LAUZHACK DL BOOTCAMP EPFL

Explainable AI

Landscape of Interpretable Models

VINITRA SWAMY
ML4ED x MLO

Vinitra Swamy

XAI Research

4th year PhD Student at EPFL

Co-advised by
Tanja Käser (ML4ED) and
Martin Jaggi (MLO)

UC Berkeley → Microsoft AI → EPFL

IC Rep, EPIC, PolyDoc, LauzHack

Outline

1. What is interpretability / explainability?
2. How do current explainability methods work?
3. Overview of our recent research!

What is interpretability?

"Interpretability is the degree to which a human can understand the cause of a decision." (Miller 2017)

Why is eXplainable AI important?

1. Building user trust in models
2. Auditing models when they make mistakes
3. Improving models with their own reasoning

What is an explanation?

An explanation is the answer to a why-question:

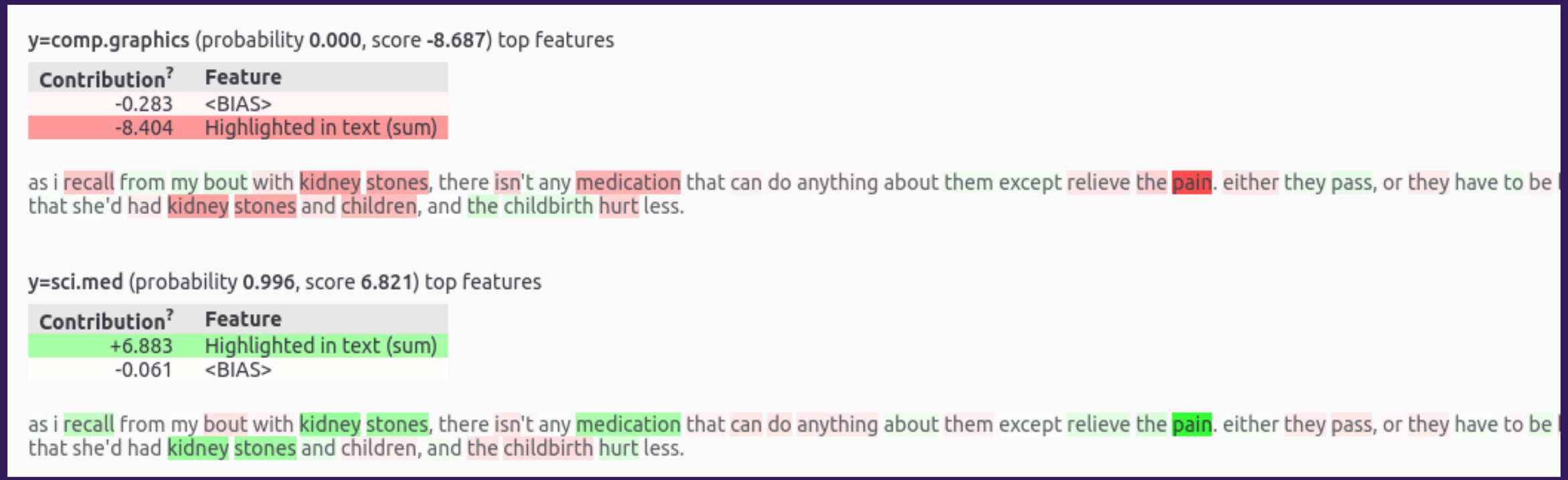
- Why did the treatment not work on the patient?
- Why was my loan rejected?
- Why have we not been contacted by alien life yet?

What is a good explanation?

contrastive, selected, social, focus on the abnormal, truthful, general and probable

Interpretability

Why did the model generate this output?
Why did the model make this decision?



y=comp.graphics (probability **0.000**, score **-8.687**) top features

Contribution[?]	Feature
-0.283	<BIAS>
-8.404	Highlighted in text (sum)

as i recall from my bout with kidney stones, there isn't any medication that can do anything about them except relieve the pain. either they pass, or they have to be
that she'd had kidney stones and children, and the childbirth hurt less.

y=sci.med (probability **0.996**, score **6.821**) top features

Contribution[?]	Feature
+6.883	Highlighted in text (sum)
-0.061	<BIAS>

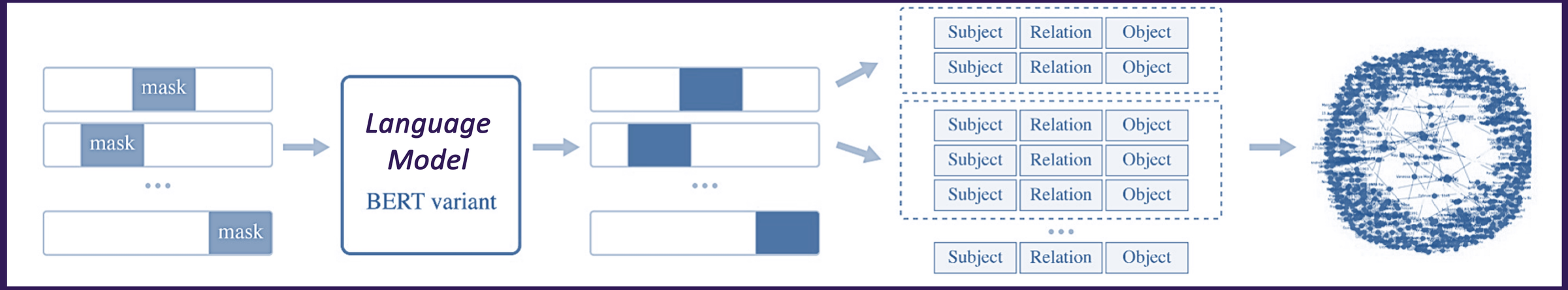
as i recall from my bout with kidney stones, there isn't any medication that can do anything about them except relieve the pain. either they pass, or they have to be
that she'd had kidney stones and children, and the childbirth hurt less.

Word importance scores for classification

[2] Ribeiro et al. (2016)

Interpretability

What knowledge does the model have?



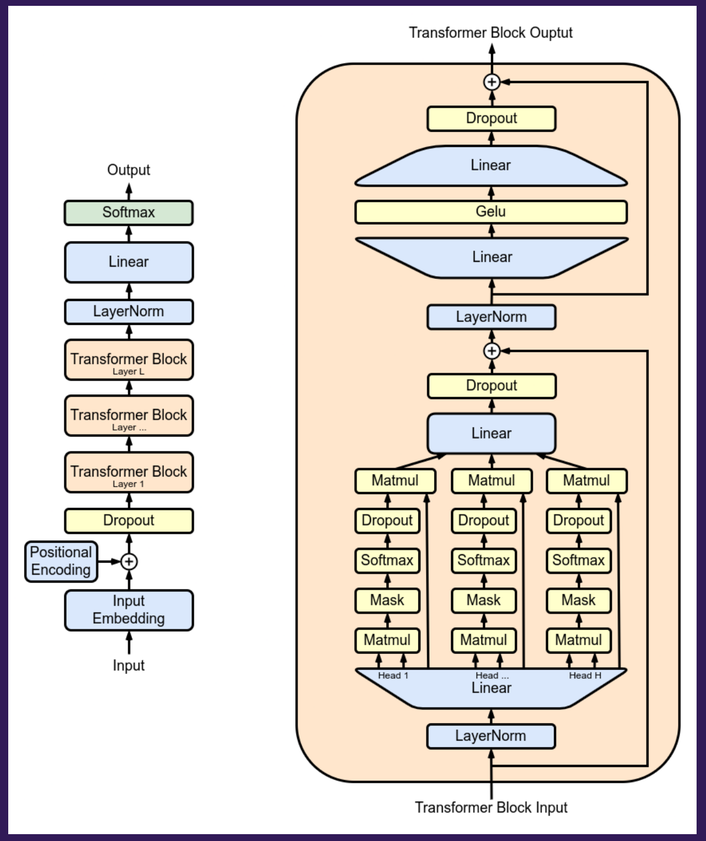
Extracting Knowledge Graphs from LLMs

[3] Swamy et al. NeurIPS XAI Workshop (2021)

Interpretability

Where in the model is this information stored?

Relationships between topics

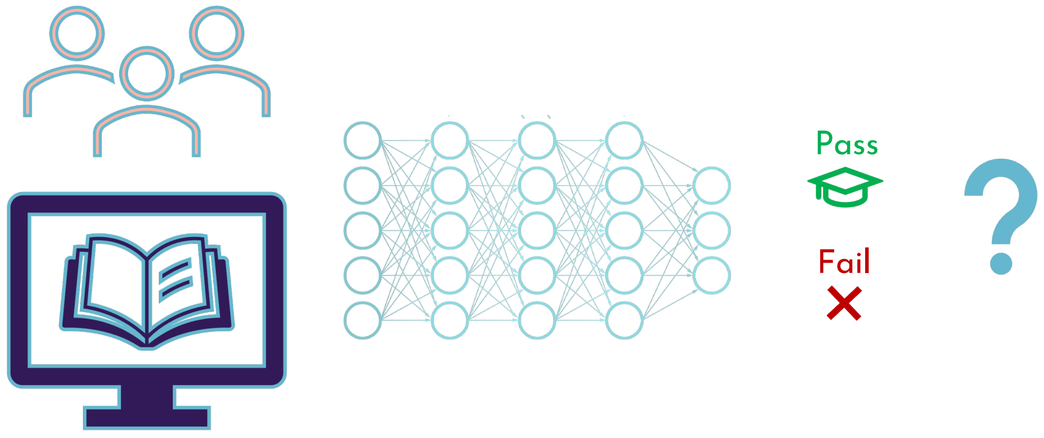


Knowledge of cities and capitals

Cost of using neural networks

DEEP LEARNING IN APPLIED FIELDS (EDUCATION, HEALTHCARE)

Problem: Deep Learning trades transparency for accuracy

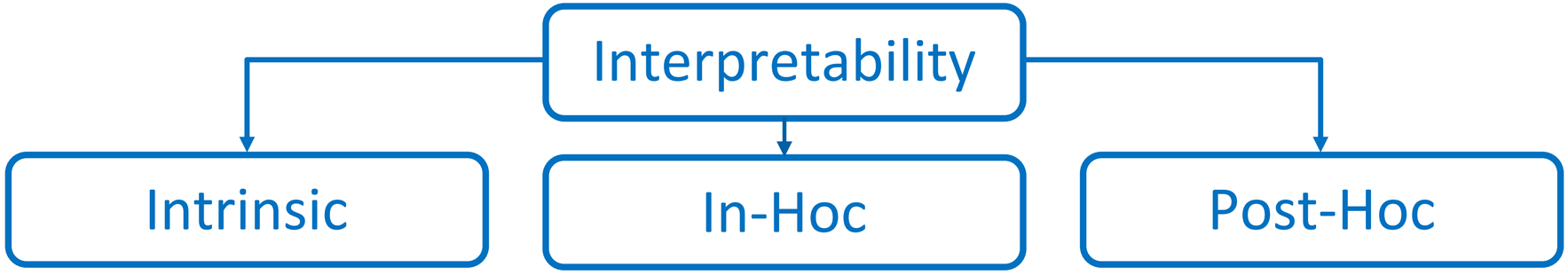


Identifying “why” is important for effective, personalized treatment

Solution: Interpretable Machine Learning / eXplainable AI

Interpretability

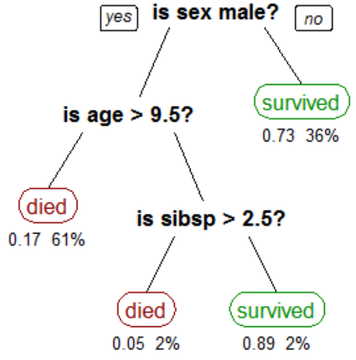
XAI Fundamentals



Decision Trees

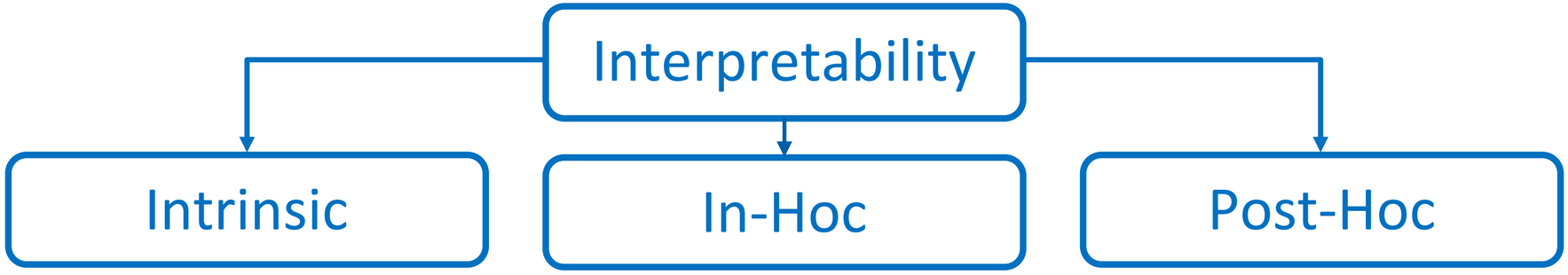
Gradient methods

LIME, SHAP

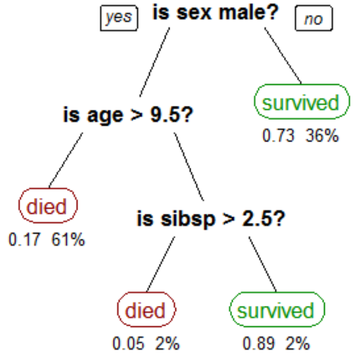


Interpretability

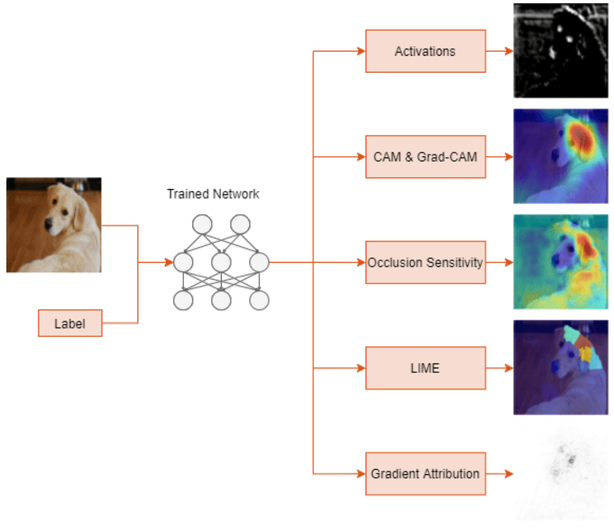
XAI Fundamentals



Decision Trees



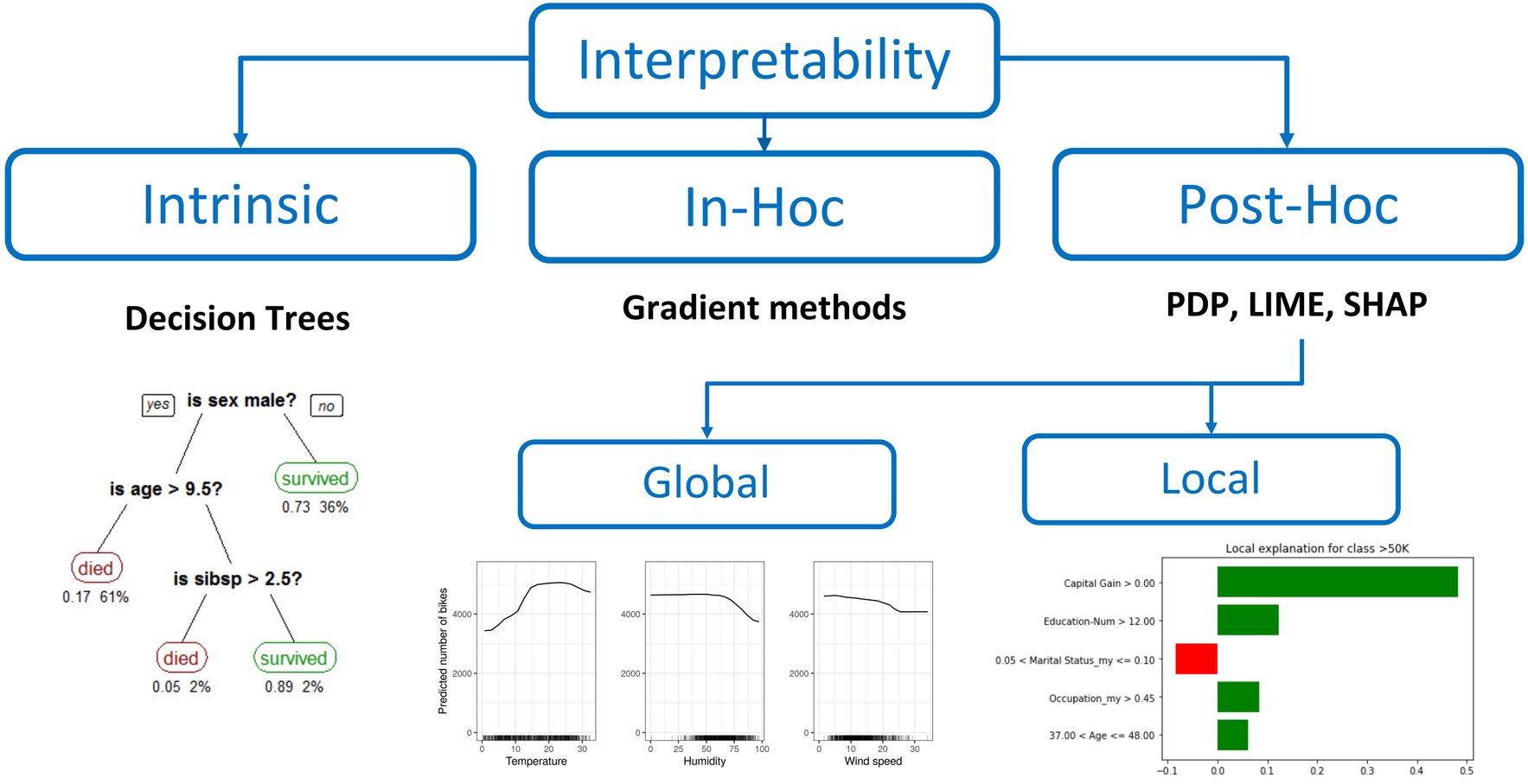
Gradient methods



LIME, SHAP

Interpretability

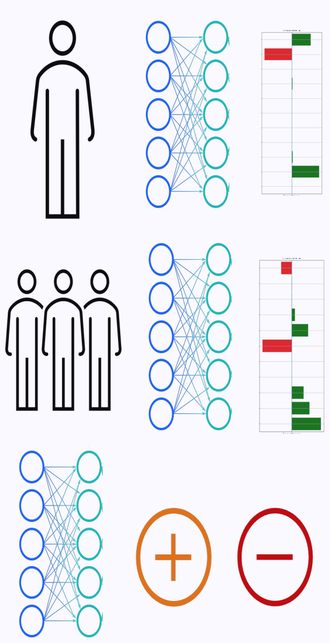
XAI Fundamentals



Interpretability

XAI Fundamentals

Local Explanation



Granularity

Global Explanation

In-Hoc Explanation

Post-Hoc Explanation

Training

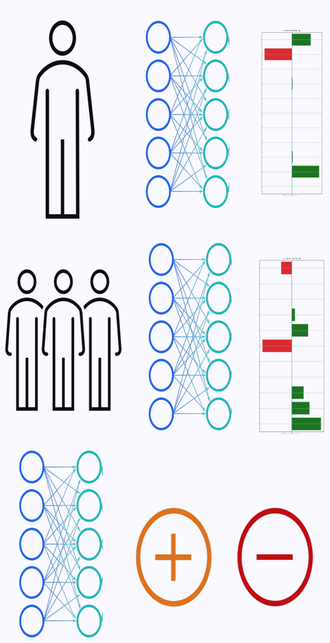
Inference

Intrinsically Interpretable Model

Interpretability

XAI Fundamentals

Local Explanation



Granularity

Global Explanation

Decision trees

In-Hoc Explanation

Post-Hoc Explanation

Training

Inference

Intrinsically Interpretable Model

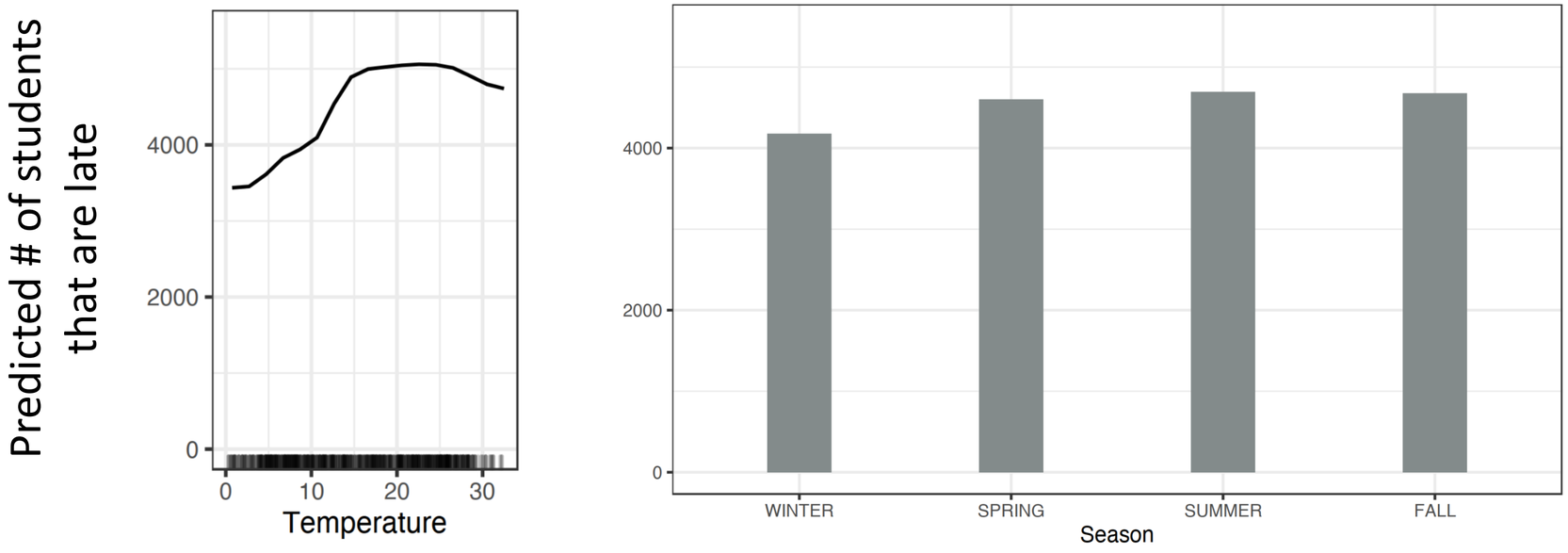
Pipeline

METHODOLOGY

PDP

Partial Dependency Plots

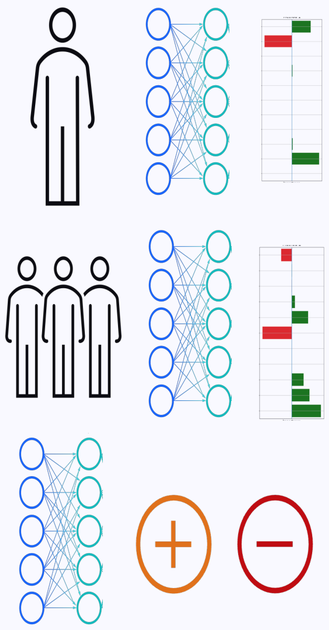
- Y denotes the number of students that will be walk into class late on a given day
- Features (X): season, weekday, temperature, class topic, prior grades …



Interpretability

XAI Fundamentals

Local Explanation



Granularity

Global Explanation

Decision trees

PDP

In-Hoc Explanation

Post-Hoc Explanation

Training

Inference

Intrinsically Interpretable Model

Pipeline

METHODOLOGY

LIME

Local Interpretable Model-Agnostic Explanations

1

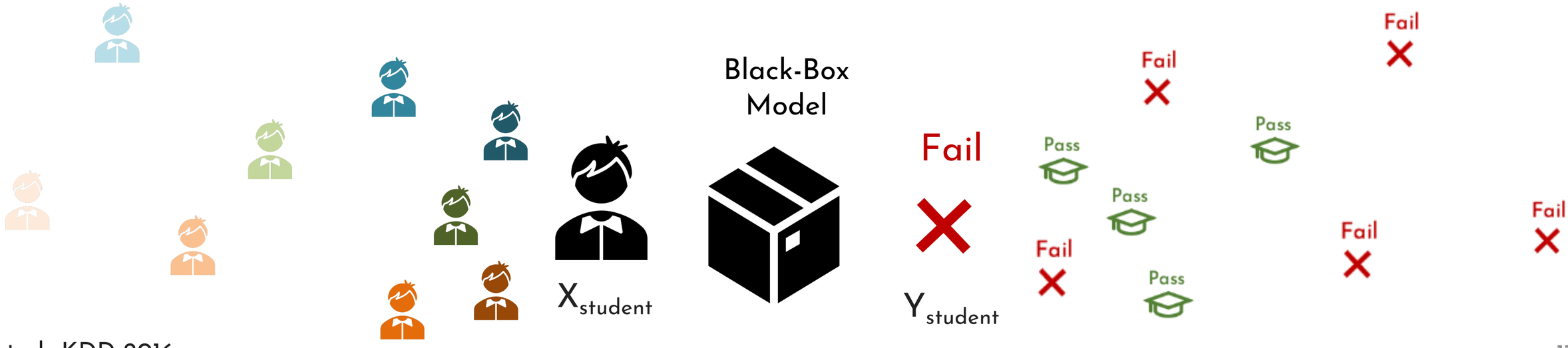
Select a specific point to explain: ($X_{student}$, $Y_{student}$)

2

Perturb features of selected point to get {$X^1_{student}$ … $X^N_{student}$} neighbors

3

Feed in $X_{student}$ neighbors to the black-box model and get predictions {$Y^1_{student}$ … $Y^N_{student}$}



Ribeiro et al., KDD 2016.

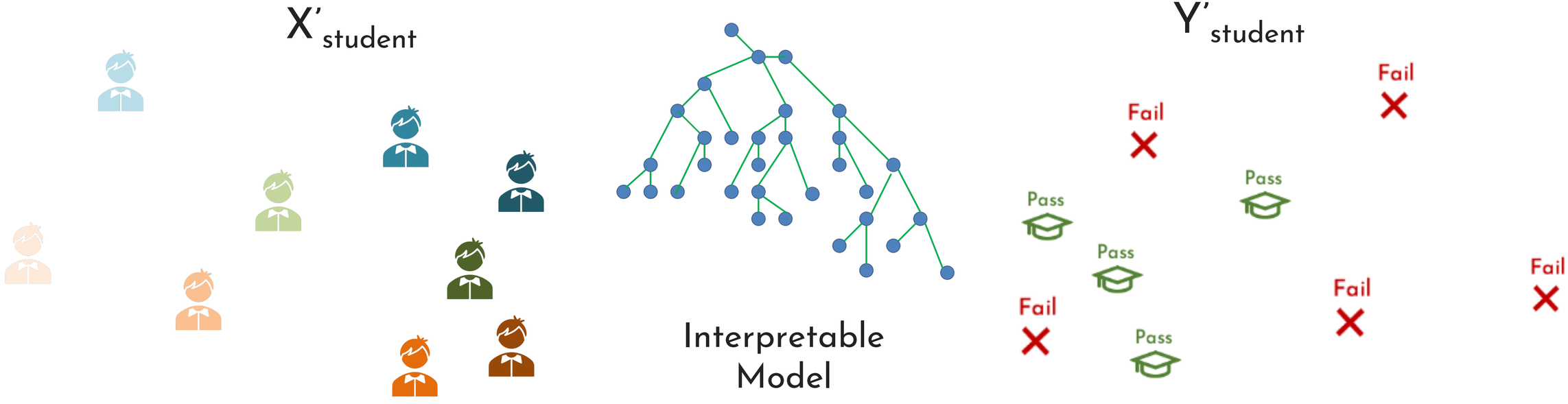
Pipeline

METHODOLOGY

LIME

Local Interpretable Model-Agnostic Explanations

4 Train an interpretable local model using (weighted) $X'_{student}$ and $Y'_{student}$



Ribeiro et al., KDD 2016.

Pipeline

METHODOLOGY

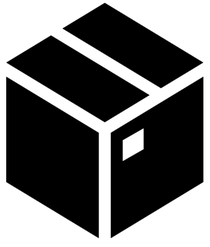
SHAP

SHapley Additive exPlanations

SHAP explains $X_{student}$ by quantifying the contribution of each feature to the prediction.

$X_{student}$

Black-Box Model

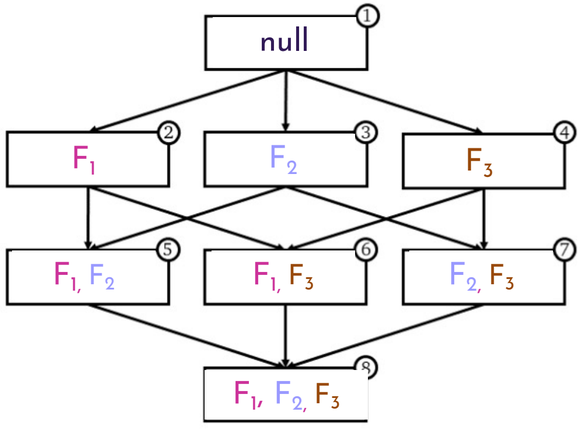


Fail

$Y_{student}$

F_1 { # of minutes watching videos , F_2 # of clicks on problems this week , F_3 # of sessions (overall) }

Power Set (coalition)



cardinality

2^3

Lundberg and Lee, NeurIPS 2017.

Pipeline

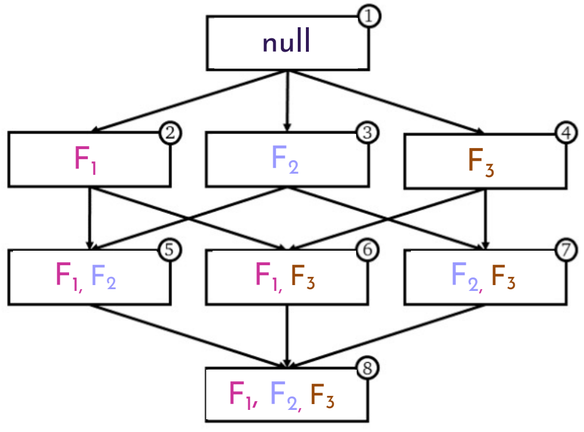
METHODOLOGY

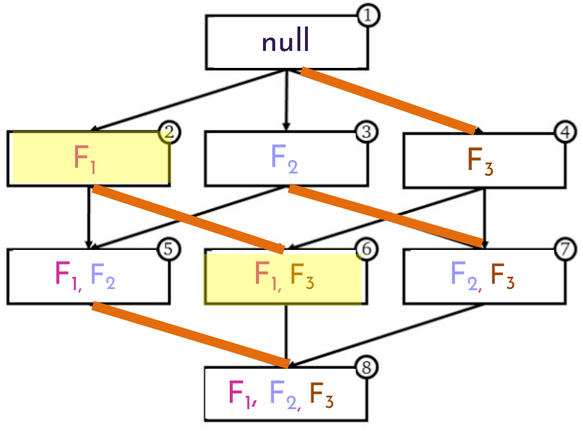
SHAP

SHapley Additive exPlanations

1 Train a model on each feature coalition.

2 Weighted sum of "marginal contributions" for each feature (i.e. F_3).





KernelSHAP

Optimizations using the **SHAP kernel function** for efficient data point construction

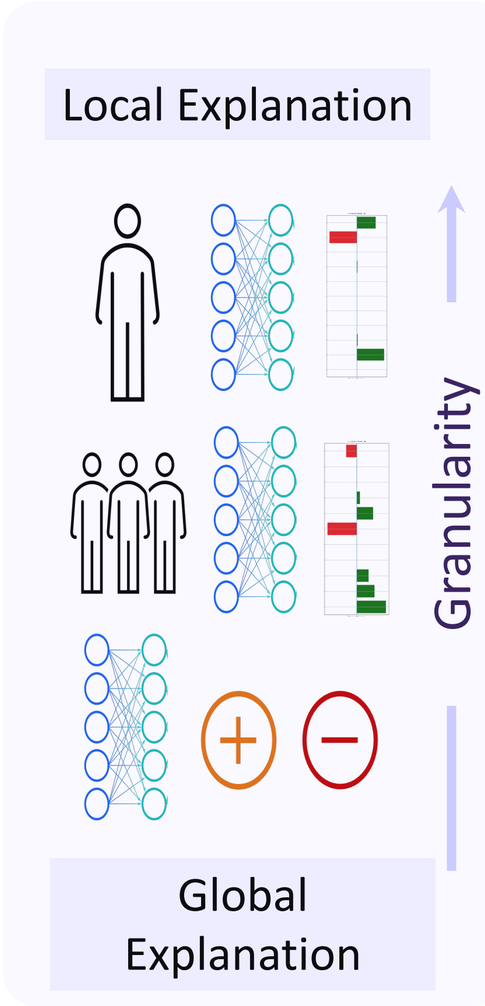
PermutationSHAP

All feature combinations in forward and reverse directions **(antithetic sampling)**

Lundberg and Lee, NeurIPS 2017.

Interpretability

XAI Fundamentals



Decision trees

LIME

SHAP

PDP

In-Hoc Explanation

Post-Hoc Explanation

Training

Inference

Intrinsically Interpretable Model

Pipeline

METHODOLOGY

CEM

Contrastive Explanation Method

$\{ F_1, F_2, F_3, F_4, \ldots F_{42}, F_{43}, F_{44}, F_{45} \}$

Pertinent Positives (PP)

X' with the minimal subset of features that should be present to maintain the prediction.

Pertinent Negatives (PN)

X' with a subset of features absent while maintaining the prediction.

Feature importance: $|X'_{student\_k} - X_{student\_k}| \times SD_{feature}$

Klaise et al., NeurIPS 2018.

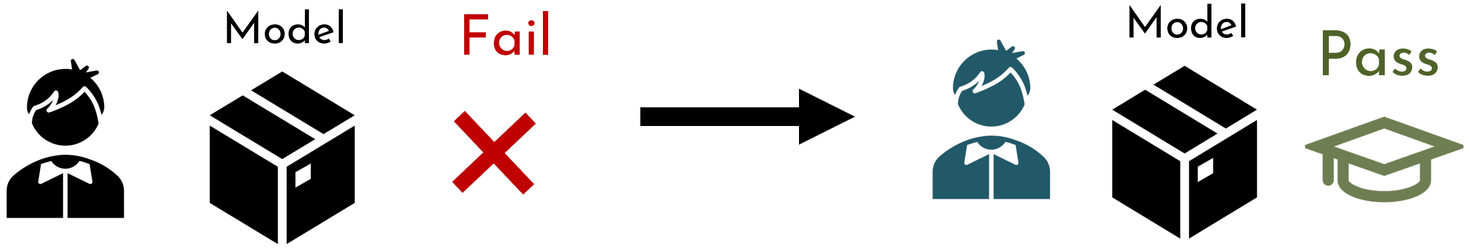
Pipeline

METHODOLOGY

DiCE

Diverse Counterfactual Explanations for ML

Generate a point with the smallest possible change to the initial instance that results in a different prediction.



Optimize DiCE loss

Determinal Point Process (DPP)
Diversity Metric

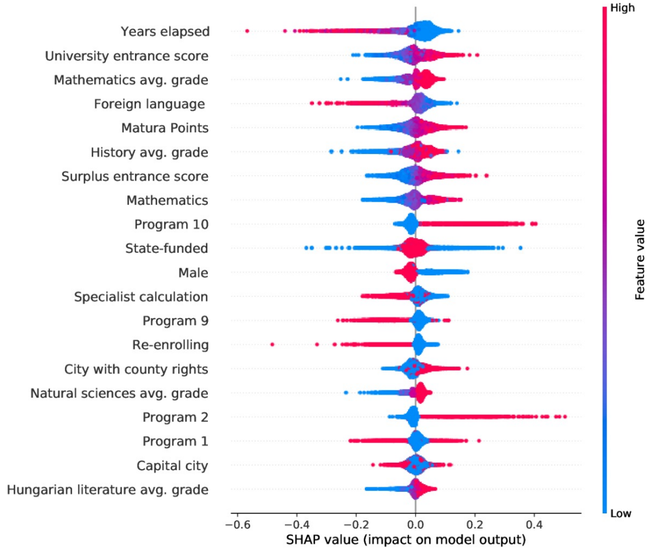
Mothilal et al., FAT* 2020.

Previous Work

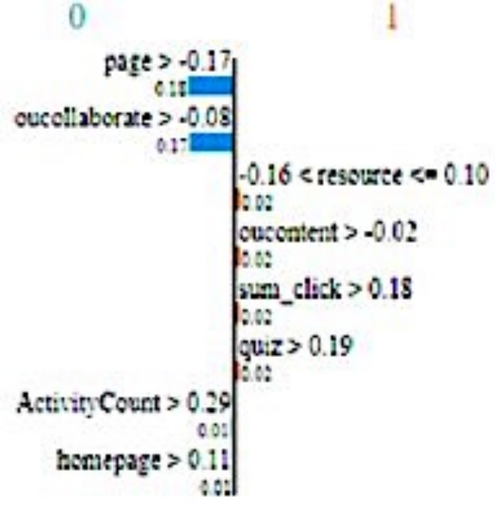
MOTIVATION

Previous work: In (minimal) related literature, only one explainability method is picked per application

SHAP for student dropout[1]



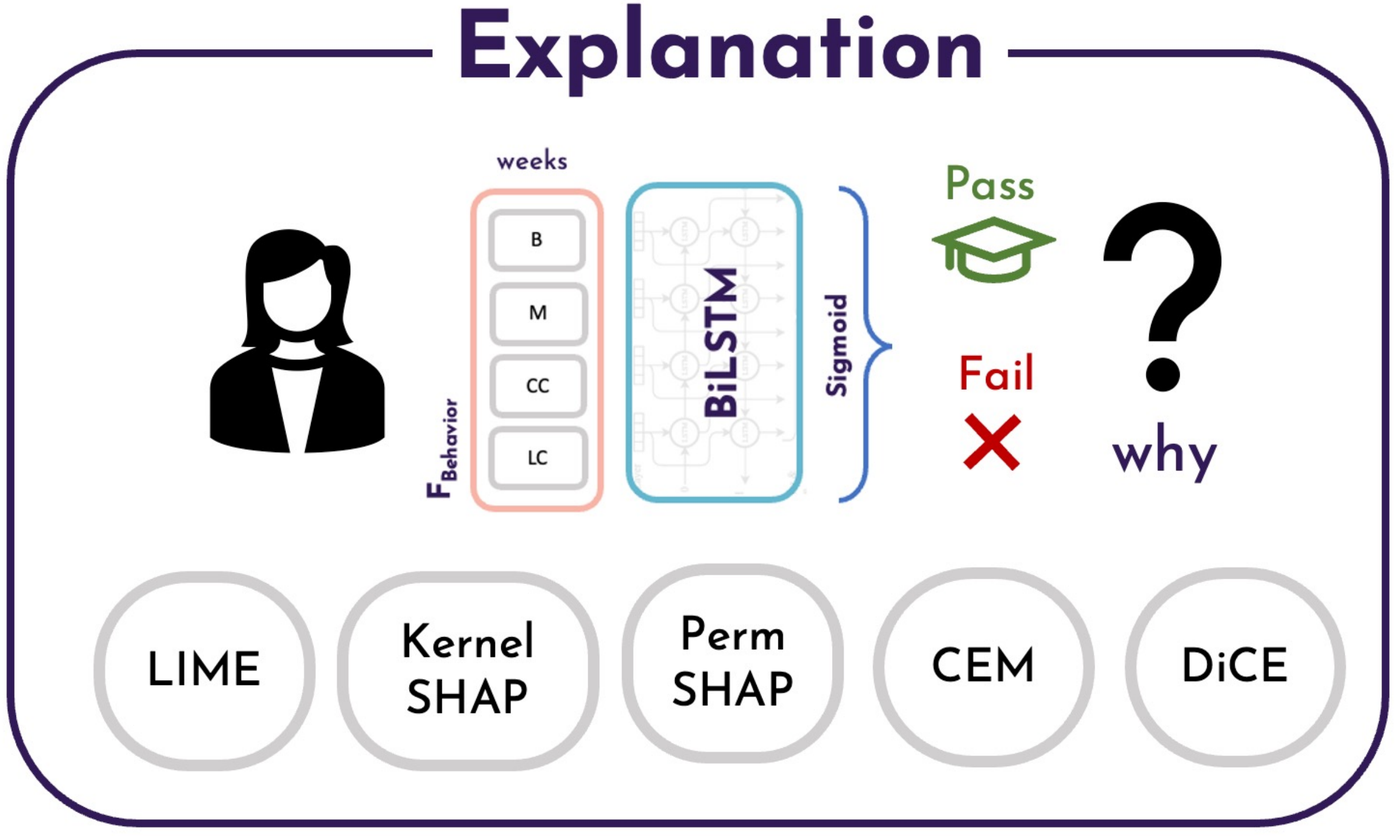
LIME for student advising[2,3]



Baranyi et al. CITE 2020; Scheers and Laet, ECTEL 2021; Pei and Xing, Journal of Edu Computing Research 2021

Pipeline

METHODOLOGY

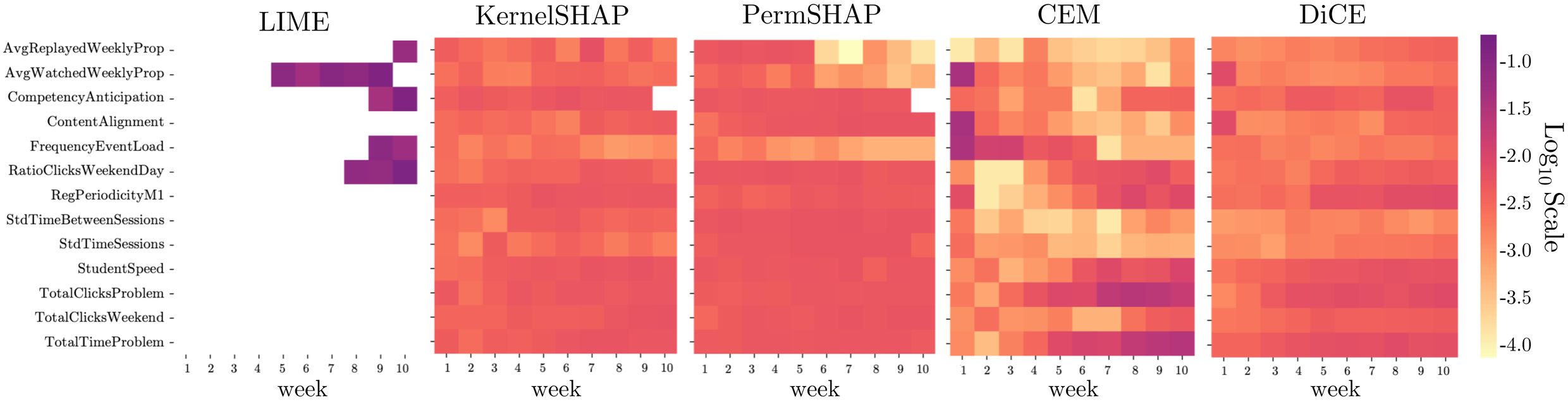


Explanation: How important is this feature to the model's prediction?

RQ1: 1 Course

RESULTS

How similar are the explanations of different explainability methods for a specific course (DSP 1)?



LIME is very sparse. CEM is significantly different.

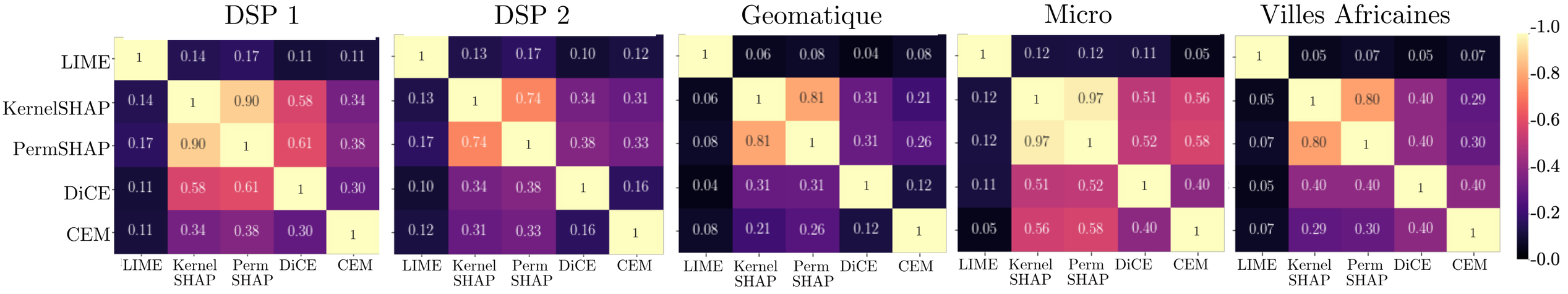
RQ2: 5 Courses

RESULTS

[EDM 2022] Evaluating the Explainers
Swamy, Radmehr, Krco, Marras, Käser

How do explanations (quantitatively) compare across courses?

Spearman's Rank Order Correlation



Big differences across explainability methods.

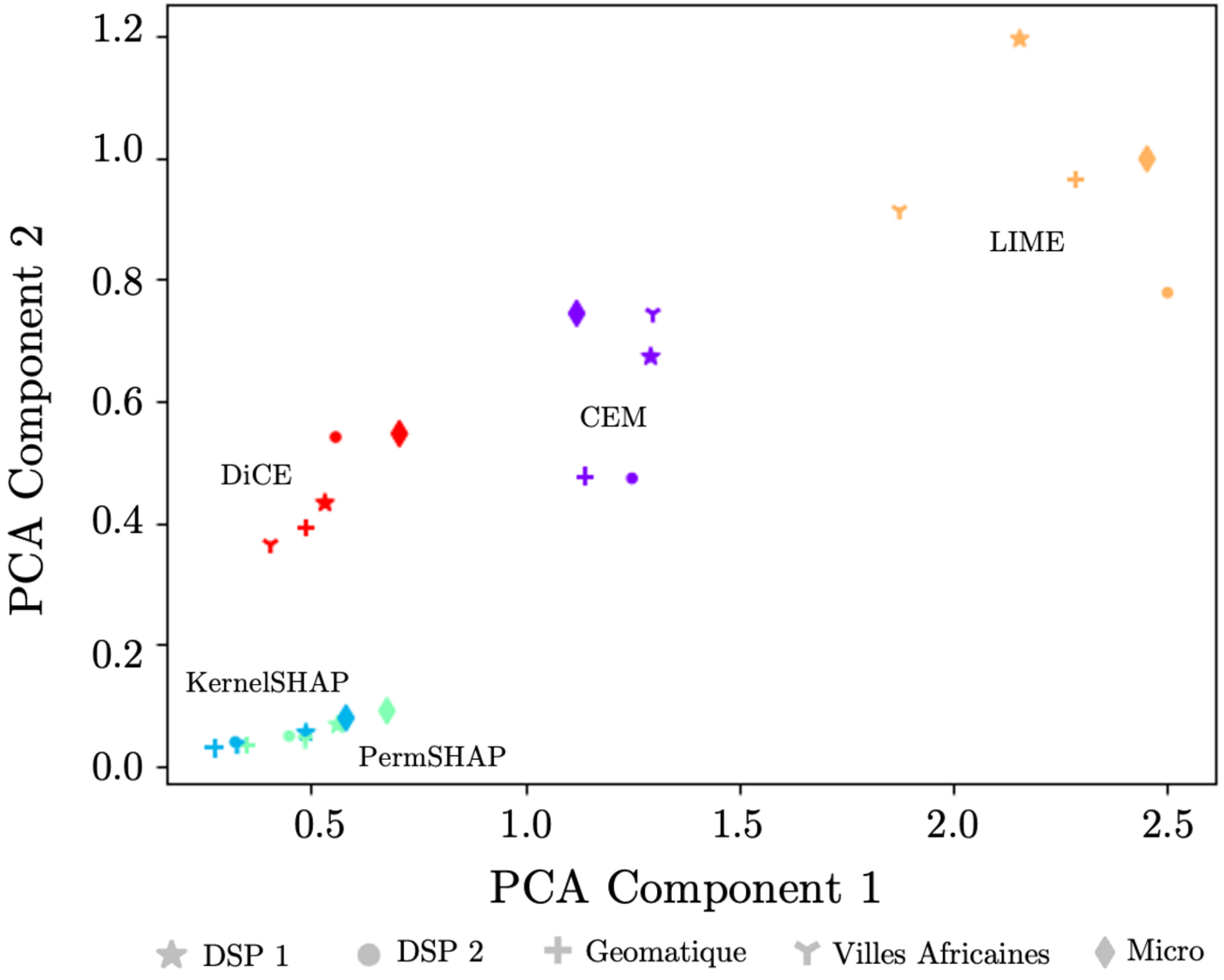
RQ2: 5 Courses

RESULTS

How do explanations (quantitatively) compare across courses?

PCA Analysis

Feature importance clusters by explainability method, not by course



XAI methods systematically disagree.

How can we build trust in them?

Human expert validation!

Study Participants

RESULTS

Who can we most trust to validate educational explainers?

Pass

Fail

26 STEM professors

45 minute semi-structured interviews

- diverse in geographic location
- 80.95% identify as male
- age: 36.5 yrs (std: 9 years)
- strong MOOC expertise

Interview Structure

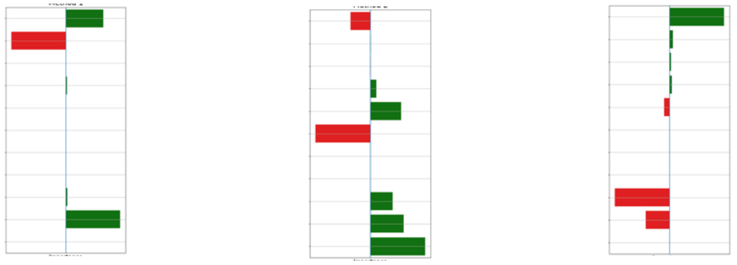
[LAK 2023] Trusting the Explainers
Swamy, Du, Marras, Käser
Best Paper Nominee

RESULTS

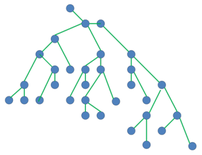
Course pairs that show differences in learning behavior

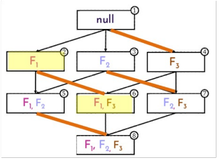
Qualitative Validation

M1 M2 M3



26 expert interviews with educators





LIME SHAP

$\{F_1, F_2, F_3, F_4, \ldots F_{42}, F_{43}, F_{44}, F_{45}\}$

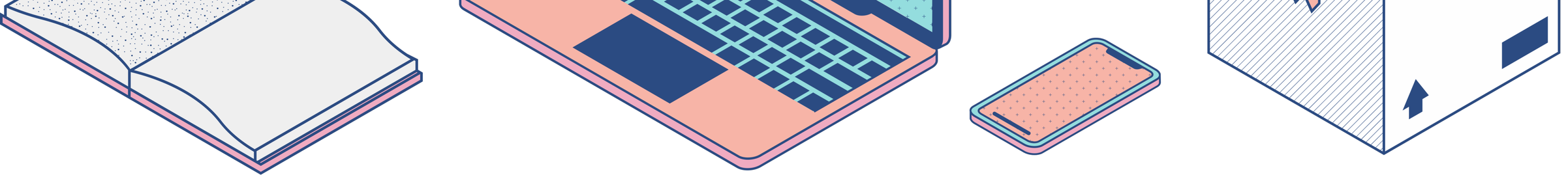
confounder

Expert Trust

RESULTS

[LAK 2023] Trusting the Explainers
Swamy, Du, Marras, Käser
Best Paper Nominee

- diverse prior perceptions of what factors enable student success and failure
- trust explanations that aligned with their beliefs (46.8%) (which led to significant disagreement across experts)
- only 3 chose the same method for both courses in the pair



Main Takeaways

EVALUATING THE EXPLAINERS: BLACK BOX EXPLAINABLE ML FOR SUCCESS PREDICTION

Explainability methods, systematically,
do not agree
on which features are important for predictions

…. and neither do human experts

Where do we go from here?

Intrinsic or in-hoc interpretable neural nets!

Framework of Explainability Needs for Human-Centric Computing

[JAIR 2023] XAI Needs
Swamy, Frej, Käser

Consistent
[multiple generated explanations are the same]

Real-Time
[next minute, next lesson, 1 week, after the course]

Accurate
[model is confident in the explanation]

Actionable
[able to take a next step based on the insight]

Human-Interpretable
[easy for a non-data scientist to understand]

[NeurIPS 2023*] MultiModN
Swamy*, Satyeva*, Frej, Vogels,
Bossy, Jaggi, Käser*, Hartley*

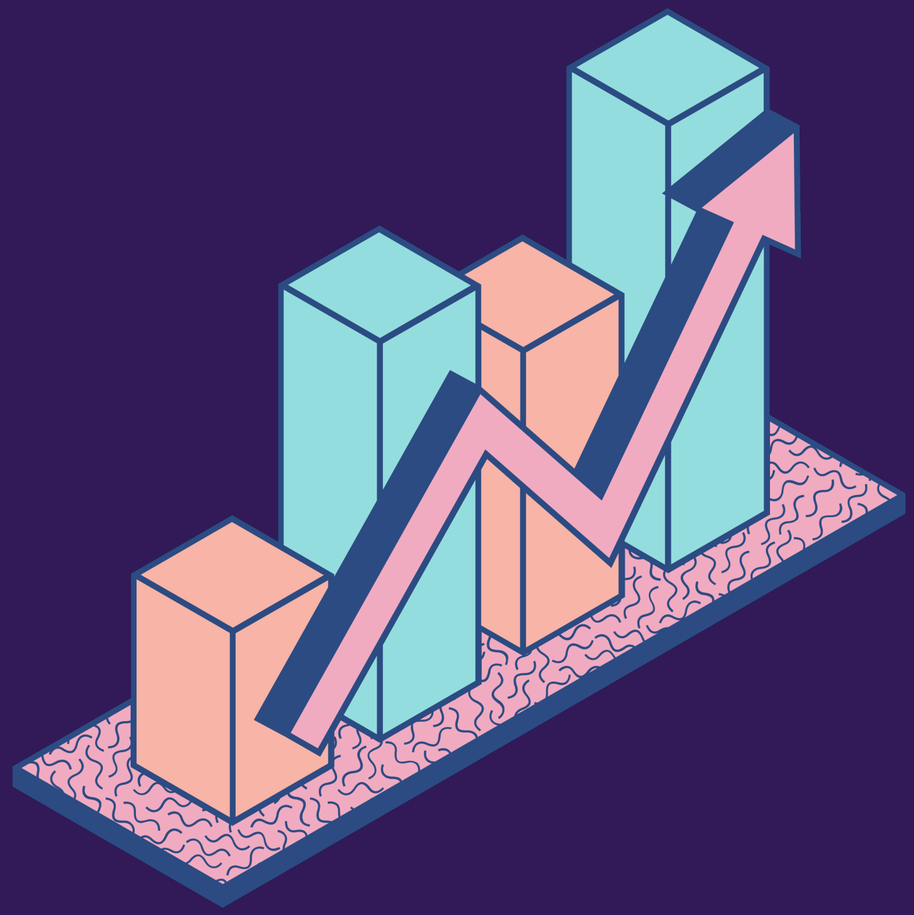
Multimodality

- synergistic predictive potential
- inputs with drastically varying sizes (i.e. images, text, sound)

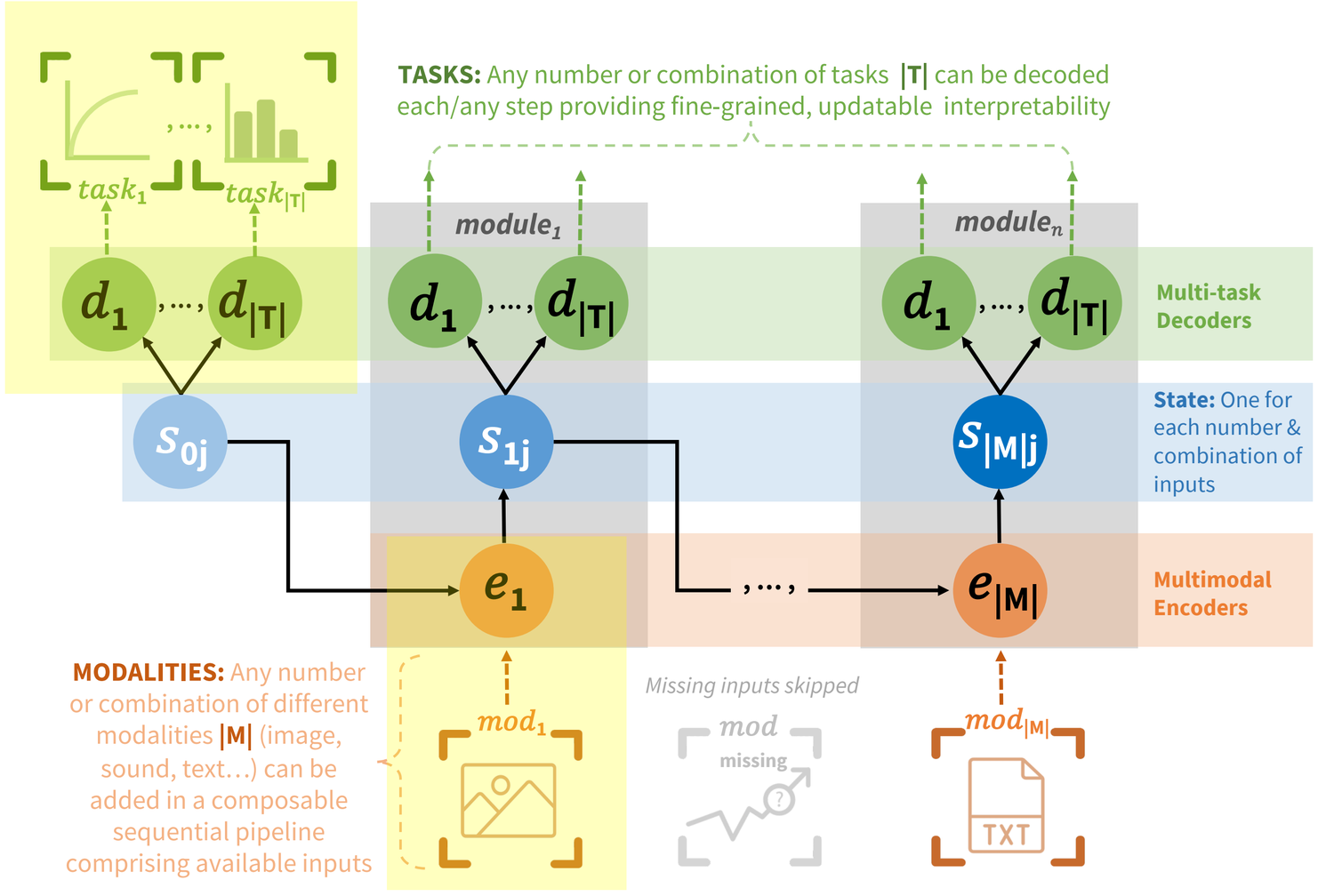
Current Approaches

fuse representations in parallel

- limits interpretability
- dependency on modality availability

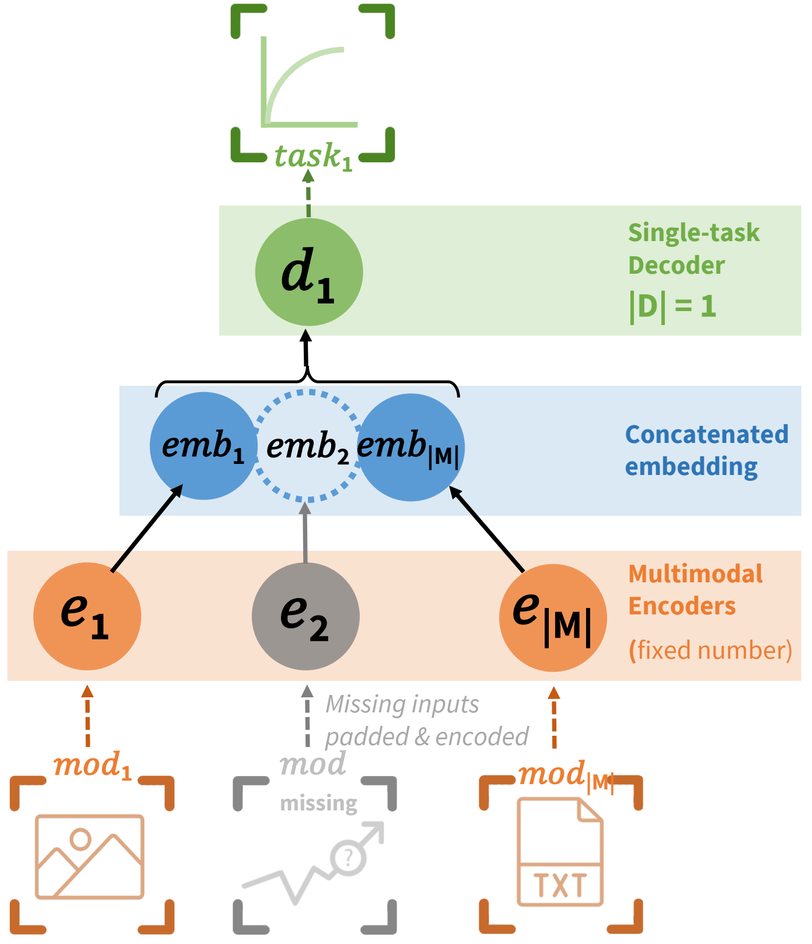


MultiModN



fusion in sequence

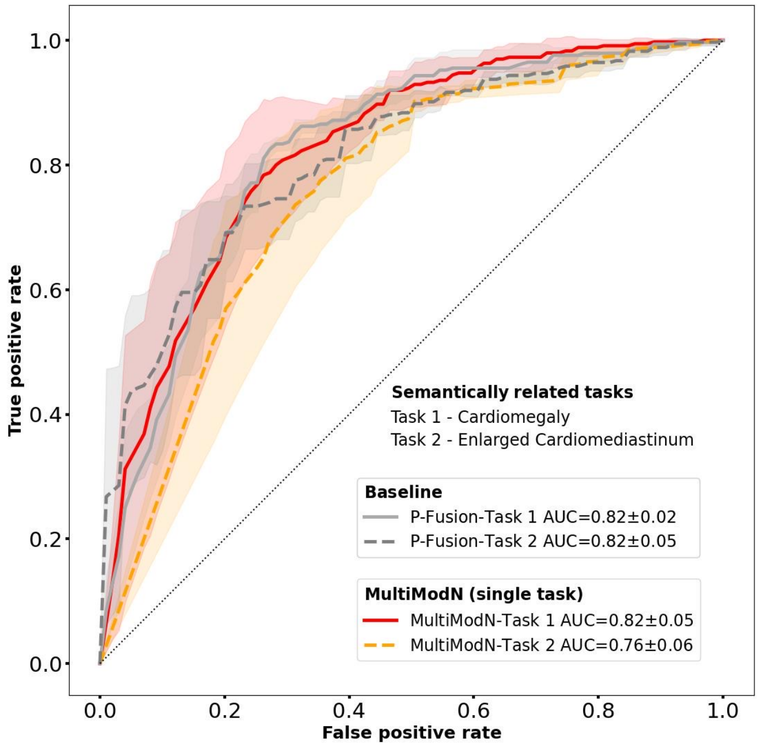
P-Fusion

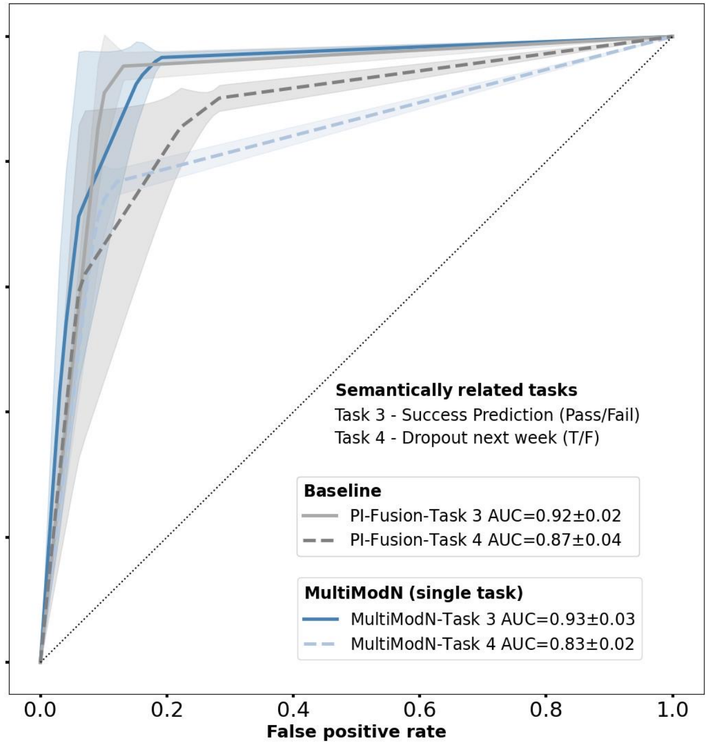


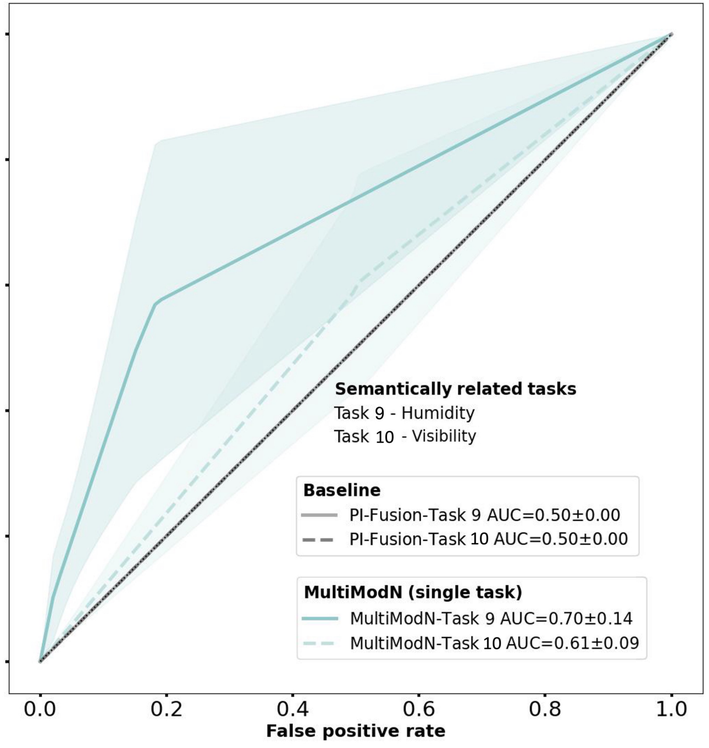
fusion in parallel

Performance

MULTIMODN VS. P-FUSION





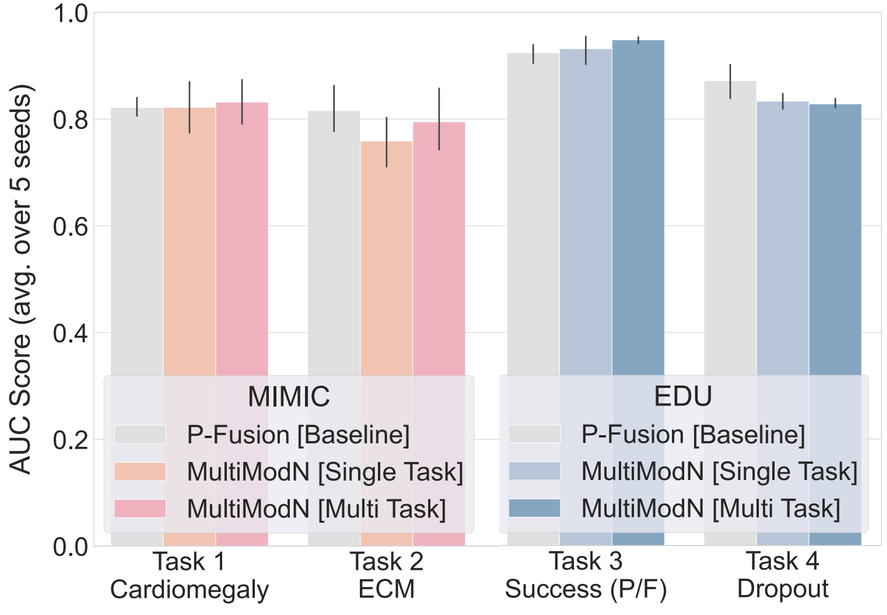


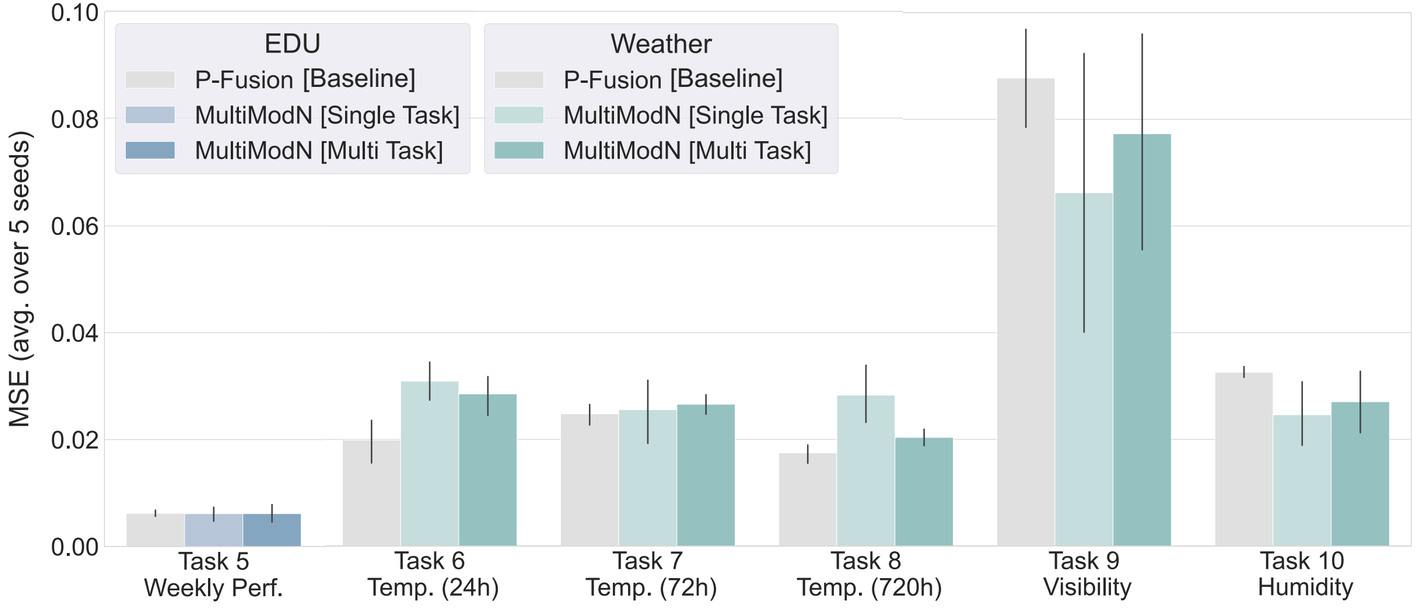
Evaluated on 10 real world tasks across multimodal healthcare (MIMIC), education (MOOC), and weather (Weather2k) benchmarks

Performance

MULTIMODN VS. P-FUSION

MultiModN is naturally extensible to the prediction of multiple tasks **without** negatively impacting the performance of individual tasks

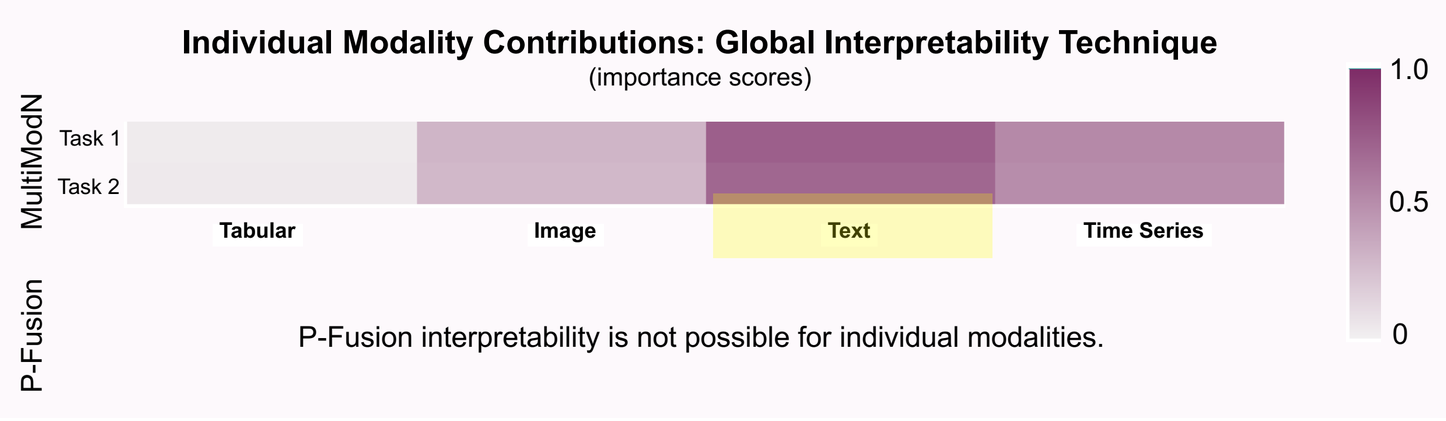


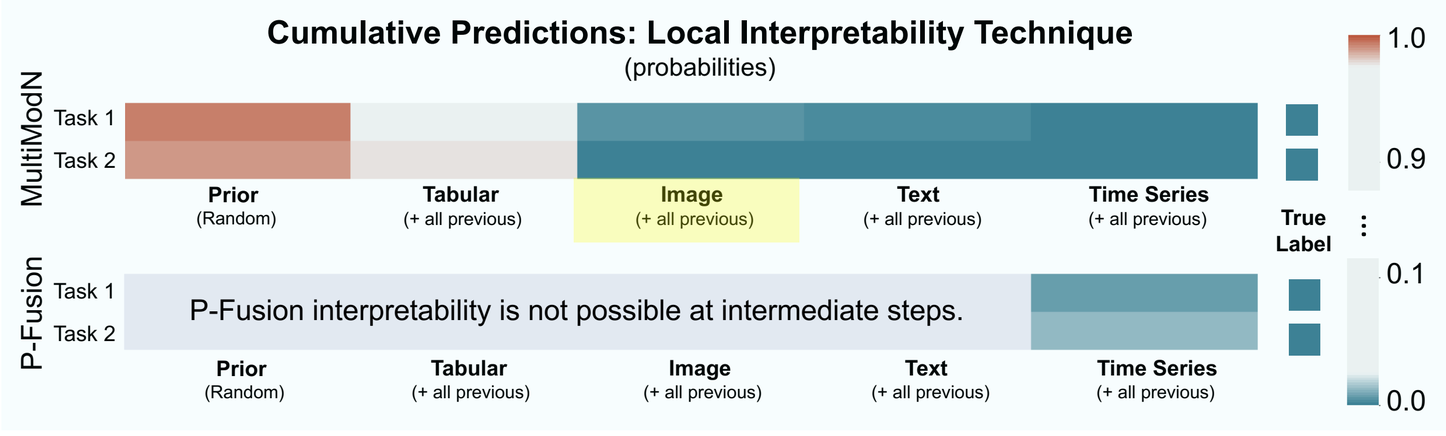


Interpretable-by-design

MULTIMODN VS. P-FUSION

Which modality is the most important for this task?





Which modality caused the decision to flip?

MultiModN has modality-specific global (IMC) and local (CP) model explainability.

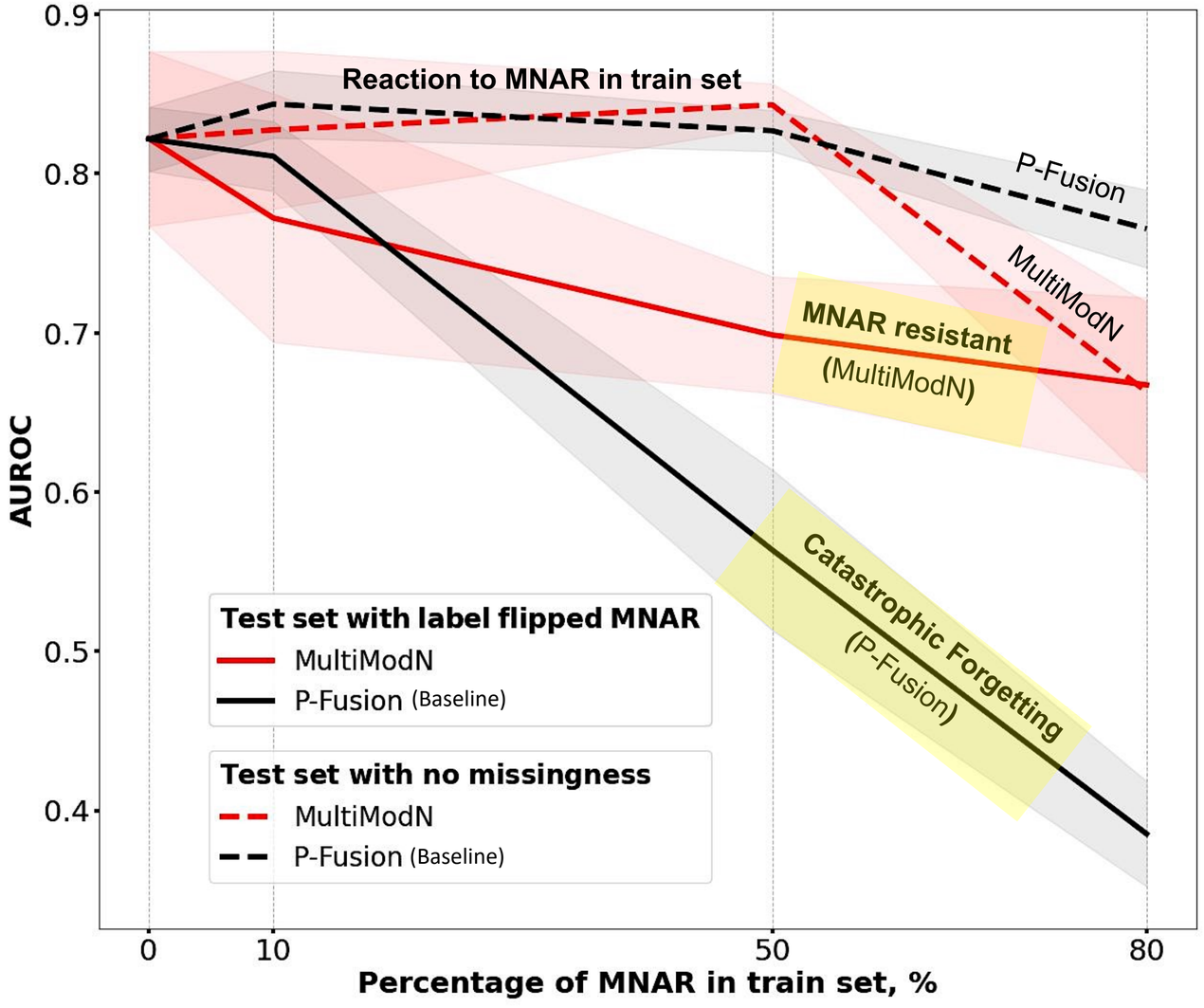
Robust to Missingness

MULTIMODN VS. P-FUSION

MultiModN is robust to bias from missing input modalities.

P-Fusion exhibits catastrophic MNAR failure when missingness patterns change.

Especially relevant in low-resource settings.



MultiModN

CONTRIBUTIONS

1) matches parallel MM fusion (P-Fusion) for a range of real-world tasks

2) composable at inference

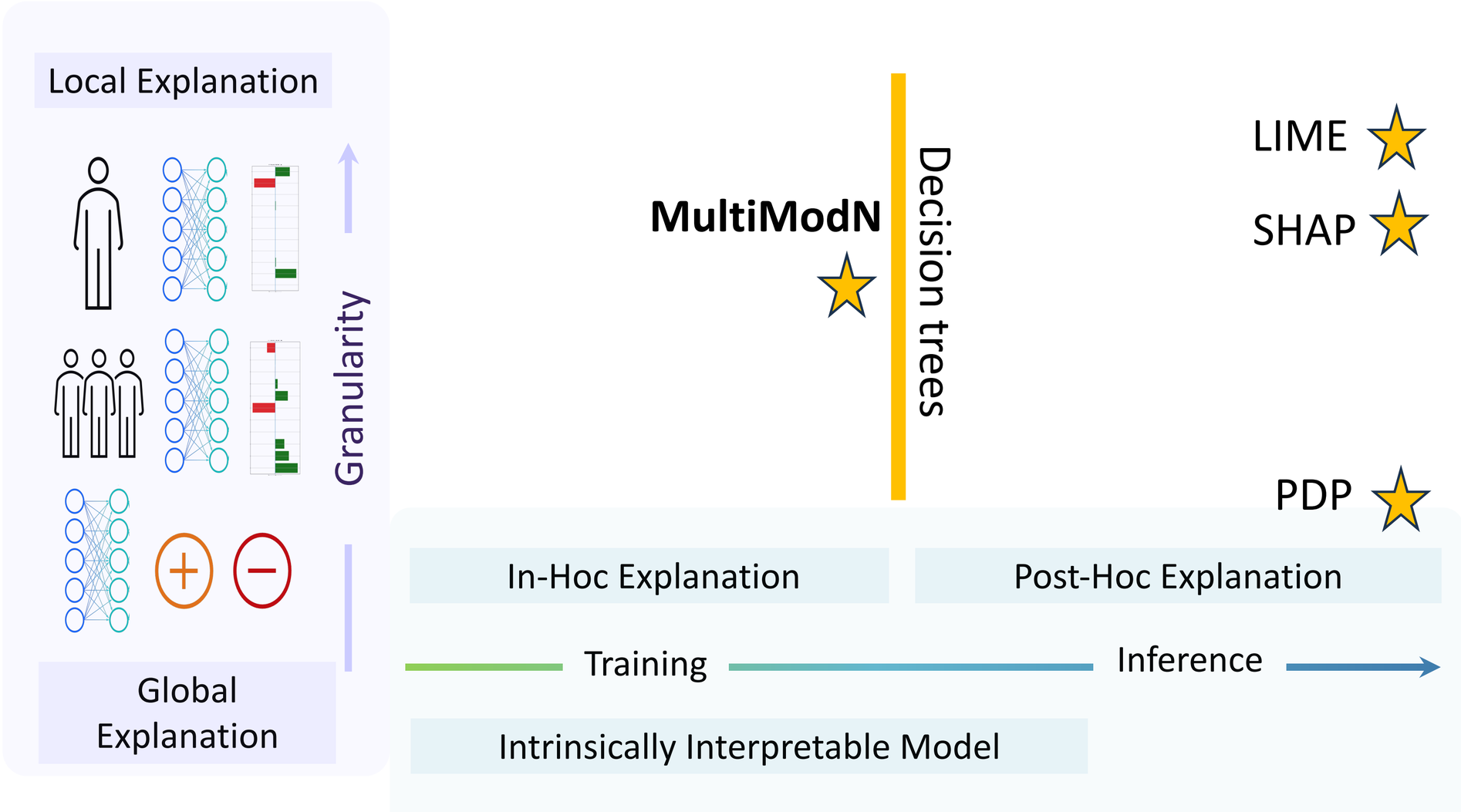
3) robust to the bias of missing not-at-random (MNAR) modalities

4) inherently interpretable

5) easily extended to any number or combination of tasks

Interpretability

XAI Fundamentals



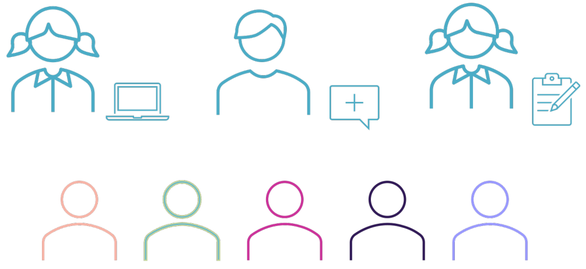
RIPPLE: In-Hoc Concept Interpretation for GNNs

[AAAI 2023] RIPPLE
Asadi, Swamy, Frej, Vignoud,
Marras, Käser

Can we obtain interpretability on raw multivariate time series?

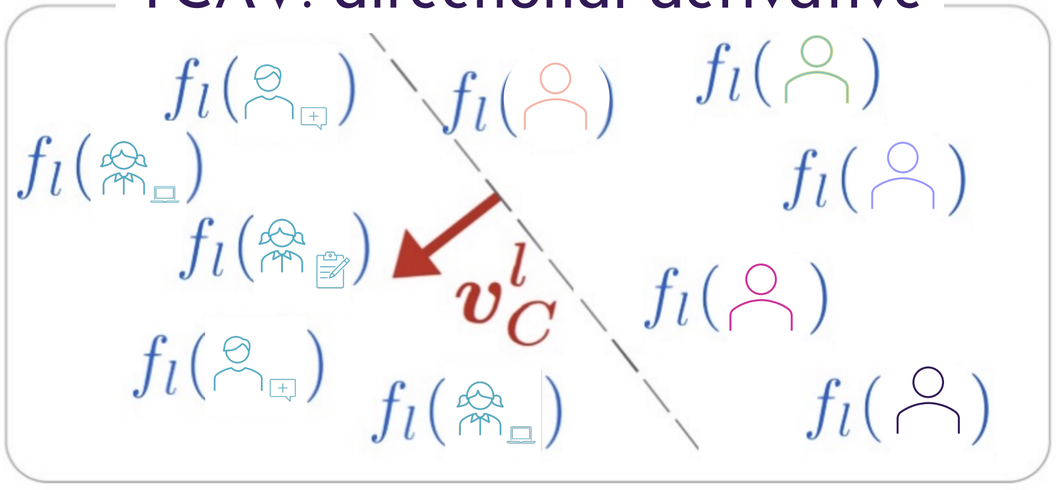
Concept Activation Vectors

concept: high effort



random students

TCAV: directional derivative

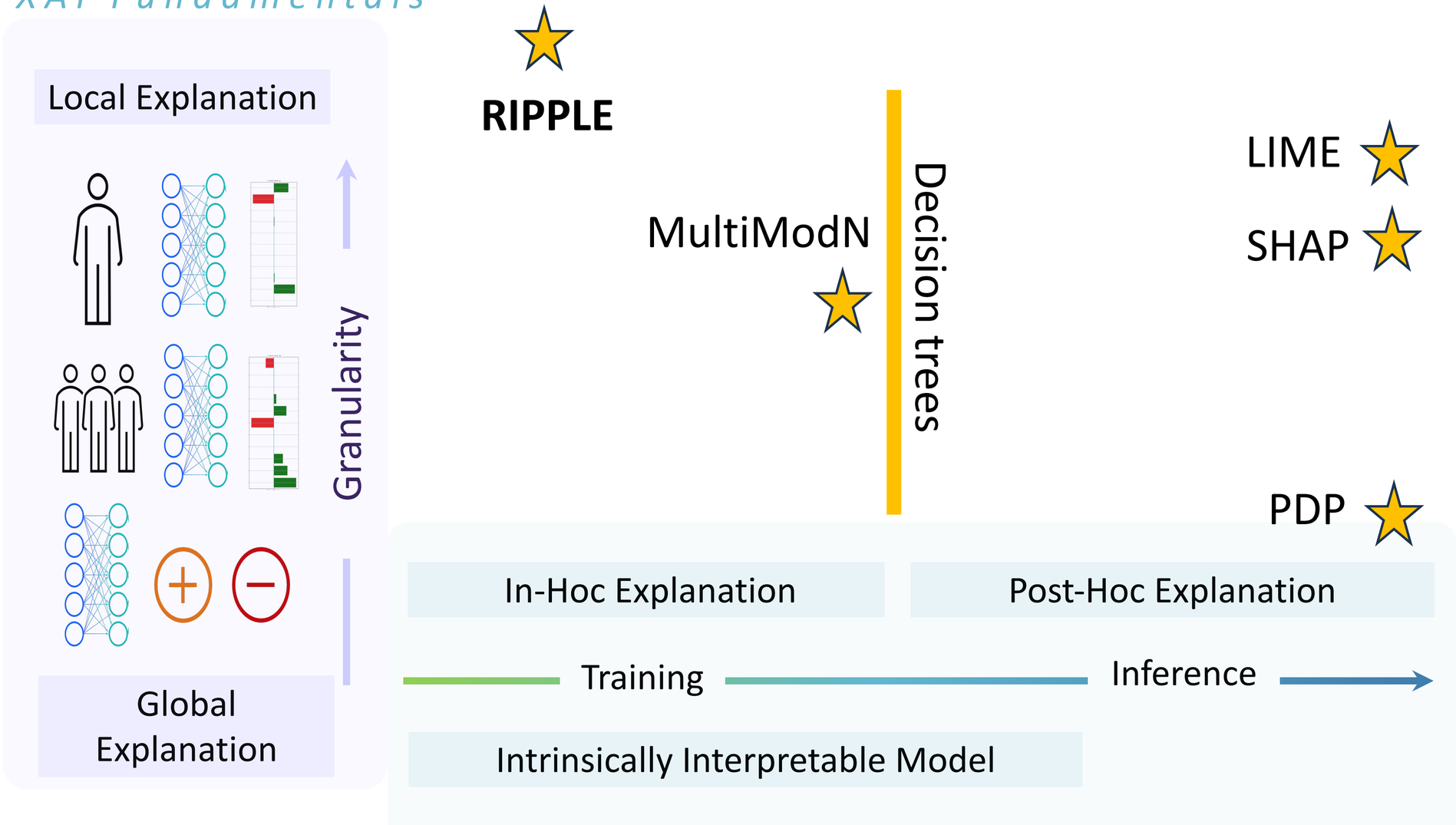


Advantages: global and local scale, user-specified example-based concepts, directly using model activations (accuracy)

Kim et al. ICML 2018

Interpretability

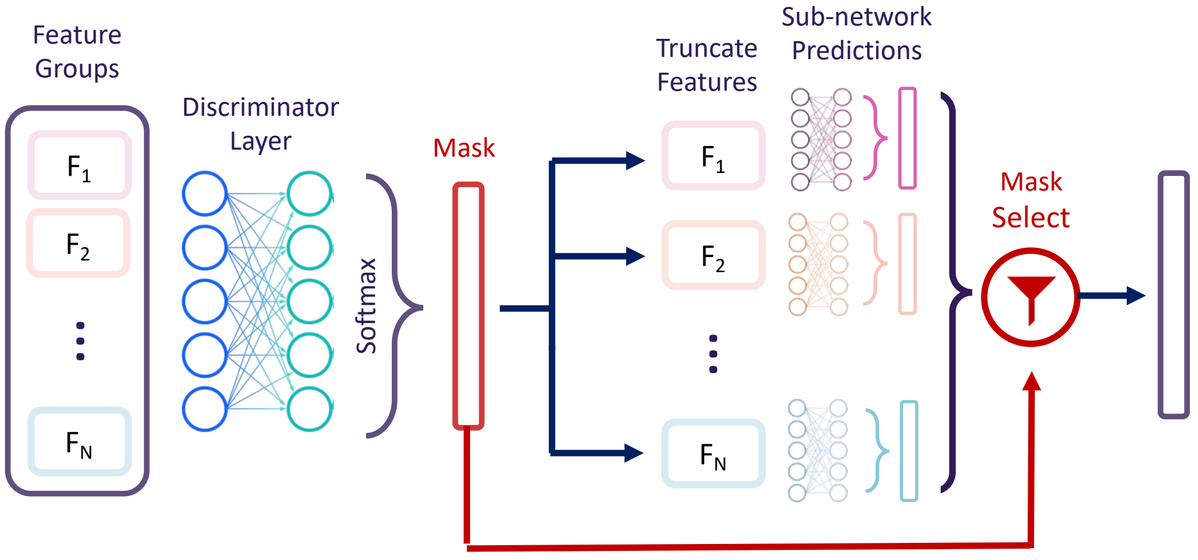
XAI Fundamentals



InterpretCC: Intrinsic User-Centric Interpretability through Global MoE

Swamy, Montariol Blackwell, Frej, Jaggi, Käser

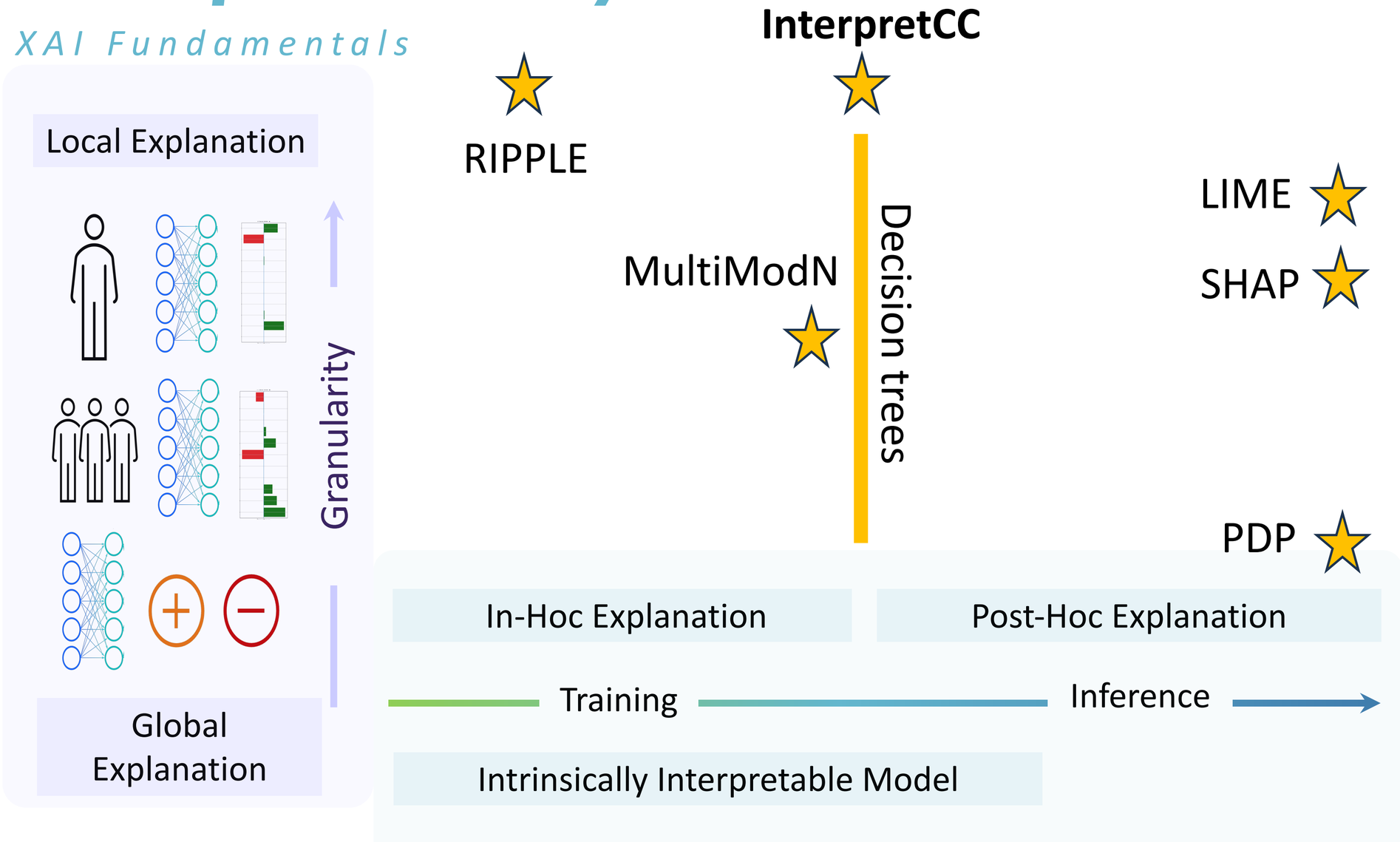
Adaptive Feature Gating – accuracy vs. interpretability tradeoff!



(based on Hinton's conditional computation)

Interpretability

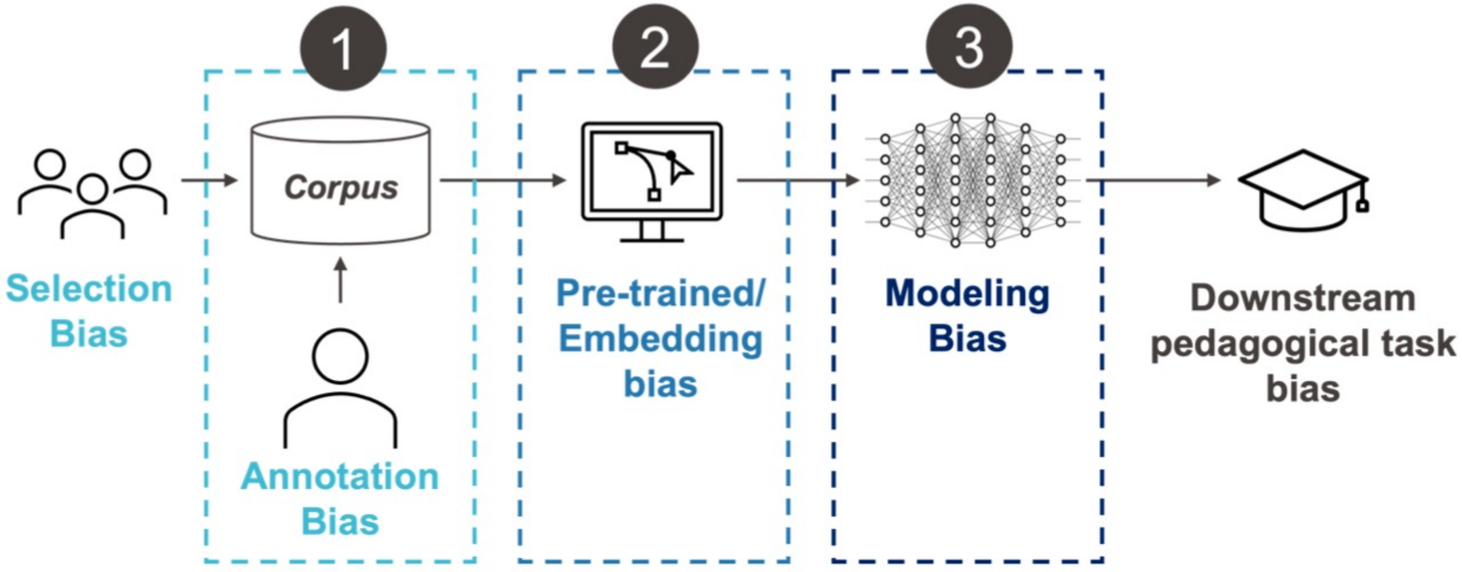
XAI Fundamentals



LLMs in the classroom

[COLING 2022, EMNLP Findings 2023]
Bias at a Second Glance
Wambsganss*, Swamy*
Rietsche, Käser

AI writing support: will bias from the model suggestions transfer to the students?



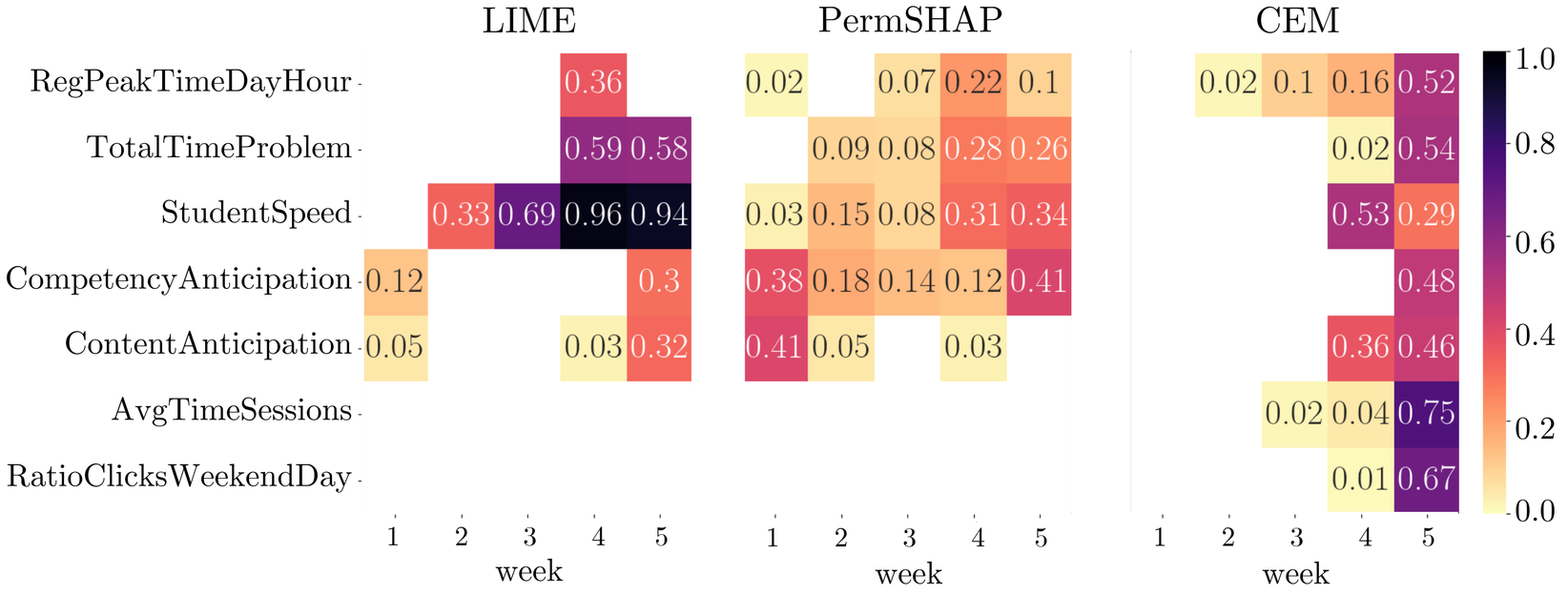
AI as a tutor: Using GPT 4 to provide written feedback on student assignments (across many courses at EPFL!)

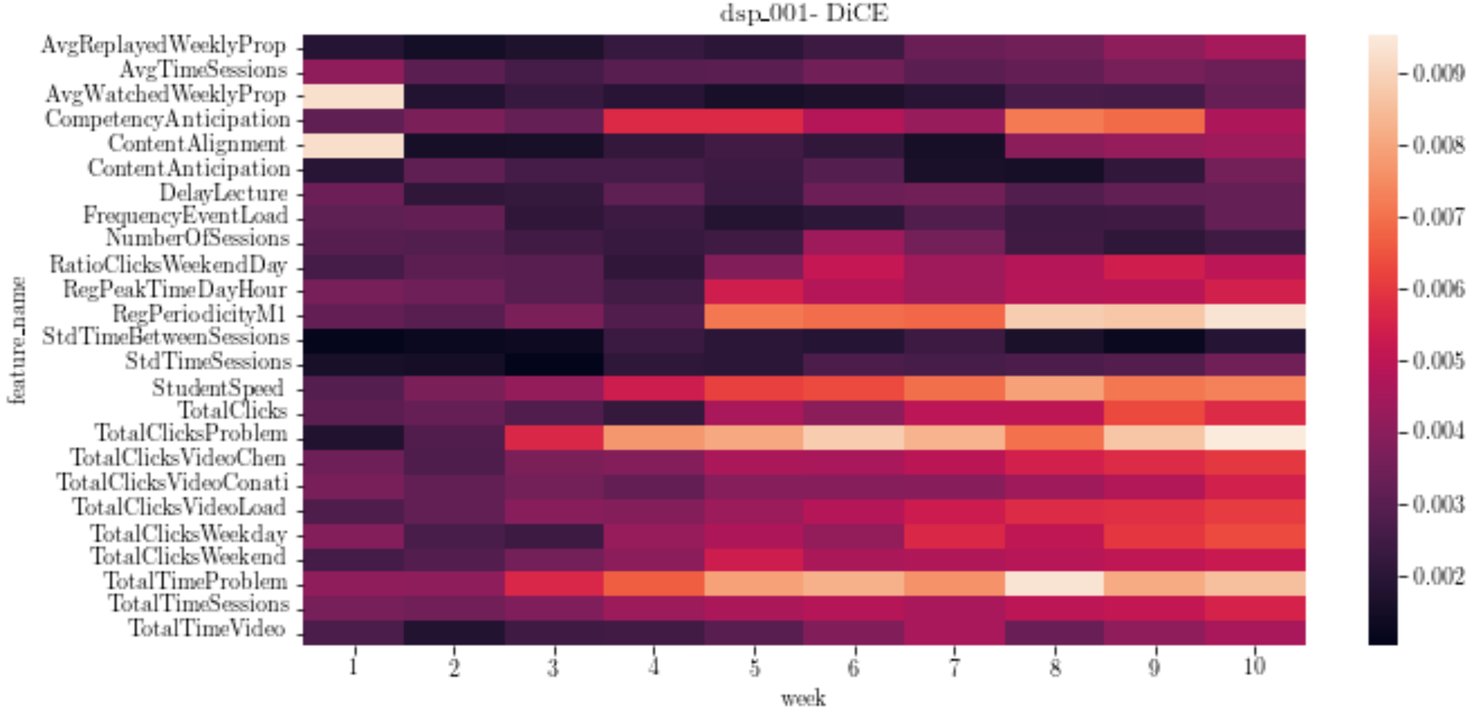
Processed Output of an Explainer

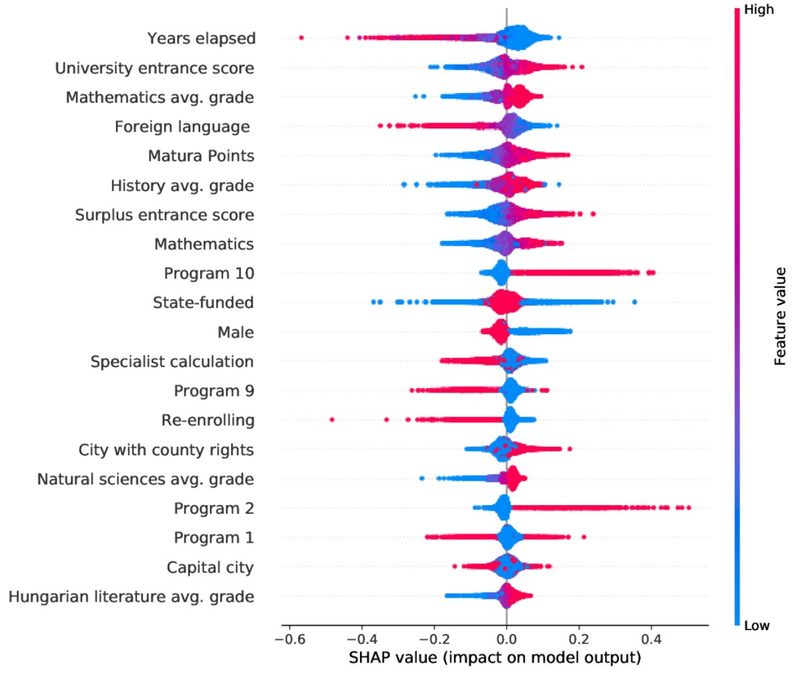
	exp_num	total_clicks_lnWeek1	number_sessions_lnWeek1	time_in__video_sum_lnWeek1	time_in__problem_sum_lnWeek1
66	1575	-4.56E-09	-2.21E-10	0.074035813	1.3699654119458948e-09
67	5252	-5.60E-10	-2.21E-10	2.1067425364992842e-11	-2.06E-10
68	881	1.7555813192071668e-09	-8.83E-10	0.14653082042868498	-2.65E-10
69	2683	-3.42E-09	-2.21E-10	0.24476348871729897	1.264484322804904e-09
70	16963	-8.50E-10	0	0.045331784	-1.61E-09
71	5931	6.354141102171695e-10	1.3576282653637861e-08	0.17988949383993513	-7.48E-10
72	12145	0	0	0.9422763586044312	0
73	11999	0	0	0.03497038	0
74	1571	6.909329608451031e-09	-8.83E-10	0.13568544287717593	1.6364350777231529e-09
75	2220	-5.60E-10	0	0.057424471615632494	0
76	9184	0	0	0.025690399	0
77	9592	-2.28E-09	-2.21E-10	0	-1.84E-10
78	12309	5.789162378644352e-10	-2.21E-10	0	-4.92E-10
79	5394	-2.28E-09	0	0.026910097	0
80	730	-2.28E-09	-2.21E-10	0.029592504314488427	8.244599122332608e-13
81	3318	2.315664951457741e-09	6.7881413268189306e-09	0.014542981	-1.48E-09
82	290	-1.12E-09	0	-3.32E-09	0
83	1959	5.789162378644352e-10	0	1.8329978368480937e-09	4.3817753245245505e-10
84	1424	-1.70E-09	-4.42E-10	-4.82E-09	4.4265635254503444e-10
85	12925	0	0	0.021664523	0
86	11139	0	0	0	0
87	12603	0	0	0	0
88	16987	1.1578324757288705e-09	-2.21E-10	0	-1.46E-10
89	10284	0	0	0	0
90	1316	1.7555813192071668e-09	-8.83E-10	0.015449408282339427	-9.00E-10
91	2601	1.447290594661088e-10	0	0	0
92	2851	-2.84E-09	-4.42E-10	0.037342517841102124	0
93	16707	2.894581189322176e-10	0	0	9.635302451738159e-11
94	11108	0	0	0	0
95	5846	1.1578324757288705e-09	-2.21E-10	-3.86E-11	0
96	16538	1.447290594661088e-10	0	0	0
97	14230	0	0	0	0

10 weeks x 45 fts
= 450 columns

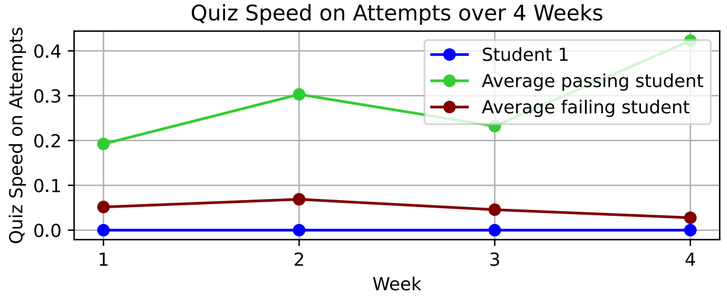
Heavily Processed Visuals

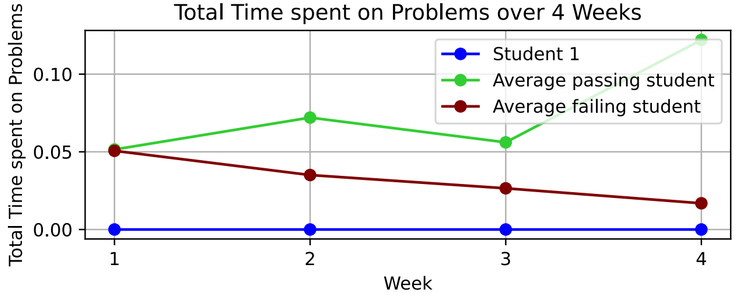


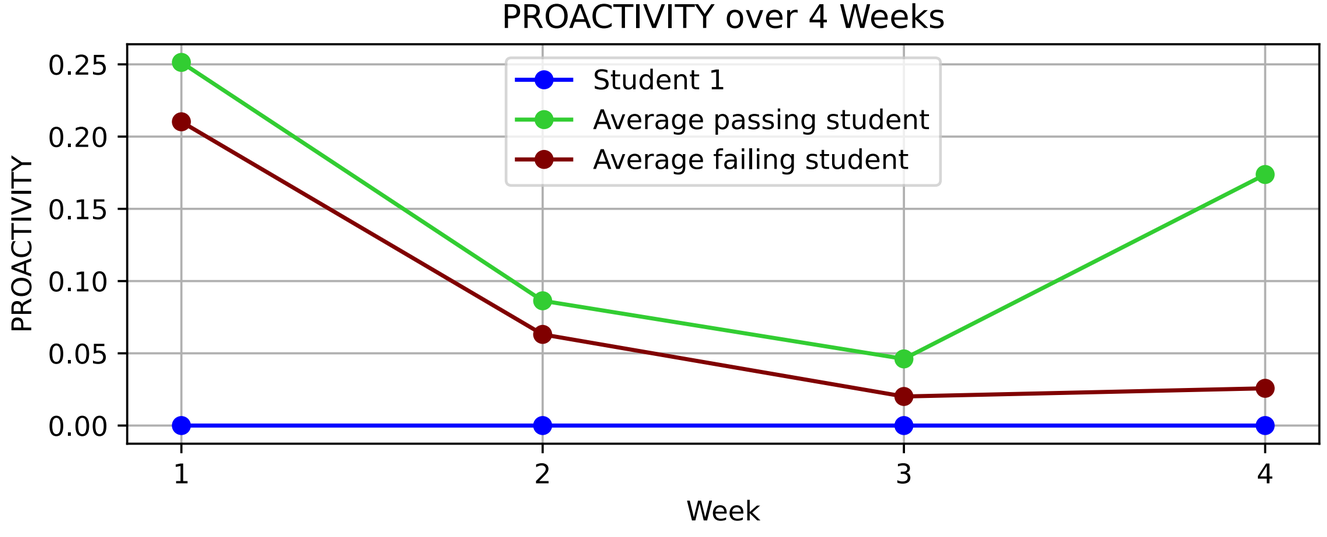


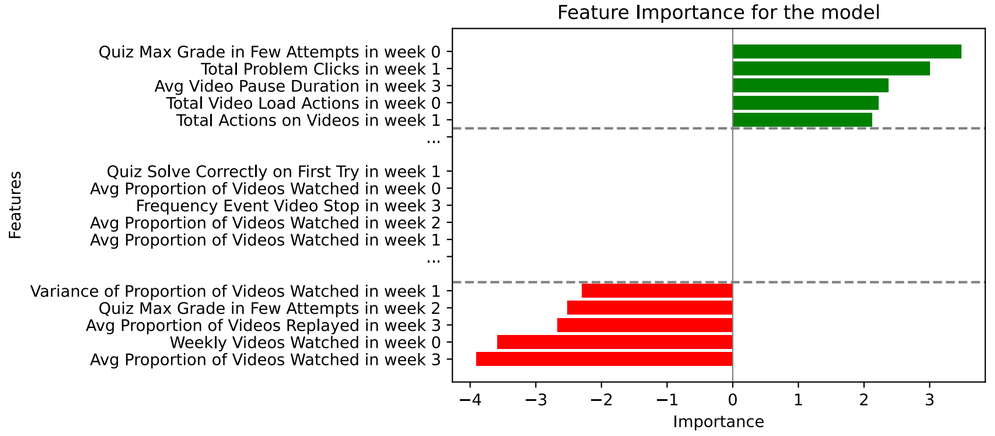


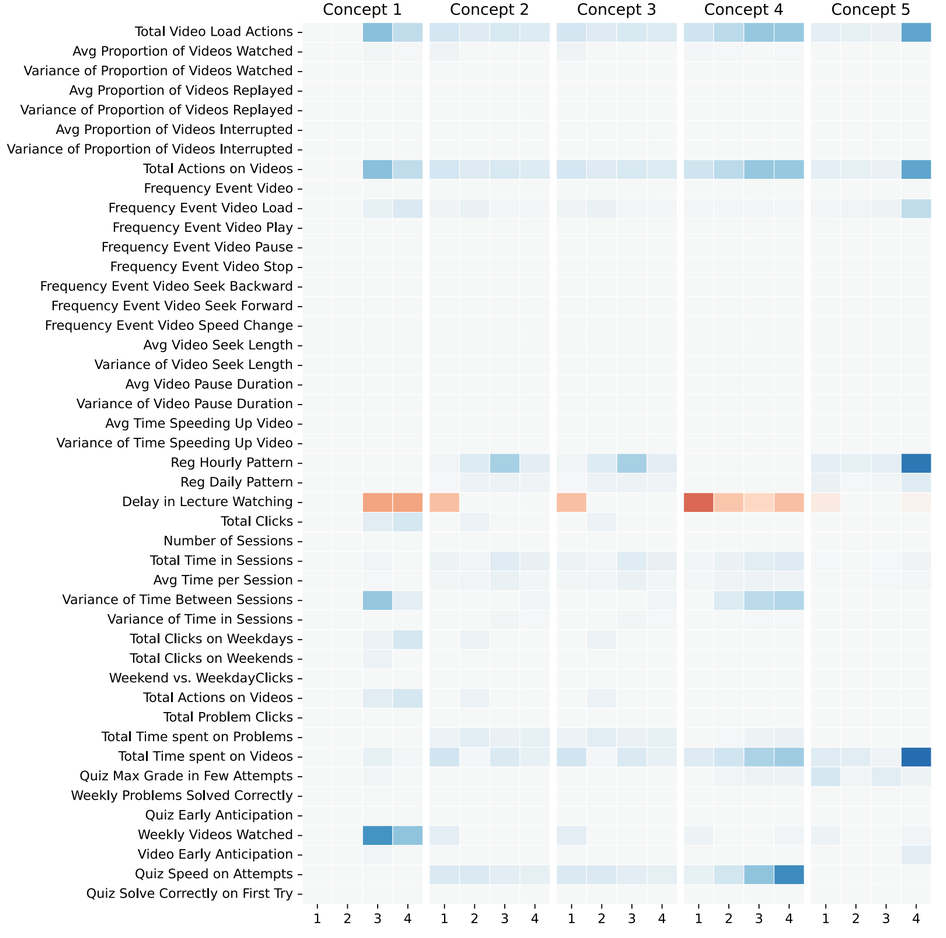
Heavily Processed Visuals

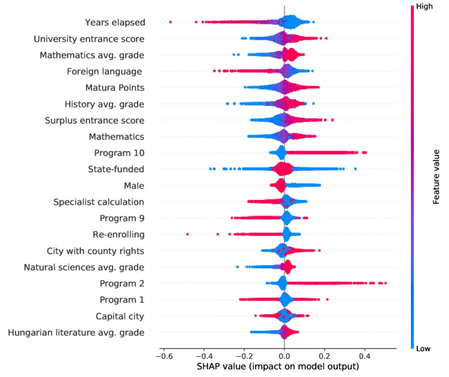


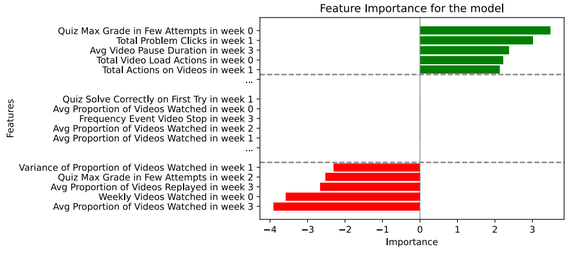






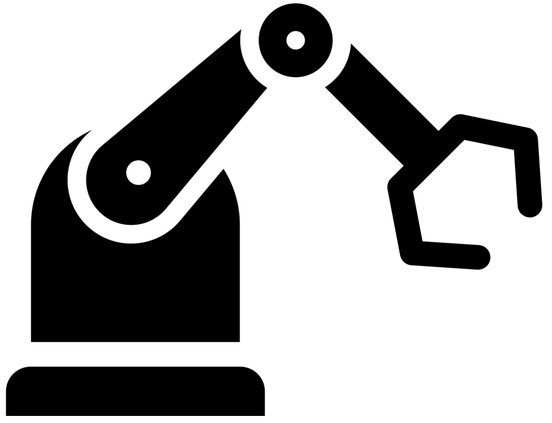


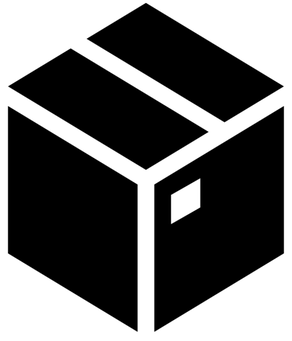




	exp_num	total_clicks_lnWeek1	number_sessions_lnWeek1	time_in_..video_sum_lnWeek1	time_in_..problem_sum_lnWeek1
66	1575	-4.56E-09	-2.21E-10	0.074035813	1.369965411945858e-09
67	5252	-5.60E-10	-2.21E-10	5.106742536496928420e-11	-2.06E-10
68	881	1.755581319207166 8e-09	-8.83E-10	0.146530820428684 98	-2.65E-10
69	2683	-3.42E-09	-2.21E-10	0.244763488717298 97	1.264484322804904e-09
70	16963	-8.50E-10	0	0.045331784	-1.61E-09
71	5931	6.354141102171695e-10	1.357628265363786 1e-08	0.179889493839935 13	-7.48E-10
72	12145	0	0	0.942276358604312	0
73	11999	0	0	0	0.03497038
74	1571	6.909329608451031e-09	-8.83E-10	0.135685442877179 3	6.636435077723152 9e-09
75	2220	-5.60E-10	0	0.057424471615632494	0
76	9184	0	0	0.025690399	0
77	9592	-2.28E-09	-2.21E-10	0	-1.84E-10
78	12309	5.789162378644352e-10	-2.21E-10	0	-4.92E-10
79	5394	-2.28E-09	0	0.026910097	0
80	730	-2.28E-09	-2.21E-10	0.029592504314488427	5.244599122332608e-13
81	3318	2.315664951457741e-09	5.788141326818930 6e-09	0.014542981	-1.48E-09
82	290	-1.12E-09	0	-3.32E-09	0
83	1959	5.789162378644352e-10	0	1.832997836848093 7e-09	1.381775324524550 5e-10
84	1424	-1.70E-09	-4.42E-10	-4.82E-09	4.426563525450344 4e-10
85	12925	0	0	0.021664523	0
86	11139	0	0	0	0
87	12803	0	0	0	0
88	16987	1.157832475728870 5e-09	-2.21E-10	0	-1.46E-10
89	10284	0	0	0	0
90	1316	1.755581319207166 8e-09	-8.83E-10	5.015449408282339427	-9.00E-10
91	2601	1.447290594661088e-10	0	0	0
92	2851	-2.84E-09	-4.42E-10	0.037342517841102124	0
93	16707	2.894581189322176e-10	0	0	5.635302451738159e-11
94	11108	0	0	0	0
95	5846	1.157832475728870 5e-09	-2.21E-10	-3.86E-11	0
96	16538	1.447290594661088e-10	0	0	0
97	14230	0	0	0	0

Select what's important



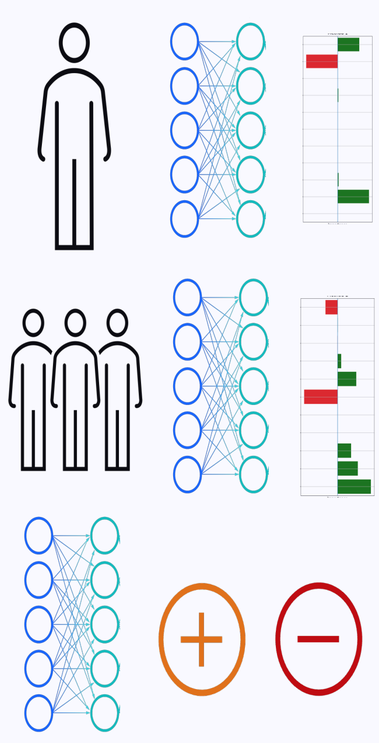


Package it up nicely

students are happy*!

*with their improved learning outcomes

Local Explanation



Granularity

Global Explanation

this project

local explanations

any method

In-Hoc Explanation

Post-Hoc Explanation

Training

Inference

Intrinsically Interpretable Model

Thank you!

Vinitra Swamy

vinitra.swamy@epfl.ch

github: epfl-ml4ed