UNITED NATIONS

Politecnico di Torino

1859

UNITED NATIONS SYSTEM STAFF COLLEGE

Gabriele Cirotto S307732

Alberto Foresti S309212

Luca Varriale S300795

Predictive Models for Humanitarian Aid

Predicting conflicts in the Sahel region

Professor: Giuseppe Rizzo
UN Staff: J.Yang, I.Arispe

ADSP Final Presentation
A.Y. 2023/24

PROJECT PROPOSITION

Help people in the Sahel region with a faster and effective response to conflicts and social tensions

OUTPUT

AI System

- Research report on standard framework and alternatives
- Process log to help future learners
- Dashboard for communication and helping users understand the results

RESEARCH QUESTIONS

What we try to answer?

RESEARCH QUESTIONS

PREDICTORS

What are the best predictors for the task?

- Action-driven predictors
- Interpretable predictors
- Informative predictors

RESEARCH QUESTIONS

PREDICTORS - EVENTS

Our goal is to predict the fatalities for a certain number of months.

Political Violence

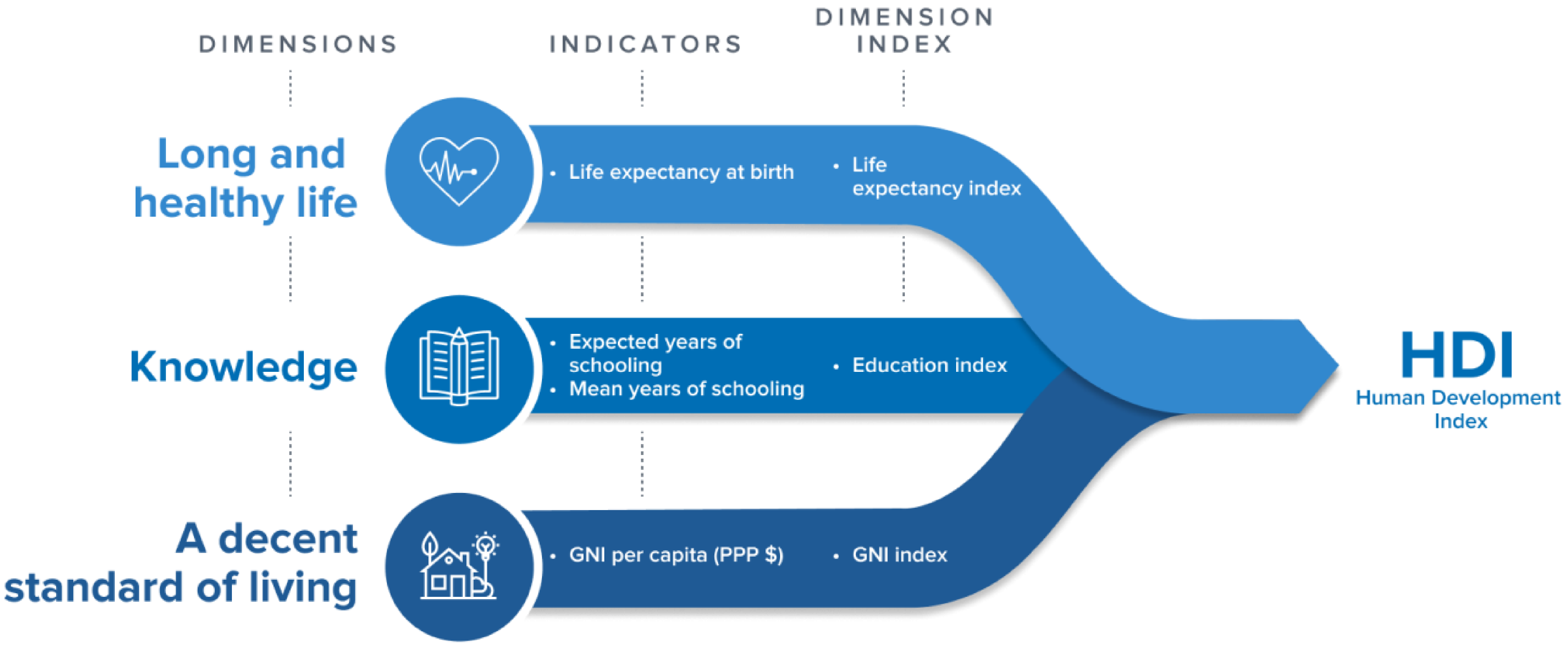
Demonstrations

Strategic developments

RESEARCH QUESTIONS

PREDICTORS - HDI

Human Development Index



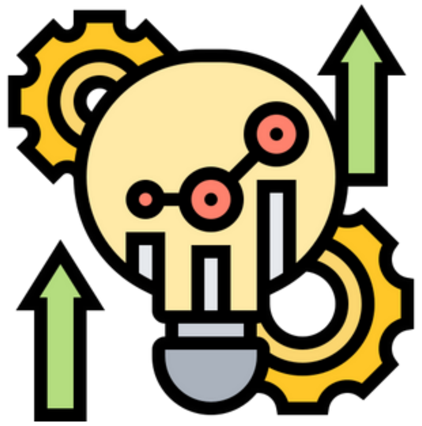
Reference to the image: https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

RESEARCH QUESTIONS

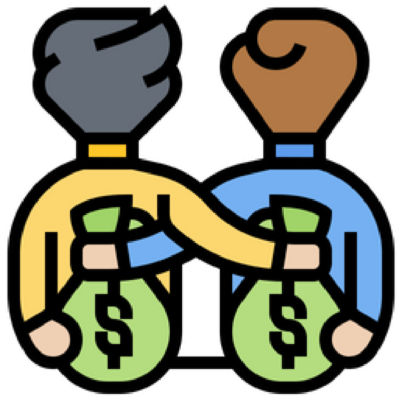
PREDICTORS - GDI

THE WORLD BANK

Voice and Accountability



Regulatory Quality



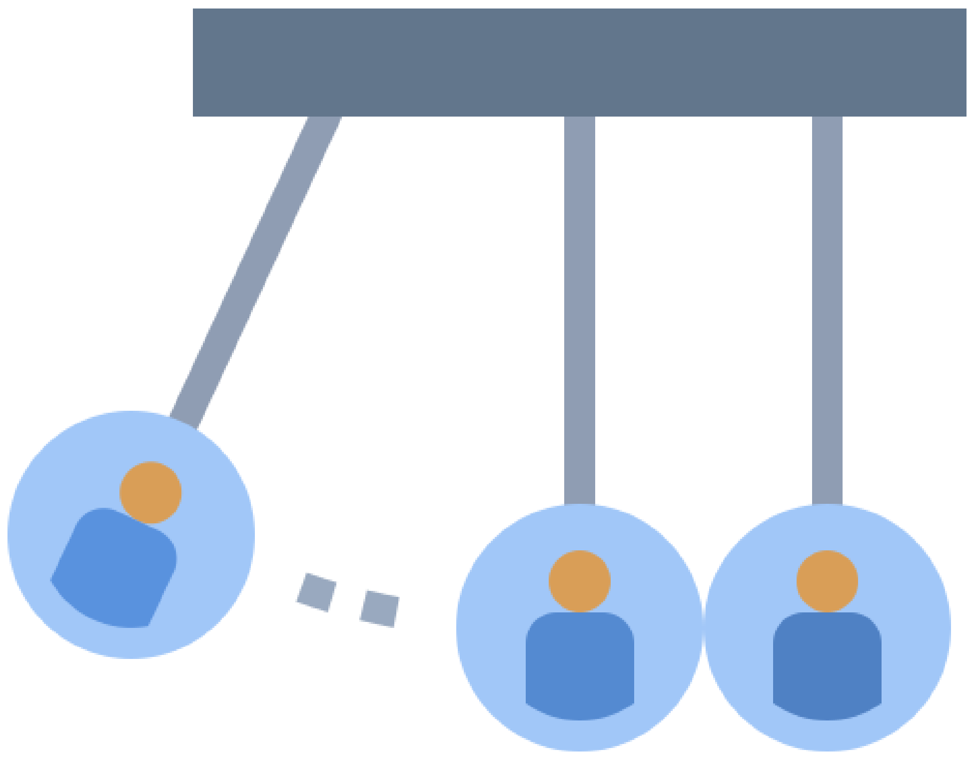
Control of Corruption

Political Stability & Absence of Violence/Terrorism

Rule of Law

RESEARCH QUESTIONS
CAUSAL DISCOVERY

Can we explore the task with causal discovery?

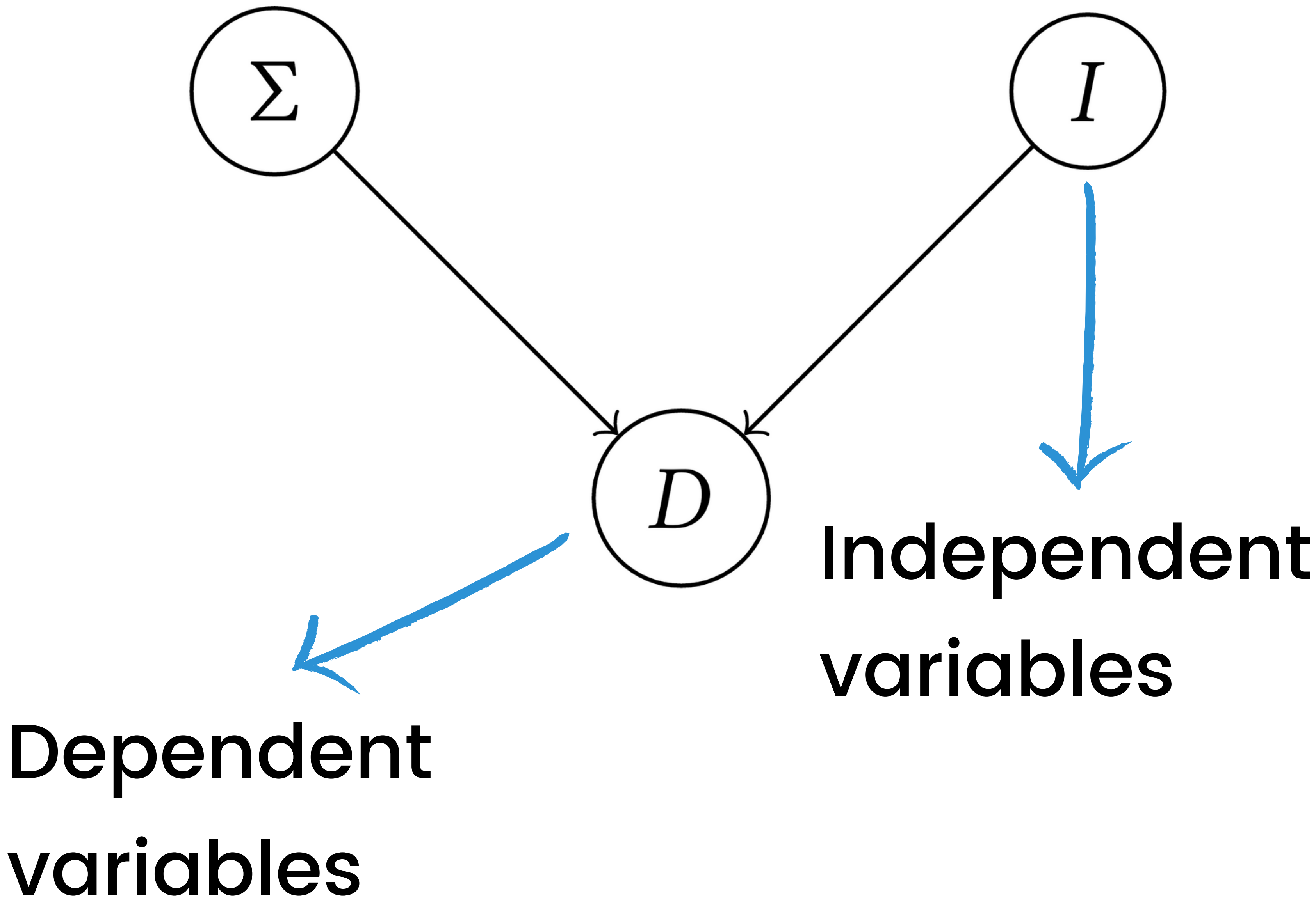


- Highlight few independent factors
- Beyond statistics: can we highlight causation directly from data?

RESEARCH QUESTIONS

CAUSAL DISCOVERY-CAUSAL GRAPHS

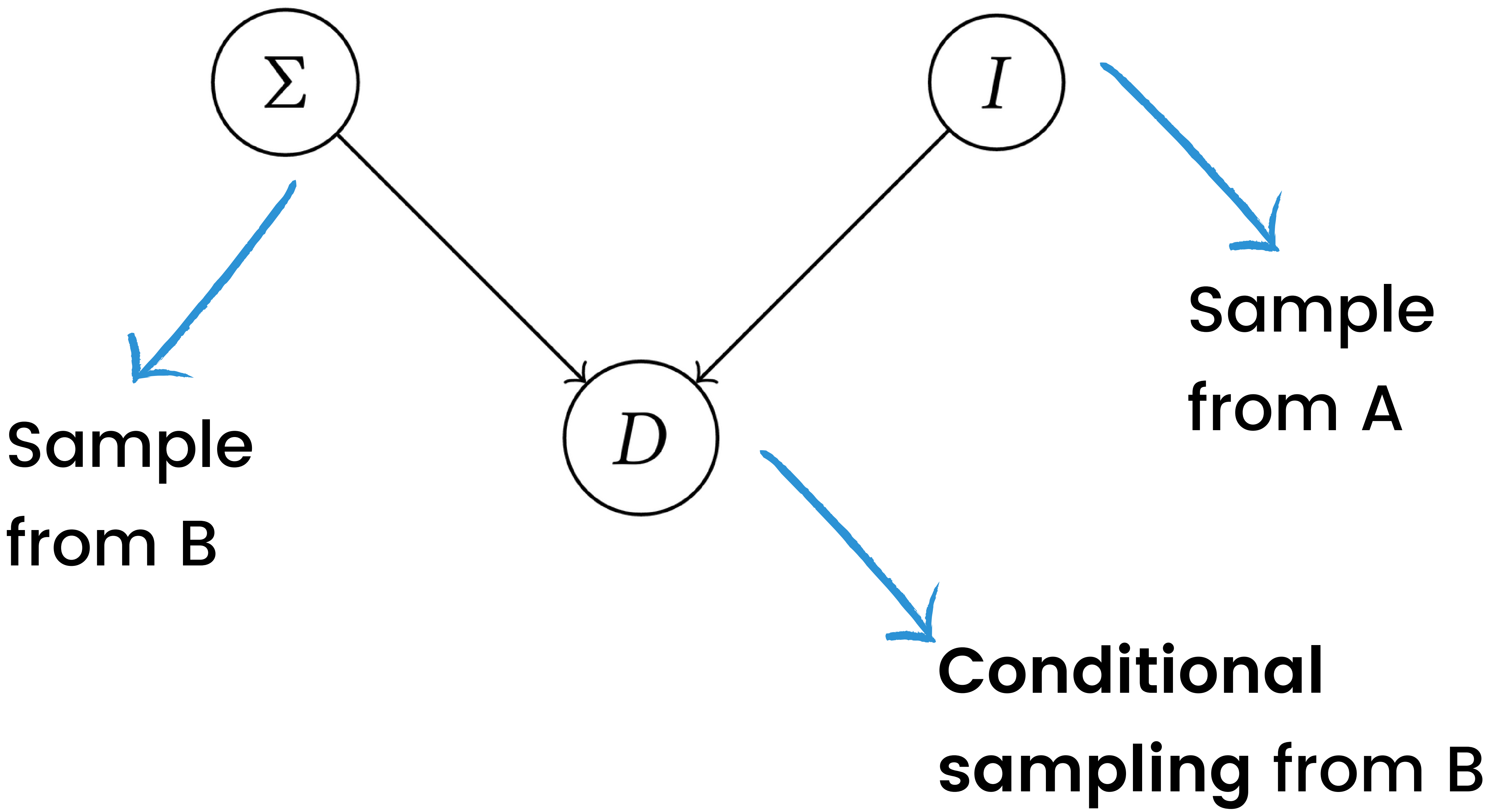
We sample new data with causal awareness



- We set a **shift variable:** Σ
- We start sampling from parent nodes: **causal parents**
- We sample new features by **conditioning** on causal parents

RESEARCH QUESTIONS
CAUSAL DISCOVERY-CAUSAL DATA AUGMENTATION

Our sampling uses a dataset split



Two datasets:

- **A**: the dataset we want to augment
- **B**: the dataset sampled from the target distribution

RESEARCH QUESTIONS

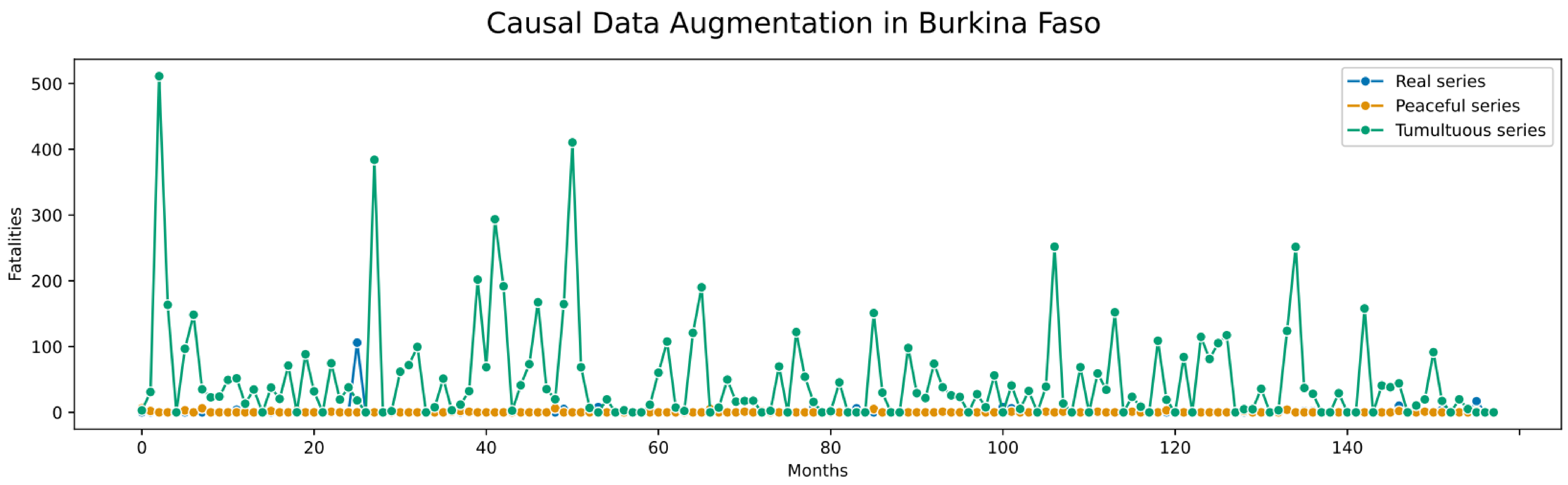
CAUSAL DISCOVERY-CAUSAL DATA AUGMENTATION

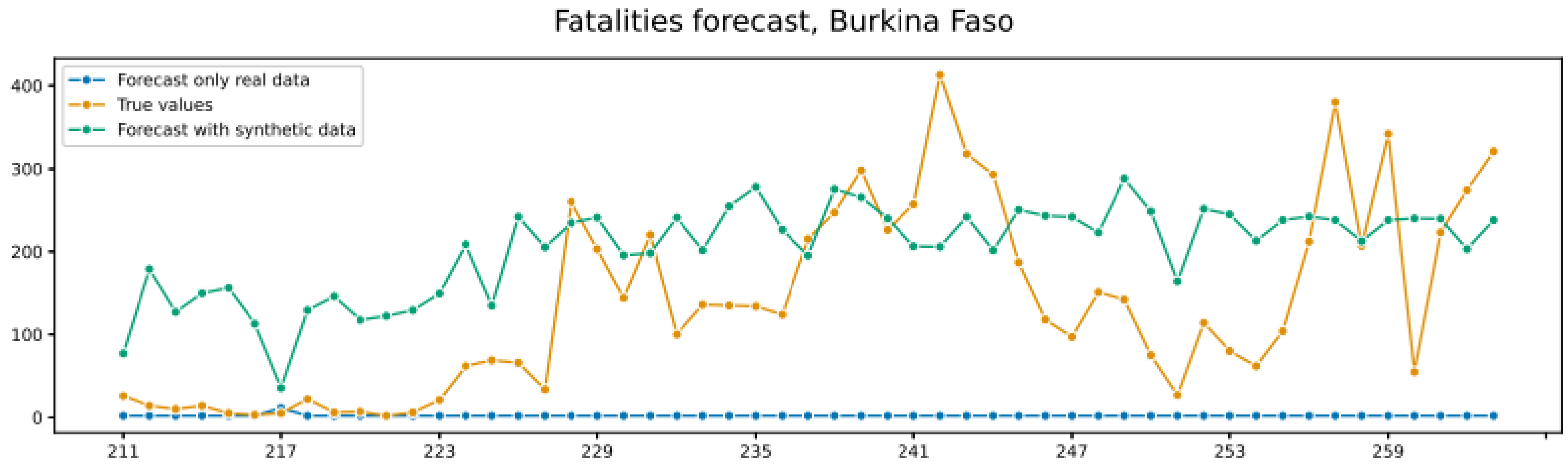
Our method:

- For every country we simulate a change of scenario using causal graphs and we create synthetic **distributions**
- We **sample** data from the new distributions
- We use synthetic and real data to train **robust** ML models

RESEARCH QUESTIONS

CAUSAL DISCOVERY-CAUSAL DATA AUGMENTATION





RESEARCH QUESTIONS

CAUSAL DISCOVERY-CAUSAL DATA AUGMENTATION

Advantages

- Model agnostic
- Regularization
- Robust

Drawbacks

- Risk of catastrophic forgetting
- Slow

RESEARCH QUESTIONS

METRICS

Can we find better metrics?

Time Series Forecast Error Metrics

Scale-Dependent Metrics

- Mean Absolute Error
- Mean Squared Error
- Root Mean Square Error

Percentage-Error Metrics

- Mean Absolute Percentage Error
- Symmetric Mean Absolute Percentage Error

Relative-Error Metrics

- Median Relative Absolute Error
- Geometric Mean Relative Absolute Error

Scale-Free Error Metrics

- Mean Absolute Scaled Error

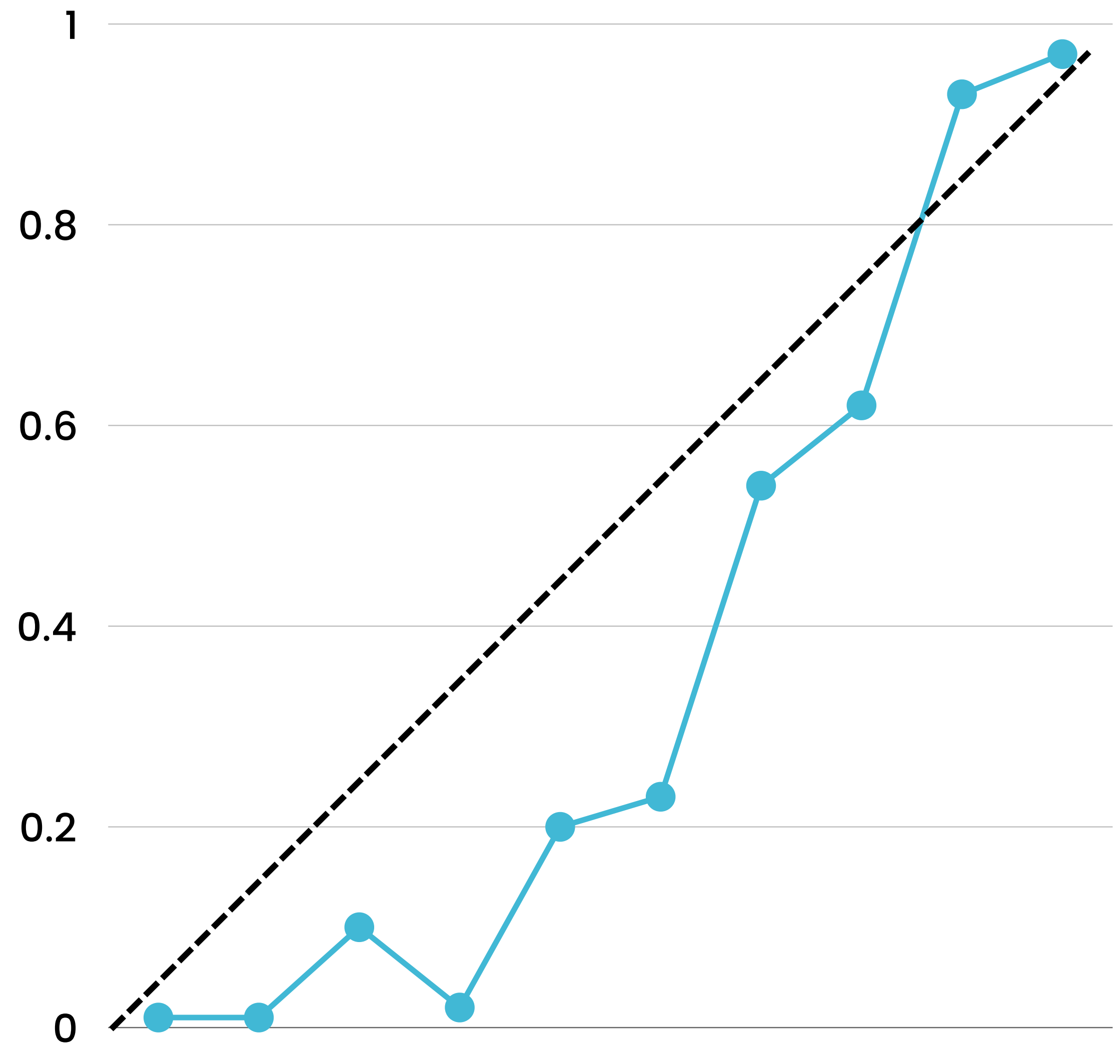
RESEARCH QUESTIONS
METRICS

- **Spike precision**: the fraction of spikes the model is able to predict
- **Spike recall**: the fraction of spikes, among the predicted, that are actually spikes

RESEARCH QUESTIONS

CALIBRATION

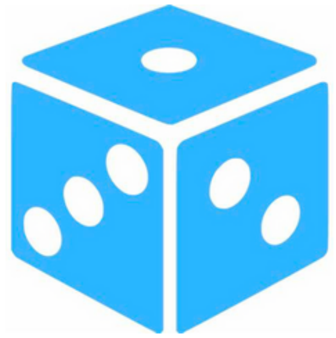
Can we find a general calibration strategy?



- Machine learning models do not provide real probabilities: we need calibration
- Drawback: calibration strategies rely on assumptions, we cannot assume much in the time series domain

RESEARCH QUESTIONS

CALIBRATION



Our tool:

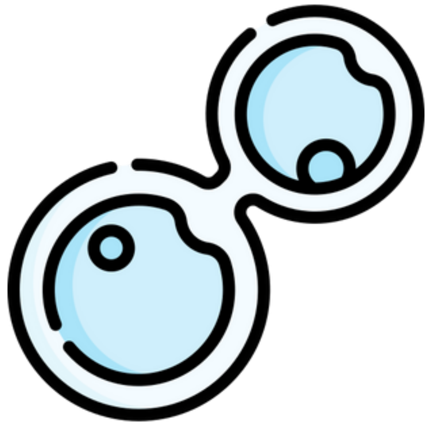
Conformal Prediction

Model agnostic framework with guarantees on real probabilities

Problem:

It relies on IID assumption

Not robust to distribution shifts

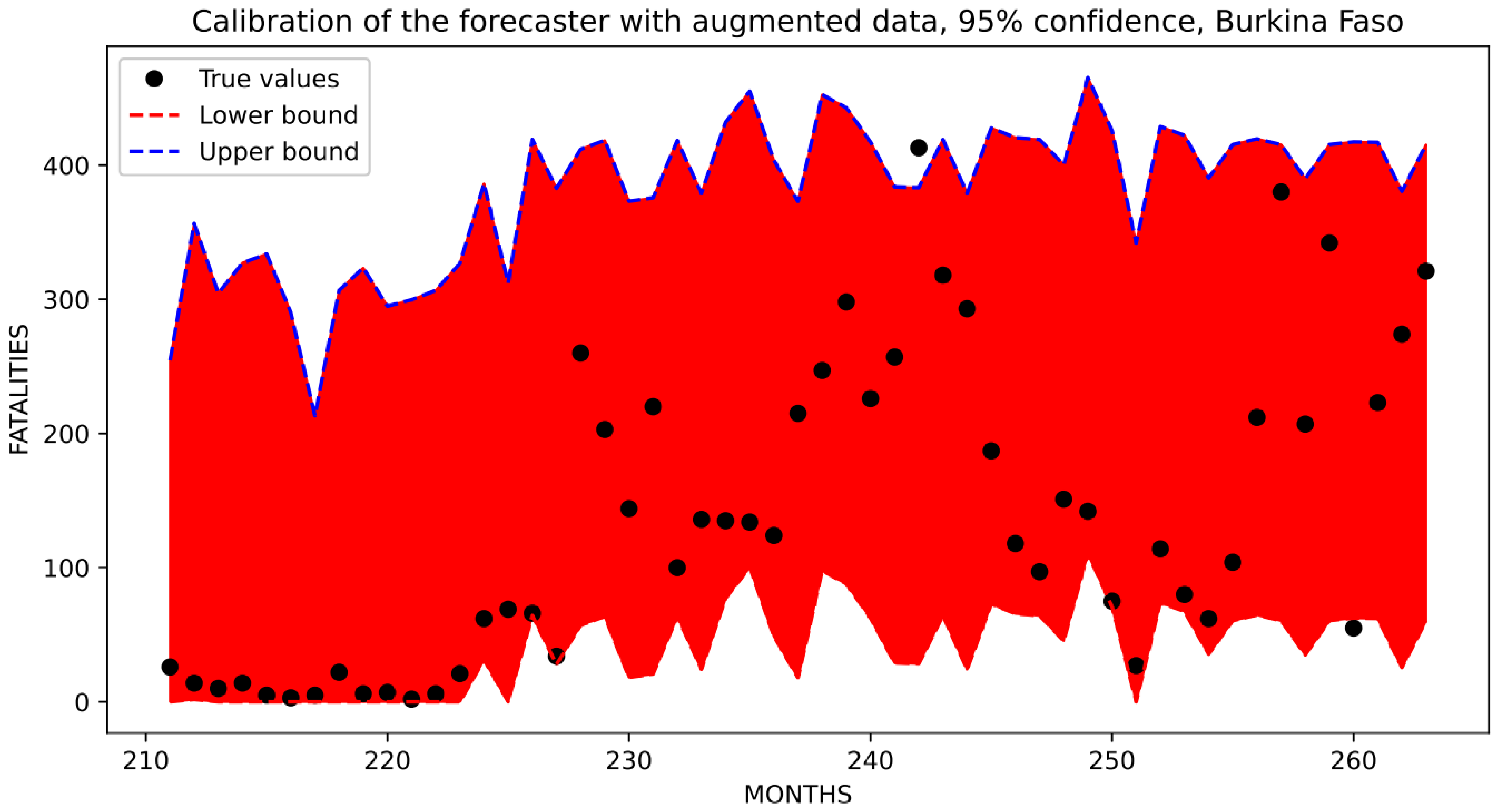


Solution:

Causal data augmentation

Augment also the calibration set

RESEARCH QUESTIONS
CALIBRATION



ACHIEVEMENTS

What did we achieve?

PERFORMANCE

	MEAN ABSOLUTE ERROR	SPIKE PRECISION	SPIKE RECALL
RANDOM FOREST - REAL DATA	116.47	0.33	0.40
RANDOM FOREST - REAL + SYNTETHIC DATA	104.48	0.37	0.43
LIGHTGBM - REAL DATA	92.82	0.33	0.55
LIGHTGBM - REAL + SYNTHETIC DATA	79.75	0.37	0.44
XGBOOST - REAL DATA	103.83	0.34	0.40
XGBOOST - REAL + SYNTHETIC DATA	76.51	0.35	0.42

PERFORMANCE

	MEAN ABSOLUTE ERROR	SPIKE PRECISION	SPIKE RECALL
RANDOM FOREST - REAL DATA	116.47	0.33	0.40
RANDOM FOREST - REAL + SYNTETHIC DATA	**104.48**	**0.37**	**0.43**
LIGHTGBM - REAL DATA	92.82	0.33	0.55
LIGHTGBM - REAL + SYNTHETIC DATA	79.75	0.37	0.44
XGBOOST - REAL DATA	103.83	0.34	0.40
XGBOOST - REAL + SYNTHETIC DATA	76.51	0.35	0.42

PERFORMANCE

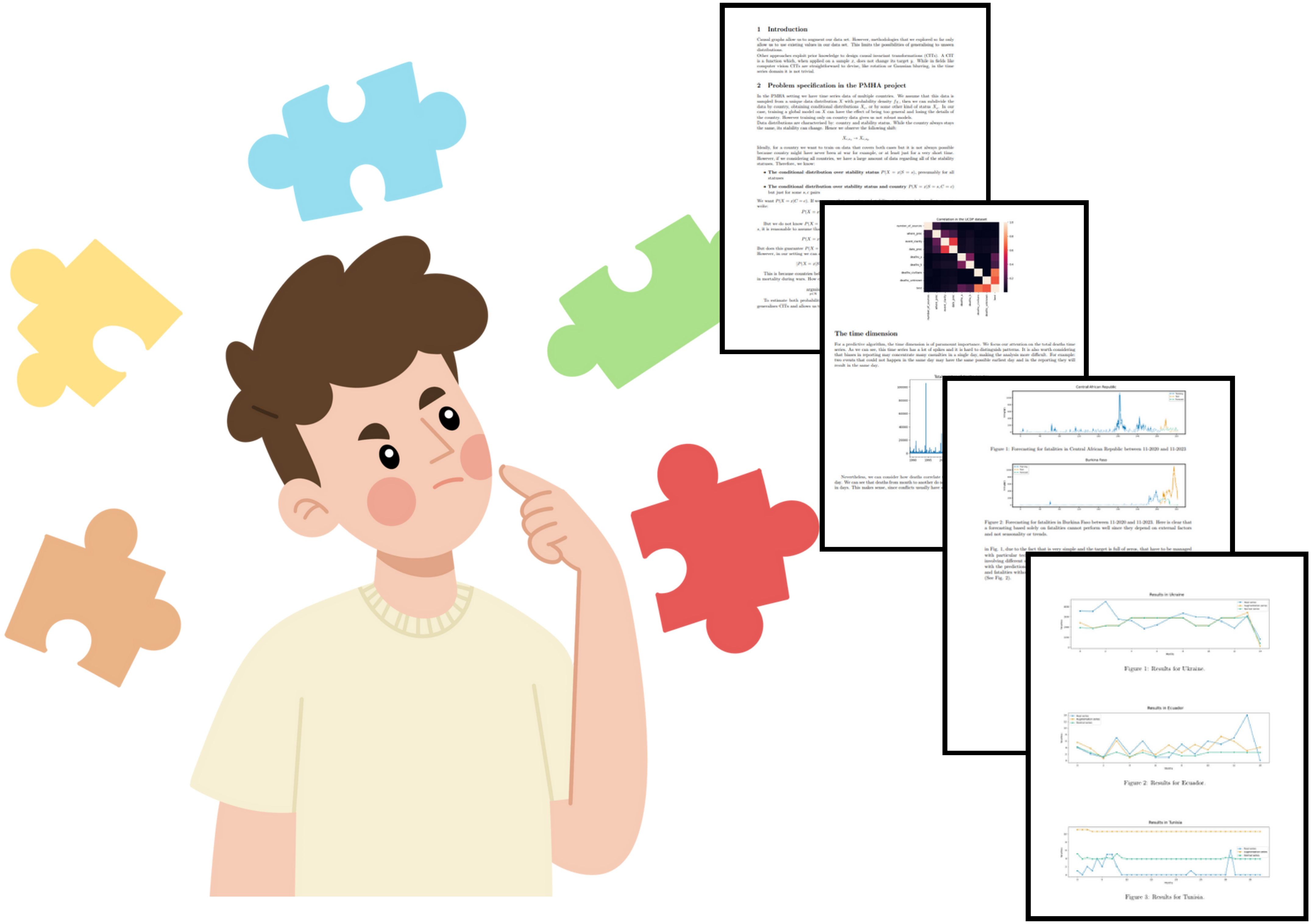
	MEAN ABSOLUTE ERROR	SPIKE PRECISION	SPIKE RECALL
RANDOM FOREST - REAL DATA	116.47	0.33	0.40
RANDOM FOREST - REAL + SYNTETHIC DATA	104.48	0.37	0.43
LIGHTGBM - REAL DATA	92.82	0.33	0.55
LIGHTGBM - REAL + SYNTHETIC DATA	79.75	0.37	0.44
XGBOOST - REAL DATA	103.83	0.34	0.40
XGBOOST - REAL + SYNTHETIC DATA	76.51	0.35	0.42

PERFORMANCE

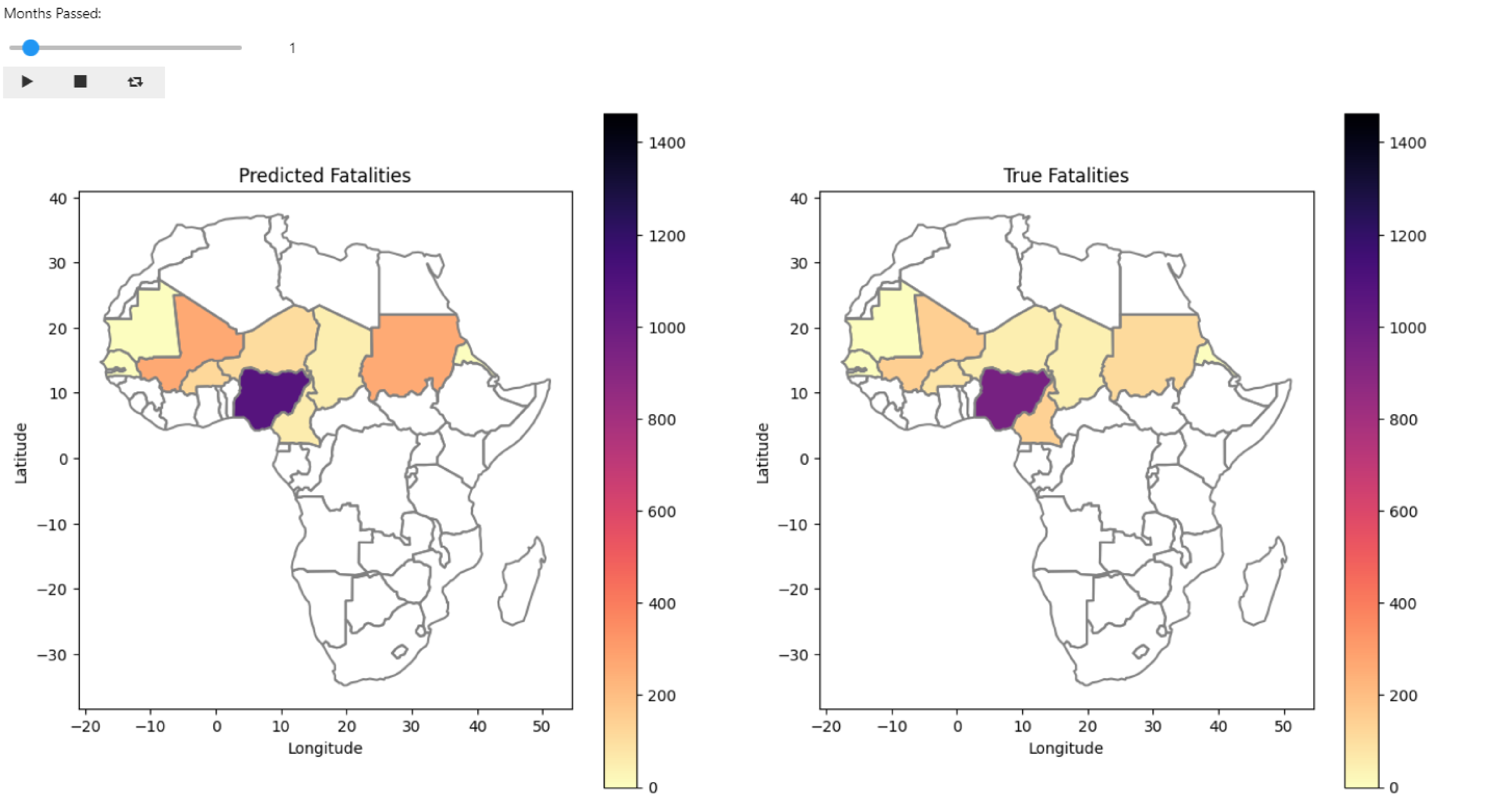
	MEAN ABSOLUTE ERROR	SPIKE PRECISION	SPIKE RECALL
RANDOM FOREST - REAL DATA	116.47	0.33	0.40
RANDOM FOREST - REAL + SYNTETHIC DATA	104.48	0.37	0.43
LIGHTGBM - REAL DATA	92.82	0.33	0.55
LIGHTGBM - REAL + SYNTHETIC DATA	79.75	0.37	0.44
XGBOOST - REAL DATA	103.83	0.34	0.40
XGBOOST - REAL + SYNTHETIC DATA	76.51	0.35	0.42

LOGS

Continuous production of short process logs for enhancing the understanding of our researches



DASHBOARD



Months Passed:
1
Predicted Fatalities
True Fatalities
Latitude
Longitude
40
30
20
10
0
−10
−20
−30
−20
−10
0
10
20
30
40
50
1400
1200
1000
800
600
400
200
0

CONCLUSIONS

What we have seen so far?

- Proposition
- Objective
- Research Questions
- Achievements

CONCLUSIONS
ACHIEVEMENTS

Dataset Construction for Sahel Region Fatality Prediction

Causal Data Augmentation for Distribution Shifts

Causal Data Augmentation for Calibration

CONCLUSIONS
FUTURE WORKS

- Integration of described methods on current early-warning systems
- Stability Enhancement for Causal Data Augmentation
- Efficiency Improvement in CDA Data Generation
- Automated Domain Split Selection

UNITED NATIONS
Politecnico di Torino
1859
UNITED NATIONS SYSTEM STAFF COLLEGE

Gabriele Cirotto S307732
Alberto Foresti S309212
Luca Varriale S300795

Thank you for your attention!

Scan me!

Questions?

Link to our repository: GitHub

ADSP Final Presentation
A.Y. 2023/24