ALGODAN

Algorithmic Data Analysis

Final view of Algodan/FDK 2011-2013

Esko Ukkonen
Director of CoE

Aalto University

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Scientific goals of the centre

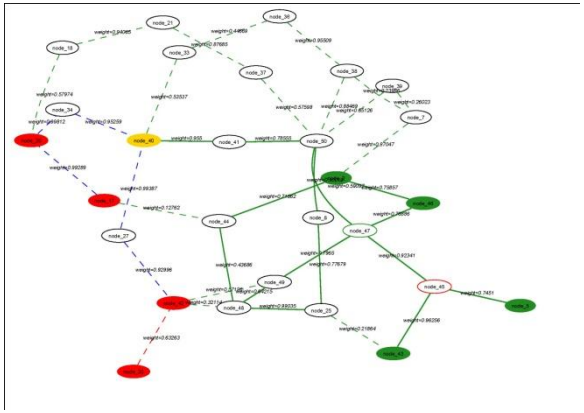
- The Algorithmic Data Analysis CoE develops new concepts, algorithms, principles, and frameworks for data analysis.
- The work combines strong basic research in computer science with interdisciplinary work in a variety of scientific disciplines and industrial problems.
- Theory <=> Applications

Four main research themes

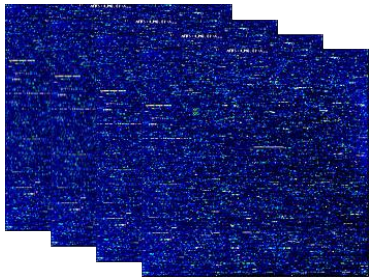
- Sequence analysis (**S**)

cgccgagtgacagagacgctaatcaggctgt
gttctcaggatgcgtac…

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

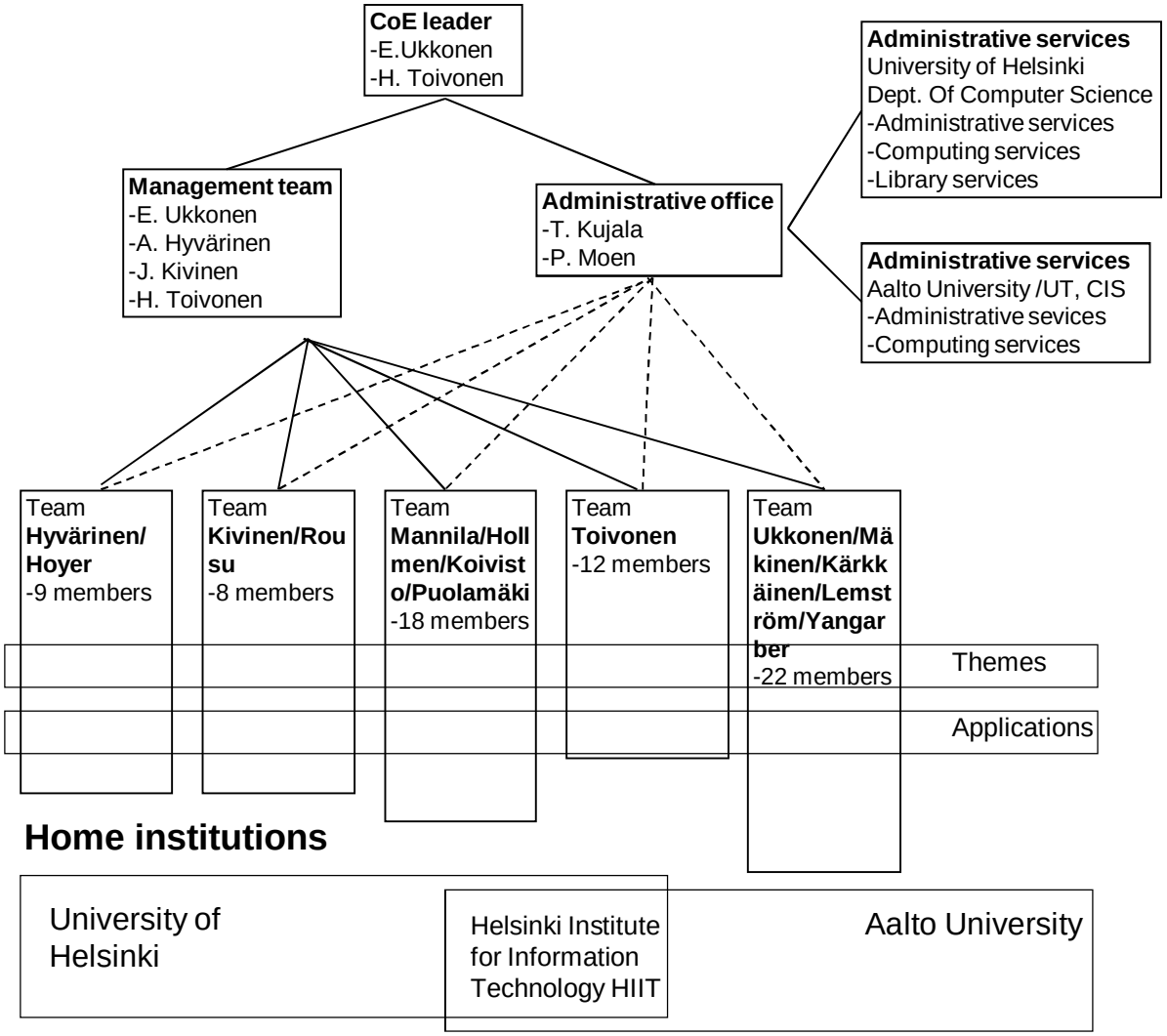


- Discovery of hidden structure in high-dimensional data (**D**)



- Foundations of algorithmic data analysis (**F**)

Organization chart of Algodan



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: infrastucture, salaries
- Academy: researcher positions
- Project funding: Academy; TEKES; EU; NIH; private foundations; industry; ...

- **Exit funding from UH for 2014-15**

Scientific activity & progress: indicators

	2008	2009	2010	2011	2012	2013	2014	Total
Journal + conf publications + books	40+68	34+46+1	34+75	31+52	29+51	32+54		200+345+1
Other publications + artistic work				6+0	20+1	4+10		30+11
PhD degrees	7	7	5	3	8	7	3	40
External funding (incl. Academy) k€	2 038	2 160	2 033	1 691	1 157			
Foreign personnel	9	19			21	20		

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)
- New professors: Veli Mäkinen (2010), Juho Rousu (2012), Petteri Kaski, with ERC! (2012), Aris Gionis (2012), Mikko Koivisto (2013)

	2008	2009	2012	2013
Prof & Senior researcher	13	15 (0 females)	13 (0)	14
PostDoc	16	19 (2)	12 (2)	12
PhD student	26	32 (7)	27 (2)	23
Student	15	20 (5)	18 (3)	7

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
  - CoE on Developmental Biology: prof. M Fortelius, prof. J Jernvall
  - 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

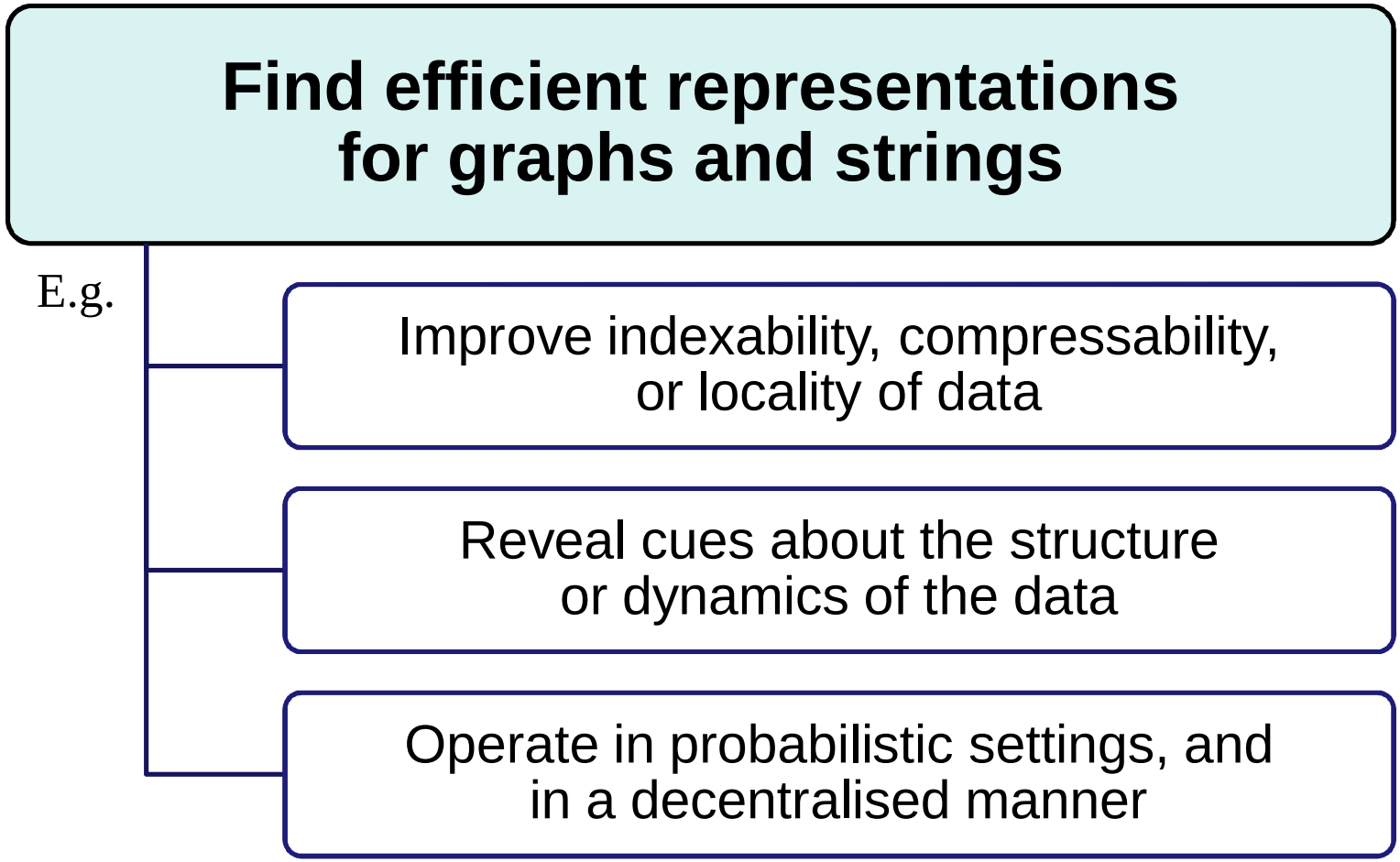
2. *Representation challenge:*

Data has implicit structure to be discovered and represented	E.g., Bayesian networks, compression, indexing, discrete ICA

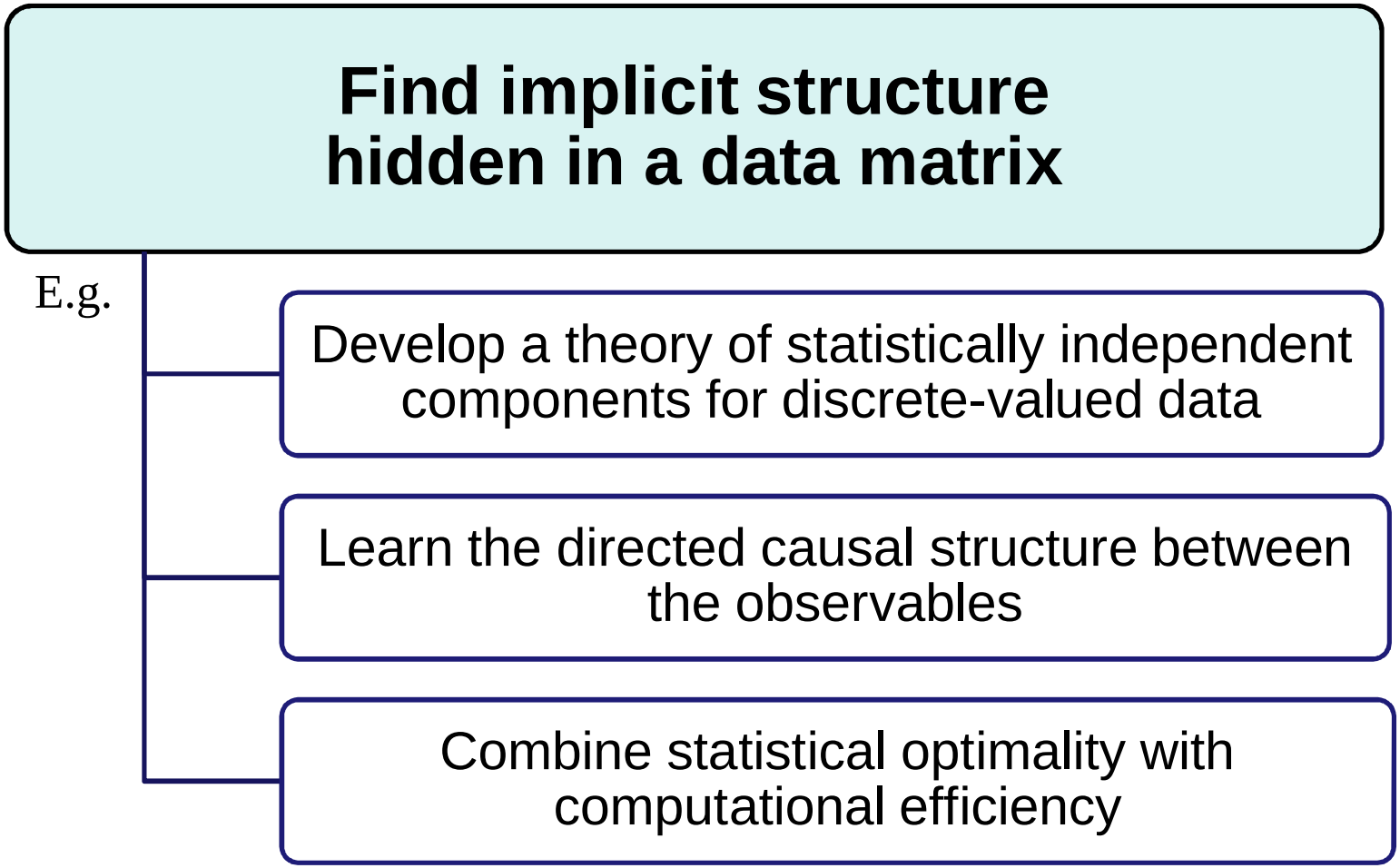
3. *Decentralisation challenge:*

Data is physically distributed and highly dynamic	E.g., Internet, social networks, sensor networks

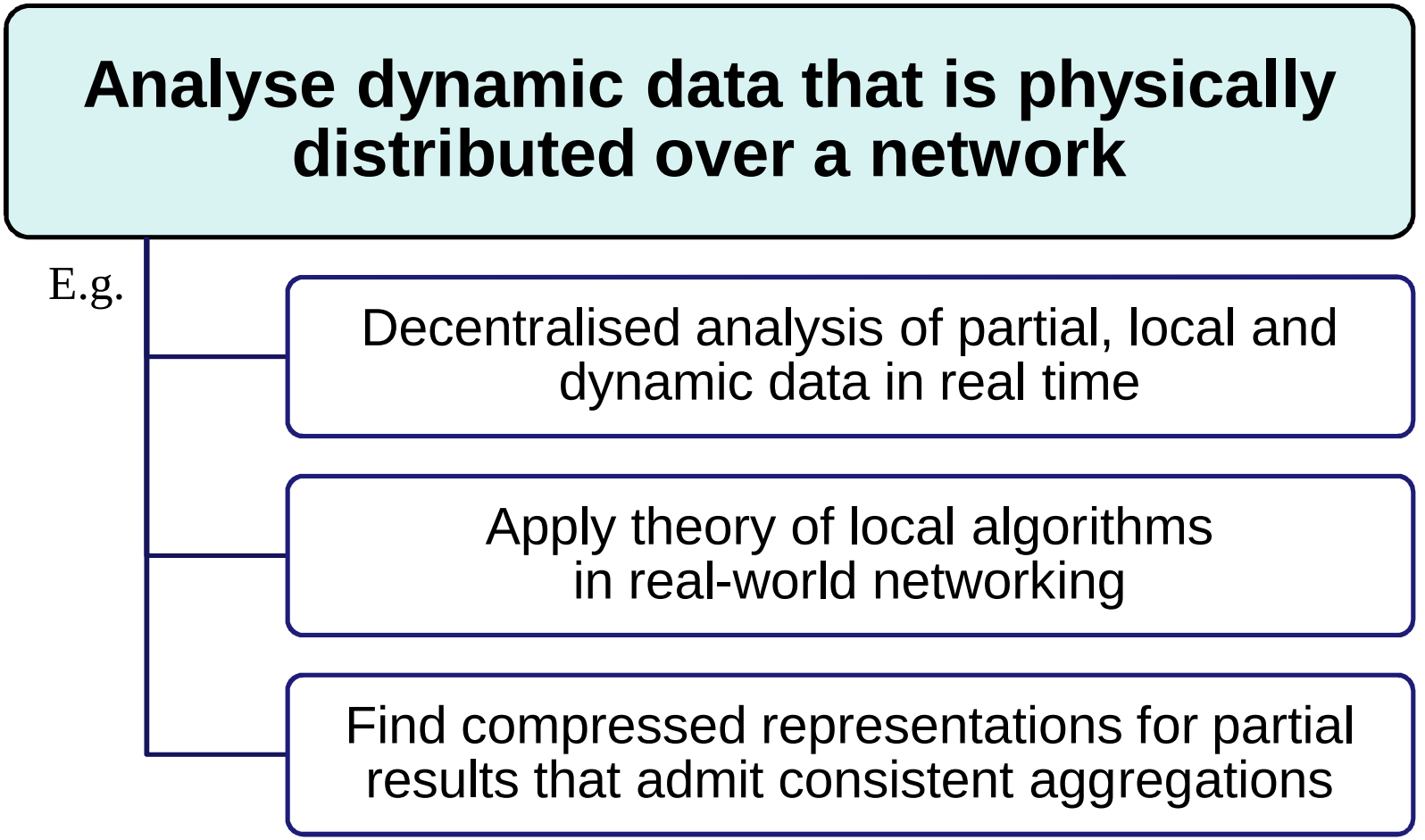
Algodan 2.0 objective 1/4



Algodan 2.0 objective 2/4



Algodan 2.0 objective 3/4



Algodan 2.0 objective 4/4

Apply data analysis in fields of technical, scientific, or social importance

E.g.

- **Networking**: maintain availability of services, allocate resources dynamically and efficiently
- **Computational biology**: develop new models for gene regulation that give better predictions
- **Neuroinformatics**: find causal connection directions in brain imaging data
- **Social media analysis**: identify information propagation patterns in highly dynamic social media
- **Computational creativity**: make computers creative beyond there programmed mechanisms

”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