SUMMER PRESENTATION

Office of the Chief AI Officer
Department of Health & Human Services
Sanja Basaric — AI Program Lead

coding it forward >

SUHAN KACHOLIA
University of Georgia
BA/MS Artificial Intelligence

PROJECT #1

Analyzing, **visualizing**, and **describing** the HHS Artificial Intelligence Use Case Inventory

cif>

PROBLEM STATEMENT

- Federal government is using artificial intelligence (AI) more than ever.
  - In HHS alone, there were **163 projects** involving AI this year.
- Federal departments, like HHS, are required to maintain and publicly upload a list of their AI projects.
- We wanted to analyze and understand this list to **guide future HHS AI strategy**

cif>

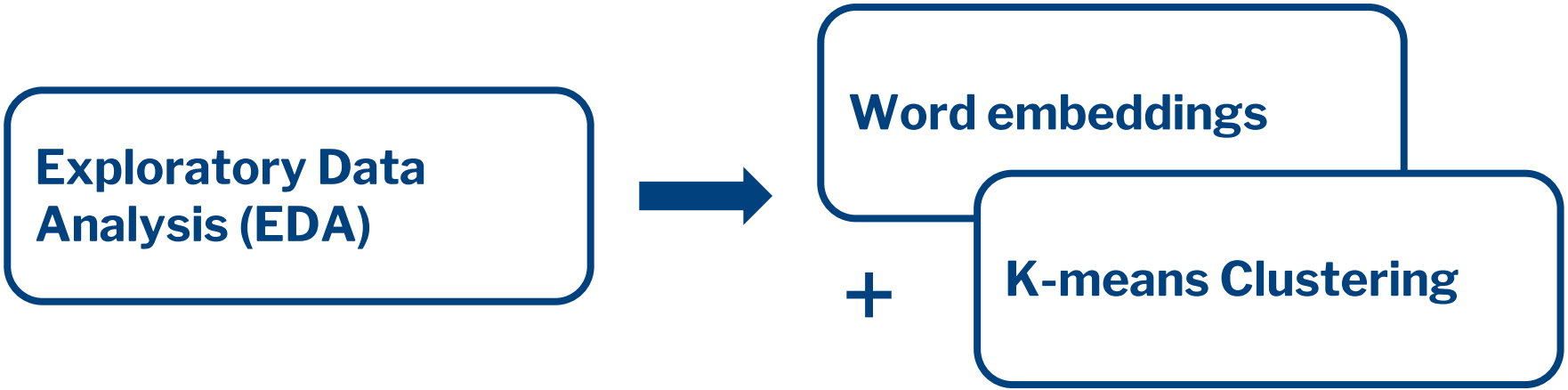
DATASET

- Two datasets:
  - 2022 Use Cases (**52 projects**) and 2023 Use Cases (**163 projects**)
  - Datasets **were not consistent** with columns and answer requirements

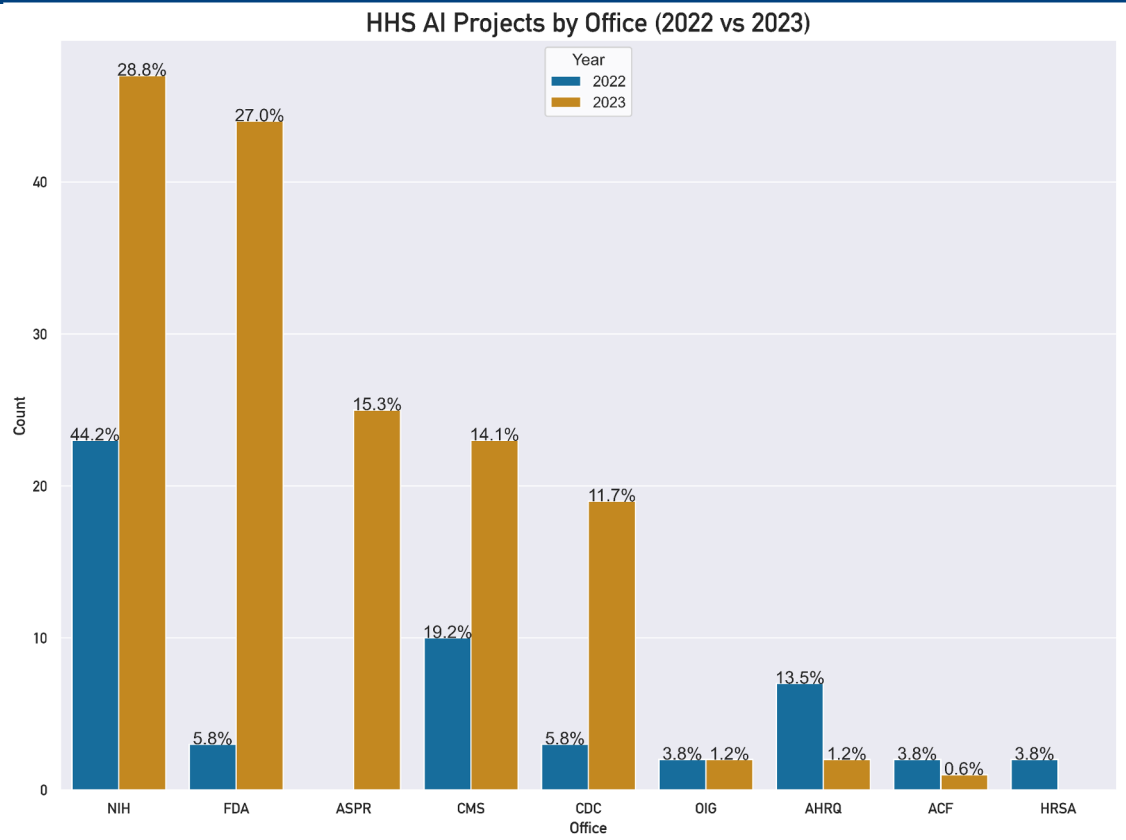
- Some columns:
  - Name, agency, description, AI techniques, stage, data/code access

cif>

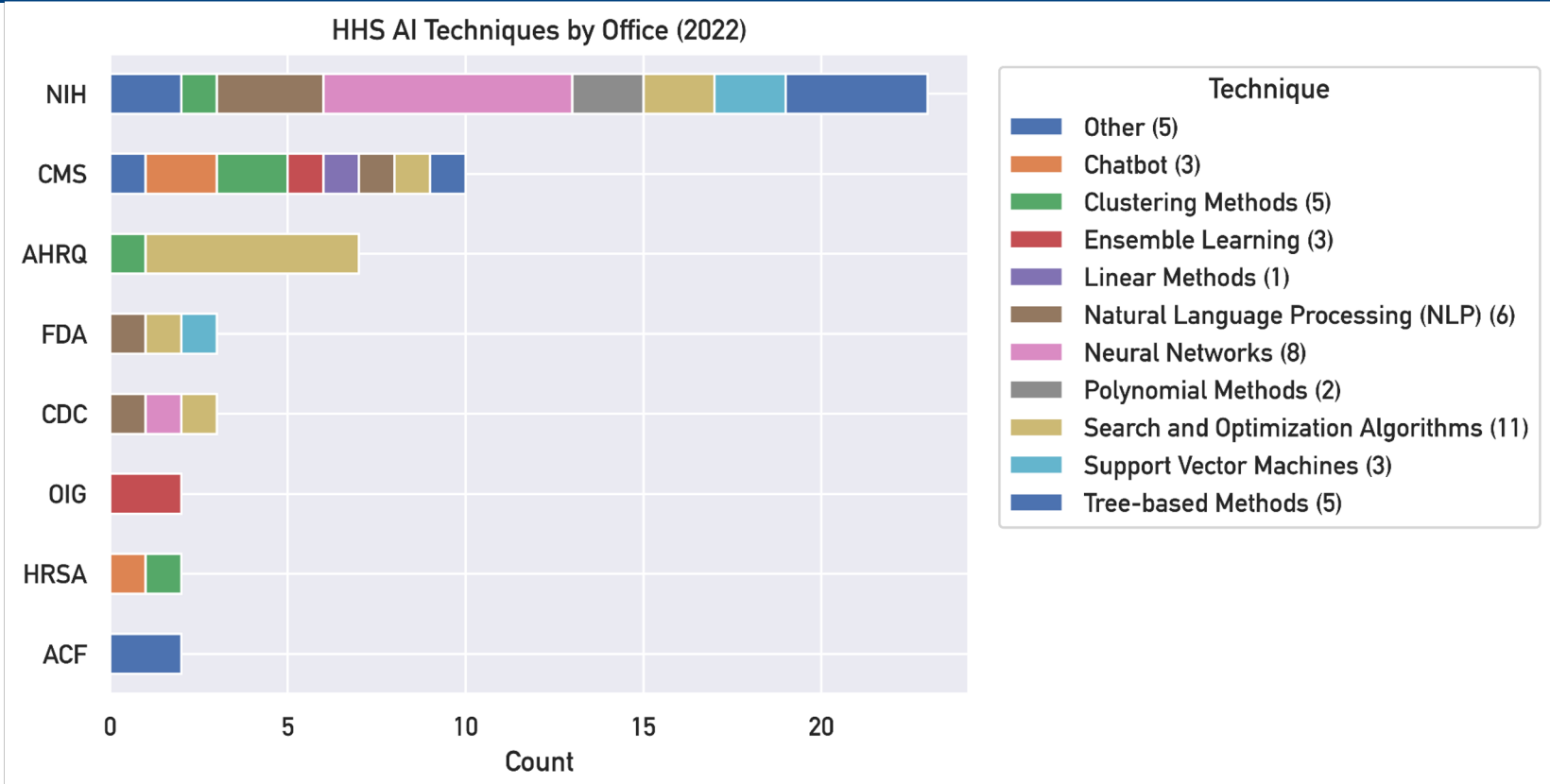
PROCESS


Exploratory Data Analysis (EDA)
Word embeddings
+
K-means Clustering

PROJECTS BY OFFICE

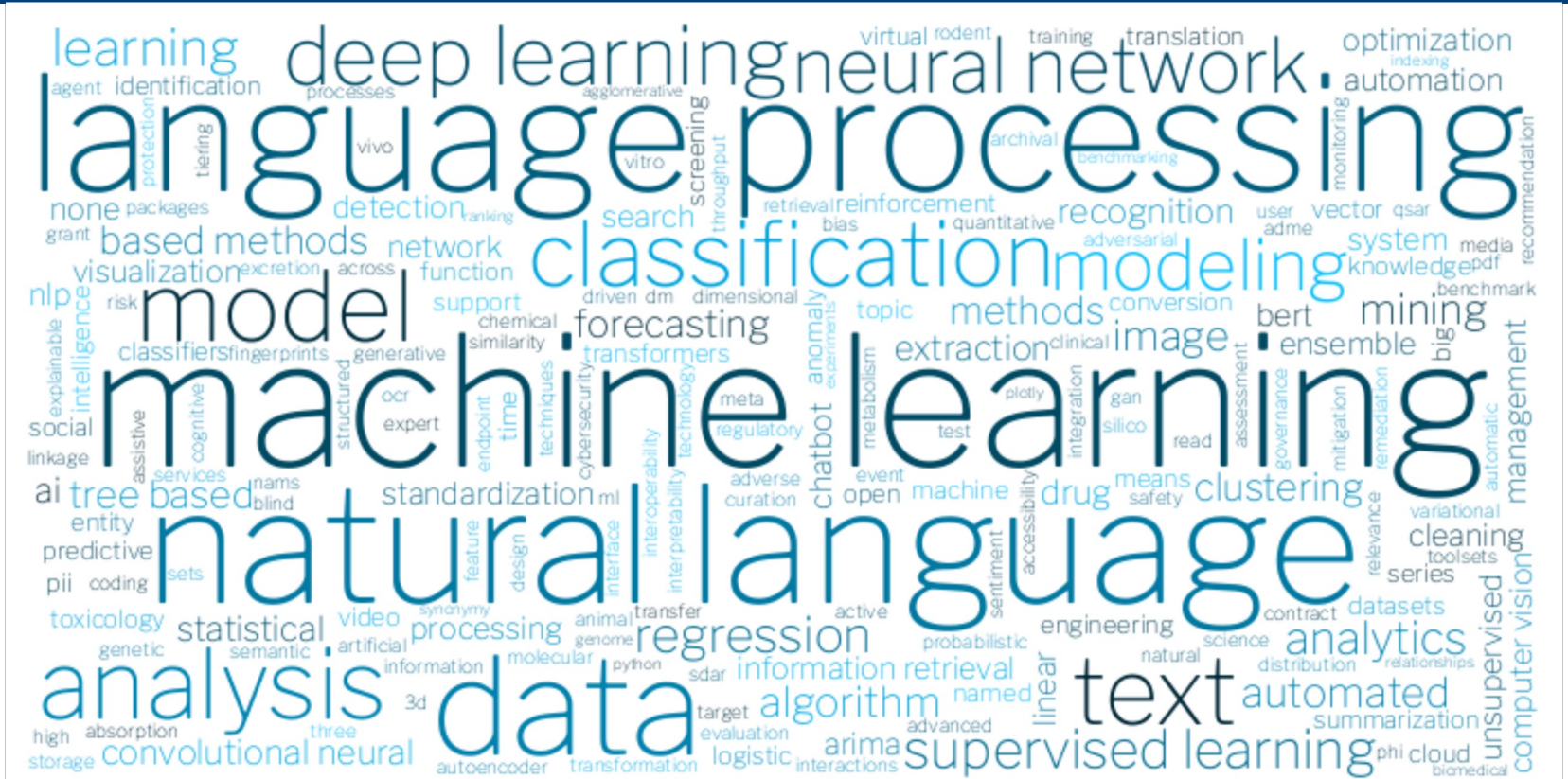


AI TECHNIQUES (2022)

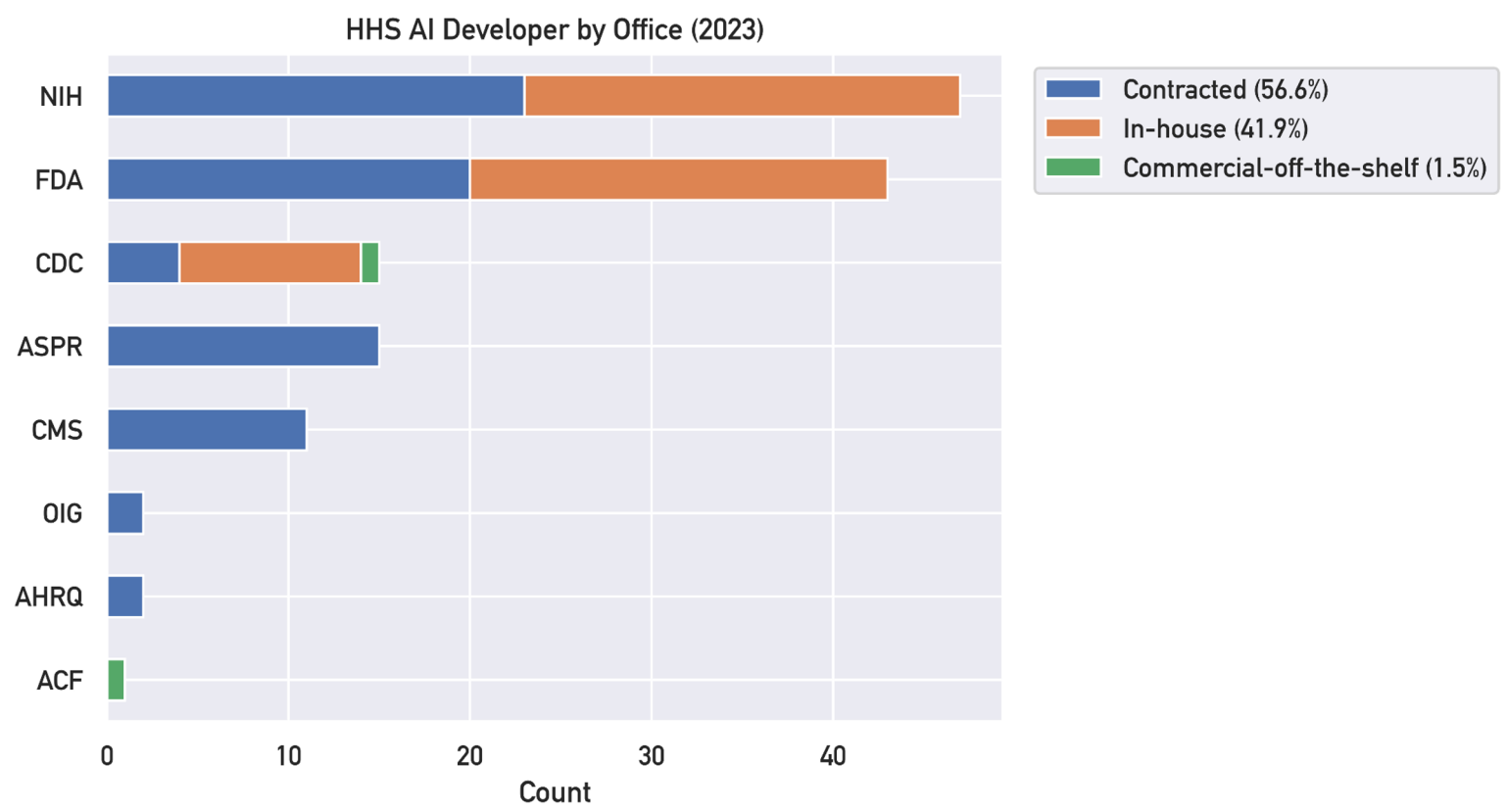


cif>

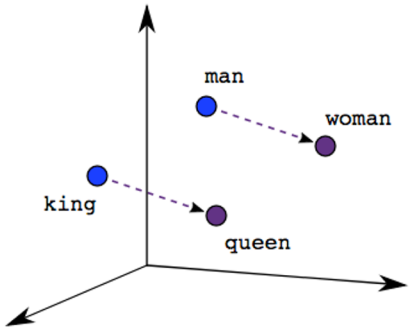
AI TECHNIQUES (2023)



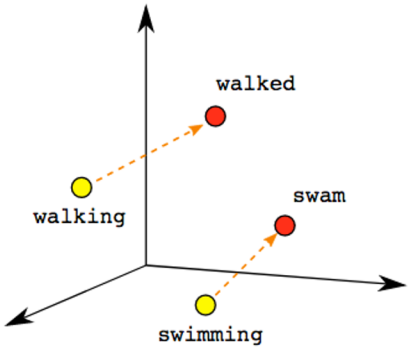
AI DEVELOPER (2023)



EMBEDDINGS



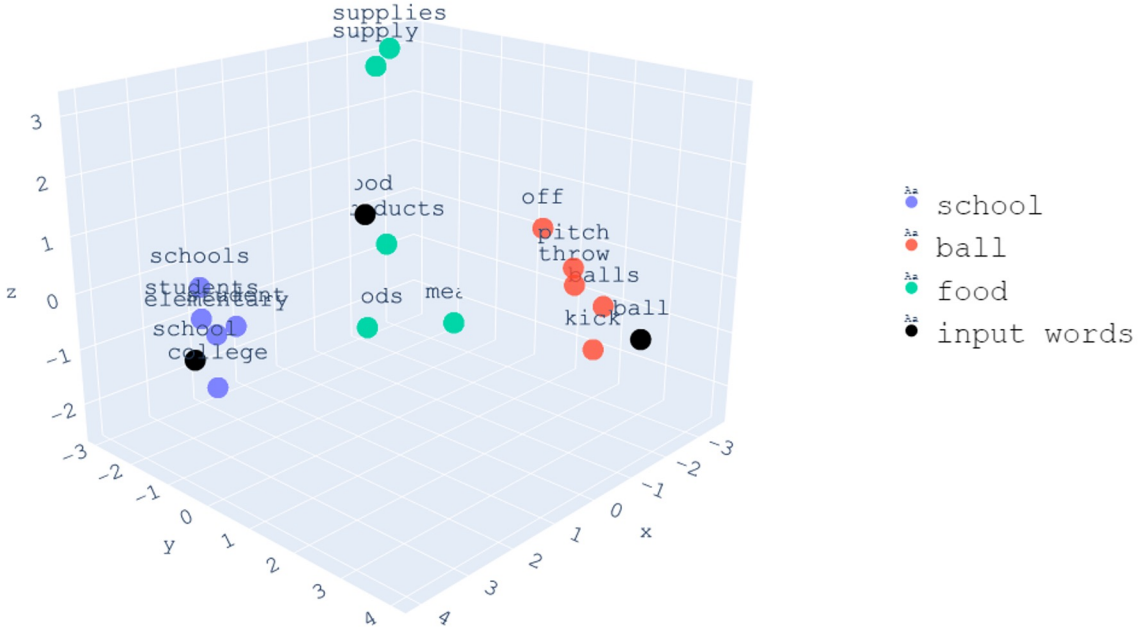
Male-Female



Verb Tense

cif>

T-SNE CLUSTERING



CLUSTERING

- **Information Management** (Keywords: grants, indexing, information, document)
- **Biology Research** (Keywords: public health, drugs, study)
- **Chatbots and NLP** (Keywords: chatbot, NLP, public interface)
- **Detection and Devices** (Keywords: detection, hardware, diagnosis)

PROJECT #2

- After this project, I worked with **Keith Bocian** (HHS OIG) to extend the clustering analysis across **all federal government AI projects**
  - **The Responsible AI Officials Council** will connect project clusters in fall

cif>

METHODOLOGY

- Retrieve and clean HHS 2023 Use Case Inventory
  - Convert “Name” and “Summary” columns to embeddings
  - Used **InstructorXL** model
- Compute cosine similarity matrix (exclude connections between same bureau)

	Information	Artificial	Information	Opioid Data	Community	C
Information		0.69993147	0.715839	0.67373579	0.69291709	
Artificial	0.69993147			0.7468066	0.70740262	
Information	0.715839			0.72019546	0.70074914	
Opioid Data	0.67373579	0.7468066	0.72019546		0.82539775	
Community	0.69291709	0.70740262	0.70074914	0.82539775		
Centers of	0.71093391	0.8075083	0.74174906	0.7474192	0.73140487	

cif>

METHODOLOGY

- Add up similarity scores for each row — this gives us a measure of how well-connected, or central, a particular project is to others
- Keep only top ~15% of those connections, to capture the most important projects
- Train a **weighted K-means** model, using cumulative similarity scores as the weight
- Apply that model to all project embeddings, and receive cluster assignments

cif>

RESULTS

	grant_application_chatbot	drug_opioid_covid	\
0	(text analytics, 1.3)	(drug repurposing, 2.559)	
1	(forecasts campus, 1.147)	(drug labeling, 2.236)	
2	(analytics portal, 1.108)	(identify drug, 2.061)	
3	(grant applications, 1.021)	(repurposing candidates, 1.858)	
4	(explorers program, 1.012)	(opioid use, 1.416)	
5	(language processing, 0.919)	(antiviral discovery, 1.345)	
6	(natural language, 0.916)	(coronavirus antiviral, 1.341)	
7	(forecaster project, 0.909)	(antiviral screening, 1.279)	
8	(intelligence explorers, 0.888)	(subcontractor coronavirus, 1.273)	
9	(tool sstat, 0.748)	(machine learning, 1.243)	
10	(sstat used, 0.739)	(influenza wearable, 1.221)	
11	(smartfind chatbots, 0.73)	(cough vocalization, 1.196)	
12	(calculator jac, 0.727)	(influenza atomwise, 1.141)	
13	(irm initiative, 0.724)	(covid influenza, 1.118)	
14	(jac jit, 0.724)	(vocalization fcv, 1.116)	

	sidewalk_anomaly_priority	pubmed_citation_indexing
0	(priority score, 1.566)	(disambiguation pubmed, 1.253)
1	(fraud prevention, 1.517)	(indexing pubmed, 1.124)
2	(anomaly detection, 1.363)	(citation indexing, 1.046)
3	(ehrs linked, 0.918)	(literature retrieval, 0.984)
4	(burn blast, 0.763)	(table extraction, 0.832)
5	(suicide trends, 0.721)	(extraction tables, 0.816)
6	(splunk system, 0.693)	(disambiguation authors, 0.702)
7	(evaluation ehr, 0.676)	(singlecite automated, 0.666)
8	(severity burn, 0.672)	(metamap widely, 0.661)
9	(endpoints cyberthreats, 0.672)	(citation singlecite, 0.653)
10	(cost anomaly, 0.666)	(citation search, 0.644)
11	(identifiable ehr, 0.658)	(mining citations, 0.644)
12	(nowcasting suicide, 0.654)	(indexing medline, 0.637)
13	(unstructured ehr, 0.653)	(uses metamap, 0.633)
14	(cciio enrollment, 0.653)	(metamap uses, 0.63)

Used KeyBERT to find representative keyphrases of each cluster

Cluster 1: Grant, application, chatbot

Cluster 2: Drug, opioid, COVID

Cluster 3: Sidewalk, anomaly, priority

Cluster 4: PubMed, citation, indexing

cif>

OTHER PROJECTS

- Gave presentation and demo on **OpenAI Code Interterpreter** to Sanja and Greg, the CDO team, the HHS AI CoP, and the ACT-IAC AI Working Group
- Helped Sanja with processing and formatting the 2023 Use Case Inventory
- Helped Anthony manage and update tags for Healthdata.gov using AI

LESSONS

- Bureaucracy is there for a reason, and sometimes adapting tech takes time because of valid concerns
- Learned about both the potential and limits of current AI tools
- Communication and people skills are just as important (if not more) than getting technology right
- Collaboration within government seems tricky and difficult to accomplish but very worth it

cif>

Thank you!

- Thank you to **Sanja Baseric** and **Greg Singleton** in the OCAIO office
- Thank you to **Kathleen Carroll, Kristen Honey, Anthony Caponiti, Ghelatia Araia,** and everyone else in the OCTO/OASH/OCIO/OCDO offices!
- Shout out to the other Fellows, **Hayley**, **Vivian**, and **Ryan**
- Thank you to Coding it Forward for making this internship possible!

cif>