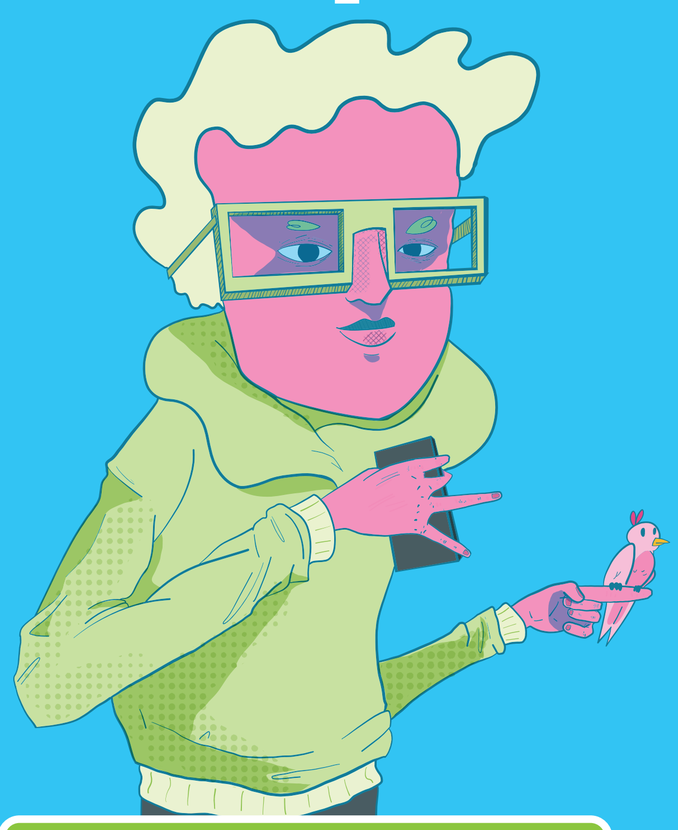
Improving Research protocols through natural language processing



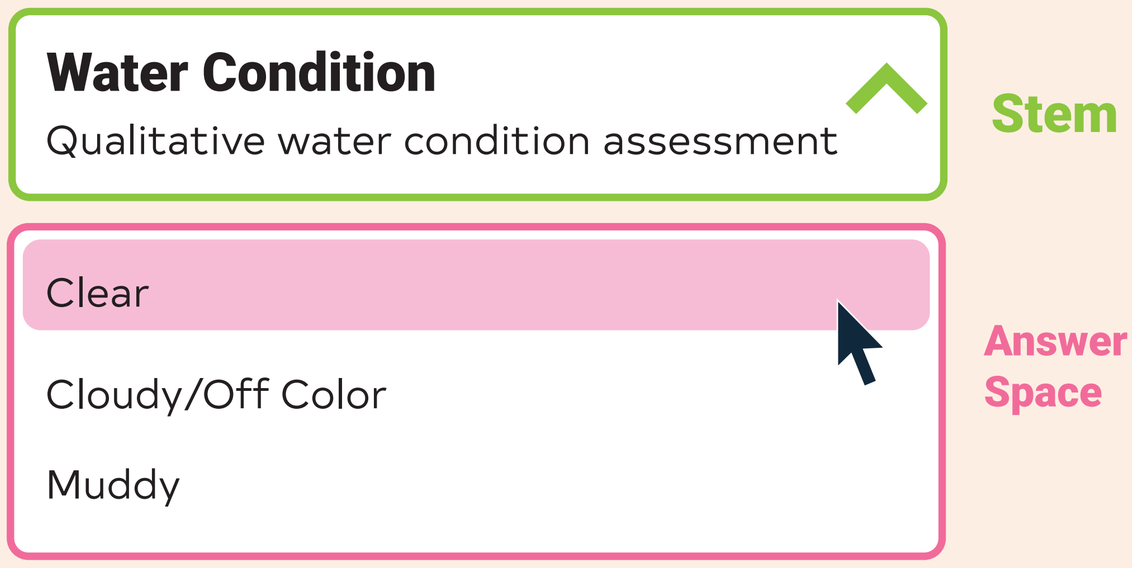
BRANDON BUDNICKI
CitSci Lead Software Developer

Data collection is the defining aspect of observational based research. Protocol designers are aided by knowledge of survey theory best practices. This project will identify the common issues in question design and identify projects asking similar questions. This will be done through classification and clustering the 13,861 questions created by projects hosted on CitSci.

Brandon Budnicki*, Hannah Phillips**, Greg Newman*, Stacy Lynn*, and Sarah Newman*
*CitSci.org | Colorado State University, **unaffiliated

ANATOMY OF A QUESTION

- The question Stem is the words which make up the question and any additional instructions
- The Answer space defines and limits possible answers



GUIDELINES FOR FORMING QUESTION STEMS

To effectively form questions, focus on clarity, precision, and purpose. Keep questions simple and direct, avoiding jargon and negatives.

Example Questions

Miles Traveled

Do you have any comments or issues?

Was water not present?

What is the water temperature?

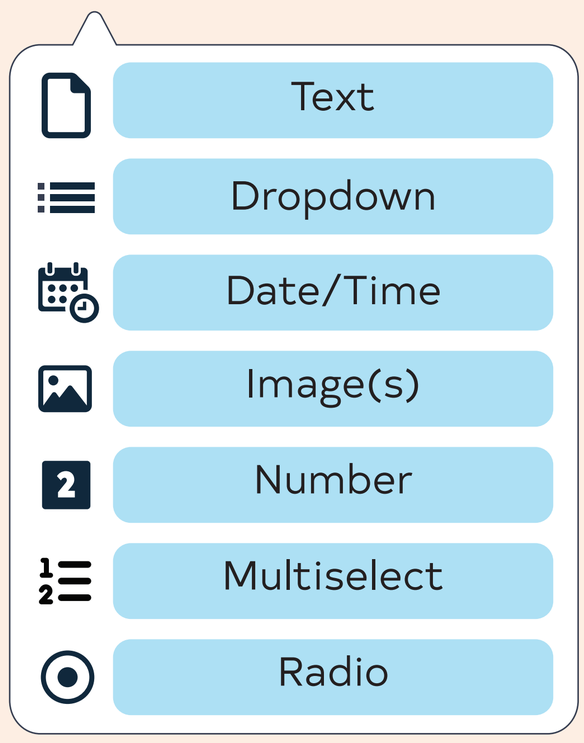
Most volunteers have feedback. Do you have feedback?

Question Stem Guideline

- Use complete sentences with simple sentence structrue.
- Avoid double barreled questions.
- Avoid double negatives.
- Include measurement units.
- Avoid loaded questions

Answer Space Supported by CitSci

Survey questions can be closed or open-ended. Closed questions offer predefined answers and are easier to analyze, while open-ended ones allow detailed responses but require more complex coding. The classification of a question about organisms can fall into either category, depending on whether participants are limited to selecting from a specific list or permitted to name any organism freely.



POSTER 139
Designed by Marvin Andino

ECOSYSTEM SCIENCE AND SUSTAINABILITY
COLORADO STATE UNIVERSITY

CitSci

2025 Conference for Advancing Participatory Sciences
Portland, OR

Methods

1 Source questions from CitSci

2,121 PROTOCOLS

Questions 13,861

Projects 1,417

2 Code question based on guidelines

3 Using SBERT TO GENERATE PARAGRAPH VECTOR FROM PRETRAINED LANGUAGE MODELS

- Miles Traveled
- Do you have any comments or issues?
- Was water not present?
- Most volunteers have feedback, Do you have feedback?

Question following STEM Guideline

TOKENIZATION

¨ What¨, ¨is¨, ¨the¨, ¨water¨, ¨temperature¨

VECTOR PARAGRAPH

[0.7, 0.6, 0.5]

[0.6, 0.1, 0.9]

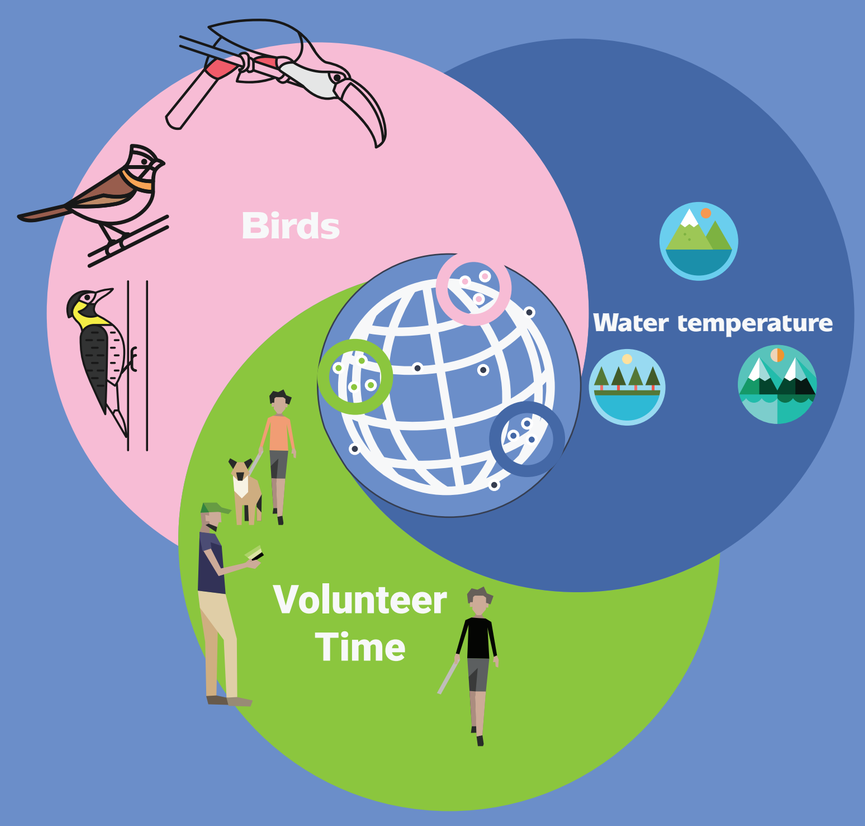
[0.1, 0.6, 0.2]

[0.2, 0.7, 0.8]

SBERT ENCODER

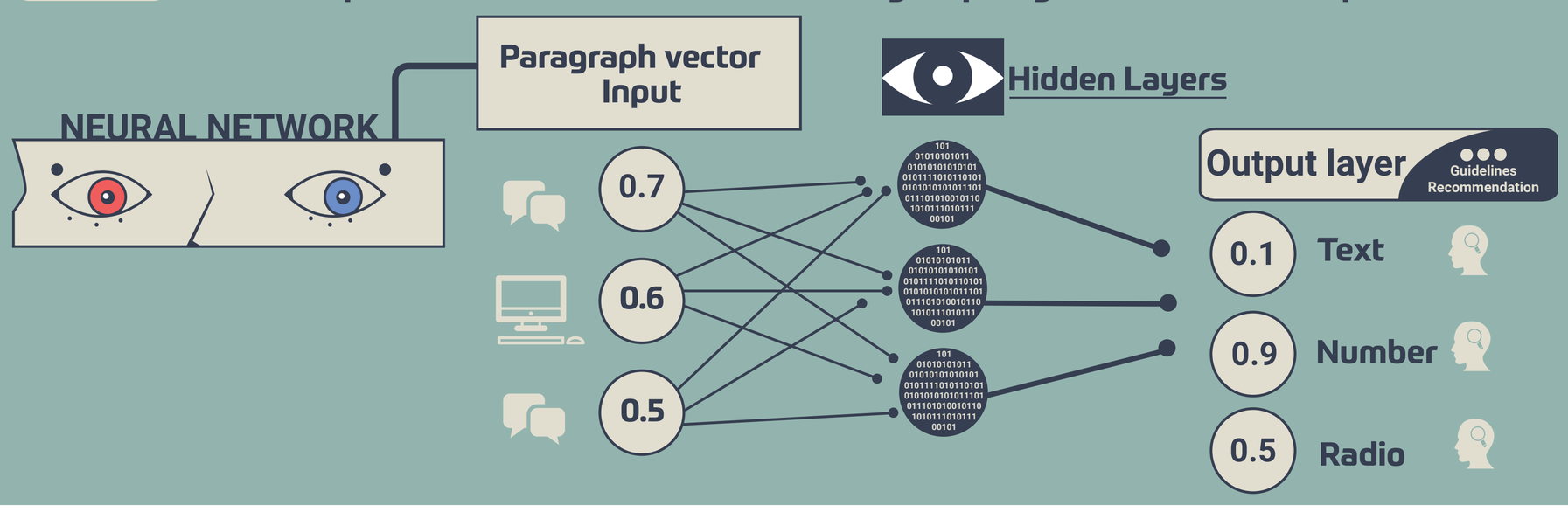
5 Cluster similar questions

HDBSCAN and BERTopic group questions by clustering their paragraph vectors in N-dimensional space, enabling unsupervised, data-driven categorization without manual input.



4 Classify questions

Develop a neural network that can identify helpful guidelines for new questions



Classification Preliminary results

	datetime	description	dropdown	image	number	organism	radio	text	text area
datetime	81.0	0.0	4.8	0.0	0.0	0.0	9.5	4.8	0.0
description	0.0	70.6	0.0	5.9	17.6	5.9	0.0	0.0	0.0
dropdown	0.0	0.6	87.3	0.6	3.8	0.0	2.5	5.1	0.0
image	0.0	2.2	2.2	93.3	0.0	2.2	0.0	0.0	0.0
number	0.5	0.0	3.6	0.5	89.1	2.3	0.9	3.2	0.0
organism	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0
radio	0.0	0.0	16.3	0.0	4.7	4.7	74.4	0.0	0.0
text	1.1	2.2	15.6	0.0	14.4	1.1	2.2	61.1	2.2
text area	0.0	0.0	22.2	16.7	5.6	0.0	11.1	5.6	38.9

Confusion Matrix - Percent Correct, Actual class in rows, Predicted class in columns

We analyzed how often the model's predicted data type aligned with the data type selected by project managers based on the question text.

In 15% of cases, there was a discrepancy between the two. Notably, 46.1% of these discrepancies could have been improved by offering a suggested data type to the project manager. To enhance alignment, we plan to implement two improvements: (1) panelist coding and (2) introducing a confidence threshold for generating recommendations.

Connect with Us!

https://citsci.org/connect

Scan me!