Data Science Project Scoping: A Guide for Social Good Organizations

Rayid Ghani

Carnegie Mellon University

Initially developed at: Center for Data Science & Public Policy
THE UNIVERSITY OF CHICAGO

Extended in collaboration with: Gob_Lab UAI
UNIVERSIDAD ADOLFO IBÁÑEZ
ITAM

Agenda

- Guide to Project Scoping for Social Good Organizations
- Scoping Curriculum
- Training Programs
- Resources

How do we scope projects that are actionable and result in (positive) social impact?

Why Scoping is Critical

- Increases the likelihood of use and impact
- Allows everyone to focus on the outcomes we should care about

Before Scoping: Initial Screening Criteria

- Real and significant problem (with clear social impact)
- Ability to act on the problem
- Priority and commitment from the institution (for people, data, validation and deployment)
- Data accessibility
- Identification of risks

Actionable and Goal-Driven Project Scope

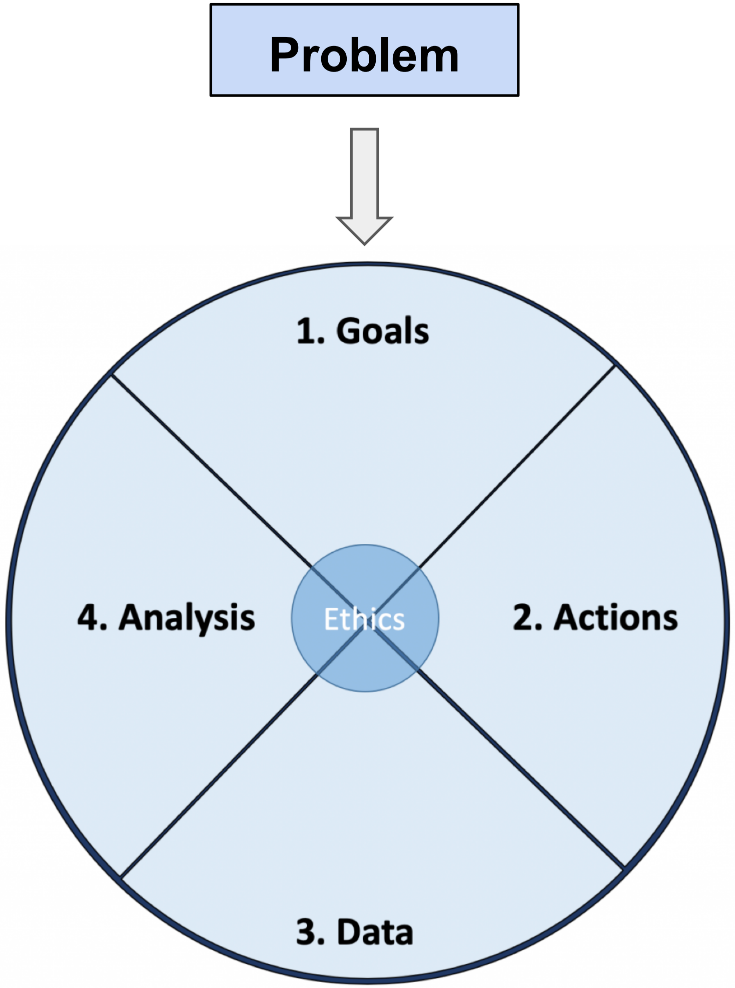
1.**Goals:** Define the goal(s) of the project

2.**Actions:** What actions/interventions will you inform?

3.**Data:** What data do you have internally? What data do you need? What can you augment from external and public sources?

4. **Analysis:** What analysis needs to be done? How will it be validated?

Ethics: What are the privacy, transparency, discrimination/equity, and accountability issues?



Step 1: Determine Goals

Efficiency

Effectiveness

Equity/Fairness

Step 1: Determine Goals

- Goals need to be measurable and concrete
- Goal is NOT to build a model, make a prediction, etc.
- What are the relative priorities and tradeoffs for each goal?
- What constraints do you face in achieving these goals?
- Stakeholder involvement from the beginning is key

Problem Templates

- Can I detect X early?
- Can I determine which X to prioritize?
- How do I improve the scheduling and assignment of X?
- Can I route X more efficiently and effectively?
- Which policies do I modify to improve X ?
- How much impact is X having?
- Can I get data that helps me X ?

Types of Problems

- Early Warning Systems
- Compliance/Inspections/Audits
- Routing (Physical or Virtual)
- Scheduling
- Policy Evaluation
- Policy Hypothesis Generation

Step 2: Identify Actions to achieve the goal

- What interventions do I have access to?
- What would we do differently if we had more information/knew where the interventions were most likely to be effective?
- Informing these actions:
  - Who? (to target for each action)
  - What? (to say to them)
  - How? (to use different communication channels)

Step 2: Identify Actions to achieve the goal

- Focus on concrete actions
- Existing vs new actions
- Consider the granularity of the actions
  - e.g. students who need help generally vs specific program
- How frequently are interventions taken/planned?
- How far out does planning occur?

Step 3: Data Sources

- What relevant data sources do you have?

- What data do you need?
    - Important to match the granularity, frequency, and time horizon of the actions to the data

- What external data can you augment this with?

Step 3: Data Sources

Types of Data

- Program Level
- Transactional
- Spatial
- Text
- Images/Audio/Video

- Nobody knows what data the entire organization has
- Don’t get intimidated by legal acronyms thrown at you
- Data is never perfect – is it useful enough to improve over status quo?

Step 3: Data Sources

- How reliable is the data?
- How current is it?
- How much of it is computer-readable?
- How much of it is stored as notes, audio, photos, videos?
- What resources and authority do you have to collect more?

Step 4: Analysis

- What analysis needs to be done?
- What type of methods should be used?
- How will the analysis be validated?

Types of Analysis Capabilities

- Description (Understand the past)
- Detection (Anomalies, Events, Patterns)
- Prediction (Predict the Future)
- Optimization
- Behavior Change (Causal Inference)

Validation and Implementation Plan

- Go back to the metrics and goals defined at the beginning of the project
- Run a Pilot/Field Trial
- Deploy
- Set up Infrastructure and allocate resources to monitor “lift”

Data and AI Ethics Issues

Privacy

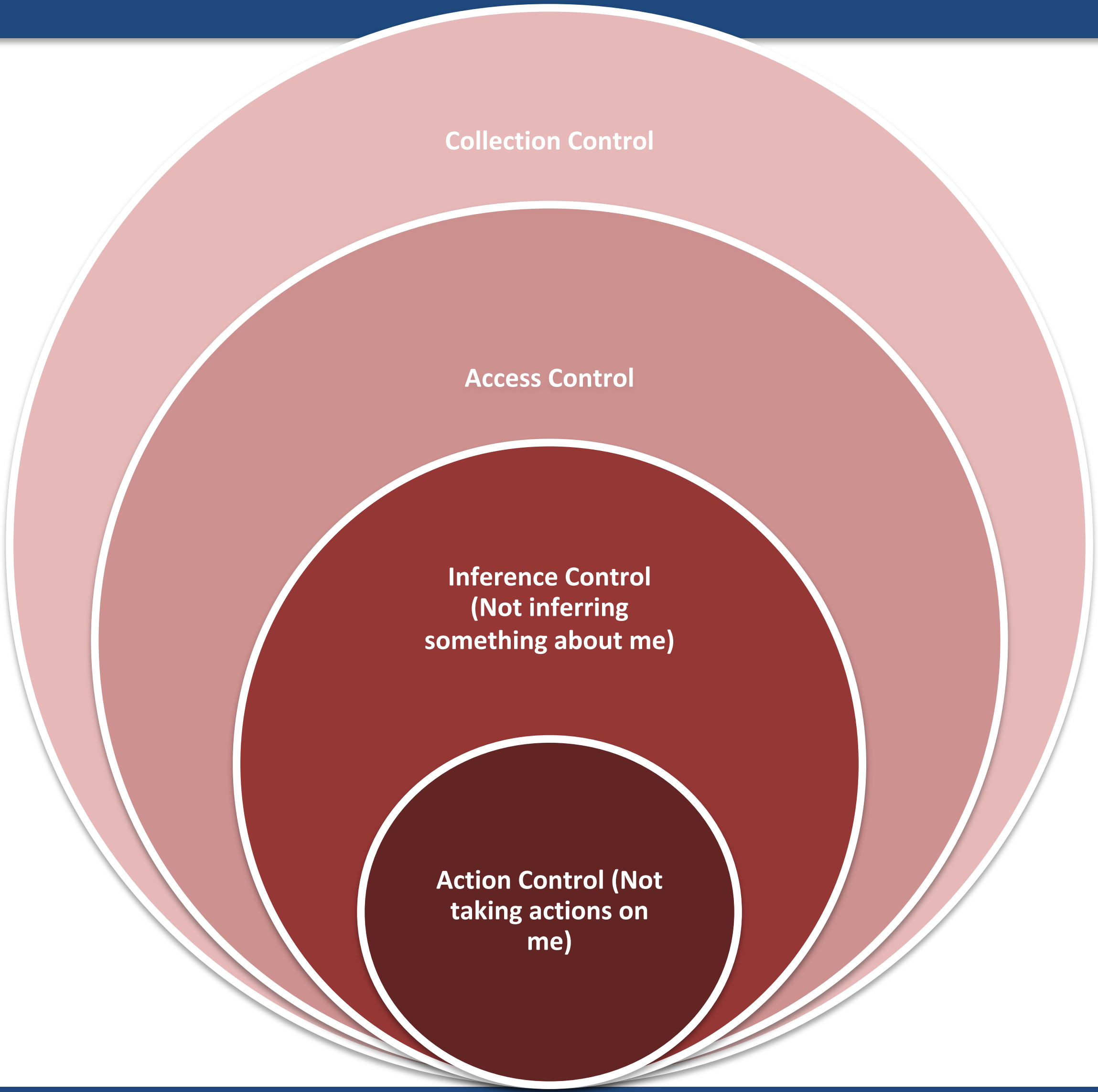
Data Ownership

Bias, Equity, & Fairness

Transparency

Trustworthiness and Accountability

Levels of control


Collection Control
Access Control
Inference Control (Not inferring something about me)
Action Control (Not taking actions on me)

Data Ethics Questions

- **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?

Data Ethics Questions

- **Transparency**
  - Do the people who “own” the data know you’re using it?
  - What actions are you taking on individuals based on this data?
  - Do the people you’re “targeting” know why and if they’re being “targeted”?
  - What recourse do they have?
  - Would it make the front page of the national newspaper if they found out what you’re doing?
  - Which stakeholders should know about which parts of the project?

Data Ethics Questions

- **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?

- **Social License**
  - If the entire population of the country finds out about your project, will they be ok with it? Why?
- **Accountability**
  - Who are the people responsible and accountable for all the things above?

- **Any other considerations such as consent, legal, etc**

Sample Curriculum for Scoping Workshops

Our curriculum is designed to teach how to:

- Determine the feasibility of using data science to address a problem facing government agencies
- Understand how to define the goals of the project and the actions that will be informed by this project
- Define a scope that can be turned into a project plan
- Understand the ethical challenges you must address during project scoping and execution.

Schedule for the pilot 3-day course

DAY 1

Analysis of Initial Criteria

Peer feedback

General overview of methodolgy

1. Defining a goal

HOMEWORK

DAY 2

Problem description

2. Actions

3. Types of data and data maturity

Iteration of actions and data

HOMEWORK

DAY 3

4. (Data) Analysis

5. Ethics

Final project iteration

How to begin a project

Conclusions

Resources

- Project Scoping Guide (Spanish version)
- Project Scope Worksheet
- Curriculum and Content for a 3-day course and video (piloted in Chile by UAI) available through creative commons license

- Upcoming courses:
  - Course in Chile
  - Course in Australia (March 9 and 12) in collaboration with University of Queensland

A Few Things to Remember

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

Summary

- Data (Science) can help build systems to improve policy and social outcomes in an efficient, effective, and equitable manner

- To get started, government agencies need to:
  - Identify policy and social goals
  - Identify actions and interventions
  - Develop data infrastructure
  - Build effective Data Science Systems
  - Validate experimentally and iterate
  - Budget for continuous monitoring & validation

Project Scoping Worksheet

http://bit.ly/datasciencescopingsheet

Credits

The scoping guide was initially developed by the Center for Data Science and Public Policy at the University of Chicago

The current version was extended (and translated to Spanish) through a collaboration with Adolfo Ibanez University (Chile) and is being maintained at Carnegie Mellon University.

Contact Information

Rayid Ghani
Carnegie Mellon University

rayid@cmu.edu
www.rayidghani.com

Carnegie Mellon University