Deep Learning for Economists

Melissa Dell

August 2025

The Deep Learning Revolution

- Deep learning has revolutionized the processing of *unstructured* data (text, images, audio, video)
- This has in turn transformed a variety of disciplines, ranging from allowing NASA to land a rover on rugged Martian terrain to changing how doctors diagnose disease
- Deep learning can similarly be used to transform economic analyses

Unstructured Data in Economics

Massive quantities of non-computable data could power economic analyses if converted into a computable format:

- Text contains massive amounts of unstructured information
- Data can be trapped in images; also audio and video
- Libraries and archives have scanned billions of pages of historical documents
- While the raw information is very different, DL methods to convert them into computable information are quite related, often drawing on the same neural net architecture

Deep Learning

- Deep neural networks map typically unstructured data - such as text, document image scans, satellite and other imagery, videos, and audio - to a continuous vector space.
- In other words, they map complex and diverse types of data into a format that is easier to process and understand
- At its core, deep learning is an approach for learning representations of data from empirical examples (LeCun, Bengio and Hinton, 2015).

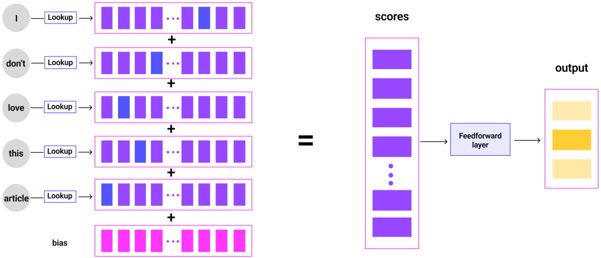
Why Deep Learning

Why would one use a neural network to transform the raw data into these vector representations, versus just working directly with raw texts or images?

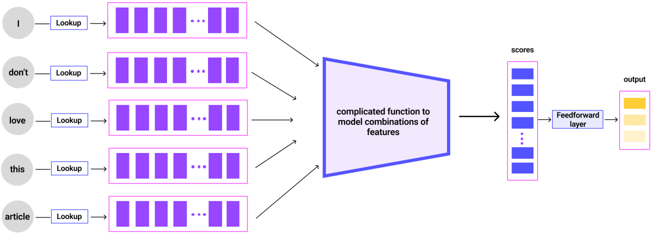
- **Transfer learning:** Deep neural nets incorporate relevant information in their parameters from exposure to massive-scale data
- **Context:** Raw pixels or words lack context. Deep neural networks provide a powerful method for computing contextualized representations
- **Scaleable:** Raw texts and images are computationally unwieldy. In contrast, there are extremely optimized tools for continuous vector computations

Context

Bag of Words



Neural Network



Computing in Economics

While computing has a long history in economics, the advent of personal computing in the 1990s revolutionized the discipline. Today, advances in GPU compute and the availability of cheap cloud compute again have the potential to transform the types of data and questions economists can study

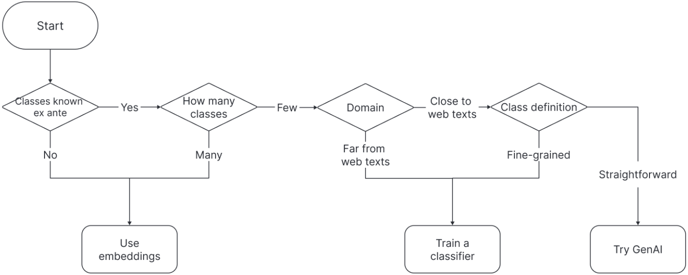
Deep Learning for Prediction

- Neural networks excel at imputing low-dimensional structured data from unstructured texts or images
- Conceptually, we can divide problems into:
    - Imputing a continuous number (regression)
    - Imputing a pre-specified discrete class (classification)
    - Imputing relationships in data when the classes are not specified ex ante

Deep Learning for Prediction

- A neural network encodes unstructured data into lower-dimensional vectors
- The researcher can use these representations to predict whether the raw data belong to pre-specified class(es) by adding a classifier layer; regression works analogously
- Broadly speaking, generative AI performs classification by predicting what word (in a pre-specified vocabulary) comes next
- Alternatively, one can work directly with the vector representations, referred to as *embeddings*

Classification Flow Chart



Outline

Introduction

Classification

Embedding Models
- Classes Unknown Ex Ante
- Many Classes
- Adding Classes at Inference Time

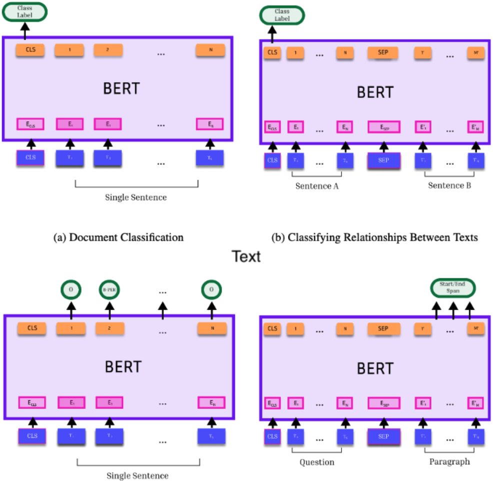
Regression

Conclusion

Classification

- In traditional classification, a neural network predicts a score for each of N classes, and the input is assigned the class with the highest score
- Central to the power of transformer neural networks is the ability to use the same pre-trained language model as the backbone for a wide variety of classification tasks

Classification with Transformers



(a) Document Classification

(b) Classifying Relationships Between Texts

Text

(c) Named Entity Recognition

(d) Text Span Classification

Classification

- Classifier training is a supervised task, and the model must see a sufficient number of examples from each class during training in order to perform well on unlabeled data.
- Alternatively, generative AI can be used for classification.
- In the JEL article, I compare custom-trained models to generative AI models for 19 different topic classification tasks. The bottom line is that customized models tend to be more accurate, particularly when domain shift is greater, but generative AI does well off-the-shelf for very straightforward topics.

Custom Classifier vs. GPT for Classification

Table: F1 scores for predictions

Topic	F1 on test set				
			GPT	Distil	RoBERTa
	GPT-3.5	GPT-4	Trained†	RoBERTa	Large
advice	0.72	0.85	0.55	0.87	**0.97**
antitrust	0.85	**0.94**	0.84	0.92	**0.94**
bible	0.52	0.81	0.10	0.85	**0.87**
civil rights	0.59	**0.87**	0.54	0.85	**0.87**
contraception	0.83	0.91	0.72	0.88	**0.97**
crime	0.85	0.80	0.85	0.85	**0.90**

Custom Classifier vs. GPT for Classification

Table: F1 scores for predictions

Topic	F1 on test set				
			GPT	Distil	RoBERTa
	GPT-3.5	GPT-4	Trained†	RoBERTa	Large
pesticides	0.58	0.91	0.71	0.89	**0.98**
polio vax	0.92	**0.99**	0.94	0.96	0.97
politics	0.67*	0.62*	0.74	**0.86**	0.85
protests	0.74	0.81	0.79	0.90	**0.91**
Red Scare	0.81	0.86	0.79	0.90	**0.91**
schedules	0.79	0.95	0.81	0.95	**0.96**
sports	0.80	0.92	0.88	**0.94**	**0.94**

Custom Classifier vs. GPT for Classification

Table: F1 scores for predictions

Topic	**F1 on test set**				
			GPT	Distil	RoBERTa
	GPT-3.5	GPT-4	Trained[†]	RoBERTa	Large
horoscope	**1.00**	**1.00**	0.92	0.96	**1.00**
labor movement	0.77	0.90	0.79	0.89	**0.94**
obituaries	0.98	**1.00**	**1.00**	0.96	**1.00**
Vietnam War	0.91	0.94	0.98	0.98	**0.99**
weather	0.94	0.92	0.94	0.94	**0.95**
WWI	0.72	0.74	0.51	0.89	**0.92**

Outline

Introduction

Classification

Embedding Models

- Classes Unknown Ex Ante
- Many Classes
- Adding Classes at Inference Time

Regression

Conclusion

Embedding models

- Embeddings can be used to group like data together, typically through clustering or knn-retrieval
- Off-the-shelf transformer (e.g., BERT, RoBERTa, etc.) embeddings have undesirable geometric properties (Ethayarajh, 2019)
- A method called contrastive training can be used to make the distances between embeddings in the vector spaced created by a neural network more meaningful (Wang, 2021)

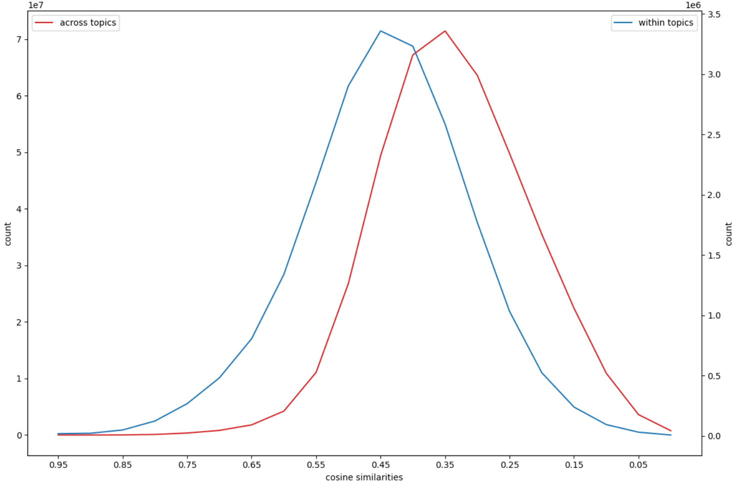
Contrastively trained embedding models

- Contrastively trained embedding models learn a mapping from unstructured data to continuous vector space, such that instances that belong to the same class have similar embeddings and instances that belong to different classes have dissimilar embeddings
- More details on contrastive learning are given in my JEL review article “Deep Learning for Economists”, or on the contrastive learning post at `https://econdl.github.io/`
- There is a lot of different information incorporated in an off-the-shelf embedding (e.g., Rogers et al., 2020)
- Hence, some fine-tuning is often necessary for the model to learn what dimension(s) of texts/images are of relevance to the task at hand

Contrastively trained embedding models

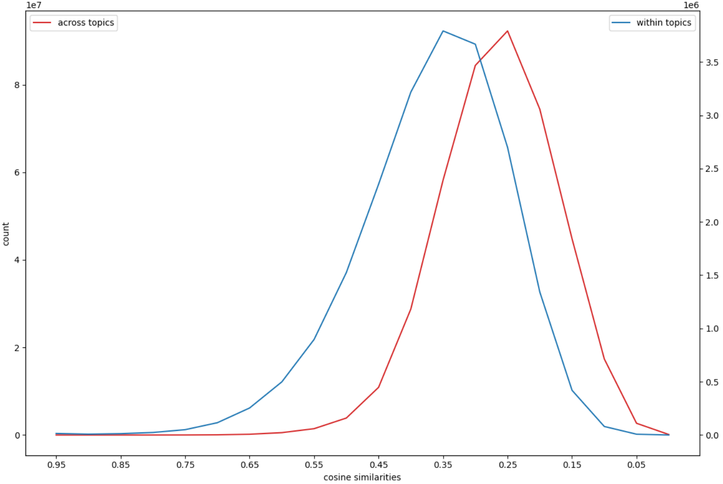
- Consider the Comparative Agendas dataset, which has high quality topic tags about legislative acts
- Off-the-shelf embeddings from Sentence Bert and OpenAI do not do a particularly good job of grouping legislation on different topics together
- However, with limited paired training data, we can contrastively tune a model to map legislation on different topics to different regions of embedding space, and legislation on the same topic nearby
- This model translates well across legislative settings

Off-the-shelf (S-BERT) embedding model and U.S. congressional bills data



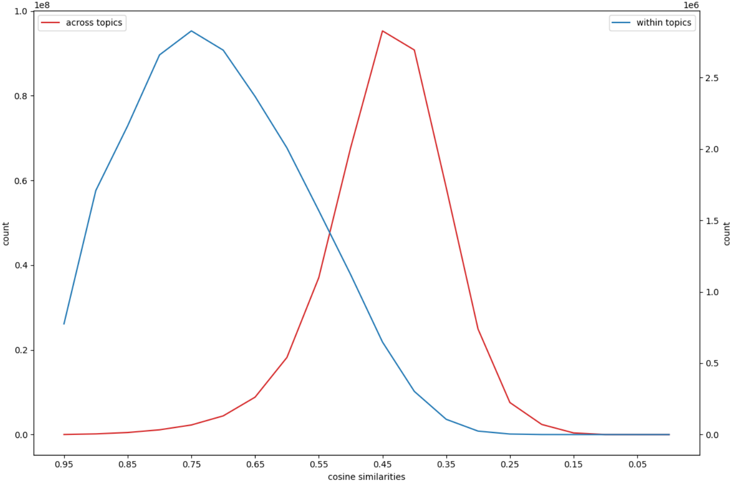
Cosine similarities between the embeddings of legislative texts, within topic (blue) and across topic (red)

Off-the-shelf (OpenAI) embedding model and U.S. congressional bills data



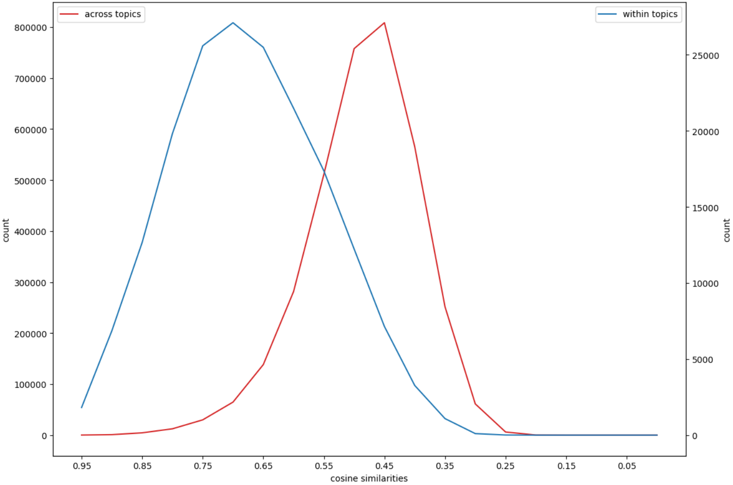
Cosine similarities between the embeddings of legislative texts, within topic (blue) and across topic (red)

Fine-tuned embedding model and U.S. congressional bills data



Cosine similarities between the embeddings of legislative texts, within topic (blue) and across topic (red)

Fine-tuned embedding model and UK Parliamentary Acts data



Cosine similarities between the embeddings of legislative texts, within topic (blue) and across topic (red)

Applications of Embedding Models

- Data exploration (classes unknown ex ante)
  - Reproduction of news articles
  - Biggest stories
  - Most similar news stories to a query
- Many classes
  - Record linkage in structured data
  - Linking individuals mentioned in unstructured texts
- Adding classes at inference time
  - Document transcription

Detecting Reproduced Content with Embeddings

- Detecting noisily reproduced content is important both for first order social science questions and for commercial applications
    - Media historian Julia Guarneri (2017) writes: “by the 1910s and 1920s, most of the articles that Americans read in their local papers had either been bought or sold on the national news market... This constructed a broadly understood American ‘way of life’ that would become a touchstone of U.S. domestic politics and international relations throughout the twentieth century.”
    - Important for de-duplicating training data and controlling test set leakage

Reproduced Content

- We develop novel methods - combining a customized embedding model and single linkage clustering - for detecting noisily reproduced content (Silcock, D'Amico-Wong, Yang, and Dell, 2022)
- We have applied this to detecting noisily reproduced historical news articles on a massive scale (Silcock, Arora, D'Amico-Wong, and Dell, 2024) and to evaluating test set leakage in foundation models (Sainz et al., 2024)

NEWS−COPY

	Neural	**Non-Neural**
Most scalable	Bi-encoder (91.5)	LSH (73.7)
Less scalable	Re-ranking (**93.7**)	*N*-gram overlap (75.0)

Table: The numbers in parentheses are the Adjusted Rand Index for four different models - a bi-encoder, a "re-ranking" strategy that combines a bi- and cross-encoder, locally sensitive hashing (LSH), and *N*-gram overlap. Hyperparameters were chosen on the NEWS−COPY validation set, and all models were evaluated on the NEWS−COPY test set.

Newswire Dataset

Description	Count
Front page articles	137,941,190
Unique articles (including singletons)	99,472,910
Unique articles reproduced $>$ 3 times	2,889,012
Unique wire articles	2,719,607
Total reproductions of wire articles	32,107,676

Table: Counts of articles meeting various criteria in our raw digitized newspaper corpus.

Domestic Datelines

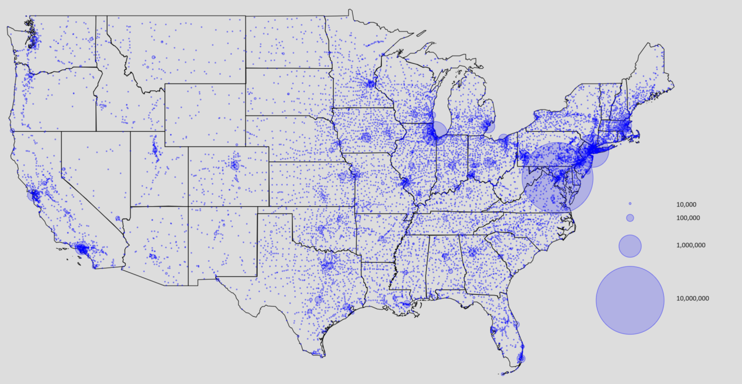


Figure: Reproduction of newswire articles with domestic datelines.

International Datelines

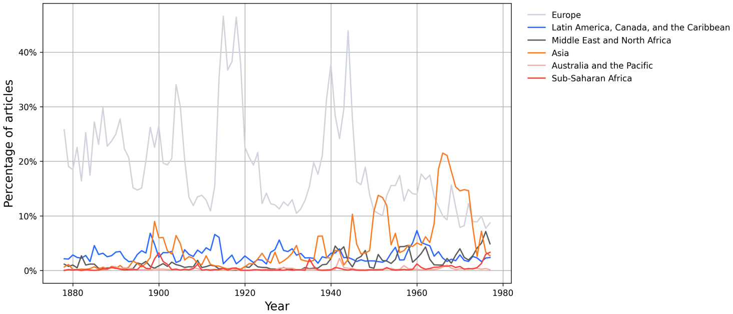


Figure: Reproduction of newswire articles with international datelines.

Newswire Topics

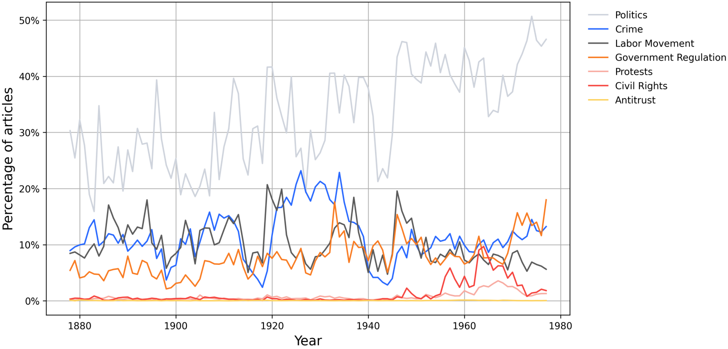


Figure: Share of reproduced newswire articles with a given binary topic tag, across time.

Newswire Topics

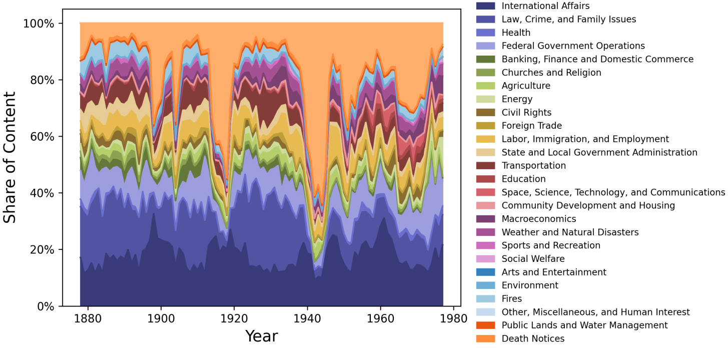
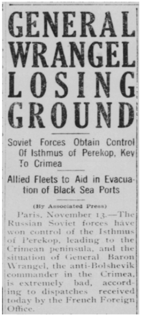


Figure: The distribution of multiclass topic tags, trained on data from the Comparative Agendas project.

Headlines



GENERAL WRANGEL LOSING GROUND

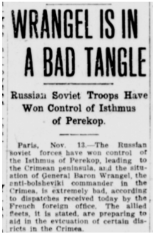
Soviet Forces Obtain Control Of Isthmus of Perekop, Key To Crimea

Allied Fleets to Aid in Evacuation of Black Sea Ports

(By Associated Press)

Paris, November 13.—The Russian Soviet forces have won control of the Isthmus of Perekop, leading to the Crimean peninsula, and the situation of General Baron Wrangel, the anti-Bolshevik commander in the Crimea, is extremely bad, according to dispatches received today by the French Foreign Office.

GENERAL WRANGEL LOSING GROUND Soviet Forces Obtain Contro Of Isthmus of Perekop, Key To Crimea Allied Fleets to Aid in Evacua tion of Black Sea Ports

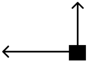


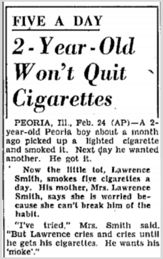
WRANGEL IS IN A BAD TANGLE

Russian Soviet Troops Have Won Control of Isthmus of Perekop.

Paris, Nov. 13.—The Russian soviet forces have won control of the Isthmus of Perekop, leading to the Crimean peninsula, and the situation of General Baron Wrangel, the anti-bolsheviki commander in the Crimea, is extremely bad, according to dispatches received today by the French foreign office. The allied fleets, it is stated, are preparing to aid in the evtcuation of certain disricts in the Crimea.

WRANGEL IS IK A BAD TANGLE
Russisi: Soviet Troops Have Won Control of Isthmus of Perekop.





FIVE A DAY

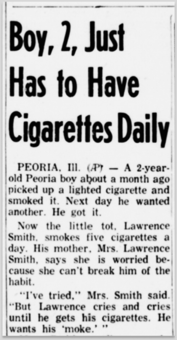
2-Year-Old Won't Quit Cigarettes

PEORIA, Ill., Feb. 24 (AP)—A 2-year-old Peoria boy about a month ago picked up a lighted cigarette and smoked it. Next day he wanted another. He got it.

Now the little tot, Lawrence Smith, smokes five cigarettes a day. His mother, Mrs. Lawrence Smith, says she is worried because she can't break him of the habit.

"I've tried," Mrs. Smith said. "But Lawrence cries and cries until he gets his cigarettes. He wants his 'moke'."

FIVE A DAY 2-Year-Old Won't Quit Cioarettes



Boy, 2, Just Has to Have Cigarettes Daily

PEORIA, Ill. (AP) — A 2-year-old Peoria boy about a month ago picked up a lighted cigarette and smoked it. Next day he wanted another. He got it.

Now the little tot, Lawrence Smith, smokes five cigarettes a day. His mother, Mrs. Lawrence Smith, says she is worried because she can't break him of the habit.

"I've tried," Mrs. Smith said. "But Lawrence cries and cries until he gets his cigarettes. He wants his 'moke.' "

Boy, 2, Just Has to Have Cigarettes Daily

HEADLINES

Decade	Headline Count	Cluster Count	Positive Pair Count	Word Count	Words Per Headline	Line Count	Lines Per Headline	Character Error Rate
1920s	4,889,942	1,032,108	28,928,226	68,486,589	14.0	18,983,014	3.9	4.3%
1930s	5,519,472	1,126,566	37,529,084	75,210,423	13.6	21,905,153	4.0	3.7%
1940s	6,026,940	1,005,342	62,397,004	61,629,003	10.2	19,538,729	3.2	2.4%
1950s	7,530,810	1,192,858	100,527,238	61,127,313	8.1	20,823,786	2.8	2.3%
1960s	6,533,071	926,819	108,415,279	46,640,311	7.1	16,408,148	2.5	3.7%
1970s	3,664,201	585,782	52,981,097	24,472,831	6.7	7,829,510	2.1	3.2%
1980s	703,052	170,507	2,857,722	5,161,537	7.3	1,502,893	2.1	1.5%
Total	**34,867,488**	**6,039,982**	**393,635,650**	**342,728,007**	**9.8**	**106,991,233**	**3.1**	

Table: Descriptive statistics of HEADLINES.

Computational Advantages

	Embed	Compute	Build	Commun.	Total	Mean Times
	Articles	Similarity	Graph	Detect.	Time	Reproduced
Bi-Encoder	8:38:52	3:04:53	0:01:23	0:00:02	11:45:10	6.41
	(GPU)	(GPU)	(GPU)	(GPU)	(GPU)	
Hashing		3:39:05	0:00:55	0:00:08	3:40:08	11.55
		(CPU)	(GPU)	(GPU)	(mostly CPU)	

Table: This table reports computational efficiency in scaling the bi-encoder and LSH methods to a 10 million article corpus. Parentheses indicate whether the calculations were run on a CPU or a single NVIDIA A6000 GPU card. *Mean times reproduced* reports the average size of duplicated article communities that each method estimates.

Biggest Stories of the Year

- We would like to know which news stories receive the most coverage
- We have no idea what those stories are ex ante - cannot use a classifier
- Instead, custom train a large language model to map articles about the same story to similar vector representations using paired data from allsides
- Cluster the resulting embeddings with single linkage clustering

American Stories Dataset

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Total	Text Bounding Boxes				Other Bounding Boxes			
	Boxes	Articles	Headlines	Captions	Bylines	Images	Ads	Tables	Mastheads
Legible	-	335M	368M	9.7M	14.7M	-	-	-	-
Illegible	-	26M	27M	0.9M	2.5M	-	-	-	-
Borderline	-	77M	22M	1.3M	1.2M	-	-	-	-
Total	1.14B	438M	417M	11.9M	18.4M	9.1M	221M	16.3M	4.9M

Table: Dell et al. (2023)

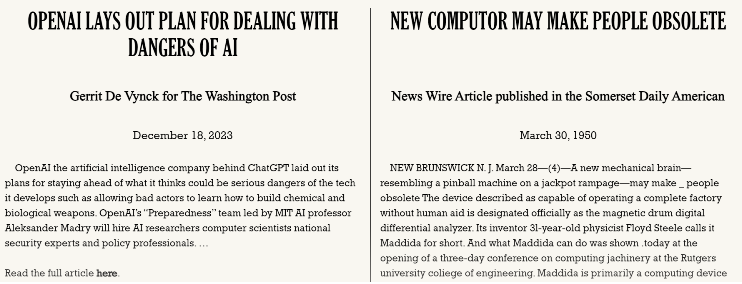
Biggest Stories

Year	Biggest story	Year	Biggest story
1885	Death of General Grant	1903	Panama Canal Treaty
1886	Southwest Railroad Strike	1904	Russo-Japanese War
1887	Vatican supports Knights of Labor	1905	Russo-Japanese Peace Process
1888	Rail strikes	1906	Hepburn Railroad Rate Bill
1889	Samoan Crisis	1907	Mining accidents
1890	1893 World's Fair planning	1908	Taft presidential victory
1891	New Orleans Lynchings	1909	Race to the North Pole
1892	Homestead Steel Strike	1910	Rail strikes
1893	World's Fair, Chicago	1911	Canadian Reciprocity Bill
1894	Wilson–Gorman Tariff Act	1912	Republican National Convention (Taft v Roosevelt)
1895	British occupation of Corinto, Nicaragua	1913	Underwood-Simmons Tariff Act
1896	Bimetallism Movement	1914	World War I
1897	Coal Miners' Strike	1915	World War I
1898	Cuban War of Independence	1916	Pancho Villa Expedition
1899	Philippine-American War	1917	World War I
1900	Anglo-Boer War	1918	World War I
1901	U.S. Steel Recognition Strike	1919	Treaty of Versailles
1902	Anthracite Coal Strike	1920	Rail strikes

Table: Dell et al. (2023)

News Déjà Vu

We use the same model (after masking named entities) to find the most similar historical story to current news stories (Franklin, Silcock, Arora, Bryan, and Dell, 2024).



OPENAI LAYS OUT PLAN FOR DEALING WITH DANGERS OF AI

Gerrit De Vynck for The Washington Post

December 18, 2023

OpenAI the artificial intelligence company behind ChatGPT laid out its plans for staying ahead of what it thinks could be serious dangers of the tech it develops such as allowing bad actors to learn how to build chemical and biological weapons. OpenAI's "Preparedness" team led by MIT AI professor Aleksander Madry will hire AI researchers computer scientists national security experts and policy professionals. ...

Read the full article **here**.

NEW COMPUTOR MAY MAKE PEOPLE OBSOLETE

News Wire Article published in the Somerset Daily American

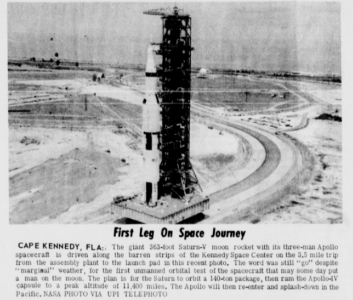
March 30, 1950

NEW BRUNSWICK N. J. March 28—(4)—A new mechanical brain—resembling a pinball machine on a jackpot rampage—may make _ people obsolete The device described as capable of operating a complete factory without human aid is designated officially as the magnetic drum digital differential analyzer. Its inventor 31-year-old physicist Floyd Steele calls it Maddida for short. And what Maddida can do was shown .today at the opening of a three-day conference on computing jachinery at the Rutgers university coliege of engineering. Maddida is primarily a computing device

Most Reproduced Images

- As with text, we can use deep learning to track noisily reproduced images
- Map different versions of the same image to similar vector representations (custom model trained on augmented data) and cluster

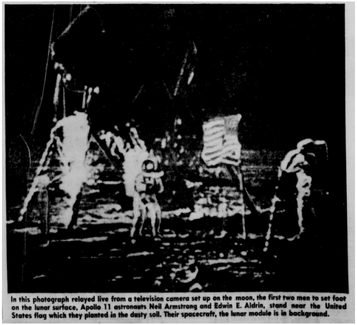
Space Program



First Leg On Space Journey

CAPE KENNEDY, FLA.: The giant 363-foot Saturn-V moon rocket with its three-man Apollo spacecraft is driven along the barren strips of the Kennedy Space Center on the 3,5 mile trip from the assembly plant to the launch pad in this recent photo. The word was still "go" despite "marginal" weather, for the first unmanned orbital test of the spacecraft that may some day put a man on the moon. The plan is for the Saturn to orbit a 140-ton package, then ram the Apollo-IV capsule to a peak altitude of 11,400 miles. The Apollo will then re-enter and splash-down in the Pacific. NASA PHOTO VIA UPI TELEPHOTO

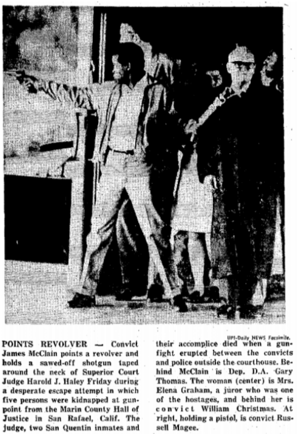
Saturn-V Launching (1967)



In this photograph relayed live from a television camera set up on the moon, the first two men to set foot on the lunar surface, Apollo 11 astronauts Neil Armstrong and Edwin E. Aldrin, stand near the United States flag which they planted in the dusty soil. Their spacecraft, the lunar module is in background.

Apollo 11 Mission (1969)

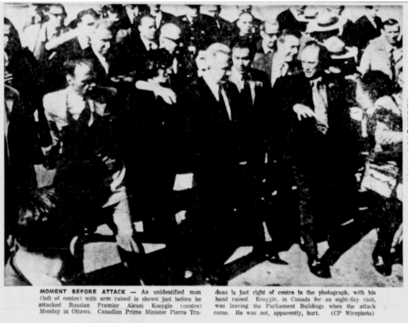
Violence in Progress



UPI-Daily NEWS Facsimile.

POINTS REVOLVER — Convict James McClain points a revolver and holds a sawed-off shotgun taped around the neck of Superior Court Judge Harold J. Haley Friday during a desperate escape attempt in which five persons were kidnapped at gunpoint from the Marin County Hall of Justice in San Rafael, Calif. The judge, two San Quentin inmates and their accomplice died when a gunfight erupted between the convicts and police outside the courthouse. Behind McClain is Dep. D.A. Gary Thomas. The woman (center) is Mrs. Elena Graham, a juror who was one of the hostages, and behind her is convict William Christmas. At right, holding a pistol, is convict Russell Magee.

James McClain (1970)



MOMENT BEFORE ATTACK — An unidentified man (left of centre) with arm raised is shown just before he attacked Russian Premier Alexei Kosygin (centre) Monday in Ottawa. Canadian Prime Minister Pierre Trudeau is just right of centre in the photograph, with his hand raised. Kosygin, in Canada for an eight-day visit, was leaving the Parliament Buildings when the attack came. He was not, apparently, hurt. (CP Wirephoto)

Attack on Russian Premier (1971)

Plane Crashes

(AP Wirephoto)

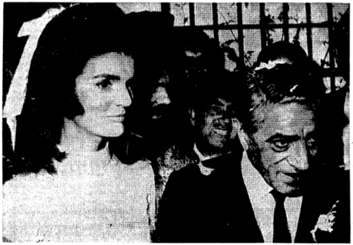
WRECKAGE SPREAD ACROSS FIELD AFTER MIDAIR COLLISION KILLED 83 PERSONS TUESDAY
Wreckage from Allegheny Airlines jet just missed trailer park near Indianapolis

Wreckage of Plane Crash (1969)

SEARCH FOR DEAD—A rescue worker looks for dead and injured after a twin-engine chartered aircraft carrying 42 persons crashed into the Colorado Rockies yesterday. The plane was carrying members of the Wichita State University football team.

Search in the Woods (1970)

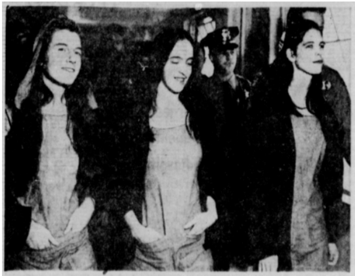
Women in the Press



(Hutchinson News-AP Wirephoto)

Jackie and AristotleArm in Arm During Wedding Sunday

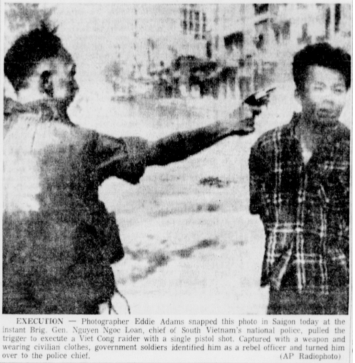
Jackie Kennedy Remarried (1968)



PATRICIA KRENWINKEL, SUSAN ATKINS AND LESLIE VAN HOUTEN FOUND GUILTY

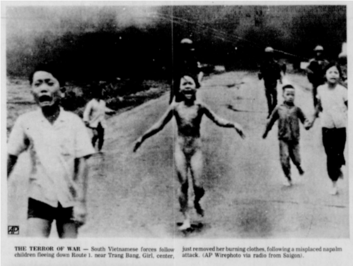
Manson Murderers (1971)

Photos that Changed History?



EXECUTION — Photographer Eddie Adams snapped this photo in Saigon today at the instant Brig. Gen. Nguyen Ngoc Loan, chief of South Vietnam's national police, pulled the trigger to execute a Viet Cong raider with a single pistol shot. Captured with a weapon and wearing civilian clothes, government soldiers identified him as a rebel officer and turned him over to the police chief. (AP Radiophoto)

Saigon Execution (1968)



THE TERROR OF WAR — South Vietnamese forces follow children fleeing down Route 1. near Trang Bang. Girl, center, just removed her burning clothes, following a misplaced napalm attack. (AP Wirephoto via radio from Saigon).

The Napalm Girl (1972)

Photos That Changed History?

- Not very widely reproduced
- No impact on abnormal stock returns of Vietnam War contractors; minimal impact on abnormal returns of Dow Chemicals
- Mentions of the Napalm Girl photo take off after the woman in it started a public advocacy charity
- What we remember about the past is filtered through the present

Record Linkage

- Linking information across diverse sources is fundamental to many pipelines
- For example, researchers and businesses frequently link individuals or firms across censuses and company records or link products/industries across datasets

Traditional Record Linkage

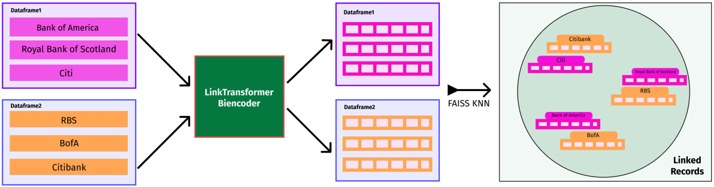
- Conventional string similarity measures rely on edit distance or n-gram overlap to determine how two strings differ - they are unable to bring in semantic information.
- For example, the distance between ABC Co. and ABC Corporation shouldn't be high, as Co. is a common abbreviation of the word Corporation.
- Yet these two string have a high Levenshtein edit distance.

LinkTransformer

A Unified Package for Scalable Record Linkage with Transformer Language Models

- LinkTransformer (Arora and Dell, 2024) brings large language models to data frame manipulation tasks like merges, deduplication, and clustering.
- It supports standard merging, merging with blocking and multiple keys, bypassing translation with cross-lingual merges, aggregation and classification, clustering, and de-duplication.
- It supports models on the Hugging Face Hub and OpenAI Embedding models. We've trained our own collection of over 20 open-source language models for 6 different languages and different tasks.

Embeddings for Record Linkage

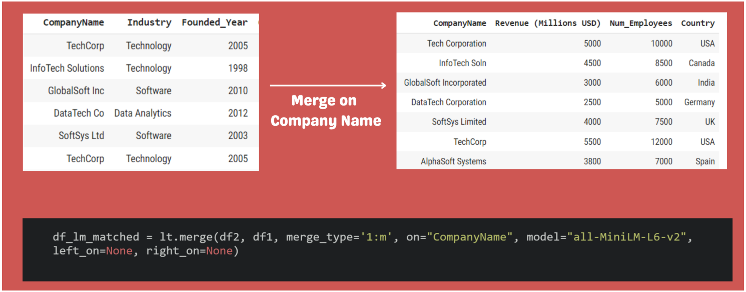


LinkTransformer

A Unified Package for Scalable Record Linkage with Transformer Language Models

- The API is designed to be familiar to practitioners coming from environments like R and Stata.
- Training your own models is as easy as one line of code.
- It also includes a classification module, to facilitate sequence-level text classification

Basic Functionality



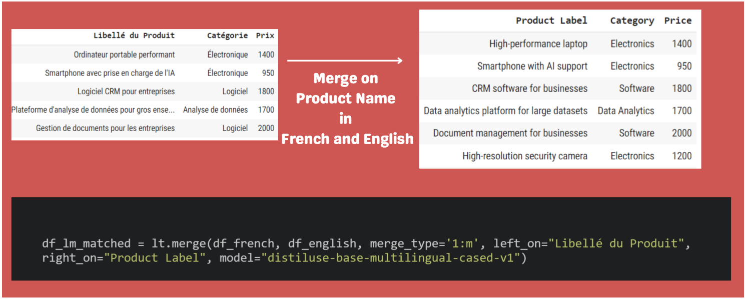
CompanyName	Industry	Founded_Year
TechCorp	Technology	2005
InfoTech Solutions	Technology	1998
GlobalSoft Inc	Software	2010
DataTech Co	Data Analytics	2012
SoftSys Ltd	Software	2003
TechCorp	Technology	2005

Merge on Company Name

CompanyName	Revenue (Millions USD)	Num_Employees	Country
Tech Corporation	5000	10000	USA
InfoTech Soln	4500	8500	Canada
GlobalSoft Incorporated	3000	6000	India
DataTech Corporation	2500	5000	Germany
SoftSys Limited	4000	7500	UK
TechCorp	5500	12000	USA
AlphaSoft Systems	3800	7000	Spain

```
df_lm_matched = lt.merge(df2, df1, merge_type='1:m', on="CompanyName", model="all-MiniLM-L6-v2",
left_on=None, right_on=None)
```

Multilingual Merging



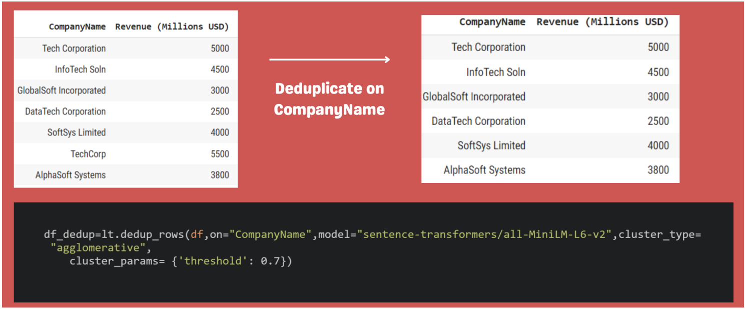
Libellé du Produit	Catégorie	Prix
Ordinateur portable performant	Électronique	1400
Smartphone avec prise en charge de l'IA	Électronique	950
Logiciel CRM pour entreprises	Logiciel	1800
Plateforme d'analyse de données pour gros ense...	Analyse de données	1700
Gestion de documents pour les entreprises	Logiciel	2000

Merge on Product Name in French and English

Product Label	Category	Price
High-performance laptop	Electronics	1400
Smartphone with AI support	Electronics	950
CRM software for businesses	Software	1800
Data analytics platform for large datasets	Data Analytics	1700
Document management for businesses	Software	2000
High-resolution security camera	Electronics	1200

```
df_lm_matched = lt.merge(df_french, df_english, merge_type='1:m', left_on="Libellé du Produit",
right_on="Product Label", model="distiluse-base-multilingual-cased-v1")
```

Deduplication



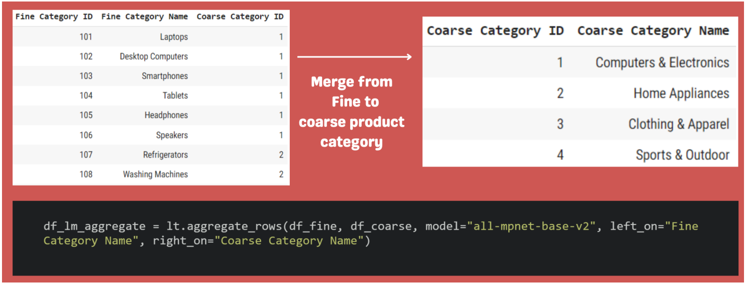
CompanyName	Revenue (Millions USD)
Tech Corporation	5000
InfoTech Soln	4500
GlobalSoft Incorporated	3000
DataTech Corporation	2500
SoftSys Limited	4000
TechCorp	5500
AlphaSoft Systems	3800

→ Deduplicate on CompanyName

CompanyName	Revenue (Millions USD)
Tech Corporation	5000
InfoTech Soln	4500
GlobalSoft Incorporated	3000
DataTech Corporation	2500
SoftSys Limited	4000
AlphaSoft Systems	3800

```
df_dedup=lt.dedup_rows(df,on="CompanyName",model="sentence-transformers/all-MiniLM-L6-v2",cluster_type=
 "agglomerative",
    cluster_params= {'threshold': 0.7})
```

Aggregation



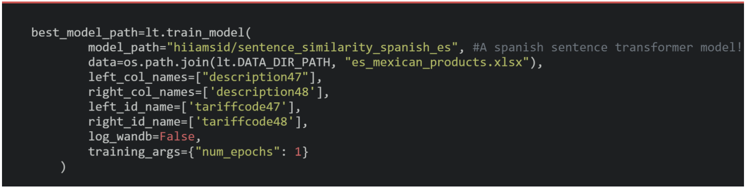
Fine Category ID	Fine Category Name	Coarse Category ID
101	Laptops	1
102	Desktop Computers	1
103	Smartphones	1
104	Tablets	1
105	Headphones	1
106	Speakers	1
107	Refrigerators	2
108	Washing Machines	2

Merge from Fine to coarse product category

Coarse Category ID	Coarse Category Name
1	Computers & Electronics
2	Home Appliances
3	Clothing & Apparel
4	Sports & Outdoor

```
df_lm_aggregate = lt.aggregate_rows(df_fine, df_coarse, model="all-mpnet-base-v2", left_on="Fine
Category Name", right_on="Coarse Category Name")
```

Model Training

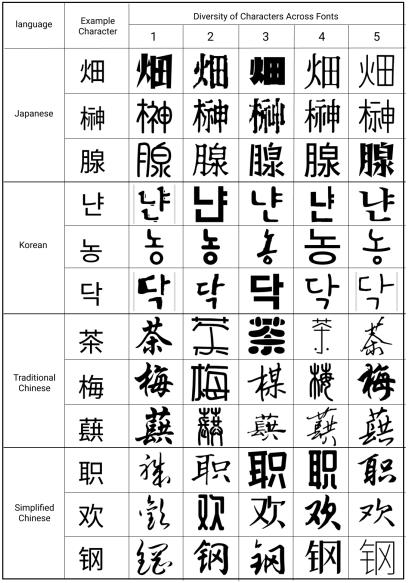


```
best_model_path=lt.train_model(
        model_path="hiiamsid/sentence_similarity_spanish_es", #A spanish sentence transformer model!
        data=os.path.join(lt.DATA_DIR_PATH, "es_mexican_products.xlsx"),
        left_col_names=["description47"],
        right_col_names=['description48'],
        left_id_name=['tariffcode47'],
        right_id_name=['tariffcode48'],
        log_wandb=False,
        training_args={"num_epochs": 1}
    )
```

Visual Record Linkage

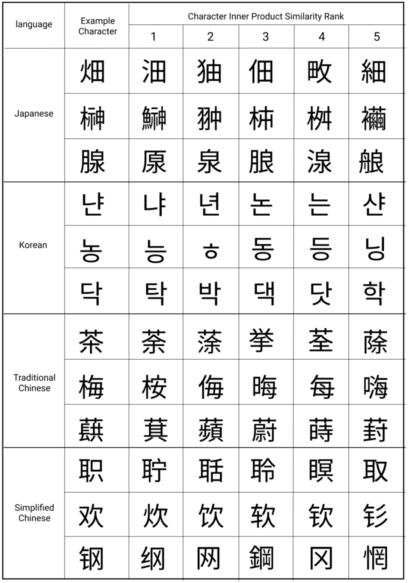
- OCR tends to make errors that are homoglyphic, confusing characters with a similar visual appearances; this can be utilized for record linkage of noisily OCR'ed texts
- Together with Xinmei Yang, Shao-Yu Jheng, and Abhishek Arora, we leverage vision transformers to measure visual similarity *across* characters and used this to improve record linkage for the CJK script
- Our HomoglyphsCJK package is available at `https://pyup.io/packages/pypi/homoglyphscjk/`

Character Variation Across Fonts



language	Example Character	Diversity of Characters Across Fonts				
		1	2	3	4	5
Japanese	畑	畑	畑	畑	畑	畑
	榊	榊	榊	榊	榊	榊
	腺	腺	腺	腺	腺	腺
Korean	냔	냔	냔	냔	냔	냔
	농	농	농	농	농	농
	닥	닥	닥	닥	닥	닥
Traditional Chinese	茶	茶	茶	茶	茶	茶
	梅	梅	梅	梅	梅	梅
	蕻	蕻	蕻	蕻	蕻	蕻
Simplified Chinese	职	职	职	职	职	职
	欢	欢	欢	欢	欢	欢
	钢	钢	钢	钢	钢	钢

Homoglyph Examples



language	Example Character	Character Inner Product Similarity Rank				
		1	2	3	4	5
Japanese	畑	沺	猫	佃	畋	細
	榊	鰰	狆	柿	枡	[illegible]
	腺	原	泉	[illegible]	淥	艆
Korean	냔	냐	년	논	는	샨
	농	능	ㅎ	동	등	닝
	닥	탁	박	댁	닷	학
Traditional Chinese	茶	荼	蒤	拳	荎	蒢
	梅	桉	侮	晦	每	嗨
	蕻	萁	蘋	蔚	蒔	葑
Simplified Chinese	职	眝	聒	聆	瞑	取
	欢	炊	饮	软	钦	钐
	钢	纲	网	鋼	冈	惘

Introduction

Classification

Embedding Models

Classes Unknown Ex Ante

Many Classes

Adding Classes at Inference Time

Regression

Conclusion

Ancient Homoglyphs

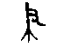
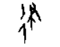
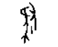
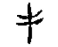
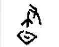
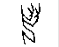
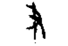
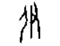
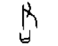
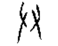
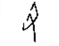
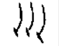
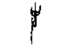
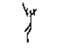
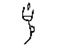
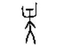
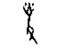
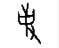
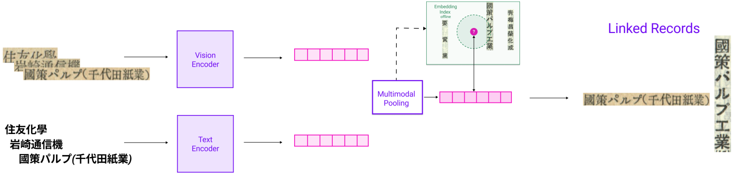
	Example Character	Character Inner Product Similarity Rank				
		1	2	3	4	5
1						
	聿	律	肄	丰	晝	丮
	Writing	Law	Learning	Thriving/ Feudal/ Abundance	Morning	Holding
2						
	尹	伊	君	廾	丑	劦
	Middle or high-level official	Name of prominent high-level official	Middle or high-level official, later referring to lord/monarch	Joining hands together/ arch/ surrounding	A raising hand	Joining/ Helping/ Assisting
3						
	史	臾	克	央	叜	吏
	History/ Historian	Tying up	Conquest	Central (shows prisoner to be executed in city center)	Searching	Governmental Official

Figure: **Ancient Homoglyphs.** This figure shows homoglyph sets constructed for ancient Chinese, with the descendant modern Chinese character and a description of the character's ancient meaning.

Image and Text Embeddings Can Be Combined



住友化學
岩崎通信機
國策パルプ(千代田紙業)

Vision Encoder

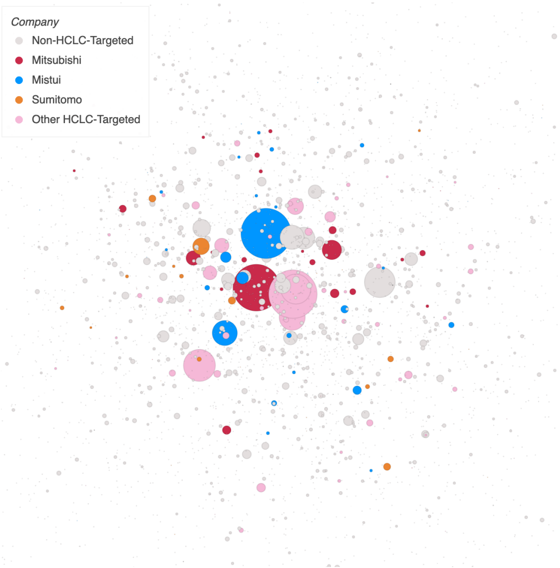
Text Encoder

Multimodal Pooling

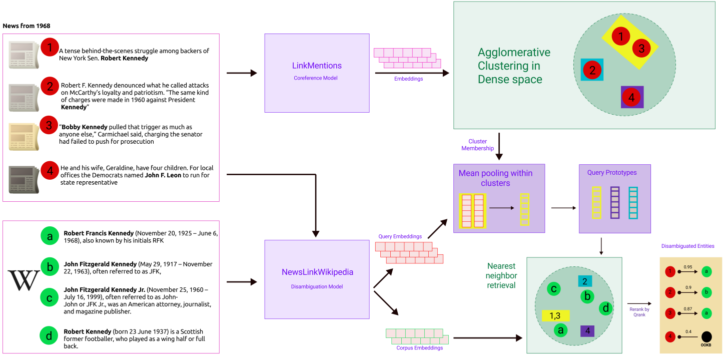
Embedding Index offline

Linked Records

Input-Output Networks



Entity disambiguation with embeddings



News from 1968

1 A tense behind-the-scenes struggle among backers of New York Sen. **Robert Kennedy**

2 Robert F. Kennedy denounced what he called attacks on McCarthy's loyalty and patriotism. "The same kind of charges were made in 1960 against President **Kennedy**"

3 "**Bobby Kennedy** pulled that trigger as much as anyone else," Carmichael said, charging the senator had failed to push for prosecution

4 He and his wife, Geraldine, have four children. For local offices the Democrats named **John F. Leon** to run for state representative

a **Robert Francis Kennedy** (November 20, 1925 – June 6, 1968), also known by his initials RFK

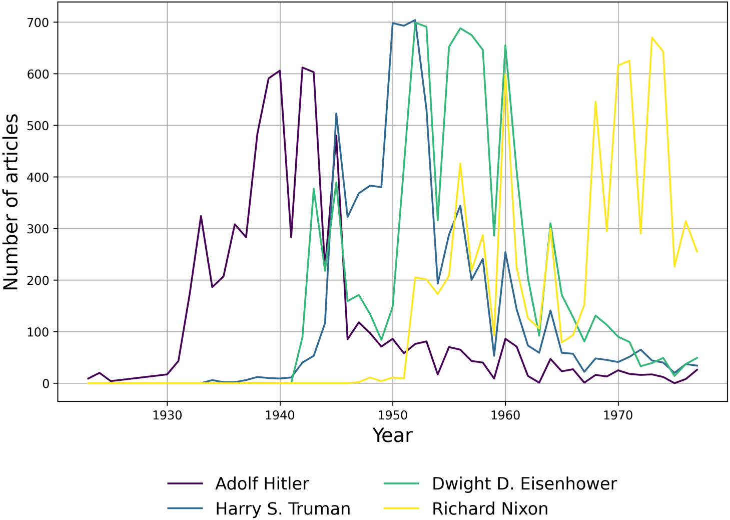
b **John Fitzgerald Kennedy** (May 29, 1917 – November 22, 1963), often referred to as JFK,

c **John Fitzgerald Kennedy Jr.** (November 25, 1960 – July 16, 1999), often referred to as John-John or JFK Jr., was an American attorney, journalist, and magazine publisher.

d **Robert Kennedy** (born 23 June 1937) is a Scottish former footballer, who played as a wing half or full back.

Arora, Silcock, Heldring and Dell (2024).

Entity disambiguation

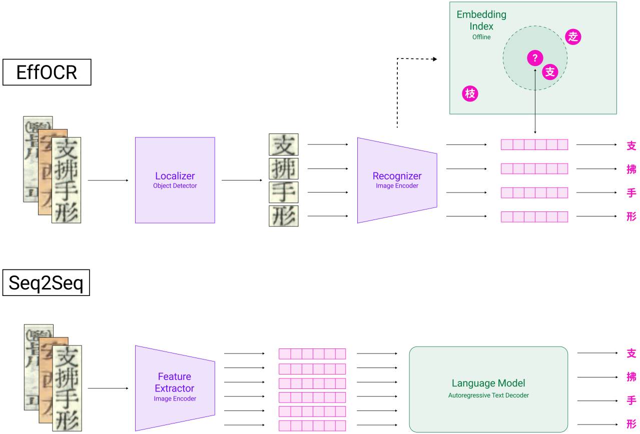


Mentions over time of entities that appeared most commonly in newswire articles.

Adding Classes at Inference Time

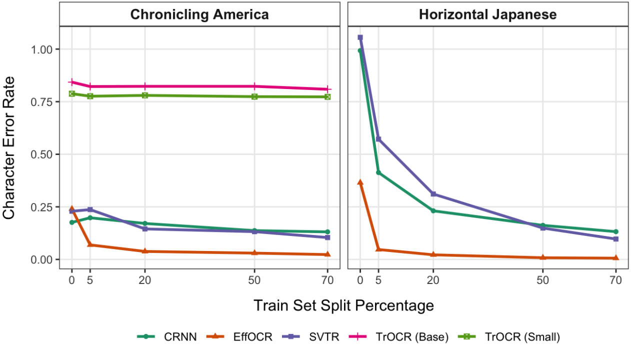
- Embedding models are also well suited to contexts where you may need to add additional classes after training the model, which cannot be done when using a classifier head
- One context where I frequently encounter this is transcribing documents

EffOCR Architecture



Carlson, Bryan, and Dell (2024)

Sample Efficiency



EfficientOCR

- Our EffOCR package (Bryan, Carlson, Arora, and Dell, 2024) makes it straightforward to tune your own custom OCR model, including for very low resource settings
- We have a demo notebook that you can use to train your own OCR model to recognize polytonic (ancient) Greek
- We show that this customized model, which can be trained with free student compute credits, beats Google Cloud Vision, the state-of-the-art for ancient Greek

Outline

Introduction

Classification

Embedding Models

- Classes Unknown Ex Ante
- Many Classes
- Adding Classes at Inference Time

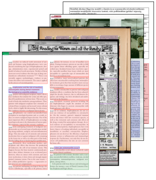
Regression

Conclusion

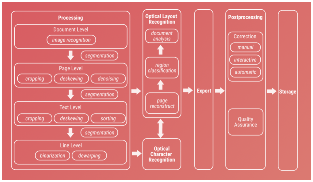
Regression Models

- In machine learning, the term regression refers to the prediction of continuous outcomes
- Regression using deep neural networks is analogous to classification
- Document image analysis is a prime example

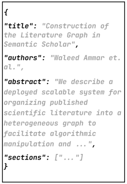
Document Image Analysis



Document Input



DIA Pipeline

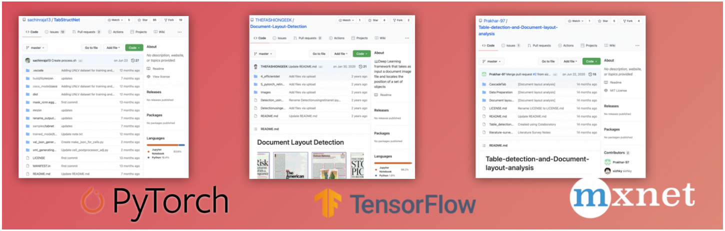


```
{
"title": "Construction of
the Literature Graph in
Semantic Scholar",
"authors": "Waleed Ammar et.
al.",
"abstract": "We describe a
deployed scalable system for
organizing published
scientific literature into a
heterogeneous graph to
facilitate algorithmic
manipulation and ...",
"sections": ["..."]
}
```

Structured Output

Designing DIA Pipelines Can Be Challenging

Previously, there was no full-fledged infrastructure for easily curating document image datasets and fine-tuning or re-training layout analysis models. Relevant resources were in different repos and used inconsistent backends and APIs



sachinraja13 / TabStructNet

sachinraja13 Create process.sh on Jun 23 27

.vscode	Adding UNLV dataset for training and...	12 months ago
build/lib/mrcnn	updates	7 months ago
coco_model/coco	Adding UNLV dataset for training and...	12 months ago
dist	Adding UNLV dataset for training and...	12 months ago
mask_rcnn.egg...	first commit	13 months ago
mrcnn	updates	7 months ago
rename_output...	updates	7 months ago
samples/tabnet	updates	7 months ago
trained_model/t...	Update note.txt	12 months ago
val_json_gener...	Create make_json_for_cells.py	2 months ago
xml_generating...	Update cell_postprocessor_adj.py	2 months ago
LICENSE	first commit	13 months ago
MANIFEST.in	first commit	13 months ago
README.md	Update README.md	7 months ago

About

No description, website, or topics provided.

Releases

No releases published

Packages

No packages published

THEFASHIONGEEK / Document-Layout-Detection

THEFASHIONGEEK Update README.md on Jan 30, 2020 21

4_efficientdet	Add files via upload	2 years ago
5_pytorch_retin...	Add files via upload	2 years ago
images	Add files via upload	2 years ago
Detection_usin...	Rename Detectionusingretinanet.ipy...	2 years ago
Detectionusinge...	Add files via upload	2 years ago
README.md	Update README.md	2 years ago

README.md

Document Layout Detection

About

Deep Learning framework that takes as input a document image file and locates the position of a set of objects

Releases

No releases published

Packages

No packages published

Prakhar-97 / Table-detection-and-Document-layout-analysis

Prakhar-97 Merge pull request #2 from siz... on Jun 22, 2020 15

CascadeTab	[Document layout analysis]	14 months ago
Data Preparation	[Document layout analysis]	14 months ago
Document layou...	[Document layout analysis]	14 months ago
LICENSE.md	Rename LICENSE to LICENSE.md	14 months ago
README.md	Update README.md	14 months ago
Table_detection...	Created using Colaboratory	14 months ago
literature-surve...	Literature Survey Notes	14 months ago

README.md

Table-detection-and-Document-layout-analysis

About

No description, website, or topics provided.

Releases

No releases published

Packages

No packages published

Contributors 2

Prakhar-97

sizhky sizhky

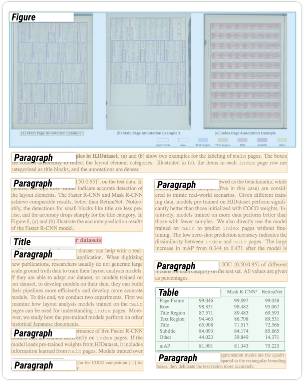
Layout Parser

A Unified Toolkit for Deep Learning Based Document Image Analysis

- I worked with pre-docs Zejiang Shen and Jake Carlson and open-source collaborators to integrate the models and tools we developed into an open-source package called `LayoutParser`
- The aim is to streamline the use of DL in document image analysis (DIA) pipelines
- Layout Parser provides simple and intuitive interfaces for applying and customizing DL models for layout detection and other document processing tasks

Layout Detection

```
>>> model =
    lp.Detectron2LayoutModel()
>>> image = load_image()
>>> layout =
    model.detect(image)
```



(a) Main Page Annotation Example 1

(b) Main Page Annotation Example 2

(c) Index Page Annotation Example

ples in HJDataset. (a) and (b) show two examples for the labeling of main pages. The boxes are colored differently to reflect the layout element categories. Illustrated in (c), the items in each index page row are categorized as title blocks, and the annotations are denser.

).50:0.95][2], on the test data. In general, the high mAP values indicate accurate detection of the layout elements. The Faster R-CNN and Mask R-CNN achieve comparable results, better than RetinaNet. Noticeably, the detections for small blocks like title are less precise, and the accuracy drops sharply for the title category. In Figure 8, (a) and (b) illustrate the accurate prediction results of the Faster R-CNN model.

r datasets

dataset can help with a real- application. When digitizing new publications, researchers usually do not generate large scale ground truth data to train their layout analysis models. If they are able to adapt our dataset, or models trained on our dataset, to develop models on their data, they can build their pipelines more efficiently and develop more accurate models. To this end, we conduct two experiments. First we examine how layout analysis models trained on the main pages can be used for understanding index pages. Moreover, we study how the pre-trained models perform on other historical Japanese documents.

ormance of five Faster R-CNN models that are trained differently on index pages. If the model loads pre-trained weights from HJDataset, it includes information learned from main pages. Models trained over

for the COCO competition [12] for y.

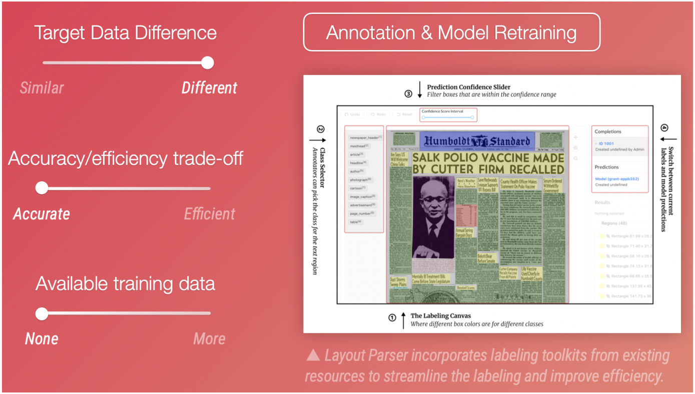
ewed as the benchmarks, while (five in this case) are considered to mimic real-world scenarios. Given different training data, models pre-trained on HJDataset perform significantly better than those initialized with COCO weights. Intuitively, models trained on more data perform better than those with fewer samples. We also directly use the model trained on main to predict index pages without fine-tuning. The low zero-shot prediction accuracy indicates the dissimilarity between index and main pages. The large increase in mAP from 0.344 to 0.471 after the model is

IOU [0.50:0.95] of different models for each category on the test set. All values are given as percentages.

		Mask R-CNN[a]	RetinaNet
Page Frame	99.046	99.097	99.038
Row	98.831	98.482	95.067
Title Region	87.571	89.483	69.593
Text Region	94.463	86.798	89.531
Title	65.908	71.517	72.566
Subtitle	84.093	84.174	85.865
Other	44.023	39.849	14.371
mAP	81.991	81.343	75.223

egmentation masks are the quadri- npared to the rectangular bounding boxes, they delineate the text region more accurately.

Model Customization



Target Data Difference

Similar — ***Different***

Accuracy/efficiency trade-off

Accurate — *Efficient*

Available training data

None — *More*

Annotation & Model Retraining

③ **Prediction Confidence Slider**
Filter boxes that are within the confidence range

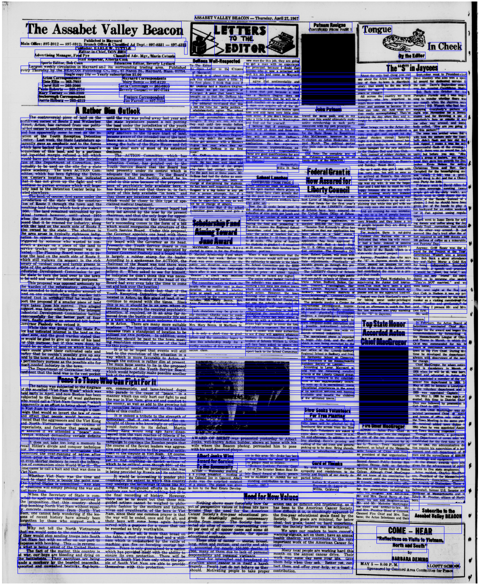
② **Class Selector**
Annotators can pick the class for the text region

④ **Switch between current labels and model predictions**

① **The Labeling Canvas**
Where different box colors are for different classes

▲ *Layout Parser incorporates labeling toolkits from existing resources to streamline the labeling and improve efficiency.*

Off-the-Shelf Digitization of Historical Newspapers Fails



The Assabet Valley Beacon

Published in Maynard
Main Office: 897-2012 — 897-8912; Branch Office & Classified Ad Dept.: 897-8881 — 897-4322
Publisher, EARLE W. TUTTLE
Editor-in-Chief, DON ERIC
Advertising Manager, Fred Pye — Classified Adv. Mgr., Maria Catania
Staff Reporter, Alberta Cook
Sports Editor, Bob Conn — Education Editor, Beverly Lydiard
Largest weekly circulation in Maynard and its surrounding trading area. Published every Thursday by the BEACON Publishing Co., Inc., 47 Nason St., Maynard, Mass. 01754.
Single copy 10¢ — Yearly subscription $5.00

Acton Correspondents	Maynard Correspondents
Elsie Ellis — 263-7932	Mary Bemis — 897-8120
Carol Hill — 263-2066	Lucia Cummings — 263-6955
Paula Roberts — 263-2734	Beverly Gaspari — 897-7744
Mary Tierney — 263-2209	
Boxborough Correspondent	Feature Correspondent
Dorris Hilberg — 263-4319	Jim Farrell — 897-7334

A Rather Dim Outlook

The controversial piece of land on the northwest corner of Route 2 and Wetherbee Street, Acton, has caromed from one political hot corner to another over recent years, and has apparently come to rest as the location of the Youth Reception Detention Center. Last week, the State Legislature apparently gave an emphatic nod to the forces which have backed the youth service board's acquisition of this land, and by a partisan and lop-sided margin refused a bill which would have put the land under the jurisdiction of the Department of Correction, presumably to be used as the site for a state police academy. The town ACTION Committee, which has been fighting the Detention Center's location here, has indicated that it has not given up the fight and will continue to pursue avenues which will hopefully lead to the Detention Center being located elsewhere.

The land, of course, first came under the jurisdiction of the state with the construction of Route 2 through the town and the resulting land-taking which took place about 1950. The land remained untroubled by political turmoil however, until about 1955 when the Acton Planning Board first proposed that it be rezoned for industry, along with the land on the south side of Route 2, also owned by the state. The abutters in the area arose in typically righteous indignation to oppose the proposal, which was triggered by someone who wanted to construct a garage on a piece of the land to service trucks, and the proposal was defeated. The Planning Board went on to rezone the land on the south side of Route 2, which still wallows (in season) in the rich luxury of verdant corn and turnip greens in spite of the arduous efforts of a once active Industrial Development Commission to get the state to turn the land over to the town to be sold and used for industrial purposes.

This proposal was opposed arduously by the warden of the reformatory, although it was amended to include a smaller amount of land on one occasion when the warden intimated (not in writing) that he would support the proposal if a smaller piece of land were taken from his control. This alleged support was later withdrawn. The intrepid Industrial Development Commission battled unsuccessfully for the better part of four years, finally getting an amended bill before Governor Peabody who vetoed it.

While this was going on, the State Police had indicated interest in the land on the other side, and the good warden stated that he would be glad to give up some of his land for this purpose, but if this were done, he would be so short of land on which his prisoners could grow their cultivated oats and barley that he couldn't possibly give up any land to the town of Acton to be used for such a perfunctory purpose as the possible site for much-needed industry in the town.

The Department of Correction felt very confident that the land was in its vest pocket until the rug was pulled away last year and the state legislature passed a bill putting the land under the jurisdiction of the youth service board. When the town, and particularly abutters to the 72-acre tract of land, became aware of the intended use of the land, there was a howl of dismay which resounded among the halls of the State House and fell on the deaf ears of most of its esteemed denizens.

The ACTION Committee, which has fought the proposed use of this land for a Detention Center, has pointed out to the Youth Service Board that the Board has land presently under its control which is adequate for the purpose. To the Board's arguments that the facility should be located in this area because of the preponderance of psychiatric help available here, it has been pointed out that there is, in fact, more of this help available "in town," and that there are locations closer to the "hub," which would be closer to this type of specialized medical treatment.

It is apparent that the present board has been led around by the nose by its present chairman, and that the only hope for opposition to the location of the Detention Center's succeeding comes in the form of a bill which would reorganize the structure of the Youth Service Board. Under this proposal, the Governor appoints one head with three advisors. These people work with an advisory board with the Governor as its head. Presently the Youth Service Board is run through its chairman who was appointed by former Governor Peabody and a board which is largely a rubber stamp for its leader. According to a spokesman for ACTION, the chairman of the Board was told there were houses near the proposed Center and didn't believe it. When asked to see for himself, he indicated he didn't think this was necessary. One would wonder if the head of the Board had ever even take the time to come out and look over the location.

There is little doubt but that if the Youth Reception Detention Center is ever located in Acton, on this piece of land, it will continue to expand with the times. Since this is so, it should be in an area where its inmates may receive the required psychiatric attention, if required, or in an area far removed from the bustle of community life and in a town where it is more welcome than it is in Acton. There are many more suitable locations. The land in question is much too valuable from a standpoint of location to be utilized for this purpose. In addition, some attention should be paid to the town meeting resolution opposing the use of the land for this purpose.

It is to be hoped that events will still lead to the resolution of the situation in a way which is more favorable to Acton, although the outlook is rather dim at present. The main hope seems to lie in the proposed reorganization of the Youth Service Board, which would hopefully make possible another look at the whole matter.

Peace To Those Who Can Fight For It

The nation was subjected to the disgrace of the so-called "Viet Nam Week" which was held early in April, and now Boston has been subjected to the bleating of wool gatherers who would call a "Viet Nam Summer," which apparently is an effort to bring about "peace" in Viet Nam by this summer through a campaign that would so invert the law of cause and effect that people would become convinced that the aggressors and the Viet Kong and North Vietnamese are the victims of imperialists, and further that peace would be assured if we attended a "peace" conference without demanding certain definite concessions from the enemy.

It does not take too long a memory to recall Hitler's divide and conquer technique nor the compromises and concessions that permitted the over-running of nation after nation prior to World War II. It requires an even shorter memory to recall the infiltration of communism since World War II. The time came to call a halt and that was done in Viet Nam.

Whether Viet Nam was the time and place to stand firm is beside the point now. The United States is committed. Any sign of weakness will simply prolong the war and cost more lives.

When the Secretary of State is compelled to spell out the nonsense involved in the proposition that this country should stop bombing North Viet Nam without equally concrete concessions from North Viet Nam, one cannot help wondering if the lessons of recent history have already been forgotten by those who suggest such a course.

Why not tell the North Vietnamese that we would come to the conference table if they would stop sending troops into South Viet Nam but with no offer on our part to dispense with bombing. This — in reverse — is what is being asked of the United States.

The fact of the matter, this country is at war, our boys are bleeding and dying on the battlefields. Their sacrifices are being made a mockery by the bearded peaceniks, assorted and unwashed beatniks, flag-burners, communists, and lame-brained dupes who parade in the cause of "peace," in a manner which can only hurt our fight to end the war in Viet Nam, give aid and comfort to the enemy, as well as add to the growing toll of casualties being recorded on the battlefields of this conflict.

It is indeed a tribute to the strength of democracy that it can withstand the onslaught of those who knowingly or otherwise would contribute to its defeat. Martin Luther King would long ago have disappeared into the vast reaches of Siberia if he, being a Soviet citizen, had launched a similar campaign to convince the Russian people that the course of the Soviet leaders was aggressive and not conducive to the peaceful settlement of the dispute in Viet Nam. Of course, this would be unlikely, since Mr. King has yet to find anything in Soviet actions of which to be critical, even though 80% of the war material needed to perpetuate the war is being provided by Russia and its satellites.

It would be interesting to ponder philosophically the extent to which this country may undergo the batterings of those like Mr. King, whose misguided efforts in the final analysis can but attribute nothing but ill to the final recording of history. However, there can be no doubt but that these misguided actions cannot be weighed in philosophic fashion by the mothers and fathers, wives and sweethearts, of the boys in Viet Nam, since their one common prayer is that the war will be ended, once and for all, and their boys will come home again having served with a purpose for a cause that can be recorded as truly worthwhile.

Peace, after all, is more than bread on the table, a roof over the head and a stillness which is undisturbed by the rattle of machine gun fire or bursting mortar barrages. Peace is only guaranteed to a nation which has provided itself with the ability to ensure its own protection. There will be peace in South Viet Nam only when the people of South Viet Nam are able to provide themselves with this protection.

ASSABET VALLEY BEACON — Thursday, April 27, 1967

LETTERS TO THE EDITOR

DeRosa Well-Respected

To The Editor
An Open Letter to the Citizens of Maynard:

I think it is about time that we pick this situation apart a little. In the first place, I think that even if Mr. DeRosa had a Masters Degree, he would not get the position of Principal. A point that hasn't been brought up is that Mr. DeRosa has enough credit hours for a degree.

The editorial mentioned that this is not the time for "petty politics." I wholeheartedly agree. The school committee are letting "petty politics" and personalities rule their thoughts. If they are not, why did they wait until two weeks before election time to make a ruling that a candidate for the office of principal must have a Masters Degree? Also, why didn't the school committee get someone with a Masters Degree last year when they appointed Mr. DeRosa in the first place?

Mr. DeRosa has been a citizen of Maynard for the past seventeen years. He has been well liked and respected by most of the people who have had contact with him. For the past two or three years, Mr. DeRosa had had the duties as assistant principal and has been in charge of disciplining the students. He has been well respected by them. Respect is a big factor in any job that a person attempts, for, if a man does not have the respect of those he supervises, he may as well not be in charge of others.

We also know that if the school committee goes outside and gets a new man for this job, they are going to get a man with no experience for principal, because I doubt that anyone who is now a principal will quit his job and come to Maynard for less pay.

I agree that sentimentality and friendship do not necessarily qualify a man as best for the job, but I know that just because a man has a Masters Degree does not necessarily mean that he will be a good teacher or a good principal. A man can acquire all of the degrees that education can bestow upon him and still have very little common sense. If you don't believe me, take a little trip down to Washington and find out for yourself.

I believe that if everyone concerned will sit down and think this over in a common sense fashion, they will see that Mr. DeRosa's qualifications overwhelm all other applications. He has experience, common sense, trustworthiness, aggressiveness, and respectability. These are good qualifications for any job that a man may undertake in his lifetime.

Respectfully,
Willis D. Dodd

School Lunches
CONTINUED FROM PAGE 1

school cafeterias must be purchased on the open market where prices on many items have risen greatly during the past two years. In addition, the wages of cafeteria workers have also gone up.

Now the 25 cents being paid by the students plus the government subsidies of nine cents per lunch and 2½ cents per carton of milk are not sufficient to cover the expense of buying the food and preparing it. For this reason the Regional School Committee voted late last year to raise the price of lunches for students in grades 7 - 12 from 25 cents to 30 cents. (The price of lunches for students in grades 1 - 6 remained at 25 cents, since their lunches have smaller portions.)

When it was determined that the price increase would still not enable the cafeterias to operate in the black, the School Committee put the $5000 subsidy item into the budget, feeling that the price of school lunches should not be raised any higher.

The entire school budget (including the subsidy) was approved at town meeting since state law makes towns which cut school budgets liable to court action. However, the town meeting also passed a motion requesting that the School Committee not use the subsidy but make the lunch program self supporting.

The School Committee is now investigating the possibility of doing this. Mrs. Parks has been asked to come up with a schedule of lunch prices high enough to cover the food and cafeteria expenses. She will also have to confer with state authorities to see if they will approve whatever increase is recommended. Superintendent of Schools William L. O'Connell has been asked to obtain this information from Mrs. Parks and report back to the School Committee.

Scholarship Fund Aiming Toward June Award

MAYNARD — Donations have started to come in for the Frediano D. Mattioli Scholarship Fund which was started last week. Before the Fund was even announced to the public, one donation was received. On Saturday, less than forty-eight hours after it was announced in the "Beacon", a check was received from a retired school teacher.

The committee wants to thank the people who so readily sent in their donation. Any donation, whether large or small will be gratefully appreciated by the committee which includes: Mrs. Mary Bemis, Mrs. Anne Francis and Mrs. Shirley Barlow.

Cash donations may be given to any member. Checks may be made payable to the Frediano D. Mattioli Scholarship Fund and sent to Mrs. Mary Bemis, 18 Marlboro St. Maynard.

It is only a short time to graduation and the committee hopes to have this scholarship ready for a worthy member of this year's graduation class.

AWARD OF MERIT was presented yesterday to Albert Jenks, well-known Acton builder, shown at home with his wife where the Beacon photog. persuaded him to pose with his well deserved plaque.

Albert Jenks Wins Award for Service To the Community

ACTON — Wednesday morning at a 9 o'clock breakfast - meeting at the Colonial Inn in Concord, Albert Jenks was the surprised recipient of a plaque. The plaque was given to him by the Real Estate Developers in this area. Mr. Jenks has been in Real Estate for about 25 years. The plaque read as follows:

"Brokers Institute Patriots Council of The Greater Boston Real Estate Board Award of Merit, presented to ALBERT R. JENKS for outstanding contribution to the community, 1967."

Need for New Values

Nothing shows more forcefully how far out of perspective values of human life have become than the need for the American Cancer Society's five-year educational program to drastically reduce unnecessary deaths from cancer. The Society has selected six sites of cancer, representing over 60 per cent of all cancer cases and some 48 per cent of all cancer deaths, for special educational emphasis.

These sites of cancer — uterus, breast, lung, colon and rectum, oral cavity and skin — accounted for nearly 150,000 deaths in 1966, many of them due to lack of personal responsibility and personal vigilance.

Need for intensified and broadened instruction about cancer is in itself a harsh tragedy. People just do not behave as they should. Motivating people to take proper action can be difficult and frustrating, as it has been to the American Cancer Society. How difficult it is, is shockingly apparent in the five-year educational goals fixed by the Society. The goals are minimal, far below ideal, but goals, based on hard experience, that the Society believes can be achieved.

We urge our readers to learn cancer's warning signals, act on them; have an annual health checkup, and contribute to the very important work of the American Cancer Society.

Many local people are working hard this month on the annual cancer drive. Their work may some day save your life. Give them help when they ask. Better yet, contact them and offer your help, or a least a contribution.

Putnam Resigns
CONTINUED FROM PAGE 1

John Putnam

travel the same path, and in my own case this would ultimately make active membership in such party consistent with my own principles.

Putnam was defeated in his bid for the State House by Republican John McGlennon; however he won re-election as Acton Moderator this past March. The political parties play little role in Acton's local elections.

Ed. Note: The complete text of Mr. Putnam's statement is on an inside page of the BEACON.

Federal Grant is Now Assured for Liberty Council

MAYNARD — Superintendent Albert Lerer of Maynard has announced that funds for the first year's operation of the LIBERTY Council of Schools have been assured by the United States Office of Education. In January the Council, in a massive 30-page application, requested $92,900 for five projects.

"Senator Ted Kennedy's office has informed us that our application has been placed on a 'hold' status," Mr. Lerer said. "This means our application has been approved in principle, but that a team from the U. S. Office will discuss certain cost estimates. The Executive Director, Robert Gaines, will meet with this group early next month to negotiate adjustments. Practically all sizeable grants go through this procedure," Mr. Lerer explained.

The LIBERTY Council of Schools is made up of the eighty-eight public and non-public schools of fifteen districts - Acton, Bedford, Bolton, Carlisle, Concord, Framingham, Harvard, Lincoln, Maynard, Stow and Sudbury, the regional districts of Acton-Boxborough, Concord-Carlisle, Lincoln-Sudbury, and Nashoba.

The projects involve a Center for Instructional Development, improved education for mentally, emotionally and physically handicapped children, a comprehensive conservation education program, work toward computer-aided instruction and a possible instructional television network, and closer ties with research oriented industries.

The operational phase is expected to begin July first, and the staff, which is now being recruited by the Director, will operate from [illegible] office of the Lincoln-Sudbury Regional School in Sudbury and from the Hathaway School on Conservation Education of the Massachusetts Audubon Society in Lincoln.

According to Lerer, Maynard, along with the remaining member schools of the Liberty Council group, is looking forward to the establishment of this supplementary centre. As soon as it is operational, many educational experiences in the areas of its specified services will become available and benefit the children of the affiliated towns.

Stow Seeks Volunteers For Tree Planting

STOW — On Saturday, May 6, fifty trees will be planted on the Smithurst Site across from the Stow Town Hall. Volunteers are asked to come down any time from 10 a.m. 'til mid-afternoon. In addition to the tree planting, there's pruning and cutting of the old trees and general picking-up to be done. Bring shovels, shears, or wheelbarrow if possible.

Card of Thanks

We wish to express our deepest sympathy and condolences to the family and friends of our dearly departed father.

Audrey B. (Leard) Southwick
Raymond D. Leard II
Willits, California

Tongue In Cheek
By the Editor

The "S" in Jaycees

About the only bad thing you can say about the Acton Jaycees is that one of their by-laws is discriminatory; that would be the one which limits the age of members to 35 years or younger.

All across the country we have teenagers who are impatient to reach their 16th birthday so they can drive and anxious to become 21 so they can vote, etc. And what are they really accomplishing? They're hurrying the day when they will no longer be eligible for the Acton Jaycees.

What brings me to the Acton Jaycees is that this past weekend a very nice girl and myself hit the rubber chicken circuit. Saturday night it was the Jaycees annual banquet and Sunday it was the dedication banquet for the new additions to the Hudson Rod and Gun Club.

Only instead of rubber chicken, on Saturday night it was Roast Prime Ribs of Beef (delicious), at Russo's and Sunday, Roast Beef.

So, as I said, this girl I was with was very nice, but still as soon as we arrived at Russo's, this Jaycee named Ed came over and assumed she was my wife. It was the happy hour and I told him he must be very happy because she is not my wife. So all evening, every five minutes on the button, he was bringing over someone to introduce to us and explaining that while this girl with me was very nice she was not my wife.

And all evening he was introducing me to a lot of people because the Jaycees are like that; they are very friendly and they know a lot of people because they are doing so many things.

For instance, there is out-going President Jim Walline who is so dedicated that when his wife recently had a baby, instead of passing out cigars he passed out tickets to the Battle of the Bands which is a state-wide Jaycees' project.

Anyway, President Jim who says the "S" in Jaycees stands for sacrifice, passed out plaques for each month of the year to the Jaycee who had contributed the most to a project that month.

July went to Paul Westphalen for supervising the Junior Golf tournament. This must have been a good choice because there was much applause and no one appeared teed-off.

August was Jack MacLeod's month for taking some under-privileged children into Fenway Park. Sacrificing an afternoon to watch the Red Sox last summer was truly standing by the "S" in Jaycees.

September went to President-elect Dave Huntley who also won a special award for dedicated service to the club. Dave's September award was for running Candidates Night.

Bob Beck won October for the Pumpkin Parade and I was beginning to wonder whether there was any month when the Jaycees were inactive. Bill Wissert who had been snapping everybody's picture all night won November for a Sweepstakes and for throwing a pie in somebody's face or maybe it was the other way around, and somebody yelled, "How are you going to get this picture?"

The issue was settled when Bill's wife took the picture only the flash bulb didn't flash and boy did that look rigged. I mean it was the only unsuccessful flash of the evening.

Well, there was no award for January but before you say, "Aha, what a group of loafers," you should know they gave two awards for December. Ernie Nelson for selling Christmas Trees and Rev. Gary Campbell for the housefighting contest (which I felt was a pretty bright thing to do) and for a Christmas Party for kids.

February went to David Allen for an open sewerage meeting (some people get all the dirty work), and March went to Griff Roser for the Battle of the Bands. Instead of a plaque, I feel he should have been given one of the bands to play in his backyard for an entire weekend (or is that against zoning regulations).

April went to Isaac Davis for selling donuts and coffee along the Jay Moody Trail...correction, that should be the other way around. It just seemed like Jay's trail after the effort required to sell 3400 donuts and 100 gallons of coffee on a miserable, wet Patriots' Day.

Anyway, Griff Roser won the award for the Project of the Year, cameraman Wissert won a bronze key, Bob Beck a silver attendance key and Harry Morse a Gold Key for being indispensable. Somebody also made a speech and promised it would be "short, but brief," and finally our new "friend" Ed introduced us to someone else saying "They're NOT married," and we went home.

The next night at the Rod and Gun Club in Hudson it was my privilege to be recognized by the toastmaster as a visiting member of the press. "Ladies and gentlemen," he said, "May I present Mr. and Mrs. Don Eric."

Oh, well.

Top State Honor Accorded Acton Chief MacGregor

Fire Chief MacGregor

ACTON — Acton Fire Chief H. Stuart MacGregor was the recipient of one of the top three municipal employee awards, presented by Allen L. Torrey, Town Manager of Amherst, on behalf of the Massachusetts League of Cities and Towns as president of that organization.

The presentation was made at the Second Annual Public Service Recognition Dinner, held on Saturday evening at the Lyons Hall of Boston College. Although Chief MacGregor did not know of his nomination for the award until the evening before, 64 friends, relatives and town officials from Acton were in attendance for the presentation including his wife, his daughter and son-in-law, Mr. and Mrs. David Barry of Vermont; his son and daughter-in-law, Mr. and Mrs. Malcolm MacGregor; his sister and brother-in-law, Mr. and Mrs. Prentice Blood of Concord; Concord Fire Chief and Mrs. Tombeno; the Acton Selectmen and their spouses.

In February, the Acton Board of Selectmen nominated Chief MacGregor for the award and began the recommendation by forwarding to the Massachusetts League of Cities and Towns in March, in which the Chief was commended for his 34 years of continuous service to the town as the Fire Chief. During this time he developed the department always with observance of the need for change.

Chief MacGregor, for whom retirement is mandatory in March of 1968 when he will be 65, was born, raised and educated in Acton. He was appointed a member of the volunteer fire department in 1924. In May of 1927 he became a Lieutenant in the volunteer department and in 1929 was appointed Captain.

On May 1, 1930 he was again elevated, this time to District Chief. This was followed by promotion to Call Chief in 1933.

In 1936 Chief MacGregor was appointed permanent Chief of Acton and tenure of office was voted by the Town Meeting in the same year. He received added town duties in 1960 when he was appointed Assistant Town Building Inspector.

Through his efforts, a permanent Fire Department of seven men was established for the first time in Acton in 1963. His administrative duties require his supervision of a permanent and call department of 60 men and the surveillance of $500,000 in buildings and equipment.

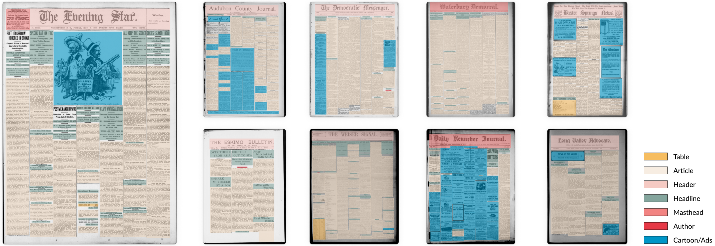
At the dinner, awards were also given for the advancement of public service to five state employees.

Acton resident, Charles MacPherson, President of the Massachusetts Selectmen's Association, presented two Awards of Distinction. These awards were presented to Leverett Saltonstall, former U.S. Senator and Governor of Massachusetts and to Edwin W. Proctor, citizen of Northboro.

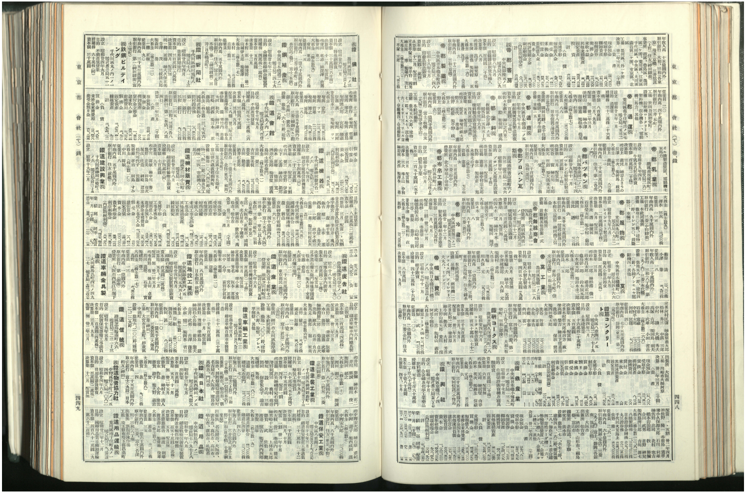
Subscribe to the Assabet Valley BEACON

COME — HEAR
"Reflections on Visits to Vietnam, North and South"
by
BARBARA DEMING
MAY 5 — 8:00 P. M. — ALCOTT SCHOOL
Sponsored by Concord Area Committee for Peace

Layouts


The Evening Star.
Audubon County Journal.
The Democratic Messenger.
Waterbury Democrat.
Baxter Springs News.
THE ESKIMO BULLETIN.
THE WEISER SIGNAL.
Daily Kennebec Journal.
Long Valley Advocate.
Table
Article
Header
Headline
Masthead
Author
Cartoon/Ads

Another Example Document



Layout Analysis

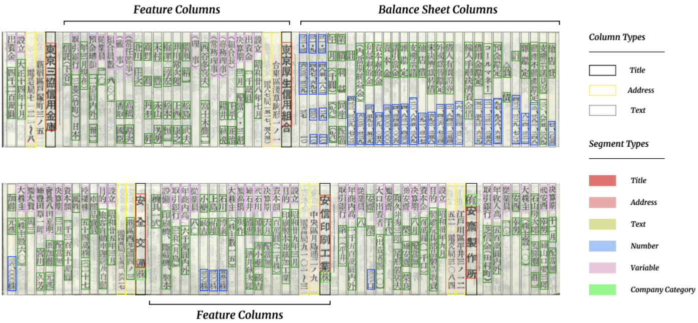


Figure: We train an object detection model (Mask R-CNN) to recognize the document layouts.

Outline

Introduction

Classification

Embedding Models

- Classes Unknown Ex Ante
- Many Classes
- Adding Classes at Inference Time

Regression

Conclusion

Conclusion

- Deep learning provides powerful tools for processing unstructured economic data, creating robust representations for downstream analyses
- Becoming familiar with deep learning methods, how they apply to economics, and how they can be implemented and debugged can entail significant startup costs
- More resources can be found at `https://econdl.github.io/`

A Unifying Framework for Robust and Efficient Inference with Unstructured Data

Jacob Carlson and Melissa Dell

August 2025

Unstructured Data in Economics

- Economists commonly use *unstructured* data—images, text, audio, and video—in empirical research.
- Unstructured data are not used directly in statistical analyses due to their high dimensionality, computational complexity, and lack of interpretability.
- Instead, researchers extract low-dimensional, *structured* features from unstructured data.

Unstructured Data in Economics

- Structured data on governance, institutions, political stability, policy uncertainty, conflict, and violence are extracted from news and other text sources.
- Researchers derive sentiment, topics, and other structured features from government transcripts, corporate filings, earnings calls, patents, and web texts.
- Nighttime satellite imagery measures economic activity, development, and urbanization.
- Remote sensing data supplement sparse ground measurements of temperature, precipitation, pollution, agriculture, land use, illicit activities, and deforestation.

Neural Networks and Unstructured Data

- Extracting structured data from unstructured sources traditionally required costly manual annotation or complex human-engineered rules.
- Large-scale initiatives were often necessary to generate such data.
- Advances in computing and deep learning have drastically reduced these costs.
- Deep neural networks are state-of-the-art for large-scale feature extraction (Goodfellow, 2016) and are widely used by individual researchers to create data.

Biased Predictions

- However, neural networks will not generically produce unbiased predictions in finite samples.
- Choices related to network architecture, the distribution of training data, and various implementation details can all introduce systematic biases.
- Moreover, the use of nonlinear transformations at each layer of the neural network and the frequent application of neural networks to binary or multiclass classification problems violate classical measurement error assumptions.

Biased Predictions

- Biases propagate to estimators that rely on these predictions, affecting both point estimates and uncertainty quantification.
- In large datasets, sampling variation is small but a poorly performing first-step predictor can introduce substantial uncertainty.
- Concerns about imputation bias are further heightened by the availability of off-the-shelf neural networks.
- Different neural networks may introduce different biases that then propagate, raising concerns that neural network-based imputations could be selectively used to produce desirable results.

Costly Investments to Improve Neural Networks

- A frequent question that arises is how accurate neural network predictions need to be for economic research.
- Improving neural network performance often involves significant costs, including training larger models, collecting more or higher-quality training data, and refining complex implementation details.
- To ensure unbiased estimates and determine whether costly investments to improve first step predictions are necessary, researchers need a framework that explicitly accounts for first-step imputation error.

Missing at Random Structured Data

- To address these challenges, we develop MAR-S (***M**issing **A**t **R**andom **S**tructured Data*).
- MAR-S provides a framework for conducting valid, efficient, and robust inference on estimands that incorporate unstructured data through their low-dimensional features.
- MAR-S frames inference with unstructured data as a missing data problem, because raw unstructured datasets commonly lack the low-dimensional summaries that are relevant to economic analyses.
- The framework builds upon the Rubin (1976) missing at random (MAR) mechanism.

MAR-S

- MAR-S leverages foundational results from semiparametric inference with missing data, which provides a well-established, broadly applicable, and assumption-light approach to debiasing estimators.
- The key idea is to collect a validation sample containing ground truth feature values, use this sample to estimate the bias in the imputed data, and adjust estimates accordingly.

Ground Truth

- Ground truth is obtained through a costly process such as annotation by highly skilled and motivated human experts or the collection of ground station data.
- While neural networks are treated as a black box, their outputs must be interpretable through establishing a clear, implementable procedure for measuring the features to be extracted from unstructured inputs.
- The validation sample must meet the ‘missing at random’ assumption: after adjusting for observables, annotated and unannotated structured data should be comparable in their ground truth values.
- This parallels the ‘selection on observables’ assumption in causal inference.

Why is “Missing at Random” so useful?

- By imposing missing at random on annotation, MAR-S minimizes restrictions on the imputation function.
- This is particularly valuable when working with deep neural networks whose biases shift in a complex way with input distributions and implementation details.
- The missing-at-random restriction is often feasible in a world where researchers are increasingly creating their own data.

Systematically missing data can be the motivation for imputation

- Departures from missing at random could be addressed with additional restrictions on the imputation function—such as assuming a stable relationship between imputed and ground truth data across annotated and unannotated samples. (Rambachan et al., 2024)
- A fundamental principle in machine learning is that predictive accuracy deteriorates when there is domain (covariate) shift away from the training data distribution (Ben-David et al. (2010)).
- Neural networks are typically trained on a subset of annotated data or - in the case of off-the-shelf models - on easily accessible data that often differs from observations where ground truth is costly to obtain.
- MAR-S directly addresses this fundamental limitation of neural networks through the missing at random requirement.

Contributions

Debiasing with a ground truth annotation sample is central to recent influential frameworks for valid statistical inference using predictions from black box AI models (Angelopoulos et. al, 2023,Egami et. al, 2023,Ludwig et. al, 2024). The contribution of MAR-S to this literature is threefold:

1. Developing a theoretical framework that unifies this work and links it to a variety of much older, familiar problems
2. Identifying estimators that are both unbiased and efficient
3. Making it feasible to apply debiasing to a wide variety of settings by deriving estimators for common scenarios that have received little attention in the existing literature

Theoretical Insights

- MAR-S provides a theoretical framework that unifies recent work on inference with black-box AI models—developed independently across disciplines with limited interactions—and connects this work to:
    - An econometrics literature on measurement error (Schennach (2016); Chen et al. (2005, 2008))
    - Widely used inference methods that incorporate machine learning-based first steps (e.g., Chernozhukov et al. (2018, 2022a,b))
    - Classical literatures on missing data and causal inference (.g., Rubin (1978); Imbens and Rubin (2015); Robins et al. (1994)).
- MAR-S improves our understanding of inference with unstructured data by connecting it to familiar problems, such as causal inference.
- Viewing inference with unstructured data and causal inference as special cases of a more general missing data problem highlights various relevant insights from semiparametric inference.

Efficiency

- We depart from existing work on black box AI by emphasizing a semiparametric approach, which provides new insights about efficiency.
- For an estimator to achieve asymptotic efficiency, the imputation of missing structured data should depend not only on unstructured data (e.g., texts or images) but also on context-specific structured variables that help estimate the target parameter (e.g., other covariates in a regression model).
- Some existing work (Angelopoulos et al., 2024) claims semiparametric methods are too complicated to be practical; unlikely to be true in economics.

Unbiased and Efficient Inference in Practice

- We derive robust and efficient estimators for descriptive moments, linear regression model coefficients, treatment effects identified through linear IV models, modern difference-in-differences estimands, and regression discontinuity estimands under local randomization.
- Because it employs a semiparametric approach, MAR-S could also be integrated with recent advances in automatic debiasing (Chernozhukov et al., 2022b; van der Laan et al., 2025) or automatic (functional) differentiation (Luedtke, 2024) to further extend its applicability.

MAR-S Extensions

- Existing de-biasing approaches assume that ground truth data are available at the same level of aggregation as the parameter of interest.
- However, economic analyses often require aggregating structured data, and non-linear aggregations are common.
- Collecting ground truth data at this aggregate level would often require labeling thousands or even millions of unstructured data instances just to obtain a single ground truth observation.
- To address this challenge, we develop approaches that leverage MAR-S for debiased inference when ground truth data are available only at a disaggregated level, while the parameter of interest is derived from structured data that are aggregated and (potentially) transformed.

MAR-S Extensions

- When working with large datasets, the structured data of interest often represent a “rare event.”
- The typical approach to this challenge, which arises frequently in empirical economics, is to only annotate data that meet some filter.
- Unless the population of interest is only instances that meet the filter, this violates the assumption that the annotation function not place zero probability on annotating certain types of observations.
- Equivalent to a strong overlap assumption in causal inference.
- The rare event estimation literature suggests optimizing the annotation function to reduce variance, e.g., by incorporating importance sampling techniques.
- We refer readers to an ML literature that develops this approach to annotation (Zrnic and Candes, 2024).

Outline

Introduction

Literature

The MAR-S Framework

Applications

Extensions

Empirical Examples

Conclusion

Literature

- **Missing data**: We build upon foundational work by Rubin (1976); Little and Rubin (2019); Robins et al. (1994, 1995); Robins and Rotnitzky (1995); Bang and Robins (2005)
- **Semiparametric inference**: (Pfanzagl (1982); Bickel et al. (1998); Kennedy (2016, 2018)) Semi-parametric inference is well-suited for missing data because it relies on relatively mild assumptions about the DGP (Tsiatis, 2006). These frameworks are supported by theories of minimax-style efficiency, which provide a benchmark for comparing the performance of different estimators (Newey, 1994; van der Vaart, 1998)
- **Black-box debiased AI:** (Angelopoulos et al., 2023, 2024; Zrnic and Candès, 2024; Zrnic and Candes, 2024, Egami et al., 2023, 2024, List et al., 2024, Ludwig et al. (2024)) Similarly combines black box AI predictions with ground truth values to construct unbiased estimators. This literature does not take a semiparametric approach or consider efficiency. We also cover many additional common empirical settings.

Literature

- **Measurement error literature:** Highlights that validation samples provide a general, model-agnostic method for correcting non-classical measurement error in nonlinear models (Chen et al., 2005).
- **Debiased machine learning:** (Chernozhukov et al., 2018, 2022a,b) MAR-S is based on the same fundamental semiparametric analysis as DML. Both lead to AIPW estimators, derived in different ways.
- **Causal inference and missing data:** (Little and Rubin, 2019; Ding and Li, 2018; Hirano et al., 2003; Imbens and Rubin, 2015; Athey et al., 2019) Causal inference is itself a missing data problem, leading to many parallels, particularly through the AIPW estimator, which is doubly robust (Robins et al., 1994; Robins and Rotnitzky, 1995; Scharfstein et al., 1999). Double robustness relaxes rate requirements on the estimation of nuisance parameters. Double robustness is a crucial property of MAR-S.

Outline

Introduction

Literature

The MAR-S Framework

Applications

Extensions

Empirical Examples

Conclusion

Missing At Random Structured data

- To perform robust and efficient inference with *unstructured data*, we recast the problem as inference on *missing structured data*.
- Structured data, $M \in \mathcal{M}$, are low-dimensional data that can be used directly in estimating equations.
- Unstructured data, $U \in \mathcal{U}$, are high-dimensional and unsuitable for direct use in estimation

Missing At Random Structured data

- Structured data are observed through “annotation”.
- Because accurate annotation is too expensive to scale, we learn a function to impute missing structured data.
- This allows the researcher to leverage the full unstructured dataset, often orders of magnitude larger than the annotated data.
- Deep neural networks increasingly serve as this imputation function $\hat{\mu}$.

Potential Outcomes for Missing Data

MAR-S is closely linked to the Rubin Causal Model (Neyman, 1923; Rubin, 1974, 1978; Imbens and Rubin, 2015), and hence we incorporate the notation of potential outcomes into MAR-S

- We observe a random variable $M \in \mathcal{M} \subseteq \mathbb{R}$ that is subject to some data missingness, given by indicator $A \in \{0, 1\}$,

$$M = A\,M^{a=1} + (1 - A)\,M^{a=0}$$

- We set $M(a = 0) = 0$ w.p.1 WLOG, and, for notational convenience, define $M^* := M(a = 1)$, and so

$$M = AM^*.$$

- We call M^* the "ground truth" potential outcome.
- We call A the "annotation indicator."
- We also assume $P(A = 1) \gg 0$, which means that we have "annotated" some non-negligible count of our unstructured data.

Assumptions

Assumption 1 (Consistency of potential outcomes)

For ground truth potential outcome of interest $M^* \in \mathcal{M}$, observed outcome of interest $M \in \mathcal{M} \times \{0\}$, and annotation indicator $A \in \{0, 1\}$,

$$M = AM^*.$$

- Annotation status needs to be well-defined, which will tend to hold trivially.
- The ground truth label for any given instance depends only on its own annotation status, not on the annotation status of other instances.

Assumptions

Assumption 2 (Missing at random structured data)

For ground truth potential outcome $M^* \in \mathcal{M}$, annotation indicator $A \in \{0, 1\}$, observed covariates $X \in \mathcal{X}$, and unstructured data $U \in \mathcal{U}$:

$$[(U, M^*) \perp\!\!\!\perp A] \mid X.$$

- After adjusting for observables X, annotated and unannotated structured data are comparable in their ground truth values.
- This is analogous to “selection on observables” in causal inference.
- If data are annotated at random, $(U, M^*) \perp\!\!\!\perp A$, the assumption is guaranteed to hold.

Assumptions

Assumption 3 (Known, bounded annotation score function)

We define the annotation score function to be

$$\pi(x) := P(A = 1 \mid X = x) \tag{1}$$

We assume that $\pi(x)$ is fixed, known, and bounded away from zero and one, i.e., $\pi(x) \in [\eta, 1 - \eta]$ for $0 < \eta \leq 1 - \eta < 1$ and for all $x \in \mathcal{X}$.

- This embeds the assumption of “strong overlap” often seen in observational causal inference settings.
- The naming convention “annotation score function” mimics the usual causal inference terminology of a “propensity score function.”
- The researcher often decides how to annotate unstructured data to produce ground truth instances of structured data, making this assumption plausible, but it can be relaxed.

Assumptions

- In observational causal inference, a strong overlap assumption becomes less plausible as the dimension of the variable granting unconfoundedness grows (D'Amour et al. 2021).
- Typically, X is fairly low-dimensional, unlike U, which is high-dimensional.
- However, X may be a low-dimensional measure of U.

Keywords and the Annotation Function

- Social scientists often use keyword-based filtering to annotate text data, assigning some probability of annotation to texts with specific keywords while excluding all others from the annotated sample.
- This violates strong overlap by assigning zero probability to some unstructured data instances.
- Intuitively, we would expect the measurement error of a language model to be systematically correlated with the terms that appear in the text.

Assumptions

Assumption 4 (MSE consistency of the imputation function)

For function $\mu(\tilde{x})$, imputation function $\hat{\mu}(\tilde{x})$, and features $\tilde{X} \in \tilde{\mathcal{X}}$, we have that

$$E\left[\left(\hat{\mu}(\tilde{X}) - \mu(\tilde{X})\right)^2\right] = o(1).$$

- Intuitively, this condition says that we need the expected square error of our estimator to go to zero as the amount of data we train the estimator with goes to infinity; in other words, the estimator is well-specified.
- Very mild in the context of deep neural networks. Recent theoretical work has shown that certain classes of deep neural networks learned with gradient descent are universally consistent (Drews and Kohler, 2024).

If the annotation score function is estimated

- We lose some robustness in our estimators, and the researcher now needs to assume that the convergence of the imputation and estimated annotation score functions occurs at a sufficiently fast rate (Chernozhukov et al., 2018; Kennedy, 2023; Wager, 2024), e.g., at order $n^{-1/4}$ rates.
- While classic semiparametric theory yields "curse of dimensionality" results for learning high-dimensional (nonparametric) conditional expectation functions, in practice, nonparametric regression using deep neural networks appears to exhibit fast rates of convergence (Klaassen et al., 2024); recent theory reveals that neural networks can be especially well suited for estimating nonparametric first-steps/nuisances for treatment effects under selection on a diverging number of confounders (Chen et. al, 2024).

Efficient Inference with MAR-S

- A key advantage of the MAR-S framework is its ability to enable valid, efficient, and robust inference in unstructured data settings.
- A semiparametrically efficient estimator of a functional achieves the lowest possible asymptotic variance among all regular, asymptotically linear (RAL) estimators of that functional (van der Vaart, 1998).
- This lower bound corresponds to the variance of the efficient influence function (EIF) associated with the functional.
- Intuitively, an influence function captures how a small change in the data distribution impacts the value of a functional.

Robust Inference with MAR-S

- Robust estimation, in the context of semiparametric inference, focuses on constructing estimators that relax the rate requirements for first-step parameter estimates in multi-step procedures.
- The first-step parameter estimators are not directly used to estimate the primary parameter of interest but are essential for constructing the final estimator.
- Intuitively, robust estimators are designed to tolerate trade-offs in estimation error among their first-step components while maintaining asymptotic normality and $\sqrt{n}$-consistency.
- In such estimators, worse performance in one component can be offset by better performance in another.

Double Robustness

- An estimator is doubly robust when there are two nuisance functions to be estimated, and the estimator can balance errors between them without losing $\sqrt{n}$-consistency.
- Like other AIPW estimators, MAR-S is (weakly) doubly robust.
- The first-step estimator $\hat{\mu}$ (e.g., a deep neural network) is subject to weak conditions because we have access to the most accurate possible first-step estimator for the annotation function π, which is π itself.

Implementing MAR-S

1. **Identification:** A researcher starts with a target functional $\theta : \mathcal{P} \to \mathbb{R}$. Assumptions 1 and 2 will allow the researcher to recover point identification for their target functional under missing structured data.
2. **Deriving the efficient influence function:** If the point identified target functional is pathwise differentiable, then it has a unique efficient influence function (EIF), computed following Kennedy (2023).
3. **Constructing the robust and efficient estimator:** The researcher may follow one of several procedures for forming a robust, efficient estimator: adding a "one-step correction" to a plug-in estimator based on the EIF; solving an "estimating equation" based on the EIF; or pursuing a targeted maximum likelihood estimation procedure.
4. **Sample splitting for estimation:** Implement estimation via data splitting (or cross-fitting).

Implementing MAR-S

1. **Identification:** A researcher starts with a target functional $\theta : \mathcal{P} \to \mathbb{R}$. Assumptions 1 and 2 will allow the researcher to recover point identification for their target functional under missing structured data.
2. **Deriving the efficient influence function:** If the point identified target functional is pathwise differentiable, then it has a unique efficient influence function (EIF), computed following Kennedy (2023).
3. **Constructing the robust and efficient estimator:** The researcher may follow one of several procedures for forming a robust, efficient estimator: adding a "one-step correction" to a plug-in estimator based on the EIF; solving an "estimating equation" based on the EIF; or pursuing a targeted maximum likelihood estimation procedure.
4. **Sample splitting for estimation:** Implement estimation via data splitting (or cross-fitting).

Implementing MAR-S

1. **Identification:** A researcher starts with a target functional $\theta : \mathcal{P} \rightarrow \mathbb{R}$. Assumptions 1 and 2 will allow the researcher to recover point identification for their target functional under missing structured data.
2. **Deriving the efficient influence function:** If the point identified target functional is pathwise differentiable, then it has a unique efficient influence function (EIF), computed following Kennedy (2023).
3. **Constructing the robust and efficient estimator:** The researcher may follow one of several procedures for forming a robust, efficient estimator: adding a "one-step correction" to a plug-in estimator based on the EIF; solving an "estimating equation" based on the EIF; or pursuing a targeted maximum likelihood estimation procedure.
4. **Sample splitting for estimation:** Implement estimation via data splitting (or cross-fitting).

Implementing MAR-S

1. **Identification:** A researcher starts with a target functional $\theta : \mathcal{P} \to \mathbb{R}$. Assumptions 1 and 2 will allow the researcher to recover point identification for their target functional under missing structured data.
2. **Deriving the efficient influence function:** If the point identified target functional is pathwise differentiable, then it has a unique efficient influence function (EIF), computed following Kennedy (2023).
3. **Constructing the robust and efficient estimator:** The researcher may follow one of several procedures for forming a robust, efficient estimator: adding a “one-step correction” to a plug-in estimator based on the EIF; solving an “estimating equation” based on the EIF; or pursuing a targeted maximum likelihood estimation procedure.
4. **Sample splitting for estimation:** Implement estimation via data splitting (or cross-fitting).

Outline

Introduction

Literature

The MAR-S Framework

Applications

Extensions

Empirical Examples

Conclusion

Applications

- We illustrate the MAR-S framework in five empirical settings of particular interest to economists: descriptive moments, linear regression, linear instrumental variables (IV) models, difference-in-differences (DiD) designs, and regression discontinuity designs (RDD) under local randomization, and then turn to common challenges such as aggregation of missing structured data.
- We develop each example by assigning a single variable to be M (missing structured data) (*e.g.*, an outcome or treatment), although MAR-S can be equally applied to settings where alternative - or multiple - variables are imputed from unstructured data.
- MAR-S is limited to pathwise differentiable functionals, in the sense that $\sqrt{n}$-consistent estimators are not guaranteed to exist for nonpathwise differentiable functionals, and as such efficiency would be ill defined.

MAR-S Mean Functional

Our core object of interest is the MAR-S mean functional

Definition 1

We define a "MAR-S mean functional" as any functional that can be written as

$$\theta(P) = E_P[\tilde{M}^*]$$

where $\tilde{M}^* = g(M^*, V)$ for a known deterministic function g and known random variable V (which is not itself a function of $\pi(X)$) with $[V \perp\!\!\!\perp A] \mid X$. g is homogeneous of degree one in its first argument.

Many functionals of missing structured data—including all the functionals of missing structured data considered in this paper—can be written as MAR-S mean functionals.

MAR-S Mean Functional

We begin by stating the following lemma, which greatly simplifies the derivations of efficient influence functions for the applications considered.

Lemma 2

The efficient influence function for a point identified MAR-S mean functional $E_P[\tilde{\mu}(\tilde{X})]$ *under a nonparametric statistical model* $\mathcal{P} \ni P$ *is the same as the efficient influence function of* $E_P[\tilde{\mu}(\tilde{X})]$ *under the semiparametric statistical model* $\mathcal{P}_\pi \ni P$ *induced by Assumption 3.*

MAR-S Mean Functional

- Under the MAR-S framework, the statistical model of the data under consideration is semiparametric when the annotation score function π is known.
- Deriving efficient influence functions under semiparametric statistical models is typically more challenging than doing so under fully nonparametric statistical models, for which there is only one influence function, which is the efficient influence function.
- Lemma 2 shows that the EIF for a MAR-S mean functional θ under a nonparametric statistical model is also the correct EIF for θ under the semiparametric model.
- Intuitively, this lemma holds because perturbing the distribution given by the annotation score of a MAR-S mean functional does not change the value of the functional.
- If we had labeled our data in a different (but valid) way, the population-level value of the parameter being estimated would remain unchanged.

Identification of a Mean with Missing Data

Consider the dataset $\{W_i\}_{i=1}^n$ where $W_i := (M_i, A_i, X_i, U_i)$, with variable of interest $M_i \in \mathcal{M} \times \{0\} \subseteq \mathbb{R}$, annotation indicator $A_i \in \{0, 1\}$, observed covariates $X_i \in \mathcal{X} \subseteq \mathbb{R}^k$, and unstructured data $U_i \in \mathcal{U} \subseteq \mathbb{R}^\ell$. We define $\tilde{X}_i := (X_i, U_i)$.

We wish to compute the mean, denoted by θ, as the expected value of M_i^*:

$$\theta := E[M_i^*].$$

Under Assumptions 1 (consistency of potential outcomes) and 2 (missing at random structured data), we can identify θ as

$$\theta = E\left[\mu(\tilde{X}_i)\right],$$

where $\mu(\tilde{X}_i) := E[M_i \mid A_i = 1, \tilde{X}_i] = E[M_i \mid A_i = 1, X_i, U_i]$.

Identification of a Mean with Missing Data

Proposition 1

The efficient influence function for functional $E_P\left[\mu(\tilde{X}_i)\right]$ is

$$\varphi(W_i) = \mu(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \mu(\tilde{X}_i)) - \theta,$$

where π is the annotation scoring function.

As such, the (doubly) robust and efficient one-step estimator is

$$\hat{\theta} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left[\hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}(\tilde{X}_i))\right],$$

where $\mathcal{I}$ is the set of indices of the data allocated to the "estimation" partition of a random data split, and $\hat{\mu}$ is estimated on the other "training" partition.

This is the AIPW estimator (because M^* is just a potential outcome).

Debiased mean estimation

Suppose that $\pi(X_i) = |\mathcal{I} \cap \mathcal{J}|/|\mathcal{I}|$, where $\mathcal{J}$ is the set of indices corresponding to annotated data points. Then we see that

$$\hat{\theta} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left[\hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)} (M_i - \hat{\mu}(\tilde{X}_i)) \right] \qquad \text{(AIPW)}$$

$$= \underbrace{\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \hat{\mu}(\tilde{X}_i)}_{A} + \underbrace{\frac{1}{|\mathcal{I} \cap \mathcal{J}|} \sum_{i \in \mathcal{I} \cap \mathcal{J}} (M_i^* - \hat{\mu}(\tilde{X}_i))}_{B} \qquad \text{(PPI)}$$

The term A in the second expression is the best imputation-based guess of $E[M_i^*]$ in the estimation sample (the naive “plug-in” estimator), and the term B is a bias correction term–an estimate of the measurement error of our imputation function in the annotated sample. This expression is reproduced in recent work on “prediction-powered inference” (Angelopoulos et al., 2023)

Debiased mean estimation

Moreover:

$$\hat{\theta} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left[\hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}(\tilde{X}_i))\right] \tag{AIPW}$$

$$= \underbrace{\frac{1}{|\mathcal{I}\cap\mathcal{J}|}\sum_{i\in\mathcal{I}\cap\mathcal{J}} M_i^*}_{C} + \underbrace{\left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\hat{\mu}(\tilde{X}_i) - \frac{1}{|\mathcal{I}\cap\mathcal{J}|}\sum_{i\in\mathcal{I}\cap\mathcal{J}}\hat{\mu}(\tilde{X}_i)\right)}_{D}. \tag{FRA}$$

The expression (FRA) is reported in recent work on “flexible regression adjustment” (List et al., 2024). Under this formulation, we can view term C as our best estimate of the quantity of interest using only ground truth data. We then leverage the imputation function $\hat{\mu}$ as a form of nonparametric regression adjustment in term D, with the same intuition as a linear regression adjustment: we adjust for systematic differences between our large unlabeled sample and our small annotated ground truth sample as summarized by $\hat{\mu}$.

MAR-S and Double/Debiased Machine Learning

- MAR-S also relates closely to the DML framework (Chernozhukov et al., 2018).
- The MAR-S AIPW estimator is equivalent to one derived via the Neyman orthogonal score for estimating a potential mean under the "selection on observables" assumption in the DML causal inference setting.
- This is not coincidental, as we saw that the expected value of missing structured data can be interpreted as an average potential outcome and deriving a Neyman orthogonal score can be viewed as an "estimating equations" approach to constructing semiparametrically efficient estimators (Kennedy, 2023).
- Because MAR-S is based on the same fundamental semiparametric analysis as DML, there are likely many ways to apply insights from DML, such as automatic debiasing corrections (Chernozhukov et al., 2022a,b).

Linear Regression

Consider the dataset $\{W_i\}_{i=1}^{n}$, where $W_i = (M_i, C_i, A_i, X_i, U_i)$, with outcome $M_i \in \mathcal{M} \times \{0\} \subseteq \mathbb{R}$, regressors $C_i \in \mathcal{C} \subseteq \mathbb{R}^d$, annotation indicator $A_i \in \{0, 1\}$, observed covariates $X_i \in \mathcal{X} \subseteq \mathbb{R}^k$, and unstructured data $U_i \in \mathcal{U} \subseteq \mathbb{R}^\ell$. We further assume that $W_i \overset{\text{iid}}{\sim} P$ for some distribution P, and that $[C_i \perp\!\!\!\perp A_i] \mid X_i$.

We assume that

$$M_i^* = C_i^{\intercal}\theta + \varepsilon_i, \quad E[\varepsilon_i \mid C_i] = 0,$$

where $\theta \in \Theta \subseteq \mathbb{R}^p$. We are interested in identifying and estimating the j-th regression coefficient θ_j.

Linear Regression

By the Frisch-Waugh-Lovell theorem, we have that

$$\theta_j = E\left[{C_{i,j}^{\perp}}^2\right]^{-1} E\left[C_{i,j}^{\perp} M_i^*\right],$$

where $C_{i,j}^{\perp} := C_{i,j} - E^*[C_{i,j} \mid 1, C_{i,1}, \ldots, C_{i,j-1}, C_{i,j+1}, \ldots, C_{i,d}]$, and where E^* is the linear projection operator.

Let $\tilde{X}_i := (X_i, U_i, C_i)$. Then, by Assumptions 1 (consistency of potential outcomes) and 2 (missing at random structured data), we can identify θ_j as

$$\theta_j := \theta_{j,\text{den}}^{-1} \theta_{j,\text{num}} = E\left[{C_{i,j}^{\perp}}^2\right]^{-1} E\left[C_{i,j}^{\perp} \mu(\tilde{X}_i)\right]$$

where $\mu(\tilde{X}_i) = E\left[M_i \mid A_i = 1, \tilde{X}_i\right]$.

Linear Regression

Let $\check{C}_{i,j} := C_{i,j} - \hat{E}^*[C_{i,j} \mid 1, C_{i,1}, \ldots, C_{i,j-1}, C_{i,j+1}, \ldots, C_{i,d}]$, where $\hat{E}^*$ is the least squares operator. We can form the efficient estimator for θ_j as

$$\hat{\theta}_j := \hat{\theta}_{j,\text{num}}\hat{\theta}_{j,\text{den}}^{-1} = \frac{\sum_{i\in\mathcal{I}} \check{C}_{i,j}\left[\hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}(\tilde{X}_i))\right]}{\sum_{i\in\mathcal{I}} \check{C}_{i,j}^2}.$$

This is just a residualized regression with pseudo-outcome $\hat{\varphi}_i := \hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}(\tilde{X}_i))$, i.e.,

$$\hat{\theta}_j := \hat{\theta}_{j,\text{num}}\hat{\theta}_{j,\text{den}}^{-1} = \frac{\sum_{i\in\mathcal{I}} \check{C}_{i,j}\hat{\varphi}_i}{\sum_{i\in\mathcal{I}} \check{C}_{i,j}^2}.$$

The Efficient Imputation Function

- Importantly, the imputation function $\hat{\mu}$ is a function of context specific variables, i.e., $\tilde{X}_i := (X_i, U_i, C_i)$.
- Under Assumption 2, the imputation function should asymptotically approximate the function $E[M \mid A = 1, X = x, U = u, C = c]$ in order for estimation to be efficient.
- This, to our knowledge, has not been emphasized in the existing literature, which treats the imputation function as a completely arbitrary black box rather than taking a semi-parametric approach that can provide insights on the optimal imputation function for efficiency.

Linear IV

The MAR-S framework is straightforward to extend to linear IV. We follow the terminology and setup of Blandhol et al. (2022). Consider dataset $\{W_i\}_{i=1}^n$ with

$$W_i = (Y_i, M_i, C_i, Z_i, A_i, X_i, U_i),$$

where $Z_i \in \mathcal{Z} \subseteq \mathbb{R}$ is a candidate instrumental variable (IV), $C_i \in \mathcal{C} \subseteq \mathbb{R}^d$ are covariates (which contain a constant), $M_i \in \mathcal{M} = \{0, 1\}$ is a treatment of interest, $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$ is an outcome of interest, $A_i \in \{0, 1\}$ is the annotation indicator, $X_i \in \mathcal{X} \subseteq \mathbb{R}^k$ are observed covariates relevant to annotation, and $U_i \in \mathcal{U} \subseteq \mathbb{R}^\ell$ are unstructured data. We further assume that $W_i \overset{\text{iid}}{\sim} P$ for some joint distribution P, and that $[(C_i, Z_i) \perp\!\!\!\perp A_i] \mid X_i$.

Liinear IV

Potential outcomes are denoted $\{Y_i^m\}_{m\in\mathcal{M}}$ and potential treatments are denoted $\{M_i^z\}_{z\in\mathcal{Z}}$. Assuming consistency of potential outcomes and treatment, we have

$$Y_i = \sum_{m\in\mathcal{M}} \mathbb{I}(M_i = m) Y_i^m, \quad M_i = \sum_{z\in\mathcal{Z}} \mathbb{I}(Z_i = z) M_i^z.$$

We are interested in identifying and estimating the average treatment effect:

$$\theta := E[Y_i^{m=1} - Y_i^{m=0}].$$

Linear IV

The efficient estimator for θ is:

$$\hat{\theta} = \hat{\theta}_{\text{num}}\hat{\theta}_{\text{den}}^{-1} = \frac{\sum_{i\in\mathcal{I}} \check{Z}_i Y_i}{\sum_{i\in\mathcal{I}} \check{Z}_i \left[\hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}(\tilde{X}_i))\right]},$$

where $\check{Z}_i$ is Z_i residualized with a least squares fit on C_i and $\tilde{X}_i := (X_i, U_i, Z_i, C_i)$.

Note that the efficient estimator is just a TSLS regression with pseudo-treatment $\hat{\varphi}_i := \hat{\mu}(\tilde{X}_i) + \frac{A_i}{\pi(X_i)}(M_i - \hat{\mu}(\tilde{X}_i))$, or

$$\hat{\theta} = \hat{\theta}_{\text{num}}\hat{\theta}_{\text{den}}^{-1} = \frac{\sum_{i\in\mathcal{I}} \check{Z}_i Y_i}{\sum_{i\in\mathcal{I}} \check{Z}_i \hat{\varphi}_i}.$$

Notice once again that the optimal imputation function is also a function of Z and C.

Differences-in-Differences

In this application, we focus on the nonparametrically founded differences-in-differences (DiD) estimator introduced in Callaway and Sant'Anna (2021).

Consider the dataset $\{W_i\}_{i=1}^{n}$:

$$W_i = (M_{i1}, \ldots M_{iT}, D_{i1}, \ldots, D_{iT}, A_{i1}, \ldots, A_{iT}, X_{i1}, \ldots, X_{iT}, U_{i1}, \ldots, U_{iT}),$$

where $M_{it} \in \mathcal{M} \times \{0\} \subseteq \mathbb{R}$ is an outcome of interest, and $D_{it} \in \mathcal{D} = \{0, 1\}$ are treatment indicators.

Let G_{ig} be a binary indicator for a unit i first being treated at time g and let C_i be an indicator for units that are never treated. Furthermore, we define $M_{it}^{g=0}$ to be the untreated potential outcome of a unit i at time t and M_{it}^{g} to be the potential outcome of a unit i at time t if they first became treated in period g. Assuming consistency of potential outcomes, we have

$$M_{it} = M_{it}^{g=0} + \sum_{g=2}^{T} \left(M_{it}^{g} - M_{it}^{g=0} \right) G_{ig}.$$

Differences-in-Differences

We are interested in the estimand

$$\theta = E\left[M_{it}^{*\,g} - M_{it}^{*\,g=0} \mid G_{ig} = 1\right],$$

the average treatment effect on those treated in cohort g at time t. For concreteness, but without loss of generality, we set $g = 2$ and $t = 2$, the canonical two-period DiD setup.

We can express θ as $E\left[M_i^* \mid G_{i2} = 1\right] - E\left[M_i^* \mid C_i = 1\right]$ for $M_i^* := M_{i2}^* - M_{i1}^*$.

Let $M_i := M_{i2} - M_{i1}$, $A_i := A_{i1}A_{i2}$, and $\tilde{X}_i := (X_{i1}, X_{i2}, U_{i1}, U_{i2})$. Under Assumptions 1 and 2, we can identify θ as

$$\theta = \theta_G - \theta_C, \quad \theta_G = E[\mu_G(\tilde{X}_i) \mid G_{i2} = 1], \quad \theta_C = E[\mu_C(\tilde{X}_i) \mid C_i = 1],$$

where $\mu_G(\tilde{X}_i) = E[M_i \mid G_{i2} = 1, A_i = 1, \tilde{X}_i]$ and $\mu_C(\tilde{X}_i) = E[M_i \mid C_i = 1, A_i = 1, \tilde{X}_i]$.

Differences-in-Differences

We can form efficient one-step estimators for θ_G and θ_C as

$$\hat{\theta}_G = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{G_{i2}}{\hat{P}(G_{i2} = 1)} \left[\hat{\mu}_G(\tilde{X}_i) + \frac{A_i}{\pi(X_i)} (M_i - \hat{\mu}_G(\tilde{X}_i)) \right],$$

$$\hat{\theta}_C = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{C_i}{\hat{P}(C_i = 1)} \left[\hat{\mu}_C(\tilde{X}_i) + \frac{A_i}{\pi(X_i)} (M_i - \hat{\mu}_C(\tilde{X}_i)) \right],$$

where $\hat{P}$ is an empirical probability, and we can combine them to form the efficient estimator for θ as

$$\hat{\theta} = \hat{\theta}_G - \hat{\theta}_C.$$

RDD Under Local Randomization

We focus on sharp RDD under the local randomization framework (Cattaneo et al., 2024), as under the continuity framework, the functionals are *not* pathwise differentiable.

Consider the dataset $\{W_i\}_{i=1}^n$ with $W_i = (M_i, R_i, D_i, A_i, X_i, U_i)$, where $M_i \in \mathcal{M} \times \{0\} \subseteq \mathbb{R}$ is an outcome of interest, $R_i \in \mathcal{R} \subseteq \mathbb{R}$ is the running variable, $D_i \in \mathcal{D} = \{0, 1\}$ is a treatment indicator, $A_i \in \{0, 1\}$ is the annotation indicator, $X_i \in \mathcal{X} \subseteq \mathbb{R}^k$ are observed covariates, and $U_i \in \mathcal{U} \subseteq \mathbb{R}^\ell$ are unstructured data. We assume $W_i \overset{\text{iid}}{\sim} P$ for some joint distribution P, and that $[(R_i, D_i) \perp\!\!\!\perp A_i] \mid X_i$.

RDD

Potential outcomes $\{M_i^d\}_{d\in\mathcal{D}}$ are related to observed outcomes via the assumption of consistency:

$$M_i = D_i M_i^{d=1} + (1 - D_i) M_i^{d=0}.$$

We are interested in the estimand

$$\theta = E\left[{M_i^*}^{d=1} - {M_i^*}^{d=0} \mid R_i \in \mathcal{B}\right]$$

for a set (or “window”) $\mathcal{B} \subset \mathcal{R}$. We can write θ as

$$\theta = E[M_i^* \mid R_i \in \mathcal{B}, D_i = 1] - E[M_i^* \mid R_i \in \mathcal{B}, D_i = 0].$$

Let $\tilde{X}_i := (X_i, U_i)$. Then, under Assumptions 1 and 2, we can identify θ as $\theta = \theta_1 - \theta_0$ where

$$\theta_d = E[\mu_d(\tilde{X}_i) \mid R_i \in \mathcal{B}, D_i = d],$$

with $\mu_d(\tilde{X}) = E[M_i \mid R_i \in \mathcal{B}, D_i = d, A_i = 1, \tilde{X}_i]$.

RDD

Forming efficient one-step estimators for θ_1 and θ_0, respectively, as

$$\hat{\theta}_1 = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{\mathbb{I}\{R_i \in \mathcal{B}, D_i = 1\}}{\hat{P}(R_i \in \mathcal{B}, D_i = 1)} \left[\hat{\mu}_1(\tilde{X}_i) + \frac{A_i}{\pi(X_i)} (M_i - \hat{\mu}_1(\tilde{X}_i)) \right],$$

$$\hat{\theta}_0 = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \frac{\mathbb{I}\{R_i \in \mathcal{B}, D_i = 0\}}{\hat{P}(R_i \in \mathcal{B}, D_i = 0)} \left[\hat{\mu}_0(\tilde{X}_i) + \frac{A_i}{\pi(X_i)} (M_i - \hat{\mu}_0(\tilde{X}_i)) \right],$$

where $\hat{P}$ again indicates an empirical probability, we can combine them to form an efficient estimator of θ as

$$\hat{\theta} = \hat{\theta}_1 - \hat{\theta}_0,$$

Outline

Introduction

Literature

The MAR-S Framework

Applications

Extensions

Empirical Examples

Conclusion

Issues that Arise in Large Datasets

- In many economic applications using unstructured data, imputed structured data are aggregated, often nonlinearly.
- The existing literature assumes that ground truth data are available at the same level of aggregation as the parameter of interest.
- However, collecting data at this level is typically infeasible, because it would require labeling a very large number of unstructured data instances.
- For the framework to be practically useful in many cases, we need it to address this challenge.

Linear Models with MAR-S First-Step Measurement Error

Consider the linear model

$$Y = X^{*}\beta + \varepsilon,$$

In particular, we consider the case that X_i^{*} is not actually observed, but has been estimated with a MAR-S first-step, X_i.

$$X_{k,i} = X_{k,i}^{*} + \eta_{k,i}, \quad (\varepsilon_i, X_{k,i}^{*}) \perp\!\!\!\perp \eta_{k,i}, \quad \eta_{k,i} \overset{d}{\approx} N(0, \sigma^2_{\eta,k}),$$

$$\sigma^2_{\eta,k} := |\mathcal{I}_k|^{-1}\mathrm{Var}(\varphi_k)$$

where φ_k is the efficient influence function associated with X_k.

As such, we are in a **classical measurement error setting**.

This setup nests the case where only some regressors are generated by a MAR-S first-step, in which case $\sigma^2_{\eta,k} = 0$ for non-MAR-S regressors.

Measurement Error Structure

We therefore assume:

$$\eta_i \perp\!\!\!\perp Y_i, \quad \eta_i \overset{\text{iid}}{\sim} N(0, \Sigma)$$

where:

$$\Sigma := \text{diag}(\sigma^2_{\eta,1}, \ldots, \sigma^2_{\eta,K})$$

Reasonable given:

- Each MAR-S first-step regressor X_k is typically estimated on a separate sample.
- For large $\mathcal{I}_k$, the normal approximation for $\eta_{k,i}$ is accurate.

We consider Σ to be known, which is reasonable when $\mathcal{I}_k$ is sufficiently large.

GMM Moment Condition for β

Under the measurement error model, the coefficient vector β satisfies the GMM moment condition:

$$E\left[g(X_i, Y_i, \beta)\right] = 0$$

where:

$$g(X_i, Y_i, \beta) = X_i(Y_i - X_i^\top \beta) + \Sigma\beta$$

Interpretation:

- First term: Moment condition from OLS
- Second term: Correction for measurement error variance Σ

Plug-in GMM Estimator

The corresponding consistent GMM estimator is:

$$\hat{\beta} = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i^{\top} - \Sigma \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i Y_i \right)$$

Equivalently:

$$\hat{\beta} = (X^{\top} X - n\Sigma)^{-1} X^{\top} Y$$

Asymptotic Distribution: Under mild regularity conditions:

$$\sqrt{n}(\hat{\beta} - \beta) \underset{d}{\rightarrow} N(0, G^{-1} \Xi G^{-1})$$

where:

$$G = -\left(E[X_i X_i^{\top}] - \Sigma \right), \quad \Xi = \mathrm{Var}(X_i(Y_i - X_i^{\top} \beta))$$

Linear Models with MAR-S First Step Measurement Error

- This setup can be straightforwardly extended to the case of clustering.
- It can also be readily specialized to the panel data case (Deaton, 1985).
- Moreover, there are various arms reach extensions of the framework for cases where, e.g., there is heterogeneity in variance for η_i across *i*, the outcome is a MAR-S first-step as well (or alone), measurement error distributions are not assumed to be normal, etc.
- The MAR-S first-step is crucial for making this linear measurement error model plausible.

Choosing an Annotation Score Function

- Choosing an appropriate annotation function is not always trivial.
- In large datasets, the structured data of interest may represent a “rare event”.
- Limiting to a subset of the content based on some rule (*e.g.,* the presence of certain keywords) violates the missing at random assumption.

Choosing the Annotation Score Function

Recall that the asymptotic variance of $\hat{\theta}$ is:

$$\sigma^2 := \text{Var}\left(\mu(\tilde{X}) + \frac{A}{\pi(X)}(M - \mu(\tilde{X}))\right)$$

which can be expressed as:

$$\sigma^2 = E\left[\frac{A}{\pi(X)^2}(M - \mu(\tilde{X}))^2\right] + c$$

where c collects all terms that do not depend on $\pi(X)$

Optimal Annotation Score Function

To minimize the variance, we solve the following constrained optimization:

$$\min_{\pi,\lambda,\kappa} \left\{ E\left[\frac{A}{\pi(X)^2}(M-\mu(\tilde{X}))^2\right] + \lambda(E[\pi(X)]-1) - \kappa E[\pi(X)] \right\}$$

where:

- λ, κ: KKT multipliers
- Constraints: $0 \leq E[\pi(X)] \leq 1$

Characterization of Optimal Score Function

Under Assumption 2, the optimal annotation score function is:

$$\pi_{IS}(X) \propto E\left[(M^* - \mu(\tilde{X}))^2 \mid X\right]^{1/2}$$

This places more weight on units that are harder to impute, capturing the idea of uncertainty-driven sampling.

However, π_{IS} is infeasible, as it depends on unobserved M^*

Feasible Annotation Score Function

Following Zrnic and Candes (2024), we may implement a feasible approximation:

$$\pi_{FIS}(X) \propto \text{err}(X)$$

where:

$$\text{err}(X) \approx E\left[(M^* - \mu(\tilde{X}))^2 \mid X\right]^{1/2}$$

Interpretation:

- err(X) is a proxy for imputation uncertainty
- Can be estimated using:
  - Model-based variance estimates
  - Cross-validation residuals
  - Ensemble disagreement (if using ML models)

Outline

Introduction

Literature

The MAR-S Framework

Applications

Extensions

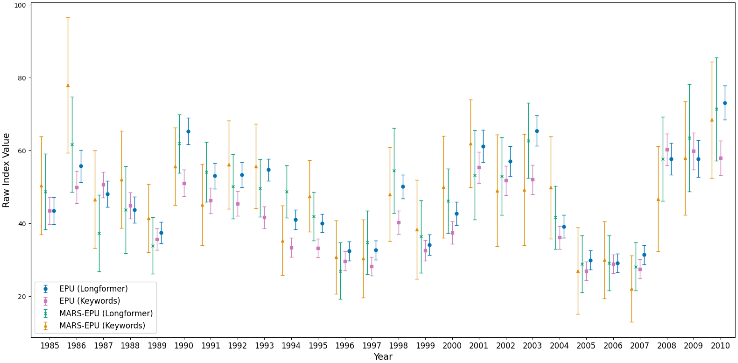
Empirical Examples

Conclusion

Debiasing Baker, Bloom, and Davis (QJE 2016)

- Baker, Bloom, and Davis (2016) construct an "economic policy uncertainty" (EPU) index by computing the share of articles in a handful of newspapers each month that also satisfy a simple keyword query, and likewise perform an audit study that collects ground truth on mentions of economic policy uncertainty.
- In the following figure, we present the original and MAR-S debiased EPU indices. We also plot debiased and naive versions of the EPU index, imputed with the long document transformer `longformer-base-4096`.
- For this figure, we utilize 25% of their audit sample ground truth labels for debiasing, and use the rest of the audit sample for imputation.

Debiased EPU



Firm-Level Regression

We reanalyze the following baseline regression from Table IV, column (5):

$$\Delta \mathrm{Emp}_{it} = \beta \, \Delta \log \mathrm{EPU}_t \times \mathrm{intensity}_{it} \\ + \gamma \, \Delta \frac{\text{Federal purchases}_t}{\mathrm{GDP}_t} \times \mathrm{intensity}_{it} + \alpha_i + \psi_t + u_{it}$$

- i: Firm index, t: Year index
- α_i, ψ_t: Firm and year fixed effects
- $\mathrm{intensity}_{it}$: Firm-year policy exposure intensity
- $\Delta \mathrm{Emp}_{it}$: Employment growth
- β: Coefficient of interest

Results

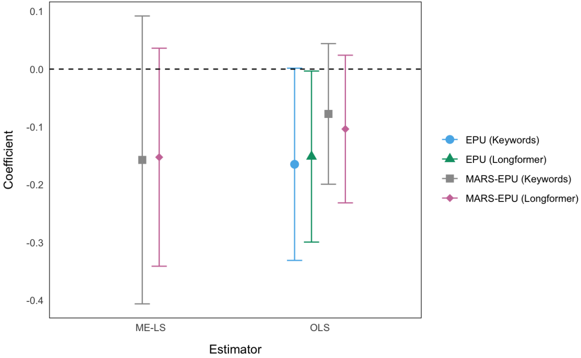
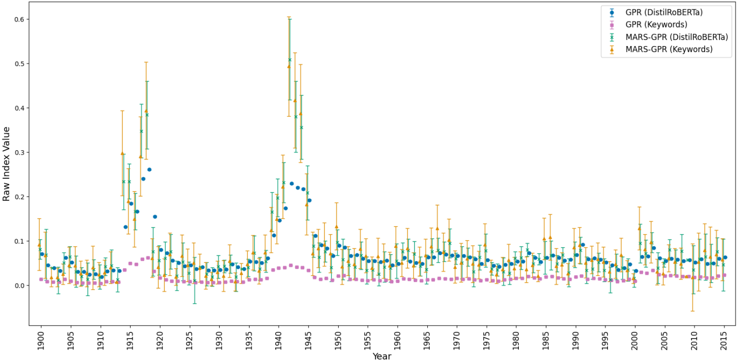


Figure: Estimates of β by estimator for EPU and MAR-S-EPU indices.

Debiasing Caldara and Iacoviello (AER 2022)

- Caldara and Iacoviello (2022) construct a “geopolitical risk” (GPR) index by computing the share of articles in a handful of newspapers each month that satisfy a keyword query.
- An inference-minded interpretation of their index is: what is the probability of an article being written on geopolitical risk period by period?
- Fortunately, Caldara and Iacoviello (2022) perform audit studies that collect ground truth on mentions of geopolitical risk.
- We utilize all of the audit sample ground truth labels for MAR-S, and unlabeled samples from Proquest for imputation.

Debiased GPR



Empirical Illustration: GPR Regression

We revisit a representative regression analysis from Caldara and Iacoviello, which estimates the impact of the GPR index on economic disaster probability:

$$D_{it} = \beta \, \mathrm{GPR}_t + \delta \, \Delta\mathrm{GDP}_{it-1} + u_{it}$$

In Figure 6.2, we compare the following estimation strategies for β:

- **ME-LS:** Measurement-error corrected least squares estimator using MAR-S-GPR index
- **Naive OLS:**
  - With MAR-S-GPR index
  - With unadjusted GPR index

Results

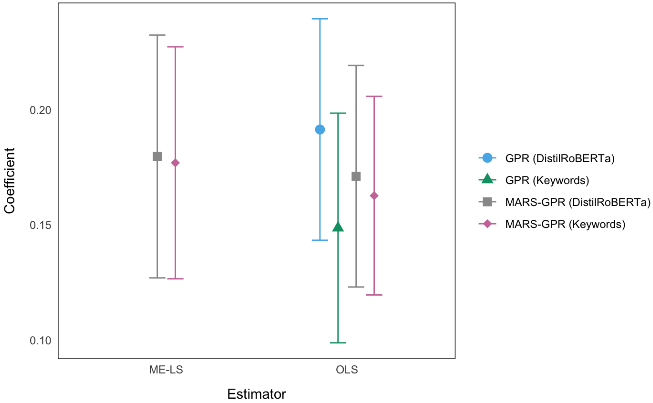


Figure: Estimates of β by estimator for GPR and MAR-S-GPR indices.

Design Choices in MAR-S

- In order to probe design choices associated with the MAR-S framework, we develop an example in which we examine the share of articles that are about politics in local U.S. newspapers across time.
- Articles were selected at random for annotation from a large sample of historical U.S. newspapers (Dell et al., 2023).
- No keyword filters were implemented beforehand, so the relevant population is all front page content in thousands of local newspapers.
- In the economics literature, annotated text audit samples are typically very small, whereas by labeling our own data, we could create a large enough set to examine how the size of the annotation set influences inference under MAR-S.

Varying the Size of Annotated Data

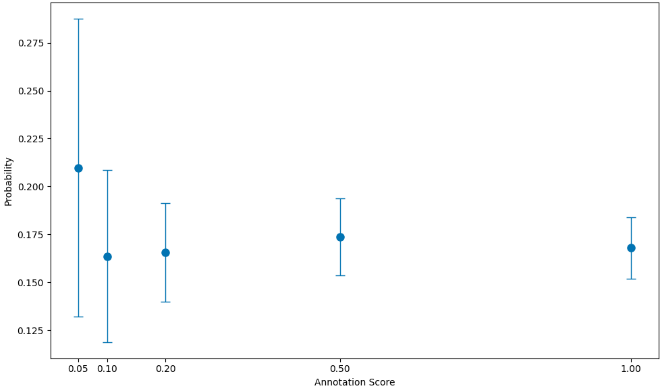


Figure: Estimating P(discussion of politics) with annotation score $\pi \in \left\{\frac{1}{20}, \frac{1}{10}, \frac{1}{5}, \frac{1}{2}, 1\right\}$.

Varying Imputation Accuracy

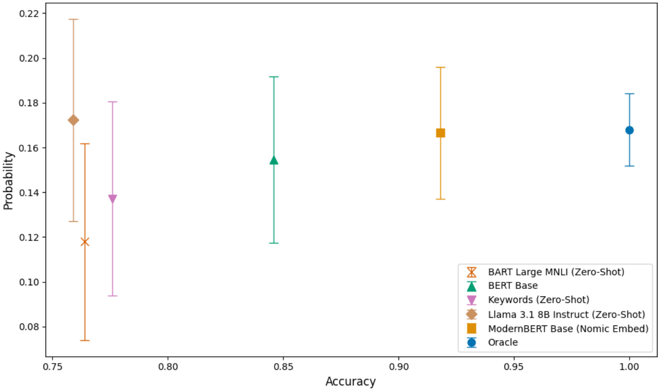


Figure: Estimating *P*(discussion of politics) across fine-tuned and zero-shot classifiers.

Outline

Introduction

Literature

The MAR-S Framework

Applications

Extensions

Empirical Examples

Conclusion

Conclusion

- Deep learning provides powerful tools for processing unstructured economic data.
- By combining a new literature on debiased inference with black box AI and an older literature on semiparametric inference for missing data, we recover efficient inference for unstructured data.
- We apply our framework to a variety of common estimators and settings in empirical economics.
- Accounting for imputation bias in inference with unstructured data can fundamentally influence both point estimates and uncertainty quantification.