Towards Interoperability and Novel Methodological Approaches for Scalable Game-Based Assessment

Hacia la interoperabilidad y nuevos enfoques metodológicos para la evaluación escalable basada en juegos

Ph.D. Thesis

Manuel Jesús Gómez Moratilla

Advisors: Félix Jesús García Clemente and
José Antonio Ruipérez Valiente

UNIVERSIDAD DE MURCIA

f SéNeCa(+)
Agencia de Ciencia y Tecnología
Región de Murcia

About me

2020 — UNIVERSIDAD DE MURCIA

BSc in Computer Science
University of Murcia

2021 — UNIVERSIDAD DE MURCIA

MSc in Big Data
University of Murcia

2021 — GOBIERNO DE ESPAÑA · MINISTERIO DE EDUCACIÓN Y FORMACIÓN PROFESIONAL

University Department Collaboration Grant
Spanish Government

2023 — f SéNeCa (+) Agencia de Ciencia y Tecnología Región de Murcia

F. Séneca Grant
Region of Murcia

2024 — MIT

4-month internship
Massachusetts Institute of Technology
Cambridge, MA (United States)

PhD Thesis Defense

2025 — UNIVERSIDAD DE MURCIA · 1272

4 years of PhD Thesis (2021-2025)

Cyber Data Lab

UNIVERSIDAD DE MURCIA

Introduction

Technology is transforming how we teach and learn

- Assessment systems are evolving alongside instruction

Games are an essential part of everyday digital life

- Part of our societies since the early days of computing

Serious Games (SGs)

Games designed for a primary purpose other than entertainment

- Education, training…

Introduction

Game-Based Assessment (GBA)

A specific application of **games**, referring to a type of **assessment** that uses players' **interactions** with the game, both digital and nondigital, as a source of **evidence** to make meaningful **inferences** about what players know and can **do** (i.e., knowledge, skills) (*Kim & Ifenthaler, 2019).*

GBAs embed evaluation with authentic and engaging experiences

- As valid as traditional assessment methods

Methodology

Lack of SOTA

Objective 1

How has GBA been employed? Which advantages and challenges?

Review the current state of GBA

M1. Literature review

WORK I

CHALLENGE Data interoperability

Lack of interoperable models

Objective 2

Can we develop a model to enable interoperability across different games?

Design an interoperable semantic model for GBA data

M2. Model design and implementation

WORK II

CHALLENGE System-level scalability

Lack of scalable architectures

Objective 3

Can we perform GBAs at scale across diverse platforms and contexts?

Build a scalable framework for interoperable GBAs

M3. Framework development

WORK III

Methodology

Lack of SOTA

Objective 1

How is GBA employed today? Which advantages and challenges?

Review the current state of GBA

M1. Literature review

WORK I

CHALLENGE Interpretability in assessment models

Lack of interpretable models

Objective 4

What strategies support useful model explanations for non-technical users?

Enhance predictive models through eXplainable AI

M4. Interpretability in predictive models

WORK IV

CHALLENGE Data labeling process

Lack of labeling frameworks

Objective 5

How can the labeling process be optimized and made more intuitive?

Optimize data labeling for AI techniques in GBA

M5. Practical tool development

WORK V

+

Additional Work

Work I ➡ Publication

1 M1

Title	A systematic literature review of game-based assessment studies: Trends and challenges
Authors	Manuel J. Gomez, José A. Ruipérez-Valiente, Félix J. García Clemente

Journal	IEEE Transactions on Learning Technologies
DOI	10.1109/TLT.2022.3226661

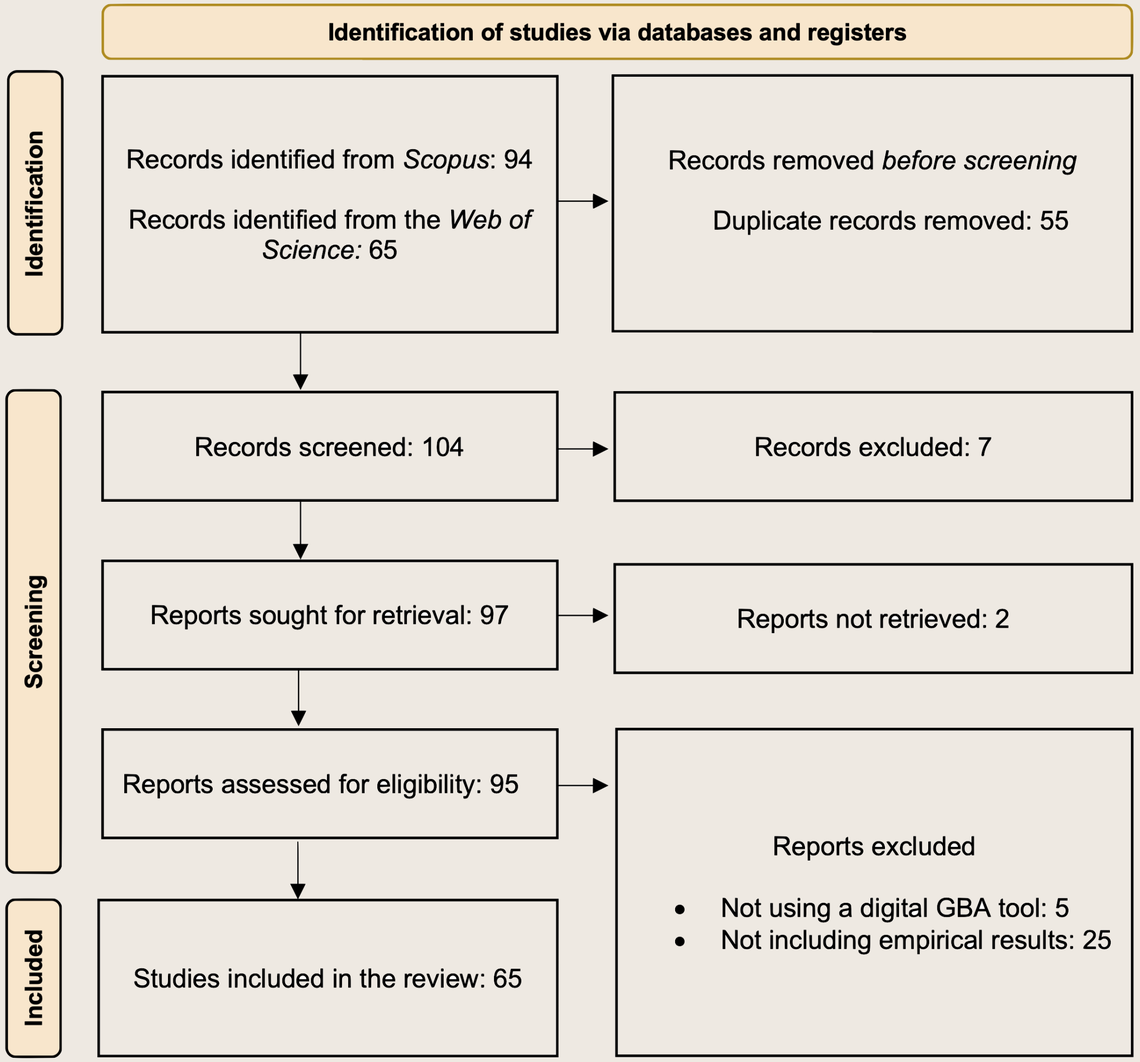
Work I ➡ Methods

PRISMA

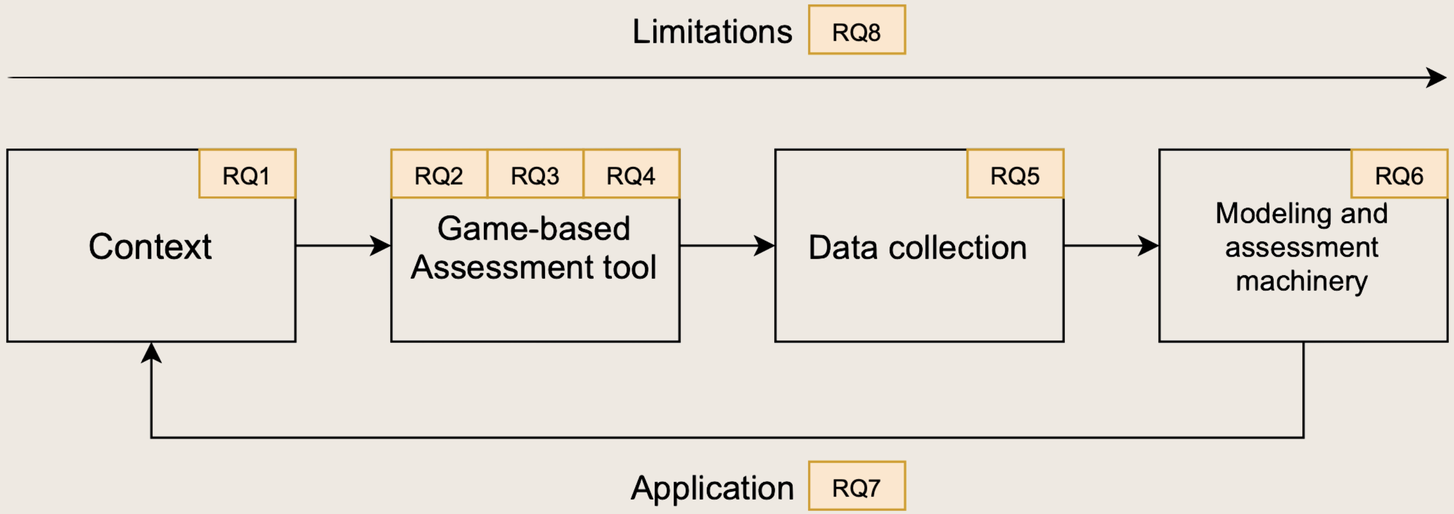
Preferred Reporting Items for Systematic reviews and Meta-Analyses

Search query:

(TITLE(“game-based assessment”) OR KEY(“game-based assessment”))



Work I ➡ Research Questions



RQ1. Context

RQ2. Primary purpose

RQ3. GBA domain

RQ4. Tool availability

RQ5. Data sample size

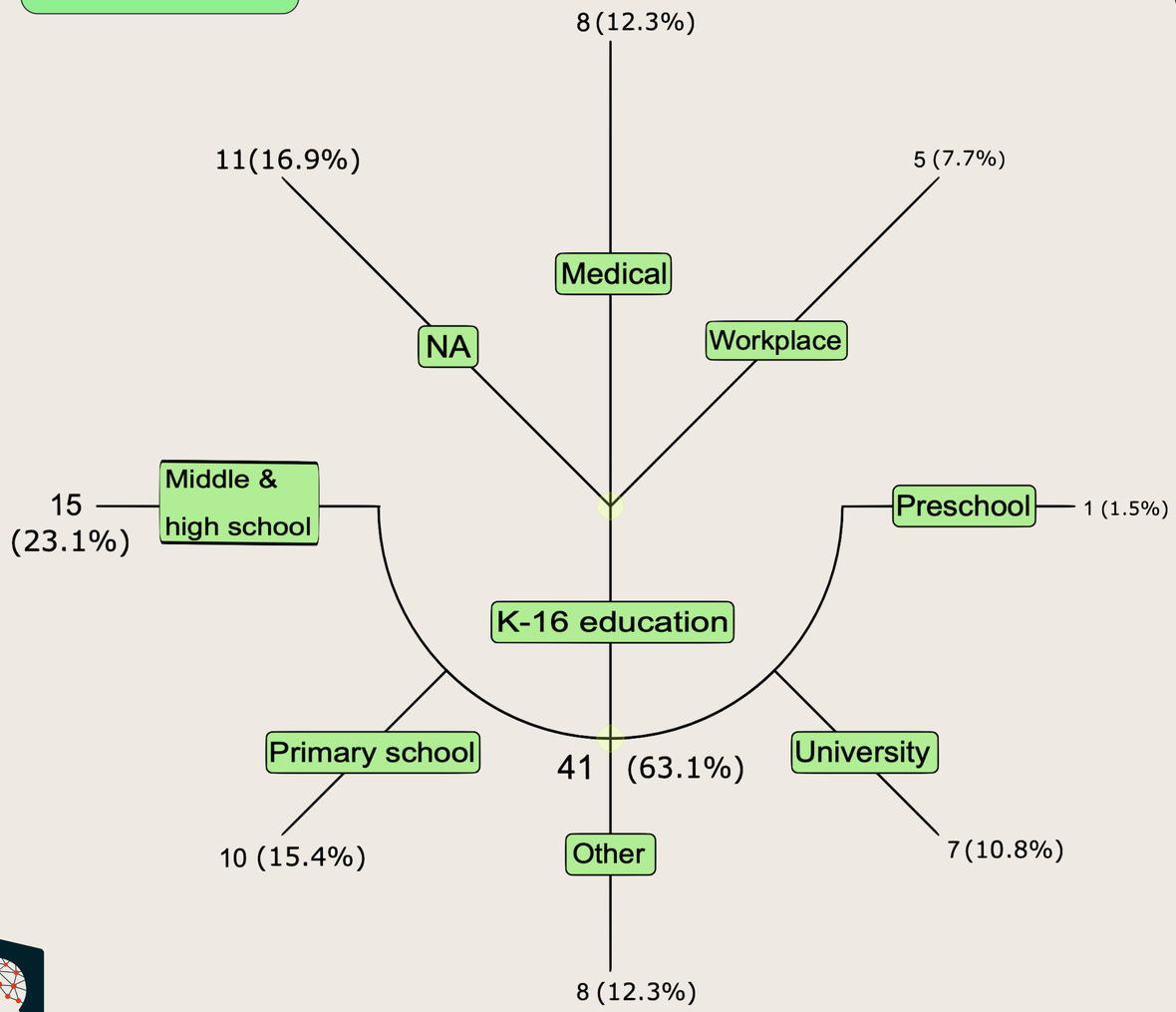
RQ6. Methods applied

RQ7. Intended stakeholder

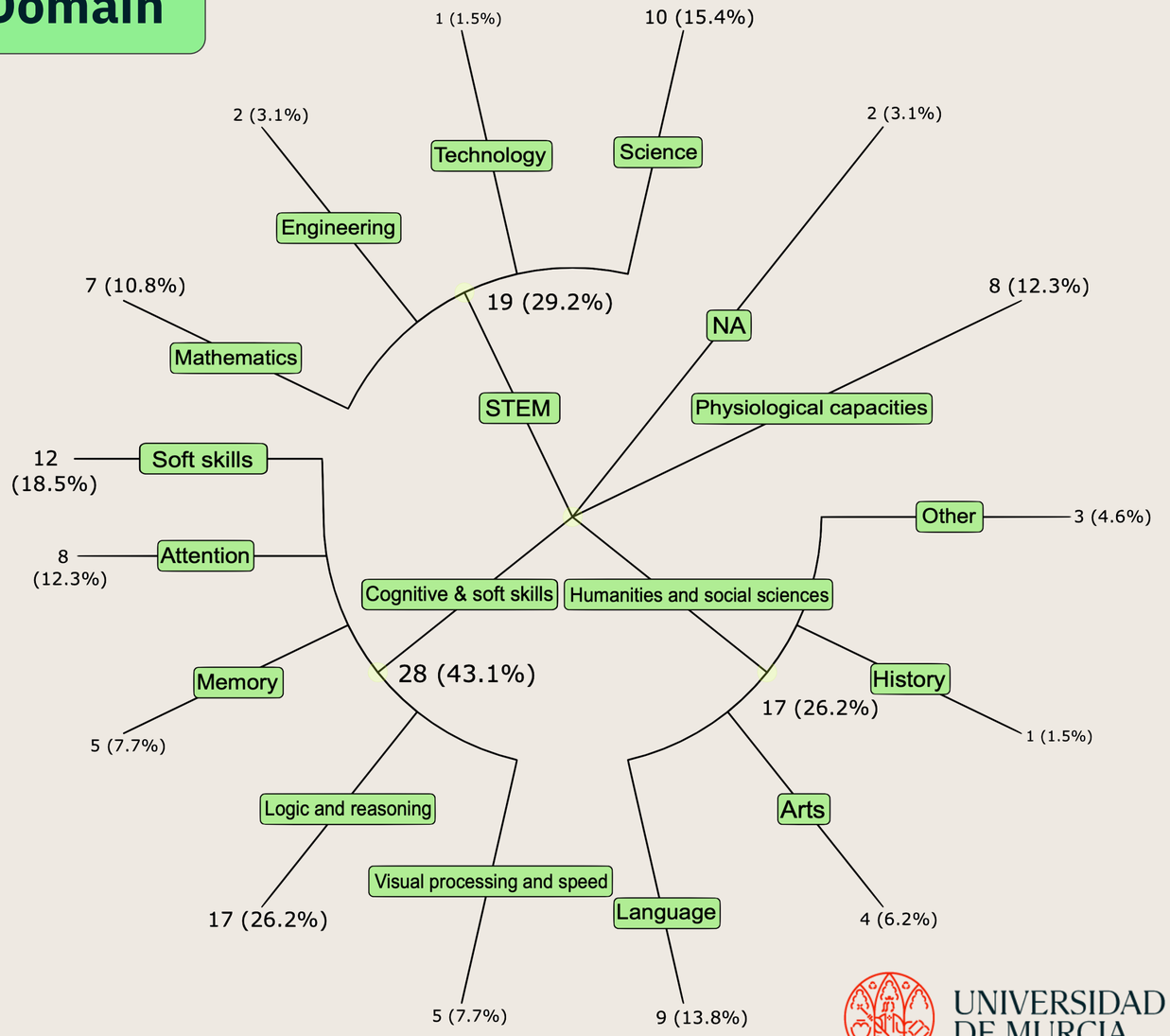
RQ8. Limitations

Work I ➡ Findings

Context



Domain



Work I ➡ Findings

Trends

- Focus on K-16 education
- Assessment and validation emphasis
- Domains beyond STEM
- Predominance of basic methods

Challenges

- Interoperability and replication
- Need for advanced and interpretable models
- Small sample sizes
- Integration into the curriculum

Work II

Objective 1

How is GBA employed today? Which advantages and challenges?

Review the current state of GBA

Work I

CHALLENGE Data interoperability

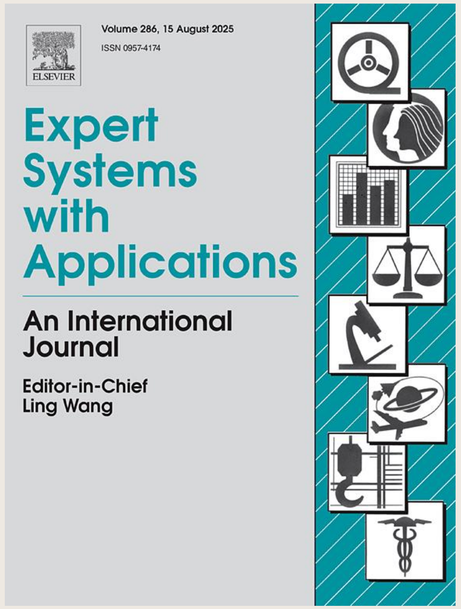
Objective 2

Can we develop a model to enable interoperability across different games?

Design an interoperable semantic model for GBA data

Work II

Work II ➡ Publication



Title	Developing and validating interoperable ontology-driven game-based assessments
Authors	Manuel J. Gomez, José A. Ruipérez-Valiente, Félix J. García Clemente

Journal	Expert Systems With Applications
DOI	10.1016/j.eswa.2024.123370

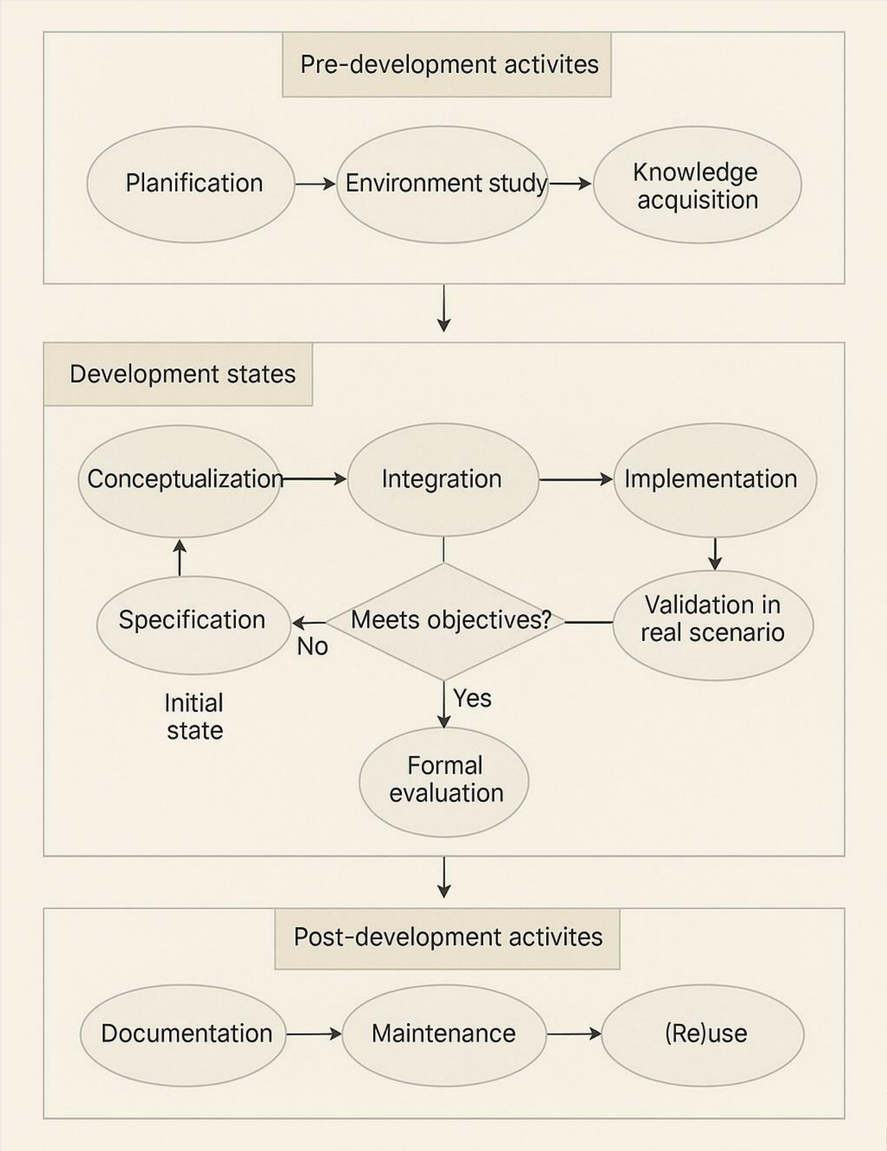
Work II ➡ Development

Ontology

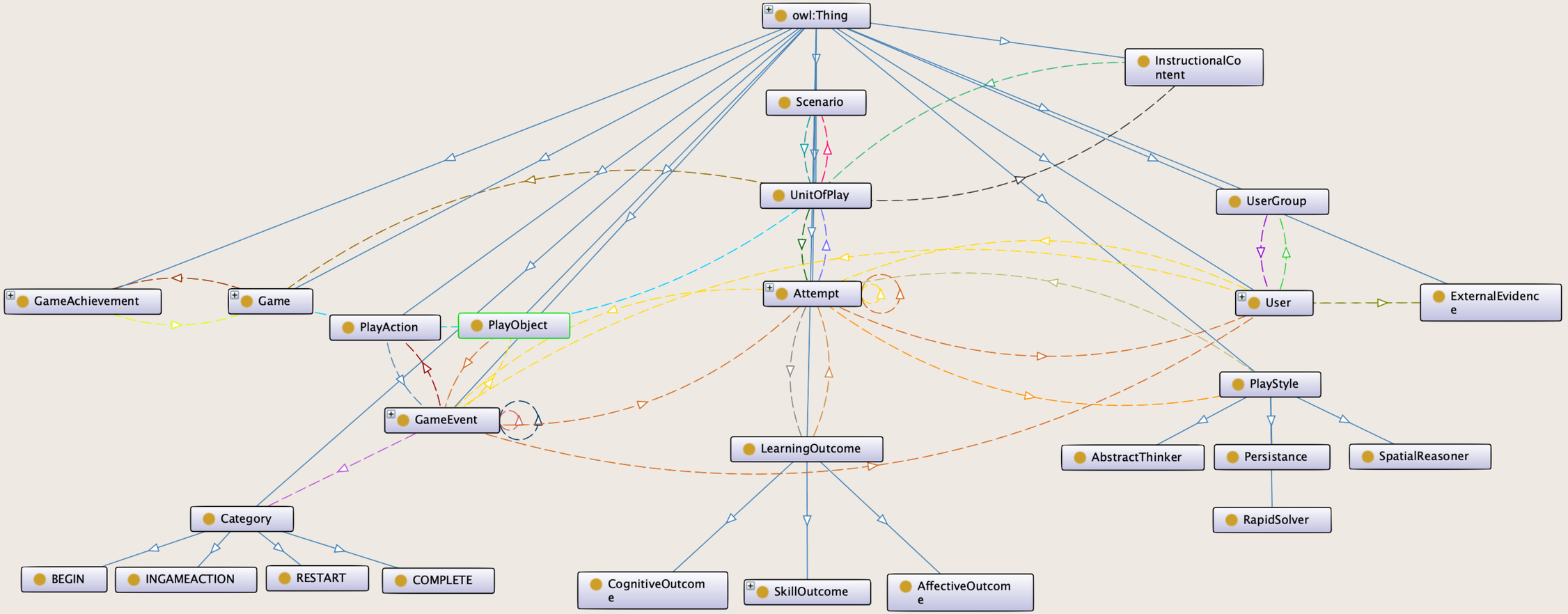
is a **structured representation of knowledge** that defines the **key concepts**, **entities**, and **relationships** within a specific domain. It provides a **shared vocabulary** and a **formal framework** for organizing and reasoning about information.

Methontology: a structured method designed to build ontologies from scratch, reusing others as they are, or by a process of re-engineering them

- Recommended by the **Foundation for Intelligent Physical Agents (FIPA)** for the ontology construction task



Work II ➡ Ontology



Work II ➡ Ontology validation

Validation computing metrics in a set of data from ten different SGs

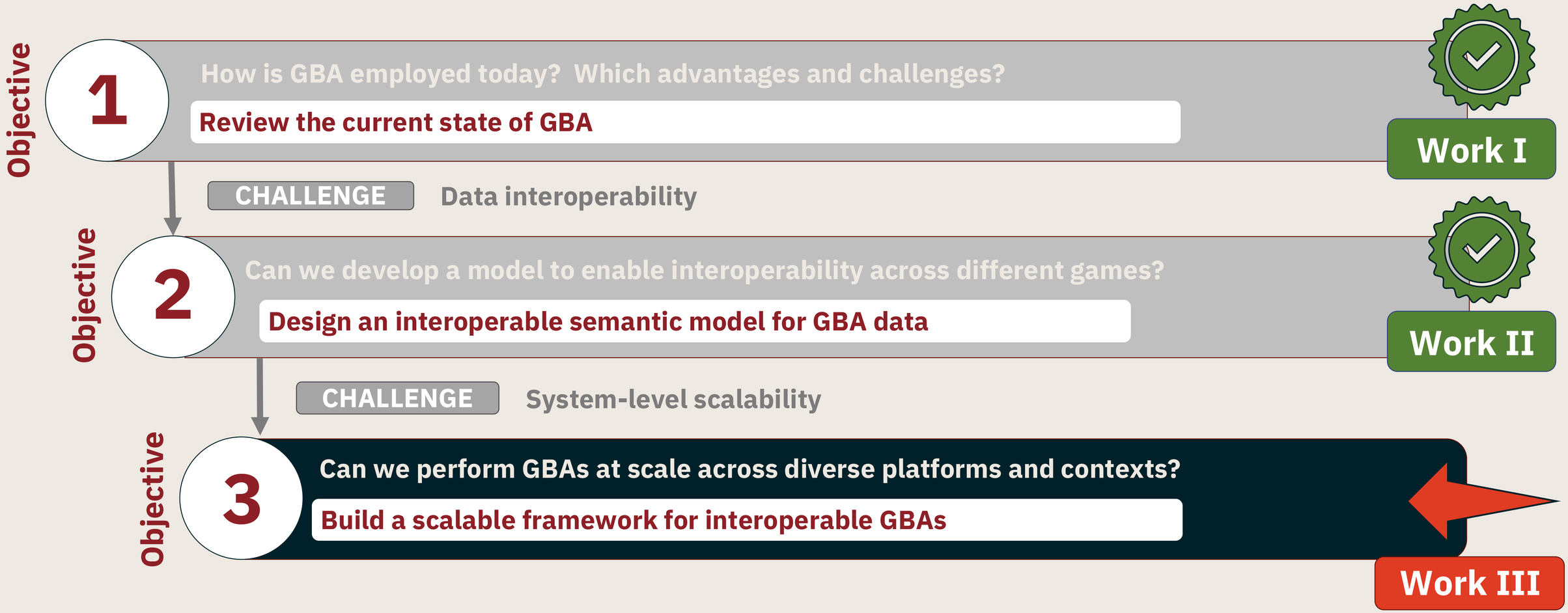
Literature metrics

- **Activity indicators**
- **Persistence indicators**
- **Event types**
- **User performance**
- **Levels of activity**
- **Funnel by user**

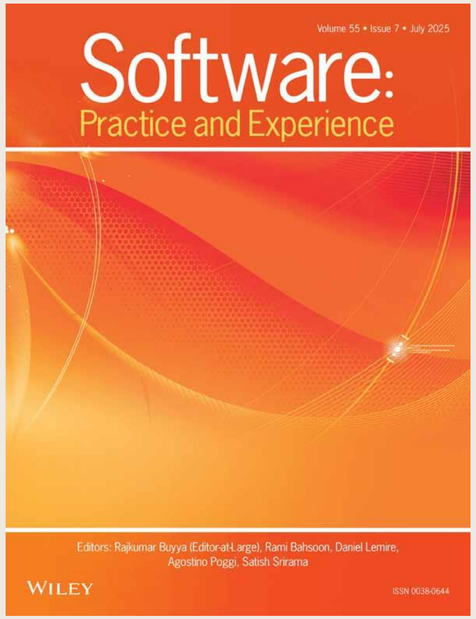
Proposal of new metrics

- **Levels of difficulty**
- **Persistence**
- **Play styles**

Work III


Objective
1
How is GBA employed today? Which advantages and challenges?
Review the current state of GBA
Work I
CHALLENGE
Data interoperability
Objective
2
Can we develop a model to enable interoperability across different games?
Design an interoperable semantic model for GBA data
Work II
CHALLENGE
System-level scalability
Objective
3
Can we perform GBAs at scale across diverse platforms and contexts?
Build a scalable framework for interoperable GBAs
Work III

Work III ➡ Publication

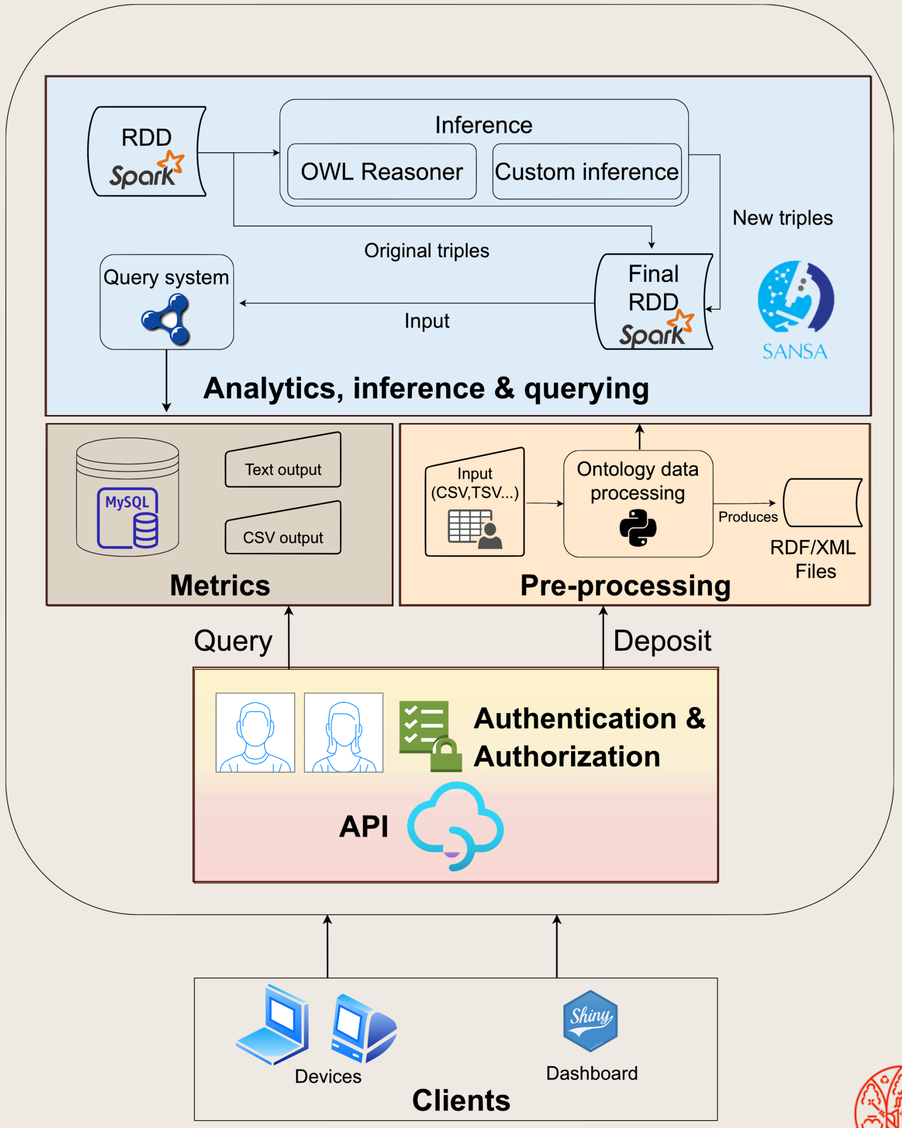


Title	A framework to support interoperable Game-based Assessments as a Service (GBAaaS): Design, development, and use cases
Authors	Manuel J. Gomez, José A. Ruipérez-Valiente, Félix J. García Clemente

Journal	Software: Practice & Experience
DOI	10.1002/spe.3254

Work III ➡ Framework

- **Requirement 1**: Semantic layer between the log data and a common knowledge model
- **Requirement 2**: Processing of large scale data
- **Requirement 3**: Interoperability for GBA metrics and visualizations
- **Requirement 4**: Easy communication with external sources
- **Requirement 5**: Privacy, authentication and authorization configurations



Cyber Data Lab

UNIVERSIDAD DE MURCIA

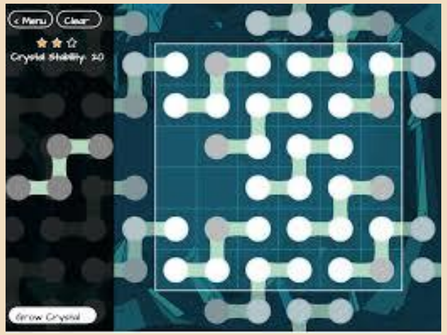
Work III ➡ Performance evaluation

FIELD DAY LEARNING GAMES

Dataset sizes

OPEN GAME DATA

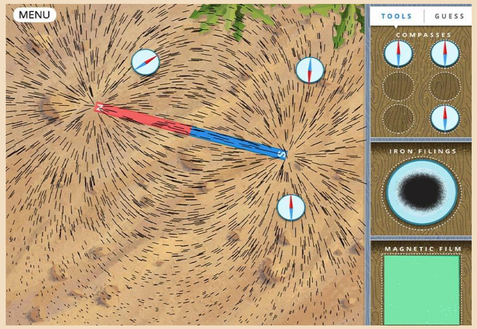
Crystal | 2.98GB



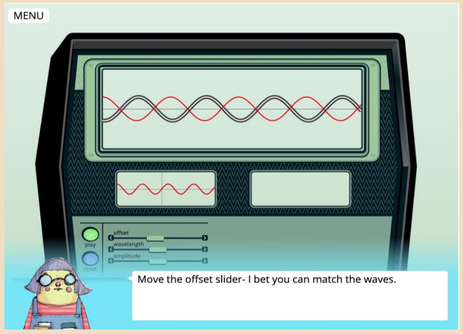
Balloon | 2.87GB

Cycle carbon | 2.93GB

Magnet | 2.76 GB



Waves | 3.01 GB



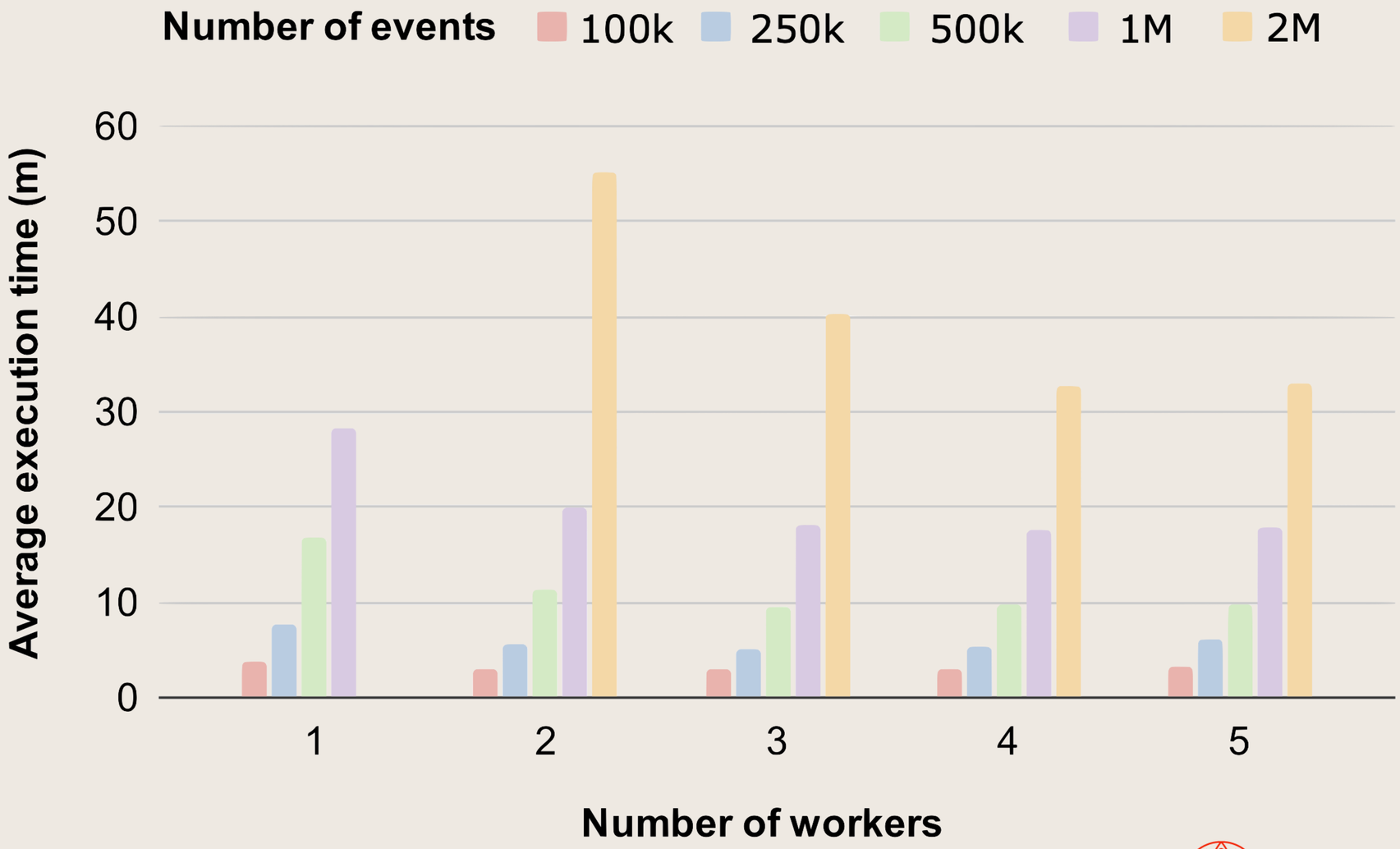
Cyber Data Lab

UNIVERSIDAD DE MURCIA

Work III ➡ Performance evaluation

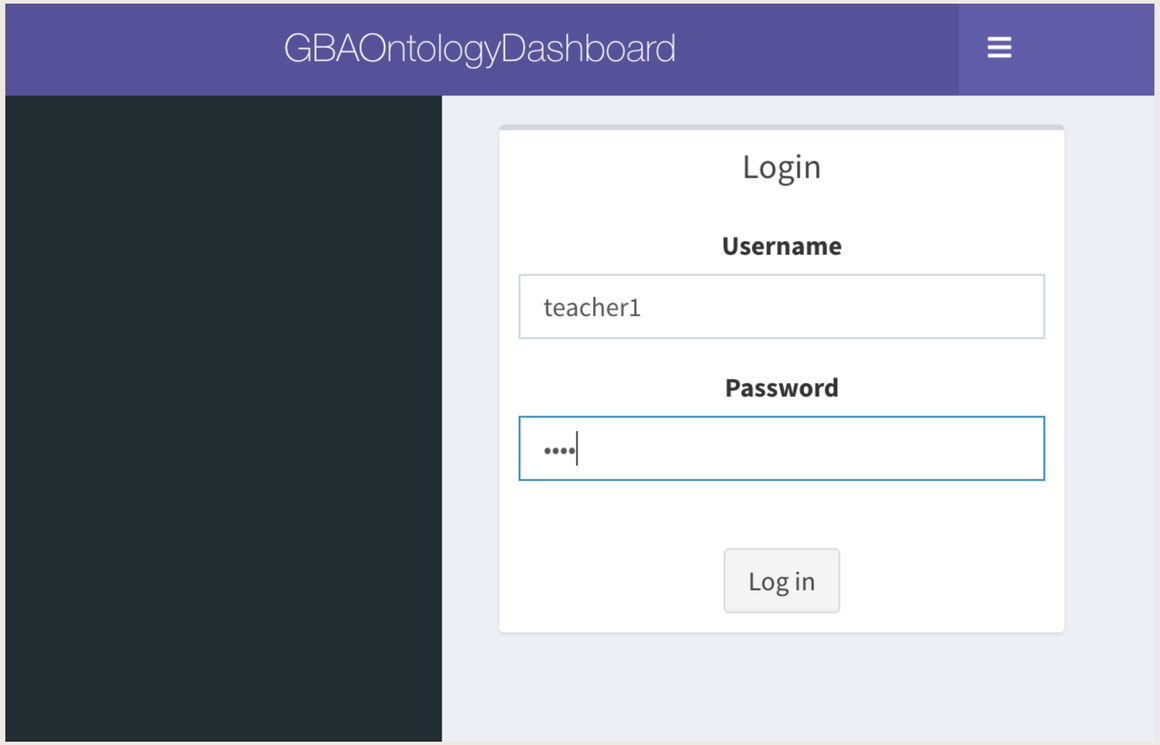
- **Tests** run with event sizes from 100k to 2M
- **Execution time** decreases as the number of workers increases
- **Best** performance gains are observed up to **3–4 workers**

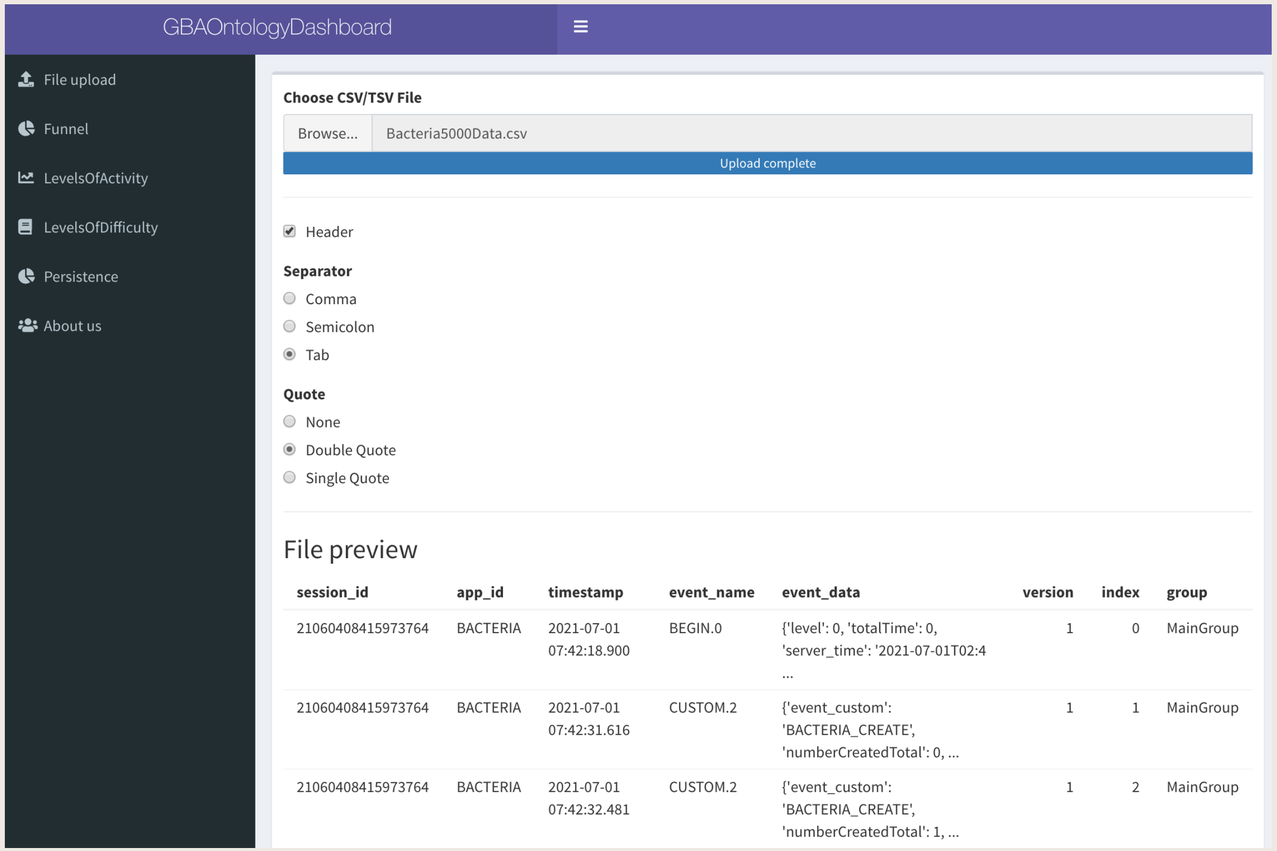
2M events ≈ 39 classrooms using a game for one hour/week for one month



Work III ➡ Use cases

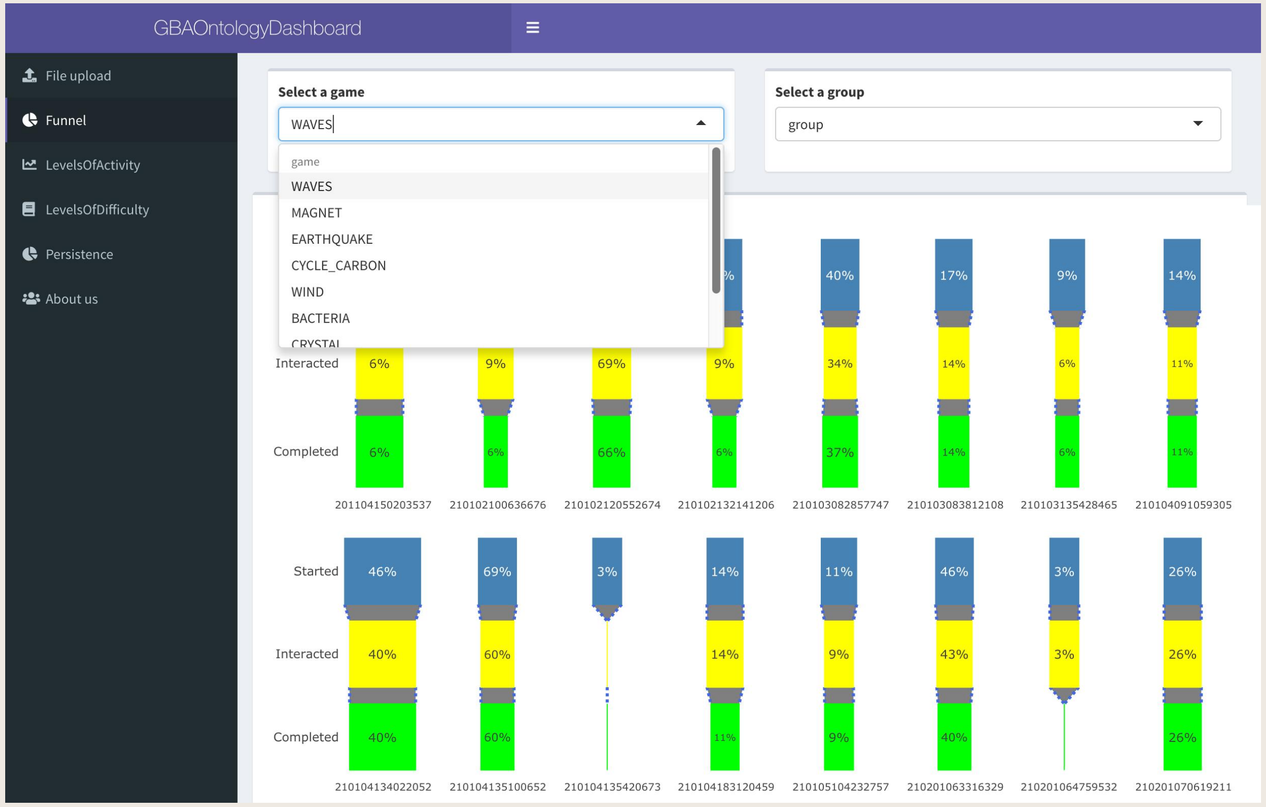
Dashboard

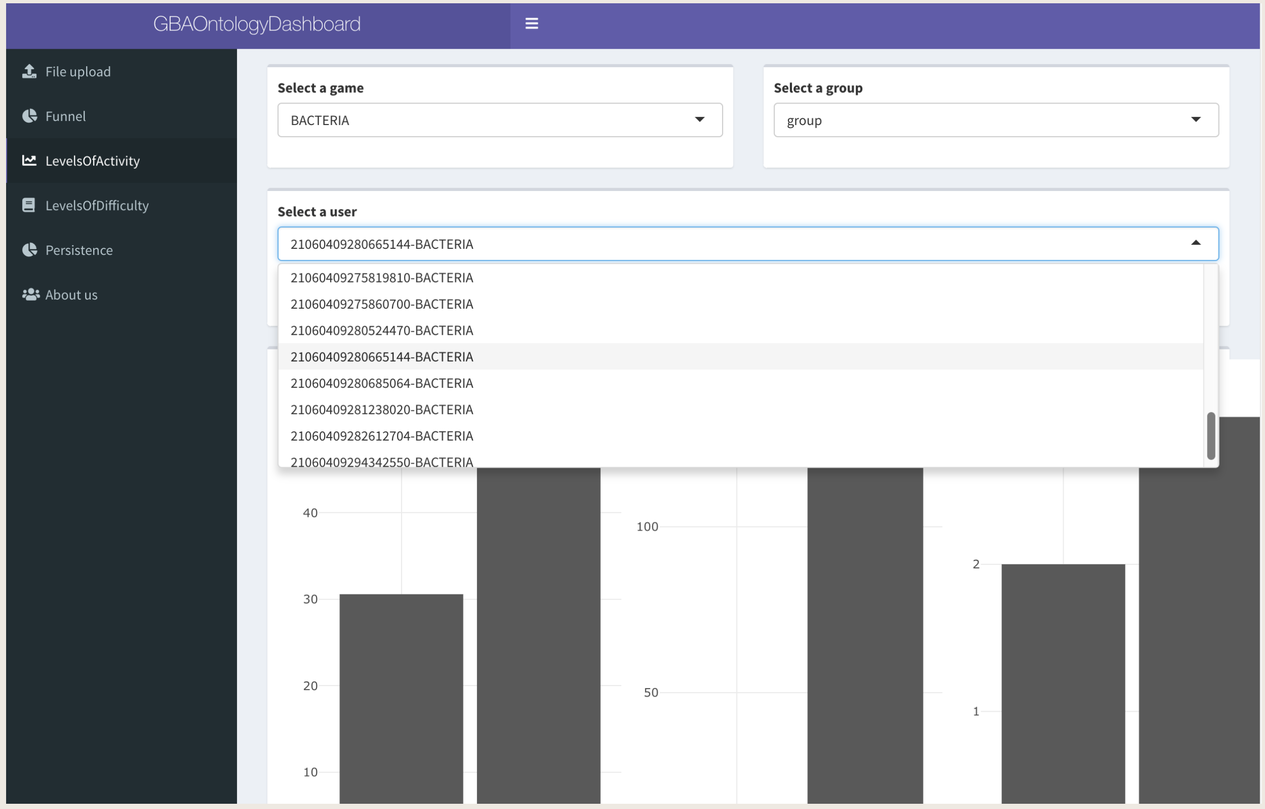




Work III ➡ Use cases

Dashboard





Work III ➡ Use cases

Reports

GBA Ontology Framework Report

This report has been generated automatically for instructor with username: teacher1

This report includes MainGroup from the game EARTHQUAKE and MainGroup from the game MAGNET.

Group summary

This has been the performance of each group for the selected games:

Summary for each group.

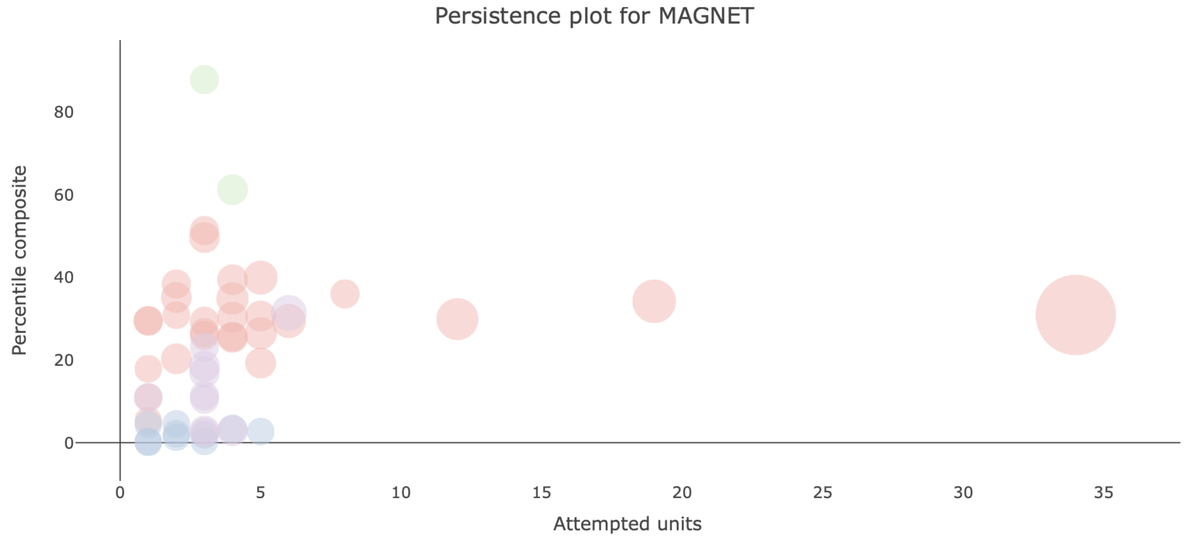
Game	Group	Avg Active Time By Attempt	Total Active Time	Avg Events By Attempt	Total Events	Total Units Started	Total Units Completed	Most Difficult Unit
EARTHQUAKE	MainGroup	108.62055	21398.25	21.954315	4325	354	16	EARTHQUAKE-36
MAGNET	MainGroup	34.69733	157005.42	8.273149	37436	5935	3318	MAGNET-58

For each game, the units that could be problematic (abandoned percentage > 75%) are:

Game	Unit	Percentage Abandoned
EARTHQUAKE	EARTHQUAKE-18	100.0000
EARTHQUAKE	EARTHQUAKE-19	100.0000
EARTHQUAKE	EARTHQUAKE-36	76.4706
EARTHQUAKE	EARTHQUAKE-47	100.0000
EARTHQUAKE	EARTHQUAKE-6	100.0000
MAGNET	MAGNET-0	100.0000

Persistence

The persistence summary for each one of the games has been:



Note that the size of each bubble represents the number of completed units for each user.

Work IV

Objective 1

How is GBA employed today? Which advantages and challenges?

Review the current state of GBA

Work I

CHALLENGE Interpretability in assessment models

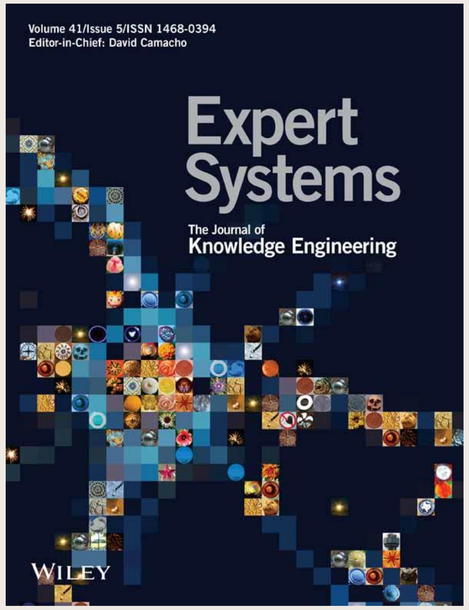
Objective 4

What strategies support useful model explanations for non-technical users?

Enhance predictive models through XAI

Work IV

Work IV ➡ Publication

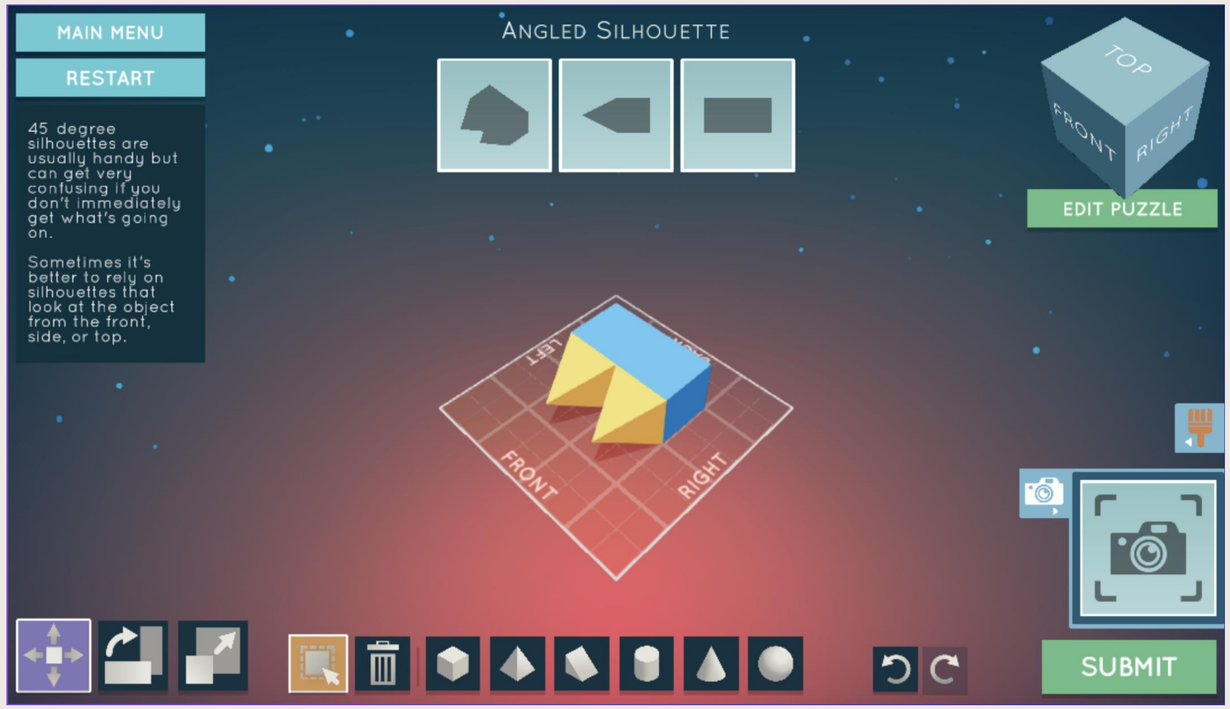


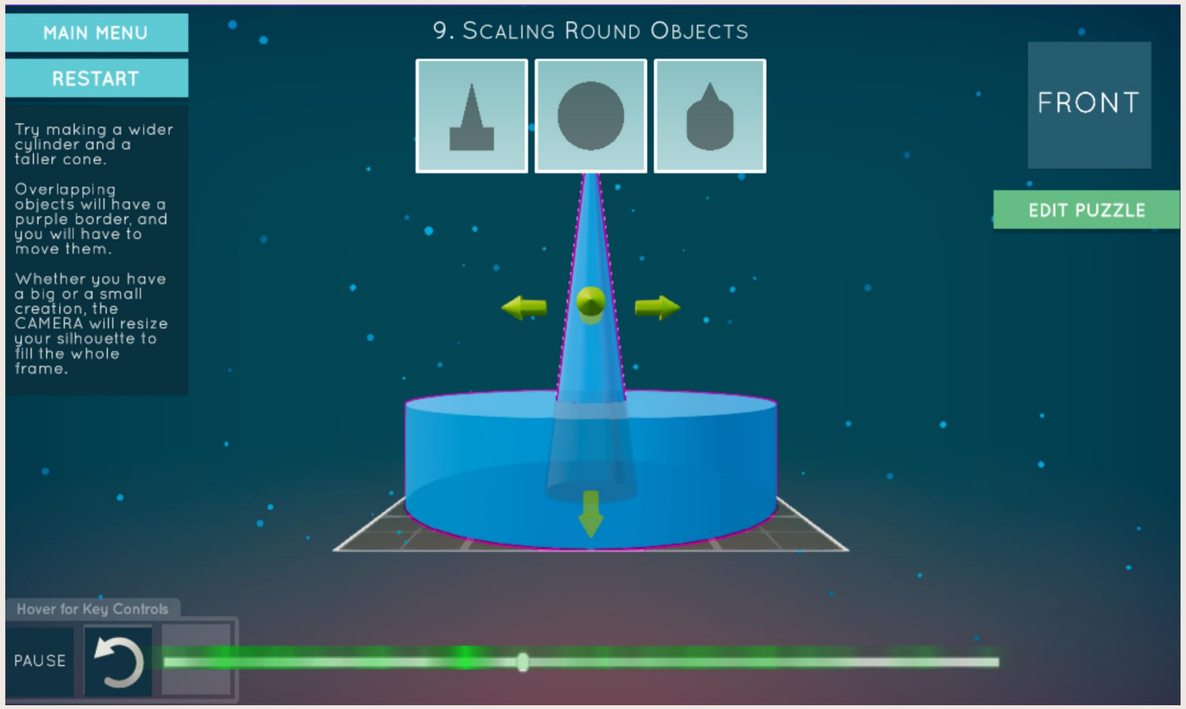
Title	Utilising Explainable AI to Enhance Real-Time Student Performance Prediction in Educational Serious Games
Authors	Manuel J. Gomez, Álvaro Armada Sánchez, Mariano Albaladejo-González, Félix J. García Clemente, José A. Ruipérez-Valiente

Journal	Expert Systems
DOI	10.1111/exsy.70008

Work IV ➡ Game used

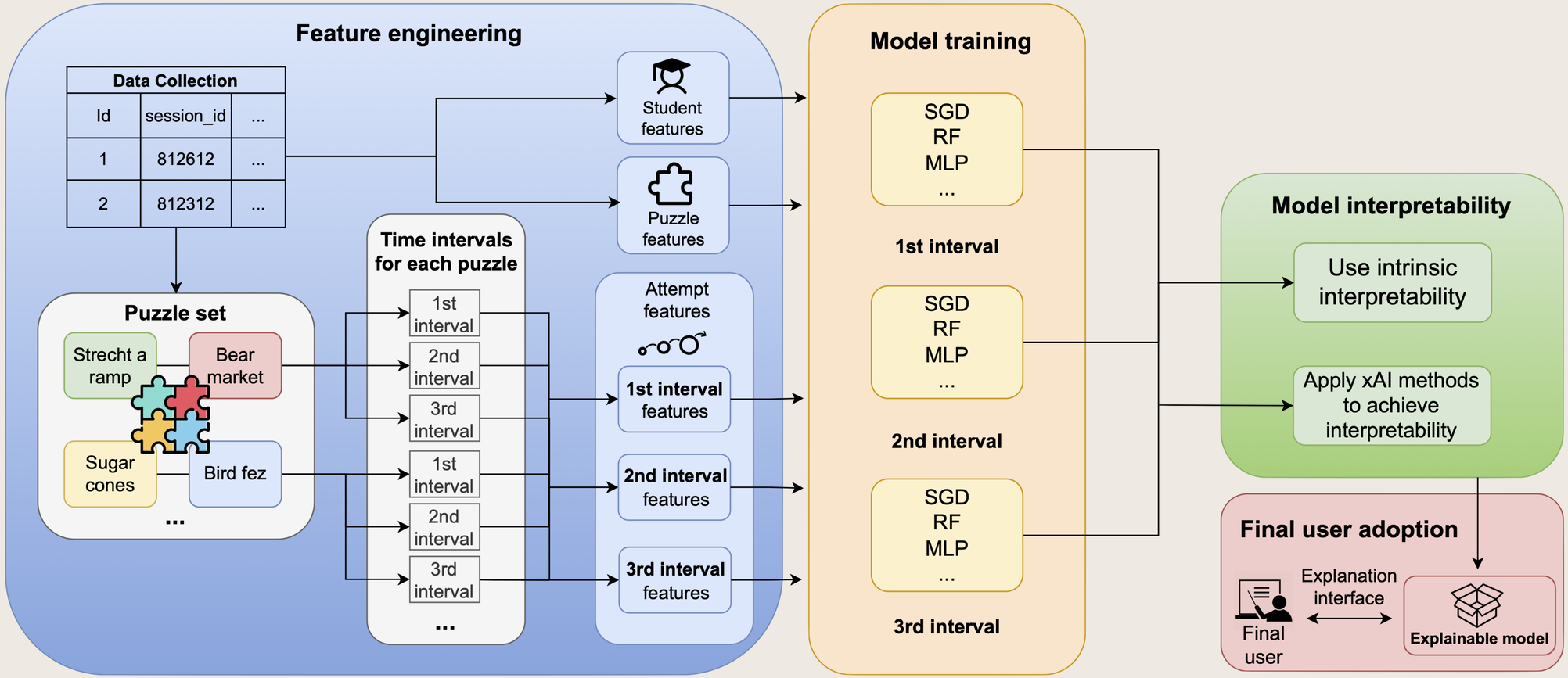
Shadowspect





Work IV ➡ Overview

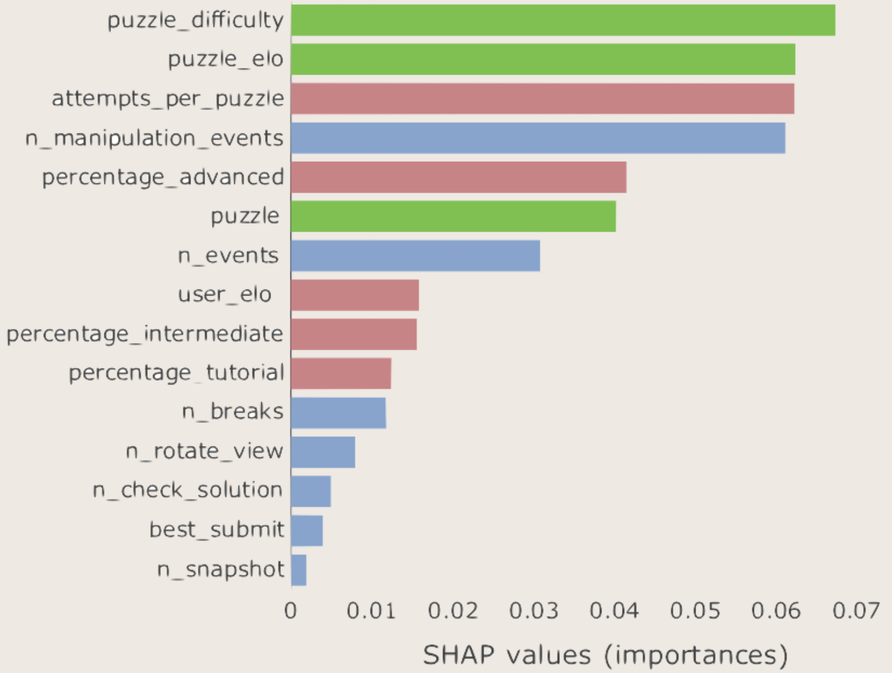
- Build a real–time **performance prediction** model for SGs
- Enable **explainability** for non-technical users



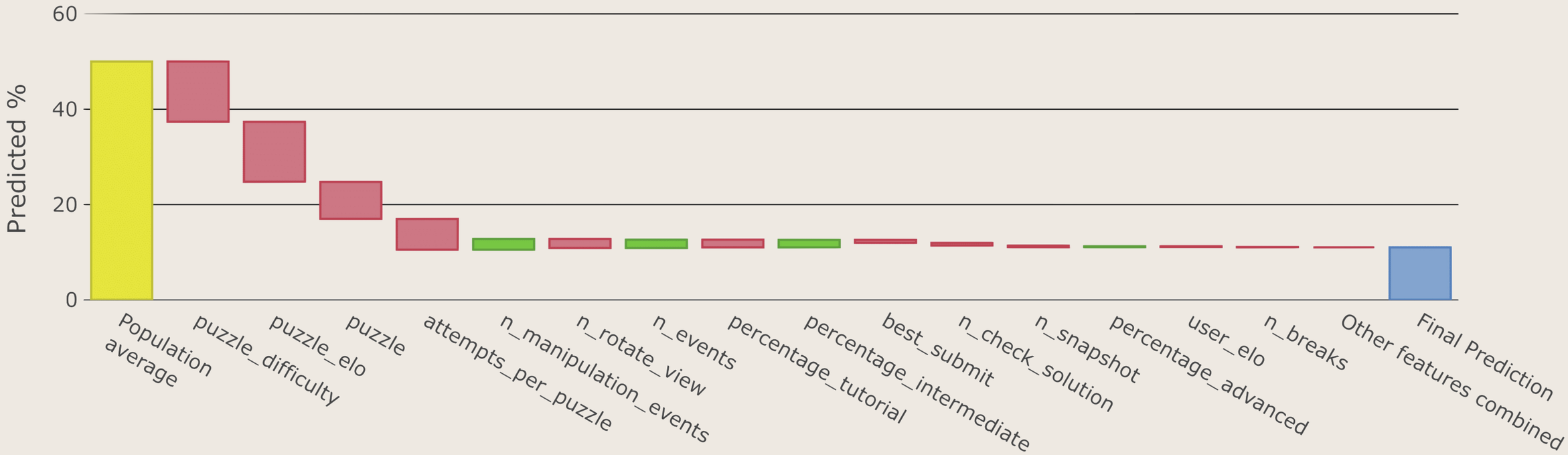
Work IV ➡ Results

Model	Balanced accuracy	F1 score	MCC	Precision	Sensitivity	Specificity
Prediction 1 (25th)	0.7721	0.8445	0.5439	0.8449	0.8441	0.7
Prediction 2 (50th)	0.7918	0.8288	0.5912	0.8026	0.8567	0.7268
Prediction 3 (75th)	0.7928	0.8020	0.5877	0.7723	0.8340	0.7516

- **Random Forest (RF)** was the best performing algorithm
- Not explainable model -> apply **XAI**
- **SHAP** values for explaining individual predictions

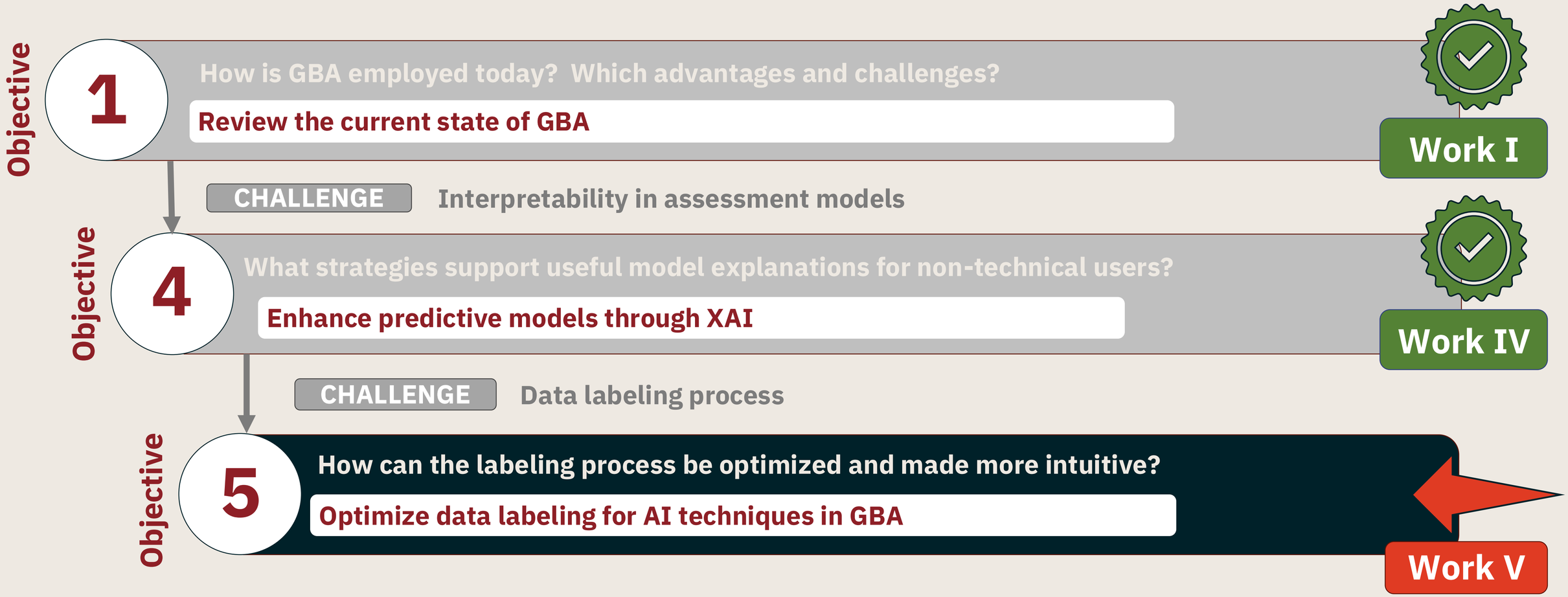


Work IV ➡ Use case

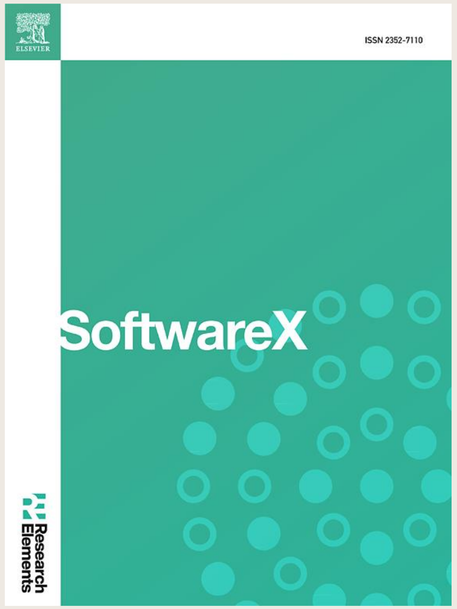


- **Contribution** of each **feature** to the prediction
- **Understanding** of the model reasoning at an individual level

Work V


Objective
1
How is GBA employed today? Which advantages and challenges?
Review the current state of GBA
Work I
CHALLENGE
Interpretability in assessment models
Objective
4
What strategies support useful model explanations for non-technical users?
Enhance predictive models through XAI
Work IV
CHALLENGE
Data labeling process
Objective
5
How can the labeling process be optimized and made more intuitive?
Optimize data labeling for AI techniques in GBA
Work V

Work V ➡ Publication



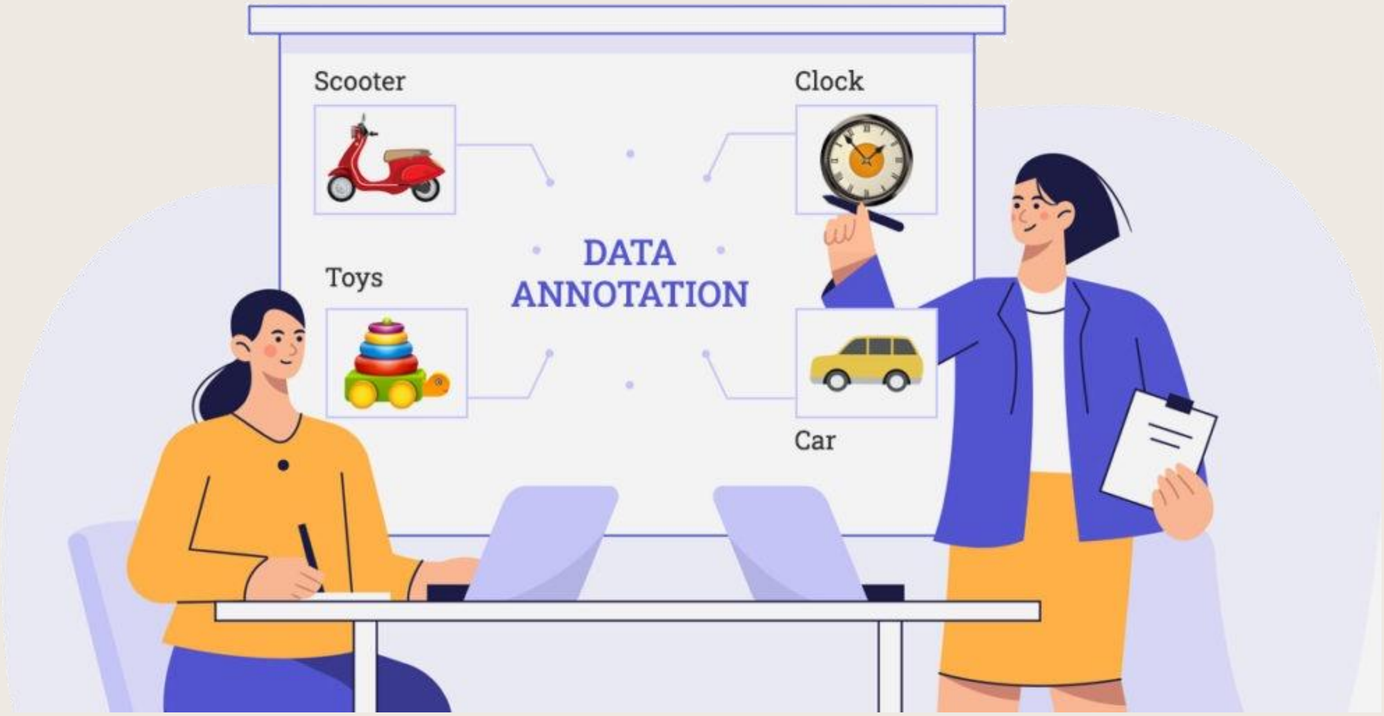
Title	Optimizing multimedia and gameplay data labeling: A web-based tool for Game-Based Assessment
Authors	Manuel J. Gomez, José A. Ruipérez-Valiente, Félix J. García Clemente

Journal	SoftwareX
DOI	10.1016/j.softx.2024.101763

Work V ➡ Data labeling

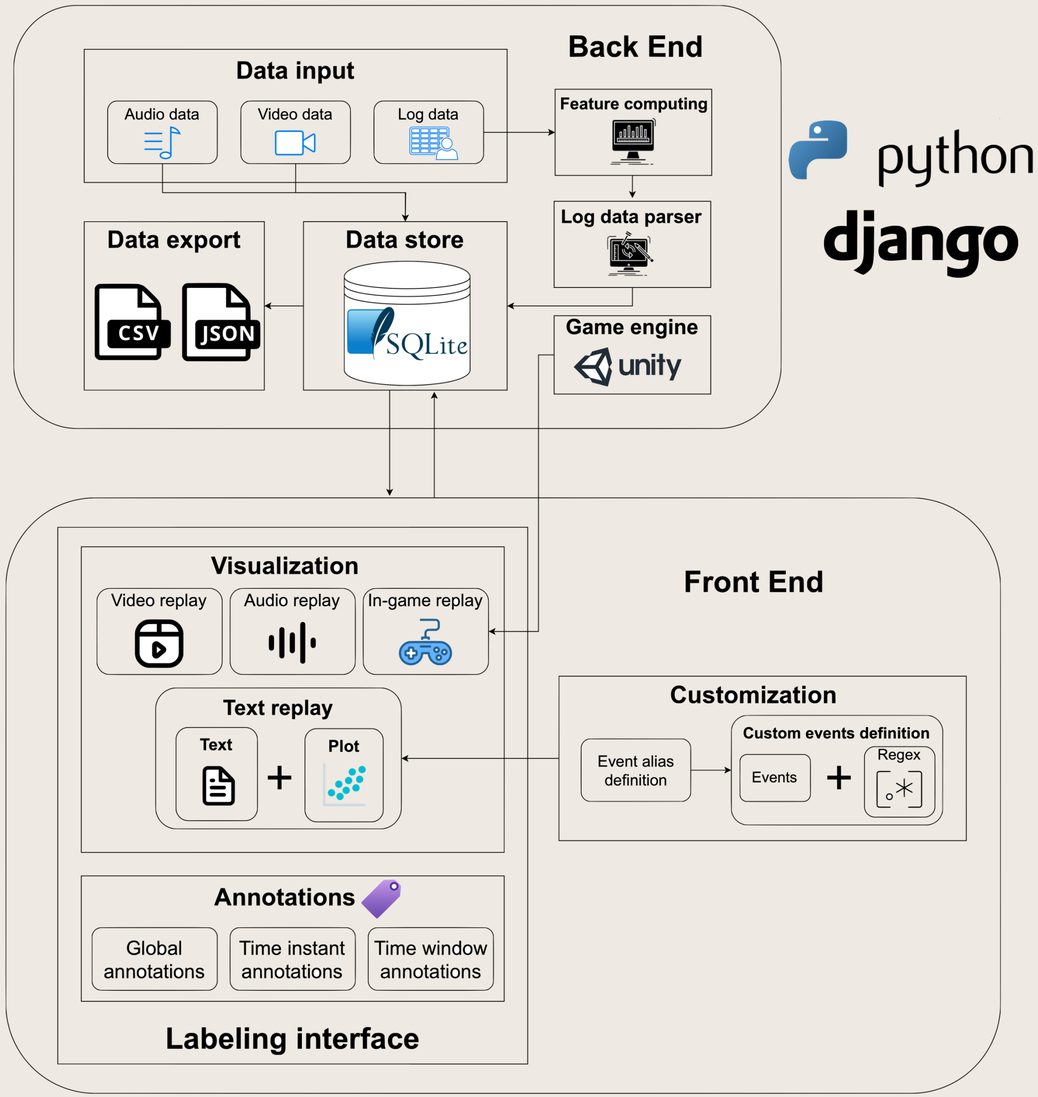
Data labeling (or annotation)

Assigning tags or categories to raw data (images, video, text, audio) so that AI can learn



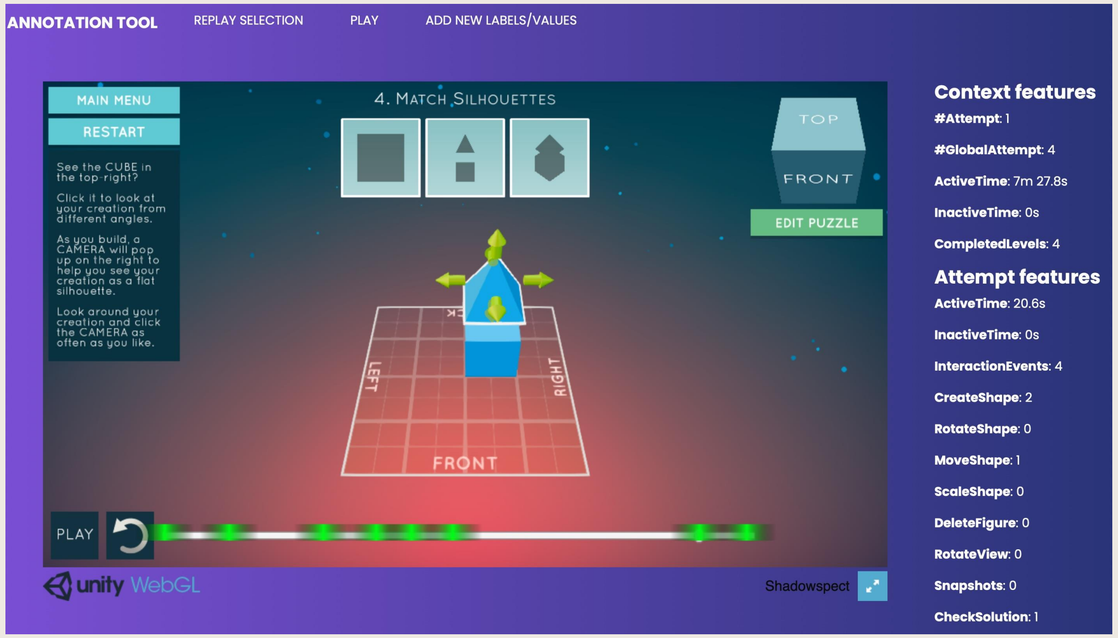
Work V ➡ Architecture

Open-source tool



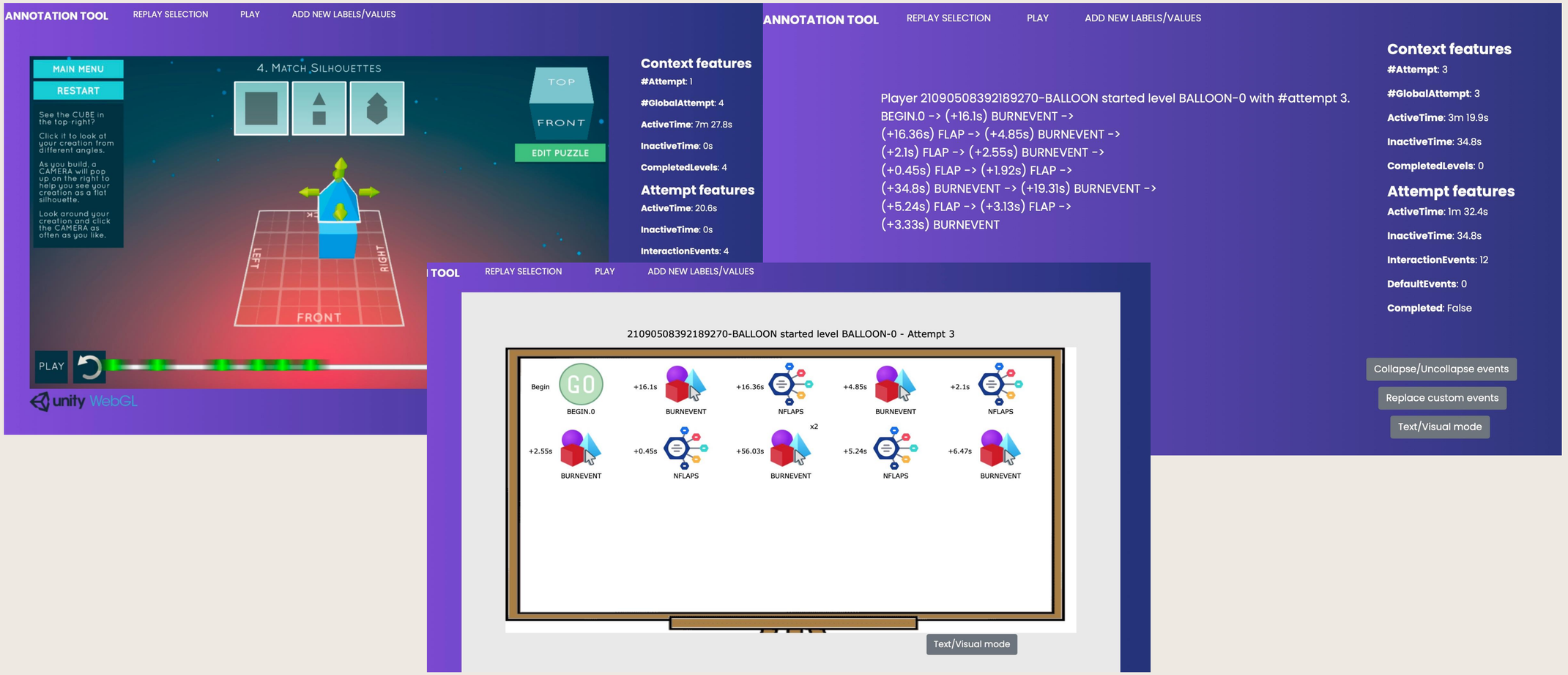
Work V ➡ Platform

Audio, video, and log-event data



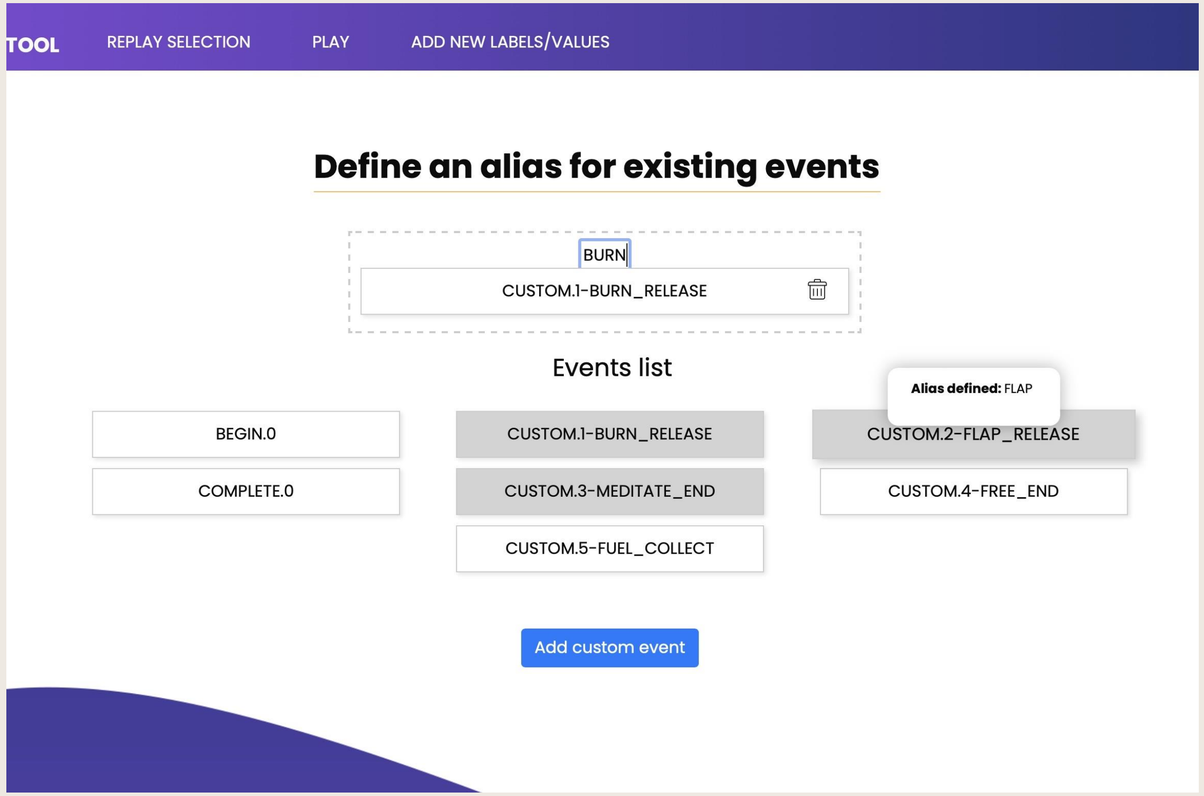
Work V ➡ Platform

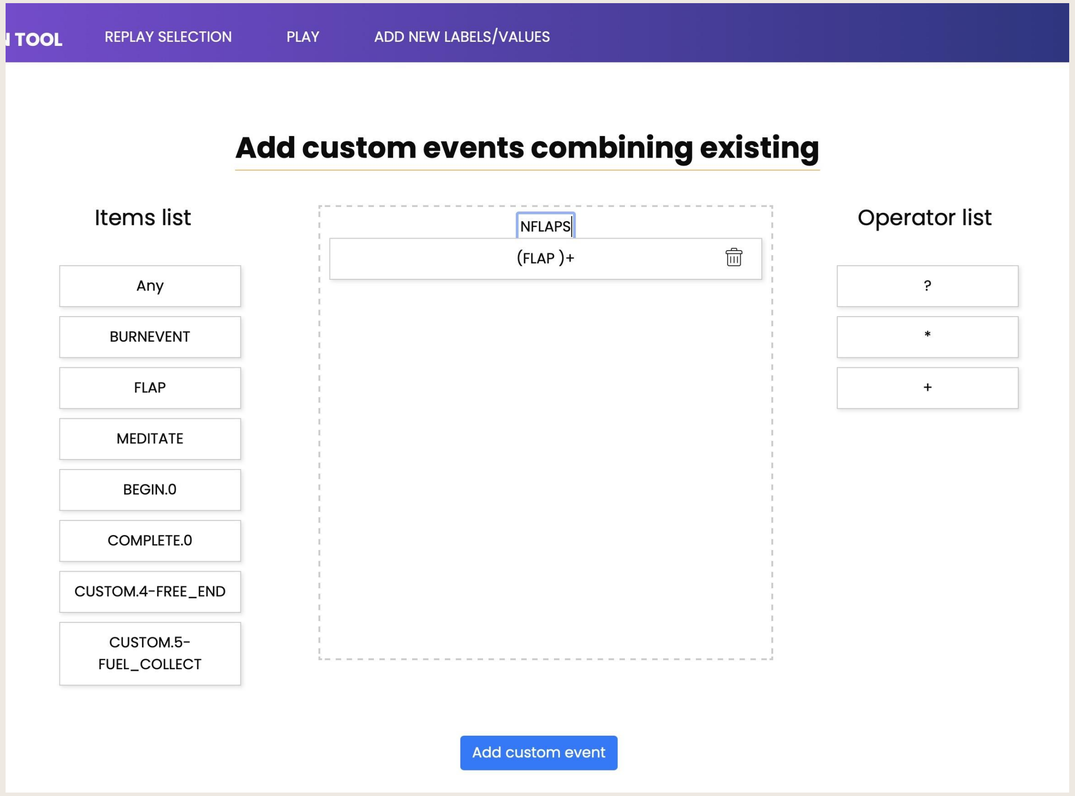
Different visualizations



Work V ➡ Platform

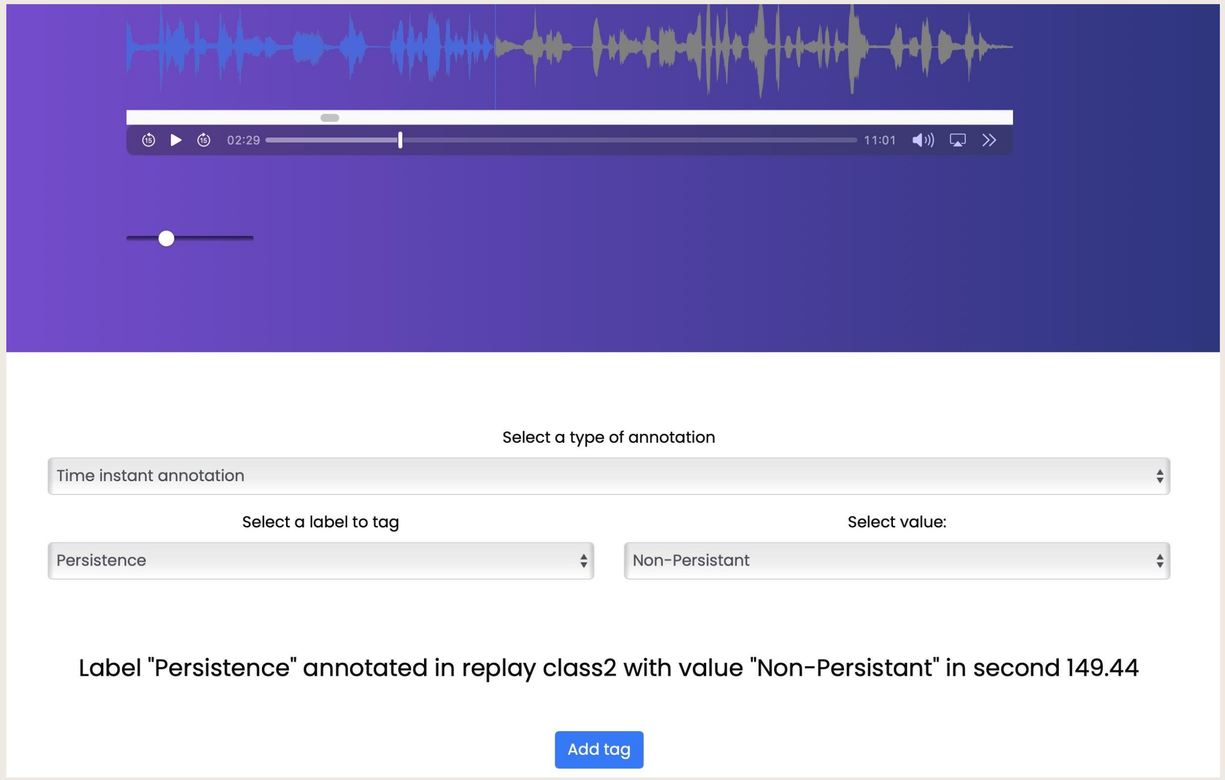
Customization

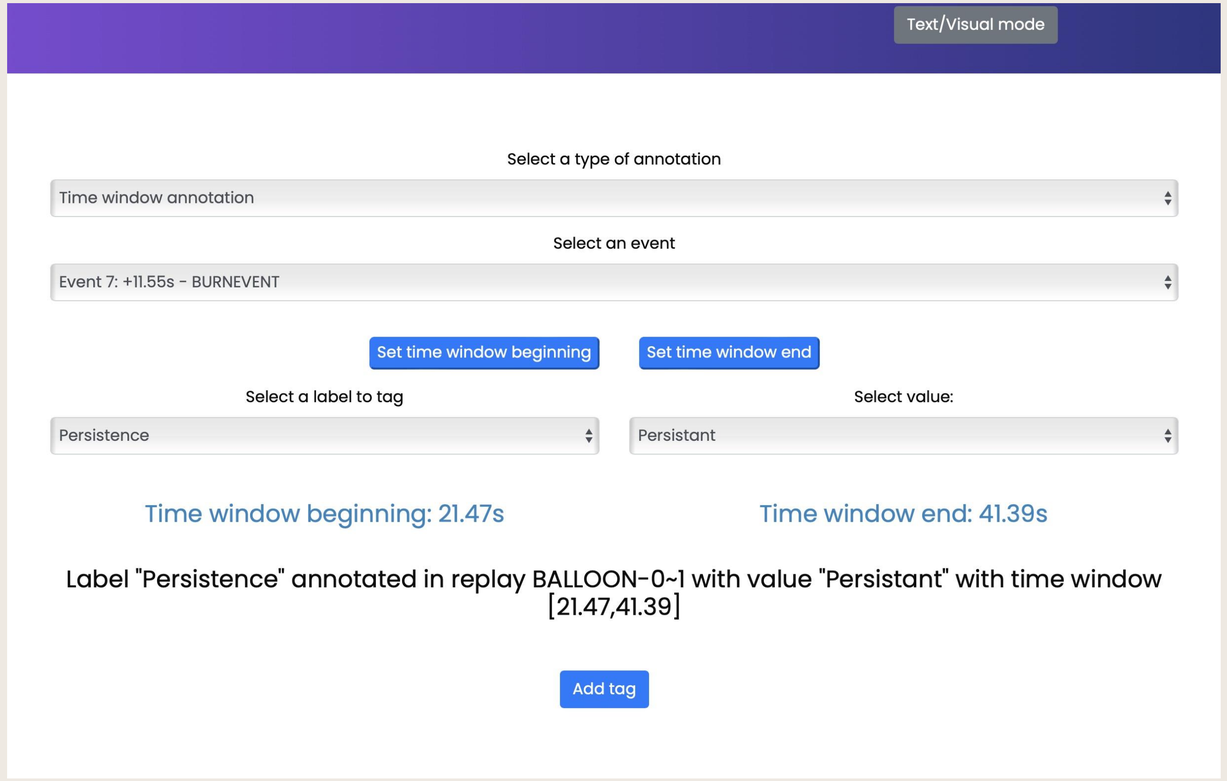




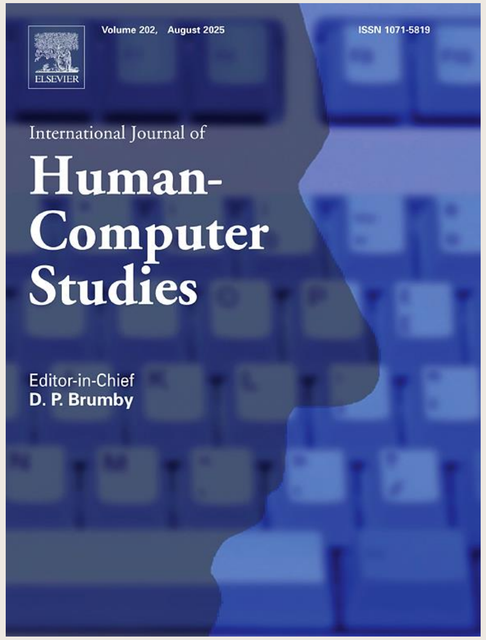
Work V ➡ Platform

Labeling options





Additional publication



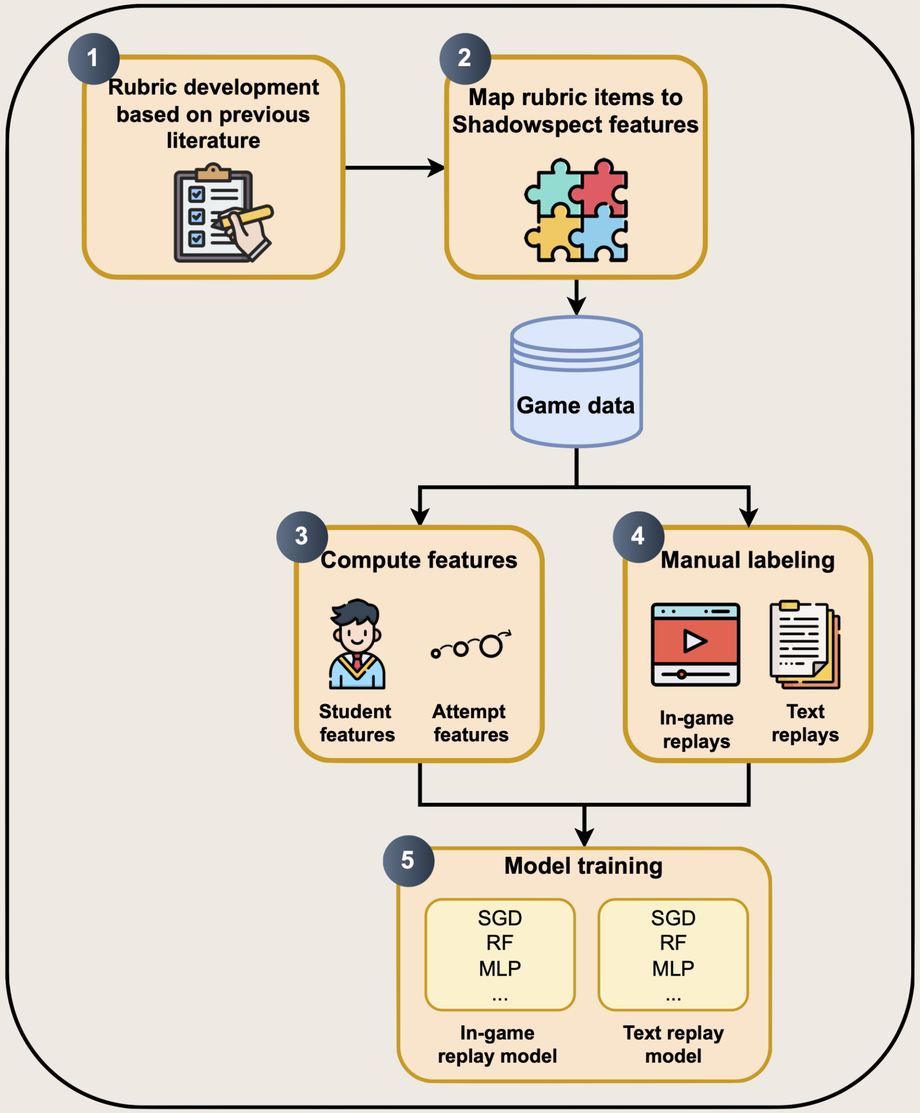
Title	*Modeling Persistence Behavior in Serious Games: A Human-Centered Approach Using In-Game and Text Replays*
Authors	Manuel J. Gomez, Mariano Albaladejo-González, Félix J. García Clemente, José A. Ruipérez-Valiente,

Journal	International Journal of Human-Computer Studies
DOI	10.1016/j.ijhcs.2025.103601

Additional publication

Four Objectives

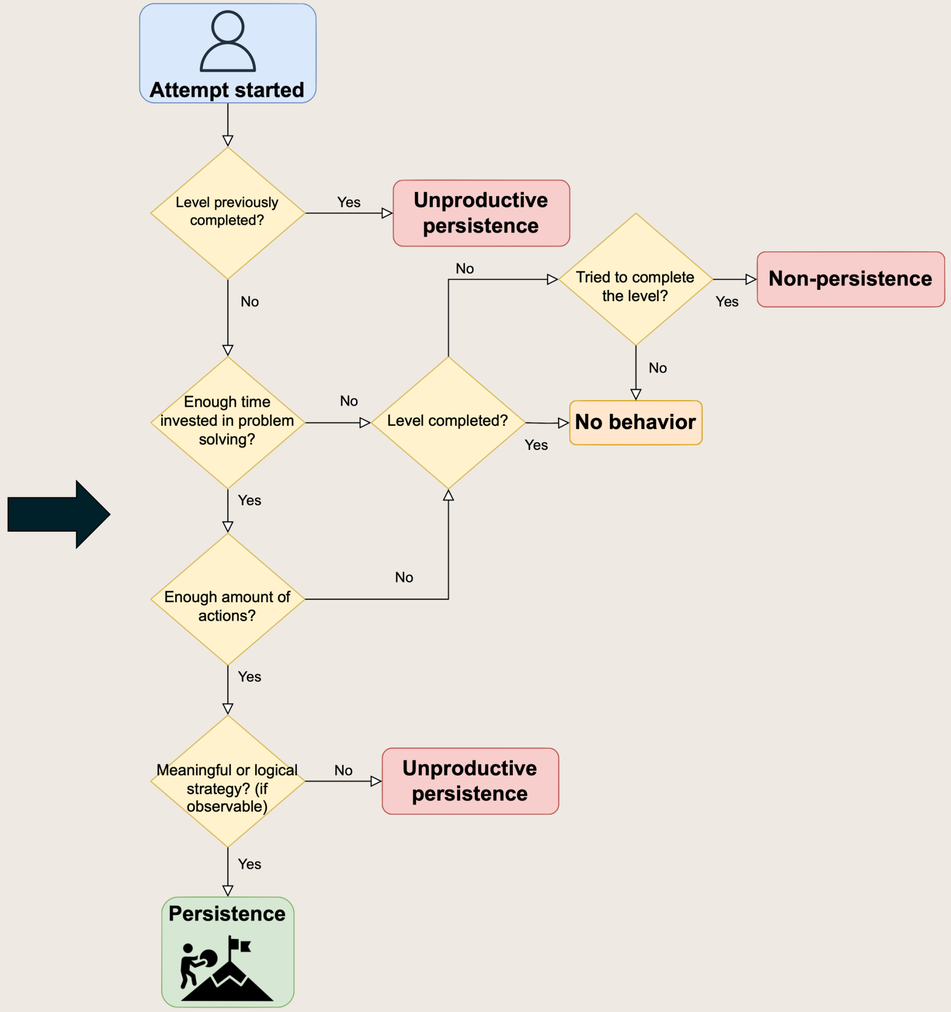
- **Persistence rubric** modeling
- **Mapping** persistence aspects **to game features**
- **Manual labeling**
- **ML model training and evaluation**



Additional publication

Rubric modeling

Category	Characteristic	Description	Source(s)
Persistence	Time Investment	Total time actively solving a problem.	DiCerbo (2014); Howard and Crayne (2019); Ventura et al. (2013); Ventura and Shute (2013); Constantin et al. (2011); Shute and Wang (2016)
	Re-attempt Indicators	Number of attempts or restarts in problem solving.	DiCerbo (2014); Constantin et al. (2011); Klein-Latucha and Hershkovitz (2024); Shute and Wang (2016); Israel-Fishelson and Hershkovitz (2020)
Non-Persistence	Early Abandonment	Abandoning a problem-solving task prematurely.	Ventura et al. (2013); DiCerbo (2014); Klein-Latucha and Hershkovitz (2024)
	Inactive Solving	Lack of active engagement in problem-solving.	DiCerbo (2014)
Unproductive Persistence	Senseless Actions	Purposeless or repetitive actions in problem-solving.	Howard and Crayne (2019); Kim and Miklasz (2021)
	Repetition after Completion	Repeating problems after successfully solving them.	DiCerbo (2014); Klein-Latucha and Hershkovitz (2024); Shute and Wang (2016)



Additional publication

Manual labeling

Count of attempts by labeled category in both types of replays.

Category	In-game replays (%)	Text replays (%)	Total (%)
No behavior	892 (64.9%)	911 (66.3%)	1803 (65.6%)
Persistence	246 (17.9%)	220 (16.0%)	466 (17.0%)
Non-persistence	221 (16.1%)	233 (17.0%)	454 (16.5%)
Unproductive persistence	15 (1.1%)	10 (0.7%)	25 (0.9%)
Total	**1374 (100%)**	**1374 (100%)**	**2748 (100%)**

ML model results

Test results for the in-game replays model.

Category	F1-score	Precision
No Behavior	0.94	0.96
Non-persistence	0.88	0.91
Persistence	0.70	0.65
Unproductive Persistence	0.33	0.22
Balanced accuracy	0.8	

Test results for the text replays model.

Category	F1-score	Precision
No Behavior	0.97	0.97
Non-persistence	0.94	0.95
Persistence	0.81	0.83
Unproductive Persistence	0.18	0.12
Balanced accuracy	0.76	

Conclusions

Conclusion 1

Need for more standard assessment frameworks

Conclusion 2

Increasing value for education and training

Conclusion 3

Increasing complexity (skills and assessment)

Conclusion 4

Interpretability as a key enabler for real world

Cyber Data Lab

UNIVERSIDAD DE MURCIA

Future lines

Future line 1

Frameworks for scalable design and integration of GBAs

Future line 2

Deployment in real environments

Future line 3

Human-in-the-Loop (HITL) GBA approaches

Future line 4

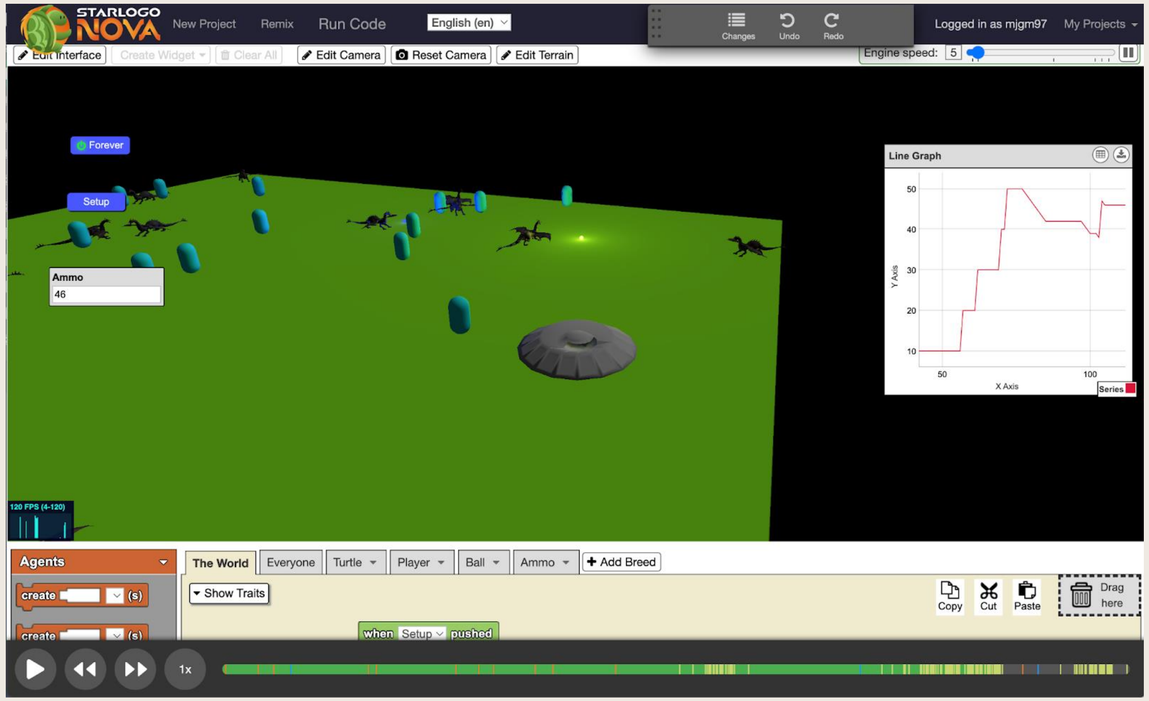
Multimodal GBAs

Beyond PhD thesis ➡ Research stay

STARLOGO NOVA

SCHELLER TEACHER EDUCATION PROGRAM

IEEE EDUCON 2025
IEEE Global Engineering Education Conference
London, United Kingdom || 22-25 April 2025 || Queen Mary University of London



Gomez, M. J., Wendel, D., & Klopfer, E. (2025, April). **Capturing and Analyzing User Interactions in Block-Based Programming With Starlogo Nova**. In *2025 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1-5). IEEE.

Capturing and Analyzing User Interactions in Block-Based Programming with StarLogo Nova

Manuel J. Gomez
Department of Information and Communications Engineering University of Murcia Murcia, Spain
manueljesus.gomezm@um.es

Daniel Wendel
Scheller Teacher Education Program Massachusetts Institute of Technology Cambridge, MA 02139, USA djwendel@mit.edu

Eric Klopfer
Scheller Teacher Education Program Massachusetts Institute of Technology Cambridge, MA 02139, USA klopfer@mit.edu

***Abstract*—Block-based programming environments, such as Scratch and StarLogo Nova, are popular in STEM education for introducing algorithmic thinking to non-technical users. However, capturing and understanding students' interactions with these environments in real-time is challenging for educators, limiting their ability to provide effective feedback and support. In this research, we developed an integrated replay system within StarLogo Nova that records and replays student interactions, providing teachers with the ability to review and reflect on these interactions in detail. Additionally, the system also includes analytics features that help educators summarize user interactions and identify key moments, offering perspectives that go beyond what a standard screen capture can provide. We conducted a preliminary evaluation of the system to assess its accuracy in capturing user interactions, confirming its effectiveness for educators in analyzing student behavior. The results are promising, showing improved analytics and a 99.65% reduction in file size compared to traditional screen capture.**

skills during programming [7]. The lack of tools for tracking and analyzing student behavior in block-based programming environments limits teachers to providing feedback only on final products, rather than on process, especially in classrooms with a large number of students. Thus, students' actions need to be detected and measured as students work so that formative feedback can be provided to promote learning [8].

Although block-based programming tools usually offer minimal support for analyzing users' coding processes, previous research has explored different approaches to address this gap. For example, Scratchlog [9] collects data from learners using Scratch, such as the history of code edits and statistics about how the Scratch user interface was used. Gross et al. [10] developed a novel feature for Looking Glass, a 3D programming environment that teaches programming fundamentals through

Cyber Data Lab

UNIVERSIDAD DE MURCIA

Beyond PhD thesis ➡ Publications

11 journal articles

7 conference papers

1 book chapter

GBA

EdTech

LA

aladejo-González, ez-Valiente, J. A. enhance real-time student performance prediction in educational serious ... *th ACM/SIGAPP* -94).

mente, F. J. G. (2022). A systematic literature review of game-based ... d challenges. *IEEE ...es*, *16*(4), 500-515.

Ruipérez-Valiente, J. A., Gómez, M. J., Martínez, P. A., & Kim, Y. J. (2021). Ideating and developing a visualization dashboard to support teachers using educational games in the classroom. *IEEE access*, *9*, 83467-83481.

Martınez, P. A., Gómez, M. J., Ruipérez-Valiente, J. A., Pérez, G. M., & Kim, Y. J. (2020). Visualizing educational game data: A case study of visualizations to support teachers. In *CEUR Workshop Proceedings* (Vol. 2671, pp. 97-111).

Gomez, M. J., Ruipérez-Valiente, J. A., & García Clemente, F. J. (2023). A framework to support interoperable Game-based Assessments as a Service (GBAaaS): Design, development, and use cases. *Software: Practice and Experience*, *53*(11), 2222-2240.

Gomez, M. J., Albaladejo-González, M., Clemente, F. J. G., & Ruipérez-Valiente, J. A. (2025). Modeling persistence behavior in serious games: A human-centered approach using in-game and text replays. *International Journal of Human-Computer Studies*, 103601.

Gomez, M. J., Calderón, M., Sánchez, V., Clemente, F. J. G., & Ruipérez-Valiente, J. A. (2022). Large scale analysis of open MOOC reviews to support learners' course selection. *Expert Systems with Applications*, *210*, 118400.

Gomez, M. J., Armada Sánchez, Á., Albaladejo-González, M., García Clemente, F. J., & Ruipérez-Valiente, J. A. (2025). Utilising Explainable AI to Enhance Real-Time Student Performance Prediction in Educational Serious Games. *Expert Systems*, *42*(3), e70008.

Gomez, M. J., Ruipérez-Valiente, J. A., & Clemente, F. J. G. (2024). Optimizing multimedia and gameplay data labeling: A web-based tool for Game-Based Assessment. *SoftwareX*, *27*, 101763.

Gomez, M. J., Ruipérez-Valiente, J. A., & Clemente, F. J. G. (2024). Developing and validating interoperable ontology-driven game-based assessments. *Expert Systems with Applications*, *248*, 123370.

Gomez, M. J., Ruipérez-Valiente, J. A., & García Clemente, F. J. (2023, July). Towards Game-based Assessment at Scale. In *Proceedings of the Tenth ACM Conference on Learning@ Scale*(pp. 297-301).

Aguilar, A. M., Gomez, M. J., & Ruipérez-Valiente, J. A. (2025). Towards the development of an assessment model of spatial reasoning with a serious game. In *Proceedings of TEEM 2024* (TEEM2024 2024, Lecture Notes in Educational Technology). Springer

Gomez, M. J., Ruipérez-Valiente, J. A., & Clemente, F. J. G. (2021, September). Bibliometric Analysis of the Last Ten Years of the European Conference on Technology-Enhanced Learning. In *European Conference on Technology Enhanced Learning* (pp. 337-341). Cham: Springer International Publishing.

Gomez, M. J., Wendel, D., & Klopfer, E. (2025, April). Capturing and Analyzing User Interactions in Block-Based Programming With Starlogo Nova. In *2025 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1-5). IEEE.

Gomez, M. J., Ruipérez-Valiente, J. A., Martinez, P. A., & Kim, Y. J. (2020, October). Exploring the affordances of sequence mining in educational games. In *Eighth international conference on technological ecosystems for enhancing multiculturality* (pp. 648-654).

Gomez, M. J., Ruipérez-Valiente, J. A., Martínez, P. A., & Kim, Y. J. (2021). Applying learning analytics to detect sequences of actions and common errors in a geometry game. *Sensors*, *21*(4), 1025.

Gomez, M. J., & Ruipérez-Valiente, J. A. (2022). Analyzing the Evolution of Digital Assessment in Education Literature Using Bibliometrics and Natural Language Processing. In *Handbook of Research on Digital-Based Assessment and Innovative Practices in Education* (pp. 178-200). IGI Global Scientific Publishing.

Gomez, M. J., Ruipérez-Valiente, J. A., & García Clemente, F. J. (2023). Exploring Technology-and Sensor-Driven Trends in Education: A Natural-Language-Processing-Enhanced Bibliometrics Study. *Sensors*, *23*(23), 9303.

Gomez, M. J., Ruiperez-Valiente, J. A., & Clemente, F. J. G. (2022). Analyzing trends and patterns across the educational technology communities using Fontana framework. *IEEE Access*, *10*, 35336-35351.

Cyber Data Lab

RSIDAD DE MURCIA

Towards Interoperability and Novel Methodological Approaches for Scalable Game-Based Assessment

Hacia la interoperabilidad y nuevos enfoques metodológicos para la evaluación escalable basada en juegos

Ph.D. Thesis

Manuel Jesús Gómez Moratilla

Advisors: Félix Jesús García Clemente and
José Antonio Ruipérez Valiente

UNIVERSIDAD DE MURCIA

f SéNeCa(+)
Agencia de Ciencia y Tecnología
Región de Murcia