Iceland
Liechtenstein
Norway grants

THESEUS
Connect the Disconnections -
from Disparate Data to Insightful Analysis

Introduction to public profiling

Understanding people

This publication was realised with the EEA Financial Mechanism 2014-2021 financial support. Its content (text, photos, videos) does not reflect the official opinion of the Programme Operator, the National Contact Point and the Financial Mechanism Office. Responsibility for the information and views expressed therein lies entirely with the author(s).

L-Università ta' Malta

SNSPA

NTNU
Kunnskap for en bedre verden

Working together for a green, competitive and inclusive Europe

LEARNING OUTCOMES

- Open data analysis ---> profiling
  - Understanding key concepts (predictive analytics, a predictive behavior, profiling)
  - Developing skills regarding the use of open data
  - Developing analytical skills – understanding the use for different types of data
  - Understanding the mechanism of behavioral prediction
  - Understanding good models of predictive behavior

PSYCHOHISTORY

Individual -> Community -> Large population

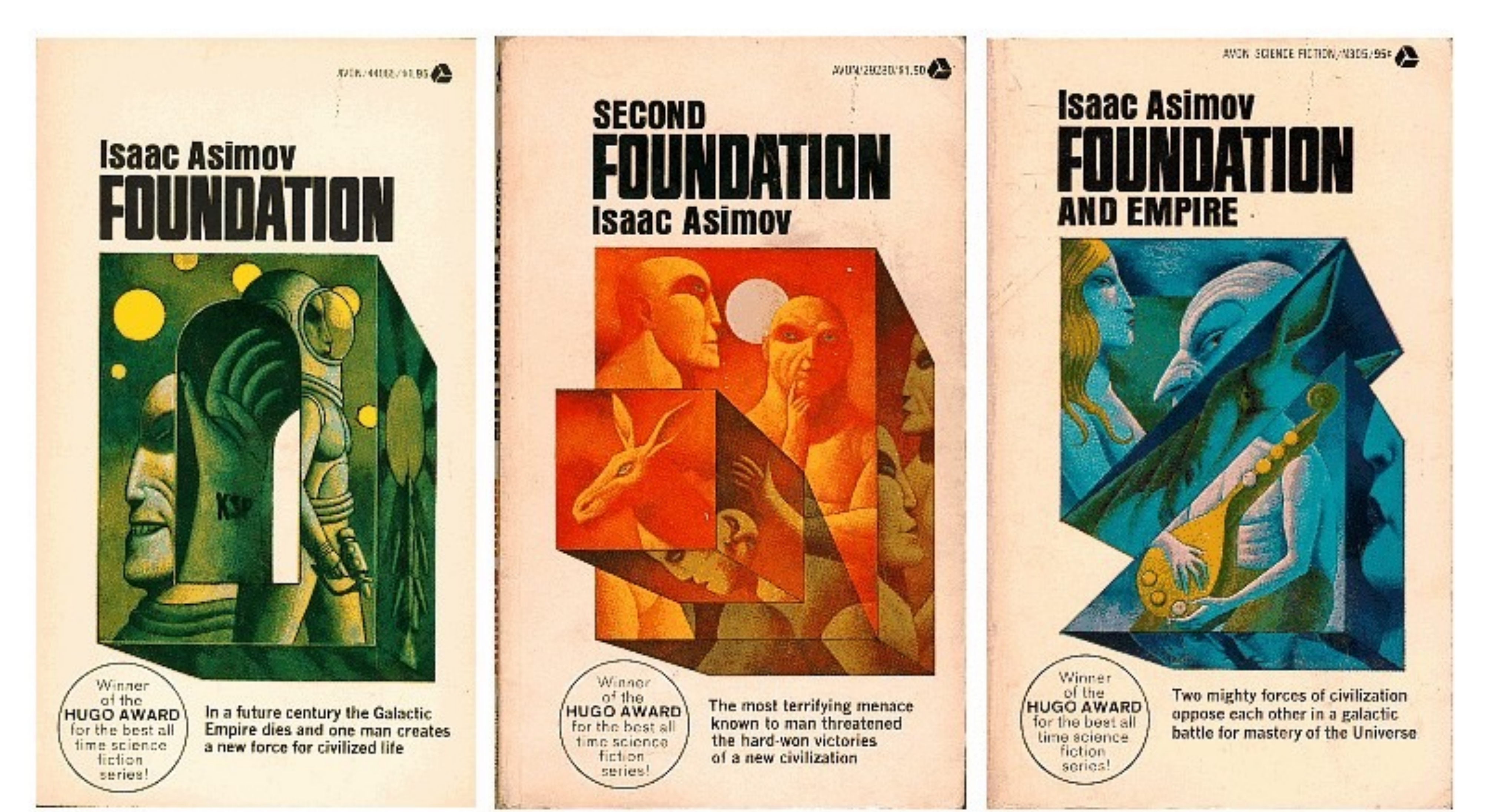

Isaac Asimov
FOUNDATION
Winner of the HUGO AWARD for the best all time science fiction series!
In a future century the Galactic Empire dies and one man creates a new force for civilized life
SECOND
FOUNDATION
Isaac Asimov
Winner of the HUGO AWARD for the best all time science fiction series!
The most terrifying menace known to man threatened the hard-won victories of a new civilization
Isaac Asimov
FOUNDATION
AND EMPIRE
Winner of the HUGO AWARD for the best all time science fiction series!
Two mighty forces of civilization oppose each other in a galactic battle for mastery of the Universe

Photo: cover art for the Isaac Asimov *Foundation* trilogy by Don Ivan Punchatz. Source: https://www.fanboy.com/2009/10/don-ivan-punchatz.html

INTRODUCTION – what can we do with data?

- How can we use data to better understand a person or a group, based on analyzing available data?
- Contemporary research and social sciences (sociology, psychology, social psychology) can offer us more insight into how people think and behave than ever before.
- This course is about how to use data efficiently (select the method / or data most appropriate for your subject) and understand people, based on survey data or other types of available data.

KEY CONCEPTS

What is profiling?

A definition

- **Profiling.** Any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour (including different types of participation), location or movements. Profiling is composed of three elements: (a) it has to be an automated form of processing; (b) it has to be carried out on personal data; and (c) the objective of the profiling must be to evaluate personal aspects about a natural person.

What types of data can predict behavior?

INDIVIDUAL

SEGMENTS

POPULATION

Why is profiling useful?

- **MARKETING** – identifying consumers for better targeting of advertising

- **POLITICAL COMMUNICATION & ELECTORAL MARKETING** – identifying voters for better targeting of messages

- **LAW ENFORCEMENT** – identifying behaviour to detect criminal aspects

Know where we can find data (open or not)

- Surveys
- Open data – social networks, public profiles (ex: Facebook)
- Comments using social network profiles, on public websites or forums
- “Cookies” (from websites that we own)
- Apps (related to own business)

Building a MODEL

BASICS based on an EXAMPLE. Understanding / predicting the voting behavior of the population – electoral participation, based on survey data

Step 1.

Define the main variable.

In our case: electoral participation (intention to vote)

Step 2.

Identify useful and relevant types of behavior as reference (there are always similar behaviors to be compared and interpreted, in order to extract useful data to predict the behavior we are interested in)

WHAT IS THE MOST RELEVANT DATA TO BE USED TO PREDICT VOTING BEHAVIOR?

Identify **primary behaviors** (past behavior involving the main variable - voting in the past)

Secondary behavior

Tertiary behavior

Socio-demographic data

Associated data (of past behavior) - from census, official government statistics or other type of data

Attitudes and orientations (values, feelings, opinions etc.)

Networks (personal, values...) or **Clusters** - relevant to identify a propensity to vote

Step 3.

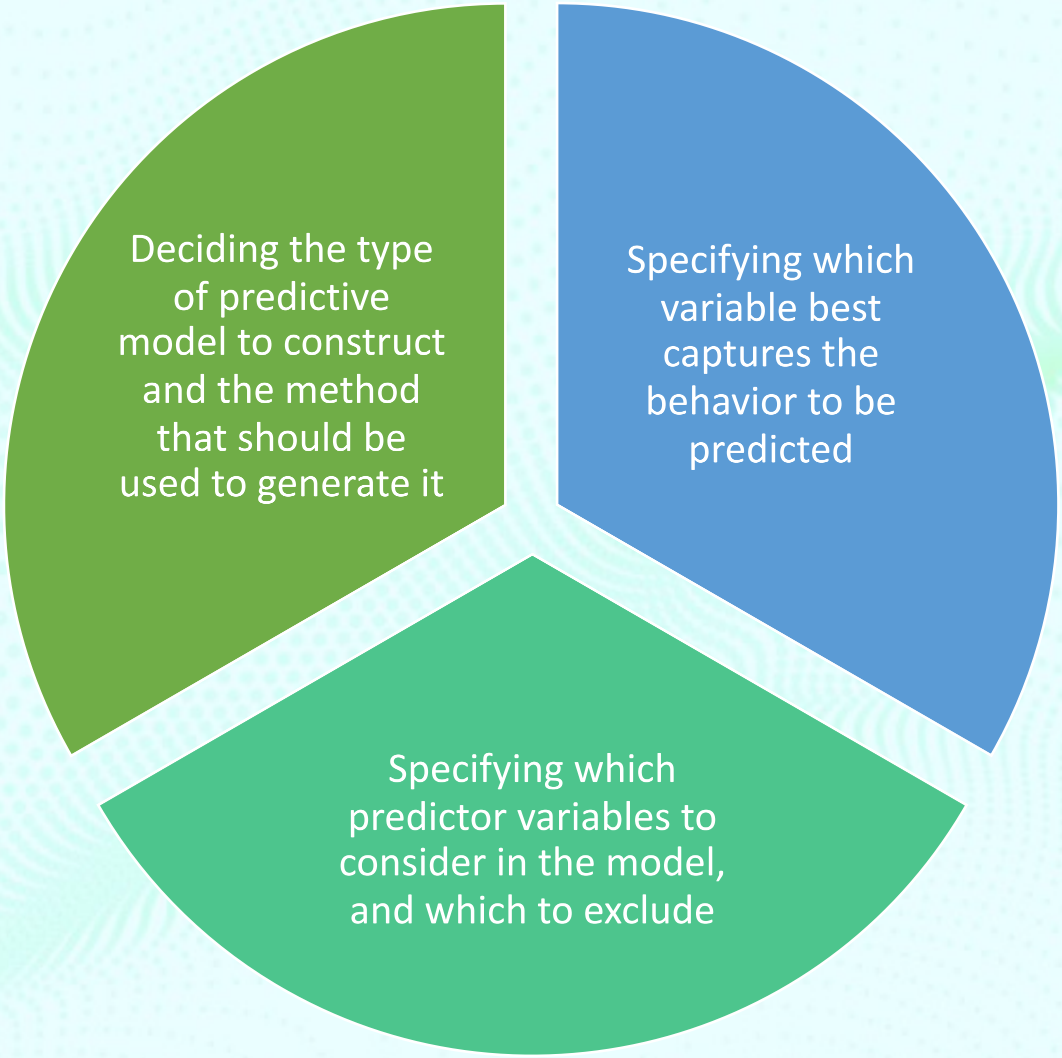
Run some data analysis on our datasets, in order to decide what works better for our objective.

- simple bivariate analysis (cross tabulations, correlations), that can create linear models between two variables.
- more complex analysis like regressions (multivariate analysis) using more relevant dependent variables to understand the relationship with your main variable.
- cluster analysis (or other type of segmentation of the populations based on complex variation of the variables: decision trees, factorial analysis, ML algorithms etc.).

Step 4.

Building the model.

Based on data analysis, we can identify the most relevant variables in a survey (or a dataset) to be taken into account when constructing a model of prediction.

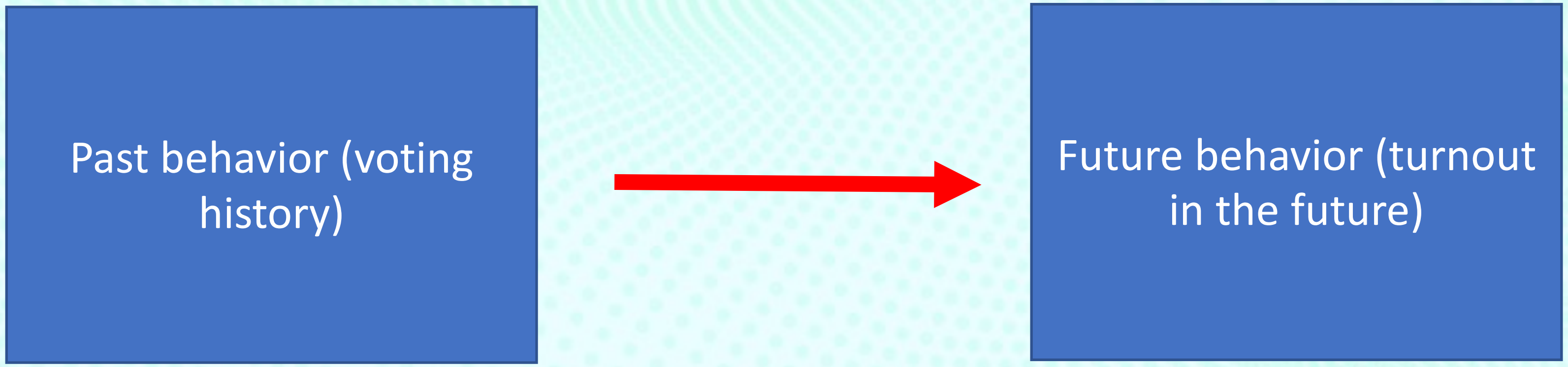


Deciding the type of predictive model to construct and the method that should be used to generate it

Specifying which variable best captures the behavior to be predicted

Specifying which predictor variables to consider in the model, and which to exclude

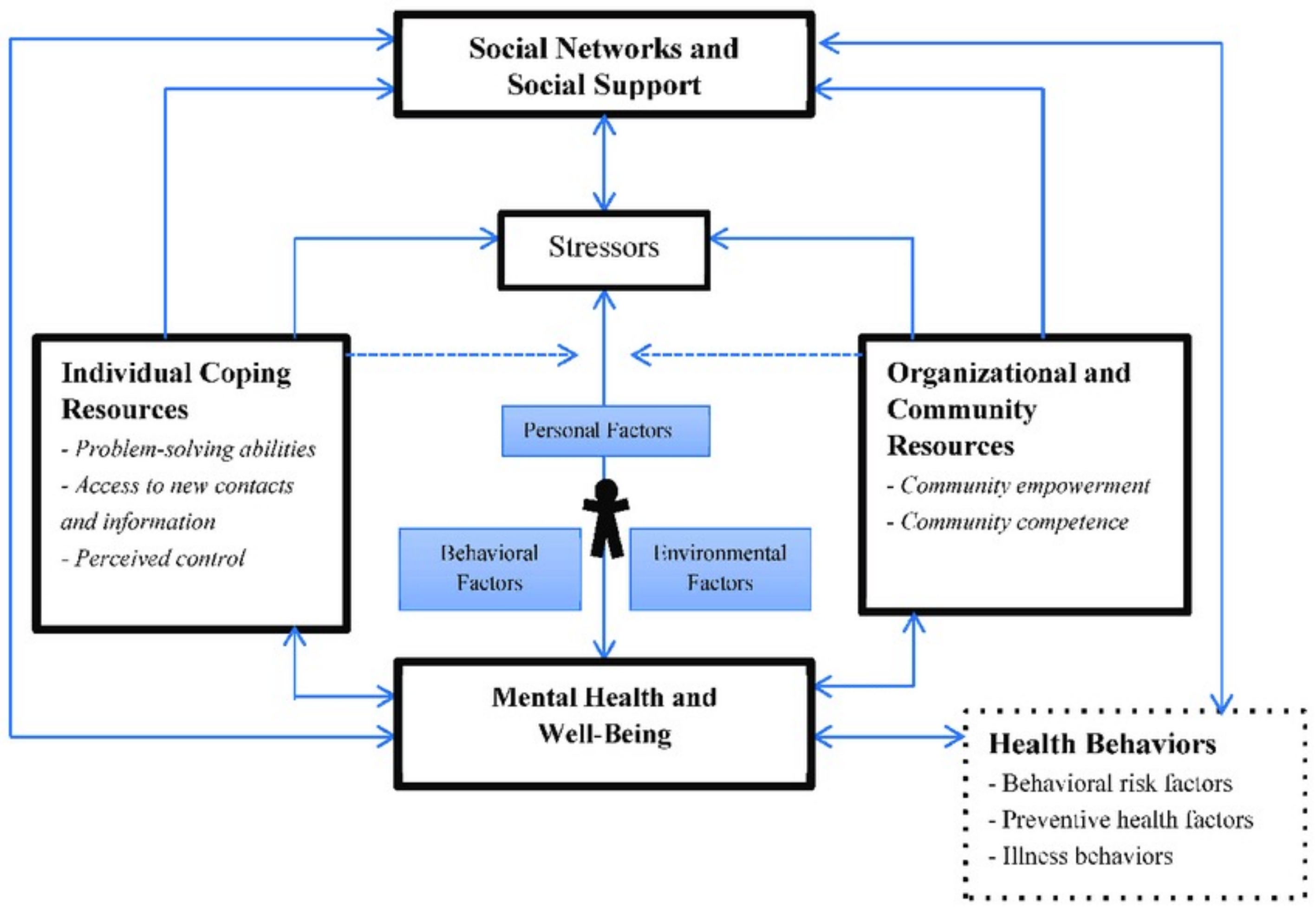
Example of a simple model



Past behavior (voting history)

Future behavior (turnout in the future)

Example of a more complex model

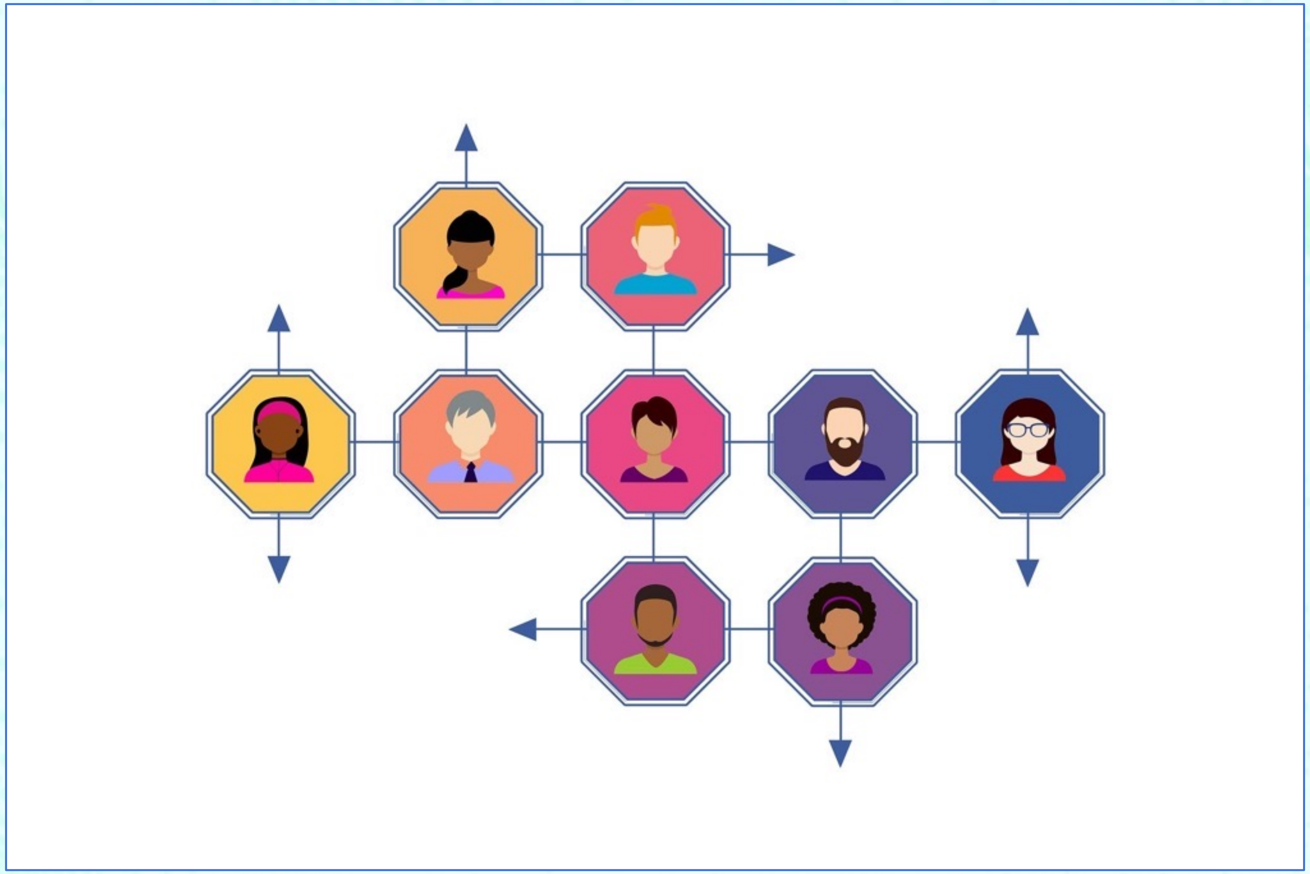

Social Networks and Social Support
Stressors
Individual Coping Resources
- Problem-solving abilities
- Access to new contacts and information
- Perceived control
Personal Factors
Behavioral Factors
Environmental Factors
Organizational and Community Resources
- Community empowerment
- Community competence
Mental Health and Well-Being
Health Behaviors
- Behavioral risk factors
- Preventive health factors
- Illness behaviors

Source: https://www.researchgate.net/publication/332429587_Improving_the_Living_Learning_and_Thriving_of_Young_Black_Men_A_Conceptual_Framework_for_Reflection_and_Projection/

Step 5.

Comparing individuals with samples or populations.

Afterwards we can make inferences about individual or groups (depending on the score of the probability we decide is relevant for the analysis)



CASE STUDY. Predicting electoral participation in Romanian presidential elections

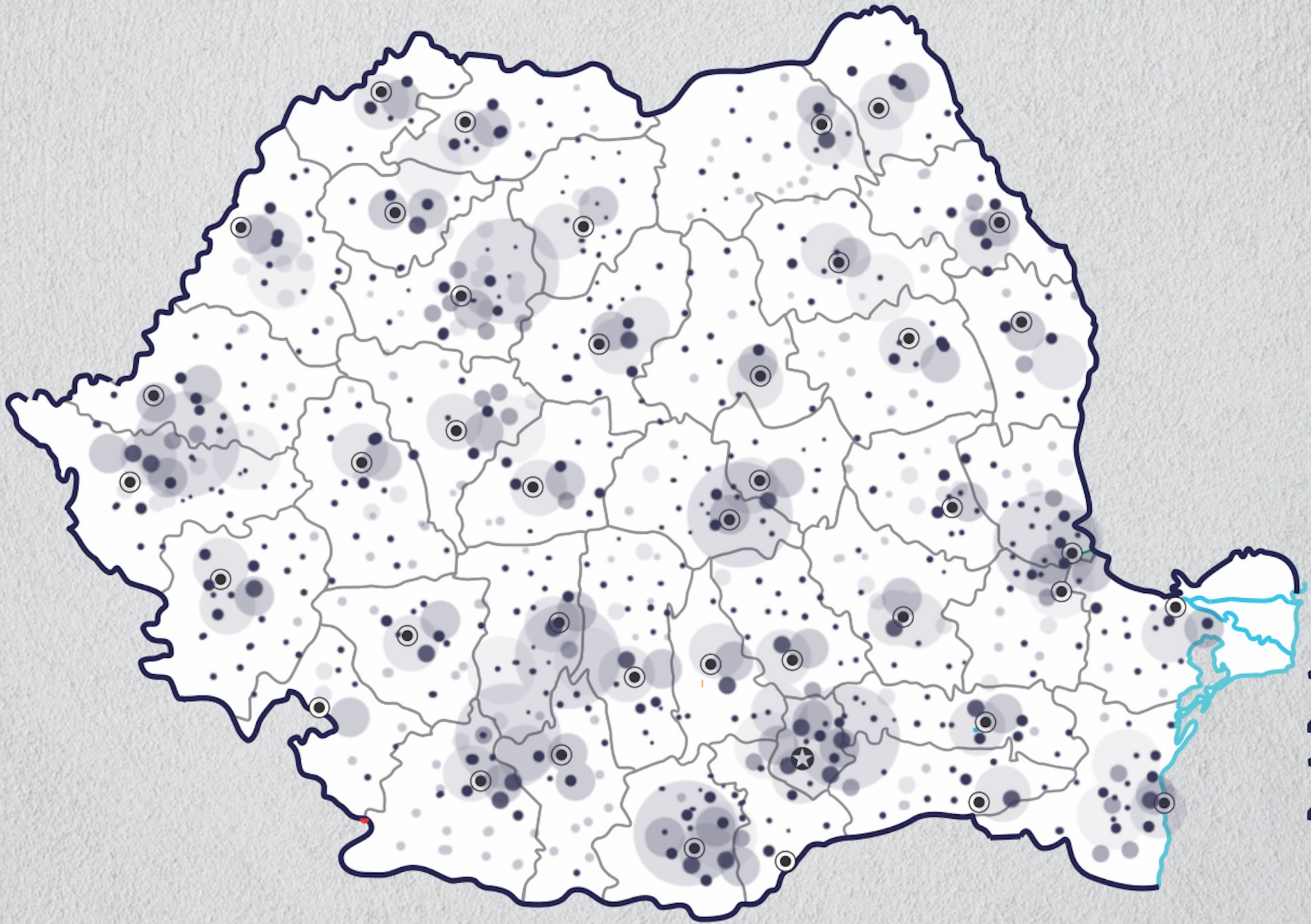
Step-by-step description of building the model

Challenges & best practices

Lessons learnt

Survey dataset used: CPD-SNSPA (2019) - survey about Romanian Presidential Elections; data collected: 1-22 June 2019 (N=977, ±3%). The analysis also uses statistics from previous elections, from the national electoral bureau in Romania

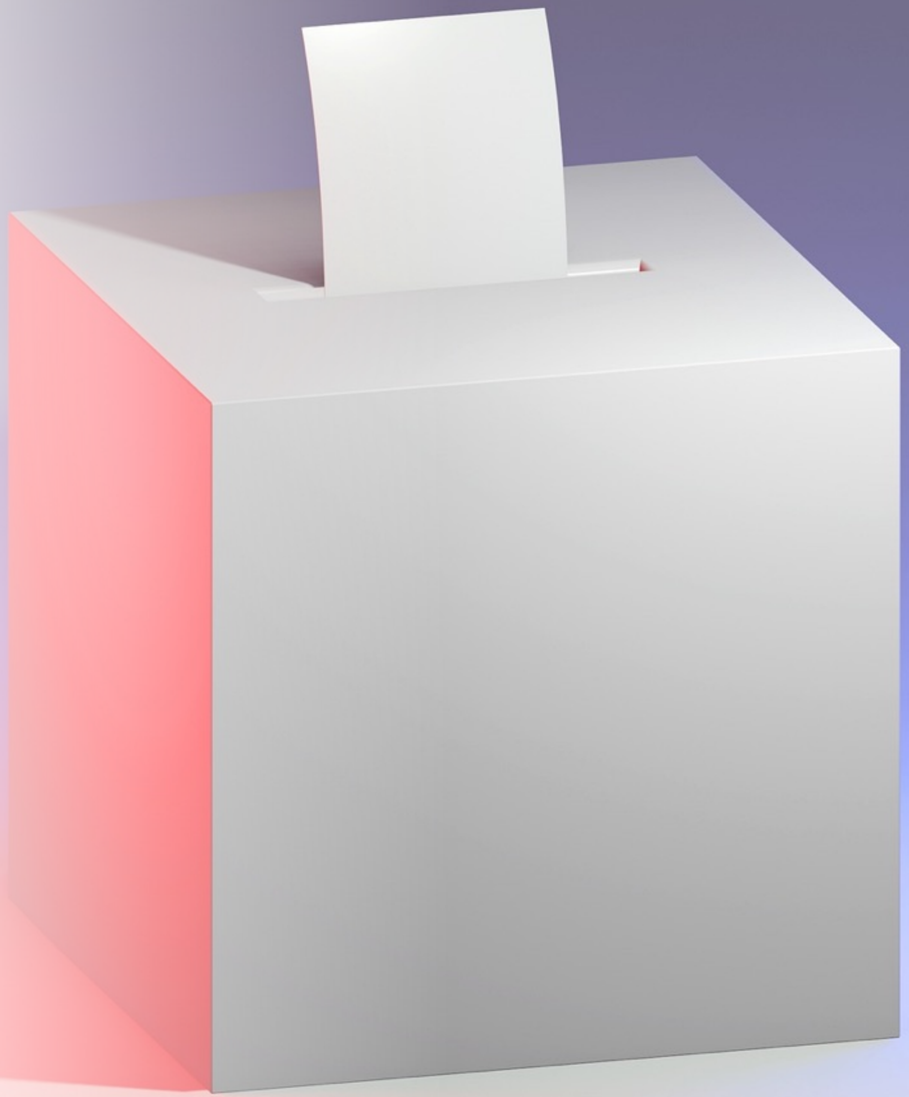
Voter turnout – areas with the highest level of electoral participation



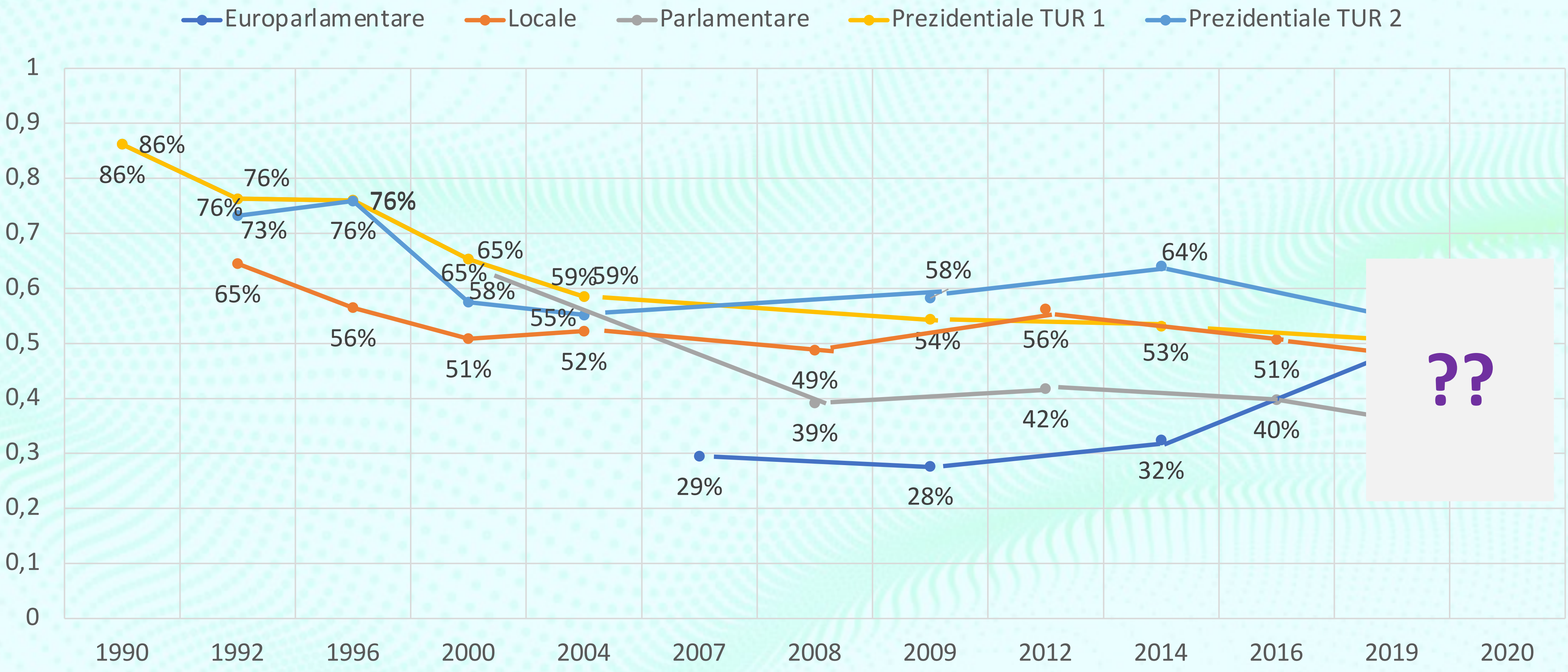
2019
2020
Romania

STEP 1. Define the variable

- The main variable: Electoral participation
- 83% of the respondents in our sample declare their intent to vote in the next elections

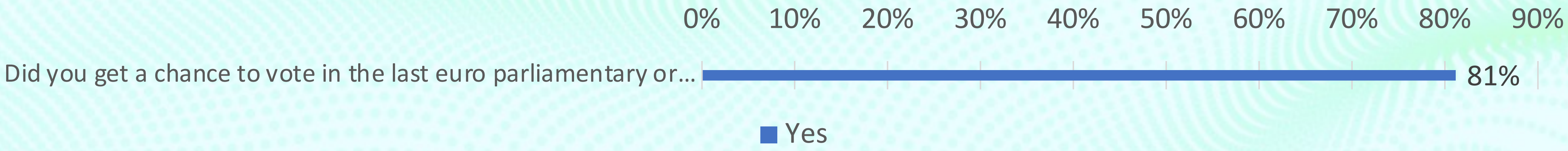


Election participation, in Romania, before 2019



STEP 2. Identify reference behavior

- **First – evaluated past behavior.**



- **Examples of past participation** (statistics at population level):
  - 26 May 2019 – European Parliament: 9069822 participants out of 18267732 potential voters (49.65%)
  - 11 December 2016 – National Parliament: 7323368 participants from 18403044 potential voters (39.44%)
  - 5 June 2016 – Local Elections: 8893687 participants from 18462528 potential voters (48.17%)
  - 2 November 2014 – Presidential elections:9723232 participants from 18284066 potential voters (53.17%)

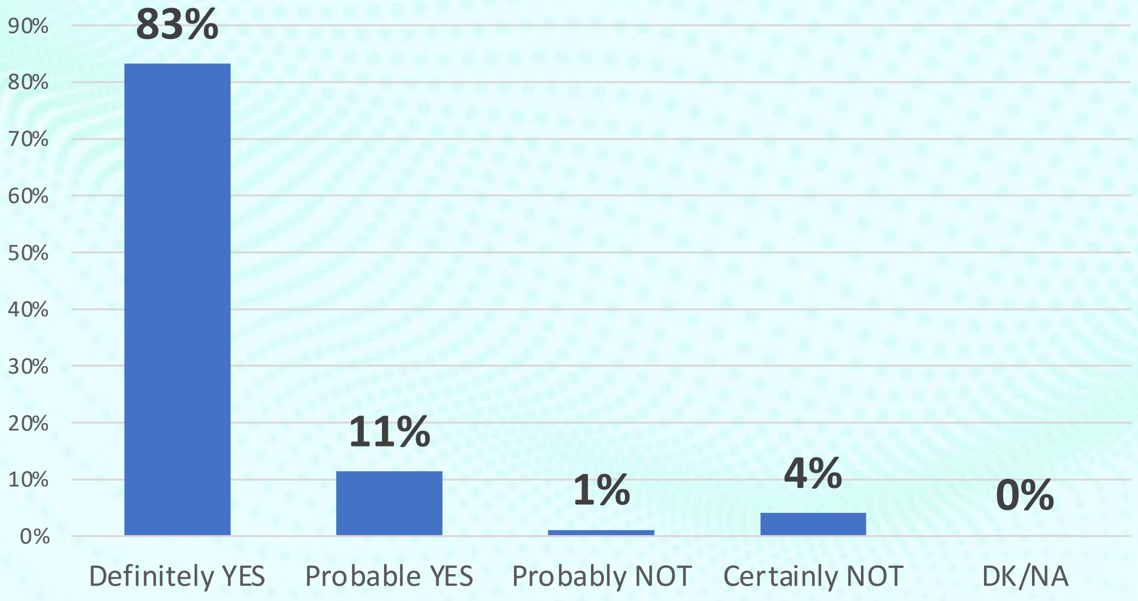
What we know so far

- In our sample we have around 83% declared intention of participation.
- We know from previous elections that participation varied between 39% and 53%.
- But we also have 81% percent in this sample saying they participated at past elections. So, we cannot assume that this indicator (declared intention of participation) is reliable.
  - Because of social pressure, people tend to give sometimes responses that are expected from them (more people will say they voted compared to actual participation figures).
- What can we do to narrow down the number of people that actually intend to vote?
- We use **more variables** related to our main variable.
- We build an alternate predictor of voting, using a combination of two indicators: **intention of participating and interest in elections**.

STEP 2. Identify reference behavior. (CONTINUED)

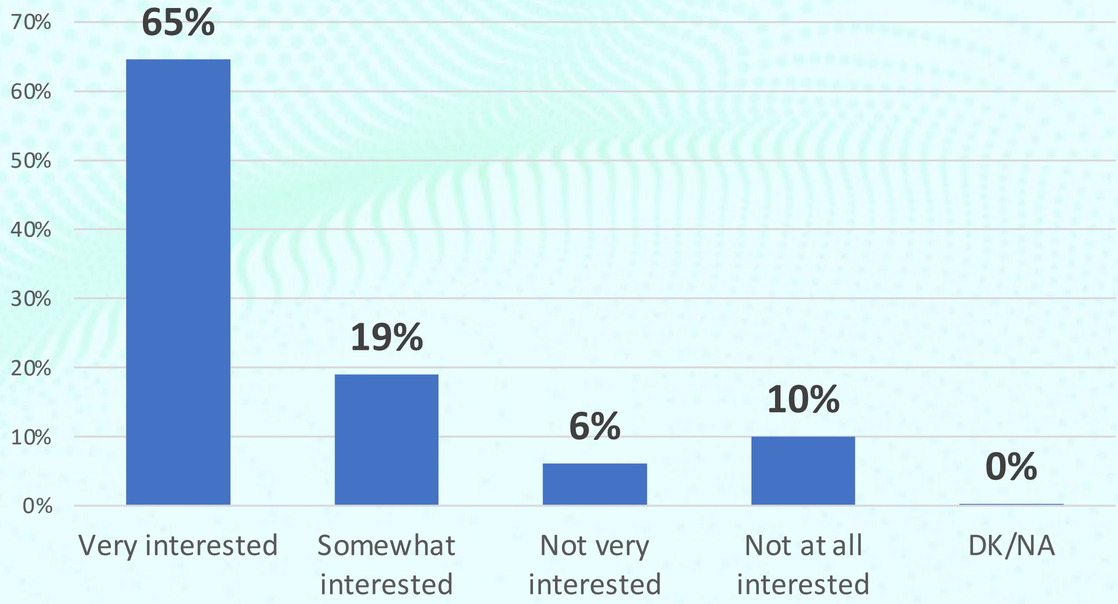
Intention of participating

If PRESIDENTIAL elections were being held next Sunday, would you go to the polls or not?



Interest in elections

If PRESIDENTIAL elections were being held next Sunday (for the election of the President of Romania), how interested would you be in these elections?

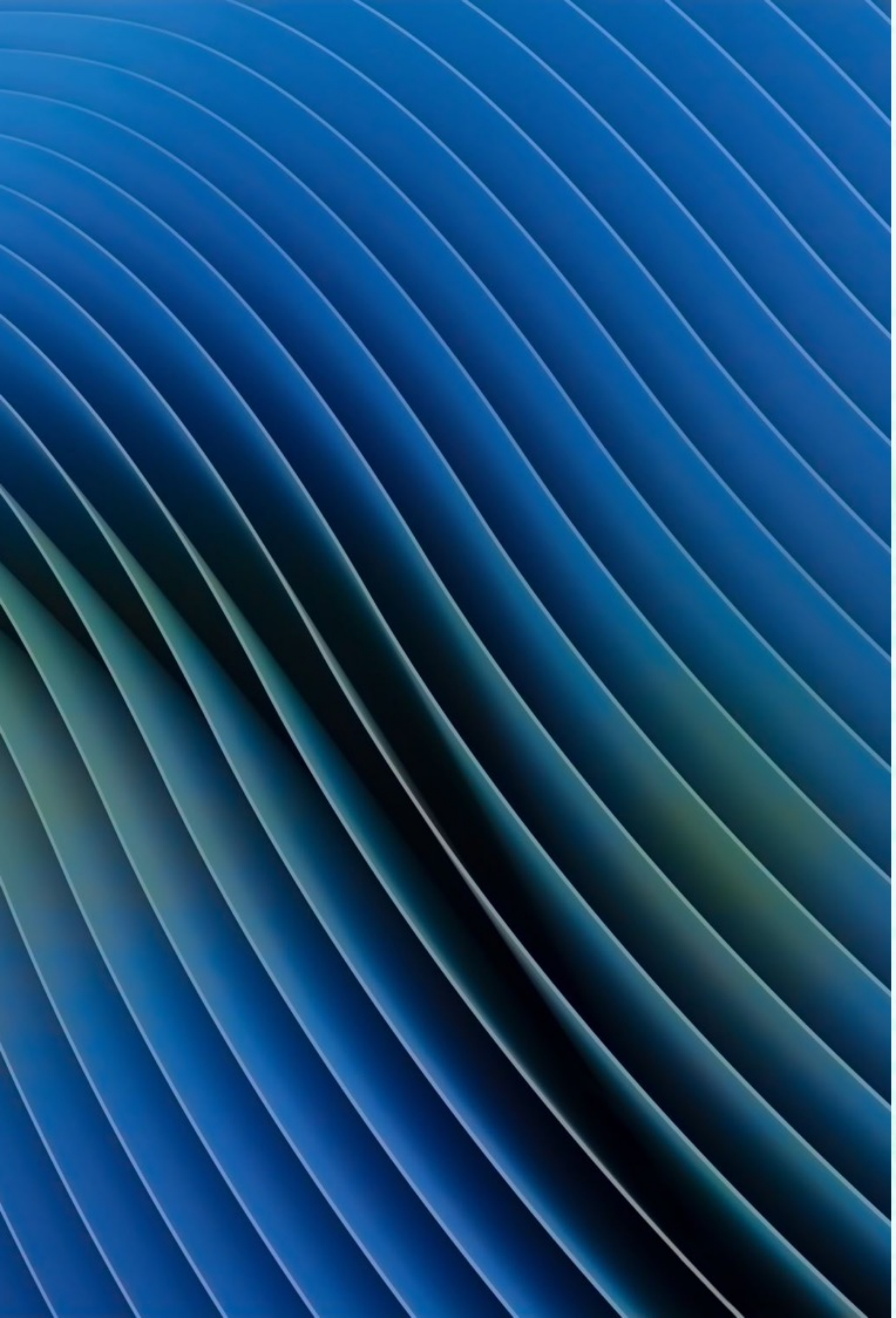


STEPS 3 & 4.
Data analysis & building the model

- A simple cross tabulation between the two variables – a statistical tool for categorical data, combining responses for both variables to analyze subsamples. The results will be a segmentation of the sample

		If PRESIDENTIAL elections were being held next Sunday, would you go to the polls or not?					Total
		Definitely YES	Probable YES	Probably NOT	Certainly NOT	DK/NA	
If PRESIDENTIAL elections were being held next Sunday (for the election of the President of Romania), how interested would you be in these elections?	Very interested	62%*	2%	0%	0%		65%
	Somewhat interested	17%	2%				19%
	Not interested	2%	3%	0%	1%	0%	6%
	Not at all interested	3%	3%	1%	3%		10%
	DK/NA	0%	0%				0%
Total		83%	11%	1%	4%	0%	100%

*Proportion of people that are highly interested in elections and say they will definitely vote.



To sum up our work so far

First, we have narrowed down from a subsample of 81% of people that said they voted in the past, to a lower sample of 62%, which is a more plausible participation rate.

This is better, but this figure is still much bigger compared to the real-life general participation behaviors (which are, usually, below 50%).

Can we do even better?

STEPS 3 & 4.
Data analysis & building the model (CONTINUED)

- We will create a cross tabulation report to see the relation between two variables, and specifically to look at how our sample of 62% (the people who seem to be very mobilized) responded regarding their voting behavior in past elections

		Did you get a chance to vote in the last euro parliamentary or presidential elections?		
		No	Yes	Total
Mobilization for presidential election	Very mobilized	7%	55%	62%
	Slightly mobilized	6%	16%	21%
	Not mobilized	6%	10%	17%
Total		19%	81%	100%

Adjusting data to reality

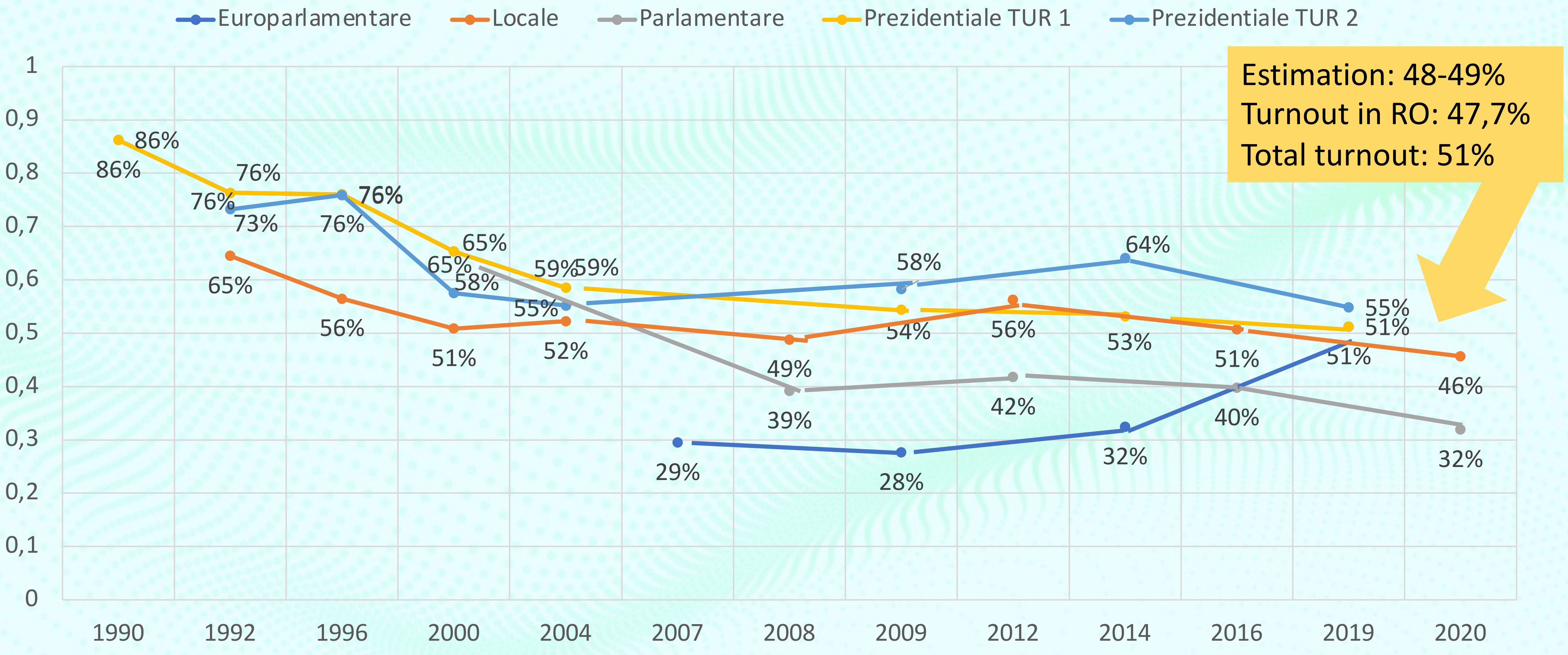
- **We must also adapt the survey data to the reality of the population!**
- People currently living in Romania only represent 85% of the adult population (thus, our sample is not representative for all 18.3 million Romanian adults, but for 85%*18.3= 15.6 million adults actually living in the country, while around 2.7 million people live abroad - the Romanian “diaspora”).
  - if the 55% turnout is representative for the “in country” population, then, when applied to **“the whole voting population”, this percentage will lower to 47%**.

Potential electoral participation abroad: 1-2% bonus to the overall turnout

Final result

- To finalize our estimation, we can predict that the turnout in these elections would be between 48% and 49% (after adding the estimate participation for voters abroad).

Election participation, in Romania, including 2019



Estimation: 48-49%
Turnout in RO: 47,7%
Total turnout: 51%

Reality check! Did the estimation work?

- Our estimation of turnout was based on data measured in June 2019, a few months before the actual election day (10 November 2019), and just after another election (for European Parliament - 26 May 2019).
- The real turnout in Presidential elections, in November 2019, was 47.66% in Romania and 3,67% - abroad (among the Diaspora). Total participation = **51,3%**.
- The margin of error for our estimation has proven to be very small for the “domestic” population (only 0,66%), and slightly larger for the turnout in diaspora (around 2%).

STEP 5. Compare results with population

- Next step: understanding more characteristics of people with higher probability to vote
- Using: official statistics of voters from the official voting bureau and information in the sample.

		TOTAL	Probability of voting						
		Sample	High	Medium	Low			BEC statistics	Difference BEC - SAMPLE
		Col N %	Col N %	Col N %	Col N %				
TOTAL	Sample	100%	100%	100%	100%				
GENDER	Male	48%	48%	45%	53%		Male	48%	0%
	Female	52%	52%	55%	47%		Female	52%	0%
	18-29 y.o.	20%	18%	22%	22%		18-29 y.o.	14%	-4%
	30-44 y.o.	26%	26%	26%	30%		30-44 y.o.	26%	1%
	45-59 y.o.	25%	26%	23%	23%		45-59 y.o.	28%	2%
	>60 y.o.	29%	31%	29%	25%		60 y.o. and over	32%	1%
UrbanRural	Urban	56%	57%	54%	56%		Urban	58%	1%
	Rural	44%	43%	46%	44%		Rural	42%	-1%

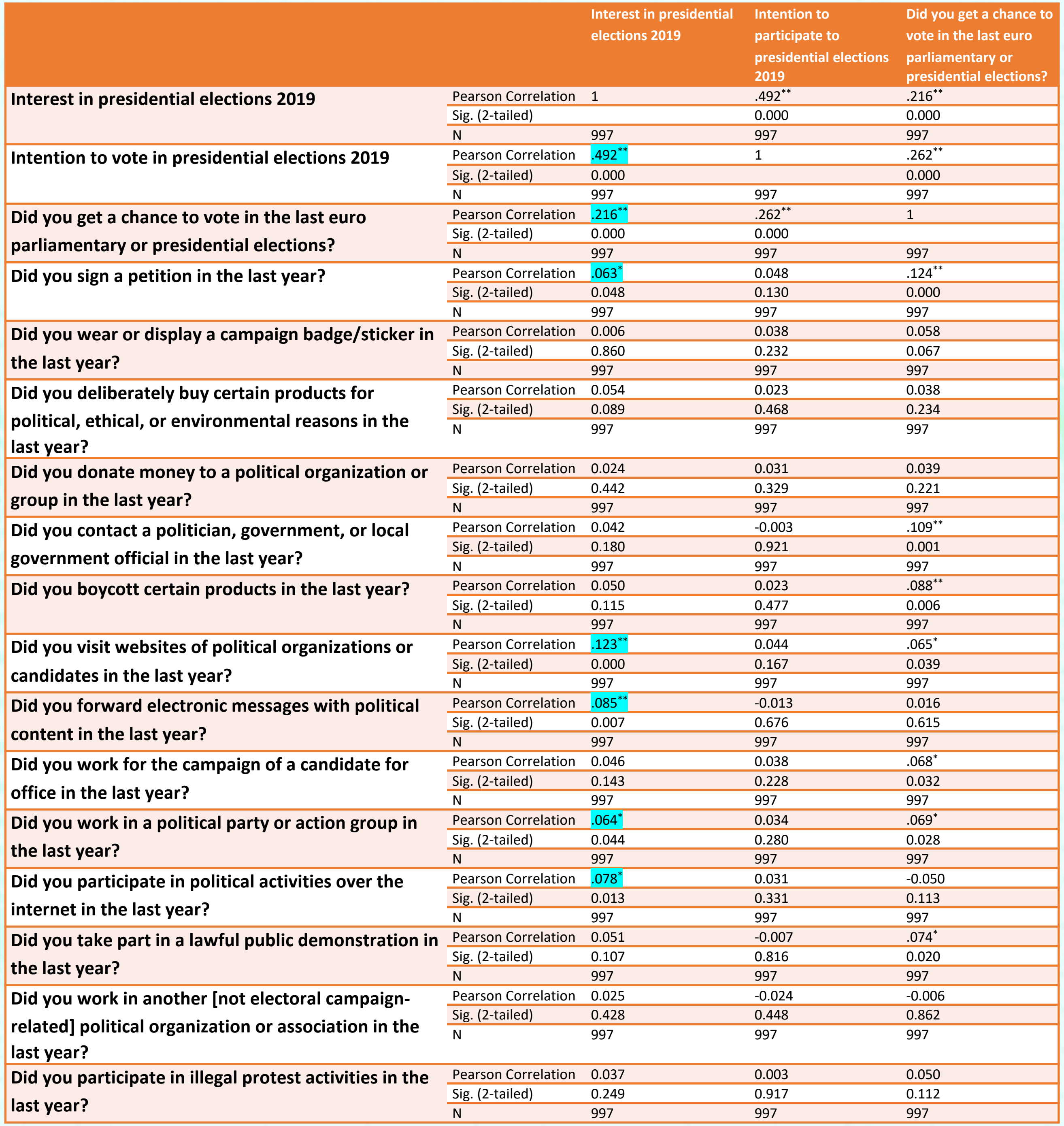
NEW STEPS

- Once we know that our model is working, we can explore further data analysis methods to understand our segment of the population...

What if we want to learn more about the population of “likely voters”?

15 PARTICIPATIVE ACTIONS	SAMPLE
Did you get a chance to vote in the last euro parliamentary or presidential elections?	81%
Did you sign a petition in the last year?	18%
Did you wear or display a campaign badge/sticker in the last year?	15%
Did you deliberately buy certain products for political, ethical, or environmental reasons in the last year?	9%
Did you donate money to a political organization or group in the last year?	8%
Did you contact a politician, government, or local government official in the last year?	7%
Did you boycott certain products in the last year?	5%
Did you visit websites of political organizations or candidates in the last year?	5%
Did you forward electronic messages with political content in the last year?	5%
Did you work for the campaign of a candidate for office in the last year?	4%
Did you work in a political party or action group in the last year?	3%
Did you participate in political activities over the internet in the last year?	2%
Did you take part in a lawful public demonstration in the last year?	2%
Did you work in another [not electoral campaign-related] political organization or association in the last year?	2%
Did you participate in illegal protest activities in the last year?	1%

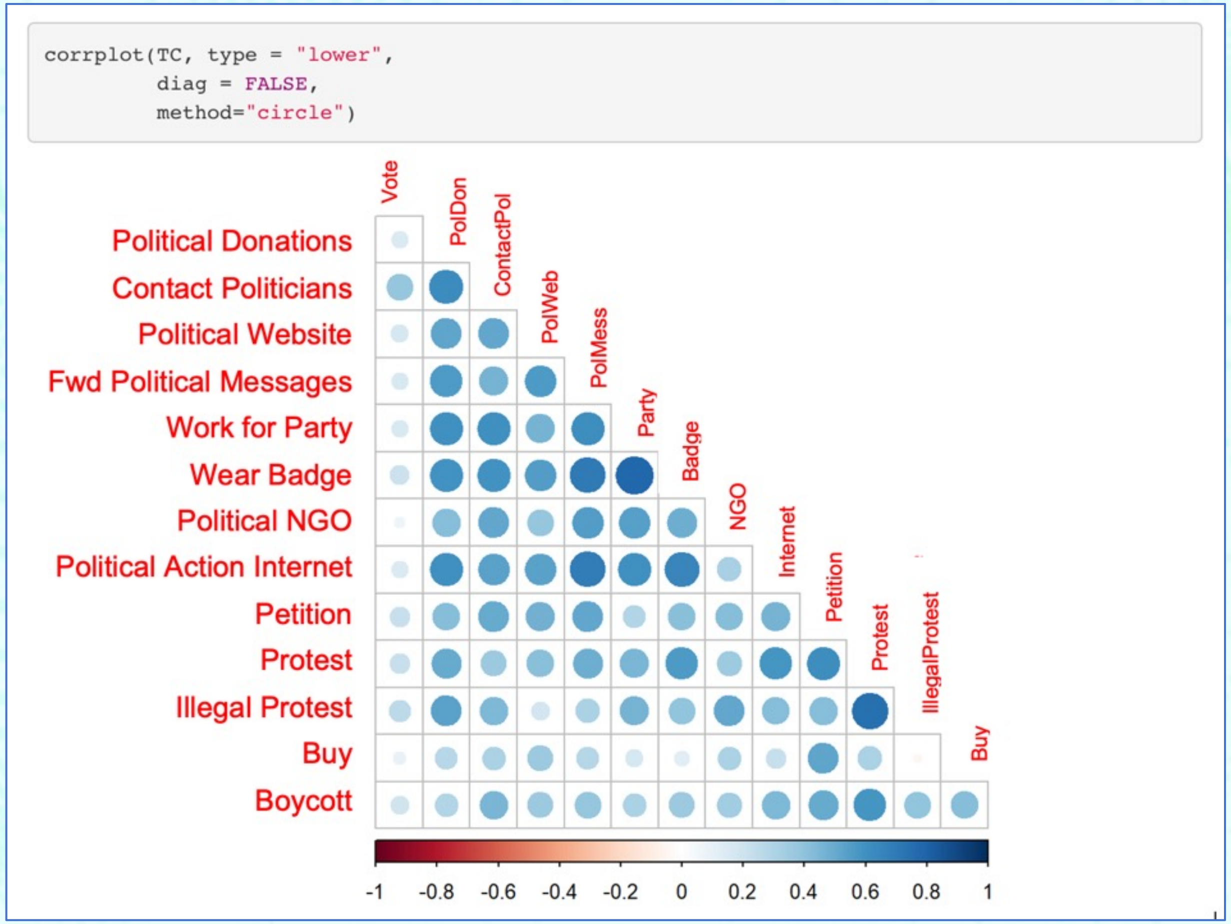
CORRELATIONS



		Interest in presidential elections 2019	Intention to participate to presidential elections 2019	Did you get a chance to vote in the last euro parliamentary or presidential elections?
Interest in presidential elections 2019	Pearson Correlation	1	.492**	.216**
	Sig. (2-tailed)		0.000	0.000
	N	997	997	997
Intention to vote in presidential elections 2019	Pearson Correlation	.492**	1	.262**
	Sig. (2-tailed)	0.000		0.000
	N	997	997	997
Did you get a chance to vote in the last euro parliamentary or presidential elections?	Pearson Correlation	.216**	.262**	1
	Sig. (2-tailed)	0.000	0.000	
	N	997	997	997
Did you sign a petition in the last year?	Pearson Correlation	.063*	0.048	.124**
	Sig. (2-tailed)	0.048	0.130	0.000
	N	997	997	997
Did you wear or display a campaign badge/sticker in the last year?	Pearson Correlation	0.006	0.038	0.058
	Sig. (2-tailed)	0.860	0.232	0.067
	N	997	997	997
Did you deliberately buy certain products for political, ethical, or environmental reasons in the last year?	Pearson Correlation	0.054	0.023	0.038
	Sig. (2-tailed)	0.089	0.468	0.234
	N	997	997	997
Did you donate money to a political organization or group in the last year?	Pearson Correlation	0.024	0.031	0.039
	Sig. (2-tailed)	0.442	0.329	0.221
	N	997	997	997
Did you contact a politician, government, or local government official in the last year?	Pearson Correlation	0.042	-0.003	.109**
	Sig. (2-tailed)	0.180	0.921	0.001
	N	997	997	997
Did you boycott certain products in the last year?	Pearson Correlation	0.050	0.023	.088**
	Sig. (2-tailed)	0.115	0.477	0.006
	N	997	997	997
Did you visit websites of political organizations or candidates in the last year?	Pearson Correlation	.123**	0.044	.065*
	Sig. (2-tailed)	0.000	0.167	0.039
	N	997	997	997
Did you forward electronic messages with political content in the last year?	Pearson Correlation	.085**	-0.013	0.016
	Sig. (2-tailed)	0.007	0.676	0.615
	N	997	997	997
Did you work for the campaign of a candidate for office in the last year?	Pearson Correlation	0.046	0.038	.068*
	Sig. (2-tailed)	0.143	0.228	0.032
	N	997	997	997
Did you work in a political party or action group in the last year?	Pearson Correlation	.064*	0.034	.069*
	Sig. (2-tailed)	0.044	0.280	0.028
	N	997	997	997
Did you participate in political activities over the internet in the last year?	Pearson Correlation	.078*	0.031	-0.050
	Sig. (2-tailed)	0.013	0.331	0.113
	N	997	997	997
Did you take part in a lawful public demonstration in the last year?	Pearson Correlation	0.051	-0.007	.074*
	Sig. (2-tailed)	0.107	0.816	0.020
	N	997	997	997
Did you work in another [not electoral campaign-related] political organization or association in the last year?	Pearson Correlation	0.025	-0.024	-0.006
	Sig. (2-tailed)	0.428	0.448	0.862
	N	997	997	997
Did you participate in illegal protest activities in the last year?	Pearson Correlation	0.037	0.003	0.050
	Sig. (2-tailed)	0.249	0.917	0.112
	N	997	997	997

Correlation
Source: Sultanescu (2020)

CORRELATIONS



```
corrplot(TC, type = "lower",
         diag = FALSE,
         method="circle")
```

Visualization - Polychoric correlation, binary relation in pairs of two variables. The stronger the correlation, the darker the color.

Source: Sultanescu (2020)

CLUSTER ANALYSIS

	2-cluster	3-cluster	4-cluster	5-cluster
Predicted class membership	0,83	0,76	0,75	0,019
AIC	5584	5437	5424	5426
BIC (Bayesian Information Criterion)	5716	5638	5692	5763
L2 (likelihood ratio)	790	615	574	548
LL (maximum log-likelihood)	-2765	-2677	-2657	-2644

LCA model – statistics for types of participators
Source: Sultanescu (2020)

cluster	label	Population weight
1	The Passives	70,2
2	The “Mainstream” participators	20,0
3	The “Protesters”	7,9
4	The „All-around” participators	1,8

4 cluster option. Their weight (frequencies in the sample for each cluster)
Source: Sultanescu (2020)

Factor analysis

Rotated Component Matrix[a]				
	Group			
	1	2	3	4
Did you work for the campaign of a candidate for office in the last year?	,777			
Did you work in a political party or action group in the last year?	,759			
Did you wear or displayed a campaign badge/sticker in the last year?	,656			
Did you contact a politician, government, or local government official in the last year?	,531	,266	,101	,170
Did you forward electronic messages with political content in the last year?	,483	,378		-,378
Did you participate in political activities over the internet in the last year?	,437	,210	,243	-,297
Did you donate money to a political organization or group in the last year?	,416		,287	
Did you work in another [not electoral campaign-related] political organization or association in the last year?	,340	,306		-,294
Did you deliberately buy certain products for political, ethical, or environmental reasons in the last year?	-,117	,693		
Did you sign a petition in the last year?	,142	,635	,218	,111
Did you boycott certain products in the last year?		,448	,427	
Did you visit websites of political organizations or candidates in the last year?	,391	,428		
Did you participate in illegal protest activities in the last year?	,148	-,118	,775	
Did you take part in a lawful public demonstration in the last year?	,115	,242	,676	
Did you get a chance to vote in the last euro parliamentary or presidential elections?	,144	,174		,843

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

Factor analysis - distribution of answers, in four factors, data measured in 2019
Source: Sultanescu (2020)

Comparing the results: cluster analysis vs factor analysis

Types of participators	LCA	Factor analysis
“All around” participators	2%	2-17%
“Protesters”	8%	9%
“Mainstream” participators	20%	20-30%
“Passives”	70%	61%

Comparison of the shares of participator types, resulting from each type of statistical analysis
Source: Sultanescu (2020)

Comparing data with other countries

Participators	LCA Romania	LCA US
“All around”	2%	6% (triple compared to Romania)
“Protesters”	8%	10% (comparable)
“Mainstream”	20%	24% (slightly more compared to Romania)
Passives	70%	60% (less compared to Romania)

Percentages for types of participators, Romania vs the US
Source: Sultanescu (2020). Data source for US: Oser (2017)

Exercises & practice

Practice

- Use **Google Trends** to compare information about the willingness to go to vote in different regions in Romania, before an election
- Use **Facebook** data (if available) to construct a model to predict if a person has a family with at least a child
- Evaluate the data from a **news feed** to select only the relevant elements (select the variables relevant to identify only the outliers)
- Use **Facebook Insights** to identify the most important news pages for a segment of the population (from a county, or from a city).
- Use **Google Trends** (or Google Ads) to understand data associations in search
- Compare data from an individual (using open data from Facebook) with his peers (people of same age, or same level of education), using survey data

Exercise session

- Use this database: https://docs.google.com/spreadsheets/d/1xMHIe20XvZOdocKHR4eft01c1oBqXfRSVEOcxKItuS0/edit#gid=230559637

- We will use the dataset to identify the characteristics of the public that is willing to go to vote, but it does not do any other participatory activities.
- We will open the data base and we will try to do some work on it. At the end, we will try to identify some new characteristics of the likely voters

Data sources

Survey used for the case study:

- CPD-SNSPA (2019). *Sondaj – participare civica, 2019.* http://civicparticipation.ro/participation/sondaj-participare-civica-2019/

[dataset available by request]

Other data sources:

- https://ro.wikipedia.org/wiki/Alegeri_pentru_Parlamentul_European_%C3%AEn_Rom%C3%A2nia,_2019
- http://europarlamentare2019.bec.ro/
- http://parlamentare2016.bec.ro/wp-content/uploads/2016/12/3_RF.pdf
- https://ro.wikipedia.org/wiki/Alegeri_parlamentare_%C3%AEn_Rom%C3%A2nia,_2016
- http://2016bec.ro/wp-content/uploads/2016/06/BEC_PVFinal.pdf
- https://ro.wikipedia.org/wiki/Alegeri_preziden%C8%9Biale_%C3%AEn_Rom%C3%A2nia,_2014
- http://bec2014.roaep.ro/
- https://prezenta.bec.ro/prezidentiale10112019/romania-stats
- https://insse.ro/cms/ro/content/studiu-exploratoriu-privind-stocurile-de-migra%c8%9bie
- https://www.libertatea.ro/stiri/dimensiunea-emigratiei-din-romania-ce-stim-si-ce-nu-despre-cat-de-mare-e-diaspora-2885018
- https://www.oecd-ilibrary.org/sites/bac53150-en/1/2/1/index.html?itemId=/content/publication/bac53150-en&_csp_=5911873c6569105028ad0a0066943c9d&itemIGO=oecd&itemContentType=book
- https://prezenta.bec.ro/parlamentare2016/
- http://bec2014.roaep.ro/wp-content/uploads/2014/11/SIAP2014_PAR_Raport-Situatie-Prezenta-la-urne.pdf
- https://prezenta.bec.ro/europarlamentare26052019/abroad-pv-final

Data sources

Recommended bibliography

- Finlay, S. (2014). *Predictive Analytics, Data Mining and Big Data*. Palgrave Macmillan, NY

- Pherson, R. (2019). *Handbook of Analytic Tools & Techniques*. Pherson Associates

Sources used for the case study:

- Oser, J., Hooghe, M., & Marien, S. (2013). Is online participation distinct from offline participation? A latent class analysis of participation types and their stratification. *Political Research Quarterly*, *66*(1), 91-101. https://doi.org/10.1177/1065912912436695
- Oser, J. (2017). Assessing How Participators Combine Acts in Their „Political Tool Kits”: A Person-Centered Measurement Approach for Analyzing Citizen Participation. *Social Indicators Research, An International and Interdisciplinary Journal for Quality-of-Life Measurement*, *133*(1), 235-238.
- Howard, M. M., Gibson, J. L., & Stolle, D. (2005). The U.S. citizenship, involvement, democracy survey. Washington, D.C.: Center for Democracy and Civil Society (CDACS), Georgetown University)
- Sultănescu, D. (2020). *Modele de participare politică în România democratică* [Models of political participation in democratic Romania]. [Unpublished doctoral dissertation]. National University of Political Studies and Public Administration, Bucharest, Romania.

Other graphics & photo source: www.freepik.com

Thank you!