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 the prevention and remediation of detrimental cyber events, the development of best practices to improve overall communications reliability, the availability & performance of communications services and emergency alerting during natural disasters, terrorist attacks, cyber security attacks or other events that result in exceptional strain on the communications infrastructure, the rapid restoration of communications services in the event of widespread or major disruptions and the steps communications providers can take to help secure end-users and servers.

Objective

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

My Approach

- Clean Data
- Exploratory Data Analysis
- Performing models on the Priority of the Attacks
- Performing Natural Language Processing on the description of the attack

Notebook 1 Models:

Grid Search:

- Naive Bayes
- Random Forest
- Adaboost
- Gradient Boost
- Keras Neural Network

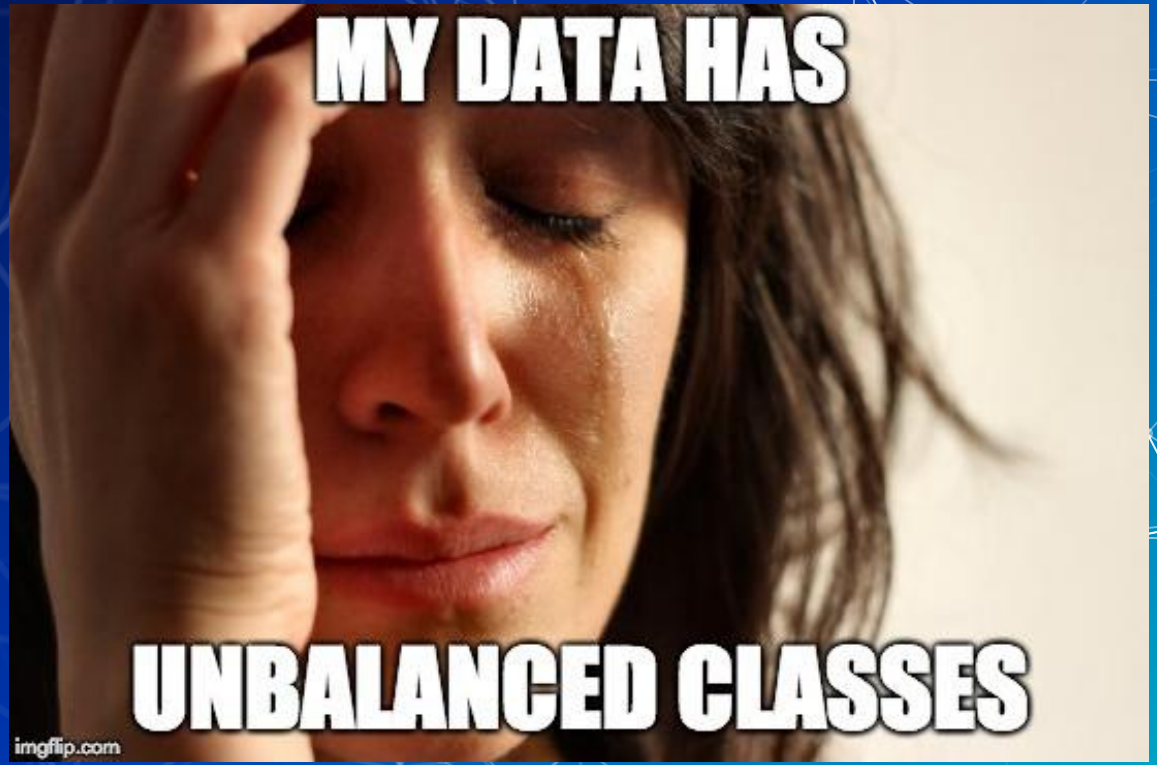
Notebook 2 Models:

- Lemmatizing,
- Stemming
- Count Vectorize
- TFIDF Vectorizer

Grid Search & Pipeline:

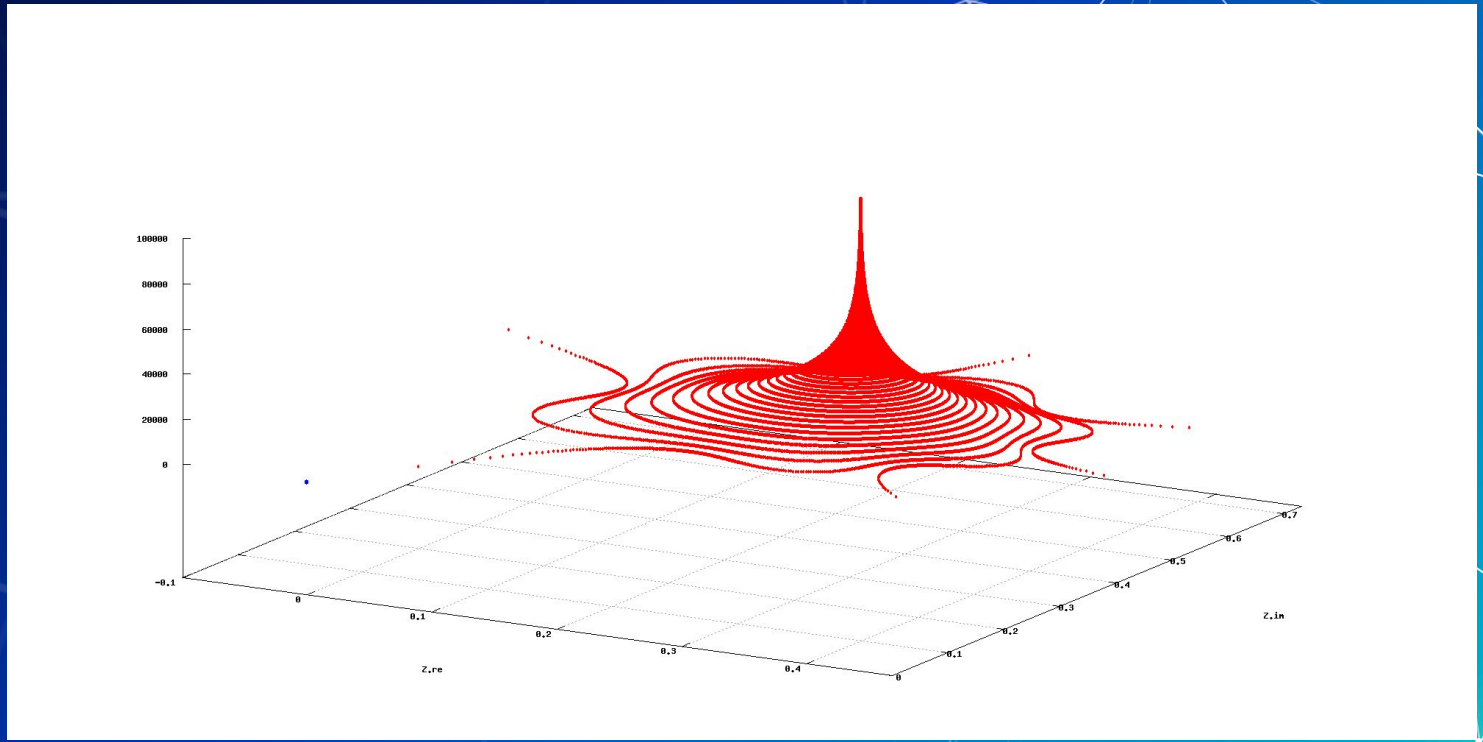
- Logistic Regression
- Random Forest

The Problem



Priorities of Network Disservice

Ranked: 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%

My NLP Process

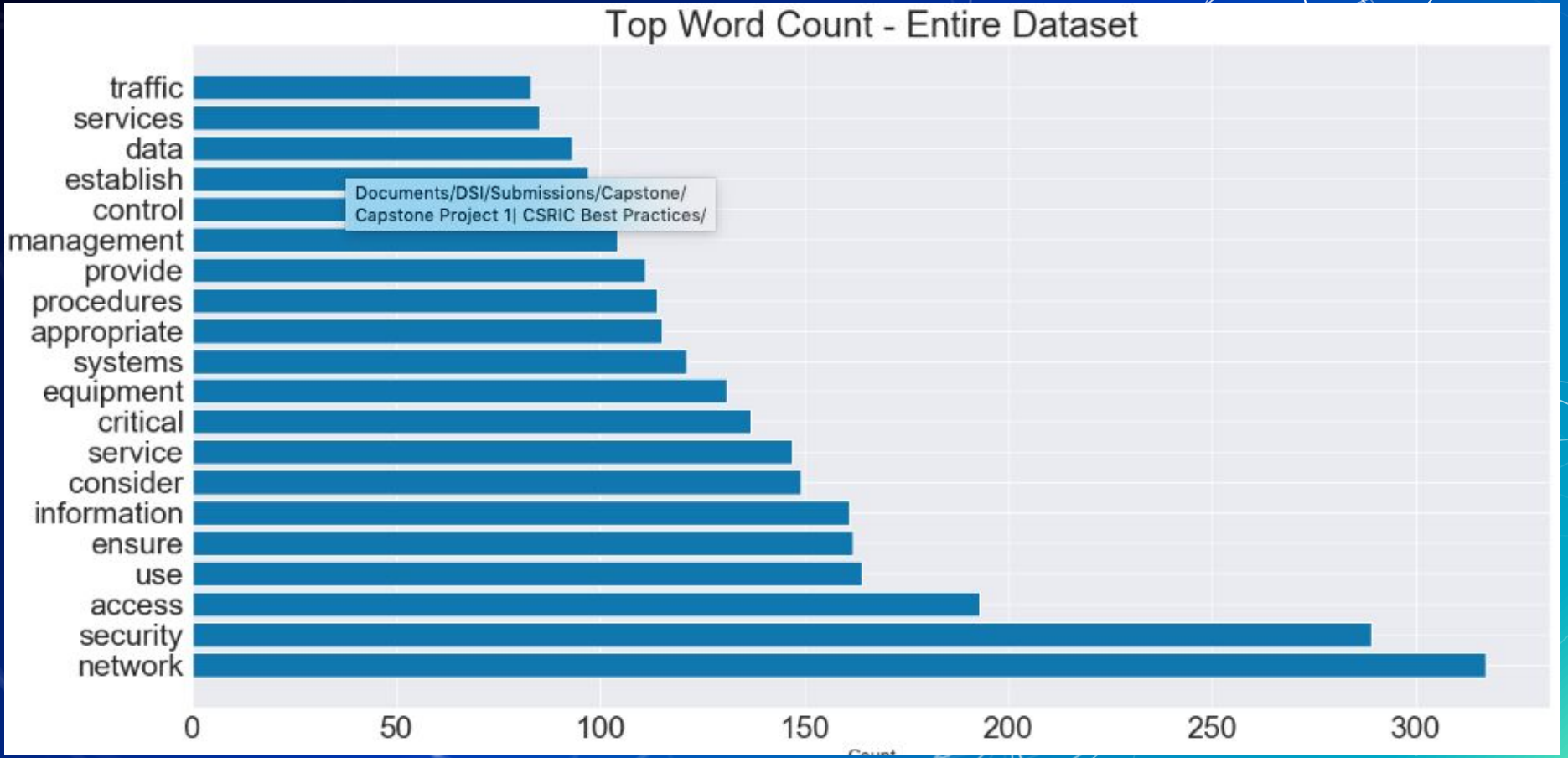
1. Pipeline & Gridsearch: Logistic Regression, Naive Bayes & Random Forest Modeling
2. Preprocessing, Tokenization, Lemmatization, Stemming
3. Data Cleaning, EDA & Regex

Understanding the NLP

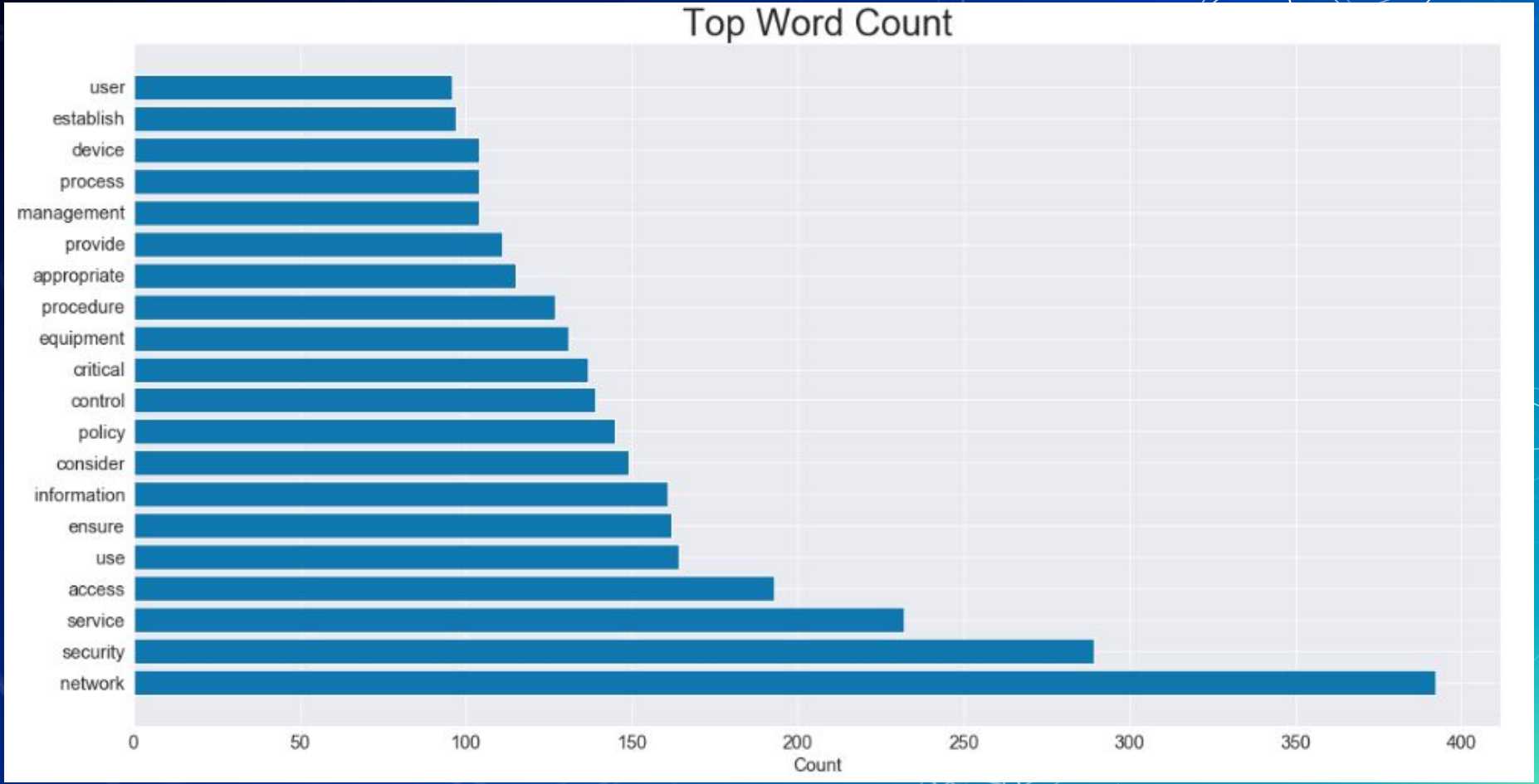
- **Preprocessing** is the technique of converting raw data into a clean data set.
- **Regex,** regular expression is a string of text that allows you to create patterns that help match, locate & manage text.
- **Lemmatizing** is the process of grouping together the inflected forms of a word so they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

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%



Count Vectorized

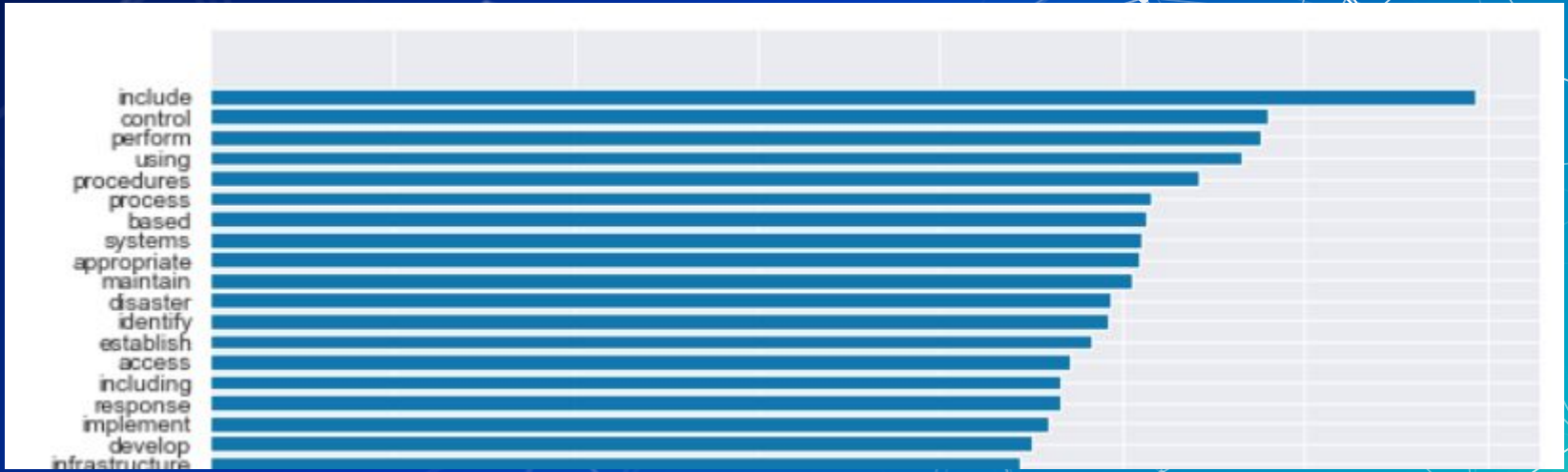


Lemmatized & Stemmed

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:

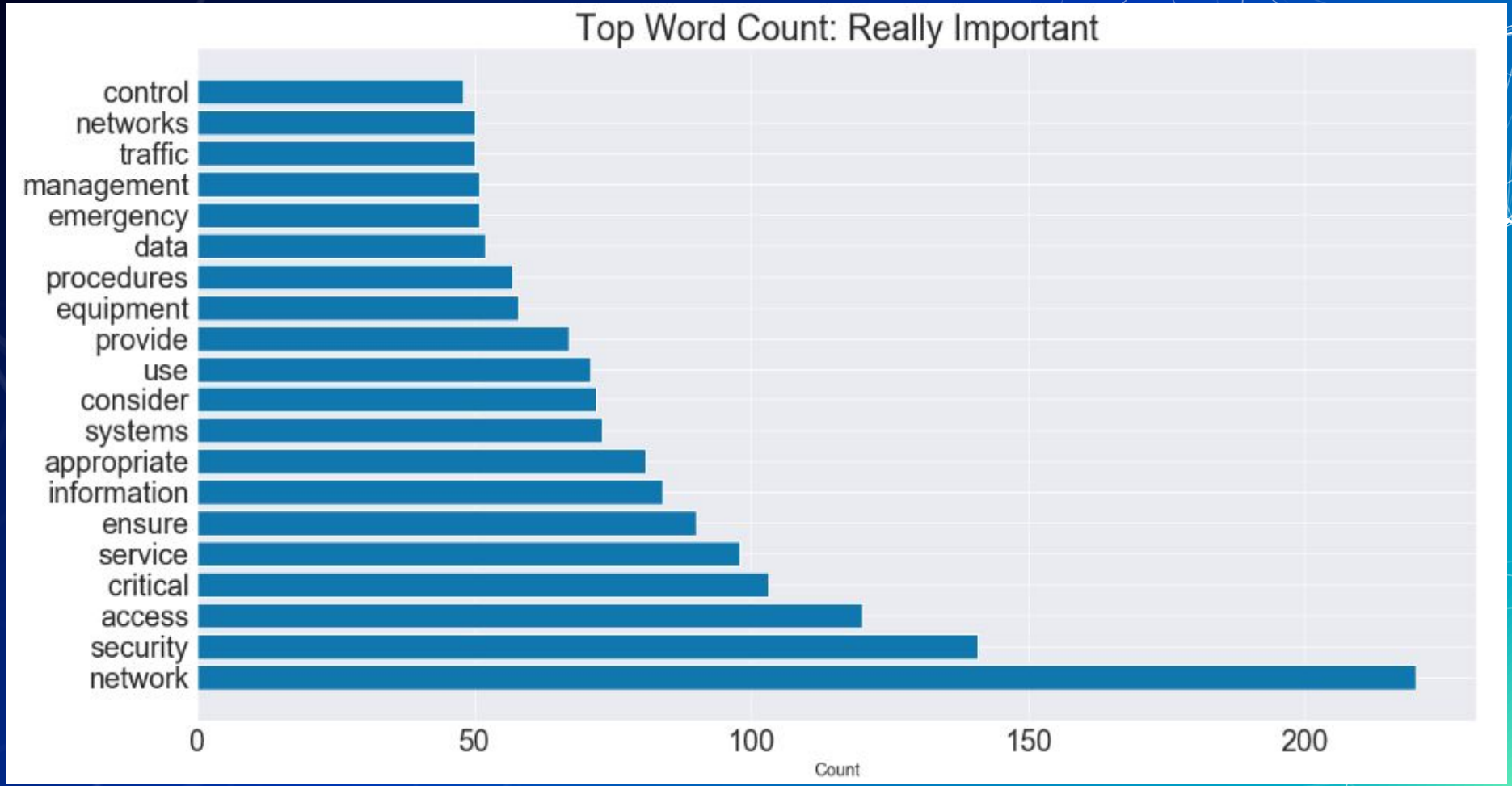
- Different Cables
- Color Code Cables
- Better ventilation of warehouse
- Increase power capacity
- Better hardware
- Spacing of antennas

Simple:

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

Complex:

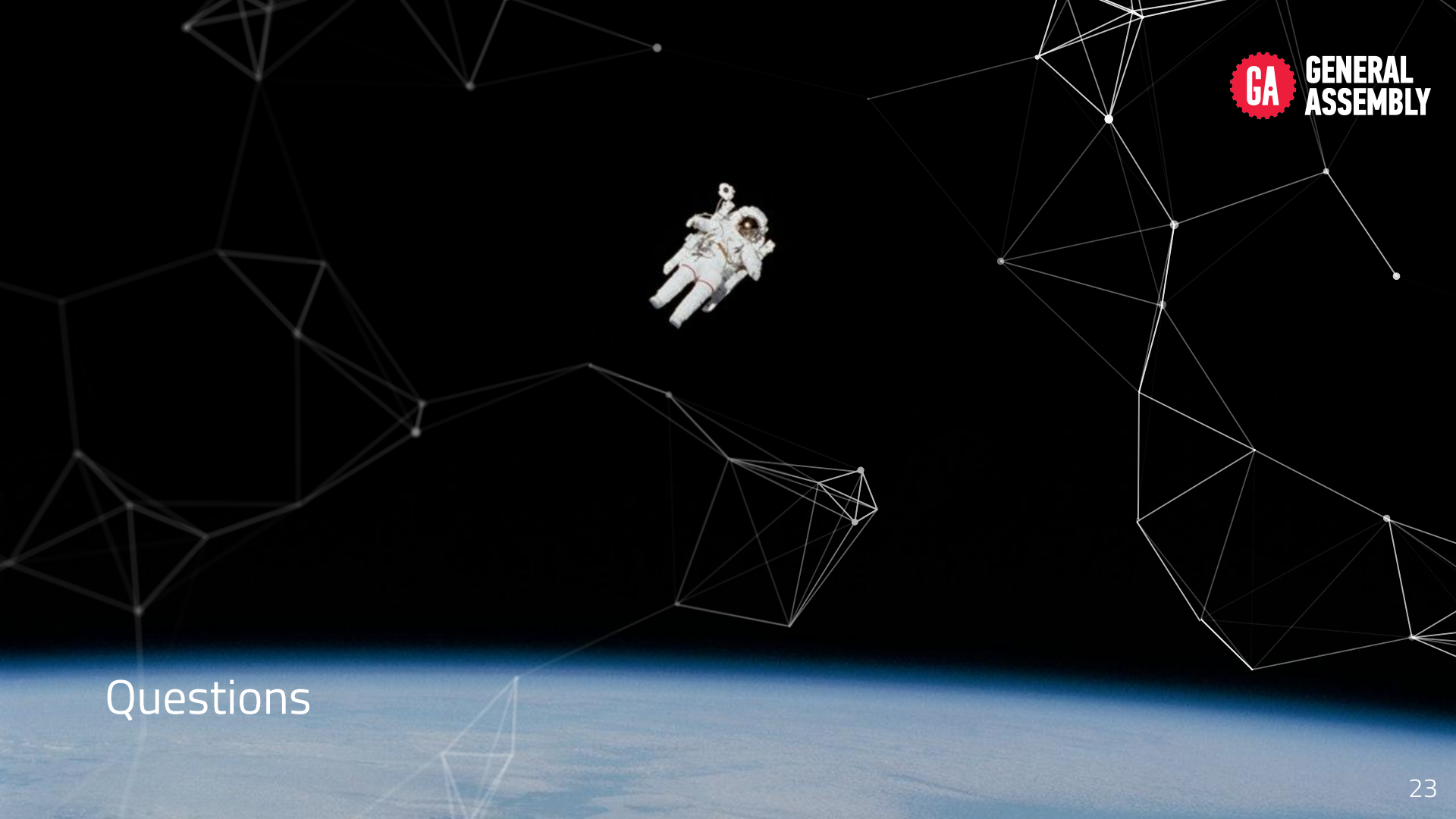
- Minimizing single points of failure
- Device Management Architecture
- Secure networks
- Encrypted systems
- Software faults


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