

INDIGO Hack and Learn: Technical and Learning Report

March 2024



Insper METRICiS
Center for Socio-Environmental Impact Management



This report is part of the GO Lab-supported International Network for Data on Impact and Government Outcomes (INDIGO). The report aims at reflecting on the learnings and challenges from participants and challenge leaders of the March 2024 edition of the Hack and Learn event. This is not an academic research report nor is it an evaluation.

Please cite this report as follows:

Nagarajan, S., Assunção, M., Chauhan, M., Grennan, E., Santos, R. & Veres, P. (2024) INDIGO Hack and Learn: Technical and Learning Report - June 2024. Government Outcomes Lab, University of Oxford; South African Medical Research Council; Insper Metricis, Insper Institute of Education and Research.

Publication date: June 2024

This work is licensed under a Creative Commons Attribution 4.0 International License: <https://creativecommons.org/licenses/by/4.0/>



Executive summary

This report reflects on the learnings and experiences of the participants of the INDIGO Hack and Learn event that took place in March 2024. Participants include attendees that participated in the team, challenge leaders (team leaders) and co-host representatives. The objective is to collect all of these experiences in a single document, forming the basis for a discussion on how to move forward with different pieces of work, and how to design and improve future Hack and Learn events.

Table of Contents

Introduction.....	5
Who should read this report?.....	6
What is a Hack and Learn event?	6
Our timeline for this Spring Hack and Learn	8
Our challenges	9
Our learnings.....	10
Looking forward	15
What are we doing next?	15
I am interested in being part of INDIGO, how can I contribute?.....	16

Introduction

The International Network for Data on Impact and Government Outcomes - INDIGO - is a community of researchers, policymakers, and data enthusiasts who share an interest in using and reusing data on social outcomes. INDIGO is set up as a data and learning collaborative where different organisations share their data on a voluntary basis and learn from each other. The [Impact Bond Dataset](#) is an example of this collaborative approach. We believe that collecting more and better data on outcomes-based contracts and offering it to the community of practitioners in an accessible format is a key part of our mission.

In that context, our Hack and Learn events are one of our INDIGO engagement activities. We run these events twice a year and strongly encourage our friends and colleagues to join us for two weeks of hacking and learning. It is a unique opportunity to meet people from different backgrounds and collectively think of potential solutions to complex social problems. Participants include a range of policymakers, practitioners, students, and senior researchers who bring their diverse perspectives to the table and co-create an output to share at the end of the event.

Figure 1. Invitation to our 2024 Spring Hack and Learn event

The screenshot shows the Government Outcomes Lab website page for the 'Hack and Learn Spring 2024' event. The page features a navigation bar with 'The basics', 'Knowledge bank', 'Toolkit', and 'Community' tabs. The main heading is 'INDIGO HACK AND LEARN EVENT Hack and Learn Spring 2024'. Below the heading, it states 'ONLINE EVENT 7-21 MARCH (GMT)' and includes a 'Set my timezone' button. A large image of Earth with data lines is displayed. The text describes the event as an international gathering for practitioners, researchers, policy makers, students, and data enthusiasts to tackle social outcomes using data. Social media links for Twitter (#indigoinitiative) and LinkedIn are provided. A 'Register for Show and Tell session' button is visible. Below the main content, there is a section titled 'About INDIGO and the Government Outcomes Lab' with a plus sign icon. The 'Overview' section explains the event's purpose: to help people understand outcomes-based projects by connecting with others and working on real-life projects. It mentions that participants choose from data-related challenges set by the Government Outcomes Lab and its partners, and receive a letter of participation. The overview also notes that participants can engage in coding and data wrangling, research and policy writing, or graphic design to create data visualisations.

Who should read this report?

If you participated at our March 2024 Hack and Learn, you may wish to reflect on your own learnings, and those of your peers. In addition, as there is little time to think about other sub challenges in the intense two-week event, this is an opportunity to have a look at the work of others.

If you did not participate, this report offers a summary of two weeks of intense teamwork. You will find out about our initial proposal to participants, the work of the sub-challenge teams, and the outputs they presented at the end of the two weeks. Some of the sub-teams might be developing ideas that align with your work and this report will help you identify them. We also hope that you will be inspired to sign up for the next Hack and Learn event.

What is a Hack and Learn event?

Hack and Learn events are one of our INDIGO engagement activities. The ‘hack’ part of the event refers to our data investigations. We test new ideas, experiment with different analyses and visualisations of the numbers, and seek to fill gaps in the data. The ‘learn’ part refers to our policy orientation and collective knowledge sharing. We bring together policymakers, practitioners, students, and senior researchers who share an interest in solving complex problems in a data-driven way.

Figure 2. Image of our kick-off session on March 7th 2024



Hack and Learn phases

BLAVATNIK SCHOOL OF GOVERNMENT | UNIVERSITY OF OXFORD | GOVERNMENT OUTCOMES LAB

Kick-off → Hack, hack, hack → Show & Tell → Reflections & Learning Report

- 1. Kick off session: 07 March 2024**
HACK HACK HACK -> use your Slack channels.
- 2. Show and tell session: 21 March 2024**
Come and show your results and share your learnings. There will be a virtual social gathering after this session, all welcome!
All invited to reflect on their experiences and contribute to co-authoring a technical and learning report 😊

4:25 / 35:31

We kick off our event with a plenary meeting where we explain the available challenge to our participants. Each challenge has a challenge leader, who gets 5 - 10 minutes to pitch their idea. Participants have the opportunity to propose their own challenge. If they have a topic and some relevant data to work on, they can get 5 minutes to pitch their idea.

The teams then have two weeks to work on their challenge. The agenda is rather open: teams can develop a potential solution, use a pre-existing dataset to better understand a topic, or develop tools or prototypes to help practitioners with a particular problem.

After two weeks, we all get together at our Show and Tell session. Every team and subteam designates a presenter, who gets 10 minutes to tell the other participants what they worked on, which challenges and difficulties they faced, and how they plan to move forward with their work.

Figure 3. Image of our Show and Tell session on 21st March 2024



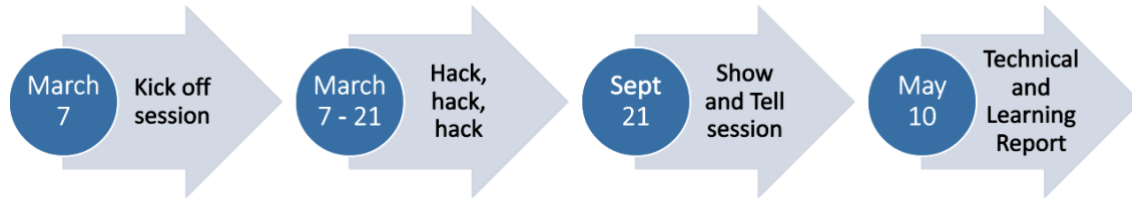
The end of the Show and Tell session is not the end of our Hack and Learn event. After the session, we invite all participants to share reflections on their experience at the Hack and Learn.

We have collated all these stories in this report, which will form the basis for a discussion on how to move forward with these different pieces of work, and how to design and improve future Hack and Learn events.

Our timeline for this Spring Hack and Learn

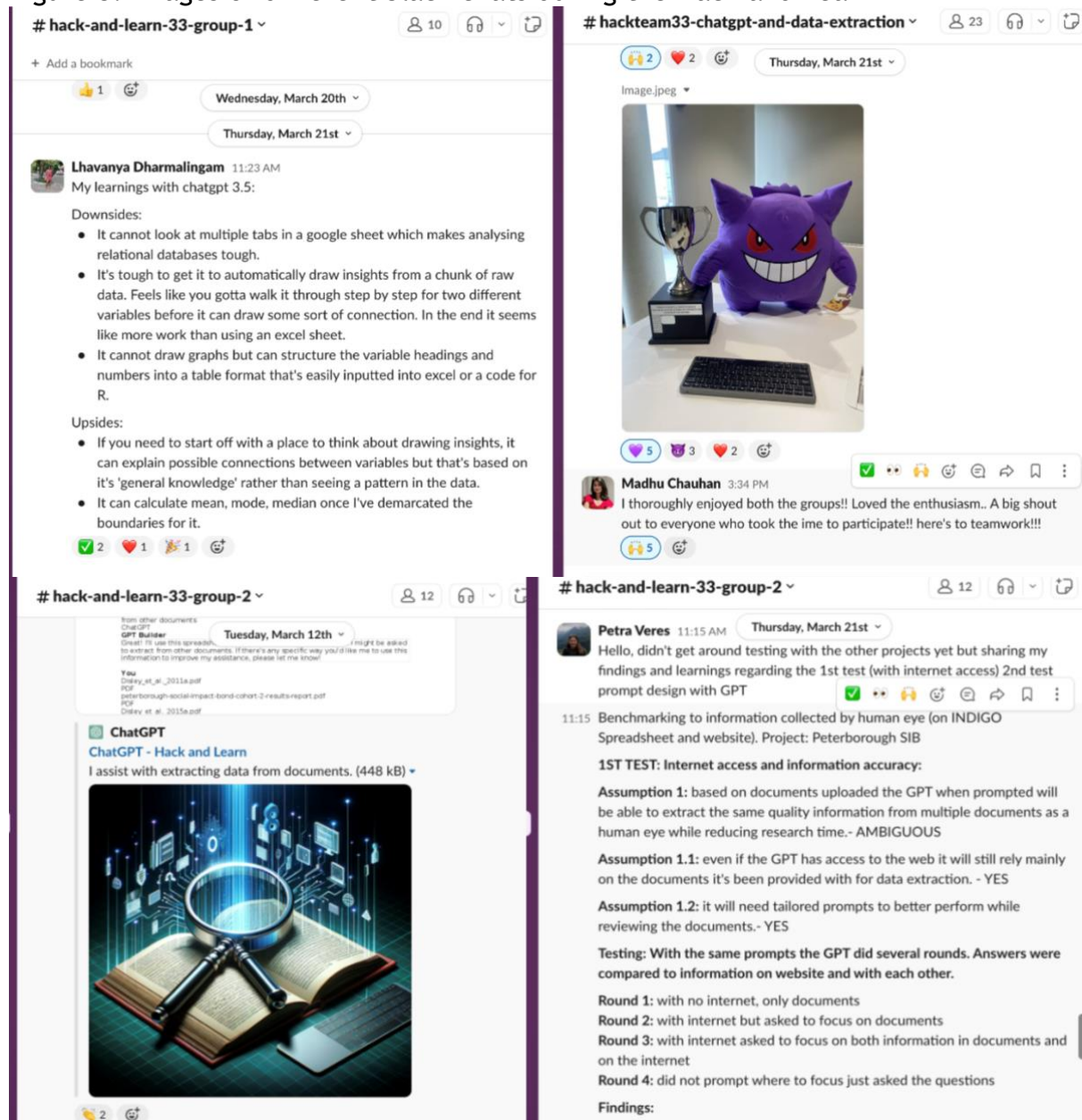
The challenge was carefully curated by the group of Hack and Learn co-hosts: INSPER Metricis (Brazil), the South African Medical Research Council (South Africa) and the Government Outcomes Lab (UK).

Figure 4. Timeline of Hack and Learn activities



The kick-off session took place on 7 March, and the Show and Tell session on 21 March. In those two intervening weeks, the different teams organised their work in such a way that every participant could contribute according to their time zone and availability. We had one challenge with two sub-challenges and every sub-challenge team had its own Slack chat where they could coordinate their asynchronous work. All the teams shared ideas, provided feedback for others, created data visualisation or undertook basic data analysis.

Figure 5. Images of different Slack chats during the Hack and Learn



Finally, before we get into the particular challenges, we want to give a special thanks our discussant who is an expert researcher on these matters, who kindly attended our final session and provided feedback for our participants: Professor [Rob Procter](#), Visiting Professor at the Blavatnik School of Government and Professor of Social Informatics and Deputy Head (Research) in the department of Computer Science at Warwick University . If you would like to watch the recording of this final session, you can access it [here](#).

Our challenges

Together with the co-host partners, we prepared a challenge for our participants. Below is our challenge description.

Hack and Learn Challenge: Data extraction using Artificial Intelligence

Embark on a mission to improve data availability on the INDIGO Impact Bond Dataset! The INDIGO Impact Bond Dataset collects data on impact bond projects from all around the world. It has immense potential, but it also has missing data for key variables. Not every impact bond project on the dataset has data for every variable of interest. Our challenge is to use ChatGPT to test if we can harness the power of AI to complete the gaps on the Impact Bond Dataset. Our goal is clear: to fortify the dataset, making it more comprehensive and reflective of global initiatives. By filling these gaps, we're not just improving the dataset's integrity but also paving the way for more insightful and accurate analyses by researchers, academics, and practitioners.

But that's not all—this challenge is about more than just patching up holes. We will also be thinking about the possibility of including AI and chat gpt in the day-to-day processes of data extraction and data entry. The Impact Bond Dataset is constantly updated by a group of data stewards, data analysts, and data officers. How would data extraction look like if ChatGPT4 were to be used as an aid?

Our learnings

The aim of Hack and Learn is not just to improve data in the field, but also to share learnings with others. At the end of the two weeks, we hosted a Show and Tell session on 21 March at 1.00pm GMT for each team to share their reflections and outputs. Along with our partners, we collated these lessons and shared them in this section.

Hack and Learn Challenge: Data extraction using Artificial Intelligence

This section was written by Madhu Chauhan, Data Officer at Government Outcomes Lab, and Matheus Assunção, Research and Policy Officer at Government Outcomes Lab

The INDIGO Hack and Learn Challenge #33 sought to leverage AI and ChatGPT to fortify the Impact Bond Dataset, addressing gaps in data completeness and currency. By enhancing the dataset's integrity, we aimed to facilitate more insightful analyses by researchers, academics, and practitioners worldwide, ultimately pioneering the future of impact bond data management.

The Impact Bond Dataset, a cornerstone of the Government Outcome Lab's data collaborative - INDIGO (International Network on Data for Impact and Government Outcomes), serves as a repository for open data assets related to outcomes-based contracting, cross-sector collaboration, and public service improvement. Housing information on 292 projects globally, this dataset holds tremendous potential to inform policy decisions and enhance social impact initiatives (Outes Velarde, J., et. al., 2023). However, the collaborative nature of its assembly renders it susceptible to vulnerabilities such as data incompleteness and timeliness issues, particularly affecting older and international projects.

Historical Challenges in Data Collection

Historically, data collection for the Impact Bond Dataset relied heavily on manual efforts by the data team at GO lab, involving collaboration between the GO lab, practitioners, and policymakers. This laborious process, comprising identifying relevant sources, information sharing, scanning, and updating data from lengthy policy documents, posed several challenges. These included time-intensive scanning processes, susceptibility to errors in manual recording, and the demanding nature of iterative updates, especially for tracking project progress over time.

Sprint 2024 Hack and Learn Goal

The growth in the use of AI-based tools such as ChatGPT has raised several concerns both in academia and in practitioners. The goal of the challenge in the spring 2024 edition of Hack and Learn was to test the efficacy of chat GPT around data extraction, data accuracy and automation of the data updating process exclusively for impact bonds.

Exploring New Questions

As we began un-wrapping the many layers hidden underneath this challenge some striking questions started to surface - Can ChatGPT optimise the time spent on data extraction? Can we rely on ChatGPT-extracted data for accuracy? Can ChatGPT serve as a substitute for traditional data scraping and updating methods?

A thorough test of efficacy of AI tools in data collection, updating, and pre-processing was a multi-fold task. The team parallelly worked in 2 groups to investigate 2 critical strands:

1. **Data Extraction and Enrichment:** Is chat GPT capable of mastering the art of extracting data associated **with** an impact bond from several policy documents and other sources and **populating** the extracted information into the INDIGO database? Furthermore, does it enhance the reliability and depth in data extraction therefore improving the overall quality of data and optimising the process of data extraction?
2. **Automated Workflow Prototyping:** Is it possible to design automated workflows that leverage both programming and AI capabilities to maintain a dynamic, up-to-date database, reflecting the latest trends and data points for various projects in the INDIGO dataset?

The challenges were designed in a way that a participant could be involved in one or both the challenges and there was a swift build-up across the level of complexity.

Two of the pre-existing projects were used as a source to emulate data extraction and automated data updating using chat GPT. [Essex County Council Multi-Systemic Therapy \(MST\) - INDIGO-POJ-0130](#) and [HMP Peterborough \(The One Service\) - INDIGO-POJ-0153](#) were selected as sources for the entire process. Using the same projects to test various dimensions helped us build reliable inferences from the analysis.

For the first strand of the challenge, Participants engaged in a comprehensive exploration to ascertain ChatGPT's capability in extracting data associated with impact bonds from policy documents and other sources in .pdf or .docx formats, with the goal of populating the extracted information into the INDIGO database. Their approach encompassed multiple facets, including extracting essential quantitative variables such as project names and key dates, querying ChatGPT to summarise qualitative information such as intervention descriptions and identifying themes. Finally, compare the results with existing data in the INDIGO database. While ChatGPT demonstrated some proficiency in retrieving details, challenges arose in achieving the required level of completeness, particularly in navigating through complex data structures such as multi-tab Excel files. Additionally, concerns regarding data credibility surfaced due to issues related to source access and data completeness.

Participants explored methods such as tone and sentiment analysis to validate the accuracy of AI-generated qualitative data, highlighting the potential dual-use case of leveraging AI for completeness checks and verification of manual recordings. Educating ChatGPT with informative resources and definitions related to impact bonds improved output verification but revealed limitations in the model's capacity to maintain performance with extended prompts. Nonetheless, the generation of a Data Completeness Report proved informative in identifying critical gaps in detailed project data and outcome metrics. Furthermore, assigning ChatGPT individual roles and summarizing data from a project demonstrated the model's ability to distil situation-specific qualitative information, fostering trust in its capabilities.

The second strand of the challenge involved a more technical experiment with the goal of prototyping an automated workflow that could leverage AI capabilities that could help in keeping the data points on the INDIGO database up to date for all the pre-existing projects. Chat GPT 3.5 allows the users to build a custom model for a bot, which thrives on the resources that the user provides. Through this exercise, we aimed to build a SIB-specific/ INDIGO-specific bot using chat GPT for data validation and pre-processing and automating the integration process to ensure the INDIGO database is always current and comprehensive.

The discussions revealed valuable insights into the application of a custom ChatGPT bot designed for SIBs and INDIGO datasets, for data extraction and analysis purposes. Participants explored various methodologies, from training the model with specific documents to experimenting with internet access. While effectiveness was observed in extracting quantitative data, limitations emerged in handling qualitative responses, particularly in the absence of access to contract documents. Strategies such as utilizing project-related documents and defining a glossary of terms enhanced the bot's performance, showcasing the importance of context in improving accuracy. Challenges related to understanding complex datasets underscored the need for further refinement in prompt design and data pre-processing techniques. Additionally, discussions surrounding the impact of non-public information highlighted the importance of ethical considerations in data extraction processes. Overall, the exercise provided valuable learning opportunities in leveraging AI tools for data-driven insights, emphasizing the iterative nature of refining methodologies for optimal outcomes.

The challenge highlighted the potential of AI, particularly ChatGPT, in streamlining data extraction processes and improving data completeness. However, it also underscored the importance of understanding the limitations and challenges associated with generic AI tools. While ChatGPT was promising in extracting quantitative data, issues arose in handling qualitative responses and navigating complex datasets. One notable aspect of the discussion was the exploration of strategies to enhance ChatGPT's performance, such as providing informative resources and defining a glossary of terms. These strategies proved effective in improving accuracy and reliability, showcasing the iterative nature of refining methodologies for optimal outcomes. An additional dimension to this could be designing an AI-based model that specifically caters to the processing of data related to social impact bonds and training it to be coherent with extraction and processing for the INDIGO dataset.

Additionally, the exercise shed light on the ethical considerations surrounding AI-driven data extraction processes, particularly concerning the handling of non-public information. This raised important questions about privacy, consent, and the responsible use of AI in data management.

Overall, the challenge served as a valuable learning experience, providing participants with insights into the potential and limitations of AI in data management and analysis.

Moving forward, it is essential to continue refining methodologies for leveraging AI in data extraction and analysis. This includes further exploration of strategies to enhance ChatGPT's performance, such as refining the prompt design and improving data pre-processing techniques. Additionally, there is a need to address ethical considerations surrounding AI-driven data extraction processes. This includes developing guidelines and best practices for the responsible use of AI in data management, with a focus on privacy, consent, and transparency.

To conclude, ongoing collaboration between researchers, practitioners, and policymakers will be crucial for advancing the field of AI-driven data management and analysis. By working together, we can continue to innovate and develop solutions that improve the integrity and efficacy of datasets like the Impact Bond Dataset, ultimately leading to better-informed policy decisions and more impactful social initiatives.

Key lessons learned from this challenge

- **Efficacy of AI in Data Extraction:** ChatGPT showed promise in extracting quantitative data from policy documents but faced challenges with completeness and navigating complex data structures like multi-tab Excel files.
- **Quality of AI-generated Data:** While the tool could extract key information, verifying the accuracy and depth of this data posed challenges, particularly for qualitative data. Strategies like tone and sentiment analysis were used to validate the data's accuracy.

- **Improving ChatGPT's Performance:** Providing ChatGPT with informative resources and a glossary of terms specific to impact bonds enhanced its output accuracy. However, maintaining performance with extended prompts revealed limitations.
- **Automated Workflow Development:** There was progress in designing automated workflows that could maintain a dynamic, up-to-date database using AI capabilities, although limitations in handling qualitative responses were noted.
- **Ethical Considerations:** The challenge highlighted the importance of addressing ethical considerations like privacy and the responsible use of AI, especially when dealing with non-public information.
- **Custom AI Development:** The potential to design AI models specifically tailored to social impact bonds and the INDIGO dataset was recognised, suggesting further exploration to improve data extraction and processing.
- **Iterative Improvement and Collaboration:** The iterative nature of refining AI methodologies was emphasised, along with the need for ongoing collaboration among researchers, practitioners, and policymakers to advance AI-driven data management.
- **Feedback and Continuous Learning:** Participant feedback underscored the benefits of an iterative approach to working with AI tools, where difficulties encountered.

Comments from participants:

'It has been a highly engaging endeavour to trial using a customised Chat GPT-4 bot, through the framework of the 'Auto Flow Fiesta' challenge. Through this experimentation, we found that the accuracy of the bot's data extraction was considerably improved when we took time to assimilate it to the specific contexts of our research, in addition to the intended data extraction process. Another key learning includes the benefit of working with the bot to understand its present flaws and improve its training. In other words, when the bot encountered difficulties in extracting data, we probed it to ascertain the reasons behind its shortcomings and utilised this feedback to refine its training regimen. Moreover, this iterative approach often highlighted the importance of framing questions in a more explicit and deliberate manner to facilitate more accurate responses from the bot. Due to the success of this challenge, moving forwards I am keen to continue trialling the use of AI to complement our data work.' *Eve Grennan, Data Officer at Government Outcomes Lab*

"I participated in the 'Auto Flow Fiesta' challenge at the Spring Hack and Learn event, where our team focused on designing and experimenting with a tailored GPT-4 bot. We trained a Peterborough SIB bot using ChatGPT's customised GPT function. We chose this old impact bond project because it offered more data and documents for our use.

After training the bot with interim and final reports and the Excel spreadsheet used by the INDIGO team, we tested our assumption that the GPT, if supplied with a

number of high-quality resources, could provide us with quality data. During this process, we found that while internet access does not greatly impact the accuracy of the information provided by the tailored GPT, the design of the prompts considerably influences the outcomes. Due to the GPT's processing time, the bot also struggles to find segmented information as it does not know where to focus across and within documents. While we experimented with different prompts and sought feedback from the GPT on why it wasn't able to find the needed information and about the prompts it would need, we did not obtain satisfactory results. We observed that identifying all relevant stakeholders and acquiring financial data presented the most significant challenges. Nevertheless, the GPT performed well in identifying location data, target population characteristics, and numbers.

"I believe that there is a great opportunity to further collaborate and explore ways we can use AI to aid in reducing the time spent on data extraction. While prompt design and data storage remain significant challenges, open-source large language models, as discussed in the Show and Tell session, can offer a way forward to design and test a closed, secure system over which the team has more control. I'm really looking forward to continuing these experiments and discussions!" - *Petra Veres, Social Entrepreneur and Researcher*

I had several highlights from my participation in the Hack and Learn. First, I was able to expand my knowledge not only in terms of the features that AI freely enables us to use but also in terms of how this could facilitate the extraction of information available in the vast world of the internet and its limits. Second, it was made clear how important it is to define a clear and specific prompt in the use of these tools. Prompts are not just about wording, but also about the role and context that will directly affect the outputs. For instance, "act as a financial analyst" can give slightly different answers from "act as a data analyst". Especially in the free version, within its limitations, a good strategy can be used to obtain similar resources. It also highlighted the importance of critical thought in analyzing the results. They can be confusing, misleading, and not 100% reliable. When working collaboratively, all these points are even more relevant. Prompts need to be shared so as to obtain the same results. Finally, the Hack and Learn also allowed me to learn more about the INDIGO platform, its projects, and the talented people behind it." - *Rafaela Santos, Master of Public Policy student at Blavatnik School of Government.*

Looking forward

What are we doing next?

At the GO Lab, we believe that learning with the community is a key part of our work. Keeping in touch with our participants is a key priority. Our [Slack channels](#) will remain active through the year, and we will keep sending invitations to learning activities or other opportunities through that channel. Those who expressed an interest in being added to our mailing list will be included in the INDIGO mailing group and receive invitations to [peer learning sessions](#) and other events.

Finally, we are now planning our upcoming Hack and Learn events. Your feedback and learnings will be used to design and improve this event. If you want to share your thoughts with us or suggest new challenges, you can send us an email to indigo@bsg.ox.ac.uk

I am interested in being part of INDIGO, how can I contribute?

INDIGO is a diverse community of peers with an interest in better data for better social outcomes. You can help us grow our community by attending our quarterly peer learning sessions, joining future Hack and Learn events and signing up for our mailing list. If you prefer social media, engage on Twitter using [@golaboxford](https://twitter.com/golaboxford) and [#indigoinitiative](https://twitter.com/indigoinitiative).

If you are involved in the delivery of an impact bond project, you can [share data with us](#). Email us at indigo@bsg.ox.ac.uk if you have any questions.