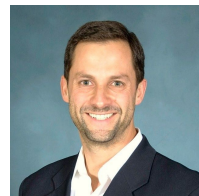
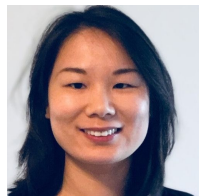


FourthBrain

GLG Project

**A match made in machine learning heaven:
*linking every client request to the best expert***

Ying Hu, Cody McCormack, Cris Fortes





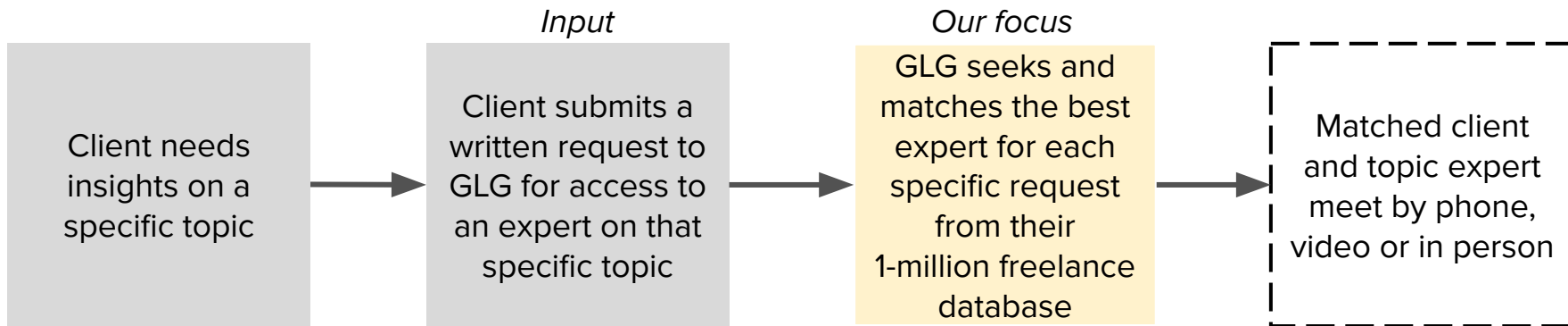
Outline

- Problem
- Solution
- Data + Model
- Demo
- MLE Stack
- Future Work
- Q&A and Feedback
- Appendix

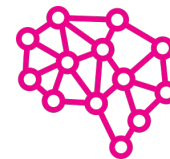


Problem

GLG's business largely revolves around **matching clients**, requesting insights on a specific topic, **with an expert** on that topic from their large database so that they can meet by phone, video or in person. Visually:

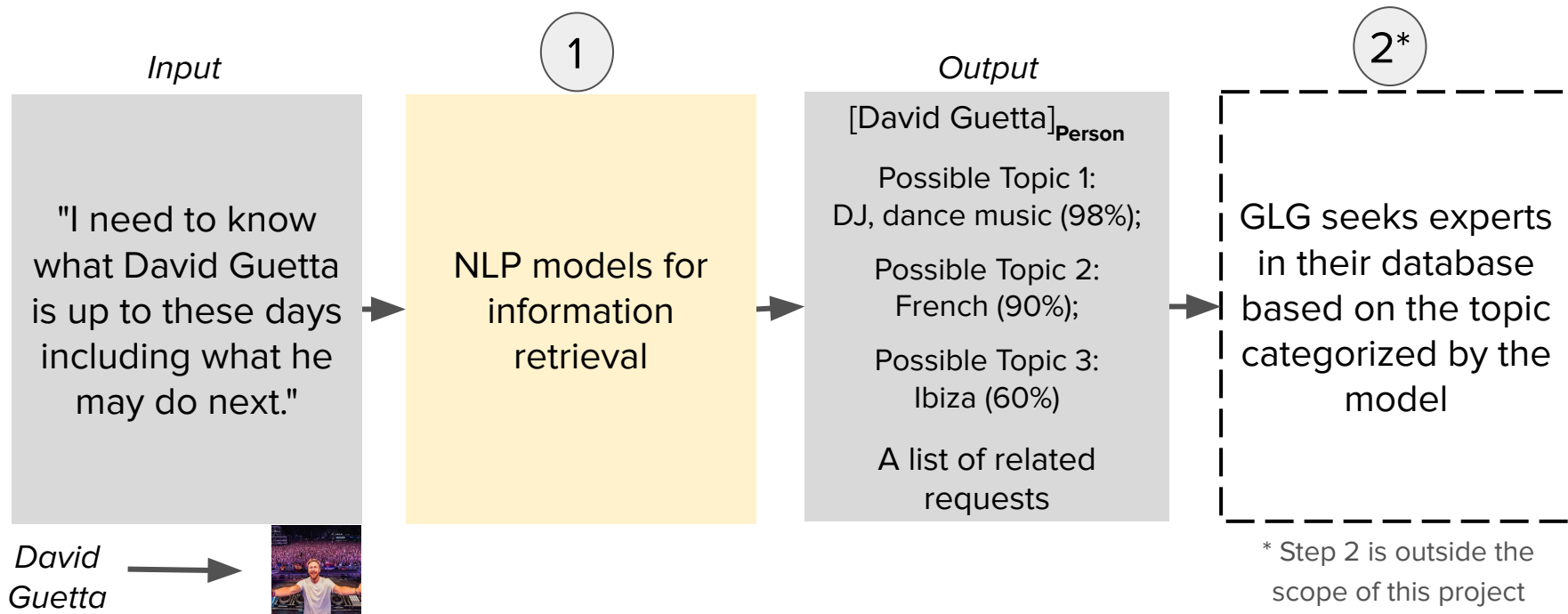


Since GLG receives **100s of these requests** per day, how can they leverage machine learning to **semi-automate the matching process at scale?**



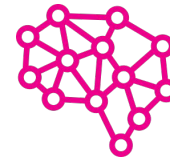
Solution

Natural Language Processing (NLP)!

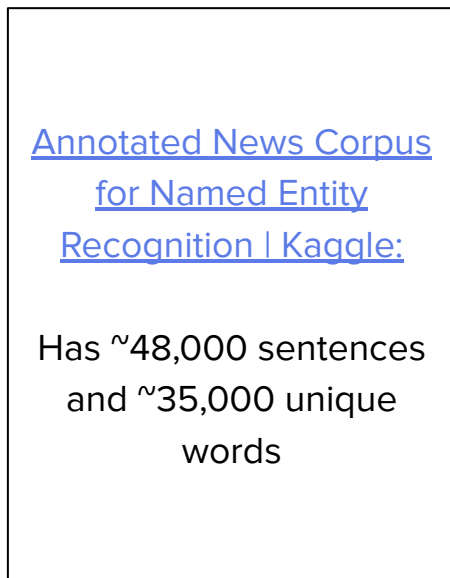


Acronyms: DJ (Disc Jockey), GLG (Gerson Lehrman Group)

Data + Model



Data



Natural Language Processing Models

Model 1, Named-Entity Recognition (NER):

- Trained our **NER** model
- Leveraged **spaCy** pre-trained **NER** model

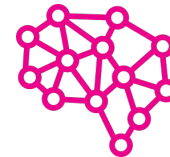
Model 2, Topic modeling:

- **Latent Dirichlet allocation (LDA)**

Model 3, Sentence matching:

- Used **SentenceTransformers** for text embedding
- Used **k-nearest neighbors (k-NN)** model to find nearby texts

Demo



[Demo URL](#)

Your Text:

As midterms near, Biden faces a nation as polarized as ever.

Named Entities Identified

['Biden']

Possibly Related Topics

52.39% of Topic 7: ['party', 'president', 'election', 'leader']
24.22% of Topic 14: ['charge', 'right', 'court', 'Iraq', 'house']
13.38% of Topic 4: ['attack', 'military', 'bomb', 'force']

Other Similar Requests


Biden is currently the chairman of the Senate Foreign Relations Committee and is a prominent critic of President Bush 's Iraq war strategy .
In a statement Friday , Mr. Biden 's office said the vice president will meet with the political leadership in all three countries , as well as U.S. officials and military personnel stationed in the region .
Biden has served in the U.S. Senate for more than 30 years .



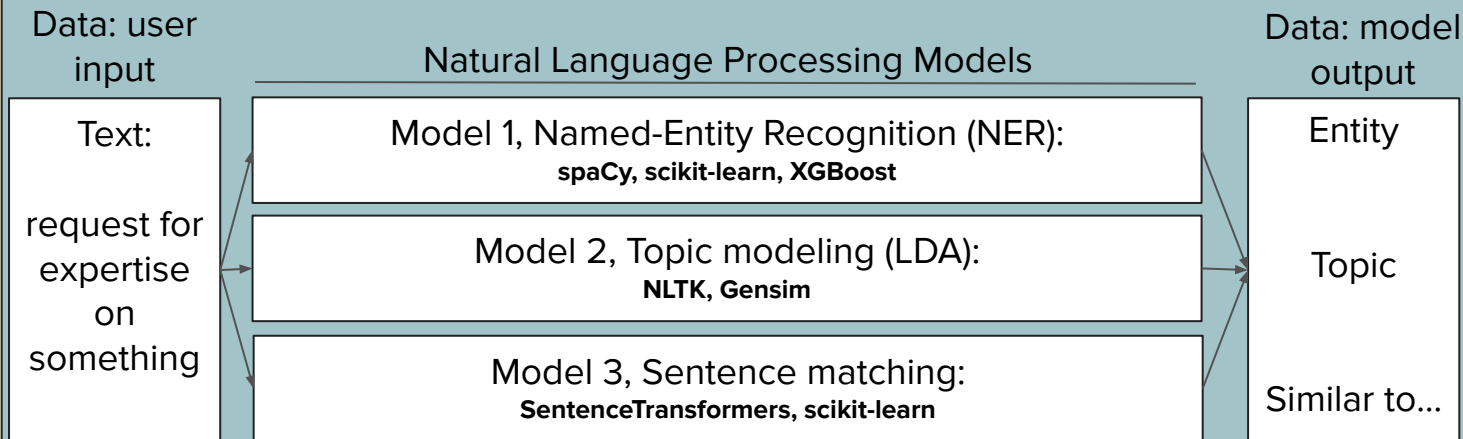
MLE Stack

Problem Solution Data + Model Demo **MLE Stack** Future Work Q&A Appendix

Cloud deployment: Amazon EC2 

Containerization: Docker 

Web framework: Flask 



Future Work

Problem	Solution	Data + Model	Demo	MLE Stack	Future Work	Q&A	Appendix
---------	----------	--------------	------	-----------	-------------	-----	----------



1. Improve the Topic Modeling:

- Training an LDA model on a **more diverse** [dataset](#)
- Using a **semi-supervised learning** method (SentenceTransformers + Label Propagation)

2. Expand the scope of the project:

- Building the expert(s) **recommendation model**
- Adapting our models to cover **non-English languages**
(GLG also has offices in Europe, Asia, and the Middle East)



Thank you

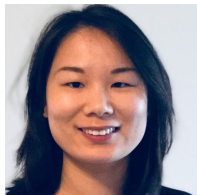


Q&A and Feedback

GLG Project

**A match made in machine learning heaven:
*linking every request to the best expert***

Ying Hu, Cody McCormack, Cris Fortes



[Problem](#)[Solution](#)[Data + Model](#)[Demo](#)[MLE Stack](#)[Future Work](#)[Q&A](#)[Appendix](#)

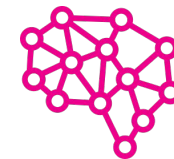
1 Named-Entity Recognition (NER) preliminary results

	Test 1: spaCY predictions	Test 2: TPOT for AutoML	Test 3: one-hot encoding		Test 4: TF-IDF encoding		Test 5: one-hot encoding with preprocessed data	
			XGB	Logistic Regression	XGB	Logistic Regression	XGB	Logistic Regression
Accuracy	0.937	Too computa tionally intense for local machine	0.959	0.932	0.935	0.921	0.959	0.932
Recall	0.619		0.906	0.761	0.881	0.612	0.906	0.761
Precision	0.753		0.755	0.659	0.644	0.638	0.758	0.659
F1 Score	0.680		0.824	0.706	0.744	0.625	0.825	0.706



2 Unsupervised clustering preliminary results

	Model 1: Bag of words + KMeans		Model 2: TF-IDF + KMeans		Model 3: Bag of words + PCA + KMeans		Model 4: Bag of words + PCA + Agglomerative
n_cluster	2	3	2	3	2	3	Aborted: It took too long to run; after 50 mins, the model was still running. The code is tested on a small portion of the dataset
Silhouette Coefficient	0.28	0.17	0.00814	0.000157	0.28	0.17	
random_states	1, 5, 10, 42	0, 1					
	Silhouette Coefficient decreases as number of cluster increases						



Did exploratory data analysis (EDA) on one dataset from Kaggle: [Annotated Corpus for Named Entity Recognition | Kaggle](#)

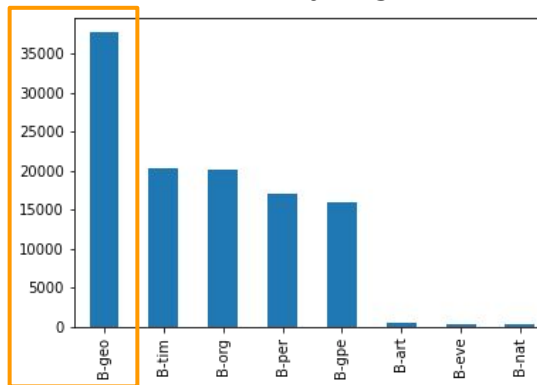
List of entity tags

- geo = Geographical Entity
- org = Organization
- per = Person
- gpe = Geopolitical Entity
- tim = Time indicator
- art = Artifact
- eve = Event
- nat = Natural Phenomenon

Example of entity tag

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

Plot of entity tag count



Capital vs. non-capital word count

