

* The program is subject to minor changes, so please check it regularly. The schedule for keynote speakers may change slightly.

GENERAL PROGRAM - SIMBig 2025

Day 1: Octobre 29, 2025

Hour	Title	Speaker
08h30 - 09h00	Reception	
09h00 - 09h15	Welcoming to SIMBig2025	Authorities of the UPC and the Co-Chairs of SIMBig 2025
Session 1 - DISE		
09h30 - 09h45	Set of Guidelines for the Development of a Web-Based Assistant Focused on the Creation of Dubbing Scripts	Ramos Ramirez, Renzo Manuel; Roca Huapaya, Orlando Arturo; Díaz Suárez, Jorge Eduardo*
09h45 - 10h00	ASSIST: A Multi-Agentic Framework for Human–Computer Interaction in Cultural Heritage Settings	Samuel Ramos Varela, Anmol Guragain, Jaime Bellver, David Aragón Díaz, Long Lin, and Luis Fernando D'Haro
Session 2 - Core Machine Learning		
10h00 - 11h00	<i>Invited Speaker:</i> Ariel PROCACCIA <i>Harvard University, USA</i>	 HARVARD UNIVERSITY
	Thinking Outside the Ballot Box: How should one design unprecedented democratic processes capable of handling enormous sets of alternatives like all possible policies, bills, or statements? I argue that this challenge can be addressed through a framework called generative social choice, which fuses the rigor of social choice theory with the flexibility and power of large language models. I then explore an application of generative social choice to the problem of identifying a proportionally representative slate of opinion statements. This includes a discussion of desired properties, an algorithm that provably achieves them, an implementation using GPT, and insights from an end-to-end pilot. By providing guarantees, generative social choice could alleviate concerns about AI-driven democratic innovation and help unlock its potential.	
11h00 - 11h15	Networking and Coffee Break	
11h15 - 11h30	Geometric goodness-of-fit measure for interval-valued data	Benavides Castillo, Dylan*; Solís Hernández, Maikol; Rodríguez Rojas, Oldemar
11h30 - 11h45	Multi-Objective Optimization for Strategic Shelter Placement: Case Study of Lima's Earthquake Preparedness	Espezua, Soledad*; Sanjinez, Alexandra Constanza; Checcllo, Amy Antalu; Rios, Alexia Shariann
11h45 - 12h00	Solving the Qubit Transmission Problem with Biologically Inspired Neural Networks and Reinforcement Learning	Fiuri, Ariel*; Osenda, Omar; Dominguez, Martín; Perón Santana, Sofía
Session 3 - Knowledge Graphs		
12h00 - 12h15	Horoyah: A Knowledge Graph-based Framework for Election Monitoring in Low and Middle-Income Countries	Diallo, Gayo*
12h15 - 12h30	Semantic Similarity Analysis of Theses Using Sentence Embeddings and Knowledge Graphs	Cusi Diaz, Ibeth Janela*; Mamani Salcedo, Lizeth Carla; Zuniga Rojas, Gabriela; Alvarez Mamani, Edwin; Era Olivera, Harley
12h30 - 12h45	Measuring Healthcare Accessibility in South Africa through Clustering Techniques and Multi-Layer Network Analysis	Nunez-del-Prado, Miguel*
12h45 - 14h15	Lunch	
14h00 - 14h15	Intelligent query system for consumer protection regulations in Perú using hierarchical vector databases	Antonio Arroyo-Paz*, Franco Cárdenas-Martínez
14h15 - 14h30	Faithfulness and Relevance in Stable Normative Domains: A Comparative Analysis of CAG vs. RAG	Collanqui Casapia, Elmer; Coaguila Mamani, Maribel; Flores Garcia, Anibal; Apaza, Honorio*
14h30 - 15h30	<i>Invited Speaker:</i> Marc NAJORK <i>Google, USA</i>	
	Generative Information Retrieval: This talk provides a brief history of Information Retrieval and of Natural Language Understanding, two areas of Computer Science that evolved separately and are now connecting to enable a new class of Question Answering systems. Generative Information Retrieval systems answer a user's information need directly, rather than referring the user to relevant sources. These systems draw both on parametric knowledge encoded in the weights of large language models, as well as external tools, such as search engines, knowledge bases or program execution sandboxes. The talk also discusses some of the challenges ahead, including the impact of such systems on the web ecosystem.	
15h30 - 15h45	Networking and Coffee Break	
Session 3 - Low-resource & Indigenous Languages		
15h45 - 16h00	A BERT-Based Approach for Quechua Dialect Identification	Santisteban, Julio*; Aranibar Solaligue, Christian
16h00 - 16h15	Quechua Speech Synthesizers	Huaman Curo, Edwin*; Pineda, Ferdinand
Session 4 - Applied NLP and Language Model Applications		
16h15 - 16h30	Detecting explicit content in Spotify song lyrics using natural language processing techniques	Navarro, Aaron*; Usurin, Fernando; Maita, Marcelino; Huarcaya, Dylan
16h30 - 16h45	Spam Detection in YouTube Comments Using a Multinomial Naive Bayes Classifier	Alzamora, Guina*; Condori-Alejo, Henry Ivan
16h45 - 17h00	Prompt Engineering in Spanish for Improving Sexism Detection in Tweets	Limaylla Lunarejo, María*
17h00 - 18h00	Poster Session	

Day 2: October 30, 2025

Hour	Title	Speaker
08h30 - 09h00	Reception	
Session 6 - Biomedicine		
09h00 - 09h15	Beyond Benchmark Accuracy: Evaluation of Ab Initio miRNA Detection Models with Genomic Screening Considerations	Opazo, Juan*; Villanueva, Edwin
09h15 - 09h30	Drug repurposing for rare diseases: A novel knowledge graph explainable approach to identify drug candidates	Drancé, Martin; Zemmari, Akka; Mougín, Fleur; Diallo, Gayo*
09h30 - 10h30	<i>Invited Speaker:</i> Pierre ZWEIGENBAUM <i>CNRS, France</i>	
	Digital Commons for Generative AI on French Clinical Text: PARTAGES is a large French consortium (10 health NLP research teams, 20 hospitals, and 2 startups including Mistral), led by the French National Platform for Health Data (the Health Data Hub). The PARTAGES project, funded by BPI France, aims to build and share open-source French resources related to large language models for health. Given the sensitivity of medical reports, GDPR-related constraints in Europe make it difficult to disseminate LLM-related resources specialized in health. To overcome these difficulties, the project is designed around open-source deliverables that include a medium-size LLM in French trained in the medical field; a corpus of 5,000 fictitious medical reports written by specialists, with annotations for several use cases, to be used for fine tuning; an algorithm specialized in data augmentation of medical reports; seven specialized algorithms corresponding to medical use cases; a platform for federated evaluation of the developed models, with 20 nodes in 20 hospitals; methodological guides to support stakeholders in adapting models to their own business context, their data, their medical reports, or to replicate the approach to other languages.	
10h30 - 10h45	Computational Prediction of Intracellular Signaling Behavior via Machine Learning	Romero, Pamela; Gourdon, Juliette; Yvinec, Romain; Jean-Alphonse, Frederic; Razzaq, Misbah*
10h45 - 11h00	MeSH Concept Relevance and Knowledge Evolution: A Data-driven Perspective	Copara, Jenny*; Naderi, Nona; Falquet, Gilles; Teodoro, Douglas
11h00 - 11h15	Networking and Coffee Break	
Session 5 - Applied NLP and LM Evaluation		
11h15 - 11h30	Fine-Tuning of Small Language Models for Ecological Monitoring: A Comparative Analysis Based on Perplexity	Monroy Barrios, Jhon Edilberto*
11h30 - 11h45	A Variance-Invariance Approach to Better Word Embeddings	Albujasim, Zainab*; Inkpen, Diana
11h45 - 12h00	Multimodal Sign Language Model	Borasino, Christian*; Mancusi, Giorgio; Vilchez, Rody
12h00 - 13h00	<i>Invited Speaker:</i> Dan GOLDSTEIN <i>Microsoft, USA</i>	
	Toward AI Interactions That Build Rather Than Erode Human Cognitive Abilities: Concerns about AI often focus on the risk that delegating cognitive work to machines may erode human thinking and learning. But not all AI interactions are alike. A more productive question might be not whether AI helps or harms learning in general, but under what conditions and through what mechanisms it can lead to improvements in human capabilities, as measured through performance on subsequent, unaided tasks. In this talk, I present findings from studies where participants received assistance from large language models (LLMs) while solving math problems, drafting professional correspondence, or interpreting legal documents. The LLMs used were lightly modified versions of commercial chatbots, not purpose-built tutoring systems. Across all tasks, participants with AI support outperformed control participants on unaided follow-up problems. I propose three mechanisms that may account for these gains. Although AI use can contribute to skill atrophy, it may also facilitate inadvertent learning.	
13h00 - 14h00	Lunch	
14h00 - 14h15	A Systematic Mapping Study on Intelligent User Interfaces: Preliminary Results	Aguilar Encarnacion, Pedro*; Iñiguez, Carlos; Sandobalín, Julio
14h15 - 14h30	Semantic search for financial products using embeddings and pgvector	Arroyo-Paz, Antonio*; Calcina-Puma, Esteven
14h30 - 14h45	Automatic Debate Generation for Spanish Educational Podcasts using LLMs	Lazo Vasquez, Ricardo*; Ochoa Luna, José
14h45 - 15h00	When One Size Doesn't Fit All: A Study on Intrinsic Iterative Self-Correction in LLMs for Hallucination Control	Esina, Svetlana*; Inkpen, Diana
15h00 - 16h00	<i>Invited Speaker:</i> Ravi KUMAR <i>Google, USA</i>	
	Algorithmic Problems in Discrete Choice: In discrete choice, a user selects one option from a finite set of available alternatives, a process that is crucial for recommendation systems applications in e-commerce, social media, search engines, etc. A popular way to model discrete choice is through Random Utility Models (RUMs). RUMs assume that users assign values to options and choose the one with the highest value from among the available alternatives. RUMs have become increasingly important in the Web era; they offer an elegant mathematical framework for researchers to model user choices and predict user behavior based on (possibly limited) observations. While RUMs have been extensively studied in behavioral economics and social sciences, many basic algorithmic tasks remain poorly understood. In this talk, we will discuss various algorithmic and learning questions concerning RUMs.	
16h00 - 16h15	Networking and Coffee Break	
Session 7 - Educational Data & Society		
16h15 - 16h30	Clustering Spatiotemporal Patterns of Infrastructure Investment in Ecuador: A Cantonal-Level Analysis (2008–2022)	Urbina, Myriam; Jerez Calero, María Elena; Gómez Medina, Helen Magdalena; Cuenca, Erick*; Córdova, Arcenio; Cuenca, Víctor

16h30 - 16h45	Where Is the Research Heading? A Text Mining Analysis of International Business Publications in Peru	Giraldo, Hugo; Farromeque, Eva; Ticse, Adriana; Ticse, Grycel; Moscoso Cuaresma, Julio Ricardo*; Nunez-del-Prado, Miguel; Acosta Mantaro, Juan Eduardo; Azabache Morán, Carlos Alberto
16h45 - 17h00	Clustering Analysis of Student Satisfaction–Performance Relationships in Higher Education Using Large-Scale Educational Data	Cutipa-Santi, Michael Newton*; Quelca-Velasquez, Daisy Crishyely; Cutipa-Santi, Yeni; Apaza-Cutipa, Renzo; Ibañez-Quispe, Vladimiro
17h00 - 17h15	Predictive Airfare Analysis Based on Machine Learning Models	Choque-Quispe, Jadyra*; Jallo-Paccaya, Yasumy; Quispe-Mamani, Thalia; Ccansaya-Soncco, Rebeca; Vera-Olivera, Harley; Alvarez-Mamani, Edwin

Day 3: Octubre 31, 2025

Hour	Title	Speaker
08h30 - 09h00	Reception	
Session 8 - Health		
09h00 - 09h15	Mobile App for Detecting Work-Related Stress Using Acoustic Biomarkers and Machine Learning	Rozas Gamero, Claudia Lucia; Heredia Miranda, Oscar; Díaz Suárez, Jorge Eduardo*; Ceh Varela, Eduardo
09h15 - 10h15	<i>Invited Speaker:</i> Diana INKPEN <i>University of Ottawa, Canada</i>	 uOttawa
	Explainable Neural Text Classification: Advances in Large Language Models (LLMs) allow us to develop highly accurate neural text classifiers. One of their major disadvantages is their lack of explainability, due to their black box nature. I am looking into neural text classifiers that are explainable, in order to open their black box architecture, at least partially. Expanability can come at the level of the classification model or at the level of the decision made for each new test data. The explanations need to look into what was learnt from the training data (unless there is no training or minimal training) and also in the pre-trained model (LLM) that was used as a basis for the classifier. To explain the individual decisions for each test data, one step is to calculate feature importance with methods such as LIME, SHAP, or Integrated Gradients. More useful full-text explanations can be generated via customized prompting, or via joint leaning of classes and explanations during training. I will show results for two case studies: applications to legal text mining and to mental health text mining. The evaluation of the generated explanations is done via automatic measures, as well as with human judges, in order to see if they find the explanations relevant and useful.	
10h15 - 10h30	Deep Learning-Based Classification of Fetal Congenital Heart Diseases Using Echocardiographic Images	Vasquez-Solis, Luis*; Anco-Copaja, Edgar; Aures-Garcia, Alvaro; Huertas-Tacchino, Erasmo; Balvin-Landeo, Adelaida
10h30 - 10h45	Evaluation of machine learning algorithms to estimate the prevalence of anemia in children at the district level in Peru	Mamani, Anthony; Gutierrez-Pachas, Daniel Alexis*; Rosero, Luis
10h45 - 11h00	Networking and Coffee Break	
Session 9 - Applied ML & Industry		
11h00 - 11h15	A Predictive Approach Based on Machine Learning for the Prevention of Obesity in University Students Through Personalized Recommendation	Leonel Alfaro Cumba, Jhan Clinton Antonio Salazar, and Sandra Analia Wong Durand
11h15 - 12h15	<i>Invited Speaker:</i> Tina HERNANDEZ-BOUSSARD <i>Stanford University, USA</i>	 Stanford University
	Patient-Specific Digital Twins for Learning Health Systems: Digital twins—virtual representations of individual patients that integrate multimodal clinical, biological, and environmental data—hold immense promise for advancing precision medicine. Yet as we prototype these models, we face the critical challenge of ensuring that patient-specific simulations are not only accurate but also safe, transparent, and ethically governed. This talk will outline ongoing work to design digital twin frameworks that couple causal and mechanistic modeling with real-world EHR data to simulate care trajectories and predict treatment response. I will highlight both the technical foundations and the guardrails necessary for responsible use. Ultimately, patient-specific digital twins represent a path toward continuously learning healthcare, where models are not static predictors but adaptive partners in improving care quality, safety, and fairness.	
12h15 - 12h30	Socioeconomic Factors Associated with Depression in Peru: A Machine Learning-Based Study Using ENDES 2016--2023	Miguel Nunez-del-Prado*, Fernando Bendezú-Galarza
12h30 - 12h45	Conjecture extraction for proof autoformalization	Sorg, Simon; Li, Wenda; Banerjee, Soumya
12h45 - 13h00	University Career Recommendation System using a Vector Database	Arroyo-Paz, Antonio*; Ticona-Hareth, Anthony
13h00 - 13h15	Counting Avocados in Orchard Videos Using a YOLO and BoT-SORT-Based Architecture	Montoya Cubas, Carlos*; Galindo Alvarez, Franco; Albitres Blondet, Diego
13h15 - 13h35	Business session: Big data in Mobile Network Operators	Diego Navarrete, Head of Network Quality, Entel
13h35 - 13h55	Business session: IA en la educación	Carlos Gamero, CDO, Growth, Digital, Laurete
13h55 - 14h15	Business session: Una mirada dentro del equipo analítico más grande del Perú	Jan Nauta, Chapter Area Leader Data Science BCP, BCP
14h15 - 14h30	SIMBig 2025 closing	Authorities of the UPC and the Co-Chairs of SIMBig 2025