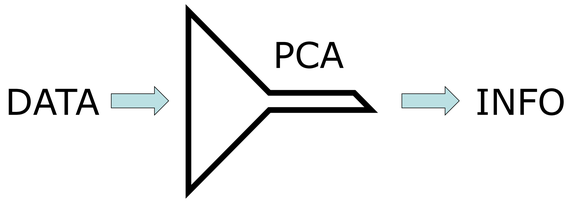
Parallelized Incremental PCA for Online Data Visualization

Callum Hepworth[1], Ariana Peck[1], Frédéric Poitevin[1]

[1]Linac Coherent Light Source (LCLS), SLAC

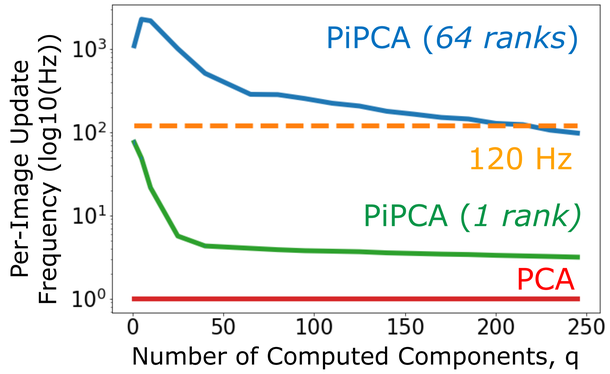
Introduction

Principal component analysis (PCA) is a lightweight exploratory data analysis technique that isolates the most significant modes of variation within a dataset.

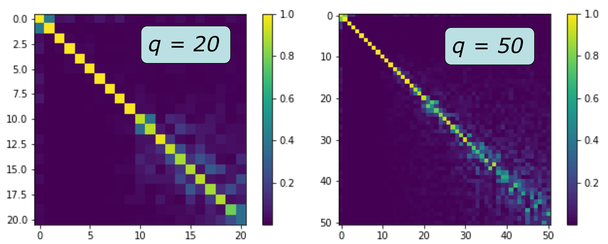


***Objective*:** Implement an online **PCA** plugin for the existing LCLS framework *'btx'* to be used for hit-detection, denoising, and real-time data exploration.

Results



PiPCA update frequency **exceeds** the 120Hz image arrival window for q < ~200 (amortized per-image update) when running over 64 ranks.



Confusion Matrix of $\boldsymbol{U}_q$ calculated using PiPCA and PCA

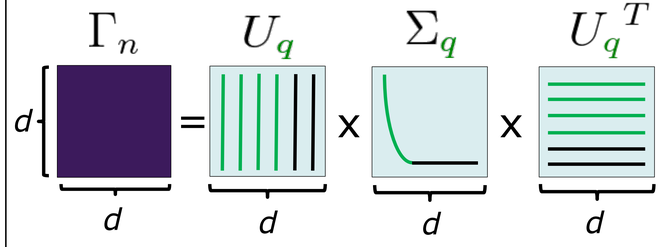
PiPCA accurately calculates and updates the first q / 2 principal axes and values, maintaining comparable explained variance to lossless PCA.

PCA

Given a *(d x n)* dataset $\boldsymbol{X}$ with *d* features and *n* samples, PCA finds the rank-*q* < *d* basis $\boldsymbol{U}_q$ that minimizes the **compression loss**:

$$L_n(\boldsymbol{U}_q) = \frac{1}{n}\|\boldsymbol{X} - \boldsymbol{U}_q\boldsymbol{U}_q^T\boldsymbol{X}\|_F^2$$

$\boldsymbol{U}_q$ and its corresponding $\boldsymbol{\Sigma}_q$ are formed from the *q* most significant eigenvectors and eigenvalues of the sample covariance matrix $\boldsymbol{\Gamma}_n = \boldsymbol{X}\boldsymbol{X}^T / (n-1)$:

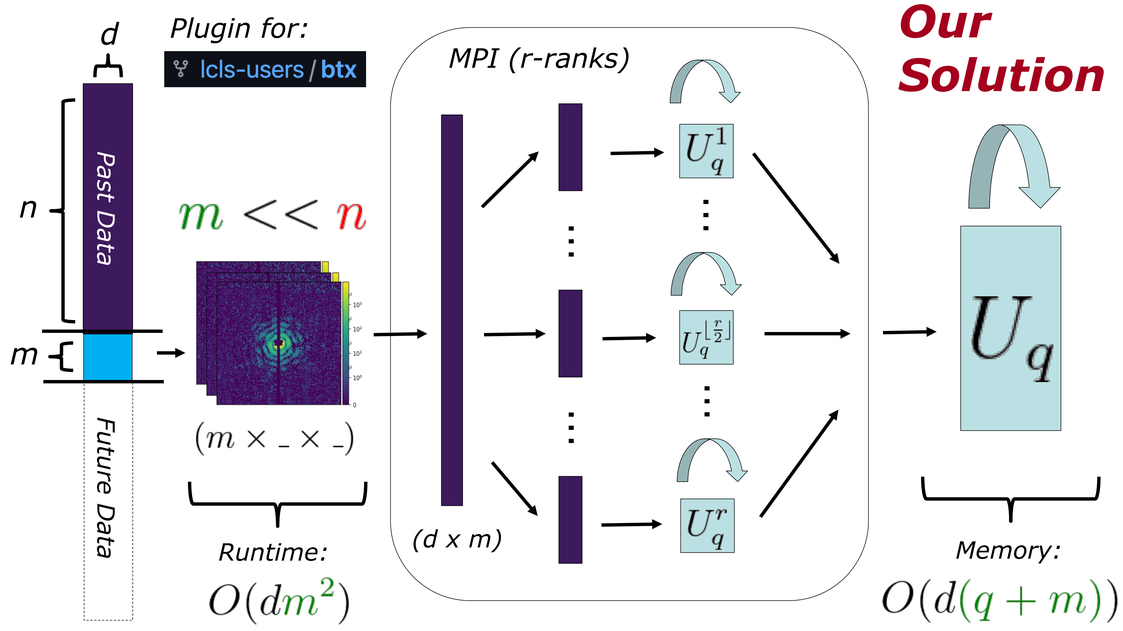


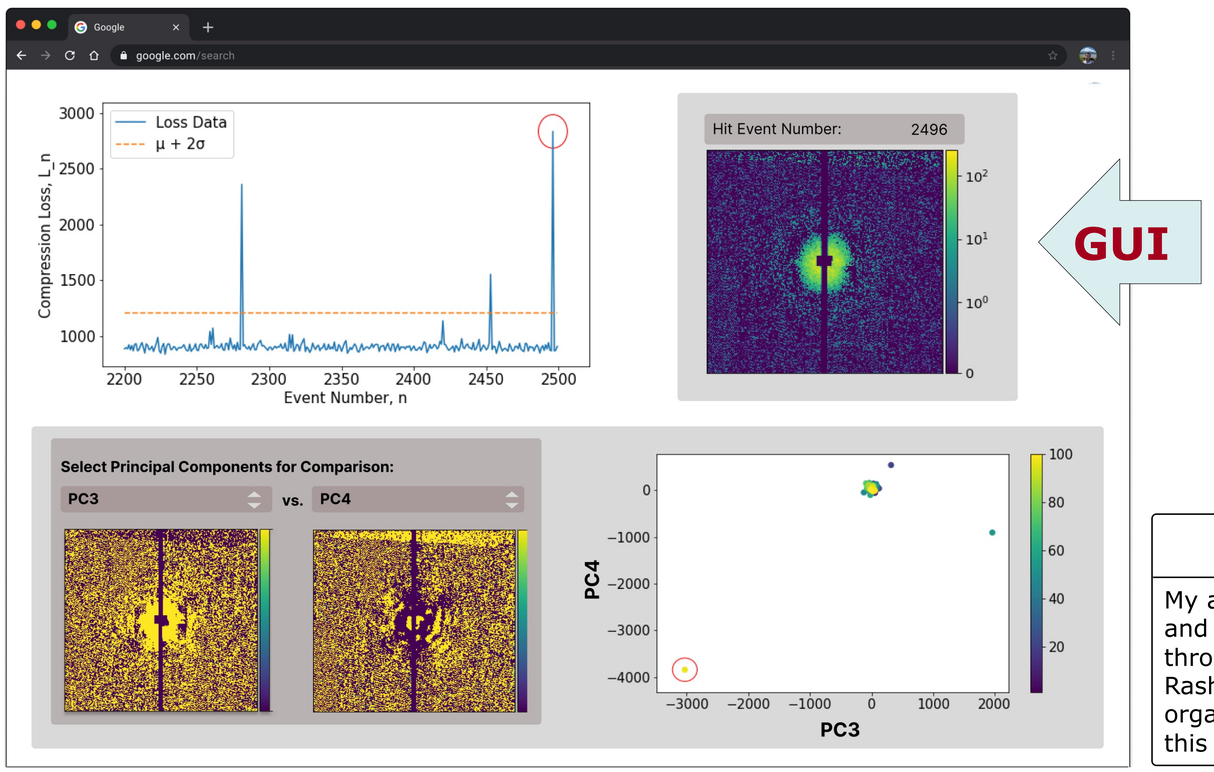
Runtime & Memory Complexity:

$O(dn^2) \sim 1e12$

Scaling Problem

Parallelized Incremental PCA (PiPCA)





Future Enhancements

Areas of enhancement and exploration include:

- Adaptive tuning of *r, q,* and *m.*
- Implementation of an interactive **GUI** for real-time data visualization.
- Research into interfacing with GPU, instead of CPU, clusters.
- Further performance improvements, including:
  - data cropping and masking
  - smart model training (pausing model updates within a stability window).

Acknowledgments

My appreciation and thanks go out to Frédéric Poitevin and Ariana Peck for their guidance and mentorship throughout this project, and to Andy Aquila, Alan Fry, Rasheida Knight, and the LCLS summer internship organizers for the opportunity to be at SLAC in person this summer!