CDW
Final Presentation

Team Members:
Rui Xue, Yanfeng Xu, Tianyi Wu, Ruoan Ni
Date: 2024.05.06

Agenda

- **Project Introduction**
- **Technical Explanation**
- **Machine Reading**
- **Web Application and CDW Flow Mapping**
- **Future Improvements**
- **Q&A**

NYU

MEET THE TEAM

Rui Xue
Meeting Moderator

Ruoan Ni
Meeting Notes

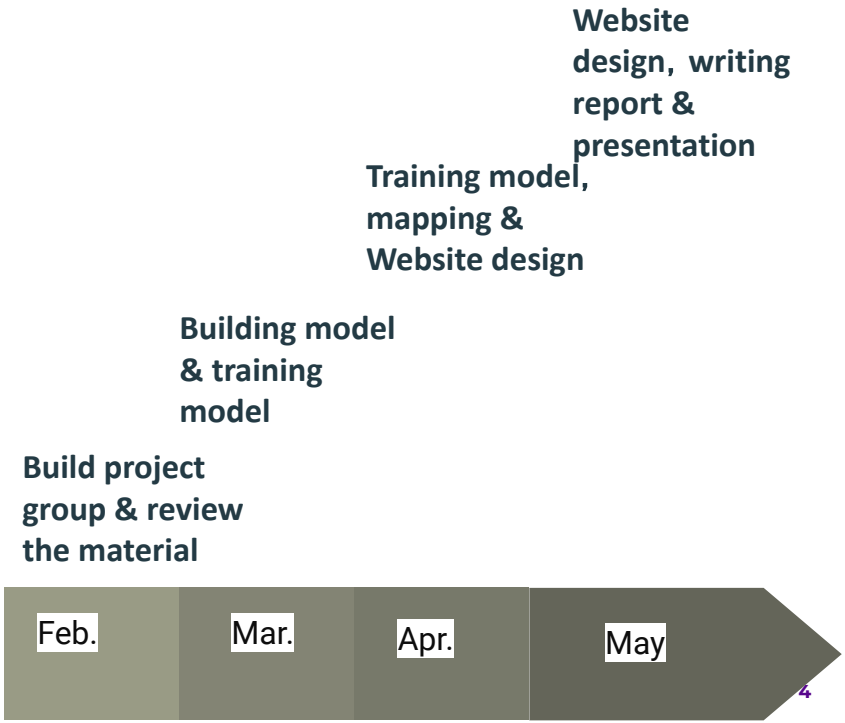
Yanfeng Xu
JIRA

Tianyi Wu
Email Communication

NYU

Project Introduction

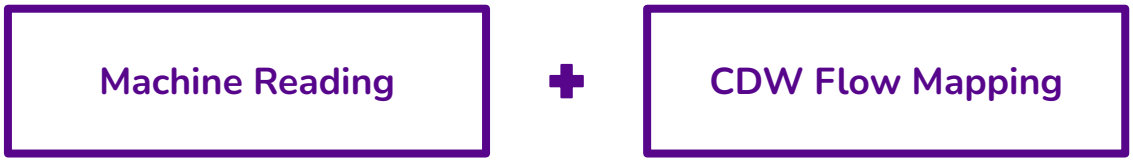
This project is the capstone project for the Spring 2024 Master of Science in Technology Management and Innovation program and focuses on the processing of Construction and Demolition (CDW) reports provided by the New York State Department of Environmental Conservation (NYS DEC) using machine reading and data mapping technologies. This project picks up and builds on the results of the Spring 2023 course project with the goal of improving and applying the Google Cloud, OCR, AI technologies used by previous teams.



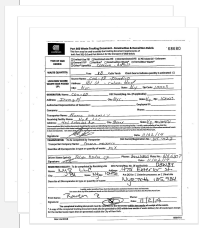
Project Introduction

In summary, our project work can be described as integrating and optimizing the work of previous teams. On one hand, we continued the approach of the previous MOT machine reading team, and utilized AI tools provided on Google Cloud Platform to implement a program that automatically extracts content from NYC DEC documents.

On the other hand, we followed the approach of the CUSP team, using the Streamlit framework to visualize the extracted data, which will ultimately be displayed on a map view. This allows users to easily see the start and end points of each waste transport, along with other information. The entire workflow, or streamline, is illustrated in the following diagram.



Project Introduction



Documents

Document AI

CSV

Extracted Data

Google Maps

Validation API

Clean and Correction

CSV

Prepared Data

+

Longitude
Latitude

Streamlit

Visualization

Our Web Application

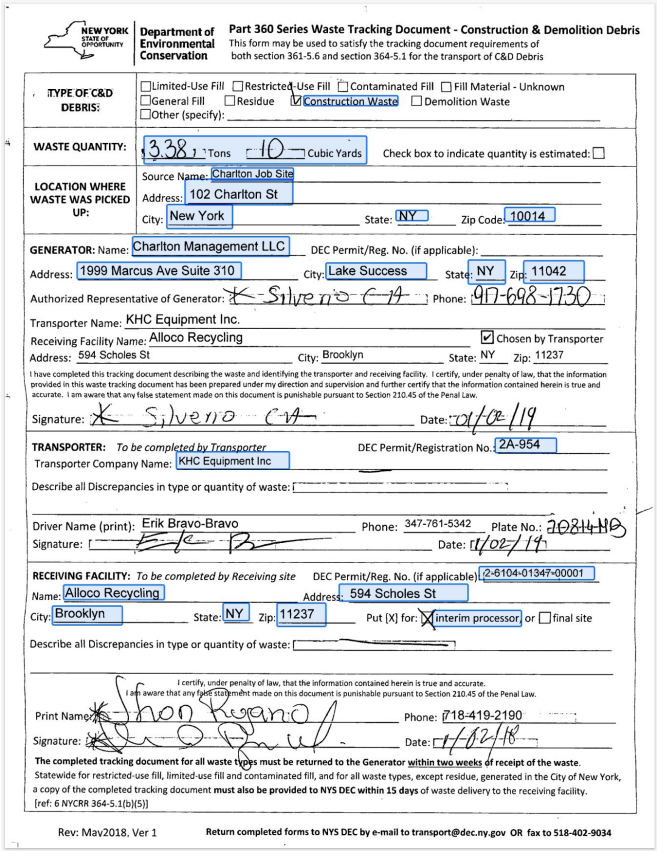
NYU

Technical Explanation

Documents

As shown in the picture on the right, this is a standard DEC waste tracking document, and the information we need to extract is highlighted in the blue box. In such a document, we focus on the type and quantity of waste, the pickup location, the entity that generated the waste, the transporter, and the facility that receives the waste.

We extract this information using tools provided on Google Cloud Platform, specifically Document AI, and then write programs for further processing.



NEW YORK STATE OF OPPORTUNITY. Department of Environmental Conservation

Part 360 Series Waste Tracking Document - Construction & Demolition Debris

This form may be used to satisfy the tracking document requirements of both section 361-5.6 and section 364-5.1 for the transport of C&D Debris

TYPE OF C&D DEBRIS: ☐ Limited-Use Fill ☐ Restricted-Use Fill ☐ Contaminated Fill ☐ Fill Material - Unknown ☐ General Fill ☐ Residue ☑ Construction Waste ☐ Demolition Waste ☐ Other (specify):

WASTE QUANTITY: 3.38 Tons 10 Cubic Yards Check box to indicate quantity is estimated: ☐

LOCATION WHERE WASTE WAS PICKED UP: Source Name: Charlton Job Site; Address: 102 Charlton St; City: New York; State: NY; Zip Code: 10014

GENERATOR: Name: Charlton Management LLC; DEC Permit/Reg. No. (if applicable):; Address: 1999 Marcus Ave Suite 310; City: Lake Success; State: NY; Zip: 11042; Authorized Representative of Generator: K Silvero CM; Phone: 917-698-1730

Transporter Name: KHC Equipment Inc.

Receiving Facility Name: Alloco Recycling ☑ Chosen by Transporter

Address: 594 Scholes St City: Brooklyn State: NY Zip: 11237

I have completed this tracking document describing the waste and identifying the transporter and receiving facility. I certify, under penalty of law, that the information provided in this waste tracking document has been prepared under my direction and supervision and further certify that the information contained herein is true and accurate. I am aware that any false statement made on this document is punishable pursuant to Section 210.45 of the Penal Law.

Signature: K Silvero CM Date: 01/02/19

TRANSPORTER: *To be completed by Transporter* DEC Permit/Registration No.: 2A-954

Transporter Company Name: KHC Equipment Inc

Describe all Discrepancies in type or quantity of waste:

Driver Name (print): Erik Bravo-Bravo Phone: 347-761-5342 Plate No.: 2084NB

Signature: E B Date: 1/02/19

RECEIVING FACILITY: *To be completed by Receiving site* DEC Permit/Reg. No. (if applicable): 2-6104-01347-00001

Name: Alloco Recycling Address: 594 Scholes St

City: Brooklyn State: NY Zip: 11237 Put [X] for: ☒ interim processor or ☐ final site

Describe all Discrepancies in type or quantity of waste:

I certify, under penalty of law, that the information contained herein is true and accurate. I am aware that any false statement made on this document is punishable pursuant to Section 210.45 of the Penal Law.

Print Name: Shon Ryan Phone: 718-419-2190

Signature: Shon Ryan Date: 1/02/19

The completed tracking document for all waste types must be returned to the Generator within two weeks of receipt of the waste. Statewide for restricted-use fill, limited-use fill and contaminated fill, and for all waste types, except residue, generated in the City of New York, a copy of the completed tracking document **must also be provided to NYS DEC within 15 days** of waste delivery to the receiving facility. [ref: 6 NYCRR 364-5.1(b)(5)]

Rev: May2018, Ver 1 **Return completed forms to NYS DEC by e-mail to transport@dec.ny.gov OR fax to 518-402-9034**

NYU

Technical Explanation

Document AI

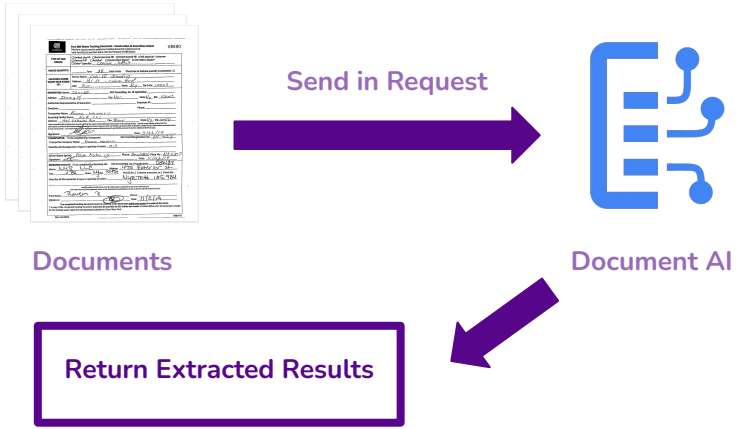
Document AI can be used to extract data from documents. Document AI supports a lot of formats, and uses generative AI to extract and structure data. Document AI has high-accuracy to extract, classify, and split. Notably, integrated with generative AI, it can be trained to improve accuracy, which also means developers should spend some time on data labeling and model training based on the foundational models. We simply need to establish a connection with GCP, obtain the necessary permissions, and then we can send documents via a request to the client in a local or other environment to call the trained model for processing. After processing, Document AI will return the extracted results.

```python
def process_page(page):
    image = page.get_pixmap()
    image_bytes = image.tobytes("png")  # Convert the page to PNG bytes

    # Create a document
    raw_document = documentai.RawDocument(content=image_bytes, mime_type='image/png')

    # Create a request
    request = documentai.ProcessRequest(name=processor_name, raw_document=raw_document)

    # Process the document
    result = client.process_document(request=request)
```



NYU

Technical Explanation

Validation API

The information extracted directly through Document AI includes many address details. However, due to various reasons such as errors in the document itself or inaccuracies in model recognition, there can be many issues with the extracted data, necessitating the need for verification and correction of these addresses. Additionally, if we aim to perform CDW flow mapping and visually represent the flow of CDW on a map, we require latitude and longitude information for the respective addresses.

This aspect of the project is particularly challenging. After much comparison and decision-making, we ultimately chose to use the validation API provided by Google Maps Platform.

	A	B	C	D	E	F	G	H	I
1	type_debris	waste_quantity	pickup_name	pickup_address	pickup_city	pickup_sta	pickup_zip	pickup_lat	pickup_lng
2	Contaminated Fill	12 Cubic Yards	Sisters of Saint Joseph Convent	1725 Brentwood Road	Brentwood	NY	11717		
3	Sand Mud	4.2 Tons	DOS SI Blue West Building	600 West Service Road	Staten Island	NY	10314		
4	Sand Mud	12.89 Tons	DOS SI Blue West Building	600 West Service Road	Staten Island	NY	10314		
5	Sand Mud	12.88 Tons	DOS SI Blue West Building	600 West Service Road	Staten Island	NY	10314		
6	Demolition Waste		Asplundh Yard	294 Old Northport Rd	Kings Park	NY	11754		
7	Sand Mud	11.7 Tons	DOS SI Blue West Building	600 West Serovice Road	Staten Island	NY	10314		
8	General Fill, Residue, Mud/Silt	12 Cubic Yards	SIMS METAL MANAGEMENT	3027 Greenpoint Ave	Long Island City	NY	11101		
9	General Fill, Residue, Mud/Silt	12 Cubic Yards	SIMS METAL MANAGEMENT	3027 Greenpoint Ave	Long Island City	NY	11101		
10	Demolition Waste	6 Tons	Asplundh Yard	294 Old Northport Rd	Kings Park	NY	11754		
11		10 Tons	Asplundh Yard	400 Charlotte Ave	Hicksville	NY	11801		
12	Demolition Waste	7 Tons	Asplundh Yard	93 Sills Road	Yaphank	NY	11980		
13	Construction Waste	7 Tons	Asplundh Yard	93 Sills Road	Yaphank	NY	11980		
14	Construction Waste	4 Tons	Asplundh Yard	294 Old Northport Rd	Kings Park	NY	11754		
15	Construction Waste	6 Tons	Asplundh Yard	93 Sills Road	Yaphank	NY	11980		
16	Construction Waste	6 Tons	Asplundh Yard	Conklin St & Broadhollow Rd	Farmigdale	NY	11735		
17	Demolition Waste	15 Cubic Yards	Ahern Rentals	45 Brook Ave	Deer Park	NY	11729		
18		15 Cubic Yards	Ahern Rentals	45 Brook Ave	Deer Park	NY	11729		
19	Contaminated Fill	15 Cubic Yards	Patchogue Fire District	25 Park Street	Patchogue	NY	11772		
20	Contaminated Fill	15 Cubic Yards	Patchogue Fire District	25 Park Street	Patchogue	NY	11772		
21	Contaminated Fill	15 Cubic Yards	Patchogue Fire District	25 Park Street	Patchogue	NY	11772		
22	Contaminated Fill	15 Cubic Yards	Crown Petroleum - Truck Roll Over	NY-878 & Burnside Ave	Inwood	NY	11096		

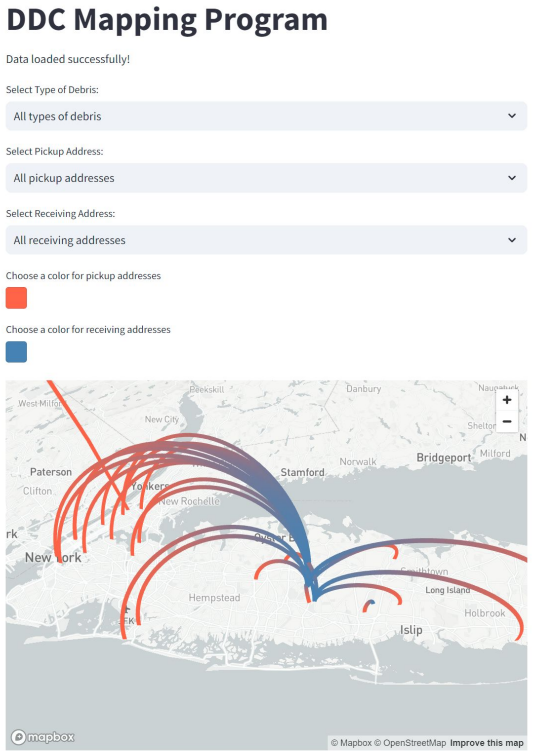
	A	B	C	D	E	F	G	H	I
1	type_debris	waste_quantity	pickup_name	pickup_address	pickup_city	pickup_state	pickup_zip	pickup_lat	pickup_lng
2	Contaminated Fill	12 Cubic Yards	Sisters of Saint Joseph Convent	1725 Brentwood Road	Brentwood	NY	11717	40.77356879	-73.24175059
3	Sand Mud	4.2 Tons	DOS SI Blue West Building	600 West Service Road	Staten Island	NY	10314	40.5856714	-74.19314981
4	Sand Mud	12.89 Tons	DOS SI Blue West Building	600 West Service Road	Staten Island	NY	10314	40.5856714	-74.19314981
5	Sand Mud	12.88 Tons	DOS SI Blue West Building	600 West Service Road	Staten Island	NY	10314	40.5856714	-74.19314981
6	Demolition Waste	Unknown	Asplundh Yard	294 Old Northport Rd	Kings Park	NY	11754	40.86997063	-73.26886954
7	Sand Mud	11.7 Tons	DOS SI Blue West Building	600 West Serovice Road	Staten Island	NY	10314	40.5856714	-74.19314981
8	General Fill, Residue, Mud/Silt	12 Cubic Yards	SIMS METAL MANAGEMENT	3027 Greenpoint Ave	Long Island City	NY	11101	40.73737917	-73.94423263
9	General Fill, Residue, Mud/Silt	12 Cubic Yards	SIMS METAL MANAGEMENT	3027 Greenpoint Ave	Long Island City	NY	11101	40.73737917	-73.94423263
10	Demolition Waste	6 Tons	Asplundh Yard	294 Old Northport Rd	Kings Park	NY	11754	40.86997063	-73.26886954
11	Unknown	10 Tons	Asplundh Yard	400 Charlotte Ave	Hicksville	NY	11801	40.76129013	-73.54961434
12	Demolition Waste	7 Tons	Asplundh Yard	93 Sills Road	Yaphank	NY	11980	40.81950867	-72.94540839
13	Construction Waste	7 Tons	Asplundh Yard	93 Sills Road	Yaphank	NY	11980	40.81950867	-72.94540839
14	Construction Waste	4 Tons	Asplundh Yard	294 Old Northport Rd	Kings Park	NY	11754	40.86997063	-73.26886954
15	Construction Waste	6 Tons	Asplundh Yard	93 Sills Road	Yaphank	NY	11980	40.81950867	-72.94540839
16	Construction Waste	6 Tons	Asplundh Yard	Conklin St & Broadhollow Rd	Farmigdale	NY	11735	40.73850107	-73.42329816
17	Demolition Waste	15 Cubic Yards	Ahern Rentals	45 Brook Ave	Deer Park	NY	11729	40.74376831	-73.30812875
18	Unknown	15 Cubic Yards	Ahern Rentals	45 Brook Ave	Deer Park	NY	11729	40.74376831	-73.30812875
19	Contaminated Fill	15 Cubic Yards	Patchogue Fire District	25 Park Street	Patchogue	NY	11772	40.75410527	-73.00863276
20	Contaminated Fill	15 Cubic Yards	Patchogue Fire District	25 Park Street	Patchogue	NY	11772	40.75410527	-73.00863276
21	Contaminated Fill	15 Cubic Yards	Patchogue Fire District	25 Park Street	Patchogue	NY	11772	40.75410527	-73.00863276
22	Contaminated Fill	15 Cubic Yards	Crown Petroleum - Truck Roll Over	NY-878 & Burnside Ave	Inwood	NY	11096	40.62302754	-73.74098061
23	Contaminated Fill	15 Cubic Yards	Crown Petroleum - Truck Roll Over	NY-878 & Burnside Ave	Inwood	NY	11096	40.62302754	-73.74098061
24	Contaminated Fill	15 Cubic Yards	Crown Petroleum - Truck Roll Over	NY-878 & Burnside Ave	Inwood	NY	11096	40.62302754	-73.74098061
25	Contaminated Fill	15 Cubic Yards	Crown Petroleum - Truck Roll Over	NY-878 & Burnside Ave	Inwood	NY	11096	40.62302754	-73.74098061
26	Contaminated Fill	Unknown	Residence	108 Candy Lane	Syosset	NY	11791	40.80829039	-73.50288001

NYU

Technical Explanation

Streamlit Framework

After processing the data, we visualize it by building a web application, similar to the work done by the CUSP team. We use the Streamlit framework to construct our web app. Streamlit provides an easy and fast way to build web applications and offers free community cloud resources. It allows applications to be hosted and run in the cloud, with the code stored in a GitHub repository. The diagram on the right shows one of the pages from the web app we developed in this project.



Machine Reading

1. Conducted in Jupyter Notebook

2. extracted data (22,501)

3. bottleneck in our work

4. future improvement

```python
def process_data(text: str) -> str:
    return text.replace("\n", " ")

def process_page(page):
    image = page.get_pixmap()
    image_bytes = image.tobytes("png")  # Convert the page to PNG bytes

    # Create a document
    raw_document = documentai.RawDocument(content=image_bytes, mime_type='image/png')

    # Create a request
    request = documentai.ProcessRequest(name=processor_name, raw_document=raw_document)

    # Process the document
    result = client.process_document(request=request)
    page_data = {col: '' for col in columns} # Initial a blank row
    for entity in result.document.entities:
        if entity.type_ in page_data: # Check if it is required
            if entity.type_ == 'type_debris':
                page_data[entity.type_] = process_type(entity.mention_text)
            else:
                page_data[entity.type_] = process_data(entity.mention_text)
    return page_data

def process_pdf(file_path, csv_file_path):
    doc = fitz.open(file_path)  # Open the PDF file
    data = []
    for page_num, page in enumerate(doc, start=1):
        print(f"Processing page {page_num}")
        page_data = process_page(page)
        data.append(page_data)

    # Create DataFrame and save it to CSV
    df = pd.DataFrame(data)
    df.to_csv(csv_file_path, mode='a', header=False, index=False)

def process_folder(folder_path, csv_file_path):
    for filename in os.listdir(folder_path):
        if filename.lower().endswith('.pdf'):
            file_path = os.path.join(folder_path, filename)
            print(f"Processing file: {file_path}")
            process_pdf(file_path, csv_file_path)
```

In [3]:

```python
csv_file_path = 'interim_csv.csv'
folder_path = "NYS Tracking Documents/R1/1A-301 PARK TRUCKING, INC"

process_folder(folder_path, csv_file_path)
```

```
Processing file: NYS Tracking Documents/R1/1A-301 PARK TRUCKING, INC\1A-301_Park_Trucking_cdd.2021-09D.wtd.pdf
Processing page 1
Processing page 2
Processing page 3
Processing page 4
Processing page 5
Processing page 6
```

NYU

Machine Reading

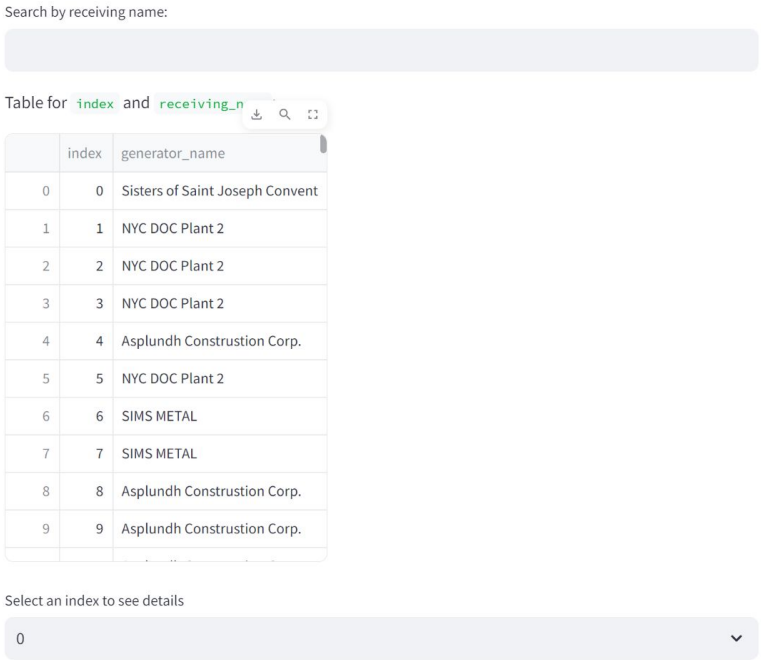
	L	M	N	O	P	Q	R	S	T	U	V
661	Long Island City	NY	11101	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
662	Long Island City	NY	11101	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
663	Long Island City	NY	11101	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
664	Long Island City	NY	11101	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
665	Long Island City, Queens	NY	11101	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
666	Long Island City	NY	11101	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
667	Long Island City	NY	11101	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
668	Long Island City	NY	11101	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
669	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
670	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
671	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
672	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
673	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
674	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
675	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
676	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
677	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
678	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
679	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
680	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
681	New York	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
682	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
683	Brooklyn	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
684	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
685	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
686	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
687	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
688	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
689	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
690	New York	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
691	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
692	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
693	Rosedale	NY	11422	Russo Development Enterprises, Inc.	Town of Babylon Recycling Center	57 Field Street	West Babylon	NY	11704	40.7321612	-73.3872518
694	Rosedale	NY	11422	Russo Development Enterprises, Inc.	Town of Babylon Recycling Center	57 Field Street	West Babylon	NY	11704	40.7321612	-73.3872518
695	Rosedale	NY	11422	Russo Development Enterprises, Inc.	Town of Babylon Recycling Center	57 Field Street	West Babylon	NY	11704	40.7321612	-73.3872518
696	Rosedale	NY	11422	Russo Development Enterprises, Inc.	Town of Babylon Recycling Center	57 Field Street	West Babylon	NY	11704	40.7321612	-73.3872518
697	Rosedale	NY	11422	Russo Development Enterprises, Inc.	Town of Babylon Recycling Center	57 Field Street	West Babylon	NY	11704	40.7321612	-73.3872518
698	Rosedale	NY	11422	Russo Development Enterprises, Inc.	Town of Babylon Recycling Center	57 Field Street	West Babylon	NY	11704	40.7321612	-73.3872518
699	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
700	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
701	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
702	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
703	Rosedale	NY	11422	Russo Development Enterprises, Inc.	Town of Babylon Recycling Center	57 Field Street	West Babylon	NY	11704	40.7321612	-73.3872518
704	Brooklyn	NY	11559	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
705	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
706	Long Island City	NY	11101	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
707	Long Island City	NY	11101	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
708	Long Island City	NY	11101	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
709	Long Island City	NY	11101	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
710	Long Island City	NY	11101	Russo Development Enterprises, Inc.	Cooper Recycling	123 Varick Ave	Brooklyn	NY	11237	40.7099207	-73.9291852
711	Long Island City	NY	11101	Russo Development Enterprises, Inc.	Allocco Recycling Corp	540 Kingsland Ave	Brooklyn	NY	11222	40.7355308	-73.9443812
712	Long Island City	NY	11101	Russo Development Enterprises, Inc.	Allocco Recycling Corp	540 Kingsland Ave	Brooklyn	NY	11222	40.7355308	-73.9443812
713	Long Island City	NY	11101	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
714	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
715	Brooklyn	NY	11559	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
716	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888
717	Lawrence	NY	11559	Russo Development Enterprises, Inc.	Russo Recycling Company	24812 Brookville Blvd	Rosedale	NY	11422	40.6384631	-73.7431888

NYU

Web Application

In this section, users can browse the data in the database.

Users can enter the keyword receiving name to search for the data content they want.



NYU

Web Application

```
  # Implement a search functionality
  search_query = st.text_input("Search by receiving name:")
```

```
# Filter data based on the search input
if search_query:
    filtered_table = small_table[small_table['generator_name'].str.contains(search_query, case=False)].copy()
else:
    filtered_table = small_table

st.write("Table for `index` and `receiving_name`:")
st.dataframe(filtered_table)
```

User can search keywords to find a specific receiving name.

Match the keywords with the receiving name in database and provide details.

Web Application

Three packages are used in mapping part.

- **Streamlit:** Implement web interfaces through code and provide interactive data presentation to users.
- **Pandas:** Read the CSV file and perform data cleaning.
- **Pydeck:** Visualize the transportation routes of CSV files in the form of maps.

Web Application

```
unique_debris_types = ['All types of debris'] + list(df['type_debris'].unique())
selected_debris = st.selectbox('Select Type of Debris:', unique_debris_types)

unique_pickup_addresses = ['All pickup addresses'] + list(df['pickup_address'].unique())
selected_pickup_address = st.selectbox('Select Pickup Address:', unique_pickup_addresses)

unique_receiving_addresses = ['All receiving addresses'] + list(df['receiving_address'].unique())
selected_receiving_address = st.selectbox('Select Receiving Address:', unique_receiving_addresses)
```

This code implements a user interaction function, where users can filter transportation information by conditions to display the routes they want to see.

Web Application

```
routes = [
    {
        "from_coordinates": [row['pickup_lng'], row['pickup_lat']],
        "to_coordinates": [row['receiving_lng'], row['receiving_lat']],
        "info": f"Type of Debris: {row['type_debris']}<br>"
                f"Waste Quantity: {row['waste_quantity']}<br>"
                f"Pickup Name: {row['pickup_name']}<br>"
                f"Pickup Address: {row['pickup_address']}<br>"
                f"Generator Name: {row['generator_name']}<br>"
                f"Generator Address: {row['generator_address']}"
    }
    for _, row in filtered_data.iterrows()
]
```

The purpose of this code is to display all information about the transportation when the user hovers the mouse over the route.

Web Application

```
routes = [
    {
        "from_coordinates": [row['pickup_lng'], row['pickup_lat']],
        "to_coordinates": [row['receiving_lng'], row['receiving_lat']],
        "info": f"Type of Debris: {row['type_debris']}<br>"
                f"Waste Quantity: {row['waste_quantity']}<br>"
                f"Pickup Name: {row['pickup_name']}<br>"
                f"Pickup Address: {row['pickup_address']}<br>"
                f"Generator Name: {row['generator_name']}<br>"
                f"Generator Address: {row['generator_address']}"
    }
    for _, row in filtered_data.iterrows()
]
```

The purpose of this code is to display all information about the transportation when the user hovers the mouse over the route.

Web Application

```
layer = pdk.Layer(
    "ArcLayer",
    routes,
    get_source_position="from_coordinates",
    get_target_position="to_coordinates",
    get_width=5,
    get_tilt=15,
    get_source_color=pickup_color_rgba,
    get_target_color=receiving_color_rgba,
    pickable=True,
    auto_highlight=True,
)
```

The purpose of this code is to make some basic settings for the map, such as color, coordinates, and line width.

Future Improvements

Although our work has concluded, there are still several shortcomings that need to be addressed by future teams, should there be any. Here, we offer some directions for improvement to the future teams.

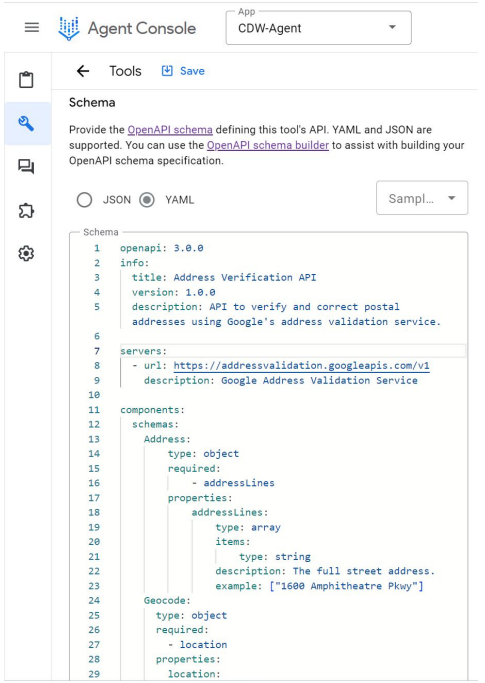
- **Correction and Valuation Improvements**
- **Database Improvements**
- **Further Model Training**

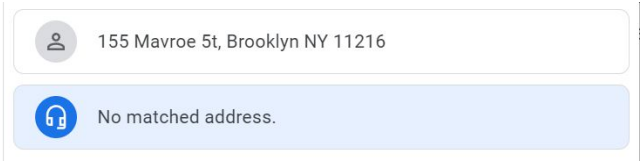
Future Improvements

Correction and Evaluation

As we previously mentioned, the validation API has its limitations, particularly its low tolerance for recognition errors. For example, it can correct a minor spelling mistake such as changing "155 Mavroe St, Brooklyn NY 11216" to "155 Monroe St, Brooklyn, NY 11216."

However, a common recognition error like "155 Mavroe 5t, Brooklyn NY 11216", where 'S' is misrecognized as '5' due to their visual similarity, cannot be corrected by the validation API. We have attempted to integrate Vertex AI with the validation API to address this issue using Generative AI, but the results were unsatisfactory and showed no significant improvement over using the validation API alone (see the AI Agent on the right). Future team need to consider this direction for improvement.

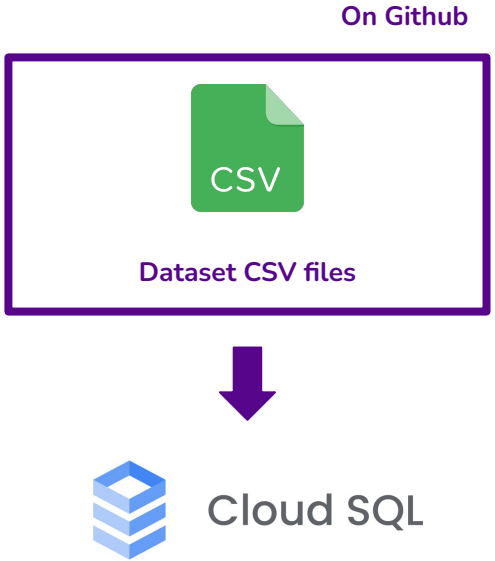




Future Improvements

Database Improvements

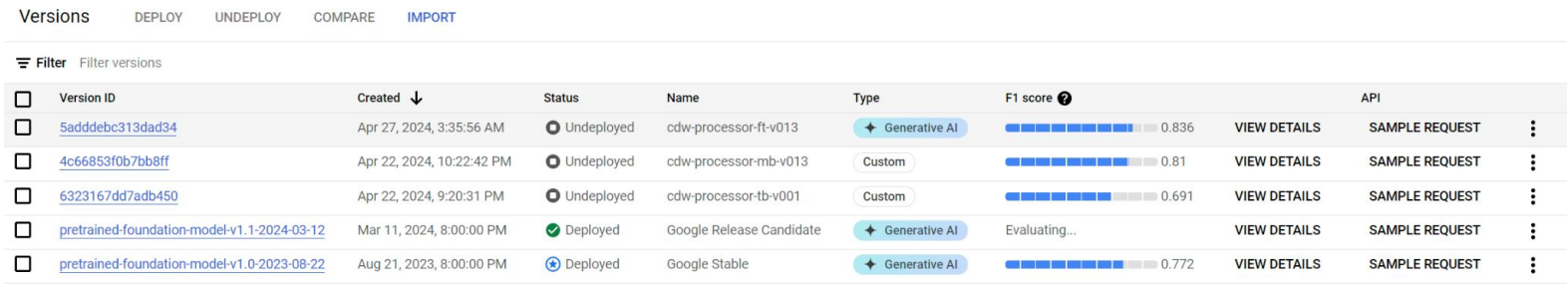
Currently, we don't have a database in the strict sense; instead, we store data in CSV files and extract data from them. It's important to note that our web application runs on code hosted on GitHub, and the CSV files are also hosted there. GitHub does not support changes to its files via non-Git commands, which means we can't directly modify the backend CSV files through the web app. We originally planned to migrate our data to a cloud database like Google Cloud SQL, but various reasons prevented this from happening. This migration is another improvement direction that future teams should consider.



Future Improvements

Further Model Training

We spent considerable time on model training, but the final results were still not satisfactory. After numerous iterations, the fine-tuned model achieved an F1 score of 0.836, indicating approximately 83.6% accuracy on our training documents. However, the actual performance still fell short of expectations. We hope future teams can further optimize and train the model to enhance its effectiveness.



Versions DEPLOY UNDEPLOY COMPARE IMPORT

Filter Filter versions

Version ID	Created	Status	Name	Type	F1 score		API	
5adddebc313dad34	Apr 27, 2024, 3:35:56 AM	Undeployed	cdw-processor-ft-v013	Generative AI	0.836	VIEW DETAILS	SAMPLE REQUEST	
4c66853f0b7bb8ff	Apr 22, 2024, 10:22:42 PM	Undeployed	cdw-processor-mb-v013	Custom	0.81	VIEW DETAILS	SAMPLE REQUEST	
6323167dd7adb450	Apr 22, 2024, 9:20:31 PM	Undeployed	cdw-processor-tb-v001	Custom	0.691	VIEW DETAILS	SAMPLE REQUEST	
pretrained-foundation-model-v1.1-2024-03-12	Mar 11, 2024, 8:00:00 PM	Deployed	Google Release Candidate	Generative AI	Evaluating...	VIEW DETAILS	SAMPLE REQUEST	
pretrained-foundation-model-v1.0-2023-08-22	Aug 21, 2023, 8:00:00 PM	Deployed	Google Stable	Generative AI	0.772	VIEW DETAILS	SAMPLE REQUEST	

NYU

One More Thing

About our report

In addition to the technical introduction and project details similar to this slide, our report will also include an analysis of the documents, roughly encompassing the following sections:

- Page counts of various types of documents and the total number of pages.
- Statistics on documents with different templates.
- Statistics and estimates of printed and handwritten documents.
- Issues and statistics for documents from which information could not be extracted.
- Statistics and estimates for documents with missing information.

We are still doing our best to analyze this collection of over one hundred thousand pages of documents.

NYU