Exploring Mathematical Strategies for Finding Hidden Features in Multi-Dimensional Big Datasets

Tri Duong[1], Fang Ren[2], Apurva Mehta[2]

[1] University of Houston, Houston, TX 77004, USA.

[2] Stanford Synchrotron Radiation Lightsources, SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA.

1. Objectives

The primary goal of this project is to develop a new algorithm using recent advances in image processing, machine learning techniques, and employing different types of distance metrics to a large amount of diffraction data collected at a synchrotron beamline in high-throughput experimentation.

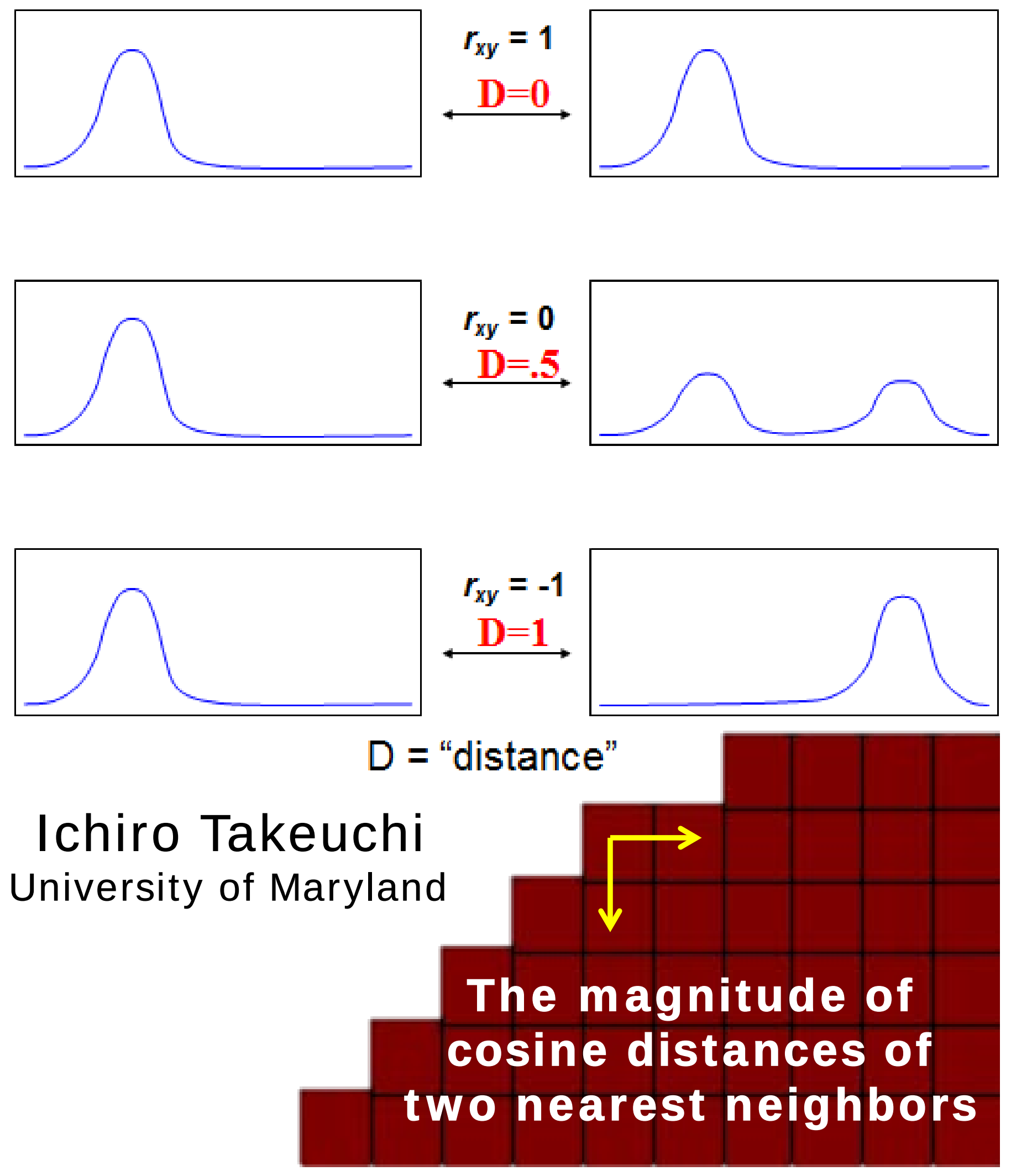
The new algorithm enables analysis and extraction of hidden features from a large multi-dimensional dataset on-the-fly and with minimal computational cost and human intervention. When the algorithm is performed on a large number of x-ray diffraction patterns, the algorithm can be used to find the **structural phase boundaries** leading to the discovery of the composition-structure relationship, which is often an end goal of many material science experiments.

3. Methods

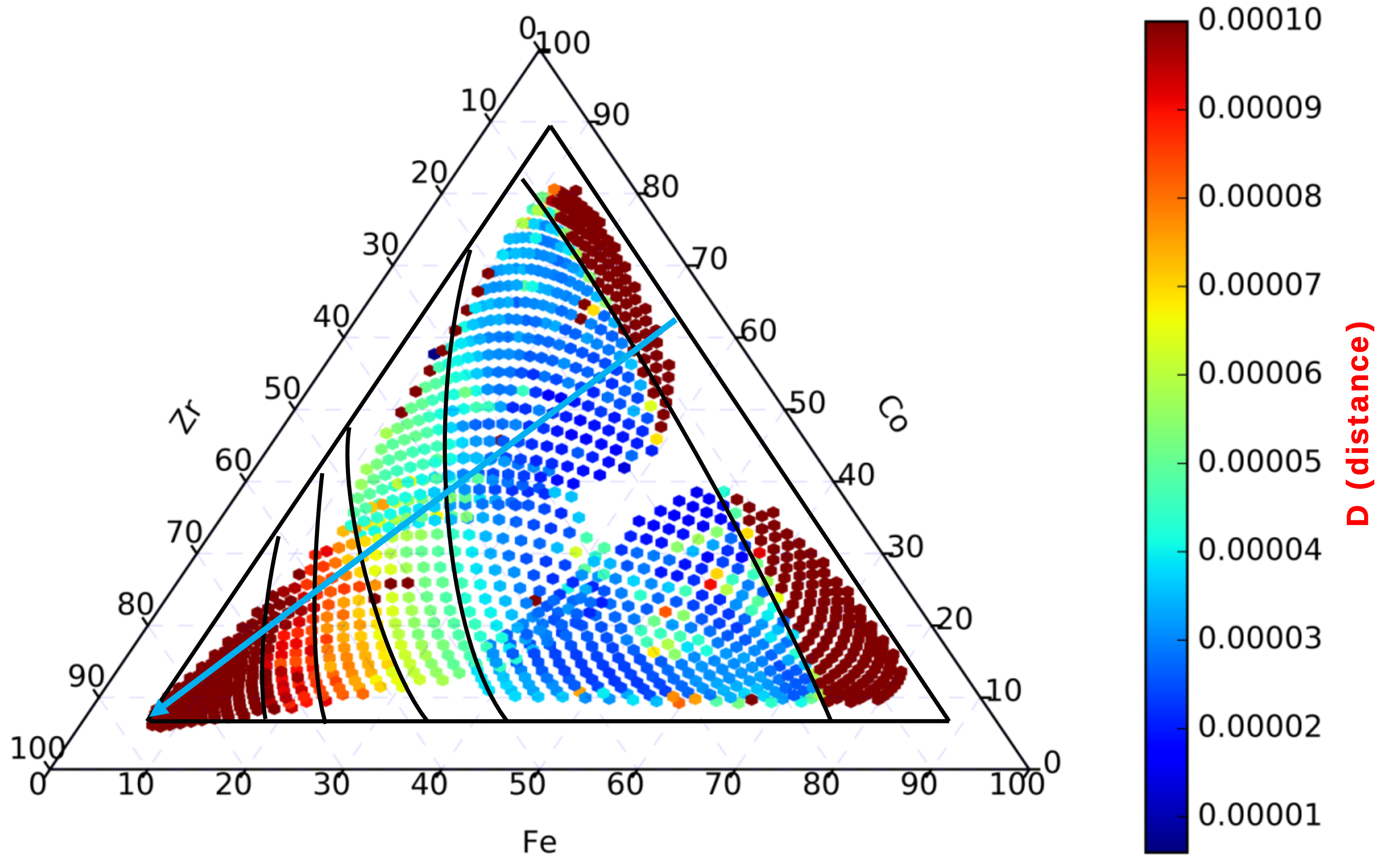
We treat each XRD spectra as a vector, and employed Euclidean Distance, Manhattan Distance, and Cosine Distance to compare each vector. We determined that Cosine distance was the best fit for our XRD spectra.

We found that it is only necessary to **compare neighbors adjacent to one another and underneath one another**, on the sputtering wafer. Then, we calculated the magnitude of both comparisons in order to identify the phase boundaries. By comparing the neighbors, we can see that along phase boundaries, XRD spectra changed dramatically.

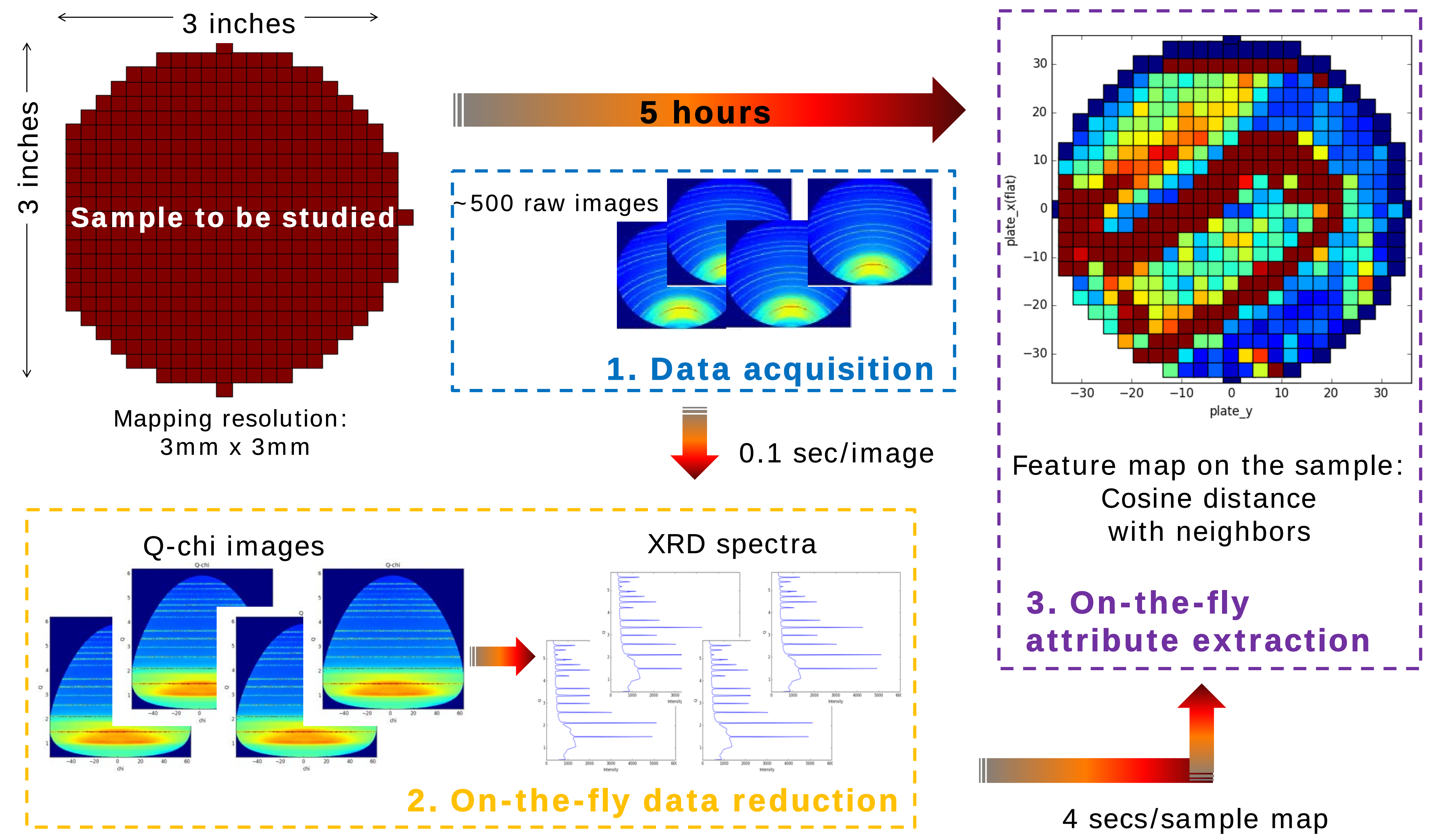
Comparing XRD patterns



5. Phase Boundaries Ternary Plot of Co, Fe, and Zr using Cosine Distance

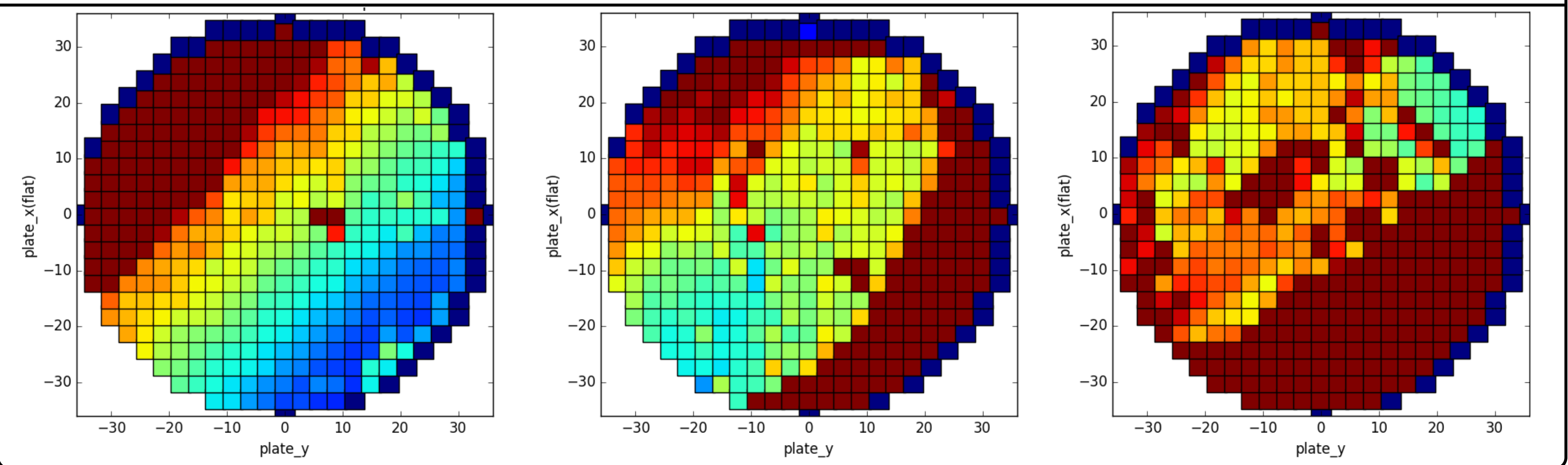


2. HiTp XRD Analysis Capability at SSRL BL1-5



Challenge: Data analysis – how do we identify phase?

4. Phase Boundaries Maps of Co, Fe, and Zr using Cosine Distance



6. Results and Discussions

The computational time of the algorithm only takes about 3 to 5 seconds for each phase boundary map and can be run on-the-fly with limited human interaction and cheap computational cost. It even works with larger datasets and only grows linearly, unlike k-mean clustering, where computational would increase as factorial.

The phase boundaries in Co-Fe-Zr ternary were identified using a spectrum of tools combining domain knowledge and unsupervised machine learning algorithm, both of which are important to find the "ground truth" of phases in ternary. The structure changed significantly with Zr composition, but not much by Co:Fe ratio.

7. Acknowledgements

This work and the use of Stanford Synchrotron Radiation Lightsource (SSRL) were supported by United States Department of Energy. This material is based upon work supported by the National Science Foundation through the Robert Noyce Teacher Scholarship Program under grant#1546150. Any opinions, finding, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The research was made possible by the California State University STEM Teacher Researcher Program.