White paper for NSF Award 1443061: "CIF21 DIBBs: Systematic Data-Driven Analysis and Tools for Spatiotemporall Solar Astronomy Data"

PI: Dr. Rafal A. Angryk (GSU/Computer Science)
Co-PI's: Dr. Petrus C Martens (GSU/Solar Physics) & Dr. Katherine Reeves (SAO/Solar Physics)

Ours is a **truly interdisciplinary project**, with intense collaboration between computer scientists and solar physicists. We are very thankful to NSF for the DIBBs award. In part, because of this award, the "Solar-Stellar Informatics Cluster" at GSU, led by Drs. Angryk and Martens, has recently received strong support from the GSU administration in being awarded eight new academic positions (three tenure track professorships, three research professors with substantial support, and two postdoc lines). The research cluster has also been awarded its own office space, on the same floor as the GSU astronomy group, and one floor below the computer science department. We insisted on this proximity to maintain the close integration with both departments and to provide truly interdisciplinary experience for our students.

Angryk and Martens have collaborated for well over a decade now in trying to apply cutting edge big data analytics approaches to the wealth of solar astronomy data from both ground- and space-based observatories. From the onset it was well understood that such a collaboration could only be successful if the research effort was fully integrated, and well-balanced with both disciplines bringing a complimentary set of skills and research ideas to the table. This we have achieved through our decade old **"Data Mining Lab", which has as its members all students, graduate and undergraduate, postdocs and research scientists that work under our direction, regularly sharing offices with personnel from other disciplines, and co-authoring papers**. The group meets weekly for presentations in both disciplines, with a focus on integrated data-driven research projects. In addition we discuss recent developments in both disciplines, such as new techniques for the prediction of solar activity, and new solar observing data and metadata becoming available. One-on-one collaborations between computer scientists and solar physicists are initiated during these meetings and later reported on.

Such an **integrated research approach** is in our view absolutely essential for obtaining credible and significant results. The literature is full of examples of data analytics performed on solar astronomy data as a one-sided approach. By that we mean either solar physicists applying machine learning methodologies to their data, that they well understand, but without much knowledge of machine learning approaches, or computer scientists applying their sophisticated data mining methods on solar data of which they do not understand the inherent biases and other limitations. In both cases the research outcomes are sub-optimal at best, often hard to reproduce and/or further utilize, and sometimes just plain misleading. Our Data Mining lab is specifically set up to avoid these pitfalls, and we **think we have succeeded in meeting that key challenge.**

A remaining and ongoing challenge that we are working on under this grant is quality control of the enormous data and metadata sets that solar observatories currently deliver, systematic verification of solar events reported by computer vision software modules as well as tracking of them. We also work hard toward employing our novel spatio-temporal data mining techniques on all of the big solar astronomy data and reporting new discoveries to the scientific communities in the areas of solar physics and computer science. We have made significant progress on all these fronts mainly thanks to the support from this NSF grant.

Our future direction is **to obtain the best possible predictions for solar flares, Coronal Mass Ejections, and Solar Energetic Particle storms,** using data from all current solar observatories, and recent cutting edge Spatio-temporal Big Data Analytics methodologies. Space weather prediction has become a national priority through a recent executive order by president Obama, and our NSF-funded work provides us with unique preparation to respond to this order.