Slides

- Title [30s]
- Outline [30s]
- Problem (i.e. Problem + Why) [1 min]
- Solution (i.e., Success + Audience) [1 min]
- Data + Model (i.e., What, 1st part) [1 min]
- Demo [2 min]
- MLE Stack/Iterations (What, 2nd part) [2 min]
- Conclusions (and lessons learned) [90s]
- Future Work [30s]
- (Thank You & Q&A) [0s]

FourthBrain

GLG: Automated Text Analysis for Improved Service Demand

Curtis Pond and Julia Nickerson

Who is GLG?

- 40% of the Fortune 100
- 7 of the 10 largest global technology firms
- 8 of the 10 largest pharmaceutical companies
- 9 of the 10 largest law firms
- 9 of the 10 leading banks
- 800+ global private equity funds
- 500 public equity investment firms
- 85+ Social Impact Fellows

Outline

- Problem
- Solution
- Data + Model
- Demo
- MLE Stack
- Conclusions (and lessons learned)
- Future Work

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)

Solution (1 min)

- Metadata (named entities) are auto extracted from intake submissions
- Intake submissions are categorized by topic (starting with healthcare and technology)
- Clustering mechanism that shows topic trends over time
- (SME routing? Hit a match and our UI says the submission should be routed to Tech or Healthcare expert?)
- Wouldn’t it be great if a user with this problem had a solution like this!?
- Paint the picture of that future world where the problem is solved.
- Objectively, what would that be like for them? Better, faster, cheaper?
- How do we get there? What would we measure to know if we got there?
- Five bullets are always better than 6.

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.)	• 27 million news articles published between 2016 and 2020 • includes date, author, title, and publication name
Size	~15 MB	~9 GB
Labels	Labeled	Unlabeled
Task	Supervised learning (named entity recognition)	Unsupervised learning (clustering)

EDA for NER Corpus

- Clean dataset; no missing data, formatting, or data type issues
- Generated pandas-profiling report
- Created histograms (number of characters per word, number of words per sentence)
- Modified the data format for easier input into models

part-of-speech

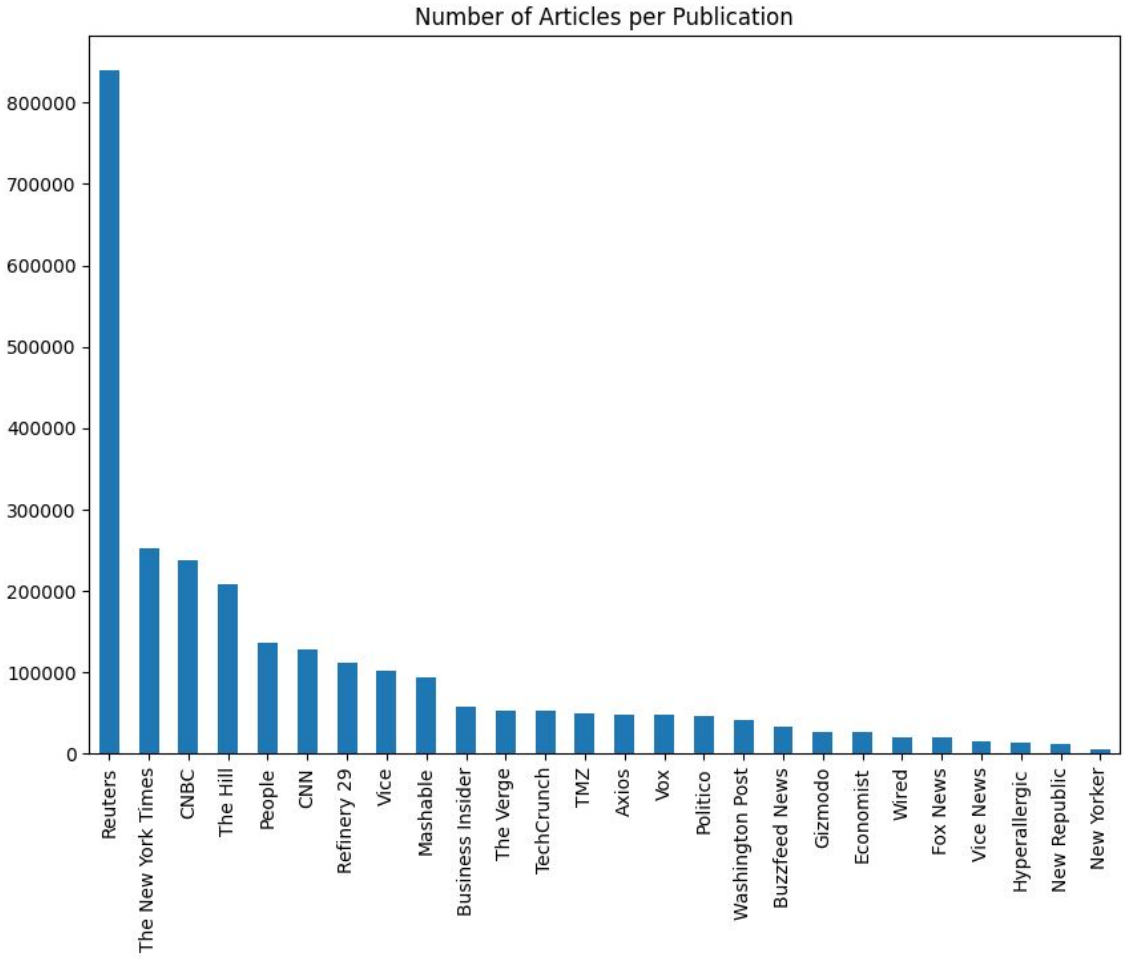
NER

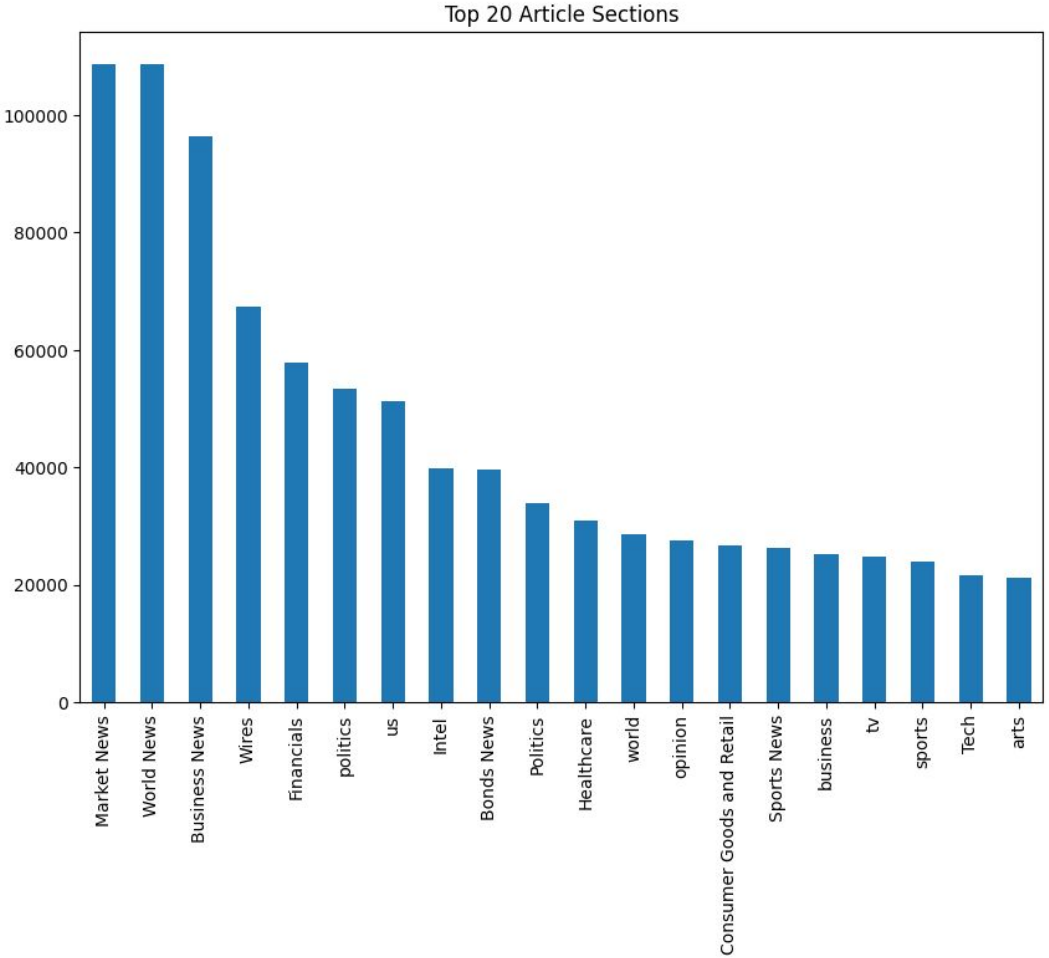
	Sentence #	Word	POS	Tag
0	1	Thousands	NNS	O
1	1	of	IN	O
2	1	demonstrators	NNS	O
3	1	have	VBP	O
4	1	marched	VBN	O
5	1	through	IN	O
6	1	London	NNP	B-geo
7	1	to	TO	O
8	1	protest	VB	O
9	1	the	DT	O
10	1	war	NN	O

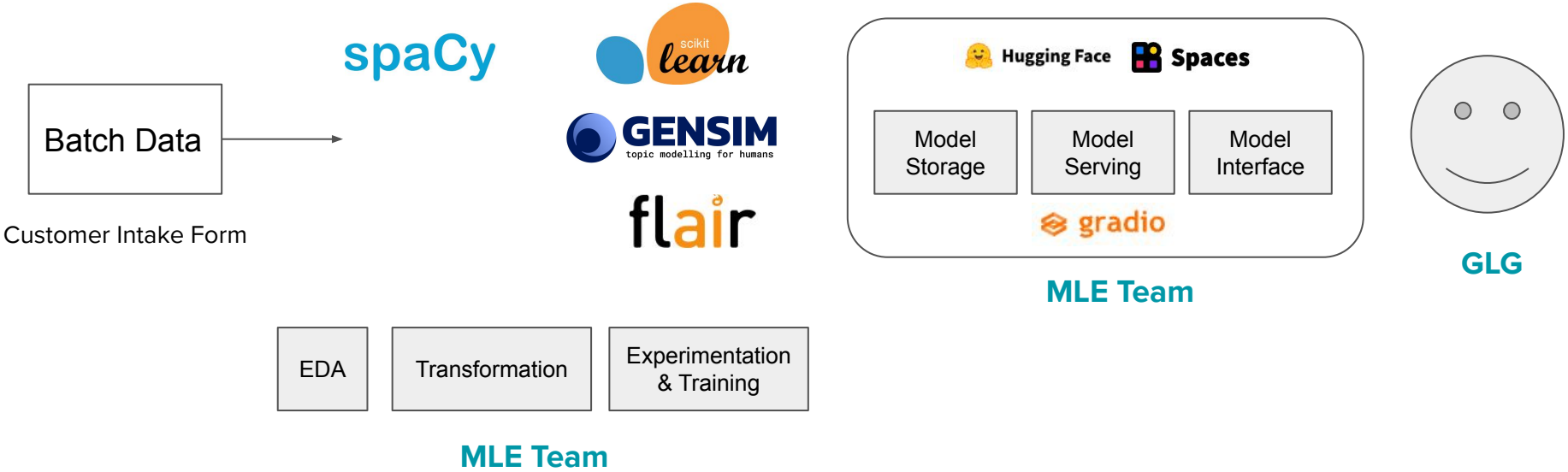
EDA for All the News 2.0

- It was challenging to handle a ~9 GB file (attempted PySpark and pandas chunking)
- Created histograms (number of articles per publication and per year, top article sections, etc.)
- Removed irrelevant columns (URL, author) and removed rows with missing articles
- Performed text analysis to count the number of words and sentences in each article

year	month	day	article	section	publication
2017	Nov	27	Nov 27 (Reuters) - Scout24 Ag: * BLOCK TRADE -...	IT Services & Consulting	Reuters
2018	Aug	11	Maryland has placed members of the football su...	Sports News	Reuters
2019	Apr	11	UBER SAYS CORE PLATFORM ADJUSTED NET REVENUE W...	Wires	CNBC
2018	Sep	28	NEW DELHI (Reuters) - India's burgeoning shado...	Business News	Reuters
2017	Feb	6	Feb 6 (Reuters) - China Child Care Corporation...	Chinese Labor Unrest	Reuters







MLE Stack (2 min)

- Exploratory Data Analysis & Wrangling
- Experimentation
- Data Engineering Pipeline
- Machine Learning Pipeline
- Deployment Pipeline

- *These ^^ are probably worth talking about*

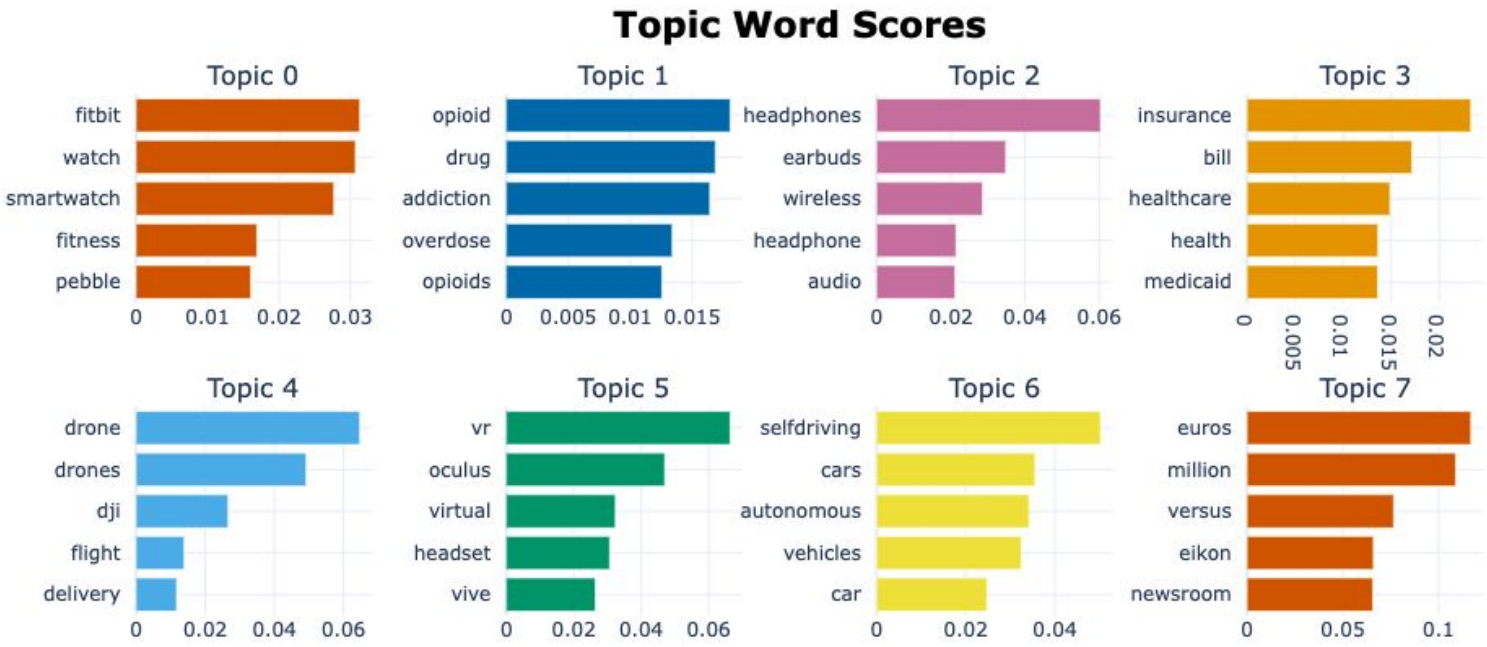
- Maybe consider...
  - Feature Store
  - Metadata store
  - Model registry
  - Model serving
  - Model Monitoring

Add your stack image here!

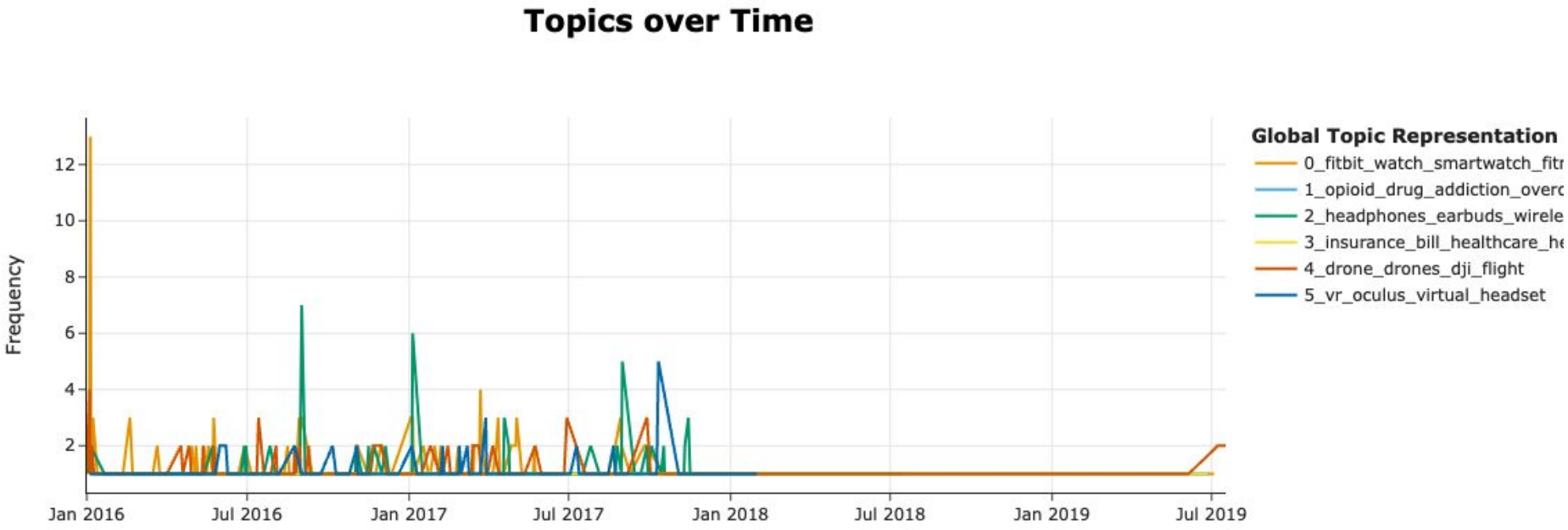
Modeling — *Supervised Learning*

- 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)

Modeling — *Unsupervised Learning*



Modeling — *Unsupervised Learning*



Data + Model (1 min)

- Any unique data wrangling/data-centric AI? If not, leave out. Lineage?
- Pre-trained model? If not, why not? Fine-tuned? If not, why not?
- Messy live data, modeling explainability aspects, ethical/responsible?
- Choose wisely; time is of the essence; four bullets always better than five.

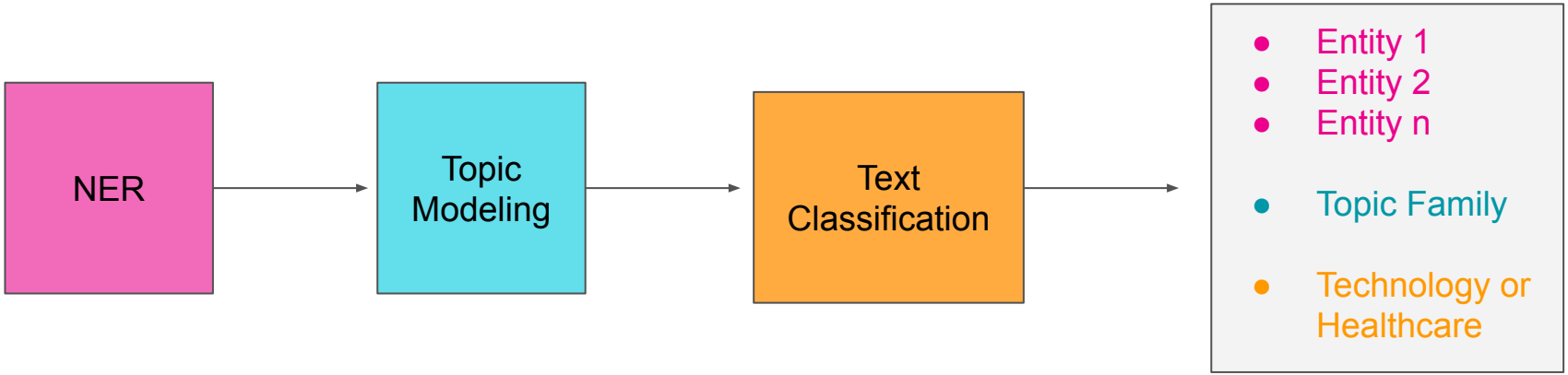
Demo (2 min)

... & screen share well!

Conclusions (90 s)

- We learned to do end-to-end ML the easy way, the hard way
- Let us tell you about it!
- Here’s a tip or two for anyone who tries to walk down a similar path!
- And the biggest lesson we’re taking with us into the future is …

Future Work - Potential Flow



Future Work

- Try Transformers to improve NER performance
- Further explore unsupervised methods to understand topics
- Web interface for users to interact with

Thank You! Questions?