Predicting and Gaining Insights on Olympic Games: A Temporal Hierarchical Bayesian Network Approach

Ramazan Ege Solak
Mehmet Emre Tiryaki
Sun Tanomkiattikun



JOHNS HOPKINS
UNIVERSITY

COMAP 2025 Mathematical Contest in Modeling JHU Team Submission

- This presentation summarizes the JHU team's submission to Problem C: **Modeling Olympic Medal Tables**, part of the COMAP MCM 2025 Contest.
- Most medal prediction models use extensive **external data** such as athlete performance indicators, GDP, Olympics budgets to predict medals.
- This contest limits the data usage to **only** the given dataset, which transforms the problem into one of solving the medal-prediction problem using only Olympics level historical medal information in a low-resource setting



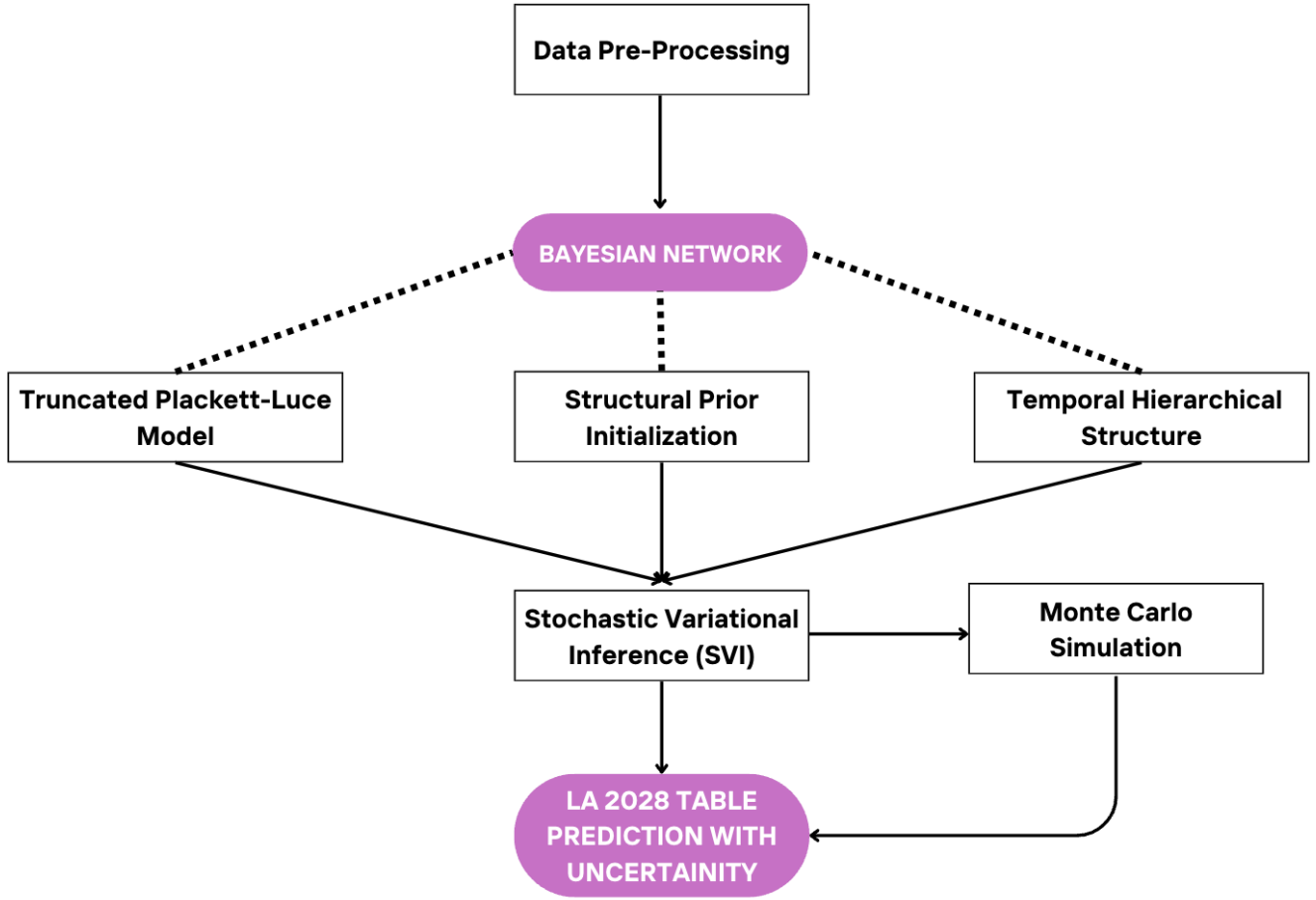
Data Description

- Athlete-level medal data (252,565 data points)
- Host Country data for each Olympics
- Medal counts per event and sport
- No data on extra-Olympics country level information (budget, population, etc.)
- No data on coaches

Name	Sex	NOC	Year	City	Sport	Event	Medal
A Dijiang	M	CHN	1992	Barcelona	Basketball	Basketball Men's Basketball	No medal
A Lamusi	M	CHN	2012	London	Judo	Judo Men's Extra-Lightweight	No medal
Gunnar Aaby	M	DEN	1920	Antwerpen	Football	Football Men's Football	No medal
Edgar Aabye	M	DEN	1900	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold

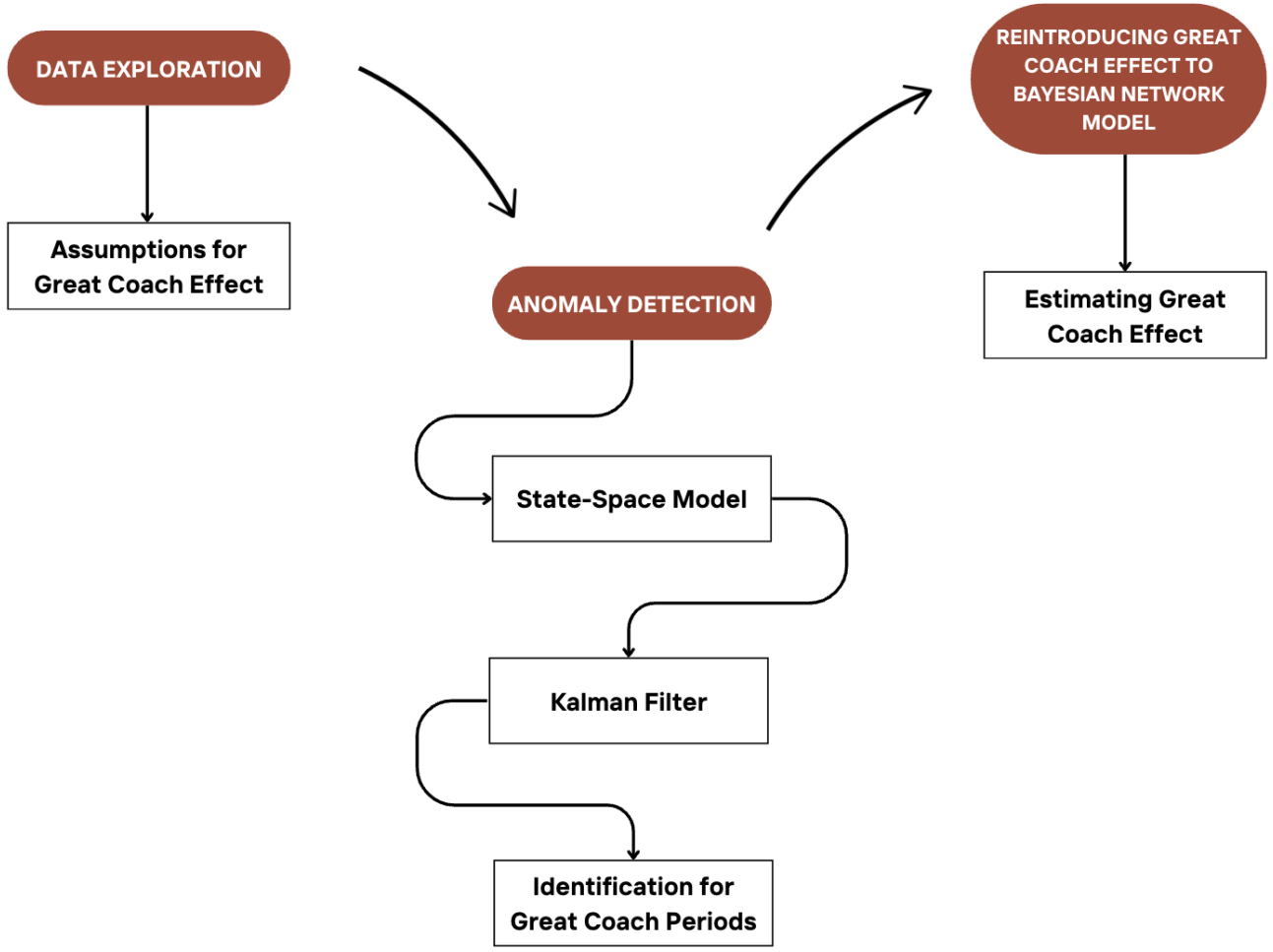
Task 1

- Construct a predictive model for Olympic medal counts.
- Estimate model accuracy and uncertainty. Predict the medal table for the 2028 LA Olympics.



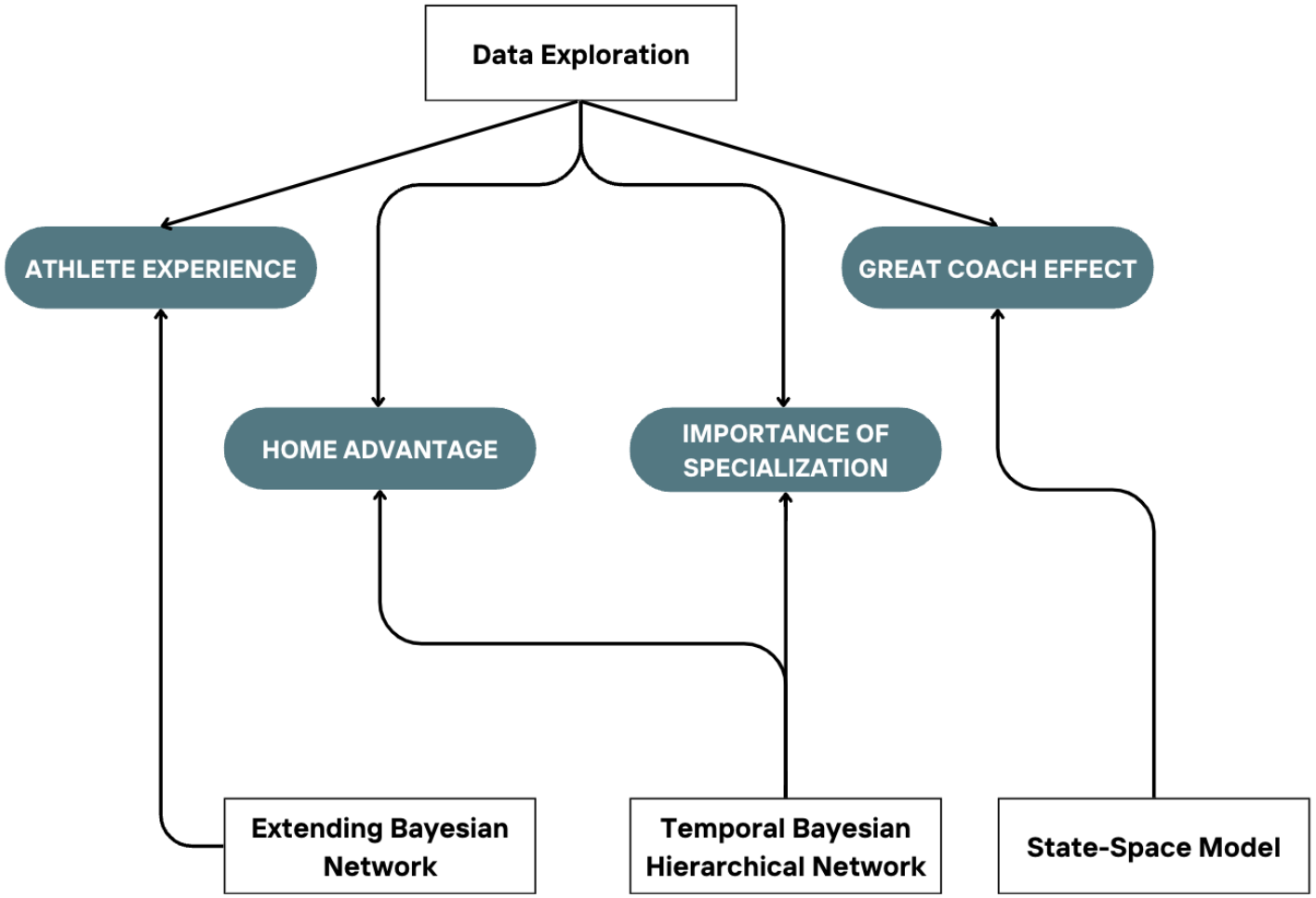
Task 2

- Investigate the potential impact of elite coaches on medal counts.
- Identify three countries and suggest sports where investing in top-tier coaching might alter their Olympic fortunes.



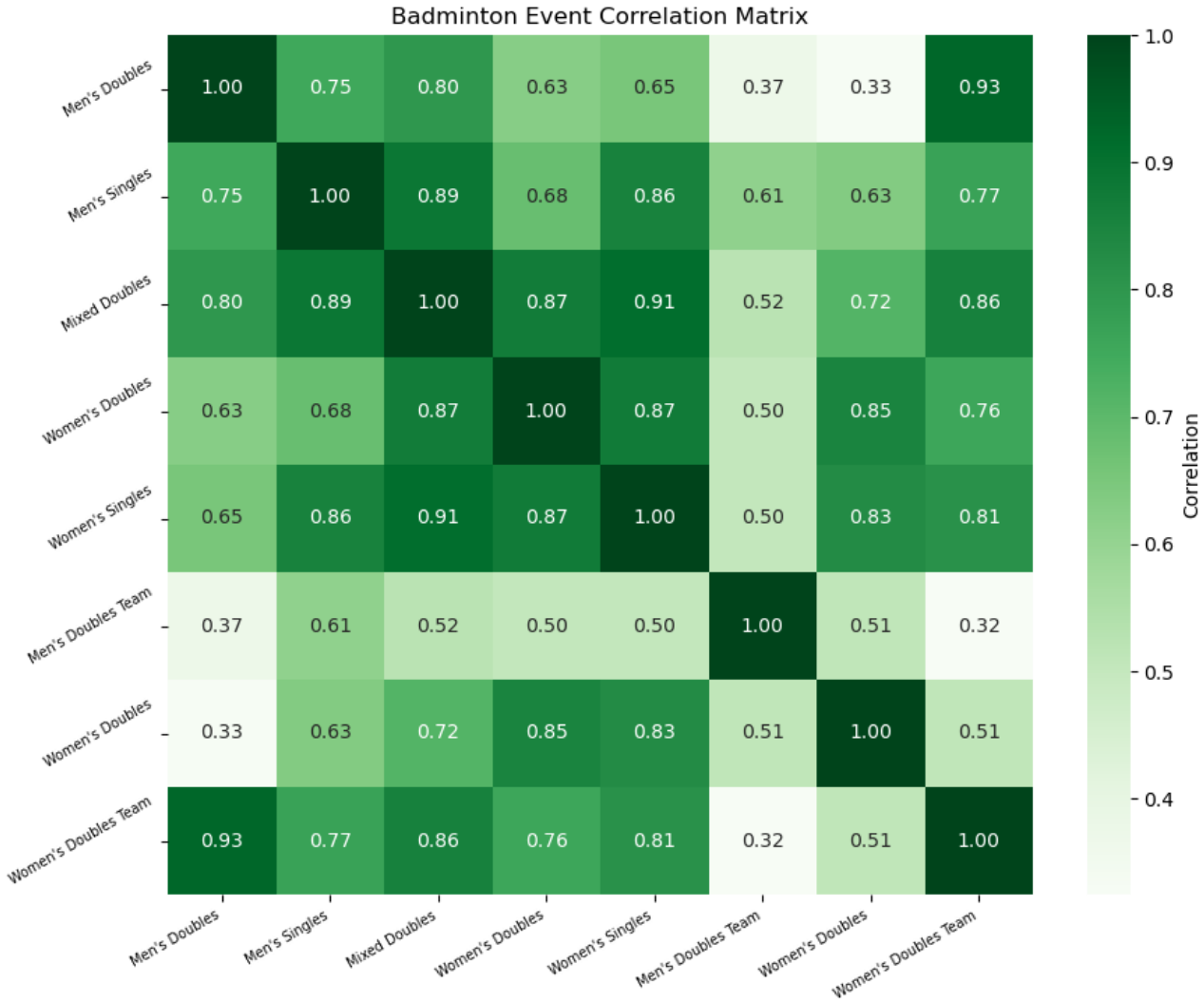
Task 3

- Gain actionable data insights for the Olympic Committees using the developed model.



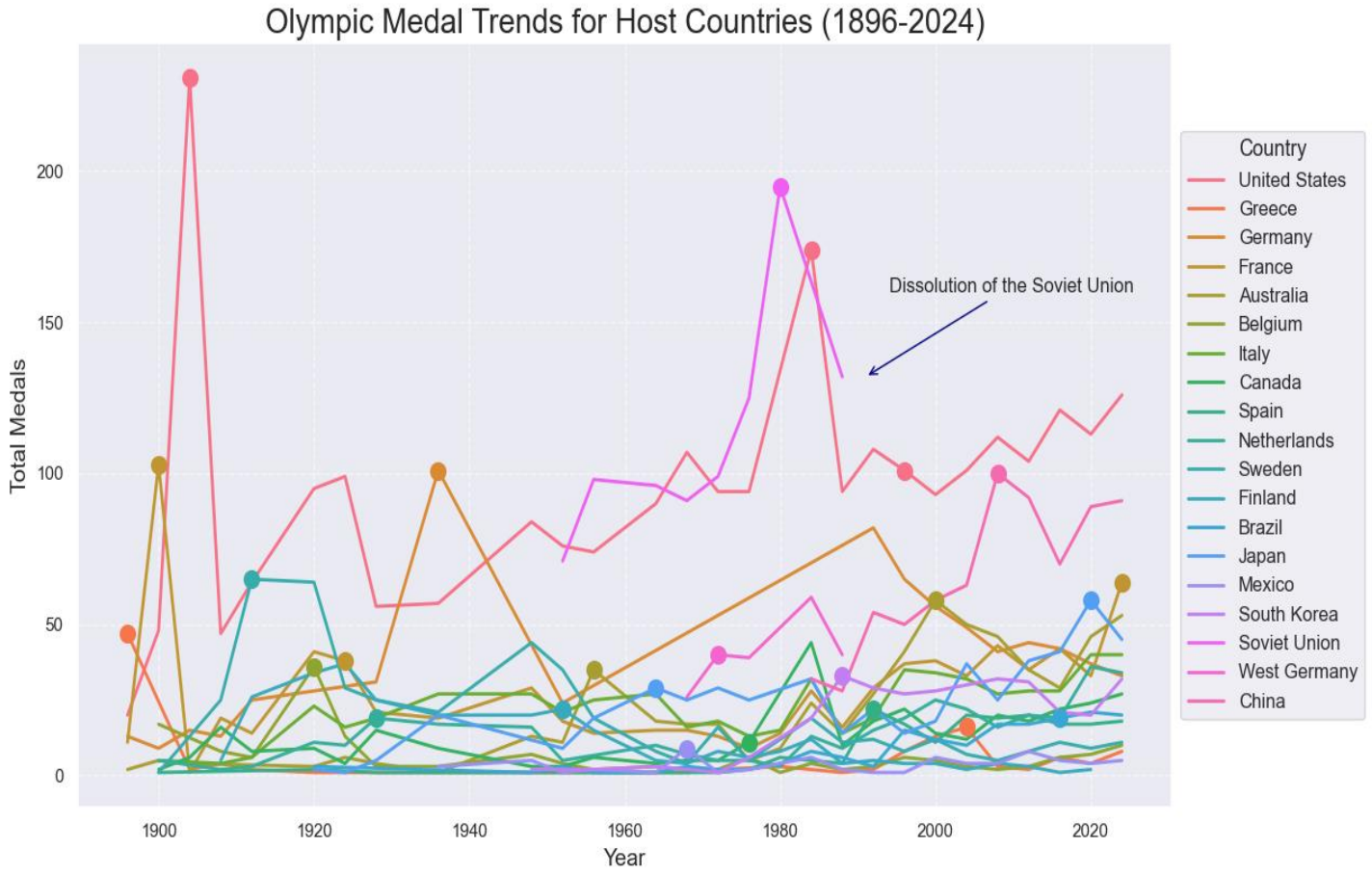
Exploratory Data Analysis: Most Events are Highly Correlated

- Examined correlation between different events within each sport
- Most events within a sport are highly correlated in medal outcomes
- This prompts modeling at the sport level rather than individual events
- For events with weak intra-sport correlation, we treated them as separate sports



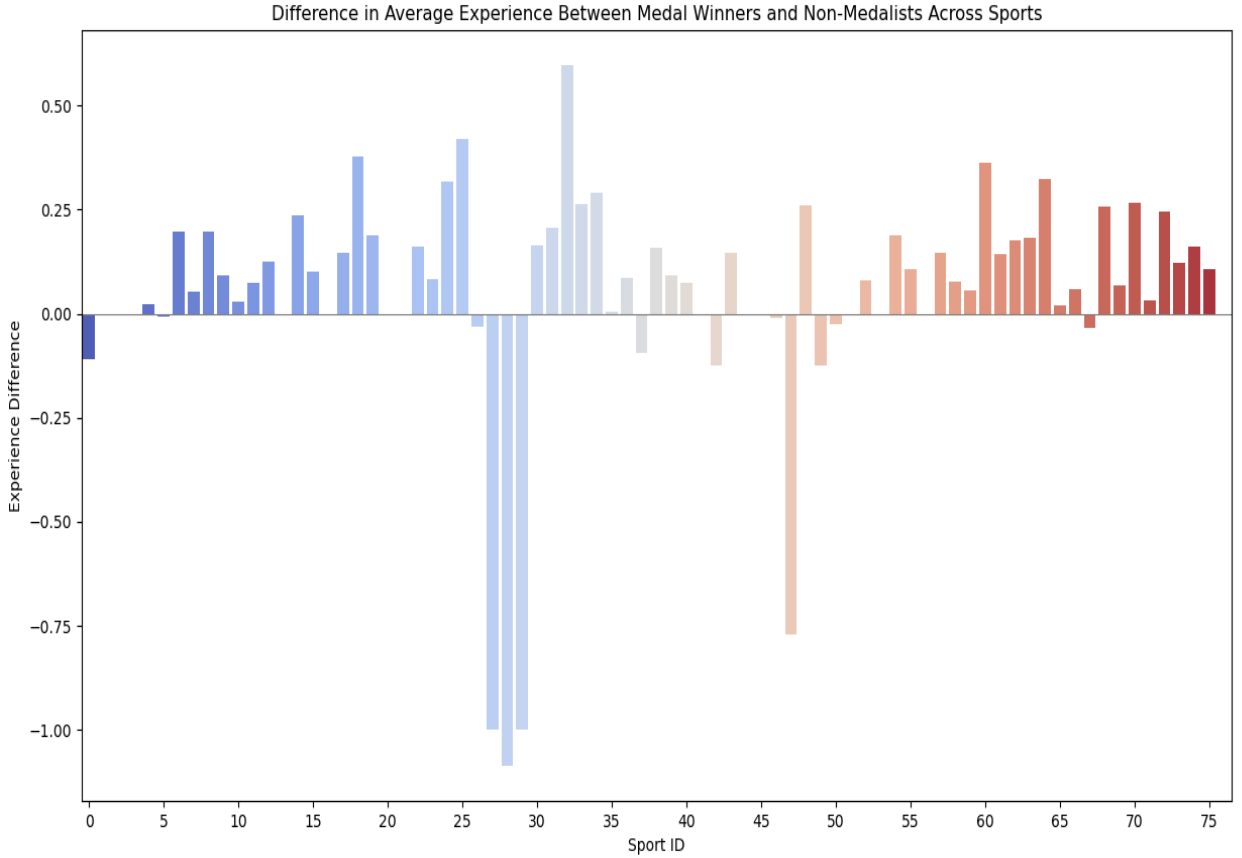
Exploratory Data Analysis: Home Effect is Important for Winning More Medals

- Medal counts of host countries in their respective years of hosting
- Being a host country greatly improves results



Exploratory Data Analysis: Athlete Experience is Important for Some Sports

- The number of times an athlete has participated is available in data
- Athlete experience has different impacts on different sports



Model: Temporal Hierarchical Bayesian Network



Motivation & Inference

- Model uncertainty estimations motivate a probabilistic model
- Random fluctuations can capture structures impossible to infer from observed data
- Possible to run inference using MCMC or Variational Inference methods
- Latent variables were established for a temporal, hierarchical Bayesian network

- Posterior distributions were estimated using Stochastic Variational Inference using the ADAM optimizer
- The model was implemented using the Pyro module in Python
- Evidence Lower Bound (ELBO) was used as the training objective

Features

$$M_{t,c,s} = 4\ n_{\text{Gold},t,c,s} + 2\ n_{\text{Silver},t,c,s} + 1\ n_{\text{Bronze},t,c,s}$$

$$\theta_{t,c,s} = \begin{pmatrix} M_{t,c,s} \\ \#\text{ of Athletes Participating} \\ \#\text{Proportion Of Medals in Sport s to Total Medals Historically} \end{pmatrix}$$

$$H_{t,c} = I(\text{At time t, host country was country c})$$

$$E_{t,c,s} = \frac{1}{\#\text{ Athletes}} \sum_{a \in \text{Athletes}} \#\text{ Participations of athlete a in Olympics}$$

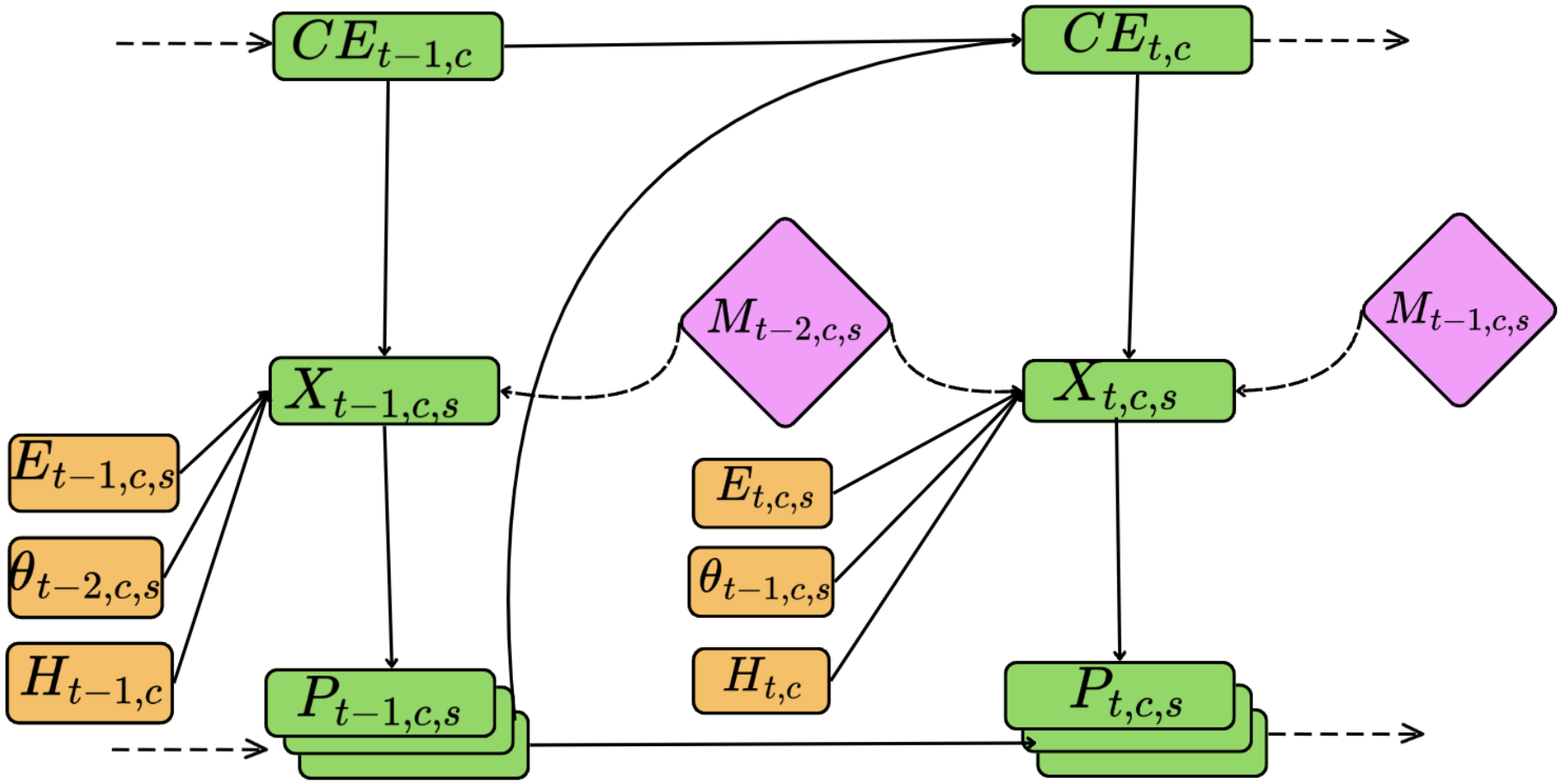
Model Latent Variables

$$P_{t,c,s} = (1-\delta)\,P_{t-1,c,s} + \delta \cdot \exp(X_{t,c,s})$$

$$X_{t,c,s} \sim \mathcal{N}\left(\begin{array}{l} w_{\mathrm{CE}} \cdot CE_{t,c} + w_{\mathrm{mom}} \cdot (M_{t-1,c,s} - M_{t-2,c,s}) + w_{\mathrm{host}} \cdot H_{t,c} \\ +\ w_{\mathrm{xp1}} \cdot k_s^{(1)} \cdot \mathrm{E}_{t,c,s} + w_{\mathrm{xp2}} \cdot k_s^{(2)} \cdot \mathrm{E}^2_{t,c,s} + \boldsymbol{w}_{\mathrm{feat}}^{\top} \boldsymbol{\theta}_{t-1,c,s} \end{array} ,\ \sigma_X^2 \right)$$

$$CE_{t,c} \sim \mathcal{N}\left(\phi \cdot CE_{t-1,c} + \lambda \cdot \sum_s P_{t-1,c,s} \ ,\ \sigma^2_{CE} \right)$$

DAG of Bayesian Network



Conversion to Medal Counts

Plackett-Luce defines a distribution over permutations:

$$\mathbb{P}(\pi) = \prod_{i=1}^{n} \frac{P_{\pi_i}}{\sum_{j=i}^{n} P_{\pi_j}}$$

Powers are translated into medals using a Plackett-Luce distribution per event

$$\text{MedalRankings}_{t,s,e} \sim \text{PlackettLuce}\left(\{P_{t,c,s}\}_{c \in \mathcal{C}_{s,e}}\right)$$

The results of countries per event are then summed up to get medal counts for each country per sport

Based on Latent Variables Country Effect and Power, Monte Carlo Simulation is Performed

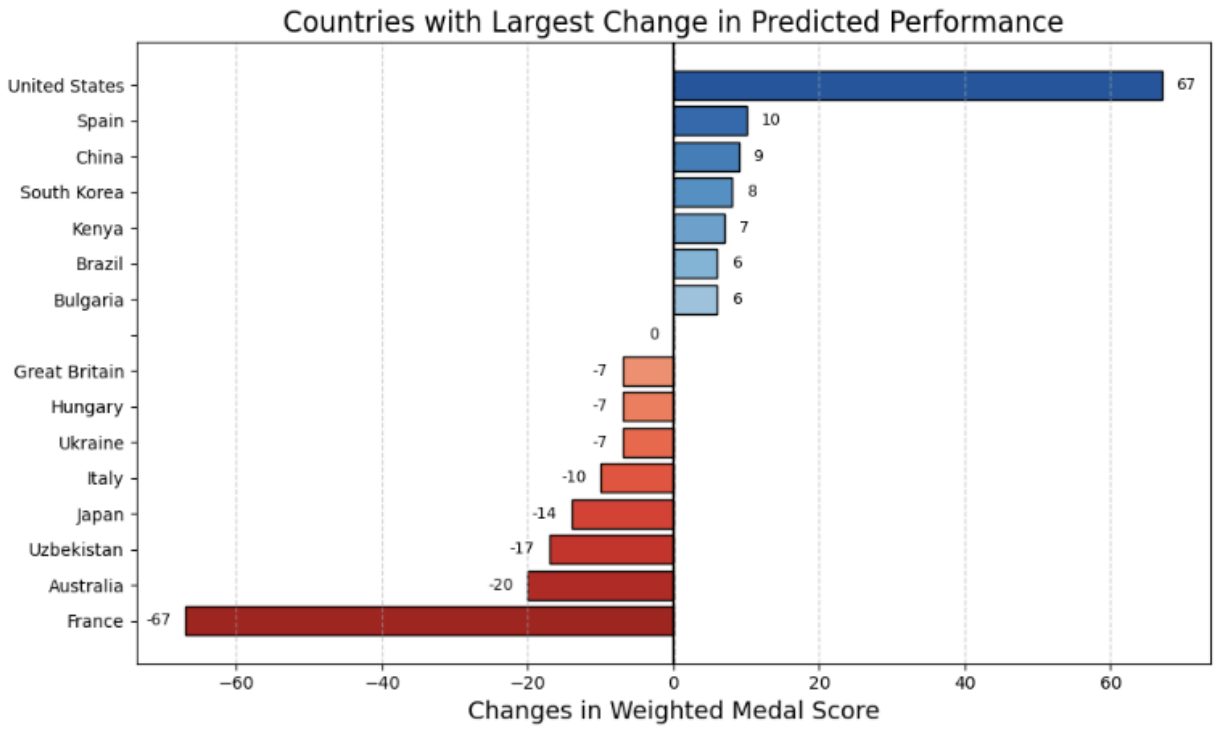
- The prediction listed in the table is the median value with uncertainty calculated by taking the standard deviation of 2000 simulated samples and assuming normal distribution.
- Median ± 1.96σ
- A country may have high uncertainty regarding winning specifically gold or specifically silver medals, but we can say with more certainty whether they will win at least silver medal or not.
- Data from 1896 to 2020 is used for training. We have achieved an accuracy of 72% when predicting 2024 results.

Rank	Country	Gold	Silver	Bronze	Total
1	United States	47±9.57	59±12.42	51±10.51	157±23.71
2	China	44±12.11	22±8.95	27±14.01	93±22.25
3	Japan	16±8.65	10±6.18	19±7.31	45±11.56
4	South Korea	15±8.89	9±4.20	10±5.51	34±8.75
5	Great Britain	14 ± 6.19	20±7.20	26±10.25	60±11.86
6	Australia	13 ± 5.11	18±7.05	18±8.25	49±10.12
7	Netherlands	13±6.90	9±5.20	12±5.10	34±11.87
8	Germany	12±6.81	11 ± 5.88	9 ±4.20	32± 12.99
9	Italy	10±4.00	13±6.15	13±5.17	46±11.24
10	Canada	9 ± 5.10	7 ± 3.50	13±6.11	29 ± 10.25
11	New Zealand	9± 4.29	7 ± 4.26	4 ± 3.26	20 ± 8.10
12	France	8 ± 5.12	12 ± 8.91	15 ± 7.53	35 ± 11.56
13	Spain	7 ± 4.21	5 ± 2.98	9 ± 4.40	21 ± 7.18
14	Hungary	5 ± 3.01	7 ± 3.78	3 ± 3.36	15 ± 5.82
15	Brazil	5 ± 3.87	5 ± 3.22	12 ± 6.31	22 ± 6.94
16	Kenya	5 ± 3.98	4 ± 4.02	4 ± 4.21	13 ± 5.30
17	Sweden	4 ± 1.98	3 ± 1.65	3 ±1.82	10 ±3.46
18	Belgium	4 ± 2.50	2 ± 1.92	3 ± 1.87	9 ± 5.87
19	Bulgaria	4 ± 2.10	2 ± 1.02	3 ± 1.67	9 ±4.94
20	Norway	4 ± 2.89	1 ± 2.41	2 ±3.96	7 ±4.99

Countries Experience Variation in Predicted Performance

- US and France are expected to experience largest improvement and decline respectively due to host advantage effect.

Country	Gold	Silver	Bronze	Total
Chad	0 ± 0.90	1 ± 1.23	1 ± 1.20	2 ± 1.88
Benin	0 ± 0.92	0 ± 0.78	1 ± 1.19	1 ± 1.02
Antigua and Barbuda	0 ± 0.89	0 ± 0.78	1 ± 1.32	1 ± 1.11



Posterior Distributions of Coefficients

Parameter	Estimate	SD of Estimator	95% Confidence Interval	p-value
ϕ	0.413	0.134	[0.132, 0.618]	0.002
λ	0.653	0.216	[0.230, 1.076]	0.003
w_{CE}	0.102	0.349	[−0.582, 0.786]	0.772
w_{mom}	0.076	0.120	[−0.159, 0.311]	0.521
w_{host}	0.476	0.181	[0.121, 0.831]	0.009
e_1	0.052	0.077	[−0.099, 0.203]	0.504
e_2	-0.0022	0.0074	[−0.0167, 0.0123]	0.773
$w_{\text{feat},1}$	0.672	0.185	[0.310, 1.034]	0.0002
$w_{\text{feat},2}$	0.065	0.053	[−0.039, 0.169]	0.220
$w_{\text{feat},3}$	0.032	0.045	[−0.056, 0.120]	0.469

Understanding “Great Coach” Effect Using Kalman Filter



The "Great Coach" Effect

- No data on coaches, assuming great coaches are anomalies in the country performance specific to a sport
- We define the rate ratio as a time series per country and per sport to keep track

$$\text{RateRatio}_{c,s,t} = \frac{P_{c,s,t} - P_{c,s,t-1}}{\frac{1}{|S|}\sum_{i \in \text{sports}}(P_{c,i,t} - P_{c,i,t-1})}$$

- Then we assume these values generated by a random walk process to create a 1D state-space model

$$\text{RateRatio}_t = \mu_t + \nu_t, \quad \nu_t \sim \mathcal{N}(0, \sigma^2_{\text{obs}})$$

$$\mu_t = \mu_{t-1} + \omega_t, \quad \omega_t \sim \mathcal{N}(0, \sigma^2_{\text{process}})$$

Kalman Filter

Prediction

$$\hat{\mu}_{t|t-1} = \hat{\mu}_{t-1|t-1}$$
$$\sigma^2_{t|t-1} = \sigma^2_{t-1|t-1} + \sigma^2_{\text{process}}$$

Update (Adaptive Noise)

$$K_t = \frac{\sigma^2_{t|t-1}}{\sigma^2_{t|t-1} + \sigma^2_{\text{obs}}}$$

$$\hat{\mu}_{t|t} = \hat{\mu}_{t|t-1} + K_t(\text{RateRatio}_t - \hat{\mu}_{t|t-1})$$

$$\sigma^2_{\text{obs},t} = \lambda \cdot \sigma^2_{\text{obs},t-1} + (1-\lambda) \cdot \text{residual}_t^2$$
$$\sigma^2_{t|t} = (1-K_t)\sigma^2_{t|t-1}$$

Normalization

$$\text{residual}_t = \text{RateRatio}_t - \hat{\mu}_{t|t-1}$$
$$z_t = \frac{\text{residual}_t}{\sqrt{\sigma^2_{t|t-1} + \sigma^2_{\text{obs}}}}$$

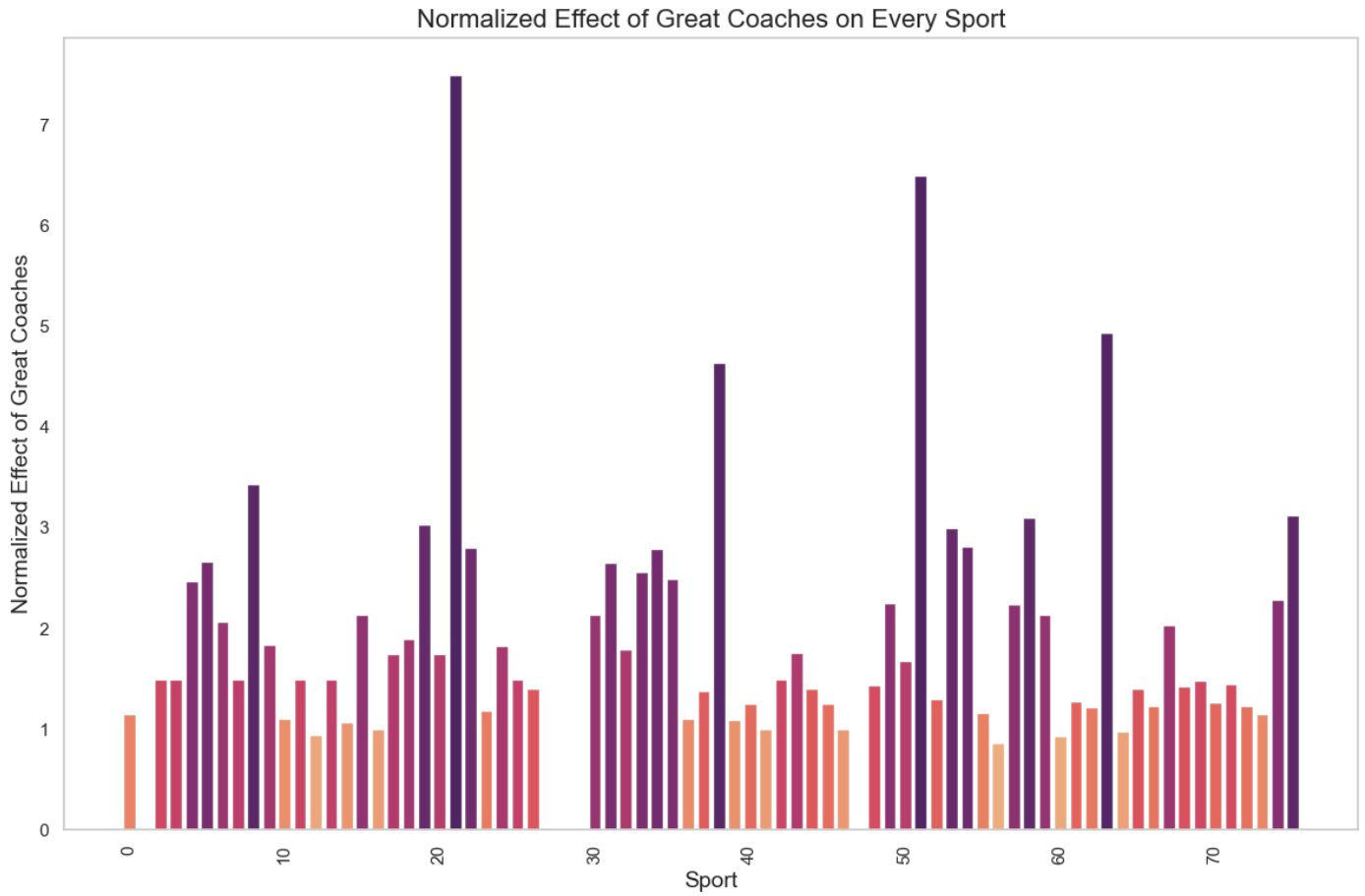
Detection

Positive anomaly: $z_t >$ threshold
Negative anomaly: $z_t <$ -threshold
or successive small values

Identify great coach periods as periods between positive and negative anomalies

Insights on Great Coach Effect

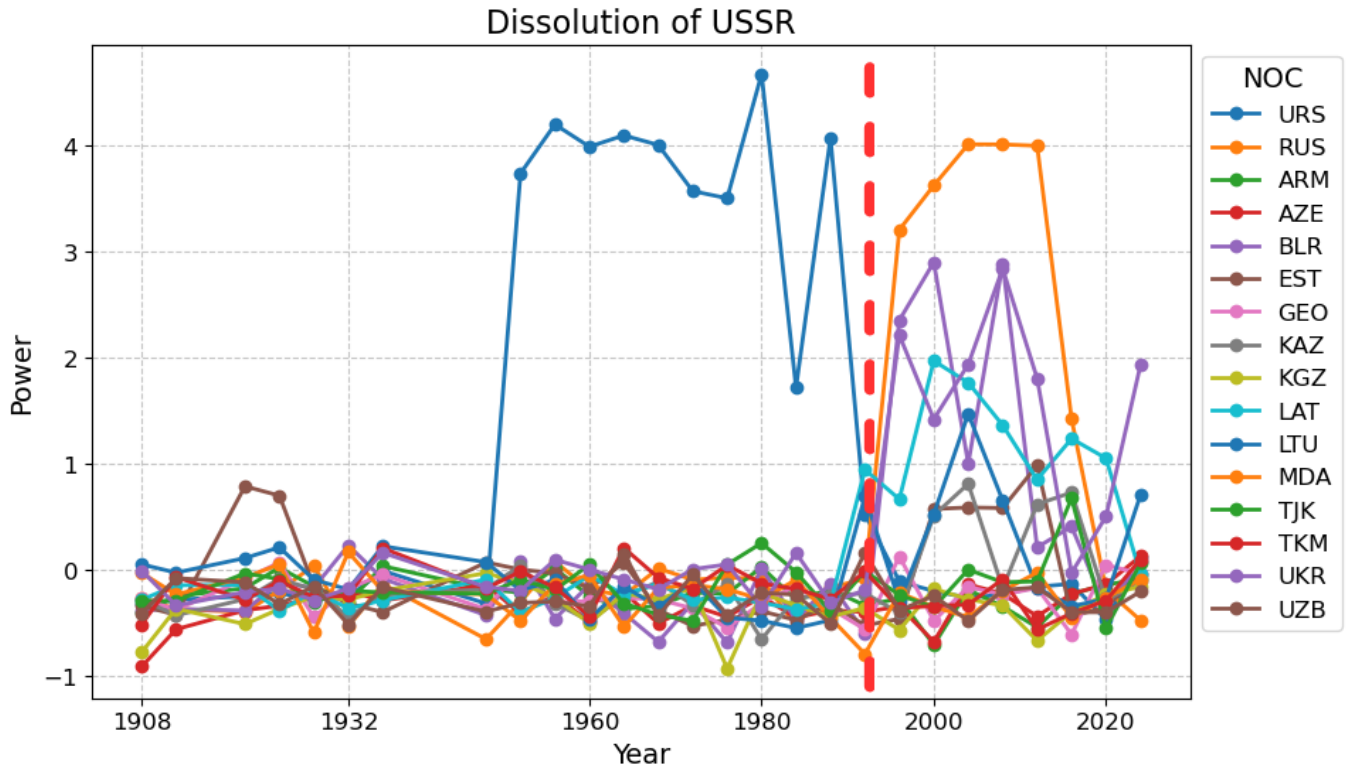
- The model identifies 332 total great coach periods, meaning an approximate average of 10.71 great coaches being there per year.
- The Great Coaches were added as another feature in the Bayesian Network and results were obtained
- Great Coaches are very effective but have high p-values



Strength & Weaknesses

- ✓ Captures both permanent and transitory effect
- ✓ Able to handle unification and dissolution

- X Change in the Number of Events
- X Neglecting economic, cultural, and social influences
- X Keeping parameters with low p-valued posteriors needed to be empirically tweaked.



Thank You!

Any Questions?

