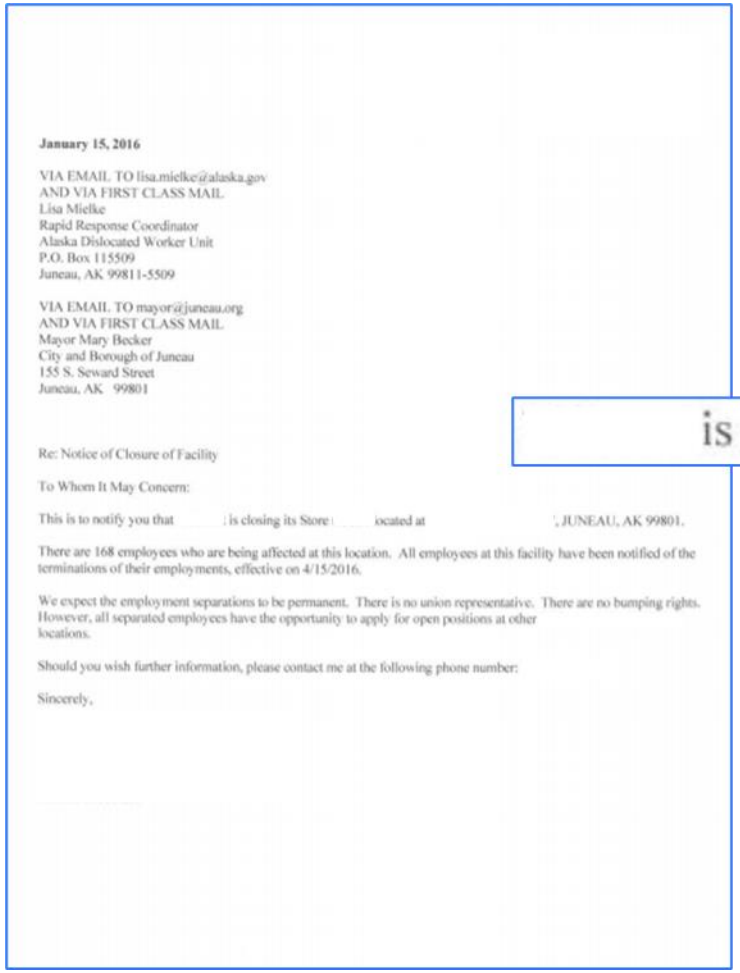
WARN Notice Archive Scraping

Lucia Korpas
CIF Data Science Fellow
OCWC

CIF End-of-Summer Presentations
August 5, 2021

What are WARN Notices?

- Worker Adjustment and Retraining Notification (WARN) Act
  - Advanced notice on mass layoffs/plant closures
  - Required above employee number thresholds
  - Archives made public by states



January 15, 2016

VIA EMAIL TO lisa.mielke@alaska.gov
AND VIA FIRST CLASS MAIL
Lisa Mielke
Rapid Response Coordinator
Alaska Dislocated Worker Unit
P.O. Box 115509
Juneau, AK 99811-5509

VIA EMAIL TO mayor@juneau.org
AND VIA FIRST CLASS MAIL
Mayor Mary Becker
City and Borough of Juneau
155 S. Seward Street
Juneau, AK 99801

Re: Notice of Closure of Facility

To Whom It May Concern:

This is to notify you that : is closing its Store : ocated at , JUNEAU, AK 99801.

There are 168 employees who are being affected at this location. All employees at this facility have been notified of the terminations of their employments, effective on 4/15/2016.

We expect the employment separations to be permanent. There is no union representative. There are no bumping rights. However, all separated employees have the opportunity to apply for open positions at other locations.

Should you wish further information, please contact me at the following phone number:

Sincerely,

Juneau, AK

is closing its Store #

168 employees

effective on 4/15/2016.

BLS Need for Data

- WARN notices used widely within BLS, but…
  - Not automatically collected for all states
  - Not held in a centralized location or common format
- How can we make these data easier for economists to use?
  - Automate collection
    - One-time full archive scrape
    - Monthly updates
  - Normalize their format
  - Centralize where they are stored

Project Goals

- Automate archive scraping (find and download all archive files)
- Automate archive parsing (read [mostly] tabular data)
- Export all parsed data in common format
- Strive to make the codebase easy to maintain

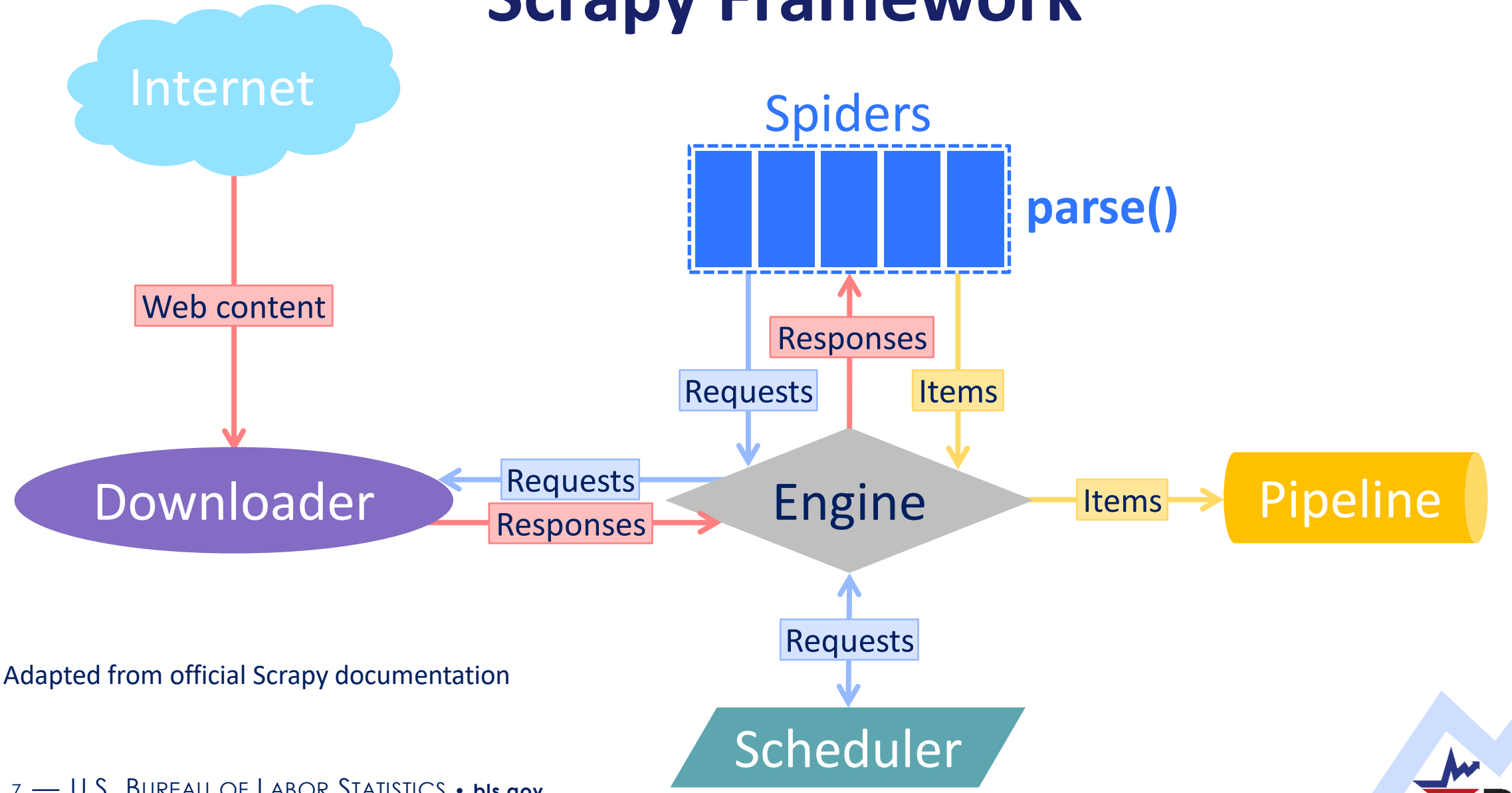
The Manual Step: Make Configuration File

- To automate scraping, need to know where to look and what to look for
  - Archive URLs
  - Data formats (HTML, Excel, etc.)
  - Fields provided (company name, layoff date, etc.)
  - Notes on other parsing considerations
- Created configuration file containing this information
  - Can be updated in Excel

Scraping and Parsing with Scrapy

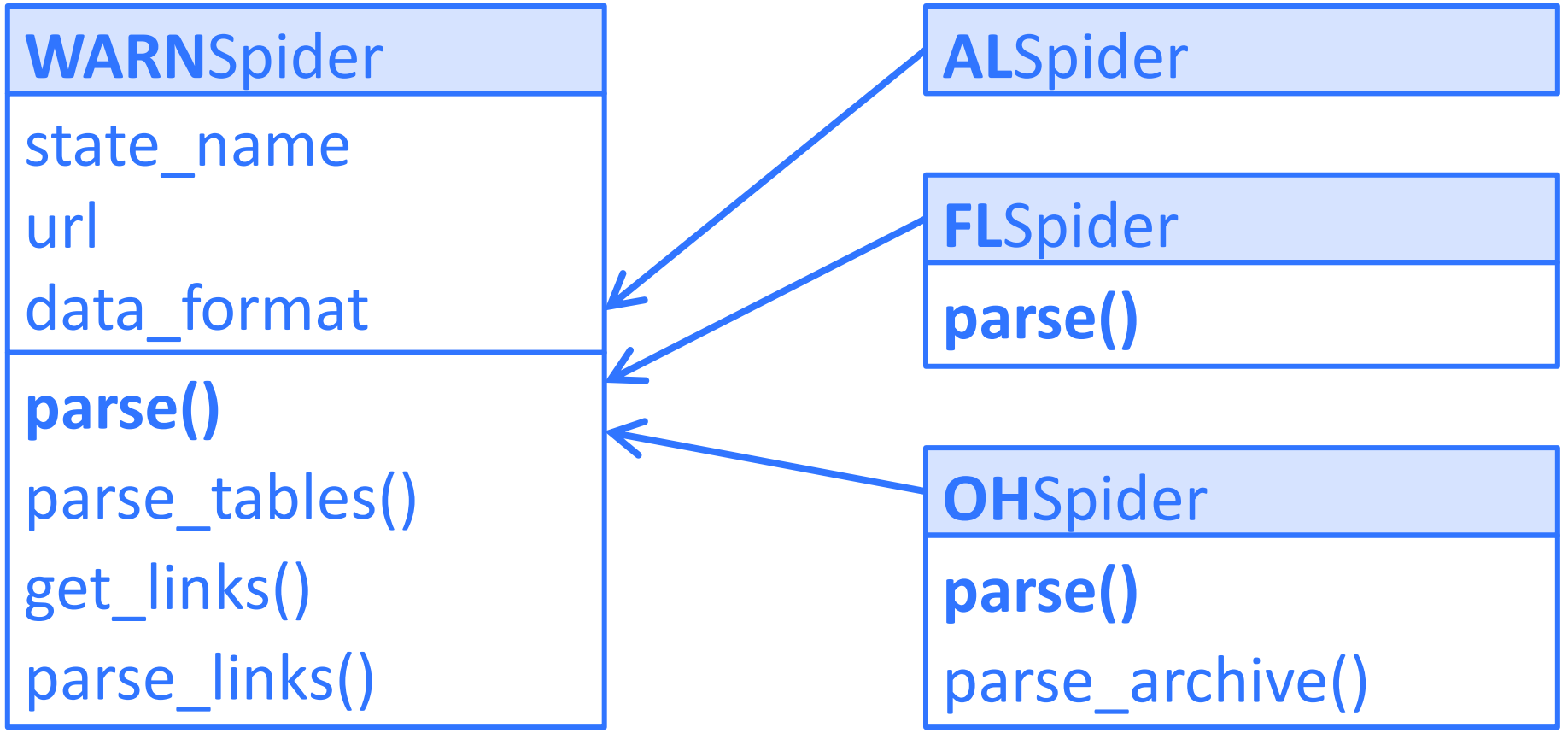
- Framework to download, clean, and save data from the internet, at scale
- For each website to scrape, create a Spider

Scrapy Framework



Adapted from official Scrapy documentation

Spider Classes



Parse Function Example

Example archive website (California)

Listing of Filed WARN Notices

WARN Report: WARN notices processed from July 1, 2021, to present (XLSX).

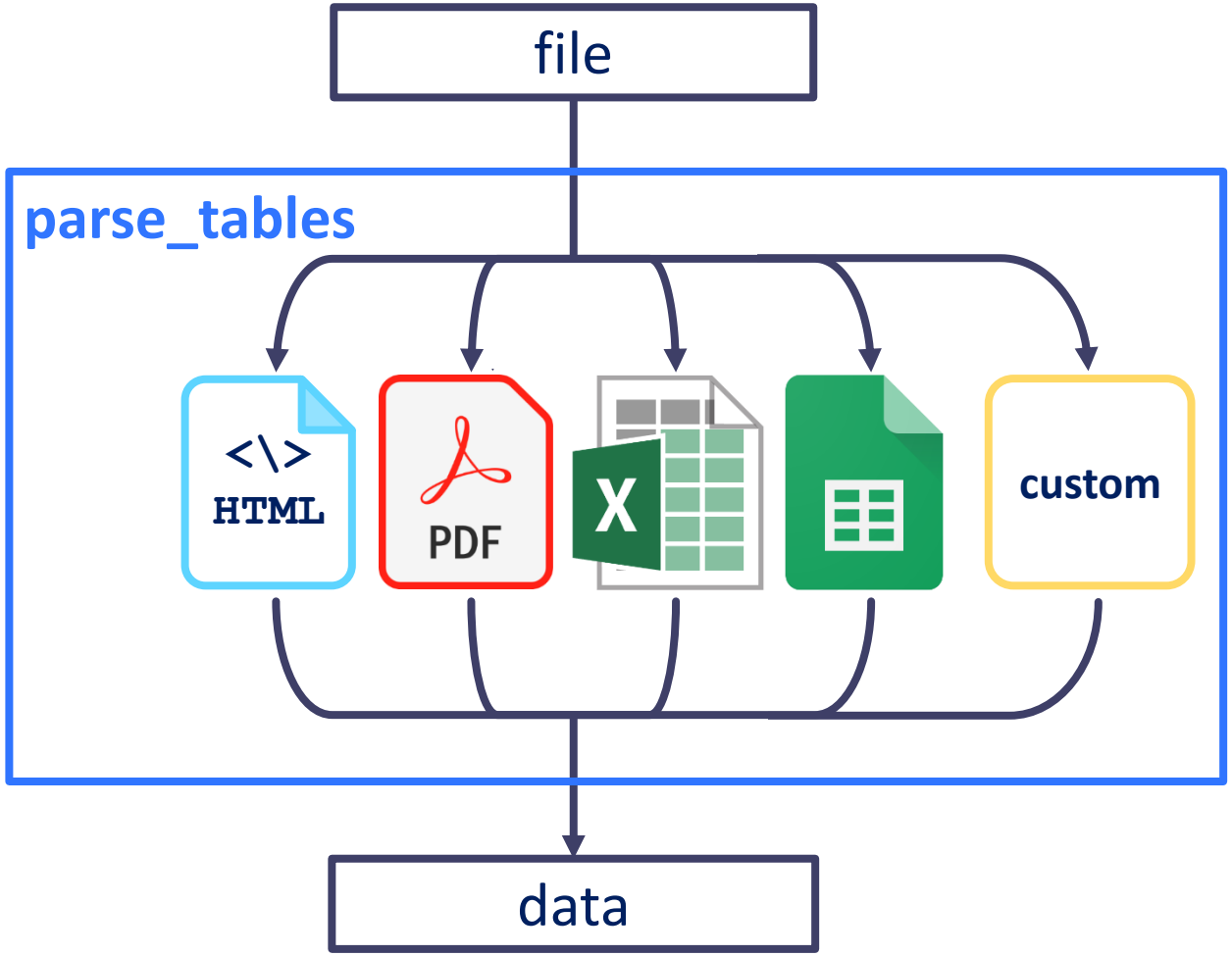
Listing of WARN Notices from previous years:

- 2020-2021 WARN Report from July 01, 2020 through June 30, 2021 (PDF)
- 2019-2020 WARN Report from July 01, 2019 through June 30, 2020 (PDF)
- 2018-2019 WARN Report from July 01, 2018 through June 30, 2019 (PDF)
- 2017-2018 WARN Report from July 01, 2017 through June 30, 2018 (PDF)
- 2016-2017 WARN Report from July 01, 2016 through June 30, 2017 (PDF)
- 2015-2016 WARN Report from July 01, 2015 through June 30, 2016 (PDF)
- 2014-2015 WARN Report from July 01, 2014 through June 30, 2015 (PDF)

Pseudocode for **CA**WARNSpider

```
def parse(response):
        parse_links(response,
                    link_has='.xlsx',
                    format=excel)
        parse_links(response,
                    link_has='.pdf',
                    format=pdf)
```

Parsing Tabular Data



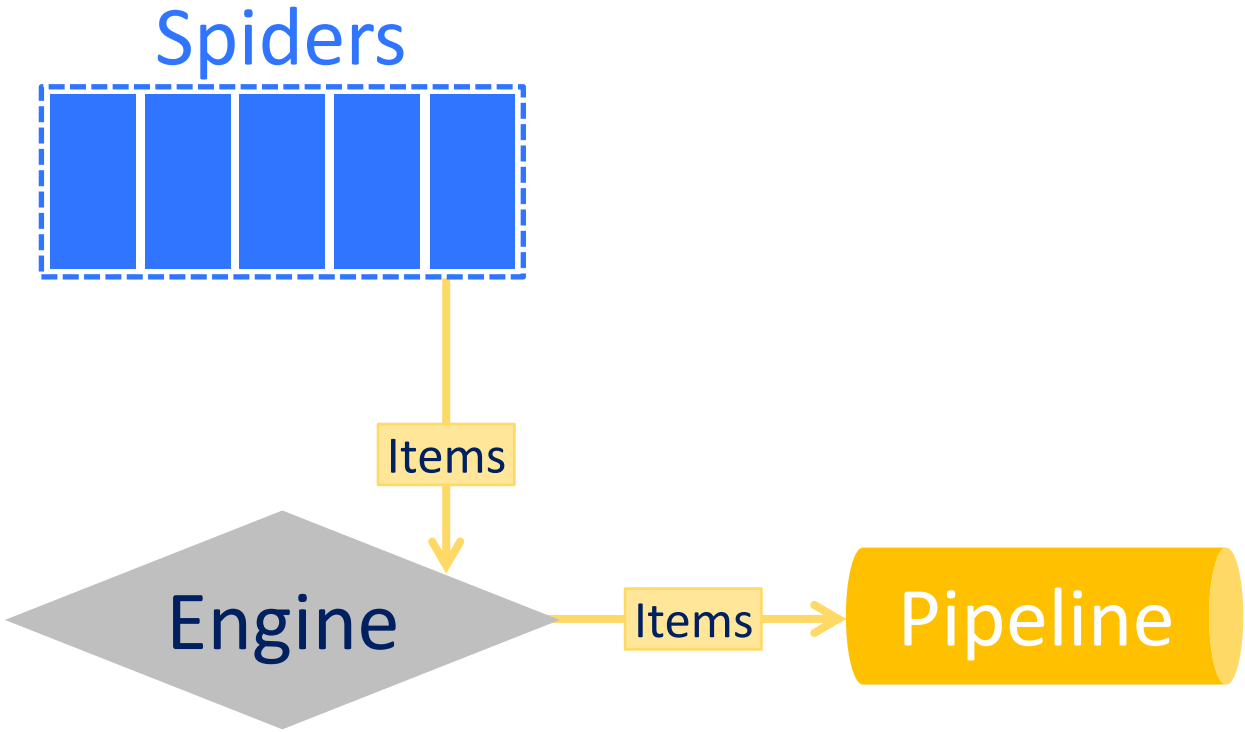
So How Do You Actually Extract Data?

- Wrote one standalone function for each format
  - Takes optional parameters to tweak parsing, if necessary
  - Return data as pandas DataFrame
- These functions work for tabular data in general!

Package	Purpose
pandas	Data manipulation
selenium	Scrape dynamic webpages
xlrd	Parse Excel (.xls) files
openpyxl	Parse Excel (.xlsx) files
lxml	Parse HTML
camelot	Parse tabular PDF files
pdfquery	Parse non-tabular PDF files

Exporting the Data: Pipelines

- Spider generates Item from each row of the DataFrame (i.e., each WARN entry)
- Items are yielded to Pipeline from Spider
- Pipeline defines export format and appends each Item to the right file



Current Output

- Scraped WARN notice data from 34 states
- Fields exported:
  - Common fields: company name, layoff date, # employees affected, etc.
  - Additional fields: link to original notice PDF, NAICS codes, unions, etc.
- Common fields also exported under normalized names
- Additional metadata included (timestamp, URL)

Output Example

Intermediate DataFrame

Example archive website (Alaska)

WARN NOTICES FILED

Company	Location	Notice Date	Layoff Date	Employees Affected	Notes
[redacted]	Prudhoe Bay	7/2/21	8/31/21	134	Loss of Contract
	Fairbanks, North Pole	4/30/21	6/30/21	185	Loss of contract
	Anchorage	4/12/21	August-November 2021	59	Business need changes/reorganization; permanent
	Anchorage and 20 smaller Alaska communities, plus Boston, Mass.	9/30/20*	8/7/20	1,234	COVID-19; permanent

Company	Employees Affected	Layoff Date	Location	Notes	Notice Date	Notice Link
[redacted]	134	8/31/21	Prudhoe Bay	Loss of Contract	7/2/21	https://jobs.alask
	185	6/30/21	Fairbanks, North Pole	Loss of contract	4/30/21	https://jobs.alask
	59	August-Novembe	Anchorage	Business need cha permanent	4/12/21	https://jobs.alask
	1,234	8/7/20	Anchorage and 20 sma	COVID-19; permar	9/30/20*	https://jobs.alask


```
{
    "state_name": "Alaska",
    "timestamp": "2021-08-05 13:00:56",
    "url": "https://jobs.alaska.gov/RR/WARN_noti
    "fields":
        {
            "Company": [redacted],
            "Employees Affected": "134",
            "Layoff Date": "8/31/21",
            "Location": "Prudhoe Bay",
            "Notes": "Loss of Contract",
            "Notice Date": "7/2/21",
            "Notice Link": "https://jobs.alaska.
        },
    "normalized_fields":
        {
            "company": [redacted],
            "affected_employees": "134",
            "layoff_date": "8/31/21",
            "city": "Prudhoe Bay",
            "notice_date": "7/2/21"
        }
}
```


Final JSONL output

Future Work

- Scrape and parse remaining state archive
- Data cleaning
- Schedule monthly scrape and automate merging new data
- Dashboard/visualization

Contact Information

Lucia Korpas
CIF Data Science Fellow
korpas.lucia@bls.gov

Supervisor: **Drake Gibson**
Data Scientist (Operations Research Analyst)
OCWC
gibson.drake@bls.gov