CAPS Case - Lecture 2
Data Collection Under Uncertainty

Leo Klenner, Henry Fung, Cory Combs

Last updated: 11/1/2019

Goal

This is the case we will discuss in the second lecture. The goal of this case is to familiarize you with the challenge of collecting relevant data for open-ended problems. Whereas we might often focus on knowing the right model, we first need to focus on acquiring the right data for solving the problem. Good machine learning models without the right data won't solve your problems. Data however has a cost - it needs to be purchased or collected in virtual or physical environments. Reasoning through what data might be most relevant to a specific problem is thus a critical skill and one that generalists can leverage to contribute to AI.

If you have questions, please contact us at capsseminar@gmail.com

Task

Your organization wants to build a machine learning model that can classify insurgent groups based on a specific factor: the extent to which a group's *organizational structure* is *centralized* or *decentralized*.

What does "centralized or decentralized organizational structure" mean? Essentially, it means whether the group is organized more like a hierarchy (centralized), or like a network (decentralized).

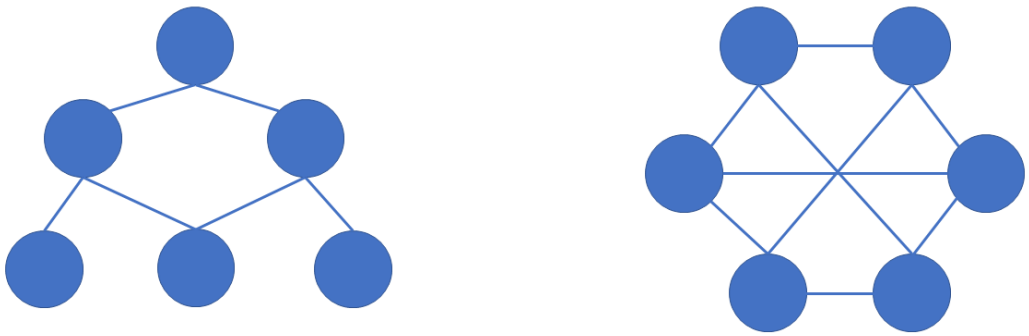

Figure 1: Centralized and decentralized organizational structure

Knowing this information is important to understanding the group's decision-making, ie. it might be easier to negotiate peace with the leader of a hierarchical group than with a member of a decentralized network.

There's a catch though - organizational structure is difficult to measure. There's no one variable you can collect data on to understand organizational structure. In addition, many insurgent groups are extremely opaque. Hence, even if there was one variable we could safely bet on, it would be hard to collect in practice.

Here, your organization turns to you for an answer.

You have the following tasks:

- Determine which *three* variables from the list below your organization should collect data on that can be used as a proxy for organizational structure
- For each selected variable, give a number between 1-10 for what you believe to be the *cost of collection* and *expected accuracy* of the variable, ie. going on the internet to check a government database is cheaper than building a network of human intelligence, but probably also leads to less accurate results
- Briefly think about the following questions: What would be the *ideal* data to collect for this problem, independent of the variables given below? Is there such a thing as an ideal dataset for this problem?

We will discuss your selection and the reasoning behind it in class. Although we'll show you a well-researched solution for estimating organizational structure from some of the variables below in class, this is an open-ended problem. What matters is not your solution but your thought process.

Variable	Description	Cost of Collection	Expected Accuracy
Allies, number of formal alliances	Number of the group's affiliations with other state or non-state actors		
Allies, number of joint operations	Number of the group's joint military operations conducted with other state or non-state actors		
Allies, economic strength	Economic strength of the group's affiliated state or non-state actors		
Attacks, number of casualties	Number of casualties caused by attacks committed by the group		
Attacks, number of VIP casualties	Number of VIP casualties (military leaders, politicians) caused by attacks committed by the group		

Variable	Description	Cost of Collection	Expected Accuracy
Attacks, locations	Locations of attacks committed by the group		
Attacks, type of targets	Type of targets (civilian, law enforcement, etc.) of attacks committed by the group		
Attacks, type of attacks	Type of attacks (ambush, improvised explosive devices, etc.) of attacks committed by the group		
Rival presence	What is the presence of local rivals of the group in the country in which the group is active?		
Counter-attacks, number of casualties	Number of casualties caused by attacks committed against the group		
Counter-attacks, number of VIP casualties	Number of VIP casualties (leaders) caused by attacks committed against the group		
Counter-attacks, locations	Locations of attacks committed against the group		
Counter-attacks, type of targets	Type of targets (buildings, group members) of attacks committed against the group		
Counter-attacks, type of attacks	Type of attacks (drone strike, raid, etc.) of attacks committed against the group		
Government, corruption	How corrupt is the government of the country in which the group is active?		
Country, infrastructure	What is the level of infrastructure (roads, etc.) in the country in which the group is active?		
Country, demographics	What are the demographics (income, religion) of the country in which the group is active?		
Country, duration of conflict	What is the duration of the conflict in the country in which the group is active?		

Variable	Description	Cost of Collection	Expected Accuracy
Diplomacy	Have there been previous attempts to establish diplomatic relations with the group, or has the group attempted to establish diplomatic relations?		
Government, political system	What is the political system of the government (democracy, dictatorship) of the country in which the group is active?		
Government, stability	How stable is the government of the country in which the group is active?		
Group, age	How long has the group been active?		
Group, hostages taken	Does the group currently hold hostages?		
Group, finances	What are the revenues and expenses of the group?		
Group, formation	How did the group form (as a splinter of another terrorist group or after a coup d-etat, etc.)?		
Group, phone metadata	Phone metadata (call logs, etc.) for parts of the group's membership		
Group, recruiting	Data on the groups online and offline recruiting efforts		
Group, religious cause	Does the group claim a religious cause?		
Prisoner of war interview	Interview a small number of prisoners of war about their motivation to fight for the group		