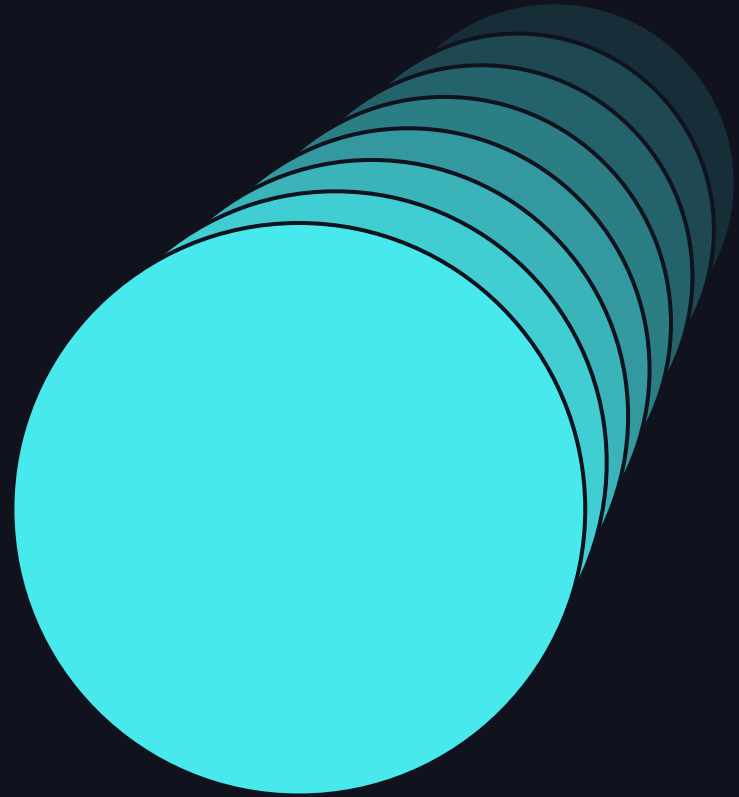


# Best Practices for Data Prep for GenAI Development

---

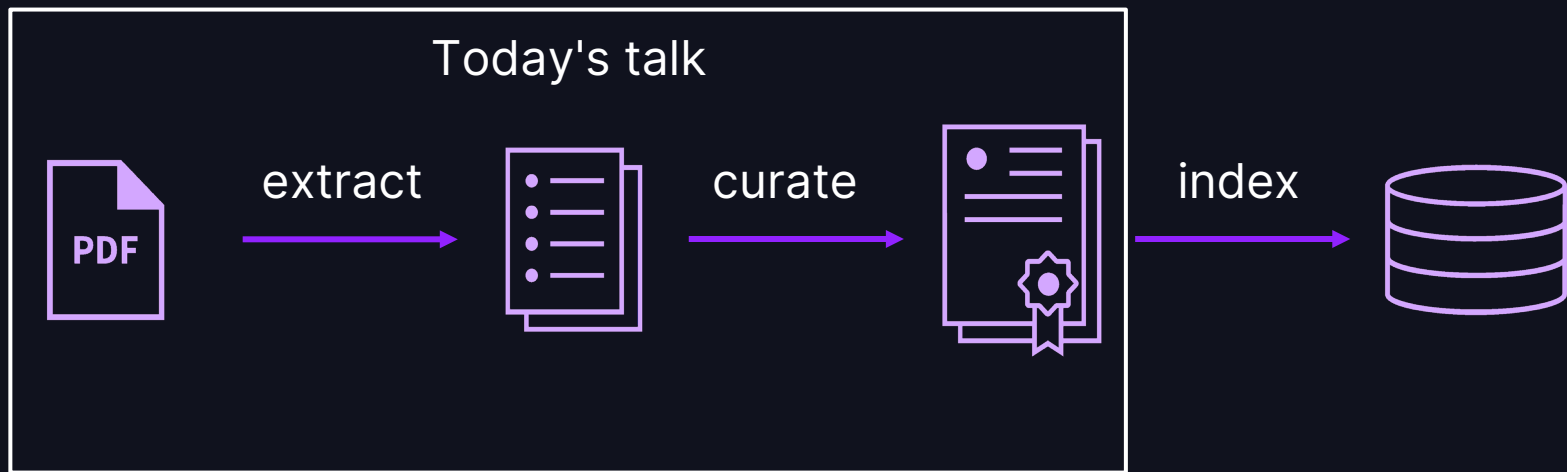
Brian Kihoon Lee



# Lilac x Databricks

- Lilac joined Databricks in March 2024
- We built LLM-superpowered tools to understand and curate text data

# A typical RAG data pipeline



# LIVE DEMO: EXTRACTION



# Extract data from PDF

## PYTHON

```
def parsed1(filepath):
    try:
        doc = pymupdf.open(filepath) # open a document
        output = '\n'.join(page.get_text() for page in doc)
    except Exception as e:
        print(e, filepath)
        output = ''
    return {'parsed': output}

dataset.map(
    parsed1, input_path='filepath', output_path='parser1', num_jobs=8, execution_type='processes', overwrite=True
)
```

# Look at your data

The screenshot shows a Databricks notebook interface. At the top, there is a navigation bar with a left arrow, the text "2 of 799", and a right arrow. On the far right of the navigation bar is a trash icon. Below the navigation bar, the notebook name "parser1" is visible. The main content area shows a table with the following text:

|    | parsed  |
|----|---|
| 1  | 383.235 Verdict -- Procedure in case of disagreement.   |
| 2  | The jurors, after hearing the evidence, shall, by their inquest, say whether the defendants,  |
| 3  | or either of them, be guilty or not guilty of the forcible entry or detainer complained of;   |
| 4  | and shall return their inquest, signed by one of their body, to the court. If the jury do not |
| 5  | agree, it may be discharged, and another be ordered to be summoned to meet, either            |
| 6  | immediately or at some future day to be then and there fixed and indorsed on the warrant;     |
| 7  | and this proceeding shall be continued until a jury agree.                                    |
| 8  | Effective: January 2, 1978  |
| 9  | History: Amended 1976 (1st Extra. Sess.) Ky. Acts ch. 14, sec. 315, effective January         |
| 10 | 2, 1978. -- Transferred 1952 Ky. Acts ch. 84, sec. 1, effective July 1, 1953, from            |
| 11 | C.C. sec. 459.  |
| 12 |   |

On the right side of the table, there are several icons: a magnifying glass (search), a downward arrow (dropdown), a fork (share), a list icon, and a document icon. A small preview of the table's content is visible next to the document icon.



# Extract data from PDF, attempt #2

PYTHON

```
def parsed2(filepath):
    try:
        doc = pymupdf.open(filepath) # open a document
        text_builder = []
        for page in doc:
            blocks = page.get_text('blocks')
            text_builder.append('\n\n'.join(block[4].replace('\n', ' ') for block in blocks))
        output = ''.join(text_builder)
    except Exception as e:
        print(e, filepath)
        output = ''
    return {'parsed': output}

dataset.map(parsed2, input_path='filepath', output_path='parser2', num_jobs=8, execution_type='processes')
```

# Don't overfit to one document

parser1

parser1 > parsed

```
1 383.235 Verdict -- Procedure in case of
disagreement.
2 The jurors, after hearing the evidence, shall, by
their inquest, say whether the defendants,
3 or either of them, be guilty or not guilty of the
forcible entry or detainer complained of;
4 and shall return their inquest, signed by one of their
body, to the court. If the jury do not
5 agree, it may be discharged, and another be ordered to
be summoned to meet, either
6 immediately or at some future day to be then and there
fixed and indorsed on the warrant;
7 and this proceeding shall be continued until a jury
agree.
8 Effective: January 2, 1978
9 History: Amended 1976 (1st Extra. Sess.) Ky. Acts ch.
14, sec. 315, effective January
10 2, 1978. -- Transferred 1952 Ky. Acts ch. 84, sec. 1,
effective July 1, 1953, from
11 C.C. sec. 459.
12
```

parser2 > parsed

↶ ↷

```
1 383.235 Verdict -- Procedure in case of
disagreement.
2 The jurors, after hearing the evidence, shall, by
their inquest, say whether the defendants, or either
of them, be guilty or not guilty of the forcible entry
or detainer complained of; and shall return their
inquest, signed by one of their body, to the court. If
the jury do not agree, it may be discharged, and
another be ordered to be summoned to meet, either
immediately or at some future day to be then and there
fixed and indorsed on the warrant; and this proceeding
shall be continued until a jury agree.
3 Effective: January 2, 1978
4 History: Amended 1976 (1st Extra. Sess.) Ky. Acts ch.
14, sec. 315, effective January
5 2, 1978. -- Transferred 1952 Ky. Acts ch. 84, sec. 1,
effective July 1, 1953, from C.C. sec. 459.
```



# Normalize whitespace and unicode

parser1

parser1 > parsed

```
1 Larson Statement on Ryan Budget | Congressman John
  Larson
2 Tuesday, 12 March 2013
3
4 (Washington) Today Congressman John B. Larson
  released the following statement on the
5 proposed budget by House Republicans:
6
7 "Paul Ryan, in a clear effort to appease the radical
  tea party members of the Republican
8 caucus, has put forth a budget even more austere than
  his last proposal to end Medicare as we
9 know it. Today Ryan refused to deal with the problems
  our nation is currently facing, and
10 proposed a budget that would end the Medicare
  guarantee, end protections for those with
```

parser2 > parsed

The character U+2013 " – " could be confused with the ASCII character U+002d " - ", which is more common in source code. [Adjust settings](#)

```
1 Larson Sta
  LarsonTues
2
3 (Washington) Today Congressman John B. Larson
  released the following statement on theproposed budget
  by House Republicans:
4
5 "Paul Ryan, in a clear effort to appease the radical
  tea party members of the Republicancaucus, has put
  forth a budget even more austere than his last
  proposal to end Medicare as weknow it. Today Ryan
  refused to deal with the problems our nation is
  currently facing, andproposed a budget that would end
  the Medicare guarantee, end protections for those
  withpre-existing conditions, and prevent children from
```

# Clean up data from PDF

## PYTHON

```
from unstructured.cleaners.core import clean_bullets, replace_unicode_quotes
import re

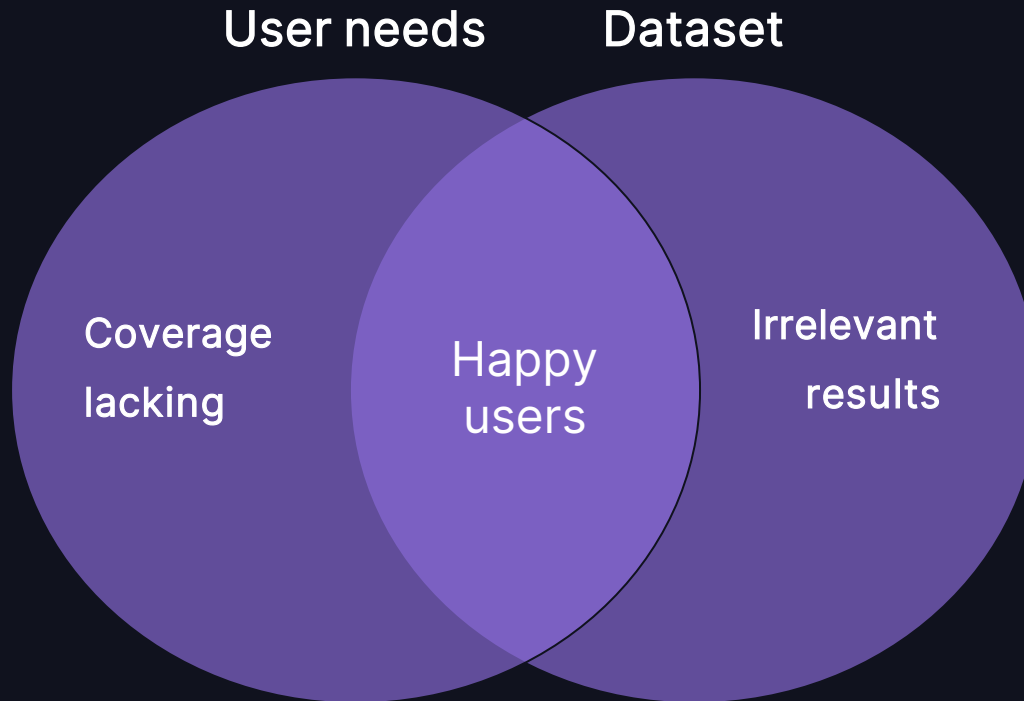
def clean_text(text):
    text = replace_unicode_quotes(text) # Homogenize quotation marks
    text = clean_bullets(text) # Homogenize bullet points
    text = re.sub(r'[\u00ad]', '', text) # Soft hyphen removal
    text = re.sub(r'[\u2013\u2212]', '-', text) # Uncommon hyphen variants
    text = re.sub(r'[ ]{2,}', ' ', text) # Normalize spacing
    text = re.sub(r'[\u00a0]', ' ', text) # Remove invisible whitespace
    return text

dataset.map(clean_text, input_path='parser2.parsed', output_path='parser2.cleaned', overwrite=True)
```

# Extraction: summary

1. Look at your data!
2. Don't overfit your pipeline to a few documents
3. Normalize whitespace and uncommon characters

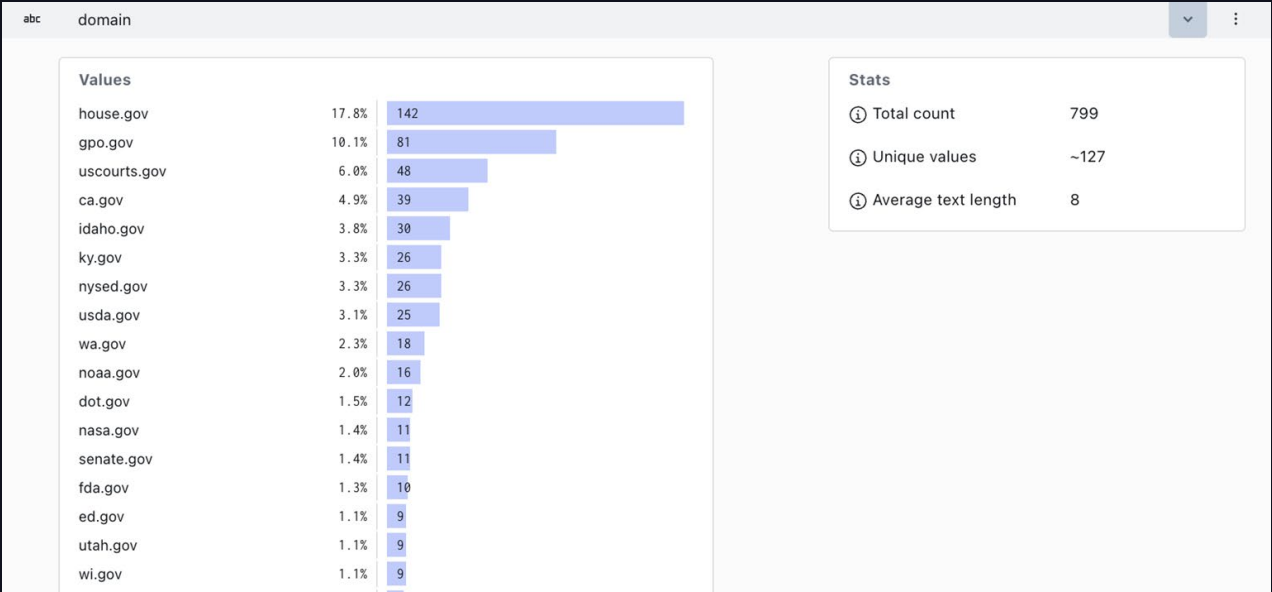
# What is curation?



# LIVE DEMO: CURATION



# Start with structured columns



# Clustering creates structure

## Regulations & Policies

22.65%

181 rows

[Explore ↗](#)

Environmental Regulations and Studies

14.36%  
26 rows



Government Contracts and Regulations

11.05%  
20 rows



Regulatory Notices and Proposals

10.50%  
19 rows



Natural Gas Piping & Safety Regulations

7.73%  
14 rows



1 of 4



## Government and Politics

21.65%

173 rows

[Explore ↗](#)

Congressional Office Services and Events

17.34%  
30 rows



Crime Victimitizations and Time Lost from Work

13.87%  
24 rows



Government Meetings Schedules and Minutes

13.29%  
23 rows



Healthcare Reform Discussions

10.40%  
18 rows



1 of 4



# Curation: summary

1. Every dataset and use case is different: you are the domain expert
2. Use histograms and text clusters to check coverage and relevancy
3. Cross-reference usage logs with your indexed data



# Takeaways

Tools used in this talk:

PDF extraction: PyMuPDF, unstructured

Text cleaning: spacy, unstructured, regex

Visualization: Lilac

## Extraction

1. Look at your data!
2. Don't overfit your pipeline to a few documents
3. Normalize whitespace and uncommon characters

## Curation

1. Every dataset is different: you are the domain expert
2. Use histograms and text clusters to check coverage and relevancy
3. Cross-reference usage logs with your indexed data