Data Science for Social Good

Scopeathon

Carnegie Mellon University

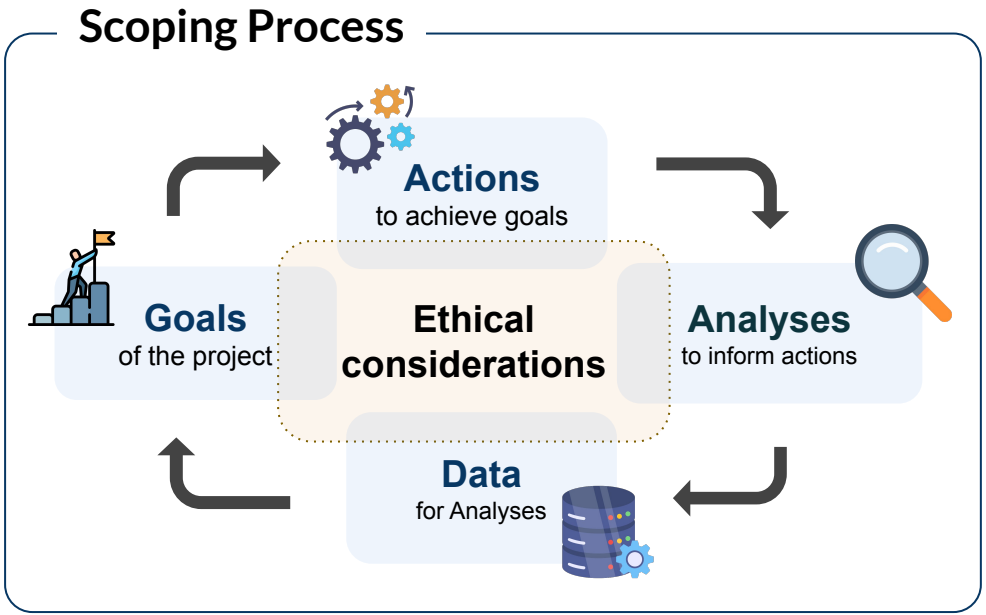
Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Why Scoping is Critical

- Increases the likelihood of use and impact
- Allows everyone to discuss, agree on, and then focus on tangible goals and identify actionable ways to achieve them.
- We start with the **problem** (and **not the data**) and get to a scope that allows us to connect the data to inform the actions to achieve our goals.

Actionable and Goal-Driven Project Scoping Process

Problem Statement

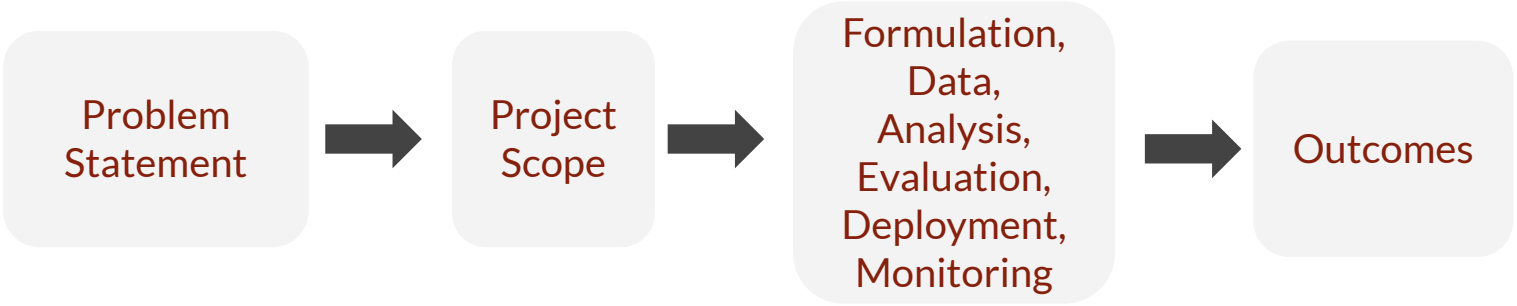


Project Scope

Iterative

Takes place over days/weeks/months

From Problem Statement to Outcomes



Conversations begin with problems an organization are facing

Available funds for the rental assistance program are set to significantly reduce. How do we target the limited funds to who need it the most ?

Eviction is a known pathway to homelessness. Can we use rental assistance funds as a homelessness prevention tool?

Current rental assistance program requires tenants who are at risk of eviction to initiate the contact to receive help. Can we proactively reach out to the vulnerable individuals and offer them help without waiting for them to contact us?

Scoping Process and its Outputs

Goals

What do we want to achieve?

How do we define success?

Actions

What actions can we take (or interventions we have) to achieve the goals?

Analysis

What analyses should we perform to inform the actions?

Data

What data do we need for the Analysis?

What data do we have?

How do we fill the gaps?

Ethical Considerations

What ethical issues do we need to consider during scoping, design, development, and deployment?

Example Project Scope

Goals

Minimize homelessness among tenants facing eviction

Efficiency: Maximize allocation of resources to individuals who will be homeless

Effectiveness: Minimize homelessness rate

Equity: Minimize disparities in homelessness rate across demographic groups

Actions

Proactively conduct outreach and provide rental assistance funds

Analysis

Predict which tenants facing eviction are most likely to interact with the homeless system in the near future.

Predict which tenants facing eviction are most likely to remain stably housed with rental assistance

Data

Individual level administrative data from the County's data warehouse:

Demographics, eviction cases, homeless services, mental & behavioral health and other county, state, federal programs and service usage

Ethical Considerations

Data Protection: Sensitive information about County residents needs to be protected

Discrimination/ Equity: Groups that are disproportionately affected by housing crises need to be considered

Trade–offs across Goals: Efficiency vs Equity

Transparency: Engaging Frontline workers & impacted communities about the new RA program

Data limitations: Does not capture non-court enforced evictions and does not capture individuals experiencing homelessness without interacting with county homelessness system

Analysis: Errors in model predictions can lead to assistance not reaching vulnerable individuals

Understanding the Problem

Before Scoping: Initial Screening

Is the problem real and significant?

High homelessness rates, low vaccination rates leading to adverse health outcomes, high prevalence of behavioral health crises

Would solving it positively impact society?

Can data play a role in the solution?

Can help decision making with additional information

Is the organization committed?

Solving the problem is a priority, data is available, willing to allocate resource to integrate the solution into their processes and assess its impact

Understanding the Problem

- What problem is the organization facing, and why is it a priority?
- What is the size of the problem: Who or what is affected? How much are they affected?
- How has the organization tackled the problem in the past/present?
- Who are the stakeholders inside and outside of the organization, and how should they be involved?
  - Policymakers/Problem or Program Leads
  - Data Team
  - Users
  - Impacted Community members

Understanding the Problem: ACDHS

Available funds for the rental assistance program are set to significantly reduce. How do we target the limited funds to who need it the most ?

Eviction is a known pathway to homelessness. Can we use rental assistance funds as a homelessness prevention tool?

Current rental assistance program requires tenants who are at risk of eviction to initiate the contact to receive help. Can we proactively reach out to the vulnerable individuals and offer them help without waiting for them to contact us?

Understanding the Problem: ACDHS

- What problem is the organization facing, and why is it a priority?
- What is the size of the problem: Who or what is affected? How much are they affected?
- How has the organization tackled the problem in the past/present?
- Who are the stakeholders inside and outside of the organization, and how should they be involved?
  - Policymakers/Problem or Program Leads
  - Data Team
  - Users
  - Impacted Community members

Use this session to get to know

- The organization (and people)
- The problem and why it's important
- Who does it impact and how much does it impact them?
- How (if) are they solving this problem today?
- Are there any big ethical issues that they know of right away that need to be considered?

web page with workspace link: **bit.ly/scopeathon25**

Defining the Goals

Step 1: Determine Goals

- Goals need to be measurable and concrete
- What outcomes would change if this project was successful?
- What constraints do you face in achieving these goals?
- What are the relative priorities and tradeoffs for each goal?

The Goal is **NOT** to build a model, make a prediction, build a dashboard, do some analysis, etc.

ACDHS Goals

- Allocate Rental Assistance Better
- Allocate Rental Assistance to those who are going to become homeless
- Maximize Allocation of Rental Assistance to those who are going to become homeless
- Maximize Reduction in Homelessness
- Allocation of Rental Assistance to Maximize Reduction in Homelessness
- Minimize disparities in rental assistance allocation
- Minimize disparities in homelessness
- Minimize disparities in homelessness-caused outcomes

Step 1: Define Goals

Efficiency

Focused on allocation of resources

Effectiveness

Focused on outcomes

Equity/Fairness

Focused on people not getting left behind

Pop Quiz: Which of these are reasonable goals for a project?

- Improve education quality in Pittsburgh Public Schools
- Teach students how to scope data science projects
- Minimize wait time for patients visiting ER
- Minimize change in average transit time for people getting to work from their homes
- Maximize the number of people who can get to a covid vaccine within 15 minutes

Examples of good goals

- maximize the likelihood that you'll be able to scope a project well in the future after attending the scopeathon today
- Improve education quality in Pittsburgh Public Schools
- Teach students how to scope data science projects
- Minimize wait time for patients visiting ER
- Minimize change in average transit time for people getting to work from their homes
- Maximize the number of people who can get to a covid vaccine within 15 minutes

Scoping Checklist for Identifying Goals

Make sure the goal is about **outcomes and impact** vs about doing something (building a dashboard or model or visualization or map)

Think about if the project was successful, how would the world change (for the people impacted)? What would improve?

Is the goal concrete and measurable vs improving some high level aim (such as education)?

Have you made a list of constraints (limited resources for example)?

Do you have explicit efficiency, effectiveness, and equity goals?

Have you discussed ethical issues around your goals?

Goals - Report Back

CORE

lower kidney non-utilization rate by 25% across the US

- reduce the number of kidney organs that are not utilized
- when?

Amachi

- maximize consumption to mentorship services for people at risk of involvement in the cj system in pittsburgh
- minimize incarceration rates

OBB

- reducing # of students involved in violence in PPS

Actions

Goals

Actions

We **achieve our goals** (and positively impact society) **only if we take actions**.

The role of the **data science/AI/ML system we may build** is to **inform those actions/interventions**

What are some examples of actions?

Health department **inspecting homes for lead hazards**

City government **doing preventative maintenance of water mains**

County mental health center **conducting mental health outreach**

A school system **enrolling students in after-school programs**

A nonprofit **sending targeted emails to potential donors**

Each action connects to a goal

Health department **inspecting homes for lead hazards** to **prevent lead poisoning**

City government **doing preventative maintenance of water mains** to **prevent breaks**

County mental health center **conducting mental health outreach** to **prevent crises**

A school system **enrolling students in after-school programs** to **support them graduate on time**

A nonprofit **sending targeted emails to potential donors** to **raise funds**

Identifying actions that help achieve the goal

Enumerate the interventions the organization has access to

Identify how those actions/interventions impact the identified goals

Get deeper into each action

- Who is taking the action?
- Is this something they already do? or setting up something new?
- Who is impacted by the action?
- Are there any resource constraints?
- Ethical issues?

Let's revisit our rental assistance example

Goals

Prevent homelessness among tenants facing eviction

Efficiency: Maximize allocation of resources to individuals who will be homeless

Effectiveness: Minimize homelessness rate

Equity: Minimize disparities in homelessness rate across demographic groups

Actions

Proactively conduct outreach and provide rental assistance funds

Who is taking the action: ACDHS and their partner organizations

Existing/New: Rental assistance program is existing, but proactive outreach will be new

Who is impacted: Individuals with eviction cases

Goals: Connects to all three goals

Constraints: Resources to reach ~30 individuals every week, Limited funds, no control over eligibility criteria

Ethical considerations: Not reaching individuals at risk of homelessness, how to help at-risk individuals who are not eligible for rental assistance?

Ethical Considerations

Checklist to identify relevant actions

Is the action concrete?

Go beyond – understand, identify, strategize, plan, observe…

Does it connect to a goal?

Who is taking the action? Does the org have control/influence over it? What information, if you gave it to the person taking the action, will improve the impact/goal?

Who/What is it performed on? What granularity? (e.g., performed on homes – individual homes vs all homes on a street)

Have you considered ethical issues?

e.g., any adverse consequences of acting on someone who do not need this intervention? Issues around excluding someone who needs the intervention?

Data

Goals

Actions

Data

What data do we have access to?

How do we identify and fill any gaps?

What data do we need?

What data do you have?

How far back does historical data go?

What is the granularity of the data?

How is the data collected and updated? With what frequency?

Our rental assistance example

Individual level administrative data from the County's data warehouse:

Demographics, eviction cases, homeless services, mental & behavioral health and other county, state, federal programs and service usage

What data do you need?

Data to inform Goals

Sufficient reliable historical data

Captures relevant outcomes

Identifies the cohort

Data to inform Actions

Refreshed at least as frequently as the action is taken

Covers the same granularity as the proposed action

Identifying Data Gaps and Readiness

Do we have enough to move forward, or do we need to focus on getting additional data?

- Do we have enough historical data?
- Is it reliable going back?
- Is it at the right level of granularity to inform the actions?
- Are different pieces of data linkable? unique identifiers for people, places, etc.
- Do we have the right data to measure outcomes (based on the goals)?

Example: preventing water main breaks in Syracuse

Key information about pipe age, diameter, and material was only available in 100-year-old handwritten notebooks of installation diagrams

Strategically digitizing these records enabled us to more accurately predict whether a water main would break

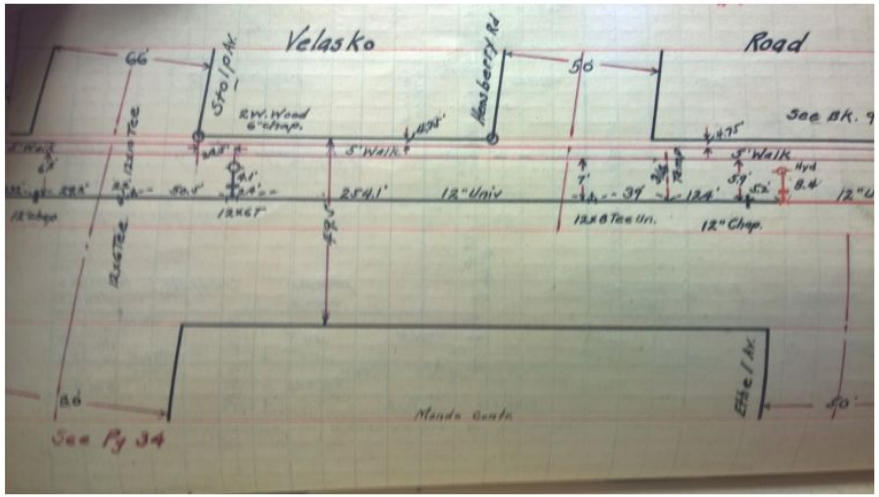


Figure 4: A hand drawn map of the water main layout on Velasko Road in the city of Syracuse from an original field notebook.

Returning to the rental assistance example

Goals

Minimize homelessness among tenants facing eviction

Efficiency: Maximize allocation of resources to individuals who will be homeless

Effectiveness: Minimize homelessness rate

Equity: Minimize disparities in homelessness rate across demographic groups

Actions

Proactively conduct outreach and provide rental assistance funds to 30 individuals every week

Analysis

Predict which tenants facing eviction are most likely to interact with the homeless system in the near future.

Predict which tenants facing eviction are most likely to remain stably housed with rental assistance

Data

Individual level administrative data from the County's data warehouse:

Demographics, eviction cases, homeless services, mental & behavioral health and other county, state, federal programs and service usage

What data do we need to evaluate the goals?
- Enough historical data to be confident in our predictions
- Data to identify outcomes: who becomes homeless?
- Data to identify cohort: who has an eviction filing?

What data do we need to take the action?
- Evictions data updated at least weekly
- Sufficient individual-level characteristics to predict the outcome
- Links between characteristics and evictions and homelessness data
- Contact information to outreach to the identified individuals

Ethical issues to consider

Data Privacy: What legal and security requirements govern how this data is used? What additional ethical considerations apply?

Data ownership and transparency: Are the people whose data it is (and is about) aware that their data is being used and how it affects them?

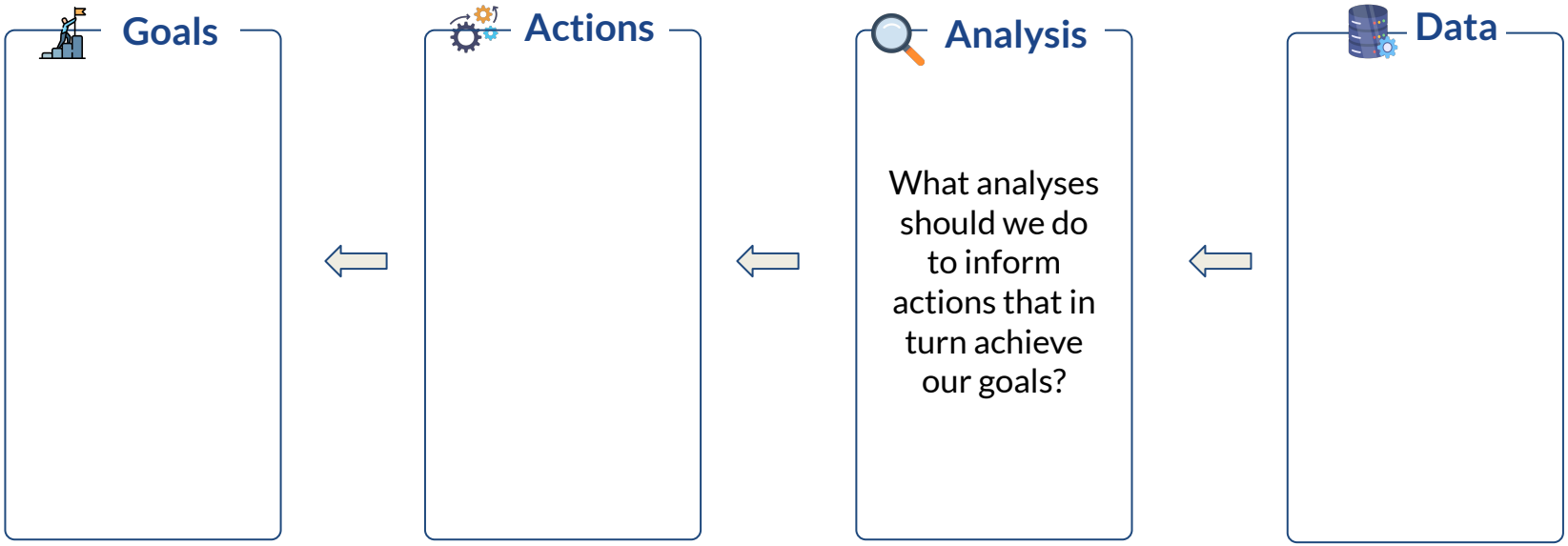
Bias, equity, and fairness: Are there biases in the data sources? Is the data equally accurate and available about everyone?

Scoping Checklist for Data

1. Have you explored data sources that you think are sufficient to tackle this problem?
2. Do you know the level of detail, history, and format of the data?
3. Have you explored ethical issues with each data source?
4. Have you identified any data gaps?

Analysis

What analyses should we do to inform actions that in turn achieve our goals?



Types of Analyses

Description - Understanding the past

e.g., Which zip codes have the most eviction cases? How many students failed to graduate on time last year?

Detection - Identifying a pattern/characteristic of the present state

e.g., Detecting fraudulent credit card transactions, identifying unsafe traffic patterns from CCTV cameras

Prediction - Predicting a future state

e.g., Predicting the likelihood of becoming homeless in the next year, Predicting the crop yield for the next harvest

Optimization - Allocate/assign/schedule resources/$/people/assets

e.g., Where place ambulances to minimize response time? Out of the people who need help, who will most benefit from the help?

Behavior change - Intervention effectiveness, behavior changes

e.g., which after-school program would be most beneficial to a particular student? Which email version would persuade a potential donor to donate money?

Typically requires additional data

Descriptions - Understanding the past states

e.g., Which zip codes have the most eviction cases? How many students failed to graduate on time last year?

Detection - Identifying a pattern/characteristic of the present state

e.g., Detecting fraudulent credit card transactions, identifying unsafe traffic patterns from CCTV cameras

Prediction - Predicting a future state

e.g., Predicting the likelihood of becoming homeless in the next year, Predicting the crop yield for the next harvest

Optimization - Allocate/assign/schedule resources/$/people/assets

e.g., Where to place ambulances to minimize response time? Out of the people who need help, who will most benefit from the help?

Behavior change - Intervention effectiveness, behavior changes

e.g., which after-school program would be most beneficial to a particular student? Which email version would persuade a potential donor to donate money?

Similar data, but different purposes

Descriptions - Understanding the past states

e.g., Which zip codes have the most eviction cases? How many students failed to graduate on time last year?

Detection - Identifying a pattern/characteristic of the present state

e.g., Detecting fraudulent credit card transactions, identifying unsafe traffic patterns from CCTV cameras

Prediction - Predicting a future state

e.g., Predicting the likelihood of becoming homeless in the next year, Predicting the crop yield for the next harvest

Optimization - Allocate/assign/schedule resources/$/people/assets

e.g., Where place ambulances to minimize response time? Out of the people who need help, who will most benefit from the help?

Causal In

effective ... ess, behavior changes

... ersion would persuade a potential donor to donate money?

Typically, a project involves multiple analyses

Identifying analyses that inform actions

Start with the action

Identify what additional information can improve the action

Identify analyses that can provide this information

This can be a combination of several analyses

What are the ethical issues with the analyses?

e.g., could your analysis be "more wrong" about historically marginalized groups? Will acting on this analysis output worsen outcomes for some groups?

Let's revisit our rental assistance example

Actions

Conduct outreach about RA and allocate RA funds (to prevent homelessness of individuals facing eviction)

Information needed:

Who is currently facing eviction?

Who is at risk of future homelessness?

Who will remain stably housed if given RA?

What's the optimal allocation of limited rental assistance funds to balance the goals of efficiency, effectiveness, and equity?

Detection

Prediction

Causal Inference/ Behavior Change

Optimization

Scoping checklist for identifying analyses

Is the analysis informing an action?

What type of analysis is it?

Description Detection Prediction Behavior Change Optimization/Allocation

Do we have access to relevant data to perform the analysis and validate it?

What are the ethical issues?

e.g., what happens when the analysis produces ‘wrong’ information? Will acting on this analysis output worsen outcomes for some groups? How can we mitigate disproportionate impacts?

Ethics

Ethical Considerations

Ethical issues are embedded continuously and in each part of the scoping process

We want to revisit them here and make sure we have identified the key ones that need to be discussed early on

Ethical Considerations

- **Privacy & Confidentiality**
  - Are you working with personal and/or sensitive data that is individually identifiable?
  - How are you protecting the data?

- **Data Ownership**
  - Do the people who “own” the data know you’re using it?
  - Do you have their permission? How was it obtained?
  - How will/Can they opt out of their data being used?

Ethical Considerations

- **Transparency**
  - Which aspects of the project do different stakeholders (policymakers, frontline workers, people who will be affected by the actions, the general public) need to be informed about?
  - Do the people who “own” the data know you’re using it?
  - Do the people you’re “targeting” know why and if they’re being “targeted”?
  - What recourse do they have to challenge a decision that comes from your analysis?

Ethical Considerations

- **Discrimination/Equity**
  - Are there any specific groups for whom you want to ensure equity of outcomes?
  - How do you define, detect, and increase equity in outcomes?
- **Accountability**
  - Who are the people responsible and accountable for all the things above?
- **Social License**
  - Would it make the front page of the national newspaper if they found out what you're doing?

Scoping Checklist for Identifying Ethical Issues

1. We don't need to have all the answers but it's important to list out the questions that need to be discussed
2. Have you identified who needs to be involved in these discussions during the scoping process and the project?
3. Have you considered how the project scope changes based on the answers to these questions?
4. What about implications of people impacted downstream and what can be done now to understand and manage them?

Wrap-Up

A Few Things to Remember

- Don’t be afraid to ask naïve questions
- We need to make sure that we tackle these problems responsibly and ethically
- Spend time discussing goals and metrics – don’t forget equity as a goal
- Understand what the current process or solution is
- Communication is critical – before, during, and after
- Data and technology does not solve problems, people do.

Useful Resources

Scoping guide:
https://datasciencepublicpolicy.org/our-work/tools-guides/data-science-project-scoping-guide/

Scoping worksheet:
https://datasciencepublicpolicy.org/wp-content/uploads/2025/02/Blank-Worksheet-January-2025.pdf