TOPIC MODELING WITH TWITTER DATA

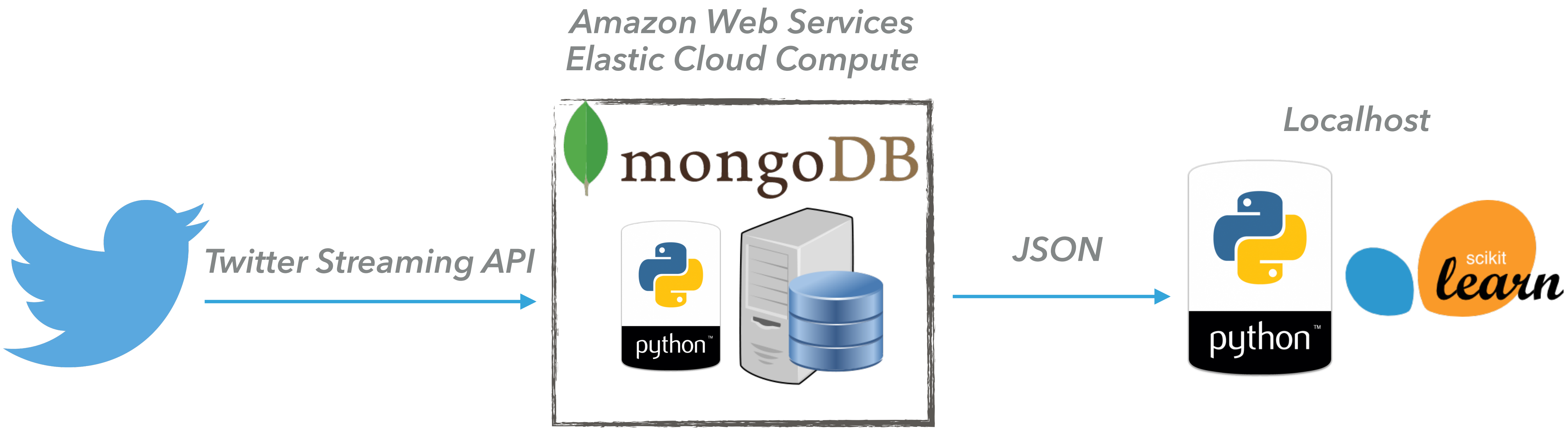
STATE OF THE (EU)NION

Jason Sypniewski
4 March 2016

PROBLEM STATEMENT

- A U.S. client, Fletcher Corp., is looking to expand operations into Europe. While the financial analysis looks promising, examine the geo-political situation using unsupervised learning and natural language processing (NLP).

ARCHITECTURE



Amazon Web Services
Elastic Cloud Compute

Twitter Streaming API

JSON

Localhost

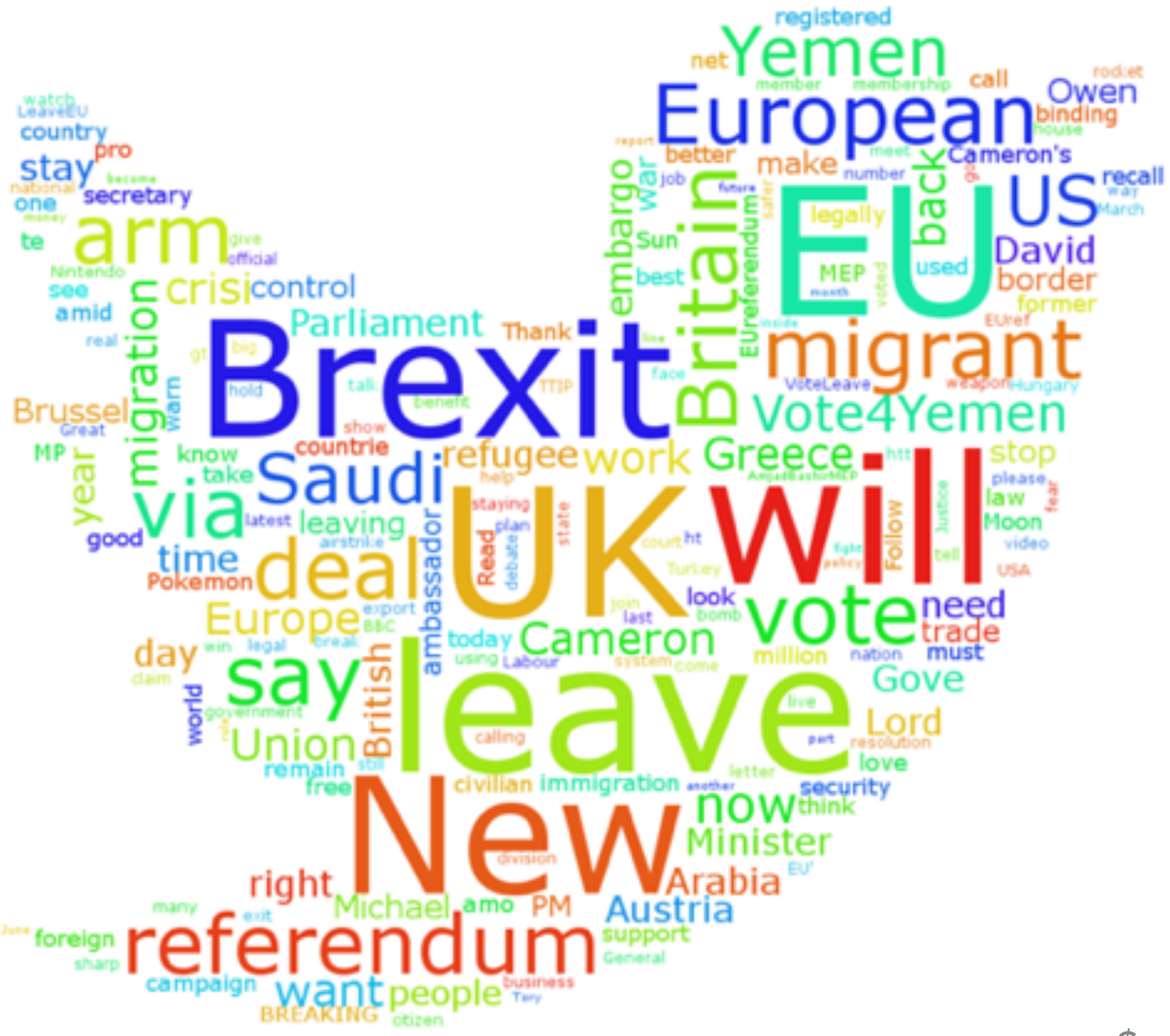
Search Terms = ‘European Union’ | ‘EU’

166,347 tweets over 36 hours (24 - 25 Feb 2016)

74 different countries (only 1,673 total tweets with location data)

20 unique hashtags

WHAT A LITTLE BIRD TOLD ME…



$ pip install WordCloud

K-MEANS CLUSTERING…AKA TWITTER NEWS FEED

Cluster 1	{ yemen, saudi, vote, arms, embargo, arabia, parliament, war, civilians, airstrikes }
Cluster 2	{ migrant, crisis, austria, ambassador, recalls, greece, amid, division, sharp, bbcbreaking }
*Cluster 3**	{ uk, referendum, brexit, leave, britain, cameron, deal, european, migration, says }
Cluster 4	{ stay, leave, vote, want, uk, like, britain, johnson, say }

* Most tweets

LDA TOPIC MODELING AND D3 VISUALIZATION

See HTML files

CONCLUSION: EUROPE…WE HAVE PROBLEMS

- With limited prior knowledge on a subject, aggregate Twitter data can be analyzed to extract relevant topics.
- Keep it clean: Twitter data is "noisy" and requires significant preprocessing.
- Technical challenges:
    - Task queueing and fault tolerance is critical in distributed environments.