Final view of Algodan/FDK 2011-2013

Esko Ukkonen
Director of CoE
Scientific goals of the centre

- The Algorithmic Data Analysis CoE develops new concepts, algorithms, principles, and frameworks for data analysis.

- The work combines strong basic research in computer science with interdisciplinary work in a variety of scientific disciplines and industrial problems.

- Theory  <=>  Applications
Four main research themes

- Sequence analysis (S)  
  cgccgagtgacagagacgctaatcaggctgt
  gttctcaggatgcgtac...

- Learning from and mining structured and heterogeneous data (L)

- Discovery of hidden structure in high-dimensional data (D)

- Foundations of algorithmic data analysis (F)
Evolution of research teams (state of 2012 vs 2013)

- **Combinatorial pattern matching**
  - Ukkonen, Mäkinen (-12/2011), Kärkkäinen, Lemström, Polishchuk, Yangarber, 4 postdocs, 8 PhD students

- **Data mining: theory and applications**
  - Mannila (-2/2012), Hollmen, Koivisto, Kaski, Puolamäki, 2 postdocs, 6 PhD students

- **Pattern and link discovery**
  - Toivonen, 1 postdoc, 7 PhD students

- **Machine learning**
  - Kivinen, Rousu, 1 postdocs, 3 PhD students

- **Neuroinformatics**
  - Hyvärinen, Hoyer, 4 postdocs, 3 PhD students

- About 70 / 60 persons in total
Funding

- Basic funding from the Academy of Finland (2010-2013): 520 k€ / year
- Basic funding from the home universities: 300 k€ / year
- Home universities: infrastructure, salaries
- Academy: researcher positions
- Project funding: Academy; TEKES; EU; NIH; private foundations; industry; ...

- Exit funding from UH for 2014-15
## Scientific activity & progress: indicators

<table>
<thead>
<tr>
<th></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><strong>Journal + conf publications + books</strong></td>
<td>40+68</td>
<td>34+46+1</td>
<td>34+75</td>
<td>31+52</td>
<td>29+51</td>
<td>32+54</td>
<td></td>
<td>200+345+1</td>
</tr>
<tr>
<td><strong>Other publications + artistic work</strong></td>
<td></td>
<td></td>
<td>6+0</td>
<td>20+1</td>
<td>4+10</td>
<td></td>
<td></td>
<td>30+11</td>
</tr>
<tr>
<td><strong>PhD degrees</strong></td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>8</td>
<td>7</td>
<td>3</td>
<td>40</td>
</tr>
<tr>
<td><strong>External funding (incl. Academy) k€</strong></td>
<td>2 038</td>
<td>2 160</td>
<td>2 033</td>
<td>1 691</td>
<td>1 157</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Foreign personnel</strong></td>
<td>9</td>
<td>19</td>
<td></td>
<td></td>
<td>21</td>
<td></td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>
Researcher career development

- Algodan PhDs (2008-14: 40 persons)
  - Academia in Finland: 9 (HIIT, FIMM, UH, Aalto, other CoEs)
  - Academia abroad: 10 (Boston, CalTech, MPI, Berkeley, INRIA, KTH)
  - Industry in Finland: 18 (NSN, Rovio, …)
  - Industry abroad: 10 (IBM, Google)


<table>
<thead>
<tr>
<th></th>
<th>2008</th>
<th>2009</th>
<th>2012</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prof &amp; Senior researcher</td>
<td>13</td>
<td>15 (0 females)</td>
<td>13 (0)</td>
<td>14</td>
</tr>
<tr>
<td>PostDoc</td>
<td>16</td>
<td>19 (2)</td>
<td>12 (2)</td>
<td>12</td>
</tr>
<tr>
<td>PhD student</td>
<td>26</td>
<td>32 (7)</td>
<td>27 (2)</td>
<td>23</td>
</tr>
<tr>
<td>Student</td>
<td>15</td>
<td>20 (5)</td>
<td>18 (3)</td>
<td>7</td>
</tr>
</tbody>
</table>
Collaboration in applications: Bioinformatics, neuroinformatics

- International & European Union projects
  - EU-Project: Systems biology of colorectal cancer (J Taipale)
  - European Bioinformatics Institute, UK: Dr Alvis Brazma
  - Center for Neurobehavioral Genetics at the University of California Los Angeles (UCLA)
  - S Luyssaert & I Janssens, Univ Antwerp (carbon balance)

- University of Helsinki:
  - CoE on Translational Genome-Scale Biology: J Taipale, L Aaltonen
  - CoE in Microbial Food Safety (A Palva)
  - prof Sakari Knuutila (genetics), prof Liisa Holm (bioinformatics), prof. A Urtti (pharmacology), P Hari & E Nikinmaa (forestry)
  - Institute for Molecular Medicine in Finland (FIMM) and National Institute of Health and Welfare (THL)
  - CoE in Experimental and Computational Developmental Biology: J Jernvall

- Aalto University
  - CoE on systems neuroscience and neuroimaging (Riitta Hari, S Vanni)

- VTT Biotechnology:
  - prof H Söderlund, prof M Penttilä (CoE)
Collaboration in applications: Environmental research

- University of Helsinki:
  - CoE on Metapopulation research: prof I Hanski
  - CoE on Physics, Chemistry and Biology of Atmospheric Composition and Climate Change: prof M Kulmala
  - ESO project with astronomers: prof. K Mattila
Collaboration in applications: Linguistics and language technology

- University of Helsinki
  - CoE on Language Variation and Changes: prof T Nevalainen
  - Univ Helsinki: prof. K Koskenniemi (computer linguistics), L Carlson (computer linguistics)

- Research Institute for the Languages of Finland:
  - prof R-L Pitkänen

- European Commission's Joint Research Centre (JRC, Ispra), EC Frontex Agency, Global Health Security Initiative (GHSI), European Center for Disease Control (ECDC), Russian Academy of Sciences
Algodan 2.0: Modern challenges of data analysis

1. **Structured data challenge:**
   - Data has structured forms such as graphs and strings
   - E.g., biological, neural, social, Internet data

2. **Representation challenge:**
   - Data has implicit structure to be discovered and represented
   - E.g., Bayesian networks, compression, indexing, discrete ICA

3. **Decentralisation challenge:**
   - Data is physically distributed and highly dynamic
   - E.g., Internet, social networks, sensor networks
Algodan 2.0 objective 1/4

Find efficient representations for graphs and strings

E.g.
- Improve indexability, compressability, or locality of data
- Reveal cues about the structure or dynamics of the data
- Operate in probabilistic settings, and in a decentralised manner
Find implicit structure hidden in a data matrix

E.g.

Develop a theory of statistically independent components for discrete-valued data

Learn the directed causal structure between the observables

Combine statistical optimality with computational efficiency
Algodan 2.0 objective 3/4

Analyse dynamic data that is physically distributed over a network

E.g.

Decentralised analysis of partial, local and dynamic data in real time

Apply theory of local algorithms in real-world networking

Find compressed representations for partial results that admit consistent aggregations
### Apply data analysis in fields of technical, scientific, or social importance

<table>
<thead>
<tr>
<th>E.g.</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Networking</strong>: maintain availability of services, allocate resources dynamically and efficiently</td>
<td></td>
</tr>
<tr>
<td><strong>Computational biology</strong>: develop new models for gene regulation that give better predictions</td>
<td></td>
</tr>
<tr>
<td><strong>Neuroinformatics</strong>: find causal connection directions in brain imaging data</td>
<td></td>
</tr>
<tr>
<td><strong>Social media analysis</strong>: identify information propagation patterns in highly dynamic social media</td>
<td></td>
</tr>
<tr>
<td><strong>Computational creativity</strong>: make computers creative beyond their programmed mechanisms</td>
<td></td>
</tr>
</tbody>
</table>
"Big Data” in Algodan

- "Too big or fast to handle with conventional means”
- Efficiency and scalability has always been a concern
  - Algorithmics for pattern matching, data mining, theory, …
  - Very Large Data Bases (VLDB) conference series since 1975

- Our vision of the big change: decentralisation
  → Focus on (1) analysis of (2) decentralised big data
  - (not e.g. on distributed computing using MapReduce)

- NB: The first IEEE BigData conference to be held in 2013,
  H. Toivonen in the Programme Committee