

# On Data: Collection & Creation

CS 485, Spring 2024

Applications of Natural Language Processing

[https://people.cs.umass.edu/~brenocon/cs485\\_s24/](https://people.cs.umass.edu/~brenocon/cs485_s24/)

*[Slides by Laure Thompson]*

# Upcoming

- This week: starting HW2

# General NLP problem

Input: some text  $\mathbf{x}$  ( e.g. sentence, document)

Output: a label  $\mathbf{y}$  (from some finite label set)

Goal: learn a mapping function  $f$  from  $\mathbf{x}$  to  $\mathbf{y}$

# General NLP problem

Input: some text  $x$  ( e.g. sentence, document)

Output: a label  $y$  (from some finite label set)

Goal: learn a mapping function  $f$  from  $x$  to  $y$

# What is a dataset?

A collection of texts and metadata



# Text sources can be used in many ways



## **Text Unit**

- Full Volume

## **Metadata / Label Examples**

- Author, Translator
- Genre, Literariness
- Publication Year

# Text sources can be used in many ways



## **Text Unit**

- Full Volume
- Passage

## **Metadata / Label Examples**

- Author, Translator
- Genre, Literariness
- Publication Year
- Elapsed Narrative Time
- Event Depiction
- Sarcasm, Irony, Suspense
- Memorability

# Text sources can be used in many ways



## **Text Unit**

- Post

## **Metadata / Label Examples**

- Popularity, Controversy
- Topic / Post Type
- Shared News Source



# Text sources can be used in many ways



## **Text Unit**

- Post
- Interaction  
(Post + Replies)

## **Metadata / Label Examples**

- Popularity, Controversy
- Topic / Post Type
- Shared News Source
- Agreement / Disagreement
- Moderator Intervention
- Condescending Language Use

# Where to get datasets?

Off the shelf



Build from  
scratch

# Where to get datasets?

Off the shelf



Build from  
scratch

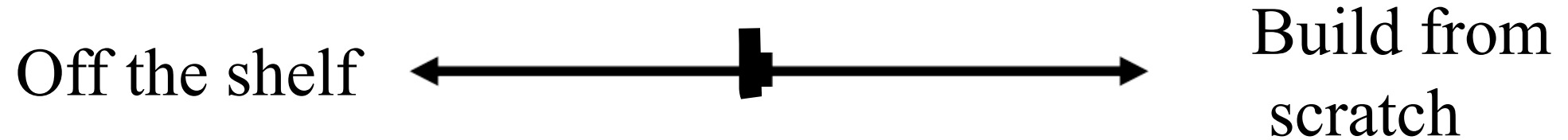
Q: What are the pros & cons of each  
strategy?

# Off - the - Shelf Datasets

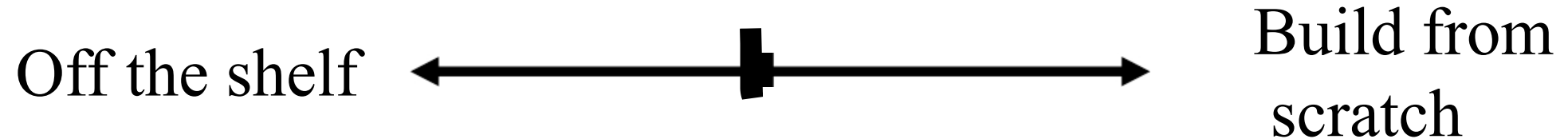
There are many publicly available datasets.

- <https://huggingface.co/datasets>
- <https://nlpprogress.com/>
- <https://index.quantumstat.com/#dataset>

# Middle ground: Augmentation & Curation



# Middle ground: Augmentation & Curation



- Use a subset of the texts based on metadata, content, etc.
- Combine texts from multiple data sources
- Align text and metadata from different sources
- Add new labels to an existing dataset

# Example: CMU Movie Summary Corpus

Dataset: 42,306 movie plot summaries extracted from Wikipedia  
+aligned metadata extracted from Freebase

<http://www.cs.cmu.edu/~ark/personas/>

# So you want to collect some data

1. Is it ethical to collect this data?
2. Is this data publicly available?
3. Is this data protected by copyright?



# U.S. Public Domain

Published works enter the public domain **95 years** after their first publication date (in the U.S.)

As of this year, works published in **1925** entered the US public domain!



# How to collect data?

- Digitization & Transcription
- APIs
- Bulk downloads / database dumps
- Web scraping

# Digitization

Watch out for OCR errors!

## THE WALWORTH MURDER.

The Story of a Member of the  
Bereaved Family.

Scenes Beside Chancellor Wal-  
worth's Coffin.

A Letter Written in Blood and Sanded  
with Powder.

A Wife's Silent Agony—Another Life To Be  
Saved at the Expense of the Memory of  
the Dead—Sad and Stern Facts  
which Must Come Out.

The Walworth tragedy, which a few days ago startled the community to its depths, has not yet lost any of its morbid interest for the public. Every little detail and surrounding of the horrible murder, the antecedents of the victim and his slayer, even everything that remotely concerns the bereaved family, is discussed with that unhealthy gusto which only a satiety of crime can produce. The demeanor of the wretched prisoner, the characteristics of the father and the sufferings of the unfortunate mother are still subjects of all-absorbing interest in gossiping circles.

## THE Wal-worth MURDER.

The Story of a Member of the  
Bereaved Family.

Scenes Beside Chancellor Wal-  
worth's Coffin.

A hotter Written in Blood and Sanded  
with Powder.

A Wife'i Bite Agony—Another Life To  
Be  
fised at the Expense of the Memory  
of  
the Dead-Sad and Stem Facts  
which Mwt Come Out.

The walworth tragedy, which a few days agox  
•turned the community to its depths,  
hag not yet toat any of its morbid  
intercut for the public. Kverjr UIUe

# APIs

Some examples:

- (now mostly closed) Twitter: twitter- api
- (now mostly closed?) Reddit: pushshift.io

---

Big problem: value of data and questions about tech company scrutiny

Q: What are the incentives for maintaining APIs?

# Q: What are the incentives for maintaining APIs?

## LibraryThing APIs

### Notice: LibraryThing APIs currently disabled

The LibraryThing APIs are disabled until further notice (last updated on 1/28/2021). If you would like to be notified when they are re-enabled, please email [info\(AT\)librarything.com](mailto:info(AT)librarything.com).

## Does Goodreads support the use of APIs?

🕒 Dec 17, 2020 · Knowledge

As part of our overall commitment to continually improve our data management, Goodreads no longer issues new developer keys for our public developer API and plans to retire the current version of these tools. While we assess the value of APIs to determine how to support in the future, we continue to support active API users who meet our terms of service. You are welcome to give your feedback on Goodreads APIs by [completing the developer API survey](#).

# Bulk Downloads

Wikipedia: [dumps.wikimedia.org/](https://dumps.wikimedia.org/)

Court Listener: [courtlistener.com/api/bulk - info/](https://courtlistener.com/api/bulk-info/)

# Web Scraping Responsibly

Be considerate

- Check Terms of Service
- Check for robots.txt
- Use low request rates



# Books



# Wikipedia derived datasets

**Wikitext** : High quality subset of Wikipedia pages.

**WikiMatrix** : Parallel sentences (across languages) of Wikipedia

**SQUAD**: Question - answer dataset that relies on Wikipedia text

There's no such thing as raw data.

