Face recognition and gallery management

Maurizio Montagnuolo
Rai – Radiotelevisione Italiana
Centre for Research, Technological Innovation and Experimentation (CRITS)

EBU Workshop on “Facial Recognition: benchmarking key technologies” – 21st June 2023

AI in media workflows

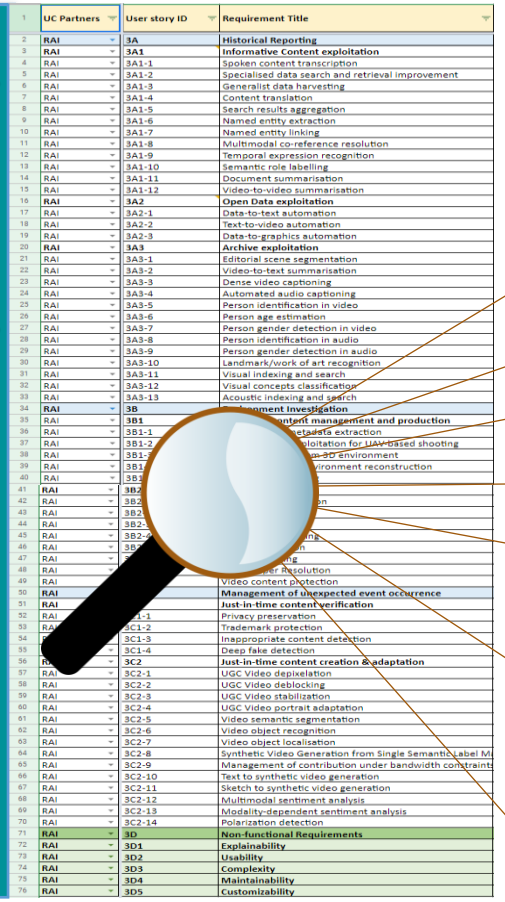
3 BUSINESS AREAS

- Content Access
- Content Production
- Content Monitoring

13 APPLICATION DOMAINS

- Language & Text Processing
- Content Summarisation
- Content Annotation
- Person Annotation
- Content Indexing & Search
- Computer-assisted Media Production
- Audio/Video Quality Enhancement
- UGC Adaptation for Broadcast
- Synthetic Video Generation
- UAV Content Production
- 3D Virtual Content Production
- Content Verification
- Social Media Analysis

59 AI FUNCTIONALITIES (user stories)

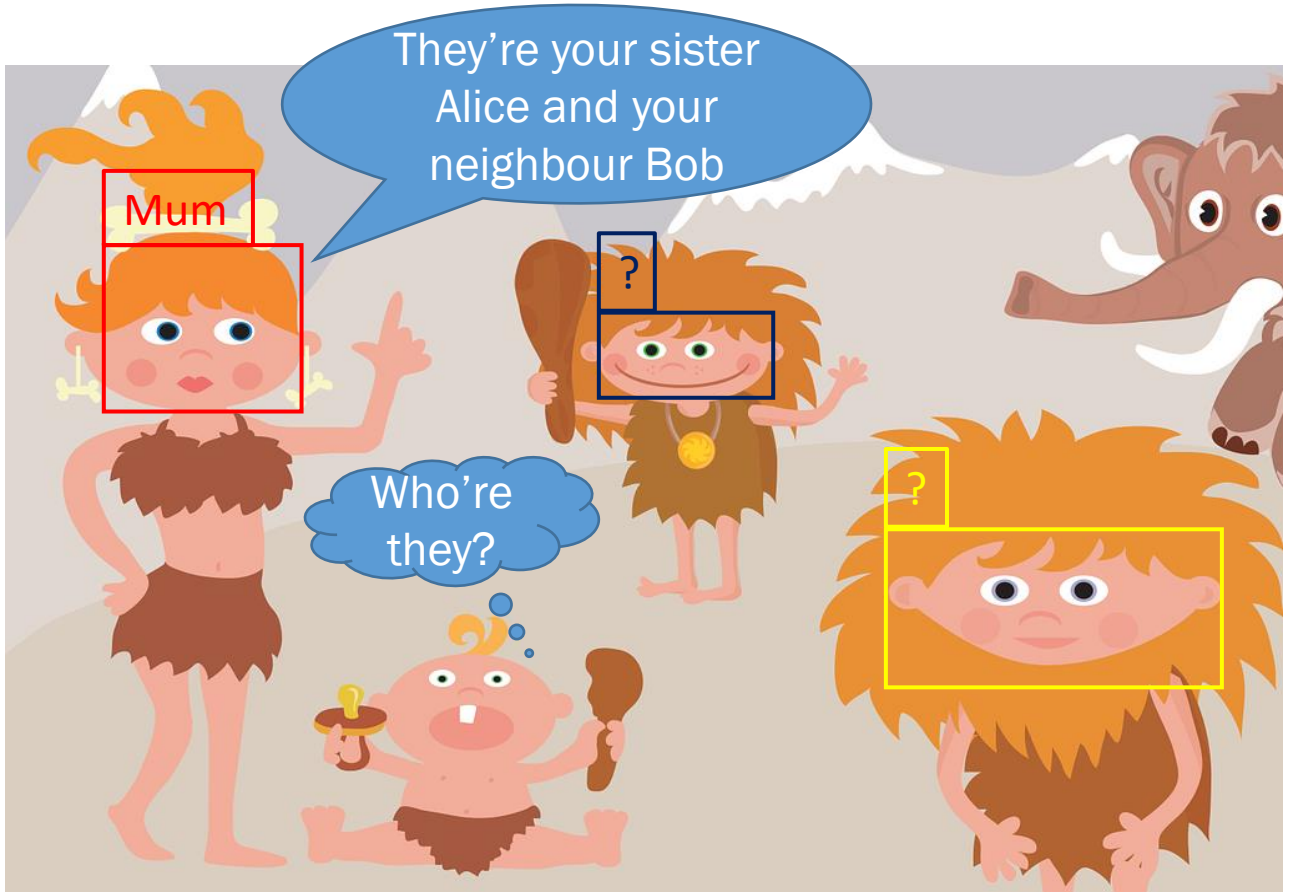


UC Partners	User story ID	Requirement Title
RAI	3A	**Historical Reporting**
RAI	3A1	**Informative Content exploitation**
RAI	3A1-1	Spoken content transcription
RAI	3A1-2	Specialised data search and retrieval improvement
RAI	3A1-3	Generalist data harvesting
RAI	3A1-4	Content translation
RAI	3A1-5	Search results aggregation
RAI	3A1-6	Named entity extraction
RAI	3A1-7	Named entity linking
RAI	3A1-8	Multimodal co-reference resolution
RAI	3A1-9	Temporal expression recognition
RAI	3A1-10	Semantic role labelling
RAI	3A1-11	Document summarisation
RAI	3A1-12	Video-to-video summarisation
RAI	3A2	**Open Data exploitation**
RAI	3A2-1	Data-to-text automation
RAI	3A2-2	Text-to-video automation
RAI	3A2-3	Data-to-graphics automation
RAI	3A3	**Archive exploitation**
RAI	3A3-1	Editorial scene segmentation
RAI	3A3-2	Video-to-text summarisation
RAI	3A3-3	Dense video captioning
RAI	3A3-4	Automated audio captioning
RAI	3A3-5	Person identification in video
RAI	3A3-6	Person age estimation
RAI	3A3-7	Person gender detection in video
RAI	3A3-8	Person identification in audio
RAI	3A3-9	Person gender detection in audio
RAI	3A3-10	Landmark/work of art recognition
RAI	3A3-11	Visual indexing and search
RAI	3A3-12	Visual concepts classification
RAI	3A3-13	Acoustic indexing and search
RAI	3B	**Enhancement Investigation**
RAI	3B1	**Content management and production**
RAI	3B1-1	Metadata extraction
RAI	3B1-2	Exploitation for UAV-based shooting
RAI	3B1-3	3D environment
RAI	3B1-4	Environment reconstruction
RAI	3B2	
RAI	3B2-1	
RAI	3B2-2	
RAI	3B2-3	
RAI	3B2-4	
RAI	3B2-5	
RAI	3B2-6	Super resolution
RAI	3B2-7	Video content protection
RAI	3C	**Management of unexpected event occurrence**
RAI	3C1	**Just-in-time content verification**
RAI	3C1-1	Privacy preservation
RAI	3C1-2	Trademark protection
RAI	3C1-3	Inappropriate content detection
RAI	3C1-4	Deep fake detection
RAI	3C2	**Just-in-time content creation & adaptation**
RAI	3C2-1	UGC Video depixelation
RAI	3C2-2	UGC Video deblocking
RAI	3C2-3	UGC Video stabilization
RAI	3C2-4	UGC Video portrait adaptation
RAI	3C2-5	Video semantic segmentation
RAI	3C2-6	Video object recognition
RAI	3C2-7	Video object localisation
RAI	3C2-8	Synthetic Video Generation from Single Semantic Label M
RAI	3C2-9	Management of contribution under bandwidth constraints
RAI	3C2-10	Text to synthetic video generation
RAI	3C2-11	Sketch to synthetic video generation
RAI	3C2-12	Multimodal sentiment analysis
RAI	3C2-13	Modality-dependent sentiment analysis
RAI	3C2-14	Polarization detection
RAI	3D	**Non-functional Requirements**
RAI	3D1	**Explainability**
RAI	3D2	**Usability**
RAI	3D3	**Complexity**
RAI	3D4	**Maintainability**
RAI	3D5	**Customizability**

- Landmark/work of art recognition
- Spoken content transcription
- Person identification in video
- Video super resolution
- Deep-fake detection
- Editorial scene segmentation
- Visual indexing and search

How do humans perceive faces?

- Infants give preferential attention to human faces; thus, they are able to discern them from other objects
- Without proper teaching, infants (and humans in general) can't associate faces to identities
- Infants recognise familiar faces and respond happily to them
- Adults teach infants to differentiate between familiar people and strangers



How do machines perceive faces?

- Face Detection:
  this is the process by which the **areas of an image containing faces** are located
- Face Representation:
  a face is described by a numerical vector (**embedding**) that **represents the features** of the face's keypoints
- Face Similarity:
  it is the process by which the **degree of likeness** between 2 faces is determined. Likeness is computed based on the face embedding and a similarity (distance) function
- Face Clustering:
  it is the process by which faces with **similar key point features** are grouped
- Face Verification:
  it computes **one to one similarity** to check if two faces are from the same individual
- Face Identification:
  it computes **one to many similarity** to determine the specific identity of an unknown face

Rai

The Face Management Framework

- It implements an end-to-end pipeline for face identification and content annotation based on few-shot or zero-shot face embedding extraction models

- The workflow of the FMF is a backend process made of two phases:
  (1) Management (create / read / update / delete) of a gallery of reference identities, i.e., people to be annotated;
  (2) detection, identification and grouping of the reference identities in unannotated video streams

- The reference gallery may be edited in an opportunistic (e.g., based on production or archival metadata) or strategic way (e.g., based on a priori list of celebrities)

Rai

Requirements

Extensibility

The Gallery must include links to knowledge databases that enrich the metadata of the enrolled identities

Scalability

The Gallery must not have any functional limitations on the number of enrolled identities, except for physical HW restrictions (e.g., disk or memory size)

Flexibility

The Gallery must provide basic logic to:

disambiguate homonyms

create custom profiles

update metadata

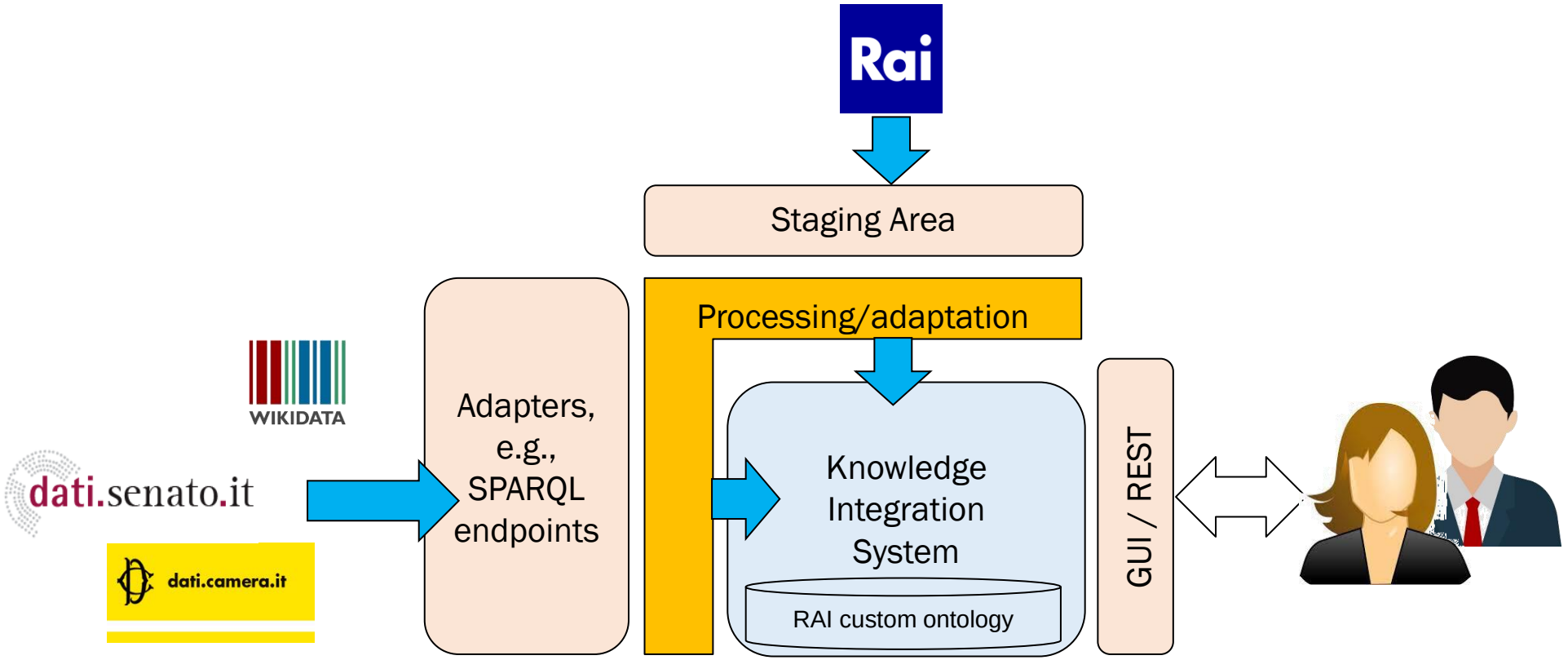
group profiles

of the enrolled identities

Variety

The Gallery must be as diversified as possible regarding attributes (e.g., age, head pose, mack-up) of the enrolled identities

High level architecture

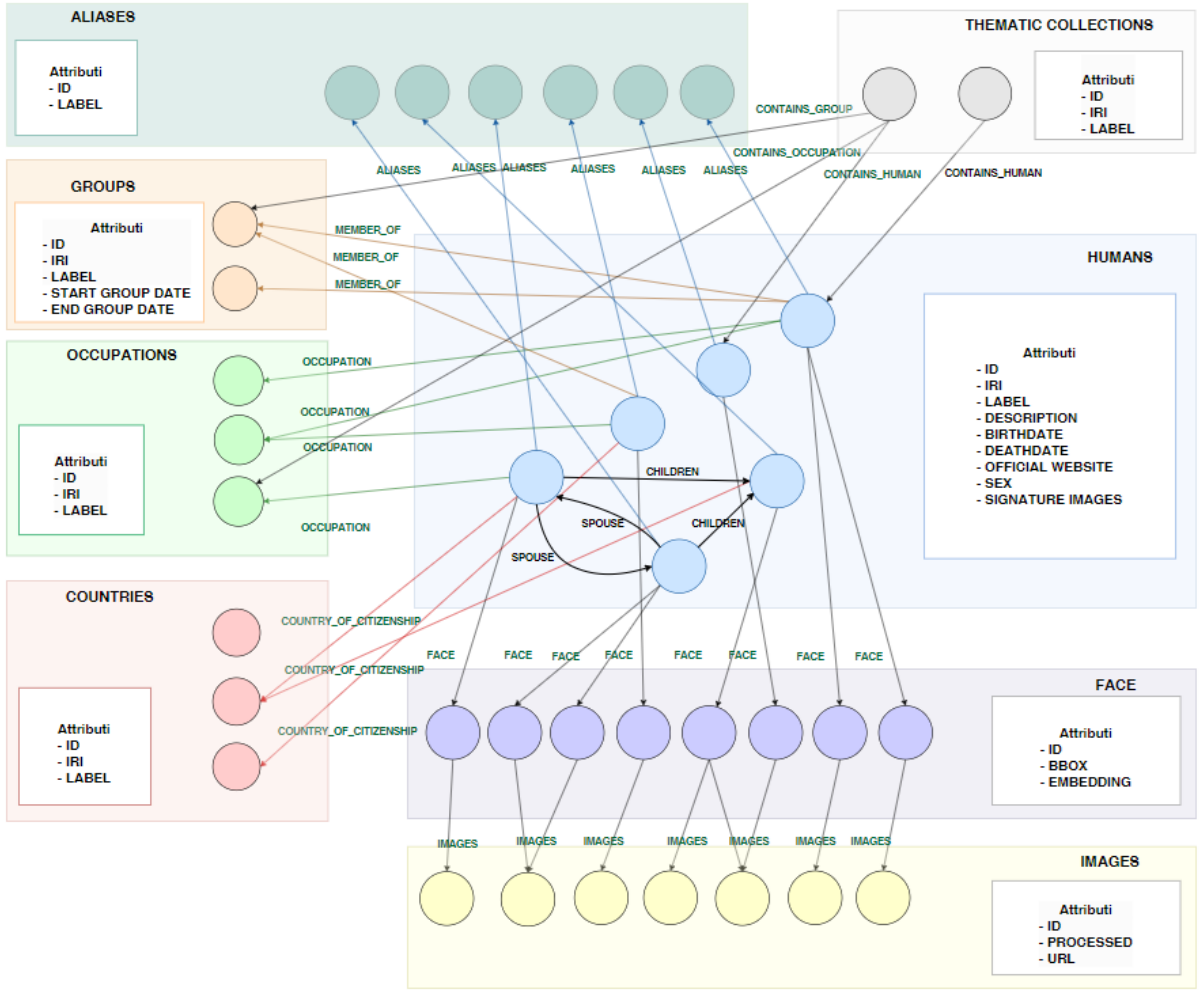


Conceptual model

PERSON:
A media personality who is notable for being either internationally (e.g., a music star), nationally (e.g., a TV programme presenter) or locally (e.g., a representative of a municipal body or institution) acknowledged

PERSON GROUP:
A collection of persons sharing some interests, memberships, or other criteria defined by the user (e.g., the members of a band)

THEMATIC COLLECTION:
A hierarchical set of person groups (e.g., Italian ESC participants)



Technical implementation

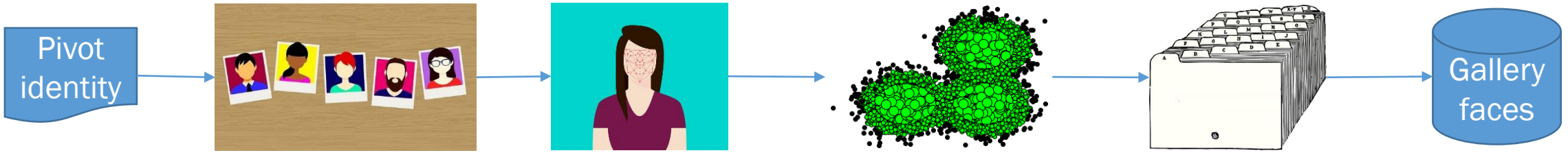


Image scraping (e.g., archives, Web,...)

Face detection, alignment and embedding extraction

Embedding clustering

Embedding indexing

Application example (1)

Planned Event

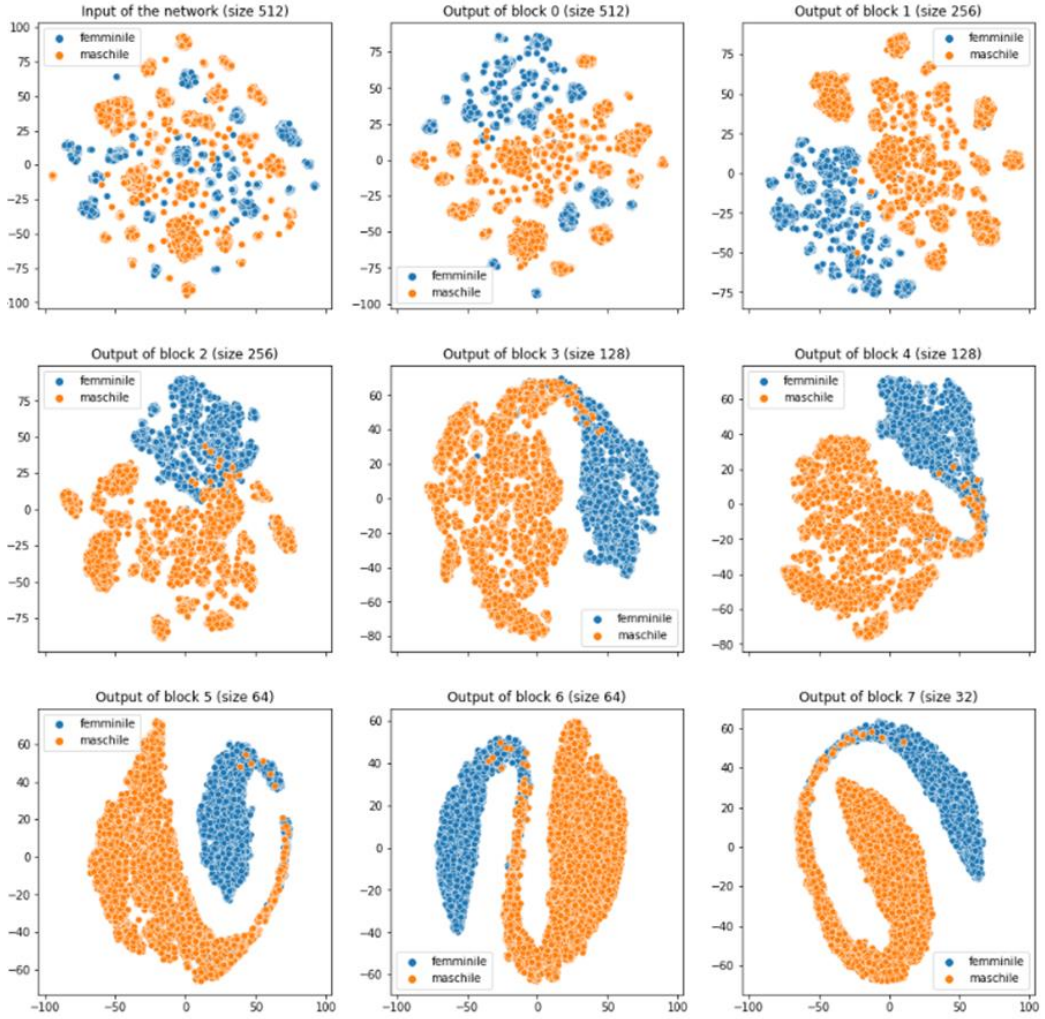
Production need:
Mashup for RaiPlay

Broadcasted show (live)

Face Management Framework

Application example (2)

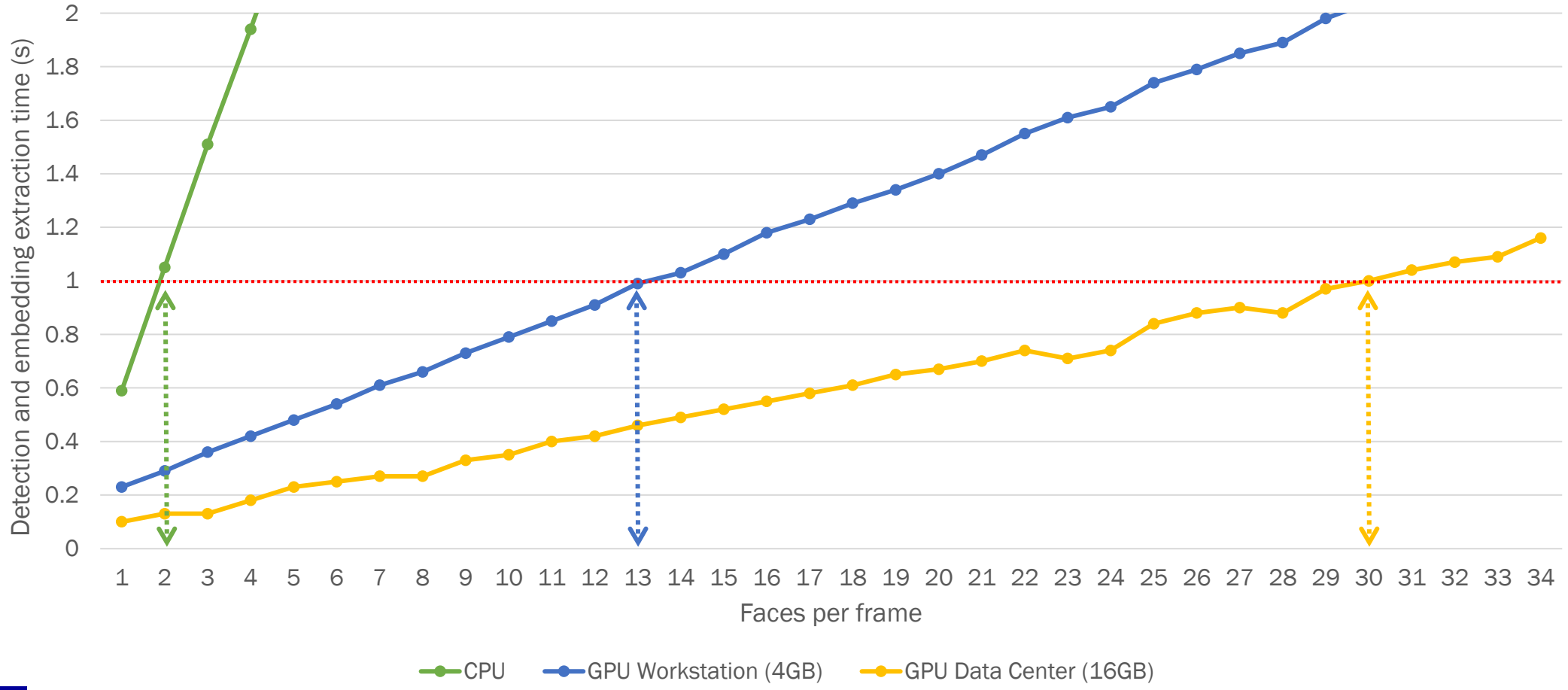
- What: analysis of facial related gender features
- Why: monitoring gender equality in TV coverage
- How: FMF face embeddings to feed an 8-block MLP network



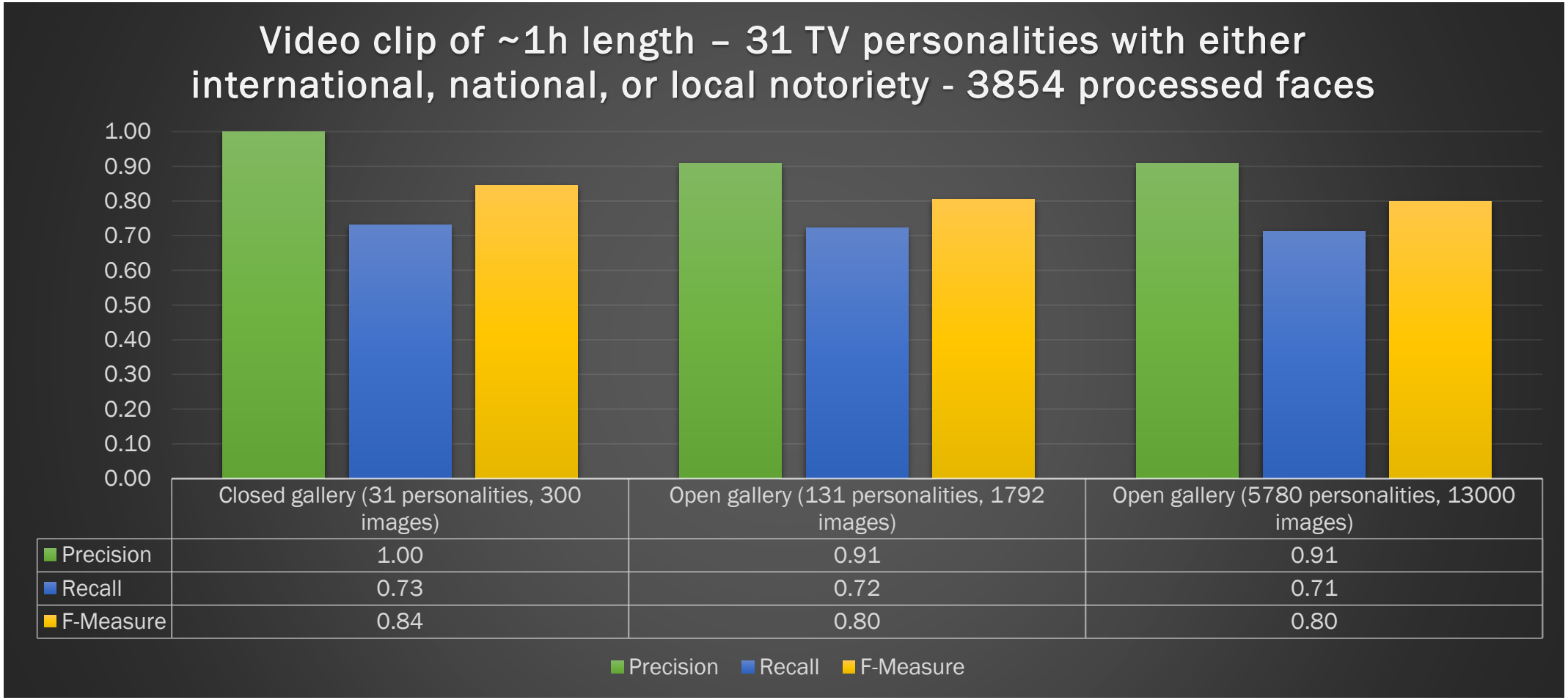
Demo

1. Gallery creation
2. Content annotation & monitoring

Benchmarking algorithms (1)



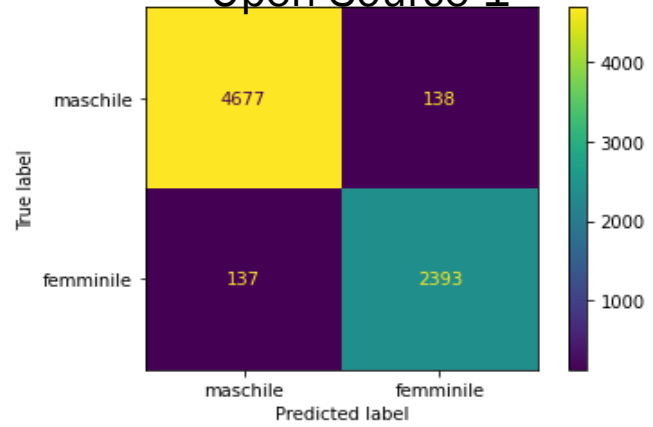
Benchmarking algorithms (2)



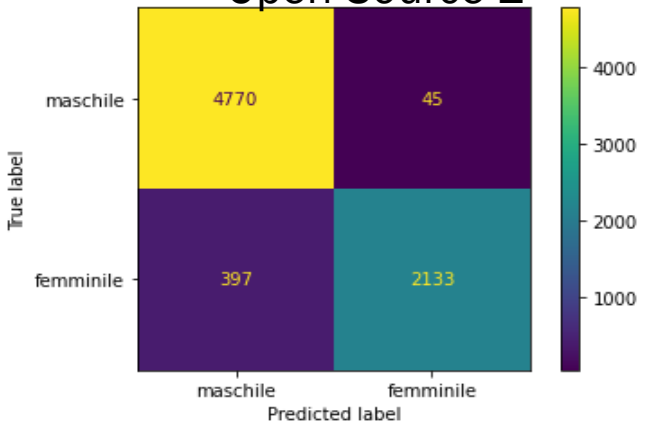
	Closed gallery (31 personalities, 300 images)	Open gallery (131 personalities, 1792 images)	Open gallery (5780 personalities, 13000 images)
Precision	1.00	0.91	0.91
Recall	0.73	0.72	0.71
F-Measure	0.84	0.80	0.80

Benchmarking algorithms (3)

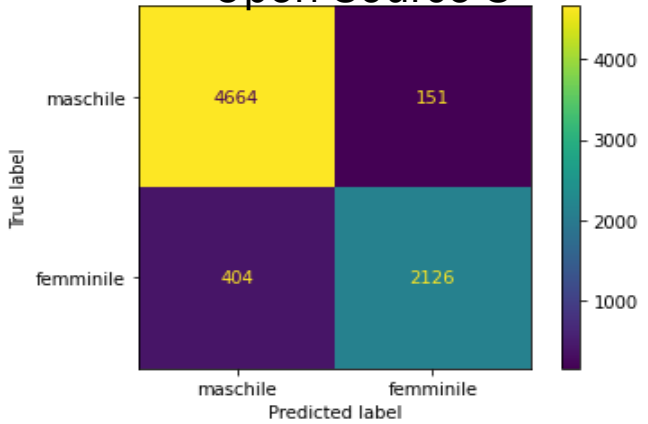
Open Source 1



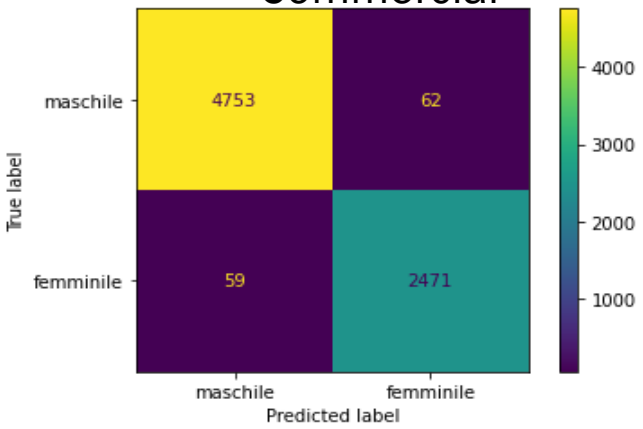
Open Source 2



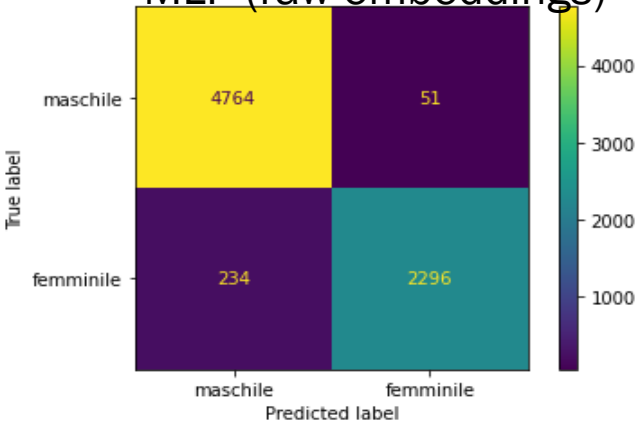
Open Source 3



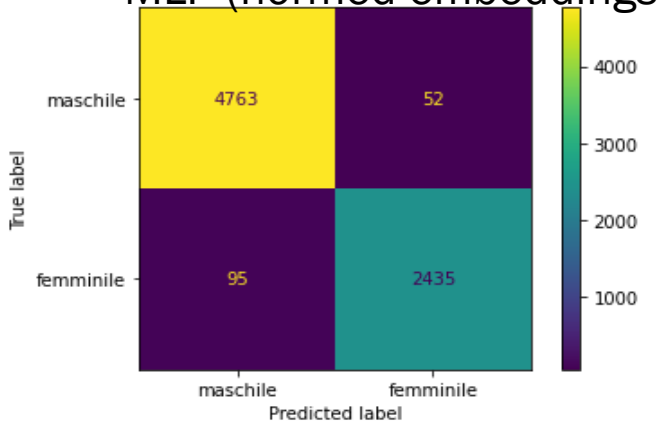
Commercial



MLP (raw embeddings)



MLP (normed embeddings)



Benchmarking processes & systems

Testing scenario: *Remote sessions*

- *Demo + interactive samples + brainstorming*
- *Questionnaires*

User data

To collect information about the professional background of the interviewees

Graphical User Interface

To ask feedback about several aspects of the GUI, such as intuitiveness and interactivity

Tool functionalities

To ask feedback about the usefulness or impact of each functionality on existing workflows following a 5-point rating scale (i.e., “strongly disagree, disagree, neutral, agree, strongly agree”)

Final section

- To collect suggestions for subsequent refinements, adaptations and improvements
- KPI evaluation (3 statements to measure results with respect to the provided KPI)

Main users' suggestions

GUI improvements

- Provide more options for filtering based on broadcast metadata, e.g., date, channel or programme's title
- Need for more information about technical metadata, e.g., video formats, resolution or size
- Display further statistics about a person/face, e.g., the percentage of time appearance Vs the video length or the face size Vs the frame size
- Ability to show the keyframes of more than one person at a time

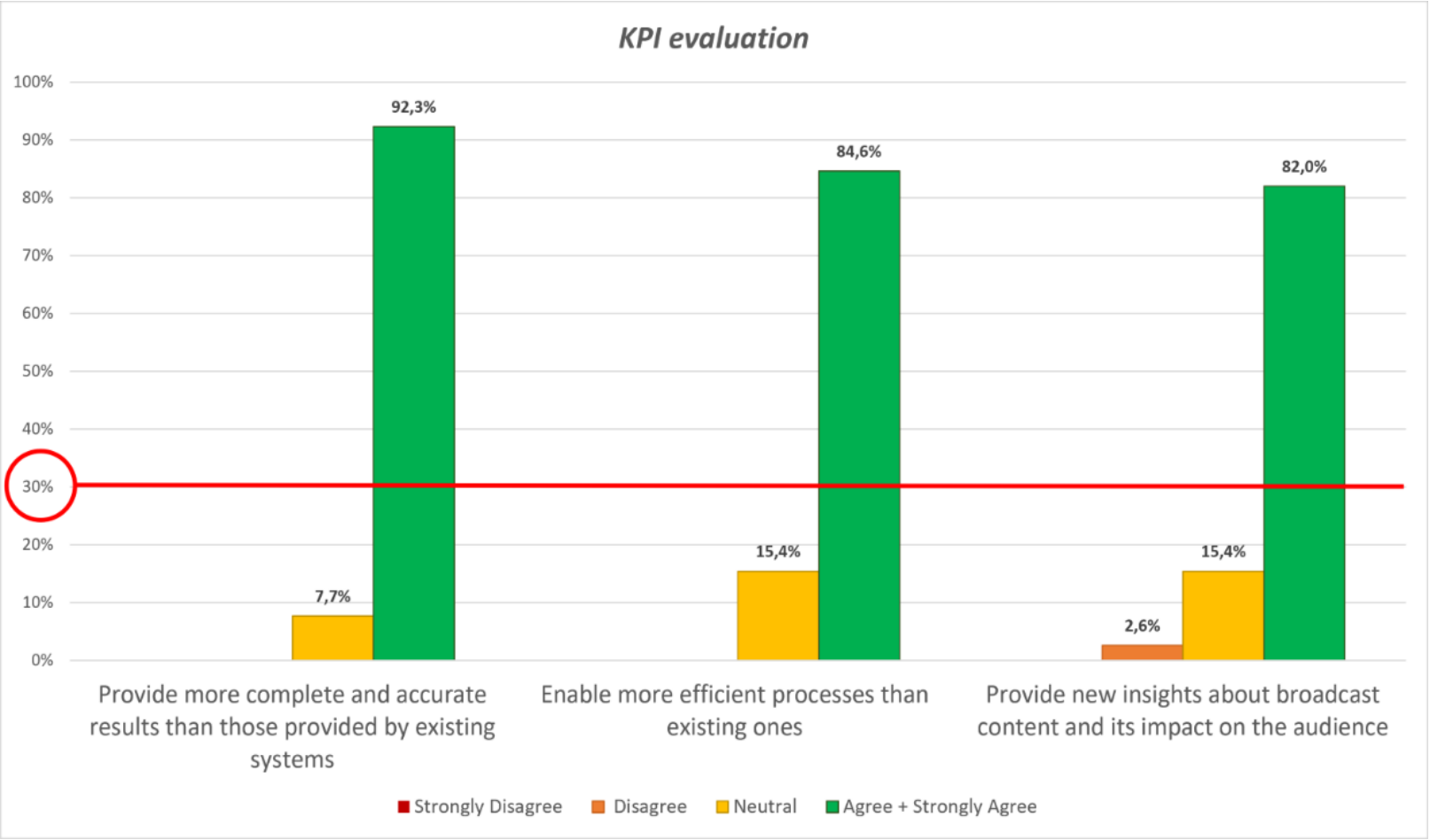
New functionalities

- Recognise speaking people
- Identify the type of camera shot for the detected face, e.g., close-up, very close-up, etc.
- Search by facial-related attributes, e.g., hairs colour, wearing hats or glasses, moustache, etc.
- Search by people and context, e.g., a certain person on the street

The most popular

- Need for **cross-referencing face annotations with further information**, such as speech transcriptions

KPI evaluation



KPI definition:

Enabling more efficient methodology for detection and identification of people within the broadcaster's production processes

Conclusions

- AI technologies for face management (i.e., detection, verification, identification, clustering) have the potential to provide significant support to the media value chain
- On the application side, there are several possible use cases where these technologies could be of benefit
- In some cases, individual components address only a relatively small part of a much bigger problem, as for example the gender analysis is only an aspect of a wider process monitoring the representation of female characters in RAI TV programmes
- Flexible, variated and informative Galleries are fundamental to success
- In the Gallery creation process, different strategies may be applied, according to the tasks to be executed
  - Closed Gallery in which the list of reference TV personalities is circumscribed, e.g., a talk show
  - Open Gallery in which the list of reference TV personalities is undefined, e.g., unsupervised annotation

Rai

Thank you!

maurizio.montagnuolo@rai.it

White paper:
"AI in Vision: high quality video production and content automation"
This work is supported by the UE Horizon 2020 (Grant Agreement 951911 – AI4Media)

Rai