LEVERAGING DATA SCIENCE TO CYBER

GENERAL ASSEMBLY

HELLO!

I am Aakash Sharma

I'm here today to discuss how data science can be applied to cybersecurity data from the government, mainly the FCC

The FCC: CSRIC

CSRIC's recommendations will address:

- prevention/remediation of detrimental cyber events
- development of best practices to improve:
  - overall communications reliability
  - availability & performance of communications services
  - emergency alerting during natural disasters, terrorist attacks, cyber security attacks
  - rapid restoration of communications services in the event of widespread major disruptions
  - steps, communications providers can take to help secure end-users & servers

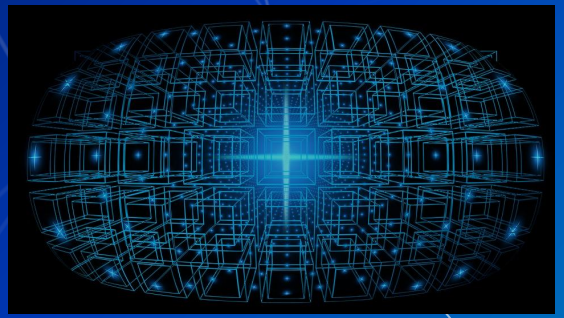
Objective

Provide recommendations to the FCC to ensure optimal security & reliability of communication systems in telecommunications, media & public safety.

My Approach

Exploratory Data Analysis

Clean Data



Model: Priority of the Attacks

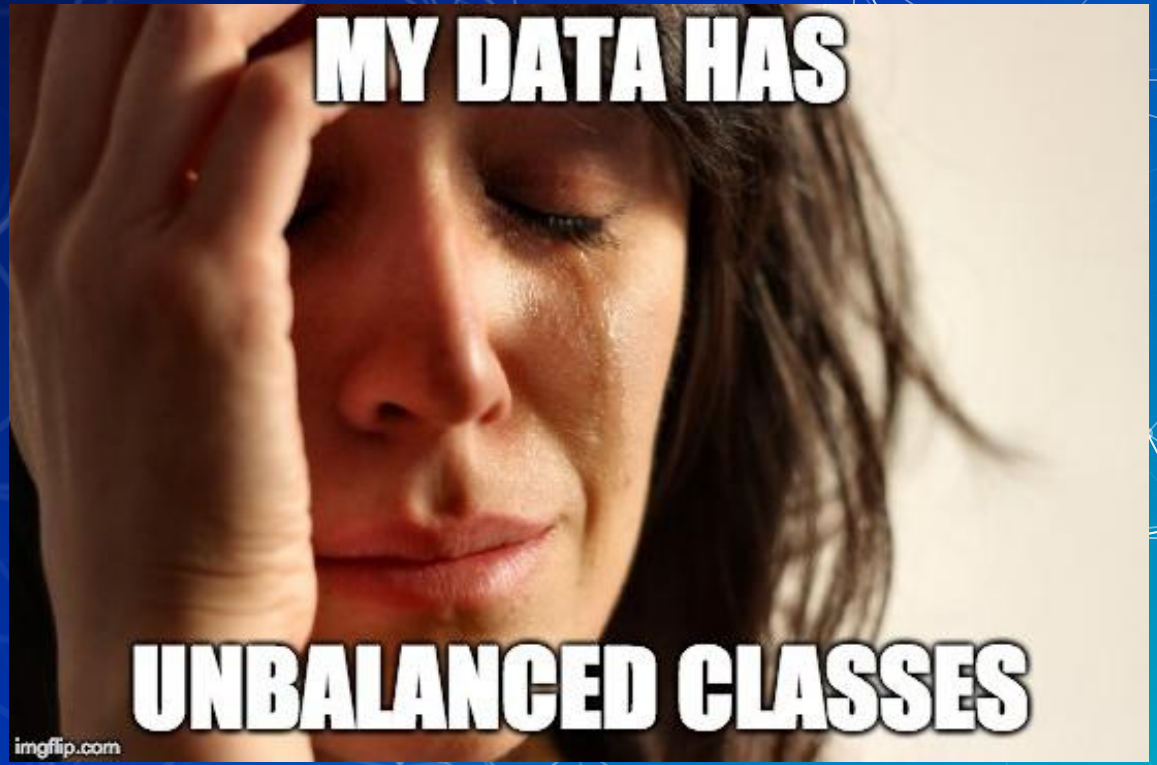
Model: NLP on description of the attack

My Approach Cont.

Diving In!

- Notebook 1 Models:
  - Grid Search (Look for the best features):
    - Naive Bayes (Test for strong independent features)
    - Random Forest (Test strong features randomly for BEST results)
    - Adaboost/Gradient Boost (Reduce error & coach our model on the strongest features)
    - Keras Neural Network (Pattern & Decision Recognition based off strongest/precise features)

The Problem



What does that mean?!

Our target, the priority, has more observations in one specific **class** than the others!

My Approach Cont.

1. Pipeline & Gridsearch: Logistic Regression, Naive Bayes & Random Forest Modeling
2. Preprocessing, Tokenization, Lemmatization, Stemming
3. Data Cleaning, EDA & Regex

Understanding the NLP

- **Preprocessing** is the technique of converting raw data into a clean data set.
- **Regex,** regular expression is a string of text that allows you to create patterns that help match, locate & manage text.
- **Lemmatizing** is the process of grouping together the inflected forms of a word so they can be analyzed as a single term.
- **Stemming** is the process of reducing inflected words to their stem, base or root form.

Priorities of Network Disservice

1 as “important”, 2 as “highly important” & 3 as “critical”

Largest Positive Correlated

- Network Operator
- Equipment Supplier
- Property Manager
- Service Provider
- Internet/Data

Largest Negative Correlated

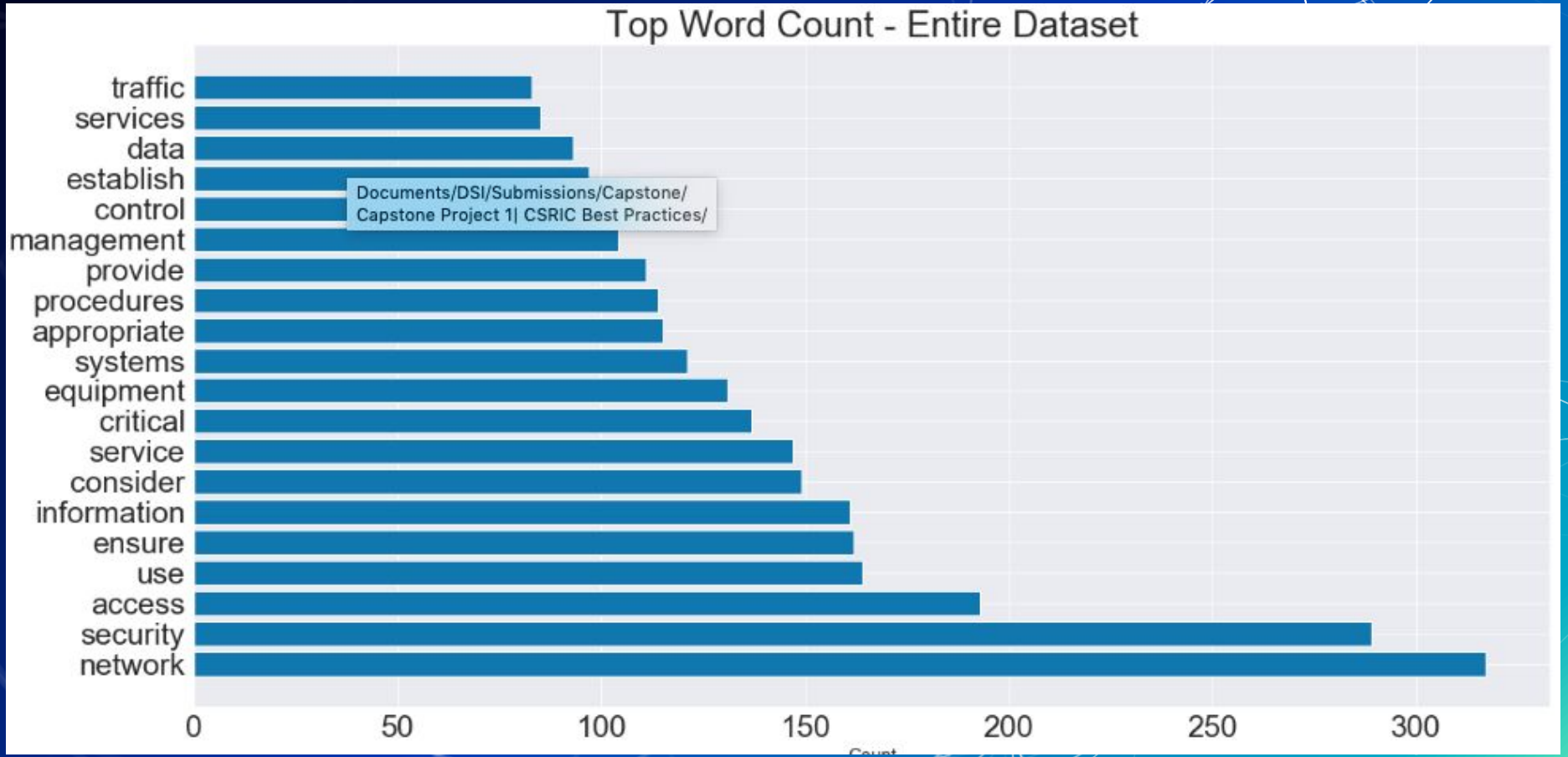
- Wireline
- Wireless
- Satellite
- Public Safety

Notebook 1 Models

	Naive Bayes	Random Forest	Adaboost	Gradient Boost	Keras Neural
Train	62.1659%	66.1040%	60.9001%	66.1040%	61.8846%
Test	57.0491%	59.0163%	56.3934%	57.3770%	57.0491%

Notebook 2 Models

	Logistic Regression Count Vectorize	Logistic Regression TFIDF	Random Forest Count Vectorize
Train	97.4683%	95.7805%	95.7805%
Test	67.5409%	66.8852%	64.5901%

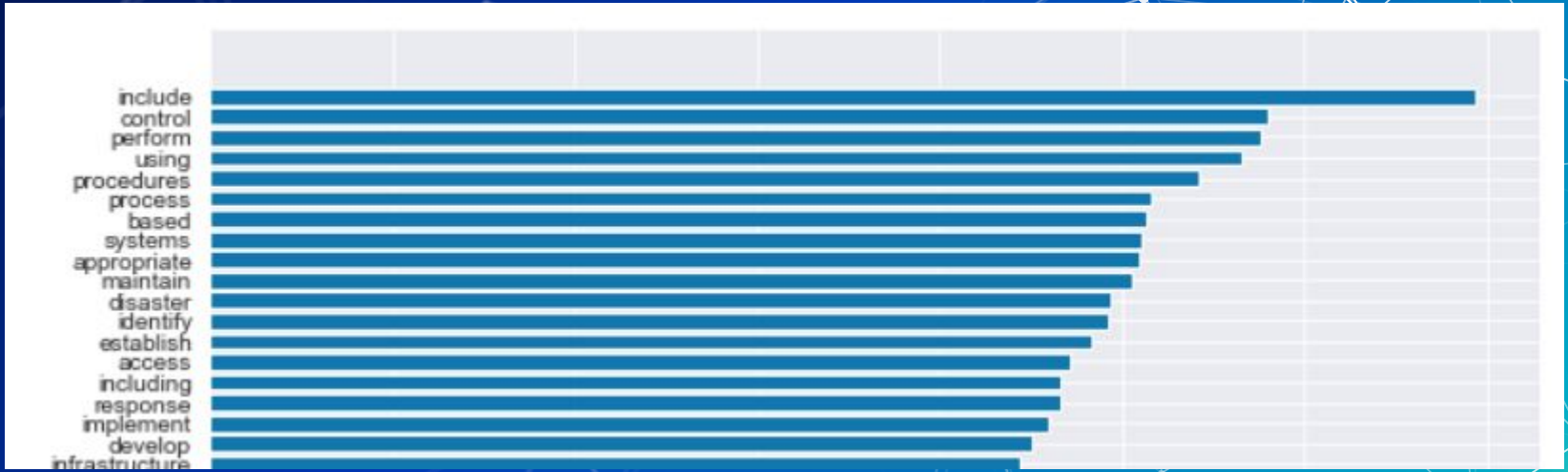


Most Common Words

Recommendations

My Process & FCC Data

Predicting Network Based Problems



My Hurdles

- Not Enough Data
- Unbalanced Classification Problem
- Not Knowing Where to Start
- Vague Data

Recommendations

Network Based

Better methodologies that improve security practices

Simple:

- Different Cables
- Color Code Cables
- Better ventilation of warehouse
- Increase power capacity
- Better hardware
- Spacing of antennas

Recommendations Cont.

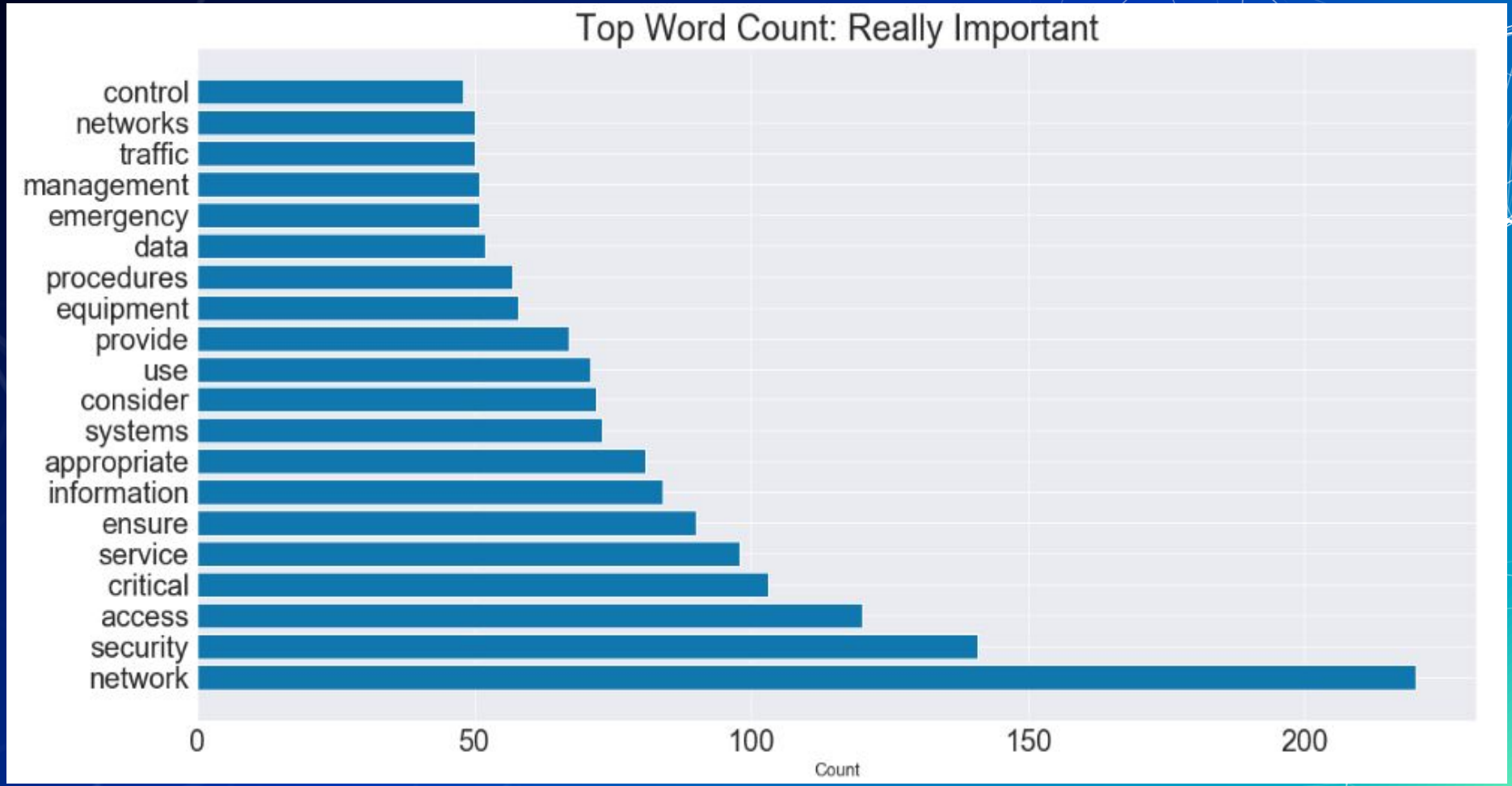
Moderate:

- Utilize Network Surveillance
- Provide secure electrical software where feasible
- Find thresholds for new hardware & software
- Virus protection

Recommendations Cont.

Complex:

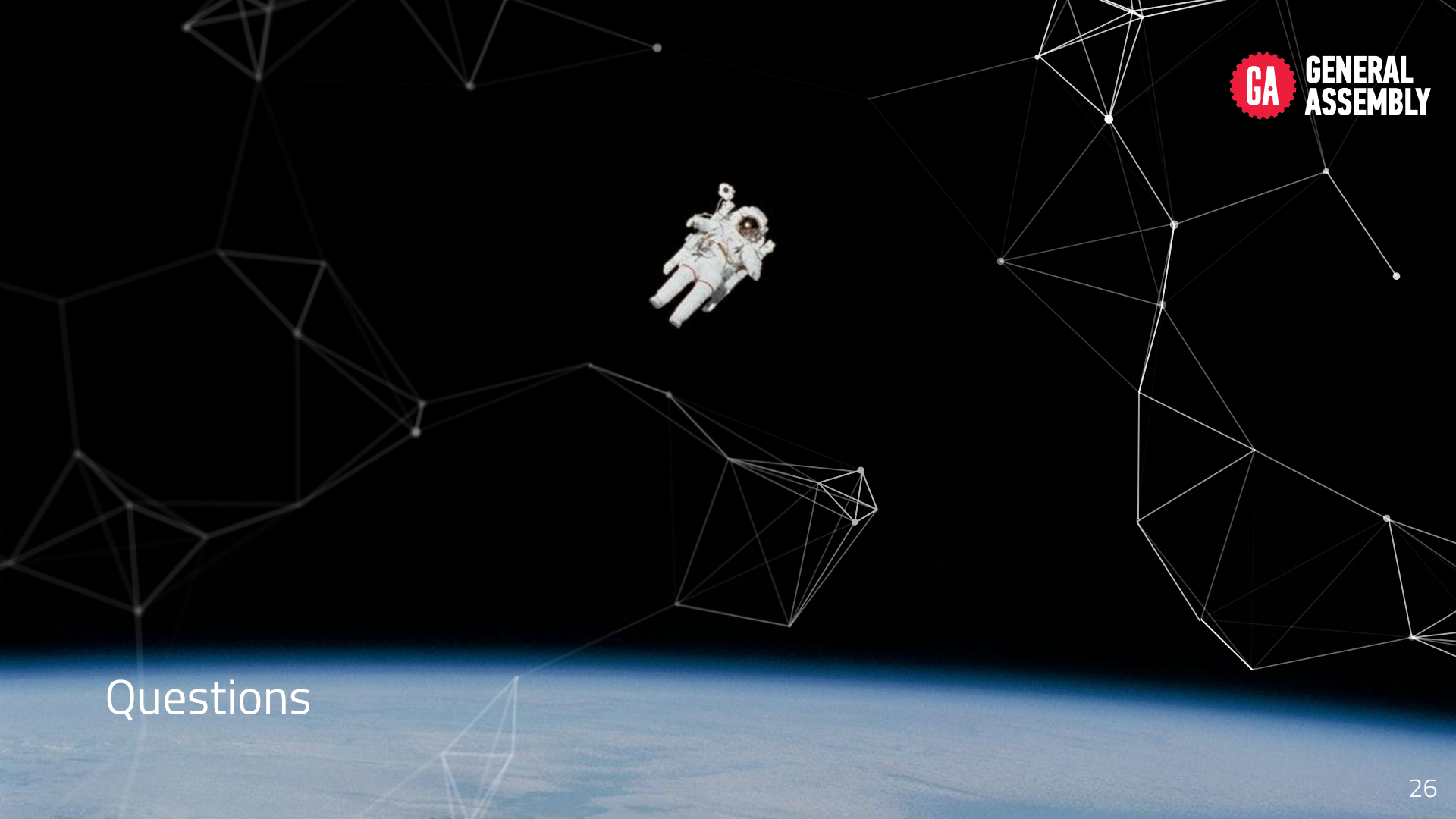
- Minimizing single points of failure
- Device Management Architecture
- Secure networks
- Encrypted systems
- Software faults


Top Word Count: Really Important
control
networks
traffic
management
emergency
data
procedures
equipment
provide
use
consider
systems
appropriate
information
ensure
service
critical
access
security
network
0
50
100
150
200
Count

“A PICTURE IS WORTH A THOUSAND WORDS”

Description Column

Lemmatized & Stemmed Pre-Processed Columns



Questions