Abstract Submitted
for the MAR14 Meeting of
The American Physical Society

Finding Structure in the ArXiv ALEXANDER ALEMI, RICKY CHACHRA, PAUL GINSPARG, JAMES SETHNA, Cornell University — We applied machine learning techniques to the full text of the arXiv articles and report a meaningful low-dimensional representation of this big dataset. Using Google's open source implementation of the continuous skip-gram model, word2vec, the vocabulary used in scientific articles is mapped to a Euclidean vector space that preserves semantic and syntactic relationships between words. This allowed us to develop techniques for automatically characterizing articles, finding similar articles and authors, and segmenting articles into their relevant sections, among other applications.

Alexander Alemi
Cornell University

Date submitted: 07 Nov 2013 Electronic form version 1.4