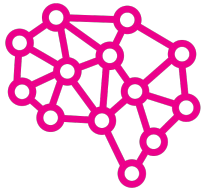


FourthBrain

GLG: Automated Text Analysis for Improved Service Demand

Curtis Pond and Julia Nickerson

Who is GLG?

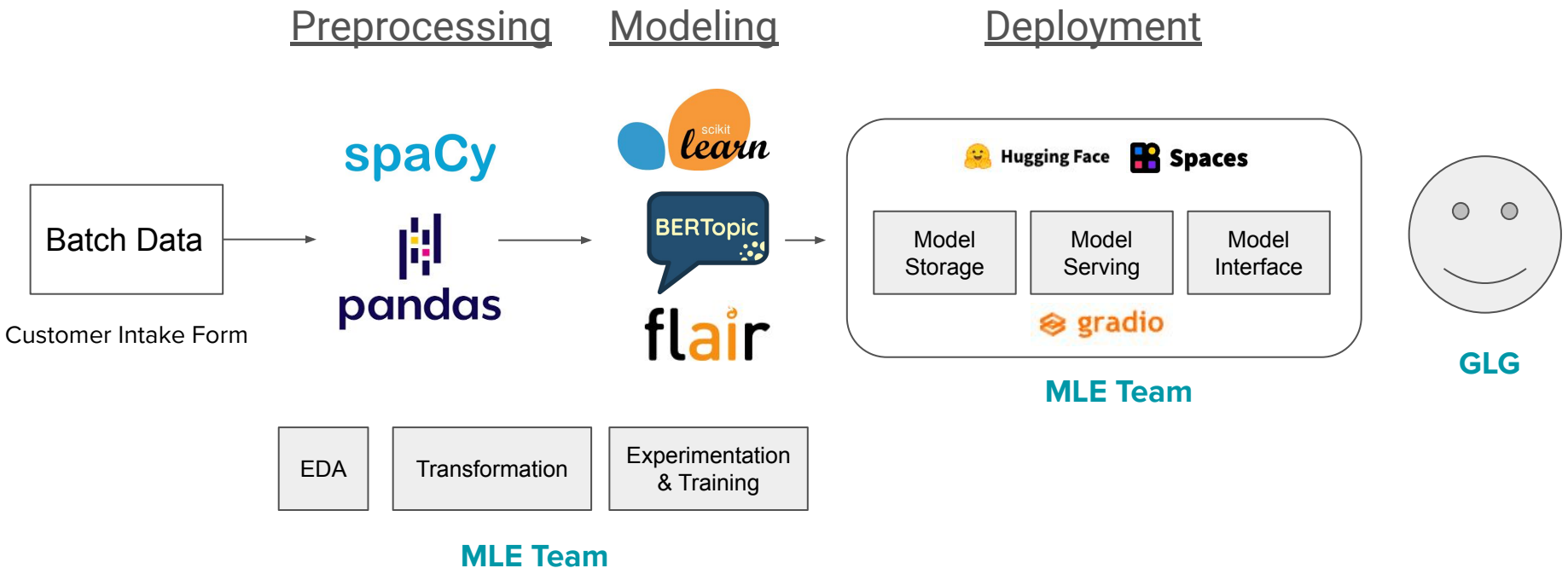
Problem

- Hundreds of requests are submitted daily to GLG via an intake form
- GLG wants to **help people reach experts faster by**:
  - Grouping common topics together
  - Grouping similar client requests together
  - Identifying underlying patterns in the data (NER, time-based patterns)

Problem & Solution

- Hundreds of requests are submitted daily to GLG via an intake form
- GLG wants to **help people reach experts faster by**:
  - Grouping common topics together
  - Grouping similar client requests together
  - Identifying underlying patterns in the data (NER, time-based patterns)
- Metadata — **named entities** — are auto-extracted from intake submissions
- Intake submissions are **categorized by topic** (healthcare and technology)
- Clustering mechanism that shows **topic trends over time**

Workflow



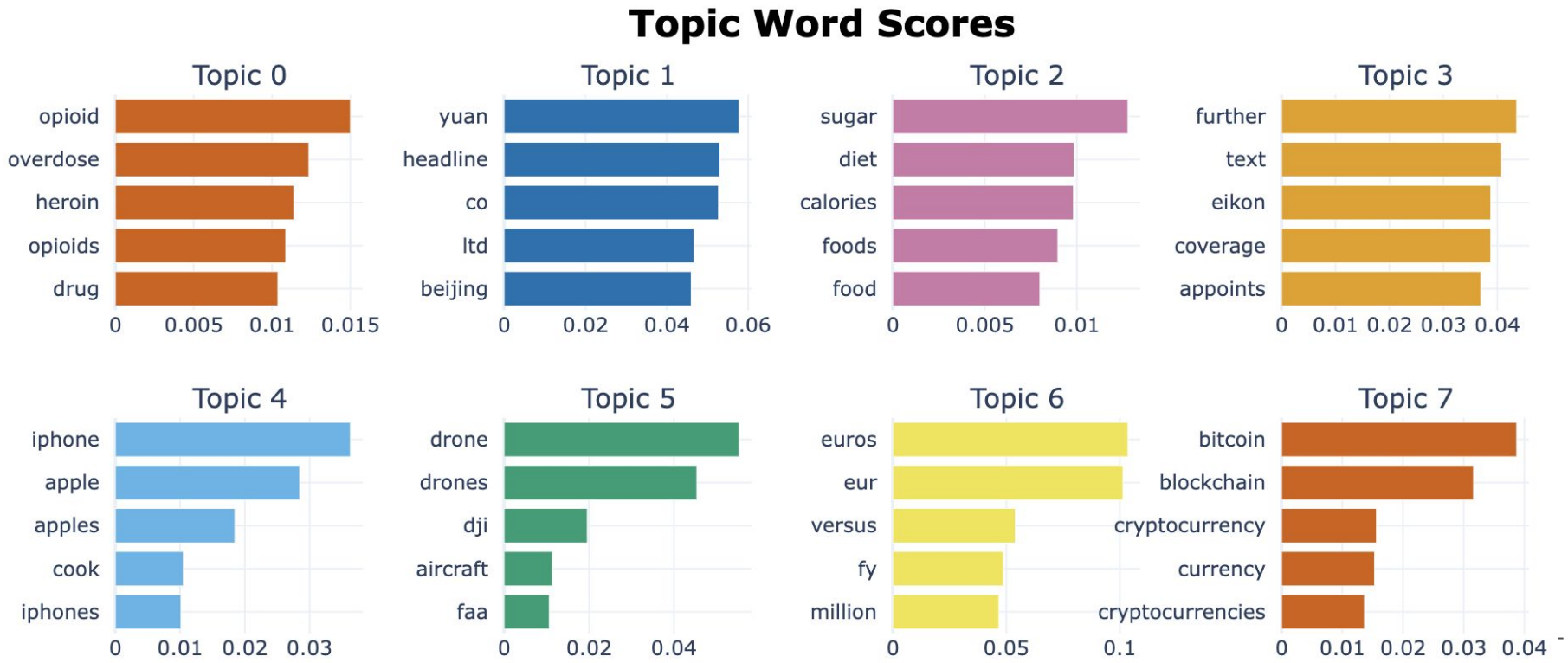
Two Datasets: NER Corpus & All the News 2.0

	NER Corpus	**All the News 2.0**
Description	• 47,959 sentences • Includes each word's part-of-speech (noun, verb, etc.) and NER (geo, org, per, etc.)	• 2.7 million news articles published between 2016 and 2020 • includes date, author, title, and publication name
Size	15 MB	9 GB
Labels	Labeled	Unlabeled
Task(s)	Supervised learning (named entity recognition)	Unsupervised (clustering) and supervised (text classification) learning

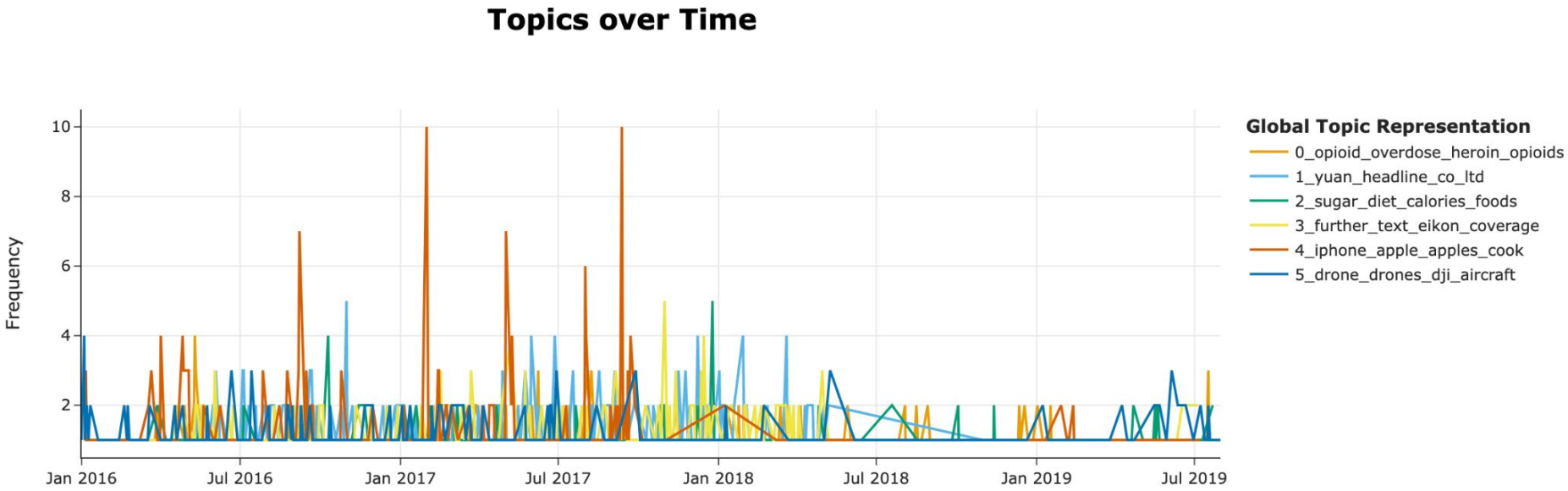
Modeling — *Supervised Learning*

- **Named Entity Recognition (NER)**
  - **Data**: NER Corpus
  - **Model**: LSTM RNN
  - **Library**: Flair
  - **Baseline Result**: 0.81 F1 score (using a subset of data)
  - **Latest Result**: 0.86 F1 score (using all data)
- **Text Classification**
  - **Data**: All the News 2.0
  - **Model**: Logistic Regression
  - **Library**: scikit-learn
  - **Baseline Result**: 0.95 F1 score (using 25k articles)
  - **Latest Result:** 0.92 F1 score (using 60k articles)

Modeling — *Unsupervised Learning*

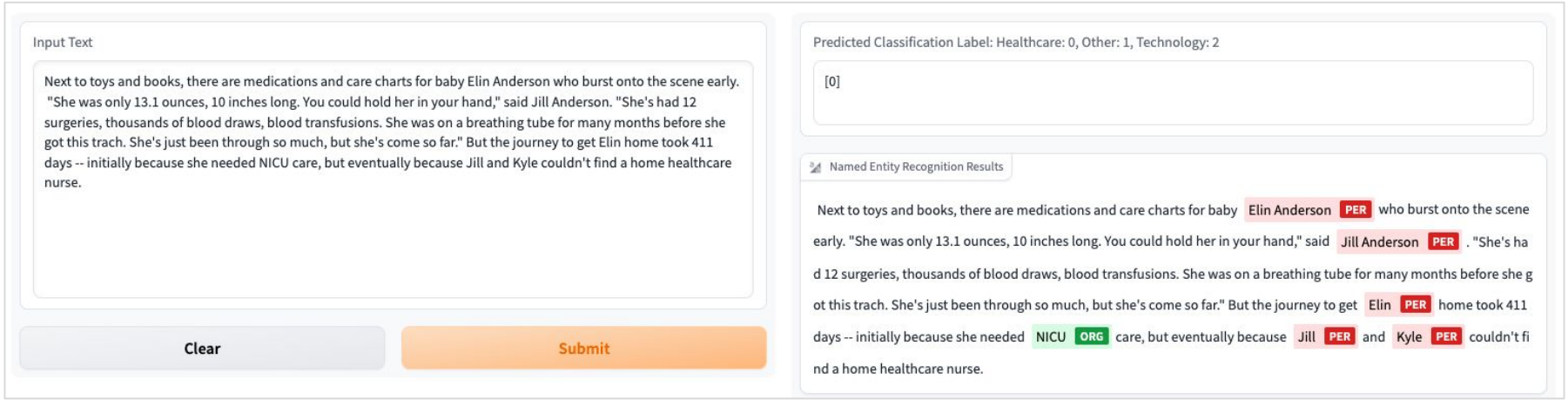


Modeling — *Unsupervised Learning*



Demo

- Our demo's web interface is built using **Gradio**, and hosted on **Hugging Face Spaces** at https://huggingface.co/spaces/curtpond/mle10-glg-demo
- Let's check it out!



Concluding Thoughts

- A challenging project involving multiple different tasks (NER, text classification, unsupervised topic modeling)
- Considered connecting different models together sequentially, but went with separate standalone models
- Text classification model explainability analysis was helpful for informing our final approach
- Would have preferred using actual GLG data instead of news articles

Thank You! Questions?