Table Extraction

What Is Table Extraction?

- Table region detection
  - Identify all tables
  - Separate tables from non-table text
  - Separate tables from each other
- Cell structure recognition
  - Partition text into cells
  - Define rows and columns
  - Find cell span and cell-to-cell overlap (along X- or Y-axis)

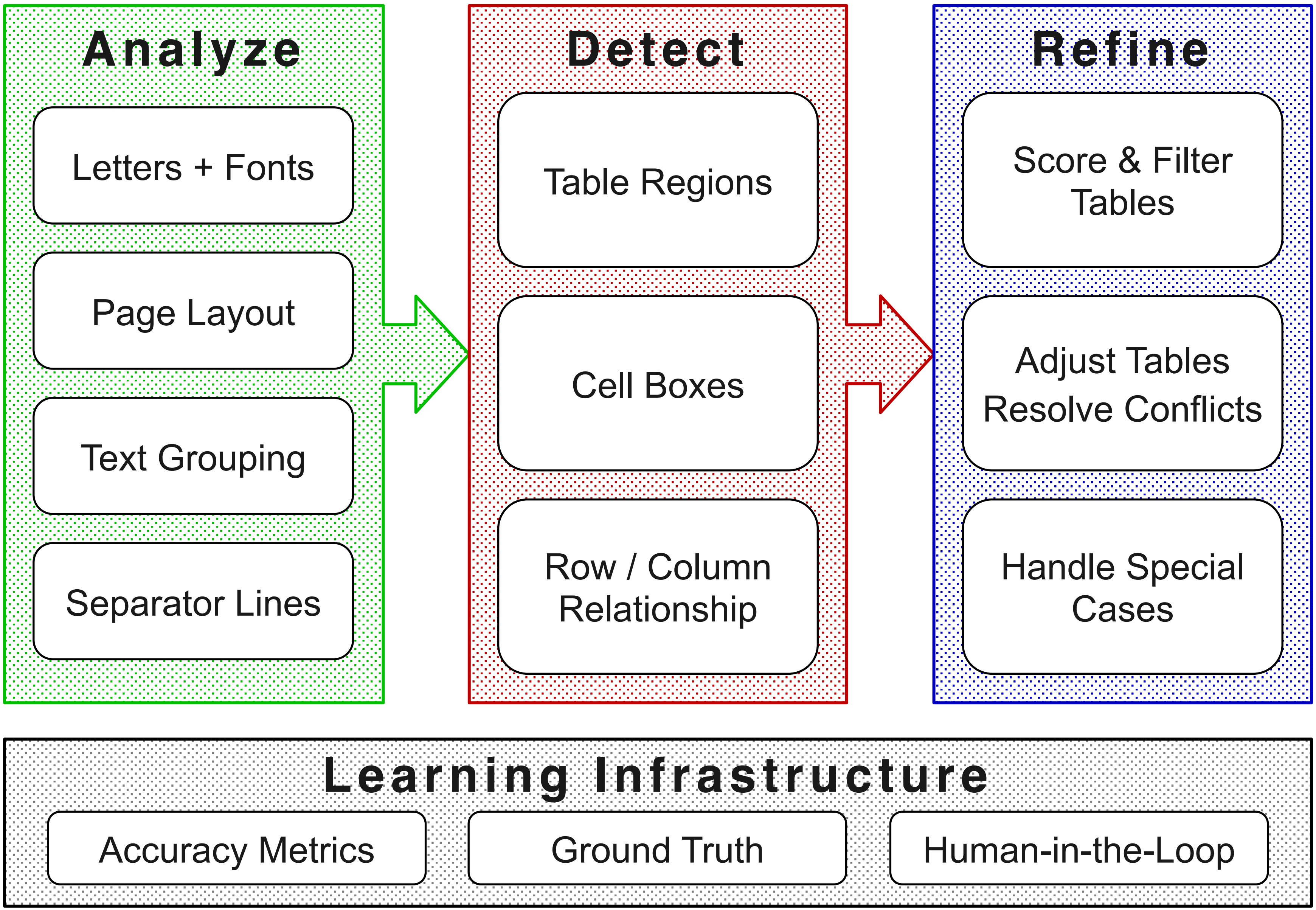
Table Extraction Timeline

- **Early 1990s : Separator based “top-down” methods**
  - Ruled line tables
  - Extend to white-space “lines”
- **1990s – early 2000s : “Bottom-up” text clustering**
  - Group text into columns (or rows), then to tables
  - Use space features (gaps, overlap, alignment) and keywords
- **2000s – early 2010s : Machine Learning (supervised or not)**
  - Classify text-rows using CRF, SVM, HMM, etc.
  - Probabilistic models for tables
  - Graph-based models for cell structure
- **Late 2010s : Deep Learning**
  - Scanned image table detection with R-CNN, YOLO, RetinaNet
  - Graph neural networks for cell structure
  - Natural language embeddings for text linkage

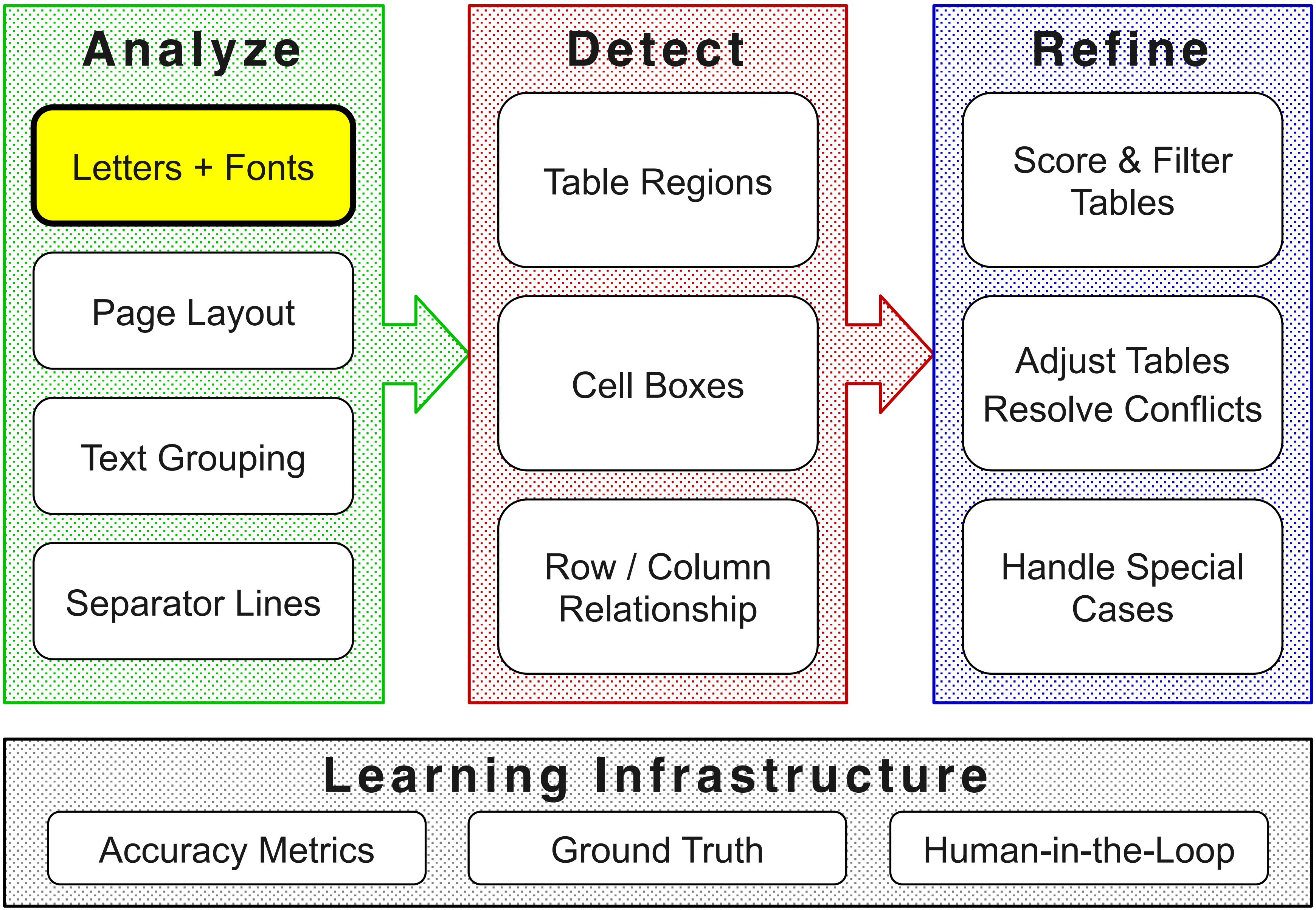
How to Build a Table Extraction System?

- Analyze Page
  - Read symbols & lines
  - Identify low-level structures & relations
  - Take shortcuts
- The Main Tasks
  - Table (region) detection
  - Cell structure recognition (given table region)
- Refine Tables
  - Discard false positives
  - Adjust table border and structure
  - Customer specific rules

Common Sub-Tasks in Table Extraction



Common Sub-Tasks in Table Extraction



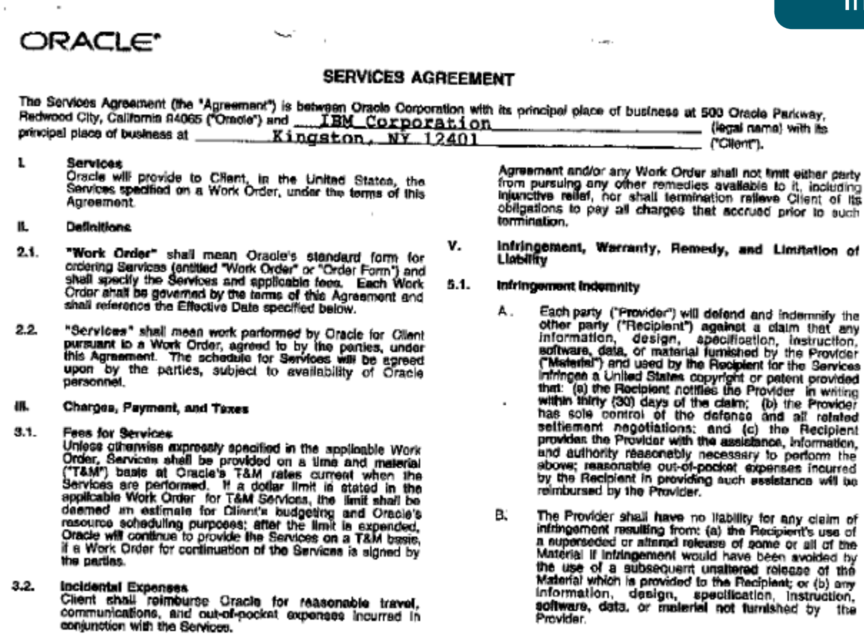
Character Features

- Documents can be:
  - **scanned**
  - **programmatic** (“born digital” PDF, TXT)
  - **hybrid**
- **Scanned** pages are noisy:
  - Reverse any rotation, distortion
  - Filter noise, sharpen if low resolution [M19]
- Augment **OCR** output:
  - Fix inconsistent fonts, bounding boxes, highlighted text
  - Detect ruled lines and boxes
    - E.g., Gaussian filter + black hat transform [K13]

[K13] T. Kasar et al. “Learning to Detect Tables in Scanned Document Images Using Line Information”, ICDAR ‘13
[M19] S. Mujumdar et al. “Simultaneous Optimisation of Image Quality Improvement and Text Content Extraction from Scanned Documents”, ICDAR ‘19

Character Features

Tilt
Fuzzy Text


ORACLE®

SERVICES AGREEMENT

The Services Agreement (the "Agreement") is between Oracle Corporation with its principal place of business at 500 Oracle Parkway, Redwood City, California 94065 ("Oracle") and IBM Corporation (legal name) with its principal place of business at Kingston, NY 12401 ("Client").

I. **Services**
Oracle will provide to Client, in the United States, the Services specified on a Work Order, under the terms of this Agreement.

II. **Definitions**

2.1. **"Work Order"** shall mean Oracle's standard form for ordering Services (entitled "Work Order" or "Order Form") and shall specify the Services and applicable fees. Each Work Order shall be governed by the terms of this Agreement and shall reference the Effective Date specified below.

2.2. "Services" shall mean work performed by Oracle for Client pursuant to a Work Order, agreed to by the parties, under this Agreement. The schedule for Services will be agreed upon by the parties, subject to availability of Oracle personnel.

III. **Charges, Payment, and Taxes**

3.1. **Fees for Services**
Unless otherwise expressly specified in the applicable Work Order, Services shall be provided on a time and material ("T&M") basis at Oracle's T&M rates current when the Services are performed. If a dollar limit is stated in the applicable Work Order for T&M Services, the limit shall be deemed an estimate for Client's budgeting and Oracle's resource scheduling purposes; after the limit is expended, Oracle will continue to provide the Services on a T&M basis, if a Work Order for continuation of the Services is signed by the parties.

3.2. **Incidental Expenses**
Client shall reimburse Oracle for reasonable travel, communications, and out-of-pocket expenses incurred in conjunction with the Services.

Agreement and/or any Work Order shall not limit either party from pursuing any other remedies available to it, including injunctive relief, nor shall termination relieve Client of its obligations to pay all charges that accrued prior to such termination.

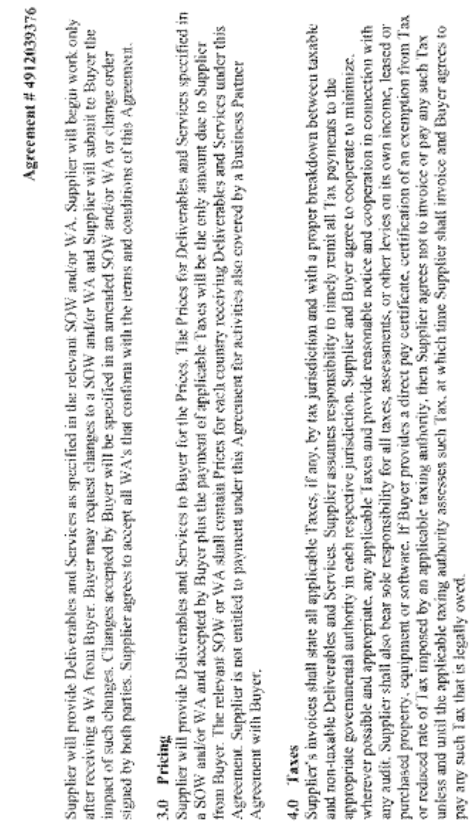
V. **Infringement, Warranty, Remedy, and Limitation of Liability**

5.1. **Infringement Indemnity**

A. Each party ("Provider") will defend and indemnify the other party ("Recipient") against a claim that any information, design, specification, instruction, software, data, or material furnished by the Provider ("Material") and used by the Recipient for the Services infringes a United States copyright or patent provided that: (a) the Recipient notifies the Provider in writing within thirty (30) days of the claim; (b) the Provider has sole control of the defense and all related settlement negotiations; and (c) the Recipient provides the Provider with the assistance, information, and authority reasonably necessary to perform the above; reasonable out-of-pocket expenses incurred by the Recipient in providing such assistance will be reimbursed by the Provider.

B. The Provider shall have no liability for any claim of infringement resulting from: (a) the Recipient's use of a superseded or altered release of some or all of the Material if infringement would have been avoided by the use of a subsequent unaltered release of the Material which is provided to the Recipient; or (b) any information, design, specification, instruction, software, data, or material not furnished by the Provider.

Rotated
Image


Technical Services Agreement

Agreement # 4912039376

Supplier will provide Deliverables and Services as specified in the relevant SOW and/or WA. Supplier will begin work only after receiving a WA from Buyer. Buyer may request changes to a SOW and/or WA and Supplier will submit to Buyer the impact of such changes. Changes accepted by Buyer will be specified in an amended SOW and/or WA or change order signed by both parties. Supplier agrees to accept all WA's that conform with the terms and conditions of this Agreement.

3.0 Pricing
Supplier will provide Deliverables and Services to Buyer for the Prices. The Prices for Deliverables and Services specified in a SOW and/or WA and accepted by Buyer plus the payment of applicable Taxes will be the only amount due to Supplier from Buyer. The relevant SOW or WA shall contain Prices for each country receiving Deliverables and Services under this Agreement. Supplier is not entitled to payment under this Agreement for activities also covered by a Business Partner Agreement with Buyer.

4.0 Taxes
Supplier's invoices shall state all applicable Taxes, if any, by tax jurisdiction and with a proper breakdown between taxable and non-taxable Deliverables and Services. Supplier assumes responsibility to timely remit all Tax payments to the appropriate governmental authority in each respective jurisdiction. Supplier and Buyer agree to cooperate to minimize, wherever possible and appropriate, any applicable Taxes and provide reasonable notice and cooperation in connection with any audit. Supplier shall also bear sole responsibility for all taxes, assessments, or other levies on its own income, leased or purchased property, equipment or software. If Buyer provides a direct pay certificate, certification of an exemption from Tax or reduced rate of Tax imposed by an applicable taxing authority, then Supplier agrees not to invoice or pay any such Tax unless and until the applicable taxing authority assesses such Tax, at which time Supplier shall invoice and Buyer agrees to pay any such Tax that is legally owed.

Per capita poultry consumption

Country	Chicken consumption (kg/capita/year)	GDP/capita[1] (US$)	Population (m)
Malaysia	37.3	10,060	28.6
Singapore	36.2	49,945	5.2
Thailand	12.5	5,070	68.2
China	9.2	5,460	1,321.0
Philippines	8.4	2,210	101.8
Vietnam	7.2	1,380	88.6
Indonesia	6.1	3,448	245.6
India	2.3	1,540	1,202.0

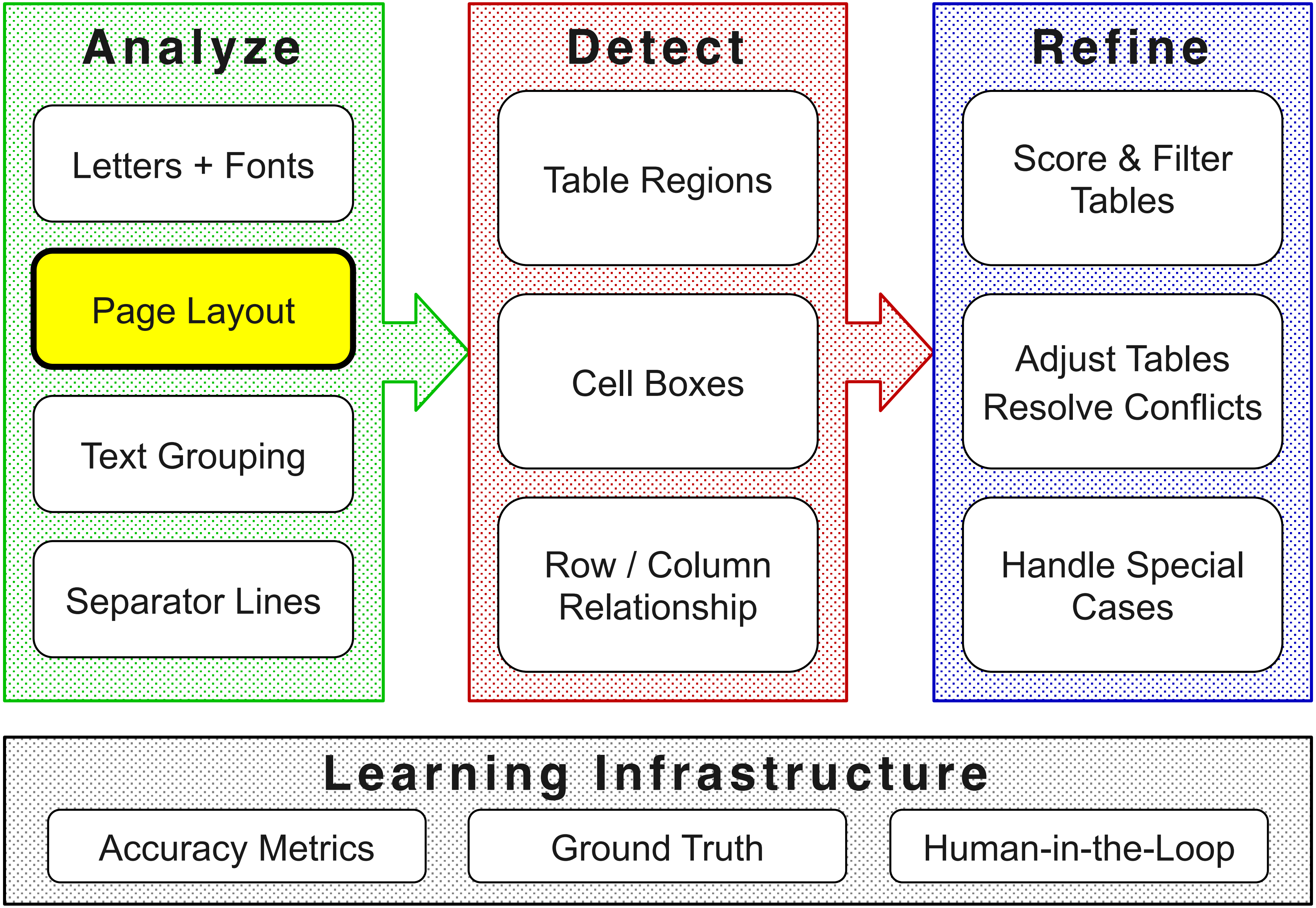
Potential upside for chicken consumption vs "matured" Malaysian market

Colored Background
Unclear Font Sizes

Character Features

- **Programmatic PDFs** (and TXTs)
  - Have letters, but **no** table markup
- May contain **spurious** (invisible) text and lines
  - White-on-white lines or text
  - Occluded or out-of-range lines or text
  - Text repeated to simulate bold font
  - **Need to filter them out**
- Deep Learning (CNN-based) methods need an image
  - Convert programmatic to scanned

Common Sub-Tasks in Table Extraction



Layout Analysis

- Plain text layout (1-column, 2-column, etc.)
  - Helps avoid false-positive “tables”
- Obvious non-tables
  - Page headers, footers, margins, numbering
  - Section headers
  - Lists, charts, highlighting
- Low-level structure
  - **Alignment** @ different box positions & tolerance levels
  - A minimum spanning tree for clustering by distance
- Deep learning features
  - Natural language embeddings

Table Extraction

Text Alignment

Tab-Stops

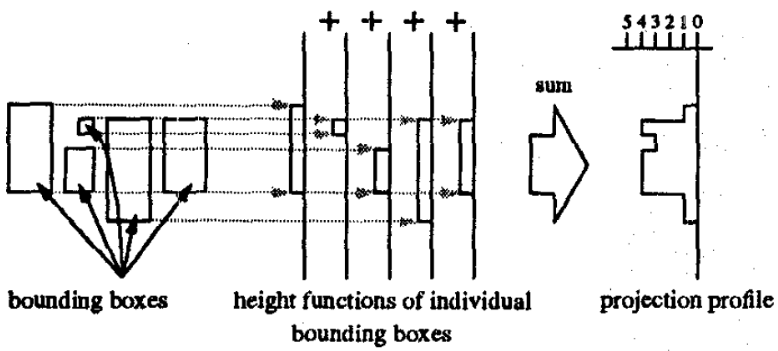
($000s)	2014 Q4	2014 Q3	2014 Q2	2014 Q1	2013 Q4	2013 Q3	2013 Q2	2013 Q1
Cash flow from operating activities	**80,866**	78,006	67,280	59,781	65,932	61,756	60,835	45,733
Change in non-cash working capital	**(18,865)**	(996)	5,452	8,923	(11,758)	(266)	1,958	1,949
Abandonment costs	**6,177**	3,189	697	1,346	1,760	814	434	962
Funds flow from operations	**68,178**	80,199	73,429	70,050	55,934	62,304	63,227	48,644
Weighted average outstanding shares *(000s)*								
- Basic	**193,497**	176,318	134,291	125,730	125,629	125,620	125,620	125,620
- Diluted	**193,497**	177,003	135,437	126,129	126,245	125,620	125,620	125,620
Funds flow from operations per share *($/share)*								
- Basic	**0.35**	0.45	0.55	0.56	0.45	0.50	0.50	0.39
- Diluted	**0.35**	0.45	0.54	0.56	0.44	0.50	0.50	0.39

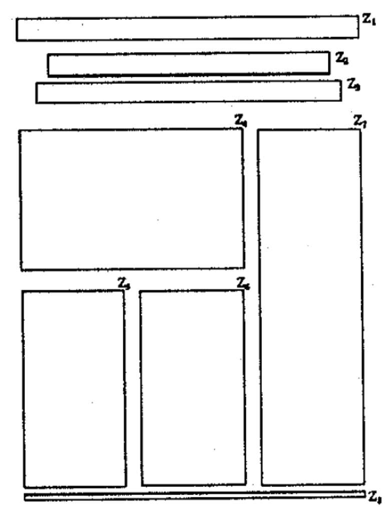
($000s)	**2014**	2013
Cash flow from operating activities	**285,933**	234,256
Change in non-cash working capital	**(5,486)**	(8,117)
Abandonment costs	**11,409**	3,970
Funds flow from operations	**291,856**	230,109
Weighted average outstanding shares *(000s)*		
- Basic	**157,697**	125,622
- Diluted	**157,697**	125,778

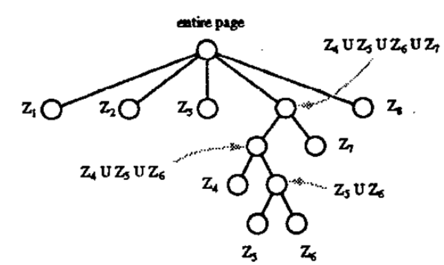
Table Source: http://iq.iradesso.ca/main/components/clients_profiles/55/financial_reports/LRE-2014-YearEnd-Combined.pdf

Recursive X-Y Cut Algorithm

- Commonly used to partition page and generate separators
  - By [C02], [W04], [K14], and others
- [H95] The algorithm recursively, for each block:
  - Computes X- and Y-axis projection profiles
  - Divides the block into sub-blocks based on dips in profiles:





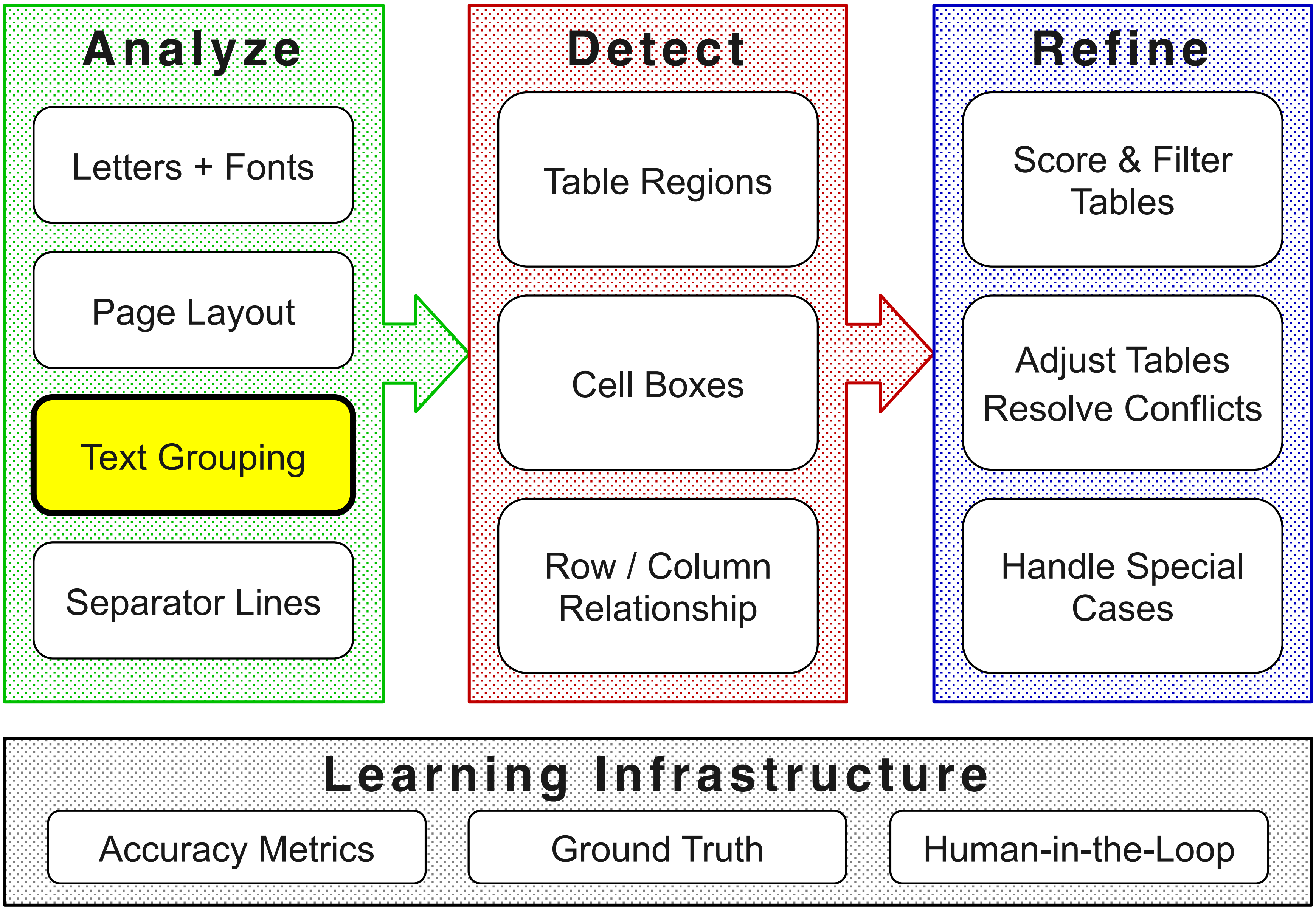


[H95] J. Ha et al. “Recursive X-Y Cut Using Bounding Boxes of Connected Components”, ICDAR ‘95
[C02] F. Cesarini et al. “Trainable Table Location in Document Images”, ICPR ‘02
[W04] Y. Wang et al. “Table Structure Understanding and Its Performance Evaluation”, Pattern Recog. ‘04
[K14] S. Klampfl et al. “A Comparison of Two Unsupervised Table Recognition Methods from Digital Scientific Articles”, D-Lib Mag. ‘14

Short-Cuts

- No tables $\Longrightarrow$ **take a short-cut**
  - Pre-trained CNNs can be slow
  - Most pages have no tables $\Longrightarrow$ major time savings
- Detect obvious non-tables
  - Solid plain text, 1- or 2-column layout
  - Frames, lists, header / footer, comments on margins
- Detect “easy” tables quickly
  - Ruling lines only tables
  - One-line-per-row aligned numerical tables
- No other structures $\Longrightarrow$ **take a short-cut**

Common Sub-Tasks in Table Extraction



Group Text into Larger Units

- Most systems group text early on
  - Table **detection** systems *may* skip text grouping
- Text is grouped in one of 3 ways:
  - Columns first
  - Rows first
  - “Blobs” or “paragraphs” first
- Some systems partition text using separator lines
  - **BUT:** “Blob” detection reduces over- / under-partitioning

Example

Two Tables

($000s)	2014 Q4	2014 Q3	2014 Q2	2014 Q1	2013 Q4	2013 Q3	2013 Q2	2013 Q1
Cash flow from operating activities	**80,866**	78,006	67,280	59,781	65,932	61,756	60,835	45,733
Change in non-cash working capital	**(18,865)**	(996)	5,452	8,923	(11,758)	(266)	1,958	1,949
Abandonment costs	**6,177**	3,189	697	1,346	1,760	814	434	962
Funds flow from operations	**68,178**	80,199	73,429	70,050	55,934	62,304	63,227	48,644
Weighted average outstanding shares *(000s)*								
- Basic	**193,497**	176,318	134,291	125,730	125,629	125,620	125,620	125,620
- Diluted	**193,497**	177,003	135,437	126,129	126,245	125,620	125,620	125,620
Funds flow from operations per share *($/share)*								
- Basic	**0.35**	0.45	0.55	0.56	0.45	0.50	0.50	0.39
- Diluted	**0.35**	0.45	0.54	0.56	0.44	0.50	0.50	0.39

($000s)	**2014**	2013
Cash flow from operating activities	**285,933**	234,256
Change in non-cash working capital	**(5,486)**	(8,117)
Abandonment costs	**11,409**	3,970
Funds flow from operations	**291,856**	230,109
Weighted average outstanding shares *(000s)*		
- Basic	**157,697**	125,622
- Diluted	**157,697**	125,778

Table Source: http://iq.iradesso.ca/main/components/clients_profiles/55/financial_reports/LRE-2014-YearEnd-Combined.pdf

Example

($000s)	2014 Q4	2014 Q3	2014 Q2	2014 Q1	2013 Q4	2013 Q3	2013 Q2	2013 Q1
Cash flow from operating activities	**80,866**	78,006	67,280	59,781	65,932	61,756	60,835	45,733
Change in non-cash working capital	**(18,865)**	(996)	5,452	8,923	(11,758)	(266)	1,958	1,949
Abandonment costs	**6,177**	3,189	697	1,346	1,760	814	434	962
Funds flow from operations	**68,178**	80,199	73,429	70,050	55,934	62,304	63,227	48,644
Weighted average outstanding shares *(000s)*								
- Basic	**193,497**	176,318	134,291	125,730	125,629	125,620	125,620	125,620
- Diluted	**193,497**	177,003	135,437	126,129	126,245	125,620	125,620	125,620
Funds flow from operations per share *($/share)*								
- Basic	**0.35**	0.45	0.55	0.56	0.45	0.50	0.50	0.39
- Diluted	**0.35**	0.45	0.54	0.56	0.44	0.50	0.50	0.39

Columns

($000s)	**2014**	2013
Cash flow from operating activities	**285,933**	234,256
Change in non-cash working capital	**(5,486)**	(8,117)
Abandonment costs	**11,409**	3,970
Funds flow from operations	**291,856**	230,109
Weighted average outstanding shares *(000s)*		
- Basic	**157,697**	125,622
- Diluted	**157,697**	125,778

Table Source: http://iq.iradesso.ca/main/components/clients_profiles/55/financial_reports/LRE-2014-YearEnd-Combined.pdf

Example

($000s)	2014 Q4	2014 Q3	2014 Q2	2014 Q1	2013 Q4	2013 Q3	2013 Q2	2013 Q1
Cash flow from operating activities	**80,866**	78,006	67,280	59,781	65,932	61,756	60,835	45,733
Change in non-cash working capital	**(18,865)**	(996)	5,452	8,923	(11,758)	(266)	1,958	1,949
Abandonment costs	**6,177**	3,189	697	1,346	1,760	814	434	962
Funds flow from operations	**68,178**	80,199	73,429	70,050	55,934	62,304	63,227	48,644
Weighted average outstanding shares *(000s)*								
- Basic	**193,497**	176,318	134,291	125,730	125,629	125,620	125,620	125,620
- Diluted	**193,497**	177,003	135,437	126,129	126,245	125,620	125,620	125,620
Funds flow from operations per share *($/share)*								
- Basic	**0.35**	0.45	0.55	0.56	0.45	0.50	0.50	0.39
- Diluted	**0.35**	0.45	0.54	0.56	0.44	0.50	0.50	0.39

Rows

($000s)	**2014**	2013
Cash flow from operating activities	**285,933**	234,256
Change in non-cash working capital	**(5,486)**	(8,117)
Abandonment costs	**11,409**	3,970
Funds flow from operations	**291,856**	230,109
Weighted average outstanding shares *(000s)*		
- Basic	**157,697**	125,622
- Diluted	**157,697**	125,778

Table Source: http://iq.iradesso.ca/main/components/clients_profiles/55/financial_reports/LRE-2014-YearEnd-Combined.pdf

Example

Multi-line "Blobs"

($000s)	2014 Q4	2014 Q3	2014 Q2	2014 Q1	2013 Q4	2013 Q3	2013 Q2	2013 Q1
Cash flow from operating activities	**80,866**	78,006	67,280	59,781	65,932	61,756	60,835	45,733
Change in non-cash working capital	**(18,865)**	(996)	5,452	8,923	(11,758)	(266)	1,958	1,949
Abandonment costs	**6,177**	3,189	697	1,346	1,760	814	434	962
Funds flow from operations	**68,178**	80,199	73,429	70,050	55,934	62,304	63,227	48,644
Weighted average outstanding shares *(000s)*								
- Basic	**193,497**	176,318	134,291	125,730	125,629	125,620	125,620	125,620
- Diluted	**193,497**	177,003	135,437	126,129	126,245	125,620	125,620	125,620
Funds flow from operations per share *($/share)*								
- Basic	**0.35**	0.45	0.55	0.56	0.45	0.50	0.50	0.39
- Diluted	**0.35**	0.45	0.54	0.56	0.44	0.50	0.50	0.39

($000s)	**2014**	2013
Cash flow from operating activities	**285,933**	234,256
Change in non-cash working capital	**(5,486)**	(8,117)
Abandonment costs	**11,409**	3,970
Funds flow from operations	**291,856**	230,109
Weighted average outstanding shares *(000s)*		
- Basic	**157,697**	125,622
- Diluted	**157,697**	125,778

Table Source: http://iq.iradesso.ca/main/components/clients_profiles/55/financial_reports/LRE-2014-YearEnd-Combined.pdf

Start with Columns

Many systems detect columns first:

- T-Recs [KD98], Pdf2table [Y05], Lixto [HB07], Tesseract [SS10], smartFIX [D11]

Example – Tesseract [SS10] :

1. Detect X-axis “tab-stops” (alignment positions)

2.	Group tokens	between “tab-stops”	horizontally	into entries
3.	Group entries	of the same font	vertically	into column fragments
4.	Group column fragments	within page columns	horizontally	into table fragments
5.	Group table fragments	if columns match	vertically	into tables

[KD98] T. Kieninger and A. Dengel. “The T-Recs Table Recognition and Analysis System”, DAS ‘98
[Y05] B. Yildiz et al. “pdf2table: A Method to Extract Table Information from PDF Files”, IICAI ‘05
[HB07] T. Hassan and R. Baumgartner. “Table Recognition and Understanding from PDF Files”, ICDAR ‘07
[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

Example

Tab-Stops

($000s)	2014 Q4	2014 Q3	2014 Q2	2014 Q1	2013 Q4	2013 Q3	2013 Q2	2013 Q1
Cash flow from operating activities	**80,866**	78,006	67,280	59,781	65,932	61,756	60,835	45,733
Change in non-cash working capital	**(18,865)**	(996)	5,452	8,923	(11,758)	(266)	1,958	1,949
Abandonment costs	**6,177**	3,189	697	1,346	1,760	814	434	962
Funds flow from operations	**68,178**	80,199	73,429	70,050	55,934	62,304	63,227	48,644
Weighted average outstanding shares *(000s)*								
- Basic	**193,497**	176,318	134,291	125,730	125,629	125,620	125,620	125,620
- Diluted	**193,497**	177,003	135,437	126,129	126,245	125,620	125,620	125,620
Funds flow from operations per share *($/share)*								
- Basic	**0.35**	0.45	0.55	0.56	0.45	0.50	0.50	0.39
- Diluted	**0.35**	0.45	0.54	0.56	0.44	0.50	0.50	0.39

($000s)	**2014**	2013
Cash flow from operating activities	**285,933**	234,256
Change in non-cash working capital	**(5,486)**	(8,117)
Abandonment costs	**11,409**	3,970
Funds flow from operations	**291,856**	230,109
Weighted average outstanding shares *(000s)*		
- Basic	**157,697**	125,622
- Diluted	**157,697**	125,778

Table Source: http://iq.iradesso.ca/main/components/clients_profiles/55/financial_reports/LRE-2014-YearEnd-Combined.pdf

Example

	2014				2013			
($000s)	**Q4**	Q3	Q2	Q1	Q4	Q3	Q2	Q1
Cash flow from operating activities	**80,866**	78,006	67,280	59,781	65,932	61,756	60,835	45,733
Change in non-cash working capital	**(18,865)**	(996)	5,452	8,923	(11,758)	(266)	1,958	1,949
Abandonment costs	**6,177**	3,189	697	1,346	1,760	814	434	962
Funds flow from operations	**68,178**	80,199	73,429	70,050	55,934	62,304	63,227	48,644
Weighted average outstanding shares *(000s)*								
- Basic	**193,497**	176,318	134,291	125,730	125,629	125,620	125,620	125,620
- Diluted	**193,497**	177,003	135,437	126,129	126,245	125,620	125,620	125,620
Funds flow from operations per share *($/share)*								
- Basic	**0.35**	0.45	0.55	0.56	0.45	0.50	0.50	0.39
- Diluted	**0.35**	0.45	0.54	0.56	0.44	0.50	0.50	0.39

($000s)	**2014**	2013
Cash flow from operating activities	**285,933**	234,256
Change in non-cash working capital	**(5,486)**	(8,117)
Abandonment costs	**11,409**	3,970
Funds flow from operations	**291,856**	230,109
Weighted average outstanding shares *(000s)*		
- Basic	**157,697**	125,622
- Diluted	**157,697**	125,778

Column Fragments

Table Source: http://iq.iradesso.ca/main/components/clients_profiles/55/financial_reports/LRE-2014-YearEnd-Combined.pdf

Example

Table Fragments

	2014				2013			
($000s)	**Q4**	Q3	Q2	Q1	Q4	Q3	Q2	Q1
Cash flow from operating activities	**80,866**	78,006	67,280	59,781	65,932	61,756	60,835	45,733
Change in non-cash working capital	**(18,865)**	(996)	5,452	8,923	(11,758)	(266)	1,958	1,949
Abandonment costs	**6,177**	3,189	697	1,346	1,760	814	434	962
Funds flow from operations	**68,178**	80,199	73,429	70,050	55,934	62,304	63,227	48,644
Weighted average outstanding shares *(000s)*								
- Basic	**193,497**	176,318	134,291	125,730	125,629	125,620	125,620	125,620
- Diluted	**193,497**	177,003	135,437	126,129	126,245	125,620	125,620	125,620
Funds flow from operations per share *($/share)*								
- Basic	**0.35**	0.45	0.55	0.56	0.45	0.50	0.50	0.39
- Diluted	**0.35**	0.45	0.54	0.56	0.44	0.50	0.50	0.39

($000s)	**2014**	2013
Cash flow from operating activities	**285,933**	234,256
Change in non-cash working capital	**(5,486)**	(8,117)
Abandonment costs	**11,409**	3,970
Funds flow from operations	**291,856**	230,109
Weighted average outstanding shares *(000s)*		
- Basic	**157,697**	125,622
- Diluted	**157,697**	125,778

Table Source: http://iq.iradesso.ca/main/components/clients_profiles/55/financial_reports/LRE-2014-YearEnd-Combined.pdf

Example

($000s)	2014 Q4	2014 Q3	2014 Q2	2014 Q1	2013 Q4	2013 Q3	2013 Q2	2013 Q1
Cash flow from operating activities	**80,866**	78,006	67,280	59,781	65,932	61,756	60,835	45,733
Change in non-cash working capital	**(18,865)**	(996)	5,452	8,923	(11,758)	(266)	1,958	1,949
Abandonment costs	**6,177**	3,189	697	1,346	1,760	814	434	962
Funds flow from operations	**68,178**	80,199	73,429	70,050	55,934	62,304	63,227	48,644
Weighted average outstanding shares *(000s)*								
- Basic	**193,497**	176,318	134,291	125,730	125,629	125,620	125,620	125,620
- Diluted	**193,497**	177,003	135,437	126,129	126,245	125,620	125,620	125,620
Funds flow from operations per share *($/share)*								
- Basic	**0.35**	0.45	0.55	0.56	0.45	0.50	0.50	0.39
- Diluted	**0.35**	0.45	0.54	0.56	0.44	0.50	0.50	0.39

Table Fragments

($000s)	2014	2013
Cash flow from operating activities	**285,933**	234,256
Change in non-cash working capital	**(5,486)**	(8,117)
Abandonment costs	**11,409**	3,970
Funds flow from operations	**291,856**	230,109
Weighted average outstanding shares *(000s)*		
- Basic	**157,697**	125,622
- Diluted	**157,697**	125,778

Table Source: http://iq.iradesso.ca/main/components/clients_profiles/55/financial_reports/LRE-2014-YearEnd-Combined.pdf

Table Extraction

Example

Tables

($000s)	2014 Q4	2014 Q3	2014 Q2	2014 Q1	2013 Q4	2013 Q3	2013 Q2	2013 Q1
Cash flow from operating activities	**80,866**	78,006	67,280	59,781	65,932	61,756	60,835	45,733
Change in non-cash working capital	**(18,865)**	(996)	5,452	8,923	(11,758)	(266)	1,958	1,949
Abandonment costs	**6,177**	3,189	697	1,346	1,760	814	434	962
Funds flow from operations	**68,178**	80,199	73,429	70,050	55,934	62,304	63,227	48,644
Weighted average outstanding shares *(000s)*								
- Basic	**193,497**	176,318	134,291	125,730	125,629	125,620	125,620	125,620
- Diluted	**193,497**	177,003	135,437	126,129	126,245	125,620	125,620	125,620
Funds flow from operations per share *($/share)*								
- Basic	**0.35**	0.45	0.55	0.56	0.45	0.50	0.50	0.39
- Diluted	**0.35**	0.45	0.54	0.56	0.44	0.50	0.50	0.39

($000s)	**2014**	2013
Cash flow from operating activities	**285,933**	234,256
Change in non-cash working capital	**(5,486)**	(8,117)
Abandonment costs	**11,409**	3,970
Funds flow from operations	**291,856**	230,109
Weighted average outstanding shares *(000s)*		
- Basic	**157,697**	125,622
- Diluted	**157,697**	125,778

Table Source: http://iq.iradesso.ca/main/components/clients_profiles/55/financial_reports/LRE-2014-YearEnd-Combined.pdf

Example

Multi-Column Headers

($000s)	2014				2013			
	Q4	Q3	Q2	Q1	Q4	Q3	Q2	Q1
Cash flow from operating activities	**80,866**	78,006	67,280	59,781	65,932	61,756	60,835	45,733
Change in non-cash working capital	**(18,865)**	(996)	5,452	8,923	(11,758)	(266)	1,958	1,949
Abandonment costs	**6,177**	3,189	697	1,346	1,760	814	434	962
Funds flow from operations	**68,178**	80,199	73,429	70,050	55,934	62,304	63,227	48,644
Weighted average outstanding shares *(000s)*								
- Basic	**193,497**	176,318	134,291	125,730	125,629	125,620	125,620	125,620
- Diluted	**193,497**	177,003	135,437	126,129	126,245	125,620	125,620	125,620
Funds flow from operations per share *($/share)*								
- Basic	**0.35**	0.45	0.55	0.56	0.45	0.50	0.50	0.39
- Diluted	**0.35**	0.45	0.54	0.56	0.44	0.50	0.50	0.39

Tables

($000s)	**2014**	2013
Cash flow from operating activities	**285,933**	234,256
Change in non-cash working capital	**(5,486)**	(8,117)
Abandonment costs	**11,409**	3,970
Funds flow from operations	**291,856**	230,109
Weighted average outstanding shares *(000s)*		
- Basic	**157,697**	125,622
- Diluted	**157,697**	125,778

Table Source: http://iq.iradesso.ca/main/components/clients_profiles/55/financial_reports/LRE-2014-YearEnd-Combined.pdf

Start with Rows

Systems ***with ML*** often detect rows first

- Pinto-McCallum [P03], e Silva [S06], TableSeer [L08], PDF-TREX [OR09]

Typical process:

1. Identify text-lines
2. Train an ML classifier to label text-lines:
   - "Table Dense", "Table Sparse", "Table Header", "Non-table", etc.
   - ML = CRF [P03], HMM [S06], SVM [L08], etc.
3. Merge sparse rows into dense rows – get full table rows:
   - Merge up, down, or cluster around, by **row alignment** [H00a]
4. Combine table rows into tables

[H00a] J. C. Handley. "Table Analysis for Multi-line Cell Identification", SPIE Doc. Recog. & Retr. '00
[P03] D. Pinto et al. "Table Extraction Using Conditional Random Fields", SIGIR '03
[S06] A. C. e Silva et al. "Design of an End-to-end Method to Extract Information from Tables", IJDAR '06
[L08] Y. Liu et al. "Identifying Table Boundaries in Digital Documents via Sparse Line Detection", CIKM '08
[OR09] E. Oro and M. Ruffolo. "PDF-TREX: An Approach for Recognizing and Extracting Tables from PDF Documents", ICDAR '09

Example

Align-ment

	2014				2013				
($000s)	**Q4**	Q3	Q2	Q1	Q4	Q3	Q2	Q1	
Cash flow from operating									Sparse Row
activities	**80,866**	78,006	67,280	59,781	65,932	61,756	60,835	45,733	Dense Row
Change in non-cash									Sparse Row
working capital	**(18,865)**	(996)	5,452	8,923	(11,758)	(266)	1,958	1,949	Dense Row
Abandonment costs	**6,177**	3,189	697	1,346	1,760	814	434	962	Dense Row
Funds flow from									Sparse Row
operations	**68,178**	80,199	73,429	70,050	55,934	62,304	63,227	48,644	Dense Row
Weighted average									Sparse Row
outstanding shares									Sparse Row
(000s)									Sparse Row
- Basic	**193,497**	176,318	134,291	125,730	125,629	125,620	125,620	125,620	Dense Row
- Diluted	**193,497**	177,003	135,437	126,129	126,245	125,620	125,620	125,620	Dense Row
Funds flow from									Sparse Row
operations per share									Sparse Row
($/share)									Sparse Row
- Basic	**0.35**	0.45	0.55	0.56	0.45	0.50	0.50	0.39	Dense Row
- Diluted	**0.35**	0.45	0.54	0.56	0.44	0.50	0.50	0.39	Dense Row

Table Header
Table Header

($000s)	**2014**	2013	Table Header
Cash flow from operating activities	**285,933**	234,256	Dense Row
Change in non-cash working capital	**(5,486)**	(8,117)	Dense Row
Abandonment costs	**11,409**	3,970	Dense Row
Funds flow from operations	**291,856**	230,109	Dense Row
Weighted average outstanding shares			Sparse Row
(000s)			Sparse Row
- Basic	**157,697**	125,622	Dense Row
- Diluted	**157,697**	125,778	Dense Row

Table Source: http://iq.iradesso.ca/main/components/clients_profiles/55/financial_reports/LRE-2014-YearEnd-Combined.pdf

Example

	2014				2013				Table Header
($000s)	**Q4**	Q3	Q2	Q1	Q4	Q3	Q2	Q1	Table Header
Cash flow from operating activities	**80,866**	78,006	67,280	59,781	65,932	61,756	60,835	45,733	Sparse Row Dense Row
Change in non-cash working capital	**(18,865)**	(996)	5,452	8,923	(11,758)	(266)	1,958	1,949	Sparse Row Dense Row
Abandonment costs	**6,177**	3,189	697	1,346	1,760	814	434	962	Dense Row
Funds flow from operations	**68,178**	80,199	73,429	70,050	55,934	62,304	63,227	48,644	Sparse Row Dense Row
Weighted average outstanding shares *(000s)* - Basic	**193,497**	176,318	134,291	125,730	125,629	125,620	125,620	125,620	Sparse Row Sparse Row Sparse Row Dense Row
- Diluted	**193,497**	177,003	135,437	126,129	126,245	125,620	125,620	125,620	Dense Row
Funds flow from operations per share *($/share)* - Basic	**0.35**	0.45	0.55	0.56	0.45	0.50	0.50	0.39	Sparse Row Sparse Row Sparse Row Dense Row
- Diluted	**0.35**	0.45	0.54	0.56	0.44	0.50	0.50	0.39	Dense Row

Align-ment

($000s)	**2014**	2013	Table Header
Cash flow from operating activities	**285,933**	234,256	Dense Row
Change in non-cash working capital	**(5,486)**	(8,117)	Dense Row
Abandonment costs	**11,409**	3,970	Dense Row
Funds flow from operations	**291,856**	230,109	Dense Row
Weighted average outstanding shares *(000s)* - Basic	**157,697**	125,622	Sparse Row Sparse Row Dense Row
- Diluted	**157,697**	125,778	Dense Row

Table Source: http://iq.iradesso.ca/main/components/clients_profiles/55/financial_reports/LRE-2014-YearEnd-Combined.pdf

Example

Align-ment

($000s)	**2014**				2013				Label
	Q4	Q3	Q2	Q1	Q4	Q3	Q2	Q1	Table Header
Cash flow from operating									Sparse Row
activities	**80,866**	78,006	67,280	59,781	65,932	61,756	60,835	45,733	Dense Row
Change in non-cash									Sparse Row
working capital	**(18,865)**	(996)	5,452	8,923	(11,758)	(266)	1,958	1,949	Dense Row
Abandonment costs	**6,177**	3,189	697	1,346	1,760	814	434	962	Dense Row
Funds flow from									Sparse Row
operations	**68,178**	80,199	73,429	70,050	55,934	62,304	63,227	48,644	Dense Row
Weighted average									Heading Row
outstanding shares									Heading Row
(000s)									Heading Row
- Basic	**193,497**	176,318	134,291	125,730	125,629	125,620	125,620	125,620	Dense Row
- Diluted	**193,497**	177,003	135,437	126,129	126,245	125,620	125,620	125,620	Dense Row
Funds flow from									Heading Row
operations per share									Heading Row
($/share)									Heading Row
- Basic	**0.35**	0.45	0.55	0.56	0.45	0.50	0.50	0.39	Dense Row
- Diluted	**0.35**	0.45	0.54	0.56	0.44	0.50	0.50	0.39	Dense Row

($000s)	**2014**	2013	Label
			Table Header
Cash flow from operating activities	**285,933**	234,256	Dense Row
Change in non-cash working capital	**(5,486)**	(8,117)	Dense Row
Abandonment costs	**11,409**	3,970	Dense Row
Funds flow from operations	**291,856**	230,109	Dense Row
Weighted average outstanding shares			Heading Row
(000s)			Heading Row
- Basic	**157,697**	125,622	Dense Row
- Diluted	**157,697**	125,778	Dense Row

Table Source: http://iq.iradesso.ca/main/components/clients_profiles/55/financial_reports/LRE-2014-YearEnd-Combined.pdf

Example

Align-ment

	2014				2013				
($000s)	**Q4**	Q3	Q2	Q1	Q4	Q3	Q2	Q1	Table Header / Table Header
Cash flow from operating activities	**80,866**	78,006	67,280	59,781	65,932	61,756	60,835	45,733	Sparse Row / Dense Row
Change in non-cash working capital	**(18,865)**	(996)	5,452	8,923	(11,758)	(266)	1,958	1,949	Sparse Row / Dense Row
Abandonment costs	**6,177**	3,189	697	1,346	1,760	814	434	962	Dense Row
Funds flow from operations	**68,178**	80,199	73,429	70,050	55,934	62,304	63,227	48,644	Sparse Row / Dense Row
Weighted average outstanding shares *(000s)*									Heading Row / Heading Row / Heading Row
- Basic	**193,497**	176,318	134,291	125,730	125,629	125,620	125,620	125,620	Dense Row
- Diluted	**193,497**	177,003	135,437	126,129	126,245	125,620	125,620	125,620	Dense Row
Funds flow from operations per share *($/share)*									Heading Row / Heading Row / Heading Row
- Basic	**0.35**	0.45	0.55	0.56	0.45	0.50	0.50	0.39	Dense Row
- Diluted	**0.35**	0.45	0.54	0.56	0.44	0.50	0.50	0.39	Dense Row

($000s)	**2014**	2013	Table Header
Cash flow from operating activities	**285,933**	234,256	Dense Row
Change in non-cash working capital	**(5,486)**	(8,117)	Dense Row
Abandonment costs	**11,409**	3,970	Dense Row
Funds flow from operations	**291,856**	230,109	Dense Row
Weighted average outstanding shares *(000s)*			Heading Row / Heading Row
- Basic	**157,697**	125,622	Dense Row
- Diluted	**157,697**	125,778	Dense Row

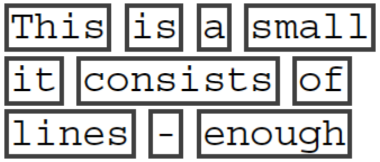
Table Source: http://iq.iradesso.ca/main/components/clients_profiles/55/financial_reports/LRE-2014-YearEnd-Combined.pdf

"Blobs" (Phrases ≤ Text-Lines ≤ Paragraphs)

- "Blob" = largest semantically bound text unit
  - Single-line or multi-line
  - If in a table, the whole "blob" must be in a single cell
- "Blob" ≠ Cell
  - Cell has **span** and **overlaps** other cells
  - Some "blobs" end up in plain text or non-table text
- "Blobs" help define table structure:
  - Trace alignment
  - Determine header cell spans
  - **Fix over-split / over-merged cells, rows, columns**
  - Reduce search space

How to Detect “Blobs”

- [KD98] Distance based clustering:
  - Merge words horizontally
  - Merge text strings vertically *if word-spans interleave*



- Problems with distance:
  - **Multi-column headers:** 1 justified phrase *vs.* ≥ 2 closely spaced phrases
  - **Row headers / text cells:** 1 multi-line cell *vs.* ≥ 2 closely spaced rows
- Example:

	Two Column Header		**Two Column Header**	
HEADER	Header	Header	Header	Header
Row 1, text line 1	0.12	1.23	2.34	3.45
Row 1, text line 2				
Row 1, text line 3				
Row 2, text line 1	4.56	5.67	6.78	7.89
Row 2, text line 2				
Row 2, text line 3				

[KD98] T. Kieninger and A. Dengel. “The T-Recs Table Recognition and Analysis System”, DAS ‘98

How to Detect "Blobs"

- [H00a], [OR09] Merge "sparse" rows into "dense" rows
  - Merge up, merge down, or cluster around
- [L09] Detect and follow reading order ← **an NLP challenge**
- [B12] [B14] Train a classifier over "blob" features:
  - Proper termination (e.g. "blobs" don't end with a dash or comma)
  - Number of numeric strings
  - Indentation, large space at the end of a string
  - Shared font properties
- Deep learning approaches ← **see later in this tutorial**
  - Cell detection over image
  - Semantic relationship detection (over text) using BERT

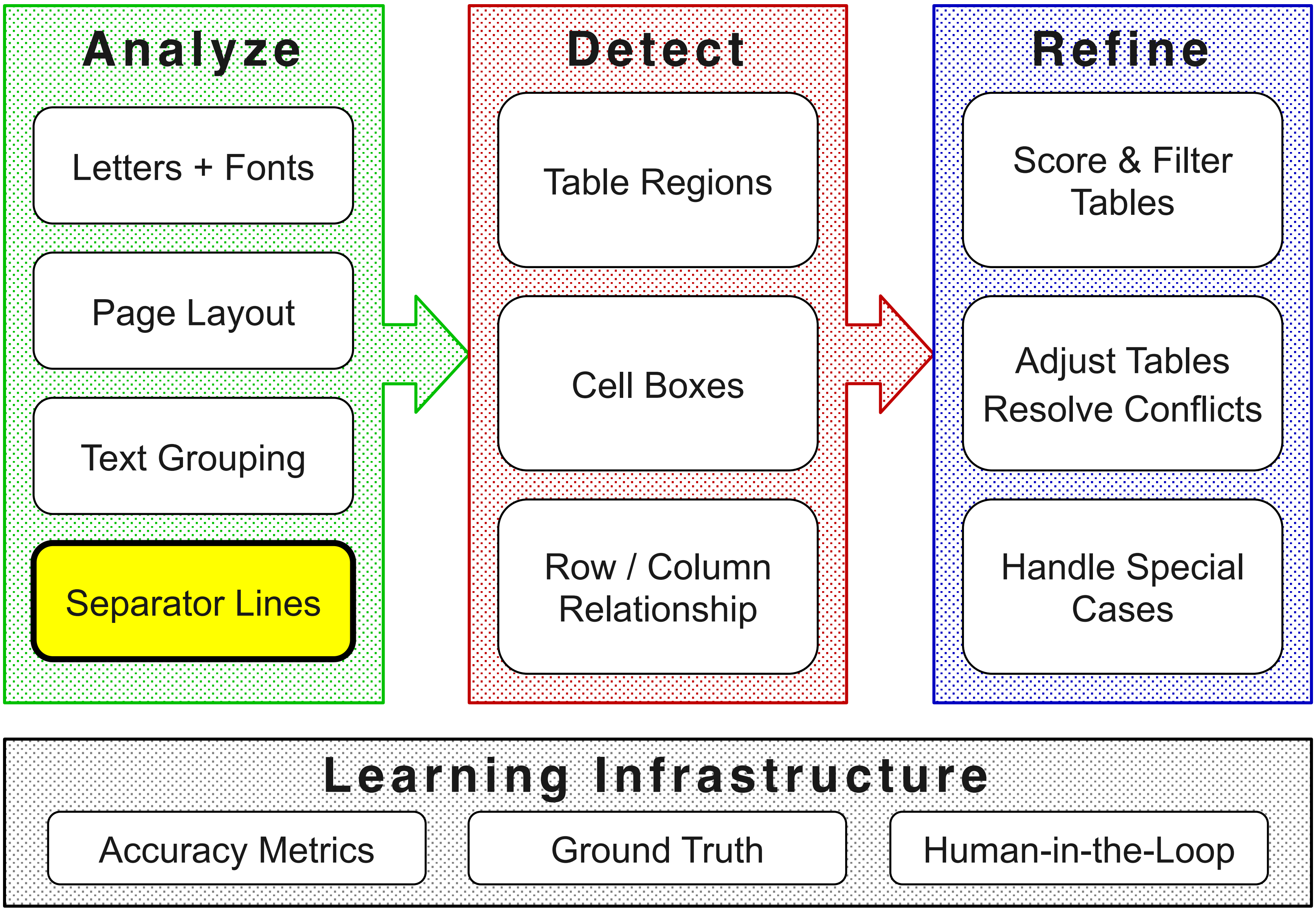
[H00a] J. C. Handley. "Table Analysis for Multi-line Cell Identification", SPIE Doc. Recog. & Retr. '00
[OR09] E. Oro and M. Ruffolo. "PDF-TREX: An Approach for Recognizing and Extracting Tables from PDF Documents", ICDAR '09
[L09] Y. Liu et al. "Improving the Table Boundary Detection in PDFs by Fixing the Sequence Error of the Sparse Lines", ICDAR '09
[B12] E. Bart. "Parsing Tables by Probabilistic Modeling of Perceptual Cues", DAS '12
[B14] A. Bansal et al. "Table Extraction from Document Images using Fixed Point Model", ICVGIP '14

Example

Name and Principal Position	Year	Salary ($)[1]	Bonus ($)[2]	Stock Awards ($)[3]	Non-Equity Incentive Plan Compensation ($)[1][4]	All Other Compensation ($)[5]	Total ($)
Bob Sasser Chief Executive Officer	2015	$1,585,577	—	$5,803,264	$2,080,320	$ 60,549	$ 9,529,710
	2014	$1,505,769	—	$4,104,531	$2,140,773	$ 63,415	$ 7,814,488
	2013	1,410,577	—	3,839,768	1,909,929	58,089	$ 7,218,363
Kevin Wampler Chief Financial Officer	2015	635,577	—	1,695,764	617,121	51,452	2,999,914
	2014	570,192	—	1,249,783	628,654	54,481	2,503,110
	2013	545,192	—	1,140,273	499,465	56,380	2,241,310
Gary Philbin President and Chief Operating Officer	2015	971,154	—	2,438,906	1,258,725	56,568	4,725,353
	2014	830,769	—	1,780,806	1,000,652	57,302	3,669,529
	2013	738,846	—	1,749,799	796,624	53,080	3,338,349
Robert H. Rudman Chief Merchandising Officer	2015	692,307	—	1,726,563	645,165	61,647	3,125,682
	2014	656,154	—	1,357,425	682,642	59,269	2,755,490
	2013	636,154	—	1,253,591	555,262	54,918	2,499,925
Michael Matacunas Chief Administrative Officer	2015	537,500	—	1,247,773	550,639	40,269	2,376,181
	2014	483,077	—	949,917	324,766	42,349	1,800,109
	2013	274,038	150,000	899,826	182,258	215,306	1,721,428
Howard Levine Former Chief Executive Officer of Family Dollar Stores	2015	666,388	—	—	—	11,838,299[6]	12,504,687
	2014						
	2013						

Table Source: https://www.dollartreeinfo.com/static-files/0c3687d8-e6ce-4566-bc89-79fc8c8b665e (2016_Proxy_Statement_Final.pdf)

Common Sub-Tasks in Table Extraction



Separator Line Detection

- Ruled Lines & Colored Boxes
  - Extend ruled lines over small gaps, “snap” together
  - Merge touching colored boxes, **then** convert into lines
  - Filter out: highlighting, underlining, boxed comments, logos, charts etc.
- **BUT:** A “perfect” ruled-line grid can be incomplete !
  - Some lines may be **missing**
  - Lines may **fail to extend** to header rows / columns

[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
[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

Table Extraction

Example 1

(Canadian dollars in millions, except where indicated)

	Third Quarter		Change	
	2015	2014	$	%
Aircraft fuel expense – GAAP	**$ 697**	**$ 939**	**$ (242)**	**(26)**
Add: Aircraft fuel expense related to regional airline operations	95	137	(42)	(31)
Total Aircraft fuel expense	**$ 792**	**$ 1,076**	**$ (284)**	**(26)**
Add: Net cash payments on fuel derivatives [1]	14	4	10	250
Economic cost of fuel – Non-GAAP [2]	**$ 806**	**$ 1,080**	**$ (274)**	**(25)**
Fuel consumption (thousands of litres)	**1,289,911**	**1,200,017**	**89,894**	**7.5**
Fuel cost per litre (cents) – GAAP	**61.4**	**89.7**	**(28.3)**	**(31.5)**
Economic fuel cost per litre (cents) – Non-GAAP [2]	**62.5**	**90.0**	**(27.5)**	**(30.6)**

(Canadian dollars in millions, except where indicated)

	First Nine Months		Change	
	2015	2014	$	%
Aircraft fuel expense – GAAP	**$ 1,937**	**$ 2,567**	**$ (630)**	**(25)**
Add: Aircraft fuel expense related to regional airline operations	278	389	(111)	(29)
Total Aircraft fuel expense	**$ 2,215**	**$ 2,956**	**$ (741)**	**(25)**
Add: Net cash payments on fuel derivatives [1]	36	6	30	500
Economic cost of fuel – Non-GAAP [2]	**$ 2,251**	**$ 2,962**	**$ (711)**	**(24)**
Fuel consumption (thousands of litres)	**3,442,909**	**3,220,893**	**222,016**	**6.9**
Fuel cost per litre (cents) – GAAP	**64.3**	**91.8**	**(27.4)**	**(29.9)**
Economic fuel cost per litre (cents) – Non-GAAP [2]	**65.4**	**91.9**	**(26.6)**	**(28.9)**

Table Source: https://www.aircanada.com/content/dam/aircanada/portal/documents/PDF/en/quarterly-result/2015/2015_MDA_q3.pdf

Table Extraction

Example 2

Minimum Number of Accessible Parking Spaces

ADA Standards for Accessible Design 4.1.2 (5)

Total Number of Parking spaces Provided (per lot)	Total Minimum Number of Accessible Parking Spaces (60” & 96” aisles)	**Van Accessible Parking Spaces with min. 96” wide access aisle**	**Accessible Parking Spaces with min. 60” wide access aisle**
	Column A		
1 to 25	1	**1**	**0**
26 to 50	2	**1**	**1**
51 to 75	3	**1**	**2**
76 to 100	4	**1**	**3**
101 to 150	5	**1**	**4**
151 to 200	6	**1**	**5**
201 to 300	7	**1**	**6**
301 to 400	8	**1**	**7**
401 to 500	9	**2**	**7**
501 to 1000	2% of total parking provided in each lot	**1/8 of Column A***	**7/8 of Column A****
1001 and over	20 plus 1 for each 100 over 1000	**1/8 of Column A***	**7/8 of Column A****

*** one out of every 8 accessible spaces** **** 7 out of every 8 accessible parking spaces**

Table Source: https://www.ada.gov/restripe.pdf

IBM

Example 3

course	*material type*	*row*						
Java	*prob. & anim. exam.*		*prob. ID*	*prob. name*	*prob. topic*	*anim. exam. topic*	*anim. exam. name*	*anim. exam. ID*
		1	14	jArrayList5	ArrayList	ArrayList	ae_arraylist2_v2	3
		2	18	jBoolean_Operators	Boolean expressions	Switch	ae_switch_demo2	44
		3	65	jMathFuc2	Arithmetic operations	Arithmetic operations	ae_arithmetic_v2	1
		4	100	jWhile1	Loops while	Loops while	ae_while_demo	49
	prob. & annot. exam.		*prob. ID*	*prob. name*	*prob. topic*	*annot. exam. topic*	*annot. exam. name*	*annot. exam. ID*
		5	37	jDowhile1	Loops do_while	Loops for	for1_v2	28
		6	57	jInterfaces1	Interfaces	Variables	PrintTester	78
		7	61	jInterfaces5	Interfaces	Objects	AccessorMutatorDemo	1
		8	63	jMathCeil	Arithmetic operations	Loops for	JavaTutorial_4_6_8	57
Python	*prob. & annot. exam.*		*prob. ID*	*prob. name*	*prob. topic*	*annot. exam. topic*	*annot. exam. name*	*annot. exam. ID*
		9	3	q_py_arithmetic1	Variables	Variables	pyt1.3	5
		10	21	q_py_nested_if_elif1	if_statements	values_references	pytt10.25	58
		11	23	q_py_obj_account1	classes_objects	Lists	pyt7.2	53
	prob. & anim. exam.		*prob. ID*	*prob. name*	*prob. topic*	*anim. exam. topic*	*anim. exam. name*	*anim. exam. ID*
		12	7	q_py_dict_access1	dictionary	loops	ae_adl_while	39
		13	29	q_py_output1	output_formatting	variables	ae_adl_arithmetics2	1
		14	10	q_py_fun_car1	functions	exceptions	ae_adl_tryexcept2	34
	prob. & pars. prob.		*prob. ID*	*prob. name*	*prob. topic*	*pars. prob. topic*	*pars. prob. name*	*pars. prob. ID*
		15	10	q_py_fun_car1	functions	exceptions	ps_python_try_adding	38
		16	12	q_py_if_elif1	if_statements	loops	combo_python_while	9
		17	35	q_py_swap1	variables	variables	combo_swap	11
	pars. prob. & annot. exam.		*pars. prob. ID*	*pars. prob. name*	*pars. prob. topic*	*annot. exam. topic*	*annot. exam. name*	*annot. exam. ID*
		18	1	combo_avg	variables	variables	pyt2.1	32
		19	14	ps_python_addition	variables	variables	pyt1.2	4
		20	41	ps_return_bigger_or_none	functions	functions	pyt10.7	30
	pars. prob. & anim. exam.		*pars. prob. ID*	*pars. prob. name*	*pars. prob. topic*	*anim. exam. topic*	*anim. exam. name*	*anim. exam. ID*
		21	1	combo_avg	variables	variables	ae_python_assignment	40
		22	12	ps_hello	variables	variables	ae_adl_arithmetics2	1
		23	43	ps_simple_params	functions	functions	ae_adl_returnvalue	29

Table Source: http://educationaldatamining.org/files/conferences/EDM2018/EDM2018_Preface_TOC_Proceedings.pdf

Separator Line Detection

- White-space separators ("virtual" lines)
  - Help define cell span / cell alignment in tables
  - **Prune false-positives** by ML or by heuristics [B12]
- How to detect white-space separators
  - Cell-unit ("blob") bounding box expansion [I93]
  - Axis projection histograms [CK93]
  - White-space cover by maximum-area white-space rectangles [F11]
- How to prune separators (features to use)
  - **Adjacent text "blobs"** : alignment, size, and content
  - **Other separators** that run parallel to, or **intersect**, the separator

[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
[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

Common Sub-Tasks in Table Extraction

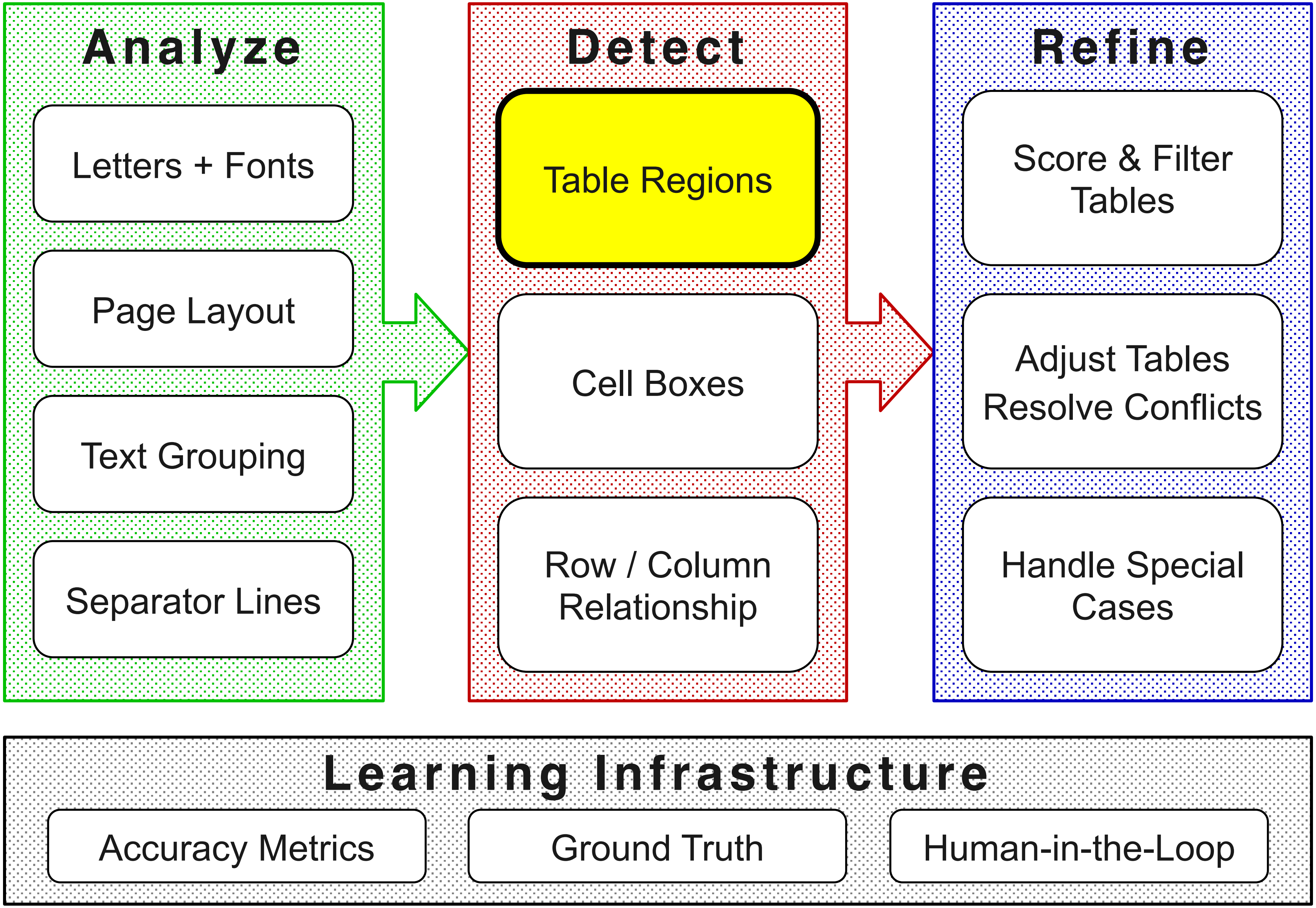


Table Detection Overview

- (**Pre-DL**) Find elements of tables and group them to find the whole table (rows/columns, blobs or lines first)

- (**CNN-based**) Try a fixed set of table region proposals from object detection
  - CNN shares computation of features across all translations of a given proposal rectangle
  - Proposal rectangle shapes / sizes are fixed as hyperparameters
  - If a proposal hits a table, a regression decides table borders

[CL12] J. Chen and D. Lopresti. "Model-Based Tabular Structure Detection and Recognition in Noisy Handwritten Documents", ICFHR '12
[B14] A. Bansal et al. "Table Extraction from Document Images using Fixed Point Model", ICVGIP '14
[G17] A. Gilani et al. "Table Detection using Deep Learning", ICDAR '17
[S18b] S. A. Siddiqui et al. "DeCNT: Deep Deformable CNN for Table Detection", IEEE Acc. '18

Detect Candidate Table Regions (pre-DL)

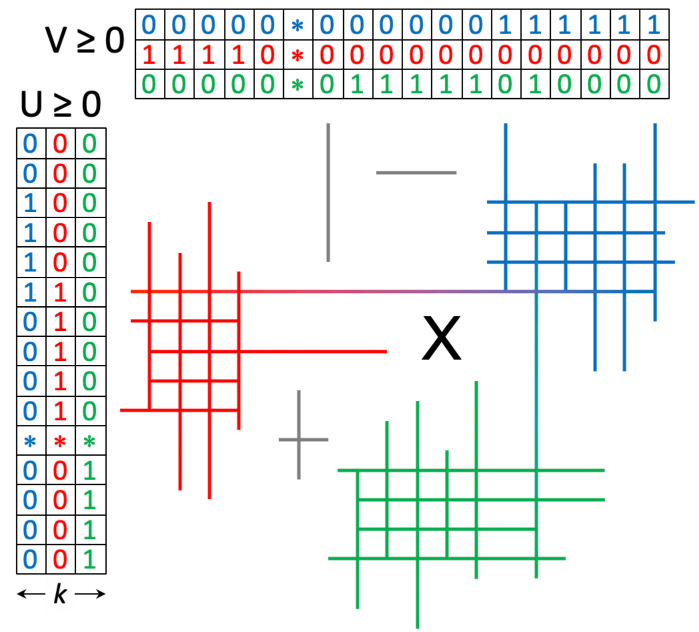
- **Ruled Line grids** / frames, connected components
- (**Rows 1st**) Stack "table" rows whose "blobs" co-align [L08], [OR09]
  - Rows are labeled by an ML-classifier (CRF, SVM, HMM)
  - Labels & matching "blob" layout → table regions
  - **NOTE:** Be sure to label "header rows" to tell tables apart !
- (**Cols 1st**) Cluster overlapping column fragments [HB07], [SS10]
  - Group table columns horizontally, staying within page layout columns (when possible)
  - Group vertically if column fragments overlap, match, or subsume
  - **NOTE:** Column header areas require special handling !

[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
[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
[K13] T. Kasar et al. "Learning to Detect Tables in Scanned Document Images Using Line Information", ICDAR '13

Detect Candidate Table Regions (pre-DL)

- (**Blobs 1st**) Classify text "blobs", cluster those labeled "table"
  - [B14] iteratively labels "blobs" **given their neighbors' labels**
  - [B14] trains a Kernel Logistic Regression classifier
- (**Lines 1st**) Find areas where "strong" separators make a grid
  - [CL12] uses Max-Flow / Min-Cut algorithm to extract grids
  - Bi-cluster the intersection matrix of horizontal *vs.* vertical separators
    - Example: Non-neg. matrix factorization for grid clustering (right)

Non-neg. Matrix Factorization for Grid Clustering



[CL12] J. Chen and D. Lopresti. "Model-Based Tabular Structure Detection and Recognition in Noisy Handwritten Documents", ICFHR '12
[B14] A. Bansal et al. "Table Extraction from Document Images using Fixed Point Model", ICVGIP '14

Deep Learning for Table Detection

Use existing object detection frameworks (Faster R-CNN or YOLO) retrained for table detection

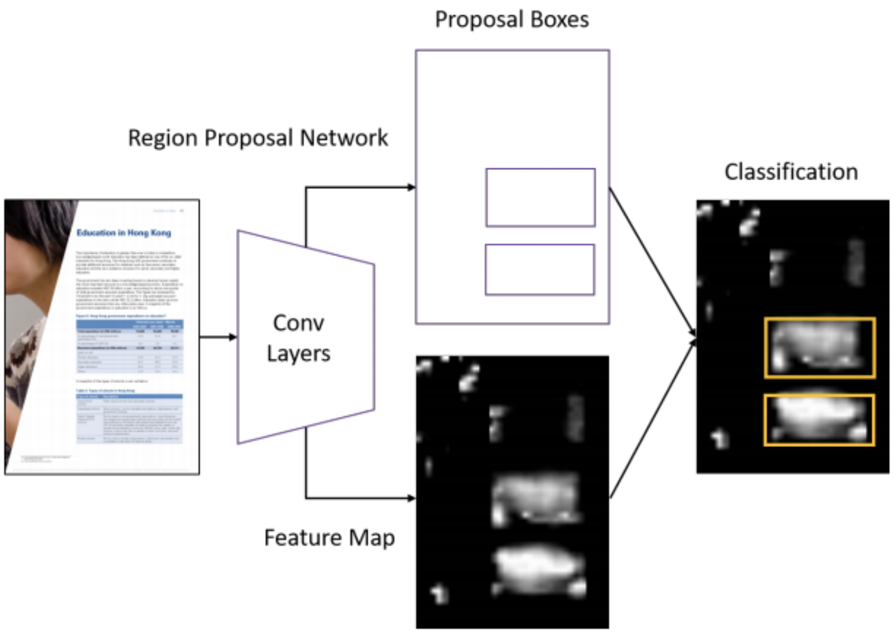
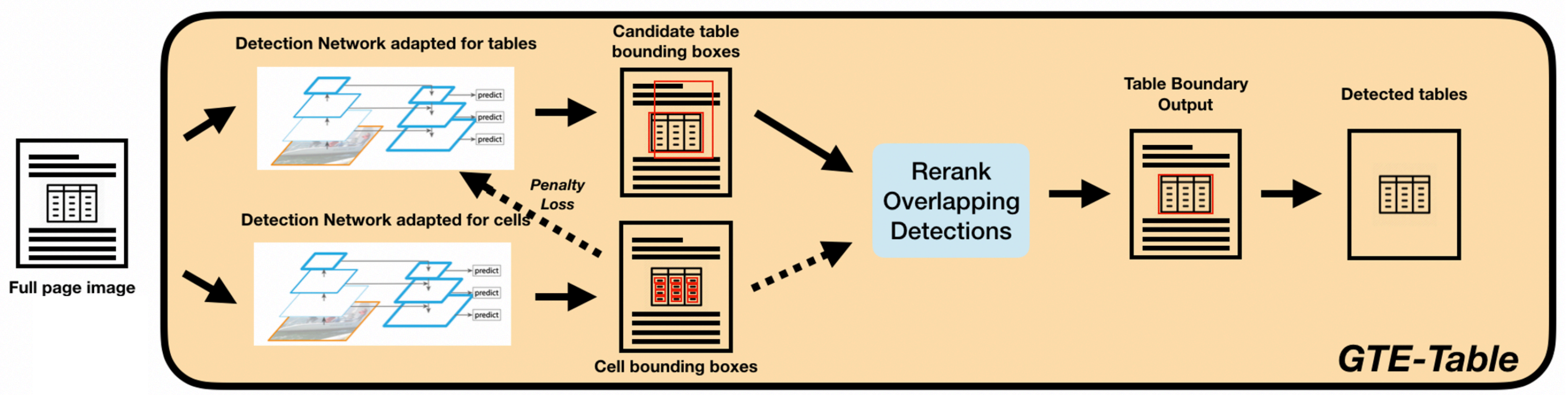


Figure 5: The Faster R-CNN model for table detection

[G17] A. Gilani et al. "Table Detection using Deep Learning", ICDAR '17
[S17] Schreiber et al. "DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images" ICDAR '17
[S18a] P. Staar et al. "Corpus Conversion Service: A Machine Learning Platform to Ingest Documents at Scale", KDD '18
[L20] Li et al. "TableBank: Table Benchmark for Image-based Table Detection and Recognition". LREC '20
[Z20a] Zheng et al. "Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context", arXiv 2020
[P20a] D. Prasad et al. "CascadeTabNet: An Approach for End to End Table Detection and Structure Recognition from Image-Based Documents", In *CVPR Workshops* 2020
[P20b] Paliwal et al. "TableNet: Deep Learning Model for End-to-end Table Detection and Tabular Data Extraction from Scanned Document Images", arXiv 2020

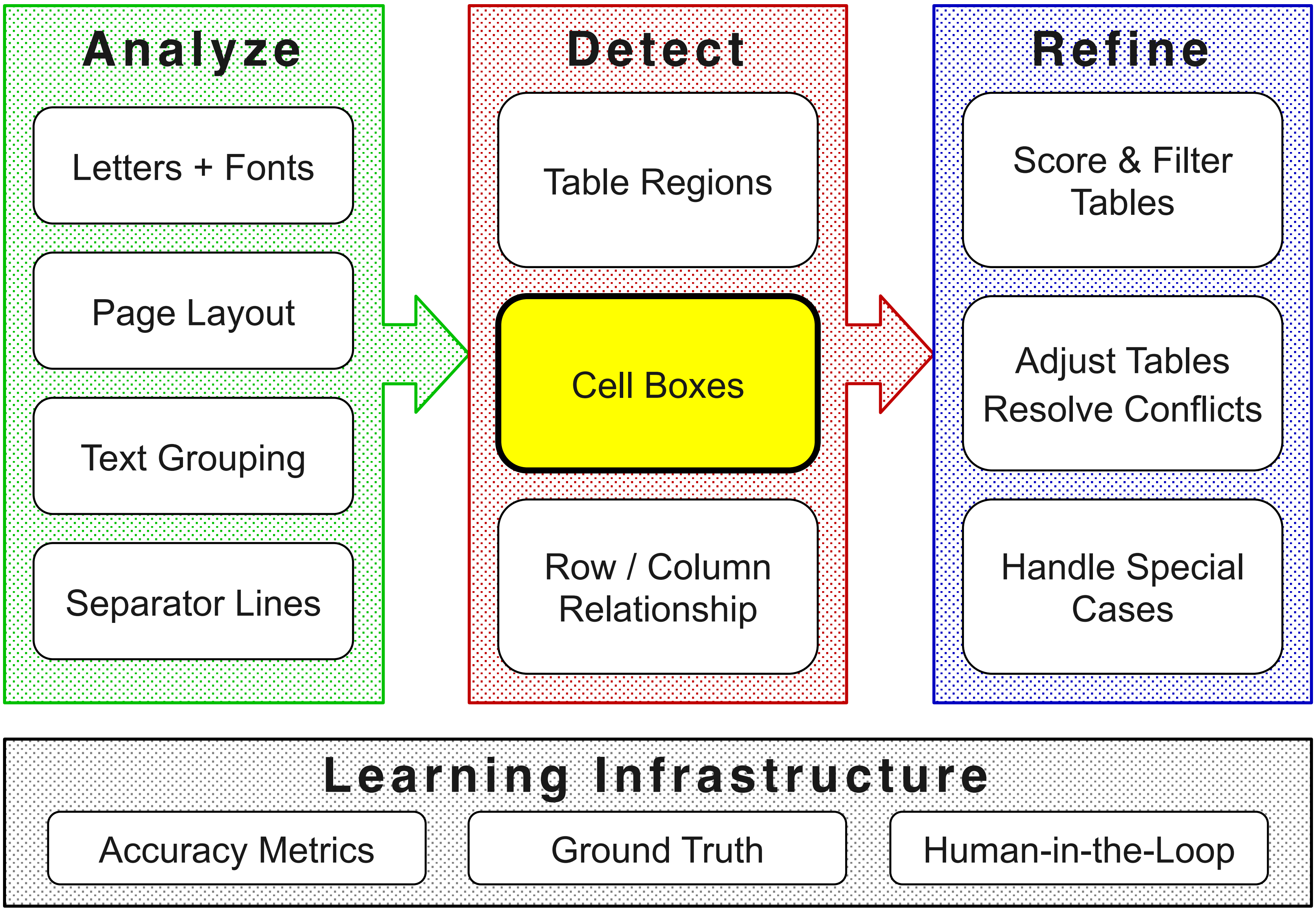
GTE-table

Leverage spatial containment relationship between tables and cells to improve table object recognition



[Zheng et al. Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context, arXiv 2020]

Common Sub-Tasks in Table Extraction



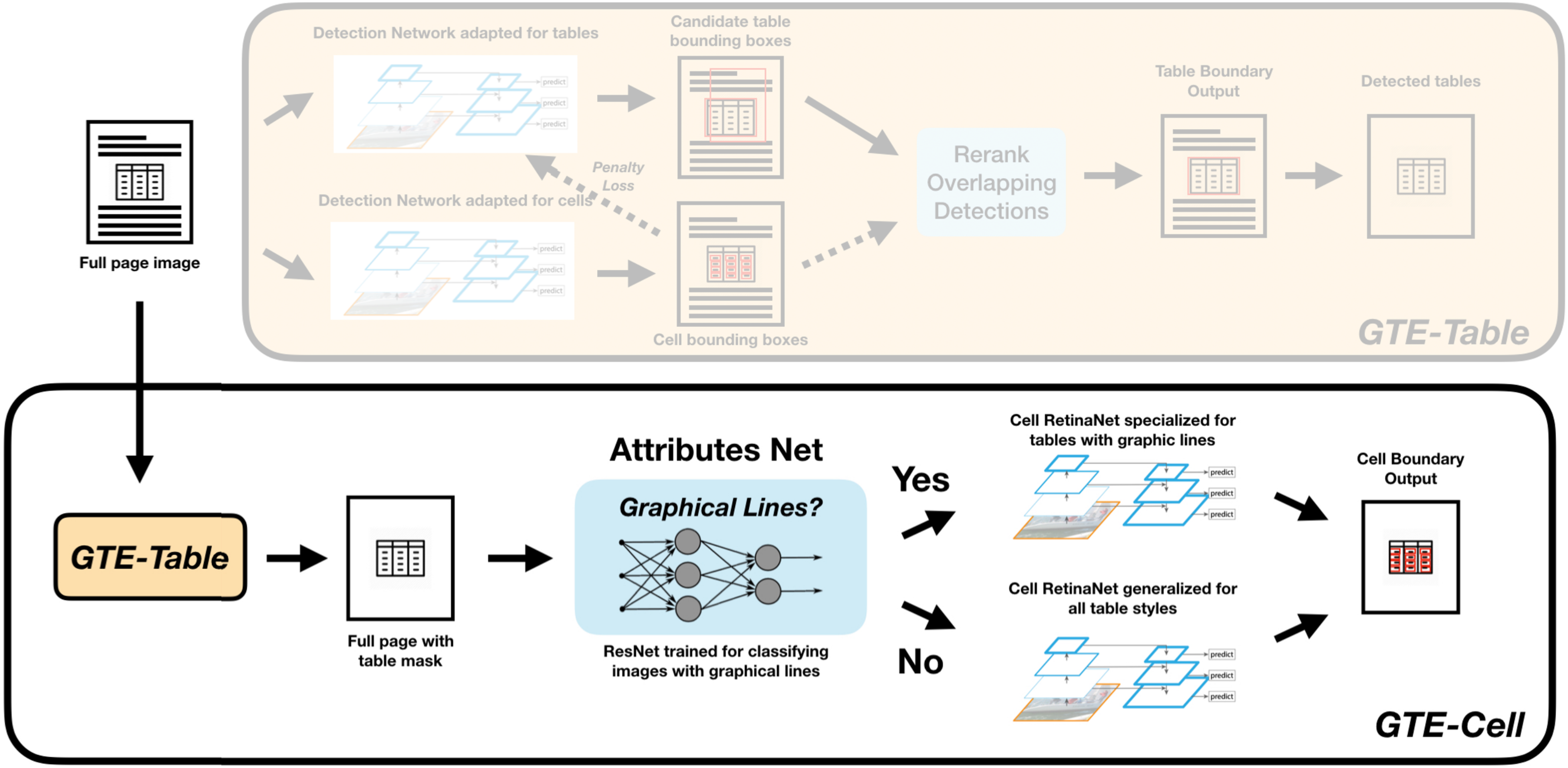
Cell Detection – Overview

- Pre-DL approaches:
  - Just use text “blobs” as cells
  - Iteratively merge “blobs” sharing columns & rows [H00a] [OR09]
  - Use separator lines to define cells [B12]
- Deep Learning approaches:
  - Detect cells over image using object detection CNNs [Z20a] [P20a]

[H00a] J. C. Handley. “Table Analysis for Multi-line Cell Identification”, SPIE Doc. Recog. & Retr. ‘00
[OR09] E. Oro and M. Ruffolo. “PDF-TREX: An Approach for Recognizing and Extracting Tables from PDF Documents”, ICDAR ’09
[B12] E. Bart. “Parsing Tables by Probabilistic Modeling of Perceptual Cues”, DAS ‘12
[Z20a] Zheng et al. Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context, arXiv 2020
[P20a] D. Prasad et al. CascadeTabNet: An approach for end to end table detection and structure recognition from image-based documents. In *CVPR Workshops* 2020.

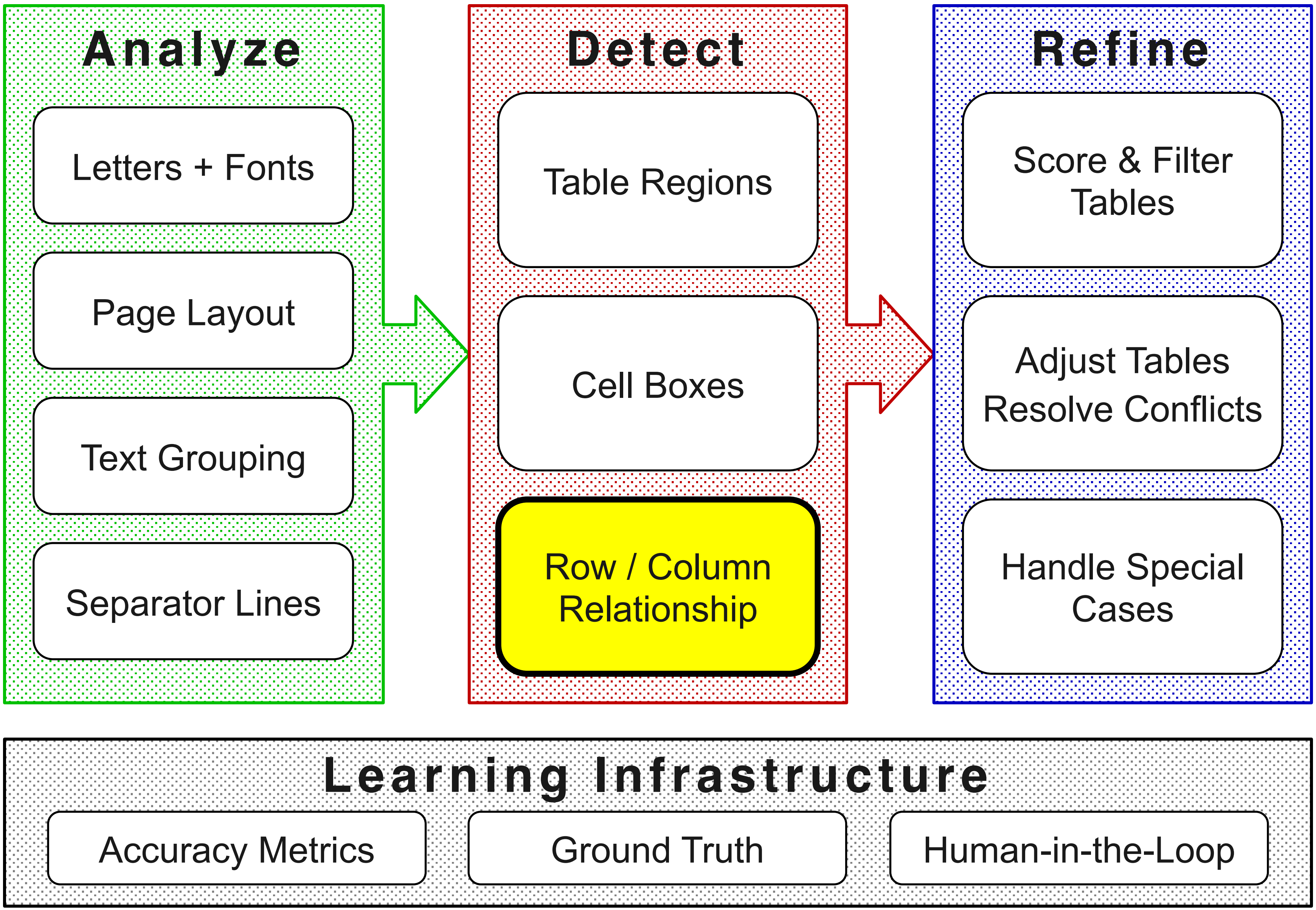
GTE Cell

Hierarchical deep learning system that pays attention to the global table style before cell detection



[Zheng et al. Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context, arXiv 2020]()

Common Sub-Tasks in Table Extraction



Cell Structure: Overview

- Cell structure defines:
  - Rows and Columns
  - Precedence order within each row and column
- Ways to specify cell structure:
  - **Separator lines:** Define cell spans across rows and columns
  - **Graphs over cells:** Define same-row and same-column relations
  - **Cell boxes:** Define row and column spans for each cell
  - **Text based:** Define cell structure using structured code output,
    - Such as HTML, XML

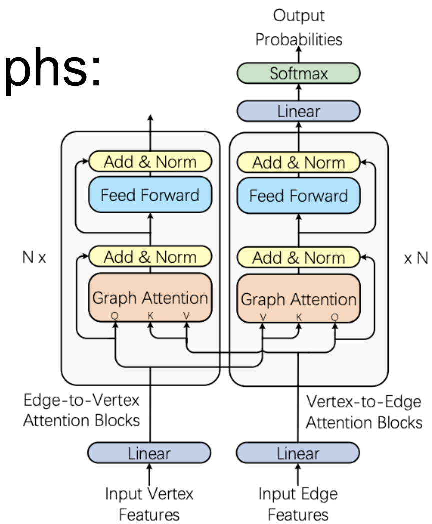
Cell Structure: Line Based

- Cell borders ← ruled lines ∪ "strong" white-space lines
  - **Extend lines** to make rectangular cells, avoid crossing "blobs"
- **Ruled-line grids:** test for incompleteness
  - Multiple numerics per cell
  - A "strong" white-space line splits text in ≥ 2 cells
  - A "mini-table" inside a ruled cell
  - Cell structure extends beyond table frame
- **White-space grids:** clean up empty cells
  - Expand header cells by merging with empty cells [S06]
  - Merge (almost-) empty rows and columns

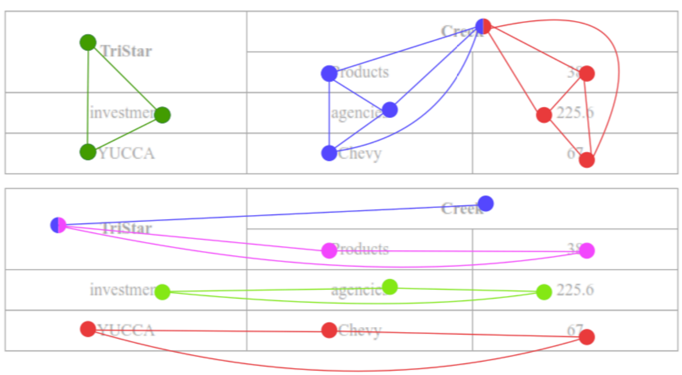
[S06] A. C. e Silva et al. "Design of an End-to-end Method to Extract Information from Tables", IJDAR '06
[B12] E. Bart. "Parsing Tables by Probabilistic Modeling of Perceptual Cues", DAS '12

Cell Structure: Graph Based

- Use **Spatial Constraints** to find an overlap DAG over cells [H03]
- Use **Graph Neural Networks** to find 2 *undirected* graphs:
  - "Same Row" graph & "Same Column" graph
  - Two cells share an edge ⇔ share a row / a column
  - [Q19] : Rows and columns = **maximal cliques**
  - [C19] : Only adjacent cells share a graph edge



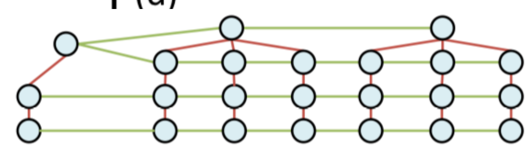
[Q19]



[C19]

Method	D1			D2		
	P	R	F1	P	R	F1
A	0.5	0.5	0.5	0.1	0.1	0.1
B	0.3	0.3	0.3	0.2	0.2	0.2

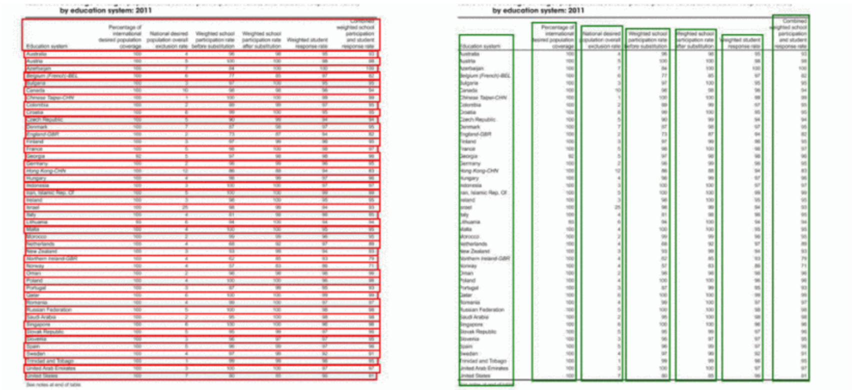
(d)



[H03] M. Hurst. "A Constraint-based Approach to Table Structure Derivation", ICDAR '03
[Q19] S. R. Qasim et al. "Rethinking Table Recognition using Graph Neural Networks", 2019
[C19] Z. Chi et al. "Complicated Table Structure Recognition", arXiv, 2019
[L20] Y Li et al. "GFTE: Graph-based Financial Table Extraction", arXiv, 2020

Cell Structure: Vision Model Based

- **Object detection networks** were also used for cell structure detection [S17][T19][P20b]
- [V20] Use Conditional Generative Adversarial Network to approximate table form first and then xy-cut and genetic algorithm to refine.
- [K19] Treat image as series of timesteps and use gated recurrent neural networks to determine column and row separation points.

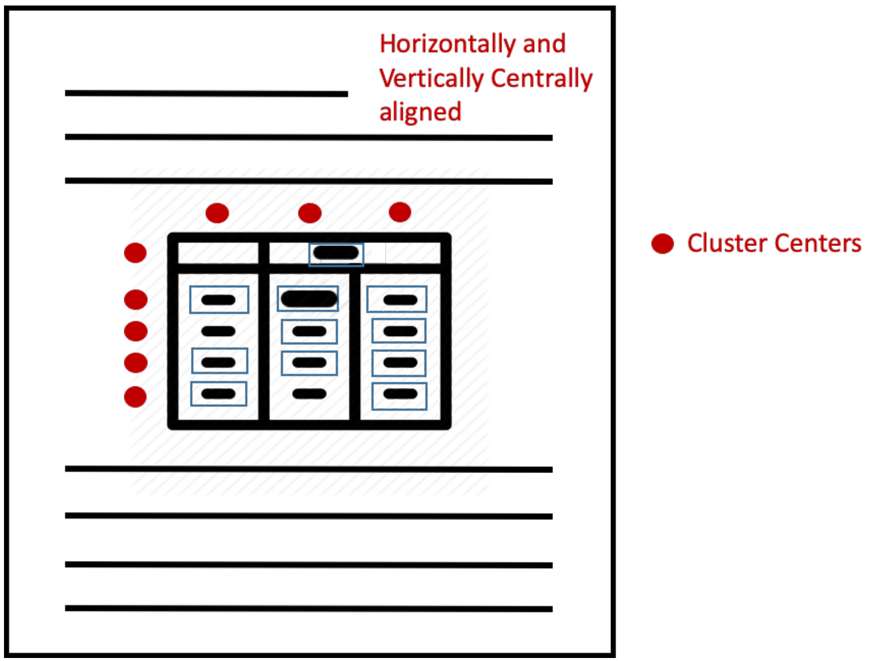


(a) Row detection, no ruling lines present (b) Column detection, no ruling lines present

[P20b] Paliwal et al. TableNet: Deep Learning model for end-to-end Table detection and Tabular data extraction from Scanned Document Images arXiv 2020
[S17] Schreiber et al. "Deepdesrt: Deep learning for detection and structure recognition of tables in document images" ICDAR 2017
[T19] Tensmeyer et al. "Deep splitting and merging for table structure decomposition" ICDAR 2019
[V20] Le Vine et al. Identifying Table Structure in Documents using Conditional Generative Adversarial Networks, arXiv 2020
[K19] Khan et al. "Table Structure Extraction with Bi-directional Gated Recurrent Unit Networks" ICDAR 2019

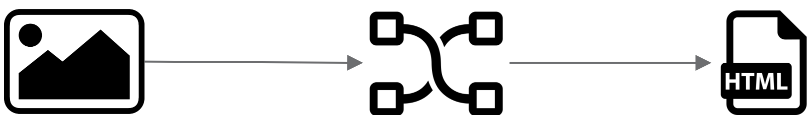
Cell Structure: Spatial clustering of cell units with language post-processing (GTE)

1. Cluster detected cells into rows and columns based on x-y coordinate and detected alignment.
2. Merge and split result based on textual clues (capitalization, special symbols etc.)



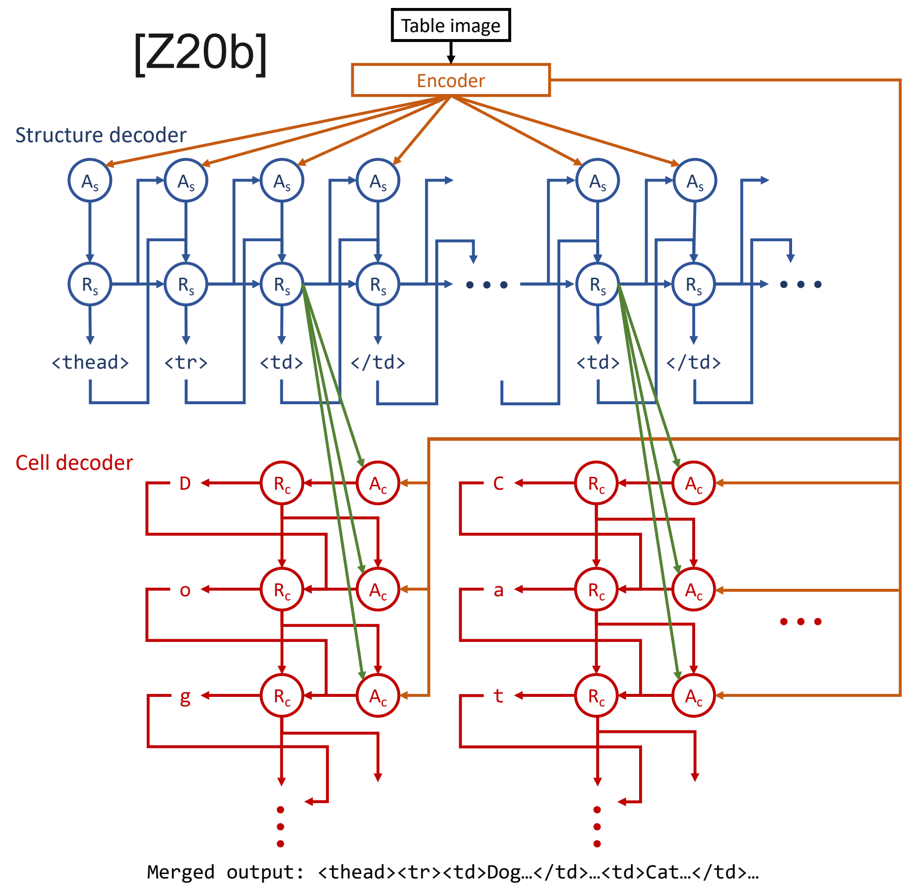
[Zheng et al. Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context, arXiv 2020]

Cell Structure: Language Generation Based



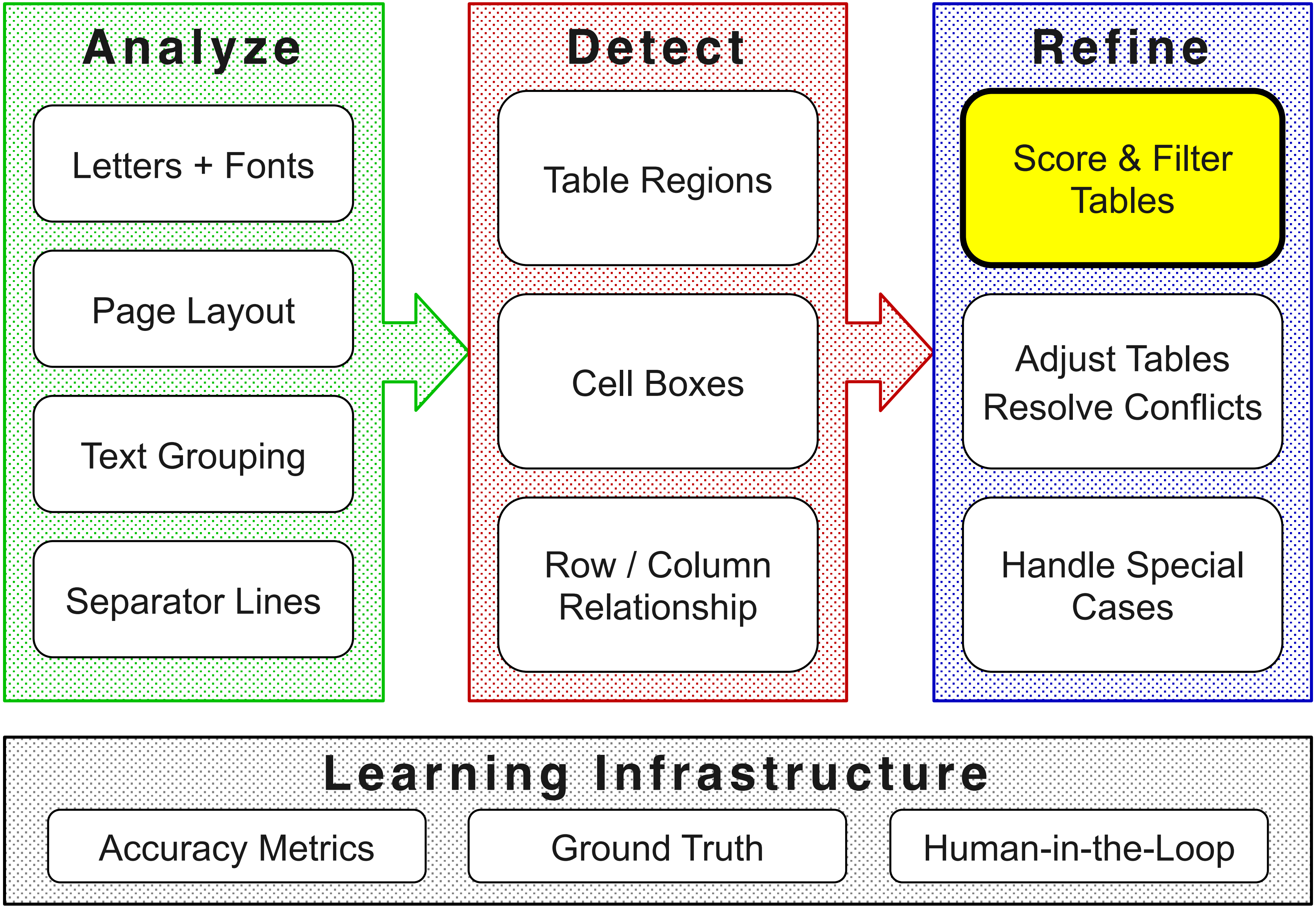
Recurrent neural network to encode image and then decode text of html representation of image

[Z20b]



[L20] Li et al. "TableBank: Table Benchmark for Image-based Table Detection and Recognition". LREC 2020
[Z20b] Zhong et al. "Image-based table recognition: data, model, and evaluation", ECCV 2020

Common Sub-Tasks in Table Extraction



Why Scoring Tables?

- Eliminate false positive tables
- Detect (and fix) malformed table regions
  - Plain text in tables
  - Missing row / column headers or split-off pieces
  - One region covers multiple tables
- Compare alternative table candidates
  - Example: Is this 1 table or 2 tables?
- Improve table region and structure
  - Pick the best adjustment out of a range of options
  - Given cell structure, fix table region

Table Scoring Challenges

- Tables are **very** diverse
  - Tiny or huge, misaligned, text in cells, key-value pairs, confusing delimiters
  - Complex row / column headers – so different, **easy to chop off !**
- What's **around** the table also matters
  - Can its columns or rows be extended? Should they be?
- One table, or ≥ 2 adjacent tables?
  - **1 table may have:** ruled bars, wide gaps, font / alignment changes
  - **2 tables may be:** fully or partly co-aligned, separated in one of many ways
- Non-table text can have structure, too
  - Page headers / footers, framed / highlighted text, hierarchical lists, …

Example 1

NOT A TABLE !

Part 2 Financial claims scheme

4AA Support that is not external support

(1) For subsection 11CA (1C) of the Act, a form of support that is entered into in the normal course of business is not to be considered external support for the purposes of subsection 11CA (1B) of the Act.

(2) For subsection 13A (1A) of the Act, a form of support that is entered into in the normal course of business is not to be considered external support for the purposes of paragraph 13A (1) (b) of the Act.

(3) For subsection 13E (3) of the Act, a form of support that is entered into in the normal course of business is not to be considered external support for the purposes of paragraph 13E (1) (b) of the Act.

4A Clearance period

For subsection 16AF (1) of the Act, 5 business days is the prescribed period of clearance.

5 Financial claims scheme — limit on payments

(1) For subsection 16AG (1) of the Act, a limit of $1 000 000 is prescribed.

(2) For the purpose of determining the prescribed limit on the payments to the account-holder, if the amount held in the account is expressed as a foreign currency, it must be converted to Australian dollars using the daily exchange rate published by the Reserve Bank of Australia.

Table Source: https://www.legislation.gov.au/Details/F2010C00607/0d99393c-5c5b-4af0-9cc1-b5c2de8632c3 (F2010C00607.pdf)

Example 2

Row headers

Column headers

A summary of the impact of these items on EPS is as follows:

(in millions, except per share data)	Pre-Tax Income/(Loss)	Tax Benefit/ (Expense)[1]	After-Tax Income/(Loss)	EPS Favorable/ (Adverse)[2]
Year Ended September 29, 2018:				
Net benefit from the Tax Act	$ —	$ 1,701	$ 1,701	$ 1.11
Gain from sale of real estate, property rights and other	601	(158)	443	0.30
Impairment of equity investments	(210)	49	(161)	(0.11)
Restructuring and impairment charges	(33)	7	(26)	(0.02)
Total	$ 358	$ 1,599	$ 1,957	$ 1.28
Year Ended September 30, 2017:				
Settlement of litigation	$ (177)	$ 65	$ (112)	$ (0.07)
Restructuring and impairment charges	(98)	31	(67)	(0.04)
Gain related to the acquisition of BAMTech	255	(93)	162	0.10
Total	$ (20)	$ 3	$ (17)	$ (0.01)
Year Ended October 1, 2016:				
Vice Gain	$ 332	$ (122)	$ 210	$ 0.13
Restructuring and impairment charges	(156)	43	(113)	(0.07)
Infinity Charge[3]	(129)	47	(82)	(0.05)
Total	$ 47	$ (32)	$ 15	$ 0.01

Table Source: https://www.thewaltdisneycompany.com/wp-content/uploads/2019/01/2018-Annual-Report.pdf

Example 3

Depreciation expense is as follows:

(in millions)	2018	2017	2016
Media Networks			
Cable Networks	$ 172	$ 137	$ 147
Broadcasting	92	88	90
Total Media Networks	264	225	237
Parks and Resorts			
Domestic	1,410	1,336	1,273
International	742	660	445
Total Parks and Resorts	2,152	1,996	1,718
Studio Entertainment	55	50	51
Consumer Products & Interactive Media	69	63	63
Corporate	218	252	251
Total depreciation expense	$ 2,758	$ 2,586	$ 2,320

Column headers

Row headers

Amortization of intangible assets is as follows:

(in millions)	2018	2017	2016
Media Networks	$ 62	$ 12	$ 18
Parks and Resorts	4	3	3
Studio Entertainment	64	65	74
Consumer Products & Interactive Media	123	116	112
Total amortization of intangible assets	$ 253	$ 196	$ 207

Table Source: https://www.thewaltdisneycompany.com/wp-content/uploads/2019/01/2018-Annual-Report.pdf

Example 4

Column headers

Row headers

As at, or for the 12-month periods ended, March 31 ($ in millions)	Objective	2019	2018
Components of debt and coverage ratios			
Net debt [1]		**$ 15,732**	$ 13,785
EBITDA – excluding restructuring and other costs [2]		**$ 5,533**	$ 5,091
Net interest cost [3]		**$ 660**	$ 582
Debt ratio			
Net debt to EBITDA – excluding restructuring and other costs	2.00 – 2.50 [4]	**2.84**	2.71
Coverage ratios			
Earnings coverage [5]		**4.3**	4.8
EBITDA – excluding restructuring and other costs interest coverage [6]		**8.4**	8.8

1 Net debt is calculated as follows:

As at March 31	*Note*	2019	2018
Long-term debt	*26*	**$ 15,775**	$ 13,990
Debt issuance costs netted against long-term debt		**90**	75
Derivative (assets) liabilities, net		**41**	59
Accumulated other comprehensive income amounts arising from financial instruments used to manage interest rate and currency risks associated with U.S. dollar-denominated long-term debt – excluding tax effects		**(86)**	(24)
Cash and temporary investments, net		**(588)**	(415)
Short-term borrowings	*22*	**500**	100
Net debt		**$ 15,732**	$ 13,785

2 EBITDA – excluding restructuring and other costs is calculated as follows:

	EBITDA (*Note 5*)	Restructuring and other costs (*Note 16*)	EBITDA – excluding restructuring and other costs
Add			
Three-month period ended March 31, 2019	$ 1,379	$ 36	$ 1,415
Year ended December 31, 2018	5,104	317	5,421
Deduct			
Three-month period ended March 31, 2018	(1,269)	(34)	(1,303)
EBITDA – excluding restructuring and other costs	$ 5,214	$ 319	$ 5,533

Table Source:
https://assets.ctfassets.net/rz9m1rynx8pv/2x3p5ompzZyrRtAHw4M3XB/be648275661795139cabcee29a730630/TELUS_Q1_2019_quarterly_report.pdf

How to Score a Table

- Rule-out patterns
  - Rule out charts, lists, signature blocks etc.
- Aggregated column / row score
  - [KD01] Aggregate the similarities that led to the table's column fragments
- Dynamic programming score
  - [H99] Score (T) = max { Score (T – line) + Merit (line) }
  - Score the best split into 2 sub-tables
- Probability of being a table (given the features)
  - [W04] Partition page into blocks labeled "table" and "plain text"
  - Compute label probability for block + **neighboring blocks**
- A scoring neural network on top of CNN [G17, S18b]

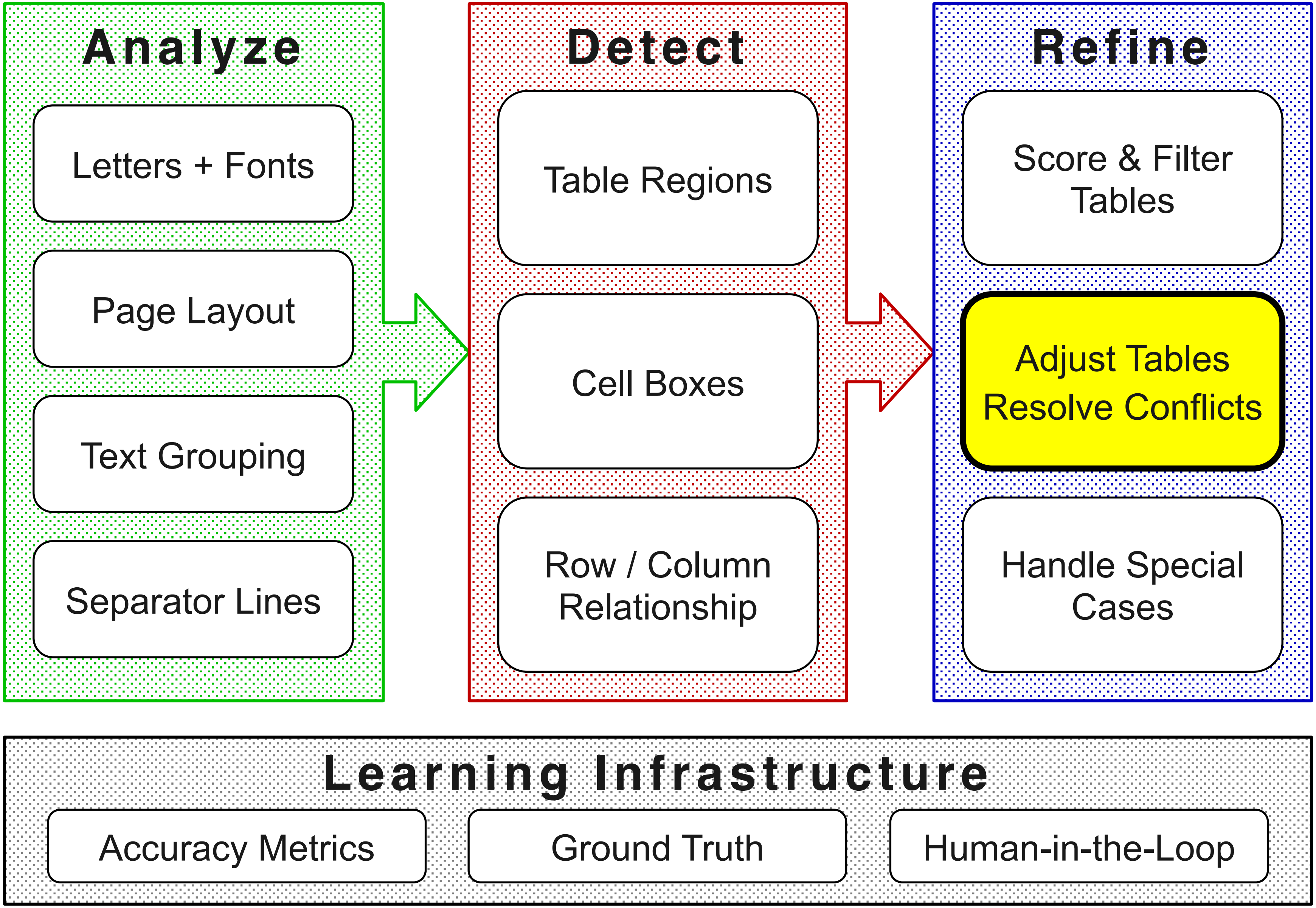
[H99] J. Hu et al. "Medium-Independent Table Detection", SPIE Doc. Recog. & Retr. '99
[KD01] T. Kieninger and A. Dengel. "Applying the T-Recs Table Recognition System to the Business Letter Domain", ICDAR '01
[W04] Y. Wang et al. "Table Structure Understanding and Its Performance Evaluation", Pattern Recog. '04
[G17] A. Gilani et al. "Table Detection using Deep Learning", ICDAR '17
[S18b] S. A. Siddiqui et al. "DeCNT: Deep Deformable CNN for Table Detection", IEEE Acc. '18

Features for Table Scoring

- Columns and rows:
  - Number, span / extent, alignment, font / content similarity
- Ruled and white-space separators:
  - Number, span / extent, width of their margins
  - If they match, reach (*good*) or cross (*bad*) table borders
- Inside vs. outside table:
  - Border crossing ruled lines, aligned blocks, or highly similar text
  - The two sides have matching structure
- Cell structure:
  - Oversized cells, misaligned pairs of cells, “runs” of empty cells
- Content:
  - Numerics, repeated words; customizable keywords
  - Domain-specific “expectations,” e.g. header dictionary [D11]
- CNN-generated features

[D11] F. Deckert et al. “Table Content Understanding in smartFIX”, ICDAR ‘11

Common Sub-Tasks in Table Extraction



Why Adjust Tables?

- Leverage table features and score
  - Specify how a well-formed *vs.* mal-formed table looks like
- Use a transparent, explainable method
  - If detection is a “black box”, adjustment uses explainable rules & features
- Correct errors quickly
  - Bypass the need for extra ground-truth data, retraining
- Customize to address specific concerns
  - Add custom features, rules, and constrains

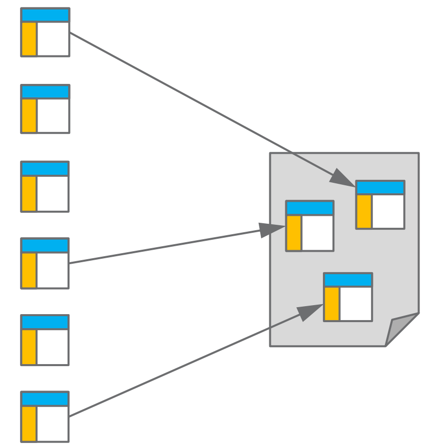
[W04] Y. Wang et al. “Table Structure Understanding and Its Performance Evaluation”, Pattern Recog. ‘04
[HB07] T. Hassan and R. Baumgartner. “Table Recognition and Understanding from PDF Files”, ICDAR ‘07
[SS10] F. Shafait and R. Smith. “Table Detection in Heterogeneous Documents”, DAS ‘10
[D11] F. Deckert et al. “Table Content Understanding in smartFIX”, ICDAR ‘11
[G17] A. Gilani et al. “Table Detection using Deep Learning”, ICDAR ‘17
[S18b] S. A. Siddiqui et al. “DeCNT: Deep Deformable CNN for Table Detection”, IEEE Acc. ‘18

How to Adjust Candidate Tables

- **Merge table** with an adjacent table or text-block [W04] [SS10]
- **Adjust table border** – add or drop rows or columns [HB07] [D11]
- **Split one table into two**, possibly with plain text between
- **Re-compute table region** by neural network regression [G17] [S18b]
- **Choose best-scoring** border (or structure) out of a range of options
- Iterate adjustment → **traverse a search tree** of candidate tables

[W04] Y. Wang et al. "Table Structure Understanding and Its Performance Evaluation", Pattern Recog. '04
[HB07] T. Hassan and R. Baumgartner. "Table Recognition and Understanding from PDF Files", ICDAR '07
[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
[G17] A. Gilani et al. "Table Detection using Deep Learning", ICDAR '17
[S18b] S. A. Siddiqui et al. "DeCNT: Deep Deformable CNN for Table Detection", IEEE Acc. '18

Select Best Tables for Output

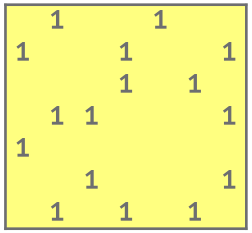


What if candidate tables overlap each other?

- [H99] uses **Dynamic Programming:**
  - Only for top and bottom line-positions: [i, j]
  - Score disjoint unions of tables:

$$score[i,j] = \max \begin{cases} tab[i,j] \\ \max_{i \le k < j} \{score[i,k] \; + \; score[k+1,j]\} \end{cases}$$

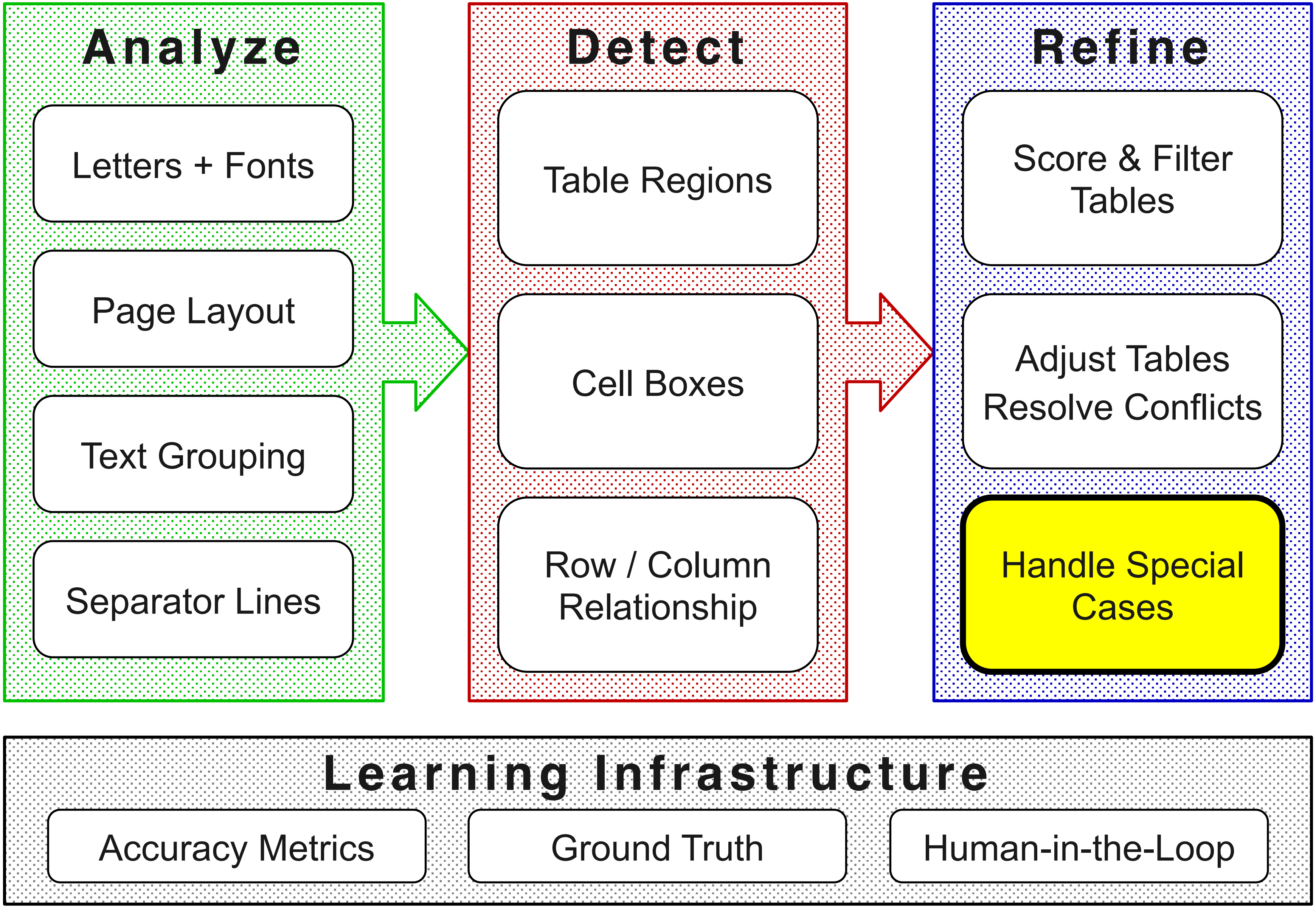
- CNN-based object detection systems:
  - **Greedy Approach:** Pick the top-scoring region, repeat
  - PROBLEM: Lower-scoring table may have a high-scoring sub-table
- **Maximum Weighted Independent Set**
  - Nodes = tables, edges = conflicts, weights = table scores
  - NP-hard even for 2-dim rectangles [RN95], but can be solved efficiently in real-life cases



Conflict = Table Overlap

[H99] J. Hu et al. “Medium-Independent Table Detection”, SPIE Doc. Recog. & Retr. ‘99
[RN95] C.S. Rim and K. Nakajima. “On Rectangle Intersection and Overlap Graphs”, IEEE Trans. on Circuits & Systems I, 42(9), 1995

Common Sub-Tasks in Table Extraction

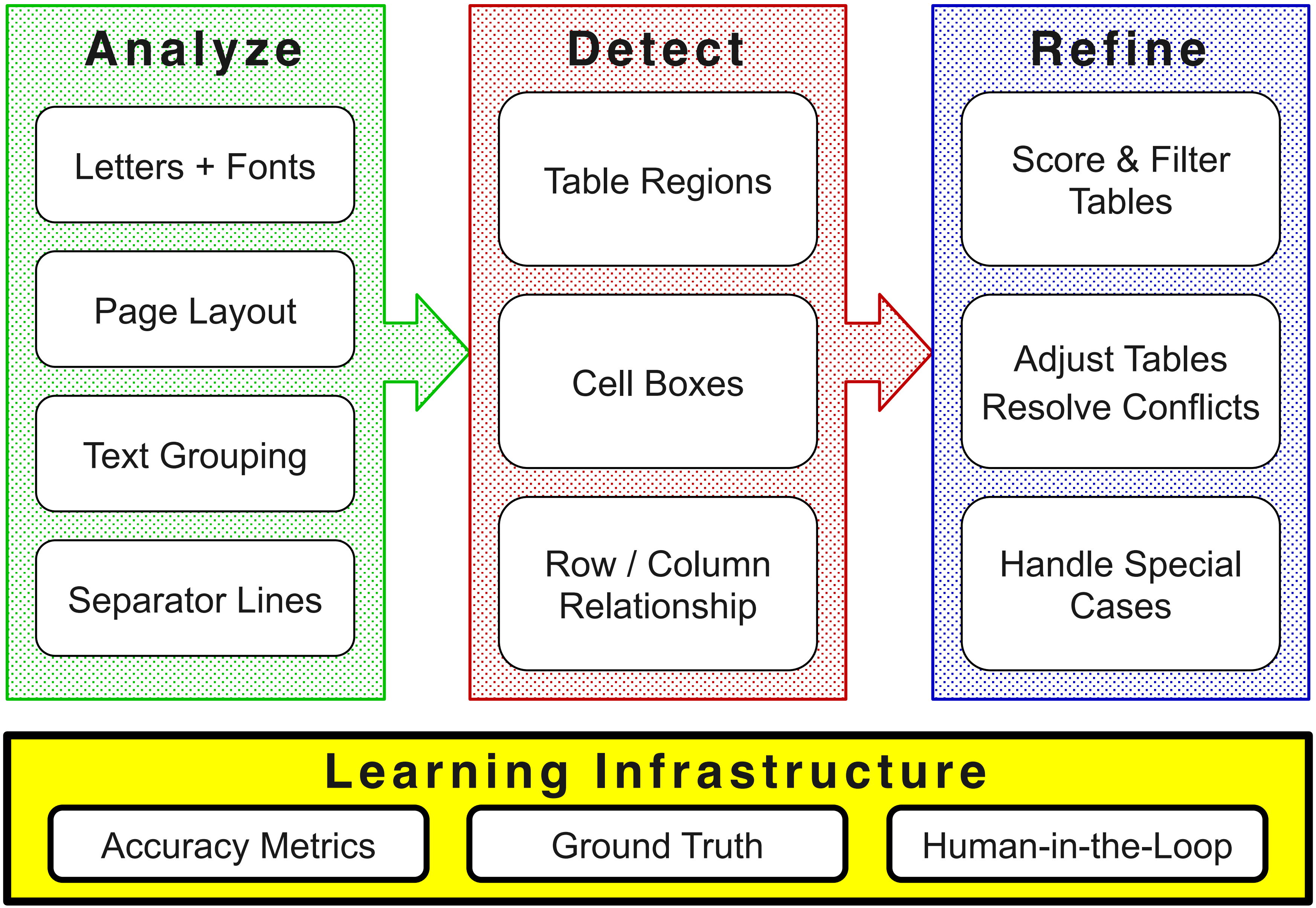


Handle Customer Specific Rules and Forms

- Customers need ~100% accuracy on specific tables
  - Invoices & financial reports
  - Healthcare forms
  - Contracts, insurance and legal documents
- Customers may only label a few examples
  - Not enough to learn a new ML / DL model
  - Learning a new model may jeopardize older correct results
- Customers want to see how decisions are made
  - Explain how a certain table is handled
  - Provide a guarantee for a (narrow) class of tables
- **Solution:** Refine results with a human readable ruleset

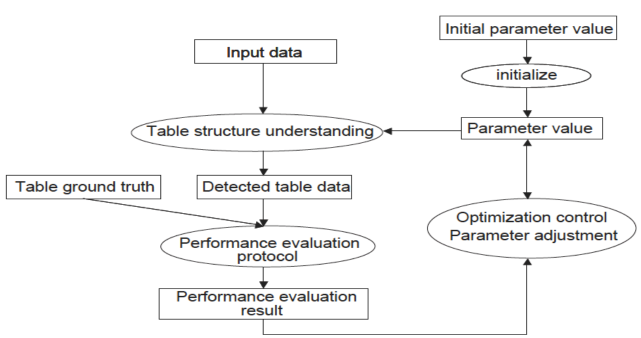
[K01] B. Klein et al. "Three Approaches to Industrial Table Spotting", ICDAR '01
[D11] F. Deckert et al. "Table Content Understanding in smartFIX", ICDAR '11

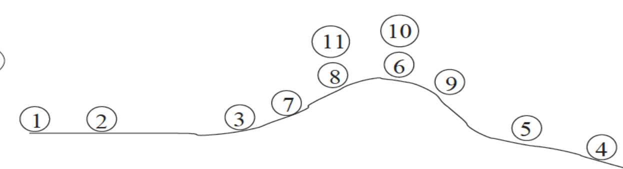
Common Sub-Tasks in Table Extraction



Learning from Data: Challenges

- **Accuracy Metrics**
  - Exact match of table region or structure is too inflexible
  - Partial match: Text? Area? Cell relationship? Functional?
- **Ground Truth Labeling**
  - Very time consuming, requires sophisticated UI tools
  - Humans disagree on what's correct
- **Optimization** (pre- deep learning)
  - Lots of discrete, non-differentiable steps
  - **Learn sub-tasks**, e.g. row labeling with CRF / SVM
  - [W04] Global parameter learning:





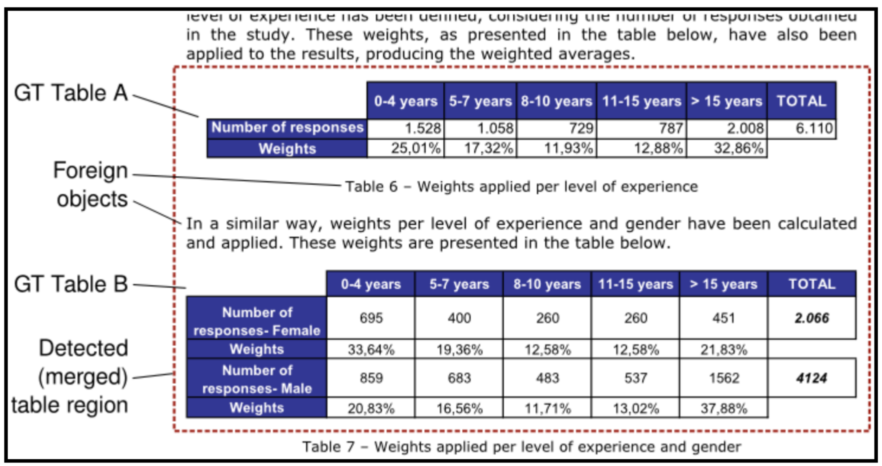
[W04] Y. Wang et al. "Table Structure Understanding and Its Performance Evaluation", Pattern Recog. '04

Accuracy Metrics

ICDAR 2013 Competition Metrics

Table Boundary

- Purity & Completeness
- Character level recall, precision and F1



level of experience has been defined, considering the number of responses obtained in the study. These weights, as presented in the table below, have also been applied to the results, producing the weighted averages.

	0-4 years	5-7 years	8-10 years	11-15 years	> 15 years	TOTAL
Number of responses	1.528	1.058	729	787	2.008	6.110
Weights	25,01%	17,32%	11,93%	12,88%	32,86%	

Table 6 – Weights applied per level of experience

In a similar way, weights per level of experience and gender have been calculated and applied. These weights are presented in the table below.

	0-4 years	5-7 years	8-10 years	11-15 years	> 15 years	TOTAL
Number of responses- Female	695	400	260	260	451	2.066
Weights	33,64%	19,36%	12,58%	12,58%	21,83%	
Number of responses- Male	859	683	483	537	1562	4124
Weights	20,83%	16,56%	11,71%	13,02%	37,88%	

Table 7 – Weights applied per level of experience and gender

Table Structure

- Recall and Precision of Cell Adjacency Relations

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

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

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

[G12] Göbel et al. “A Methodology for Evaluating Algorithms for Table Understanding in PDF Documents”. DocEng '12

Accuracy Metrics

ICDAR 2019 Competition Metrics

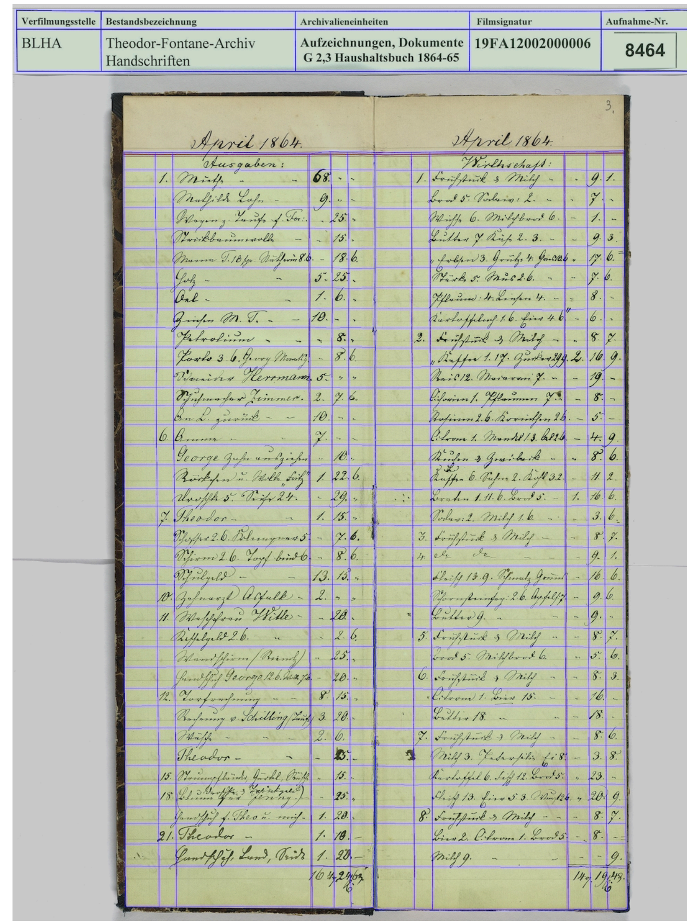
Two Document types, modern and archival, in image format only.

Table Boundary

Intersection over union (IOU) at varying thresholds (0.6,0.7,0.8,0.9) and weighted average comparing ground truth and predicted table bounding boxes

Table Structure

Adjacency relationship like ICDAR 2013 but cell accuracy is based on IOU of cell bounding boxes instead of text content.



[G19] Gao et al. “ICDAR 2019 Competition on Table Detection and Recognition (cTDaR)”, ICDAR ‘19

Accuracy Metrics

ICDAR 2020 Competition Metrics

Task A
Document layout recognition

- Dataset: PubLayNet
- Task: Identifying the position and category of document layout elements, including title, text, figure, table, and list.
- Metric: Mean Average Precision @ IoU
- Important dates:
  - **20th July, 2020:** Open for submission
  - **31st March, 2021**: Submission close
  - **1st May, 2021:** Announcement of winning team

Task B
Table Structure Recognition

- Dataset: PubTabNet
- Task: Converting table images into HTML code
- Metric: Tree-edit-distance-based similarity (TEDS)
- Important dates:
  - **20th July, 2020:** Open for test submission
  - **28th March, 2021:** Open for final evaluation submission
  - **31st March, 2021:** Submission close
  - **1st May, 2021:** Announcement of winning team

https://icdar2021.org/competitions/competition-on-scientific-literature-parsing/

Accuracy Metrics

Functional Metrics

- Measure **what actually matters** downstream
- Capture accuracy of access paths to each cell
- Need **header annotation** as well as cell structure

		Turnover ($bn)		
		2008	2009	2010
AA	American Airlines	17.5	18.1	17.2
AF	Air France	11.6	10.8	11.9
KL	KLM Royal Dutch Airlines	8.3	9.5	9.4
LH	Lufthansa	12.8	14.1	13.8
NA	New Airline		2.1	2.4

Functional representation:

[AA],[Turnover ($bn).2008] → [17.5],
[American Airlines],[Turnover ($bn).2008] → [17.5],
[AA],[Turnover ($bn).2009] → [18.1],
[American Airlines],[Turnover ($bn).2009] → [18.1],
. . . ,
[NA],[Turnover ($bn).2008] → [],

Göbel et al. “A Methodology for Evaluating Algorithms for Table Understanding in PDF Documents”. DocEng '12

Ground Truth Datasets

Complete Datasets with table boundary, cell boundary, and cell structure:

- ICDAR-2013 competition (PDF Format) [G12]
- ICDAR-2019 competition (Image Format) [G19]
- SciTSR 2019 (Generated from LaTeX files)[C09]
- PubXNet 2020 (PDF Format) [Z20a]
- FinTabNet 2020 (PDF Format) [Z20b]

Incomplete Datasets

- Table-bank (table boundary information and cell structure only)[L20]
- PubLayNet (table boundary information only)[Z19]
- PubTabNet (Cell structure information only)[Z20b]
- PDF-Trex (Financial Table dataset without ground truth Labels)[O09]
- Marmot (Only ground truth for table boundary, cells inaccessible)
- UNLV , UW-3 (Table structure and boundary annotations for scanned documents)

[C09] Chi et al. "Complicated Table Structure Recognition" arXiv 2019
[OR09] E. Oro and M. Ruffolo. "PDF-TREX: An Approach for Recognizing and Extracting Tables from PDF Documents", ICDAR '09
[G12] Göbel et al. "A Methodology for Evaluating Algorithms for Table Understanding in PDF Documents". DocEng '12
[L20] Li et al. "TableBank: Table Benchmark for Image-based Table Detection and Recognition". LREC 2020
[Z20a] Zheng et al. Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context, arXiv 2020
[Z19] Zhong et al. **Publaynet**: largest dataset ever for document layout analysis, ICDAR2019
[Z20b] Zhong et al. Image-based table recognition: data, model, and evaluation, ECCV 2020
[G19] Gao et al. Icdar 2019 competition on table detection and recognition(ctdar), ICDAR2019

Table Annotation

- Labeling ground truth tables & cells is labor-intensive [W04]
- **Manual annotation:** requires
  - Sophisticated user interface tool [FK15] [HL19] [Z20a]
  - Lots of time and human labor
  - Detailed agreement on how to handle ambiguous cases
- **Automated annotation:** requires
  - HTML and PDF versions of the same documents
  - An automated text matching algorithm [Z20a]
  - Manual editing to fix matching errors (much less labor)

[W04] Y. Wang et al. "Table Structure Understanding and Its Performance Evaluation", Pattern Recog. '04
[FK15] M. Frey and R. Kern. "Efficient Table Annotation for Digital Articles", D-Lib Mag. '15
[HL19] J. Hoffswell and Z. Liu. "Interactive Repair of Tables Extracted from PDF Documents on Mobile Devices", CHI '19
[Z20a] Zheng et al. Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context, arXiv 2020