Identifying Galaxy Blends with Gaussian Processes

James Buchanan, Michael Schneider, Robert Armstrong,
Amanda Muyskens, Benjamin Priest

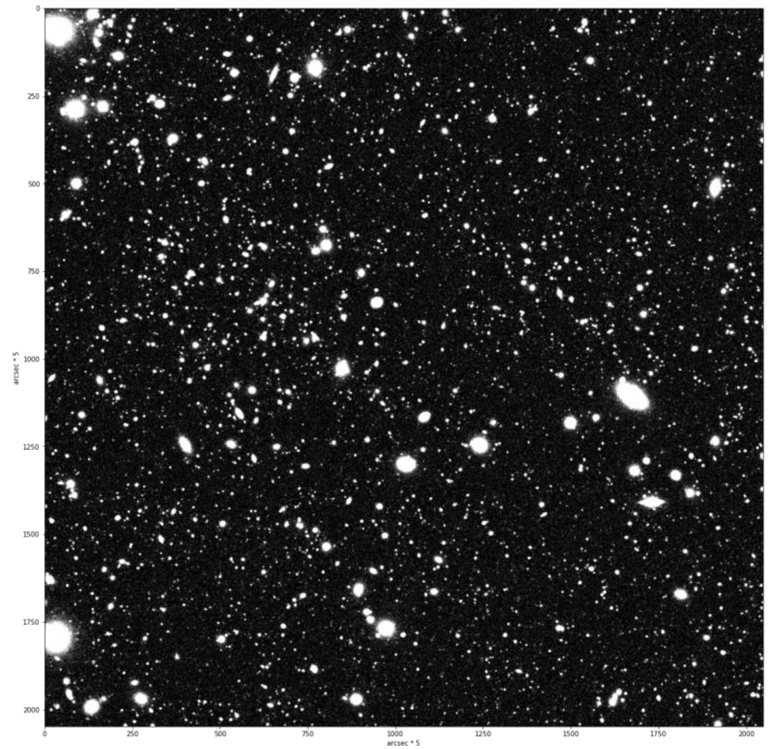
April 26, 2021

LLNL-PRES-XXXXXX

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Galaxy scene simulation

- **Sersic model bulge and disk for each galaxy**
  - Sersic parameters, ellipticity components, relative component fluxes from cosmoDC2 catalog; overall flux in each band and lensed RA,Dec from DESC DC2 truth catalog
- Weak lensing shear and magnification
  - Gamma components and convergence from cosmoDC2 catalog
- Kolmogorov PSF
  - FWHM = 0.7 (+- 10% per exposure)
- Random sub-pixel-scale scene offset (‘dither’)
- Photon shooting
- Silicon sensor
  - ‘lsst_itl_32’ in galsim
- Sky background
  - Dark sky magnitudes from smtn-002.lsst.io
  - +- 5% mean flux per exposure
  - Poisson noise in each pixel
- **100 separate exposures simulated, then added together**



i-band, 2048^2 pixels

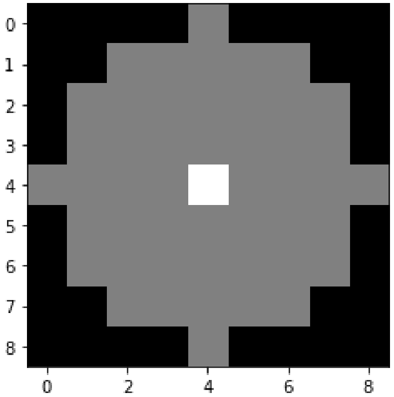
(409.6^2 arcsec)

Footprint Construction

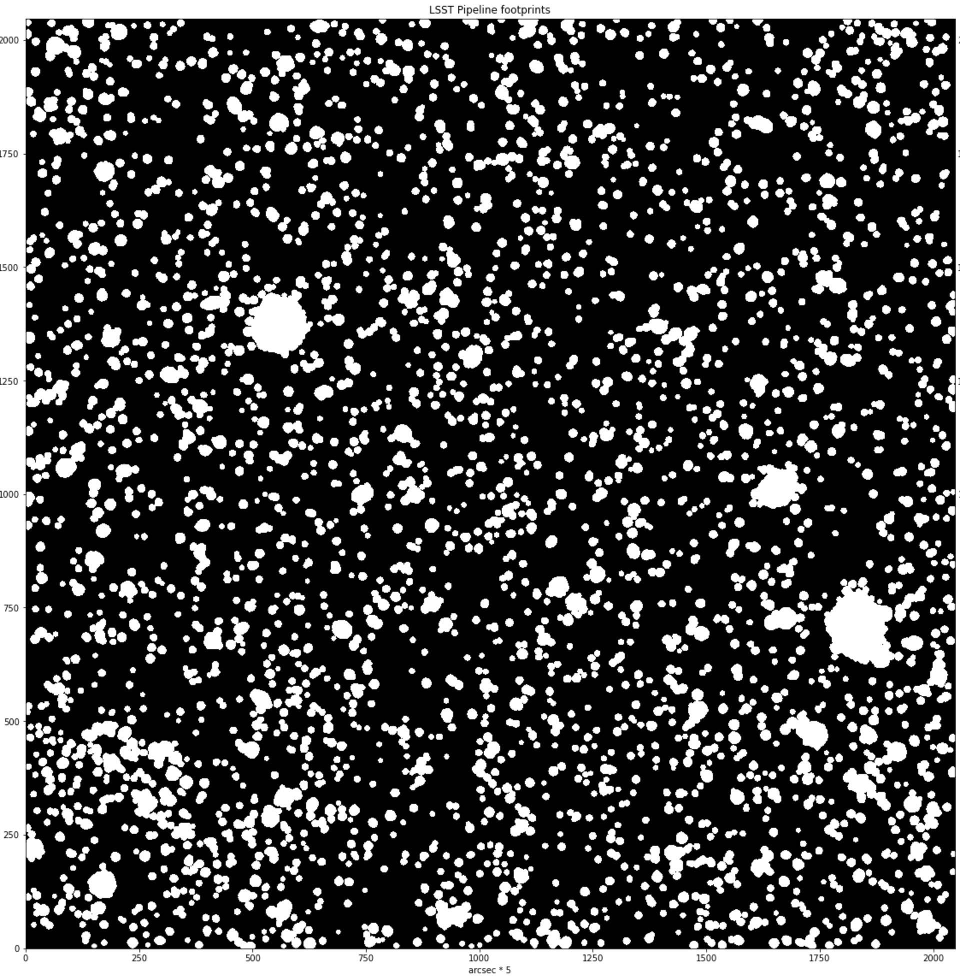
- Subtract estimated sky background
- Convolve with Gaussian approximation of PSF
- Threshold each pixel at S/N >~ 5 to get initial footprints
  - In the background-subtracted, PSF-convolved image, single-pixel S/N = pixel intensity / sqrt(sky)*sqrt(A)

    where A = sum over pixels of (integrated, normalized PSF)^2

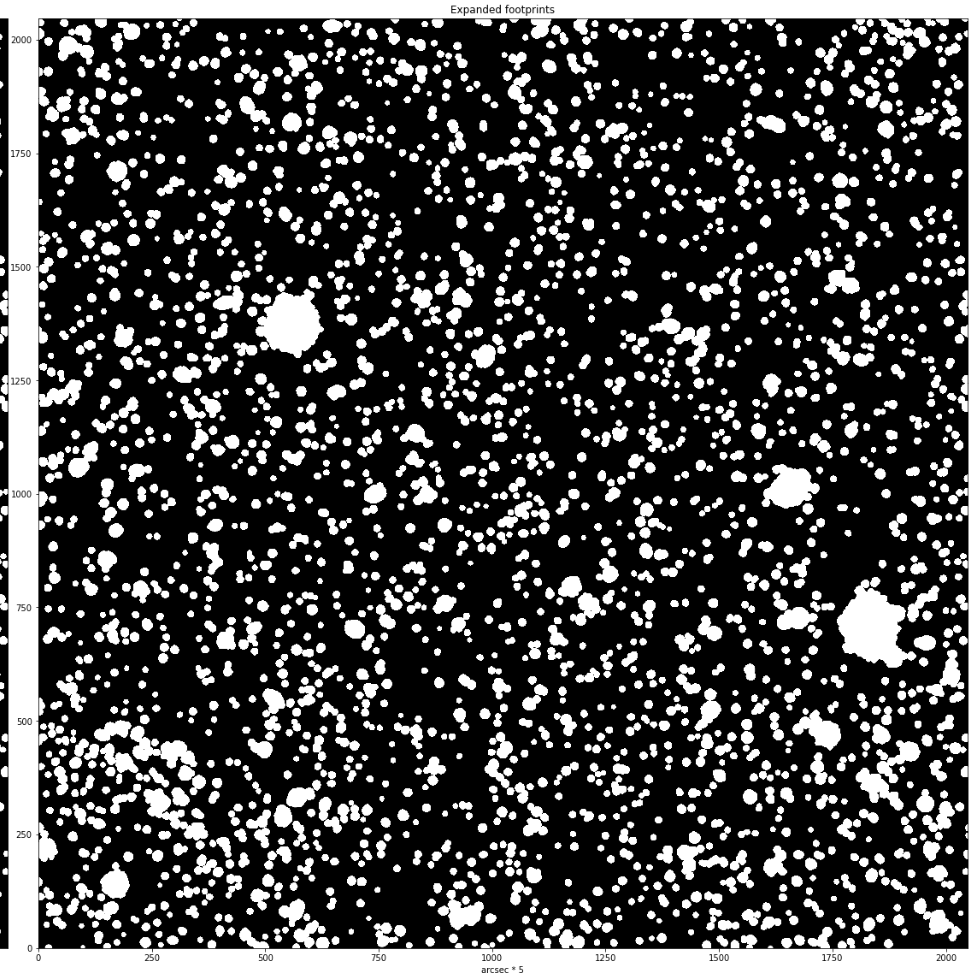
    [*doi:10.1093/pasj/psx080*]
- Expand these initial footprints by ~2.4*PSF width
- Merge the expanded footprints



LSST Pipeline footprints



My replication



Dataset

- **Define an i-band footprint as blended** if it **contains the center** of > 1 galaxy with **5-sigma i-band flux**

- Across 10 total scenes:
  - 65299 total galaxies with i-band flux >= 5 sigma
  - 64.3% of these galaxies are contained in i-band footprints
  - **8107 blended footprints**
  - **15137 unblended footprints**
    - For model training/evaluation: Choose a random subset of unblended footprints so that datasets are balanced
  - 0.4% of footprints contain no galaxies
    - These are on the scene boundaries, cut off at the edges
    - Ignoring these here

Preprocessing

For each footprint:

- Make a **cutout** of a fixed size, centered on that footprint
  - >= 23 pixels to a side
  - Specific centering strategy doesn't matter much
- Zero out any pixels that aren't part of the footprint
- Flatten the pixel array and normalize
  - Specific normalization doesn't matter much as long as values are constrained to lie between 0 and 1
- **PCA embedding** to reduce dimensionality
  - PCA dimension between 7 and 10

Gaussian Process Model

- Gaussian process: An infinite collection of random variables, any finite subset of which is Gaussian-distributed
- The random variables: **For each possible value of the PCA-embedded data vectors, yield a number specifying the "blendedness"**
  - If that number is > 0, classify the footprint as blended
- The Gaussian distribution: Prior mean of 0; covariance matrix is a function of the observed data vectors (**kernel**)
  - Common kernel choice: RBF
    - One hyperparameter – length scale
  - Generalization: Matérn
    - Additional hyperparameter – smoothness

Gaussian Process Model

- For each training example i, define y_i **= +1 if blended, -1 if unblended**
- Let $\mathbf{f}$ denote the **model-estimated blendedness** of **training** examples, $\mathbf{f}^*$ for **test** examples
- **Matérn kernel:**

$$k_{\text{Matérn}}(\vec{x}, \vec{x'}) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{\|\vec{x} - \vec{x'}\|_2^2}{\ell} \right)^{\nu} K_{\nu} \left(\sqrt{2\nu} \frac{\|\vec{x} - \vec{x'}\|_2^2}{\ell} \right)$$

- **Kernel matrices:**

$$(K_{\mathbf{ff}})_{i,j} \equiv k(x_i^{\text{train}}, x_j^{\text{train}})$$

$$(K_{\mathbf{f}*})_{i,j} \equiv k(x_i^{\text{train}}, x_j^{\text{test}}) = (K_{*\mathbf{f}})_{j,i}$$

$$(K_{**})_{i,j} \equiv k(x_i^{\text{test}}, x_j^{\text{test}})$$

Gaussian Process Hyperparameters

- Kernel length scale (ℓ)
  - Between 1e1 and 1e2
- Kernel smoothness (ν)
  - At least 1

  (Note: As ν -> inf, Matérn -> RBF)
- Assume that $y_i \sim N(f_i, \sigma^2)$
  - σ between 1e-6 and 1e-4

More math

- **Given the PCA encodings of train and test examples,** assert Bayesian **prior** on the joint distribution of **blendedness of training and test sets**:

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

- Additionally **given the actual blendedness of the training examples**, we can **analytically compute the posterior** joint distribution of **blendedness of test set**:

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

- Classify test example as blended if $\bar{\mathbf{f}}^* > 0$

Model Comparison: Replication of LSST Pipeline Footprints

- **GP classifier**
  - Balanced accuracy = **0.884**
  - Unblended acc**: 0.827,** Blended acc**: 0.940**

- **Logistic regression** with l2 regularization
  - Balanced accuracy = **0.827**
  - Unblended acc**: 0.786,** Blended acc**: 0.868**

- **Peak counting**
  - Balanced accuracy = **0.848**
  - Unblended acc**: 0.982,** Blended acc**: 0.713**

- *Binomial uncertainty: 0.001-4*

- *Variability due to random training data selection ~ Bin. unc.*

Model Comparison: Fainter+smaller footprints

- **GP classifier**
  - Balanced accuracy = **0.863**
  - Unblended acc**: 0.810,** Blended acc**: 0.915**

- **Logistic regression** with l2 regularization
  - Balanced accuracy = **0.786**
  - Unblended acc**: 0.723,** Blended acc**: 0.850**

- **Peak counting**
  - Balanced accuracy = **0.759**
  - Unblended acc**: 0.997,** Blended acc**: 0.522**

- *Binomial uncertainty: 0.003-4*

- *Variability due to random training data selection ~ Bin. unc.*

Topics for further study

- Posterior uncertainties
- Multi-class classification (e.g. 1 vs. 2 vs. >=3)
- Maybe combine GP and peak counting into one better classifier
- Galaxy localization
- Incorporate multiple bands

Lawrence Livermore National Laboratory