Preface

The Finnish Centre of Excellence in Algorithmic Data Analysis Research (Algodan) has completed its six year term. This report describes the activities of the Centre in 2011-13, and also looks forward into the future of the research groups after Algodan. Moreover, the report presents an extensive bibliometric analysis of the publications produced by the Centre during its total term 2008-13. Research activities are reported according to the teams and their research groups. The groups present their members, mission and main results, cooperation and societal impact as well as a list of most important publications. Complete lists of publications and PhD degrees are given at the end of the report.

I would like to thank all members of the Centre for their efforts over the years to make our Algodan a success.

Helsinki, April 16, 2014

Esko Ukkonen

www.cs.helsinki.fi/research/algodan/
Contents

1. Introduction .......................................................................................................................................... 1
   Summary of Algodan centre as described in the original application .................................................. 1
   Main research themes ............................................................................................................................. 2


3. Funding of ALGODAN in 2008-2013 ....................................................................................................... 4

4. Reports from the Teams and their Groups............................................................................................. 5
   Team Data Mining: Theory and Applications ....................................................................................... 5
   Data mining – theory and applications ................................................................................................. 5
   Parsimonious Modelling Group ............................................................................................................. 8
   Combinatorics, Algebra, and Computing (CO-ALCO) ............................................................................ 11
   Phenomics Group ................................................................................................................................ 14
   Team Combinatorial Pattern Matching ................................................................................................. 16
   Combinatorial pattern matching algorithms and applications .............................................................. 16
   Succinct Data Structures (SuDS) ........................................................................................................... 20
   Practical Algorithms and Data Structures on Strings (PADS) ................................................................. 22
   Computational Geometry .................................................................................................................... 25
   C-BRAHMS Group ............................................................................................................................... 27
   Computational Linguistics Group ........................................................................................................... 30
   Team Link and Pattern Discovery ......................................................................................................... 35
   Discovery Group: Data Mining and Computational Creativity .............................................................. 35
   Team Machine Learning ................................................................................................................... 38
   Machine Learning Group ...................................................................................................................... 38
   Team Neuroinformatics ......................................................................................................................... 41
   Neuroinformatics Group ....................................................................................................................... 41

5. Publications ........................................................................................................................................ 44
   2013 ....................................................................................................................................................... 44
   2012 ....................................................................................................................................................... 49
   2011 ....................................................................................................................................................... 55

6. PhD degrees ........................................................................................................................................ 62

APPENDIX 1: Analysis of ALGODAN Publications 2008-2013 ................................................................. 63
1. Introduction

Summary of Algodan centre as described in the original application

The importance of data analysis in science and in industry is increasing continuously, as our ability to measure and store data grows. While data analysis is as old as science itself, the new methods of collecting raw data pose unprecedented challenges and opportunities to data analysis and to the algorithms of data analysis.

The Algorithmic Data Analysis (Algodan) Centre of Excellence develops new concepts, algorithms, principles, and frameworks for data analysis. The work combines strong basic research in computer science with interdisciplinary work in a wide variety of scientific disciplines and industrial problems.

The research of the Algodan CoE lies in the areas of combinatorial pattern matching, data mining, and machine learning. The work in Algodan is strongly interdisciplinary: we cooperate constantly with application experts in various application areas, formulating novel computational concepts and ways of attacking the scientific and industrial problems of the application areas. Developing new concepts and algorithms is an iterative process consisting of interacting extensively with the application experts, formulating computational concepts, analyzing the properties of the concepts, designing algorithms and analyzing their performance, implementing and experimenting with the algorithms, and applying the results in practice. The main application areas of the Algodan CoE are in biology, medicine, telecommunications, environmental studies, linguistics, and neuroscience.

The formulation of new computational concepts, their analysis, and the design of algorithms are some key ingredients that make the Algodan CoE unique. First, rather than concentrating on improvements to existing problems and methods, the CoE focuses on defining new tasks where significant impact can be made by introducing new concepts. Second, we emphasize the need for analyzing the performance of the algorithms, instead of just relying on heuristic approaches. Third, we use our strong background in algorithmic and probabilistic methods to guarantee that our algorithms perform well both in terms of modelling accuracy and robustness, and in terms of computational complexity and practical efficiency.

The research in Algodan is grouped under four interacting themes: sequence analysis, learning from and mining complex and heterogeneous data, discovery of hidden structure in high-dimensional data, and foundations of algorithmic data analysis. All these themes combine aspects of combinatorial pattern matching, data mining, and machine learning.

The host organizations of the Algodan CoE are University of Helsinki and Aalto University\(^1\). The CoE is in part a continuation of the “From Data to Knowledge” CoE (2002-2007), and consists of about 70 persons. The director of the Algodan CoE is Professor Esko Ukkonen and the vice-director is Vice President, Professor Heikki Mannila\(^2\).

---

1 Until 31\(^{st}\) of December 2009 Helsinki University of Technology.
2 From 1\(^{st}\) of March 2012 Professor Heikki Mannila was appointed as the President of the Academy of Finland.
Main research themes

The main research themes of the Algodan CoE are the following.

- S – Sequence analysis
- L – Learning from and mining structured and heterogeneous data
- D – Discovery of hidden structure in high-dimensional data
- F – Foundations of algorithmic data analysis

There is considerable overlap between the themes: certain algorithmic and probabilistic techniques occur in many themes. In the same way, several themes can be used for a single application. We next describe the themes briefly.

Sequence analysis considers the algorithmic techniques for sequential data. The key methods in the theme are string algorithms, pattern discovery techniques, dynamic programming, and probabilistic modelling. Examples of the algorithmic tasks in the area are approximate string matching, episode discovery, and finding motifs and orders from data. The techniques of sequence analysis have numerous applications in, for example, gene mapping, finding regulatory regions in genomes, telecommunications, linguistics, and paleontology.

Most applications have multiple types of data objects, many different types of data, etc., instead of the classical situation of a single table with observations and variables. Learning from and mining structured and heterogeneous data looks for techniques for data analysis tasks involving such data sets. The methods studied are pattern discovery, prediction of structured objects, the analysis of flows, etc. The applications include biological data analysis, information retrieval, telecommunications, and environmental studies. Algorithmic techniques for probabilistic modelling are crucial in this theme.

The high dimensionality of many datasets causes interesting modelling problems and leads to extremely challenging algorithmic questions. The third theme, discovery of hidden structure in high-dimensional data, looks at how to find latent structure in high-dimensional data sets. The latent structure can be in the form of components, as in independent component analysis, or cluster-like structures, or it can be a parsimonious model giving weight only to a small fraction of the observed variables. The techniques in this theme are based on probabilistic modelling, with a strong algorithmic component.

The theme on foundations of algorithmic data analysis looks at the frameworks of algorithmic data analysis. What can be said about the limitations of pattern discovery? What are the fundamental bounds on the efficiency of string algorithms? What is the computational complexity of fitting probabilistic models of a certain type? Questions such as these abound in algorithmic data analysis, and they are fascinating problems in core computer science.
2. Analysis of ALGODAN Publications 2008-2013

The total number of citations to Algodan publications in 2008-2013 found in Scopus is 2673 and in Publish or Perish (Google Scholar) 6166. The analysis was made by Helsinki University Library bibliometrics team in April 2014 using the publication list provided by the research group. For full report see Appendix 1.

Summary of publication statistics

There are 661 publications listed for 2008-2014. The types and annual statistics are summarized in Table 1.

Table 1: ALGODAN publications by year and type. See explanation below for type classification.

<table>
<thead>
<tr>
<th>Year</th>
<th>A1</th>
<th>A3-A4</th>
<th>B-E</th>
<th>C</th>
<th>F</th>
<th>G</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>40</td>
<td>68</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>124</td>
</tr>
<tr>
<td>2009</td>
<td>34</td>
<td>46</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>97</td>
</tr>
<tr>
<td>2010</td>
<td>34</td>
<td>75</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>129</td>
</tr>
<tr>
<td>2011</td>
<td>31</td>
<td>52</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>92</td>
</tr>
<tr>
<td>2012</td>
<td>29</td>
<td>51</td>
<td>20</td>
<td>0</td>
<td>1</td>
<td>8</td>
<td>109</td>
</tr>
<tr>
<td>2013</td>
<td>32</td>
<td>54</td>
<td>4</td>
<td>0</td>
<td>10</td>
<td>8</td>
<td>108</td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>346</td>
<td>64</td>
<td>1</td>
<td>11</td>
<td>39</td>
<td>661</td>
</tr>
</tbody>
</table>

The publications are classified according to the following scheme:

- **A1**: Articles in refereed scientific journals
- **A3-A4**: Refereed conference articles and articles in edited books
- **B-E**: Technical reports and other publications
- **C**: Books
- **F**: Artistic works
- **G**: Theses

Citation analysis with Scopus

The ALGODAN publications found in Scopus received 2673 citations by beginning of April 2014. A more detailed view for each year is seen in Table 2.

Table 2: Citations in Scopus by year

<table>
<thead>
<tr>
<th>Year</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2008</td>
<td>22</td>
<td>96</td>
<td>162</td>
<td>158</td>
<td>180</td>
<td>133</td>
<td>26</td>
<td>777</td>
</tr>
<tr>
<td>2009</td>
<td>37</td>
<td>152</td>
<td>167</td>
<td>159</td>
<td>146</td>
<td>146</td>
<td>27</td>
<td>688</td>
</tr>
<tr>
<td>2010</td>
<td>2</td>
<td>53</td>
<td>243</td>
<td>263</td>
<td>266</td>
<td>39</td>
<td>39</td>
<td>866</td>
</tr>
<tr>
<td>2011</td>
<td>1</td>
<td>0</td>
<td>8</td>
<td>69</td>
<td>92</td>
<td>13</td>
<td>13</td>
<td>183</td>
</tr>
<tr>
<td>2012</td>
<td></td>
<td></td>
<td>26</td>
<td>80</td>
<td>14</td>
<td>30</td>
<td>8</td>
<td>120</td>
</tr>
<tr>
<td>2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>30</td>
<td>8</td>
<td>38</td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>136</td>
<td>367</td>
<td>576</td>
<td>697</td>
<td>747</td>
<td>127</td>
<td>2673</td>
</tr>
</tbody>
</table>
Citation analysis with Publish or Perish

Citations from Publish or Perish (based on Google Scholar) are listed in Table 3. The quality of the citations has not been controlled in any way.

Table 3: Citations by year and type according to PoP

<table>
<thead>
<tr>
<th>Year</th>
<th>A1</th>
<th>A3-A4</th>
<th>B-E</th>
<th>C</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>1111</td>
<td>752</td>
<td>23</td>
<td>0</td>
<td>1886</td>
</tr>
<tr>
<td>2009</td>
<td>1006</td>
<td>532</td>
<td>9</td>
<td>208</td>
<td>1755</td>
</tr>
<tr>
<td>2010</td>
<td>1102</td>
<td>386</td>
<td>22</td>
<td>0</td>
<td>1510</td>
</tr>
<tr>
<td>2011</td>
<td>270</td>
<td>166</td>
<td>0</td>
<td>0</td>
<td>436</td>
</tr>
<tr>
<td>2012</td>
<td>175</td>
<td>136</td>
<td>25</td>
<td>0</td>
<td>336</td>
</tr>
<tr>
<td>2013</td>
<td>149</td>
<td>65</td>
<td>29</td>
<td>0</td>
<td>243</td>
</tr>
<tr>
<td>Total</td>
<td>3813</td>
<td>2037</td>
<td>108</td>
<td>208</td>
<td>6166</td>
</tr>
</tbody>
</table>

3. Funding of ALGODAN in 2008-2013

Table 4: Funding spent by the unit by year and source

<table>
<thead>
<tr>
<th>Funding agency</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Finland</td>
<td>1 548 000</td>
<td>1 319 000</td>
<td>1 518 000</td>
<td>1 604 000</td>
<td>1 178 000</td>
<td>1 304 000</td>
</tr>
<tr>
<td>Tekes</td>
<td>46 000</td>
<td>223 000</td>
<td>215 000</td>
<td>5 000</td>
<td>108 000</td>
<td>65 000</td>
</tr>
<tr>
<td>EU</td>
<td>362 000</td>
<td>340 000</td>
<td>59 000</td>
<td>69 000</td>
<td>113 000</td>
<td>235 000</td>
</tr>
<tr>
<td>Other</td>
<td>79 000</td>
<td>212 000</td>
<td>241 000</td>
<td>13 000</td>
<td>3 000</td>
<td>0</td>
</tr>
<tr>
<td>Own funding</td>
<td>1 055 000</td>
<td>1 324 000</td>
<td>1 452 000</td>
<td>1 103 000</td>
<td>1 197 000</td>
<td>810 000</td>
</tr>
<tr>
<td>Ministry</td>
<td>2 000</td>
<td>66 000</td>
<td>0</td>
<td>0</td>
<td>25 000</td>
<td>50 000</td>
</tr>
<tr>
<td>TOTAL</td>
<td>3 092 000</td>
<td>3 484 000</td>
<td>3 485 000</td>
<td>2 794 000</td>
<td>2 624 000</td>
<td>2 464 000</td>
</tr>
</tbody>
</table>
4. Reports from the Teams and their Groups

Team Data Mining: Theory and Applications

Data mining – theory and applications

Members
- Kai Puolamäki, PhD, Docent, Group leader (-2012)
- Aristides Gionis, Associate Professor, Group leader (2013-)
- Heikki Mannila, Professor (on leave of absence)
- Panagiotis Papapetrou, PhD, Postdoctoral researcher (-2012)
- Nikolaj Tatti, PhD, Postdoctoral researcher (2013-)
- Michael Mathioudakis, PhD, Postdoctoral researcher (2013-)
- Aleksi Kallio, Part-time doctoral student
- The following people finished their PhD between 2011–2013:
  - Esa Junttila, August 2011
  - Markus Ojala, November 2011
  - Niko Vuokko, February 2012
  - Sami Hanhijärvi, May 2012
  - Jefrey Lijffijt, December 2013

Mission of the group
The Data Mining: Theory and Applications group at Aalto University conducts research on finding local patterns and global models in discrete high-dimensional data. Techniques for this task include both algorithmics in the traditional computer science sense and probabilistic methods. The group was founded by Professor Heikki Mannila who was later appointed the Vice-President of Aalto and then President of the Academy of Finland, and who still contributes to doctoral student supervision in the group. In 2013, Aristides Gionis, new faculty member in Aalto, joined the activities of the group and broadened the research agenda to new areas such as graph mining, social-network analysis, analysis of information networks.

Research activities

S - Sequence analysis
The group has been very active in analysis of sequential data. Research highlights include novel approaches to the problem of discovering episodes in sequential data [2] (continuing on a topic that was introduced by the seminal work of professor Mannila in the 90’s), segmentation algorithms that scale to very large data for a wide family of models [4], methods for discovering surprising sequential patterns [5], and methods for selecting the values of parameters used in sequential pattern mining [7].

In December 2013 Jefrey Lijffijt defended his PhD thesis titled "Computational methods for comparison and exploration of event sequences". The dissertation obtained an honorary grade and it received the award for the "Best doctoral dissertation of 2013" in the Aalto University School of Science.

F - Foundations of algorithmic data analysis
The group is internationally known for contributions in pattern discovery. In the recent years the group continued successfully the research on this area. Highlights include work on sampling methods for finding robust itemsets [1] and methods to evaluate the statistical significance of pattern sets [3].
A - Applications
As the name of the group suggests, there is strong emphasis on applications and multidisciplinary research. This is demonstrated with the active collaboration of the group members with scientists from different domains (biology, paleontology, linguistics), and cross-disciplinary publications [6,8,9].

Future plans
The group is internationally very well known in the field of data mining and knowledge discovery. Additionally, the group has had a substantial impact on application areas – ecology and linguistics being the recent focuses – where the group has introduced new computational concepts and developed methods together with the application area experts.

Prof Heikki Mannila, who established the group, was nominated in 2009 as the Vice President of then newly established Aalto University. Since 2012, Prof Mannila has been serving as the President of the Academy of Finland.

In 2013 Aristides Gionis joined Aalto University as an associate professor, assumed the responsibility for the group, broadened the research agenda, and recruited new students and postdocs. The emphasis of the group in the next years will be in establishing its international reputation, building its collaboration network, recruiting top-notch postdoctoral researchers, and guiding students in research.

Societal, economic, and technical impact
The main societal impact of the work is in the use of the methods in other sciences. This is demonstrated with successful cross-disciplinary publications in the domains of biology, ecology, and linguistics.

There is a high degree of mobility and high rate of renewal. Aristides Gionis joined the group after being a senior research scientist in Yahoo! Research. Nikolaj Tatti returned to Aalto after a successful postdoc position in Belgium, and Michael Mathioudakis obtained his PhD in 2013 from University of Toronto.

The alumni of the group have positioned themselves well in industry and academia. Prof Mannila is the President of the Academy of Finland. Kai Puolamäki is a research scientist at the Finnish Institute of Occupational Health, and Panagiotis Papapetrou obtained an associate professor position in Stockholm University. Recent PhD graduates, Niko Vuokko, Sami Hanhijärvi, Esa Junttila, and Markus Ojala have all obtained attractive positions in the pioneering Helsinki data-science sector.

Cooperation
The group works heavily with other teams of the Algodan center, and there is also a wide international cooperative network, as evidenced by the publication list.

Selected publications


**Parsimonious Modelling Group**

**Members**
- Jaakko Hollmén, Chief Research Scientist, Group leader
- Indrė Žliobaitė, PhD, Postdoctoral researcher (2013- )
- Jesse Read, PhD, Postdoctoral researcher (2013- )
- Mika Sulkava, Academy Postdoctoral Researcher ( -2011)
- Miguel Angel Prada, PhD ( -2010)
- Mikko Korpela, MSc (Tech.), Doctoral student
- Janne Toivola, MSc (Tech.), Doctoral student ( -2013)
- Prem Raj Adhikari, MSc (Tech.), Doctoral student
- Olli-Pekka Rinta-Koski, M.Sc. (Tech.), Doctoral student ( -2013)

**Mission of the group**
The research group Parsimonious Modelling develops novel computational data analysis methods and applies these methods on two application fields: cancer genomics and environmental informatics. Parsimonious modeling aims at simple, compact, or sparse models as a result of learning from data in the presence of very little or no a priori information about the modeled problem. Simplicity of the models facilitates understanding of the problem domain by humans.

**Research activities**

In the area of *cancer genomics*, the research concentrated on the analysis of high throughput microarray data, such as gene expression data and array-based chromosomal genomic hybridization (aCGH) data [13, 16]. A clear emphasis is on the aCGH data measuring gene-specific genomic aberrations, whereas gene expression data has been employed when integrating data sets together in joint analysis scenario. Multi-resolution data has been an active topic of research [1, 14, 15, 19]. Health and wellbeing applications have been surveyed [20].

Methodologically, the research concentrates on biomarker selection problems [13], model selection criteria in search-based feature selection, as well as modeling of multiresolution data [1, 14, 15, 19].

The second research area of the group, *environmental informatics* is understood as the analysis of time series from the natural environment (such as forests, trees, and climate) as well as the man-made, built environment. The analysis of the data from the built environment is strengthened by the acquisition of a research project *TrafficSense* - Energy efficient traffic with crowdsensing.

Projects on the natural environment focused on the forests and their role in the carbon balance [8], environmental monitoring [10], understanding factors behind tree growth [5] and the analysis of proxy time series for climate reconstructions. The man-made environment currently embodies structures, such as buildings and bridges, for instance, which can be equipped with measurement sensors to yield large data bases reflecting health of the structures. The analysis is concerned with identifying or discovering abstract states for the structure and the problem is to detect abnormal states and diagnose faults [9, 11, 12].

Generic methodological research in time-series and sequence analysis has been conducted with other Algodan researchers [6] and with others [7, 17]. The research group has been actively involved in conference organization activities [3, 4].
Publication activity in the group has been very good. The group has hosted several visits between 2011 and 2013. Good balance between applications and methodologies has been achieved.

**Future plans**
Research will be continued in all areas, with more emphasis on the environmental applications. The recent project acquisition of TrafficSense- Energy efficient traffic with crowdsensing will undoubtedly guide the research into transportation related themes. Multiresolution data analysis is now investigated in both cancer genomics and environmental informatics, which could bring synergetic effects.

**Societal, economic, and technical impact**
The research group has taken an active role in conference organization of the IDA and DS conference series for many years. The application oriented research papers contribute to the fields of environmental sciences and transportation research.

**Cooperation**
- Sakari Knuutila, Laboratory of Cytomolecular Genetics (CMG), University of Helsinki, Finland: Joint projects on cancer genomics
- Harri Mäkinen and Pekka Nöjd, Finnish Forest Research Institute, Vantaa, Finland: Joint research on forest growth and proxy time series
- Pertti Hari and Eero Nikinmaa, University of Helsinki, Department of Forestry, Helsinki, Finland: Joint research on forest growth and proxy time series
- Sebastiaan Luysaert and Ivan Janssens, University of Antwerp, Belgium: Joint research on carbon balance and the role of forests
- Dimitrios Gunopulos, University of Athens, Greece: Joint research on sequence analysis, string matching and analysis of the built environment

**Selected publications**


Combinatorics, Algebra, and Computing (CO-ALCO)

Members
- Mikko Koivisto, Academy Research Fellow (8/2008-10/2013), Assistant Professor (1/2013-), Co-leader
- Petteri Kaski, Academy Research Fellow (9/2011-), Associate Professor (1/2012-), Co-leader
- Pekka Parviainen, Doctoral student (-6/2012, PhD 3/2012)
- Janne Korhonen, Doctoral student (PhD 2/2013)
- Juho-Kustaa Kangas, Doctoral student (1/2012-)
- Teppo Niinimäki, Doctoral student

The Phenomics group was merged to the CO-ALCO group in the beginning of 2013.

Mission of the group
The group develops and applies combinatorial and algebraic tools for computational problems, focusing on exact deterministic algorithms. Applications range from fundamental combinatorial problems to computational tasks associated with established probabilistic models in machine learning and data mining.

Research activities

D - discovery of hidden structure in high-dimensional data
We have continued our research on algorithmic foundations of learning graphical models from data. On one hand, we have studied parameterized problems. For special classes of graphical models called polytrees, we showed both tractability and intractability results [6]. For the class of Bayesian networks parameterized by the treewidth (of the moralized graph), we showed an intractability result but also gave a novel dynamic programming algorithm that is practical for small problem instances [8]. On the other hand, we have introduced a partial ordering based framework for learning Bayesian networks (see, e.g., the JMLR article [10] and Parviainen's PhD thesis). We have also studied more efficient ways to learn Bayesian networks using sampling-based estimators [9].

Complementary to learning graphical models, we have pursued parameterized solutions to locate connected motifs in graph data; our randomized algorithms run in time linear in the size of the host graph, and scale exponentially only in the size of the motif [3].

F - Foundations of algorithmic data analysis
We have continued our work on combinatorially flavored variants of zeta and Möbius transforms and on improved time-space tradeoffs for hard combinatorial problems. For example, we showed that every lattice with \( v \) elements, \( n \) of which are nonzero and join-irreducible, has FFT-like arithmetic circuits of size \( O(vn) \) for computing the zeta transform and its inverse, thus enabling fast multiplication in the Möbius [2]. Another highlight concerns the classic subset sum problem: we gave a randomized algorithm that currently yields the best known space-time tradeoff [1]. We have also studied to what extent some recent results obtained using negation (or subtraction) can be achieved with only monotone computation, i.e., without negation: we proved both positive and negative results (see, e.g., ref. 7 and Korhonen's PhD thesis). Very recently, we show that a class of parameterized counting problems can be solved faster than “meet-in-the-middle time” by a combination of fast Möbius inversion and fast rectangular matrix multiplication [4].
**Future plans**
The Phenomics group was merged to the CO-ALCO group in the beginning of 2013.

**Societal, economical and technical impact**
As the group focuses on foundations of algorithmic data analysis, we do not expect to see high societal impact within the next five years. However, we invest substantial efforts to high-risk, high-yield research problems of relatively broad theoretical interest. We expect that some of our results will quickly prove useful for our research community and have high impact in the long run, say within the next fifty years. A specific example of visibility in the broader community is the recent review article in the Communications of the ACM [5].

**Cooperation**
The members of the groups are active also in research projects and study groups initiated by or shared with other groups in Algodan: combinatorial structures in binary data (the Data Mining group); genotype and phenotype analysis (the Phenomics group); causal networks, inference, and discovery (Hyvärinen, Hoyer); local algorithms (Polishchuk).

The current members of the group are all affiliated also with the Helsinki Institute for Information Technology (HIIT). There is active cooperation with other researchers, nationally and internationally:

- P. Austrin; KTH, Sweden; joint publications.
- P. Floréen (J. Suomela), HIIT; local algorithms; joint publications.
- F. Fomin, University of Bergen, Norway; algorithm theory; joint research, manuscript, visits.
- A. Hulpke, Colorado State University, USA; computational algebra; joint publications.
- T. Husfeldt (A. Björklund), Lund University, Sweden & IT University of Copenhagen, Denmark; algorithm theory; joint publications.
- M. Järvisalo; HIIT; relations to satisfiability and constraint satisfaction problems; joint publications.
- J. Nederlof, Utrecht University, the Netherlands; algorithm theory; joint publications, visits.
- S. Szeider, TU Vienna, Austria; parametrized algorithms and complexity in the context of probabilistic models; joint publications, visits.
- P. R. J. Östergård, Aalto University, Helsinki; combinatorics; joint publications.

**Selected publications**


**Phenomics Group**

**Members**
- Heikki Mannila, Professor, Group leader (-2/2012)
- Mikko Koivisto, Academy Research Fellow (8/2008-10/2013), Assistant Professor (1/2013- ), Co-leader
- Pekka Parviainen, Doctoral student (-6/2012, PhD 3/2012)
- Jaana Wessman, Doctoral student (-12/2011, PhD 4/2012)
- Teppo Niinimäki, Doctoral student

The Phenomics group was merged to the CO-ALCO group in the beginning of 2013.

**Mission of the group**
The group develops and applies data mining techniques to identify new phenotypic and genotypic associations in population sample databases.

**Research activities**

**D - Discovery of hidden structure in high-dimensional data**
We have completed a project that aimed at detecting new, biologically more meaningful phenotypic associations using data from the Northern Finland Birth Cohort of 1966 (NFBC66); this is part of the larger Consortium for Neuropsychiatric Phenomics (coordinated by the University of California in Los Angeles). Clustering the subjects according to a set of temperament phenotypes using a mixture model method revealed four coherent clusters and interesting dependencies between the so-called temperament and character inventory subscales; the results are published in PLoS ONE [1, 5]; a thorough description and discussion of the methodology, with application to other data sets are published in the PhD thesis of Jaana Wessman (2012).

We have also progressed on a related theme of analyzing causal and statistical dependencies within and between phenotypes and genotype using Bayesian network models. On one hand, in collaboration with domain experts we have prepared parts of the NFBC66 data for the analyses by pruning, merging, and discretizing variables. On the other hand, we have introduced a novel Markov chain Monte Carlo method that uses our recently developed algorithmic techniques (reported by the CO-ALCO group) to significantly improve the reliability of the network discovery results; we have also studied the robustness of such analyses in the presence of unobserved variables. These methodological results were published at the UAI’11, UAI’12, and ECML-PKDD’11 conferences, respectively [2, 3, 4].

**Future plans**
The Phenomics group was merged to the CO-ALCO group in the beginning of 2013.

**Societal, economical and technical impact**
The data analysis results have direct impact on the hypothesis formation by the domain experts. We expect this further lead to new studies and knowledge that have impact on practices relevant for public health.

Our contributions to data analysis methods more generally are expected to have impact on data analysis in other domains and on further development on the methods. In addition, the methodological issues raise research problems that motivate and steer the research on algorithm theory, especially in our CO-ALCO group.
Cooperation

Within Algodan, the group works in close collaboration with the CO-ALCO group, partly because of the shared researchers (Koivisto and Parviainen). The group was merged to CO-ALCO in the beginning of 2013.

National and international collaborations:

- Center for Neurobehavioral Genetics at the University of California Los Angeles (UCLA), USA; analysis of genotype and phenotype data; joint publications.
- Institute for Molecular Medicine in Finland (FIMM) and National Institute of Health and Welfare (THL); analysis of genotype and phenotype data; joint publications.
- Departments of Psychiatry and of Public Health and General Practice at University of Oulu, Finland; analysis of the NFBC66 data; joint publications.

Selected publications

Team Combinatorial Pattern Matching

Combinatorial pattern matching algorithms and applications

Members

- Esko Ukkonen, Professor, Group leader
- Leena Salmela, Postdoctoral researcher (8/2009 - )
- Emanuele Giaquinta, Postdoctoral researcher (2/2012 - )
- Simon Puglisi, Postdoctoral researcher (1/2012 - )
- Jarkko Toivonen, Doctoral student
- Otto Solin, Doctoral student
- Dominik Kempa, Doctoral student
- Antti Laaksonen, Doctoral student
- Johannes Ylinen, MSc student (2010–2012)

Mission of the group

The group develops theoretical concepts, models, and algorithms for sequence-related problems from biological sequence analysis and other areas. The algorithm-theoretic research is complemented by application-oriented work which is done in close collaboration with many groups of biologists who provide up-to-date problems and new data to be analyzed using the new methods developed in the group.

Research activities

$S$ - Sequence analysis

Prediction of gene regulatory motifs in DNA and de novo sequencing has been the major applied research topics in the group.

The activity of genes is regulated by so-called regulatory modules that are complexes of proteins called transcription factors (TF). Such complexes are formed by TFs that bind to each other and to the DNA on specific TF binding sites. The position weight matrix (PWMs) is the standard probabilistic model for the binding affinity between a transcription factor (TF) and DNA. Accurate PWM models are an essential component of gene regulation models. Our collaborator Jussi Taipale (Univ Helsinki and Karolinska Institute, Sweden) has developed a novel technology, based on the so-called SELEX procedure, for high-throughput sampling of the binding sites of all TFs and of pairs of TFs. SELEX yields very large training data for learning PWMs and more advanced models. We have developed a novel 'multinomial' learning algorithm for PWMs that gives accurate models [2, 14]. We have also developed learning algorithms for finding more accurate models that consist of multiple PWMs for the same TF. The binding affinity of a TF may have several local peaks within the possible binding sites. Each peak has its characteristic sequence that we use as a seed to construct a PWM for that local maximum [22]. Moreover, we have developed new co-operative binding models and learning algorithms for them, to model binding sites of TF complexes that consist of two factors (dimers) [14]. Such dimeric models will be used as building blocks for a general Markov chain model that we are currently developing for chains of binding sites that constitute a putative regulatory module.

In de novo sequencing, the group participated in Professor Ilkka Hanski’s (Metapopulation Research Unit, Department of Ecology and Evolutionary Biology) major project of sequencing of the entire genome of the Glanville fritillary butterfly (Melitaea cinxia); this is the first eukaryote sequencing project in Finland. The genome has been sequenced, assembled and annotated [23]. The assembled draft genome consists of
8.262 scaffolds of total length 390 Mbp. The genome has been compared to other published butterfly genomes and an exceptionally high level of synteny is found in the gene order between the species. The Granville fritillary is the first sequenced butterfly with the ancient karyotype of 31 which also allows us to find patterns in the chromosome fusion events that have occurred in other species. We have developed the data analysis pipeline for high-throughput sequencing, with novel algorithms for error correction [8] and scaffolding [5]. For the most time-consuming step of finding overlaps between sequence segments we use the very fast compressed suffix-array implementation developed by Veli Mäkinen's group. A new metric, normalized N50, and a method for evaluating it is proposed for the quality of a set of contigs or scaffolds produced by a genome assembler [11].

In basic research in algorithms for sequences, the group has produced several results. For example, we proved a practical result about the optimality of a class of string searching algorithms in the average case [9]. We also introduced an alphabet sampling technique to speed up string searching algorithms [10]. In [17, 19] we presented a novel method and fast algorithms to find, in a given DNA sequence, the binding sites of a TF motif described using a generalized Position Weight Matrix that also models dependencies between the motif positions.

\textit{D - Discovery of hidden structure in high-dimensional data}

In a joint project with the Division of Astronomy of the university (Doc Lauri Haikala), we applied Gaussian mixture modeling to locate stellar clusters (potential formation areas of new stars) in the recent data of the United Kingdom Infrared Telescope Infrared Deep Sky Survey. Our search found 137 previously unknown cluster candidates and 30 previously unknown sites of star formation [6]. In the Vía Láctea Survey (VVV) catalogue data the search located 88 previously unknown candidates, most of which are embedded stellar cluster candidates, and 39 previously unknown sites of star formation [7].

\textit{Future plans}

We are working on a few projects whose current status and future plans are as follows:

- Analysis of SELEX data to synthesize models for DNA binding sites of transcription factor complexes (such as dimers); joint work with Jussi Taipale’s group. This work will be utilized in a joint EU-funded project with Jussi Taipale and Lauri Aaltonen on systems biology of colorectal cancer. Our role is to develop computational tools for modeling gene regulatory relations and disorders in them. Several yet unpublished results have been obtained.

- The genome project of the Glanville fritillary butterfly with Ilkka Hanski has almost been finished. The final publications are under writing or review.

- We have started with Acad. Prof. Jukka Jernvall a joint work on the evolution of the gene regulatory structures. We utilize a multiple genome variant of our earlier analysis method EEL (Enhances Element Locator) to find conserved regulatory patterns and their evolutionary relations.

- The research on basic algorithmics on strings will be continued very actively. The Algodan CoE has raised a flourishing community of string algorithmics researchers working at the Department of Computer Science. They are expected to have an internationally leading role in the development of this field.
**Societal, economical and technical impact**

Some of our new algorithms (for the genome assembly [5, 8] or for learning PWMs and their generalizations [14, 22], for example) have potential to become into wide use as they offer improved performance. On application side, for example the new PWM models that are more accurate than the earlier ones [14, 22] have potential for a significant impact in their fields. The new butterfly genome [23] is a major achievement of basic research in its field and may have a strong impact on future research. Some our papers [1, 2, 14] are attracting a rapidly growing number of citations.

**Cooperation**

Cooperation within Algodan: Collaboration and joint publications with Petteri Kaski, Heikki Mannila, and Veli Mäkinen.

Cooperation within University of Helsinki: Joint ongoing or planned projects with professors Lauri Aaltonen (Biomedicum), Ilkka Hanski (Department of Ecology and Evolutionary Biology), Jussi Taipale (Biomedicum & Karolinska Institutet, Sweden), Jukka Jernvall (Institute of Biotechnology).

International cooperation:

- Professor Alberto Apostolico & Dr. Cinzia Pizzi, University of Padua, Italy and Georgia Tech, USA; analysis of motifs in strings; joint publications.

- Collaborations within EU project SYSCOL with several European and US partners; joint publications.

**Selected publications**


Succinct Data Structures (SuDS)

Members
- Veli Mäkinen, Professor, Group leader (−2011)
- Niko Välimäki, Doctoral student (−2012)
- Jouni Sirén, Doctoral student (−2012)

Mission of the group
The study of succinct data structures extends traditional data compression with the functionality preserving property: data structure functions need to be efficiently computable directly from the compressed representation. In addition to providing and analyzing new succinct data structures, the group contributes by engineering open source implementations targeted to applications especially in biological sequence analysis and information retrieval.

Research activities

S - Sequence analysis, S.1 String algorithms /
F - Foundations of algorithmic data analysis, F.1 Theory of string matching

Compressed representations for highly repetitive sequence collections, such as version histories and collections of genomes of individuals within same species, are developed in [1]. This extensive study includes combinations of static cases, dynamic cases, different models to measure high repetitiveness, tradeoffs, and extensions to suffix tree representation. This is the first study beyond the familiar k-th order model in compressed text indexes. In [2] we propose, implement, and experiment a compressed solution for XML indexing. The solution supports XPath query language together with full-text predicates such as prefix, suffix, contains, less-than, etc. In principle, the solution is a carefully designed merge of existing solutions from compressed tree representations and compressed text representations, but it also contains new insights into XPath query evaluation. On an existing benchmark, the new index is faster on all queries than its competitors. Space requirement is better or similar to its competitors. We have recently added the support for XML documents representing a genome annotation database, enabling queries by annotation restrictions (e.g. organism type, gene function, promoter, etc.) and sequence content (PWM matrix and approximate search support). This work with some other improvements by co-authors is now submitted to a journal.

In [3] we extended our previous results on exact substring searches to approximate search, giving some new insights into the timely DNA sequencing read alignment problem. Then we worked on new solutions to the classical de novo fragment assembly problem using our new scalable approach to approximate overlap alignment [4]. Jouni Sirén continued his work on compressed suffix arrays, extending the ideas to longest common prefix arrays [3]. Then later he got in contact with Paolo Ferragina and Rossano Venturini to develop improved suffix array sampling scenarios [7]. Our most recent developments include an extension of Burrows-Wheeler transform to finite automaton representing reference genome together with its common variations among the population [6]. This enables a space-efficient index structure to be constructed to support efficient DNA sequencing read alignment to a rich model of the population.

Future plans
Starting from 2012, the group is no longer part of ALGODAN. See below cooperation section for the reason.
**Societal, economical and technical impact**

Our new developments in [6] are very useful in the variation calling application and led to the collaboration on cancer genetics research (see below).

**Cooperation**

Cooperation within Algodan: Collaboration with the group of Esko Ukkonen on and de novo fragment assembly, with the group of Juha Kärkkäinen on text index construction algorithms, and with Petteri Kaski on space-efficient traversals on huge implicit graphs.

Cooperation within University of Helsinki: Starting from 2012 the group moved to the new Center of Excellence in Cancer Genetics Research, changing the name to genome-scale algorithmics. The new center is led by Professor Lauri Aaltonen (Biomedicum), with whom we had earlier joint work in the analysis of next-generation sequencing data (Riku Katainen from our group moved to Aaltonen's group in 2010). Then we work also with Professor Ilkka Hanski (Department of Ecology and Evolutionary Biology) on the assembly of the genome of Glanville fritillary butterfly (Melitaea cinxia).

International cooperation (during 2010-2012):

- Professor Gonzalo Navarro, University of Chile, Theory of string matching, joint publications, software development, exchange of researchers
- Dr. Johannes Fischer, Karlsruhe Institute of Technology, String mining & compressed suffix trees, joint publications, software development, exchange of visits
- Senior Researcher Sebastian Maneth, NICTA Kensington Research Lab, Sydney, Australia, XML indexing, joint publications, software development, exchange of visits
- Professor Paolo Ferragina, University of Pisa, Theory of string matching, joint publications, exchange of researchers

**Selected publications**

Practical Algorithms and Data Structures on Strings (PADS)

**Members**
- Juha Kärkkäinen, University Researcher, Group leader
- Simon Puglisi, Postdoctoral researcher (4/2012- )
- Dominik Kempa, Doctoral student (7/2011- )
- Pekka Mikkola, MSc student (5/2010 – 8/2012)

**Mission of the group**
The group develops fast and practical algorithms and data structures for fundamental problems arising in sequence analysis. The research is based on thorough understanding of both the combinatorial properties of the problems and the properties of modern computers. The goal is not only to obtain better algorithms but to understand why they are better.

**Research activities**

5. **Sequence analysis**  
F. **Foundations of algorithmic data analysis**

When indexing and analysing massive amounts of sequential data, conventional data structures such as the suffix tree are increasingly being replaced by text indexes based on compressible representations of sequences, particularly the Burrows-Wheeler transform (BWT) and the Lempel-Ziv factorisation (LZF), and context free grammars (CFG). We have studied several aspects related to these new indexes.

1. **Index construction.** We have developed several new algorithms for computing the LZF that are simultaneously faster and more space efficient than previous algorithms [4,8,9]. To deal with ever larger amounts of data, we have recently designed and implemented practical external memory algorithms for computing the LZF [16] as well as the suffix array [17] and the LCP array [19], which are key components in constructing BWT-based text indexes. Of more theoretical interest are in-place algorithms for computing the BWT and its inverse [6].

Some of the subproblems arising in text index construction are interesting in their own right. We have indentified and studied two new string matching variants, longest prefix matching [10] and string range matching [18]. Algorithms developed in these studies are used as a part of the index construction algorithms mentioned above. We have also achieved a big theoretical improvement to the problem of sparse suffix sorting [14] that arises in some suffix array construction algorithms.

2. **Index components.** The FM-index is the best known BWT-based text index and we have developed new techniques for implementing the basic components of the FM-index reducing both the space requirement and the query time [1,15].

3. **Bidirectional indexes.** We have developed compressed text indexes that support extending and contracting the pattern from both ends during a search, which allows a much richer set of pattern matching and pattern discovery operations on the index [11,13].

4. **Document retrieval.** We have developed techniques for using compressed text indexes for document retrieval, which is one of the most important, complex and demanding application of text indexes [7,12].
5. Data compression. We have also developed practical techniques for pure (non-indexing) data compression [2, 3, 5], some of which are included in an experimental, open source compressor available at https://github.com/pjmikkol/bwtc.

Future plans
The group continues to study all aspects of compressed text indexes but particular focus areas are practical techniques for index construction and the basic components of the indexes. For index construction the next direction is handling ever larger amounts of data using external memory, parallel and distributed computing. For basic components, the focus will be on highly repetitive data, where the difficulty is achieving good compression and fast queries simultaneously.

Societal, economical and technical impact
Many of the algorithms and techniques developed by the group are simple and practical and have the potential for being included in many applications. When dealing with massive data, faster and more space efficient techniques can provide substantial economic benefits too.

Cooperation

Cooperation within Algodan and the University of Helsinki:
We have worked on bidirectional indexes and information retrieval with the group of Veli Mäkinen, which left Algodan in 2012 to join the Center of Excellence in Cancer Genetics Research.

National and international cooperation:

- Kalle Karhu, Aalto University, Finland; bidirectional indexes and document retrieval, joint publications.
- Maxime Crochemore and Costas Iliopoulos, King's Collage London, UK; BWT computation; joint publication, research visits.
- Hideo Bannai and Tomohiro I, Kyushu University, Japan; sparse suffix sorting; joint publication, research visits.
- Simon Gog, University of Melbourne, Australia; bidirectional indexes; joint publication, research visit.
- Gonzalo Navarro and Jouni Siren, University of Chile; document retrieval; joint publications.
- Roberto Grossi, University of Pisa, Italy; BWT computation; joint publication.
- Gad M. Landau, University of Haifa, Israel: BWT computation; joint publication.
- German Tischler, The Wellcome Trust Sanger Institute, UK; data compression; joint publication.
- Peter Sanders, Karlsruhe Institute of Technology, Germany; research visit.

Selected publications
Computational Geometry

Members
- Valentin Polishchuk, Academy Postdoctoral Researcher (1/2011-12/2013), Docent (4/2012- ), Group leader
- Mikko Nikkilä, MSc student (3/2012- )
- Mikko Sysikaski, MSc student (6/2010-9/2013)
- Topi Talvitie, MSc student (6/2013- )
- Juha-Antti Isojärvi, MSc student ( -2012)
- Sylvester David Eriksson-Bique, Doctoral student ( -2012)

Mission of the group
Geometric data analysis, visualization and processing are inherent to numerous domains ranging from motion planning to VLSI to geographic information systems to robotics. We design, analyze, and implement computational-geometry algorithms applicable to current and future tasks in intelligent path design, cartography, shape reconstruction and sensor networks.

Research activities
F - Foundations of algorithmic data analysis
We extended classical algorithmic and combinatorial results from discrete network flows to continuous geometric domains; we will continue investigations into the mapping between discrete graph notions and their continuous analogues in geometry, motivated by motion planning and coordination challenges arising in air traffic industry. Our most recent results were presented at ACM SIGSPATIAL GIS'13 [1] and are to appear in ACM/SIGGRAPH Symposium on Computational Geometry 2014 [2].

In our cross-disciplinary work bridging shape modeling and geophysics, we developed novel shape reconstruction methods robust to noise and outliers; our first paper on application of the tools to seismic data analysis has just been accepted to Geophysical Journal International [3]. Our other recent publications on the theme include a paper [4] in IEEE Transactions on Visualization and Computer Graphics -- a top journal in the field.

Sensor network is a source of a bulk of measurements taken by the sensors over time; we look at a variety of computational-geometry questions that arise from processing sensor-network data. Our publications in the area include a paper [5] in MobiHoc -- a top conference in the field.

Future plans
We will continue research in geometric methods for a variety of applications. On the motion planning frontier, we will investigate further fundamental geometric problems of planning paths under a multitude of constraints and requirements. Research on shape approximation will lead to automatic shape reconstruction approaches for seismic data analysis. For sensor networks we will develop efficient data gathering methods.

Societal, economical and technical impact
The research on air traffic motion coordination provides decision-support tools for air traffic industry’s humans-in-the-loop – traffic controllers, dispatchers, managers; given the amount of the world air traffic, even small improvements to the current procedures, even implemented on a local scale, lead to huge savings in operating costs, to decrease in the environmental impact of air traffic, and to increased safety
and efficiency of flight management. The high level of theoretical abstraction pertinent to our algorithmic work allows one to use our results also in other domains – nanostructure design, crowd evacuation, robotics, computer games. Efficient processing of geospatial data representing real-world terrains may enable faster and less costly search-and-rescue operations.

Analyzing geophysical data ultimately leads to better prediction of seismic events. In general, shape approximation and simplification tools are applicable in motion planning, object recognition and data analysis.

Sensor networks are in use for surveillance, monitoring and tracking of objects of very different types – from wildlife to goods in a warehouse; improved algorithms for the networks imply savings in the network management and data handling.

Cooperation

Cooperation within the universities
- Our group is part of the New Paradigms in Computing group at HIIT

International cooperation:
Our group collaborates with researchers all around the world, coauthoring both with academia (Carnegie Mellon, Stony Brook, University of Arizona, Institute of Dynamics of Geospheres, many European colleagues) and industry (IBM, Google, Mathworks, Metron Aviation). We regularly visit our colleagues and host visitors in Helsinki.

Participation in European research networks
- ComplexWorld.eu. Mastering system complexity.
- Toward Higher Levels of Automation in ATM.

Selected publications:
C-BRAHMS Group

Members

- Kjell Lemström, PhD, Docent, Group leader
- Teppo Ahonen, Doctoral student
- Antti Laaksonen, Doctoral student
- Mika Laitinen, MSc student (2010–2011)
- Simo Linkola, MSc student (2011)
- Lari Rasku, MSc student (2012–2013)

Mission of the group
The C-BRAHMS project aims at designing and developing efficient methods for computational problems arising from music comparison, retrieval, and analysis. The main concentration is on retrieving polyphonic music in large-scale music databases containing symbolically encoded music. The project utilizes various algorithmic techniques together with findings in musicology and music psychology to achieve efficient, musically meaningful results.

Research activities

S: Sequence analysis
We have studied and developed new algorithms for analyzing, classifying and retrieving musical sequences. In order to find efficient and effective tools for various tasks, we have used a variety of different modelings of music, similarity measures and methods, with the aim of finding the optimal combination for the given task. For finding occurrences or recurrences in symbolically-encoded polyphonic music, computational-geometry-based techniques seem very promising: we have extended the setting behind our original sweep-line music-retrieval algorithm [1] to new musical point-pattern matching problems [2, 3, 10, 11] and have also adapted the framework of mathematical morphology to this problem area [7]. Most recently, we have adapted the method to work directly with audio music [16].

Another research branch of the group is in applying normalized compression distance (NCD) in content-based music retrieval (CBMR) tasks. In his PhD project, MSc Ahonen have focused on CBMR using NCD both in audio [4, 5, 11] and in symbolic [6] domains. In addition to studying how the musical information should be represented for compression-based similarity measuring, the work has presented several novel ideas for extending the pairwise NCD similarity measuring to sets of objects [5] and explored how NCD can be used in classification and clustering instead of more commonly used distance metrics [6,11]. Recently, we have also suggested an adaptive representation for music features that both allows efficient similarity measuring with any common similarity metric while preserving essential distinguishing power [12].

A new research branch of the group is automatic music transcription from a musician's viewpoint. In his PhD project, MSc Laaksonen, currently concentrates on chord transcription with the aim of finding new ways to utilize the musical context in the transcription. His latest findings suggest that the melody context, which has not been used in chord transcription so far, plays an important role in cases where the pure chord content is ambiguous. Laaksonen aims at creating transcriptions which sound good as a whole in a real musical performance rather than creating exact transcriptions from individual chords. Results of this project have been reported in publications [13-15].
**Future plans**
Dr. Lemström is co-editing a text-book (working title “The Oxford Handbook of Automated Knowledge Discovery in Music”) with two internationally leading scientists in the field, Professor Geraint Wiggins and Professor Roger Dannenberg.

MSc Teppo Ahonen’s PhD thesis, *Compression-based Similarity Measuring for Practical Applications in Content-based Music Information Retrieval*, is set to be defended in fall 2012. The thesis focuses on using NCD and other compression-based similarity measures to measure similarity between tonal features extracted from music data. The thesis provides insight into (1) what features are essential when measuring tonal similarity between pieces of music, (2) how the features should be represented for a compression-based similarity metric, (3) what are the advantages and disadvantages of using NCD for measuring tonal similarity, and (4) how the methodology can be applied for retrieval and machine learning tasks.

**Cooperation**
- Gerant Wiggins, Goldsmiths College, University of London, UK, visits, joint book project
- Roger Dannenberg, Carnegie Mellon University, USA, joint book project
- David Rizo, Jose Manuel Iñesta, University of Alicante, joint publications
- David Meredith, Aalborg University, visits, joint publications

**Selected publications**


Computational Linguistics Group

Members

- Roman Yangarber, University Researcher, Group leader
- Mian Du, Doctoral student
- Lidia Pivovarova, Doctoral student (2011-2012 visiting from St Petersburg State University; at UH 11/2012- )
- Javad Nouri, MSc student
- Guowei Lv, MSc student (MSc 2014)
- Matthew Pierce, MSc student
- Natalia Ostapuk, Doctoral student (2013-2014 visiting from St Petersburg State University)
- Silja Huttunen, Doctoral student (Linguistics) (-2011)
- Hannes Wettig, Doctoral student (-2012)
- Peter von Etter, MSc, (-2013)
- Arto Vihavainen, MSc student (-2011)
- Suvi Hiltunen, MSc student (-2011)
- Mikhail Novikov, MSc student (-2011)

Mission of the group

The group works on various problems in analysis of linguistic data. We investigate how language conveys information, how information can be extracted from linguistic data, and how hidden, underlying structure can be learned from observed linguistic data. The research programme combines empirical, applied and theoretical approaches to these problems.

Research activities

PULS

The PULS Project builds tools for analysis of plain text, and specifically for surveillance of on-line news media. The group conducts research in the field of information extraction: i.e., identifying pre-defined types of events in text. PULS participates in projects on several different knowledge domains: business intelligence, epidemiological surveillance, and cross-border security.

For example, in the domain of epidemic surveillance the system identifies, in each news article, how many people have been affected with what condition, where, when, etc. The system is operational (at puls.cs.helsinki.fi with access to detailed analysis obtainable from puls@cs.helsinki.fi). The project is developed in collaboration with international partner organisations, who act as research partners and users. A central research theme in the project is automating the acquisition of domain knowledge from plain text. Machine learning methods, especially weakly-supervised learning, are at the core of the methods to bootstrap new systems for analysing documents in new domains quickly and accurately. An important objective of PULS is to investigate and model linguistic phenomena in the context of real-world applications. The key benefit of engaging end-users is that they provide high-quality feedback, as well as annotated data for training and testing, which allows us to experiment on a large scale with supervised and weakly-supervised methods. Prior work deals with text analysis in toy-like laboratory settings, where the amount of data is limited; shortage of data is a bottleneck in NLP research in general, and in news surveillance in particular. PULS has been carefully designed and coordinated over the last 4 years with a view toward obtaining good data in collaboration with real-world end users.
**Etymon**

The Etymon Project aims to develop computational methods for studying the origin and development of languages and language families, i.e., genetically related groups of languages. We use databases of lexical material from these languages and developing computational methods for studying language relationships. Currently these methods are applicable generally in principle, and are currently being tested on the Uralic, Turkic, Indo-European and Xoisan families. Etymon brings together an international, inter-disciplinary team of researchers, with complementary expertise in the areas of computational, comparative and historical linguistics.

Research on language evolution and inter-linguistic relationships has gone on for over two centuries, during which linguists devised methods for discovering patterns of correspondence, and for testing hypotheses. Because linguistic data is often scarce and incomplete, modern non-computational methods for investigating linguistic relationships leave a large number of “grey areas” – questions that current theory is unable to answer with certainty. In particular, the Uralic data is strikingly uncertain, with conflicting theories in current scholarship. Earlier research on computational etymology over the last decade has focused largely on the Indo-European language family (traditionally at the center of historical-linguistic research) with less emphasis on other families. Etymon has developed novel computational approaches based on the information-theoretic MDL (minimum description length) Principle, to model the etymological correspondences and evolution within the family. One of our aims is to illuminate some of the uncertainties in existing data sets. Another is to compare results from the computational methods with those obtained by traditional, manual methods, directly from the data -- and using all available data in an objective, unbiased fashion. Relationships among words from a group of related languages are captured by discovering regular rules of derivation, from parent to child language, or among sibling languages. The main idea is that the “correct” set of relationships (or a set of rules that efficiently describe the derivation) among a group of related languages will yield a compact encoding for the totality of observable data in these languages.

The key research objectives are:

1. develop novel computational methods for discovering and investigating the relationships, using rich sets of relevant data;
2. apply and test the methods in a wider setting to investigate specific language families, and to the study of relations beyond established language families.

We do not aim to replace the human computational linguist by a model, but rather to aid the linguist by providing additional objective sources of evidence. This will enable one to formulate and quantify evidence that supports the hypotheses and conclusions in ways that were not possible before.

We use data collections provided by partner organisations: – KOTUS (Institute for research on languages of Finland) and the Russian Academy of Sciences.

**FinUgRevita**

The Computational Linguistics Group has begun a new project (jointly funded by Finland and Hungary) to build novel tutoring and authoring tools for supporting revitalisation of severely endangered languages. The initial focus is on the Uralic languages. There are several difficult research questions. We aim to

1. utilise existing low-level pre-processing resources to provide fundamental automated analysis of arbitrary text in a target language;
b. define a set of states of "competence" -- or "linguistic items" which an expert teacher would consider as helpful for assessing user proficiency,
c. automatically infer a partial order of dependencies among the postulated states -- a dependency network.

The goal is then to create a teaching/tutoring system that is able to use dynamic content, and use the learned dependency network of competence states to create a reliable model of competence of a given user. This is to assure that the user will not be bored with test examples that are too easy, nor discouraged by test examples that are too difficult.

**Future plans**

In PULS, one established direction that the group is exploring is the acquisition of different types of knowledge bases in parallel, where they provide mutual constraints to maintain high precision. An important new direction is extending the linguistic coverage of PULS; most of the work in the field still focuses on English-language text, motivated by the abundance of lower-level processing tools available for English. PULS has been working to extend the coverage to Russian, which is notably much more poor in NLP resources, yet has high value in the news scenarios covered to date. The methods employed include bootstrapping techniques, together with methods for "projecting" existing (English) resources/tools onto the target languages, and using similar documents in English and Russian, so-called "comparable" data (i.e., not strictly "parallel" data).

In Etymon, the methods we devise for examining relationships are being applied in a wider context:

a. extending alignment of observed data to performing reconstruction of the corresponding forms in the unobserved, ancestral languages.
b. to apply the methodology to other language families, especially the less-studied families, to help obtain novel results.
c. collaboration with population-genetics experts, to i. explore cross-pollination on the methodological level (i.e., applicability of algorithms to different kinds of data), and ii. explore complementary information carried in linguistic vs. genetic data. This is available (to a degree) for certain population, and presents an exciting opportunity for deeper insight into the inter-relationships among the related groups.

**Societal, economic and technical impact**

In the domain of Epidemic Surveillance, PULS collaborates with major National and international Health Agencies, including Health Ministries of several European Countries (France, UK, and Spain), the European Center for Disease Control (ECDC), in Stockholm, Sweden, and the WHO (World Health Organisation). Computational methods developed in PULS help specialists at these agencies perform their tasks more efficiently, in protecting the European citizen from dangerous epidemics. Users of PULS in the domain of business intelligence include international companies.

The results of Etymon provide insights into the structure and development of language families. These relationships have been studied for two centuries using manual methods. Computational methods for analysing these fascinating data are only now beginning to emerge. The results will have wide-ranging implications for the understanding of the common origins of language.

In FinUgRevita, we work with interested parties, teachers and schools, in areas where there is interest in supporting or revitalising languages in danger of imminent disappearance, which has the potential to slow or reverse language decay.
Cooperation

Cooperation within Algodan
Link and Pattern Discovery Group: PULS exploits the tools developed by the Biomine group for analysing the complex graphs resulting from the analysis of news in the domain of business intelligence.

Cooperation within the University
Etymon project collaborates with the Population Genetics group, lead by Prof. Jukka Corander, (joint publications currently in preparation), as well with researchers at the Department of Modern Languages, and the Department of Finno-Ugric Studies.

International cooperation
PULS and Etymon have very strong on-going international collaboration.

In PULS we have been collaborating with the Text Mining Research Unit at the European Commission's Joint Research Centre (JRC, in Ispra, Italy) for several years. In the domain of border security, PULS collaborates with the EC Frontex Agency for the protection of the European External Frontiers. PULS provides an on-line feed into MedISys, the system for global news surveillance developed by the JRC. Our results and links to our databases are served in real-time on the MedISys platform at http://medusa.jrc.it – JRC sends raw articles that it mines from the Web to PULS, and PULS returns the results to JRC. Medlsys has thousands of users every day, government and private; our results are documented in several joint publications. Some of these users – Public Health professionals – use PULS on a daily basis. PULS participates in the Global Health Security Initiative (GHSI)/Global Health Security Action Group’s Early Alerting and Reporting project. The GHSI is an international consortium created to strengthen health preparedness and response globally to biological, chemical, radiological/nuclear threats. This initiative was launched in 2001 by Canada, the EU, France, Germany, Italy, Japan, Mexico, the UK, and the USA. PULS is the only European academic partner within the GHSAG consortium (the other two are from University of Tokyo and Harvard). Other members in the consortium are user organisations – those who need high-quality analysis of news for surveillance of thousands of on-line sources, in real time. The partners include the European Center for Disease Control (ECDC) in Stockholm, Sweden, the WHO (World Health Organisation), as well as several large national health ministries. As mentioned, these users are important for our work for providing challenging scenarios and good data.

Etymon is an inter-disciplinary project, where in addition to local collaborators; we work closely with the Russian Academy of Sciences, in Moscow, Russia.

Selected publications


Team Link and Pattern Discovery

Discovery Group: Data Mining and Computational Creativity

Members
- Hannu Toivonen, Professor, Group leader
- Alessandro Valitutti, PhD, Postdoctoral researcher ( -2013)
- Laura Lango, Doctoral student
- Oskar Gross, Doctoral student
- Jukka Toivanen, Doctoral student
- Fang Zhou, Doctoral student ( -2012)
- Esther Galbrun, Doctoral student, co-supervised with Mikko Koivisto ( -2013)
- External doctoral students
  - Joonas Paalasmaa, Doctoral student, employed by Beddit.com Ltd. ( -2014)
  - Mika Timonen, Doctoral student, employed by VTT, the Technical Research Centre of Finland ( -2013)

Mission of the group
The Discovery group develops novel methods and tools for data mining and computational creativity. Our focus is on algorithmic methods for discovering links and patterns in data, and especially on their use in creative systems. Application areas range from computational generation of poetry to link discovery in bioinformatics and to sleep analysis.

Research activities
A methodological focus area has been in analysis and exploration methods for weighted (biological or word co-occurrence) graphs, i.e., mainly in the theme “L - learning from and mining structured and heterogeneous data” of Algodan.

In 2011, we initiated work on computational poetry and have since expanded our efforts to computational humor, music, and fine arts [3, 4, 7]. Especially in linguistic creativity we develop novel methods that learn from existing texts and thereby minimize the need for manually coded or language specific knowledge, making the methods more easily adaptable to different languages. Artistic results of the computational creativity work have been published in print and exhibited in several galleries (see the section on Societal, economical and technical impact below).

Various (other) data mining method development activities have also been continued, covering graph mining [12] and its applications in biology [1, 9, 11], sleep analysis [7, 8], discovery of functional dependencies [5], subgroup discovery [6], redescription mining [10], and news analysis.

Future plans
The group will focus on computational creativity, using data mining and graph mining as the methodological basis. We will primarily develop methods that either are “knowledge poor” (in the sense of not relying on knowledge bases or manually coded knowledge) or that help make sense of data [7].

In the past two years we have established strong networks with computational creativity researchers internationally (see below). In the future, we plan to establish contacts and collaboration with scientists in applied fields (literature, cinema), also domestically.
**Societal, economical and technical impact**

With our shift of research focus towards computational creativity, our work has reached two novel forms of impact.

First, we have a societal and cultural impact via contributions to *artistic acts* using computational creativity. Since 2012, we have contributed computational poetry to an art book, given several performances of biomusic, participated in installations of brain wave inspired poetry and of biophysical sensor based music, and contributed to paintings, among others. Most of these works have been collaborations with professional artists and have been exhibited or performed in galleries. (For more information, see http://www.cs.helsinki.fi/en/discovery/art .)

Second, the impact of this work is indirectly evidenced by the attention received from public media. Since 2012, our work has been covered on TV (YLE, Finland), radio (CBC, Canada), popular press (e.g., The Times, UK; New Scientist, USA; Helsingin Sanomat, Finland), leading technology web sites (e.g., CNET, Engadget), and hundreds of other websites in at least fifteen languages.

Economical and technical impact is primarily via collaboration with companies (see above). Sleep research carried out with Beddit Ltd. seems highly promising based on the interest to their newest product.

**Cooperation**

**Cooperation within Algodan**

Co-supervision of students, exchange of information, and other informal co-operation.

**Cooperation within the universities**

- Prof. Markku Partinen, Institute of Clinical Medicine: joint research in sleep analysis, joint publications.
- Dr. Juhani Huovelin, Department of Physics: joint research on news analysis and aggregation, joint publications.
- Dr. Liisa Ilomäki, Institute of Behavioural Sciences: joint proposal to research educational uses of computational creativity in schools.
- Dr. Alina Leminen, Institute of Behavioural Sciences: starting joint research on empirical cognitive neuroscience studies for linguistic creativity.

**Other national cooperation**

- Beddit Ltd: research collaboration on sleep analysis methods, supervision of PhD studies, joint publications.
- Research collaboration with a number of companies in the Future Media program of the ICT cluster of the Finnish Strategic Centres for Science, Technology and Innovation (Tivit Ltd).

**International cooperation**

- Concept Creation Technology, ConCreTe, EU project with 7 partners, 2013-16.
- Promoting the Scientific Exploration of Computational Creativity, PROSECCO, EU coordination action with 7 partners, 2013-16.
- Professor Nada Lavrac, Jozef Stefan Institute, Slovenia: joint research on data and text mining for bioinformatics. Joint publications.
- Professor Ross D. King, University of Manchester, UK: joint research on graph mining in bioinformatics, research visits, joint publication.
• Professor Jiuyong Li, University of Southern Australia: joint research on dependency mining. Joint publications, research visits.
• Assoc. Professor Antoine Doucet, University of Caen, France: joint research on text mining, student exchange, research visits, joint publications.

Selected publications
1. The Use of Weighted Graphs for Large-Scale Genome Analysis. Fang Zhou, Hannu Toivonen, Ross D. King. PLoS One, accepted for publication.
7. Sleep Musicalization: Automatic Music Composition from Sleep Measurements. Aurora Tulilaulu, Joonas Paalasmaa, Mikko Waris, Hannu Toivonen. Eleventh International Symposium on Intelligent Data Analysis (IDA), LNCS 7619, 392-403, Helsinki, Finland, October 2012. (See sleepmusicalization.net for an implementation and example music.) (Winner of the IDA 2012 Frontier Prize as the ‘most novel and visionary contribution’ of the conference.)
Team Machine Learning

Machine Learning Group
Machine Learning group consists of two subgroups:

- Kernel Machines, Pattern Analysis and Computational Biology (until 2011: Computational Systems Biology and Bioinformatics), and
- Learning Theory.

Members
- Jyrki Kivinen, Professor, Principal Investigator, Group leader (subgroup Learning Theory)
- Juho Rousu, Professor, Principal Investigator, Group leader (subgroup Kernel Machines, Pattern Analysis and Computational Biology)
- Jana Kludas, Postdoctoral researcher (2012-)
- Esa Pitkänen, Postdoctoral researcher (-2011)
- Markus Heinonen, Doctoral student (-2012)
- Panu Luosto, Doctoral student (-2013)
- Hongyu Su, Doctoral student (2011-)
- Huibin Shen, Doctoral student (2012-)
- Anna Cichonska, Doctoral student (2013-)

Mission of the group
The group develops machine learning methods, models and tools for computational sciences, in particular computational biology. The methodological backbone of the group is kernel methods and regularized learning. The group particularly focusses in learning with multiple and structured targets, multiple views and ensembles. Applications of interest in computational biology include network reconstruction, gene functional classification as well as metabolite identification.

Research activities

- Learning from and mining structured and heterogeneous data
  Metabolite identification from tandem mass spectra is an important problem in metabolomics, underpinning metabolic modelling and network analysis. With the support of Academy of Finland research grant (MIDAS, 2013-2017), we develop methods for automatic identification of metabolites from tandem mass spectra using machine learning. In our most recent approach, fragmentation trees are computed from the spectra, and turned into a collection of kernels capturing similarities of trees from different viewpoints. The kernels are then combined using state-of-the-art multiple kernel learning methods. Our method is currently the most accurate fully computational approach for metabolite identification [1].

  In metabolic network analysis we developed methods for simultaneous reconstruction of metabolic networks for a set of related organism, connected by a phylogenetic tree. The method generalizes the Fitch-Hartigan algorithm to discover phylogenetic tree with minimum number of reaction mutations so that each ancestral node corresponds to a gapless metabolic network [2]. Another example of our recent work is biomarker discovery from plasma proteomics and clinical data using sparse canonical correlation analysis. In the method L1-penalization is used for the proteomics data while the clinical data is kernelized [4].
F - Foundations of algorithmic data analysis

In machine learning, we have developed new methods for multilabel classification, relying on ensemble learning on a collection of random output graphs imposed on the multilabels, and a kernel-based structured output learner as the base classifier. Theoretically, we have been able to show that our multilabel ensemble not only benefits from diversity in the base classifier predictions (phenomenon already known for single target ensembles), but also from the covariance of predictions of microlabel pairs. Experimentally, our method has robust performance on a variety tasks, matching or exceeding the performance of commonly used multilabel/task learning approaches [3].

Issues related to the Minimum Description Length (MDL) principle have been studies in joint work between Panu Luosto and Dr. Petri Kontkanen. The main conceptual novelty is a systematic way of dealing with cases where the parametric complexity of the model class is infinite. The general principle has been specifically applied mainly in clustering and histogram models. There has also been some work on more specific computational issues that arise in these applications [5].

Future plans

Major themes for future research of the group include:

- Kernel methods for structured data. We will develop kernel representations for structured objects such as sequences, trees and graphs, and efficient algorithms for mapping data into kernels and back (aka pre-image problem). Especially we focus on predicting structured output, a setup that aims to leverage the structure of the data to increase efficacy of learning and prediction. Learning with multiple kernels and ensembles is another direction of importance.

- Network labeling. In many applications, the data to be predicted has a network form. In particular, network labeling problems involve a known network structure, and a set of data instances that each activates a set of nodes. The prediction task is to learn which nodes are to be activated for a given input data. Examples of such prediction problems are frequent in document management (e.g. hierarchical text classification) as well as information and computer networks (e.g. resource placement).

- Computational biology. A major application field of our methods is in biological sciences. Our core competence there is in biological network reconstruction and analysis, especially in metagenomic context, as well as prediction problems involving small molecules (e.g. metabolomics, mass spectrometry).

Social, economical and technical impact

Our research is geared towards building computational tools and methods for the analysis of data arising in biotechnology and biomedicine. The impact of our work is in making applied research better targeted, faster and more cost-effective. An example of this is our metabolite identification technology, which speeds up research in metabolomics by using less time in verifying false positive predictions and may have ramifications to applications such as doping control and forensics.

Cooperation

Cooperation within Algodan

The group collaborates with the Succinct Data Structures (currently: Genome Scale Algorithmics) group in algorithms and index structures for graph data and data mining group in network labeling problems.
National cooperation

- University of Helsinki, Finland/Professor Liisa Holm. Collaboration in enzyme function prediction and metabolic reconstruction [2].
- VTT Technical Research Center of Finland/Professor Merja Penttilä, Doc. Merja Itävaara. Collaboration in pathway modelling in industrial microbes (EU FP7 STREP BIOLEDGE) and deep biosphere microbiota (GEOBIOINFO project, 2011-).
- University of Helsinki (Complex Systems Computation Group). Ongoing collaboration on MDL with Dr. Petri Kontkanen, which has lead to three published articles.

International cooperation

- University College London, United Kingdom/Professor John Shawe-Taylor. Collaboration in kernel methods and structured output learning. The collaboration has resulted in several new methods for machine learning for structured data, including sequences, taxonomies, and general graphs.
- Friedrich-Schiller Universität Jena/Professor Sebastian Böcker. Collaboration in Metabolite identification method development.
- ETH Zurich, Institute of Molecular Systems Biology/Dr. Nicola Zamboni. Collaboration in Mass Spectrometric data analysis

Selected publications

Team Neuroinformatics

Neuroinformatics Group

Members
- Aapo Hyvärinen, Professor, Group leader
- Patrik Hoyer, Academy Research Fellow, Co-leader (-7/2013)
- Michael Gutmann, Postdoctoral researcher (-5/2012)
- Jun-ichiro Hirayama, Postdoctoral researcher (-6/2011)
- Cristina Campi, Postdoctoral researcher (-8/2011)
- Hugo Eyeherabide, Postdoctoral researcher (10/2011- )
- Jukka-Pekka Kauppi, Postdoctoral researcher (6/2011- )
- Doris Entner, Doctoral student (-11/2013)
- Antti Hyttinen, Doctoral student (-5/2013)
- Jouni Puuronen, Doctoral student
- Miika Pihlaja, MSc/Doctoral student (-12/2011)

Mission of the group
Our mission is to develop statistical data analysis methods, with the particular applications of neuroscience in mind. In some areas of neuroscience, such as brain imaging, measurement devices provide huge amounts of data and new methods are needed to analyze the data. On the other hand, modelling perception can be approached from a Bayesian viewpoint, as probabilistic modelling of typical stimuli. General-purpose statistical learning methods are naturally developed at the same time.

Research activities

D - Discovery of hidden structure in high-dimensional data

Testing for ICA. In the ICA research community, the almost exclusive focus has been on estimation, and testing of the results has received almost zero attention. However, in any practical application it would be extremely important to be able to test some kind of statistical significance of the components. Otherwise, we don't know if the results are just due to random noise, or local minima. We have developed a new testing framework for ICA based on the idea of having many datasets (or splitting one dataset into many), applying ICA separately on each dataset, and then investigating if the results are similar enough in the different datasets. We were able to formulate a proper null hypothesis and use the conventional machinery of the theory of statistical hypothesis testing. The test was developed for two different cases: testing the mixing matrix [1], or testing the actual values of the independent components [2].

Causal discovery. Our project on estimating causal relations from continuous-valued data continued to be successful and lead to many new methods. Highlights include: very simple measures of causal direction for two variables, i.e. does x cause y or does y cause x [3], a procedure to estimate the strength of causal effects in the presence of hidden variables [4], and identifiability results of linear cyclic causal models based on randomized experiments [5].

Neuroimaging data analysis. In our joint neuroimaging project with Prof. Riitta Hari of Aalto University, we investigated new variants of ICA for analysis of spontaneous brain activity (e.g. during rest) [6], as well as "decoding" in MEG, i.e. using classification methods to infer what kind of stimulation was given to the
brain, using only the MEG data as input to the classifier [7]. An important focus was the very deep topic of two-person neuroscience. This project, related to Prof. Hari’s ERC Advanced Grant of the same title, attempts to open completely new vistas in social neuroscience by measuring two subjects’ brain activities at the same time. This is an extremely challenging, high-risk project since almost everything has to be built from scratch: instrumentation (connecting to MEG systems with audio and video links), experimental design (since such experiments have hardly been conducted before), and data analysis (which is our responsibility in this joint project). Initial developments in the machine learning methods were published in [8], but much remains to be done.

**Foundations of algorithmic data analysis**

**Estimation of non-normalized probabilistic models.** Our project on estimating non-normalized probabilistic models culminated in a long JMLR paper [9] which shows a deep connection between supervised and unsupervised learning, and how it can be utilized to estimate such intractable models.

**Future plans**

The end of the Algodan funding period coincides with a changepoint in the structure and focus of the neuroinformatics group, since the co-leader, Patrik Hoyer, left academia to start his own company in autumn 2013. At the same time, the long sabbatical that Aapo Hyvärinen spent in ATR, Japan, in 2013-2014 led to many new ideas. In the future, causal discovery and estimation of non-normalized probabilistic models will be more or less discontinued, and replaced by new projects related to modeling spontaneous brain activity. Measuring spontaneous brain activity (i.e. during rest) is now popular in brain imaging, but functional models of the information processing happening in spontaneous brain activity are very rare. In other words, we don’t really know very well what the brain is doing during rest, and, most importantly, why. Theoretical models, based on machine learning and probabilistic modeling, would be extremely useful here.

**Societal, economical and technical impact**

Several of the new methods have been made publicly available as software packages distributed on the internet.

**Cooperation**

**Cooperation within Algodan**

In recent years, one of the main strands of work in the group has been on the topic of methods for learning directed graphical models from data. This family of models includes Bayesian networks, well studied in the fields of machine learning and artificial intelligence. Thus, the group has recently benefited greatly from the expertise of Dr. Mikko Koivisto on exact algorithms for structure learning of Bayesian networks. Recently, we have also benefited from discussions with Dr. Koivisto and Dr. Petteri Kaski on problems in combinatorics derived from our identifiability conditions on learning cyclic models from randomized experiments.

**National cooperation**

- Riitta Hari, Aalto University, Finland. Topic: Brain imaging. Joint project funded by Academy of Finland; joint Doctoral student, joint Postdoctoral researchers

**International cooperation**

- Stephen M. Smith, Oxford University, UK. Topic: causal discovery in fMRI. Aapo Hyvärinen visited him for two months in 2011. One joint publication.
• Frederick Eberhardt, Peter Spirtes, et al, Carnegie Mellon University, USA. Topic: theory of causal discovery. Several visits (2-3 weeks each) by Patrik Hoyer. Several joint publications.

• Dominik Janzing, Max Planck Institute for Cybernetics, Germany. Topic: causal inference among high-dimensional variables. Joint publication.

• Alessio Moneta and Alex Coad, Max Planck Institute for Economics, Germany. Topic: causal inference in time-series. Joint publications.

• Shin Ishii, Kyoto University and Motoaki Kawanabe, ATR, Japan. Topic: Brain imaging signal analysis. Aapo Hyvärinen visited ATR for 10 months in 2013-2014. We have a joint publication and a couple in preparation.

• Klaus-Robert Müller, Berlin Institute of Technology, Germany. Topic: Decoding biomedical signals. Jukka-Pekka Kauppi spent 2 months in Berlin in 2013 to start the collaboration. One joint article in preparation.

• Masashi Sugiyama, Tokyo Institute of Technology, Japan. Topic: Image processing. We have a joint project based on joint post-doc supervision. One joint article submitted.

Selected publications


5. Publications

2013

Articles in refereed scientific journals


**Technical reports and other publications**


Artistic work

1. O. Gross (editor) and H. Toivonen. Organization of Computational Creativity Club. 2013.


3. O. Holmqvist, D. Murphy and J. Paalasmaa. Biomusic performance, the 8th World Conference of Scientists Journalists, 2013.


2012

Articles in refereed scientific journals


Refereed conference articles and articles in edited books


Technical reports and other publications


**Artistic work**


**2011**

**Articles in refereed scientific journals**


Refereed conference articles and articles in edited books


Technical reports and other publications


6. **PhD degrees**

Members of Algodan obtained 21 PhD degrees, listed below.

**2014**
1. **Galbrun, Esther.** Methods for Redescription Mining. *University of Helsinki.*
2. **Korhonen, Janne.** Graph and Hypergraph Decompositions for Exact Algorithms. *University of Helsinki.*

**2013**
1. **Entner, Doris.** Causal Structure Learning and Effect Identification in Linear Non-Gaussian Models and Beyond. *University of Helsinki.*
4. **Lijffijt, Jeffrey.** Computational methods for comparison and exploration of event sequences. *Aalto University.*
5. **Luosto, Panu.** Normalized Maximum Likelihood Methods for Clustering and Density Estimation. *University of Helsinki.*

**2012**
1. **Hanhijärvi, Sami.** Multiple hypothesis testing in data mining. *Aalto University.*
2. **Heinonen, Markus.** Computational Methods for Small Molecules. *University of Helsinki.*
3. **Parviainen, Pekka.** Algorithms for Exact Structure Discovery in Bayesian Networks. *University of Helsinki.*
4. **Sirén, Jouni.** Compressed Full-Text Indexes for Highly Repetitive Collections. *University of Helsinki.*
5. **Vuokko, Niko.** Testing the Significance of Patterns in Complex Null Hypotheses. *Aalto University.*
6. **Välimäki, Niko.** Applications of Compressed Data Structures on Sequences and Structured Data. *University of Helsinki.*
7. **Wessman, Jaana.** Mixture Model Clustering in the Analysis of Complex Diseases. *University of Helsinki.*

**2011**
1. **Junttila, Esa.** Patterns in Permuted Binary Matrices. *University of Helsinki.*
2. **Hintsanen, Petteri.** Simulation and Graph Mining Tools for Improving Gene Mapping Efficiency. *University of Helsinki.*
3. **Ojala, Markus.** Randomization Algorithms for Assessing the Significance of Data Mining Results. *Aalto University.*
APPENDIX 1: Analysis of ALGODAN Publications 2008-2013

The analysis was made by Helsinki University Library, Kumpula Campus Library bibliometrics team in April 2014 using the publication list provided by the research group.

Summary of publication statistics

There are 661 publications listed for 2008-2014. The types and annual statistics are summarized in Table 1.

Table 2: ALGODAN publications by year and type. See explanation below for type classification.

<table>
<thead>
<tr>
<th>Year</th>
<th>A1</th>
<th>A3-A4</th>
<th>B-E</th>
<th>C</th>
<th>F</th>
<th>G</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>40</td>
<td>68</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>124</td>
</tr>
<tr>
<td>2009</td>
<td>34</td>
<td>46</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>97</td>
</tr>
<tr>
<td>2010</td>
<td>34</td>
<td>75</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>129</td>
</tr>
<tr>
<td>2011</td>
<td>31</td>
<td>52</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>92</td>
</tr>
<tr>
<td>2012</td>
<td>29</td>
<td>51</td>
<td>20</td>
<td>0</td>
<td>1</td>
<td>8</td>
<td>109</td>
</tr>
<tr>
<td>2013</td>
<td>32</td>
<td>54</td>
<td>4</td>
<td>0</td>
<td>10</td>
<td>8</td>
<td>108</td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>346</td>
<td>64</td>
<td>1</td>
<td>11</td>
<td>39</td>
<td>661</td>
</tr>
</tbody>
</table>

The publications are classified according to the following scheme:

- A1: Articles in refereed scientific journals
- A3-A4: Refereed conference articles and articles in edited books
- B-E: Technical reports and other publications
- C: Books
- F: Artistic works
- G: Theses

The large share of conference articles and articles in edited books, as shown for A3-A4, is usual for computer science.

Coverage in bibliometric databases

The coverage of computer science publications in Thomson & Reuters Web of Science is known to be insufficient, and based on this, Web of Science was excluded from this analysis.

A search for all relevant publications (from A1 to C) was made in Elsevier Scopus\(^3\). The hits are listed in Table 2. Likewise, a search was made with the Publish or Perish tool\(^4\) (based on Google Scholar). The hits are also listed in Table 2.

---

\(^3\) [http://www.scopus.com](http://www.scopus.com)

\(^4\) [http://www.harzing.com/pop.htm](http://www.harzing.com/pop.htm)
Comparing the hits to the data in Table 1, we get the following coverage for ALGODAN publications:

- Scopus 71.5% – 437 out of 611 (artistic works and theses excluded)
- Publish or Perish 90.8% – 590 out of 650 (artistic works excluded)

Publish or Perish (PoP) coverage is good, but identifying the hits was considerably slow and in some cases, it was difficult to identify the hits reliably, so the PoP results should be seen as a good guess.

The coverage in Scopus is better than anticipated, and reflects the efforts by Elsevier to increase the number of conference publications in Scopus. The publication types (for more than 1 hits) are listed in Table 3. The most popular publication channel was LNCS as seen from Table 4.

### Table 4: Scopus publication types with > 1 hits

<table>
<thead>
<tr>
<th>Publication Type</th>
<th># of titles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conference Paper</td>
<td>256</td>
</tr>
<tr>
<td>Article</td>
<td>163</td>
</tr>
<tr>
<td>Review</td>
<td>7</td>
</tr>
<tr>
<td>Book Chapter</td>
<td>4</td>
</tr>
<tr>
<td>Letter</td>
<td>3</td>
</tr>
</tbody>
</table>

### Table 5: Publication channels with >4 hits according to Scopus

<table>
<thead>
<tr>
<th>SOURCE TITLE</th>
<th># of titles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture Notes in Computer Science</td>
<td>125</td>
</tr>
<tr>
<td>Journal of Machine Learning Research</td>
<td>13</td>
</tr>
<tr>
<td>Information Processing Letters</td>
<td>12</td>
</tr>
<tr>
<td>Proceedings IEEE International Conference on Data Mining</td>
<td>9</td>
</tr>
<tr>
<td>ACM International Conference Proceeding Series</td>
<td>7</td>
</tr>
<tr>
<td>BMC Bioinformatics</td>
<td>6</td>
</tr>
<tr>
<td>Proceedings of the ACM SIGKDD</td>
<td>6</td>
</tr>
<tr>
<td>Neurocomputing</td>
<td>5</td>
</tr>
<tr>
<td>Bioinformatics</td>
<td>5</td>
</tr>
<tr>
<td>Plos One</td>
<td>5</td>
</tr>
</tbody>
</table>
Scopus seems to be a little inconsistent with subject areas. A few publication series that probably should be classified under computer science are classified e.g. under mathematics only. A rough view of subject area provided by Scopus looks like in Figure 1.

Figure 1: ALGODAN subject areas in Scopus

Citation analysis with Scopus

The ALGODAN publications found in Scopus received 2673 citations by beginning of April 2014. A more detailed view for each year is seen in Table 5.

Table 6: Citations in Scopus by year

<table>
<thead>
<tr>
<th>Year</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2008</td>
<td>22</td>
<td>96</td>
<td>162</td>
<td>158</td>
<td>180</td>
<td>133</td>
<td>26</td>
<td>777</td>
</tr>
<tr>
<td>2009</td>
<td>37</td>
<td>152</td>
<td>167</td>
<td>159</td>
<td>146</td>
<td>27</td>
<td>688</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>2</td>
<td>53</td>
<td>243</td>
<td>263</td>
<td>266</td>
<td>39</td>
<td>866</td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>1</td>
<td>0</td>
<td>8</td>
<td>69</td>
<td>92</td>
<td>13</td>
<td>183</td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td></td>
<td>26</td>
<td></td>
<td></td>
<td>80</td>
<td>14</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td></td>
<td></td>
<td>30</td>
<td>8</td>
<td></td>
<td>0</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>127</td>
<td>2673</td>
<td></td>
</tr>
</tbody>
</table>

As University of Helsinki does not have access to Scopus raw data, we could only do a partial analysis of highly cited (top 10%) publications. The annual volume of Computer Science publications in Scopus is around 300,000 while Scopus only allows downloads of max 20,000 items. However, it was possible to
extrapolate the citation numbers to get the cut-off value for the top 10% for each year for computer science. As for publications that were classified under mathematics, the annual volume was low enough and no extrapolation was needed.

For biosciences and medicine, it was not possible to estimate a 10% cut-off value due to high publication volumes. Also, the citation rates for these subject areas differ strongly from one subfield to another.

Table 6 lists the number of highly cited publications for the two dominant subject areas. The mathematics values cover only those publications not listed under computer science.

Table 7: Number of highly cited publications

<table>
<thead>
<tr>
<th>Year</th>
<th>Scopus hits</th>
<th>10% cutoff</th>
<th># of10% items</th>
<th>10% cutoff</th>
<th># of10% items</th>
<th># of10% items</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>76</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>2009</td>
<td>60</td>
<td>9</td>
<td>12</td>
<td>11</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>2010</td>
<td>99</td>
<td>6</td>
<td>17</td>
<td>9</td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td>2011</td>
<td>70</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2012</td>
<td>77</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2013</td>
<td>52</td>
<td>1</td>
<td>9</td>
<td>1</td>
<td>2</td>
<td>11</td>
</tr>
</tbody>
</table>

A list of highly cited publications can be found in Appendix 1.

Citation analysis with Publish or Perish

Citations from Publish or Perish are listed in Table 7. The quality of the citations has not been controlled in any way.

Table 8: Citations by year and type according to PoP

<table>
<thead>
<tr>
<th>Year</th>
<th>A1</th>
<th>A3-A4</th>
<th>B-E</th>
<th>C</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>1111</td>
<td>752</td>
<td>23</td>
<td>0</td>
<td>1886</td>
</tr>
<tr>
<td>2009</td>
<td>1006</td>
<td>532</td>
<td>9</td>
<td>208</td>
<td>1755</td>
</tr>
<tr>
<td>2010</td>
<td>1102</td>
<td>386</td>
<td>22</td>
<td>0</td>
<td>1510</td>
</tr>
<tr>
<td>2011</td>
<td>270</td>
<td>166</td>
<td>0</td>
<td>0</td>
<td>436</td>
</tr>
<tr>
<td>2012</td>
<td>175</td>
<td>136</td>
<td>25</td>
<td>0</td>
<td>336</td>
</tr>
<tr>
<td>2013</td>
<td>149</td>
<td>65</td>
<td>29</td>
<td>0</td>
<td>243</td>
</tr>
<tr>
<td>Total</td>
<td>3813</td>
<td>2037</td>
<td>108</td>
<td>208</td>
<td>6166</td>
</tr>
</tbody>
</table>
Co-operation and co-authorship

Table 7 lists the countries and Table 8 the institutional affiliations of the contributing authors.

Table 9: Countries of ALGODAN authors according to Scopus (>4)

<table>
<thead>
<tr>
<th>COUNTRY</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finland</td>
<td>425</td>
</tr>
<tr>
<td>United States</td>
<td>79</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>28</td>
</tr>
<tr>
<td>Germany</td>
<td>26</td>
</tr>
<tr>
<td>Italy</td>
<td>20</td>
</tr>
<tr>
<td>Sweden</td>
<td>19</td>
</tr>
<tr>
<td>Japan</td>
<td>18</td>
</tr>
<tr>
<td>Spain</td>
<td>16</td>
</tr>
<tr>
<td>France</td>
<td>15</td>
</tr>
<tr>
<td>Canada</td>
<td>10</td>
</tr>
<tr>
<td>Chile</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COUNTRY</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greece</td>
<td>10</td>
</tr>
<tr>
<td>Poland</td>
<td>10</td>
</tr>
<tr>
<td>Netherlands</td>
<td>9</td>
</tr>
<tr>
<td>Denmark</td>
<td>9</td>
</tr>
<tr>
<td>Australia</td>
<td>9</td>
</tr>
<tr>
<td>Belgium</td>
<td>6</td>
</tr>
<tr>
<td>Israel</td>
<td>5</td>
</tr>
<tr>
<td>New Zealand</td>
<td>5</td>
</tr>
<tr>
<td>Norway</td>
<td>5</td>
</tr>
<tr>
<td>Switzerland</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 10: Institutional affiliations of ALGODAN authors according to Scopus (>4)

<table>
<thead>
<tr>
<th>Affiliation</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Helsinki</td>
<td>342</td>
</tr>
<tr>
<td>Aalto University</td>
<td>146</td>
</tr>
<tr>
<td>Helsinki Institute for Information Technology</td>
<td>23</td>
</tr>
<tr>
<td>Stony Brook University State University of New York</td>
<td>17</td>
</tr>
<tr>
<td>Universidad de Chile</td>
<td>12</td>
</tr>
<tr>
<td>Helsinki University Central Hospital</td>
<td>10</td>
</tr>
<tr>
<td>Osaka University</td>
<td>9</td>
</tr>
<tr>
<td>University of Texas at Arlington</td>
<td>8</td>
</tr>
<tr>
<td>University of Helsinki Institute of Biotechnology</td>
<td>8</td>
</tr>
<tr>
<td>Massachusetts Institute of Technology</td>
<td>8</td>
</tr>
<tr>
<td>Lunds Universitet</td>
<td>8</td>
</tr>
<tr>
<td>VTT Technical Research Centre of Finland</td>
<td>8</td>
</tr>
<tr>
<td>Yahoo Research Barcelona</td>
<td>8</td>
</tr>
<tr>
<td>University of Athens</td>
<td>7</td>
</tr>
<tr>
<td>King’s College London</td>
<td>7</td>
</tr>
<tr>
<td>IT-Universitetet i København</td>
<td>7</td>
</tr>
<tr>
<td>University of Arizona</td>
<td>7</td>
</tr>
<tr>
<td>Carnegie Mellon University</td>
<td>6</td>
</tr>
<tr>
<td>Jozef Stefan Institute</td>
<td>5</td>
</tr>
<tr>
<td>European Commission Joint Research Centre, Ispra</td>
<td>5</td>
</tr>
</tbody>
</table>