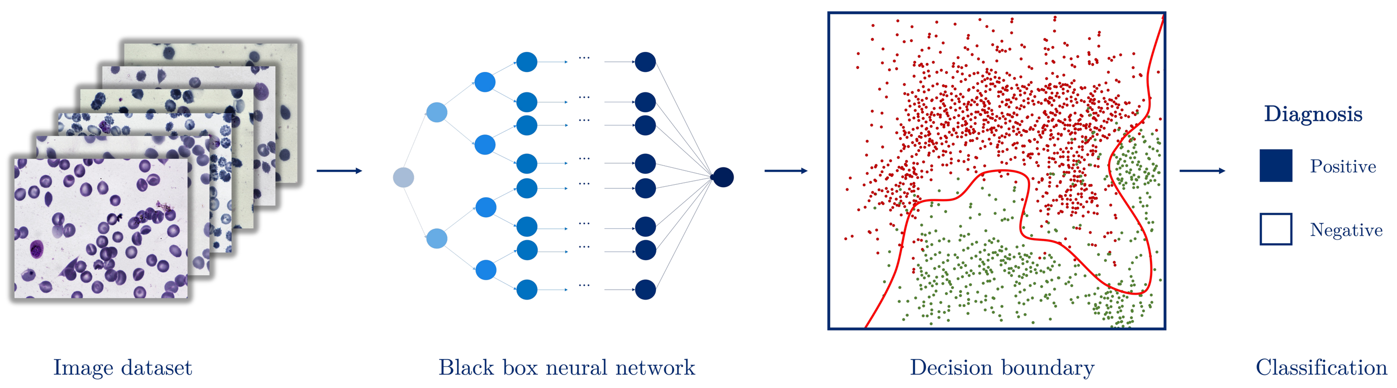
Generative approaches to Shapley-based explanations

Ishan Kalburge, Siyu Wang, Kuai Yu
Department of Biomedical Engineering, Johns Hopkins University

Background

The black box problem

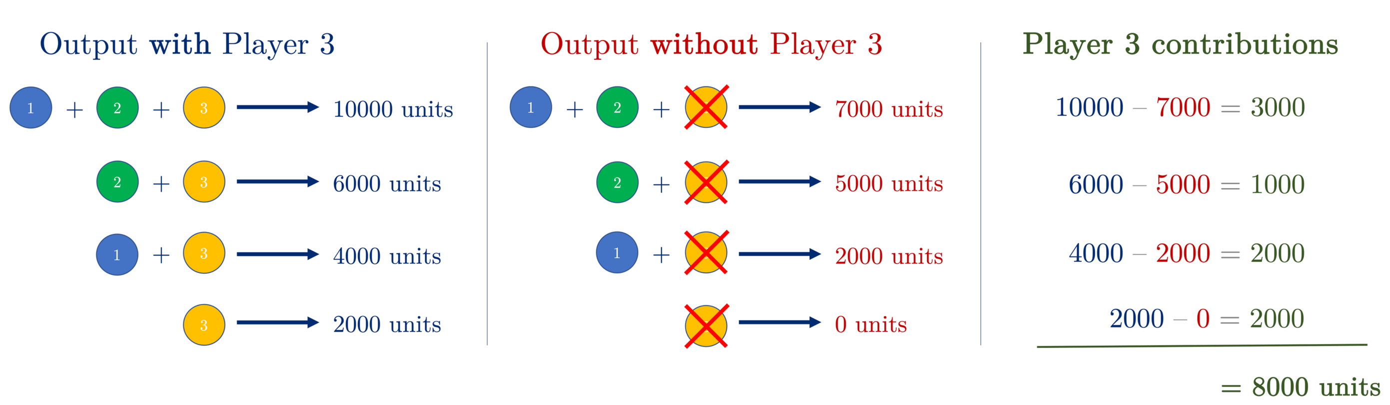
In clinical computer vision settings, deep neural networks (DNNs) can diagnose diseases by creating complex mathematical relationships between image data, sometimes estimating millions of parameters. But how can clinicians trust and verify the conclusions of these DNNs, especially when these mathematical relationships are so complicated?



A posteriori methods: Shapley values

Because DNNs are valuable for their raw predictive power, we can implement methods that can interpret the results of a DNN for us!

Shapley values, a tool inherited from cooperative game theory, offer a simple but elegant solution. A Shapley game computes the value added for each member of a team – or, likewise, each feature of an image.



If we treat each feature of an image as players in a Shapley game as shown above, we can summarize the predictive values for a feature using a weighted sum to determine its overall importance:

$$\phi_j(v) = \phi_j = \sum_{S \subseteq M \setminus \{j\}} \frac{|S|!(M-|S|-1)!}{M!}(v(S \cup \{j\}) - v(S)), \quad j = 1, ..., M,$$

Problems

<u>But what constitutes a player in a Shapley game?</u>

Computational feasibility

- How many players can we have before the problem becomes intractable?

Semantic Relevance

- How do we pinpoint features that have conceptual importance? Which **"things,"** rather than pixels, does a DNN latch onto?

<u>And how should we treat the players we remove?</u>

Statistical Accuracy

- If we remove part of the image and replace it with a black space, we are creating an image that does not truly exist in our distribution. How do we fix this?

Our Approach

<u>**Reduce**</u> the number of players

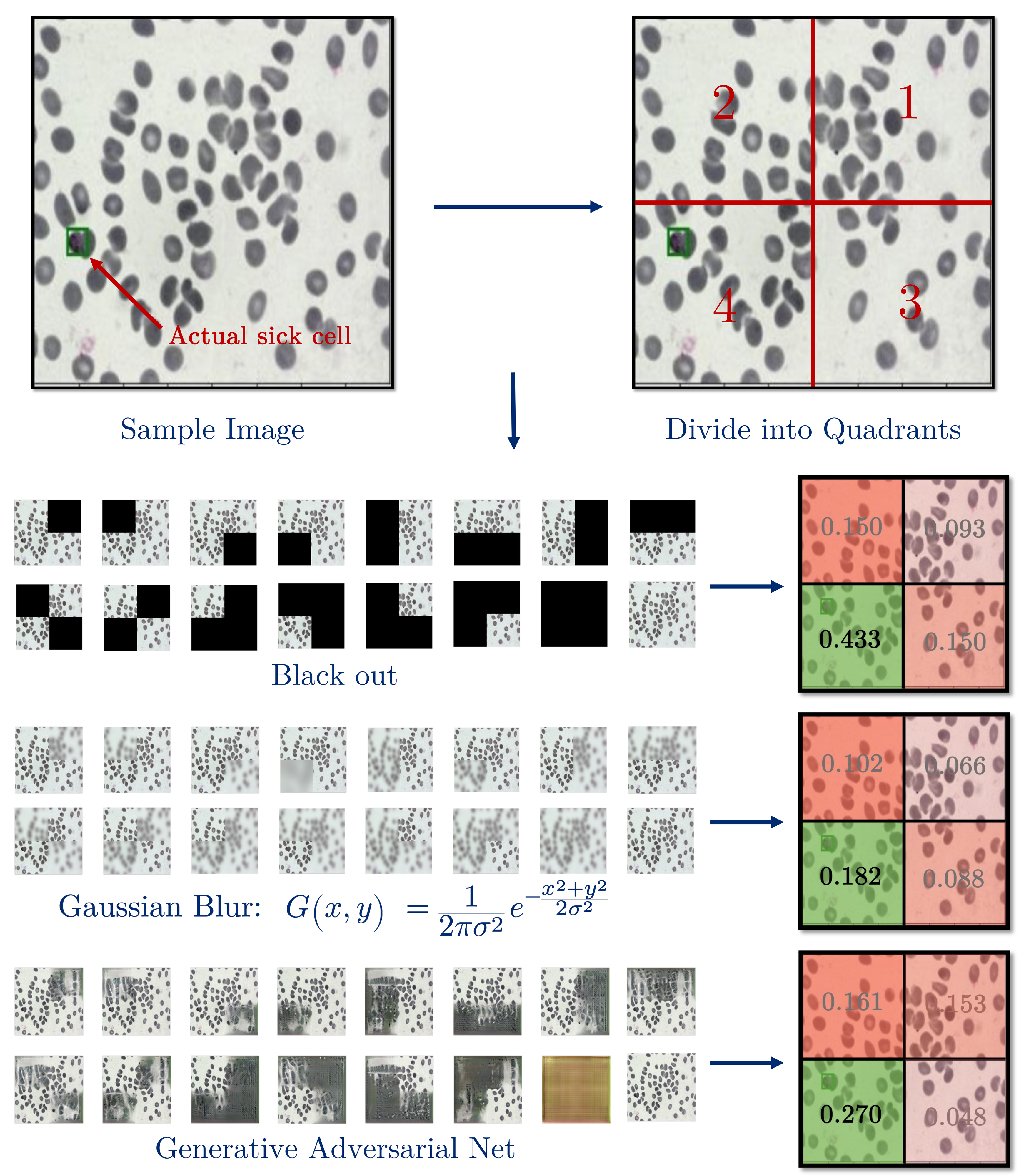
Partition

Segmentation

Unsupervised clustering

<u>**Inpaint**</u> during Shapley games

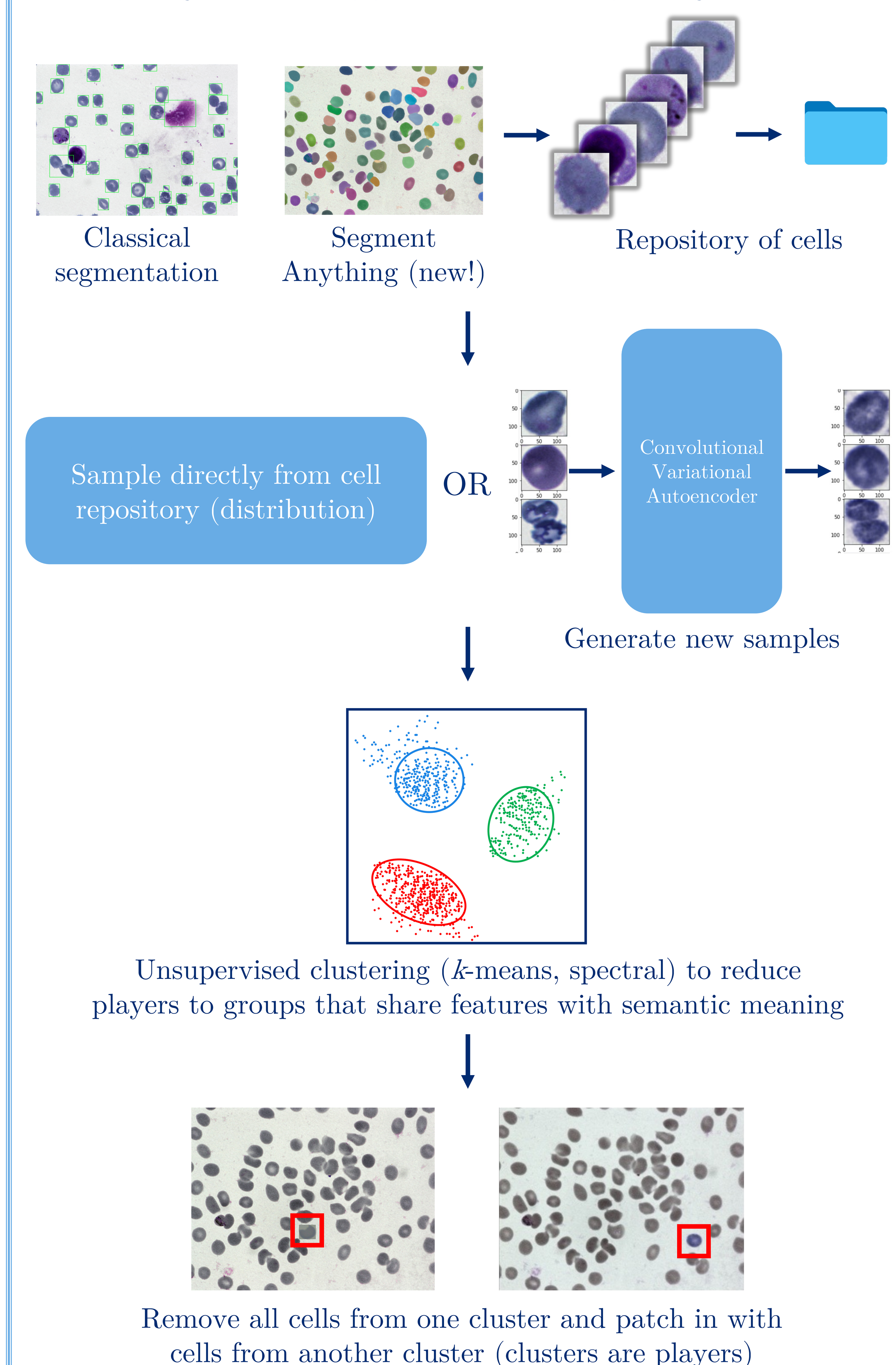
Partition Pipeline Comparison



Segmentation + Clustering Pipeline

Alternatives to Generative AI

Even when Generative Adversarial Nets (GANs) are trained on our image set, they poorly inpaint removed portions of images – which is not surprising. We offer a pipeline for manual inpainting:



Next Steps

Segmentation and patching process

- Use Segment Anything? How can we patch for inconsistent dimensions?

Clustering process

- Should we use handcrafted features? Should we cluster within or over entire dataset?

Codebase and documentation