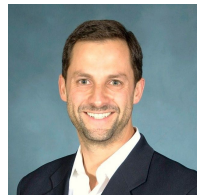
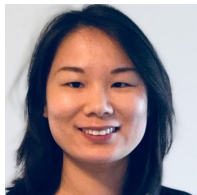


FourthBrain

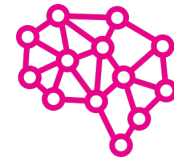
GLG Project

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

Ying Hu, Cody McCormack, Cris Fortes



Outline

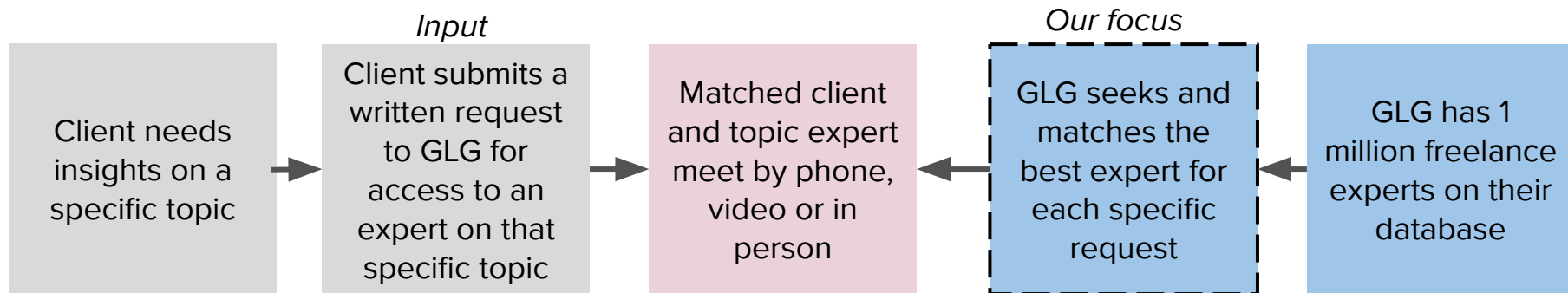


- Problem
- Solution
- Data + Model
- Demo
- MLE Stack
- Conclusions (and lessons learned)
- Future Work
- Q&A and Feedback
- Appendix

Problem

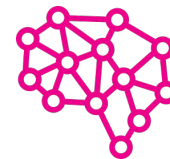


GLG's business largely revolves around **matching clients**, requesting insights on a specific topic, **with an expert** on that topic from their large database so that they can meet by phone, video or in person. Visually:

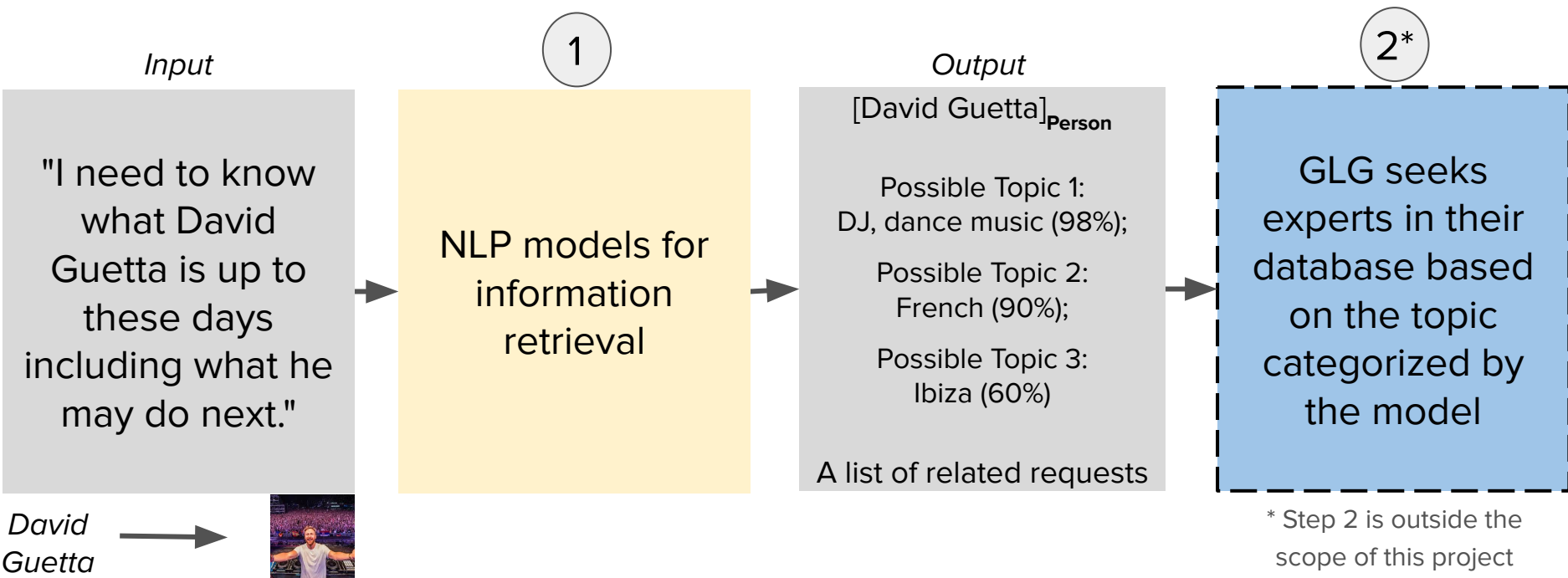


Since GLG receives **100s of these requests** per day, how can they leverage machine learning to **semi-automate the matching process at scale?**

Solution



Natural Language Processing (NLP)!



* Step 2 is outside the scope of this project

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

Data + Model



Data

[Annotated News Corpus
for Named Entity
Recognition | Kaggle:](#)

Has ~48,000 sentences
and ~35,000 unique
words

Natural Language Processing Models

Model 1, Named-Entity Recognition (NER):

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

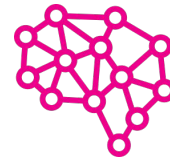
Model 2, Topic modeling:

- Latent Dirichlet allocation (LDA)

Model 3, Membership Classifier:

- Used SentenceTransformers for text embedding
- Used K-Nearest Neighborhood (KNN) model to find nearby texts

Demo



Placeholder for demo day URL: <http://54.221.36.219:8000/>

Replace it with
Flask web app screenshot
and/or recorded demo *after*
demo day

Enter Text to Get Matched!

Let's Connect You!

The sentence entered is:
David Guetta is a DJ, who sometimes plays for rich kids in Ibiza, Spain.

The entities from the text are:
['David Guetta', 'Ibiza', 'Spain']

The text may be related to the following topics:

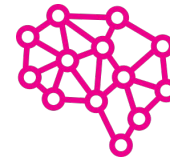
- 28.78% of Topic 2: ['north', 'south', 'korea', 'prime', 'minister']
- 16.42% of Topic 3: ['Beijing', 'Britain', 'France', 'gas', 'German', 'Middle', 'East', 'Russian']
- 15.67% of Topic 14: ['charge', 'right', 'court', 'Iraq', 'house']
- 14.94% of Topic 15: ['oil', 'company', 'market', 'demand', 'power', 'government']

The text is close to the following texts:


- Antonio Banderas was born in Spain and is an accomplished actor , writer , singer and producer .
- The concert will include a number of well-known Hispanic performers including Gloria Estefan , Marc Anthony , Jose Feliciano , George Lopez and Thalía .
- Musicians - particularly those from Mexico - have struck a cord with US audiences .


MLE Stack

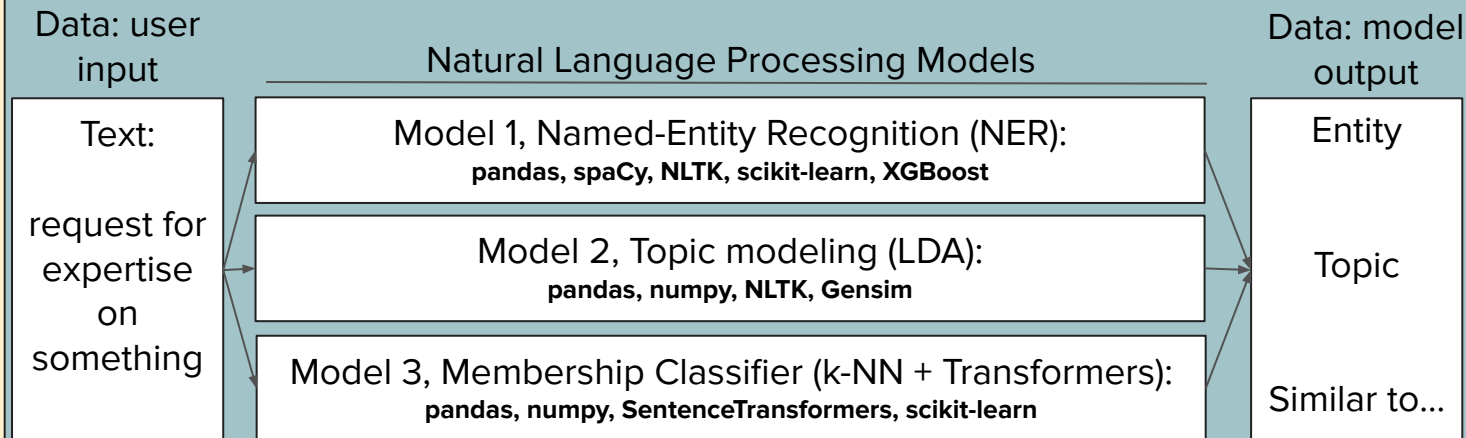
Problem Solution Data + Model Demo **MLE Stack** Conclusions Future Work Q&A Appendix



Cloud deployment: Amazon EC2 

Containerization: Docker 

Web framework: Flask 



Conclusions

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



- Natural Language Processing (NLP) models work!
- Any NLP model is only as good as the data it was trained on
- Quickly jumping into the web app (Flask), even before the NLP models were working properly, was the right thing to do (MVP mindset)
- Seeing a live, working, deployed model that addresses a real business problem is priceless

Future Work

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



- Training our NLP models on larger and more diverse datasets should yield better results. For example, using this other 2.7-million news articles dataset: [All the News 2.0 - Components](#)
- Adapting our models to cover non-English languages would come in handy (GLG also has offices in Europe, Asia, Japan and the Middle East)
- Building a GLG topic expert(s) recommendation model with input from our NLP models would be a natural next step for this project

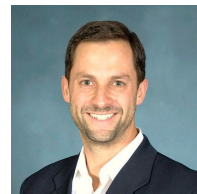
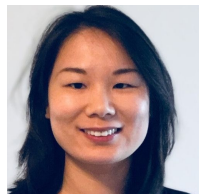


Q&A and Feedback

GLG Project

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

Ying Hu, Cody McCormack, Cris Fortes





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

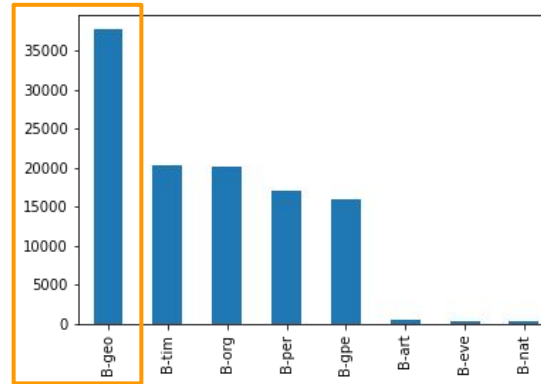
List of entity tags

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

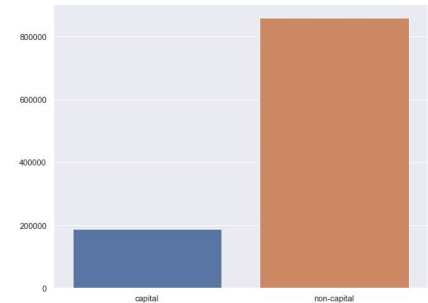
Example of entity tag

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





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		Test 6: TF-IDF encoding with preprocessed data	
			XGB	Logistic Regression	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	0.935	0.921
Recall	0.619		0.906	0.761	0.881	0.612	0.906	0.761	0.881	0.612
Precision	0.753		0.755	0.659	0.644	0.638	0.758	0.659	0.644	0.638
F1 Score	0.680		0.824	0.706	0.744	0.625	0.825	0.706	0.744	0.625



2 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	Model 5: Bag of words + LDA (to be tested further)
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	So far, with topic number = 10, the model seemingly outputs the most sensible list of topics
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 n_cluster increases							