# Human-interpretable and deep features for privacy classification

Darya Baranouskaya    Andrea Cavallaro

Centre for Intelligent Sensing, Queen Mary University of London, UK
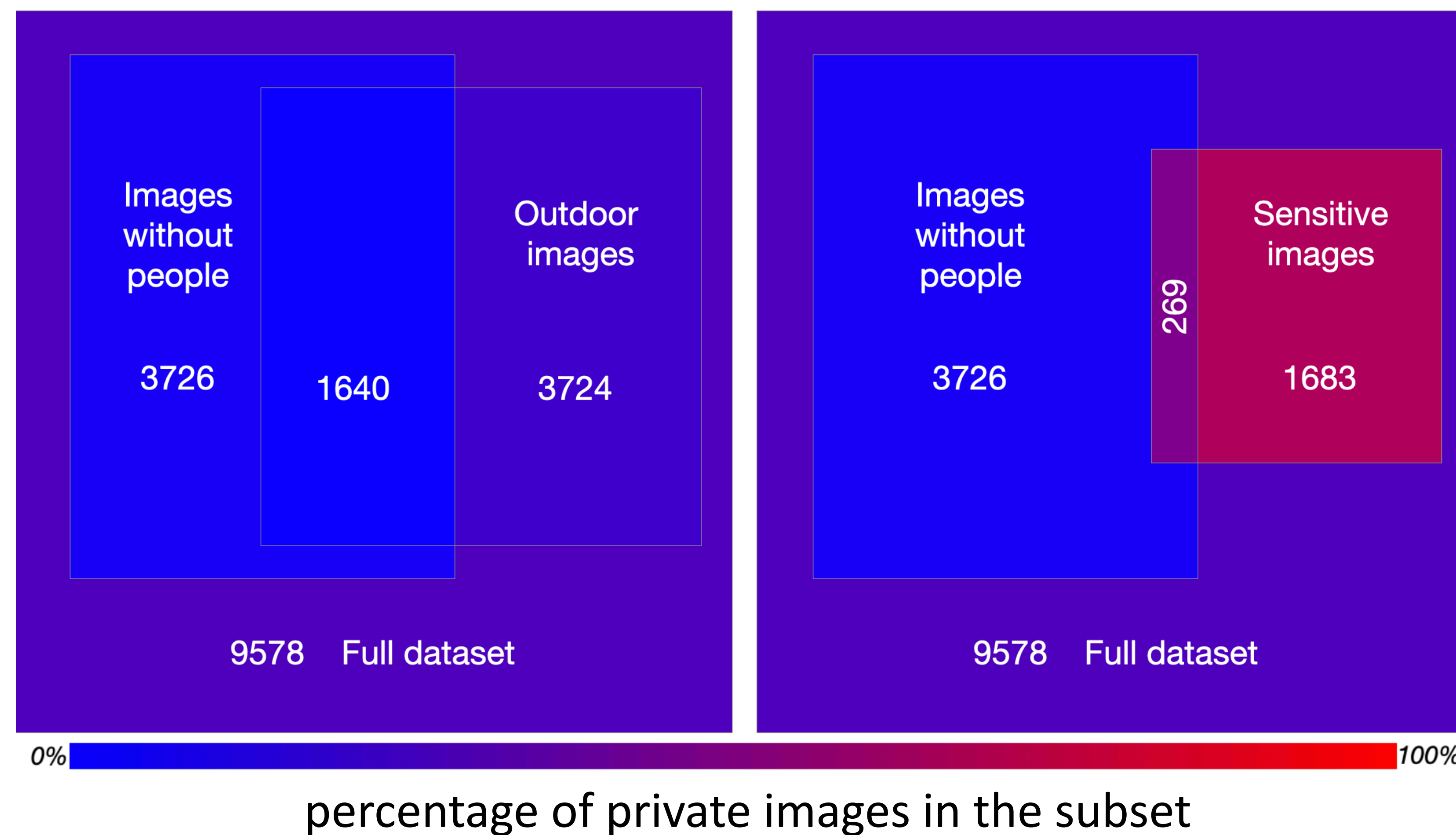
## Goal

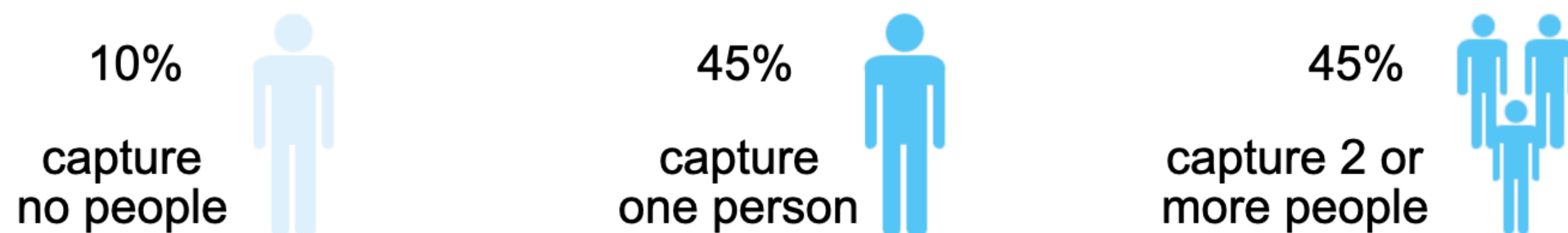An interpretable and extensible method for privacy classification

## Contribution

- Analysis of privacy datasets and properties of controversially labelled images
- Propose eight privacy-specific and human-interpretable features for privacy classification (8PS)
  - Improved performance over higher dimensional deep features
  - Improve the performance of deep features
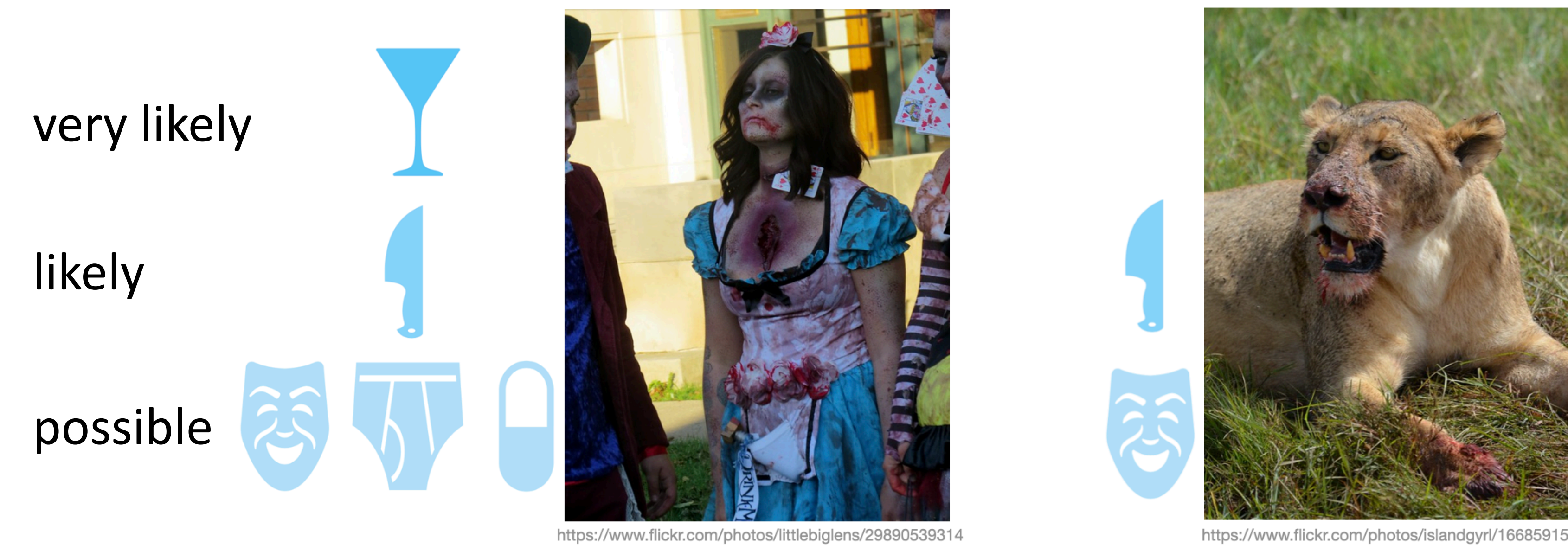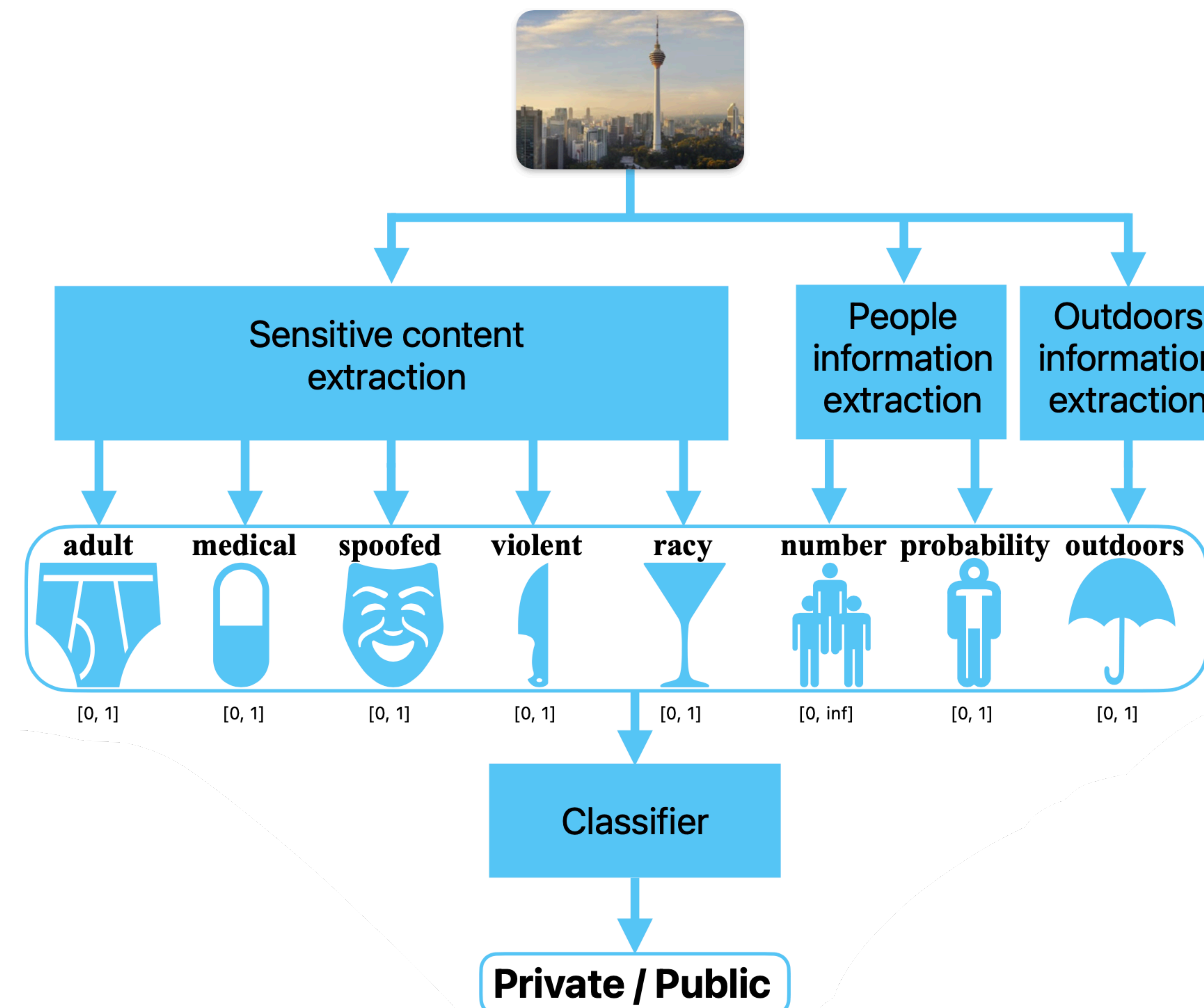
## Dataset content



percentage of private images in the subset

## Controversial images

No privacy label was chosen by more than 65% of annotators



10% capture no people

45% capture one person

45% capture 2 or more people

## Sensitive & Public



very likely

likely

possible

https://www.flickr.com/photos/littlebiglens/29890539314

https://www.flickr.com/photos/islandgyrl/16685915265

## Method



Sensitive content extraction

People information extraction

Outdoors information extraction

adult [0, 1]  medical [0, 1]  spoofed [0, 1]  violent [0, 1]  racy [0, 1]  number [0, inf]  probability [0, 1]  outdoors [0, 1]

Classifier

**Private / Public**

## Results

LogReg on different subsets of features

| Sensitive | People | Outdoors | Places | Deep features | BA | F1 |
|---|---|---|---|---|---|---|
| ✓ | | | | | 80.04 | 66.46 |
| | | | ✓ | | 73.01 | 56.41 |
| | | | | ResNet18 | 78.99 | 63.97 |
| | | | | ResNet50 | 81.37 | 66.99 |
| | | | | ResNet101 | 81.51 | 67.46 |
| ✓ | ✓ | | | | 81.23 | 67.76 |
| | ✓ | ✓ | | | 74.69 | 57.58 |
| | ✓ | | ✓ | | 75.56 | 59.34 |
| ✓ | ✓ | ✓ | | | 80.96 | 66.54 |
| ✓ | ✓ | | ✓ | | 81.22 | 67.91 |
| | ✓ | ✓ | ✓ | | 74.23 | 57.74 |
| ✓ | ✓ | ✓ | ✓ | | 80.83 | 67.21 |
| ✓ | ✓ | ✓ | | ResNet18 | 80.39 | 65.82 |
| ✓ | ✓ | ✓ | | ResNet50 | **81.93** | 67.71 |
| ✓ | ✓ | ✓ | | ResNet101 | 81.80 | **67.79** |

BA - balanced accuracy, F1 - f1 score for private class

MLP on deep and 8PS features, F1 score

| Deep features | *without 8PS* | ⊕ *8PS* |
|---|---|---|
| - | - | 69.72 |
| ResNet101 | 70.64 | 71.07 |
| ConvNext tiny | 69.03 | 70.63 |
| ConvNext small | 70.00 | 70.47 |
| ConvNext base | 70.35 | **72.63** |
| Swin tiny | 69.67 | 71.79 |
| Swin base | 71.44 | 72.18 |

## Conclusion

We proposed a set of privacy-specific, human-interpretable features that achieves comparable performance to higher-dimensional features

## References

[1] A.Tonge and C. Caragea, "Image Privacy Prediction Using Deep Neural Networks," ACM Trans. Web, vol. 14, no. 2, 2020.
[2] S. Zerr, S. Siersdorfer, J. Hare, and E. Demidova, "Privacy-aware image classification and search," 35th Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2012.
[3] C. Zhao, J. Mangat, S. Koujalgi, A. Squicciarini, and C. Caragea, "PrivacyAlert: A Dataset for Image Privacy Prediction," Proc. 16th Int. AAAI Conference on Web and Social Media, 2022.

Now the authors are at the Idiap Research Institute darya.baranouskaya@idiap.ch, andrea.cavallaro@idiap.ch