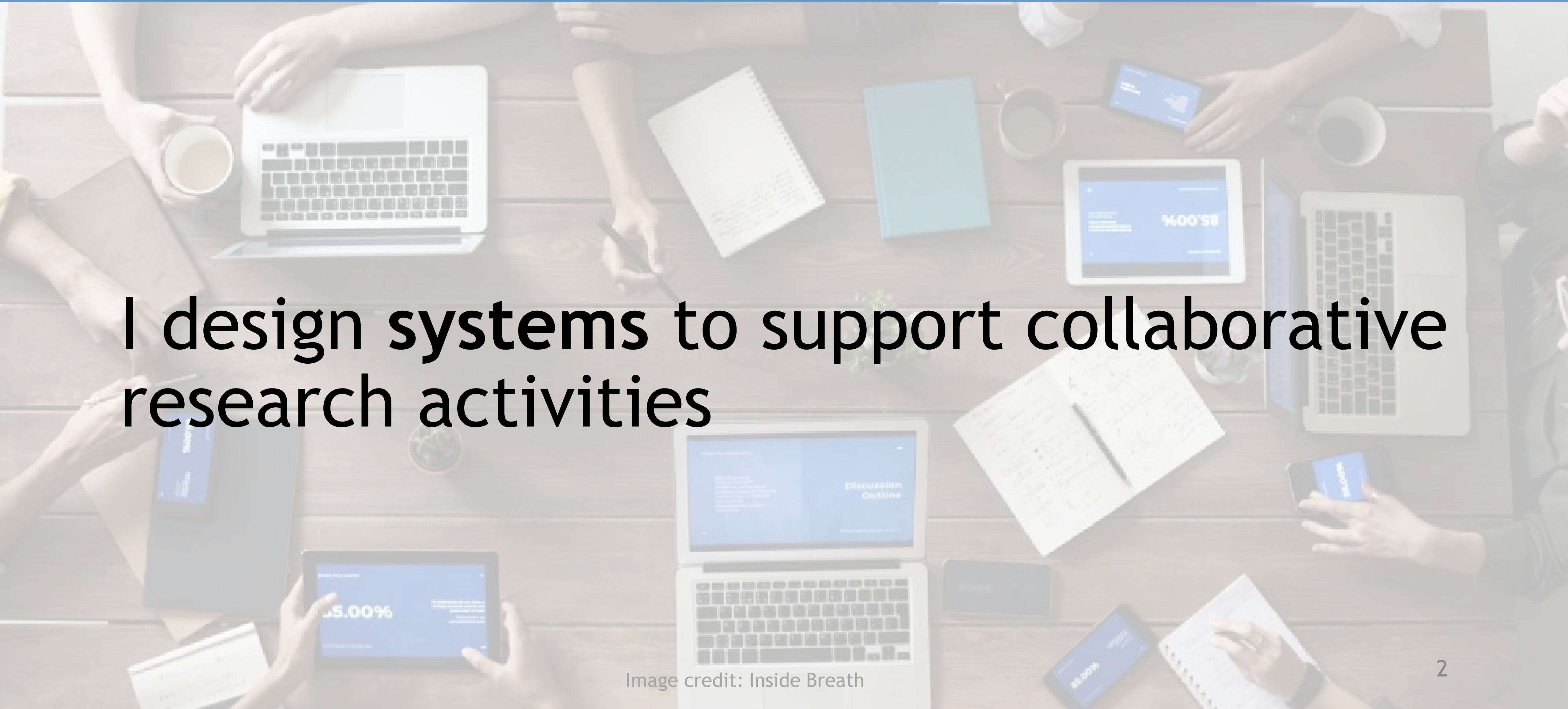
Supporting Multidisciplinary Collaborative Research Activities

Dr. Sarah Morrison-Smith

Hamilton College

Summary



I design **systems** to support collaborative research activities

Image credit: Inside Breath

Research Relies on Collaboration

Collaboration is the keystone of modern research, without which problems with high societal impact are **difficult if not impossible** to solve (Olson and Luo, 2007)

Enables significant scientific breakthroughs

- Ex: Developing the COVID-19 vaccine
  - Medical professionals collecting samples
  - Lab technicians extracting DNA
  - Bioinformaticians assembling genomes
  - Etc.

These projects have unique challenges that are not adequately addressed by technology (Morrison-Smith et al., 2015)

Image credit: St. Louis Community College

Problem: Insufficient Tools for Scientific Collaboration

Important: getting the technology to work

- Ex: Developing assembly algorithms (i.e., De Bruijn graphs (Compeau et al., 2011))

Really important: getting the **human-system interaction** right

- Affects collaborative activities:
  - Past work: Collaboration activities in genomics research are insufficiently supported
  - Past work: Prioritization of projects is difficult to perceive when working remotely
  - Current work: Precarious balance between data control and collaboration
  - Current work: Supporting collaborative qualitative data analysis
  - Future work: Metadata organization can help or hinder data sharing efficiency

Outline

Introduction

Background

Slides 3-7

- Problem overview
- HCI methods

2

Past Work

Slides 8-27

- Collaboration in genomics
- Project Prioritization

3

Current Work

Slides 28-53

- Data control vs collaboration
- Assisting qualitative data analysis

4

Future Work

Slides 54-63

- Large Language Models for qualitative data analysis
- Future collaborations

5

Conclusion

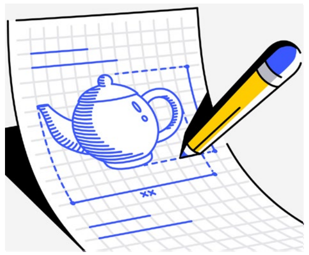
Slides 64-66

- Conclusion
- References

HCI Methods

Step 1: Understand user needs and requirements

- Methods: Interviews, observations, surveys
- Results: Research findings and design recommendations



Step 2: Prototype system

- Methods: low fidelity/paper prototyping, proof-of-concept systems, interviews, quantitative evaluation (NASA TLX, SUS)
- Results: Design & opportunities for improvement

Step 3: Iterate and refine

- Methods: system implementation, evaluation
- Results: working system

Image credit: Connect Carolina, Ecwid, ClipartMax

Qualitative Approach to Computer Science

Consider: you want to investigate student problem-solving strategies when implementing sort algorithms (Hazzan et al., 2006)

Quantitative research: compare students who implement algorithms in Python vs Java

- Tells us that there are differences in problem-solving performance (“**what**”)

Qualitative research: interview students in Python and Java groups

- Reveals mental processes that may explain how students in both groups use different sort algorithms (“**why**”)

Challenge 1: Collaboration Activities in Genomics are Insufficiently Supported

Challenges in Large-Scale Bioinformatics Projects. Morrison-Smith et al. Humanities and Social Science Communications ‘22

Problem: Collaboration is Insufficiently Supported

Genomic research relies on large, multi-institutional collaborations to generate and analyze large data (Morrison-Smith et al, 2015)

Scientific projects depending on many institutions and disciplines are less successful than those relying on fewer
(Cummings and Kiesler, 2005; Kiesler and Cummings, 2002)

RQ: Have advances in technology circumvented coordination issues in this research area?

Image Credit: MIT Sloan Management Review

Method

Qualitative methods

- Semi-structured interviews
- Observed software use

20 experienced life science researchers

Sessions

- Conducted at researchers' workplaces or via Skype/Zoom
- Lasted ~ 60 minutes

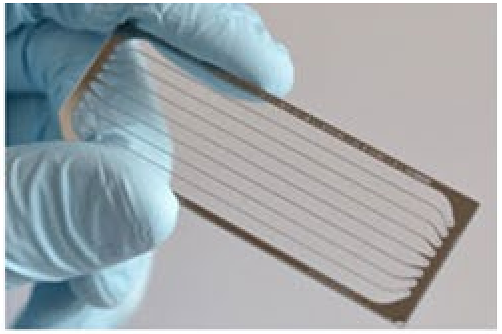
PID	Position	Department
P1	Adjunct	Epidemiology
P2	Faculty	Animal Sciences
P3	Post-Doc	Epidemiology
P4	Faculty	Microbiology
P5	Faculty	Plant Biology
P6	Faculty	Epidemiology
P7	Faculty	Biology
P8	Faculty	Bioinformatics
P9	Faculty	Animal Sciences
P10	Post-Doc	Immunology

PID	Position	Department
P11	Post-Doc	Industrial Hygiene
P12	Faculty	Veterinary
P13	Faculty	Biology
P14	Faculty	Epidemiology
P15	Laboratory Director	Proteomics and Metabolomics
P16	Faculty	Agriculture and Food Systems
P17	Faculty	Plant Biology
P18	Post-Doc	Plant Biology
P19	Faculty	Animal Sciences
P20	Chair	Animal Sciences

Context: Genomics Workflow

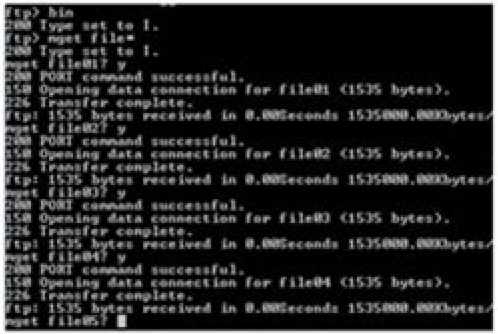
Collection

Extraction



Preparation

Sequencing



```
ftp> bin
200 Type set to I.
ftp> mget file*
200 Type set to I.
mget file01? y
200 PORT command successful.
150 Opening data connection for file01 (1535 bytes).
226 Transfer complete.
ftp: 1535 bytes received in 0.00Seconds 1535000.00Kbytes/
mget file02? y
200 PORT command successful.
150 Opening data connection for file02 (1535 bytes).
226 Transfer complete.
ftp: 1535 bytes received in 0.00Seconds 1535000.00Kbytes/
mget file03? y
200 PORT command successful.
150 Opening data connection for file03 (1535 bytes).
226 Transfer complete.
ftp: 1535 bytes received in 0.00Seconds 1535000.00Kbytes/
mget file04? y
200 PORT command successful.
150 Opening data connection for file04 (1535 bytes).
226 Transfer complete.
ftp: 1535 bytes received in 0.00Seconds 1535000.00Kbytes/
mget file05?
```

Transferring

Analysis

Findings: Multi-institutional Research is Unavoidable

Large, collaborative interdisciplinary and multi-institutional research is unavoidable in genomics

The need for expertise outweighs the difficulties associated with remote collaboration

Image Credit Force the Cloud

Findings: Coordination Challenges

Large teams are difficult to coordinate

- Researchers lose interest in large team meetings
- Collaborators feel under-appreciated and under-prioritized

Image Credit TT&CC

Findings: Scientific Communication

Multidisciplinary teams experience language barriers

- Complex ideas
- Technical jargon

Scientific communication, which relies on transmitting knowledge rather than information

“Sometimes it’s taking something that is complicated and explaining so it’s understandable. That’s difficult. You have to speak in terms of their language” (P7)

Image by Dmitry Baranovskiy

Recommendations for Design

Support the discussion and negotiation process as methodologies and data sharing standards evolve

Provide mechanisms for facilitating, simplifying, and documenting conversations surrounding scientific knowledge

- Allow users to search for abstract representations of information

Facilitate explicit management in large scale projects

Challenge 2: Prioritization of Projects is Difficult to Perceive in Remote Teams

AmbiTeam: Providing Team Awareness Through Ambient Displays. Morrison-Smith et al. Graphics Interface ’21

Facilitating Team Awareness Through Ambient Displays. Morrison-Smith et al. Microsoft New Future of Work Symposium ’20

Problem: Maintaining Awareness of Remote Teams

Research is increasingly conducted remotely without the benefit of informal interactions that help maintain awareness of each collaborator (Olson and Olson, 2014)

Without informal interactions, it's difficult to stay aware of our team's priorities

Researchers resort to adverse coping strategies to determine whether their collaborators are prioritizing the project

- Ex: gauge priorities using email

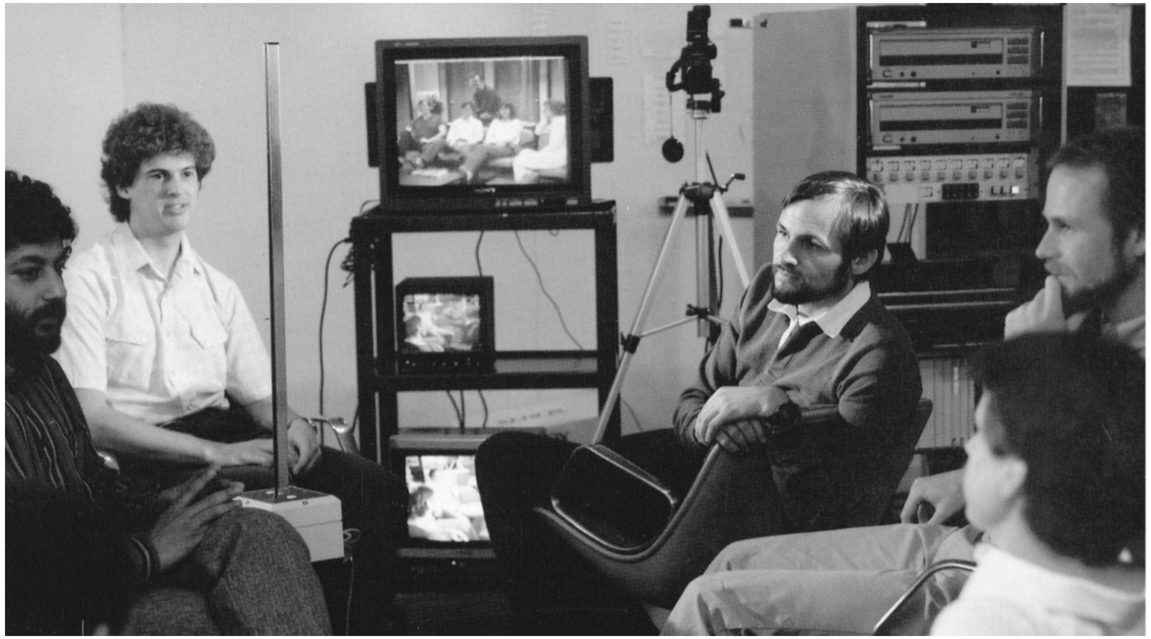
"if they're not physically in the same space...they are not necessarily on your radar in the same way." (P5)

Image Credit: Dmitry Baranovskiy

Existing Awareness Systems

Existing systems rely on disruptive notifications or require deliberate attention to use

RQ: How can we unobtrusively convey information about teammate's work efforts through ambient displays?



Media Spaces: Bringing People Together in a Video, Audio, and Computing Environment

Ambient Displays

- Communicate **contextual** or **background** information via peripheral awareness
- Don’t require constant attention
- Techniques:
  - Light (Brewer 1979,Dahley et al. 1998, Ishii et al. 1998)
  - Music (Barrington et al. 2006)
  - Art (Heiner et al. 1999)

Image Credit: The Verge

Tracking Project-Related Activity

- Untrackable Activities
  - Thoughts about the project
- Private Activities
  - Emails
  - Phone calls
- Potentially Acceptable
  - File Activity

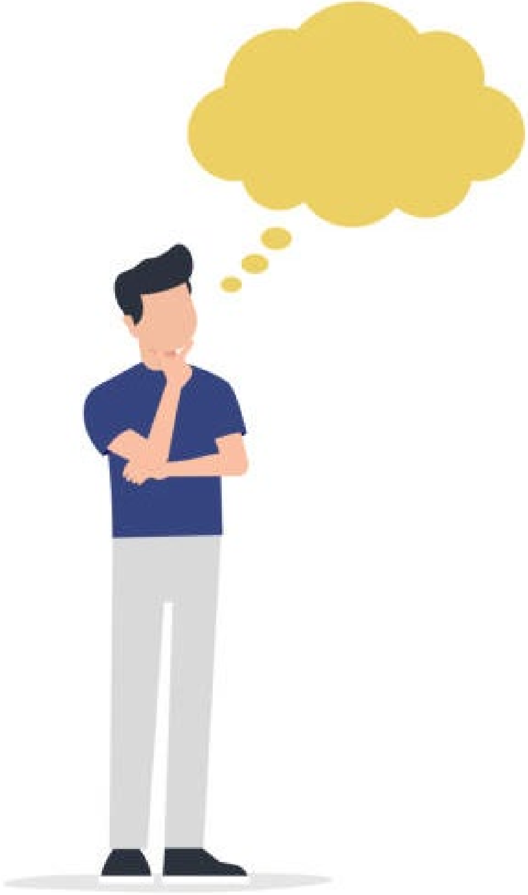


Image Credit: Ali Kerem

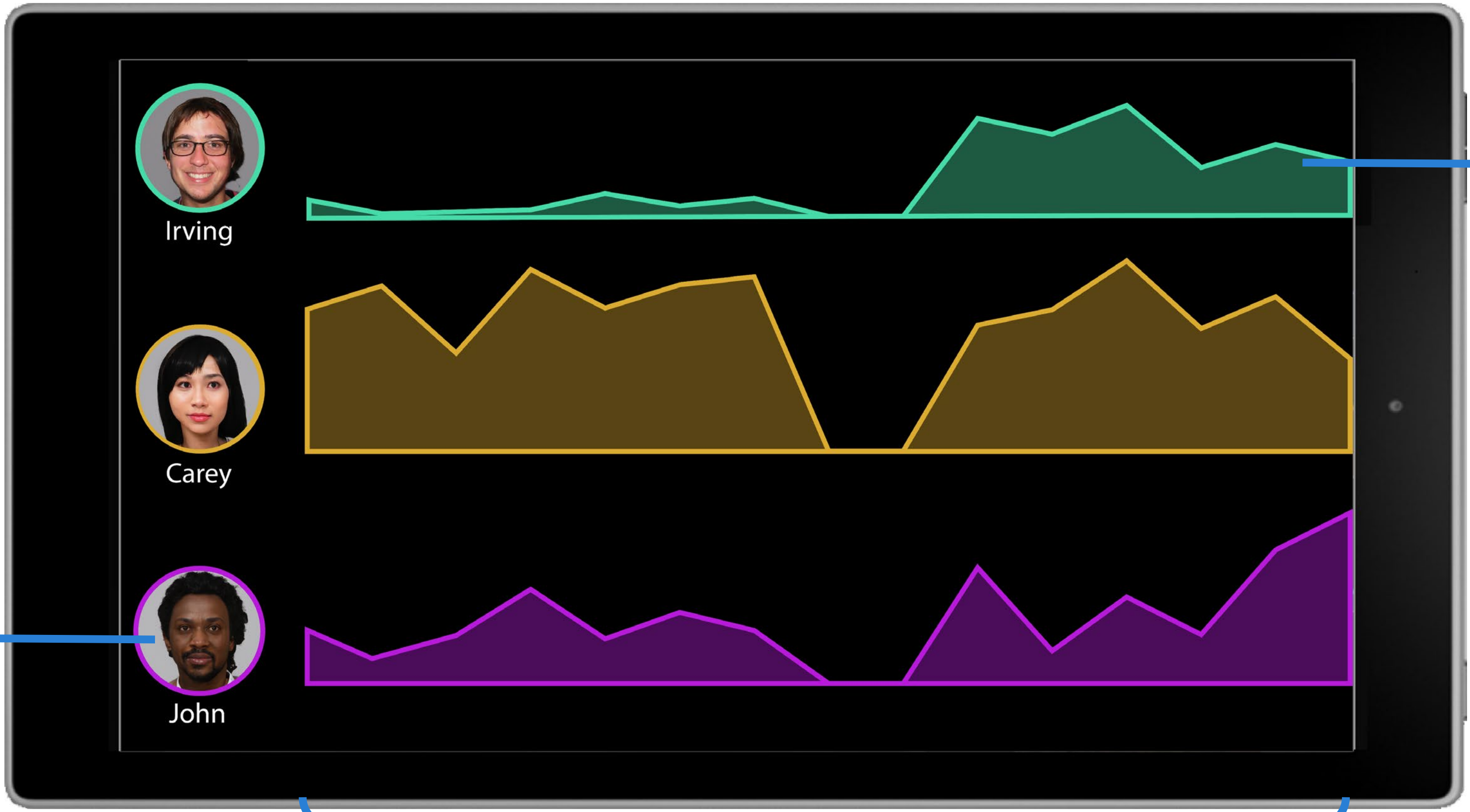
Tracked Activity

File creations, deletions, and changes

Abstracted to whether a change occurred, not what changes occurred

- Leveraged OS tracking
- Did not require viewing contents of files

AmbiTeam



# of Files Changed

Teammate

Window of time

Evaluation Method

10 researchers in teams of 2-4

- Teams ranged from fully co-located to fully remote

Used AmbiTeam for 4 weeks

Twice a week rated the effort everyone put into the project

- Correlation between visualization data and perceived effort

Finished with semi-structured interviews

AmbiTeam sitting at A1's desk

Findings (Quantitative)

RQ: Did information displayed by AmbiTeam is related to perceptions of effort?

- No correlation (r=0.09), p> 0.5) between displayed activity and personal scores
- Weak positive correlation (r=0.22, p=0.011) between activity and scores given to teammates

Conclusion: AmbiTeam **may** affect user’s perceptions of their teammate’s effort

Findings (Qualitative)

Many participants felt more motivated to work on specific projects while using AmbiTeam

Attributed to the “motivational sense of the presence of others”

(Olson and Olson, 2014)

“it felt strangely like a reward even though it was such a simple aesthetic visualization” (A1)

“Every single time that happened I was like, oh he’s working, I should probably work on it too.” C2

Image Credit: Dmitry Baranovskiy

Findings (Limitations)

Several activities that part of participants’ workflows were not tracked by AmbiTeam during the study

- Some are easy to add
    - e.g., number of papers in a literature library
- Some require changes to behavior
    - e.g., handwritten notes in a digital notebook
- Others would involve significant privacy violations
    - e.g., in-person conversations, emails, and internet browsing history

Image Credit: pixabay.com

Recommendations for Design

I recommend that future awareness systems consider:

- Using file activity to measure effort
- Implementing ambient displays that do not interrupt the user’s workflow

Image Credit: Lingford Consulting

Challenge 3: Balancing Data Control and Collaboration

Part 1: Investigating Data Sharing

DriveGroups: Using Group Perspective for Efficient and Secure Data Sharing in Life Sciences. Frazier et al. Under review

Background: Data Sharing

Data sharing is a hallmark of modern science
(Merton and Merton, 1968)

Data is shared both publicly and within collaborations

Data sharing methods depend on **who** the data is being shared with

- Ex: Collaborators at outside institutions
- Ex: Students
- Ex: Non-collaborators

Image credit: Hootsuite

Problem: Navigating Different Methods and Standards

- Data sharing is considered one of the biggest challenges faced by life-science researchers (Darby et al. 2012)
- Data from various disciplines differ in
  - Data characteristics
  - Collection methods
  - Sharing needs and criteria (Velden 2013)
- Data sharing needs of life-scientists are not adequately supported by existing technology (Morrison-Smith et al. 2015)
- RQ: What are the challenges and barriers to data sharing in life-science collaborations?

Image credit: Juno College

Method

Adapted contextual inquiry

- Semi-structured interviews
- Observed software use

26 experienced life science researchers

Sessions

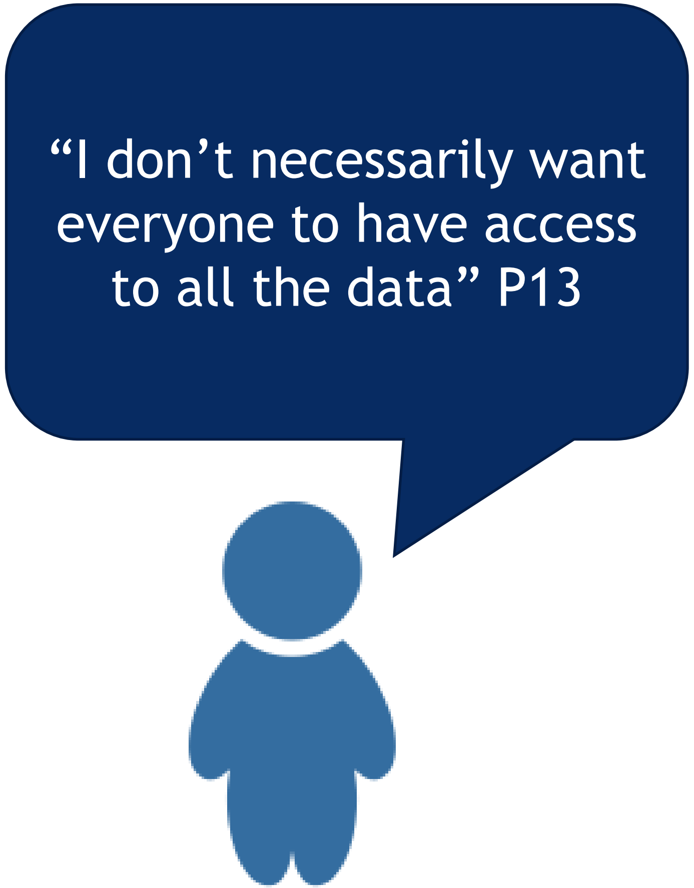
- Conducted at researchers' workplaces or via Skype/Zoom
- Lasted ~ 60 minutes

PID	Position	Department
P1	Adjunct	Epidemiology
P2	Faculty	Animal Sciences
P3	Post-Doc	Epidemiology
P4	Faculty	Microbiology
P5	Faculty	Plant Biology
P6	Faculty	Epidemiology
P7	Faculty	Biology
P8	Faculty	Bioinformatics
P9	Faculty	Animal Sciences
P10	Post-Doc	Immunology
P11	Post-Doc	Industrial Hygiene
P12	Faculty	Veterinary
P13	Faculty	Biology
P14	Faculty	Epidemiology
P15	Laboratory Director	Proteomics and Metabolomics
P16	Faculty	Agriculture and Food Systems
P17	Faculty	Plant Biology
P18	Post-Doc	Plant Biology
P19	Faculty	Animal Sciences
P20	Chair	Animal Sciences
P21	Faculty	Biology
P22	Faculty	Biology
P23	Faculty	Biology
P24	Lab Technician	Biology
P25	Post-Doc	Biology
P26	Faculty	Genomic Medicine

Findings: Trust and Control over Data

- Concerns about data control affects data sharing platforms
  - The passive nature of shared drives may lead to accidental sharing

- Institutions may not allow outside collaborators to have access to certain data



"I don't necessarily want everyone to have access to all the data" P13

Image Credit: Dmitry Baranovskiy

Findings: Data Protection

- Users are not always aware of available access controls or their affect on data sharing
- Users sometimes lack awareness of who has access to what files
- Users often use methods that they perceive as being insecure
  - Feel there is no good alternative

“There’s probably mechanisms that allow access to server data that you can control and that allow you to still maintain protection of sensitive data on the same server. But I don’t know what those systems are.” (P2)

Image Credit: Dmitry Baranovskiy

Findings: Relationship between Data Size and Expertise

- Data sharing platforms are determined by data size and user familiarity
- Users rely on email and off-the-shelf methods due to familiarity until they run out of space
- Users often lack the expertise to use file servers with more space

> “It would take me a lot longer to teach them how to access the server than for me to download to a 3-terabyte drive…it’s just easier to hand them a disk.” - P8

Image Credit: Dmitry Baranovskiy

Recommendations for Design

- Emphasize accessibility for protecting data
  - Provide transparent mechanisms for access control
- Support fine-grain levels of control
  - Users don’t necessarily want to share all their data with all their collaborators all the time
- Facilitate communication about methods and sharing protocols
  - Especially prior to steps that are costly to reproduce

Part 2: Developing Systems

DriveGroups: Using Group Perspective for Efficient and Secure Data Sharing in Life Sciences. Frazier et al. Under review

Problem: Balancing data control and collaboration

Balancing data access control and collaboration is a challenge

- Systems like Google Drive lack fine grained control
- Frameworks for usable access control models lack widespread implementation

Image credit: Boxcryptor

RQ: How can we promote discoverability and accessibility of access control?

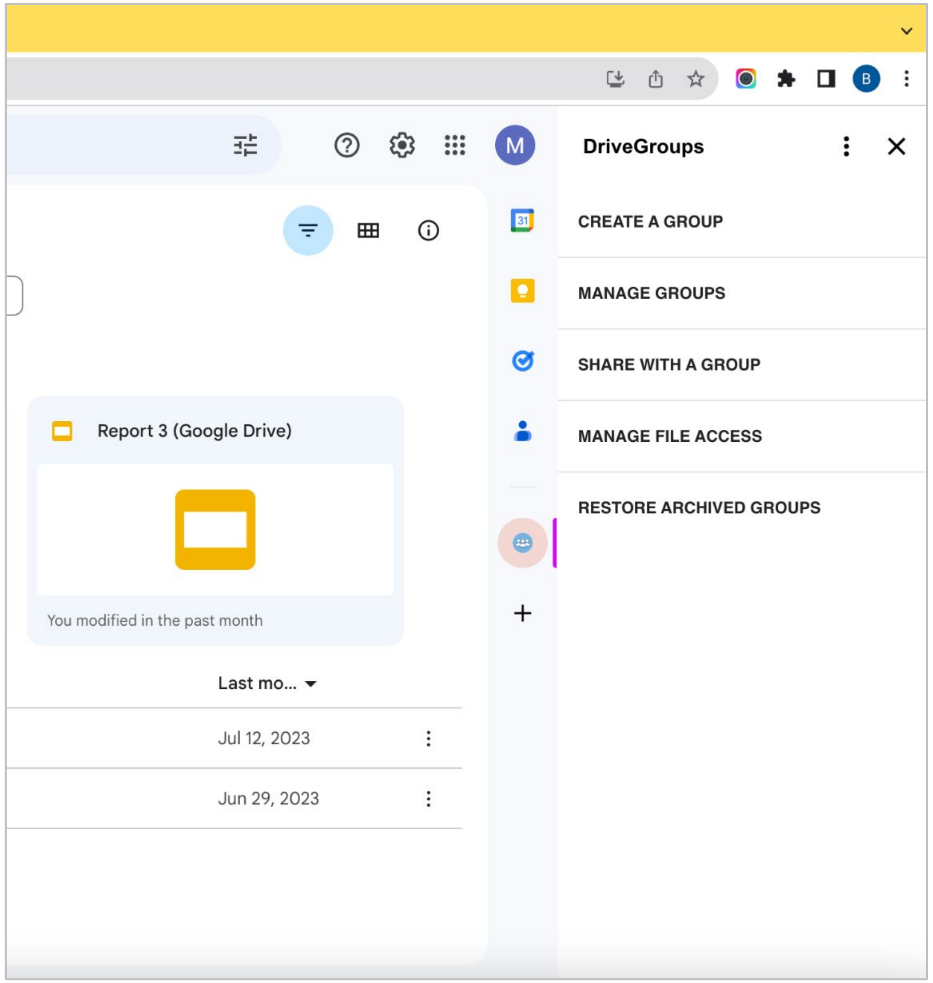
DriveGroups

Manages groups of people and controls who has access to data

Controls group privileges

Facilitates quick, specific changes in sharing permissions

Runs as an add-on to Google Drive, making it familiar and accessible



Evaluation Method

20 local undergraduate students

- Variety of majors and experience with Google Drive

Used DriveGroups and Google Drive to complete sharing tasks

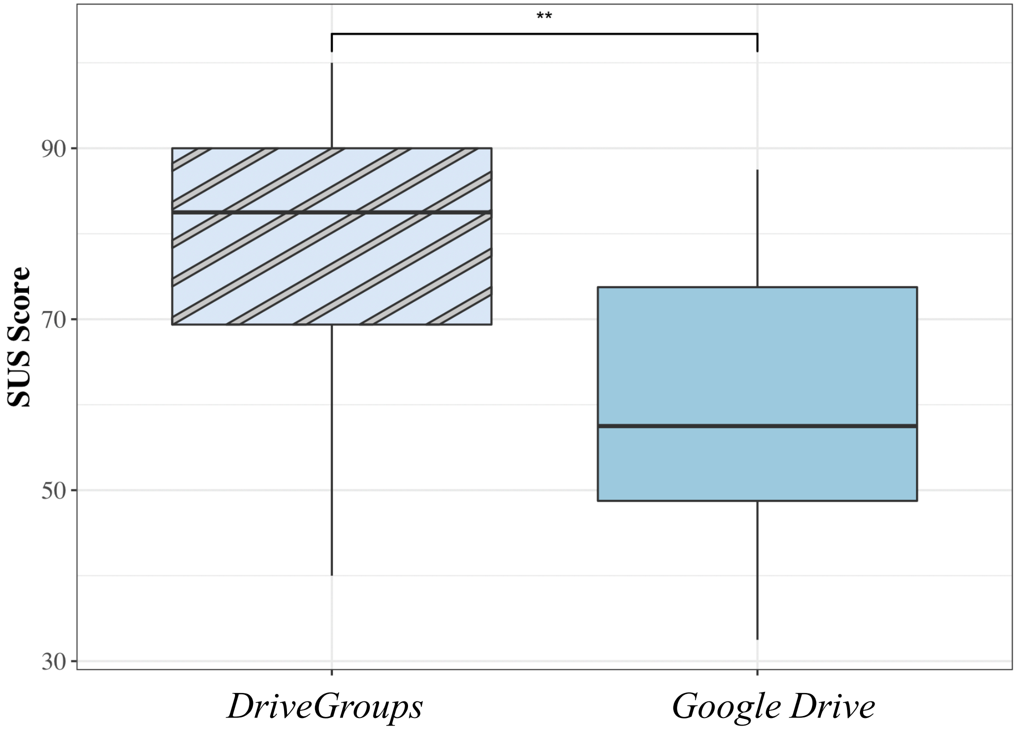
Evaluated both systems quantitatively through a survey

Finished with semi-structured interviews

Findings

- DriveGroups significantly improved on Google Drive's usability
- Moderate success with improving transparency
- Mild success at improving awareness of access control features

System Usability Scale Scores: DriveGroups vs. Google Drive



Upcoming Work

- Planning in-situ study with life science researchers from all over the US

- Participants will use DriveGroups for 4 weeks with their own files and computers

- Once a week rate system via emailed survey

- Finish with semi-structured interviews

Challenge 4: Collaborative Qualitative Data Analysis Relies on Inefficient Methods

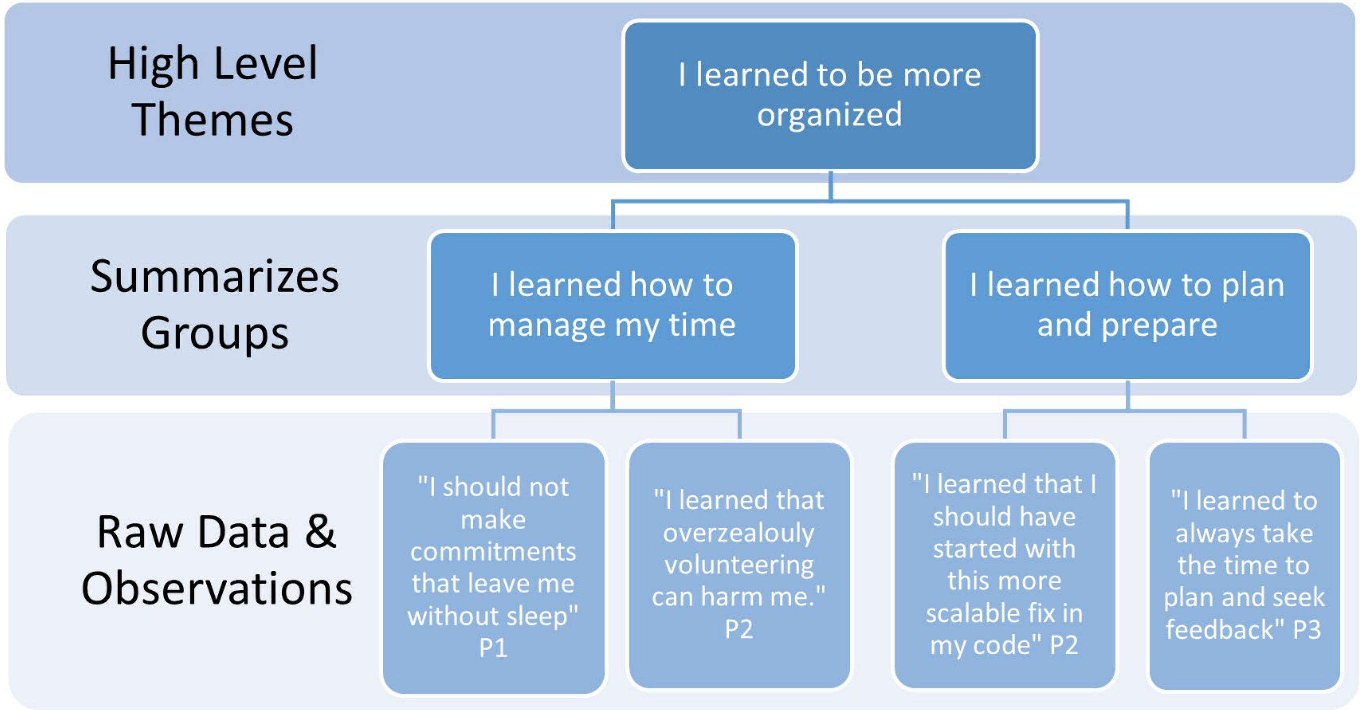
Affinity Diagramming

Affinity diagramming is a method for analyzing qualitative data

(Beyer and Holzblat, 1998)

Typically used by teams of researchers

This method is ideal for wide-ranging, unstructured data



High Level Themes

I learned to be more organized

Summarizes Groups

I learned how to manage my time

I learned how to plan and prepare

Raw Data & Observations

"I should not make commitments that leave me without sleep" P1

"I learned that overzealouly volunteering can harm me." P2

"I learned that I should have started with this more scalable fix in my code" P2

"I learned to always take the time to plan and seek feedback" P3

Challenge & Goal

Challenge: Affinity Diagramming is inefficient for large datasets

Solution: We can harness computing to provide suggestions for automatic data grouping

Goal: Design, implement, and evaluate a system that recommends groups of datum during affinity diagramming

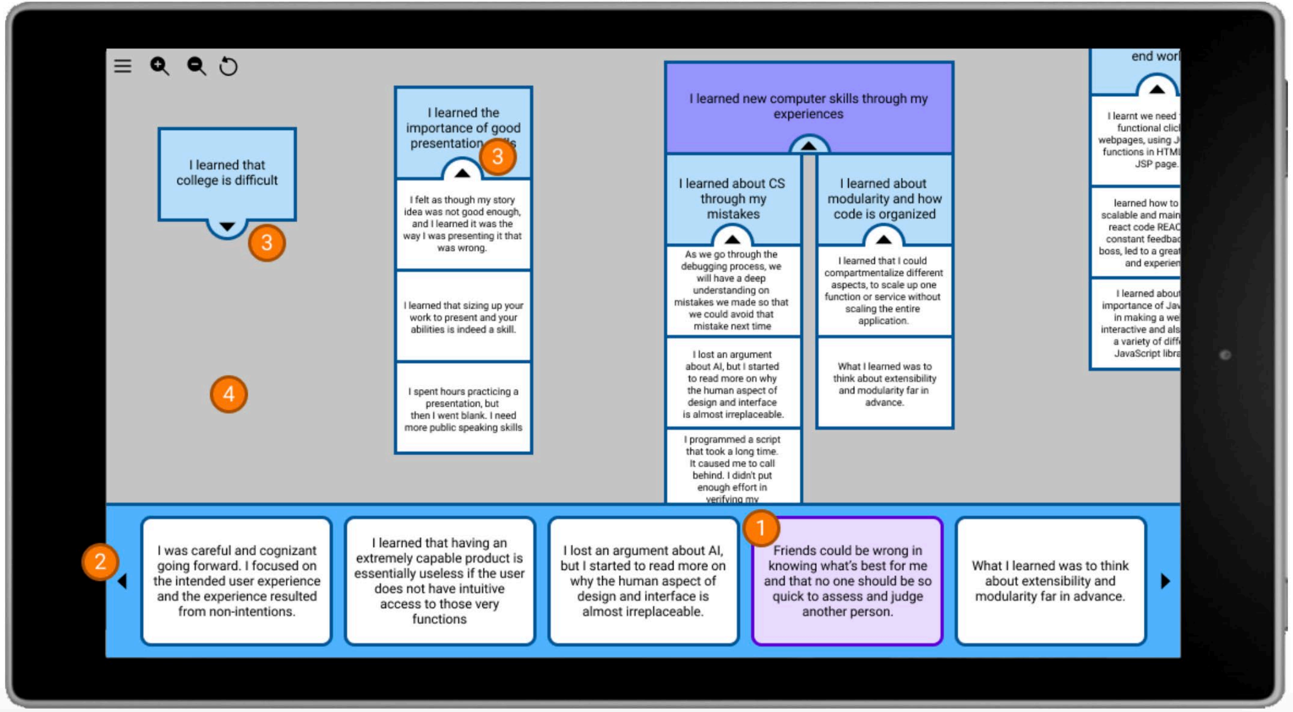
Part 1: QuAD - Qualitative Affinity Diagrammer

Computer Assisted Qualitative Data Analysis Tool: Qualitative Affinity Diagrammer (QuAD). Goldman et al. CHI EA 2022

QuAD (Qualitative Affinity Diagrammer)

QuAD scales qualitative data analysis by:

- Pre-grouping similar notes
- Providing labeling suggestions for the notes
- Providing a digital layout that is conducive to large diagrams



I learned that college is difficult

I learned the importance of good presentation skills

I felt as though my story idea was not good enough, and I learned it was the way I was presenting it that was wrong.

I learned that sizing up your work to present and your abilities is indeed a skill.

I spent hours practicing a presentation, but then I went blank. I need more public speaking skills

I learned new computer skills through my experiences

I learned about CS through my mistakes

As we go through the debugging process, we will have a deep understanding on mistakes we made so that we could avoid that mistake next time

I lost an argument about AI, but I started to read more on why the human aspect of design and interface is almost irreplaceable.

I programmed a script that took a long time. It caused me to call behind. I didn't put enough effort in verifying my

I learned about modularity and how code is organized

I learned that I could compartmentalize different aspects, to scale up one function or service without scaling the entire application.

What I learned was to think about extensibility and modularity far in advance.

end worl

I learnt we need functional clicl webpages, using J functions in HTM JSP page.

learned how to scalable and main react code REAC constant feedbac boss, led to a great and experien

I learned about importance of Jav in making a we interactive and als a variety of diff JavaScript libra

I was careful and cognizant going forward. I focused on the intended user experience and the experience resulted from non-intentions.

I learned that having an extremely capable product is essentially useless if the user does not have intuitive access to those very functions

I lost an argument about AI, but I started to read more on why the human aspect of design and interface is almost irreplaceable.

Friends could be wrong in knowing what's best for me and that no one should be so quick to assess and judge another person.

What I learned was to think about extensibility and modularity far in advance.

Determining Data Similarity

Step 1: Determine similarity using BERT

(Bidirectional Encoder Representations from Transformers)

- Developed for language modeling and next sentence prediction

Image credit: Jenny Bristol

Determining Data Similarity

BERT Can be used to determine sentence similarity

1. Convert sentences into vectors

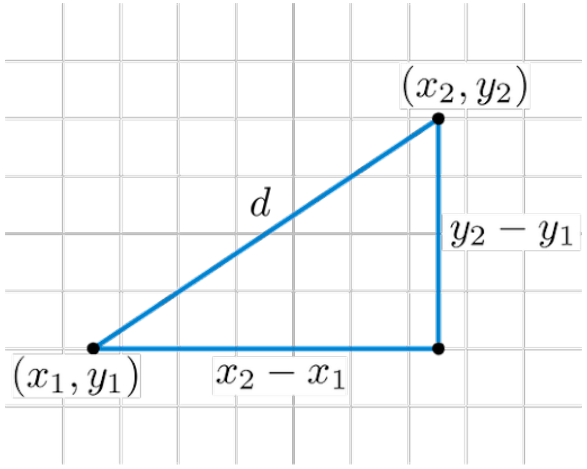
   $V1 = (X_1, X_2 \dots X_n)$, $V2 = (Y_1, Y_2 \dots Y_n)$

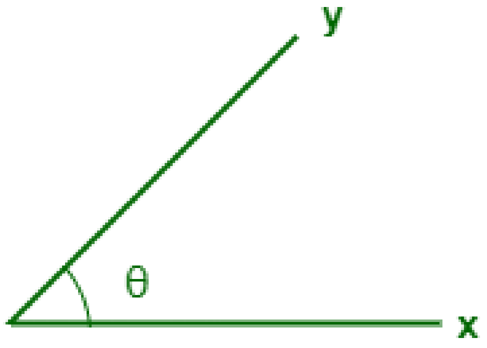
1. Calculate either:
   A. Smallest (Euclidian) distance between vectors

$$\sqrt{(x1 - y1)^2 + (x2 - y2)^2 + \dots + (xN - yN)^2}$$

   B. Smallest angle between vectors (cosine similarity)

$$\cos(x, y) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \, \|\vec{y}\|}$$





Data Clustering

Step 2: Group similar data using Girvan-Newman

- Detects when nodes in a network are densely connected internally
- Progressively removes edges from an original network of nodes
- Used to identify clusters of similar data
    - Create network of data using BERT similarity scores to weigh edges
    - Remove edges based on weight
    - Results in a tree that splits along different communities

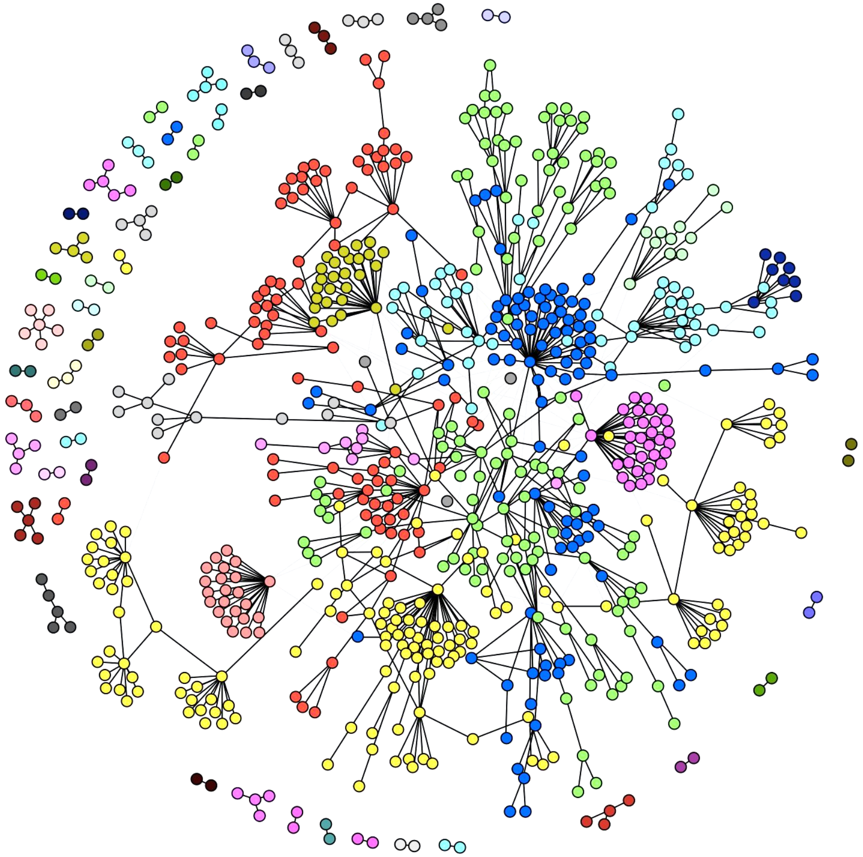


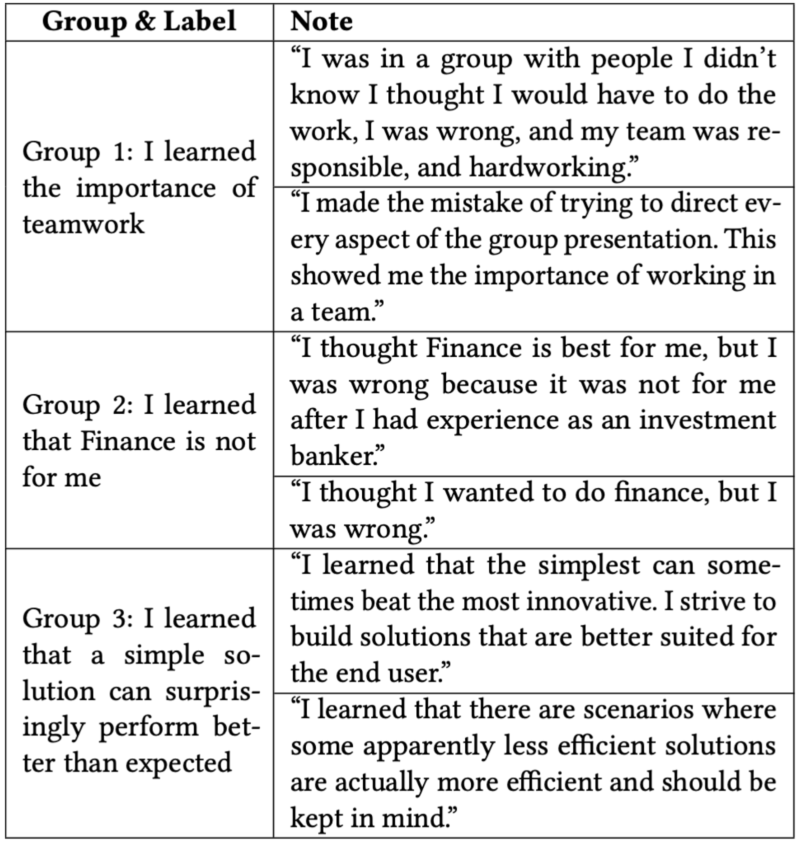
Image credit: Kernix

Preliminary Evaluation

- Ran QuAD’s grouping algorithm on 431 notes from a data set of upper-level computer science student survey
  - “Tell us about a time that you were wrong about something and learned from it.”
- Grouping process took 34.02 seconds and resulted in 15 automatic groups
  - 9 were of a high enough quality that they are likely to be accepted by a user and included in an affinity diagram

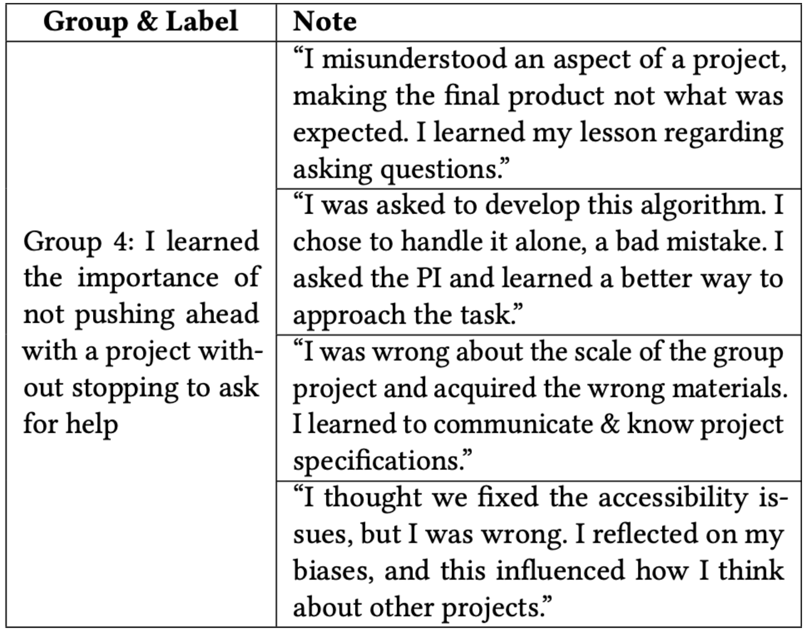
Findings: Example Groups

Example groups that are likely to be accepted by a QuAD user



Group & Label	Note
Group 1: I learned the importance of teamwork	"I was in a group with people I didn't know I thought I would have to do the work, I was wrong, and my team was responsible, and hardworking."
	"I made the mistake of trying to direct every aspect of the group presentation. This showed me the importance of working in a team."
Group 2: I learned that Finance is not for me	"I thought Finance is best for me, but I was wrong because it was not for me after I had experience as an investment banker."
	"I thought I wanted to do finance, but I was wrong."
Group 3: I learned that a simple solution can surprisingly perform better than expected	"I learned that the simplest can sometimes beat the most innovative. I strive to build solutions that are better suited for the end user."
	"I learned that there are scenarios where some apparently less efficient solutions are actually more efficient and should be kept in mind."

Example group that are likely to be edited, then accepted by a QuAD user



Group & Label	Note
Group 4: I learned the importance of not pushing ahead with a project without stopping to ask for help	"I misunderstood an aspect of a project, making the final product not what was expected. I learned my lesson regarding asking questions."
	"I was asked to develop this algorithm. I chose to handle it alone, a bad mistake. I asked the PI and learned a better way to approach the task."
	"I was wrong about the scale of the group project and acquired the wrong materials. I learned to communicate & know project specifications."
	"I thought we fixed the accessibility issues, but I was wrong. I reflected on my biases, and this influenced how I think about other projects."

Recommendations for Design

- Computation can be used to create coherent groups of qualitative data
    - About the same level as an intern who isn’t an expert
- Humans in the loop are necessary
    - Perfection is not possible. Humans are necessary to catch mistakes
- Future work should focus on fine tuning
    - Allow human feedback to improve system recommendations

Moving Forward

Future Work Directions

Focus: user interaction for multidisciplinary research to better support the scientific efforts that have important social impact

- Example: Large Language Models can Assist in Qualitative Data Analysis

Enter Large Language Models

- Large Language Models (LLM’s) such as GPT-4 have potential
  - No need to first identify similarity and then group!
  - Prompt can jump straight to forming and labeling groups
  - Systems like ChatGPT can even provide reasons why notes are grouped together

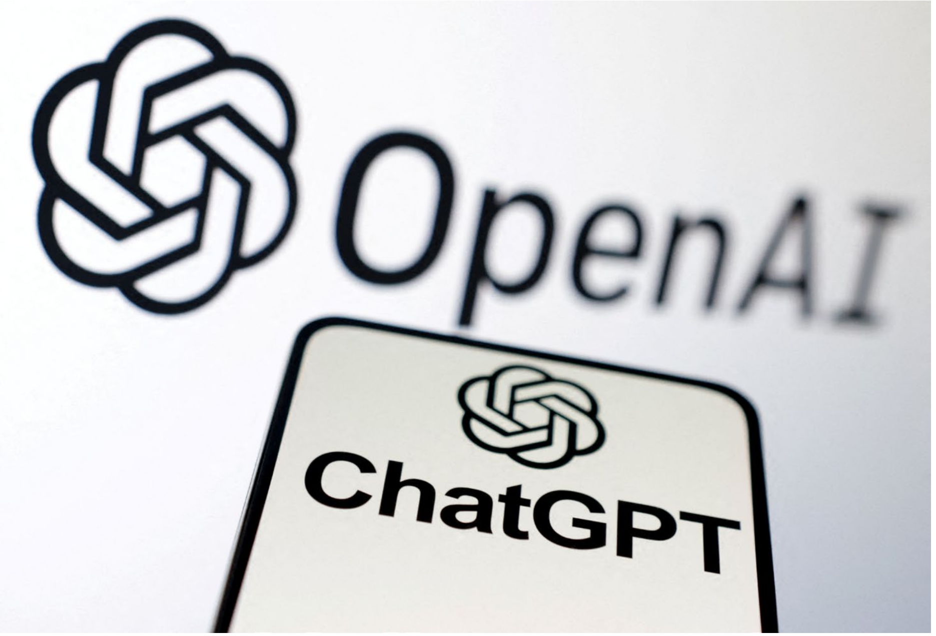


Image credit: Reuters

What is a Large Language Model?

Large language models (LLMs) are advanced AI systems designed to process and generate human-like text

They assist with tasks like text generation, translation, summarization, and more

These models are trained on vast amounts of text data, enabling them to understand and generate text in multiple languages

Image credit: Infinity Networking

Training LLMs

Language models are trained on massive datasets containing text from the internet, books, and other sources

LLMS use deep neural networks, often employing Transformer architectures, which allow them to capture complex patterns in language

Models are fine-tuned through a two-step process - pre-training (learning grammar and facts) and fine-tuning (tailoring for specific tasks)

Training these models requires powerful hardware, often using thousands of GPUs or TPUs

Image credit: Getty Images

Neural Networks

A neural network (NN) is a computational model inspired by the structure and functioning of the human brain

It's composed of interconnected artificial neurons organized into layers

NNs learn to approximate complex functions by adjusting its internal parameters (weights) during training

NNs can capture patterns and relationships in data, making them powerful tools for a wide range of machine learning tasks

A simple neural network

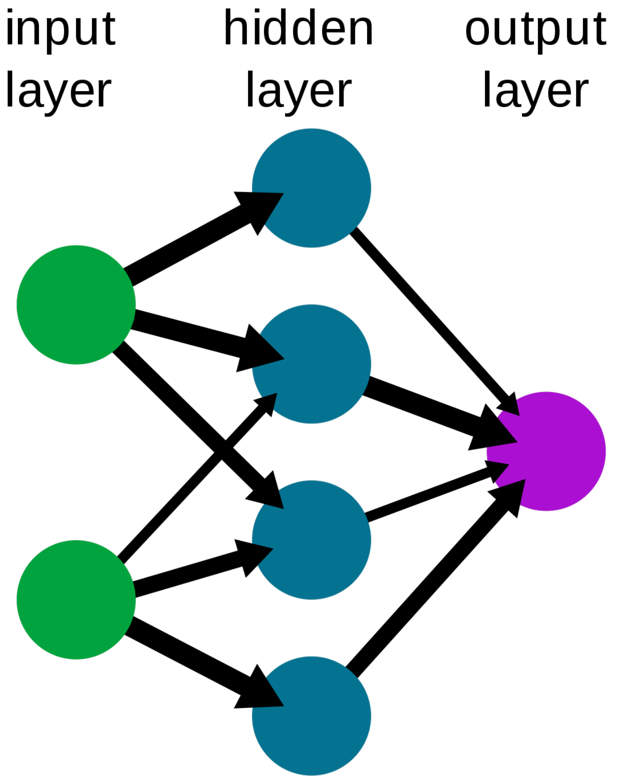


Image credit: Wikimedia Commons

Text Generation Process

Users provide a prompt or context, which the model uses to generate text

The model calculates the probability of the next word based on the context and its training data

Random sampling or other techniques are used to select the next word, balancing creativity and coherence

The model generates text, which can be further refined or customized based on user feedback

Image credit: Adobe Stock

Upcoming Work

- Develop Systematic Qualitative Information Diagrammer (SQuID) which uses OpenAI GPT-4 to group and label qualitative data

- Planning remote study to assess SQuID's effectiveness with qualitative researchers across the US

- Participants will use SQuID on test data as a warmup, then use it to analyze their own data

- We want to see if SQuID changes the affinity diagramming process for the better

Future Collaborations

Excited to collaborate on ongoing research in access control when collaboratively sharing data

Potential collaborators:

Conclusion

Conclusion

Getting the **human-system interaction** right is a key challenge in computer science

Through qualitative research, we can:

- Understand how to employ new technology
- Ensure that the systems we design will be used in practice

My goal is to use qualitative research to **design**, **implement**, and **evaluate systems** to support multidisciplinary research collaborations

Contact and Acknowledgements

Sarah Morrison-Smith, Ph.D.
Assistant Professor
Hamilton College

✉ smorriso@hamilton.edu
🌐 www.sarahmorrisonsmith.com
🌐 www.mochiresearch.com

DriveGroups Team

- Perrin Anto (Google)
- Julia Chang
- Nazaret Cuadros
- James Frazier*
- Iris Izydorczak*
- Catherine O'Brien
- Hariti Patel
- Emily Ringel
- Dipashreya Sur
- Yiyun Wang
- Emily Weinstein*
- Yifan Wu*
- Morgan Zee

* Current student

QuAD + SQuID Teams

- Francesca Cavuoti
- Jade Chen
- Alexandra Cheng
- Dr. Lydia Chilton (Columbia)
- Cindy Espinosa
- Sebastian Favela*
- Ariel Goldman
- Luiza Leschziner
- Matthew Maillet*
- Sabrina Meng
- Shivani Patel
- Aditi Patil
- Yifan Wu*

References

1. G. M. Olson, and A. Luo, "Intra-and inter-cultural collaboration in science and engineering." In *International Workshop on Intercultural Collaboration*, 249–259. (2007)
2. S. Morrison-Smith, C. Boucher, A. Bunt, J. Ruiz. "Elucidating the role and use of bioinformatics software in life science research." Proc. of British HCI. (2015)
3. P. E. C. Compeau, P. A. Pevzner, and G. Tesler. "How to apply de Bruijn graphs to genome assembly." *Nature biotechnology* 29.11 (2011)
4. O. Hazzan et al. "Qualitative Research in Computer Science Education." ACM SIGCSE Bulletin. (2006)
5. J.N. Cummings and S. Kiesler. "Collaborative research across disciplinary and organizational boundaries." *Social studies of science*. (2005)
6. S. Kiesler and J.N. Cummings. "What do we know about proximity and distance in work groups? A lgegacy of research." *Distributed work*. (2002)
7. J.S. Olson and G. M. Olson. "Bridging Distance: Empirical studies of distributed teams." Human-Computer Interaction and Management Information Systems: Applications. Advances in Management Information Systems. (2014)
8. J. Devlin et al. "BERT: Pre-training of deep bidirectional transformers for language understanding" arXiv preprint. (2018)
9. R. Gonçalves, M. Musen. "The variable quality of metadata about biological samples used in biomedical experiments." Sci Data 6, 190021 (2019).
10. S. Morrison-Smith, A. Sarcevic, C. Boucher, and J. Ruiz. "Challenges in Large-Scale Bioinformatics Projects." Under review. (2021)
11. A. Dahley, C. Wisneski, and H. Ishii. Water lamp and pinwheels: ambient projection of digital information into architectural space. (1998)
12. M. B. Brewer. "In-group bias in the minimal intergroup situation: A cognitive-motivational analysis." Psychological bulletin (1979).
13. H. Ishii et al. ambientROOM: Integrating Ambient Media with Architectural Space. In CHI 98 (1998).
14. Luke Barrington, Michael J Lyons, Dominique Diegmann, and Shinji Abe. "Ambient display using musical effects." (2006).
15. Jeremy M. Heiner, Scott E. Hudson, and Kenichiro Tanaka. "The information percolator: ambient information display in a decorative object." UIST '99. (1999)
16. M. Girvan and M.E.J. Newman. "Community structure in social and biological networks." PNAS. (2002)

Deep Dive Into QuAD

Converting notes into vectors

How to Convert Sentences to Vectors

1. **Tokenization:**
   1. Begin by breaking down the sentence into individual words or tokens.
   2. Example: "The quick brown fox" becomes ["The", "quick", "brown", "fox"].

2. **Vocabulary Creation:**
   1. Create a vocabulary of all unique words (tokens) present in your corpus.
   2. Example: ["The", "quick", "brown", "fox"].

3. **Vectorization:**
   1. Represent each sentence as a vector of fixed length, where each element corresponds to a word from the vocabulary.
   2. The value of each element indicates the frequency of the corresponding word in the sentence.

4. **Counting Word Frequencies:**
   1. For each sentence, count the number of times each word from the vocabulary appears in that sentence.
   2. Example: "The quick brown fox" might be represented as [1, 1, 1, 1] in the vector if each word appears once.

5. **Normalization (Optional):**
   1. You can optionally normalize the vectors by dividing the word frequencies by the total number of words in the sentence. This helps account for sentence length variations.

BONUS: Text Preprocessing

1. Remove noise
   - Text file header, footer, HTML, XML, markup data etc.
   - HTML & XML → BeautifulSoup library (Python), markup & headers → RegEx
2. Tokenization
   - Split data into small chunk of words
   - “Ross 128 is earth like planet.” → [‘Ross’, ‘128’, ‘is’, ‘earth’, ‘like’, ‘planet’, ‘.’]
   - NLTK word_tokenize() (Python)
3. Normalization
   - Use stemming and lemmatization to convert tokens to meaningful words
   - Punctuation, stop words(is, in, that, can, etc.), upper → lower case words
   - [‘Ross’, ‘128’, ‘is’, ‘earth’, ‘like’, ‘planet’, ‘.’] →
     [“ross’, ‘128’, ‘earth’, ‘like’, ‘planet’]

BONUS: Vectorization with Bag of Words

- “There used to be Stone Age”
- “There used to be Bronze Age”
- “There used to be Iron Age”
- “There was Age of Revolution”
- “Now it is Digital Age”

BONUS: Vectorization with Bag of Words

- “There used to be Stone Age”
- “There used to be Bronze Age”
- “There used to be Iron Age”
- “There was Age of Revolution”
- “Now it is Digital Age”

Word
There
was
to
be
used
Stone
Bronze
Iron
Revolution
Digital
Age
of
Now
it
is

BONUS: Vectorization with Bag of Words

- “There used to be Stone Age”
- “There used to be Bronze Age”
- “There used to be Iron Age”
- “There was Age of Revolution”
- “Now it is Digital Age”

Word	Count
There	1
was	0
to	1
be	1
used	1
Stone	1
Bronze	0
Iron	0
Revolution	0
Digital	0
Age	1
of	0
Now	0
it	0
is	0

BONUS: Vectorization with Bag of Words

- “There used to be Stone Age”
- “There used to be Bronze Age”
- “There used to be Iron Age”
- “There was Age of Revolution”
- “Now it is Digital Age”

→ [1,0,1,1,1,1,0,0,0,0,1,0,0,0,0]

BONUS: Vectorization with Bag of Words

- “There used to be Stone Age”
- “There used to be Bronze Age”
- “There used to be Iron Age”
- “There was Age of Revolution”
- “Now it is Digital Age”

[1,0,1,1,1,1,0,0,0,0,1,0,0,0,0]

[1,0,1,1,1,0,1,0,0,0,1,0,0,0,0]

[1,0,1,1,1,0,0,1,0,0,1,0,0,0,0]

[1,1,0,0,0,0,0,0,1,0,0,1,0,0,0]

[0,0,0,0,0,0,0,0,0,0,1,0,1,1,1]