Afghan Profiles:
Finding Structure in Survey Data to Better Understand the Human Terrain

Thomas E. Powell, Ph.D.

University of Amsterdam
Faculty of Social and Behavioral Sciences
t.e.powell@uva.nl

Philip T. Eles, Ph.D.

NATO Communications and Information Agency
Operational Analysis Service Line
philip.eles@ncia.nato.int

Purpose

- To provide an example of advanced analytics conducted at NCI Agency in the area of Big Data and Artificial Intelligence
- To demonstrate how **unsupervised machine learning** techniques can extract structure from a complex dataset to obtain **operationally relevant results** in order to **support military decision makers**

Agenda

- **Background**:
  - NATO's public opinion polling in Afghanistan
- **Under the hood**:
  - Our machine learning approach
- **So What?**
  - Operationally relevant results

Background and Motivation

- NATO sponsors public opinion polling in Afghanistan
  - Program run by Joint Forces Command Brunssum for NATO Resolute Support Mission with NCI Agency as the technical lead
- NATO Surveys:
  - Inform "*Operational Assessments*" of progress against mission objectives (strategic to tactical)
  - Provide situational awareness of the operating environment to inform planners
- Surveys collect data on:
  - **Public perceptions** on a range of issues: Afghan government, security, economic well-being, public services and infrastructure + hot topics
  - **Demographic data** including age, gender, ethnicity, employment status, geographic region, urbanicity (urban/rural), literacy, etc.

Collecting Public Opinion Data in Afghanistan

- Household interviews are conducted by a 3rd party
- Interviews administered by local Afghans using methods sensitive to local culture/customs
- Survey designed to give nationally representative results
  - Interviews conducted in all 34 provinces
  - Both men and women over age of 18
- Survey sample size ~ 13,000 interviews
- NATO surveys conducted quarterly since 2008
  - 39 surveys completed to date
  - Cumulative dataset ~450,000 interviews

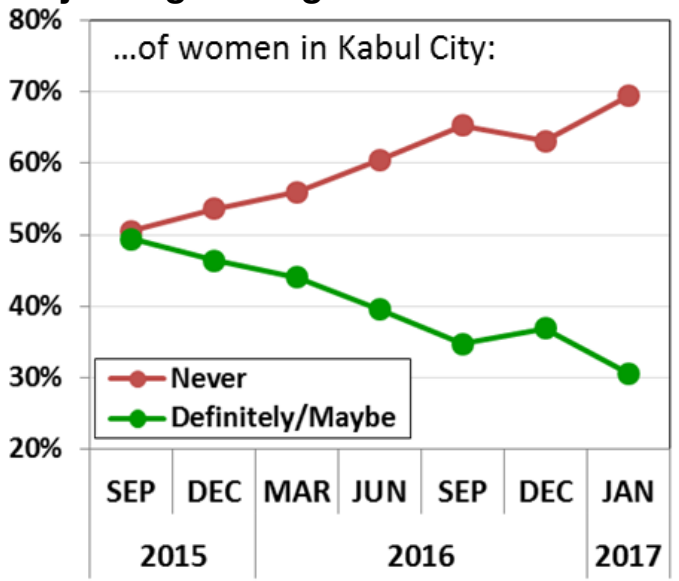
Survey Data Analysis

- Majority of data analysis is descriptive
  - Summary statistics; trends over time
- Manual disaggregation of data
  - By gender, ethnicity, region, etc.
- Richly **structured dataset** offers opportunities for application of **data mining and AI techniques** to help identify **operationally-relevant trends**
- Some initial development of exploratory techniques at NCI Agency

Employment of women in the Afghan National Police

		Women shouldn't work outside the home	An ANP Job is Inappropriate	An ANP Job Is Appropriate
Male	Pashtun	61%	28%	11%
	Tajik	33%	44%	22%
	Hazara	15%	61%	24%
	Other	35%	47%	18%
	AFG-Wide	45%	39%	17%
Female	Pashtun	33%	48%	19%
	Tajik	16%	60%	25%
	Hazara	12%	62%	25%
	Other	19%	62%	19%
	AFG-Wide	22%	57%	22%

Of you were looking for a job, would you consider joining the Afghan National Police?



Our Aim:

- Develop novel analytical techniques to support decision-makers
- Use unsupervised machine learning techniques to segment the population to…

  …identify demographics groups with common opinions…

  …to help better understand the human terrain…

  …and inform operations

Top-down approach

Qualitative desk analysis

Tailored population intervention

enables

Data-driven understanding

analytical scope

generates

Individual demographics

informs

supplemented by

Population grouping

generates

Assumption-based understanding

enables

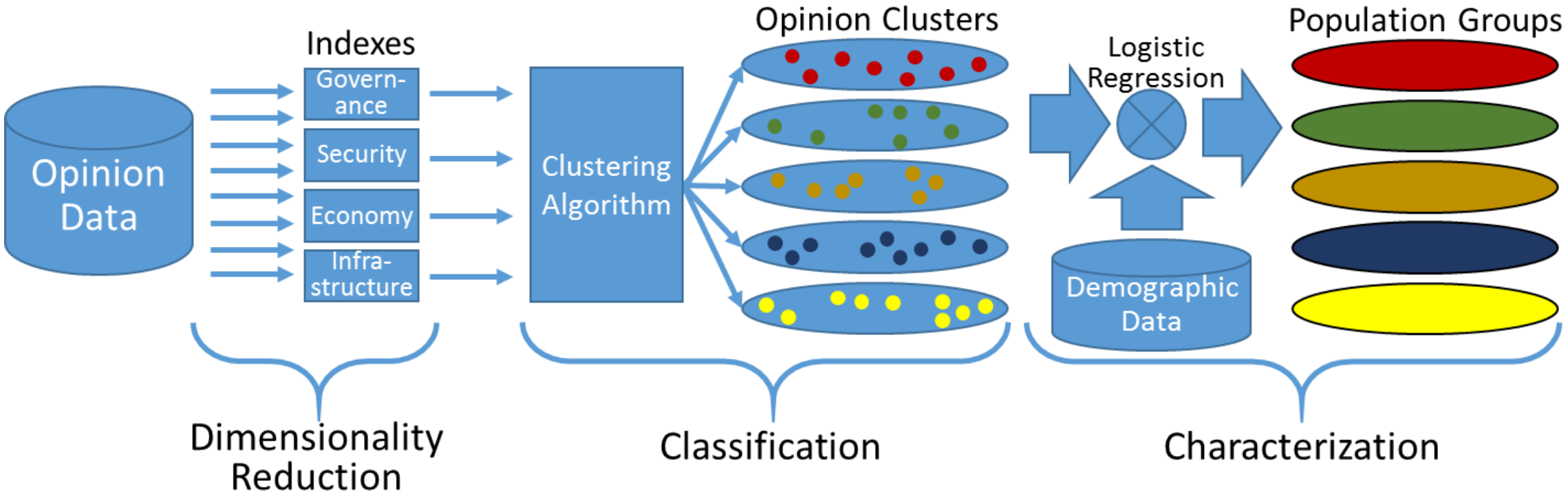
Coarse population intervention

Bottom-up approach

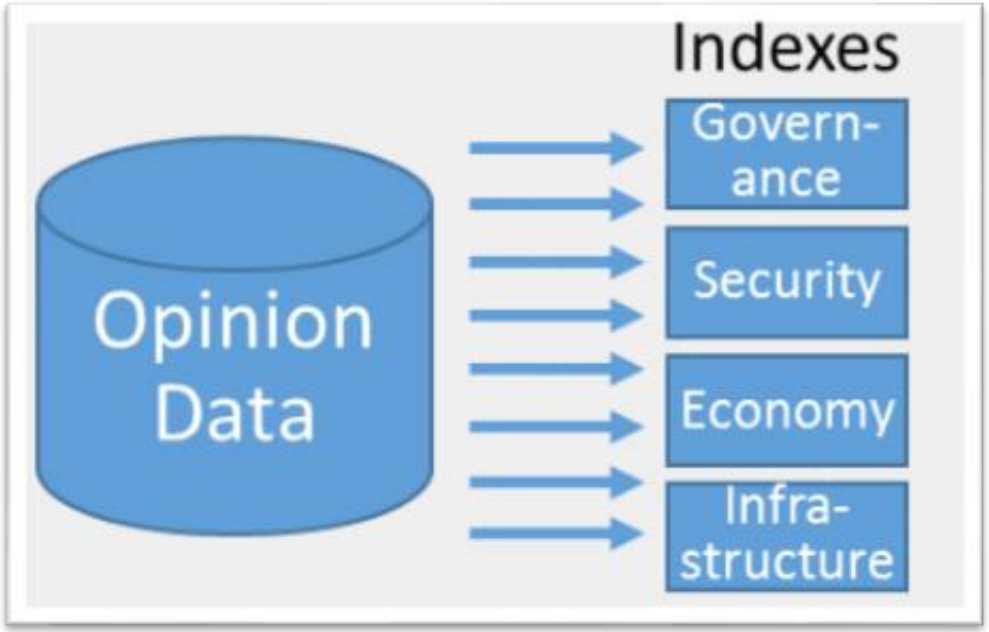
analytical scope

Quantitative cluster analysis

Population Segmentation through Unsupervised Machine Learning: "Under the Hood"



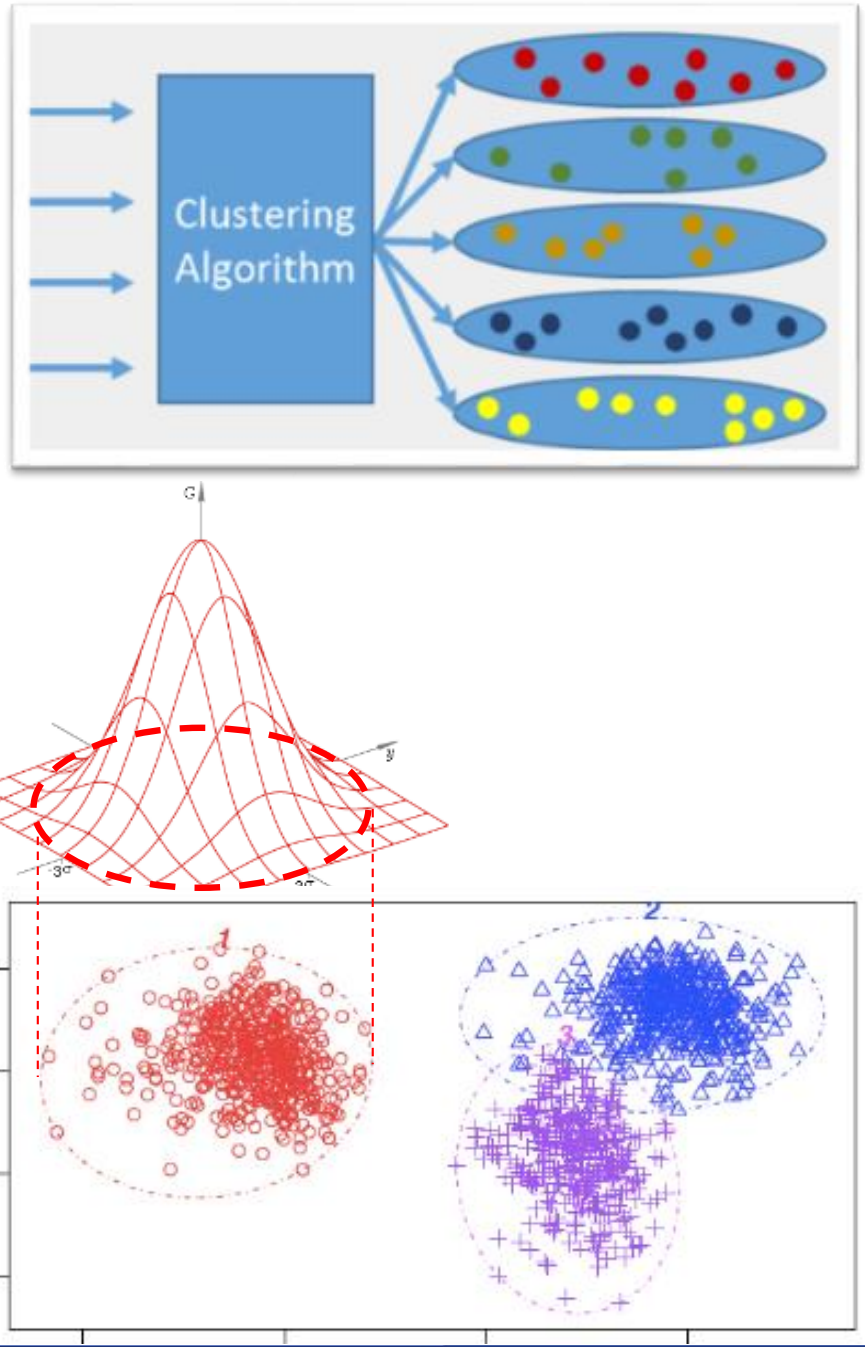
Dimensionality Reduction



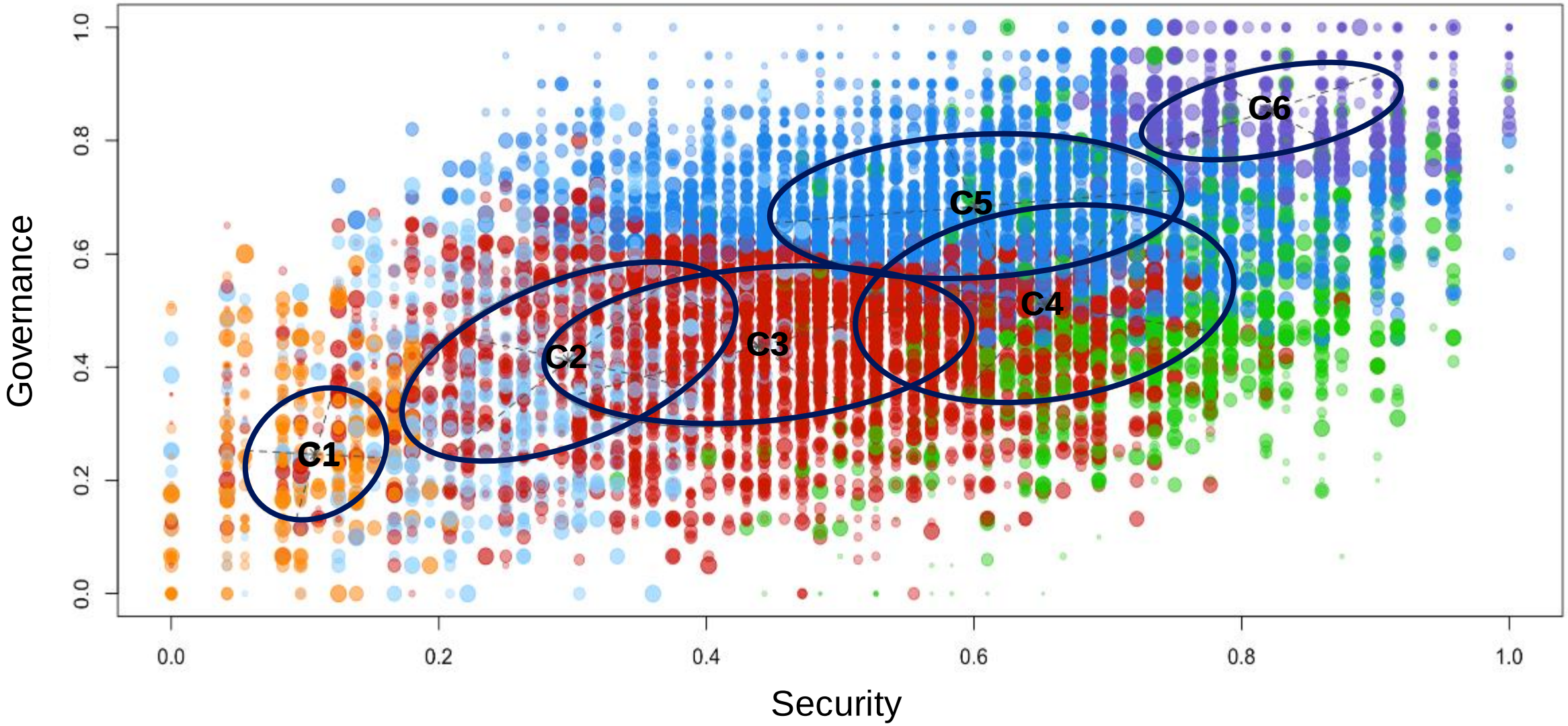
- Surveys include 100+ questions; each with 3-5 discrete response options
- HQ Resolute Support uses a core set of 38 questions in 4 topic areas:
  - Governance; Security; Economy; Infrastructure
- Core questions are **mapped onto 4 "indexes"**, one per topic area:
  - on a scale from 0 to 1
  - approximately continuous (resolution from 0.03 – 0.005)
- Indexes are a convenient way to map 38 discrete variables to a few continuous variables

Classification

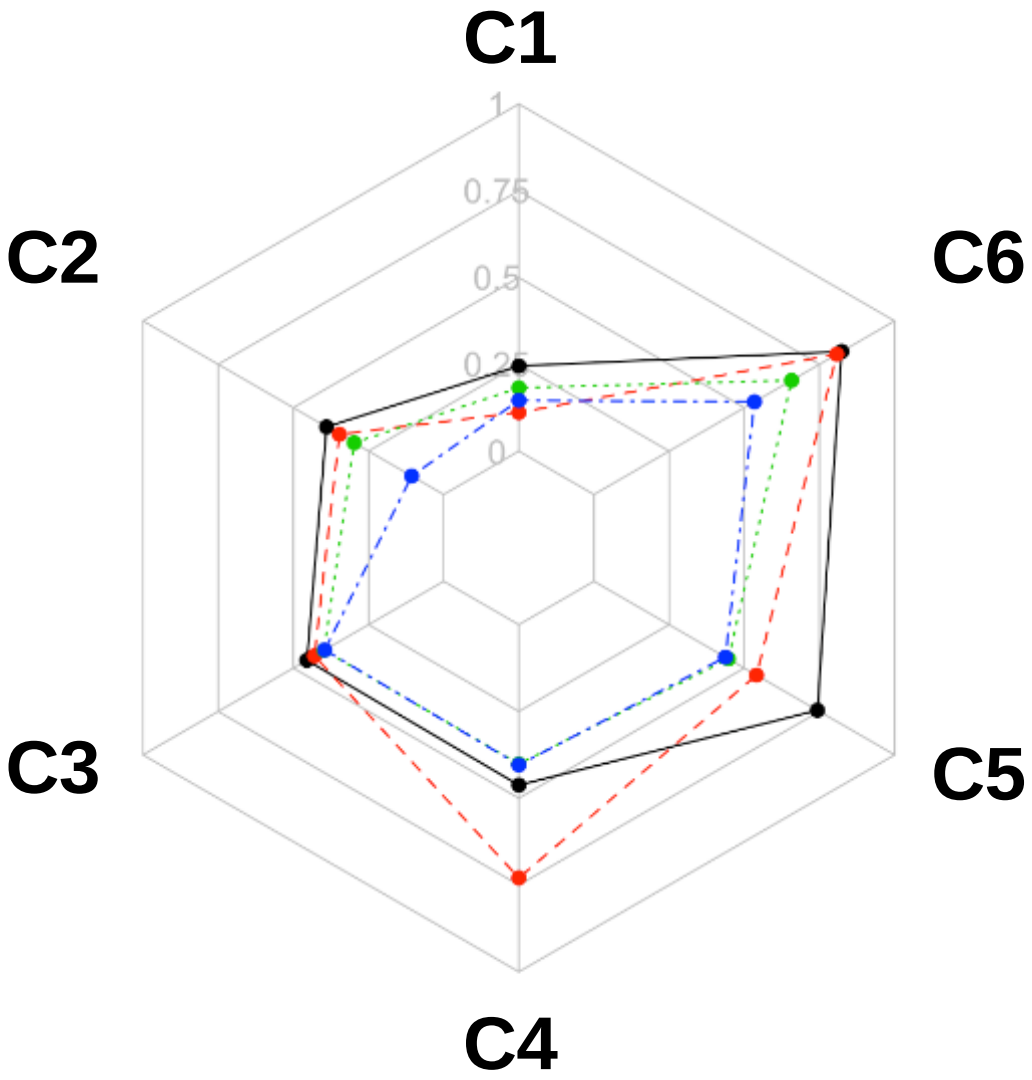
- Clustering algorithm applied to response data to identify "***opinion clusters***"
- Model Based Clustering Analysis (MCBA)
  - Assumes data clustered around two or more centroids, each with a statistical distribution
  - Shape, size, and number of distributions optimized through maximum likelihood fit



Results of Classification



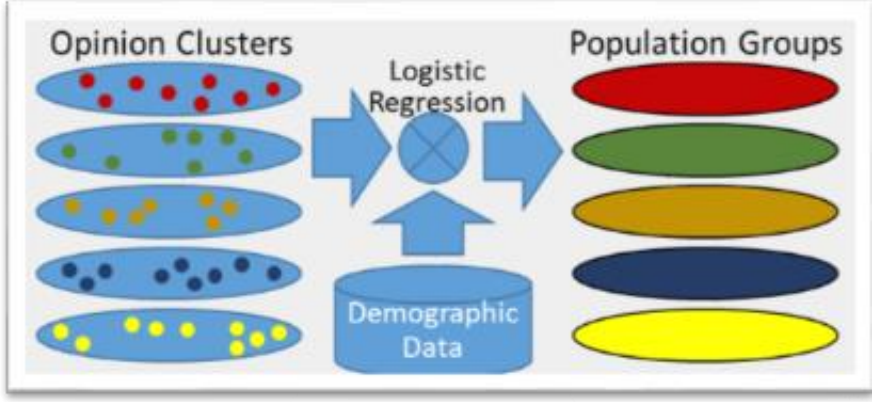
Results of Classification (2)



KEY

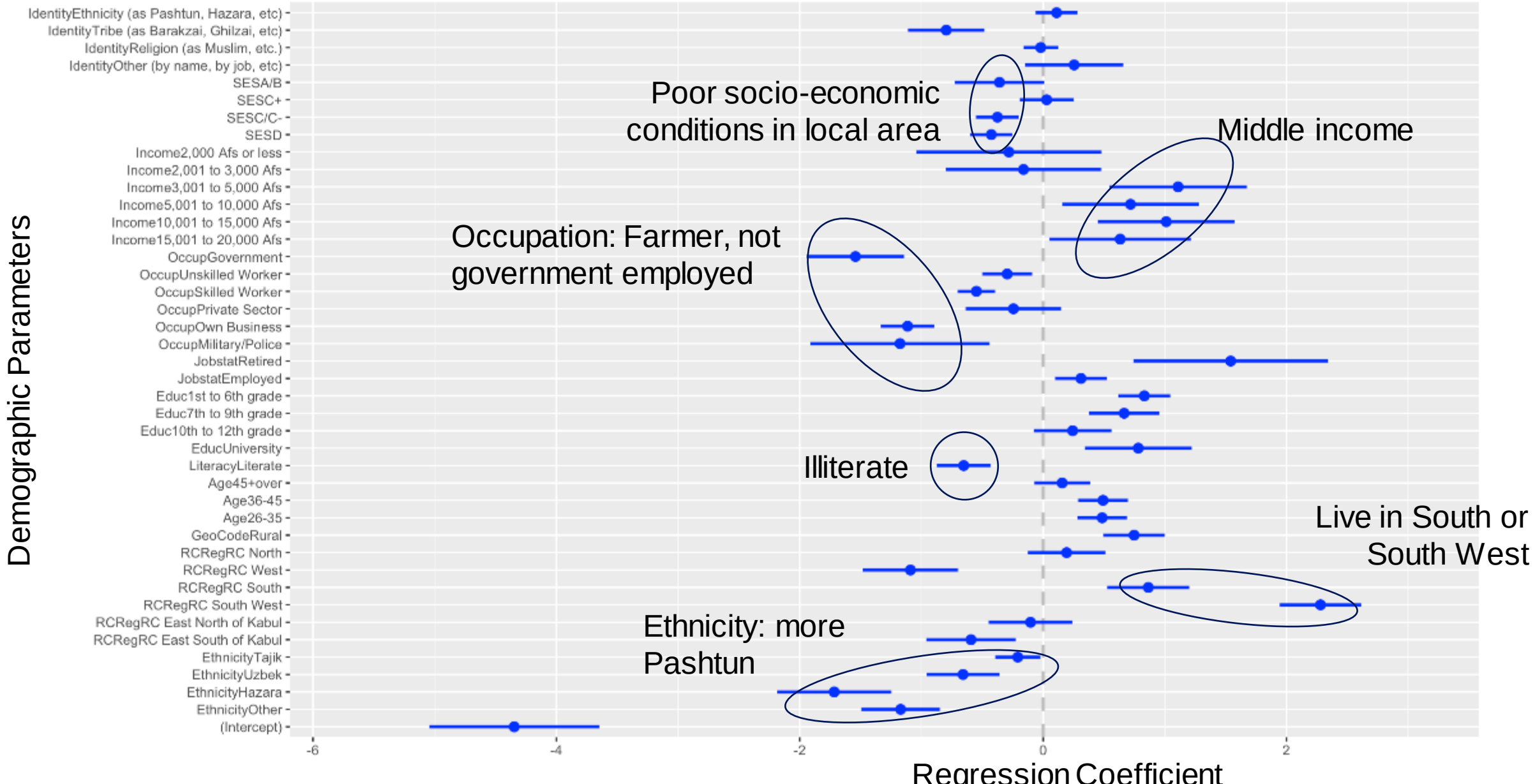
- Governance
- Security
- Economic
- Infrastructure

Characterization



- Demographics of each cluster are compared to the demographics of the whole population
- Generalized Logistic Regression Model (GLM)
  - Cluster membership as dependent variable
  - Demographic data as independent variables
    - age; gender; ethnicity; region; rural or urban residence;
    - literacy; education level; job status; income; social-economic-status;
    - dominant social identity (Afghan, Ethnic, tribe, religion, other)
    - predominant source of news and information
- Result is demographic trends for each cluster

Demographic Characteristics of Cluster C1



8% of sample

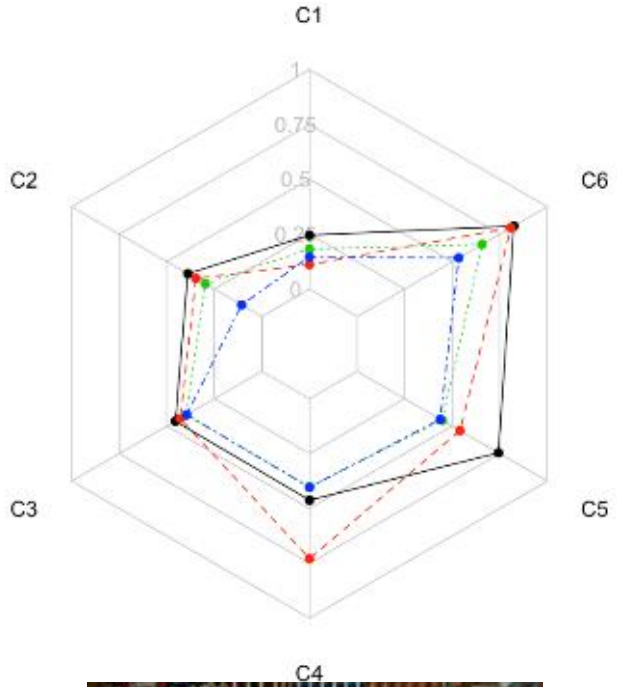
The forgotten minorities

4% of sample

Disgruntled Pashtun farmers

9% of sample

Well-heeled government affiliates



37% of sample

Common Kabul tradespeople

9% of sample

Diverse young starters

33% of sample

Decent-living provincial retirees

So What?

Population profiles can help operators and planners to…

- **Tailor Actions and Messaging** to "target audiences"
  - Security efforts target to Disgruntled Farmers; Infrastructure efforts to Forgotten Minorities
- Disaggregate survey data to examine opinions of each group
  - e.g. **Migration**: "*If you had the resources and opportunity to leave Afghanistan would you leave or would you stay?*"
    - Most likely to want to leave: Young Starters and Forgotten Minorities
    - Least likely to want to leave: Disgruntled Pashtuns and Well-Heeled Government Affiliates
- Choose **appropriate media** to message intended audience
  - e.g. television messages preferentially reach Young Starters; mosque messages tend to reach Pashtun Farmers
- Assess effectiveness of friendly force actions

What Next?

- As a by-product of our process, we have trained a machine learning algorithm to profile ***new*** cases solely based on their demographics (not their opinions)
  - Examine changes over time in a group's opinions
    - Link to operational effectiveness of actions
  - Requires cross-validation (training vs test sets)
- Make techniques and results available to military planners and operators
  - Current methods are exploratory
  - Initial results well received by Resolute Support Info Ops staff
  - Investigate potential mechanisms to get methods & results into the field

Conclusion

- Demonstrated capability to extract structure from complex dataset
- Data-driven, bottom-up approach complements analyst-driven, top-down approach
- Operationally relevant results can support military planners and operators

Questions?