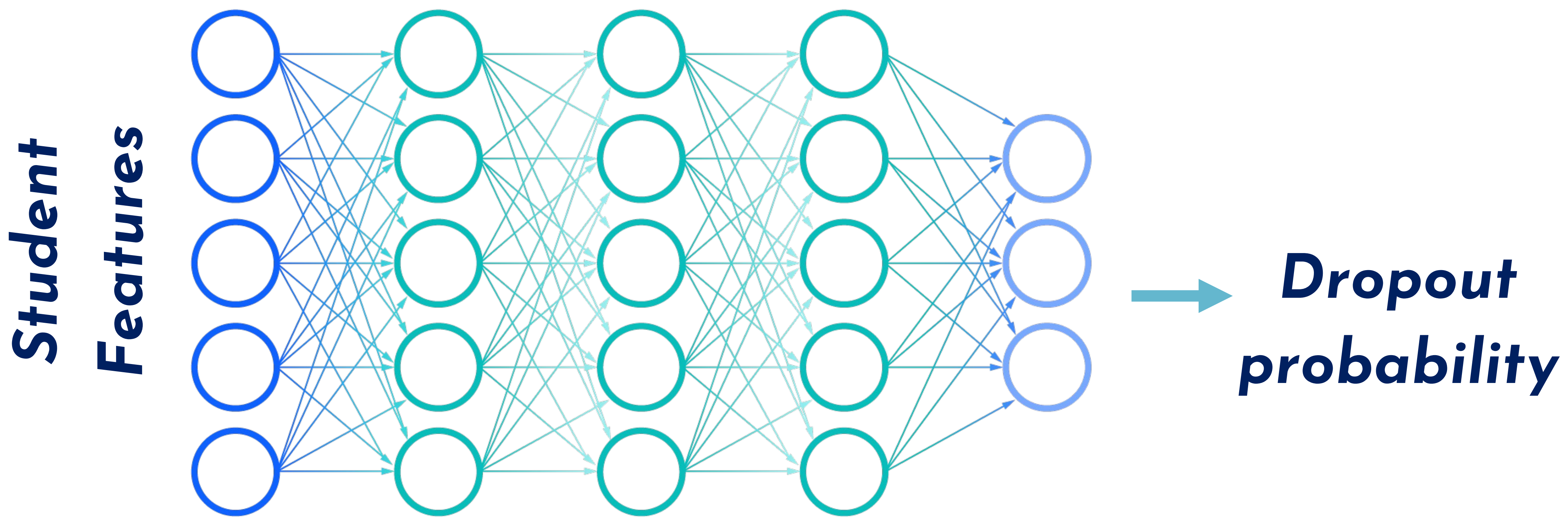
EDUCATIONAL DATA MINING 2022

Evaluating the Explainers: Black Box Explainable ML for Student Success Prediction in MOOCs

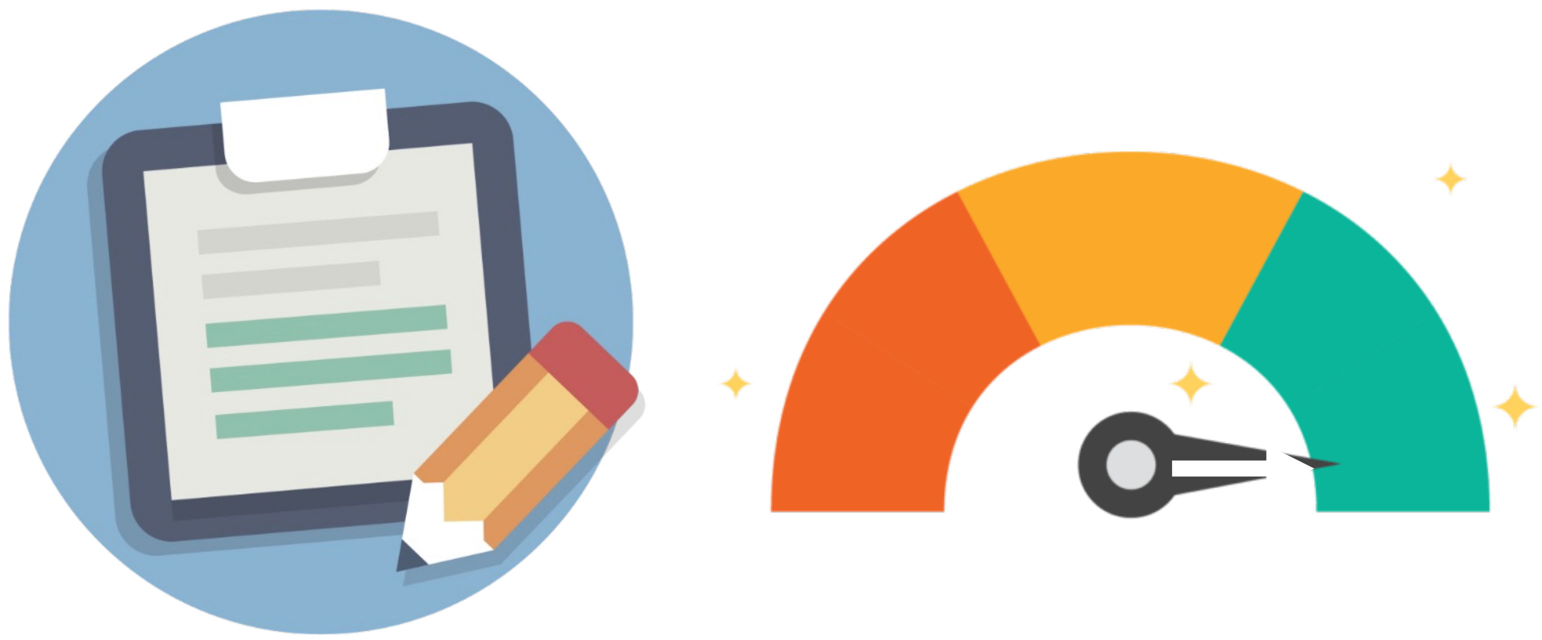
VINITRA SWAMY, BAHAR RADMEHR, NATASA KRCO, MIRKO MARRAS, TANJA KÄSER

Deep Learning has been increasingly researched in digital learning environments

(MOOCs) Dropout Prediction



(LMS) Autograding, Plagiarism detection



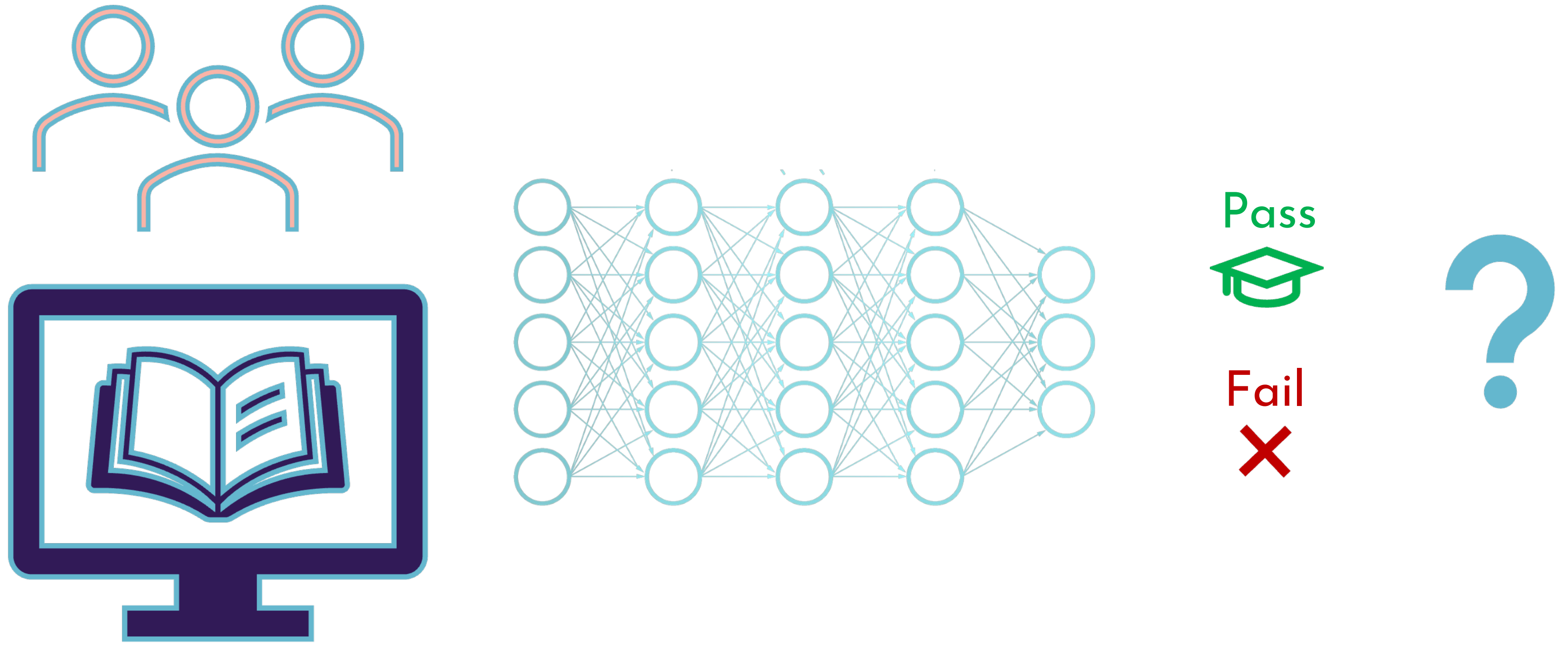
(OELEs) Student Knowledge Tracing

Imran et al. ICCAI 2019; Xing and Du, Journal of Edu Computing Research 2019; Piech et al. NeurIPS 2015.

Cost of using neural networks

DEEP LEARNING IN EDUCATION

Problem: Deep Learning trades transparency for accuracy



Identifying “why” is important for effective, personalized interventions

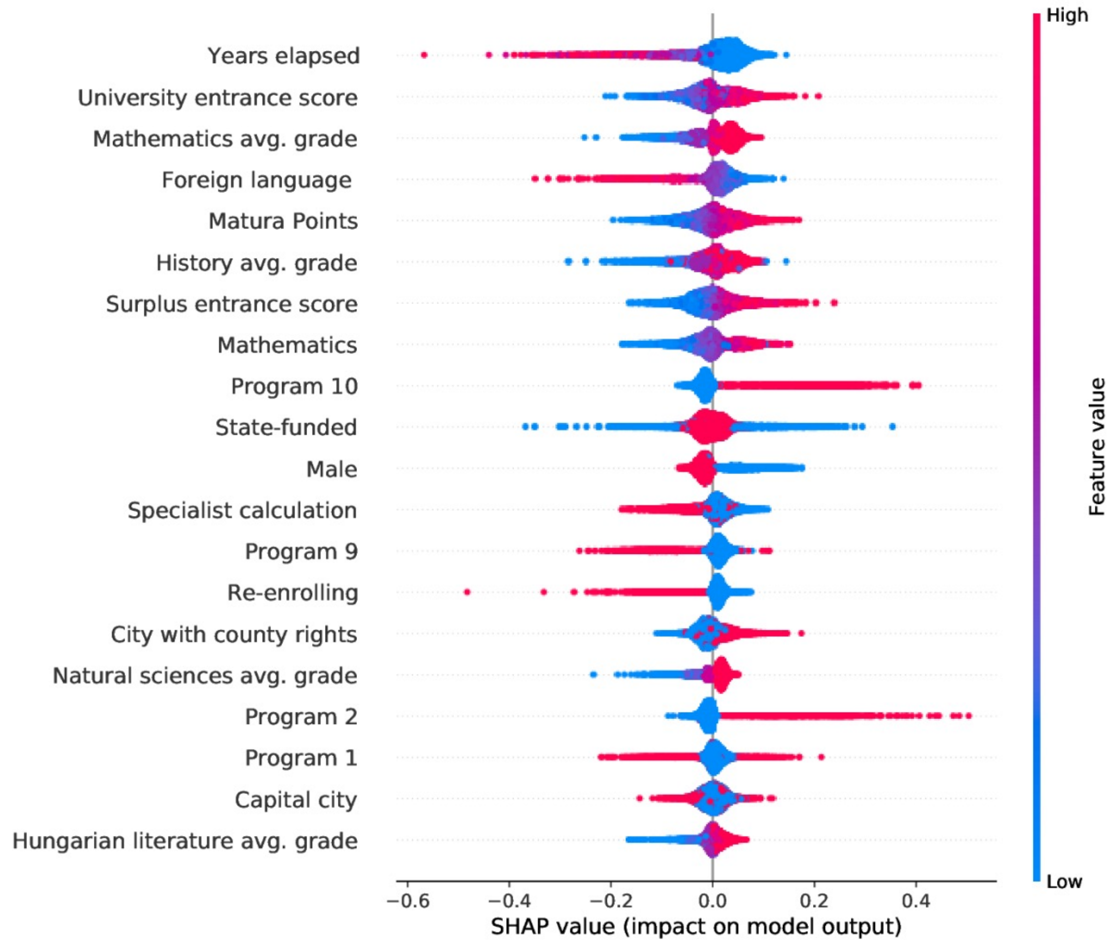
Solution: Explainable Machine Learning

Previous Work

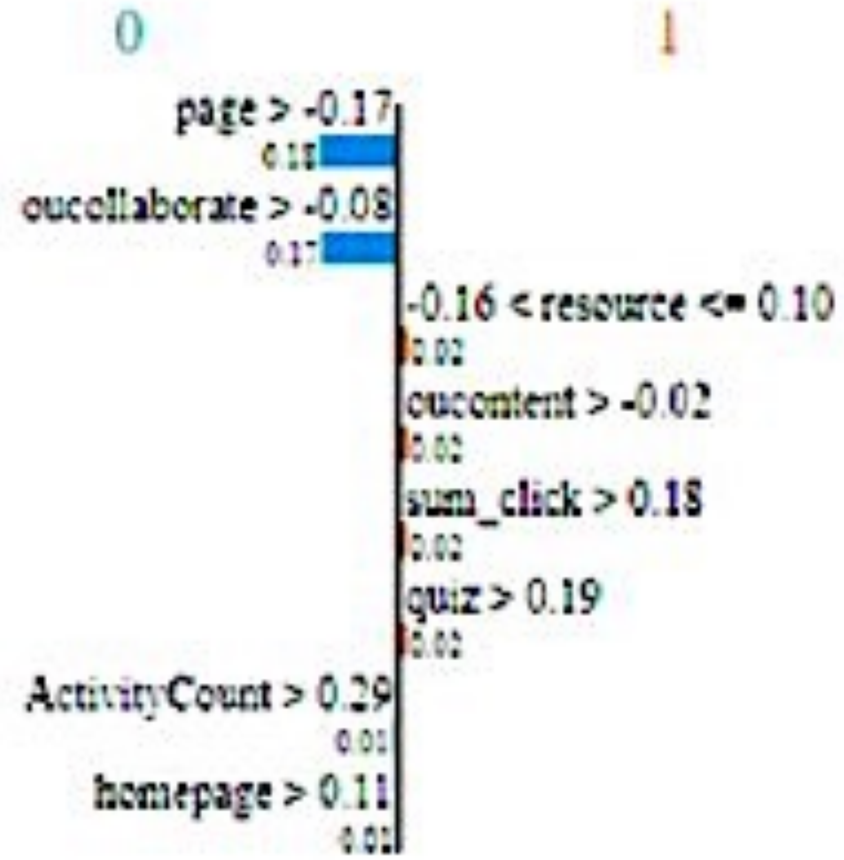
MOTIVATION

Previous work: In (minimal) related literature, only one explainability method is picked per ML for Edu paper

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

Objectives

MOTIVATION

The objective of this paper is therefore
to evaluate strengths and weaknesses of explainable
AI methods across 5 models

5 diverse courses

5 different methods

Dataset: 20,000 MOOC enrollments, hundreds of thousands of interactions

EPFL coursera

Research Questions

MOTIVATION

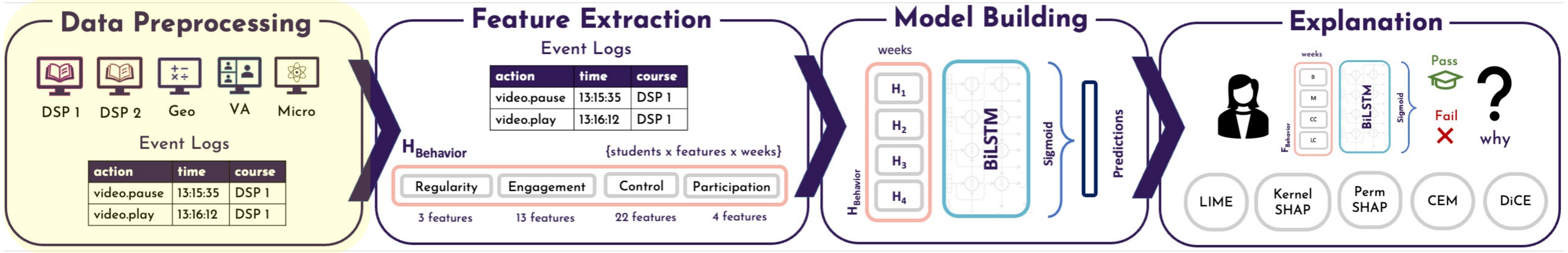
1) How similar are the explanations of different explainability methods for a specific course?

2) How do explanations (quantitatively) compare across courses?

3) Do explanations align with prerequisite relations in a course curriculum?

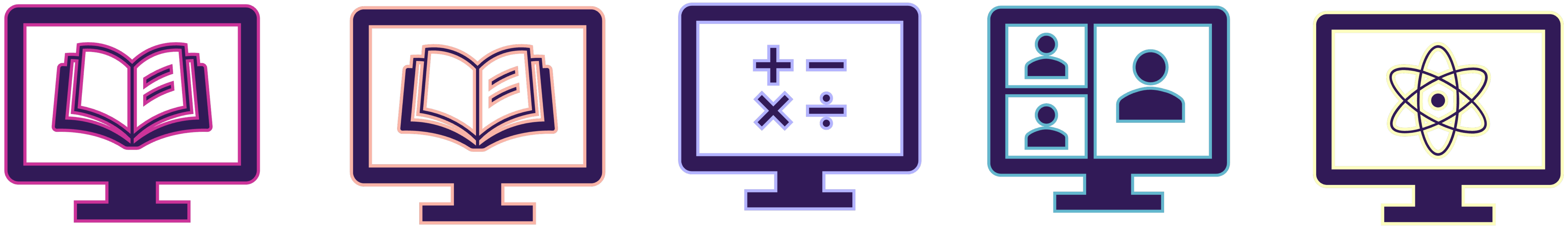
Pipeline

METHODOLOGY



Pipeline

METHODOLOGY



Digital Signal Processing 1

Digital Signal Processing 2

Villes Africaines

Geomatique

Microcontrôleurs

Languages: English / French

Student Level: BSc / MSc

 # Students: 452 - 5.6k

 # Weeks: 10 - 15

Pass Ratio: 5% - 45%

 # Quizzes: 17 - 27

Pipeline

METHODOLOGY

Data Preprocessing

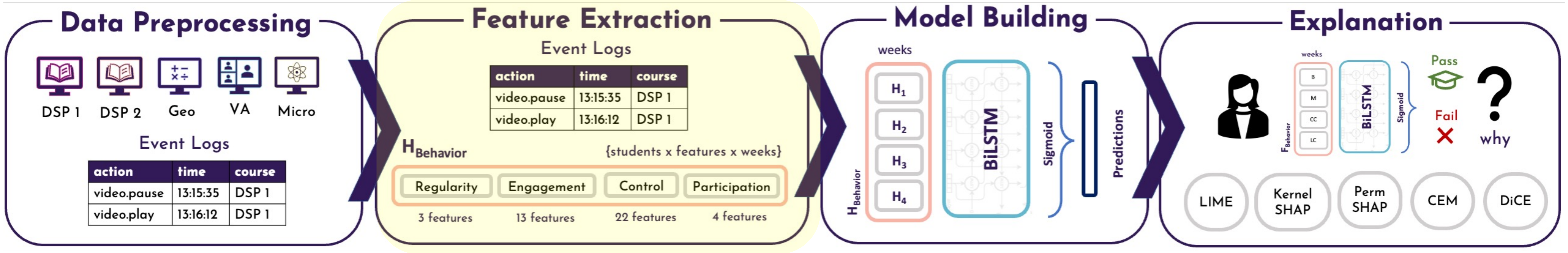
Event Logs

action	time	course
video.pause	13:15:35	DSP 1
video.play	13:16:12	DSP 1

Easy-to-Predict: Filter out easy-to-predict failing students, as there is no need for a complex model if a LogReg is sufficient!

Pipeline

METHODOLOGY



Pipeline

METHODOLOGY

Feature Extraction

Event Logs

action	time	course
video.pause	13:15:35	DSP 1
video.play	13:16:12	DSP 1

$H_{Behavior}$ {students x features x weeks}

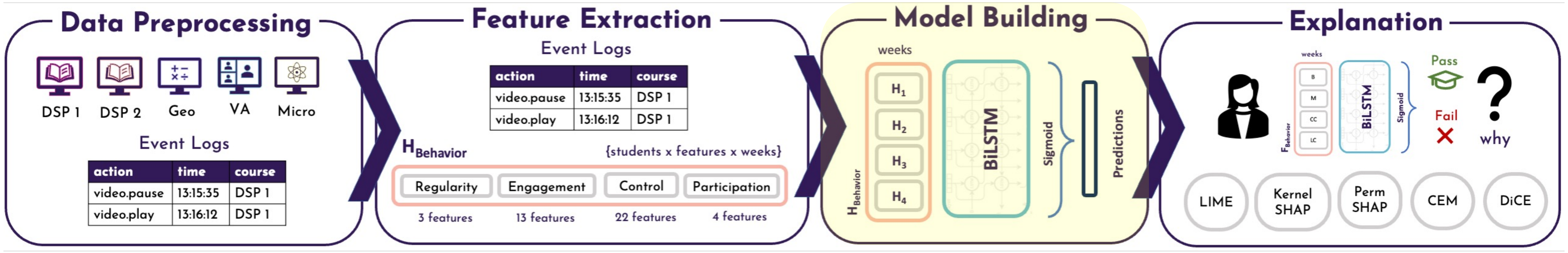
Regularity	Engagement	Control	Participation
3 features	13 features	22 features	4 features

All features are derived from previous work.

(Boroujeni et al., Marras et al., Chen Cui, Lalle Conati)

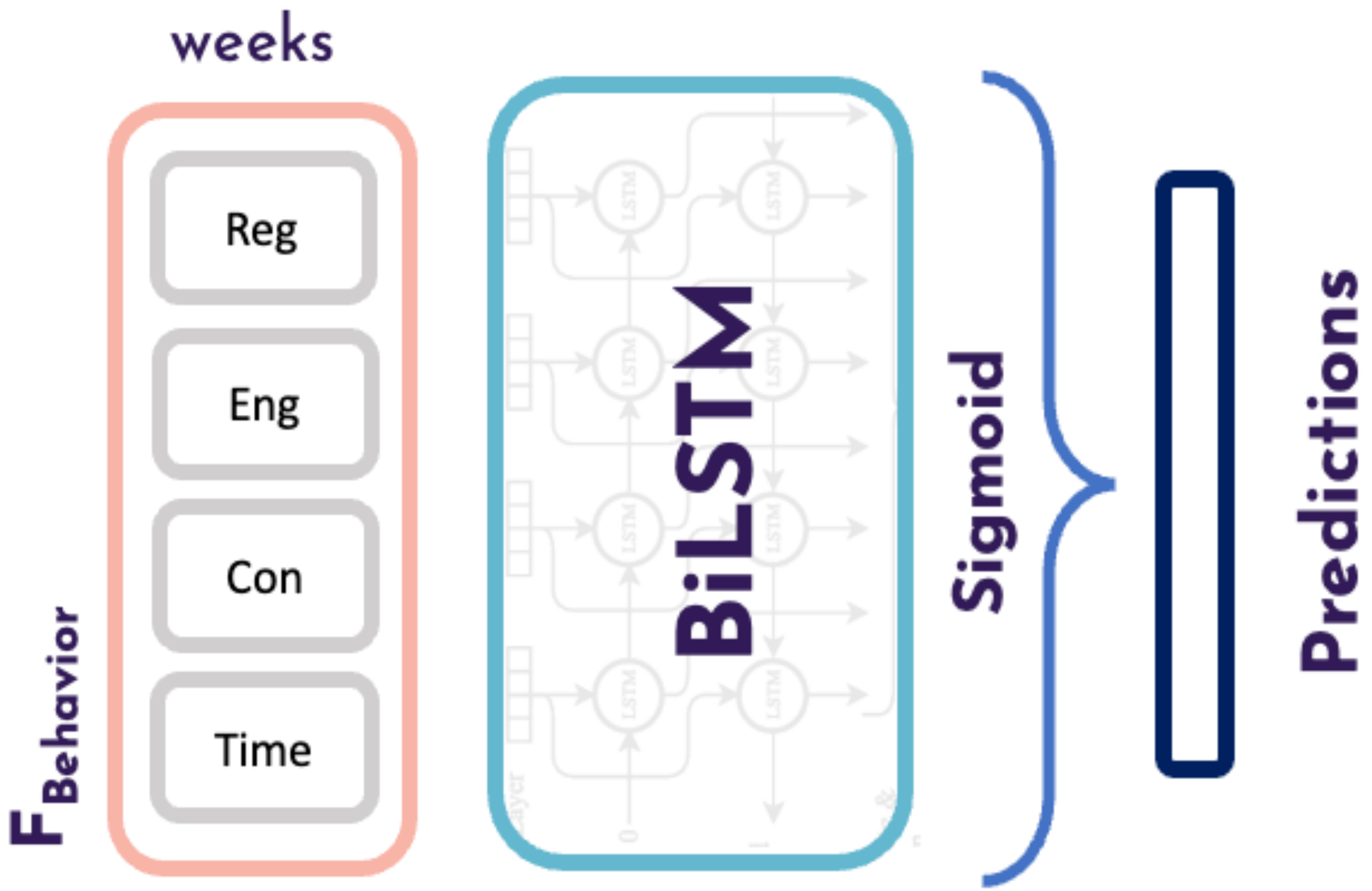
Pipeline

METHODOLOGY



Pipeline

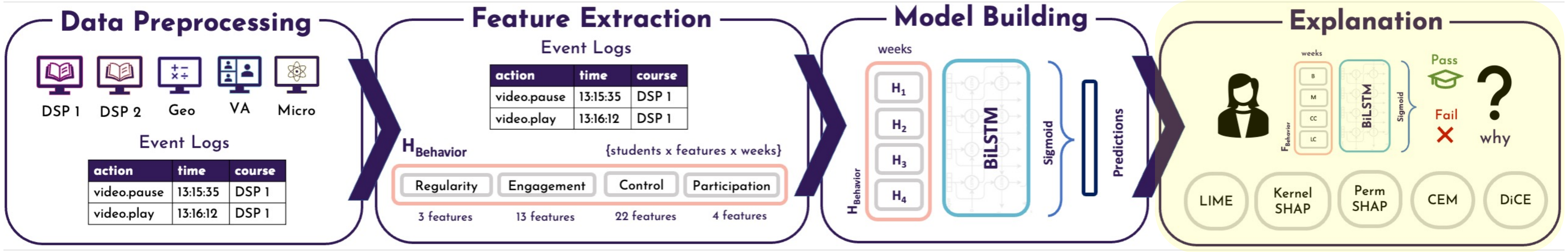
METHODOLOGY



Student Performance Prediction

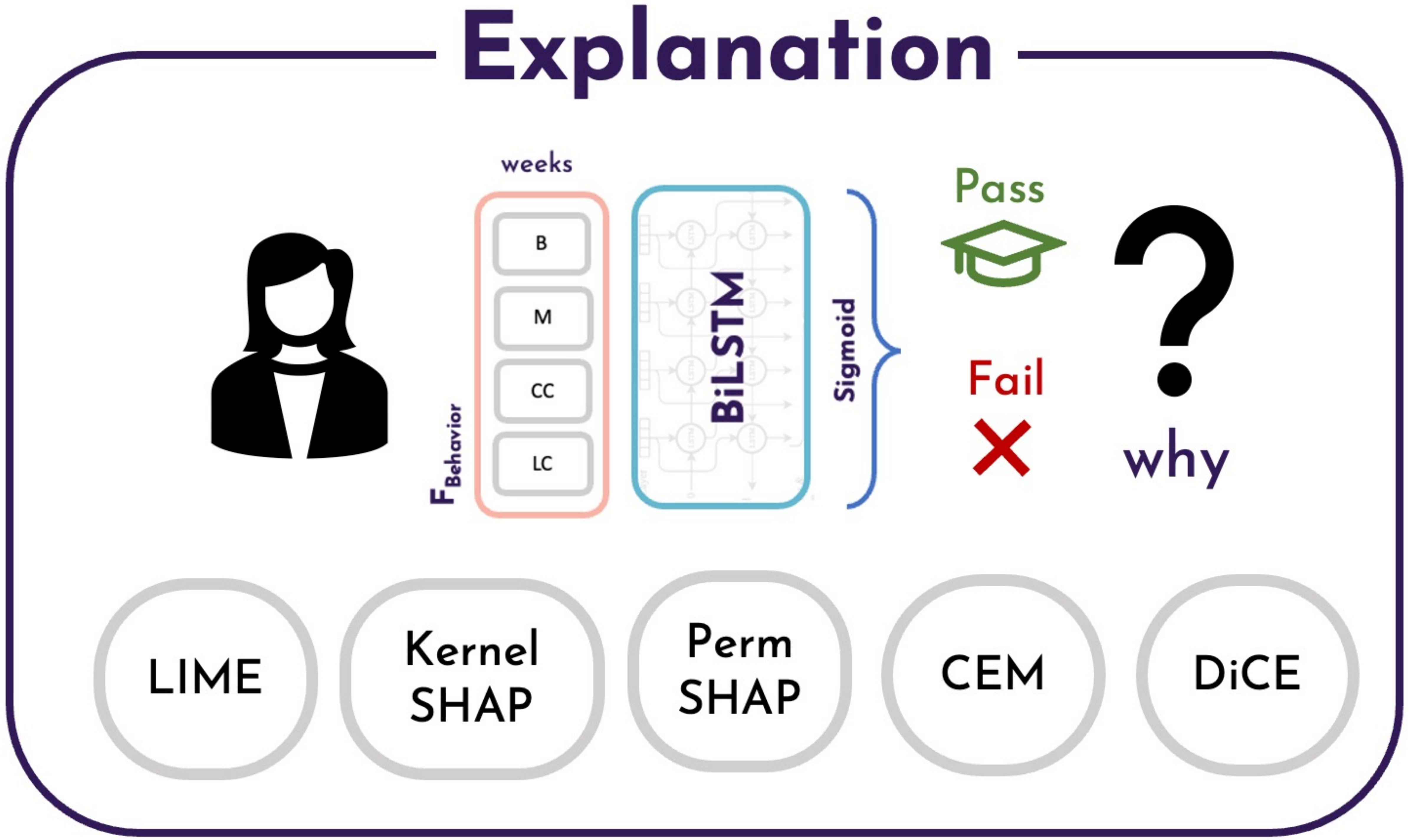
Pipeline

METHODOLOGY



Pipeline

METHODOLOGY



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

Pipeline

METHODOLOGY

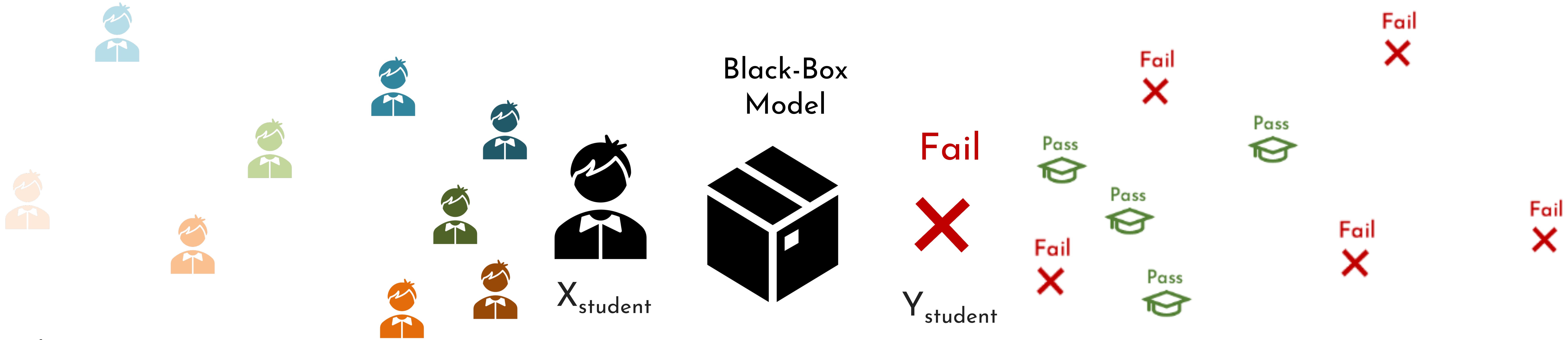
LIME

Local Interpretable Model-Agnostic Explanations

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

2. Perturb features of selected point to get $\{X^1_{student} \dots X^N_{student}\}$ neighbors

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



Ribeiro et al., KDD 2016.

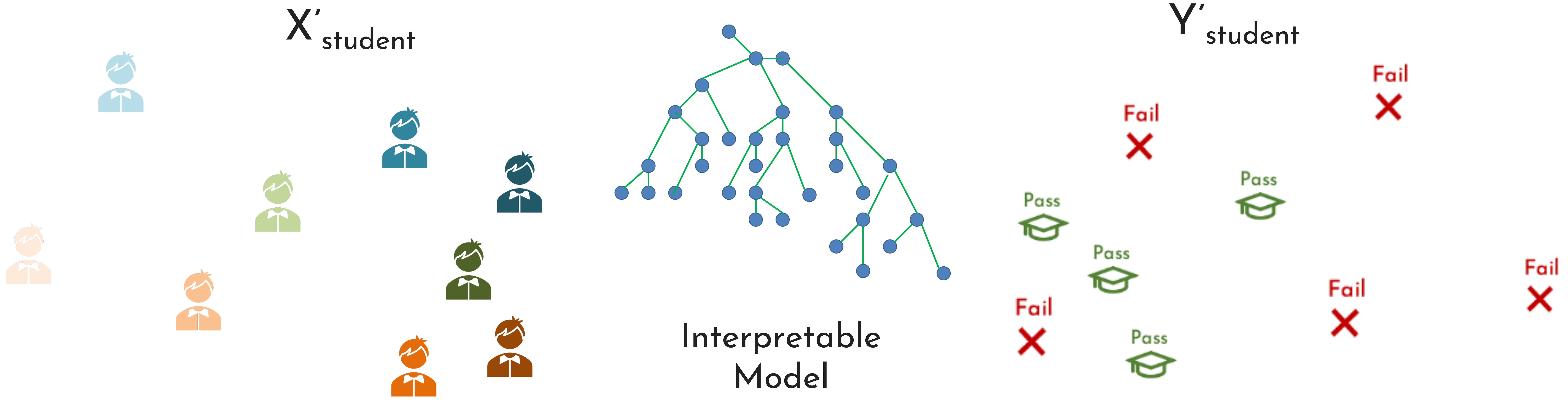
Pipeline

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

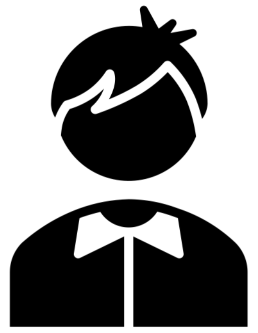
METHODOLOGY

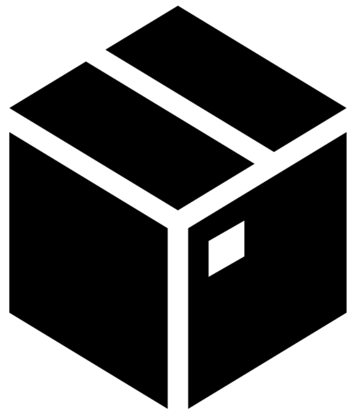
SHAP

SHapley Additive exPlanations

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

Black-Box Model





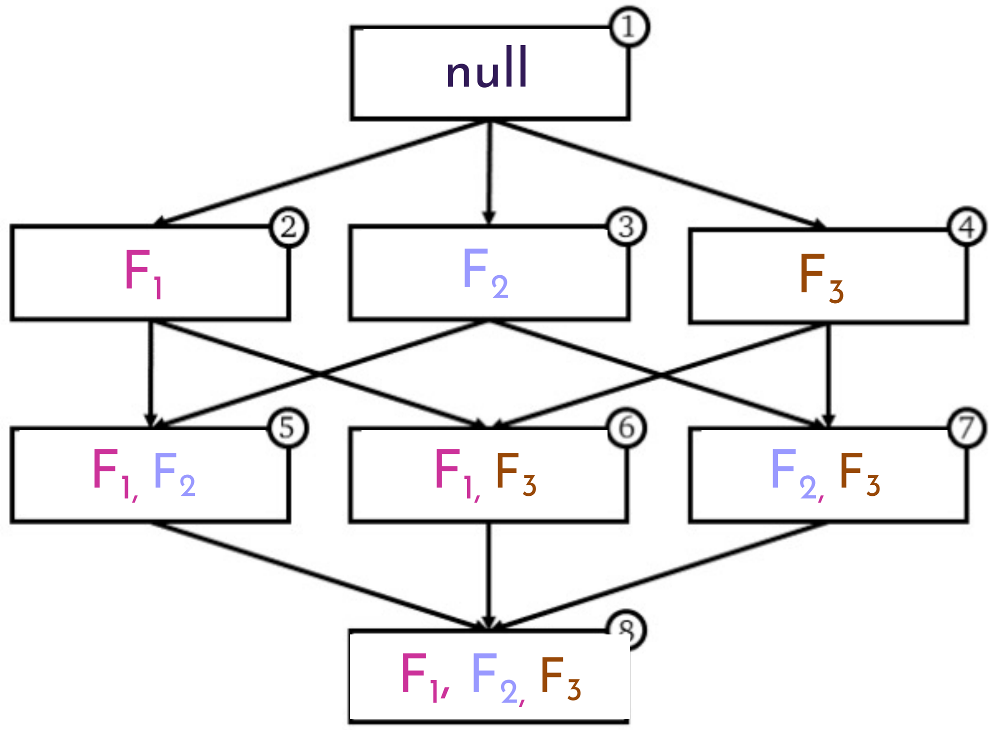
Fail

$X_{student}$ $Y_{student}$

F_1 F_2 F_3

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

Power Set (coalition)



cardinality

2^3

Lundberg and Lee, NeurIPS 2017.

Pipeline

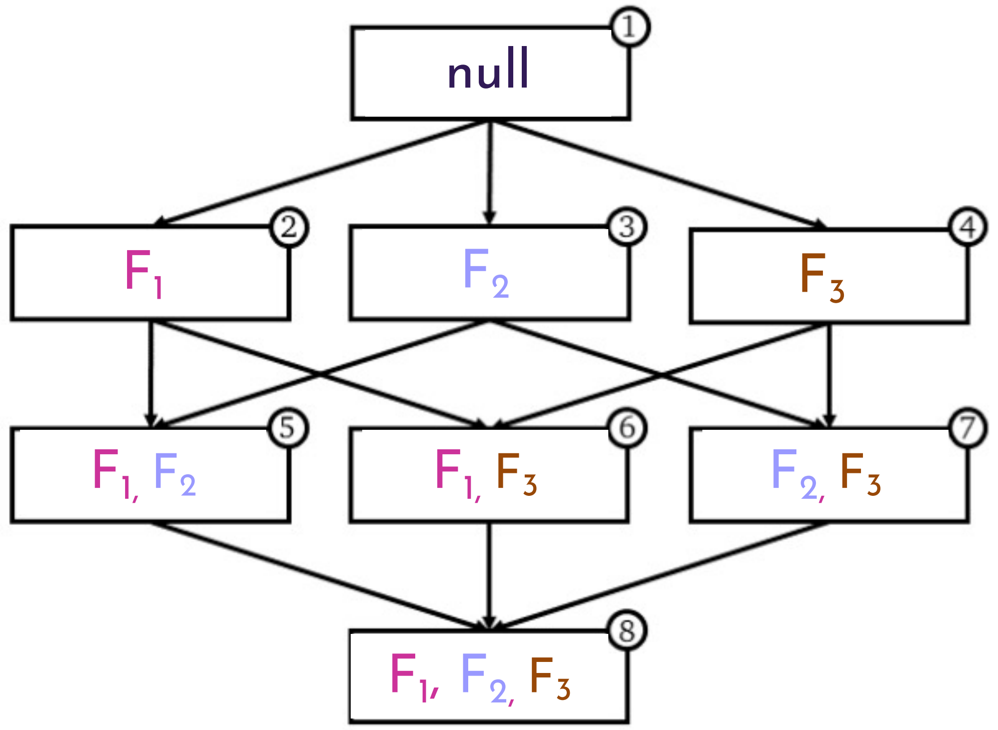
METHODOLOGY

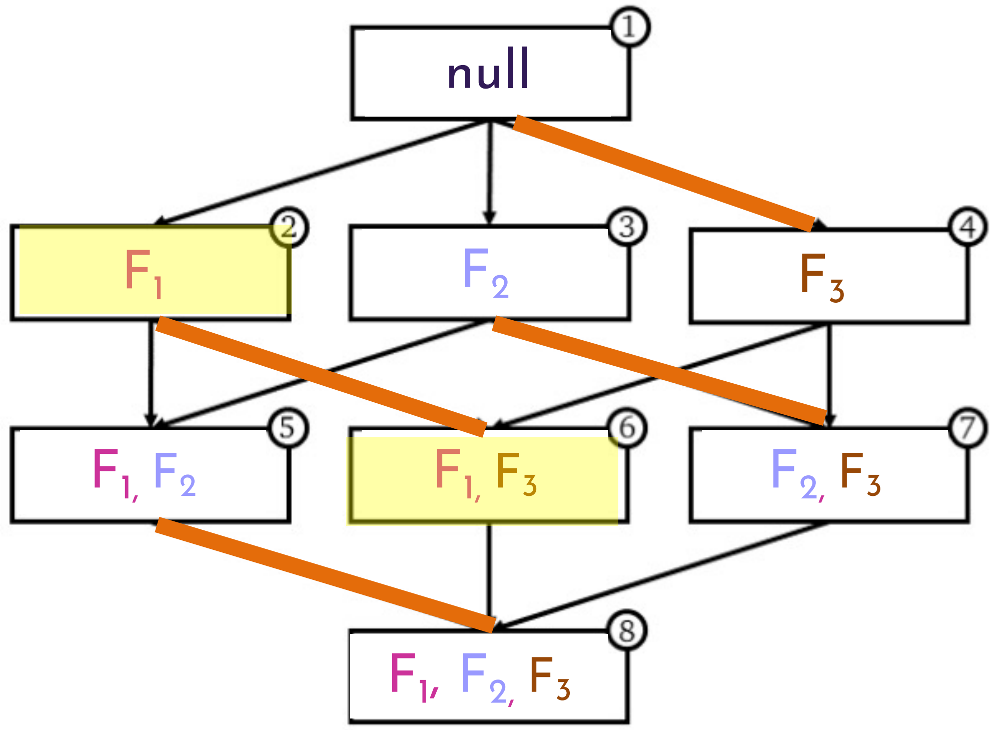
SHAP

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.

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.

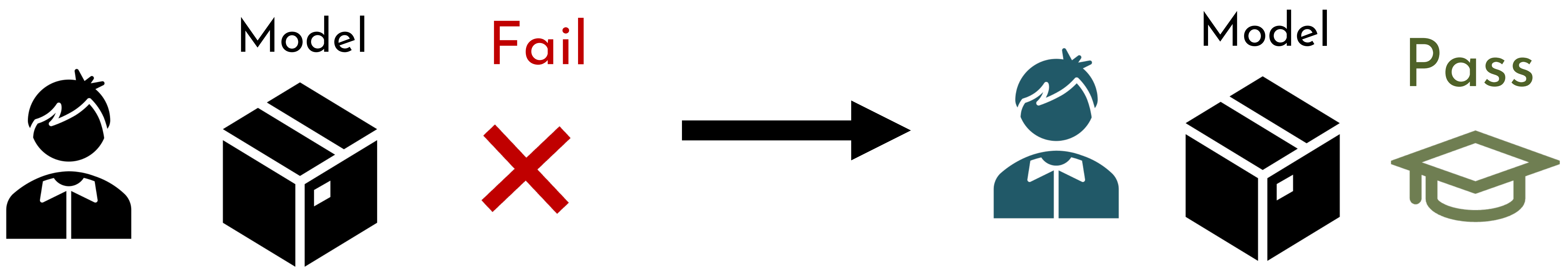
Pipeline

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

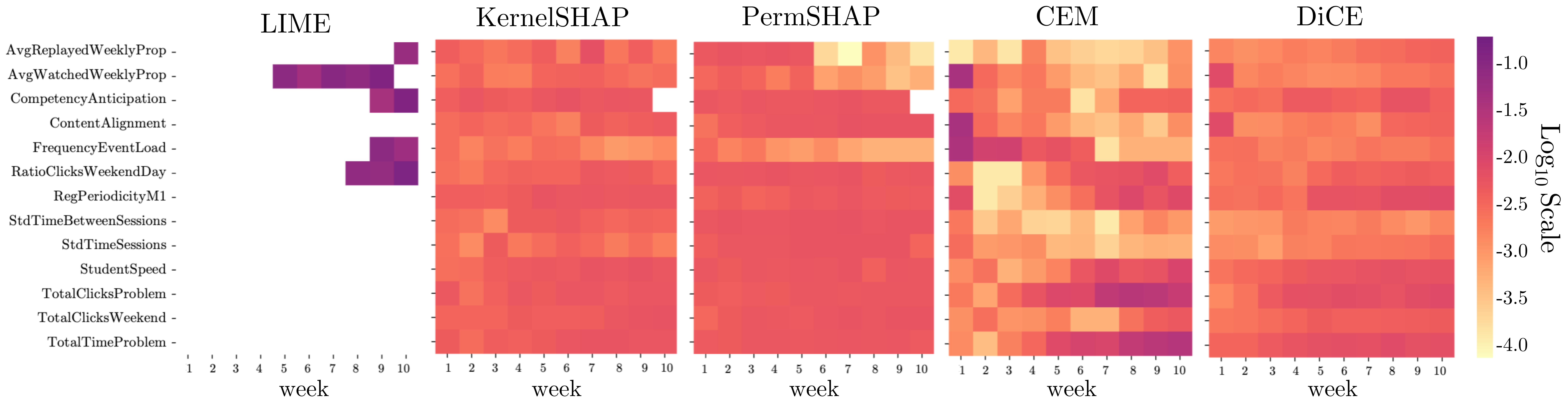
Determinal Point Process (DPP)
Diversity Metric

Mothilal et al., FAT* 2020.

RQ1: 1 Course

RESULTS

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



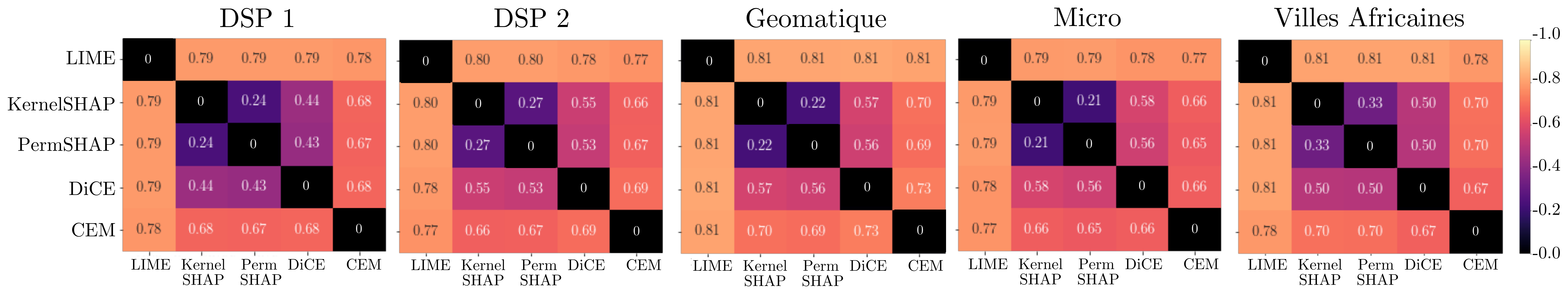
LIME is very sparse. CEM is significantly different.

RQ2: 5 Courses

RESULTS

How do explanations (quantitatively) compare across courses?

Jensen-Shannon Distance



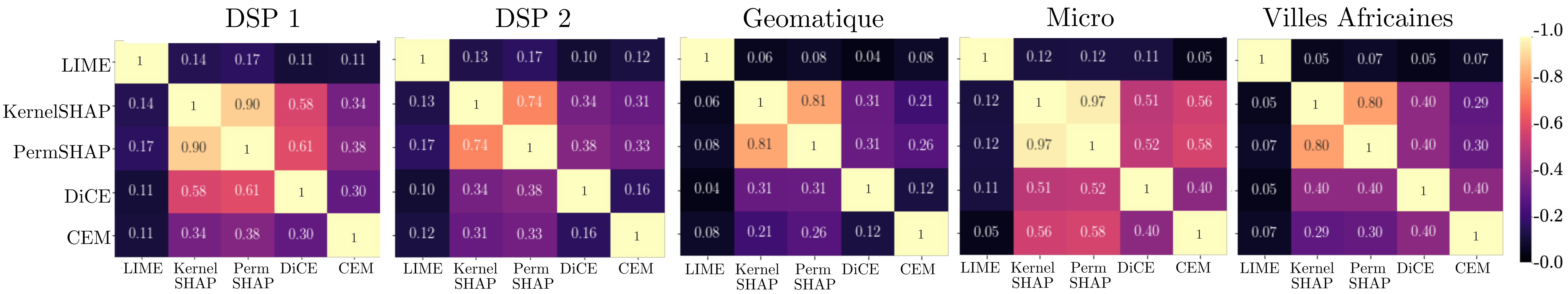
Big differences across explainability methods.

RQ2: 5 Courses

RESULTS

How do explanations (quantitatively) compare across courses?

Spearman's Rank Order Correlation



Again, big differences across explainability methods.

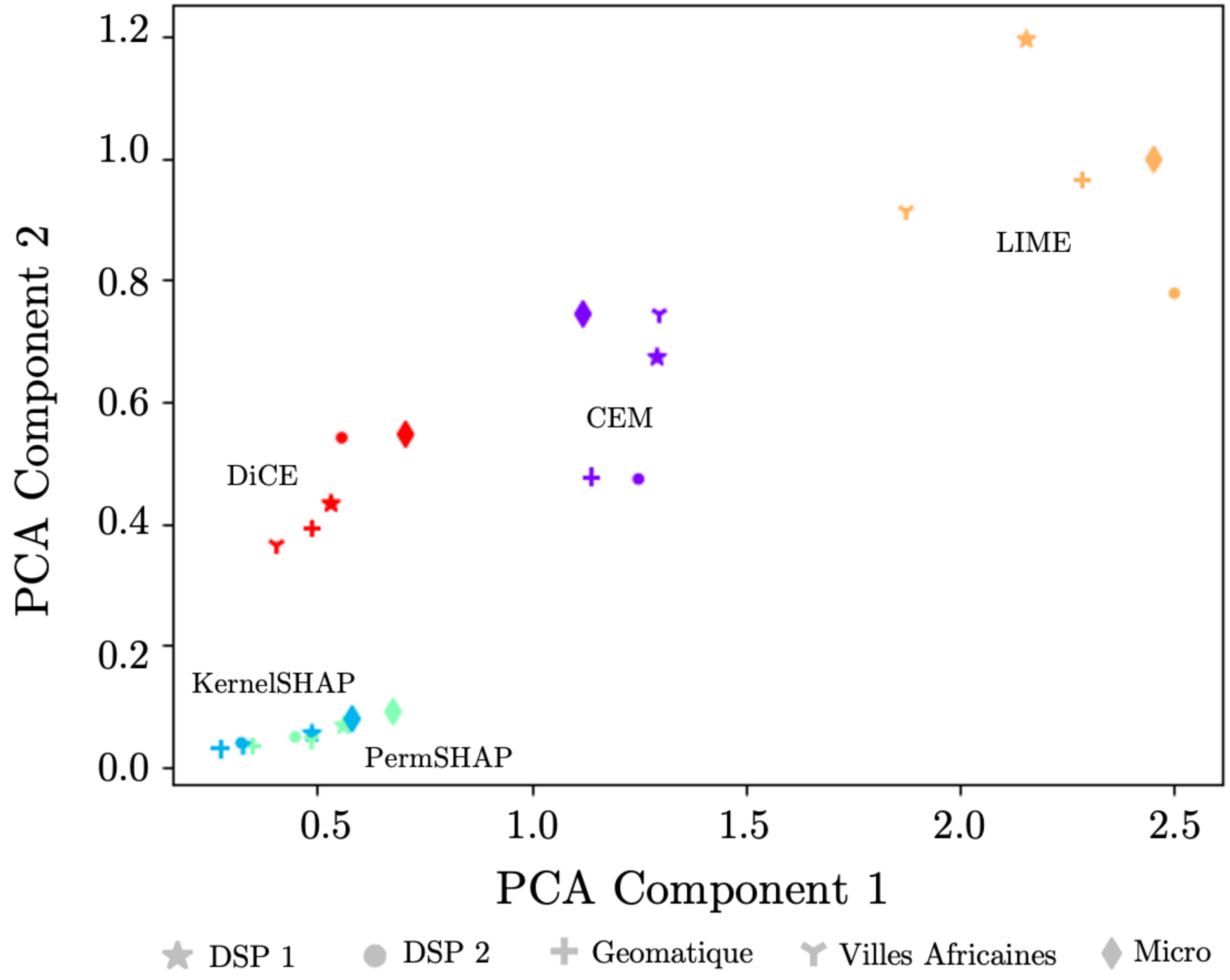
RQ2: 5 Courses

RESULTS

How do explanations (quantitatively) compare across courses?

PCA Analysis

Feature importance clusters by explainability method, not by course

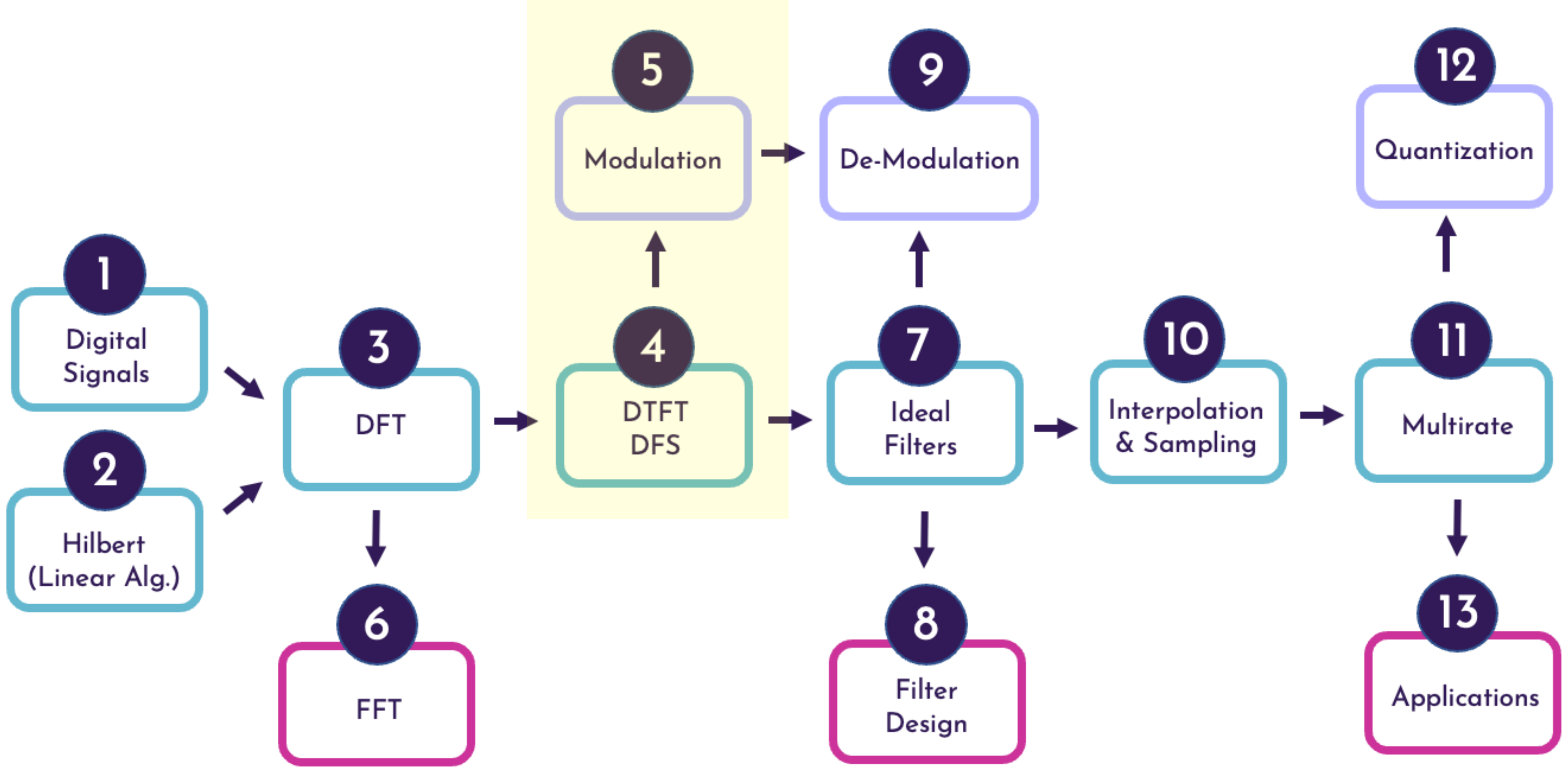


RQ3: Validation

RESULTS

Do explanations align with prerequisite relations in a course curriculum (DSP 1)?

1. Train a model to predict **Week 5** quiz performance.
2. Examine if **Week 4** features are found important.

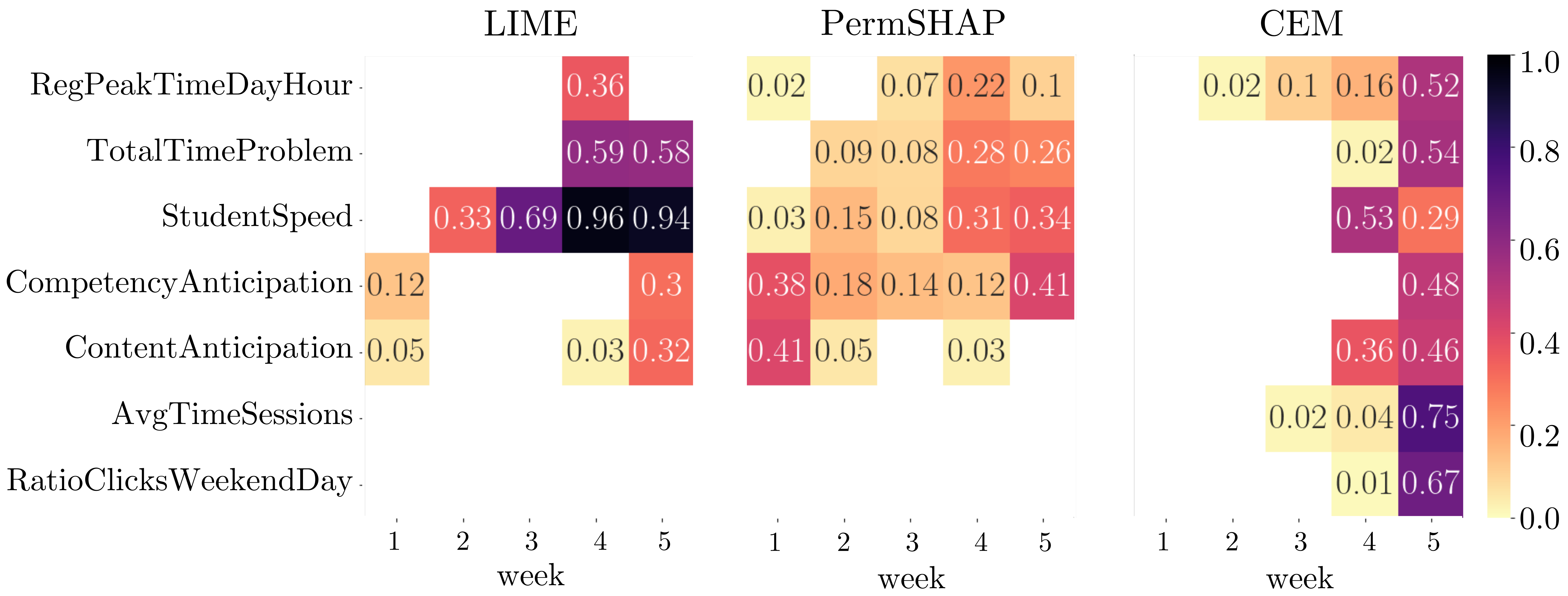


DSP 1: SKILL MAP

RQ3: Validation

RESULTS

Do explanations align with prerequisite relations in a course curriculum (DSP 1)?



Partially! However, each method identifies different important features.

Implications

DISCUSSION

Explainability methods are imperfect and biased.

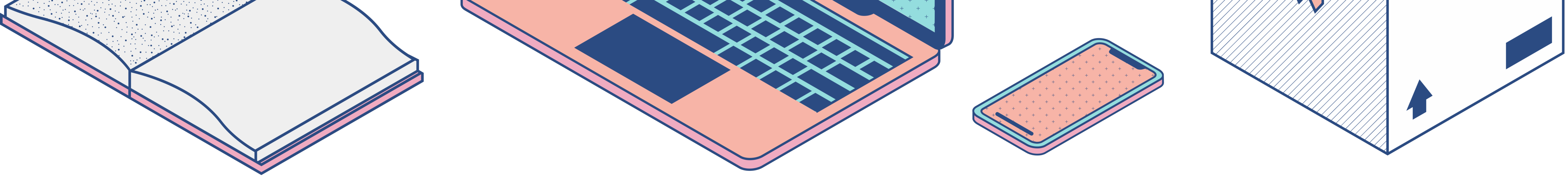
We urge data scientists to:

- Carefully select an appropriate explainability method based on a downstream task
- Keep potential biases of the explainer in mind while analyzing interpretability results

Extensions

FUTURE WORK

- Extend to different tasks (i.e. dropout) and modalities (i.e. flipped, ITS)
- Explore black-box model architectures to see if explainability method effectiveness differs across predictors
- Which explanations lead to the most effective interventions for improved learning outcomes?



Main Takeaways

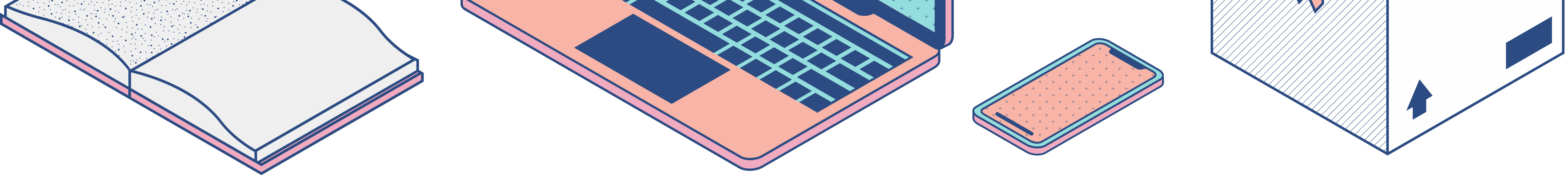
EVALUATING THE EXPLAINERS: BLACK BOX

EXPLAINABLE ML FOR SUCCESS PREDICTION

Explainability methods, systematically,

do not agree

on which features are important for predictions



Main Takeaways

EVALUATING THE EXPLAINERS: BLACK BOX EXPLAINABLE ML FOR SUCCESS PREDICTION

epfl-ml4ed/evaluating-explainers

Using our insights, educators can
be aware of the bias
of their chosen explainability technique

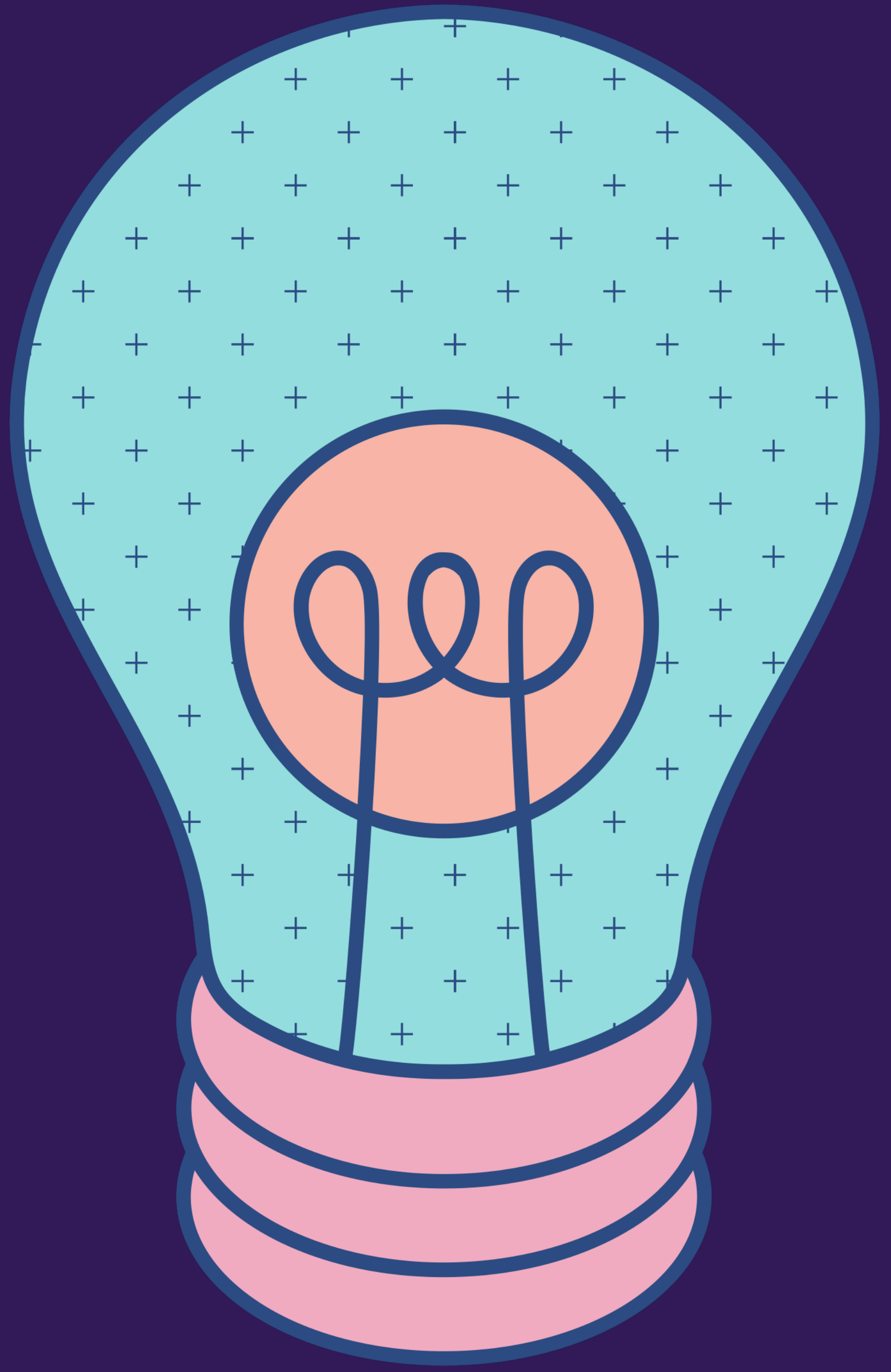
Thank you!

EVALUATING THE EXPLAINERS: BLACK BOX EXPLAINABLE ML FOR SUCCESS PREDICTION

Questions?

EVALUATING THE EXPLAINERS: BLACK BOX EXPLAINABLE ML FOR SUCCESS PREDICTION

Vinitra Swamy
epfl-ml4ed/evaluating-explainers
vinitra.swamy@epfl.ch



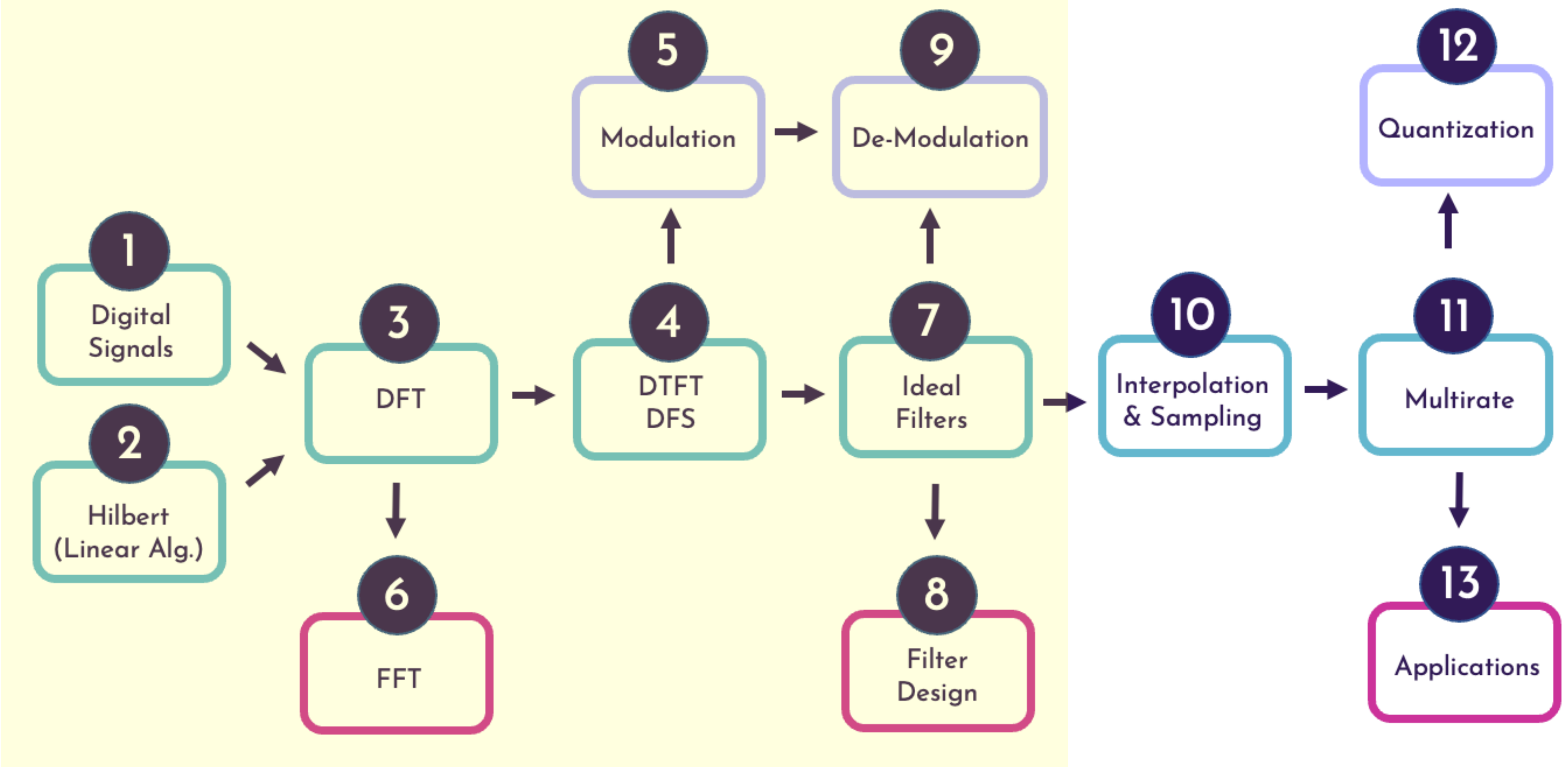
RQ3: Validation

RESULTS

Do explanations align with prerequisite relations in a course curriculum (DSP 1)?

1 Train a model to predict **Week 9** performance.

2 Examine which weeks' features are found important.

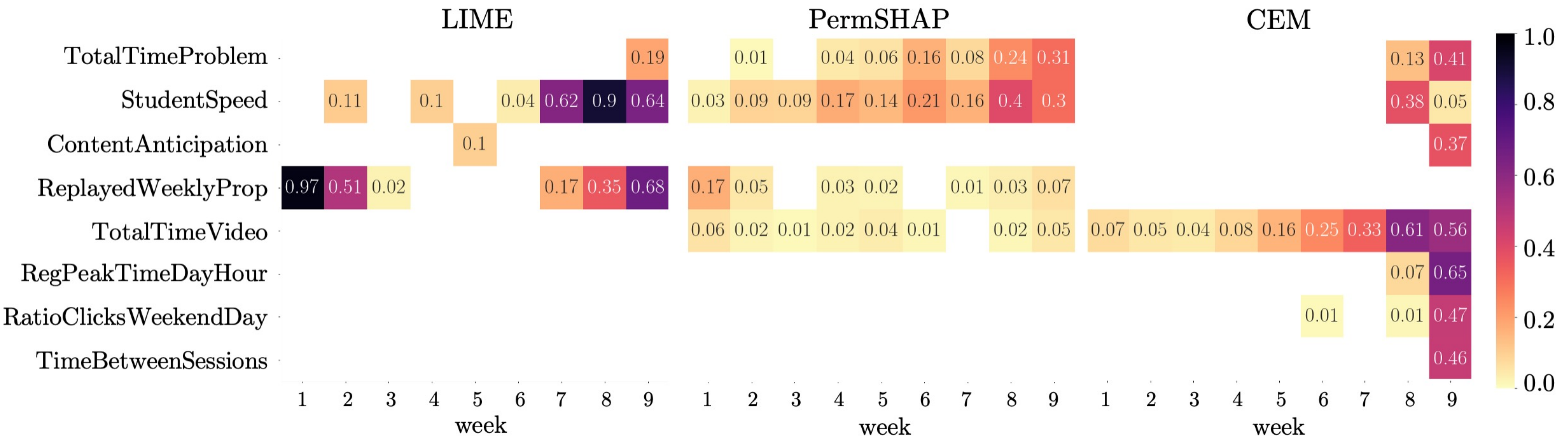


DSP 1: SKILL MAP

RQ3: Validation

RESULTS

Do explanations align with prerequisite relations in a course curriculum (DSP 1)?



Partially! However, each method identifies different important features.