%	Q3 2014	Q3 2015
10.7%	10.7%	Q3 2014	Q3 2015
...	...	...	...

Input: 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

Output: Table content representation:

1) Captures semantic information
2) Amenable to post-processing

Outline

- Problem definition
  - Table Extraction
  - Table Understanding
- Challenges
  - Limited document format support for table structure
  - Table variety
- Takeaways

Challenge: Table structure representation varies across document formats

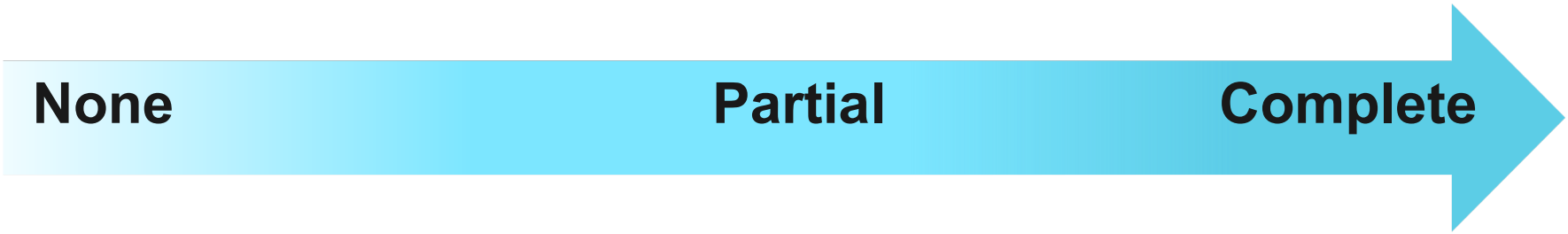


Table Understanding still required for all document types

- M. Cafarella et al. " WebTables: exploring the power of tables on the web". VLDB '08
- Z. Chen et al. "Spreadsheet Property Detection With Rule-assisted Active Learning". CIKM '17
- H. Dong et al. "TableSense: Spreadsheet Table Detection with Convolutional Neural Networks". AAAI '19

HTML completely represents table structure

None **Partial** **Complete**

HTML

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	286,817	(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	43,082	(5.8%)	0.59	0.66	(10.6%)
Employee profit share	32,974	19,039	73.2%	0.48	0.29	65.5%
Total operating expenses	**885,537**	**883,898**	**0.2%**	**12.83**	**13.60**	**(5.7%)**
Total, excluding fuel and profit share	**645,639**	**578,042**	**11.7%**	**9.35**	**8.90**	**5.1%**

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 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	377,178	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	285,304	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	45,388	102.9%	0.46	0.24	91.7%
Total operating expenses	**2,613,484**	**2,646,294**	**(1.2%)**	**12.83**	**13.78**	**(6.9%)**
Total, excluding fuel and profit share	**1,889,071**	**1,754,392**	**7.7%**	**9.27**	**9.13**	**1.5%**

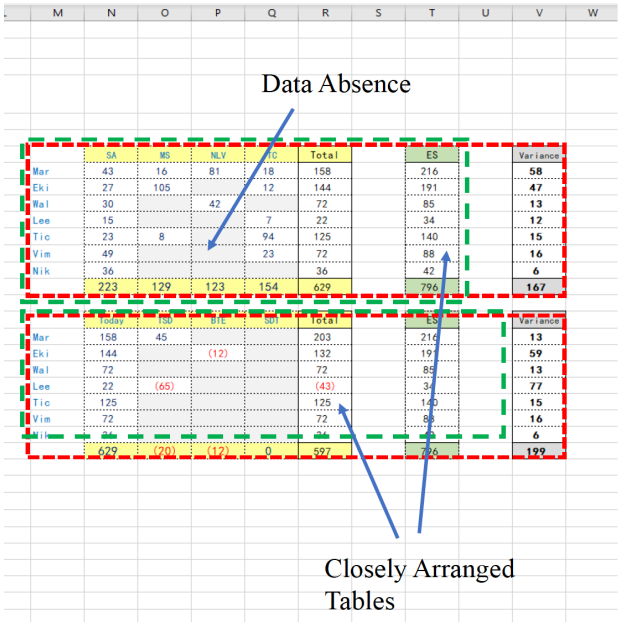
During the nine months ended September 30, 2015, operating expenses decreased by 1.2 per cent to $2,613.5 million as compared to $2,646.3 million in the same period in 2014, primarily driven by the year-over-year decrease in aircraft fuel expense, partially offset by the increase in our employee profit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

On an ASM basis, operating expenses for the nine months ended September 30, 2015 decreased by 6.9 per cent to 12.83 cents from 13.78 cents in the same period in 2014. Included in the first half of 2014 is a fuel value added tax recovery (VAT recovery) of $20.1 million, $17.6 million in aircraft fuel expense and $2.5 million in airport operations expense, related to the periods 2009 to 2013. Excluding the VAT recovery from the comparable period, CASM would have decreased by 7.6 per cent and CASM, excluding fuel and profit share would have increased by 1.3 per cent over the same period in 2014.

Table structure representation varies across Excel documents

None → **Partial** → **Complete**

MS Excel



Multiple tables defined in single sheet

Characteristic	2017 [1]			2018			Percent change* in real median income (2018 less 2017)	
	Number (thousands)	Median income (dollars)		Number (thousands)	Median income (dollars)			
		Estimate	Margin of error [2] (±)		Estimate	Margin of error [2] (±)	Estimate	Margin of error [2] (±)
HOUSEHOLDS								
All Households	127,669	62,626	542	128,579	63,179	691	0.9	1.06
Type of Household								
Family households	83,523	79,693	884	83,482	80,663	664	*1.2	1.14
Married-couple	61,869	93,556	863	61,959	93,654	1,125	0.1	1.16
Female householder, no spouse present	15,303	42,669	862	15,043	45,128	1,116	*5.8	3.00
Male householder, no spouse present	6,351	59,636	2,072	6,480	61,518	1,246	3.2	3.90
Nonfamily households	44,146	37,229	512	45,096	38,122	825	*2.4	2.38
Female householder	23,316	31,915	593	23,515	32,007	667	0.3	2.53
Male householder	20,830	43,843	1,680	21,582	45,754	868	*4.4	3.98
Race [3] and Hispanic Origin of Householder								
White	100,113	66,413	862	100,528	66,943	646	0.8	1.25
White, not Hispanic	84,706	69,851	1,136	84,727	70,642	652	1.1	1.57
Black	17,019	40,324	1,430	17,167	41,361	906	2.6	3.67
Asian	6,750	83,376	1,822	6,981	87,194	2,805	*4.6	3.66
Hispanic (any race)	17,336	51,389	776	17,758	51,450	735	0.1	1.83
Age of Householder								
Under 65 years	94,703	70,944	1,018	94,423	71,659	573	1.0	1.40
15 to 24 years	6,223	39,901	1,663	6,199	43,531	2,689	*9.1	8.00
25 to 34 years	20,258	62,732	852	20,611	65,890	1,075	*5.0	1.95
35 to 44 years	21,609	80,768	1,893	21,370	80,743	1,071	Z	2.47
45 to 54 years	22,566	82,111	1,365	22,071	84,464	1,845	*2.9	2.43
55 to 64 years	24,047	70,576	1,603	24,172	68,951	1,444	-2.3	2.95
65 years and older	32,966	42,303	808	34,156	43,696	816	*3.3	2.48

Each sheet separate table

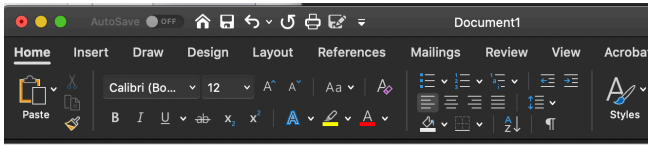
H. Dong et al. "TableSense: Spreadsheet Table Detection with Convolutional Neural Networks". AAAI '19

Z. Chen et al. “Spreadsheet Property Detection With Rule-assisted Active Learning”. CIKM ‘17

Table structure representation varies across Word documents

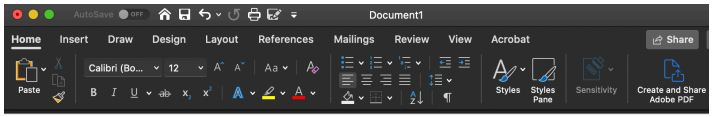
None → **Partial** → **Complete**

MS Word



NAME	GENDER	AGE	HEIGHT	WEIGHT	SMOKER
Alice	F	38	63	116	0
Bob	M	43	71	176	1
Carol	F	40	67	140	0
Dave	M	27	74	225	0

Omit Office Table Object



NAME	GENDER	AGE	HEIGHT	WEIGHT	SMOKER
Alice	F	38	63	116	0
Bob	M	43	71	176	1
Carol	F	40	67	140	0
Dave	M	27	74	225	1

Use Office Table Object for all tables

Document formats with no native table representation

None → **Partial** → **Complete**

Image

TXT

PDF

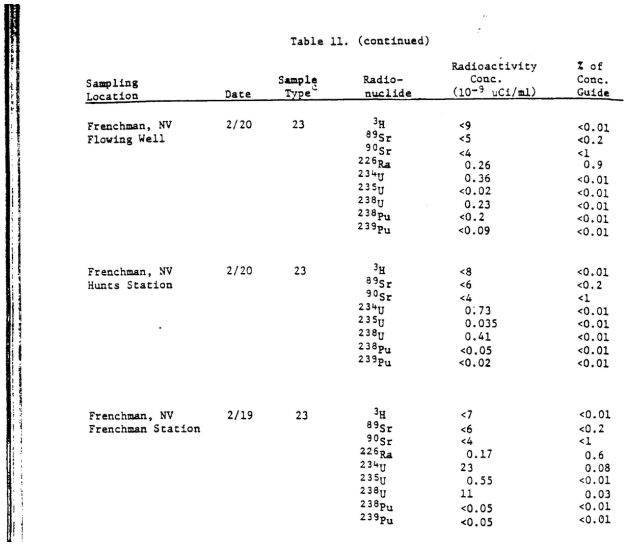


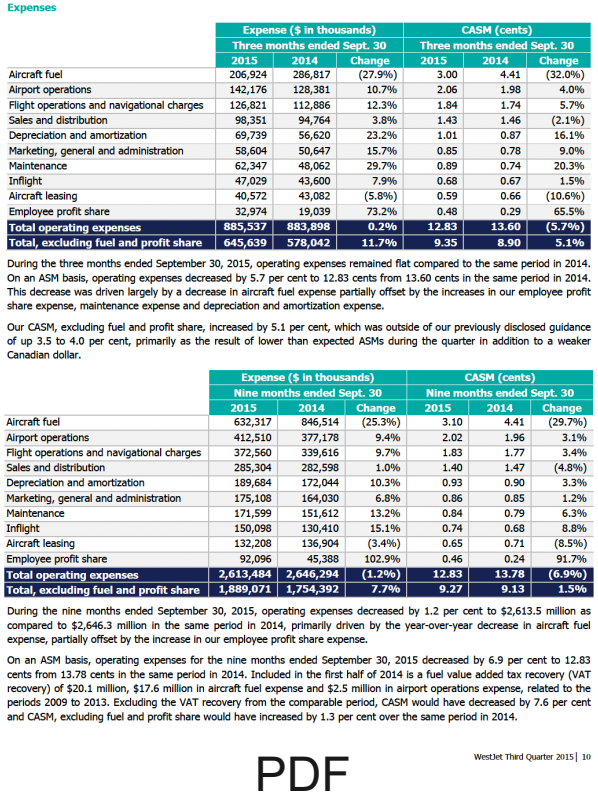
Table 11. (continued)

Sampling Location	Date	Sample Type	Radio-nuclide	Radioactivity Conc. (10^{-9} µCi/ml)	% of Conc. Guide
Frenchman, NV Flowing Well	2/20	23	^{3}H	<9	<0.01
			^{89}Sr	<5	<0.2
			^{90}Sr	<4	<1
			^{226}Ra	0.26	0.9
			^{234}U	0.36	<0.01
			^{235}U	<0.02	<0.01
			^{238}U	0.23	<0.01
			^{238}Pu	<0.2	<0.01
			^{239}Pu	<0.09	<0.01
Frenchman, NV Hunts Station	2/20	23	^{3}H	<8	<0.01
			^{89}Sr	<6	<0.2
			^{90}Sr	<4	<1
			^{234}U	0.73	<0.01
			^{235}U	0.035	<0.01
			^{238}U	0.41	<0.01
			^{238}Pu	<0.05	<0.01
			^{239}Pu	<0.02	<0.01
Frenchman, NV Frenchman Station	2/19	23	^{3}H	<7	<0.01
			^{89}Sr	<6	<0.2
			^{90}Sr	<4	<1
			^{226}Ra	0.17	0.6
			^{234}U	23	0.08
			^{235}U	0.55	<0.01
			^{238}U	11	0.03
			^{238}Pu	<0.05	<0.01
			^{239}Pu	<0.05	<0.01

Image

NAME	GENDER	AGE	HEIGHT	WEIGHT	SMOKER
Alice	F	38	63	116	0
Bob	M	43	71	176	1
Carol	F	40	67	140	0
Dave	M	27	74	225	0

TXT



Expenses

	Expense ($ in thousands) Three months ended Sept. 30			CASM (cents) Three months ended Sept. 30		
	2015	2014	Change	2015	2014	Change
Aircraft fuel	206,924	286,817	(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	43,082	(5.8%)	0.59	0.66	(10.6%)
Employee profit share	32,974	19,039	73.2%	0.48	0.29	65.5%
Total operating expenses	**885,537**	**883,898**	**0.2%**	**12.83**	**13.60**	**(5.7%)**
Total, excluding fuel and profit share	**645,639**	**578,042**	**11.7%**	**9.35**	**8.90**	**5.1%**

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 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) Nine months ended Sept. 30			CASM (cents) 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	377,178	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	285,304	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	45,388	102.9%	0.46	0.24	91.7%
Total operating expenses	**2,613,484**	**2,646,294**	**(1.2%)**	**12.83**	**13.78**	**(6.9%)**
Total, excluding fuel and profit share	**1,889,071**	**1,754,392**	**7.7%**	**9.27**	**9.13**	**1.5%**

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.

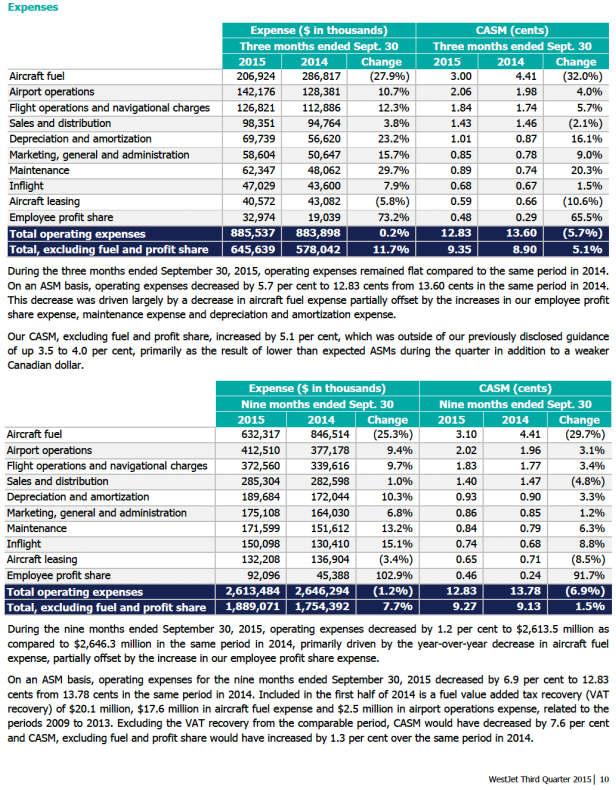
On an ASM basis, operating expenses for the nine months ended September 30, 2015 decreased by 6.9 per cent to 12.83 cents from 13.78 cents in the same period in 2014. Included in the first half of 2014 is a fuel value added tax recovery (VAT recovery) of $20.1 million, $17.6 million in aircraft fuel expense and $2.5 million in airport operations expense, related to the periods 2009 to 2013. Excluding the VAT recovery from the comparable period, CASM would have decreased by 7.6 per cent and CASM, excluding fuel and profit share would have increased by 1.3 per cent over the same period in 2014.

WestJet Third Quarter 2015 | 10

PDF

PDF Document Format

Rendered PDF



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	286,817	(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	43,082	(5.8%)	0.59	0.66	(10.6%)
Employee profit share	32,974	19,039	73.2%	0.48	0.29	65.5%
Total operating expenses	**885,537**	**883,898**	**0.2%**	**12.83**	**13.60**	**(5.7%)**
Total, excluding fuel and profit share	**645,639**	**578,042**	**11.7%**	**9.35**	**8.90**	**5.1%**

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 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	377,178	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	285,304	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	45,388	102.9%	0.46	0.24	91.7%
Total operating expenses	**2,613,484**	**2,646,294**	**(1.2%)**	**12.83**	**13.78**	**(6.9%)**
Total, excluding fuel and profit share	**1,889,071**	**1,754,392**	**7.7%**	**9.27**	**9.13**	**1.5%**

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.

On an ASM basis, operating expenses for the nine months ended September 30, 2015 decreased by 6.9 per cent to 12.83 cents from 13.78 cents in the same period in 2014. Included in the first half of 2014 is a fuel value added tax recovery (VAT recovery) of $20.1 million, $17.6 million in aircraft fuel expense and $2.5 million in airport operations expense, related to the periods 2009 to 2013. Excluding the VAT recovery from the comparable period, CASM would have decreased by 7.6 per cent and CASM, excluding fuel and profit share would have increased by 1.3 per cent over the same period in 2014.

WestJet Third Quarter 2015 | 10

PDF Binary

```
...
BT
  0.0503 Tc
  8.503556 0 0 8.52 503.2795 688.92 Tm
  /Tc2 1 Tf
  [ ( m) 16 (o) 21 (n) 17 (t) 39 (h) 16 (s) 29 (
) 28 (e) 28 (n) 17 (d) 24 (e) 28 (d) 24 ( ) ] TJ
  0 Tc
ET
```

Draw ”m” at (503, 688) in 8.5 point font in color white
Draw “o” at …
Draw “n” at ….
Draw “t” at ...
Draw “h” at …
Draw “s” at …
….

```
...
Q
q
 46.91952 776.52 m
 242.04 776.52 l
 242.04 729.96 l
 144.48 729.96 l
 46.91952 729.96 l
 h
...
```

Draw green line segment from (46, 776) to (242, 776)
Draw green line segment from (242, 776) to (242, 729)
…

- Programmatic PDF collection of instructions to draw characters and line segments to page with visual formatting information
- **2 – 4 trillion** PDFs in existence and rapidly growing

Challenge: Variety in Tables

Fig.7: Sensitivity to price and volume movements

	Net Profit	
For each +/-1% chg in coal price	FY14F	FY15F
Indo Tambang	-0.40%	-1.80%
Adaro	-0.50%	-1.40%
Harum	-1.00%	-1.80%
For each +/-1% chg in volume		
Indo Tambang	-1.00%	-1.10%
Adaro	-1.10%	-1.10%
Harum	-1.40%	-1.20%

Source: Various, AllianceDBS

Table with graphic lines

		Previously	*Now (under proposed new notes)*
-	FCCR (limitation on indebtedness)	3x	2.5x
-	Debt/EBITDA	4x	-
-	Bekasi Power debt basket	5mn	-
-	General debt basket	10mn	20mn
-	Permitted investment basket	-	7.5% of total assets
-	Interest reserve account establishment	Yes	No

Table with visual clues only

Rogers Communications Inc.
Business Solutions - Historical Trend
(unaudited)

(In millions of dollars, except margins)	2014 Total	Q3'14	Q2'14	Q1'14	2013 Total	Q4'13	Q3'13	Q2'13	Q1'13
Operating revenue									
Next generation	200	69	67	64	213	63	54	52	44
Legacy	82	26	27	29	149	34	38	37	40
Service revenue	282	95	94	93	362	97	92	89	84
Equipment sales	3	1	1	1	12	1	1	1	9
Operating revenue	285	96	95	94	374	98	93	90	93
Operating expenses	(197)	(64)	(67)	(66)	(268)	(69)	(64)	(65)	(70)
Adjusted operating profit	**88**	**32**	**28**	**28**	**106**	**29**	**29**	**25**	**23**
Adjusted operating profit margin	30.9%	33.3%	29.5%	29.8%	28.3%	29.6%	31.2%	27.8%	24.7%
Additions to property, plant and equipment	93	28	39	26	107	41	20	31	15

Complex tables – graphical lines can be misleading – is this 1, 2 or 3 tables ?

Multi-row, multi-column column headers

($ in thousands)	Three months ended September 30			Nine months ended September 30		
	2015	2014	Change	2015	2014	Change
Salaries and benefits	171,586	153,103	12.1%	526,255	468,468	12.3%
Employee share purchase plan	22,101	19,809	11.6%	64,274	57,953	10.9%
Employee profit share	32,974	19,039	73.2%	92,096	45,388	102.9%
Share-based payment plans	4,125	4,548	(9.3%)	12,875	14,310	(10.0%)
Total	**230,786**	**196,499**	**17.4%**	**695,500**	**586,119**	**18.7%**
Presentation on the Condensed Consolidated Statement of Earnings:						
Airport operations	30,916	28,145	9.8%	88,292	81,055	8.9%
Flight operations and navigational charges	67,549	59,891	12.8%	206,913	184,607	12.1%
Sales and distribution	19,436	17,512	11.0%	55,928	53,358	4.8%
Marketing, general and administration	27,058	22,818	18.6%	85,419	73,941	15.5%
Inflight	36,116	33,838	6.7%	116,294	101,921	14.1%
Maintenance	16,737	15,256	9.7%	50,558	45,849	10.3%
Employee profit share	32,974	19,039	73.2%	92,096	45,388	102.9%
Total	**230,786**	**196,499**	**17.4%**	**695,500**	**586,119**	**18.7%**

Nested row headers

Example: claim 2 – Progressive Condition

18 months after Michael's initial diagnosis of early stage melanoma, despite treatment, it has recurred and has been detected at a higher stage.

It now meets the defined criteria for benefit category D under the *Health Event* category for solid tumour cancers (see the Cancer body system on page 44).

As this recurrence of the melanoma is a *Progressive Condition* to the previous claim, we will pay the difference between the percentage payable for benefit category of the current claim and that of the previous claim.

In this case, the current benefit category of D provides a benefit of 20%, whilst the previous claim under benefit category E provided a benefit of 5%, therefore the amount payable is 15% of the *Initial Amount of Cover*, which is $150,000. This is the amount that will be paid, as it is not greater than the *Maximum Amount Payable* for benefit category D ($200,000). For this claim, an amount of $150,000 is paid to Michael.

Following this claim, the *Remaining Amount of Cover* under the policy is reduced, resulting in the *Maximum Amount Payable* for each benefit category as shown below.

Benefit category	Type of cover	*Maximum Amount Payable*
AA	Death and *terminal illness*	$1,300,000
A	*Health Events*	$800,000
B	*Health Events*	$650,000
C	*Health Events*	$400,000
D	*Health Events*	$200,000
E	*Health Events*	$50,000

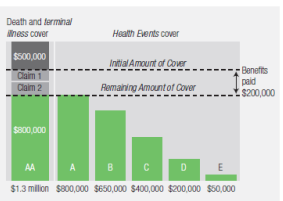


Table interleaved with text and charts

Event	Description	Comprehensive	Third Party Property Damage, Fire and Theft	Third Party Property Damage
Storm or flood (continued)	• you – had entered into a contract of sale to purchase your vehicle, or a contract to lease your vehicle, and – took out your insurance with us for the vehicle prior to taking possession of the vehicle.	✓	✗	✗
Theft or attempted theft	We will cover your vehicle for loss or damage as a result of theft or attempted theft. We also cover theft of the keys to your vehicle and certain items in your car, refer to policy benefits on pages 12-20. The theft or attempted theft must be reported to the Police.	✓	✓	✗

Tables with Textual content

Outline

- Problem definition
  - Table Extraction
  - Table Understanding
- Challenges
  - Limited document format support for table structure
  - Table variety
- Takeaways

Decades of work across multiple communities

- Information Management Community
  - Focus on “(semi-)structured” setting
    - Usually given table border and structure (e.g., HTML tables)
  - S
  - T

Goal: Synthesize work across these communities for both table extraction and understanding in context of complex business and scientific documents

- Doc
  - F
    t
  - Interpret values in individual tables by identifying table headers and linking with cell values

- Others (AI, NLP)
  - Leverage object detection (e.g., YOLO) to detect tables in images of document pages
  - Multi-modal (NLP approaches augmented with position / format information)
  - Focus on complex documents

Great Deal of Recent Interest in Table Processing

- Academic:
  - VLDB 2018: 10-year Test of Time for WebTables
  - 16 papers on table extraction alone at ICDAR 2019

- Commercial: Table processing key component of many recent commercial offerings
  - Amazon Textract / Kendra
  - Google Document AI
  - Microsoft Form Recognizer
  - IBM Watson Discovery

M. Cafarella et al. *"Exploring the Power of Tables on the Web"*. VLDB '08

M. Cafarella et al. *"Ten Years of WebTables"*. VLDB '18

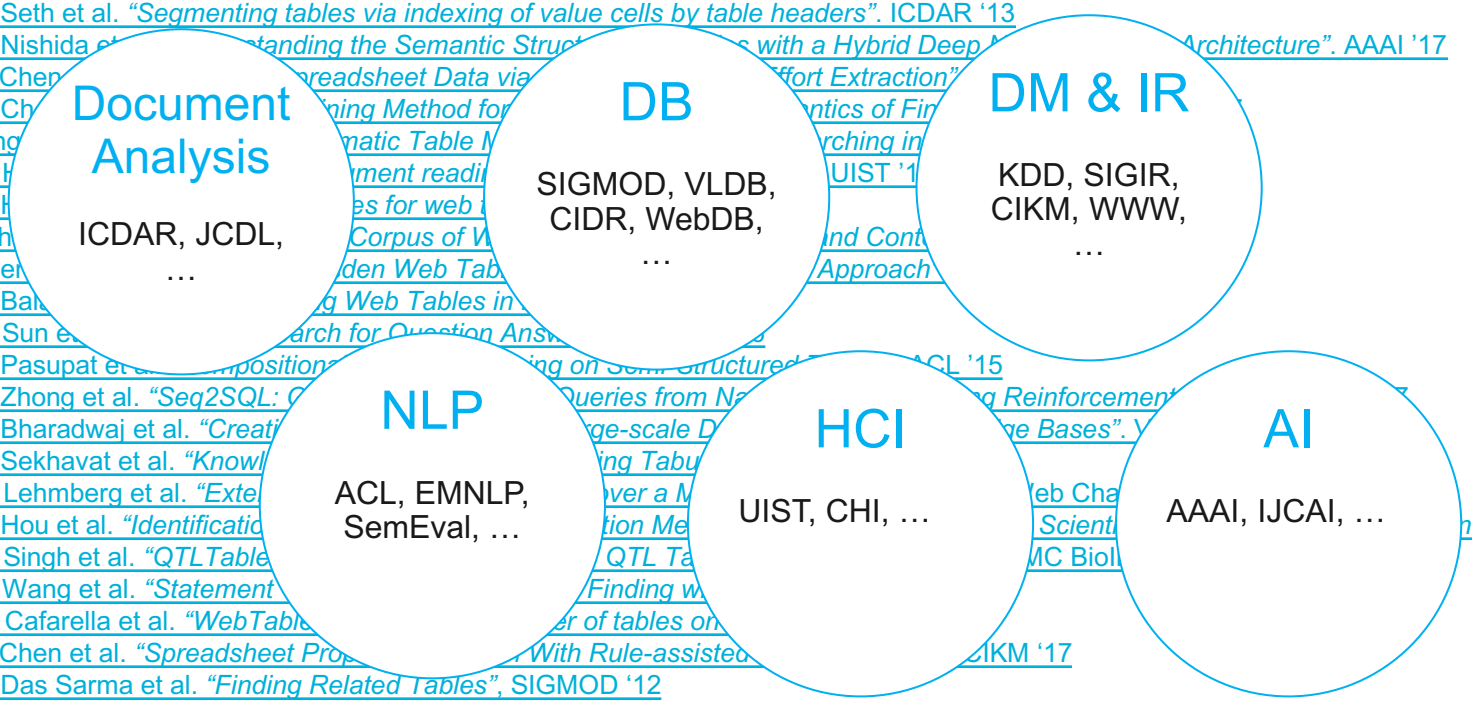
Table Extraction: A Sample of Prior Work

[CK93] S. Chandran and R. Kasturi. “Structural Recognition of Tabulated Data”, ICDAR ‘93
[I93] K. Itonori. “Table Structure Recognition Based on Textblock Arrangement and Ruled Line Position”, ICDAR ‘93
[H95] J. Ha et al. “Recursive X-Y Cut Using Bounding Boxes of Connected Components”, ICDAR ‘95
[KD98] T. Kieninger and A. Dengel. “The T-Recs Table Recognition and Analysis System”, DAS ‘98
[H99] J. Hu et al. “Medium-Independent Table Detection”, SPIE Doc. Recog. & Retr. ‘99
[H00a] J. C. Handley. “Table Analysis for Multi-line Cell Identification”, SPIE Doc. Recog. & Retr. ‘00
[H00b] J. Hu et al. “Table Structure Recognition and Its Evaluation”, SPIE Doc. Recog. & Retr. ‘00
[K01] B. Klein et al. “Three Approaches to Industrial Table Spotting”, ICDAR ’01
[KD01] T. Kieninger and A. Dengel. “Applying the T-Recs Table Recognition System to the Business Letter Domain”, ICDAR ‘01
[C02] F. Cesarini et al. “Trainable Table Location in Document Images”, ICPR ‘02
[P03] D. Pinto et al. “Table Extraction Using Conditional Random Fields”, SIGIR ‘03
[H03] M. Hurst. “A Constraint-based Approach to Table Structure Derivation”, ICDAR ‘03
[W04] Y. Wang et al. “Table Structure Understanding and Its Performance Evaluation”, Pattern Recog. ‘04
[Y05] B. Yildiz et al. “pdf2table: A Method to Extract Table Information from PDF Files”, IICAI ‘05
[S06] A. C. e Silva et al. “Design of an End-to-end Method to Extract Information from Tables”, IJDAR ‘06
[M06] S. Mandal et al. “A Simple and Effective Table Detection System from Document Images”, IJDAR ‘06
[L07] Y. Liu et al. “TableSeer: Automatic Table Metadata Extraction and Searching in Digital Libraries”, JCDL ‘07
[HB07] T. Hassan and R. Baumgartner. “Table Recognition and Understanding from PDF Files”, ICDAR ‘07
[L08] Y. Liu et al. “Identifying Table Boundaries in Digital Documents via Sparse Line Detection”, CIKM ’08
[L09] Y. Liu et al. “Improving the Table Boundary Detection in PDFs by Fixing the Sequence Error of the Sparse Lines”, ICDAR ‘09
[OR09] E. Oro and M. Ruffolo. “PDF-TREX: An Approach for Recognizing and Extracting Tables from PDF Documents”, ICDAR ‘09
[SS10] F. Shafait and R. Smith. “Table Detection in Heterogeneous Documents”, DAS ‘10
[D11] F. Deckert et al. “Table Content Understanding in smartFIX”, ICDAR ‘11
[F11] J. Fang et al. “A Table Detection Method for Multipage PDF Documents via Visual Separators and Tabular Structures”, ICDAR ‘11
[B12] E. Bart. “Parsing Tables by Probabilistic Modeling of Perceptual Cues”, DAS ‘12
[CL12] J. Chen and D. Lopresti. “Model-Based Tabular Structure Detection and Recognition in Noisy Handwritten Documents”, ICFHR ‘12
[G12] Göbel et al. “A Methodology for Evaluating Algorithms for Table Understanding in PDF Documents”. DocEng '12
[K13] T. Kasar et al. “Learning to Detect Tables in Scanned Document Images Using Line Information”, ICDAR ‘13
[K14] S. Klampfl et al. “A Comparison of Two Unsupervised Table Recognition Methods from Digital Scientific Articles”, D-Lib Mag. ‘14
[B14] A. Bansal et al. “Table Extraction from Document Images using Fixed Point Model”, ICVGIP ‘14
[R15] F
[FK15]
[T16] T
[G17] A
[S17] S
[S18a]
[S18b]
[Q19] S
[C19] Z
[G19] Ga
[K19] Khan et al. “Table Structure Extraction with Bi-directional Gated Recurrent Unit Networks“ ICDAR 2019
[L19] M. Li et al. “TableBank: Table Benchmark for Image-based Table Detection and Recognition”, LREC, 2020
[M19] S. Mujumdar et al. “Simultaneous Optimisation of Image Quality Improvement and Text Content Extraction from Scanned Documents”, ICDAR ’19
[HL19] J. Hoffswell and Z. Liu. “Interactive Repair of Tables Extracted from PDF Documents on Mobile Devices”, CHI ’19
[T19] Tensmeyer et al. “Deep splitting and merging for table structure decomposition” ICDAR 2019
[L20] Li et al. “TableBank: Table Benchmark for Image-based Table Detection and Recognition”. LREC ’20
[P20a] D. Prasad et al. “CascadeTabNet: An Approach for End to End Table Detection and Structure Recognition from Image-Based Documents”, In *CVPR Workshops* ’20
[P20b] Paliwal et al. “TableNet: Deep Learning Model for End-to-end Table Detection and Tabular Data Extraction from Scanned Document Images”, arXiv 2020
[V20] Le Vine et al. Identifying Table Structure in Documents using Conditional Generative Adversarial Networks, arXiv 2020
[Z20a] Zheng et al. “Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context”, arXiv 2020
[Z20b] Zhong et al. “Image-based table recognition: data, model, and evaluation”, ECCV 2020

🔥 ICDAR 2019 has ≥ **16** new papers on table extraction!

- ICDAR = International Conference on Document Analysis and Recognition

Table Understanding & Applications: A Sample of Prior Work



J. Fang et al. *"Table Header Detection and Classification"*. AAAI '12
S. Seth et al. *"Segmenting tables via indexing of value cells by table headers"*. ICDAR '13
D. Pinto et al. *"Table Extraction Using Conditional Random Fields"*. SIGIR '03
R. Rastan et al. *"TEXUS: A unified framework for extracting and understanding tables in PDF documents"*. Information Processing & Management
S. Seth et al. *"Segmenting tables via indexing of value cells by table headers"*. ICDAR '13
K. Nishida et ... *standing the Semantic Struc... with a Hybrid Deep ... Architecture"*. AAAI '17
Z. Chen ... *preadsheet Data via ... ffort Extraction"*
X. Ch... *ning Method for ... ntics of Fin...*
Ying ... *matic Table M... rching in...*
D. H... *ment readi... UIST '1...*
B. H... *es for web ...*
Leh... *Corpus of W... nd Cont...*
Eber... *den Web Tab... Approach*
S. Bal... *g Web Tables in ...*
H. Sun e... *arch for Question Answ...*
P. Pasupat et ... *mpositio... ing on Semi-Structured ... ACL '15*
V. Zhong et al. *"Seq2SQL: ... Queries from Na... g Reinforcement ...*
S. Bharadwaj et al. *"Creati... ge-scale D... ge Bases"*. V...
Y. Sekhavat et al. *"Knowl... ing Tabu...*
O. Lehmberg et al. *"Exte... ver a M... eb Cha...*
Y. Hou et al. *"Identificatio... tion Me... Scienti... n"*. ACL '19
G. Singh et al. *"QTLTable... QTL Ta... MC Biol...*
N. Wang et al. *"Statement ... Finding w...*
M. Cafarella et al. *"WebTabl... er of tables on...*
Z. Chen et al. *"Spreadsheet Prop... With Rule-assisted ... CIKM '17*
A. Das Sarma et al. *"Finding Related Tables"*, SIGMOD '12
H. Dong et al. *"TableSense: Spreadsheet Table Detection with Convolutional Neural Networks"*. AAAI '19
P. Yin et al. "Neural Enquirer: Learning to Query Tables in Natural Language", IJCAI '16
Z. Yan...
E. Cre...
M. Ca...
Y. Wa...
J. Hoffswell et al. "Interactive Repair of Tables Extracted from PDF Documents on Mobile Devices". CHI '19
M. Yakout et al. *"InfoGather: Entity Augmentation and Attribute Discovery by Holistic Matching with Web Tables"*, SIGMOD '12
T. Müller et al. *"Answering Conversational Questions on Structured Data without Logical Forms"*, EMNLP '19
J. Herzig et al. *"TAPAS: Weakly Supervised Table Parsing via Pre-training"*, ACL 2020

Document Analysis
ICDAR, JCDL, …

DB
SIGMOD, VLDB, CIDR, WebDB, …

DM & IR
KDD, SIGIR, CIKM, WWW, …

NLP
ACL, EMNLP, SemEval, …

HCI
UIST, CHI, …

AI
AAAI, IJCAI, …

🔥 Substantial work distributed across multiple sub-communities

Takeaways

- Widely used document formats have limited table representation
  - Limits of document format: Image, PDF
  - How documents authored: Word, Excel
- Wide variety of tables makes general model construction difficult
  - Tables are form of art
  - Diverse visual encoding of semantic information
  - Different domains
- Once tables are identified and interpretable, many downstream applications can be enabled