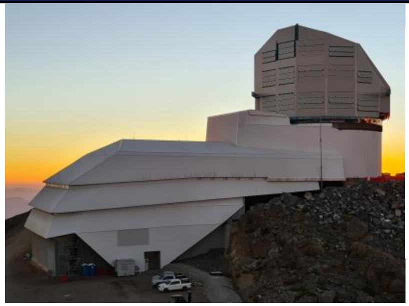
Lawrence Livermore National Laboratory

Gaussian Processes for Galaxy Blend Identification in LSST

James J. Buchanan[1], Michael Schneider[1], Robert Armstrong[1], Amanda Muyskens[2], Benjamin Priest[3], Ryan Dana[3]

[1]Physics Division, LLNL; [2]Engineering Division, LLNL; [3]Center for Applied Scientific Computing, LLNL

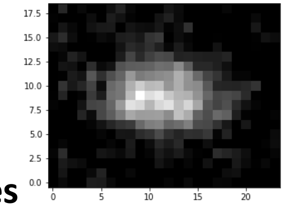
LSST



VERA C. RUBIN OBSERVATORY

- The Vera C. Rubin Observatory, currently under construction in Chile, will commence the 10-year Legacy Survey of Space & Time (LSST) beginning in October 2023
- **LSST will catalog an unprecedented number of galaxies**

PROBLEM: BLENDING

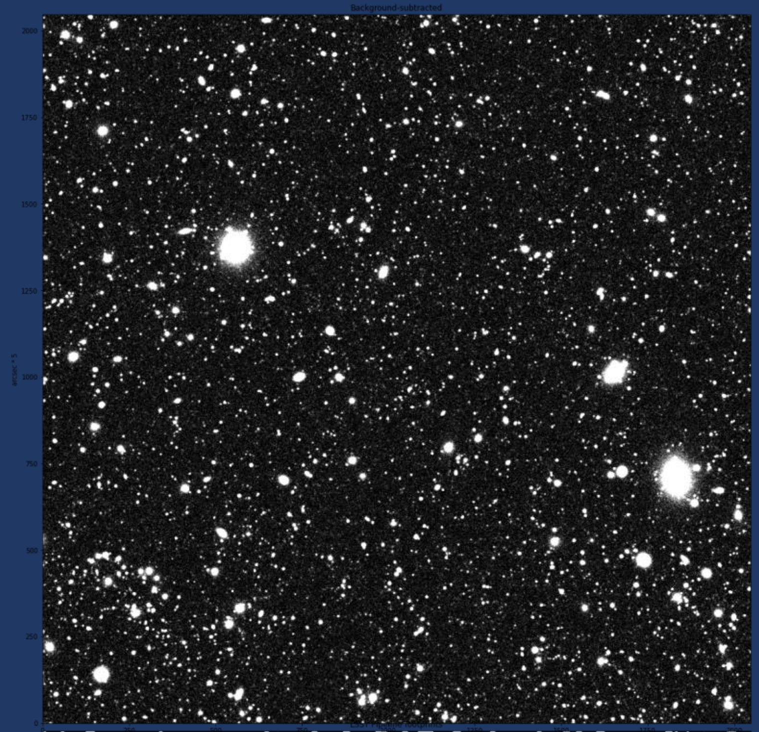


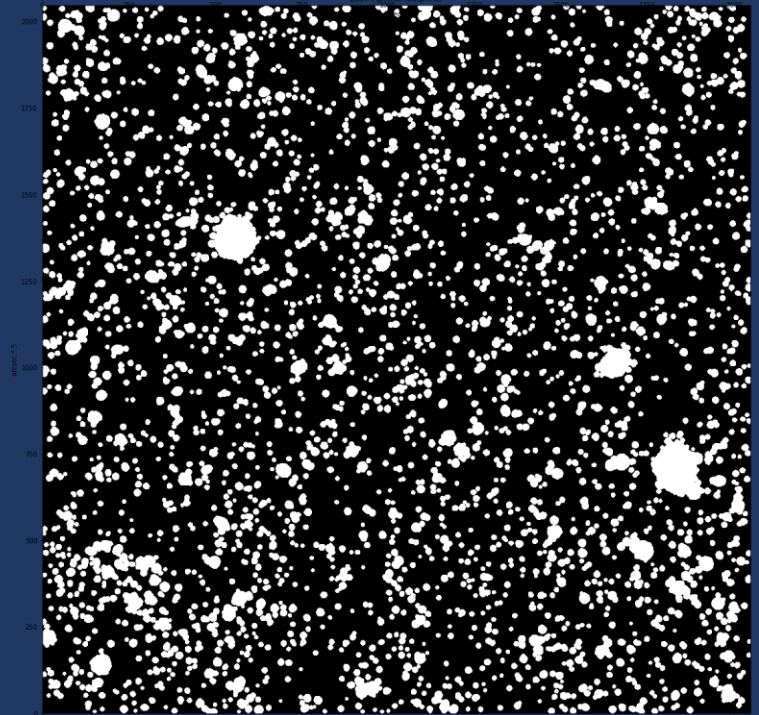
- Roughly half of all observed galaxies will overlap another galaxy along the same line of sight: “**blending**”
- Blending makes it difficult to measure galaxy shapes
- **Need to reliably identify instances of blending in images containing billions of galaxies**

Lawrence Livermore National Laboratory

IMAGE SIMULATION

- **cosmoDC2 + DESC DC2 catalogs**: comprehensive simulation of galactic light propagation, with distribution of galaxies based on the Outer Rim N-body simulation
  - Galaxy locations and light profiles taken from these catalogs
- Images rendered using GalSim, configured using expected LSST parameters
- Simulated 20 scenes this way, 2048^2 pixels each, containing 134,493 galaxies
  - i-band only for now
- Result: **Realistic images with realistic distributions of galaxies**
- Provides a **science-relevant training set** and allows for straightforward **estimates of classifier performance on realistic data distributions**





Lawrence Livermore National Laboratory

IMAGES TO FOOTPRINTS

For each simulated scene:

- Estimate and subtract **background**
- Construct footprints
  - **Convolve** scene with point spread function
  - Identify bright pixels with **S/N > 5**
  - Contiguous blobs of bright pixels: "**footprints**"
  - **Expand** footprints by a few pixels to pick up diffuse edges of galaxies

- 66% of galaxies are contained in footprints

Define a footprint as blended if it contains more than 1 galaxy; unblended otherwise

- 62% of footprints are blended
- 38% of footprints are unblended

FOOTPRINTS TO MODEL INPUT

For each footprint:

- Focus on a "**cutout**" – a small square array of image pixels centered on footprint
- **Zero** out pixels not part of footprint
- **Normalize** pixel values
- **Flatten** cutout into 1D vector
- Reduce dimensionality using **PCA** embedding

PEAK FINDING

- Convolve each footprint by PSF
- In smoothed footprint, **count number of intensity peaks**
- Classify a footprint as blended if > 1 peak; unblended otherwise

Current default method in the Hyper Suprime-Cam data reduction pipeline and LSST Science Pipelines

GAUSSIAN PROCESS MODEL

Gaussian process: A collection of random variables, any finite subset of which is Gaussian-distributed

The random variables: For each possible value of the PCA-embedded footprint vectors, yield a number specifying "blendedness"

The Gaussian distribution: Assert a **prior** on the joint training+testing blendedness distribution:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(0, \sigma^2 \begin{bmatrix} K_{\mathbf{ff}} + \tau^2 I_n & K_{\mathbf{f}*} \\ K_{*\mathbf{f}} & K_{**} \end{bmatrix}\right)$$

Lawrence Livermore National Laboratory

GAUSSIAN PROCESS CLASSIFICATION

Given the known blendedness values y of training footprints (+1 or -1), we can **analytically compute** the **posterior** distribution for the blendedness f* of test footprints:

$$\mathbf{f}^* | X_{train}, X^*_{test}, y \sim \mathcal{N}(\bar{\mathbf{f}}^*, \sigma^2 C)$$
$$\bar{\mathbf{f}}^* \doteq K_{*\mathbf{f}}(K_{\mathbf{ff}} + \tau^2 I_n)^{-1}\mathbf{y}$$
$$C \doteq K_{**} - K_{*\mathbf{f}}(K_{\mathbf{ff}} + \tau^2 I_n)^{-1}K_{\mathbf{f}*}$$

Classify footprint as blended if posterior mean f* > 1, unblended otherwise.

Given a random **blended** footprint,

GP classification	has a	80.0%	chance of classifying it correctly.
Peak counting		78.9%	

Given a random **unblended** footprint,

GP classification	has a	94.2%	chance of classifying it correctly.
Peak counting		75.4%	

Unlike peak finding, the **Gaussian process model naturally assigns probability estimates** to its predictions

Relatively well-calibrated compared to other models