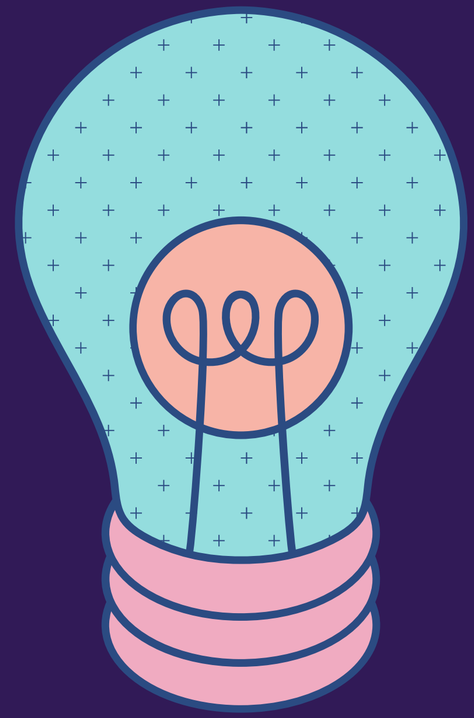
ICLR 2025

InterpretCC

github.com/epfl-ml4ed/InterpretCC

Intrinsic User-Centric Interpretability through Global MoE

EPFL




Vinitra Swamy


Syrielle Montariol

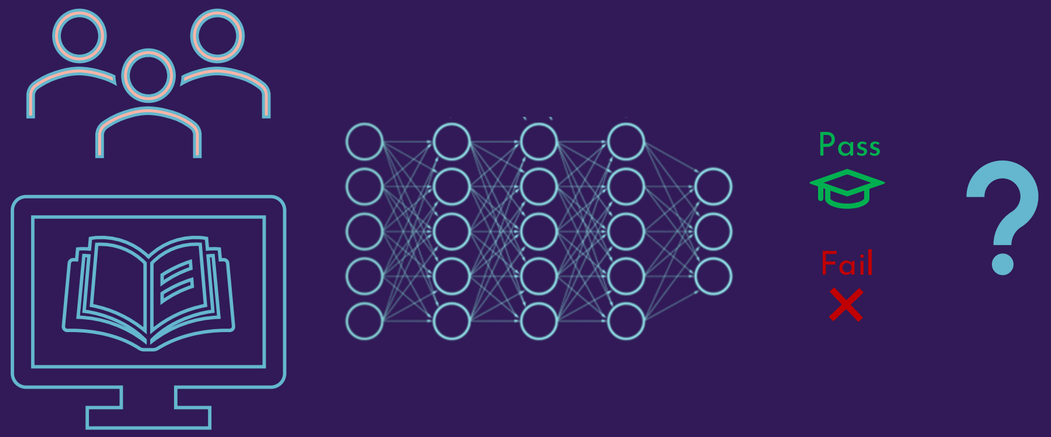

Julian Blackwell


Jibril Frej


Martin Jaggi


Tanja Käser

Explainable AI is crucial in human-centric settings



Identifying “why” is important for effective, personalized interventions

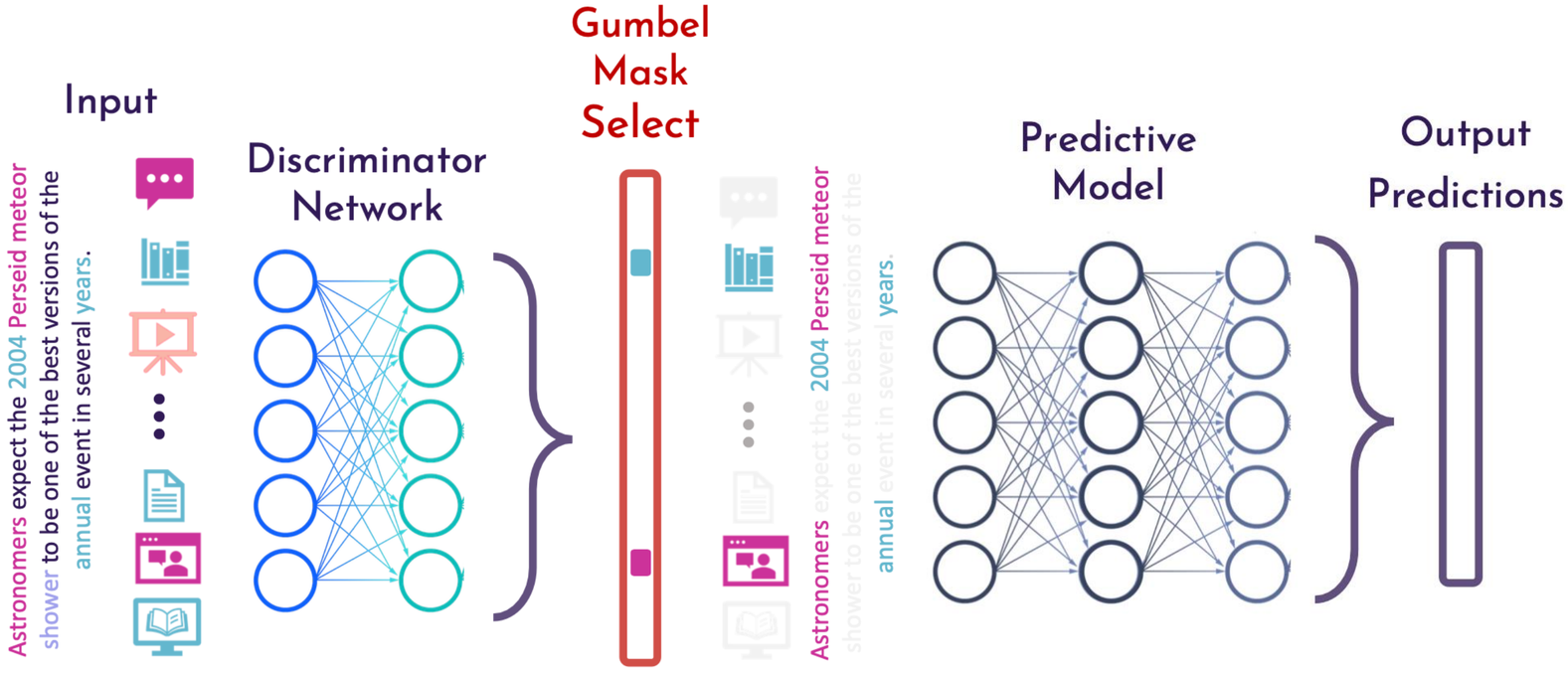
Current XAI approaches

- interpretations are not faithful (post-hoc)
- interpretations are faithful, but not user-friendly (intrinsic)

How can we design an intrinsically interpretable model that maintains performance while prioritizing users' needs?

InterpretCC: Feature Gating for Interpretability

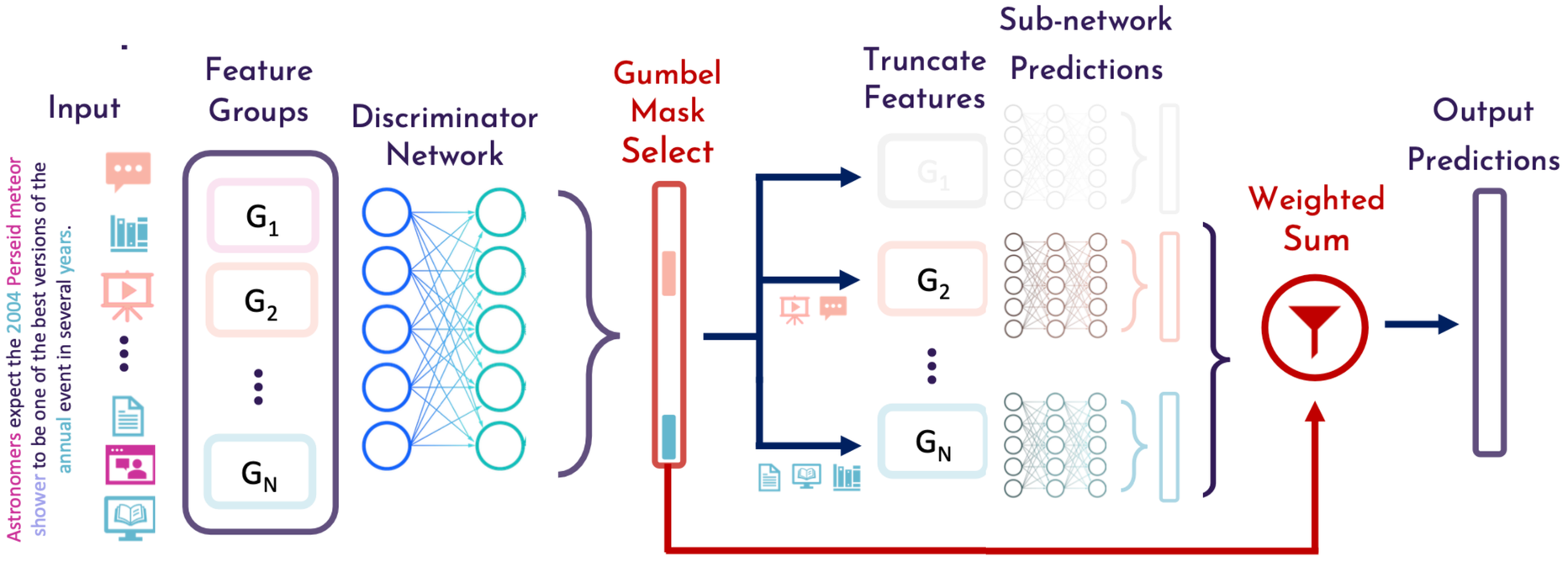
Adaptive Feature Gating – accuracy vs. interpretability tradeoff



Intrinsic, User-Centric Interpretability through Global Mixture of Experts. Swamy V. et al. (ICLR 2025)

InterpretCC: Mixture-of-Experts for Interpretability

Filter the feature space and send relevant parts to relevant experts



The student's regularity and video watching behavior were the only two aspects used to make the prediction that the student will fail the course.

Intrinsic, User-Centric Interpretability through Global Mixture of Experts. Swamy V. et al. (ICLR 2025)

InterpretCC: Results

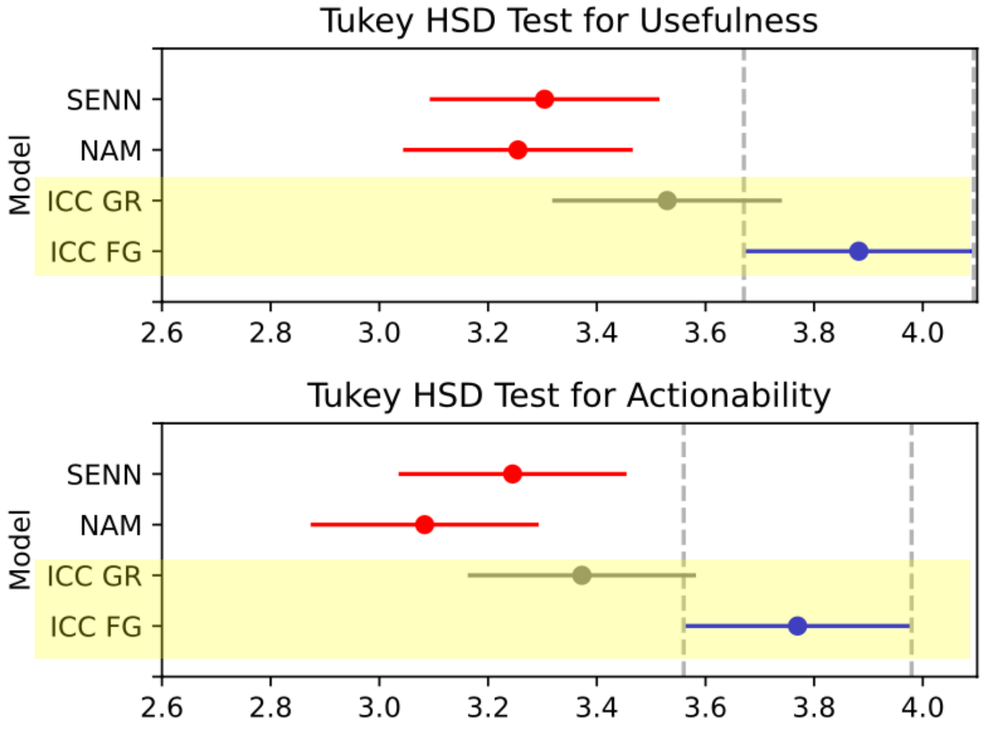
1) Maintains model performance (BAC, F1)

	Dataset	Non-interpretable Base Module	Feature-Based Interpretability			Concept-Based Interpretability		
			NAM	SENN Features	InterpretCC Feature Gating	SENN Concepts	InterpretCC Top K Routing	InterpretCC Group Routing
Education	DSP	82.81 ± 2.61	85.20 ± 0.64	71.70 ± 0.95	90.75 ± 0.01	81.50 ± 2.26	83.08 ± 1.10	84.90 ± 7.59
	Geo	72.96 ± 1.59	65.12 ± 4.07	57.90 ± 2.69	71.92 ± 0.01	70.90 ± 2.45	80.44 ± 3.19	81.58 ± 0.57
	HWTS	73.93 ± 3.76	73.11 ± 2.13	68.63 ± 3.78	82.89 ± 0.04	75.10 ± 11.67	72.59 ± 2.84	78.34 ± 0.95
	VA	74.90 ± 5.28	71.39 ± 3.38	74.37 ± 1.11	77.80 ± 0.01	69.99 ± 8.83	71.43 ± 1.11	72.08 ± 3.71
Health	B. Cancer	89.70 ± 1.05	88.77 ± 7.31	80.52 ± 6.21	78.19 ± 3.54	85.26 ± 1.03	84.66 ± 3.02	94.85 ± 1.25
Text	AG News	89.93 ± 3.32	Not Supported	Not Supported	85.72 ± 5.31	Not Supported	87.25 ± 2.48	90.35 ± 1.07
	SST	91.12 ± 2.03			88.21 ± 3.41		92.98 ± 0.88	91.75 ± 1.86
Synthetic	OpenXAI	86.67 ± 0.31	87.85 ± 1.31	83.67 ± 1.86	89.51 ± 0.51	84.67 ± 4.04	90.83 ± 1.93	89.47 ± 2.89

Intrinsic, User-Centric Interpretability through Global Mixture of Experts. Swamy V. et al. (ICLR 2025)

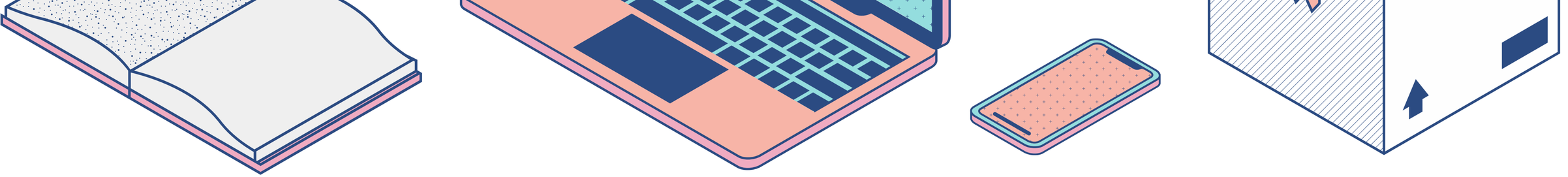
InterpretCC: Results

2) Preferred by 56 teachers over other interpretable-by-design approaches



	NAM	SENN	ICC GR	ICC FG	Weight
Usefulness	3.25 ±0.98	3.3 ±1.11	3.53 ±1.11	**3.88 ±0.94**	0.28
Trustworthiness	3.28 ±0.93	3.64 ±0.92	3.36 ±1.06	**3.78 ±0.9**	0.23
Actionability	3.08 ±0.96	3.25 ±1.06	3.37 ±1.04	**3.77 ±0.95**	0.21
Completeness	3.18 ±1.02	**3.76 ±1.09**	3.1 ±1.19	3.67 ±1.07	0.16
Conciseness	3.13 ±1.06	2.82 ±1.31	**3.72 ±1.06**	3.68 ±1.05	0.12
Global	3.2 ±0.81	3.38 ±0.85	3.41 ±0.88	**3.78 ±0.77**	

Intrinsic, User-Centric Interpretability through Global Mixture of Experts. Swamy V. et al. (ICLR 2025)



Main Takeaways

INTERPRETCC

With interpretable-by-design NNs,
guaranteed interpretability
does not have to come at the cost of performance
or human-understandability

Thank you!

Vinitra Swamy
vinitra.swamy@epfl.ch
github: epfl-ml4ed