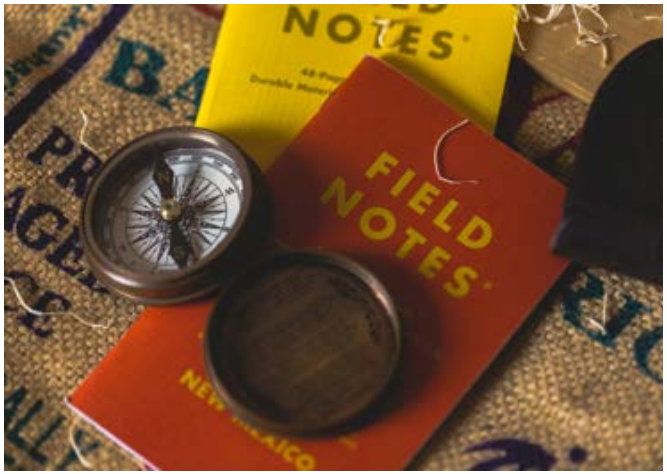
Good Morning!

Grab a coffee

Say hello!

Add your experience stickers to the Skills Map

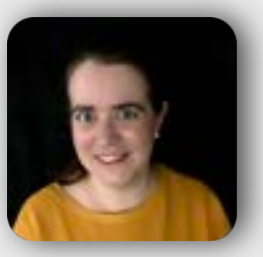
Place colored stickers based on your skills & background.



Pick up your Traveler's Notebook

Find your participant sheet & jot down initial thoughts

Today's Workshop



Corina Paraschiv

Private sector

Analyst, mixed methods researcher, product owner

Carnegie Mellon University alumna

Planning with/for Data Errors

- Ice-Breaker
- Breakout Session
- Debrief
- In-the-Field Project

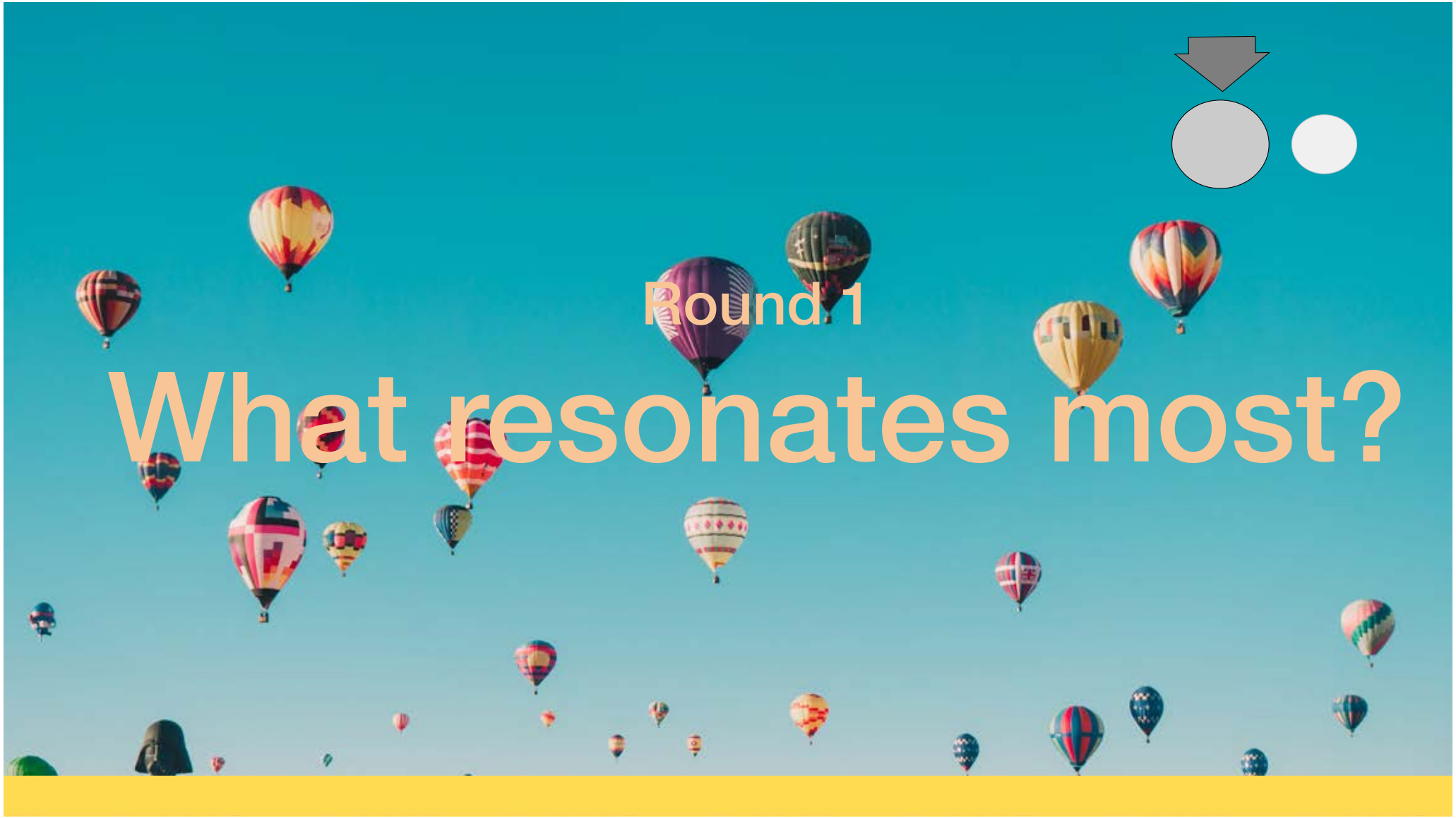
Automated Text Analysis on Travel Reviews

Participant Sheet - Section 1

03:00

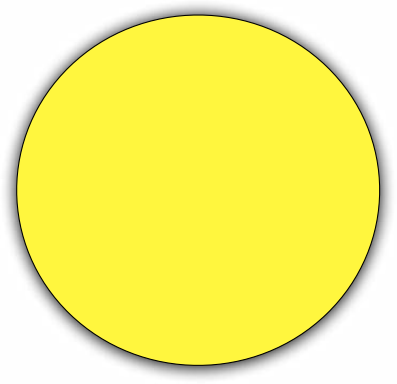
J.CREW

“Travel Tales!”
“Tell Me More!”

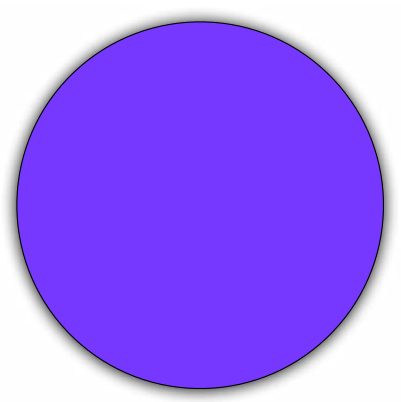


Round 1

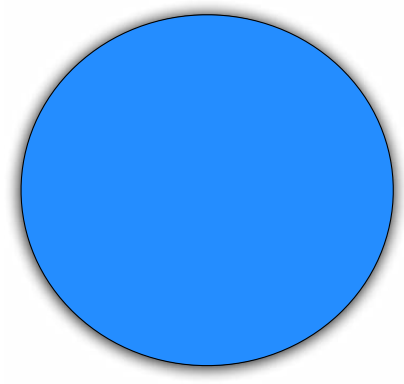
What resonates most?



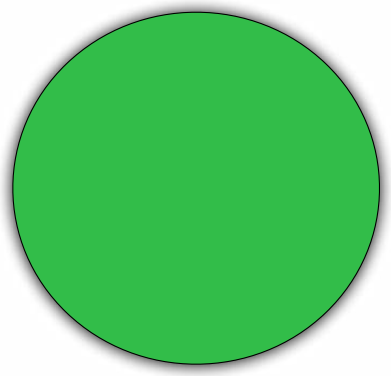
“My computer has multiple language keyboards installed on it”



“I often use auto-correct on my phone”

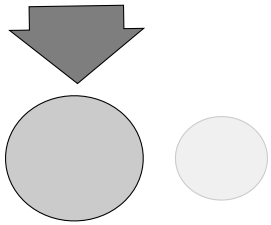


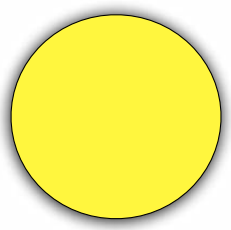
“I have never used Grammarly”



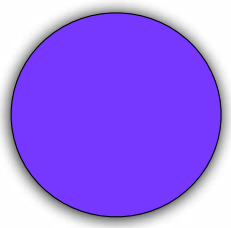
"Emoticons are a part of my vocabulary”

Round 1
What resonates most?

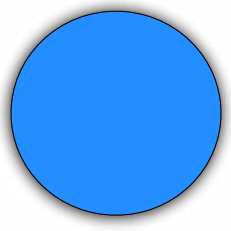




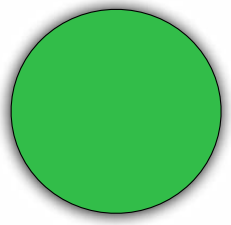
“My computer has multiple language keyboards installed on it”



“I often use auto-correct on my phone”



“I have never used Grammarly”



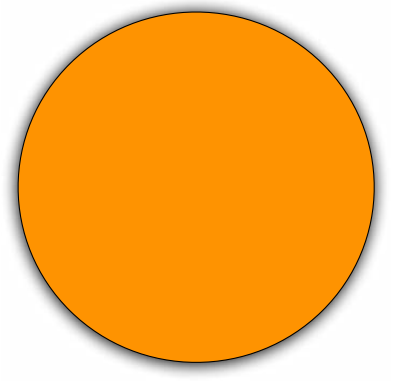
"Emoticons are a part of my vocabulary”

Ice-Breaker Q1

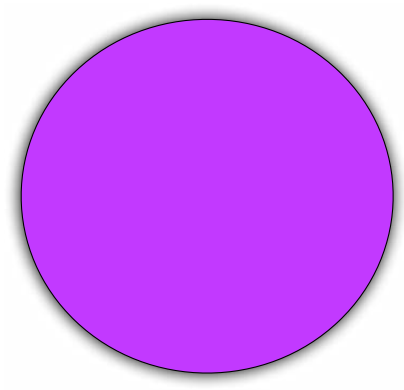
03:00

Round 2

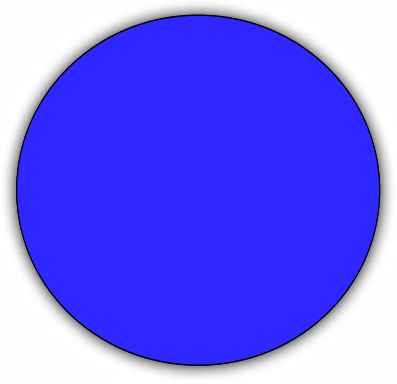
What resonates most?



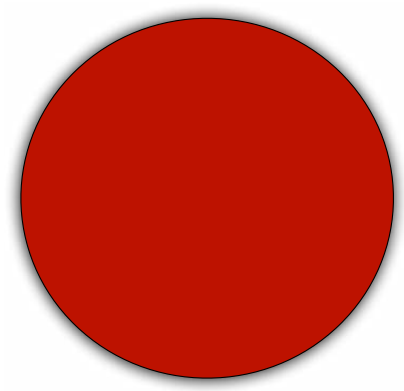
“Literature, poetry and crossword puzzles are little pleasures of life”



“Traveling is the greatest way to spend a summer vacation”

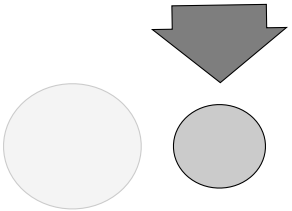


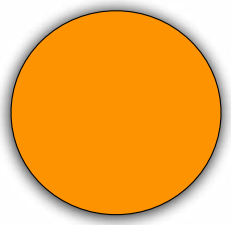
“I have grown-up or lived in a small town or village for at least one year in my life”



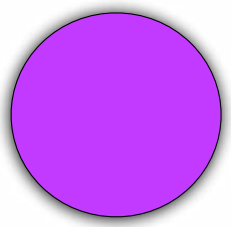
“You can say anything, it's just how you say it”

Round 2
What resonates most?

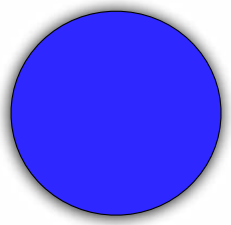




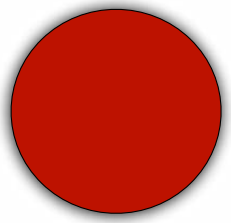
“Literature, poetry and crossword puzzles are little pleasures of life”



“Traveling is the greatest way to spend a summer vacation”



“I have grown-up or lived in a small town or village for at least one year in my life”



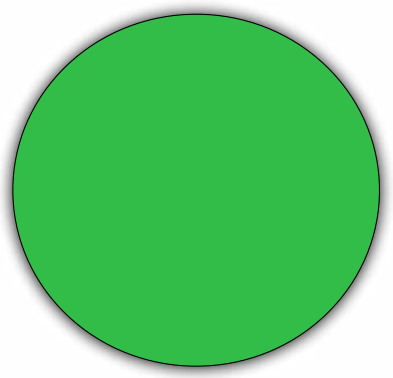
“You can say anything, it's just how you say it”

Ice-Breaker Q2

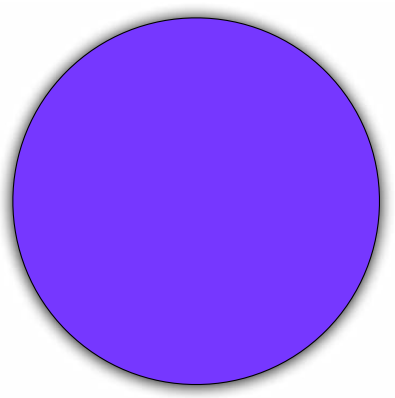
03:00

Round 3

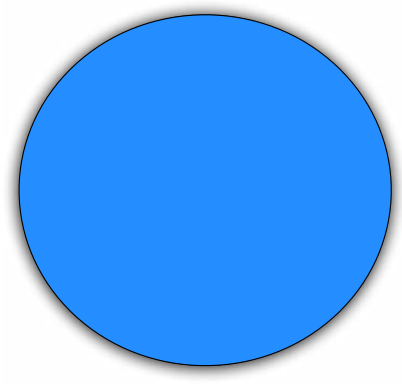
What resonates most?



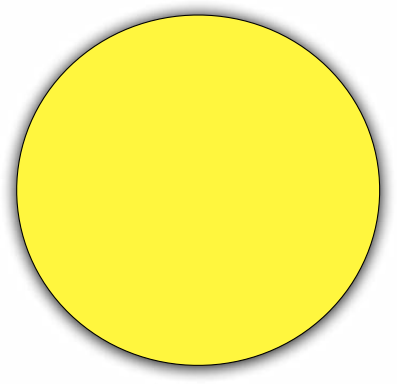
“I love programming”



“I am good with numbers”

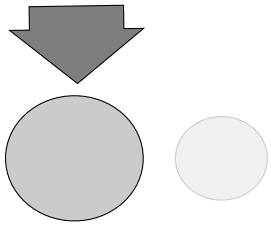


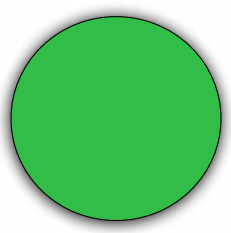
“I excel at language arts”



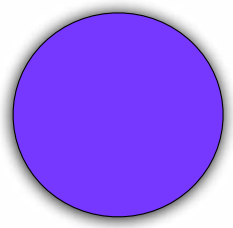
“I'm great with people”

Round 3
What resonates most?

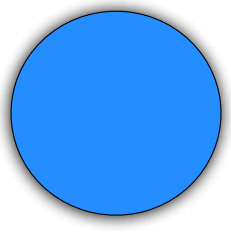




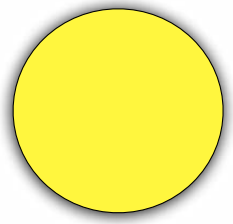
“I love programming”



“I am good with numbers”



“I excel at language arts”



“I'm great with people”

Ice-Breaker Q3

03:00

Break-Out Session

Breakout Session

Participant Sheet - Section 2

03:00

Break-Out Session

First

Use the clues you've been given to find the error

Second

Report the error found to the TA

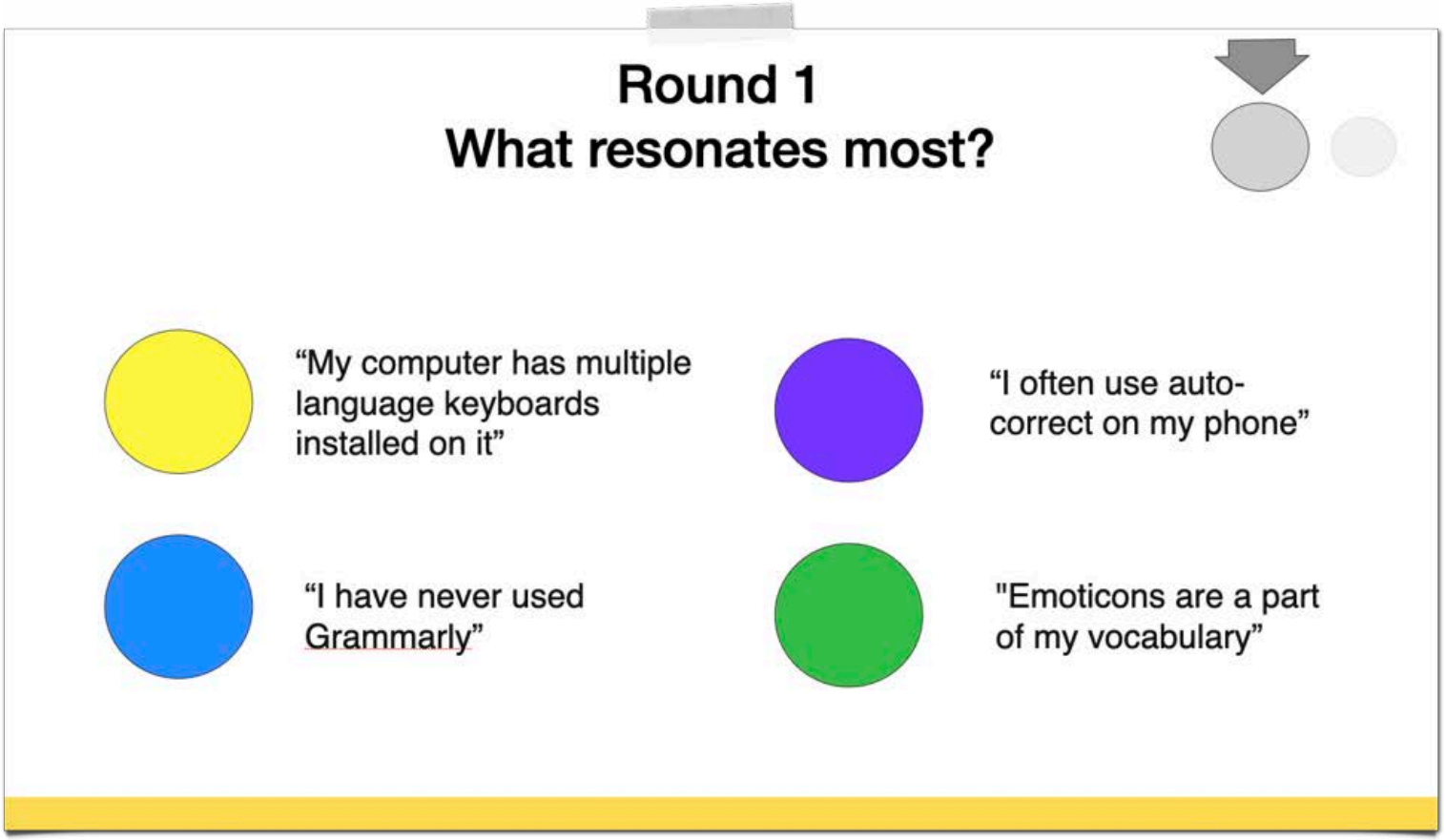
Third

Read the collectively identified errors and update your participant sheet

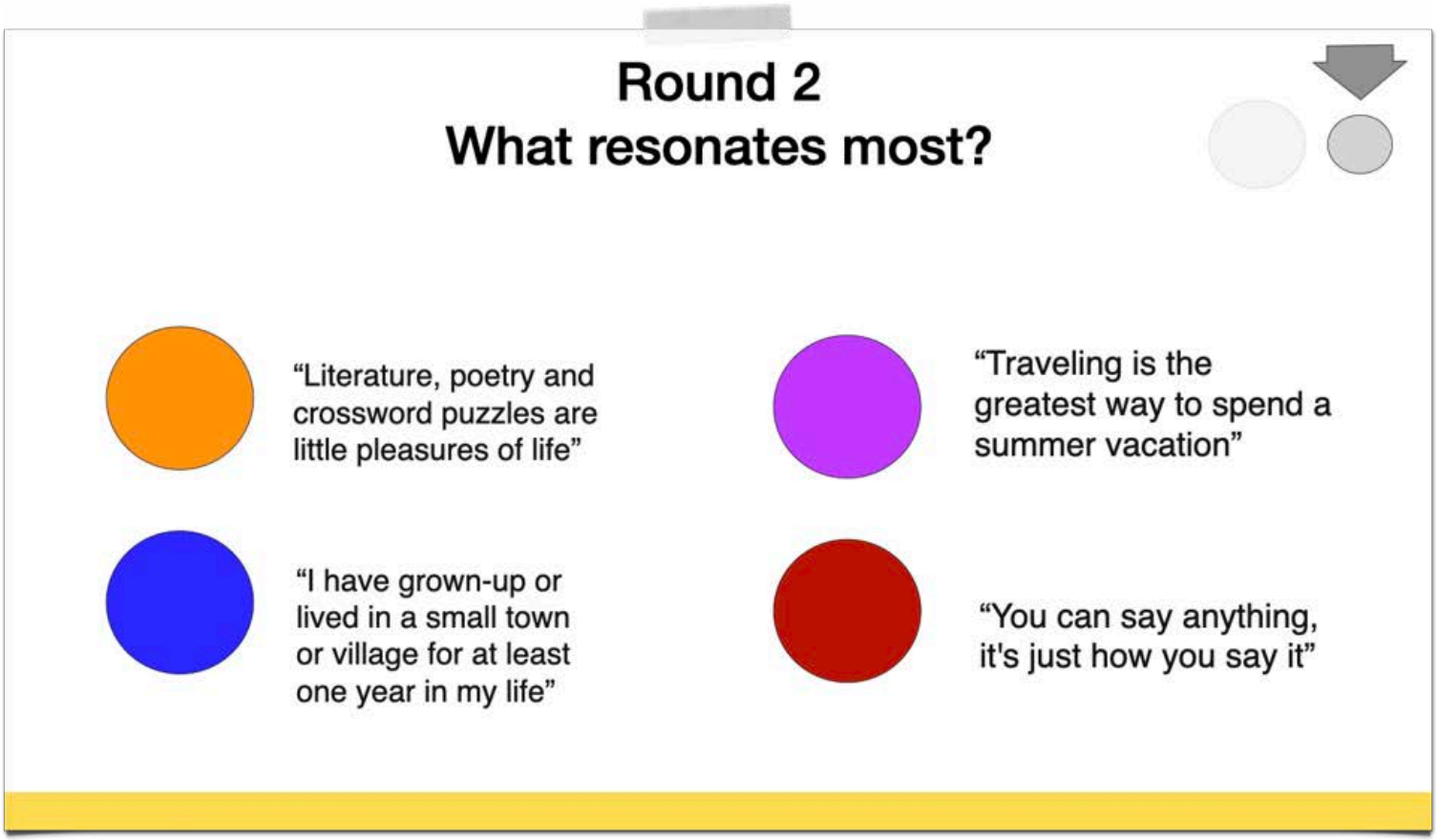

ENJOY LUNCH!

Participant Sheet - Section 3

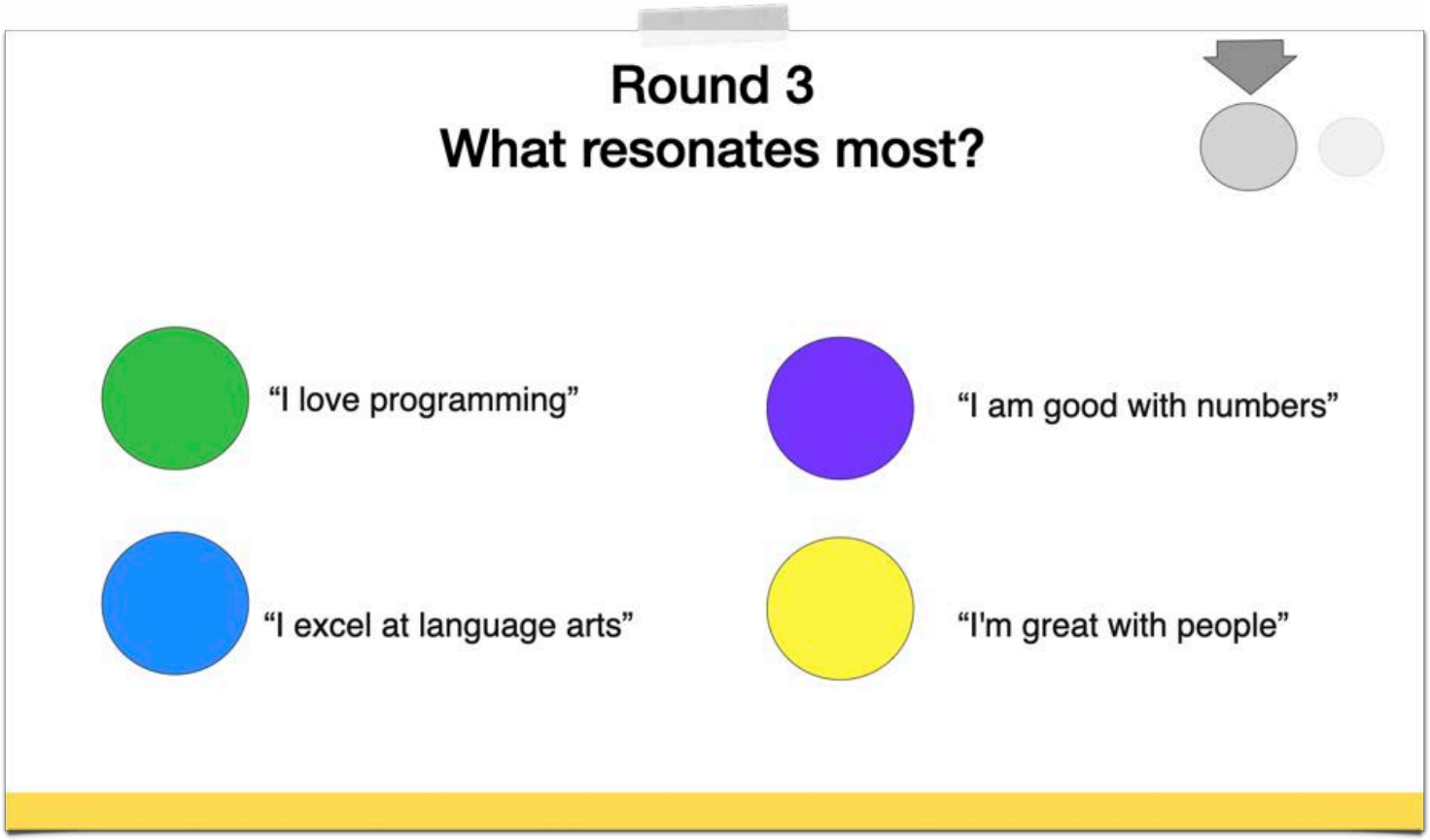
Inspiration from the Ice Breaker



Inspiration from the Ice Breaker



Inspiration from the Ice Breaker



Collective Findings

Inspiration from the R Code

Functions are useful!

Possible Sources of Errors

gsub(pattern, replacement, x)

- Punctuation
- Spaces
- Trimming beg/end spaces
- Replacing values (ex. Emoticons, abbreviations, jargon, etc)
- Encompassing regex values

grep(pattern, x)

- Truncating and use of wildcards

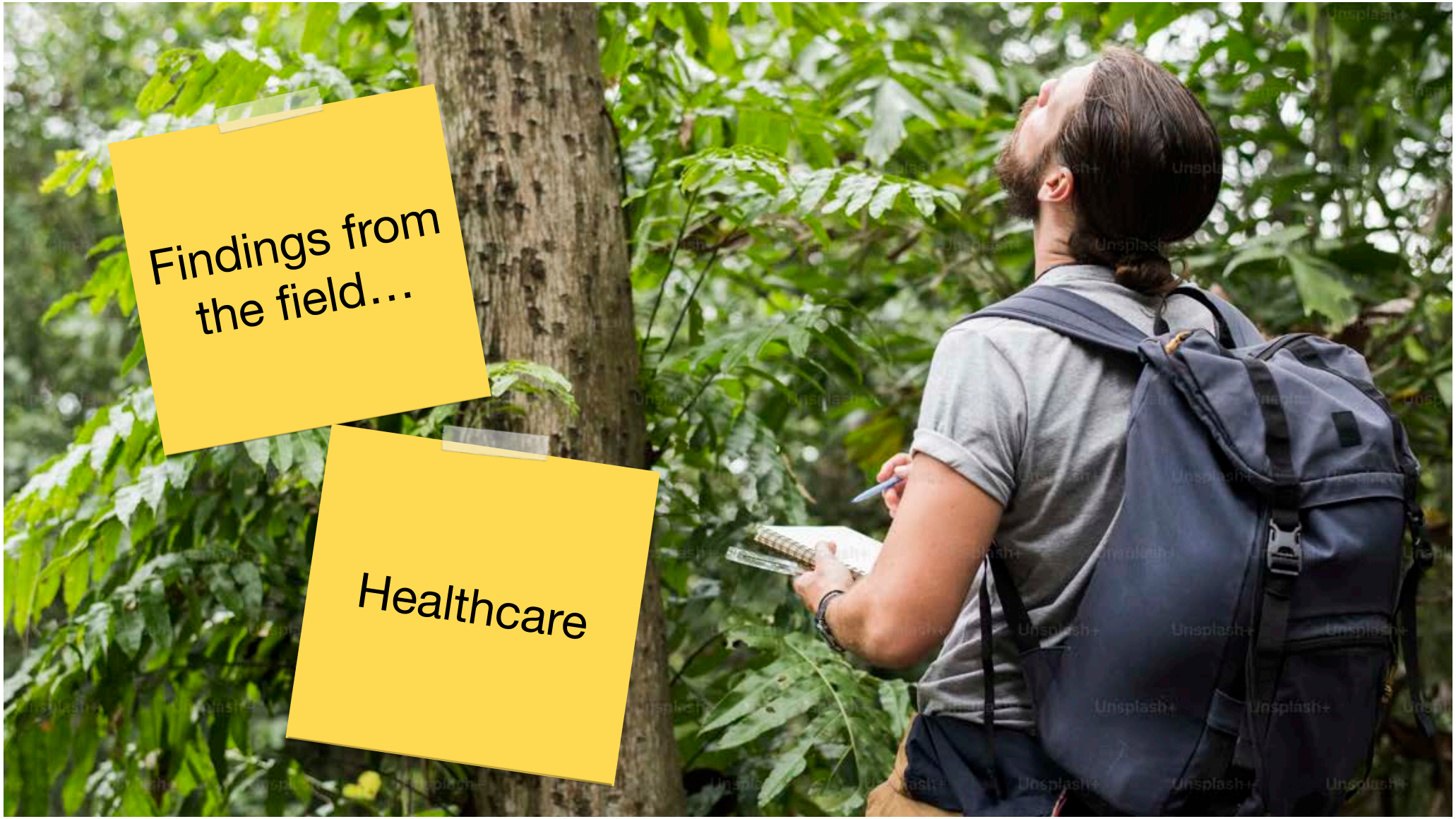
tolower(text_data)

- Upper and lower case normalization

Inspiration from the Reviews

Common Error Types

Jargon	Homographs	Cultural Reference	N-grams
Dialects	Outdated Dictionary	Typos/Misspellings	Cross-cultural communication
Idioms	Biases	Connotation	Emoticons
Neologisms	Context-Specific	Socio-Political Context	Punctuation and Upper/Lower Case
Homonyms	Registers	Niche words / Borrowed words	Truncation


Findings from the field…
Healthcare

Although we manually review program feedback every week, it can be daunting to accurately summarize these themes in annual reports.

What are participants talking about? What are recurrent discussion topics common amongst participants?

What needs emerge? Are there new areas worth developing in our program, based on discussion trends?

What can we do to help participants? Are participants struggling in specific areas?

HOW MIGHT WE KNOW AT A GLANCE THE TOPICS DISCUSSED BY PARTICIPANTS IN OUR PROGRAM REVIEWS?

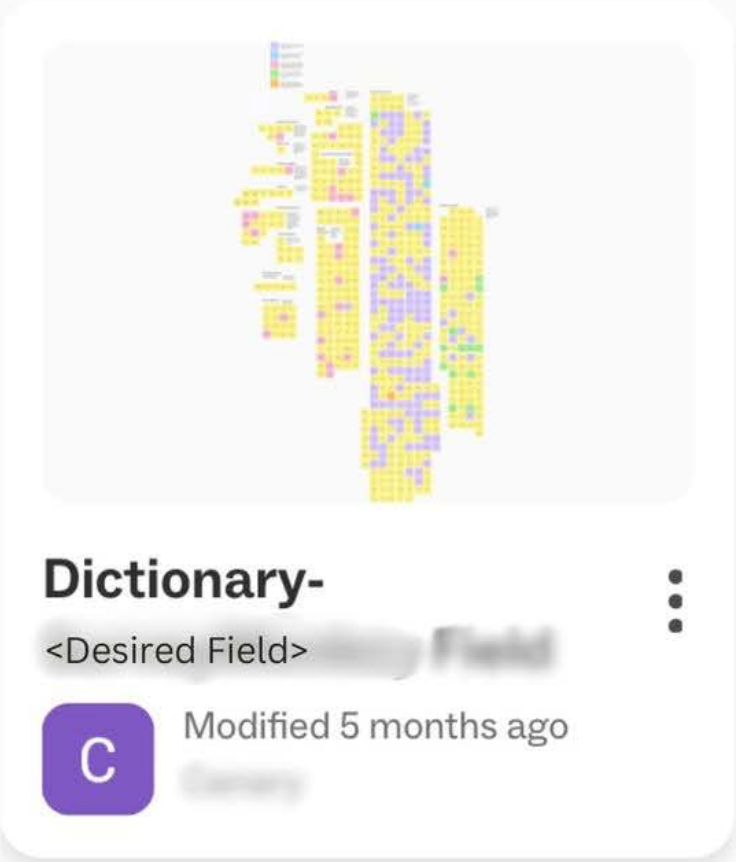
Using Inductive Logic

1. Starting from the data generated by participants, I do not force a theoretical construct upon their words.

2. This approach allows me to remain flexible, as more participants come through the program. New categories can be identified and added to the dictionary.

3. Using a semantic approach allows the treatment of a large set of data relatively fast. It is a reliable method that is less laborious than manual coding.

CLUSTERING VOCABULARY



Creating Categories

I used an inductive approach to create categories for the dictionary. Looking at the collected data, I clustered terms into clusters based on common themes. This activity was manual, as it relied heavily on the precise use of each of these terms in the contexts being discussed by participants. I privileged a manual approach to classification for this portion of the project, because of the nature of our project (see limitations of Commercial Dictionaries in the slides above).

* Uses blurred data to protect confidentiality

LIMITATIONS : CLASSIFYING USING GROUNDED THEORY

Interviewee's Knowledge

When looking at the text provided by participants we are limited to their knowledge in terms of topics chosen for discussion. This means omitted themes we might have liked to observe, but which were left undiscussed by some participants.

Interviewee's Biases

The texts collected from the survey may reflect an interviewee's own biases or expectations. While these are precisely what we seek to capture in some instances, it may be difficult to interpret responses in other circumstances, as we lack knowledge of the interviewee's views or background.

Refreshing Classifications

Due to the fact the dictionary is based on collected data, the dictionary will need to be updated with incoming data, to maintain its accuracy and completeness over time. This is especially true if a different population joins (ex. English as second language, paid vs family caregivers, etc.). This is because lexical terms are often connected to participants' backgrounds.

Frequency Approximations

Frequencies are an approximate measurement. If a participant mentions a topic multiple times, it may not necessarily be representative of "a more important topic". Frequency measures must therefore be taken with caution. Looking at document-level frequency can partly help.

Creating our Custom Dictionary

Classifying unstructured data from participant surveys

DATA COLLECTION	TEXT PROCESSING	TEXT ANALYSIS	CATEGORIZATION	DICTIONARY CONSTRUCTION
• Caregiver survey • Data cleansing	• Tokenizing punctuation • Removing stop words • Stemming	• Text statistics	• Clustering words into categories	• Constructing categories-based dictionary in R language

Limitations of Commercial Dictionaries

1. Since we service people from different cultural backgrounds, there were often words inspired by other languages. Speaking foreign languages myself, I was able to capture their meaning and add them to the dictionary, despite them not being English words.

2. Common typing errors or spelling mistakes occurred, which I also chose to include in our custom dictionary.

3. Certain words were used as a connotation, rather than their denotation. Others were consistently used in a specific context, giving them a special meaning. I categorized these words accordingly, and added them to a list. This list is meant to verify the words in context the next time the dictionary gets updated with new data, ensuring consistency.

Anticipating Errors

Better coding for in-house textual analysis tool

Critical Skills

Modeling Skills