Using Data Analytics to Map the Corrosion Science and Technology Landscape

Dr. Elissa Trueman, PE

Naval Surface Warfare Center, Carderock Division

CAPT Todd Hutchinson

Commanding Officer, NSWCCD

Lawrence Tarasek

Technical Director, NSWCCD

AUG 2022

Outline

NAVSEA WARFARE CENTERS Carderock Division

- What is topic modelling?
- Keyword Query
- Data sources and Retrieving Relevant Documents
- Visualizations
- Data Analysis
- Conclusions

Disclaimer: The opinions and data interpretations in this document are solely those of the author and not of the Navy, Office of Naval Research, Naval Surface Warfare Center Carderock Division, or Digital Science and Research Solutions, Inc. The approval for public release does not constitute endorsement of the findings presented herein.

Topic Modelling

End goals: (i) Capture documents based on specific keywords and data sources; (ii) Identify primary investment areas and possible gaps; (iii) Identify potentials collaborations

STEP 1: Keyword Query
What terms to include?
How to accurately capture the field of research?

```
"molecular engineering" OR bionano* OR
 OR fullerene* OR fullerite* OR
d* OR nanodiamond* OR "condensed
s" OR "2D material" OR "2D materials" OR
ion metal dichalcogenides" OR nanowire* OR
"porous carbon" OR "optical metamaterial" OR…
```

STEP 2: Retrieve **Relevant** Documents
What time period to include? What document types?

```
query = f"""
    search publications in title_abstract_only for "{keywords_boost}"
    where year in [{min_year}:{max_year}] and type = "article"
    return publications
    sort by relevance desc
    limit 50000
"""

df = dsl.query_iterative(query)
```

STEP 3: Visualize Data
How large is the field? How large is the investment? What are topics in the field? How quickly are they growing? Who are current funders? Who are current performers?

This slide is courtesy of Danielle Paynter.

Step 1: Keyword Query

corrosion AND NOT (OR justice OR ethic* OR violence OR "social mission" OR legal OR law OR Clinton OR envy OR humanities OR fixity OR spirit OR "traditional values" OR "language" OR "corrosion of identity" OR "journalistic impartiality" OR "of the sentences" OR juxtaposes OR gendered OR soul OR personality OR politic* OR micropolitic* OR philosophical OR pension* OR morte OR "public spaces" OR social OR goat OR brainstem)

* is the wildcard operator

~ is the ambiguity operator (not used here)

Quotation marks (“phrase”) allow grouping of terms

Query goal for this work was to keep the results as broad as possible while excluding non-relevant results.

Step 2: Retrieving Relevant Documents

Data Source: Dimensions provided by Digital Science & Research Solutions, Inc. for the Office of Naval Research

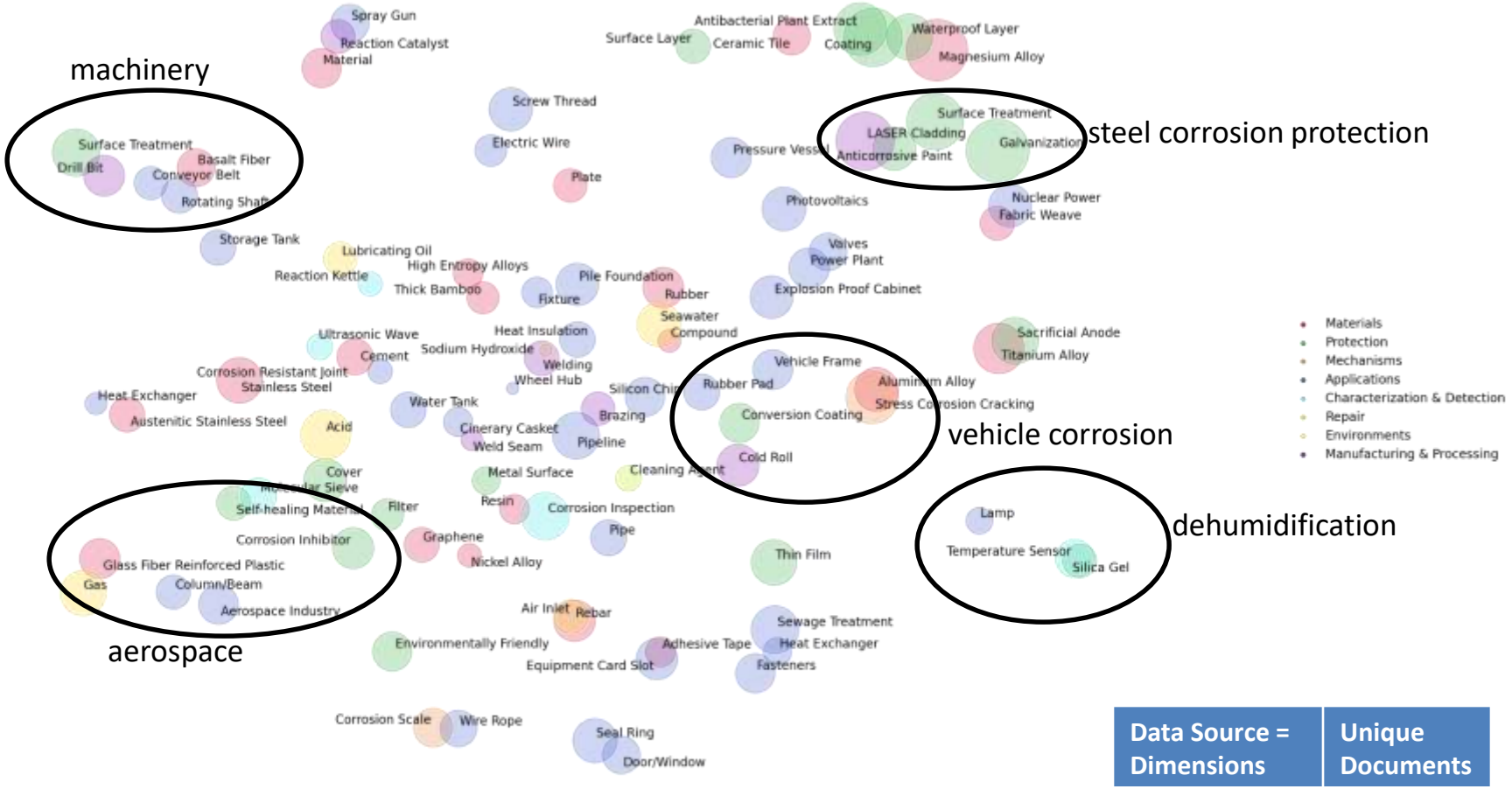
- Publically available sources
- Grants, Patents, and Publications
- Only using Title and Abstract fields from 2017 – 2021
- Filter results to remove unwanted documents
    - From academic areas Human Society, Law & Legal, Creative Arts & Writing, Language, and Philosophy
    - Documents with no abstract were also removed

Other Data Sources: some data sources internal to specific agencies are available and can be added into the search

Retrieving Relevant Documents

- Natural language processing algorithm
- Query is matched to document fields
- Returns a list of documents
- Top topics are automatically assigned based on the language in the document fields
    - Use coherence measure to determine “Junk” topics
    - Analyst review topics for sense making

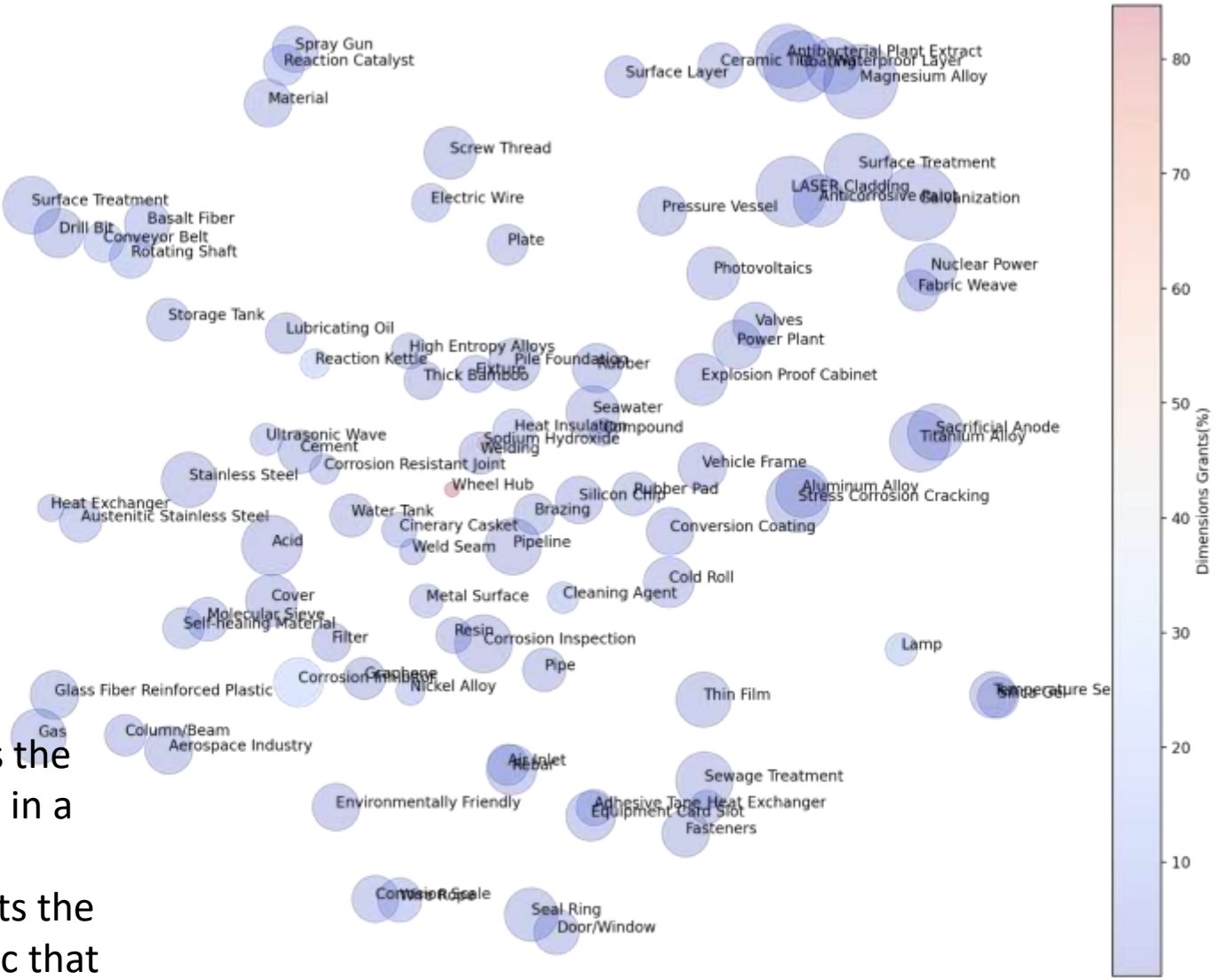
Step 3: Visualize Data



Plot of 95 topics with coordinates determined by term/phrase overlap between topics.
Topic size is proportional to Size. Color is based on Super Label.

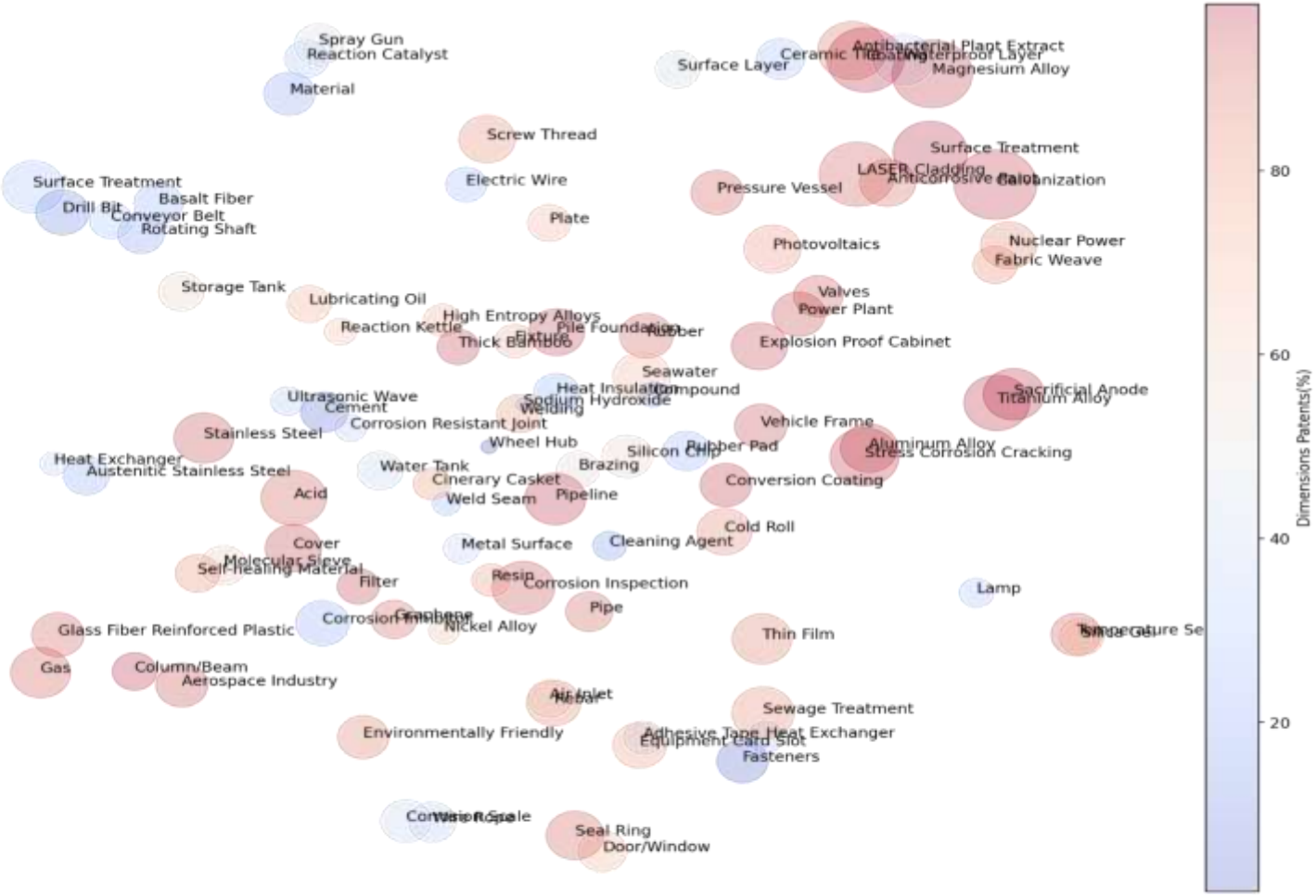
Data Source = Dimensions	Unique Documents
Grants	21079
Patents	239985
Publications	76828

Step 3: Visualize Data



- Bubble size represents the number of documents in a topic
- Bubble color represents the percentage of the topic that are grants

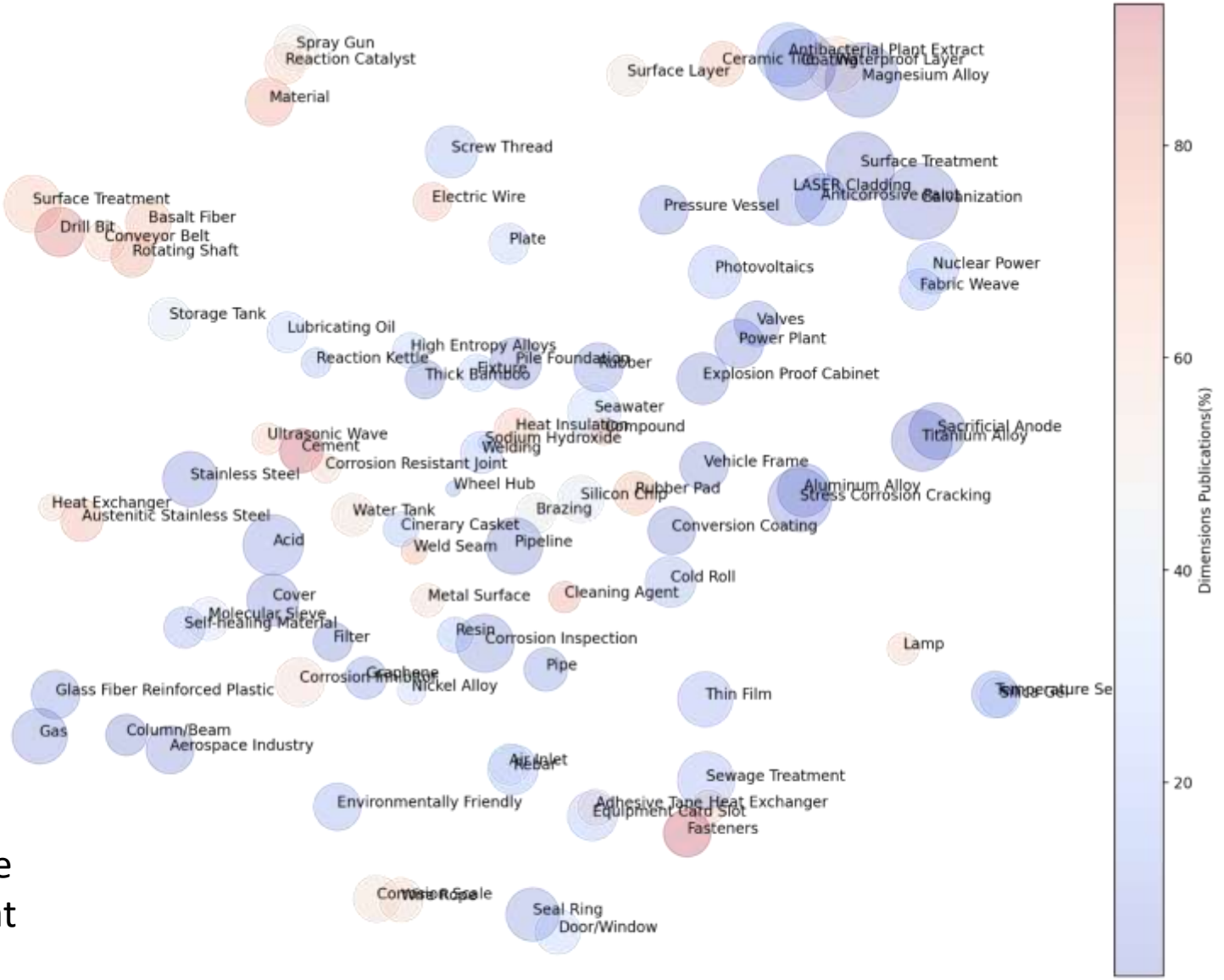
Step 3: Visualize Data



Plot of 95 topics with coordinates determined by term/phrase overlap between topics.
Topic size is proportional to Size. Color is based on Dimensions Patents(%).

- Bubble size represents the number of documents in a topic
- Bubble color represents the percentage of the topic that are patents

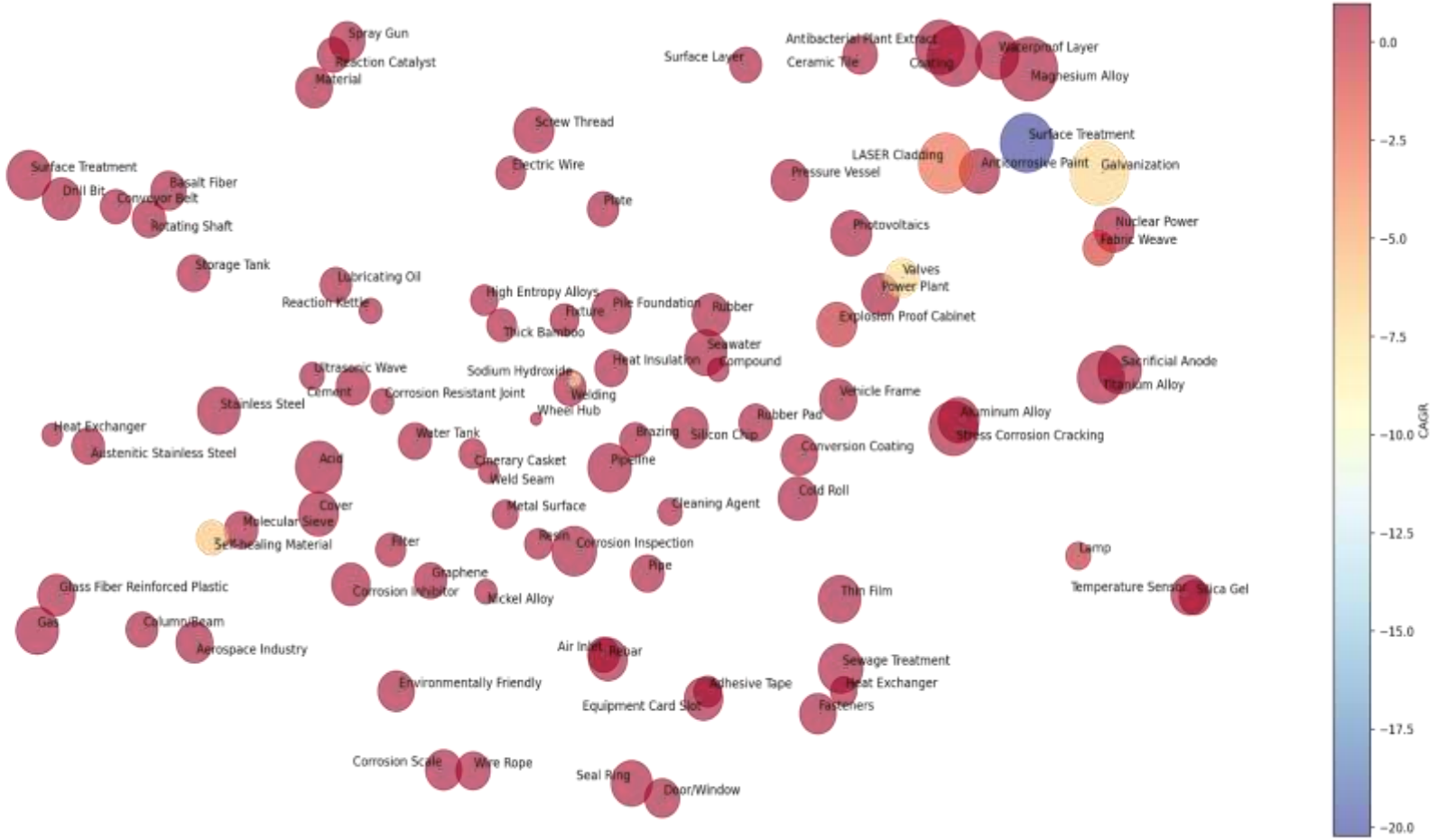
Step 3: Visualize Data



- Bubble size represents the number of documents in a topic
- Bubble color represents the percentage of the topic that are publications

Step 3: Visualize Data

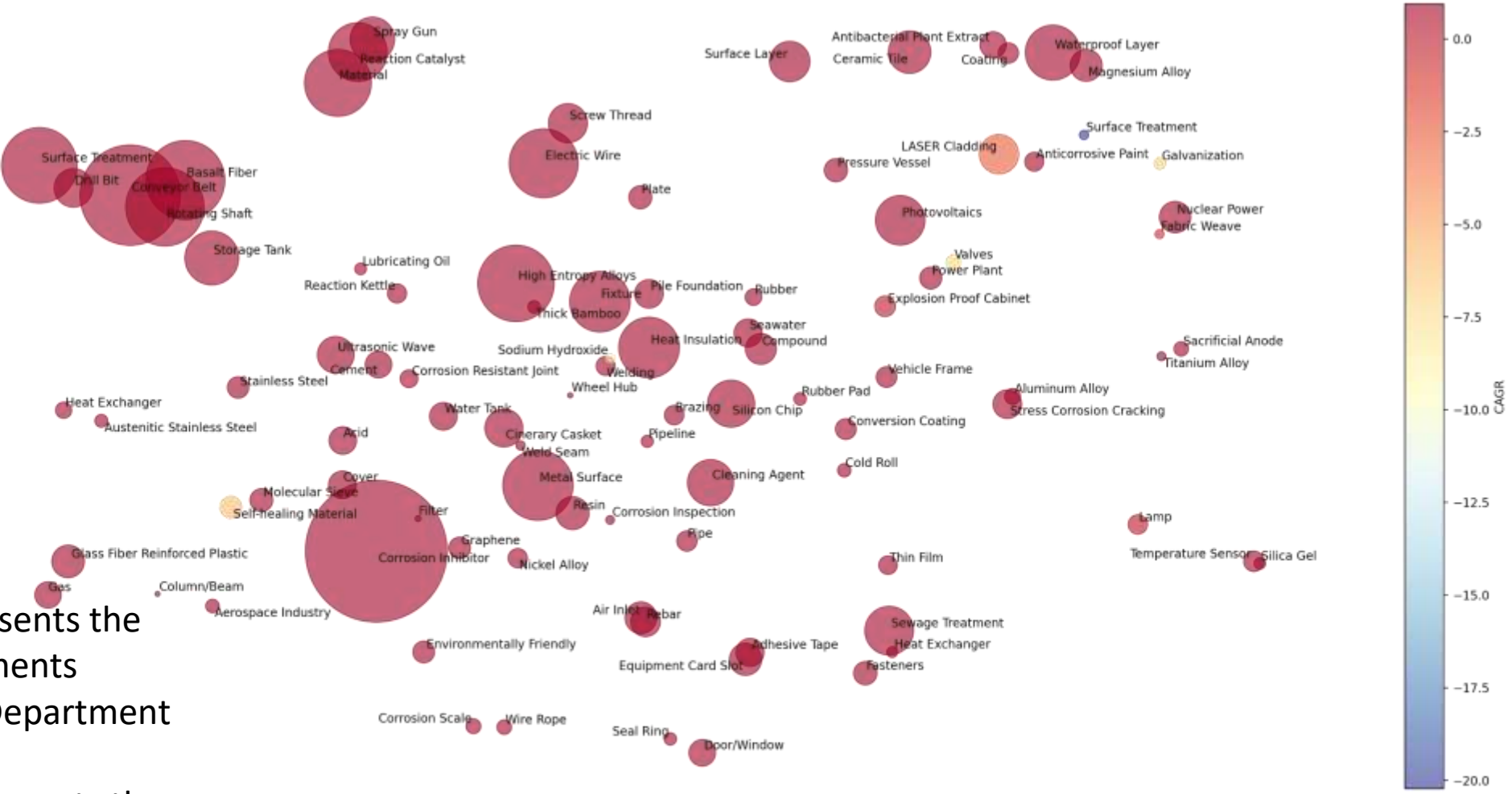
Compound **A**nnual **G**rowth **R**ate



- Bubble size represents the number of documents in a topic
- Bubble color represents the compound annual growth rate of a topic 2017 to 2021

$$\mathrm{CAGR} = \left(\frac{V_{\mathrm{final}}}{V_{\mathrm{begin}}}\right)^{1/t} - 1$$

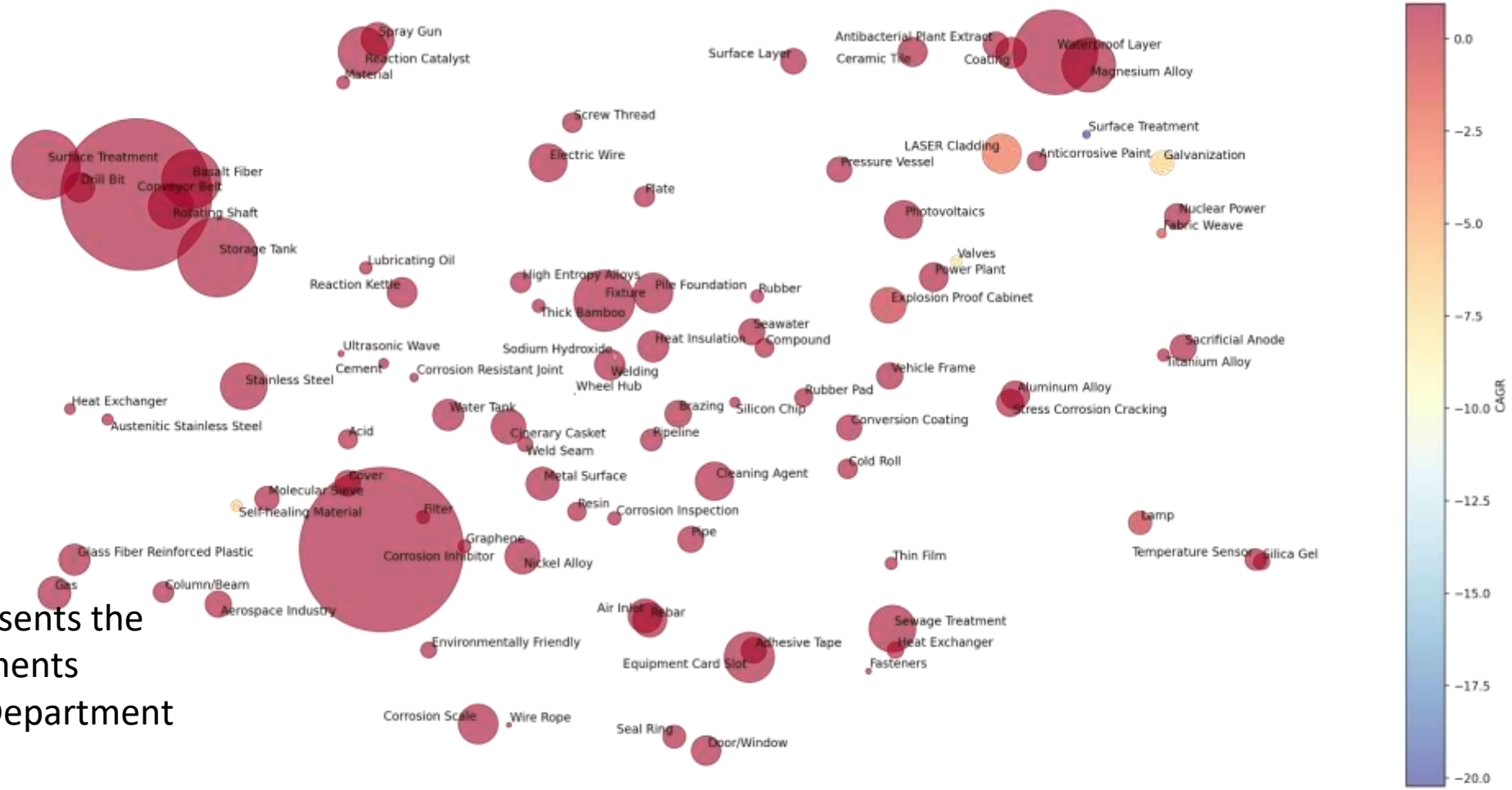
Step 3: Visualize Data



- Bubble size represents the number of documents associated with Department of the Navy
- Bubble color represents the compound annual growth rate of a topic 2017 to 2021

Plot of 95 topics with coordinates determined by term/phrase overlap between topics.
Topic size is proportional to United States Department of the Navy. Color is based on CAGR.

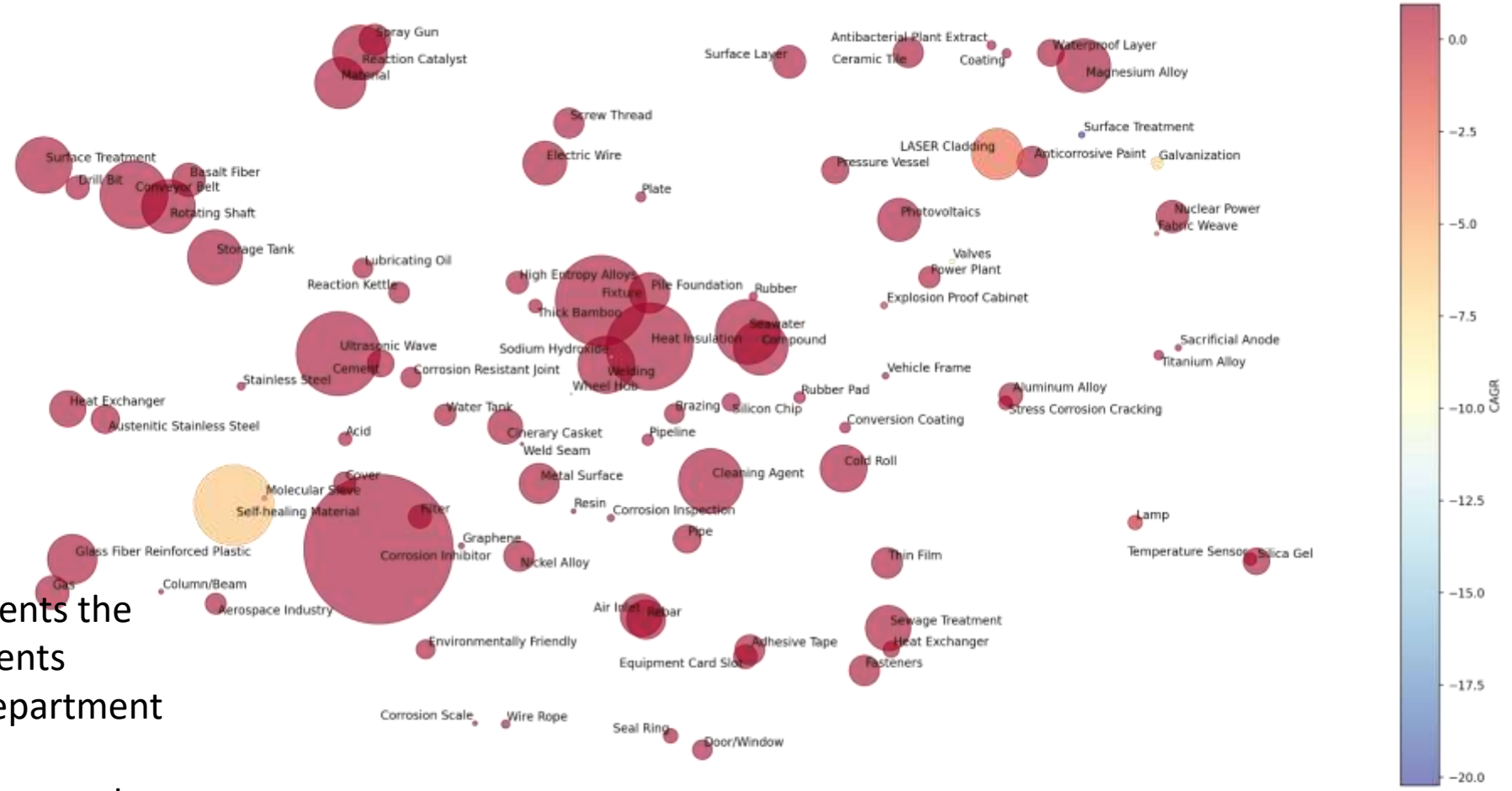
Step 3: Visualize Data



- Bubble size represents the number of documents associated with Department of the Air Force
- Bubble color represents the compound annual growth rate of a topic 2017 to 2021

Plot of 95 topics with coordinates determined by term/phrase overlap between topics.
Topic size is proportional to United States Department of the Air Force. Color is based on CAGR.

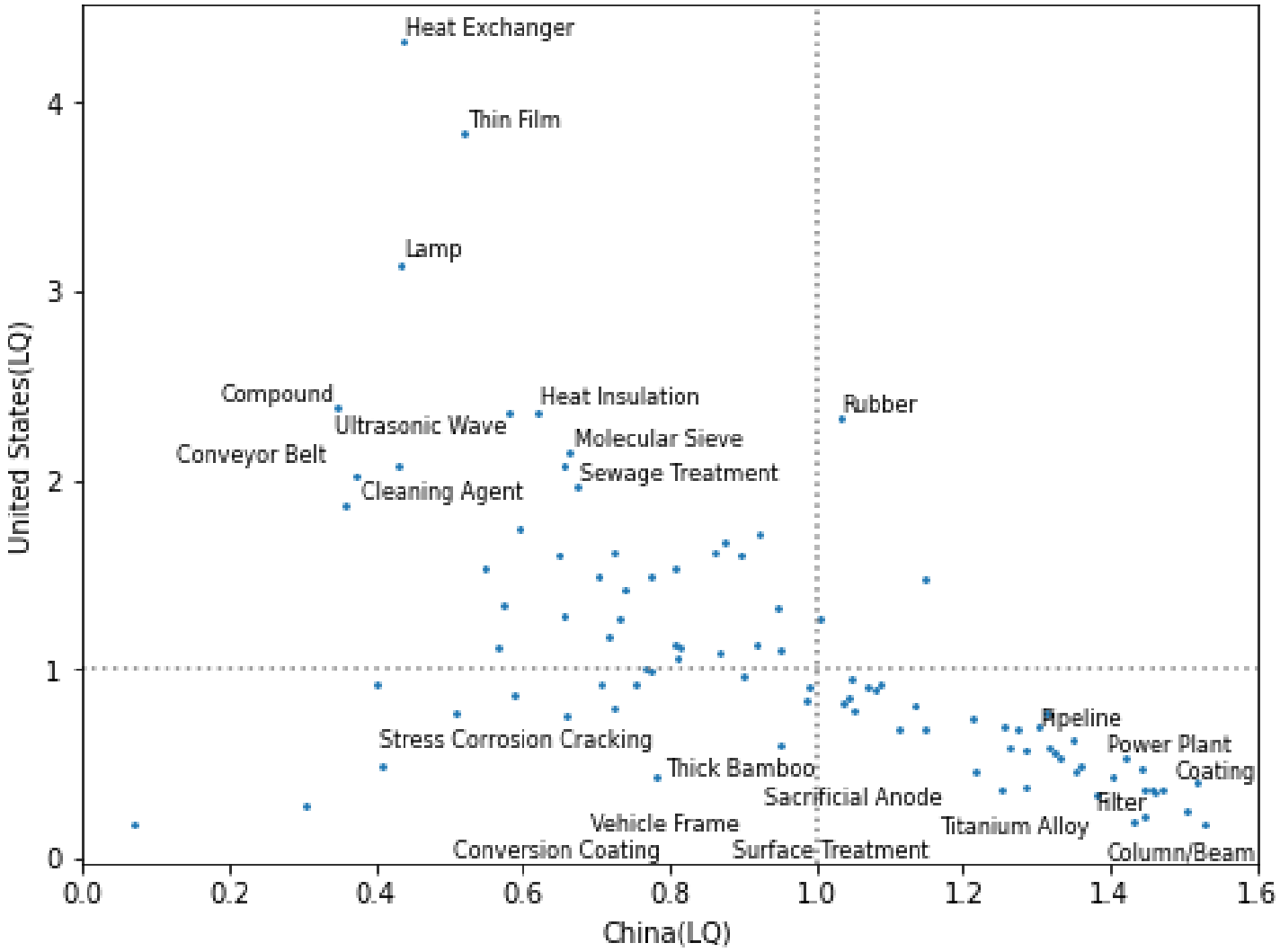
Step 3: Visualize Data



- Bubble size represents the number of documents associated with Department of the Army
- Bubble color represents the compound annual growth rate of a topic 2017 to 2021

Plot of 95 topics with coordinates determined by term/phrase overlap between topics.
Topic size is proportional to United States Department of the Army. Color is based on CAGR.

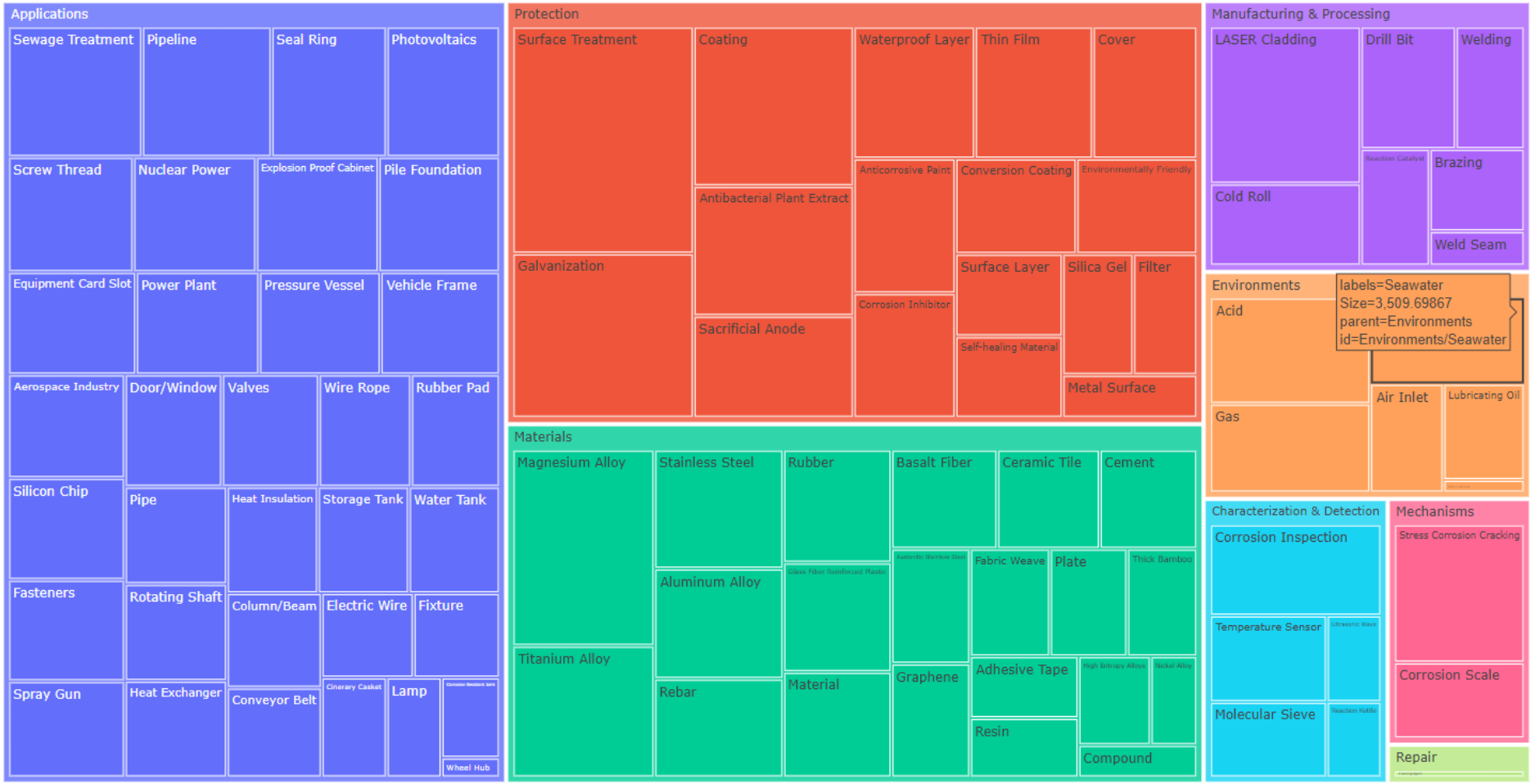
Step 3: Visualize Data



Location **Q**uotient

$$LQ = \frac{X_i / \sum X_i}{N_i / \sum N_i}$$

Step 3: Visualize Data



Step 3: Visualize Data

Top Companies By Patents	Top Universities By Publications	Top Funders By Publications	Top Research Organizations
Sinopec China	University of Science & Technology Beijing China	National Natural Science Foundation China	Institute of Metals Research China
Nippon Steel & Sumitomo Metal Japan	University of Chinese Academy of Sciences China	Ministry of Science & Technology China	Chinese Academy of Sciences China
JFE Holdings Japan	Central South University China	European Commission European Union	Institute for Color Science Iran
Pohang Iron & Steel South Korea	Anna University, Chennai India	China Postdoctoral Foundation China	Institute of Oceanology China
China National Petroleum Corporation China	Tianjin University China	Chinese Academy of Sciences China	Lanzhou Institute of Chemical Physics China

Step 3: Visualize Data

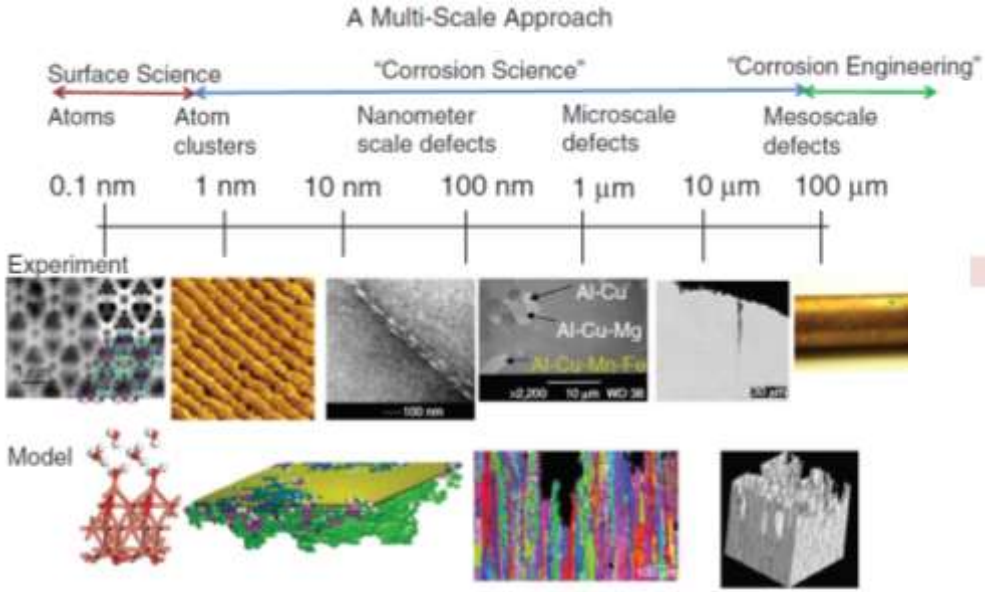
Top US Companies By Patents	Top US Universities By Publications	Top US Funders By Publications	Top US Gov't Research Organizations By Publications
Battelle Ohio	The Ohio State University Ohio	Department of Energy Washington D.C.	Oak Ridge National Laboratory Tennessee
PPG Industries Pennsylvania	University of Virginia Virginia	National Science Foundation Virginia	Pacific Northwest National Laboratory Washington (state)
Honeywell North Carolina	Ohio University Ohio	Office of Naval Research Virginia	Argonne National Laboratory Illinois
Sikorsky Aircraft Corporation Connecticut	University of Michigan Michigan	Battelle Ohio	Sandia National Laboratories New Mexico
IBM New York	University of Akron Ohio	Department of Transportation Washington D.C.	Lawrence Berkley National Laboratory California

Step 4: Analyze Data

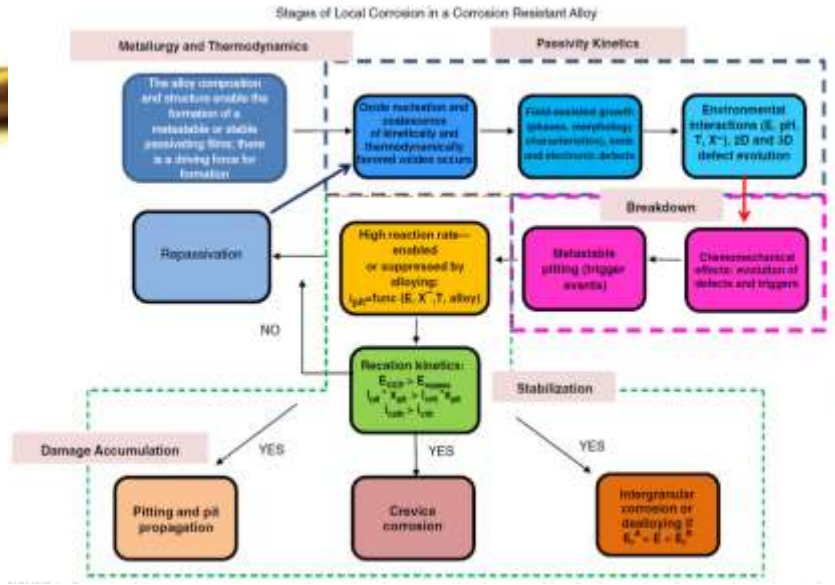
- **What data is used and how it is analyzed depends on the questions being asked of the data.**
  - Given a reduction in resources, we need to understand where to partner and collaborate, rather than lead.
  - What topics/areas are missing from the landscape? What are reasons these might be missing?
  - What areas might need more investment?
- **Compare data to policy documents, forward thinking editorials, etc.**
  - DoD Corrosion Policy Office Roadmaps
  - Reliance21 Community of Interest Roadmaps
  - DoD Modernization Priorities
  - National Research Council/ National Academies
  - Corrosion Journal editorial series on “Future Frontiers of Corrosion Science and Engineering”

Step 4: Analyze Data

Future Frontiers



Scully, J. R. (2018). Future frontiers in corrosion science and engineering, part I. *Corrosion, 74*(1), 3-4. doi:http://dx.doi.org/10.5006/2734

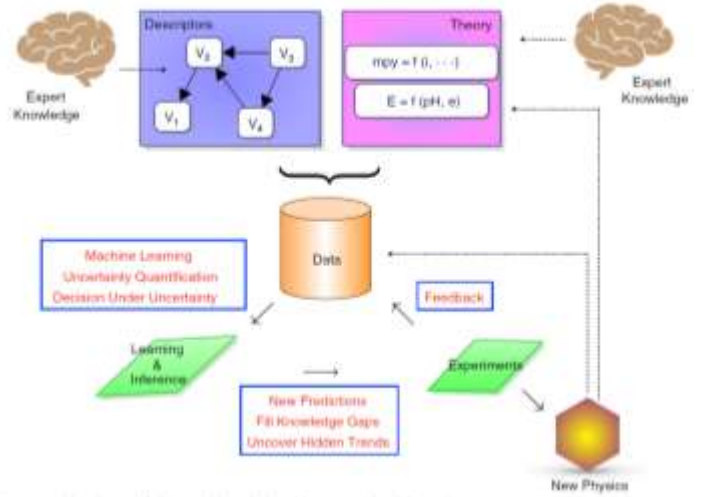


Scully, J. R. (2019). Future frontiers in corrosion science and engineering, part II: Managing the many stages of corrosion. *Corrosion, 75*(2), 123-125. doi:http://dx.doi.org/10.5006/3132

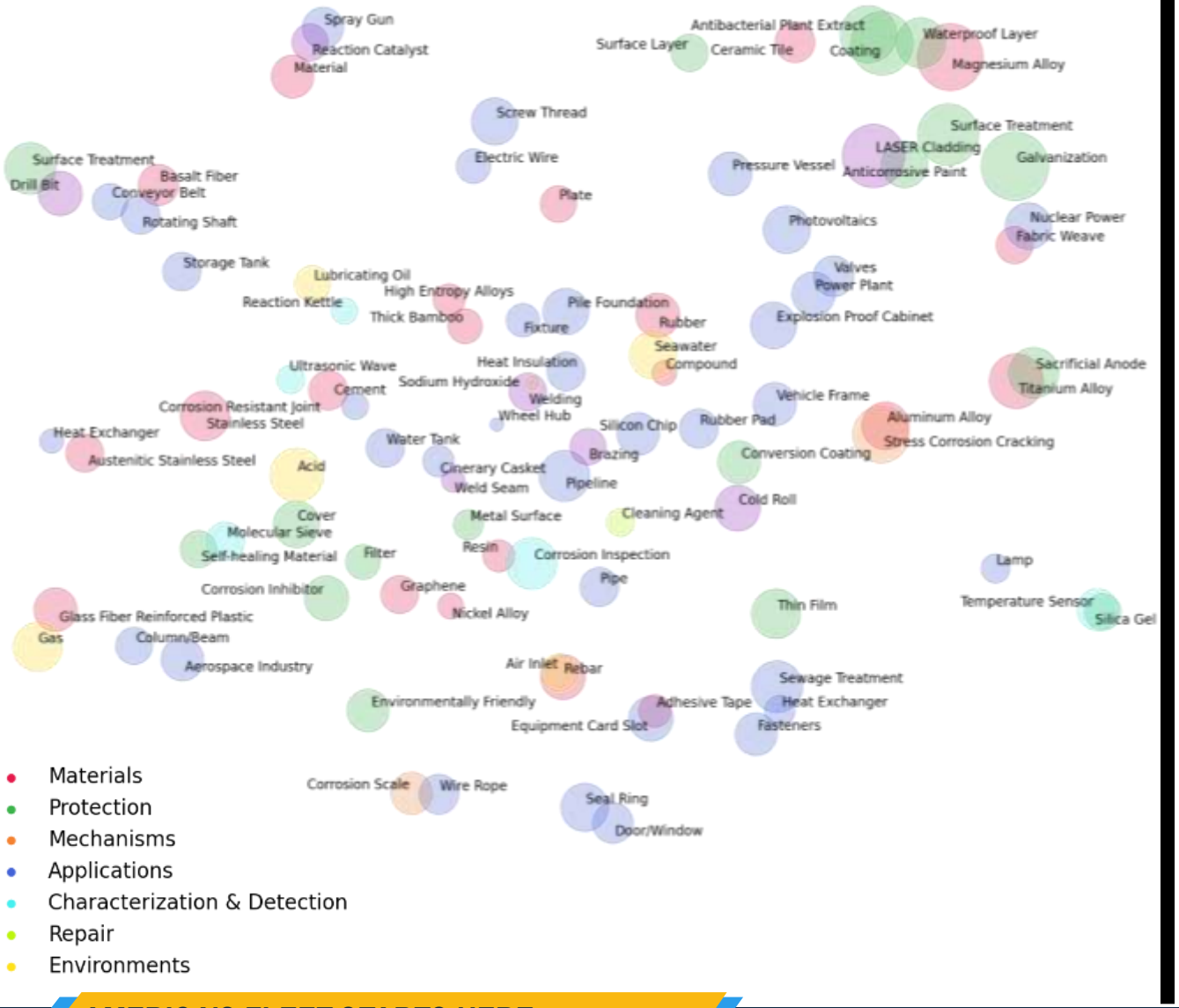
Multi-scale
Multi-physics
Multi-stage
Data analytics

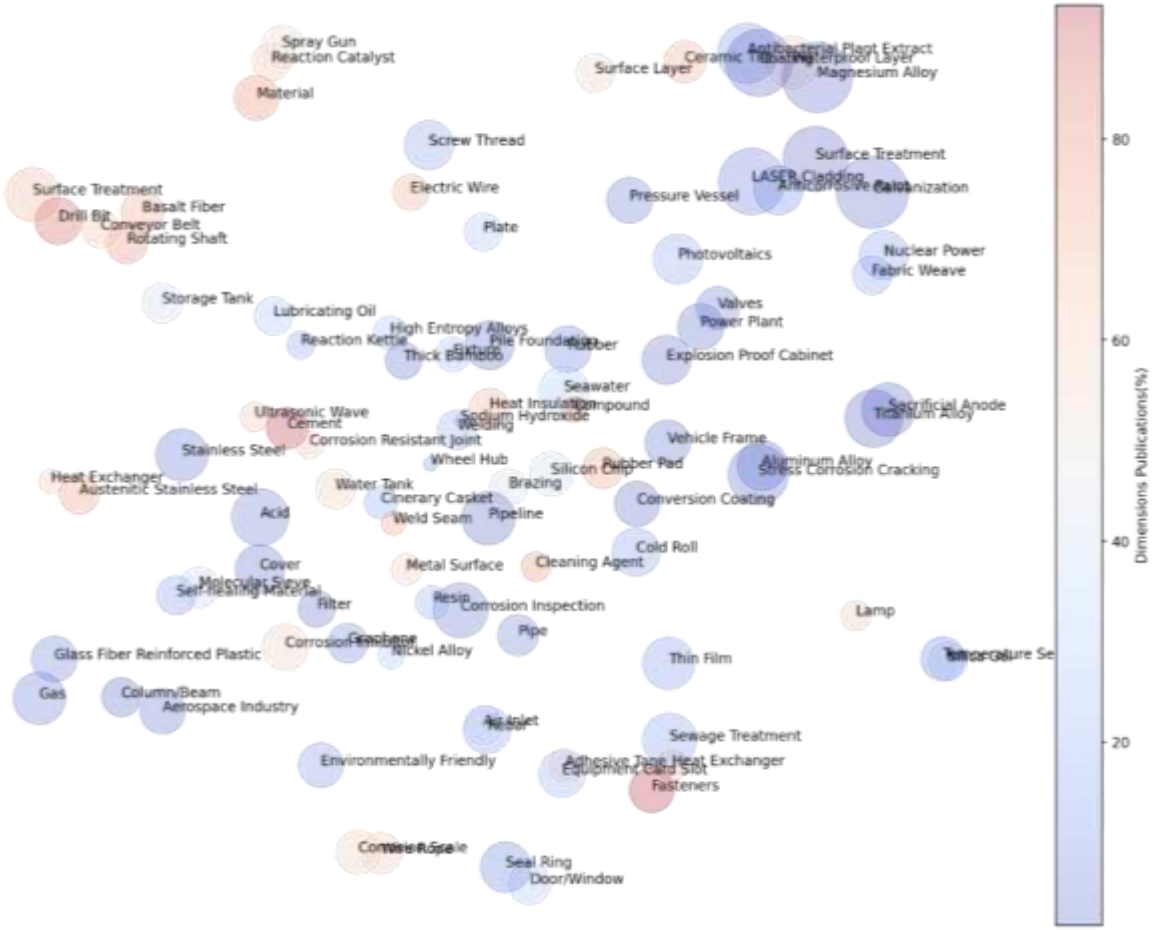
Scully, J. R., & Balachandran, P. V. (2019). Future frontiers in corrosion science and engineering, part III: The next "Leap ahead" in corrosion control may be enabled by data analytics and artificial intelligence. *Corrosion, 75*(12), 1395-1397. doi:http://dx.doi.org/10.5006/3432

Corrosion Informatics



Step 4: Analyze Data





Step 4: Analyze Data

NAVSEA WARFARE CENTERS Carderock Division

Other things to think about:

1. **This analysis doesn't look at quality of publications.**
   - There are techniques that can incorporate impact factors, H-factors, etc. for the publications and authors.
2. **The analysis is limited by the data sources.**
   - Non-public data sources can be added (of course those are not presented here).
3. **This is a bibliometric heavy analysis.**
   - What about areas where reports, articles, etc. might not be available?
4. **Other taxonomies can be assigned.**
   - Automation of reporting requirements

Conclusions

- Landscape analyses can provide insight into the past and current state of a technology area such as corrosion.
- The tools available are very flexible and can be tailored to the questions being asked.
- Landscape analysis does not provide technology forecasting or backcasting, but can put results from these activities in context.
- Using landscape analysis with subject matter expert review, can be helpful for understanding where to lead, follow, or partner.