

Data collection methodologies, processes, and ethical considerations within current digital environments

RISK
Summer
School
2025

Dr Nicholas Micallef, Swansea University

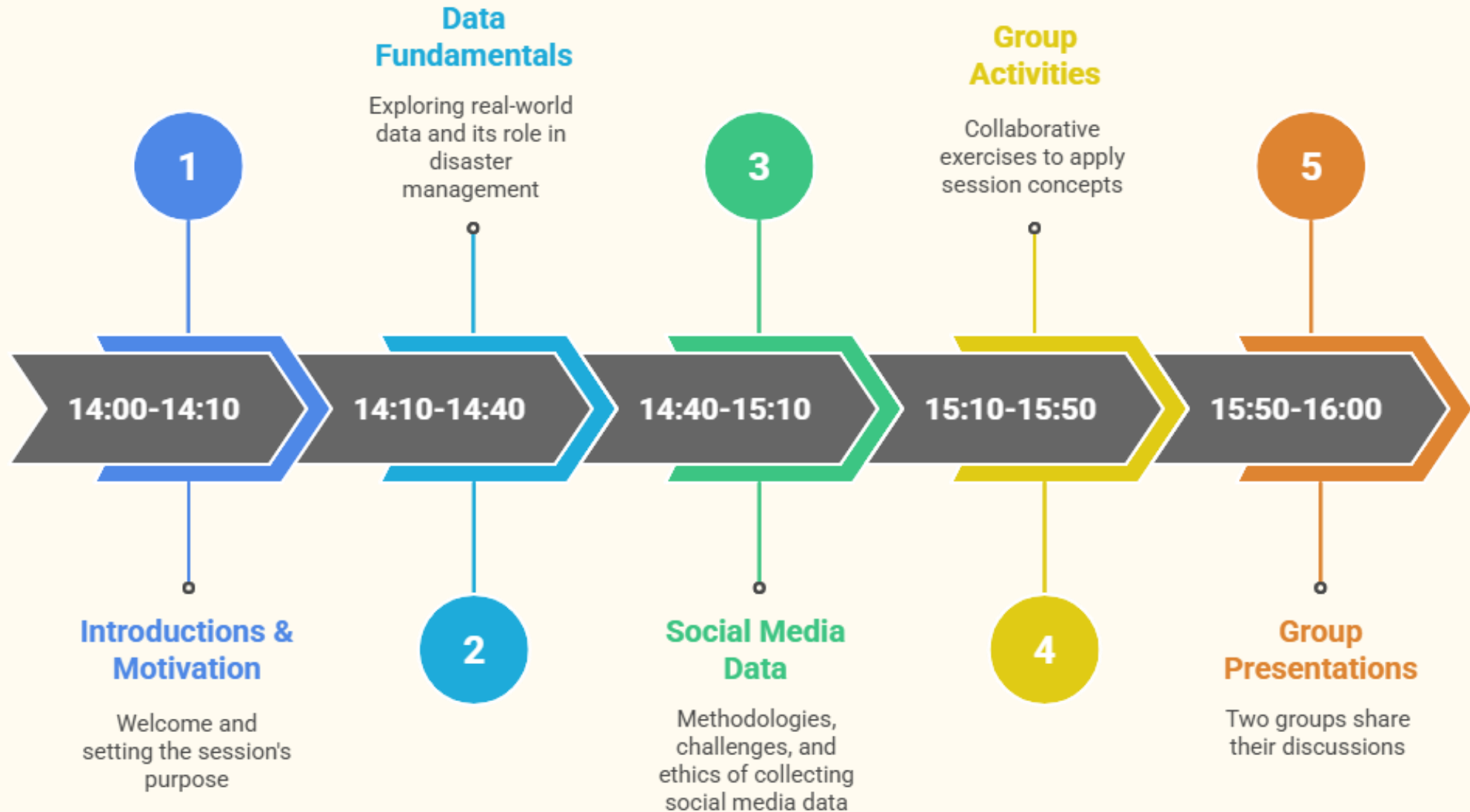
Dr Rouba Iskandar, LIG



Risk Institute
Université Grenoble Alpes



Session Plan



Introductions and Motivation

Dr Nicholas Micallef

Senior Lecturer in Computer Science

Research areas:

- Human-Computer Interaction (HCI)
- Usable Security
- Social Media Analysis for Misinformation/Disinformation
- Human-Centered AI



Dr Rouba Iskandar

Postdoctoral researcher in Computer Science

Research areas:

- Natural hazards
- Seismic risk modeling
- Human behavior in crisis
- Agent-based modeling and simulation





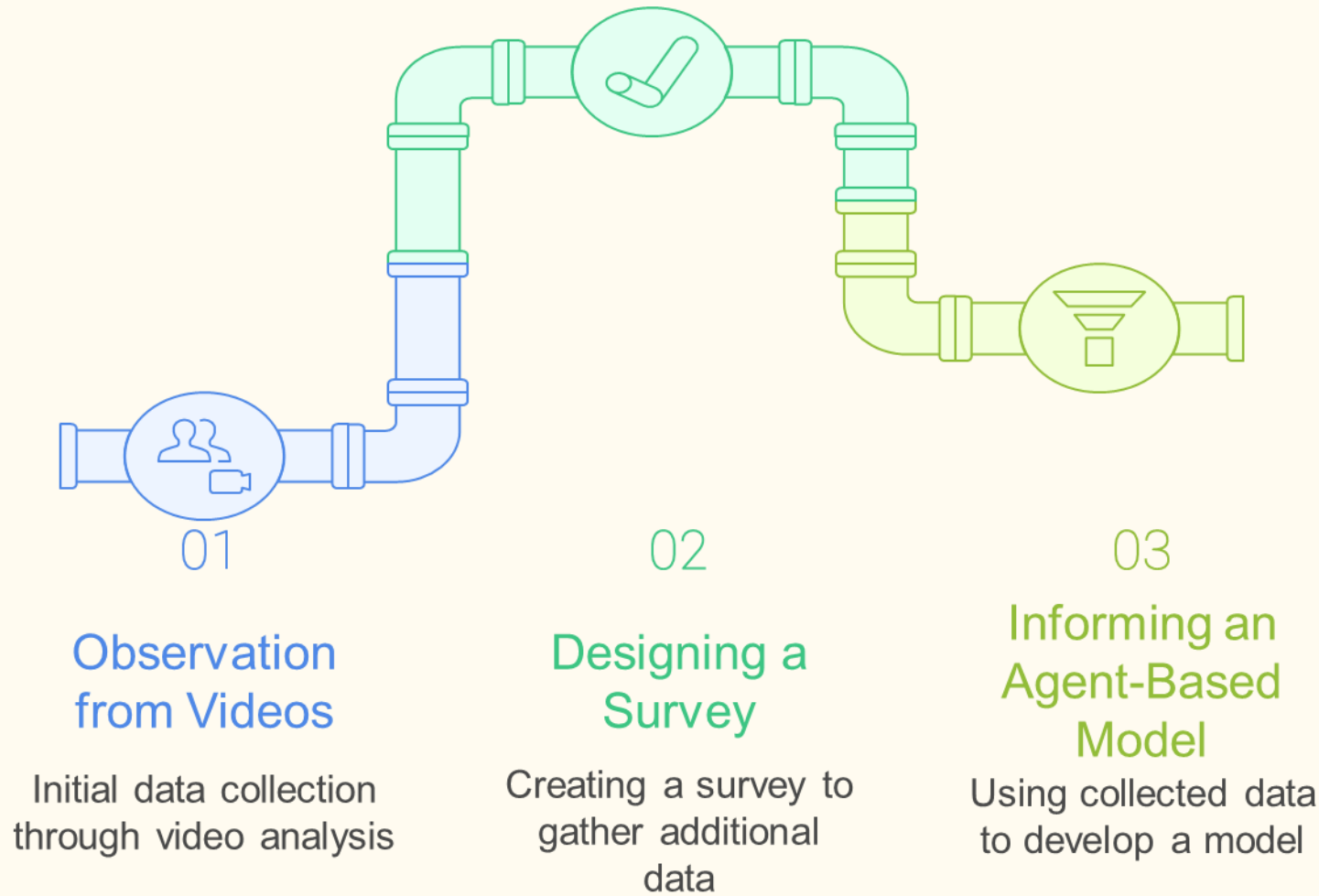
Data- From real-world observations to disaster management

What did we learn from the blast?

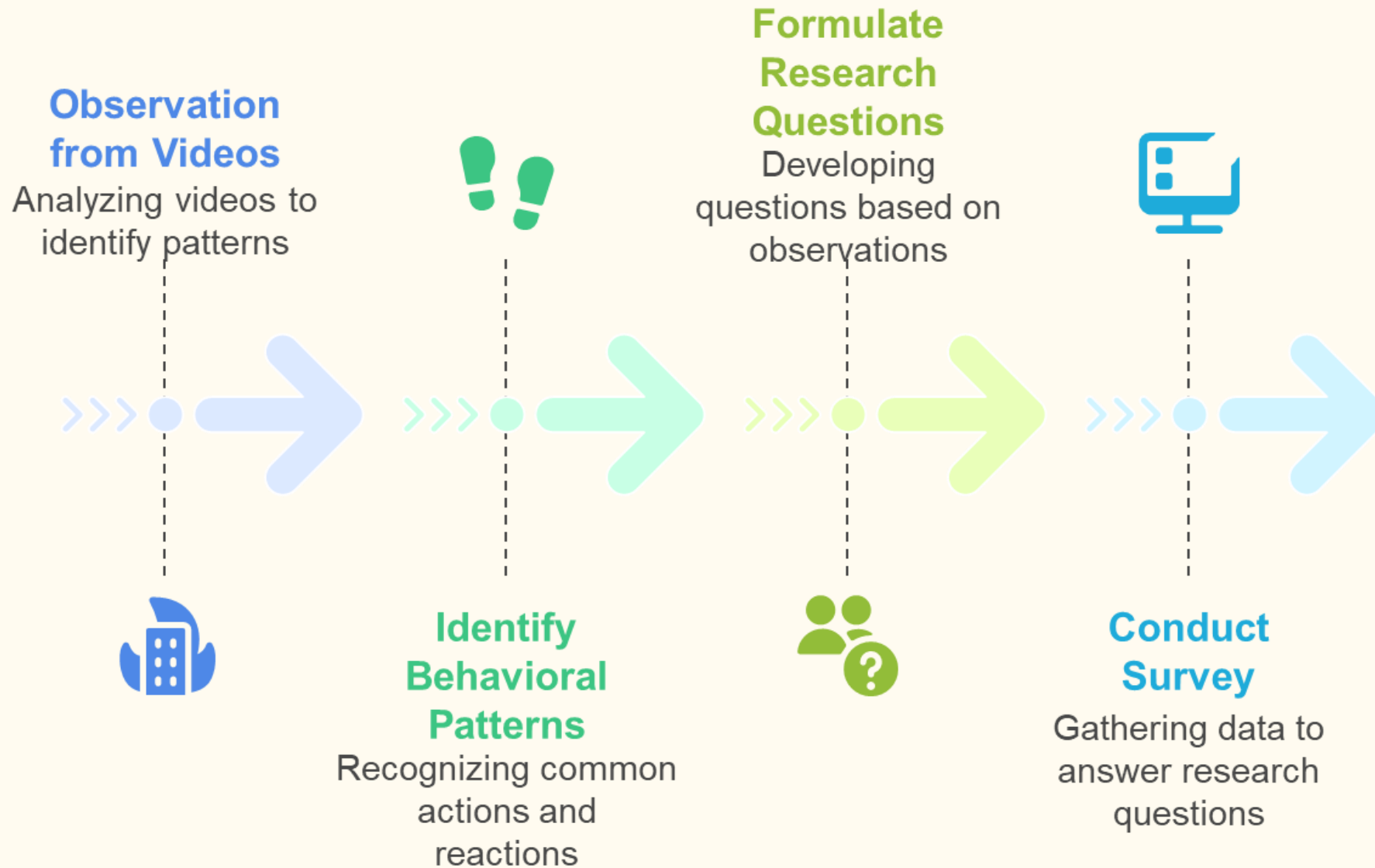


- Disasters generate real-time, human centered data
- Videos capture movement, reactions, before official reports

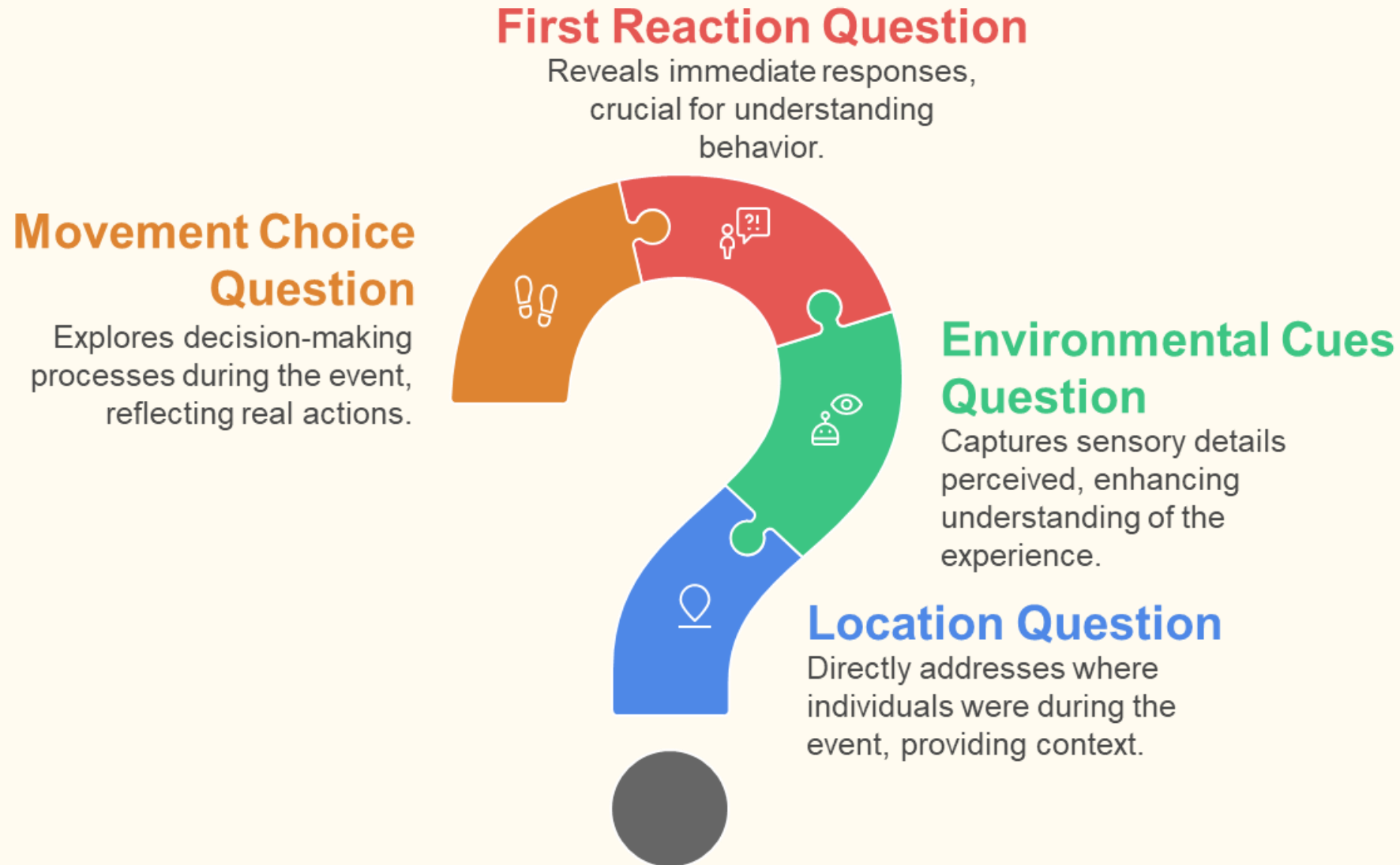
Turning Observation into Research: A Data Pipeline



Step 1: Observation from Videos



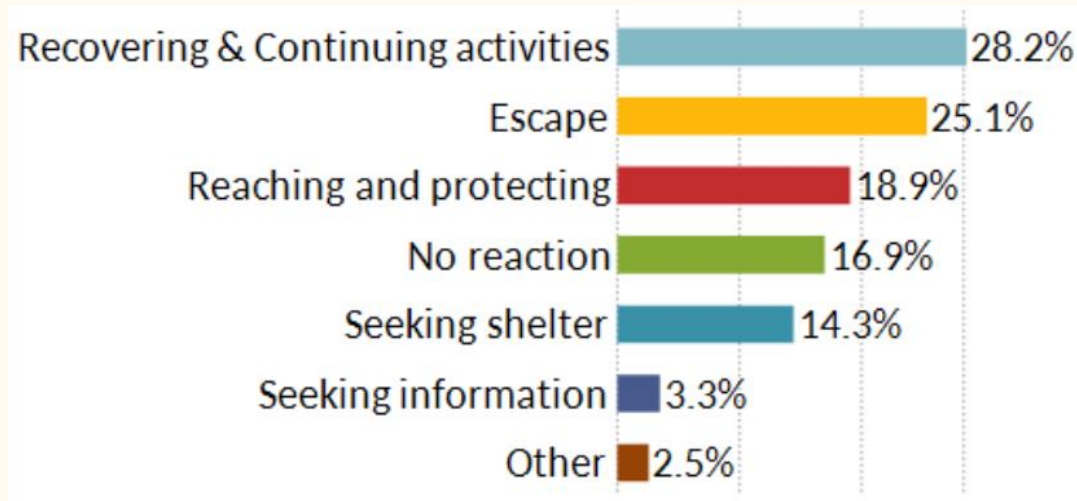
Step 2: Designing a Survey



Survey responses example

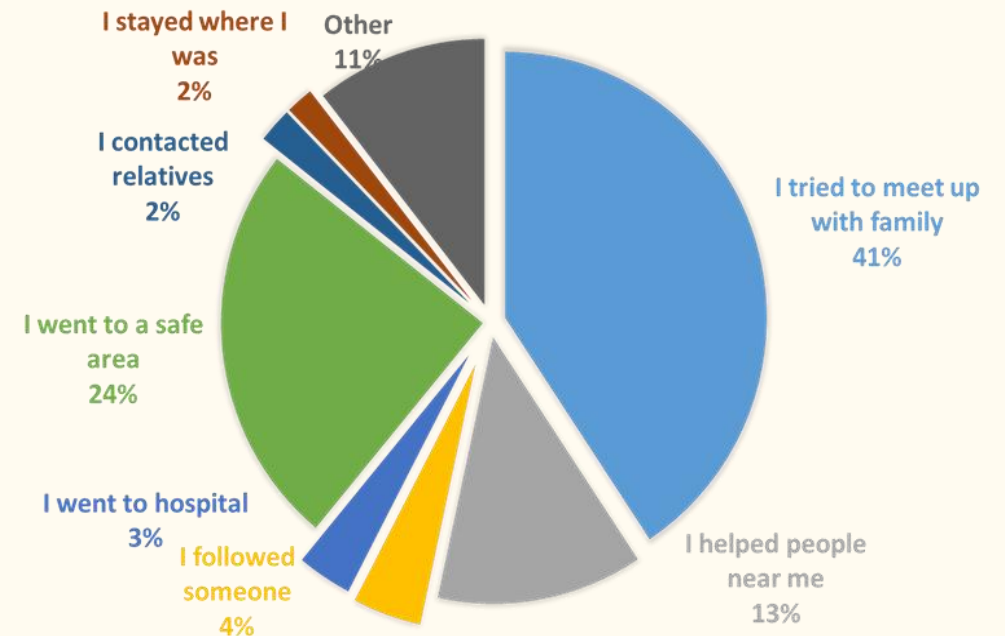
Reactions to the explosion – Indoor

N= 482

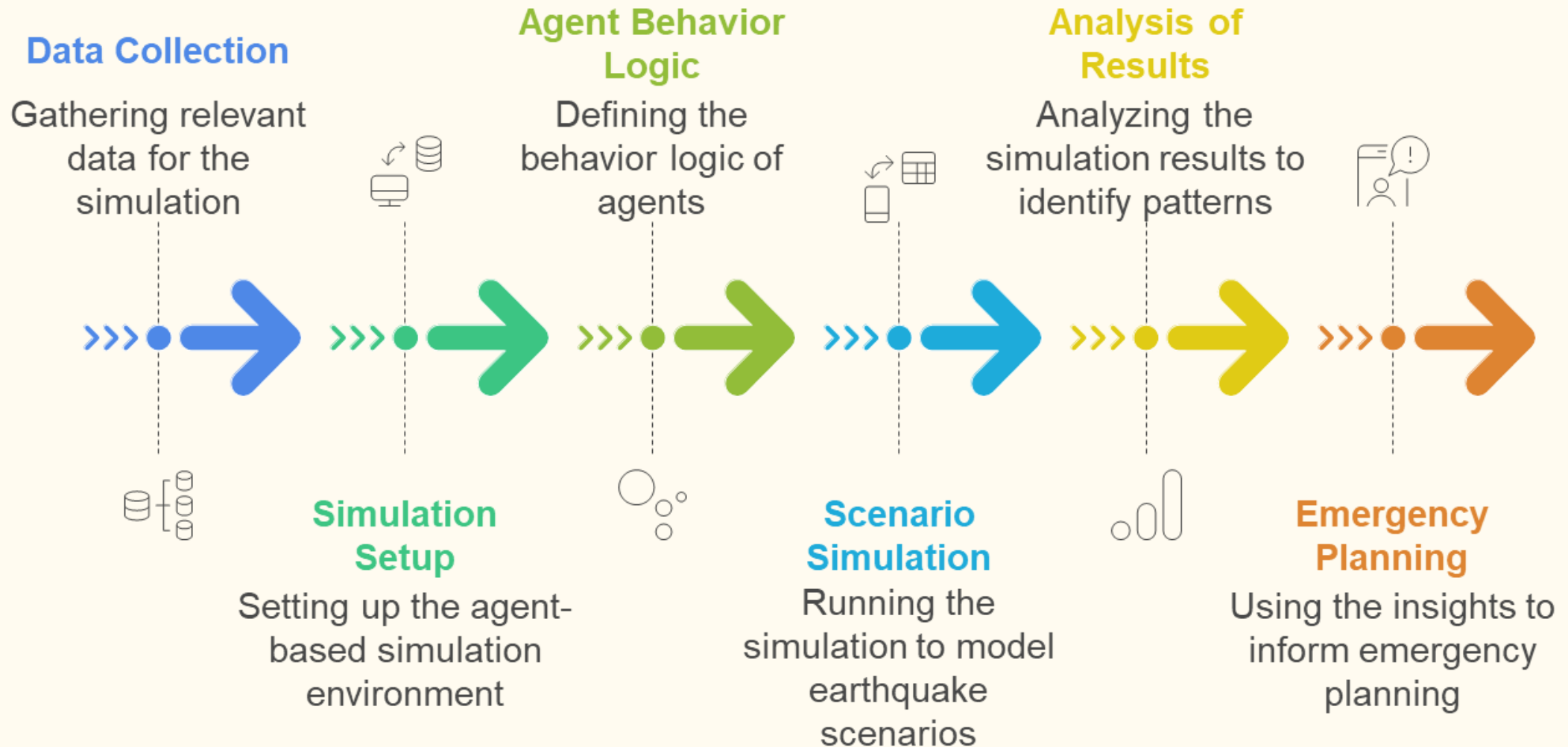


Reactions to the explosion – Outdoor

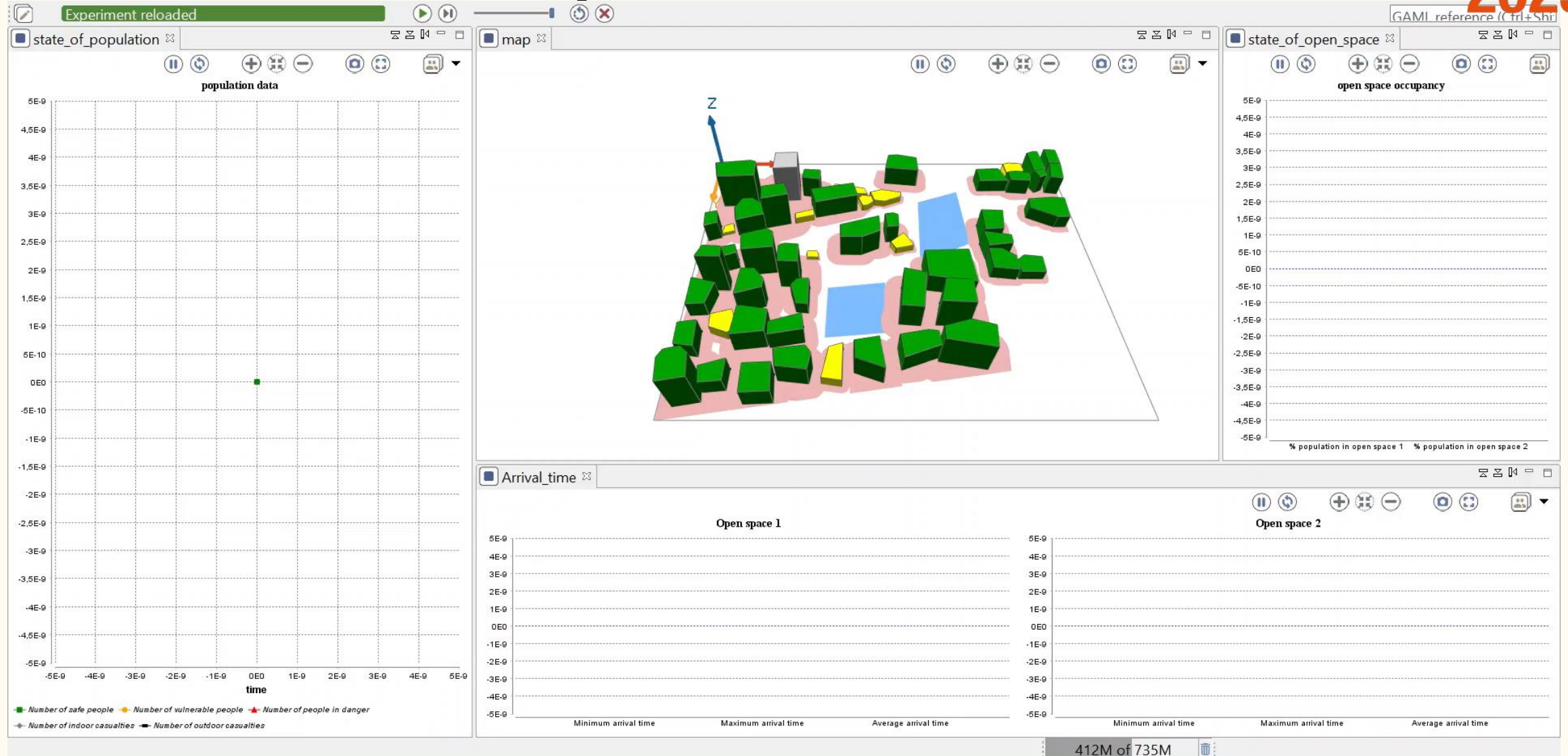
N= 192



Step 3: Informing an Agent-Based Model



Simulation example



What-Ifs : Using Simulations to Shape Policy



Early Flood Warning

Test different timings to optimize safety outcomes.

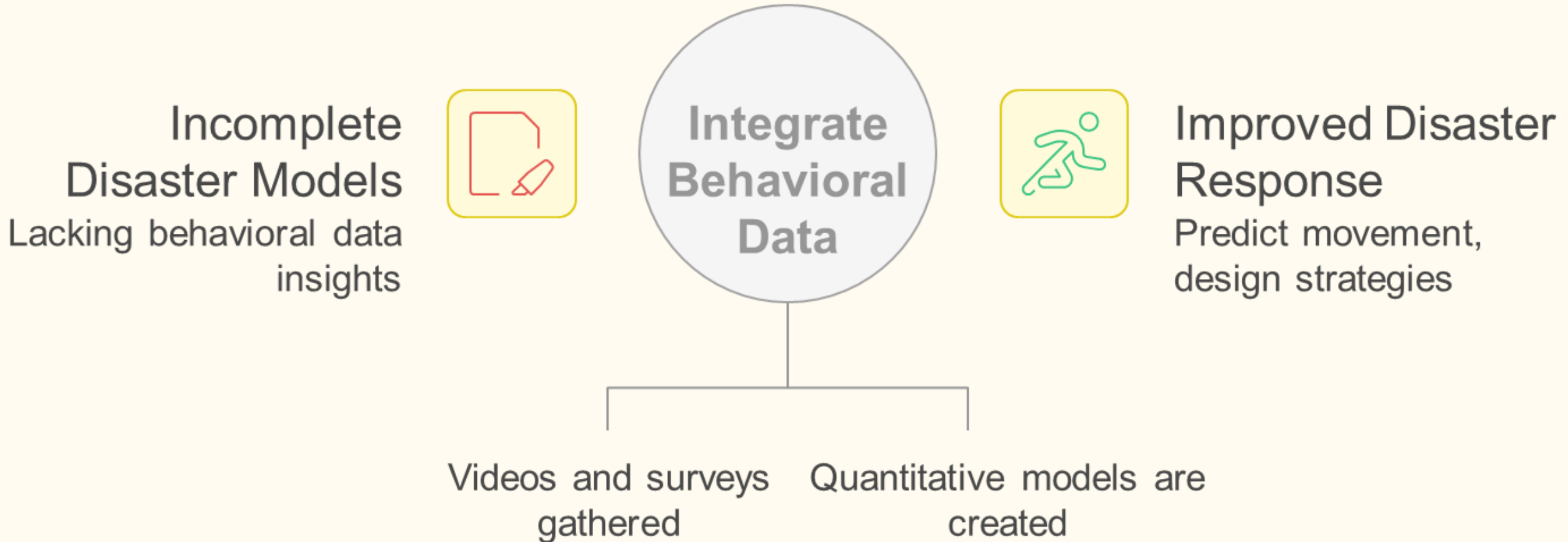
Urban Planning for Earthquakes

Evaluate the impact of replacing unsafe buildings with safe spaces.

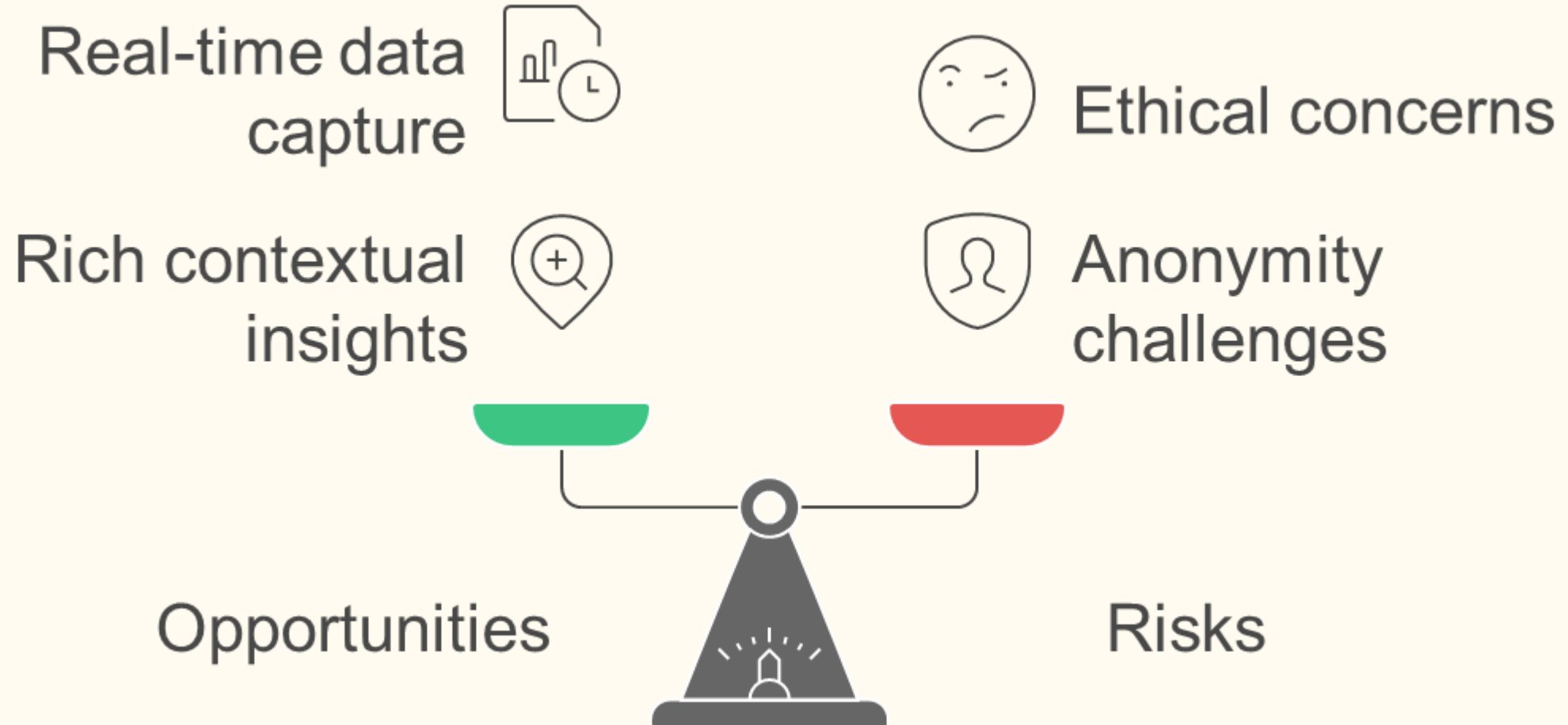
Fire Mitigation Strategies

Assess the effectiveness of firebreaks and evacuation routes.

Behavioral Data is Disaster Data



Social Media, Ethics, and Beyond



Social Media Data Collection: Methodologies, Challenges, and Ethics

Social Media: A Rich Data Repository



Methodologies for Data Collection

Social Listening Tools

Commercial or academic platforms for social data analysis.

Web Scraping

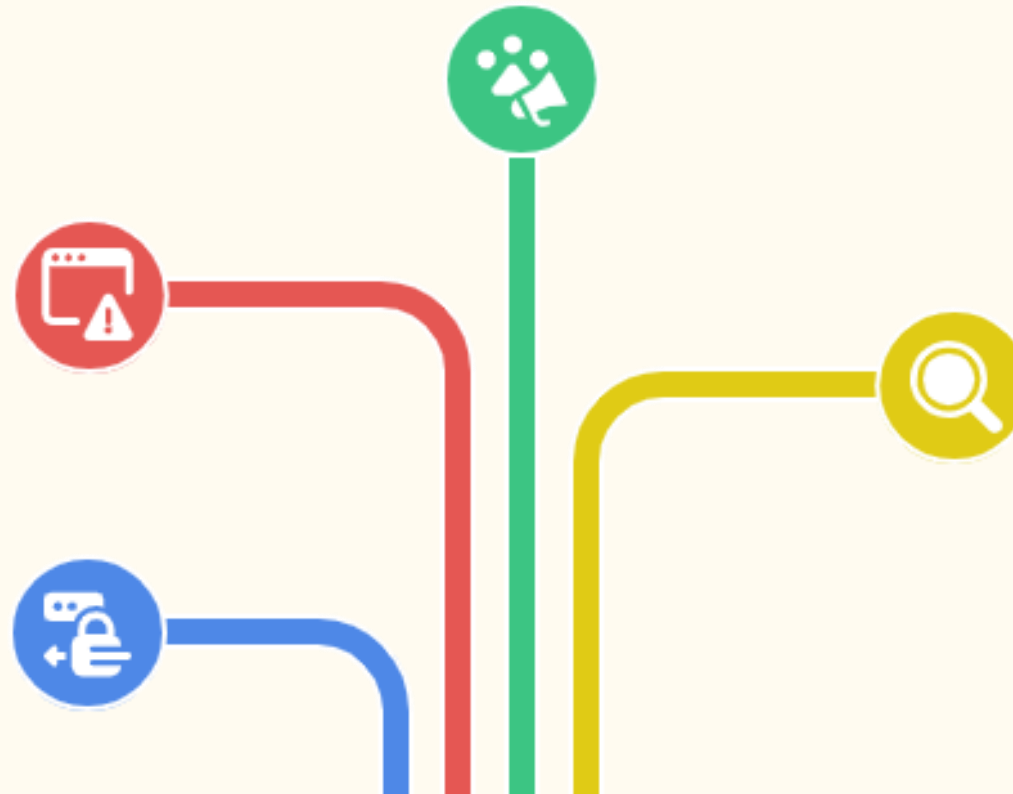
Risky with ethical and legal concerns, suitable for specific needs.

API Access

Structured and legitimate, ideal for platform-approved data collection.

Manual Collection

Best for small-scale, qualitative data gathering.



API Access



What is API

Definition of Application Programming Interface. Controlled gateway to platform data.



Why Use API

Structured and legitimate access. Platform approved data access.



API Examples

Examples include Twitter/X, YouTube, and Facebook APIs.



API Limitations

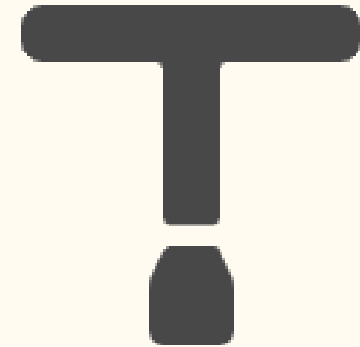
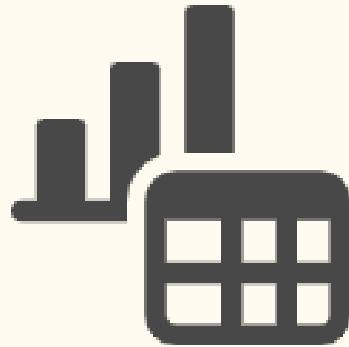
Rate limits, data sampling, and policy changes.

Web Scraping

Definition: Programmatically extracting data directly from websites.



Social Listening Tools



What

Platforms that aggregate and analyse social media data.

Function

Use API access to provide advanced analytics.

Use Case

Large-scale monitoring, brand management, market research.

Examples

Brandwatch, Meltwater, Sprinklr.

Manual Collection



What

Direct observation
and recording of
posts, comments, or
interactions.

Best For

Small-scale, in-depth
qualitative studies.

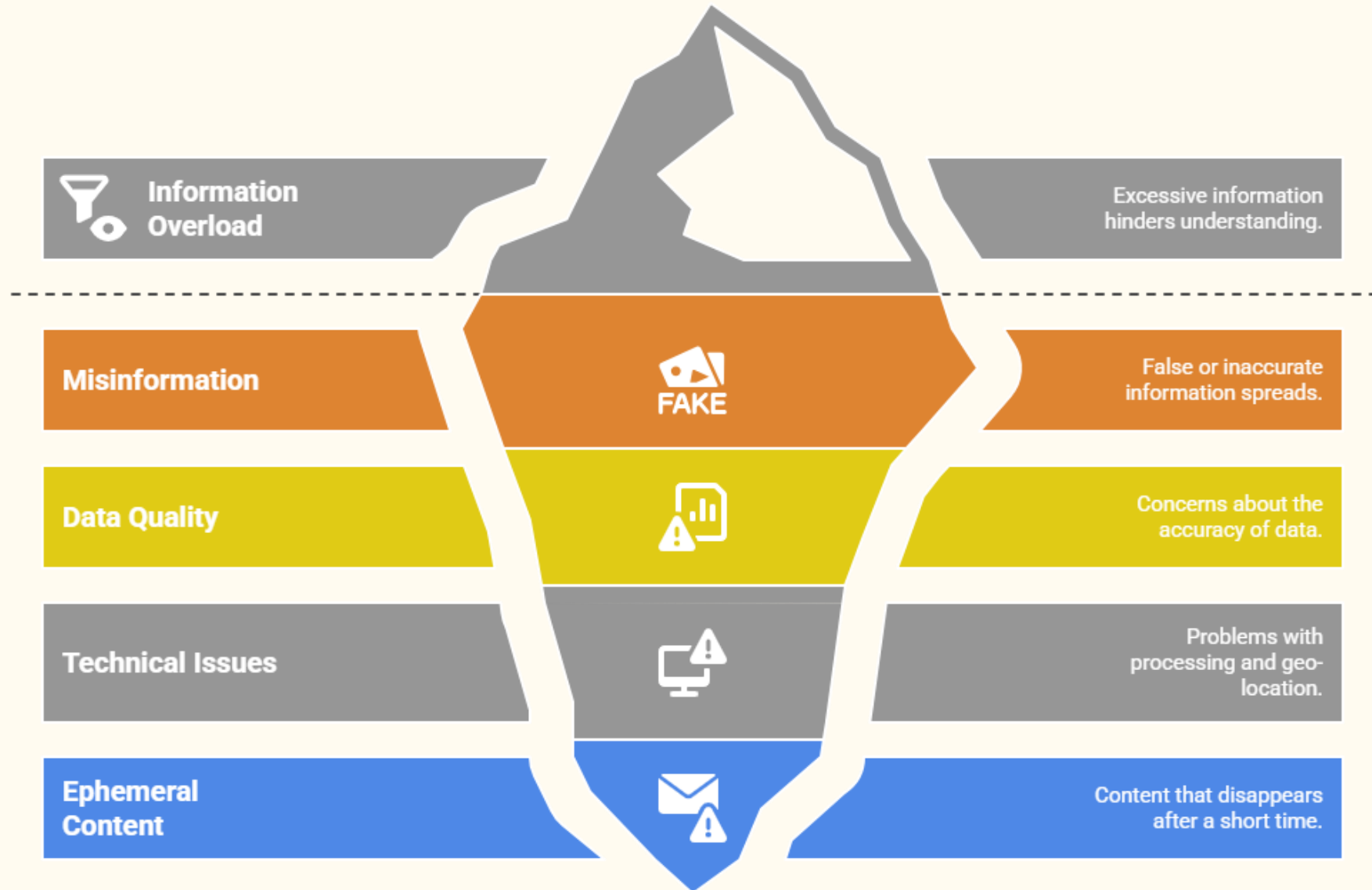
Focus

Understanding
context, nuance, and
specific community
dynamics.

Limitation

Not scalable for "Big
Data" analysis.

Challenges & Barriers: General



Challenges & Barriers: Platform-Specific



TikTok Challenges

Analyzing videos,
bias in algorithms,
extracting data.



Twitter Challenges

API limitations, bot
presence, verifying
information.



WhatsApp Challenges

Encryption methods,
user privacy, user
consent.

X/Twitter (Example)

SPRINGER NATURE Link Log in

[Find a journal](#) [Publish with us](#) [Track your research](#) [Search](#) [Cart](#)

[Home](#) > [Social Networks Analysis and Mining](#) > [Conference paper](#)

Dissecting the Advocacy Discourse Behind the #StopAsianHate Movement on X/Twitter

Conference paper | First Online: 24 January 2025
pp 211–229 | [Cite this conference paper](#)

✓ Access provided by Swansea University / Prifysgol Abertawe

[Download book PDF](#) [Download book EPUB](#)

The book cover for 'Social Networks Analysis and Mining' (ASONAM 2024) is shown. It features a red and white design with the title and Springer logo.

[Yuze Sha](#), [Nicholas Micallef](#) & [Yan Wu](#)

Part of the book series: [Lecture Notes in Computer Science](#) ((LNCS, volume 15213))

Included in the following conference series:
[International Conference on Advances in Social Networks Analysis and Mining](#)

Sections

Figures

References

[Abstract](#)
[Keywords](#)
[Introduction](#)
[Related Work](#)
[Methodology](#)

WhatsApp (Example)



|16

Automating Data Collection from Public
WhatsApp Groups

Challenges and Solutions

NICHOLAS MICALLEF, MUSTAQUE AHAMAD,
NASIR MEMON, AND SAMEER PATIL

The WhatsApp messaging service is one of the most popular mediums for broadening the reach of information dissemination (Srivastava and Singh 2021). Unlike other messaging platforms, the collection of real-world messaging data from WhatsApp is challenging and complicated because of end-to-end encryption, closed source code, and lack of publicly accessible application programming interfaces (APIs). Researchers have used two main ways to circumvent these issues and collect information propagated via WhatsApp: (1) setting up a dedicated WhatsApp number to which people can forward information and (2) joining public WhatsApp groups connected to information about topics of interest (e.g., health, elections, etc.). The latter of the two approaches has been the most popular technique reported in prior work involving WhatsApp data (Garimella and Tyson 2018; Melo et al. 2019; Reis et al. 2020; Resende et al. 2019b). However, such an approach requires considerable manual labor to curate the data collection (Melo et al. 2019; Reis et al. 2020), limiting the scale of the data collection efforts. In our research, we addressed the challenge of scale by investigating the barriers to automating data collection from public WhatsApp groups.

To achieve our research goals, we began with an exploratory investigation with one mobile device that the lead researcher used to manually discover, join, and observe the activities of several public WhatsApp groups. During this initial exploration, we uncovered various challenges that we classified into five broad categories: group discovery, group

Privacy & Ethical Considerations



Key Takeaways



Group Activities

Group Activities

Activity 1: Disaster
Scenario - Data
Collection Strategy

Activity 2: Ethical
Dilemmas in Social
Media Data

Activity 3: Platform-
Specific Data
Challenges

Activity 4:
Misinformation and
Data Veracity

Activity 5: Designing a
Consent Process for
Citizen Data (Social
Media Focus)

Activity 6: Data for
Decision-Making:
Bridging the Gap

Activity 7: The Future
of Social Media Data
in Disaster Research

Activity 8: Designing a
"Rapid Response"
Survey

Activity 9: "What If?" -
Brainstorming for an
Agent-Based Model

Activity 1: Disaster Scenario - Data Collection Strategy

- **Scenario:** A sudden, major earthquake has just struck a densely populated urban area. Communication lines are partially down, but social media (Twitter/X, TikTok, Facebook) is buzzing with activity.
- **Task:** As a research team, outline a high-level strategy for collecting social media data to understand the immediate impact and needs.
 - What platforms would you prioritize and why?
 - What keywords/hashtags would you track?
 - What types of data would you aim to collect (e.g., text, images, location)?
 - What are the immediate challenges you anticipate in collecting this data in real-time?
- **Discussion Prompt:** How would your strategy change if the disaster was a slow-onset event like a prolonged drought?

Activity 2: Ethical Dilemmas in Social Media Data

- **Scenario:** You've collected a large dataset of public tweets related to a recent flood, including some highly emotional and personal accounts from individuals seeking help or expressing distress. Your research aims to map affected areas and identify unmet needs.
- **Task:** Discuss the ethical considerations involved in using this data.
 - How do you balance the potential for public good (identifying needs) with individual privacy?
 - What steps would you take to anonymize or protect sensitive information?
 - Is it ethical to use data from individuals who might be in a vulnerable state and not fully aware their public posts are being collected for research?
 - What are your responsibilities if you come across direct pleas for help or evidence of harm?
- **Discussion Prompt:** Should researchers be held to a higher ethical standard than news organizations or aid agencies when using public social media data during a crisis?

Activity 3: Platform-Specific Data Challenges

- **Scenario:** Your team wants to understand public sentiment and information flow during a public health crisis using social media. You decide to look at TikTok, Twitter/X, and WhatsApp.
- **Task:** For each platform, identify one unique challenge in data collection and suggest a potential (even if partial) solution or mitigation strategy.
 - **TikTok:** (e.g., analyzing video content, trends)
 - **Twitter/X:** (e.g., API limitations, bot activity, misinformation)
 - **WhatsApp:** (e.g., privacy/encryption, group dynamics, consent)
- **Discussion Prompt:** Which platform do you think is the most challenging for systematic data collection in a crisis, and why?

Activity 4: Misinformation and Data Veracity

- **Scenario:** During a rapidly unfolding event, your social media monitoring identifies several viral posts that appear to be false or misleading (e.g., incorrect evacuation routes, unverified claims of casualties).
- **Task:** Discuss how you, as a data collector/analyst, would handle this misinformation within your dataset.
 - How would you identify and flag potentially false information?
 - What are the risks of including or excluding such data from your analysis?
 - What responsibility, if any, do researchers have in combating misinformation during a crisis?
- **Discussion Prompt:** How can collaboration between researchers, social media platforms, and official agencies help address the spread of misinformation during disasters?

Activity 5: Designing a Consent Process for Citizen Data (Social Media Focus)

- **Scenario:** You are planning a research project that involves collecting social media data from a specific community affected by a long-term environmental issue. You want to ensure ethical data collection and gain informed consent where possible.
- **Task:** Design a simplified "consent process" for social media data collection in this context.
 - What information would you need to provide to potential participants?
 - How would you attempt to obtain consent (e.g., direct messaging, community meetings, platform-specific features)?
 - What are the limitations of obtaining consent for publicly available social media data?
- **Discussion Prompt:** When is it *not* necessary to obtain individual consent for social media data, and what are the ethical justifications for that?

Activity 6: Data for Decision-Making: Bridging the Gap

- **Scenario:** You've successfully collected and analysed social media data related to a recent localized flood, identifying areas with significant damage and immediate needs for shelter and medical aid.
- **Task:** How would you present this data to decision-makers (e.g., local government, NGOs) to ensure it is actionable and effectively informs their response?
 - What format would be most effective (e.g., maps, dashboards, brief reports)?
 - What key insights would you highlight?
 - What caveats or limitations of the social media data would you need to communicate?
- **Discussion Prompt:** What are the biggest barriers to translating social media data insights into effective real-world decisions during a crisis?

Activity 7: The Future of Social Media Data in Disaster Research

- **Scenario:** Imagine it's 2035. New social media platforms have emerged, and AI capabilities for data analysis are far more advanced.
- **Task:** Brainstorm how social media data collection and its use in disaster management might evolve.
 - What new opportunities might arise (e.g., predictive analytics, automated needs assessment)?
 - What new ethical or technical challenges might emerge?
 - How might the role of citizen data collectors change?
- **Discussion Prompt:** What single technological advancement related to social media data would have the most significant positive impact on disaster response and decision-making?

Activity 8: Designing a "Rapid Response" Survey

- **Scenario:** A major flash flood has just occurred in a city centre. Your team has seen dozens of TikTok and Twitter videos showing people trapped in cars, wading through water, and helping each other. Now, you need to design a survey to systematically collect data from affected individuals to understand their decision-making.
- **Task:**
 - Based on the scenario, identify three key behavioural themes you would want to investigate (e.g., "Decision to evacuate," "Choice of shelter," "Pro-social behaviour like helping others").
 - For each theme, write two well-formulated survey questions that could be deployed digitally (e.g., via SMS or a web link) to the affected population. The questions should be clear, concise, and aim to capture specific behaviors or decisions.
 - Example for "Pro-social behaviour":
 - "During the flood, did you actively help someone you didn't know? (Yes/No/I was not in a position to help)."
 - "If you saw someone in need of help, what was the biggest factor that influenced your decision to assist or not assist? (e.g., Personal safety, Lack of skills, Assumed someone else would help, etc.)."
- **Discussion Prompt:** What are the biggest challenges in deploying a survey in the immediate aftermath of a disaster, and how might you overcome them?

Activity 9: "What If?" - Brainstorming for an Agent-Based Model

- **Scenario:** Your team has successfully collected and analysed survey data on how people behaved during a recent earthquake. You found that 70% of people hesitated for more than 30 seconds before moving, and 40% tried to use the main stairwell instead of emergency exits.

- **Task:**

You are now advising a team building an agent-based model to simulate evacuations.

- Based on the findings above, brainstorm three "what if" scenarios that the model could test to improve safety.
- For each scenario, describe the change you would make in the simulation and what outcome you would measure.
- Example Scenarios:
 - Scenario 1: "What if a new public address system reduces hesitation time from 30 seconds to 5 seconds? How does that affect building evacuation time and congestion at exits?"
 - Scenario 2: "What if the main stairwell is blocked in the simulation? Where do the 40% of people go, and what new bottlenecks are created?"
 - Scenario 3: "What if we add 'helper' agents who guide others to emergency exits? How many 'helper' agents are needed to significantly change the flow of people away from the main stairwell?"
- **Discussion Prompt:** "The data you collect today becomes the simulation that shapes tomorrow's disaster response." What is the biggest ethical responsibility a researcher has when their data is used to make real-world policy decisions?