Project Scoping Worksheet

This worksheet is designed to scope actionable analytics and data science projects for organizations (businesses, government agencies, non-profits, social enterprises, and others).

1. **Project Name:**

2. **Organization Name:**

3. **Names and Titles of Main Organizational Contacts:**

4. **Problem Description:**

    - **What is the problem you are facing?**

    - **Who/what is affected by this problem?**

    - **How many people/organizations/places/etc. and how much are they affected?**

    - **Why is solving this problem a priority for your organization?**

5. **Goals: What business goals will be accomplished by solving this problem and what constraints do you have? (In order of priority)**

    - The technical solution that will be built (e.g., predictive model or dashboard or map) is not the business goal - that is the tool that will achieve your goal
    - The goal should be specific and measurable
    - Achieving the goal should help solve the problem you're tackling
    - Typical goals include improving/maximizing/increasing or decreasing/mitigating/reducing some outcome or metric.
    - Typical constraints include budget, lack of human capital, legal restrictions, etc.
    - Consider tradeoffs between conflicting goals.
    - What is the value of a perfect solution? Is it worth the effort required?

	Goal	**Constraints**
1		
2		
3		

6. Data

- The data has to connect to the goals.
- Typical data science projects use administrative data as the primary data source, and enhance it with publicly available data sources (Census, other open data).

- **What data sources do you have internally?**

	Data Source 1	**Data Source 2**	**Data Source 3**
Name e.g., Hospital Admissions database			
What does it contain? Describe the attributes included in the data source. *e.g., admission and discharge records, patient sociodemographic data, insurance, physician, etc.*			
What level of granularity? *e.g., transaction, person, organization, location*			
How frequently is it collected/updated after it's captured? *e.g., real time, daily, weekly, monthly, yearly, one off*			
Does it have reliable and unique identifiers that can be linked to other data sources? *e.g., SSN, National identifier, patient identifier, insurance number, etc.*			
Who's the internal owner of the data? *e.g., IT department, individual, etc.?*			
How is it stored? *e.g., in a database, pdfs, excel, SPSS*			
Additional comments			

- **What data can you get from external, private or public sources?**

	Data Source 1	**Data Source 2**	**Data Source 3**
Name *e.g., Air Quality database*			
What does it contain? Describe the attributes included in the data source. *e.g., distinct pollution's particle concentration*			
What level of granularity? *e.g., geolocalized hourly sensor data*			
How frequently is it collected/updated after it's captured? *e.g., daily*			
Does it have unique identifiers that can be linked to other data sources? *e.g., sensor identifier*			
Who's the internal owner of the data? *e.g., NOAA*			
How is it stored? *e.g., API endpoint from an open data portal*			
Additional comments			

- **In an ideal world, is there additional data would you want to get/gather that would be relevant to his problem?** (surveys, phone records, etc.)

7. Analysis

- Typical data science projects include a combination of analysis, typically including description, detection, prediction, optimization, and/or behavior change.
- Again, the analysis is not the goal of the project - the **analysis** helps you use the **data** you have to inform the **actions** you have access to in order to achieve your **goals.**
- Choose the right set of analyses for each problem
- You must validate the analysis and use a validation process that matches how your analysis will be used in practice

	Analysis 1:	**Analysis 2:**	**Analysis 3:**
Analysis type *e.g., Description, Prediction, Detection, Behavior Change*			
Purpose of the analysis *e.g., understand historical behavior of individuals, probability of purchase, identify at risk for failure, etc.*			
How will you validate this analysis using existing data? What methodology and what metrics will you use? *e.g., Validate on future data using RMSE*			

8. Ethical considerations

Privacy Are you working with personal and/or sensitive data that is individually identifiable? Mention them.	
Transparency Which stakeholders should know about which parts of the project? *Stakeholders typically include leaders, managers, frontline workers, people who will be affected by the actions, etc*	
Discrimination/Equity Are there any specific groups for whom you want to ensure equity of outcomes? *e.g., groups of interest defined by gender, age, localization, social class, educational level, urban/rural, ethnicity*	
Social License If the country's entire population finds out about your project, will they be ok with it? Why?	
Accountability Who are the people responsible for all the things above?	
Any other considerations such as consent, legal, etc.	

9. **Who are the external organizations and internal departments that will need to be involved?**

(Typically, data science projects need involvement from data owners, IT infrastructure owners, problem owner, analytics people)

Organization/Department	**Description of desired involvement**	**Name/role of counterpart**
IT department	*Provide Data infrastructure*	*Head of IT department*
Statistics agency	*Provide population data*	*Head of Department of Statistics*

10. **Can you develop an experiment to validate the project in real life?**
    - Define how you will measure the success of the project in real life.
    - Describe how the model will be tested.
    - Define the duration of the experiment.