

# Training AI on digitised portraits

**dr. sc. Hrvoje Stančić, full. prof.**

Faculty of Humanities and Social Sciences

University of Zagreb, Croatia

[hstancic@ffzg.unizg.hr](mailto:hstancic@ffzg.unizg.hr)

InterPARES Trust AI Symposium  
Honolulu, Hawaii, USA  
23 February 2024

# Contents

1. Introduction
2. The study
3. The Team
4. Research results
5. Description of the training set
6. Workflow
7. Future work
8. Conclusion
9. Dissemination

# Introduction

- Archival institutions
  - information society challenges
- Emerging technologies
  - change information landscape
  - new user habits and expectations
  - redesign of the relationships between users and institutions
  - traditional practices of archiving are being transformed



# Introduction

- Disruptive technologies
  - artificial intelligence
  - blockchain
  - big data
  - crowdsourcing
  - gamification etc.
- Positive disruption of current archival processes (service improvement)



# The study

- Identification of critical archival challenges which are the best candidates for improvement by AI technologies
  - particular interest in the context of retention and preservation of digital records



# The Team

- **Hrvoje Stančić, lead**, Arian Rajh + GRAs: Željko Trbušić, Vladimir Bralić, Patricija Gligora, Vita Jozić, Josipa Sumpor, Faculty of Humanities and Social Sciences (FHSS), University of Zagreb, **Croatia**
- Alicia Barnard, Universidad Nacional Autónoma de **México** - ENES-Morelia
- Gabriele Bezzi, Regione Emilia-Romagna, **Italy**
- Meltem Dişli, Hacettepe University, **Turkey**
- Pat Franks, San Jose State University - School of Information, **USA**
- Arien Gonzales Crespo, El Colegio de **México**
- Claudia Lacombe Rocha, National Archives of **Brazil**
- Lungile Luthuli-Ngidi, University of **South Africa**
- Patricia (Pat) Moore, Carleton University, **Canada**
- Samir Musa, European University Institute - Historical Archives of the European Union, **Italy**
- Rosely Rondinelli, Institute of Technology and Society, **Brazil**



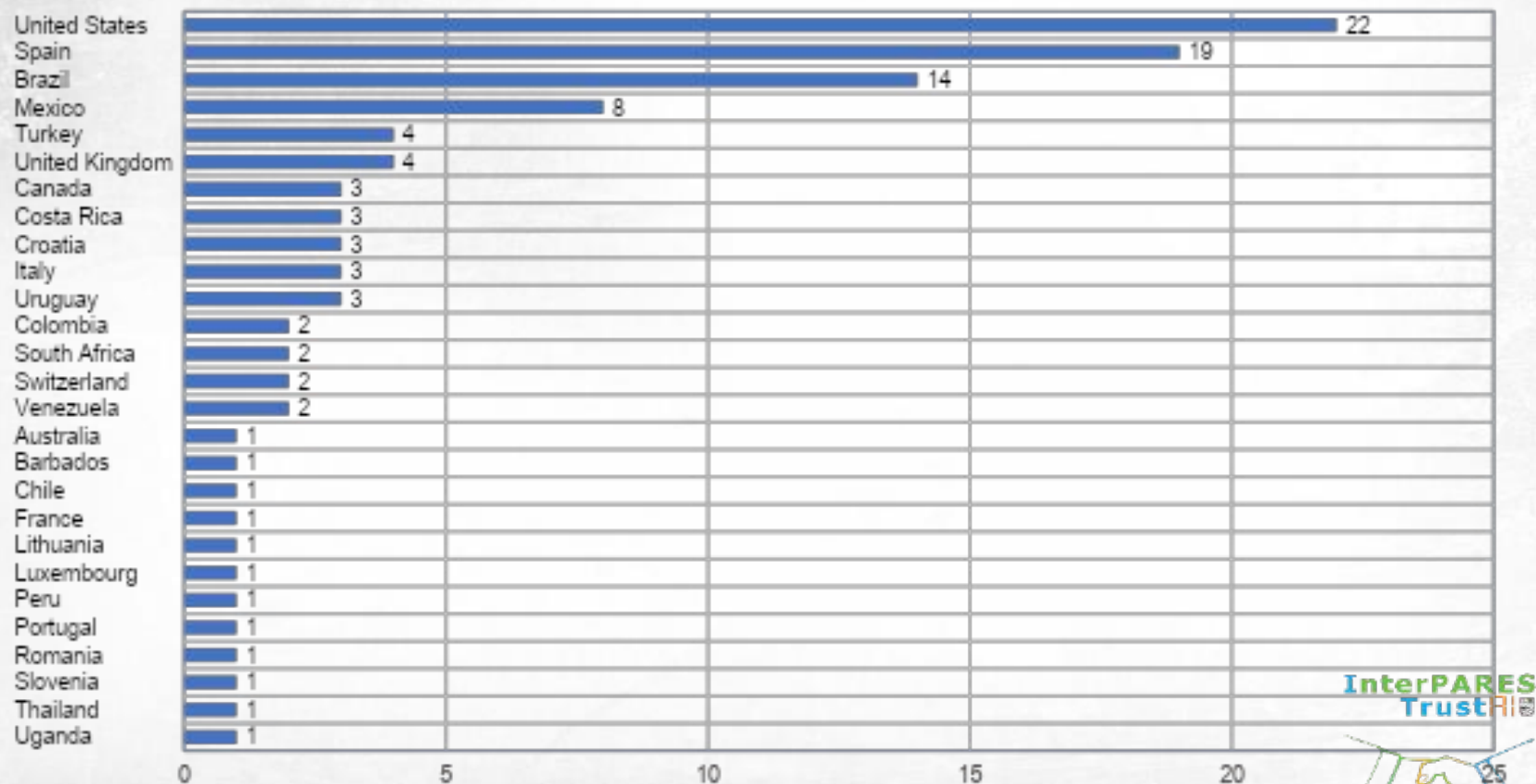
# Research results

- Phase 1 methodology
- Online survey
  - targeted archival practitioners and experts in the field
- Follow up in-person interviews



# 106 respondents from 27 countries

In which country is the institution/organization at which you work located  
(n=106)?

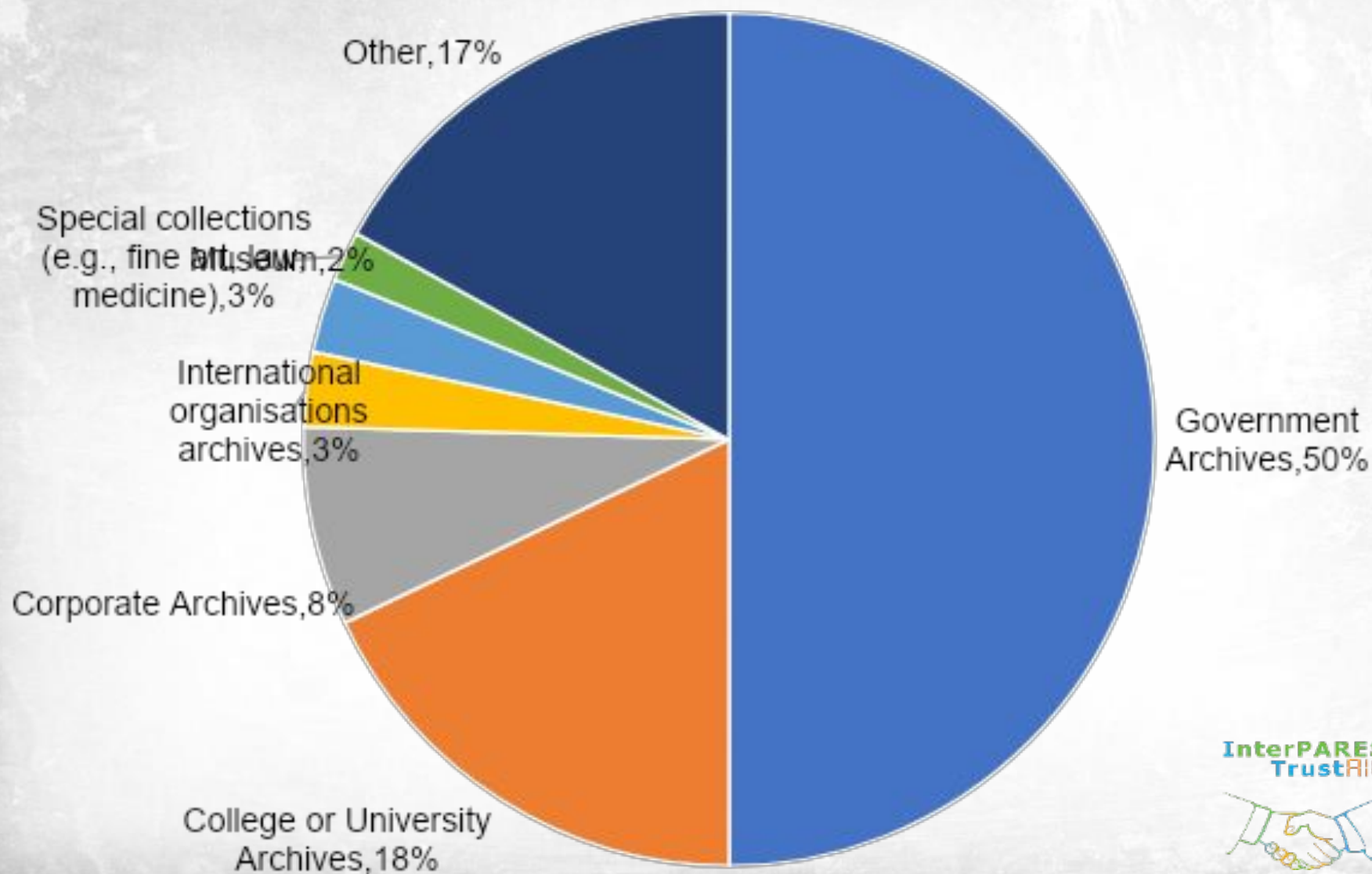


InterPARES  
TrustRI





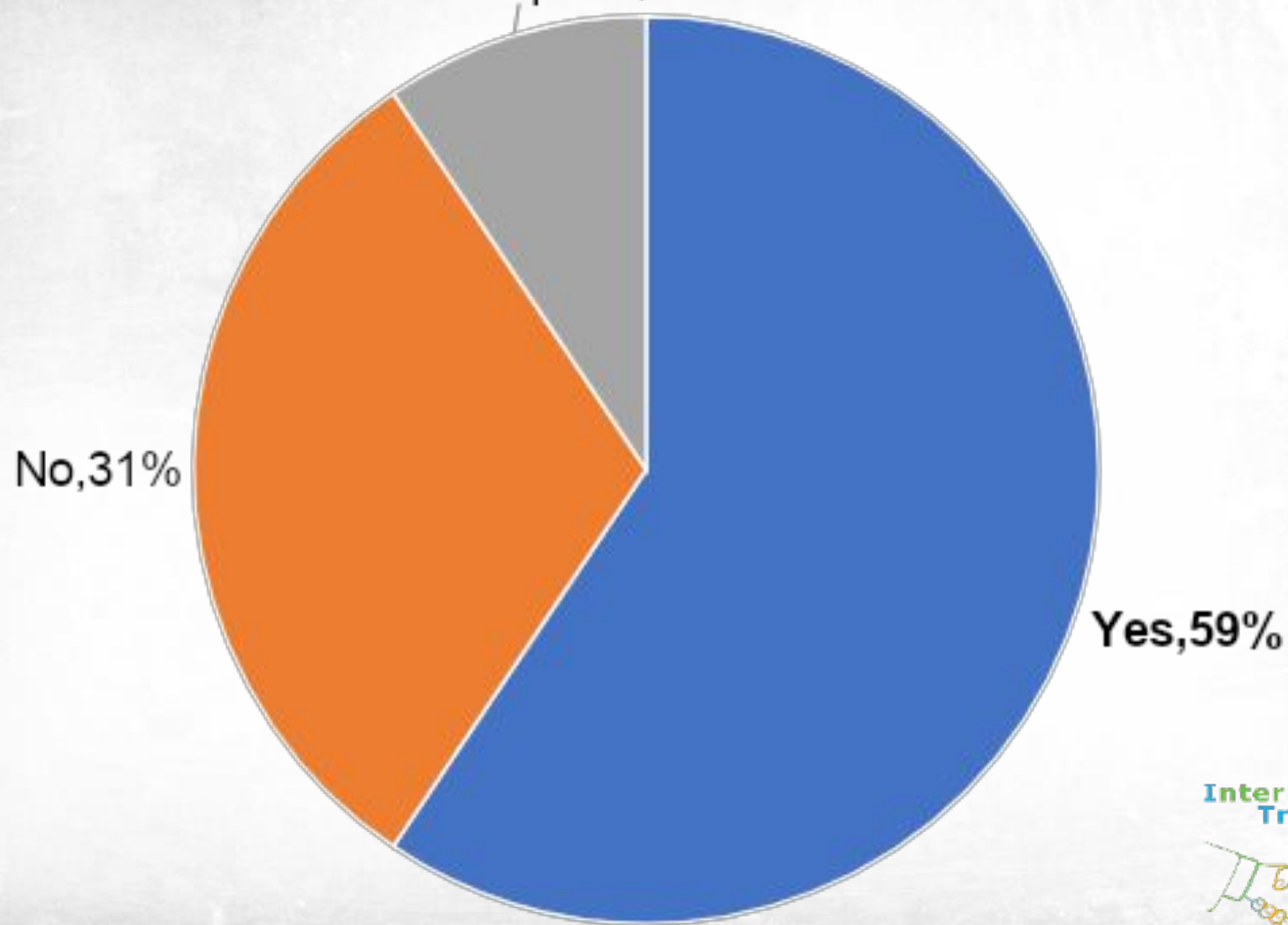
In which type of institution/organization do you work (n=106)?



InterPARES  
TrustAll

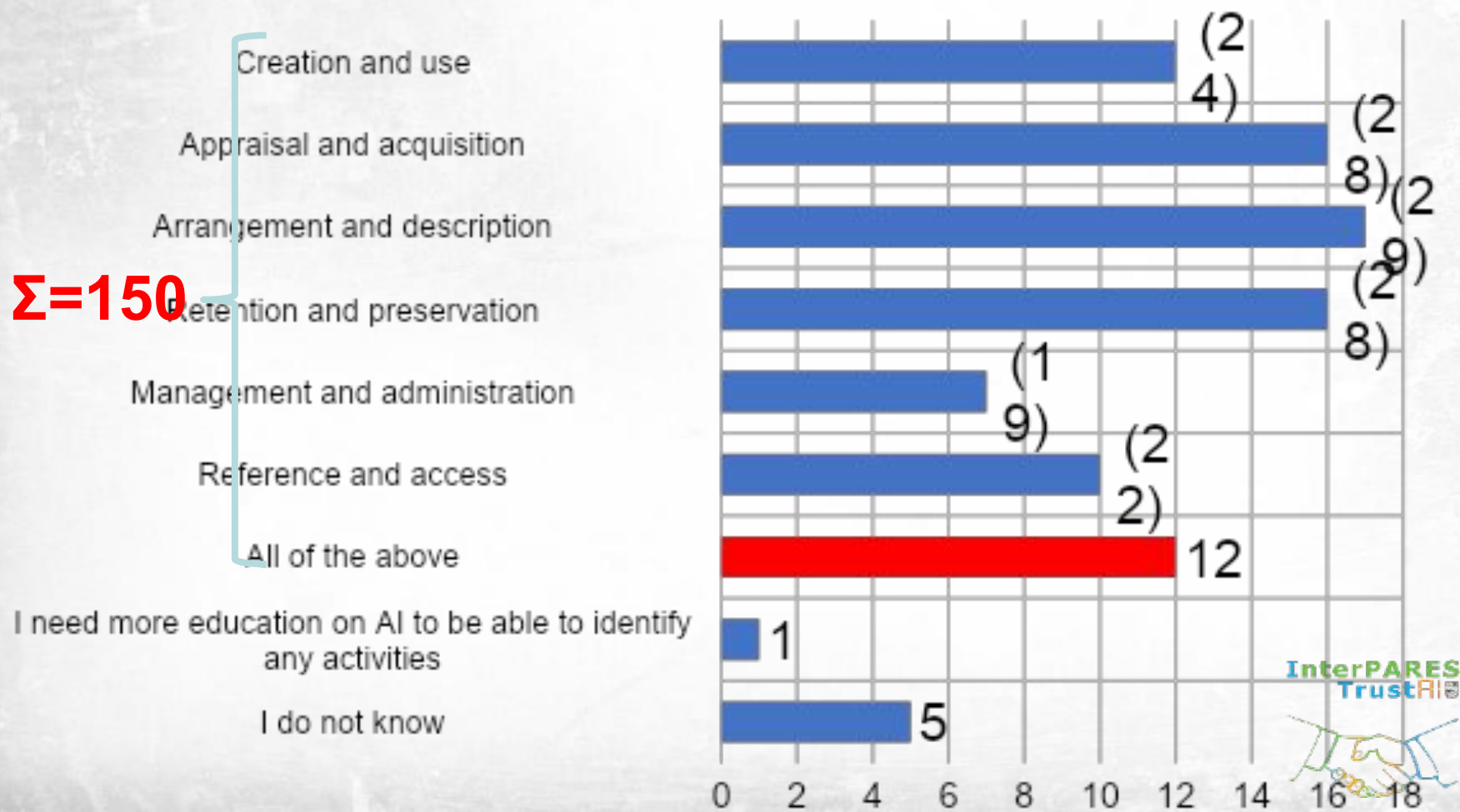


Do you have any particular processes which can be (additionally) improved by AI-related technologies (n=106)?

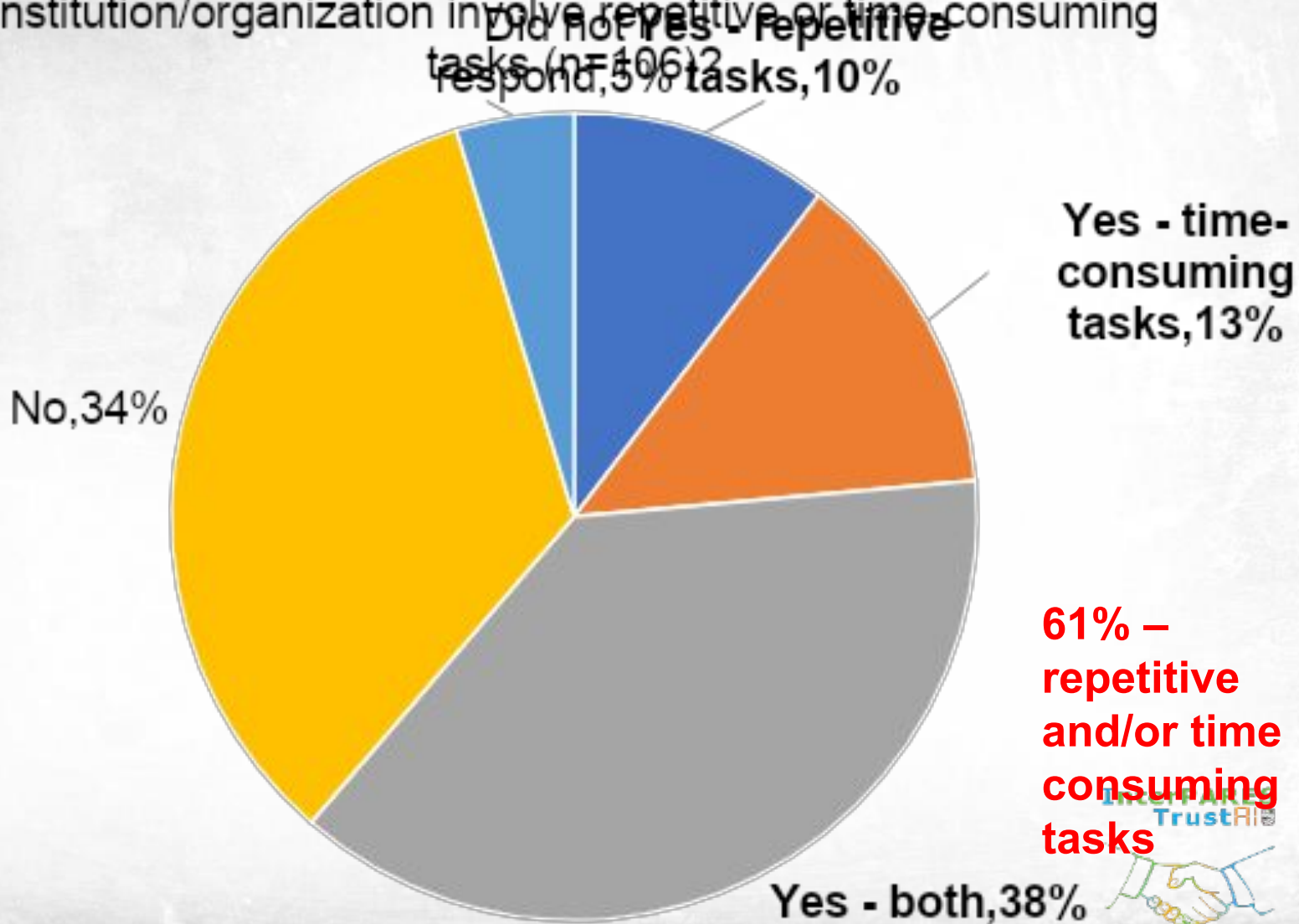


To which group of activities the identified processes which can be improved by AI-related technologies best relate to (n=63)?

$\Sigma=150$



Do any of the digital preservation processes in your institution/organization involve repetitive or time-consuming tasks (n=106)?



# Identified repetitive and/or time-consuming tasks (30 in total, showing 3+)



Adding, gathering, extracting metadata	11
Digitization	10
Capture / ingest	7
File integrity check	6
Indexing	5
Records management	5
Appraisal	4
Backup	3
Renaming files (based on their content)	3





# In-person interviews analysis – selection of responses

- *Will you please explain how you think AI might help you to solve your records and archival issues for retention and preservation.*

Various answers, mostly not related to retention and preservation (RP), i.e., **it can help with** transcription, acquisition, **description**, classification etc.

**RP** – AI **can help with** file format analysis and monitoring of obsolescence and integrity of preservation.



# In-person interviews analysis – selection of responses

- *Are you afraid of losing your job because of AI?*

No, no, no, ..., and – hell no!

... because AI needs human supervision

... because my position will grow and advance along with the technology

... because we are the ones who train the machine

... people who work description and transcription may be endangered

InterPARES  
TrustAI





# Description of the training set

- Cooperation with the State Archives in Osijek, Croatia
  - AI training set
    - collection of (set – studio and outdoor) portraits
      - total: 1,417 images (recto: 708 images)
      - from 1870s to the beginning of the 20<sup>th</sup> century
    - 9 categories
      - male, female, children, group – studio + outdoors
      - school groups + groups of members of associations
    - description of the set





# Description of the training set

- Description of the set – example 1.11.
  - Name: Marija Sudić, married Tadijanović
  - Description: young female person shot in studio (**dress, jewelry, flower, pillars**)
  - Photographer: Georg Knittel
  - Atelier: Georg Knittel, Osijek
  - Date: 1865. – 1907.
  - Place: Osijek
  - Format: 14.50 x 10.50 cm
  - Medium: paper
  - Technique: black and white
  - Polarity: positive



# Workflow

1. Dataset preparation
2. Annotation
3. Training
4. Results and testing



# Workflow – dataset preparation

- Implicit data extraction
  - male / female
  - age: child, young, adult, old
  - place: studio, outdoor
- Explicit data extraction
  - extract all labels
  - create a set of unique labels

C	
OPIS FOTOGRAFIJE	
Mlada muška osoba snimljena u studiju (odijelo, kravata, ograda).	
Mlada muška osoba snimljena u studiju (vojna odora, brkovi).	
Muška osoba srednjih godina snimljena u studiju (vojna odora, brkovi, šešir, pero, mač, stolac, stolić).	
Starija muška osoba snimljena u studiju (leptir mašna, odijelo, brkovi).	
Muška osoba srednjih godina snimljena u studiju (vojna odora, brkovi).	

**Male portraits:** "suit", "military uniform", "hunting uniform", "priest dress", "mustache", "coat", "fur", "beard", "glasses", "hat", "cap", "tie", "butterfly bow", "bow", "scarf", "decoration", "decorations", "sword", "staff", "jewelry", "pipe", "watch", "gloves", "chair", "table", "small table", "fence", "cabinet", "mirror", "stand", "haircut and shaving accessories", "pen", "cigarette", "flower", "flowers", "vase", "umbrella", "books", "box", "bag", "chair back", "armchair", "deer statue", "cross", "number"



# Workflow – annotation

- Label selection
  - full scope: > 100 labels
  - pilot training: 5 labels

0 uniform

1 suit

2 dress

3 hat

4 flowers



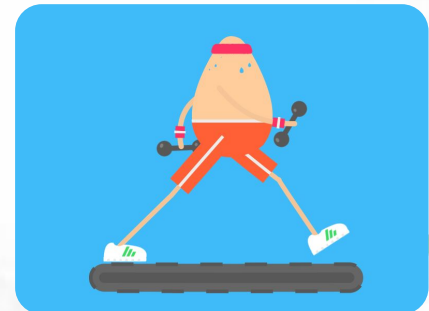
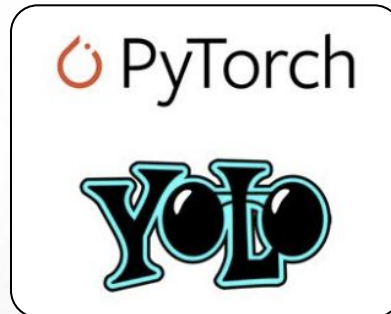




# Workflow – training

## Training environment

- YOLOv5 (computer vision) and PyTorch (machine-learning) utilities
- Google Colab + local environment



PARES  
ustAI

# Workflow – results and testing

## Training



## Validation



# Workflow – results and testing

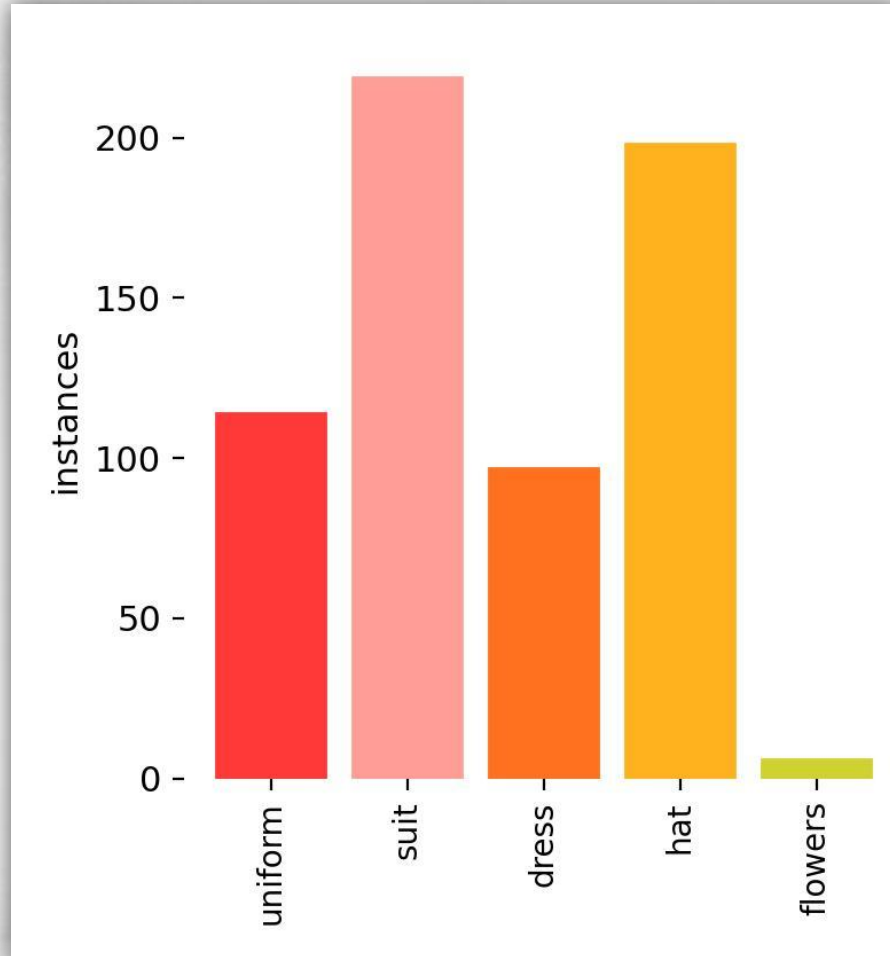
## Training



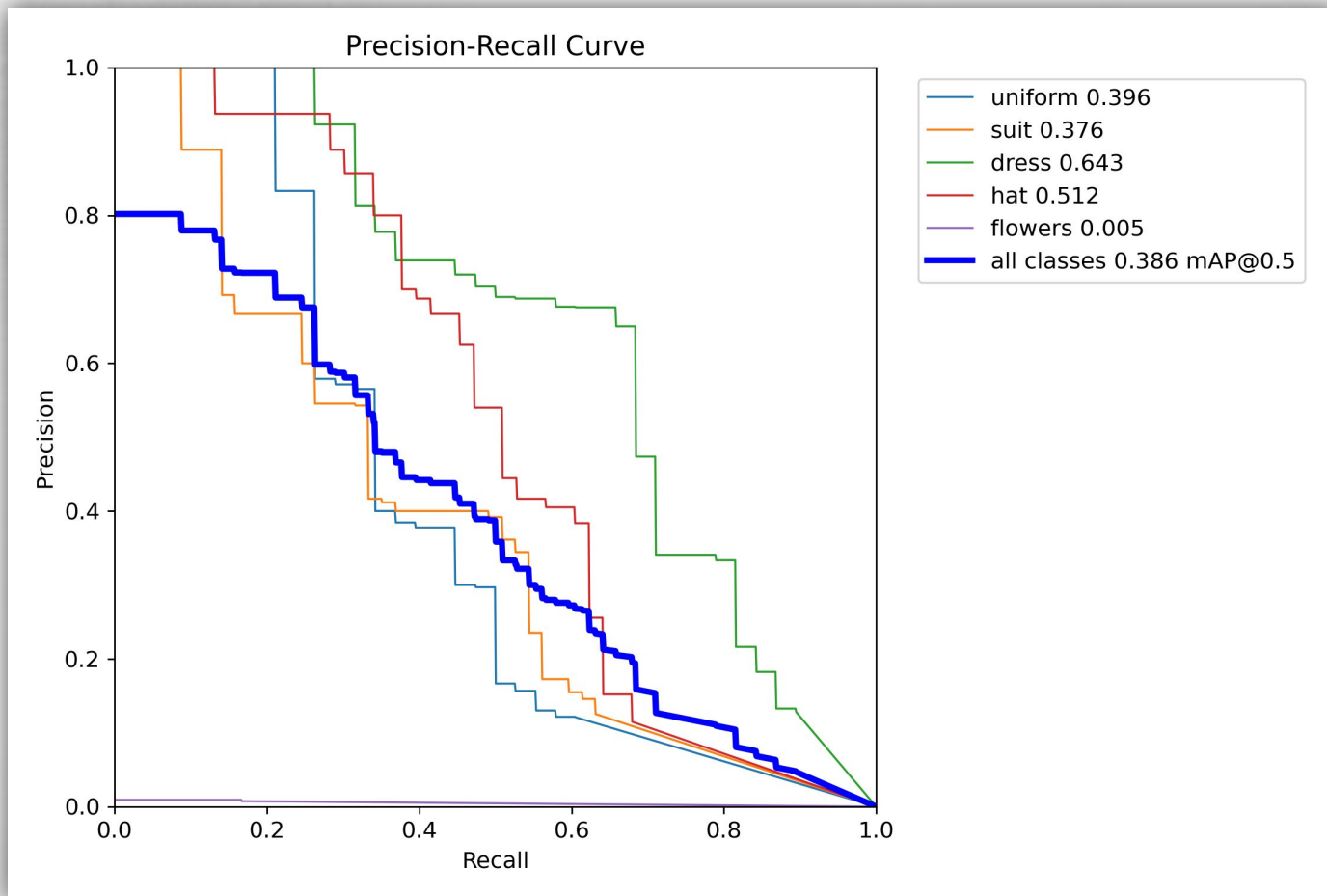
## Validation



# Workflow – results and testing



# Workflow – results and testing

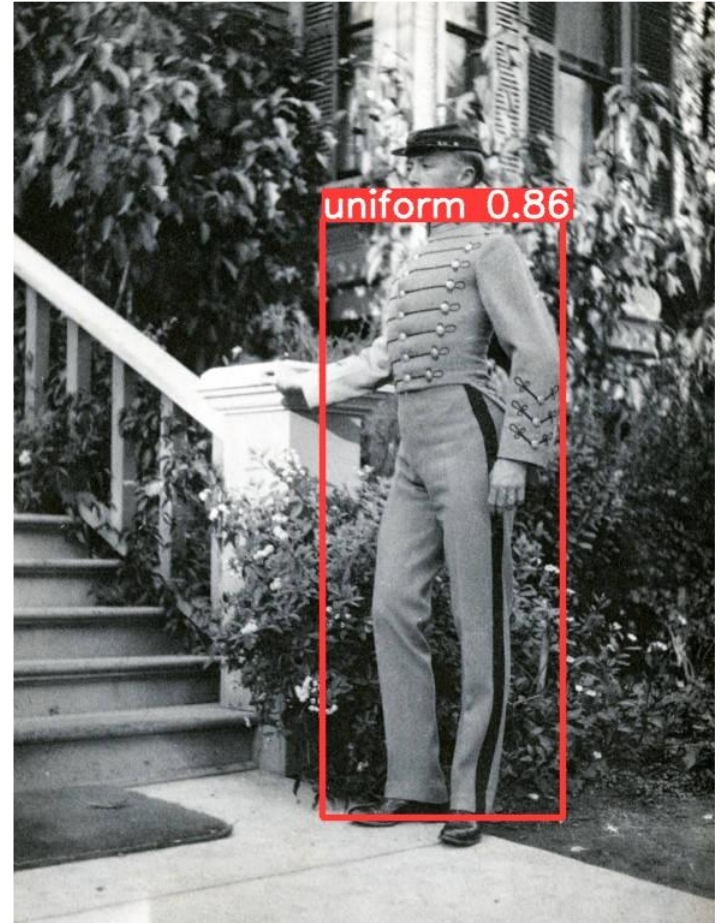




# Workflow – results and testing

- Testing on untrained data
- Model is surprisingly successful in recognition of the trained objects on unlabeled images given the size of the model

Young general Patton



# Workflow – results and testing



Sheraton Princess Kaiulani - reception wallpaper



# Future work

- Setting up a server environment with the trained model
- Allowing anyone to
  - label (annotate) their collections (of digitized portraits) by pointing the model trained on archival images to their set of (archival) images
  - effectively reduce the time needed for repetitive and time-consuming “addition, gathering, and extracting metadata”



# 7. Conclusion

- Trying to get as straightforward as possible
- Aiming to create a product (trained AI model) to be easily used by an average archivist
- Develop a guidance how to replicate the training process on other collections



# THANK YOU!



## Training AI on digitised portraits

**dr. sc. Hrvoje Stančić, full. prof.**

Faculty of Humanities and Social Sciences

University of Zagreb, Croatia

[hstancic@ffzg.unizg.hr](mailto:hstancic@ffzg.unizg.hr)

InterPARES Trust AI Symposium

Honolulu, Hawaii, USA

23 February 2024