#SOMOS2030

INFORMATION ACCESS THROUGH EVALUATION

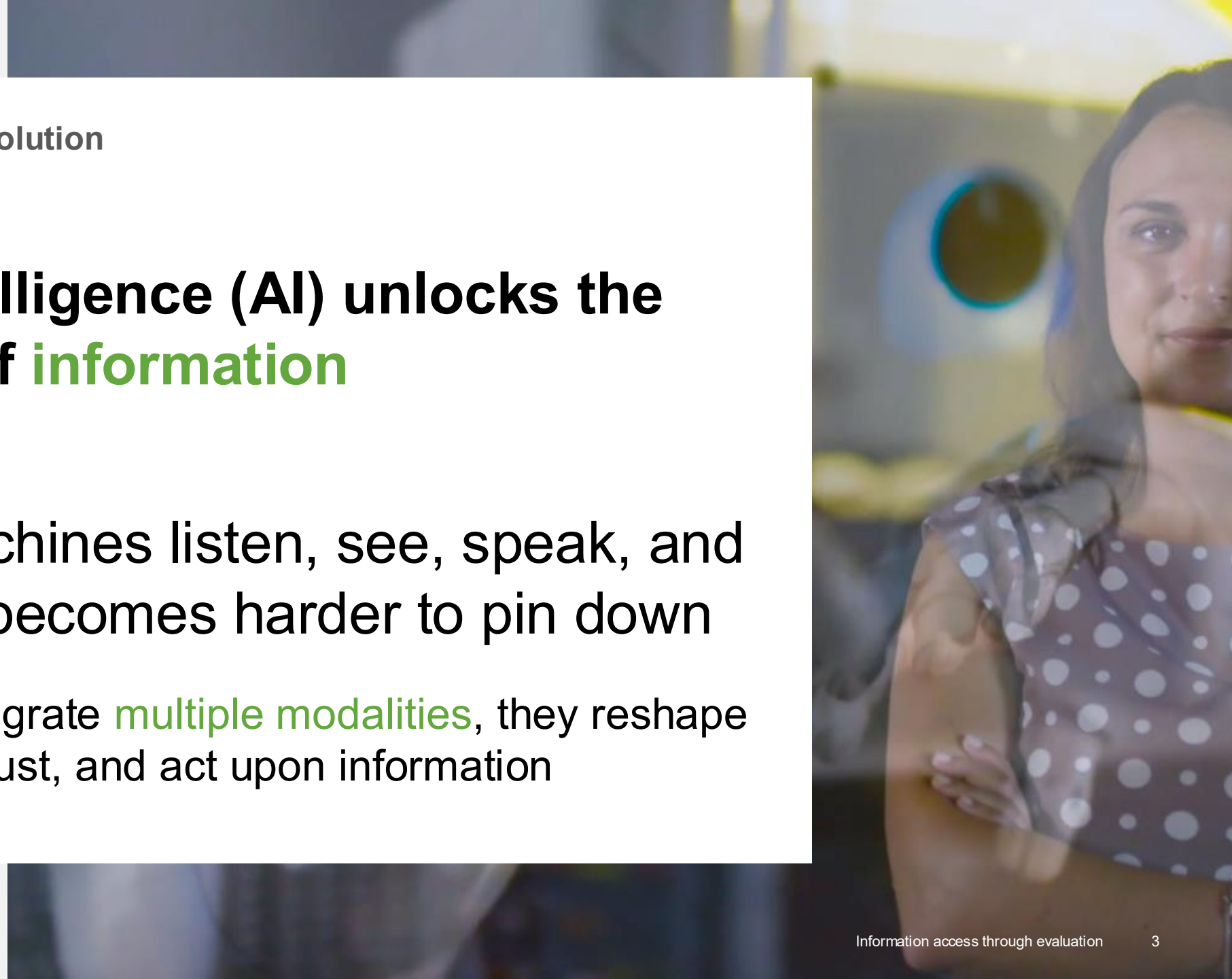
Alba García Seco de Herrera

Truth is in the Eyes of the Machines

Artificial Intelligence (AI) unlocks the true power of information

But when machines listen, see, speak, and decide, truth becomes harder to pin down

As AI systems integrate multiple modalities, they reshape how we access, trust, and act upon information



Why evaluation?

Making observations of all aspects of one's research

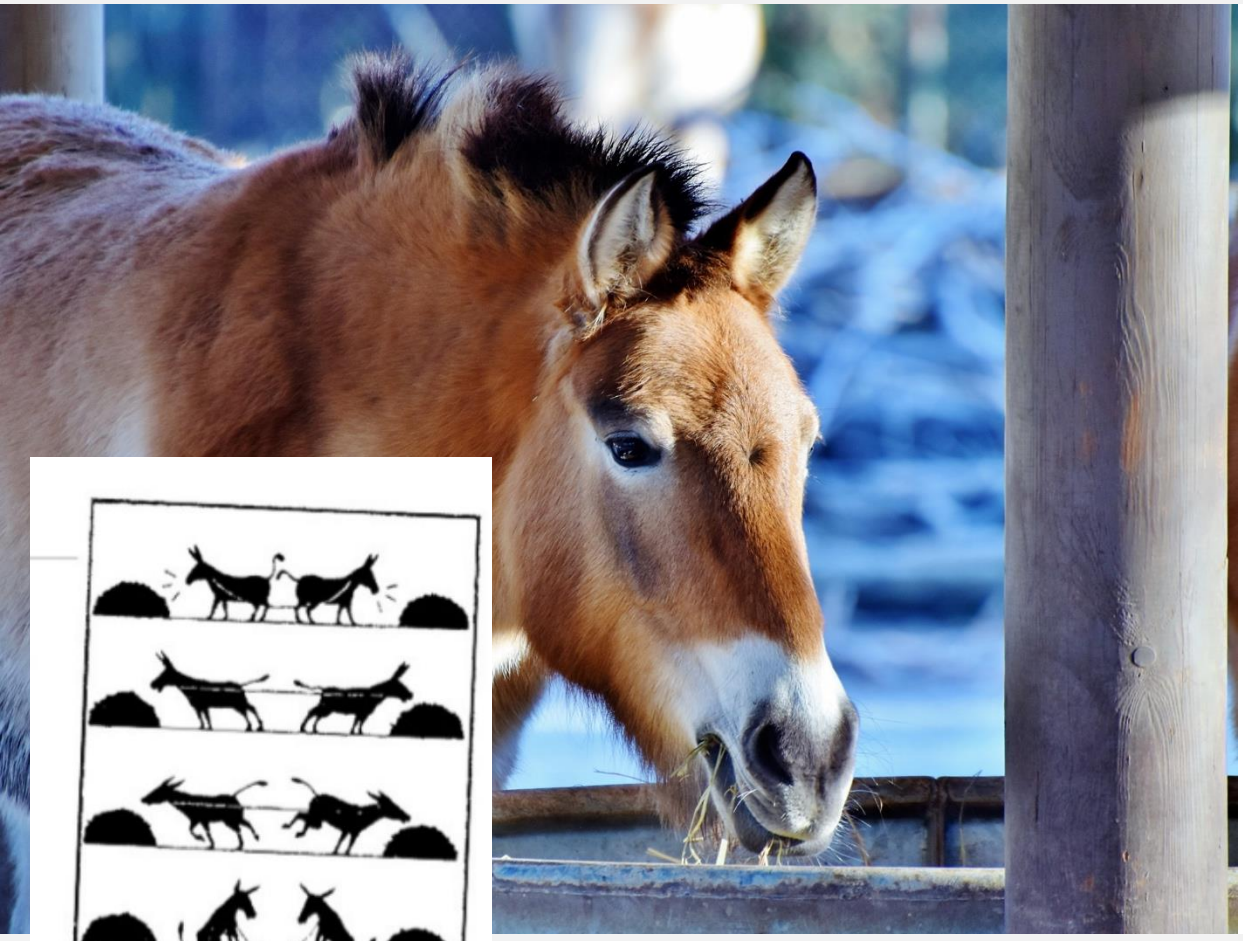
Why Evaluation Matters in the Multimodal AI Era

- Multimodal AI integrates text, image, audio, video, metadata,…
- LLMs + vision models = powerful, but also opaque and vulnerable.
- **Evaluation is not just scoring**—it's shaping development, revealing biases, and guiding responsible innovation

Beyond Performance

Cohen, P.R. and Howe, A.E., 1988. How evaluation guides AI research. *AI magazine*, *9*(4), pp.35-35.

- **How well AI systems perform**
- **Not performance measures:**
  - why we are doing the research?
  - why our tasks are particularly illustrative?
  - why our views and methods are a step forward?
  - whether the system performance is likely to increase or has reached a limit (and why)?
  - what problems we encounter at each stage of our research?



Benchmarking

A community-based and (preferably) community-driven activity involving consensus-based decisions on how to make reproducible, fair, and relevant assessments

Collaborative Evaluation Framework

- Contribute to the success of the task
- Systematic and quantitative evaluation
- Shared tasks on shared resources
  - Reproducible and comparative

Evaluation as a Driver of Progress

- Shared tasks provide standardised benchmarks, community collaboration, and longitudinal tracking
- Examples: CLEF, TREC, MediaEval
- Real-world inspired tasks

CLEF: Conference and Labs of the Evaluation Forum

www.clef-initiative.eu

- Over two decades of promoting reproducibility, collaboration, and innovation in Information Access
  - Realistic benchmarking
  - Multilingual & multimodal focus
  - Involves 200+ research groups worldwide
  - Annual Conference: Combines presentations of lab results, invited talks, and workshops

CLEF 2025

https://clef2025.clef-initiative.eu/

- 9–12 September, Madrid, Spain
- Important dates:
  - Conference abstract submission: 11 May
  - Evaluation cycle ends: Mid May
  - Working notes submission: June

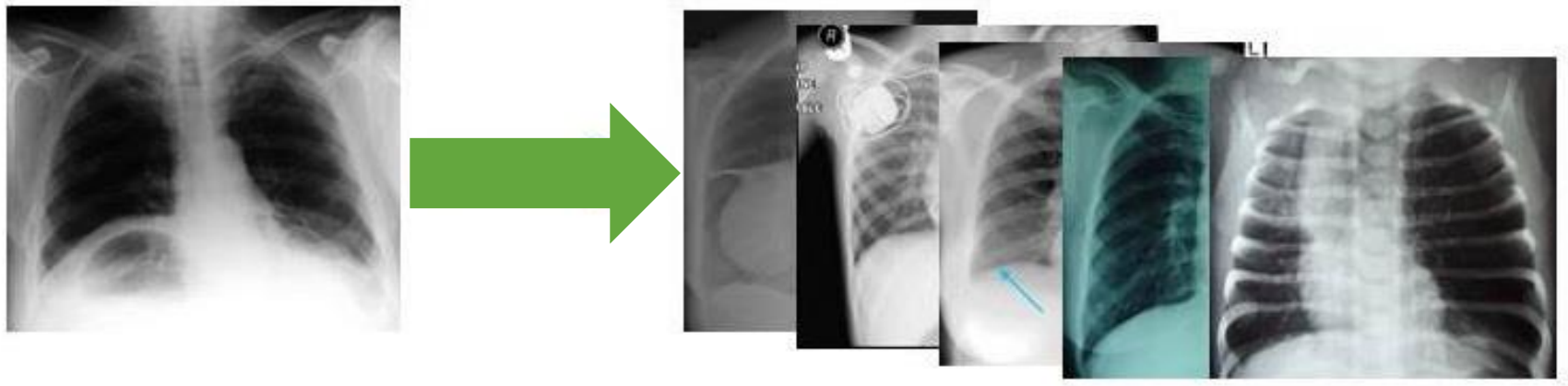
www.imageclef.org

- A long-running lab for evaluating multimodal medical systems
- 2025 tasks:
  - Automatic Image Captioning
  - VQA with synthetic gastrointestinal (GI) data
  - QA for Multimodal And Generative TelemedICine
  - Controlling the Quality of Synthetic Medical Images created via GANs

Are we asking the right question?

ImageCLEF medical task: Image retrieval

Goal: retrieve similar images where similarity is based on the relevance of the retrieved images



ImageCLEF medical task: Multimodal Image retrieval

- Textual queries: in English, Spanish, French and German
- 1-7 images per topic

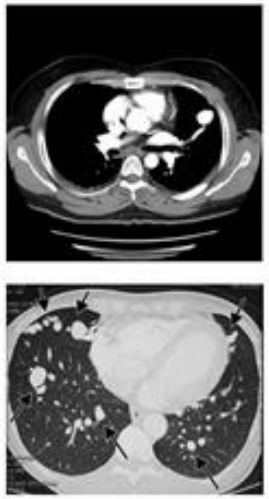



EN: osteoporosis X-ray images
ES: imágenes de rayos X de osteoporosis
FR: radiographie d'ostéoporose
DE: Röntgenbild Osteoporose

ImageCLEF medical task: Medical case-based retrieval

Goal: retrieve articles from the biomedical literature that are useful in differential diagnosis

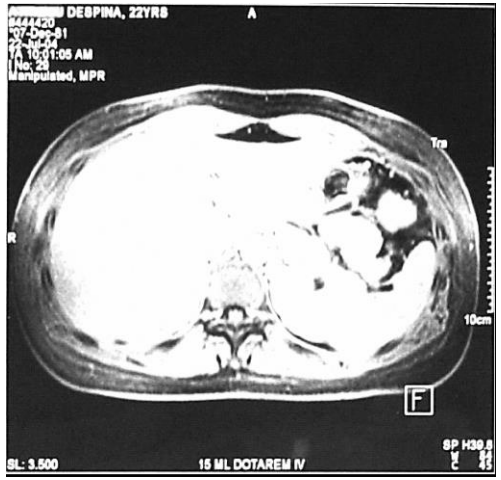
A woman in her mid-30s presented with dyspnea and hemoptysis. CT scan revealed a cystic mass in the right lower lobe. Before she received treatment, she developed right arm weakness and aphasia. She was treated, but four years later suffered another stroke. Follow-up CT scan showed multiple new cystic lesions.



Cystic nephroma: a case report and review of the literature.
2008 Case
ABSTRACT
cystic renal
RESULTS:
Cardiovascular magnetics resonance diagnosis of cystic tu
2009 J Cardiovasc Magn Reson
Late gadolinii
nature of can
Hydatid cyst disease of the lung as an unusual c
2009 J Med Case Reports
ABSTRACT: INTRODUCTION: Echinococcosis and/or hyd
however, hydatid disease of the lung is uncommon and us
presented with massive hemoptysis due to hydatid disease
Article in
Authors:
http://www.
Article in
Authors:
http://www.jc
Article in PDF-Version View all images / Visual searc
Authors: Tekinbas, C. Turedi, S. Gunduz, A. Erol, M. M.
http://www.jmedicalcasereports.com

ImageCLEF medical task: caption analysis

Goal: image understanding by aligning visual content and textual descriptors to interpret medical



Magnetic resonance imaging. After intravenous injection of gadolinium, the mass showed a progressive, heterogeneous, and delayed enhancement.

- C0016911: Gadolinium
- C0021485: Injection of therapeutic agent
- C0024485: Magnetic Resonance Imaging
- C0577559: Mass of body structure
- C1533685: Injection procedure

Information access in social media

From Detecting Harmful Content to Predicting What We Remember

EXIST: sEXism Identification in Social neTworks lab

https://nlp.uned.es/exist2025/

Goal: automatic detection of sexism

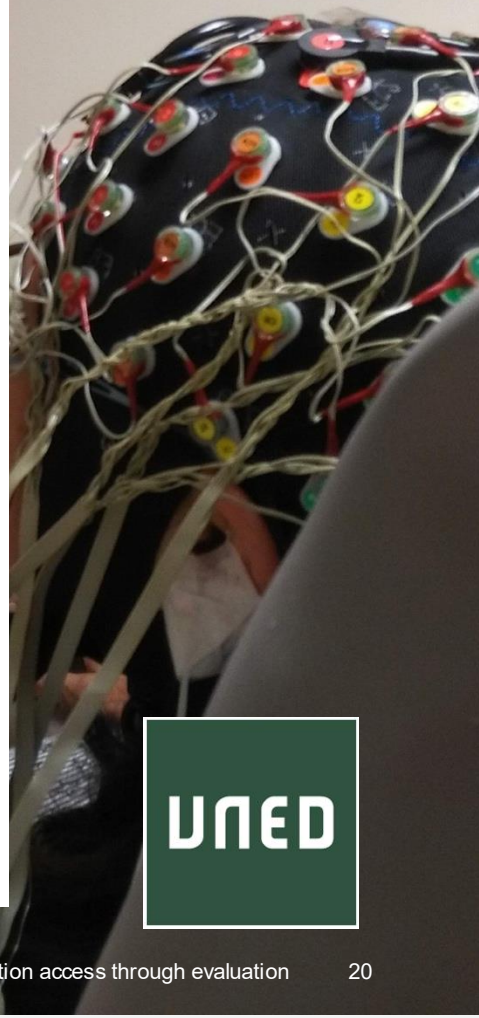
2025 tasks

	Subtask	Categories	Content	Source
Task 1: Tweets	Subtask 1.1: Sexism Identification	*YES* *NO*	Textual	Twitter
	Subtask 1.2: Source Intention	*DIRECT* *REPORTED* *JUDGEMENTAL*	Textual	Twitter
	Subtask 1.3: Sexism Categorization	*IDEOLOGICAL AND INEQUALITY* *STEREOTYPING AND DOMINANCE* *OBJECTIFICATION* *SEXUAL VIOLENCE* *MISOGYNY AND NON-SEXUAL VIOLENCE*	Textual	Twitter
Task 2: Memes	Subtask 2.1: Sexism Identification	*YES* *NO*	Textual (OCR) Image	Google Search
	Subtask 2.2: Source Intention	*DIRECT* *JUDGEMENTAL*	Textual (OCR) Image	Google Search
	Subtask 2.3: Sexism Categorization	*IDEOLOGICAL AND INEQUALITY* *STEREOTYPING AND DOMINANCE* *OBJECTIFICATION* *SEXUAL VIOLENCE* *MISOGYNY AND NON-SEXUAL VIOLENCE*	Textual (OCR) Image	Google Search
Task 3: Videos	Subtask 3.1: Sexism Identification	*YES* *NO*	Textual (OCR) Video	Tiktok Videos
	Subtask 3.2: Source Intention	*DIRECT* *JUDGEMENTAL*	Textual (OCR) Video	Tiktok Videos
	Subtask 3.3: Sexism Categorization	*IDEOLOGICAL AND INEQUALITY* *STEREOTYPING AND DOMINANCE* *OBJECTIFICATION* *SEXUAL VIOLENCE* *MISOGYNY AND NON-SEXUAL VIOLENCE*	Textual (OCR) VIdeo	Tiktok Videos

https://multimediaeval.github.io/editions/2025/tasks/memorability/

Memorability task

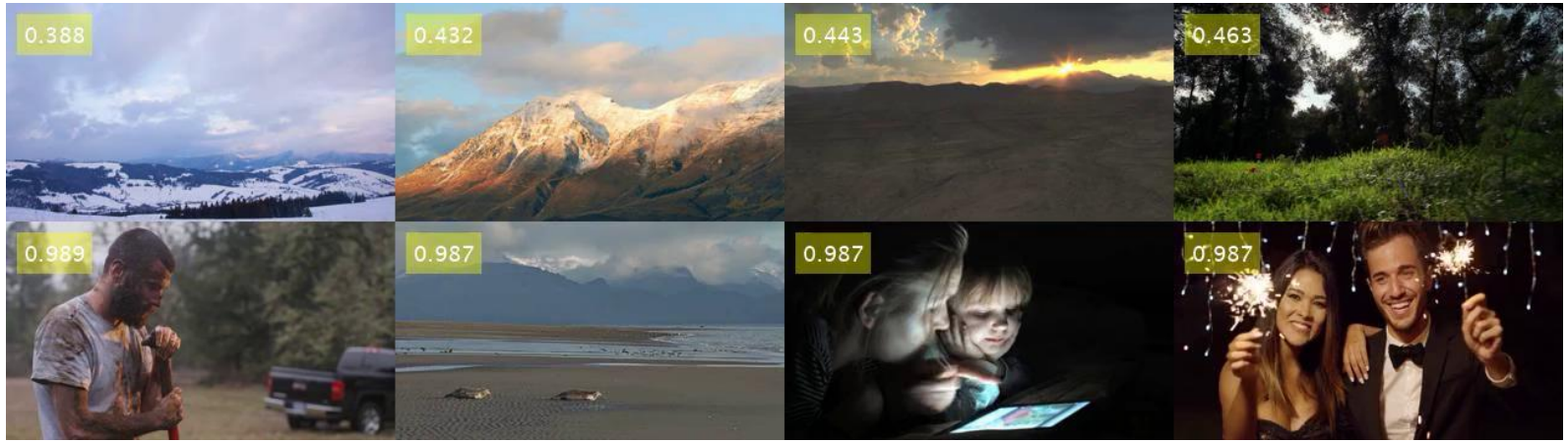
- 2025 challenges:
  - Is this person familiar with this video? - EEG-based detection of recall
  - *How memorable is this movie/commercial video?* - Video-based prediction
  - *Can you predict the brand memorability?* - Video-based prediction



Memorability task

https://multimediaeval.github.io/editions/2025/tasks/memorability/

Goal: Study memory performance when recognising videos videos



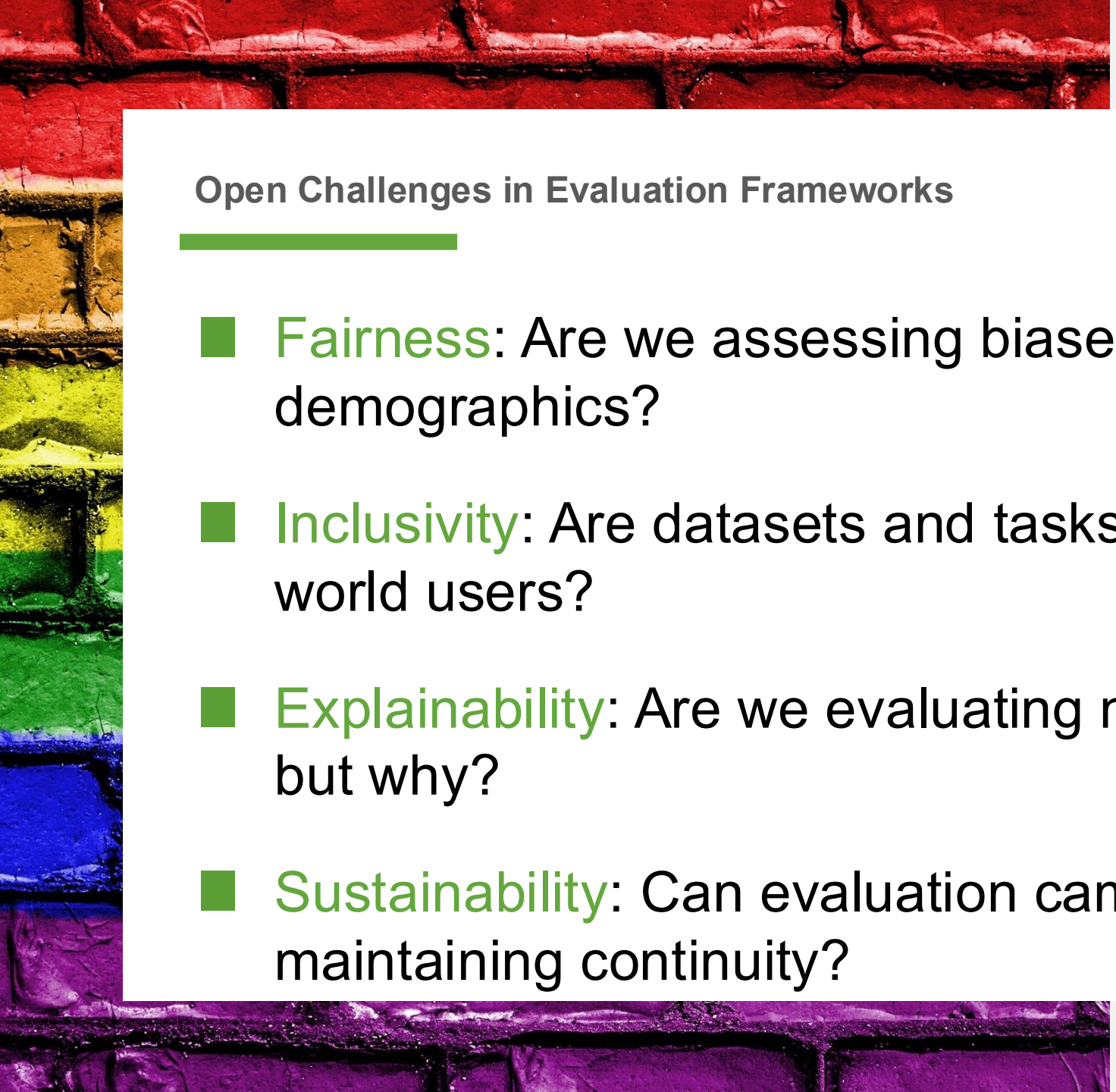
MediaEval (Multimedia Evaluation Benchmark) 2025

https://multimediaeval.github.io/editions/2025/

- 25–26 October, Dublin, Irleand
- Important dates:
  - Registration jut opened!
  - Evaluation Cycle Ends: 24 September
  - Working Notes Submission: 8 October

Rethinking Evaluation for Ethical and Inclusive Information Access

Fairness, Inclusivity, and Explainability in Multimodal Benchmarks



Open Challenges in Evaluation Frameworks

- Fairness: Are we assessing biases across modalities and demographics?
- Inclusivity: Are datasets and tasks diverse enough to reflect real-world users?
- Explainability: Are we evaluating not just what models predict, but why?
- Sustainability: Can evaluation campaigns evolve while maintaining continuity?

Shaping the Future of Ethical Information Access Evaluation

The way we evaluate systems shapes the systems we build

Key Takeaways & the Road Ahead

- Evaluation campaigns (e.g. CLEF) are essential to fostering progress and collaboration
- Realistic tasks and datasets enable meaningful, application-driven assessment
- Community-driven initiatives ensure relevance, reproducibility, and sustainability

Let's design evaluation frameworks that are not only technically robust,but also fair, inclusive, and aligned with human values

Thank you!

alba.garcia@lsi.uned.es