

## Education

- 2004 - 2008 | PhD in Computer Science with minor in Statistics, Cornell University. Advised by Rich Caruana.
- 1998 - 2003 | Diplom with Honors in Applied Mathematics and Computer Science  
Department of Computational Mathematics and Cybernetics, Moscow State University

## Research interests

Data mining and machine learning algorithms.

- Domain knowledge extraction: feature ranking/selection, variable interaction detection, visualization of feature effects.
- Algorithms involving decision/regression trees, in particular, ensembles - methods to combine different models to improve overall predictive performance.
- Algorithms based on modelling additive structure such as additive models, logistic/linear regression, gradient boosting.
- Applications to large real-world data sets.

## Research experience

- from 2009* | Postdoctorate Fellow at Auton Lab, Carnegie Mellon University. Primarily worked in a joint project with USDA and SAIC on Salmonella risk prediction at meat-processing establishments. Investigated different approaches for feature selection in logistic regression in the case when data is limited and has high variance. Also participated in a number of other Auton Lab projects including nuclear threat detection and development of public health information system.
- 2005 - 2008 | PhD thesis research in Cornell University: "Modelling Additive Structure and Detecting Interactions with Groves of Trees." Joint work with Rich Caruana and Mirek Riedewald. Invented a new type of ensembles for regression and classification, Additive and Gradient Groves. These algorithms are strong prediction models as well as tools for detecting non-additive interactions in large data sets. The application of this work is a part of a joint project with Cornell Lab of Ornithology on tracking environmental change based on birds abundance data.
- summer 2007* | Intern at Google Pittsburgh. Worked with Scott Larsen and Jeremy Kubica on fast large scale feature evaluation in logistic regression.
- 2004 - 2006 | Worked with Johannes Gehrke and Simeon Warner on detection of plagiarism and self-plagiarism in arXiv.org collection of research papers.
- summer 2005* | Intern at Fraunhofer IPSI, Darmstadt, Germany. Internship project: "Scalable Approximate Spectral Clustering." Joint work with Ulrike von Luxburg and Thomas Hofmann.
- 2003 - 2004 | Junior researcher at the Institute for Information Transmission Problems of the Russian Academy of Sciences. Worked on GeoTime project: geoinformation system for detection of earthquake precursors.

- 2001 - 2003 | Diplom thesis research at Moscow State University under the guidance of Mikhail Petrovskiy. Created a modification of the fuzzy decision tree algorithm for the classification of objects described as multidimensional datacubes.
- 2000 - 2001 | Worked on S5M project in the Laboratory of Programming Technologies, Moscow State University. Designed and implemented disassembler and a part of interpreter for a cross-platform programming system for S5M, a task-oriented processor used in spacecrafts.

## Awards/Honours

- 2009 | Third place in the Supervised Learning Challenge at ICDM Data Mining Contest
- 2007 - 2008 | Supported by a fellowship from the Leon Levy Foundation.
- 2007 | Best Student Paper award at European Conference on Machine Learning (ECML'07)  
D. Sorokina, R. Caruana, M. Riedewald, "Additive Groves of Regression Trees."

## Open Source Software

- 2009 | **TreeExtra** package — a set of command line tools implementing algorithms such as Additive Groves for regression and classification, multiple counts feature evaluation, interaction detection with Additive Groves and a number of other supplemental tools. Implemented in C++ / STL. Code, binaries and manuals are available at
- [www.cs.cmu.edu/~daria/TreeExtra.htm](http://www.cs.cmu.edu/~daria/TreeExtra.htm)
- TreeExtra is currently used in a number of projects both in academia and industry.

## Publications

- 2009 | *Daria Sorokina, Rich Caruana, Mirek Riedewald, Wesley M. Hochachka, Steve Kelling*  
Detecting and Interpreting Variable Interactions in Observational Ornithology Data. In proceedings of the ICDM Workshop on Domain Driven Data Mining (DDDM'09).
- 2009 | *Lujie Chen, Artur Dubrawski and Daria Sorokina*  
Multivariate Analysis for Predicting Risk of Microbial Contamination of Food. In proceedings of 8th Annual International Society for Disease Surveillance Conference (ISDS'09).
- 2009 | *Daria Sorokina*  
Application of Additive Groves Ensemble with Multiple Counts Feature Evaluation to KDD Cup'09 Small Data Set. In JMLR Workshop and Conference Proceedings vol. 7: proceedings of KDD Cup'09 competition.
- 2009 | *Sameer Singh, Jeremy Kubica, Scott Larsen, Daria Sorokina*  
Parallel Large Scale Feature Selection for Logistic Regression. In proceedings of SIAM International Conference on Data Mining (SDM'09).
- 2008 | *Daria Sorokina, Rich Caruana, Mirek Riedewald, Daniel Fink*  
Detecting Statistical Interactions with Additive Groves of Trees. In proceedings of 25th International Conference on Machine Learning (ICML'08).
- 2007 | *Daria Sorokina, Rich Caruana, Mirek Riedewald*  
Additive Groves of Regression Trees. In proceedings of the 18th European Conference on Machine Learning (ECML'07). (**Best Student Paper award.**)

- 2007 | *W. Hochachka, R. Caruana, A. Munson, M. Riedewald, D. Sorokina, D. Fink, S. Kelling*  
Data-Mining Discovery of Pattern and Process in Ecological Systems. *Journal of Wildlife Management*, 71(7): 2427–2437
- 2006 | *Daria Sorokina, Johannes Gehrke, Simeon Warner, Paul Ginsparg*  
Plagiarism Detection in arXiv. In proceedings of the 6th IEEE International Conference on Data Mining (ICDM'06)
- 2006 | *R. Caruana, M. Elhawary, A. Munson, M. Riedewald, D. Sorokina, D. Fink, W. Hochachka, S. Kelling*  
Mining Citizen Science Data to Predict Prevalence of Wild Bird Species. In proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)

## Technical Reports

- 2006 | *Daria Sorokina, Johannes Gehrke, Simeon Warner, Paul Ginsparg*  
Plagiarism Detection in arXiv. Technical Report TR2006-2046, Computing and Information Sciences, Cornell University. (Full version of ICDM'06 paper.)
- 2003 | *Daria Sorokina, Mikhail Petrovskiy*  
Adaptation of Fuzzy Decision Tree Algorithm for Application in Multidimensional Databases. *Collected Articles on Software Systems and Tools*, CMC MSU publishing, Moscow, Russia
- 2003 | *Daria Erofeyeva*  
Fuzzy Approach to Classification for Multidimensional Datacubes. Diplom thesis, Moscow State University

## Technical skills

| C++, STL, MatLab

## Teaching Experience

- Fall 2007* | Teaching assistant for CS 578 (Empirical Methods in Machine Learning and Data Mining), Cornell University
- Fall 2004* | Teaching assistant for CS 212 (Java Practicum), Cornell University