doi:10.1017/S1431927617001684
Microsc. Microanal. 23 (Suppl 1), 2017

Processing a Five Dimensional X-ray Image: Big Data Challenges and Opportunities

Jeffrey M. Davis[1,*], Julia Schmidt[1], Martin Huth[1], Robert Hartman[2], Heike Soltau[1], Lothar Strüder[2,3]

[1.] PNDetector GmbH, Otto-Hahn-Ring 6, 81739 München, Germany
[2.] PNSensor GmbH, Otto-Hahn-Ring 6, 81739 München, Germany
[3.] University of Siegen, Walter-Flex-Strasse 3, 57072 Siegen, Germany

A three dimensional X-ray spectrum image (XSI) can be recorded by scanning a focused electron or X-ray beam across a two dimensional array of points on a sample and saving an X-ray spectrum at each point. The X-ray spectrum is typically a vector of integers recording the intensity (i.e., "counts") at a particular X-ray energy. Thus, recording one dimensional X-ray spectra across a two dimensional array results in a three dimensional XSI, since each intensity is a function of X, Y and energy. The sizes of early XSIs were limited by the X-ray detector hardware and the practical limitations of hard drive space on personal computers, but modern instrumentation and computers can easily produce XSIs of 10 GB or more [1].

Even larger data sets are possible if, rather than recording an X-ray spectrum at each point, a full XSI is recorded at each point in a two dimensional stage scanning array on a sample. This five dimensional imaging is made possible through the use of an imaging spectrometer, such as the pnCCD based Color X-ray Camera (CXC) [2]. The CXC records the position (with a position resolution down to $3\,\mu\mathrm{m}$) and energy (with an energy resolution of approximately 145 eV at Mn Kα) of each X-ray event on the detector. From a single analysis, the full XSI can be recorded with variables of (X_{camera}, Y_{camera}, energy). The final five dimensional data can be constructed by recording an array of XSIs, where each intensity is a function of (X_{sample}, Y_{sample}, X_{camera}, Y_{camera}, energy), as shown in Figure 1.

One of the major challenges of this imaging mode is that the resulting data often requires several terabytes of hard disk space. It is tempting, therefore, to significantly reduce the data by simply summing along three of the dependent axes (e.g., X_{camera}, Y_{camera}, energy), resulting in a two dimensional image. But this strategy can reduce or eliminate the information content of the data. The focus of this work is to demonstrate how some "Big Data" techniques can be applied to these enormous X-ray data sets. Here, a particular emphasis has been placed on dimensionality reduction such that the maximum amount of information is preserved. This technique also makes free use of noise reduction, missing data imputation and machine learning, as provided in the Anaconda data analysis environment from Continuum Analytics [3].

An example of "Big Data" style dimensionality reduction is demonstrated in Figure 1. A sample of manganese doped lead zirconate titanate (PMZT) was scanned with a collimated X-ray beam to create a five dimensional X-ray image as described above. First, the data are summed along the energy axis to create a four dimensional data set (i.e., an array of images). A region of interest is selected on the images and the counts in that region are summed and plotted. The result reveals structures on the sample. In the second case, the data are summed along the X_{camera} axis. This image contains both diffraction information (seen as curves in the image) and fluorescence (seen as vertical lines). Using Bragg's law, the curves on the image can be modeled and removed, and the missing fluorescence data imputed using various methods. This process is highlighted in Figure 2.

References:

[1] Newbury, D.E., and Ritchie, N.W.M., Elemental mapping of microsctructures by scanning electron microscopy-energy dispersive X-ray spectrometry (SEM-EDS): extraordinary advances with the silicon drift detector (SDD), Journal of Analytical Atomic Spectrometry **28** (2013) p. 973-988
[2] Scharf, O., Ihle, S., Ordavo, I., *et al*, Compact pnCCD-based X-ray camera with high spatial and energy resolution: A color X-ray camera, Analytical Chemistry, **83** (2011), p. 2532:2538.
[3] Continuum Analytics, *Anaconda Software Distribution*. Computer software Vers. 2-2.4.0 (Nov. 2016) Web. <https://continuum.io>.

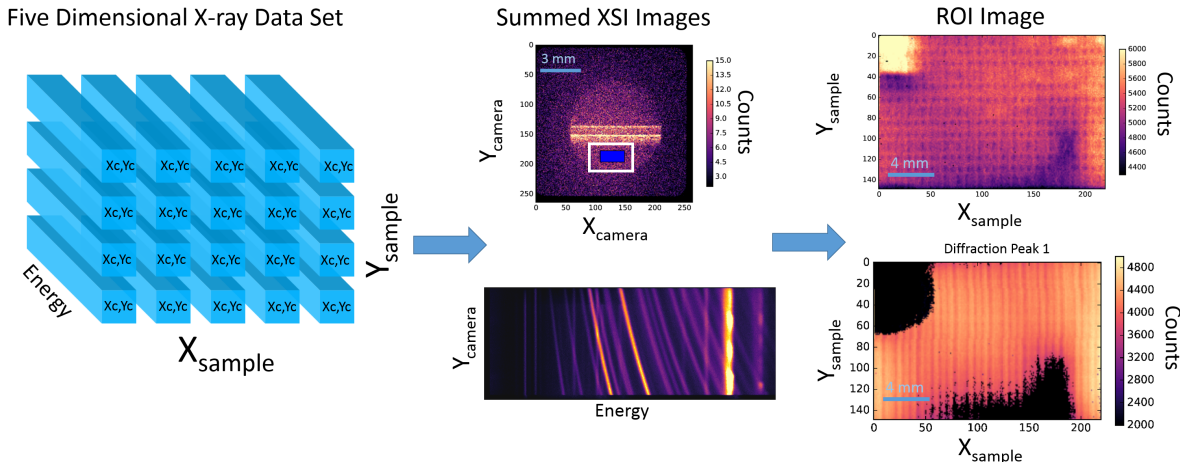

Figure 1: A diagram representing two data reduction strategies. From the original five dimensional data set (left), two projected images are made by summing the individual XSIs along one axis (middle). These images can be further processed to create derived sample images (right) by either summing along a region of interest (top) or modeling the diffraction curves (bottom).

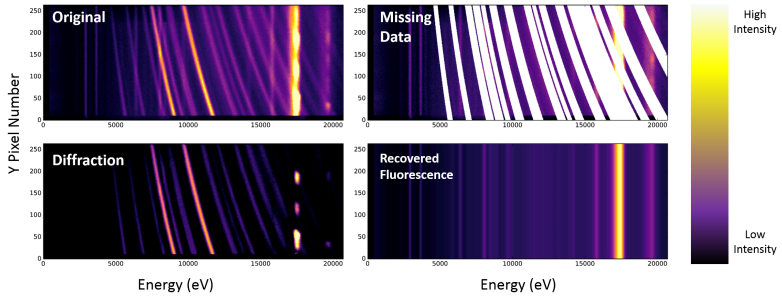

Figure 2: A diagram showing the differentiation of XRD and XRF signals in the CXC. The original image (top left) shows both diffraction and fluorescence signals, which appear as curves and vertical lines, respectively. Braggs law is used to model the diffraction data, but the process of removing them from the original image results in an image with missing data. To recover the fluorescence, an expectation maximization algorithm is used to impute the missing data (bottom right).

https://doi.org/10.1017/S1431927617001684 Published online by Cambridge University Press