CS4221
Modern Databases III. Data Curation and RAGs

Yao LU

2024 Semester 2

National University of Singapore
School of Computing

Data curation and RAGs: outline

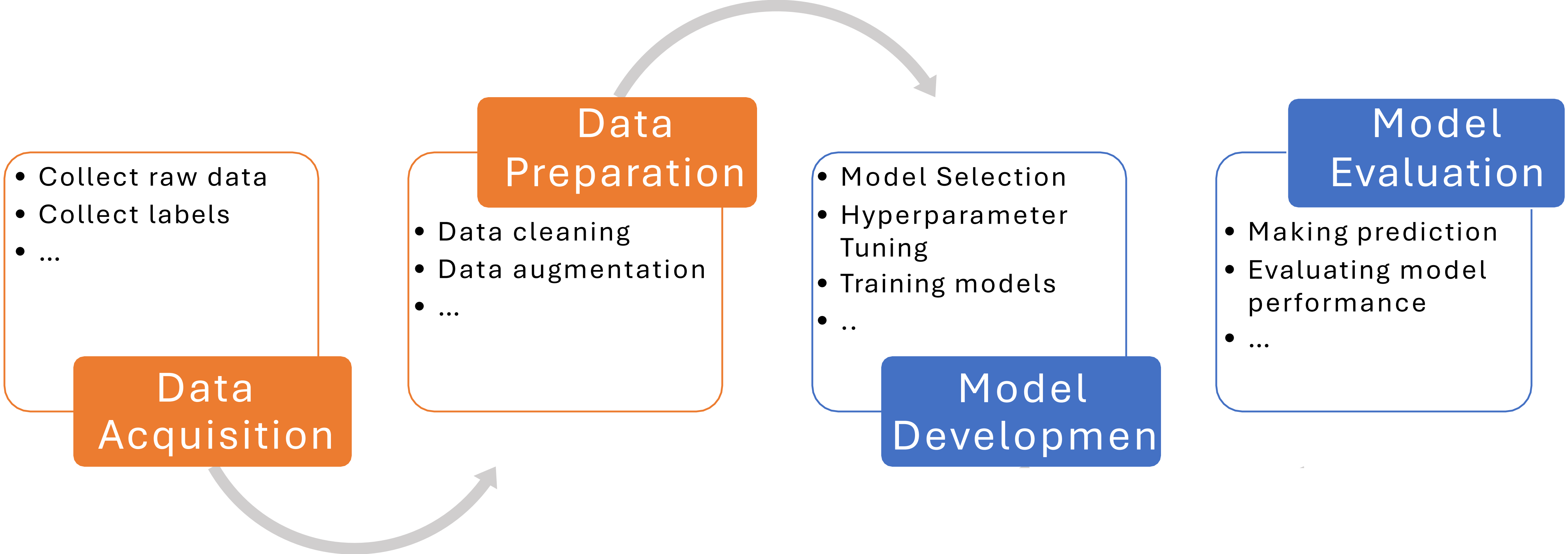
- Data curation and preparation for DB/ML
  - Data parsing
  - Data cleaning
  - Data labeling

- Retrieval Augmented Generation (RAG)

The ML lifecycle

"Only a fraction of real-world ML systems is composed of ML code" [1]

ML ≈ Model + Data



[1] Sculley, David, et al. "Hidden technical debt in machine learning systems." NeurIPS 2015

Data is the bottleneck

ML ≈ Model + Data

Model is gradually commoditized

- Transformers for “all” tasks
- Out-of-the-box invocation of ML libraries gives decent results

Data remains the bottleneck

- Collecting and storing raw data is becoming cheaper
- Turning them into ML-ready datasets is not

https://www.techspot.com/news/98600-chatgpt-powered-hidden-army-contractors-making-15-hour.html

Parsing unstructured data

- **Parsing: unstructured >> structured data**

LlamaIndex
UNSTRUCTURED
Airparser.
Amazon Textract
DOCSUMO Document AI
parsio
Nanonets

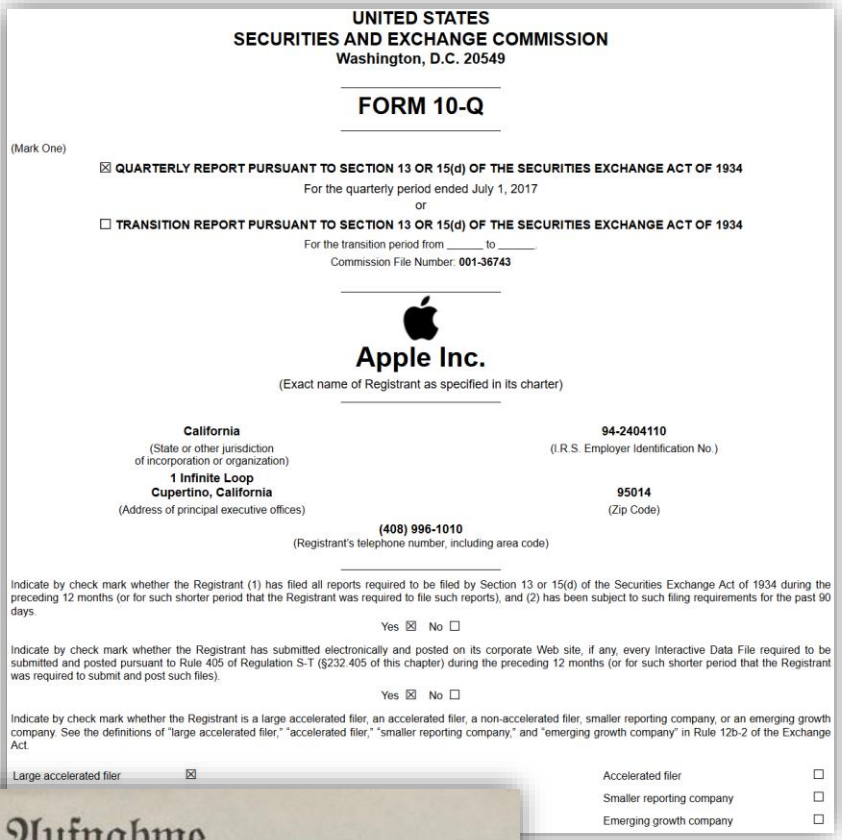
- Common approaches:
  - Rule based parsing: regex, HTML tags
  - Computer-vision-based parsing
  - NLP based parsing
  - LLM based parsing

```python
import re

def extract_emails(text):
    pattern = r'\b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.[A-Z|a-z]{2,7}\b'
    return re.findall(pattern, text)

sample_text = "Contact us at john.doe@example.com or support@company.org for assistanc
emails = extract_emails(sample_text)
print(emails)
# Output: ['john.doe@example.com', 'support@company.org']
```

Rule-based parsing

- Using per-template, pre-defined rules
  - E.g., name = row 2 char 4 to char 10
  - Pixel(10, 10) to Pixel(100, 200)
  - Search keyword = “Zip Code”
- How to define the rules?
  - Manual scripting (when there IS a template)
  - For dynamic/noisy inputs:

    ML based vision, NL solutions



UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549

FORM 10-Q

(Mark One)

☒ QUARTERLY REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the quarterly period ended July 1, 2017
or
☐ TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
For the transition period from ______ to ______
Commission File Number: 001-36743

Apple Inc.
(Exact name of Registrant as specified in its charter)

California
(State or other jurisdiction of incorporation or organization)

94-2404110
(I.R.S. Employer Identification No.)

1 Infinite Loop
Cupertino, California
(Address of principal executive offices)

95014
(Zip Code)

(408) 996-1010
(Registrant's telephone number, including area code)

Indicate by check mark whether the Registrant (1) has filed all reports required to be filed by Section 13 or 15(d) of the Securities Exchange Act of 1934 during the preceding 12 months (or for such shorter period that the Registrant was required to file such reports), and (2) has been subject to such filing requirements for the past 90 days.

Yes ☒ No ☐

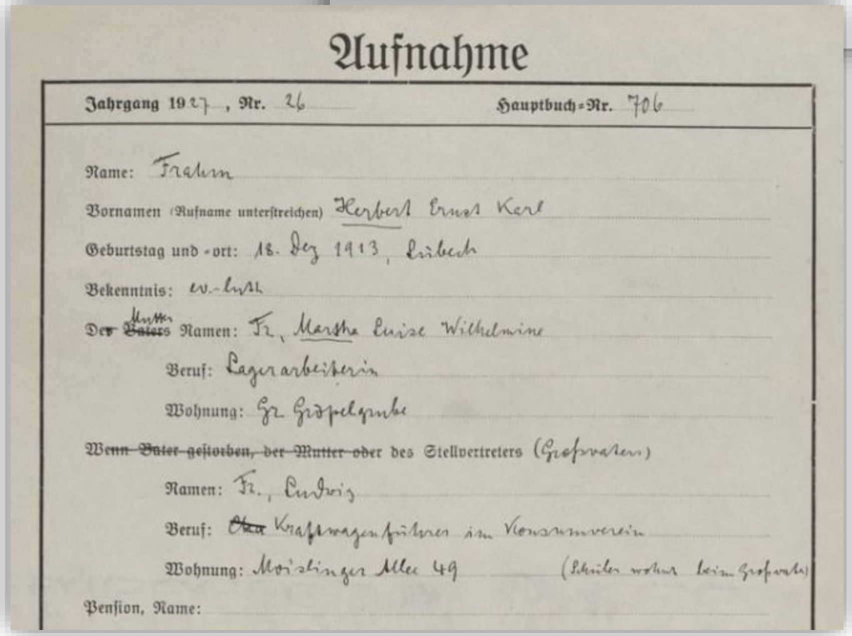
Indicate by check mark whether the Registrant has submitted electronically and posted on its corporate Web site, if any, every Interactive Data File required to be submitted and posted pursuant to Rule 405 of Regulation S-T (§232.405 of this chapter) during the preceding 12 months (or for such shorter period that the Registrant was required to submit and post such files).

Yes ☒ No ☐

Indicate by check mark whether the Registrant is a large accelerated filer, an accelerated filer, a non-accelerated filer, smaller reporting company, or an emerging growth company. See the definitions of "large accelerated filer," "accelerated filer," "smaller reporting company," and "emerging growth company" in Rule 12b-2 of the Exchange Act.

Large accelerated filer ☒
Accelerated filer ☐
Smaller reporting company ☐
Emerging growth company ☐

Template



Aufnahme

Jahrgang 1927, Nr. 26 Hauptbuch-Nr. 706

Name: Frahm

Vornamen (Rufname unterstreichen) Herbert Ernst Karl

Geburtstag und -ort: 18. Dez 1913, Lübeck

Bekenntnis: ev.-luth.

Der Mutter ~~Vaters~~ Namen: Fr. Martha Luise Wilhelmine

Beruf: Lagerarbeiterin

Wohnung: Gr. Gropelgrube

~~Wenn Vater gestorben, der Mutter oder~~ des Stellvertreters (Großvaters)

Namen: Fr. Ludwig

Beruf: ~~Obm~~ Kraftwagenführer im Konsumverein

Wohnung: Moislinger Allee 49 (Schüler wohnt beim Großvater)

Pension, Name:

Handwritten

Parsing unstructured data

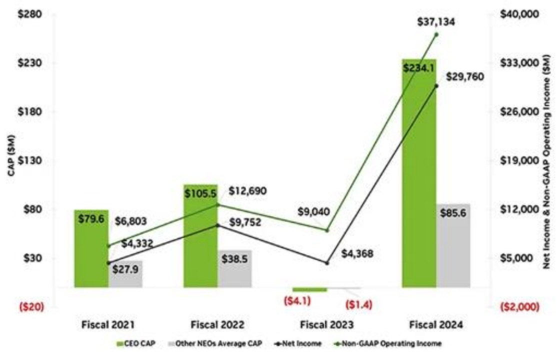
Original Document

Relationships Between CAP and Financial Performance

The following graphs illustrate how CAP for our NEOs aligns with the Company's financial performance measures as detailed in the Pay Versus Performance table above for each of Fiscal 2021, 2022, 2023, and 2024, as well as between the TSRs of NVIDIA and the Nasdaq100 Index, reflecting the value of a fixed $100 investment beginning with the market close on January 24, 2020, the last trading day before our Fiscal 2021, through and including the end of the respective listed fiscal years.

NEO CAP versus TSR

NEO CAP versus Net Income & Non-GAAP Operating Income



All information provided above under the "Pay Versus Performance" heading will not be deemed to be incorporated by reference into any filing of the Company under the Securities Act of 1933, as amended, or the Securities Exchange Act of 1934, as amended, whether made before or after the date hereof and irrespective of any general incorporation language in any such filing, except to the extent the Company specifically incorporates such information by reference.

63

Parsing Results

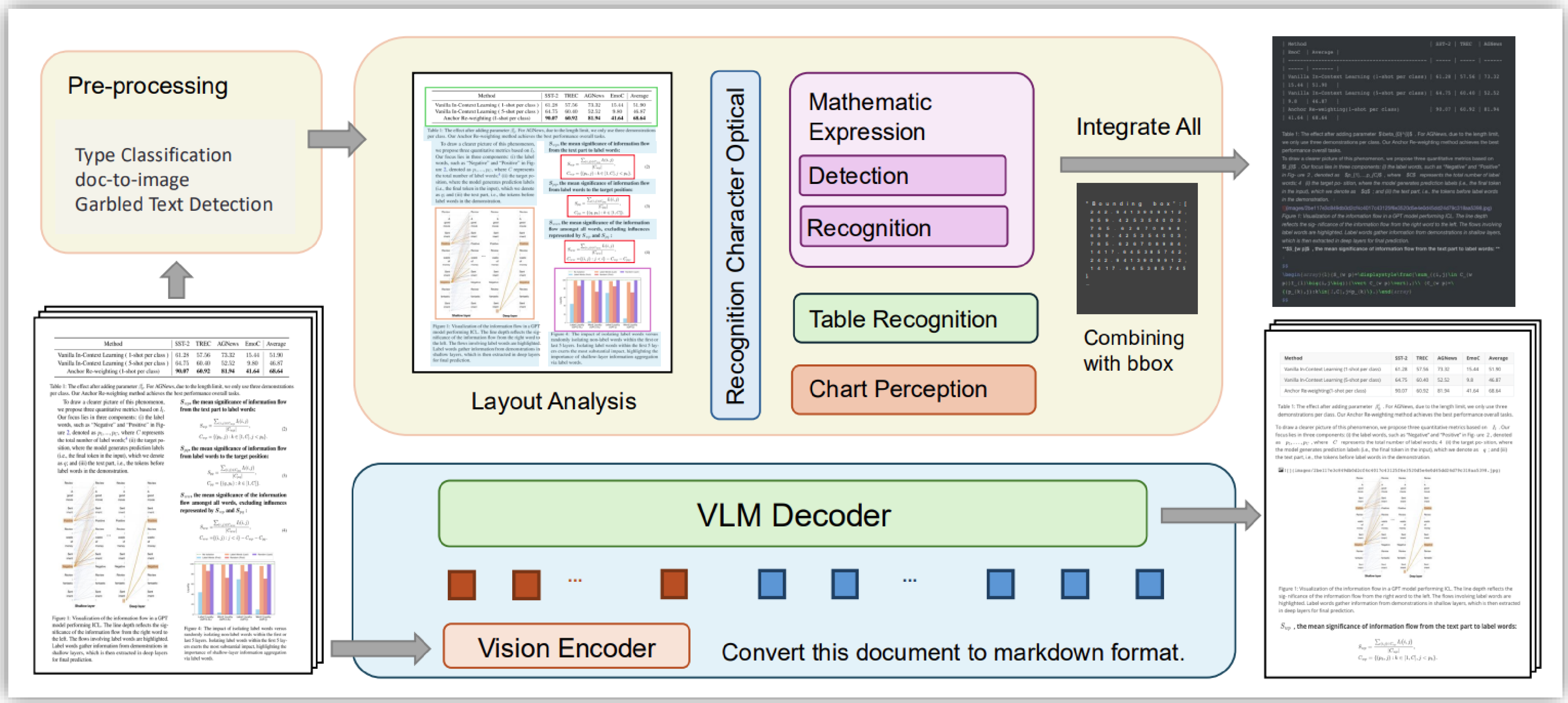
```
# Relationships Between CAP and Financial Performance

The following graphs illustrate how CAP for our NEOs aligns with the Company's financial
performance measures as detailed in the Pay Versus Performance table above for each of
Fiscal 2021, 2022, 2023, and 2024, as well as between the TSRs of NVIDIA and the Nasdaq100
Index, reflecting the value of a fixed $100 investment beginning with the market close on
January 24, 2020, the last trading day before our Fiscal 2021, through and including the
end of the respective listed fiscal years.

## NEO CAP versus TSR

| Fiscal Year | CEO CAP | Other NEOs Average CAP | NVIDIA TSR | Nasdaq100 Index TSR |
|-------------|---------|------------------------|------------|---------------------|
| Fiscal 2021 | $79.6   | $27.9                  | $207.79    | $141.39             |
| Fiscal 2022 | $105.5  | $38.5                  | $365.66    | $158.12             |
| Fiscal 2023 | ($4.1)  | ($1.4)                 | $326.34    | $133.09             |
| Fiscal 2024 | $234.1  | $85.6                  | $978.42    | $190.57             |

*Note: Values on right y-axis range from ($20) to $1,120*

## NEO CAP versus Net Income & Non-GAAP Operating Income

| Fiscal Year | CEO CAP | Other NEOs Average CAP | Net Income | Non-GAAP Operating Income |
|-------------|---------|------------------------|------------|---------------------------|
| Fiscal 2021 | $79.6   | $27.9                  | $4,332     | $6,803                    |
| Fiscal 2022 | $105.5  | $38.5                  | $9,752     | $12,690                   |
| Fiscal 2023 | ($4.1)  | ($1.4)                 | $4,368     | $9,040                    |
| Fiscal 2024 | $234.1  | $85.6                  | $29,760    | $37,134                   |

*Note: Values on right y-axis range from ($2,000) to $40,000*

All information provided above under the "Pay Versus Performance" heading will not be
deemed to be incorporated by reference into any filing of the Company under the Securities
Act of 1933, as amended, or the Securities Exchange Act of 1934, as amended, whether made
before or after the date hereof and irrespective of any general incorporation language in
any such filing, except to the extent the Company specifically incorporates such
information by reference.

63
```

Example from LlamaParse

More complex parsing:

- Tables, figures, charts
- Complex layouts
- Large multi-modal models

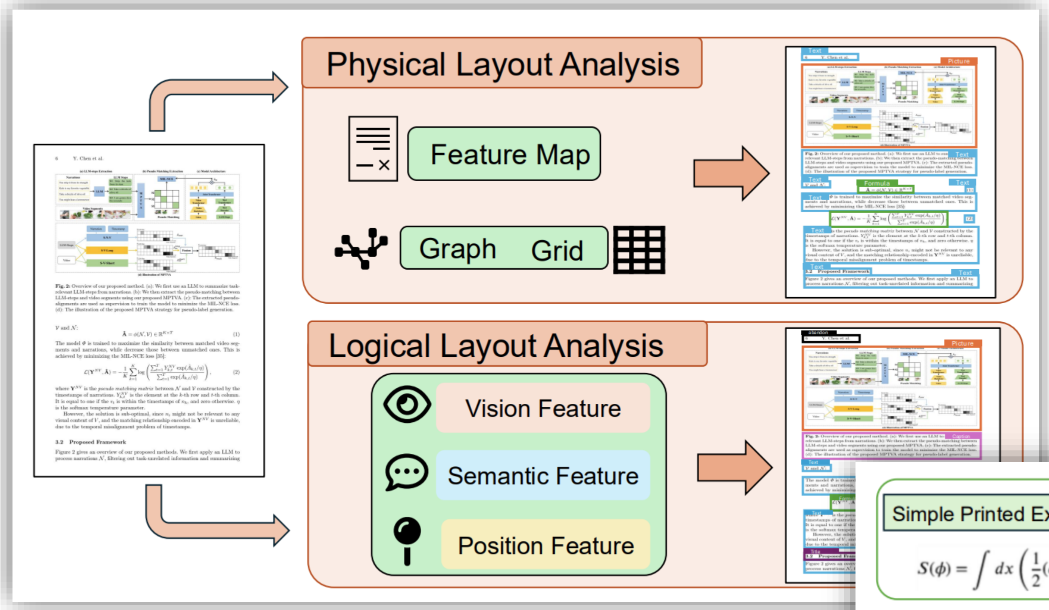
Computer-vision-based parsing



CV-based parsing uses pretrained models to extract structural information from images

Zhang et al. Document Parsing Unveiled: Techniques, Challenges, and Prospects for Structured Information Extraction. arXiv 2410.21169v2

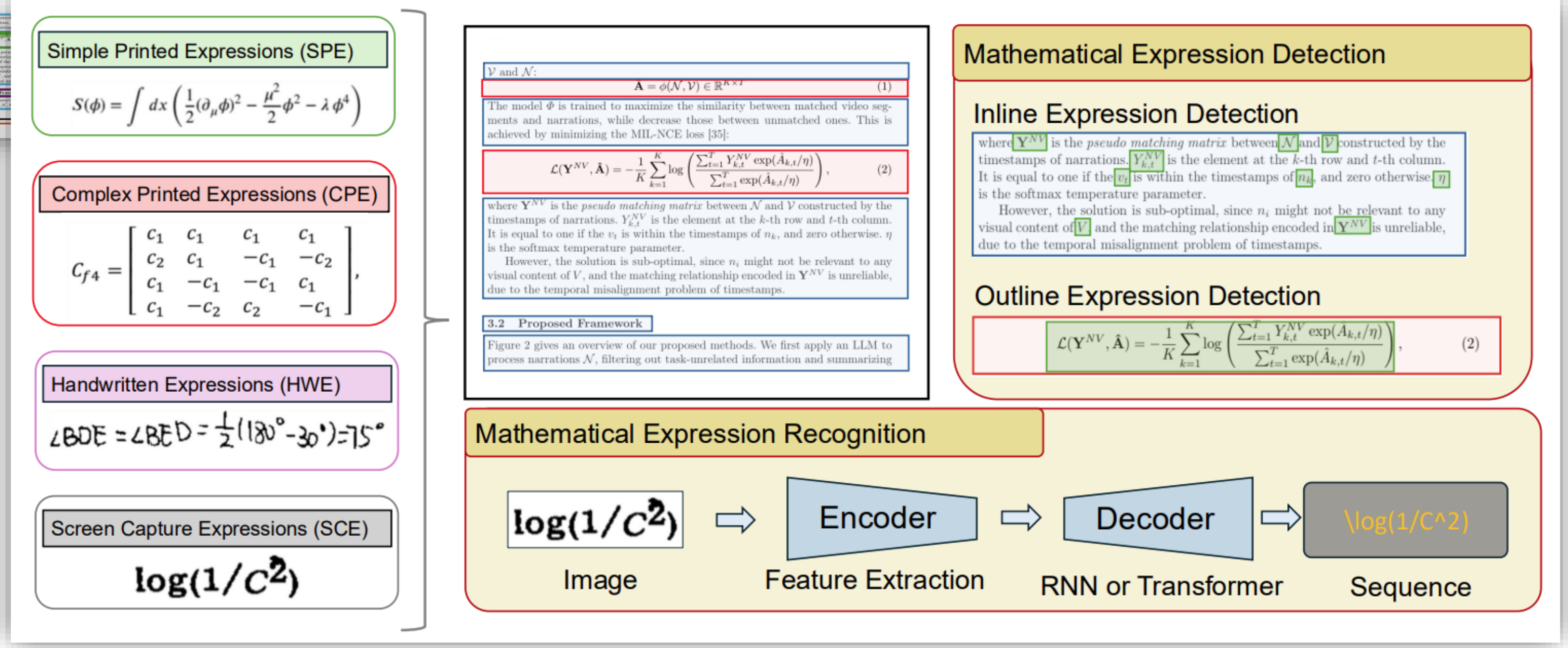
Computer-vision-based parsing



Layout analysis

Some tasks use standalone, specific models:

- Layout analysis (extract bounding boxes)
- Optical Character Recognition (OCR)
- Math formula recognition (OCR)
- Table and chart recognition



Computer-vision-based parsing

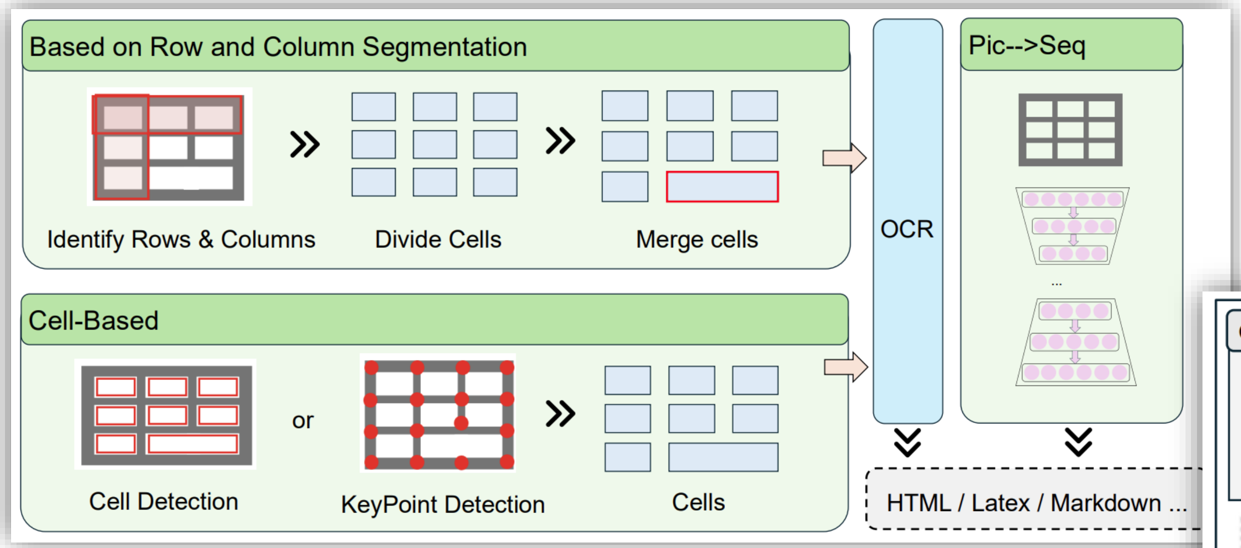
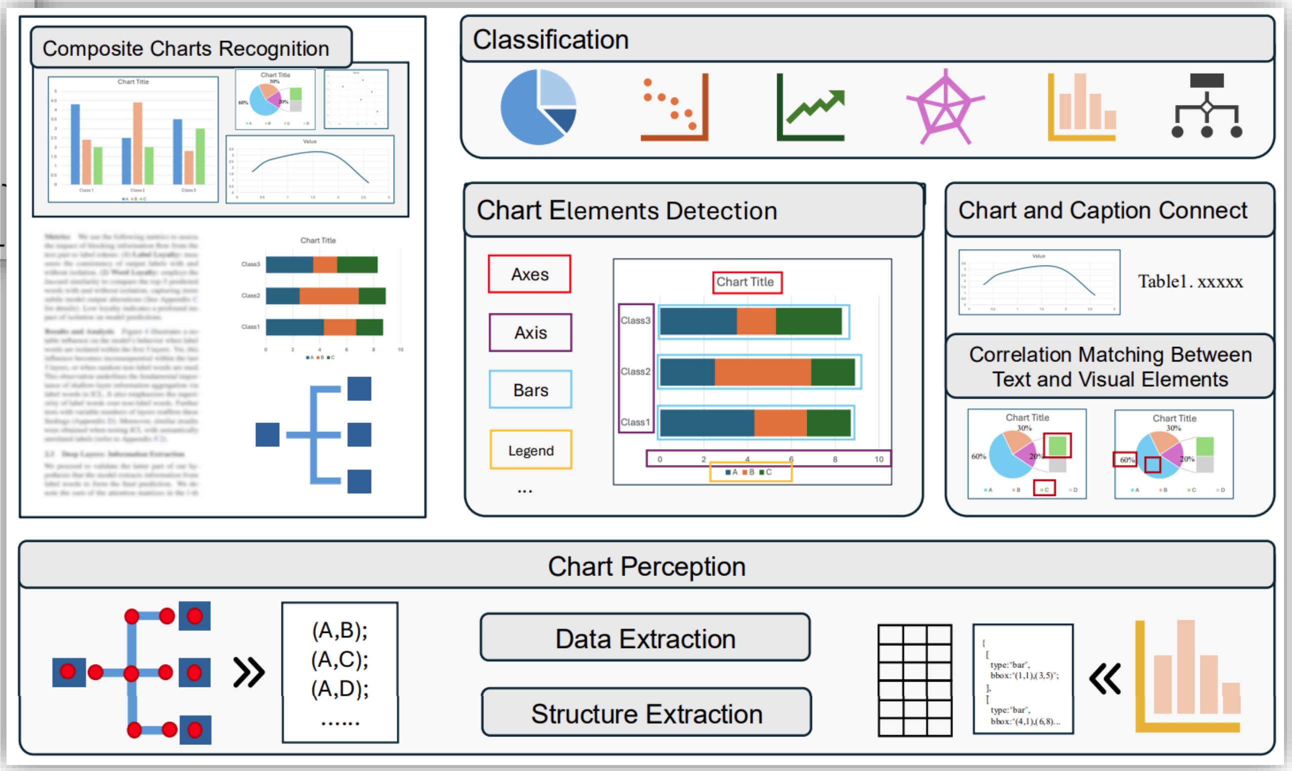


Table recognition

Chart recognition



NL-based data parsing

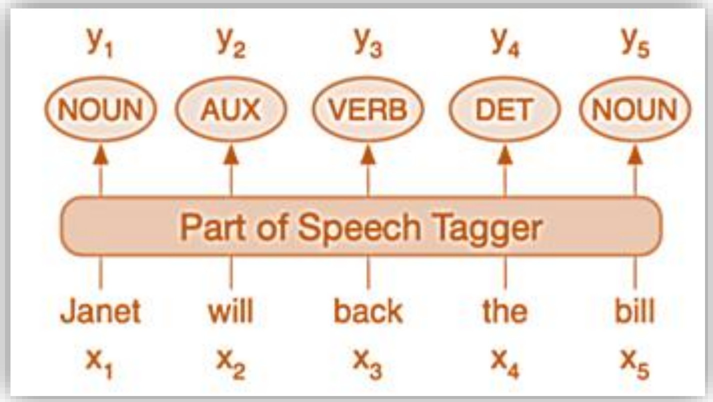
- Common text pre-processing
  - Cleaning (removing words like stopwords, emojis, punctuation, etc.)
  - Normalization
  - Lemmatization & stemming
- Tools: Regex, NLTK, spaCy, OpenNLP

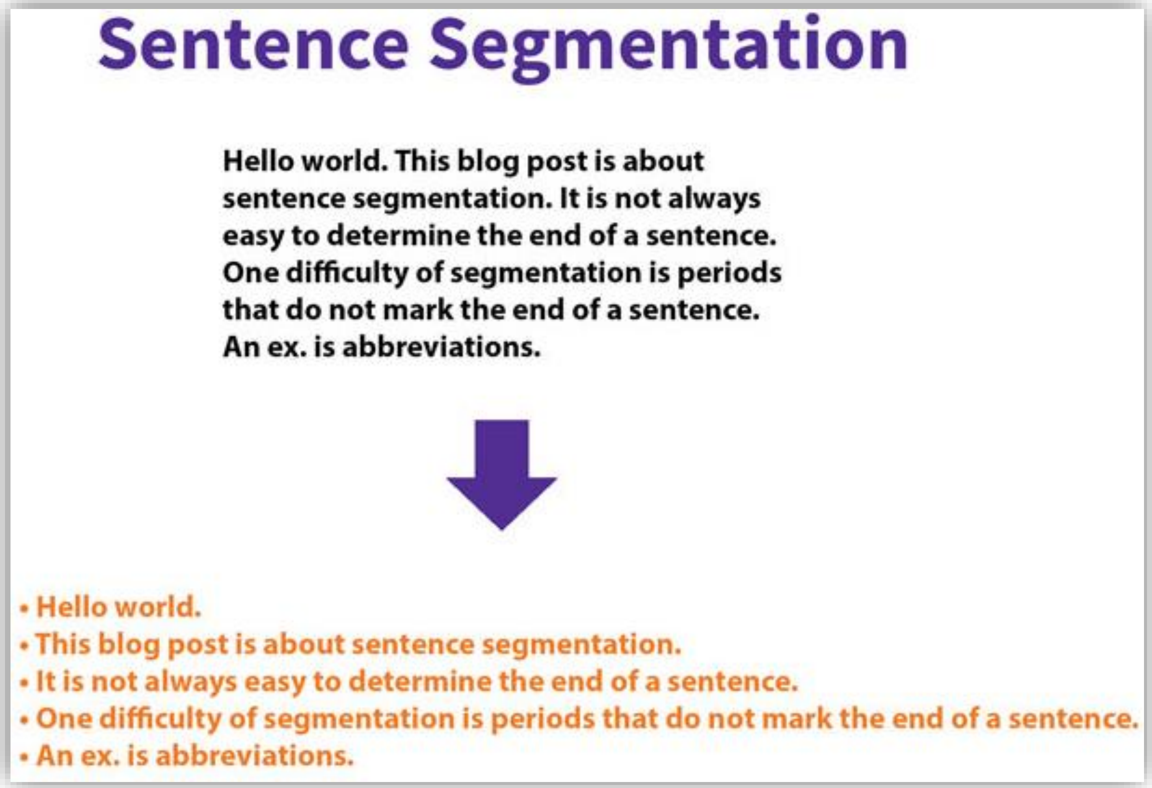
	original_word	stemmed_word
0	trouble	troubl
1	troubled	troubl
2	troubles	troubl
3	troublemsome	troublemsom

```
sample = "Hello @gabe_flomo 👋🏾, I still want us to hit that new sushi spot??? LM
K when you're free cuz I can't go this or next weekend since I'll be swimming!!!
#sushiBros #rawFish #🍣"
print(pipeline(sample))

# output"
hello still want us hit new sushi spot lmk free cuz cant go next weekend since i
ll swim"
```

NL-based data parsing

- Segmentation & tagging
  - Some useful applications: detecting title etc.





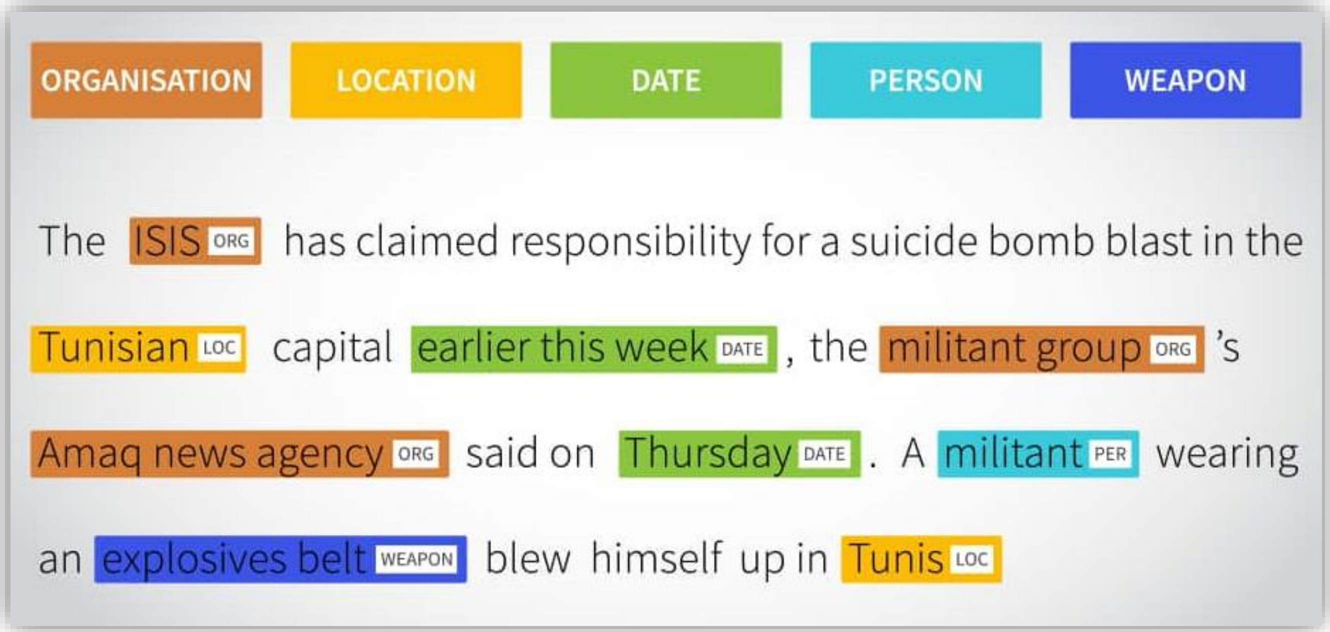
NL-based data parsing

- Segmentation & tagging
  - Some useful applications: detecting title etc.

- Name entity recognition
  - Person: Steve Jobs
  - Company: Apple
  - Location: California
  - Column names often use entity names

NL-based data parsing

- Segmentation & tagging
  - Some useful applications: detecting title etc.
- Name entity recognition
  - Person: Steve Jobs
  - Company: Apple
  - Location: California
  - Column names often use entity names
- Extraction (column = value)
  - Rule-based
  - RAGs (later)



Real documents are complex

- Complex layout
- Complex tables
- Noisy data
- Variance

Contingencies

The Company is subject to various legal proceedings and claims that have arisen in the ordinary course of business and that have not been fully adjudicated, as further discussed in Part II, Item 1 of this Form 10-Q under the heading "Legal Proceedings" and in Part II, Item 1A of this Form 10-Q under the heading "Risk Factors." In the opinion of management, there was not at least a reasonable possibility the Company may have incurred a material loss, or a material loss in excess of a recorded accrual, with respect to loss contingencies for asserted legal and other claims. However, the outcome of litigation is inherently uncertain. Therefore, although management considers the likelihood of such an outcome to be remote, if one or more of these legal matters were resolved against the Company in a reporting period for amounts in excess of management's expectations, the Company's consolidated financial statements for that reporting period could be materially adversely affected.

Apple Inc. v. Samsung Electronics Co., Ltd., et al.

On August 24, 2012, a jury returned a verdict awarding the Company $1.05 billion in its lawsuit against Samsung Electronics Co., Ltd. and affiliated parties in the United States District Court, Northern District of California, San Jose Division. On March 6, 2014, the District Court entered final judgment in favor of the Company in the amount of approximately $930 million. On May 18, 2015, the U.S. Court of Appeals for the Federal Circuit affirmed in part, and reversed in part, the decision of the District Court. As a result, the Court of Appeals ordered entry of final judgment on damages in the amount of approximately $548 million, with the District Court to determine supplemental damages and interest, as well as damages owed for products subject to the reversal in part. Samsung paid

		Three Months Ended		Nine Months Ended	
Comprehensive Income Components	**Financial Statement Line Item**	**July 1, 2017**	**June 25, 2016**	**July 1, 2017**	**June 25, 2016**
Unrealized (gains)/losses on derivative instruments:					
Foreign exchange contracts	Revenue	$ (148)	$ (131)	$ (657)	$ (785)
	Cost of sales	(73)	106	(630)	(419)
	Other income/(expense), net	(364)	(112)	(127)	(123)
Interest rate contracts	Other income/(expense), net	—	3	3	10
		(585)	(134)	(1,411)	(1,317)
Unrealized (gains)/losses on marketable securities	Other income/(expense), net	(48)	(20)	(37)	129
Total amounts reclassified from AOCI		$ (633)	$ (154)	$ (1,448)	$ (1,188)

LLM-based parsing

- **One model for all?**
- Large multi-modal models, e.g., GPT-4o

Note 7 – Shareholders' Equity

Dividends

The Company declared and paid cash dividends per share during the periods presented as follows:

	Dividends Per Share	Amount (in millions)
2017:		
Third quarter	$ 0.63	$ 3,281
Second quarter	0.57	2,988
First quarter	0.57	3,042
Total cash dividends declared and paid	$ 1.77	$ 9,311
2016:		
Fourth quarter	$ 0.57	$ 3,071
Third quarter	0.57	3,117
Second quarter	0.52	2,879
First quarter	0.52	2,898
Total cash dividends declared and paid	$ 2.18	$ 11,965

Future dividends are subject to declaration by the Board of Directors.

Parse into markdown

- Drawbacks:
  - Expensive
  - Hard to instruct

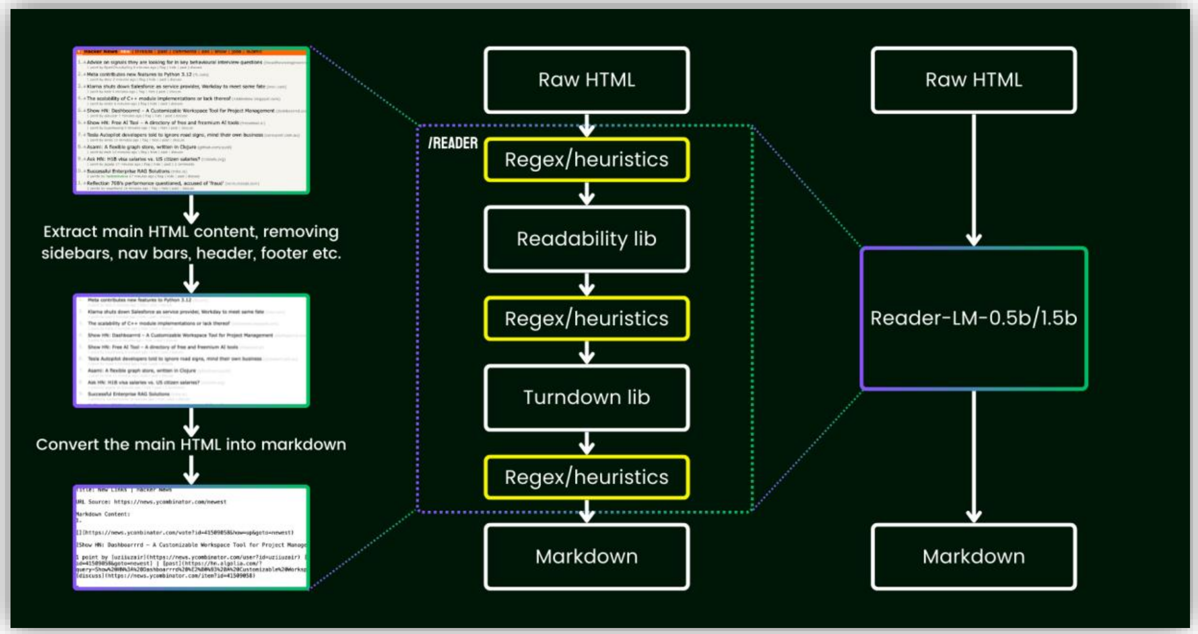
ChatGPT-4o

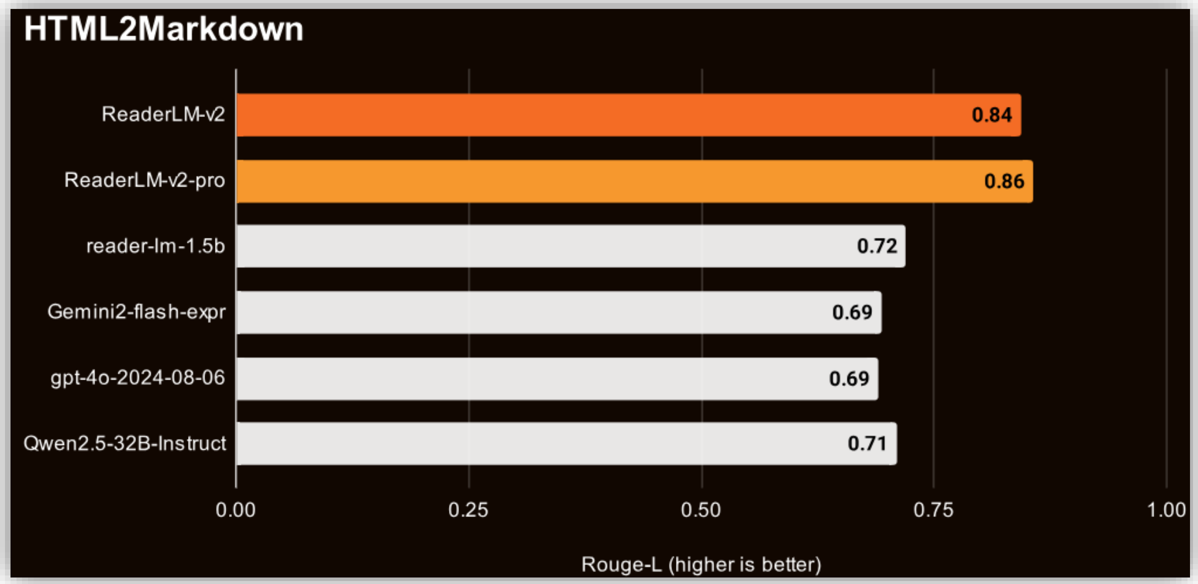
markdown Copy Edit

```
# Note 7 – Shareholders’ Equity

## Dividends

The Company declared and paid cash dividends per share during the periods presented as follow

### 2017:
| Quarter        | Dividends Per Share | Amount (in millions) |
|--------------|--------------------|---------------------|
| Third quarter | $0.63               | $3,281               |
| Second quarter | $0.57               | $2,988               |
| First quarter | $0.57               | $3,042               |
| **Total cash dividends declared and paid** | **$1.77** | **$9,311** |

### 2016:
| Quarter        | Dividends Per Share | Amount (in millions) |
|--------------|--------------------|---------------------|
| Fourth quarter | $0.57               | $3,071               |
| Third quarter  | $0.57               | $3,117               |
| Second quarter | $0.52               | $2,879               |
| First quarter  | $0.52               | $2,898               |
| **Total cash dividends declared and paid** | **$2.18** | **$11,965** |

Future dividends are subject to declaration by the Board of Directors.
```

LLM-based parsing

- **One model for all?**
- Large multi-modal models, e.g., GPT-4o
  - Expensive
  - Hard to instruct
- Small Language Models (SLMs)
  - Small = cheap
  - Instruction tuned for data parsing
  - E.g., ReaderLM-v2 from Jina AI





https://jina.ai/news/readerlm-v2-frontier-small-language-model-for-html-to-markdown-and-json

There is no free lunch

- No single method can guarantee 100% correct
- Hard to verify

- There are ML/AI solutions to alleviate these problems
  - Human-in-the-loop systems and applications design
  - Multi-agent framework to cross validate
  - Active learning to reduce annotation
  - Synthetic data generation to improve parsing robustness

Data curation and RAGs: outline

- Data curation and preparation for DB/ML
  - Data parsing
  - Data cleaning
  - Data labeling

- Retrieval Augmented Generation (RAG)

Data cleaning and ML

Cleaning "before" ML:

- Perform cleaning independently of the downstream ML applications; leverage user-specified signals or data-driven approaches
- Example: HoloClean: Holistic Data Repairs with Probabilistic Inference
  - Also an example of using ML for data cleaning

Reading: From Cleaning Before ML to Cleaning For ML

Data cleaning and ML

Cleaning "for" ML:

- Leverage the downstream ML model or application to define cleaning signals that incorporates high-level semantics
- Why is this a good idea?
  - Clean datasets that contain fully correct attributes are rarely available
  - Data cleaning can sometimes negatively impact the performance of ML models
    - CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks
- Example: BoostClean: Automated Error Detection and Repair for Machine Learning

Reading: From Cleaning Before ML to Cleaning For ML

Common data problems

Incomplete

Country	UN R/P 10%[4]	UN R/P 20%[5]	World Bank Gini (%)[6]	WB Gini (year)	CIA R/P 10%[7]	Year	CIA Gini (%)[8]	CIA Gini (year)	GPI Gini (%)[9]
Seychelles			65.8	2007					
Comoros			64.3	2004					
Namibia	106.6	56.1	63.9	2004	129.0	2003	59.7	2010	
South Africa	33.1	17.9	63.1	2009	31.9	2000	65.0	2005	
Botswana	43.0	20.4	61.0	1994			63	1993	
Haiti	54.4	26.6	59.2	2001	68.1	2001	59.2	2001	
Angola			58.6	2000					62.0
Honduras	59.4	17.2	57.0	2009	35.2	2003	57.7	2007	

Adapted from Intro to Data Cleaning lecture from Xu Chu

Common data problems

Inconsistent

Financial

Employee	Salary
John	1000

Employee → Salary

Human Resources

Employee	Salary
John	2000
Mary	3000

Employee → Salary

Target Database

Employee	Salary
John	**1000**
John	**2000**
Mary	3000

Employee → Salary

Mapping

Financial(e,s) $\subseteq$ Global(e,s)
HumanRes(e,s) $\subseteq$ Global(e,s)

Common data problems

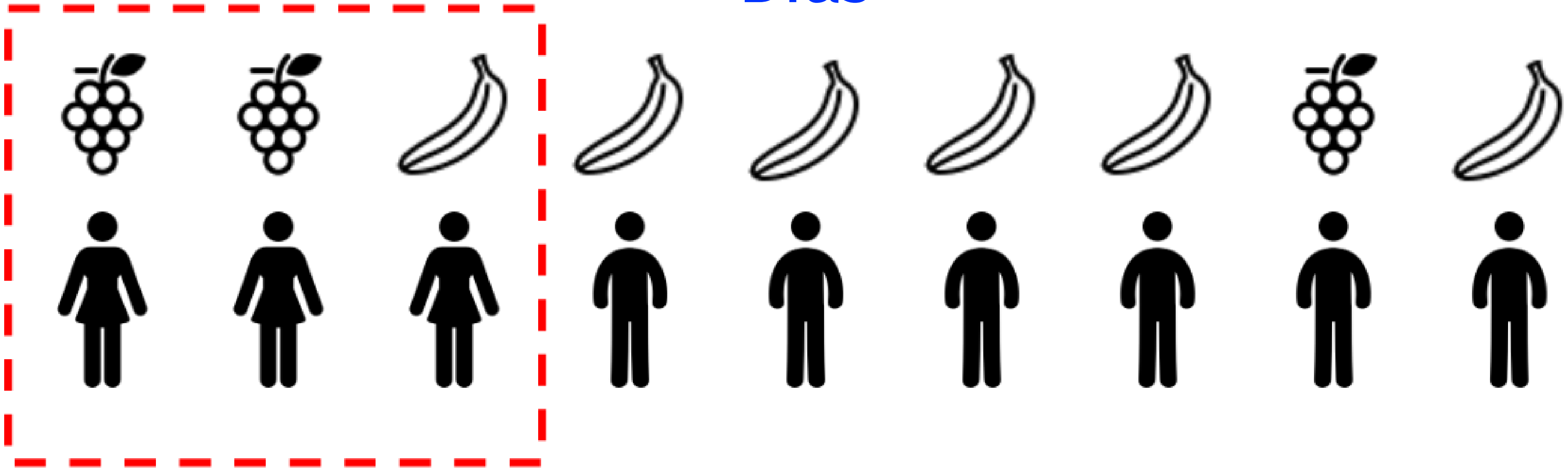
Inaccurate

Common data problems

Outliers

Common data problems

Bias



Model amplifies data biases
Example: Buolamwini and Gebru (2018). Gender Shades

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Dirty data is costly

UNITED STATES POSTAL SERVICE

- Address errors caused 6.8 billion undelivered mails in 2013
- Estimated $1.5 billion spent on processing
- At least $3.4 billion wasted postage

Harvard Business Review

DATA

Bad Data Costs the U.S. $3 Trillion Per Year

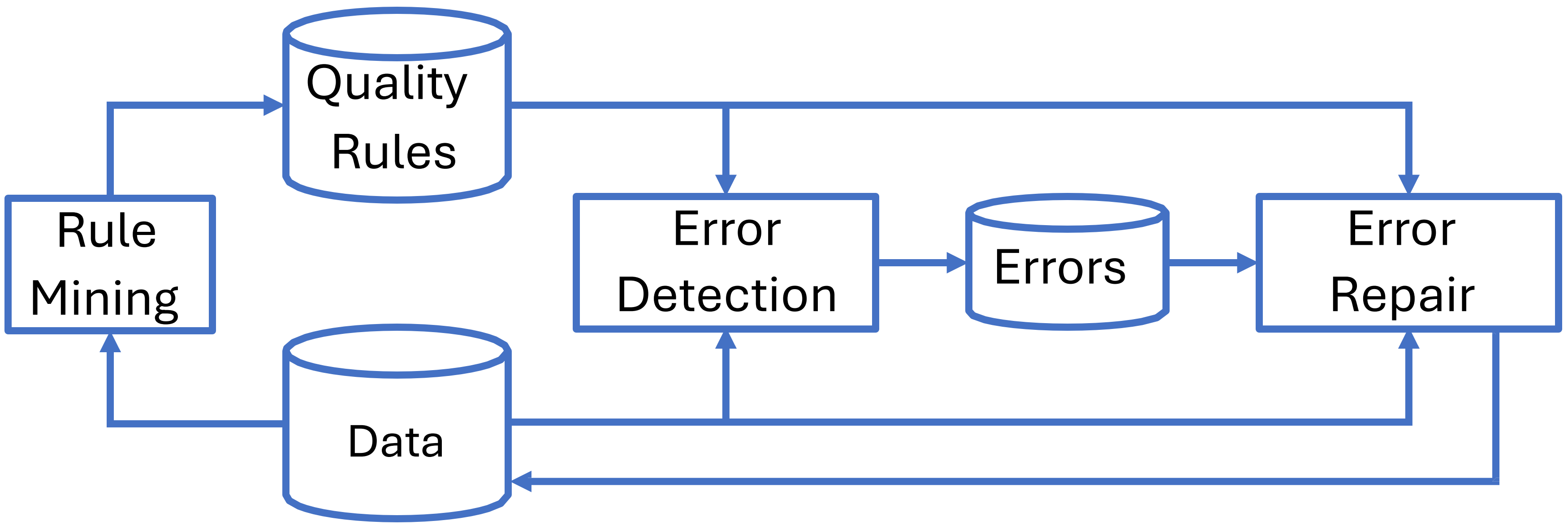
by Thomas C. Redman

SEPTEMBER 22, 2016

Data cleaning for structured data

- Detect and repair errors in a structured dataset
  - Discovering denial constraints. [VLDB’13]
  - HoloClean: Holistic Data Repairs with Probabilistic Inference. [VLDB’17]
- Data cleaning and machine learning
  - Cleaning before ML
  - Cleaning for ML

Two tasks in data cleaning



- Detection: A minimal set of cells that cannot coexist together
- Repair: A set of cell updates to resolve the violations

Data quality rules

	Name	ID	LVL	ZIP	ST	SAL
t_1	Alice	ID1	5	10001	NM	90K
t_2	Bob	ID2	6	87101	NM	80K
t_3	Chris	ID3	4	10001	NY	80K
t_4	Dave	ID4	1	90057	CA	20K
t_5	Frank	ID5		90057	CA	50K

R1: Two persons with the same ZIP live in the same ST

Data quality rules

	Name	ID	LVL	ZIP	ST	SAL
t_1	Alice	ID1	5	10001	NM	90K
t_2	Bob	ID2	6	87101	NM	80K
t_3	Chris	ID3	4	10001	NY	80K
t_4	Dave	ID4	1	90057	CA	20K
t_5	Frank	ID5		90057	CA	50K

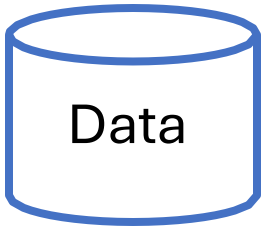
R2: LVL should not be empty

Data quality rules

	Name	ID	LVL	ZIP	ST	SAL
t_1	Alice	ID1	5	10001	NM	90K
t_2	Bob	ID2	6	87101	NM	80K
t_3	Chris	ID3	4	10001	NY	80K
t_4	Dave	ID4	1	90057	CA	20K
t_5	Frank	ID5		90057	CA	50K

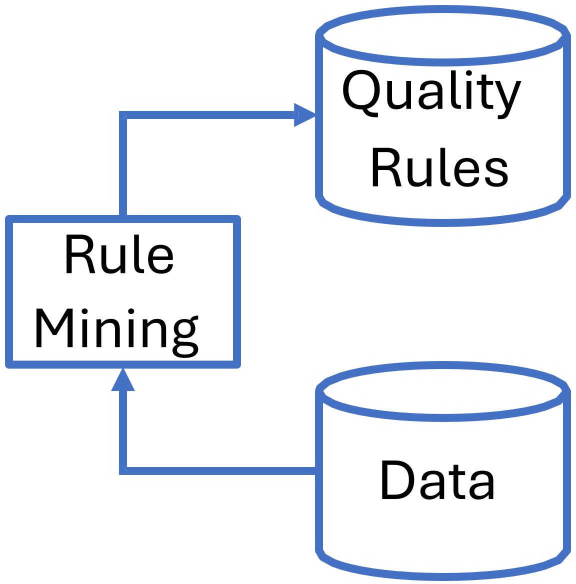
R3: People with a higher LVL earn more SAL in the same ST

Rule-based data cleaning



Name	ZIP	ST
Alice	10001	NM
Bob	87101	NM
Chris	10001	NY

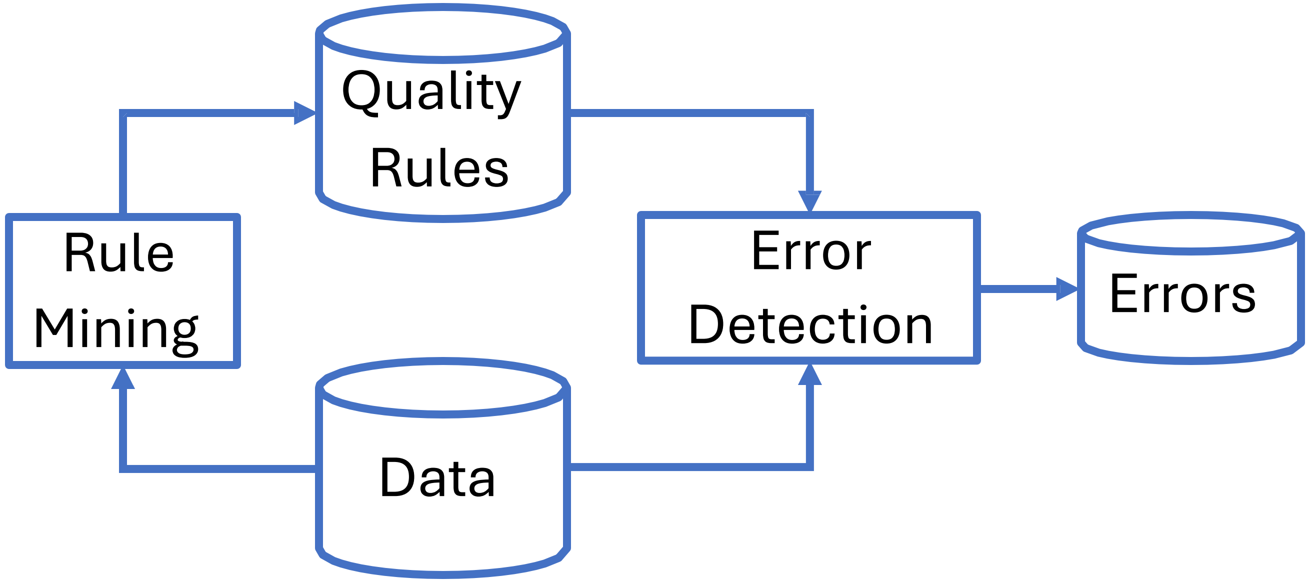
Rule-based data cleaning



Name	ZIP	ST
Alice	10001	NM
Bob	87101	NM
Chris	10001	NY

Two persons with the same ZIP live in the same ST

Rule-based data cleaning

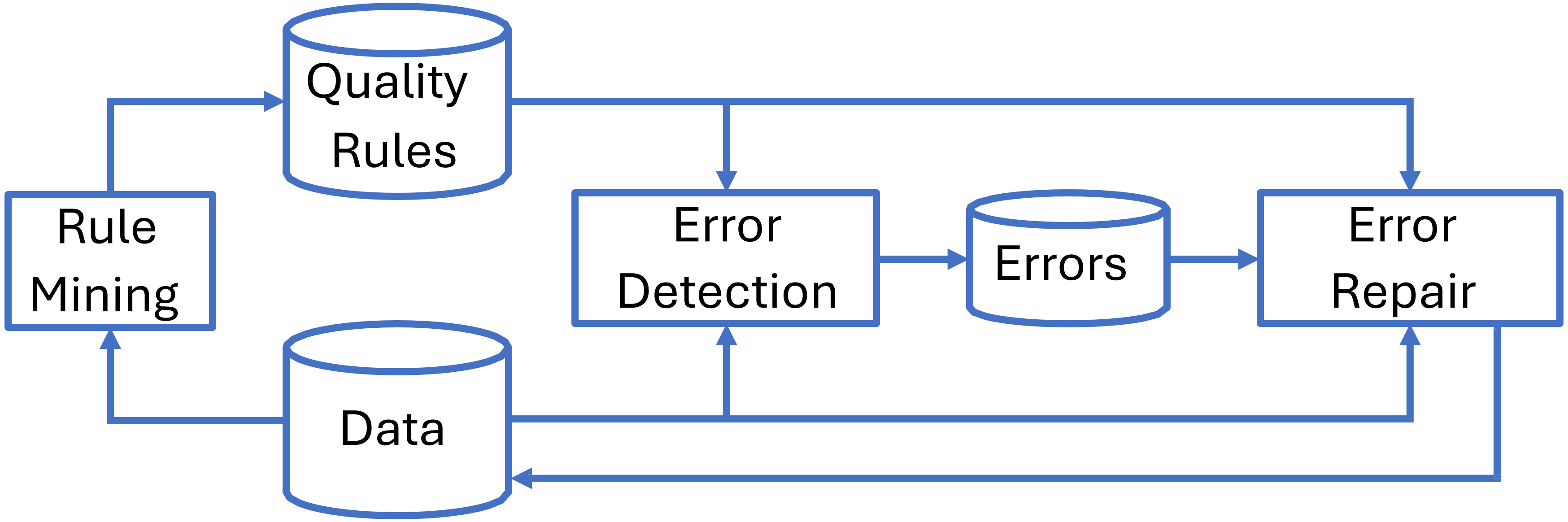


Name	ZIP	ST
Alice	10001	NM
Bob	87101	NM
Chris	10001	NY

Two persons with the same ZIP live in the same ST

Adapted from Intro to Data Cleaning lecture from Xu Chu

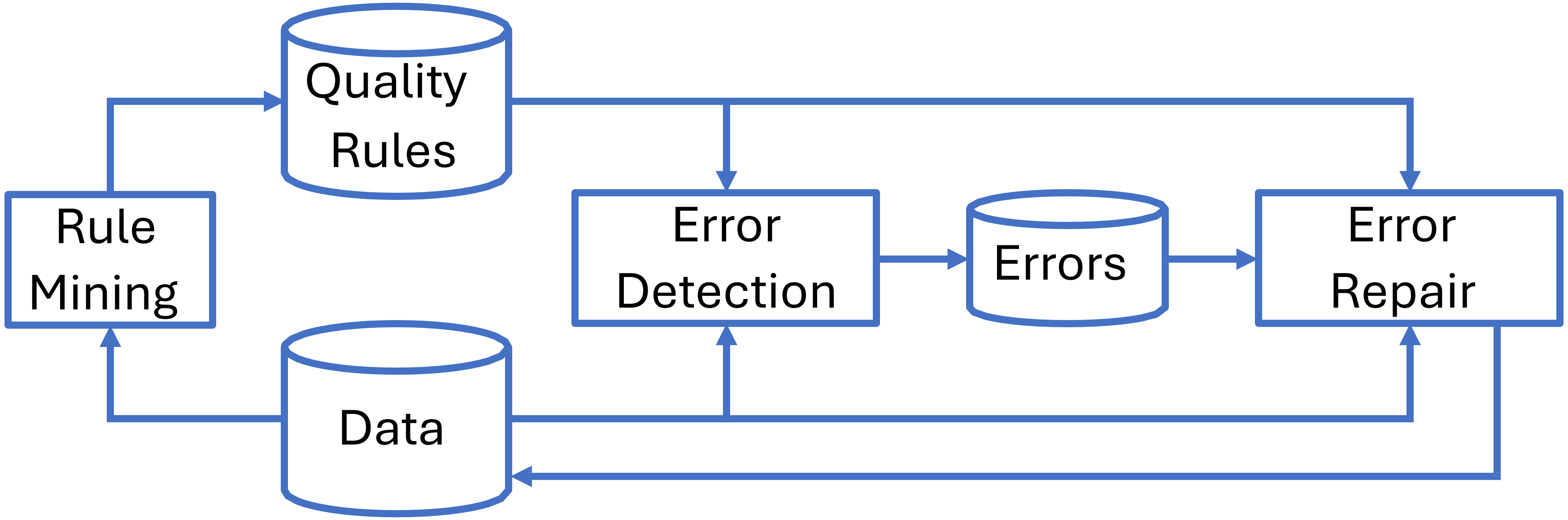
Rule-based data cleaning



Name	ZIP	ST
Alice	10001	NY
Bob	87101	NM
Chris	10001	NY

Two persons with the same ZIP live in the same ST

Discovering denial constraints [VLDB’13]



Can ask a domain expert, but takes too much time

Automatically discover quality rules in the form of Denial Constraints

R1: Two persons with the same ZIP live in the same ST

$$\forall t\alpha, t\beta \; \neg(t\alpha.ZIP = t\beta.ZIP \wedge \quad t\alpha.ST \neq t\beta.ST)$$

Examples of discovered DCs

On a tax dataset

$\forall t\alpha\ \neg(t\alpha.ST = \text{“FL”} \wedge \quad t\alpha.ZIP < 30397)$

State Florida’s ZIP code cannot be lower than 30397.

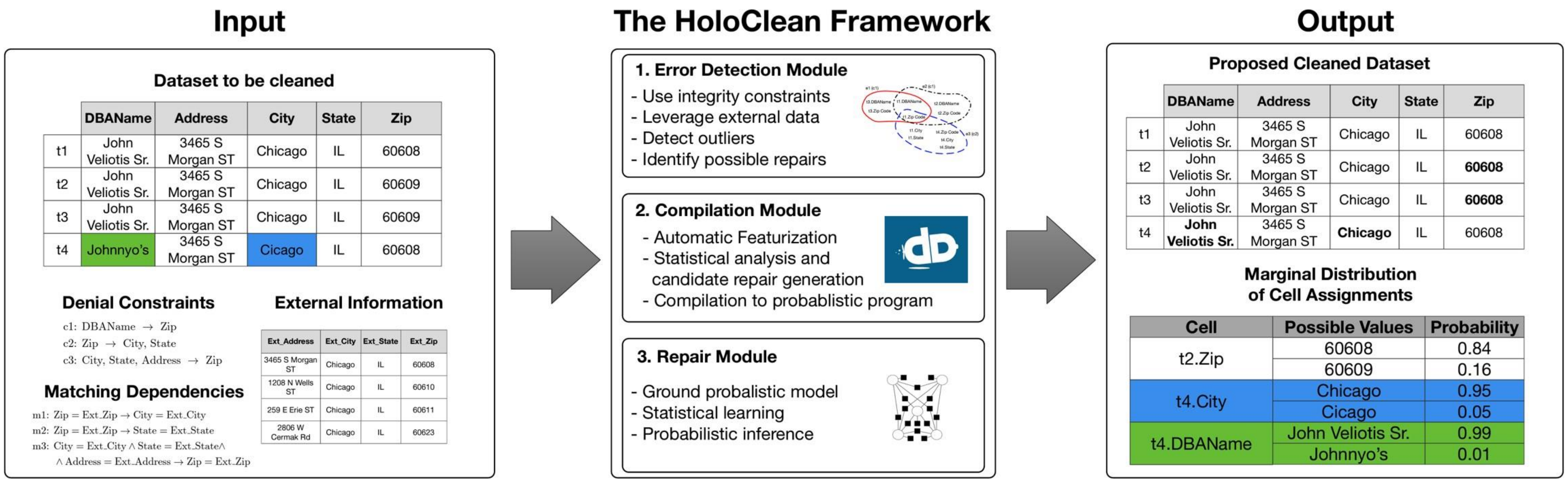
$\forall t\alpha\ \neg(t\alpha.MS \neq \text{“Single”} \wedge \quad t\alpha.STX \neq 0)$

One has to be single to have any single tax exemption.

$\forall t\alpha, t\beta\ \neg(t\alpha.ST = t\beta.ST \wedge \quad t\alpha.SAL < t\beta.SAL \wedge \quad t\alpha.TR > t\beta.TR)$

There cannot exist two persons who live in the same state, but one person earns less salary and has higher tax rate at the same time.

HoloClean: Holistic Data Repairs with Probabilistic Inference [VLDB’17]



Probabilistic model that unifies different signals for repairing a dataset.

Constraints and minimality

Functional dependencies

$$\text{c1: DBAName} \rightarrow \text{Zip}$$
$$\text{c2: Zip} \rightarrow \text{City, State}$$
$$\text{c3: City, State, Address} \rightarrow \text{Zip}$$

	DBAName	AKAName	Address	City	State	Zip
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	***Chicago***	IL	***60608***
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t4	***Johnnyo's***	Johnnyo's	3465 S Morgan ST	***Cicago***	IL	60608

Bohannon et al., 2005, 2007; Kolahi and Lakshmanan , 2005; Bertossi et al., 2011; Chu et al., 2013; 2015 Fagin et al., 2015

Constraints and minimality

Functional dependencies

c1: $\text{DBAName} \rightarrow \text{Zip}$

c2: $\text{Zip} \rightarrow \text{City, State}$

c3: $\text{City, State, Address} \rightarrow \text{Zip}$

	DBAName	AKAName	Address	City	State	Zip
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	**60609**
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	**60609**
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	**60609**
t4	**Johnnyo's**	Johnnyo's	3465 S Morgan ST	**Cicago**	IL	60608

Action: Fewer erroneous than correct cells; perform minimum number of changes to satisfy all constraints

Constraints and minimality

Functional dependencies

$$c1:\ \text{DBAName} \rightarrow \text{Zip}$$
$$c2:\ \text{Zip} \rightarrow \text{City, State}$$
$$c3:\ \text{City, State, Address} \rightarrow \text{Zip}$$

	DBAName	AKAName	Address	City	State	Zip
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	**60609**
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	**60609**
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	**60609**
t4	**Johnnyo's**	Johnnyo's	3465 S Morgan ST	**Cicago**	IL	60608

Error; correct zip code is 60608

Does not fix errors and introduces new ones.

External information

Matching dependencies

m1: $Zip = Ext\_Zip \rightarrow City = Ext\_City$

m2: $Zip = Ext\_Zip \rightarrow State = Ext\_State$

m3: $City = Ext\_City \wedge State = Ext\_State \wedge$
$\wedge\ Address = Ext\_Address \rightarrow Zip = Ext\_Zip$

External list of addresses

Ext_Address	Ext_City	Ext_State	Ext_Zip
3465 S Morgan ST	Chicago	IL	60608
1208 N Wells ST	Chicago	IL	60610

	DBAName	AKAName	Address	City	State	Zip
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	***Chicago***	IL	***60608***
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t4	***Johnnyo's***	Johnnyo's	3465 S Morgan ST	***Cicago***	IL	60608

Fan et al., 2009; Bertossi et al., 2010; Chu et al., 2015

External information

Matching dependencies

$$\text{m1: } \mathrm{Zip} = \mathrm{Ext\_Zip} \rightarrow \mathrm{City} = \mathrm{Ext\_City}$$

$$\text{m2: } \mathrm{Zip} = \mathrm{Ext\_Zip} \rightarrow \mathrm{State} = \mathrm{Ext\_State}$$

$$\text{m3: } \mathrm{City} = \mathrm{Ext\_City} \wedge \mathrm{State} = \mathrm{Ext\_State} \wedge \wedge \mathrm{Address} = \mathrm{Ext\_Address} \rightarrow \mathrm{Zip} = \mathrm{Ext\_Zip}$$

External list of addresses

Ext_Address	Ext_City	Ext_State	Ext_Zip
3465 S Morgan ST	Chicago	IL	60608
1208 N Wells ST	Chicago	IL	60610

	DBAName	AKAName	Address	City	State	Zip
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	60608
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	**60608**
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	**60608**
t4	**Johnnyo's**	Johnnyo's	3465 S Morgan ST	**Chicago**	IL	60608

Action: Map external information to input dataset using matching dependencies and repair disagreements

External information

Matching dependencies

m1: $Zip = Ext\_Zip \rightarrow City = Ext\_City$

m2: $Zip = Ext\_Zip \rightarrow State = Ext\_State$

m3: $City = Ext\_City \wedge State = Ext\_State \wedge$

$\wedge\ Address = Ext\_Address \rightarrow Zip = Ext\_Zip$

External list of addresses

Ext_Address	Ext_City	Ext_State	Ext_Zip
3465 S Morgan ST	Chicago	IL	60608
1208 N Wells ST	Chicago	IL	60610

	DBAName	AKAName	Address	City	State	Zip
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	60608
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	**60608**
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	**60608**
t4	**Johnnyo's**	Johnnyo's	3465 S Morgan ST	**Chicago**	IL	60608

External dictionaries may have limited coverage or not exist altogether

Quantitative statistics

Reason about co-occurrence of values across cells in a tuple

Estimate the distribution governing each attribute

	DBAName	AKAName	Address	City	State	Zip
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	***Chicago***	IL	***60608***
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t4	***Johnnyo's***	Johnnyo's	3465 S Morgan ST	***Cicago***	IL	60608

Example: Chicago co-occurs with IL

Hellerstein, 2008; Mayfield et al., 2010; Yakout et al., 2013

Quantitative Statistics

Reason about co-occurrence of values across cells in a tuple

Estimate the distribution governing each attribute

	DBAName	AKAName	Address	City	State	Zip
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	60608
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	60609
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	60609
t4	**John Veliotis Sr.**	Johnnyo's	3465 S Morgan ST	**Chicago**	IL	60608

Again, fails to repair the wrong zip code

Combining everything

Constraints and minimality

	DBAName	AKAName	Address	City	State	Zip
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	**60609**
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	**60609**
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	**60609**
t4	**Johnnyo's**	Johnnyo's	3465 S Morgan ST	**Cicago**	IL	60608

External data

	DBAName	AKAName	Address	City	State	Zip
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	60608
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	**60608**
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	**60608**
t4	**Johnnyo's**	Johnnyo's	3465 S Morgan ST	**Chicago**	IL	60608

Quantitative statistics

	DBAName	AKAName	Address	City	State	Zip
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	60608
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	**60609**
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	**60609**
t4	**John Veliotis Sr.**	Johnnyo's	3465 S Morgan ST	**Chicago**	IL	60608

Different solutions suggest different repairs

HoloClean: Holistic Data Repairs with Probabilistic Inference [VLDB’17]

	Address	City	State	Zip
t1	3465 S Morgan ST	*Chicago*	IL	*60608*
t2	3465 S Morgan ST	Chicago	IL	*60609*
t3	3465 S Morgan ST	Chicago	IL	*60609*
t4	3465 S Morgan ST	*Cicago*	IL	*60608*

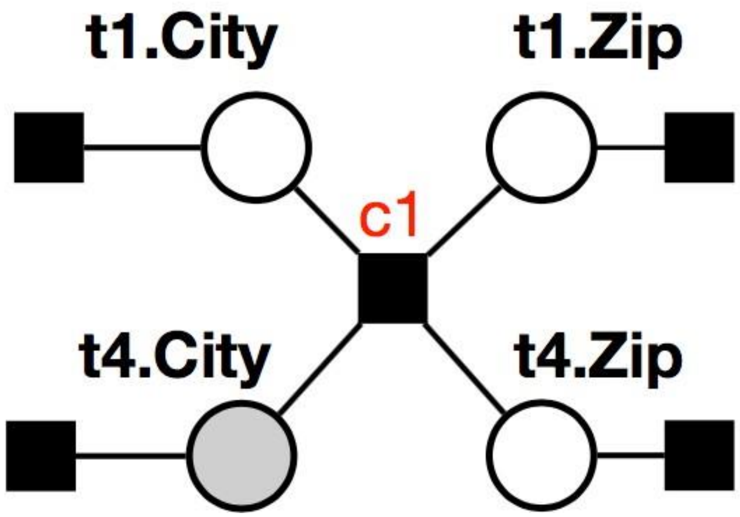
Each cell is a random variable

Value co-occurences capture data statistics

Constraints introduce correlations

$$\mathrm{c1: Zip} \rightarrow \mathrm{City}$$

“Address= 3465 S Morgan St”



: Unknown (to be inferred) RV

: Observed (fixed) RV

: Factor (encodes correlations)

Data curation and RAGs: outline

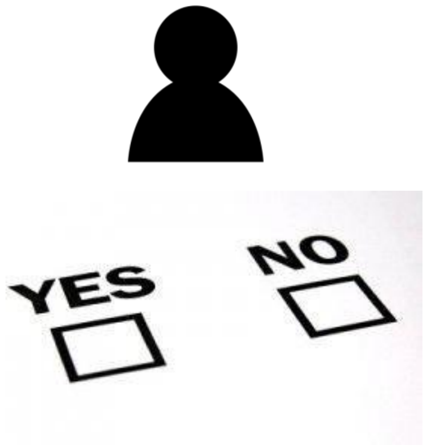
- Data curation and preparation for DB/ML
  - Data parsing
  - Data cleaning
  - Data labeling

- Retrieval Augmented Generation (RAG)

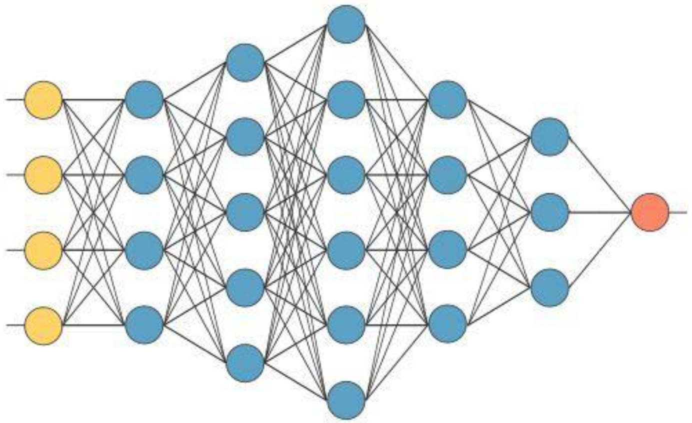
Data & labels are everything

Data Acquisition

Data Labeling



Representation Learning
and Training



A core pain point today, lots of time spent in labeling data.

Training data

- Collecting training data is **expensive** and **slow**.
- We are overfitting to our training data. [Recht et al., 2018]
  - Hand-labeled training data does not change
- Training data is the point to inject domain knowledge
  - Modern ML is too complex to hand-tune features and priors

How do we get training data (with labels) more effectively?

Weak supervision

Definition: Supervision with noisy (much easier to collect) labels; prediction on a larger set, and then training of a model.

Semi-supervised learning and ensemble learning

Examples:

- use of non-expert labelers (crowdsourcing),
- use of curated catalogs (distant supervision)
- use of heuristic rules (labeling functions)

Weak supervision

Definition: Supervision with noisy (much easier to collect) labels; prediction on a larger set, and then training of a model.

Related to semi-supervised learning and ensemble learning

Examples: use of non-expert labelers (crowdsourcing), use of curated catalogs (distant supervision), use of heuristic rules (labeling functions)

Methods developed to tackle data integration problems are closely related to weak supervision.

Learning from crowds [Raykar et al., JMLR'10]

Setup: Supervised learning but instead of gold groundtruth one has access to multiple annotators providing (possibly noisy) labels (no absolute gold standard).

Task: Learn a classifier from multiple noisy labels.

Learning from crowds [Raykar et al., JMLR'10]

Example Task: Binary classification

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$$

N examples, with labels $\mathbf{y}_i = y_i^1, \ldots, y_I^R$ provided by R different annotators

Annotator performance:

Sensitivity (true positive rate)

$$\alpha^j = \Pr[y^j = 1 | y = 1]$$

Specificity (1 - false positive rate)

$$\beta^j = \Pr[y^j = 0 | y = 0]$$

Learning from crowds [Raykar et al., JMLR'10]

$$\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$$

N examples, with labels $\mathbf{y}_i = y_i^1, \ldots, y_I^R$ provided by R different annotators

Example Task: Binary classification

Annotator performance:

Sensitivity (true positive rate)

$$\alpha^j = \Pr[y^j = 1 | y = 1]$$

Specificity (1 - false positive rate)

$$\beta^j = \Pr[y^j = 0 | y = 0]$$

Learning:

$$\Pr[\mathcal{D}|\theta] = \prod_{i=1}^{N} [a_i p_i + b_i(1 - p_i)]$$

$$\boxed{p_i := \sigma(\boldsymbol{w}^\top \boldsymbol{x}_i).}$$

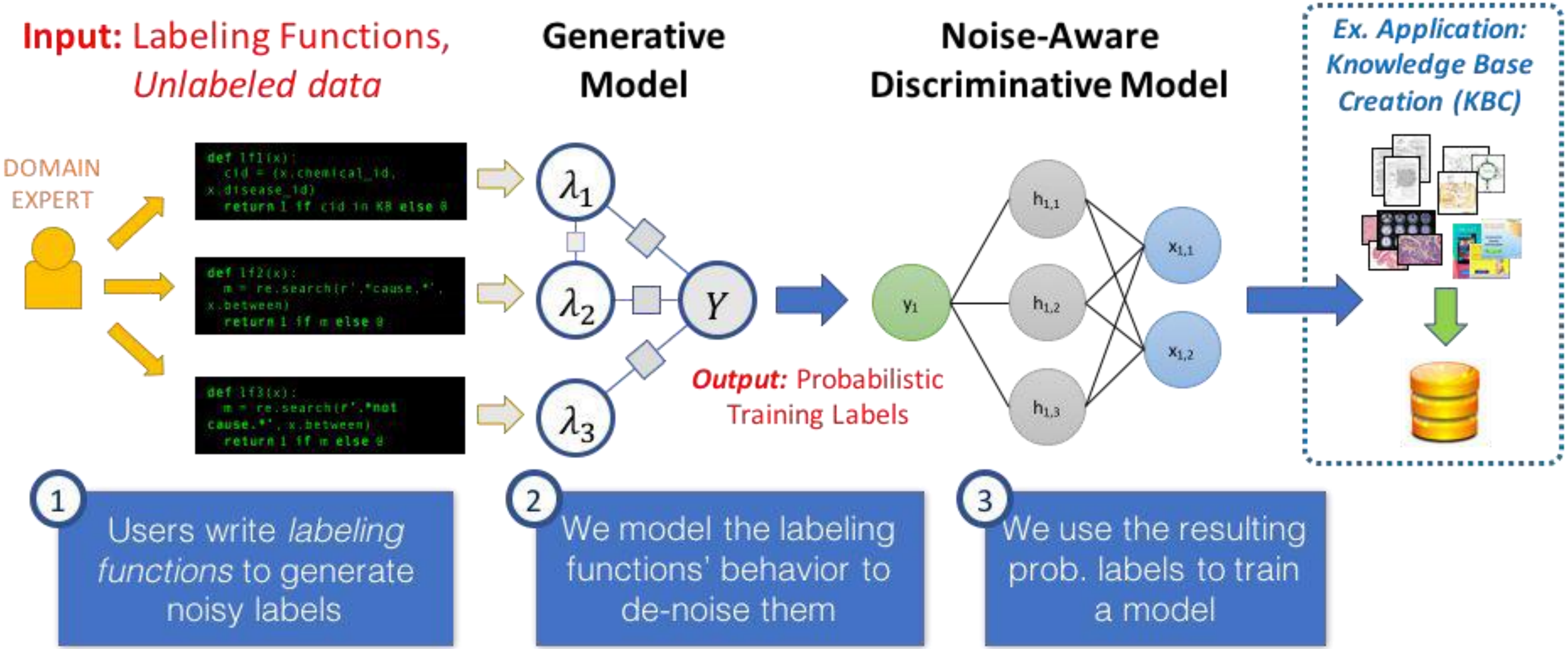
$$a_i := \prod_{j=1}^{R} [\alpha^j]^{y_i^j} [1 - \alpha^j]^{1 - y_i^j}.$$

$$b_i := \prod_{j=1}^{R} [\beta^j]^{1 - y_i^j} [1 - \beta^j]^{y_i^j}.$$

Model parameters $\{\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$

EM algorithm to obtain maximum-likelihood estimates.

Snorkel: Code as supervision [Ratner et al., NIPS'16, VLDB'18]

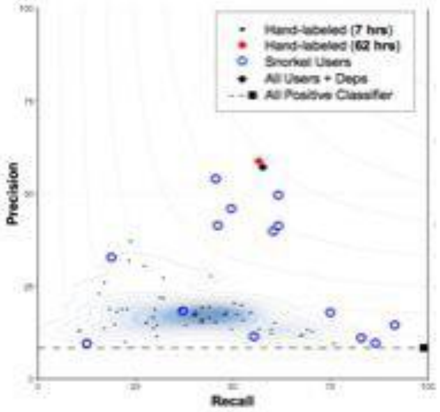


Snorkel: Code as supervision [Ratner et al., NIPS'16, VLDB'18]

Snorkel biomedical workshop in collaboration with the NIH Mobilize Center

15 companies and research groups attended

How well did these new Snorkel users do?



71% New Snorkel users matched or beat 7 hours of hand-labeling

2.8x Faster than hand-labeling data

45.5% Average improvement in model performance

3rd Place Score

No machine learning experience

Beginner-level Python

Challenges in creating training data

- Richly-formatted data is still a challenge. How can attack weak supervision when data includes images, text, tables, video, etc.?
- Combining weak supervision with other data enrichment techniques such as data augmentation is an exciting direction. How can reinforcement learning help here (http://goo.gl/K2qopQ)?
- How can we combine weak supervision with techniques from semi-supervised?

Use LLMs to label data?

- Pretrained LLMs for labelling

EmpId	ManagerId	Name	Department	Salary	City
1	1	Alex Smith	Admin	$90,000	Boulder
2	1	Amy Mars	Admin	$50,000	Longmont
3	1	Logan Mars	Admin	$70,000	Longmont
4	1	James Mont	Marketing	$55,000	
5	6	John Smith	Marketing	$60,000	Boulder
6	6	Lily Mars	Marketing	$95,000	
7	6	Ravi Grace	Database	$75,000	Longmont
8	6	Tara Frank	Database	$80,000	Longmont
9	6	Tom Ford	Database	$65,000	
10	6	William Cruze	Database	$85,000	Longmont

Approve credit?

Use LLMs to label data?

- Pretrained LLMs for labelling

EmpId	ManagerId	Name	Department	Salary	City
1	1	Alex Smith	Admin	$90,000	Boulder
2	1	Amy Mars	Admin	$50,000	Longmont
3	1	Logan Mars	Admin	$70,000	Longmont
4	1	James Mont	Marketing	$55,000	
5	6	John Smith	Marketing	$60,000	Boulder
6	6	Lily Mars	Marketing	$95,000	
7	6	Ravi Grace	Database	$75,000	Longmont
8	6	Tara Frank	Database	$80,000	Longmont
9	6	Tom Ford	Database	$65,000	
10	6	William Cruze	Database	$85,000	Longmont

- Similarly, apply pretraind LLMs in NL, image, video data
- Could be a good idea, but too expensive, and may not work with domain knowledge. Also, chicken-and-egg problem in how to get the initial model.

Use LLMs to label data?

- Pretrained LLMs for labelling

EmpId	ManagerId	Name	Department	Salary	City
1	1	Alex Smith	Admin	$90,000	Boulder
2	1	Amy Mars	Admin	$50,000	Longmont
3	1	Logan Mars	Admin	$70,000	Longmont
4	1	James Mont	Marketing	$55,000	
5	6	John Smith	Marketing	$60,000	Boulder
6	6	Lily Mars	Marketing	$95,000	
7	6	Ravi Grace	Database	$75,000	Longmont
8	6	Tara Frank	Database	$80,000	Longmont
9	6	Tom Ford	Database	$65,000	
10	6	William Cruze	Database	$85,000	Longmont

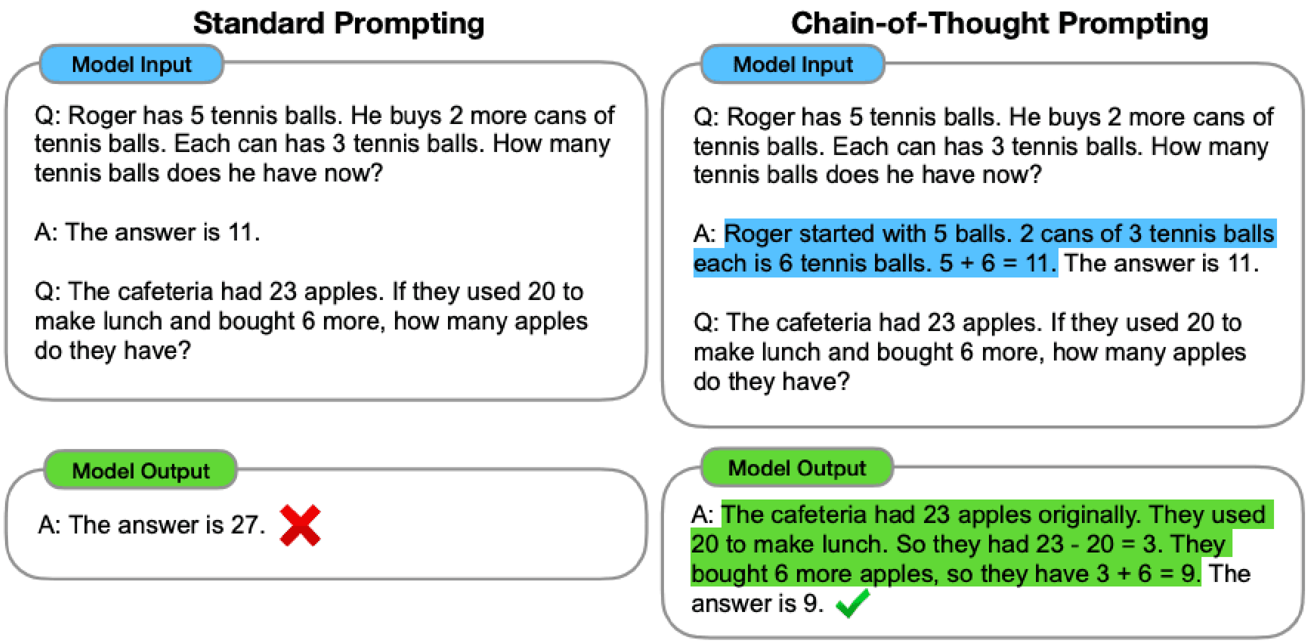
Approve credit?

- Similarly, apply pretraind LLMs in NL, image, video data
- Could be a good idea, but too expensive, and may not work with domain knowledge
  Also, chicken-and-egg problem in how to get the initial model.
  - Use distilled, fine-tuned model
  - Reorder columns to maximize KV cache reuse

OPTIMIZING LLM QUERIES IN RELATIONAL DATA ANALYTICS WORKLOADS. MLSys25'.

Obtaining labelled language data

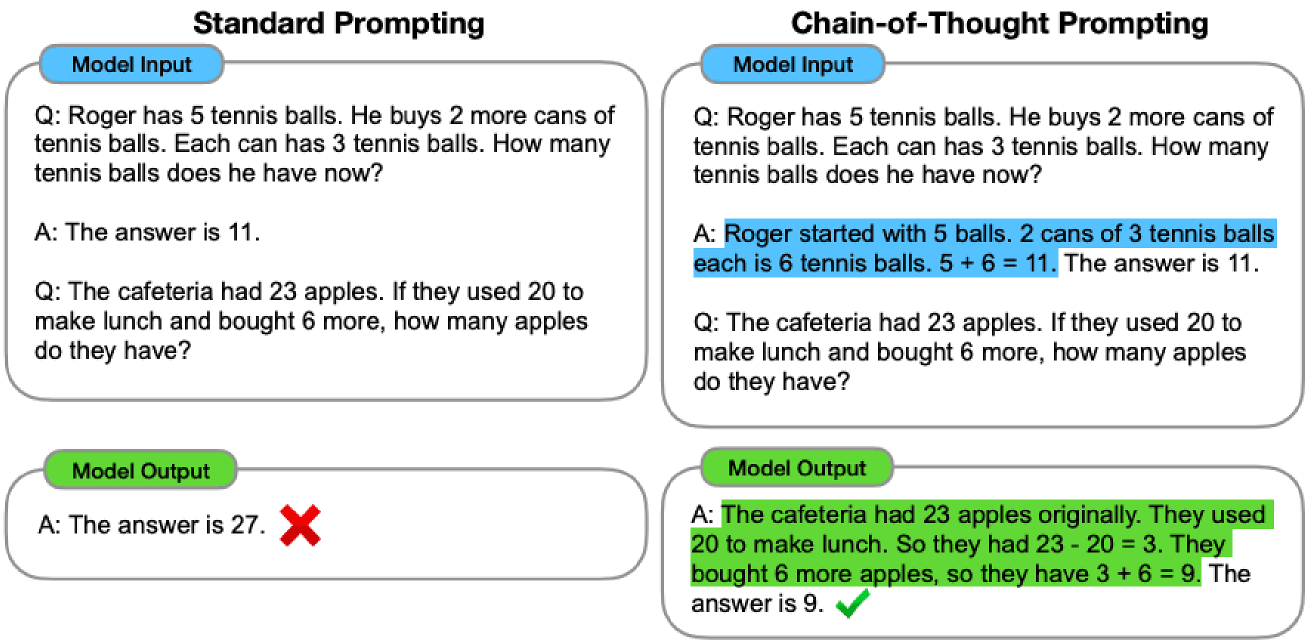
- Pretrained LLMs that generate NL labels
  - Chain-of-thought, or “deep-think” prompting



- Use OpenAI GPT-o1 or DeepSeek-R1

Obtaining labelled language data

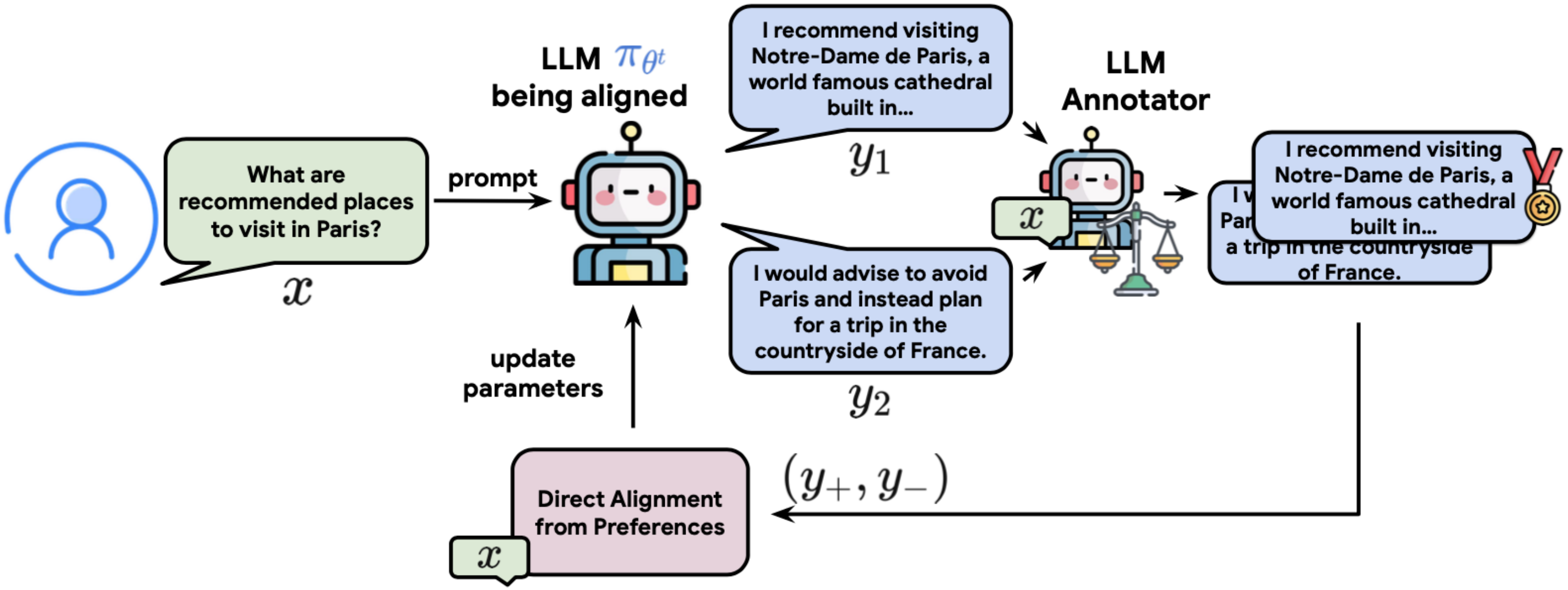
- Pretrained LLMs that generate NL labels
  - Chain-of-thought, or “deep-think” prompting



- Use OpenAI GPT-o1 or DeepSeek-R1
  - But LLMs are good at bluffing (hallucinations). How to verify results?

Obtaining labelled language data

- Verifying LLM generations
  - Use human experts (RLHF) > too costly
  - Use other LLM(s) or AI agents? Voting, debating, etc.



Direct Language Model Alignment from Online AI Feedback [arXiv 2024]

Data curation and RAGs: outline

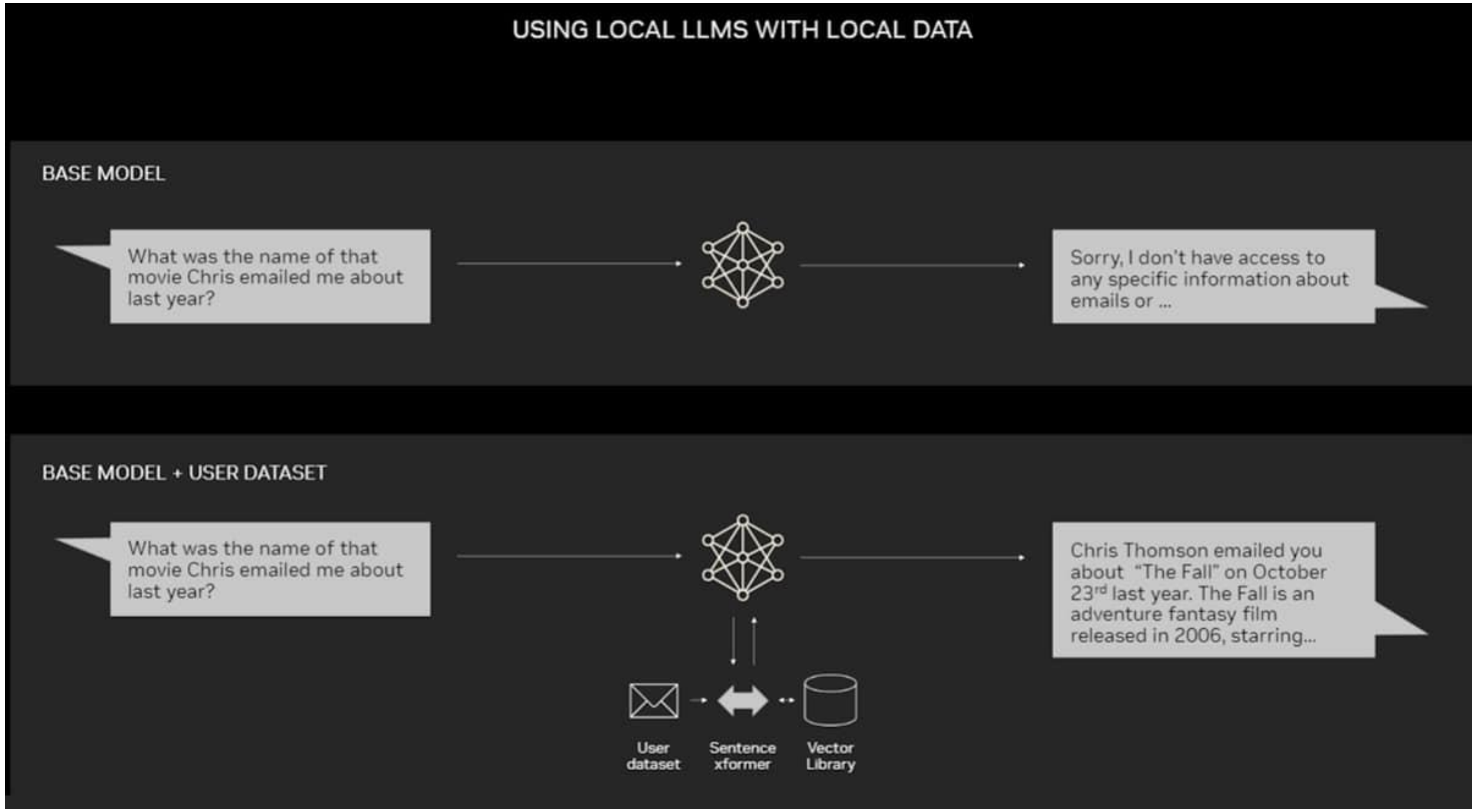
- Data curation and preparation for DB/ML
  - Parsing
  - Cleaning
  - Labeling

- Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG)

Directly using LLMs faces problems

- Information lag
- Model hallucination
- Hard to incorporate proprietary data



Retrieval Augmented Generation (RAG)

LangChain

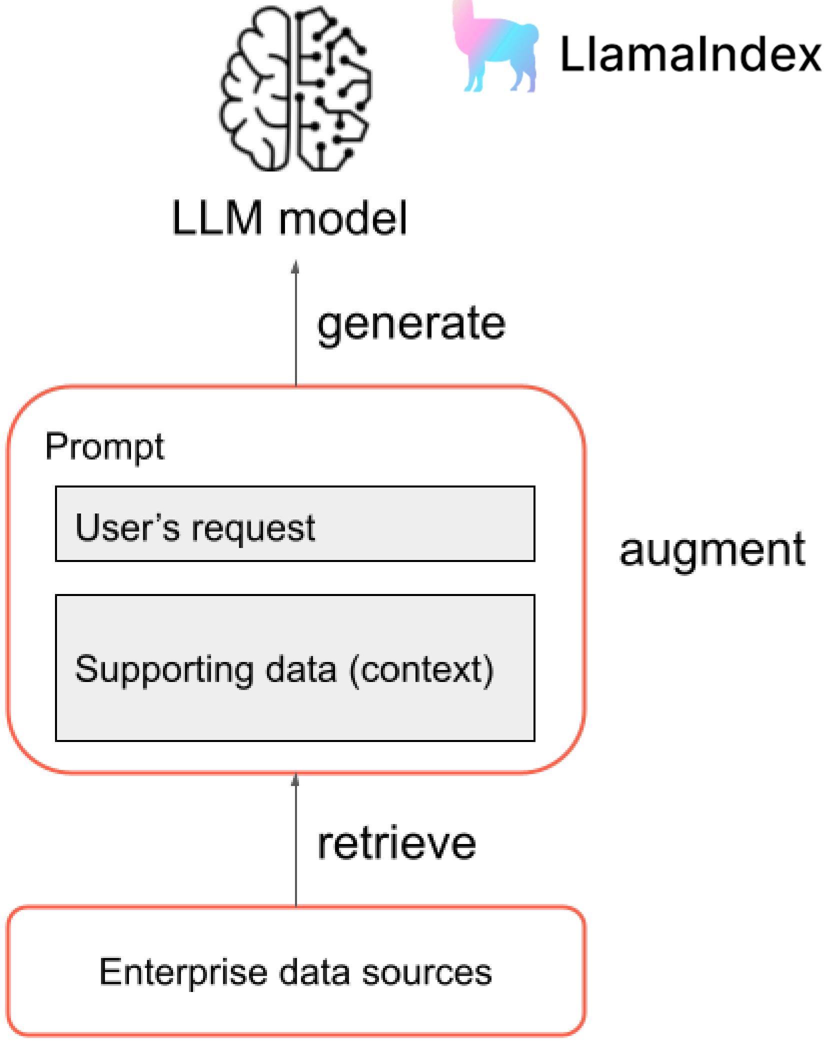
LlamaIndex

Directly using LLMs faces problems

- Information lag
- Model hallucination
- Hard to incorporate proprietary data

Instead, we need RAG =

- **Retrieval**: The user's request is used to query some external info - querying a vector store, a keyword search over text, or querying a database. This is to obtain supporting data / context that helps the LLM provide a useful response.
- **Augmentation**: The supporting data / context is combined with the user request, often using a template with instructions to the LLM, to create a prompt.
- **Generation**: The LLM generates a response to the prompt.



With an LLM alone	Using LLMs with RAG
No proprietary knowledge: LLMs are generally trained on publicly available data, so they cannot accurately answer questions about a company's internal or proprietary data.	**RAG applications can incorporate proprietary data:** A RAG application can supply proprietary documents such as memos, emails, and design documents to an LLM, enabling it to answer questions about those documents.
Knowledge isn't updated in real time: LLMs do not have access to information about events that occurred after they were trained. For example, a standalone LLM cannot tell you anything about stock movements today.	**RAG applications can access real-time data:** A RAG application can supply the LLM with timely information from an updated data source, allowing it to provide useful answers about events past its training cutoff date.
Lack of citations: LLMs cannot cite specific sources of information when responding, leaving the user unable to verify whether the response is factually correct or a hallucination.	**RAG can cite sources:** When used as part of a RAG application, an LLM can be asked to cite its sources.
Lack of data access controls (ACLs): LLMs alone can't reliably provide different answers to different users based on specific user permissions.	**RAG allows for data security/ACLs:** The retrieval step can be designed to find only the information that the user has credentials to access, enabling a RAG application to selectively retrieve personal or proprietary information.

RAG workflow

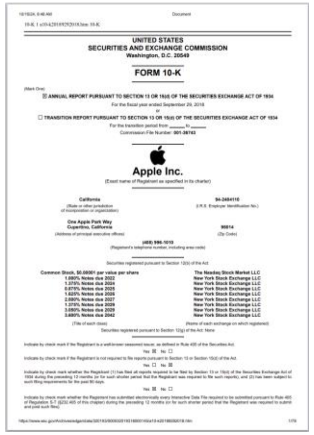
(Offline) Preprocess

- Chunking documents with simple heuristics (1)
- Compute embeddings w/ a pre-trained model (2)
- Indexing & store the embeddings in a database (2)

(Online) When a user query comes

- Compute embedding for the user query (3)
- Retrieve relevant embeddings from the database (4)
- Assemble a prompt, send it to LLM for result (5-7)

Example: Ask "How many employees?" to an SEC filing


UNITED STATES
SECURITIES AND EXCHANGE COMMISSION
Washington, D.C. 20549
FORM 10-K
Apple Inc.

~100 pages, tables, text

"Retrieved" context from the document:

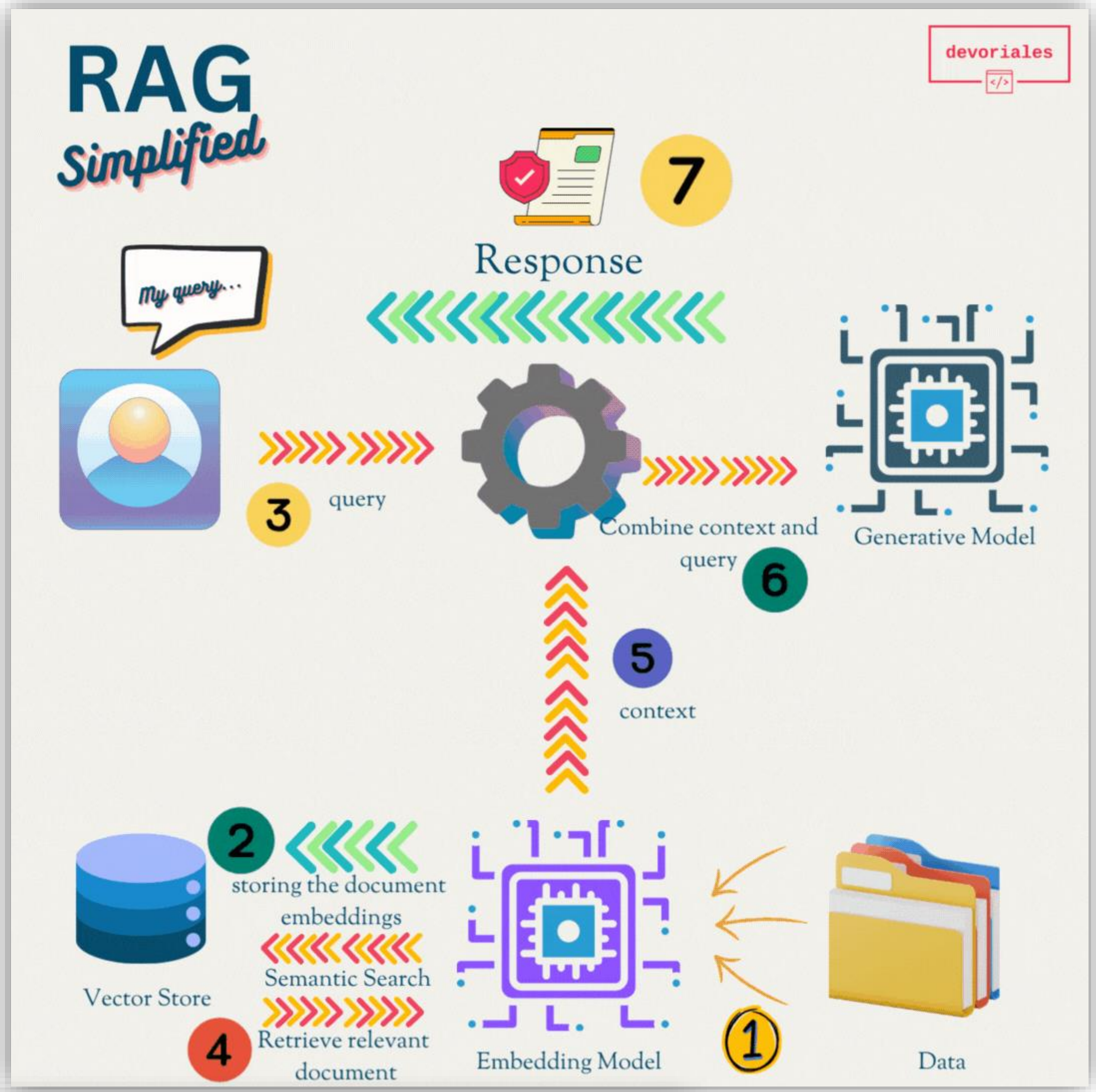
Backlog

In the Company's experience, the actual amount of product backlog at any particular time is not a meaningful indication of its future business prospects. In particular, backlog often increases immediately following new product introductions as customers anticipate shortages. Backlog is often reduced once customers believe they can obtain sufficient supply. Because of the foregoing, backlog should not be considered a reliable indicator of the Company's ability to achieve any particular level of revenue or financial performance.

Employees

As of September 29, 2018, the Company had approximately 132,000 full-time equivalent employees.

Apple Inc. | 2018 Form 10-K | 6



Credits: devoriales.com

Drawbacks of RAG

- **What if retrieval goes wrong?**
  - Raw documents are highly nonstructured
  - Documents are too long
  - Complex retrieval
  - Ranking is wrong

- **What if generation goes wrong?**
  - Prompt is too complex / long
  - Generation doesn't follow instruction / format requirement

Note 3 – Financial Instruments

Cash, Cash Equivalents and Marketable Securities

The following tables show the Company's cash, cash equivalents and marketable securities by significant investment category as of December 31, 2022 and September 24, 2022 (in millions):

	December 31, 2022						
	Adjusted Cost	Unrealized Gains	Unrealized Losses	Fair Value	Cash and Cash Equivalents	Current Marketable Securities	Non-Current Marketable Securities
Cash	$ 17,908	$ —	$ —	$ 17,908	$ 17,908	$ —	$ —
Level 1 (1):							
Money market funds	818	—	—	818	818	—	—
Mutual funds	330	2	(40)	292	—	292	—
Subtotal	1,148	2	(40)	1,110	818	292	—
Level 2 (2):							
U.S. Treasury securities	24,128	1	(1,576)	22,553	13	9,105	13,435
U.S. agency securities	5,743	—	(643)	5,100	—	310	4,790
Non-U.S. government securities	17,778	14	(1,029)	16,763	—	9,907	6,856
Certificates of deposit and time deposits	2,025	—	—	2,025	1,795	230	—
Commercial paper	237	—	—	237	—	237	—
Corporate debt securities	85,895	14	(7,039)	78,870	1	10,377	68,492
Municipal securities	864	—	(26)	838	—	278	560
Mortgage- and asset-backed securities	22,448	3	(2,405)	20,046	—	84	19,962
Subtotal	159,118	32	(12,718)	146,432	1,809	30,528	114,095
Total (3)	$ 178,174	$ 34	$ (12,758)	$ 165,450	$ 20,535	$ 30,820	$ 114,095

	September 24, 2022						
	Adjusted Cost	Unrealized Gains	Unrealized Losses	Fair Value	Cash and Cash Equivalents	Current Marketable Securities	Non-Current Marketable Securities
Cash	$ 18,546	$ —	$ —	$ 18,546	$ 18,546	$ —	$ —
Level 1 (1):							
Money market funds	2,929	—	—	2,929	2,929	—	—
Mutual funds	274	—	(47)	227	—	227	—
Subtotal	3,203	—	(47)	3,156	2,929	227	—
Level 2 (2):							
U.S. Treasury securities	25,134	—	(1,725)	23,409	338	5,091	17,980
U.S. agency securities	5,823	—	(655)	5,168	—	240	4,928
Non-U.S. government securities	16,948	2	(1,201)	15,749	—	8,806	6,943
Certificates of deposit and time deposits	2,067	—	—	2,067	1,805	262	—
Commercial paper	718	—	—	718	28	690	—
Corporate debt securities	87,148	9	(7,707)	79,450	—	9,023	70,427
Municipal securities	921	—	(35)	886	—	266	620
Mortgage- and asset-backed securities	22,553	—	(2,593)	19,960	—	53	19,907
Subtotal	161,312	11	(13,916)	147,407	2,171	24,431	120,805
Total (3)	$ 183,061	$ 11	$ (13,963)	$ 169,109	$ 23,646	$ 24,658	$ 120,805

(1) Level 1 fair value estimates are based on quoted prices in active markets for identical assets or liabilities.

(2) Level 2 fair value estimates are based on observable inputs other than quoted prices in active markets for identical assets and liabilities, quoted prices for identical or similar assets or liabilities in inactive markets, or other inputs that are observable or can be corroborated by observable market data for substantially the full term of the assets or liabilities.

(3) As of December 31, 2022 and September 24, 2022, total marketable securities included $13.6 billion and $12.7 billion, respectively, that were restricted from general use, related to the European Commission decision finding that Ireland granted state aid to the Company, and other agreements.

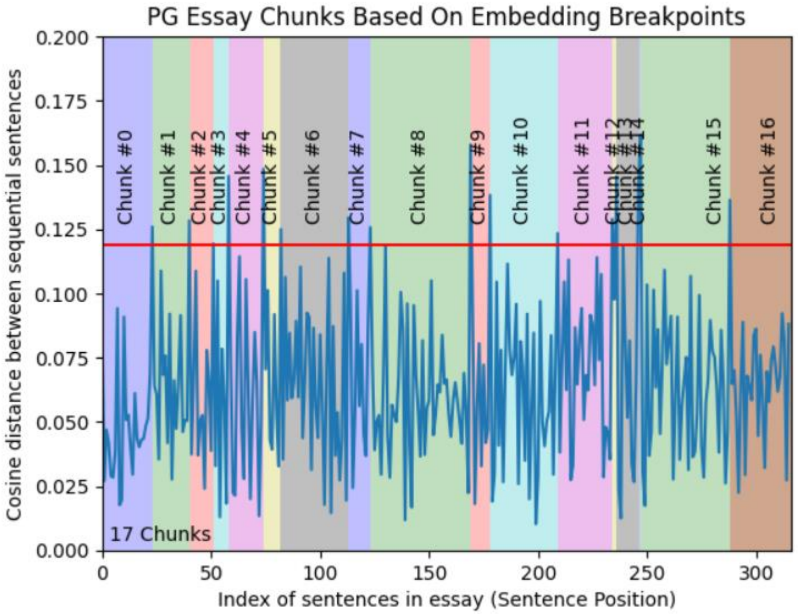
Looking back on the info retrieval literature

Many IR techniques can be applied to RAG

- Better chunking mechanisms
- Prompt compression
- Learning to rank / re-ranking
- Model selection, finetuning & distillation
- Multi-way retrieval
- Graph RAG
- Combine with full-text search

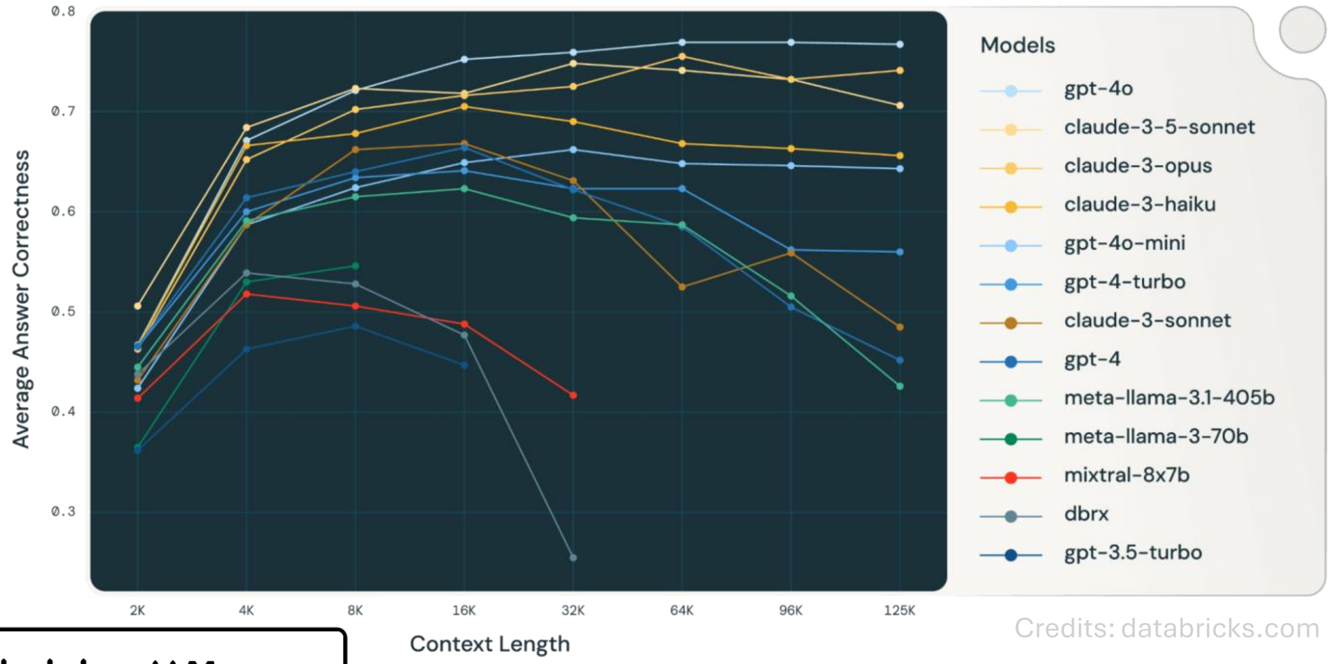
Better chunking mechanisms

- Besides the simple fix-length chunking, there are many other ways:
  - **Overlapping windows** to make sure information is captured in some windows
  - **Structure-aware chunking** to avoid breaking in the middle of paragraphs and sentences
  - **Document based chunking** that leverages the document property (Markdown, HTML, LaTeX etc.)
  - **NLP/Semantic chunking** to detect topic changes
  - **Agentic chucking** uses AI agents to decide if a sentence should be added to the previous chunk.

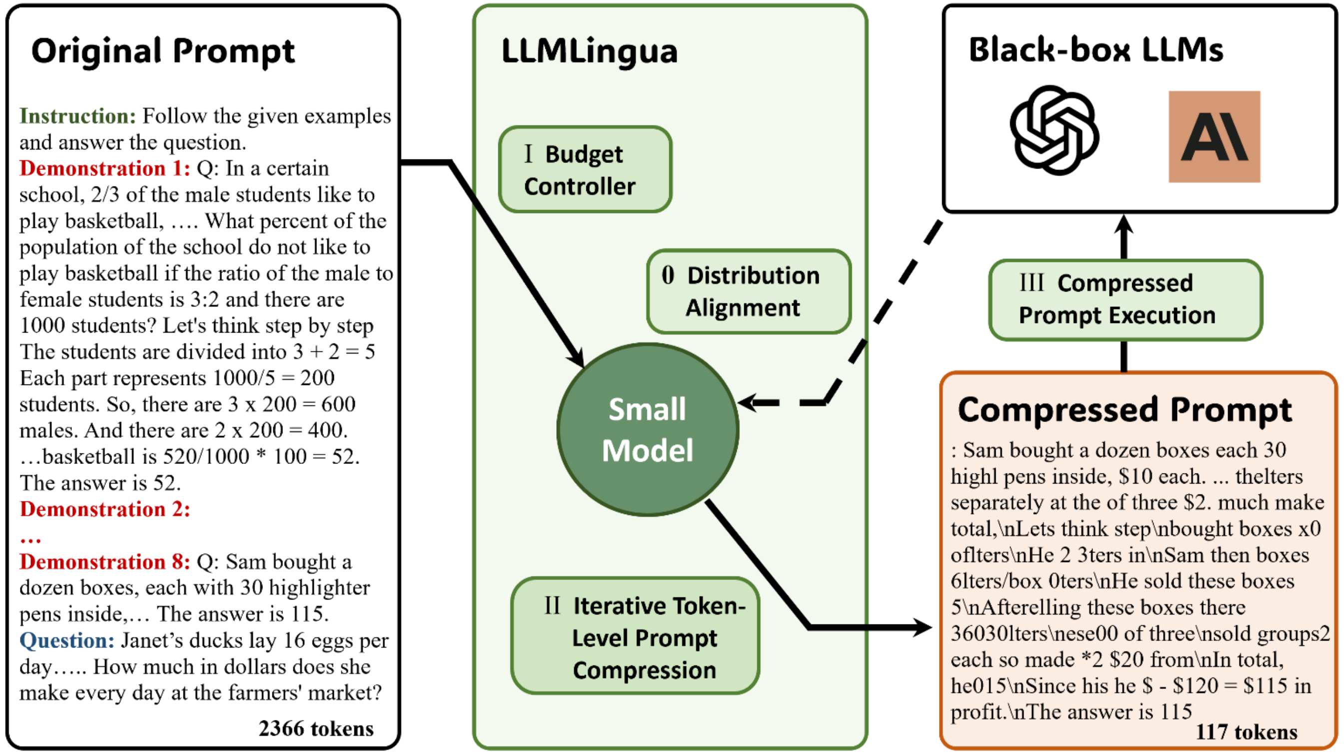


Prompt compression

- More context = more accurate (at cost)
- LLMLingua EMNLP 2023 (Instruction tuning!)



Credits: databricks.com



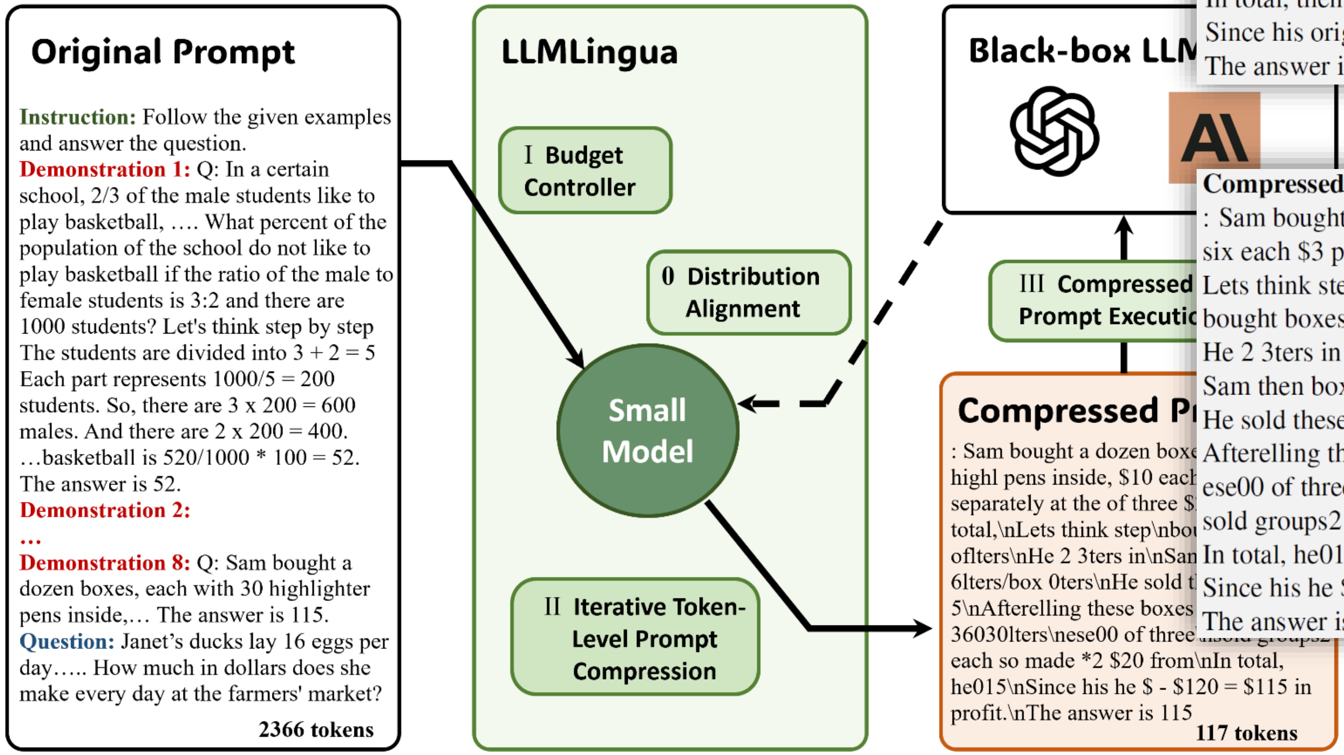
Prompt compression

- More context = more accurate (at cost)
- LLMLingua EMNLP 2023 (Instruction tuning!)

Original Prompt(9-steps Chain-of-Thought):
Question: Sam bought a dozen boxes, each with 30 highlighter pens inside, for $10 each box. He rearranged five of these boxes into packages of six highlighters each and sold them for $3 per package. He sold the rest of the highlighters separately at the rate of three pens for $2. How much profit did he make in total, in dollars?

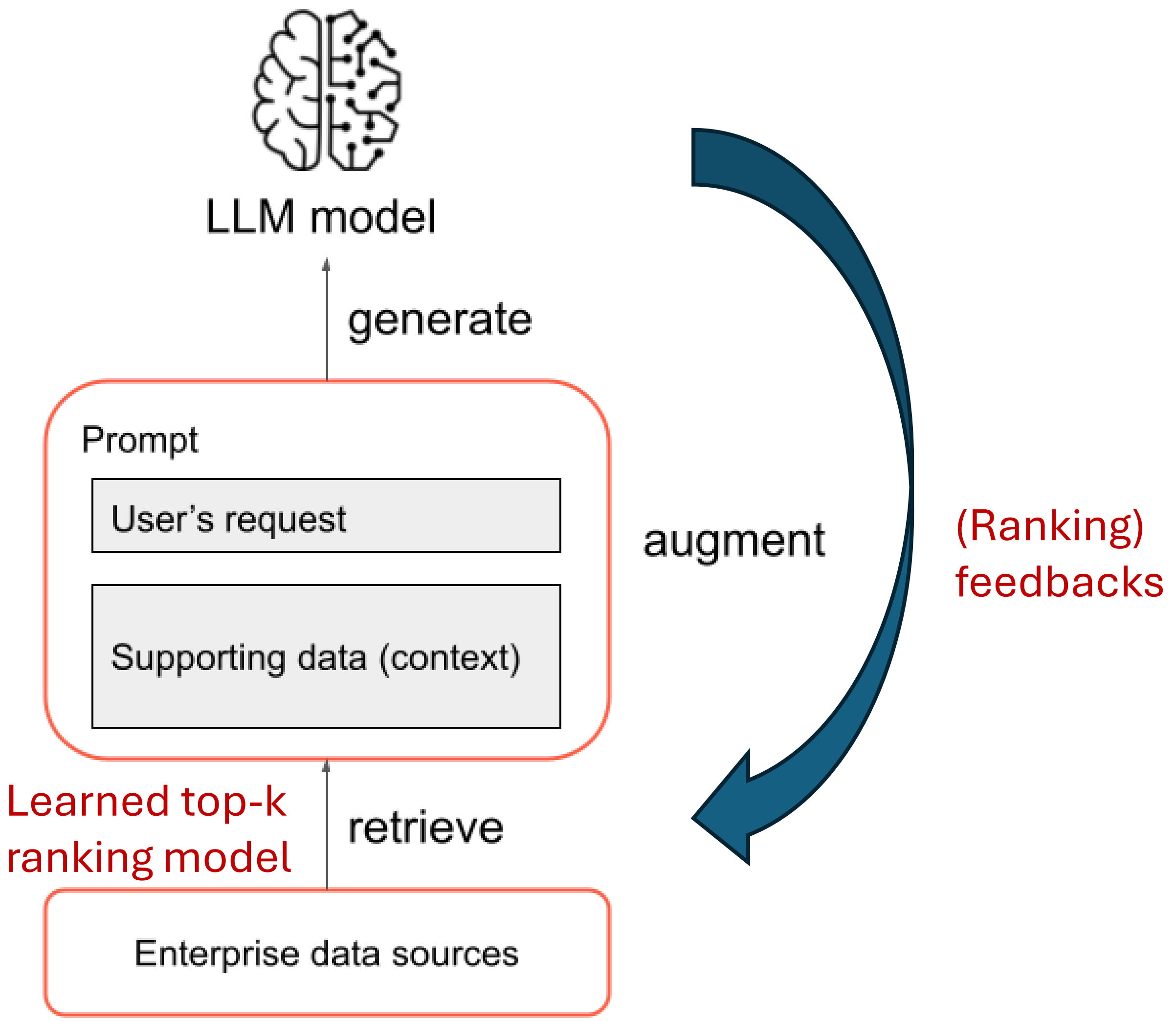
Let's think step by step
Sam bought 12 boxes x $10 = $120 worth of highlighters.
He bought 12 * 30 = 360 highlighters in total.
Sam then took 5 boxes × 6 highlighters/box = 30 highlighters.
He sold these boxes for 5 * $3 = $15
After selling these 5 boxes there were 360 - 30 = 330 highlighters remaining.
These form 330 / 3 = 110 groups of three pens.
He sold each of these groups for $2 each, so made 110 * 2 = $220 from them.
In total, then, he earned $220 + $15 = $235.
Since his original cost was $120, he earned $235 - $120 = $115 in profit.
The answer is 115

Compressed Prompt:
: Sam bought a dozen boxes each 30 highl pens inside, $10 each. He reanged five of boxes into of six each $3 per. He sold the thelters separately at the of three $2. much make total,
Lets think step
bought boxes x0 oflters
He 2 3ters in
Sam then boxes 6lters/box 0ters
He sold these boxes 5
Afterelling these boxes there 36030lters
ese00 of three
sold groups2 each so made *2 $20 from
In total, he015
Since his he $ - $120 = $115 in profit.
The answer is 115



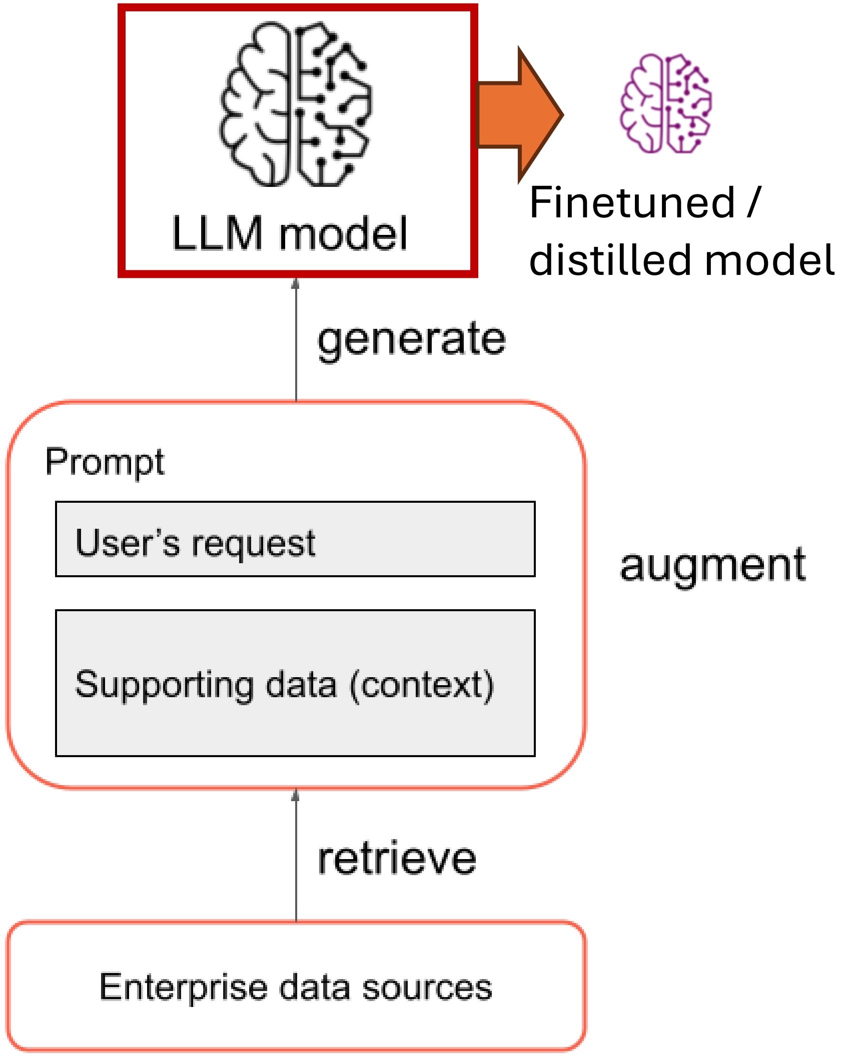
Learning to rank / re-ranking

- The “retrieval” part can be improved by using a learned top-k ranking model (should be cheaper than the later LLM)
- Automatic and free labels from previous runs
- Reduces context length requirements (improve P@K)

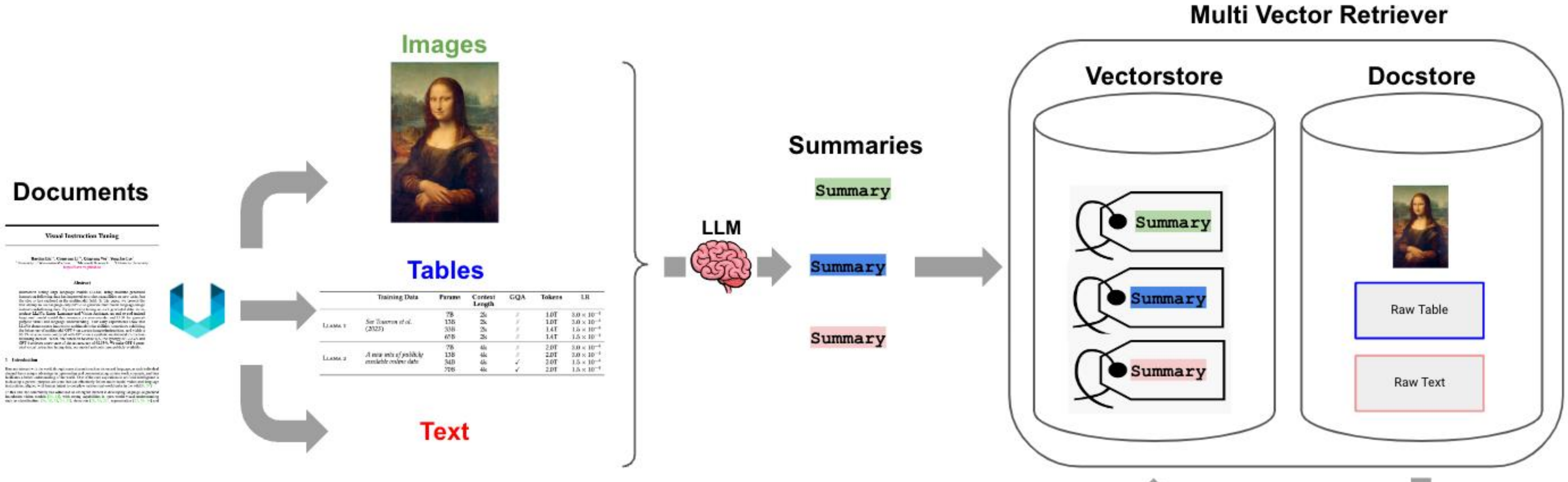


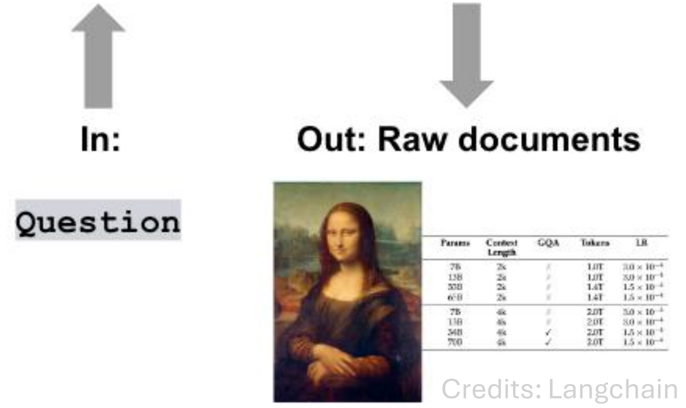
Model selection, finetuning & distillation

- Finetune or distill the generation model in order to reduce size, adapt to formatting requirements. e.g., collect RAG outputs from Llama 70b and send them to finetune Llama 13b
- Or for different queries, use different generation models
- Further, we can propagate the gradients to the embedding phrase, **and finetune embedding models**



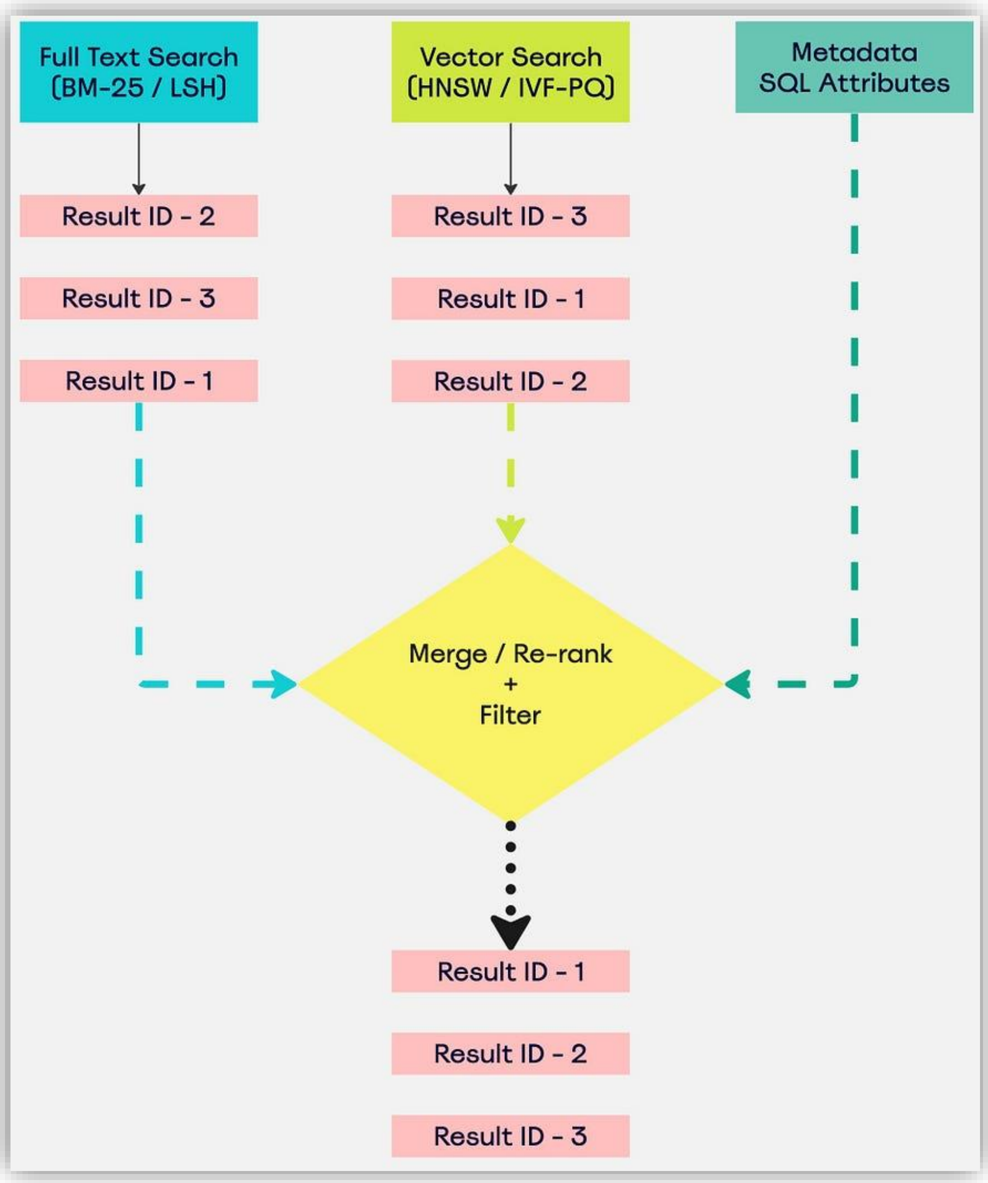
Multi-vector retrieval





- Classic RAG falls short for complex, multi-modal datasets
- Use different embedding models for inputs of different modality

Combine with full-text search



- Embedding has “needle-in-the-hay” problem.
- To improve, RAGs can be combined with full-text search or external tools (SQL, search engine) to boost accuracy
- Full-text search: BM-25 or LSH.

Data curation and RAGs

- Data curation and preparation for DB/ML
  - Data parsing
  - Data cleaning
  - Data labeling

- Retrieval Augmented Generation (RAG)

Credits

- Luna Xin Dong, Meta
- Kexin Rong, Galtech