Hypothesizing and Evaluating GenAI in Your Newsroom

SRCCON 2023

**(aka:
My boss thinks AI is cool. What do I do now?)**

Who we are

Eric Ulken

Vice president of product,
The Baltimore Banner

Building sustainable business models for local news through technology

Sachita Nishal

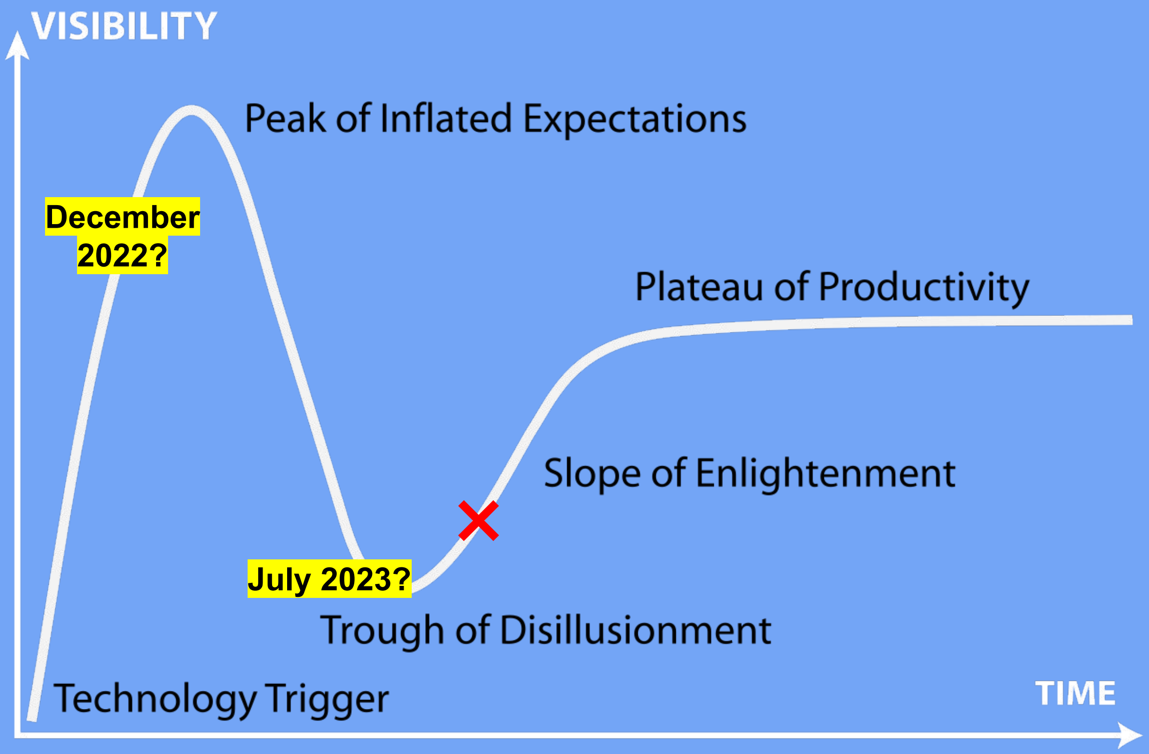
4th year PhD student @ Northwestern
Communications & Computer Science

Designing and studying AI tools to support journalists

Goals for this session

- Develop testable hypotheses around AI use cases
- Consider human factors of interacting with these tools
- Determine which applications are ripe for pursuit

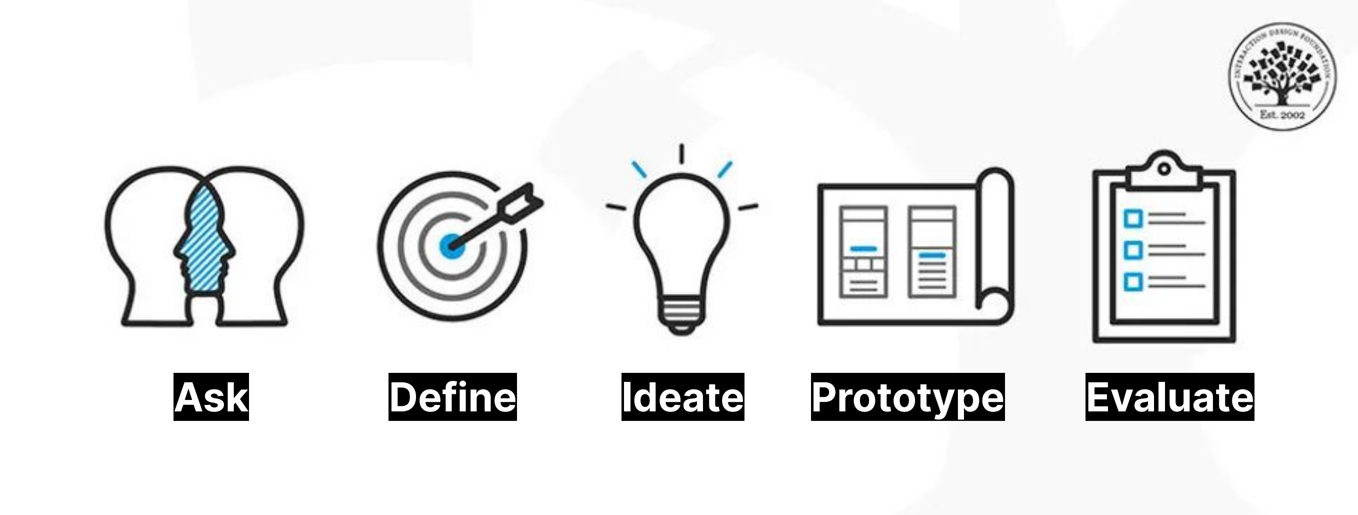
Where are we on the hype cycle?



“How can we use GenAI for something useful?”

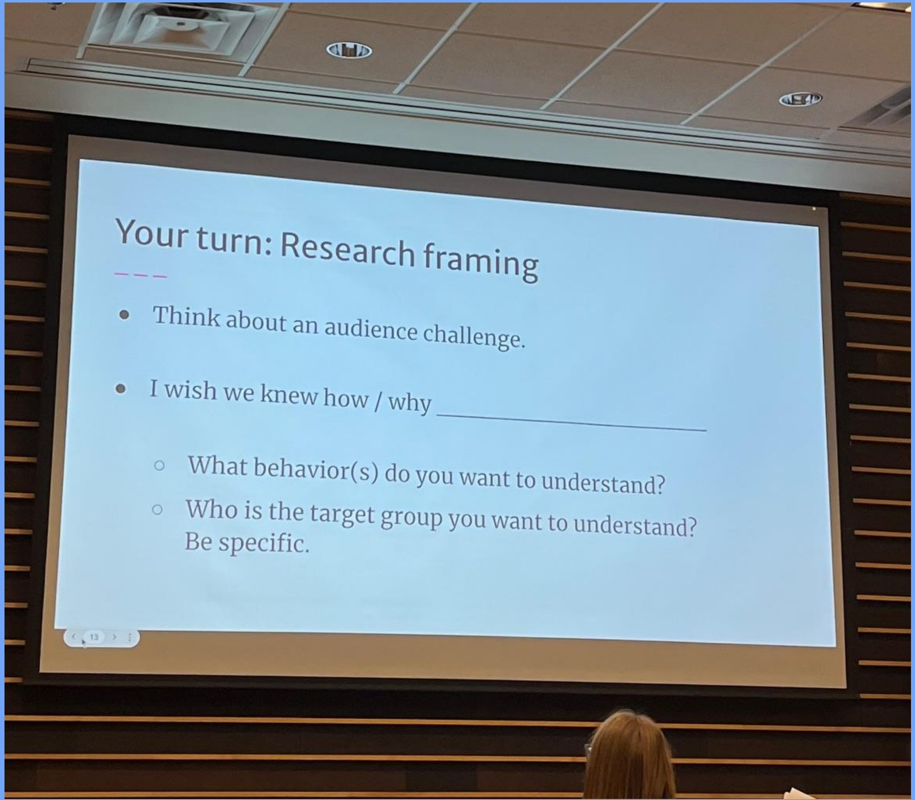
This has felt like a bit of a strange question to ask over the past year …

How we generally do this: human-centered design



Source: Interaction Design Foundation
(https://www.interaction-design.org/literature/article/5-stages-in-the-design-thinking-process)

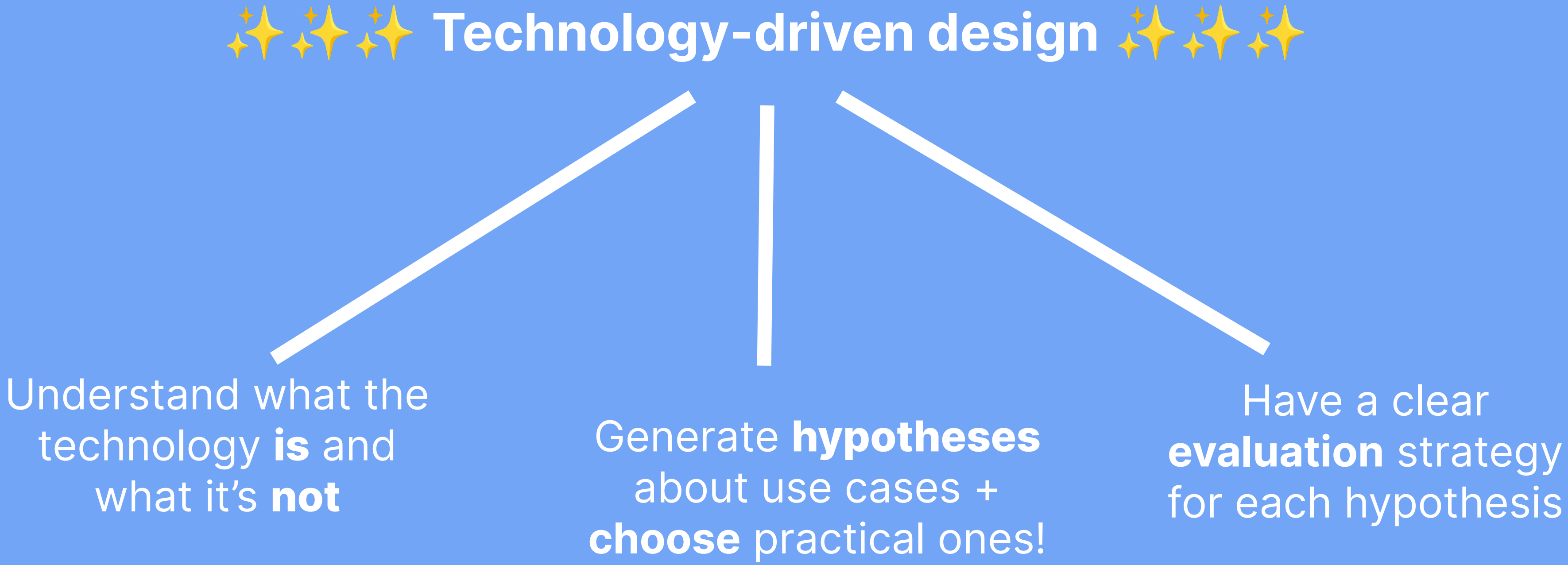
How we generally do this: human-centered design



Your turn: Research framing

- Think about an audience challenge.
- I wish we knew how / why ______________________
  - ○ What behavior(s) do you want to understand?
  - ○ Who is the target group you want to understand? Be specific.

Why generative AI feels different



✨✨✨ **Technology-driven design** ✨✨✨

Understand what the technology **is** and what it's **not**

Generate **hypotheses** about use cases + **choose** practical ones!

Have a clear **evaluation** strategy for each hypothesis

Let’s do a group exercise …

Grab some scratch paper, or you can use the session doc
Pick a task! And let’s generate some **hypotheses**!

We will also use a sample case to walk through the example

Categories/examples of use cases for AI in news

Newsgathering

→Surfacing leads from structured or unstructured text
→Generating news angles
→Summarize/QA dense data
→Transcription

Production

→Writing summaries of stories
→Editing partner
→Retrieve archive data that might be relevant

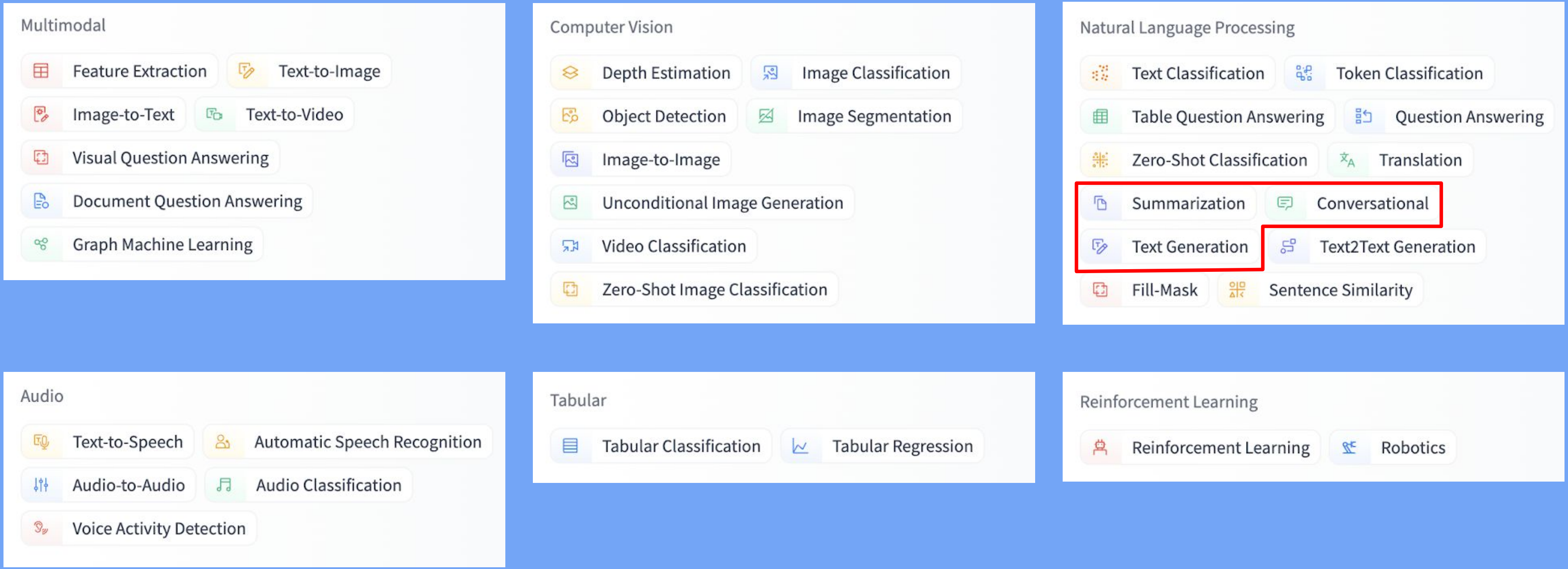
Dissemination

→Summaries for social media or SEO
→Personalize/localize studies
→Alt text

Collaboration and Coordination

→Summarize meeting minutes
→Deliver targeted analytics insights

Ways to do this



Multimodal

- Feature Extraction
- Text-to-Image
- Image-to-Text
- Text-to-Video
- Visual Question Answering
- Document Question Answering
- Graph Machine Learning

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Image-to-Image
- Unconditional Image Generation
- Video Classification
- Zero-Shot Image Classification

Natural Language Processing

- Text Classification
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Conversational
- Text Generation
- Text2Text Generation
- Fill-Mask
- Sentence Similarity

Audio

- Text-to-Speech
- Automatic Speech Recognition
- Audio-to-Audio
- Audio Classification
- Voice Activity Detection

Tabular

- Tabular Classification
- Tabular Regression

Reinforcement Learning

- Reinforcement Learning
- Robotics

Source: Hugging Face tasks taxonomy

[An attempt at] generating practical hypotheses

What **task** would it be used for?

→Specificity of task? General purpose use? More customizable?

What is the **goal** for this task?

→Supplement activity (e.g. glue work)?
→Expand capacity (e.g. summarization for scanning)?
→Add a new activity (e.g. new way to brainstorm)?

What is the level of **oversight** necessary?

→Converting data formats vs. generating an article summary
→Time needed for this oversight?

[An attempt at] generating practical hypotheses

What **task** would it be used for?

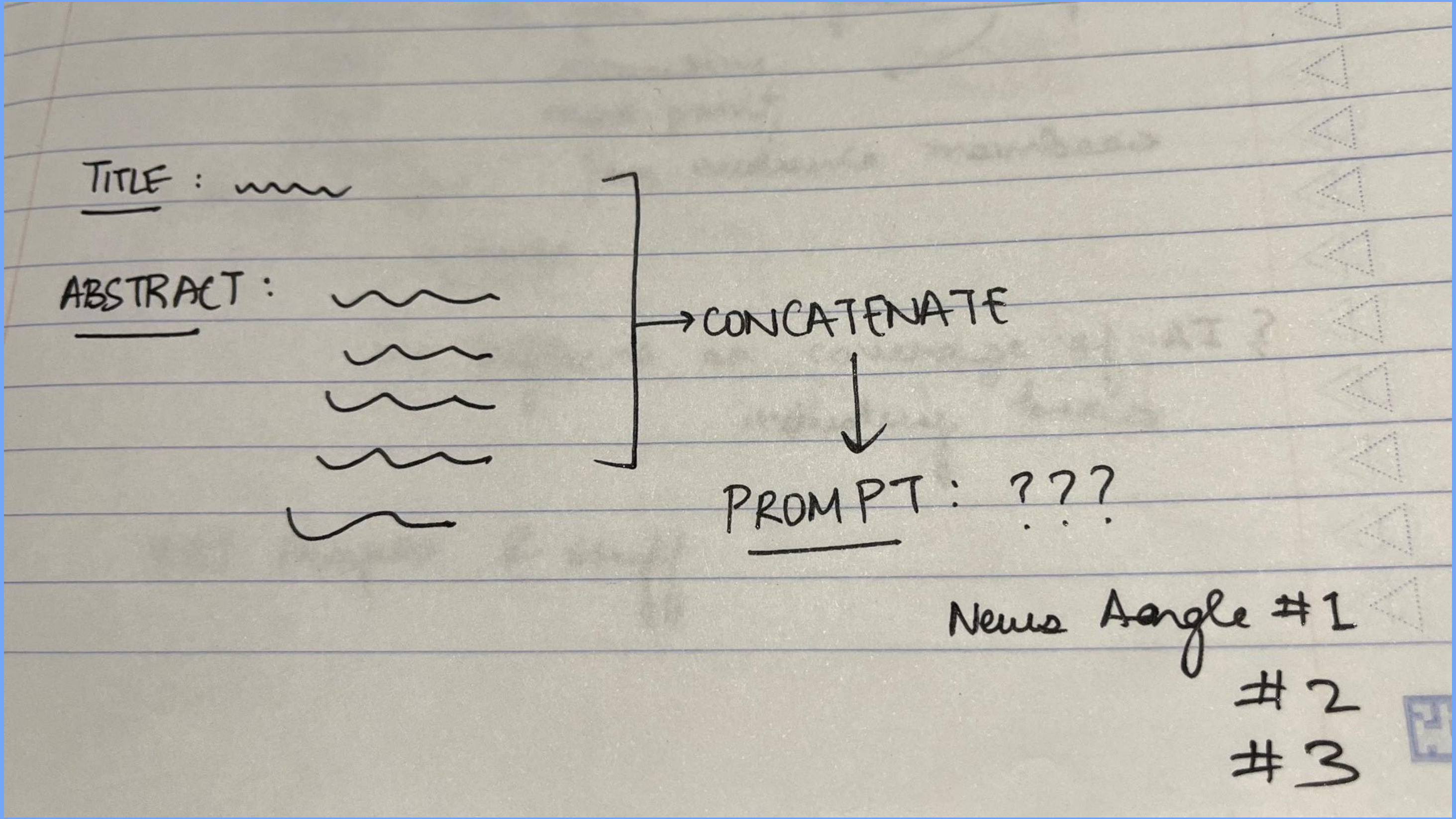
→Generate news angles from science papers, no personalization

What is the **goal** for this task?

→Expand capacity, allow people to filter and brainstorm without getting bogged down by jargon

What is the level of **oversight** necessary?

→Not too much honestly, although factuality is a must or time is wasted



Old sketch from when I started doing this

[An attempt at] generating practical hypotheses

Data **privacy** concerns, given this task + goal + oversight?
→Don’t feed it personal/confidential data

What are the stakes if it makes an error? **Margin** for error?
→C/f oversight that is needed + it will make errors, yes
→Cost of false positives vs. false negatives?

What **resources and training** will you/users need?
→Startup resources vs. longer-term needs

[An attempt at] generating practical hypotheses

Data **privacy** concerns, given this task + goal + oversight?

→Open-access, public data, not too worried about this

What are the stakes if it makes an error? **Margin** for error?

→Wasted time with false positives - how much?

→Risk of missing out on important info - how much?

What **resources and training** will you/users need?

→Documentation around LLMs+how to use

→Mechanisms to archive/bookmark things

→Report errors and mistakes

→Explanation???

[An attempt at] generating practical hypotheses

Would you need a specialised **model training/tuning/interface**?
→ UI? Sheets? In CMS?
→ What other tech stack? Scraping?
→ Why not **simpler, cheaper** NLP?

Push vs. pull type of system?
→Who seeks out whom? Alignment to time?

Communicating **uncertainty** with system output?
→Do you need to do this? Where would it help?

[An attempt at] generating practical hypotheses

Would you need a specialised **model training/tuning/interface**?

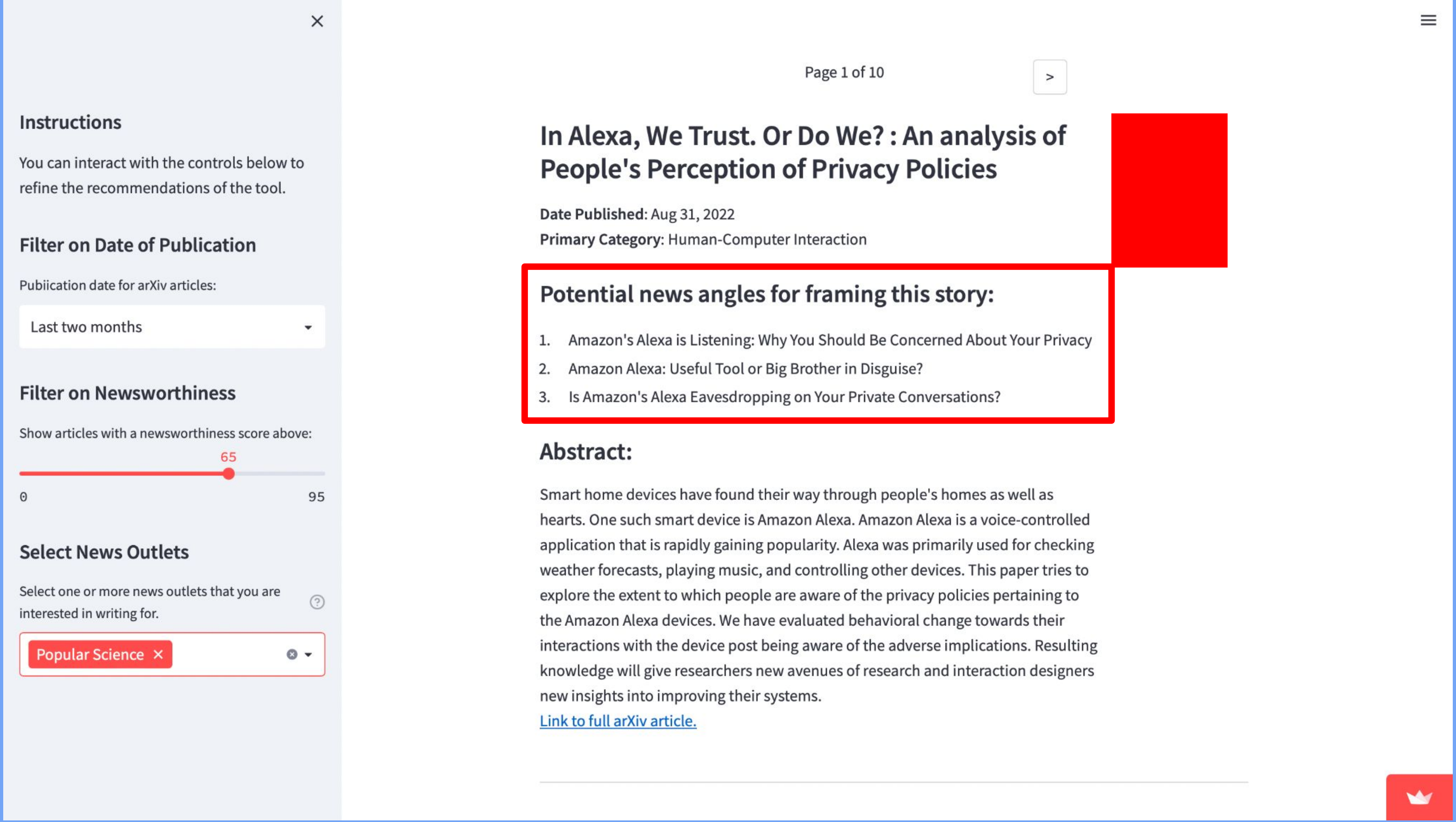
→Fine-tuning because science is a specialized domain
→Interface unnecessary, but allowed interoperability

Push vs. pull type of system?

→Pull system, unsure if reporters actually dig through individual preprints a lot

Communicating **uncertainty** with system output?

→Still very speculative when we were doing this, but if I were to do this again, I'd think about output probabilities for text



Instructions

You can interact with the controls below to refine the recommendations of the tool.

Filter on Date of Publication

Pubiication date for arXiv articles:

Last two months

Filter on Newsworthiness

Show articles with a newsworthiness score above:

65

0 95

Select News Outlets

Select one or more news outlets that you are interested in writing for.

Popular Science

In Alexa, We Trust. Or Do We? : An analysis of People's Perception of Privacy Policies

[redacted]

Date Published: Aug 31, 2022

Primary Category: Human-Computer Interaction

Potential news angles for framing this story:

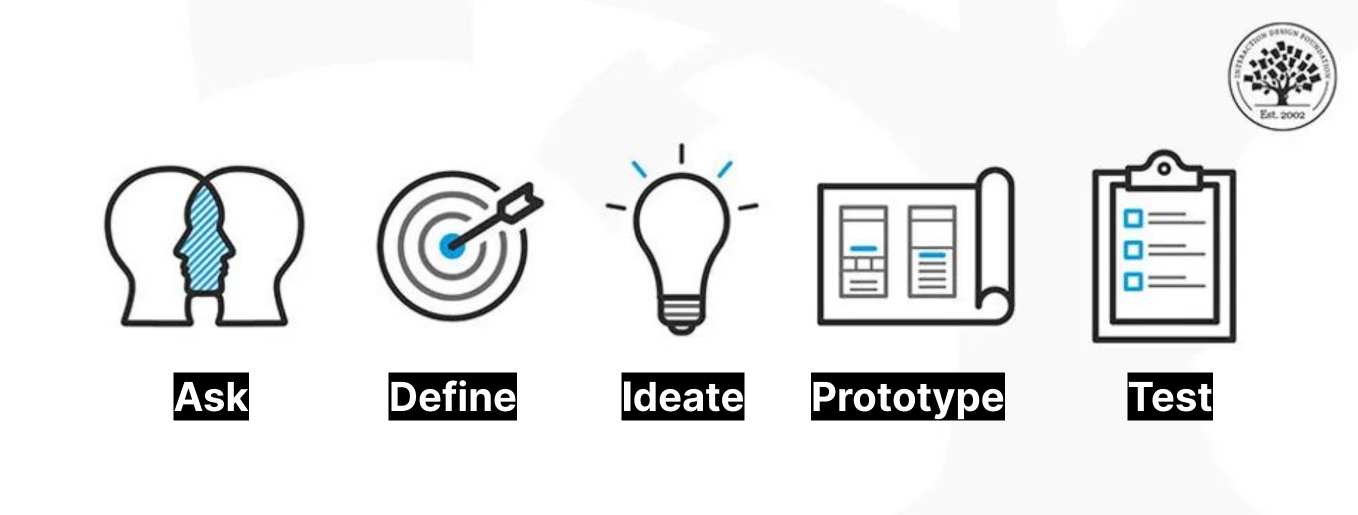
1. Amazon's Alexa is Listening: Why You Should Be Concerned About Your Privacy
2. Amazon Alexa: Useful Tool or Big Brother in Disguise?
3. Is Amazon's Alexa Eavesdropping on Your Private Conversations?

Abstract:

Smart home devices have found their way through people's homes as well as hearts. One such smart device is Amazon Alexa. Amazon Alexa is a voice-controlled application that is rapidly gaining popularity. Alexa was primarily used for checking weather forecasts, playing music, and controlling other devices. This paper tries to explore the extent to which people are aware of the privacy policies pertaining to the Amazon Alexa devices. We have evaluated behavioral change towards their interactions with the device post being aware of the adverse implications. Resulting knowledge will give researchers new avenues of research and interaction designers new insights into improving their systems.

Link to full arXiv article.

So there’s a hypothesis!



Source: Interaction Design Foundation
(https://www.interaction-design.org/literature/article/5-stages-in-the-design-thinking-process)

Worth asking over the rest of the process

Conflict in **responsibilities** to audiences given the ethical positions and issues the profession navigates vs. GenAI **biases**/**inaccuracies**?

Is this actually going to **expand** capacity, or creativity, or speed? **Should** it?

□□□?

How do you decide what projects to pursue?

↖

Capability-out?

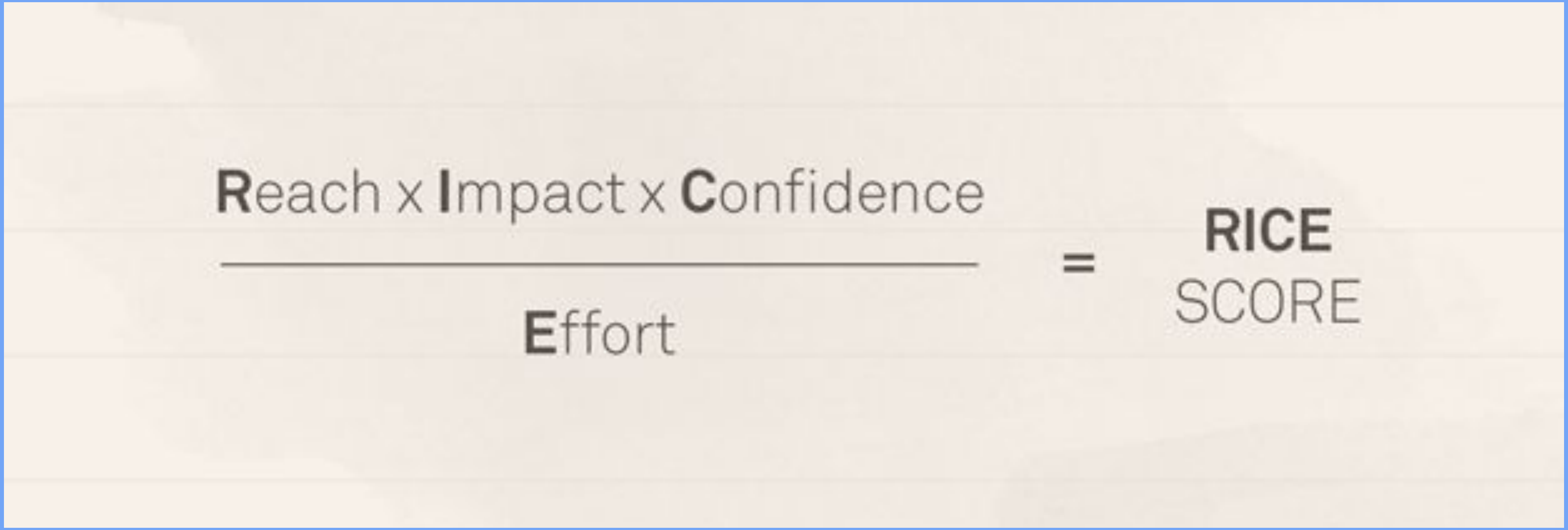
We can do a thing.
What uses are
there for it?

↙ **Use-case-in?**

We have a need.
How can we
address it?

- Discuss and share

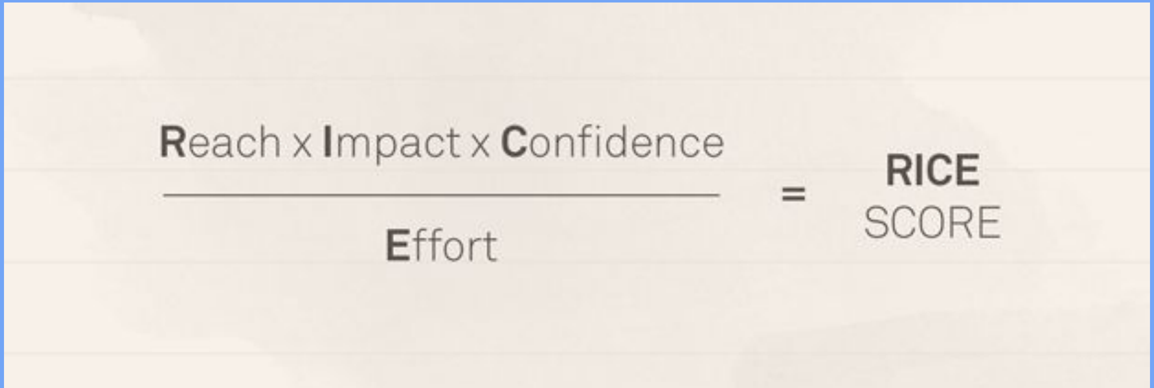
What’s the value?



$$\frac{\text{Reach} \times \text{Impact} \times \text{Confidence}}{\text{Effort}} = \text{RICE SCORE}$$

Source: RICE framework from Intercom

What’s the value?



$$\frac{\text{Reach x Impact x Confidence}}{\text{Effort}} = \text{RICE SCORE}$$

$$\frac{\text{Reach} \times \text{Impact} \times \text{Confidence}}{\text{Effort} \times \boxed{\text{Risk}}}$$

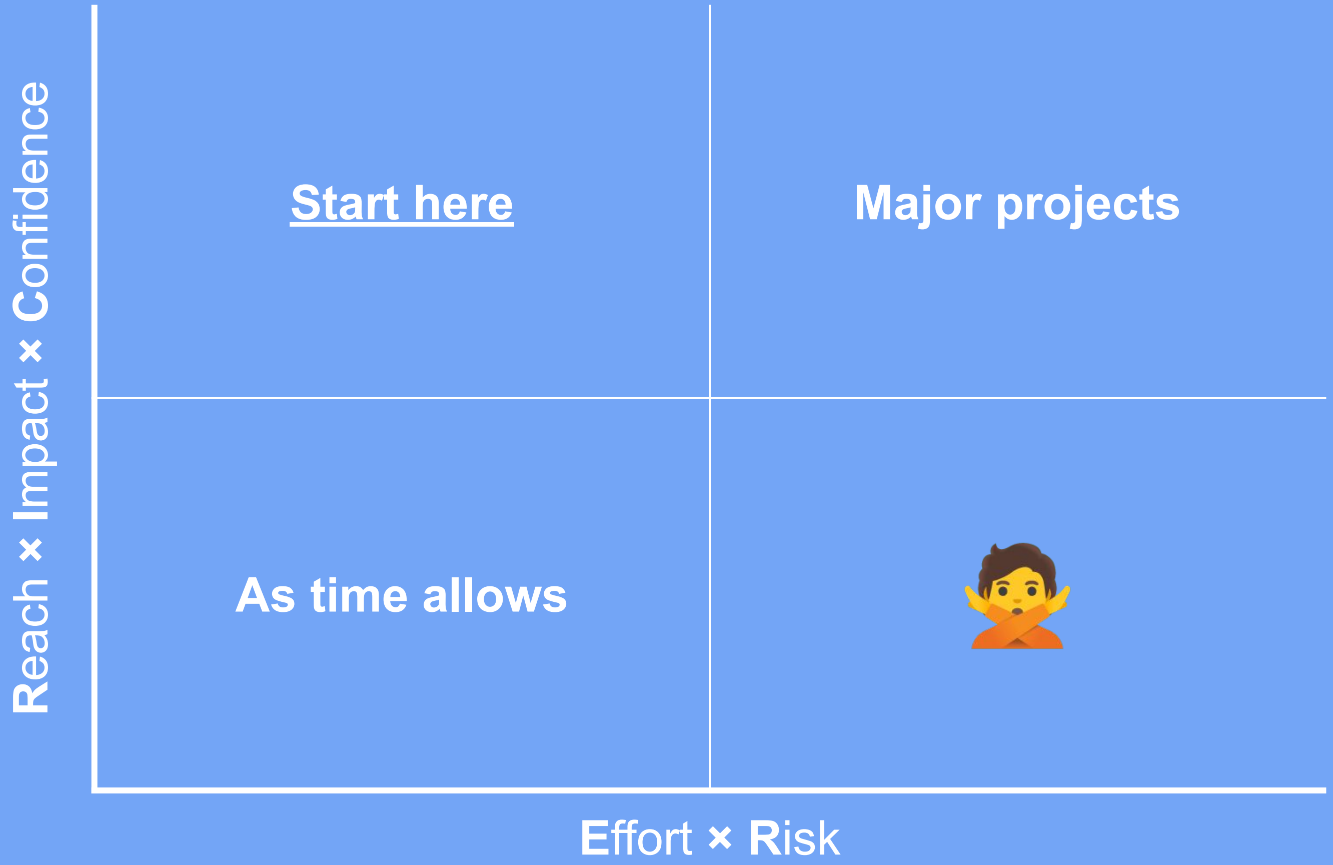
RICER: Proposed framework for evaluating AI efforts

Reach	How many people will this solution help?	Internal users End users
×		
Impact	How often and by how much?	Time saved New functionality enabled
×		
Confidence	How sure are we that…	…we can build it? …the output will be of sufficient quality? …people will use it? …it will deliver the expected benefit?
Effort	How much time/expense will it take to…	…build it? …maintain it? …train the model and oversee the output?
×		
Risk	If something goes wrong, what is the potential harm to…	…our reputation/brand? (internal + external) …our security?

Let’s try it out: Take your use case through these

Reach	How many people will this solution help?	Internal users End users
×		
Impact	How often and by how much?	Time saved New functionality enabled
×		
Confidence	How sure are we that…	…we can build it? …the output will be of sufficient quality? …people will use it? …it will deliver the expected benefit?

Effort	How much time/expense will it take to…	…build it? …maintain it? …train the model and oversee the output?
×		
Risk	If something goes wrong, what is the potential harm to…	…our reputation/brand? (internal + external) …our security?



Reach × Impact × Confidence

Effort × Risk

Start here

Major projects

As time allows

🙅

Questions?

Thanks!