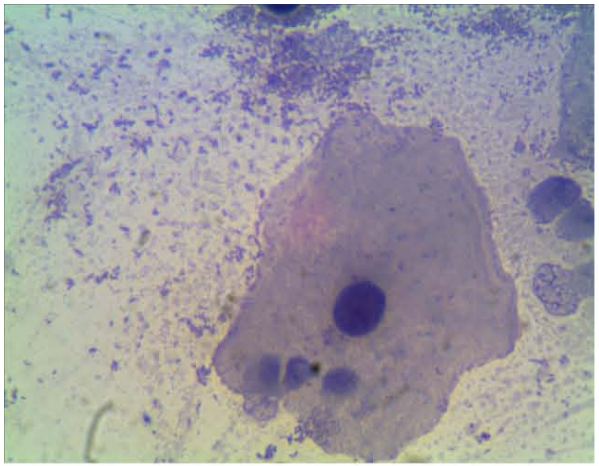
# Automated Pap Smear Analysis

Nicki Dell and Waylon Brunette

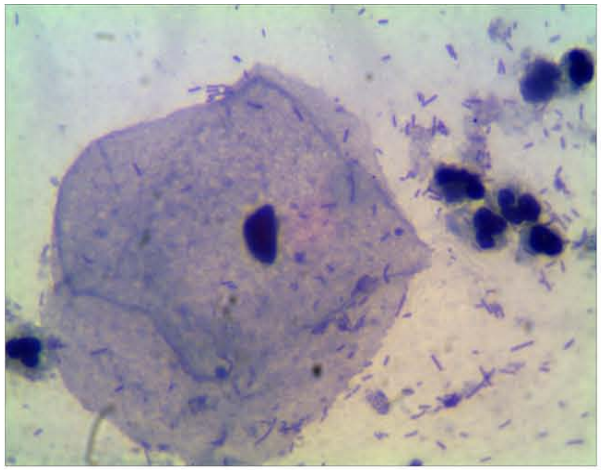
# Cervical Cancer

- The second most common cancer in women worldwide.
- Kills 270 000 women every year, mostly in developing countries.
- Can be cured in almost all patients if detected and treated early.

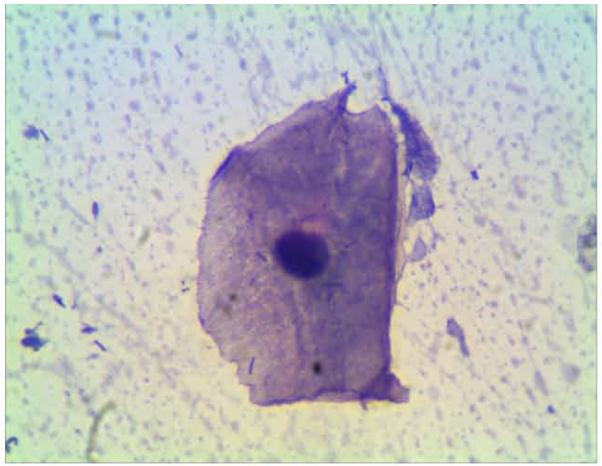
# Pap Smear Screening



Normal

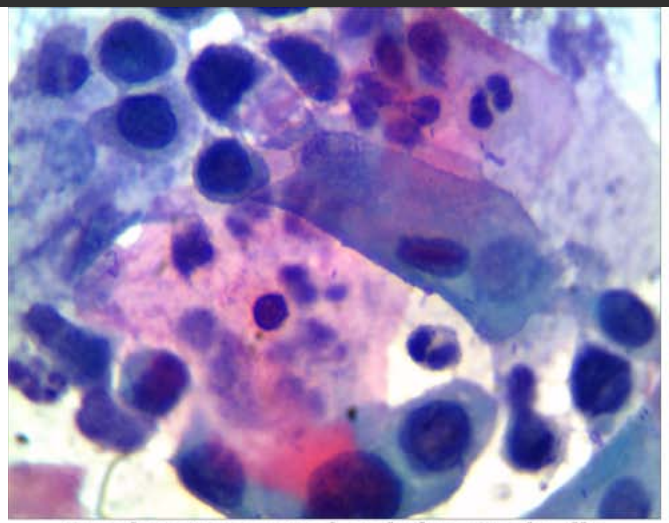


Abnormal - LSIL

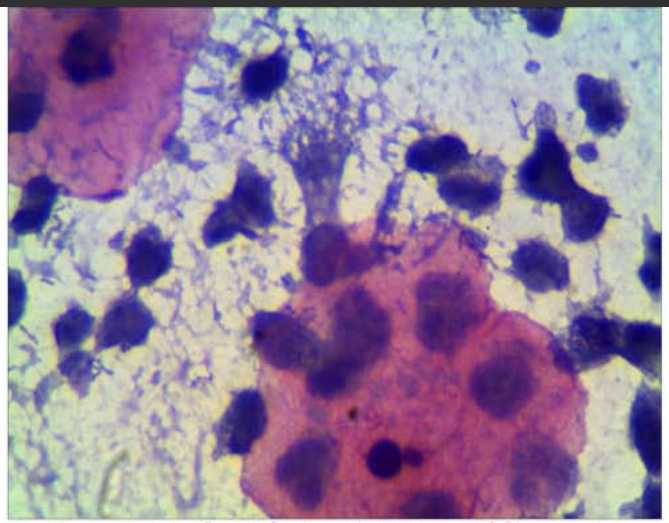


Abnormal - HSIL

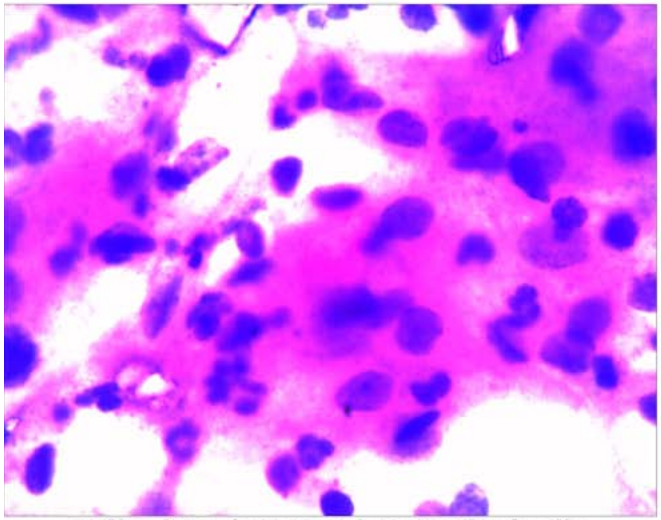
# Why is automated analysis hard?



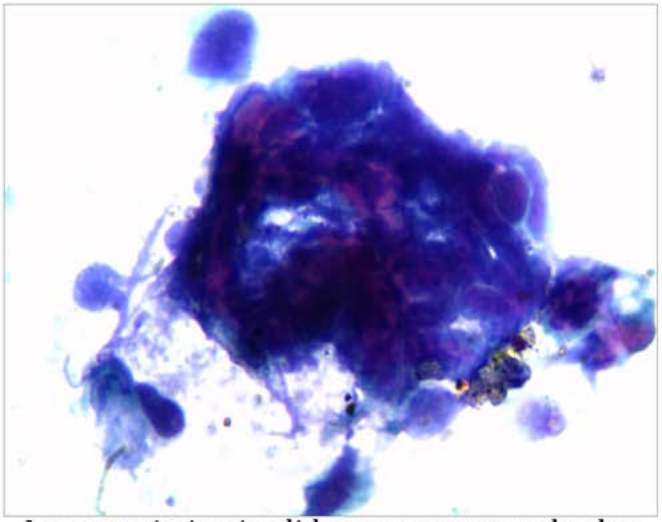
Overlapping normal and abnormal cells



Presence of artifacts, viruses and bacteria



Difficult to distinguish individual cells



Large variation in slide appearance and color

# Objective

- Perform preliminary experiments to determine if abnormal cells can be detected automatically.
- Hopefully lead to a larger research project.
- Segment and extract features from slide images.
- Classify images as normal or abnormal.

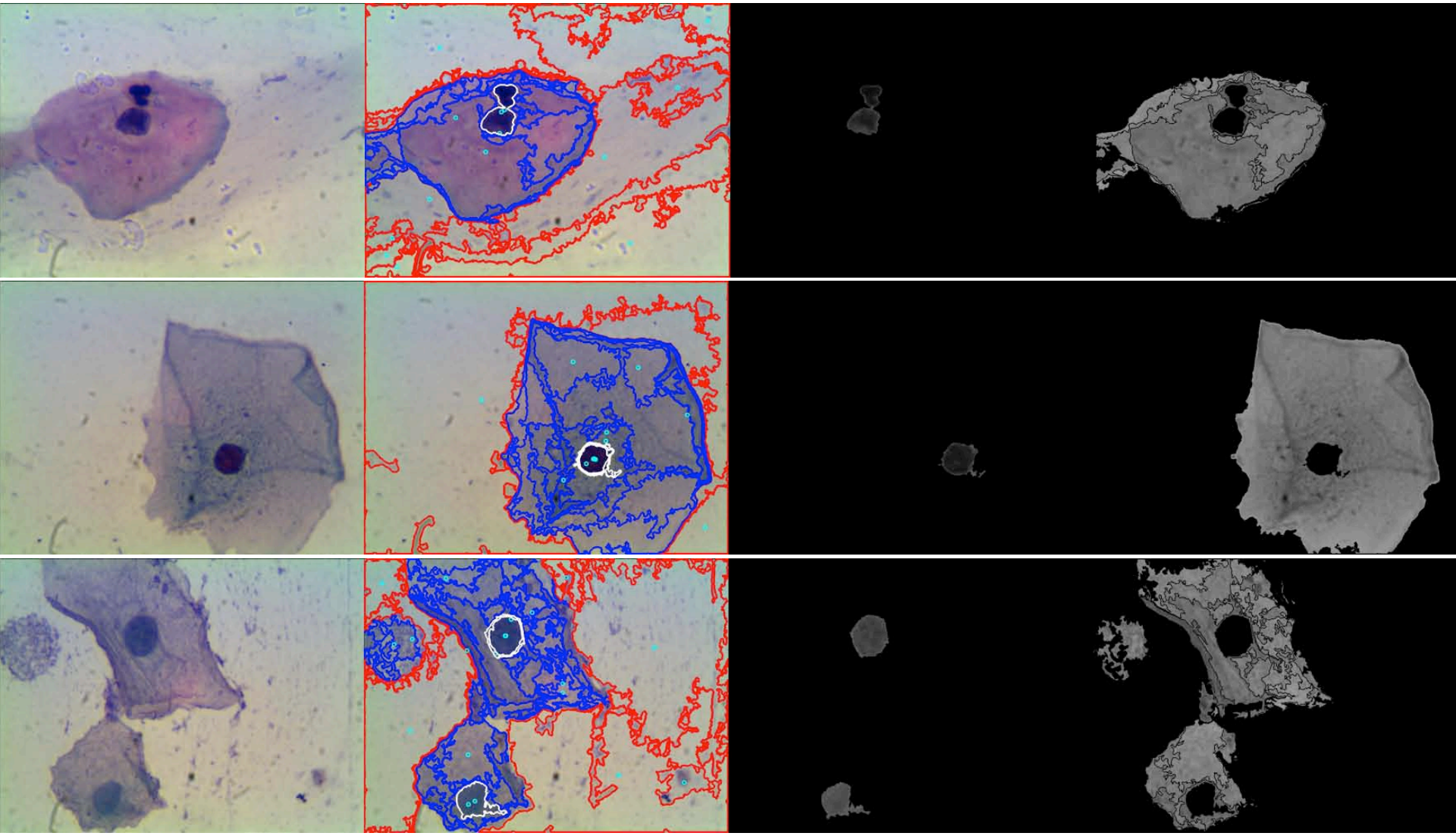
# Dataset

- Ainbo Project
  - Women in the Peruvian jungle
  - 341 anonymous slides
  - Courtesy: Dr. Magaly Blas & Dr. Isaac Alva
- Collection of 200 digital images
  - Zeiss Microscope, 100x, Max illumination, 3MP digital camera
  - 100 normal images
  - 100 abnormal images - 50 LSIL and 50 HSIL
  - Subset of 35 simpler “single-cell” images (18 normal, 17 abnormal)
  - Courtesy: Mabel Raza

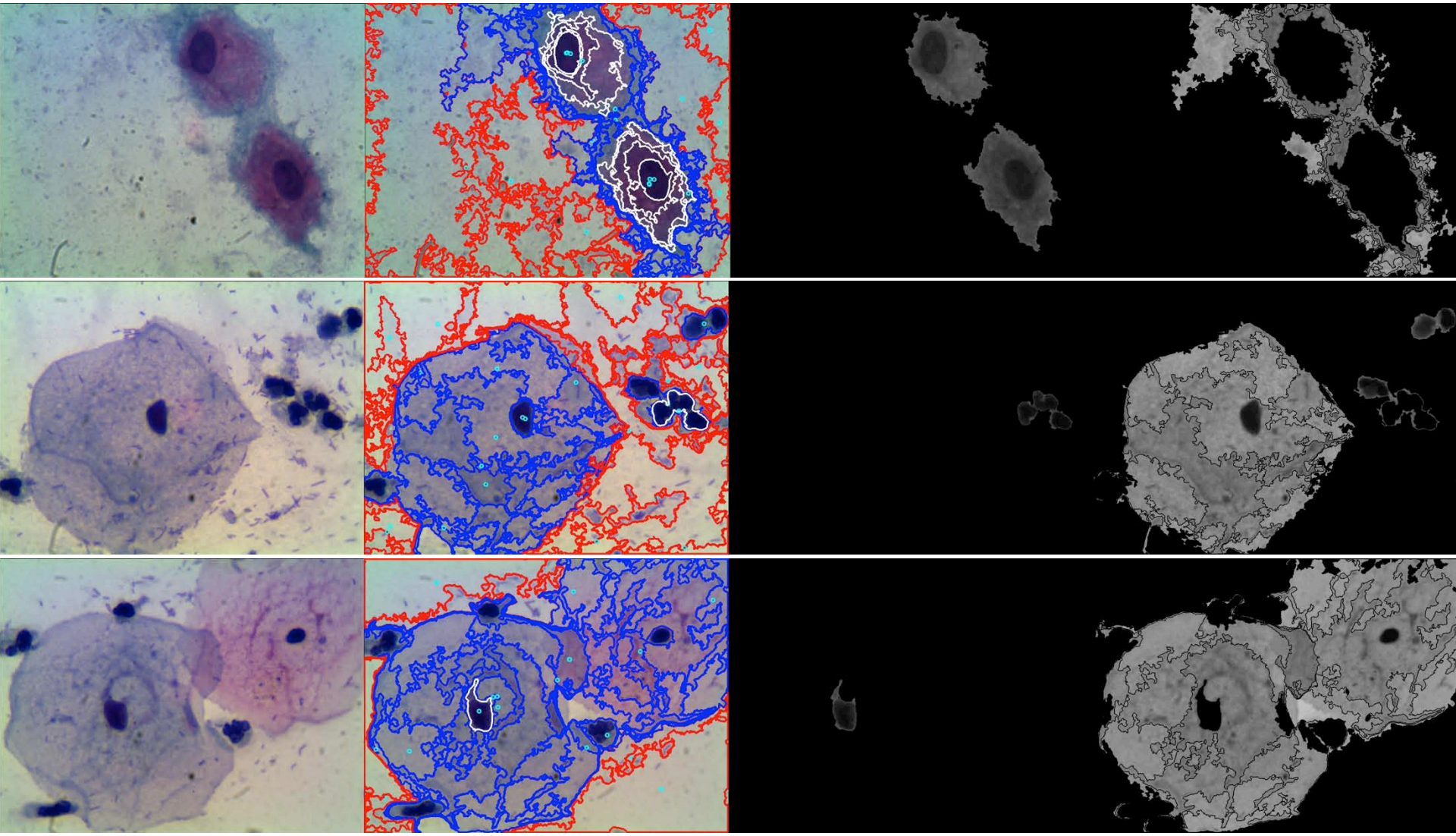
# Method

- K-means clustering in grayscale
- Connected components
  - Label regions as nucleus, cytoplasm, background
- Feature extraction
  - Nucleus color, cytoplasm color, K-means color
  - Nucleus area, cytoplasm area
  - Nucleus roundness, cytoplasm roundness
  - Nucleus entropy, cytoplasm entropy
  - Nucleus texture, cytoplasm texture
  - Nucleus:Cytoplasm ratio
- Image Classification
  - Multiple classifiers from WEKA.

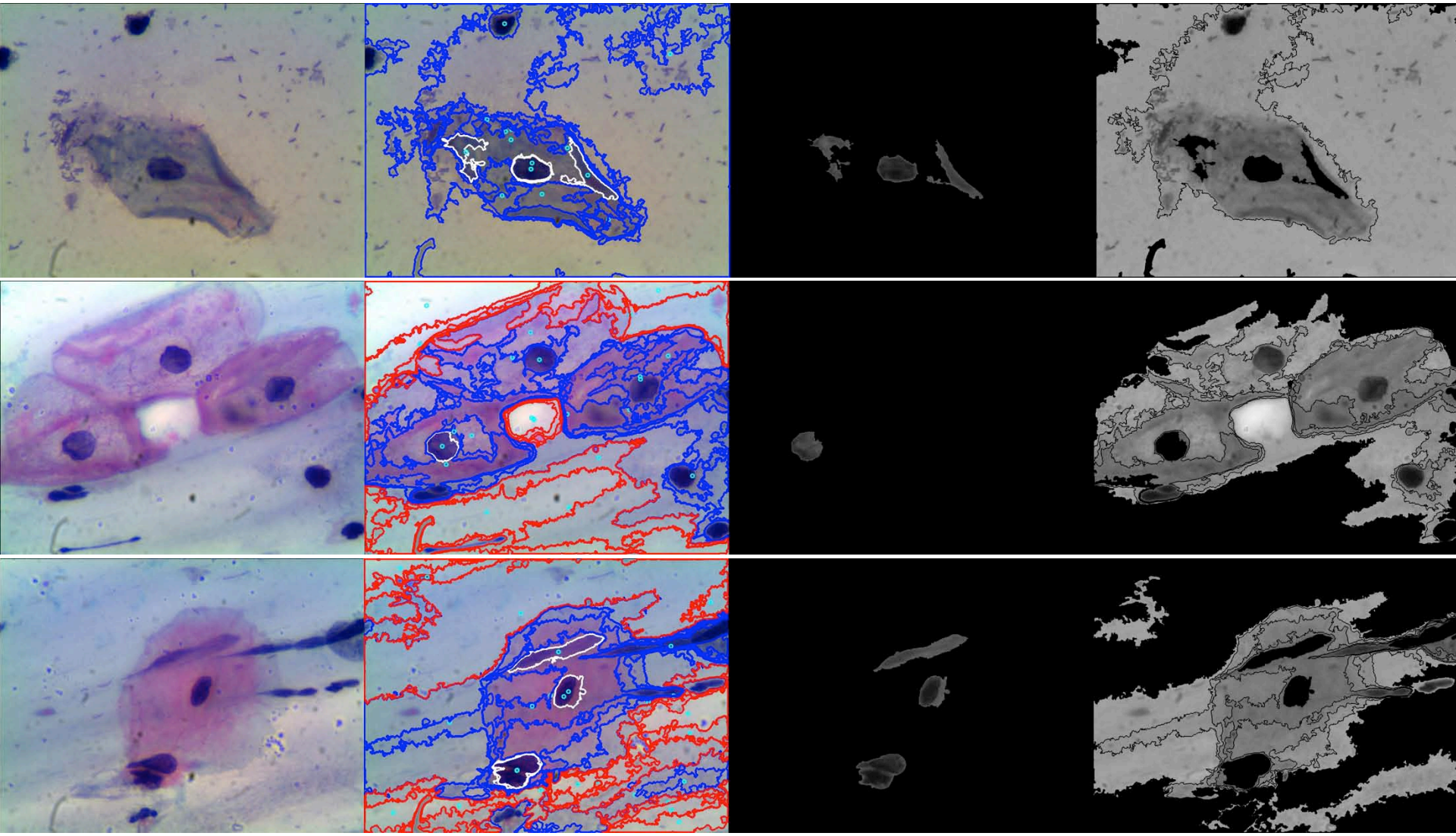
# Results – Images



# Results – Examples of nucleus error



# Results – Examples of cytoplasm error



# Preliminary classification results

| WEKA CLASSIFIER | ALL | | K-Means & Connected Comp | | Entropy & Connected Comp | |
|---|---|---|---|---|---|---|
| | K=6 | K=8 | K=6 | K=8 | K=6 | K=8 |
| NaiveBayes | 35.3 | 82.4 | 58.8 | 47.1 | 58.8 | 82.4 |
| Classification Via Clustering | 70.6 | 52.9 | 52.9 | 82.4 | 64.7 | 76.5 |
| DecisionTable | 58.8 | 76.5 | 64.7 | 52.9 | 52.9 | 52.9 |
| Logistic | 52.9 | 76.5 | 64.7 | 70.6 | 52.9 | 58.8 |
| MultilayerPerceptron | 58.8 | 64.7 | 64.7 | 70.6 | 47.1 | 64.7 |
| RandomForest | 64.7 | 76.5 | 88.2 | 52.9 | 76.5 | 64.7 |
| SMO | 47.1 | 70.6 | 64.7 | 70.6 | 52.9 | 70.6 |
| ThresholdSelector | 47.1 | 82.4 | 70.6 | 52.9 | 47.1 | 47.1 |

# Limitations and Future work

- Better segmentation
  - Slides with multiple cells and overlapping cells
- Better feature extraction
  - More features
  - Better methods for finding features
- Better Classification
  - Work out classifier parameters
- Larger data set
  - Handle multiple cells better
  - Handle overlapping cells better

# Conclusions

- Overall results are promising.
- Larger research project - working with collaborators at Cayetano Heredia University in Peru.
- Need a LOT more work.