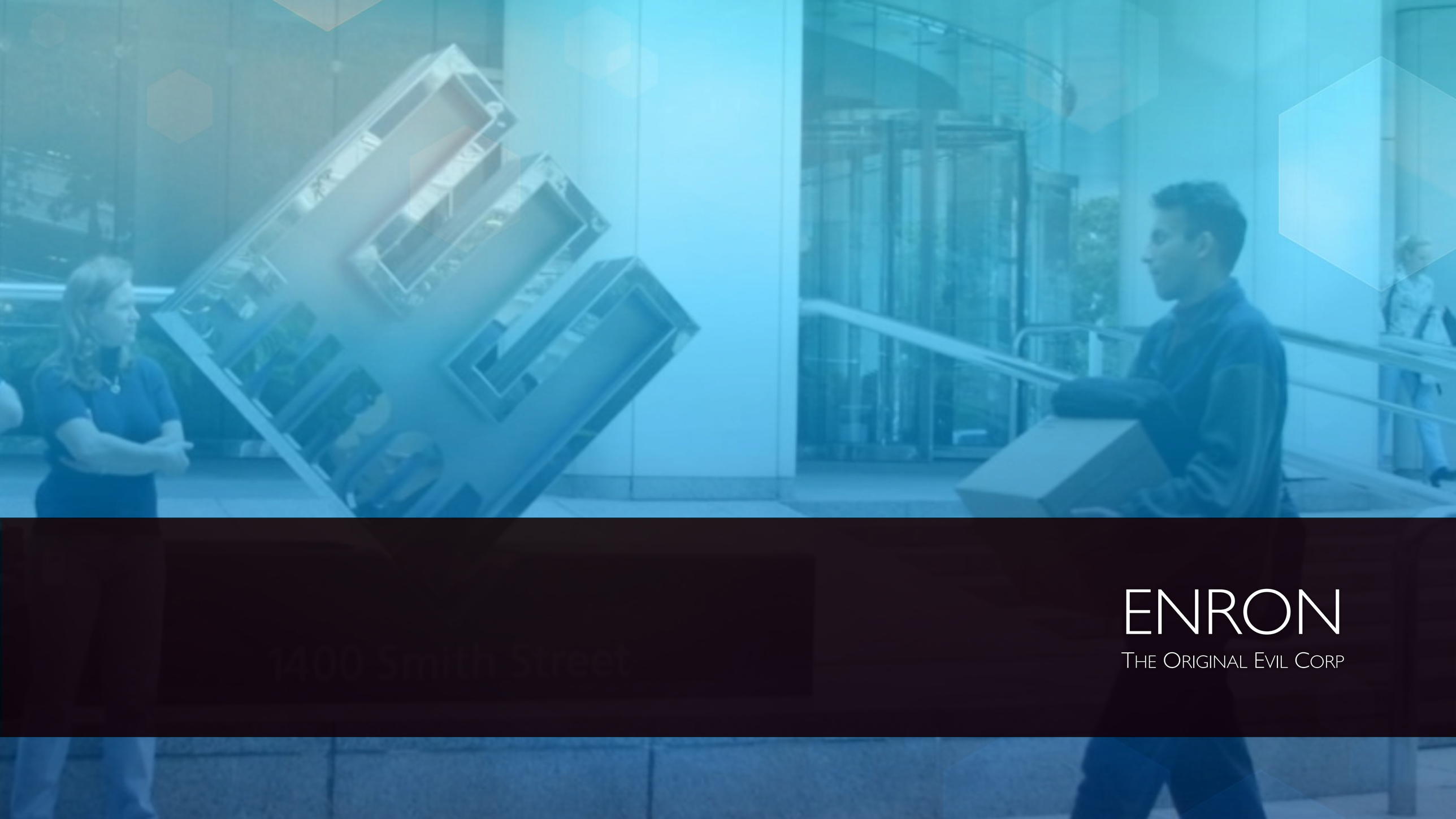
# Enron Email Corpus

## A K-Means Cluster Analysis



# ENRON

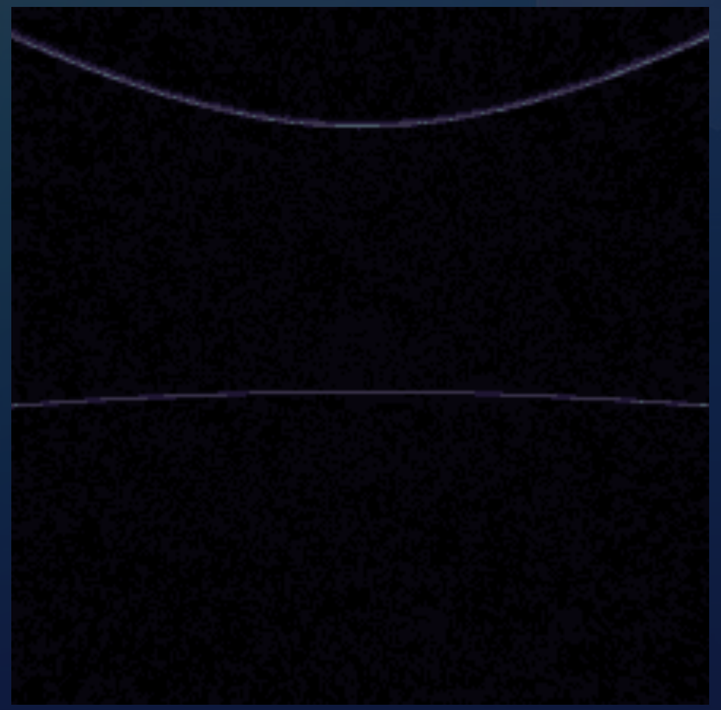
THE ORIGINAL EVIL CORP

# Back Story

- Houston based company founded in 1985
- Before bankruptcy on Dec 3, 2001
  - Had 29,000 employees
  - Major electric, natural gas, communications & pulp & paper co.
  - Revenue of 101 Billion in 2000
  - named "America's Most Innovative Co." 6 consecutive years - Fortune.

# Evil Corp.

- They cooked their books
- Which caused the Enron Scandal
- Sarbanes-Oxley Act of 2002
- Killed Arthur Anderson
- Intentionally encouraged removal of power during CA energy crisis
- But most importantly…



# The People

- 144 Senior Managers
- 29,000 Staff
- Thousands lost their 401K's and Pensions
- Your life's work

jupyter enron_email_analysis-all-countvectorize Last Checkpoint: a day ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted

Python 3

Run

Markdown

In [2]:
```
df = pd.read_csv('./data/enron_email_20_percent.csv')
```

In [3]:
```
df.shape
```

Out[3]: (103480, 52)

## Cleaning Data

In [193]:
```
# show 55 columns and rows to see my column data

pd.set_option('display.max_columns', 55)
pd.set_option('display.max_rows', 55)
```

In [5]:
```
df.head(1)
```

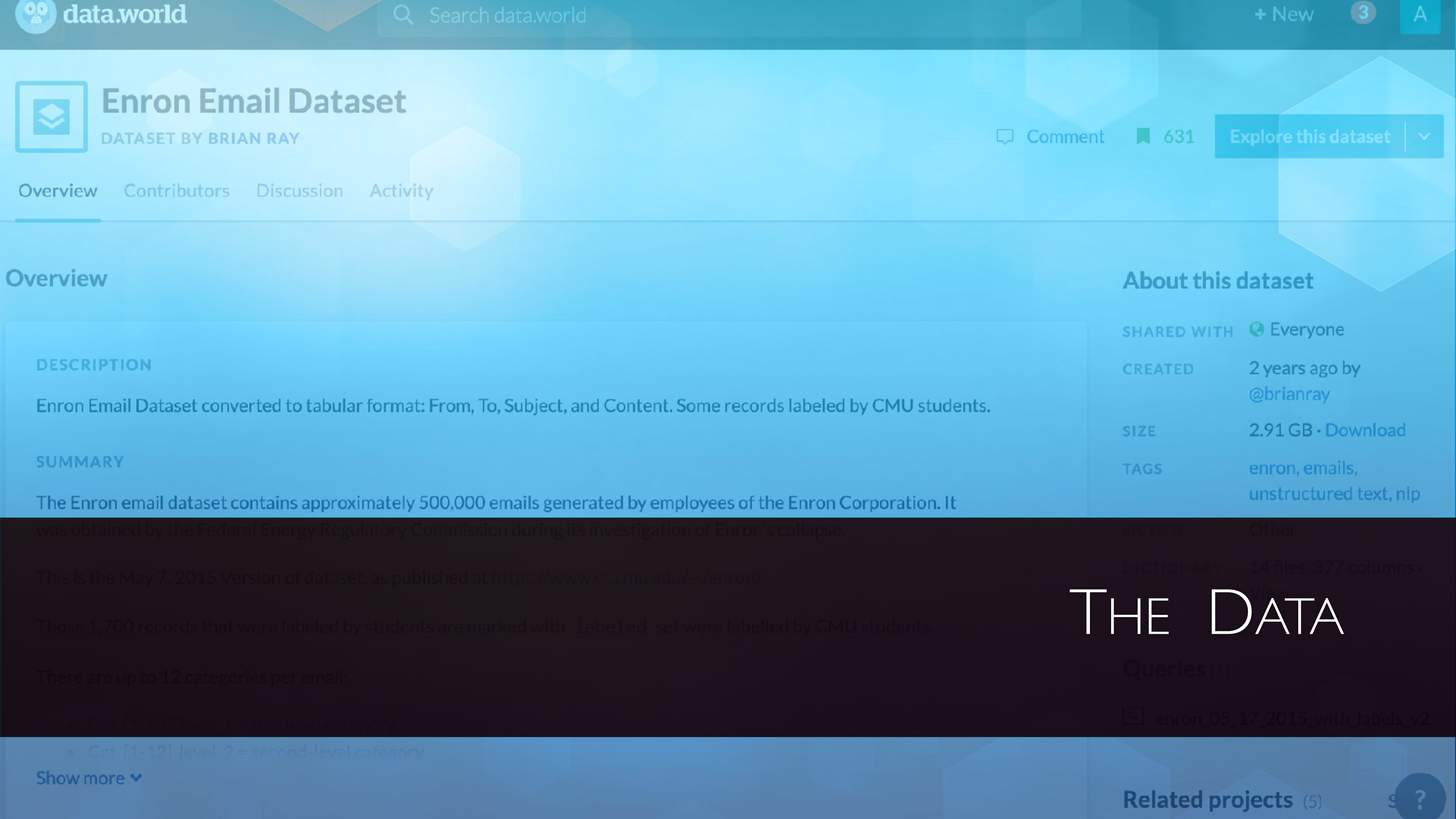
Out[5]:

| | Unnamed: 0 | Message-ID | Date | From |
|---|---|---|---|---|
| 0 | 403470 | <117238.1075846810566.JavaMail.evans@thyme> | 2000-12-11 09:22:00 | frozenset({'susan.scott@enron.com'}) | frozenset({'alicia.perkins@ |

# The Project

# Enron Email Analysis

- Data - Enron Emails
- Culpability vs Probable Cause vs Save Time ???
- Researched & Acquired Our Data
- Cleaned the Data
- Explored the Data
- Modeled Data Through K-Means Cluster
- Present Data


data.world
Search data.world
+ New
3
A
Enron Email Dataset
DATASET BY BRIAN RAY
Comment
631
Explore this dataset
Overview
Contributors
Discussion
Activity
Overview
DESCRIPTION
Enron Email Dataset converted to tabular format: From, To, Subject, and Content. Some records labeled by CMU students.
SUMMARY
The Enron email dataset contains approximately 500,000 emails generated by employees of the Enron Corporation. It was obtained by the Federal Energy Regulatory Commission during its investigation of Enron's collapse.
This is the May 7, 2015 Version of dataset, as published at https://www.cs.cmu.edu/~./enron/
Those 1,700 records that were labeled by students are marked with 'labeled' set were labelled by CMU students.
There are up to 12 categories per email:
Show more
About this dataset
SHARED WITH
Everyone
CREATED
2 years ago by @brianray
SIZE
2.91 GB · Download
TAGS
enron, emails, unstructured text, nlp
LICENSE
Other
DICTIONARY
14 files, 377 columns
Queries
enron_05_17_2015_with_labels_v2
Related projects (5)

# The Data

# Enron Email Corpus

- Originally 1.7 MM emails
  - Most have been removed because of privacy issues
- 500,000 remain publicly available
- Several Kaggle projects on just cleaning the data
- I used the data set by Brian Ray found at data.world
- Removed unnecessary NaN's, duplicates, Also misc. columns

jupyter enron_email_analysis-all-countvectorize Last Checkpoint: a day ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3

Run Markdown

## Countvectorize

In [156]:
```
stopwords = ENGLISH_STOP_WORDS
my_stopwords = ENGLISH_STOP_WORDS.union(['ect', 'hou', 'com', 'recipient', 'sent', 'enron',
                                         'forwarded', 'corp', 'said', 'attached', 'ect',\
                                         'recipient', 'email', 'original', 'doc', 'pm', 'ma
                                         'enronxgate', 'na', 'year'])


# tf = TfidfVectorizer(analyzer='word', stop_words=my_stopwords, min_df=2, max_df=0.5)
cv = CountVectorizer(analyzer='word', stop_words=my_stopwords, max_features=2000)

# vectorizer = TfidfVectorizer(stop_words='english')
X = cv.fit_transform(df['content']) # this is my sparce matrix
```

# Data Modeling

In [157]:
```
# Additional EDA on number only data that I don't believe is mission critical
pd.set_option('display.max_columns', 500)
```

In [158]:
```
# New DataFrame created to remove vectorized numbers
X_df = pd.DataFrame(X.toarray(), columns=cv.get_feature_names()) # this is my dense matrix
```

In [159]:
```
# Looking at shape of new dataframe
```

# Enron Email Analysis - Modeling

- Term Frequency-Inverse Document Frequency was initially used
- I ended up using CountVectorizor
  - Stopwords
    Mystopwords
  - CountVectorizer
    max_features = 2000

# Enron Email Analysis - Modeling

- K-Means Cluster Algorithm
- Clustered groups of words more closely related
  - K = 5 Clusters

# Enron Email Analysis - Initial Results

- Whole Corpus Frequently Occurring Words / Number of Occurrences

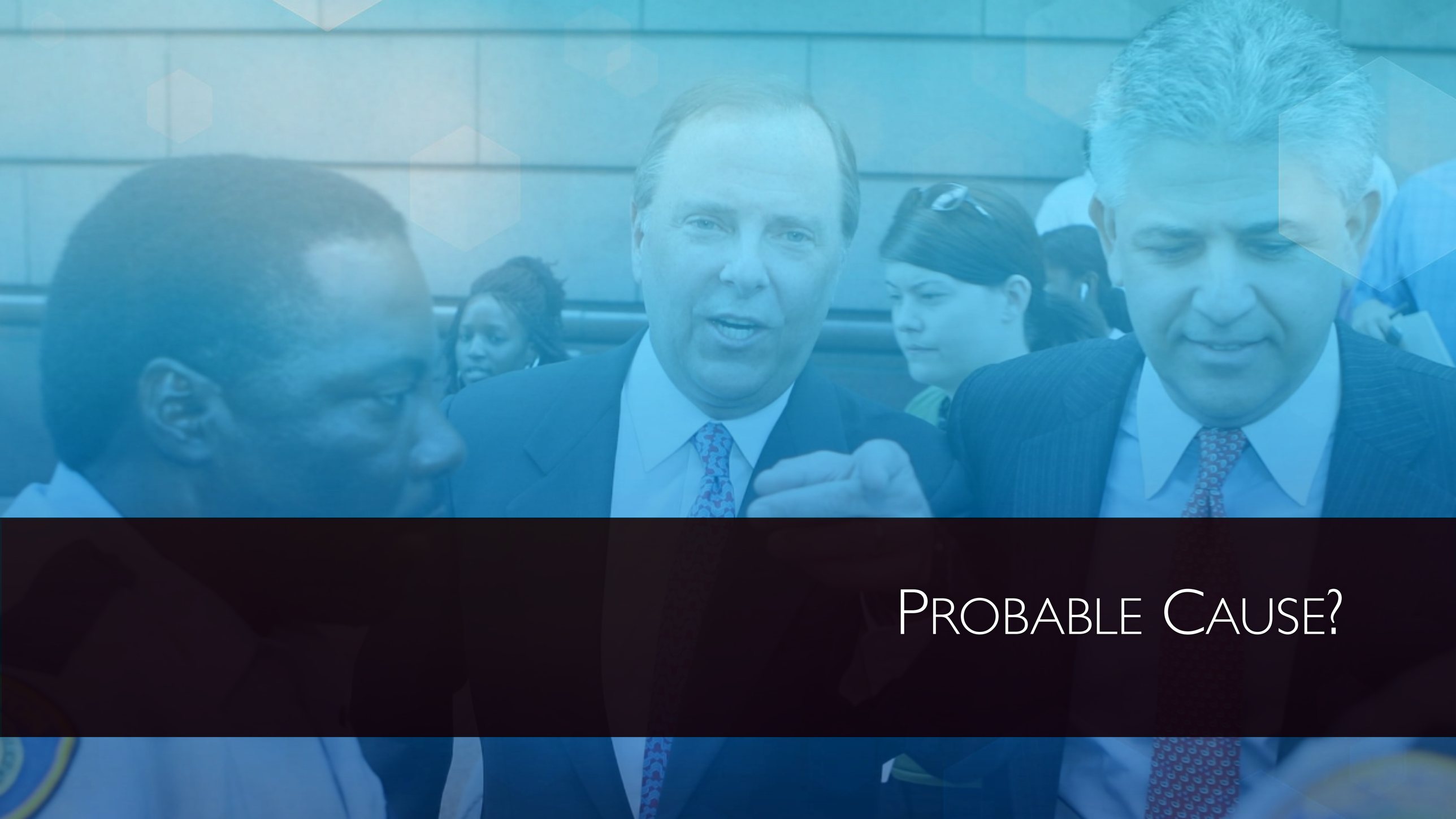
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| power | 61521 | ees | 35715 | just | 25315 | make | 20947 | services | 18356 |
| energy | 59937 | california | 31829 | houston | 24148 | question | 20779 | think | 18186 |
| new | 55393 | need | 31753 | john | 24042 | use | 20611 | work | 18158 |
| time | 43476 | state | 30168 | mark | 23802 | today | 20337 | million | 17956 |
| compan | 42883 | business | 29407 | agreem | 23129 | jeff | 20170 | report | 17948 |
| gas | 42459 | like | 28488 | meeting | 22331 | td | 19901 | monday | 17805 |
| thanks | 40003 | day | 28418 | image | 22274 | contact | 19495 | service | 17636 |
| know | 39739 | let | 27730 | deal | 21771 | don | 19424 | credit | 17589 |
| market | 37091 | week | 27554 | group | 21493 | date | 19286 | electricit | 17371 |
| informat | 36272 | price | 26530 | trading | 21250 | font | 19209 | friday | 17235 |

# Enron Email Analysis - Modeling

- K-Means Cluster Algorithm
- Clustered groups of words more closely related
  - K = 5 Clusters

# Enron Email Analysis - Clusters

| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|---|---|---|---|---|
| Businesses | sources | manager | meter | option |
| option | engine | stake | bpa | season |
| Defense | stories | station | impacted | defense |
| manager | role | including | quantity | attend |
| linda | applicationes | defense | heat | debbie |
| evening | 30 | option | software | commodities |
| known | body | day | assistance | evening |
| assignment | trades | gov | mkt_type | team |
| attend | quickly | evening | ancillary | pager |
| analysis | 75 | main | appear | page |
| season | doug | known | option | notice |
| crude | south | investigation | members | analysis |
| hot | follow | involved | shapiro | dave |
| sellers | site | commodity | kenneth | red |
| business | annual | businesses | industry | manager |
| efforts | computer | tools | sense | known |
| debbie | person | holding | holding | clean |
| students | light | assignment | read | teams |
| charge | break | hub | judy | brown |
| dale | alternative | knowledge | dear | office |



# Probable Cause?

# Enron Email Analysis - Cluster 3

| | |
|---|---|
| manager | known |
| stake | investigation |
| station | involved |
| including | commodity |
| defense | businesses |
| option | tools |
| day | holding |
| gov | assignment |
| evening | hub |
| main | knowledge |

# Practical Applications

# Practical Applications

- 1.7 Million Emails
- Avg email read rate - 20 seconds
- 34,000,000 seconds (that's 34 million seconds)
- 393 - 24 hour days to read through 1.7 MM emails
- 1,180 - 8 hour work days (3.23 years)

# Practical Applications

- Sort through a cluster of emails
- Less time consuming
- Potentially more productive
- For Example…

# Consultation Gig

- Pull out my trusty Titanium PowerBook G4
  - Circa 2001
  - 128 MB Ram
  - 128 MB Graphics Card
  - 10 Gigs of disk space

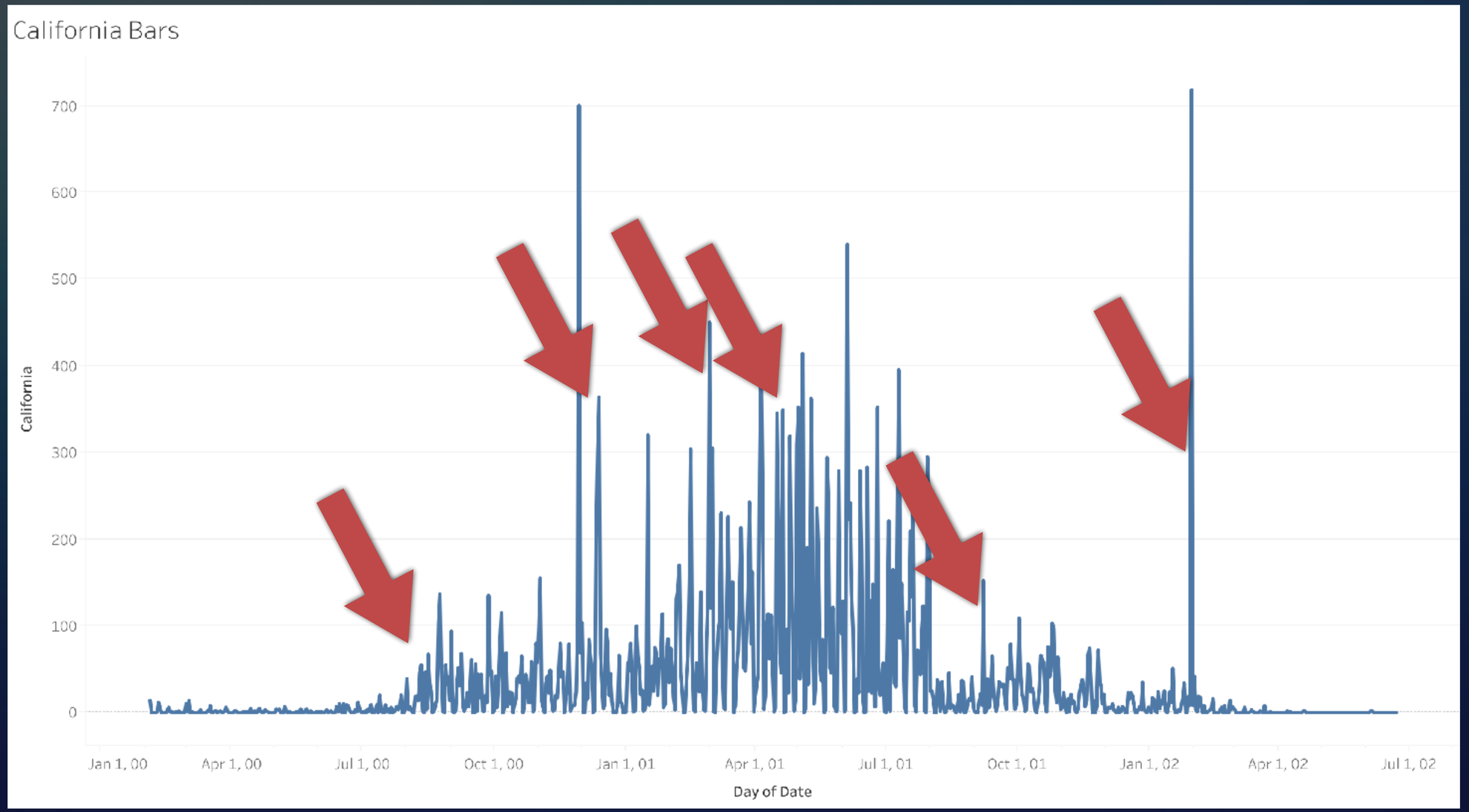
# Practical Applications

# Enron Email Analysis - Cluster 3

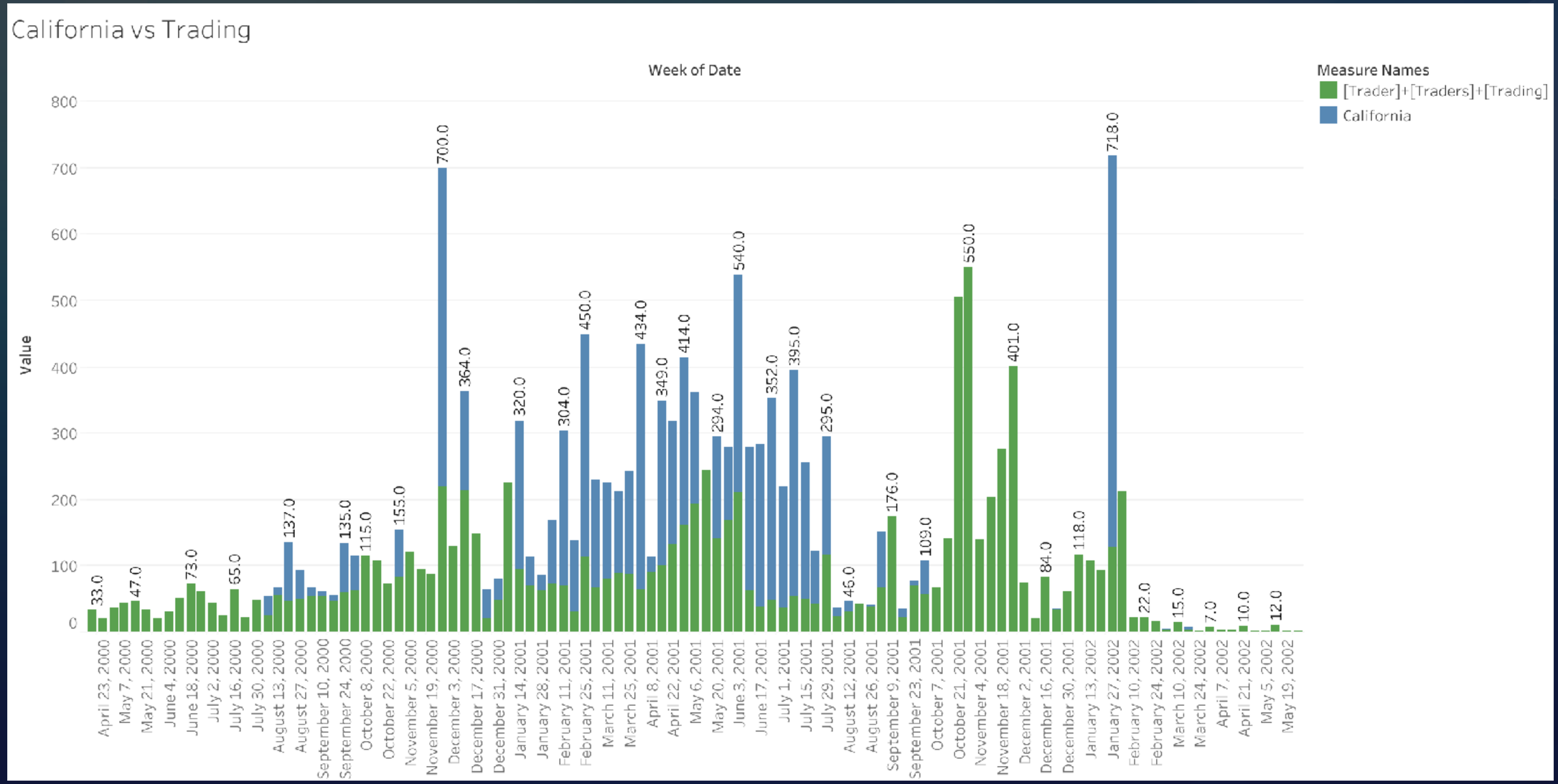
| | |
|---|---|
| manager | known |
| stake | investigation |
| station | involved |
| including | commodity |
| defense | businesses |
| option | tools |
| day | holding |
| gov | assignment |
| evening | hub |
| main | knowledge |

# Q & A

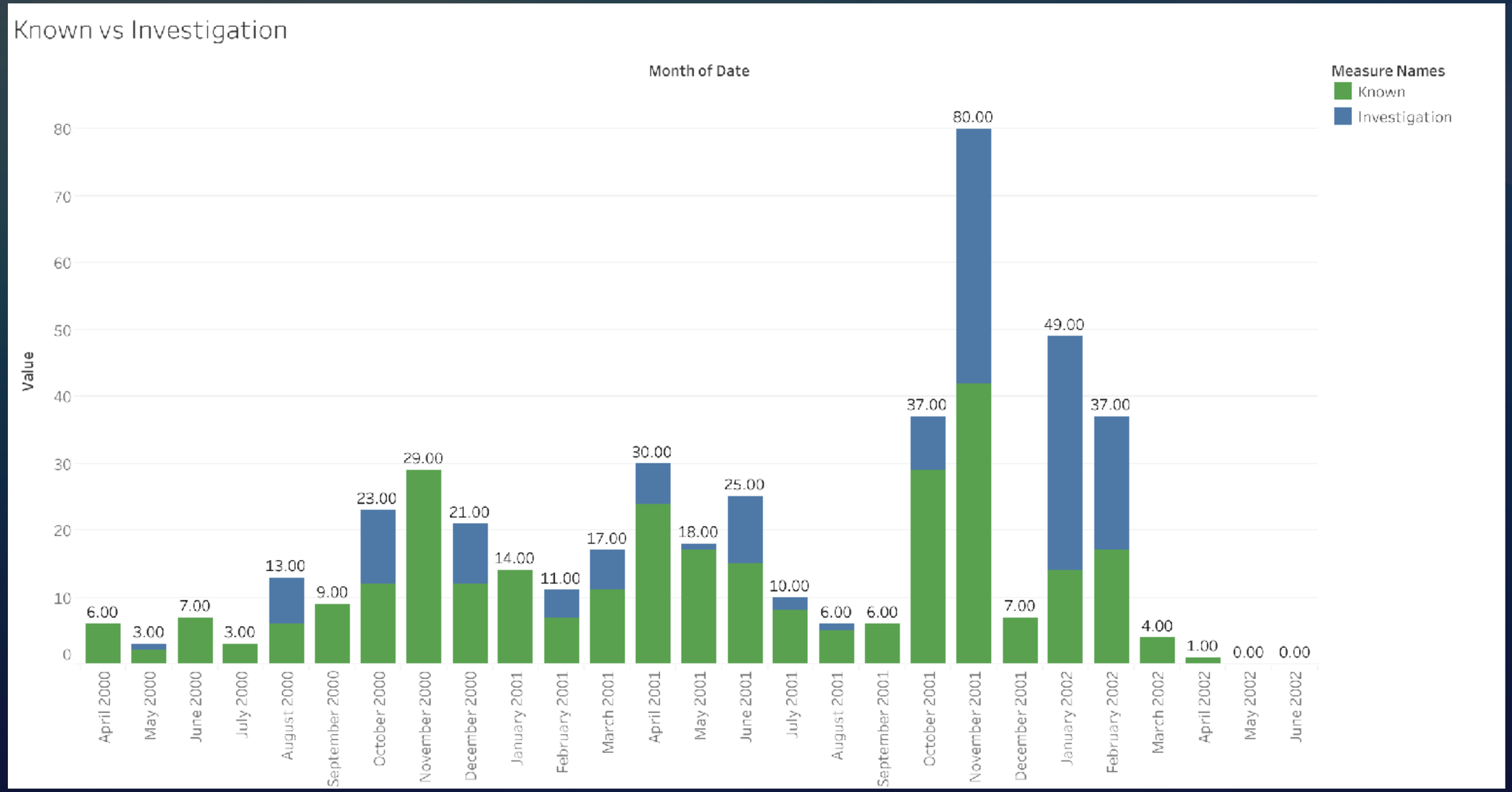
# Interesting Finds



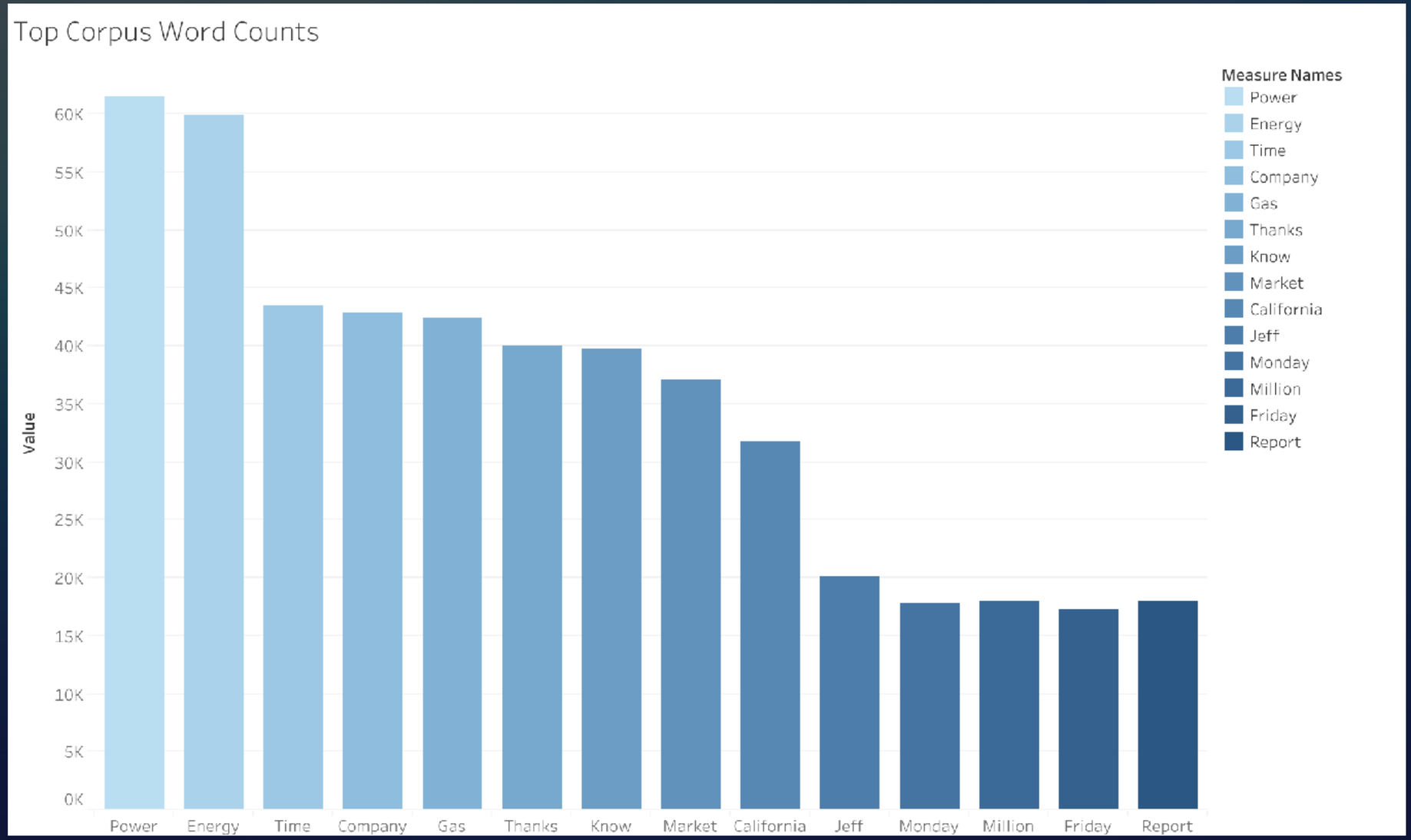

# California Word Count




# California vs Trading

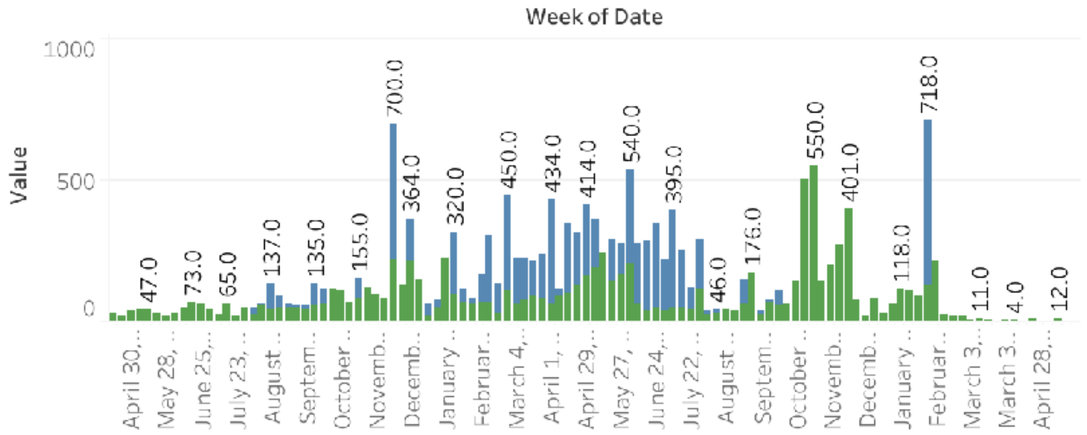



# Known vs Investigation

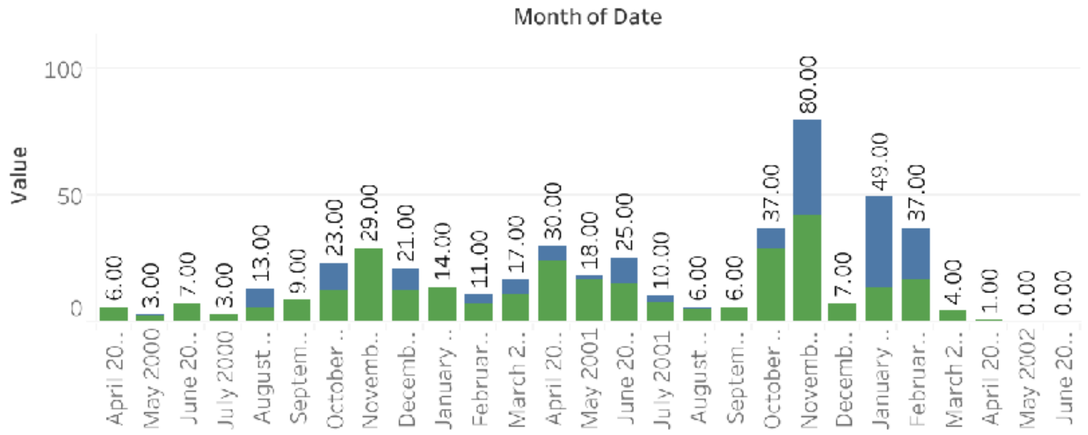



# Top Words For Whole Corpus

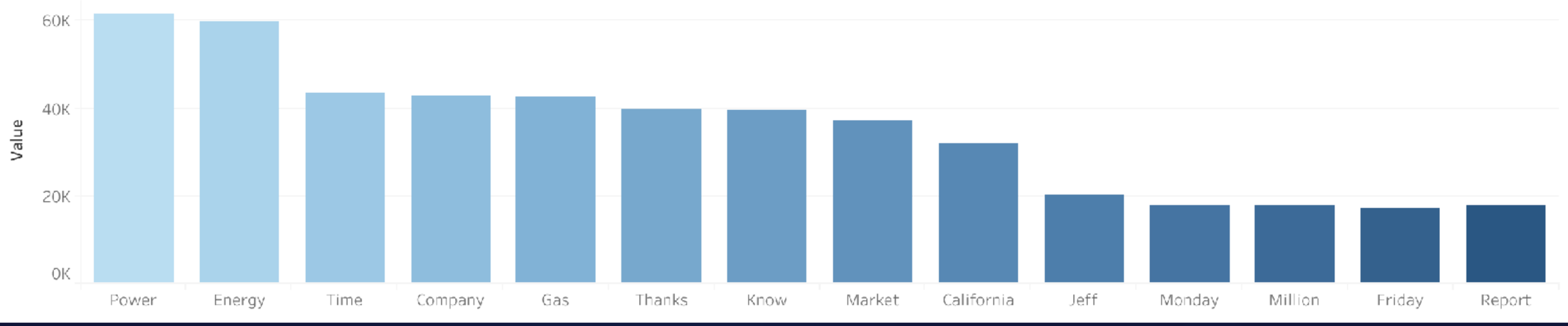
## California vs Trading




## Known vs Investigation




## Top Corpus Word Counts




# MINI DASHBOARD - HTTP://ADRIANCAMPOS.ME

# Q & A

# Originally Terms & Conditions

- Hard to find
- Hard to scrape
- Not easily accessible
- Begin signing up for everything
- No, not really
- But to identify what privacy we may or may not be giving up

# Enron

- Rights given up terms and conditions landed on the Enron Data Set from 2003
- Learned that siri was trained on it…
- You're probably thinking, oh, that's why Siri is so bad
- Only publicly available corpus of that type available to the public
  - What does that mean…. How many of you have gmail…
    - In July 2017 Google promised not to search your emails - don't worry they would only search for what to sell you.They don't do it any more
    - They now search your pictures that you upload to google photos For free for potential marketing opportunites. So again, you're good

# Enron

- If you're not paying for the product, you are the product