TRANSCRIPT ANALYTICS

A Supervised Approach

100% → HEY SIRI

Speaker 0: Are you a fan of Google or Microsoft?
Speaker 1: Both are excellent technology they are helpful in many ways. For the security purpose both are super.
Speaker 0: I'm not a huge fan of Google, but I use it a lot because I have to. I think they are a monopoly in some sense.
Speaker 1: Google provides online related services and products, which includes online ads, search engine and cloud computing.

DATA PREPARATION

—Getting data in standardized format which the company prefers

JSON FORMAT

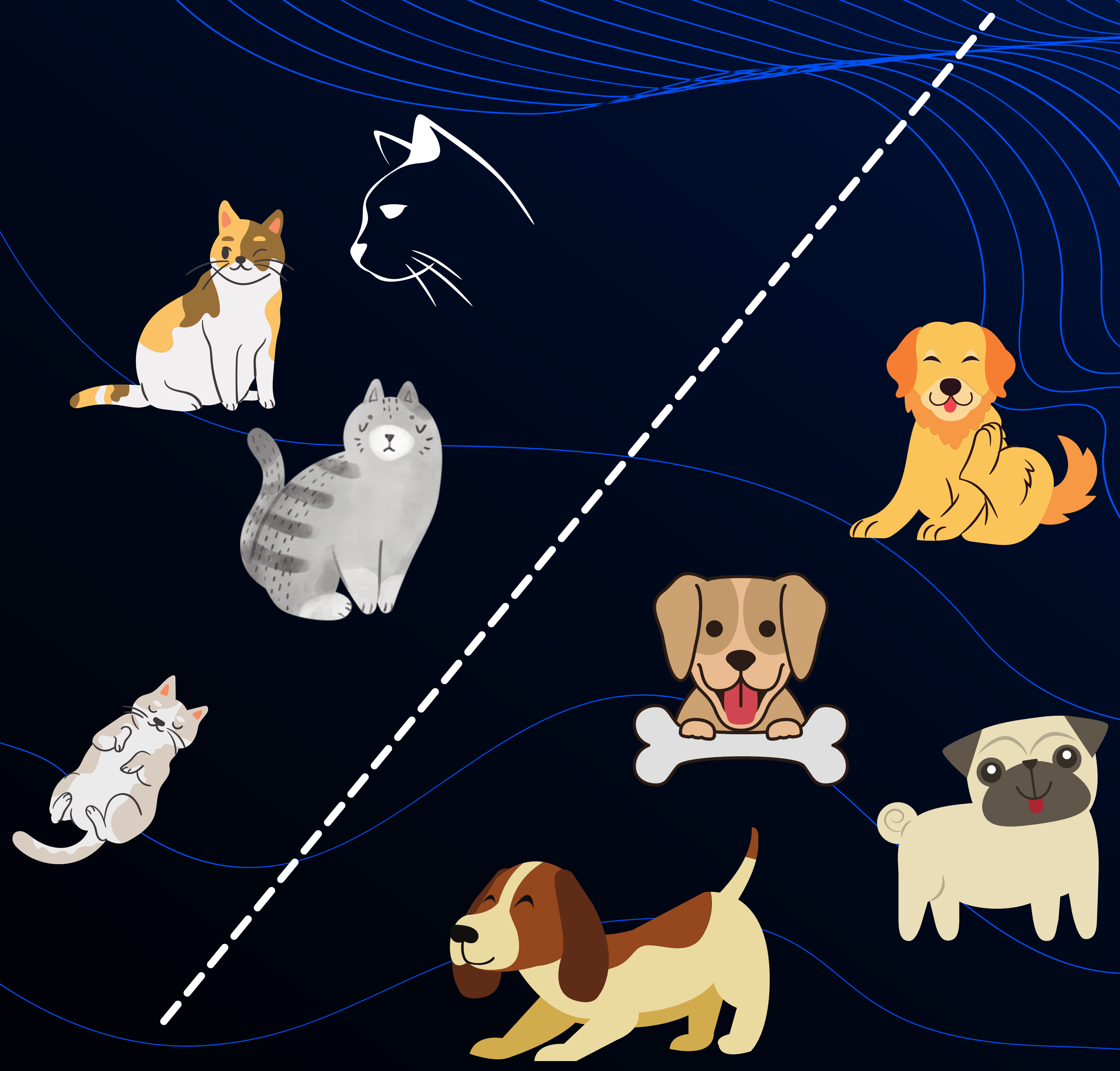
```
{
  "message": "Are you a fan of Google or Microsoft?",
  "agent": "agent_1",
  "sentiment": "Curious to dive deeper",
  "knowledge_source": [
    "FS1"
  ],
  "turn_rating": "Good"
},
```

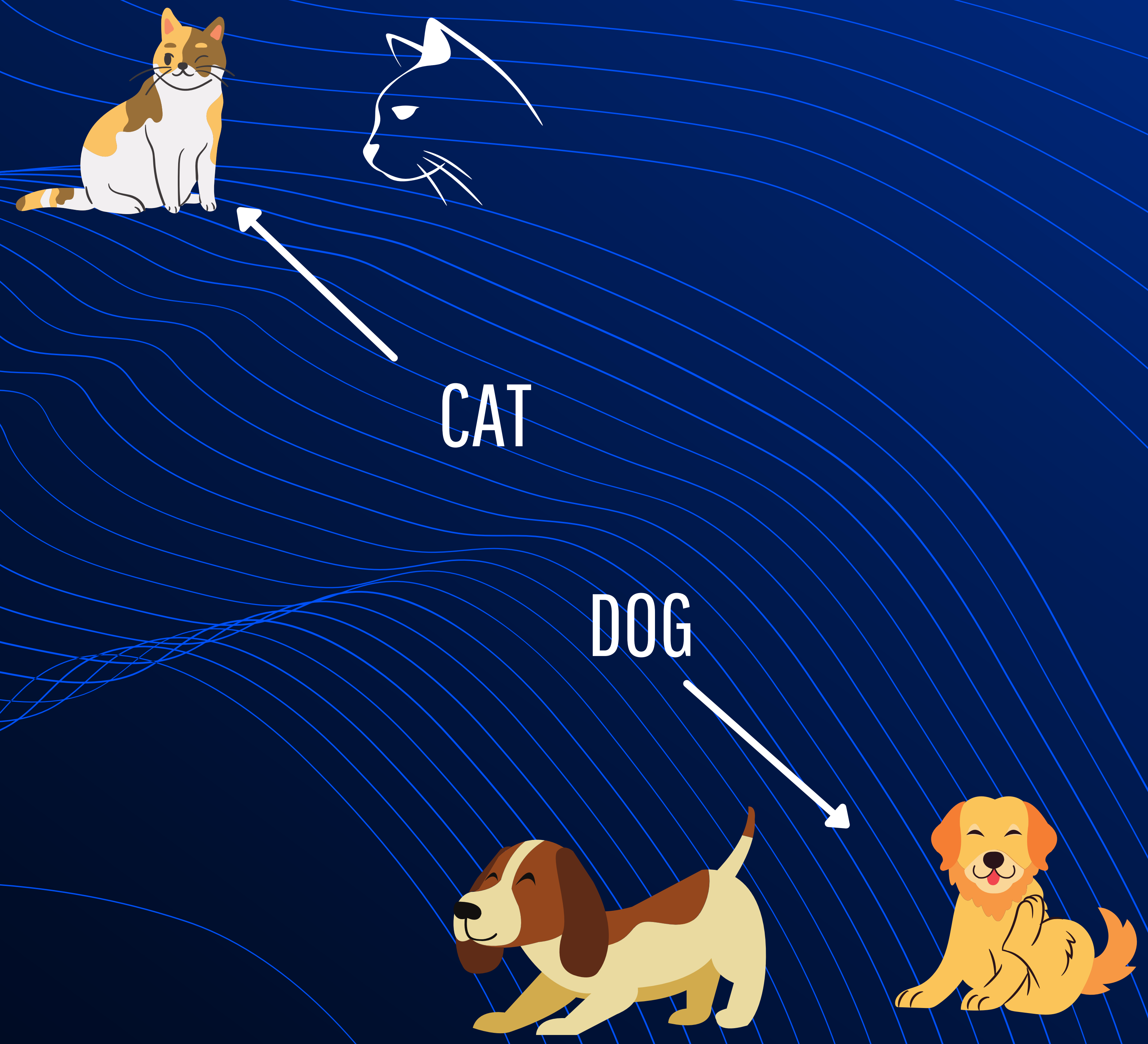
→

SPEAKER FORMAT

Speaker 0: Are you a fan of Google or Microsoft?
Speaker 1: Both are excellent technology they are helpful in many ways. For the security purpose both are super.
Speaker 0: I'm not a huge fan of Google, but I use it a lot because I have to. I think they are a monopoly in some sense.
Speaker 1: Google provides online related services and products, which includes online ads, search engine and cloud computing.

UN/SUPERVISED





TOPIC MODELING

—*S*tatistical model for discovering the abstract "topics" that occur in a collection of documents.

ISSUE

If only it was straightforward

NO SUPERVISED WAY EXISTS

- Found theorotical ways of approaching such problems
- Built a semi supervised way for topic modeling

CONSISTENCY AND UNIFORMITY

- Tagging every sensible word with the same consistency would have been difficult considering it was SIRI's data
- Different word & tags combination for each one

MANUALLY IMPOSSIBLE

- Taging 1.8 Mn conversations would have taken over 2 months single handedly
- Why not automate it?

SPACY - A NLP LIBRARY

- Used predefined library to generate tags for 1.8Mn conversations
- 1.8Mn conversations ----> 18 Tags
- 13,951 unique words ---> 18 Tags
- Avg length ---> 775 words

DICTIONARY SNAPSHOT

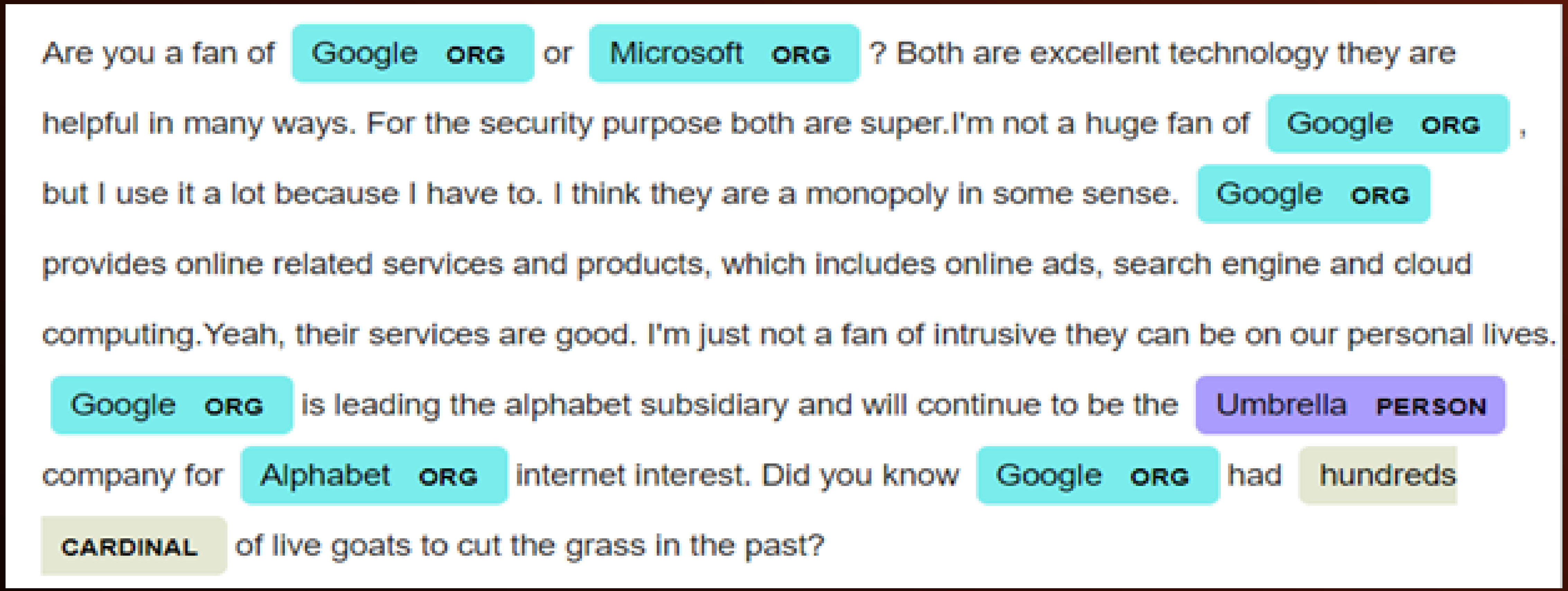
'CARDINAL':'100','4','One','four','only one','only 1','1','two'
'DATE':'2015','a good day','1958','a week','1998','daily','every day',
'MONEY':'$1.09 billion','84 per cent','billion dollar','over $8.5 billion dollars',

DEFINE A DICTIONARY

—a label attached to something for the purpose of identification or to give other information tag

ADDING TAGS TO TRAIN AND VALIDATION DATA

- Final data preparation phase
- Applying the tags generated to train and validation data sets



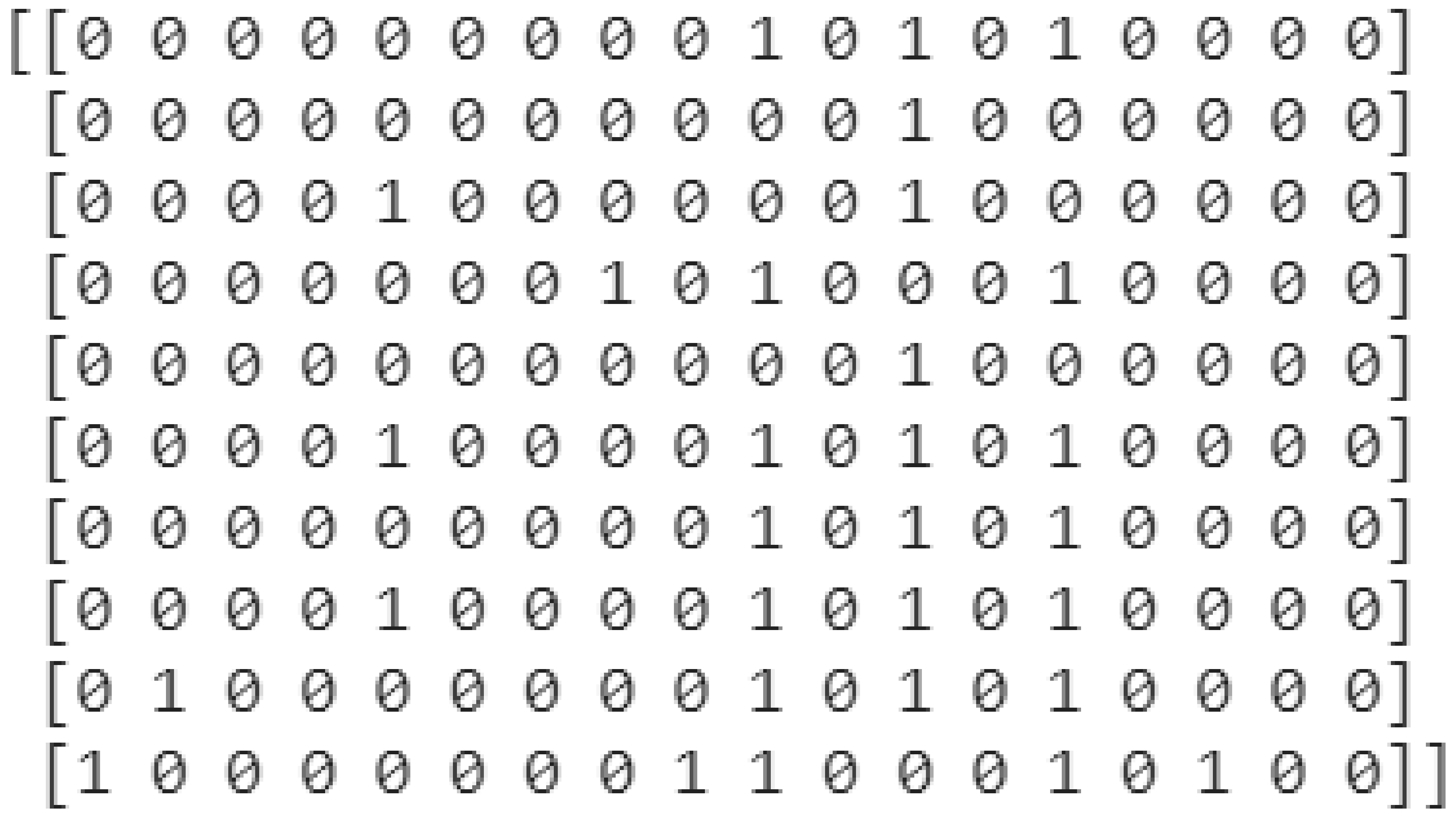
Are you a fan of Google ORG or Microsoft ORG ? Both are excellent technology they are helpful in many ways. For the security purpose both are super.I'm not a huge fan of Google ORG , but I use it a lot because I have to. I think they are a monopoly in some sense. Google ORG provides online related services and products, which includes online ads, search engine and cloud computing.Yeah, their services are good. I'm just not a fan of intrusive they can be on our personal lives. Google ORG is leading the alphabet subsidiary and will continue to be the Umbrella PERSON company for Alphabet ORG internet interest. Did you know Google ORG had hundreds CARDINAL of live goats to cut the grass in the past?

OUTPUT - MULTI LABELED DATA

—Multi-label data has zero or more class labels making it difficult as the size of output is undefined

GOAL : BINARIZE

[('PERSON', 'ORG', 'NORP'),
('ORG',),
('GPE', 'ORG'),
('PERSON', 'LOC', 'NORP'),
('ORG',),
('GPE', 'PERSON', 'ORG', 'NORP'),
('PERSON', 'ORG', 'NORP'),
('GPE', 'PERSON', 'ORG', 'NORP'),
('PERSON', 'ORG', 'DATE',
'NORP')]



```
[[0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0]
 [0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0]
 [0 0 0 0 0 0 0 1 0 1 0 0 0 1 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0]
 [0 0 0 0 1 0 0 0 0 1 0 1 0 1 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0]
 [0 0 0 0 1 0 0 0 0 1 0 1 0 1 0 0 0 0]
 [0 1 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0]
 [1 0 0 0 0 0 0 0 1 1 0 0 0 1 0 1 0 0]]
```

OVR → CREATE ONE VS REST PIPELINE

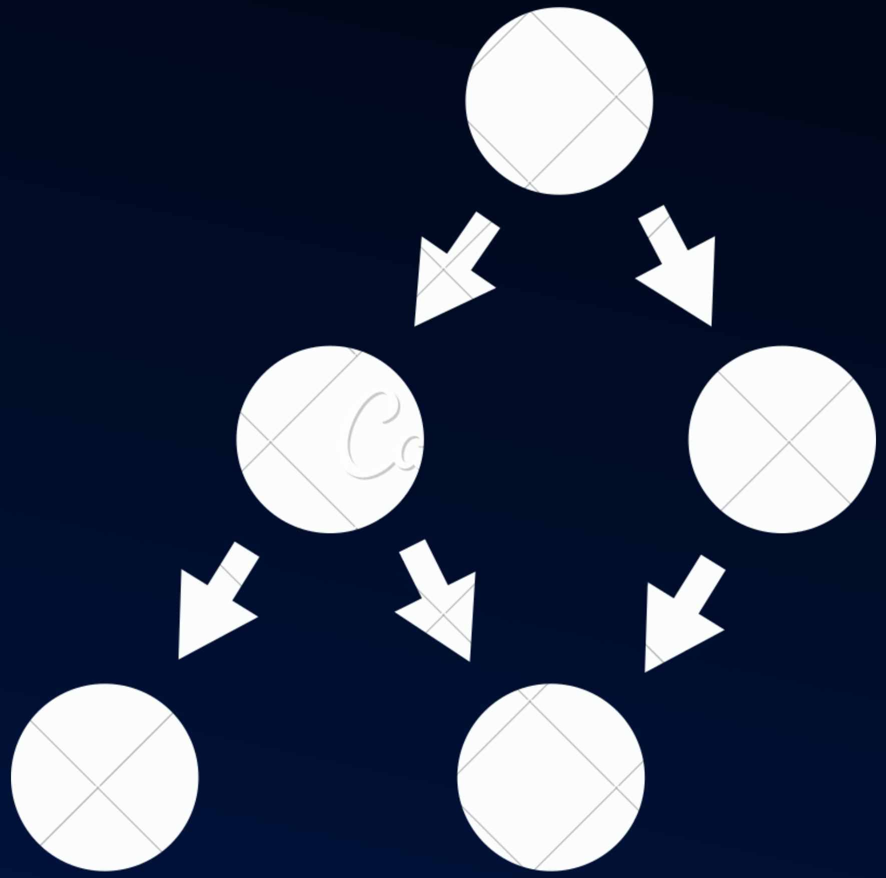
- Method for using binary classification algorithms for multi-class classification.

MODEL DEVELOPMENT

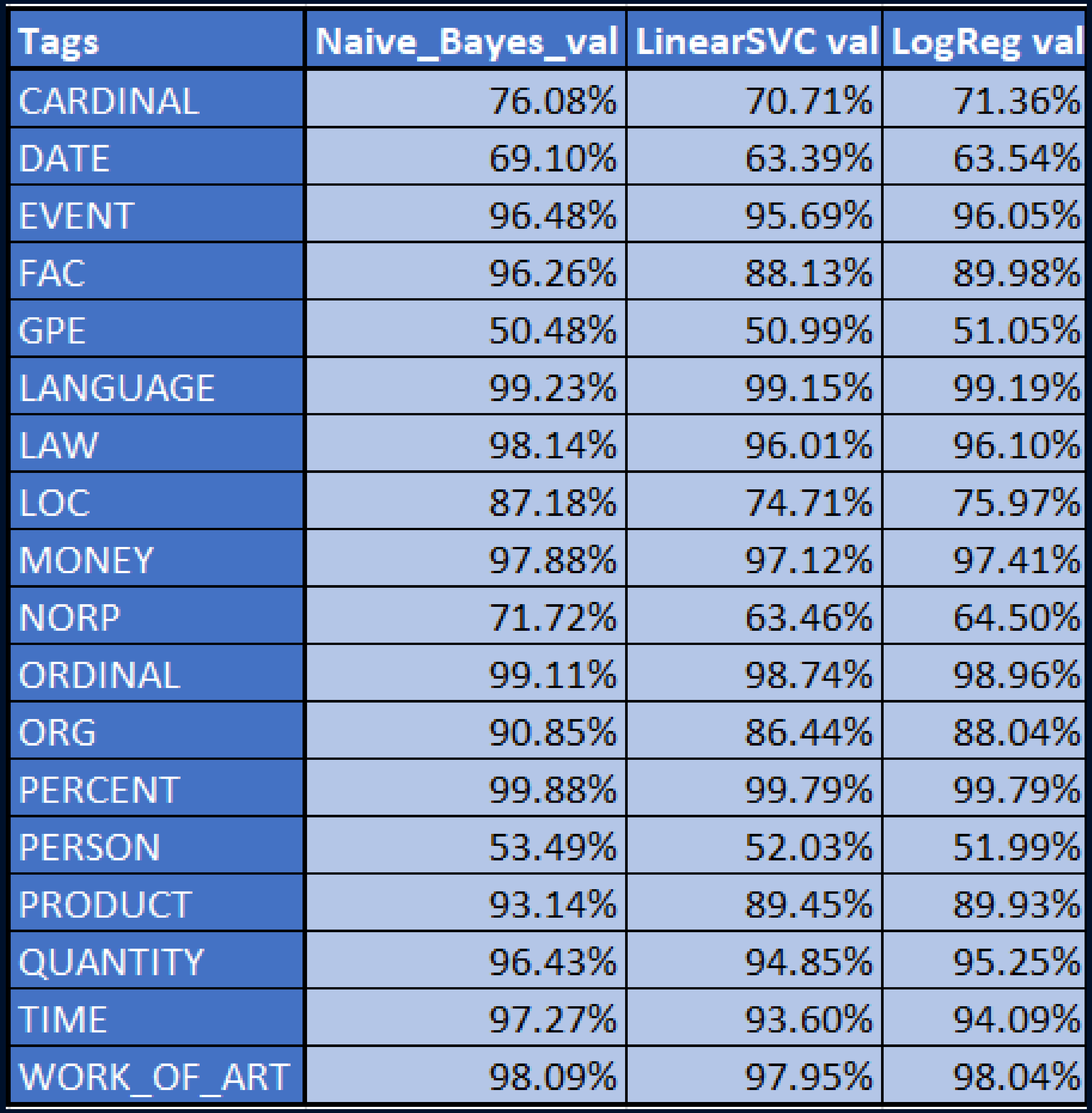
— Applied Naive Bayes, Linear SVC, Logistic Regression

ALGO'S

- Naive Bayes - classifies based on probabilities of events
- Linear SVC - performs well but needs equally distributed classes
- Logistic Regression - Using As we are opting for OVR



MODEL COMPARISION



Tags	Naive_Bayes_val	LinearSVC val	LogReg val
CARDINAL	76.08%	70.71%	71.36%
DATE	69.10%	63.39%	63.54%
EVENT	96.48%	95.69%	96.05%
FAC	96.26%	88.13%	89.98%
GPE	50.48%	50.99%	51.05%
LANGUAGE	99.23%	99.15%	99.19%
LAW	98.14%	96.01%	96.10%
LOC	87.18%	74.71%	75.97%
MONEY	97.88%	97.12%	97.41%
NORP	71.72%	63.46%	64.50%
ORDINAL	99.11%	98.74%	98.96%
ORG	90.85%	86.44%	88.04%
PERCENT	99.88%	99.79%	99.79%
PERSON	53.49%	52.03%	51.99%
PRODUCT	93.14%	89.45%	89.93%
QUANTITY	96.43%	94.85%	95.25%
TIME	97.27%	93.60%	94.09%
WORK_OF_ART	98.09%	97.95%	98.04%

ANY QUESTIONS?

Thank you!

Srishti Patil
RBA 009