Political bias and diversity in recommendation algorithms' representations

Tim Faverjon

SciencesPo
MÉDIALAB

LEARNING PLANET INSTITUTE

State of the art

- Recommender systems predict new *interactions* from past *interactions*

Research suggests :

- *Interactions* are related to political attitudes
- Recommendations can be politically biased
- Recommendations can have impact on polarization, diversity…

Regulation (DSA) :

- Ask to reveal which features are important for recommendations
- Ask to not discriminate users by political opinion (for ads)

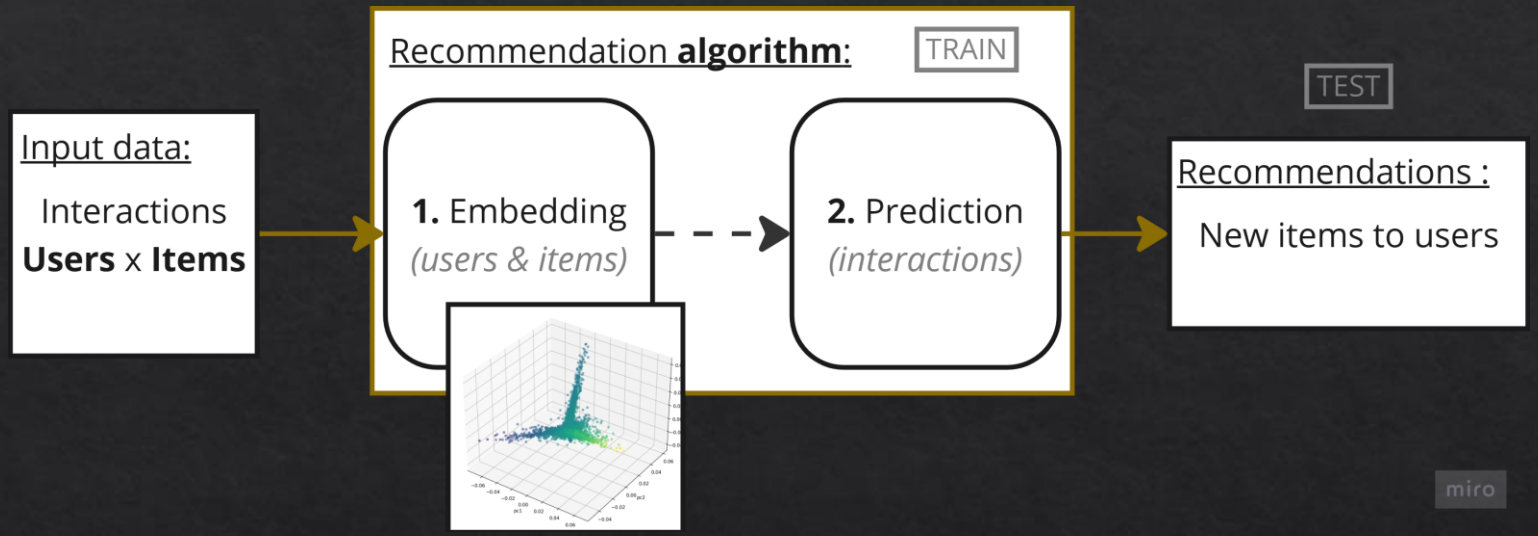
Research Questions

- **Q1.a**: What political information is captured by the models of recommendation algorithms?
- **Q1.b**: What is the impact of learned political information on the recommendations received by users?

- **Q2**: How would recommendations of political content change if we removed political information from the models?

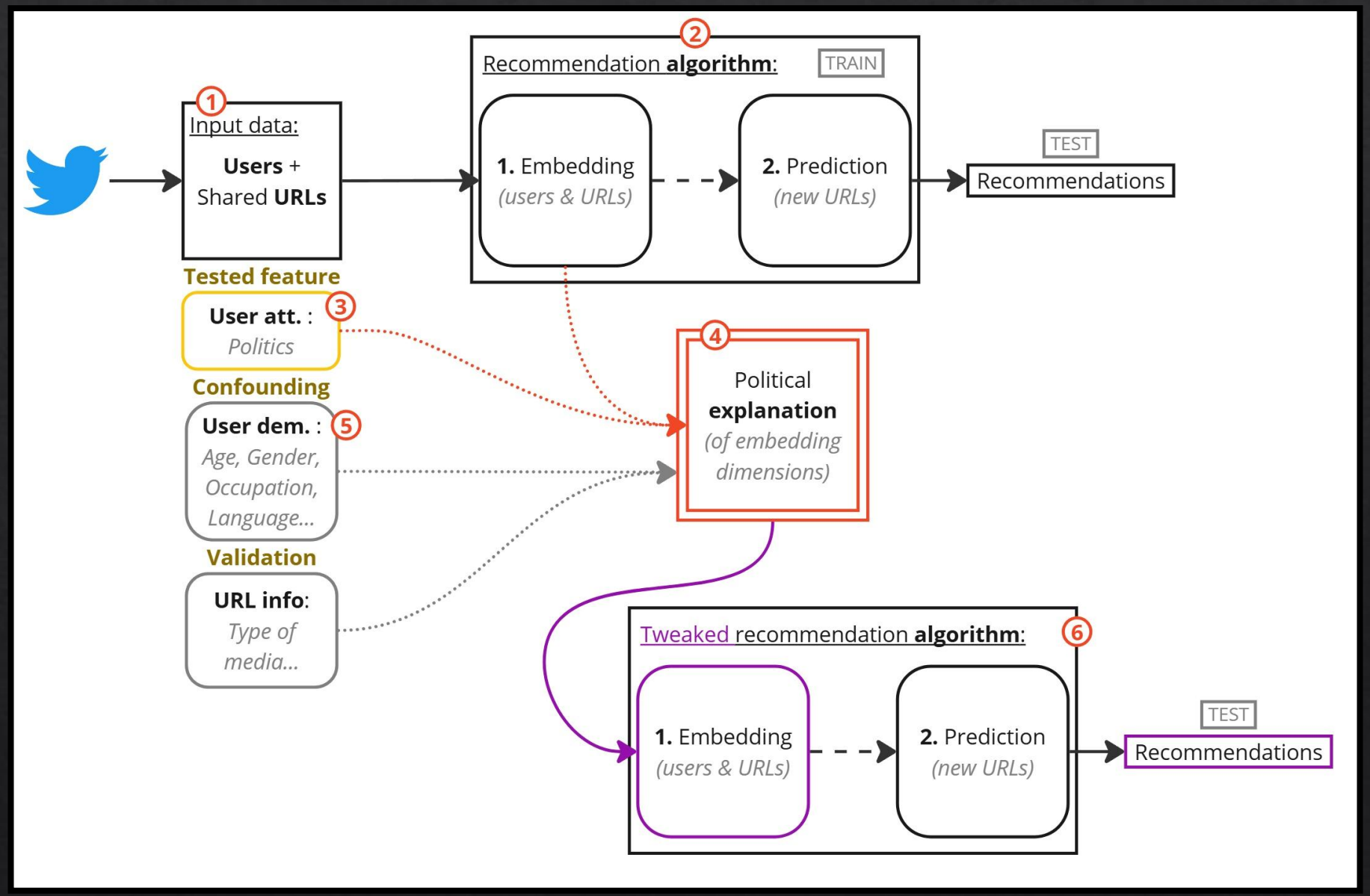
Methods and Tools

- **Observation :** Recommendation systems usually contain embedding layers



- **Hypothesis :** Some dimensions of the embedding will specialize on political attitudes
- **Tool :** We can use *hidden semantic explanation* methods
  - Identify dimensions learning politic
  - Observe the result of embedding manipulation

Method : a case study on Twitter



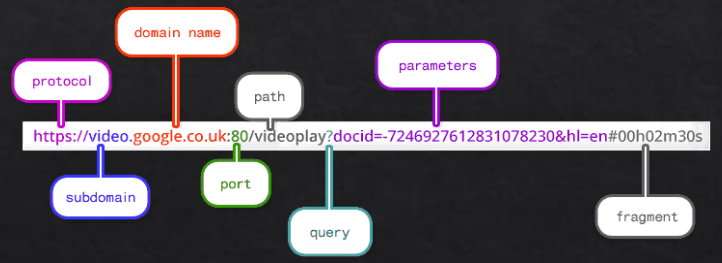
1. Data collection

Users : ***Twitter users***

- Active[1] users
- following French MPs[2]
- $\approx 10^4$: randomly selected

Items : ***URLs***

- Shared on Twitter
- $\approx 10^5$: shared by at least 10 users



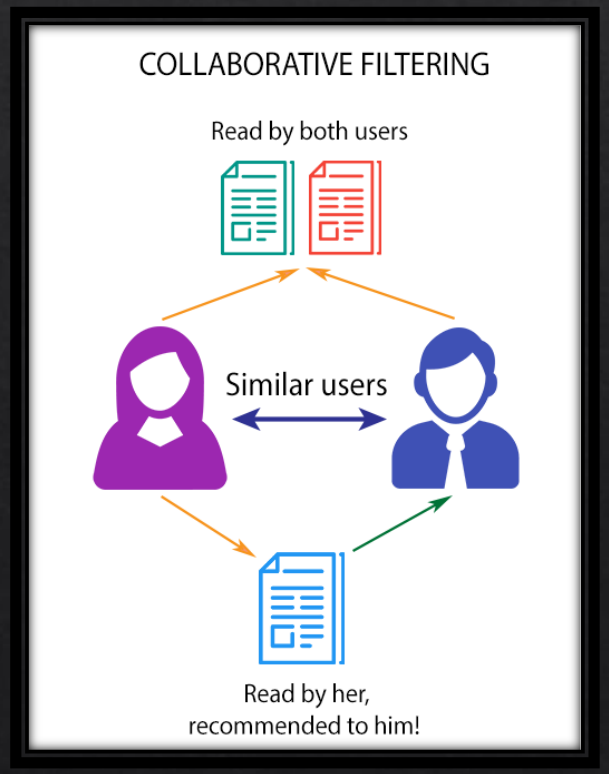
[1] : Sharing links, following at least 3 MPs and 25 users and with at least 25 followers (to avoid bots)
[2] : Members of Parliament

2. Recommendation algorithm

2.1. Representation Space :

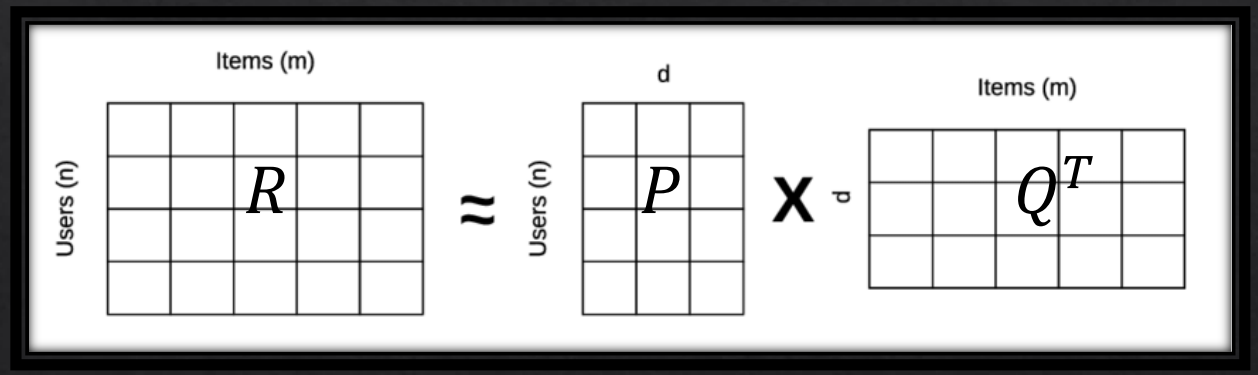
From : interaction data - *To* : vector space

The idea :



The method :

Non-negative Matrix Factorization



The Loss-function :

$$\mathcal{L}(P,Q) = \frac{1}{2} \cdot \|R - PQ^T\|_2^2$$

$$+ Regularisation \ldots$$

2. Recommendation algorithm

2.2. Prediction :

From : vector space - *To* : prediction of new Items for each User

The predition :

- Predicted rating for user u and item i is a **scalar product** : $\widehat{r_{ui}} = P_u \cdot Q_i$
- We recommend the best rated new items

The accuracy metric : Hits@10

The results : Hits@10 = 0.34

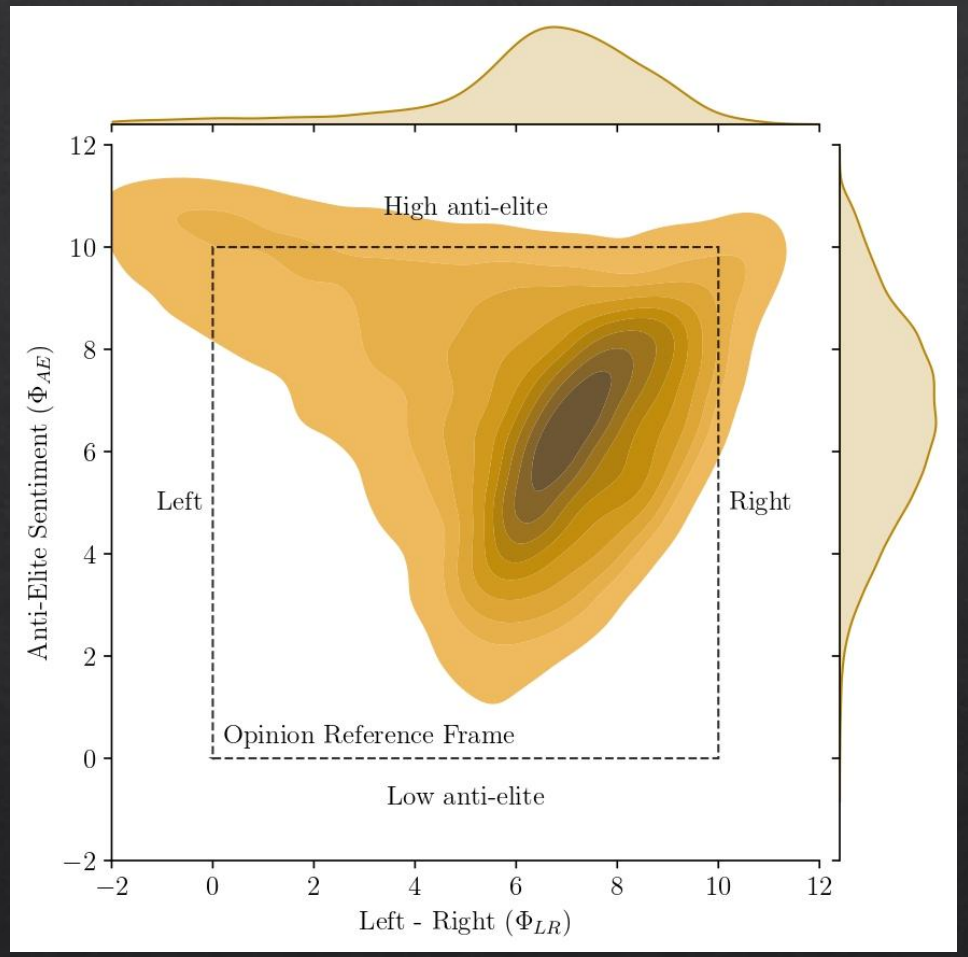
3. Users political attitudes

- Spatialisation **following network of French MPs**
- Scaling from ideology to attitudes based on CHES survey

<u>France - 2 most explicative dimensions</u>

Left – Right : classical left-right

Anti-Elite Sentiment : negative attitude toward élites and insitutions

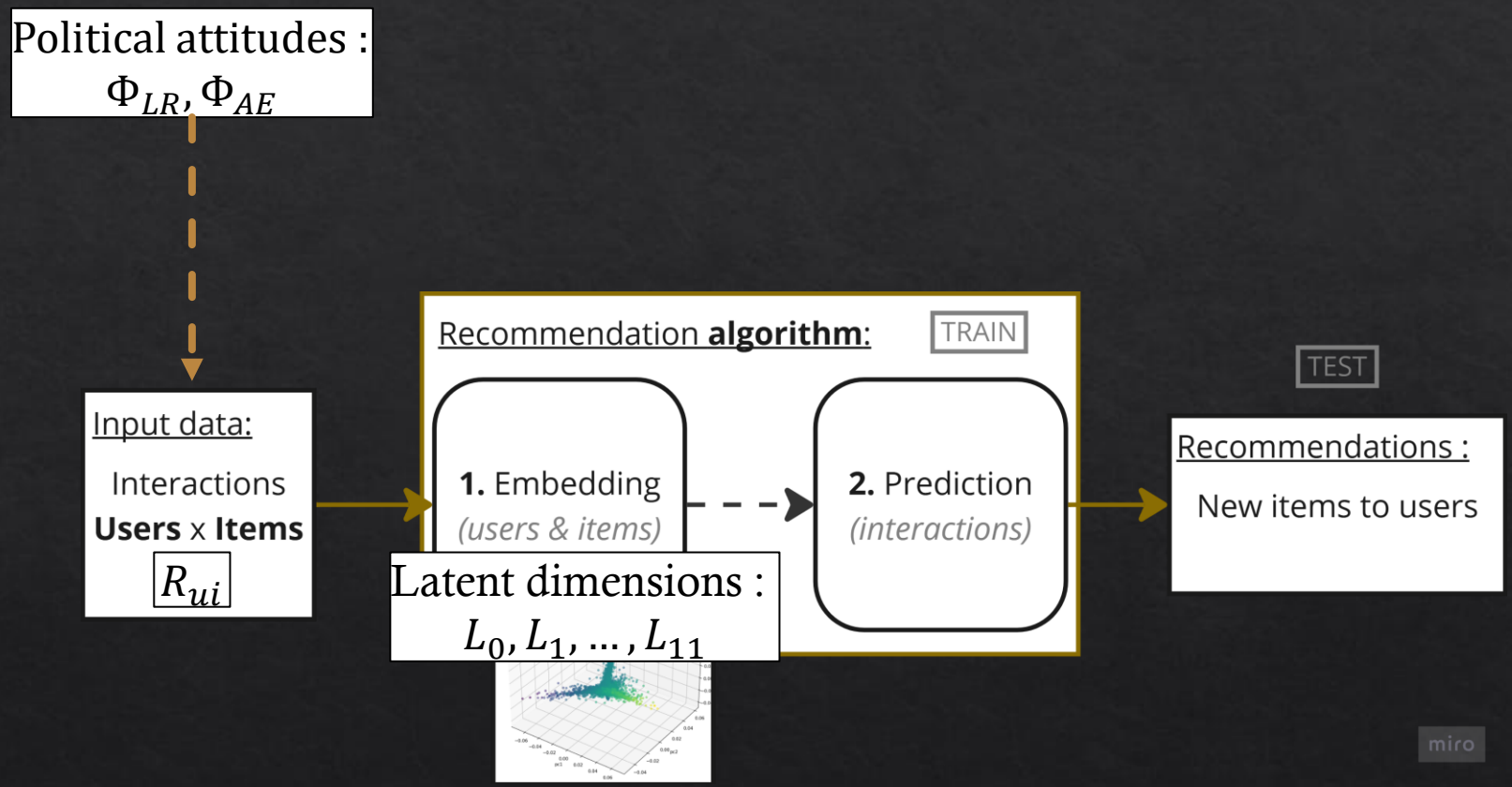


4. Explanation

- Φ_{LR} : Left - Right attitude
- Φ_{AE} : Anti-Elite sentiment

- $L_0, \ldots, L_{11}$: Position in the latent dimensions of the embedding

Statistical relation between Φ_{LR}, Φ_{AE} and $L_0, L_1, \ldots, L_{11}$?

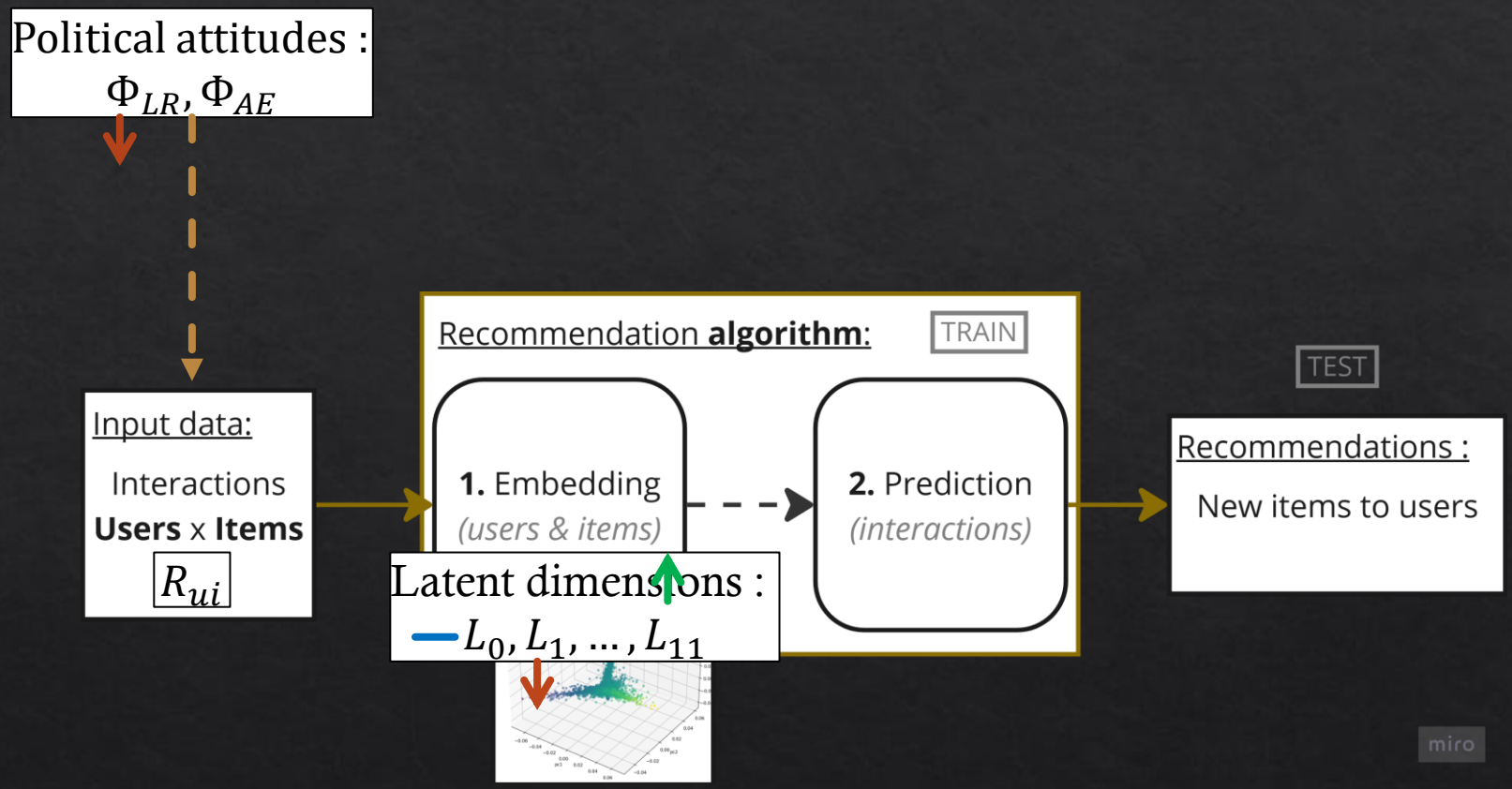


4. Explanation: Attribution

- Φ_{LR} : Left - Right attitude
- Φ_{AE} : Anti-Elite sentiment

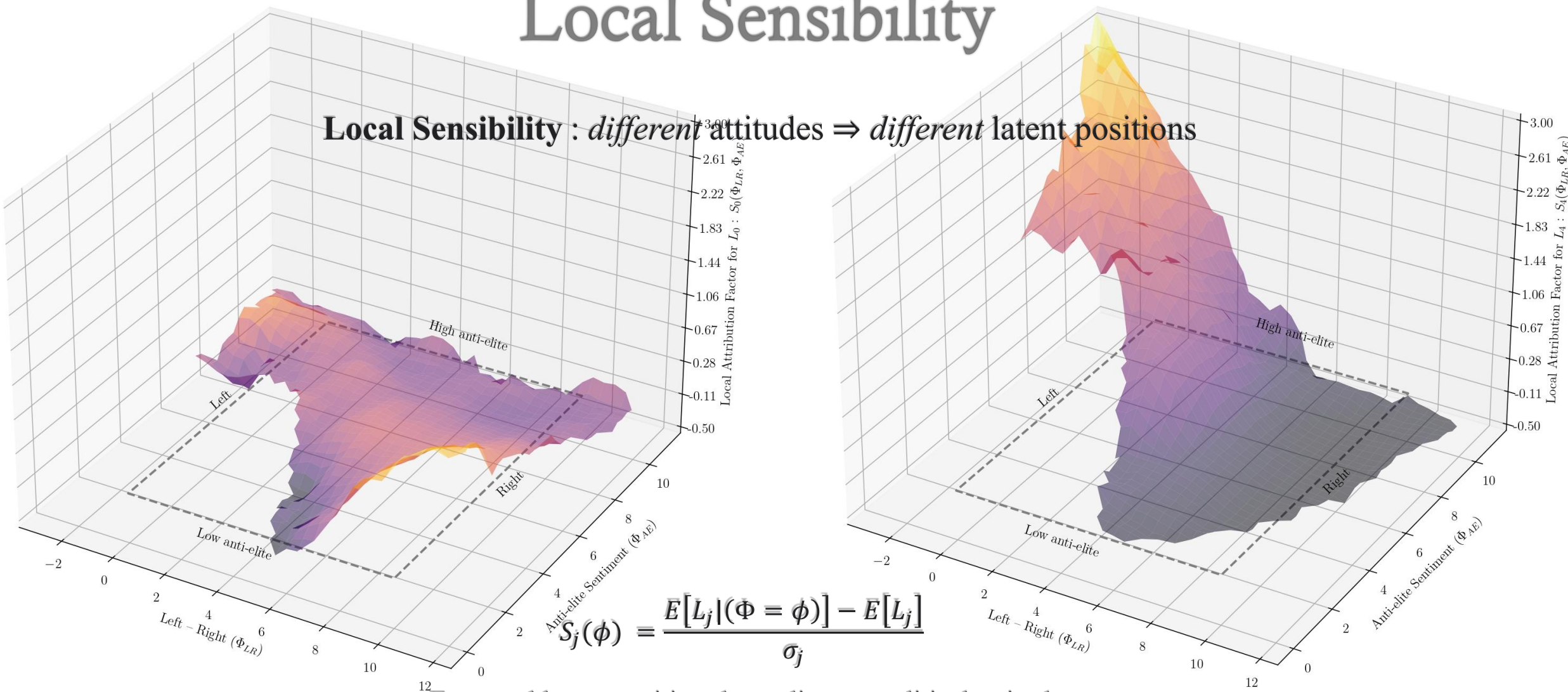
- $L_0, \ldots, L_{11}$: Position in the latent dimensions of the embedding

Statistical relation between Φ_{LR}, Φ_{AE} and $L_0, L_1, \ldots, L_{11}$?



Local Sensibility

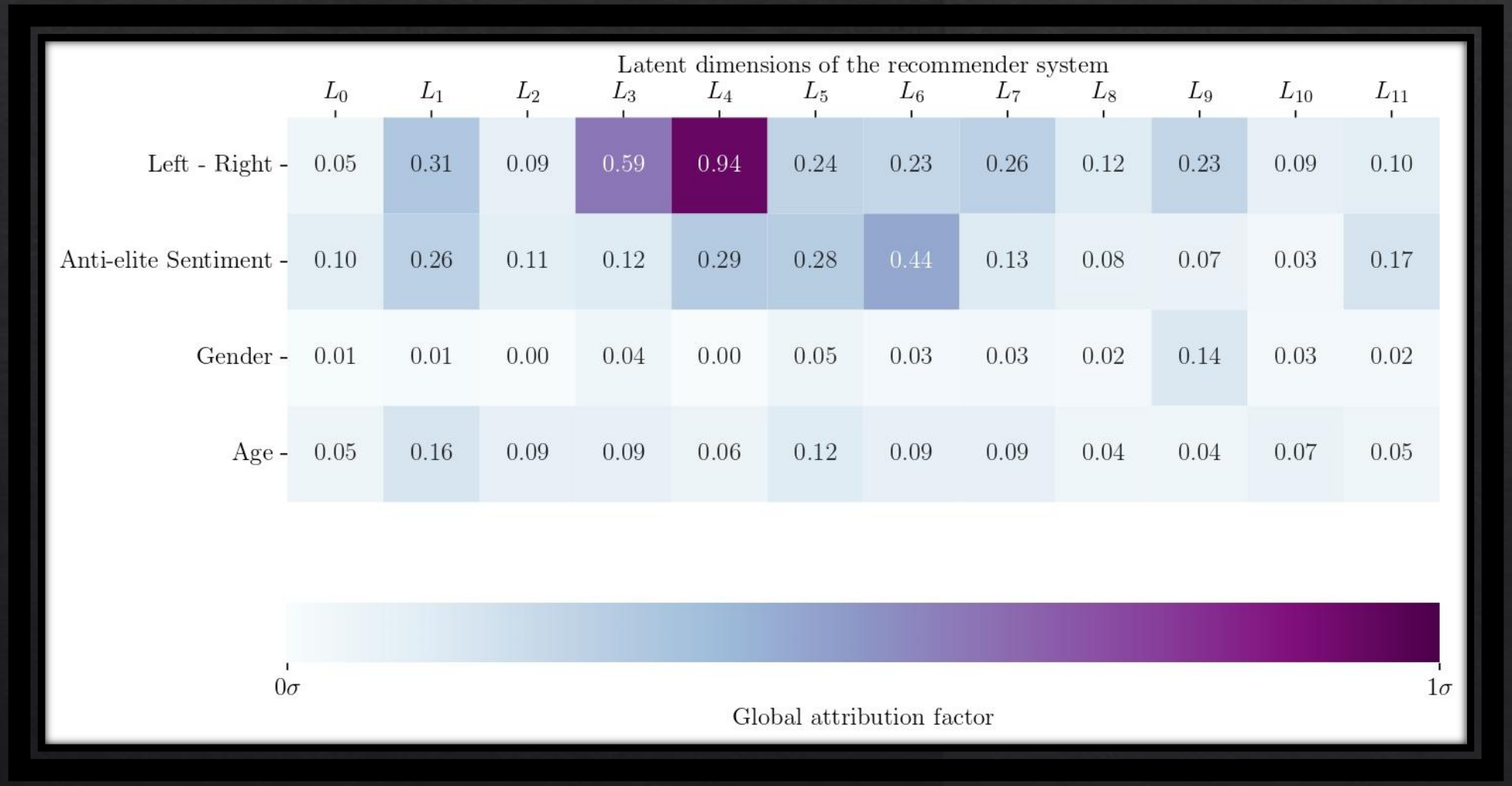
Local Sensibility : *different* attitudes ⇒ *different* latent positions



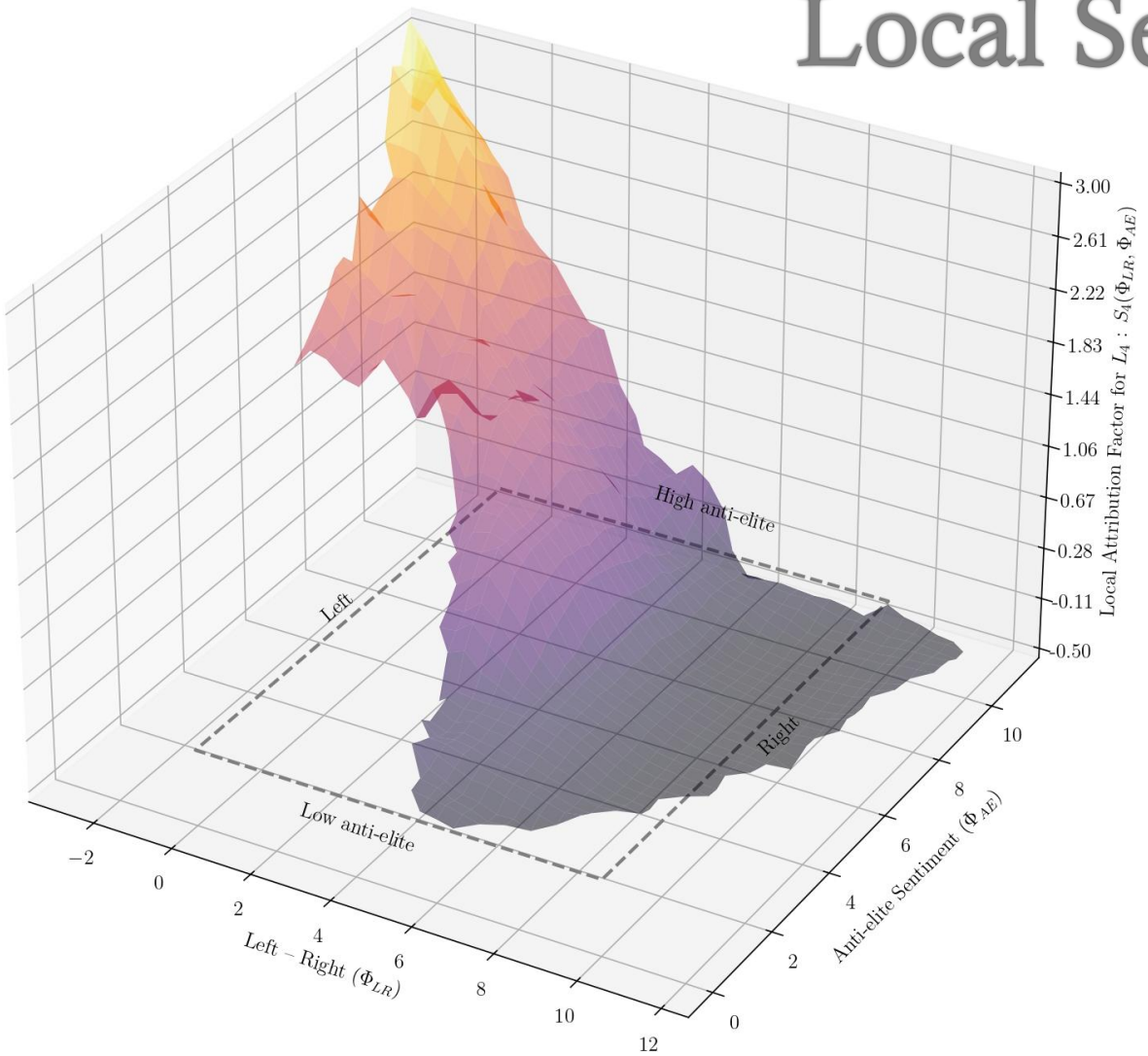
$$S_j(\phi) = \frac{E[L_j|(\Phi = \phi)] - E[L_j]}{\sigma_j}$$

Expected latent position depending on political attitude

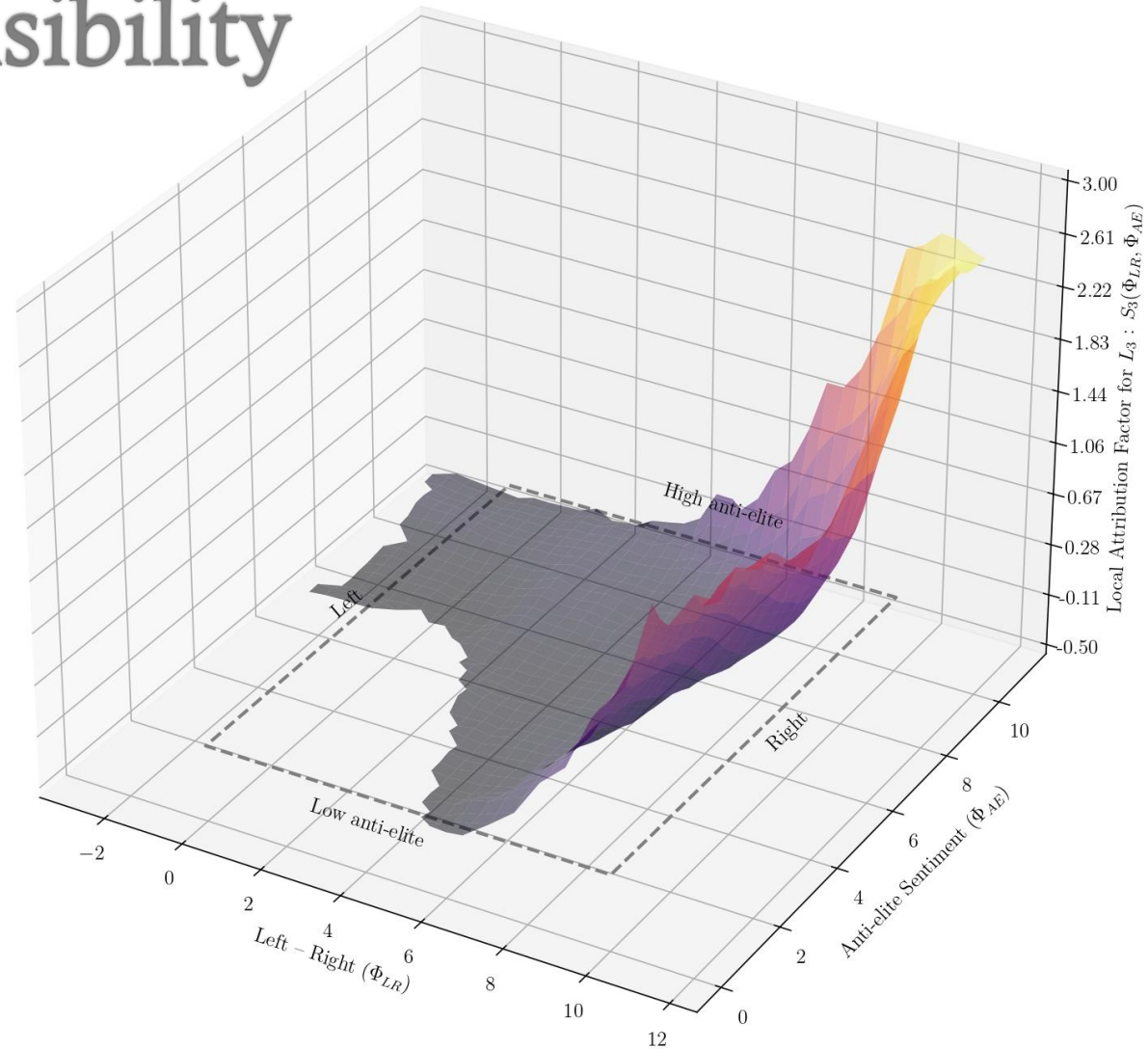
Global sensibility : average sensibility



Local Sensibility



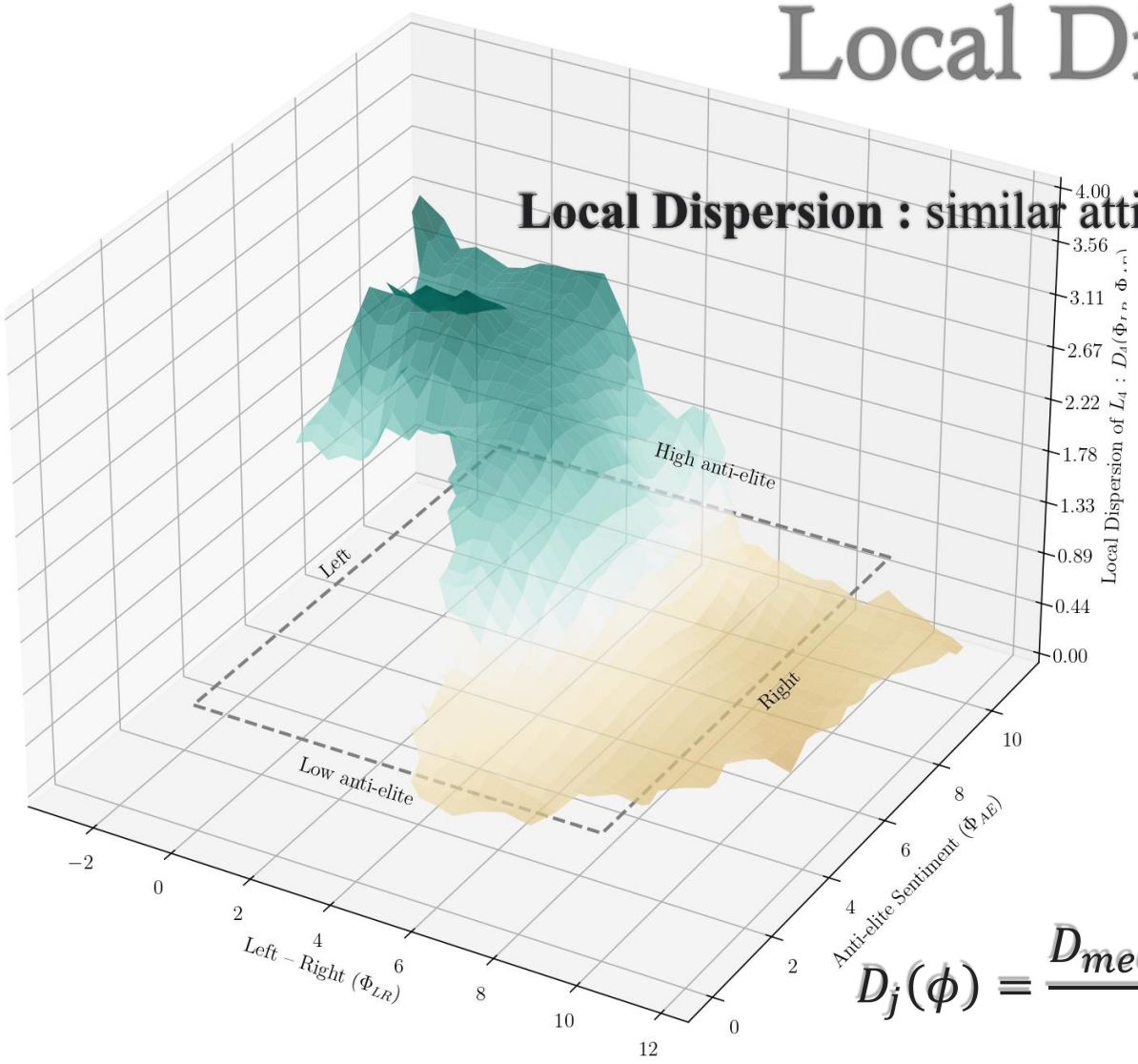
$L_4 : S_{4,\Phi_{LR}} = 0.94\sigma$

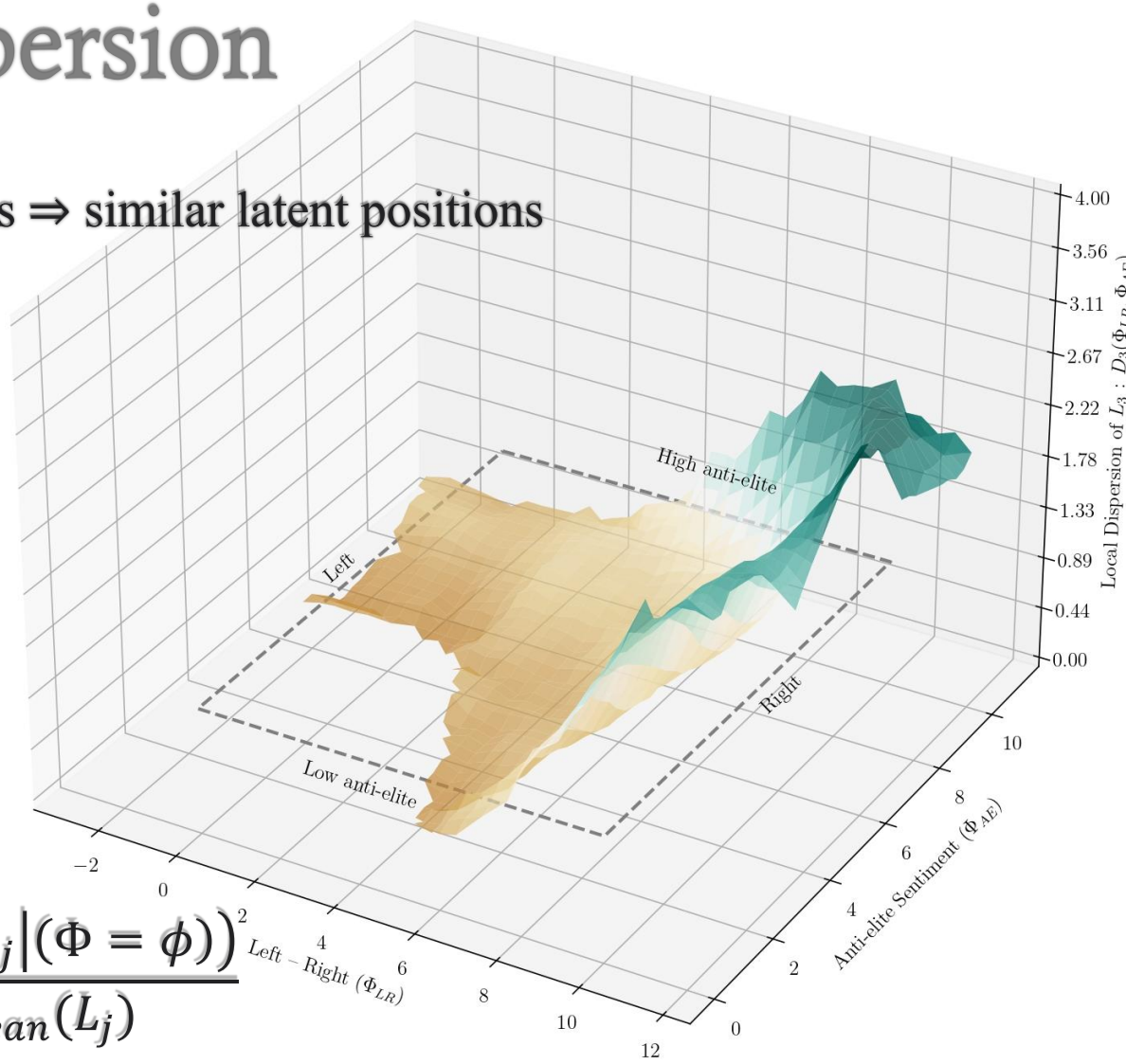


$L_3 : S_{3,\Phi_{LR}} = 0.59\sigma$

Local Dispersion

Local Dispersion : similar attitudes ⇒ similar latent positions





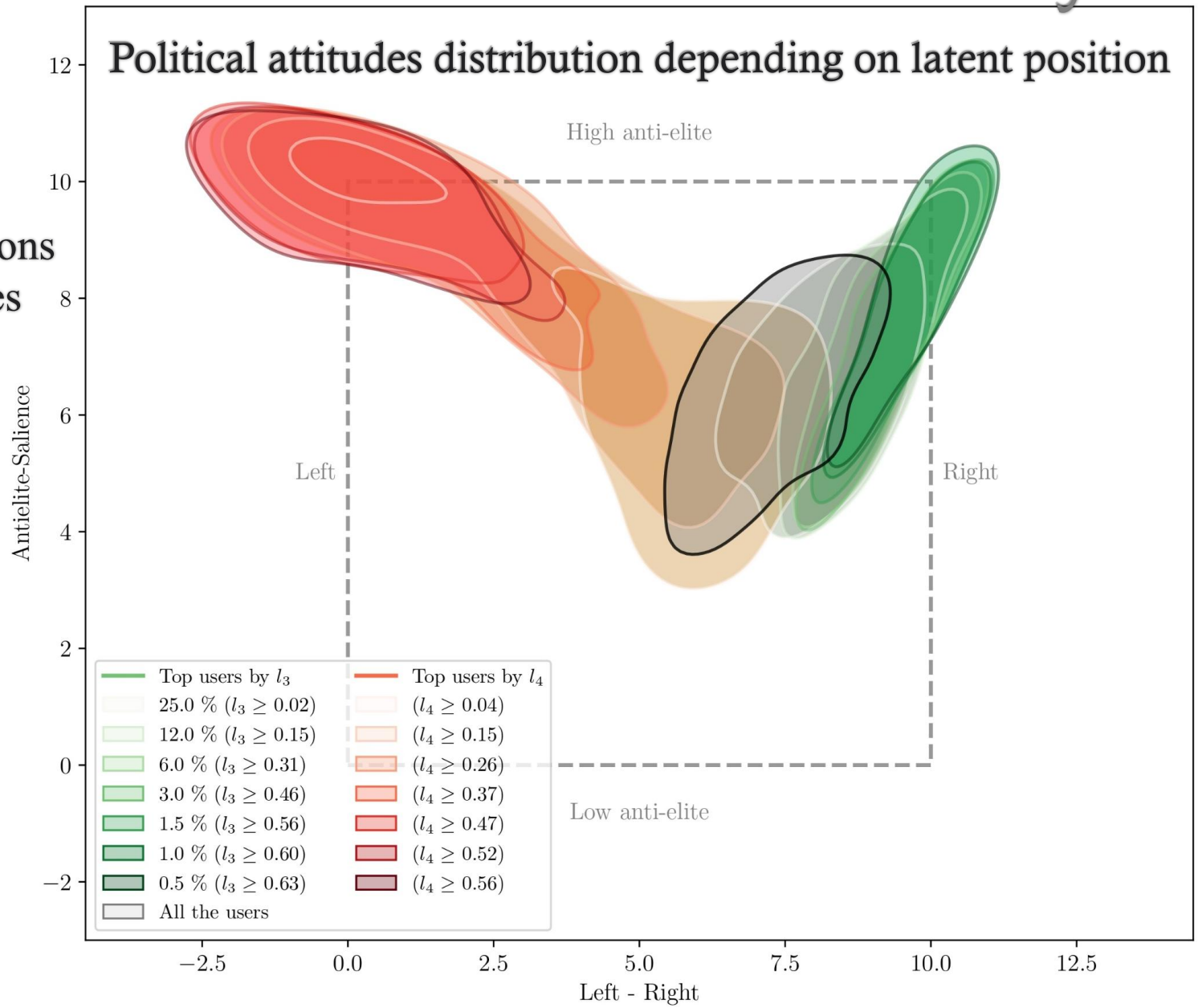
$$D_j(\phi) = \frac{D_{mean}\left(L_j \middle| (\Phi = \phi)\right)^2}{D_{mean}(L_j)}$$

Distance to mean latent position depending on political attitude

Latent Bias and Diversity

Latent bias :

different latent positions
⇒ different attitudes



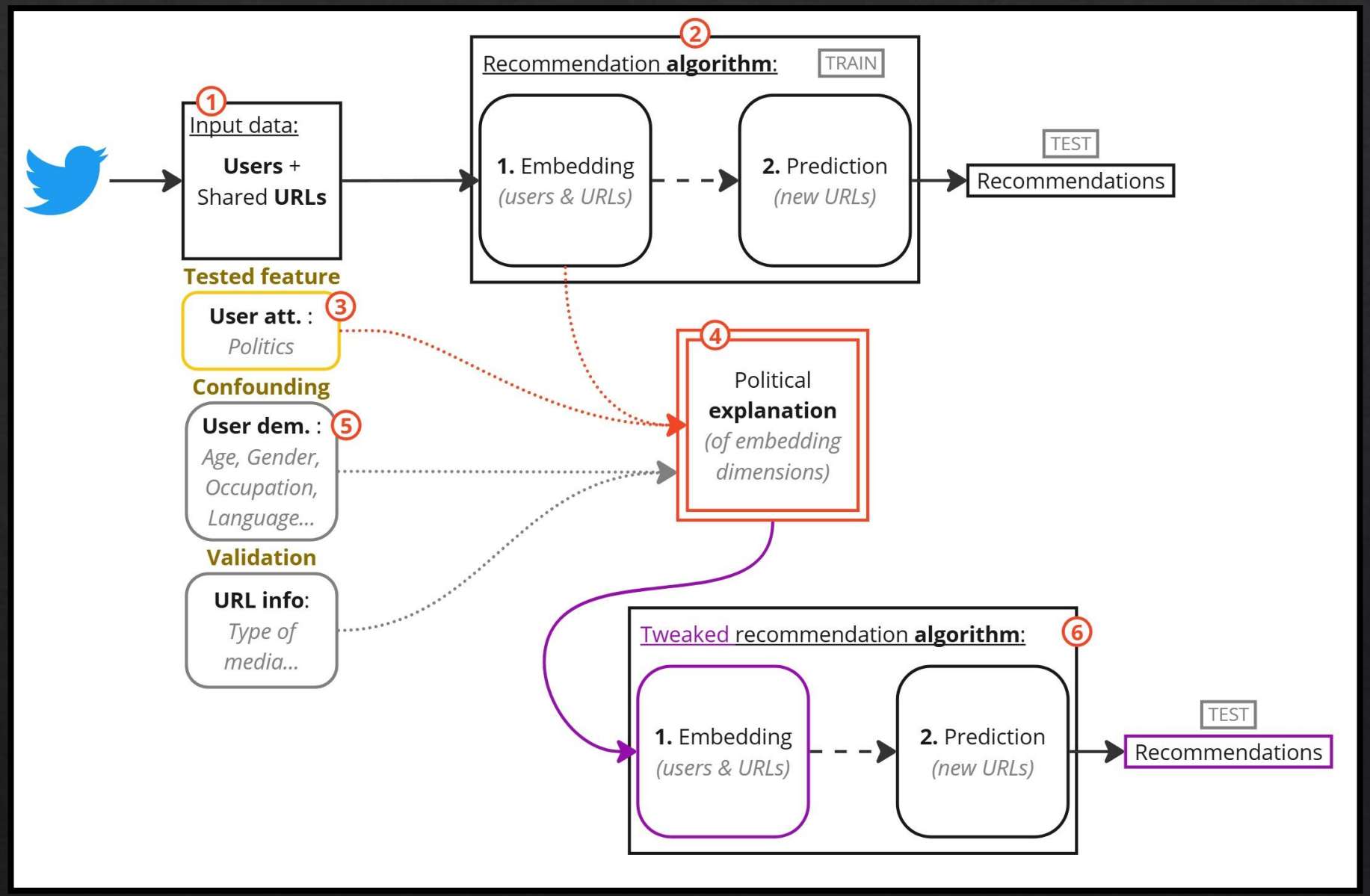
Latent diversity :

similar latent positions
⇒ similar attitudes

4. Explanation : Results

- The model can learn political attitudes
- The model identifies specific groups of only Left-wing or only Right-wing users
- We can identify the dimension carrying the Left-Right information: L_3, L_4

Method : a case study on Twitter



5. Covariates : Socio-demographic factors

Machine Learning (from bios and profile picture):

- Age
  - ≤ 18
  - $19 - 29$
  - $30 - 39$
  - ≥ 40
- Gender (M/F)
- Organization (Org/ Non-org)

Language detection on bios :

- Language

Online french media citation network :

- Media category

Manually annotated keywords on bios :

- Occupation (CSP)
  - Information, arts and entertainment professions
  - Business, IT and administration professionals
  - Elected officers and political representatives
  - Professors and higher scientific professions
  - Legal professions
- Declared interest (co-occurring keyword groups)
  - Politics
  - Academia
  - Business
  - Journalism
  - Ecology

5. Covariates

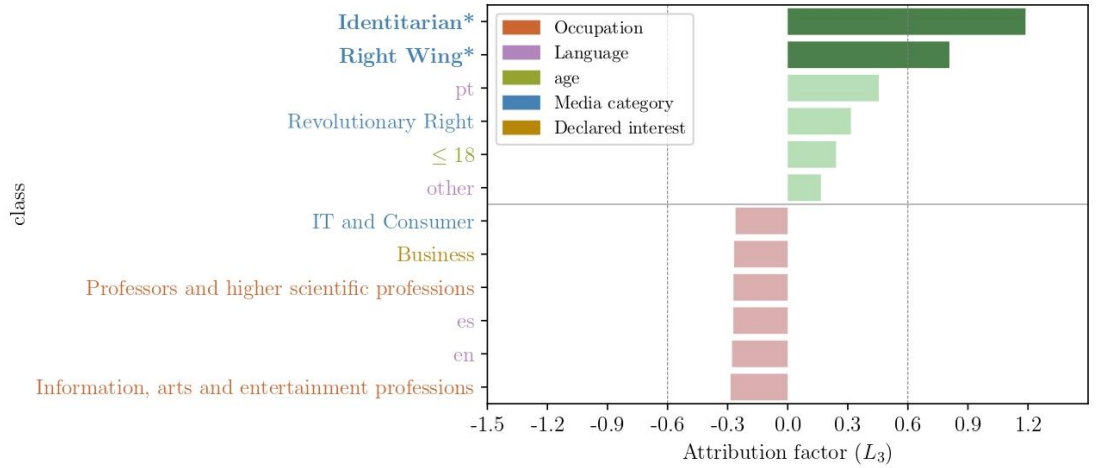
Covariates considered :

- Age
- Gender
- Organisation
- Language
- Occupation
- Declared interest

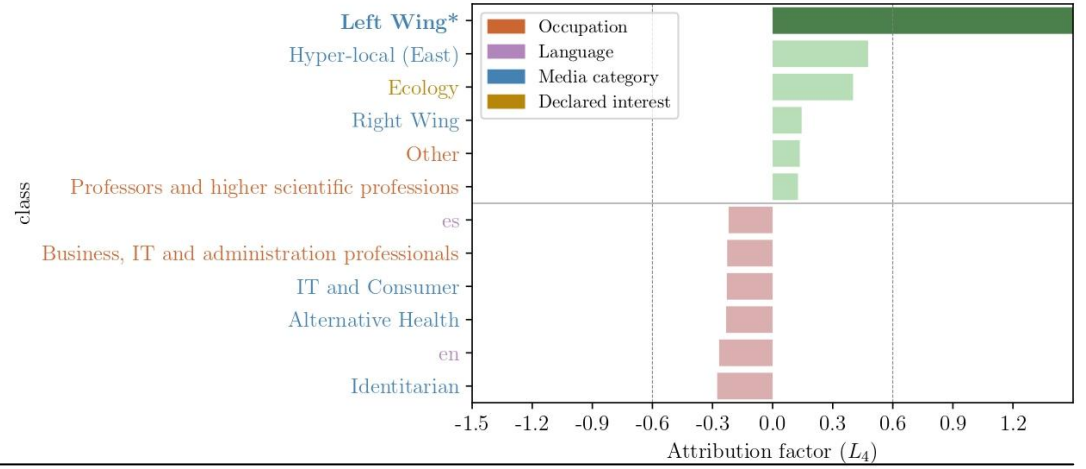
Validation :

- Media category

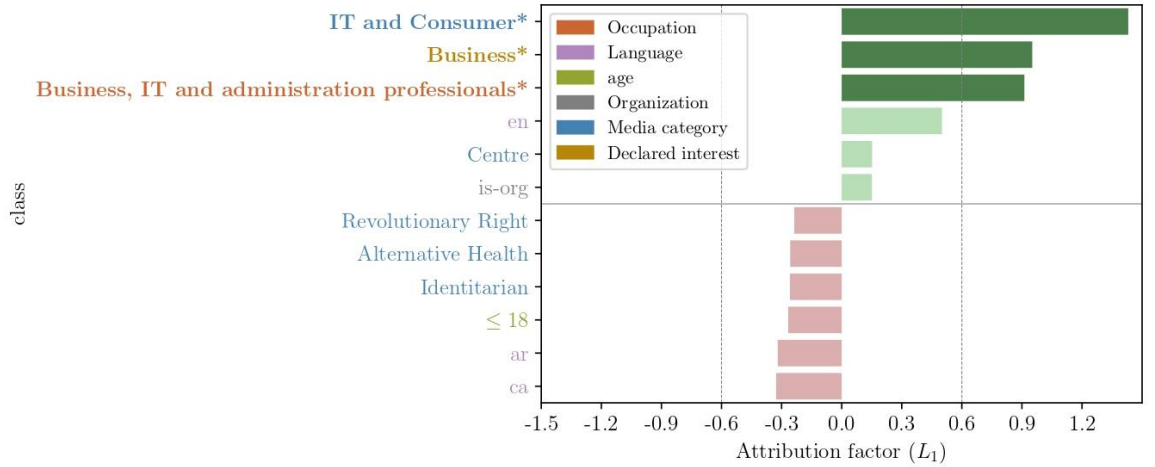
L_3 : Right leaning tendency



L_4 : Left leaning tendency



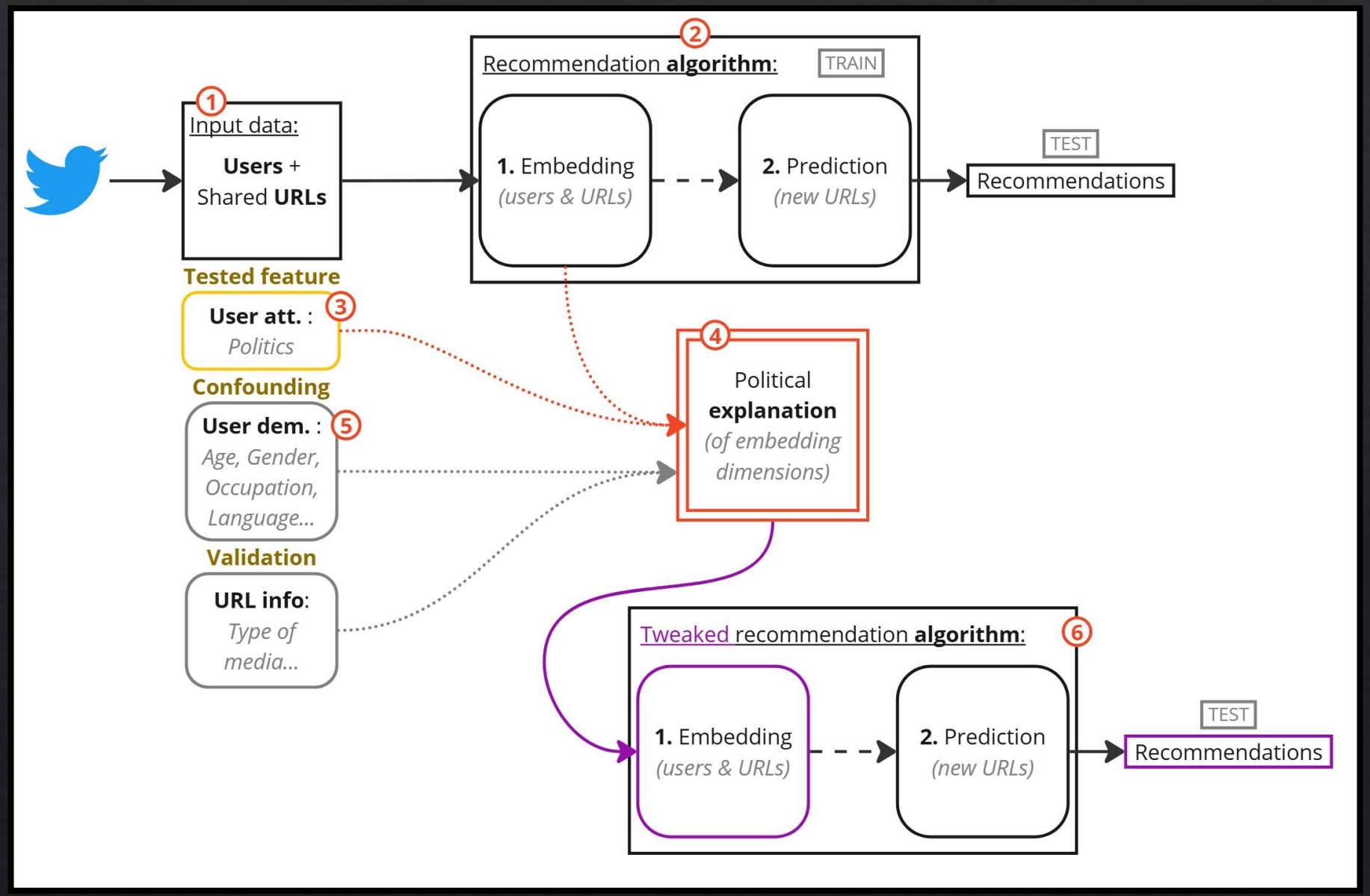
L_1 : Business and IT tendency

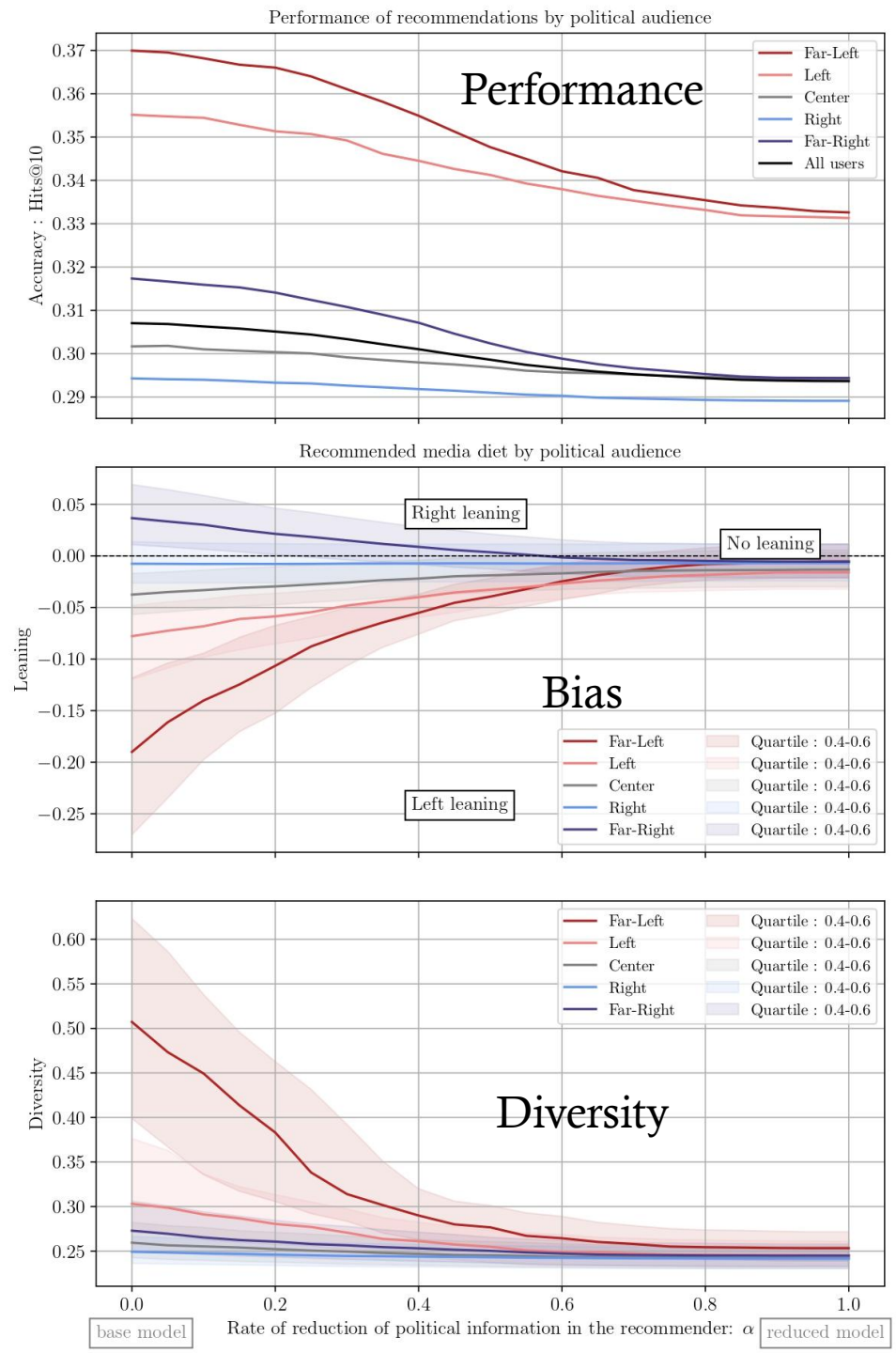


5. Covariates: Results

- The model can learn political attitudes
- The model identifies specific groups of only Left-wing or only Right-wing users
- We can identify the dimension carrying the Left-Right information: L_3, L_4
- Socio-demographic factors impact the model, but not the political dimensions

Method : a case study on Twitter





6. Political information reduction

Reducing political dimensions :

- Reduced leaning
- Reduced diversity

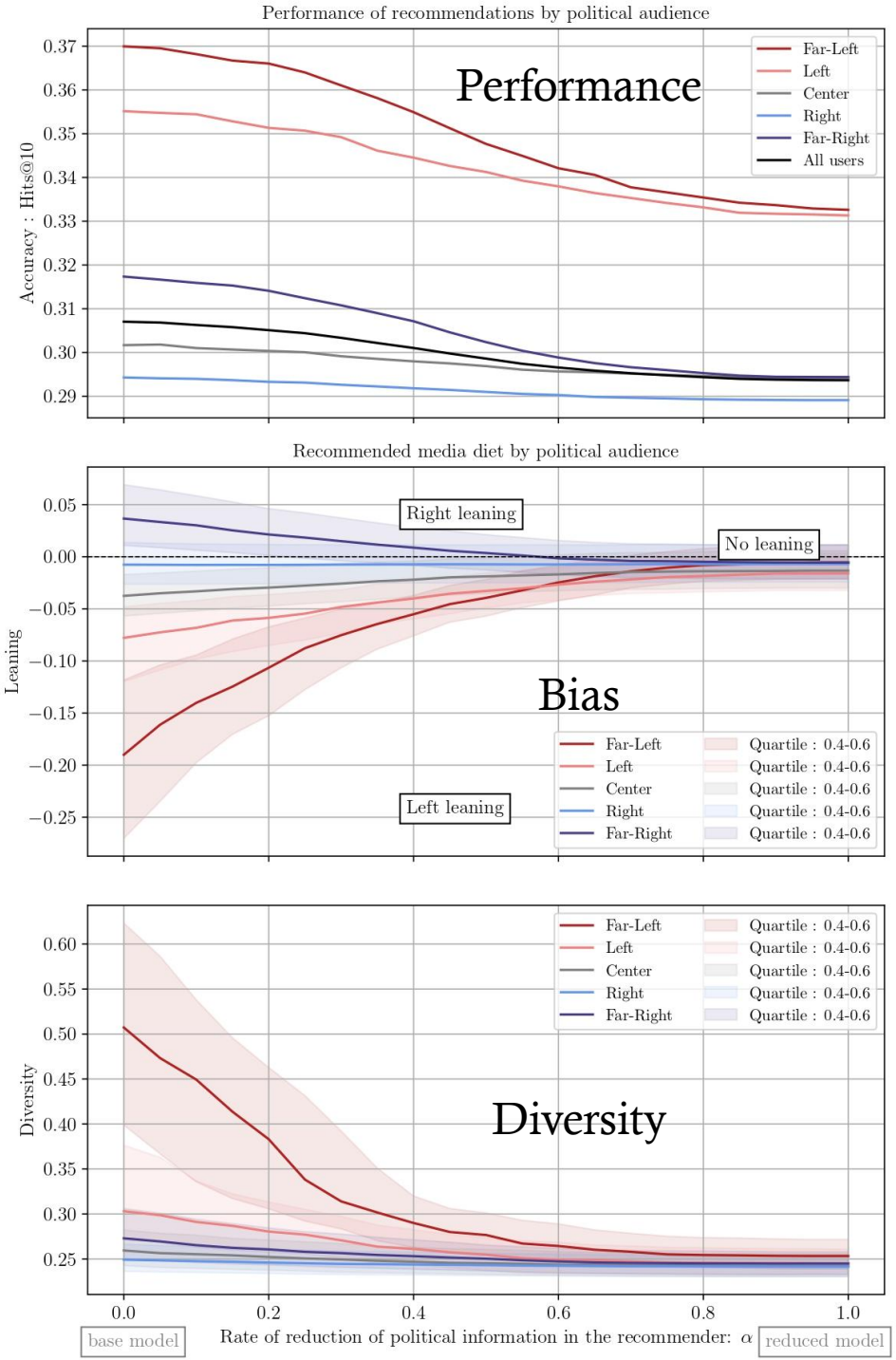
6. Political information reduction

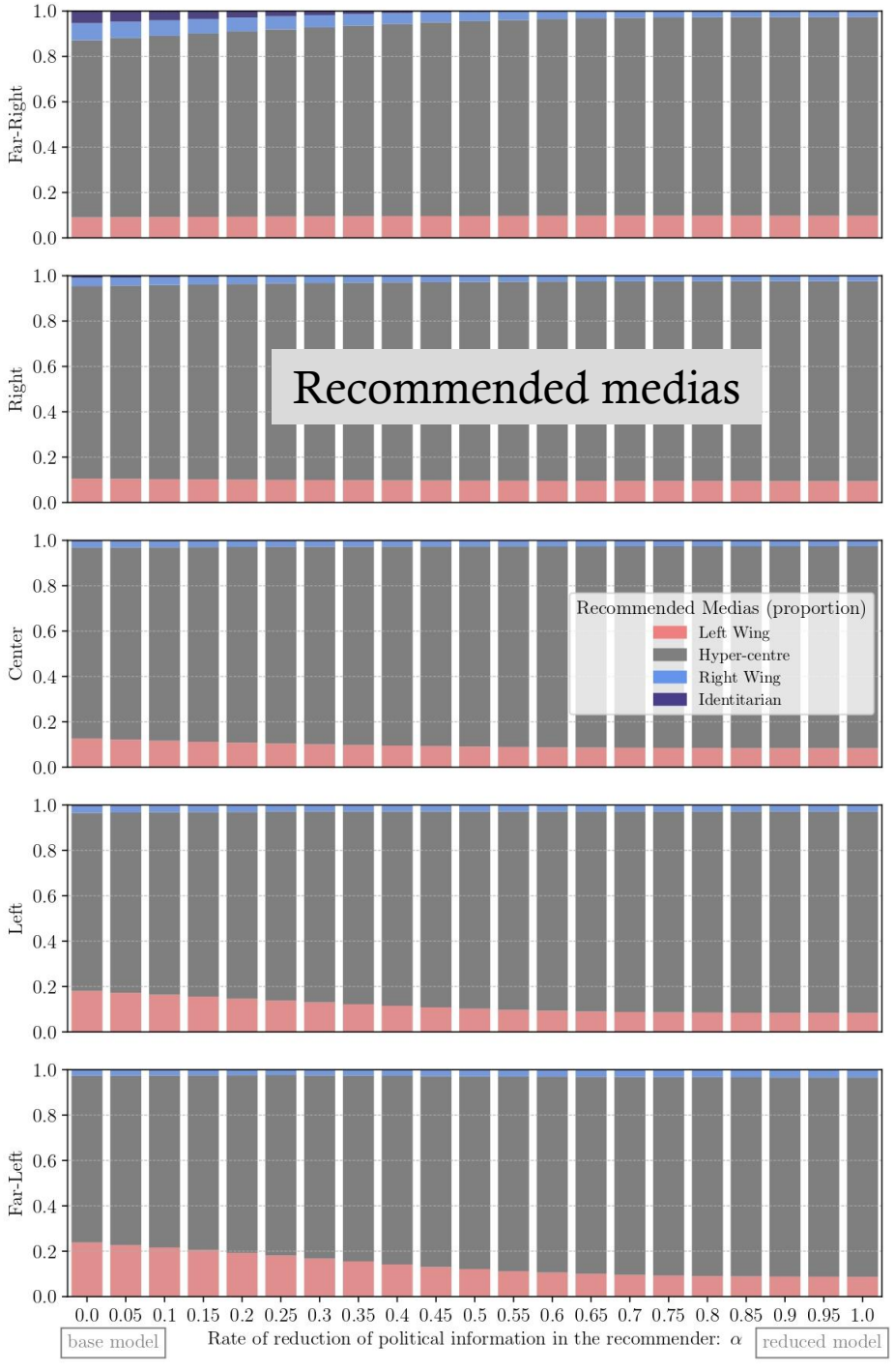
Reducing political dimens

- Reduced leaning
- Reduced diversity

Mechanism :

- Mainstream media recommendation





Conclusion

Tools :

- Political explanation for recommendation algorithms
- Manipulation of model representation

Results :

- Algorithm learn specific political features
- Reducing political information in models reduces bias while also reducing diversity

Are those results generalizable

Algorithm political explanation :

- Politicized users – *Number of dimensions*
- Sharing data – *Social media type*
- Collaborative filtering – *Complex recommenders*

Political information reduction :

- Depend on top items recommendation
- Depend on “filter bubbles”

Research Questions

- **Q1.a**: What political information is captured by the models of recommendation algorithms?
- **Q1.b**: What is the impact of learned political information on the recommendations received by users?

- **Q2**: How would recommendations of political content change if we removed political information from the models?

Supplementary Information

Discussion

- Role of mainstream medias (Benkler 2018)
- Identity radicalisation online (Bail 2021)
- “Distinction” on online consumption (Bourdieu)
- Polarization studies

Need for qualitative research :

- Difference between uses and practices

4. Explanation : design choices

- Latent embedding explanation
  - Generalization (model agnostic)
  - Statistical significance
  - Safety engineering
- Post training
  - Governance
- User based
  - Politicized content
- Global
  - Systemic phenomena evaluation
- Multi-indicators
  - Account for specific learning