## Fundamentals in Data Science: Active Deep Learning-Based Annotation of Electroencephalography Reports for Patient Cohort Identification

## Sanda Harabagiu, Professor, Department of Computer Science, University of Texas at Dallas

## Abstract:

### Electronic medical records (EMRs) collected at every hospital in the country contain a staggering wealth of biomedical knowledge. EMRs can include unstructured text, temporally constrained measurements (e.g., vital signs), multichannel signal data (e.g., EEGs), and image data (e.g., MRIs). This information could be transformative if properly harnessed. When processing the clinical text from EMRs, state-of-the-art natural language processing (NLP) methods need to consider new forms of deep learning to face the challenges posed by the complexity and volume of Big Data repositories of medical records. By combining the advantages of active and deep learning, we show how a large collection of electroencephalography (EEG) reports can be annotated to capture a variety of clinical concepts and their attributes. The automatic identification of EEG-specific medical concepts allows us to produce a novel representation of EEG knowledge, namely the EEG-Qualified Medical Knowledge Graph (EEG-QMKG), which is generated using BigData solutions through MapReduce. Because we are representing the EEG-QMKG as a probabilistic graphical model, we shall be able to use this probabilistic knowledge in many applications, including (a) medical inference, e.g. estimation of likelihoods of clinical correlations, interpretations of the EEG tests or (b) discovery of patient cohorts. When using the EEG-QMKG for patient cohort discovery, we rely on a novel index that captures not only information discerned from the EEG reports, but also information available in the EEG signal recording. This multi-modal index is accessed by relevance models that decide which patients satisfy the inclusion/exclusion criteria of a clinical study.

## About The Speaker:

Sanda Harabagiu is a Professor of Computer Science and the Erik Jonsson School Research Initiation Chair at the University of Texas at Dallas. She is also the Director of the Human Language Technology Research Institute at University of Texas at Dallas. She received the Ph.D. degree in Computer Engineering from the University of Southern California in 1997 and a Ph.D. in Computer Science from the University of Rome, “Tor Vergata”, Italy in 1994. She is a past recipient of the National Science Foundation Faculty Early CAREER Development Award for studying coreference resolution. Her research interests include Natural Language Processing, Information Retrieval, Knowledge Processing, Artificial Intelligence and more recently Medical Informatics. She has been interested for a long time in Textual Question-Answering, reference resolution and textual cohesion and coherence. In 2006 she co-edited a book entitled “Advances in Open Domain Question Answering”. Prof. Harabagiu is a member of AMIA, AAAI, IEEE and ACM. See [www.hlt.utdallas.edu/~sanda](http://www.hlt.utdallas.edu/~sanda) to learn more about her research and teaching.

Sanda Harabagiu is a co-PI on an NIH BD2K grant titled “Automatic discovery and processing of EEG cohorts from clinical records” which is a collaboration between Temple University and the University of Texas at Dallas.