

Unlocking the Potential of Unstructured Data in Finance Through Document Intelligence

Himanshu S. Bhatt

Emerging AI Data Products & Services, American Express AI Labs



Credit and
Fraud Risk

AI Research @ American Express AI Labs

Vision Statement

Develop cutting edge AI/ML capabilities *to drive growth in enterprise products/platforms* or *to drive efficiencies in enterprise processes* while *pushing the state-of-the-art*.

Research Areas

Natural
Language
Processing

Document AI

Machine
Learning
Algorithms

Ethical
AI

Document AI Team @ Amex AI Labs



Saikiran Peketi



Dhruv Premi



Rohit Bhogade



Tarun Kumar



Tamanna Agarwal



Chinesh Doshi

Agenda

- What is Document Intelligence?
- Information Extraction from Various Document Types
- Overview of Extraction Approaches
- Use Case Highlight – Bank Statements, Digital Auditor
- Future Research Directions

What is Document Intelligence?

Quick Primer on Document Intelligence

It allows us to tap into the opportunities offered by unstructured document data and unlock the potential for:



How does it work?



1. Extract what's there



2. Understand it



3. Make it useful

Documents available in Financial Industry

Structured

Application Forms

Tax/Claim Forms

Semi-Structured

Bank Statements

Balance sheet example

TEDDY FAB INC.		BALANCE SHEET	
December 31, 2100			
ASSETS		LIABILITIES AND SHAREHOLDERS' EQUITY	
Current assets		Current liabilities	
Cash and cash equivalents	\$ 100,000	Accounts payable	\$ 30,000
Accounts receivable	20,000	Notes payable	10,000
Inventory	15,000	Accrued expenses	5,000
Prepaid expense	4,000	Deferred revenue	2,500
Investments	10,000	Total current liabilities	47,000
Total current assets	149,000	Long-term debt	200,000
Property and equipment		Total liabilities	247,000
Land	24,300	Shareholders' Equity	
Buildings and improvements	250,000	Common stock	10,000
Equipment	50,000	Additional paid-in capital	20,000
Less accumulated depreciation	(5,000)	Retained earnings	197,100
Other assets		Treasury stock	(2,000)
Intangible assets	4,000	Total liabilities and shareholders' equity	\$ 472,100
Less accumulated amortization	(200)		
Total assets	\$ 472,100		

Balance Sheets

Invoices

Contracts

Policies

IDs

Passport

Receipts

Financial / Earnings Reports

Marketing Creatives

Unstructured

Sample Enterprise Use Cases & Types of Documents

Efficiency gained by AI powered document intelligences fuels revenue growth or helps in loss savings

Form Type

[Company Name]

[Street Address]
[City, ST ZIP]
Phone: [000] 000 0000

INVOICE

INVOICE # DATE
2034 2/17/2018

CUSTOMER ID TERMS
584 Due Upon Receipt

DESCRIPTION	QTY	UNIT PRICE	AMOUNT
Service Fee	1	289.99	289.99
Labor 1/ hour at 37.1/hr	5	75.99	375.95
New client discount		(24.99)	(24.99)
SUBTOTAL			525.90
TAX RATE			4.250%
TAX			22.31
TOTAL			\$ 547.31

FIRST BANK OF WIKI
1425 JAMES ST, PO BOX 4000
VICTORIA BC V8X 3X4 1-800-555-5555

CHEQUING ACCOUNT STATEMENT
Page : 1 of 1

JOHN JONES
1643 DUNDAS ST W APT 27
TORONTO ON M6K 1V2

Statement period Account No.
2003-10-09 to 2003-11-08 00005-123-456-7

Date	Description	Ref.	Withdrawals	Deposits	Balance
2003-10-08	Previous balance				0.55
2003-10-14	Payroll Deposit - HOTEL			694.81	695.36
2003-10-14	Web Bill Payment - MASTERCARD	9685	200.00		495.36
2003-10-16	ATM Withdrawal - INTERAC	3990	21.25		474.11
2003-10-16	Fees - Interac		1.50		472.61
2003-10-20	Interac Purchase - ELECTRONICS	1975	2.99		469.62
2003-10-21	Web Bill Payment - AMEX	3314	300.00		169.62
2003-10-22	ATM Withdrawal - FIRST BANK	0064	100.00		69.62
2003-10-23	Interac Purchase - SUPERMARKET	1559	29.08		40.54
2003-10-24	Interac Refund - ELECTRONICS	1975		2.99	43.53
2003-10-27	Telephone Bill Payment - VISA	2475	6.77		36.76
2003-10-28	Payroll Deposit - HOTEL			694.81	731.57
2003-10-30	Web Funds Transfer - From SAVINGS	2620		50.00	781.57
2003-11-03	Pre-Auth. Payment - INSURANCE		33.55		748.02
2003-11-03	Cheque No. - 409		100.00		648.02
2003-11-06	Mortgage Payment		710.49		-62.47
2003-11-07	Fees - Overdraft		5.00		-67.47
2003-11-08	Fees - Monthly		5.00		-72.47
*** Totals ***			1,515.63	1,442.61	

<https://upload.wikimedia.org/wikipedia/commons/c/cb/BankStatementChequing.png>

- Understand spend pattern from invoices
- Get cash-flow insights from bank statements

example only – not real data

Verbose Type

My Account Cards Travel Insurance Rewards Business Help Log In

CONSTRUCTION CONTRACT AGREEMENT

This Construction Contract Agreement (the "Agreement") is made as of the 17 day of January, 2018 by and between Anthony C. Cummings, an individual located at 4900 DeHaven Road, San Francisco, CA 94109 ("Owner") and David C. Ortiz an individual located at 2318 Garfield Road, Phoenix, AZ 85004 ("Contractor"). Owner and Contractor may each be referred to in this Agreement individually as a "Party" and collectively as the "Parties."

WHEREAS, Contractor is a duly licensed general contractor in the State of Illinois, in good standing, with contractor's license number 1234567; and

WHEREAS, Owner owns the property located at 3790 Carriage Lane, Greendale, PA 17935 (the "Property") and desires to have certain work performed by Contractor at the Property;

NOW, THEREFORE, in consideration of the mutual promises and for other good and valuable consideration exchanged by the Parties as set forth in this Agreement, the Parties, intending to be legally bound, hereby mutually agree as follows:

- 1. Description of Work.** Contractor shall perform the work described in Exhibit A (the "Work"), in accordance with Owner's contract plans and specifications, this Agreement and any Change Order, as defined herein, (collectively, the "Contract Documents") at the Property.
- 2. Contract Price and Payments.** Owner agrees to pay Contractor for the Work the total amount of \$10,000.00 USD (the "Contract Price"). Payment of this amount is subject to additions or deductions in accordance with any mutually agreed to changes and/or modifications in the Work, and the other documents to which this Agreement is subject. Payment for the Work will be by wire transfer, according to the following schedule:
 - \$10,000.00 balance due upon completion of the Work.
- 3. Certificate of Completion.** Upon completion of the Work, Contractor shall notify Owner that the Work is ready for final inspection and acceptance. When Owner finds the Work acceptable and this Agreement fully performed, Contractor shall issue Owner a "Certificate of Completion" stating that the Work has been completed in accordance with the Contract Documents and the entire balance of the Contract Price is due and payable. Owner shall make the final payment within 7 days after receiving a Certificate of Completion. Owner by making final payment waives all claims except those arising out of: (a) any faulty Work appearing after completion; (b) any Work that does not comply with the Contract Documents; and (c) outstanding claims or liens. Contractor, by accepting final payment, waives all claims except those previously made in writing, and which remain unsettled at the time of acceptance.
- 4. Materials and Labor.** Contractor shall provide and pay for all labor and equipment, including tools, construction equipment, machinery, transportation and all other facilities and services, and all materials necessary for the completion of the Work. All materials shall be good quality and new, unless the Contract Documents require or permit otherwise. Contractor may substitute materials only with the prior written approval of Owner.

Construction Contract Agreement (Rev. 12/24/2012) 1 / 6

<https://legaltemplates.net/form/construction-contract-agreement/>

- Review marketing creatives before campaign launch
- Highlight key clauses from contract documents

Information Extraction from Documents

Extraction Challenges in Verbose Documents

Critical Care Research and Practice 5

text A summary of the performance of eGFR equations in critically ill patients with AKI whose ⁵¹CrCrCl was less than 60 mL·min⁻¹ per 1.73 m² and whose urine output was greater than 0.2 mL·kg⁻¹ per min during the study period (37 patients).

	Creatinine	Cystatin C	aMDRD	MDRD ⁶	MDRD ⁷	Cystatin C ₂	Cystatin C ₃	Cystatin C ₄
Mean (SD) (mL·min ⁻¹ per 1.73 m ²)	27.2	33.4	35.5	33.3	35.5	28.8	32.3	43.2
Range (mL·min ⁻¹ per 1.73 m ²)	8-51	13-109	11-63	9-87	9-79	8-71	9-88	17-85
r ² correlation (P < 0.0001)	0.64	0.82	0.72	0.75	0.71	0.70	0.71	0.70
Bias (1.96 × SD)	-26.3	-8.4	-6.2	-5.4	-1.6	-5.2	-16.1	-13.9
Percentage error (precision)	52	39	56	47	58	57	46	47
Accuracy (%)								
30%	3	36	36	27	36	24	31	36
50%	5	46	57	49	57	57	50	50
70%	22	68	78	76	86	81	66	79

text which is perhaps more useful comparison hence its use in this analysis.

text recent study of eGFR performance in renal transplant patients [39] used Bland-Altman analysis and described the bias of CrCl, aMDRD, and MDRD 7 as 15.2, 9.2, and 7.4 and worse than this study. The precision (25.4%, 23.9%, and 10%, resp.) however, was better and within a range suggests previously [38]. The percentage of values within 30% of the ⁵¹CrCl (P₃₀) (37, 60, and 67.4, resp.) was comparable to the data from this study, and the use of the equations in renal transplant recipients is recommended.

text in introduced, the CKD-EPI equation [7] had a bias of 2.1 mL·min⁻¹ per 1.73 m² and a P₃₀ of 79.9%, which are better than data presented in this study and comparable only to the MDRD7 equation.

text g methods based on cystatin C when compared with methods incorporating serum creatinine have shown a higher correlation and improved accuracy in predicting GFR in patients with various degrees of renal function, liver disease, and spinal cord injuries [17]. However, results in patients with diabetes, paediatric patients, and those with early renal impairment did not show a significant difference between cystatin C and creatinine based eGFR, indicating that the performance may be patient population specific [40-43]. Human studies also suggest that cystatin C can predict the development of AKI [44] and the requirement for renal replacement therapy [45], although its superiority over serum creatinine has not been a universal finding [46].

text presented in this study demonstrate a very broad range of both ⁵¹CrCl and cystatin C measurement across each of the AKIN/RIFLE criteria. Figure 3 shows that serum cystatin C increased with worsening renal function measured by ⁵¹CrCl, but the correlation coefficient is not compelling and the confidence intervals are wide. When originally derived the equations which incorporate cystatin C showed minimal bias and excellent accuracy with P₃₀ of 81%, 85%, and 89% for cystatin C₂, C₃, and C₄ equations, respectively [16]. These

text were not reproduced in this study and the cystatin C equations actually performs worse than the original MDRD equations in patients with AKI.

s. Limitation: title

text ing rapid changes in renal function accurately in critically ill patients is difficult and there is no gold standard method. A useful, routine exogenous marker has remained elusive and there are well-described difficulties when interpreting creatinine clearance. Tubular secretion and extrarenal elimination of creatinine increases as GFR deteriorates thus exaggerating the discrepancy between the clearance of creatinine and true renal function [47]. In addition, serum creatinine concentrations are influenced by muscle mass, protein intake, gender, and age, limiting the precision further. The influence of these factors in the acute setting is not clear. However, over a period of hours and days, as the renal function deteriorates in AKI, one would anticipate that these other factors would remain relatively constant.

text re of its limitations, in the absence of an accepted gold standard, the ⁵¹CrCl was piloted as a baseline standard. It incorporates both changes in creatinine and urine output and is supported by an evidence base. A small study of eighteen critically ill patients used correlation coefficients to compare clearance of DTPA or insulin (their gold-standard measure) to 2-hour creatinine clearance (⁵¹CrCl) [48]. The authors conclude that a ⁵¹CrCl is not an accurate descriptor of insulin clearance, in this population, when the GFR is <30 mL·min⁻¹. However, reanalysis of the published raw data reveals a correlation coefficient (r) between DTPA and 2-hour creatinine clearance of 0.92 (P < 0.001) though this is not discussed in the original paper. Perhaps the more encouraging conclusion should include the close relationship with DTPA clearance. There is no mention of urine volume during the study time period and patients with very low DTPA clearances (2 mL·min⁻¹) were included.

	Years ended*		
	September 29, 2018	September 30, 2017	September 24, 2016
Net sales	\$ 205,595	\$ 229,234	\$ 215,639
Cost of sales	163,756	141,048	131,376
Gross margin	101,839	88,186	84,263
Operating expenses:			
Research and development	14,236	11,581	10,045
Selling, general and administrative	16,705	15,261	14,194
Total operating expenses	30,941	26,842	24,239
Operating income	70,898	61,344	60,024
Other income/(expense), net	2,005	2,745	1,348
Income before provision for income taxes	72,903	64,089	61,372
Provision for income taxes	13,372	15,738	15,685
Net income	\$ 59,531	\$ 48,351	\$ 45,687

Financial Statement

#Country	Average number of authors per publication by countries' groups (n=2330)	Number of articles (%)					Total
		Journal article	Review	Clinical trial	Case report	Others	
Kingdom of Saudi Arabia	2.94	814 (74.5)	68 (6.2)	35 (3.2)	156 (14.3)	20 (1.8)	1093 (100)
Other GCC countries	3.08	183 (79.7)	11 (4.2)	5 (1.9)	60 (23.2)	0 (0)	259 (100)
Arab and African countries	2.9	306 (79.7)	8 (2.1)	31 (8.1)	38 (9.9)	1 (0.3)	384 (100)
Asian countries	3.27	39 (66.1)	1 (1.7)	2 (3.4)	17 (28.8)	0 (0)	59 (100)
Iran	3.54	91 (84.3)	1 (0.9)	9 (8.3)	7 (6.5)	0 (0)	108 (100)
Turkey	4.51	257 (80.8)	2 (0.6)	11 (3.5)	47 (14.8)	1 (0.3)	318 (100)
West countries, South America and Japan	3.07	83 (76.1)	12 (11)	0 (0)	14 (12.8)	0 (0)	109 (100)

Medical journal table

Tables have different types of cell formats

* - cell spanning three columns
- cell covering multiple lines

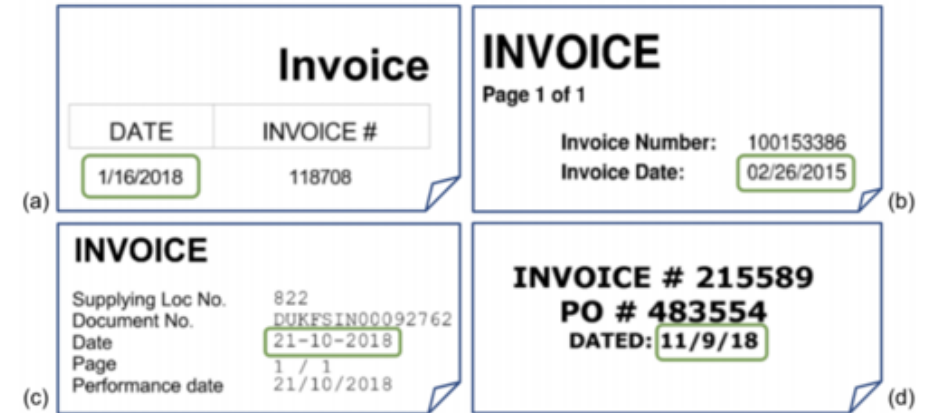
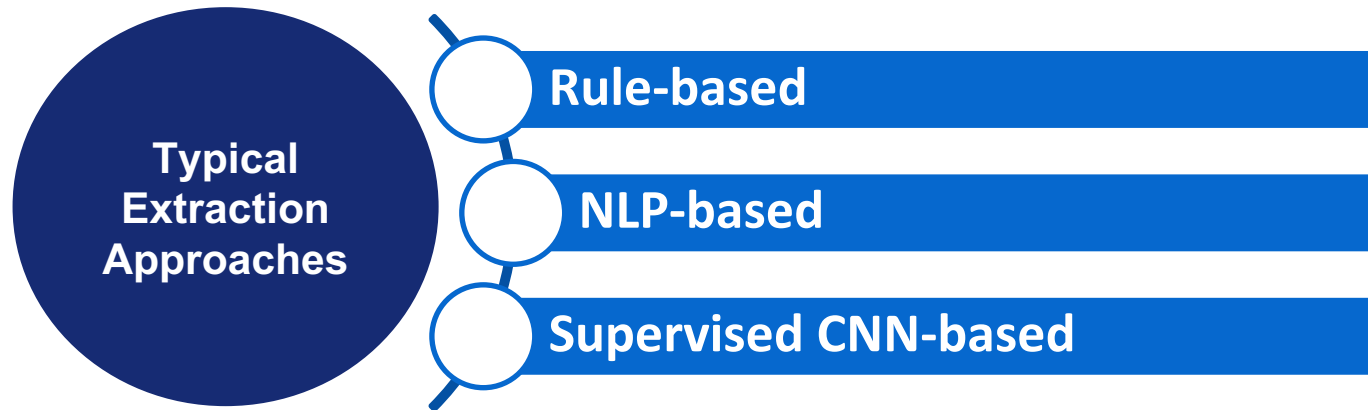
Simple rule-based or template-based approach will fail to extract table types that vary across different documents

Text Title

example only – not real data

Extraction Challenges in Form Type Documents

- Form type documents also have diverse templates



- Rule-based approaches can't handle unseen templates and are difficult to manage
- NLP-based approaches assign tags to each portion of the text while CNN-based approaches can capture irrespective of variations in templates
- Both NLP/CNN approaches have limitations for the cases where information is embedded in the spatial arrangement of the layout, not in the text itself

Impact of Extraction on Downstream Systems/Processes

[Company Name]

INVOICE

Street Address
City, ST or Prov
Phone (888) 888 8888

Invoice #	Invoice Date
10000001	09/15/20

Client Address
City, ST or Prov
Phone #
Fax #

Customer ID	Customer Name
10000001	10000000

Item Description	QTY	UNIT PRICE	AMOUNT
Accessories	4	200.00	800.00
Label - 3 Item x 2 1/4"	5	35.00	175.00
Label - 3 Item x 1 1/2"	5	20.00	100.00
Subtotal			1075.00
Tax (GST)			4.29%
Total			\$ 1,079.29

Automated payment/claim processing by understanding merchant details, line item description, pricing details etc.

FIRST BANK OF WIKI
1425 JAMES ST. PO BOX 4000
VICTORIA BC V8X 3X4 1-800-555-5555

CHEQUING ACCOUNT STATEMENT
Page 1 of 1

JOHN JONES
1843 DUNDAS ST W APT 27
TORONTO ON MKK 1V2

Date	Description	Ref.	Withdrawals	Deposits	Balance
2003-10-08	Previous Balance				61.53
2003-10-14	Payroll Deposit - HOTEL			694.81	695.36
2003-10-14	Web Bill Payment - MASTERCARD	9685	200.00		495.36
2003-10-16	ATM Withdrawal - INTERAC	3890	21.25		474.11
2003-10-16	Fees - Interac		1.50		472.61
2003-10-20	Interac Purchase - ELECTRONICS	1975	2.99		469.62
2003-10-21	Web Bill Payment - AMEX	3314	300.00		169.62
2003-10-22	ATM Withdrawal - FIRST BANK	0064	100.00		69.62
2003-10-23	Interac Purchase - SUPERMARKET	1559	29.06		40.56
2003-10-24	Interac Refund - ELECTRONICS	1975		2.99	43.53
2003-10-27	Telephone Bill Payment - VISA			35.76	79.29
2003-10-28	Payroll Deposit - HOTEL	2475	6.77		72.52
2003-10-30	Web Funds Transfer - From SAVINGS	2620		50.00	122.52
2003-11-03	Pre-Auth. Payment - INSURANCE		33.55		88.97
2003-11-03	Cheque No. - 409		100.00		88.97
2003-11-06	Mortgage Payment		716.49		-627.52
2003-11-07	Fees - Overdraft		5.00		-632.52
2003-11-08	Fees - Monthly		5.00		-637.52
*** Totals ***			1,515.63	1,442.61	

Automated verification of bank account ownership as well as summarizing the total credit and debit amount from transactions

FINE HOTELS + RESORTS

You go make memories. We make sure they're everlasting.

Enjoy hotels that are destinations unto themselves. Book Fine Hotels + Resorts™ with American Express Travel and get an average total value of \$500 in perks. Like daily breakfast for two, 4pm late check-out, and a \$100 experience credit to have something unique at each property. Terms apply.

- Complimentary Breakfast: Wake up to daily breakfast for two and 16:30, all-included.
- Pay with Flexible: Use Membership Rewards™ Pay with Points to lighten your wallet.
- Late Check-Out: Guaranteed late check-out with each booking.
- Experience Credits: Treat yourself with \$100 experience credit.

Automated validation of terms and conditions, spelling checks and adherence to branding guidelines

Overview of Extraction Approaches:

General Component Detection & Extraction Approaches

Extraction for Verbose Documents

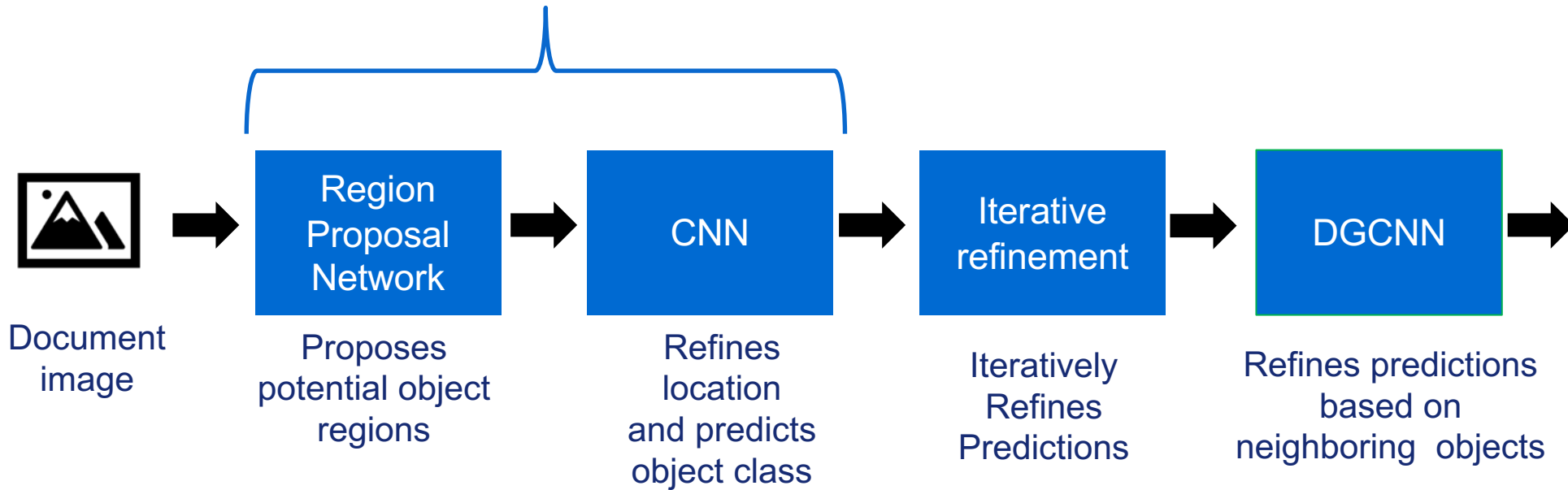
Extraction for Form Type Documents



Credit and
Fraud Risk

A General Component Detection & Extraction Pipeline

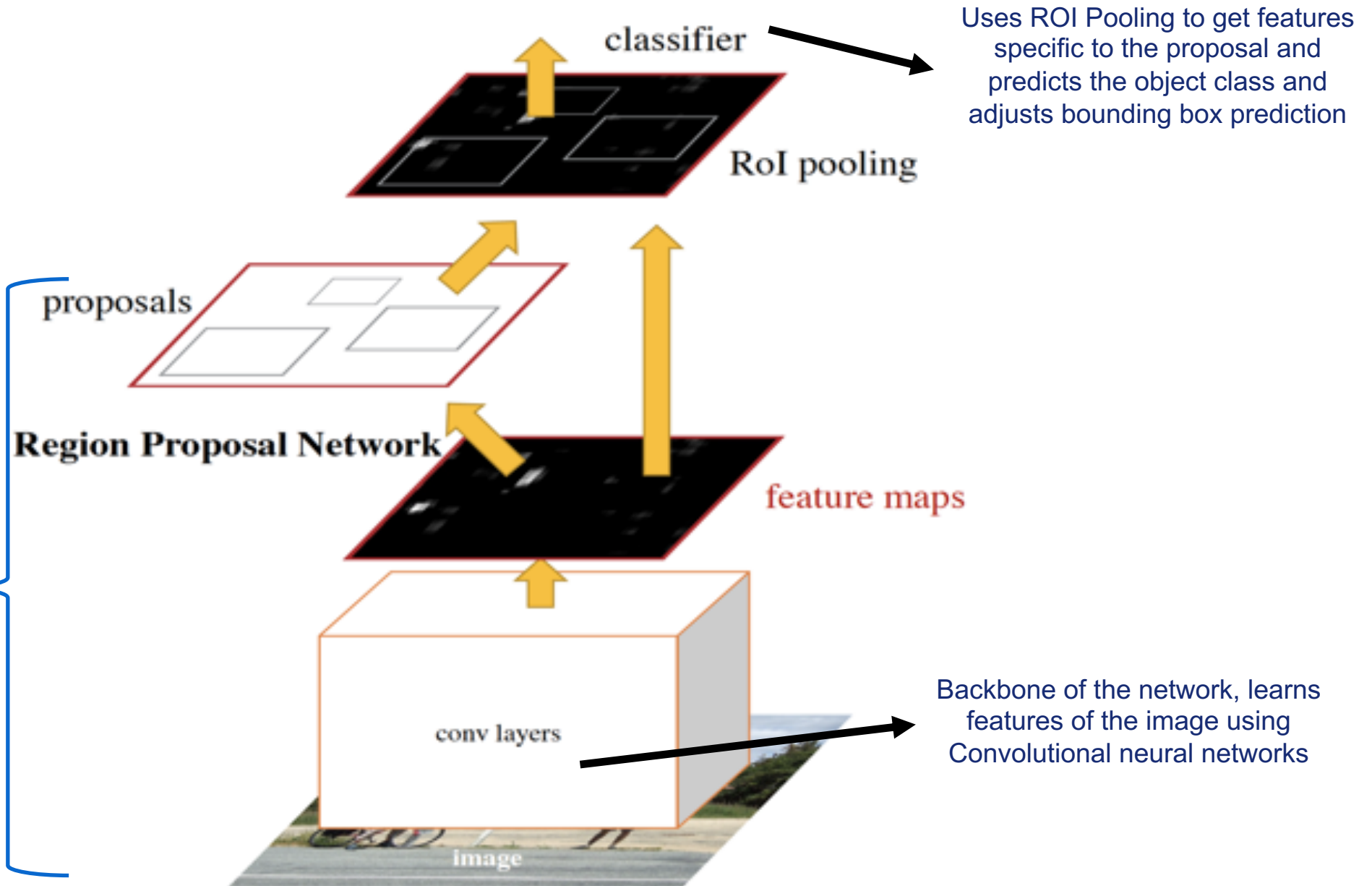
Faster RCNN



Bounding Boxes
+
Class prob.

Faster RCNN

For each predefined region known as Anchor Boxes, gives "objectness" score which indicates the probability of an object being present in the box.



Iterative Refinement

Idea: Objects can be located accurately by iteratively refining predictions

Relation	Question Template
<i>educated_at(x, y)</i>	Where did x graduate from?
	In which university did x study?
	What is x 's alma mater?
<i>occupation(x, y)</i>	What did x do for a living?
	What is x 's job?
	What is the profession of x ?
<i>spouse(x, y)</i>	Who is x 's spouse?
	Who did x marry?
	Who is x married to?

Figure 1: Common knowledge-base relations defined by natural-language question templates.

We show that it is possible to reduce relation extraction to the problem of answering simple reading comprehension questions. We map each re-

1st Iteration

Relation	Question Template
<i>educated_at(x, y)</i>	Where did x graduate from?
	In which university did x study?
	What is x 's alma mater?
<i>occupation(x, y)</i>	What did x do for a living?
	What is x 's job?
	What is the profession of x ?
<i>spouse(x, y)</i>	Who is x 's spouse?
	Who did x marry?
	Who is x married to?

Figure 1: Common knowledge-base relations defined by natural-language question templates.

We show that it is possible to reduce relation extraction to the problem of answering simple reading comprehension questions. We map each re-

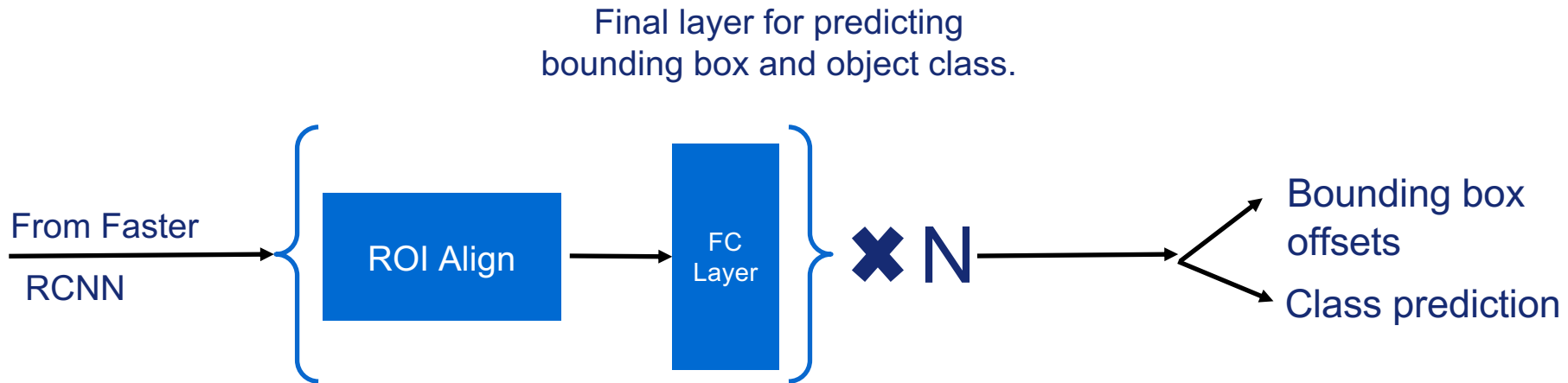
After n iterations

Relation	Question Template
<i>educated_at(x, y)</i>	Where did x graduate from?
	In which university did x study?
	What is x 's alma mater?
<i>occupation(x, y)</i>	What did x do for a living?
	What is x 's job?
	What is the profession of x ?
<i>spouse(x, y)</i>	Who is x 's spouse?
	Who did x marry?
	Who is x married to?

Figure 1: Common knowledge-base relations defined by natural-language question templates.

We show that it is possible to reduce relation extraction to the problem of answering simple reading comprehension questions. We map each re-

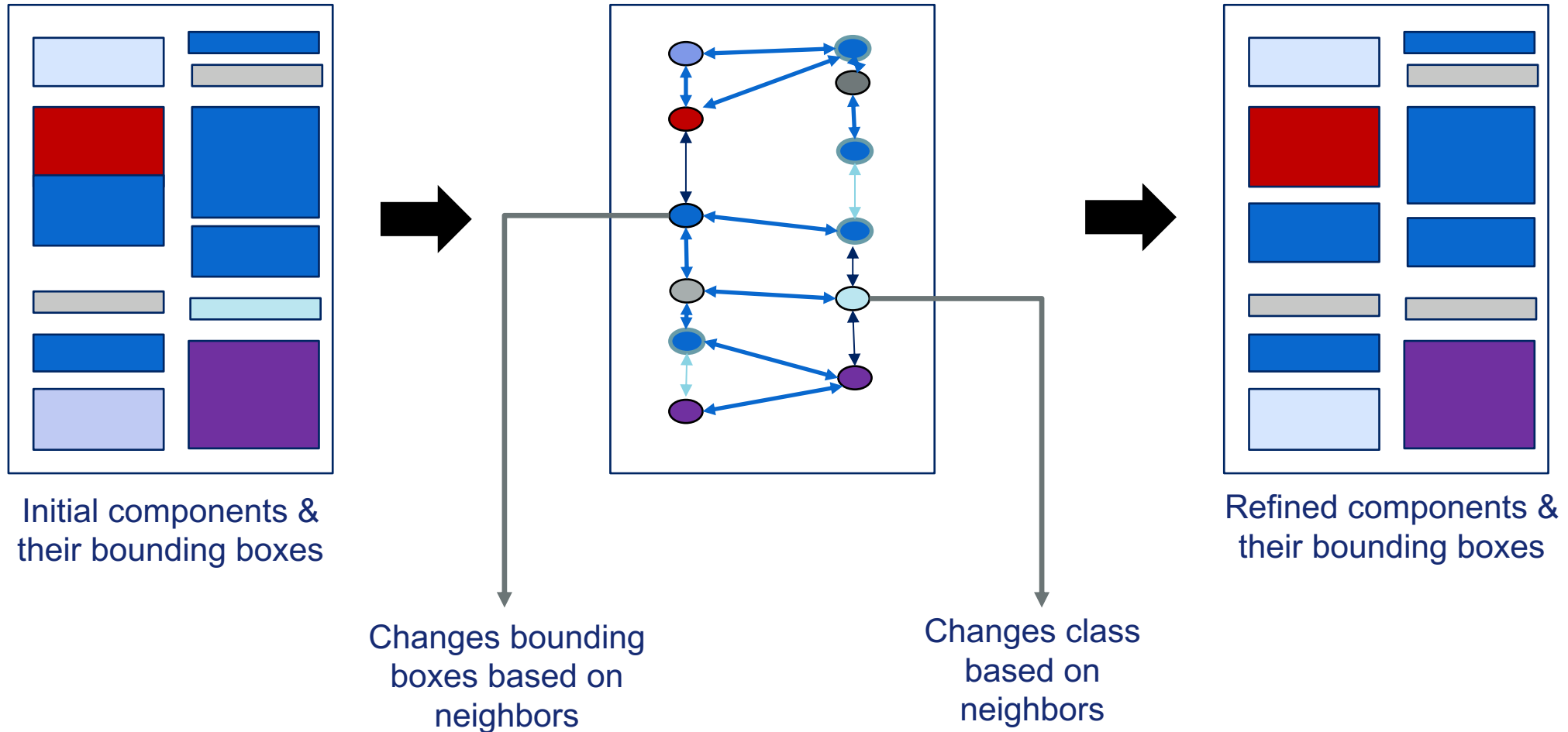
Under the Hood: Iterative Refinement



Bounding box features are extracted by using ROI Align layer. Iteratively bounding box and object class predictions are made.

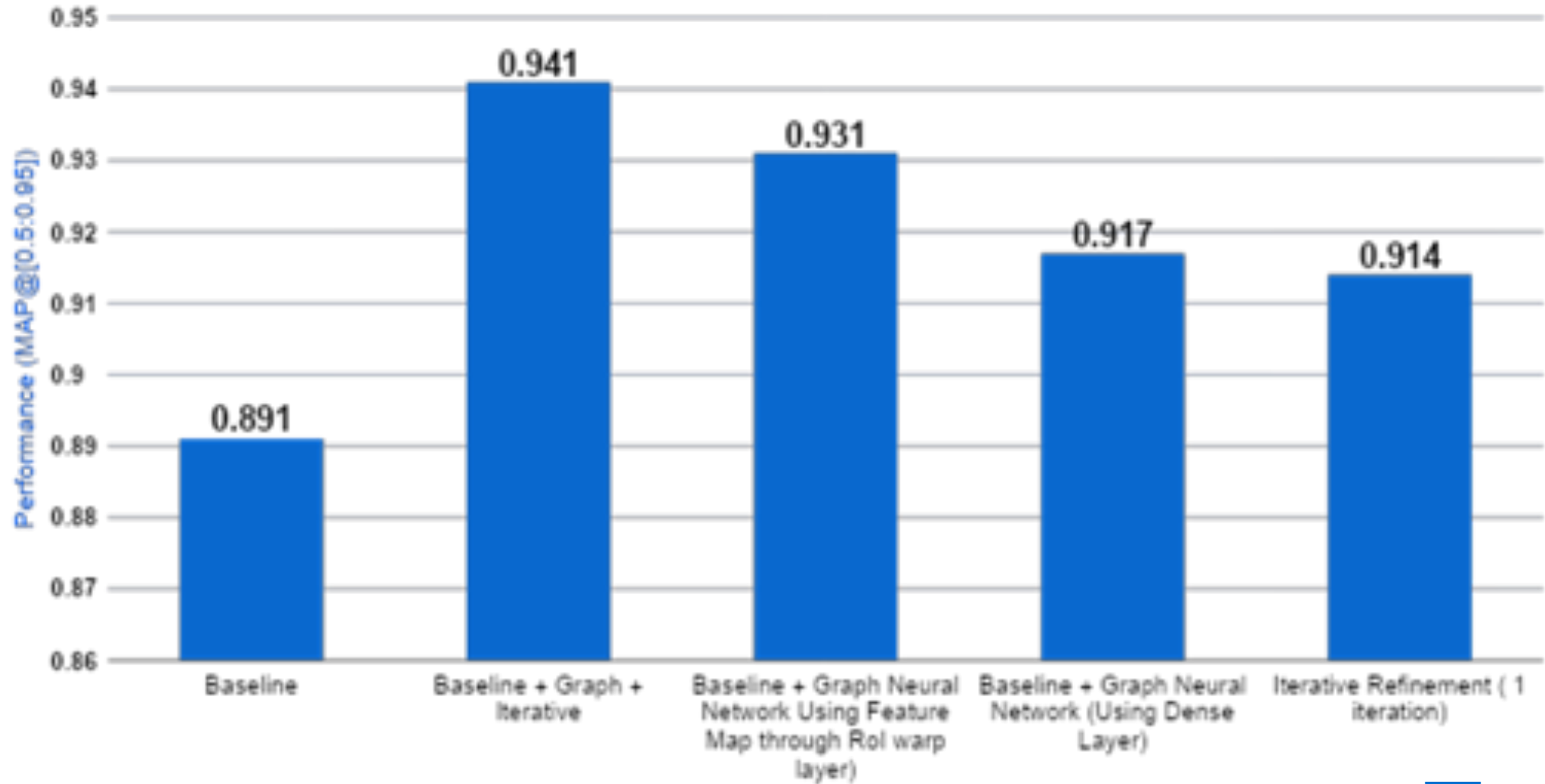
DGCNN

Idea: Components can be better understood by looking at others in its proximity



Results

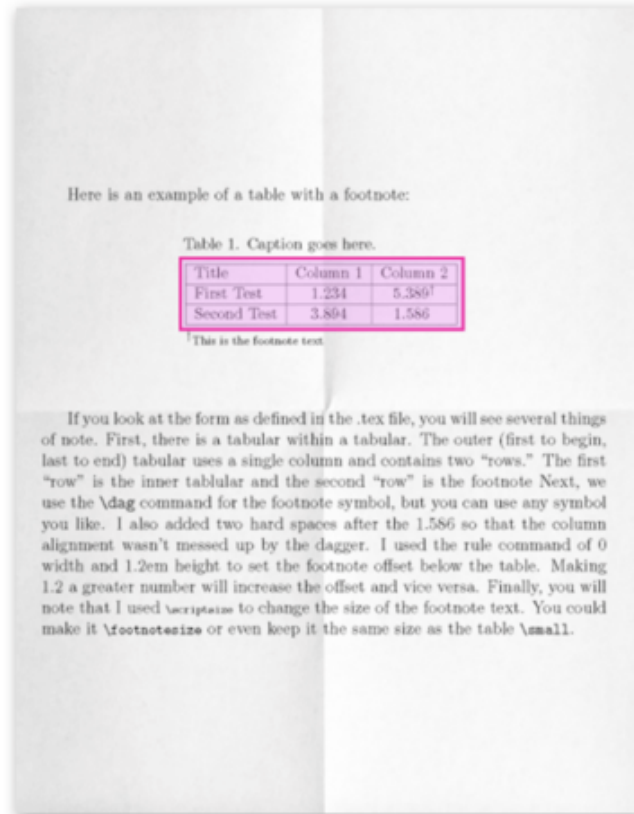
- We leverage publicly available [publaynet](#) dataset that has ~350K annotated images



Tabular Data Extraction

Table Extraction is the task of detecting and decomposing table information in a document.

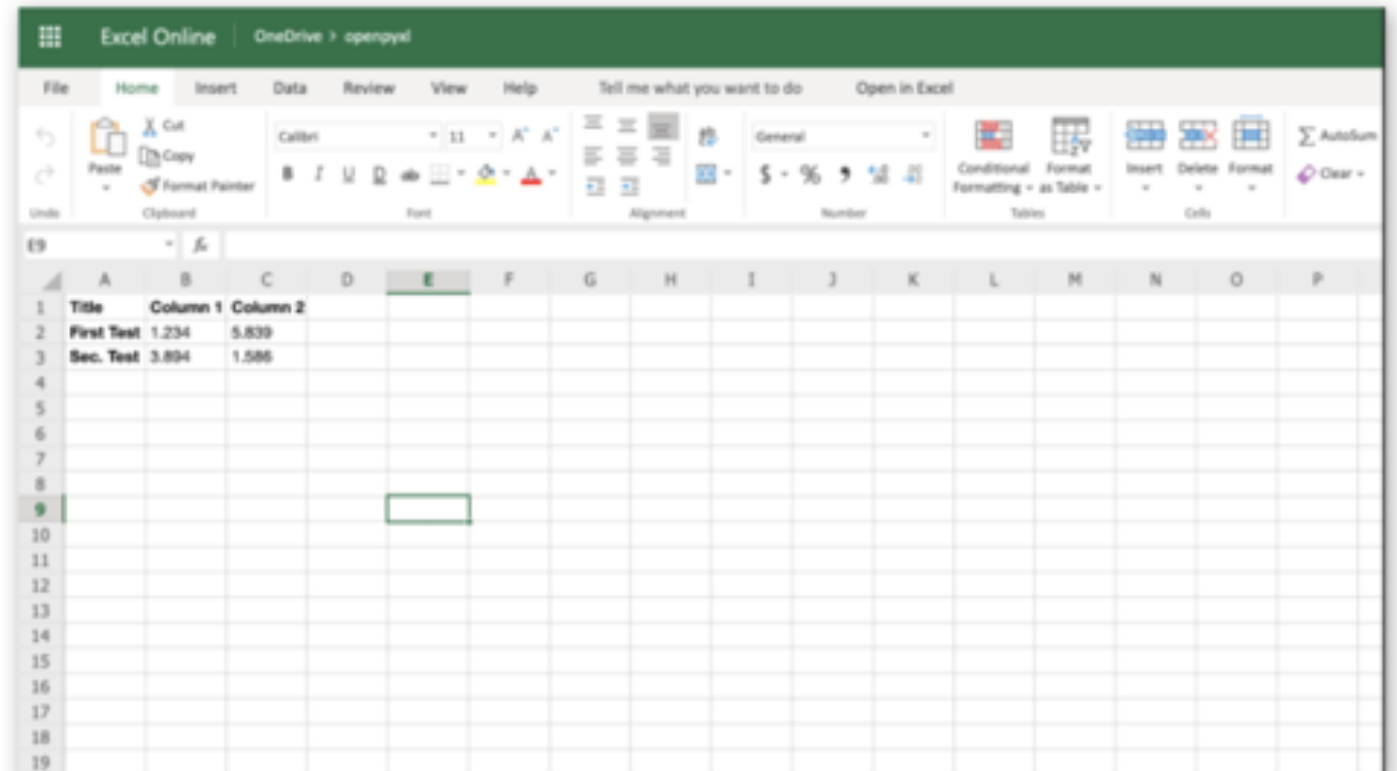
1



2

Title	Column 1	Column 2
First Test	1.234	5.389 ¹
Second Test	3.894	1.586

3



Extraction of Tables

- Existing approaches can be broadly classified into two categories;
 - Top-down: Detect row/column first, followed by formation of cells.
 - Bottom-up: Detect cells first, followed by formation of row/column.
- Top-Down: Advantages/Disadvantages*
 - Straightforward and exploit alignment of different rows/columns to make decisions.
 - Cannot handle spanning cells, because these cells are part of multiple rows/columns.
- Bottom-Up: Advantages/Disadvantages**
 - Bottom-up methods are complex and make decisions based on locality of cells. Since, process is started from cell-level, these methods are more flexible and can handle a wide variety of tables.
 - This flexibility sometimes causes to generate meaningless predictions.

*

1. Khan, Saqib Ali, et al. "Table structure extraction with bi-directional gated recurrent unit networks." *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019.
2. Schreiber, Sebastian, et al. "Deepdesrt: Deep learning for detection and structure recognition of tables in document images." *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*. Vol. 1. IEEE, 2017.

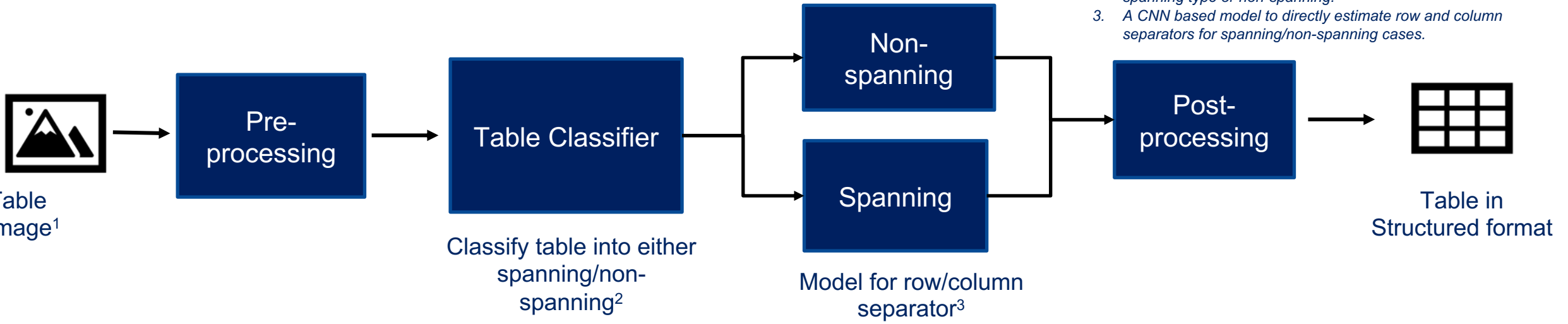
**

1. Zhong, Xu, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. "Image-based table recognition: data, model, and evaluation." *arXiv preprint arXiv:1911.10683* (2019).
2. Qasim, Shah Rukh, Hassan Mahmood, and Faisal Shafait. "Rethinking table recognition using graph neural networks." *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019.

Table Extraction Approach

Note:

1. Input: table location along with complete image to crop out the required region.
2. Image Classification Model to classify a table into either spanning type or non-spanning.
3. A CNN based model to directly estimate row and column separators for spanning/non-spanning cases.



Visualization of sample output

Type of support	Group		P
	CMV (n = 8)	AV-ECMO (n = 9)	
Lactated Ringer's (10 ml/kg)	2	8	0.015
Epinephrine (0.5–2 µg/kg per min)	1	6	0.049
Dopamine (5 µg/kg per min)	1	6	0.049
Bicarbonate (1 mEq/kg bolus)	1	6	0.049
Surviving/nonsurviving	6/2	4/5	0.333
Total number of resuscitative measures in surviving lambs	2 (n = 6)	12 (n = 4)	0.001
Cause of death	Prolonged hypotension with MAP <30 mmHg	Prolonged hypotension with MAP <30 mmHg and AV shunt <5% of CO	

Input: Spanning table with column span=2 in 1st row

Type of support	Group		P
	CMV (n = 8)	AV-ECMO (n = 9)	
Lactated Ringer's (10 ml/kg)	2	8	0.015
Epinephrine (0.5–2 µg/kg per min)	1	6	0.049
Dopamine (5 µg/kg per min)	1	6	0.049
Bicarbonate (1 mEq/kg bolus)	1	6	0.049
Surviving/nonsurviving	6/2	4/5	0.333
Total number of resuscitative measures in surviving lambs	2 (n = 6)	12 (n = 4)	0.001
Cause of death	Prolonged hypotension with MAP <30 mmHg	Prolonged hypotension with MAP <30 mmHg and AV shunt <5% of CO	



Prediction

- Horizontal blue lines indicate row separators
- Vertical small line segments indicate column separators for every cell

Information Extraction from Verbose Documents

Information Extraction from Verbose Documents

? NER extracts the entity but do not generate the labels

Person
ORG

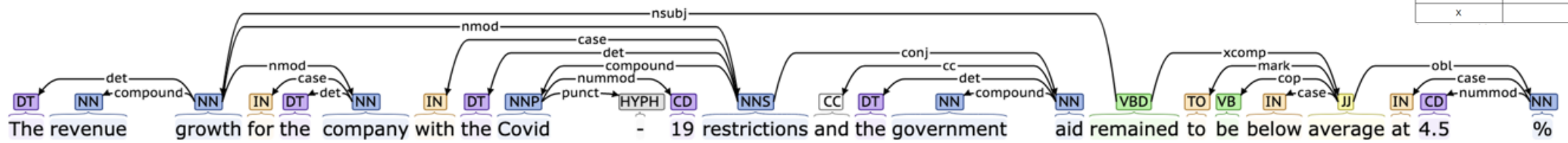
Stephen J. Squeri is the CEO of Amex, a multinational financial services corporation, which is is headquartered at 200 Vesey Street in New York city.

Location

Tag	Description
ADJ	Adjective
ADV	Adposition
ADP	Adverb
AUX	Auxiliary
CCONJ	Coordinating Conjunction
DET	Determiner
INTJ	Interjection
NOUN	Noun
NUM	Numeral
PART	Particle
PRON	Pronoun
PROPN	Proper Noun
PUNCT	Punctuation
SCONJ	Subordinating Conjunction
SYM	Symbol
VERB	Verb
X	Other

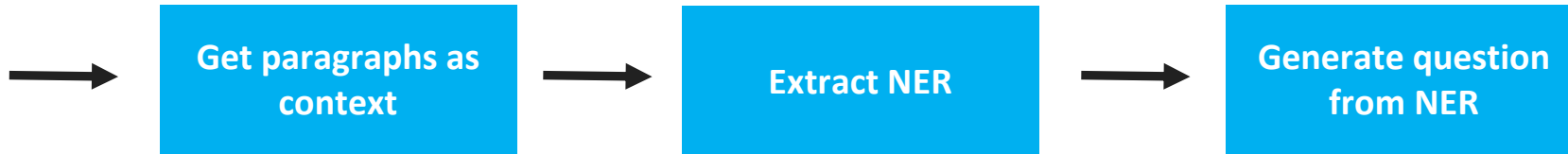
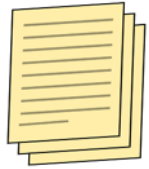
? Dependency parse tree or proximity-based methods fail to capture long dependencies

? E.g. The revenue growth for the company with the Covid-19 restrictions and the government aid remained to be below average at 4.5%



Key-value Pair Extraction

Information extraction formulated as a question answering problem

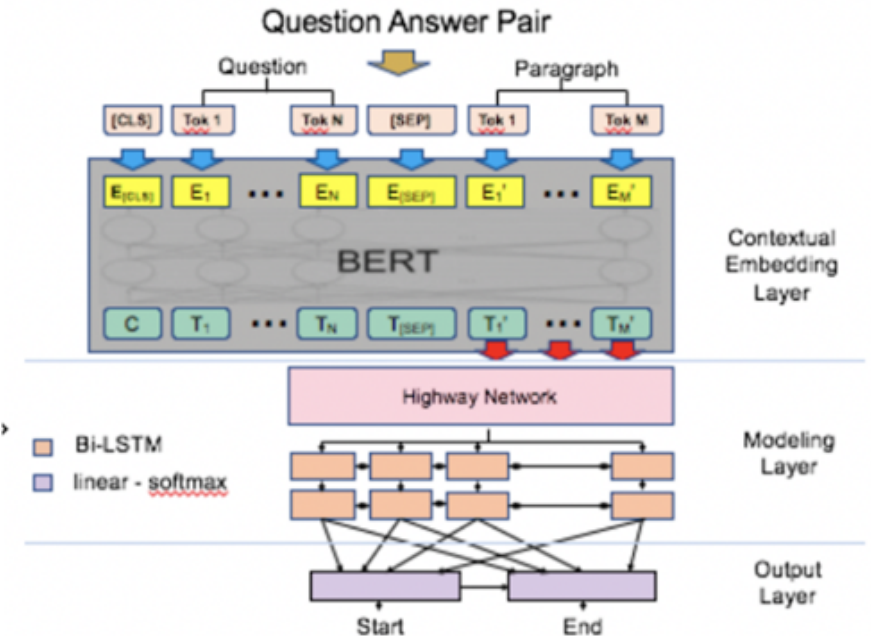


Context: Stephen J. Squeri is the CEO of Amex, a multinational financial services corporation, which is headquartered at 200 Vesey Street in New York City.

Question: Who is Stephen J. Squeri?



{KEY: Answer, VALUE: NER}
{CEO of Amex : Stephen J. Squeri}



Example (1/2)

Input to the System

In October 2019, the Company increased the borrowing capacity on the revolving credit loan by \$33,000 increasing the available credit facility from \$60,000 to \$93,000....If the loans paid during months 13-24 or 25-36 and then a penalty of 2% and 1%, respectively, of the loan balance will be charged on the date of repayment... The weighted-average remaining lease term and discount rate related to the Company's lease liabilities as of September 26, 2020 were 10.3 years and 2.0%, respectively.

Output of the system

Sentences	Entity	Entity Type	Associated text
In October 2019, the Company increased the borrowing capacity on the revolving credit loan by \$33,000 increasing the available credit facility from \$60,000 to \$93,000.	\$33,000	Money	capacity on the revolving credit loan
	\$60,000 to \$93,000	Money	available credit facility
If the loan is paid during months 13-24 or 25-36 and then a penalty of 2% and 1%, respectively, of the loan balance will be charged on the date of repayment.	13-24 or 25-36	Date	loan is paid during months
	2% and 1%	Percent	penalty of the loan balance
The weighted-average remaining lease term and discount rate related to the Company's lease liabilities as of September 26, 2020 were 10.3 years and 2.0%, respectively	10.3 years	Date	remaining lease term
	2.0%	Percent	discount rate

Example (2/2)

Input to the System

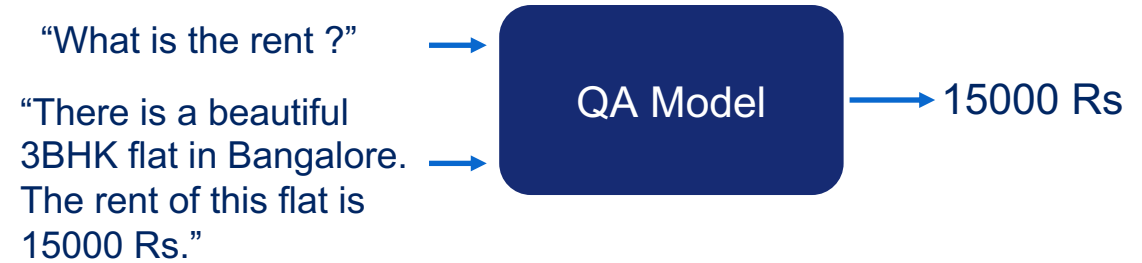
In October 2019, the Company increased the borrowing capacity on the revolving credit loan by \$33,000 increasing the available credit facility from \$60,000 to \$93,000....If the loans paid during months 13-24 or 25-36 and then a penalty of 2% and 1%, respectively, of the loan balance will be charged on the date of repayment... The weighted-average remaining lease term and discount rate related to the Company's lease liabilities as of September 26, 2020 were 10.3 years and 2.0%, respectively.

Output by State-of-the-Art Open IE Systems

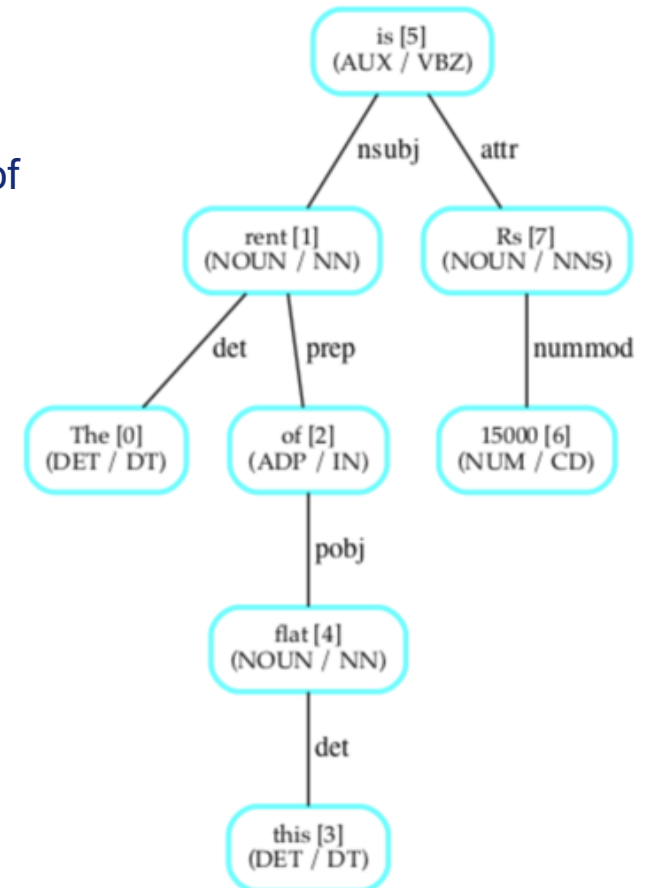
Sentence 1			
	Subject	Relation	Object
Stanford Open IE	Company	increased	borrowing capacity
Allen NLP Open IE	the Company	increased	the borrowing capacity on the revolving credit loan
Sentence 2			
	Subject	Relation	Object
Stanford Open IE	loan	is	If paid
Allen NLP Open IE	the loan	paid	NIL
Sentence 3			
	Subject	Relation	Object
Stanford Open IE	NIL	NIL	NIL
Allen NLP Open IE	NIL	remaining	lease term

Information Extraction as Question Answering System

- A system which ingests a document, generates relevant questions, retrieves answers which are focused finding relevant information. Since the we are dealing in question answer pairs, the relation problem is eliminated, and insight generation is easier.
- **QA Model** : It is an open-source framework for NLP. This model takes question and a paragraph as an input and searches the answer from the paragraph.
- **Syntactic Map** : It is the representation that analyses the grammatical structure of a sentence based on the dependencies between the words in a sentence.



Syntactic Map of the sentence



Question Generation System – Sentence Level

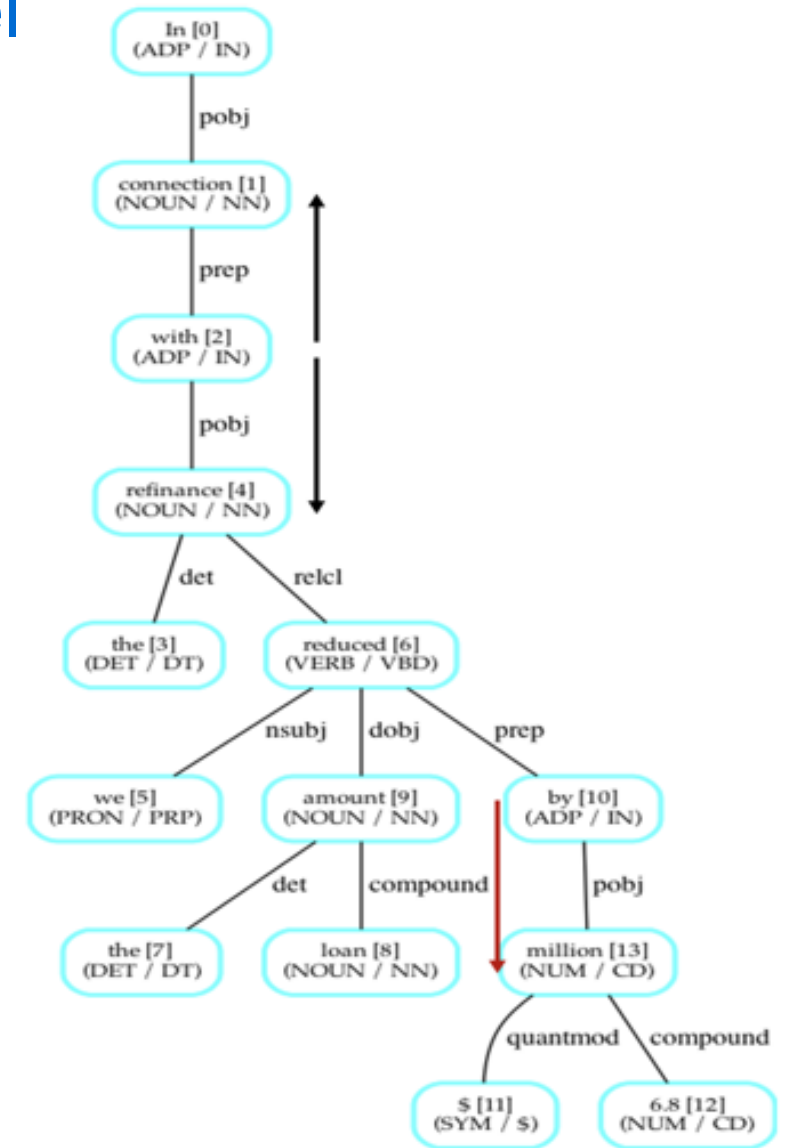
Sentence: “In connection with the refinance we reduced the loan amount by \$6.8 million.”

Noun Phrase Questions: Each sentence comprises subject-object and verb connecting them where Subject or Object is usually a noun or pronoun. Initially we search for a noun and pronoun, after that we check for any noun compound or adjective.

- **Loan Amount | What is loan amount ?**

Preposition Phrase Questions: For complex phrase extraction we first start with preposition extraction. We then follow similar steps as in simple phrase extraction to look for phrases in both left and right of the preposition.

- **Connection with refinance | What is Connection with refinance ?**



Sentence level question generation example

Sentence : The weighted-average remaining lease term and discount rate related to the Company's lease liabilities as of September 26, 2020 were 10.3 years and 2.0%, respectively

Noun Phrases Extracted:

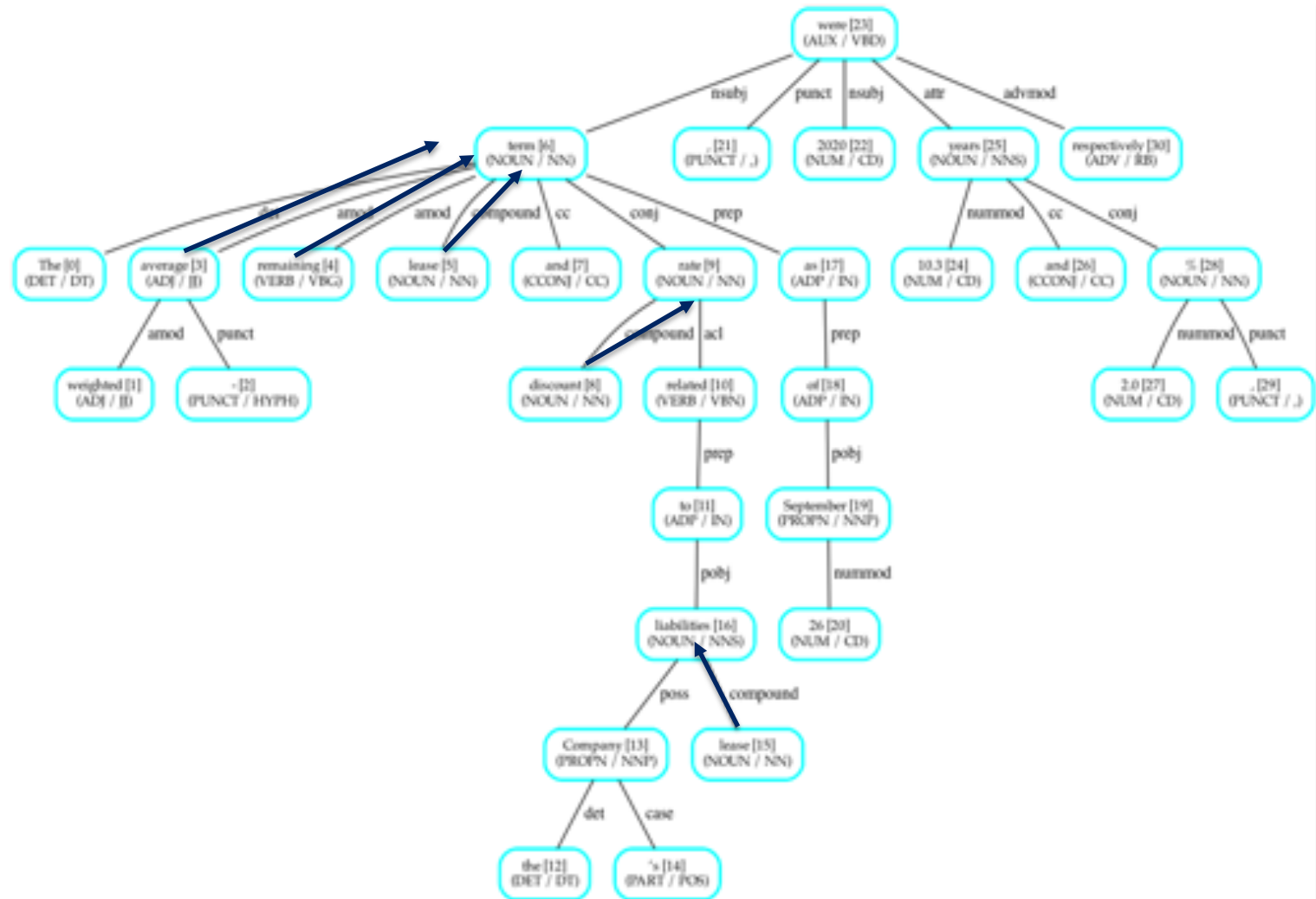
- average lease term
- lease liabilities
- discount rate

Preposition Phrases Extracted:

- average remaining lease term

Questions Created :

- What is average remaining lease term
- What is lease liabilities
- What is discount rate



Sentence level question generation example

Sentence : In October 2019, the Company increased the borrowing capacity on the revolving credit loan by \$33,000 increasing the available credit facility from \$60,000 to \$93,000.

Noun Phrases Extracted:

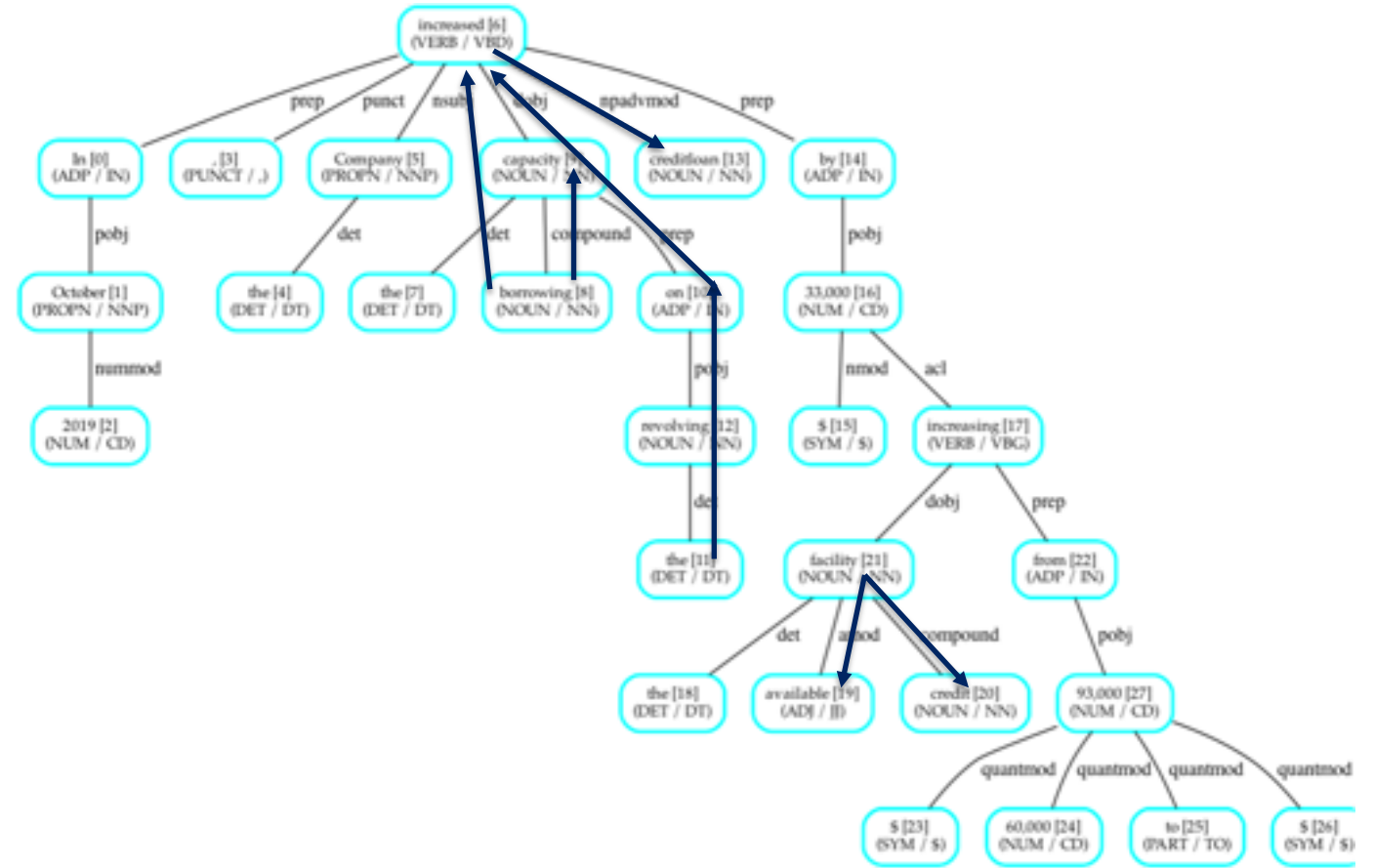
- borrowing capacity
- revolving credit loan
- available credit facility

Preposition Phrases Extracted:

- borrowing capacity on revolving credit loan

Questions Created :

- What is borrowing capacity on revolving credit loan
- What is available credit facility
- What is revolving credit loan
- What is borrowing capacity

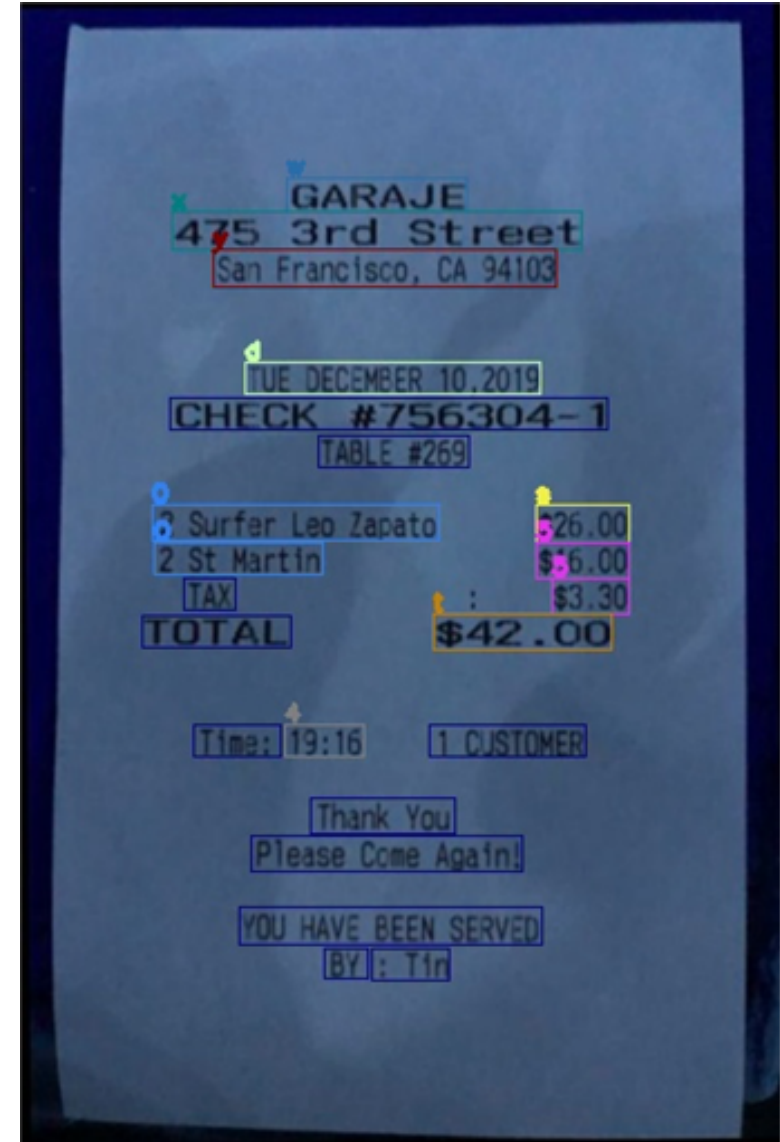


Information Extraction from Form Type Documents

Sample Input Output



INPUT



OUTPUT

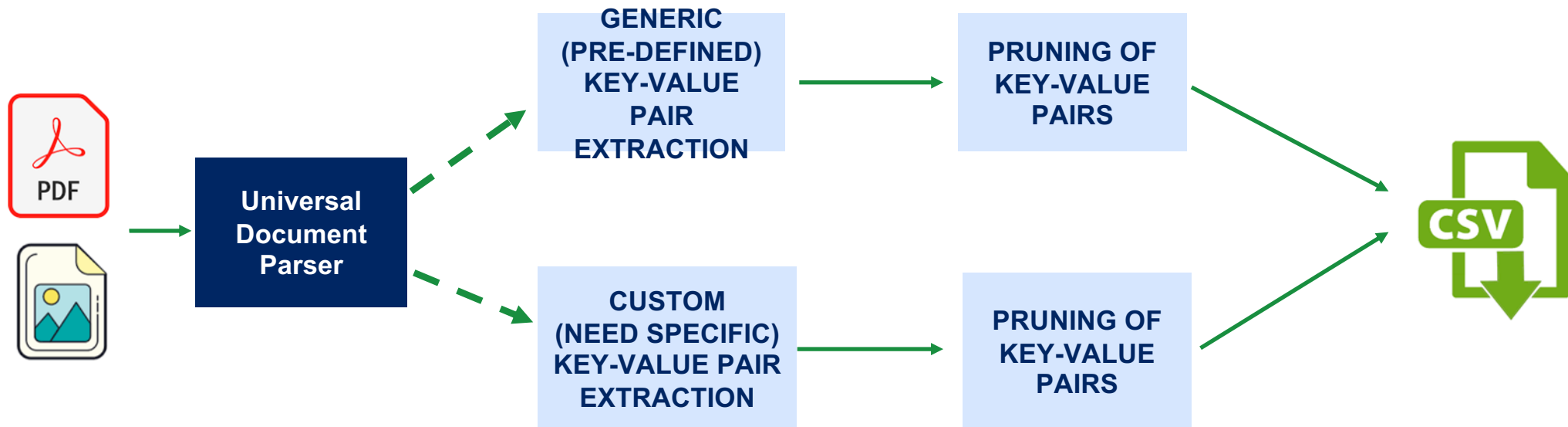


Information Extraction from Form Type Documents

Objective: Extract data elements from Templated type documents into structured key-value pairs

Solution:

- To be exposed as a service that can be consumed by partners in their end-to-end process without the need for manual heuristics



Existing Methods for information extraction (Form-Type Documents)

1 Template-based :

Rules to conclude what is the type of information contained in each position on the image

CONS:

1. Cumbersome and error prone process
2. Fails on unseen templates
3. Does not take spatial features into account

Example

(a) Invoice form with fields: DATE (1/16/2018), INVOICE # (118708)

(b) Invoice form with fields: Invoice Number: 100153386, Invoice Date: 02/26/2015

(c) Invoice form with fields: Date (21-10-2018), Performance date (21/10/2018)

(d) Invoice form with fields: INVOICE # 215589, PO # 483554, DATED: 11/9/18

example only – not real data

2 NLP Based :

The goal of assigning tags to each portion of the text

PROS: Able to perceive unseen layout

CONS:

1. Breaks down with multiline text, like addresses and tables
2. The cases where information is embedded in the spatial arrangement of the layout, not in the text itself

Example

C111032
3 World Financial Center
MC 01-06-12
New York NY 10285
United States
Terms Net 30
Cust #: C111032

Existing Methods and Motivation for using GCN

3 Deep Neural Networks (Supervised CNNs)

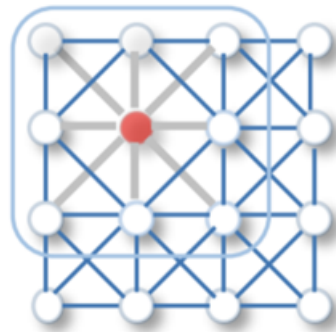
PROS: Able to capture local patterns irrespective of variability

CONS:

1. Limitation to capturing local pattern through vertices which all have equal weightage.
2. Addresses and tables, where the information is embedded in the spatial arrangement of the layout, not in the text itself.

1. Nodes are spatially related to each other by their Euclidean distance, nodes will have an absolutely uniform structure: each node has equally-weighted edges to its 4 immediate neighbours

2. Filters analyse patterns in the locality of each pixel, e.g.: changes in colour that indicate borders.



Graph Neural Networks

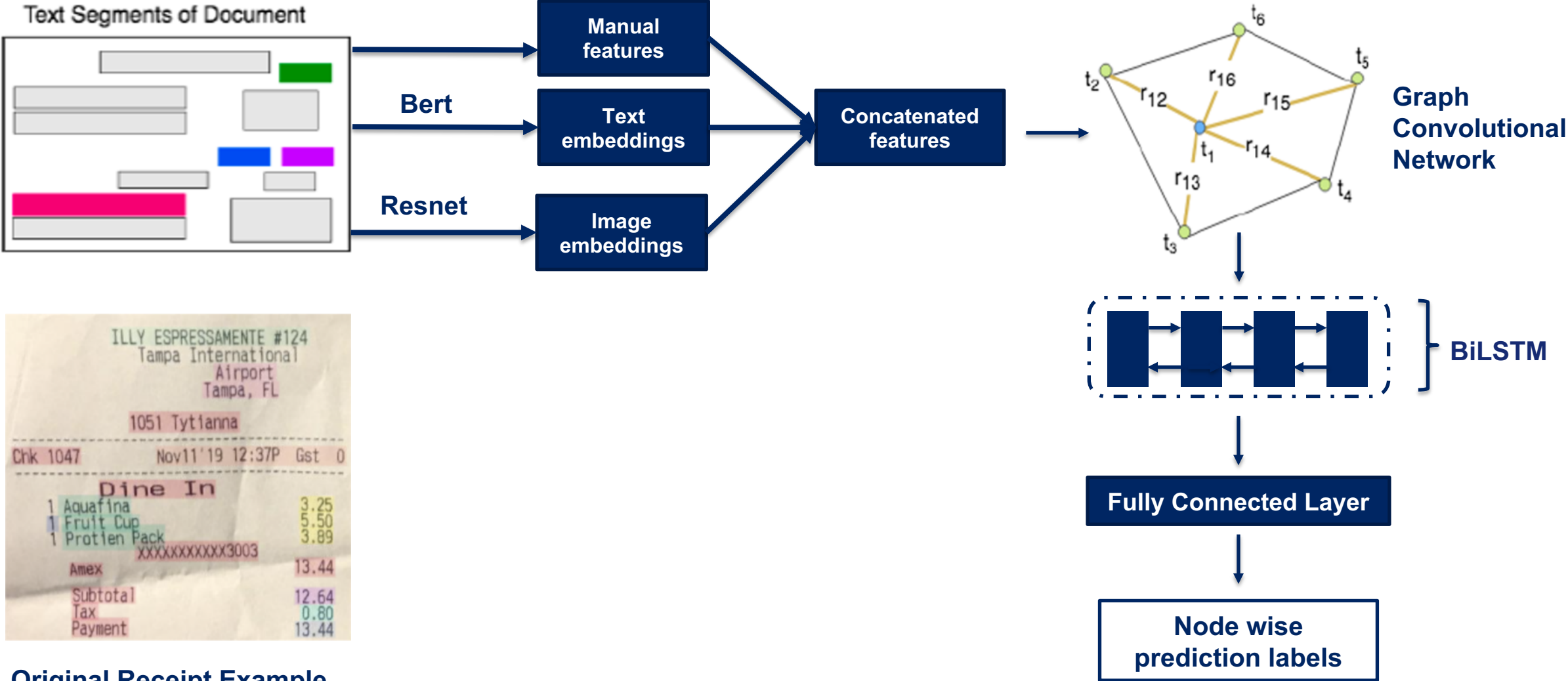
1. The need to recognize local patterns in graphs, a GCN could start by capturing local patterns between neighbouring nodes in a graph
2. Graphs are locally connected structures, which makes them a good candidate for the type of analysis supported by a stack of convolutions.



1. No implicit uniformity assumption

2. The edges between nodes only exists if they are explicitly defined

Solution Overview



Original Receipt Example

example only – not real data

Modelling approach

1. Text Embedding

- [Bert](#)

2. Image Embeddings

- [Resnet](#)

3. Manual Features specific to dataset

- Date,
- Bounding Box Co ordinates
- Currency

4.Edge Embeddings

- X,Y : Horizontal and Vertical Distance between boxes
- W,H : Width and Height of the text boxes

$$\mathbf{r}_{ij} = [x_{ij}, y_{ij}, \frac{w_i}{h_i}, \frac{h_j}{h_i}, \frac{w_j}{h_i}],$$

Node wise classification based on Node and edge embeddings

References: “Graph Convolution for Multimodal Information Extraction from Visually Rich Documents”

Use Case Highlight: Bank Statements

Key Fields for Extraction

FIRST BANK OF WIKI
 1425 JAMES ST, PO BOX 4000
 VICTORIA BC V8X 3X4 1-800-555-5555

CHEQUING ACCOUNT STATEMENT
 Page : 1 of 1

JOHN JONES
 1643 DUNDAS ST W APT 27
 TORONTO ON M6K 1V2

Statement period	Account No.
2003-10-09 to 2003-11-08	00005-123-456-7

Date	Description	Ref.	Withdrawals	Deposits	Balance
2003-10-08	Previous balance				0.55
2003-10-14	Payroll Deposit - HOTEL			694.81	695.36
2003-10-14	Web Bill Payment - MASTERCARD	9685	200.00		495.36
2003-10-16	ATM Withdrawal - INTERAC	3990	21.25		474.11
2003-10-16	Fees - Interac		1.50		472.61
2003-10-20	Interac Purchase - ELECTRONICS	1975	2.99		469.62
2003-10-21	Web Bill Payment - AMEX	3314	300.00		169.62
2003-10-22	ATM Withdrawal - FIRST BANK	0064	100.00		69.62
2003-10-23	Interac Purchase - SUPERMARKET	1559	29.08		40.54
2003-10-24	Interac Refund - ELECTRONICS	1975		2.99	43.53
2003-10-27	Telephone Bill Payment - VISA	2475	6.77		36.76
2003-10-28	Payroll Deposit - HOTEL			694.81	731.57
2003-10-30	Web Funds Transfer - From SAVINGS	2620		50.00	781.57
2003-11-03	Pre-Auth. Payment - INSURANCE		33.55		748.02
2003-11-03	Cheque No. - 409		100.00		648.02
2003-11-06	Mortgage Payment		710.49		-62.47
2003-11-07	Fees - Overdraft		5.00		-67.47
2003-11-08	Fees - Monthly		5.00		-72.47
*** Totals ***			1,515.63	1,442.61	

Ownership details, e.g., account name, number for authorization

Transaction details for understanding of financial risks

example only – not real data

Opportunities



Existing cloud or vendor-based solutions are limited in quality and cover only generic fields.

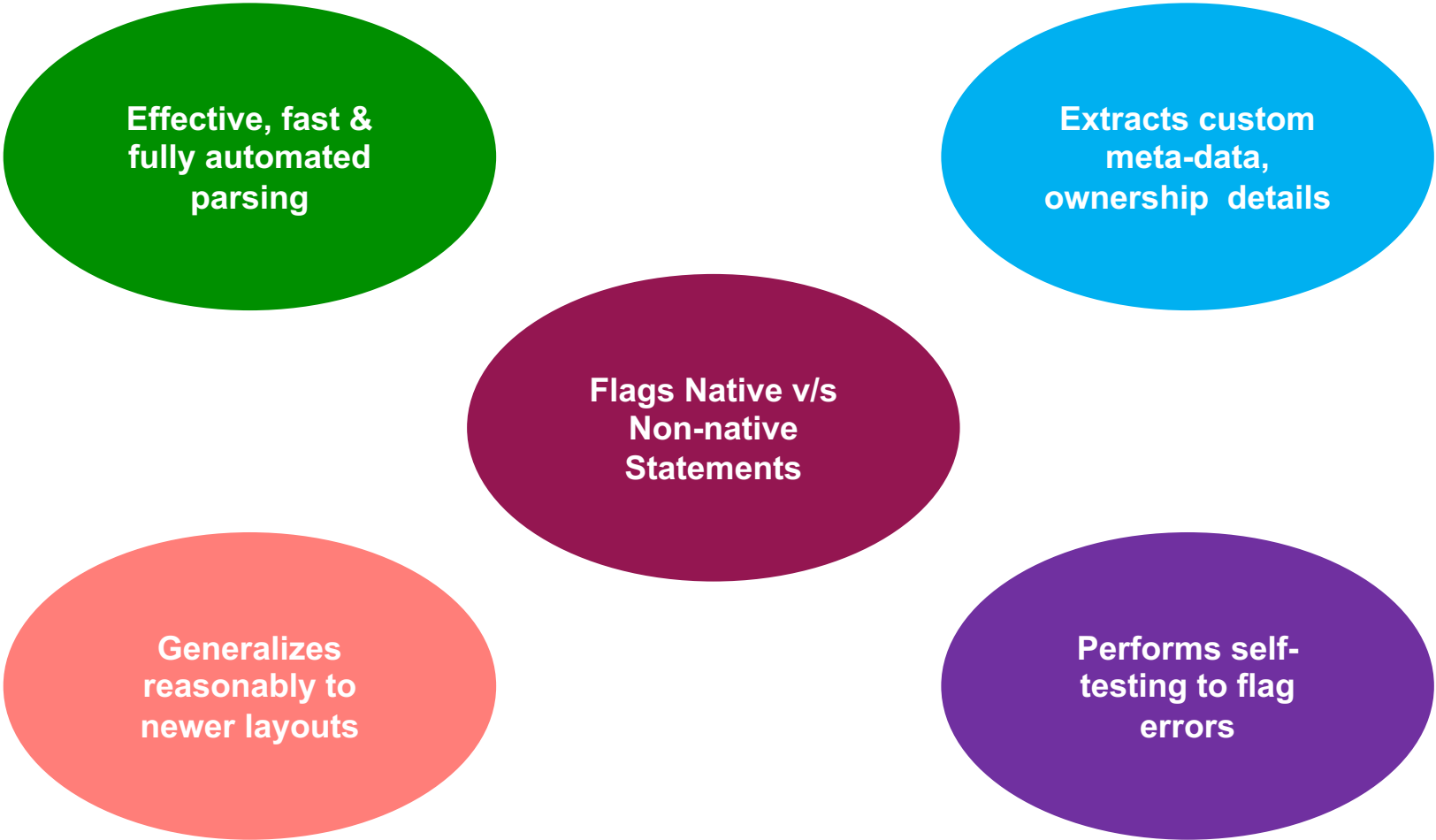


Vendor-based solutions can not provide near real-time response and thus compromise customer experience

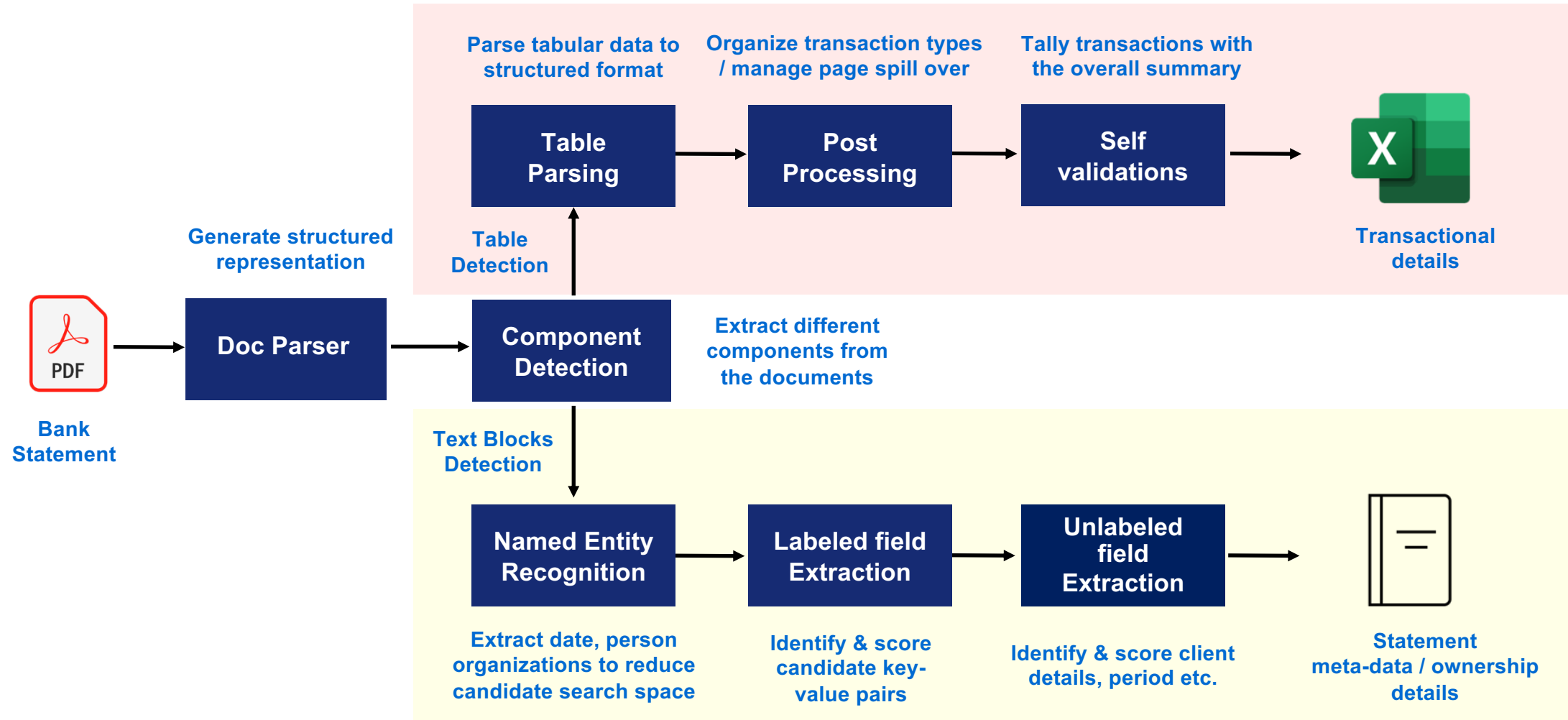


Huge opportunity to mitigate the risk of financial loss, operational expenses and protects brand reputation

Desired Features for a Bank Statement Extraction Solution



Pipeline for Bank Statement Extraction



Use Case Highlight: Digital Auditor

Digital Auditor

Business Problem : Procure-to-Pay operations validates if invoices are legitimate, unique, and had not been previously financed, which is extensively laborious and subjective



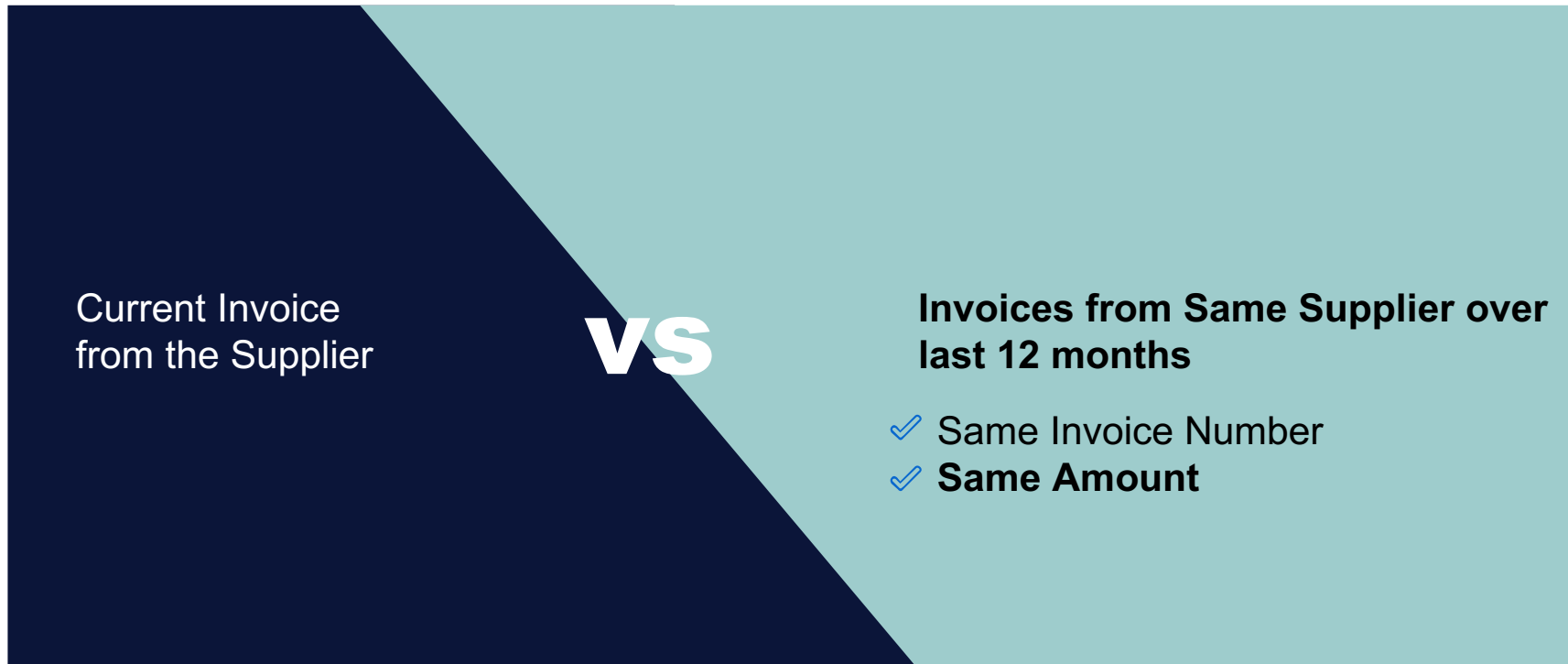
In its 2014 annual report, the US Government Accountability Office (GAO) disclosed that it was involved in preventing such improper payments of \$124.7 billion within just 22 federal agencies

Duplicate invoices occur far more often than organisations realise (around 0.1% total invoice payments) and the overlapping invoice scenario is the big fraction of duplicate invoices.

Why Digital Auditor ?

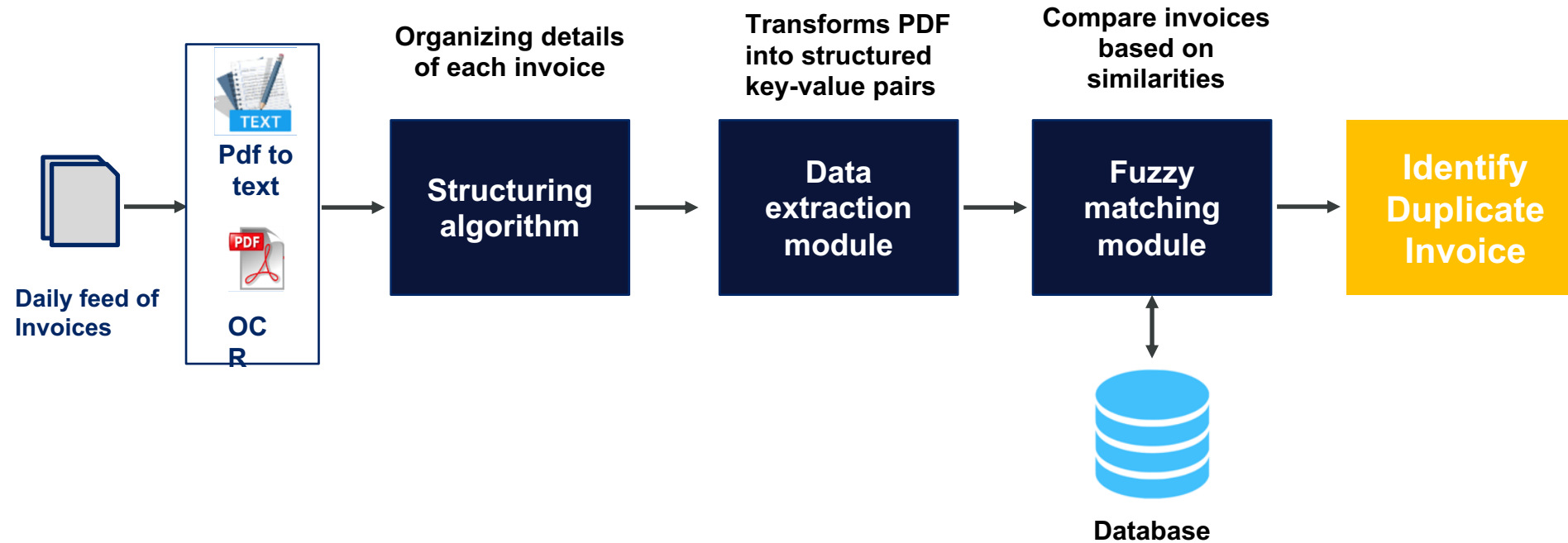
How ERP Solutions detect Potential Duplicate Invoices

Due to limitations in current ERP solutions, organizations must invest manual effort in identification of duplicate invoices and prevent them from payment



Current ERP Solutions are **NOT** completely equipped to detect fraudulent / incorrect Invoices

Digital Auditor Solution



Stage 1: Data Extraction Module

Information Extraction & NLP (3)

Regular expressions (REGEX):

- capture different fields like invoice no, amount, date, PO, period etc. including variations across vendors
- Multiple regex to capture all different variations

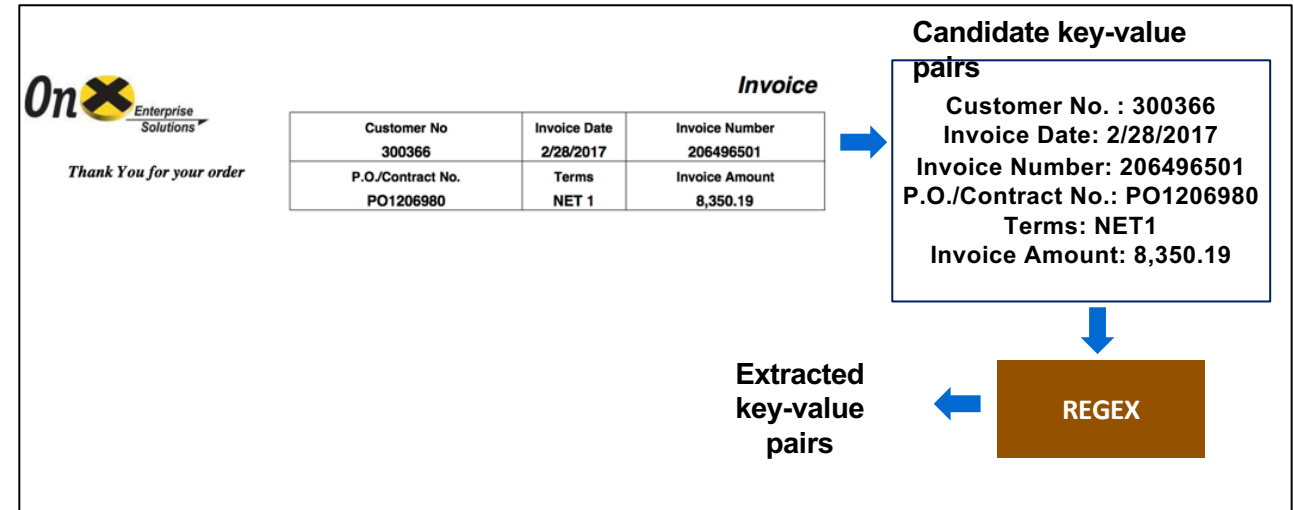
Regex:

Invoice No: `[i][n][v][o][i][c][e]\s*[\s#:=~]+\s*(\w-)+`
Invoice Date: `date\s*[#-:]*\s*\d{1,2}\s*\d{1,2}\s*\d{4}`
Invoice PO: `[p][o]\s*[:#-]*\s*(\w+)`

Invoice	INV127541
Date	4/30/2018
PO #	PO1239977
Terms	Net 15
Payment Due	5/15/2018

Structure-aware extraction:

- Based on keywords from field lookup dictionary
- leverage the field context and the structural information to create candidate field-value pairs
- run the candidate pairs through *REGEX*



Stage1: Data Extraction Module

Information Extraction & NLP (3)

Table parsing:

- triggered by keywords from a look-up dictionary
- associating values to field column based on the structural position of field and value
- robust to blank columns or missing column headers


Item	Description	Hours	Rate	From	Till date	Amount
AMEX 110	Bhaves Jain	160	68.50	10/22/2016	11/18/2016	10,960.00
Amx293	Pradyumna Poddar	160	70.00	10/22/2016	11/18/2016	11,200.00

↓

```
[{'from': ' 10/22/2016', 'description': ' bhaves jain', 'amount': ' 10,960.00', 'till date': ' 11/18/2016', 'hours': ' 160', 'item': ' amex 110', 'rate': ' 68.50'},  
{'from': ' 10/22/2016', 'description': ' pradyumna poddar', 'amount': ' 11,200.00', 'till date': ' 11/18/2016', 'hours': ' 160', 'item': ' amx293', 'rate': ' 70.00'}]
```

Address parsing:

- triggered by keyword from address look-up dictionary
- or pyap(python address parser) based on regular expressions to validate valid components like street number, street name followed by a street identifier, city state name abbreviation.

**Invoice**

Thank You for your order

Customer No 300366	Invoice Date 2/28/2017	Invoice Number 206496501
P.O./Contract No. PO1206980	Terms NET 1	Invoice Amount 8,350.19

Bill To:
American Express
TRS Co Inc
20022 N 31st
Phoenix, AZ 85027

Ship To:
American Express
Accertify
2 Pierce Place, Ste 900
Itasca, IL 60143
Attn: Greg Consier

↓

```
[['u' bill to: american express trs co inc 20022 n 31st phoenix, az 85027'],  
['u' ship to: american express accertify 2 pierce place, ste 900 itasca, il 60143 attn: greg consier']]
```

Stage 1: Data Extraction Module

Examples of data extraction

Data
Extraction
Module

```

Invoice_No: cj4019917
Invoice_Date: 7/01/2017
Invoice_Due_Date: 8/31/2017
Invoice_Amount: 462,537.50
Invoice_Address: [
  'bill to : american express consumer affiliate,200 vesey street, 44th floor, new york, ny 10285',
  'attention : michal flejsierowicz, michal.flejsierowicz@aexp.com,; cc: amalia asan, amanda noodell, armand lamhing, michael arlia,,matt santini'
]
Invoice_Description: [
  'Description_Text': [],
  'Line_Items': [
    {'price': 'None', 'campaign name / description': 'june balance'},
    {'price': '158,125.00', 'campaign name / description': 'bankrate q2 placement fees'},
    {'price': '126,412.50', 'campaign name / description': 'offers quarterly fee'},
    {'price': '178,000.00', 'campaign name / description': 'dealmoon q2 placement package'}
  ],
  'Description_Explored': [],
  'Overall_Description': [
    u'campaign name / description           price
    june balance
    bankrate q2 placement fees           158,125.00
    offers quarterly fee                 126,412.50
    dealmoon q2 placement package       178,000.00']
  ]
}
    
```

Extracted data as json (key-value pairs)

Data
Extraction
Module

```

Invoice_No:m5200535386
Invoice_Date:20-oct-2014
Invoice_Due_Date:19-nov-2014
Invoice_Amount:3,350.00
Invoice_Po_No:po1123339
Invoice_account_no:None
Invoice_Address: [
  [u'bill to:', u'attn ':, u' american express', u' 20022 n 31st ave', u' phoenix az 85027']]
  [[u'ship to:', u'american express', u'sandra edwards', u'3202 w behrend dr', u'phoenix az 85027']]
]
Invoice_Description: [
  'Description_Text': [],
  'Line_Items': [
    {'description': ' backup and recovery manager avamar', 'qty shipped': ' 1', 'tax': ' 0.00', 'item': '456-103-950', 'line number': '1', 'unit price': ' 0.00', 'extended price': ' 0.00'},
    {'description': 'brs sol architect 4 hours qs', 'qty shipped': ' 2', 'tax': ' 0.00', 'item': 'ps-bas-sabrs', 'line number': '2', 'unit price': ' 1,675.00', 'extended price': ' 3,350.00'},
  ],
  'Description_Explored': [],
  'Overall_Description': [
    u'line number           item                description                qty
    shipped                unit price          extended price            tax\1
    backup and recovery manager avamar           1                0.00
    0.00\2
    2                1,675.00           3,350.00                0.00']
  ]
}
    
```

Extracted data as json (key-value pairs)

example only – not real data

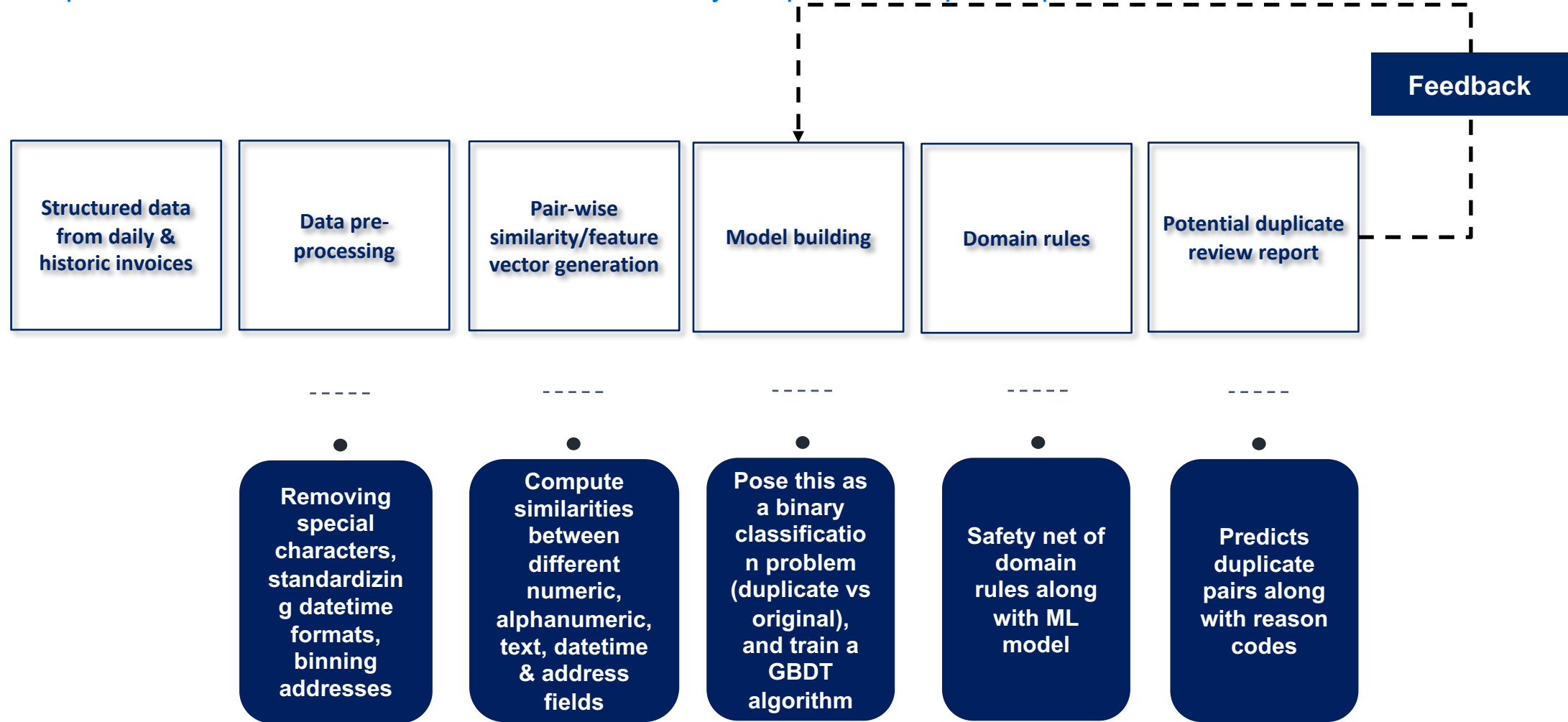
Invoice
e



Credit and
Fraud Risk

Stage2: Duplicate Detection Module

Objective: Compare an invoice with historic invoices and identify the potential duplicate pairs



Feature Extraction Examples

Invoice1

Item	Description	Hours	Rate	From	Till date	Amount
AMEX 110	Bhavesh Jain	160	68.50	12/24/2016	01/20/2017	10,960.00



```
[{'from': '12/24/2016', 'description': 'bhavesh jain', 'amount': '10,960.00', 'till date': '01/20/2017', 'hours': '160', 'item': 'amex 110', 'rate': '68.50'}]
```

Extracted description text

Invoice2

Item	Description	Hours	Rate	From	Till date	Amount
AMEX 110	Bhavesh Jain	160	68.50	10/22/2016	11/18/2016	10,960.00
Amx293	Pradyumna Poddar	160	70.00	10/22/2016	11/18/2016	11,200.00

2a
2b



```
[{'from': '10/22/2016', 'description': 'bhavesh jain', 'amount': '10,960.00', 'till date': '11/18/2016', 'hours': '160', 'item': 'amex 110', 'rate': '68.50'}, {'from': '10/22/2016', 'description': 'pradyumna poddar', 'amount': '11,200.00', 'till date': '11/18/2016', 'hours': '160', 'item': 'amx293', 'rate': '70.00'}]
```

Extracted description text

1a -> 2a



Derived period field

Amex 110	Amex 110
Bhavesh Jain	Bhavesh Jain
160	160
68.50	68.50
12/24/2016-01/20/2017	10/22/2016-11/18/2016
10,960.00	10,960.00

1a -> 2b



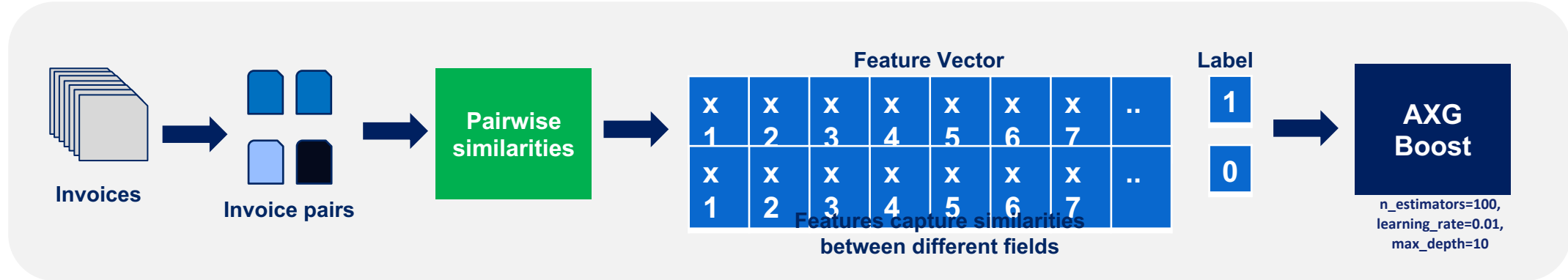
Derived period field

Amex 110	Amx239
Bhavesh Jain	Pradyumna Poddar
160	160
68.50	70.00
12/24/2016-01/20/2017	10/22/2016-11/18/2016
10,960.00	11,200.00

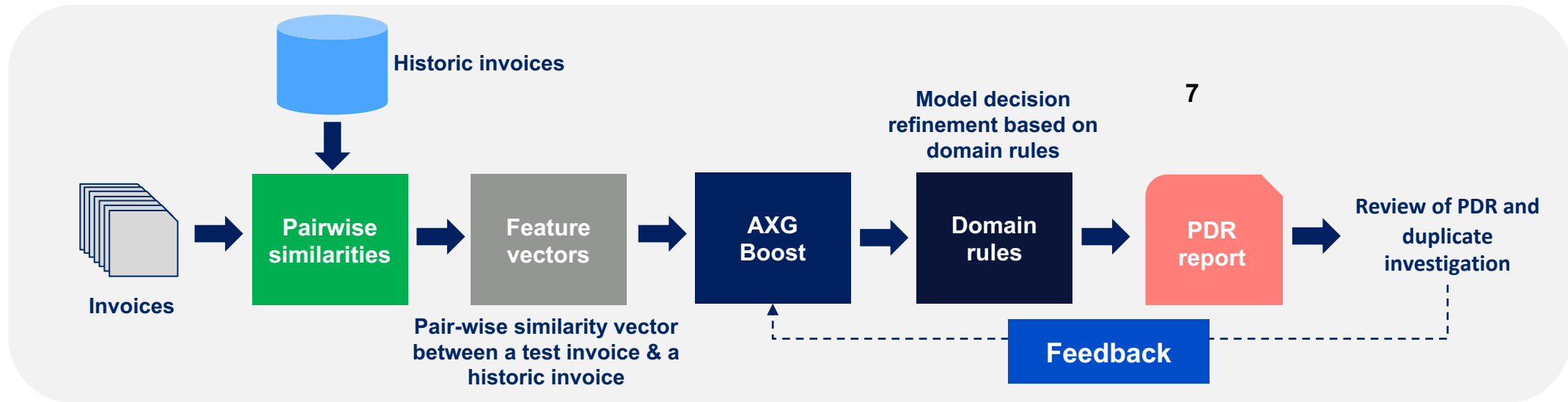
Segregate columns, compare corresponding fields and pick the best matching line-item

Stage2: Duplicate Detection Module

Model Building (6)



Prediction and Feedback Loop (8)



Feedback provides additional instances to train the model

Future Research

Key Areas

Information Extraction

- Generic information extraction i.e. move beyond the predefined list of elements to be extracted

Document Authenticity

- Documents needs to be authenticated against forgery & tampering if they have to support critical decisions such as underwritings, KYC etc

DocVQA

- Look beyond textual information to make sense out of visual elements of documents such as figures and tables

VQA on Document Images



How many females are affected by diabetes: 3.6%

What percentage of cases can be prevented: 60%

What could lead to blindness or stroke: diabetes

VQA on Images

Who is wearing glasses?

man

woman



Where is the child sitting?

fridge

arms



Is the umbrella upside down?

yes

no



How many children are in the bed?

2

1



VQA on Financial Document Images

Springfield Psychological Services

Balance Sheets
December 31, 2004 and 2003

	2004	2003		2004	2003
Assets			Liabilities		
Current assets:			Current liabilities:		
Cash	\$ 12,587	\$ 8,179	Notes payable	\$ 4,200	\$ 4,702
Short-term investments	5,363	3,517	Accounts payable	375	15
Accounts receivable	2,314	3,790	Accrued wages	1,079	1,140
Prepaid rent	2,000	2,000	Taxes payable	5,200	4,722
Total current assets	22,264	18,486	Total current liabilities	10,854	10,679
Property, plant and equipment:			Long-term debt		
Land and building	65,553	28,369		16,898	26,898
Machinery and equipment	5,000	3,211	Owner's Equity		
	70,553	31,580	Total owners' equity	27,791	26,158
Less accumulated depreciation	3,775	4,205	Total liabilities and owners' equity		
Property and equipment, net	66,778	27,375		44,645	52,856
Long-term investments	1,763	4,887			
Other assets	280	218			
Total assets	\$ 89,328	\$ 58,792			

Conclusions

- Documents are one of the key sources of unstructured data in the context of an enterprise
- AI powered document intelligence can understand the structure of a document and extract contents which leads to significant process efficiency
- While information extraction from documents can be performed using naïve methods or rule based approaches – they tend to fail when the document structure dynamically changes
- Deep learning approaches borrowed from image processing, computer vision and other related disciplines can significantly outperform naïve rule based approaches
- A high accuracy approach for information extraction from documents can lead to speed and delightful customer experience in an industry setting

THANKS!!