

Understanding & Designing for Data Literacies in Online Communities

Ruijia “Regina” Cheng
HCDE PhD candidate
April 26th, 2022

Committee: Jennifer Turns, Benjamin Mako Hill,
Sayamindu Dasgupta, Amy Zhang



Roadmap

- Motivation & Background
 - Dissertation Overview & Anticipated Contributions
- Existing and Ongoing Work:
Formative case studies
 - Study 1 - 3
- Proposed Work:
Design-based Research
 - Study 4 & 5
 - Timeline & Anticipated Chapters



Motivation & Background

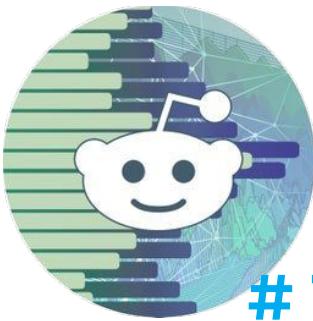


The abilities to understand and work with data are important for everyone.

Only a small proportion of people have access to formal training.

Working with data requires a range of skills, yet only a small part is taught in formal education.

Can we find an accessible context for the learning of multi-faceted skills to work with data?



TidyTuesday

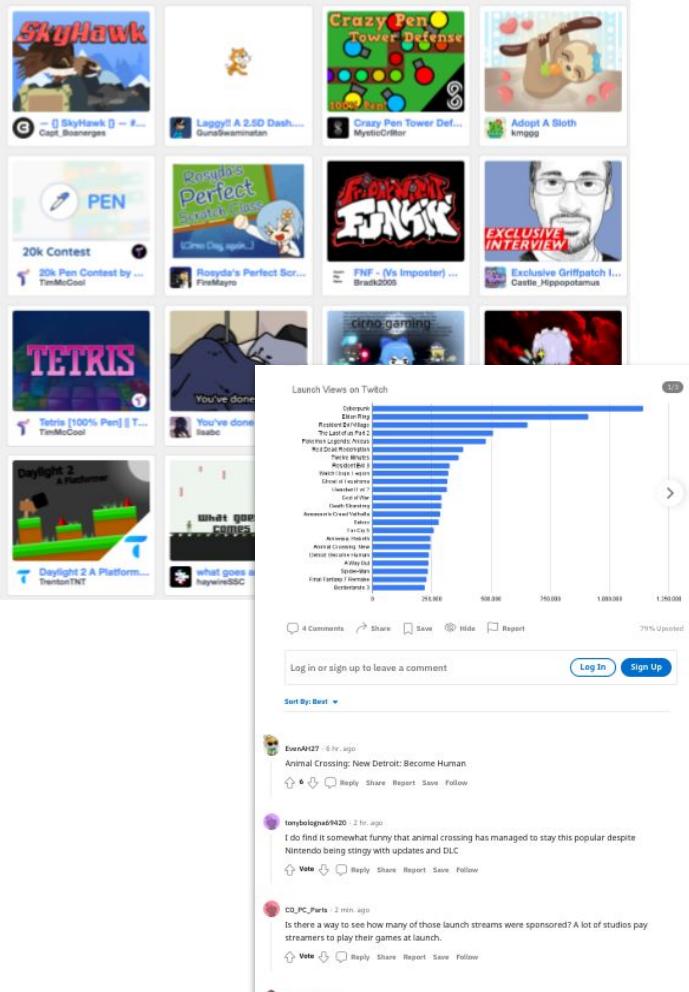
Communities of Practice

Community of Practice (CoP):

- A group that "share a concern or a passion for something they do and learn how to do it better as they interact regularly" (Wenger-Trayner 2015)
- Learning through Legitimate Peripheral Participation (Lave & Wenger 1991)

Online CoPs:

- Affinity spaces (Gee 2005)
- Distributed mentoring (Campbell et al. 2016)
- Plural learning pathways (Cheng & Hill 2022)



Computational participation

- Computational learning via making and sharing projects in socially supported context (Kafai 2016)
- Online communities for computational participation (e.g., Bruckman 1998, Resnick et al. 2009, Fiesler et al. 2017)

Online communities can potentially be a good learning context, but how well does it support data literacies?

What do youth data literacies mean to you?



"Data Literacies"

Data literacy -> Data literacies

My initial attempt to define data literacies:

- Technical: choose and apply relevant programming and statistical tools to work with data
- Discursive: recognize, reflect and properly communicate the process of working with data
- Critical: interpret and argue for the impact of data in a larger social context

My dissertation will add:

Empirical knowledge
of how people
develop data
literacies in online
communities & the
challenges they face



Theoretical frameworks
that describe:
- Different kinds of data
literacies
- How features of online
communities impact
each type



Design knowledge &
prototype examples of
how to promote
data literacies in
online communities



Dissertation Overview

Formative case studies



RQ1: How do online communities support learners to develop a range of skills related to data literacies?



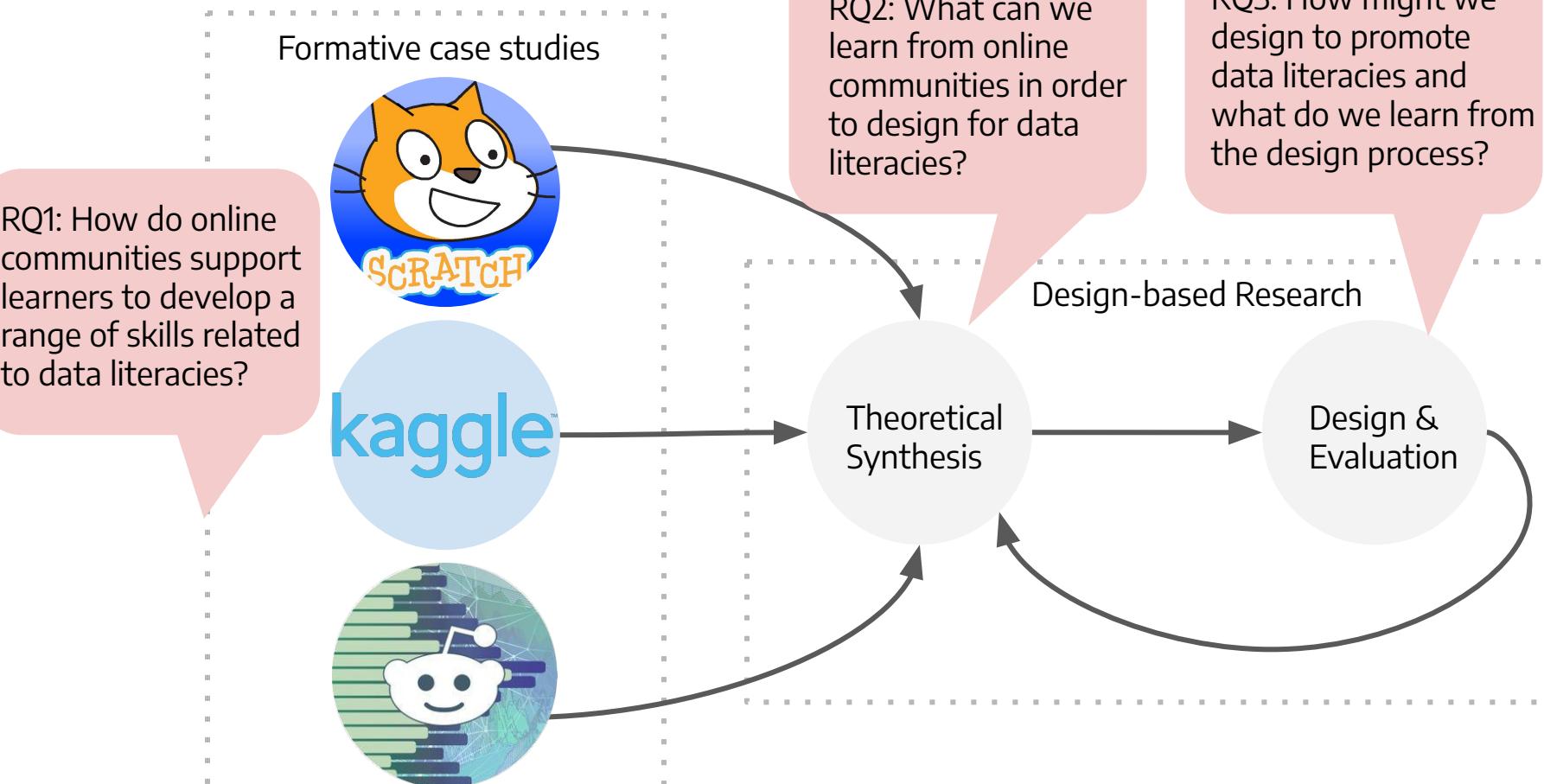
RQ2: What can we learn from online communities in order to design for data literacies?

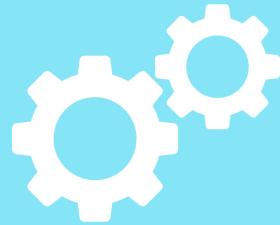
RQ3: How might we design to promote data literacies and what do we learn from the design process?

Design-based Research

Theoretical
Synthesis

Design &
Evaluation





Existing & Ongoing Work:

Formative Case Studies of 3 Communities



Study 1

How Scratch users learn data structures in the community?

Published in CHI 2022

Cheng, R., Dasgupta, S., Hill, B. How Interest-Driven Content Creation Shapes Opportunities for Informal Learning in Scratch: A Case Study on Novices' Use of Data Structures. 2022. Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2022).

Community: Scratch

Scratch Online community

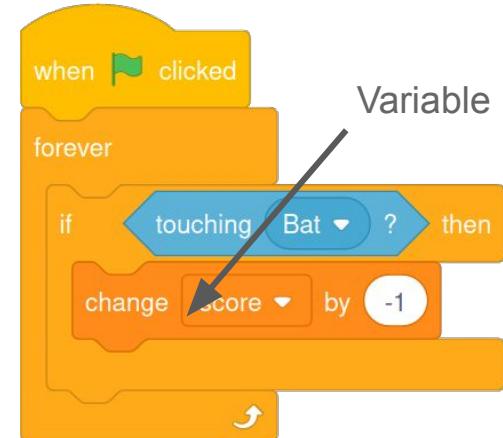
- Scratch: block-based programming language for novices
- Largest interest-driven computational learning community

The image shows three main sections of the Scratch website:

- Homepage:** Features a banner with "Create stories, games, and animations Share with others around the world". It includes icons for a cat, a blue character, and a sun, with links to "TRY IT OUT", "SEE EXAMPLES", and "JOIN SCRATCH (It's free)". Below this, it says "A creative learning community with 4,608,204 projects shared" and links for "ABOUT SCRATCH | FOR EDUCATORS | FOR PARENTS".
- Project Editor:** Shows a project titled "The Water Cycle!" with a script of orange blocks and a stage showing a landscape with a volcano and a sun.
- Forum:** A "Welcome to Scratch" forum with 304 posts, 119340 topics, and the last post from "Today 06:25:33 by". It also includes sections for "Announcements", "New Scratches", and "Help With Scripts".

Scratch Data structures

- Scalar variables and lists
- < 15% users have ever used data structures



Method: Mixed-method analysis

Study 1: Qualitative analysis of Scratch forum discussions on data structures

- Sampled 400 threads
- Grounded theory analysis

Proposed Theory

Study 2: Quantitative analysis using large-scale data on Scratch projects and user activities

Hypothesis testing analyses on 5 years of Scratch data:
(241,634 projects created by 75,911 users)

Key Findings

Novices use specific use cases to teach each others about variables and list

- Many are game specific

User generated tutorials are often framed in game-making scenarios

"Create a Variable with the name **lives** so that every time the main character touches a ghost, it would lose one life."

Key Findings

Certain examples of how to use data structures become archetypes, which can constrain innovative usage

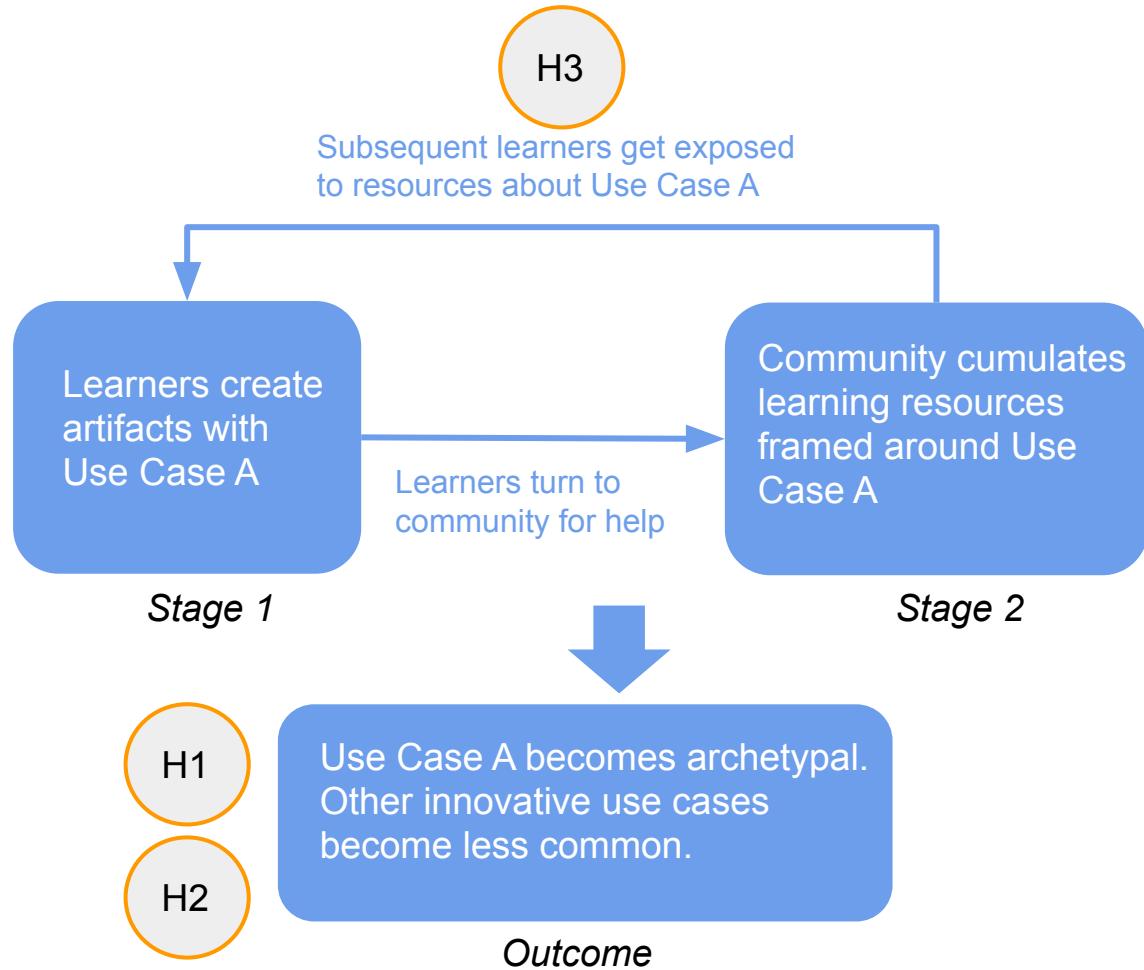
Q: "I want to use variables in my project but I don't know how?"

A: "Here is an example of how to make a **score counter** in your game!"

Q: "But my project is not a game. I am making a storytelling project..."

Key Findings

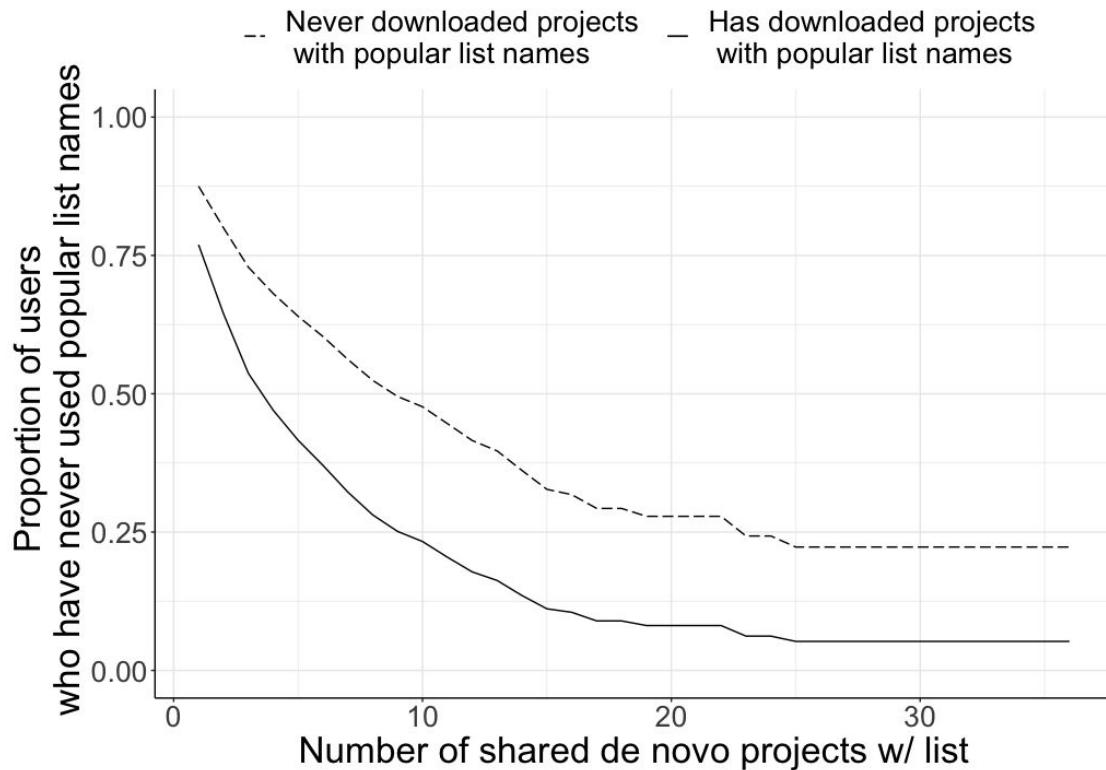
A social feedback loop of how community-generated resources may constrain innovative computational participation



Key Findings

H3: Users who have been exposed to popular variable and list names will be more likely to use those names in their own projects

(Partially supported by
Projects with List)



Main Takeaway

Trade-off of online community for learning of technical concepts:

- User-generated explanation and understanding that fit common interests
- Risk of superficial, restricted learning resources



Study 2

How and why Kaggle users share procedures and rationales of data analysis in the community?

Published in CSCW 2020

Cheng, R., Zachry, M. Building Community Knowledge in Online Data Science Competitions: Motivation, Practices and Challenges. 2020. Proceedings of the ACM Human Computer Interaction, Computer Supported Cooperative Work and Social Computing (CSCW 2020).

Community: Kaggle

Largest data science competition platform

Health Insurance prediction using TensorFlow
Python notebook using data from [multiple data sources](#) • 20 views • 2h ago • insurance, tensorflow, keras, +1 more

RandomOverSampler and SMOTE(Synthetic Minority Oversampling Technique) are used to treat imbalanced datasets (which is the case here as we can see in the first figure).

RandomOverSampler duplicates the minority class data until minority class data reaches specified proportion of majority class data.

SMOTE generates synthetic data of minority classes and ensures that the data doesn't overfit

```
In [1]:  
import numpy as np # linear algebra  
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)  
import csv  
import tensorflow as tf  
from tensorflow.keras.preprocessing.image import ImageDataGenerator  
import seaborn as sns  
import matplotlib.pyplot as plt  
import matplotlib as mpl  
import os  
import tempfile  
import sklearn  
from sklearn.model_selection import train_test_split  
from sklearn.model_selection import cross_val_score  
from sklearn.preprocessing import StandardScaler  
from sklearn.linear_model import LogisticRegression  
from sklearn.metrics import accuracy_score, f1_score, confusion_matrix  
  
from imblearn.over_sampling import RandomOverSampler,SMOTE  
from imblearn.under_sampling import RandomUnderSampler  
  
import plotly.express as px  
from plotly.subplots import make_subplots  
import plotly.graph_objs as go  
  
In [2]:  
mpl.rcParams['figure.figsize'] = (12, 10)  
colors = plt.rcParams['axes.prop_cycle'].by_key()['color']
```

Notebook (publicly shared executable code)

Abhishek Bhat • (1008th in this Competition) • a day ago • Options • Reply

Similar to what you have mentioned in point-1, I tried blending models by finding optimal weights using OOF predictions. Blending gives a slight boost to the CV.

StackNet might be tricky to implement here given that most of the conventional models are not doing well in this problem and its also going to be a challenge to find a lot of uncorrelated models. Any thoughts on this?

Trigram • (534th in this Competition) • 19 hours ago • Options • Reply

most of the conventional models are not doing well in this problem

Since most people are using NNs or a variant, I doubt they'd be much interested in using StackNet. LightGBM in a public kernel was pretty decent, but haven't seen much of LGB after that.

Uncorrelated models too might be a bit of a challenge, so you'll need to run a lot of (computationally expensive) experiments in order to get the best possible models from your experiments.

Mehrdi Gakhramanian • (121st in this Competition) • 2 days ago • Options • Reply

Thank you for sharing.
In my previous competition, I used <https://www.kaggle.com/mekhdigakhramanian/post-processing-v2> with public submission outputs and 3 my own kernels but it was overfitted, be careful

Ajay Chaudhary • (105th in this Competition) • 3 days ago • Options • Reply

I tried out stacking but notebook ran out of time. Any results you want to share?

Discussion

		251 Grandmasters	1,744 Masters	7,712 Experts	67,916 Contributors	96,702 Novices
Rank	Tier	User	Medals	Points		
1		Dieter	joined 4 years ago	27 12 3	244,490	
2		bestfitting	joined 6 years ago	37 11 1	223,350	
3		Guanshu Xu	joined 6 years ago	23 20 2	201,534	
4		Psi	joined 10 years ago	25 7 0	183,315	
5		Qishen Ha	joined 7 years ago	17 8 2	143,571	
6		Giba	joined 10 years ago	59 47 31	140,706	
7		ilu	joined 3 years ago	14 6 3	122,625	
8		Ahmet Erdem	joined 6 years ago	21 25 4	122,426	

Public user ranking

Method:
Interview with 14
Kaggle users of a
various experiences in
data science

Key Findings

Experts share procedures and rationales that will be used and appreciated by other experts.

Experts' "niches": partial procedures and convoluted discussions.

"The snippets are better because they are focused into one topic... I don't want end-to-end solutions because it's kind of a copy paste." (E4)

"Use notebooks to communicate ideas on an inspiration level, but not submittable solutions." (E2)



Key Findings

Beginners are left out - challenge understanding and participating in experts' niche.

They also get little attention and encouragement for sharing their procedure and rationales.

"Some people only put a snippet, then sometimes I don't exactly understand what they did after that... and When I want to so the same thing, it just doesn't work." (B8)

"Here's always some anxiety that everyone else is more experienced at this, and they're gonna think my post is stupid." (B2)



Main Takeaways

Community recognition can potentially encourage members to share procedures and rationales in data analysis.

However, experts and beginners have very different experience and practices in sharing about analysis in the community.



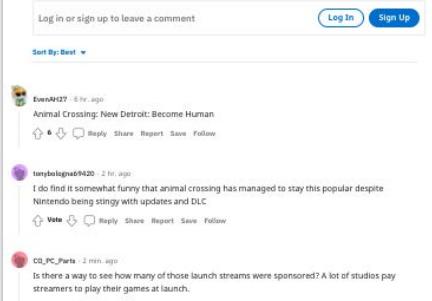
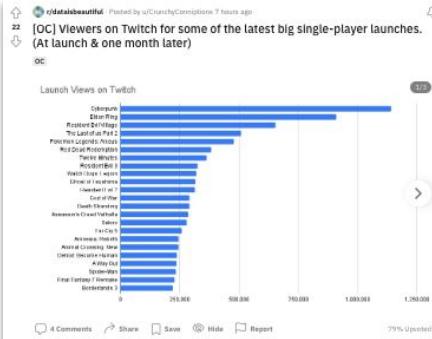
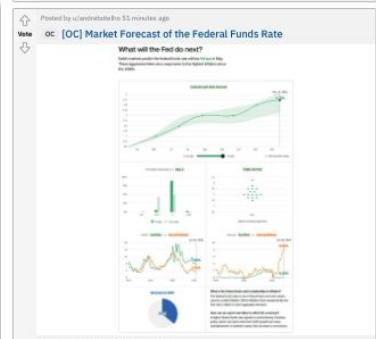
Study 3

How interaction with community members shapes how people present, interpret, and argue with data.

Ongoing

Community: r/dataisbeautiful

Largest community dedicated to data visualizations on Reddit



Method: Interview with 22 members

- How and why users share their insights with data
- How and why they engage with other members
- What they learn about data science, domain topics and the community in the process.

Finished data collection and initial grounded theory analysis

Key Findings (tentative)

How to make visualizations & how to make visualizations *in and for* the community.

- Make sense of the community and social impact.
- Tailoring for audience's needs and interests.

Emerging trade-offs:

Information consumption experience vs critical exchanges and learning.

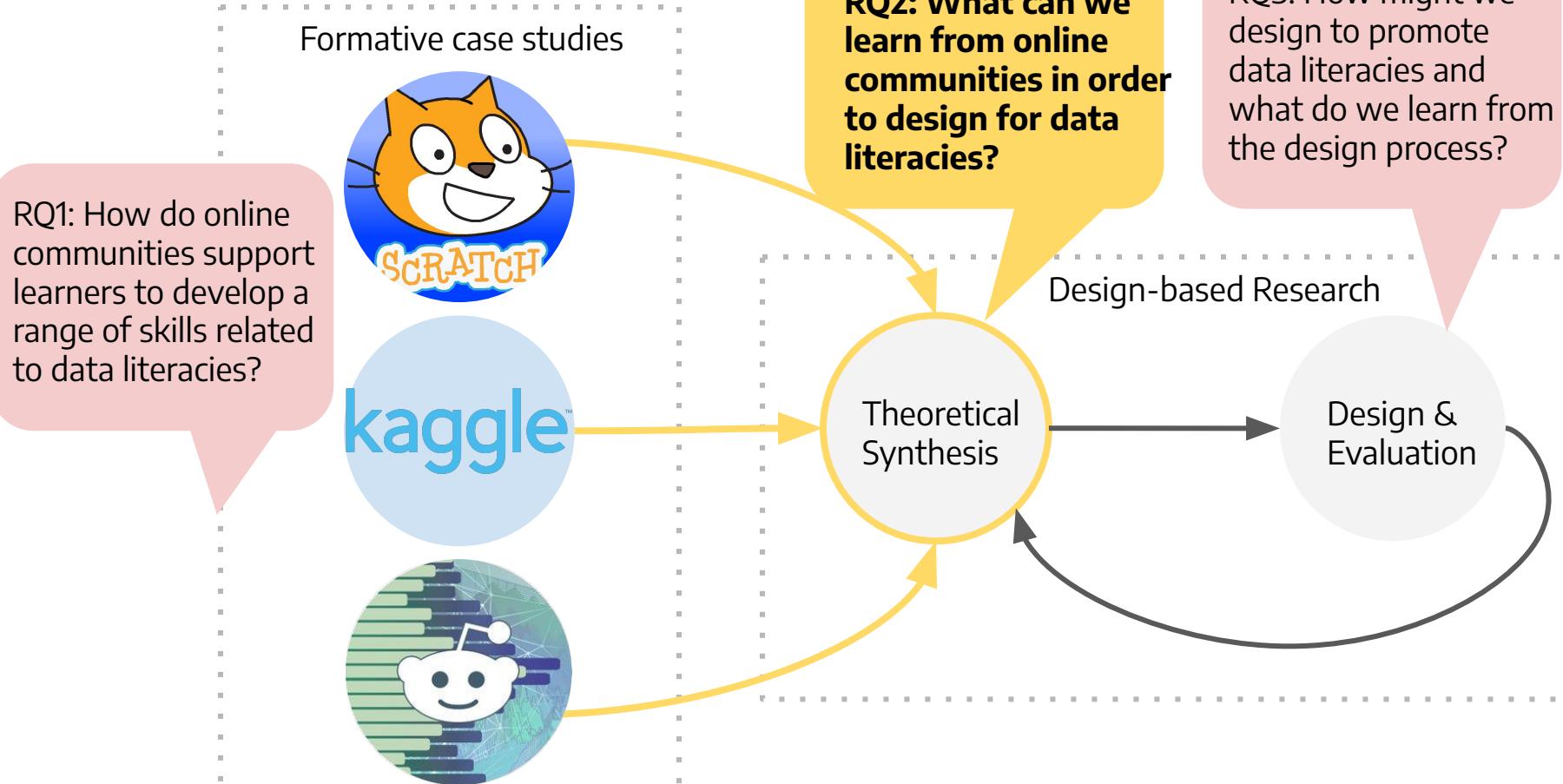
Awareness of audiences can constrain how learners work with and present data.



Proposed Work:

Design-based Research

Dissertation Overview



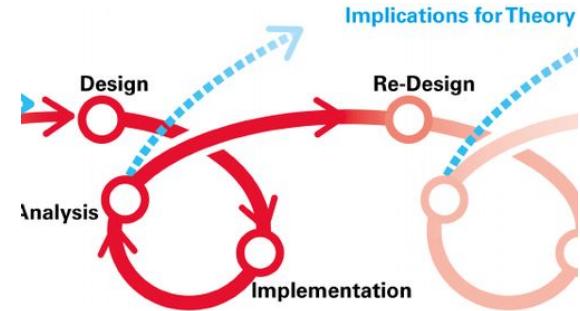


Study 4

Theorizing Data Literacies in
Online Communities

Design-based Research (DBR)

- A methodology in learning science
- Relationships b/w theories, design, and empirical evidences of learning
- Iteration



Conjecture maps

- An analysis instrument of DBR
- Developing theoretical conjectures:

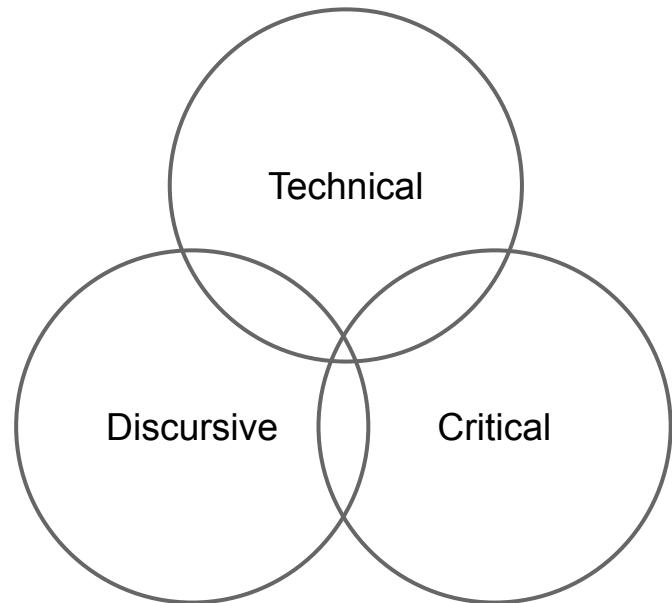


Theorizing Data Literacies in Online Communities

Step 1: Build a framework of data literacies in online communities:

- Extend the three types of data literacies
- Add empirical evidence from formative studies
- Grounded with literature on data science education and workplace

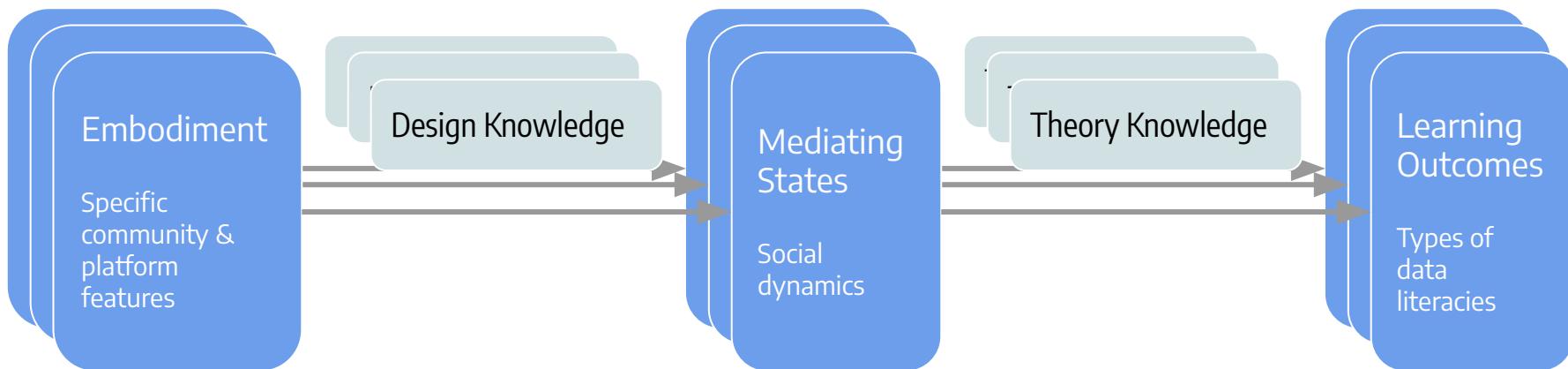
Desired framework: definition, skills and competencies, theoretical grounding, and connection to online communities.



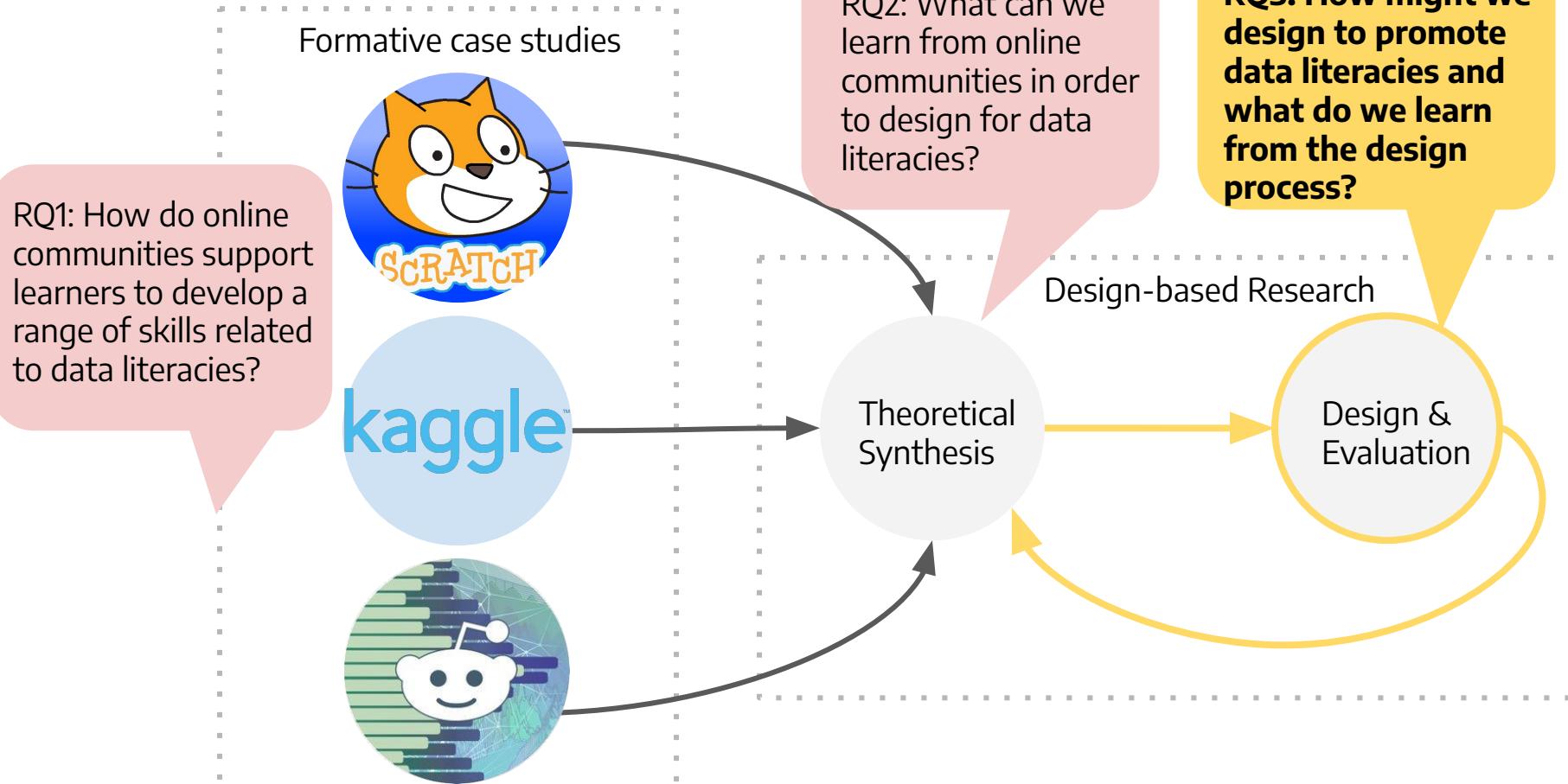
Theorizing Data Literacies in Online Communities

Step 2: Conjecture Maps

- Mechanisms of how features of online communities support or not support the different types of data literacies.
- Output: design directions



Dissertation Overview





Study 5

Design, evaluation, and iteration
on theories

- Design and build 3-5 prototypes
- Use the prototypes as elicitations to conduct interview-based user studies.

Design

- Specific design directions will be based on the results of theoretical synthesis.
- Designs can range in fidelity and forms.
- Some examples of design directions:

Extension to Jupyter notebooks that prompts users to enter their rationales and decisions in data analysis and automatically forms a story that users can edit and share in online communities.

Online discussion interface with multiple parallel comment sections tailored for different purposes under a single post of data visualization.

A bot that identifies and explains underlining programming and statistical concepts that appear in posts about data exploration.

User studies

Recruitment:

- 20 - 30 participants
- Must have experience sharing or data analysis projects online and/or learning from others' projects.
 - A balanced mix of skill levels.
- Recruitment from multiple channels:
 - Online data hobbyist communities
e.g., Kaggle, Tableau Public, and r/dataisbeautiful.
 - Classes such as HCDE 410, HCDE 530, and DATA 512.
 - Previous participants of formative research.



User studies

Procedures:

Pre-interview survey

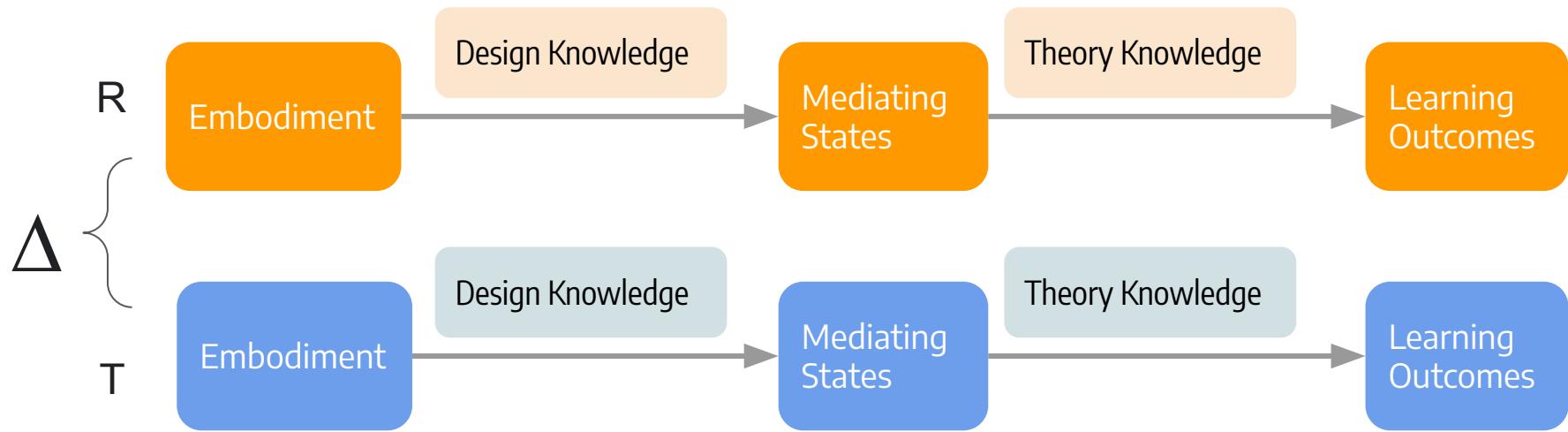
- Share a most memorable project that they posted or learned from in online communities
- Questions about the experience and challenges

Semi-structured interviews

- Online, 90 mins long, video audio-recorded
- Interact and think-out-loud with each of the prototypes.
- Imagine a scenario where they could use the prototypes for the project shared in survey.

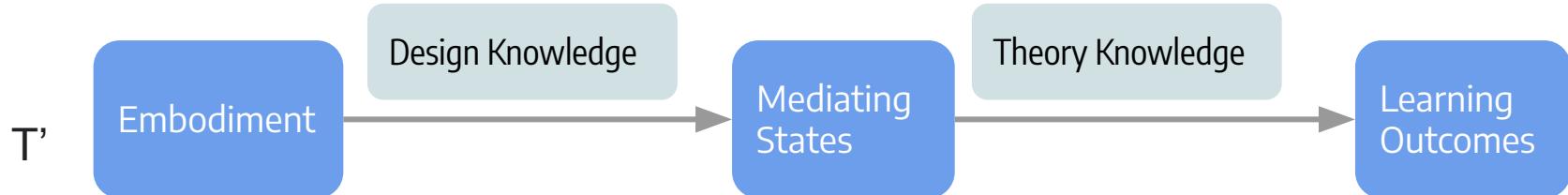
Analysis and Iteration on theories

- Grounded Theory analysis on interview transcripts
- New “real” conjecture maps (R), compared to the theoretical maps (T)



Analysis and Iteration on theories

- Grounded Theory analysis on interview transcripts
- New “real” conjecture maps (R), compared to the theoretical maps (T)
- Refine the theory (T')

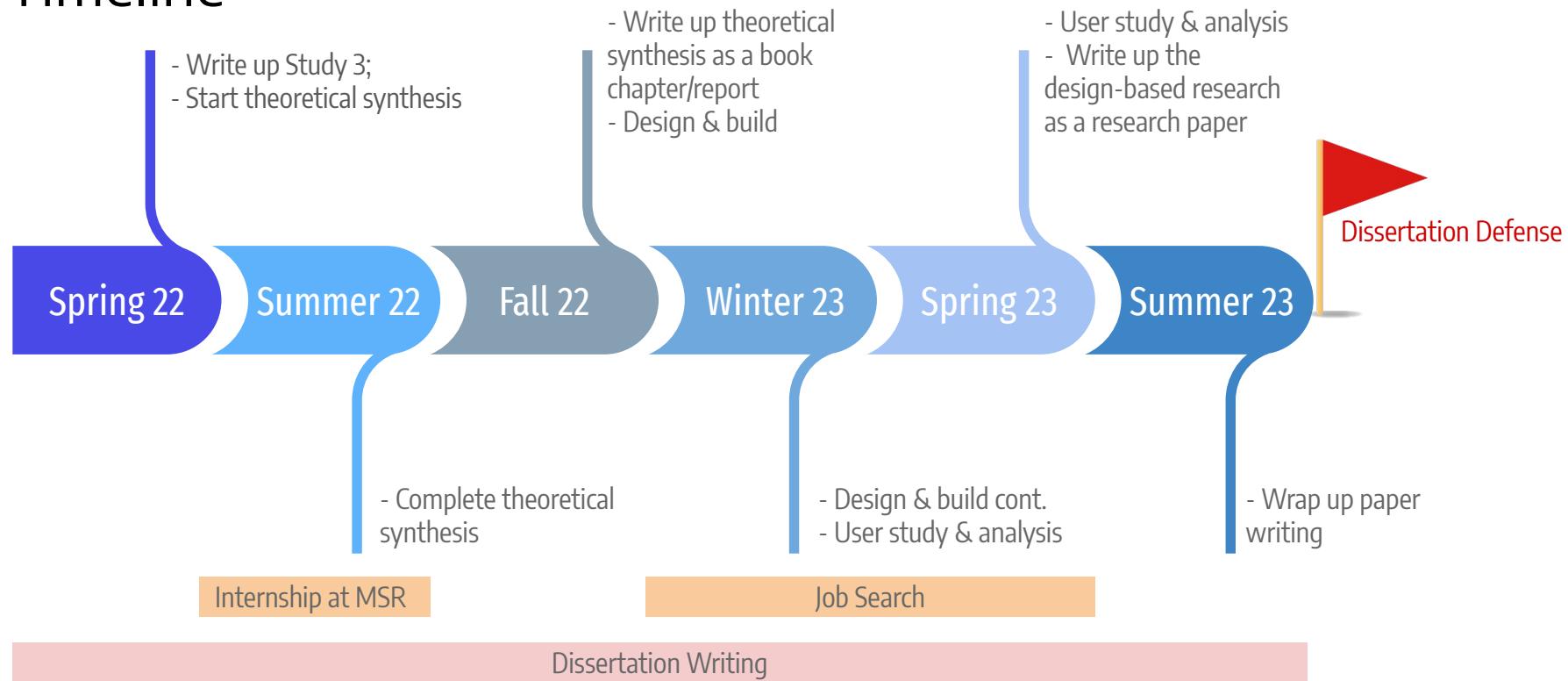


Anticipated Chapters

- Chapter 1: Introduction.
- Chapter 2: Case Study 1 - Learning about Data Structures in Scratch.
- Chapter 3: Case Study 2 - Sharing Analysis Procedures in Kaggle.
- Chapter 4: Case Study 3 - Representing and Arguing with Data in r/dataisbeautiful
- Chapter 5 - Theorizing Data Literacies in Online Communities.
- Chapter 6 - Design, Evaluation, and Iteration on the Theories.
- Chapter 7 - Implications and Conclusion.



Timeline



Acknowledgements

My advisors, committee members & mentors, my co-authors, humans of CDSC & HCDE, DRG students, my friends, my partner, and my parents - thank you all!

