MLM25

Exploratory Data Analysis Presentations

Slides: go.wisc.edu/kbx667

WISCONSIN INSTITUTE FOR DISCOVERY

data science HUB

ML+X

Slides: go.wisc.edu/kbx667

Tonight's Agenda

1. **Announcements**

2. **EDA Presentations!**

Slides: go.wisc.edu/kbx667

MLM25 — Looking Ahead

1. **10/9 (Thur), 4:30-7:30pm**: Intro to AWS SageMaker for ML/AI
2. **10/16 (Thur), 4:30-6:30PM**: Sprint 3 + RAG with Watsonx.ai
   a. Request access to Watsonx.ai (50 person cap): **https://forms.gle/JkcMFgXx7PfpfjrM9**
3. **10/23 (Thur), 4:30-6:30PM**: Sprint 4

Extras: The ML+X Forum and ML+Coffee sessions next week may of interest. RSVP.

Full schedule: ml-marathon.wisc.edu/schedule/

Slides: go.wisc.edu/kbx667

MLM25 — EDA Presentations

- **4 minutes per team**
  - 1-2 questions/comments following each presentation.
- **I'll hold up a sign to give you your 1-minute warning.**
- **Presentations ordered by challenge type.**
  - As we go along, you can quickly gloss over insights that were already shared by other teams.

Digit Recognizer: Data Exploration

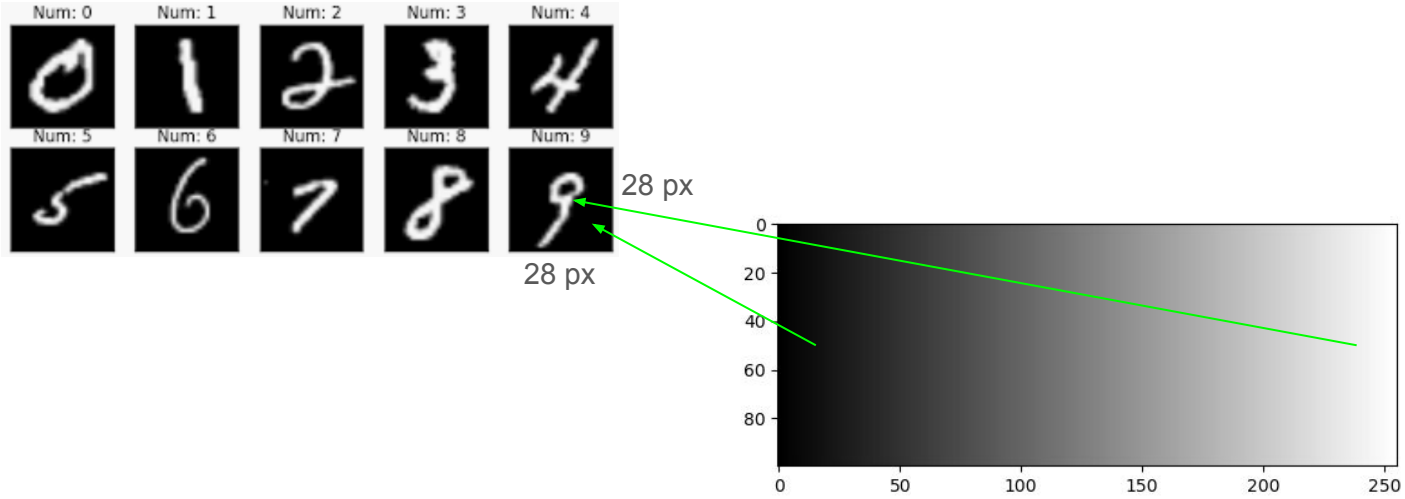
The Overfitters

MLM25 - October 02, 2025

MNIST - a dataset of handwritten digit images

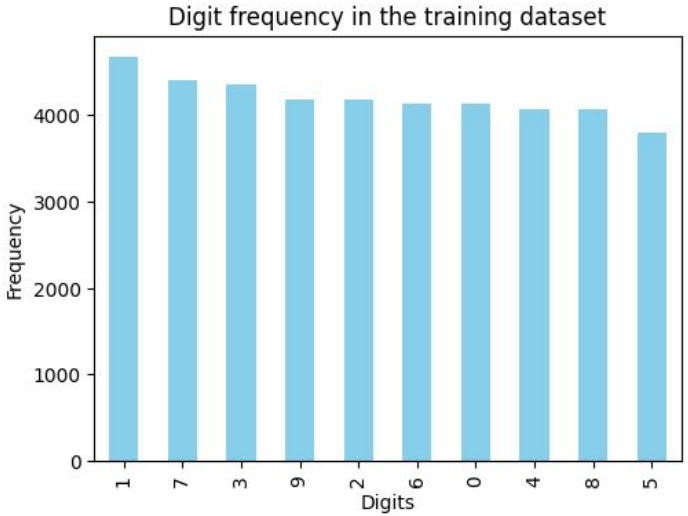
Train + **Test** dataset:

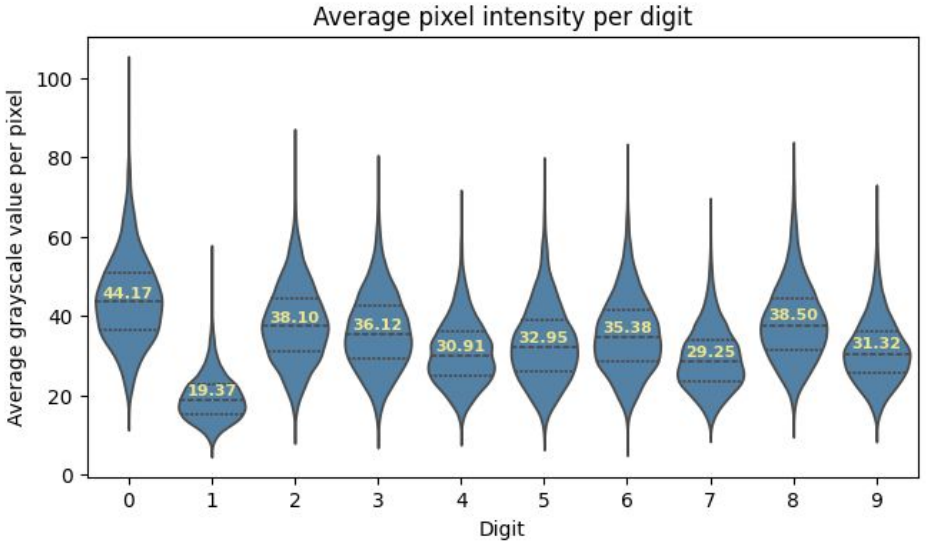
- Gray-scale images of hand-drawn digits, from zero through nine.
- Digit image dimension: 28 px x 28 px = 784
- Pixel-value: 0 - 255 (gray scale)
- 70,000 rows (images) x 786 columns (px/image)



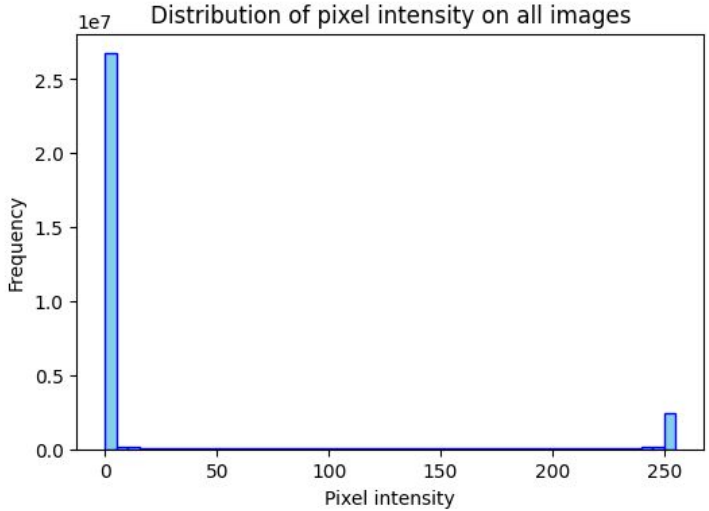
Data exploration

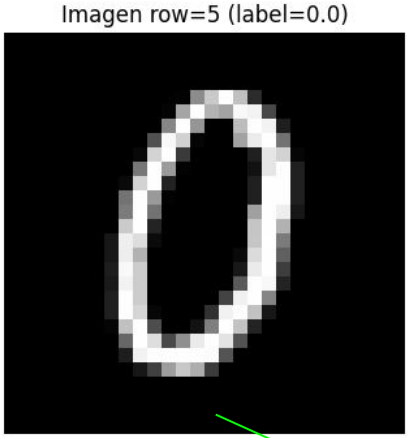
- No NAs
- Values within the 0 - 255 grayscale range (no outliers)



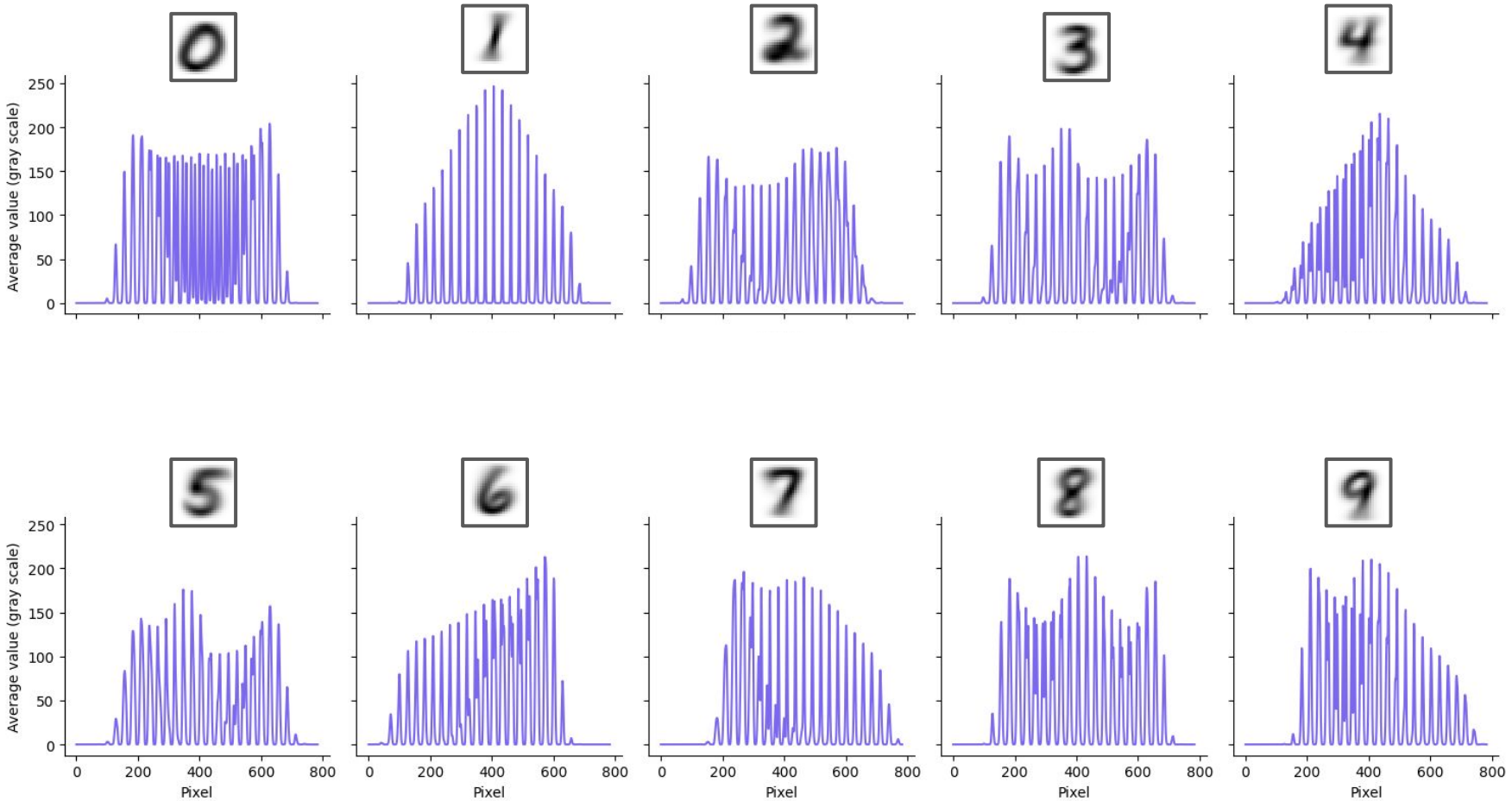


- Most pixels are black!





Average pixel intensity:
(650*0 + 135*250) / 784 = **43.05**

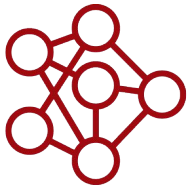

0
1
2
3
4
5
6
7
8
9
Average value (gray scale)
0
50
100
150
200
250
0
200
400
600
800
Pixel

- **Observations**:
  The average pixel intensity value is not sufficient to identify each number.
  More complex models that account for stroke patterns fit the dataset better.

- **Next steps:**
  Test clustering methods to evaluate how similar the datasets are.

Exploratory Data Analysis: MNIST Digit Recognizer

Inkvestigators: Andrew, Lily, Quincy, Ryan

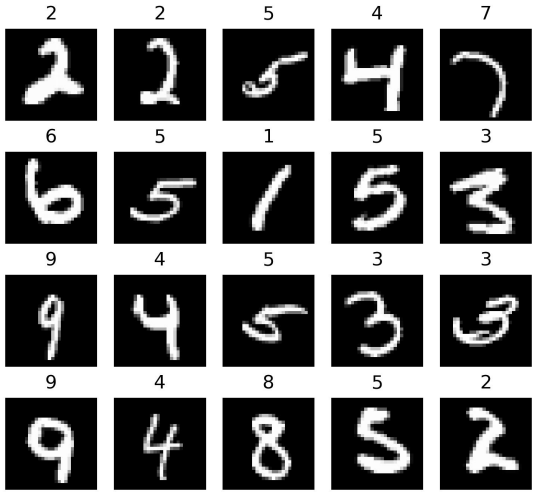


ML+X

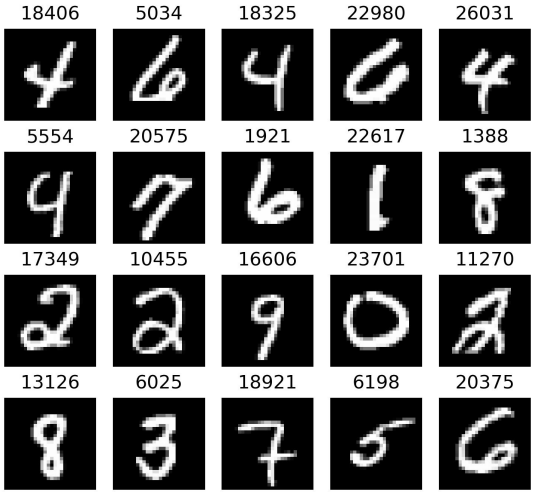
Data Introduction - MNIST dataset from Kaggle

- Training dataset (n = 42,000), Test dataset (n = 28,000)
- Data format:
  - Row: samples
  - Columns: label (for training dataset only), 784 columns of pixel values

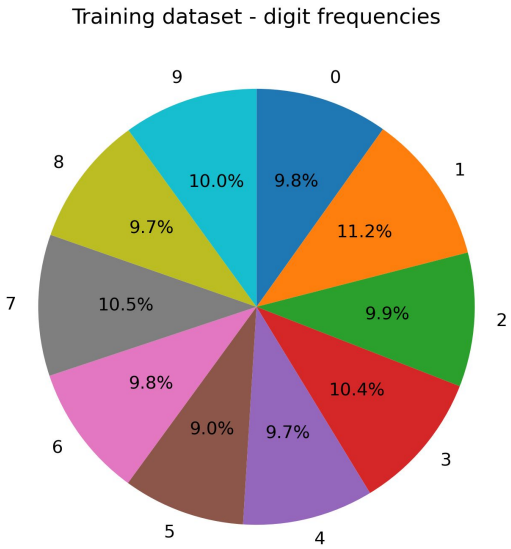
Training dataset - random 12



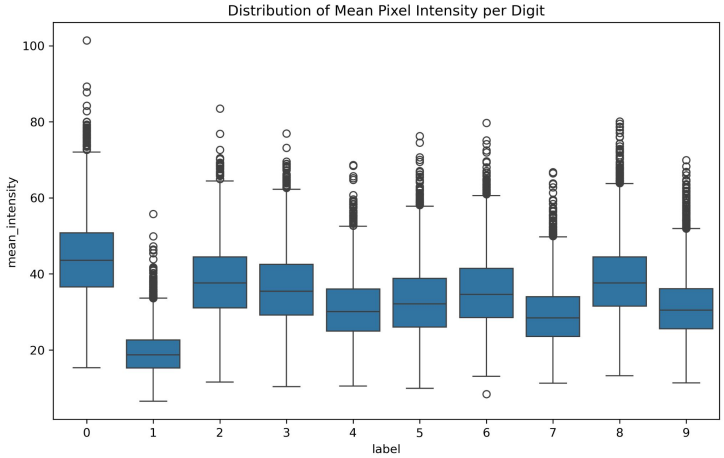
Test dataset - random 42



Insights / Patterns / Challenges

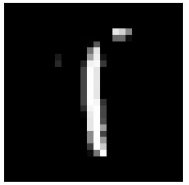


- Roughly equal
- Digit 5 has fewest images



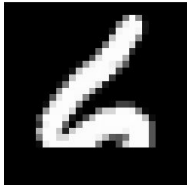
- Digit 1 has lowest mean
- Other means are similar

Some images have ‘debris’ →

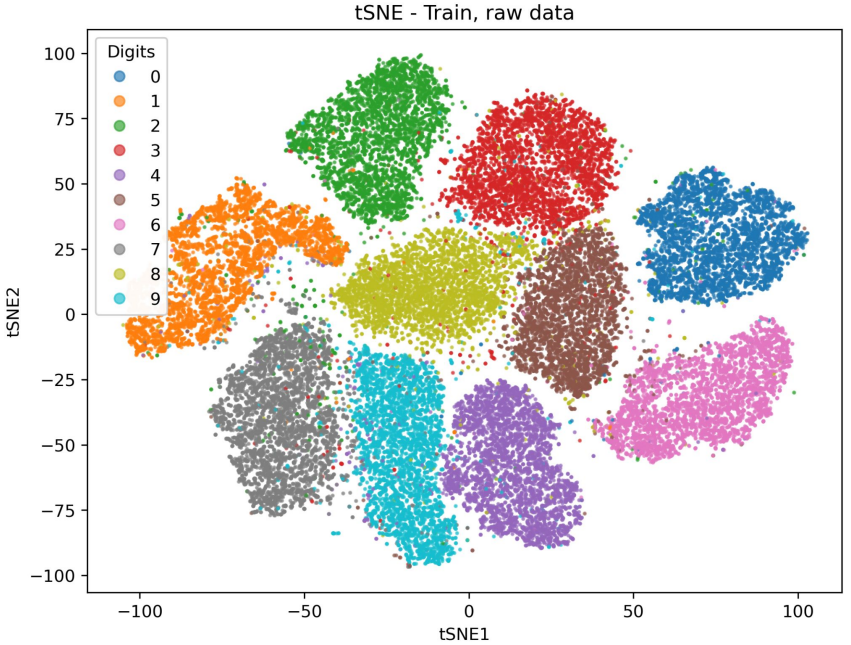


Some digits are not a single object →

Some digits are cut off →



Baseline models



The PCA plot showed some clusters (digit ‘1’ being most obvious), but also a lot of overlaps. The tSNE plot was able to cluster the digits well, and represents a good starting point. Binarized dataset (0-127 = 0; 128-255 = 1) did not improve PCA or tSNE.

Dataset Cleaner (prototype)

```
# 1. Check missing values
```

Reason: A missing pixel means incomplete or corrupted data.
Function: Ensures every image has valid information before training.

```
# 2. Pixel value ranges
```

Reason: Pixels should only be grayscale (0–255). Out-of-range = corruption.
Function: Verifies images are valid and comparable.

```
# 3. Check shape
```

Reason: Each image must have exactly 784 pixels (28×28).
Function: Guarantees the dataset fits the model's input shape.

```
# 4. Class balance
```

Reason: If some digits dominate, the model becomes biased.
Function: Confirms fair representation of all 0–9 digits.

```
# 5. Check duplicates
```

Reason: Too many duplicate images reduce diversity, hurt generalization.
Function: Identifies redundancy in the dataset.

```
# 6. Check empty images
```

Reason: Blank (all 0) or uniform images add noise, no useful features.
Function: Filters out images that don't help learning.

```
# 7. Normalize (scale to 0-1)
```

Reason: Neural networks train more stably when inputs are small and standardized.
Function: Rescales pixels into `[0, 1]` for faster, more accurate training.

```
# 8. Save cleaned dataset
```

Next steps

- Try tSNE on test dataset, submit result on Kaggle, and get % accuracy - as baseline.
- Run Dataset Cleaner, then re-run PCA and tSNE.
- Run logistic regression, SVM, Random Forest on cleaned dataset
- Apply Deep Learning models on the training and test datasets. Options:
  - TensorFlow/Keras (model= Sequential)
  - PyTorch (nn.Sequential): compare using linear layers to using convolutional layers:
    - LeNet 5 (uses convolutional layering)
  - ResNet / RetinaNet

Data Set

metadata.csv (32 rows x 6 columns)

- Information about references

	id	type	title	year	citation
1					
2	**amazon2023**	report	2023 Amazon Sustainability Report	2023	Amazon Staff. (2023). Amazon Sustainability Report. https://sustainability.aboutamazon.com/2023-amazon-sustainability-report.pdf

train_QA.csv (41 rows x 9 columns)

- Question-answer with references

id	question	answer	answer_value	answer_unit	ref_id
q003	What is the name of the benchmark suite presented in a recent paper for measuring inference energy consumption?	The ML.ENERGY Benchmark	ML.ENERGY Benchmark	is_blank	['chung2025']
q009	What were the net CO2e emissions from training the GShard-600B model?	4.3 tCO2e	4.3	tCO2e	['patterson2021']
q054	What is the model size in gigabytes (GB) for the LLaMA-33B model?	64.7 GB	64.7	GB	['chen2024']

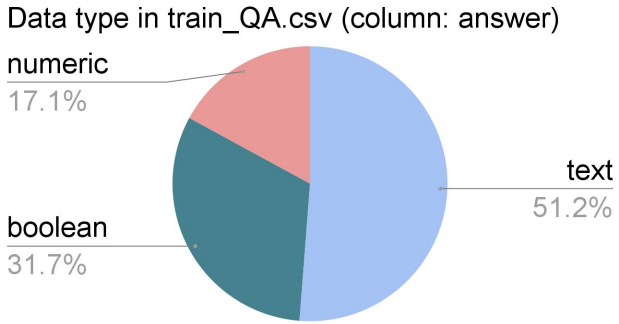
ref_url	supporting_materials
['https://arxiv.org/pdf/2505.06371']	We present the ML.ENERGY Benchmark, a benchmark suite and tool for measuring inference energy consumption under realistic service environments...
['https://arxiv.org/pdf/2104.10350']	"Training GShard-600B used 24 MWh and produced 4.3 net tCO 2 e."
['https://arxiv.org/pdf/2405.01814']	Table 3: Large language models used for evaluation.

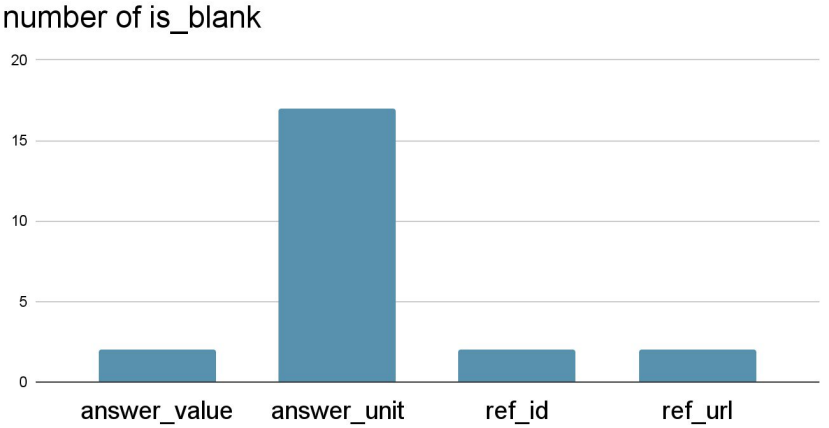
test_Q.csv (282 rows x 9 columns)

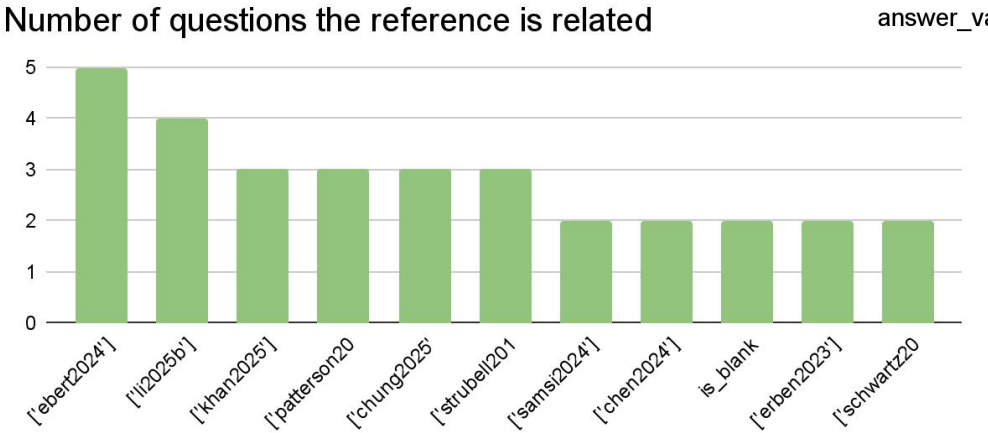
- Supposed to predict answers here

question	answer	answer_value	answer_unit	ref_id	ref_url	supporting_materials	explanation
What was the average increase in U.S. data center electricity consumption between 2010 and 2014?			percent				

Data exploration







Data Clean

True or False: Hyperscale data centers in 2020 achieved more than 40% higher efficiency compared to traditional data centers.

Answer: **TRUE** → `True`

Python doesn't recognize `TRUE`

original	cleaned
Interval [0.02, 0.1]	List `[0.02, 0.1]`
is_blank	NaN

Packages: `pandas, numpy`

Functions:

- `isna()`
- `value_counts()`
- `copy()`

Baseline Model

For all test questions, we find the **most similar** question in the training set (based on text overlap) and copy its answer.

Test question	Predicted answer
What was the average increase in U.S. data center electricity consumption between 2010 and 2014?	Unable to answer with confidence based on the provided documents.
How many data centers did AWS begin using recycled water for cooling in 2023?	0.18 L/kWh

WATTBOT: ESTIMATING AI EMISSIONS & COSTS WITH RAG

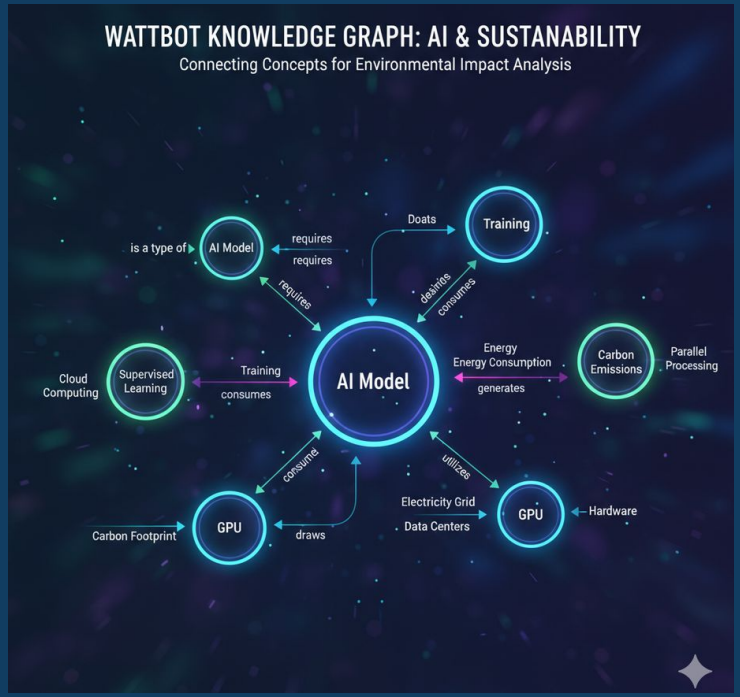
Ishaan Kharbanda

Arunjay Agrawal

Aadya Ganjigunta

Diya Kothari

Devanshi Jain



INTRODUCTION

- **WattBot**: a system that extracts credible environmental impact estimates of AI and data centers from peer-reviewed sources.
- To train and test it, we were given three datasets:
  - **Metadata** (32 research papers & reports from 2019–2025)
  - **Training Q&A** (41 questions with answers and references)
  - **Test Questions** (282 new questions, no answers provided)
- These datasets cover topics like AI energy use, carbon emissions, and hardware efficiency. They include a mix of answerable and unanswerable questions and require both numeric and categorical answers.

KEY STEPS

- **Fixed Encoding:** standardized file formats (handled “latin-1” encoding issues)
- **Checked for missing values**: none in key columns.
- **Answer types**: handle "Unable to answer" cases properly (fallback system needed) and parse different answer types (numeric, categorical, blank)
- **Extract data**: find evidence from tables and figures
- **References**: almost every training question links to a document — making this a highly evidence-based dataset.

TOOLS WE USED

- **Data analysis & cleaning: Pandas, NumPy**
- **Visualization: Matplotlib, Seaborn, WordClouds**
- **PDF parsing: PyMuPDF, PDFPlumber (to extract text from papers)**
- **Language tools:**
  - **LangChain for splitting text**
  - **Sentence Transformers + FAISS for search**
  - **Hugging Face Transformers (Flan-T5) for Q&A**

INSIGHTS

- Training questions are mostly factual “What” questions
- **Frequent keywords:** model, energy, training, gpu, emissions
- About ⅓ of the dataset is environmental-focused
- **Baseline Model:**
  - ***PDF Parsing***: PyMuPDF for text, Camelot for tables, PyTesseract for figures
  - ***Workflow***:

Input PDF → code separates chunks (figures, tables, captions) → JSON → RAG pipeline

  - ***Retrieval***: FAISS index with MiniLM embeddings (vector DB testing in progress)
  - ***Q&A:*** Flan-T5 (handles simple answers, weaker on detailed reasoning)
  - ***Future Direction***: Hybrid approach combining Vector DB + Knowledge Graphs for richer context

NEXT STEPS

Stress testing the PDF parser

Build a sample hybrid model combining Knowledge Graph (KG) and Vector DB

- KGs are strong for factual lookups
- Vector DBs excel at semantic questions

Hybrid approach leverages strengths of both

Mitigates weaknesses by letting each cover the other's gaps

THANK

YOU!

WattBot BAAKE team

WattBott Data

32 papers and reports that cover:

- Measurement of environmental impact of AI: energy consumption, water usage, CO2 emissions etc.
- Corporate reports (i.e. Google, Facebook)
- Global policy and regulations that impact measurement requirements (actual vs. estimates), reporting, and impact (i.e. goals for Power Usage Effectiveness (PUEs)
- Ethical discussions related governance and transparency

Training data

- Q&A pairs with answer types that include: True/False, quotes, calculations based on table/graph data, or blank if answer is not in the corpus.

Key Steps for the Text Data

1. Load the corpus
2. Split text into chunks
3. Create embeddings
4. Retrieve relevant chunks (iterative improvement)
5. Generate answer using retrieved text by asking a language model

Key Steps for the Tables and Graphs?

Possible Tools/Packages/Functions

Tools

- Starter code, Chris’s Romeo and Juliet example

Packages

- SentenceTransformer
- PyPDF2

Functions

- ?

Challenges

Confidence is relatively low (mean around 50%), and high proportion of is_blank (80/282)

The PDF conversion tool provided does not parsing table data very well.

Look for trends in posted output related to low confidence?

Images? Layout aware model?

Next Steps

- *Review sample code to evaluate outputs with low confidence to determine if we need to adjust chunking strategy?*
- *Integrate more efficient PDF parsing tools to extract data from tables.*
- *Try sophisticated models and run on GPU*

LabRAGs - WattBot: Estimating AI Emissions with RAG

Exploratory Data Analysis

Team Members: Anders Kvalsvik, Can Yi, Jackson Conrad, Parith Reddy, Yuehao Yang

9/30/2025

1 Data Overview – Training & test sets

- **323 QA pairs** on AI sustainability & efficiency metrics
- Split: **41 training**, 282 testing
- Each entry includes:
  - id, question, answer
  - answer_value, answer_unit
  - ref_url, ref_id
  - supporting_materials, explanation

1 Data Overview – Training & test sets

Field Name	Explanation	Sample
id	Unique identifier for the QA pair	q139; q102;
question	The natural language query	As of 2023, what was the water use effectiveness (WUE) for AWS data centers, in L/kWh?; True or False: The AI Act makes energy consumption data from providers publicly available to NGOs, analysts, and the general public;
answer	The direct answer to the question	0.18 L/kWh; FALSE;
answer_value	The numerical value extracted from the answer	0.18; 0;
answer_unit	The unit of measurement for the answer value	L/kWh; is_blank;
ref_id	Citation ID linking to the source document	['amazon2023']; ['ebert2024'];
ref_url	URL pointing to the full source document	['https://sustainability.aboutamazon.com/2023-amazon-sustainability-report.pdf']; ['https://arxiv.org/pdf/2410.06681'];
supporting_materials	Specific references within the source	0.18 Liters of water per kilowatt-hour (L/kWh) water use effectiveness (WUE) for AWS data centers; Section 4.3 Transparency: 'Where the Act does mandate disclosure... this information is restricted to authorities and is not accessible to downstream providers... or the general public.';
explanation	Justification for how the answer was derived from the source	Quote; Quote;

1 Data Overview – Metadata

- **33 documents** in total
  - 32 academic papers
  - +1 Amazon Sustainability Report 2023
- Including:
  - Extensive textual information
  - Multimodal data (*figures, tables, images*)

2 Key Steps – QA Pairs Exploration, Question Type

Question Type	Description	Example from Dataset	Count	Percentage
Quantitative	Questions asking for a specific numerical value, measurement, or statistic. Often start with "How much", "How many", "What percentage", or "By what factor".	q124: What is the estimated total operational water consumption for training GPT-3...?	158	48.90%
Verification (True/False)	Questions that require a binary verification of a given statement.	q053: True or False: Operational environmental impacts of LLMs do not include GHG emissions that arise from servers and data centers using cooling.	42	13.00%
Factual / Definition	Questions asking for a specific term, name, concept, or definition.	q282: What is the term for the amount of water evaporated, transpired, or incorporated into products...?	40	12.40%
Comparative	Questions that explicitly request a comparison between entities (models, metrics, time periods) or ask for the "highest", "lowest", or "most efficient" option.	q166: Which of the following five large NLP DNNs has the highest energy consumption...?	36	11.10%
Procedural / Methodological	Questions about the methods, frameworks, software, or tools used in a process.	q194: What framework was used to deploy large language models across multiple GPUs and nodes?	26	8.00%
Out-of-Scope / Unanswerable	Questions that are completely unrelated to the domain of the knowledge base (AI sustainability, models, data centers) and cannot be answered by the provided documents.	q164: How much does an elephant weigh? q079: How many miles is the Earth from the Sun?	13	4.00%
Causal	Questions exploring the cause-effect relationship or the reason behind a phenomenon.	(Less common in this dataset; an inferred example would be "Why does model inference consume less power than training?")	6	1.90%
Multi-step / Calculation	Questions that require combining multiple pieces of numerical information from the context to perform a calculation.	q066: ...estimate the daily energy consumption in MWh. (Requires using the provided queries/day and kWh/query rates)	2	0.60%
Total			**323**	**100.00%**

2 Key Steps – Metadata Exploration, Document Topic Analysis

1. AI Environmental Impact & Sustainability

- Societal and environmental impact of digital technologies
- Carbon emissions, energy consumption, and water footprint of AI training and inference
- Green AI and efficient algorithm design
- Sustainable development frameworks and policy recommendations

2. Model Efficiency & Cost Optimization

- Quantification of energy and financial costs for training and inference
- Model compression (e.g., quantization, sparsification), hardware-software co-design
- Low-cost model training (e.g., FLM-101B, JetMoE)
- Cloud-based distributed training and resource scheduling strategies

3. Challenges Specific to Large Language Models (LLMs)

- Performance and energy analysis of LLM inference
- Impact of prompt engineering on energy efficiency
- Model architecture optimization (e.g., Mixture of Experts, attention mechanism improvements)
- Decoding strategies and cache management

4. Ethics & Social Responsibility

- AI ethics and fairness (e.g., bias, privacy, transparency)
- Technology accessibility and inclusivity (e.g., unequal resource distribution)
- Environmental justice and public health burdens
- Corporate sustainability reporting and policy compliance

5. Policy & Governance

- AI regulations (e.g., EU AI Act)
- Standardization and disclosure requirements for environmental data
- Multi-stakeholder collaborative governance models
- Addressing Jevons Paradox (where efficiency gains lead to increased consumption)

6. Cross-domain Applications & Future Prospects

- Potential impact of AI in domains like transportation, healthcare, and education
- Linking technological development with UN Sustainable Development Goals (SDGs)
- Long-term predictions (e.g., life scenarios in 2030)

2 Key Steps – Training Set Exploration, Supporting Material Location

Location Category	Count	Percentage
Main Body / Section	20	48.80%
Table	8	19.50%
Figure	5	12.20%
Abstract	3	7.30%
Introduction / Background	3	7.30%
Appendix	1	2.40%
Conclusion / Recommendations	1	2.40%
Total	41	100%

2 Key Steps – Data Preprocess Pipeline

1. Text Extraction

- Extracts raw text from each page using the LlamaParse library.
- Fills JSON files with extracted text, metadata, and document id

2. Text Chunking

- Segments PDF content into analysis-ready text chunks.
- Splits the text from each valid page into chunks of approximately 220 words*
- Consolidates all text chunks and their source metadata (ID, title, page number) into a single, efficient Parquet file.

*We will experiment using various other chunking methods later to find the optimal method

3 Useful Tools

Exploratory Data Analysis

- **Ima.copilot & DeepSeek V3.1:** Used to deeply analyze QA samples and document content to obtain insights.
- **Gemini 2.5 Pro:** Helped calculate statistics and create tables.

Page Extraction

- **LlamaParse**: Used to open **PDFs** and extract raw text from each page.
- **pandas**: Used to create a **DataFrame** and save the page number and text length to a **CSV** file.
- **tqdm**: A utility used to display a progress bar during processing.

Encoding Standardization

- **pandas**: Used to read CSV files with various encodings and re-save them in the standard **UTF-8** format.

Text Chunking

- **pandas**: Used to load and merge source data, then to store the final text chunks into a single **Parquet** file.
- **pymupdf**: Used to extract page text from the **PDFs** before the chunking process.
- **re**: Python's built-in library used to split text into paragraphs via regular expressions.

4 Challenges

Small Training Set → Limits Generalization

```
RangeIndex: 32 entries, 0 to 31
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   id        32 non-null     object
 1   type      32 non-null     object
 2   title     32 non-null     object
 3   year      32 non-null     int64
 4   citation  32 non-null     object
 5   url       32 non-null     object
dtypes: int64(1), object(5)
memory usage: 1.6+ KB
```

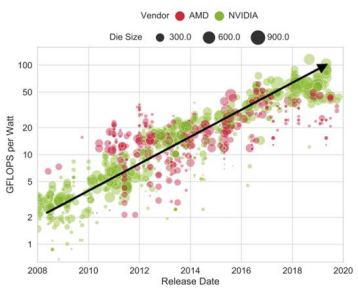
We currently have only about 32 papers as our corpus.

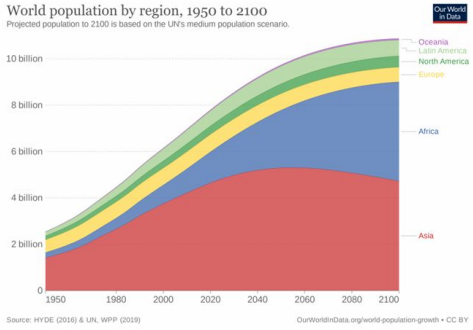
Because the corpus is small and the coverage is limited, new queries may not find relevant information in the database → leading to insufficient retrieval recall.

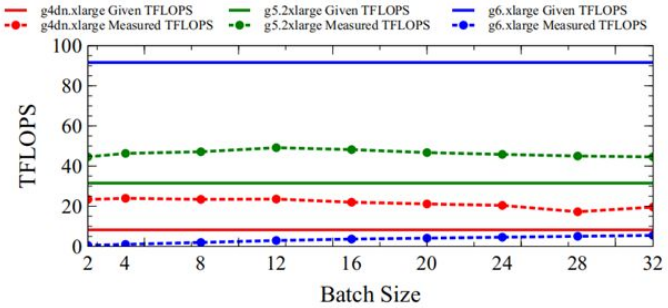
Tables & charts: unstructured, hard to parse

Tables are often not standardized in format, making them difficult to parse automatically.

Many figures lack numeric labels and only show curves or bars → since we do not have access to the original data sources, it is uncertain whether the extracted values are accurate.







5 Baseline Model

Planned Method:

- **all-MiniLM-L6-v2 → embeddings**
- **FAISS → vector database for retrieval**
- **GPT-3.5 → reader / answer generation**

Expected Performance:

- **Should handle single-paragraph questions reasonably well**
- **Likely to struggle with cross-file reasoning and unstructured data (tables, figures)**
- **Retrieval may be too broad at times → risk of including irrelevant context**

6 Next Steps

- **Stronger embeddings**

Explore larger embedding models to improve retrieval accuracy and reduce irrelevant matches.

- **Cross-file reasoning**

Implement multi-hop QA so the system can combine evidence from multiple documents, rather than relying on single-passage retrieval.

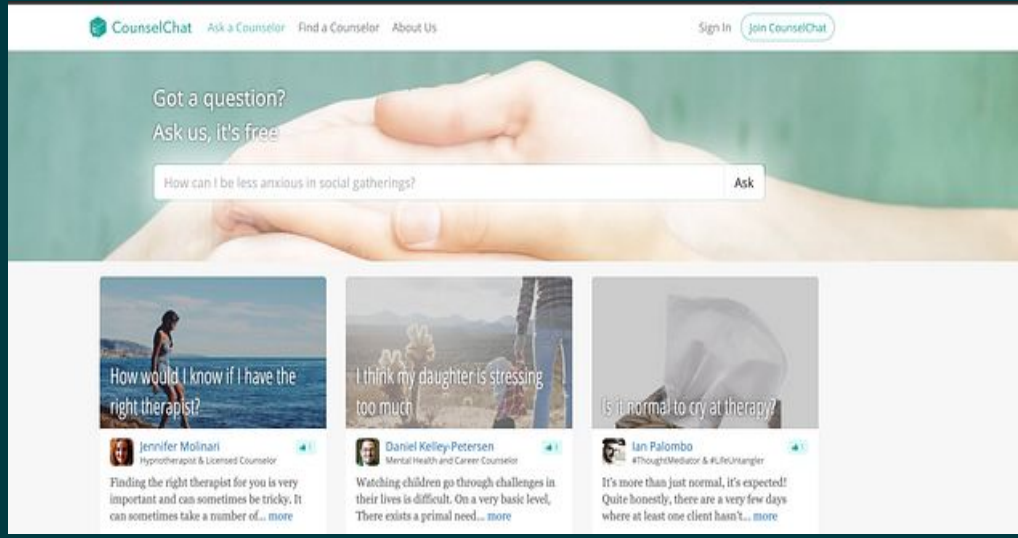
Presented by
Khine Thant Su
and Alexandra Wong

CRISIS COMPANION: EDA

TEAM: DR.CHATBOT

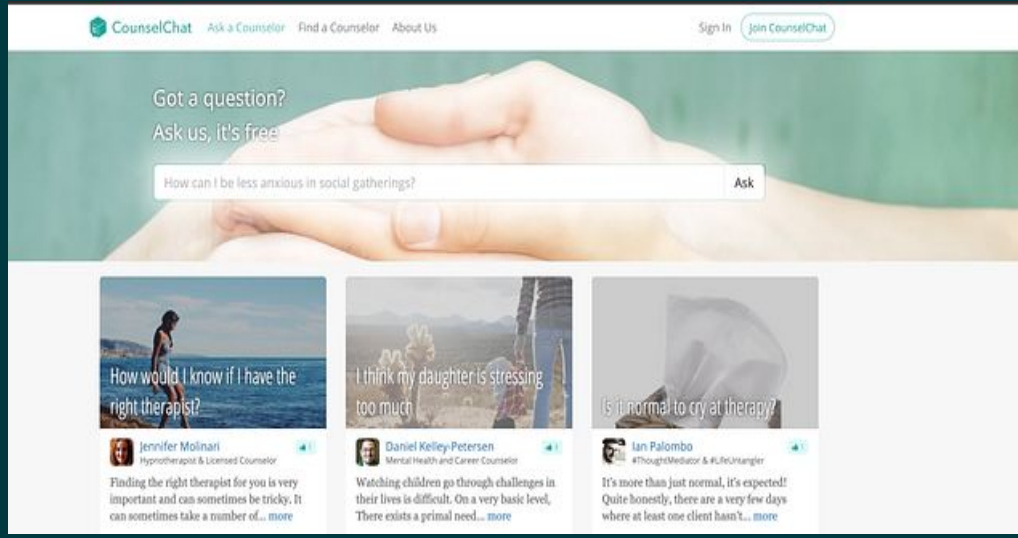
* NOT A LICENSED DOCTOR

The data



questionID	questionTitle	questionText	questionLink	topic	therapistInfo	therapistURL	answerText	upvotes	views
0	Do I have too many issues for counseling?	I have so many issues to address. I have a his...	https://counselchat.com/questions/do-i-have-to...	depression	Jennifer MolinariHypnotherapist & Licensed Cou...	https://counselchat.com/therapists/jennifer-mo...	It is very common for people to have multiple ...	3	1971
0	Do I have too many issues for counseling?	I have so many issues to address. I have a his...	https://counselchat.com/questions/do-i-have-to...	depression	Jason Lynch, MS, LMHC, LCAC, ADSIndividual & C...	https://counselchat.com/therapists/jason-lynch...	I've never heard of someone having "too many i...	2	386
0	Do I have too many issues for counseling?	I have so many issues to address. I have a his...	https://counselchat.com/questions/do-i-have-to...	depression	Shakeeta TorresFaith Based Mental Health Couns...	https://counselchat.com/therapists/shakeeta-to...	Absolutely not. I strongly recommending worki...	2	3071

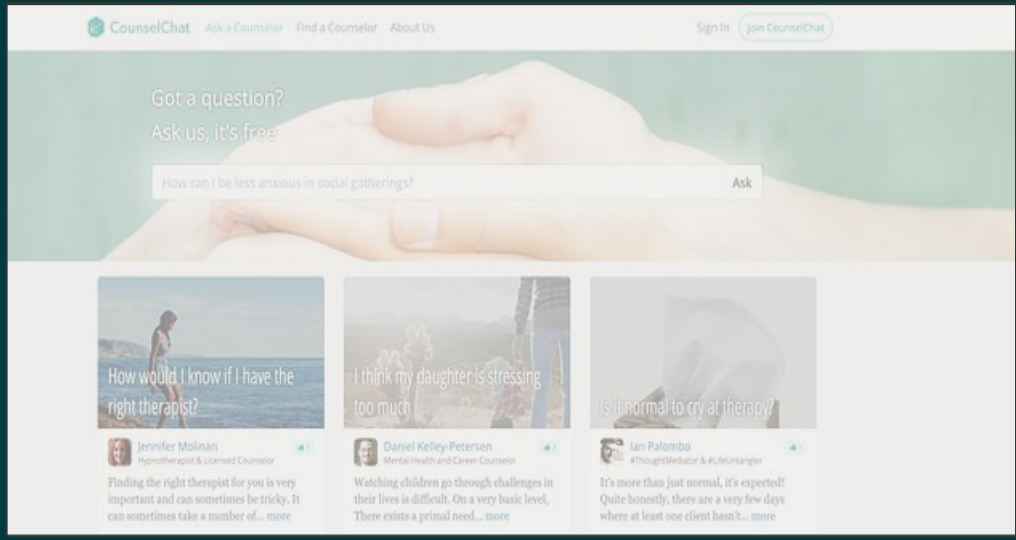
The data



questionID	questionTitle	questionText	questionLink	topic	therapistInfo	therapistURL	answerText	upvotes	views
0	Do I have too many issues for counseling?	I have so many issues to address. I have a his...	https://counselchat.com/questions/do-i-have-to...	depression	Jennifer MolinariHypnotherapist & Licensed Cou...	https://counselchat.com/therapists/jennifer-mo...	It is very common for people to have multiple ...	3	1971
0	Do I have too many issues for counseling?	I have so many issues to address. I have a his...	https://counselchat.com/questions/do-i-have-to...	depression	Jason Lynch, MS, LMHC, LCAC, ADSIndividual & C...	https://counselchat.com/therapists/jason-lynch...	I've never heard of someone having "too many i...	2	386
0	Do I have too many issues for counseling?	I have so many issues to address. I have a his...	https://counselchat.com/questions/do-i-have-to...	depression	Shakeeta TorresFaith Based Mental Health Couns...	https://counselchat.com/therapists/shakeeta-to...	Absolutely not. I strongly recommending worki...	2	3071

The rationale

- Answers are from licensed therapists
- Each post comes categorized into a topic, like depression, anxiety, intimacy
- “Upvotes” can be used as feedback signal for how helpful an answer is



questionID	questionTitle	questionText	questionLink	topic	therapistInfo	therapistURL	answerText	upvotes	views
0	Do I have too many issues for counseling?	I have so many issues to address. I have a his...	https://counselchat.com/questions/do-i-have-to...	depression	Jennifer MolinariHypnotherapist & Licensed Cou...	https://counselchat.com/therapists/jennifer-mo...	It is very common for people to have multiple ...	3	1971
0	Do I have too many issues for counseling?	I have so many issues to address. I have a his...	https://counselchat.com/questions/do-i-have-to...	depression	Jason Lynch, MS, LMHC, LCAC, ADSIndividual & C...	https://counselchat.com/therapists/jason-lynch...	I've never heard of someone having "too many i...	2	386
0	Do I have too many issues for counseling?	I have so many issues to address. I have a his...	https://counselchat.com/questions/do-i-have-to...	depression	Shakeeta TorresFaith Based Mental Health Couns...	https://counselchat.com/therapists/shakeeta-to...	Absolutely not. I strongly recommending worki...	2	3071

Example question and response from the dataset

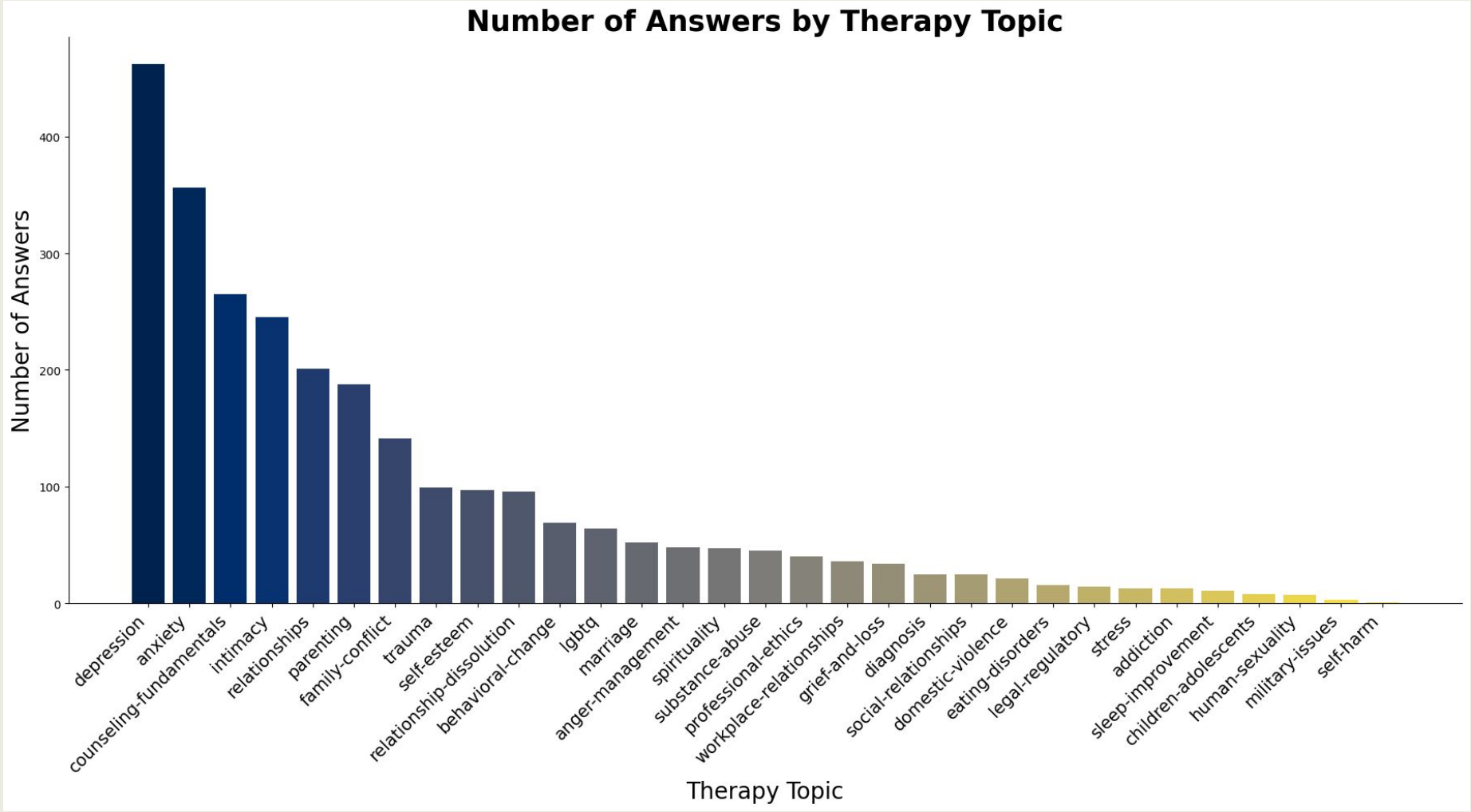
Question:
I have bipolar II disorder, I'm addicted to alcohol and weed, and I'm hopeless. I keep drinking even though it's harming myself and others.

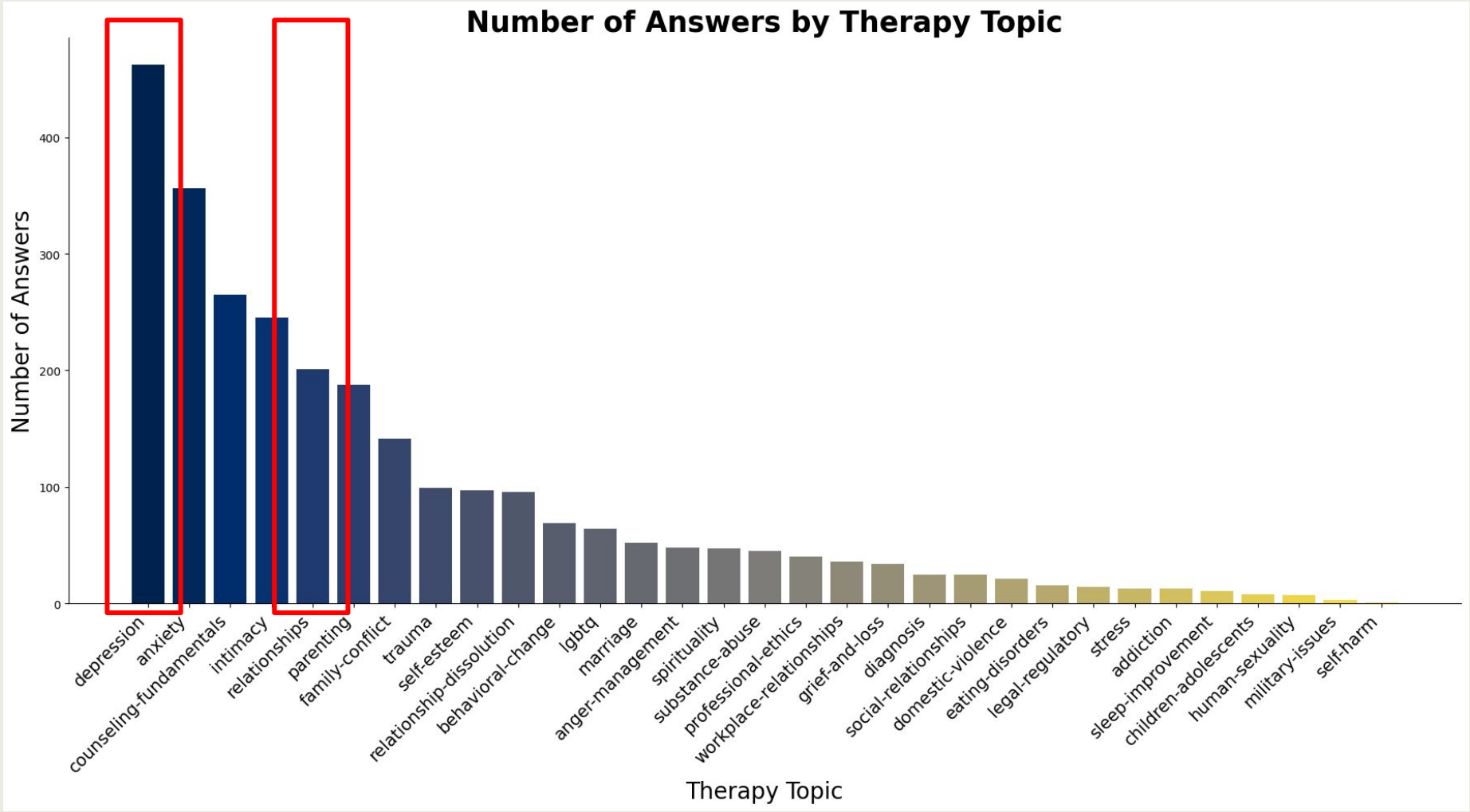
Answer:
What an important question. I'm hearing your hopelessness and fear about the damage of your substance use and it sounds like you have reason for concern. While there can be negative stigma about using substances, for the most part substance use is an attempt to cope with emotional distress in the absence of sufficient coping strategies. We all look for comfort when we are in pain and this may be the way that you are getting comfort - even though it is also hurting you. In order to stop using alcohol and weed you will need a lot of support and you will need to learn other ways of getting comfort when you are in pain or struggling with bipolar related symptoms. There is nothing to be ashamed of and we all need help when we are struggling. I would encourage you to reach out for support in any way that you can. Have you talked with your health care providers about your concerns or friends or family members? Here are some links of resources in Whistler that may be helpful: http://www.vch.ca/locations-and-services/find-health-services/?program_id=11035http://redbookonline.bc211.ca/service/9509054_9509054/whistler_mental_health_and_addictionshttp://mywcss.org/programs/counselling-assistance/http://redbookonline.bc211.ca/organization/9489472/alcoholics_anonymous_aa___squamishwhistler

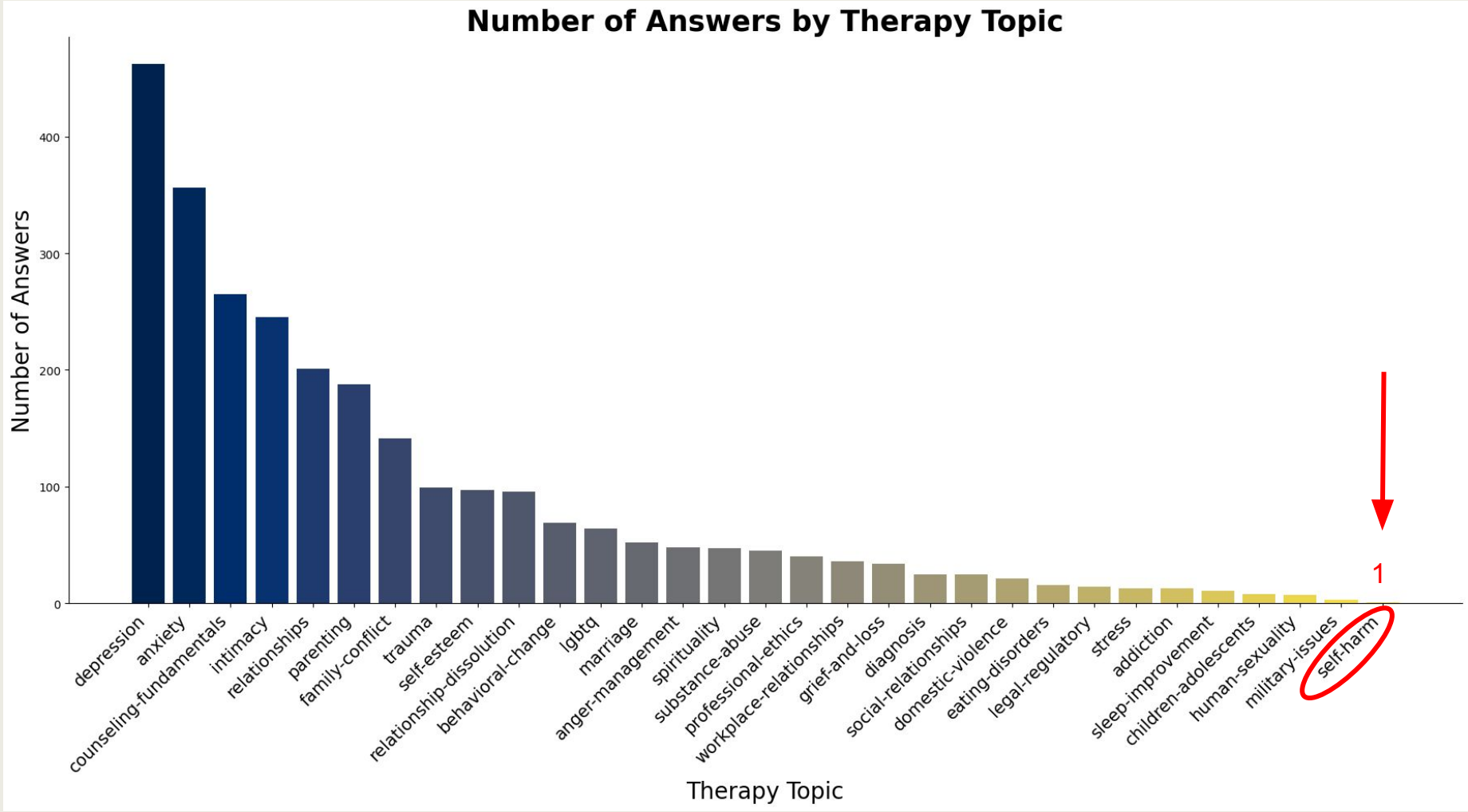
Dataset description

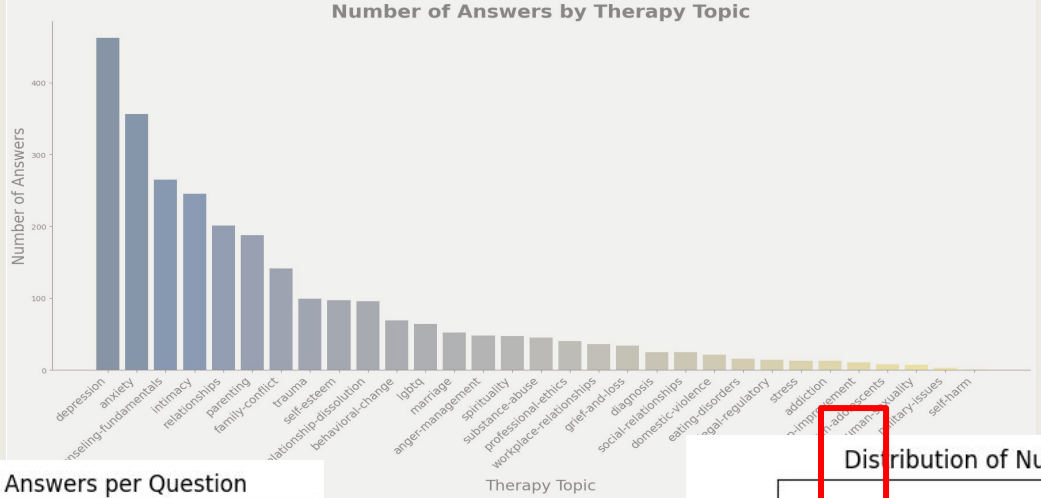
- Unique therapy questions: 933
- Unique therapy responses: 2739
- Number of questions with no responses: 24
- Unique number of therapists contributing to responses: 437
- Number of therapy topics discussed: 31
- Average question character length: 49
- Average answer character length: 960

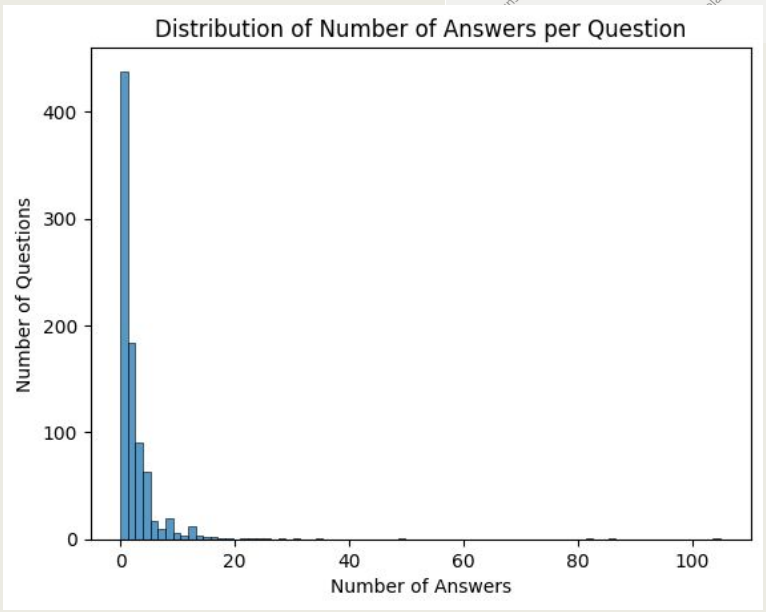
```
Topics in this dataset:
['depression',
 'anxiety',
 'parenting',
 'self-esteem',
 'relationship-dissolution',
 'workplace-relationships',
 'spirituality',
 'trauma',
 'domestic-violence',
 'anger-management',
 'sleep-improvement',
 'intimacy',
 'grief-and-loss',
 'substance-abuse',
 'family-conflict',
 'marriage',
 'eating-disorders',
 'relationships',
 'lgbtq',
 'behavioral-change',
```


Number of Answers by Therapy Topic
Number of Answers
0
100
200
300
400
depression
anxiety
counseling-fundamentals
intimacy
relationships
parenting
family-conflict
trauma
self-esteem
relationship-dissolution
behavioral-change
lgbtq
marriage
anger-management
spirituality
substance-abuse
professional-ethics
workplace-relationships
grief-and-loss
diagnosis
social-relationships
domestic-violence
eating-disorders
legal-regulatory
stress
addiction
sleep-improvement
children-adolescents
human-sexuality
military-issues
self-harm
Therapy Topic

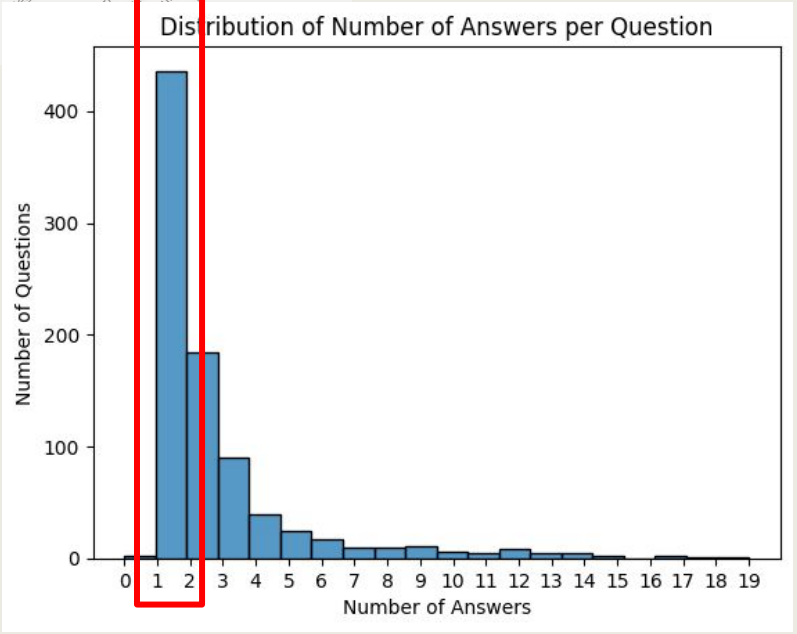

Number of Answers by Therapy Topic
Number of Answers
0
100
200
300
400
depression
anxiety
counseling-fundamentals
intimacy
relationships
parenting
family-conflict
trauma
self-esteem
relationship-dissolution
behavioral-change
lgbtq
marriage
anger-management
spirituality
substance-abuse
professional-ethics
workplace-relationships
grief-and-loss
diagnosis
social-relationships
domestic-violence
eating-disorders
legal-regulatory
stress
addiction
sleep-improvement
children-adolescents
human-sexuality
military-issues
self-harm
Therapy Topic

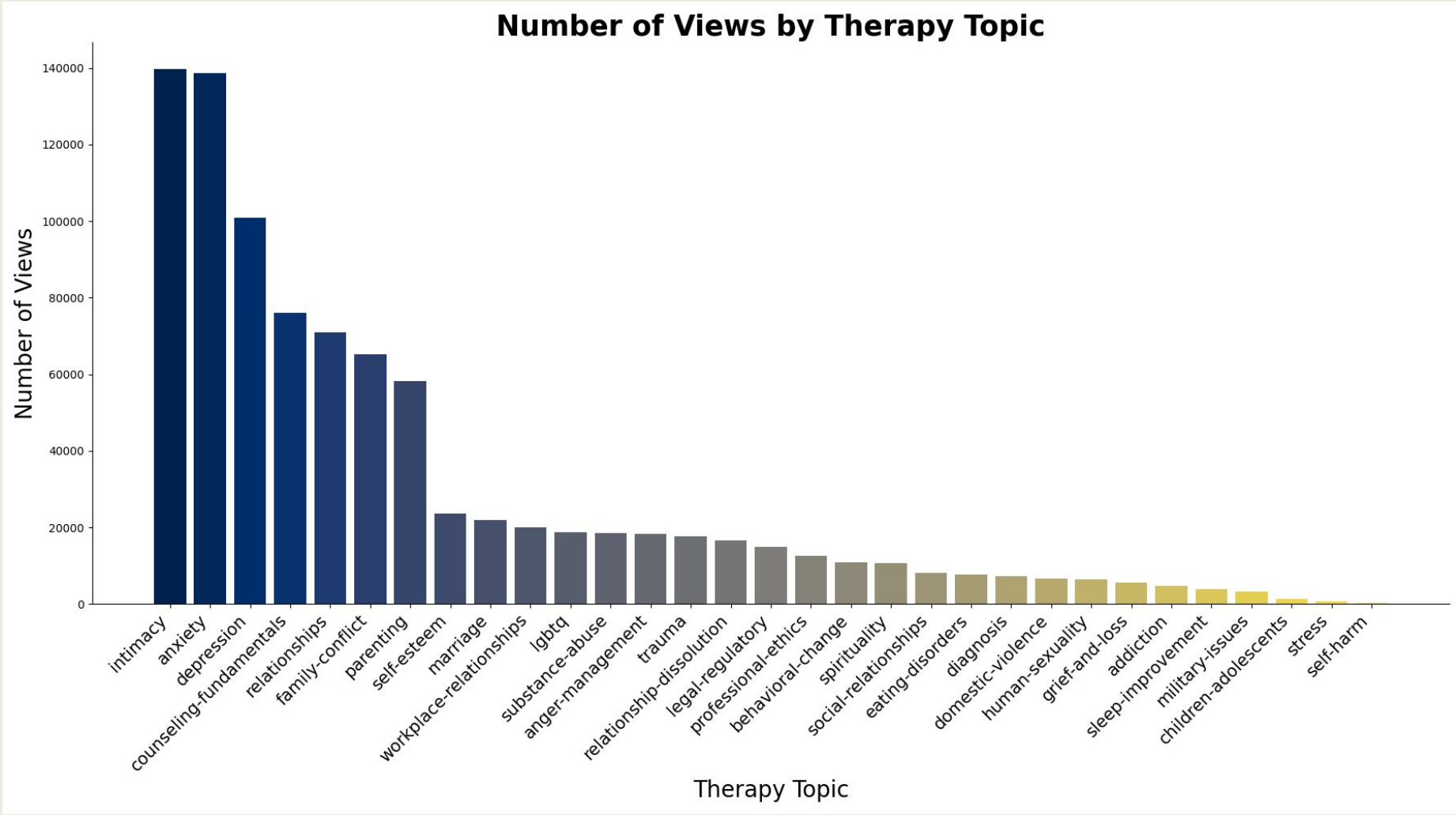

Number of Answers by Therapy Topic
Number of Answers
0
100
200
300
400
1
depression
anxiety
counseling-fundamentals
intimacy
relationships
parenting
family-conflict
trauma
self-esteem
relationship-dissolution
behavioral-change
lgbtq
marriage
anger-management
spirituality
substance-abuse
professional-ethics
workplace-relationships
grief-and-loss
diagnosis
social-relationships
domestic-violence
eating-disorders
legal-regulatory
stress
addiction
sleep-improvement
children-adolescents
human-sexuality
military-issues
self-harm
Therapy Topic


Number of Answers by Therapy Topic
Number of Answers
0
100
200
300
400
depression
anxiety
intimacy
relationships
parenting
family-conflict
trauma
self-esteem
relationship-dissolution
behavioral-change
lgbtq
marriage
anger-management
spirituality
substance-abuse
professional-ethics
workplace-relationships
grief-and-loss
diagnosis
social-relationships
domestic-violence
eating-disorders
stress
addiction
self-harm
Therapy Topic

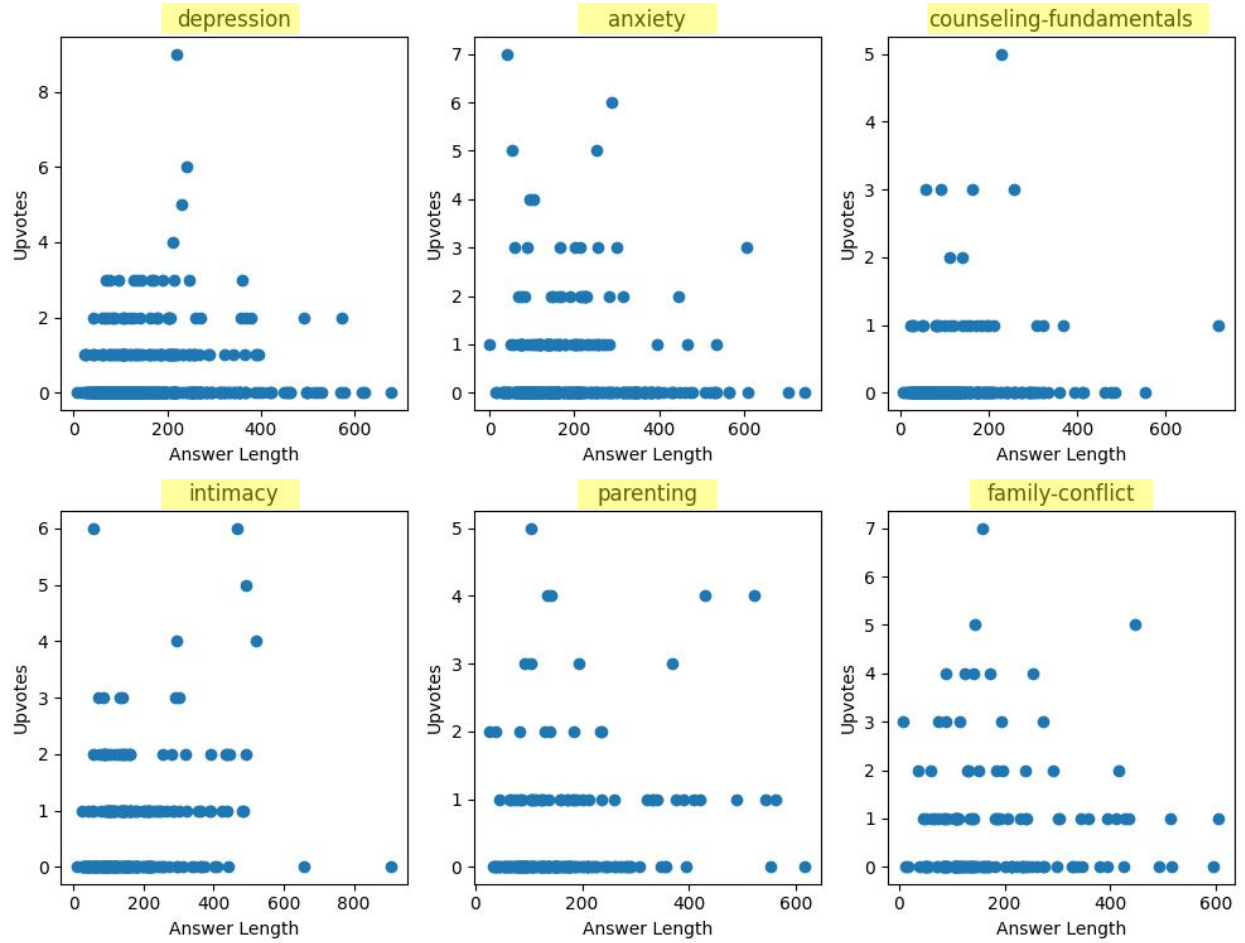

Distribution of Number of Answers per Question
Number of Questions
0
100
200
300
400
0
20
40
60
80
100
Number of Answers

Zoomed in

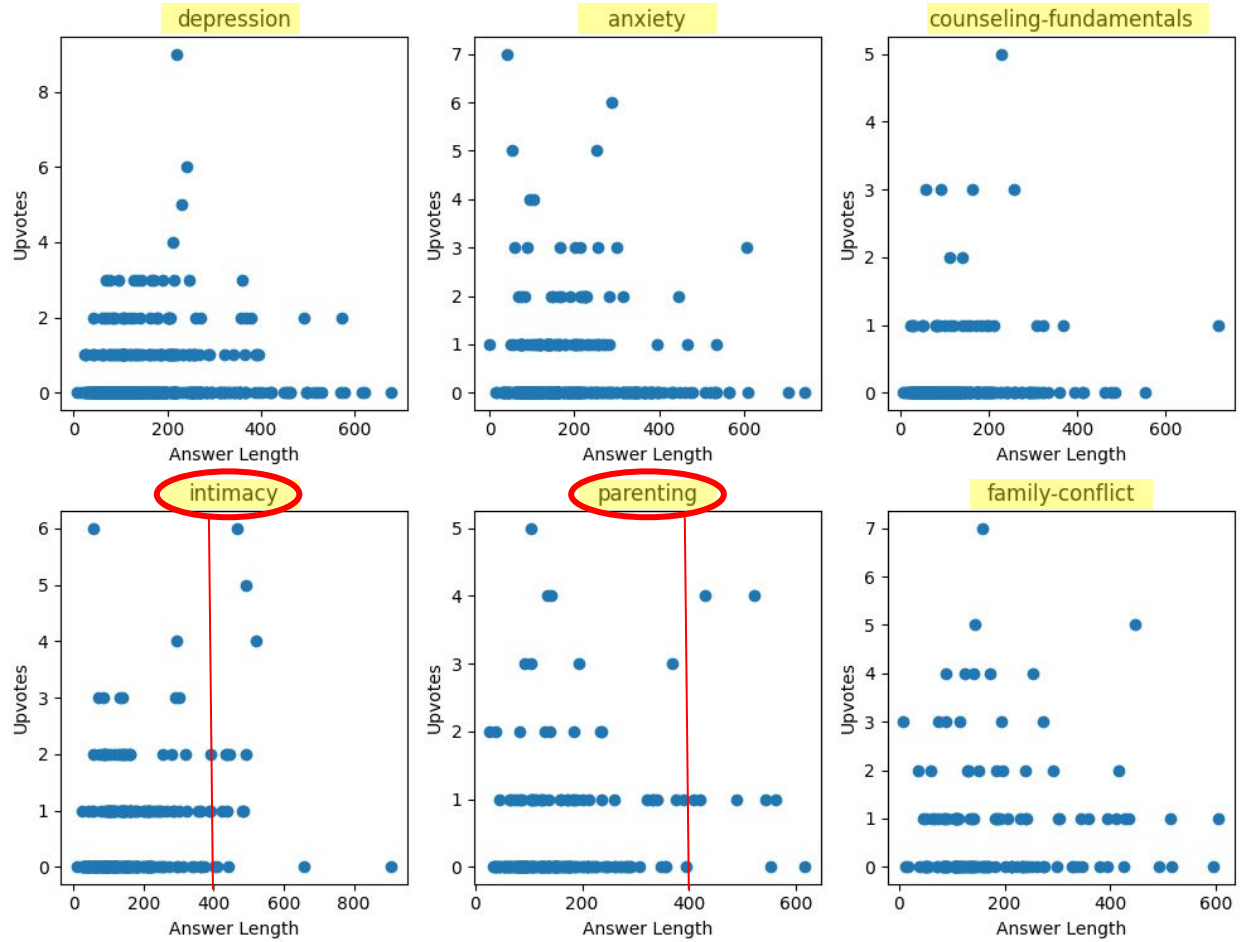

Distribution of Number of Answers per Question
Number of Questions
0
100
200
300
400
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
Number of Answers


Number of Views by Therapy Topic
Number of Views
0
20000
40000
60000
80000
100000
120000
140000
intimacy
anxiety
depression
counseling-fundamentals
relationships
family-conflict
parenting
self-esteem
marriage
workplace-relationships
lgbtq
substance-abuse
anger-management
trauma
relationship-dissolution
legal-regulatory
professional-ethics
behavioral-change
spirituality
social-relationships
eating-disorders
diagnosis
domestic-violence
human-sexuality
grief-and-loss
addiction
sleep-improvement
military-issues
children-adolescents
stress
self-harm
Therapy Topic

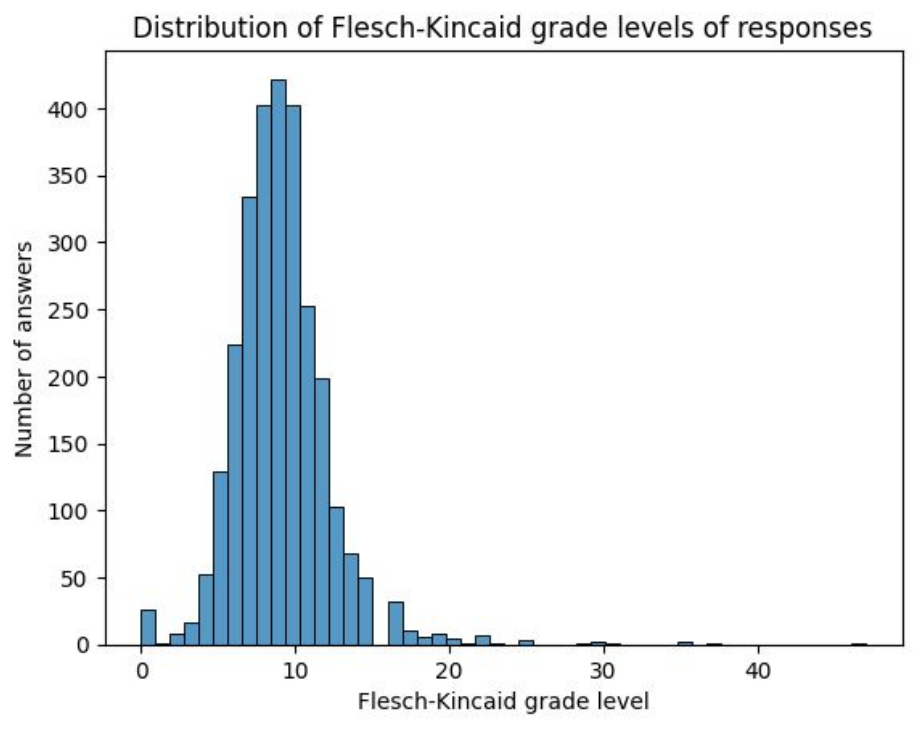
Shorter answer length (in words) is associated with more upvotes



Shorter answer length (in words) is associated with more upvotes



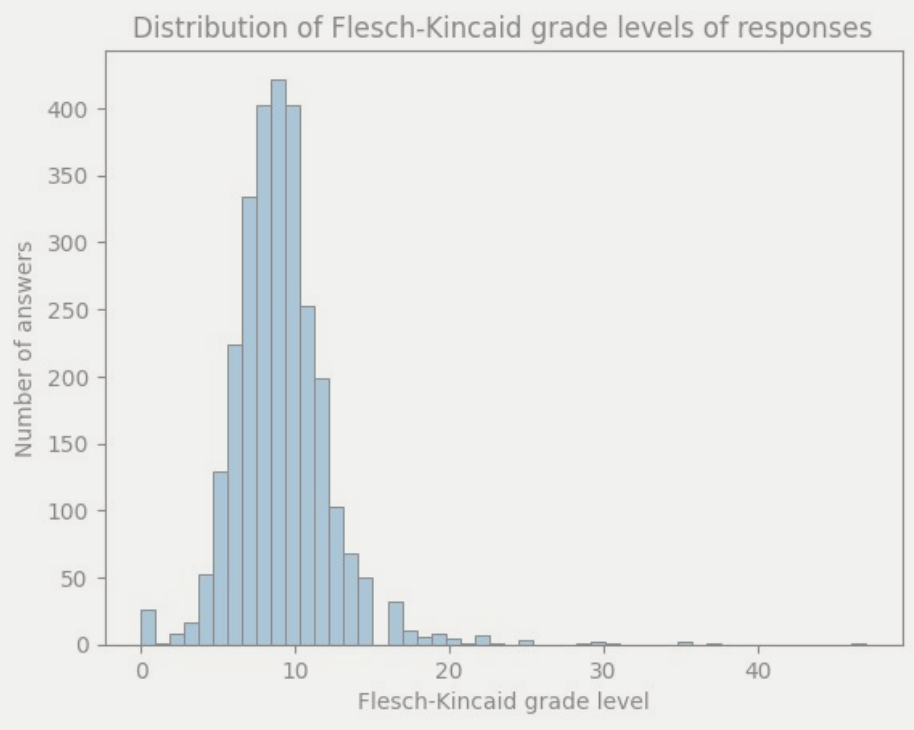
How readable are the responses?

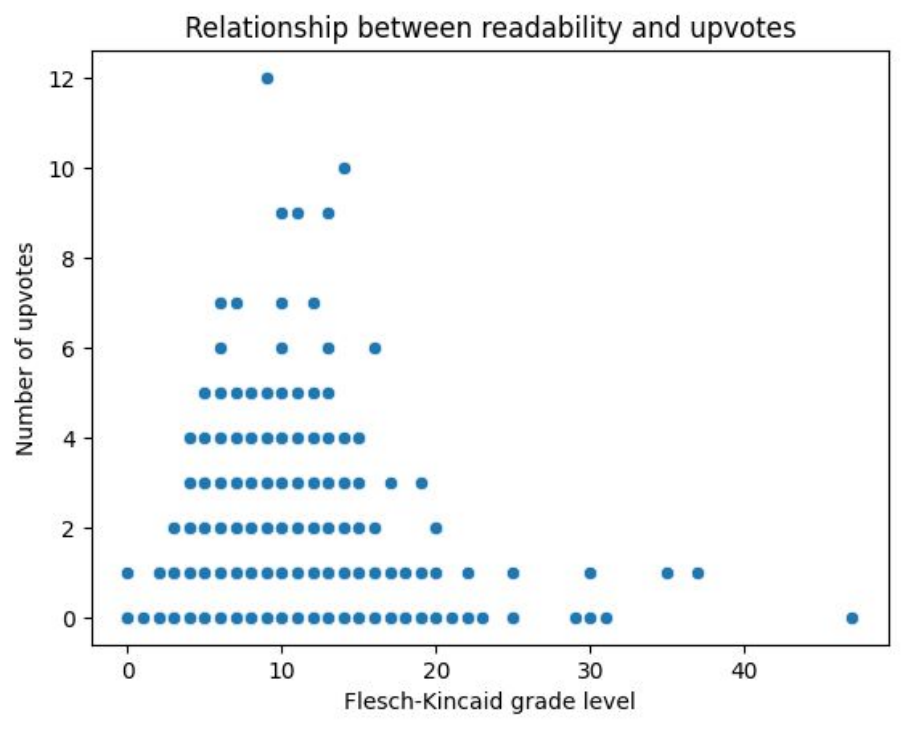


Flesch-Kincaid grade level roughly corresponds to a U.S. grade level.

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right)+11.8\left(\frac{\text{total syllables}}{\text{total words}}\right)-15.59$$

Do more readable responses get more upvotes?

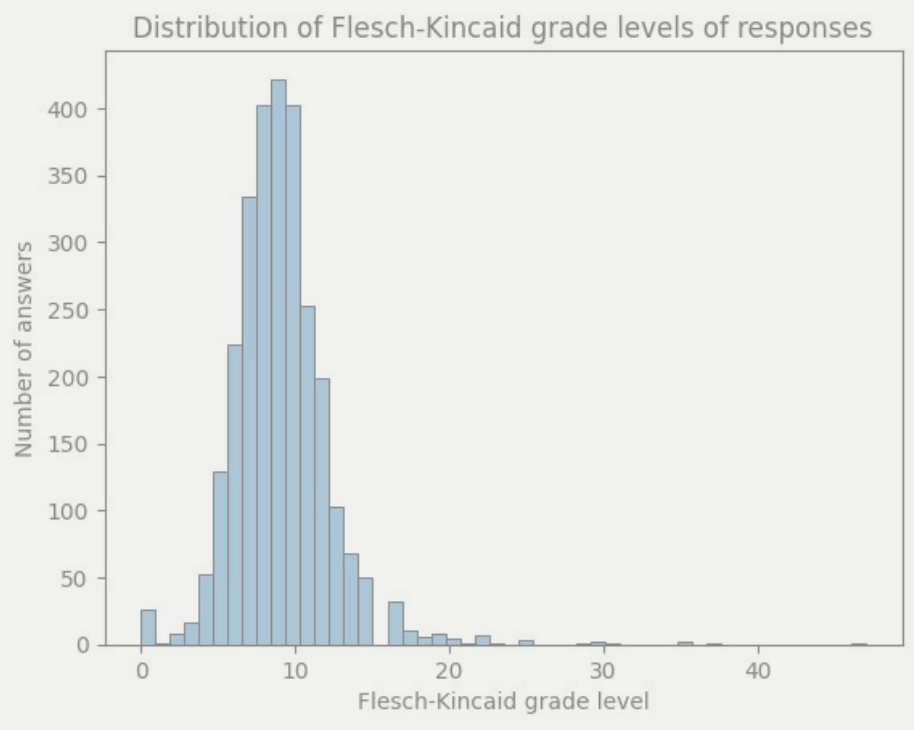


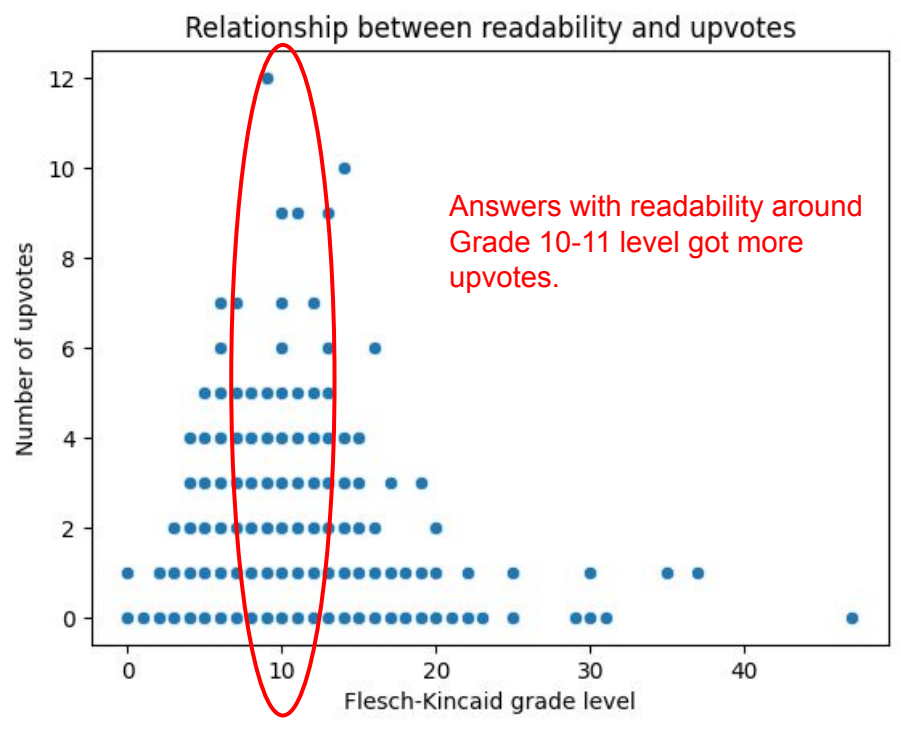


Flesch-Kincaid grade level roughly corresponds with a U.S. grade level.

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right)+11.8\left(\frac{\text{total syllables}}{\text{total words}}\right)-15.59$$

Do more readable responses get more upvotes?





Flesch-Kincaid grade level roughly corresponds with a U.S. grade level.

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right)+11.8\left(\frac{\text{total syllables}}{\text{total words}}\right)-15.59$$

Next steps

- Use this CounselChat data as an instruction tuning dataset to fine-tune a base large language model.

- Base LLM that we're considering:
  - QWEN2-7B-Instruct
    - Has good benchmarks for language understanding & generation
    - Smaller model (less than 10B) suitable for instruction tuning
    - Can operate efficiently in resource-constrained environments

- We will access this model through Hugging Face using the Transformers library.

Slides: go.wisc.edu/7w28y

EDA Presentations on 10/2, 5:30-7:30pm

- With ~20 teams, each team will have 4 minutes to present
  1. **Introduce your data**: Briefly explain what kind of data you're working with.
  2. **Highlight key steps**: Mention the most important data exploration or cleaning actions you've taken.
  3. **Useful tools/packages/functions**: Mention any useful tools/libraries you used for your analysis
  4. **Share insights**: Discuss any early patterns or challenges you've discovered so far.
  5. **Baseline model**: Discuss results of baseline model. How hard is the task?
  6. **Next steps**: Include ideas for next steps

- Send google slides link (5 slides max) by **9/30, 11:59pm** to endemann@wisc.edu.
  - **Format slides as Widescreen 16:9** (file -> page setup)
  - **Synced slides**: Slides can be *polished* up until presentation on 10/2. However...
    - No rearranging, adding, or removing slides after 9/30. These changes will not sync!

Cache Us If You Can

Sanya Gupta, Xixi Liu, Veda Poranki, Tanya Devireddy

Introducing Store Sales Dataset

train.csv: data used to train the model

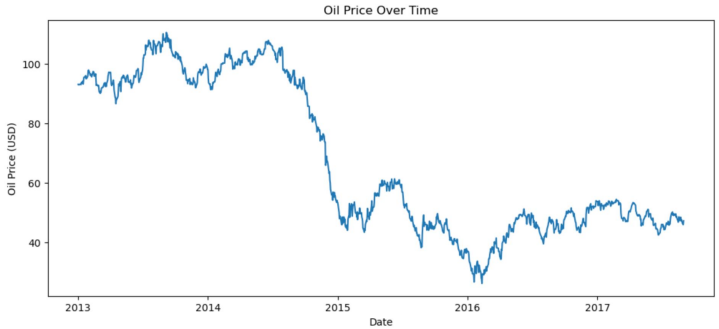
test.csv: data used to test the final model

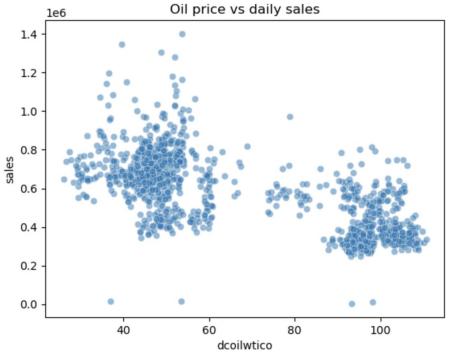
stores.csv: includes metadata like city, state, type and cluster of stores

oil.csv: oil dependent economy makes the country volatile to oil prices fluctuations, affecting transportation

holiday_events.csv: holiday season creates an influx in customers, affecting sales data

Oil Prices vs. Effects on Sales



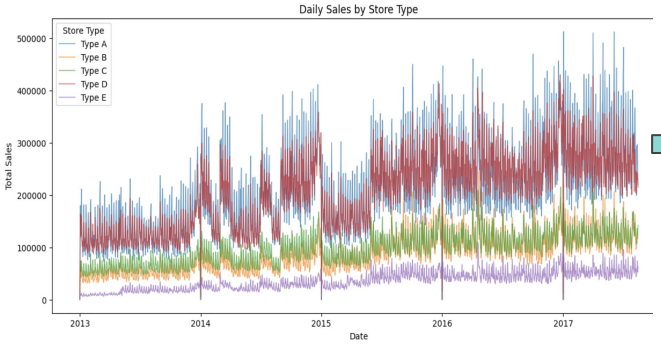


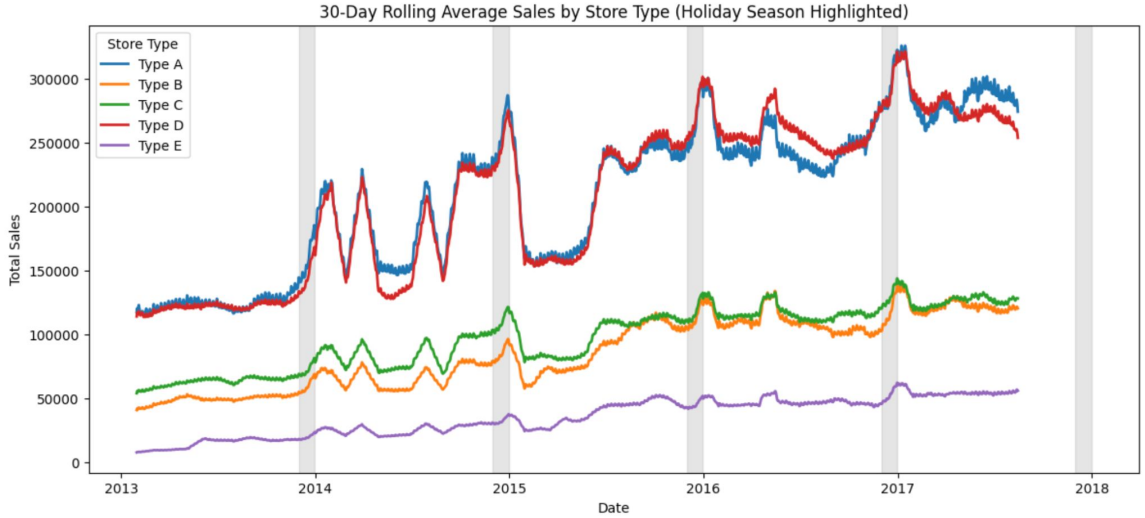
- Cleaned oil dataset
  - 43 NaN values due to non price changes on weekends and holidays
  - changed date from object string to datetime and sorted
- Merged Daily Sales with Oil Price
- Visualized using Seaborn and Matplotlib'

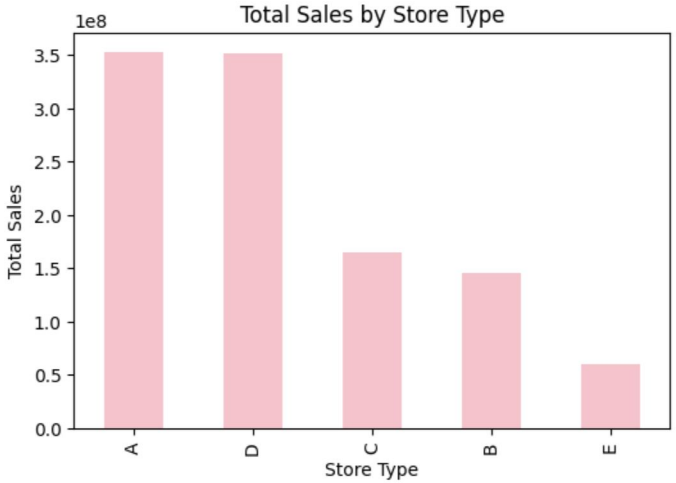
CONCLUSIONS

- As oil prices rise, sales decrease

Store Type vs. Performance

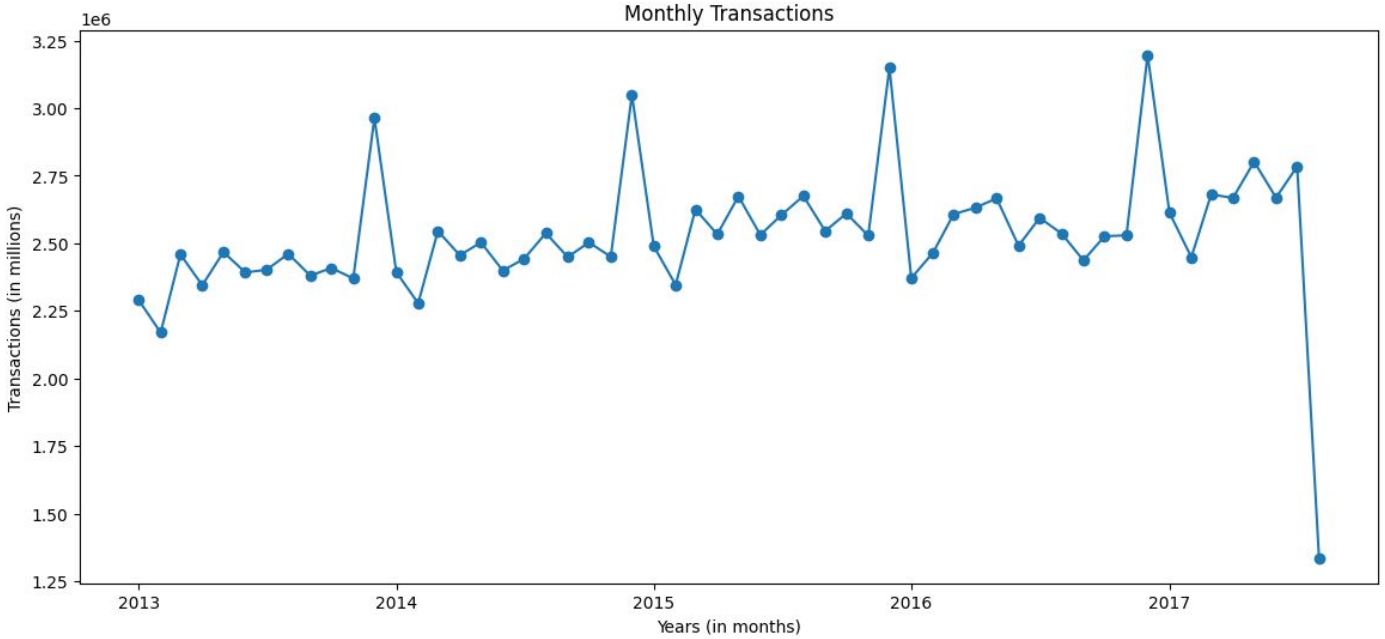






- First graph - raw representation of daily sales vs date.
- Second graph - to see the trend more clearly, we used 30-day rolling averages for each store instead of daily sales
- Third graph - shows sales in by store

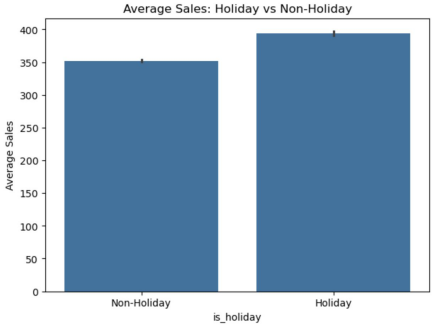
Seasons vs. Transactions

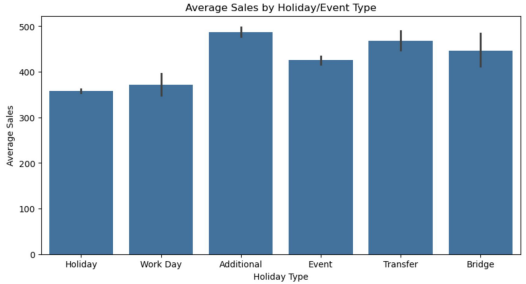


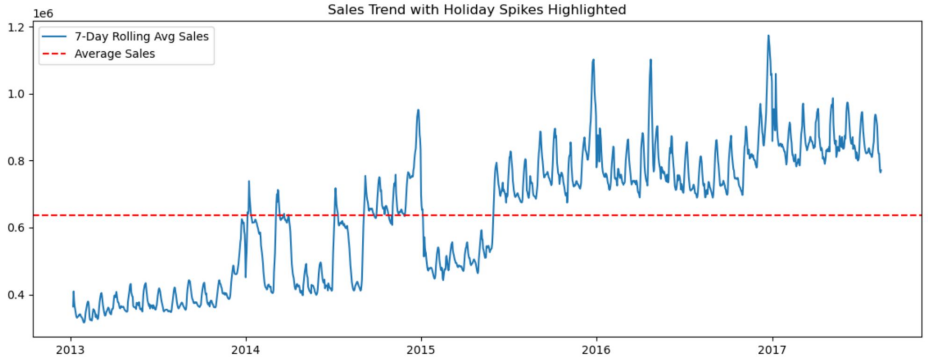
- No cleaning needed due to no null values
- Used transactions and date column of transactions.csv to see relationship of sales over time
- Visualized with matplotlib, pyplot and date modules

- Can see spikes during late months of the year like December most likely due to the holiday season

Holiday vs. Sales







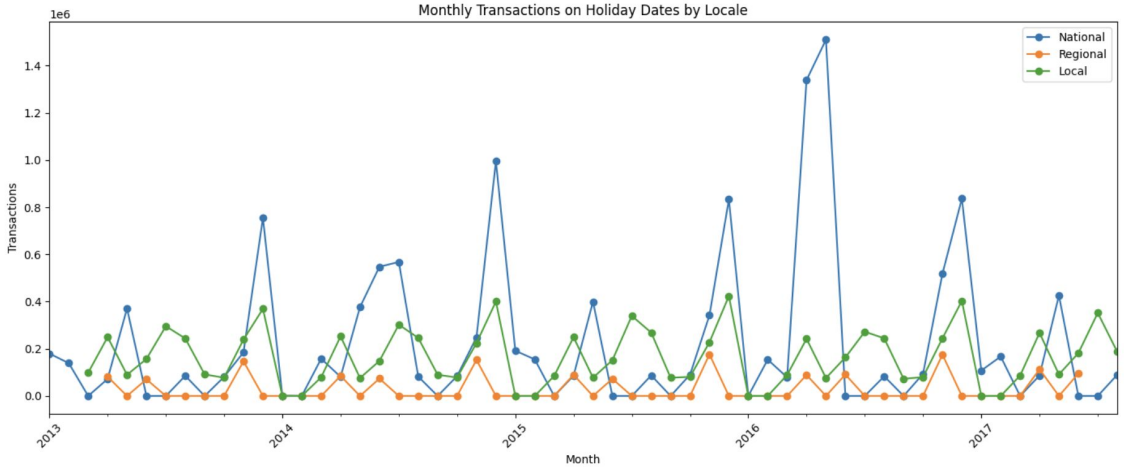
Findings (Top 2 Graphs)

- National Holidays in Ecuador show highest sales spikes
- Additional Holidays columns unpredictable sales
- People shop more on bridge and transfer holidays

Bottom Graph

- There are clear recurring spikes yearly
- Sales return to baseline quickly after spikes

Holidays vs. Transactions



- Cleaned holiday + transactions dataset
  - Converted date fields to datetime
  - Merged holiday flags (National, Regional, Local) with daily sales
- Aggregated daily sales to monthly totals
- Split holiday impacts by National, Regional, and Local
- Visualized trends using Matplotlib

CONCLUSIONS:

- National holidays show the strongest spikes in sales
- Local holidays are frequent but produce smaller shifts
- Regional has less impact on transactions

Baseline Model

Because we do not have much ML experience, we had to do a lot of research about models and Python packages and kept our baseline very simple...

We predict next week's sales will be the same as this week’s.

This will be our basepoint and what we plan on beating with our model.

Next Steps

- Decide model
  - Light-GBM (Light Gradient Boosting Machine)
    - Excels in dealing with tabular data
- Utilizing the chosen model:
  - Deciding features to run through the model
    - Features —> Target
  - Deciding how to run our model (recursive or side-by-side)
  - Training the Model in order to optimize RMSLE
    - RMSLE: Root Mean Squared Logarithmic Error

Store Sales

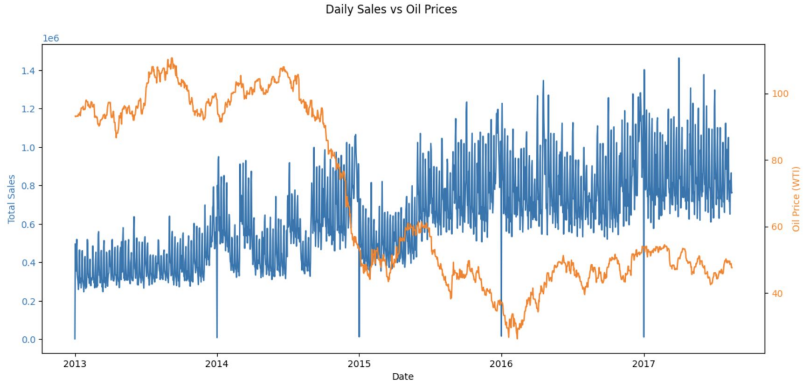
By: Camryn, Alex, Noor, Anrric

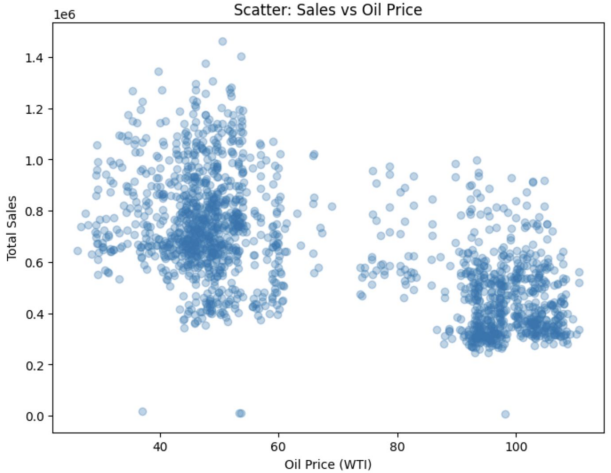
Introducing our data

- **Data Sources**
  - Train.csv → Store #, Product Family, Promotions, Target Sales
  - Stores.csv → City, State, Type
  - Oil.csv → Daily Oil Prices
  - Holidays_events.csv → Holidays & Events

- **Goal**
  - to predict the target sales for the dates in our given test.csv file (the next 15 days after the last date in our training data)

Data Cleaning Steps we've taken





Missing Values: Fill oil gaps, align transactions, check sales NaNs

Holidays: Adjust transferred holidays, simplify to single is_holiday flag

Outliers: Remove negative/zero oil prices & sales

Useful Tools

Pandas

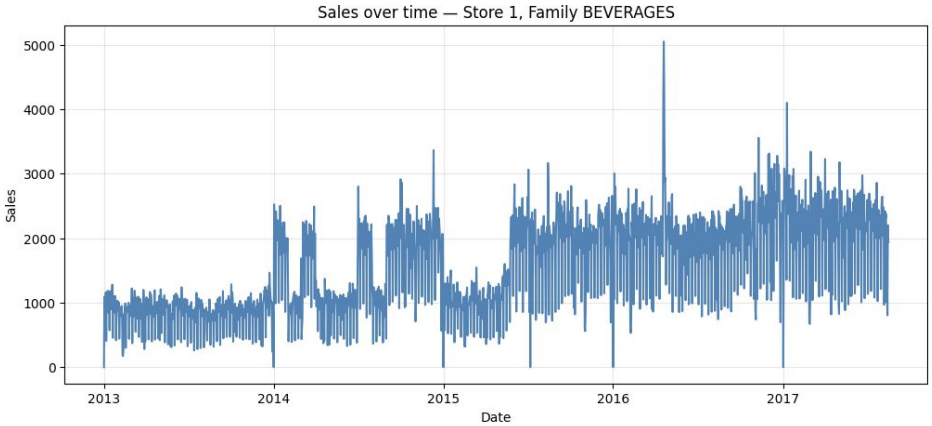
Numpy

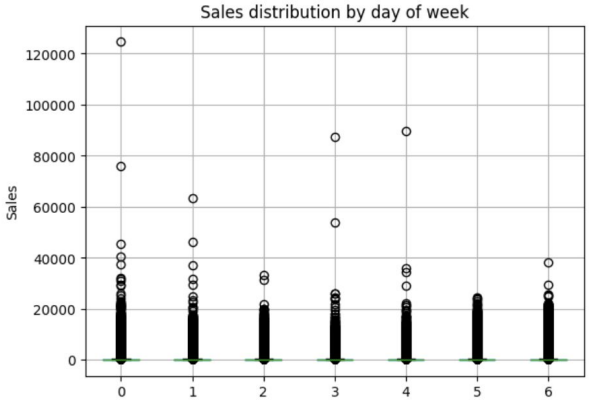
Matplotlib

Datetime to get month, day of week, payday info

Sklearn for linear regression and RMSLE

Patterns so far





1. Sales Over Time (Line Chart)

- Pattern: Peaks and dips repeat in a regular rhythm, especially for categories like Beverages.
- Challenge: hard to anticipate sudden spikes from promotions or holidays.

2. Sales by Day of Week (Boxplot)

- Pattern: Sales vary systematically by day of the week. Some product families sell more on weekends, while others dip.
- Challenge: A naïve average model ignores these weekday effects, meaning predictions may consistently may predict too under or over depending on the day.

Baseline Model

Model: Linear Regression using store state, city, family, promotions, oil prices, holidays, and previous day sales

Results: RMSLE = 1.749113, which isn't great, but it's a decent first step

We've since explored LGBM, giving us a better RMSLE of ~0.54

Next Steps

1. More Data Exploration

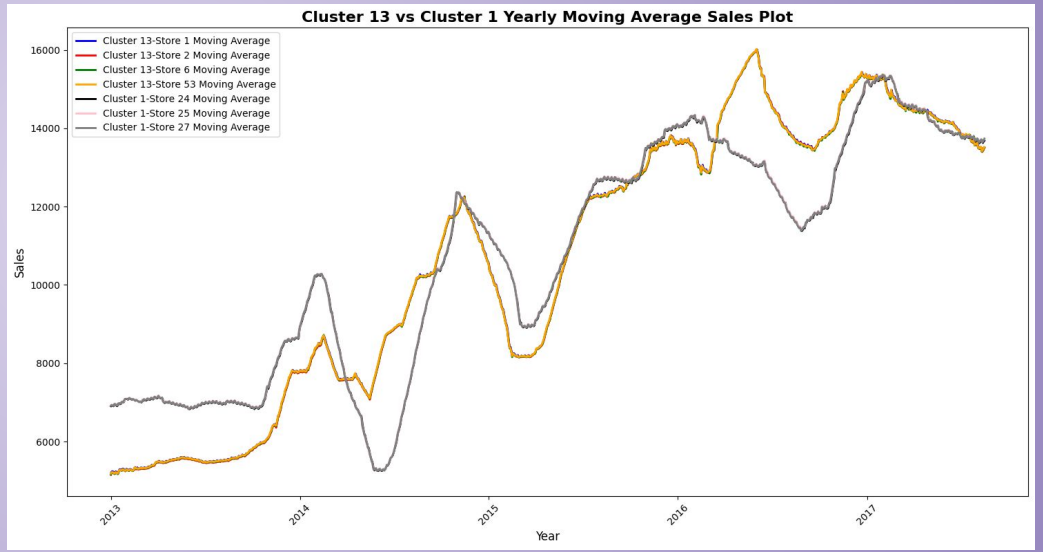
2. Research into different models

Store Sales EDA

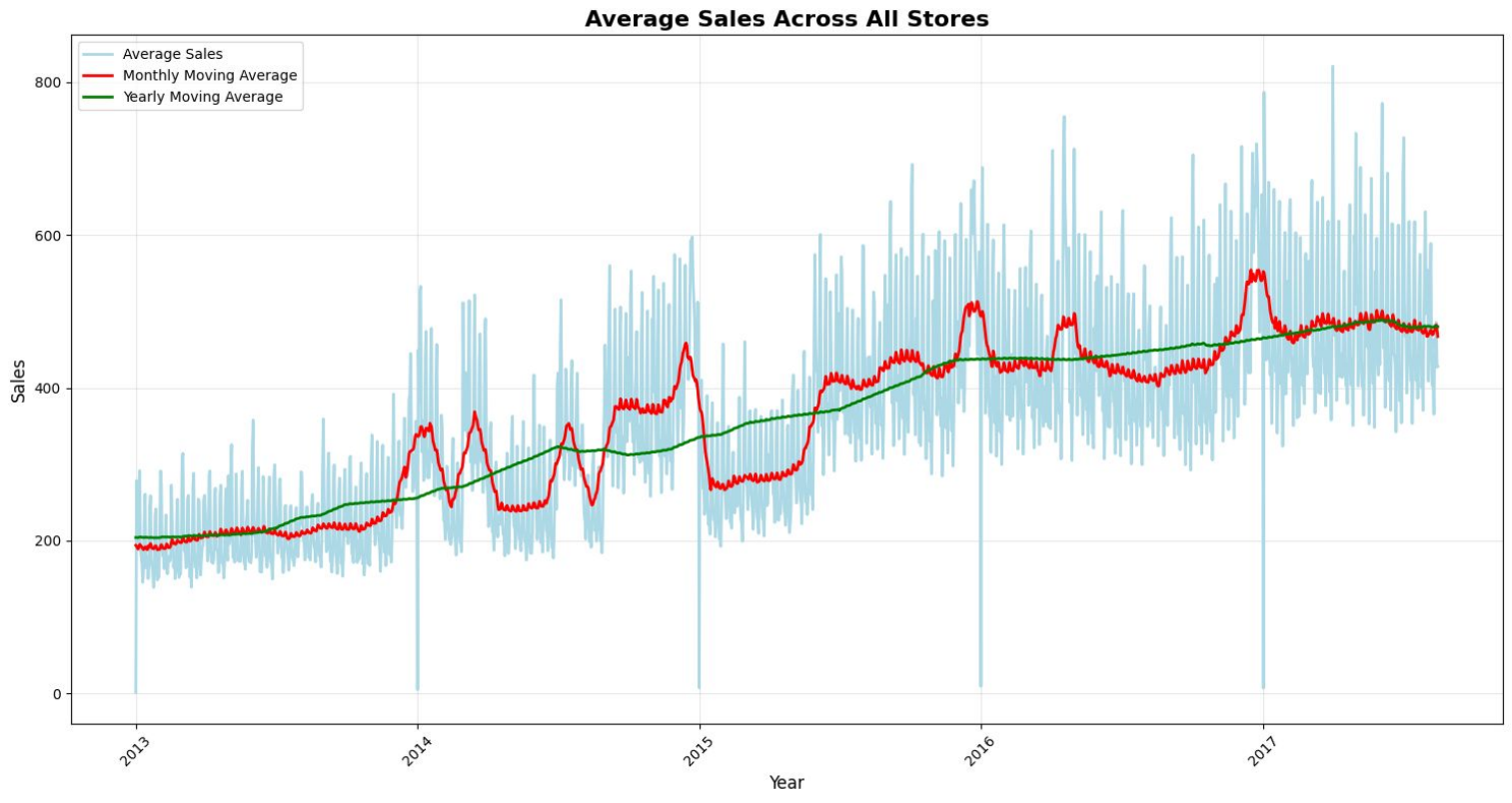
Zibo Shen, Brian Hepler, Daksh Kacham, Alex Kubiak, Andrew Piela

Dataset Descriptions

- **Train.csv**–main time series data with product-level store metadata
- **Test.csv**
- **Stores.csv**–store metadata
- **Oil.csv**– daily oil prices
- **Holidays_events.csv**–local, regional, and national holidays
- **Transactions.csv**–store-level transaction data over time

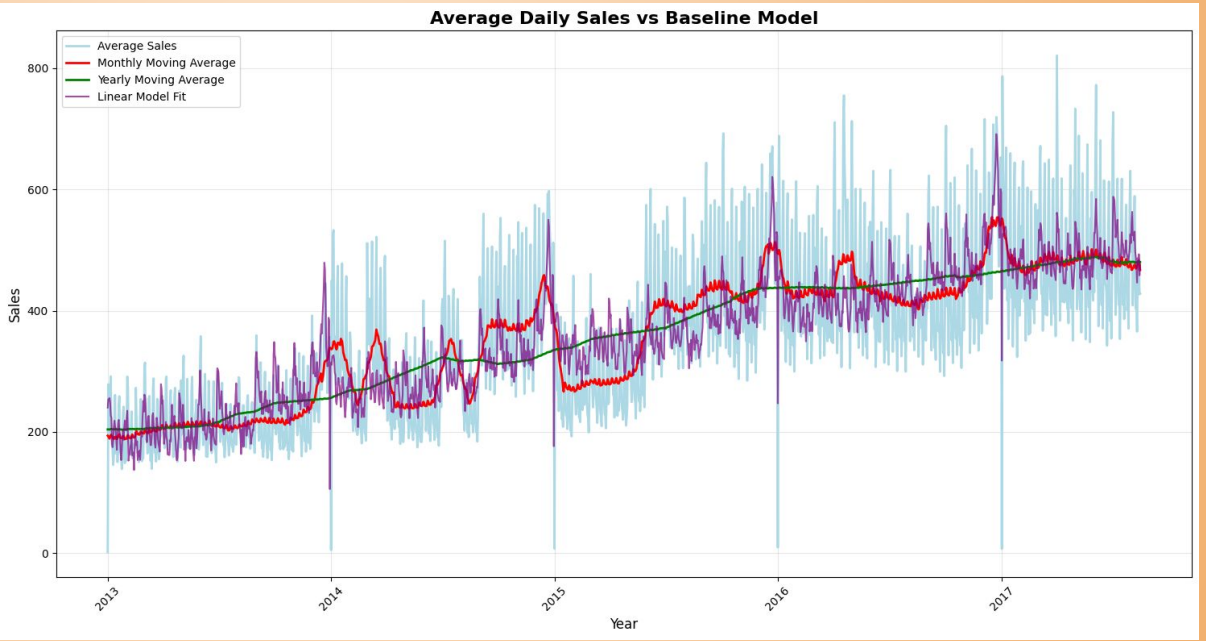


Daily Average Sales time plot

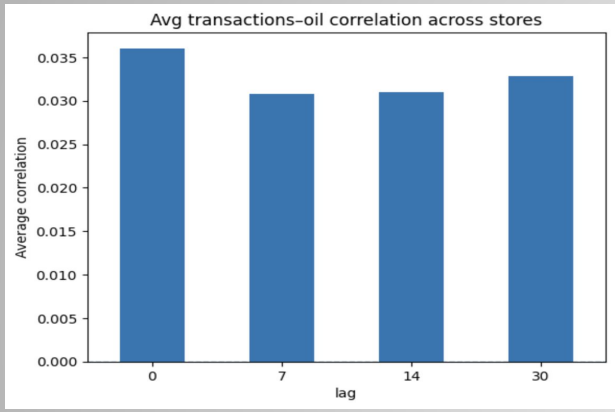


Baseline Model: Linear Regression

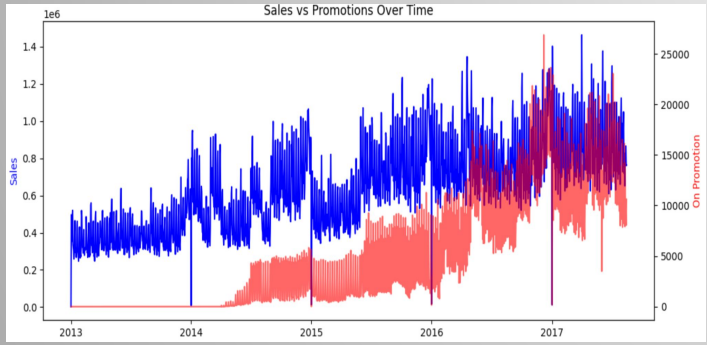
- Just using yearly seasonality
- Ignores daily volatility, weekly seasonality, holidays, oil price data, etc.
- Achieves RMSLE of 0.31 (predictions are off by ~30-35% on average, in relative terms)



Interesting relationships and data behavior



Low Correlation



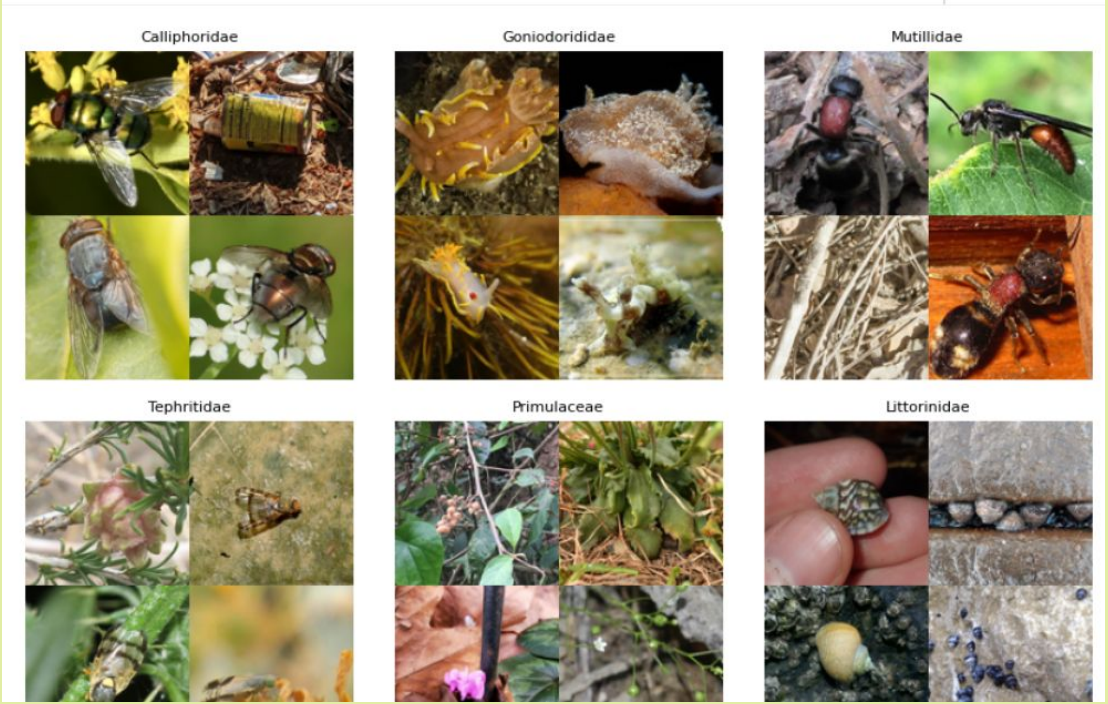
High Correlation

MLM 25
Clustering BioTrove

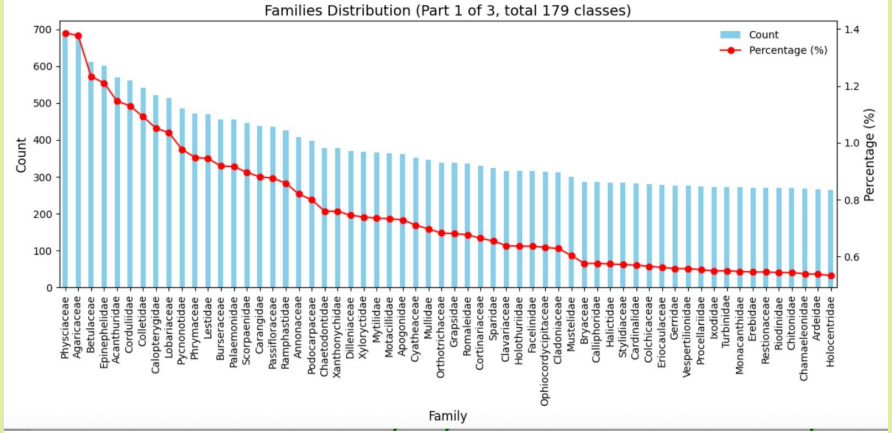
Chenchen Ding, Michal Laszkiewicz, Zhixing Xu

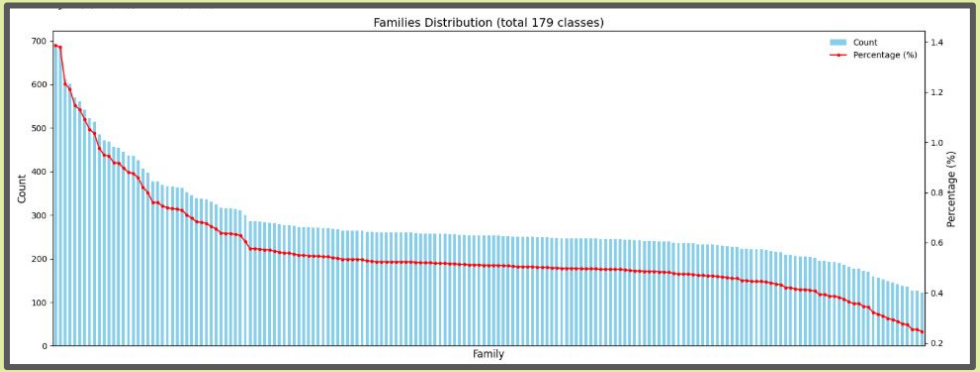
Dataset

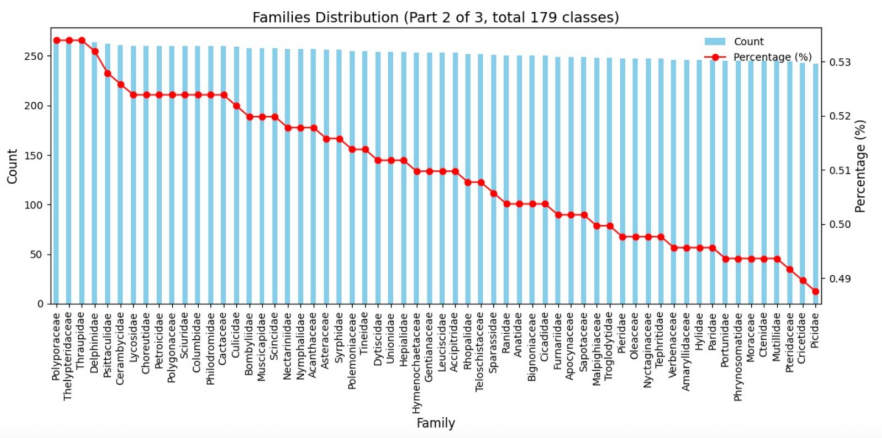
- Original dataset: 161.9 million images
  - Curated from iNaturalist platform and vetted
  - “Largest publicly accessible dataset designed to advance AI applications in biodiversity”
- MLM 25 BioTrove subset dataset: **49,633 images**
  - Image files (jpeg, png)
  - Metadata file (49,633, 2)
    - hash_id
    - family

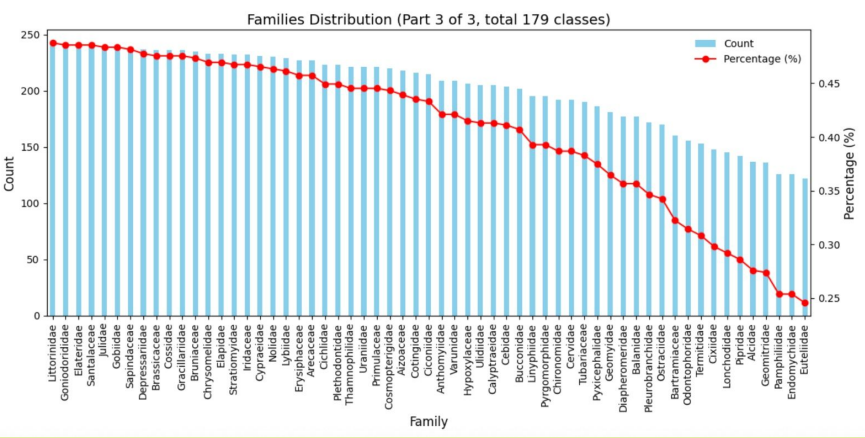


Data Distribution: Family labels









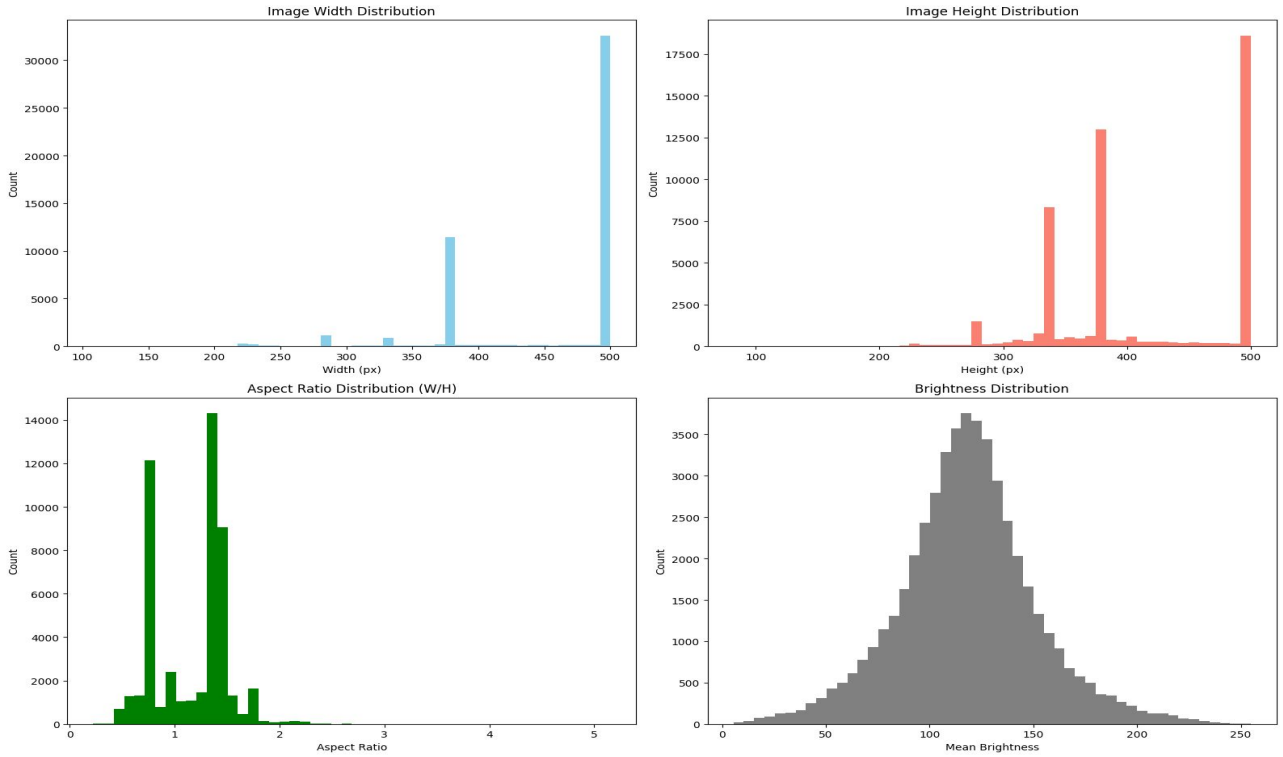
Families Distribution

- Most-represented (highest count)
  - Physciaceae

- Least-represented (lowest count)
  - Euteliidae

- Coolest? (subjective)
  - Chamaeleonidae

Image Dataset Quality Check



1. Very small images (width < 100 or height < 100): 3
2. Very large images (width > 500 or height > 500): 0
3. Abnormal aspect ratios (W/H < 0.2 or > 5): 1
4. Very dark images (brightness < 20): 118
5. Very bright images (brightness > 240): 21

Base Model

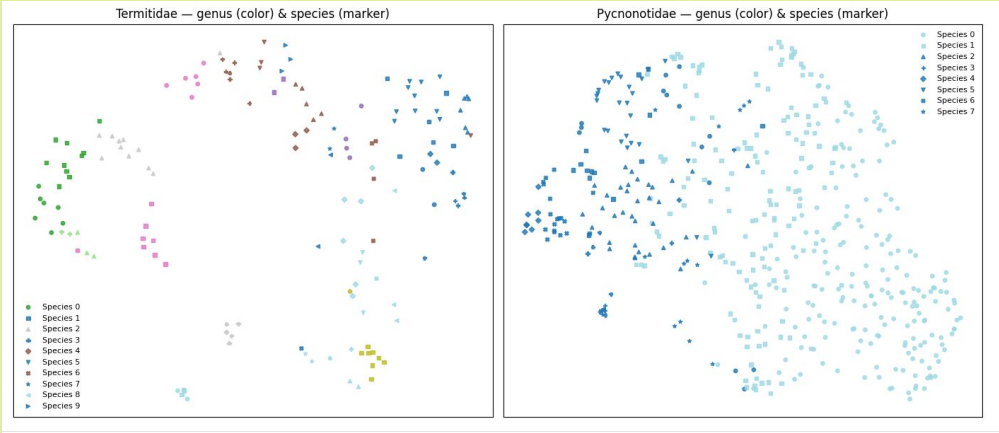
- Pre-trained CNN (ResNet 50) for image embeddings
- UMAP to visualize image embeddings
- 2 Approach on the embeddings from CNN:
  - HDBScan:
    - 93% of the dataset reported as noise
  - K-means:
    - Used silhouette_score to achieve auto k value
    - Used function to make sure all samples in one genus had the same family and all the samples in a species had the same genus during clustering

```
for fam in E["family"].unique():
    fam_mask = (E["family"] == fam).values
    fam_idx = E.loc[fam_mask, "idx"].to_numpy()
    Xf = X[fam_idx]
```

```
Kg = choose_k_silhouette(Xf, k_min=2, k_max=kmax_genus)
kmg = KMeans(n_clusters=Kg, n_init=10, random_state=0)
genus_labels = kmg.fit_predict(Xf)
E.loc[fam_mask, "genus_id"] = genus_labels
```

Base Model - Results/Analysis

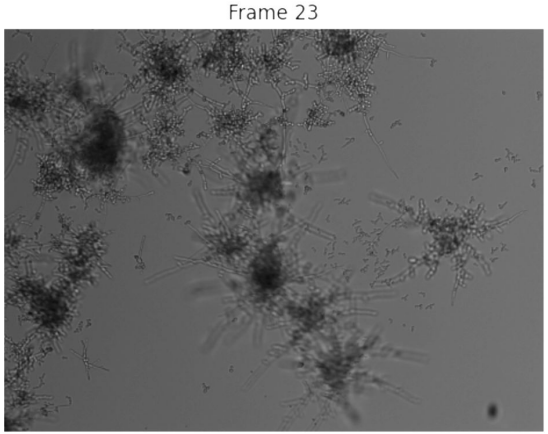
- First attempt evaluation score = 0.42582
- Next steps:
  - Improve embedding quality with different pre-trained CNN model.
  - Explore clustering choice and structure:
    - K Means restrictions:
    - Assumes clusters are spherical, similar in sizes, and no outliers.

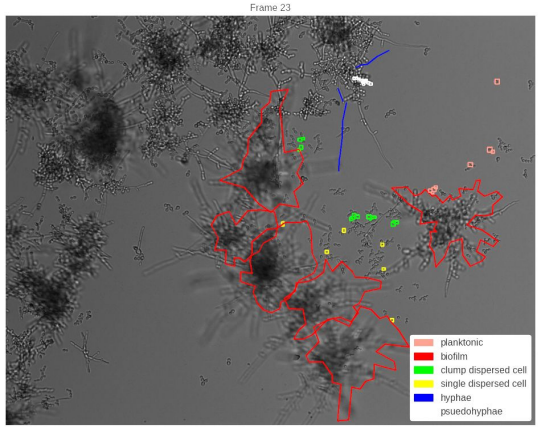


Cellular State Trackers

Ryan Bemowski
Moshi Fu
Khadijeh Masumnia-Bisheh
Yura Oh

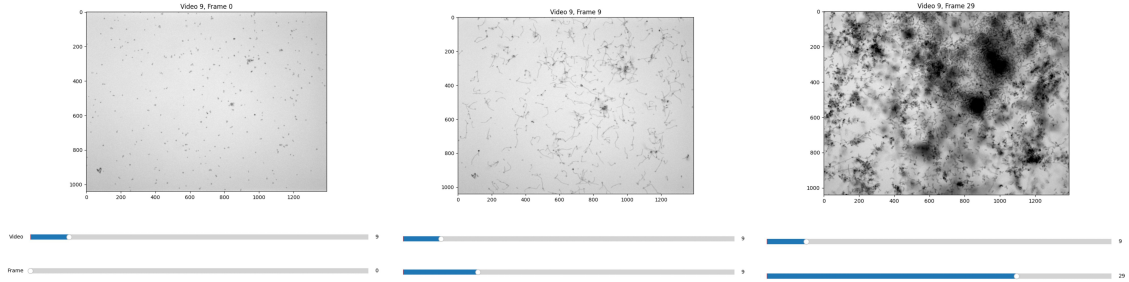
Quantifying Fungal Biofilm Dynamics Using Time Lapse Microscopy Imaging





Cell Type	Image ID	Count	…
A	01	2	…
B	01	4	…
C	01	3	…
…	…	…	…

Exploratory Analysis



Exploration plan:

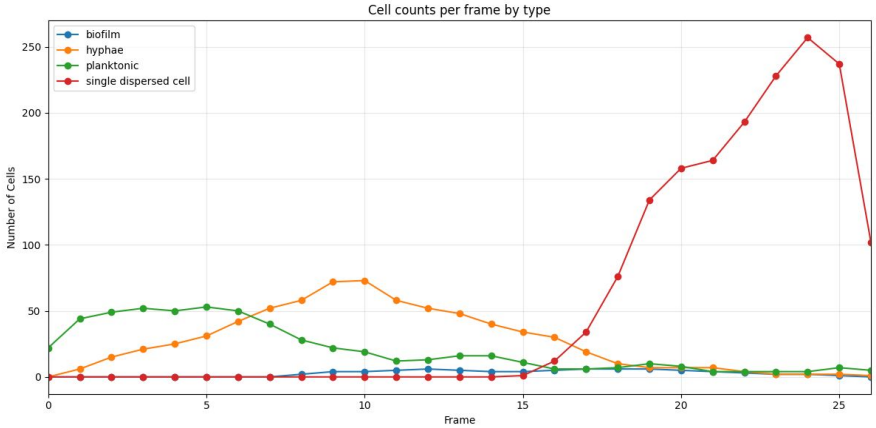
- Identify contents of .nd2 files
- Identify data in label files (.xml)

Exploration findings:

- .nd2 files are images in time lapse video format and can be converted to tif
- Labeled data is segmentation mask with additional information, such as individual cells.
- Cell counts vary drastically over time.

Tools used:

- Python
- Pandas
- Matplotlib
- nd2
- tifffile



Challenges and baseline

- One of the challenges is its low magnification, which makes it difficult to identify individual cells visually.

- As biofilms begin to form, they tend to lift above the imaging plane, creating a three-dimensional structure that causes the cells to appear out of focus (shown below)

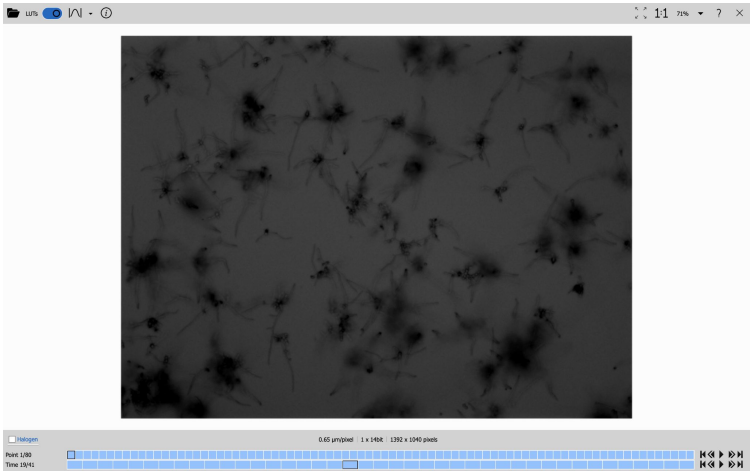
- We are still working on understanding the labeled data and plan to build a baseline prediction model to count cells using polynomial regression of labeled cell counts

EDA Presentation – Cellular State Image Analysis

By Brain Buddies:
Aman, Ainesh, Diya, Aarav, Aviaditya

Data

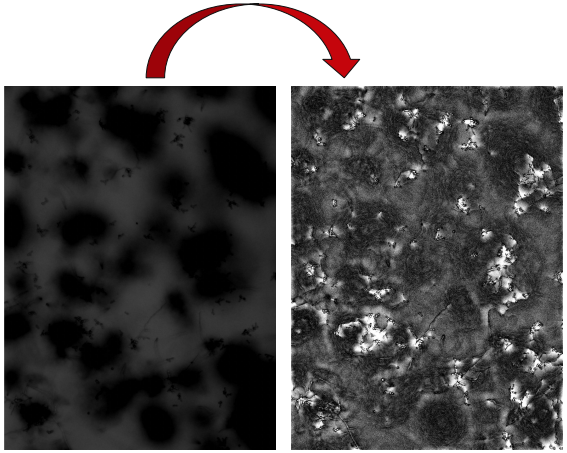
- Nd2 files with 20 hours of image data taken from a Nikon microscope
- Data was also provided through .tif and .xml files



Cellular progression

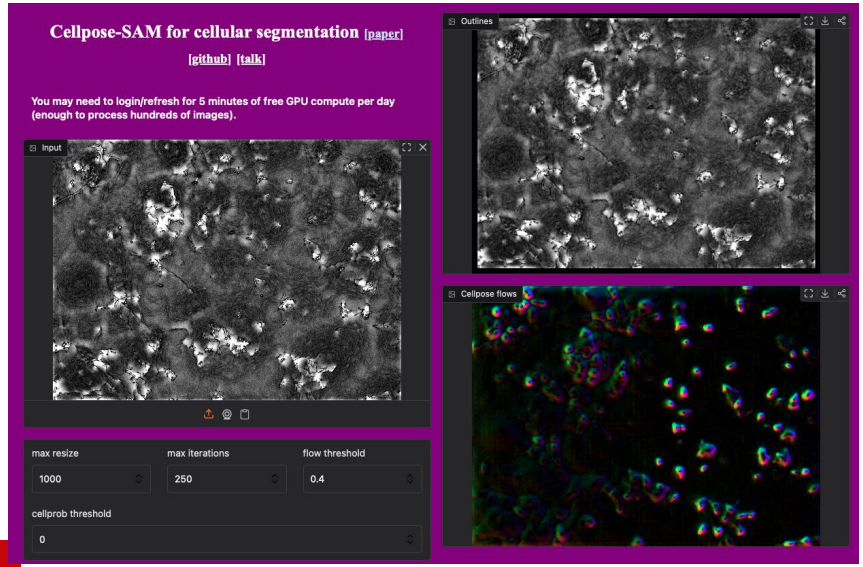
- Planktonic
- Hyphae
- Biofilm
- Dispersed

Exploratory Data Analysis



Cleaning & Pre-Processing

1. Loaded Nd2 files into a viewer and converted to .tif files
2. Pre-processing: noise reduction, gaussian smoothing, rescaling, and other Scikit packages



Helper packages

- Nd2File
- tifffile
- OpenCV/CV2
- Matplotlib
- Scikit-Image

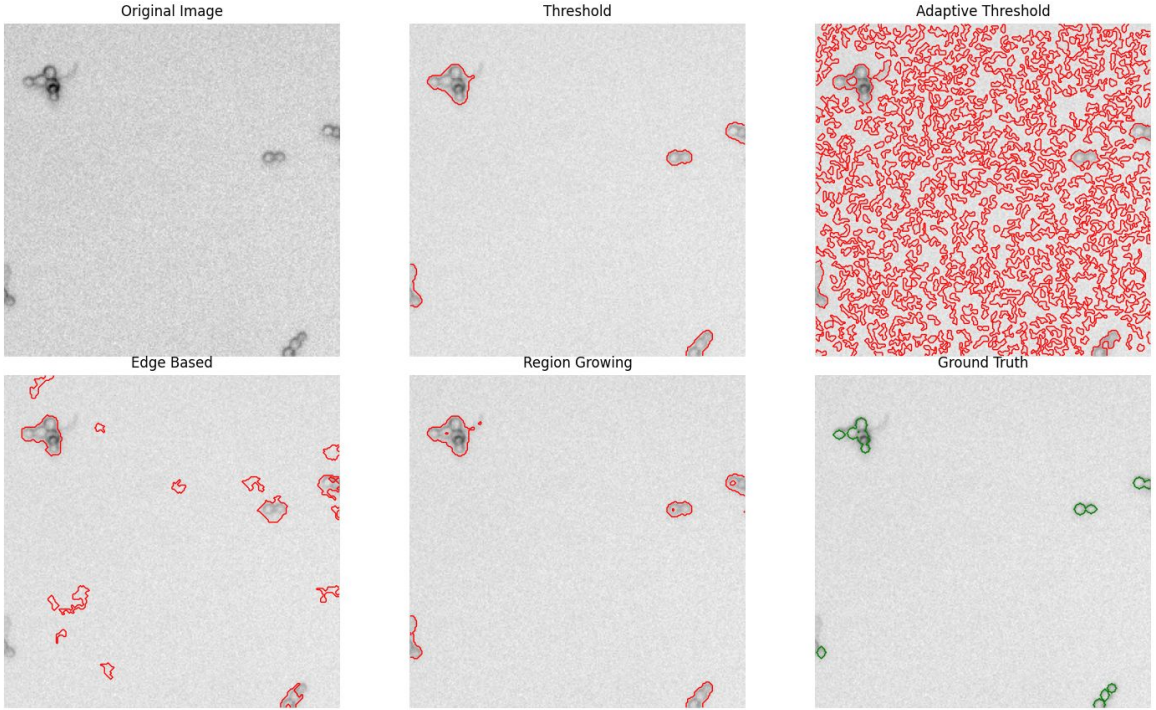
Insights & Challenges

- After doing research and comparing current existing model performance, they tend to work with highly magnified cells.

- We faced challenges choosing the combination of image pre-processing techniques and segmentation models for different frames in the test .tif file.

Baseline Model

- Edge based segmentation had the most consistent F1-score of around 0.4.
- There were issues with the granularity of the image which interfered with the segmentation.
- Some of the other models we tested was able to find regions o planktonic clusters, however, it overestimated certain areas of darker pixels.



Next Steps

- We aim to improve our image pre-processing techniques:
  - Adaptive Background Subtraction
  - Multi-Scale Enhancement

- Test new image segmentation techniques:
  - Advanced Watershed
  - Contour Refinement

[NeurErgo] Brain-to-text '25 Exploratory Data Analysis

MLM Challenge 2025
Madison, Wisconsin USA
10/02/2025

Jeevan Jayasuriya, David Nartey, Yinsu Zhang

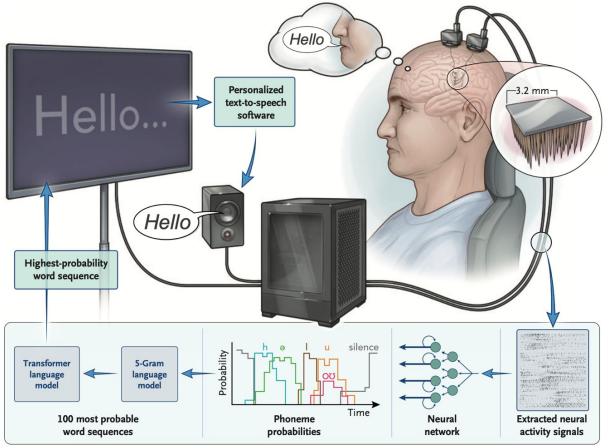
Brain-to-Text Background

Background:

- People with ALS or brainstem stroke may lose the ability to move and speak
- Speech BCIs restore communication by decoding intended speech directly from brain activity
- Previous work: BCI decoded attempted speech of a man with ALS with 97.5% accuracy

Project Objective:

- Decode intracortical neural activity from the speech motor cortex into words



Data Overview

Dataset Components:

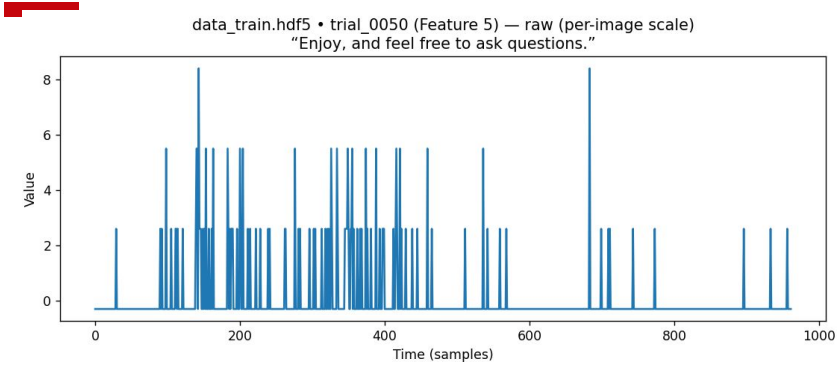
- *t15_copyTask.pkl*: Online Copy Task results
- *t15_personalUse.pkl*: Conversation Mode data
- *t15_copyTask_neuralData.zip*: Neural data for Copy Task

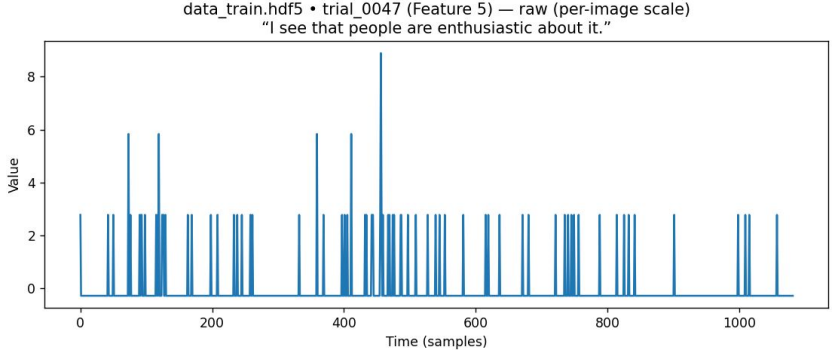
Neural Data

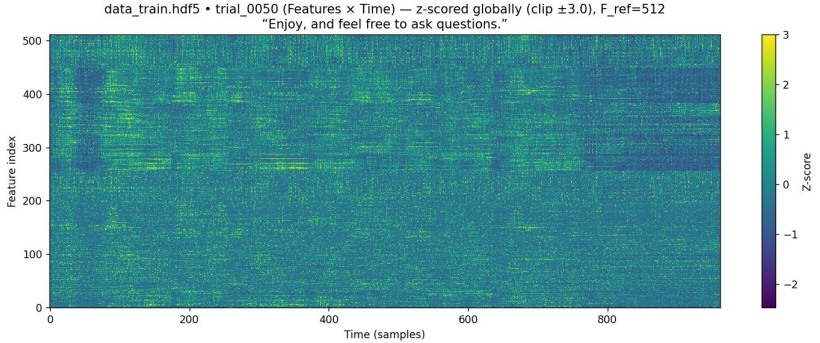
- **10,948** sentences from **45** sessions spanning **20** months
- Each trial includes session date, block number, trial number
- **512** neural features per trial (256 electrodes), binned at **20 ms**

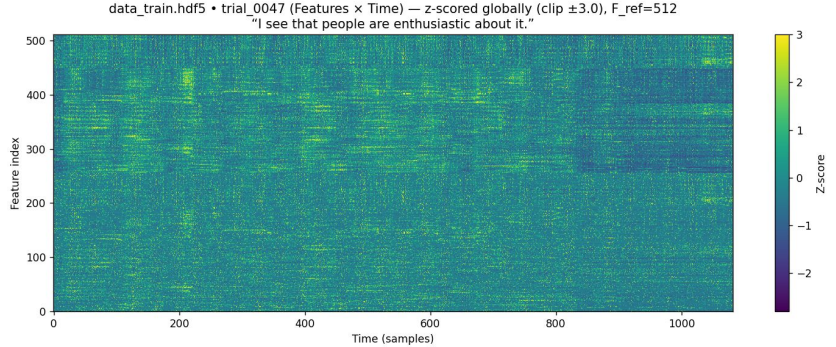
Raw Neural Signals

Visualizing raw data from S15 - some features are more active others across time

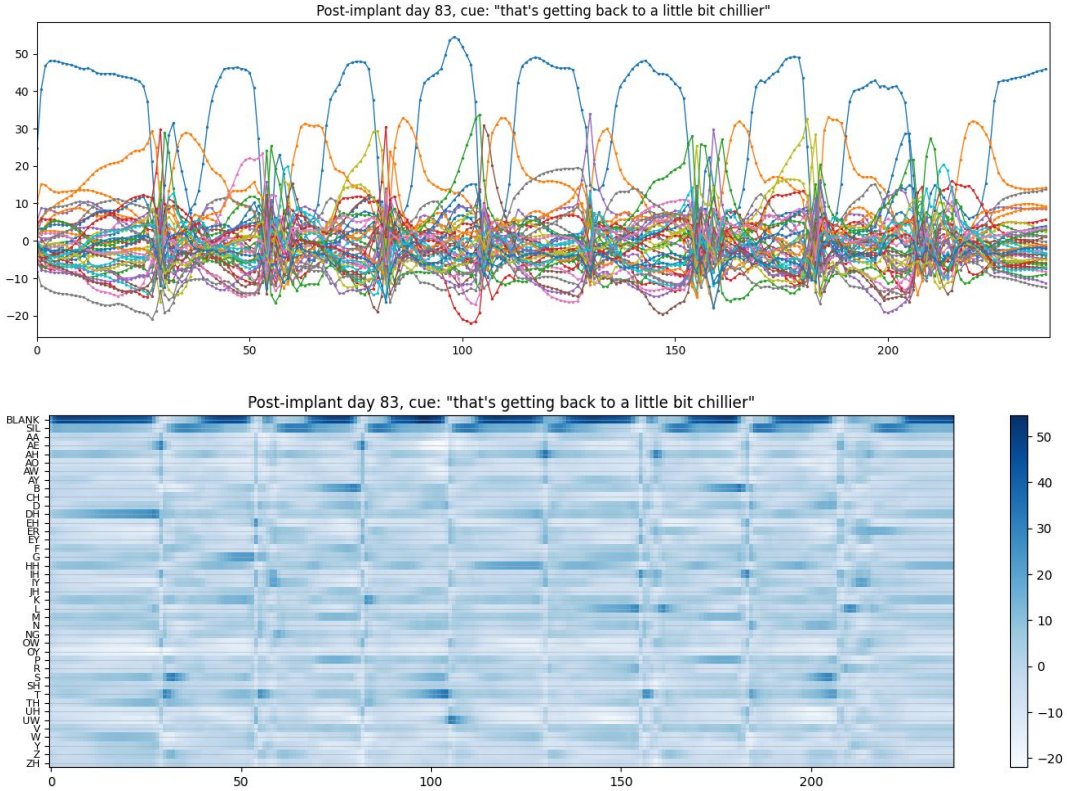








Neural Network Outputs (logits) for phonemes



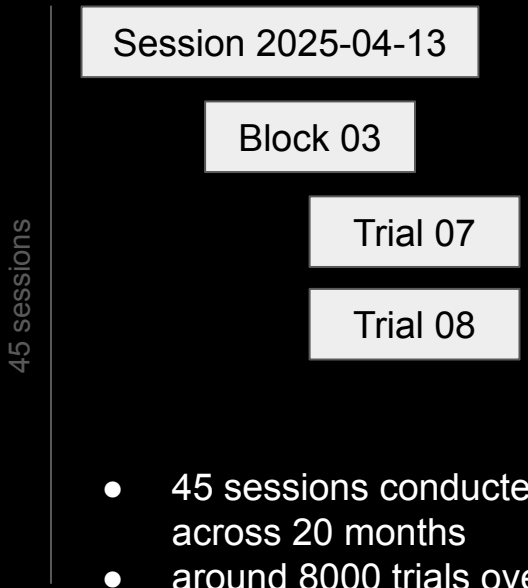
Brain-to-text EDA

4K

Eugene Kim, Taeyeon Lee, Woonggi Yoon, Yoonho Park

Exploring the dataset structure

Used Packages and Tools: h5py, pathlib / os, pandas, numpy, matplotlib, scikit-learn

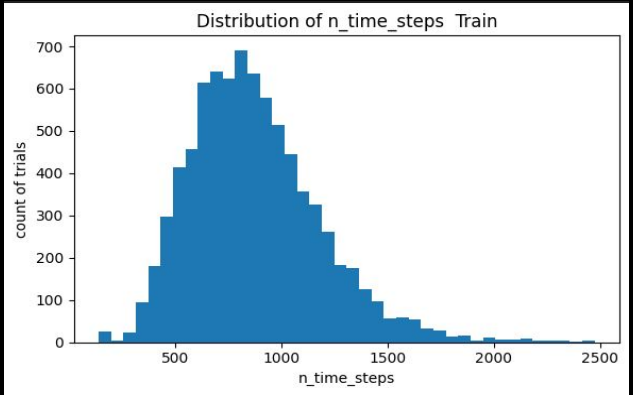


- 45 sessions conducted across 20 months
- around 8000 trials overall

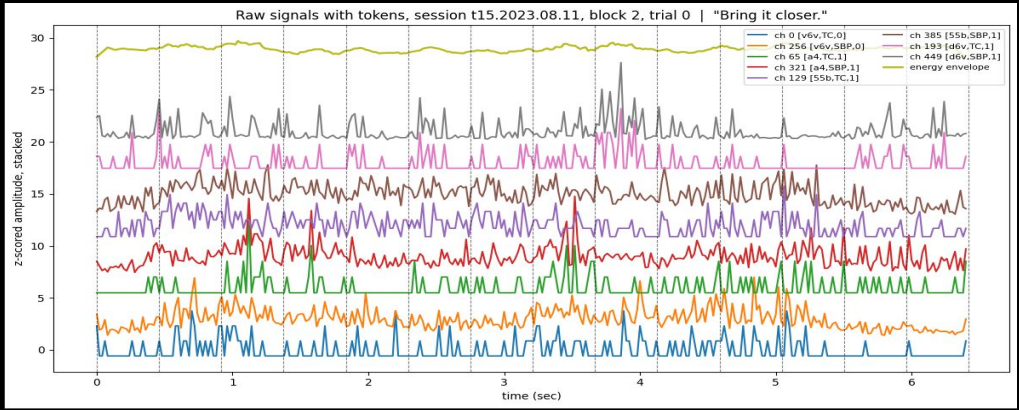
For each trials:

- Dataset
  - input features (n_time_steps, 512) | float
    - 2 features (Threshold Crossings, Spike Band Power) for each electrodes
    - 64 electrodes in 4 areas (ventral, area 4, 55b, dorsal)
  - seq_class_ids: ids of phonemes
  - transcription: ids of chars of sentence label
- Attributes
  - sentence label, session, block_num
  - trial_num: trial num within the block
  - n_time_steps: 20ms per time step
  - seq_len: num of phonemes

Key Steps (EDA): Dataset + Neural Activity

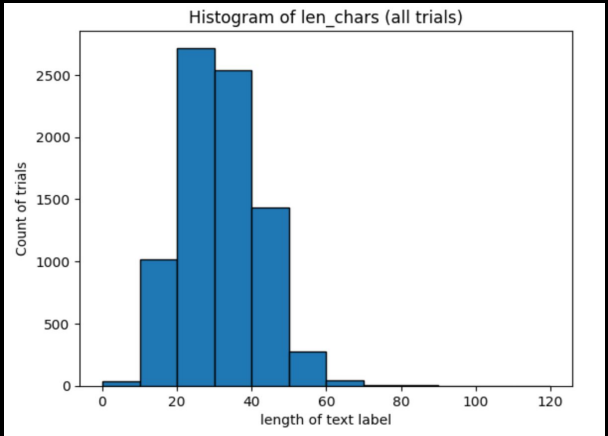


- **Sessions/Blocks**: 2–9 blocks, 60–350 trials per session
- **Number of Time steps**: Most last **600–1200 steps (~12–24s)** within range 138-2475, shows right skew
- **Implication**: Imbalance → Applied 400~1600 steps (q05 ~ q95) to remove extreme short/long trials and stabilize training

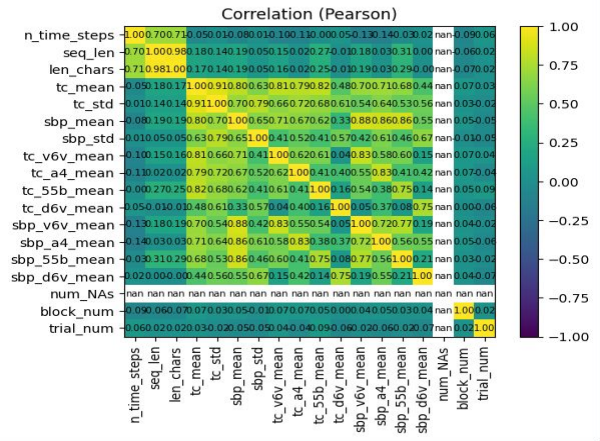


- **Neural signals (**z-scored amplitudes) shown for selected electrodes.
- Token boundaries marked with dashed lines, **aligned** to speech.
- Energy envelope (purple) tracks overall speech intensity.

Key Steps (EDA): Statistics & Correlations

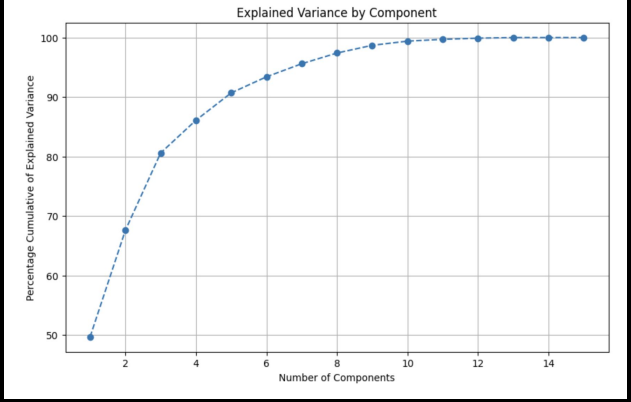


- **Sentence length**: Most transcripts fall between **20–40 characters**; only a few exceed 60.
- **Vocabulary diversity**: char TTR and phoneme TTR around 0.5-0.6 is most common
- **Implication**: Models must handle variable-length inputs and address vocabulary imbalance.



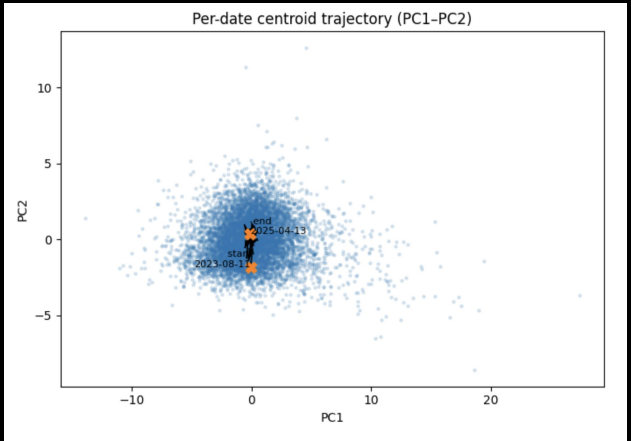
- **seq_len & len_chars** strongly correlated (longer sentences → more phonemes to pronounce).
  a. seq_len and len_chars also have moderately high correlations to n_time_steps
- **Electrode features (TC/SBP means)** show **moderate–high correlations**, indicating redundancy across channels/areas. → PCA for feature extraction

PCA and Domain shift over time



5 PCs explain 91% variance, 7 PCs reach 95% → Chose 5 PCs

- 5 PCs variance: [0.49653528 0.17922831 0.12992007 0.05482748 0.04621366]



Levene's test and Kruskal-Wallis test suggested that variance and median differs by date

Centroid trajectory for PC2 vs PC1 plot from old to new session

- **Concluded there is domain shift across dates / sessions, mainly in PC2**

RNN Baseline model

```
model:
  n_input_features: 512 # number of input features in the neural data.
  n_units: 768 # number of units per GRU layer
  rnn_dropout: 0.4 # dropout rate for the GRU layers
  rnn_trainable: true # whether the GRU layers are trainable
  n_layers: 5 # number of GRU layers
  patch_size: 14 # size of the input patches (14 time steps)
  patch_stride: 4 # stride for the input patches (4 time steps)

  num_training_batches: 120000 # number of training batches to run
  epsilon: 0.1 # epsilon parameter for the Adam optimizer
  weight_decay: 0.001 # weight decay for the main model
```

```
2025-09-30 09:41:09,221: Train batch 58200: loss: 0.06 grad norm: 4.22 time: 0.505
2025-09-30 09:43:16,181: Train batch 58400: loss: 0.08 grad norm: 6.46 time: 0.594
2025-09-30 09:45:23,507: Train batch 58600: loss: 0.05 grad norm: 3.40 time: 0.486
2025-09-30 09:47:34,354: Train batch 58800: loss: 0.04 grad norm: 2.12 time: 0.585
2025-09-30 09:49:41,844: Train batch 59000: loss: 0.03 grad norm: 1.83 time: 0.529
2025-09-30 09:51:55,419: Train batch 59200: loss: 0.03 grad norm: 3.04 time: 0.500
2025-09-30 09:54:02,648: Train batch 59400: loss: 0.02 grad norm: 1.47 time: 0.678
2025-09-30 09:56:12,038: Train batch 59600: loss: 0.03 grad norm: 2.96 time: 0.688
2025-09-30 09:58:23,527: Train batch 59800: loss: 0.07 grad norm: 15.85 time: 0.531
2025-09-30 10:00:34,533: Train batch 60000: loss: 0.09 grad norm: 5.53 time: 0.724
2025-09-30 10:00:34,536: Running test after training batch: 60000
2025-09-30 10:01:17,624: Val batch 60000: PER (avg): 0.1055 CTC Loss (avg): 22.7981 time: 43.087
2025-09-30 10:01:17,625: t15.2023.08.13 val PER: 0.0665



2025-09-30 10:01:17,656: t15.2025.01.12 val PER: 0.1001
2025-09-30 10:01:17,656: t15.2025.03.14 val PER: 0.2959
2025-09-30 10:01:17,658: t15.2025.03.16 val PER: 0.1545
2025-09-30 10:01:17,658: t15.2025.03.30 val PER: 0.2230
2025-09-30 10:01:17,659: t15.2025.04.13 val PER: 0.1897
2025-09-30 10:01:17,659: New best test PER 0.1067 --> 0.1055
2025-09-30 10:01:17,661: Checkpointing model
2025-09-30 10:01:19,138: Saved model to checkpoint: /kaggle/working/trained_models/baseline_rnn/checkpoint/best_checkpoint
```

Baseline RNN provided by Kaggle competition

50% trained base model

- avg PER (Phoneme Error Rate) of 10.55

Next Steps

Choosing Pipeline

finalizing data preprocessing method (filtering, feature extraction etc)

Fine Tuning RNN

adjust hyperparameters

Future Integration

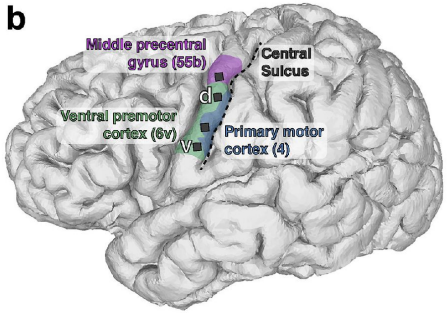
choose n-gram model to integrate with

Exploratory Data Analysis

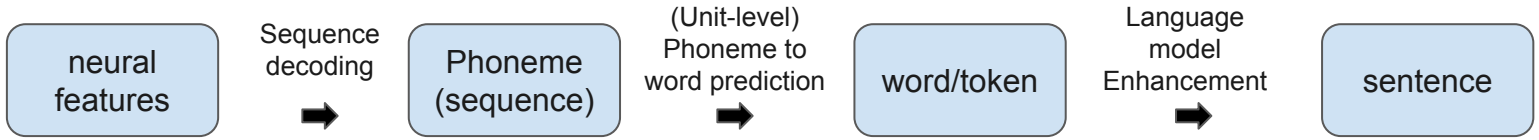
BuckyBrain (Brain-to-Text 25') 🧠 🦡

Overview

- Input variables
  - neural_features: Neural features per timestep
    - Channels: 512 = 4(Groups)*2(Channels)*64(Micro electrodes)
    - Shape: 2D array, (timestep, channels)
- Outcome variables
  - text: Target sentences
  - seq_class_ids: Phonemic sequence
  - transcriptions: Character sequence
- Meta data
  - session: Date
  - block_num: In each session there were multiple blocks
  - trial_num: A single sentence per trial
  - seq_len: Phonemic seq length
- Baseline model workflow

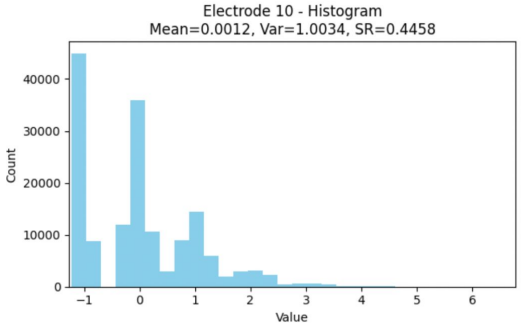


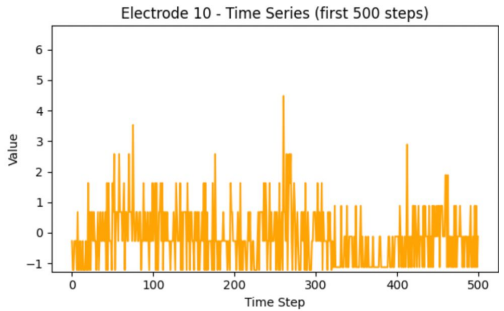
https://www.nejm.org/doi/full/10.1056/NEJMoa2314132



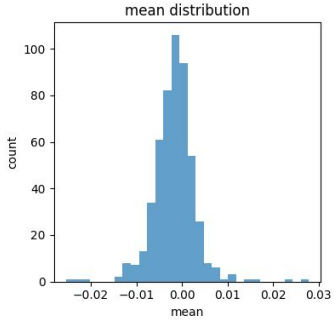
Statistics of Neural Features

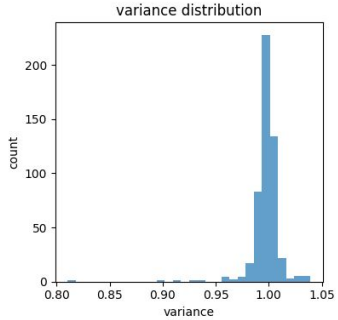
- Zoom-in single electrode

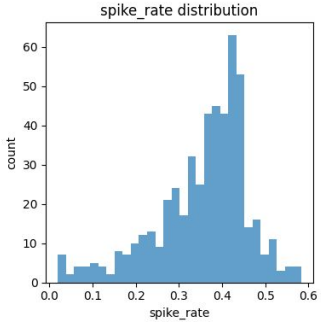




- Overall





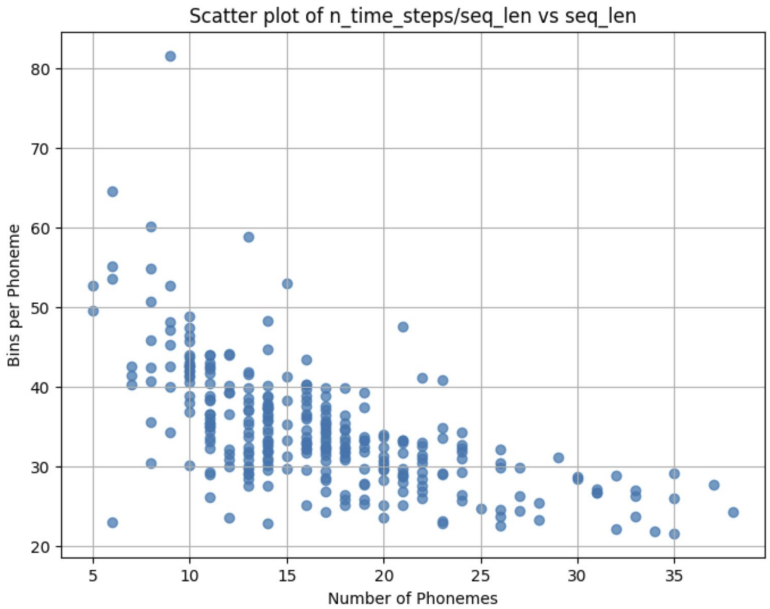


Analyzing how the neural data aligns with the sentence text

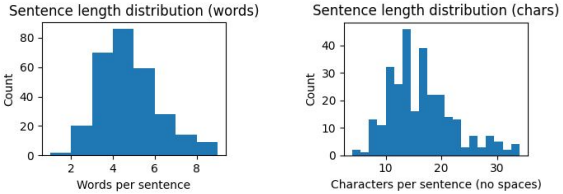
Divide the number of bins by the number of phonemes in the sentence

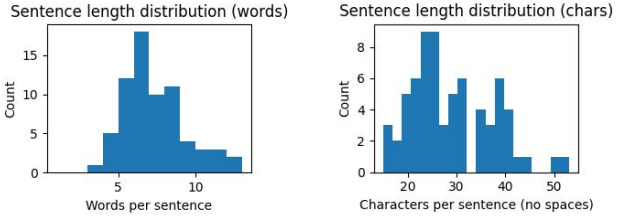
The average number of bins per phoneme = 34.70188631148757

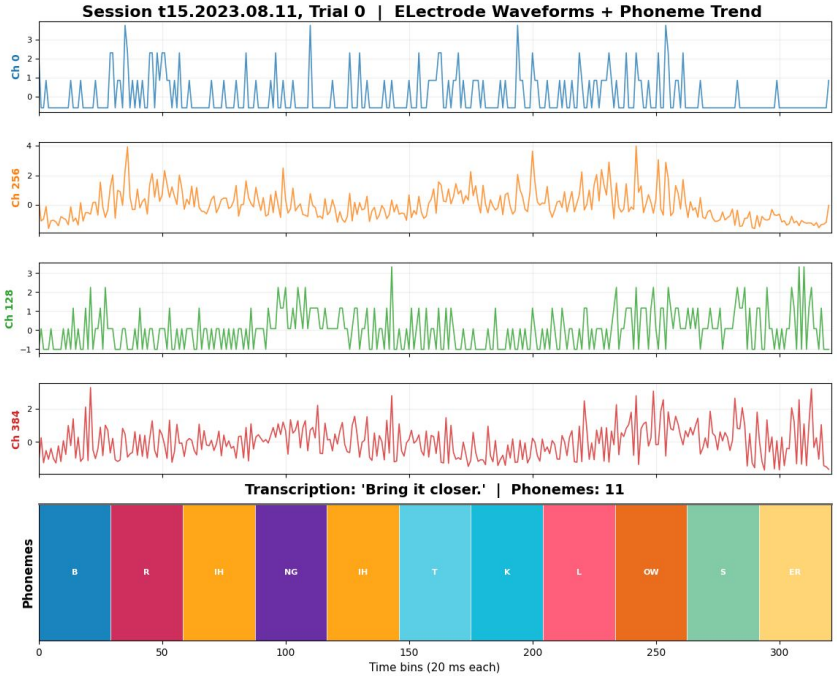
However, looking at the scatter plot it is very evident that this split is not accurate, so we need to find a better mechanism to split the number of bins for each phoneme!

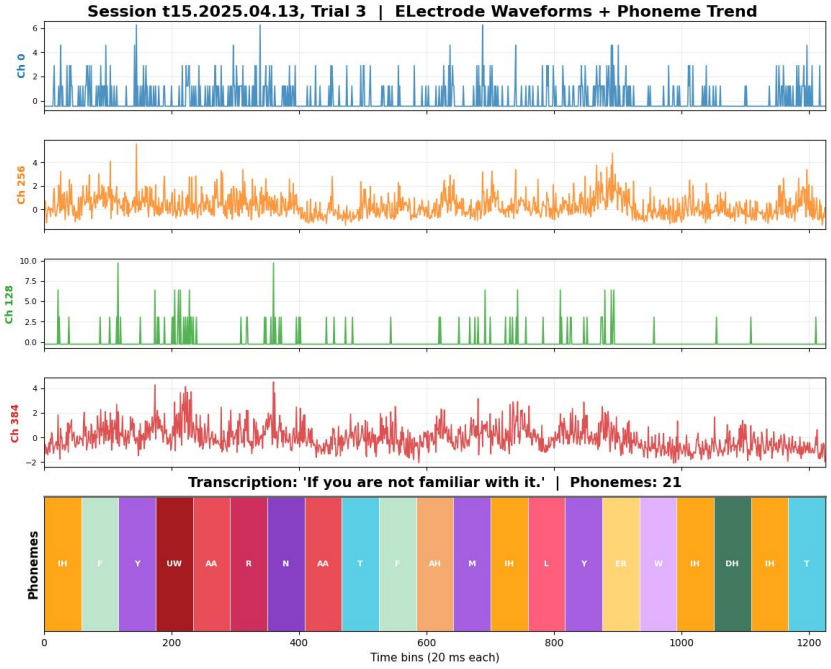


Neural activity patterns







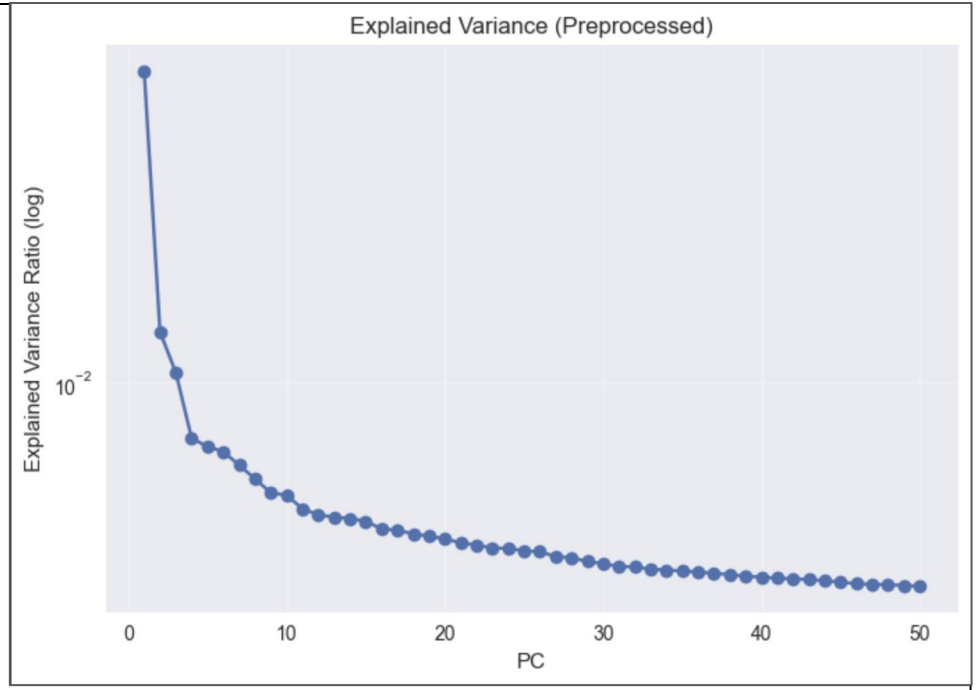


PCA Reveals the Complexity of Brain Signals

```
Total variance explained by first 3 components: 8.2%

Total variance explained by first 10 components: 13.1%

Number of components needed for 95% variance: 385
```



Thank You!

October 2nd, 2025

Neural Navigators

Brain-To-Text Neuroprosthesis EDA

Nils Matteson, Daniel Yang, George Yu, Carl Kashuk, Rohini G

The Challenge

THE CHALLENGE:

- Convert neural activity from speech motor cortex into text
- Data from NEJM 2024 study: first successful real-time speech BCI
- 1.7 years of neural recordings from 256 electrodes across 4 brain regions
- Goal: Build the best possible decoder

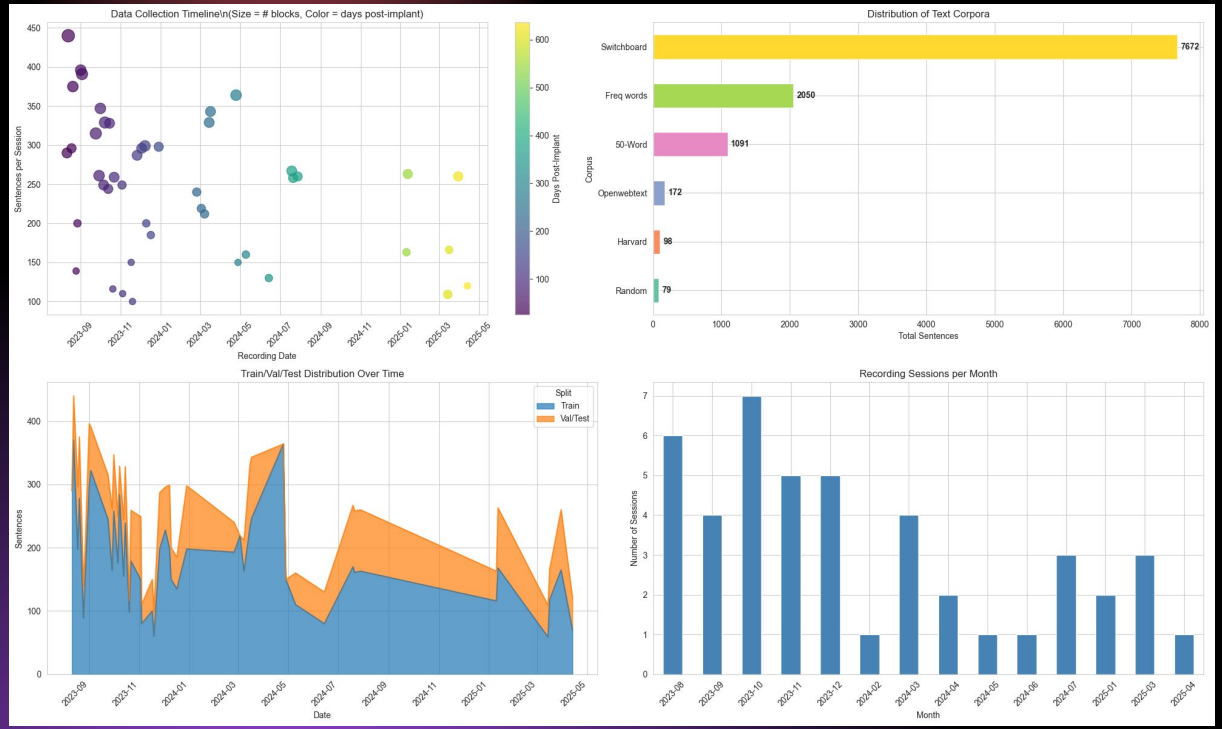
WHAT MAKES THIS HARD:

- High-dimensional neural data (512 features/trial)
- Variable sentence lengths and content
- Individual differences across recording sessions
- Need for real-time processing

Dataset Overview

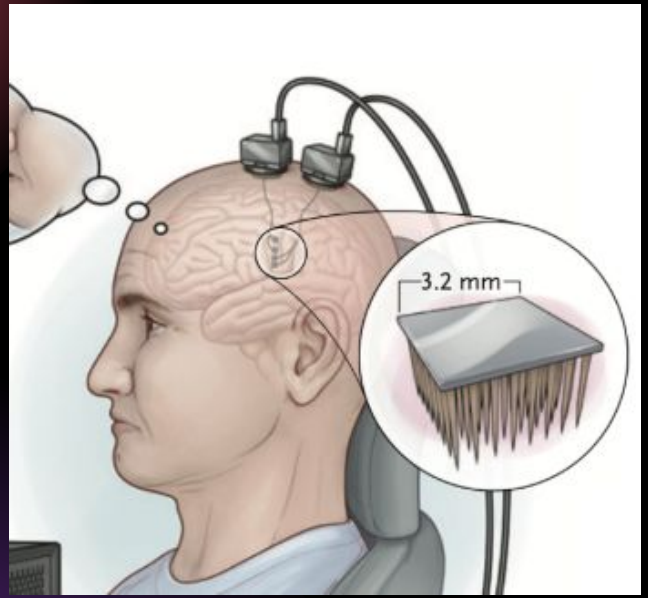
WHAT WE DISCOVERED:

- 45 sessions, 11,162 sentences over 20 months
- Multiple text corpora (conversations, controlled vocab, etc.)
- Remarkably stable data quality across time
- Clear train/validation/test splits
- Highlights data imbalance across corpora, which could affect model performance and require strategies like weighted loss or oversampling for underrepresented corpora



Neural Recording System

HARDWARE SETUP:
- 256 microelectrodes across 4 brain regions
- 512 neural features (2 per electrode)
  - Threshold crossings (spikes)
  - Spike band power (local field potentials)
- 20ms time resolution for real-time decoding



BRAIN REGIONS:
Ventral 6v (E10): Premotor cortex - speech planning and preparation
Area 4 (E75): Primary motor cortex - direct muscle control for speech
55b (E140): Parietal region - sensory-motor integration for speech monitoring
Dorsal 6v (E210): Higher-order premotor - cognitive control and sound selection

RESULT: High-density neural recording from speech motor cortex

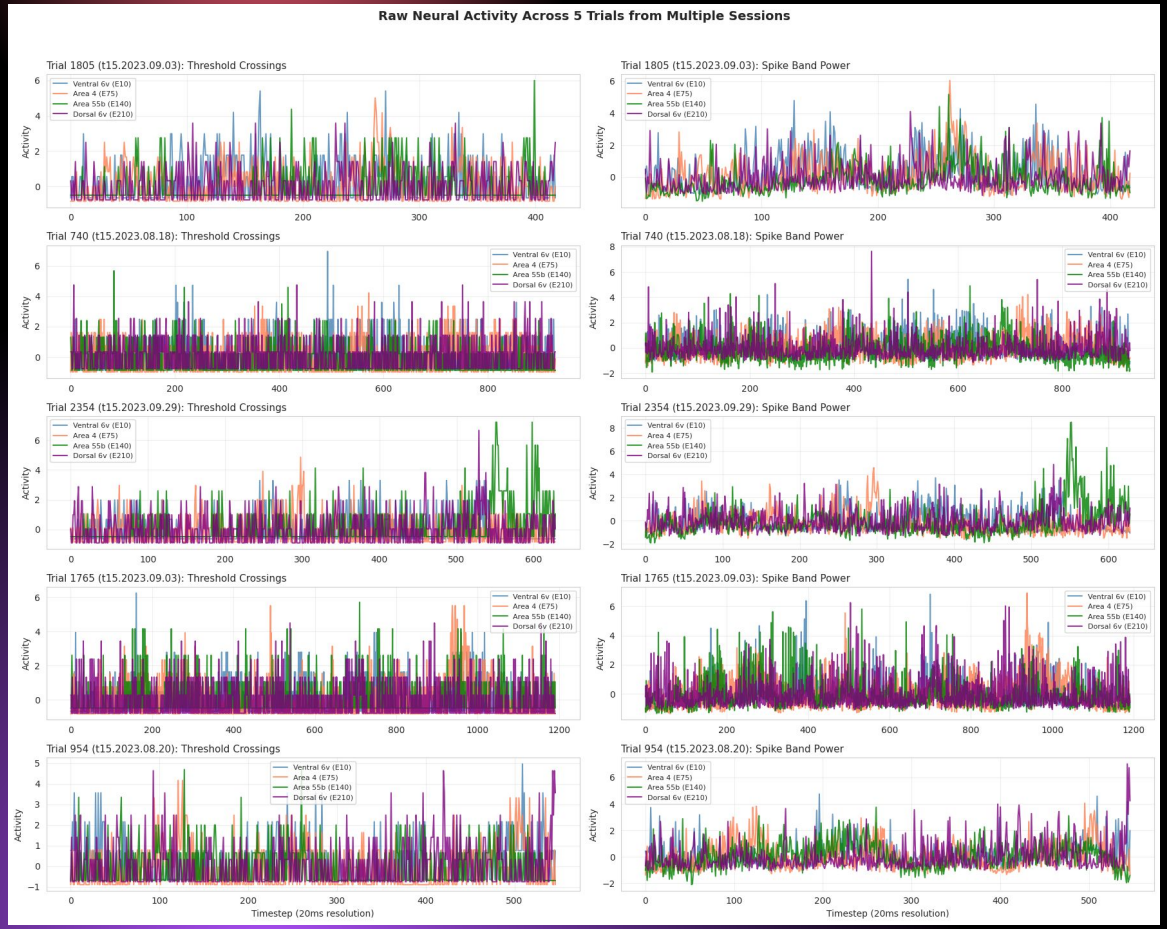
Neural Signal Patterns Across Brain Regions

- Threshold crossings (spikes): Discrete neural firing events, sharply time-locked, capturing the precise timing of neuron communication.

- Spike band power (SBP): Continuous, smoother signal reflecting the collective energy of neural populations, sensitive to sustained engagement.

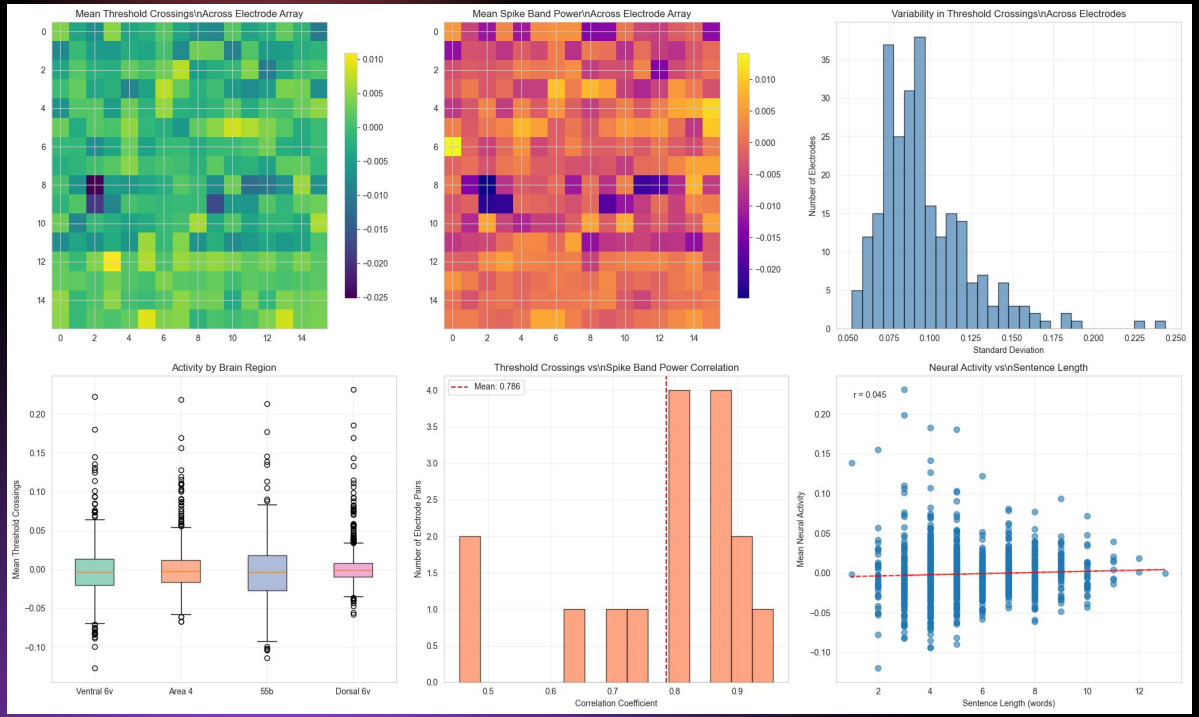
- Both features are complementary: spikes highlight rapid changes, SBP captures ongoing processes.
- Clear activity patterns emerge across different brain regions:

- Signals remain stable across long trials (~25s), showing that models can leverage both rapid events and sustained trends.



Electrode-Level Patterns and Cross-Feature Relationships

- Heatmaps of average activity show distinct spatial “hotspots,” meaning certain electrodes are consistently more active and informative.
- Variability analysis shows most electrodes cluster around moderate variability, but a subset are highly noisy, important for feature selection.
- Regional comparisons reveal: Area 4 and 55b produce more variable activity (heterogeneous signals), while Dorsal 6v is more consistent.
- Cross-feature correlations: Threshold crossings and SBP are strongly correlated (mean ~0.79), meaning they capture related but not identical processes.
- Sentence length vs activity: Correlation is essentially flat (r ≈ 0.045) → sentence complexity is not encoded by raw amplitude, but likely through distributed spatiotemporal patterns.



Electrode-Level Patterns and Cross-Feature Relationships

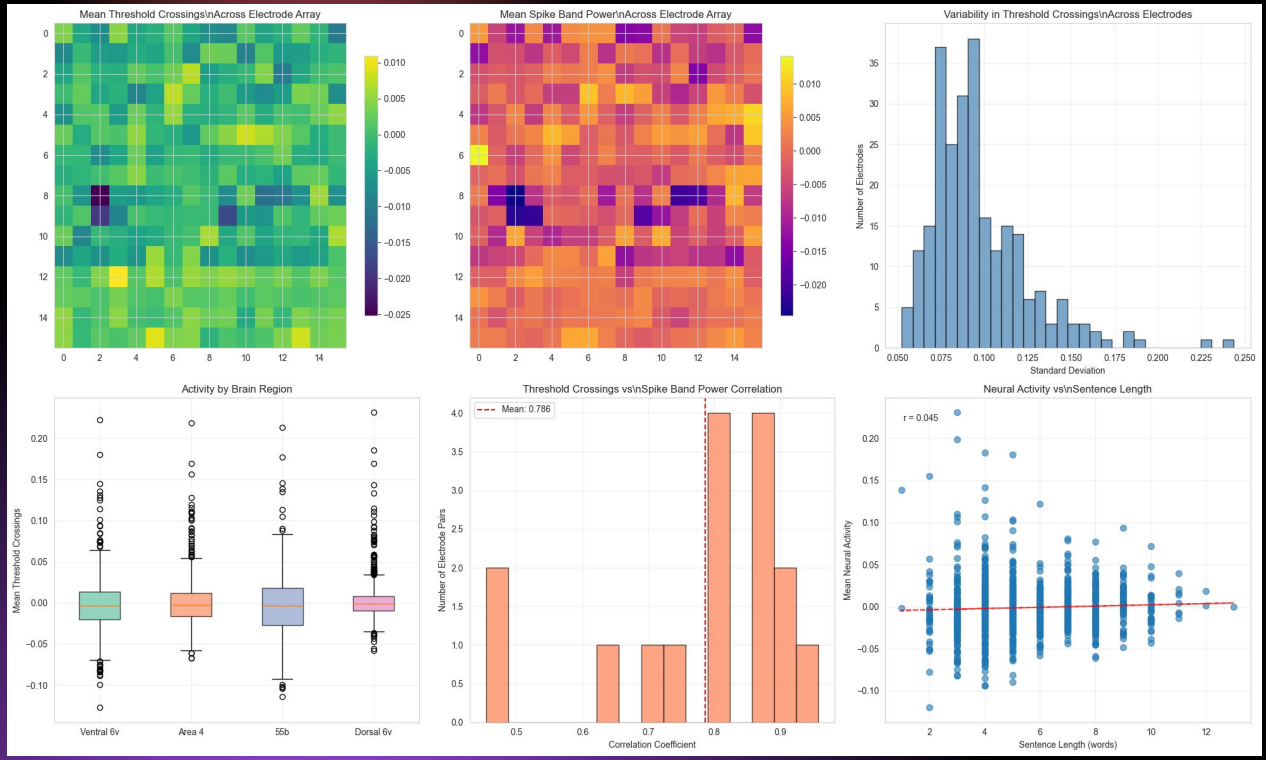
Modeling Implications:
- Feature redundancy: High TC-SBP correlation suggests potential for feature reduction

- Regional weighting: Area 4 and 55b need higher weights due to higher activity

- Complex encoding: Simple linear models insufficient for sentence length decoding

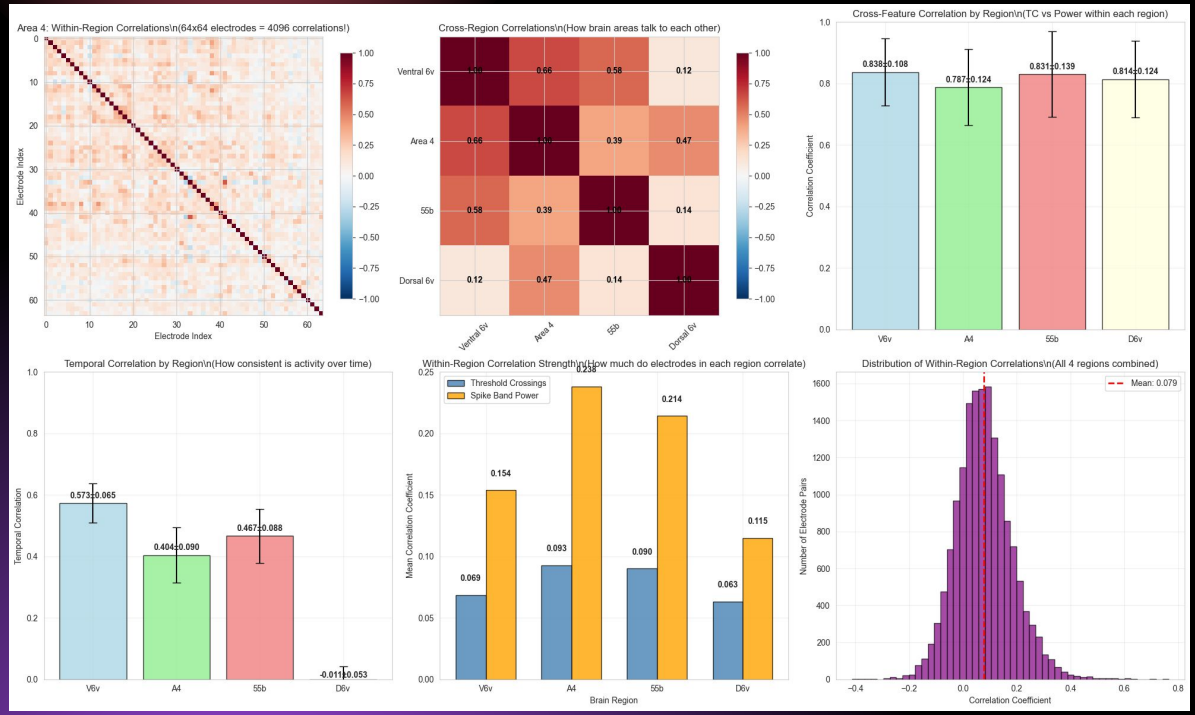
- Spatial patterns matter: Electrode location important, not just individual activity

- Robustness required: High variability necessitates noise-resistant models



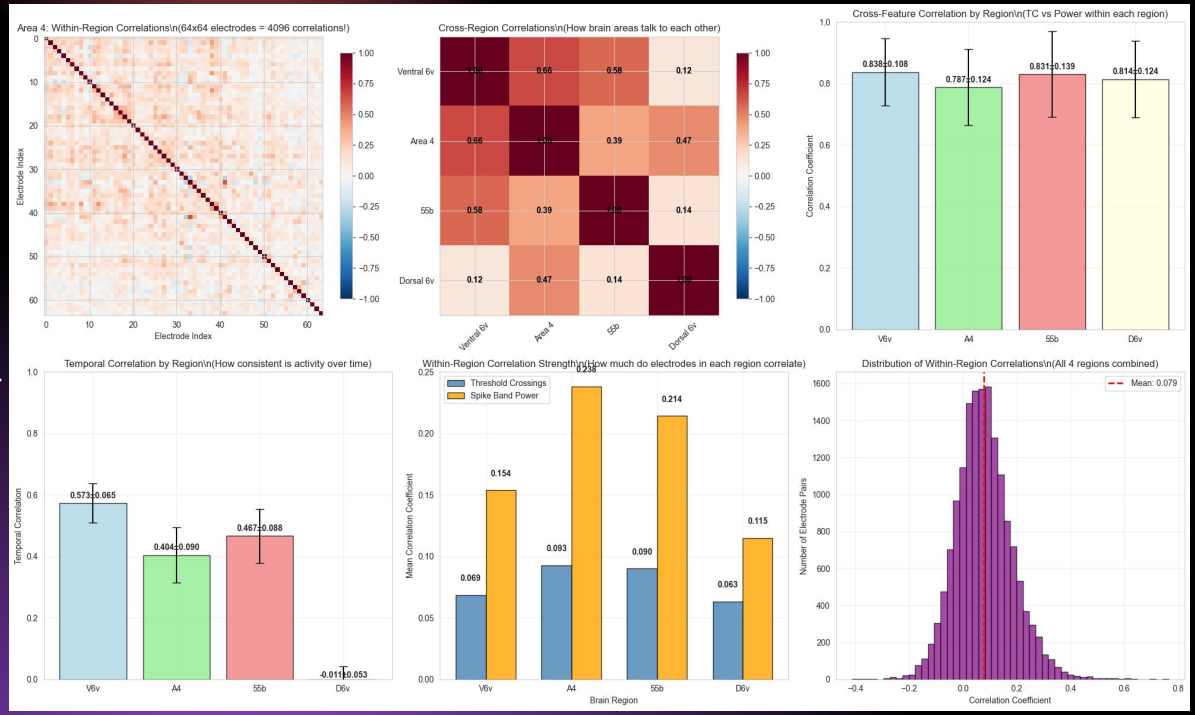
Correlation Analysis

- Within-region correlations: Electrodes in the same area show moderate coordination → activity is not random but regionally organized.
- Cross-region communication: Strongest links between Ventral 6v ↔ Area 4 (r=0.66) and Ventral 6v ↔ 55b (r=0.58). Dorsal 6v shows weak links, suggesting a more isolated role.
- Feature-type validation: Across all regions, SBP correlates slightly more strongly than spikes (0.79–0.84), reinforcing its value for decoding.
- Temporal stability by region: Ventral 6v and 55b remain stable across months; Dorsal 6v is far less reliable.
- Overall correlation distribution is broad: many weak links, some strong pairs → encoding is a complex network rather than a uniform signal.



Correlation Analysis

Modeling Implications:
- Regional teamwork: Focus on regional averages rather than individual electrodes
- Cross-region networks: Model interactions between Ventral 6v, Area 4, and 55b
- Feature simplification: High TC-SBP correlation allows feature reduction
- Temporal modeling: Leverage stable regions (Ventral 6v, 55b) for consistent decoding
- Spatial hierarchy: SBP more informative for regional coordination than TC



Linguistic Content Analysis

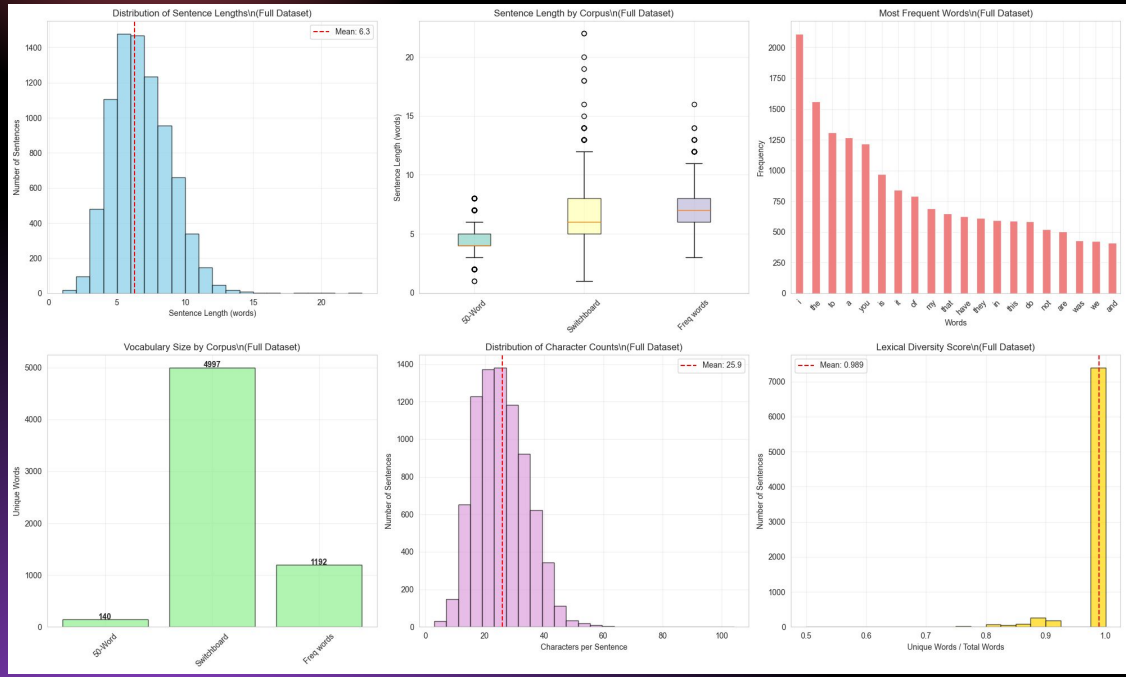
- Sentence lengths are short (avg 6.3 words), but variable (1–22). Switchboard shows the widest range; the 50-word corpus is tightly centered around ~4–5 words.

- Most frequent words are function words (“I,” “the,” “to”), meaning the decoder must reliably distinguish high-frequency, low-information tokens.

- Vocabulary complexity is high: from a fixed 140 words (50-Word list) to nearly 5,000 in Switchboard.

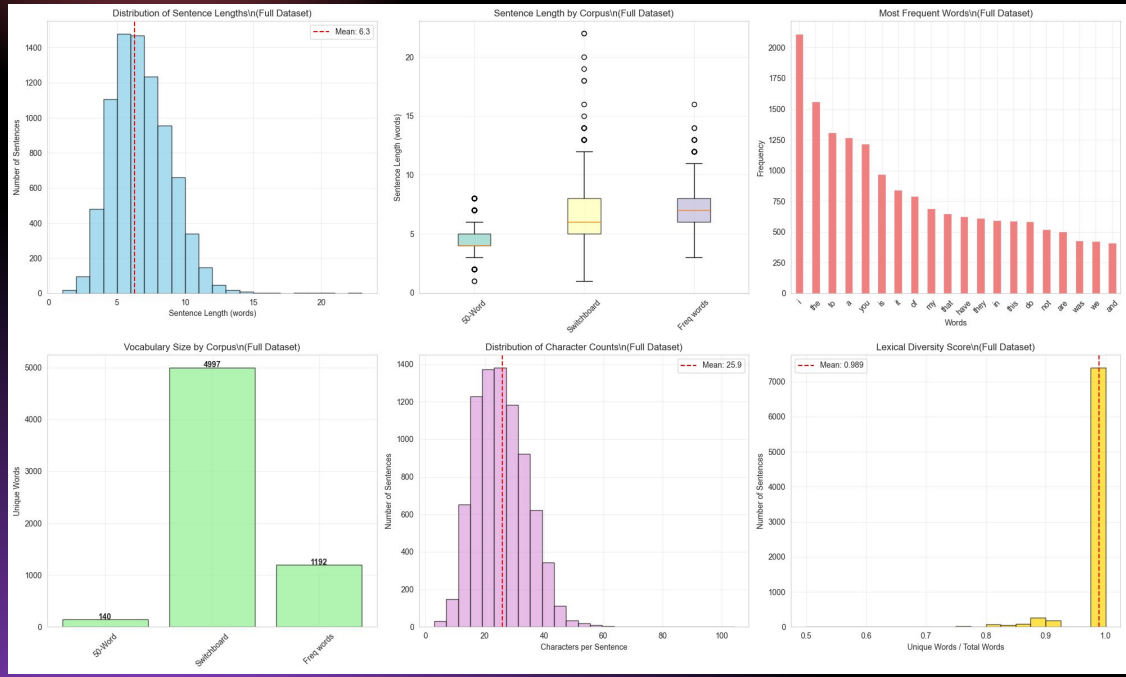
- Character counts also cluster around short utterances (~26 characters), which supports the feasibility of character-level decoding.

- Lexical diversity is nearly maximal (~0.99), meaning sentences rarely repeat words internally. Every word matters, limiting redundancy that might help decoding.



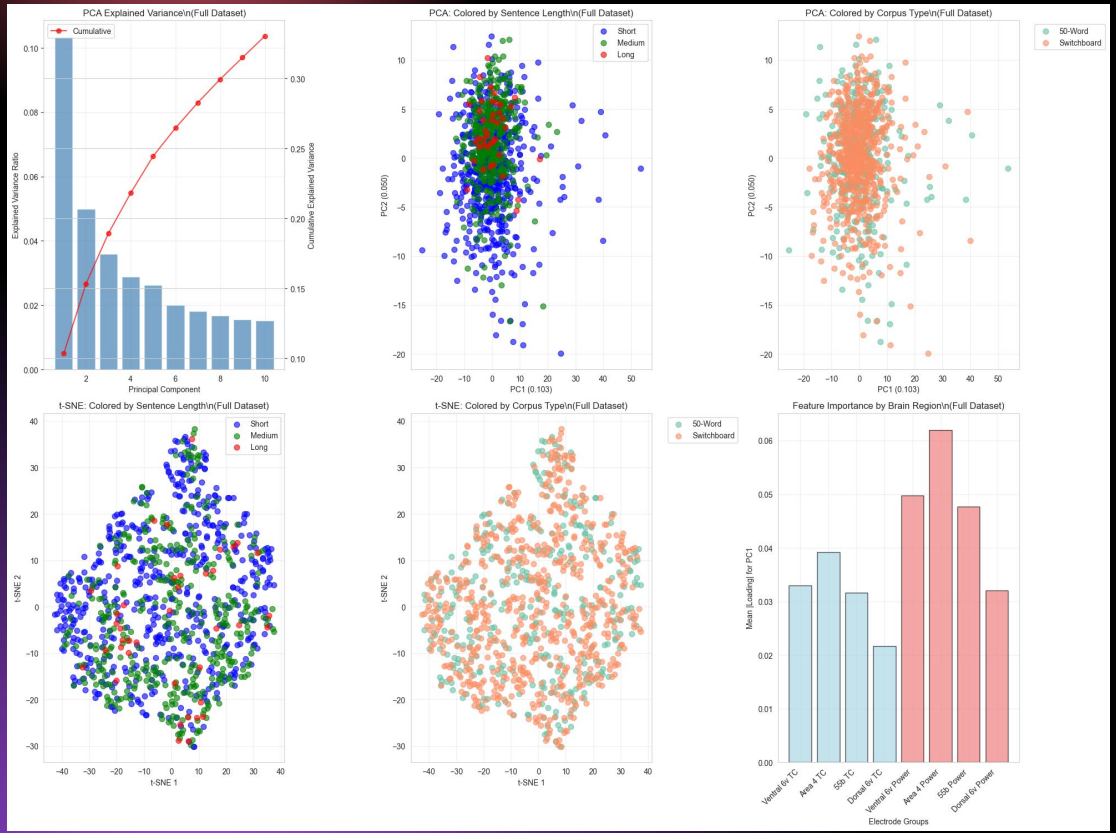
Linguistic Content Analysis

Modeling Implications:
- Multi-granularity needed: Both word-level and character-level decoding approaches
- Corpus-aware training: Significant differences require careful data handling
- High vocabulary challenge: 5,199 words requires sophisticated output layers
- Unique word encoding: High diversity suggests rich neural representations
- Function word focus: Common words need special attention in neural decoding



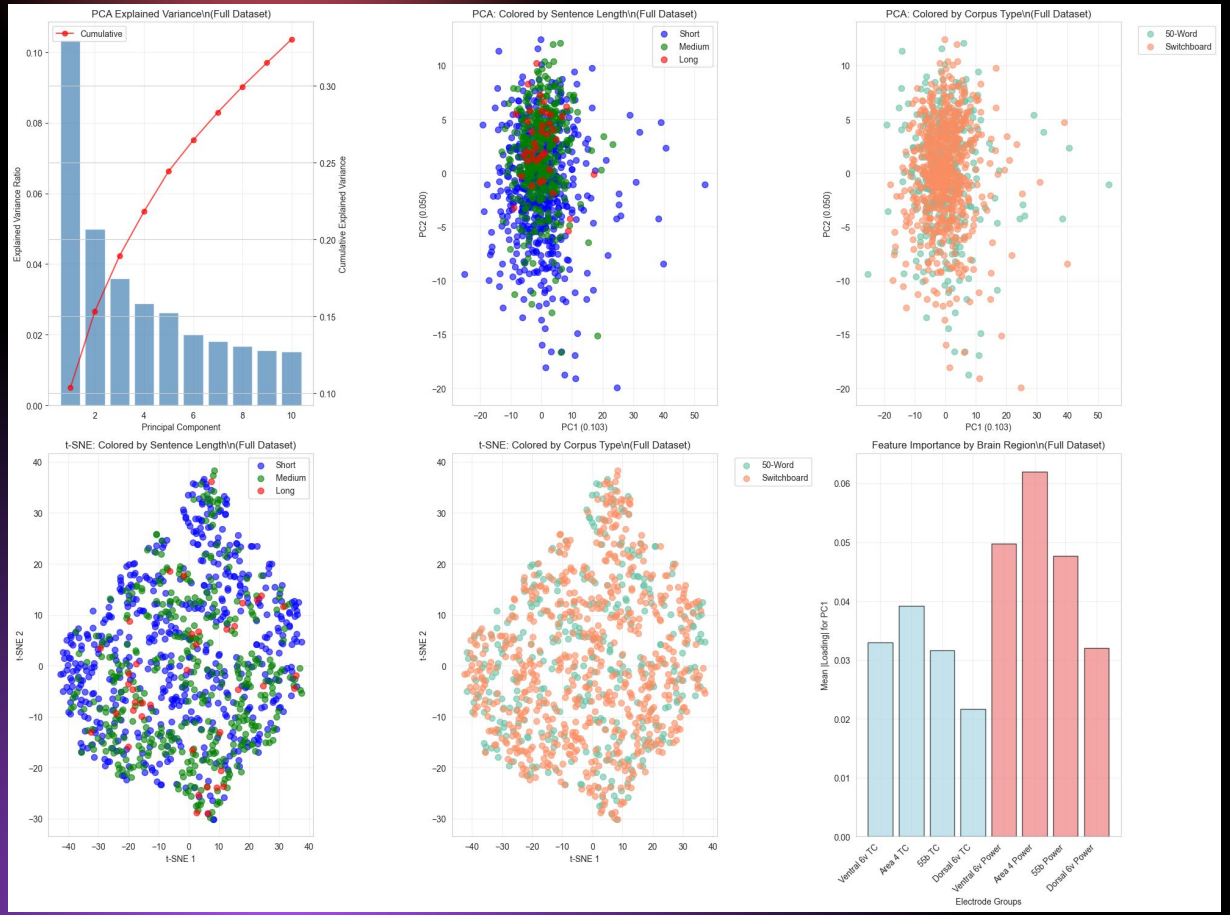
Dimensionality Analysis - Is There Structure?

- PCA results: PC1 explains ~10%, first 5 PCs explain ~24%, first 10 PCs ~30%. Variance is broadly distributed → signals are high-dimensional.
- By sentence length: Short, medium, and long sentences overlap heavily in low-dimensional projections → complexity is encoded diffusely, not separable linearly.
- By corpus type: Switchboard and 50-word sentences intermingle, showing corpus identity is not strongly represented. Good for cross-corpus generalization.
- Feature loadings: Power features (esp. Area 4, Ventral 6v, 55b) dominate top PCs, suggesting SBP features capture the most variance.
- t-SNE: Confirms lack of distinct clustering → brain encoding reflects speech content directly, not dataset source or sentence length.



Dimensionality Analysis - Is There Structure?

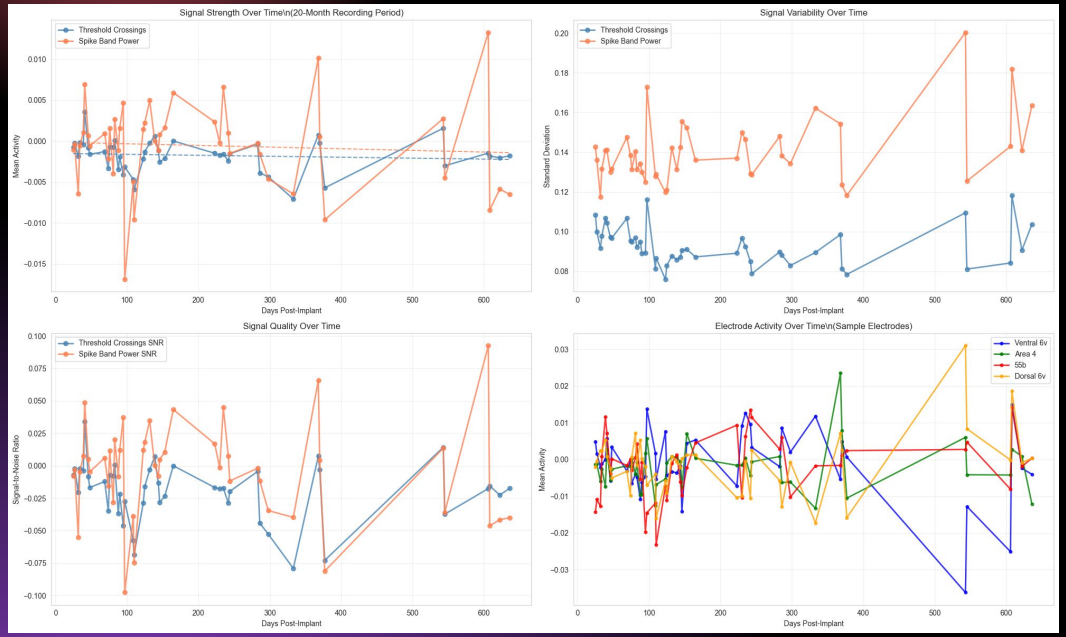
MODELING IMPLICATIONS:
The lack of distinct clustering by corpus type suggests that the brain's encoding primarily reflects the content of speech rather than its source, meaning models may not need to explicitly handle corpus differences (?)



Temporal Stability Analysis

This line plot tracks the Signal-to-Noise Ratio (SNR) for both feature types across recording sessions over the 20-month period

Demonstrates the stability of signal quality over time, which is a positive indicator for the feasibility of training a single model that performs well across the entire recording duration without frequent recalibration (suggests our model should work consistently and good news for clinical applications)



Our Baseline Approach

Our Baseline Approach:
We trained simple Logistic Regression models on our comprehensive dataset to establish fundamental decodability of linguistic features from neural signals.

Classification Tasks:
- Sentence Length Classification: Binary classification distinguishing "Short" (≤6 words) vs "Long" (>6 words) sentences
- Corpus Classification: Binary classification identifying source corpus ("50-Word" vs "Switchboard")

Why These Tasks Matter:
- Sentence Length: Tests if neural activity encodes linguistic structure and complexity
- Corpus Classification: Tests if neural signals contain corpus-specific patterns or are generalizable
- Baseline Performance: Establishes minimum expected performance for more complex models

Model Configuration:
- Algorithm: Logistic Regression with L2 regularization
- Features: 512 neural features (256 TC + 256 SBP) averaged across trial duration
- Train/Test Split: 70/30 stratified split preserving class distributions
- Cross-validation: 5-fold CV for robust performance estimates
- Preprocessing: StandardScaler normalization across all features

Feature Engineering:
- Trial-averaged features: Mean activity across 25-second trial duration
- Regional grouping: Features organized by brain region (Ventral 6v, Area 4, 55b, Dorsal 6v)
- Dual feature types: Both Threshold Crossings and Spike Band Power included
- No temporal dynamics: Static features only, testing spatial pattern decodability

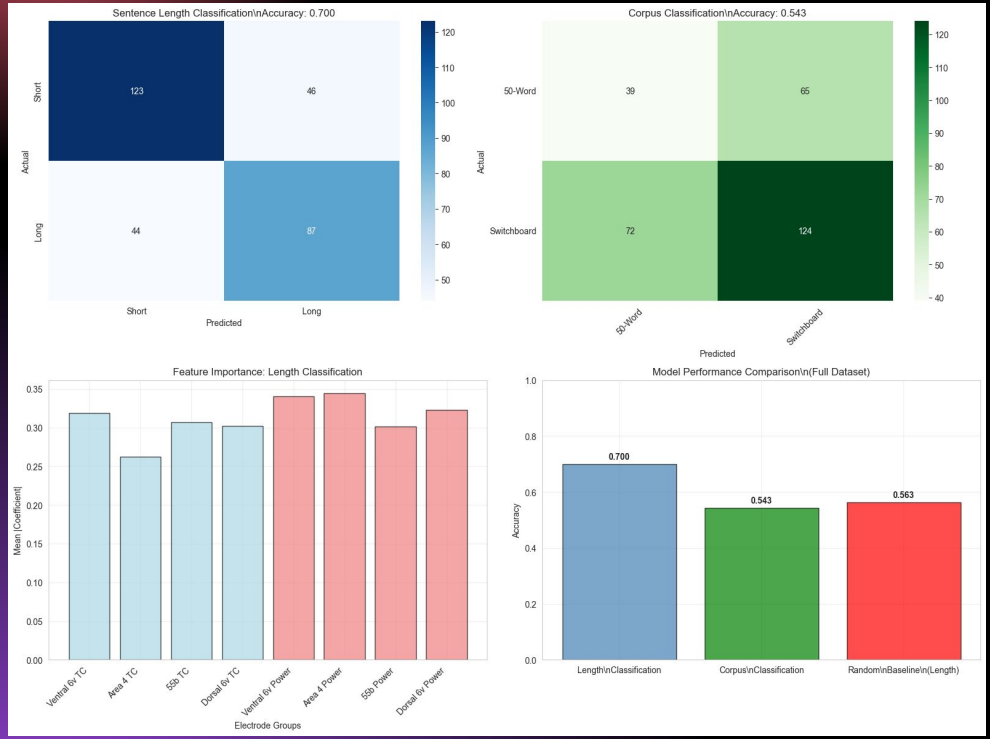
Critical Insights from baseline training !

What the Brain Encodes:

- Sentence length: Strongly encoded in neural activity patterns
- Linguistic structure: Brain processes sentence complexity, not just individual words
- Cross-corpus consistency: Neural representations generalizable across text sources
- Spatial patterns: Regional activity differences important for decoding

What the Brain Doesn't Encode:

- Corpus identity: Neural signals don't distinguish between different text sources
- Simple amplitude changes: Sentence length not encoded as overall activity level
- Individual electrode signals: No single electrode sufficient for classification
- Linear relationships: Complex non-linear patterns required for decoding



Hard Problems We're Still Figuring Out

Next Steps:

- Implement regional feature extraction (Ventral 6v, Area 4 focus)
- Develop temporal models (RNNs/Transformers for time series)
- Build attention mechanisms (weight features by importance)
- Target sentence length decoding (improve from 70% to 90%+)
- Validate cross-corpus generalization (ensure robust performance)

THANK YOU

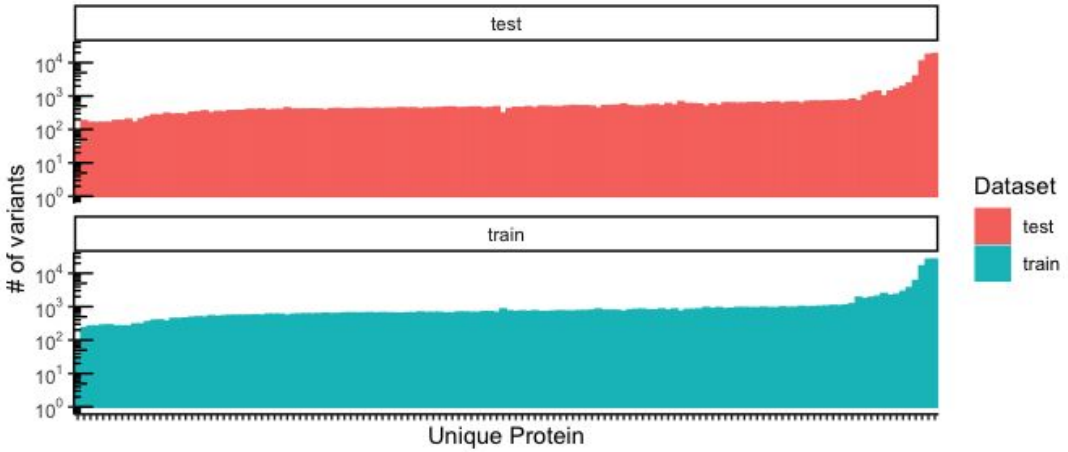
Any questions?

MaveDB Exploratory Data Analysis

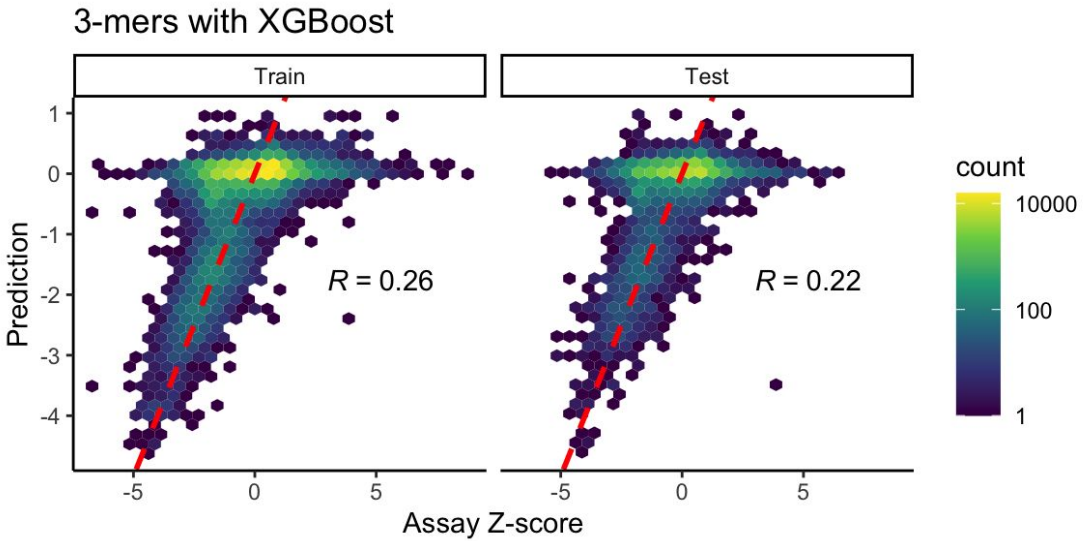
Team Off the Deep End
2025/10/02

MaveDB protein function prediction

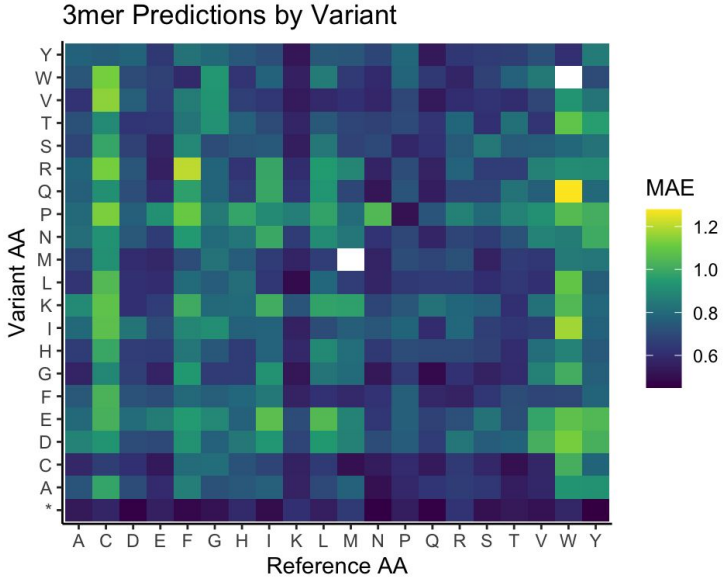
- 136 Unique proteins
- 204 Unique scoresets
- All proteins are in training and testing set
- All but 3 assays are in both training and testing set
- Median variant x scoreset ~ 700
- Max variant x scoreset ~ 25,000
- Min variant x scoreset ~ 30
- Train/test is a 60/40 split on average



Baseline Modeling reveals challenge of the problem



- Z-score the outcome by assay
- Encode proteins as 3mers
- Train tree-based XGboost models, tuned with 3-fold CV



- Task is easy for stop codons
- Hard for mutations to Proline
- Hard for mutations from Cysteine, Phenylalanine, and Tryptophan

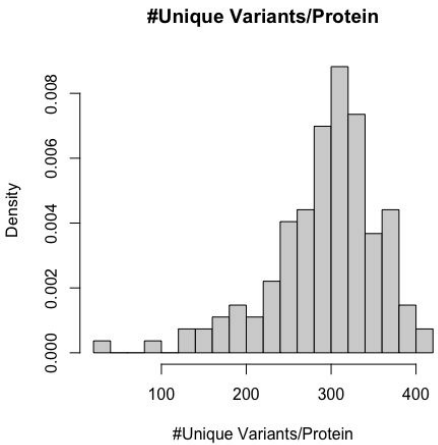
ESM enables embedding of proteins based on a large set of known protein sequences

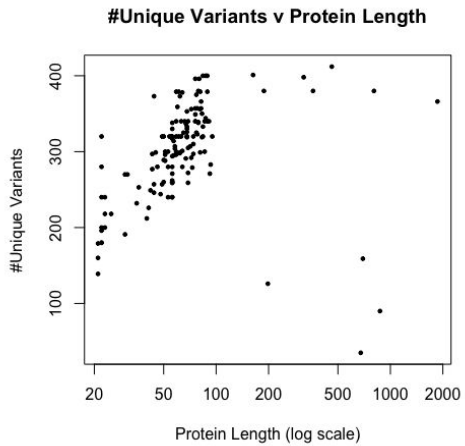
- Sequence Access:
  - Parse out of json using mavedb web interface for experiment sets
    - https://api.mavedb.org/api/v1/score-sets/urn:mavedb:00000069-a-2
  - Use ensembldb to on “ensp” column (eg. ENSP00000252519.3)

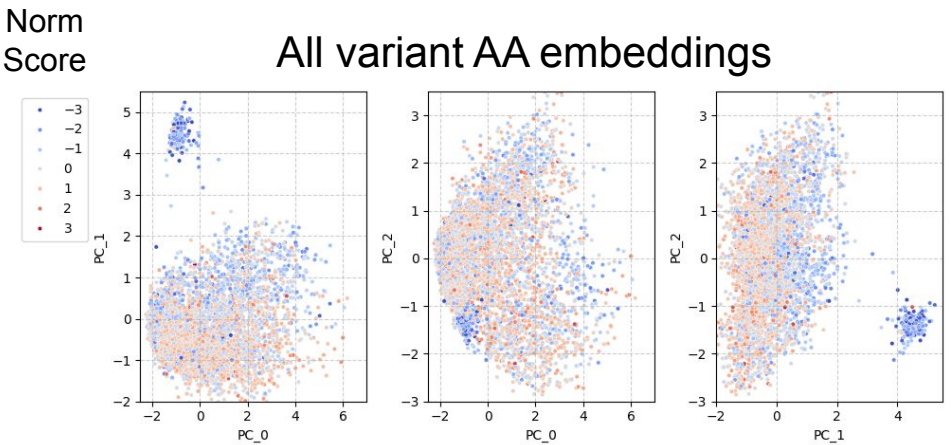
- ESM Based Sequence Embedding
  - Required Packages: transformers (Hugging Face), pytorch
  - Load model and tokenizer from Hugging Face via. Transformers
    - We used facebook/esm2_t6_8M_UR50D, use larger in the future
  - Tokenize and input original and variant sequences to ESM model
  - Save both variant AA and pooled full sequence embeddings

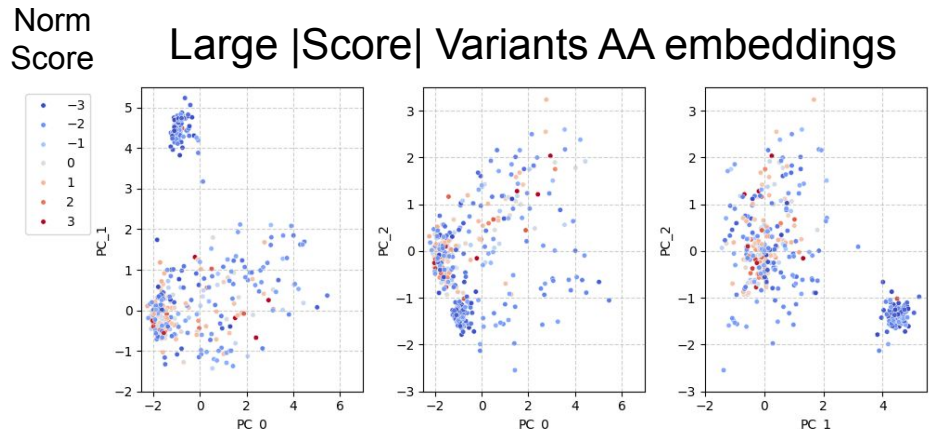
Proteins Variants separate in the ESM space by score

- Around 300 variants/protein
- In general, positive increase in #variants with protein length
- Using ESM-2 embeddings show promise to improve models to predict variant scores



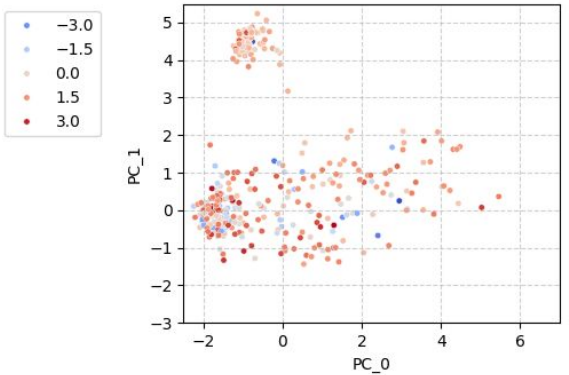


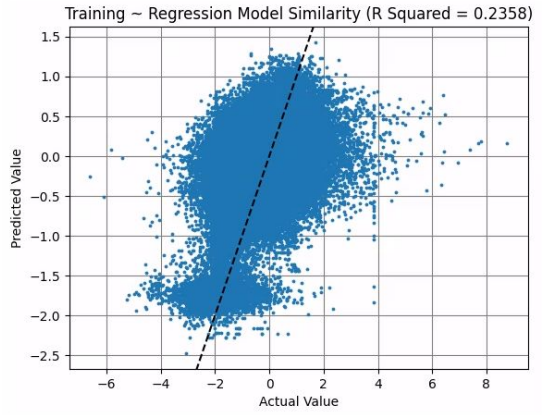




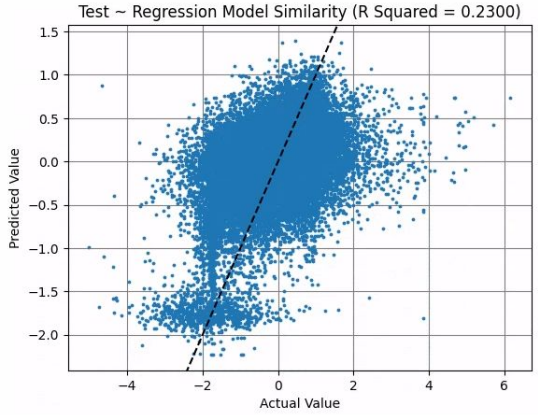
Baseline Linear Modeling of ESM Embeddings

- ESM-8B creates 320 parameter embeddings
- Simple Linear regression with sklearn:
  - 20 tests (chosen randomly) as out of distribution hold out
  - Remaining tests used to produce train test split
    - 4:1 train:test splits balanced on test type

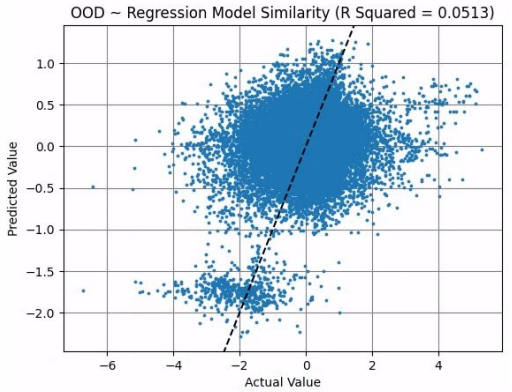




R = .4832



R = .4805



R = .2822

Next steps

- Iterate on our modeling to include the ESM embeddings as features
  - Explore larger ESM model embeddings
  - Fine tune ESM models
- Look into other features
  - Molecular descriptors (Ex. protr)
- Try other regression models
  - Partial least squares
  - Random forest & extensions
  - Neural networks with fine tuning (ESM)
- Further debugging of hard to predict variants

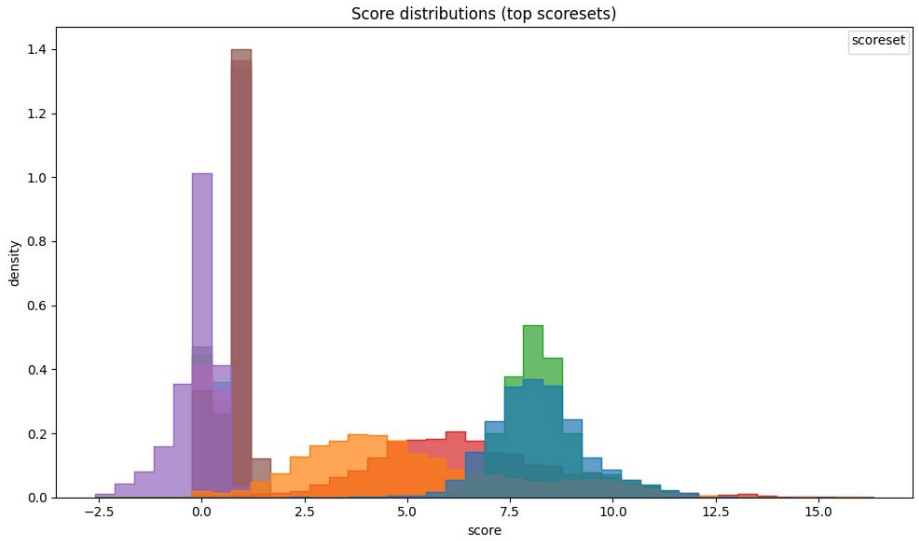
Data

an initial look

Raw data:
- Train: 178,554 rows, Test: 118,421 rows
- Slightly imbalanced data (39% of data is from 3/200 datasets)
- Data consists of IDs: features can't be used for training

Cleaning:
- No missing, n/a, or duplicate data

Processing:
- Ensembl API: full sequence, variant effect predictor (VEP)
  - Retrieve more data based on the substitution (consequence type, impact, biotype, etc.)
- ESM C embeddings: pretrained model for embedding full protein sequences



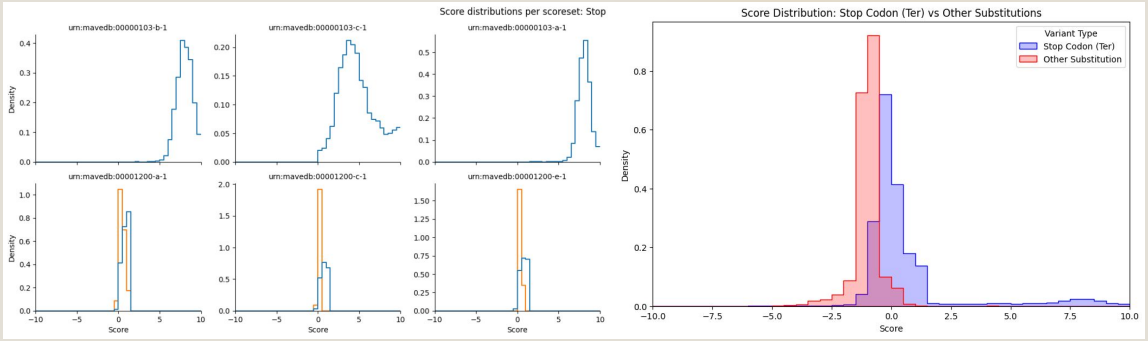
Scoreset Analysis

- **What is a scoreset?**
  - Grouping of experimental measurements.
  - ID for the protein being tested and the measurement being taken.
- **What we noticed:**
  - Strong correlation between scoreset and score range.
  - Each scoreset appears to function somewhat independently.
- **Applications:**
  - Vectorized scoresets as a major parameter.
  - Baseline model

Features with Significance

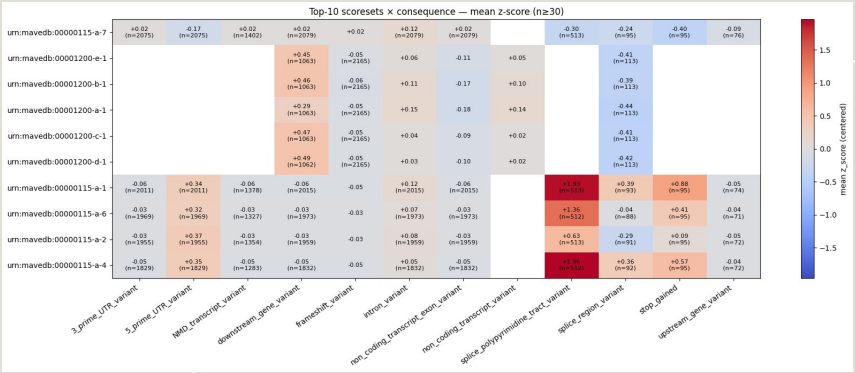
1 Termination

- Some experiments terminate part of the protein instead of swapping amino acids.
- Termination codon substitution correlates with an increase in score (relative to other substitutions).



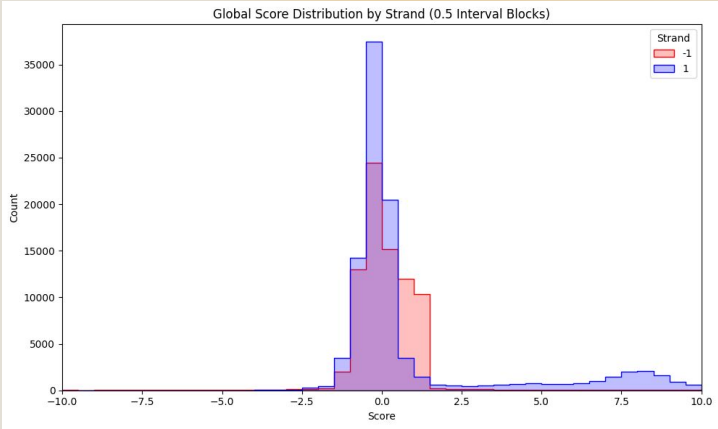
2 Consequence Variant

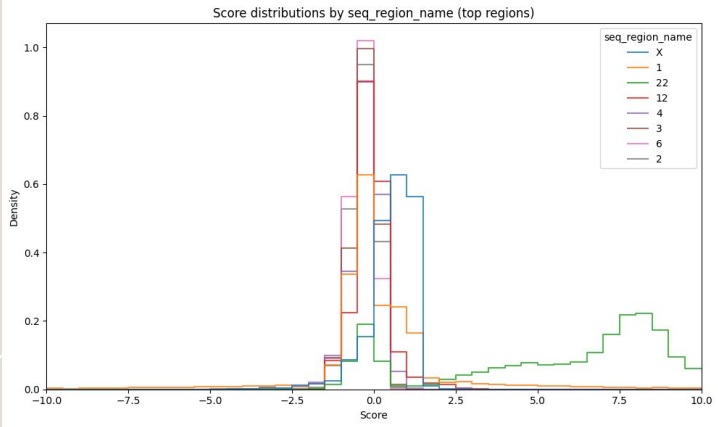
- Each row is a scoreset, each column a consequence variant predicted by Ensembl's Variant Effect Predictor (VEP); colors show mean z-score relative to that scoreset (red=above avg, blue=below).



Features without Significance

- **Strand (–1 or 1)**
  - Encodes whether a AA is positioned within a β-strand of a protein's secondary structure
  - The values –1 and 1 indicate the relative orientation of the strand within a β-sheet
- **Impact**
  - From Ensembl's VEP data—presumed impact of this specific substitution
  - Almost 100% of rows were listed as "High" or "Modifier," but neither category correlated with score
- **Sequence Region Name**
  - Although it has a strong correlation with score, this factor is already identified within the scoreset.
  - Most likely not necessary as a model parameter.





Baselines & Next Steps

Baseline 0: Random Standard Deviation

50.3 MSE

Random numbers taken from a standard deviation determined by the “score” label

Baseline 1: Mean of Scoresets and Substituting Protein

8.3 MSE

Uses mean score from training data of scoreset and alt_short. If encountering a (scoreset, alt_short) pair that was not in training data, use scoreset mean score.

Next Steps:

Finish data analysis for a few remaining features

Finalize data (VEP, Embeddings) into single files for training our first machine learning model

MaveDB

Exploratory Data Analysis

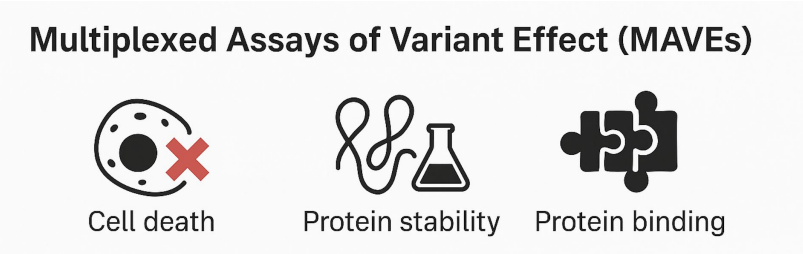
Team DBMavens
MLM25

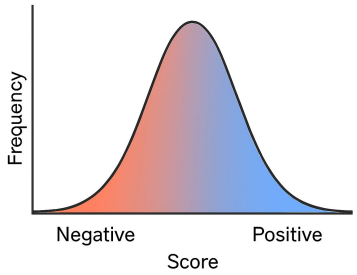
Introduction

- Multiplexed Assays of Variant Effect (MAVEs) provide high-throughput measurements of how thousands of variants impact gene function. MaveDB is a growing repository of these datasets.

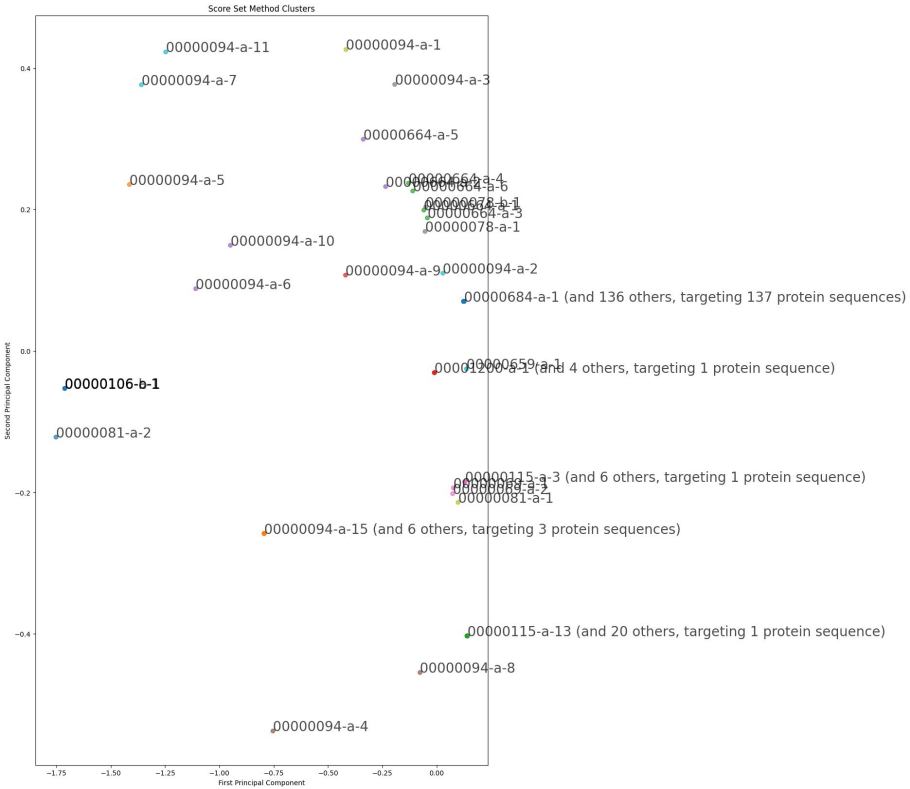
- Each MaveDB entry describes their use of different experimental procedures to understand the functional outcomes of individual substitutions on a protein of interest.

- The numerical score to every amino acid substitution, represents how that change affects - cell death, protein stability, binding capability, etc.



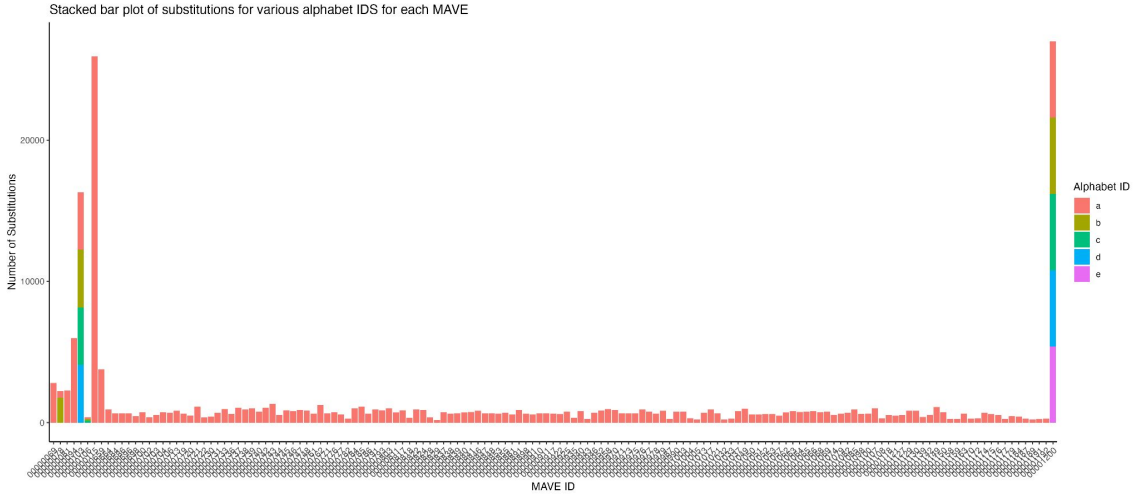


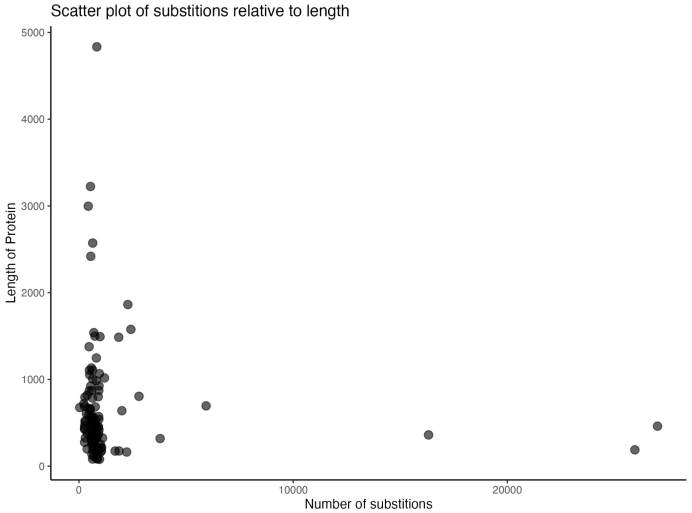
Methods text embeddings, clustered

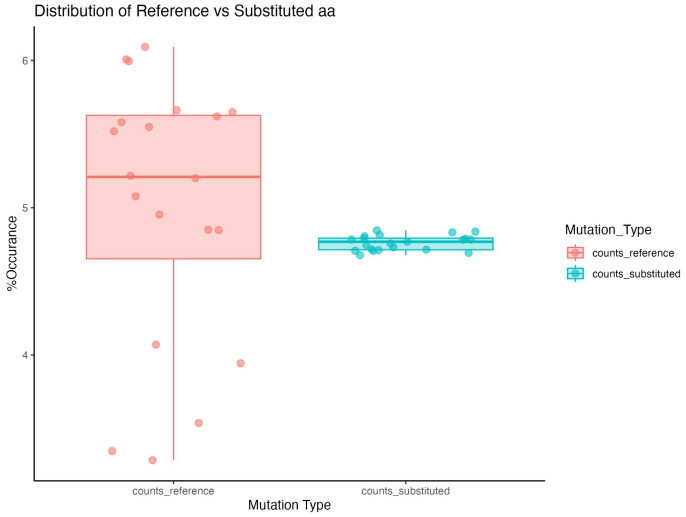


- We have 204 method texts, of which 32 are distinct method texts. 7 are empty. One is repeated 137 times.
- Very short method texts with small differences can end up in different clusters
- Longer method texts with larger differences can end up in same cluster
- Hard to assess automatically the semantic significance of small differences

Exploratory Data Analysis





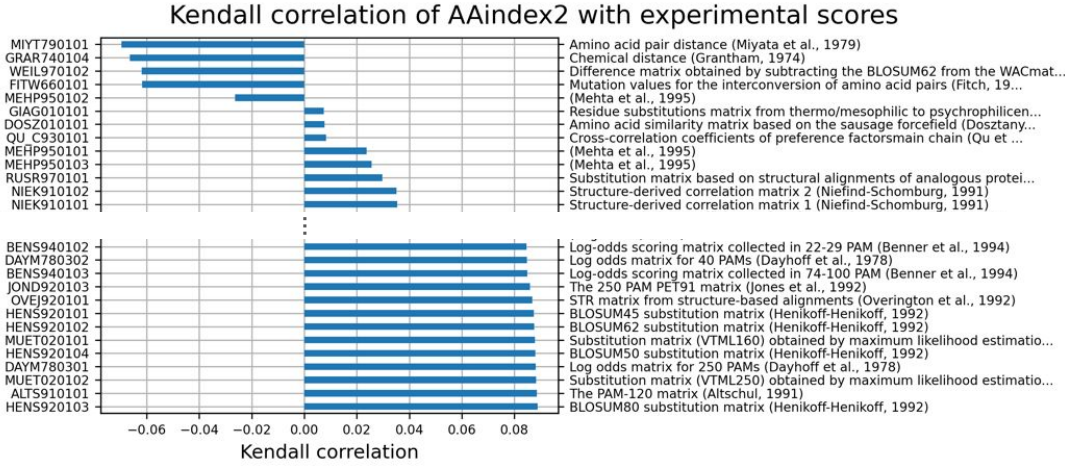


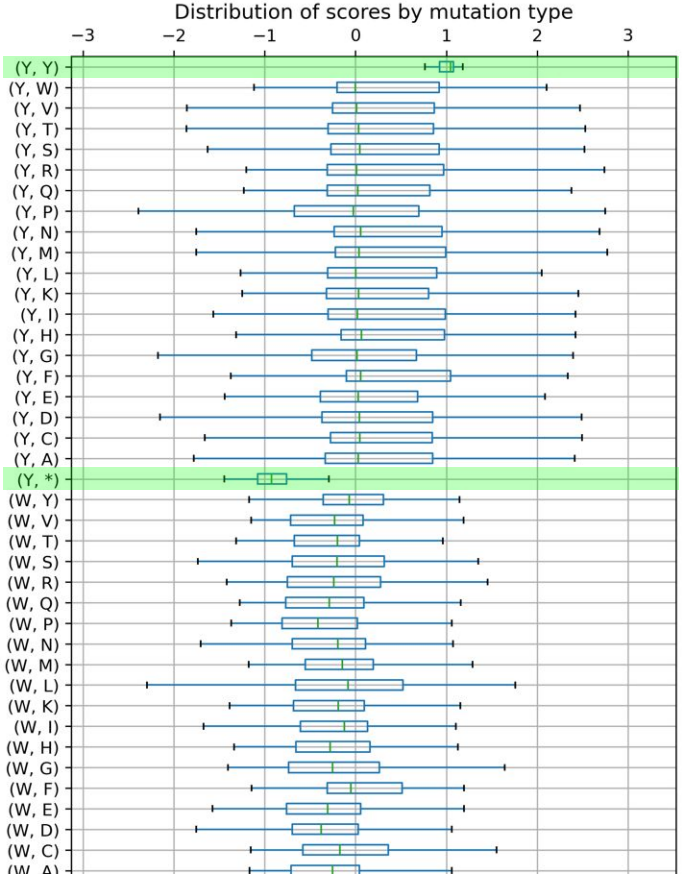
Except for a few MaveDB entries number of substitutions are in the few 100s range.

The number of substitutions are not a function of protein size.

% Counts of amino-acid substituted out has a large variation compared to amino-acid substituted into the protein.

Property correlations and mutation analysis



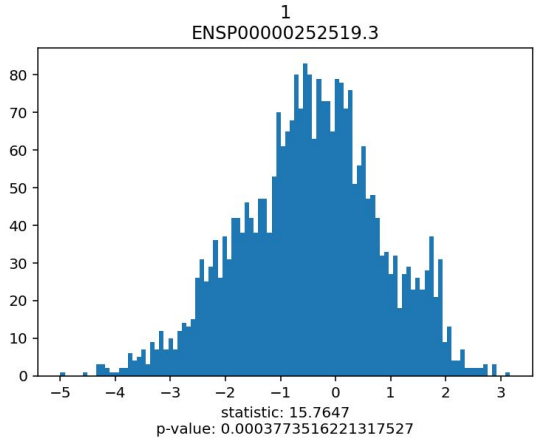


From the training data, every synonymous mutation is scored close to +1 and every truncation is scored close to -1.

When correlating substitution matrix scores to MAVE scores, we find that some are decent (but none are great) at correlating the relative ordering of MAVE scores for the training mutations.

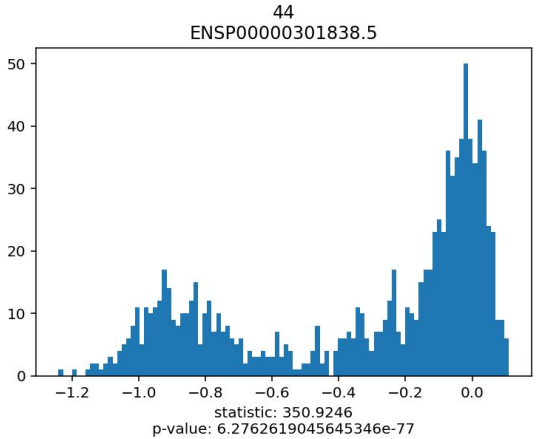
Score Distributions

Unimodal



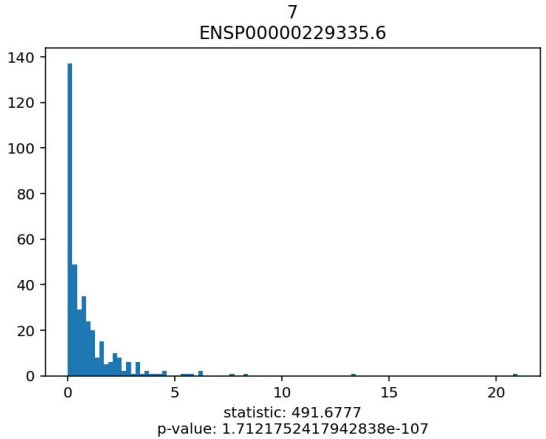
“By using deep mutagenesis, mutations in ACE2 that increase S binding are found across the interaction surface”

Multimodal



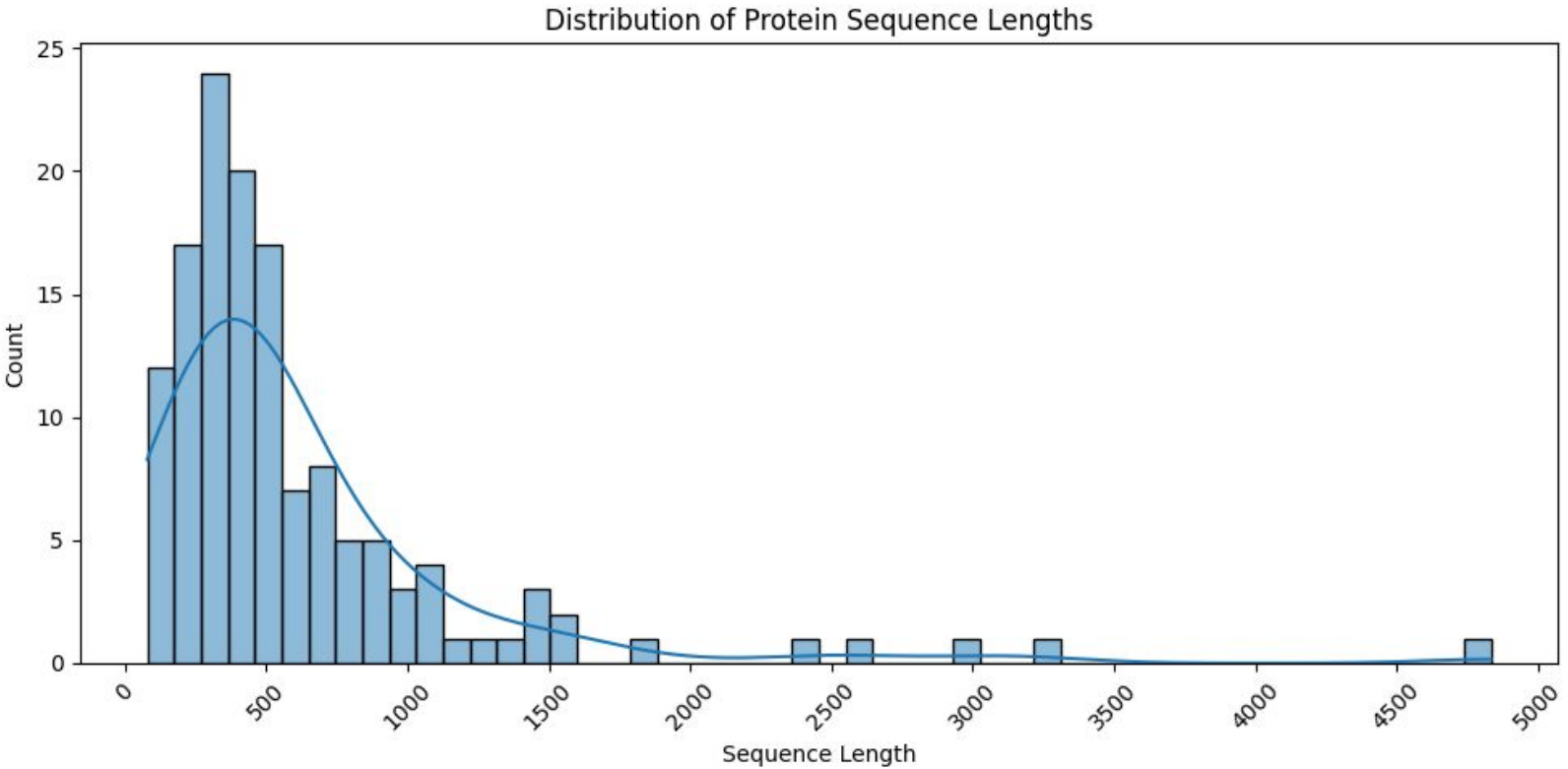
“Pathogenic missense variants reduce protein stability”

Power law



“Mutational analysis revealed dominant characteristics for residues within the loop and additionally yielded enzymatic variants that enhance deaminase activity”

Protein Sequence Length Distributions



Thank you

Mavnificent

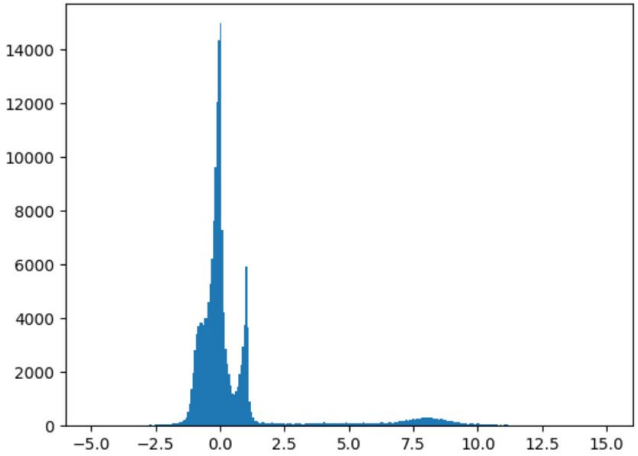
Tyler Thompson, Tejvir Mann, Shenyan Zhang, Lekshmi Thulasidharan

Training Data

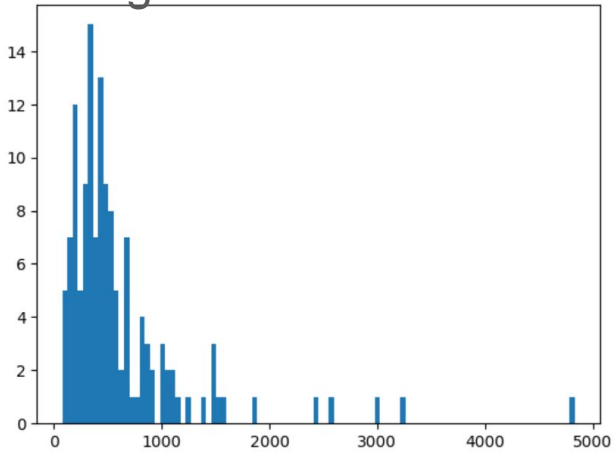
- Sequence, mutation, and functional score
- `Score` is overloaded with different protein functions

Scoreset	Sequence	Mutation_pos	Mutation	Score
urn:mavedb:00000069-a-2	QPFLLRNGANEGFHEAVGEIMSI	76	*	0.54
urn:mavedb:00000069-a-2	QPFLLRNGANEGFHEAVGEIMSI	75	W	-0.36
urn:mavedb:00000069-a-2	QPFLLRNGANEGFHEAVGEIMSI	74	V	-2.44
urn:mavedb:00000069-a-2	QPFLLRNGANEGFHEAVGEIMSI	73	A	0.21

Score Distribution



Length Distribution



Important Protein Info

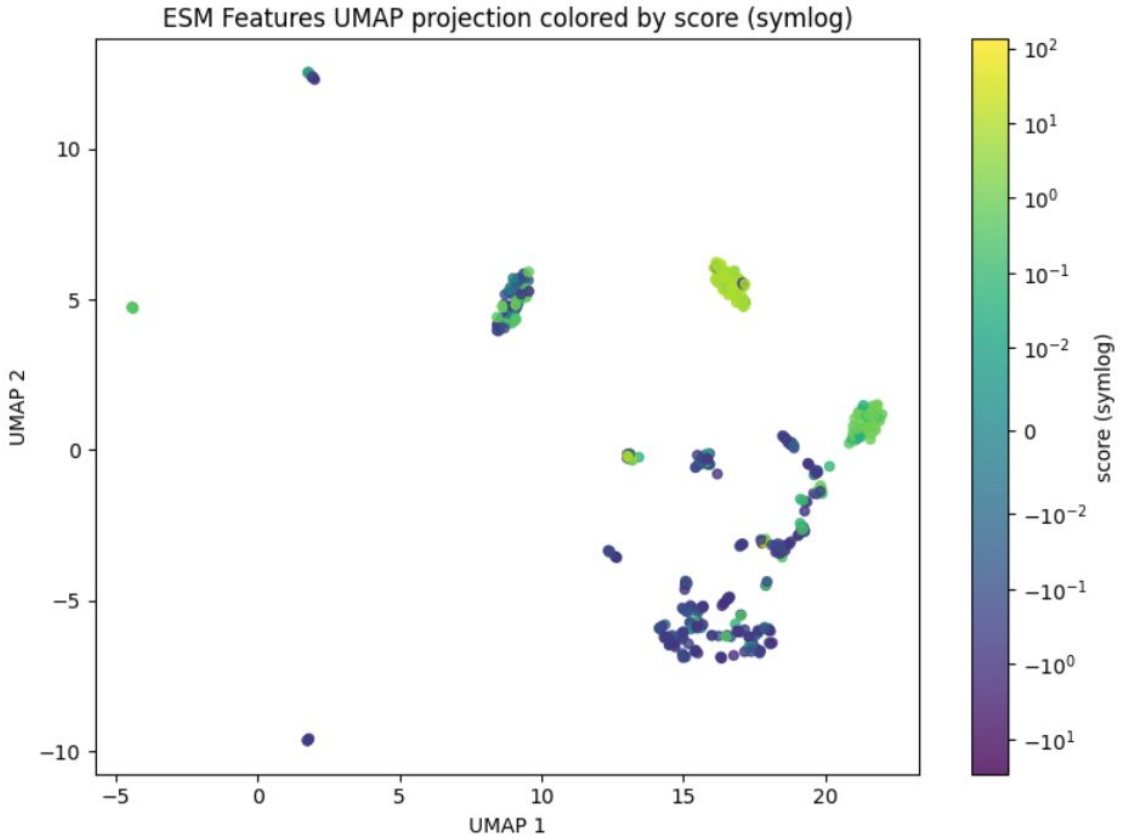
- Every protein in initial dataset targets other proteins
-

Current tools

- One-hot encoding
- ESM2
- Linear regression / MLP

UMap of ESM2 Embedding

- Esm features do well at classifying strong performers out of the box



Next Steps

- Huggingface deepdive.
- Add more diversity of proteins to training data
- bioembeddings.com Toolkit
  - https://docs.bioembeddings.com/v0.2.3/
- MaveDB API/Tools to get more distinguishing info for proteins