

# Knowledge discovery using data mined from Nuclear Magnetic Resonance spectral images

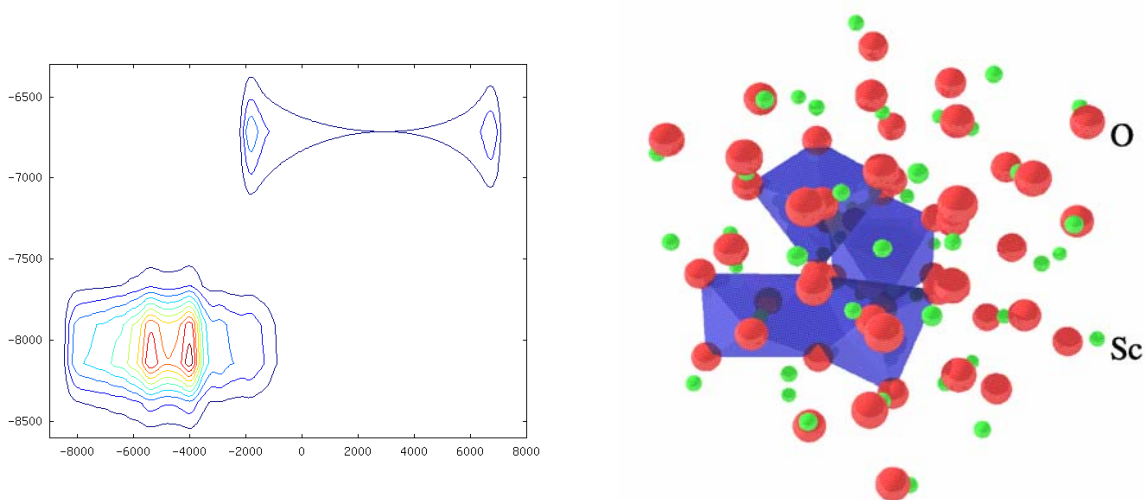
William J Brouwer<sup>1</sup>, Saurabh Kataria<sup>2</sup>, Prasenjit Mitra<sup>2</sup>, Karl Mueller<sup>1</sup>, C. Lee Giles<sup>2</sup>

<sup>1</sup>*Department of Chemistry,* <sup>2</sup>*Department of Information Sciences and Technology,*  
*Pennsylvania State University, University Park PA 16802*

Scientific journal document images are a wealthy yet largely overlooked source of data, suited to extraction and further analysis by researchers. This particular work is devoted to performing accurate data extraction from Nuclear Magnetic Resonance images in documents. The intention is to couple this data with further information mined from text, in order to create a database suitable for machine learning and knowledge discovery.

Researchers in the sciences frequently report data in two dimensional figures, largely for visual impact and posterity. These figures represent significant details, and the extraction of this information is invaluable to researchers wishing to perform data analysis.

Additionally, metadata may be created from images, allowing the figure information to be stored and retrieved in a database. Aspects of the Chem<sub>x</sub>Seer project[1] have been devoted to recognizing and classifying figures of this nature[2,3]. Further, custom applications for data extraction from images have been created, most recently for 2D Nuclear Magnetic Resonance (NMR) spectra. NMR is a valuable local probe used by the spectroscopist to characterize and understand the microscopic structure of materials. Since the early 1950's, NMR spectra have become ubiquitous in the literature, for a large variety of nuclei and materials. There is in addition a vast array of possible experiments that may be performed, with different types of data representation. For the purposes of



**Figure 1:** (left) <sup>45</sup>Sc MQMAS spectra for scandium oxide and (right) corresponding atomic structure displaying scandium in an octahedral oxygen environment

this work we restrict attention to 2D NMR spectra, where intensity is represented as contour levels and  $x$  and  $y$  axes represent two frequency dimensions. Nuclei studied in the liquid state have relatively simple spectra in that frequencies may be ‘picked’ using the spectral center of mass alone. Nuclei studied in the solid state are generally quite complicated, suffering the effects of anisotropic broadening. In terms of information extraction, in the former case kernel density estimation (KDE) combined with methods developed by the group are sufficient. Solid state spectra (for instance high resolution MQMAS, figure 1) would require some means of topological simplification[4], in order to be useful from a machine learning perspective.

However in both cases, there is the great potential of combining the spectral information with further data extracted from document text, in order to perform structural prediction using machine learning. Since the earliest days of NMR, empirical relationships have been established between variables measured via NMR and microscopic details such as atomic coordination, bond lengths and angles as well as frequencies of microscopic motion. If information from NMR images and document text may be extracted in parallel and effectively stored, a growing database of this nature could be used in conjunction with machine learning to provide unprecedented insight into atomic structure.

Initial work is being devoted to testing the accuracy of information extraction from 2D NMR spectra, as well as investigating means for topological simplification of solid state NMR spectra. Supervised machine learning is being conducted, training using input spectral data extracted from images, coupled with structural details extracted from document text.

[1] **Chem<sub>x</sub>Seer: A Chemistry Web Portal for Scientific Literature and Datasets**, Bolelli, L., Lu, X., Liu, Y., Jaiswal, A., Bai, K., Councill, I., Mitra, P., Wang, J.Z., Mueller, K., Kubicki, J., Garrison, B., Bandstra J., Giles, C.L. *Open Repositories Conference 2007*

[2] **Automatic Information Extraction from 2-Dimensional Plots in Digital Documents**, William Brouwer, Saurabh Kataria, Sujatha Das, Prasenjit Mitra and C. Lee Giles, *Joint Conference on Digital Libraries 2008 (JCDL-08)*

[3] **Automatic Extraction of Data Points and Text Blocks from 2-Dimensional Plots in Digital Documents**, Saurabh Kataria, William Brouwer, Prasenjit Mitra, C. Lee Giles, *Twenty-Third AAAI Conference on Artificial Intelligence (AAAI-08)*

[4] **Topological Persistence and Simplification**, Herbert Edelsbrunner, David Letscher, Afra Zomorodian, *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*