The Power of Bad Data

Ms. Alicia Scott

CAPT Thomas McKay, USN
Commanding Officer

Dr. Angela Lewis, SES
Technical Director

NAVSEA
WARFARE CENTERS
CRANE

NAVSEA
WARFARE CENTERS
CRANE

Personal Introduction

ALICIA SCOTT

Work Experience

Currently leads a Data Science Team for NSWC Crane. Started in Industry for a Market Research Firm.

Patents

Patent for Natural Language Processing for Short Text Data
Patent Pending For Data Preprocessing Algorithm.

Skills & Languages

- Data Centric Design
- Statistical Analysis
- NLP
- Python
- R
- SQL

Education

Pursuing Masters in Information and Data Science at UC Berkeley.
Bachelors in Theoretical Mathematics

Hobbies

Reading copiously, traveling, hiking and plant care

Contact Info

alicia.scott1@navy.mil

The Power of Data

We have seen the revolutionary effects of data in the modern age

NAVSEA
WARFARE CENTERS
CRANE

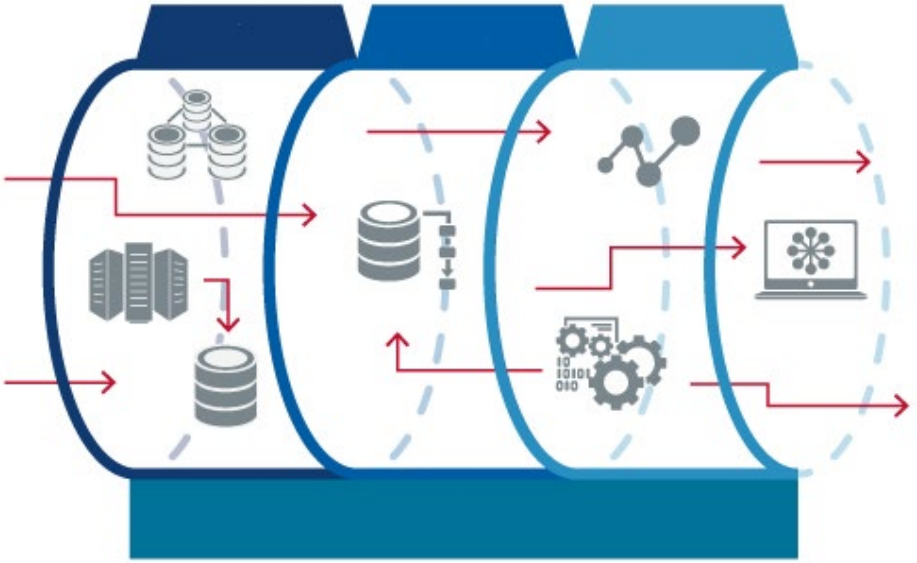
Not Every Business is New Age

Not every business is built on new age data and can't be expected to have the same level of data strategy

NAVSEA
WARFARE CENTERS
CRANE

Waiting for Good Data to Happen

- Waiting for the perfect opportunity for building the perfect data perspective is a sliding timeline
- Good Data doesn't just happen
- Lacks the Realization of the benefits that could be had presently



The Benefits of Bad Data

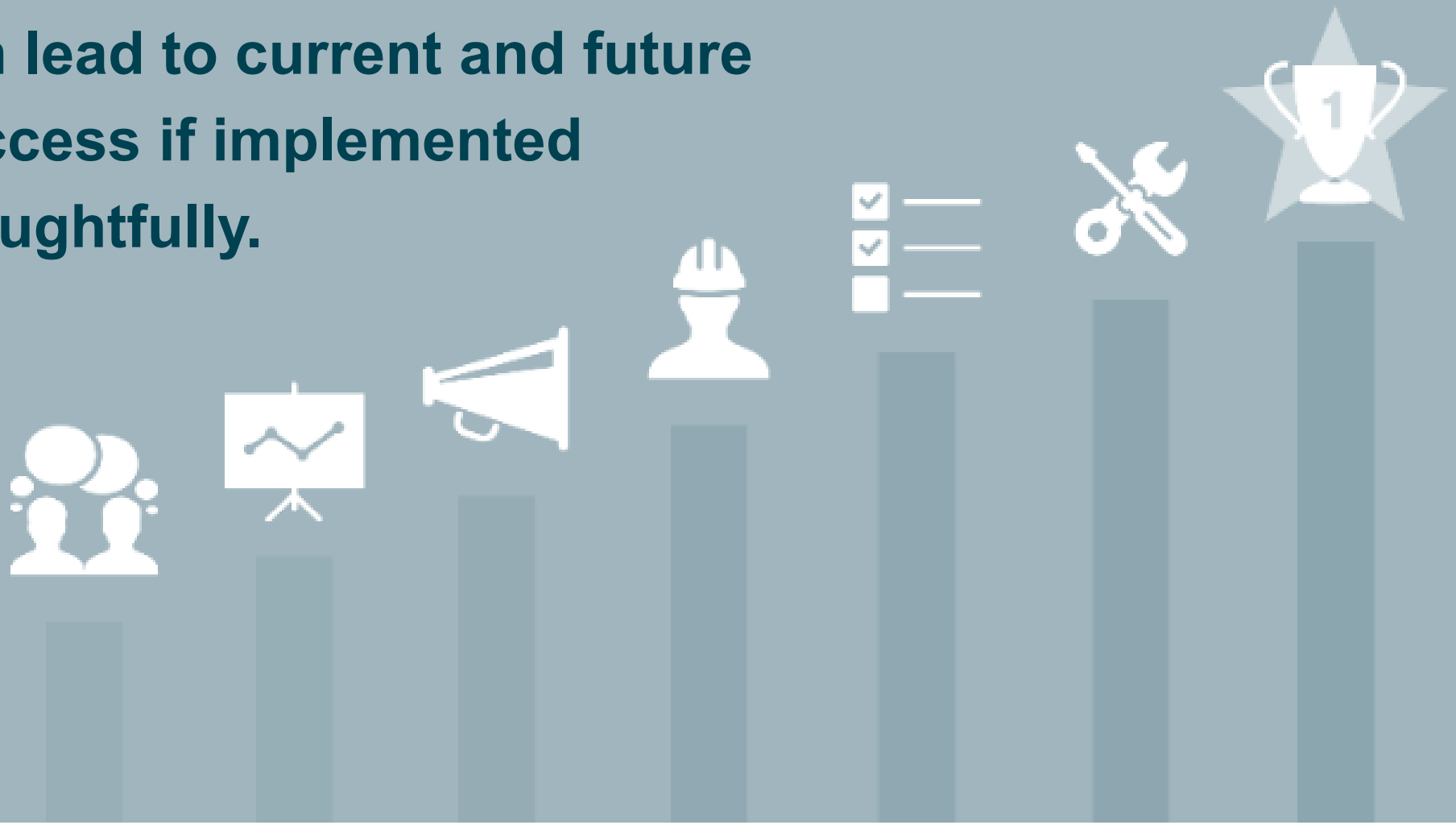
The untapped resource of 'bad data' could be a strong asset for organizations with legacy information or unedited elements.

Data Quality Does Matter

- This is not to say that the 5 Vs of Data don't matter.
- Organizations and Analysts need to have an honest perspective of the state of the data
- Bad Data should be used as a piece of the puzzle, but not the ruling authority on all
- There is good to have even outside of perfection

Working with the Data You Have Now

Working with the present possibilities can lead to current and future success if implemented thoughtfully.



Tactical Options for Bad Data

Data Preprocessing and a Data Cleaning Pipeline

Collection
- Grab Anything and Everything Approach
- Continuation Approach

Data
- Pare Down to Relevant Information
- Grab a Subset from the Source

Sorting and Cleaning
- The 60-70%
- The Make or Break Point

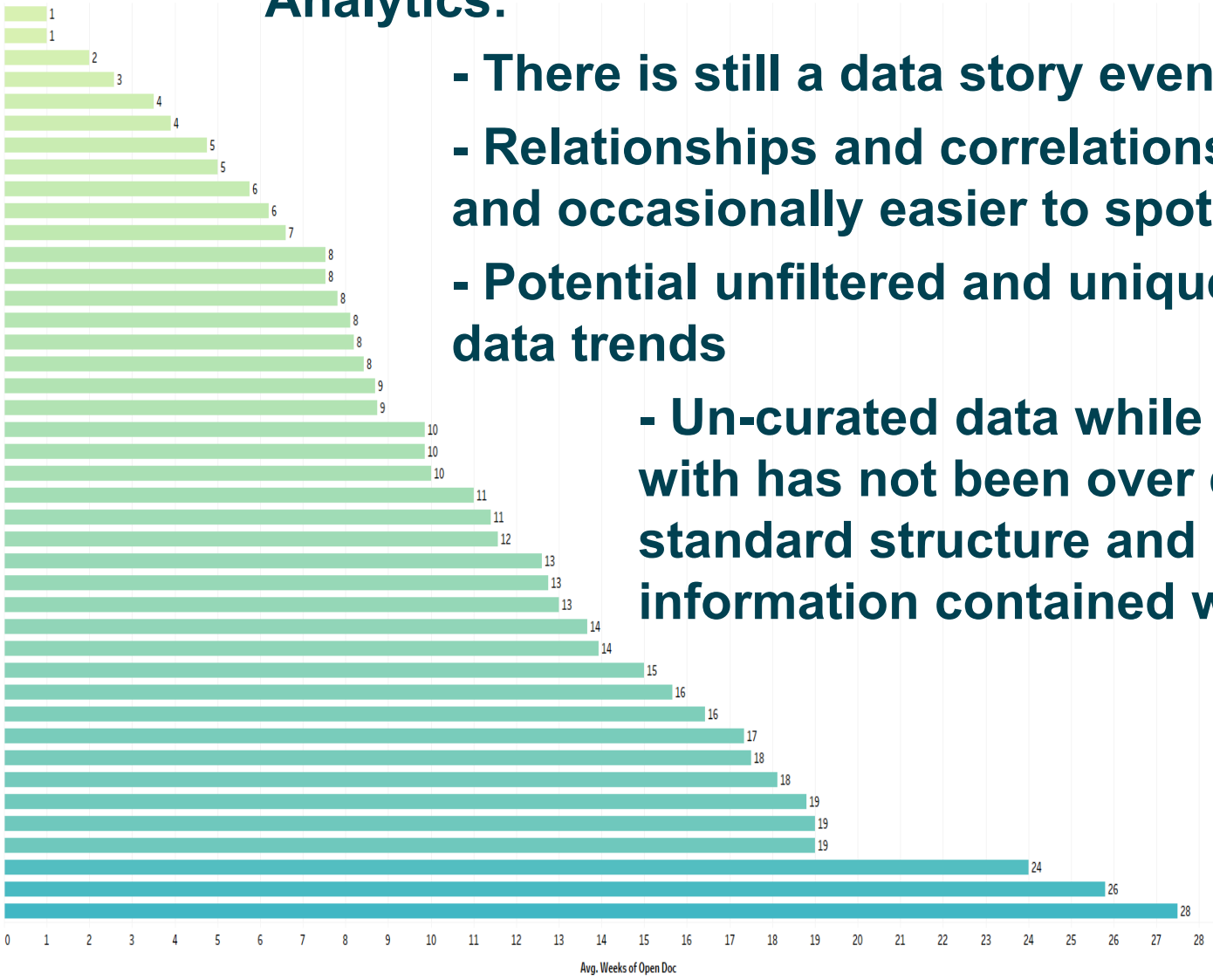
Arranging
- Aligning for the Specific Purpose
- Formatting and Preprocessing

NAVSEA
WARFARE CENTERS
CRANE

Good Data Tools for Bad Data

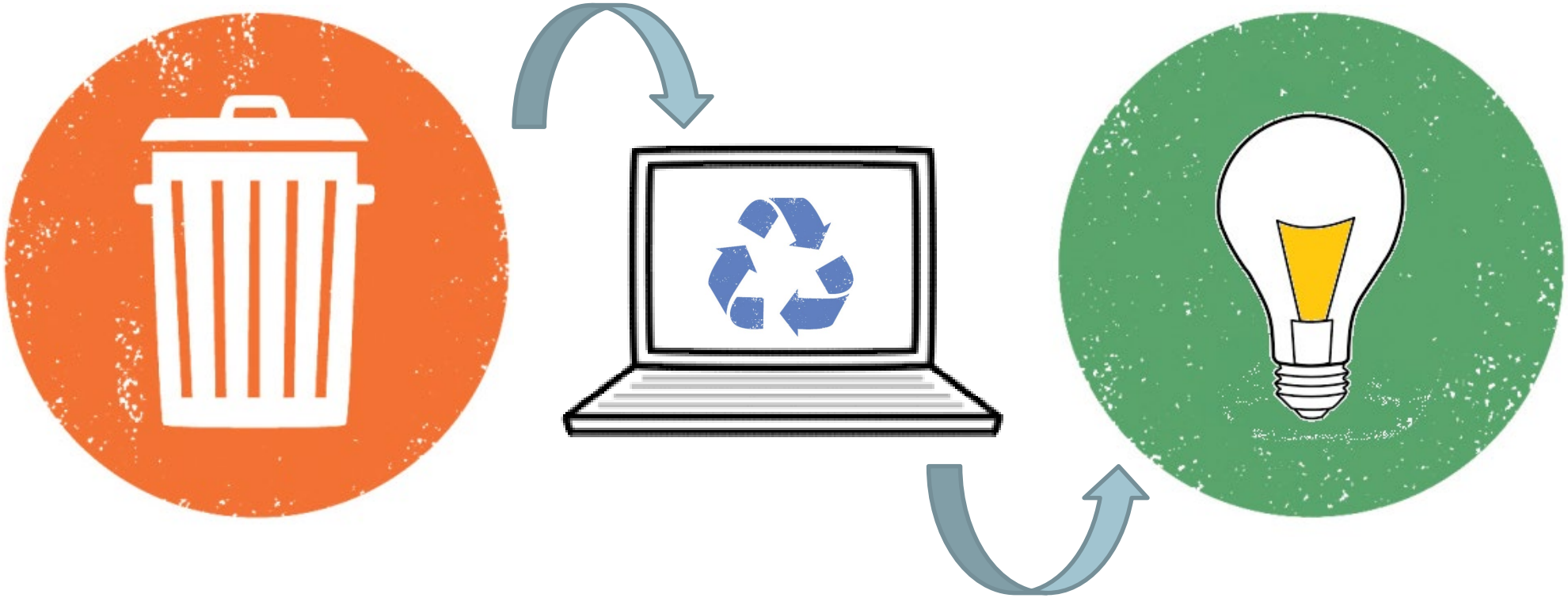
Analytics:

- There is still a data story even in imperfect data.
- Relationships and correlations are still available and occasionally easier to spot in ‘bad data’
- Potential unfiltered and unique perspective into data trends
  - Un-curated data while more difficult to work with has not been over designed into a standard structure and often has more information contained within



AI for Bad Data

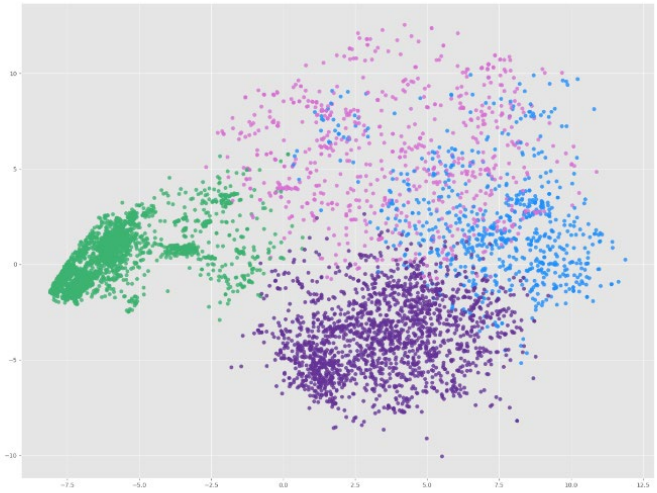
AI with 'bad data' must have a larger focus on Trust and Verification

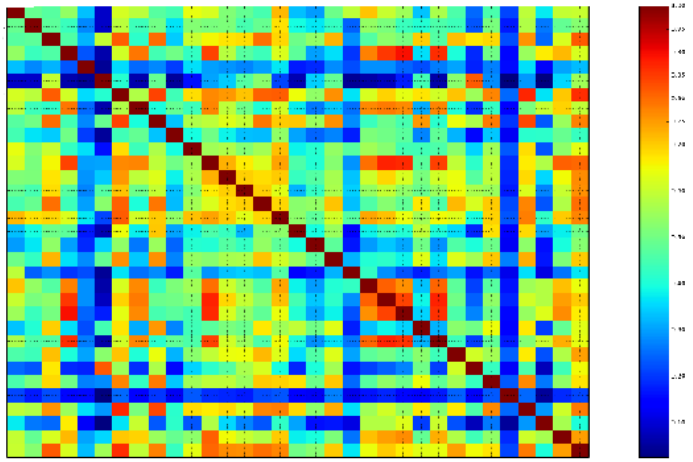


NAVSEA
WARFARE CENTERS
CRANE

ML for Bad Data

- **Using Machine Learning as a prompt for deeper analysis**
  - **Pinpoint trends or clusters in the data for a more rigorous statistical analysis to confirm**
- **Using Machine Learning to fix issues in the data**
  - **Depends on the Reliability level of original data**
- **Using Machine Learning or algorithmic coding to reformat or populate data in a new way**
  - **Automation / Storage / Pipeline**





NLP for Bad Data

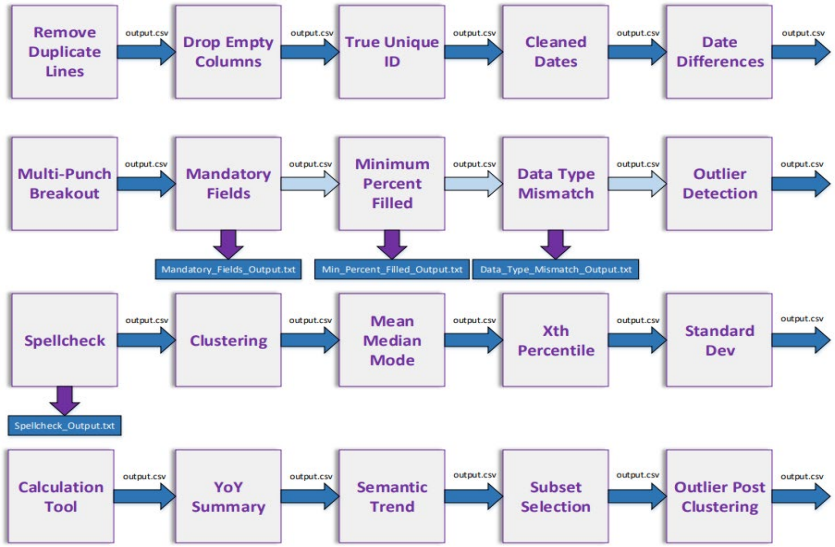
- Found the most success with NLP used for ‘bad data’
- Relates concepts through language or linguistic trends
- Able to pull a more detailed picture through mining text data
- NLP as a ‘bad data’ Powerhouse

Data Science Team Toolkit

Standard Open Source tools were often built to work with good data or perfect conditions. We didn't have those. So we built our own.

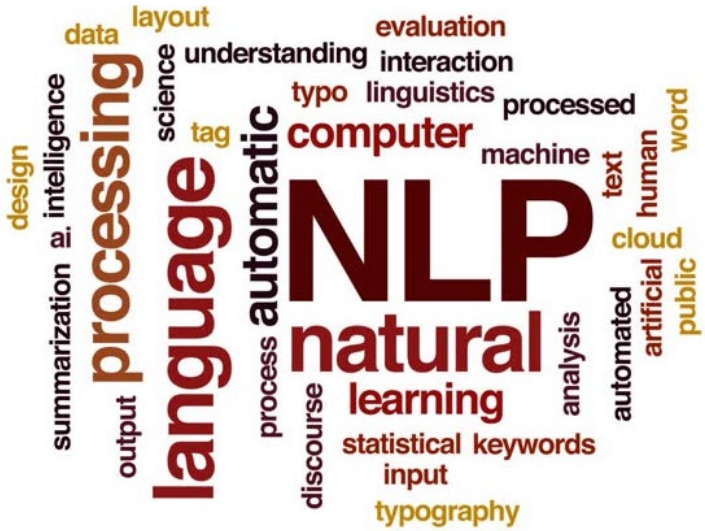
Data Science Team Toolkit

- **Data Preprocessing with ML and NLP**
  - Create a modular and agnostic algorithm for data cleaning, data transformation, analytics, and artificial intelligence solutions
  - For analysts and researchers to mitigate the current issues of spending up to 80% of their research time on data wrangling
  - Bringing AI to a more accessible level



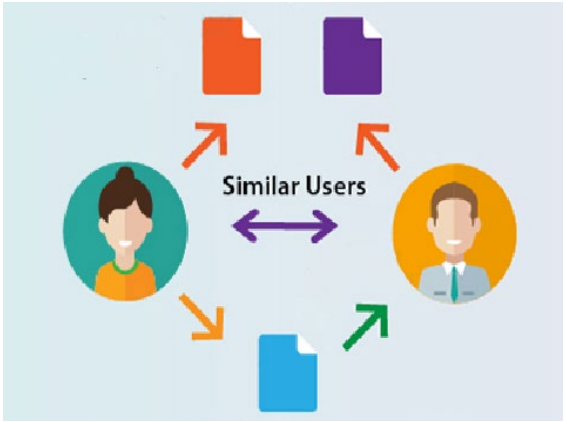
Data Science Team Toolkit

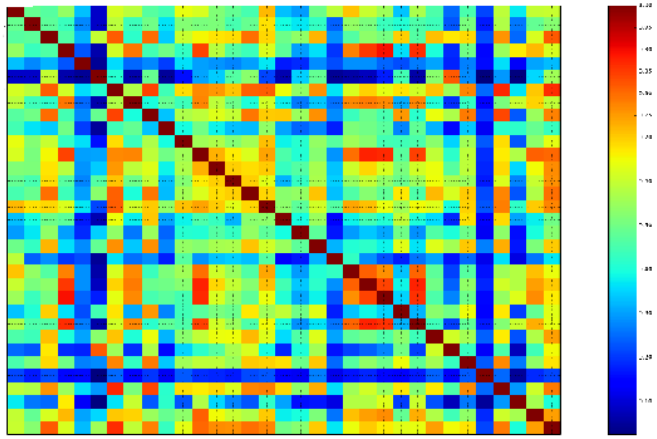
- **NLP for Reliability**
  - Solves current labor intensive process of mining failure records
    - In-House developed algorithm
    - 85% accuracy in Top 10 Comparison
  - Able to automatically identify trends
    - Deciphers plurals, parts of speech, and different capitalization
    - Maps Linguistic Uses of words to identify commonality
    - Scalable and Multi-use



Data Science Team Toolkit

- **ARISE - Automated Recommendation Investigator for Systems Engineering**
  - **Crane, Corona, and Port Hueneme will work together to create a topic modeling tool that utilizes NLP and ML to establish likeness in disparate maintenance or failure records**
  - **Scalable and able to produce a map of similarities mined from the text and actions to give suggestions.**





NAVSEA
WARFARE CENTERS
CRANE

Conclusion

Bad Data can be an asset to an organization as long as the correct practices and tools are applied to it.

Learning to work in a 'Bad Data' World can enable your team and organization to be more impactful

NAVSEA
WARFARE CENTERS
CRANE

Questions and Discussion