



WOMEN IN DATA SCIENCE
AMERICAN UNIVERSITY
OF BEIRUT

 #WiDSAUB2023



#WiDSAUB2023



Transforming the Public Sector with Tech & AI



Carole Alsharabati

Professor, Université Saint Joseph
Research Director, Siren Associates

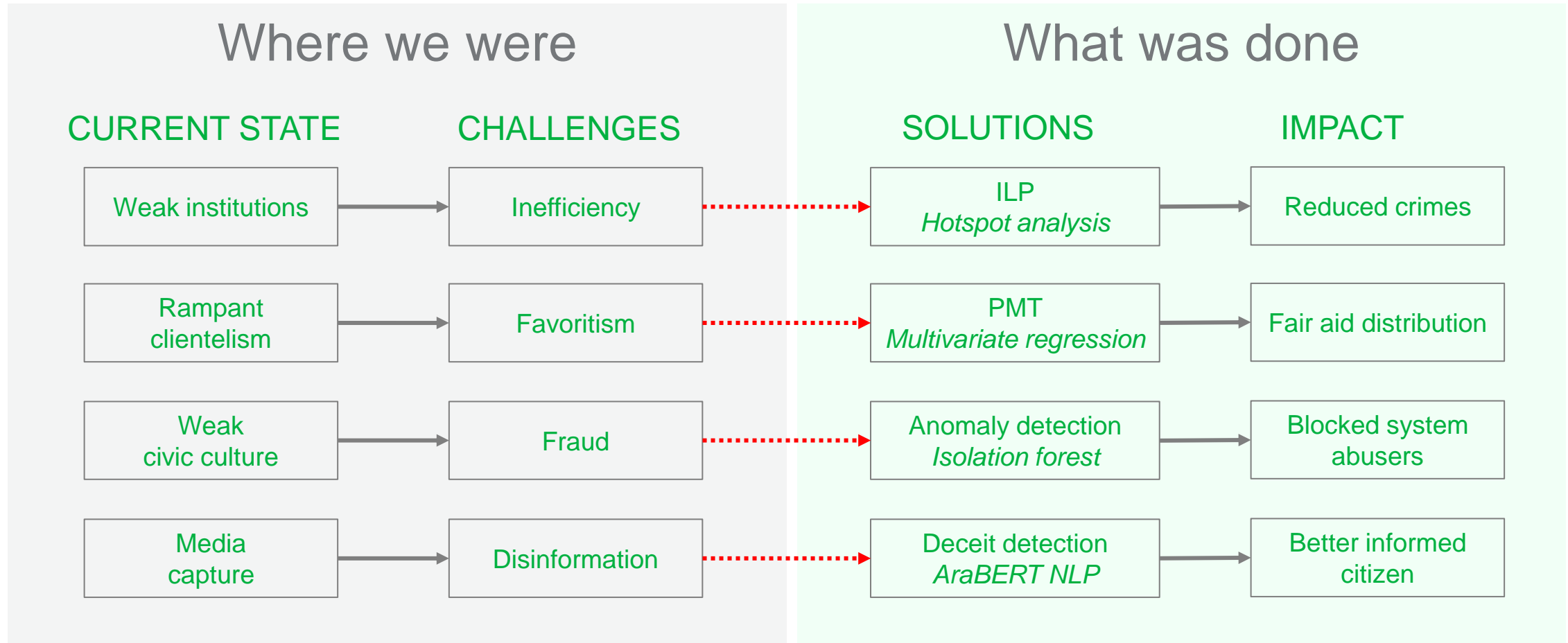


OUTLINE

1. The Big Picture
2. Improving safety with hotspot analysis
3. Strengthening civic culture with anomaly detection
4. Ensuring inclusive aid distribution through PMT
5. Constructive journalism with AraBERT and ChatGPT
6. Concluding Remarks



LEBANON'S AI EXPERIENCES



01

Improving safety with ILP

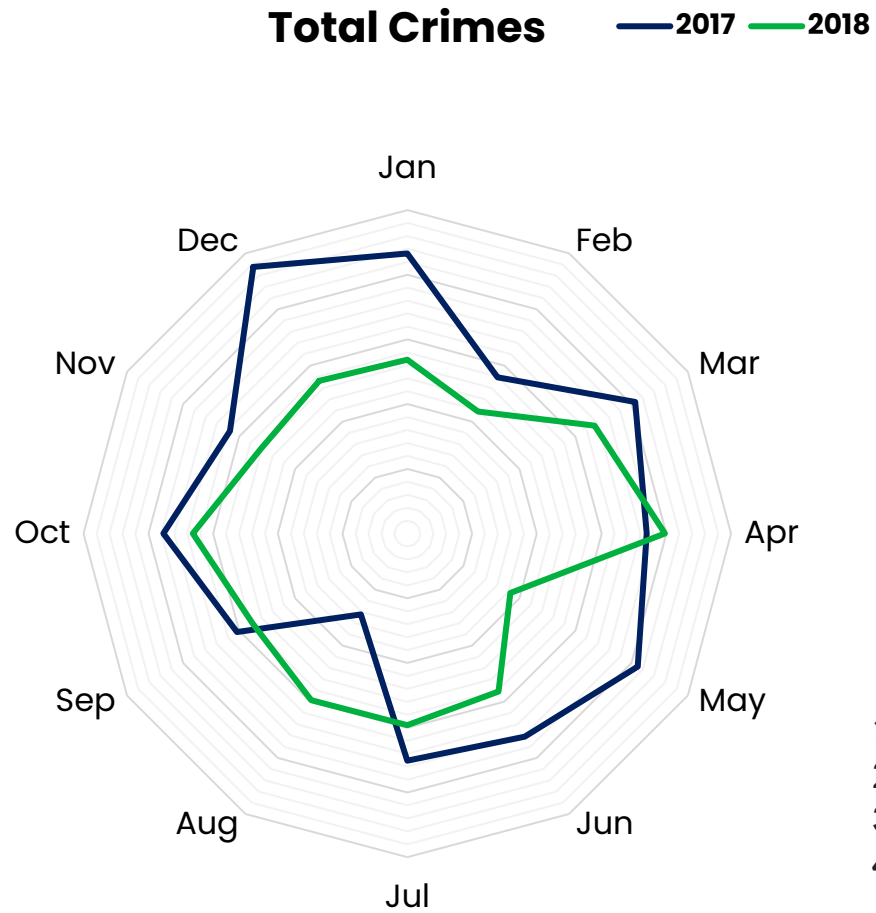
Hotspot Analysis



ILP to prevent crimes



Streamlined analytical products for decision-making



1. Space-time clustering
2. Probability of occurrence based on past incidence
3. Identification of crime prone zones
4. Deploy patrols as preventive measures

Strategic Assessment

Tactical Assessment

Problem Profile

Table 1: Different categories of crimes

Crime Type	2017	2018
Traffic	100	120
Vehicle crimes	80	90
Crimes committed by juveniles	50	60
Violent crimes	30	40
Crimes with weapons	20	30
Firearms	10	15
Other	100	110

Table 2: Bike brands

Brand	Count
Brand A	15
Brand B	10
Brand C	8
Brand D	5
Brand E	3
Brand F	2
Brand G	1
Brand H	1
Brand I	1
Brand J	1
Brand K	1
Brand L	1
Brand M	1
Brand N	1
Brand O	1
Brand P	1
Brand Q	1
Brand R	1
Brand S	1
Brand T	1
Brand U	1
Brand V	1
Brand W	1
Brand X	1
Brand Y	1
Brand Z	1

02

Strengthening civic culture with Machine Learning

Isolation Forest



Detecting and blocking exploits

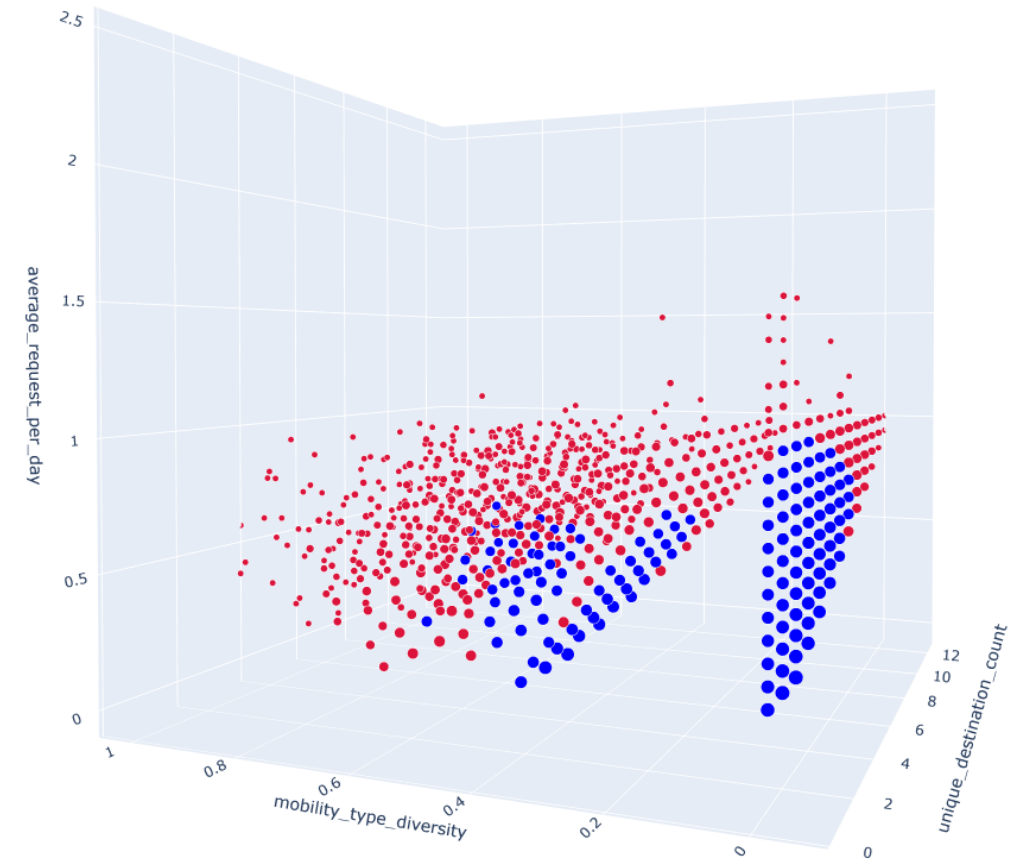
- Government mandated lockdowns in Lebanon, with permits processed by IMPACT
- 1M mobility requests per day with spikes of 1K requests per second
- Excessive use and exploits of excuses that gave 4-5 hours of allowed mobility
 - i.e. hospitals, pharmacy, doctors...
- Need for a validation system
- Objective: Block requests from users that attempt to exploit the platform

The screenshot shows the IMPACT mobile application interface. At the top, there is a header in Arabic: "طلب إذن خاص خلال فترة إعادة الفتح التدريجي" (Request a special permit during the gradual reopening period) and "Permissions during gradual reopening". The IMPACT logo is visible. Below the header, the text "طلب إذن خروج للأماكن القابلة للإكتظاظ" (Request a permit to exit crowded places) is displayed. The form contains several input fields: "الإسم *First Name" (Name), "الشهرة *Last Name" (Surname), "سنة الولادة *Year of Birth" (Year of Birth), "رقم الهاتف *Mobile Number" (Mobile Number), "بلدية الموقع الحالي *Current Municipality" (Current Municipality), "نوع المكان المقصود *Destination Type" (Destination Type), "تاريخ التنقل المرتقب - Expected Mobility Date" (Expected Mobility Date) with a calendar icon, and "إسم المكان المقصود *Destination Address Name" (Destination Address Name).



Machine Learning & Anomaly Detection

- Automatically flag anomalous behavior based on the user's request history (14 days).
- Assumptions:
 - Anomalous behavior consist of a small percentage of the total requests traffic
 - Anomalous behavior are more unique and diverse than normal user behavior.
- 4 key behavioral patterns were identified
 - Unique destination count
 - Requests frequency
 - Diversity of mobility types (car, bus, by foot...)
 - Unique Kadaa count
- Isolation forest: an unsupervised Tree-based Algorithm



03

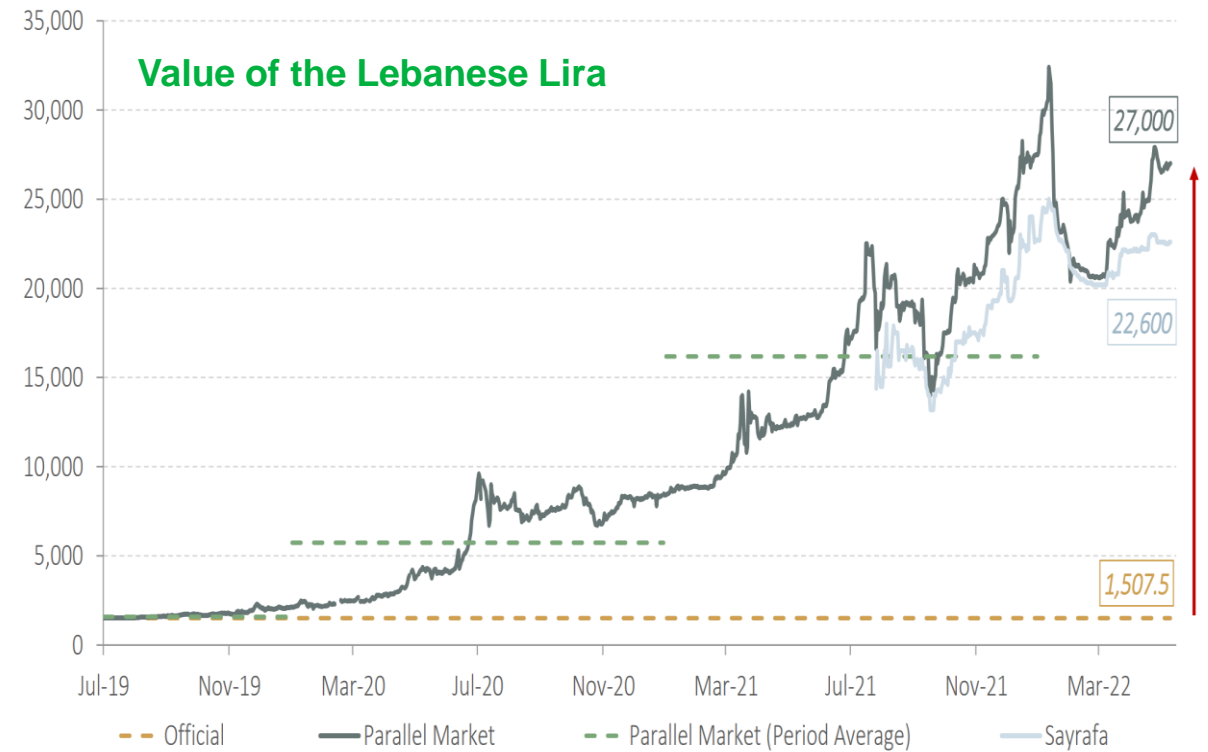
Ensuring inclusive aid distribution with PROXY MEANS TEST

Multivariate Regression



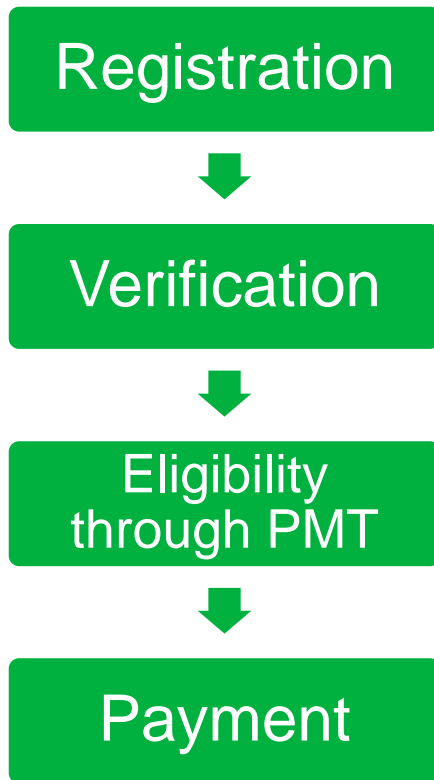
ESSN to the Rescue

Since 2019, Lebanon has endured a severe and prolonged economic and financial crisis. The Emergency Social Safety Network (ESSN), under MoSA and the PCM, was created as an add-on to the National Poverty Targeting Program to further extend social assistance programming.



PROXY MEANS TEST

Based on a statistical analysis of the population, different weights are assigned to the variables depending on their influence on household consumption. The model includes more than 40 variables including income, assets, housing quality, occupation, and demographic characteristics correlated with poverty.



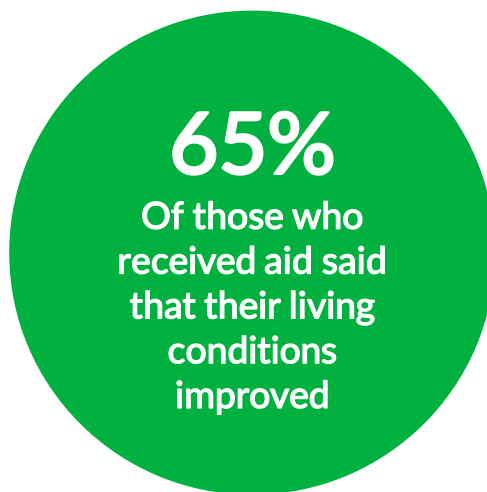
<i>variable</i>	<i>coefficient</i>
<i>Household owns the house</i>	100
<i>One child in the household</i>	40
<i>Two children in the household</i>	30
<i>Three or more children in the household</i>	20
<i>Household owns cattle</i>	200
<i>Household owns a bicycle</i>	300
<i>Household owns a car</i>	800
<i>Dwelling walls made of brick</i>	100
<i>Dwelling walls made of tin</i>	0
<i>Dwelling walls made of clay</i>	-100
Constant	1000



Did we target the extreme poor?

550,000 households registered, 250,000 visited, and 80,000 enrolled through PMT.
 Survey was done on a sample of 1600 randomly picked respondent to validate the PMT results (Y4G summer of 2022).

The World Bank has defined extreme poverty as people living on less than \$2.15 a day, measured using the international poverty line.



	ESSN Paid	Registered	Not Registered
At least one member with disability	24.6%	16.2%	12.4%

	ESSN Paid	Registered	Not Registered
Average dependency ratio	3.91	3.16	2.88

04

Constructive journalism with NLP

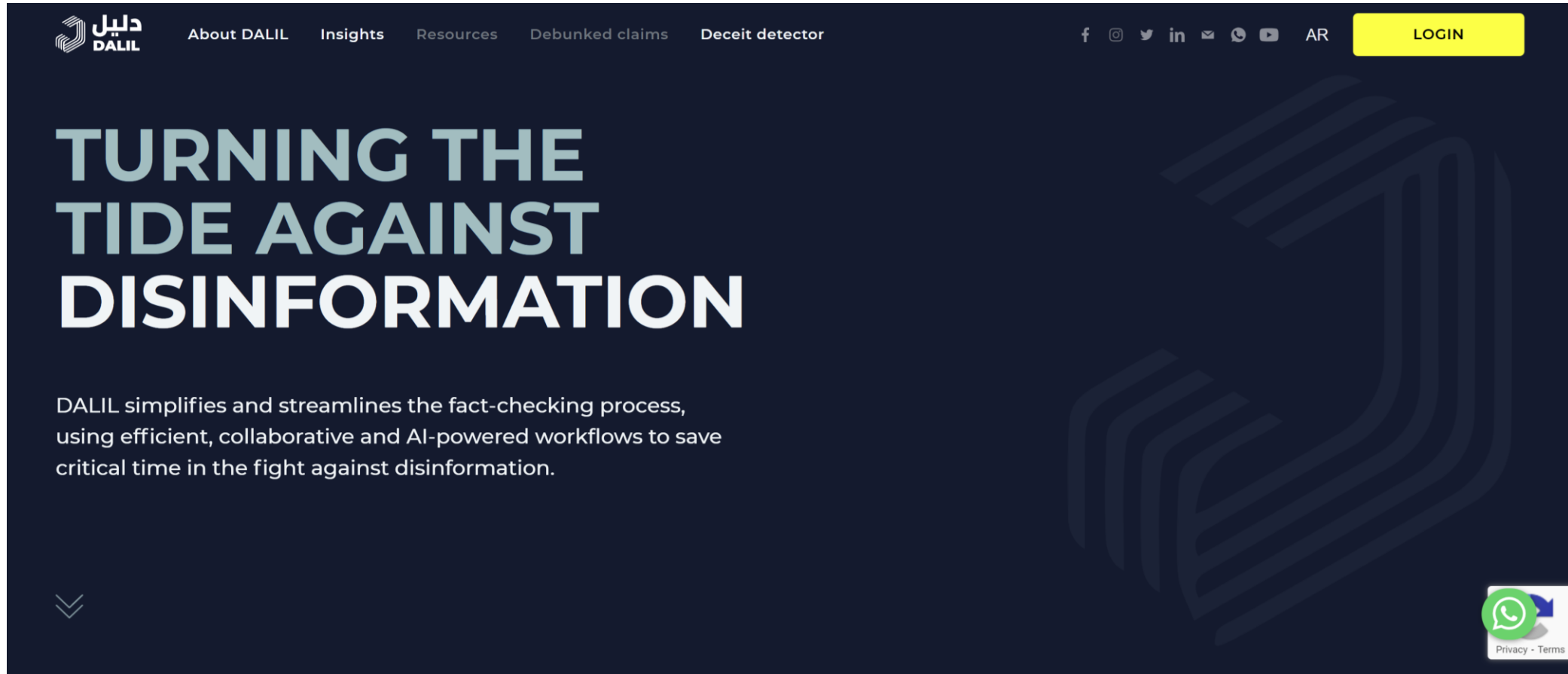
When ChatGPT met AraBERT



Towards a constructive role of media: Detecting deceit

Media in Lebanon are serving politics.

The objective is to provide tools for journalists and citizens to detect disinformation.



The screenshot shows the homepage of the DALIL website. The header includes the DALIL logo (دليل DALIL) and navigation links: About DALIL, Insights, Resources, Debunked claims, and Deceit detector. Social media icons for Facebook, Instagram, Twitter, LinkedIn, Email, Telegram, YouTube, and AR are also present, along with a yellow LOGIN button. The main content area features the headline "TURNING THE TIDE AGAINST DISINFORMATION" in large, bold, white and light blue text. Below the headline, a paragraph states: "DALIL simplifies and streamlines the fact-checking process, using efficient, collaborative and AI-powered workflows to save critical time in the fight against disinformation." A large, faint fingerprint graphic is visible in the background. At the bottom right, there is a WhatsApp icon and a "Privacy - Terms" link.



Chat GPT and AraBERT Duo

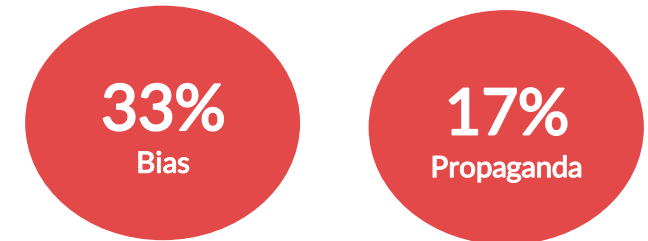
Reliance on AraBERT (Antoun, Wissam, et al.)

More particularly AraBERTv02 large model, which contains 371 million parameters, equivalent to 1.38 gigabytes of data, and was pre-trained on 77 gigabytes of Arabic corpus.

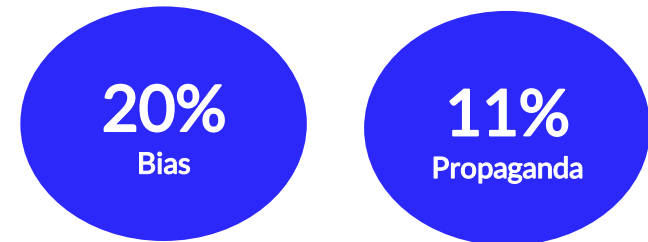
- Labeling through open source, human, ChatGPT
- 7200 articles from various topics in the MENA Labeled
- Trained AraBERT to recognize Arabic language propaganda and Subjectivity indicators
- Applied to local media
- Daily KPIs

```
Consider yourself an annotator.  
I am providing you with this text:  
  
{text_input}  
  
I want you to label for me this text for two models, first by checking if the given text contains any propaganda.  
If yes, label for me the propaganda by providing the Span containing the propaganda and the technique of the propaganda by choosing one of the techniques  
inside this list [Loaded Language, Name Calling/Labeling, Repetition, Doubt, Exaggeration/Minimization, Others]  
For the second model I want you to check if the given text is subjective or objective by returning a span on the subjective part of the text if found.  
And I want your output to be in this json format and without explanations  
{  
  "Propaganda" : bool,  
  "Propaganda Span" : [list of text containing propaganda],  
  "Propaganda Technique" : [list of technique for each span],  
  "Subjective" : bool,  
  "Subjectivity Span" : [list of text containing subjectivity]  
}  
Whenever a value is empty write NULL without ""
```

Newspaper x



Newspaper y



The way forward: DALIL journalism

Going forward, the platform will be opened up to media with a dedicated space for journalists where they will be able, among other things, to run their content through bias, propaganda and consensus detectors in order to check it before publication; think Turnitin meets fact-checking.

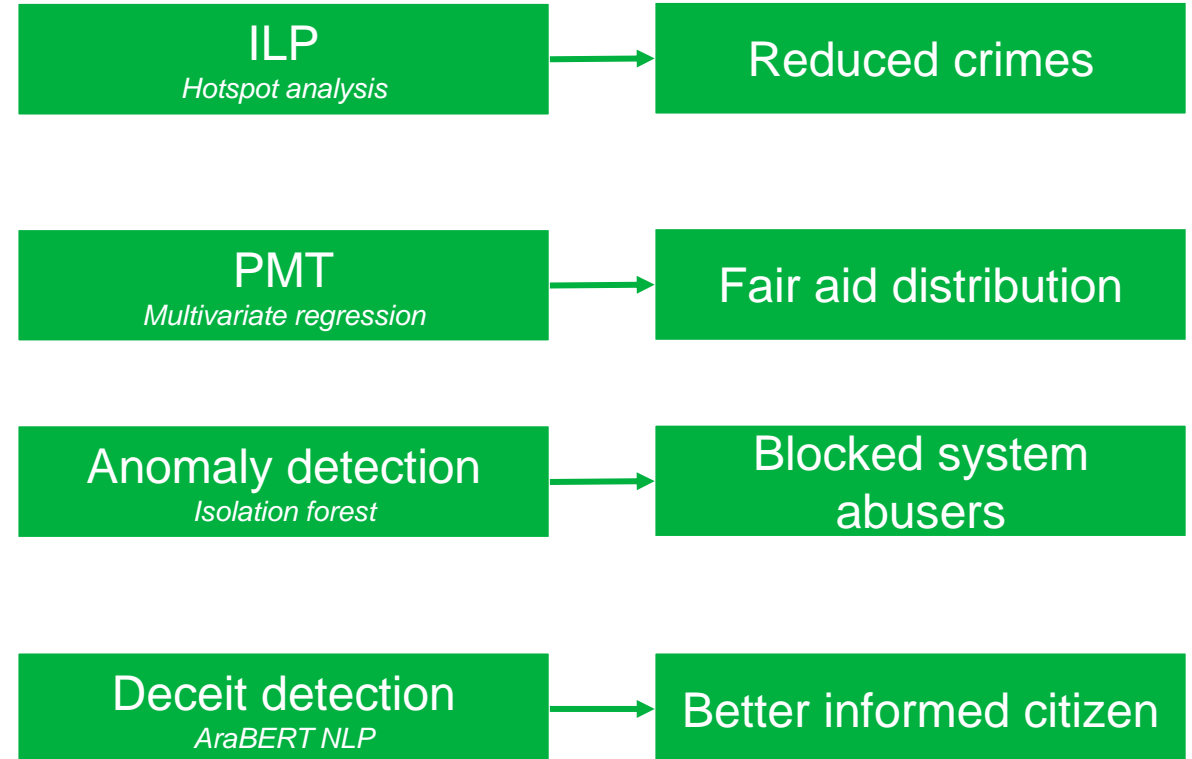


Concluding Remarks

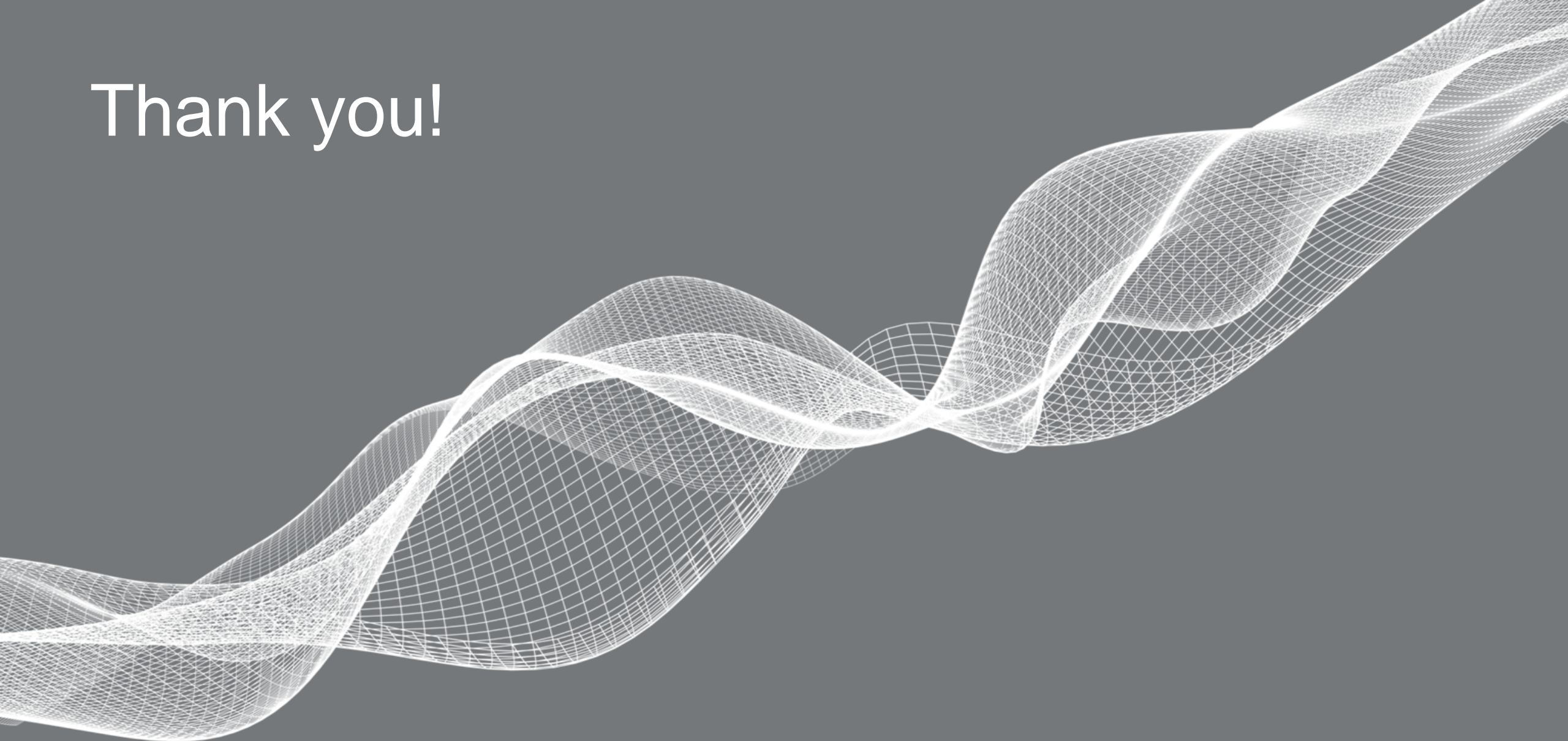
- Data-driven decision making processes have huge potential in low resource contexts
- There are champions of change in the public sector who will quickly embrace data-driven decision processes
- Once integrated into the processes, AI can help gain efficiency, bring fairness, control fraud, preempt deceit, and much more.

SOLUTIONS

IMPACT



Thank you!



Hotspot with Getis-Ord G_i^*

$$G_i^* = \frac{\sum_{j=1}^n w_{ij} x_j - \bar{x} \sum_{j=1}^n w_{ij}}{S \sqrt{\frac{n \sum_{j=1}^n w_{ij}^2 - \left(\sum_{j=1}^n w_{ij} \right)^2}{n-1}}}$$

1-Spatial weights matrix for relationship between data points

2-Incorporate temporal component in weights.

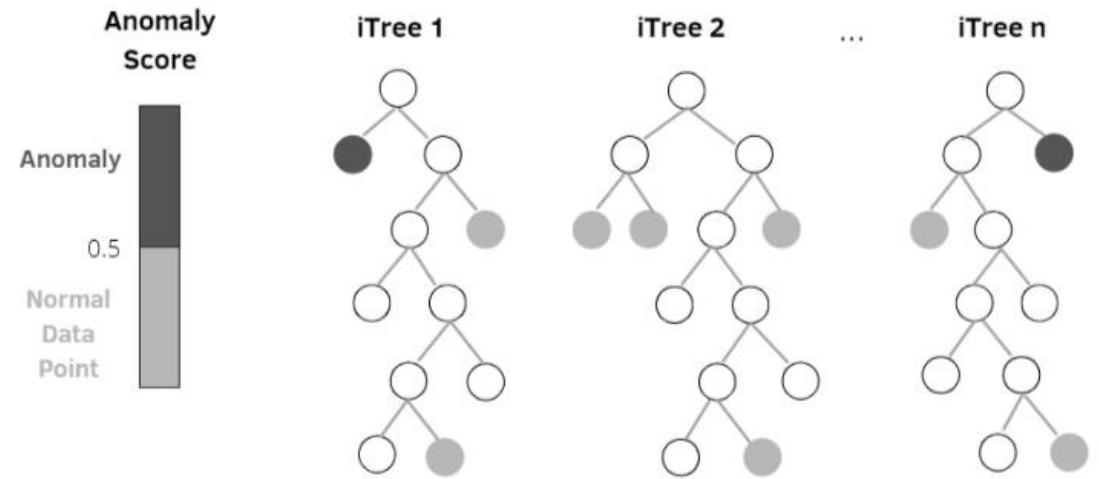
3-Getis-Ord G_i^* statistic degree of surrounding by other points with high or low values.

3-Calculate z-score

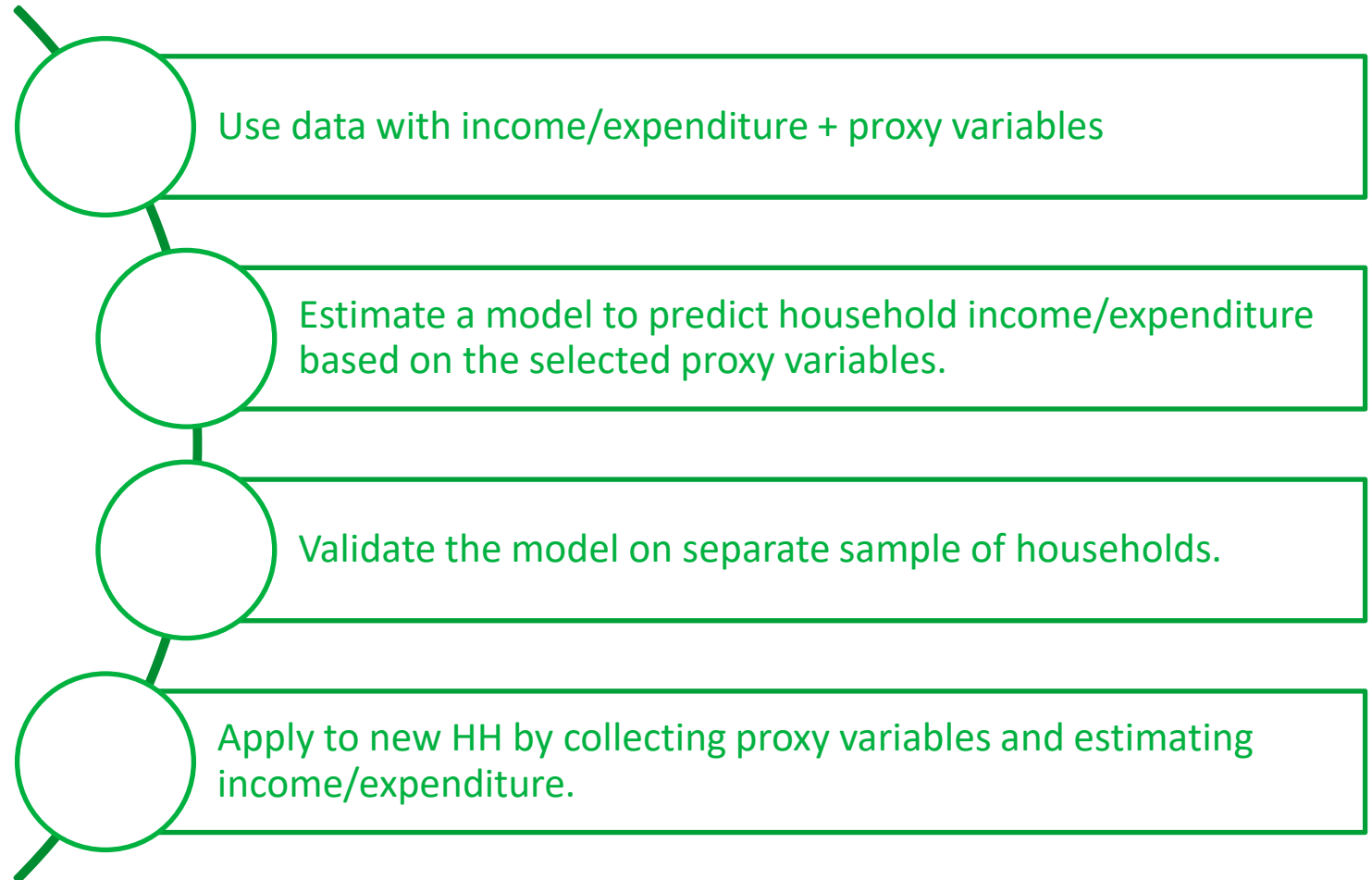
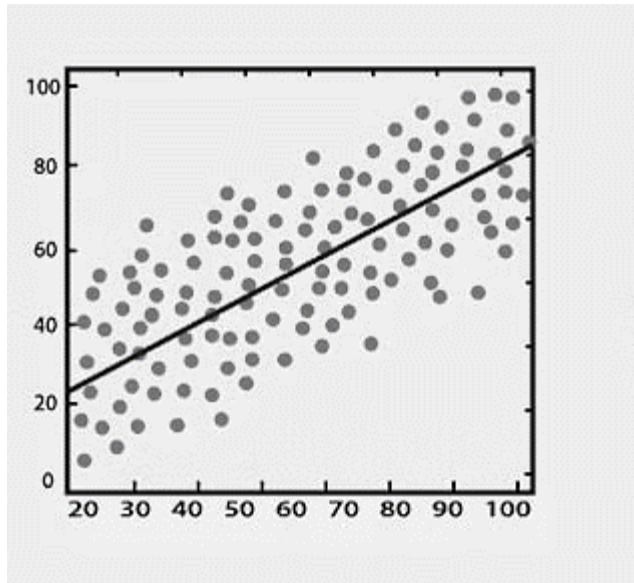


Isolation Forest

1. Select a random subset of data
2. Select a random feature
3. Choose a random split point
4. Partition the data
5. Repeat the process recursively for each subset until all the data points are isolated
6. Calculate anomaly score based on the number of partitions required to isolate it
7. Anomalies are identified as data points with lower anomaly scores.
8. Set threshold and classify anomalies.



PMT



BERT and AraBERT (Bidirectional Encoder Representations from Transformers)

- Bi-directional encoding (vectors) to process text in both directions & capture the context and meaning based on the entire text.
- Self-attention mechanism to weigh the importance of different words in a sentence based on the context of the sentence.
- Transformer architecture, a neural network to capture long-range dependencies and relationships between words and sentences.

