# Analyst Workflow Dataset

## Introduction

This dataset represents a detailed log of about 8166 actions taken by 25 intelligence analysts with some degree of professional training who used a handful of simple tools to investigate a fictional scenario and uncover a crime. The crime occurred at a fictional company named Roadeez, a small startup company that develops self-driving car software. They have experienced an insider attack, meaning one of its employees collaborated to steal the company's software. Police have seized several computers owned by the company, and the player's job is to investigate the documents on these computers to conclude who committed the attack. Various tools are provided which attempt to mimic real analyst tools, including a database query search, a fictional web search, a way to recover deleted documents from employee computers, and tools to visualize documents on a timeline and on a map. Players use these tools to explore the collection of documents and then assign which documents they think indicate that each employee is guilty or innocent of the crime. Before performing a search or marking a document as important, analysts are prompted to explain their intentions.

## Population

This data was collected in two rounds, first between September and December 2020 and the second later, on a modified scenario, between August and November 2022. Most participants were intelligence analysts working at the Laboratory for Analytic Sciences. Only subjects who completed the exercise are represented, 20 from the first round and 5 from the second. Specific details on each subject's experience is given in `demographics.csv` (discussed below). Data collection was organized by Jascha Swisher and Christine Brugh of North Carolina State University and the Laboratory for Analytic Sciences.

## Scenarios

In the first version of the game, launched in 2020, the software has fallen into the hands of Roadeez's largest domestic competitor, CARGO. Michelle, the owner of Roadeez, is the culprit who committed the insider attack. With her company only days from bankruptcy, she accepts an offer from a CARGO executive to steal Roadeez's software in exchange for a job at CARGO. Though she deleted the emails in which she admits to this crime, the emails can be recovered. There are also other clues to be found, such as when Michelle calls in sick but then sends emails from San Francisco (the home of CARGO) when she should be in Boston.

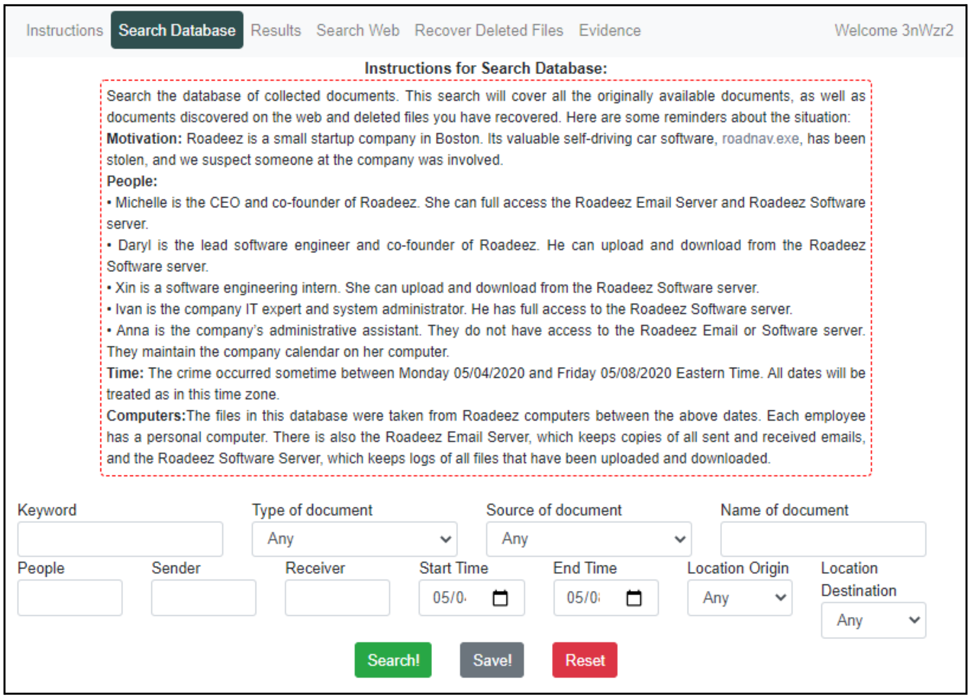
In the second version of the game, launched in 2022, the scenario was modified. The software has fallen into the hands of Karikampani, a foreign competing company. This time it is Ivan, the IT manager for Roadeez, who committed the crime. Like in the first scenario, an agent from Karikampani reaches out, but Ivan never admits to the crime in writing. The most important clue is that he downloaded the stolen software when he has no particular reason to do so. He deletes the network log showing this download, but the player can recover it.

## Instrument

The exercise was developed as a serious game by four researchers at the University of Kentucky during the 2020 calendar year and later expanded in 2022 as part of the Laboratory for Analytic Sciences project *Explainable Interventions for Analyst Workflow*:

- Brent Harrison Ph.D., professor of Computer Science and director of the CORGI Lab
- Anton Vinogradov, research assistant in the CORGI Lab
- Chengxi Li, research assistant in the CORGI Lab
- Stephen G. Ware Ph.D., professor of Computer Science and director of the Narrative Intelligence Lab

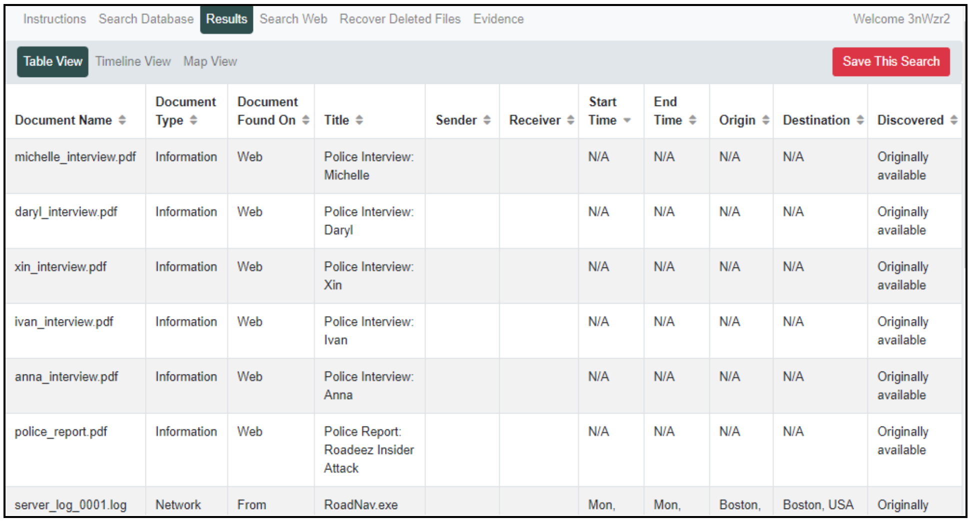
The exercise was deployed via a web browser. Subjects who were invited to participate gave informed consent and then watched an 11 minute video explaining the fictional scenario and how to use the tools provided in the game. Some images of the tools are provided below.



**Instructions for Search Database:**

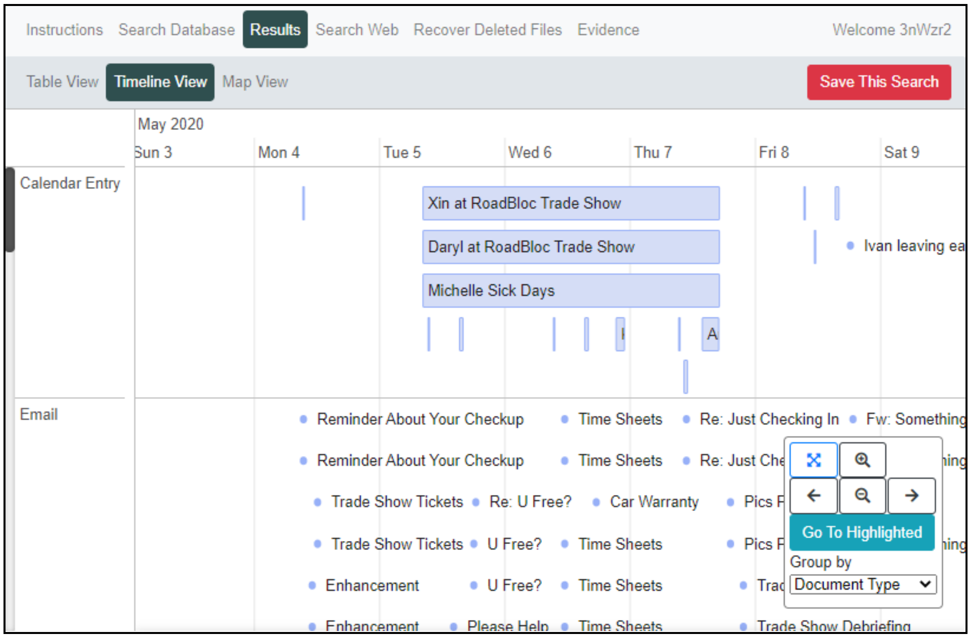
Search the database of collected documents. This search will cover all the originally available documents, as well as documents discovered on the web and deleted files you have recovered. Here are some reminders about the situation:
**Motivation:** Roadeez is a small startup company in Boston. Its valuable self-driving car software, roadnav.exe, has been stolen, and we suspect someone at the company was involved.
**People:**
• Michelle is the CEO and co-founder of Roadeez. She can full access the Roadeez Email Server and Roadeez Software server.
• Daryl is the lead software engineer and co-founder of Roadeez. He can upload and download from the Roadeez Software server.
• Xin is a software engineering intern. She can upload and download from the Roadeez Software server.
• Ivan is the company IT expert and system administrator. He has full access to the Roadeez Software server.
• Anna is the company's administrative assistant. They do not have access to the Roadeez Email or Software server. They maintain the company calendar on her computer.
**Time:** The crime occurred sometime between Monday 05/04/2020 and Friday 05/08/2020 Eastern Time. All dates will be treated as in this time zone.
**Computers:**The files in this database were taken from Roadeez computers between the above dates. Each employee has a personal computer. There is also the Roadeez Email Server, which keeps copies of all sent and received emails, and the Roadeez Software Server, which keeps logs of all files that have been uploaded and downloaded.

The above image shows the database query search tool. It allows players to find a subset of documents based on certain criteria, including keywords, document type, the people mentioned, etc. Search queries can be saved. Saved search can be easily repeated later or can be used as evidence that employees are guilty or innocent.

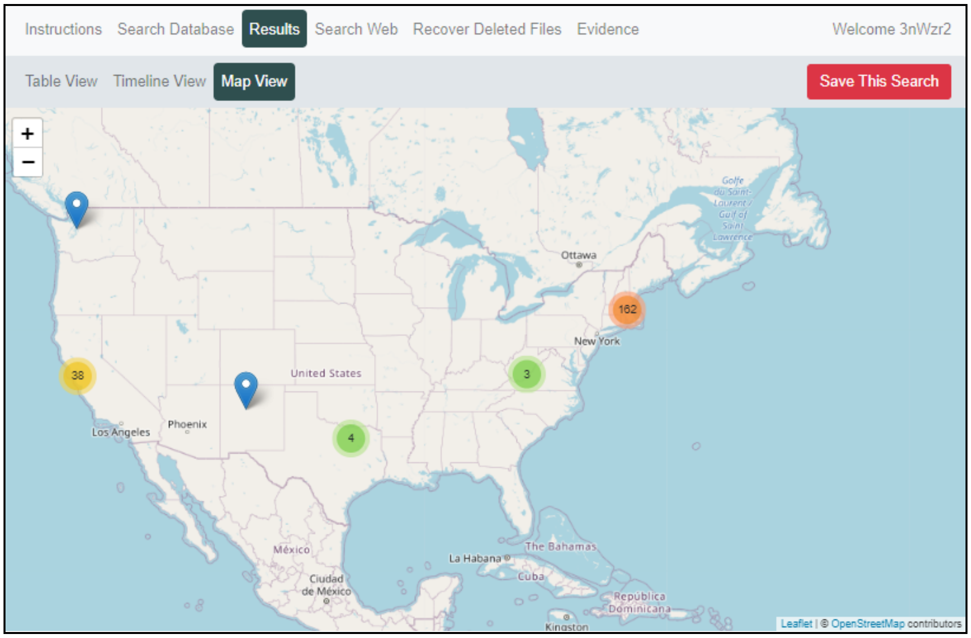


| Document Name | Document Type | Document Found On | Title | Sender | Receiver | Start Time | End Time | Origin | Destination | Discovered |
|---|---|---|---|---|---|---|---|---|---|---|
| michelle_interview.pdf | Information | Web | Police Interview: Michelle | | | N/A | N/A | N/A | N/A | Originally available |
| daryl_interview.pdf | Information | Web | Police Interview: Daryl | | | N/A | N/A | N/A | N/A | Originally available |
| xin_interview.pdf | Information | Web | Police Interview: Xin | | | N/A | N/A | N/A | N/A | Originally available |
| ivan_interview.pdf | Information | Web | Police Interview: Ivan | | | N/A | N/A | N/A | N/A | Originally available |
| anna_interview.pdf | Information | Web | Police Interview: Anna | | | N/A | N/A | N/A | N/A | Originally available |
| police_report.pdf | Information | Web | Police Report: Roadeez Insider Attack | | | N/A | N/A | N/A | N/A | Originally available |
| server_log_0001.log | Network | From | RoadNav.exe | | | Mon, | Mon, | Boston, | Boston, USA | Originally |

The above image shows the results of a search, shown as a table of documents.

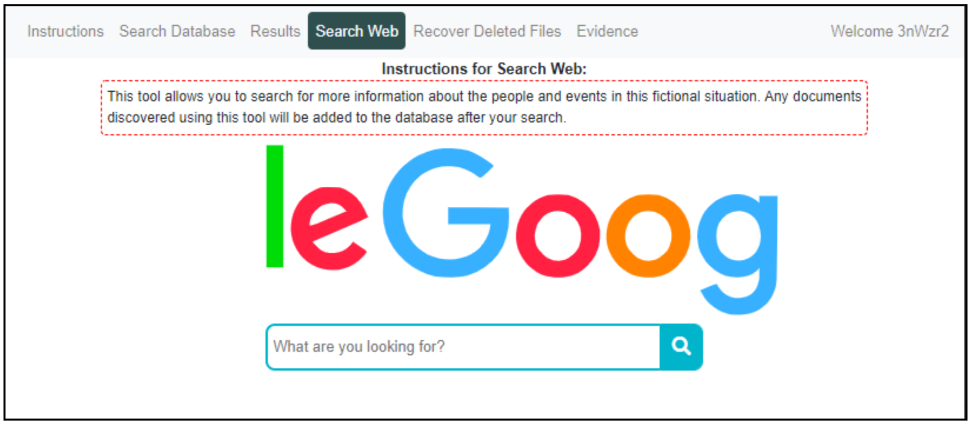


The above image shows the results of a search visualized as a timeline. Blue bars represent documents with duration (e.g. calendar events). Dots represent documents without a meaningful duration.



The above image shows the results of the search visualized as a geographic map.

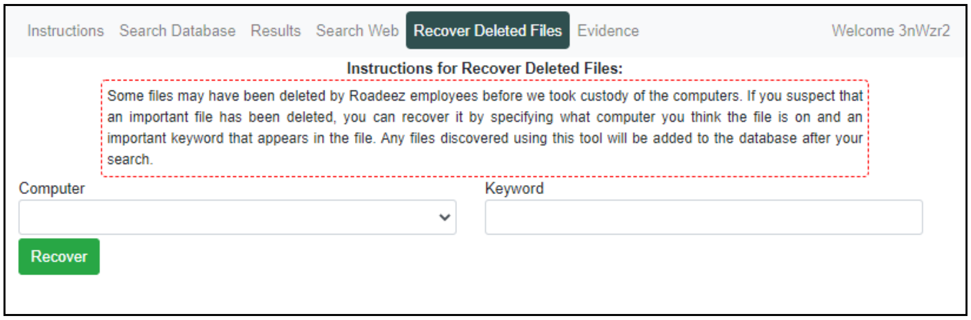
Two tools allow players to discover new documents that are not originally available at the start of the game.



**Instructions for Search Web:**

This tool allows you to search for more information about the people and events in this fictional situation. Any documents discovered using this tool will be added to the database after your search.

The above image shows the fictional web search engine that can be used to find new information documents about specific keywords. Plays can use this to learn more about the fictional people, companies, and events mentioned in the originally available documents.



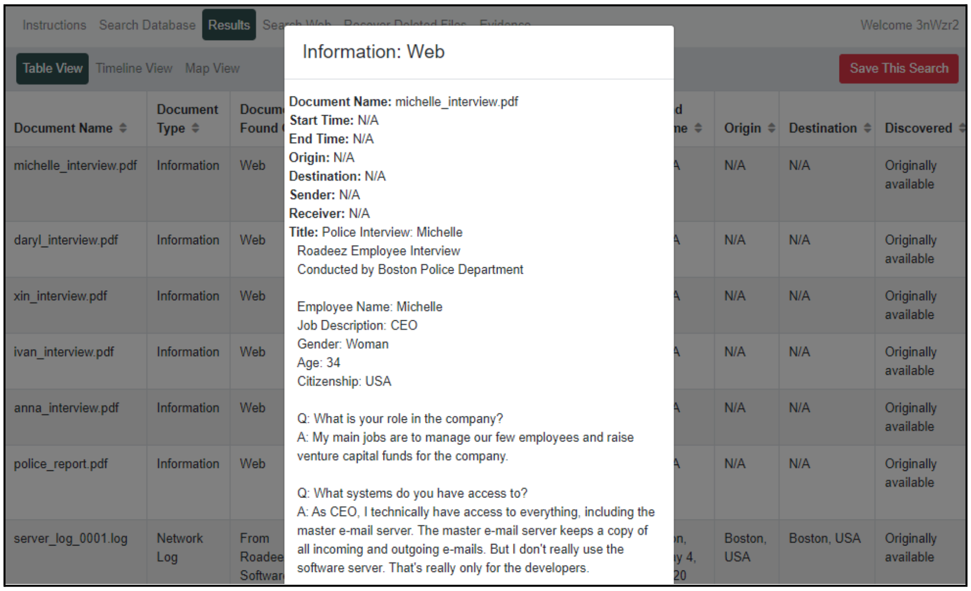
**Instructions for Recover Deleted Files:**

Some files may have been deleted by Roadeez employees before we took custody of the computers. If you suspect that an important file has been deleted, you can recover it by specifying what computer you think the file is on and an important keyword that appears in the file. Any files discovered using this tool will be added to the database after your search.

Computer

Keyword

The above image shows the tool for recovering deleted files. The user must specify which computer to recover the file from and one or more keywords contained in the file to be recovered.



| Document Name | Document Type | Docum Found | d ne | Origin | Destination | Discovered |
|---|---|---|---|---|---|---|
| michelle_interview.pdf | Information | Web | A | N/A | N/A | Originally available |
| daryl_interview.pdf | Information | Web | A | N/A | N/A | Originally available |
| xin_interview.pdf | Information | Web | A | N/A | N/A | Originally available |
| ivan_interview.pdf | Information | Web | A | N/A | N/A | Originally available |
| anna_interview.pdf | Information | Web | A | N/A | N/A | Originally available |
| police_report.pdf | Information | Web | A | N/A | N/A | Originally available |
| server_log_0001.log | Network Log | From Roadee Softwar | n, y 4, 20 | Boston, USA | Boston, USA | Originally available |

Information: Web

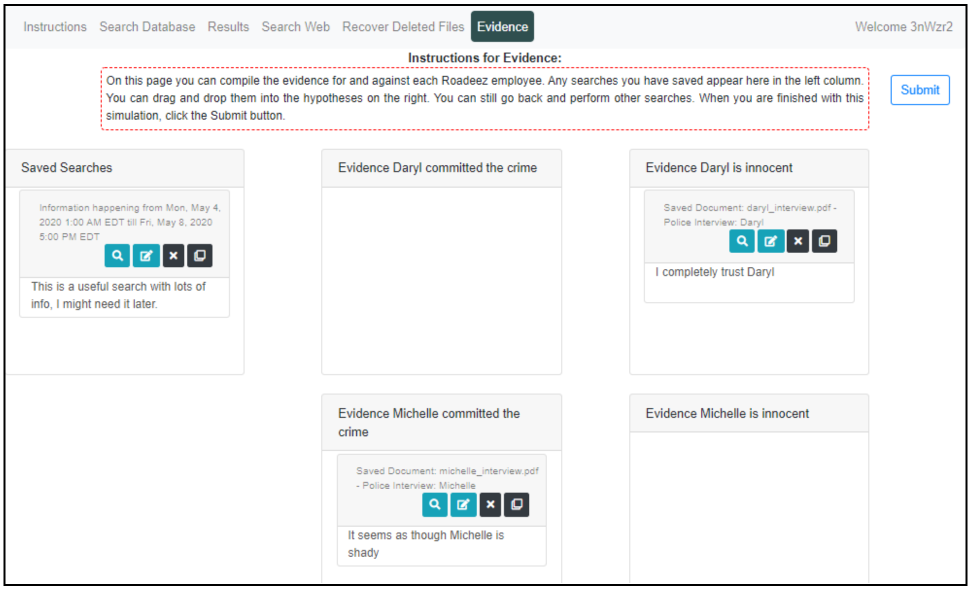
**Document Name:** michelle_interview.pdf
**Start Time:** N/A
**End Time:** N/A
**Origin:** N/A
**Destination:** N/A
**Sender:** N/A
**Receiver:** N/A
**Title:** Police Interview: Michelle
Roadeez Employee Interview
Conducted by Boston Police Department

Employee Name: Michelle
Job Description: CEO
Gender: Woman
Age: 34
Citizenship: USA

Q: What is your role in the company?
A: My main jobs are to manage our few employees and raise venture capital funds for the company.

Q: What systems do you have access to?
A: As CEO, I technically have access to everything, including the master e-mail server. The master e-mail server keeps a copy of all incoming and outgoing e-mails. But I don't really use the software server. That's really only for the developers.

The above image shows the player viewing an individual document. Individual documents can also be saved as important and used as evidence later.



**Instructions for Evidence:**

On this page you can compile the evidence for and against each Roadeez employee. Any searches you have saved appear here in the left column. You can drag and drop them into the hypotheses on the right. You can still go back and perform other searches. When you are finished with this simulation, click the Submit button.

Saved Searches

Information happening from Mon, May 4, 2020 1:00 AM EDT till Fri, May 8, 2020 5:00 PM EDT

This is a useful search with lots of info, I might need it later.

Evidence Daryl committed the crime

Evidence Daryl is innocent

Saved Document: daryl_interview.pdf - Police Interview: Daryl

I completely trust Daryl

Evidence Michelle committed the crime

Saved Document: michelle_interview.pdf - Police Interview: Michelle

It seems as though Michelle is shady

Evidence Michelle is innocent

The above image shows the evidence page. Any searches or individual documents that the player has saved as important are shown in the leftmost column called “Saved Searches.” The two columns on the right contains 10 hypotheses. For each of the 5 employees of the fictional company, there is a hypothesis that the employee is innocent and one that they are guilty. Saved searches and saved documents can be dragged and dropped into these hypotheses to provided support for each one. When the player is finished, the click the “Submit” button at the top right. They are then asked which employee they think committed the insider attack and to report their demographic information (age range, years of experience as an analyst, etc.).

## Documents

There are four kinds of fictional documents used in the game:

- **Email:** emails sent or received by company employees.
- **Calendar Entry:** events from on the company calendar.
- **Network Log:** details about files being uploaded or downloaded to the company server.
- **Information:** miscellaneous documents, such as police reports, web pages, etc.

The fictional documents used in the game are provided in `documents.csv`, which is in comma separated value format. Note that some documents were only available in the first scenario and some were only available in the second scenario. Some documents that are only available in the second scenario are slightly modified versions of documents from the first scenario. For example, `email_031` and `email_045` are identical in every way, except that `email_031` is sent from San Francisco and is only available in the first scenario, whereas `email_045` is sent from Boston and is only available in the second scenario.

The values of each column in `documents.csv` are:

- **Name:** The file name of the document.

- **Type:** `Email`, `Calendar Entry`, `Network Log`, or `Information`.
- **Title:** The title of the document.
- **Body:** The contents of the document, such as the body of an email or the text of a webpage.
- **Source:** Where the document was found. Documents can be found on `Roadeez Email Server`, which keeps a log of all emails sent and received at the company, `Roadeez Software Server`, where the company's software is stored, or any employee's personal computer, for example, `Anna's Computer`. Anna is the company administrative assistant, so all calendar entries have `Anna's Computer` as the source. Information documents and those found via the web search tool have the source `Web`.
- **Keywords:** A list of keywords relevant to the document, separated by spaces.
- **People:** A list of people involved in the document, separated by spaces. All fictional characters in the game are referred to only by their first names.
- **Start Time:** The start time of the document, as a Unix timestamp (number of seconds since January 1, 1970). The earliest possible time for any document is `1588564800`, which represents May 4th, 2020 at 12:00 AM US Eastern Time. All documents assume US Eastern Time. Start times for calendar entries represent when the event began. Start times for emails are the same as end times and represent when the email was sent or received. Start times for network logs are the same as end times and represent when the event was recoded. Start times for information documents are always `1588564800`.
- **End Time:** The end time of the document. Calendar events are the only documents whose end times are different from their start times, and this represents when the event ended.
- **Sender:** The person who sent an email. Blank for non-email documents.
- **Receiver:** The person who received an email. Blank for non-email documents.
- **Origin (Latitude / Longitude / City / Country):** The latitude, longitude, city, and country of the location from which the document was sent or where it began (e.g. where an email was sent from or where a calendar event began). Information documents do not have an origin. When documents (e.g. information) have no meaningful origin Latitude and Longitude are blank and city and country are given as `None`.
- **Destination (Latitude / Longitude / City / Country):** The latitude, longitude, city, and country of the location to which the document was sent or where it ended (e.g. where an email was received or where a calendar event ended). Information documents do not have a destination. When documents (e.g. information) have no meaningful destination Latitude and Longitude are blank and city and country are given as `None`.
- **Status:** `available` if the document was available to the player at the start of the game, `discoverable` if the document was not originally available but could be

discovered via the web search tool, or deleted if the document was not originally available but could be discovered via the recover deleted files tool.

- **Michelle Cargo Scenario**: true if the document is available in the first scenario, otherwise false.
- **Ivan Karikampani Scenario**: true if the document is available in the second scenario, otherwise false.

Note that there are many documents where are nearly duplicates of one another. This happens when the same document is found in multiple places. For example, if Ivan sends and email to Anna, there will be three copies of that document, each with a different Source: one on Ivan's Computer, one on Anna's Computer, and one on the Roadeez Email Server.

## Events

All event data is provided in events.csv, which is in comma separated value format. Most events in the game revolve around searching for, examining, and saving sets of documents. Search queries have many possible parameters corresponding to the various features of documents described above, including the type of document, where it was found, keywords, origin, designation, etc. An individual document can be represented as a query with a specific value in the Document Name field (i.e. a query that returns a single document). Analysts can save queries to use on the game's evidence tab. These saves queries are than used as evidence that each of the company's five employees are either guilty or innocent of the insider attack.

There are six kinds of events:

- **search:** The analyst uses the database query tool to find a set of documents.
- **discover:** The analyst uses the web search tool or the recover deleted files tool to find new documents. After documents are discovered, they are added to the database, and henceforth they can be returned as part of a database search.
- **switch:** The analysts switches between the tools and visualizations available in the game.
- **read:** The analyst reads a specific document.
- **save:** The analyst saves their most recent search query as important. If the analyst is reading a specific document, there is a shortcut to save that document as important by saving a query with that document's name in the Document Name field.
- **revisit**: The analyst uses the evidence tool to revisit a previously saved search or document.
- **assign:** The analyst assigns a saved query as evidence that an employee is guilty or innocent of the crime.

The meanings of the columns are:

- **Scenario:** Indicates which scenario this user belongs to: michelle-cargo for the original scenario and ivan-karikampani for the second scenario.

- **User ID:** A unique index number (starting at 0) for each user.
- **Step:** The step index number (starting at 0) for this event for this specific user. This number is unique and sequential for each user, but not unique in the whole dataset.
- **Timestamp:** The time the event occurred, as a Unix timestamp (number of seconds since January 1, 1970).
- **Event Type:** `search`, `discover`, `revisit`, `switch`, `read`, `save`, or `assign`.
- **Visualization:** The tool the analyst was using during this event. For `switch` events, this is the tool the analyst is switching to. Possible values are the `instructions` tab at the start of the game, the `table` view of documents, the `timeline` view of documents, the `map` view of documents, the `web` search tool, the `recover` deleted files tool, and the `evidence` tab where analysts assign queries as evidence of guilt or innocence. This feature is especially important for `discover` events, because it indicates which tool the analyst is using: `web` for the web search and `recover` for the recover deleted files tools.
- **Document Name:** For `search` events, this is the value the analyst entered into the "Name of document" field, if any. For `read` events, this is the name of the document being read. For `save` and `assign` events, this is either the name of the individual document being saved/assigned or the value of the "Name of document" field of the search query. For `switch` events, this is the value from the most recent `search`, `discover`, `revisit`, or `save` event or blank if no such events have happened yet.
- **Document Type:** Values include `Email`, `Calendar Entry`, `Network Log`, and `Information`. The value `Any` covers all possible document types. For `search`, `save`, and `assign` events, this is the type of document chosen from the drop-down menu on the search query tool, or `Any` by default. For `read` events, this is the type of document being read. For `discover` events, this will always be `Any`. For `switch` events, this is the value from the most recent `search`, `discover`, `revisit`, or `save` event or blank if no such events have happened yet.
- **Document Source:** For `read` events, this is the source of the document being read. For `discover` events, this is the source from which documents are potentially being discovered. When using the web search tool to discover new documents, the source will always be `Web`. When using the recover deleted files tool, the source will be the computer the analyst is searching for deleted files on (e.g. `Michelle's Computer`). For `search`, `revisit`, `save`, and `assign` events, this is the value entered in the "Source of document" field of the query, or `Any` by default. For `switch` events, this is the value from the most recent `search`, `discover`, `revisit`, or `save` event or blank if no such events have happened yet.
- **Keywords:** For `read` events, this is the keywords associated with the document being read. For `discover` events using the web search tool, this is the search query. For `discover` events using the recover deleted files tool, this is the value entered into the

"Keyword" field, which is required to perform a recover deleted files operation. For `search`, `revisit`, `save`, and `assign` events, this is the value entered in the "Keyword" field of the query, if any. For `switch` events, this is the value from the most recent `search`, `discover`, `revisit`, or `save` event or blank if no such events have happened yet.

- **People:** For `read` events, this is the people associated with the document being read. For `search`, `revisit`, `save`, and `assign` events, this is the value entered in the "People" field of the query, if any. For `switch` events, this is the value from the most recent `search`, `discover`, `revisit`, or `save` event or blank if no such events have happened yet. This will always be blank for `discover` events.
- **Sender:** For `read` events, this is the sender of the document being read, if any. For `search`, `revisit`, `save`, and `assign` events, this is the value entered in the "Sender" field of the query, if any. For `switch` events, this is the value from the most recent `search`, `discover`, `revisit`, or `save` event or blank if no such events have happened yet. This will always be blank for `discover` events.
- **Receiver:** For `read` events, this is the receiver of the document being read, if any. For `search`, `revisit`, `save`, and `assign` events, this is the value entered in the "Sender" field of the query, if any. For `switch` events, this is the value from the most recent `search`, `discover`, `revisit`, or `save` event or blank if no such events have happened yet. This will always be blank for `discover` events.
- **Document Start Time:** For `read` events, this is the start time of the document being read, if any. For `search`, `revisit`, `save`, and `assign` events, this is the value entered in the "Start Time" field of the query, if any. For `switch` events, this is the value from the most recent `search`, `discover`, `revisit`, or `save` event or blank if no such events have happened yet. This will always be blank for `discover` events.
- **Document End Time:** For `read` events, this is the end time of the document being read, if any. For `search`, `revisit`, `save`, and `assign` events, this is the value entered in the "End Time" field of the query, if any. For `switch` events, this is the value from the most recent `search`, `discover`, `revisit`, or `save` event or blank if no such events have happened yet. This will always be blank for `discover` events.
- **Origin (Latitude / Longitude / City / Country):** For `read` events, this is the origin of the document being read. Latitude and longitude are only available for `read` events. For all other events, only city and country are available. For `search`, `revisit`, `save`, and `assign` events, this is the value chosen from the "Location Origin" drop-down menu in the query, with was a list of cities or `Any` by default. For `read` events for documents with no origin, the values will be `None`. It was also possible to choose `None` from the drop-down menu in the query in order to search for documents with no origin (e.g. information). For `switch` events, this is the value from the most recent `search`,

discover, revisit, or save event or blank if no such events have happened yet. This will always be Any for discover events.

- **Destination (Latitude / Longitude / City / Country):** Same as origin, but destination.
- **Reasoning Type:** Before executing a search query or a discover event (via the web search or recover deleted files tools), analysts were prompted to explain why they were performing that search. This is a discrete value from a drop-down menu whose values include "Refining or repeating a previous query," "Just exploring the data," "Looking for a specific document or specific information," "Testing the search tool," and "Not Listed." Before save events, analysts were prompted to explain why they considered the query or document important. This is a discrete value from a drop-down menu whose values include, "I found what I was looking for," "I did not find the thing I was looking for," "I found something useful on accident," and "I'm saving this search to come back to later." For read and switch events, this is the value from the most recent search, discover, revisit, or save event or blank if no such events have happened yet. For revisit events this value is the reasoning type of the saved event that the analyst is revisiting.
- **Reasoning Text:** Before a search, discover, or save event, analysts were prompted to explain their intentions (see above). Besides choosing an option form a drop-down menu, they were also allowed to write a free text response. This is the text of that response. For read and switch events, this is the value from the most recent search, discover, revisit, or save event or blank if no such events have happened yet. For revisit events this value is the reasoning text of the saved event that the analyst is revisiting.
- **Suspect:** For assign events, the analyst uses a saved query as evidence that one of the five employees of the company is either guilty or innocent. This is the suspect chosen. Values include Michelle, Daryl, Xin, Ivan, and Anna. This will always be N/A for all other events.
- **Hypothesis:** For assign events, the analyst uses a saved query as evidence that one of the five employees of the company is either guilty or innocent. Values include Guilty and Innocent. This will always be N/A for all other events.
- **Readable:** Contains a human readable summary of the event.

## Demographics

Results from the post-survey on subject demographics is provided in demographics.csv, which is in comma separated value format. The meanings of each column are:

- **Scenario:** Indicates which scenario this user belongs to: michelle-cargo for the original scenario and ivan-karikampani for the second updated scenario. Only users 20 through 24 used the second scenario.

- **User ID:** A unique index number (starting at 0) for each user. This number matches the User ID in the events file.
- **Duration:** The total time that subject spent on the exercise, in seconds.
- **Education Level:** The subject's level of education. Options included "High school diploma or less," "Some college or university, up to associate's or bachelor's degree," "Some graduate education or more," and "Prefer not to disclose."
- **Age Range:** The subject's age. Options included "30 or younger," "31 to 49," "50 or older," and "Prefer not to disclose."
- **Type of Analyst:** The subject's answer to the question, "Which of these best describes the occupational role that you held or have held for the largest part of your professional career?" Options included, "IA: Intelligence Analyst," "LA: Language Analyst," "Other Analyst," "Researcher, data scientist, or developer," "Manager," "Other role not listed here," and "Prefer not to disclose."
- **Experience Level:** The subject's answer to the question, "How many years of experience do you have in the role you selected above?" Options included, "3 or fewer years," "3 to 7 years," "7 to 15 years," and "15 or more years," and "Prefer not to disclose."
- **Training:** The subject's answer to the question, "Have you ever had formal training as an intelligence analyst or language analyst?" Options included, "Yes," "No," "Uncertain how to respond," and "Prefer not to disclose."

## Images

Images of each event have been generated to approximate what the screen looked like for the user during each event. These images are not actual recordings of the user's session, but generated after the fact, so they are not a perfect rendering of what the user saw.

The format for an image's file name is `<UserID>_<StepID>.png`, where `<UserID>` matches the User ID from the events and demographics file, and `<StepID>` matches the Step ID in the events file. For example, `002_0052.png` is an image of what the user 2's screen looked like after their 52nd event.

## License and Classification

This serious game and its data was generated by researchers at the University of Kentucky and North Carolina State University for a project funded by the Laboratory for Analytic Sciences, which is part of the Department of Defense. This data is unclassified and may be used for other projects.